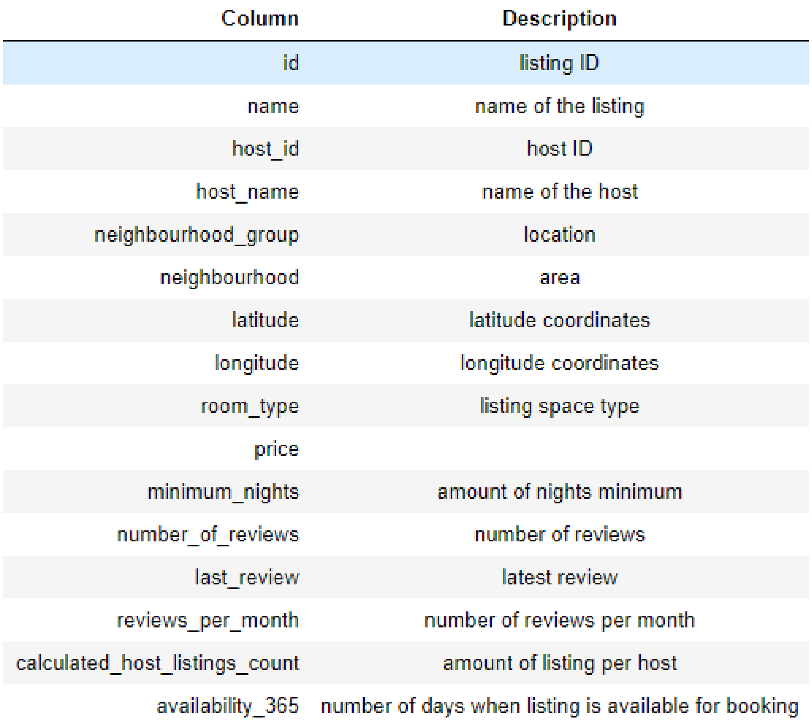
Methodology

Importing Data

- All the important libraries for data analysis were imported into the Jupyter Notebook
- Libraries Imported – NumPy, Pandas, Matplotlib, Seaborn, os
- The Airbnb dataset which was in csv format was imported

Data Understanding

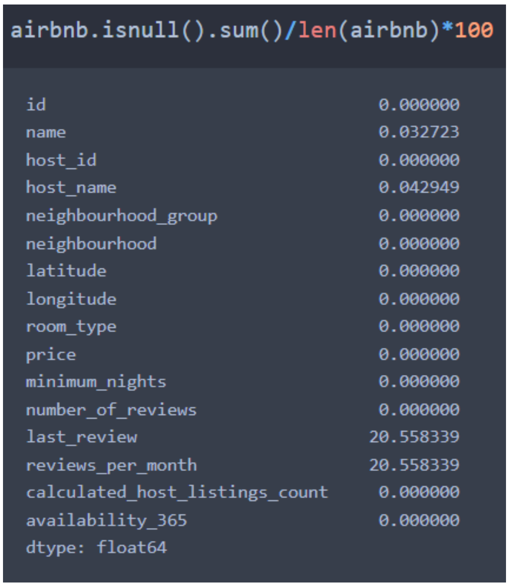


Column	Description
id	listing ID
name	name of the listing
host_id	host ID
host_name	name of the host
neighbourhood_group	location
neighbourhood	area
latitude	latitude coordinates
longitude	longitude coordinates
room_type	listing space type
price	
minimum_nights	amount of nights minimum
number_of_reviews	number of reviews
last_review	latest review
reviews_per_month	number of reviews per month
calculated_host_listings_count	amount of listing per host
availability_365	number of days when listing is available for booking

- Methods such as info(), describe(), dtypes, etc. were used to get basic understanding of data

Data Cleaning

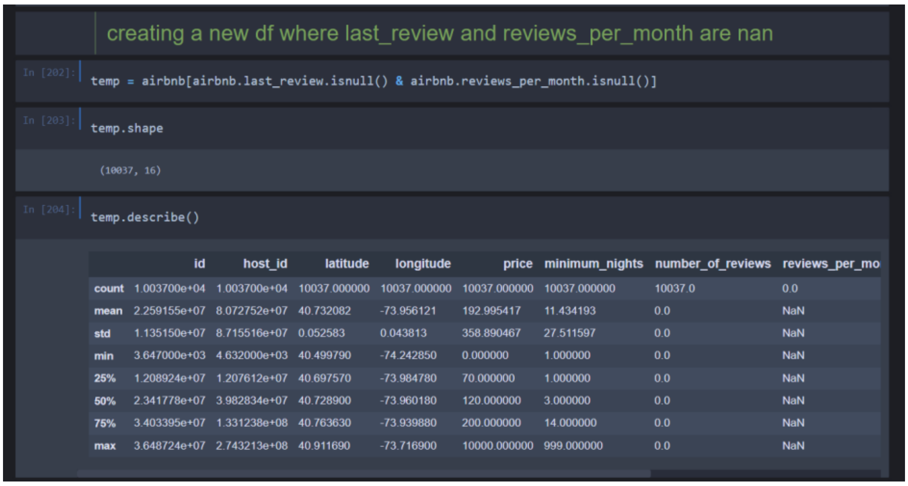
- A total of 4 columns had null values which were ‘name’, ‘host_name’, ‘last_review’, ‘reviews_per_month’.



```
airbnb.isnull().sum()/len(airbnb)*100

id                                 0.000000
name                               0.032723
host_id                            0.000000
host_name                          0.042949
neighbourhood_group                0.000000
neighbourhood                      0.000000
latitude                           0.000000
longitude                          0.000000
room_type                          0.000000
price                              0.000000
minimum_nights                     0.000000
number_of_reviews                  0.000000
last_review                       20.558339
reviews_per_month                 20.558339
calculated_host_listings_count     0.000000
availability_365                   0.000000
dtype: float64
```

Imputing null values

- The columns ‘name’ and ‘host_name’ were of object data type and had listing names and host names.
- As ‘host_name’ has 0.033% null values and ‘name’ has 0.043% null values we directly dropped these rows as imputing them with any modal value does not makes sense.
- After dropping these rows, 99.9% of data was retained.



creating a new df where last_review and reviews_per_month are nan

```
In [202]: temp = airbnb[airbnb.last_review.isnull() & airbnb.reviews_per_month.isnull()]
```

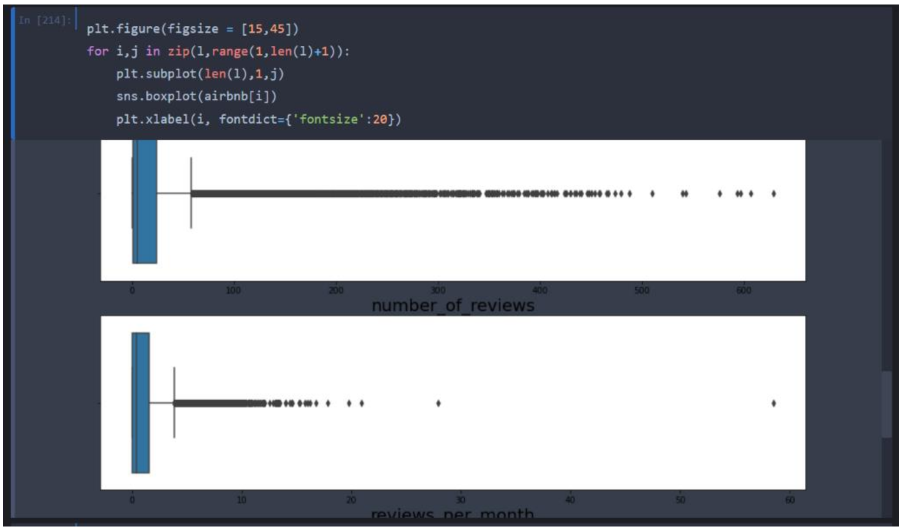
```
In [203]: temp.shape

(10037, 16)
```

```
In [204]: temp.describe()
```

