AirBNB Pricing Report by Mike Li

Problem:

As we see positive reports on economy in 2018, price inflation has also been creeping up across US. For many homeowners who deal with burden from ever increasing living expenses, Airbnb serves as a viable source of partial income. However it's a big decision to open up your house for rental. Goal of this analysis is to serve as a useful guideline on setting the right price for the rental and evaluate if he/she can expect a profit.

Target Audiences:

Target audiences are the host who is looking to get advice on the best listing price to optimize their profit while to ensure the unit will be leased in reasonable time. The homeowner can derive the overall gross profit based on the rental price and the duration. Factoring in the costs (eg mortgage payments, insurance, etc), he/she can derive the net profit and make the final decision whether it's worth to make the house available for short-term rental or it is better to repurpose.

Dataset:

As of now, Airbnb does not release any data on the listings under the site, a separate group "Inside Airbnb" has extracted data on a sample of the listings for many of the major cities on the website. The analysis is based on the 8/16/18 Seattle data that is made available by Inside AirBNB site. Here is the link.

The analysis will mainly use data from 5 areas:

- Basic data such as the address/location, neighborhood and pricing (price, weekly_price, monthly_price, security_deposit, cleaning_fee)
- Rental's features such as the property type, room type, bathrooms, bedrooms, square_feet, etc
- Policy data such as security deposit, cleaning fee, guests_included, extra_people, minimum_nights, maximum_nights, cancellation_policy
- Host-related data such as 'host_since', 'host_is_superhost', 'host_location', 'host_response_time', 'host_response_rate', etc
- Review data such as price', 'number_of_reviews', 'first_review', 'last_review', 'review_scores_rating', 'review_scores_accuracy',

Data Cleansing:

To prepare the data for EDA, a subset of the available data is identified for the analysis. The data is mainly attributes around 4 areas: the rental configurations, the pricing data, the host data and policies such as security deposit, cleaning fee, etc.

The subset of data is then cleansed before the EDA. The data cleansing includes transforming dollar amounts to numeric values, filling empty pricing amounts (price, weekly price, etc) with 0. Values with bad formats are removed (eg postal code).

The amenities column is being parsed to a feature matrix for analysis. Some attributes are formatted to corresponding data types to aid the analysis. 3 date attributes (host_since, first_review, last_review) is assigned with date type and attributes property_type, room_type, bed_type and cancellation_policy is assigned with category type.

Finally, few outlier data points above rental price $1650 are identified and removed from the final data set.

Initial Findings

Summary:

1) Seattle listings have mean price around 150 with 75% of listing below 200.

2) Location wise, there is a heavy cluster of listings around Capitol Hill area. No specific area stands out as higher-priced area.

3) Features wise, it seems price has higher correlations with accommodates, bedrooms and square_feet.

4) Policy wise, only cleaning fee seems to have impace on listing price and the fee mostly varies between 0 to 300 over different levels of listing prices.

5) Host attributes and Review scores do not seem to have strong correlation with the listing prices

The Analysis:

Initial Statistics:

The mean is around $150 with total records of 8417. The distribution is shown as below

```
print(r_cleaned['price'].describe())
```

```
count    8417.000000
mean      150.919568
std       114.207548
min         0.000000
25%        80.000000
50%       119.000000
75%       189.000000
max      1500.000000
Name: price, dtype: float64
```

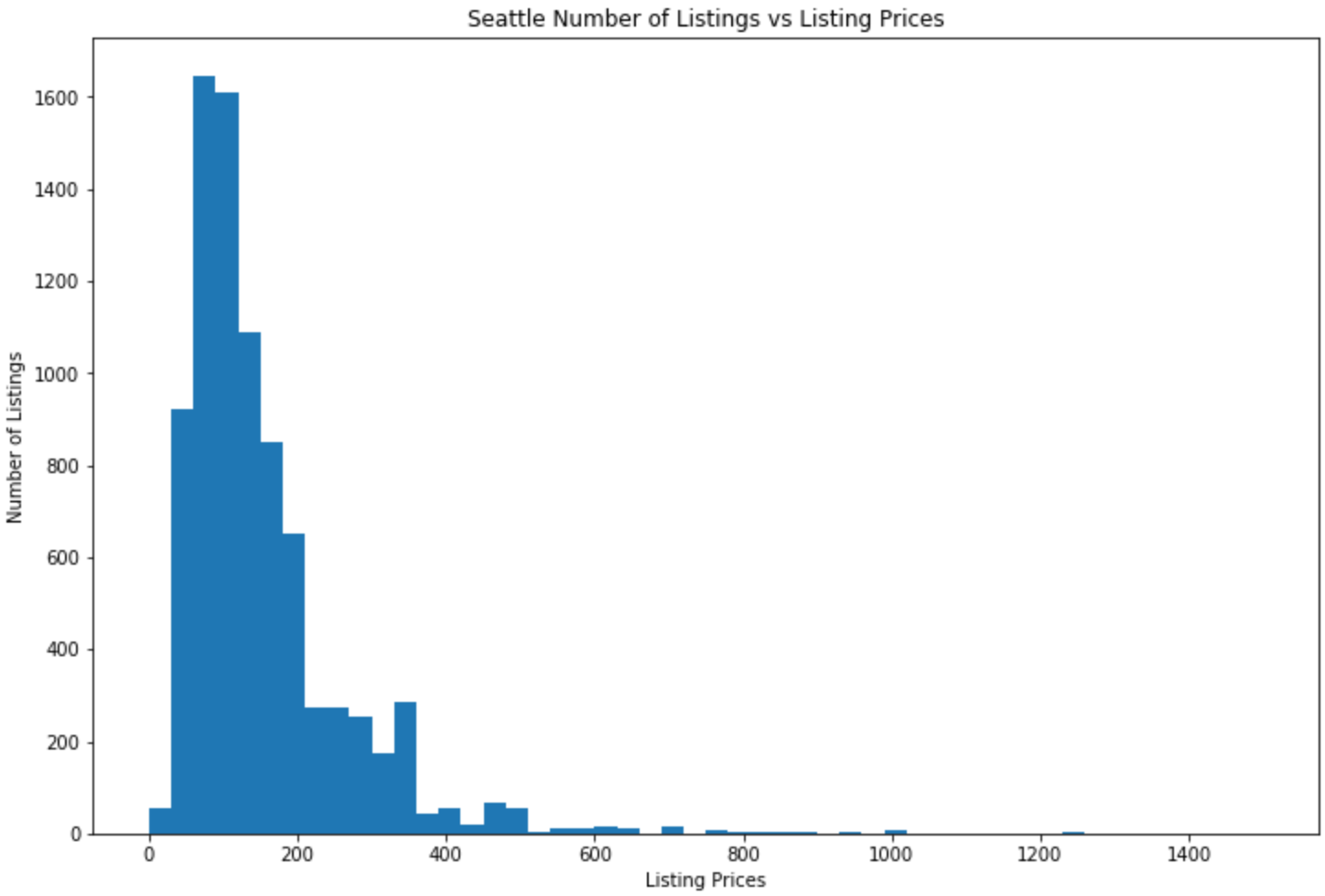
Analysis on any trend between number of listings and listing prices?

