Bike Sharing Demand Prediction Assignment

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: From our analysis of the categorical variables from the dataset we can predict the formula for the best fit line equation:

cnt = 0.091994 + (yr * 0.237944) + (holiday * (-0.069918)) + (temp * 0.576176) + (windspeed * (-0.124141)) + (season_2 * 0.069603) + (season_4 * 0.115836) + (weathersit_3 * (-0.246621))

- **YR** – The coefficient of yr indicates that a unit increase in **yr** variable, will increase bike hirings by 0.237 values.
- **HOLIDAY** - Coefficient of holiday indicates that a unit increase in holiday variable, will decrease the bike hiring by 0.0699 values.
- **TEMP** - Coefficient of temp indicates that a unit increase in temp variable, will increase the bike hiring by 0.5761 values.
- **WINDSPEED** – The coefficient of the windspeed indicates that a unit increase in windspeed data, will decrease the bike hiring by 0.1241 values.
- **SEASON_2** - The coefficient of season_3 indicates that a unit increase in the season_2, will increase the bike hiring by 0.0696 values.
- **SEASON_4** - The coefficient of season_4 indicates a unit increase in the season_4, will increase the bike hiring by 0.1158 values.
- **WEATHERSIT_3** - Coefficient of weathersit_3 indicates that a unit increase in the weathersit_3, will decrease the bike hiring by 0.2466 values.

Q2. Why is it important to use **drop_first=True** during dummy variable creation?

Ans: It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

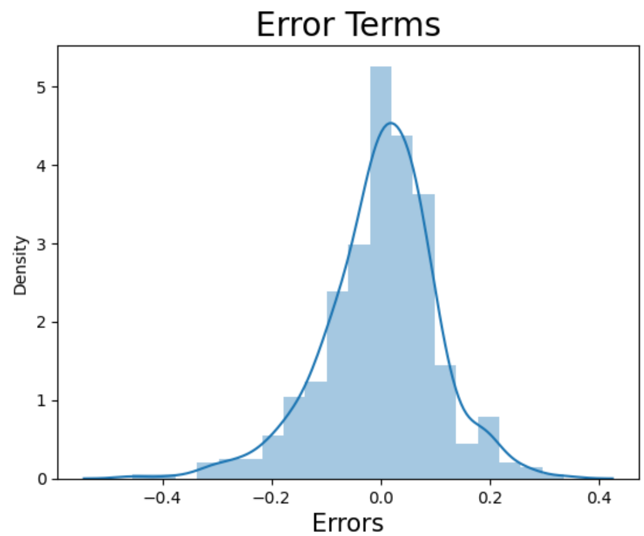
Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: By looking at the pair-plot, **‘TEMP’** has the highest correlation among the other numerical variables with the **‘CNT’** as the target variable.

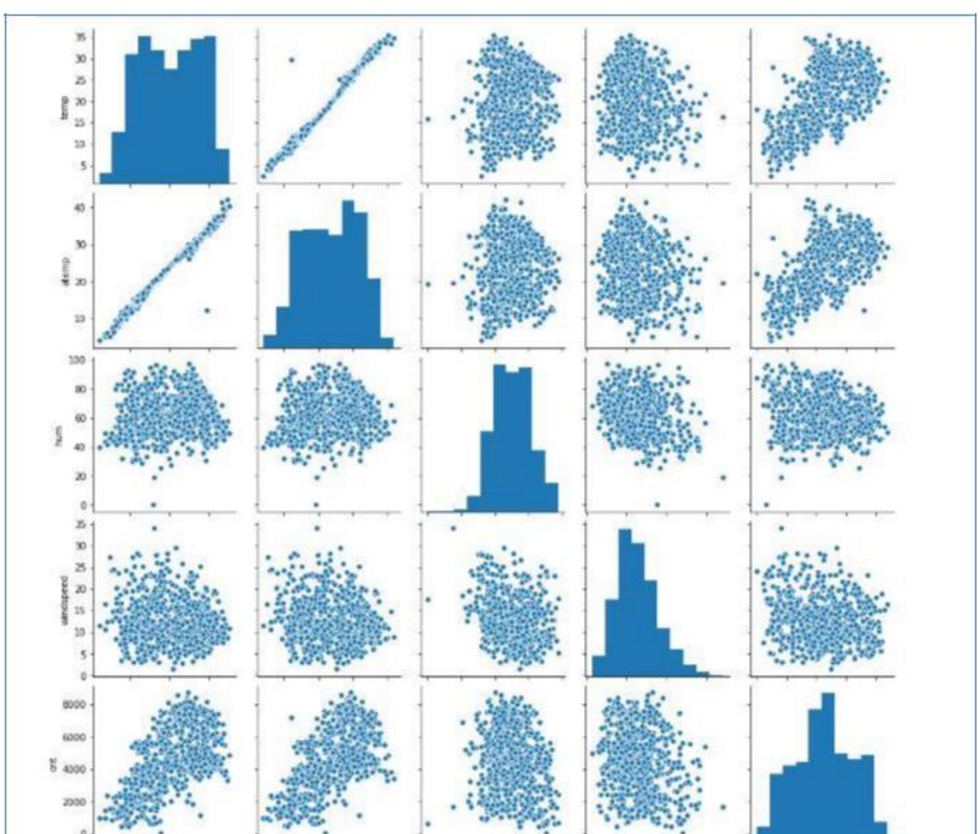
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: After building the model we can validate the assumptions of Linear Regression,

