22.01.2024

Data-Oriented Programming Paradigms 2023W: Exercise 2

Topic 6:
Airbnb & European cities. Berlin.

GROUP 12

STUDENTS

Viktoriia Ovsianik (12217985)

Jandl Christoph (11808249)

Ohanian Artur (12239164)

Obukhov Evgeny (12302378)

Table of Contents

I Research Background & Motivation

Motivation:

With the increasing influx of tourism and rising demands from visitors, European cities are experiencing high pressure. Airbnb positions itself as a P2P platform, aiming to enable local residents to benefit from tourism simultaneously reducing pressure on city centers by redirecting tourists towards residential areas. Our focus is **to evaluate the actual impact of Airbnb on the city of Berlin** in comparison to the traditional hotel industry.

Research questions:

- What socio-economic impact do Airbnb apartments have on cities in Europe?
- Is the position or the popularity of hotels/Airbnb apartments related to Points of Interest, public transport or other features of the city?
- How does the location of AirBnB's differ from Hotels, and what are the consequences?
- How well can good locations for a new Airbnb be predicted?

I Data

AirBnb data

Source: Insideairbnb.com

- Detailed information about all listings available at the platform on **18 December 2023** in Berlin - csv file
- Geographical shapes of Berlin neighbourhoods - geojson file

Points of interest data

Source: Tourist guide for Berlin

- Coordinates of main POIs were extracted from the text of the article

Additional data:

- Hotel locations (Source: tourpedia.com)
- Restaurants & bars locations (Source: tourpedia.com)
- Attractions locations (Source: tourpedia.com)
- Public transport stops (Source: daten.berlin.de)
- Health and social structure (Source: daten.berlin.de)

I Data cleaning & Feature engineering

AirBnb data

Dataset shape (13.327; 75) -> (9370; 26)

- **Deleted** unnecessary features (# of beds, detailed availability info, etc.)
- Converted variables to appropriate **data types** (string -> float/boolean)
- Created **new variable** "isHotel" based on "property_type" using regex
- Cleaned the data (excl. price outliers & postings with missing prices)

Hotels data

- **Deleted** all instances that were not of type "hotel/hostel/guest house"

More about additional data preprocessing steps in "Data Modelling" section ...

I Exploratory data analysis

7.5% listings of the whole dataset are hotel-related properties.

25% listings of non-hotels are properties of shared type.

Locations:

- Hotels mostly located mid-west, Airbnbs - mid-east
- Airbnbs are penetrating more into residential areas compared to hotels
- % of listings of shared type increases towards outskirts (**15 %** in Mitte vs **46%** in Reinickendorf)

Ownership structure:

- **50 %** of listings are owned by hosts who have more than 1 posing (those hosts likely using platform for business purposes)

Prices:

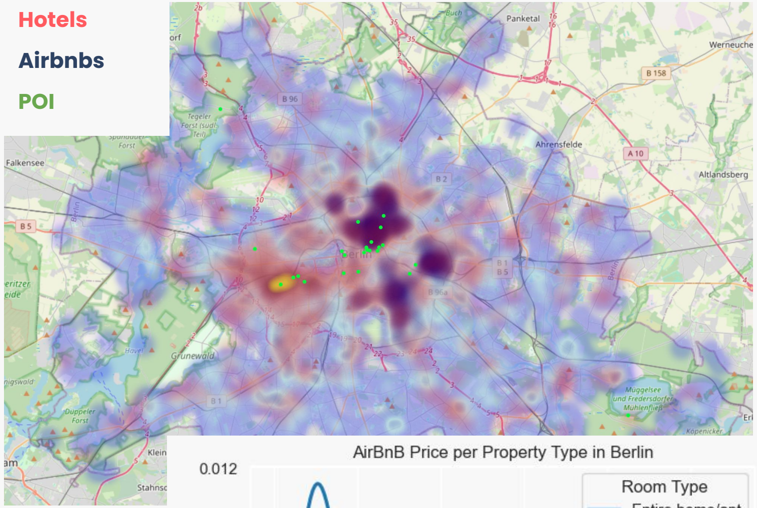
- Average price per night lies in the range **76 € - 142 €** with "Mitte" being the most expensive and "Spandau" the cheapest

