Data Understanding and Exploratory Data Analysis (EDA) on Airbnb Dataset

Arthur Gubaidullin, B22-DS-01

March 9, 2025

1 Data Understanding

1.1 Data Collection

The dataset was initially collected by a Kaggle enthusiast to analyze how Airbnb competes with the residential housing market. The data set was extracted from the Airbnb website and published on Kaggle platform, from which our team downloaded it. No problems were encountered during the data collection phase.

1.2 Data Description

The data were acquired in tabular format as a ".csv" file. It is also possible to acquire the data in the ".geojson" format. The basic "surface" features of the dataset are as follows.

- Quantity of data: 494,954 unique datapoints;
- Number of features: 89, including one target variable;
- Feature data types distribution:
  - Object/Categorical: 55;
  - Float: 33;
  - Int: 1 (ID column).
- The data set has 7,793,336 missing values (17.7% of all values).

Table 1: Top 10 Features with Missing Values

Feature	**Missing Values**	**Percentage (%)**
Has Availability	485,647	98.0
Square Feet	482,745	97.5
License	480,358	97.0
Host Acceptance Rate	452,696	91.5
Monthly Price	398,863	80.6
Weekly Price	397,207	80.3
Neighbourhood Group Cleansed	392,791	79.4
Jurisdiction Names	360,401	72.8
Notes	297,364	60.0
Security Deposit	290,942	58.8

I decided to drop all variables with the amount of missing values $\geq 15\%$. In general, the data quality satisfies the requirements, and it is assumed that the data loss should not be substantial.

1.3 Data Exploration

Our initial target variable was the "Review Scores Rating". However, almost 26% of these values are missing. From my perspective, we can predict the review scores using relevant features. I do not think that using simple imputing techniques would suit us here, since most NaNs mean that no scores were given to a particular place.

From a business perspective, predicting the daily price might be more relevant by helping landlords set the price in an appropriate range. For your information, the percentage of missing Reviews vs. Price is 0.26 / 0.01. So, from that very moment on, I decided to change my business objective as well as the target variable to the daily price ("Price"). By accurately predicting the price, we will make the lives of both customers and landlords easier. This will guarantee the price adequacy of a particular property, ensuring that neither customers nor landlords are fooled by incorrect pricing.

1.4 Types of Data

The provided dataset will be split into three sets of data based on its type:

- **Categorical**: Contains categorical data with the amount of unique values ≤ 50 (e.g., "Property Type", "Bed Type", etc.);
- **Object**: Contains non-numerical data with the amount of unique values > 50 (e.g., "Summary", "Description", etc.);
- **Numerical**: Contains numerical data.

As a result, we have the following datasets:

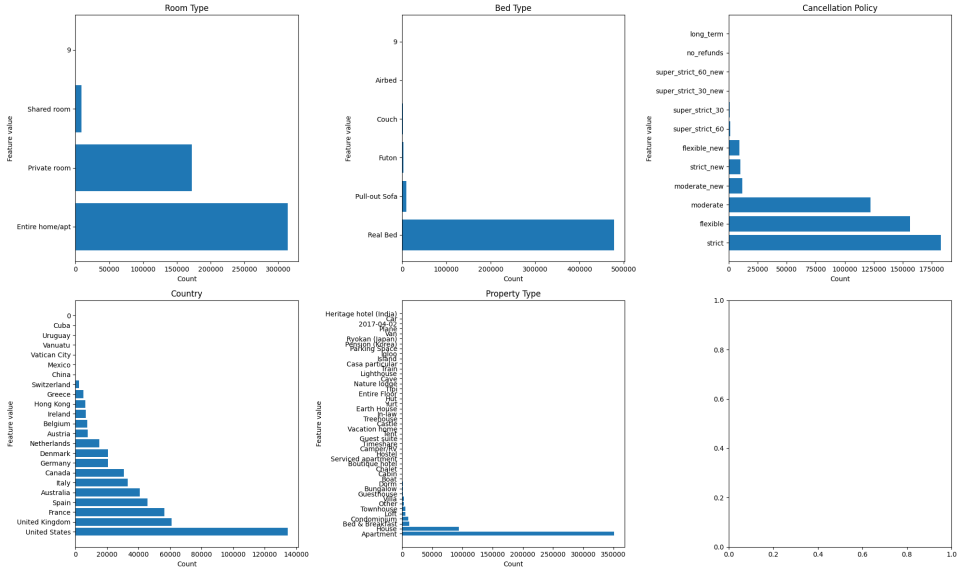
- **Categorical**:
  - Room Type;
  - Experiences Offered;
  - Bed Type;
  - Cancellation Policy;
  - Country;
  - Property Type.
- **Object**:
  - ID;
  - Listing Url;
  - Scrape ID;
  - Last Scraped;
  - Name;
  - Summary;
  - Description;
  - Picture Url;
  - Host URL;
  - Host Name;
  - Host Since;
  - Host Location;
  - Host Thumbnail Url;
  - Host Picture Url;
  - Host Verifications;
  - Street;
  - Neighbourhood Cleansed;

  - City;
  - State;
  - Zipcode;
  - Market;
  - Smart Location;
  - Country Code;
  - Amenities;
  - Calendar Updated;
  - Calendar last Scraped;
  - Geolocation;
  - Features.

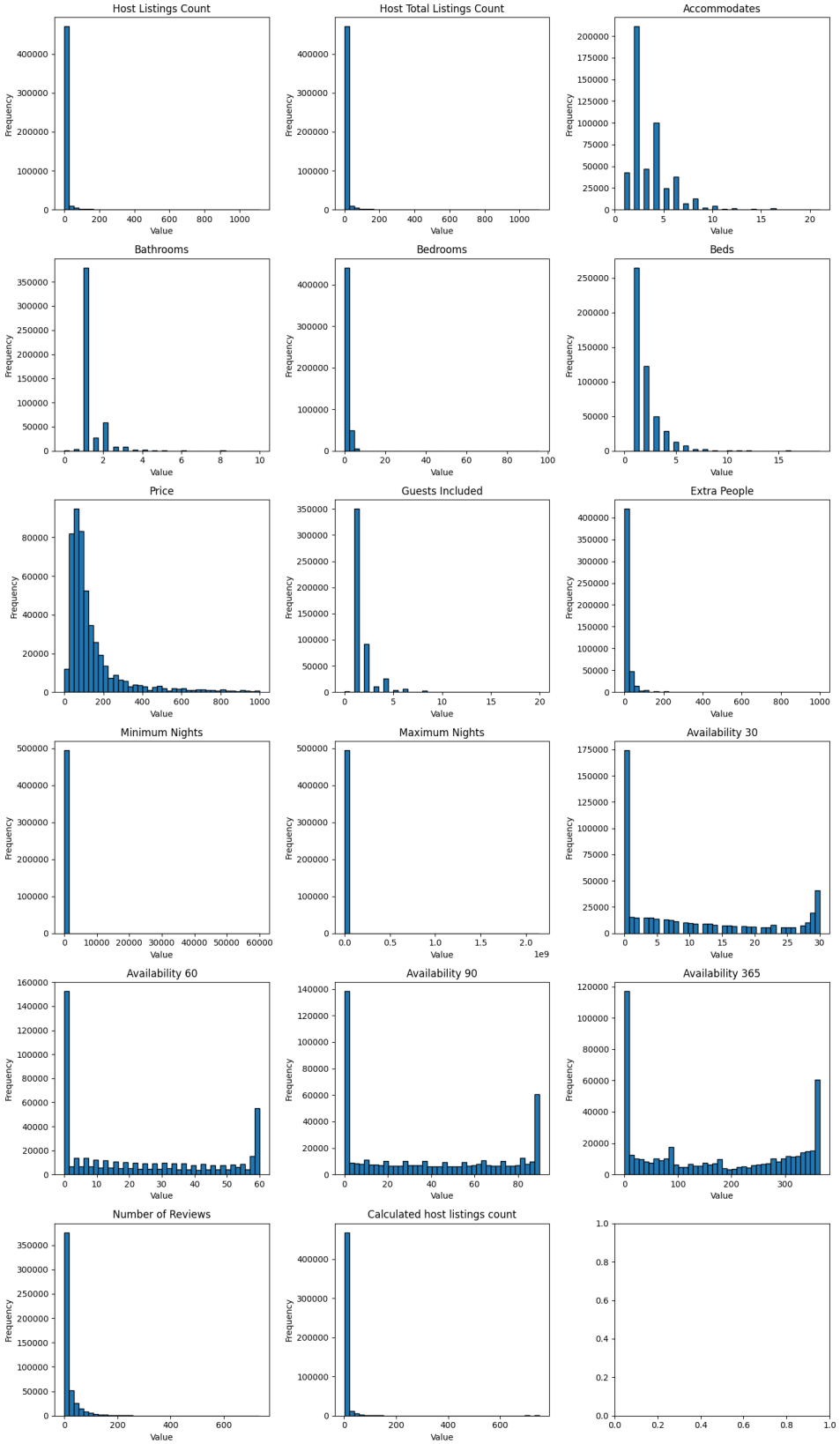
- **Numerical**:
  - Host Listings Count;
  - Host Total Listings Count;
  - Accommodates;
  - Bathrooms;
  - Bedrooms;
  - Beds;
  - Price;
  - Guests Included;
  - Extra People;
  - Minimum Nights;
  - Maximum Nights;
  - Availability 30;
  - Availability 60;
  - Availability 90;
  - Availability 365;
  - Number of Reviews;
  - Calculated host listings count.

