

Q4

1.

For poisson regression, the likelihood function can be written as

$$p(X, Y, \theta) = \prod_{i=1}^N \frac{e^{y_i \theta x_i} \cdot e^{-\theta x_i}}{y_i!}$$

That implies log likelihood as

$$\log(p(\theta|X, Y)) = \sum_{i=1}^N (y_i \theta x_i - e^{\theta x_i} - \log(y_i!)) = \sum_{i=1}^N (y_i \theta x_i - e^{\theta x_i}) + \text{constant}$$

MLE estimate of  $\theta$

$$\text{argmax}_{\theta} \log(p(\theta|X, Y))$$

i.e., theta that maximizes the log-likelihood

Therefore, the loss function that needs to be minimized is

$$-\log(p(\theta|X, Y)) = \sum_{i=1}^N (e^{\theta x_i} - y_i \theta x_i)$$

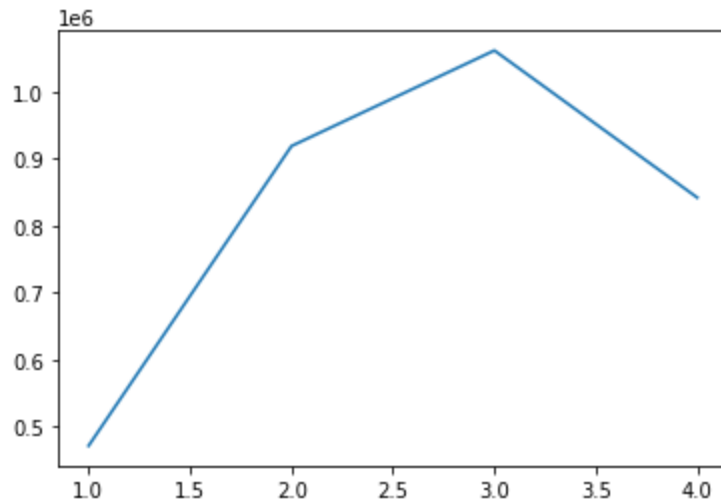
2. To understand the statistics of the count of bikes in the given Bike Sharing Demand dataset, mean count of bikes used per hour, per month, per year is calculated. The below table compiles the findings

	Mean value
Per year	1646339.5
Per month	182926.6111
Per hour	189.463

3. The dataset consists of 16 dimensions out of which the date as string is removed, to ensure the data is all numerical. Alternatively, we can parse the string, but since the details are present at the level of an hour for each day, parsing a date will only cause redundancy  
Thus the dataset used for our experiments consists of 15 dimensions.

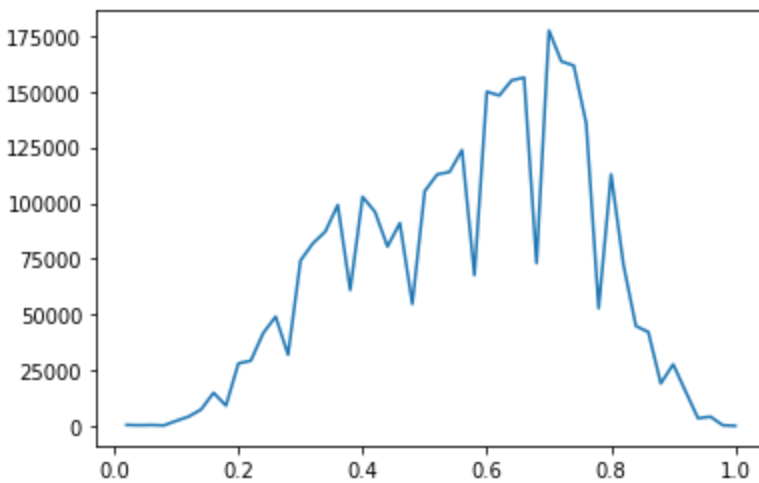
We tried understand the behaviour of 5 selected features (season, working day, temp, windspeed, casual) out of 15 and plotted count against each of these features

Below are the findings and inference we can draw from the graphs  
season vs count



