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)

Ans) The demand for shared bikes increases in the months of “June, July, August” which is nothing but summer season (season_summer).

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

Ans) In case of transforming the categorical variables into dummy variables, adding columns for all the levels of the categorical variable will introduce multicollinearity. Moreover with the (drop_first=True) we still be able to represent all the information with n-1 columns , by setting all the columns to “0” to represent the first dropped column.

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

Ans) The target filed “cnt” has highest correlation with “temp” variable.

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

Ans) To validate the assumption of a Linear Regression

i) We check the residual errors are normally distributed by plotting a histogram for the error terms (Yactual-Ypred) and making sure the errors are normally distributed.

ii) Residual errors should not follow any pattern. We can validate this by doing a scatter plot of (len(X_train) , residuals)

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

Ans) The top 3 factors that influence the Shared bikes demand are

temp - Every unit increase in temperature will increase the demand by 3875 times

year - Every year the demand for shared bikes is growing by 1979 times

Weathersit_LightRain - The demand for shared bikes decreases by 2034 times in case of a light rainfall

General Subjective Questions

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

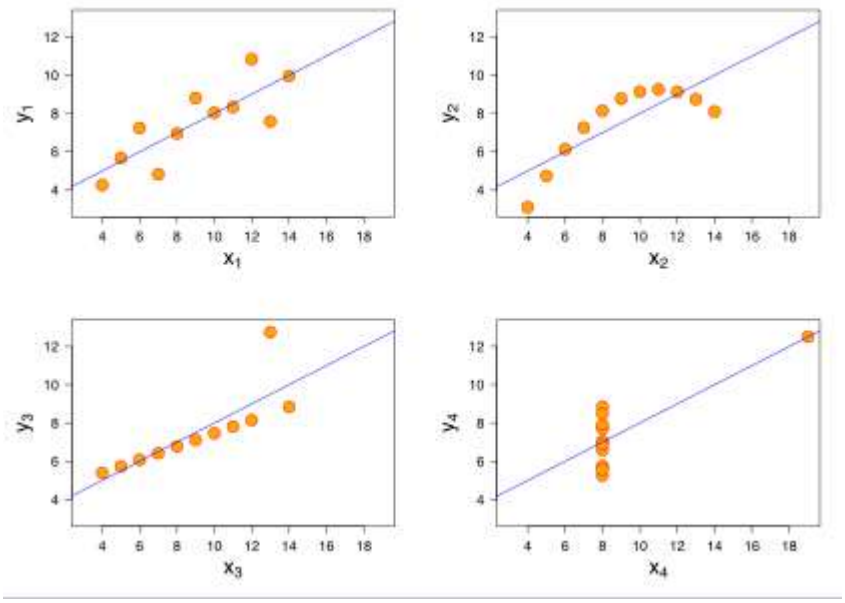
Ans) Linear regression algorithm mainly aims to establish the relationship between the target variable and independent variables in the dataset by fitting a straight line Y= mx+c .

The straight line is fitted such that the absolute square distance between the Y actual values from training dataset to the Y predicted values from the straight line using Residual sum of Squares.

To minimize this cost function (Residual sum of Squares), we follow the gradient descent model to find the optimal solution that minimizes the error terms.

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

Ans) Scatter Plot of Anscombe's Quartet dataset



Anscombe's Quartet contains four datasets with same descriptive statistics like mean, variance when calculated for all the four sets of variables.

Based on the mean, variance numbers of these datasets they may look similar but they indeed have a very different distribution of data which is only apparent when we visualize the data. The main purpose of this experiment is to indicate the importance of data visualisation for determining the outliers, check data distribution and take necessary actions like outlier removal during the EDA analysis.

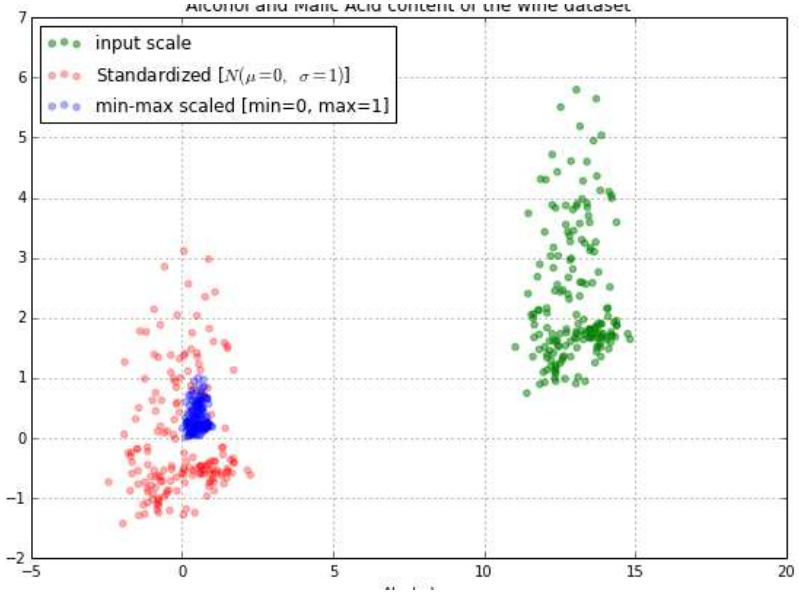
3. What is Pearson's R? (3 marks)

Ans) Pearson's R is a correlation co-efficient which signifies the linear correlation between the two variables in a dataset and is obtained by the ratio of covariance to standard deviation of both the variables . The value of the Pearson' R coefficient lies between -1 and 1 . Where the value of 1 signifies that both the variables are directly proportional to each other and a negative value towards -1 signifies inverse relation between two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans) Scaling is a step to normalize the independent variables being used for ML modelling. This is mainly performed to bring all the variables in a scale between 0 and 1 . Leaving the variables with the default scales like sales in Crores and sale_quanty in thousands which is very small leaving the sales variable will having smaller co-efficient and sale quantity having larger co-efficient. To avoid this bias, we rescale the variables using Normalization or standardization methods. Another reason for scaling all the variables is the predictor algorithms finish faster if all the variables are in same range.

With Normalized scaling there would be a loss out outliers and all the values of the variables are fitted between 0 and 1. Standardized scaling does not cause loss to outlier information. An example scatter plot representing the distribution of variables after MinMax scaling vs Standardized scaling.



5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

A large VIF value indicates the variable is highly correlated with the rest of the variables in the dataset and is a redundant variable in the model prediction. Based on the other factors like p-value we can make a call to drop such variables during the multi-linear regression ML models.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans) Q-Q plot is used to visually determine how close the sample data is to the normal distribution. In case of Linear regression , during the residual analysis we can plot the residual error terms using

QQ plot to make sure the error terms (Yactual – Ypred) are normally distributed to validate the linear regression assumptions.