**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?**

During the analysis of the categorical variables from the dataset, like season, weather situation, month, weekday and holiday the following are the insight were observed.

Season: This variable indicates the favourable weather conditions which encourages more bike rentals.

Weather Situation: some kind of weather situations discourage the people from bike rentals

Month: month is differentiated like in the start of the year the rentals will be high, like in the month of June to Sep this goes into peak.

Weekday: there will chances of increase in bike rentals on the weekdays

Holiday: Due to less commutators the demand will be low in Holidays.

Year: The demand got increased based on years with respect to the populations.

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

Using drop\_first-True during dummy variable creation which automatically drops one dummy column. Which removes redundancy, keeps the model interpretable, prevents multicollinearity.

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

Among all the numerical variables which has highest correlation with the target variable is Temperature.

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

The major assumption of Linear Regression after building the model on the training set. The residuals were normally distributed and independent.

The model of linear relationship between predictors and targets, it was free from multicollinearity.

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

The top 3 features contributing significantly towards explaining the demand of the shared bikes are Working Day, Year and Temperature.

**General Subjective Questions**

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

Here the Linear regression models is supervisor machine learning algorithm, which this model is used to identify the relationship between dependent variable and independent variable.

The basic algorithm is all about assuming of linear regression between the x variable and y variable. Here x is independent variable and y is dependent variable. The equation can be stated as x variable results in proportional to y variable.

This algorithm finds the coefficient value which predicts the closest values of the y variable.

The Linear Regression is a best fit line through scattered points such that the distance can be plotted between the actual point and the error.

1. **What is Pearson’s R?**

Pearson’s R is also called as Pearson Correlation Coefficient; this quantifies the relationship between the two continuous variables. The R denotes the values ranges from +1 to -1.

This measures only the linear regression, this only shows the correlation, which is very sensitive in outliers.

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

The process of transforming the numerical values which it fits into the specific ranges and have the similar statistical properties.

The scaling is performed for below reasons.

* To improve the model performance
* These features are scaled for optimising the algorithms
* Prevents computational errors and slow trainings

Difference between normalized scaling and standardized scaling.

The range between the normalized scaling will be -1 to +1, but the standardized scaling is centred at 0 and there is fixed range.

The normalized scaling is sensitive to outliers, where the standardised scaling is very less sensitive to outliers.

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

The Variance Inflation Factor are inflated due to multicollinearity among the independent variable.

The VIF becomes infinite when R values are squared and results in exactly 1.

The infinite VIF indicates perfect multicollinearity which causes regression coefficient. This also caused inflated standard errors.

The VIF becomes infinite when the independent variable is predicted by the other values which is perfectly multicollinearity.

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

The Q-Q plot is a tool which distinguishes of a actual data set and the graphical representation using normal distribution.

Q-Q stands for Quantile – Quantile, this quantile one distribution against the quantiles the other.

The Q-Q plot is used to check the assumption visually.

X-Axis (Theoretical)

Y-Axis (Sample)

Importance of Q-Q plot linear regression

Linear Regression assumes residuals are normally distributed.

If residuals deviates the transformation may be applied to achieved.