Pre -processing Data:

1. After the data has been retrieved to get the necessary details like those of restaurants in the area of AZ and OR.
2. Once, the data for the restaurants has been filtered
3. Print the schema of the data set (fig a)
4. Check for the row count in the data set. (fig b)
5. The next operation performed is to understand the pair-wise frequency on the columns (not restricted to numeric columns) (fig b)
6. Next step, is to put the necessary columns which are required for modelling into a data frame and work on the pre-analysis stage of the modelling (fig c)

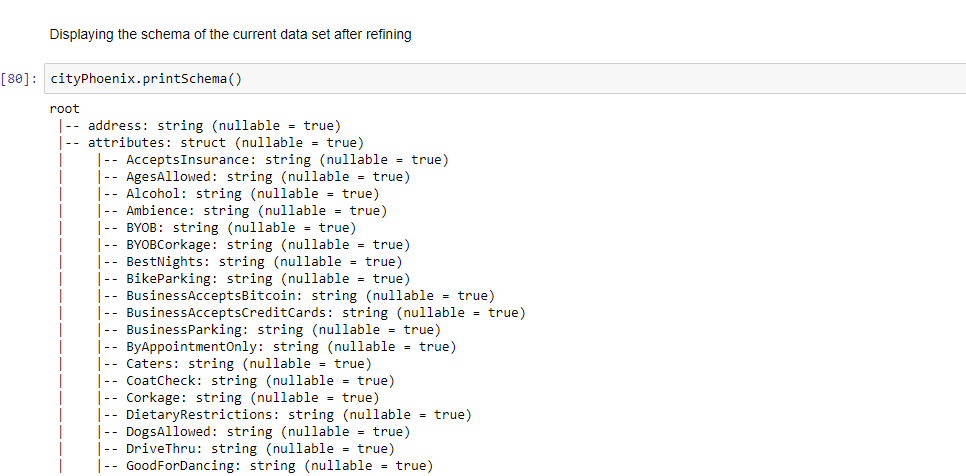


Figure a

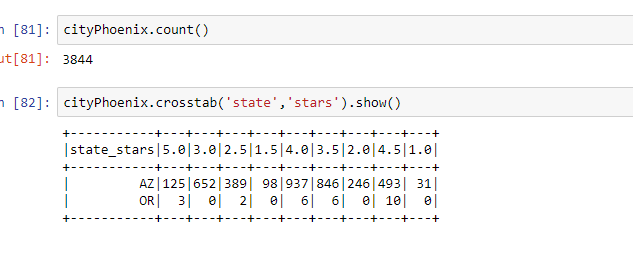


Figure b

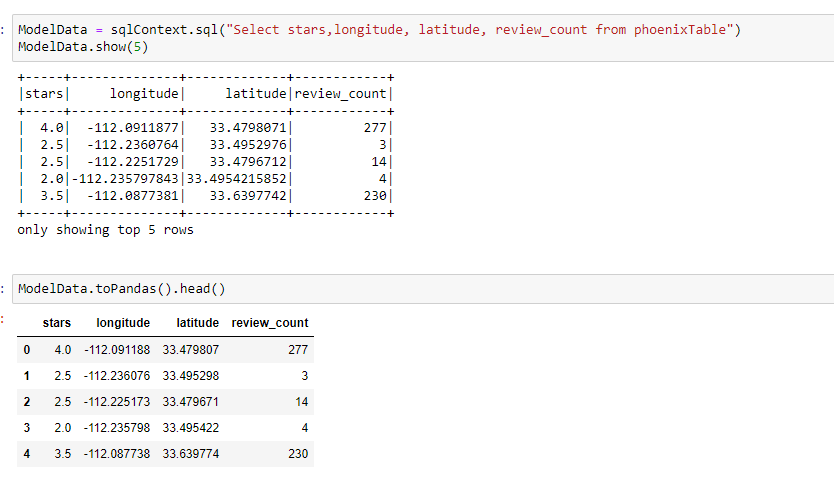


Figure c

Pre- Analysis

Analysis

1. Transform the data from a data frame into an RDD of Labelled points. (fig d)
2. Perform statistical operations of the data, to find the mean and variance along with non-zero rows (fig e)
3. Find the correlation between the variables we are going to work (fig e)
4. Plot a box plot to understand the distribution of the variable -review\_count along with checking the number of outliers, mean etc (fig f)
5. Plot between the variables star and review\_count to understand the relation and the interaction between the variables (fig g)
6. Plot histogram of the variables- stars and review\_count to eye-ball if they are normally distributed or not (fig h)

