Data 602 – Final Project Proposal

By: Mubashira Qari

Project Title: **Airbnb Data Analysis and Price Prediction Using Machine Learning**

Describing Dataset: This is an Airbnb listing dataset. This dataset is acquired from the Airbnb website. It has details about the host, location, type of listing, and reviews, along with more interesting features in different columns. It consists of (18337, 18) rows and columns and needs thorough data cleaning. The analysis will provide interesting insights and guidelines on choosing Airbnb for your next vacation trip.The metrics for the columns in the dataset are as below:

id	float64	Airbnb's unique identifier for the host/user
name	object	name of the listing
host_id	int64	unique identifier for the host
host_name	object	name of the host
neighbourhood_group	float64	next town zip code
neighbourhood	int64	Field that can be filled using the lat & Long
latitude	float64	latitude
longitude	float64	longitude
room_type	object	type of rental space
price	int64	price
minimum_nights	int64	minimum nights available
number_of_reviews	int64	count of reviews

last_review	object	date of last review
reviews_per_month	float64	per month reviews
calculated_host_listings_count	int64	total host's listing count
availability_365	int64	Availability for the number of days
number_of_reviews_ltm	int64	Count of reviews in last 12 months
license		permits for listing

Data Source Link:

Data Source: http://insideairbnb.com/get-the-data/

Data Dictionary: https://docs.google.com/spreadsheets/d/1iWCNJcSutYqpULSQHlNyGInUvHg2BoUGoNRIGa6Szc4/edit#gid=1322284596

Justification for Dataset Selection:

Whenever it comes to traveling, there is always a search for a good rating Airbnb listing. There are many important factors that need to consider. For example, the room type, comfort rating, activity, top hosts, rent, and many others. And paying close attention to your selection adds a lot to the enjoyment of the trip. For this reason, I decided to choose the Airbnb dataset.Now there are couple of questions related to electric vehicle that require analysis and those are my research questions.

Research Questions & Objectives:

The research questions for the electric vehicle’s dataset are the following:

1- Which listing has the best reviews?

2- What is the availability in days?

3- What is the type of room?

4- What is the correlation between the variables.

5- What is the price of the listing?

5- What are the description of the listing?

6- Which factors affect the price of the listing?

7- Upload the dataset in the SQL database for further analysis.

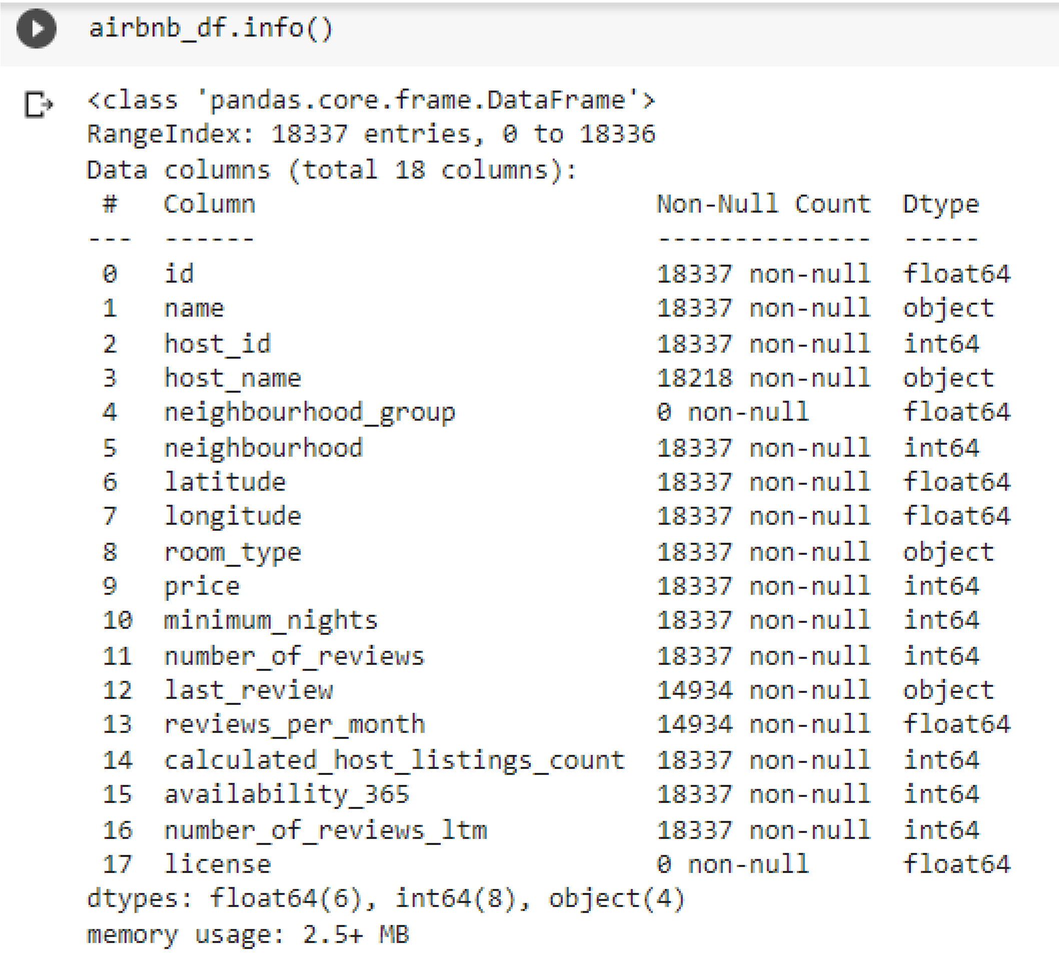
8- Predict the price of Airbnb listing using machine learning.