Conclusion:

1) Listing prices are highly concentrated around the mean price around $150

Approach

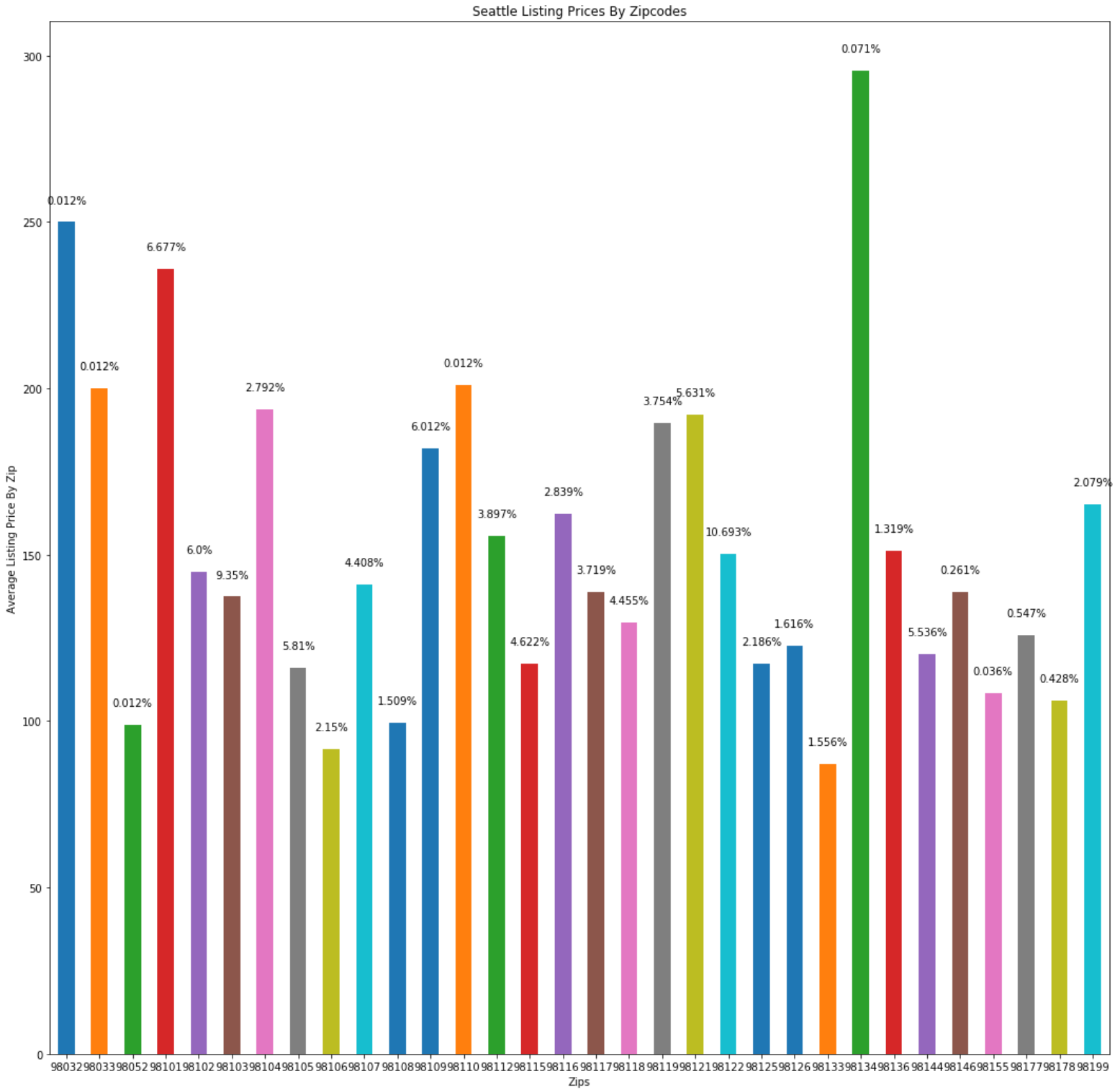
1) Break down price ranges to 50 buckets
2) Build dataframe with num of listings vs average price under each price bucket
3) Plot a histogram with these data and try to identify correlations
4) Identify further questions from the analysis



