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: The categorical variables given are the seasons, months, weathersit, holiday, weekday, workingday.

***Seasons**: The Fall season has the greatest number of bookings than the other seasons. Also, the winter has the least bookings.*

***Month**: The months in the middle of the year (5^{th},6^{th} and 7^{th} month) and slightly at the end of the year (8^{th} and 9^{th} month) has more number of bike sharing. It could be due to the season or holiday time.*

***Weather:** Clear, Few clouds, Partly cloudy and Partly cloudy has the most count in bike sharing. This could be completely due to the nice weather.*

***Holiday:** More number of bike sharing is observed in the non-holidays. This could be due to people who are going to office, colleges and schools. Sometimes, people prefer to take it regularly.*

***Weekday**: There is no significant differences in the Weekday and WeekEnd*

***Workingday:** Working day has more number of Bookings. As we know, This could be due to people moving from place to place for their works.*

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

Ans: Creating the dummy always end up with (n-1)th variable. For example, if we have three categories furnished, unfurnished and semi-furnished. It is good to go with two choices. (1 ,1) for furnished, (0,0) for unfurnished and obviously (1,0) for semi-furnished. So, Dropping the first variable would help us proceed with relevant features or columns.

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

Ans: registered is the variable having the highest correlation value.

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

Ans :

- *Taken the Error Terms to analyze the data and plotted it for Normal distribution.*
- *This was done by the residuals of the training data. The Residuals were normally distributed. So concluded the model is significant.*
- *There is also Linear Relationship between the X and Y. This was made by the numerical data.*
- *There is Linear Relationship between the temp, atemp and the 'cnt' variable.*
- *And there is no multi-collinearity. This was analysed by calculating the VIF.*

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: The top three features were concluded as weathersit, yr and temp

General Subjective Questions

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

Ans : Linear Regression is a supervised learning in Machine Learning. This algorithm is used to find the linear relationship between two variables i.e., X and Y. X is the independent variable which is also called as the predictor variables. Y is the dependent variables. X is the variable which decides the changes in the y.

Together , it is explained in the equation of y= mx+c

Where m is the slope or the co-efficients and

C is the constant. And this forms the straight line if the variables are completely related to each other.

Linear Regression can be two types – Simple Linear Regression and Multiple Linear Regression.

***Simple Linear Regression** - This is analyze the data single variables. Y = mx +C*

***Multiple Linear Regression** - This is to analyze the data with multiple variables. Y = c + m1x + m2x+..mNx*

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

Ans: *Anscombe's quartet were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.*

It is associated with four types:

1. Simple Linear Regression - represents two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
2. Non -Linear - The graph would be slightly curved , explains the regression and the corresponding coefficient of determination.
3. Tight Linear Relationship between x and y except for one large outlier.
4. The value of x remains constant , except for one outlier as well.

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

- *Pearson's R is a measure of linear correlation between two data sets. It also states about the ratio between the covariance of two variables and the product of their standard deviations. Hence it is a significant normalized measurement of the covariance, And Hence the value always between -1 and 1.*

- *The pearson's measure reflects only in the Linear correlation and not with types like relationships and correlations.*

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

Ans:

Scaling is a method of normalising the numbers from 0 and 1.(or within a range) Or in simple terms to make it concised.

Why Scaling ? . If scaling is not done , then the data become redundant. collected data set may different units and measurement. So scaling is performed to normalize it.

normalisation brings all the data between 0 and 1. Formula = x-xmin/xmax-xmin

Standardiztion is to keep the data in z scores. it brings all the data into a standard normal distribution. Formula = x-mean(x)/sd(x)

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

   Ans : VIF will result in infinity when the correlation is perfect. It means there is extremely good correlation between the x and y. In this case the R2 value will be 1. this gives the 1/(1-R2) to infinity. So dropping one of those values would give us the better results by avoiding the multi-collinearity.

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 called Quantile-Quantile plots. These are plots of two quantiles against each other. which means, A quantile is a fraction where some of the values will lies below that quantile. considering an example, where the median is a quantile where 50 percent of the data is lying below that point and 50 percent lies above it. So, the Q-Q plots figures out two data sets come from the same distribution. A 45 degree angle is plotted on QQ plot to find the points falling on the reference line when the two data sets come from a common distribution.