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 Variables, Weekday, Season and Weathersit are having significant effect on the dependent variable. On Weekday = 1 to Weekday = 5 there are more bookings than holiday and weekday = 0 and weekday = 6. Coming to Season, there are more booking when season = fall

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

Ans. The main reason for using drop_first = True, is to make the number of dummy values to n-1. Hence, to remove the redundancy and remove the co-relation.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation

with the target variable? (1 mark)

Ans. Casual and Registered columns are having the high correlation compared to others.

4. How did you validate the assumptions of Linear Regression after building the model on the

training set? (3 marks)

Ans. We can validate the assumptions of linear Regression, by looking into the summary, where we see the P value and R squared values. Low P values shows the high significance. A good model will have high adjusted R Square.

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. Temp, Season and weekday seems to be contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

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

Ans. Linear regression is a supervised machine learning algorithm. It is used to predict the target varaible based on the independent variables. The simple linear regression expression is $y = \beta 0 + \beta 1x$ where m is the coeffecint of x and c is the intercept. we train the model to get the best fit line to predict the y. This is done by finding m and c values. We can perform Simple linear regression which is done using only one independent variable. Multiple linear regression done by using multiple independent variable.

The equation for multiple linear regression is $y = \beta 0 + \beta 1x1 + \beta 2x + \varepsilon$ … ….

here $\beta 0$ = intercept

$\beta 1, \beta 2, \beta 3$= coeffects of x1, x2, x3...

ε = y - y_{pred} --→ error

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

Ans. Anscombe's quartet will consist of four data sets, which has identical simple descritptive statistics but it will have different distributions and different graphs when plotted. There will be some pecularities in the data sets that fools the Linear regression model if build. when we check the staistics like standard deviation, variance and mean will be same. It describes the importance of visualizing the data before analyzing and building the model. All the important features should be visualized before implementing any machine learning algorithm.

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

Ans. Pearson's correlation coefficient r or Pearson's r is a measure to determine the relationship between two quantitative variables and the extent to which two variables are linearly related. A correlation coefficient value 1 means the wo variables are positively related and -1 means the variables are negatively related. Value zero indicates that there is no relation between the variables.

Equation:

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

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

Ans. Scaling means transforming the data so that it fits within a specific scale. For example, between 0 – 100 or 0-1 or -1 to 1. Scaling is performed in order to get rid of the uncertainty in the outcome due to large difference between the datapoints. Scaling makes the model to learn and understand the problem.

Normalization is the technique in which values are rescaled so that there range lies between 0 and 1.

Equation: $$X' = \frac{X - Xmin}{(Xmax - Xmin)}$$

The minimum value in the column becomes 0, the maximum value becomes 1. The values between max and min will be a value in between 1 and 0

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attributes becomes 0 and the resulted distribution will have unit standard deviation

Equation: $$X' = \frac{X - \mu}{\sigma}$$

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

Ans. When VIF is infinite, it means that there is a perfect correlation between the two variables. When two variables are perfectly corelated then the R2 will be 1. So, 1/1-R2 become infinite. Whenever we see the value of VIF is infinite, we should drop one of the variables.

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

Ans. Quantile-Quantile plot is called as Q-Q plot, this is used assess if a set data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. It is mainly used in

Linear regression to check if the training data set and test data set are from same populations and are with the same distribution or not, when we received them separately.

Main advantages of using Q-Q plot:

1) We can detect the shift in location, shifts in scale, change in symmetry and presence of outliers
2) I case of two data sets, we can use it to find
    a. If they came from population with common distribution
    b. Have common location and scale
    c. Having similar distribution scale
    d. Having similar tail behavior