Analysis on the Locations

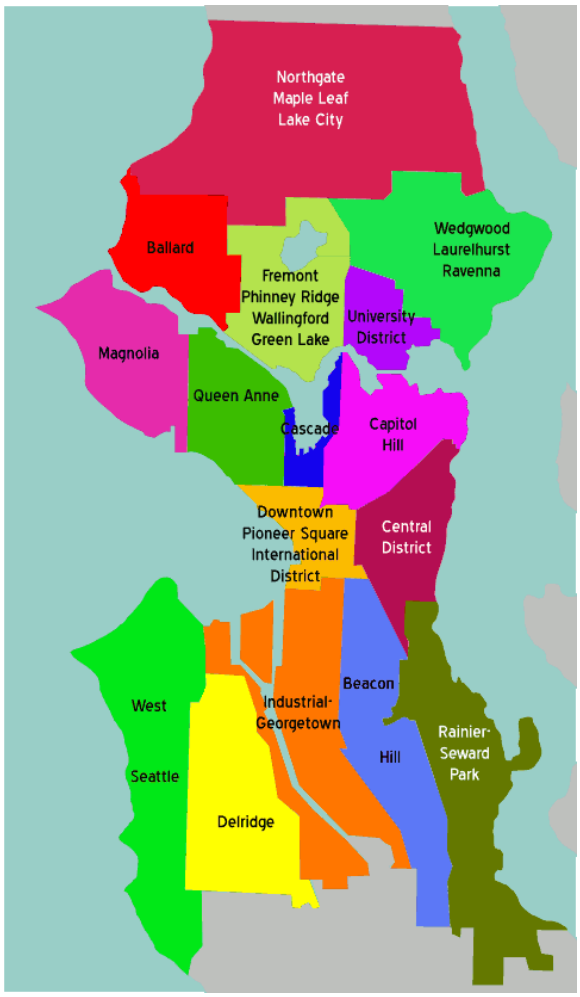
Initial Observations

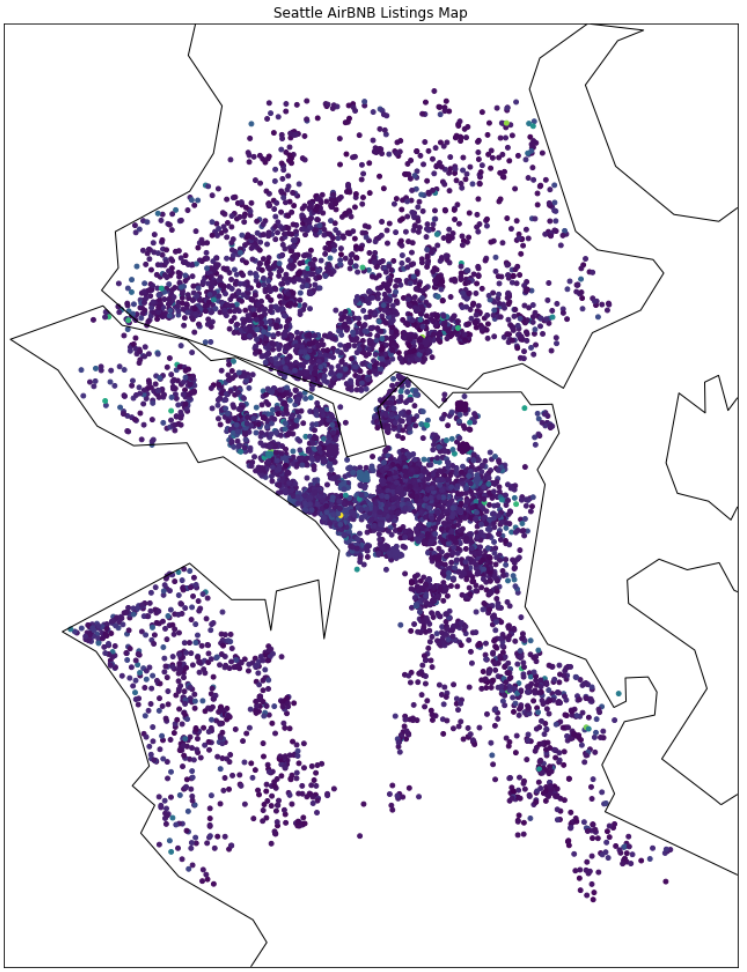
1) From the chart result, listing prices are in a narrow range except for a specific zip 98134
2) Zip 98122 contains the most listings while zip 98101 has the highest mean price



Seattle AirBNB Listings Distribution

1) From below plots, we can see heavy cluster of listings around the Capitol Hill and Downtown neighborhoods.
2) The mean price goes from the highest in the Downtown neighborhood to the lowest in the Delridge neighborhood


Northgate
Maple Leaf
Lake City
Ballard
Wedgwood
Laurelhurst
Ravenna
Fremont
Phinney Ridge
Wallingford
Green Lake
University
District
Magnolia
Queen Anne
Cascade
Capitol
Hill
Downtown
Pioneer Square
International
District
Central
District
West
Seattle
Industrial-
Georgetown
Beacon
Hill
Rainier-
Seward
Park
Delridge


