Bike Sharing Demand Prediction Assignment

Assignment based Subjective Questions:

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

Equation for the best fit line: Model Co-efficient values cnt = 0.2967 + 0.2468(yr) + (-0.1001 holiday) + (-0.0219 hum) + (-0.1750 windspeed) + 0.2565(season_2) + 0.2969(season_3) + 0.2223 (season_4) + 0.0771(mnth_9) + (-0.0856 weathersit_2) + (-0.2901 weathersit_3)
Interpretation of Equation:

yr = Bike hiring will increase by 0.22468 values.

holiday = Bike hiring will decrease by 0.1001 values.

hum = Bike hiring will decrease by 0.0219 values.

windspeed = Bike hiring will decrease by 0.1750 values.

season_2 = Bike hiring will increase by 0.2565 values.

season_3 = Bike hiring will increase by 0.2969 values.

season_4 = Bike hiring will increase by 0.2223 values.

mnth_9 = Bike hiring will increase by 0.0771 values.

weathersit_2 = Bike hiring will decrease by -0.0856 values.

weathersit_3 = Bike hiring will decrease by -0.2901 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.

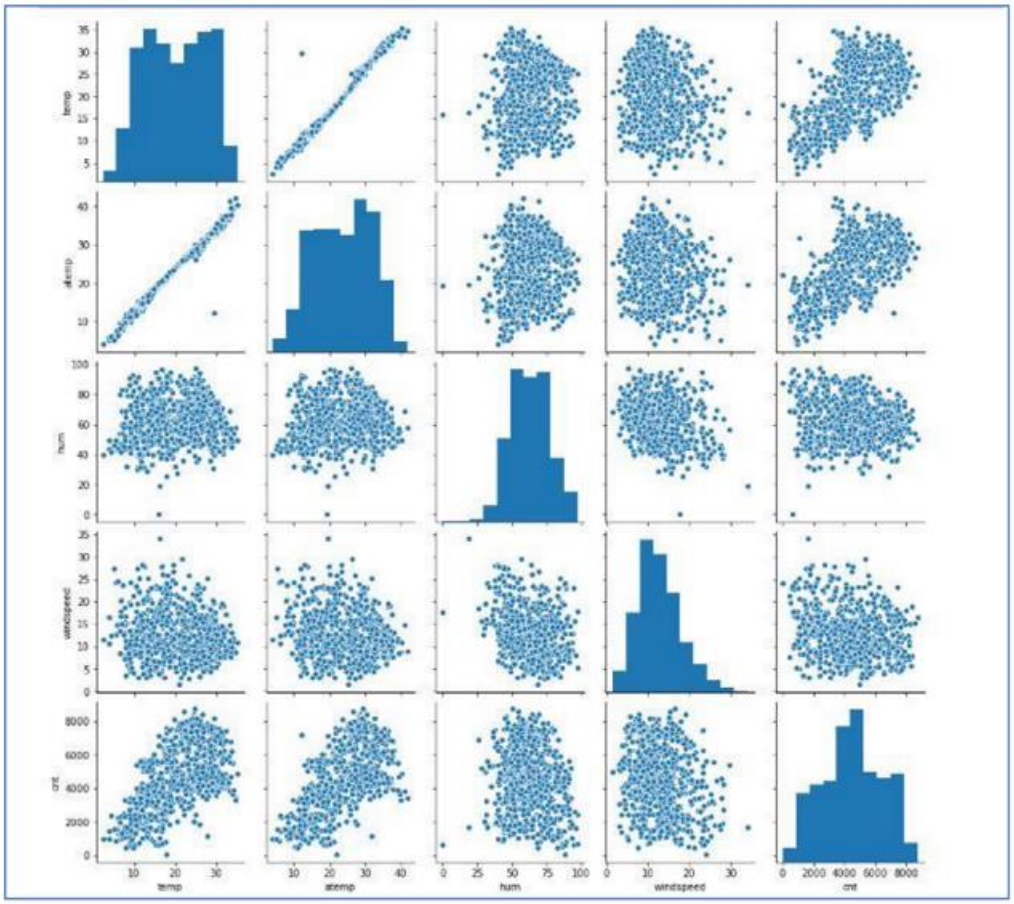
Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: ‘TEMP’ has the highest correlation among rhe other numerical variables with the ‘CNT’ as the target variable.

