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?

- Temperature is the most important factor determining the demand. Higher the temperature higher the demand
- Demand is high in Winter
- Demand is high in month of september
- Being a holiday negatively impacts the demand
- Weather conditions also negatively impact the demand. Snow, high wind, cloudy misty weather reduces the demand

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

Using drop_first=True when creating dummy variables is important to prevent multicollinearity. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, and this can lead to issues in estimating the regression coefficients.

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

We had created a pair plot during our analysis, and it was clear that temperature had the highest correlation with the target variable

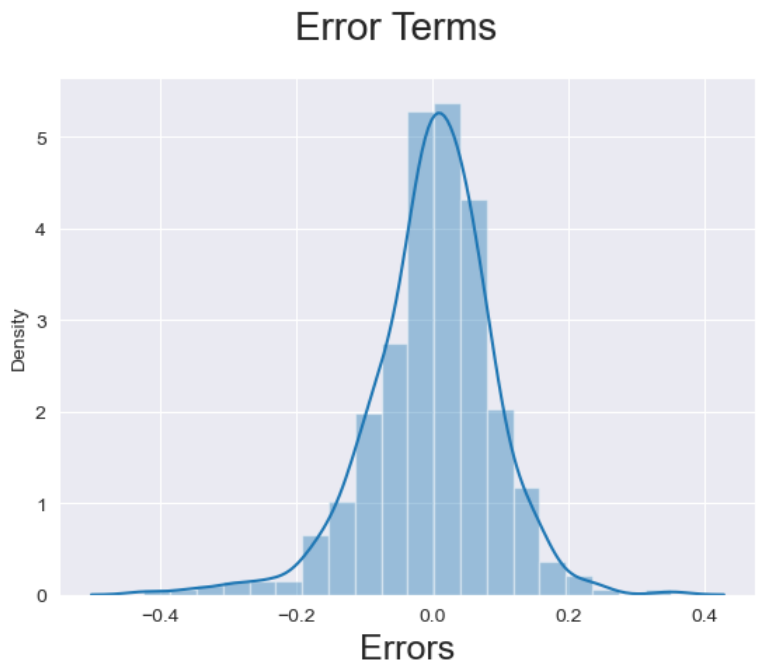
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Validating the assumptions of linear regression is crucial to ensure the reliability and accuracy of the model.

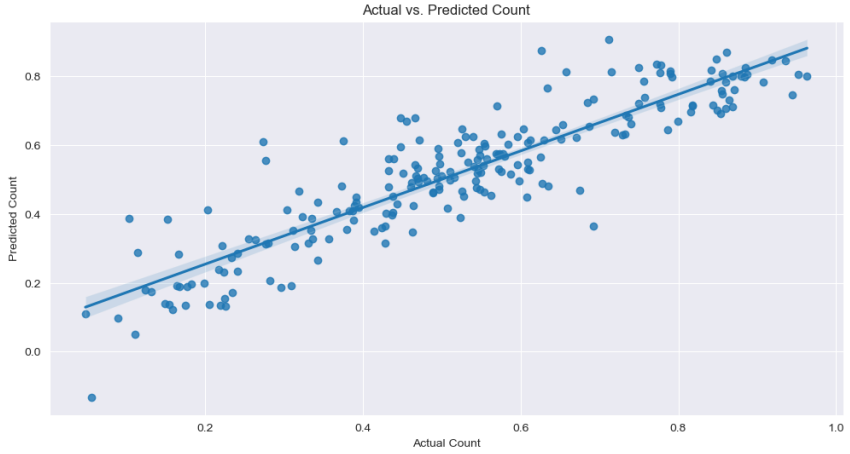
So after the model was built, validated it using

- Residual Analysis
- QQ Plot

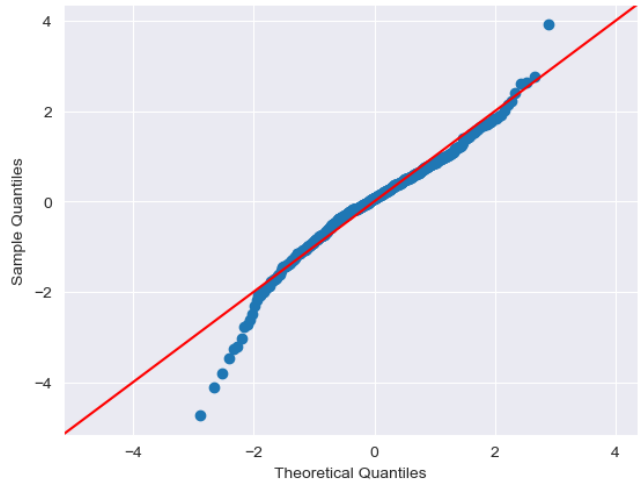
Error terms are normally distributed



Error Terms have constant Variance



QQ Plot



The points should all be around a line

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

The top 3 features contributing significantly towards the demand of the shared bikes are the temperature, the year and the winter variables.

General Subjective Questions

1. Explain the linear regression algorithm in detail?

Linear regression is a statistical method used for modelling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The goal is to find the best-fitting line that describes the relationship between the variables. This line is also known as the regression line.

There are two types of linear regression
- Simple Linear Regression: A linear regression model with one independent and one dependent variable.
- Multiple Linear Regression: A linear regression model with more than one independent variable and one dependent variable.

The basic idea is to represent the relationship between the dependent and independent variables using a linear equation of the form:

$Y=mX+b$

Y = Dependent variable

X = Independent variable
m = How much Y changes with one unit change in X
b = Value of Y when X is 0

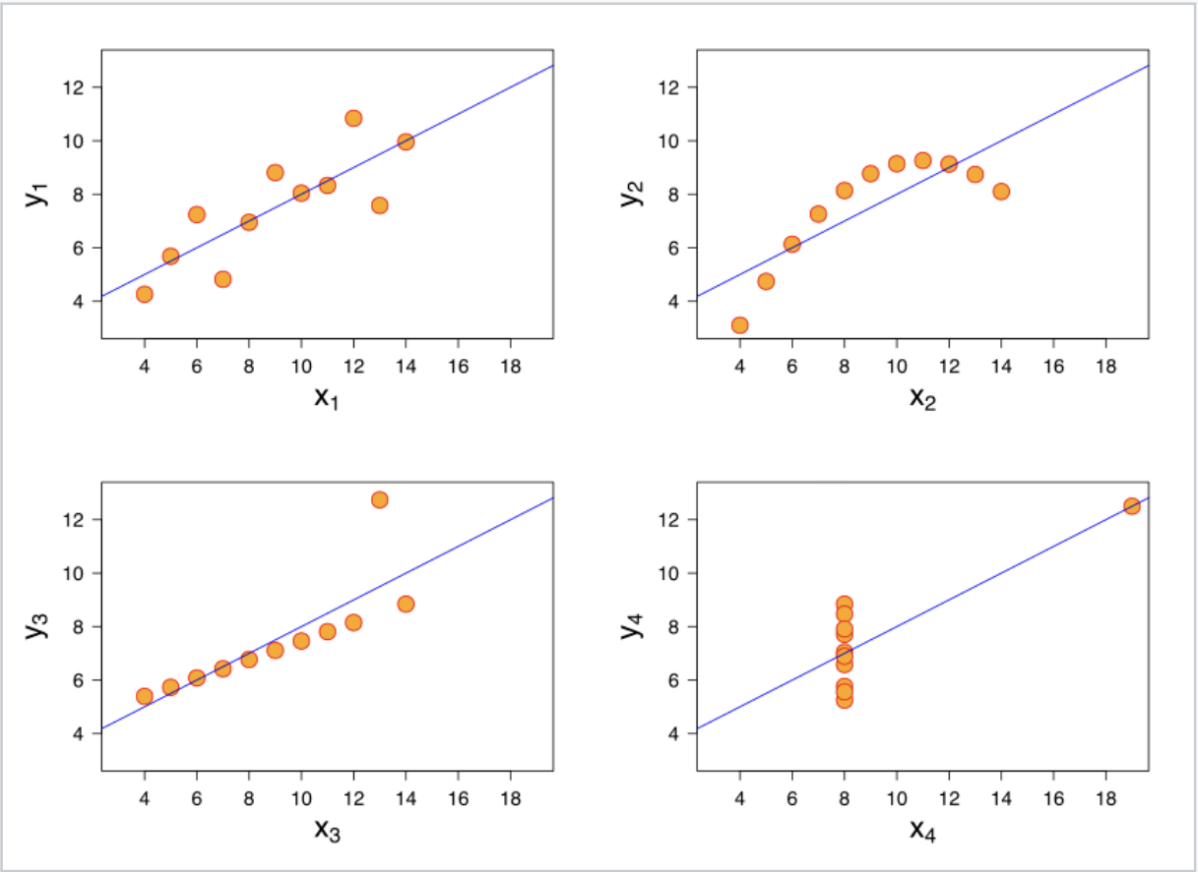
The objective of linear regression is to find the values of m and b that minimise the sum of the squared differences between the actual values and the values predicted by the linear equation.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but differ significantly when graphed. It emphasises the importance of visualising data and not relying solely on summary statistics. The datasets share the same mean, variance, correlation coefficient, and linear regression line, but they have distinct patterns when plotted.

Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



As we can clearly see from above, the 4 data sets if analyse using summary statistics would look identical, but when we plotted them, they have distinct patterns

3. What is Pearson's R

Pearson's correlation coefficient(r) is a statistical measure of the strength and direction of a linear relationship between two continuous variables.

It quantifies the degree to which a change in one variable is associated with a change in another variable.

- The coefficient ranges from -1 to 1, where,

  r=1 indicates a perfect positive linear relationship,
  r=−1 indicates a perfect negative linear relationship, and
  r=0 indicates no linear relationship.

- The closer r is to 1 or -1, the stronger the linear relationship. The closer it is to 0, the weaker the linear relationship.

- The sign of r indicates the direction of the relationship:
  positive (r>0) means as one variable increases, the other tends to increase,
  negative (r<0) means as one variable increases, the other tends to decrease.

4. What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?

Scaling is the process of transforming or standardising the values of variables in a dataset. The goal of scaling is to bring all variables to a similar scale, which can be important for certain machine learning algorithms and statistical techniques. The primary reasons for scaling are to ensure that no variable dominates due to its scale, and to improve the convergence and performance of certain algorithms.

There are 2 types of scaling that are normally used

Normalised Scaling

- Scales the data to a specific range, often [0, 1].
- Preserves the shape of the original distribution but compresses it into the specified range.

Standardised Scaling

- Transforms the data to have a mean (μ) of 0 and standard deviation (σ) of 1.
- Makes the data more suitable for algorithms that assume normal distribution or those that are sensitive to variable scales.

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

The Variance Inflation Factor (VIF) is a measure used to quantify the severity of multicollinearity in a regression analysis. Multicollinearity occurs when independent variables in a regression model are highly correlated with each other. The VIF assesses how much the variance of an estimated regression coefficient increases when the predictors are correlated.

When the VIF is infinite for a particular variable, it means that there is perfect multicollinearity involving that variable. Perfect multicollinearity occurs when one or more independent variables in a regression model can be exactly predicted by a linear combination of the other variables. In other words, the correlation between the variable in question and the remaining independent variables is perfect, leading to a situation where the VIF becomes infinite.

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

As we can see from the above formula for VIF, the (R-squared) is the coefficient of determination obtained by regressing the ith variable against all the other independent variables.

If R-squared is very close to 1, it indicates that the variance of the ith variable is very high compared to its covariance with the other independent variables. In such cases, the VIF approaches infinity, and this suggests a severe multicollinearity problem.

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

A Q-Q plot, which stands for Quantile-Quantile plot, is a graphical tool used to assess whether a given dataset follows a particular theoretical distribution. In the context of linear regression, Q-Q plots are often used to check the normality assumption of the residuals.

If the points on the Q-Q plot fall approximately along a straight line, it suggests that the residuals (the differences between the observed and predicted values) follow a normal distribution. Deviations from a straight line may indicate departures from normality.

In linear regression, the normality assumption is crucial for making valid statistical inferences, such as hypothesis testing and constructing confidence intervals. If the residuals are not normally distributed, it can affect the accuracy and reliability of the statistical tests associated with the regression model.