A logo of a company

Description automatically generated

**Cairo University**

**Faculty of Engineering Credit Hours System**

Data Mining, Big Data Analysis

CMPS451

**Airbnb Data Analytics Project**

Subjected to: Eng. Hannah Hatem

Submitted by: Amr Ahmed Abd ElZaher ID: 1190074

Karim Yasser Ali Ragab ID: 1190175

Mostafa Osama Abd Elzaher ID: 1190173

Omar Mohamed Ahmed Amin ID: 1190204

Contents

[I. Problem defenition. 3](#_Toc166448377)

[II. Project pipeline. 5](#_Toc166448378)

[III. Analysis and solution of the problem 5](#_Toc166448379)

[1) Data preprocessing 5](#_Toc166448380)

[Analysis, Visualization, and Insights 6](#_Toc166448381)

[Cities 6](#_Toc166448382)

[Unique cities are 6](#_Toc166448383)

[Distribution of Airbnb listings per city 6](#_Toc166448384)

[Distribution of average price vs the city 7](#_Toc166448385)

[Prices 8](#_Toc166448386)

[Room Type 10](#_Toc166448387)

[Price categories vs Room Type (\*) 10](#_Toc166448388)

[Property Type 12](#_Toc166448389)

[Examples of property types 12](#_Toc166448390)

[Neighborhood 13](#_Toc166448391)

[The neighborhoods that have the most Airbnb listings 13](#_Toc166448392)

[The top 3 neighborhoods in avg price in each city 13](#_Toc166448393)

[Number of accommodates 14](#_Toc166448394)

[Distribution of Airbnb listings vs number of accommodates 14](#_Toc166448395)

[Amenities 15](#_Toc166448396)

[The most frequent amenities using **word cloud** 15](#_Toc166448397)

[The correlation between multiple features 16](#_Toc166448398)

[Reviews 17](#_Toc166448399)

[Price correlated features 18](#_Toc166448400)

[Uncorrelated features 19](#_Toc166448401)

[Model and Ai: 20](#_Toc166448402)

[First: Price Prediction model 20](#_Toc166448403)

[Second: City prediction model 20](#_Toc166448404)

[Future Work: 20](#_Toc166448405)

# Problem definition

The project aims to analyze a comprehensive dataset of Airbnb accommodations across various U.S. cities. The objective is to extract insights that can guide property owners and business stakeholders in making informed decisions regarding property investments, pricing strategies, and market understanding within the Airbnb ecosystem.

The rise of Airbnb as a popular platform for short-term rental accommodations has created a dynamic and competitive market landscape across various cities in the United States. For property owners and business stakeholders, understanding the intricacies of this market is crucial for making strategic decisions regarding property investments, pricing strategies, and market positioning.

**Challenges in the Airbnb Market**:

* **Property Selection**: Property owners need to identify optimal locations and property types that are likely to attract guests and yield high returns.
* **Pricing Strategies**: Determining competitive and profitable pricing strategies based on location, property features, and market demand is essential for maximizing occupancy rates and revenue.
* **Market Insight**: A comprehensive understanding of the Airbnb ecosystem, including guest preferences, amenities, and regional trends, is vital for making informed investment decisions.

**Project Objectives**

The primary goal of this project is to leverage big data analytics techniques to extract actionable insights from a comprehensive Airbnb dataset. These insights will empower property owners and business stakeholders by providing:

1. **Market Understanding**:
   * Identify lucrative regions and property types based on historical Airbnb data.
   * Analyze the impact of various attributes (e.g., property type, amenities, location) on rental success metrics (e.g., occupancy rates, guest satisfaction).
2. **Decision Support**:
   * Enable data-driven property investment decisions by recommending optimal property types and locations.
   * Offer pricing recommendations based on market dynamics and competitive analysis within specific regions.
3. **Competitive Advantage**:
   * Equip property owners with the knowledge to position their listings strategically to maximize market competitiveness and guest satisfaction.

**Dataset Overview**

The Airbnb dataset comprises over 70,000 entries encompassing diverse attributes related to accommodations, hosts, pricing, and guest reviews across multiple U.S. cities. Key dataset features include:

* **Property Attributes**: Property type, room type, amenities, and listing descriptions.
* **Host Characteristics**: Host profile and hosting history.
* **Geographical Details**: Location coordinates, neighborhood information, and city.
* **Guest Reviews**: Ratings, comments, and feedback from previous guests.

