DS 412: Statistical Data Analysis Lab

# Lab - Report

Prepared By:

**Meher Durdana Khan**

ID:192-35-2818

Sec-A

Department of Software Engineering

Course Teacher:

**Musabbir Hasan Sammak**

**Lecturer**

Department of Software Engineering

Daffodil International University

# Breakdown of this notebook:

1. **Importing Libraries**
2. **Loading the dataset**
3. **Data Cleaning:**
    - Deleting redundant columns.
    - Dropping duplicates.
    - Cleaning individual columns.
    - Remove the NaN values from the dataset
    - Some Transformations
4. **Data Visualization: Using plots to find relations between the features.**

    a.Histogram

    b.Density Plot

    c.Boxplot

    d.Scatter Plot

    e.Heatmap

    f.Correlogram

    g.Bubble Chart

    h.Bar Plot

    i.Word Cloud

    j.Grouped Bar Chart

    k.Stacked Bar Chart

# Import Libraries

We'll first need to import the relevant libraries.

In [153]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

# Load Data

Next, we'll need to load our AirBnb dataset.

In [177]:
```
df=pd.read_csv("DS 332 Lab Final Dataset - DS 332 Lab Final Dataset.csv")
df
```

Out[177]:

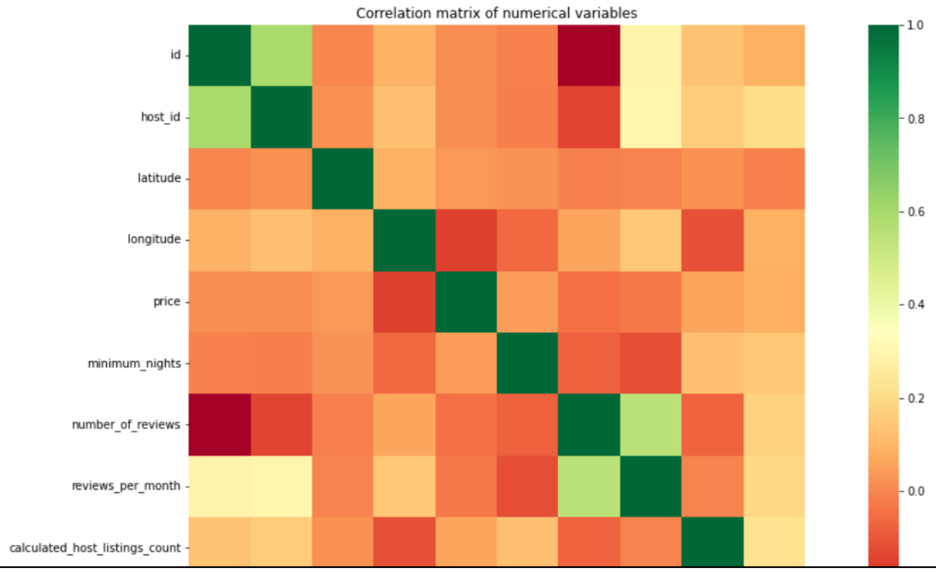
| | Serial_No | Degree | GRE_Score | TOEFL_Score | University_Rating | SOP | LOR | CGPA | Research | Chance_of_Admit |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | B.Sc | 337.0 | 118.0 | 4.0 | 4.5 | 4.5 | 9.65 | 1.0 | 0.92 |
| **1** | 2 | B.Sc | 324.0 | 107.0 | 4.0 | 4.0 | 4.5 | 8.87 | 1.0 | 0.76 |
| **2** | 3 | B.Sc | 316.0 | 104.0 | 3.0 | 3.0 | 3.5 | 8.00 | 1.0 | 0.72 |
| **3** | 4 | B.Sc | 322.0 | 110.0 | 3.0 | 3.5 | 2.5 | 8.67 | NaN | 0.80 |
| **4** | 5 | B.Sc | 314.0 | 103.0 | 2.0 | 2.0 | 3.0 | 8.21 | 0.0 | NaN |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **395** | 396 | M.Sc | 324.0 | 110.0 | 3.0 | 3.5 | 3.5 | 9.04 | 1.0 | 0.82 |
| **396** | 397 | M.Sc | 325.0 | 107.0 | 3.0 | 3.0 | 3.5 | 9.11 | 1.0 | 0.84 |
| **397** | 398 | M.Sc | 330.0 | 116.0 | 4.0 | 5.0 | 4.5 | 9.45 | 1.0 | 0.91 |
| **398** | 399 | B.Sc | 312.0 | 103.0 | 3.0 | 3.5 | 4.0 | 8.78 | 0.0 | 0.67 |
| **399** | 400 | NaN | 333.0 | 117.0 | 4.0 | 5.0 | 4.0 | 9.66 | 1.0 | 0.95 |