Libraries Used for Project Implementation:

- Python Pandas
- Python NumPy
- Python Matplotlib
- SQL Database
- Python sklearn
- Python seaborn

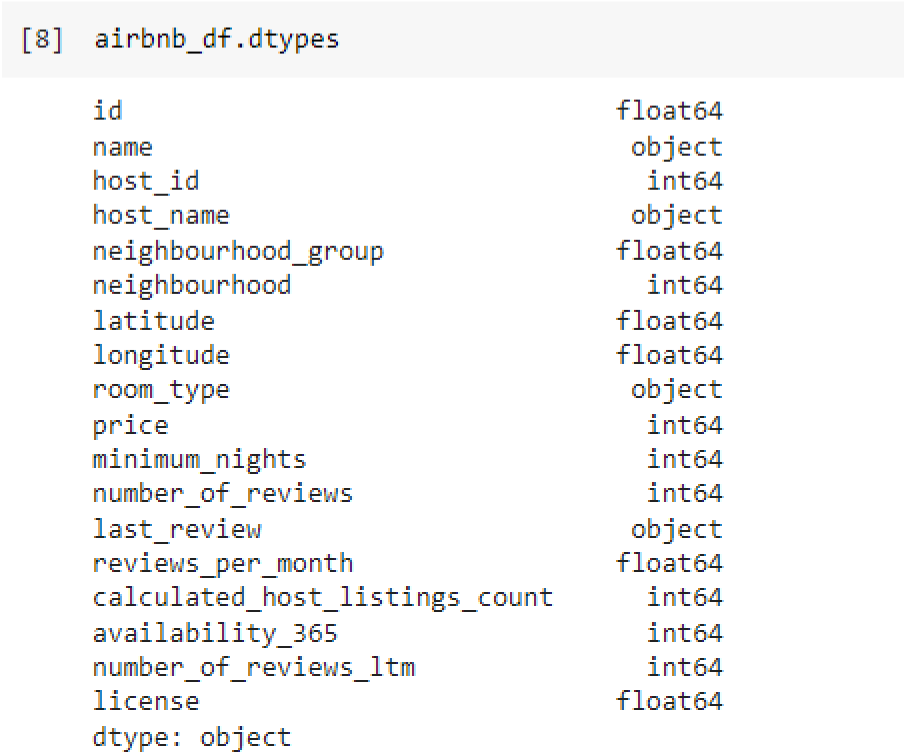
EDA and Summary Statistics:

Below are the images of exploratory data analysis and summary statistics:



