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?

→ Season, weathersit , holiday, mnth, workingday and weekday were the categorical variables in the dataset. I have used boxplot to visualise these categorical variables. These variables influenced our dependent variable (y variable/ target variable) that is "cnt" in the following ways:

1) Season : The boxplot shows that the median of 1 i.e., spring season is lowest or had the lowest value of cnt (y variable/ target variable), whereas the median of 3 i.e., fall season is highest or had the highest value of cnt (y variable/ target variable).

2) Weathersit : Whenever there is thunderstorm [3] cnt value (y variable/ target variable) is extremely low, indicating that the weather is not in favour; on the other hand when weather forecast is clear or few clouds [1] bike rental users are the highest making cnt value (y variable/ target variable) is highest when sky is clear.

3) Holiday : The boxplot is indicating that rentals are the lowest on 0 i.e., holidays, whereas highest on 1 i.e., non-holidays.

4) Mnth : 9th month that is September had the most cnt (y variable/ target variable) values, while 12th month that is December had fewest cnt value (y variable/ target variable). The same can be compared with weathersit, weather in 12th month is cold and snowy.

5) Weekday : While observing the boxplot I came to acknowledge that weekends had more cnt value (y variable/ target variable) than compared to weekdays

6) Workingday : While observing the boxplot, it seems like workingday had little effect on cnt value (y variable/ target variable).

7) yr : The boxplot shows that the median of 1 i.e., year 2019 has a higher cnt value (y variable/ target variable) is higher than of 0 i.e., year 2018.

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

→ Using drop_first = True during dummy variable creation is important as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variable.

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

→ While looking at the pair-plot among the numerical variables "temp" is the numerical variable which had the highest correlation coefficient of 0.627044.

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

→ The distribution of residuals should be normal and it's mean should be 0. When we test this residuals assumption by plotting displot from sns library to see if they follow a normal distribution or not. Through my observation, I came to conclusion that the residual are around mean = 0 which shows normal distribution.

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

→ Based on the final model, the top 3 features contributing significantly towards the demands of the shared bikes:

a) weathersit (3rd - light snow, light rain, thunderstorm) : as in weathersit 3 median is extremely low which tells us that it reduces the demand of bikes.

b) yr : we can see increase in yr variable in 1 (2019 year) which leads to increase in demand of bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

→ Linear regression is a supervised machine learning algorithm which predict a continuous entity / Y entity / Target variables. Linear regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model. Linear regression is based on the popular equation : **"y = mx+c"**. In regression, we calculate the best-fit line which describes the relationship between the dependent variable (y) and the independent variable/s (x). Regression is performed when the dependent variable is of continuous data type and independent variables could be of any data type such as continuous, categorical, etc. Regression method tries to find the best fit line by minimizing the difference between predicted and actual values using a cost function such as mean squared error. In regression, the dependent variable is the function of an independent variable and the coefficient and the error term.
Regression is divided into 2 types:

1. Simple Linear Regression : It is used when the dependent variable is predicted using only one independent variable. **"y = mx+ c"**

2. Multiple Linear Regression : It is used when the dependent variable is predicted using multiple independent variable. **"y = m1x1 + m2x2 + m3x3 +…. + mnxn + c"**.

2. Explain the Anscombe's quartet in detail.

→ Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, yet differ significantly when visualized. The quartet was introduced by the statistician Francis Anscombe in 1973 to emphasize the importance of data visualization in statistical analysis. Each dataset in the quartet consists of 11 paired x and y values. When examining the summary statistics of these datasets, they appear quite similar, with nearly identical means, variances, correlations, and regression lines. However, the plots of the

datasets tell a different story. Dataset I exhibits a linear relationship, fitting a simple linear regression model accurately. Dataset II forms a non-linear pattern and requires a polynomial regression model. Dataset III appears to have an outlier, greatly influencing the linear regression line. Finally, Dataset IV demonstrates a strong relationship between the variables but is undermined by an outlier that drastically alters the regression line. Anscombe's quartet highlights the importance of visualizing data to gain a deeper understanding of relationships and potential outliers. It serves as a reminder that summary statistics alone may not capture the true nature of the data, emphasizing the need for exploratory data analysis.

3. What is Pearson's R?

→ Pearson's R also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It was developed by Karl Pearson in the late 19th century and is widely used in various fields, including statistics, social sciences, and data analysis. The value of Pearson's R ranges between -1 and 1. A value of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable increases proportionally. Conversely, a value of -1 indicates a perfect negative linear relationship, where one variable decreases as the other increases. A value of 0 suggests no linear relationship between the variables. The Pearson's R coefficient is computed by dividing the covariance of the two variables by the product of their standard deviations. It measures the strength of the linear association and provides insights into how closely the data points align around the best-fit line. Pearson's R is a valuable tool for analyzing and understanding relationships between variables, assisting in predictive modeling, and identifying dependencies in data analysis.

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

→ Scaling refers to the transformation of data to a specific range or distribution. It is performed to ensure that all variables are on a similar scale and have a comparable impact on the analysis or model training process.

Scaling is necessary for several reasons:

a. Preventing variable dominance: Variables with larger magnitudes or wider ranges can dominate the learning algorithm, leading to biased results. Scaling helps avoid this issue by bringing all variables to a similar scale.

b. Enhancing convergence: Many machine learning algorithms, such as gradient descent-based optimization methods, converge faster when the variables are on a similar scale. Scaling facilitates efficient convergence and improves algorithm performance.

c. Supporting distance-based calculations: Scaling is crucial when using distance-based algorithms, such as clustering or k-nearest neighbours. These algorithms heavily rely on the distance between data points, which can be influenced by variables with different scales.

There are two commonly used scaling techniques: normalized scaling and standardized scaling.

Normalized Scaling: Normalized scaling, also known as min-max scaling, transforms the data to a specific range, typically between 0 and 1. It preserves the relative relationships and proportions of the data. The formula for normalized scaling is:

X_normalized = (X - X_min) / (X_max - X_min)

Standardized Scaling: Standardized scaling, also called z-score scaling, transforms the data to have a mean of 0 and a standard deviation of 1. It centers the data around the mean and adjusts the scale based on the standard deviation. The formula for standardized scaling is:

X_standardized = (X - X_mean) / X_std

The main difference between normalized and standardized scaling lies in the resulting range and distribution of the scaled data. Normalized scaling transforms the data to a specific range, whereas standardized scaling centers the data around the mean and adjusts the scale based on the standard deviation. Normalized scaling is useful when preserving relative relationships is important, while standardized scaling is suitable when comparing variables with different means and standard deviations or when assumptions of normality are desired.

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

→ **VIF – The Variance Inflation Factor:** The occurrence of an infinite value of VIF (Variance Inflation Factor) typically happens when there is perfect multicollinearity among the predictor variables. Perfect multicollinearity means that one or more variables in a regression model can be perfectly predicted from a linear combination of other variables. In this case, the correlation matrix used to calculate VIF becomes singular, resulting in a division by zero and producing an infinite VIF value. Perfect multicollinearity can arise when variables are linearly dependent or when there is redundancy in the set of predictors. To address this issue, it is necessary to identify and handle the multicollinearity problem by removing or transforming variables or applying other techniques such as ridge regression.

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

→ A Q-Q (quantile-quantile) plot is a graphical tool used to assess the distributional similarity between a given dataset and a theoretical distribution, often the normal distribution. It compares the quantiles of the dataset against the corresponding quantiles of the theoretical distribution. In linear regression, Q-Q plots are useful for evaluating the assumption of normality of residuals. Residuals are the differences between the observed values and the predicted values from the linear regression model. The Q-Q plot of the residuals helps determine if they follow a normal distribution or deviate from it. If the residuals follow a normal distribution, the points on the Q-Q plot will fall approximately along a straight line. Deviations from a straight line indicate departures from normality.

For instance - heavy tails, skewness, or outliers can be identified by examining the shape of the Q-Q plot.

The importance of Q-Q plots in linear regression lies in assessing the validity of the assumption of normally distributed residuals. If the residuals violate this assumption, it can affect the validity of statistical inferences, such as hypothesis testing or confidence intervals, based on the regression model. In such cases, transformations or alternative modeling techniques may be necessary to account for the non-normality of residuals and improve the accuracy of the model.