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?

Ans1. Following inferences were drawn from dataset:

- Season-Fall has highest demand for rental bike.
- Second year the demand of bikes has increased.
- Bike demand goes high in summers like September has maximum demand and drops in winters.
- No clarity on working day
- Good weather condition raise the demand of bikes.
- Demand of bike rental increases with temp and atemp.
- The bike demand is negatively correlated with windspeed and humidity.
- cnt is highly correlated with day_count.
- Data of windspeed is scattered and doesn't seem to contribute much.

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

Ans2. Dropping the first dummy variable helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Ans3. Temp and atemp.

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

training set?

Ans4. By testing the Linear regression model built on the test dataset. The predicted dependent had a very good coherence with the train dataset.

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?

Ans5. Based on final model following 3 features are the top contributors:

- Temp
- Weathersit_bad
- yr

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans1. Linear regression is a supervised learning algorithm, which means that it is trained on a dataset of labeled data. The labels tell the algorithm what the output should be for each input. In linear regression, the output is a continuous value, such as the price of a house or the number of customers who visit a store.

The algorithm works by finding the line that best fits the data. The line is called the regression line. The regression line is the line that minimizes the sum of the squared errors between the predicted values and the actual values.

2. Explain the Anscombe’s quartet in detail.

Ans2. Anscombe's quartet is a collection of 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 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".

The four datasets are:

- Dataset 1: The first dataset is a linear relationship between x and y, with a slope of 1 and a y-intercept of 0.
- Dataset 2: The second dataset is a quadratic relationship between x and y, with a parabolic curve.
- Dataset 3: The third dataset is a linear relationship between x and y, with a slope of 1 and a y-intercept of 1. However, there is one outlier point that is far from the other points.
- Dataset 4: The fourth dataset is a non-linear relationship between x and y, with a sinusoidal curve.

When the four datasets are graphed, they appear very different. The first dataset shows a clear linear relationship between x and y. The second dataset shows a parabolic curve. The third dataset shows a linear relationship, but with one outlier point. The fourth dataset shows a non-linear sinusoidal curve.

Despite their different appearances, the four datasets have nearly identical simple descriptive statistics. They all have the same mean, standard deviation, and correlation coefficient. This shows that summary statistics can be misleading, and that it is important to graph data before analyzing it.

Anscombe's quartet is a useful tool for teaching the importance of data visualization. It is also a reminder that summary statistics can be misleading, and that it is important to look at the data itself before drawing any conclusions.

3. What is Pearson’s R?

Ans3. Pearson's R, also known as the Pearson correlation coefficient, is a statistical formula that measures the strength and direction of the relationship between two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans. Scaling is the process to convert the numerical data on the same scale keeping the conversion ratio constant. This is performed to bring the homogeneity or uniformity in the distribution of the data without losing their actual sense.

Standardization centers data around a mean of zero and a standard deviation of one, while normalization scales data to a set range, often [0, 1], by using the minimum and maximum values.

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

Ans5. Infinite VIF indicates a perfect collinearity between the variables. This happens due to very high correlation between the variables/features.

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

Ans6. Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian distribution, uniform distribution, exponential distribution or even a Pareto distribution.

Type of distribution can be understood using the power of the Q-Q plot just by looking at it.