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)

Ans. Categorical variable did place a role in predicting bike rentals in this case. Categorical variables were encoded into dummy variables and then it was used for model prediction. It was observed following categorical are predictive variable for data set:

- Monday from 'weekday'
- August and September from 'mnth'
- Summer and Winter from 'season'
- Sunny and Light Rain from 'weathersit'

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

Ans.

- I initially started my model building process by converting categorical variables to dummy variables, n dummy variable columns for n categorical values.
- When VIF was calculated for this model, it comes out to be infinite. For linear regression, we look for VIF less then 5, and infinite is not a good value for a model.
- When 'drop_first=True' was used in dummy encoding, the resulted in finite VIF value.

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

Ans. atemp

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

Ans.

- By performing residual analysis, assumptions of linear regression were validated.
- To check for the normal distributions, histogram of error terms is plotted. A normal distributions of error term is the expected result.

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

Ans. 'atemp', 'winter season' and 'sept month'

General Subjective Questions

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

Ans.

Linear regression algorithm is based on supervised learning method. It is regression model used for predictive analysis. It is a method of fitting the best fit straight line for the dataset, which means it is about finding the best linear relationship between dependent and independent variable. It is mostly done by the Sum of Squared Residual method.

Linear regression is classified two parts:

1. Simple linear regression
2. Multiple linear regression

Simple linear regression – This explains the relationship between dependent variable and one independent variable.

Multiple linear regression – This explains the relationship between dependent variable and more than one independent variable.

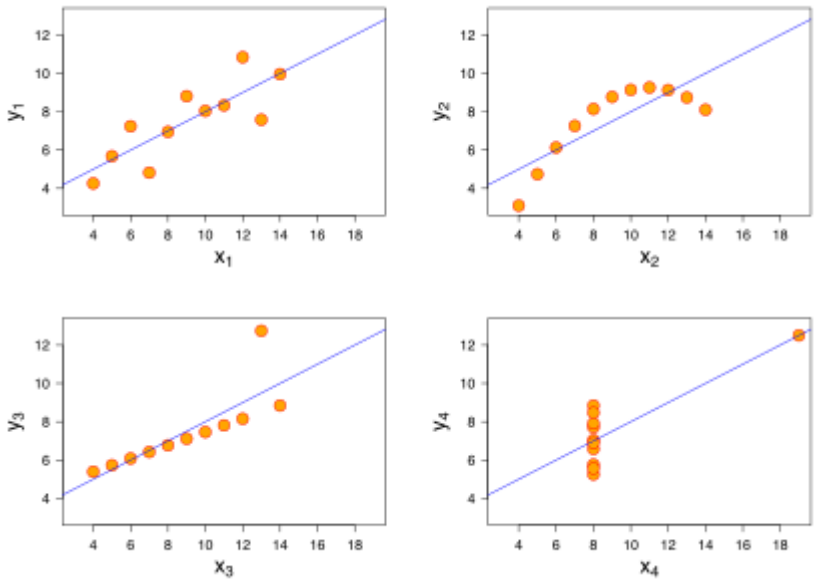
3. Explain the Anscombe's quartet in detail. (3 marks)

Ans.

There are few shortcomings of simple linear regression model as follows:

1. It is sensitive to outliers.
2. It predicts only a linear relationship
3. Certain assumptions are made to make inference.

These outcoming of linear regression can be better explained using Anscombe's quartet



Here, it can be observed that all four quartet have very different distribution and yet produce identical statistics.

1. First graph does have a linear relationship
2. Second graph – It is not at all a linear graph, it is rather a curve.
3. Third graph – This graph looks linear, but there is an outlier in this.
4. Fourth graph – In this graph, data points does not possess any relation with independent variable.

Hence, it is always advised to have a graphical visual of your data before considering it for linear regression.

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

Ans. Pearson's R is correlation coefficient in linear regression.

Formula for pearson's R is:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

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

Ans. Scaling means putting feature variables in same range scale for prediction. When scaling is not performed on feature variables, it does not impact F-statistics and R squared value, but it does effects correlation coefficient values of feature variable.

Scaling is performed to normalize correlation coefficient of feature variables. In your predictive model if you are also concerned to find the correlation coefficient between dependent and independent variable scaling is required.

Normalizing scaling: It is scaling technique in which values are ranged between 0 and 1.

Formula:

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardized scaling: In this scaling technique data points centered around the mean with a unit standard deviation.

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
(3 marks)
Ans. That happens when dummy encoding is performed on categorical variable and all the dummy variables are used for modelling. One need to use n-1 dummy variables for n dummy encoded variables.

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

Ans.

- Q-Q plots are used to find the type of distribution for a random variable.
- It is graphical method for comparing two probability distributions by **plotting.**