Fig 1. Hotels vs Airbnbs locations

Hotels

Airbnbs

POI



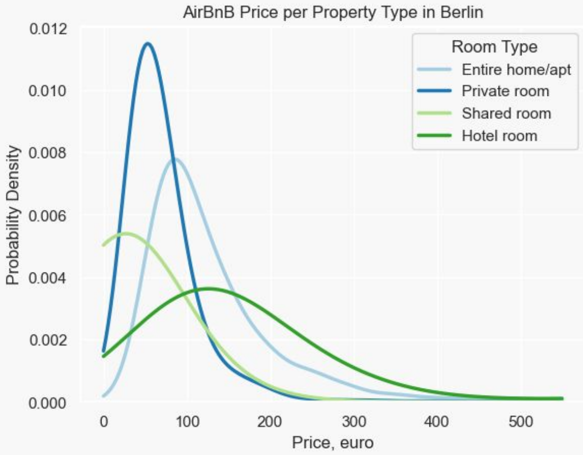


Fig 2. Price distributions by property types

I Spatial data analysis - Autocorrelation

- Spatial autocorrelation understand the degree to which one object is similar to other nearby objects.
- Positive spatial autocorrelation is when similar values cluster together on a map.
- Negative spatial autocorrelation is when dissimilar values cluster together on a map.

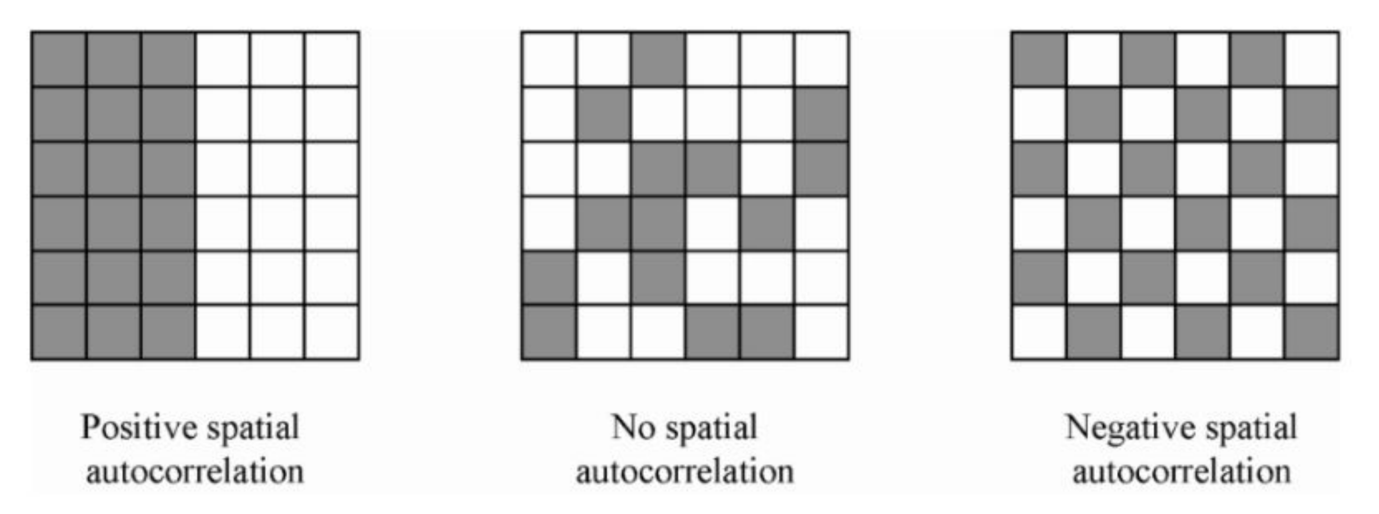
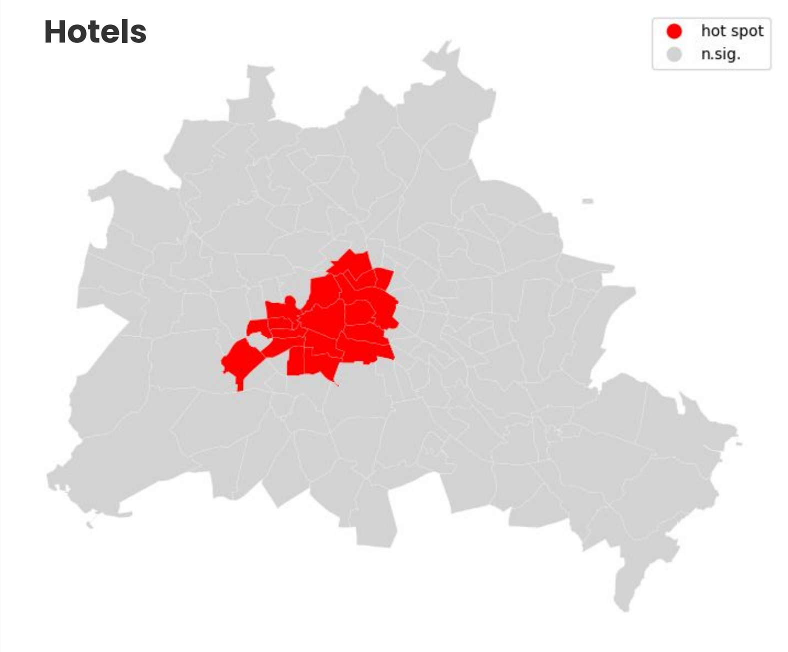