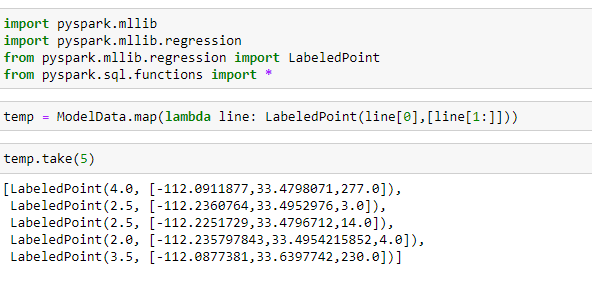


Figure D

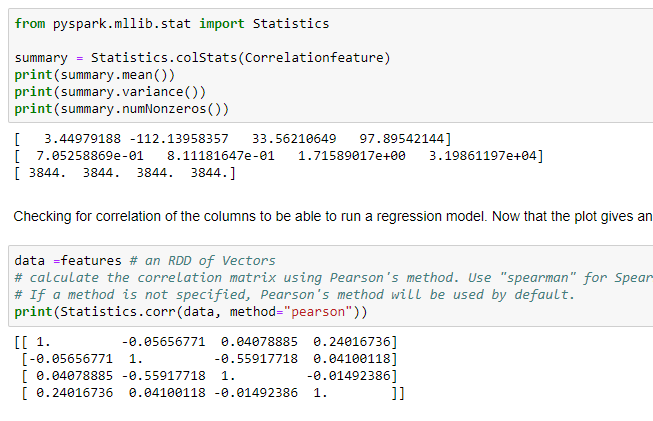


Figure E

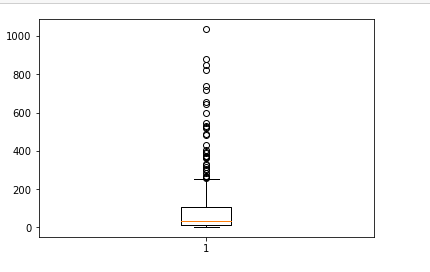


Figure F

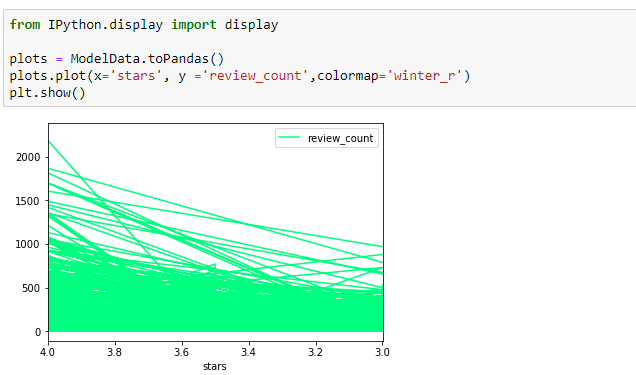


Figure G

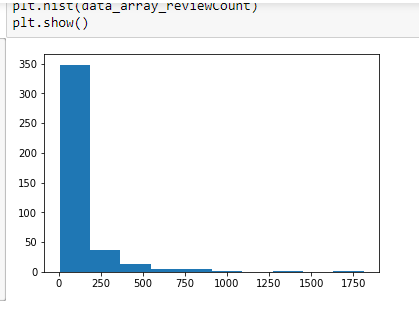


Figure h.1

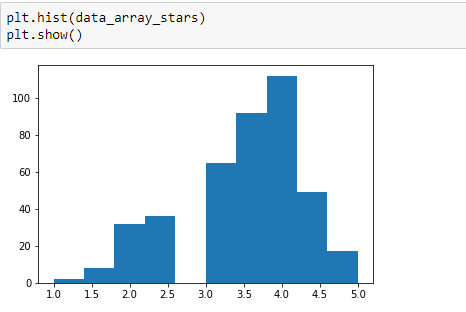


Figure h.2

Modelling -Linear Regression Model (fig I)

1. The Data transformed into the required format – RDD of Labelled point
2. Divide the data into train and test data set
3. Apply the linear regression with SGD command available in the Machine learning package
4. Check for the output
5. Use the generated output, to predict future values.

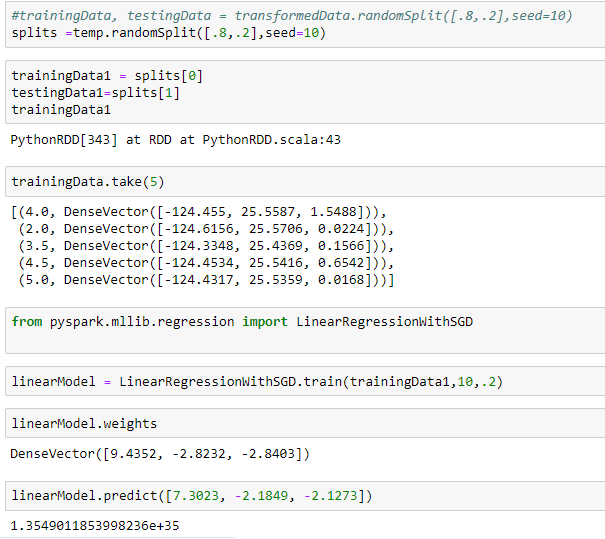


Figure I

Challenges Faced:

The first challenge was to do find the command for transforming the data

Once the data has been transformed into an RDD of labelled points, trying to execute the commands referenced by google did not help in executing the task.

Had to search a lot to find the right kind of input for a certain command

Though google, has some references, since they are not belonging to the version we are currently, working a lot of effort went into arriving at the result.