The Airbnb Revenue Model:

Analyzing KPI Accomplishments

CONTRIBUTORS

- Siddharth Susarla
- Tapan Pradyot
- Shahid Shakil
- Hailey Thanki
- Sajid Hussain Rafi Ahamed

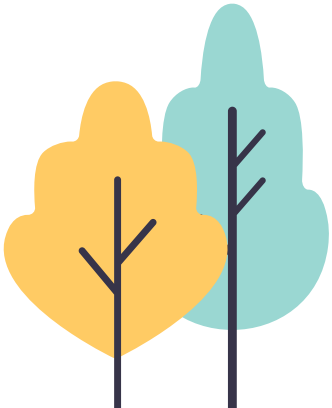
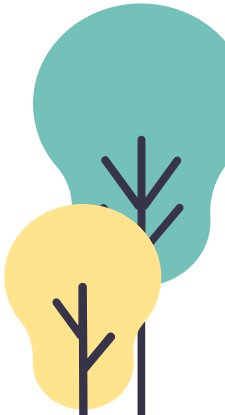


TABLE OF CONTENTS



MOTIVATION

- To study the Airbnb revenue model and key performance indicators regionally.
- Over 150 million worldwide users have booked over one billion stays.
- The hotel industry loses approximately $450 million in direct revenue per year to AirBnb.
- 400,000 companies directly engage with Airbnb to manage travel for their employees.
- Hosts have collectively earned over $110 billion.
- 6 guests check into an Airbnb listing every second

GOAL

- **Our goal is to predict the availability of these lodgings using machine learning and also identify key performance indicators which drive revenue.**

- **Our study aims to help the individuals in the hospitality business by predicting accommodation availability during the renters target timeframe.**

BUSINESS PROBLEMS BEING ADDRESSED

- **Availability prediction to direct customers to AirBNB**
- **Which attributes help to make an AirBNB listing stand out?**
- **Which Time Frame is more popular?**
- **Which guests are more likely to book an AirBNB?**
- **Formulating a strategy using our findings to increase earnings.**

EXPECTED RESULTS

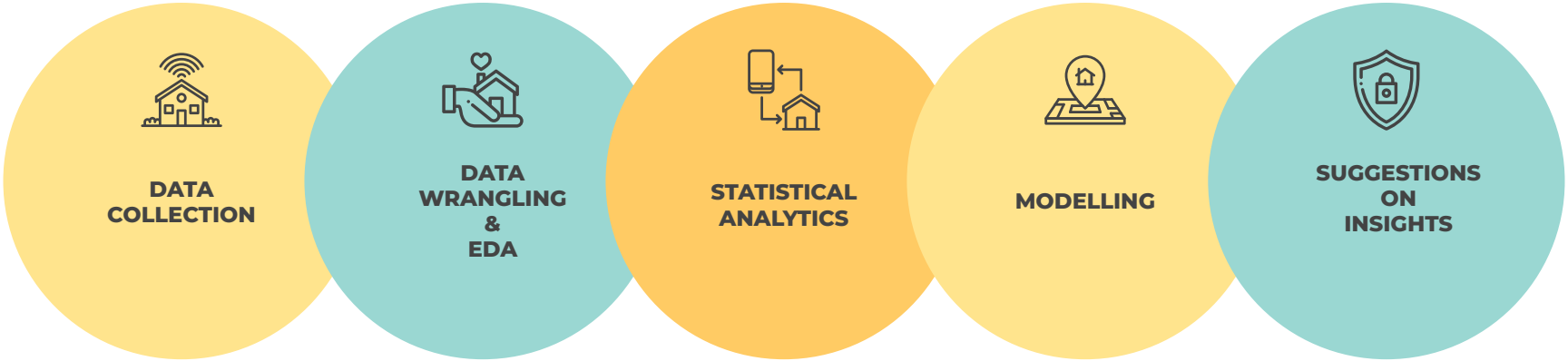
Provide robust approaches to estimate occupancy.

Highlight important factors and predictors driving growth.

Demonstrate which data is unreliable or is a red herring hindering good recommendations

Showcase key trends such as most profitable months, popular types of accommodation units and other volume driven variables.

PROJECT LIFECYCLE



DATA

- **The dataset has daily availability and pricing for AirBNB listings in the Phoenix market from 4/1/18 to 5/31/18.**
- **The two data tables provided as flat CSV files are structured as craped_listings.csv and scraped_data.csv.**
- **There are 1,630,0175 observations in the dataset.**

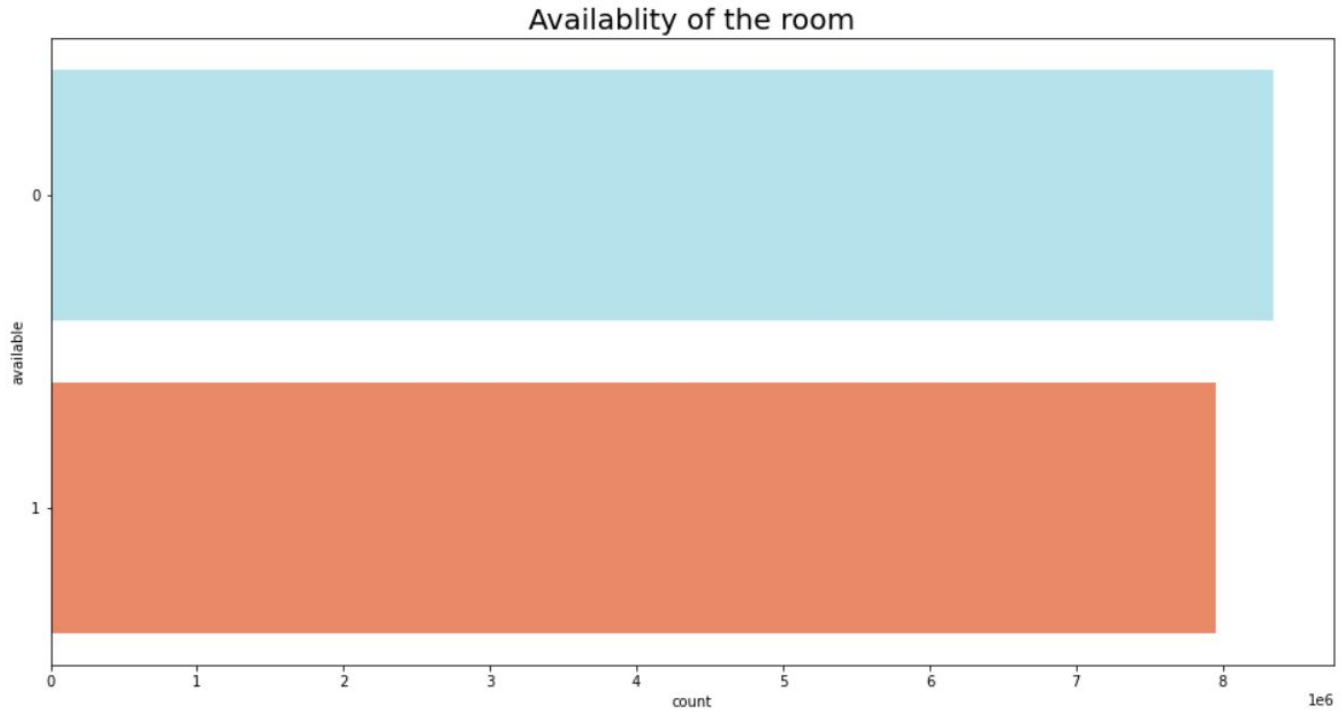
FEATURES

Feature	Description
scraping_id	ID Key
listing	URL Link to Airbnb posting
city	Name of City within Phoenix market
lon	Longitude of Unit Rented
lat	Latitude of Unit Rented
mapped_location	A Google Maps URL of location
name	Posting Name
capacity	Number of People it can accomodate
bedrooms	Number of bedrooms at unit
bathrooms	Number of bathrooms at unit
Has_pool	1 if unit has pool listed; 0 if does not
Cleaning_fee	The amount in dollars to cover cleaning
Is_superhost	1 if the host is a superhost; 0 if not
Hostname	Name of Host

