

Data Science Project

An overview of rental prices in Switzerland

Conceptual Design Report

02 october 2024

Abstract

This project focuses on analyzing the rental market in Switzerland by scraping data from homegate.ch, a popular real estate platform. The primary objective was to gather and process data on rental apartments across Swiss municipalities to derive insights into rental price variations and to develop a rent simulator.

Table of Contents

Abstract	0
Table of Contents	1
1 Project Objectives	2
2 Methods	2
3 Data	3
4 Metadata	4
5 Data Quality	4
6 Data Flow	5
7 Data Model	6
8 Documentation	7
9 Risks	8
Technical Risks	8
Analytical Risks	9
Legal and Ethical Risks	9
Quality Risks	10
10 Preliminary Studies	11
11 Conclusions	17
Statement	18

1 Project Objectives

The primary goal of this project is to scrape the website **homegate.ch** to collect data about rental apartment listings across Switzerland. This data is then analyzed to generate statistics that provide insights into the rental market. The project focuses specifically on rental apartments, excluding other categories such as parking spaces, houses, and commercial properties. The goal is to offer a snapshot of the rental market by using the data collected from the website during a single scraping session.

The specific objectives are:

- **Data Collection:** Scrape and collect data on rental apartments from the website homegate.ch.
- **Data Analysis:** Analyze the scraped data to extract insights such as average rental prices by municipality and canton.
- **Visualization:** Visualize the data through maps and charts to make the findings more accessible and comprehensible.
- **Prediction:** Build a rent simulator based on the average rent per square meter in different municipalities, allowing users to estimate rents for specific locations and property sizes.

Given the constraints of scraping the website in one go, this project aims to deliver an accurate but momentary picture of the rental market, recognizing the limitations of capturing dynamic market data over time.

2 Methods

The project utilizes a variety of tools and libraries, as well as statistical methods, to achieve its objectives. The primary infrastructure used includes:

- **Local Python Environment:** Due to restrictions on Google Colab (e.g., blocked requests by homegate.ch), the project was executed locally using a Jupyter notebook.
- **Libraries:**
 - **fake-useragent:** This was used to generate fake user agents, making it harder for homegate.ch to block scraping requests.
 - **geopandas:** Used for geospatial analysis and mapping Swiss municipalities and cantons.
 - **matplotlib:** Employed for generating visualizations, including heatmaps of rent prices.

- **pandas**: Utilized for data manipulation, cleaning, and analysis.

Data collection involves sending requests to the homegate.ch website using the aforementioned libraries. Data is then cleaned, processed, and visualized using Python's data science libraries. The project is designed to be run locally due to limitations of cloud platforms in handling large-scale scraping and the need for custom rate limiting to avoid being blocked by Cloudflare's anti-bot measures.

3 Data

The data for this project is sourced from **homegate.ch**, a Swiss real estate platform that provides listings for rental apartments. Due to the aggressive rate-limiting imposed by Cloudflare on homegate.ch, scraping was performed in a controlled manner with a delay of 10 seconds between requests. Only apartment listings were considered, while other categories (e.g., houses, parking spaces) were excluded.

The data contains fields such as:

- **Location**: Municipality, canton.
- **Price**: Rental price.
- **Size**: Surface area in square meters.
- **Number of rooms**: To differentiate between apartment sizes.

A couple of example rows might include information like:

Municipality	Canton	Rent Price (CHF)	Surface (sqm)	Number of Rooms
Zurich	ZH	2500	80	3.5
Geneva	GE	3200	70	4.0

Security considerations include ensuring compliance with homegate.ch's terms of use and the data collection being limited to publicly available information.

4 Metadata

To reproduce the analysis, several metadata elements are required:

- **Scraping details:** This includes the list of URLs accessed, the timing of the scraping process, and the fake user agents used.
- **Data cleaning processes:** Documenting the steps taken to clean and process the data, such as handling missing values or filtering non-apartment listings.
- **Codebase:** The Python code used for scraping, analysis, and visualization.

This metadata is stored alongside the project files and can be accessed via the local environment.

5 Data Quality

The data quality requirements are crucial to ensure the accuracy of the rental market analysis:

- **Completeness:** The dataset must contain information from all relevant municipalities; however, it is noted that some municipalities may lack data due to scraping limitations.
- **Consistency:** Data cleaning ensures consistency across the dataset, handling missing values and erroneous entries.
- **Timeliness:** The data represents a snapshot of the market at a single point in time, which could impact the accuracy of any long-term predictions.

Measures to improve data quality include regular checks during the scraping process and ensuring a consistent methodology is followed throughout the data collection and cleaning phases.

