INTRODUCTION TO DATA SCIENCE AND ARTIFICIAL INTELLIGENCE MINI PROJECT

Palak Porwal

Scope Of The Presentation

- Chosen Dataset and Problem Formulation
- Data Cleaning
- Exploratory Data Analysis
- Regression: Linear and Multivariate

Chosen Dataset

Dataset 1 : AirBnB Open Data from Seattle

Problem Formulation

‘Which aspect or characteristics of the host will enable them to secure more bookings resulting in higher revenue?’

With the help of data analysis, identifying key factors can result to a increase in revenue for future potential hosts. AirBnB hosts can gain better understanding of how to improve their profiles and develop higher number of bookings to maximize their profits rates in return.

Data Cleaning

Data Cleaning

- The Initial number of Predictors was 92
- Final Number of Predictors after cleaning was 33
- The object Data types were converted to Category
- All the NAs in the data columns were replaced with 0
- Estimated Revenue was calculated for the listings.

Dataset Cleaning

```
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   listing_id                    3818 non-null   int64
 1   name                          3818 non-null   category
 2   host_response_time            3818 non-null   category
 3   host_is_superhost             3818 non-null   category
 4   host_listings_count           3818 non-null   float64
 5   host_total_listings_count     3818 non-null   float64
 6   host_has_profile_pic          3818 non-null   category
 7   host_identity_verified        3818 non-null   category
 8   availability_30               3818 non-null   int64
 9   reviews_per_month             3818 non-null   float64
 10  number_of_reviews             3818 non-null   int64
 11  price                         3818 non-null   float64
 12  minimum_nights                3818 non-null   int64
 13  review_scores_rating          3818 non-null   float64
 14  review_scores_accuracy        3818 non-null   float64
 15  review_scores_cleanliness     3818 non-null   float64
 16  review_scores_checkin         3818 non-null   float64
 17  review_scores_communication   3818 non-null   float64
 18  review_scores_location        3818 non-null   float64
 19  review_scores_value           3818 non-null   float64
 20  neighbourhood_group_cleansed  3818 non-null   category
 21  latitude                      3818 non-null   float64
 22  longitude                     3818 non-null   float64
 23  is_location_exact             3818 non-null   category
 24  host_location                 3818 non-null   category
 25  property_type                 3818 non-null   category
 26  room_type                     3818 non-null   category
 27  accommodates                  3818 non-null   int64
 28  bathrooms                     3818 non-null   float64
 29  bedrooms                      3818 non-null   float64
 30  beds                          3818 non-null   float64
 31  amenities                     3818 non-null   category
 32  guests_included               3818 non-null   int64
 33  neighbourhood_cleansed        3818 non-null   category
dtypes: category(12), float64(16), int64(6)
```

Estimated Revenue for each listing

listing_id	estimated_revenue
4291	5740.0
5682	42768.0
6606	9360.0
7369	3400.0
9419	14220.0
...	...
9995551	79.0
10012724	50.0
10020221	55.0
10118341	210.0
10248139	22.0

The Estimated Revenue was calculated by multiplying (Price x Minimum Night)

Data Analysis

Exploratory Analysis

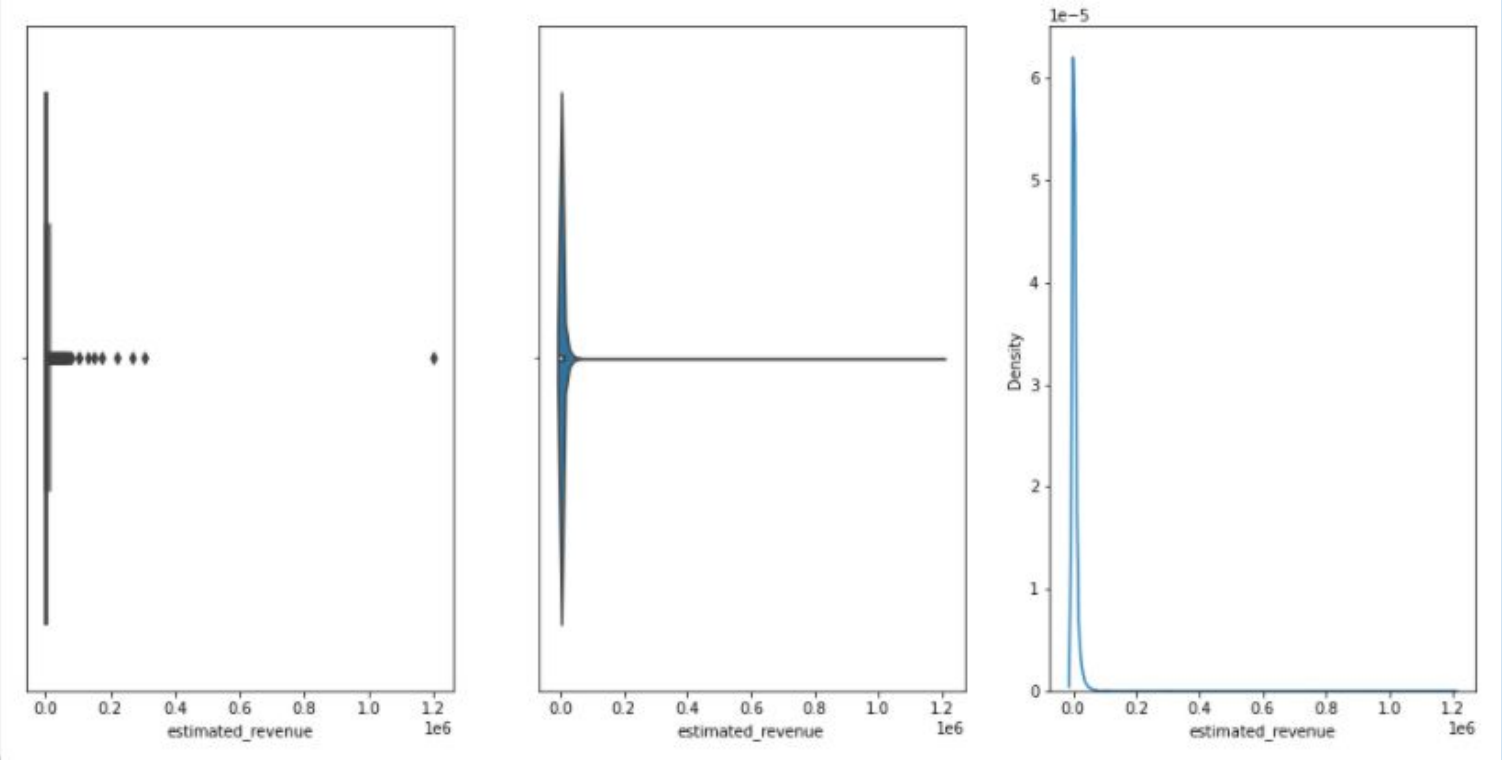
These are the various types to exploratory data analysis carried out

- Categorical Variables vs Estimated Revenue
- Numeric Variable vs Estimated Revenue
- Neighbourhood vs Estimated Revenue
- Accommodates vs Estimated Revenue
- Which type of Amenities should the host include?

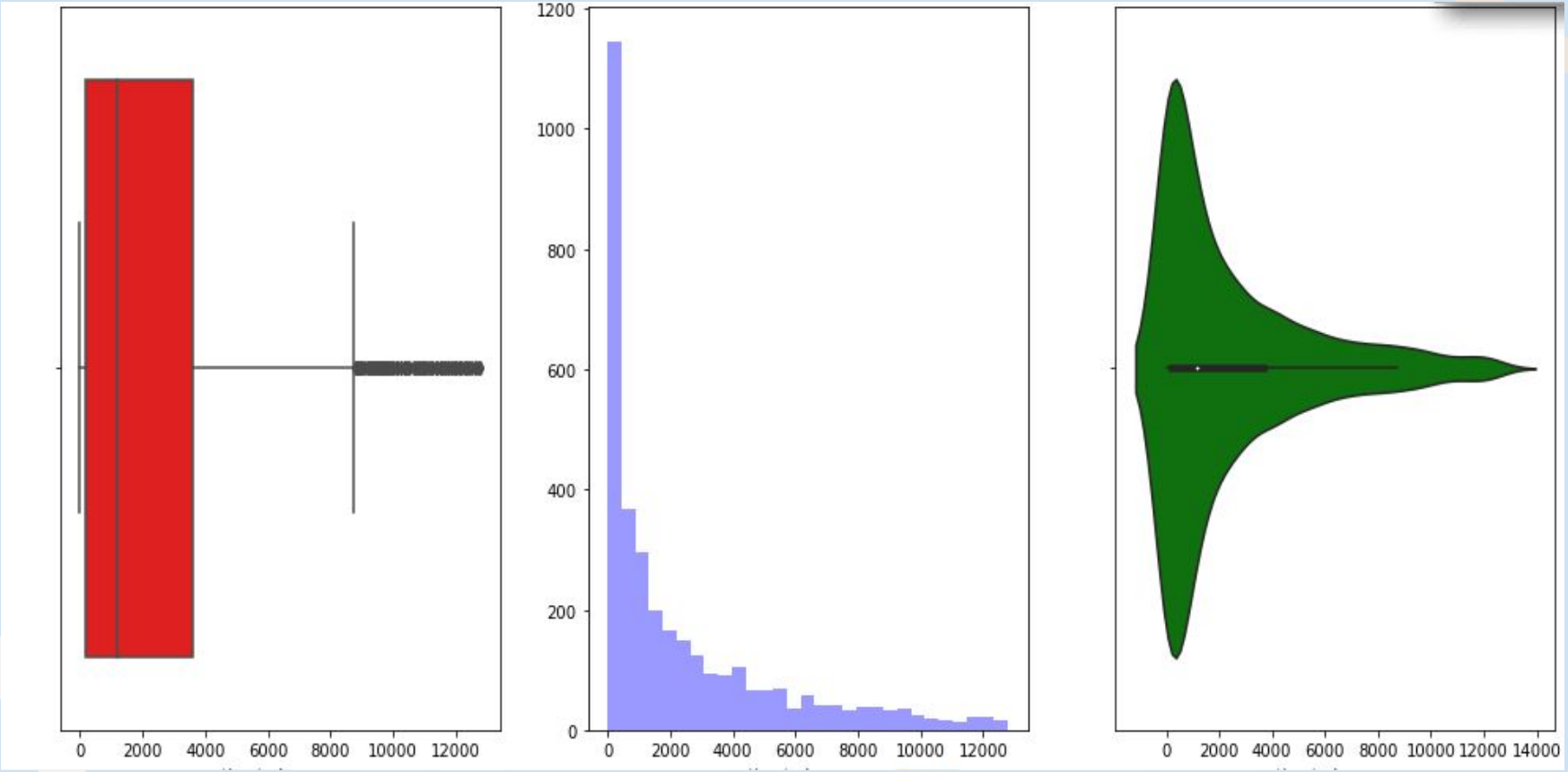
Results for highest revenue listings

	listing_id	number_of_reviews	minimum_nights	accommodates	bedrooms	beds	estimated_revenue
2617	3594885	8	1000	4	1.0	1.0	1200000.0
2107	5056580	100	31	2	1.0	1.0	306900.0
1500	4009508	38	20	5	2.0	2.0	266000.0
1537	1954452	71	14	2	1.0	1.0	218680.0
1519	3971934	48	20	3	1.0	1.0	171840.0
...	...	...	...	...	...	...	...
2982	9463729	0	1	2	1.0	1.0	0.0
2983	8866331	0	1	2	1.0	1.0	0.0
1146	8829089	0	1	1	1.0	1.0	0.0
2986	8484705	0	4	4	1.0	1.0	0.0
3817	10208623	0	1	3	2.0	1.0	0.0

Raw plots for Estimated Revenue before IQR

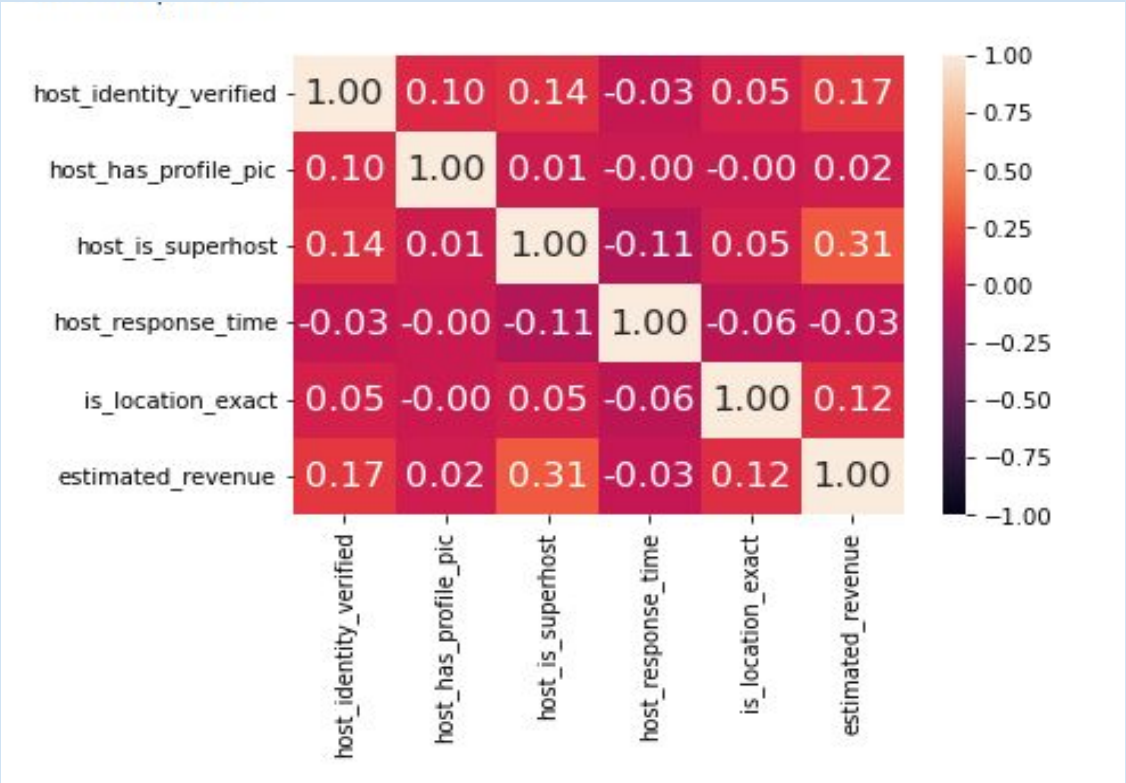


Estimated Revenue after IQR



Estimated Revenue vs Category Variable

Relation between Categorical Variables and Estimated Revenue

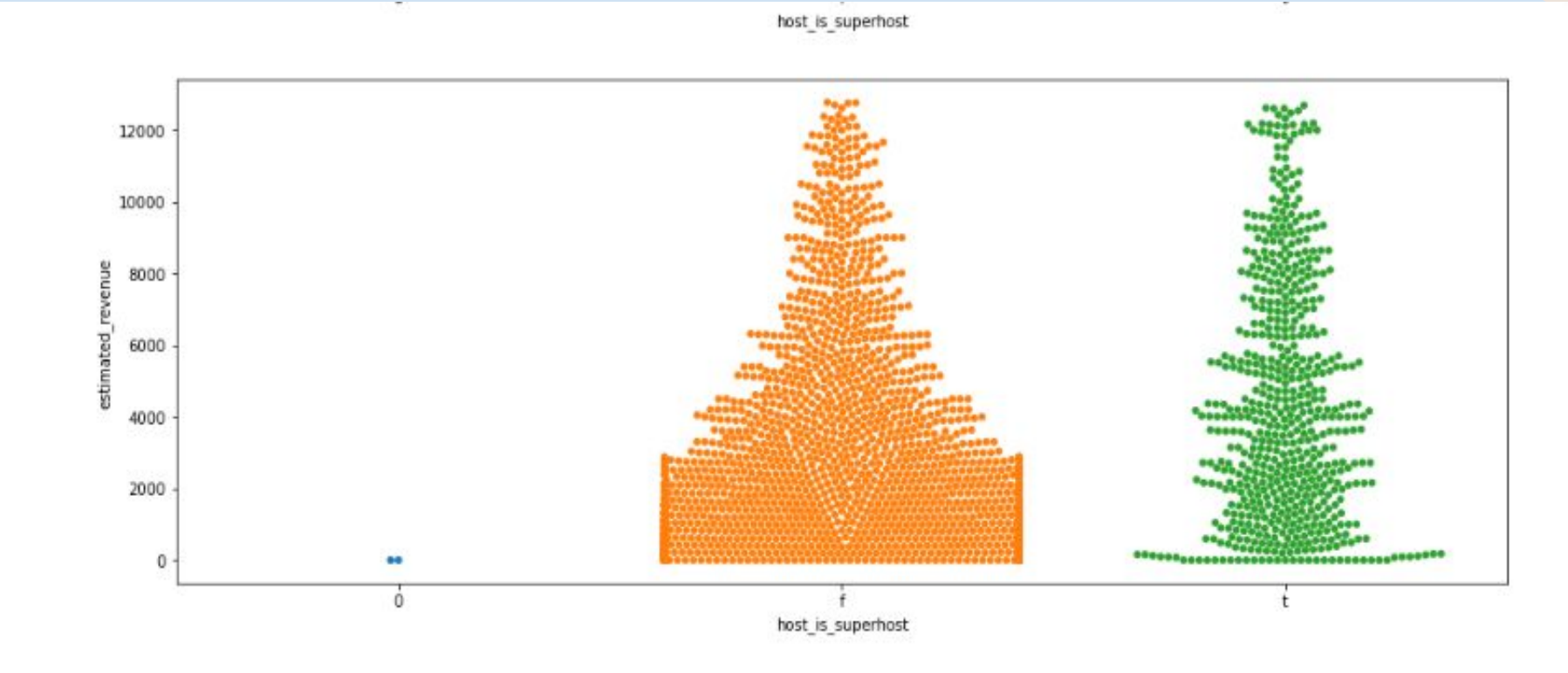


Relation between Categorical Variables and Estimated Revenue

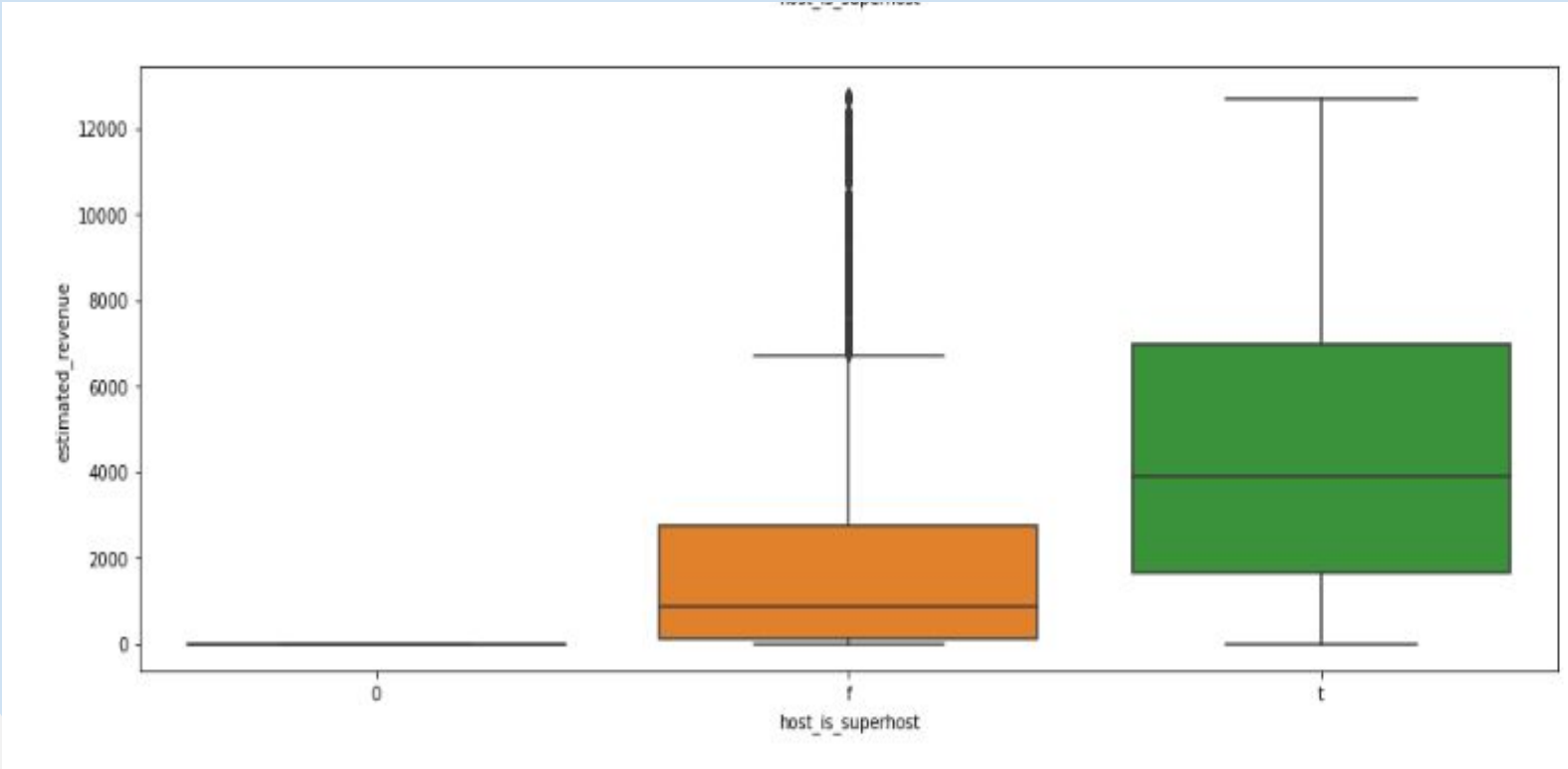
- **From the heatmap, we can infer that the following predictors have the highest correlation with Estimated Revenue**
  - **Superhost Status of the host**
  - **Identity verification of the host**
  - **Accuracy of the location**
- **Further Data Analysis presents an accurate representation of the relationship of the above variables with the Estimated Revenue**

