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)

Answer) 1. Season: 3: fall has highest demand for rental bikes

2. I see that demand for next year has grown

3. Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing

4. When there is a holiday, demand has decreased.

5. Weekday is not giving clear picture abount demand.

6. The clear weathershit has highest demand

7. During September, bike sharing is more. During the year end and beginning, it is less, could be due to extereme weather conditions.

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

Answer) drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation.

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

Answer) temp

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

Answer) Plotting different graphs used for those assumptions. Kindly refer the notebook for reference.

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

Answer) temp, holiday and season

General Subjective Questions

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

Answer) We started with splitting the data set into training and test dataset. We built the model using both RFE and Automated approach for selecting features as well as manual approach to eliminate and select features. This was achieved by checking VIG and p-values.

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

Answer) Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.

3. What is Pearson’s R? (3 marks)

Answer) Pearson correlation coefficient — also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient — is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

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

Answer) Machine learning algorithm just sees number — if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So these more significant number starts playing a more decisive role while training the model.

The machine learning algorithm works on numbers and does not know what that number represents. A weight of 10 grams and a price of 10 dollars represents completely two different things — which is a no brainer for humans, but for a model as a feature, it treats both as same.

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

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

Answer) Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. First, the set of intervals for the quantiles is chosen. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile.