AIRBNB HOTEL BOOKING ANALYSIS

NAME : SEJAL KARNWAL
AICTE ID : STU64548ff77b0d11683263479

PROBLEM STATEMENT

Airbnb is a widely-used platform that allows property owners to rent out their homes or apartments to travelers, creating a dynamic marketplace for short-term lodging. A significant challenge for hosts is determining the optimal price for their listings. This pricing decision is complex because the value of a listing depends on numerous factors, including but not limited to, the property's size, location, amenities, cleanliness, and the host's communication quality.

Without a data-driven approach, hosts may underprice their listings and lose potential revenue, or overprice them and fail to attract guests. This project aims to analyze a dataset of Airbnb listings to identify the key factors that influence pricing. By uncovering the relationships between a listing's price and its various features, we can provide hosts with actionable insights to help them set competitive and effective prices, ultimately optimizing their profitability and improving the overall marketplace efficiency.

Project Description

This project focuses on building a machine learning model to predict the price of Airbnb listings. 🏡 Correctly pricing an Airbnb property is vital for both hosts and travelers. Hosts aim to maximize their earnings and occupancy, while travelers seek fair and competitive prices. Using historical Airbnb data, we'll develop a regression model. This model will learn the complex relationships between various listing attributes, such as the number of bedrooms and bathrooms, and the guest ratings, and the final price of a listing. 🤖
The completed model can then be used to predict prices for new or hypothetical listings, helping property owners make data-driven, informed pricing decisions. This approach moves beyond guesswork, ensuring a more efficient and profitable experience for hosts and a better value for guests.

WHO ARE THE END USERS?

- **Airbnb Hosts**
  To optimize pricing of their listings based on property features and guest reviews.

- **Travelers**
  To evaluate whether a listing is overpriced or reasonably priced.

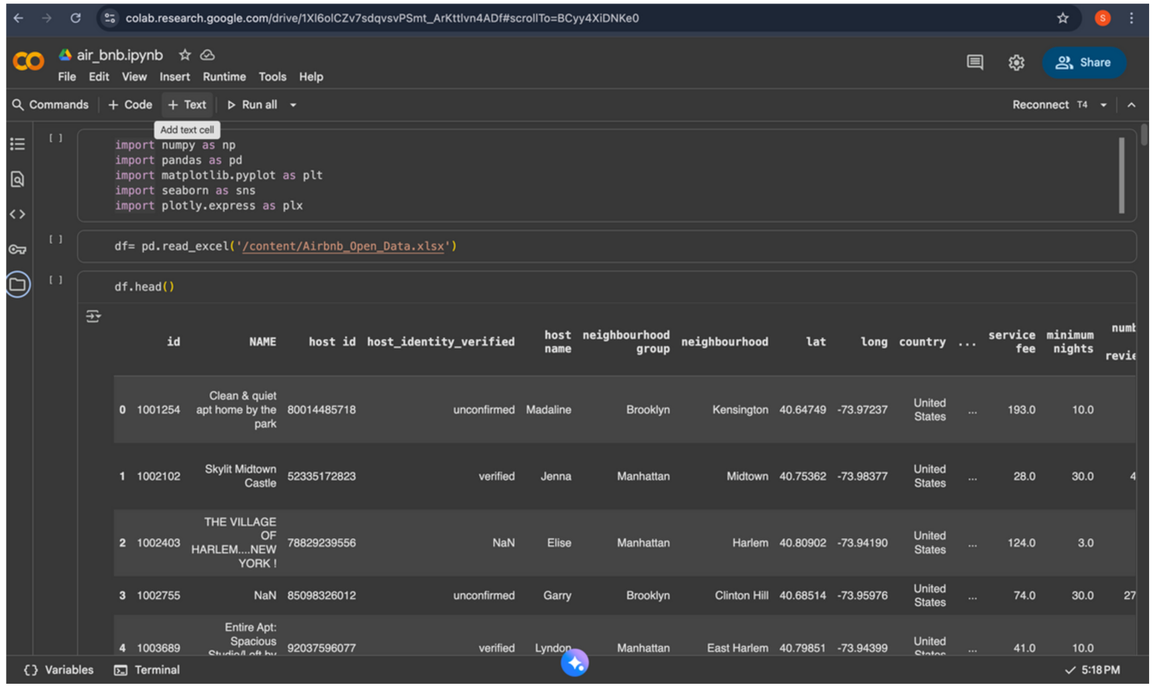
- **Airbnb Platform Analysts**
  To improve automated pricing suggestions and increase platform trust.

- **Researchers/Students**
  To study the impact of property features and reviews on rental pricing.

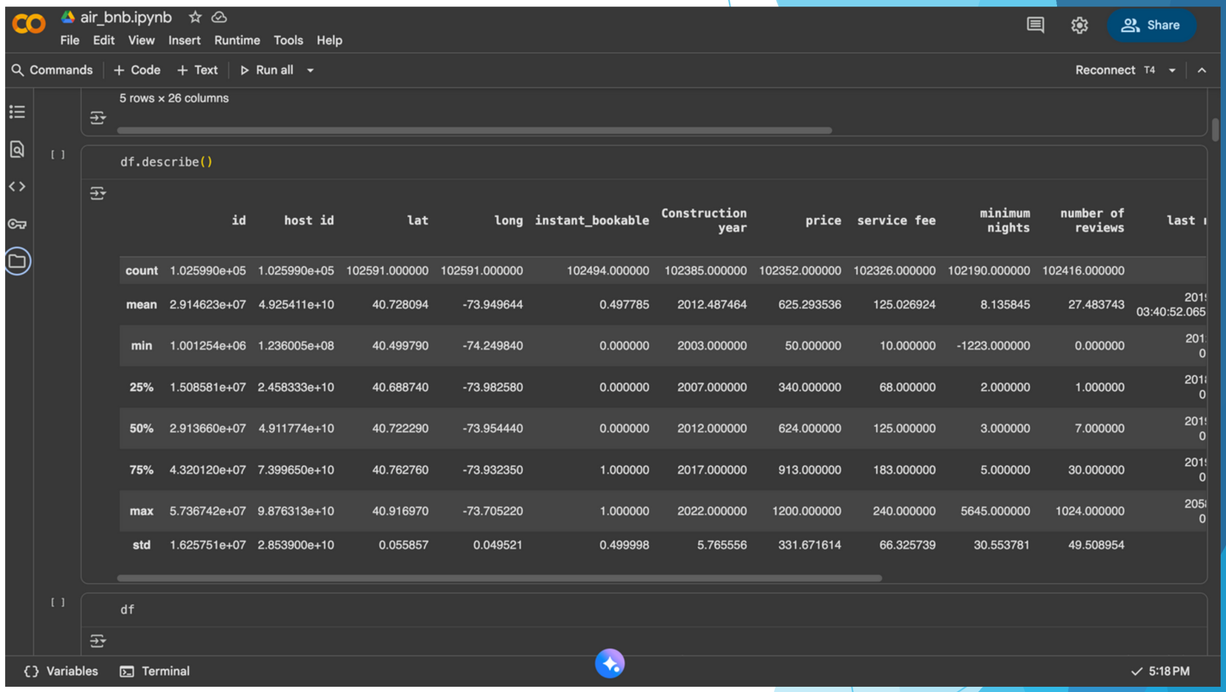
Technology Used

- **Python -** Core programming language
- **Pandas & NumPy** – Data cleaning and preprocessing
- **Scikit-learn** – Machine learning (model training, regression, evaluation)
- **Matplotlib/Seaborn** – Data visualization and feature importance
- **Google Colab** – Cloud-based environment for running the project
- **File handling libraries** – openpyxl (for Excel) and built-in CSV handling