400 rows × 10 columns

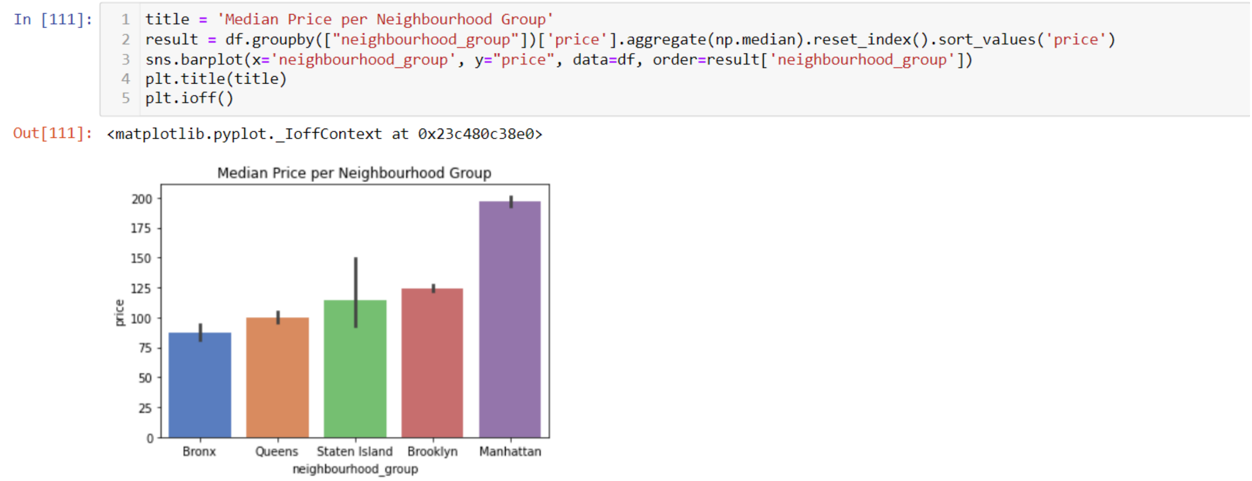
# Get Correlation between different variables

In [87]:
```
plt.figure(figsize=(20,10))
title = 'Correlation matrix of numerical variables'
sns.heatmap(df.corr(), square=True, cmap='RdYlGn')
plt.title(title)
plt.ioff()
```

Out[87]:
```
<matplotlib.pyplot._IoffContext at 0x23c3e9f2df0>
```

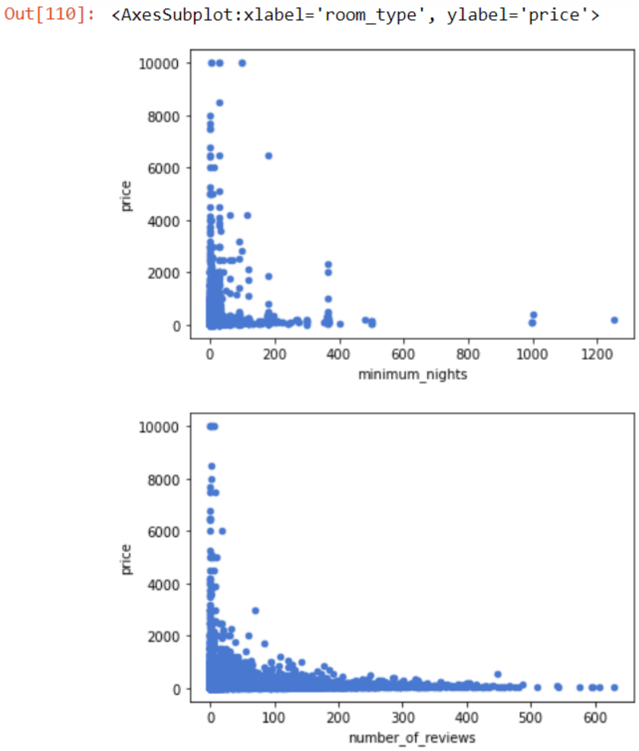


## Visualize data using appropriate graphs and charts using matplotlib/seaborn/plotly.

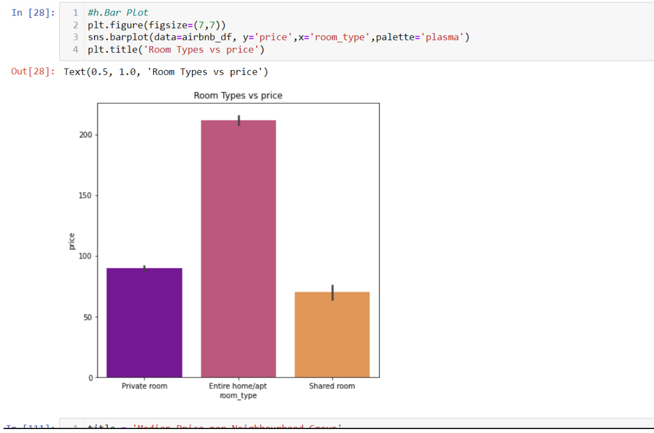


```
In [111]:
title = 'Median Price per Neighbourhood Group'
result = df.groupby(["neighbourhood_group"])['price'].aggregate(np.median).reset_index().sort_values('price')
sns.barplot(x='neighbourhood_group', y="price", data=df, order=result['neighbourhood_group'])
plt.title(title)
plt.ioff()

Out[111]: <matplotlib.pyplot._IoffContext at 0x23c480c38e0>
```



```
Out[110]: <AxesSubplot:xlabel='room_type', ylabel='price'>
```



```
In [28]:
#h.Bar Plot
plt.figure(figsize=(7,7))
sns.barplot(data=airbnb_df, y='price',x='room_type',palette='plasma')
plt.title('Room Types vs price')

Out[28]: Text(0.5, 1.0, 'Room Types vs price')
```

