Module 3 Final Project – Milestone 1: EDA

JinYoung Lee, Peter Yang, Ruitong Guo, Chaitra Naik, Snehal Sham Tikone

Northeastern University - Seattle

ALY 6040: Data Mining

Dr. Shanu Sushmita

May 2nd, 2022

Introduction

The data set we are going to investigate is the “Seattle Airbnb Open Data” (“Listings.csv”). The data set contains 92 features with 3,818 entries. To perform the exploratory data analysis, following questions are addressed in the data set study:

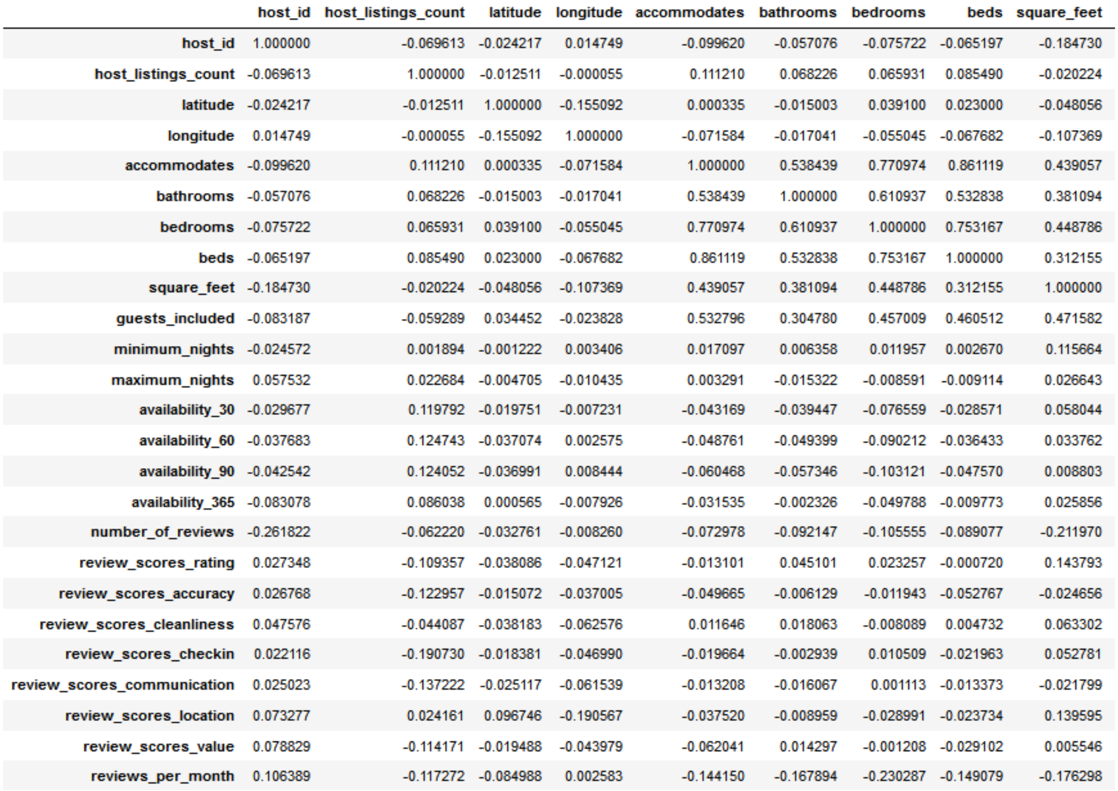
- What did you do with the data in the context of exploration?
- How many entries are in the dataset?
- Was there missing data? Duplications? How clean was the data?
- Were there outliers or suspicious data?
- What did you do to clean the data?
- What did you find? What intrigued you about the data? Why does that matter?
- What would your proposed next steps be?
- What business questions do you plan to answer with your data mining?

Analysis

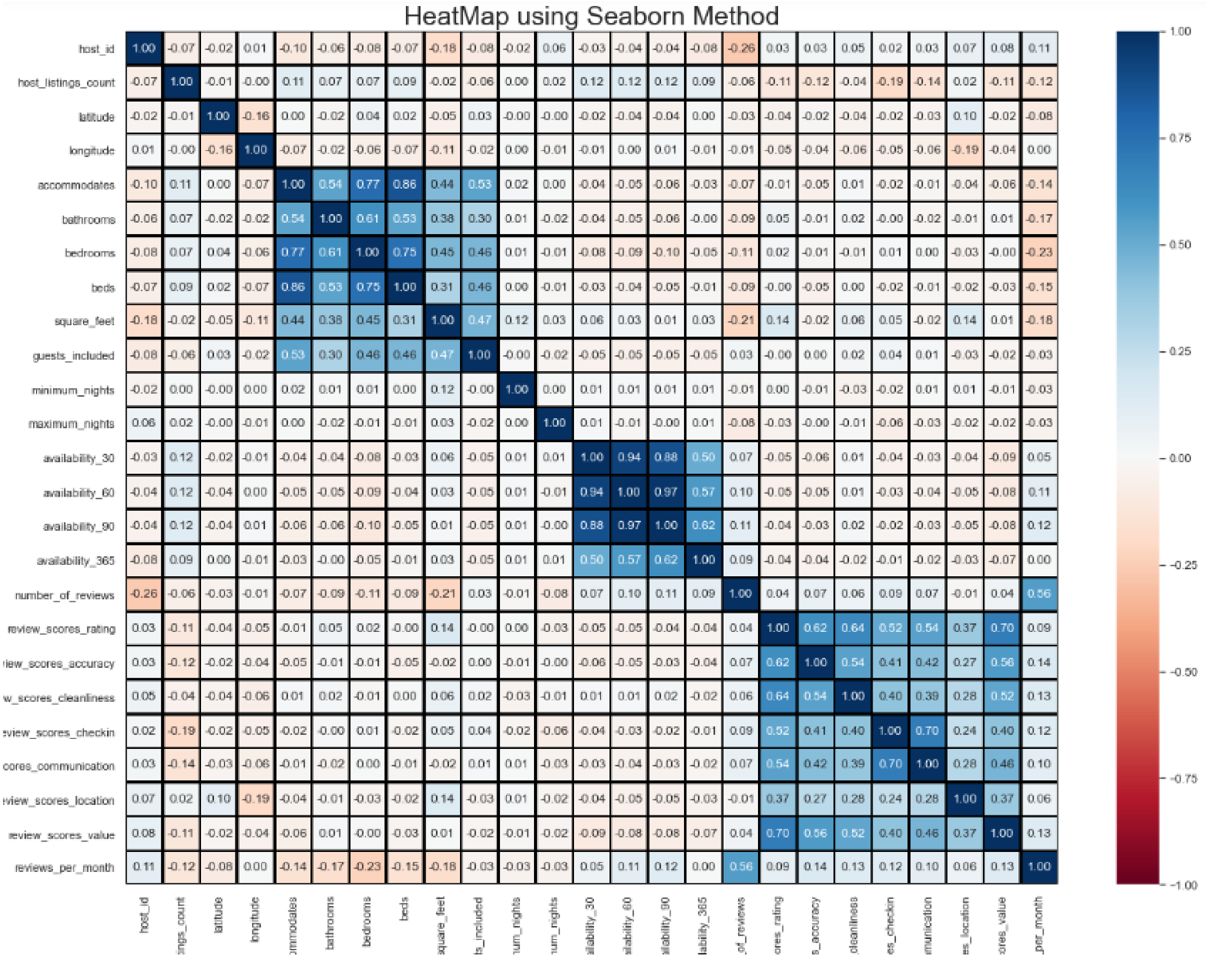
★ **What business questions do you plan to answer with your data mining?**

- We want to understand whether the reviews support the ratings or not. Is there any discrepancy between ratings and reviews?

We are going to dive into the correlation section to see which factor had the most impact on “review_scores_accuracy” by using correlation charts along with heatmap.



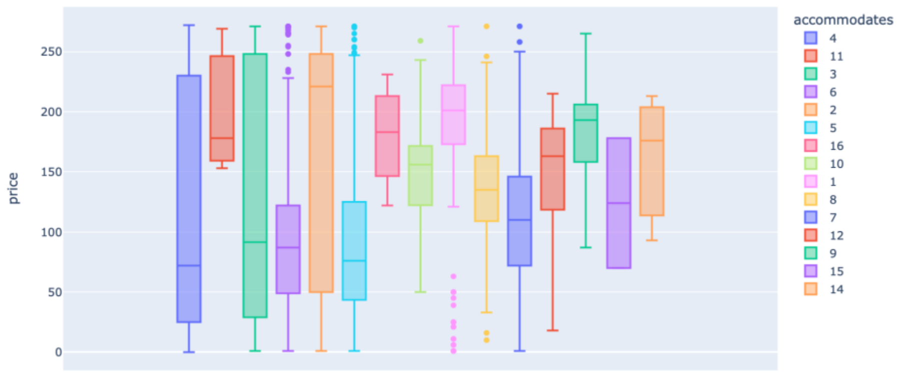
	host_id	host_listings_count	latitude	longitude	accommodates	bathrooms	bedrooms	beds	square_feet
host_id	1.000000	-0.069613	-0.024217	0.014749	-0.099620	-0.057076	-0.075722	-0.065197	-0.184730
host_listings_count	-0.069613	1.000000	-0.012511	-0.000055	0.111210	0.068226	0.065931	0.085490	-0.020224
latitude	-0.024217	-0.012511	1.000000	-0.155092	0.000335	-0.015003	0.039100	0.023000	-0.048056
longitude	0.014749	-0.000055	-0.155092	1.000000	-0.071584	-0.017041	-0.055045	-0.067682	-0.107369
accommodates	-0.099620	0.111210	0.000335	-0.071584	1.000000	0.538439	0.770974	0.861119	0.439057
bathrooms	-0.057076	0.068226	-0.015003	-0.017041	0.538439	1.000000	0.610937	0.532838	0.381094
bedrooms	-0.075722	0.065931	0.039100	-0.055045	0.770974	0.610937	1.000000	0.753167	0.448786
beds	-0.065197	0.085490	0.023000	-0.067682	0.861119	0.532838	0.753167	1.000000	0.312155
square_feet	-0.184730	-0.020224	-0.048056	-0.107369	0.439057	0.381094	0.448786	0.312155	1.000000
guests_included	-0.083187	-0.059289	0.034452	-0.023828	0.532796	0.304780	0.457009	0.460512	0.471582
minimum_nights	-0.024572	0.001894	-0.001222	0.003406	0.017097	0.006358	0.011957	0.002670	0.115664
maximum_nights	0.057532	0.022684	-0.004705	-0.010435	0.003291	-0.015322	-0.008591	-0.009114	0.026643
availability_30	-0.029677	0.119792	-0.019751	-0.007231	-0.043169	-0.039447	-0.076559	-0.028571	0.058044
availability_60	-0.037683	0.124743	-0.037074	0.002575	-0.048761	-0.049399	-0.090212	-0.036433	0.033762
availability_90	-0.042542	0.124052	-0.036991	0.008444	-0.060468	-0.057346	-0.103121	-0.047570	0.008803
availability_365	-0.083078	0.086038	0.000565	-0.007926	-0.031535	-0.002326	-0.049788	-0.009773	0.025856
number_of_reviews	-0.261822	-0.062220	-0.032761	-0.008260	-0.072978	-0.092147	-0.105555	-0.089077	-0.211970
review_scores_rating	0.027348	-0.109357	-0.038086	-0.047121	-0.013101	0.045101	0.023257	-0.000720	0.143793
review_scores_accuracy	0.026768	-0.122957	-0.015072	-0.037005	-0.049665	-0.006129	-0.011943	-0.052767	-0.024656
review_scores_cleanliness	0.047576	-0.044087	-0.038183	-0.062576	0.011646	0.018063	-0.008089	0.004732	0.063302
review_scores_checkin	0.022116	-0.190730	-0.018381	-0.046990	-0.019664	-0.002939	0.010509	-0.021963	0.052781
review_scores_communication	0.025023	-0.137222	-0.025117	-0.061539	-0.013208	-0.016067	0.001113	-0.013373	-0.021799
review_scores_location	0.073277	0.024161	0.096746	-0.190567	-0.037520	-0.008959	-0.028991	-0.023734	0.139595
review_scores_value	0.078829	-0.114171	-0.019488	-0.043979	-0.062041	0.014297	-0.001208	-0.029102	0.005546
reviews_per_month	0.106389	-0.117272	-0.084988	0.002583	-0.144150	-0.167894	-0.230287	-0.149079	-0.176298



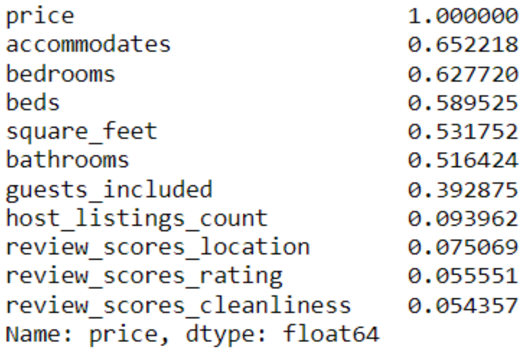
```
review_scores_accuracy         1.000000
review_scores_rating           0.621257
review_scores_value            0.562878
review_scores_cleanliness      0.543345
review_scores_communication    0.423189
review_scores_checkin          0.407238
review_scores_location         0.267605
reviews_per_month              0.143613
number_of_reviews              0.066618
host_id                        0.026768
Name: review_scores_accuracy, dtype: float64
```

The result tells us that ″review_scores_rating″ is moderately correlated to ″review_scores_accuracy″ with 62%. The next leading factors are ″review_scores_value″, and ″review_scores_cleanliness″.

★ Which variables can affect the price?