Fig 3. Spatial autocorrelation types

Source: https://rpubs.com/laubert/SACtutorial

Spatial data analysis - hotspots (α=0.05)

Statistically confirmed that hotels mostly located middle-west, Airbnbs - exactly in the middle.



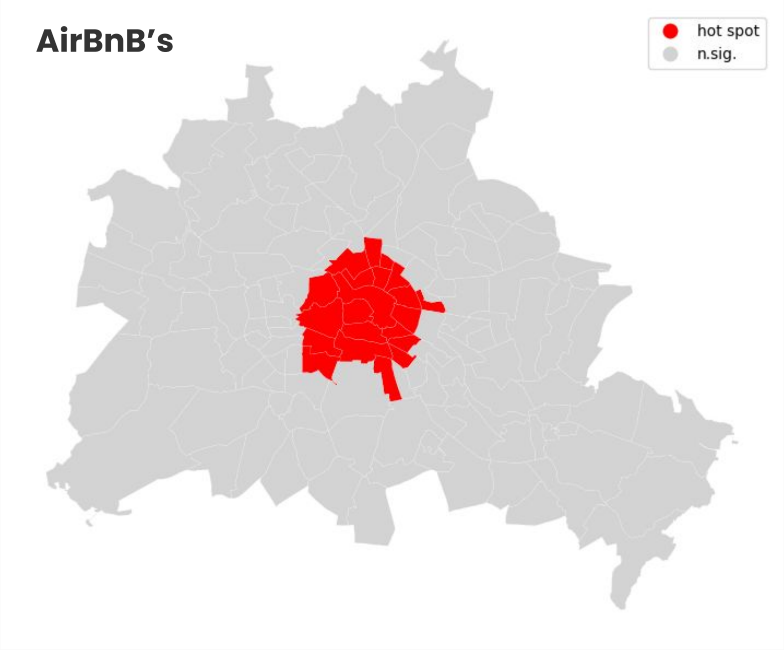
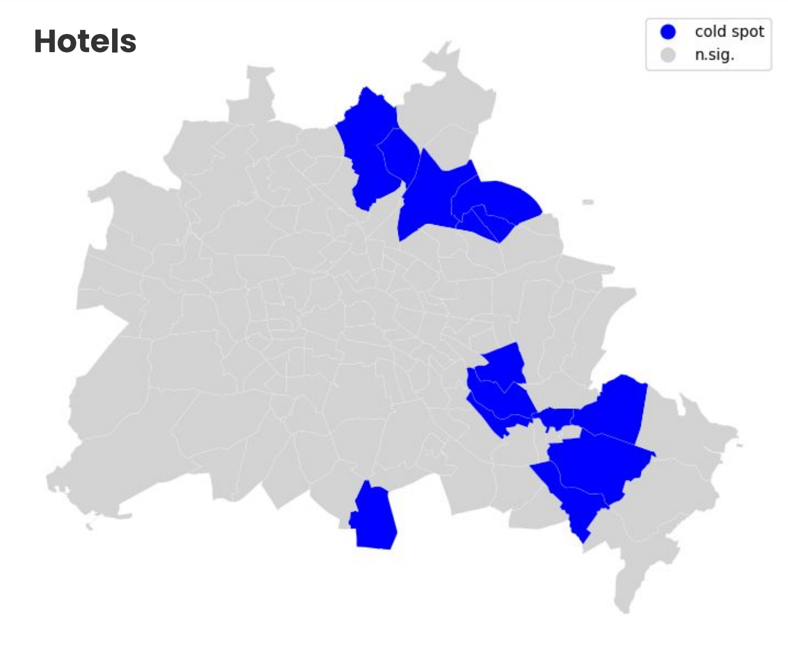


Fig 4. Spatial autocorrelation for hotels & Airbnbs - hot clusters

Spatial data analysis - cold spots

Statistically confirmed clusters with small amount of hotels/airbnb are on the periphery.



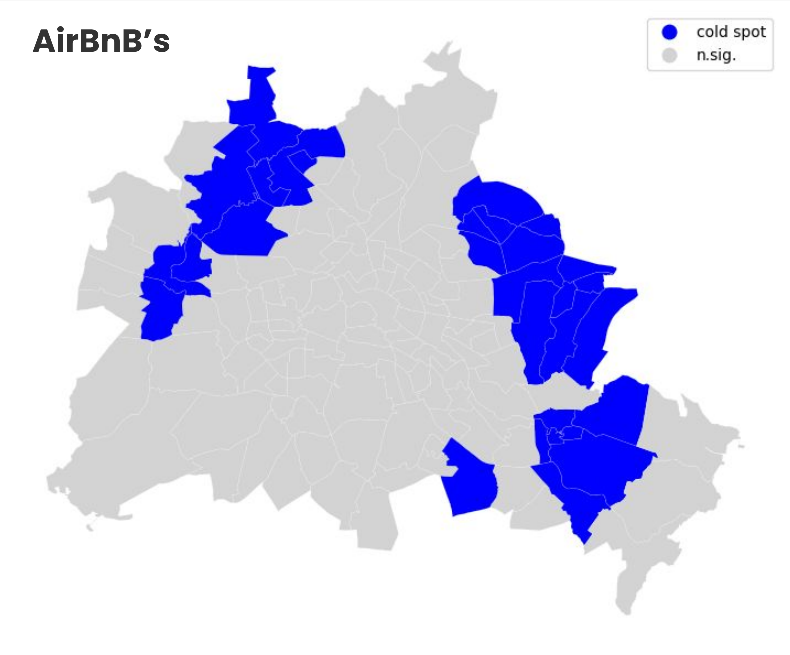


Fig 5. Spatial autocorrelation for hotels & Airbnbs - cold clusters

I Spatial data analysis - interpretations

Distributions of **AirBnBs** fundamentally differs from Hotels, **putting additional pressure on** already highly priced **residential neighbourhoods** that are close to the center.

