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?**

A → There are 6 categorial variables in dataset, those are Season, Month, Holiday, Weekday, working day, Weathersit, Boxplots are used to visualize these categorial variable as these are independent variable.

SEASON: -From boxplot we can say Spring season has lowest rental & fall has highest rental whereas summer and winter has rental values in mid.

MONTH: - September month has highest rentals where January has lowest rental among all

HOLIDAY: - Low rentals in holidays

WEEKDAY: -high rentals in Friday

WORKINGDAY: - no significant difference

WEATHERSIT: - highest rental in good weather condition and lowest in bad condition

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

A →To avoid multicollinearity between dummy variables, even if we keep first dummy it is a redundant column.

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

A → 'temp' and 'atemp' are highly correlated with cnt.

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

A → From error distribution we saw that the distribution of residuals are normally distributed and centered around mean (0)

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

A → from final model the top 3 features that contributing significantly are Temperature (0.556023) , Humidity (-0.305790) and year (0.229612).

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A→ Linear regression is a type of supervised Machine learning algorithm used to predict numeric values. Regression most commonly used predictive analysis model

Equation of Linear regression is 'y=mx+c'

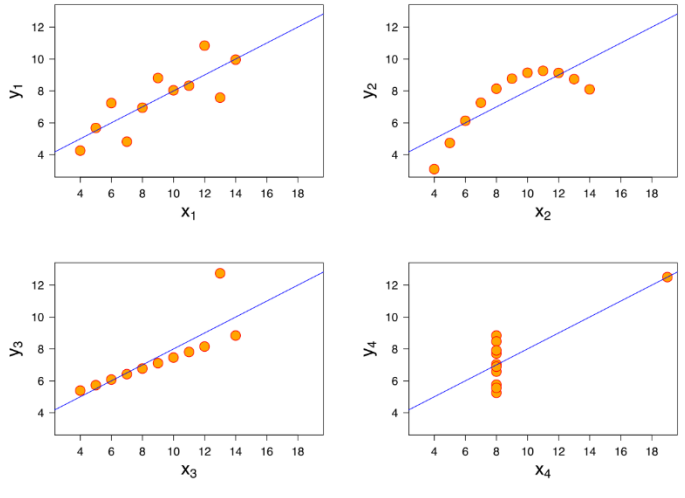
there is a linear relation between the dependent variable(y) and the predictors or independent variable(x). In regression, we calculate the best fit line which describes the relation between the dependent and independent variable.

Regression is classified into simple linear regression and multiple linear regression.

- Simple linear Regression: SLR is used when the dependent variable is predicted using only one independent variable.
- Multiple linear regression: MLR is used when the dependent variable is predicted using multiple independent variables.

2. Explain the Anscombe's quartet in detail.

A→Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.



- In dataset I it appears to have clean and well-fitting linear models.
- In dataset II is not distributed normally.
- In dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- In dataset IV shows that one outlier is enough to produce a high correlation coefficient.

3. What is Pearson's R?

A→It is a numerical representation of linear relationship between the variables it values ranges from -1 to +1. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

r = 1 (the data is perfectly linear with a positive slope)
r = -1 (the data is perfectly linear with a negative slope)
r = 0 (there is no linear correlation)

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

A→Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. 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, irrespective of the units of the values.

- Normalization is generally used when you know that the distribution of your data does not follow Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-nearest neighbors and neural networks.

- Standardization, on the other hand can be helpful in cases where the data follows a Gaussian distribution.

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

A→ If there is perfect correlation, then VIF is infinity. A large value of VIF indicates that there is a correlation between the variables. When the value of VIF is infinite it shows a perfect correlation between two independent variables. To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

A→The purpose of the quantile-quantile (QQ) plot is to show if two data sets come from the same distribution. Plotting the first data set's quantiles along the x-axis and plotting the second data set's quantiles along the y-axis is how the plot is constructed. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.