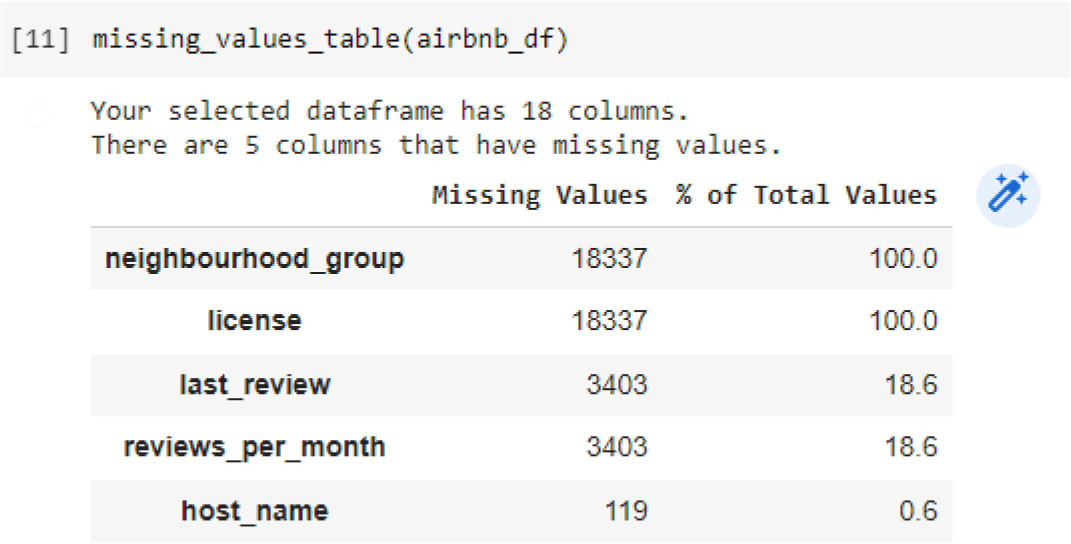
```
airbnb_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18337 entries, 0 to 18336
Data columns (total 18 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              18337 non-null  float64
 1   name                            18337 non-null  object 
 2   host_id                         18337 non-null  int64  
 3   host_name                       18218 non-null  object 
 4   neighbourhood_group             0 non-null      float64
 5   neighbourhood                   18337 non-null  int64  
 6   latitude                        18337 non-null  float64
 7   longitude                       18337 non-null  float64
 8   room_type                       18337 non-null  object 
 9   price                           18337 non-null  int64  
 10  minimum_nights                  18337 non-null  int64  
 11  number_of_reviews               18337 non-null  int64  
 12  last_review                     14934 non-null  object 
 13  reviews_per_month               14934 non-null  float64
 14  calculated_host_listings_count  18337 non-null  int64  
 15  availability_365                18337 non-null  int64  
 16  number_of_reviews_ltm           18337 non-null  int64  
 17  license                         0 non-null      float64
dtypes: float64(6), int64(8), object(4)
memory usage: 2.5+ MB
```



```
[8]  airbnb_df.dtypes
```

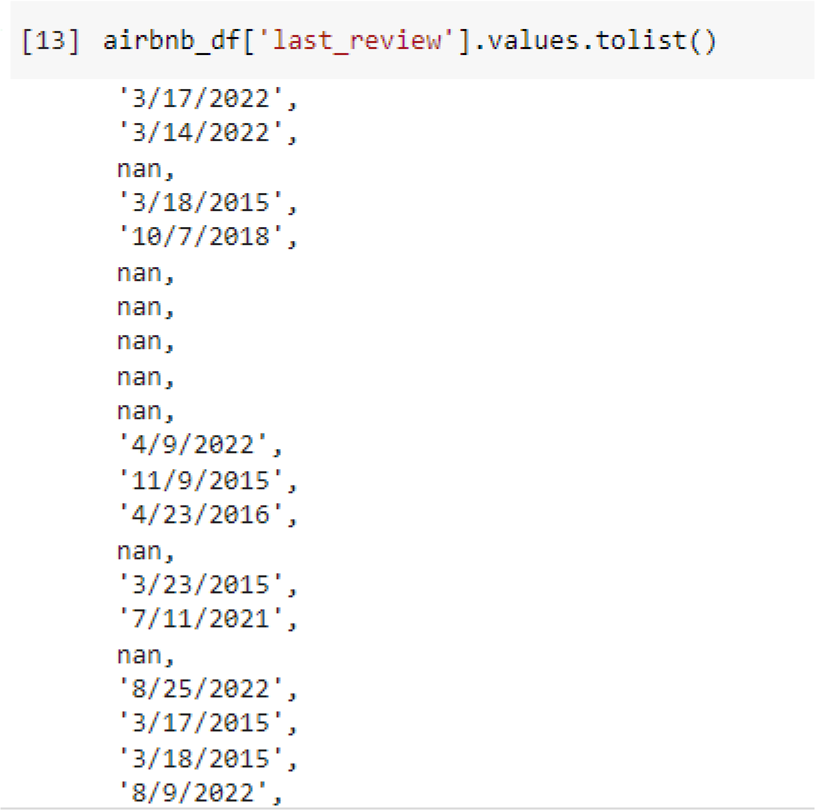
```
id                                 float64
name                                object
host_id                              int64
host_name                           object
neighbourhood_group                float64
neighbourhood                        int64
latitude                           float64
longitude                          float64
room_type                           object
price                                int64
minimum_nights                       int64
number_of_reviews                    int64
last_review                         object
reviews_per_month                  float64
calculated_host_listings_count       int64
availability_365                     int64
number_of_reviews_ltm                int64
license                            float64
dtype: object
```



```
[11] missing_values_table(airbnb_df)
```

