Impact of open data on living conditions in the cities

By
Rezo Heorhadze

Supervised by Dmytro Krukovets PhD

June 2025

CONTENT

ABSTRACT

With the continuation of the war in Ukraine, cities struggle to find resources for day-to-day operations, not to mention post-war reconstruction. One of the key ways to improve the situation and facilitate the city's quality evolution is through data-driven decision-making, which will enable us to achieve more with less. Many cities worldwide, including Ukraine, have already created municipal Open Government Data (OGD) portals based on the open-source platform CKAN to harness the benefits of openness. Due to the limited number of research studies in this domain and the scarcity of quantitative analysis on the impact of open data, cities are struggling with efficient and effective data management and understanding of data quality.

Using an automated data gathering and analytical framework that gathered datasets from OGD platforms of European cities and consolidated them into a general dataset containing features of the downloaded datasets. Using cluster analysis, it was experimentally demonstrated that objective properties of data can be utilized as a supplement to existing open data quality assurance frameworks based on metadata analysis. The limitations of the research scale produced inconclusive results regarding the quantitative calculation of the impact of open data on living conditions in European cities.

Key words: open data, data quality, quantitative analysis, OGD portal
Word count: 11452

INTRODUCTION

The Open Government Data (OGD) initiative is gaining momentum among cities worldwide to enhance transparency. According to Karolis Granickas, access to public information and open government data has preventive effect on corruption, and recognised as undisputed public good that improves government management quality and introduces infrastructure for broader citizen participation and improve quality of decision making (Granickas 2014, 2), tap into possible economic effects of easily accessed and free-to-use data, “Open data—public information and shared data from private sources—can help create $3 trillion a year of value in seven areas of the global economy.” (Manyika et al. 2013, 1), improving public service, through smoothing inter govermental interaction on all levels and speeding up intrinsically slow local or state government entities. This approach is widely recognized as a foundational element of modern democratic governance. (House of Commons 2014, 15) Moreover, many other beneficial aspects of opening government data to the general public.

Nevertheless, the privacy problem remains; not all data should be open, especially if individual privacy is at stake, and many countries are trying to keep up with modern advancements and balance openness with common sense in (Witzleb 2023, 1-2) paper review current state of legal framework of personal data protection in Austration and anticipating future encroachment into personal privacy which will grow large with each year of technological advancements.

In a more practical context, open data can be utilized to anticipate and respond to potential challenges that may impact cities and their citizens. For example, the Town of Cary, North Carolina, has implemented an advanced flood prediction and management system by partnering with SAS and Microsoft. Such predictive analytics allowed the city to mitigate the risks of flooding by actively monitoring the state of the water supply with the help of publicly available data from stream sensors and water gauges, enabling city dwellers to assess the risk of flooding and its effects on their health and property (SAS 2020). Using IoT (Internet of Things) solutions is becoming increasingly widespread with the development of low-power, low-frequency transmitters and their increased availability. More and more cities are introducing increasingly sophisticated systems to address any possible problem related to urban areas. Gathering data from its source in a controlled and continuous manner provides insight into the inner workings of the city, offering a unique perspective that creates previously unseen challenges, which require ad-hoc solutions to reap all possible benefits.

Although technological positivity can undermine the magnitude of risks, especially present in Ukraine, early adoption is the only way forward for speedy recovery and post-war reconstruction, followed by steady development.

Open data initiatives are not limited by central governments, local also jump the train of open data with varying degree of success e.g. Spain where Open Government Data (here in

after OGD) projects became the primary source for good publicity for local governments (Muñoz, Rodríguez Bolívar, and Arellano 2022, 11), Estonia where local government struggling to keep up with state level of digitalization and open data initiatives (Rajamae-Soosaar and Nikiforova 2024, 5), USA where research uncover major problems in existing approach to open data which struggles from lack of feedback from data users and public engagement. (Wilson and Cong 2020, 6). Quality engagement of stakeholders is a significant impact metric used by some researchers in this field. As was stated before, citizen participation is crucial for the creation of a meaningful impact of raw data on the real world, creating opportunities for fast adoption of sophisticated systems that are becoming in demand.

The problem of understanding how to manage open data in the most effective way pushed researchers to find ways to evaluate open data quality in order to know how to improve findability, accessibility, interoperability, re-usability described by industry as a best practice indicator for data quality (Wilkinson et al. 2016, 4) creating problems of qualitative evaluation that require a lot of resources and produce often biased results. Nevertheless, recent developments showed that OGD portals can be evaluated quantitatively using an automated framework that focuses on reading datasets metadata and comparing them with industry accepted standards (Neumaier, Umbrich, and Polleres 2016, 6), however a significant problem remains, evaluation of metadata is problematic due it descriptive nature and proneness to human error and no academic research on the matter is quite limited, it is crucial to employ dataset analysis in order to highlight the quality of qualitative research in the area of OGD.

The analytical questions that will stay in focus throughout the research

- How to distinguish “good” data from “bad” based on the objective features of the datasets?
- Is it possible to quantitatively measure the effect of open data on the city's living conditions?

This research will employ quantitative methods to ensure scalability and facilitate the objective assessment of questions. When analytical questions necessitate handling large volumes of raw data, quantitative research methods are the only feasible option.

The paper is structured as follows: the literature review will clarify the evolution of the open data initiative, the Ukrainian context of open data usage, the latest developments in this field concerning urban management, and advancements regarding the evaluation of the impact of open data; the analytical framework comprises a thorough description of the intentions and justification for the selected research design, theoretical basis, and a hypothesis; the methodological design consists of three stages: data harvesting, cluster analysis, and regression analysis, which describe in detail the approaches to automating data collection, consolidating the general dataset, and the basis for city selection; the results presenting, namely, key findings of the paper with results with graphical representation and key limitations of data harvesting methods used; the conclusion weaving research results into existing academical discussion and

present description of the results, denoting possibility for real world application and future opportunities.

LITERATURE REVIEW

The year 2009 can be considered the starting point for the open data movement worldwide. Prior governments, companies, international organizations, and institutions kept data locked. After 2009, vast amounts of data became publicly available to anyone with internet access, jump-starting new industries created based on open data, such as data journalism. New domains for scientists to research and regular citizens to analyze and hold governments accountable. Developments in information technology have made it more accessible to regular consumers, allowing for lower entry barriers. Additionally, advancements in supporting infrastructure, such as processing capacity and bandwidth, have enabled people to share information on a much greater scale, creating capacity for previously cited industries.

Such changes were made possible with the development of the Internet and computer technologies. The more physical data digitized, the bigger the pressure from the science community, large enterprises, political actors, governments, and development agencies exercises on data holders. Consensus among all actors that data can be used instrumentally to unlock economic, social, and political benefits, but not limited to it.

“Small enterprises and social enterprises seeking to innovate with public data sets; Technological communities inspired by decentralized and collaborative models of production and problem-solving in open source, focusing on government data, and believing in the value of open sharing of corporate data; Open science advocates believing that sharing data is essential for accountable research and solving complex new research challenges (Murray-Rust 2008); Political actors supporting the potential of open data for increased transparency and accountability; Governments and development agencies exploring the role of open data in a country's development. All are interested in the instrumental value of open access to data and in the economic, political, and social benefits that this will unlock. ”(Davies and Edwards 2012, 2-4)

Another separation of Open Data can be traced back to the start of the open data movement, where data was often "curated." Some data holders present it as a way to make data more accessible; in other cases, curation is used for ethical considerations, such as personal data or data that could be used by rogue actors. With regards to contemporary times, data users are craving “raw”, unedited data. In case of Ukraine open data debate is in constant risk reevaluation, currently open data is curated I found out about it during my course “R for beginners” when I decided to find insights in data records of doctor appointments from National Health Service of Ukraine (NHSU) (Колесник 2021), during my research I tried to calculate number of injuries sustained due to military activities starting from full-scale invasion but every classification with designation Y-36 - Damage caused by military

actions(Міністерство охорони здоров'я України 2021, 59) was nowhere to be found, however frustrating it is understandable why data was sanitized in such way, military risks remains acute for decades to come.

Nevertheless, open data remains a crucial amenity during russian aggression against Ukraine, allowing citizens to access up-to-date information about air-raid alerts. "Air Alarm is a mobile app that helps you stay informed about air, chemical, man-made, and other threats to the civil defence system. The app is available for download on Google Play Market and the App Store. It does not require registration, does not collect any personal data from users, and does not track geolocation."(Visit Ukraine 2025) was developed by Ajax Systems Inc. pro bono, literally saving lives. Another similar project, "DeepStateMap" (ГО «ДіпСтейтЮА», n.d.), provides detailed information about frontline changes based on OSINT, which is the second-best option for warfare threat analysis, as the Ukrainian government does not provide official data about changes on the front in GIS format. Such a solution helps people during war, but an open data application is not limited to wartime; it can also aid in the post-war reconstruction stage as a catalyst for transparency and a means of robust planning for rebuilding cities that suffered during the war (Samokhodskyi 2023). Ukraine already has a plethora of urban open data portals, some of which are configured and supported by councils themselves, such as Vinnytsia (Вінницька міська рада, n.d.). Others use a provided platform (Міністерство цифрової трансформації України, n.d.), where city councils can be found by organisation attribute.

Importance of open data is uncontested, but impact evaluation related to cumbersome in article by An Yan and Nicholas Weber, who analyzed citations to OGD portals in academic papers, which showed steady growth from 2009 to 2016, with data.gov.uk and data.gov being the most frequent sources (Yan and Weber 2018, 6-8). Additionally, the results regarding the top 10 areas of science that use the most citations to OGDs were eye-opening, not even close to the top-ranking areas. Such a gap in knowledge poses a threat to urban development. Since anything related to urban development typically operates on a decade-level scale, the price of a mistake is immeasurable, and the cost of correcting the error will take another decade. Therefore, to avoid wasting time and resources, any decision should be thoroughly peer-reviewed to identify any significant issues at the planning stage; the only way to achieve this is through transparency. Open data, being a significant carrier of transparency, should be created, provided to everyone for free, and made easily accessible, so that these cities can engage more people in participatory decision-making and produce the best possible solution under the presented circumstances.

Problems with the development of strong institutions can also be mitigated with the help of public data initiatives, in cases when all decisions are made through deep analysis of data in a fast and repeatable manner that accommodates future development in the city. Using data to predict the future can be the answer for operating over a decades-long timeframe of the city's

lifecycle. The conclusion is straightforward: the only viable approach to urban management is transparency through publicly available data and scalable predictive analytics.

Regarding the use of open data in urban areas, it encompasses all aspects of city operations and lifecycle. The open data movement presents tremendous opportunities for all city stakeholders, from transportation to public service automation. For example, a use case in the transportation domain is real-time information about public transport's geolocation, which is quite widespread among Ukrainian cities. Businesses and local governments could provide a new kind of service with the help of open data from GPS sensors. The best part is that businesses, non-profit organizations, or local governments can reuse it. One such example is the EasyWay (Eway, n.d.) service, which provides real-time information on public transport locations in more than 50 cities in Ukraine. In the case of Kyiv City, “Kyiv Digital” includes similar functionalities for public transport monitoring, essentially utilizing the same data as EasyWay, but not limited to it.

“The best example of open government data promoting public-private-people partnership, according to R4 is the app Waze (outsmarting traffic application). It is a two-way exchange that affects people in general. When the pope went to Rio de Janeiro, agents of the city hall informed trough Waze about the roads that would be blocked for the app developers do not include those in the routes to the users. In exchange, Waze began to inform the center about the fluidity of streets, accident reports, and pavement and road damage reports, among other things. This is very useful in the daily activities, and illustrates one of the impacts of technology in management. By using a private application, citizens are consuming government information without knowledge of that and they are unconsciously doing an act of city management.” (Pereira et al. 2016, 9) Transportation is a significant area of urban management, and as presented in the case of Rio de Janeiro's usage of the Wase app that consolidates data from different sources into a single map, moreover, allowing sharing of data from user to user and technically can be independent from a centralized municipal data store about road.

“The reliability of information in both ways (from government and from external actors) must be considered to avoid the misuse of data or waste of resource in a non-emergency situation. The digital divide is also a concern, considering that everyone must receive and understand government alerts in a critical situation.”(Pereira et al. 2016, 14). With the development of more advanced technologies based on reinforcement learning and general transformers, rogue actors can influence the perception of information, particularly in cases where policies are overly permissive. Such risks are becoming increasingly apparent with each passing year, making it harder and harder to strike a balance between openness and risk management.

