Assignment Based Subjective Questions & Answers:

BIKE SHARING ASSIGNMENT

BY:

BHUMIKA JAIN

Question : 1

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

Ans:

- In Fall there is significant increase than other season.
- There is dip in spring season.
- The business is gaining popularity year by year and hence there is increase in 2019.
- There is not much impact of weekdays but it is performing little well on Fridays.
- More bike rentals took place during non-holiday time.
- It is getting more used in case of clear weather.
- If it is working day people are using it more might be for going office, college or school.

Question : 2

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

Ans:

- Using the drop_first=True during dummy variable creation helps in reducing the extra columns created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
- For Example: We have four variables: 'Spring’, 'Summer’ , 'Fall’ , 'Winter’. We can only take 3 variables as 'Spring’ will be 1-0-0, 'Summer’ as 0-1-0 and 'Winter’ as 0-0-1 , so we don’t need 'Fall’ as we know 0-0-0 will indicate 'Fall’. So we can remove it.

Question : 3

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

Ans:

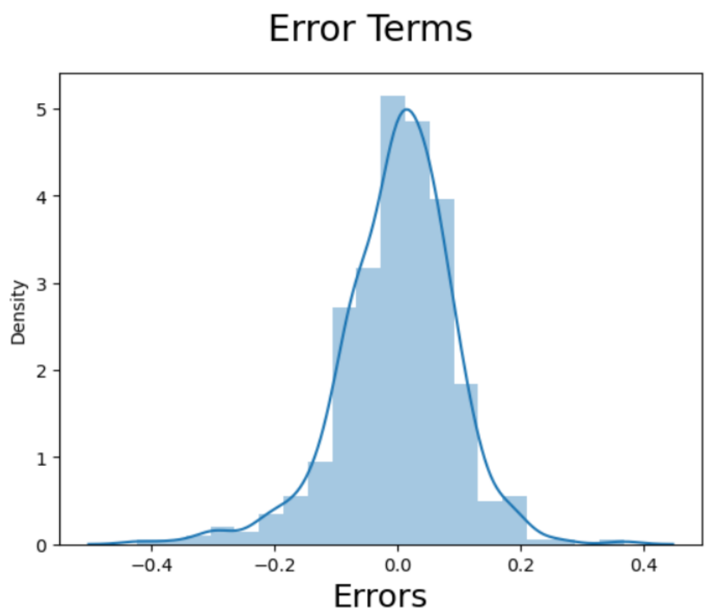
- ‘temp and ‘atemp’ both are highly correlated with target variable of 0.63, which is highest among all the other variables.

Question : 4

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

Ans:

- The simple way to determine if this assumption is met or not is by creating a scatter plot x vs y. If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds.
- Fitted regression line is linear .
- Error terms came out normally distributed with mean as 0.



Question : 5

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

Ans:

The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

- ➢ Weather sit (Light snow) – Negatively correlated with target variable (-0.24)
- ➢ Year – Positively correlated with target variable (0.57)
- ➢ Temp - Positively correlated with target variable (0.63)

General Subjective Questions & Answers:

Question : 1

Explain the linear regression algorithm in detail.

Ans:

Linear Regression is a type of supervised machine learning algorithm that is used for the prediction of numeric values. Linear Regression is 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”.

Regression is divided into two models:

- Simple Linear Regression: The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points. The standard equation of the regression line is given by the following expression: $Y = \beta_0 + \beta_1 X$.