RESULTS1

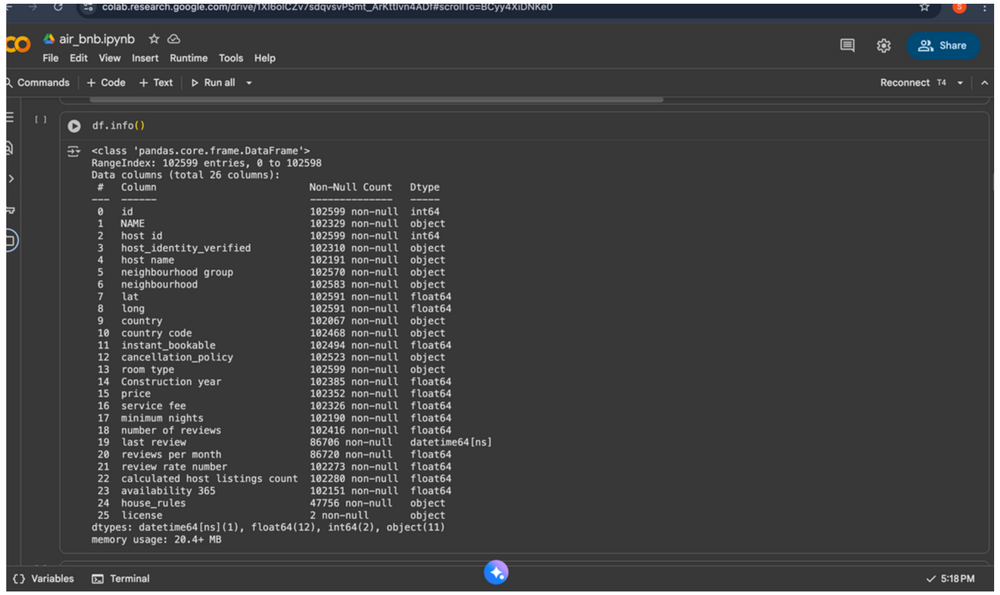


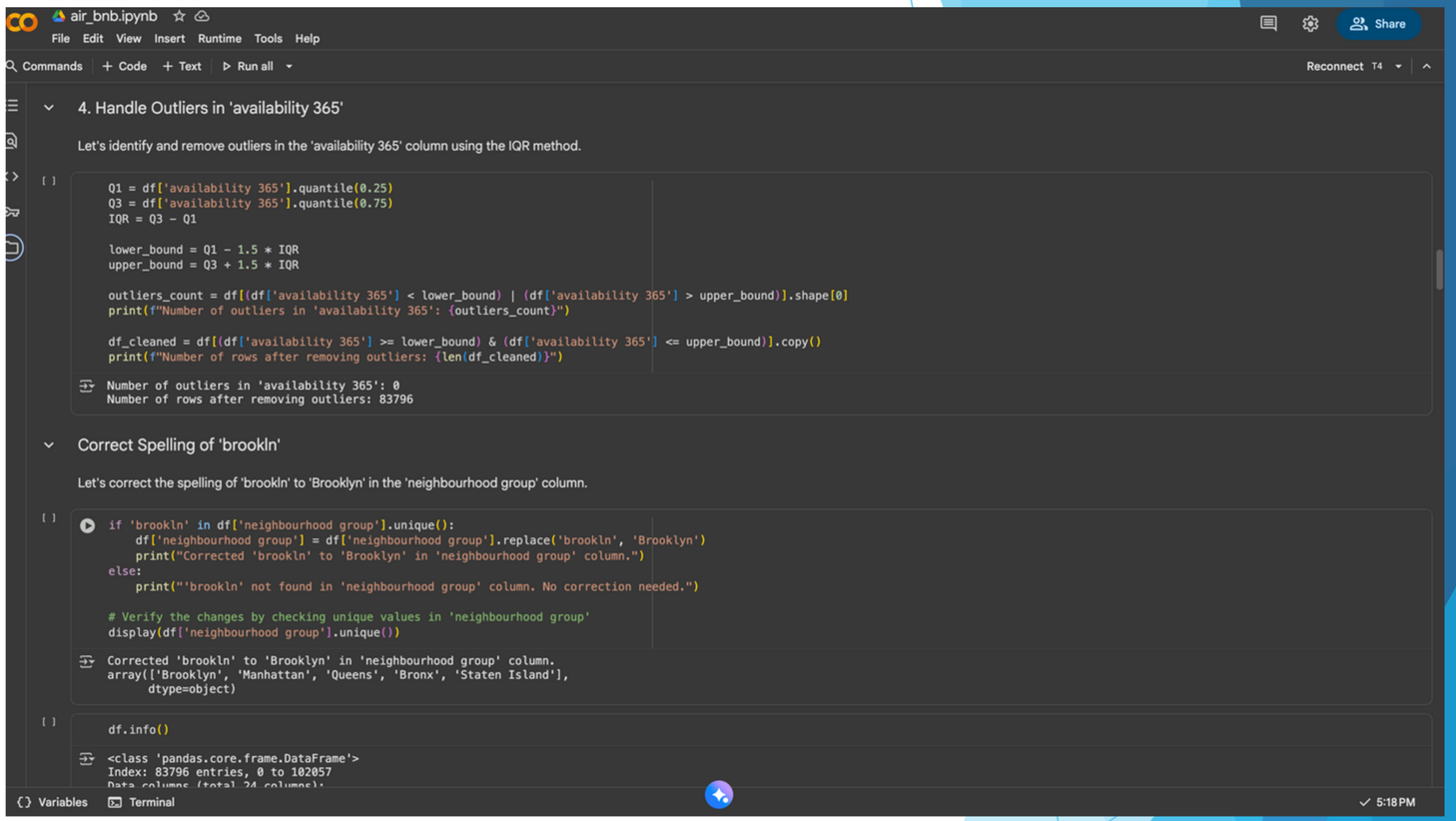
	id	NAME	host id	host_identity_verified	host name	neighbourhood group	neighbourhood	lat	long	country	...	service fee	minimum nights	numb revie
0	1001254	Clean & quiet apt home by the park	80014485718	unconfirmed	Madaline	Brooklyn	Kensington	40.64749	-73.97237	United States	...	193.0	10.0	
1	1002102	Skylit Midtown Castle	52335172823	verified	Jenna	Manhattan	Midtown	40.75362	-73.98377	United States	...	28.0	30.0	4
2	1002403	THE VILLAGE OF HARLEM....NEW YORK !	78829239556	NaN	Elise	Manhattan	Harlem	40.80902	-73.94190	United States	...	124.0	3.0	
3	1002755	NaN	85098326012	unconfirmed	Garry	Brooklyn	Clinton Hill	40.68514	-73.95976	United States	...	74.0	30.0	27
4	1003689	Entire Apt: Spacious Studio/Loft by	92037596077	verified	Lyndon	Manhattan	East Harlem	40.79851	-73.94399	United States	...	41.0	10.0	



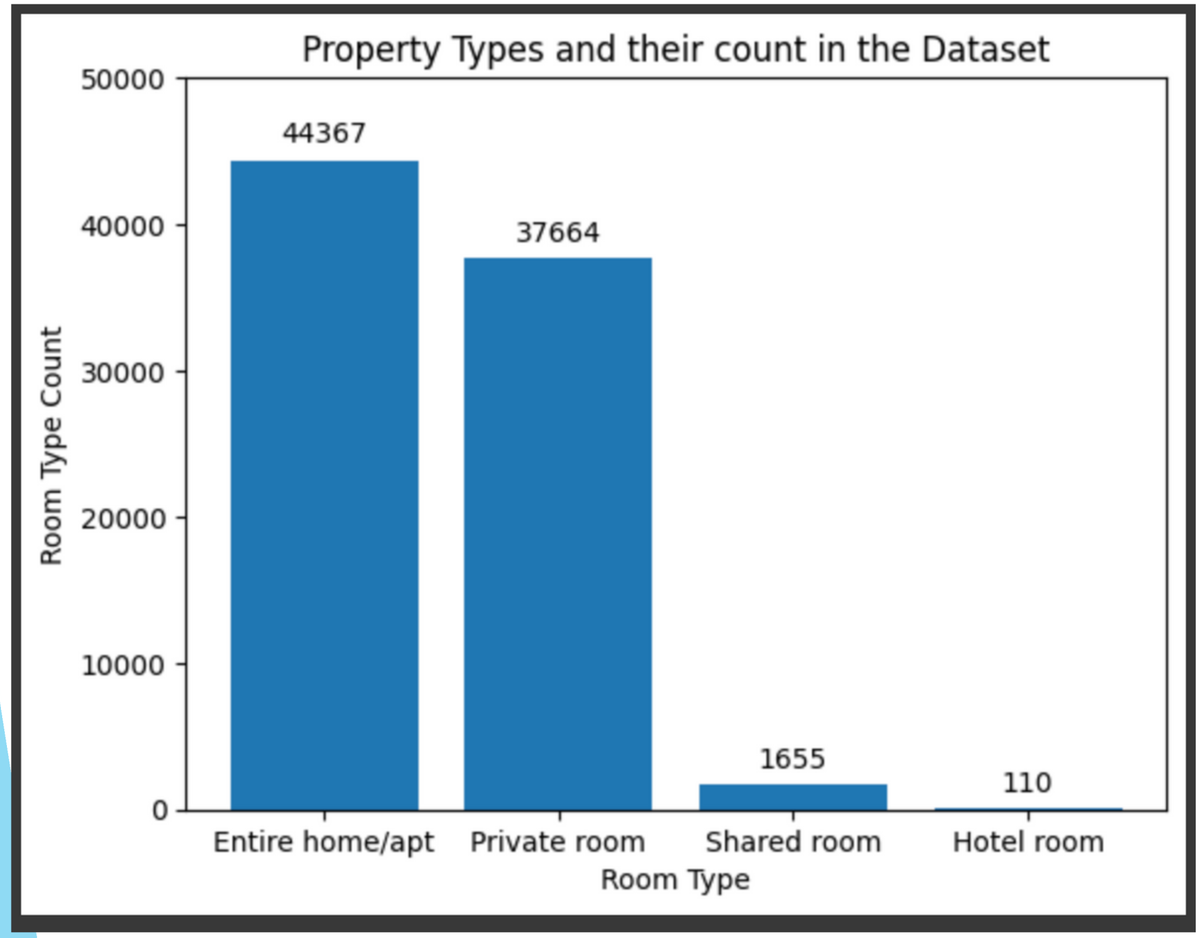
	id	host id	lat	long	instant_bookable	Construction year	price	service fee	minimum nights	number of reviews	last
count	1.025990e+05	1.025990e+05	102591.000000	102591.000000	102494.000000	102385.000000	102352.000000	102326.000000	102190.000000	102416.000000	
mean	2.914623e+07	4.925411e+10	40.728094	-73.949644	0.497785	2012.487464	625.293536	125.026924	8.135845	27.483743	2015 03:40:52.065
min	1.001254e+06	1.236005e+08	40.499790	-74.249840	0.000000	2003.000000	50.000000	10.000000	-1223.000000	0.000000	201 0
25%	1.508581e+07	2.458333e+10	40.688740	-73.982580	0.000000	2007.000000	340.000000	68.000000	2.000000	1.000000	201 0
50%	2.913660e+07	4.911774e+10	40.722290	-73.954440	0.000000	2012.000000	624.000000	125.000000	3.000000	7.000000	201 0
75%	4.320120e+07	7.399650e+10	40.762760	-73.932350	1.000000	2017.000000	913.000000	183.000000	5.000000	30.000000	201 0
max	5.736742e+07	9.876313e+10	40.916970	-73.705220	1.000000	2022.000000	1200.000000	240.000000	5645.000000	1024.000000	205 0
std	1.625751e+07	2.853900e+10	0.055857	0.049521	0.499998	5.765556	331.671614	66.325739	30.553781	49.508954	

