Project Overview: Predicting Airbnb Listing Prices in Berlin

About Airbnb: Airbnb is an online platform that connects people renting out their homes with those seeking accommodations. Launched in 2008, it has grown into a global marketplace offering unique stays, from single rooms to entire homes, in over 220 countries.

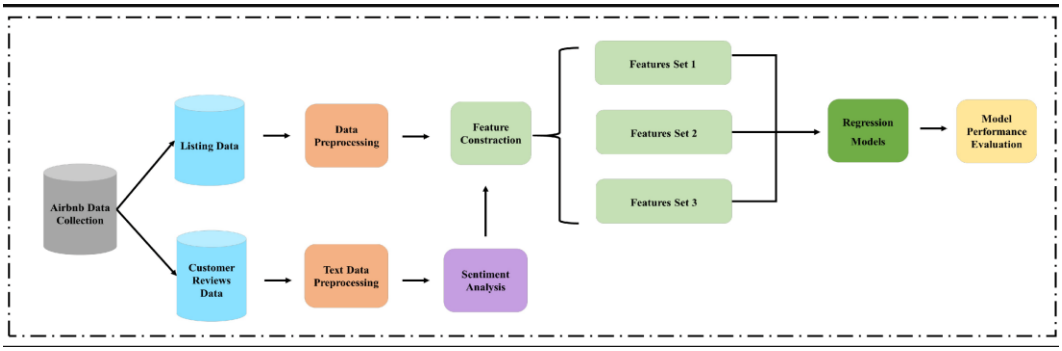
Dataset for the project: Berlin Airbnb Ratings

Objective: The project aims to evaluate and predict listing prices for Airbnb properties in Berlin, considering various factors such as host ratings, property features, and GIS data.

WHAT WE KNOW:

- The dataset includes information on Airbnb listings in Berlin, featuring host ratings, property types, location details, and review metrics.
- Previous studies on Airbnb price prediction have used machine learning models like Lasso, Ridge, and SVR regression to forecast listing prices. Key factors influencing prices include property characteristics (e.g., number of bedrooms) and customer reviews. Sentiment analysis of reviews and clustering techniques have also been employed to enhance accuracy.
  Relevant sources: (Sustainable Price Prediction Model, Airbnb price prediction using machine learning and sentiment analysis)
- The project consists of six key phases: data collection, data preprocessing, sentiment analysis of reviews, feature construction, regression modeling, and model performance evaluation.

source: A Sustainable Price Prediction Model for Airbnb Listings Using Machine Learning and Sentiment Analysis



BUSINESS GOALS:

- Platform Improvement: Airbnb can use listing price predictions to suggest optimal pricing to hosts, thereby improving overall market efficiency.
- Pricing Optimization: Hosts can utilize the model to set competitive prices, maximizing occupancy rates and revenue.

ANALYTICS TOOLS:

- Build a regression model that accurately predicts listing prices based on the prepared dataset.
- Success Metrics: Achieve high model performance in terms of accuracy and robustness, measured by metrics such as RMSE (Root Mean Squared Error) and RMSLE (Root Mean

Squared Log Error). The model should achieve at least 85% accuracy using MAPE (Mean Absolute Percentage Error).

INFLUENCING FACTORS:

- Data Features: Property attributes, host ratings, review counts, and geographical information are critical factors influencing price prediction.
- Data Quality: The accuracy of predictions depends on the quality and completeness of the data, including proper handling of missing values and outliers.

NEW APPROACHES:

- Feature Engineering: Introduce new features like distance to a central point, hierarchical property type categories, and sentiment conversion of textual data. Additionally, NLP models can be used to extract further characteristics about the properties.
- Advanced Modeling: Enhance model performance through techniques like hyperparameter tuning and ensemble methods.

PROJECT STAGES:

DATA PREPARATION.

Objective: Prepare and clean Airbnb Berlin dataset for analysis or modeling.

1. Data Cleaning and Transformation:
    - Date Conversion: Convert to datetime. (review_date, Host Since, First Review, Last Review)
    - Convert columns to binary (1/0), (Is Superhost, Is Exact Location).
    - Postal Code Cleaning: Convert Postal Code to integer, removing non-numeric characters.
    - Host Response Rate: Convert Host Response Rate from percentage to float.
    - Column Drop: Remove unnecessary columns (e.g., Country, Business Travel Ready)
    - Top 10 Neighbourhoods: Create a new column Top10Neighbourhood for the top 10 most frequent neighborhoods.
    - Handling Negative Values: Replace negative values with NaN in columns like Accomodates, Bathrooms, etc.
2. Data Preparation for Analysis:
    - Extract relevant columns into flat table removing duplicates rows. for analysis.
    - Extract review-related columns into separate table.
3. Data Filtering:
    - Filter the panel to include only for **apartments** in berlin (the population for the research).