6 Data Flow

The data flow of the project follows a well-structured process from data collection to analysis and visualization. Here is a simplified linear version of the data flow:

Website Scraping --> Data Cleaning --> Data cross-reference --> Data Analysis --> Data Visualization --> Rent Simulator

Here's an explanation of each step in the data flow process using bullet points:

- **Website Scraping:**
 - Scraping data from homegate.ch, focusing specifically on rental apartment listings.
 - Utilizing a custom Python script that sends HTTP requests to the website while handling rate limiting and bypassing bot protection measures.
 - Collecting information such as rental prices, apartment sizes, number of rooms, and location information (municipality and sometimes canton).
 - Storing the scraped data in a structured format such as a CSV or a database for further processing.
- **Data Cleaning:**
 - Removing duplicates and irrelevant listings (e.g., commercial properties or non-apartment rentals).
 - Addressing missing values, especially for critical fields like rental prices and surface areas.
 - Standardizing formats for various fields (e.g., converting currency and surface units to a consistent format).
 - Identifying and handling outliers, such as exceptionally high or low rent values that may be erroneous.
- **Data Cross-reference:**
 - After cleaning the data, cross-referencing the listings to add missing canton information.
 - This involves mapping each listing to the correct canton based on the municipality name, leveraging a reference table that associates municipalities with cantons.
 - Ensuring that each listing has a canton assigned for accurate analysis and visualization, as this information may not always be present in the raw scraped data.
- **Data Analysis:**
 - Conducting statistical analysis to extract meaningful insights, such as calculating average rent prices per square meter by municipality and canton.

- Segmenting the data by various categories (e.g., apartment size, number of rooms) to explore trends in rental pricing across different types of properties.
- Identifying geographic patterns, such as variations in rent prices between urban and rural areas.
- **Data Visualization:**
 - Creating visual representations of the data, including maps and charts.
 - Heatmaps of Switzerland are generated to illustrate rental price distributions across different municipalities and cantons.
 - Additional visualizations like bar charts or line plots may be used to show trends over time or across different apartment characteristics (e.g., size, number of rooms).
- **Rent Simulator:**
 - Building a tool that uses the analyzed data to predict rental costs based on user input (e.g., desired location, apartment size).
 - The simulator relies on the average rental price per square meter calculated during the analysis phase.
 - Users can input the surface area and municipality they are interested in, and the simulator provides an estimated rent based on the available data.

7 Data Model

At the **conceptual level**, the data model is based on rental listings that contain attributes such as price, size, location, and number of rooms.

At the **logical level**, the columns include:

- **Municipality**
- **Canton**
- **Price per sqm**
- **Total Price**
- **Size in sqm**
- **Number of rooms**

At the **physical level**, the infrastructure is primarily a local Python environment, leveraging Jupyter notebooks for execution. The project requires sufficient memory to handle large datasets and perform geospatial analysis.

The use of a relational database to store results was considered but not implemented, due to the time it would have taken.

8 Documentation

Proper documentation is a crucial aspect of any data science project, ensuring that the work is reproducible, understandable, and maintainable by others, including future versions of the team or external collaborators. The documentation for this project will be divided into several key components, ensuring that each step of the workflow is clearly explained and that the project can be easily understood, executed, and extended by others.

Here are all the types of documentation used in this project:

- **In-line Code Comments:** Throughout the Jupyter notebook and Python scripts, detailed comments are provided to explain the logic behind each block of code. This includes descriptions of function purposes, explanations of complex algorithms, and rationale for design decisions.
- **Function Docstrings:** All custom functions and classes are documented using Python docstrings, which explain the function's purpose, inputs, outputs, and exceptions.
- **Notebook Structure:** The Jupyter notebook is structured with clear section titles using markdown cells. Each section, such as Data Cleaning, Data Analysis, and Visualization, includes a brief description of what the upcoming code will do. This ensures that a reader can follow along easily without needing to understand the code immediately.
- **Installation Instructions:** Detailed steps for setting up the local environment, including instructions for installing necessary dependencies.
- **Usage Instructions:** Guidance on how to execute the Jupyter notebook or Python scripts.
- **Metadata Documentation:** Information regarding how the data was collected (e.g., scraping methodology), what transformations were applied (e.g., cleaning procedures), and any assumptions made during the analysis will be documented thoroughly.
- **Scraping Methodology:** A clear description of the web scraping process, including which endpoints were accessed on homegate.ch, the structure of the HTML elements that were scraped, and how rate limiting and Cloudflare protections were handled.
- **Data Cleaning Methodology:** An overview of the cleaning steps applied to the raw data, such as handling missing values, outliers, and any assumptions or transformations applied to the data.
- **Cross-Referencing Process:** The methodology used to cross-reference municipalities with their respective cantons to ensure accurate geographical analysis.

9 Risks

Risk management is a crucial aspect of any data science project, ensuring that potential issues are identified early and mitigated to avoid significant disruptions to the project's objectives, timeline, and quality of outcomes. Below, the risks for this project are divided into several categories, with corresponding mitigation strategies and potential impacts.