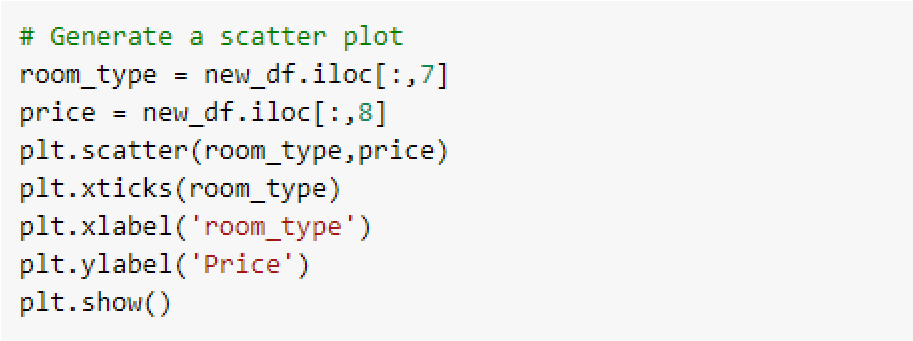
```
Your selected dataframe has 18 columns.
There are 5 columns that have missing values.
```

	Missing Values	% of Total Values
neighbourhood_group	18337	100.0
license	18337	100.0
last_review	3403	18.6
reviews_per_month	3403	18.6
host_name	119	0.6