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_mo
count	1.003700e+04	1.003700e+04	10037.000000	10037.000000	10037.000000	10037.000000	10037.0	0.0
mean	2.259155e+07	8.072752e+07	40.732082	-73.956121	192.995417	11.434193	0.0	NaN
std	1.135150e+07	8.715516e+07	0.052583	0.043813	358.890467	27.511597	0.0	NaN
min	3.647000e+03	4.632000e+03	40.499790	-74.242850	0.000000	1.000000	0.0	NaN
25%	1.208924e+07	1.207612e+07	40.697570	-73.984780	70.000000	1.000000	0.0	NaN
50%	2.341778e+07	3.982834e+07	40.728900	-73.960180	120.000000	3.000000	0.0	NaN
75%	3.403395e+07	1.331238e+08	40.763630	-73.939880	200.000000	14.000000	0.0	NaN
max	3.648724e+07	2.743213e+08	40.911690	-73.716900	10000.000000	999.000000	0.0	NaN

- The columns ‘last_review’ and ‘reviews_per_month’ had huge number of null values so to understand and underlying cause for it; creating a new dataframe where the aforementioned rows have null values
- Also, wherever ‘last_review’ had null value, ‘reviews_per_month’ had null value and vice-versa.
- After analyzing the new dataframe, the reason for null vales was listing had received no reviews at all so it had ‘last_review’ and ‘reviews_per_month’ empty.
- After knowing the reason, all missing values in ‘reviews_per_month’ column were imputed with ‘0’ and ‘last_review’ column with a dummy date ‘1970-1-1’

Detecting Outliers

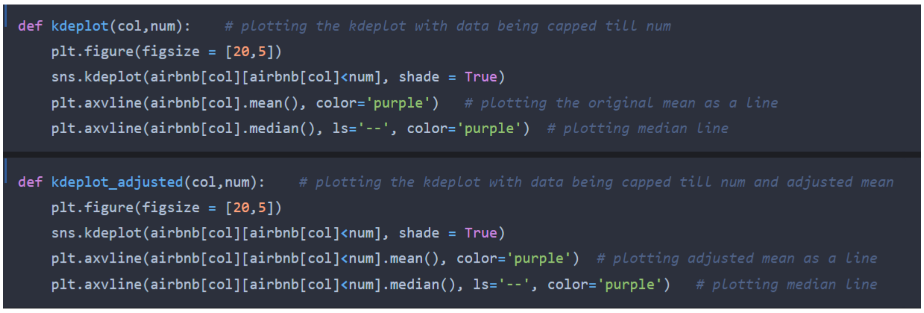


```
In [214]:  plt.figure(figsize = [15,45])
           for i,j in zip(l,range(1,len(l)+1)):
               plt.subplot(len(l),1,j)
               sns.boxplot(airbnb[i])
               plt.xlabel(i, fontdict={'fontsize':20})
```

- Boxplots were used to detect outliers.
- The columns like longitude, latitude, id, host_id, etc. are not expected to have a range so ignoring these columns.
- The 'price', 'minimum_nights', 'number_of_reviews', 'reviews_per_month' and 'calculated_host_listings_count' had huge number of outliers.

Binning

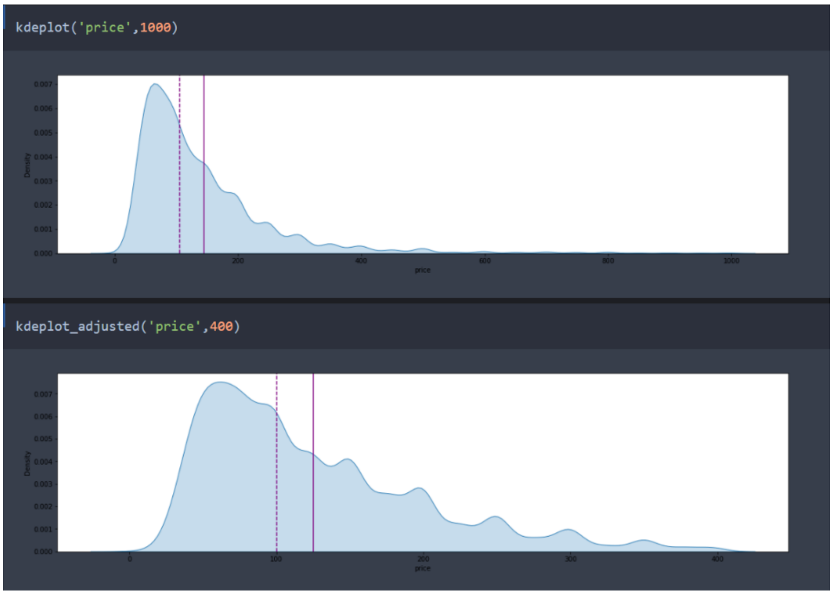
- All the above-mentioned column were binned into proper bins and new columns were created by analyzing their data distribution.
- To analyze the data distribution kdeplots and describe() methods were used.
- Functions were made to plot kdeplots and adjusted kdeplots



```
def kdeplot(col,num):    # plotting the kdeplot with data being capped till num
    plt.figure(figsize = [20,5])
    sns.kdeplot(airbnb[col][airbnb[col]<num], shade = True)
    plt.axvline(airbnb[col].mean(), color='purple')   # plotting the original mean as a line
    plt.axvline(airbnb[col].median(), ls='--', color='purple')  # plotting median line
```

```
def kdeplot_adjusted(col,num):    # plotting the kdeplot with data being capped till num and adjusted mean
    plt.figure(figsize = [20,5])
    sns.kdeplot(airbnb[col][airbnb[col]<num], shade = True)
    plt.axvline(airbnb[col][airbnb[col]<num].mean(), color='purple')  # plotting adjusted mean as a line
    plt.axvline(airbnb[col][airbnb[col]<num].median(), ls='--', color='purple')   # plotting median line
```

