DOCUMENTATION

Exploratory Data Analysis (EDA) on Airbnb Listing

Understanding the data:

I have downloaded the data source from the link that is provided in the task document link for the dataset.

Firstly I have downloaded the .CSV file and took and overview of the no. of entries and attributes and their nature.

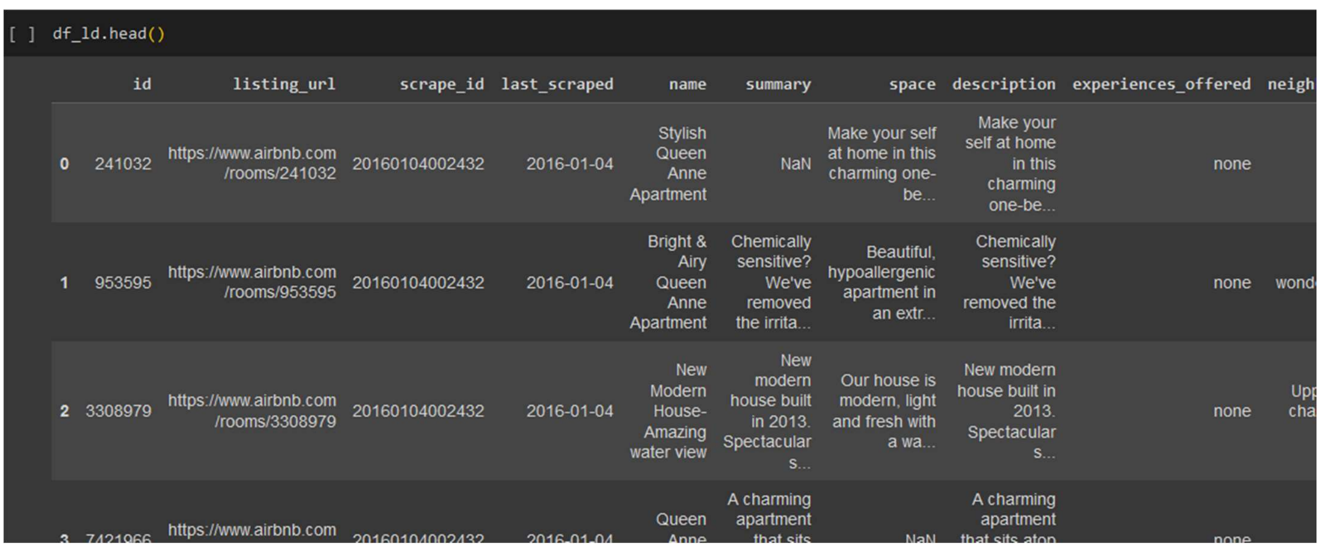
Checked for id or unique data columns which are not specifically important for the analysis.

Also observed that many columns had '$' symbol in price type attributes.

Because it is an Airbnb data I looked into the columns that are most important for the data analysis.

Data Exploration:

1. Importing libraries – import python libraries like pandas, matplotlib, seaborn, regx required for the task.
2. Data from the .CSV file is loaded and converted into a Dataframe.
3. Path of the listing data is taken as a variable listing_data and data from this path is loaded into a Dataframe.
4. Use head() function to get an overview of the data that we are working with.



```
[ ] df_ld.head()
```

	id	listing_url	scrape_id	last_scraped	name	summary	space	description	experiences_offered	neigh
0	241032	https://www.airbnb.com /rooms/241032	20160104002432	2016-01-04	Stylish Queen Anne Apartment	NaN	Make your self at home in this charming one-be...	Make your self at home in this charming one-be...	none	
1	953595	https://www.airbnb.com /rooms/953595	20160104002432	2016-01-04	Bright & Airy Queen Anne Apartment	Chemically sensitive? We've removed the irrita...	Beautiful, hypoallergenic apartment in an extr...	Chemically sensitive? We've removed the irrita...	none	wond
2	3308979	https://www.airbnb.com /rooms/3308979	20160104002432	2016-01-04	New Modern House-Amazing water view	New modern house built in 2013. Spectacular s...	Our house is modern, light and fresh with a wa...	New modern house built in 2013. Spectacular s...	none	Upp cha
3	7421966	https://www.airbnb.com	20160104002432	2016-01-04	Queen Anne	A charming apartment that sits	NaN	A charming apartment that sits atop	none	

5. Use shape( ) function to get the no. of entries and attributes in the dataset.
6. In this task the shape of the dataset is (3818,92) i.e, it has 3818 rows and 92 columns or attributes.
7. Used info( ) function to get the data-types of the attributes and also the number of non-null entries in each column.
8. Used describe( ) function to get basic statistical details like mean, standard deviation, percentile etc.

```
[ ] (df_ld.select_dtypes(include=['int64', 'float64'])).describe()
```

	id	scrape_id	host_id	host_listings_count	host_total_listings_cou
count	3.818000e+03	3.818000e+03	3.818000e+03	3816.000000	3816.0000
mean	5.550111e+06	2.016010e+13	1.578556e+07	7.157757	7.1577
std	2.962660e+06	0.000000e+00	1.458382e+07	28.628149	28.6281
min	3.335000e+03	2.016010e+13	4.193000e+03	1.000000	1.0000
25%	3.258256e+06	2.016010e+13	3.275204e+06	1.000000	1.0000
50%	6.118244e+06	2.016010e+13	1.055814e+07	1.000000	1.0000
75%	8.035127e+06	2.016010e+13	2.590309e+07	3.000000	3.0000
max	1.034016e+07	2.016010e+13	5.320861e+07	502.000000	502.0000

Data Cleaning:

1. Firstly, I have checked for columns having minimum one missing value using isnull( )and sum( ) functions.
2. And observed that there are columns that have more than 75% of null values.
3. So, I calculated the percentage of null values in each column and created a Dataframe listing the columns having more than 75% of null values.
4. Since 75% of missing values in the data will add a lot of noise while doing analysis, it is ideal to drop those columns.
5. Also I observed that there is no use of scrape_id column in the analysis, so I have dropped that column.
6. Also observed that there are columns that have currency data that are in object type and have more null values.
7. As I have mentioned before in the understanding data section that there are '$' characters in the columns making them object type, so we have to remove the special characters and also replace null values with zeroes.

```
[ ] df_ld[['security_deposit','weekly_price','cleaning_fee','monthly_price','price']]
```

	security_deposit	weekly_price	cleaning_fee	monthly_price	price
0	NaN	NaN	NaN	NaN	$85.00
1	$100.00	$1,000.00	$40.00	$3,000.00	$150.00
2	$1,000.00	NaN	$300.00	NaN	$975.00
3	NaN	$650.00	NaN	$2,300.00	$100.00
4	$700.00	NaN	$125.00	NaN	$450.00
...	...	...	...	...	...
3813	NaN	NaN	$230.00	NaN	$359.00

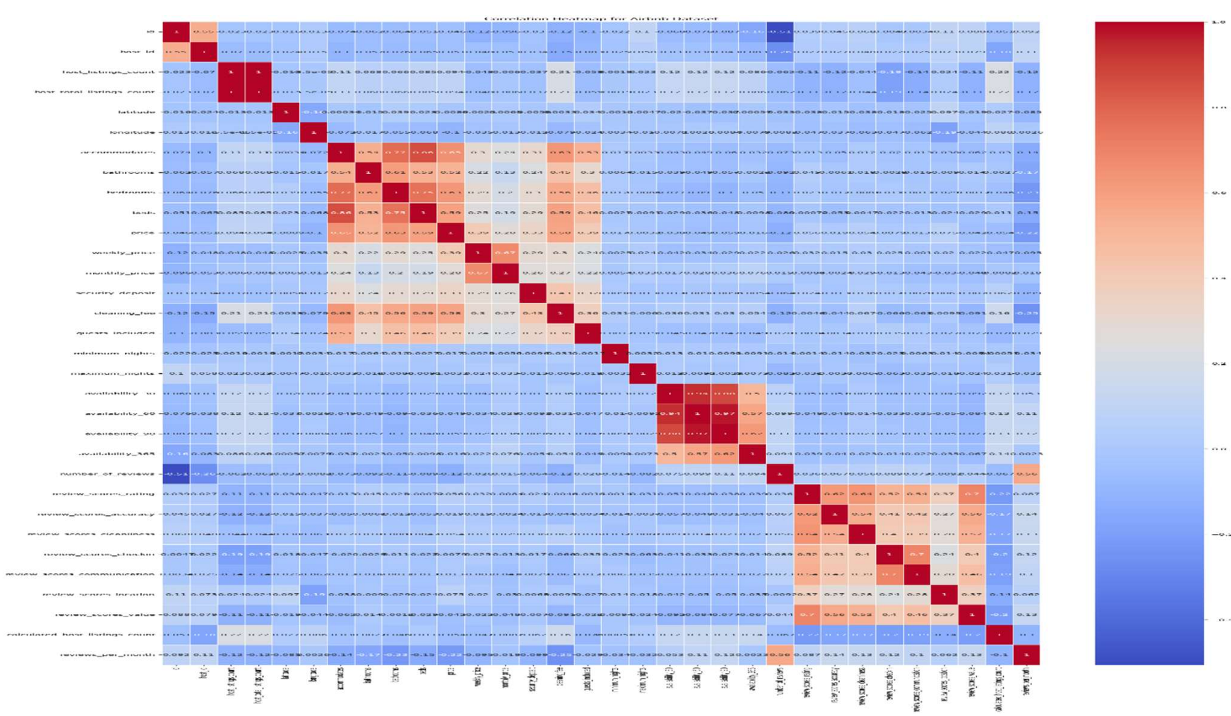
8. After removing the special characters and null values the columns automatically changes into numeric type.

```
[ ] df_ld[['security_deposit','weekly_price','cleaning_fee','monthly_price','price']]
```

	security_deposit	weekly_price	cleaning_fee	monthly_price	price
0	0.0	0.0	0.0	0.0	85.0
1	100.0	1000.0	40.0	3000.0	150.0
2	1000.0	0.0	300.0	0.0	975.0
3	0.0	650.0	0.0	2300.0	100.0
4	700.0	0.0	125.0	0.0	450.0
...	...	...	...	...	...
3813	0.0	0.0	230.0	0.0	359.0
3814	500.0	0.0	50.0	0.0	79.0

Correlation Analysis:

1. After doing the necessary cleaning in the data I've created a correlation matrix to check the relation between the attributes.
2. For a careful observation, a heat-map is plotted to clearly check for the correlation between the attributes.



3. From the above correlation matrix we can observe that the listing price is influenced or effected by number of bed, bedrooms, bathrooms and accommodates etc.

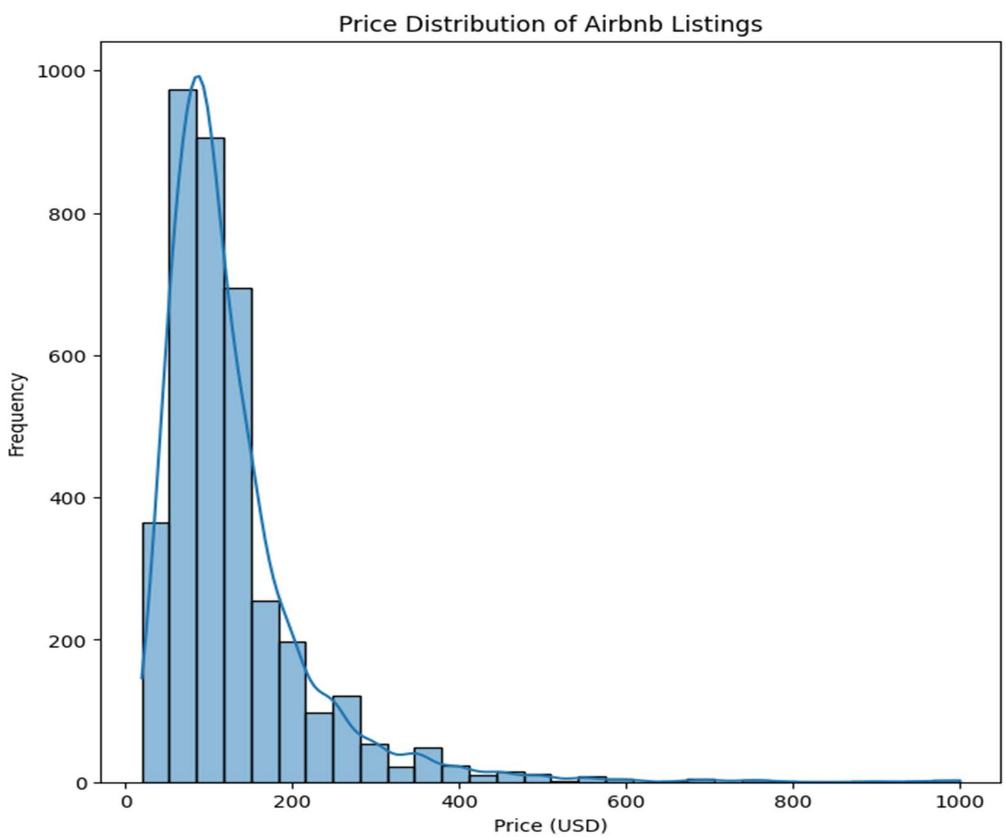
4. Also observed that the number of reviews doesn't actually effect the increase and decrease in prices.

Visualization and Statistics:

1. Calculated the average prices of rooms grouped by neighbourhood.

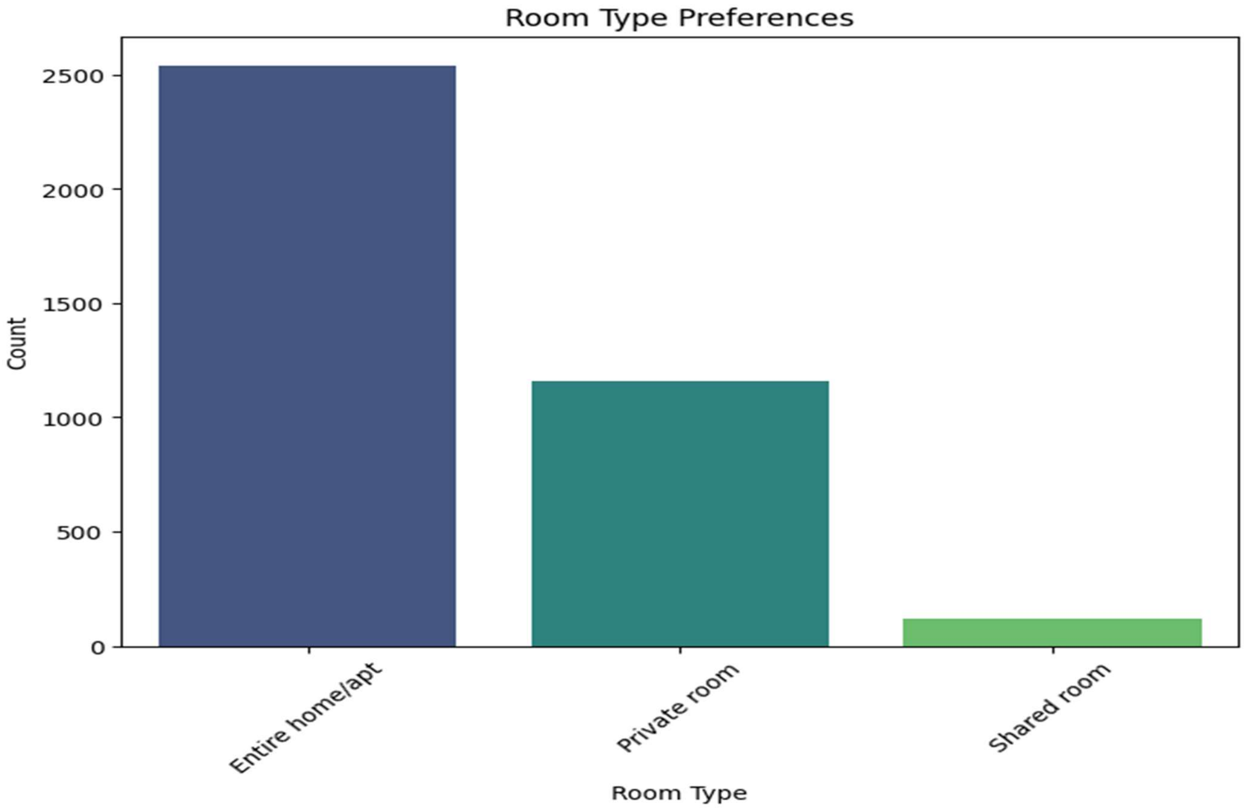
```
] # Group the data by neighborhood and calculate the mean price
  mean_price_by_neighb = df_ld.groupby('neighbourhood')['price'].mean()
  mean_price_by_neighb.sort_values(ascending=False)

  neighbourhood
  Fairmount Park         370.000000
  Industrial District    245.000000
  Portage Bay            241.428571
  Westlake               197.000000
  Alki                   196.652174
                            ...
  Georgetown              77.000000
  Rainier Beach           76.722222
  Dunlap                  75.461538
  Olympic Hills           63.666667
  Roxhill                 60.000000
  Name: price, Length: 81, dtype: float64
```