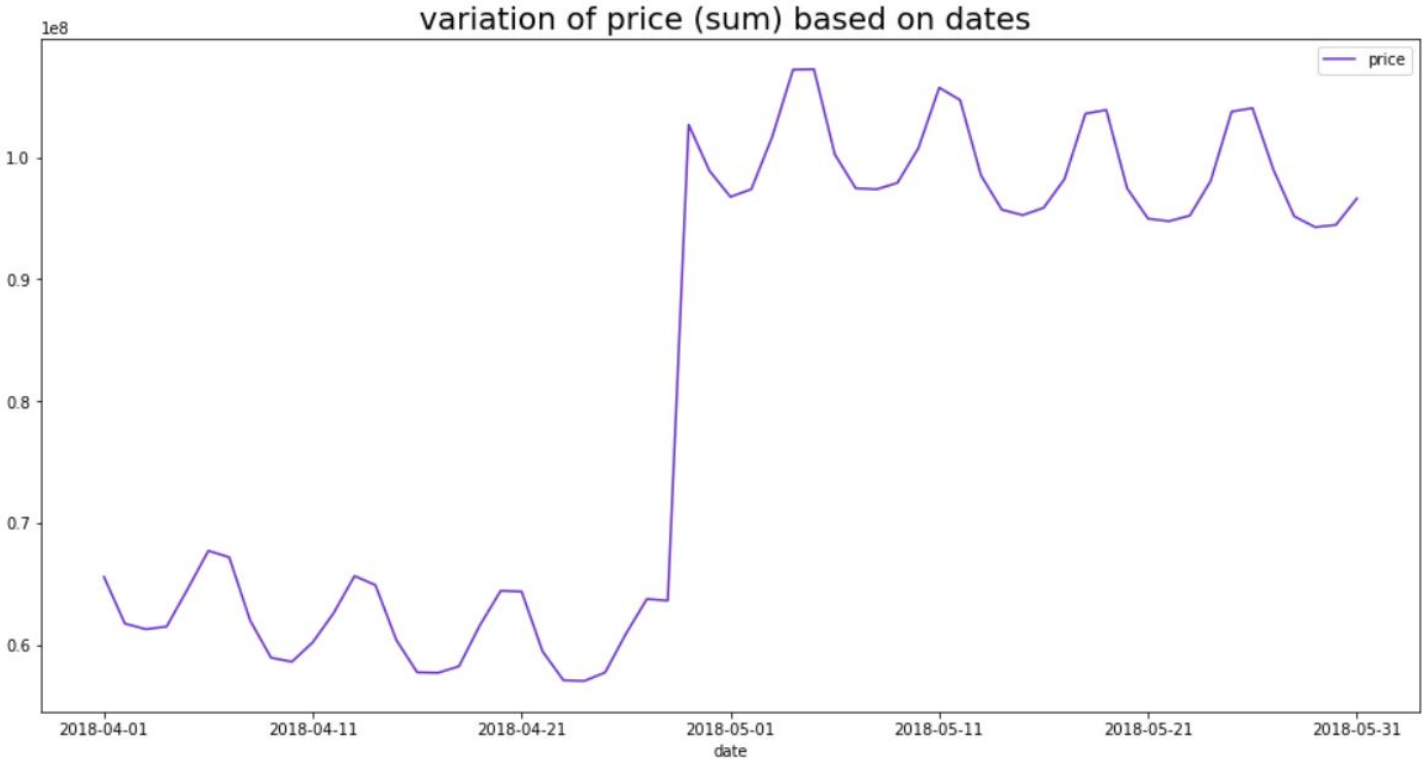
Feature	Description
scraping_id	ID key
as_of_date	Date the information was scraped
date	Date of the night to be booked
price	Price in dollars of the night
available	Availability

