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)

- Season 3 has high count followed by the season 2 and 4 has the high values
- mnth 5,6,7,8,9,10 has significant count. Median between 4000 to 6000. So, month also a good predicator for the dependant variables
- weathersit has median value more than 4000 for 1. 2 and 3 has significant difference. This also a good predicator for the dependant variables
- Huge bike booking happened on the working day than holiday. So, holiday is not a good predicator for the dependant variables
- There is no significant change seen across the weekday and working day. This may or may not be a good predicator for the dependant variables

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

- It helps to reduce the number of dummy variable creation by 1. For example, if there are 3 variables, it will create 2 features instead of 3 if drop_first=True applied. So, reducing one feature here.
- Also, this will help to reduce the correlations created among the dummy variables.

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

- temp, and atemp has the high correlation with the target variable.

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

- Validate the residuals using the histogram plot with the y_train, and y_train_pred differences. The Residuals are normally distributed.
- Validated the VIF values are less than 10 to make sure there is less or no multicollinearity

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

- Temperature - When temperature increase one units, the bike renting numbers will increase 0.5499 units
- Weather - When weathersit - 3 increase one units, the bike renting numbers will decrease by 0.288 units Note: weathersit - 3 is Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- Year - Based on this model outcome, the bike renting numbers will decrease by 0.233 units on every year

General Subjective Questions

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

Linear regression is a one of the models in Machine Learning. In this model, we can predict the behaviour of the data set using some variables. This model will work only on the variables which are linearly correlated each other.

This linear regression straight line equation is formulated as y = mx + c

Here,

Y = how far the value up

m = slope (or) Gradient

x = how for long

c = Intercept (Value of y when x=0)

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

Anscombe's quartet is to illustrate how to not relay on the statistical measures during the data analysis. Anscombe's created 4 data set which gives nearly identical statistics measures. However, when graphically the data was not scattered linearly. So, we need to see the graph along with the data coefficient.

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

Pearson's R is also known as the bivariate correlation. It is the covariance of two variables, divided by the product of their standard deviations; As it is a normalised measurement of the covariance, the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope
- r = -1 means the data is perfectly linear with a negative slope
- r = 0 means there is no linear association

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

- Scaling is helps to normalize the data into a particular range.
- Scaling is performed to make the feature to fit in a range, so that algorithms consider the features units as well as the values. Otherwise, the model will consider only the values and the model might go wrong.
- normalised scaling is brings the values between 0 and 1. It is also called min-max scaling.
    - Min-max Scaling X = x-min(x) / max(x) – min(x)
- Standardized scaling is brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
    - Standardization: X = x – mean(x) / sd(x)

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

Value of VIF is infinite when there is perfect correlation between two independent variables. In this case the R^2 = 1. So VIF is calculated as 1/(1-R^2) which is 1/0 = infinite.

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

Q-Q plot is a scattered plot create between two different quantiles against each other. The first quantile is the variable that testing the hypothesis for the second one is actual distribution that is test it against.

The data points align closely in the straight 45-degree line. If the data points are lie approximately in the straight line, then it is normally distributed. If the data points are not lies on the straight line, then it is not normally distributed.

Also, this helps to test distribution among two different data set as well.