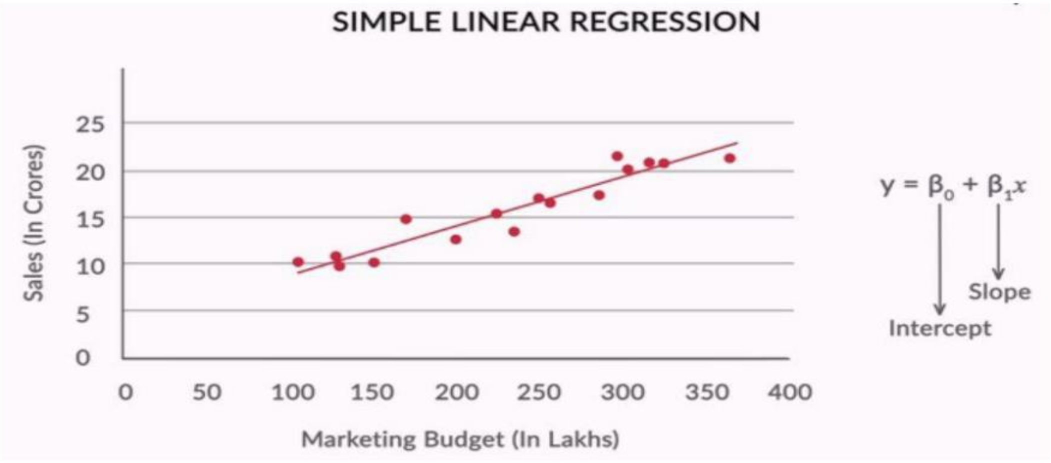
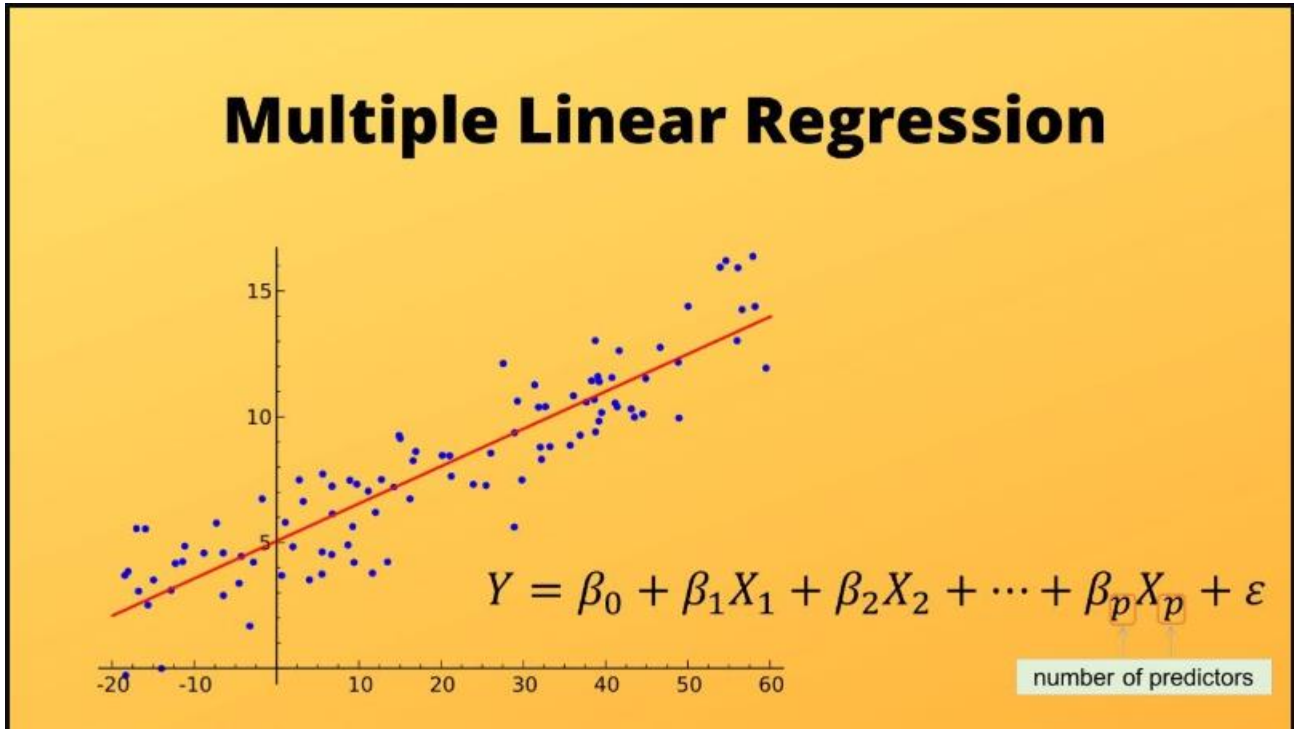


Figure 3 - Regression Line

- Multiple Linear Regression: Multiple Linear Regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. The formula now can be simply given as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$



Question : 2

Explain the Anscombe's quartet in detail.