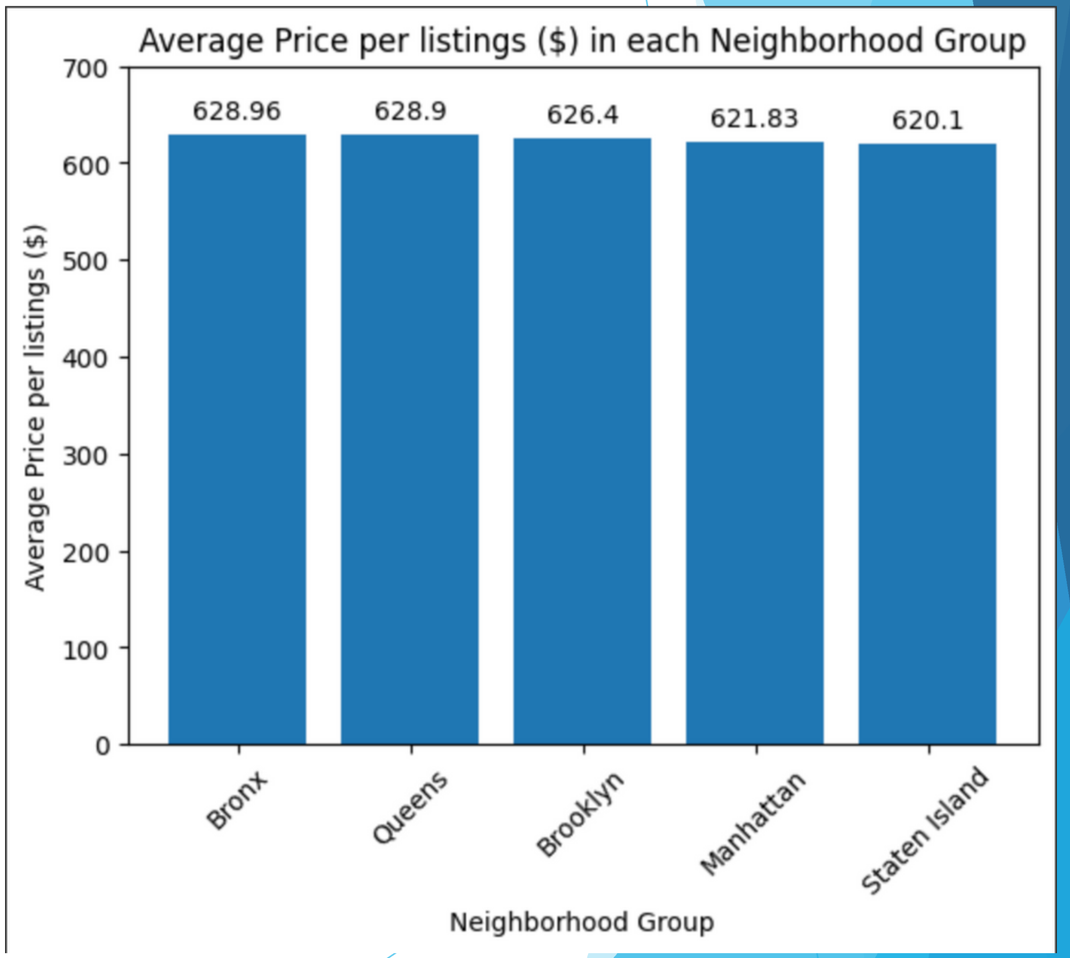
RESULTS2


air_bnb.ipynb
File Edit View Insert Runtime Tools Help
Share
Commands + Code + Text Run all
Reconnect T4
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 102599 entries, 0 to 102598
Data columns (total 26 columns):
Column Non-Null Count Dtype
0 id 102599 non-null int64
1 NAME 102329 non-null object
2 host id 102599 non-null int64
3 host_identity_verified 102310 non-null object
4 host name 102191 non-null object
5 neighbourhood group 102570 non-null object
6 neighbourhood 102583 non-null object
7 lat 102591 non-null float64
8 long 102591 non-null float64
9 country 102067 non-null object
10 country code 102468 non-null object
11 instant_bookable 102494 non-null float64
12 cancellation_policy 102523 non-null object
13 room type 102599 non-null object
14 Construction year 102385 non-null float64
15 price 102352 non-null float64
16 service fee 102326 non-null float64
17 minimum nights 102190 non-null float64
18 number of reviews 102416 non-null float64
19 last review 86706 non-null datetime64[ns]
20 reviews per month 86720 non-null float64
21 review rate number 102273 non-null float64
22 calculated host listings count 102280 non-null float64
23 availability 365 102151 non-null float64
24 house_rules 47756 non-null object
25 license 2 non-null object
dtypes: datetime64[ns](1), float64(12), int64(2), object(11)
memory usage: 20.4+ MB
Variables Terminal
5:18 PM

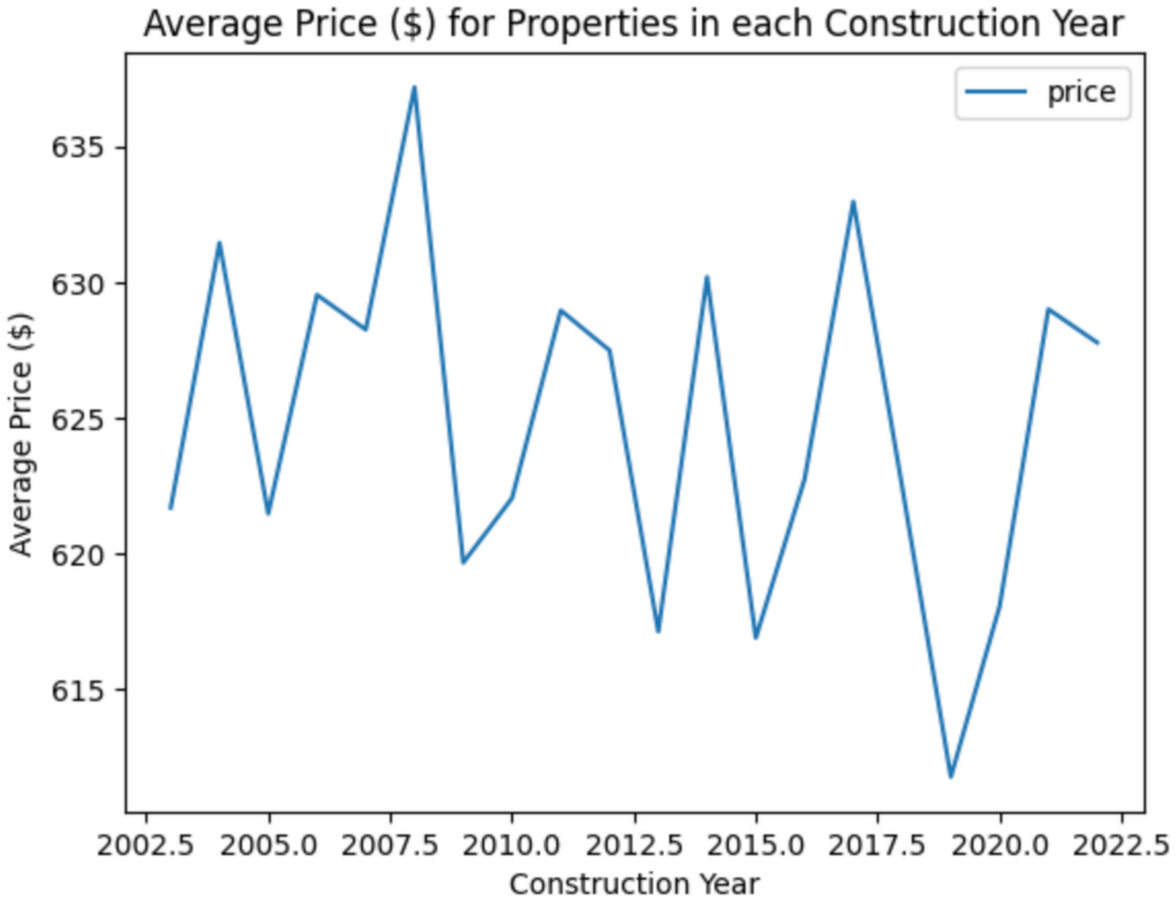

air_bnb.ipynb
File Edit View Insert Runtime Tools Help
Share
Commands + Code + Text Run all
Reconnect T4
4. Handle Outliers in 'availability 365'
Let's identify and remove outliers in the 'availability 365' column using the IQR method.
Q1 = df['availability 365'].quantile(0.25)
Q3 = df['availability 365'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
outliers_count = df[(df['availability 365'] < lower_bound) | (df['availability 365'] > upper_bound)].shape[0]
print(f"Number of outliers in 'availability 365': {outliers_count}")
df_cleaned = df[(df['availability 365'] >= lower_bound) & (df['availability 365'] <= upper_bound)].copy()
print(f"Number of rows after removing outliers: {len(df_cleaned)}")
Number of outliers in 'availability 365': 0
Number of rows after removing outliers: 83796
Correct Spelling of 'brookln'
Let's correct the spelling of 'brookln' to 'Brooklyn' in the 'neighbourhood group' column.
if 'brookln' in df['neighbourhood group'].unique():
df['neighbourhood group'] = df['neighbourhood group'].replace('brookln', 'Brooklyn')
print("Corrected 'brookln' to 'Brooklyn' in 'neighbourhood group' column.")
else:
print("'brookln' not found in 'neighbourhood group' column. No correction needed.")
# Verify the changes by checking unique values in 'neighbourhood group'
display(df['neighbourhood group'].unique())
Corrected 'brookln' to 'Brooklyn' in 'neighbourhood group' column.
array(['Brooklyn', 'Manhattan', 'Queens', 'Bronx', 'Staten Island'],
dtype=object)
df.info()
<class 'pandas.core.frame.DataFrame'>
Index: 83796 entries, 0 to 102057
Variables Terminal
5:18 PM