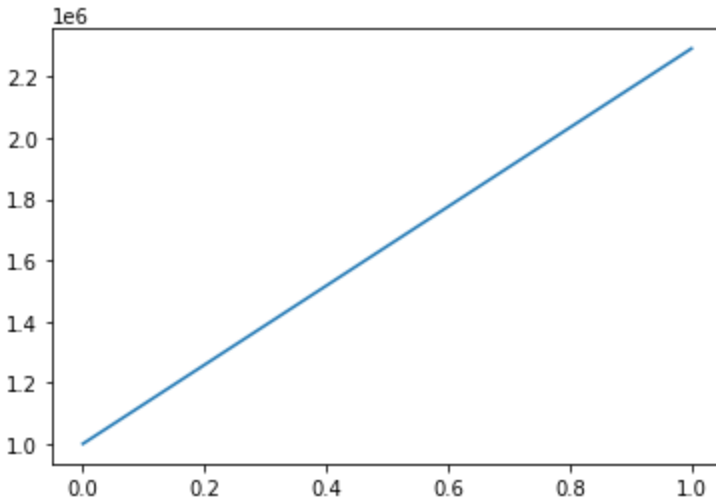
There are four seasons out of which 3rd season is most favourable for bike riding and 1st season is the least favourable

temp vs count



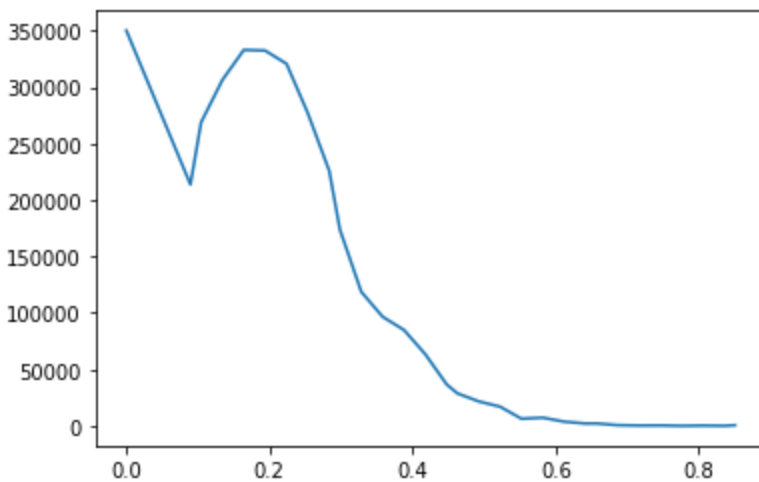
We can infer that temperature effect on the bike demand is not a fixed behaviour, but more sunny is preferred

working day vs count



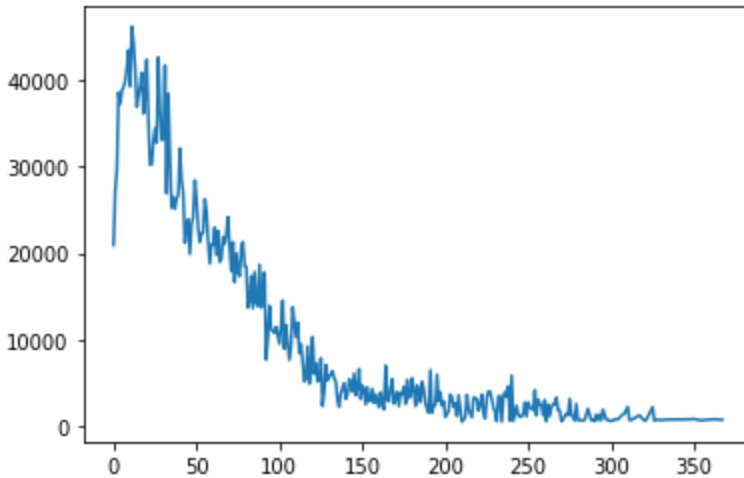
The above graph shows that bike demand is higher during the working day rather than holiday. This implies that most of the people might be transiting to work on bikes.

windspeed vs count



Lower the wind speed higher the demand. But we can observe a falling steep from 0 to 0.1 units. People preferring to have a little breeze during bike driving might be a reason

casual vs count



The above graph shows that most of the demand comes from the people who registered in a bike sharing system instead of people who apply casually now and then without registration.

Thus, the data might be collected at a place where people prefer more to transit using bikes.

4.