Technical Risks

- **Website Structure Changes:**
 - **Risk:** The website structure of homegate.ch may change, breaking the scraping script. Websites often update their HTML structure or use dynamic loading techniques that can interfere with scraping.
 - **Impact:** If the structure changes, the scraping process may fail, resulting in incomplete or no data being collected.
 - **Mitigation:** Regularly monitor the website for changes and maintain flexibility in the scraping code to adapt to potential changes. Consider implementing an automated test that checks if the critical HTML elements for scraping still exist before running the full script. Additionally, versioning the scraping script can help ensure that previous versions of the scraper are available as a backup if needed.
- **Cloudflare Rate Limiting and Anti-Bot Measures:**
 - **Risk:** Cloudflare's aggressive rate-limiting and anti-bot protection could block further scraping attempts if too many requests are made in a short time.
 - **Impact:** The scraper could be blocked for several hours, delaying data collection and possibly leading to incomplete datasets.
 - **Mitigation:** Use custom rate limiting with extended delays between requests to avoid triggering Cloudflare's protective measures. Introducing randomized user-agent strings and rotating proxies may help prevent blocks. In the event of an IP block, consider having backup IPs or the ability to switch networks to continue the process.
- **Inconsistent Data Availability:**
 - **Risk:** Incomplete data could be collected due to network interruptions, website downtimes, or rate limiting, which could lead to some municipalities or cantons missing from the dataset.
 - **Impact:** The missing data might affect the accuracy of analysis and visualization, leading to biased results that do not fully represent the Swiss rental market.

- **Mitigation:** Implement error handling and logging during the scraping process to catch and retry failed requests. Additionally, cross-reference other public data sources to fill gaps where feasible. Communicate any data gaps clearly in the final report.

Analytical Risks

- **Cross-Referencing Errors:**
 - **Risk:** Errors may occur during the cross-referencing step, where municipalities are mapped to cantons, resulting in incorrect geographical data.
 - **Impact:** Misclassification of municipalities could affect the accuracy of rent analysis by canton, skewing the geographic distribution of rental prices.
 - **Mitigation:** Use reliable external reference tables or APIs that map Swiss municipalities to their correct cantons. Double-check mappings manually, especially for municipalities that may straddle canton borders or have ambiguous names.
- **Data Bias:**
 - **Risk:** The data collected is based on a single snapshot in time and may not fully capture market dynamics, leading to biased results that reflect short-term trends or anomalies (e.g., temporary price spikes due to high demand).
 - **Impact:** Predictions and visualizations may not accurately represent long-term market conditions, which could mislead users or stakeholders.
 - **Mitigation:** Acknowledge the limitations of the data in the analysis and final report. Where possible, compare the findings with historical data or other sources to validate the conclusions. For future work, consider implementing automated data collection over longer periods to account for fluctuations.

Legal and Ethical Risks

- **Legal Issues with Web Scraping:**
 - **Risk:** Scraping websites without proper consent could violate the website's terms of service or lead to legal action, especially if automated scraping causes server strain or collects sensitive data.
 - **Impact:** The project could be subject to legal challenges, requiring cessation of data collection or deletion of previously collected data.
 - **Mitigation:** Ensure that the scraping process complies with homegate.ch's terms of use and relevant data protection laws (e.g., GDPR). Limit scraping to publicly available data and take care not to overload the server. If necessary, seek explicit permission from the website owner before proceeding with large-scale scraping.

- **Data Privacy Concerns:**

- **Risk:** Although rental listings are public information, privacy concerns may arise if personal details (e.g., contact information) are inadvertently scraped.
- **Impact:** Privacy violations could damage the project's reputation or lead to legal consequences.
- **Mitigation:** Ensure that no personal information is scraped or retained during data collection. Focus only on relevant listing attributes such as location, price, and apartment features.

Quality Risks

- **Accuracy of Predictions:**

- **Risk:** The rent simulator's predictions may not always align with real-world rental prices, especially if the dataset is incomplete or biased.
- **Impact:** Users relying on the simulator may receive inaccurate rent estimates, potentially leading to frustration or misuse of the tool.
- **Mitigation:** Clearly communicate the limitations of the simulator in the documentation and provide confidence intervals or disclaimers on the predictions. Regularly update the model with new data to improve its accuracy over time.

10 Preliminary Studies

Preliminary visualizations have been created, including heatmaps of average rents by canton and municipality, as well as descriptive statistics on the collected data. These visualizations provide insights into rental price variations across Switzerland.

The first chart we'll display is a heatmap of average rental prices by municipality. You will notice that some municipalities are missing from the map. This is because there are listings for only 929 municipalities in total. Municipalities without any listings are not displayed. A future improvement could be to display the missing municipalities and assign them the average rent of their canton.

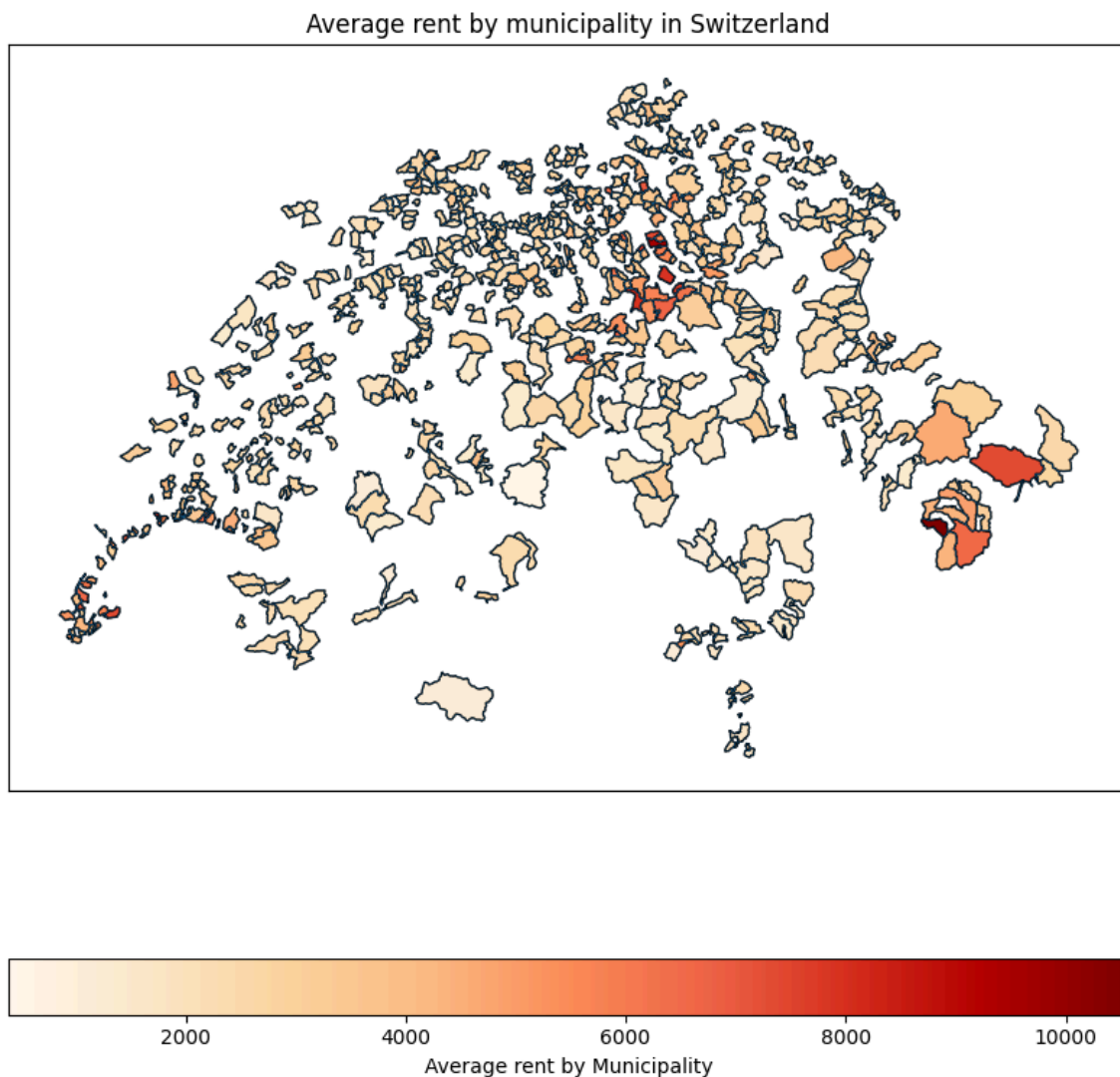


Figure 1: Municipalities heatmap

We will now display a heatmap showing the average rent for each canton:

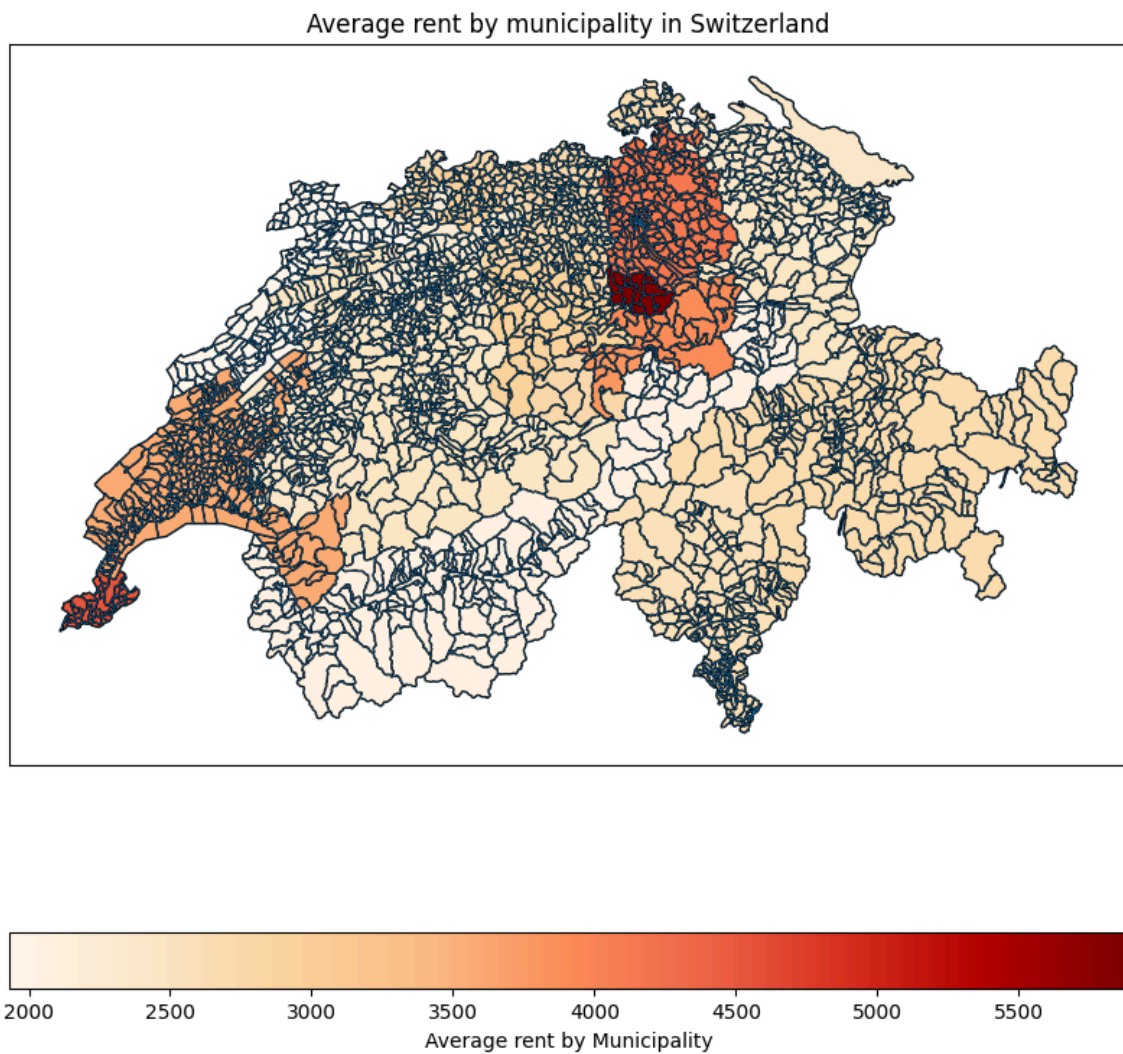


Figure 2: Cantons heatmap

We're now displaying the average rent per canton:

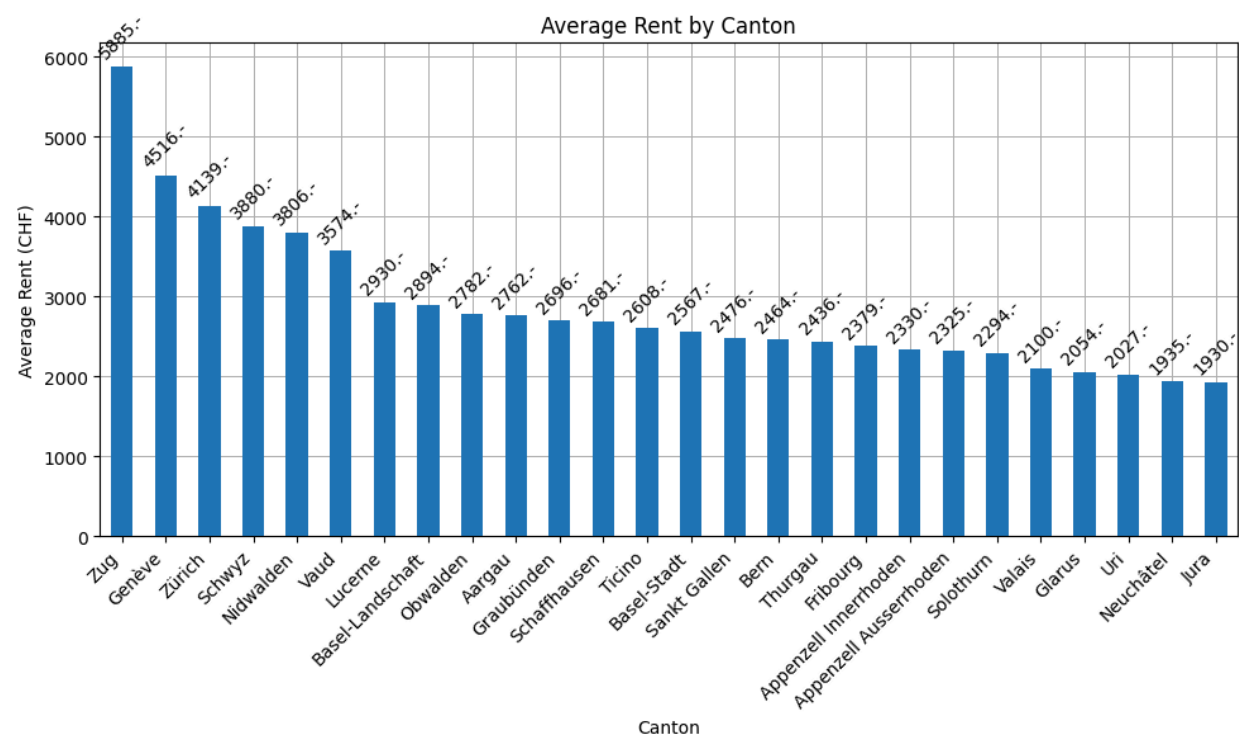


Figure 3: Average rent per Canton

We’re now displaying the 30 most expensive cities to live in switzerland:

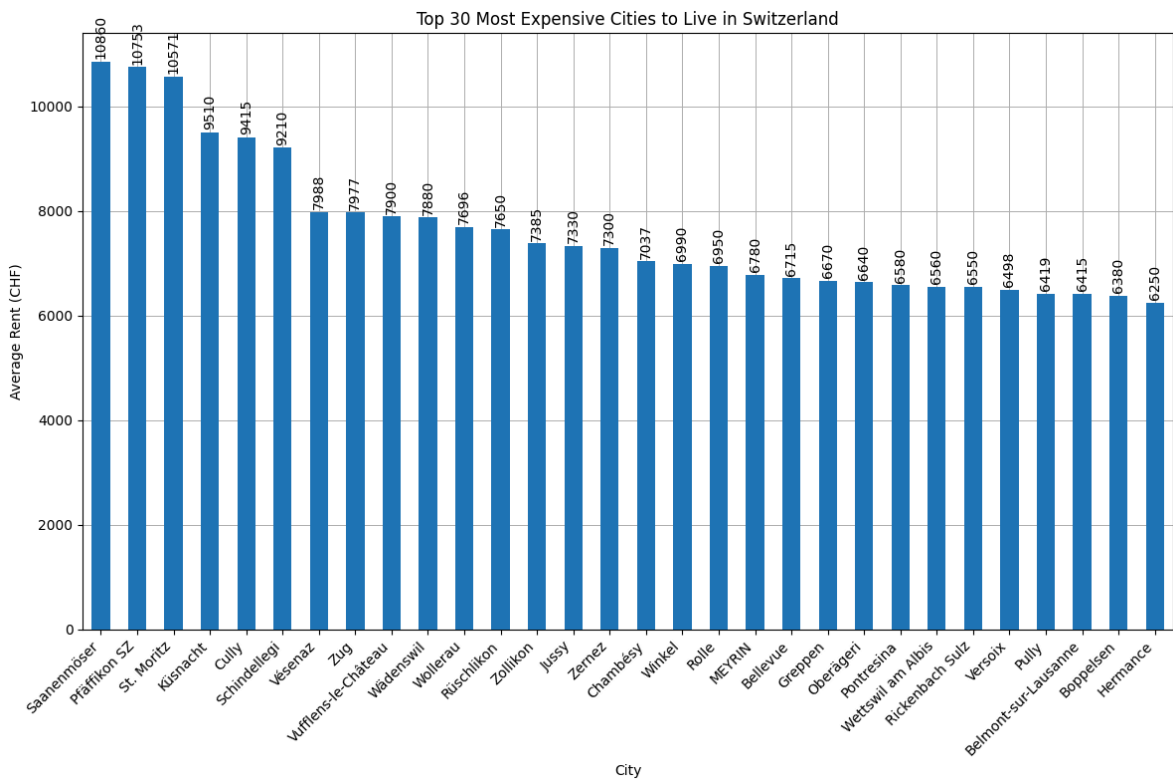


Figure 4: 30 most expensive Swiss cities

We're now displaying the 5 most expensive cities to live in per Canton. As the chart is way too big to be displayed here, we'll show the results for only 10 cantons:

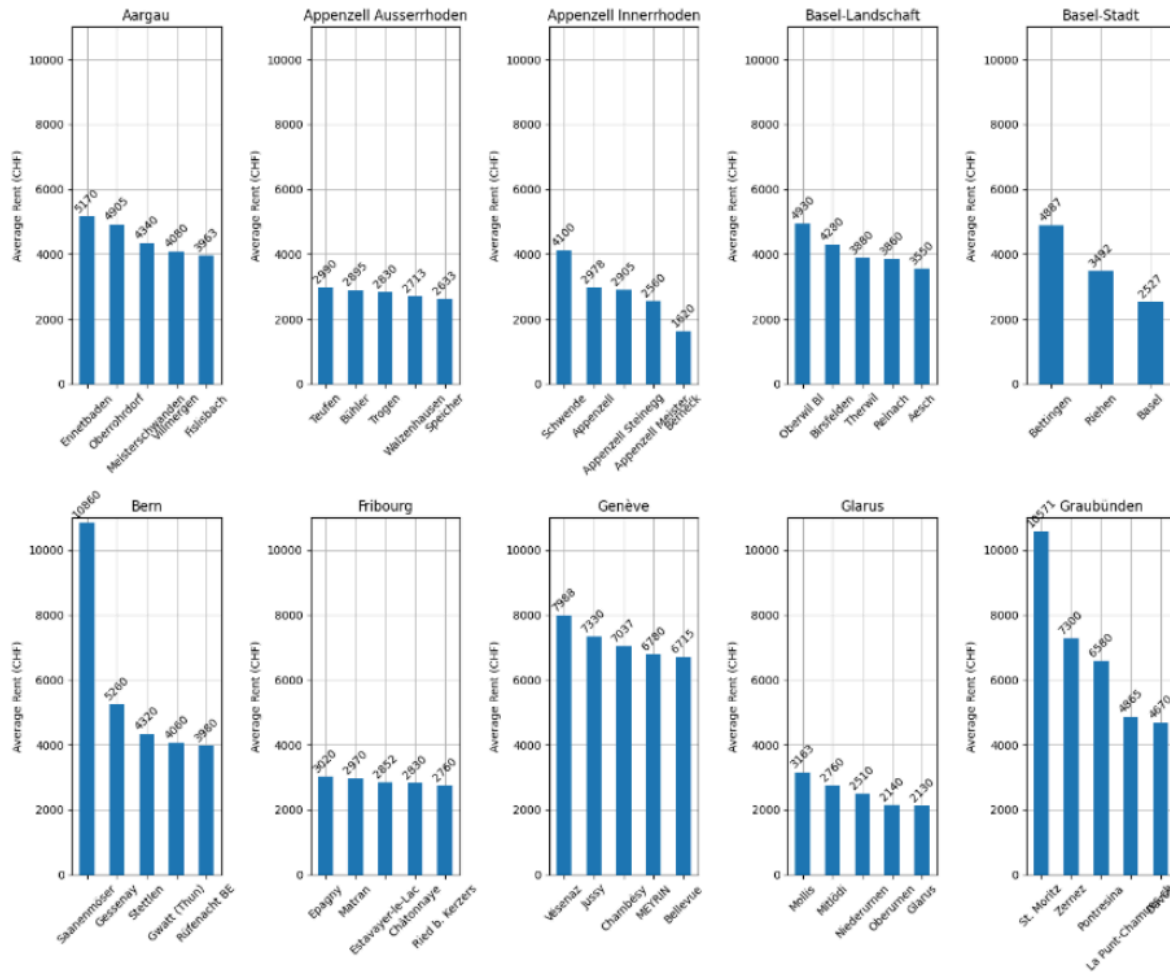


Figure 5: Top 5 most expensive cities per Canton

We'll then display the number of listing per Canton:

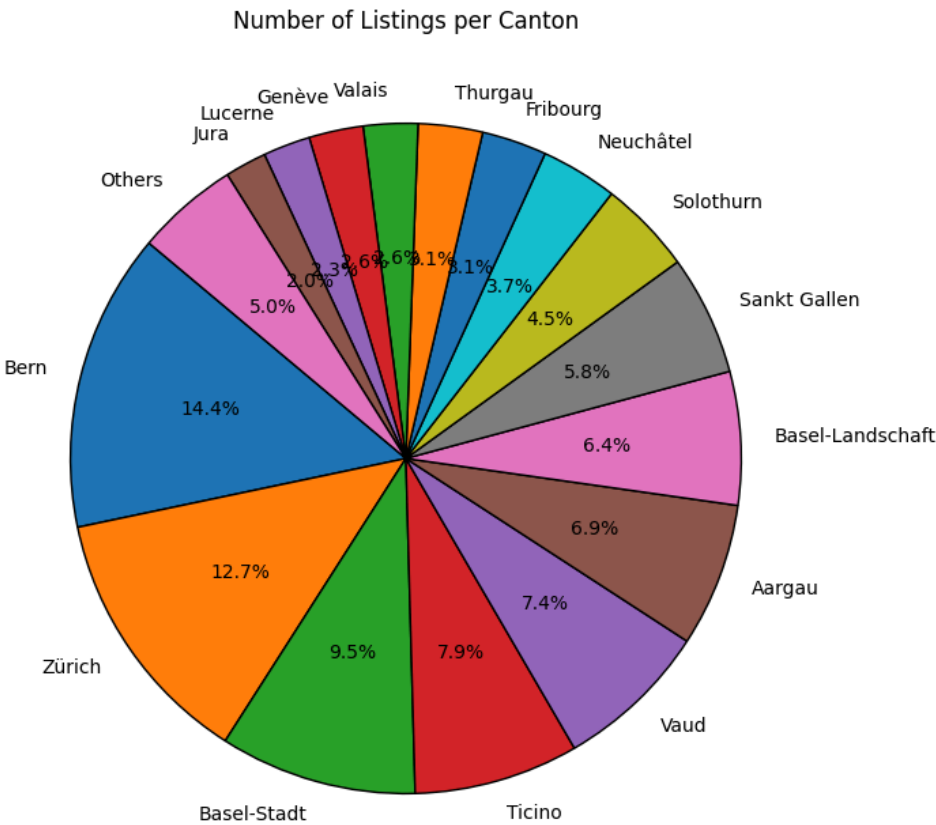


Figure 6: Number of listings per Canton

We have now finished displaying all the charts, and we'll move on to presenting a small rent simulator.

We'll start by calculating the average rent per square meter for each municipality. This will allow us to predict rent based on the municipality and the desired area.

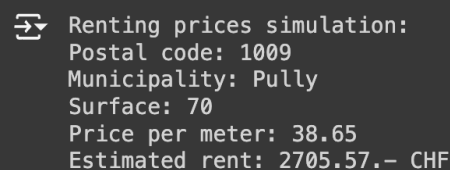
For example, if you want to live in Pully and are looking for a 70-square-meter apartment, the simulator will give you an estimate of how much you would pay.

Please note, the simulator does not work for all municipalities as there are no listings for some of them. Additionally, the estimated price may not reflect the current market conditions since we only gathered data at one point in time. More frequent data collection over a longer period would be needed for greater accuracy.

```
[ ] # The postal code of the municipality where you want to live
    postal_code = 1009
    # The size of your apartment you wishes
    square_meter = 70

    try:
        details = prices_per_minicipality[postal_code]

        print(f"Renting prices simulation: ")
        print(f"Postal code: {postal_code}")
        print(f"Municipality: {details['municipality']}")
        print(f"Surface: {square_meter}")
        print(f"Price per meter: {details['price_per_square_meter']:.2f}")
        print(f"Estimated rent: {details['price_per_square_meter'] * square_meter:.2f}.- CHF")
    except Exception:
        print(f"We couldn't find details for this municipality. Please try wth another one.")
```



```
➦ Renting prices simulation:
Postal code: 1009
Municipality: Pully
Surface: 70
Price per meter: 38.65
Estimated rent: 2705.57.- CHF
```

Figure 7: Renting simulator

11 Conclusions

The project successfully achieved its primary goal of providing a detailed snapshot of the Swiss rental market by collecting and analyzing rental apartment data from homegate.ch. Through a combination of web scraping, data cleaning, geospatial analysis, and statistical modeling, the project delivered valuable insights into rental price distributions across Switzerland. The development of the rent simulator tool also allows users to make rental cost estimates based on location and apartment size, contributing to practical applications for individuals and businesses looking to navigate the rental market.

The project uncovered significant variations in rental prices across Switzerland, with urban areas like Zurich and Geneva showing higher rates compared to rural regions. A trend of decreasing rent prices per square meter with increasing apartment size was also observed. By cross-referencing municipalities with cantons, the project improved data accuracy despite inconsistencies during scraping.

Key achievements include successfully scraping and cleaning rental data from homegate.ch, creating insightful visualizations like heatmaps, and developing a rent simulator tool that estimates costs based on location and apartment size. However, the data represents only a single snapshot in time, limiting the ability to capture trends, and some municipalities lacked data, which may have introduced bias. The rent simulator is also simplified, focusing only on average prices per square meter without considering additional factors like amenities.

Future work could involve continuous data collection to capture trends over time, enhancing the rent simulator with more features, expanding the scope to include other property types, and developing a user-friendly interface. This would improve the accuracy and practical utility of the analysis and predictions.

Statement

The following part is mandatory and must be signed by the author or authors.

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Date: 02.10.2024

Signature(s):

A handwritten signature in black ink, consisting of a stylized 'M' followed by a large, sweeping flourish that ends in a small loop.