Another example is E-Democracy, which utilizes transparent data. Citizens can directly influence local authorities through the mobile application, making management more transparent for residents. KyivDigital is an example of an app where the local city government can create questionnaires to gather information from locals about their thoughts on the

initiative. Based on my brief evaluations, an average petition gathers around 2,000 digital signatures, removing barriers and intermediaries to promote democracy. Usage of E-democracy is still limited, due to trust problems and security concerns; therefore, the questionnaire's scope is limited to specific areas, in which case open data can be a significant enabler, because the higher the degree of transparency leads the higher levels of trust in the systems and as a consequence pushing us closer to direct democracy atleast on the level of city.

However, to achieve the required level of trust, future developers will need to develop a new domain of programming that is as mature as the Saturn 5 rocket guidance computer program, but within the scope of security. The only way forward for this is absolute, unconditional transparency.

Sustainable development in the era of big data urban planning is an emerging trend that enables such analysis methods as:

- Visualisation
- Geospatial analysis
- Machine learning
- Data Mining (extracting hidden patterns from data)

Each of these methods creates relatively significant challenges for implementation, infrastructure for such analysis tools requires a lot of human, computational, and other resources, which are rather inaccessible to poorer cities. However, major cities absolutely need them because the bigger the city is, the more data is generated, and human capabilities have obvious limits. For example, transportation data helps to optimize traffic therefore reduce congestion and improve public transport, infrastructure monitoring with predictive maintenance allows city services to reduce downtime with timely repairs, housing market analysis enables informed housing policies future prediction about real estate supply and demand (Santanu et al. 2025, 4) Main benefit of data driven approaches in urban planning and management can be summarized as follows, cost reduction, the better and more sustainable decision making which will be utmost importance during post war reconstruction, different think tanks, consulting group and international organisation presenting different visions about how war will end in short and long term prediction.

A significant gap often exists between the emergency relief phase and the onset of reconstruction. This “gap” can cause delays in decision-making required for rapid response to challenges created by post-war response, misallocation of resources, and missed opportunities for sustainable development. Implementing long-term planning into the initial emergency response is crucial for ensuring smoother transitions and more resilient outcomes. So, in order to streamline the process of early adoption of data-driven planning and management, open data and sound data management practices will lay the foundation for fast and predictable reconstruction. (Lloyd-Jones and, as well as Lockpracticalt, the University of Westminster 2006, 7). Another essential requirement for post-war reconstruction is community engagement and the effective and sustainable development of urban areas in the future. Due to

the urban scale, it is usually challenging to engage a significant number of participants in a meaningful way, especially in a timely manner and without overspending. A solution for such a problem could be curated open data that is constantly updated and constantly in use by a significant part of the city population. In this case, stakeholders are already familiar with the topic, and facilitators do not have to spend a considerable part of their time on bringing everyone up to speed. Also, community engagement can be more streamlined with the help of OGD solutions, one such case in DISPAS(https://www.dipas.org/en), open-source solution for digital participation, a successful case of Humburg that allows citizens to access digital maps, 3D models, and other geospatial data from their smartphones, on which city dwellers can precisely map their feedback, creating infinitely scalable.

Digital participation by citizens is a significant milestone in urban development, with unlimited potential in the contemporary world and beyond. A platform with curated data enables users to influence decisions made regarding city development directly. In the Ukrainian case, uninterrupted access to participatory decision-making enables cities to tap into the untapped potential of local people and their street knowledge, thereby accelerating post-war recovery more effectively. War has disrupted traditional governance structures, dislocated populations, and created a pressing need for rapid yet inclusive reconstruction. The solution to this challenge would be widespread digital participation through electronic interfaces, fostering inclusivity and lowering the bar for engagement from citizens and directly affected stakeholders.

Moreover, direct inclusion of citizens at the initial planning stages provides direct and effective feedback from those on the ground. Additionally, hidden benefits may include the fact that the feeling of involvement in the planning process will incentivize people to care for, support, and maintain the selected plan.

Open data can become a tool to achieve a sustainable and rapid reconstruction, as speed will be of utmost importance, as it could affect the number of Ukrainian refugees returning to Ukraine.

Another positive effect is facilitation of transition to smart city in the book review (Hay 2019, 1-2) Author stresses pros of smart cities like sustainability, facilitation of a circular economy, cutting time on emergency response, transport regulation, availability of education, efficient usage of resources on city scale and cons loss of jobs after implementation of “super automation”, security risks of wide implementation of IT solutions into city, exposure of smart cities to emergencies related to resources that providing daily operation of city IT solutions etc.

The implementation of smart city techniques presents numerous cybersecurity risks, particularly for critical infrastructure at all levels, including hardware. We cannot be sure that the supplier has not been compromised, and there is virtually no way to confirm that the hardware is secure. The same applies to software; the only way to ensure that software does not contain malicious subroutines is to write it yourself or thoroughly review the entire code in search of such issues. Unfortunately, many software providers do not make their source code

public. This situation raises the question of how to ensure the trustworthiness of smart city IT systems. One of the solutions the book suggests is to increase the transparency of IT solutions, which could provide timely insights about what is happening with the system and allow for a more effective response. Another solution is to invest in security training for all individuals who operate smart city systems, as a disproportionately large number of cyberattacks occur with the assistance of personnel with privileged access to the system, often without their knowledge that they are aiding an attacker. However, because achieving security in the complete sense of this word is an almost impossible task, the risk management approach is more suitable for real-world applications. Furthermore, open data is the best solution for security concerns, because there is nothing to steal if access is free.

Ukraine already has a law regulating public information that includes open data, the Law of Ukraine “On Access to Public Information”.

“The key feature of public information in the form of open data is its digital format and the possibility of automated processing by electronic means. Thus, open data does not include responses to inquiries in the format of paper letters, scanned documents, tables, graphs, dashboards, interactive maps posted on websites.” (Юридичні питання відкритих даних, n.d.), but law cannot enforce good data management practices.

To effectively manage and utilize data, it must be supported by a proper infrastructure that adheres to the FAIR principle, which is a best practice in scholarly data management. Acronym translates to: Findability, Accessibility, Interoperability, Reusability.

The argument about “fairness” was formulated by (Wilkinson et al. 2016, 4), where the authors highlighted the need to improve support for the reuse of scholarly data through formulating requirements(benchmarks) for good data management. The FAIR principle puts emphasis on the readability of data by machines, meaning that data must be published in such a way that will allow computers to find, process, and reuse it in a controlled and predictable manner, but not limited by it. Machine readability is the major property that OGD portal operators and stakeholders should understand, as the amount of data produced and processed will continue to increase with each passing year, and without adaptation to the required, in this case FAIR principles are the capacity requirement for the sake of the future proofing, the foundation for everything else.

Additionally, supporting its reuse by individuals is also essential; some data cannot be and should not be machine-readable. One such example is historic data like old documents, physical maps, etc. In this case, machine readability is not required, as this type of OGD service is designed to archive and preserve information. In the context of Ukraine, numerous initiatives are digitizing artifacts from museums and even buildings for the sake of preservation, as the war has created a challenge of looting by the invading forces. (Motorevska 2025)

Nevertheless, the two most valuable arguments are that, first, well-established data stores and platforms have begun to move towards “fairness,” e.g., FAIRDOM, Dataverse, ISA, bioCADDIE, and others. Second, the notion that implementation principles are not limited by

metadata, but by data itself, should follow the same road. Presented examples of platforms that adopted FAIR principles are dedicated to a specific area; it is quite possible that a similar specialization is required for a municipal OGD portal. In this paper, we aim towards improved reusability by finding additional “metadata” within the data itself.

The EU Data Quality Guidelines completeness indicator requires validation of datasets, e.g., empty values, and the completeness indicator points towards good practice of removing duplicates. (Publications Office of the European Union 2021, 26,35). Such comprehensive guidelines on data quality emphasize the importance of data, allowing us to conclude that open data features are as important, if not more important, than metadata. The presented standardisation approach at the European Union level creates enormous possibilities for its members, as the better quality of data will directly impact academic communities, including those in the urban domain, creating endless opportunities, provided that guidelines are implemented at all levels of local governments.

Although, to confirm that wide-scale research will be needed to evaluate the effect on academic performance after implementation of these guidelines. Nevertheless, standardization at this scale will likely lay the foundations for effective quality data management.

Open data platforms available for governments and other organizations were jump-started by data.gov, created in 2009 by the US government. The success of data.gov prompted other countries to develop their own open government portals or reuse the open-source platform CKAN, on which data.gov is based. In paper (Ali, Charalabidis, and Alexopoulos 2022, 6)presented comprehensive review of popular open data platforms amongst which CKAN and DKAN(fork of the CKAN) scored highest grade, main advantages of which is customizability meaning operators of CKAN can add features as they please using mature plugin ecosystem meaning that any open source features shared in the CKAN community and can be added to your own instance of CKAN seamlessly or incase you require features that are not available in community you can write your own, such architecture allow platform to be future-proof and allowing OGD platform operators to release new features for their users driven by demand.

The API (Application Programming Interface) also contributed to high scores, as the ability to interact with the platform programmatically allows data users to save time on integration and automation. Support for deep metadata standards, such as ISA (Interoperability Solutions for European Public Administrations), which enables interoperability between open data portals and provides a benchmarking framework. Allowing for sharing and interaction with data in the most efficient way possible, saving man-hours and resources required for data mining and preparations, which are not possible without a tremendous amount of program development to customize code to accommodate all possible variance in the API framework.

The ISA defines five levels of metadata maturity for data portals presented in Table 1.

Table 1. ISA metadata maturity levels		
Level	Name	Description
1	Basic	Only minimal metadata is provided (title, description, format).
2	Structural	Metadata includes structure and content details (fields, types, schemas).
3	Semantic	Describes meaning and relationships using controlled vocabularies or ontologies.
4	Organizational	Metadata includes information about data custodians, ownership, and update cycles.
5	Legal & Policy	Metadata specifies licensing, access rights, privacy, and legal constraints.

Note: adapted and summaries based on (European Commission 2011, 4)

API is the main gateway for interoperability between OGD portals. A programmatic interface enables the seamless transfer of data between different portals on demand, allowing data to be accessed without the need for long-term storage on the part of the requester. Possibility for automation enabling data use by third parties that can build a business based on publicly available data, creating added value on the local and state levels and opportunities of interoperability also created with help of standardized API, e.g. DCAT intermediate vocabulary can be integrated in to the OGD platform back making integration easier and seamless. Not talking about intrinsic features, like detailed usage statistics that can be useful for impact evaluation and other valuable statistical data about engagement of stakeholders in public data analysis or enrichment.

The evolution of modern cities into smart cities is fundamentally driven by the need to enhance urban life through data-driven technologies. This transformation depends on the availability of data, its interoperability, and cooperation between citizens and the government. The paper places great importance on Linked Open Data (LOD). (Conde et al. 2022, 3) The article by Javier Conde and his colleagues addresses the barrier hindering the development of smart city approaches and impeding the improvement of urban life through technological innovation and availability. The significance of this paper cannot be underestimated because, through systematic analysis and practical demonstration, the authors propose an open-source, scalable framework that facilitates better publication, discovery, and utilization of LOD for urban innovation, particularly in the field of transportation. Interoperability is a significant

impediment to data sharing and interoperability, because open data movements are a relatively new development in the urban domain, a lot of cities and metropolitan areas started developing custom-tailored solutions that are serving their needs, but because open data requires infrastructure, small towns could not afford open data initiatives in this case open-source solutions enabled even the smallest of cities to join global trend and rip the benefits.

Although open source solutions have helped alleviate impediments such as finability, interoperability, and practical usage, they remain limited to larger cities and are rarely available in smaller ones. Possible solutions for smaller towns include consolidating resources and coordinating open data initiatives, which would enable data analysts and data scientists to focus on a range of small towns united under common data standards and approaches for data harvesting and curation. In this case, a small town could become a trend-setting entity; however, there is a possibility that multiple such “unions” could occur on local and international levels in this case main strategic question will be “What open data platform should we use?” this is the primary question that will continue influencing everything related to open data in the city or urban areas. The solution presented in a paper by (Conde et al. 2022, 4) is to use intermediary metadata vocabularies like DCAT, which would enable a much greater level of interoperability between OGD portals. However, this is major improvement, but in reality interoperability is highly dependent on data, what kind of data is it timeseries, is it structured as table or hash-table and the most important translation and interpretation of data classification, because currently it is wild wild west on the plains of open data initiative creation common vocabulary for descriptive classification of data still unreachable without international cooperation and coordination and development of urban specific classification of even forked version of OGD portal customized for city specific problem and time frames.

Benchmarking allows operators of data portals to evaluate the “quality” of metadata of their open data portal and act accordingly in order to improve accessibility, findability, interoperability, and reusability of their data.

Usage of open data portals are not limited to central government, local governments also moving towards openness(Wilson and Cong 2020, 6) paper attempted to find open data impact on the city using semi structured interviews amongst nine US cities, however researcher struggled to find tangible impact other insights presented themselves, like intergovernmental exchange of information which greatly helps employees, where in cases data is not publicly available process of requesting and approving must be followed which could take some time or even denied. Another impediment is the limited standardization of data formats amongst cities. Similar problems occurred during my research when, during the experiment, we tried to analyze Excel files that have .xlsx, .xlsm, .xlsb, .xls, and many more extensions.

Moreover, the problem that plagues almost all municipal data portal is engagement of the public participation, partially it a skill issue, because analysis of raw data require some data science skills on part of user, but also problem of limited resources on part of data owner,

because curation and presentation of data in human readable form require human resources and data analyst which frequently limited.

Solid research, but it is limited by its method, which is challenging to scale. To assess the impact of this methodology, the same experiments should be conducted over multiple years using robust, structured interviews It is hard to generalize these results onto other cases in a reproducible way and without changes in the methodology.

Benchmarking is an another problem of the open data portals, one of the solution presented by (Neumaier, Umbrich, and Polleres 2016, 14) is to test metadata against standart in this case DCAT(Data Catalog Vocabulary) efficient quality assessment and monitoring framework that is able to process hundreds of data portals overtime periodically. Such an approach allows for a quick assessment of the "quality" of metadata, but is limited in assessing data itself to what authors point out themselves. The paper is that files require downloading, which is very time-consuming, resource-intensive, and cumbersome at such a scale. This is the main limitation of Neumaier's paper, as the analysis of metadata imposes strict limits on the results and should not be the sole source of data quality gate, mainly due to its descriptive nature. Features and properties of the dataset should be included to achieve better results that are more closely aligned with the real world.

The paper by Wirtz et al. provided valuable insights into the state of academic research on OGD, with results showing a disproportionate use of qualitative methods in the evaluated papers, as well as an apparent number of published papers on this matter in general (Wirtz et al. 2022, 8). This disproportion with regard to the used methods is obvious, but not justified, and can be attributed to the stale time in the broader Wirtz research from 2011 to 2020. Nevertheless, data from 2020 showed that qualitative and quantitative methods are at parity in the wider adoption of open data initiatives by different actors around the world. With more and more research in these domains, we will be more confident in the management of OGD, allowing anyone to be an early adopter with its benefits and drawbacks.

A problem with quantitative research in finding the impact of open data is that the findings are limited. However, it is exists in e.g. (Chu, Dai, and Zhong 2023, 4)research performed in China attempting to find correlations between downloads and "quality of data" showing promising results using download statistics as dependent variable e.g. OGD effectiveness is more tied to practical platform features and user engagement rather than broad policy goals or economic indicators. Focus on engagement with open data and its effects sheds light on an efficient way of connecting usage and participation of citizens to real-world events and quality metrics attributed to local government entities.

However, this research is not without its drawbacks, because experiments were performed in China. Global generalization is either not possible or very limited, with a heavy reliance on official and third-party data, such as the OGD quality score, whose calculation methodology is not presented in the paper. The limitation of this paper is that it highlights widespread problems common to similar academic inquiries, including this one, such as issues

with data classification and generalization across countries and cities, which hinder the quality of the groundwork in this area. The root cause lies in a lack of standardization and modern web protection tactics, which require sophisticated bypassing tactics of even strategies.

Attempt in finding impact of open data initiatives on the cities so far were focused on qualitative methods like paper on (Neves, Neto, and Aparicio 2020, 14) with all its benefits and drawbacks, but findings in this paper consistent with my literature review where there is a lack of quantitative approaches for open data urban impact evaluation. Interestingly, during previous research, we found confirmation for the event “open data dumps” which the author attributes to open data initiatives pushing data owners to dump everything at once without considering why it is important to curate and prepare data.

However, because this paper heavily focuses on literature reviews and methodological developments, such as Randomized Controlled Trials (RCTs), it requires experimental confirmation as a theoretical model in its own right.

Nevertheless, the paper fills a considerable knowledge gap in open data research and evaluation of open data impact via a thorough literature review.

Based on this literature, I formulated several questions:

- How to distinguish “good” data from “bad” based on the objective features of the datasets?
- Is it possible to quantitatively measure the effect of open data on the city's living conditions?

Which will be verified with hypotheses:

H1: *Datasets can be grouped into discrete classes based on record size and number of fields, denoting the quality of the dataset;*

H2: *Living conditions(private households) positively correlated with record size and the number of fields available in the OGD portal;*

H3: *The correlation strength with living conditions (private households) differs for record size and number of fields.*

The main idea behind these questions is to find a tangible metric of data that is independent of metadata and cannot be affected by human error, such as record size, total field number, format, and number of fields by type. Developing such a benchmark will enable us to establish a waypoint towards which all OGD portals in Ukraine must strive, thereby generating the most benefits for city dwellers and laying the groundwork for effective smart city management. Finding quantitatively measured effects of open data will allow us to create

robust feedback loops that iteratively improve the quality of open data linked to objective metrics.

ANALYTICAL FRAMEWORK

The challenge of quantitative research of open data impacts remains. My theory is based on the simple notion that the "quality" of data can be calculated based on its features and can supplement metadata with important insights, which will be used in my research to identify effect patterns. Finding an approach that would link data "quality" to real impacts in the city would enable the creation of a feedback loop between data and its impact, answering the question of how to improve open data in order to maximize its positive effects and minimize adverse effects. In the case of Ukraine, good data management is crucial for post-war reconstruction for multiple reasons, like transparency, because it will be a major requirement for donors for example USAID, before limiting operation in Ukraine, funded transparentcities.in.ua initiative focused on evaluation of transparency index of ukrainian cities by primary metrics presented in Tables 2.

Table 2. Transparent cities initiative metric for the transparency index

Metric	Description
Open data policy	the regulatory framework governing city council operations, availability of the city's Open Data Portal, and the completeness of the open data section on the official city council website;
Publication of open data	availability of data sets included in the list approved by the municipality and required for mandatory publication under the Regulation;
Open data quality	availability of data sets included in the list approved by the municipality and required for mandatory publication under the Regulation;
Open data impact	The use of datasets in various services, the promotion of open data initiatives by city councils, and public engagement in their publication.

Note: list sourced from (Transparent Cities 2024)

Such a comprehensive report is immensely valuable for researchers and government employees. The yearly published report, starting from 2017, reveals an important trend of change and understanding of the open data movement taking hold in Ukraine, as well as its impact on cities. But the limitation of such an approach is that it requires a lot of human resources, because all the data for the yearly research is gathered manually using qualitative

methods which is problematic, because without constant funding research will not be possible and without continuity, ranking will lose its meaning and additionally limitation for generalization, because ranking covered only by ukrainian cities and I did not find at least similar initiatives in EU members states other countries. In my humble opinion, data inherently points towards quantitative methods of research, allowing for better generalization, and does not depend on qualitative methods with limited interoperability, which can create dead ends in specific research domains.

In our research, Data was gathered from European cities for the sake of generalization with regards to Ukrainian cities, movements towards European union membership remained waypoints throughout the whole history of modern Ukraine; therefore, limiting research to a single subcontinent is sufficient. Another decision was made regarding the OGD platform of municipalities CKAN open source platform was used due to its use in Ukraine for central government OGD and for municipal ones and easy to scale api integration with my automation which included such programming languages JavaScript, python and R. JavaScript was used only due to problems with cyber defence tactics on part of OGD portals like WebBrowser validation which validates that portal API used by real user and not automated program a.k.a bot e.g. Vilnius OGD portal used such a tactic for all unauthenticated users, however I tried to contact portal administrators so that they provided me with temporary credentials, which would greatly help me in my research, inquiry remained unanswered. JavaScript proved to be of great help because it allowed me to execute code on behalf of my browser. This meant that after navigating to the portal and passing the Web browser validation, I could interact with the portal's API to gather a list of all packages, also known as datasets.

Theory was based on the work of (Neumaier, Umbrich, and Polleres 2016, 14), where in their research, they focused on metadata as a carrier for the quality analysis of data in the OGD portals. However, metadata is limited by e.g. data can be miss labeled or in case of content length can be set after compression therefore they decided not to download files for further analysis to make their framework scalable where in this case research focuses on data, and therefore, a limited number of OGDare included in the research.

In research supported format for analysis were limited by:

- JSON(JavaScript Object Notation) is a text format for storing and transporting data;
- CSV(Comma-Separated Values) - a text file format for storing tabular data;
- geoJSON(geographic data structures JavaScript Object Notation) - It allows you to store spatial data, like points, lines, and polygons, along with attributes.

These three formats were used, because they are machine readable and supported by python “pandas” library that were used in my automation, other machine readable formats were excluded like Microsoft excel due to problems with analysis stage and a lot of legacy formats which were not supported by latest versions of excel or google spreadsheets, it is problematic for Vinnitsa because, rather large amount of file were using this format(~548), for others is not a problem due to limited use of this format or duplication of file for each format in which case

CSV was used. Another problematic machine-readable format is XML(eXtensible Markup Language) due to its customizable nature; it is hard to predict the notation that is used, therefore, it is hard to standardize an approach for its analysis across different OGD portals. Removal of this format did not affect research due to its limited use in the OGD of the selected cities.

Hypothesis regarding data impact on city based on results of that amount of PE in the city positively affecting data quality (Chu, Dai, and Zhong 2023, 4), therefore I assumed that effect can work both ways and in order to validate my theory I used the amount of private households (excluding institutional households) as a dependent variable as a cornerstone of a metric that shows if a city in general is doing good or bad. There is also a supported notion of open data effects in a paper by (Gurstein 2011, 8), hinting that open data has a meaningful and supportive impact on the poor and marginalized, meaning that background characteristics like the number of private households can also be interlinked with open data.

Prior consideration in literature review describing the term “data quality " indicates that previous attempts were based on metadata evaluation, which is error-prone due to its dependency on human inputs; therefore, my intention is to mitigate human factors in the evaluation of “data quality” using objective metrics of data itself.

METHODOLOGICAL DESIGN

In this section, the research methodology design will be presented in three sections: data harvesting, cluster analysis, and regression analysis.

Data harvesting

Quantitative analysis methods was chosen due to small amount researches of that kind in the domain of open data published by cities all over Europe including one Ukrainian city, in order to create an objective metric which is derived from data features like dataset size(number of records), types of fields, amount of field and date of publication. In order to validate the viability of the analytical question “Is it possible to quantitatively measure the effect of open data on the city's living condition?” and “How to distinguish “good” data from “bad” based on objective features of the data?” for that dataset was designed from automated analysis of files downloaded from Vinnitsa sister cities. A general dataset was designed containing such fields in the form of a table structure as presented in Table 3.

Table 3. General dataset design.

Field name	Description	Type
city	City name	string

Field name	Description	Type
dataset_format	Format of the downloaded file	string
dataset_id	Unique identifier of the file	string
Field name	Description	Type
date_created	Date of publication	timestamp
date_modified	Date of last modification in the file	timestamp
f_size	Size of the file in bytes	int
r_size	Number of records in the file	int
n_fields	Total number of fields	int
n_text_fields	Number of text fields,	int
n_numeric_fields	Number of numeric fields	int
n_datetime_fields	Number of timestamp fields	int
n_boolean_fields	Number of boolean fields	int
n_geometry_fields	Number of geometry fields	int
tags	Tags provided by the owner of the dataset	array[string]

Detection of field type was performed through the family of Python libraries “pandas”. Such fields “n_” were chosen because they cover the majority of data types that were available in the datasets. Data about field numbers by type and tags was gathered, but not used in this research.

To gather datasets for analysis, JavaScript and Python automation were used to bypass web protection tactics employed by some open data portals, such as DDoS protection and BotNet protection. CKAN/DKAN OGD platforms were chosen as targets due to the proliferation of these platforms in the EU and Ukraine, and similarities in programmatic interface, gathering of the data, separated into two steps:

1. Web-based metadata scraping - gathering all metadata about all available datasets, which includes the download URL on the CKAN/DKAN platform;
2. REST API downloading and analysis - previously gathered metadata used as input for a Python script, which downloads, if possible, files and analyzes them using previously presented methodology and supported data formats, CSV, GeoJSON, and JSON. The script's output creates a dataset that is used for further analysis in R.

List of cities Peterborough(England), Birmingham(England), Kielce(Poland), Burca(Turkey), Bat Yam(Israel), Panevėžys(Lithuania), Karlsruhe(Germany), Nancy(France)

Cities were chosen based on these criteria: being sister cities of Vinnytsia, as sister city agreements were typically made with cities that shared similar characteristics (Ruth 2023, 1). We found it sufficient to mitigate research bias. The pre-analysis stage consisted of removing outliers - overly small or big datasets that could skew further analysis. The limiting factor of the harvester is that it cannot distinguish duplicates of files with different formats, meaning that harvester could add the same data multiple times if the OGD portal operator decided to store various formats of the same file in the general datasets this issue is deemed acceptable for the sake of limited research scope.

Cluster analysis

Cluster analysis using k-means unsupervised classification determines how close different elements in a dataset are based on a user-defined set of *k* clusters. This approach aims to determine the distribution of datasets (each dot representing a single file or dataset) and identify discrete factors that reveal hidden groupings among the analyzed datasets. The optimal number of centroids should be determined using the elbow method. In case the elbow methods do not show a distinct “elbow” pattern, multiple experiments can be performed for different numbers of clusters.

The input dataset for cluster analysis is composed by combining all general datasets, where each observation is represented by one file downloaded from OGD, from each city, into a single dataset. This dataset is then mapped to scaled (normalized) r_size and n_fields for each dataset, and cluster analysis is performed.

After clustering is performed, another analysis should be conducted by counting the percentage of datasets in each cluster for each city.

Regression analysis

Regression analysis aims to find dependencies between the number of fields and record size, the independent variable, and the population living in private households, the dependent variable.

The dataset for regression is a design presented in Table 4.

Table 4. Dataset design used for regression analysis

Field name	Description	Type
city	City name	string
date	Year	timestamp

Field name	Description	Type
cumulative_r_size	Number of records in datasets available for the given year	int
cumulative_f_n	Number of fields/columns in datasets available for the given year	int
private_hh_perc	Percentage of private households excluding institutional properties against the population in the given year	int

Dependent variables in private_hh_perc and independent variables are cumulative_r_size and cumulative_f_n. Ideally, regression should show differentiation between what property of the dataset affects city living conditions, the number of records, or the number of fields. Dependent and independent variables should be logarithmically normalized to perform analysis on variables on the same scale.

RESULTS

Due to missing statistical data, such as living conditions in Eurostat (Eurostat, n.d.), for the initially chosen cities, problems with data accessibility, and limitations of data gathering automation, other cities were considered, including Vigo, Vilnius, Leipzig, Stuttgart, Vinnytsia, Málaga, and Rostock. Only Vigo and Malaga cities had a wide enough dataset during the continuous time frame that allowed me to proceed with further regression analysis, forcing me to reconsider the list of cities chosen for the study. Data were gathered using a harvesting framework consisting of JavaScript (JS) and Python scripts, which can be found in Appendix 1. Harvesting consisted of two stages of metadata scraping, meaning that information about all datasets was gathered into a file for each city. The second stage involves data harvesting and analysis. The links to the dataset field were cycled through and downloaded into a separate folder. Each file was renamed using the following template: <unique_identifier_of_dataset>.<file_format> After downloading is done, the script is reconfigured to analyze files of supported formats - JSON, CSV, and GeoJSON. Each file analysis result was added to the general dataset in the form of a table(CSV), the design of which is presented in Table 5.

Table 5. General dataset design.		
Field name	Description	Type
city	City name	string
dataset_format	Format of the downloaded file	string
Field name	Description	Type
dataset_id	Unique identifier of the file	string
date_created	Date of publication	timestamp
date_modified	The date of last modification in the file was not used in the analysis	timestamp
f_size	Size of the file in bytes	int
r_size	Number of records in the file	int
n_fields	Total number of fields	int
n_text_fields	The number of text fields was not used in the analysis	int
n_numeric_fields	The number of numeric fields was not used in the analysis	int
n_datetime_fields	The number of timestamp fields was not used in the analysis	int
n_boolean_fields	The number of Boolean fields was not used in the analysis	int
n_geometry_fields	The number of geometry fields was not used in the analysis	int
tags	The tags provided by the owner of the dataset were not used in the analysis	array[string]

The type of the field is distinguished using the Python “pandas” library, and the script stages are as follows:

- Loading file with metadata about datasets;
- Cycling through metadata and downloading files in the “data” folder;
- Loading downloaded file;
- Counting the number of records;

- Counting the number of fields by type and total number, evaluating the size of the file;
- Data like city, dataset_format, dataset_id, data_created, data_modified, tags are loaded from metadata;
- After the analysis is ready, the data is appended to the general dataset and saved to the file.

Files that failed to be downloaded are added to separate files for statistical analysis. Due to limitations of the harvesting framework, files with unsupported formats are not recorded into the general dataset or a separate file. Statistics regarding the gathered data and failure rate can be checked in Table 6.

Table 6. Response rate

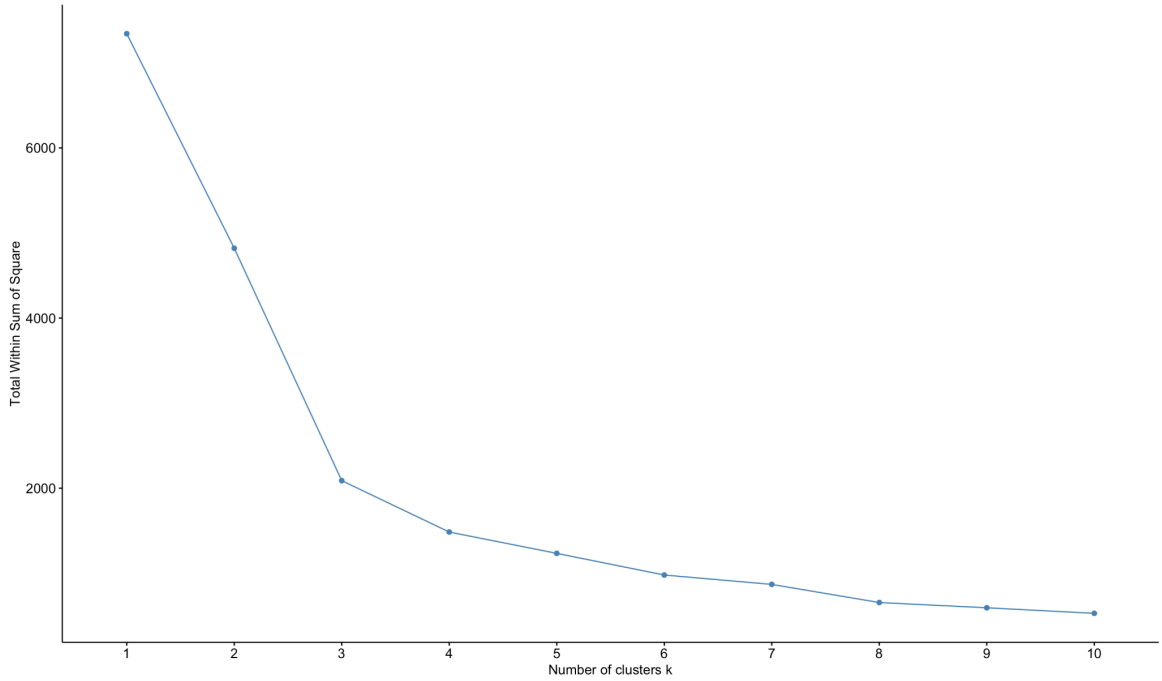
City	Total available files/datasets, n	Success rate, n/%	Unsupported format, n/%	Failed to download, n/%	Source
Vinnytsia(Ukraine)	3842	1960 / 51%	1822 / 47.5%	60 / 1.5%	https://opendata.gov.ua/
Leipzig(Germany)	539	472 / 87.6%	61 / 11.3%	6 / 1.1%	https://opendata.leipzig.de/
Rostock(Germany)	505	442 / 87.7%	56 / 11%	7 / 1.3%	https://www.opendata-hro.de/
Stuttgart(Germany)	157	88 / 56%	59 / 37.7%	10 / 6.3%	https://opendata.stuttgart.de/
Vilnius(Lithuania)	288	262 / 91%	8 / 2.7%	18 / 6.3%	https://opendata.vilnius.lt/
Malaga(Espania)	801	736 / 92%	60 / 7.4%	5 / 0.6%	https://datosabiertos.malaga.eu/
Vigo(Espania)	1111	547 / 49.3%	553 / 49.7%	11 / 1%	https://datos.vigo.org/

Note: The Total number of observations in the general dataset is 4507

After the composition of the general dataset, outliers were removed using the IQR anomaly detection method using the R script that can be found in Appendix 2. In order to find

the optimal number of centroids required by cluster analysis with the K-Means approach, the Elbow method is used, and the results are provided in Graph 1.

Graph 1. Optimal number of clusters: “Elbow Method”



Note: source General dataset, data gathered by the author in 2025

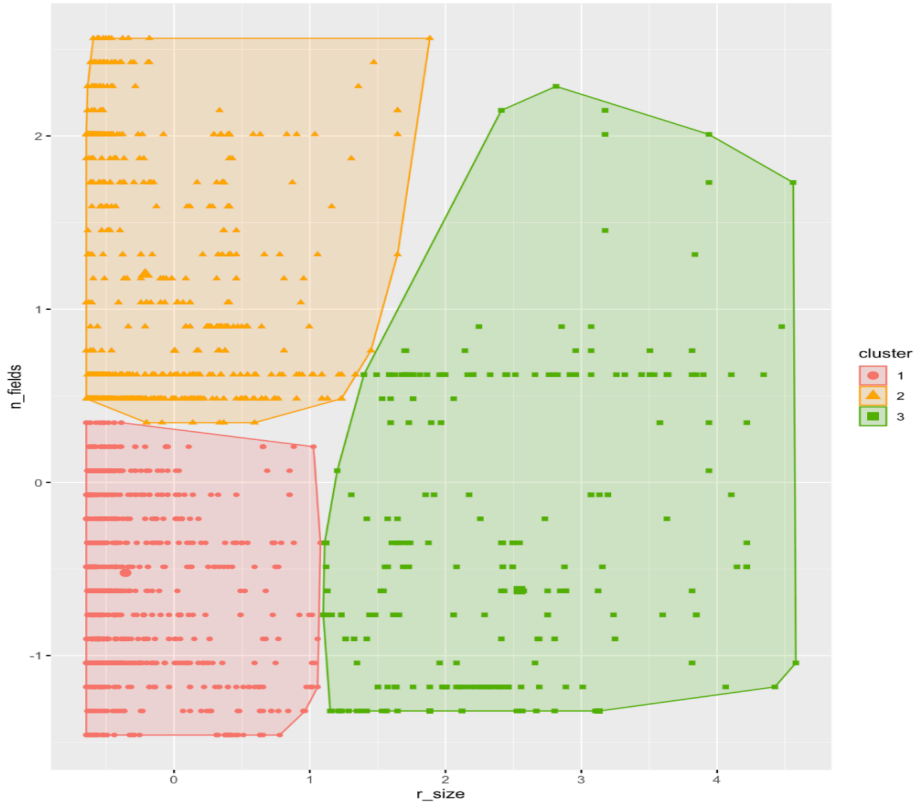
A very distinct “elbow” share line points towards the optimal number of clusters, which is 3, similar to a textbook graph, indicating with a high degree of certainty that the cluster analysis result will be without intersections.

As a result, clustering analysis was performed using the means of the factoextra library. The first plot shows the differentiation of the cities' datasets into three classes:

- 1 Class - **undesirable**, because it has a small amount of records in the dataset and simultaneously a small number of fields in them;
- 2 Class - **desirable**, datasets in this class also have a small amount of records, but drawbacks are compensated by a larger number of fields in them;
- 3 Class - **desirable**, datasets in this class have a large number of records and a decent number of fields.

A graphical representation of the result is shown in Graph 2. The graph was built based on a general dataset containing 4507 observations (datasets/files) mapped on a scaled y-axis, n_field - the number of fields in the dataset, and the x-axis, r_size - the number of records in the dataset.

Graph 2. Cluster plot

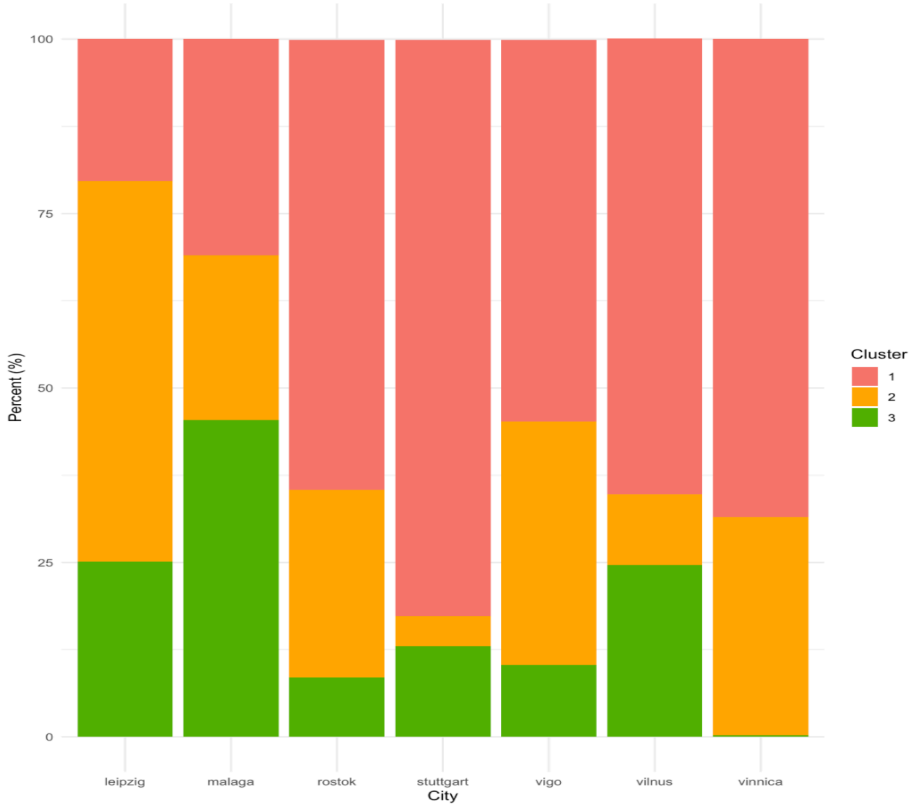


Note: source General dataset, data gathered by the author in 2025

Another significant result is the comparative analysis between cities shown in Graph 3. Where cities are mapped onto bar charts with a percentage of the dataset in each found cluster, e.g., Vinnitsa has almost no datasets in the third cluster, indicating that the municipality either does not have or has not published data with a high number of records. It can hinder the data user experience and limit the potential benefits of open data. Compared to European cities, where the number of datasets in cluster 3 is at least 8.5% in Rostock, the low percentage in Vinnitsa suggests problems with data quality published by the council, requiring further investigation to understand the root causes.

Nevertheless, the number of datasets in the second cluster (31.3%) outweighs the limited number of datasets in the third cluster, making Vinnitsa's open data quality similar to that of other cities, pointing towards similarities with other European cities in the domain of open data management. In Table 7, you can find exact statistical data on the comparative analysis of cluster analysis results among selected cities.

Graph 3. Percentage Distribution of City and Cluster



Note: source General dataset, data gathered by the author in 2025

Table 7. Percentage of datasets in each class for each city			
City	Class 1, %	Class 2, %	Class 3, %
leipzig	20.3	54.6	25.1
malaga	31	23.6	45.4
rostock	64.5	26.9	8.5
stuttgart	82.6	4.3	13
vigo	54.7	34.9	10.3
vilnius	65.3	10.2	24.6
vinnica	68.5	31.3	0.2

To prepare data for regression analysis, the general dataset was updated for regression that can be found in Table 8.

Table 8. Regression analysis dataset design.		
Field	Description	Type
city	City name	string
date	Year	timestamp
cumulative_r_size	Number of records in datasets available for the given year	int
cumulative_f_n	Number of fields/columns in datasets available for the given year	int
private_hh_perc	Percentage of private households excluding institutional properties against the population in the given year	int

In order create dataset general dataset was transformed through these stages: group number of fields and record size by year, denoting amount of records and fields available at the given year; Adding approximate population sourced from eurostat (Eurostat 2025), state statistic sites (Oficialiosios statistikos portalas 2025) and censuses ("Spanish Statistical Office - Population" 2025) to calculate the percentage of private households against population at the given year; Remove observations with empty values. The negative correlation between open data parameters and the number of households, presented in Table 9, suggests problems with data completeness and scale, particularly skewed residuals that indicate the issue of scaling the data.

Tables 9. Experiment #1					
#	Model	Residuals	Estimate	Pr(>|t|)	R-squared
1	private_hh_perc ~ cumulative_r_size	Median: 0.8528 1Q: -5.4483 3Q: 8.2311	-2.387e-07	0.00897 **	0.3084
2	private_hh_perc ~ cumulative_f_n	Median: -4.2 1Q: -6.343 3Q: 9.752	-0.001028	0.354	0.04532
3	private_hh_perc ~ cumulative_r_size + cumulative_f_n	Median: 0.9144 1Q: -5.7611 3Q: 7.8094	cumulative_r_size: -2.342e-07	cumulative_r_size: 0.0173 *	0.3093
			cumulative_f_n: -1.521e-04	cumulative_f_n: 0.8812	

Note: The Total number of observations in the general dataset is 21, used in the regression
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

An additional experiment was conducted on a dataset limited to a single city, which provided much broader data on private households. The previous experiment was then repeated using a logarithmic scale. For the sake of the new experiment, the initial dataset was to be reconfigured with new variables presented in Table 10.

Table 10. Regression analysis dataset design on a logarithmic scale

Field	Description	Type
city	City name	string
date	Year	timestamp
log(cumulative_r_size)	Number of records in datasets available at the given year on a logarithmic scale	double
log(cumulative_f_n)	Number of fields/columns in datasets available at the given year on a logarithmic scale	double
log(private_hh)	Percentage of private households excluding institutional properties against population at the given year in logarithmic scale	double

Results of experiment #2 showed better results, with an R-squared value of 0.6, indicating significant correlations in model 6 with multiple independent variables. Still, a negative correlation is problematic because it suggests that the amount of open data has a detrimental effect on the number of private households, thereby undermining decades of research in this field.

Detailed results are presented in Table 11.

Tables 11. Experiment #2

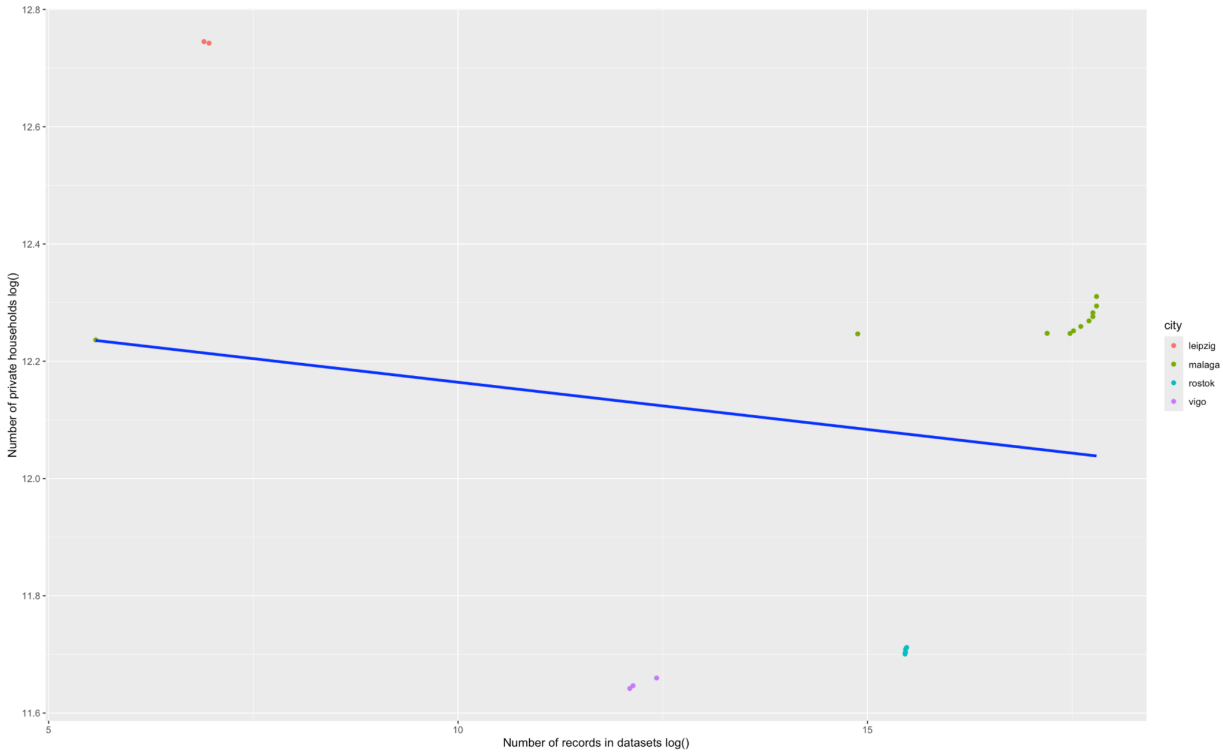
#	Model	Residuals	Estimate	Pr(>|t|)	R-squared
4	log(private_hh) ~ log(cumulative_r_size)	Median: 0.1994 1Q: -0.3725 3Q: 0.2369	0.01612	0.444	0.03118

#	Model	Residuals	Estimate	Pr(>|t|)	R-squared
5	log(private_hh) ~ log(cumulative_f_n)	Median: 0.1624 1Q: -0.2959 3Q: 0.2667	-0.0998	0.00441 **	0.3544
6	log(private_hh)~ log(cumulative_f_n) + log(cumulative_r_size)	Median: 0.06631 1Q: -0.24034 3Q: 0.17530	log(cumulative_f_n): -0.22737	log(cumulative_f_n): 2.84e-05 ***	0.643
			log(cumulative_r_size): 0.08504	log(cumulative_r_size): 0.00127 **	

Note: The Total number of observations in the general dataset is 21, used in the regression
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

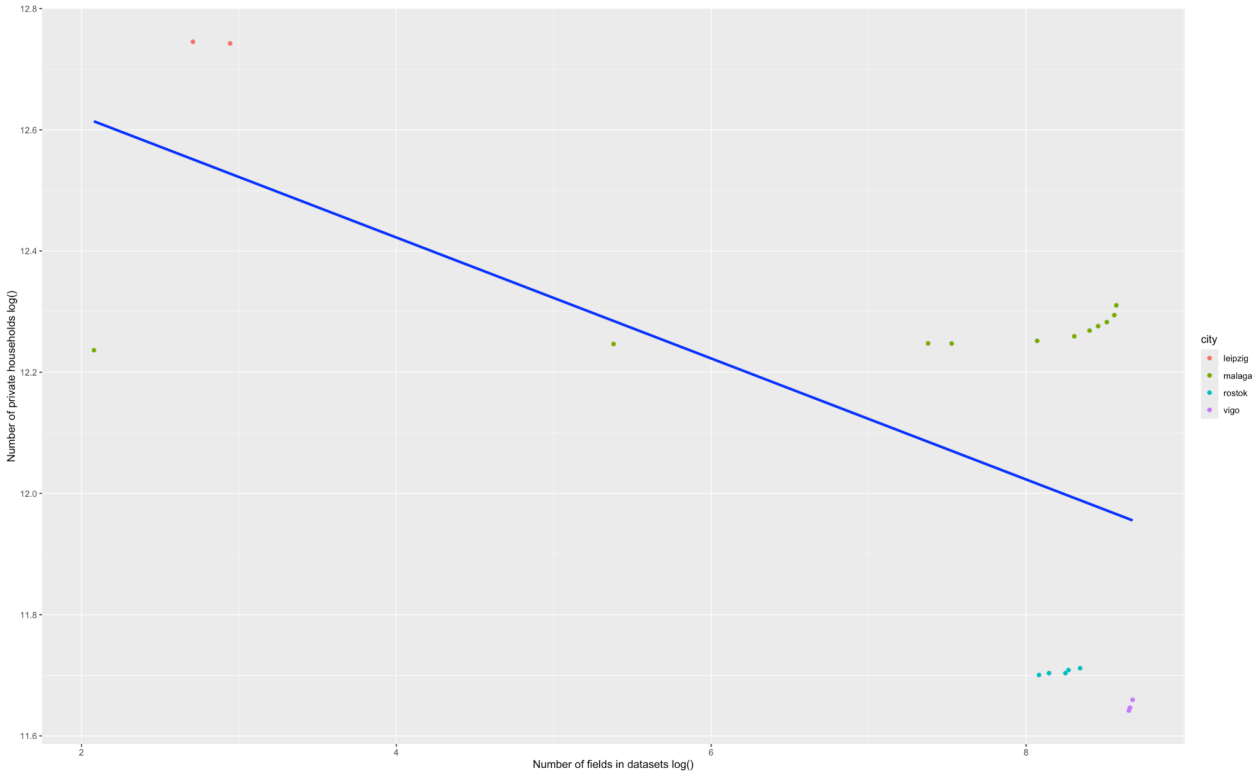
Graphical representations of models 4 and 5 can be found in Graphs 4 and 5.

Graph 4. Model #4 Regression Log() number of private households against a number of records available for the given year



Note: R-squared 0.03

Graph 5. Model #5 Regression Log() number of private households against a number of fields available in the given year



Note: R-squared 0.35

In order to validate limitations of the dataset, the same experiment was performed on a single city(Malaga), because it had the largest number of observations among the selected cities from the general dataset.

Results are presented in Table 12.

Tables 12. Experiment #3(Malaga)

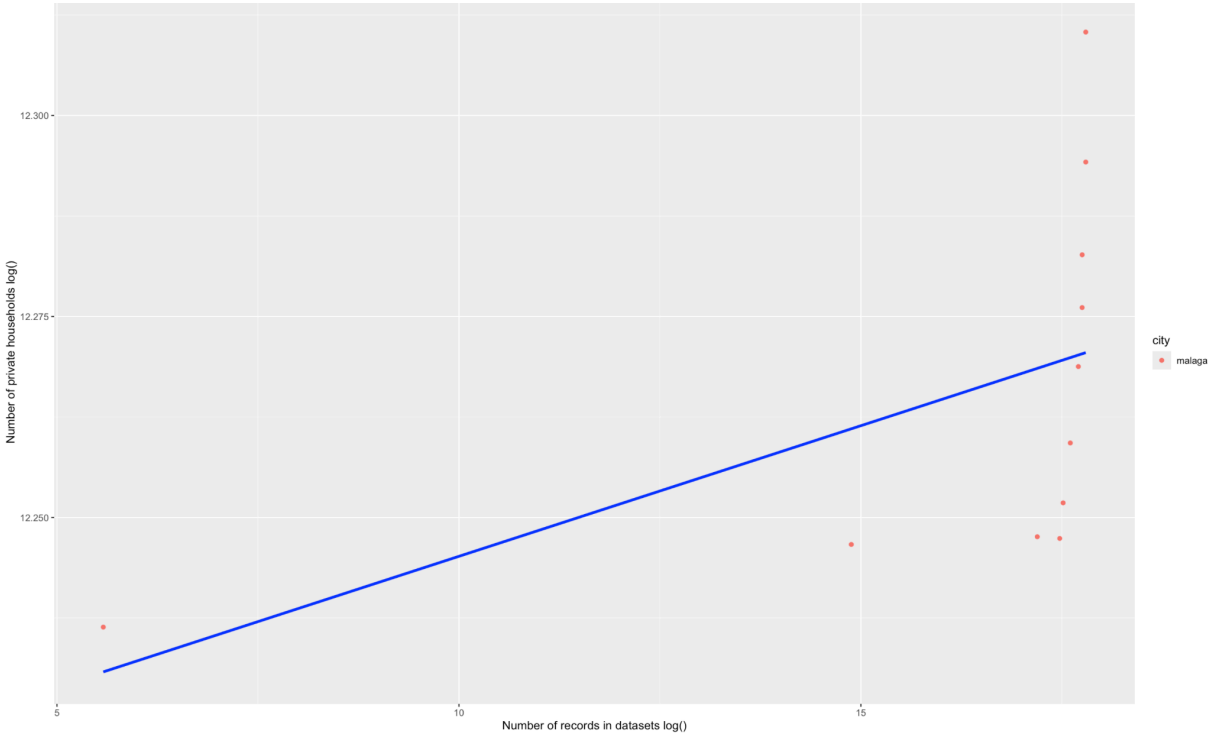
#	Model	Residuals	Estimate	Pr(>|t|)	R-squared
7	log(private_hh) ~ log(cumulative_r_size)	Median: -0.001435 1Q: -0.016079 3Q: 0.009031	0.003249	0.106	0.2633
8	log(private_hh) ~ log(cumulative_f_n)	Median: -0.003989 1Q: -0.015523 3Q: 0.009510	0.007443	0.033 *	0.4129

#	Model	Residuals	Estimate	Pr(>|t|)	R-squared
9	log(private_hh)~ log(cumulative_f_n) + log(cumulative_r_size)	Median: -0.002445 1Q: -0.009655 3Q: 0.009285	log(cumulative_f_n): 0.024157	log(cumulative_f_n): 0.0431 *	0.5719
			log(cumulative_r_size): -0.009479	log(cumulative_r_size): 0.1230	

Notes: The Total number of observations in the general dataset is 11, used in the regression
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

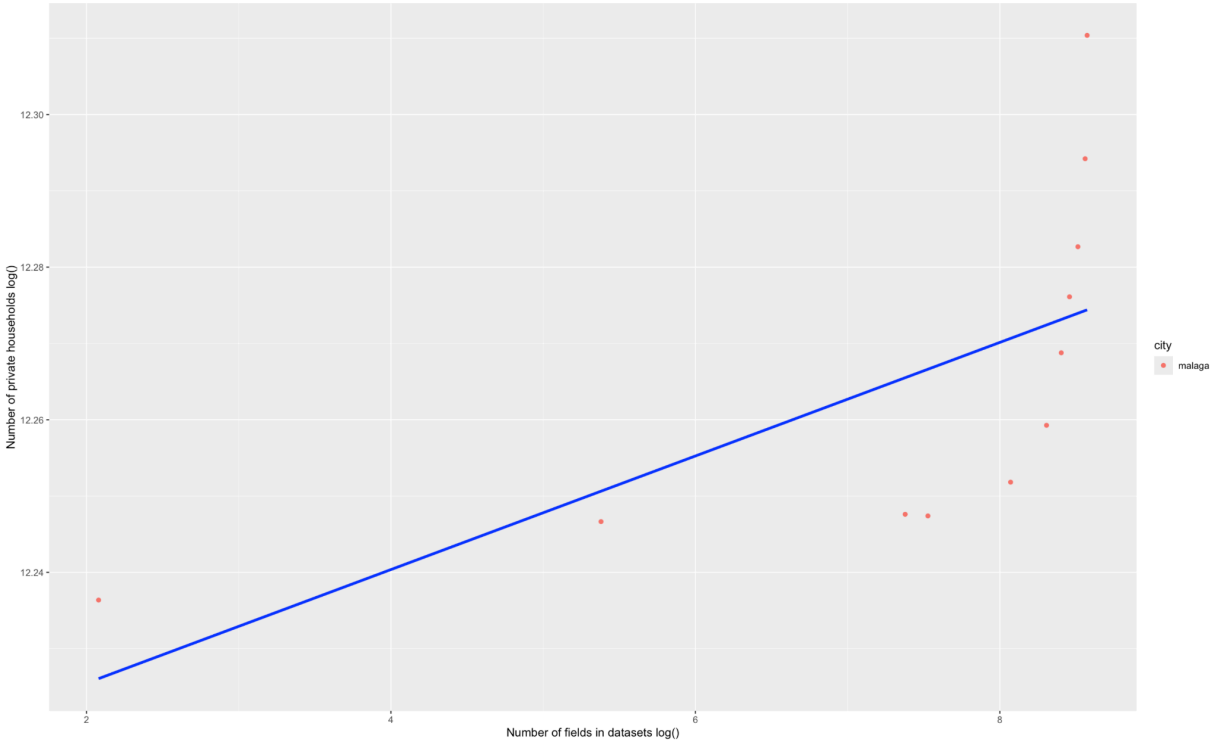
However, models 7 and 8 showed positive R-squared values of 0.26 and 0.41, indicating a weak correlation. Regression lines can be found for these models in Graphs 6 and 7.

Graph 6. Model #7 Regression Log() number of private households against a number of records available for the given year for the city of Malaga



Note: R-squared 0.26

Graph 7. Model #8 Regression Log() number of private households against a number of fields available in the given year for the city of Malaga



Note: R-squared 0.41

CONCLUSION

Research showed promising results especially in automated quality assessment of OGD portal, expanding of experiment performed by (Neumaier, Umbrich, and Polleres 2016, 12), using cluster analysis through introduction of evaluation quality of dataset via objective features of data and not metadata, which is prone to human errors, and introducing new optics on quality of dataset confirming H1 that expected hidden patterns of data based on record size and number of fields. In the presented analysis, three distinct clusters were identified and classified by desirability based on the simple notion that more data is better than less data. Another insight is that cluster analysis showed drawbacks of approach to open data management in Vinnytsia where most(68.5%) of files fall under Class 1 type which is undesirable due to the low amount of data and because Vinnytsia open data management strategy is one of the best in Ukraine according to (Transparent Cities 2024) on local level, it sheds light on what kind of problems other ukrainian city could have and because qualitative methods is hard to generalize onto other cities of the world my approach allows more precise, scalable and cheaper alternative in case more OGD platform would be supported and addition of broader support for file formats.

Although the initial plan of research had a different list of cities initially, the list used still has shared features like OGD platforms used by these cities, which is CKAN, similar timeframe(2020-2024) of file uploads meaning that used cities start open data policy implementation around same time and similar population size.

There is a problem with a considerable number of files that are not machine-readable; therefore, a high error rate is observed for some cities. However, automation is limited by the supported formats (CSV, JSON, GeoJSON). The high error rate indicates that a massive array of files is not machine-readable and, consequently, useless for data analysis. The quality of the data and the number of files point to the previously mentioned problem of an “open data dump” (Neves, Neto, and Aparicio 2020, 14). Significant limitations of the research include the small number of supported file formats for analysis, ideally all machine-readable formats mentioned in the Data Quality Guidelines (Publications Office of the European Union, 2021, p. 99), as well as more open government data (OGD) platforms, such as Socrata, ArchGIS Open Data, and others.

With regards to regression analysis in an attempt to find direct link between open data quality and background characteristics of the city, results are limited, only model #6 has R2 more than 0.6, but due to tiny sample and a lot of empty values in dataset model #6 also could not be considered insignificant. The presented results of the graphs for models 4 and 5 suggest that the experiment should have at least 100 cities analyzed to be significant. However, my time and computational resource limitations are incompatible with such a scale.

Although models 7 and 8 show promising graphical results, the R-squared values are still low, at 0.2633 and 0.4129, respectively. Inconclusive results can again be attributed to small sample sizes and missing data, such as the number of private households for certain years and cities. Therefore, I cannot definitively deny the validity of H2 for now, as more research is needed and on a much larger scale. However, the presented regression models allowed us to conclude with a high degree of certainty that the dataset record size and the number of fields with different correlation strengths with living conditions in the city confirmed H3, opening possibilities for further research. Further research is absolutely necessary, with recent enhancements in the domain of data science, it is possible to process a lot more data in a unit of time than ever before, and if possible, to employ cloud computing for the sake of scaling the scope of research, therefore improving the results.

Standardisation remains a major impediment to this kind of scientific inquiry, although a workaround is possible, but it is hardly scalable; the only proper way forward is international standardisation and classification.

REFERENCES

Міністерство охорони здоров'я України. 2021. "Класифікатор хвороб та споріднених проблем охорони здоров'я," НК 025:2021. Київ. https://www.dec.gov.ua/wp-content/uploads/2021/11/naczionalnyj-klasyfikator-nk-025.pdf.

ГО «ДіпСтейтЮА». n.d. DeepStateMAP | Мапа війни в Україні. Accessed June 1, 2025. https://deepstatemap.live/#6/49.4383200/32.0526800.

Вінницька міська рада. n.d. Портал відкритих даних Вінницької МР. Accessed June 1, 2025. https://opendata.gov.ua.

Міністерство цифрової трансформації України. n.d. Портал відкритих даних. Accessed June 1, 2025. https://data.gov.ua/.

"Юридичні питання відкритих даних." n.d. Дія.Відкриті дані. Accessed June 1, 2025. https://diia.data.gov.ua/info-center/falq.

Колесник, Микола В. 2021. "Інформація про створені та виконані електронні направлення в ЕСОЗ." Портал відкритих даних. https://data.gov.ua/dataset/005286ef-ec37-4ed4-b262-9a7597f146e0.

Ali, Mohsan, Yannis Charalabidis, and Charalampos Alexopoulos. 2022. "A Comprehensive Review of Open Data Platforms, Prevalent Technologies, and Functionalities." 15th International Conference on Theory and Practice of Electronic Governance, (November). http://dx.doi.org/10.1145/3560107.3560142.

Chu, Jinhua, You-Yu Dai, and Anyuan Zhong. 2023. “Factors Influencing the Effectiveness of Open Government Data Platforms: A Data Analysis of 61 Prefecture-Level Cities in China.” SAGE Open 13, no. 3 (August): 1-13. 10.1177/21582440231194207.

Conde, Javier, Andres Munoz-Arcentales, Johnny Choque, Gabriel Huecas, and Alvaro Alonso. 2022. “Overcoming the Barriers of Using Linked Open Data in Smart City Applications.” Computer 55, no. 12 (December): 109 - 118. 10.1109/MC.2022.3206144.

Davies, Tim, and Duncan Edwards. 2012. “Emerging Implications of Open and Linked Data for Knowledge Sharing in Development.” IDS Bulletin 43, no. 5 (September): 117–127. 10.1111/j.1759-5436.2012.00372.x.

European Commission. 2011. “Towards Open Government Metadata.” (September), 1–6. https://joinup.ec.europa.eu/sites/default/files/24/4c/14/towards_open_government_metadata_0.pdf.

Eurostat. 2025. “Population change - Demographic balance and crude rates at regional level (NUTS 3).” Eurostat. https://ec.europa.eu/eurostat/databrowser/view/demo_r_gind3__custom_16643664/default/table?lang=en.

Eurostat. n.d. Accessed June 1, 2025. https://ec.europa.eu/eurostat/databrowser/view/urb_clivcon/default/table?lang=en&category=urb.urb_cgc.

Eway. n.d. Easyway. Accessed June 1, 2025. https://www.eway.in.ua.

Granickas, Karolis. 2014. “Open Data as a Tool to Fight Corruption.” Topic Report No. 2014/04. European Public Sector Information Platform.

https://data.europa.eu/sites/default/files/report/2014_open_data_as_a_tool_to_fight_corruption.pdf.

Gurstein, Michael B. 2011. “Open data: Empowering the empowered or effective data use for everyone?” First Monday 16, no. 2 (January). https://doi.org/10.5210/fm.v16i2.3316.

Hay, Brian. 2019. “Smart Cities of Today and Tomorrow: Better Technology, Infrastructures and Society.” Journal of Tourism Futures 5, no. 3 (October): 303–304. http://dx.doi.org/10.1108/JTF-09-2019-092.

House of Commons. 2014. “Statistics and Open Data: Harvesting unused knowledge, empowering citizens and improving public services.” Tenth Report of Session 2013-14. London: House of Commons. https://publications.parliament.uk/pa/cm201314/cmselect/cmpubadm/564/564.pdf.

Lloyd-Jones, Tony, and Max Lock Centre at the University of Westminster. 2006. “Mind the Gap! Post-disaster reconstruction and the transition from humanitarian relief.” Prevention Web. https://www.preventionweb.net/files/9080_MindtheGapFullreport1.pdf.

Manyika, James, Michael Chui, Diana Farrell, Steve V. Kuiken, Peter Groves, and Elizabeth A. Doshi. 2013. “Open data: Unlocking innovation and performance with liquid information.” October 1, 2013. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information.

Motorevska, Yevheniia. 2025. “Investigation: Uncovering the secret Russian FSB operation to loot Ukraine's museums.” The Kyiv Independent. https://kyivindependent.com/investigation-uncovering-fsbs-secret-operation-to-steal-ukraines-valuable-art/.

Muñoz, Laura A., Manuel P. Rodríguez Bolívar, and Cinthia L. Arellano. 2022. “Factors in the adoption of open government initiatives in Spanish local governments.” Government Information Quarterly 39, no. 4 (July): 11. https://doi.org/10.1016/j.giq.2022.101743.

Neumaier, Sebastian, Jürgen Umbrich, and Axel Polleres. 2016. “Automated Quality Assessment of Metadata across Open Data Portals.” Data and Information Quality 8, no. 1 (November): 1-29. 10.1145/2964909.

Neves, Fátima T., Miguel d. Neto, and Manuela Aparicio. 2020. “The impacts of open data initiatives on smart cities: A framework for evaluation and monitoring.” Cities 106, no. 6 (November): 1-15. https://doi.org/10.1016/j.cities.2020.102860.

Oficialiosios statistikos portalas. 2025. “Nuolatinių gyventojų skaičius apskrityse ir savivaldybėse metų pradžioje.” Oficialiosios statistikos portalas. https://osp.stat.gov.lt/statistiniu-rodikliu-analize?hash=684e50e2-6cf6-426f-8d20-8b3e3856bdd2#/.

Pereira, Gabriela V., Marie A. Macadar, Edimara M. Luciano, and Maurício G. Testa. 2016. “Delivering public value through open government data initiatives in a Smart City context.” Information Systems Frontiers 19 (July): 213–229. 10.1007/s10796-016-9673-7.

Publications Office of the European Union. 2021. “data.europa.eu data quality guidelines.” https://data.europa.eu/doi/10.2830/79367.

Rajamae-Soosaar, Katrin, and Anastasija Nikiforova. 2024. “Exploring Estonia's open government data development as a journey towards excellence: Unveiling the progress of local governments in Open data provision.” Proceedings of the 25th Annual

International Conference on Digital Government Research, (June), 920 - 931. 10.1145/3657054.3657161.

Ruth,, Michael. 2023. “Twin towns and sister cities.” EBSCO. https://www.ebsco.com/research-starters/history/twin-towns-and-sister-cities.

Samokhodskyi, Ihor. 2023. “Can Open Data Form the Basis for a Transparent Recovery Process in Ukraine?” The Royal United Services Institute. https://www.rusi.org/explore-our-research/publications/commentary/can-open-data-form-basis-transparent-recovery-process-ukraine.

Santanu, Roy, Durga K. Sri, A.V.N. Murty, and M. Killedar. 2025. “Applications of big data analytics in urban planning and development: Current trends and future directions.” Journal of Applied Bioanalysis 11, no. 1 (January): 46–54. http://doi.org/10.53555/jab.v11i1.059.

SAS. 2020. “The Town of Cary, NC, teams up with SAS and Microsoft Azure to protect citizens from flooding, safeguard watersheds and support environmentally sound development.” SAS. https://www.sas.com/en_gb/customers/townofcary-flood-prediction.html.

“Spanish Statistical Office - Population.” n.d. INE. Accessed June 1, 2025. https://www.ine.es/en/.

Transparent Cities. 2024. “Transparent Cities: Methodology for Evaluating Open Data in Cities.” Transparent Cities. https://transparentcities.in.ua/en/news/yak-prozori-mista-budut-otsiniuvaty-vidkryti-dani-u-mistakh--prezentatsiia-metodolohii.

Visit Ukraine. 2025. “Air alarm: an app that alerts you to dangers in a specific region on your smartphone.” Visit Ukraine. https://visitukraine.today/blog/143/air-alert-app-that-notifies-about-danger-in-certain-region-in-your-smartphone?srsltid=AfmBOoo4n9v07g69zsGHLx1uJKivhDN9vgjvnTDHzEH9NPbGq9S0aDwO#the-air-alarm-app-what-is-it-and-how-can-you-download-it-to-your-smartphone.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, and Appleton Gabrielle. 2016. “The FAIR Guiding Principles for scientific data management and stewardship.” Sci Data 3 (March). 10.1038/sdata.2016.18.

Wilson, Bev, and Cong Cong. 2020. “Telematics and Informatics.” Beyond the supply side: Use and impact of municipal open data in the U.S 58 (November). https://doi.org/10.1016/j.tele.2020.101526.

Wirtz, Bernd W., Jan C. Weyerer, Marcel Becker, and Wilhelm M. Müller. 2022. “Open government data: A systematic literature review of empirical research.” Electron Markets 32, no. 2381–2404 (September). https://doi.org/10.1007/s12525-022-00582-8.

Witzleb, Normann. 2023. “Responding to Global Trends? Privacy Law Reform in Australia.” In Data Disclosure: Global Developments and Perspectives edited by Moritz Hennemann, Kai von Lewinski, Daniela Wawra and Thomas Widjaja, 147–168. Berlin, Boston: De Gruyter, 2023. https://doi.org/10.1515/9783111010601-009.

Yan, An, and Nicolas Weber. 2018. “Mining Open Government Data Used in Scientific Research.” (February). 10.48550/arXiv.1802.03074.

APPENDIX 1

JavaScript:

```
(async () => {
  //const listUrl = 'https://opendata.vilnius.lt/api/3/action/package_list';
  //const showUrlBase = 'https://opendata.vilnius.lt/api/3/action/package_show?id=';
  const city = "rostok"
  const listUrl = `https://opendata-hro.de/api/3/action/package_list`;
  const showUrlBase = `https://opendata-hro.de/api/3/action/package_show?id=`;

  const datasetDetails = [];

  try {
    const listResponse = await fetch(listUrl);
    const listData = await listResponse.json();

    if (!listData.success) {
      throw new Error('Failed to retrieve dataset list');
    }

    const datasetIds = listData.result;
    console.log(`Found ${datasetIds.length} dataset IDs.`);

    for (const id of datasetIds) {
      const showUrl = `${showUrlBase}${encodeURIComponent(id)}`;

      try {
        const response = await fetch(showUrl);
        const data = await response.json();

        if (data.success) {
          datasetDetails.push(data.result);
          console.log(`Fetched: ${id}`);
        } else {
          console.warn(`Failed to fetch details for dataset: ${id}`);
        }

      } catch (err) {
        console.error(`Error fetching dataset ${id}:`, err);
```

```
            }
        }

        console.log(`Fetched ${datasetDetails.length} dataset details.`);

        // Convert to JSON and create a downloadable file
            const blob = new Blob([JSON.stringify(datasetDetails, null, 2)], { type: 'application/json' });
        const url = URL.createObjectURL(blob);
        const a = document.createElement('a');
        a.href = url;
        a.download = `${city}_datasets.json`;
        document.body.appendChild(a);
        a.click();
        document.body.removeChild(a);
        URL.revokeObjectURL(url);

        //console.log('Download triggered: vilnius_datasets.json');
        console.log(`Download triggered: ${city}_datasets.json`);

      } catch (error) {
        console.error('Error:', error);
      }
    })();
```

Python

```
import json
import csv
import geopandas as gpd
import pandas as pd
import requests
import time
import os

## os.path.getsize("/path/to/file.mp3")

## vincitycouncil
```

```
        fieldnames = ['city', 'dataset_format', 'dataset_id', 'date_created', 'date_modified', 'f_size','r_size','n_fields','n_text_fields', 'n_numeric_fields', 'n_datetime_fields', 'n_boolean_fields','n_geometry_fields', 'tags']
        filename = "rostok_raw_data.csv"
        city = "rostok"

        def find_delimiter(filename):
          sniffer = csv.Sniffer()
          with open(filename) as fp:
             delimiter = sniffer.sniff(fp.read(5000)).delimiter
          return delimiter

        def count_excel_features(excel_path):
          try:
             df = pd.read_json(excel_path)
             dtypes = df.dtypes
          except Exception as e:
             return f"{excel_path} - csv wrong type"
          type_counts = {
             'numeric': 0,
             'text': 0,
             'datetime': 0,
             'boolean': 0,
             'geometry': 0,
             'other': 0
          }
          columns_by_type = {
             'numeric': [],
             'text': [],
             'datetime': [],
             'boolean': [],
             'geometry': [],
             'other': []
          }
          try:
             for col_name, dtype in dtypes.items():
                if pd.api.types.is_numeric_dtype(dtype):
                   type_counts['numeric'] += 1
                   columns_by_type['numeric'].append(col_name)
```

```
            elif pd.api.types.is_string_dtype(dtype) or dtype == 'object':
                type_counts['text'] += 1
                columns_by_type['text'].append(col_name)
            elif pd.api.types.is_datetime64_any_dtype(dtype):
                type_counts['datetime'] += 1
                columns_by_type['datetime'].append(col_name)
            elif pd.api.types.is_bool_dtype(dtype):
                type_counts['boolean'] += 1
                columns_by_type['boolean'].append(col_name)
            else:
                type_counts['other'] += 1
                columns_by_type['other'].append(col_name)
        result = {
            'total_columns': len(dtypes),
            'size': len(df),
            'counts': type_counts,
            'columns_by_type': columns_by_type
        }
    except Exception as e:
        print(e)
    return result

def count_json_features(json_path):
    try:
        df = pd.read_json(json_path)
        dtypes = df.dtypes
    except Exception as e:
        return f"{json_path} - csv wrong type"
    type_counts = {
        'numeric': 0,
        'text': 0,
        'datetime': 0,
        'boolean': 0,
        'geometry': 0,
        'other': 0
    }
    columns_by_type = {
        'numeric': [],
        'text': [],
```

```
        'datetime': [],
        'boolean': [],
        'geometry': [],
        'other': []
    }
    try:
        for col_name, dtype in dtypes.items():
            if pd.api.types.is_numeric_dtype(dtype):
                type_counts['numeric'] += 1
                columns_by_type['numeric'].append(col_name)
            elif pd.api.types.is_string_dtype(dtype) or dtype == 'object':
                type_counts['text'] += 1
                columns_by_type['text'].append(col_name)
            elif pd.api.types.is_datetime64_any_dtype(dtype):
                type_counts['datetime'] += 1
                columns_by_type['datetime'].append(col_name)
            elif pd.api.types.is_bool_dtype(dtype):
                type_counts['boolean'] += 1
                columns_by_type['boolean'].append(col_name)
            else:
                type_counts['other'] += 1
                columns_by_type['other'].append(col_name)
        result = {
            'total_columns': len(dtypes),
            'size': len(df),
            'counts': type_counts,
            'columns_by_type': columns_by_type
        }
    except Exception as e:
        print(e)
    return result

def count_csv_features(csv_path):
    try:

        # df = pd.read_csv(csv_path, encoding='utf-8-sig')
        # dtypes = df.dtypes
        # print(dtypes)
        delimeter = find_delimiter(csv_path)
```

```
            try:
                df = pd.read_csv(csv_path, encoding='utf-8-sig', sep=f'{delimeter}')  # Note the
separator ';'
            except UnicodeDecodeError:
                df = pd.read_csv(csv_path, encoding='latin1', sep=f'{delimeter}')
            dtypes = df.dtypes
        except Exception as e:
            return f"{csv_path} - csv wrong type"
        type_counts = {
            'numeric': 0,
            'text': 0,
            'datetime': 0,
            'boolean': 0,
            'geometry': 0,
            'other': 0
        }
        columns_by_type = {
            'numeric': [],
            'text': [],
            'datetime': [],
            'boolean': [],
            'geometry': [],
            'other': []
        }
        try:
            for col_name, dtype in dtypes.items():
                if pd.api.types.is_numeric_dtype(dtype):
                    type_counts['numeric'] += 1
                    columns_by_type['numeric'].append(col_name)
                elif pd.api.types.is_string_dtype(dtype) or dtype == 'object':
                    type_counts['text'] += 1
                    columns_by_type['text'].append(col_name)
                elif pd.api.types.is_datetime64_any_dtype(dtype):
                    type_counts['datetime'] += 1
                    columns_by_type['datetime'].append(col_name)
                elif pd.api.types.is_bool_dtype(dtype):
                    type_counts['boolean'] += 1
                    columns_by_type['boolean'].append(col_name)
                else:
```

```
            type_counts['other'] += 1
            columns_by_type['other'].append(col_name)
        result = {
            'total_columns': len(dtypes),
            'size': len(df),
            'counts': type_counts,
            'columns_by_type': columns_by_type
        }
    except Exception as e:
        print(e)

    return result

def count_geo_features(geojson_path):
    try:
        gdf = gpd.read_file(geojson_path)
        dtypes = gdf.dtypes
    except Exception as e:
        pass
        return f"{geojson_path} - wrong type"
    type_counts = {
        'numeric': 0,
        'text': 0,
        'datetime': 0,
        'boolean': 0,
        'geometry': 0,
        'other': 0
    }
    columns_by_type = {
        'numeric': [],
        'text': [],
        'datetime': [],
        'boolean': [],
        'geometry': [],
        'other': []
    }
    try:
        for col_name, dtype in dtypes.items():
            if pd.api.types.is_numeric_dtype(dtype):
```

```
            type_counts['numeric'] += 1
            columns_by_type['numeric'].append(col_name)
        elif pd.api.types.is_string_dtype(dtype) or dtype == 'object':
            type_counts['text'] += 1
            columns_by_type['text'].append(col_name)
        elif pd.api.types.is_datetime64_any_dtype(dtype):
            type_counts['datetime'] += 1
            columns_by_type['datetime'].append(col_name)
        elif pd.api.types.is_bool_dtype(dtype):
            type_counts['boolean'] += 1
            columns_by_type['boolean'].append(col_name)
        elif dtype.name == 'geometry':
            type_counts['geometry'] += 1
            columns_by_type['geometry'].append(col_name)
        else:
            type_counts['other'] += 1
            columns_by_type['other'].append(col_name)
    result = {
        'total_columns': len(dtypes),
        'size': len(gdf),
        'counts': type_counts,
        'columns_by_type': columns_by_type
    }

  except Exception as e:
    print(e)
  print(f'Result of analisys {result}')
  return result

def download_datasets_files(url,file,format):
  try:
    #time.sleep(1)
    response = requests.get(url)
    response.raise_for_status()  # Raise exception for HTTP errors
    with open(f"data/{file}.{format}", 'wb') as f:
        f.write(response.content)
        print(f"File saved as {file}")
  except requests.exceptions.RequestException as e:
    print(f"Error downloading the file: {e}")
```

```
        with open('data/failed', 'a') as file:
            file.write(f"{url}-{format}\n")

def analise_datasets(file,format):
    match format:
        case "json":
            print("Disable json")
            #return(count_json_features(file))
        case "csv":
            return(count_csv_features(file))
        case "geojson":
            print("Disable geojson")
            #return(count_geo_features(file))
        case "xlsx":
            print("Disable xlsx")
            #return(count_excel_features(file))
        case "txt":
            print("TXT")
        case "api":
            print("API")
        case _:
            print("Unknow format")


with open('rostok_datasets.json') as f:
    ckan_metadata = json.load(f)

# Prepare a list to store all data rows
data = []

# Assuming ckan_metadata is a list of datasets or a single dataset object
if isinstance(ckan_metadata, dict):
    # If it's a single dataset
    dataset_list = [ckan_metadata]
else:
    # If it's a list of datasets
    dataset_list = ckan_metadata

for dataset_metadata in dataset_list:
```

```
            # Extract tags from dataset level
            tags_array = [tag['name'] for tag in dataset_metadata.get('tags', [])]
            # Now properly access the resources list
            for resource in dataset_metadata.get('resources', []):

#download_datasets_files(resource.get('url'),resource.get('id'),resource.get('format').lower())
                                                                    data_counter =
analise_datasets(f"data/{resource.get('id')}.{resource.get('format').lower()}",resource.get('form
at').lower())
                print(resource['format'].lower(),resource['id'])
                print(data_counter)
                try:
                    temp = {
                        'city': city,
                        'dataset_format': resource['format'].lower(),
                        'dataset_id': resource['id'],
                        'date_created': resource['created'],
                        'date_modified': resource['last_modified'],
                                                                    'f_size':
os.path.getsize(f"data/{resource.get('id')}.{resource.get('format').lower()}"),
                        'r_size': data_counter['size'],
                        'n_fields': data_counter['total_columns'],
                        'n_text_fields': data_counter['counts']['text'],
                        'n_numeric_fields': data_counter['counts']['numeric'],
                        'n_datetime_fields': data_counter['counts']['datetime'],
                        'n_boolean_fields': data_counter['counts']['boolean'],
                        'n_geometry_fields': data_counter['counts']['geometry'],
                        'tags': tags_array
                    }
                # Add this row to our data list
                    data.append(temp)
                except Exception as e:
                    pass
                    print(e)

        # Write all rows to CSV
        with open(filename, 'w', newline='') as csvfile:
          writer = csv.DictWriter(
            csvfile,
```

```
            fieldnames=fieldnames,
            restval='Missing',  # Use 'N/A' for missing values
            extrasaction='ignore'  # Ignore fields not in fieldnames
        )
        writer.writeheader()
        writer.writerows(data)

    print(f"CSV with data written to {filename}")
```

APPENDIX 2

R

```
library(tidyverse)
library(factoextra)
library(ggplot2)
library(patchwork)


setwd("~/Current/diplom-research/r-analisys")



vilnus <- read_csv("./data/vilnus_raw_data.csv")
leipzig <- read_csv("./data/leipzig_raw_data.csv")
stuttgart <- read_csv("./data/stuttgart_raw_data.csv")
vinnica <- read_csv("./data/vinnica_raw_data.csv")
malaga <- read_csv("./data/malaga_raw_data.csv")
rostok <- read_csv("./data/rostok_raw_data.csv")
vigo <- read_csv("./data/vigo_raw_data.csv")

comb_raw_dataset <- rbind(vinnica,vilnus,leipzig,stuttgart,vigo,rostok,malaga)


comb_raw_dataset_mod <- comb_raw_dataset %>%
  select(city,dataset_id,r_size,n_fields)

Q_r_size <- quantile(comb_raw_dataset_mod$r_size, probs=c(.25, .75), na.rm = FALSE)
Q_f_n <- quantile(comb_raw_dataset_mod$n_fields, probs=c(.25, .75), na.rm = FALSE)

iqr_r_size <- IQR(comb_raw_dataset_mod$r_size)
iqr_f_n <- IQR(comb_raw_dataset_mod$n_fields)

up_r_size <-  Q_r_size[2]+1.5*iqr_r_size # Upper Range
```

```
low_r_size<- Q_r_size[1]-1.5*iqr_r_size # Lower Range

up_f_n <-  Q_f_n[2]+1.5*iqr_f_n # Upper Range
low_f_n<- Q_f_n[1]-1.5*iqr_f_n # Lower Range
#comb_means = select(comb_raw_dataset_quality,c(1,2,3,6,7,8,9,10,11,13,15)) ### without boolean


comb_raw_dataset_mod_eliminated <- comb_raw_dataset_mod %>%
  subset(r_size > low_r_size & r_size < up_r_size) %>%
  subset(n_fields > low_f_n & n_fields < up_f_n)
## Scale

comb_means_scale <- scale(select(comb_raw_dataset_mod_eliminated,c(r_size,n_fields)))


## Distance

comb_means_dist <- dist(comb_means_scale)

## WSS plot

fviz_nbclust(comb_means_scale, kmeans, method = "wss")+  # 3
  labs(subtitle="Elbow Method")

## k-mean

km.out <- kmeans(comb_means_scale,centers=3,nstart=100) #73.5
print(km.out)

## viz


#comb_means_scale_city %>%
#  ggplot(aes(x=r_size...1, y=f_n...2,colour = city)) +
#  geom_point()

#comb_means_scale <- comb_means_scale %>%
#  tibble::rownames_to_column("city")

#comb_means_scale %>%
#  ggplot(aes(x=r_size, y=f_n, color=variable_name)) +
#  geom_point()
  #scale_x_discrete(labels=c("0-10", "10-20", "20-30", "30-40"))

km.clusters<-km.out$cluster
```

```
comb_raw_dataset_mod_eliminated_analisys <- comb_raw_dataset_mod_eliminated |>
mutate(cluster = km.out$cluster)
#comb_raw_dataset_mod_eliminated_analisys %>%
#  ggplot(aes(x=r_size, y=n_fields, colour =as.factor(city) ,shape=as.factor(cluster))) +
#  geom_point()


#rownames(comb_means_scale) <- paste(comb_raw_dataset_mod_eliminated$city)
#rownames(comb_means_scale) <- paste(,1:dim(comb_raw_dataset_mod_eliminated)[1], sep = "_")

cluster_analiz <- fviz_cluster(ellipse = TRUE,geom="point",
  list(data=comb_means_scale, cluster = km.clusters,main = "Cluster plot field n and record n")
  )+
  scale_colour_manual(values = c("#F8766D", "orange","#53B400")) +
  scale_fill_manual(values = c( "#F8766D", "orange","#53B400"))

clster_analiz_city <- comb_raw_dataset_mod_eliminated_analisys %>%
  ggplot(aes(x=r_size, y=n_fields, colour =as.factor(city) ,shape=as.factor(cluster))) +
  labs(title = "Plot field n and record n",
       x = "Number of records",
       y = "Number of fields",
       color = "City",
       shape = "Cluster") +
  geom_point()

combined_plot <- cluster_analiz + clster_analiz_city

### Claculate datasets in each cluster
cluster_counts <- comb_raw_dataset_mod_eliminated_analisys %>%
  group_by(city, cluster) %>%
  summarize(
    count = n(),                    # Number of observations
    .groups = 'drop'
  )

cluster_counts_percent <- cluster_counts %>%
  group_by(city) %>%
  mutate(city_total = sum(count)) %>%
  # Step 2: Calculate the percentage for each cluster within the city
  mutate(percent = round(count / city_total * 100, 1)) %>%
  # Step 3: Optional - remove the helper column if not needed
  select(-city_total)

# View the result
print(n=100,cluster_counts_percent)
```

```
cluster_analiz_calc <- ggplot(cluster_counts_percent, aes(x = city, y = percent, fill = factor(cluster))) +
  geom_bar(stat = "identity", position = "stack") +
  labs(
    title = "Percentage Distribution by City and Cluster",
    x = "City",
    y = "Percent (%)",
    fill = "Cluster"
  ) +
  theme_minimal() +
  scale_fill_manual(values = c("1" = "#F8766D", "2" = "orange", "3" = "#53B400"))

combined_plot <- cluster_analiz + cluster_analiz_calc
```