Seattle AirBNB Listings Map

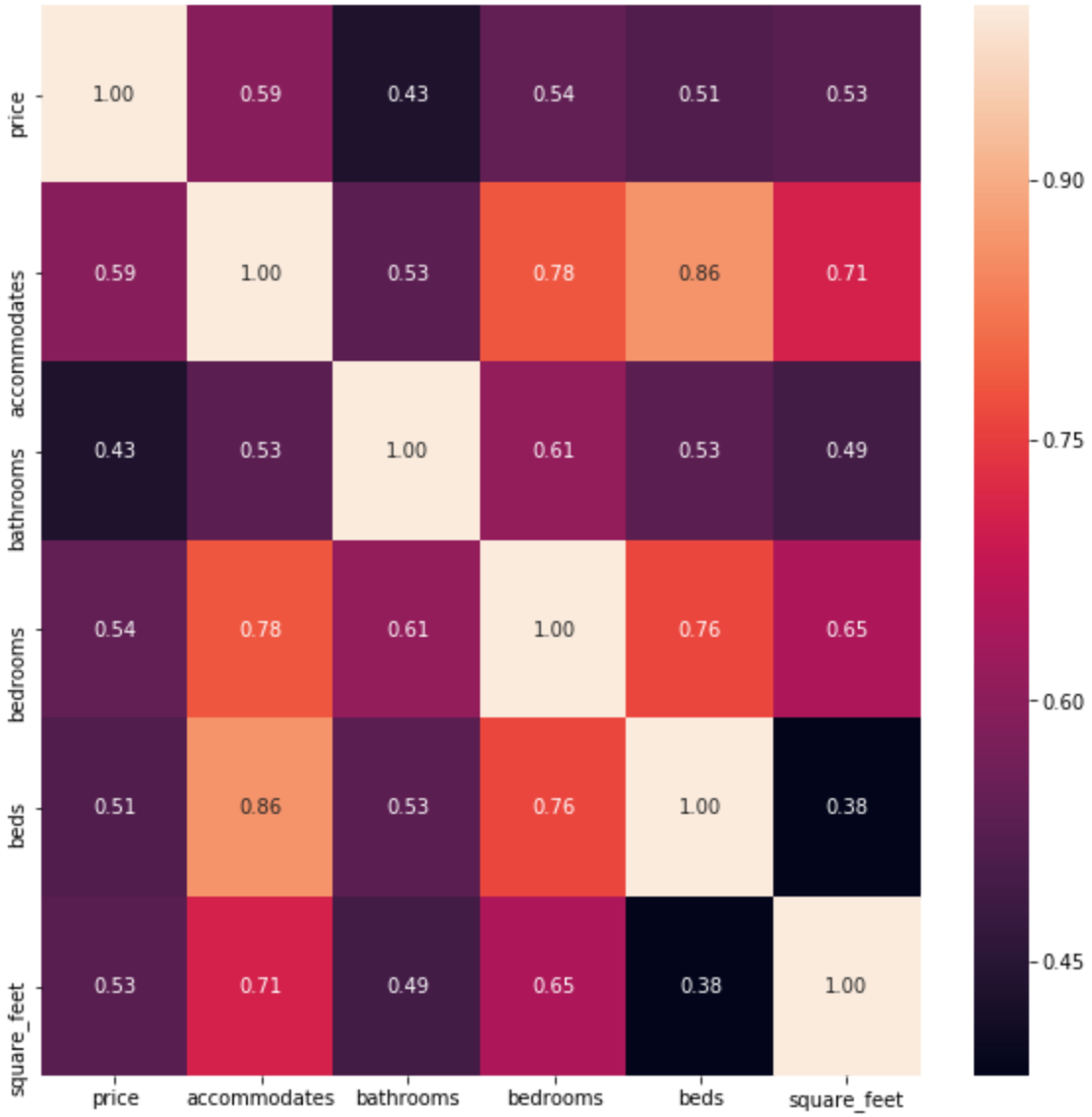
Analysis on the Features

Observation 1:

1) From below plot, it seems Price has higher correlations with accommodates, bedrooms and square_feet.

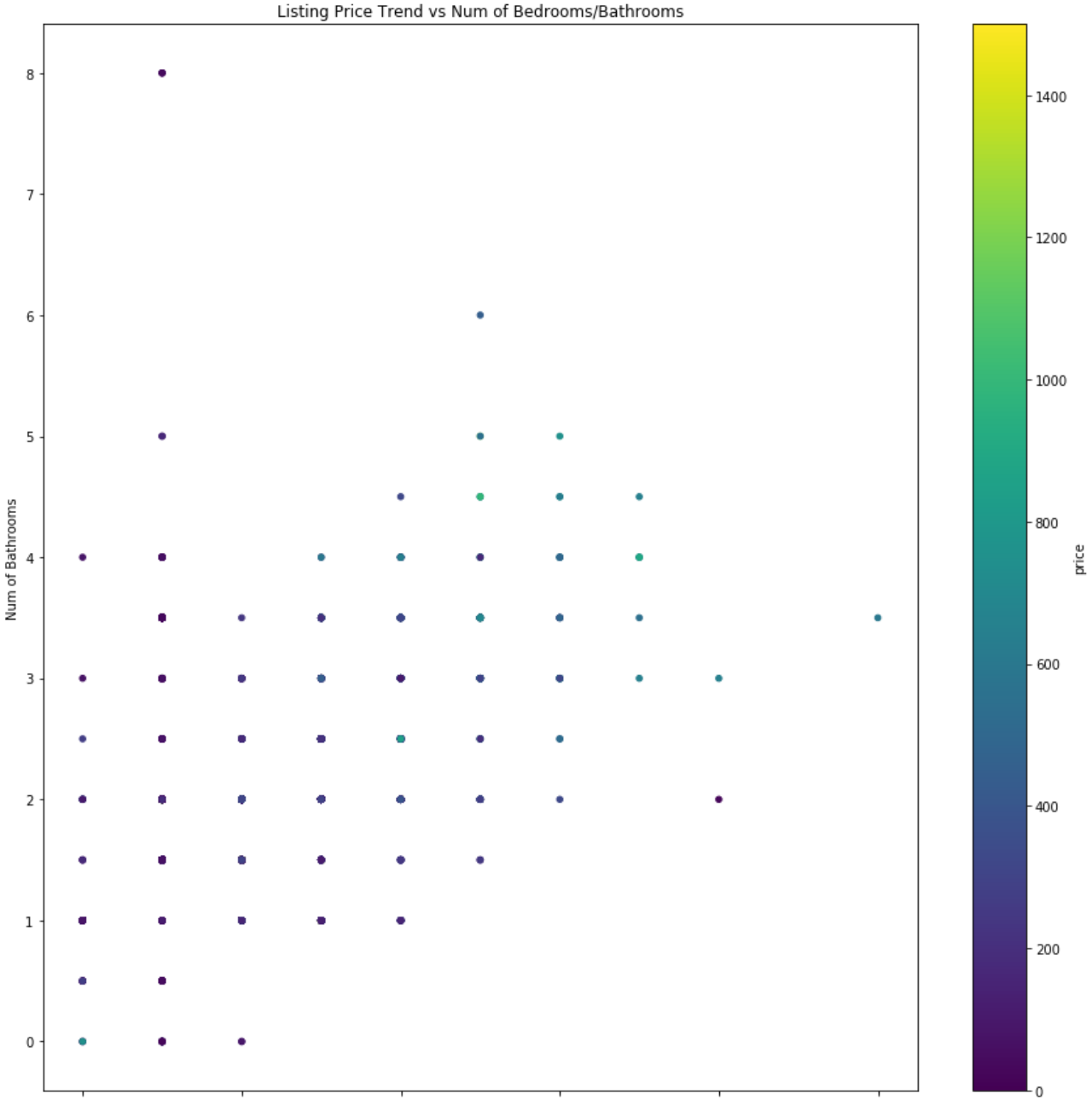
	price	accommodates	bathrooms	bedrooms	beds	square_feet
price	1.000000	0.594116	0.433696	0.538710	0.514003	0.525060
accommodates	0.594116	1.000000	0.531002	0.781323	0.864283	0.708531
bathrooms	0.433696	0.531002	1.000000	0.612668	0.531016	0.493908
bedrooms	0.538710	0.781323	0.612668	1.000000	0.761920	0.648296
beds	0.514003	0.864283	0.531016	0.761920	1.000000	0.384354
square_feet	0.525060	0.708531	0.493908	0.648296	0.384354	1.000000

Below is a heat map that illustrates above data:



Observation 2:

1) From below plot, it seems price trends higher when number of bedrooms is 3 or more

