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? (3 marks)

Below are the categorical variables identified and their impact

Season: there is no demand in winter as well as goes down when weather little rainy and light snow.

Yr: the demand is growing in subsequent year, also the month plot suggests the same.

Mnth: demand is growing, comes down only when weather is not good.

Weekday: surprisingly mid of the week, demand goes high it starts slow with start of the week.

Workingday: on a known working day demand is higher than the working days.

Holiday: holidays demand is higher, but surprisingly on non-holiday days mean is higher.

Weathersit: demand goes higher when weather is good.

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

Ans:

This is needed to reduce correlations created among dummy variables. other unwanted extra variables will be created while creating dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

if we talk about among all predictive variables then temp and atemp.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

GENERAL SUBJECTIVE QUESTIONS

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is used to predict the target variable values, as name suggest it follows the line-based prediction approach. When we predict target variable based on single independent variable it is called simple regression as well and the model is defined as below

y = mX + c

Where:

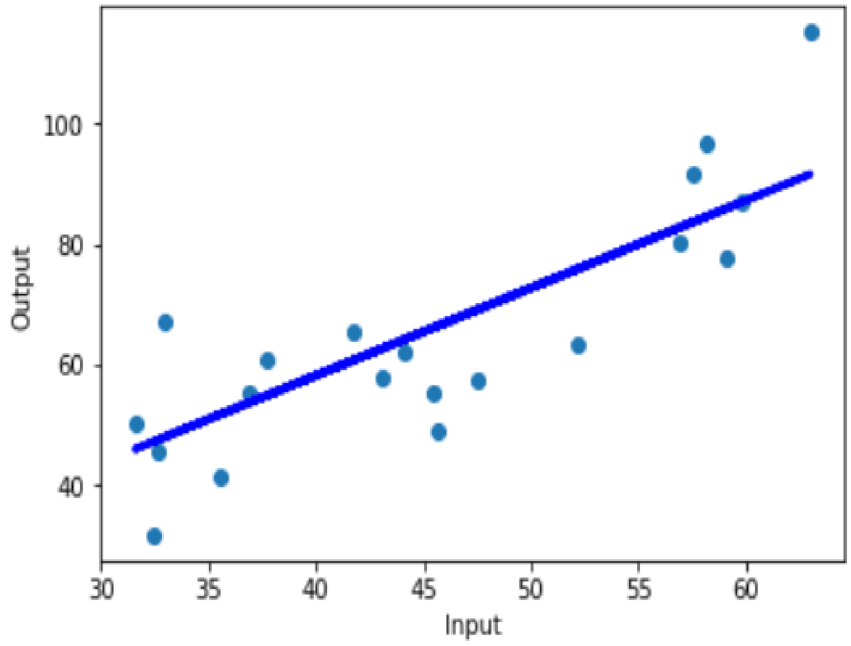
m: is correlation coefficient or slope.

c: is intercept, where the line touches the y axis.

X: independent variable.

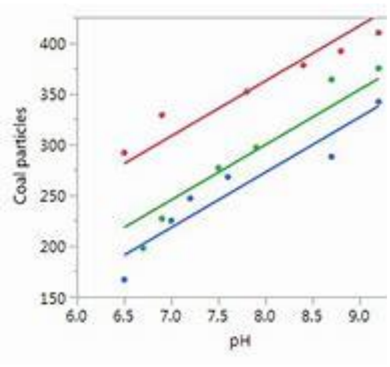
Y: target variable.

Linear regression is used for the supervised learning, as it uses historical data to train and test the model.



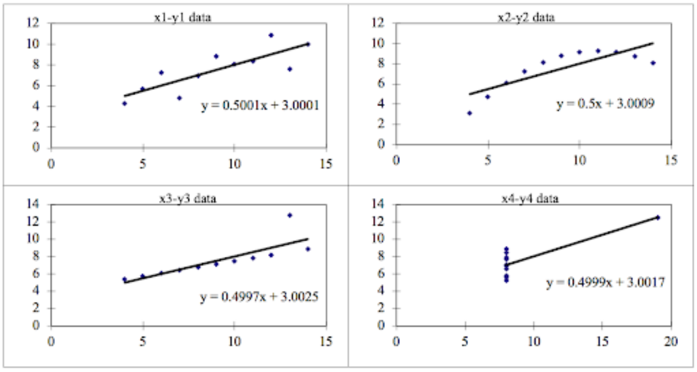
When target variable is dependent on multiple variable then it becomes then we use Multiple Linear Regression model to predict the target variable. It is similar to simple linear regression and defined as below

Y = C + m1.x1 + m2.x2 + m3.x3 + ……. + e



2. Explain the Anscombe’s quartet in detail. (3 marks)

Anscombe a statistician, has created four datasets which has similar statistical properties like mean, standard deviation, and correlation between x and y, but when plotted scatter plot for visualization they were completely different plots.



This visual behaviour is the motivation to view the datasets in graphs, to identify the nature of the data and validate if the dataset is good for linear regression model or not. So this quartet explains the importance of visualization in data analysis.

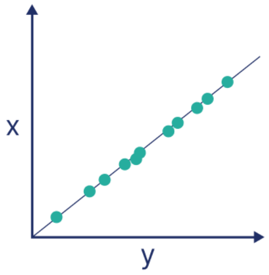
3. What is Pearson’s R?

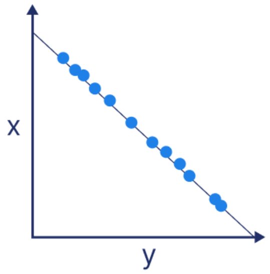
The Pearson correlation coefficient, also called Pearson’s R, is a statistical calculation of the strength of two variables’ relationships. It is used to measure, how two variables are dependent on each other.

It is used to identify linear correlation between variables, most importantly between Target and Predictor variable. Value ranges from 0 to 1, if value is towards 0 then the variables are less correlated and if the value is close towards 1 means they are highly correlated and value 1 means perfect correlation between variables.

It is the basic for linear regression model.

If the plotted line is upward then the correlation is considered as positive and if the line is downward which means the correlation is negative.





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

Scaling is step in data pre-processing to make it ready for machine learning Normalized and Standardized scaling used to do that. Data collected from different sources shows different units and range, so to make them useful so that the algorithms processes them in better manner, scaling is needed.

Normalized scaling transforms feature to be on same scale and new point is calculated as below

X_new = (X - X_min)/(X_max – X_min)

Normalized scaling also known as Min Max scaling.

Standardized scaling transforms feature data against the mean. It subtracts the value from mean and divides it by standard deviation.

X_new = (X - mean)/Std

Differences:

1. Normalized scaling usage Min and Max values for scaling, whereas Standard scaling uses mean and Standard deviation.
2. Normalized scaling makes the unit and range same for all values, where as standardization cares more for zero mean.
3. Normalized scaling gets affected by outliers easily, standardized scaling is less affected by outliers.

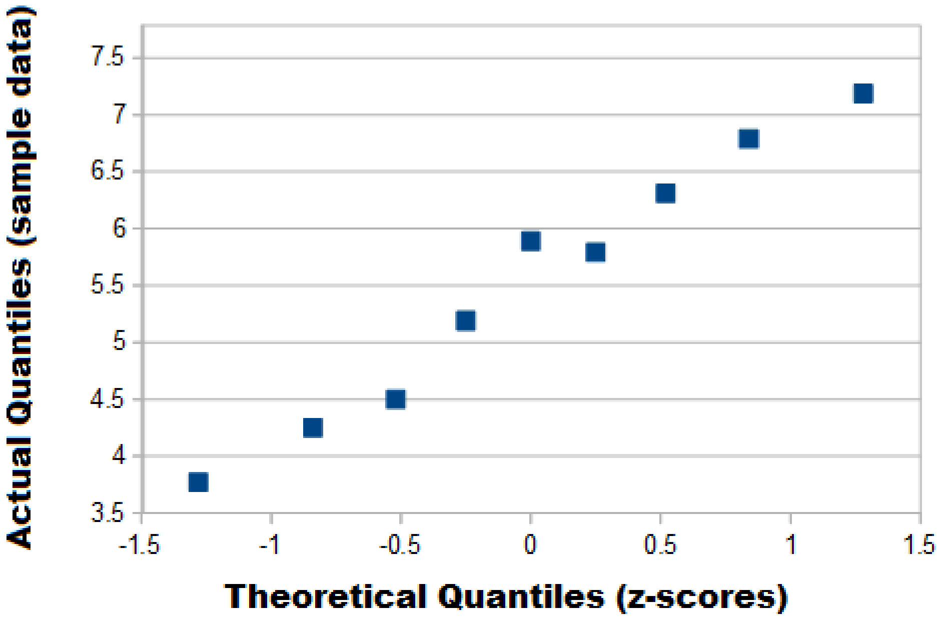
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Yes, have observed infinite VIP, and this happens when two variables are showing perfect correlation. Which means they are highly linearly related to each other. And this is very important to take the decision what variables need to be omitted for linear regression model.

VIF is used to determine multicollinearity among predictor variables. And high value shows correlation between two predictor variable. Same like we have in this assignment between **temp** and **atemp.**

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

Ans: Q-Q (Quantile-Quantile) plot is used to visually determine how close a sample is to normal distribution. The Q-Q plot orders the z-scores ascending and plots each z-score value on the y-axis. The x-axis is the corresponding quantile of a normal distribution for that value rank. If the plot show diagonal line, then the sample distribution can be considered as normally distribution.



Good example for the Q-Q plot would be daily stock returns of Netflix.

Usage:

a. If two samples are drawn from the same population.
b. Samples have the same tail and distribution shape.

Importance:

When we use the QQ plot to compare the samples, it does not depend on the sample sizes also as it works normalized data so dimensions of values does not impact.

It is important in linear regression to validate the train data and test data whether follow the same distribution and population or not, and do determine that we can make good use of Q-Q plot.