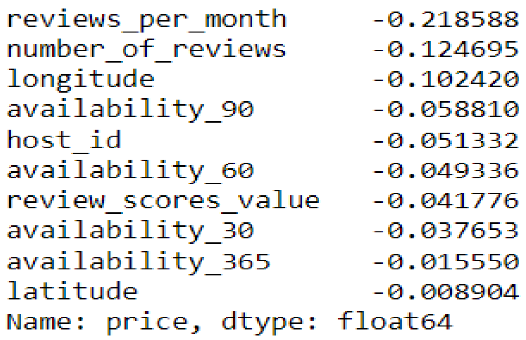


First, we do a boxplot for the price, along with accommodates to see the distribution of the price.



```
price                          1.000000
accommodates                   0.652218
bedrooms                       0.627720
beds                           0.589525
square_feet                    0.531752
bathrooms                      0.516424
guests_included                0.392875
host_listings_count            0.093962
review_scores_location         0.075069
review_scores_rating           0.055551
review_scores_cleanliness      0.054357
Name: price, dtype: float64
```

The results show that accomodates, bedrooms, and beds are the top three variables that have the most correlation with price.

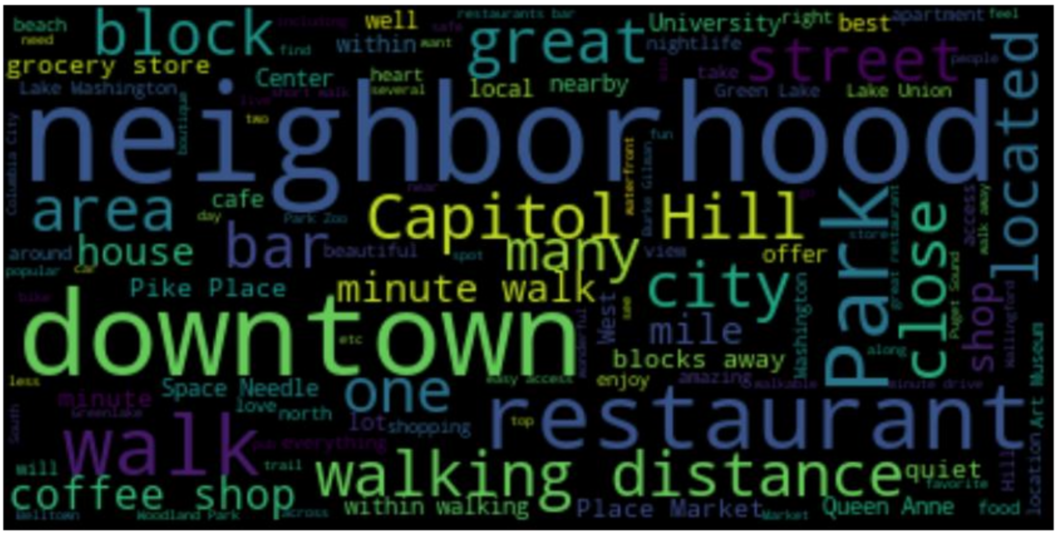


```
reviews_per_month      -0.218588
number_of_reviews      -0.124695
longitude              -0.102420
availability_90        -0.058810
host_id                -0.051332
availability_60        -0.049336
review_scores_value    -0.041776
availability_30        -0.037653
availability_365       -0.015550
latitude               -0.008904
Name: price, dtype: float64
```

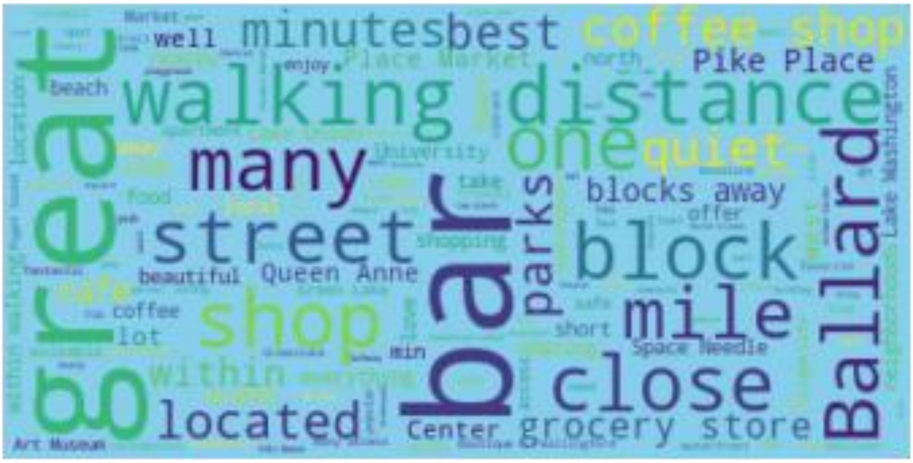
The results show that reviews_per_month, Number_of_reviews and longitude are the top three variables that have the least correlation with price.

★ **Can you describe the vibe of each Seattle neighborhood using listing descriptions?**

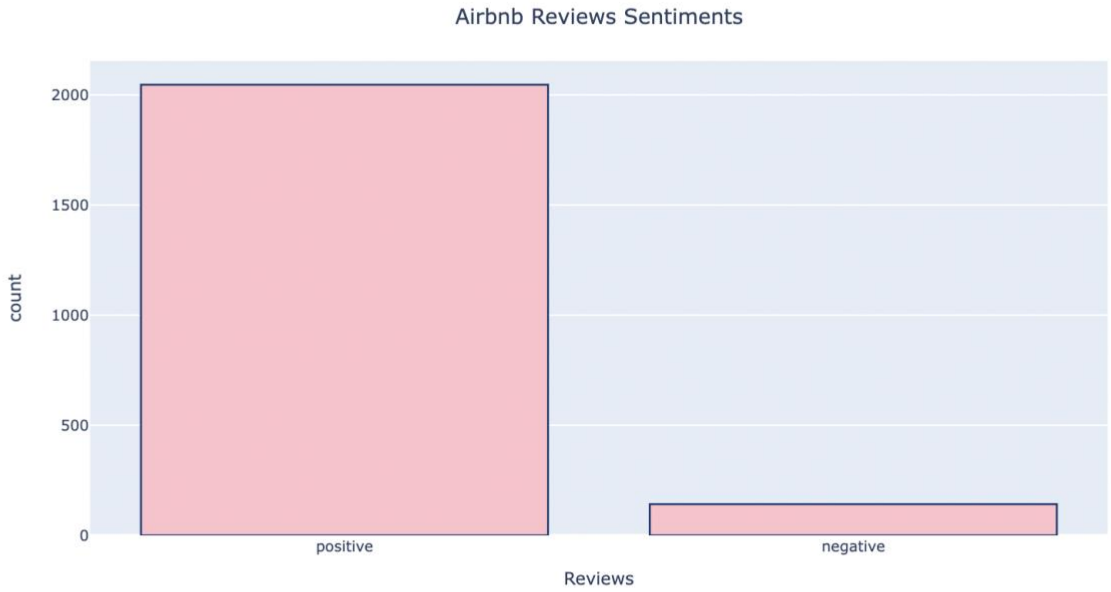
- Wordcloud below shows the most common words used in the neighborhood overview mentioned in the listsing.csv.
- Words like ‘great’ , ‘well’ , ‘best’ , ‘right’ , ‘love’ , ‘enjoy’ , ‘popular’ are protruding. We don’ t notice many negative words in the word cloud right now.
- There are many other words in the word cloud which cannot be considered as something that would define a review. But in this initial stage we have categorized the reviews as positive and negative.



