By Nandish Jani

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?

- Fall and Misty weather have a positive effect on the dependent variable. Also the months from March to October have a positive effect on the dependent variable.

- Months Jan, Feb and Light rain/snow weather lead to negative effect on the dependent variable.

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

- drop_first=True is important to use, because it helps reduce the extra column created during dummy variable creation.

- This helps in reducing the correlations created among dummy variables.

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

- Temp variable has the highest correlation with the target variable. As for registered variable, the variance keeps on increasing.

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

- P-value of variables was less than 0.05 and their VIF below 5.

- The Prob (F-statistic) was almost 0 while F-statistic was significant. Both these are good indicators for the model.

- Residuals/error terms were calculated and graphed. They showed normal distribution.

- High r-squared and the difference between r-squared and adjusted r-squared also happened to be very negligible and therefore good at just 0.05.

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

- Temp
- Holiday
- Weather

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear regression algorithm is a machine learning algorithm based on supervised learning. Supervised learning is the type of machine learning in which machines are trained using labelled training data, and on the basis of that data of input-output pairs, machines predict the output.

- Linear regression algorithm is a statistical model that attempts to show the linear relationship between ‘target’ and ‘feature’ variables with a linear equation.

- Linear regression algorithm-Target: This is the dependent variable of the dataset whose predictions are found using the linear regression.

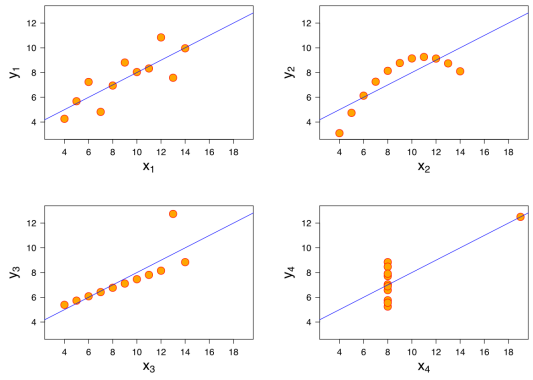
- Linear regression algorithm-Feature: These are single or set of independent variables used to build linear regression models.

2. Explain the Anscombe's quartet in detail.

- Anscombe's quartet is a set of 4 datasets with nearly identical basic statistical properties. It was developed by the statistician Francis Anscombe.

- While the datasets contained nearly identical basic statistical properties, they appeared very different when graphed.

- This was used to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.



Source: Wikipedia

3. What is Pearson's R?

Pearson's correlation (also called Pearson's R, bivariate correlation), is one of the several types of correlation coefficients.

It is one of the more popular types of correlation coefficients. It is commonly used in linear regression as it is a measure of linear correlation between two sets of data.
The values of the coefficient will always be between -1.0 to +1.0, where -1.0 to 0 suggest negative correlation and 0 to +1.0 suggest positive correlation. Values nearer to -1 and +1 suggest stronger linear relationships.

However this doesn't necessarily reflect causation as it only looks at the linear relation between the two sets of continuous variables and ignores other possible correlations between them.

Formula for Pearson's r is as follows:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where,
r = Pearson correlation coefficient
x = Values in the first set of data
y = Values in the second set of data
n = Total number of values.

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

- Scaling is a technique of standardizing data in variables before building models for machine learning.

- This is necessary because otherwise the data with higher values will be considered as having more weightage or importance than variables with smaller values/numbers, thus leading to improper model building.

- Normalized scaling, also known as MinMax scaling, is the scaling technique where the values are scaled to the range between 0 and 1.

- Unlike normalized scaling, standardized scaling doesn't scale in a bound range. Instead, here the values are centered around the mean with a unit standard deviation.

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

- Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables.

- Generally, VIF between 1 and 5 is considered low multicollinearity while VIF above 5 is considered as high.

- VIF of 1 indicates there is no multicollinearity with other variables, on the other hand, VIF infinity indicates a perfect multicollinearity with other variables.

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 probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

- The Q-Q plot is used to easily see in graphical form, the normal distribution of the error terms from the linear regression model.

- It also helps to check for skewness in the data, like negative skew or positive.Can also be a measure of tailedness.