Estimated Revenue vs Superhost Status

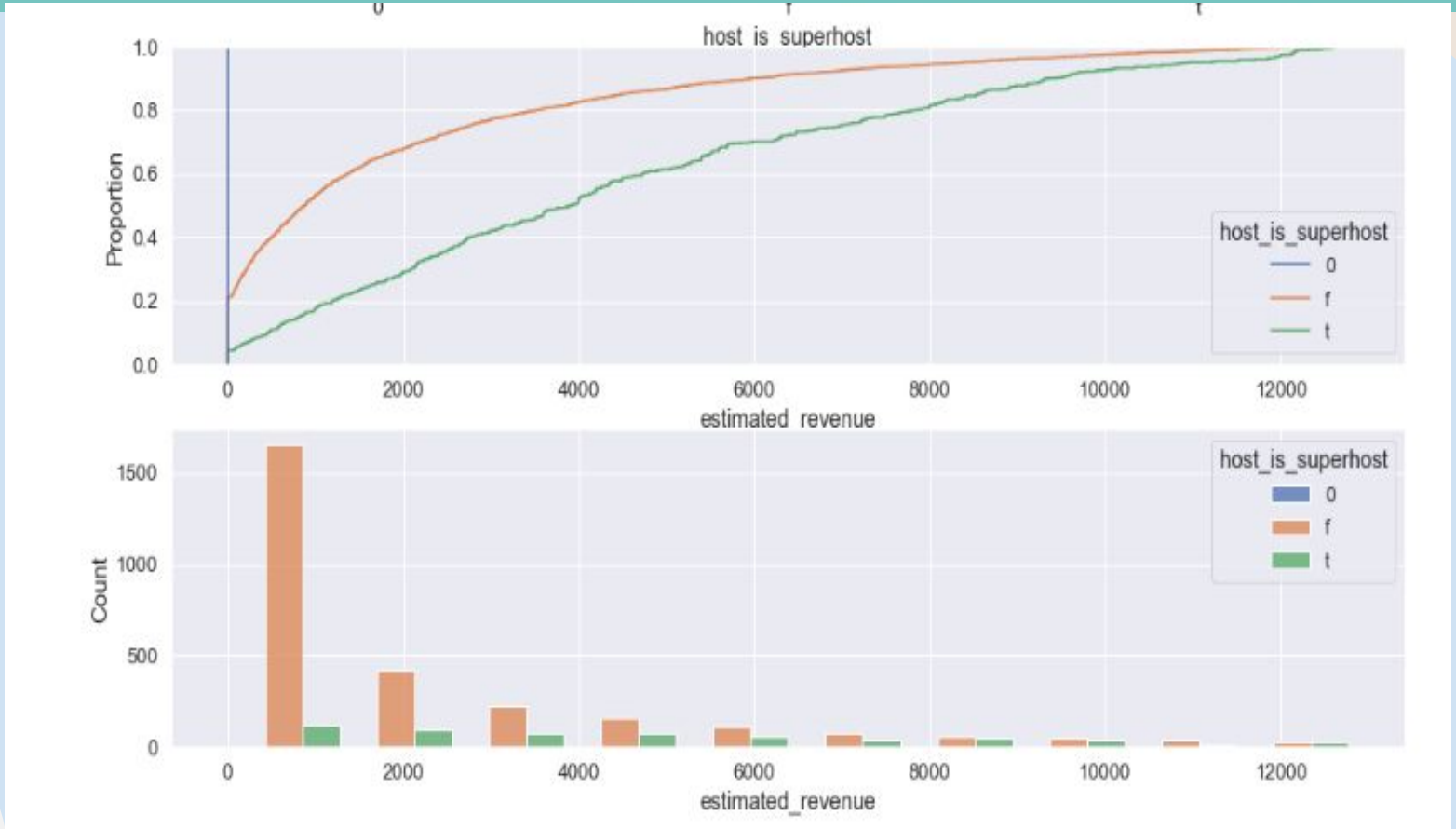
Superhost status vs Estimated Revenue



Superhost status vs Estimated Revenue

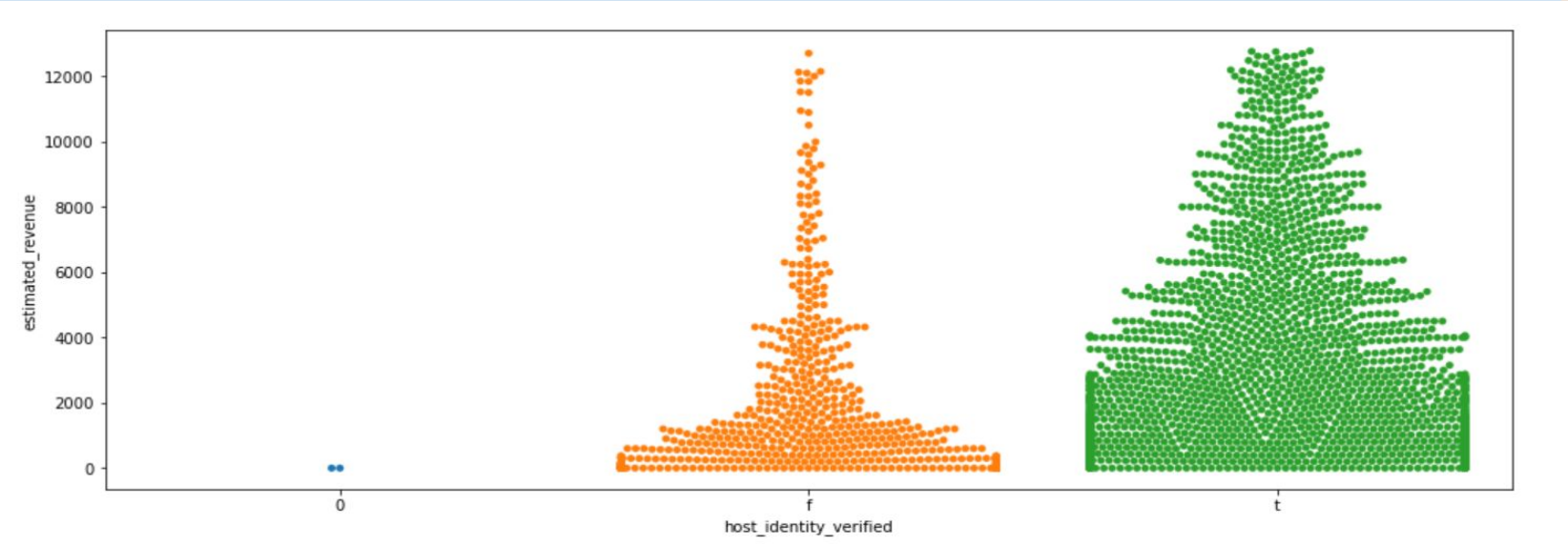


Superhost Status vs Estimated Revenue

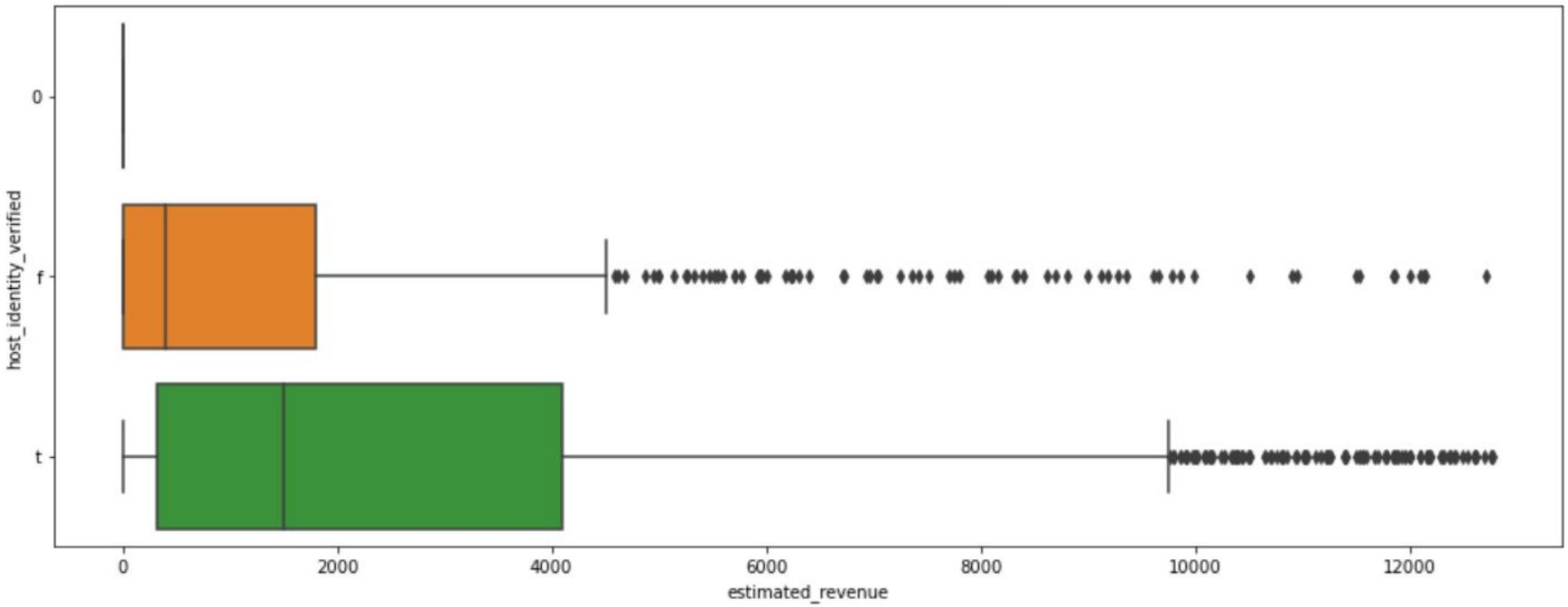


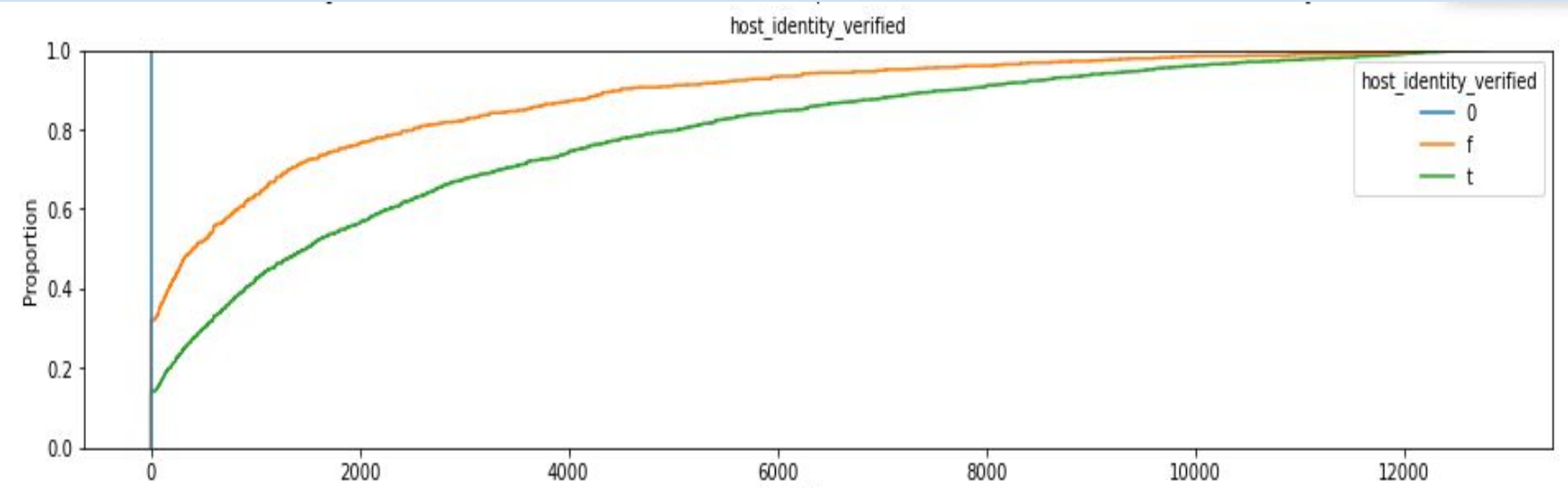
Estimated Revenue vs Host Identity Verification

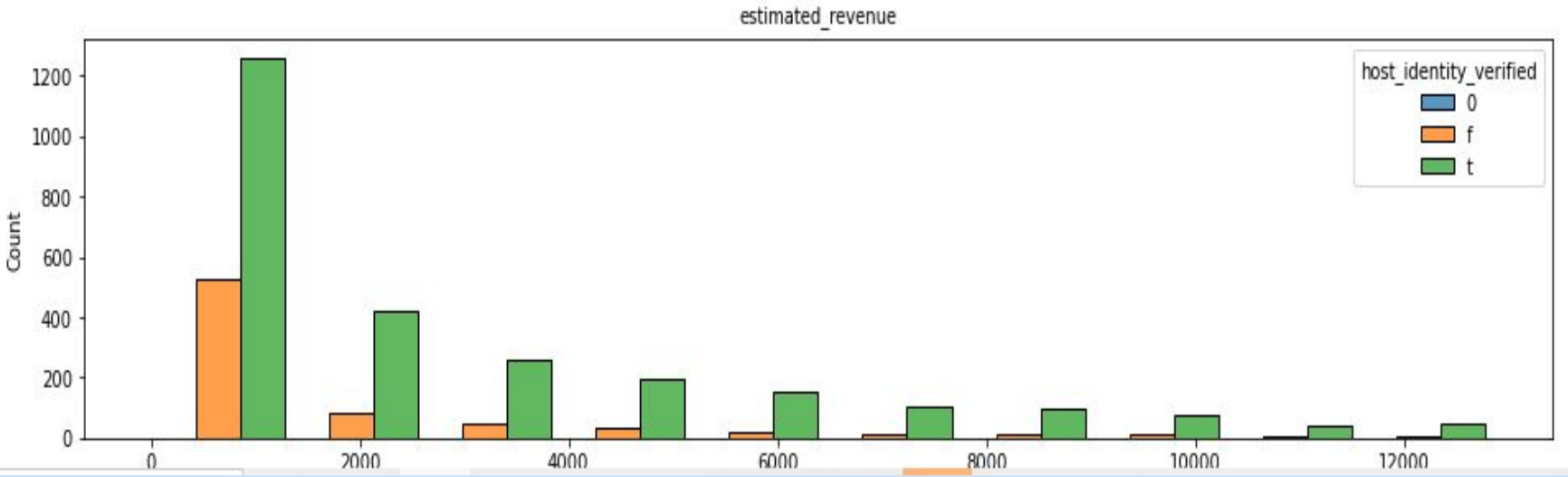
Host Identity Verification vs Estimated Revenue



Host Identity Verification vs Estimated Revenue

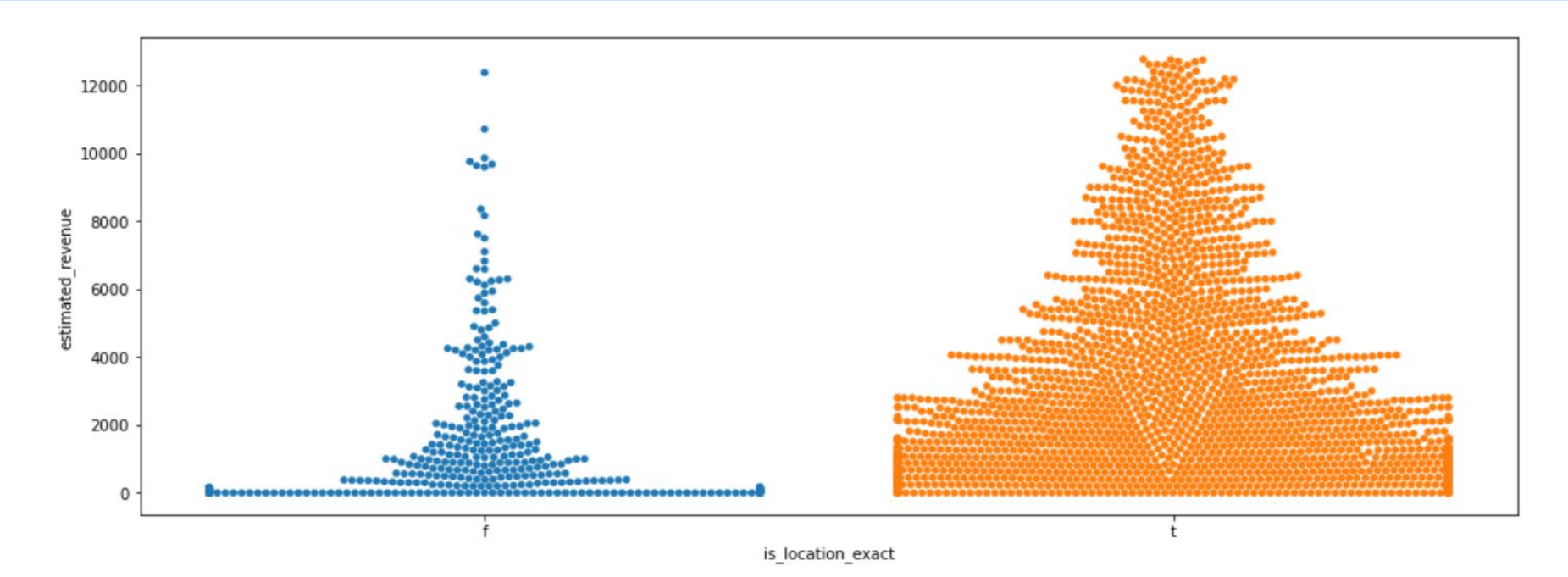



host_identity_verified
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
2000
4000
6000
8000
10000
12000
host_identity_verified
0
f
t

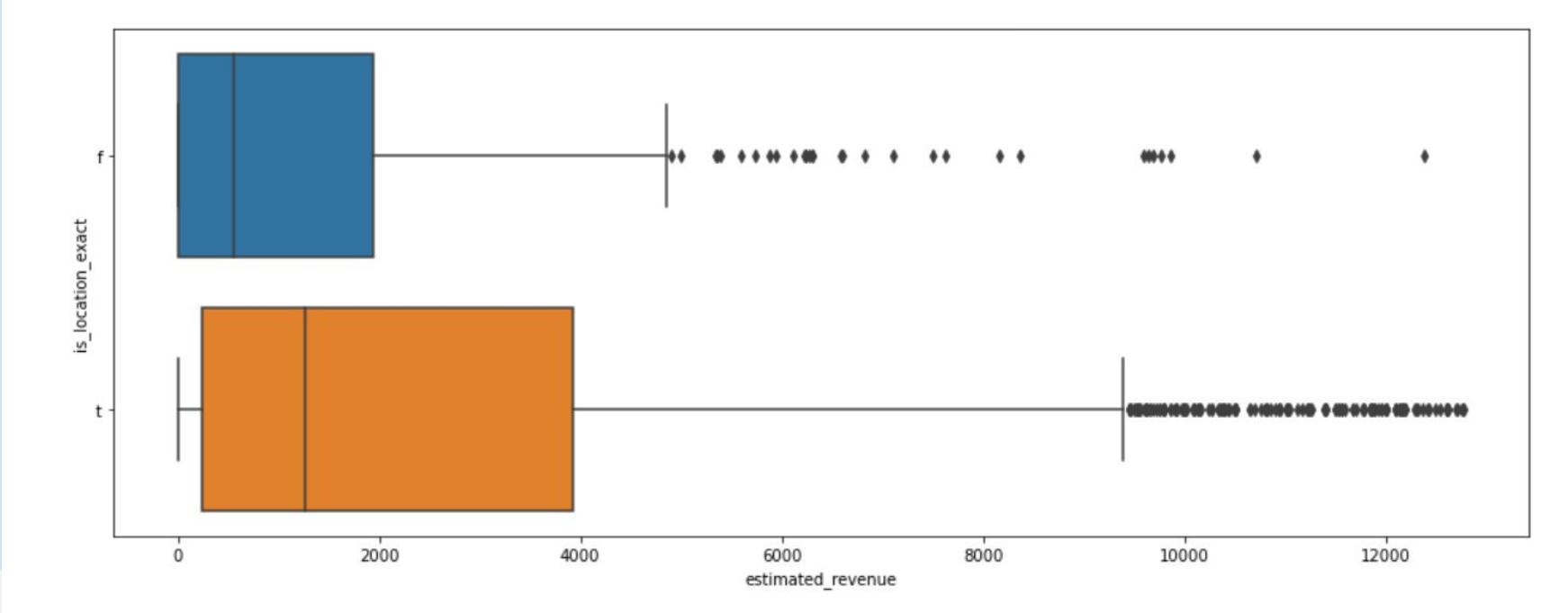

estimated_revenue
Count
1200
1000
800
600
400
200
0
0
2000
4000
6000
8000
10000
12000
host_identity_verified
0
f
t

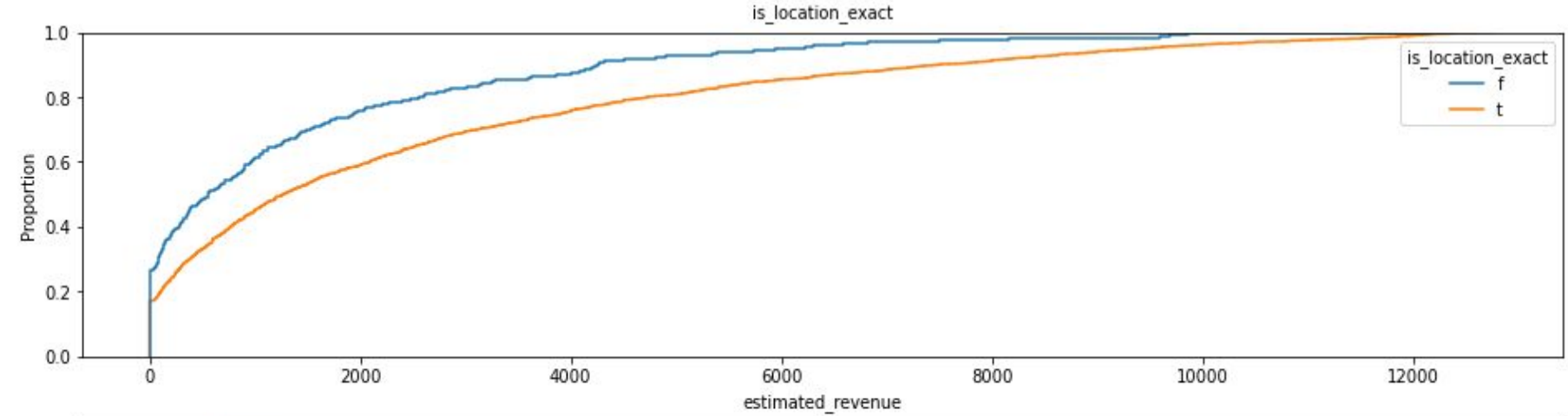
Estimated Revenue vs Accuracy of the location

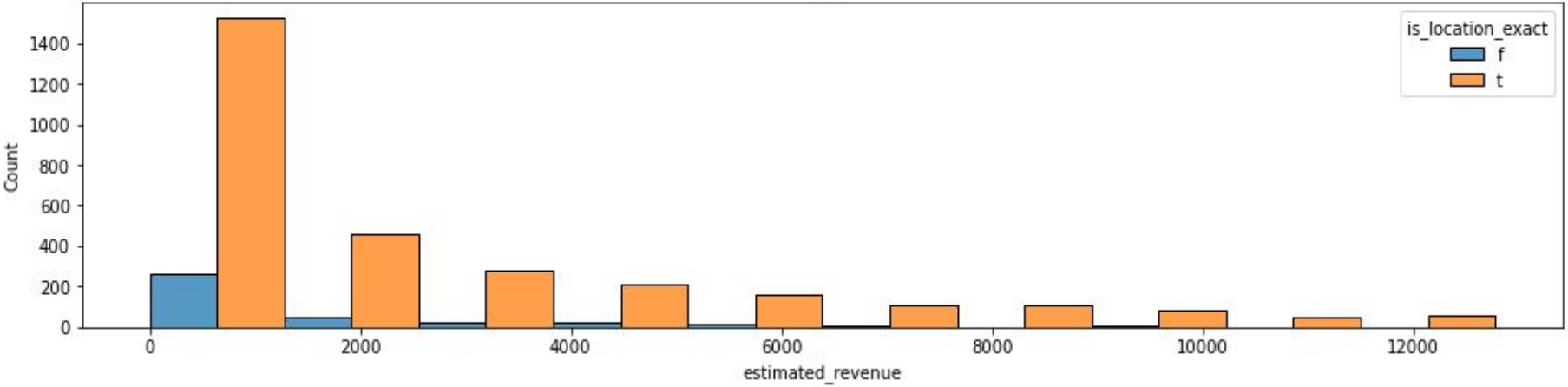
Swarmplot for Accuracy of the location vs Estimated Revenue



Box plot for the Accuracy of the location vs Estimated Revenue

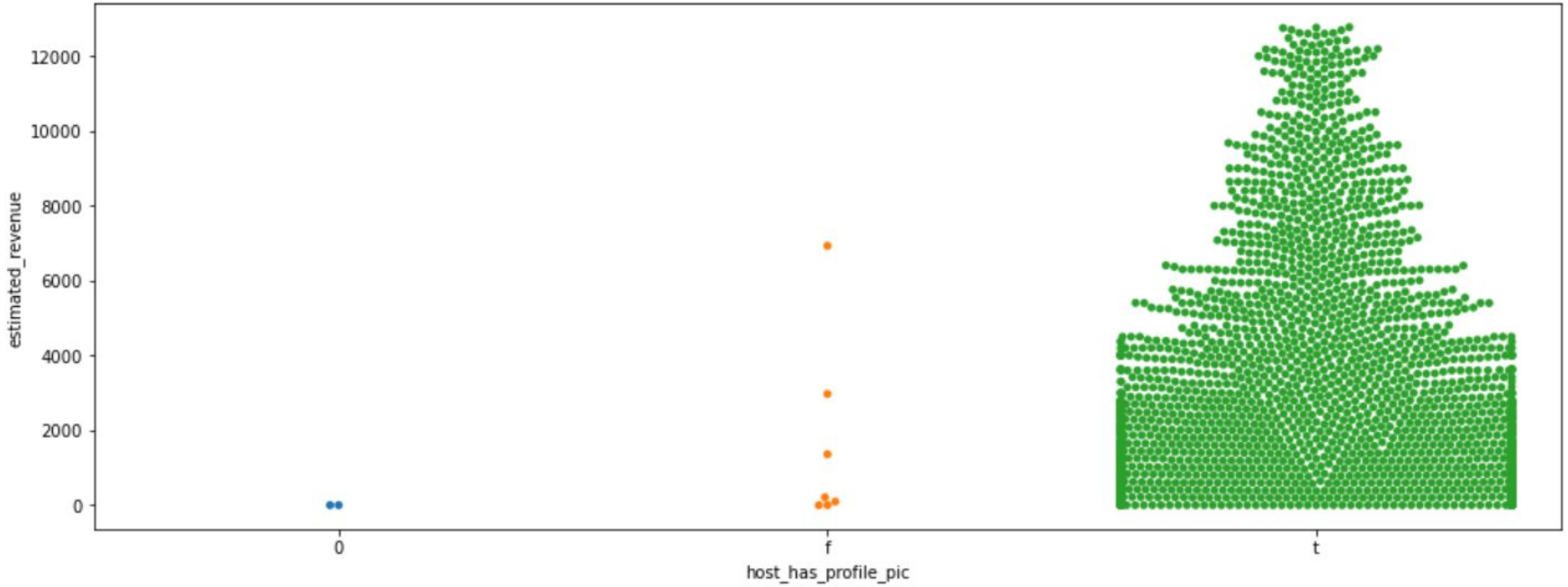



is_location_exact
Proportion
estimated_revenue
is_location_exact
f
t
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
12000

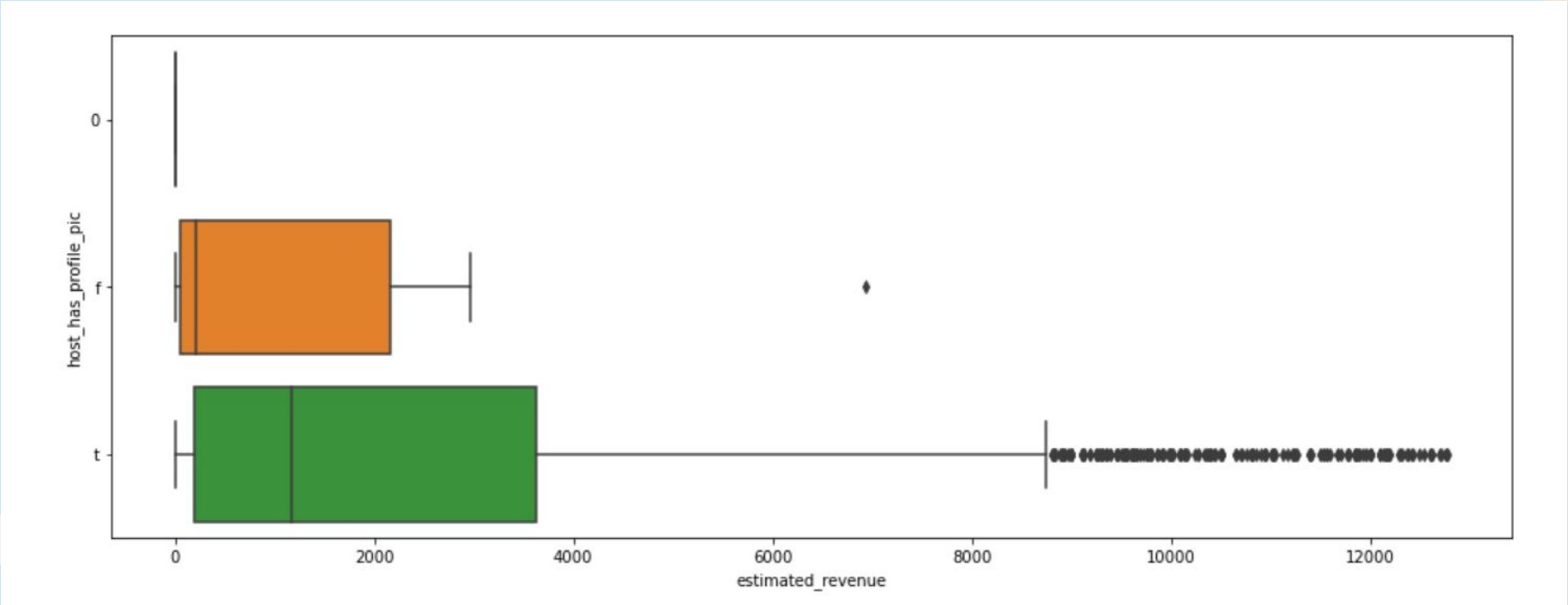

Count
estimated_revenue
is_location_exact
f
t
0
200
400
600
800
1000
1200
1400
0
2000
4000
6000
8000
10000
12000

Estimated Revenue vs Host’s profile pic

Swarmplot for Host's profile pic vs Estimated Revenue

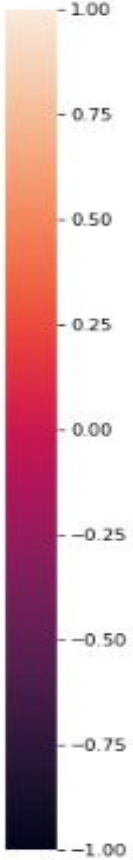


Boxplot for Host’s profile pic vs Estimated Revenue

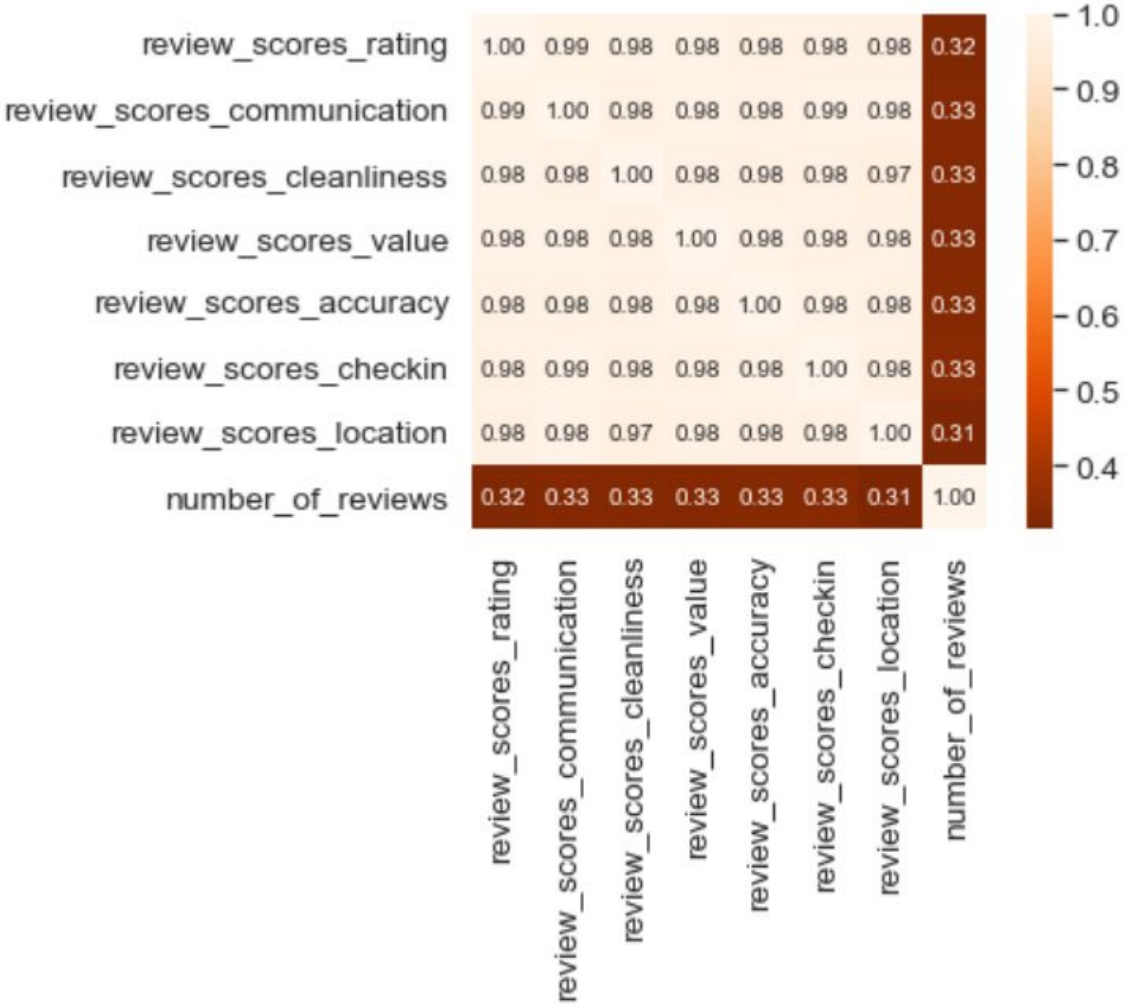


Relation between Numerical Variable and Estimated Revenue

	host_response_rate	host_acceptance_rate	host_listings_count	host_total_listings_count	availability_30	reviews_per_month	number_of_reviews	price	minimum_nights	review_scores_rating	review_scores_accuracy	review_scores_cleanliness	review_scores_checkin	review_scores_communication	review_scores_location	review_scores_value	estimated_revenue
host_response_rate	1.00	0.72	0.09	0.09	0.16	0.31	0.19	-0.09	-0.02	0.26	0.26	0.27	0.26	0.26	0.26	0.26	0.09
host_acceptance_rate	0.72	1.00	0.11	0.11	0.15	0.34	0.20	-0.13	-0.07	0.28	0.29	0.29	0.29	0.28	0.29	0.29	0.07
host_listings_count	0.09	0.11	1.00	1.00	0.12	-0.09	-0.06	0.09	0.03	-0.03	-0.03	-0.02	-0.04	-0.03	-0.01	-0.03	0.00
host_total_listings_count	0.09	0.11	1.00	1.00	0.12	-0.09	-0.06	0.09	0.03	-0.03	-0.03	-0.02	-0.04	-0.03	-0.01	-0.03	0.00
availability_30	0.16	0.15	0.12	0.12	1.00	0.06	0.07	-0.04	-0.03	0.02	0.02	0.03	0.03	0.03	0.03	0.02	0.06
reviews_per_month	0.31	0.34	-0.09	-0.09	0.06	1.00	0.60	-0.19	-0.13	0.43	0.44	0.44	0.44	0.43	0.43	0.44	0.23
number_of_reviews	0.19	0.20	-0.06	-0.06	0.07	0.60	1.00	-0.12	-0.06	0.27	0.28	0.27	0.28	0.27	0.26	0.27	0.48
price	-0.09	-0.13	0.09	0.09	-0.04	-0.19	-0.12	1.00	0.10	-0.02	-0.02	-0.02	-0.02	-0.03	-0.01	-0.03	0.19
minimum_nights	-0.02	-0.07	0.03	0.03	-0.03	-0.13	-0.06	0.10	1.00	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	0.42
review_scores_rating	0.26	0.28	-0.03	-0.03	0.02	0.43	0.27	-0.02	-0.02	1.00	0.98	0.98	0.98	0.99	0.98	0.98	0.07
review_scores_accuracy	0.26	0.29	-0.03	-0.03	0.02	0.44	0.28	-0.02	-0.02	0.98	1.00	0.98	0.98	0.98	0.98	0.98	0.07
review_scores_cleanliness	0.27	0.29	-0.02	-0.02	0.03	0.44	0.27	-0.02	-0.02	0.98	0.98	1.00	0.98	0.98	0.97	0.98	0.08
review_scores_checkin	0.26	0.29	-0.04	-0.04	0.03	0.44	0.28	-0.02	-0.02	0.98	0.98	0.98	1.00	0.99	0.98	0.98	0.07
review_scores_communication	0.26	0.28	-0.03	-0.03	0.03	0.43	0.27	-0.03	-0.02	0.99	0.98	0.98	0.99	1.00	0.98	0.98	0.07
review_scores_location	0.26	0.29	-0.01	-0.01	0.03	0.43	0.26	-0.01	-0.02	0.98	0.98	0.97	0.98	0.98	1.00	0.98	0.06
review_scores_value	0.26	0.29	-0.03	-0.03	0.02	0.44	0.27	-0.03	-0.02	0.98	0.98	0.98	0.98	0.98	0.98	1.00	0.04
estimated_revenue	0.09	0.07	0.00	0.00	0.06	0.23	0.48	0.19	0.42	0.07	0.07	0.08	0.07	0.07	0.06	0.04	1.00



Review Heatmap
Which aspect of the ratings matter most to the visitors?


review_scores_rating 1.00 0.99 0.98 0.98 0.98 0.98 0.98 0.32
review_scores_communication 0.99 1.00 0.98 0.98 0.98 0.99 0.98 0.33
review_scores_cleanliness 0.98 0.98 1.00 0.98 0.98 0.98 0.97 0.33
review_scores_value 0.98 0.98 0.98 1.00 0.98 0.98 0.98 0.33
review_scores_accuracy 0.98 0.98 0.98 0.98 1.00 0.98 0.98 0.33
review_scores_checkin 0.98 0.99 0.98 0.98 0.98 1.00 0.98 0.33
review_scores_location 0.98 0.98 0.97 0.98 0.98 0.98 1.00 0.31
number_of_reviews 0.32 0.33 0.33 0.33 0.33 0.33 0.31 1.00
review_scores_rating
review_scores_communication
review_scores_cleanliness
review_scores_value
review_scores_accuracy
review_scores_checkin
review_scores_location
number_of_reviews
1.0
0.9
0.8
0.7
0.6
0.5
0.4

Review Heatmap

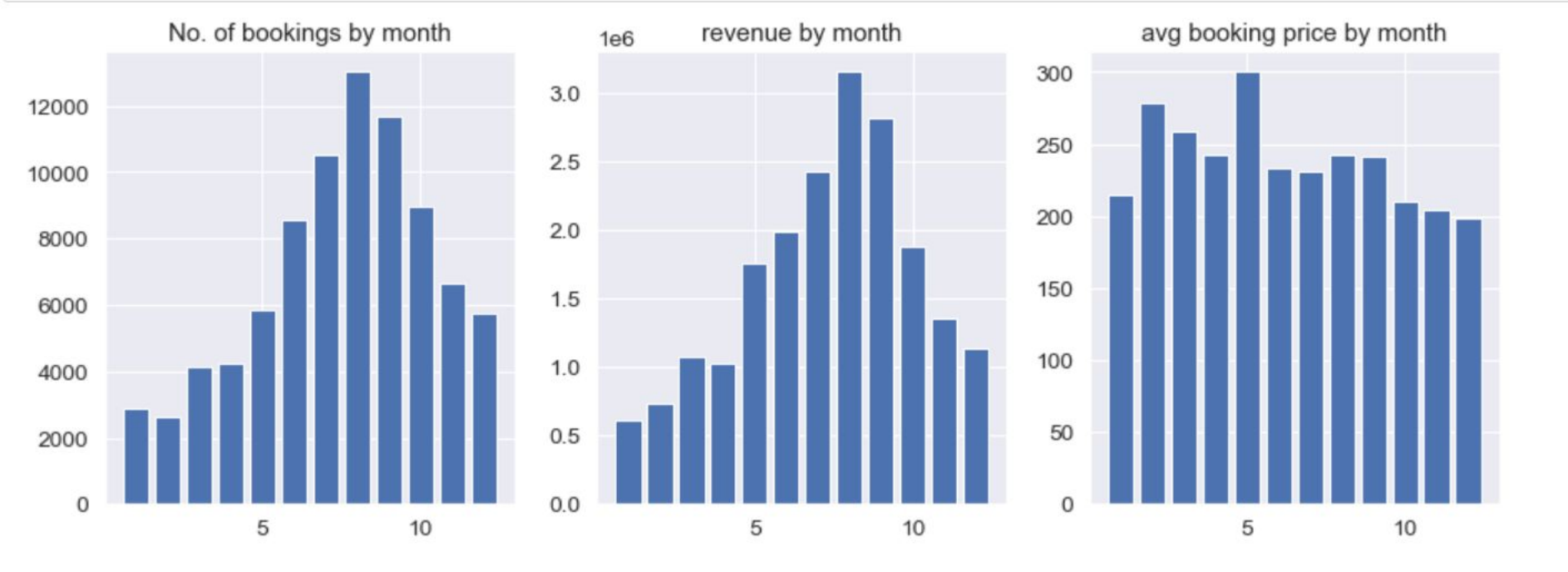
Which aspect of the ratings matter most to the visitors?

Following are the most correlated columns

- **review_scores_communication**
- **review_scores_cleanliness**
- **review_scores_value**
- **review_scores_accuracy**
- **review_scores_checkin**
- **review_scores_location**
- **number_of_reviews**

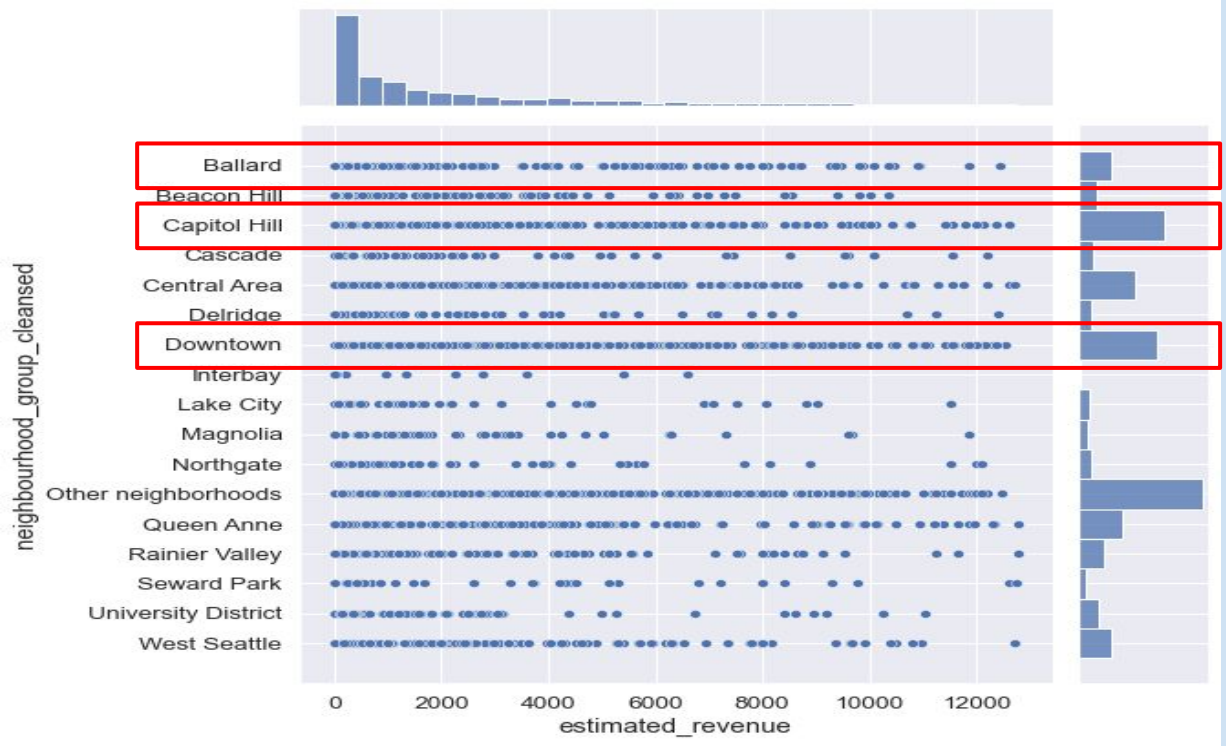
Which month is the best for renting out properties in Seattle ?

Which month is best for the renting out properties in Seattle?



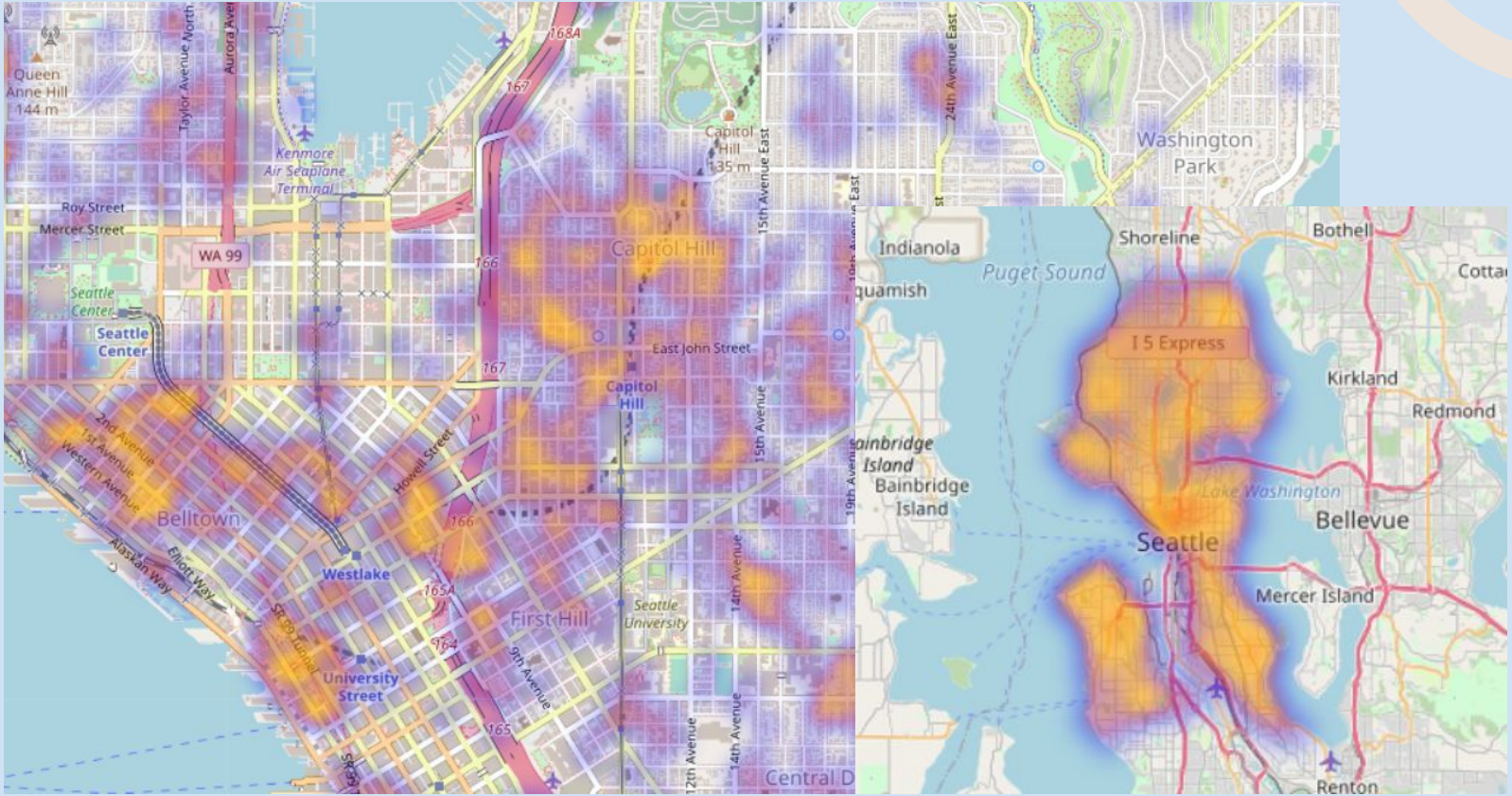
Which neighbourhoods will have higher estimated revenues?

Which neighbourhoods will have higher estimated revenues?

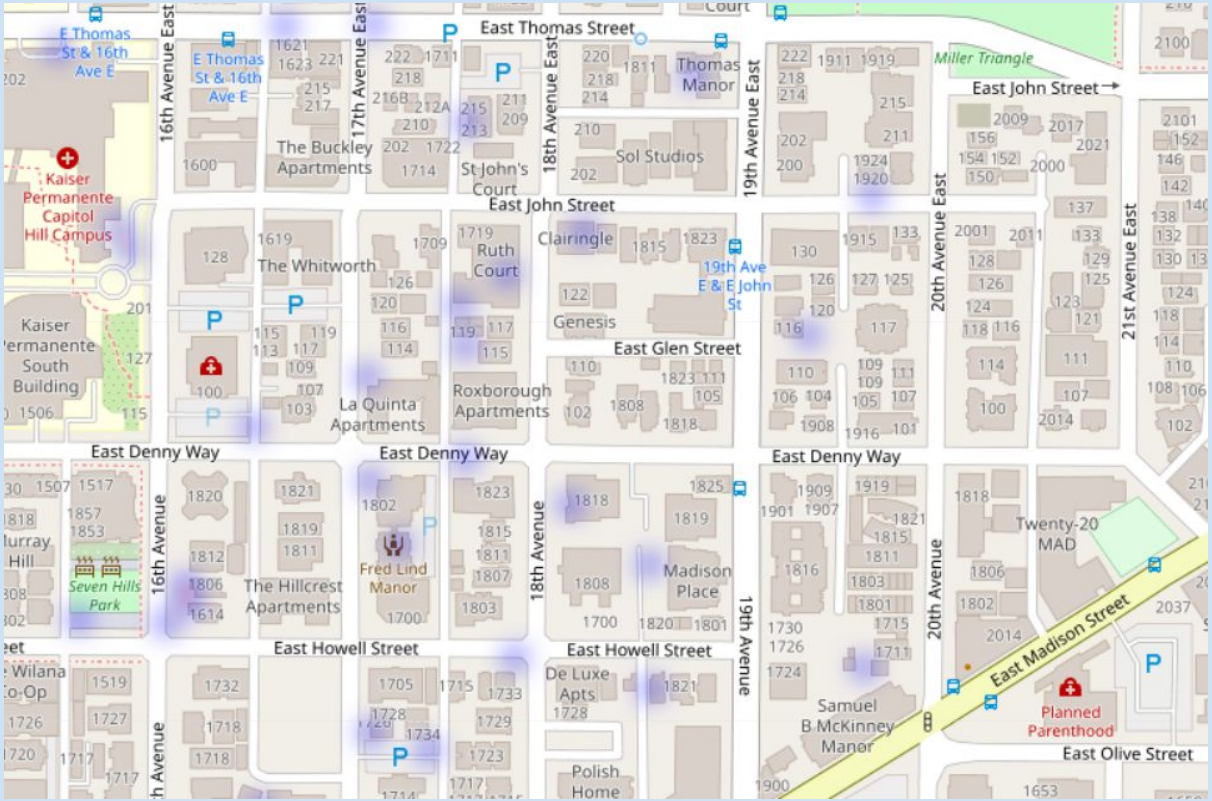


Relationship between the neighbourhood and the estimated revenue

Additional Feature (Heatmap to show listings around seattle)



Additional Feature (Heatmap to show listings around seattle)



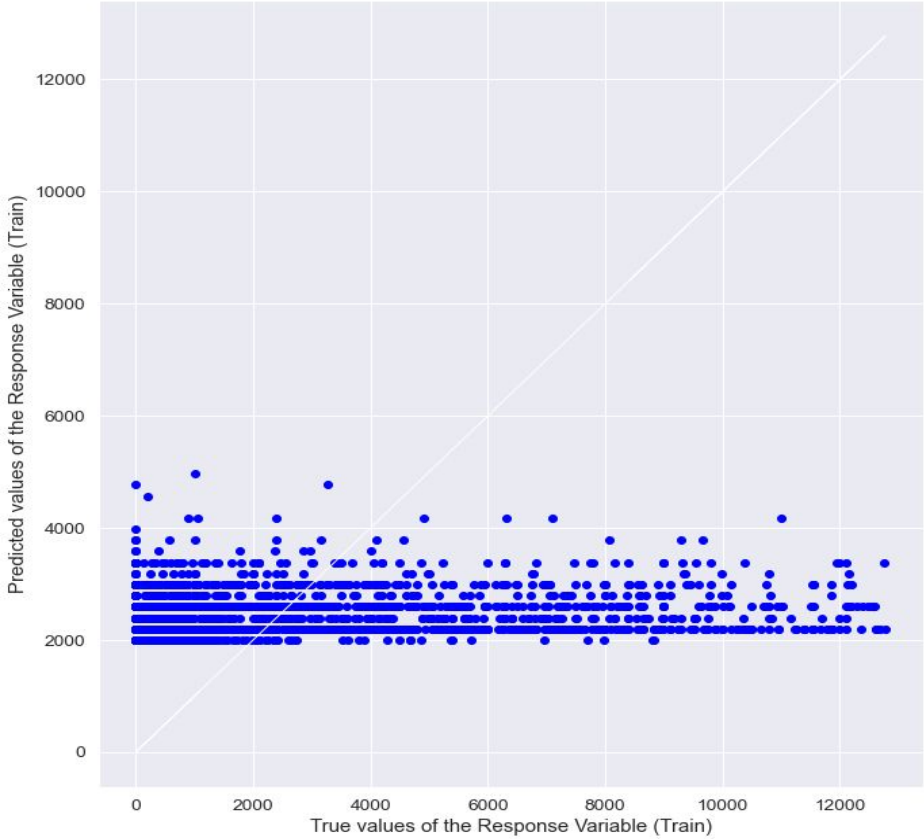
How does accommodation affects the booking rate?

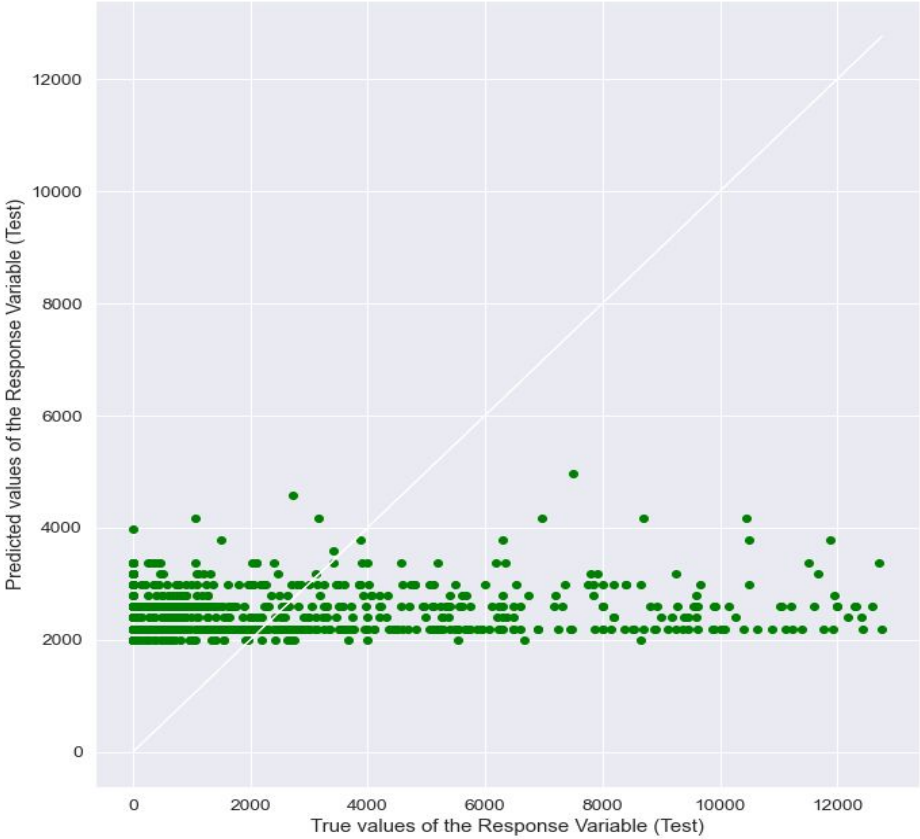
What kind of accommodation has highest booking

accommodates	no. of listings	no. of bookings	ratio
14	2	83	41.500000
10	16	520	32.500000
2	1498	42821	28.585447
3	358	10170	28.407821
16	2	48	24.000000
4	692	16041	23.180636
7	47	956	20.340426
5	159	3221	20.257862
6	281	5580	19.857651
12	12	229	19.083333
8	98	1501	15.316327
1	252	3542	14.055556
9	9	98	10.888889
11	2	20	10.000000
15	2	19	9.500000

Regression Models

Univariate Regression: Accommodates vs Estimated Revenue





Univariate Regression: Accommodates vs Estimated Revenue

Due to the nature of how we define expected revenue, MSE and R2 appear to be quite off. As a result, we choose to do regression against price.

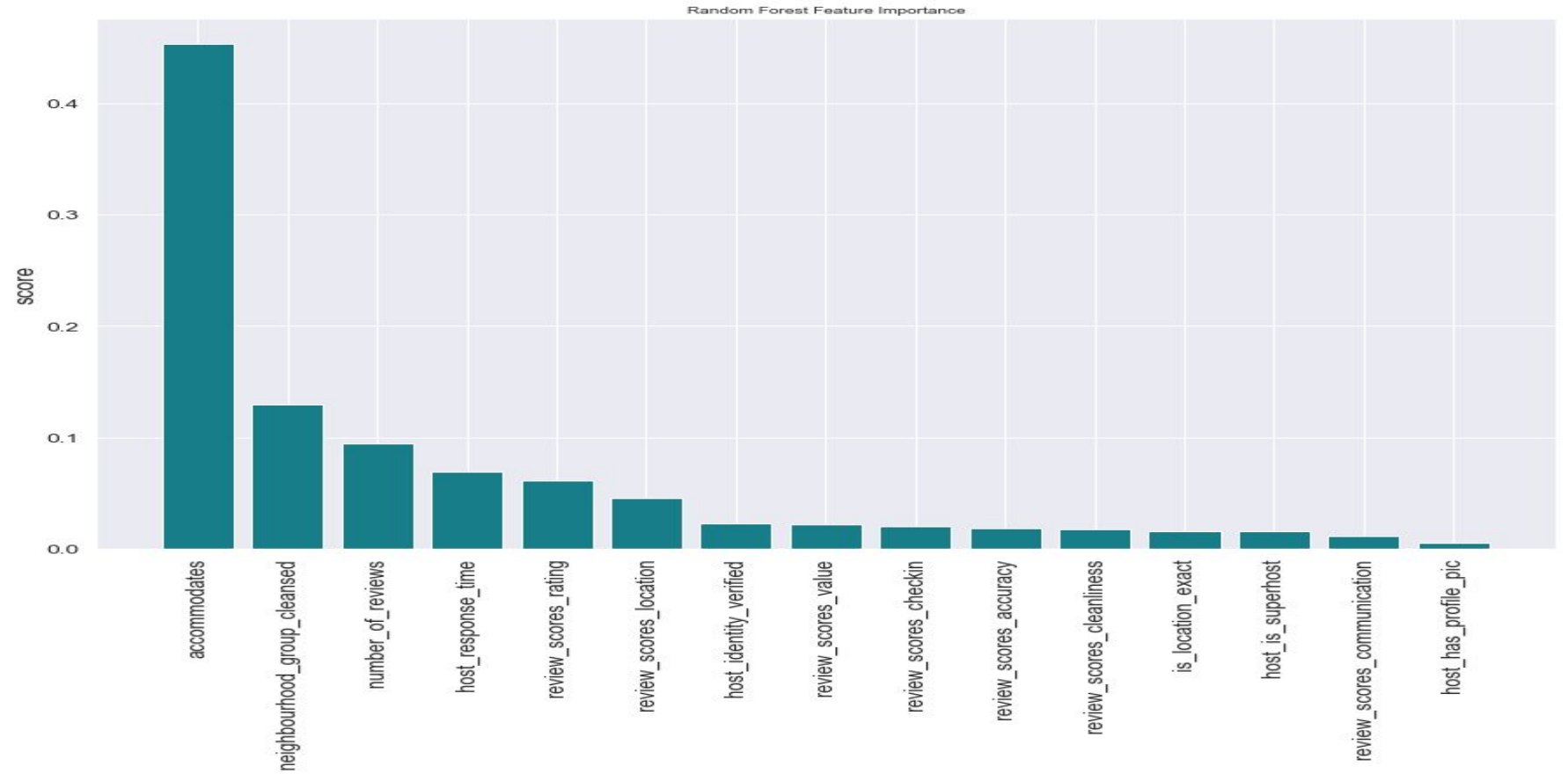
```
Goodness of Fit of Model          Train Dataset
Explained Variance (R^2)          : 0.015467425303916427
Mean Squared Error (MSE)          : 8701618.652038839

Goodness of Fit of Model          Test Dataset
Explained Variance (R^2)          : 0.03152956214841107
Mean Squared Error (MSE)          : 8986714.080592982
```

Multivariate Regression: Features vs price

```
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   neighbourhood_group_cleansed  3430 non-null   int32
 1   host_is_superhost             3430 non-null   int32
 2   host_has_profile_pic          3430 non-null   int32
 3   host_identity_verified        3430 non-null   int32
 4   host_response_time            3430 non-null   int32
 5   is_location_exact             3430 non-null   int32
 6   accommodates                  3430 non-null   int64
 7   number_of_reviews             3430 non-null   int32
 8   review_scores_rating          3430 non-null   float64
 9   review_scores_accuracy        3430 non-null   int32
 10  review_scores_cleanliness     3430 non-null   int32
 11  review_scores_checkin         3430 non-null   int32
 12  review_scores_communication   3430 non-null   int32
 13  review_scores_location        3430 non-null   int32
 14  review_scores_value           3430 non-null   int32
dtypes: float64(1), int32(13), int64(1)
```

Variables Used in Multivariate Regression

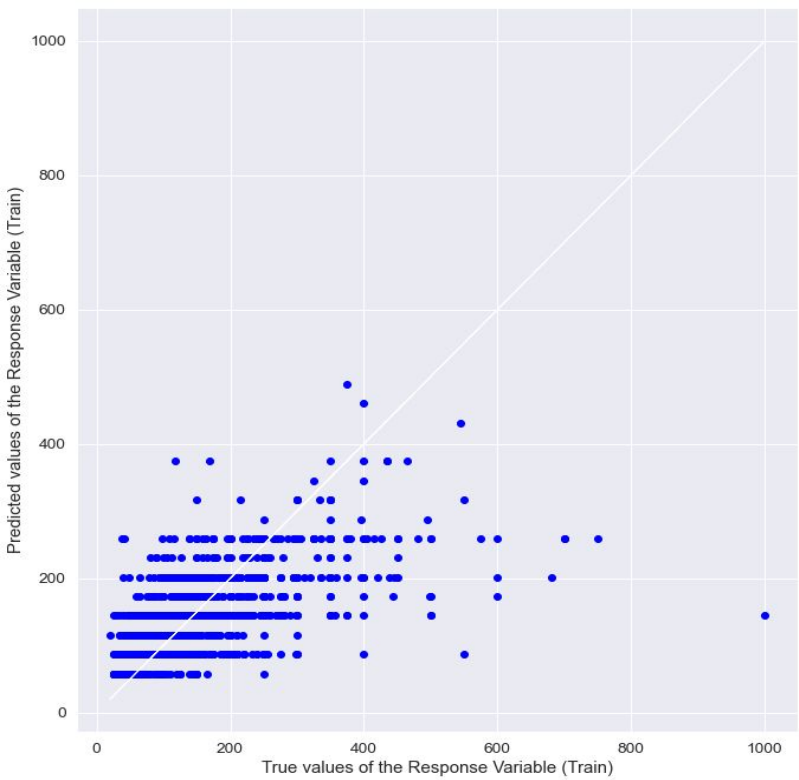
Multivariate Regression: Features vs price

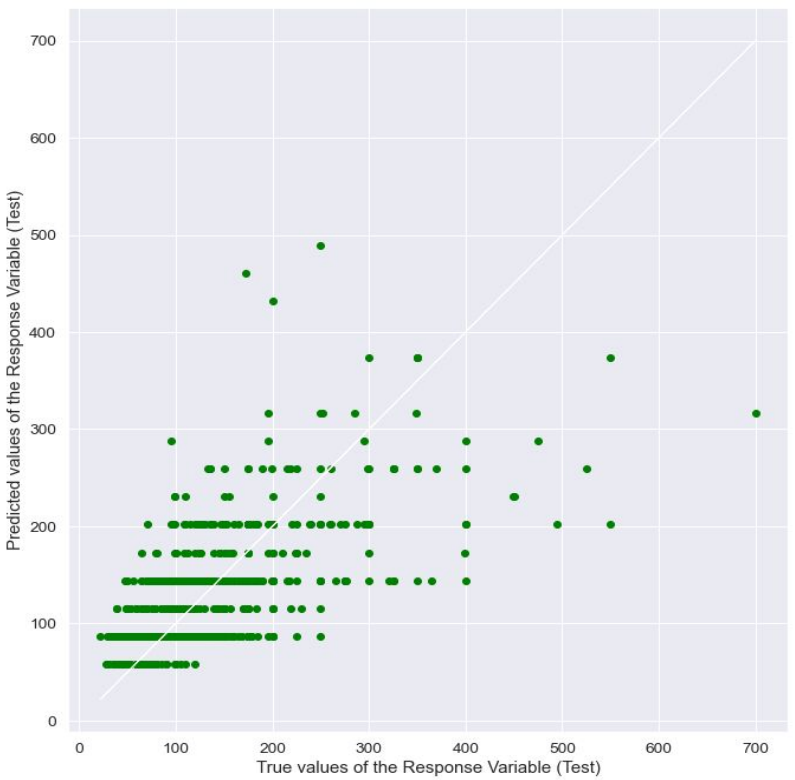


Multivariate Regression: Features vs Price

The graph depicts the importance of attributes as assessed using Linear Regression coefficients. The most essential factors in calculating the price of a listing are the number of rooms, the neighborhood group, and the number of reviews.

Univariate regression:Price vs Accommodates

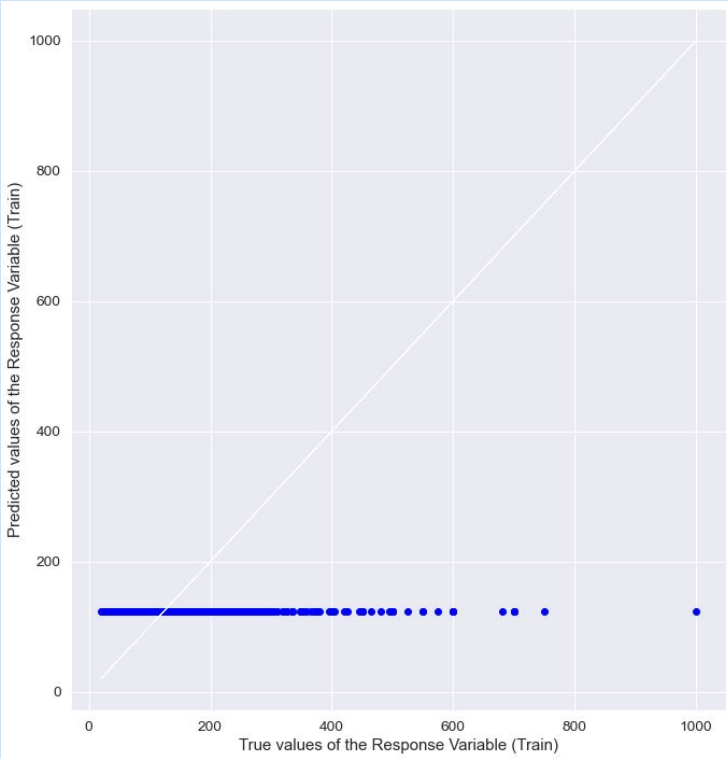


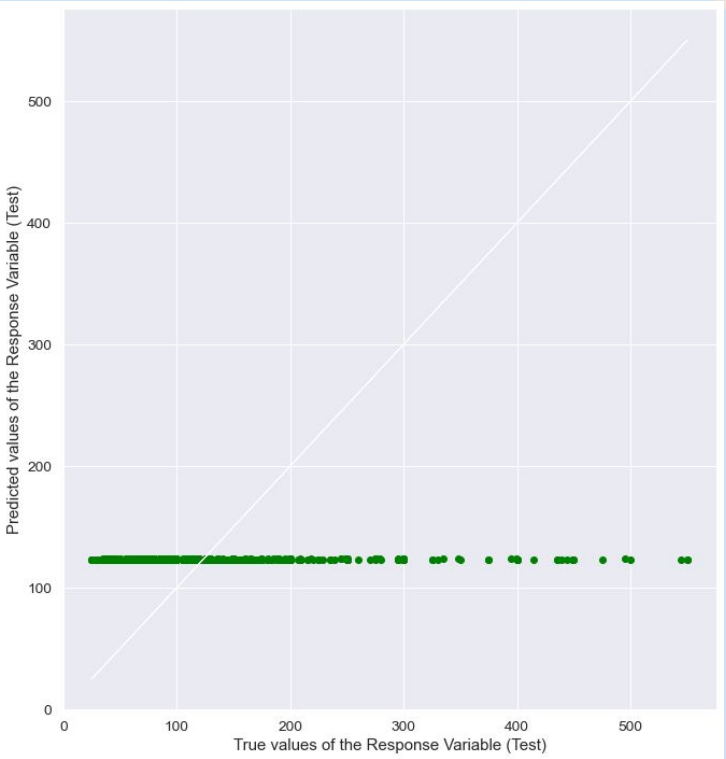


Univariate regression: Price vs Accommodates

```
Goodness of Fit of Model          Train Dataset
Explained Variance (R^2)          : 0.4167255748151911
Mean Squared Error (MSE)          : 3980.0397477050733

Goodness of Fit of Model          Test Dataset
Explained Variance (R^2)          : 0.43457336221525855
Mean Squared Error (MSE)          : 3528.5576108849796
```

Univariate regression: Price vs Neighbourhood Group

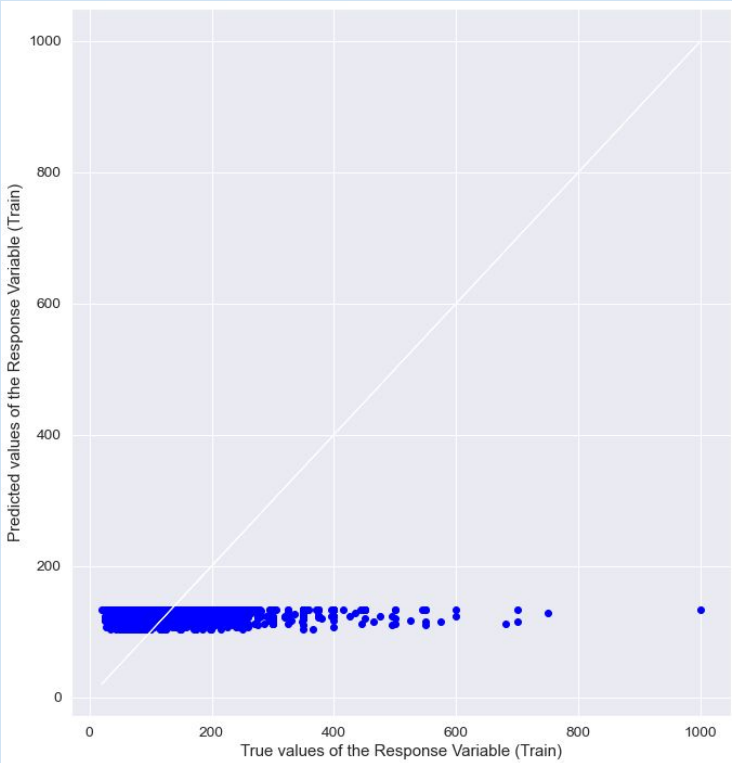


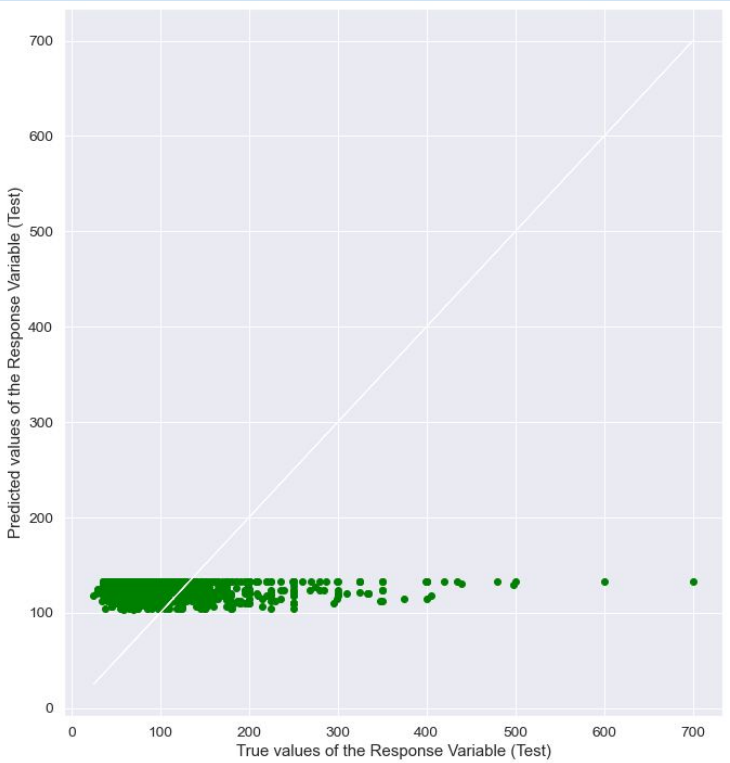


Univariate regression: Price vs Neighbourhood Group

```
Goodness of Fit of Model          Train Dataset
Explained Variance (R^2)          : 6.780549885565534e-06
Mean Squared Error (MSE)          : 6776.0556138386855

Goodness of Fit of Model          Test Dataset
Explained Variance (R^2)          : -0.0009679292014677099
Mean Squared Error (MSE)          : 6387.714078516547
```