- The image below shows a word cloud for POSITIVE reviews:
- Reviews can be considered as positive when words like ‘great’ , ‘quiet’ , ‘well’ , ‘beautiful’ are mentioned.

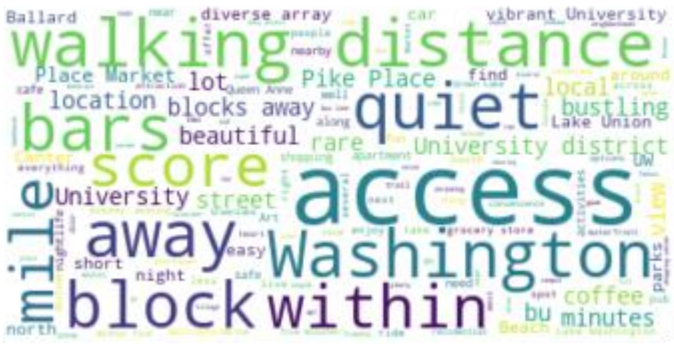


The plot shows the count of positive reviews vs the negative reviews:

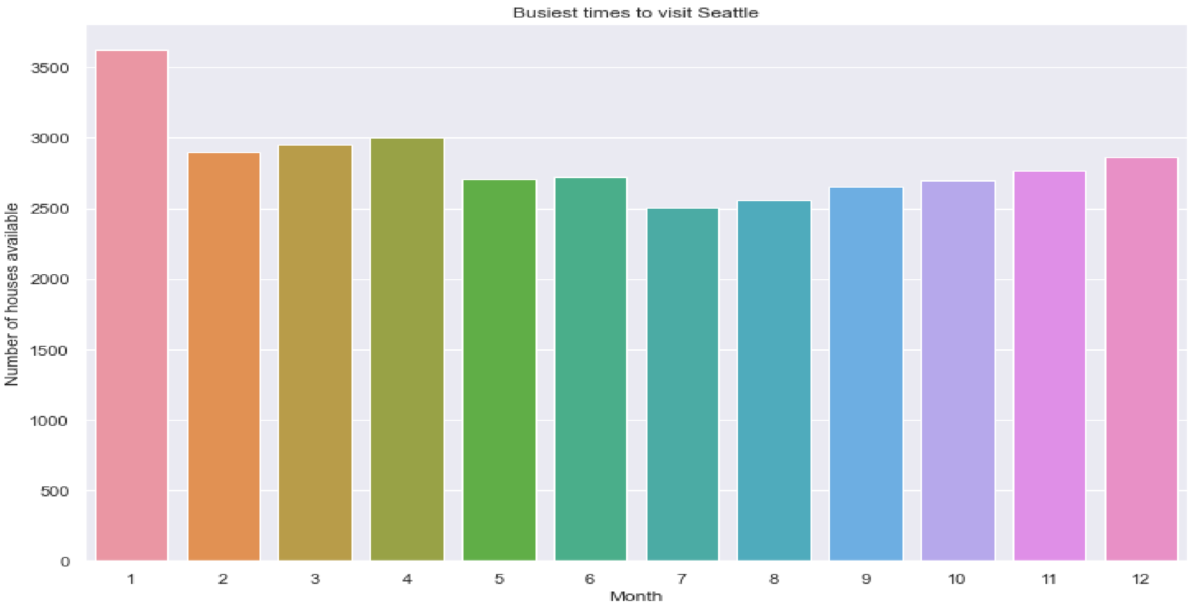


After eliminating the null values our dataset contained 2189 rows.

Approximately **93.51% of the reviews and ratings are positive**. While **6.4% are negative**. It can be said that our reviews align with the ratings because when we look at the word cloud below for the negative reviews we hardly see any words that convey negative reviews.



- What are the busiest times of the year to visit Seattle? By how much do prices spike?

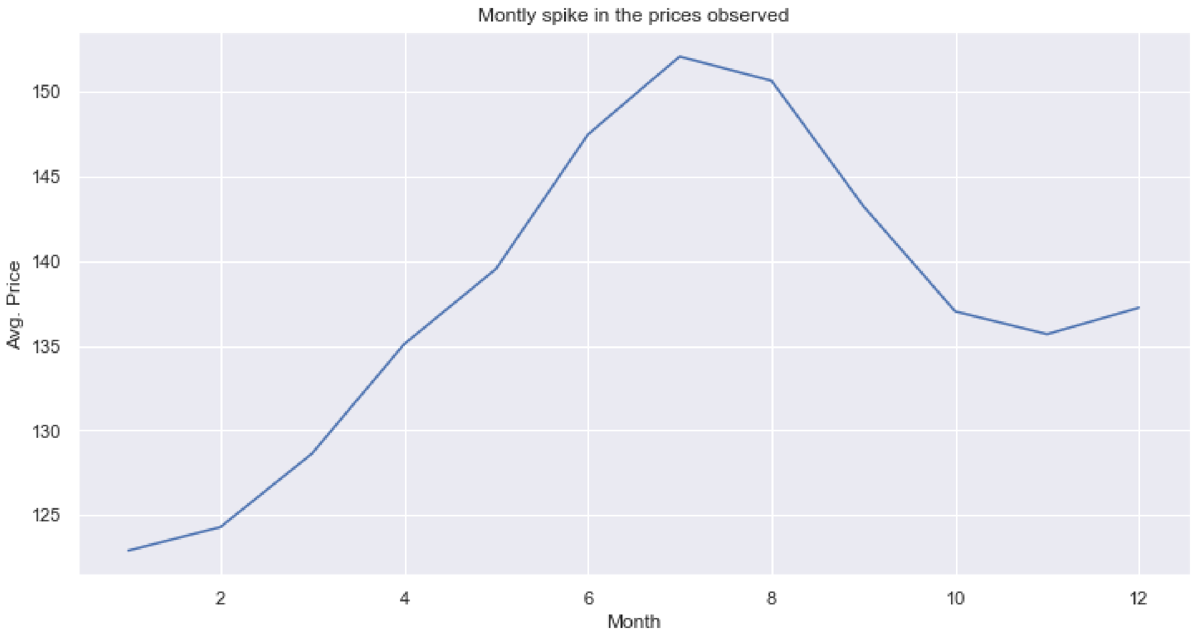


We observe that:

- From a total of 3818 houses that were listed on Airbnb, 3600 houses are available in Jan for at least a day.
- We also notice that July and August have the minimum number of houses ~2500 available for at least a day.

Hence, we can conclude that July and August are the busiest times to visit Seattle. As observed from the trend,Summer generally attracts more people hence is the busiest period of the year to visit Seattle.

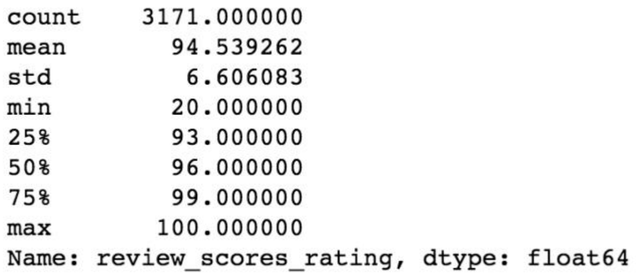
Spike in the prices per month:



- Trends show that average price increased during Summer(June,July,Aug) with a peak observed during July with average price of around $152.
- After the peak in July, the average price continues to decrease steadily until November after that it starts to increase again.
- This goes in line with our previous conclusion that Summer attracts more people and hence is the busiest time to visit Seattle and thus shows a spike in the prices.