DATA ANALYSIS AND VISUALIZATION

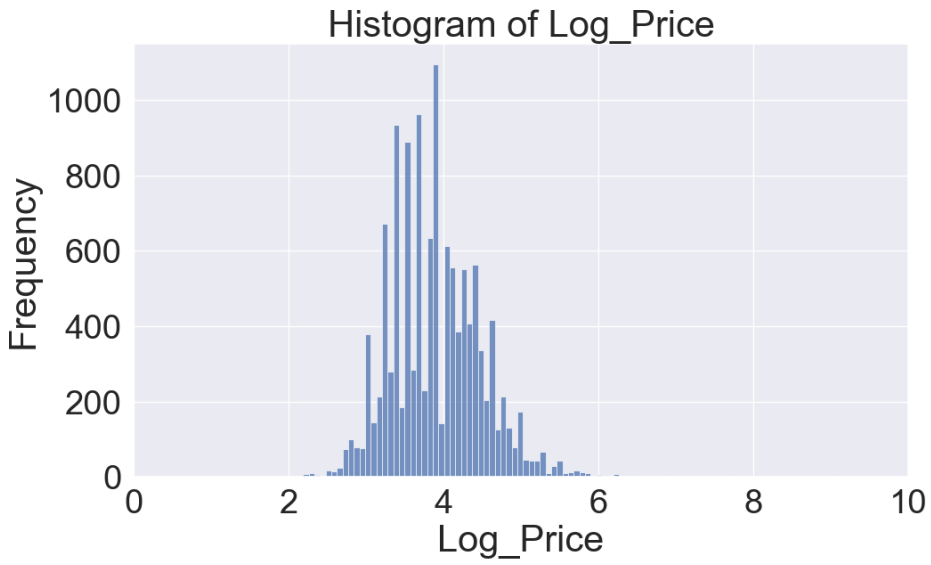
Data Analysis:

- Export data information (data types, min/max, missing values) to CSV.
- Visualize data using AutoViz, count plots, histograms, correlation heatmaps, scatter plots, and bar plots.

Key points from the analysis:

i. The `room_type` column includes `Entire home/apt`, which shows a significant difference in average price. This raises the question of whether it might be better to predict prices for rooms and apartments separately.

ii. The target variable, listing price, exhibits strong left skewness, so we should consider applying a log transformation when evaluating the best model.



DATA CLEANSING AND OUTLIERS

Objective: Clean data, handle missing values, and treat outliers.

1. Outlier Detection and Treatment:
   - Visualized and capped outliers using the Interquartile Range (IQR) method.
   - Rather than removing outliers, I addressed the impact of extreme values in various columns:
     1. I chose to retain the ratings columns, treating them as categorical variables, since all values between 1-10 are valid.
     2. For the columns `Accommodates`, `Bathrooms`, `Bedrooms`, `Beds`, `Guests Included`, and `Min Nights`, I created new binned columns. This method helps reduce the influence of outliers by grouping extreme values together (as detailed in the feature engineering section).

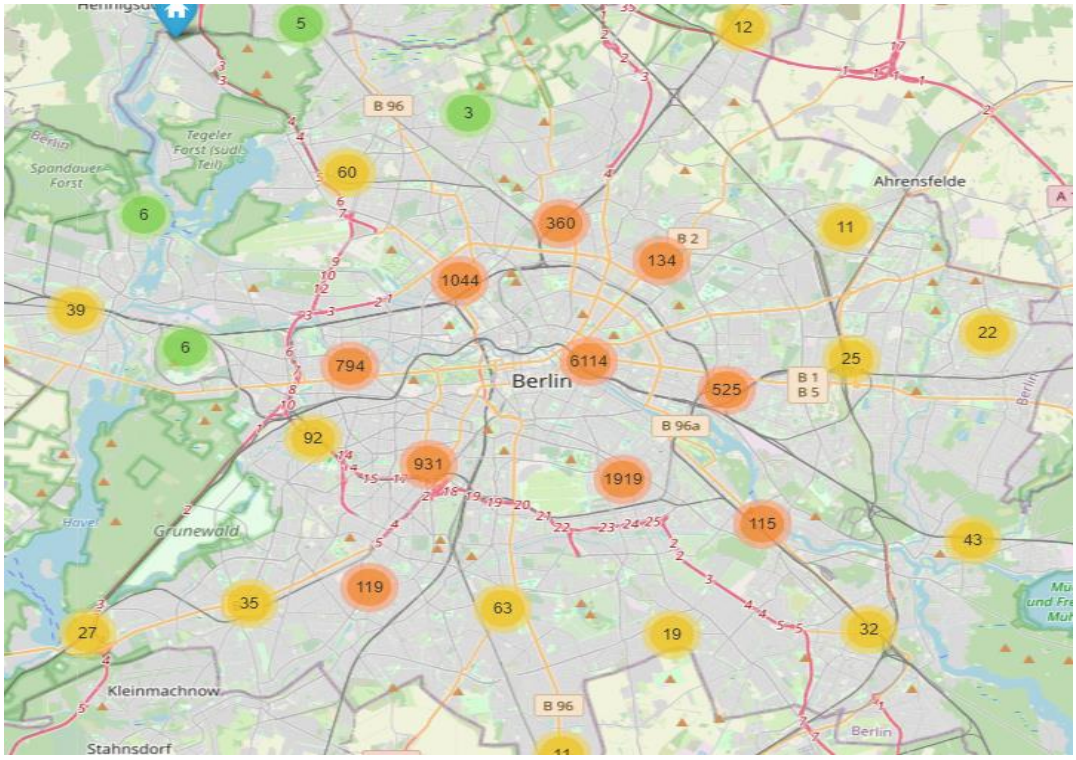
2. Handling Missing Values:

   - For all rating columns, I filled the missing values with a constant and added an indicator column to flag the null values.
   - For the 'is_Superhost' and 'reviews' columns, I filled the missing values with zero.
   - For the numerical columns Beds, Bathrooms, and Bedrooms, I applied KNN imputation to fill the missing values.
   - Dropped unnecessary columns, such as Postal Code, which had 60% missing data.

FEATURE ENGINEERING

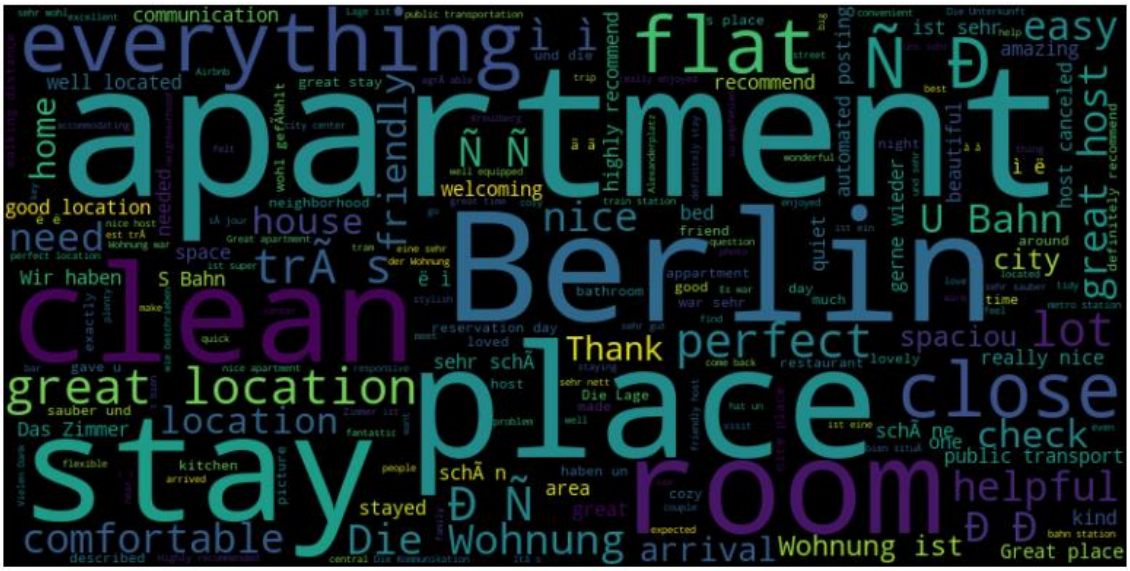
Objective: Engineer additional features for better analysis and modeling.

1. Geographic Features: Calculate distances from properties to a central point and plot on a map. (see below)



A map of Berlin with airbnb properties

2. Neighborhood Features: Compute and rank average location ratings by neighborhood.
3. Date Features: Create date-related features and indicators for missing values.
4. Reviews Processing: Analyze review comments for sentiment and polarity.



FEATURE SELECTION

Objective: Select important features for predicting the target variable Price.

1. Prepare data, encode string columns, and fit various regression models (Lasso, SVM, etc.).

2. Compile results and select features based on model performance. All of my features were selected in 4 out of 6 models, resulting in a total of 58 features before encoding.

LISTING PRICE REGRESSION MODELS SELECTION AND FINE-TUNING

Objective: Train, evaluate, and fine-tune regression models for predicting listing prices.

1. Data Preprocessing:
    - Load and preprocess data, apply one-hot and ordinal encoding, and scale numerical features.
2. Model Training:
    - Train and evaluate models like Linear Regression, Decision Trees, Random Forests, etc.
    - Perform Grid Search for hyperparameter tuning on GBM and Ridge (liner regression).

model performance metrics comparison

Model	MSE	RMSE	MAE	RMSLE	MAPE
Ridge_FT	2764.326	52.577	18.104	0.382	30.013
Linear Regression	2770.162	52.632	18.151	0.383	30.122
GBM_FT	2771.140	52.642	18.716	0.393	32.193
GBM	2780.154	52.727	18.118	0.378	29.891
XGB	3060.060	55.318	18.891	0.395	31.576
ADABoost	3099.431	55.673	23.849	0.489	47.879
Random Forest	3353.932	57.913	18.560	0.384	31.002
KNN	3637.386	60.311	28.611	0.592	53.554
Mode Baseline	3650.161	60.417	27.027	0.563	49.470
Mean Baseline	3668.914	60.572	27.017	0.563	48.512
SVM	3681.491	60.675	26.563	0.557	45.362
Decision Tree	50496.833	224.715	34.127	0.570	67.617

3. Feature Selection:
    - Use stepwise selection to finalize the model based on significant features.

    Key Features:

    - Features: Host details (e.g., Host Since, Host Response Time), property details (e.g., Property Type, Room Type), neighborhood info (neighbourhood, Neighborhood Group), and ratings (Overall Rating, Accuracy Rating).

MAJOR CONCLUSION:

After analyzing the model results using various metrics, it is clear that the accuracy of the best model, Gradient Boosting Machine (GBM), is insufficient for predicting listing prices. The Mean Absolute Percentage Error (MAPE) measures prediction accuracy as a percentage, and for my best model, the MAPE is 30%. This means that, on average, the model's predictions deviate by 30% from the actual values, which is not accurate enough for production use.

STEPS TO IMPROVE THE MODEL:

1. Add Additional Features: Incorporate more features related to listing characteristics, such as size, kitchen amenities, parking availability, Wi-Fi, air conditioning, and laundry facilities.
2. Apply NLP Techniques: Utilize natural language processing (NLP) models to analyze comments and listing names, extracting valuable textual data.
3. Split the Dataset: Divide the dataset into two groups based on property type: apartments and rooms. Develop and evaluate separate models for each group.

APPENDIX:

List of features:

1. Features Related to the Listing

- **Listing Name**: The name of the listing. (String)
- **Listing URL**: The URL of the listing. (String)
- **Property Type**: The type of property. (String)
- **Room Type**: The type of room. (String)
- **Accomodates**: The number of people the property can accommodate. (Integer)
- **Bathrooms**: The number of bathrooms. (Float)
- **Bedrooms**: The number of bedrooms. (Integer)
- **Beds**: The number of beds. (Integer)
- **Square Feet**: The square footage of the property. (Float)
- **Price**: The price of the listing. (Float)
- **Guests Included**: The number of guests included in the price. (Integer)
- **Min Nights**: The minimum number of nights required to stay. (Integer)

- **Reviews**: The number of reviews the listing has. (Integer)
- **Overall Rating**: The listing's overall rating. (Float)
- **Accuracy Rating**: The listing's accuracy rating. (Float)
- **Cleanliness Rating**: The listing's cleanliness rating. (Float)
- **Checkin Rating**: The listing's checkin rating. (Float)
- **Communication Rating**: The listing's communication rating. (Float)
- **Location Rating**: The listing's location rating. (Float)
- **Value Rating**: The listing's value rating. (Float)

2. Features Related to the Host

- **Host Name**: The name of the host. (String)
- **Host URL**: The URL of the host. (String)
- **Host Since**: The date the host joined Airbnb. (Date)
- **Host Response Time**: The host's response time. (String)
- **Host Response Rate**: The host's response rate. (String)
- **Is Superhost**: Whether or not the host is a Superhost. (Boolean)

3. Features Related to the Location

- **neighbourhood**: The neighbourhood of the listing. (String)
- **Neighborhood Group**: The neighbourhood group of the listing. (String)
- **City**: The city of the listing. (String)
- **Postal Code**: The postal code of the listing. (String)
- **Country Code**: The country code of the listing. (String)
- **Latitude**: The latitude of the listing. (Float)
- **Longitude**: The longitude of the listing. (Float)
- **Is Exact Location**: Whether or not the location is exact. (Boolean)

4. Features Related to the Reviews

- **review_date**: The date of the review. (Date)
- **Reviewer Name**: The name of the reviewer. (String)
- **Comments**: The reviewer's comments. (String)
- **First Review**: The date of the first review. (Date)
- **Last Review**: The date of the last review. (Date)