- Using histogram as we see the Residuals are normally distributed and maximum of the Error Terms are revolving around Zero.



- Using the pair plot, we could see there is a linear relation between temp and atemp variable with the predictor 'cnt'.



Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

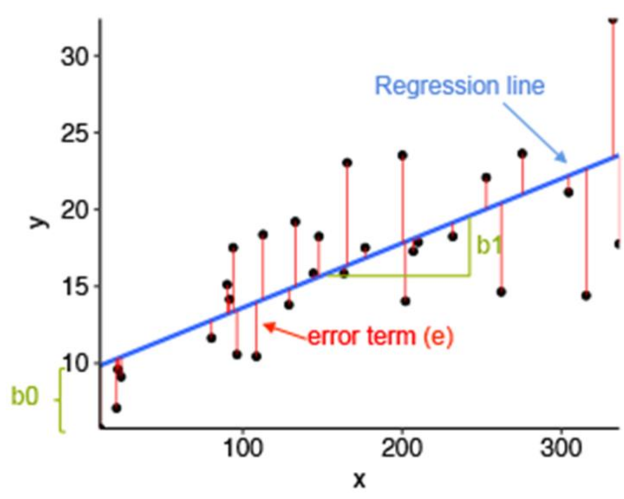
Ans: As per the Final Model, the top 3 Predictor Variables that are needed for the prediction purposes are:

- **YR** – The coefficient of yr indicates that a unit increase in **yr** variable, will increase bike hirings by 0.237 values.
- **HOLIDAY** - Coefficient of holiday indicates that a unit increase in holiday variable, will decrease the bike hiring by 0.0699 values.
- **TEMP** - Coefficient of temp indicates that a unit increase in temp variable, will increase the bike hiring by 0.5761 values.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans: Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.



Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable.

The line in the above graph is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. The line can be modelled based on the linear equation shown below:

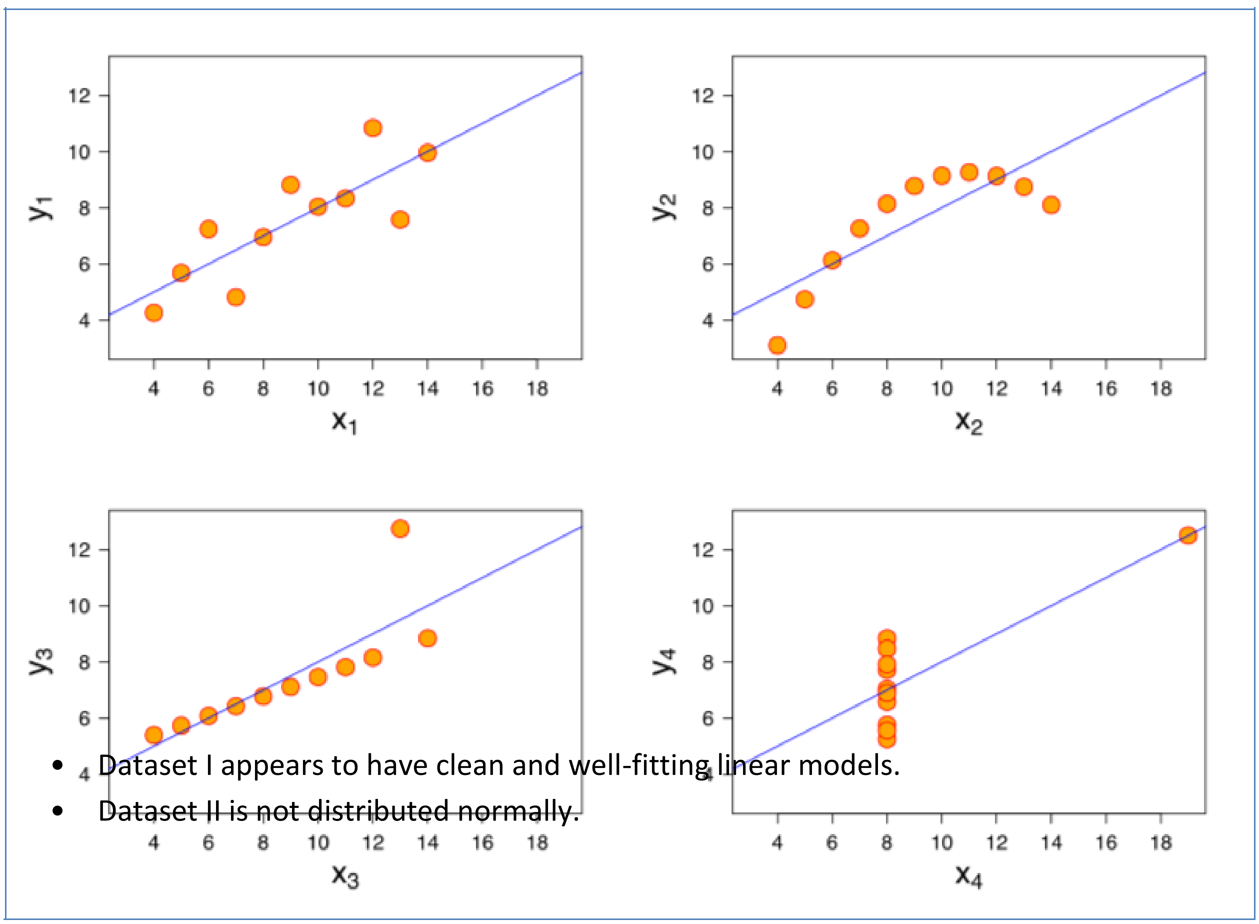
Y = m*x + c

Where, **m** = Slope of the line and **c** = Intercept.

Q2. Explain the Anscombe’s quartet in detail.

Ans: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

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.



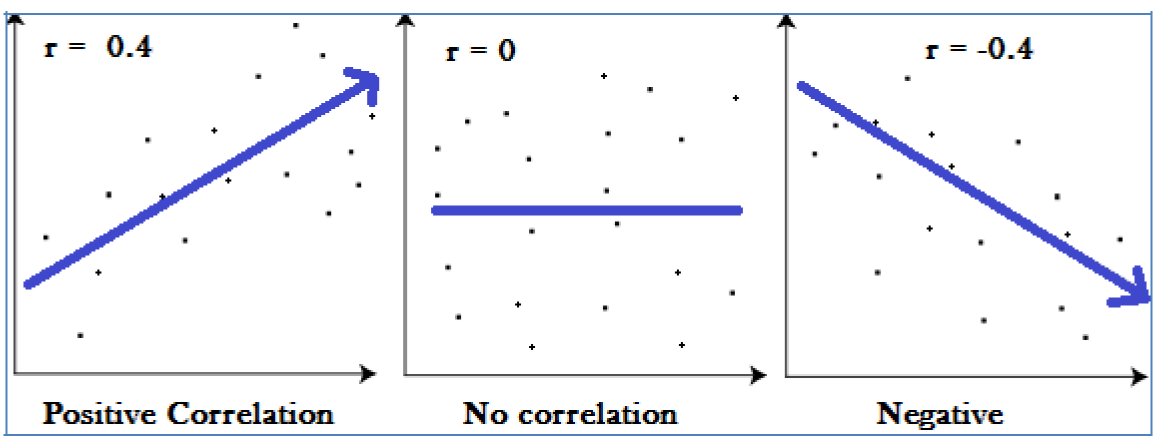
- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.

- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis.

Q3. What is Pearson's R?

Ans: Correlation coefficients are used to measure how strong a relationship is
between two variables. There are several types of correlation coefficient, but the
most popular is Pearson's. Pearson's correlation (also called Pearson's *R*) is a
correlation
coefficient commonly used in linear regression. If you're starting out in statistics, you'll
probably
learn about Pearson's *R* first. In fact, when anyone refers to the correlation coefficient, they are usually talking about Pearson's.



- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: **Scaling:** Feature Scaling is a technique to standardize the independent features present in the data in a fixed range.

It is performed during the data pre-processing to handle highly varying magnitudes or values or units. 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, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Difference between Normalization and Standardization:

- **Min-Max Normalization:** This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- **Standardization:** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

X_1=C+ α_2 X_2+α_3 X_3+⋯

〖VIF〗 _1=1/(1-R_1^2)

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

$$X_2=C+\alpha_1 X_1+\alpha_3 X_3+\cdots$$

$$\text{VIF}_2=1/(1-R_2^2)$$

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: 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.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:

a) It can be used with sample sizes also.

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

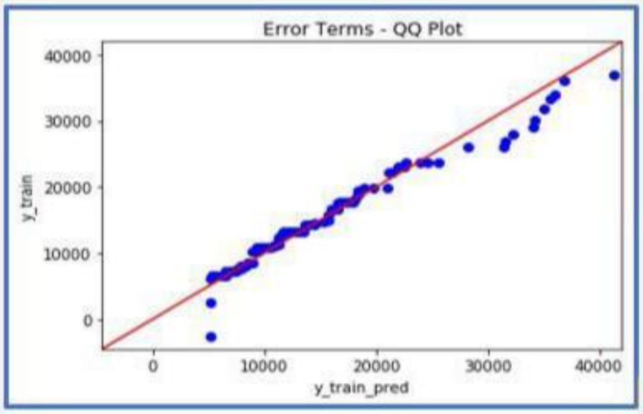
If two data sets —Come from populations with a common distribution.

- Have common location and scale.
- Have similar distributional shapes.
- Have similar tail behavior.

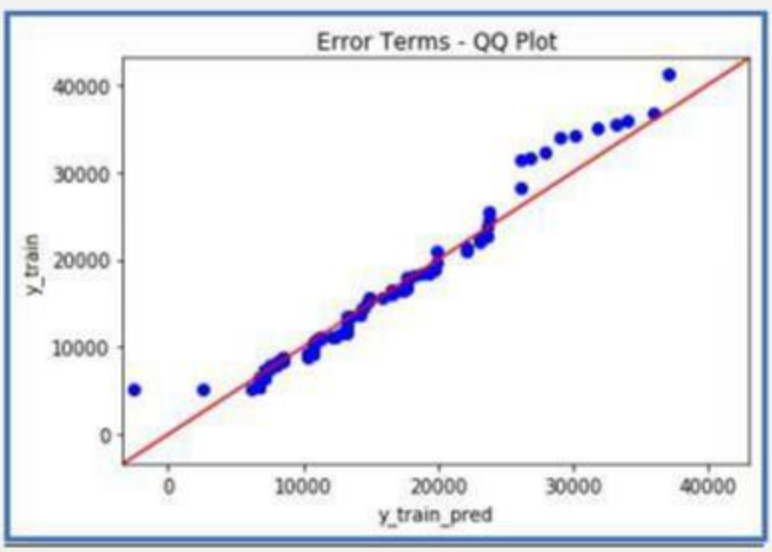
Below are the possible regressions for two data sets:

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

b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x –axis.