★ What are the characteristics of the high rating properties?

Overview

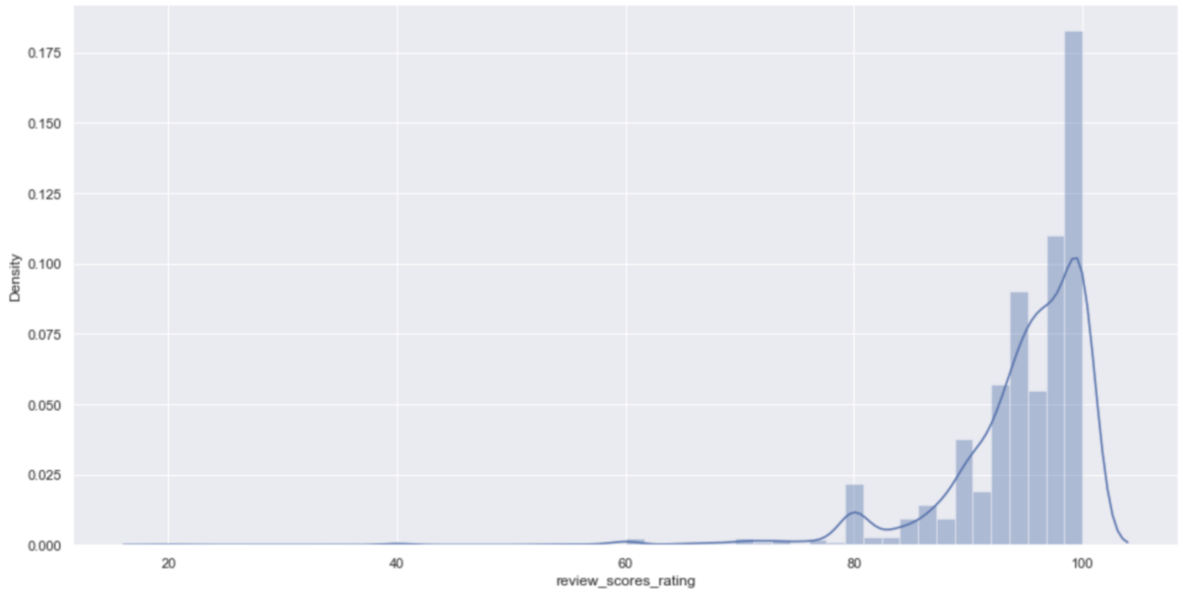


```
count    3171.000000
mean       94.539262
std         6.606083
min        20.000000
25%        93.000000
50%        96.000000
75%        99.000000
max       100.000000
Name: review_scores_rating, dtype: float64
```

- In the data set, 3171 properties have received rating scores.

- The average rating score is 94.53.

- The rating score range is between 20 points to 100 points.



- According to the frequency distribution above, more than 17.5% properties received 100 points.

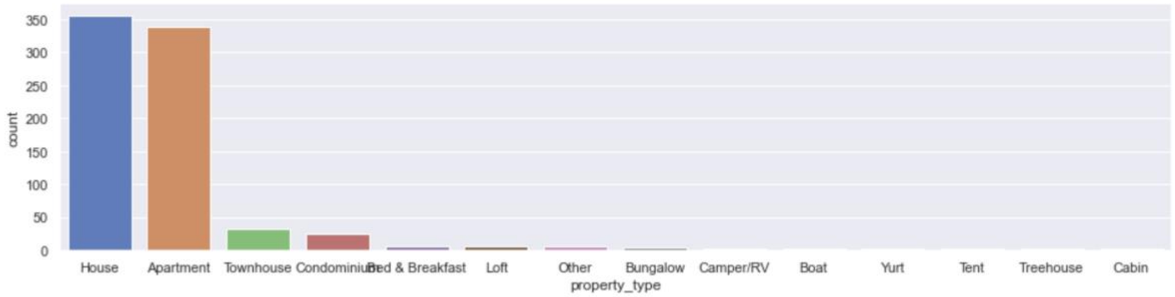
Properties that received 100 points are set as high rating properties.

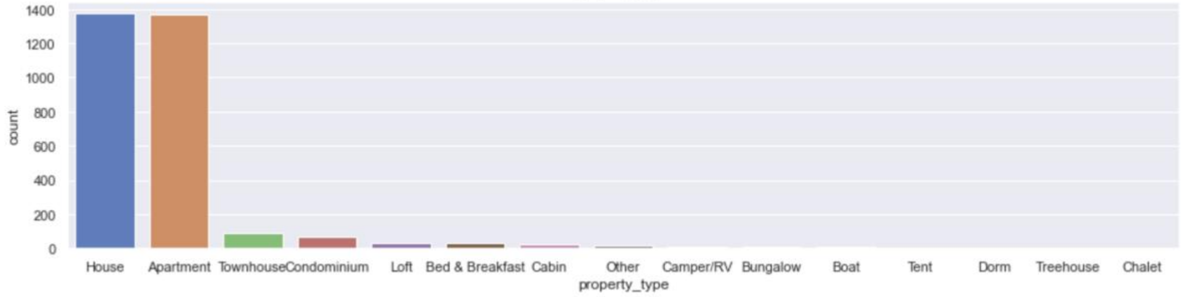
To find out the high rating properties' characteristics, we are going to compare the statistics between properties that received full points to those properties that did not receive full marks from property type, capacity, price, booking and cancel policies, availability, and property geographic locations.

Statistic comparisons

Note: First figure is for high rating properties, second figure is for non-high rating properties

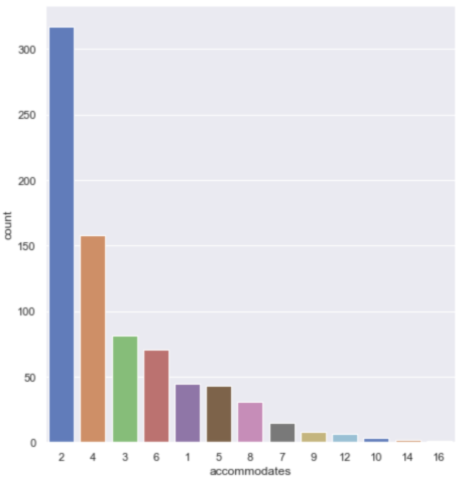
- Property type

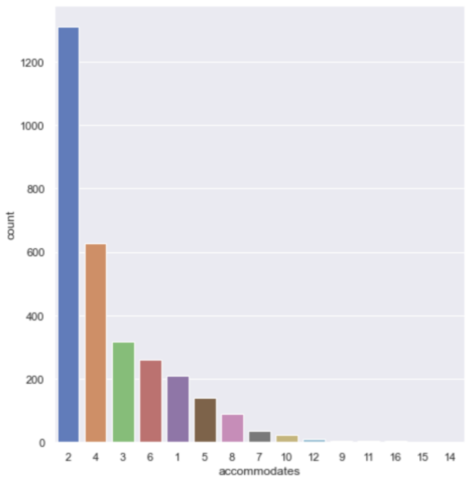




Most of the high-rating and non-high rating properties are houses, apartments, townhouses and condos.

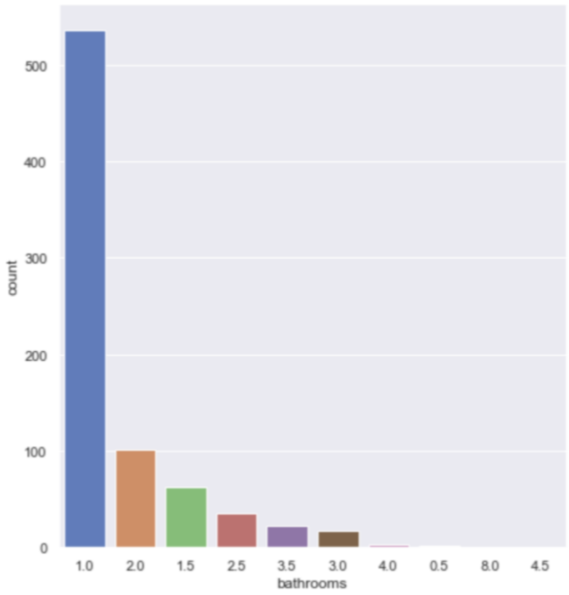
- Capacity
- Accommodates

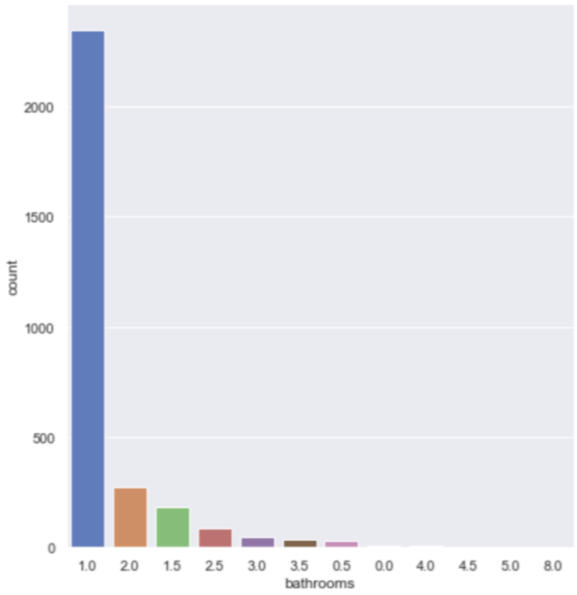




The difference between high-rating and non-high rating properties is trivial.

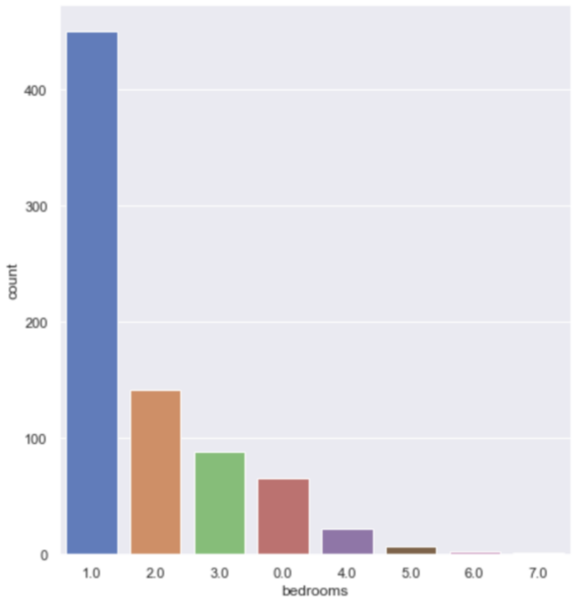
- Number of bathrooms

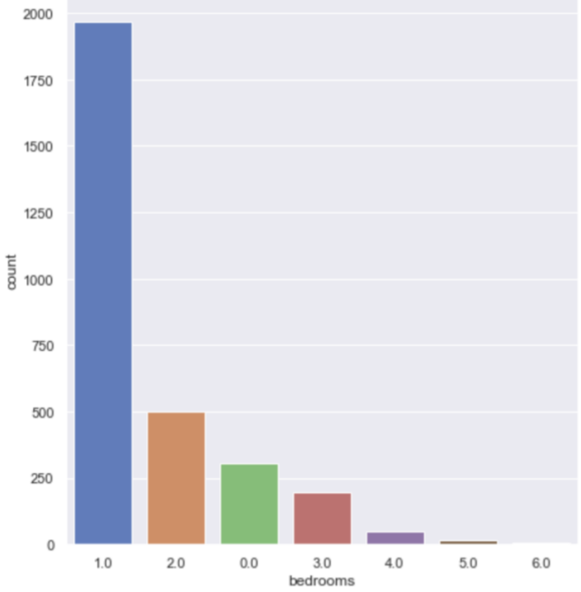




Regardless of the review scores, most of the properties have 1 bathroom.

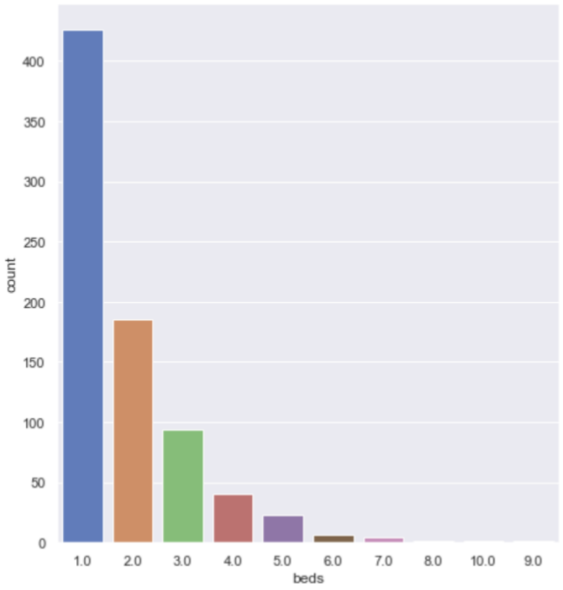
- Number of bedrooms

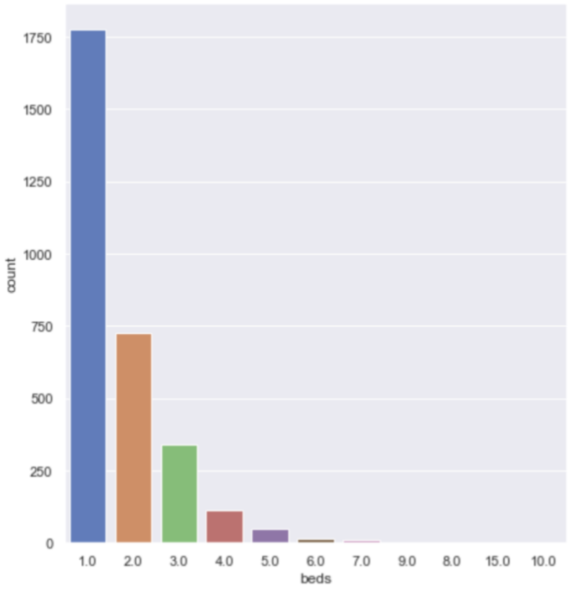




Regardless of the review scores, most of the properties have 1 bedroom.