- Numeric columns were binned based on data distribution seen in kdeplots like below
- The 4 new columns which were created are 'price_range', 'minimum_nights_range', 'number_of_reviews_range', 'number_of_listings_range'

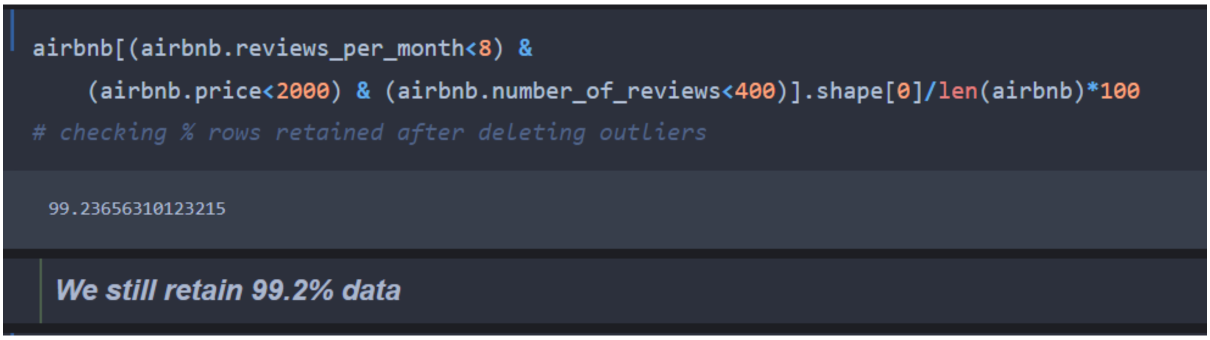


```
kdeplot('price',1000)
```

```
kdeplot_adjusted('price',400)
```

Outlier Treatment

- After creating binned columns, outliers were handled by deleting the outliers which were extremely far from the expected range



```
airbnb[(airbnb.reviews_per_month<8) &
    (airbnb.price<2000) & (airbnb.number_of_reviews<400)].shape[0]/len(airbnb)*100
# checking % rows retained after deleting outliers
```

```
99.23656310123215
```

We still retain 99.2% data

- After Data cleaning and creating necessary columns the cleaned dataframe was exported in csv format so to analyze data in Tableau.

The exported Clean Dataset
Data cleaning Jupyter Notebook

Data Analysis

- Data Analysis was performed in Tableau.
- The Calculated fields created in it are as follows
    1. Available/ Unavailable {Categorical – specifies if a listing is available or not i.e., ‘available 365’ > 0}
    2. Availability = 0 {Flag - 1 if listing is unavailable else 0}
    3. Availability > 0 {Flag - 1 if listing is available for booking at least once a year}
    4. Review per month (availability > 0) {Filters reviews per month where if a listing is unavailable then 0 else keeps the original data}
- Insights were derived using bar plots, dual axis charts, line charts and heatmaps with median reviews per month being the popularity metric
- The Tableau Notebook used for analysis – Airbnb Storytelling