- **AirBnBs** tend to cluster less around Points of Interest (POIs) and instead concentrate more in central locations
- **AirBnBs** show more cold spots on the outskirts of the city, confirming the central location bias
- **Hotels** are located closer to main sights but do not spread as evenly as AirBnBs in the center

Question 1:

What socio-economic impact do hotels or Airbnb apartments have on cities in Europe? Berlin

Question 1 - Data preprocessing

Additional **preprocessing** steps:

1. **Standardized district names** to align with all 138 Berlin district regions ("Bezirksregionen")
2. **Imputed missing values** using median for complete data in regression analysis
3. **Integrated** Health and Social Data with Airbnb and hotel **datasets** for comprehensive socio-economic impact assessment

I Question 1 - Models

Regression Analysis to assess Airbnb's Socio-Economic Impact:

1. **Regression Setup:** Implemented Ordinary Least Squares (**OLS**) regression to evaluate the **significance levels** of **Airbnb density** on socio-economic factors
2. **Key Variables:** Utilized a range of **socio-economic indices** as **predictors**, with **Airbnb density** as an additional key variable, in varied model configurations, with different **socio-economic indices** as **response** variables
3. **Model Robustness:** Ensured **model reliability** by checking **R^2**, **residuals' normality**, and **matrix conditioning**. Conducted **Shapiro-Wilk test** for residuals and checked for multicollinearity through **correlation matrices**.

I Question 1 - Models

Graphs for different models

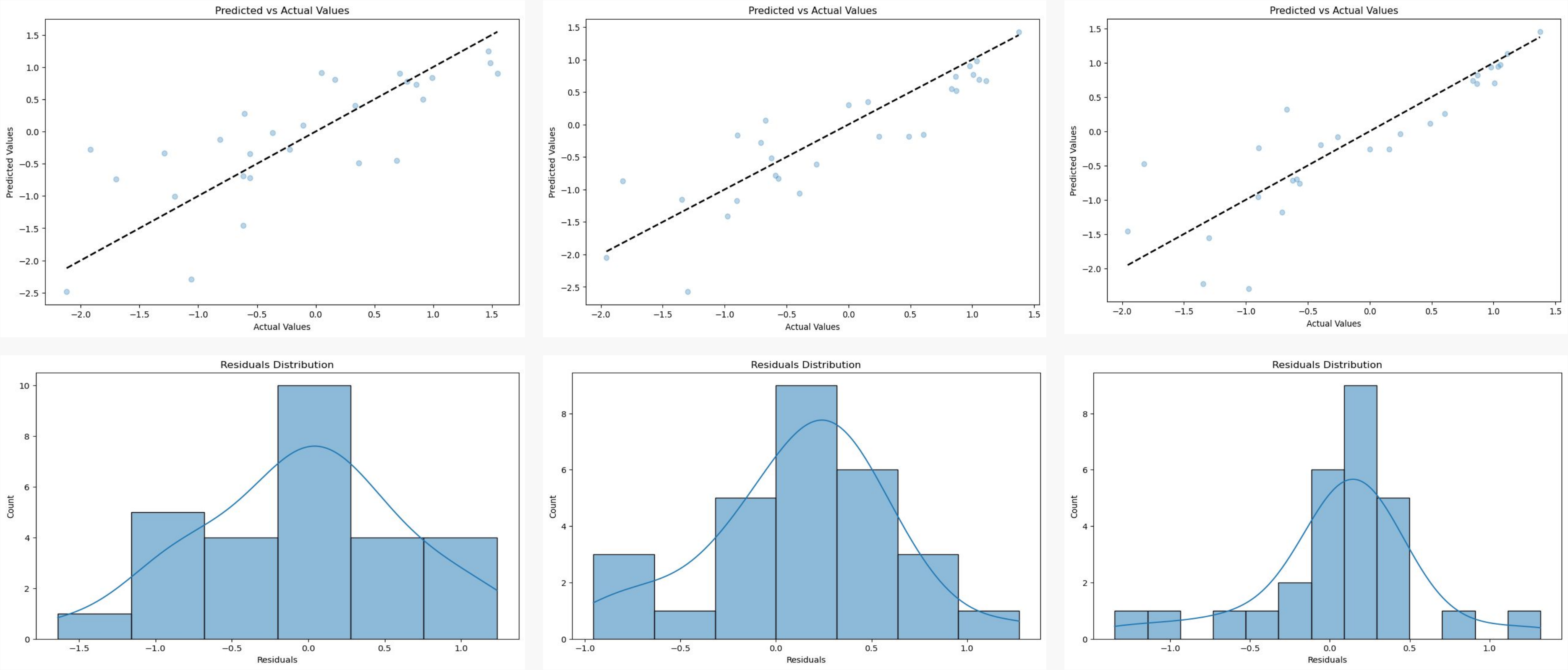


Fig 6. Graphs for different models

Question 1 - Results

1. **No Significant Impact** on **Health and Social Factors:**
   Airbnb presence does not significantly affect public health, mortality, healthcare needs, social environment, education, and social support

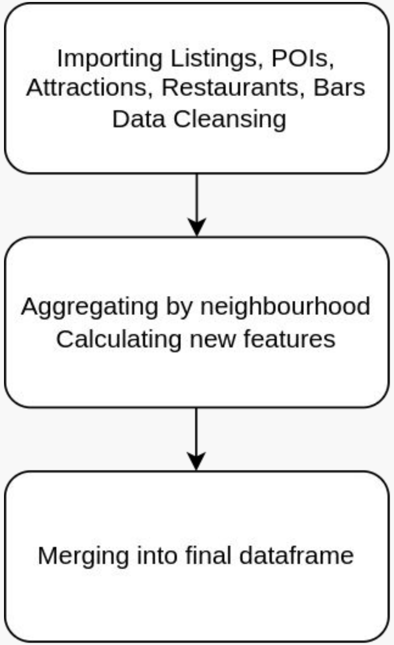
2. **Positive Influence** on **Employment:**
   Airbnb shows a significant positive impact on job market and employment conditions

Question 2:

Is the position or the popularity of Airbnb apartments related to Points of Interest, public transport or other features of the city?

Question 2 - Data preprocessing

1. Loading neighbourhood shapes to **GeoPandas** dataframe
2. **Aggregating AirBnB** listings by neighbourhood and calculating:
   a. Mean price
   b. Mean number of reviews
3. **Aggregating Tourpedia** data by and calculating for bars, restaurants, additional POIs:
   a. Mean likes
   b. Mean number of reviews
   c. Quantity per neighbourhood
   d. Mean users
   e. Mean check ins
   f. Mean tips count
4. Counting number of **public transport** stops in each neighbourhood
5. **Train/test split** (target variable - number of apartment reviews)



Question 2 - Models

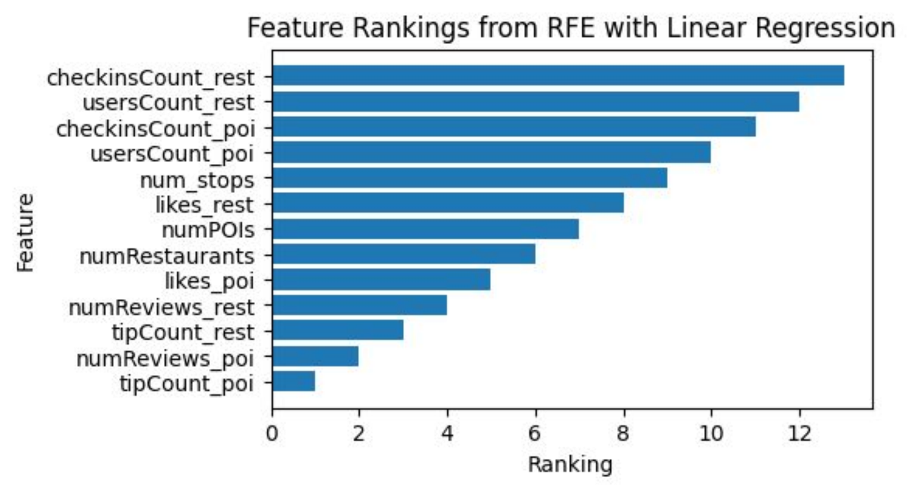


Fig 7. Linear regression with recursive feature elimination

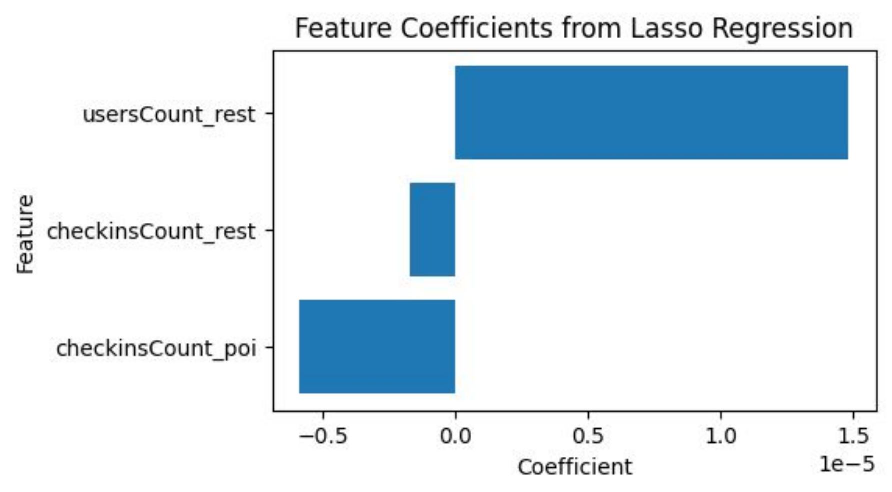


Fig 8. Linear regression with Lasso regularization

Question 2 - Results

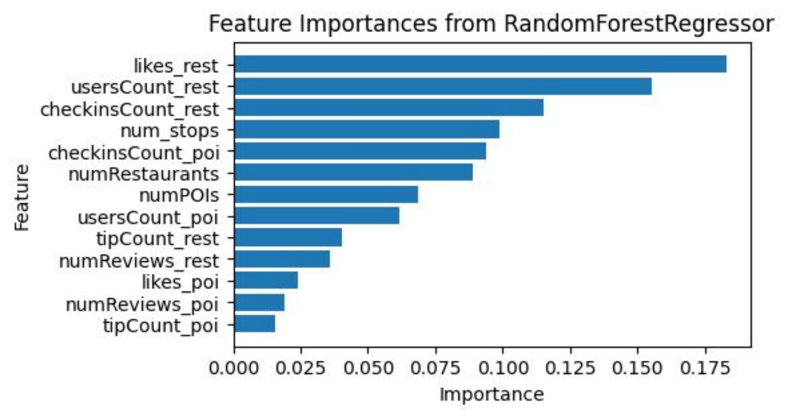


Fig 9. Feature importance from Random Forest Regressor

- **AirBnB popularity is closely linked to** data from **restaurants and points of interest**, notably data provided by Foursquare.
- The presence of transport stops shows a weak correlation with the popularity of AirBnB apartments.
- Restaurants have a more significant impact on popularity compared to points of interest, as indicated by Lasso Regression analysis.

Question 3:

How well can good locations for a new Airbnb be predicted?

Question 3 - Results

- "Good location" was interpreted as the product of the month number of reviews and the price.
- The answer is: **Quite well,** especially when using linear methods.

Model	*MAE*	*RMSE*	*R^2*	*Adj. R^2*
Linear Regression	26.57	21.12	0.92	0.87
Lasso Regression	31.78	28.14	0.86	0.78
Decision Tree Regressor	14.52	42.7	0.67	0.48
Random Forest Regressor	20.19	37.86	0.74	0.59
Support Vector Regressor	52.03	71.65	0.08	-0.46
KNeighborsRegressor	38.44	57.38	0.41	0.07

I Conclusions

Impact on the city:

- Airbnb apartments in Berlin are primarily concentrated in central districts, **adding to the overall tourism pressure** there.
- Airbnbs also exhibit a greater presence in peripheral districts compared to traditional hotels. This includes new districts, contributing to the tourism economy and **positively influencing job market.**.
- High portion of Airbnb listings comprises commercial offers, leading to a **reduction in housing supply for locals** and a decrease in taxes paid by these businesses.

Optimal locations:

- Optimal locations for Airbnbs are associated with Points of Interest (POIs) and restaurants, while they do not demonstrate a strong connection to the public transportation network.