**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?

Ans: 1) The count of rental bikes is highest during the fall and summer season

2) The count of rental bikes is highest in the year 2019

3) The count of rental bikes again is highest during the months from April to September

4) The count of renting is lesser if it’s a holiday

5) The count of rental bikes is higher if it’s clear weather

1. Why is it important to use drop\_first=True during dummy variable creation?

Ans: It is important to use drop\_first= True because this parameter specifies if the first category of dummy variable that is encoded must be dropped or not. If it is set to false then there is a high chance that the variables are correlated to themselves and will result in multicollinearity

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

Ans: Temperature has the highest co-relation with the target variable count

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

Ans: 1) Linear Model: The assumption between dependent and independent variables should be linear

1. No Multicollinearity: If the variables used for prediction are co-related then there is multicollinearity present
2. Equal Variances: The residuals obtained from the linear regression model must be spread evenly/homogenously, then it is called a satisfactory model. If the residuals are spread unevenly then it is not a satisfactory model.
3. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: The equation for best fitted line could be calculated as follows count = 0.5448 + 0.2477 X yr + 0.0708 X mnth\_Sep + 0.0563 X weekday\_Saturday + 0.0480 X workingday -0.0098 X mnth\_Jul - 0.044 X season\_Summer - 0.0571 X holiday - 0.0748 X season\_Winter - 0.0896 X Weathersit\_mist & Cloudy - -0.10244 X mnth\_Jan - 0.18783 X windspeed - 0.2602 X season\_spring -0.3044 X Weathersit\_light Snow and Rain

All the positive coefficients like year, workingday indicate that an increase in these values will lead to an increase in the value of Count and all the negative coefficients indicate that an increase in these values will lead to a decrease in the value of count.

Bike Sharing is higher in the month of September and also higher on a Saturday and working day. Rentals go down on a holiday and during winter months.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression algorithm explains the linear relationship between a dependent and one or more independent variables. Since it is linear it shows how the value of dependent variable changes based on the changes in independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables.

Mathematically linear regression is represented as:

y= a0+a1x+ ε

where y = dependent variable

X = Independent variable

A0 = interception of the line

A1 = Linear regression co-efficient

ε = random error

there are two types of linear regression:

1. Simple linear regression: There is one single independent variable used to predict the value of numerical dependent variable
2. Mulitple Linear Regression: More than one independent variable is used to predict the value of numerical dependent variable
3. Explain the Anscombe’s quartet in detail.

Ans: **Anscombe’s quartet** comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on a graph.

The four datasets that make up Anscombe’s quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe’s quartet is mainly used to illustrate exploratory data analysis and drawbacks of the same. It also emphasizes on the importance of data visualization

1. What is Pearson’s R?

Ans: The **Pearson correlation coefficient** is a correlation coefficient that measures 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. With the covariance the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations.

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

Ans: Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

|  |  |
| --- | --- |
| Min max Scaling | Standardized Scaling |
| First, we find the minimum and the maximum value of the column. |  First, we calculate the mean and standard deviation of the data we would like to normalize. |
| Then we will subtract the minimum value from the entry and divide the result by the difference between the maximum and the minimum value | Then we subtract the mean value from each entry and then divide the result by the standard deviation |

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

Ans: If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity

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

Ans: Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members’ age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.  
Since this is a visual tool for comparison, results can also be quite subjective nonetheless useful in the understanding underlying distribution of a variable(s)

Q-Q plot can also be used to test distribution amongst 2 different datasets. It can be particularly used where we split data into train-validation-test to see if the distribution is indeed the same. It is also used in the post-deployment scenarios to identify covariate shift/dataset shift/concept shift visually.

A Q-Q plot helps you compare the sample distribution of the variable at hand against any other possible distributions graphically.