Assignment-based Subjective Questions

Q. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A. Fall in seasons had the highest demand of riders, 2019 saw increase in demand as compared to 2018. The months from June until September saw a higher demand in riders.

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

A. The drop_first allows us to reduce the number of predictors and can also help in reducing multicollinearity. N variables can be identified using n-1 dummy variables and hence the first value can be dropped.

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

A. temp and atemp have the highest correlation with the target variable.

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

A. We can create a histogram or distplot in the residual analysis. The distribution should be a normal distribution and sum of residuals should be centered around 0.

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

A. temp, yr and season_4 are the top 3 features for the model based on their coefficients

General Subjective Questions

Q. Explain the linear regression algorithm in detail.

A. Linear Regression is a supervised learning algorithm. It's a regression model that predicts target value based on independent variable. It is mandatory that atleast 1 or more independent variables have a linear relationship with the target variable in order to process the linear regression algorithm.

It is represented by:

$y = c + mx$ where c is the intercept and m is coefficient of the independent variable

Q. Explain the Anscombe's quartet in detail.

A. Anscombe's quartet comprises four datasets that have nearly identical simple statistical, properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.

Q. What is Pearson's R?

A. Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

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

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

A. Scaling is the process of reducing all data set within the same minimum and maximum levels. It is performed to reduce the anomalies between the types and variations between the data sets. Normalised scaling keeps all the factors between 0 and 1 whereas standardized scaling only standardizes and there can be outliers even after scaling.

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

A. Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.