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 - Number of Bikes rented is high:

a. Season - Fall season is the top season where number of bikes rented is high.

b. Weather - Clear weather with few clouds number of bikes rented is high

c. Weekdays - Mid week bike rental is high

d. Month - In Mid year number of bikes rented is high

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

Answer- Drop_first=True is important to use as it helps reduce creation of extra columns during dummy variable creation hence decreasing the correlations created among dumb variables.

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

Answer- 'temp' and 'atemp' variables have a relationship.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Answer- We can validate the assumptions by below methods:

A. Fitted regression line is linear

B. Error terms were normally distributed when histogram was plotted with mean as 0.

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

Answer- The top 3 features include: Temperature(temp),weather situation(weathersit_3),Year(yr). The chances of increasing the number of rented bikes increases during the working day, in fall season and in clear weather, whereas demand for bikes is negatively affected by windspeed.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Answer- Linear Regression is a method of finding the best straight line fitting to the given data or finding the best linear relationship between the dependent and independent variables.

Its a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. Its used to predict values within a continuous range and is of two types :

A. Simple linear regression :(y=mx+c) : deals with one dependent variable.

B. Multivariable Linear Regression: Is a statistical technique that uses more than one variable to predict the outcome of a response variable.

2. **Explain the Anscombe's quartet in detail.**

Answer- Anscombe's quartet is a group of datasets that have same mean, standard deviation, regression line but having different representations when we plot them on a scatter plot. Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on statistical summaries. It emphasises the importance of using data visualisation to spot trends, outliers and other crucial details which are not very obvious from statistics summary alone.

3. **What is Pearson's R?**

Answer- Its the most common way of measuring a linear correlation. The value ranges between -1 and 1. It measures the strength and direction of relationship between two variables. Its also a measure of how close the observations are to a line of best fit. Its used when -

a. Both variables are quantitative

b. Variables are normally distributed

c. Data have no outliers

d. The relationship is linear.

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

Answer- Scaling is a way to transform your data into a common range of values. Scaling is important as it helps improve model performance, reduce the impact of outliers and ensures that the data is on the same scale.

Scaling is of two types: Standardizing and normalizing.

In Normalisation the values are shifted and rescaled so that they end up ranging between 0 and 1. Its also known as MIN-MAX scaling whereas in STANDARDIZATION values are centered around mean with a unit standard deviation.

Normalisation formula:

x=x-min(x)/max(x)-min(x)

Standardised scaling:

x=x-mean(x)/sd(x)

Normalisation is sensitive to outliers whereas standardization is less sensitive to outliers.

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

Answer- The value of VIF is infinite when there is a perfect correlation. The value of VIF is calculated using -

VIFi=1/(1-Ri^2) . If r-squared value=1 then the denominator becomes 0 and the overall value becomes infinite. It denotes perfect correlation.

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

Answer- The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution. Its a Q-Q plot is scatterplot created by plotting two sets of quantiles against one another.

The Q-Q plot is used to see if the points lie approximately on the line. If they don't it means residuals are not normal(Gaussian), and our errors are also not Gaussian. In the Q-Q plot sample sizes do not need to be equal, many distributional aspects can be simultaneously tested and it provides more insights into the nature of the difference than analytical methods.