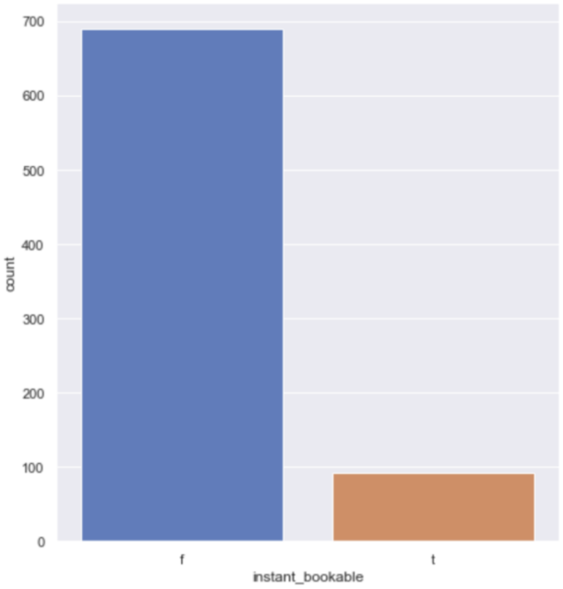
- Number of beds

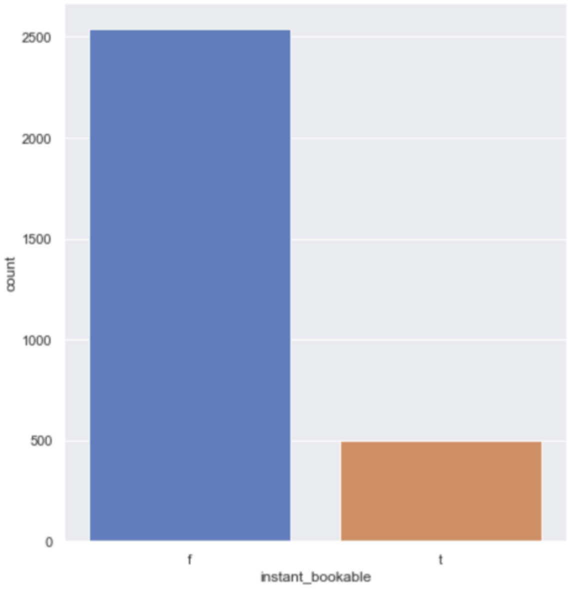




Regardless of the review scores, most of the properties have 1 bed.

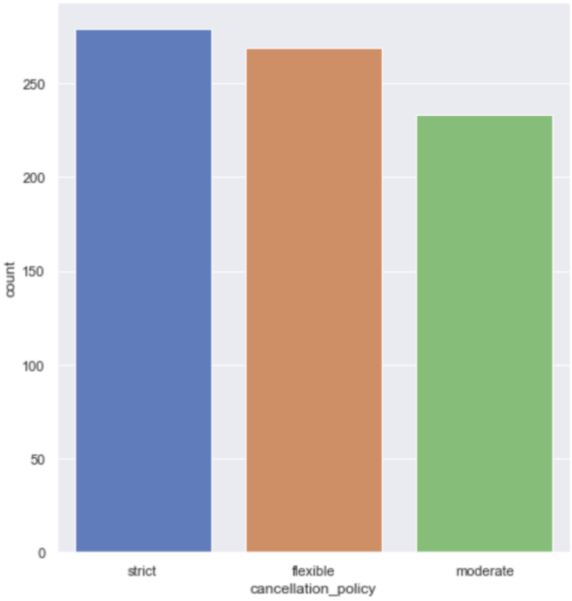
- Square feet
    - The average surface of high-rating properties: 970.25 ft^2
    - The average surface of high-rating properties: 838.29 ft^2
    - The average surface of high-rating properties is about 130 ft^2 larger than non-ratings. However, there are only 12 and 85 available entries for each subset. Since the data is small, we may not consider square feet as an informative feature in the question.
- Booking and cancel policies
- Instant booking

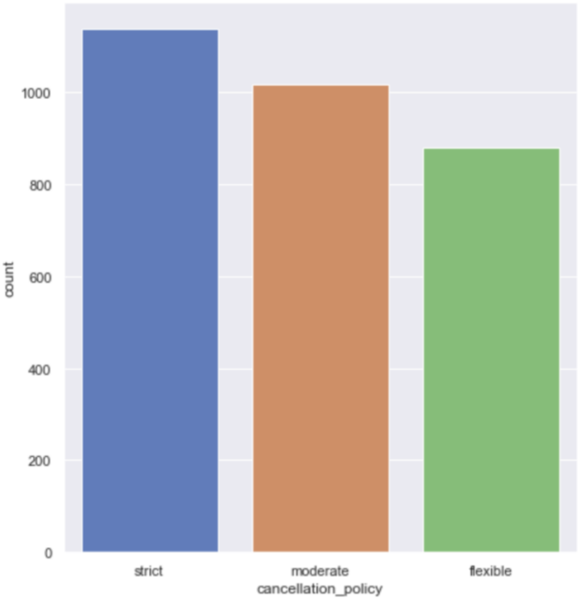




Compared to non-high rating properties, less proportion of high-rating properties support instant booking.

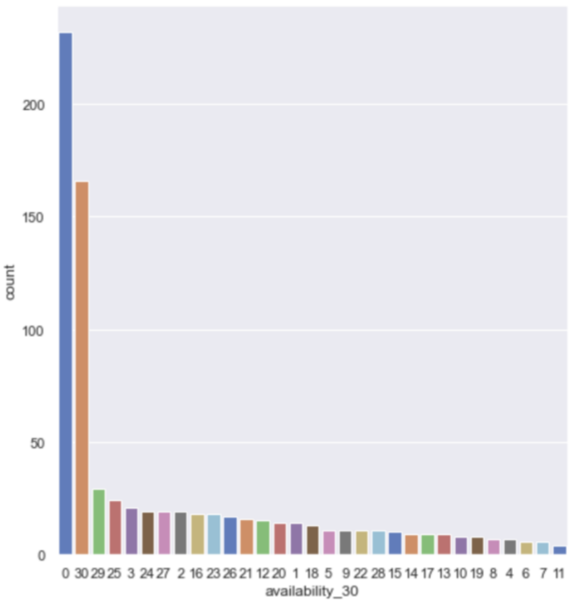
- Cancellation policy

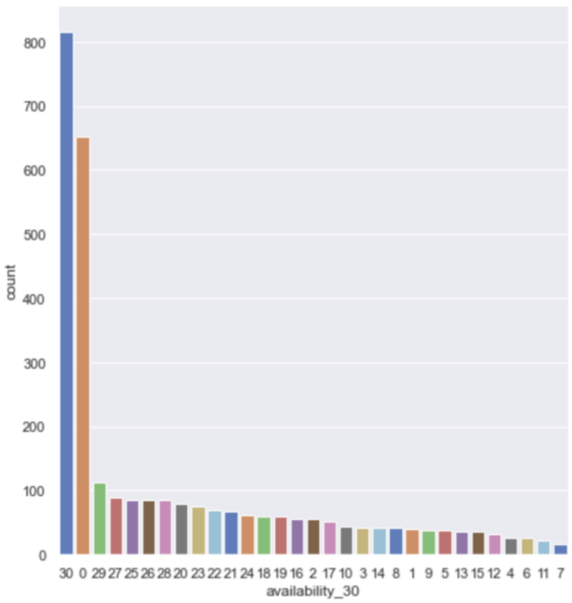




Compared to non-high rating properties, more proportion of high-rating properties are flexible and less proportion are moderate. The overall cancellation policy of high-rating properties is less strict compared to non-high ratings.

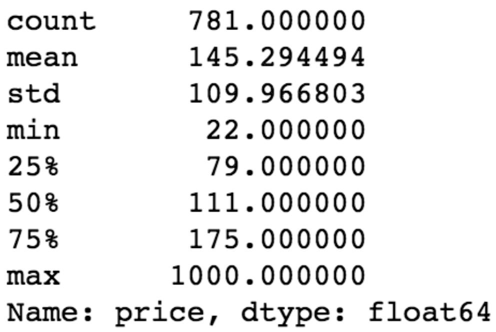
- Availability in 30 days



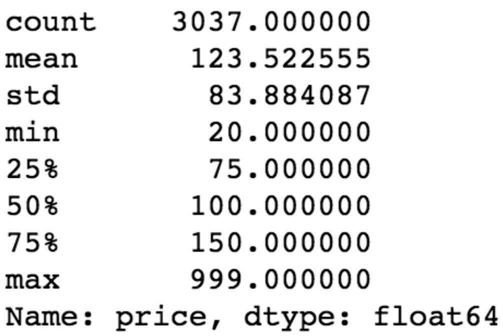


Most of the high-rating properties are available to book on the same day.

- Price



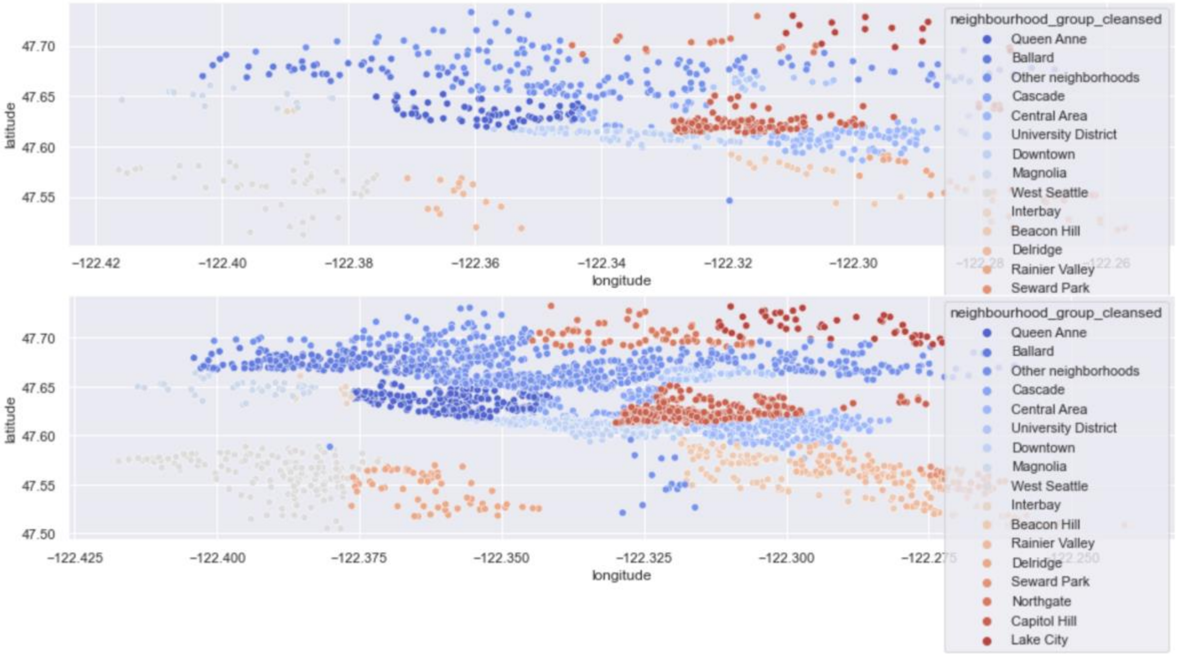
```
count      781.000000
mean       145.294494
std        109.966803
min         22.000000
25%         79.000000
50%        111.000000
75%        175.000000
max       1000.000000
Name: price, dtype: float64
```



```
count    3037.000000
mean      123.522555
std        83.884087
min        20.000000
25%        75.000000
50%       100.000000
75%       150.000000
max       999.000000
Name: price, dtype: float64
```

The average price of high-rating properties is $145.29 each night, about $22 higher than non-high ratings. The difference of price range between two groups is not obvious.

- Property geographic distribution



Most of the high-rating properties are located in West Seattle, Queen Anne, Ballard, Central Area, Cascade and other neighborhoods.

Summarization of high-rating property characteristics

- More than 17.5% properties received 100 points in total review scores and are set as high-rating properties.
- Less high-rating properties support instant booking.
- The cancellation policy of high-rating properties is less strict.
- The average price (per night) of high-rating properties is $22 above others.
- Most of the high-rating properties are located in West Seattle, Queen Anne, Ballard, Central Area and Lake City.

References

1. Chen, B. (2021, December 24). *How to convert JSON into a Pandas DataFrame - Towards Data Science*. Medium. https://towardsdatascience.com/how-to-convert-json-into-a-pandas-dataframe-100b2ae1e0d8
2. *Python Word Clouds Tutorial: How to Create a Word Cloud*. (n.d.). DataCamp Community. https://www.datacamp.com/community/tutorials/wordcloud-python
3. Real Python. (2021, September 24). *Sentiment Analysis: First Steps With Python’s NLTK Library*. Real Python NLTK. https://realpython.com/python-nltk-sentiment-analysis/
4. Saxena, N. (2021, December 15). *Extracting Keyphrases from Text: RAKE and Gensim in Python*. Medium. https://towardsdatascience.com/extracting-keyphrases-from-text-rake-and-gensim-in-python-eefd0fad582f
5. Selvaraj, N. (2021, December 16). *A Beginner’s Guide to Sentiment Analysis with Python*. Medium. https://towardsdatascience.com/a-beginners-guide-to-sentiment-analysis-in-python-95e354ea84f6

6. Singh, D. (2019, June 27). *Visualizing Text Data Using a Word Cloud*. Pluralsight. https://www.pluralsight.com/guides/natural-language-processing-visualizing-text-data-using-word-cloud

7. *"wordcloud" is not defined in python3.6*. (2017, December 26). Stack Overflow. https://stackoverflow.com/questions/47980656/wordcloud-is-not-defined-python3-6

8. Z. (2021, November 9). *How to Keep Certain Columns in Pandas (With Examples)*. Statology. https://www.statology.org/pandas-keep-columns/