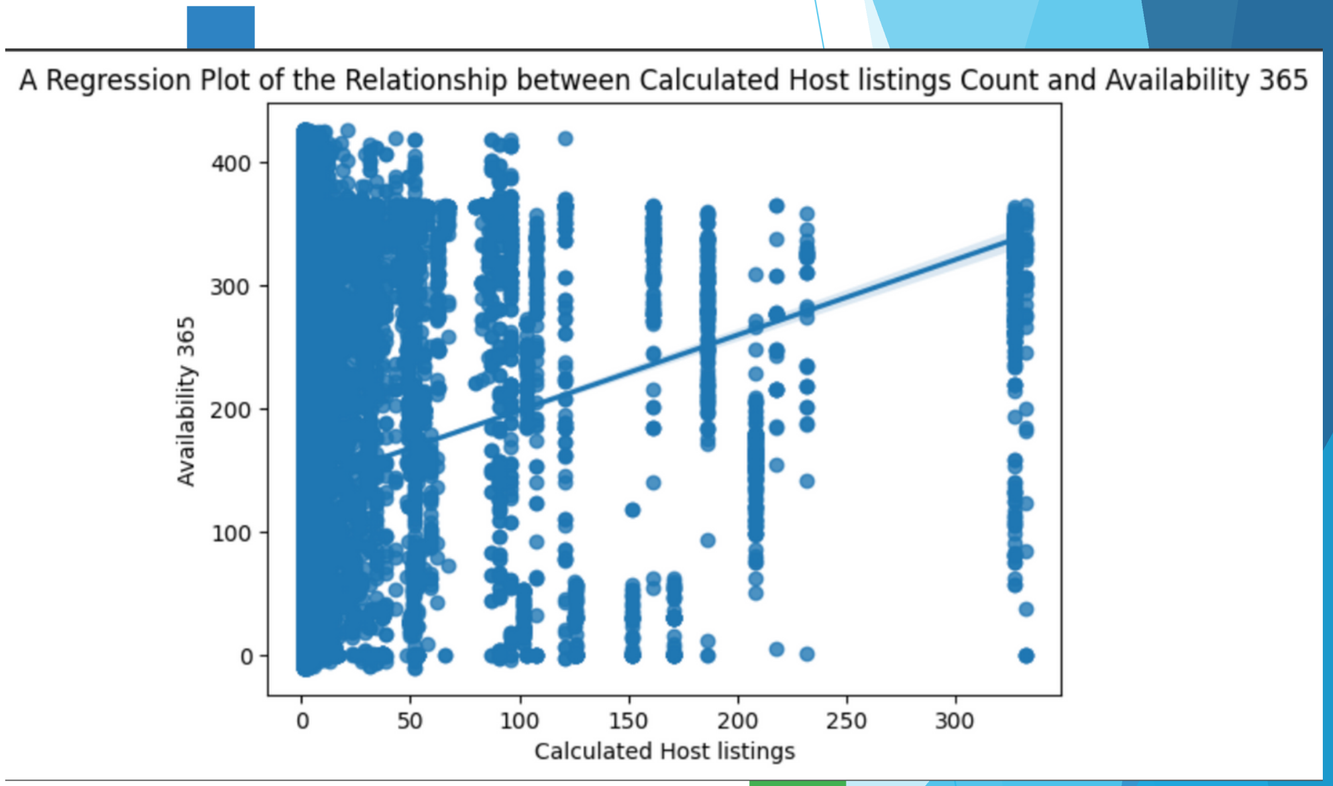
RESULTS3


Property Types and their count in the Dataset
Room Type Count
50000
40000
30000
20000
10000
0
44367
37664
1655
110
Entire home/apt
Private room
Shared room
Hotel room
Room Type