Dataset is split into 3 parts

Firstly for total dataset 80-20 split is done into training set and testing set

The obtained training data is further is split 80-20 into training and validation sets

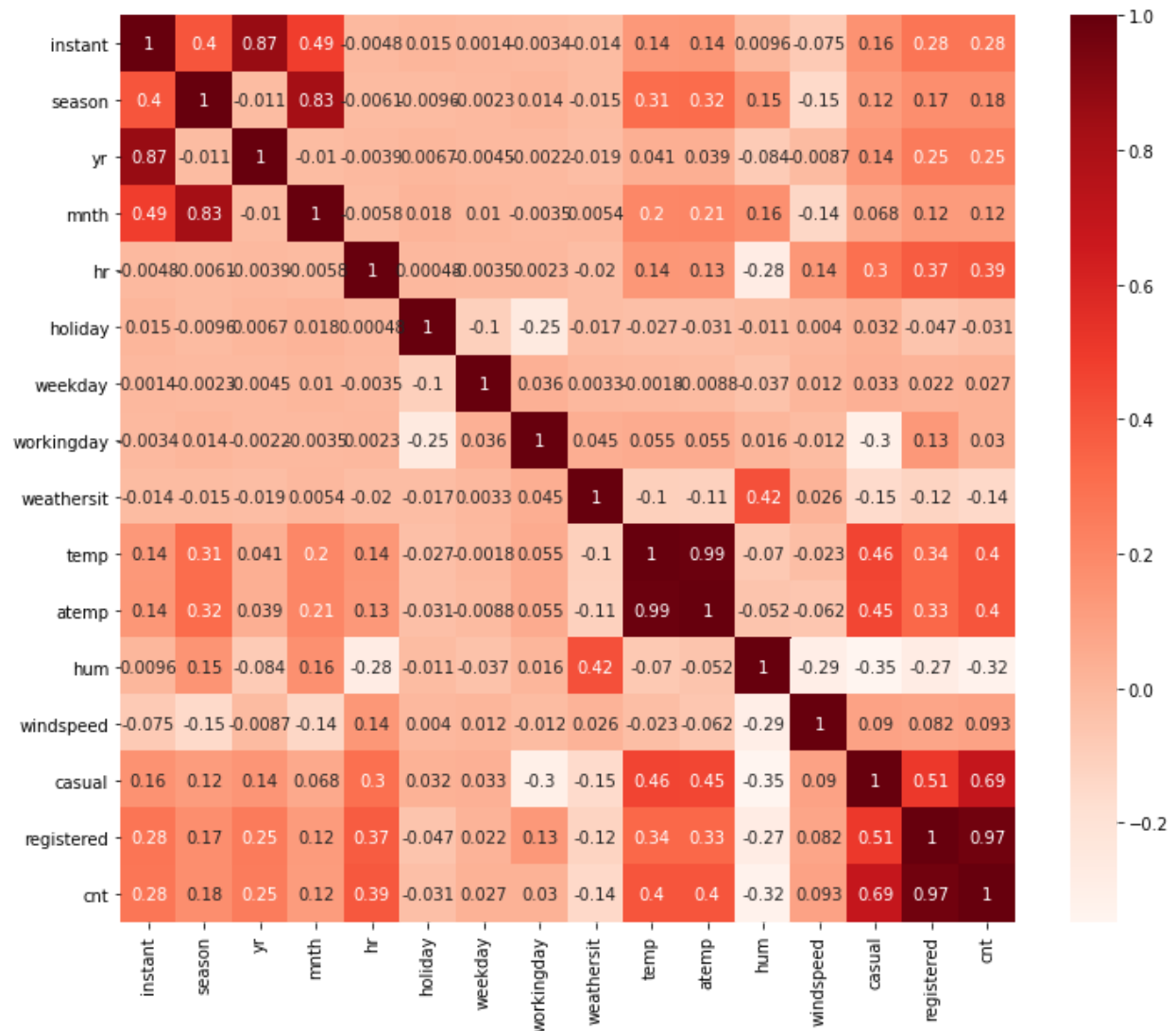
The accuracy is reported in terms of the RMSE values on the testing set. Higher the RMSE value, lower the accuracy

For l1-norm and l2-norm regularizations, values are reported based on the best hyperparameter found on the validation set.

	RMSE values
No regularization	9059.956
l1-norm regularization	8834.383
l2-norm regularization	8847.06
both l1 and l2 norm regularization	8913.57

5.

A heatmap is collected to check the correlation between features of the dataset.



This shows that the count of bikes is more correlated with registered, temp, hr, yr, season. Thus we can conclude these form some of the important features to which count is highly correlated with (based on the dataset)