**Significance of Analysis**

The insights derived from this analysis will have profound implications for:

* **Property Owners**: Optimize property investments and improve rental performance based on data-driven strategies.
* **Business Owners**: Enhance revenue generation through informed pricing decisions and market positioning strategies.
* **Industry Stakeholders**: Gain a deeper understanding of the evolving Airbnb market landscape and emerging trends.

By conducting a comprehensive analysis of the Airbnb dataset, this project aims to bridge the gap between data analytics and real-world decision-making in the dynamic short-term rental market, ultimately fostering success and competitiveness for property owners and stakeholders.

# Project pipeline

A diagram of a system

Description automatically generated with medium confidence

Figure : Project pipeline.

**Data Preprocessing**: Cleanse and preprocess the dataset to handle missing values, outliers, and ensure uniform data formatting for analysis.

**Exploratory Data Analysis (EDA)**:

* + **Visualization**: Utilize visualization techniques to understand data distributions, correlations, and trends.
  + **Insights Extraction**: Identify key patterns and insights from the dataset to inform property investment strategies and pricing decisions.

**Model Training**:

* + Develop predictive models (Polynomial regression, Random Forest) to forecast prices
  + Develop a city prediction model using SVM

# Analysis and solution of the problem

## Data preprocessing

* + - Convert the log\_price column to price by taking the exponent of each value
    - Categorize the price column to ease the analysis
    - Remove useless columns
    - Remove null values

# Analysis, Visualization, and Insights

The Following section shows the EDA done and insights found in the data

## Cities

### Unique cities

* San Fransisco - SF
* New York - NYC
* Washington - DC
* Los Angeles - LA
* Boston
* Chicago

### Distribution of Airbnb listings per city

A graph with different colored squares

Description automatically generated

* **NYC** is the city with the most Airbnb offerings with over 32,300 Airbnb listings.
* **Boston** has the minimum Airbnb offerings with count of 3,468 Airbnb listings.

### Distribution of average price vs the city

A graph of different colored squares

Description automatically generated

* **San Francisco SF** is the city with highest average price which is equal to 227.37 USD.
* **Chicago** is the city with the lowest average price, with avg of 132.47 USD.

## Prices

The mean value: 160.37 usd

25% 75.0 50% 111.0 75% 185.0

Maximum price: 1999.0 usd Minimum price: 1.0 usd

The std: 168.580415

A graph with blue and white lines

Description automatically generated

After removing the Airbnb listings with price > 500 USD.

A graph of blue bars

Description automatically generated with medium confidence

We decided to categorize the prices ranges to ease collecting insights.

A graph of different colored bars

Description automatically generated

N.B: each category spans an incremental range of 10 USD, Color code of price categories:

A graph of a graph

Description automatically generated with medium confidenceA graph of a graph

Description automatically generated with medium confidence

## Room Type

Unique room types are:

* Shared Room
* Private Room
* Entire Home/Apt

### Price categories vs Room Type (\*)

#### Shared Room

A graph of a graph

Description automatically generated with medium confidence

* We could see that the most listings with shared room type are concentrated in low budget end of the price categories

#### Private Room

A graph of a graph

Description automatically generated with medium confidence

* The price distribution for private room is larger than the price distribution of shared room,

#### Entire home/apt

A graph of different colored lines

Description automatically generated

* The price distribution for entire home/apt spans more expensive categories.

## Property Type

### Examples of property types

Apartment, House, Condominium, Hostel, Dorm, Camper/RV, Villa, Boutique hotel, Timeshare, Boat, Castle, Treehouse, Tipi,..

The correlation between property type, room type and average price (color code column)

A screenshot of a data

Description automatically generated

* If we look at shared room column, we could see that the **avg price** is very low (31 - 180) usd,

Which shows that shared rooms offerings have low budget price.

* For the private room

## Neighborhood

Each city has a number of neighborhoods, the count and avg price of airbnbs in each neighborhood differ.

### The neighborhoods that have the most Airbnb listings

A graph with blue and white stripes

Description automatically generated

### The top 3 neighborhoods in avg price in each city

A screenshot of a table

Description automatically generated

* Chicago city
  + Neighborhood of highest avg price: Old Town – 214.7usd
  + Neighborhood of lowest avg price: South Chicago – 25 usd
* NYC city
  + Neighborhood of highest avg price: Mill Basin – 500 usd
  + Neighborhood of lowest avg price: Morris Park – 43.6 usd
* Boston city
  + Neighborhood of highest avg price: Cambridge – 324.5 usd
  + Neighborhood of lowest avg price: Somerville – 54 usd
* SF city
  + Neighborhood of highest avg price: Sea Cliff – 797 usd
  + Neighborhood of lowest avg price: West Portal – 125 usd
* DC city
  + Neighborhood of highest avg price: Chevy Chase, MD – 1250 usd
  + Neighborhood of lowest avg price: Suitland-Silver Hill, MD – 37 usd
* LA city
  + Neighborhood of highest avg price: Wilmington – 1300 usd
  + Neighborhood of lowest avg price: La Puente – 40 usd

## Number of accommodates

The number of accommodates is the maximum number of people a listing can accommodate.

### Distribution of Airbnb listings vs number of accommodates

A graph of a bar graph

Description automatically generated with medium confidence

* The most common Airbnb offerings are for places that allow up to **2 accommodates**.
* The mean is 3.155 accommodates.

## Amenities

### The most frequent amenities using **word cloud**

A close up of words

Description automatically generated

* Wireless Internet, Smoke detector, Air conditioning are the most frequent amenities in airbnb listings.
* Few Airbnb allow pets.

## The correlation between multiple features

A screenshot of a computer screen

Description automatically generated

From this heat map we can conclude that,

* The correlation between the number of accommodates and the number of beds is very high (0.81)
* The correlation between the number of accommodates and the number of bedrooms is very high (0.71)
* The correlation between the number of accommodates and the number of bathrooms is not high.

## Reviews

A graph of a bar chart

Description automatically generated with medium confidence

* We could conclude that most of the reviews are for positive sentiment

(70 – 100)%

* As we said it is not correlated with the price
* Few reviews are less than 60%

A graph showing a number of reviews

Description automatically generated

* From this graph, it is clear that the number of reviews does not depend on the price.

## Price correlated features

A chart with different colored squares

Description automatically generated

A graph of different colored squares

Description automatically generated

## Uncorrelated features

A graph with different colored bars

Description automatically generatedCorrelation between city and cancellation policy

A screenshot of a computer screen

Description automatically generatedWe conclude that the cancellation policy does not vary from a city to another, and it has almost the same distribution among all cities.

The number of reviews, reviews scores and price are not correlated

# Model and Ai:

## First: Price Prediction model

We used 80% of the dataset for training and 20% for testing

We developed 2 model to predict the price, we dropped the null enteries, and used the following features: ['beds', 'bedrooms', 'city', 'number\_of\_reviews', 'cancellation\_policy', 'review\_scores\_rating', 'room\_type', 'property\_type', 'neighbourhood', 'cleaning\_fee', 'instant\_bookable', 'accommodates', 'amenities', 'latitude', 'longitude']

For the categorial features we factorized them in order to use them in the training

First model was Polynomial Regression model it gave accuracy equal to 0.67463935

Second model was a Random Forest which gave a better performance and accuracy equal to 0.71992

## Second: City prediction model

We developed an SVM model that takes ['price','bathrooms', 'cleaning\_fee', 'bedrooms', 'number\_of\_reviews', 'cancellation\_policy', 'review\_scores\_rating', 'latitude', 'longitude'] as features

And predicts the city

It gave an accuracy of 99.89% and F1 score of 0.9985209

Failed Trials: before getting those results, we tried decision tree and logistic regression models for the price, they both gave very low accuracies, we tried using SVM and Random forest but it still didn’t improve much, it was fixed by using the log(price) instead of the price in training.

## Future Work:

If we can get more data from different Airbnbs around the world we can find relations between the features that our models and analysis found uncorrelated such as price and cancellation policy and review score vs cleaning fee, as those insights have a very high chance of affecting each other.

Also having more diverse data can help business owners decide where to start their business no only within the US major cities but also around the world