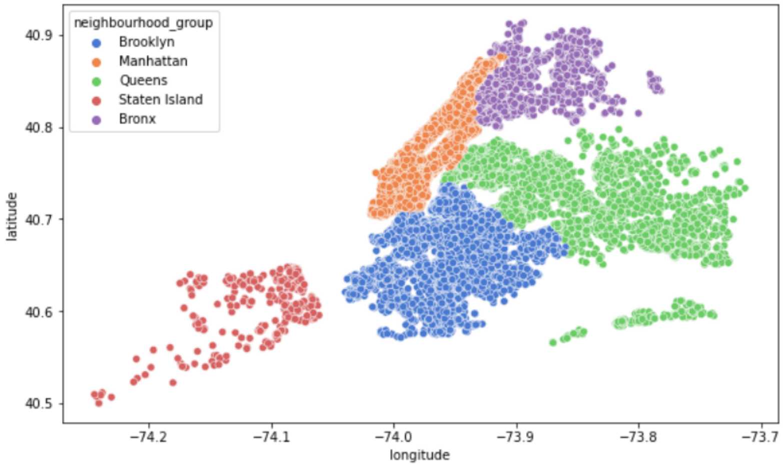
```
In [111]:
title = 'Median Price per Neighbourhood Group'
```

In [112]:

```
plt.figure(figsize=(10,6))
sns.scatterplot(df.longitude,df.latitude,hue=df.neighbourhood_group)
plt.ioff()
```

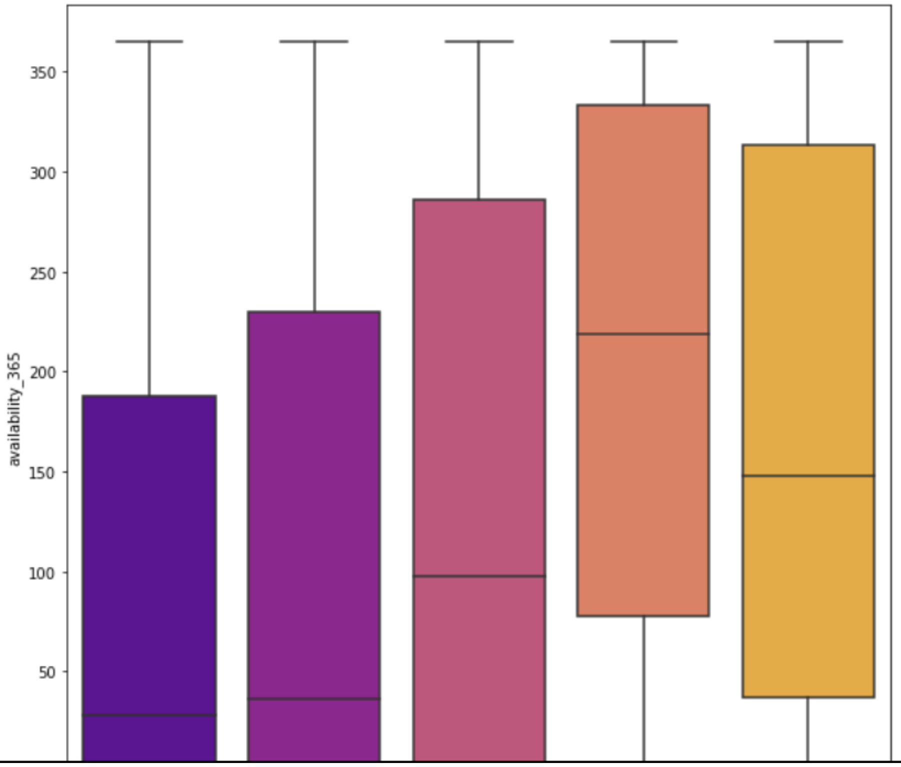
Out[112]:

```
<matplotlib.pyplot._IoffContext at 0x23c47b0c340>
```



In [62]:

```
#c.Boxplot
plt.figure(figsize=(10,10))
ax = sns.boxplot(data=df, x='neighbourhood_group',y='availability_365',palette='plasma')
```


availability_365