EXPLORATORY DATA ANALYSIS

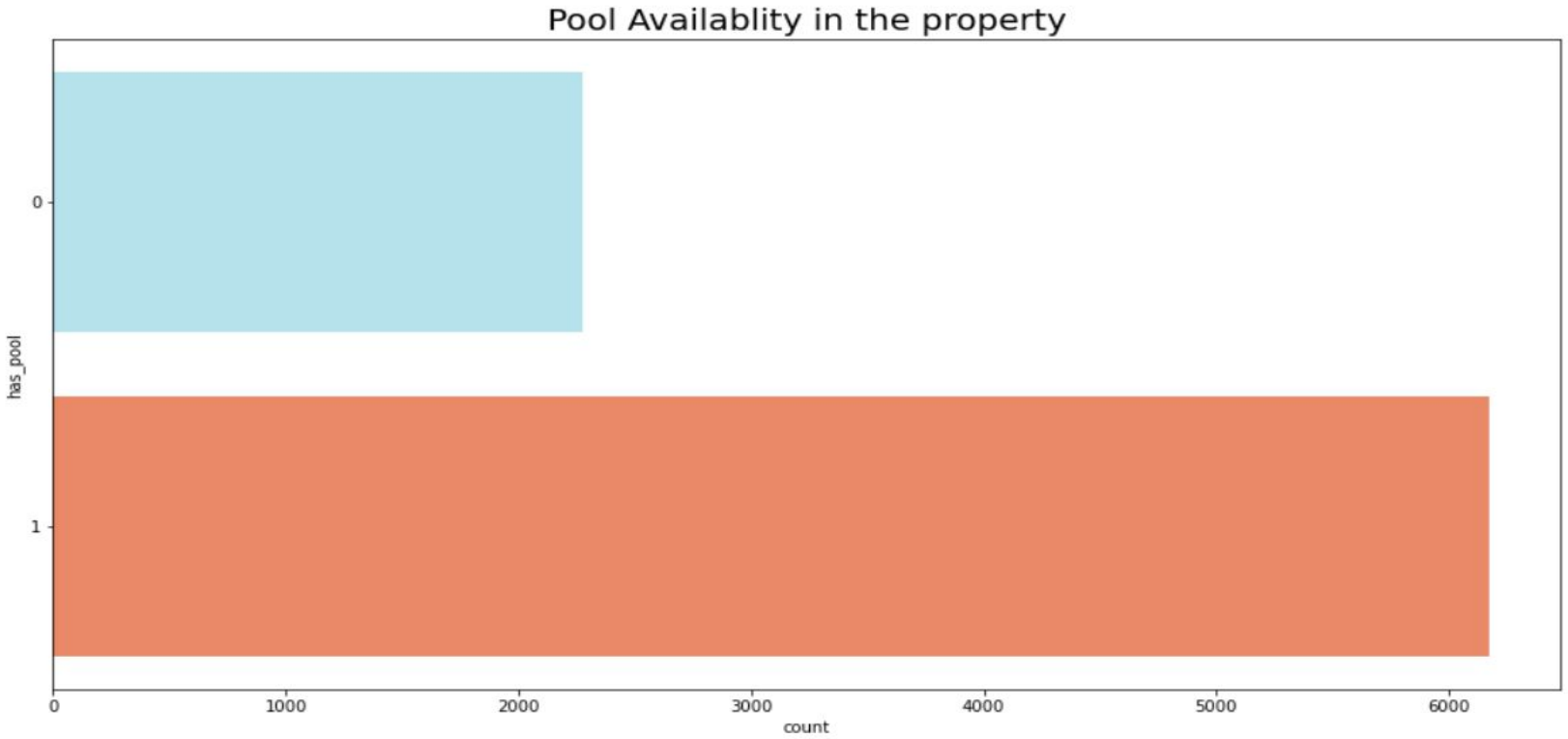
DISTRIBUTION OF TARGET VARIABLE



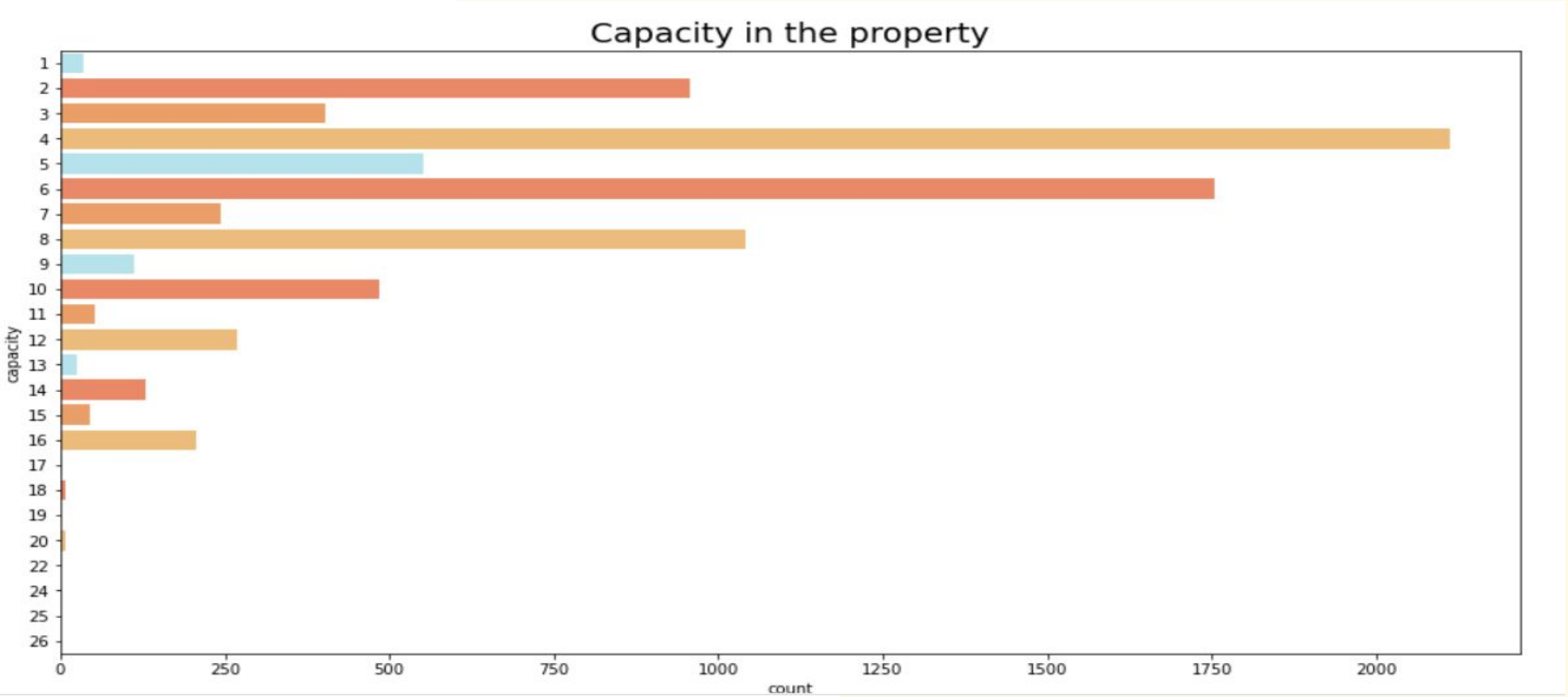
VARIATION OF PRICE vs TIME



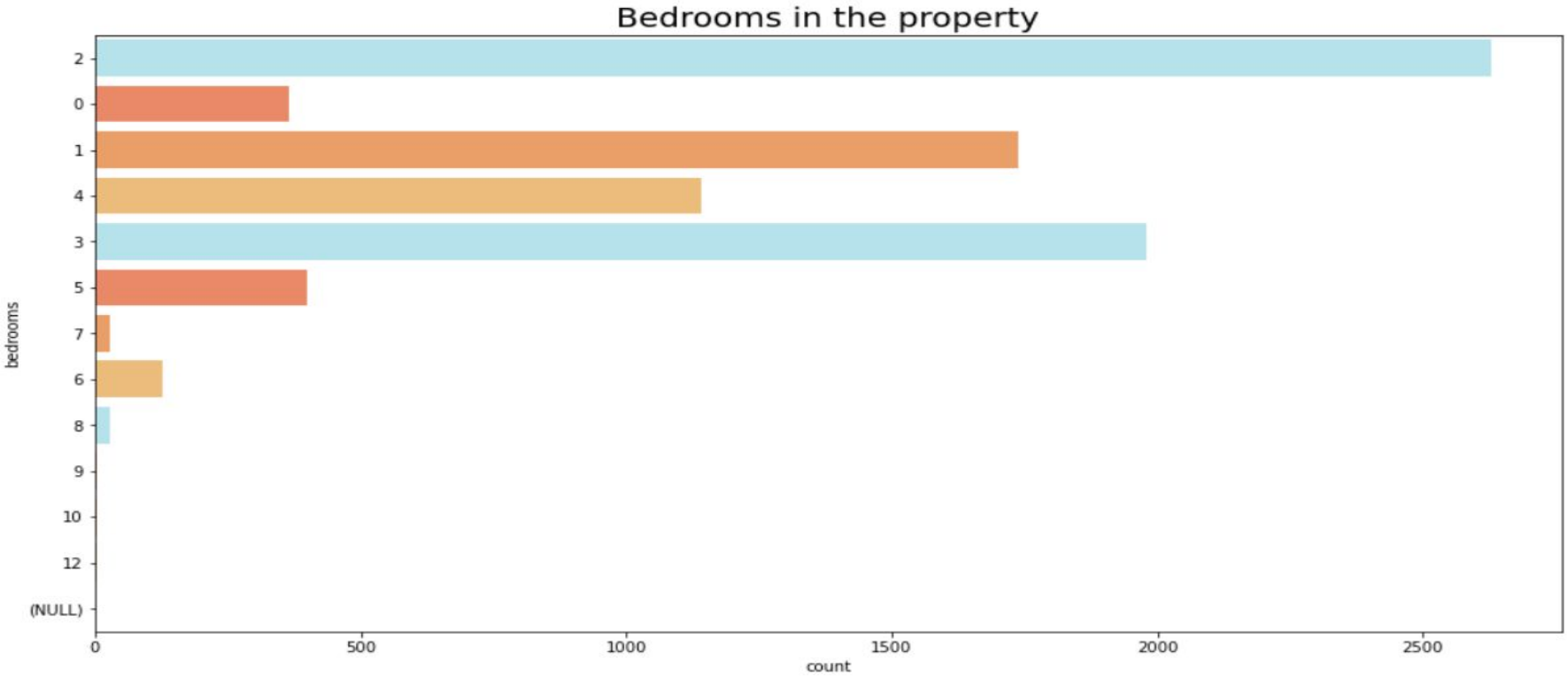
POOL AVAILABILITY IN PROPERTY