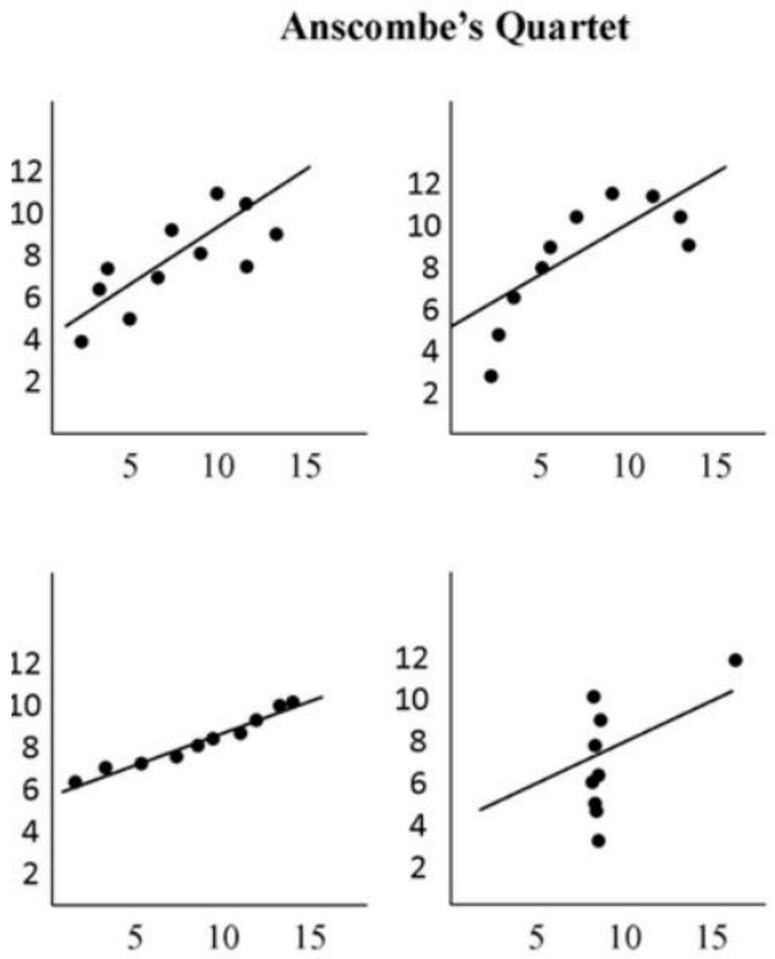
Ans:

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of mean, variance, R-squared, correlation and linear regression lines but having different representations when we scatter plot on graphs.

The datasets were created by statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

Statistical Properties:

- Find mean for X and y for all four datasets.
- Find correlation with their corresponding pair of each datasets.
- Find standard deviations for X and y for all four datasets.
- Find slope and intercept for each datasets.
- Find R-square for each dataset.
- Create a statistical summary by using all these data and print it.



Question : 3

What is Pearson’s R?

Ans:

Pearson coefficient is a type of correlation coefficient that represents the relationship between two variables that are measured on the same interval. Its value ranges from -1 to +1. In layman’s terms, it asks if we can 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.

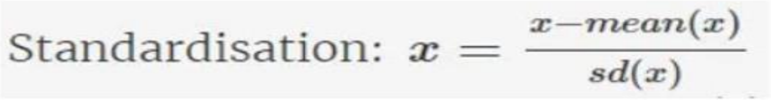
Question : 4

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

Ans:

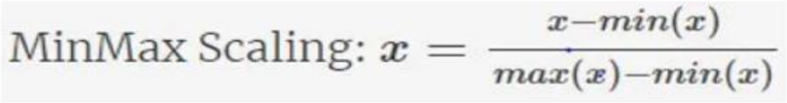
It is a step of data pre-processing which is applied to independent variable to normalize the data with in a particular range. It also helps in accelerate the calculation in an algorithm. Scaling facilitates meaningful comparisons between features, improves model convergence, and prevents certain features from overshadowing others based solely on their magnitude.

Standard Scaling: : The variables are scaled in such a way that their mean is zero and standard deviation is one.



$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data. This is also known as normalized scaling.



$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Question : 5

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

Ans:

The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where, ‘i’ refers to the ith variable.

If R-Squared value is equal to 1 than the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

Question : 6

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

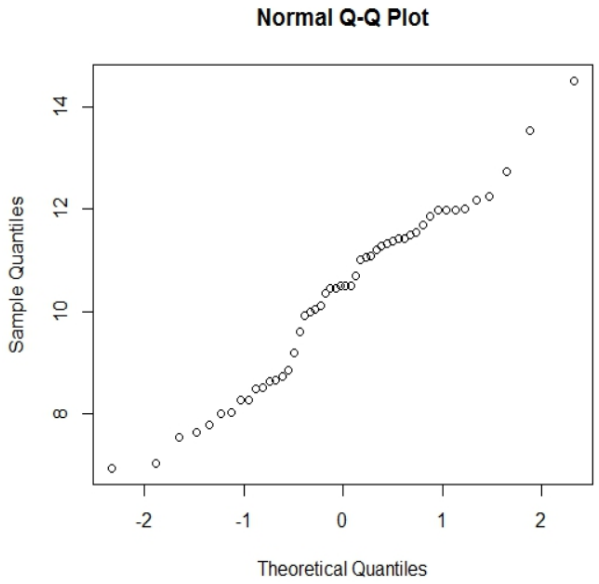
Ans:

A Q-Q plot is a scatter plot 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.

Use of Q-Q plot in linear regression: It is used to see if the points lie approximately on the line. If they don't , it means our residuals are not normal and thus, our error are also not normal.

Importance of Q-Q plot in linear regression:

- The sample size do not needs to be equal.
- Many distributional aspects can be simultaneously tested. For example, shifts in scale, changes in symmetry, shifts in locations and the presence of the outliers.
- The Q-Q plot can provide more insights into the nature of the difference than analytical methods.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles