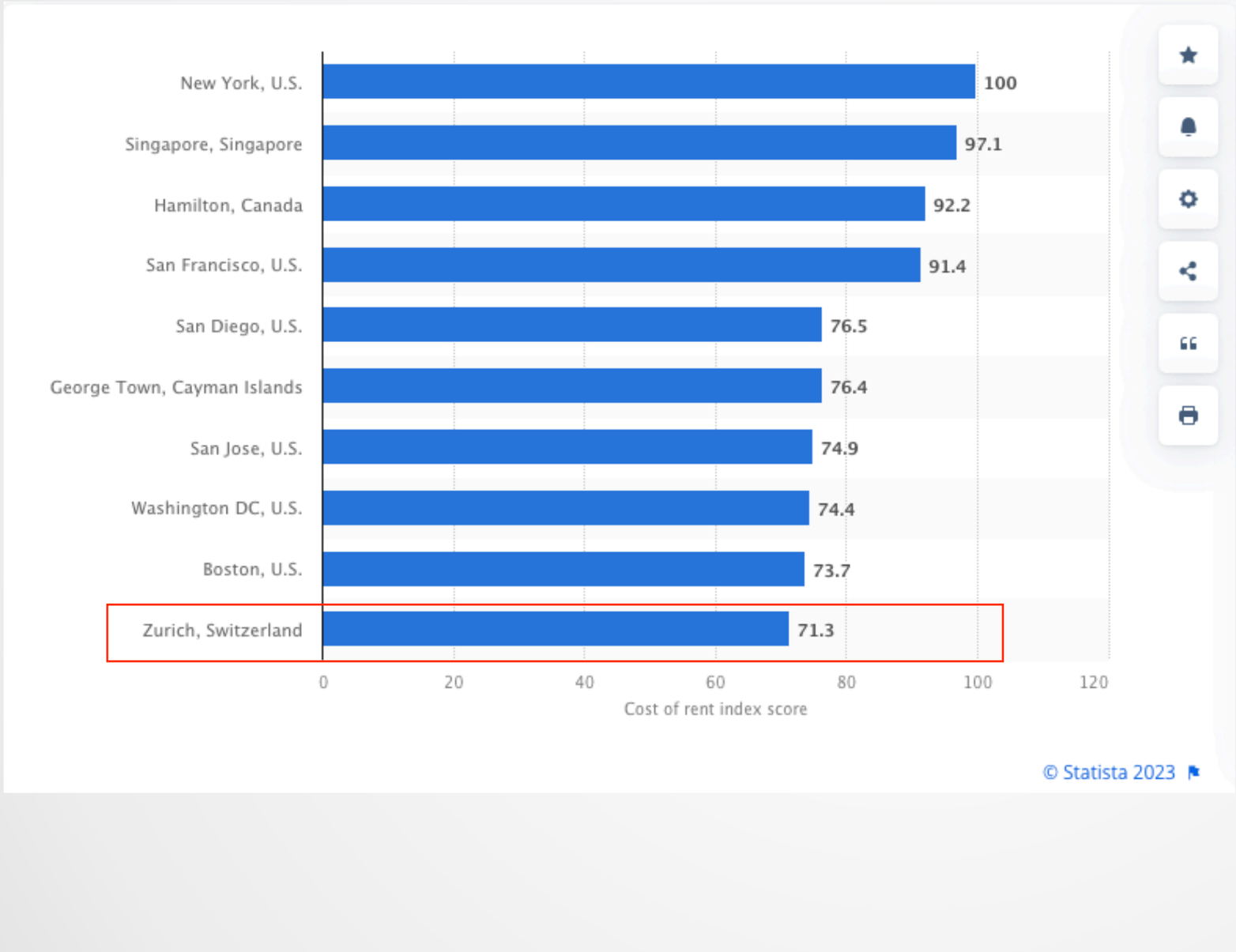
Predictive Analytics for Rental Prices: Zurich Case Study

Jidapa Lengthaisong
Quentin Leoni

Smart Data Analytics Project
Autumn 2023
University of St. Gallen

Institute Digital Assets

Zurich is one of the world's most expensive city to live in



Making the right decision on rent can save money

Goal:

- Discover the Most Effective Predictive Models
- Find Key Covariates Influencing Rental Prices in Zurich

Outline

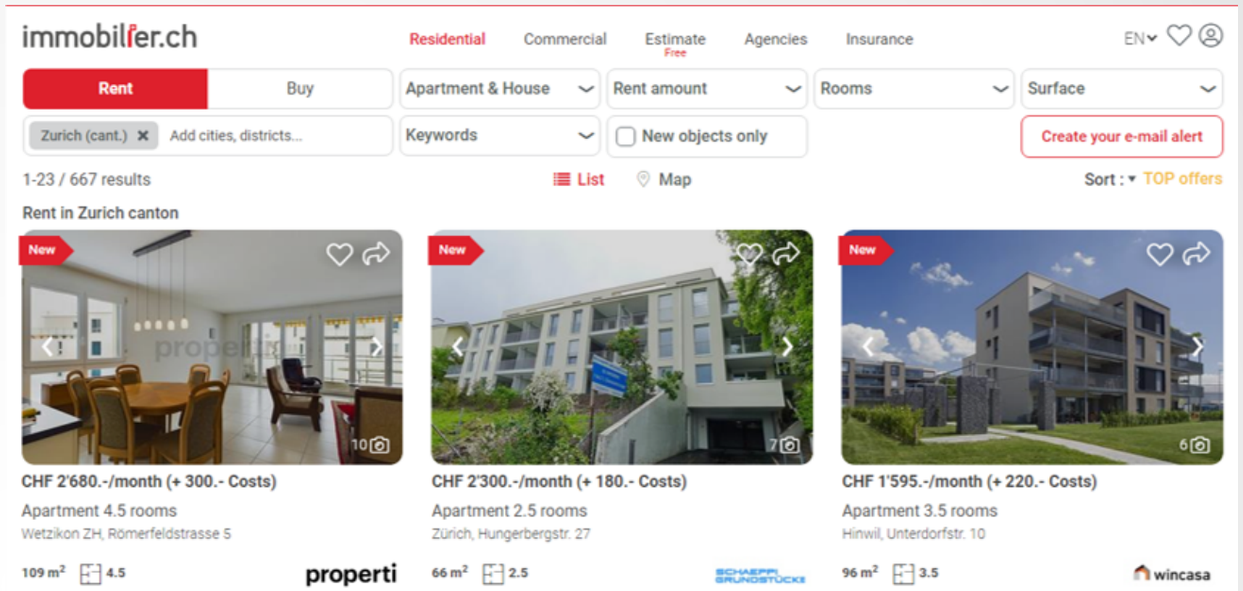
1. Motivation and Objective of the Study ✓
2. Data Collection and Preparation
3. Data Overview
   - Spatial Analysis
   - Exploratory Data Analysis (EDA)
   - Word Cloud
4. Feature Engineering and Selection
5. Predictive Models
6. Findings and Model Evaluation
7. Conclusion and Future Implication

Data Collection - Webscraping

- **Housing Advertisement Data**
  - https://www.immobilier.ch/
  - Canton of Zurich filter
  - As of November 2023

- **Web Scraping**
  - Load **“Selenium”** library
  - Download ChromeDriver
  - Run Chrome in headless mode
  - Define results page URL
  - Use WebDriverWait functions
  - Look for href links
  - Use class name selectors

Website

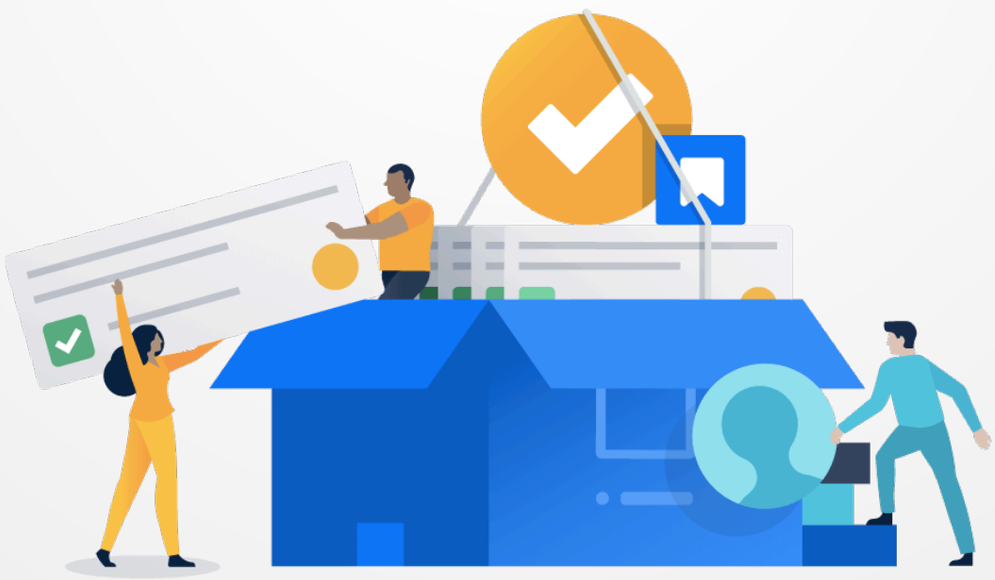


Selenium Webscraping

Unnamed: 0	description	type	Rent	Rooms	Living space	Complete Address
0		Duplex	1786	3.5	100	Marktgasse 16, 8302 Kloten
1	Schone Kloten 3.5 ZW with cellar and WT. Cl…	Furnished flat	3694	2	110	Flurstrasse, 8000 Zürich
2	Beautiful apartment with a cozy balcony	Apartment	1401	3.5	69	Grüningerstrasse 6, 8133 Esslingen
3	- NO balcony	Apartment	1620	3	64	Riedenhaldenstr. 30, 8046 Zürich
4	www.alte-landstrasse-kloten.ch	Apartment	1655	1.5	41	Alte Landstrasse 1/3/5, 8302 Kloten
5	The pictures are sample photos of a compari…	Apartment	2750	4.5	100	Rainstrasse 7, 8800 Thalwil
6	- Wooden stairs	Duplex	2250	5.5	129	Studenrain 7, 8122 Binz
7		Apartment	2153	2.5	68	Seewartweg 1, 8810 Horgen
9	- Kitchen and bathroom with slab floors - Kitchen with gas stove and oven	Apartment	970	1	32	Stöcklerstr. 8, 8610 Uster
10	Cozy 3.5 room apartment in Bachtelblick	Apartment	1590	3.5	78	Langfurrenstrasse 1a, 8623 Wetzikon ZH
11	- Parquet in the living room/bedroom - Slab flooring in kitchen and bathroom	Apartment	1045	1	30	Gutstrasse 11, 8400 Winterthur

Data Cleaning Process

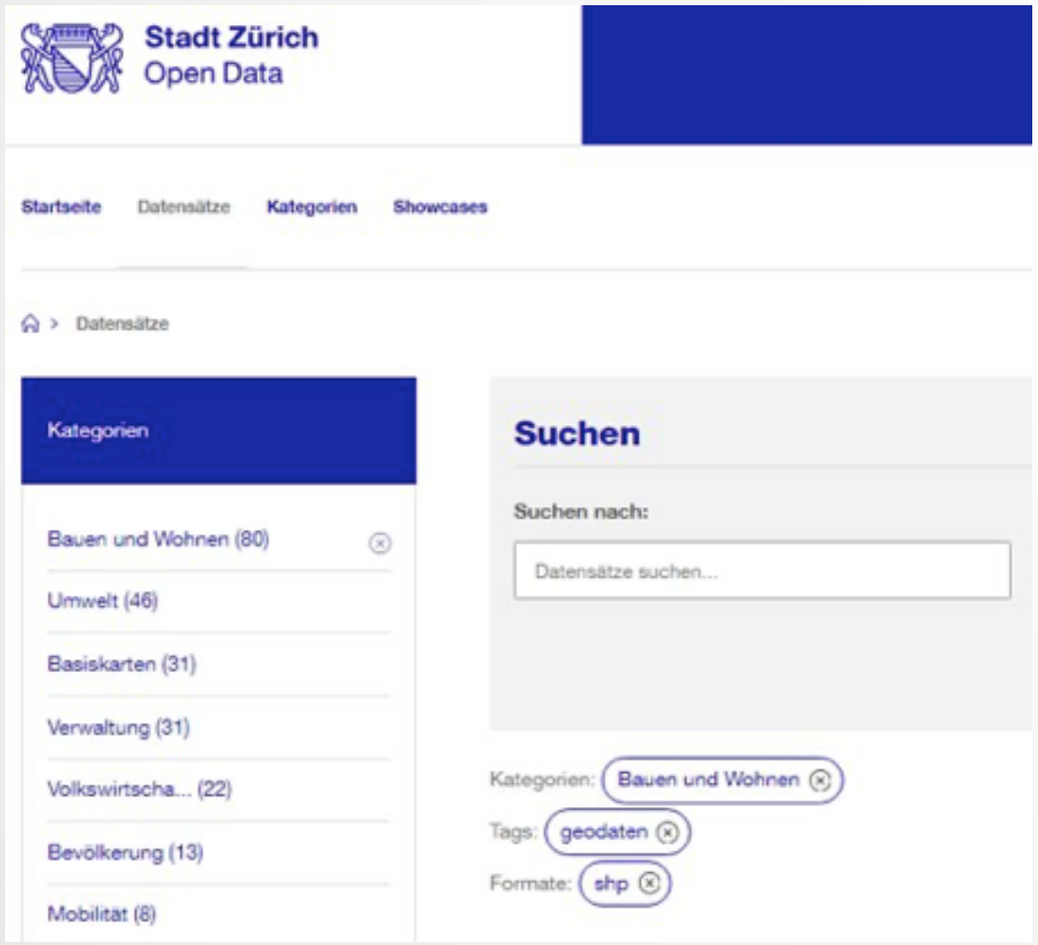
- Extract numeric values
- Remove redundant texts
- Merge address and location into a single ‘Complete Address’ column
- Remove misplaced and unnecessary information (i.e., na, nan, and etc.)
- Use the minimum common features to keep as many rows as possible
- Perform text cleaning such as removing text notation (i.e., \n, /, [‘ ‘], and etc.)
- Create new variable (i.e., price_per_sqm)
- Apply One-end-coding on Type, District, and City variables



Data Collection - Spatial Data

Geographical information is gathered from **https://data.stadt-zuerich.ch/**

- Use **“GeoPandas”** Packages to load shapefiles (Point, Linestring, Polygon)
- Use **“Nominatim”** API to convert addresses into coordinates



GeoPandas

Nominatim

```
                    geometry
POINT (8.44209 47.52614)
POINT (8.45251 47.52562)
POINT (8.45966 47.48242)
POINT (8.48854 47.47829)
POINT (8.49909 47.49290)
```

```
                                                geometry
MULTILINESTRING ((8.59864 47.36086, 8.59858 47...
LINESTRING (8.49621 47.42617, 8.49618 47.42618...
LINESTRING (8.52285 47.42373, 8.52291 47.42371...
LINESTRING (8.51109 47.36190, 8.51111 47.36187...
LINESTRING (8.56054 47.36191, 8.56050 47.36192...
```

```
                                                geometry
POLYGON ((2675196.564 1237091.031, 2675196.775...
POLYGON ((2673969.502 1238324.209, 2673988.444...
POLYGON ((2690672.926 1240725.647, 2690673.356...
POLYGON ((2712915.028 1241517.383, 2712922.843...
POLYGON ((2683476.938 1243307.523, 2683478.274...
```

Variable Creation from Spatial Data

- Calculate the distance between amenities and properties
- Obtain **Public_Transport, Gardens, PubliBike, and Park variables**

Creating New Variables from Spatial Data

```python
# Function to calculate distance to the nearest point in 'gardens'
def calculate_nearest_distance(row, gardens):
    distances = gardens.geometry.distance(row['geometry'])
    min_distance = distances.min()
    return min_distance

# Apply the function to each row in the 'data' GeoDataFrame
data['Public_transport'] = data.apply(lambda row: calculate_nearest_distance(row, public_transports), axis=1)

# Function to return the year of the building (nearest point in building_age)
def calculate_nearest_GBAUJ(row, building_age):
    distances = building_age.geometry.distance(row['geometry'])
    min_distance_idx = distances.idxmin()
    nearest_GBAUJ = building_age.at[min_distance_idx, 'GBAUJ']
    return nearest_GBAUJ

# Add a new column 'nearest_GBAUJ' to store the GBAUJ values of the nearest geometries
data['Year'] = data.apply(calculate_nearest_GBAUJ, building_age=building_age, axis=1)

# Create a new column 'Neighborhood' and set the default value to None
data['City'] = None

# Iterate through each polygon and assign the neighborhood name
for index, row in city.iterrows():
    city_name = row['GEMEINDENA']
    data.loc[data['geometry'].within(row['geometry']), 'City'] = city_name

#data.to_csv('Immobilier.ch_database_ZH_canton_completed.csv')

data = data.drop(columns=['Complete Address', 'latitude','longitude','geometry'])
data = data.loc[:, ~data.columns.str.contains('^Unnamed')]
#data.to_csv('Immobilier.ch_database_ZH_7_cov.csv')
```

Data Integration

- Create new columns in the datasets with calculated values
- More spatial covariates in City of Zurich dataset as many shapefiles cover only this area