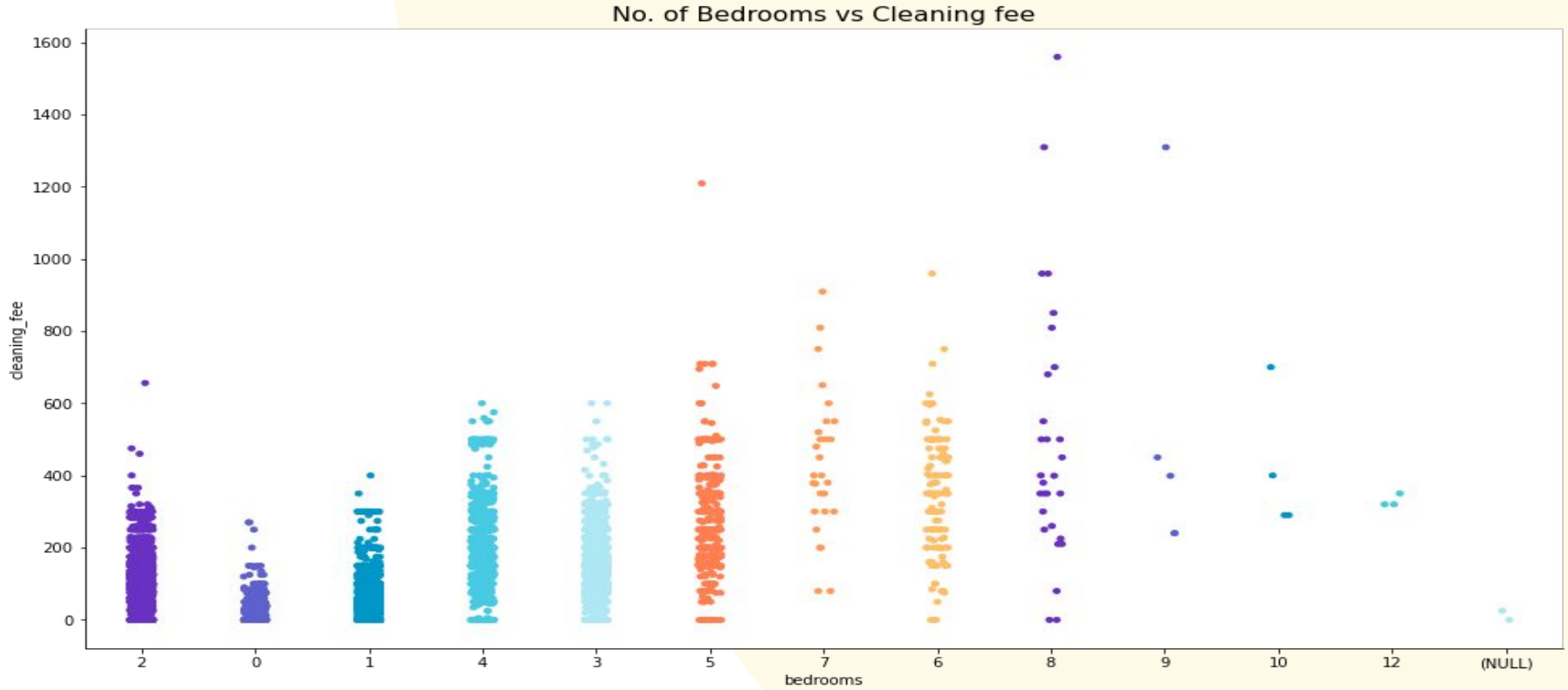
CAPACITY IN THE PROPERTY



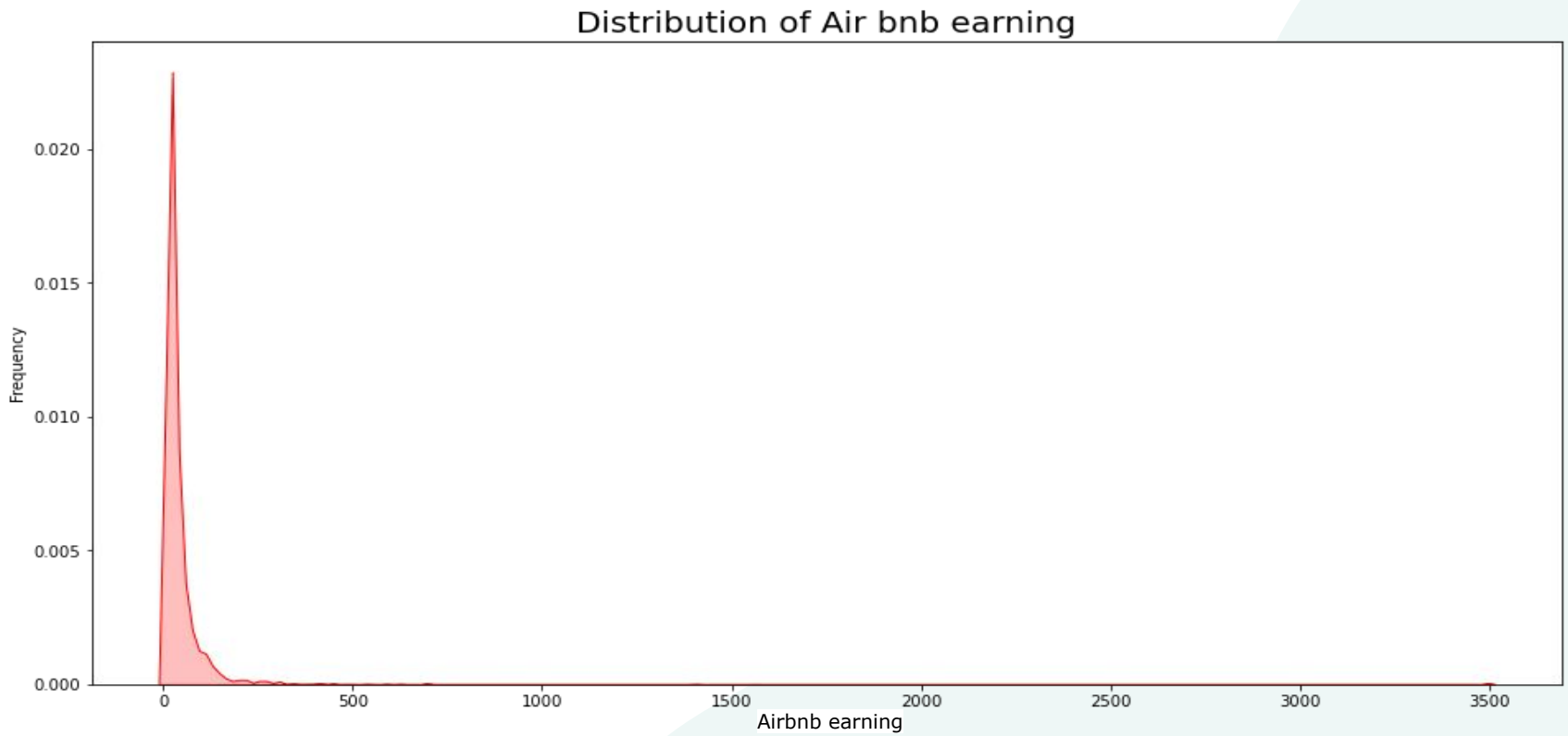
BEDROOMS IN THE PROPERTY



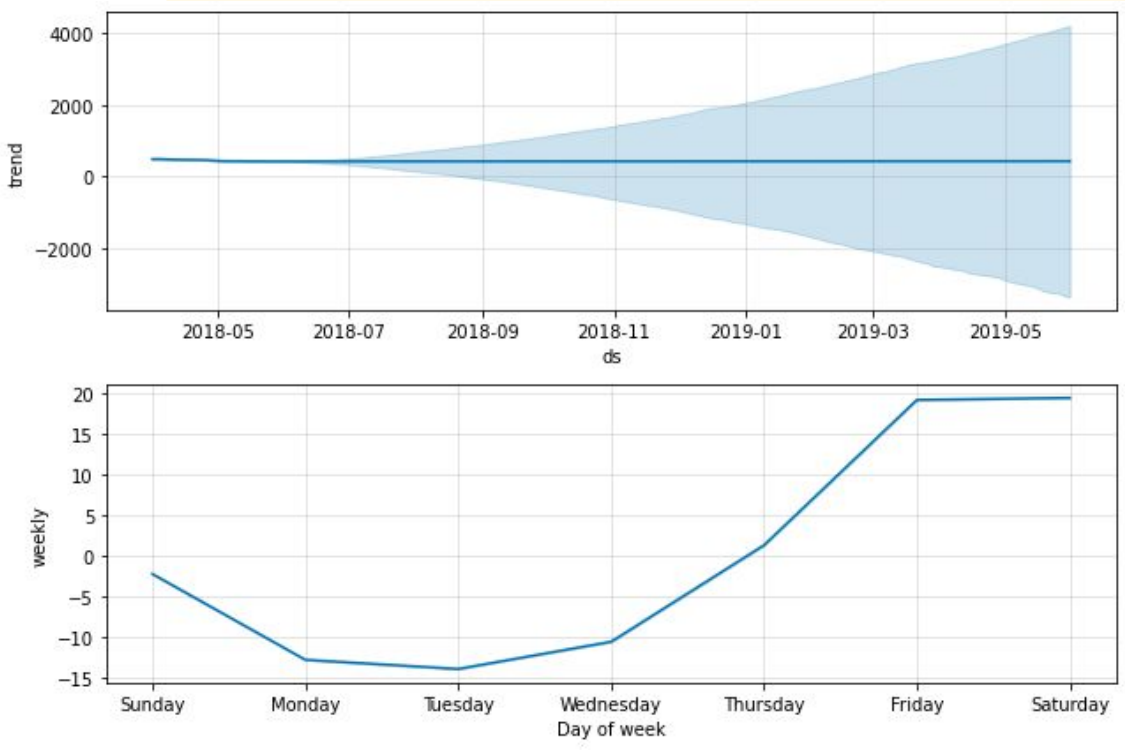
Number OF BEDROOMS vs CLEANING FEES