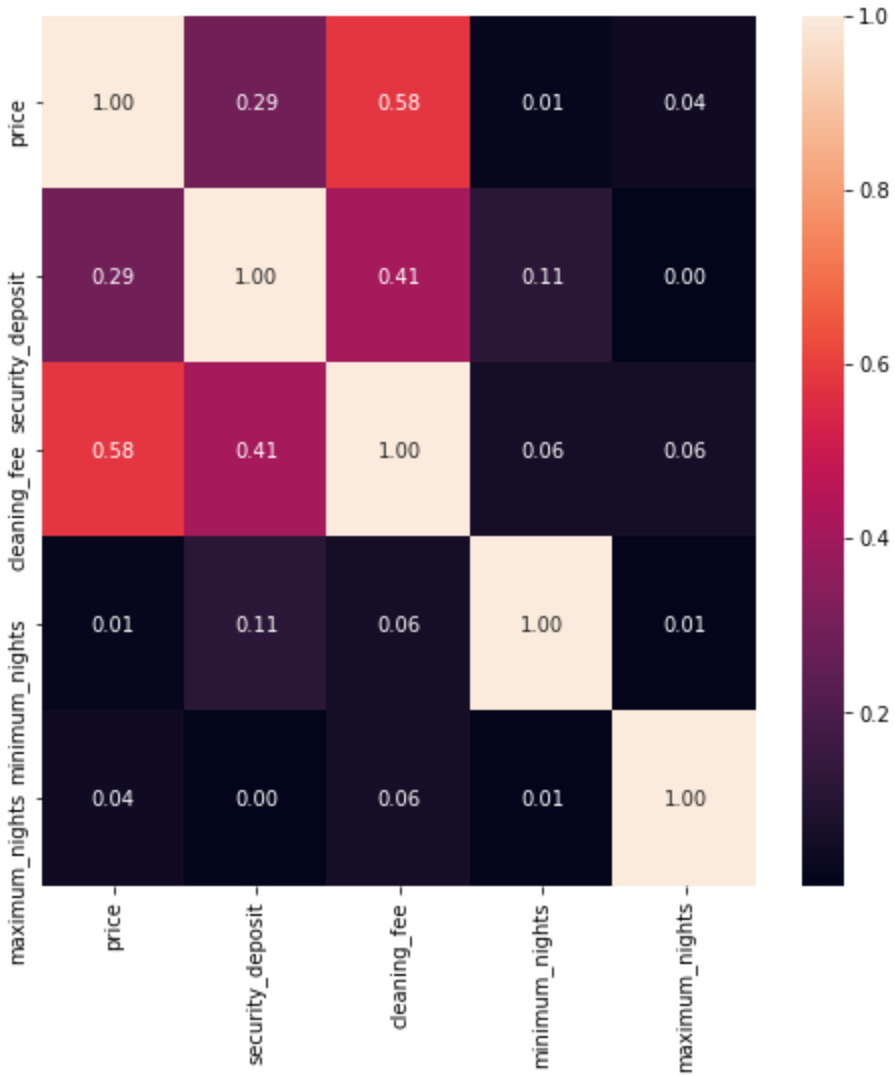


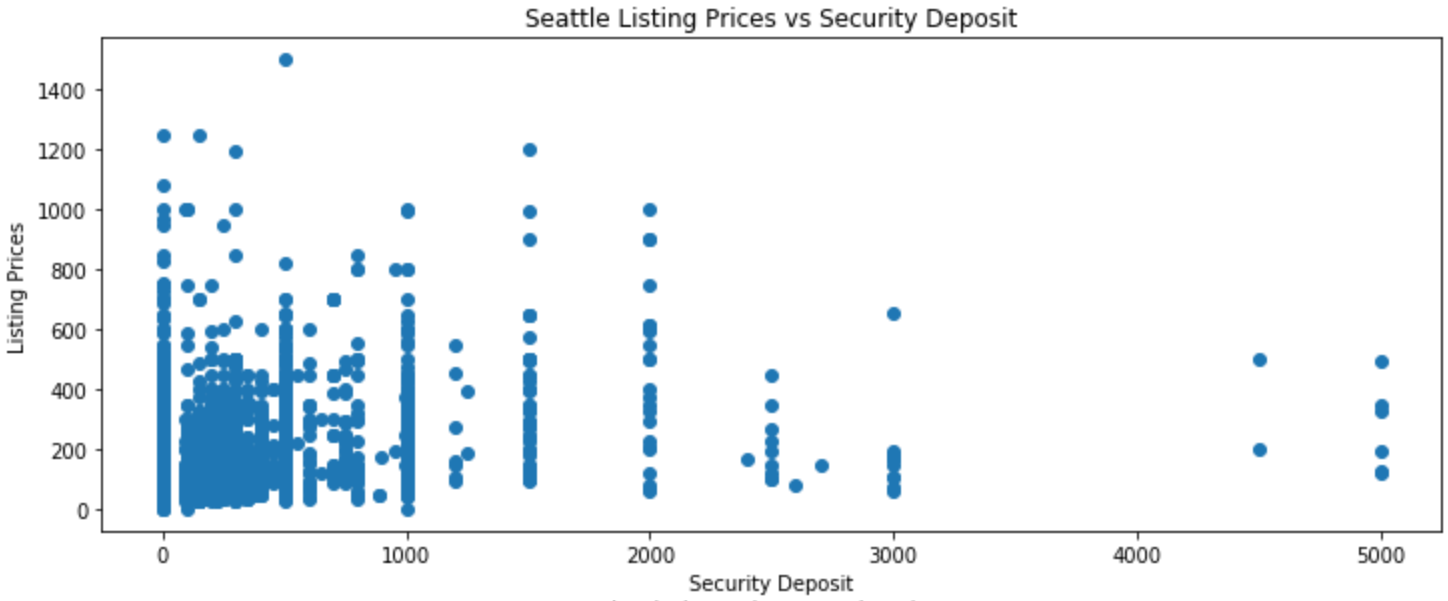
Analysis on Policy data (eg Security Deposit and Cleaning Fee)

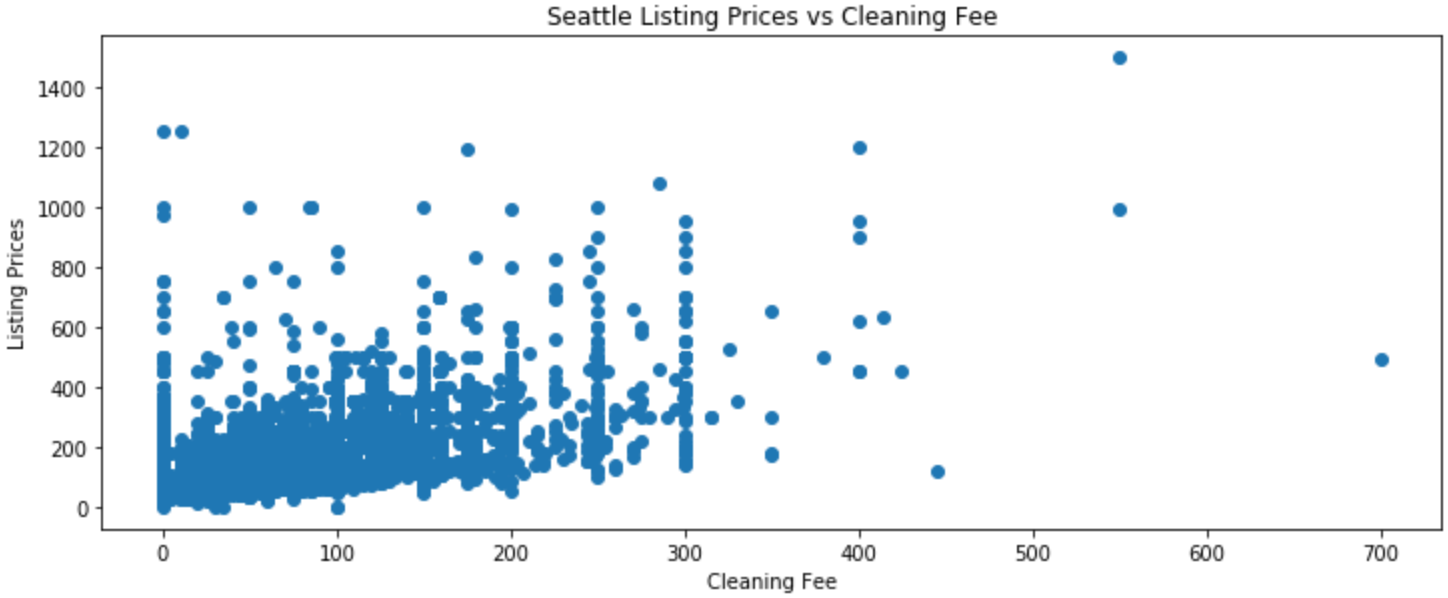
Observation 1:

1) From the analysis, cleaning Fee is only policy data that seems to have impact on listing price.
2) The cleaning fee mostly varies between 0 to 300 over different levels of listing prices.

	price	security_deposit	cleaning_fee	minimum_nights	maximum_nights
price	1.000000	0.292949	0.582248	0.014751	0.040795
security_deposit	0.292949	1.000000	0.414735	0.111534	0.001236
cleaning_fee	0.582248	0.414735	1.000000	0.056225	0.056962
minimum_nights	0.014751	0.111534	0.056225	1.000000	0.007729
maximum_nights	0.040795	0.001236	0.056962	0.007729	1.000000

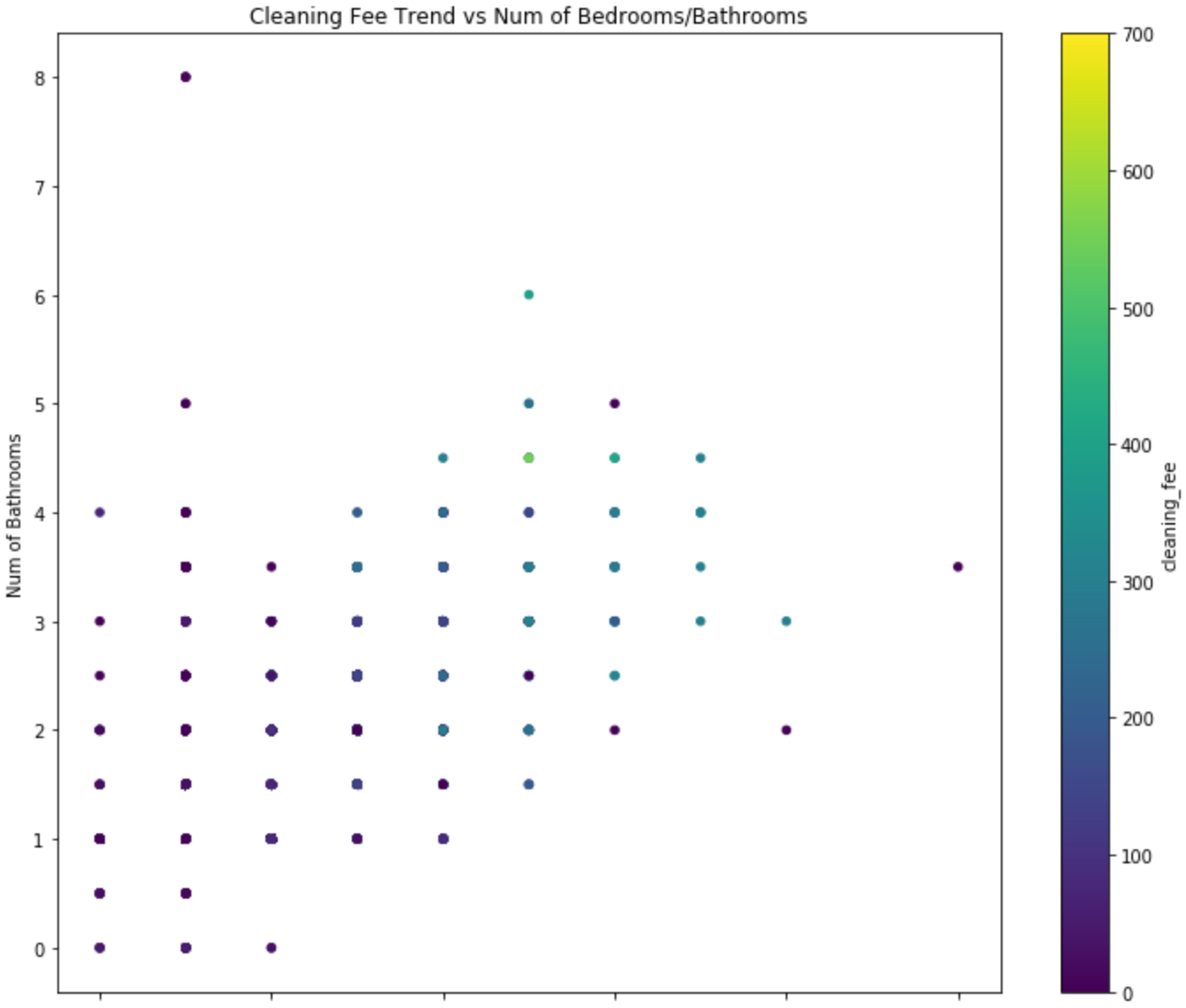






Cleaning Fee vs Bedroom/Bathrooms

1) Num of Bedrooms has more impact on Cleaning Fee than num of bathrooms (based on color changes vs num of rooms)


Cleaning Fee Trend vs Num of Bedrooms/Bathrooms
Num of Bathrooms
0
1
2
3
4
5
6
7
8
cleaning_fee
0
100
200
300
400
500
600
700

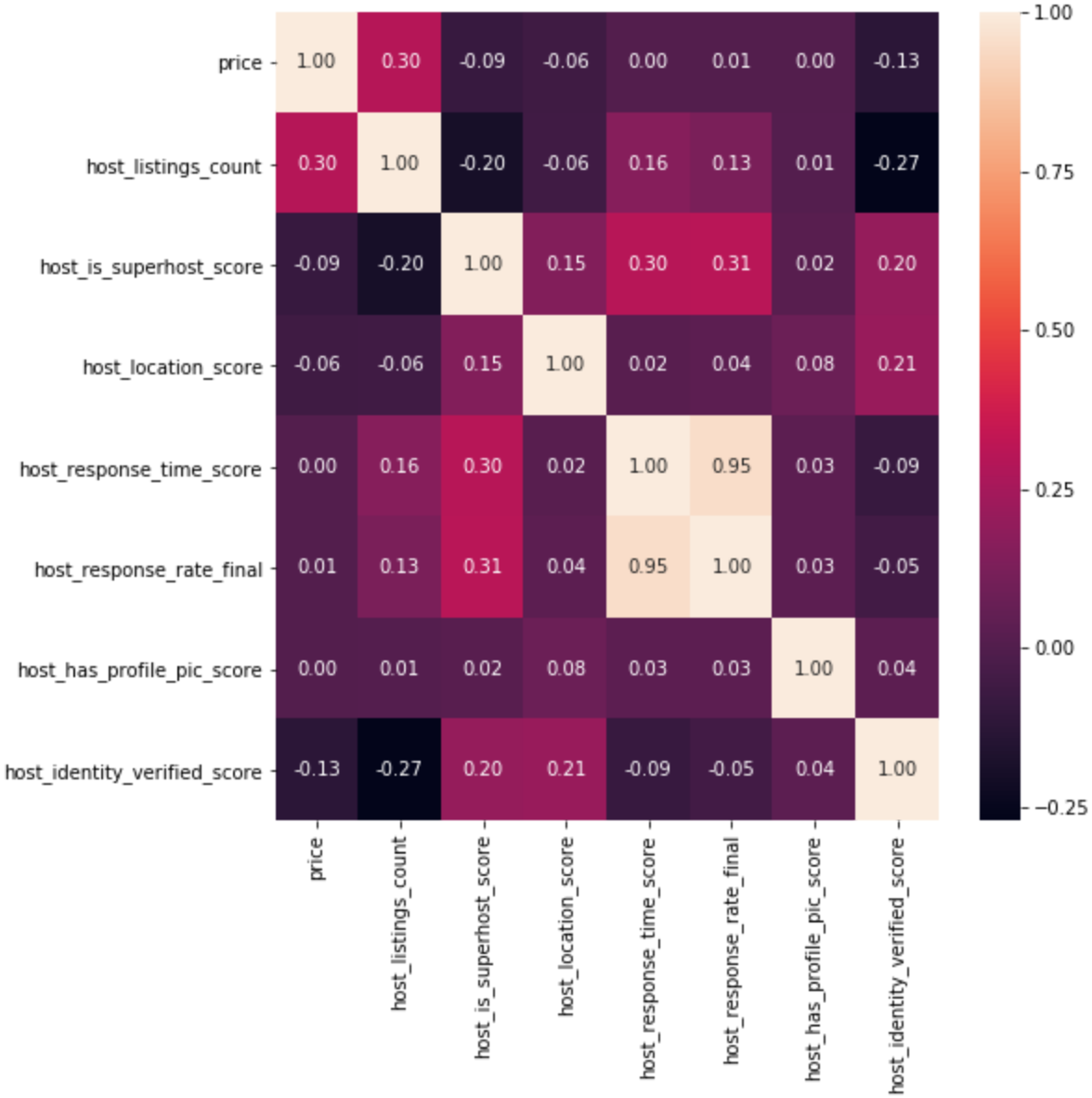
Analysis on the Host

Conclusion:

1) After examination of different hosts' attributes against the price, none of the attributes has a significant correlation on the price.

Analysis Details

1) Build correlations with all host properties against price (assign scores for some categorical data)
2) The values of interest:
a) host_since (if above 0.5 correlation, derive years value and find if above certain value impacts price)
b) is_superhost (t - super, f - not)
c) host_location (derive numeric scale for in_city, in_state, in_country, out_country)
d) host_response_time (convert to numeric scale:
e) host_response_rate (remove % sign)
f) host_listings_count (above or below certain count will impact price?)
g) host_has_profile_pic
h) host_identity_verified



Analysis on the Reviews

Conclusion:

1) After examination of different review attributes against the price, none of the attributes has a significant correlation on the price.