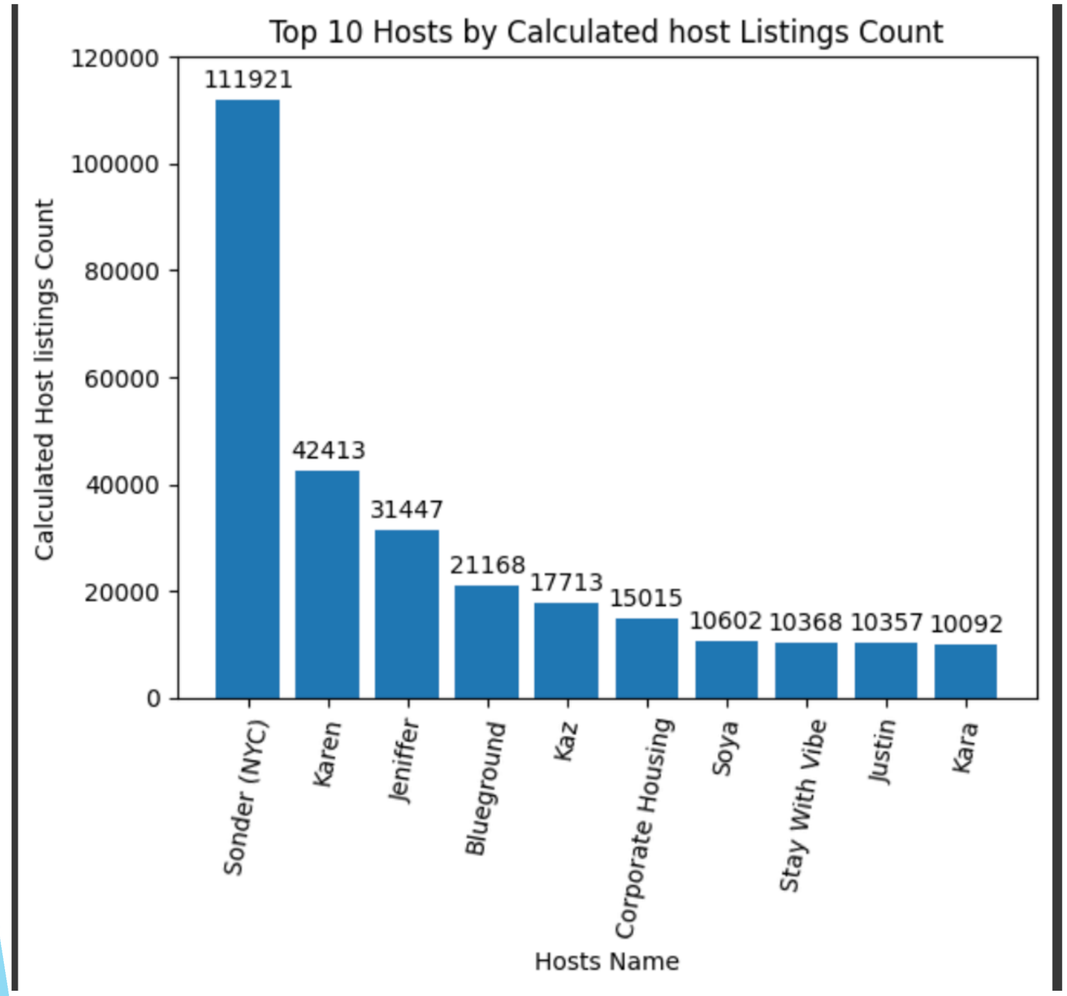

Average Price per listings ($) in each Neighborhood Group
Average Price per listings ($)
700
600
500
400
300
200
100
0
628.96
628.9
626.4
621.83
620.1
Bronx
Queens
Brooklyn
Manhattan
Staten Island
Neighborhood Group