Distribution of AirBnB earning

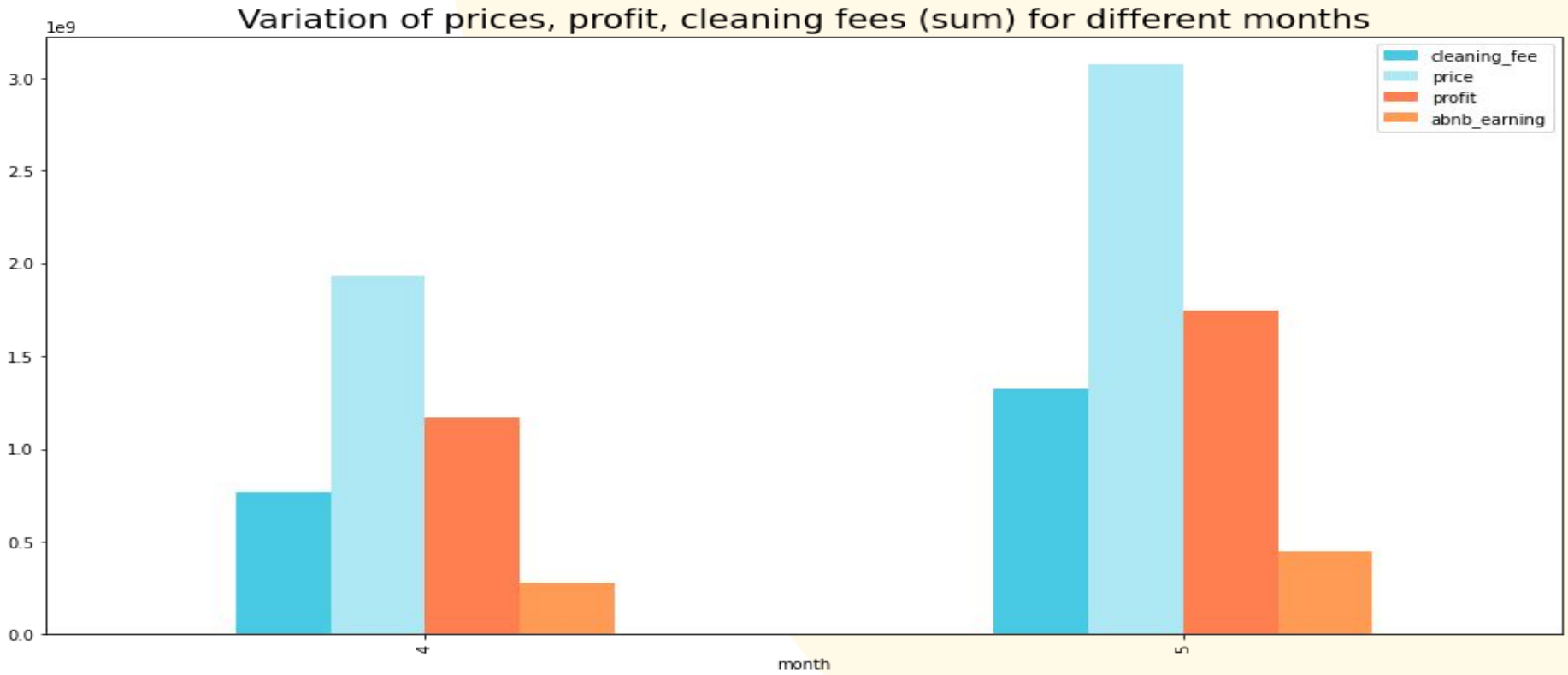


Variation of prices based on dates

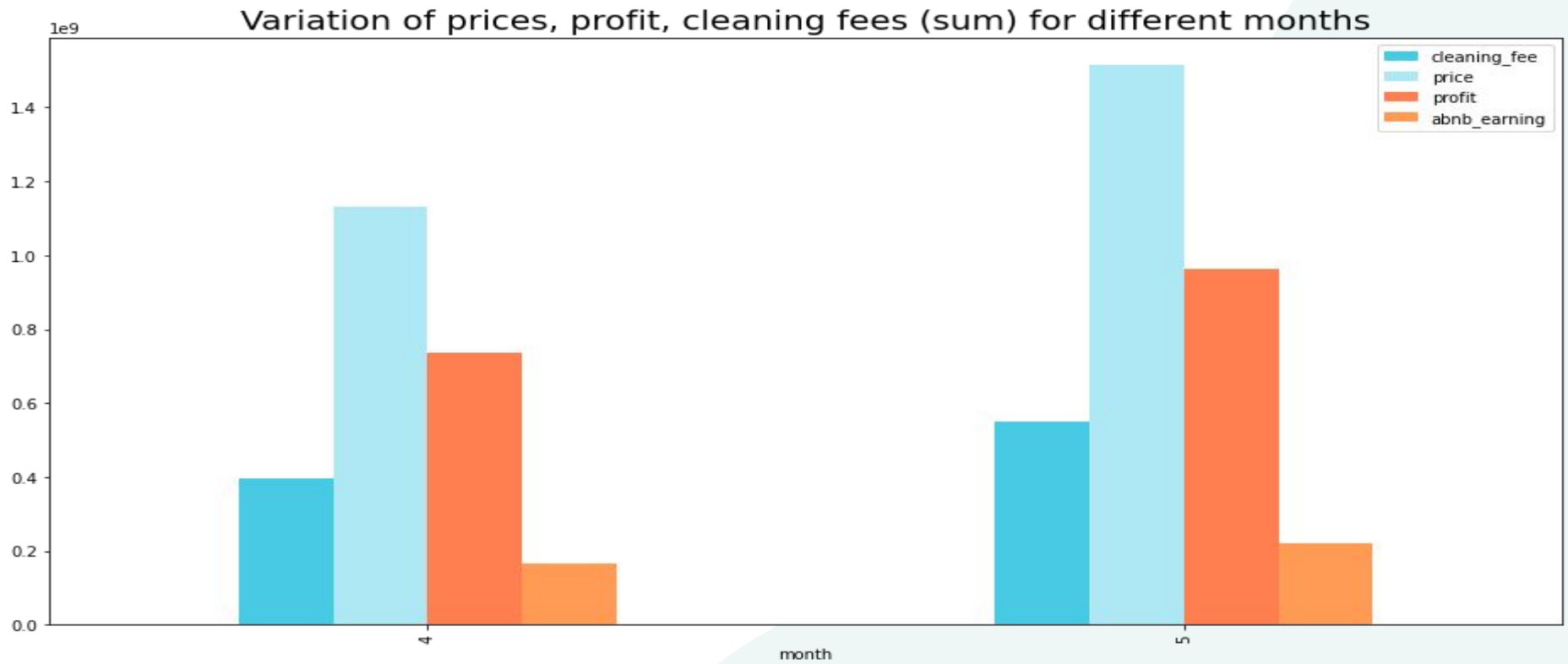


FINDINGS FROM EDA

WHICH MONTH DO PROPERTIES APPEAR TO GENERATE MORE REVENUE, APRIL? OR MAY?

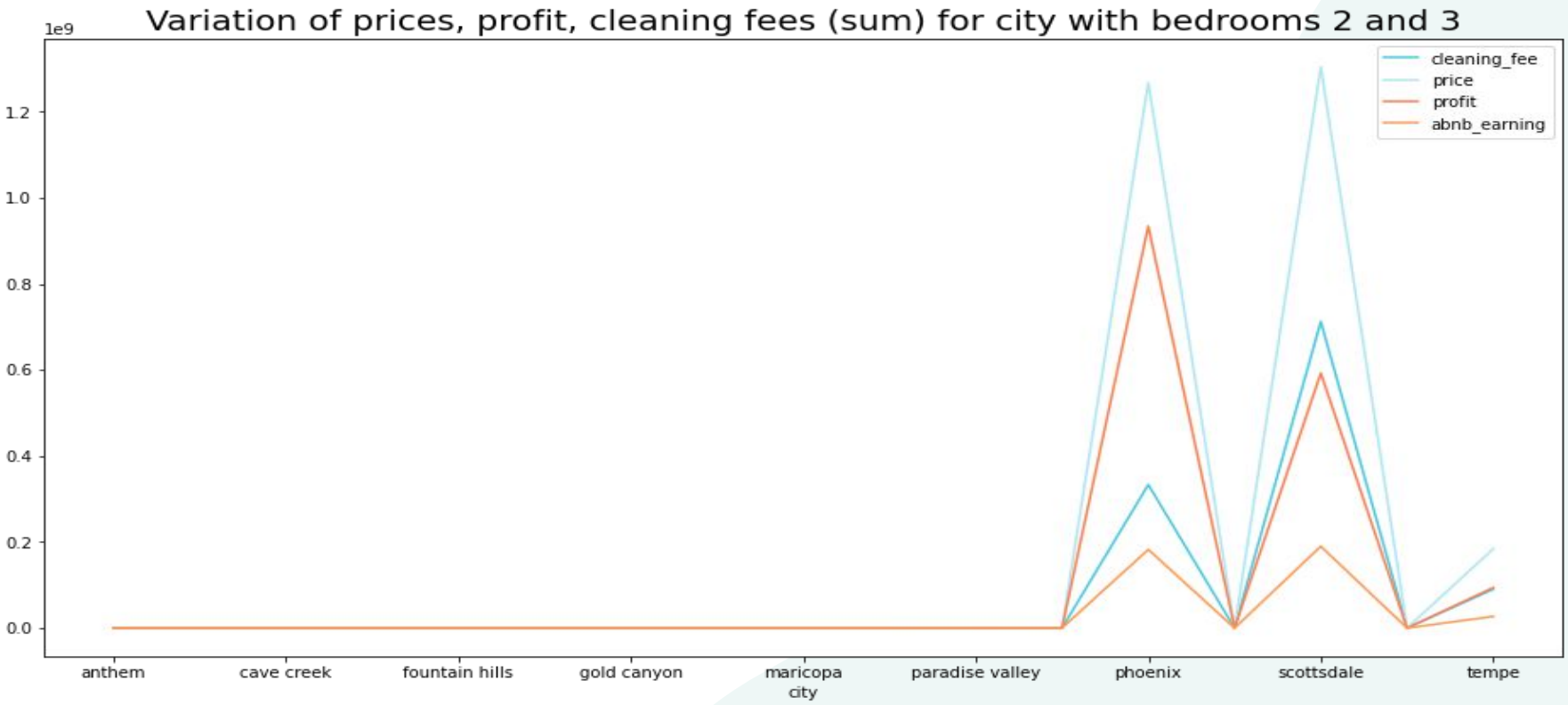


WHICH MONTH DO PROPERTIES POTENTIALLY GENERATE MORE REVENUE, CONSIDERING LISTINGS

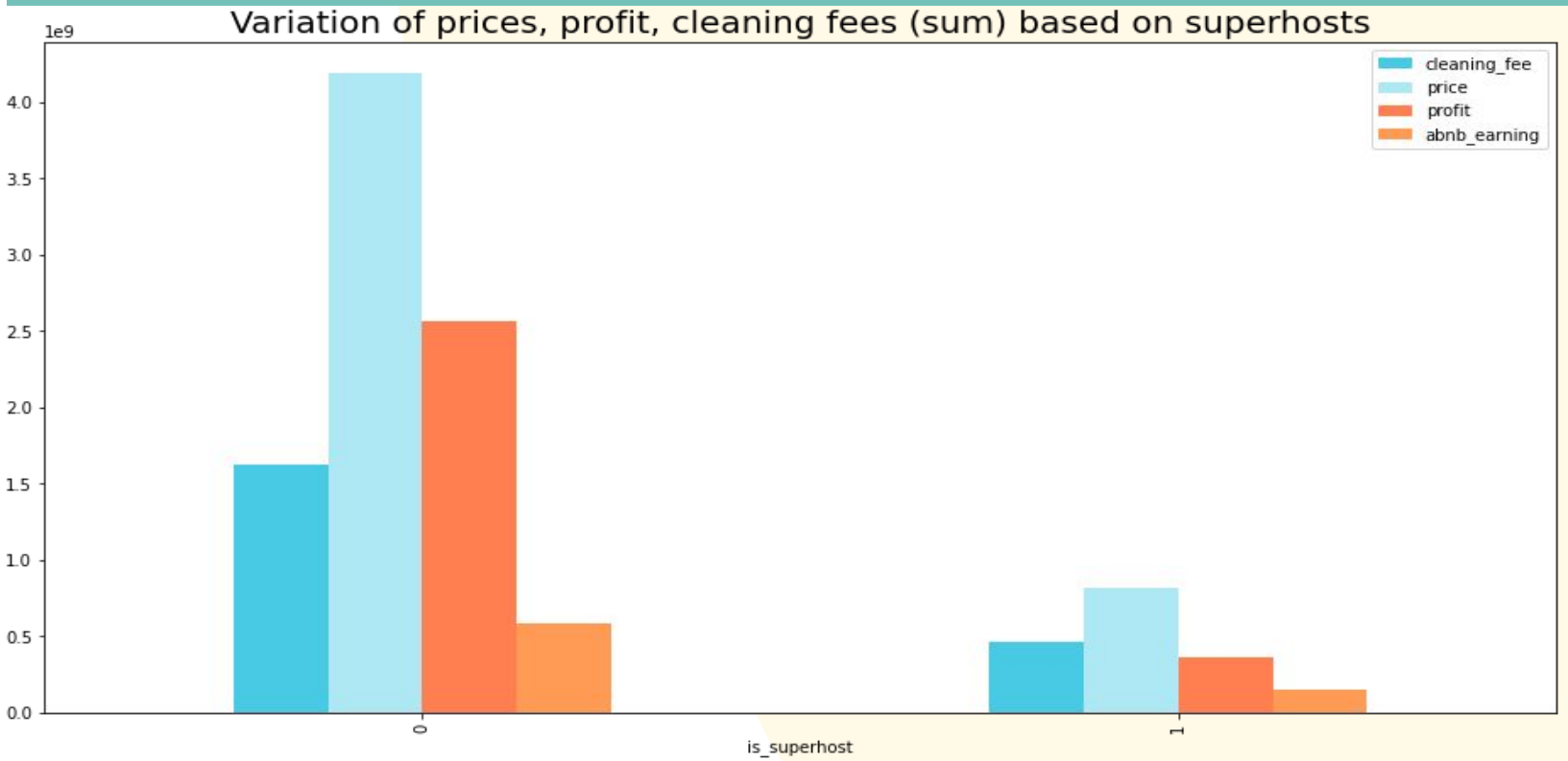


HOW MUCH REVENUE DO PLACES WITH 3 BEDROOMS MAKE VS. PLACES WITH 2 BEDROOMS?

WHAT ARE THE MOST VALUABLE CITIES FOR 3-BEDROOM OR 2-BEDROOM APARTMENTS?



How much of a difference does being a super host, on average, have on the price of a listing?



STATISTICAL TESTS

CHI SQUARE TEST

Price Category vs Availability

	available		
available	0	1	All
price_cat			
Cheap	3.87	1.62	5.49
Mid-Range	44.07	45.15	89.22
Expensive	2.78	1.75	4.53
Very Expensive	0.45	0.30	0.76
All	51.18	48.82	100.00

	Chi-square test	results
0	Pearson Chi-square (3.0) =	185617.6413
1	p-value =	0.0000
2	Cramer's V =	0.1070

The Cramer's V value is 0.1006. This indicates a moderate correlation between the price range of the Airbnb and their availability.

CHI SQUARE TEST

Capacity Category vs Availability

	available		
available	0	1	All
capacity_cat			
1-2 guests	1.96	1.16	3.11
3-6 guests	33.00	30.02	63.02
More than 6 guests	16.23	17.63	33.86
All	51.19	48.81	100.00

	Chi-square test	results
0	Pearson Chi-square (2.0) =	56525.4237
1	p-value =	0.0000
2	Cramer's V =	0.0589

The Cramer's V value is 0.0589. This indicates a weak correlation between the capacity of the Airbnb and their availability.

CHI SQUARE TEST

Number of Bathroom Category vs Availability

	available		
available	0	1	All
bathrooms_cat			
1 bathroom	12.67	9.80	22.47
2 bathrooms	26.81	25.86	52.67
3 bathrooms	8.90	9.18	18.08
More than 3 bathrooms	2.82	3.96	6.78
All	51.19	48.81	100.00

	Chi-square test	results
0	Pearson Chi-square (3.0) =	85719.9232
1	p-value =	0.0000
2	Cramer's V =	0.0725

The Cramer's V value is 0.0725. This indicates a weak correlation between the number of bathrooms in the AirBnBs and their availability.

CHI SQUARE TEST

Number of Bedroom Category vs Availability

	available		
available	0	1	All
bedrooms_cat			
1 bedroom	8.46	6.65	15.12
2 bedrooms	19.75	18.53	38.27
3 bedrooms	12.92	12.57	25.49
4 bedrooms	6.83	7.43	14.25
5-8 bedrooms	2.97	3.68	6.66
More than 8 bedrooms	0.03	0.18	0.22
All	50.96	49.04	100.00

	Chi-square test	results
0	Pearson Chi-square (5.0) =	68249.1068
1	p-value =	0.0000
2	Cramer's V =	0.0654

The Cramer's V value is 0.0654. This indicates a weak correlation between the number of bedrooms in the AirBnBs and their availability.

CHI SQUARE TEST

Pool vs Availability

	available		
available	**0**	**1**	**All**
has_pool			
0	12.08	10.88	22.96
1	39.11	37.93	77.04
All	51.19	48.81	100.00

	Chi-square test	results
0	Pearson Chi-square (1.0) =	3980.5553
1	p-value =	0.0000
2	Cramer's phi =	0.0156

The Cramer's phi value is 0.0156. This indicates there is little to no correlation between their being a pool at the AirBnBs and their availability.

CHI SQUARE TEST

Price Category vs Pool Category

	has_pool		
has_pool	0	1	All
price_cat			
Cheap	2.45	3.04	5.49
Mid-Range	19.72	69.51	89.22
Expensive	0.77	3.76	4.53
Very Expensive	0.09	0.66	0.76
All	23.03	76.97	100.00

	Chi-square test	results
0	Pearson Chi-square (3.0) =	264443.6349
1	p-value =	0.0000
2	Cramer's V =	0.1277

The Cramer's V value is 0.0670. This indicates there is a weak correlation between the price of AirBnBs and there being a pool at the AirBnBs.

ANOVA TEST

Price vs Availability

available	N	Mean	SD	SE	95% Conf.	Interval
available	7955740	439.8585	1052.3477	0.3731	439.1272	440.5897
not available	8344435	430.2618	688.7338	0.2384	429.7945	430.7291

	df	sum_sq	mean_sq	F	PR(>F)
available	1.0	3.750803e+08	3.750803e+08	478.81773	3.877369e-106
Residual	16300173.0	1.276869e+13	7.833467e+05	NaN	NaN

The p value obtained from ANOVA analysis is significant ($p < 0.05$), and therefore, we conclude that there are significant differences among the mean of the prices of available and unavailable AirBnBs.

ANOVA TEST

Capacity vs Availability

available	N	Mean	SD	SE	95% Conf.	Interval
available	7955740	6.7174	3.2850	0.0012	6.7152	6.7197
not available	8344435	6.3264	3.1045	0.0011	6.3243	6.3285

	df	sum_sq	mean_sq	F	PR(>F)
available	1.0	6.226615e+05	622661.517768	61039.628337	0.0
Residual	16300173.0	1.662771e+08	10.200939	NaN	NaN

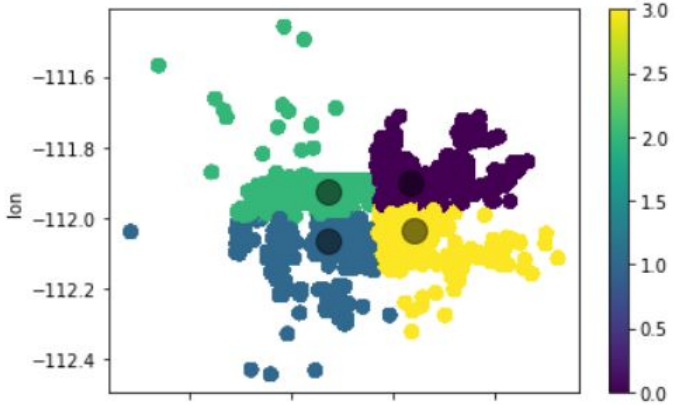
The p value obtained from ANOVA analysis is significant ($p < 0.05$), and therefore, we conclude that there are significant differences among the mean of the capacity of available and unavailable AirBnBs.

PREDICTIVE MODELLING

DATA PRE PROCESSING

Handling Latitude and Longitude

We used K-means Clustering to group the coordinates of the hotel together into 4 Clusters before feeding the data to the model



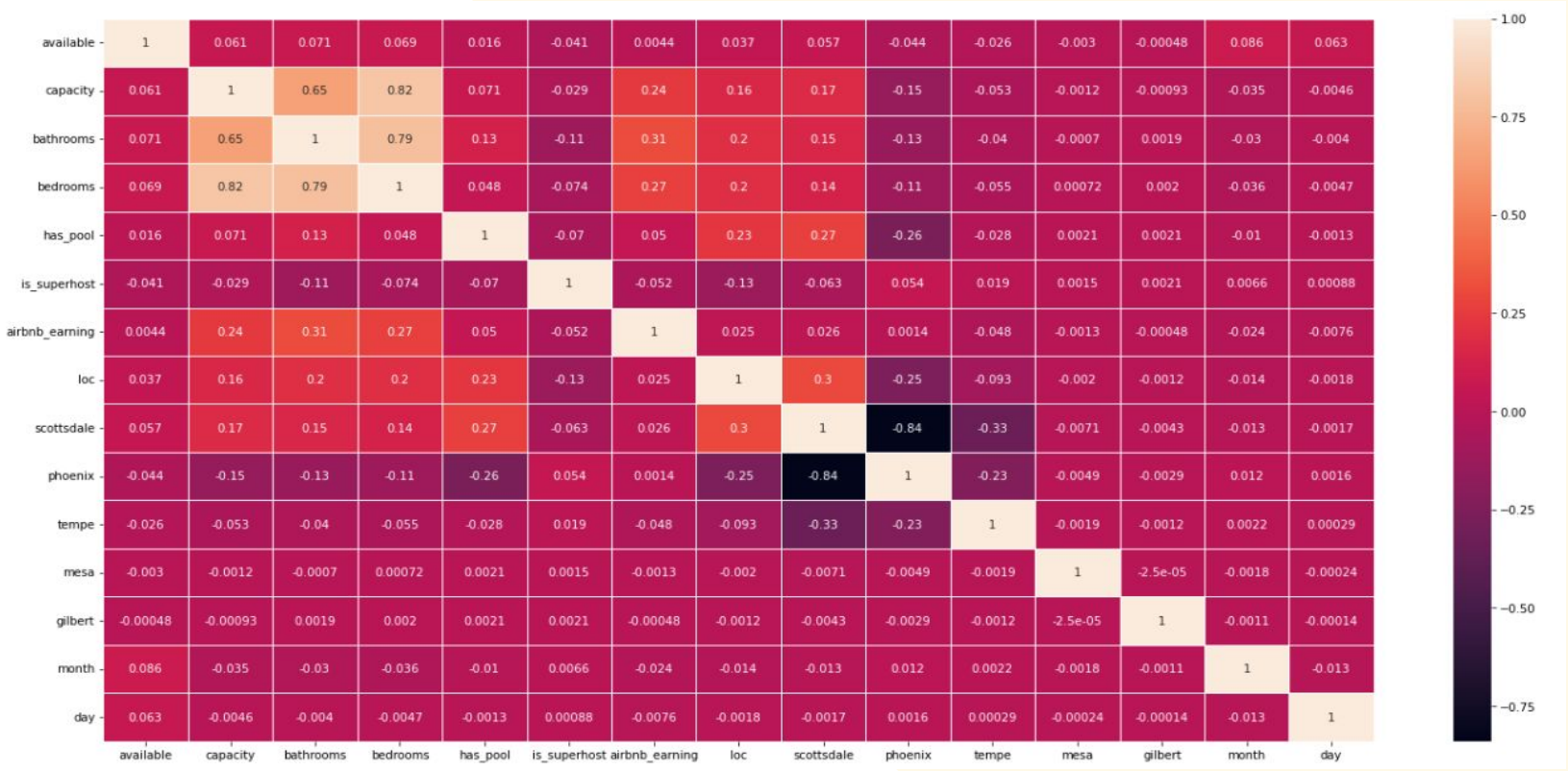
Scaling the Data

As the data was not uniform, we had to resort to scaling it. The data was scaled using Min-Max Scalar.

Splitting of data

The data was split into train and test in a ratio of 90:10.