Univariate regression: Price vs Number of Reviews





Univariate regression: Price vs Number of Reviews

```
Goodness of Fit of Model          Train Dataset
Explained Variance (R^2)          : 0.010099297449802647
Mean Squared Error (MSE)          : 6858.114232576753

Goodness of Fit of Model          Test Dataset
Explained Variance (R^2)          : 0.004405698498070265
Mean Squared Error (MSE)          : 5898.254816191866
```

Univariate regression comparison

Accommodates vs Price

```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.4167255748151911
Mean Squared Error (MSE)        : 3980.0397477050733

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.43457336221525855
Mean Squared Error (MSE)        : 3528.5576108849796
```

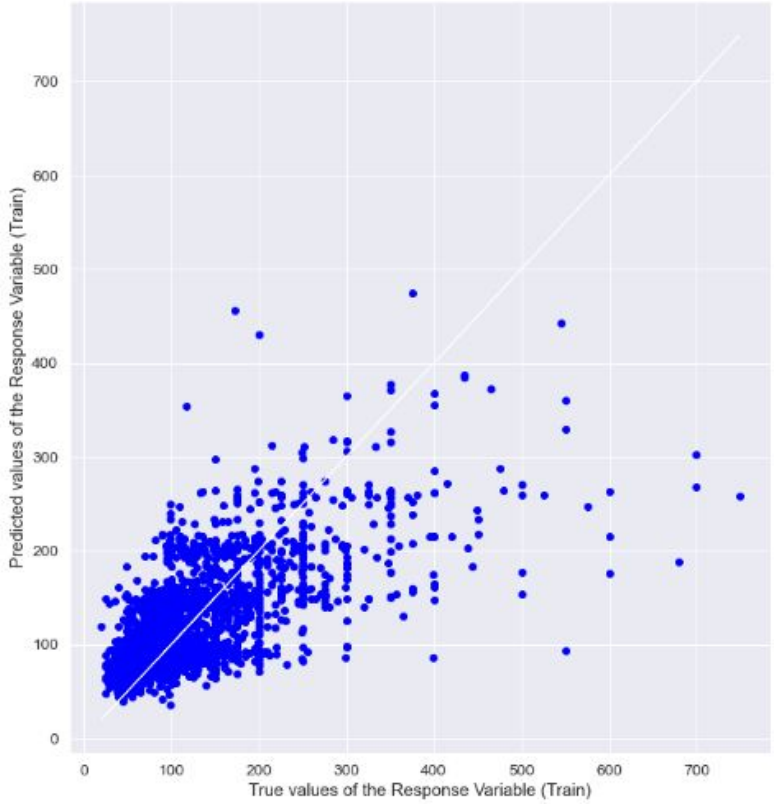
Price vs Number of Reviews

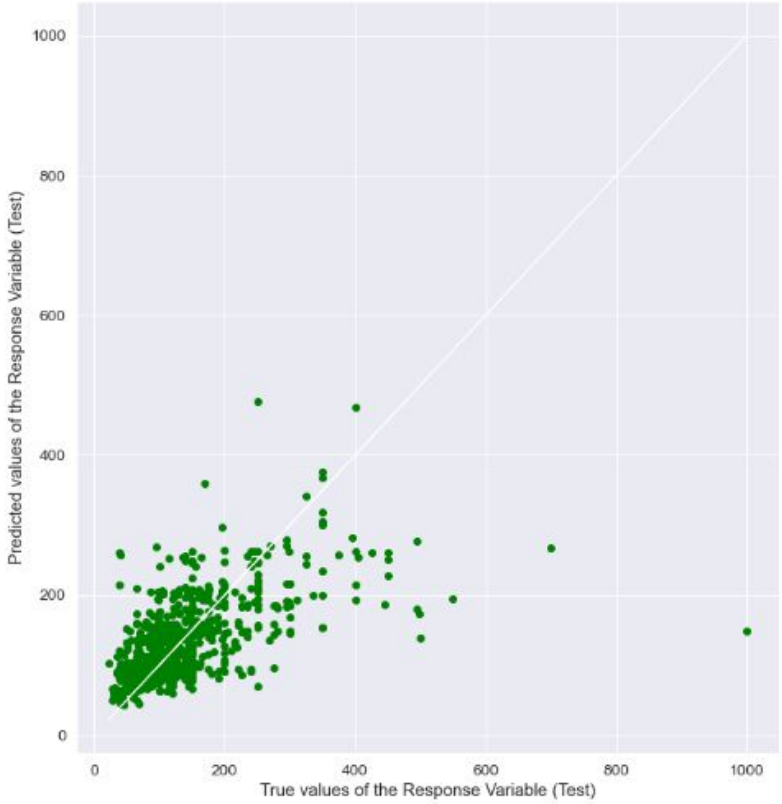
```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.010099297449802647
Mean Squared Error (MSE)        : 6858.114232576753

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.004405698498070265
Mean Squared Error (MSE)        : 5898.254816191866
```

Price vs Neighbourhood Group

```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 6.780549885565534e-06
Mean Squared Error (MSE)        : 6776.0556138386855

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : -0.0009679292014677099
Mean Squared Error (MSE)        : 6387.714078516547
```

Multivariate regression: 3 Multiple Features vs Price

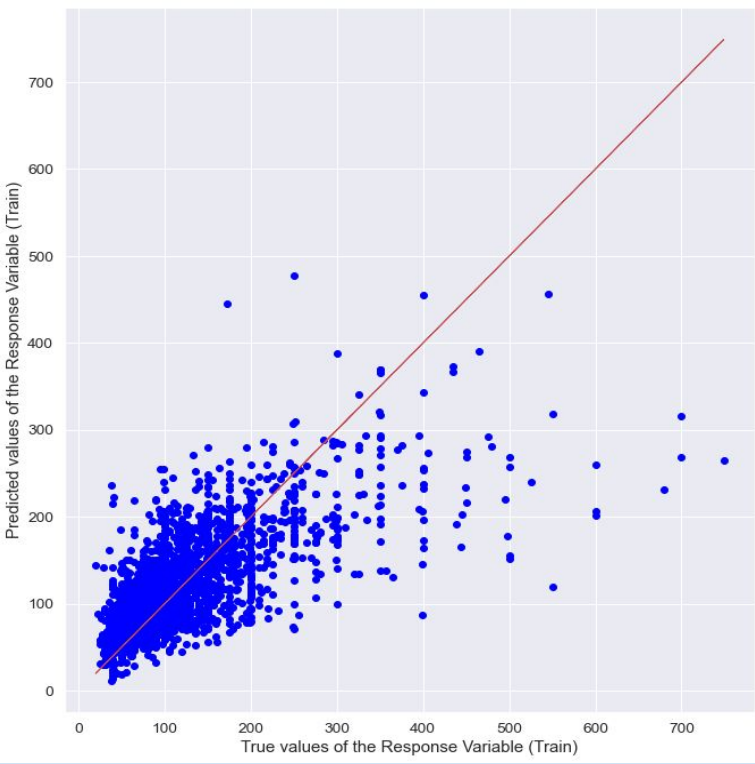


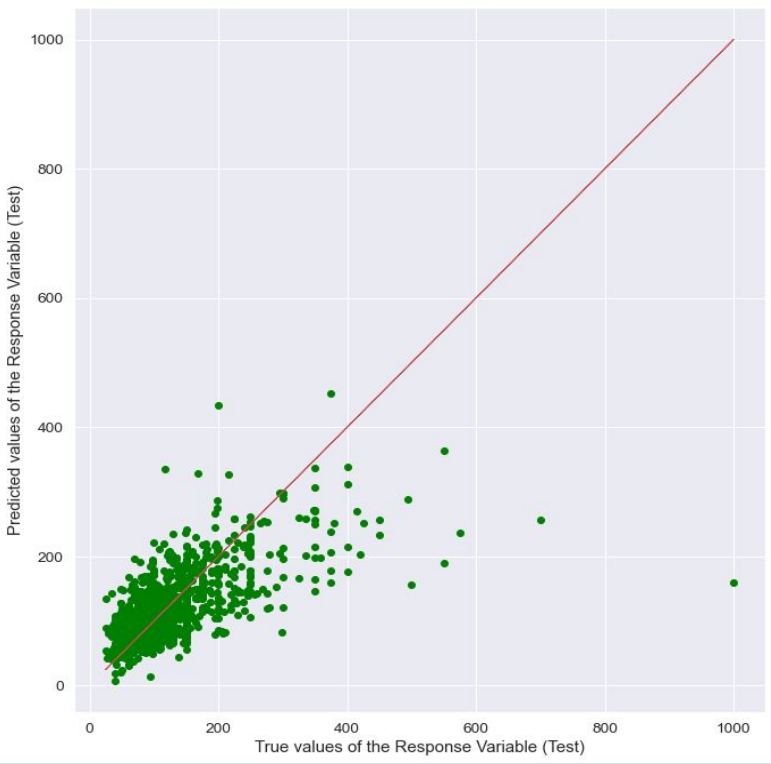


Multivariate regression: 3 Multiple Features vs Price

```
Goodness of Fit of Model          Train Dataset
Explained Variance (R^2)          : 0.4499745637924851
Mean Squared Error (MSE)          : 3520.254015420042

Goodness of Fit of Model          Test Dataset
Explained Variance (R^2)          : 0.38608173959846726
Mean Squared Error (MSE)          : 4604.068900781523
```

Multivariate regression: All Multiple Features vs Price





Multivariate regression: All Multiple Features vs Price

```
Goodness of Fit of Model          Train Dataset
Explained Variance (R^2)          : 0.477219833817900 56
Mean Squared Error (MSE)          : 3518.4876420771743

Goodness of Fit of Model          Test Dataset
Explained Variance (R^2)          : 0.4381180857550053
Mean Squared Error (MSE)          : 3684.0665221282284
```

Multivariate regression comparison

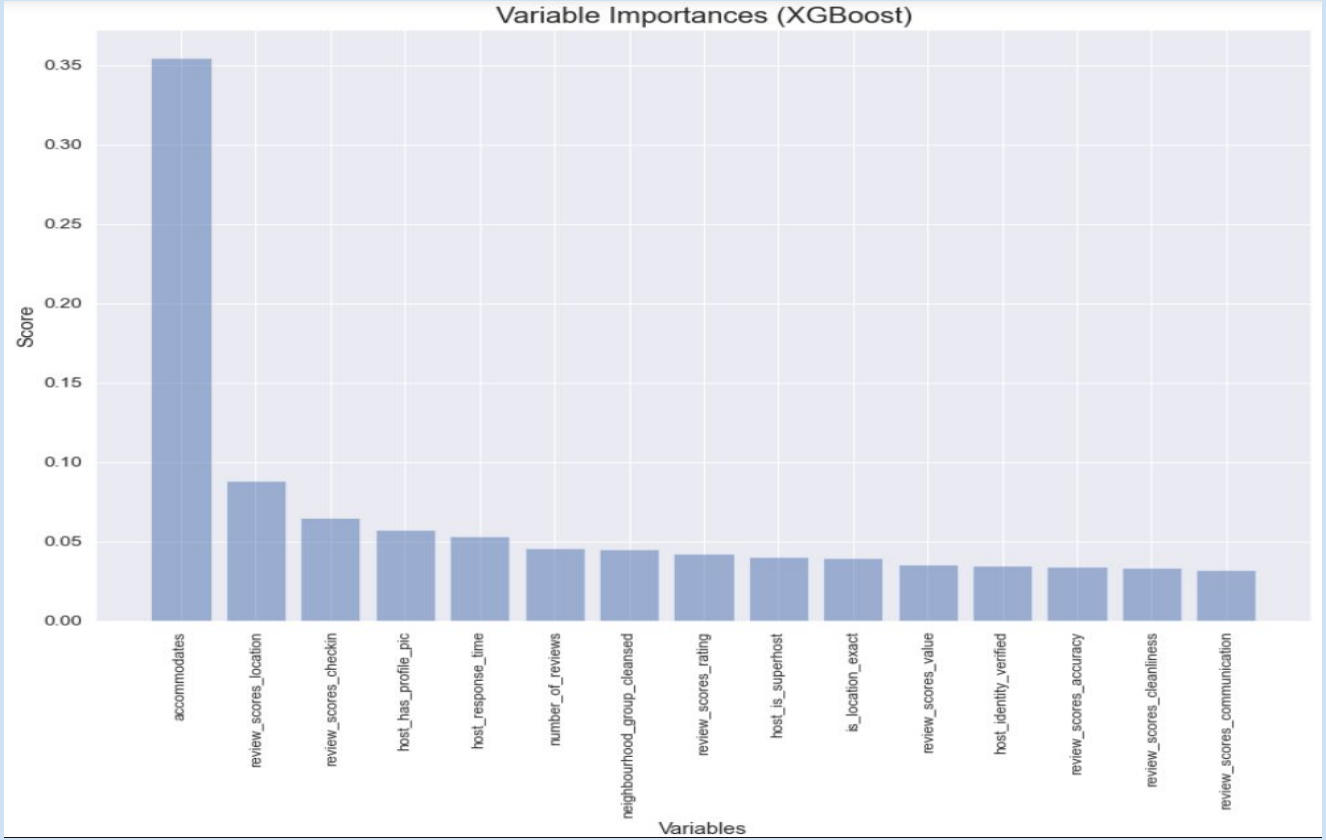
All Multiple Features vs Price

```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.47721983381790056
Mean Squared Error (MSE)        : 3518.4876420771743

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.4381180857550053
Mean Squared Error (MSE)        : 3684.0665221282284
```

3 Multiple Features (Accommodates, Number of review, Neighbourhood group cleansed) vs Price

```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.4499745637924851
Mean Squared Error (MSE)        : 3520.254015420042

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.38608173959846726
Mean Squared Error (MSE)        : 4604.068900781523
```

XGboost Regressor: Most Important Feature



Conclusion

- Verification of profile
  - (superhost, verified account & accurate location)
- Peak periods
  - July - September
- Popular accommodates
  - 1 - 4 rooms

Conclusion

- Reviews of profile & listings
  - Most important is communication
  - Other factors are equally important as well
- Linear regression
  - Allows an estimate for listing prices for host' reference
- Importance ranking for AirBnB host
  1. Accommodates
  2. Neighbourhood
  3. Number of reviews