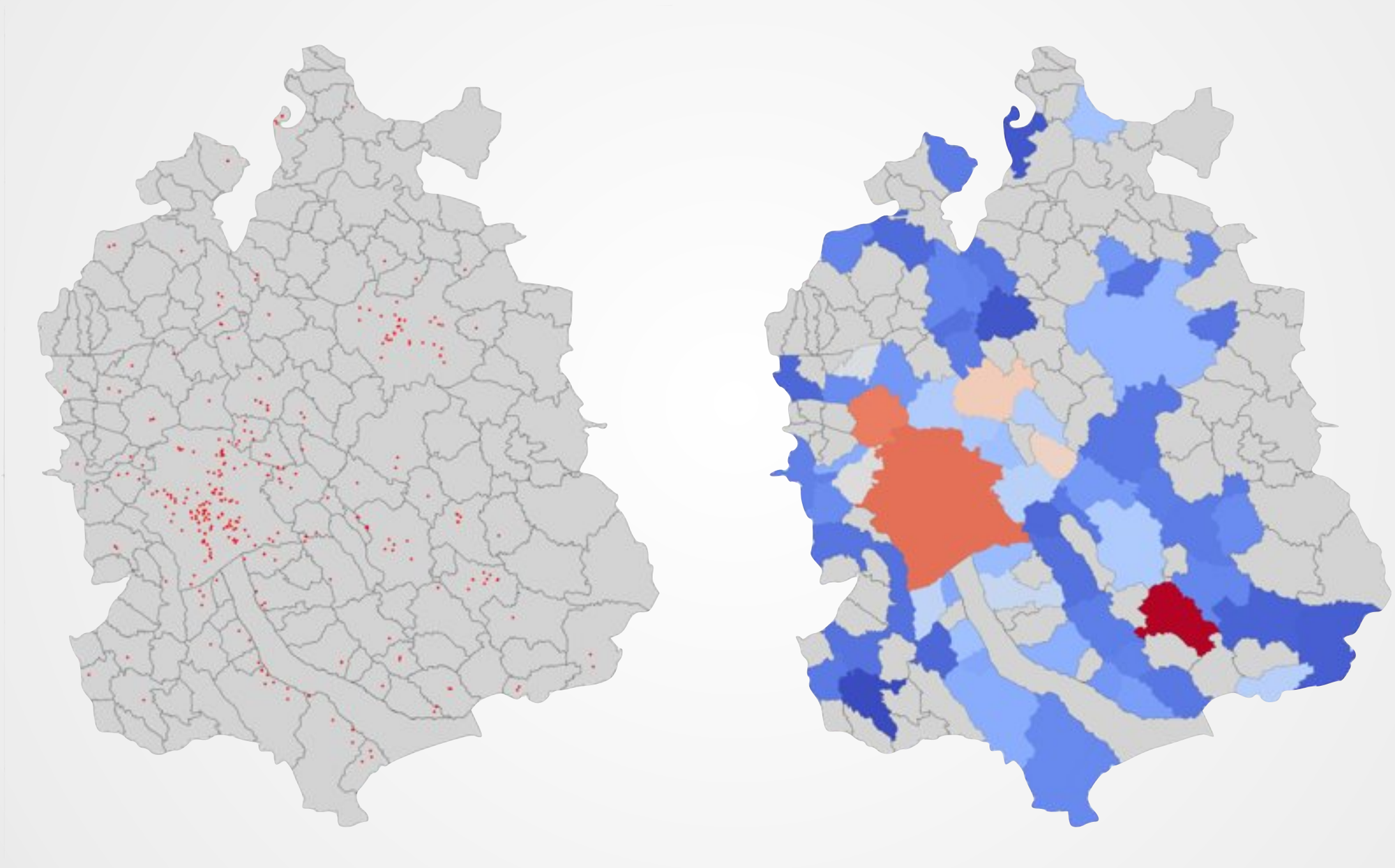
Canton of Zurich

```
In [3]: data_zh.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 401 entries, 0 to 409
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   description      401 non-null    object
 1   type             401 non-null    object
 2   Rent             401 non-null    float64
 3   Rooms            401 non-null    float64
 4   Living space     401 non-null    float64
 5   Public_transport 401 non-null    float64
 6   Year             401 non-null    float64
 7   City             401 non-null    object
 8   price_per_sqm    401 non-null    float64
dtypes: float64(6), object(3)
memory usage: 31.3+ KB
```

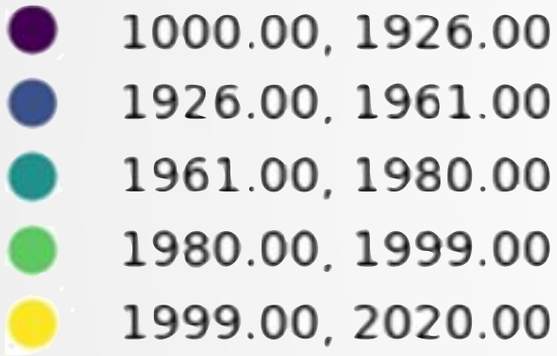
City of Zurich

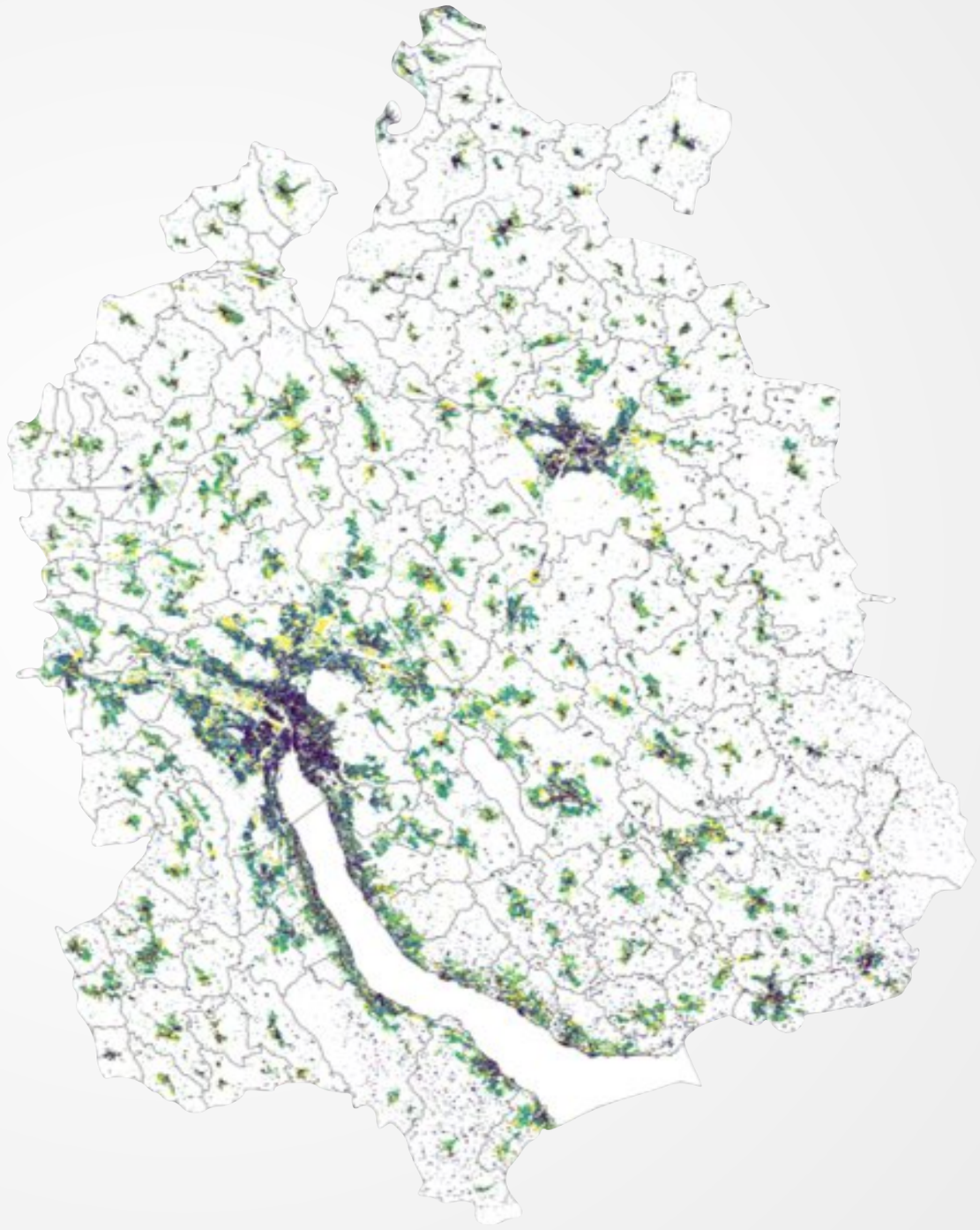
```
In [4]: data_czh.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 185 entries, 0 to 184
Data columns (total 13 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   description      185 non-null    object
 1   type             185 non-null    object
 2   Rent             185 non-null    float64
 3   Rooms            185 non-null    float64
 4   Living space     185 non-null    float64
 5   geometry         185 non-null    object
 6   Gardens          185 non-null    float64
 7   PubliBike        185 non-null    float64
 8   Park             185 non-null    float64
 9   Public_transport 185 non-null    float64
 10  Year             185 non-null    float64
 11  District         185 non-null    object
 12  price_per_sqm    185 non-null    float64
dtypes: float64(9), object(4)
memory usage: 18.9+ KB
```