1.5 Numerical and Categorical Distributions

For categorical data, we have the following distribution:



For numerical data, we have the following distribution:



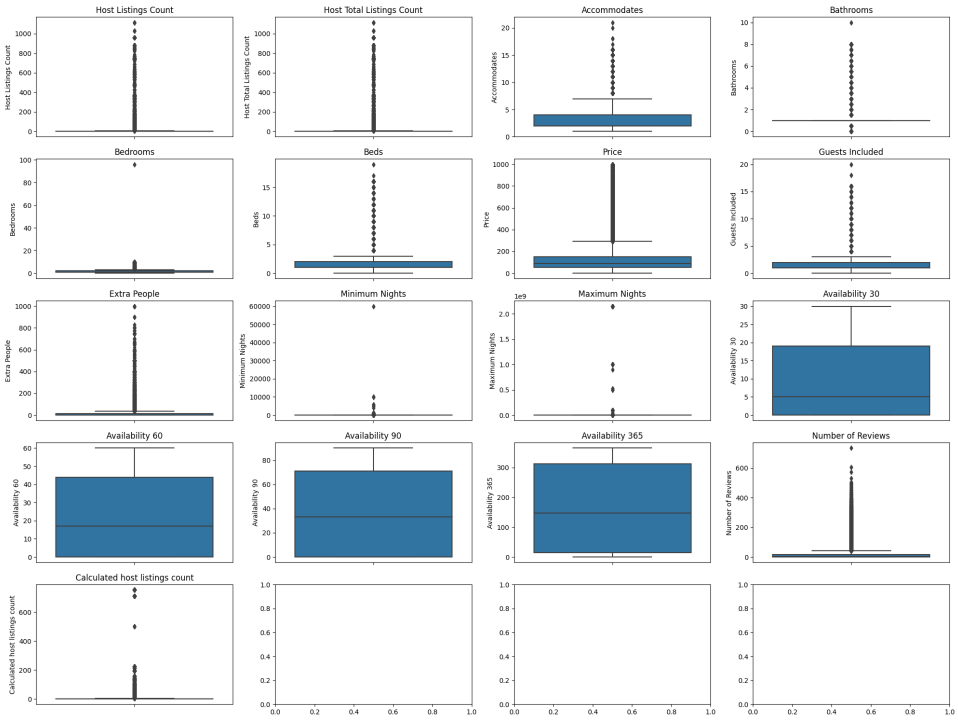
1.6 Outliers

In the numerical dataset, we have the following number of outliers:

Box plots:

Table 2: Outliers in Numerical Features

Feature	**Outliers Count**
Bathrooms	114,272
Calculated host listings count	89,056
Host Listings Count	63,362
Host Total Listings Count	63,362
Beds	57,328
Number of Reviews	55,580
Price	48,906
Guests Included	40,629
Minimum Nights	38,496
Extra People	34,474
Accommodates	24,433
Bedrooms	18,020
Maximum Nights	703
Availability 30	0
Availability 60	0
Availability 90	0
Availability 365	0



1.7 Data Quality and Possible Considerations

The following transformations can be made on categorical variables:

- Room type can be combined into 2 or 3 categories;
- Experiences offered can be dropped due to the appearance of a single value in 99% of the cases;
- A small amount of countries can be combined into a single group;
- Most property types can be combined into a single group.

In addition, latitude and longitude can be used to calculate the distance between the city center and each property in our dataset. I assume that there might be a significant correlation between price and distance from the city center to the property.

Moreover, it is possible to plot by-pair distributions for target vs. categorical features to see the possible patterns and use them to our advantage.

Due to the complexity of object data, it is possible to use Natural Language Processing (NLP) techniques to extract useful features from this data set.