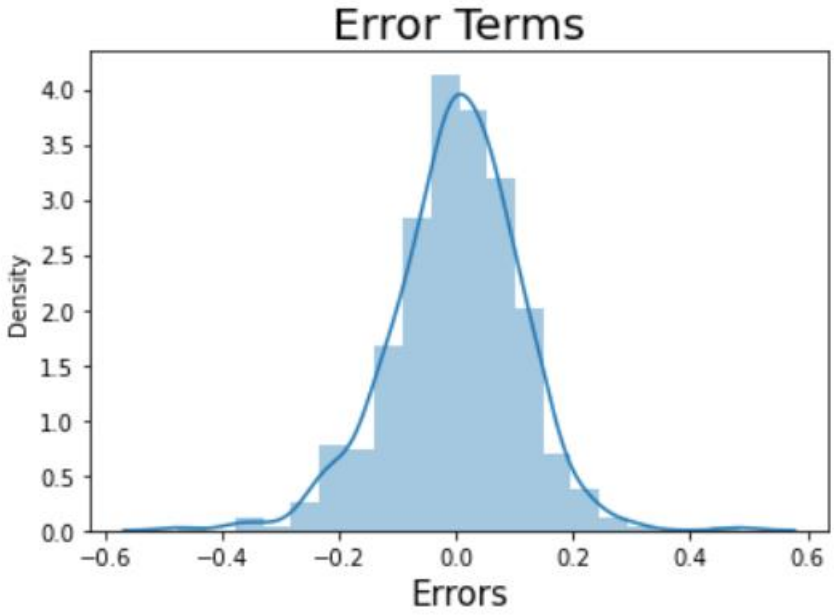
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: After building the model we can validate the assumptions of linear regression:

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



Using Histogram, the residuals are normally distributed and maximum of the error terms are revolving around zero.



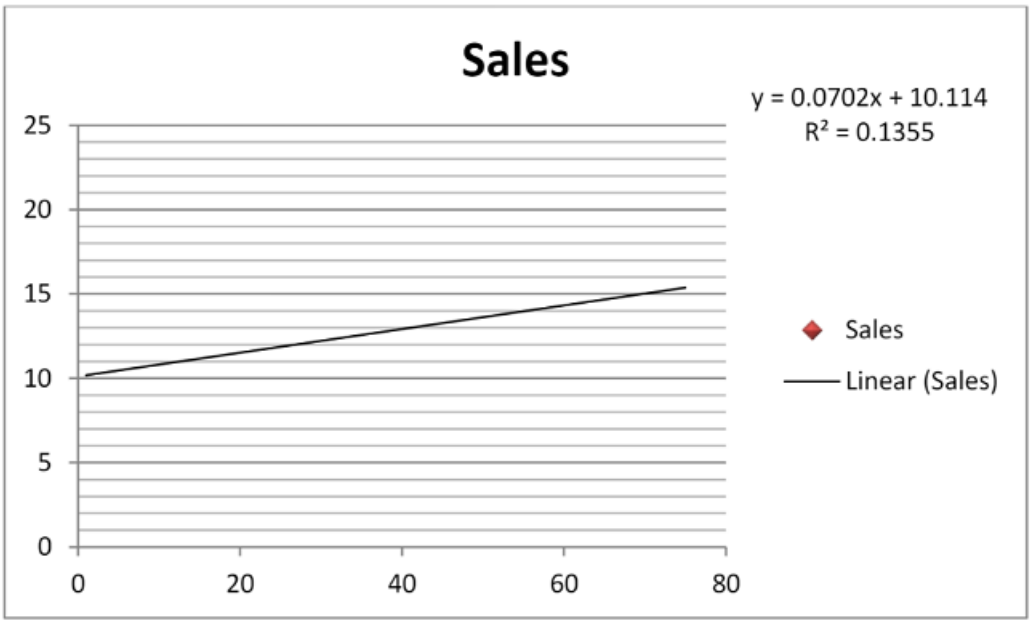
Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

- YR - Coefficient of yr indicates that a unit increase in yr variable, will increase bike hirings by 0.2468 values.
- HOLIDAY - Coefficient of holiday indicates that a unit increase in holiday variable, will decrease the bike hiring by -0.1001 values.
- season_3 = Bike hiring will increase by 0.2969 values.
- season_2 = Bike hiring will increase by 0.2565 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 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 and dependent variable.

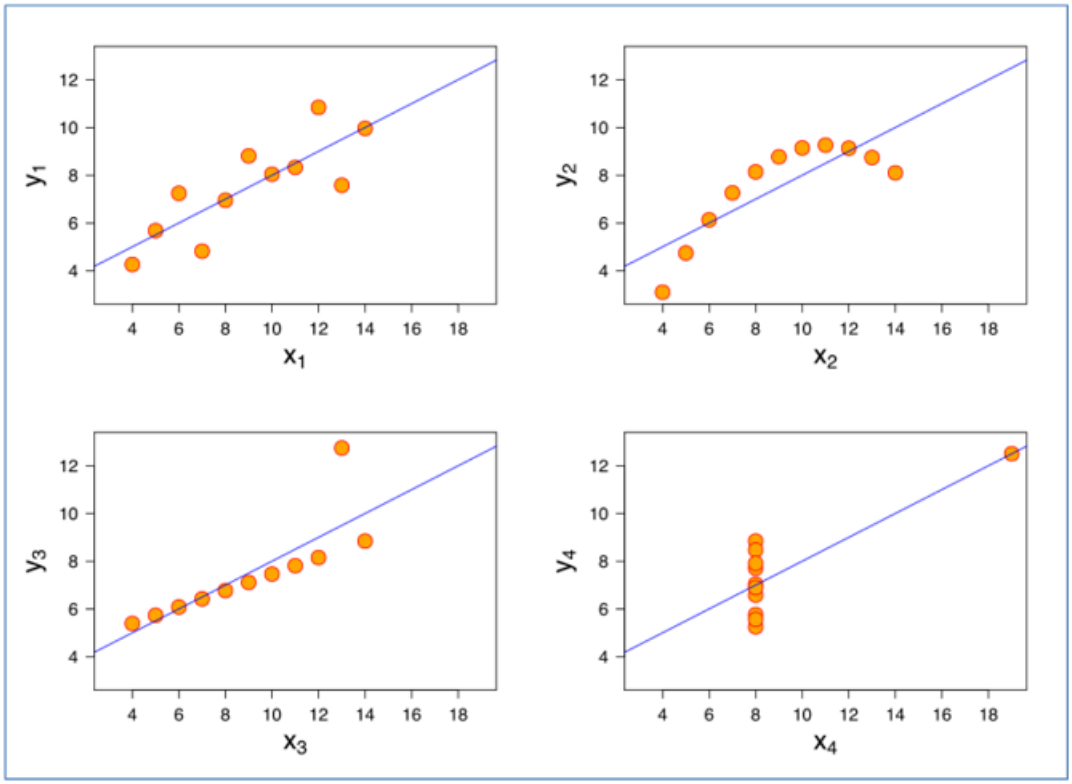
The line in the above graph is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. The line can be modelled based on the linear equation like Y = m*x + c

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 distribution 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.

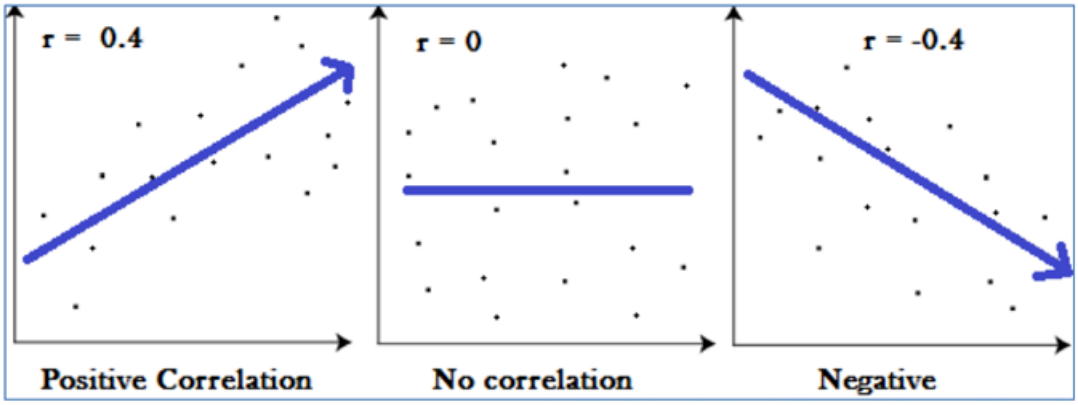


- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear.
- Dataset IV shows that one outlier produce a high correlation coefficient.

This is 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 popular is Pearson's. Pearson's correlation is correlation coefficient commonly used in linear regression. If you're starting out in statistics, you will probably learn about Pearson's R first.



- 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.
- 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.
- Zero means that for every increase, there isn't a positive or negative increases. 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 ML algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

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

$X = Xi - \min(X)/\max(x) - \min(x)$

Standardization: It is very effective technique which re scales feature value. It has distribution with 0 mean value and variance equals to 1.

X = Xi – X(mean) / 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:

$X_1 = C + \alpha_2 X_2 + \alpha_3 X_3 + \ldots$

$(VIF)_1 = 1/(1-R_1^2)$

Next, we fit the model between X2 and other independent variables.

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

$(VIF)_2 = 1/(1-R_2^2)$

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.

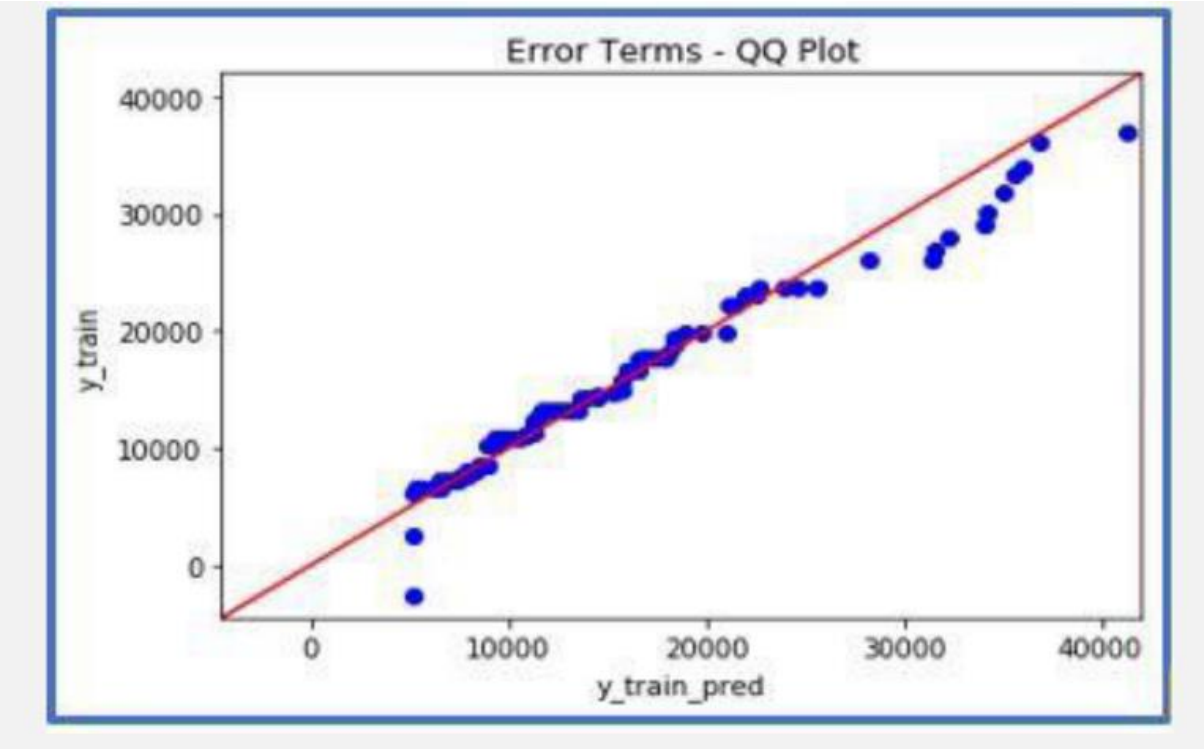
This helps in scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Advantages:

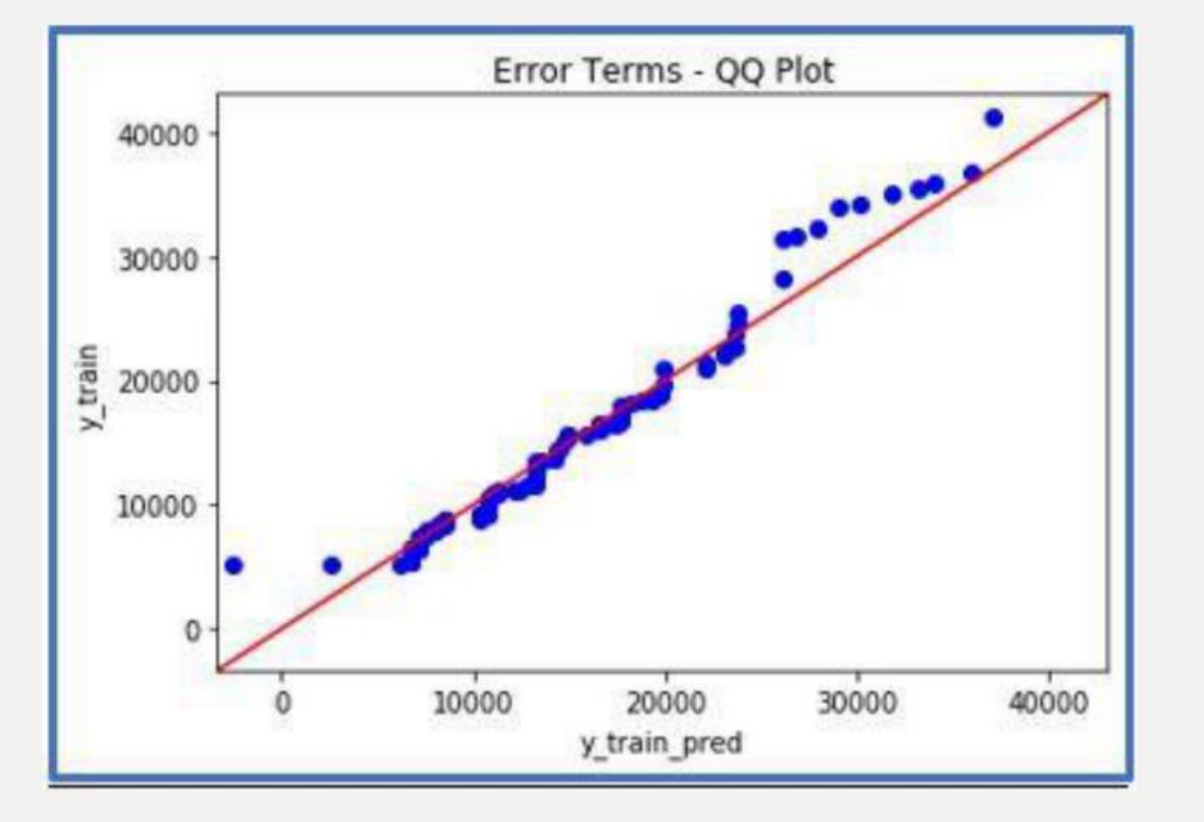
- It can be used with sample sizes also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
    - Come from population with a common distribution.
    - Have common location and scale.
    - Have similar distributional shapes.
    - Have similar tail behaviour.

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 al point of quantiles lies away from the straight line at an angle of 45 degree from x-axis.