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)**

The categorical variable in the dataset were season , yr , holiday, weekday ,workingday, and weathersit and mnth .

Season - Spring season had least value of cnt whereas fall had maximum value of cnt.

Weathersit - Highest count was seen when the weathersit was' Clear, Partly Cloudy'.

Yr - The number of rentals in 2019 is more than 2018

Mnth - September saw highest no of rentals while December saw least.

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

drop_first=True is important to use, because it helps in reducing the extra column created during dummy variable creation. So, it reduces the correlations created among dummy variables. Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

"temp" and "atemp" are the two numerical variables which are highly correlated with the target variable

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

The following tests were done to validate the assumptions:

1. We visualised the numeric variables using a pairplot to see if the variables are linearly related or not.
2. We validated assumption about residuals by plotting a distplot of residuals and saw if residuals are following normal distribution or not.
3. Linear regression assumes that there is little or no multicollinearity in the data. We calculated the VIF to get the idea about how much the feature variables are correlated with each other and dropped the features with high VIF

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

   Temp, humidity and year have the highest contribution

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric continuous values. It is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model.

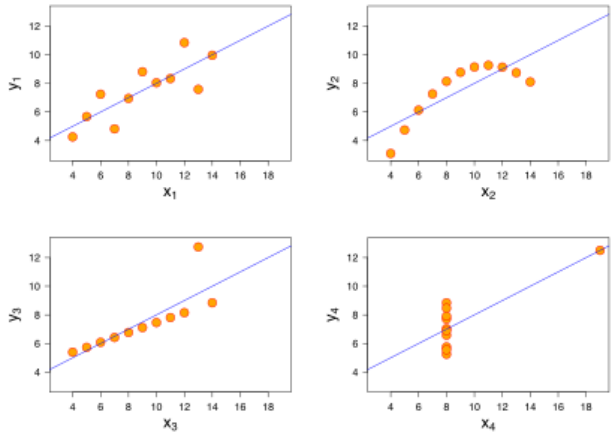
Linear regression is based on the popular equation “y = mx + c”. It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x).

In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Anscombe's quartet comprises four data sets having identical simple statistics and yet have very different distributions and appear very different when graphed. Each dataset consists of eleven points. They were constructed in 1973 to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."



3. What is Pearson's R? (3 marks)

Pearson's r is a numerical summary of the strength of the linear association between the variables.It value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us " can we draw a line graph to represent the data?

r = 1 means the data is perfectly linear with a positive slope

r = -1 means the data is perfectly linear with a negative slope

r = 0 means there is no linear association

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. 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 used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks. In this, generally the data is scaled between 0 to 1

● **Standardization**, is helpful generally where the data follows a Gaussian distribution. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization. The data will be standardized to a mean of 0 and standard deviation of 1

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

VIF - Variance Inflation Factor, VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = infinity. It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model.

$$VIF = \frac{1}{1 - R^2}$$

If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity"

A rule of thumb for interpreting the variance inflation factor:

• 1 = not correlated. • Between 1 and 5 = moderately correlated. • Greater than 5 = highly correlated

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

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. 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.

Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis