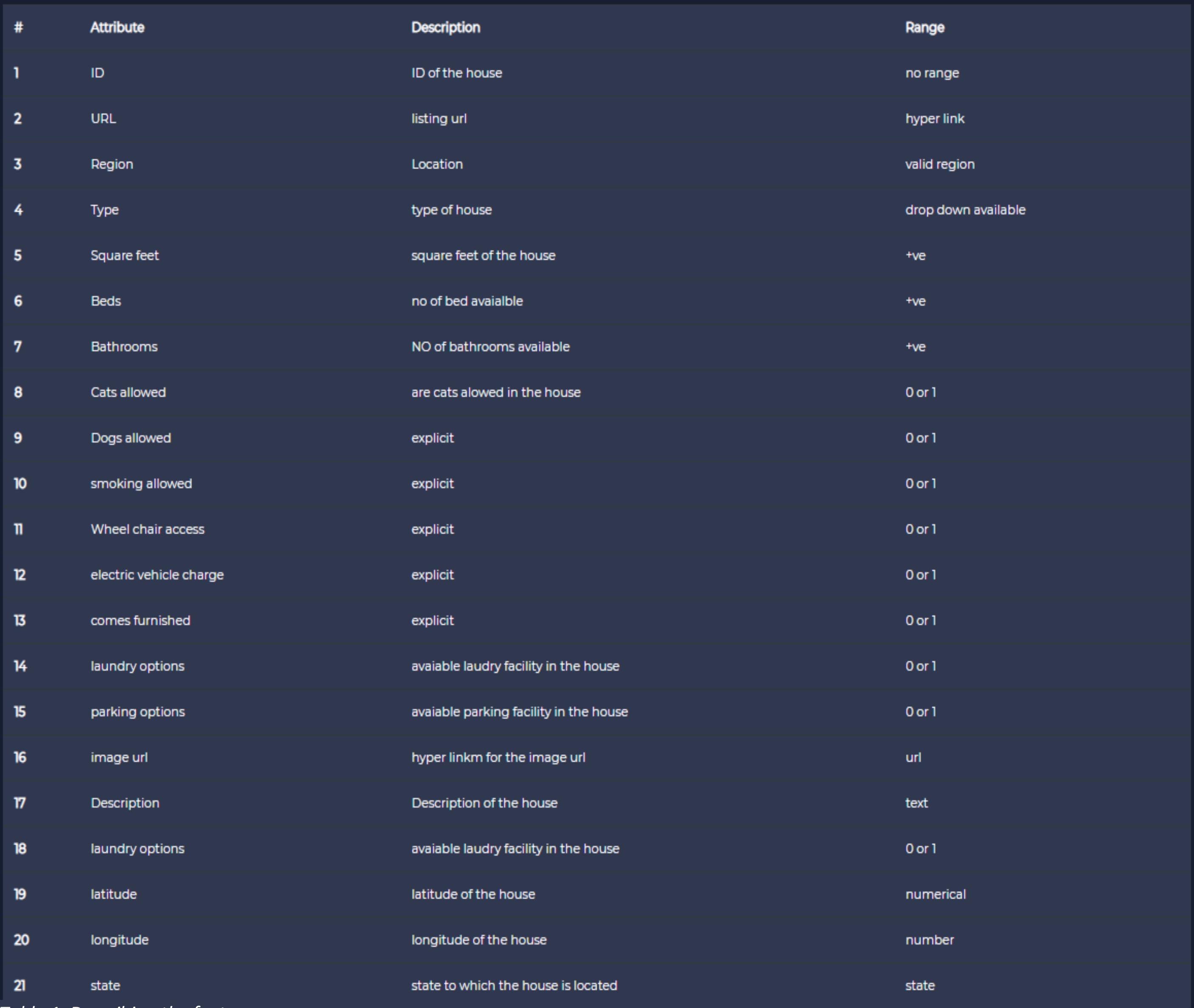
House Rent Prediction

Data descriptive analysis

The training data has 265190 data points and22 variables. Data matix has 152287 missing cells(2.6%). There are 7 categorial data, 6 boolean, 2 url and 7 continuous   data features. The table descrbing the features are showin in the table 1.



#	Attribute	Description	Range
1	ID	ID of the house	no range
2	URL	listing url	hyper link
3	Region	Location	valid region
4	Type	type of house	drop down available
5	Square feet	square feet of the house	+ve
6	Beds	no of bed avaialble	+ve
7	Bathrooms	NO of bathrooms available	+ve
8	Cats allowed	are cats alowed in the house	0 or 1
9	Dogs allowed	explicit	0 or 1
10	smoking allowed	explicit	0 or 1
11	Wheel chair access	explicit	0 or 1
12	electric vehicle charge	explicit	0 or 1
13	comes furnished	explicit	0 or 1
14	laundry options	avaiable laudry facility in the house	0 or 1
15	parking options	avaiable parking facility in the house	0 or 1
16	image url	hyper linkm for the image url	url
17	Description	Description of the house	text
18	laundry options	avaiable laudry facility in the house	0 or 1
19	latitude	latitude of the house	numerical
20	longitude	longitude of the house	number
21	state	state to which the house is located	state

Table 1: Describing the features

Heat map showing the null values are shown in the figure 1. As stated above there are a total of 2.6% missing cells.

The considerable number of missing values are found in parking_options and laundry_options. Since the features may have large weightage in the prediction model. The missing values are filled with mode values.

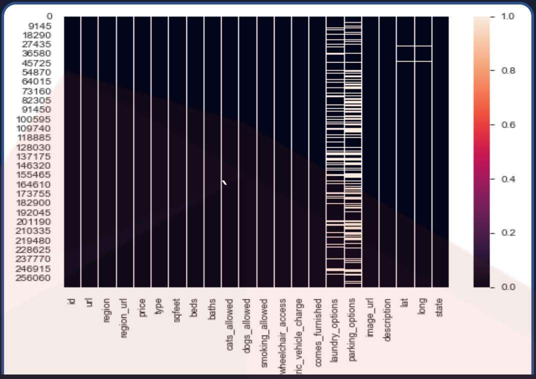


Figure1. heatmap depicting the null value in the data matrix

Feauture analysis

Type: There is a considerably large number of houses of type apartment. 82%of the houses are of tyoe apartments.

Beds: average number of bedrooms in around 2.

Bathroom: averagre number of bathrooms is 1.5.

Cats and dogs: 71% of all houses allow both and dogs in the house.

Smoking allowed:around 70% of houses allow smoking in the house.

Wheelchair access: 92 %f the houses doesn't have wheel chair access.

Electrc vehicle charge:98% of the houses still doesn't have electric charge facility. ELeon musk has to work harder.

Lat-long parameters: both the parameters have a conectrated value in a particular region. Lat-long values are clustered using kmeans classifier and clubbed into single variable cluster. We use 3 clusters to classify the locations using the elbow curve.

Figure 2 showing the lebow curev is shown on the right.

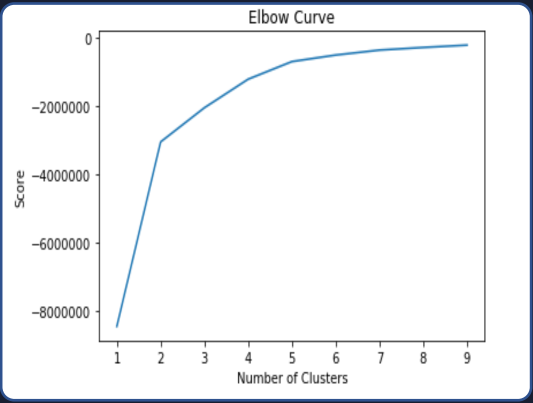


Figure2.Elbow curve showing the number of clusters were number of clusters and scores are compared

Correlation matrix

Heat map showing the Pearson correlation is shown in the figure 3. It is evdent that there is no correlation between the features. And we need to high number of estimators to fit non linear data.

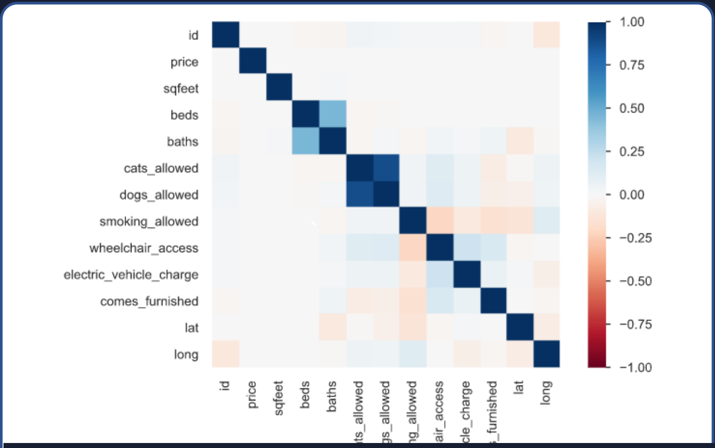


Figure3. heatmap showing the correlation between features