CORRELATION OF THE VARIABLES



LOGISTIC REGRESSION

```
Accuracy Score of Logistic Regression is : 0.5581465971541418
Confusion Matrix :
[[511746 323377]
 [396852 398043]]
Classification Report :
              precision    recall  f1-score   support

           0       0.56      0.61      0.59    835123
           1       0.55      0.50      0.53    794895

    accuracy                           0.56   1630018
   macro avg       0.56      0.56      0.56   1630018
weighted avg       0.56      0.56      0.56   1630018
```

K NEAREST NEIGHBORS

```
Accuracy Score of KNN is : 0.9104408662971182
Confusion Matrix :
[[763672  71451]
 [ 74532 720363]]
Classification Report :
              precision    recall  f1-score   support

           0       0.91      0.91      0.91    835123
           1       0.91      0.91      0.91    794895

    accuracy                           0.91   1630018
   macro avg       0.91      0.91      0.91   1630018
weighted avg       0.91      0.91      0.91   1630018
```

RANDOM FOREST

```
Accuracy Score of Random Forest is : 0.9205082397863091
Confusion Matrix :
[[775221  59902]
 [ 69671 725224]]
Classification Report :
              precision    recall  f1-score   support

           0       0.92      0.93      0.92    835123
           1       0.92      0.91      0.92    794895

    accuracy                           0.92   1630018
   macro avg       0.92      0.92      0.92   1630018
weighted avg       0.92      0.92      0.92   1630018
```

ADA BOOST CLASSIFIER

Accuracy Score of Ada Boost Classifier is : 0.9203800203433337

Confusion Matrix :

```
[[775422  59701]
 [ 70081 724814]]
```

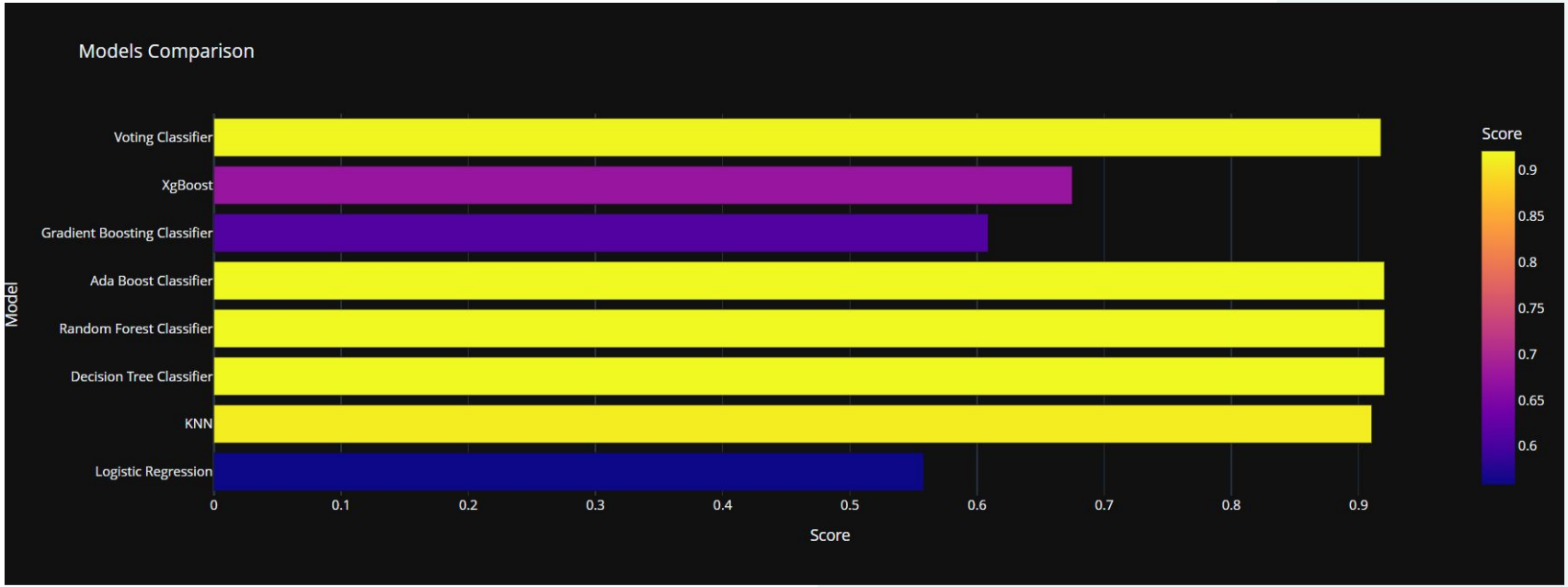
Classification Report :

	precision	recall	f1-score	support
0	0.92	0.93	0.92	835123
1	0.92	0.91	0.92	794895
accuracy			0.92	1630018
macro avg	0.92	0.92	0.92	1630018
weighted avg	0.92	0.92	0.92	1630018

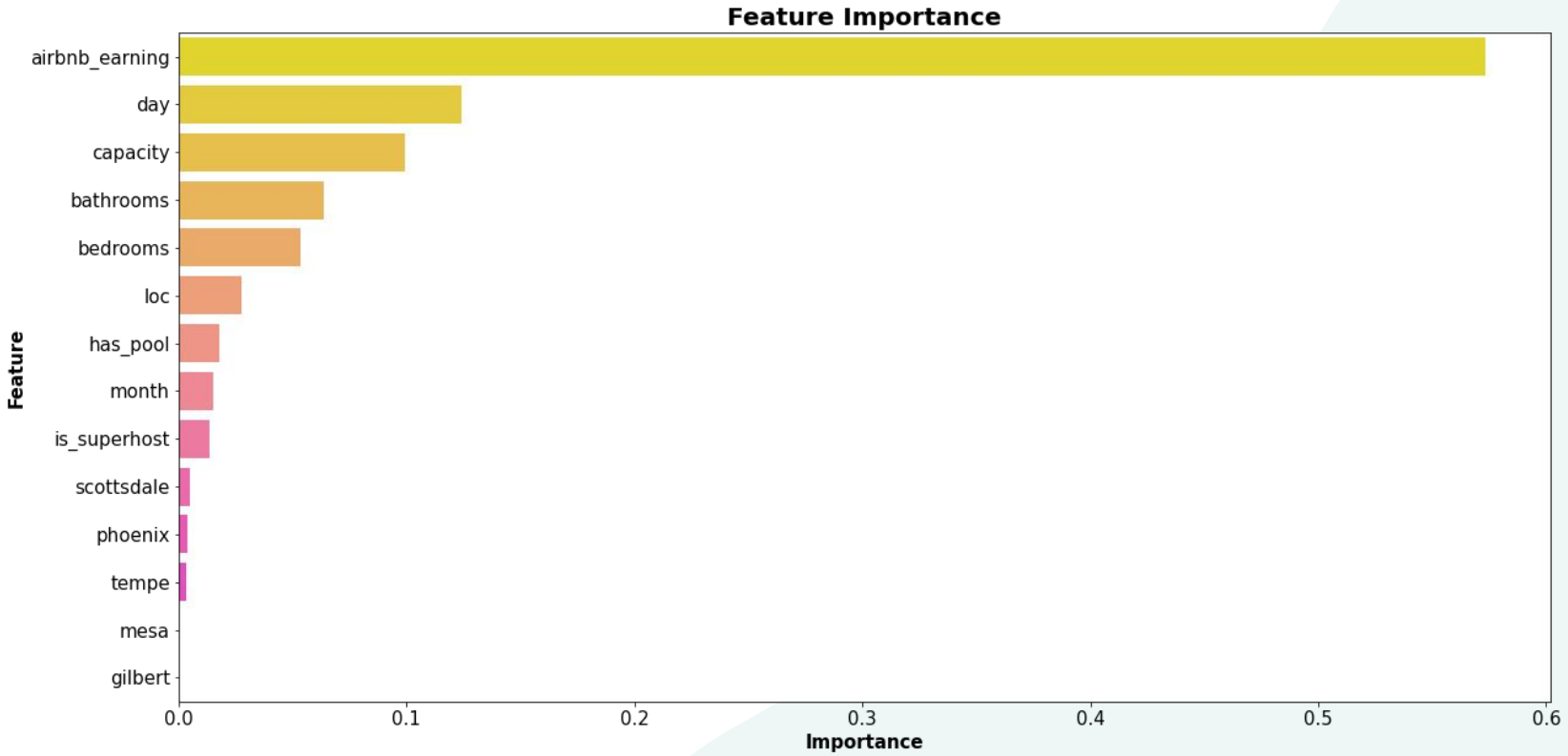
VOTING CLASSIFIER

```
Accuracy Score of Voting Classifier is : 0.9176045908695487
Confusion Matrix :
[[771814  63309]
 [ 70997 723898]]
Classification Report :
              precision    recall  f1-score   support

           0       0.92      0.92      0.92    835123
           1       0.92      0.91      0.92    794895

    accuracy                           0.92   1630018
   macro avg       0.92      0.92      0.92   1630018
weighted avg       0.92      0.92      0.92   1630018
```

COMPARISON OF MODEL PERFORMANCE



FEATURE IMPORTANCE



TOP FOUR FEATURES

CONCLUSION

- Airbnb earning, day, capacity, number of bathrooms are the features that impact the availability of the Airbnb
- We identified that there is a moderate correlation between the price range of the Airbnb and their availability.
- It was established that there is high correlation between the property having a pool, high number of rooms and its availability.
- We tried several traditional machine learning models but the best performing model was Random Forest, and we achieved an accuracy of 92.5%

THANKS