Spatial Analysis - Cities Borders

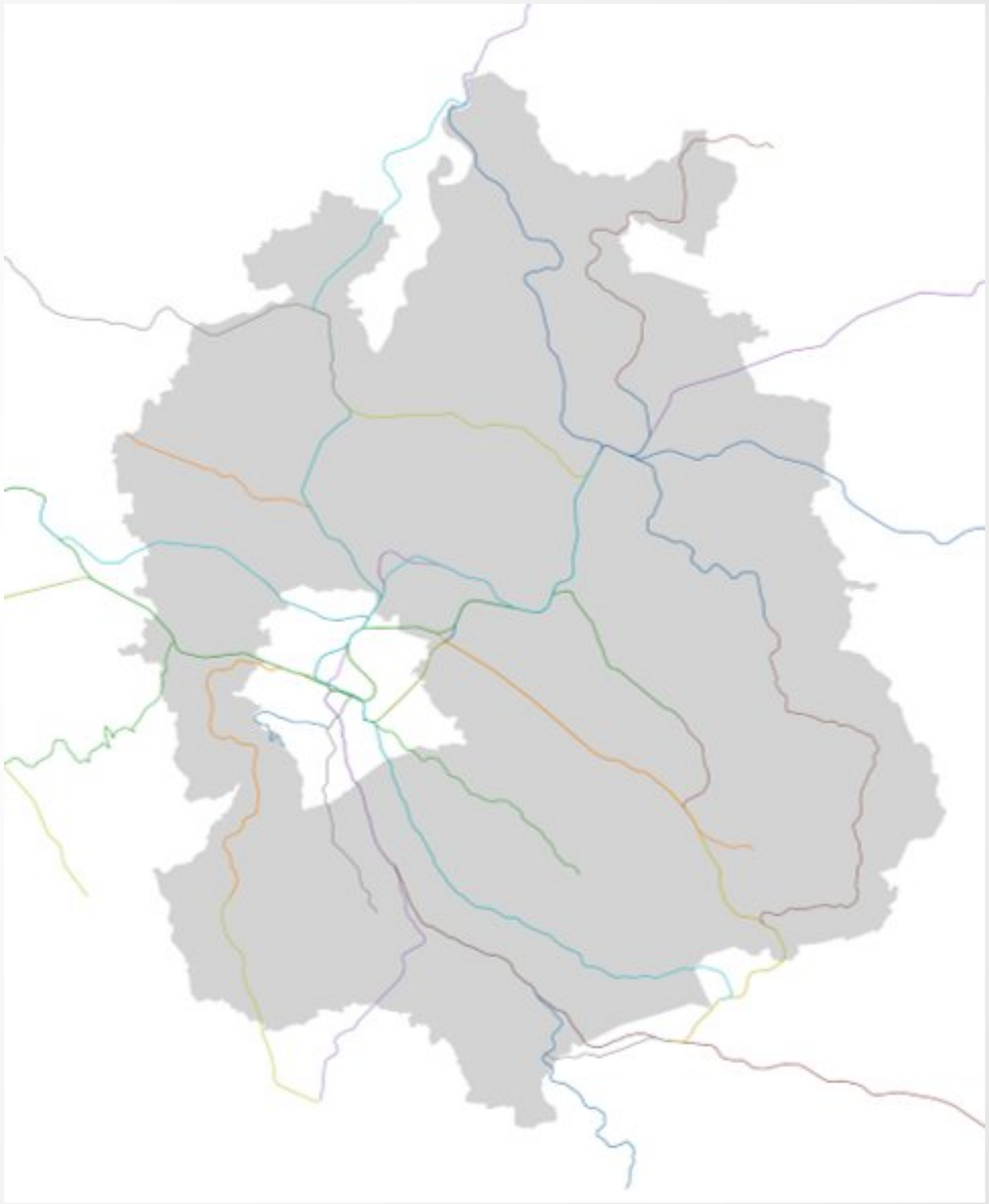


Spatial Analysis - Age of the buildings

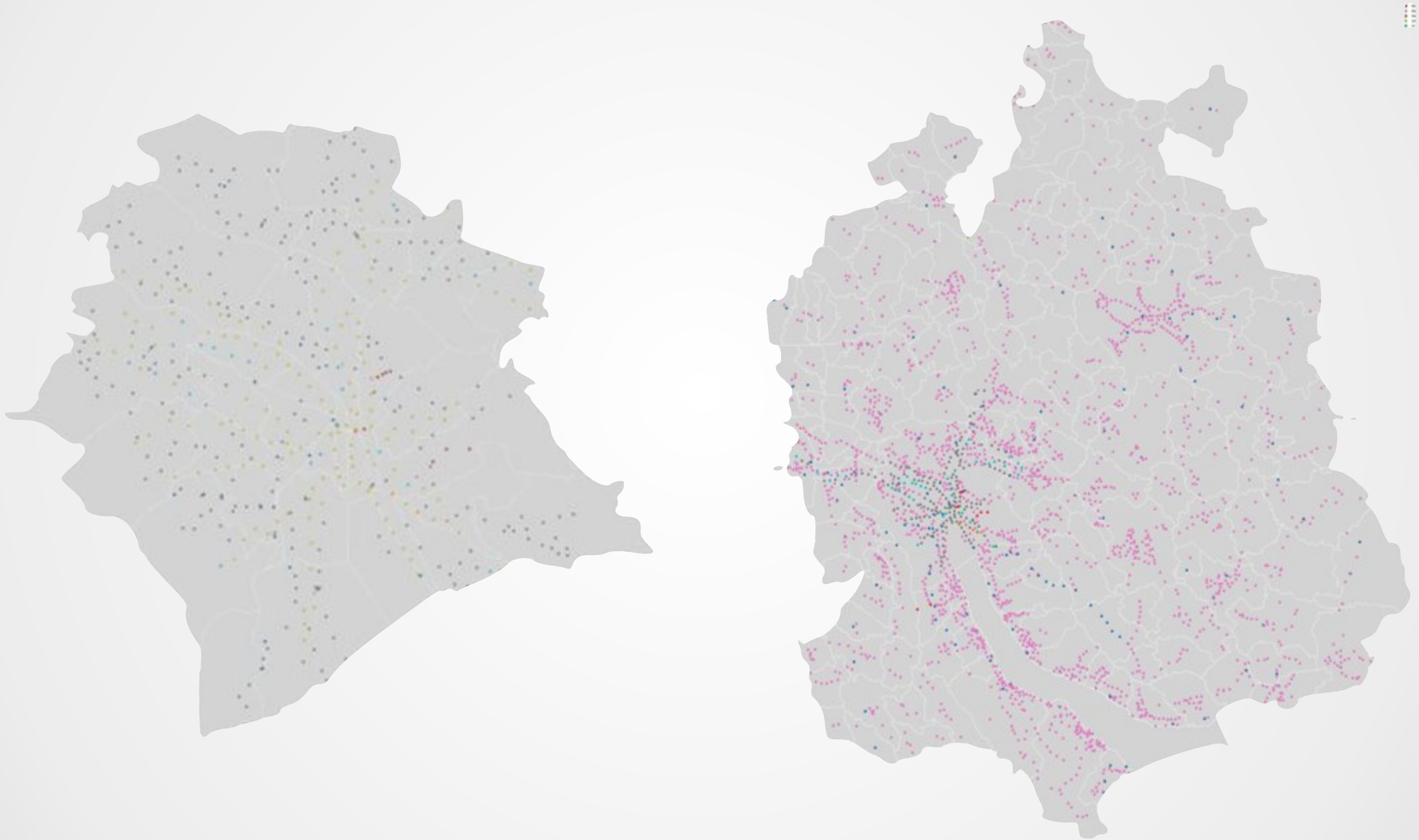




Spatial Analysis - Public Transports Line

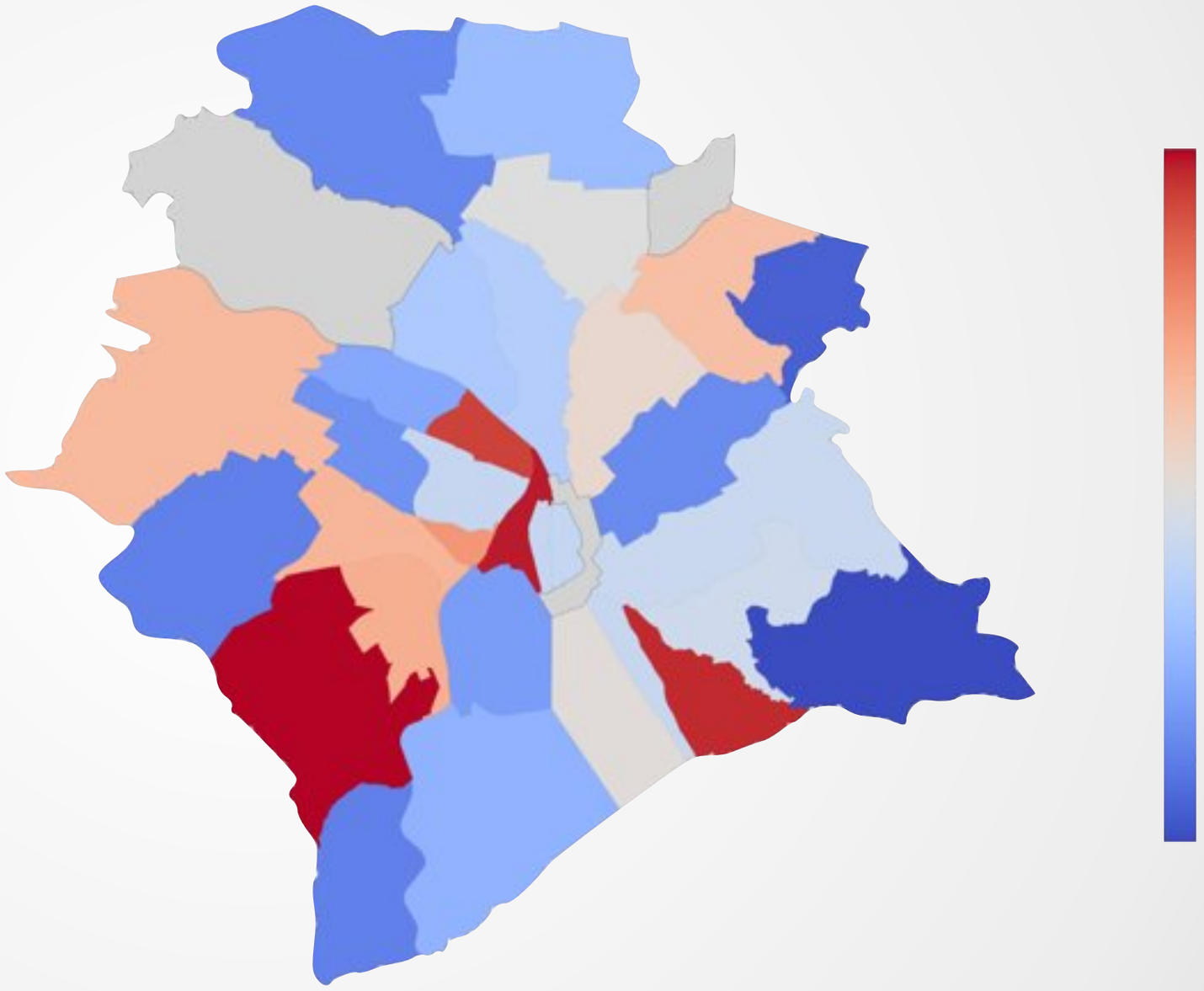


Spatial Analysis - Public Transport Stop

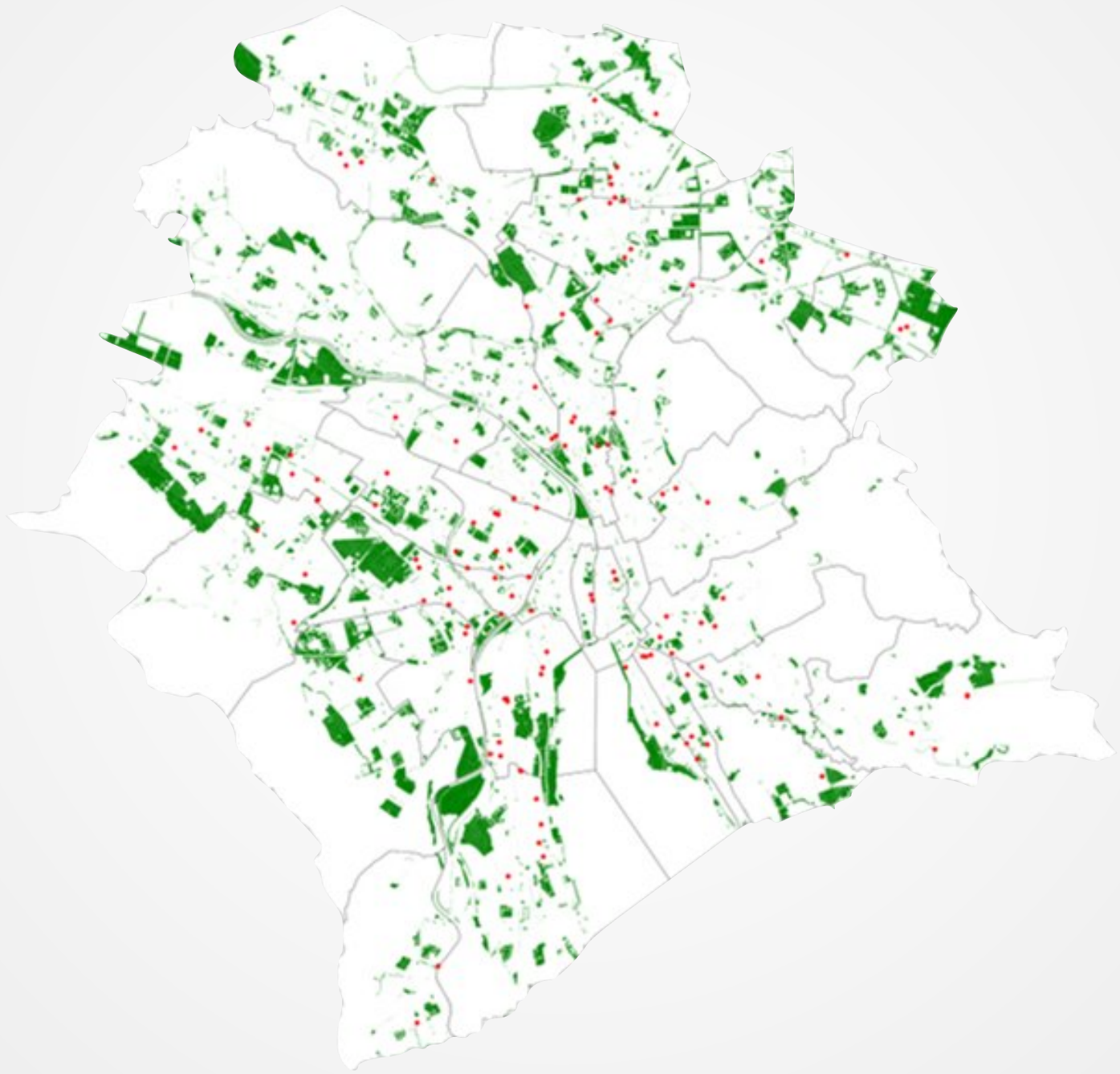


Spatial Analysis - Zurich district

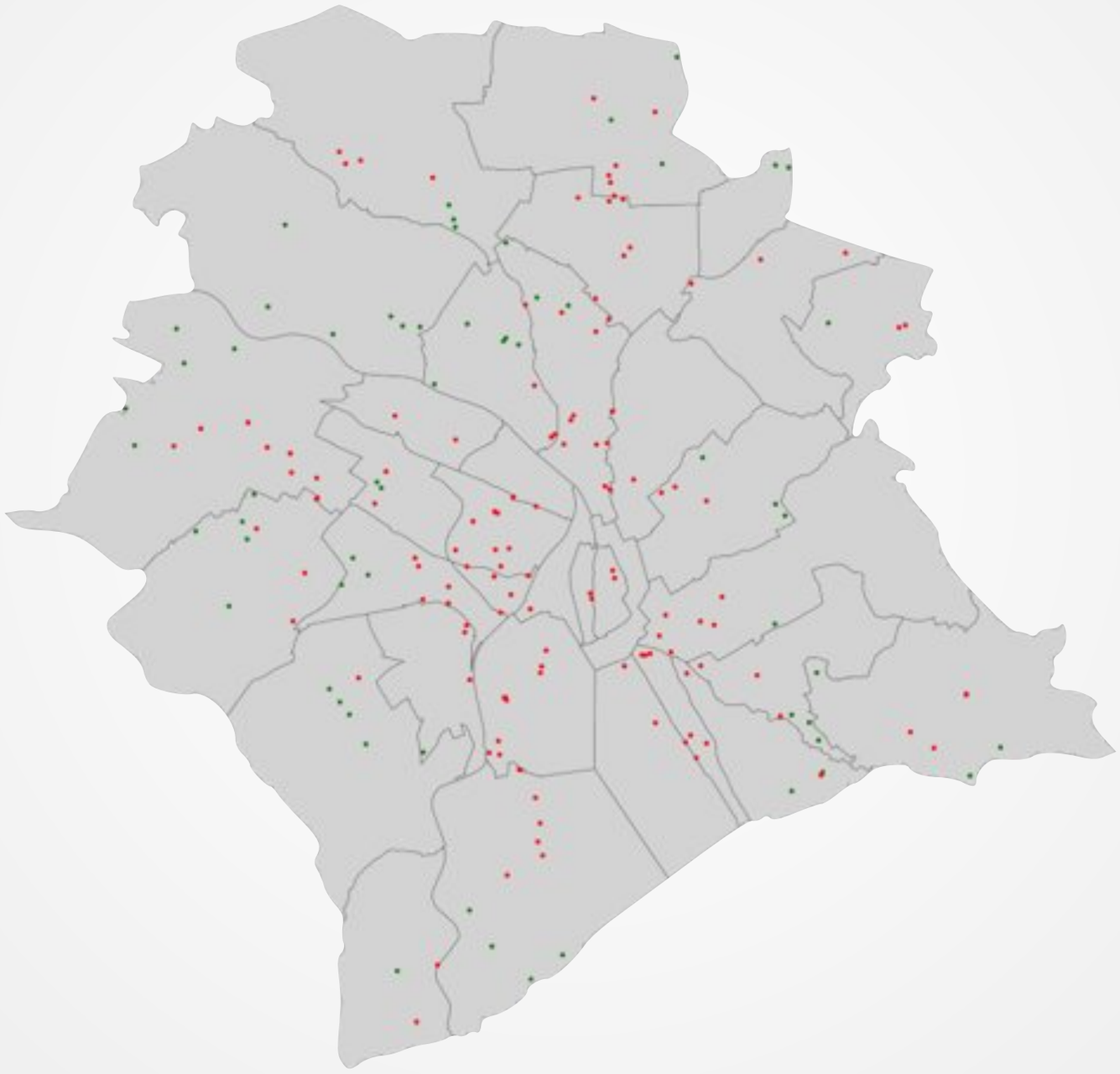
count	31.000000
mean	746.569201
std	304.251110
min	266.776268
25%	495.773223
50%	718.264615
75%	965.634515
max	1340.727273



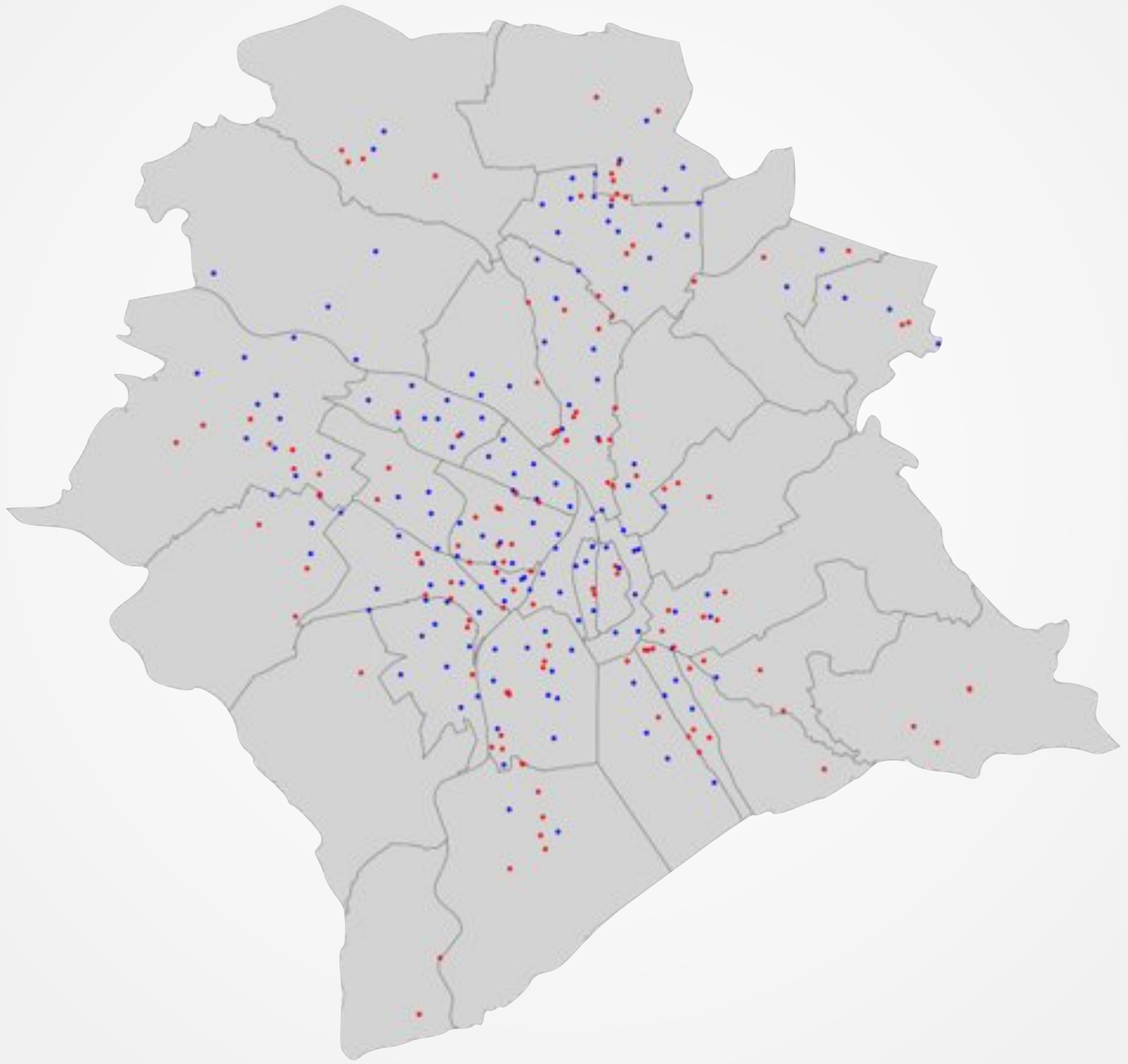
Spatial Analysis - Public Parks



Spatial Analysis - Family Gardens

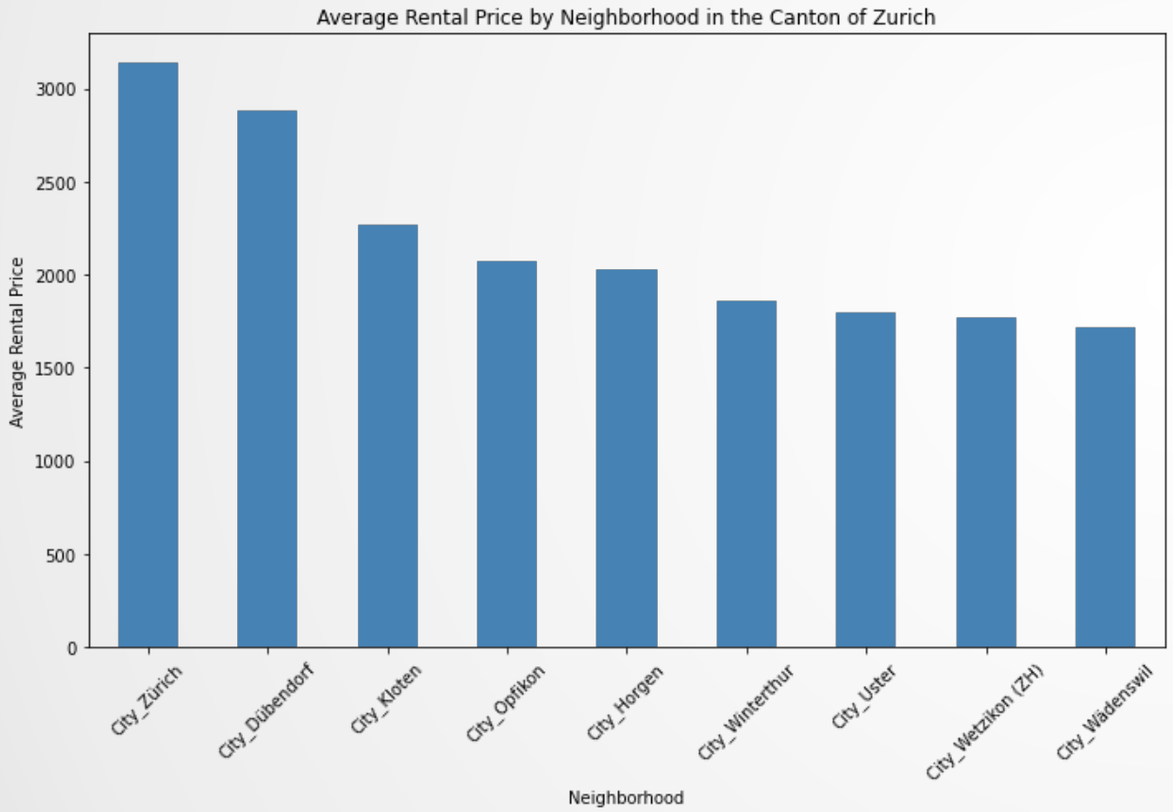


Spatial Analysis - PubliBike

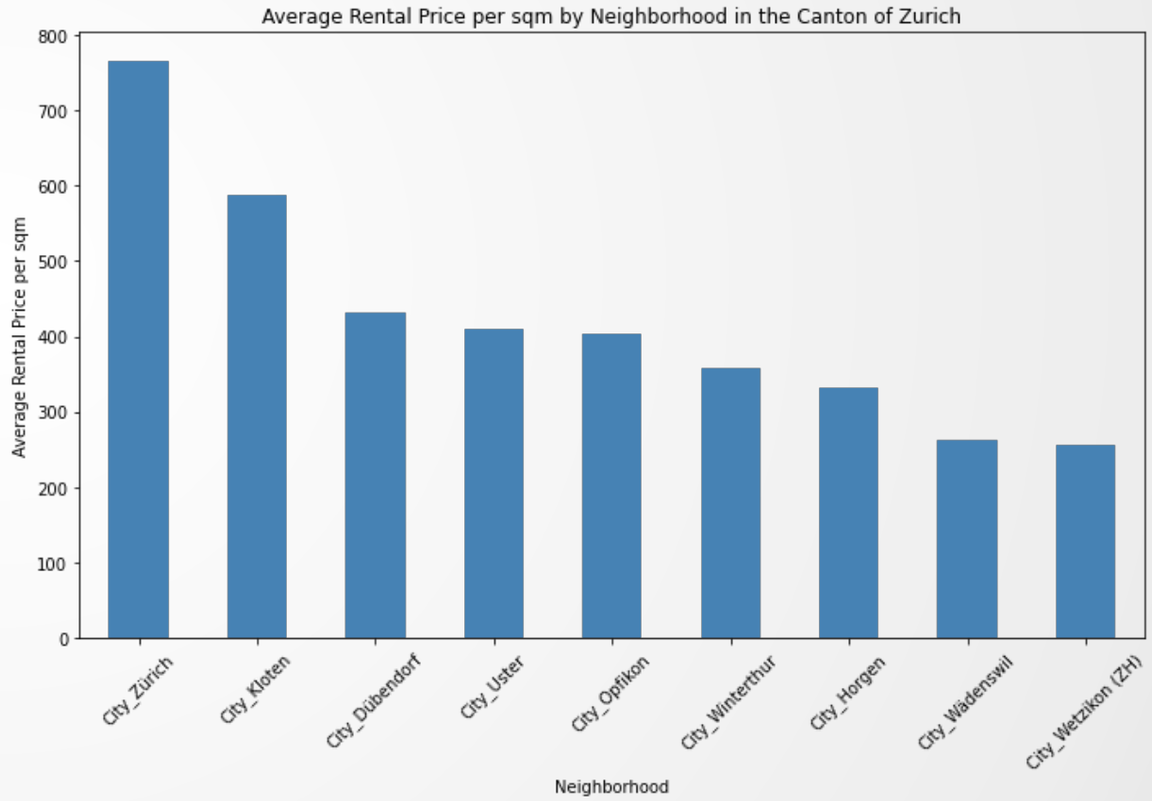


Exploratory Data Analysis (EDA)

Average Rental Prices by Neighborhood

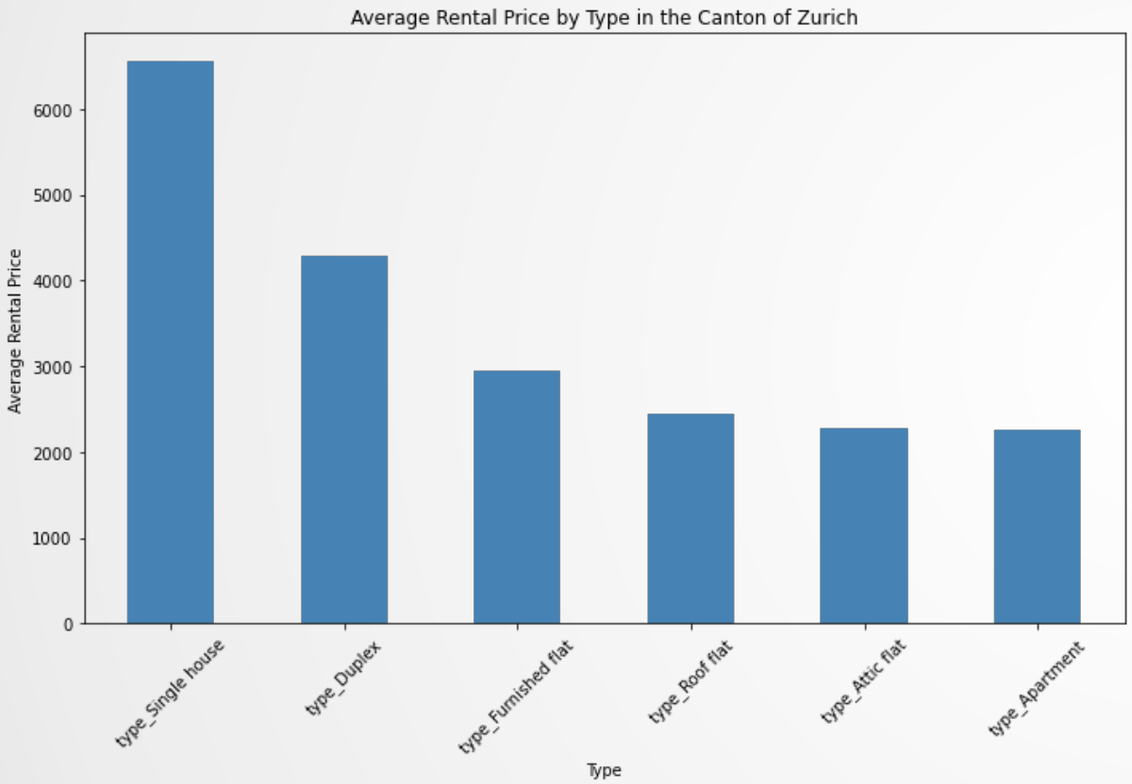


Average Rental Prices per sqm by Neighborhood

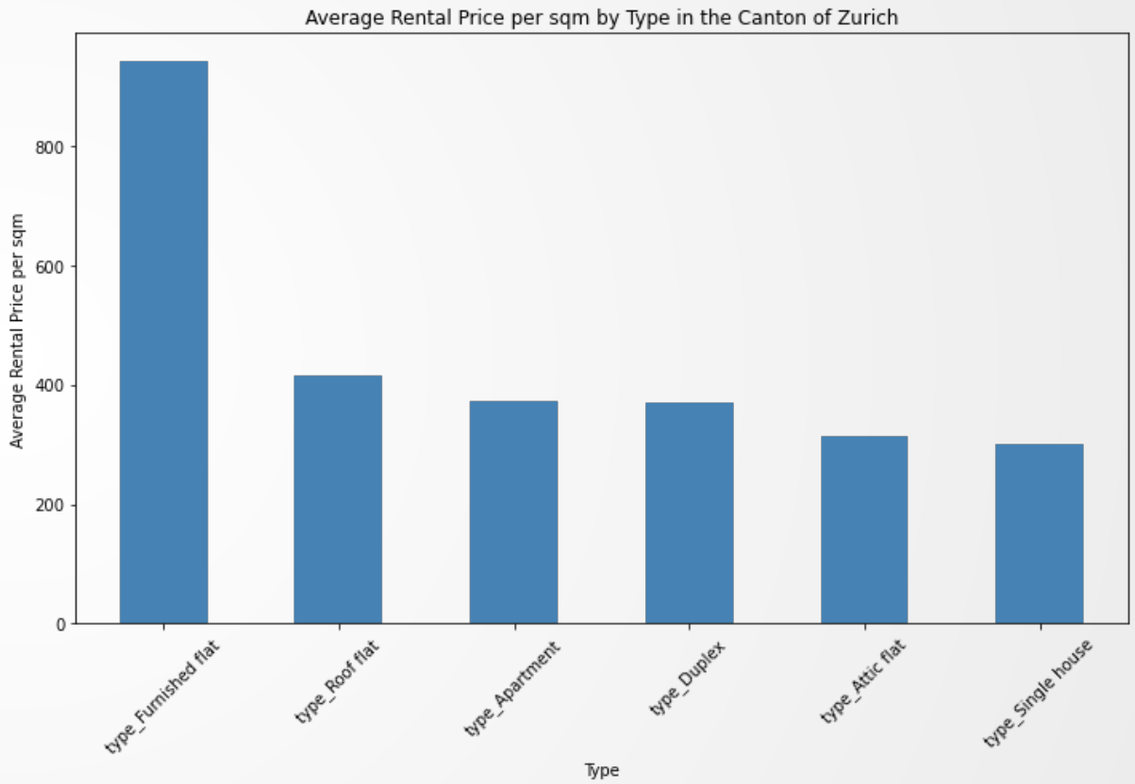


Exploratory Data Analysis (EDA)

Average Rental Prices by Types



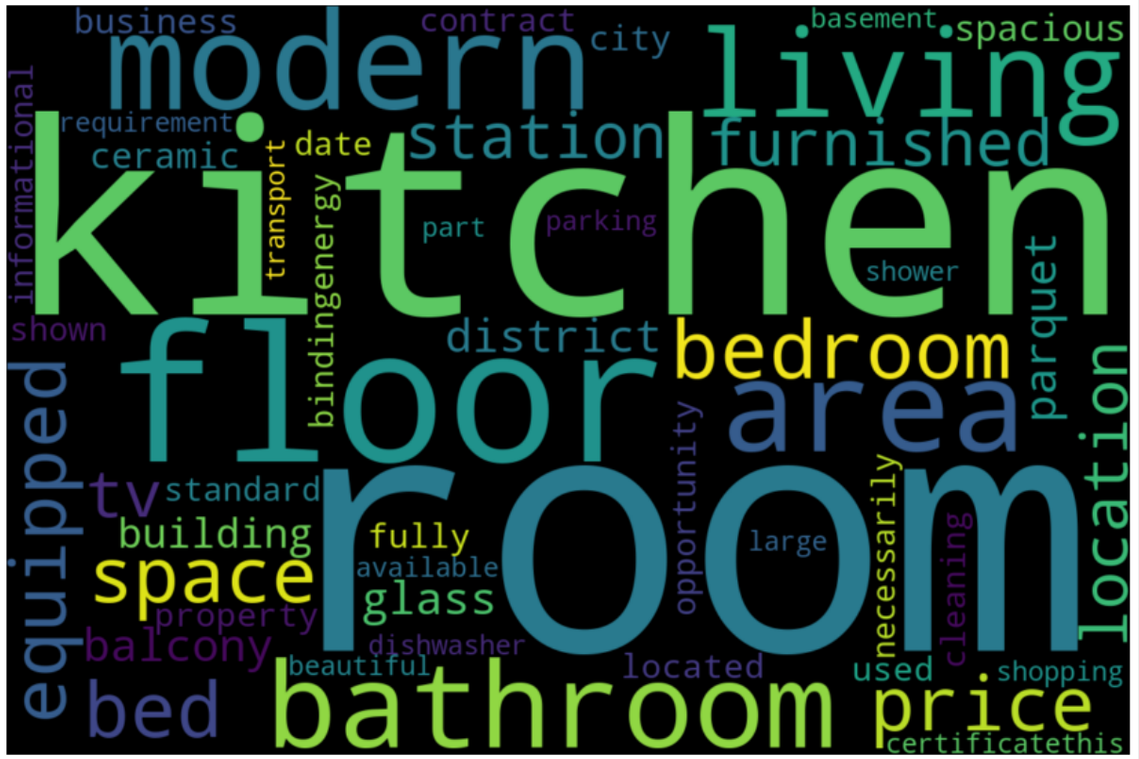
Average Rental Prices per sqm by Types



Word Cloud

Canton of Zurich

City of Zurich



Keywords: Kitchen, floor, bathroom, bedroom, modern, space, balcony, and etc.

Variables that are used in this study

Variable	Description	Class Type	Role in Our model
Rent	The monthly rental cost for each property	Numeric	Dependent Variable (ZH and CZH)
Rooms	Total count of rooms in each property	Numeric	Independent Variable (ZH and CZH)
Living space	The total area of each property (square meters)	Numeric	Independent Variable (ZH and CZH)
Public transport	The shortest distance from each property to the nearest public transport	Numeric	Independent Variable (ZH and CZH)
Type	The type of each property (One-hot encoding)	Character	Independent Variable (ZH and CZH)
City/District	The location of each property (One-hot encoding)	Character	Independent Variable (ZH and CZH)
Year	The year in which each property was built or last renovated	Numeric	Independent Variable (ZH and CZH)
Garden	The shortest distance from each property to the nearest garden	Numeric	Independent Variable (CZH)
PubliBike	The shortest distance from each property to the nearest Publibike station	Numeric	Independent Variable (CZH)
Park	The shortest distance from each property to the nearest park	Numeric	Independent Variable (CZH)
Description	The textual description of each property	Character	Excluded (ZH and CZH)
Price_per_sqm	Price per square meter	Numeric	Excluded (ZH and CZH)

Chosen Models

Eight different models are utilized within this paper for predicting rental prices including:

- Hedonic Pricing Model
- Lasso Regression Regularization
- Ridge Regression Regularization
- Decision Trees
- Random Forest Regression
- Gradient Boosting Models (i.e., XGBoost)
- Support Vector Regression (SVR)
- Neural Network (NN)

Reasons of Use for Each Model

Models	Reasons
Hedonic Pricing Model	Ideal for analyzing real estate markets by assessing the impact of property features on rental prices.
Lasso Regression Model	Enhances feature selection and model interpretability, preventing overfitting by using the absolute value of coefficients.
Ridge Regression Model	Addresses multicollinearity in models with correlated predictors, stabilizing estimates through squared coefficients.
Decision Trees	Provide interpretable decision-making structures, easily managing non-linear feature-target relationships.
Random Forest	Offers greater accuracy and robustness by averaging predictions from multiple decision trees.
XGBoost	Combines weak models (often decision trees) for enhanced performance and speed in predictive modeling.
SVR	Uses techniques like Gaussian or RBF to handle non-linear relationships, offering flexibility and accuracy in various datasets.
NN	Suitable for complex, high-dimensional datasets, adapting to intricate data patterns effectively.

The Performance of Predictive Models - Canton of Zurich

Models	MSE	R2
Hedonic Pricing Model	1.146277E+06	0.524280
Lasso Regression Model	1.220380E+06	0.493526
Ridge Regression Model	1.198624E+06	0.502555
Decision Trees	1.013893E+06	0.579221
Random Forest	8.393573E+05	0.651655
XGBoost	**8.168799E+05**	**0.660984**
SVR	2.548098E+06	-0.057495
NN	2.366691E+06	0.017791

For the **Canton of Zurich dataset**, **XGBoost** yields the lowest MSE and the highest R-squared of all techniques.

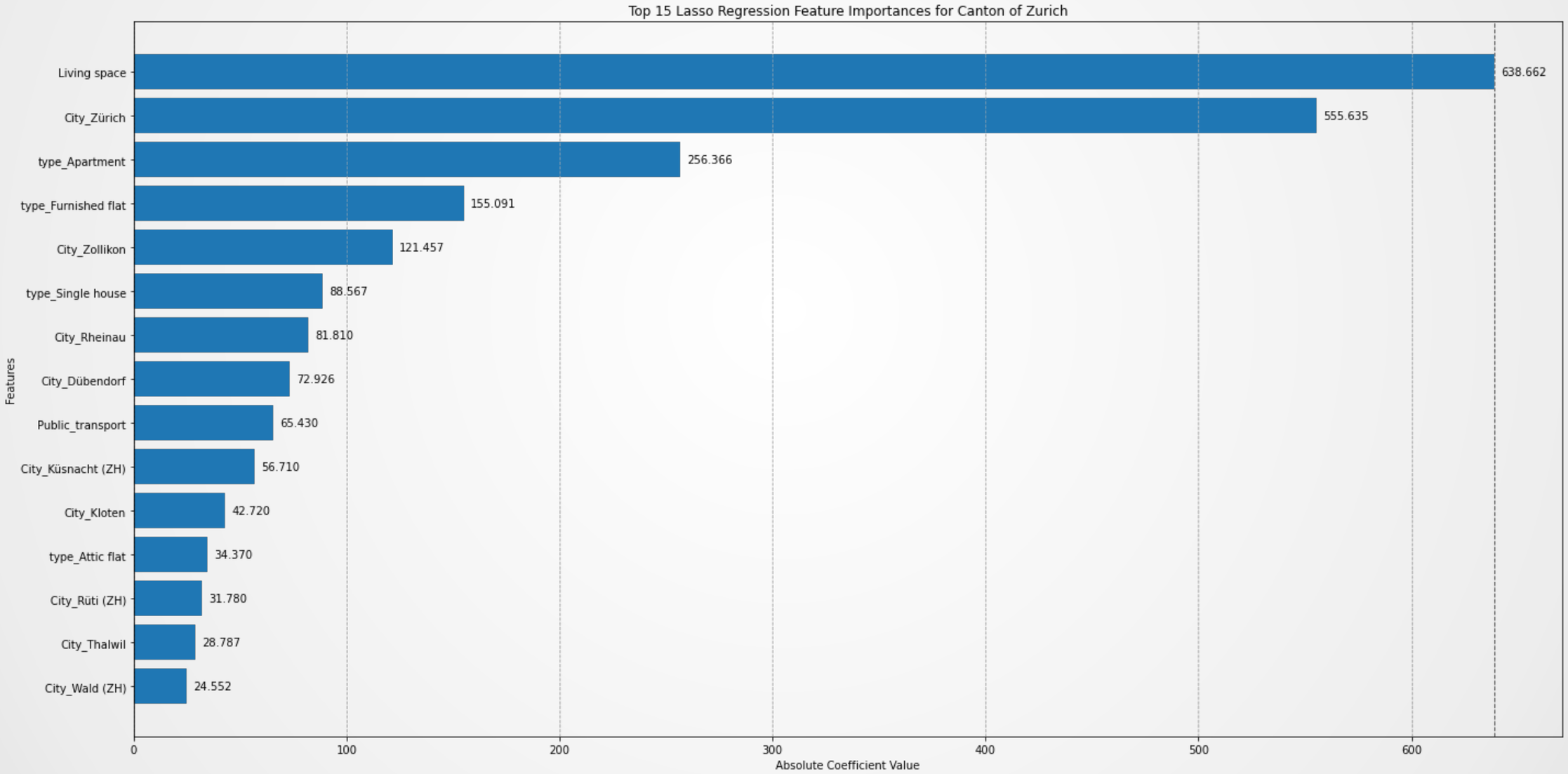
The Performance of Predictive Models - City of Zurich

Models	MSE	R2
Hedonic Pricing Model	5.369046E+05	0.653879
Lasso Regression Model	5.250769E+05	0.661504
Ridge Regression Model	4.953976E+05	0.680637
Decision Trees	7.401345E+05	0.522865
Random Forest	**4.861917E+05**	**0.686572**
XGBoost	5.818541E+05	0.624902
SVR	1.551433E+06	-0.000146
NN	1.555887E+06	-0.003017

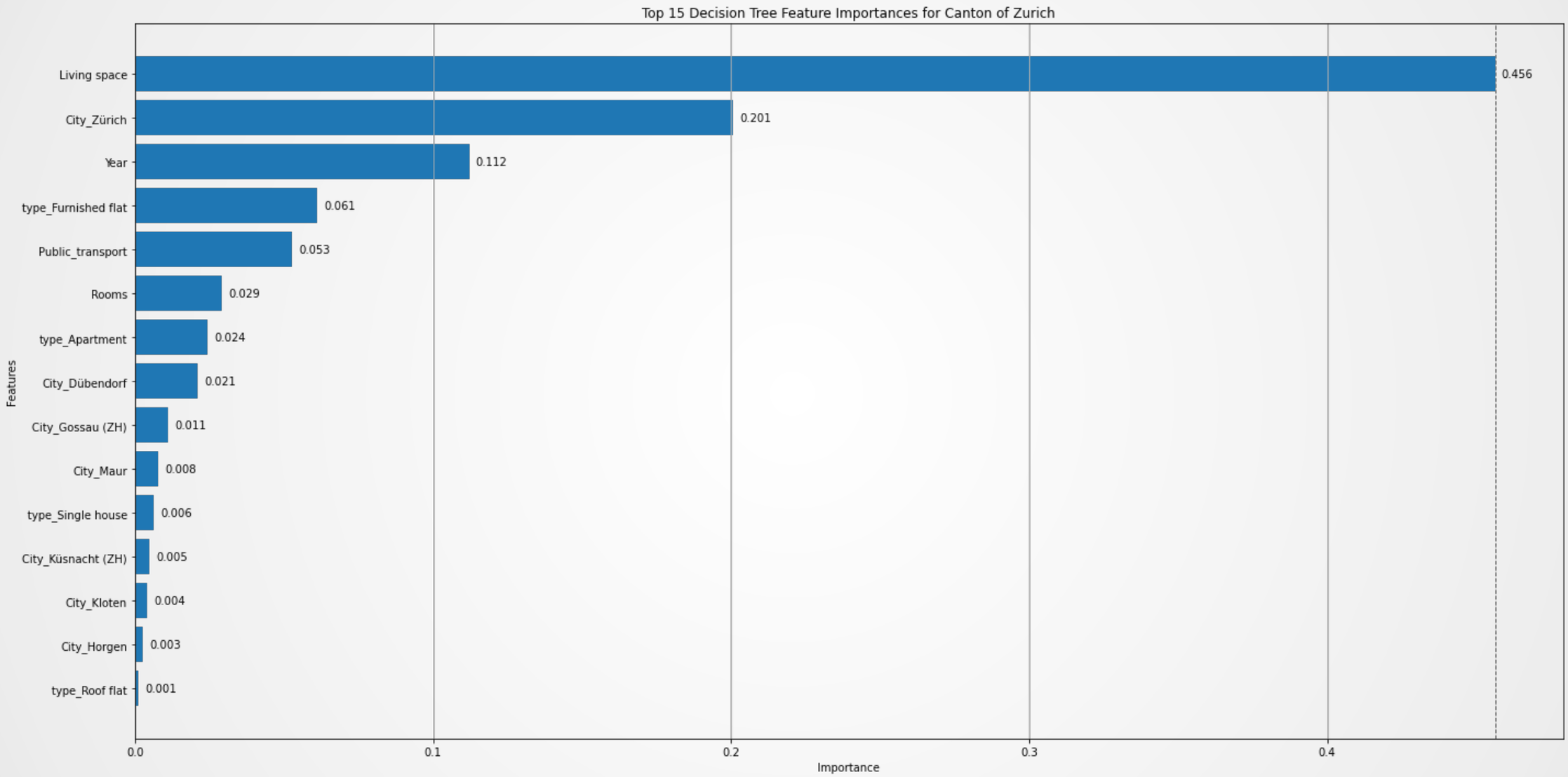
For the **City of Zurich dataset**, the **Random Forest** technique exhibits the lowest MSE and the highest R-squared of all the techniques

Variable Importance Analysis - Canton of Zurich

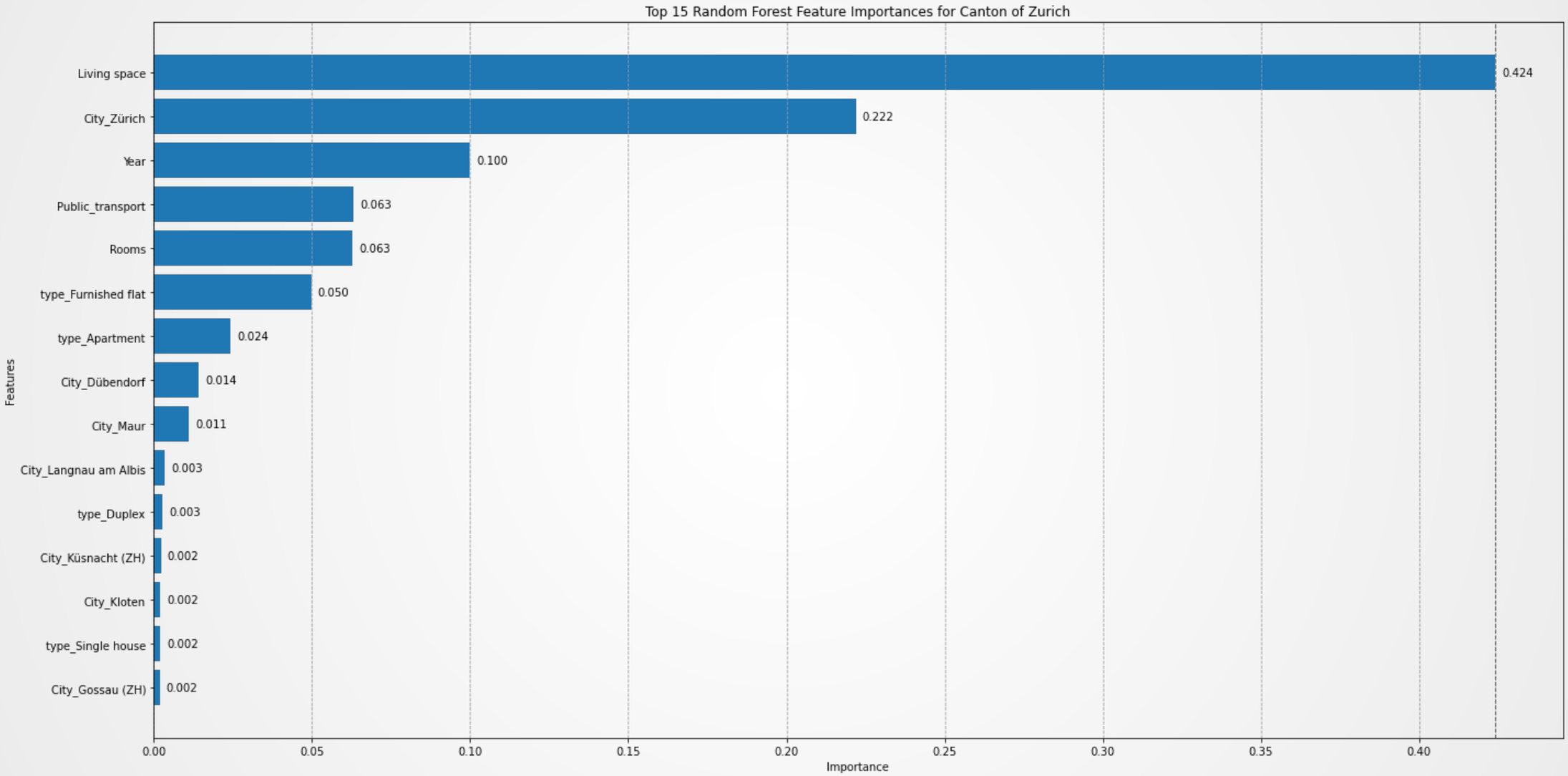
Top 15 Lasso Regression Variable Importance Analysis



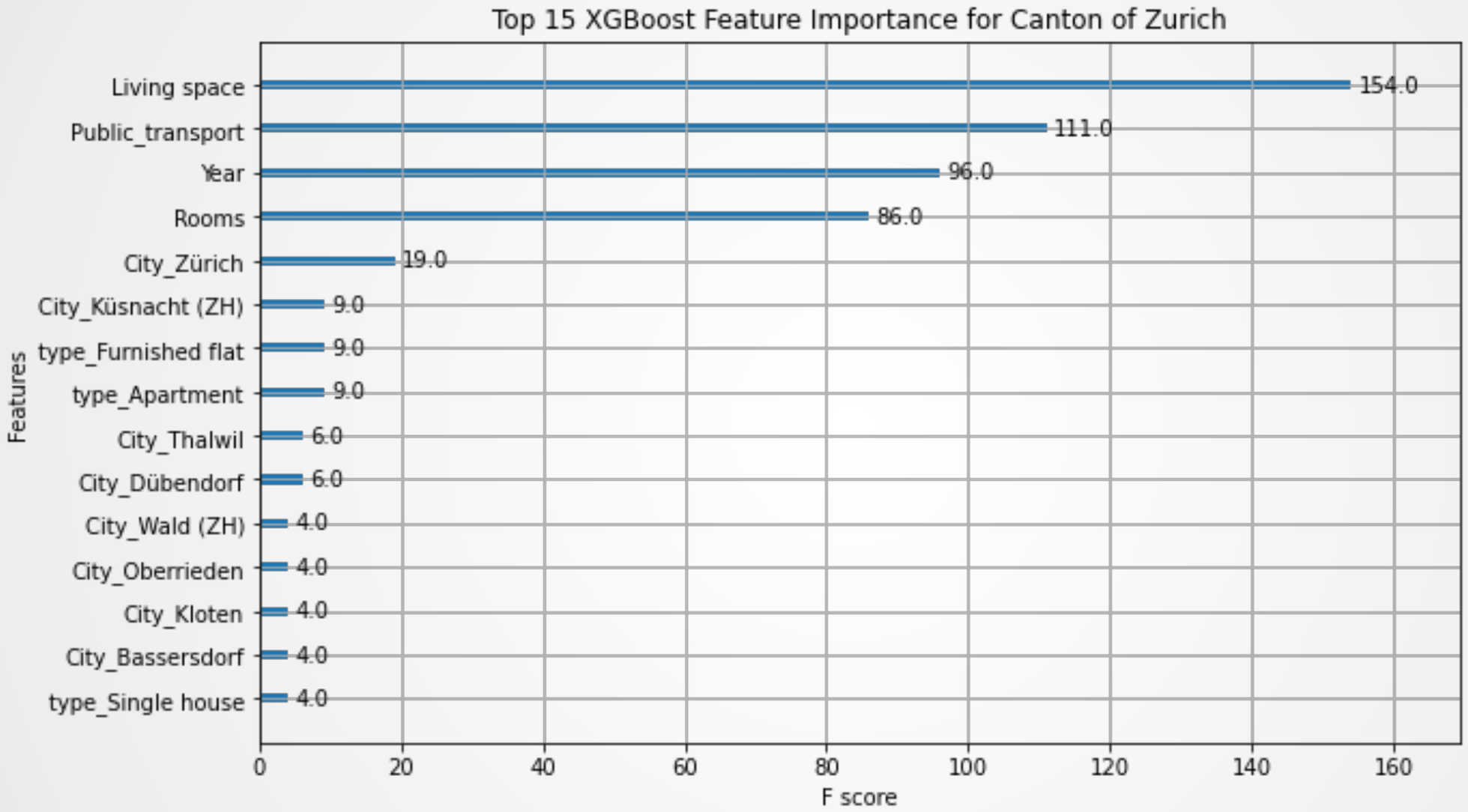
Top 15 Decision Tree Variable Importance Analysis



Top 15 Random Forest Variable Importance Analysis

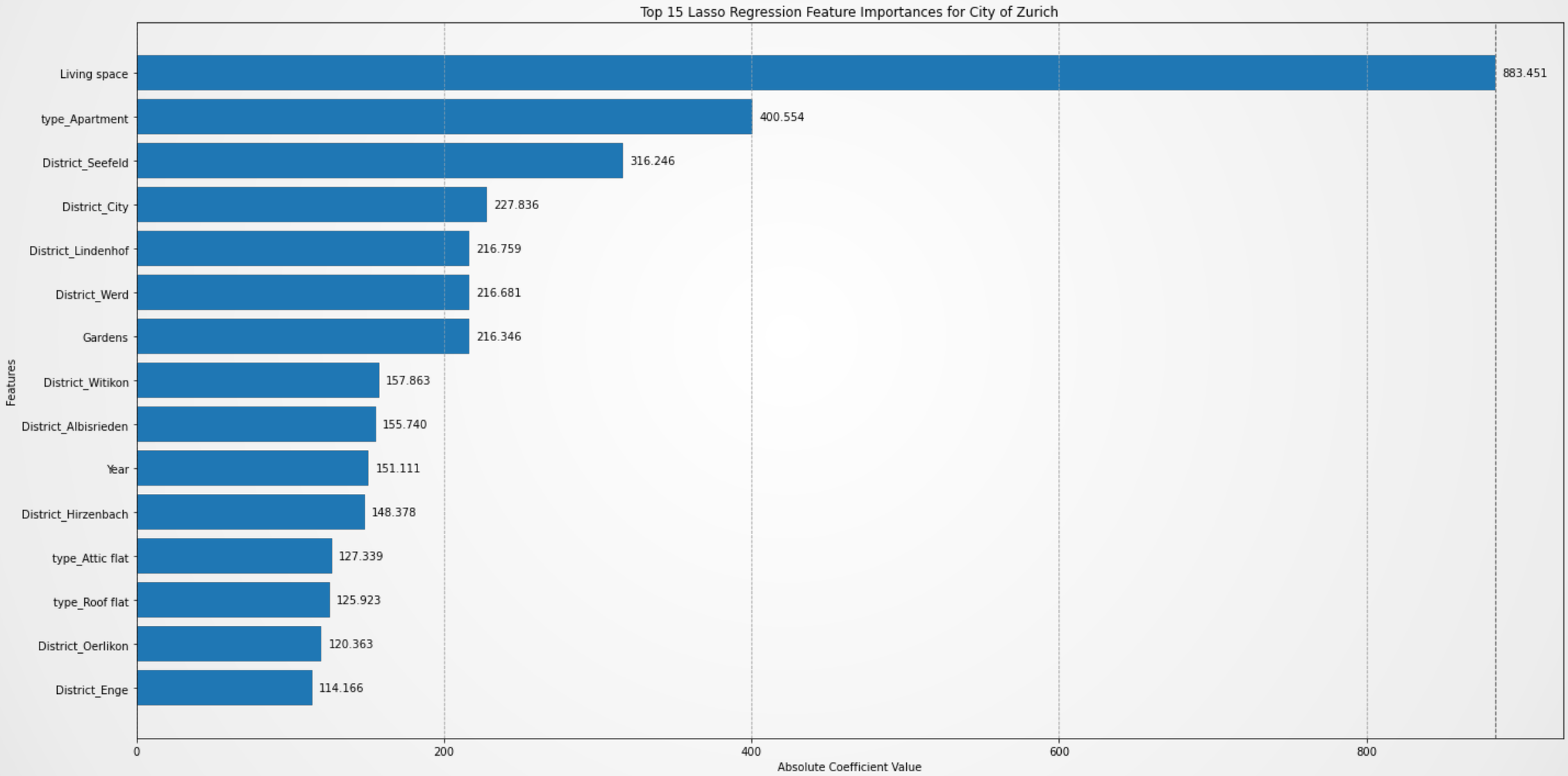


Top 15 XGBoost Variable Importance Analysis

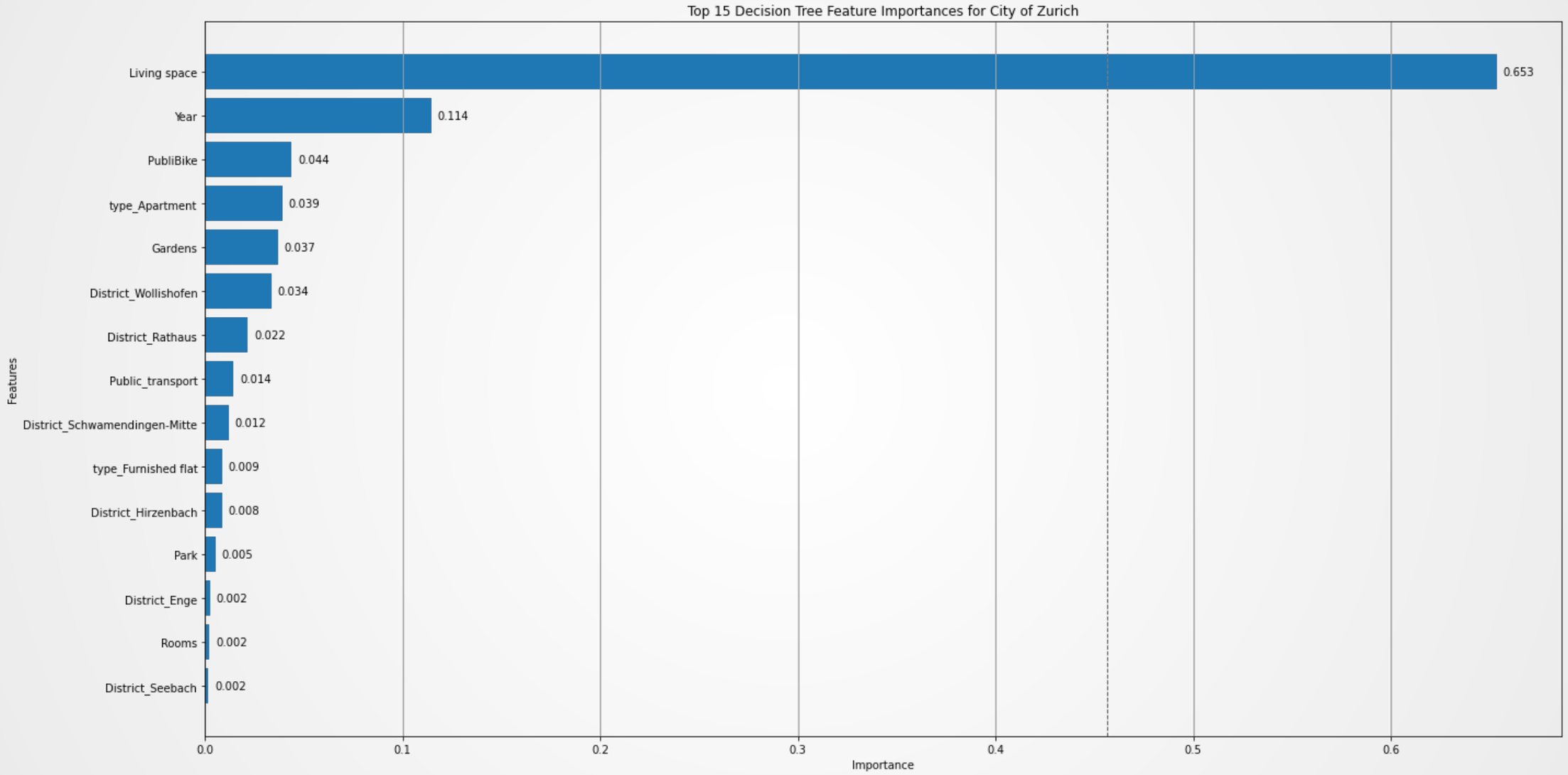


Variable Importance Analysis - City of Zurich

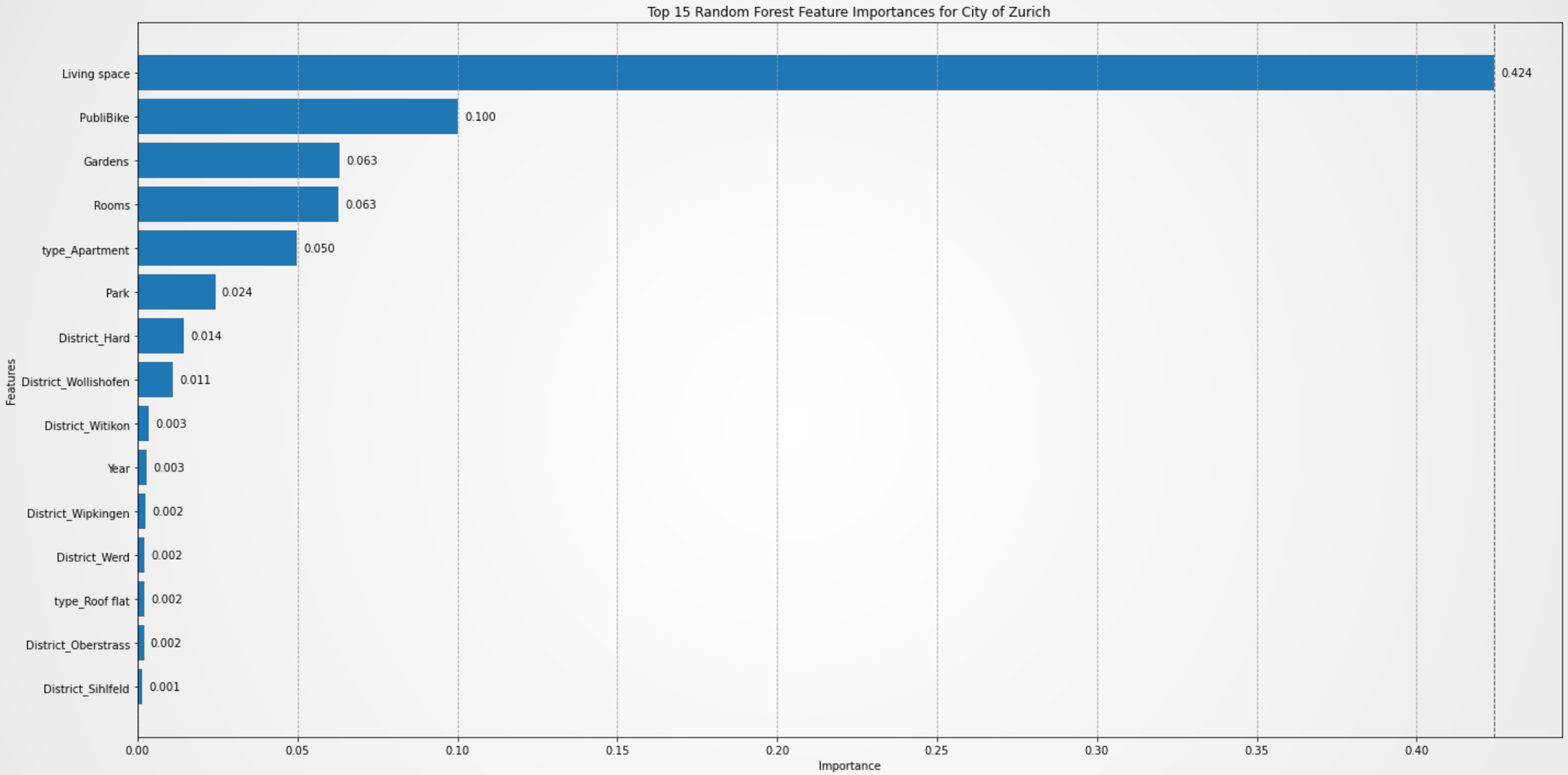
Top 15 Lasso Regression Variable Importance Analysis



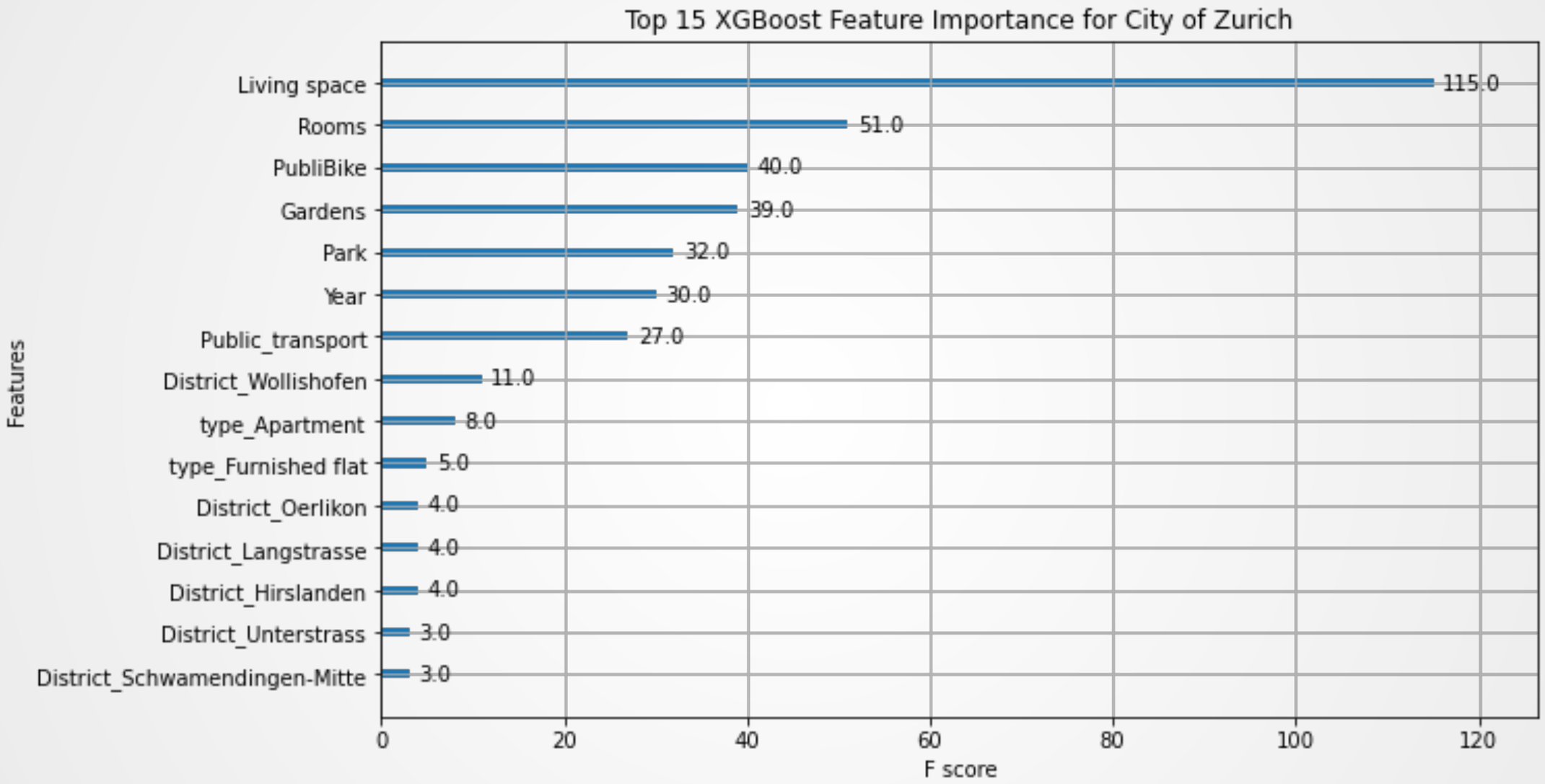
Top 15 Decision Tree Variable Importance Analysis



Top 15 Random Forest Variable Importance Analysis



Top 15 XGBoost Variable Importance Analysis



Conclusion

- **Best Method:**
  - **XGBoost** excels in the Canton of Zurich dataset, and **Random Forest** in the City of Zurich dataset, highlighting tree-based models' effectiveness in rental price prediction.
- **Important Covariates and Effect of adding covariates:**
  - Key factors include **"living space"** and the impactful addition of spatial data, especially in the City dataset
- **Generalization:**
  - Models trained on the Canton dataset might be more generalizable across the region
  - Those trained on the City dataset could be more accurate for urban-specific predictions.

In conclusion, the analysis shows that property values are determined not only by internal attributes, but also significantly by surrounding environment and neighborhood amenities.

Potential Improvement

- **Adding covariates:** Enhancing the model with macroeconomic factors, policy impacts, demographics, and market dynamics
- **Text mining techniques or Sentimental Analysis:** Employing text mining for deeper insights
- **Normalized distribution:** Applying transformations such as logarithmic or square root scaling
- **Other machine learning techniques:** Utilizing other kernel trick for SVR or Any other advanced machine learning models for future research.

Q & A

If you have any queries, please feel free to ask

Reference

Annamoradnejad, R., Annamoradnejad, I., Safarrad, T., & Habibi, J. (2019). Using Web Mining in the Analysis of Housing Prices: A Case study of Tehran. In 2019 5th International Conference on Web Research (ICWR). https://doi.org/10.1109/ICWR.2019.8765250

De Nadai, M., & Lepri, B. (2018). The economic value of neighborhoods: Predicting real estate prices from the urban environment. Proceedings of IEEE DSAA, 2018. https://doi.org/10.48550/arXiv.1808.02547

Dyvik, E. (2023). Cities with highest rents worldwide 2023. Statista. Retrieved from https://www.statista.com/statistics/275372/local-rent-cities/

Hromada E. (2015). Mapping of real estate prices using data mining techniques. Czech Technical University in Prague, Faculty of Civil Engineering, Thakurova 7, 166 29 Prague, Czech Republic.

Jiang, X., Jia, Z., Li, L., & Zhao, T. (2022). Understanding Housing Prices Using Geographic Big Data: A Case Study in Shenzhen. Sustainability, 14(9), 5307.

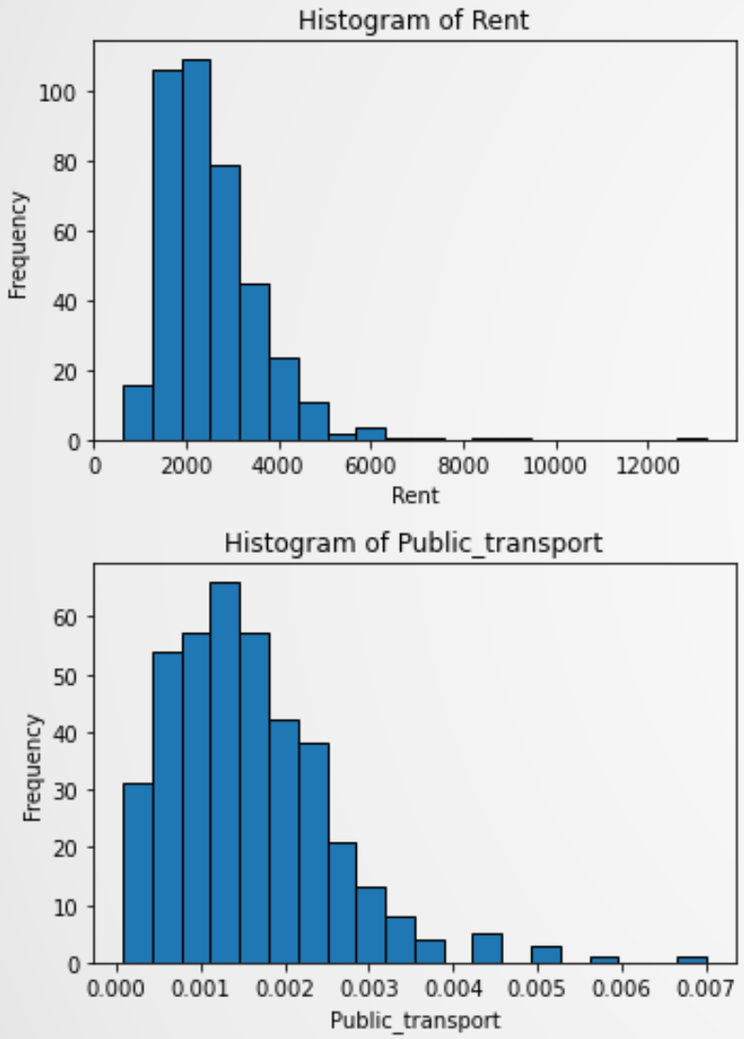
Kuvalekar, A., Manchewar, S., Mahadik, S., & Jawale, S. (2020). House Price Forecasting Using Machine Learning. Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST) 2020. https://doi.org/10.2139/ssrn.3565512

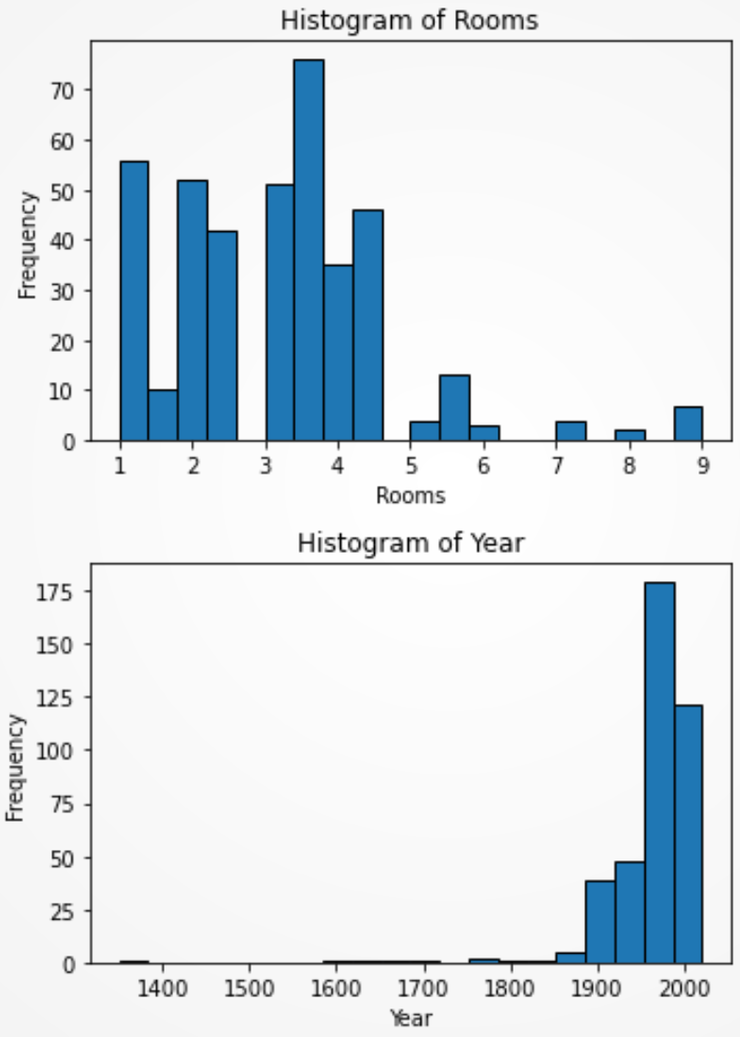
Nowak, A., & Smith, P. (2016). Textual Analysis in Real Estate. College of Business and Economics, West Virginia University.

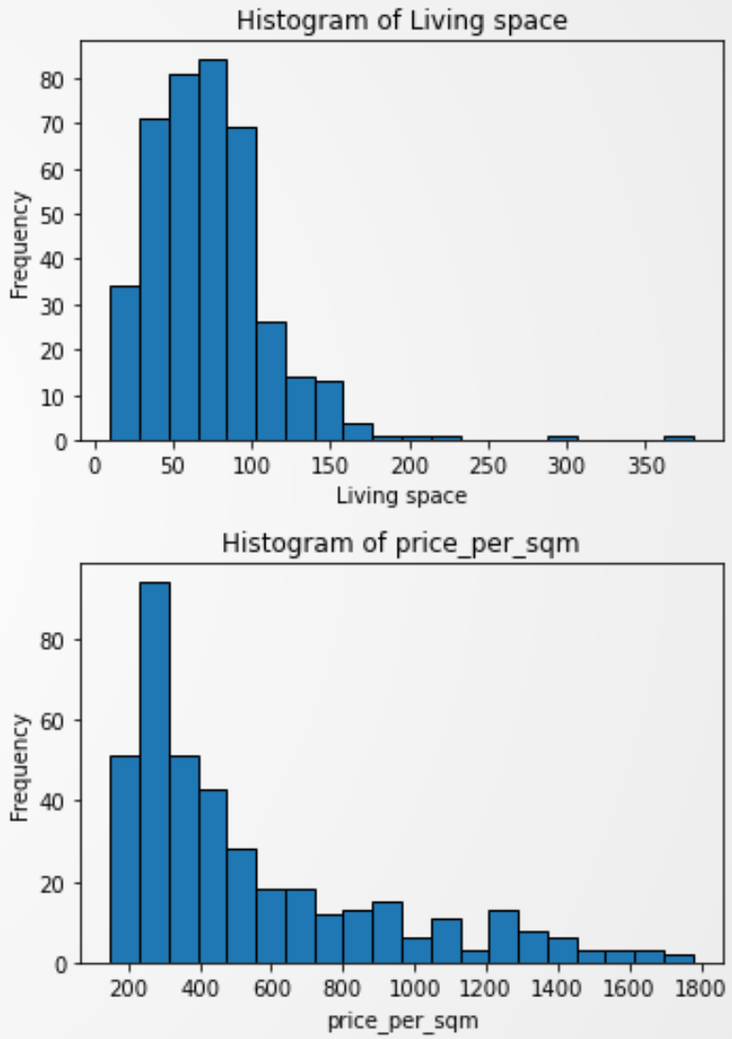
Zhang, H., Li, Y., & Branco, P. (2023). Describe the house and I will tell you the price: House price prediction with textual description data. School of Electrical Engineering and Computer Science, University of Ottawa.

Appendix

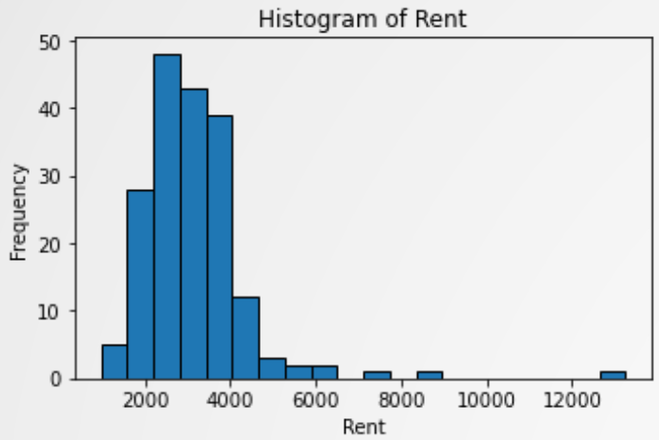
Histrograms for continuous variables - Canton of Zurich

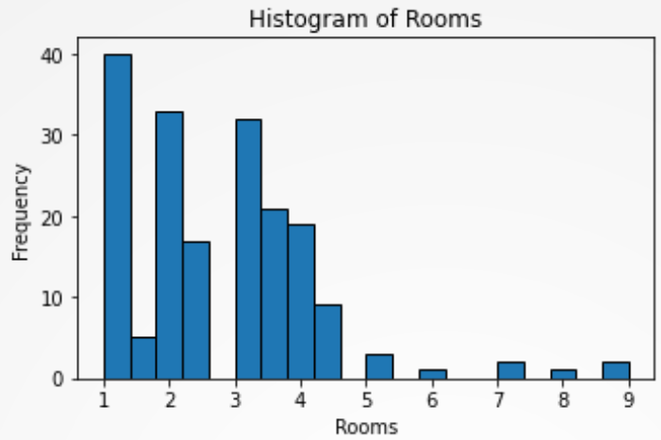


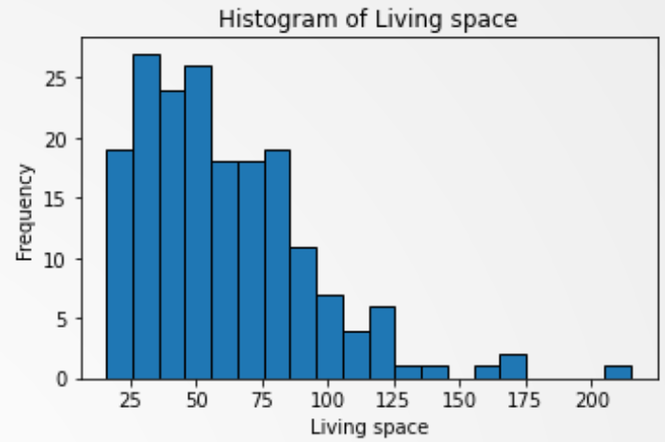


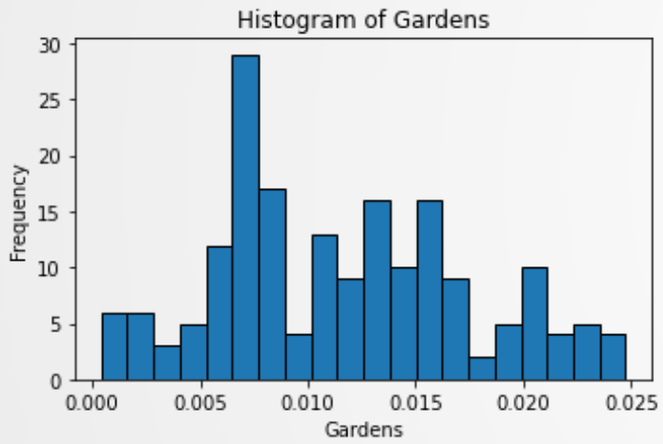


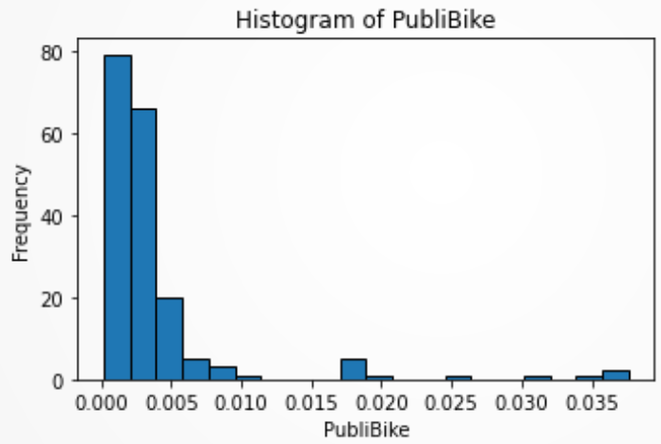
Histrograms for continuous variables - City of Zurich

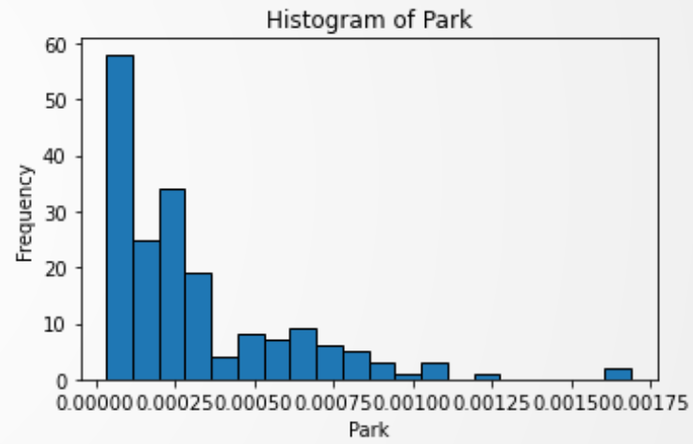


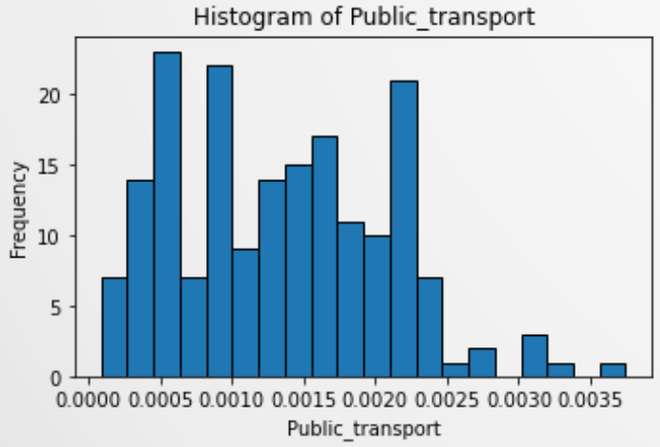


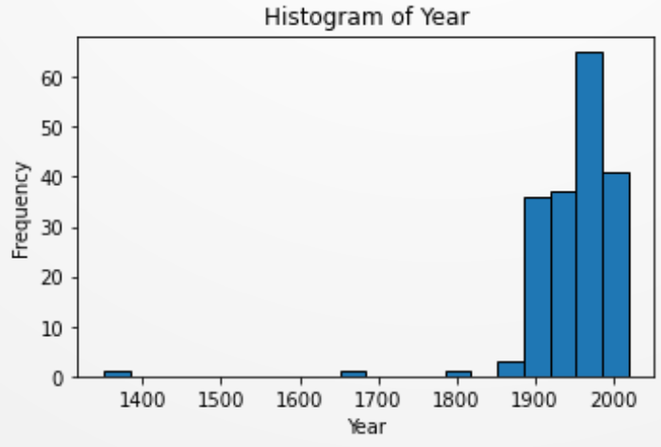


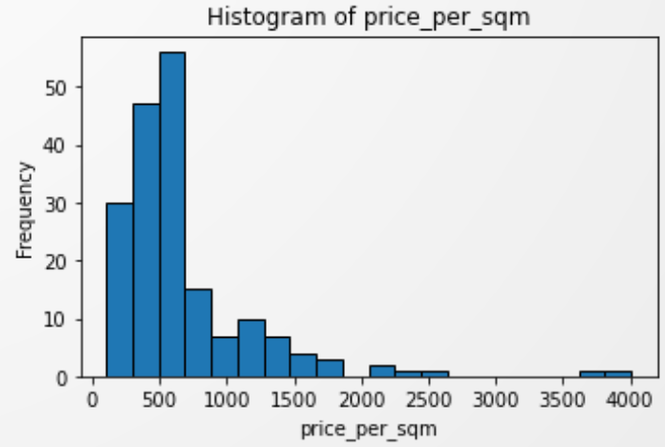








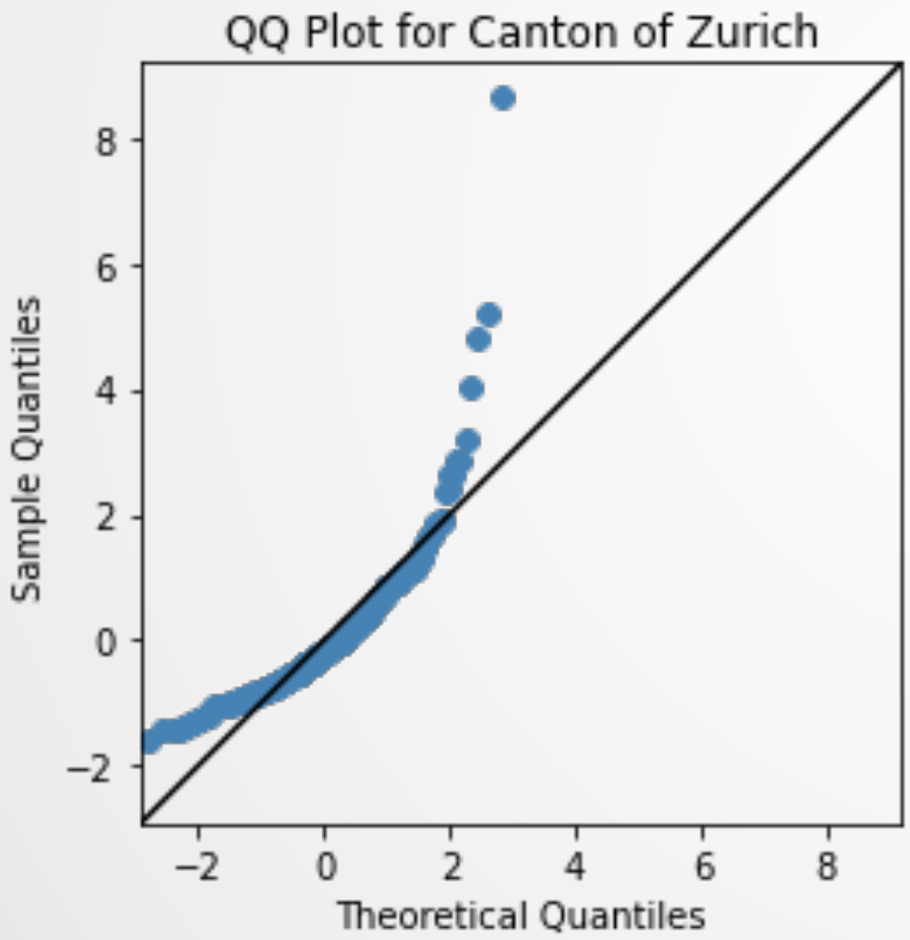




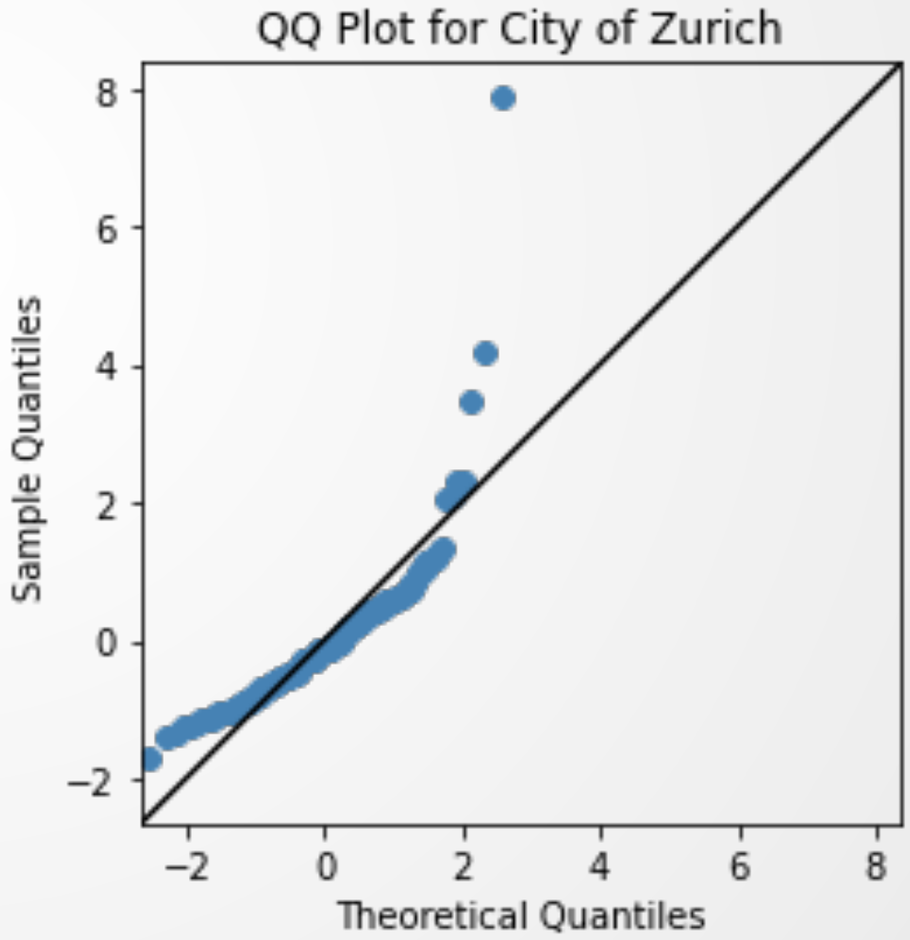
The Reasonableness of rental price data distribution

A QQ-plot is a scatterplot created by plotting two sets of quantiles against one another, to assess a normal distribution.

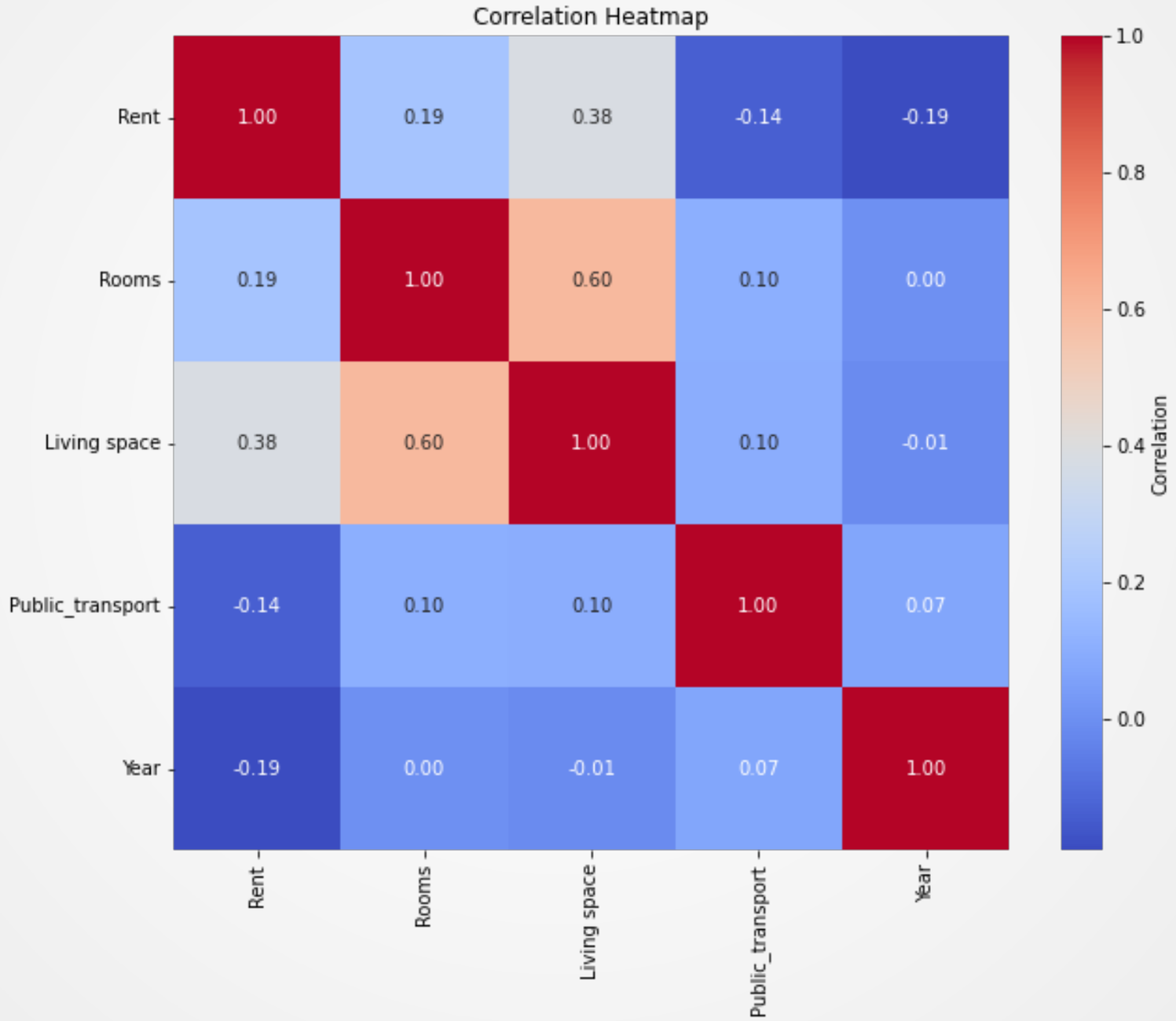
Canton of Zurich



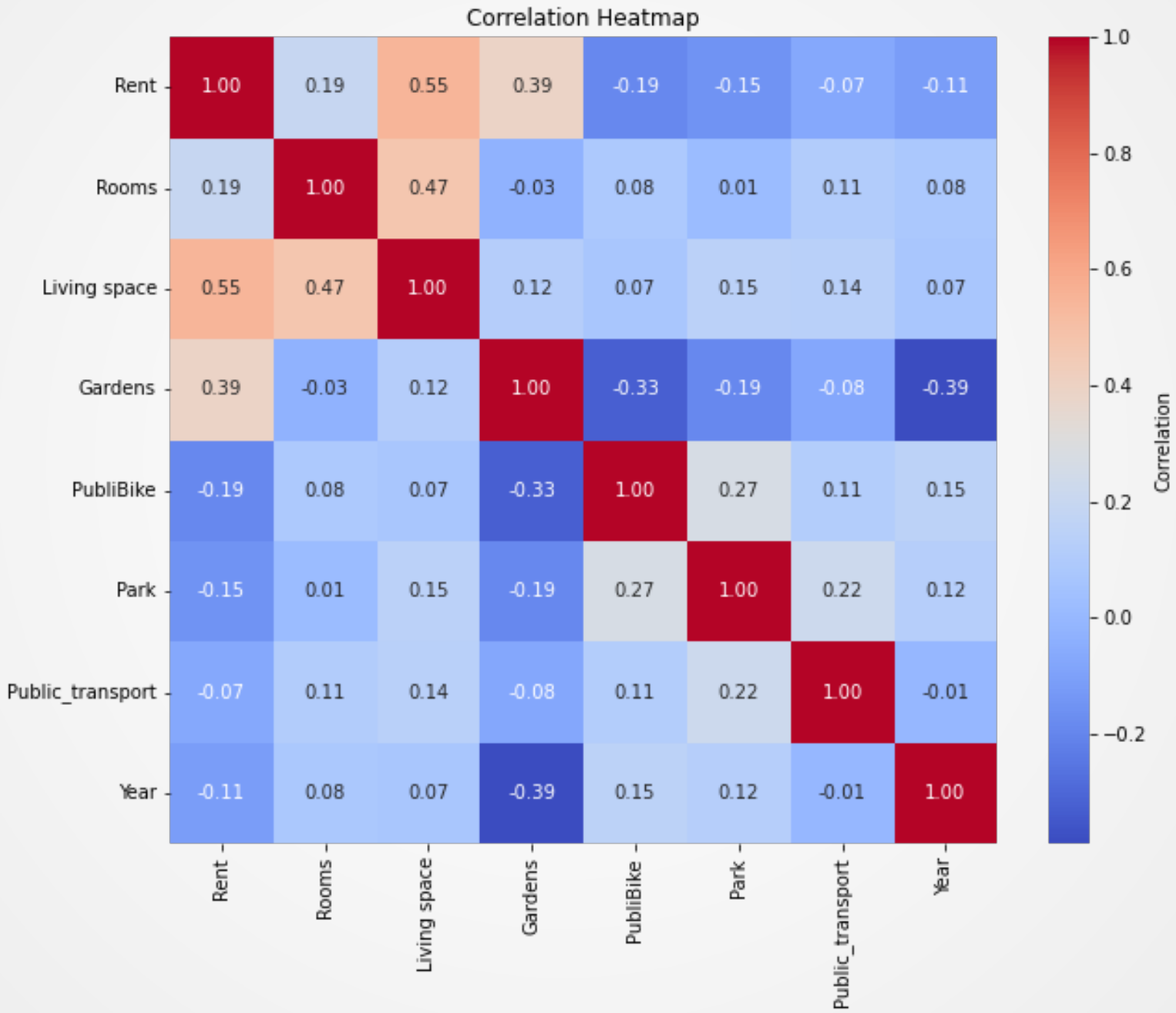
City of Zurich



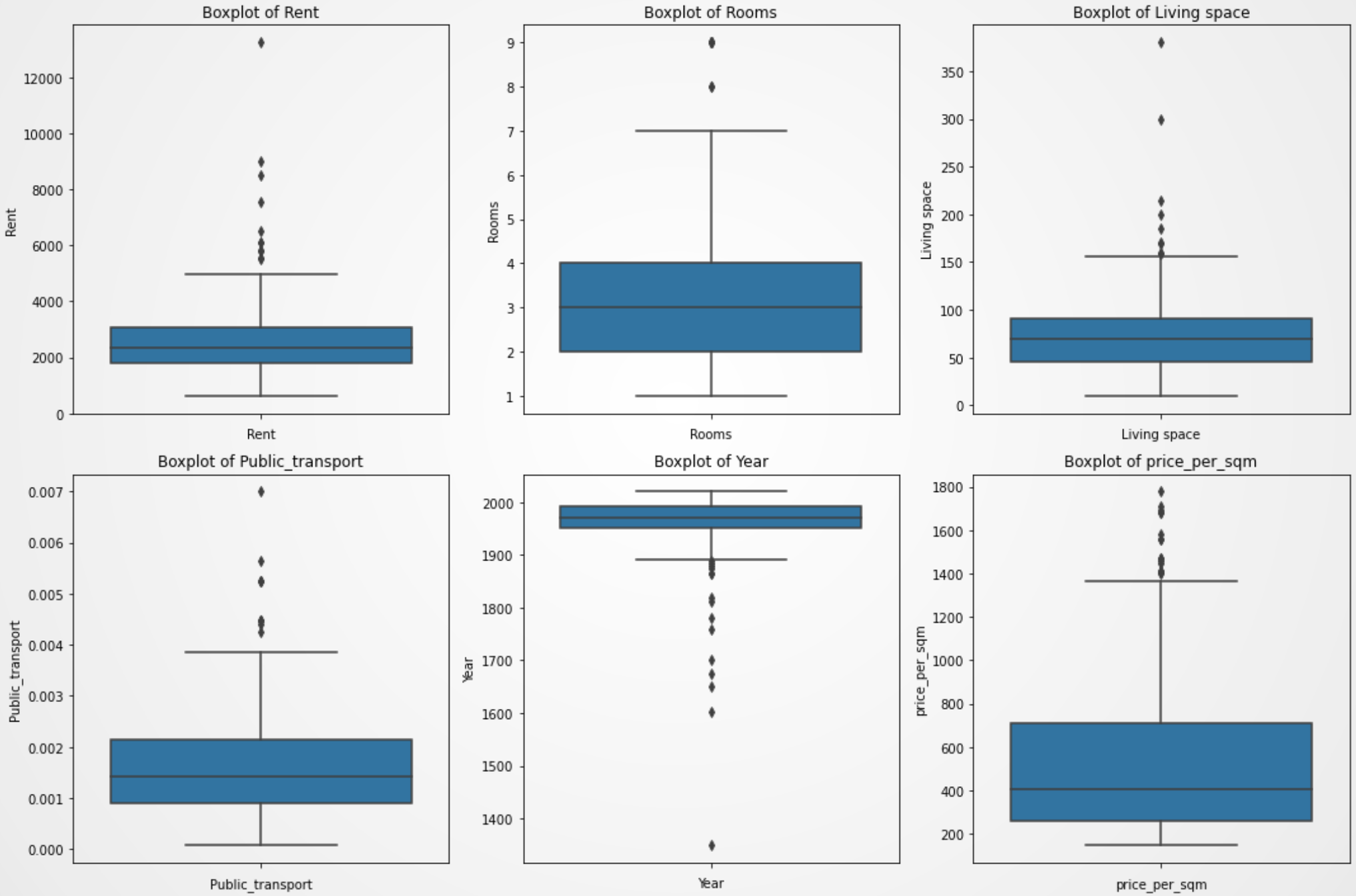
Correlation Heatmap - Canton of Zurich



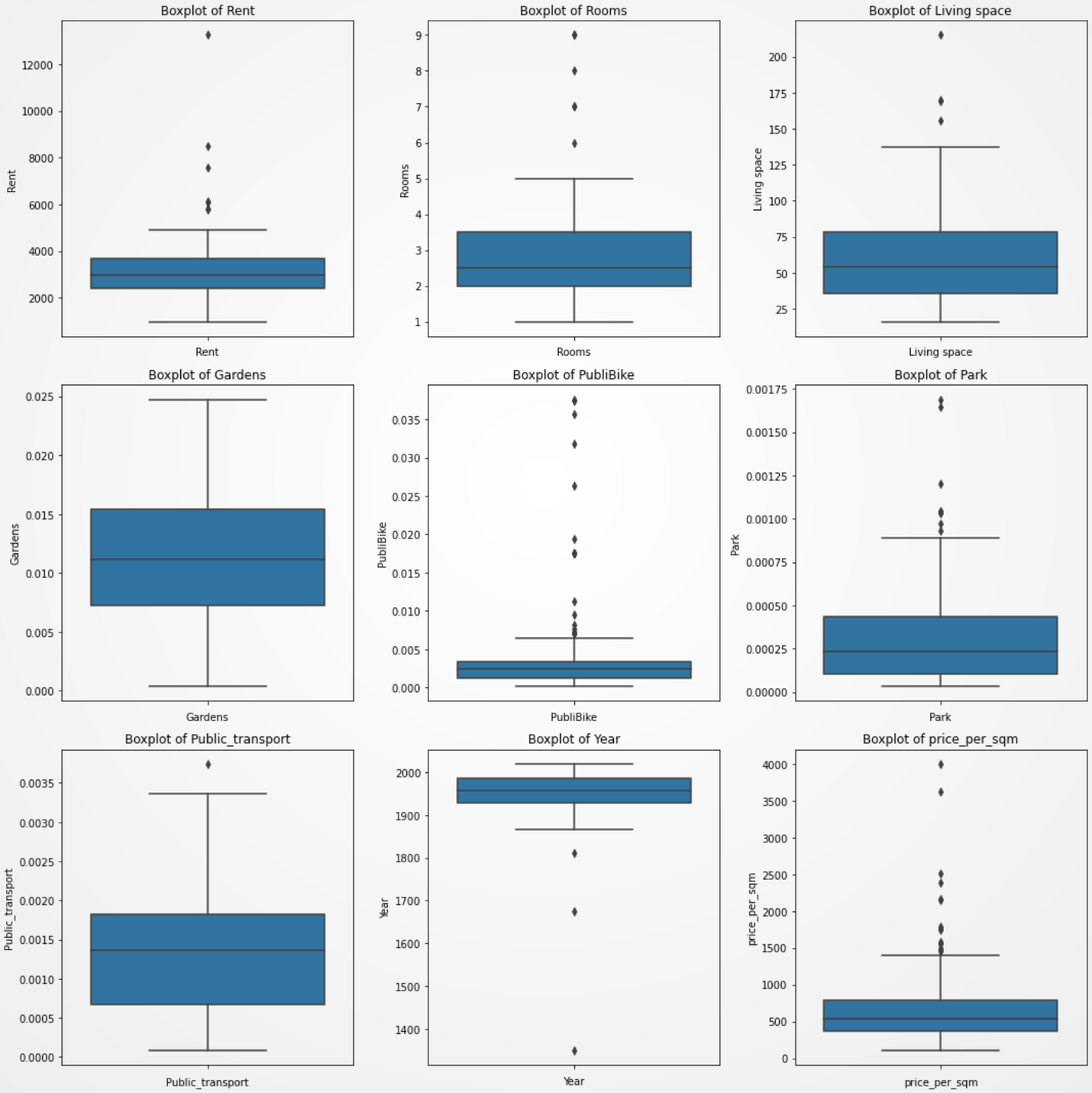
Correlation Heatmap - City of Zurich



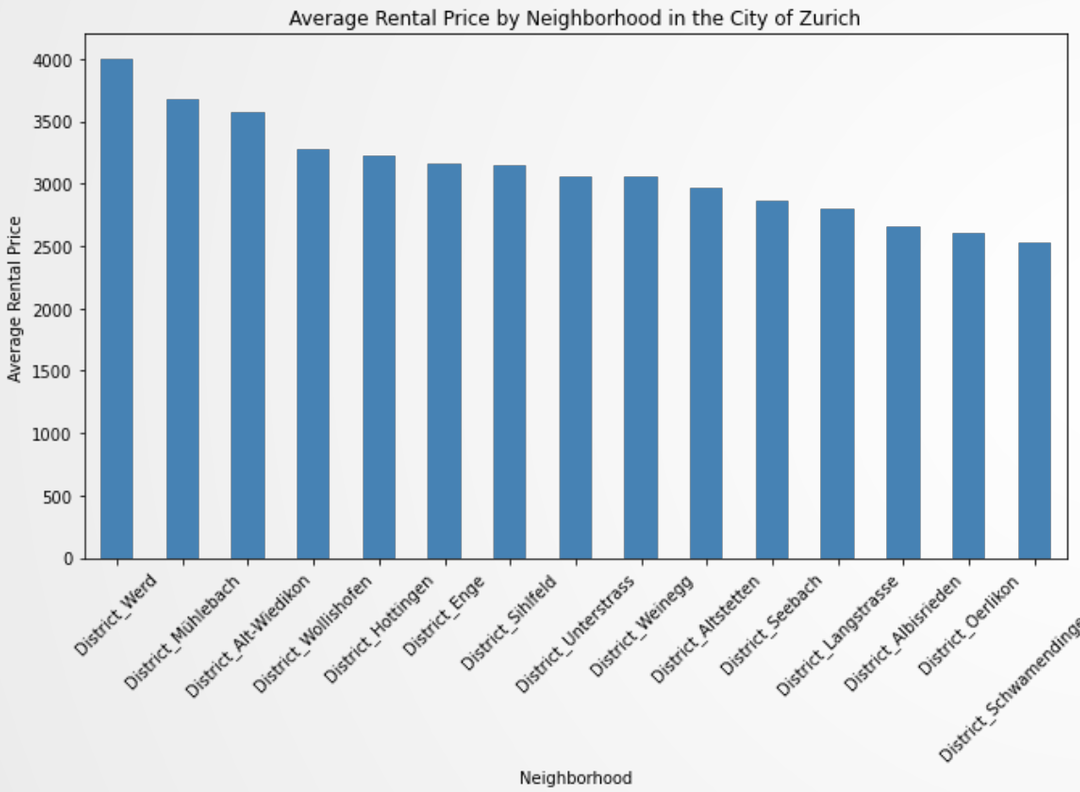
Boxplot for continuous variables - Canton of Zurich



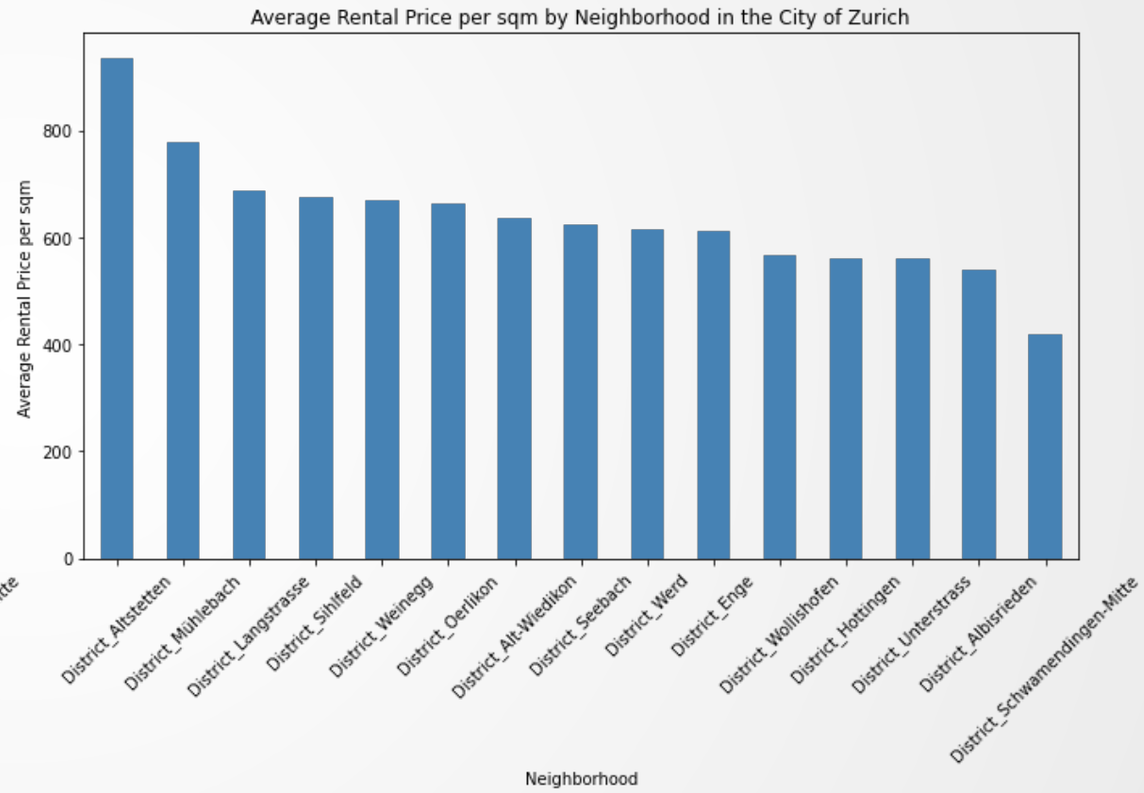
Boxplot for continuous variables - City of Zurich



Average Rental Prices by Neighborhood



Average Rental Prices per sqm by Neighborhood



Histrograms for continuous variables - Canton of Zurich