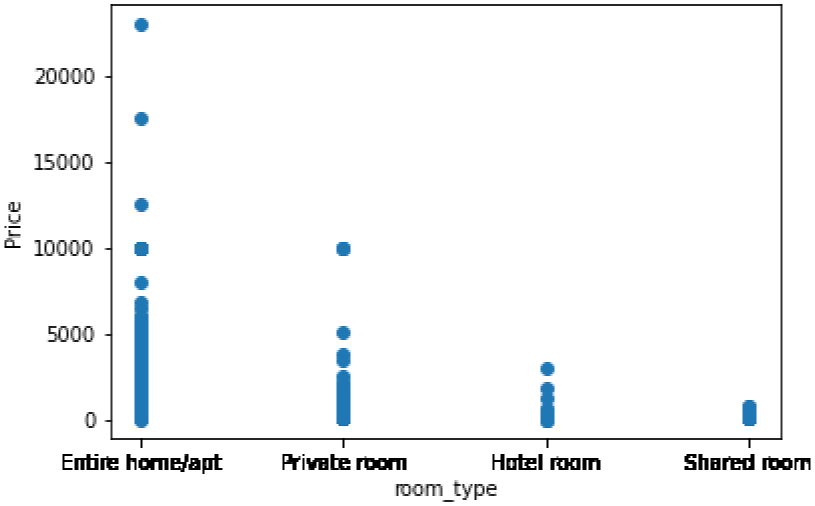


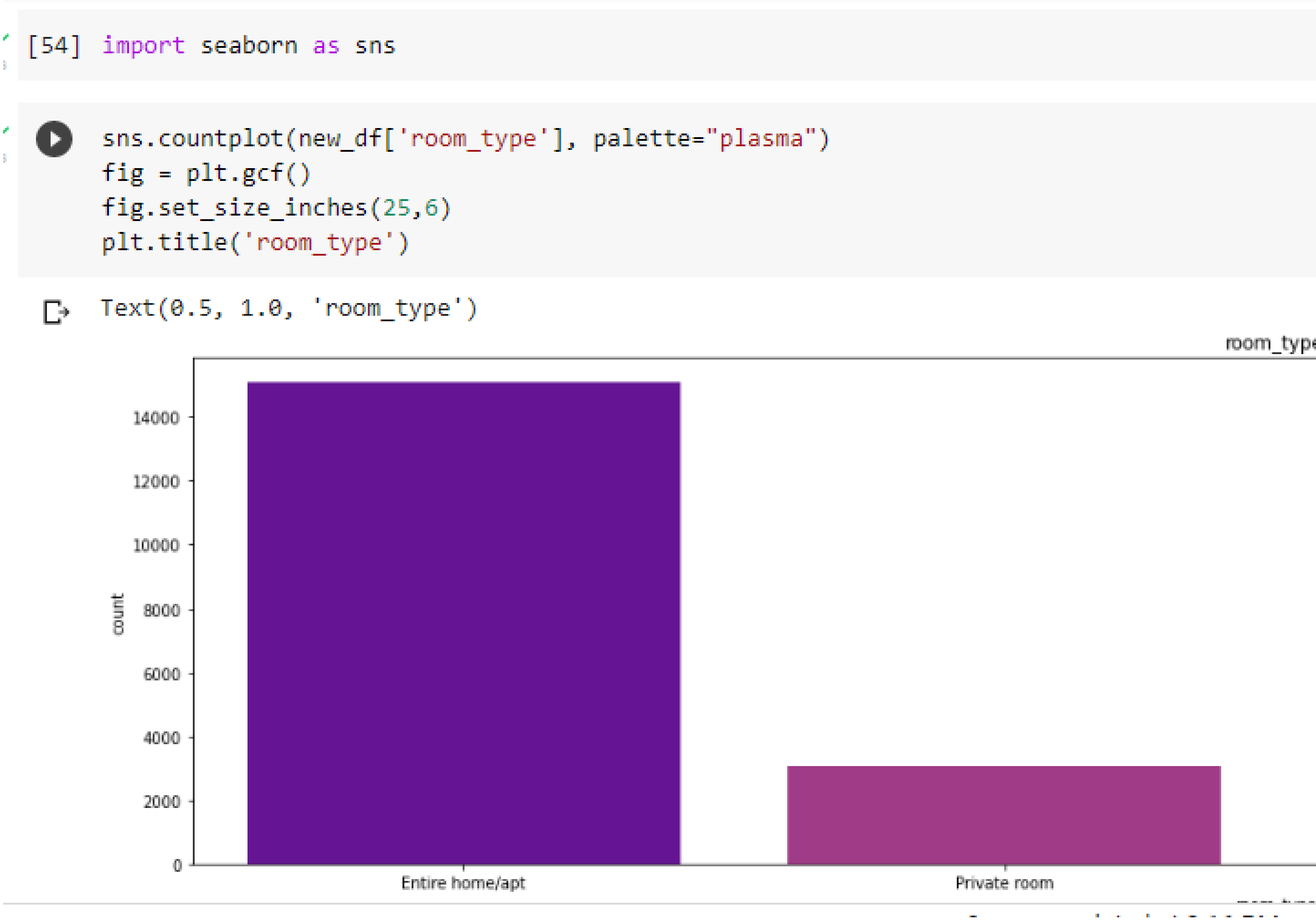
```
[13] airbnb_df['last_review'].values.tolist()
```

```
       '3/17/2022',
       '3/14/2022',
       nan,
       '3/18/2015',
       '10/7/2018',
       nan,
       nan,
       nan,
       nan,
       nan,
       '4/9/2022',
       '11/9/2015',
       '4/23/2016',
       nan,
       '3/23/2015',
       '7/11/2021',
       nan,
       '8/25/2022',
       '3/17/2015',
       '3/18/2015',
       '8/9/2022',
```



```
# Generate a scatter plot
room_type = new_df.iloc[:,7]
price = new_df.iloc[:,8]
plt.scatter(room_type,price)
plt.xticks(room_type)
plt.xlabel('room_type')
plt.ylabel('Price')
plt.show()
```





```
[54] import seaborn as sns
```

```
sns.countplot(new_df['room_type'], palette="plasma")
fig = plt.gcf()
fig.set_size_inches(25,6)
plt.title('room_type')
```

```
Text(0.5, 1.0, 'room_type')
```