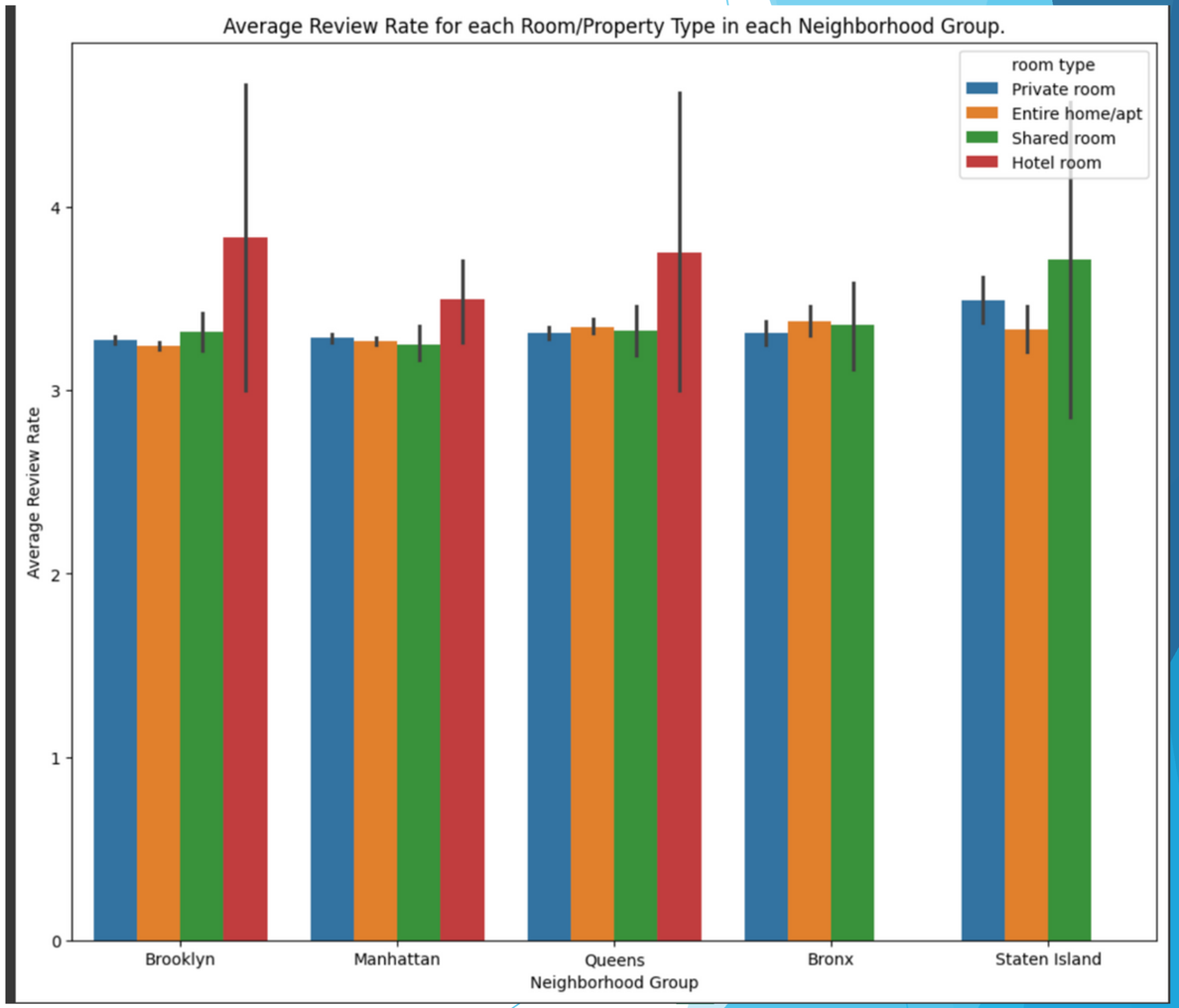
RESULTS4


Average Price ($) for Properties in each Construction Year
price
Average Price ($)
635
630
625
620
615
2002.5
2005.0
2007.5
2010.0
2012.5
2015.0
2017.5
2020.0
2022.5
Construction Year


A Regression Plot of the Relationship between Calculated Host listings Count and Availability 365
Availability 365
400
300
200
100
0
0
50
100
150
200
250
300
Calculated Host listings

RESULTS4


Top 10 Hosts by Calculated host Listings Count
Calculated Host listings Count
120000
100000
80000
60000
40000
20000
0
111921
42413
31447
21168
17713
15015
10602
10368
10357
10092
Sonder (NYC)
Karen
Jeniffer
Blueground
Kaz
Corporate Housing
Soya
Stay With Vibe
Justin
Kara
Hosts Name

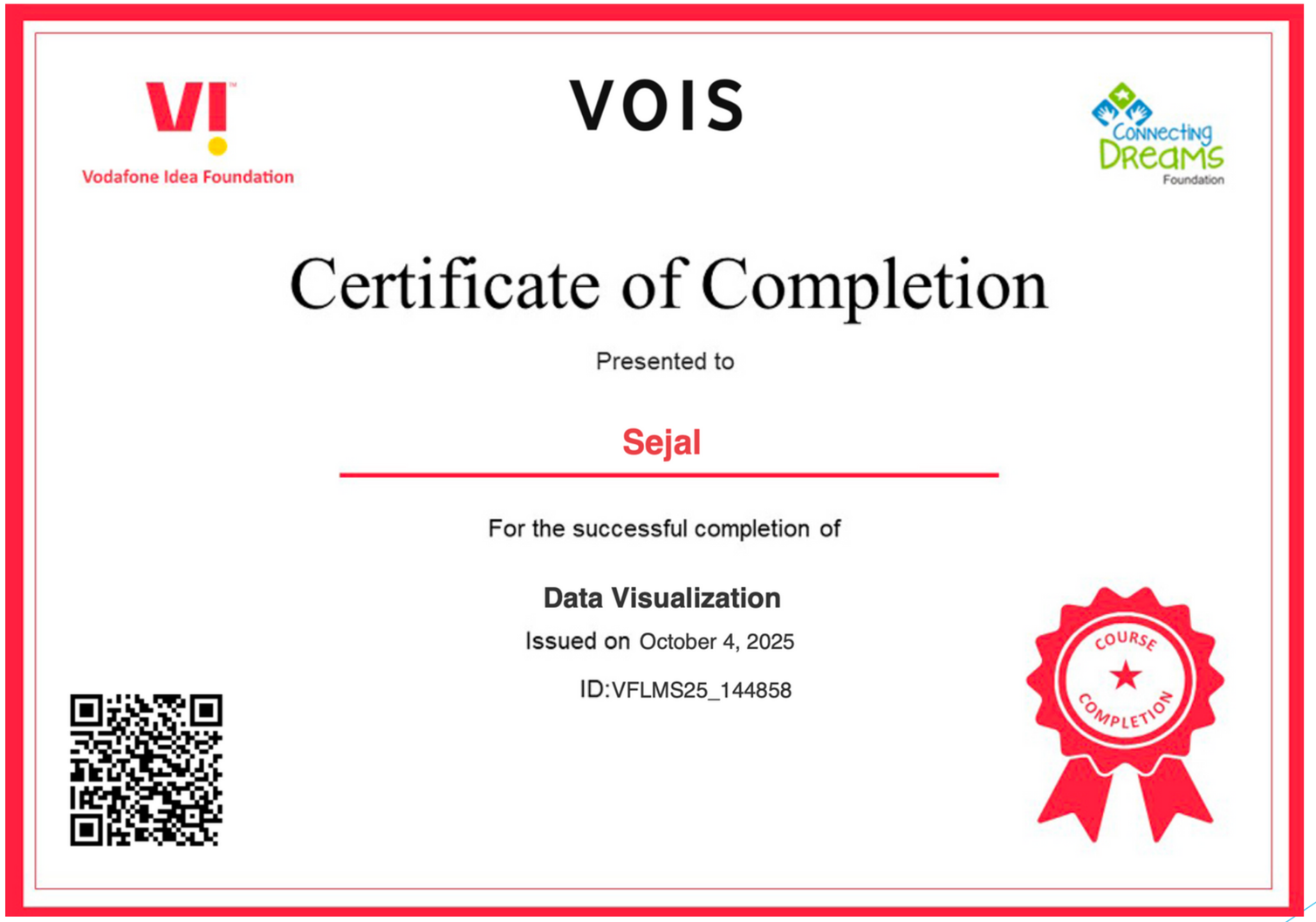

Average Review Rate for each Room/Property Type in each Neighborhood Group.
room type
Private room
Entire home/apt
Shared room
Hotel room
Average Review Rate
4
3
2
1
0
Brooklyn
Manhattan
Queens
Bronx
Staten Island
Neighborhood Group

GitHub repository

https://github.com/sejal180602/VOIS_AICTE_Oct2025_SEJAL.git

Getting started with Basics of Python Certificate

Data Visualization Certificate




Thank
you