Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

There are a large number of categorical variables in the data set and so it was evident that some categorical variables will have an effect on the dependent variable. The categorical variables has been used to classify things like seasons, weather, day of the week, holidays etc., so when we are estimating the demand of bikes, some of these variables logically should impact the demand which is the dependent variable.

Once the model was developed it was identified that season (specially summer and winter season); month (specially July and September) and weather (specially the rains) has an impact on the demand of the bike.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

One categorical variable can be represented by (n-1) dummy variables where n stands for the number of unique values that the categorical variable can take.

When we use drop_first=True while creating dummy variable from a category variable, it drops the first column of the 'n' dummy columns created for one categorical variable resulting in 'n-1' dummy variable columns. So, we achieve in creating n-1 dummy variables for one categorical variable.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

'temp' , 'atemp' --- which stands for temperature in Celsius and feeling temperature in Celsius.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

By performing residual analysis and showing that residual value (or error) is normally distributed with a mean of zero. This validates the assumption of homoscedasticity of error and also the normal distribution of error term for linear regression.

Residual analysis is performed by estimating the error which is the difference between the actual value of the dependent variable and the prediction of the dependent variable based on the linear regression model. This difference or the error is then plotted in the form of a histogram and if the error term is normally distributed with a mean value of zero, it validates that the residual is normally distributed and homoscedastic.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

i) 'temp' – Temperature in Celsius

ii) Value of 2 in season i.e. whether it is winter season or not

iii) 'hum' – humidity

General Subjective Questions

1. Explain the linear regression algorithm in detail? (4 marks)

Liner regression is an algorithm to fit a linear relationship between a dependent variable (y) and a set of independent variable (x1, x2,x3.....). This is achieved by first identifying the coefficients (b0, b1, b2, b3 ....) for all the independent variables through a process of minimizing the mean squared error, to identify the best fit possible for the relationship:

Y = b0 + b1 * x1 + b2 * x2 + b3 * x3.....

The mean squared error is the squared difference between the predicted value of Y (dependent variable) and the actual value of Y, divided by the no. of independent variables (say m).

So, for a set of coefficients (bo#, b1#, b2#.....) Y predicted = bo# + b1# * x1 + b2# * x2....

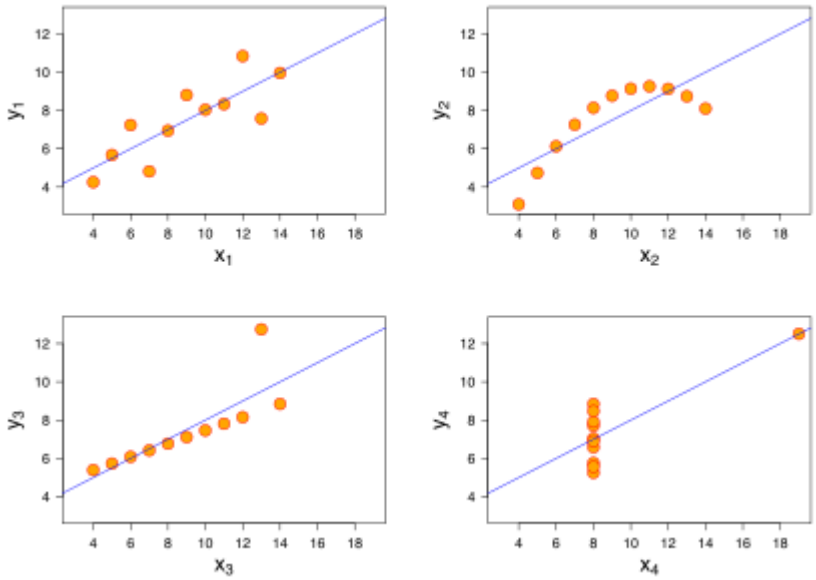
Mean squared error = (1/m) * sum of squares of (Y – Y predicted)

The mean squared error is the cost function and algorithm runs to predict the values of the coefficients (b0, b1, b2, b3.....) such that the cost function is minimal – which is the best fit of the linear regression between the independent and the dependent variable. This is called the OLS method of identifying the best fit for a linear regression.

There are some assumptions in this: (i) The error (Y – Y predicted) should be normally distributed with a mean of zero (ii) collinear relationship between the variables x1, x2, x3 .... should be removed wherever present (iii) preferably the data x1, x2, x3.... should be scaled to be between 0 and 1 or standardized to a mean of 0 and std. deviation of 1 (iv) Any outliers in the data should preferable be removed as outliers affect the linear regression fit.

2. Explain the Anscombe's quartet in detail. (3 marks)

The Anscombe's quartet represents a set of 4 different datasets which have similar descriptive statistics however the distributions are quite different. The following graph helps to explain the same:



The importance od Anscombe's quartet is specifically significant in understanding the limitation of linear regression. All the 4 data sets have the same linear regression line that can fit the data. Amongst the top 2 data sets (in the graph); 1st data set is a good fit for the linear regression model however for the 2nd data set linear regression is not the right model because of non-linear nature of the distribution; the bottom 2 graphs indicate that the linear regression is not the right model because of the presence of outliers.

3. What is Pearson's R? (3 marks)

Pearson's R is the correlation coefficient that measures the linear correlation between 2 datasets. It is the ratio between the co-variance between the two variable and the product of their standard deviation.

So, Pearson's R = cov(X,Y)/ sigma(X) * sigma (Y)

Pearson's R correlation is widely used in regression. However one needs to understand that if two variables are correlated, it is very much possible that they have some other sort of relationship and not just a linear one.

Pearson's R correlation coefficient is designed for linear relationships and it might not be a good measure for correlation if the relationship between the variables is non-linear.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is done on an input data to bring all the features of the data (all variables) to a similar scale.

Real-world datasets often contain features that are varying in degrees of magnitude, range, and units. Therefore, for machine learning models like linear regression algorithms to interpret these features, they need to be on a similar scale.

There are 2 types of scaling – normalized and standardized

Normalized scaling, also called MinMax scaling helps to get all the datapoints of a feature between 0 and 1. It is performed by identifying the min value and the max value of a feature. Then each value x is scaled as: (x-min)/(max-min).

Standardized scaling is the way of scaling the dataset of a feature to a mean of zero and standard deviation of 1. If the mean of a dataset in a feature is M and standard deviation sigma; a value x in the dataset is scaled as: (x-M)/sigma

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

It happens when there is an extremely high multicollinearity between one variable and other independent variables in a dataset (to the extent that they are almost the same). Multicollinearity between variables are measured with VIF. Where VIF for variable x is defined as 1/(1-Rsquared for regressing x with all other independent variables). Now suppose x is highly related with other independent variables, the value of "Rsquared for x regressing with other independent variables" will be almost 1. So VIF will become 1/(1-almost 1) = 1/ almost 0 = tending to infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A Q–Q plot (quantile–quantile plot) is a graphical method for comparing two datasets by plotting their quantiles against each other. A point (x, y) on the plot corresponds to one of the quantiles of the second dataset (y-axis) plotted against the same quantile of the first dataset (x-axis).

Q-Q plots are a good plot to identify linear relationship between 2 datasets (each dataset being represented by each axis) and so to identify if a linear regression can be plotted between the 2 datasets. If the two datasets being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the two datasets are linearly related, the points in the Q–Q plot will approximately lie on a line (not necessarily on the line y = x).