2. Plotted a Histogram to get the price distribution of the Airbnb Listings.

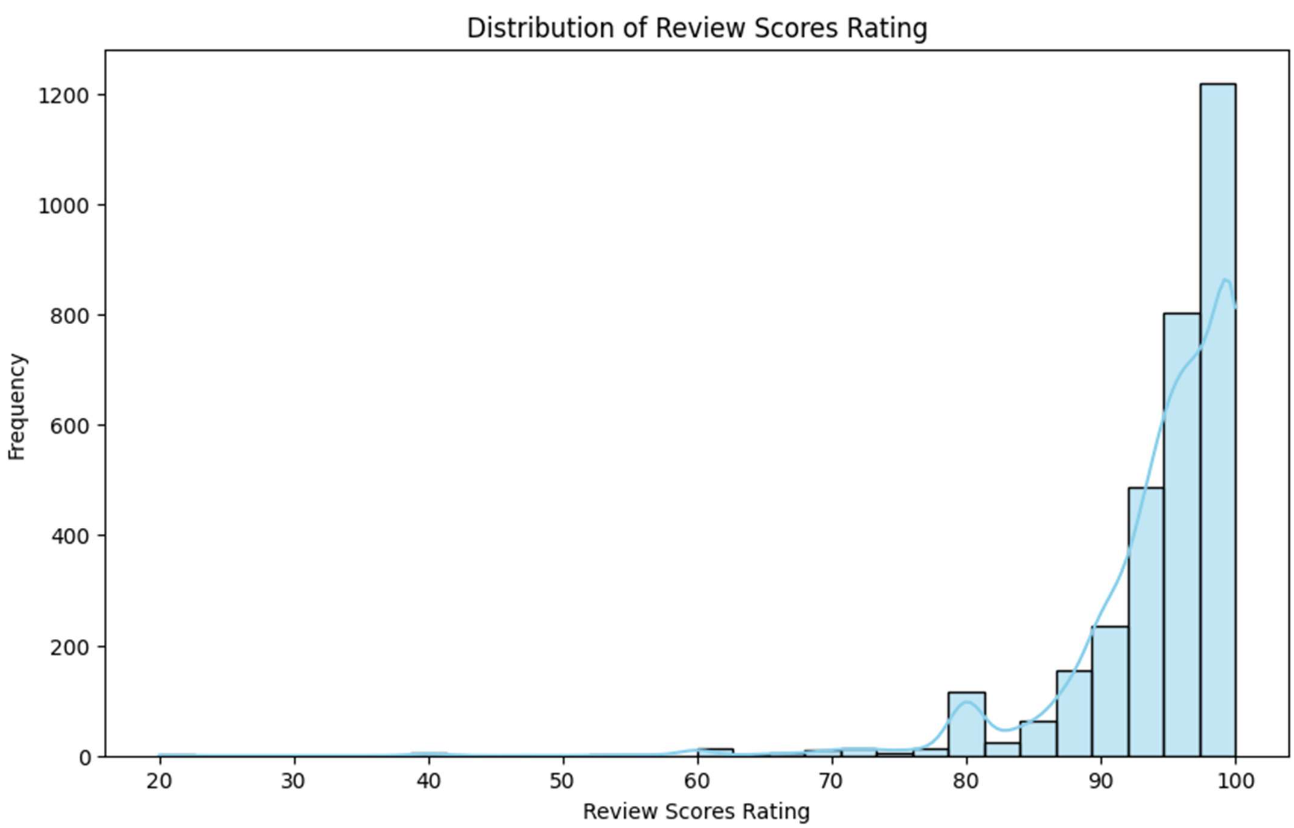


3. The plot above shows that most of the bookings takes place in price range of 60-200 dollars.

4. A count plot is plotted to get the count of bookings preferring certain kind of rooms.

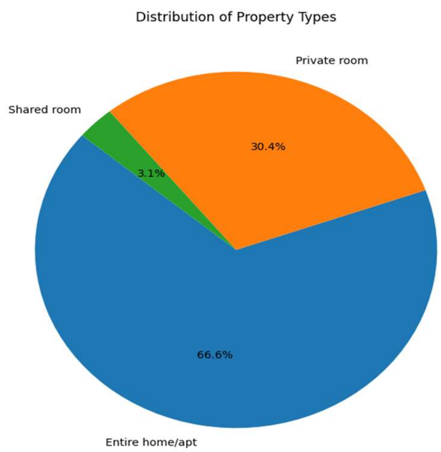


5. The plot above shows that there are more number of bookings preferring the Room Type as Entire home/ Apt.
6. Plotted a Histogram to get the distribution of Review score rating



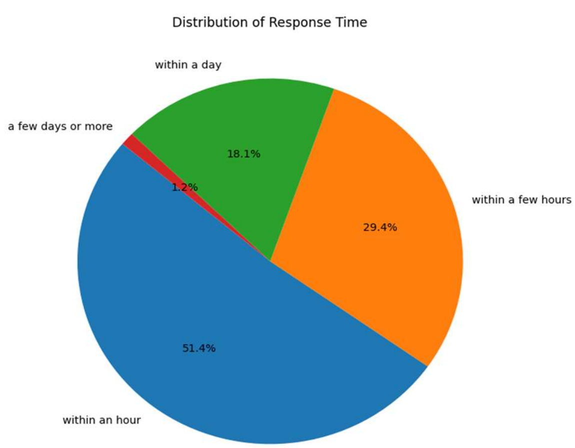
7. The above plot shows that more number of reviews are above 85%

8. Plotted a pie-chart to know the distribution of property types.



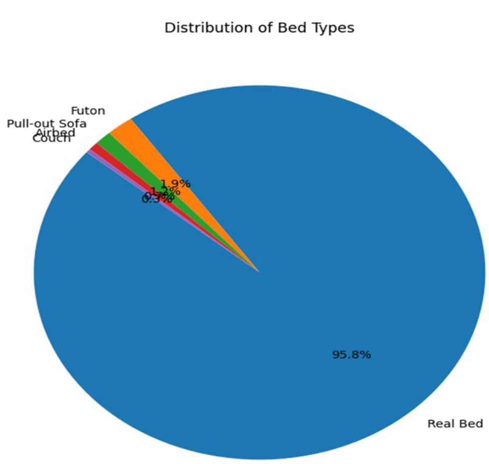
9. The above visual shows the property distribution in seattle region.
66.6% of property is Entire home/apt, 30.4% private room, 3.1% shared rooms

10. Plotted a pie-chart to know the distribution of Response Times.



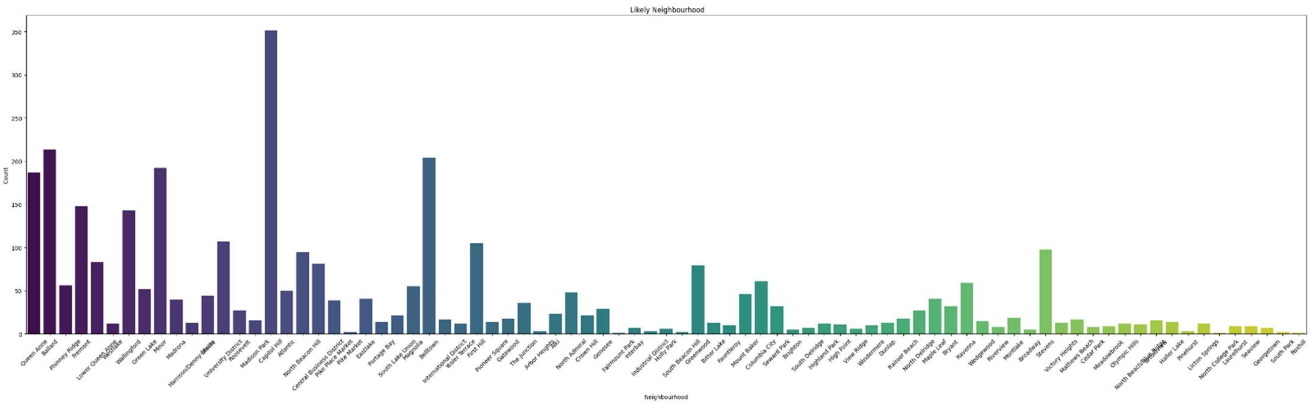
11. The above visual shows the time distribution in which the responses are received.
Within an hour 51.4%, within a few hours 29.4%, within a day 18.1%, A few days or more 1.2%.

12. Plotted a pie-chart to know the distribution of Bed types.



13. The above visual shows the pie chart in which distribution of bed Type preferences are shown. Real bed as preference 95.8%, Couch 0.3%, Airbed 0.7%, Pull-out sofa 1.2%, futon 1.9%.

14. Plotted a count-plot to know the popularity of the neighbourhood in the listings.



15. The above visual shows that the most popular neighbourhood from the Listings is Capital hill and then Ballard, Bell town etc.