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.1- Some of the variables like yr, weathersit, season and holiday gives some insight into the data. Yr 1 has high percentiles than yr 0 , weathersit 1 is the best for demand and weathersit 3 is the worst, holiday seems to bring some reduction in demand.

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

Ans2 - We can drop first because only n-1 variables are required to explain n levels. For example, to explain male and female only one feature is sufficient, male 0,1 means female and male respectively.

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

Ans3- Temp has the highest correlation with the target variable. Well, temp and atemp are highly correlated. So, they appear to have same correlation with the target variable.

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

Ans4- Assumptions were validated by plotting a distribution plot of error terms. They were seen to be normally distributed with mean corresponding to 0.

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

Ans5- Temperature, year and weathersit_3 are the top 3 features.

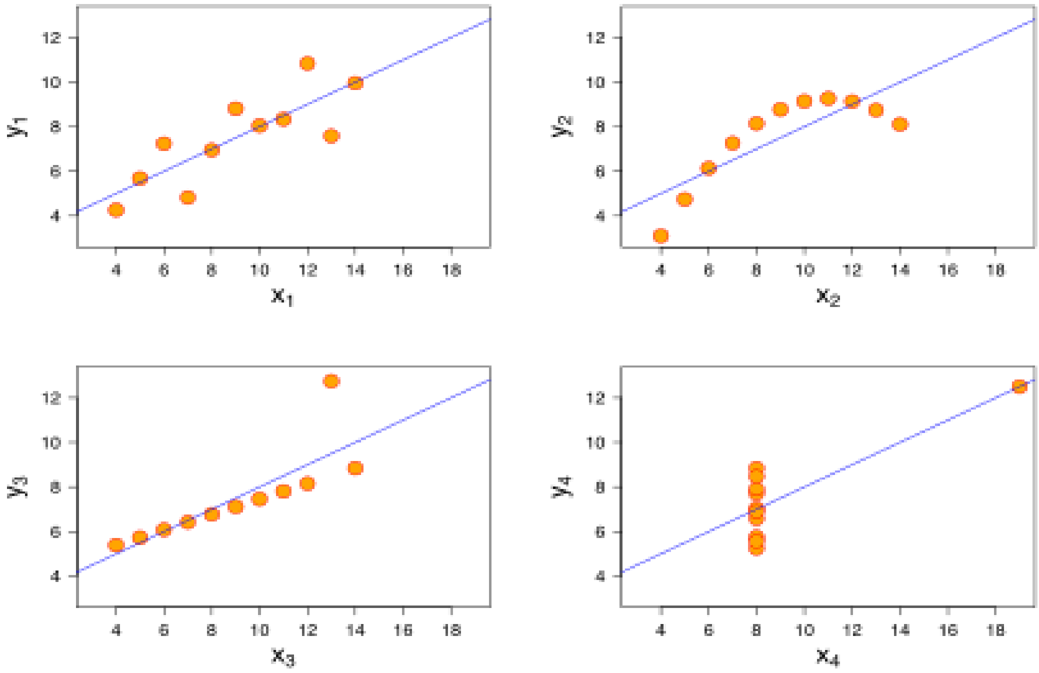
1. Explain the linear regression algorithm in detail. (4 marks)

Ans- Linear regression is used to predict continuous variables. In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.

Here y is called a dependent or target variable and X is called an independent variable also known as the predictor of Y. A linear function is the simplest type of function. Here, X may be a single feature or multiple features representing the problem.

Linear regression model – $y = \beta_0 + \beta_1x_1 + \ldots\ldots\ldots + \beta_nx_n + \text{error}$

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



Here, we can see 4 different dataset distribution but very similar distribution. Here, we can see numerical calculations are exact, but graphs can be deceptive. Just because of some outliers the graphs produced are identical.

These four datasets have the same mean, regression line and standard deviation but qualitatively very different. The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

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

Ans – Pearson's correlation coefficient R is a measure of correlation between two variables. It ranges from -1 to 1, where 0 means no relationship between the two variables, negative R means the variables are inversely correlated, i.e. decrease in one increases the other, and positive R means they are directly correlated. The closeness to 1 or -1 depicts the intensity of correlation. Higher the number, higher the correlation.

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

Ans- a. Scaling is a preprocessing step of data modelling, where the independent variables are normalised.

b. Scaling is done because the data vary in magnitude and scale which may result in coefficients being greatly manipulated by the magnitude, so to avoid this and to let model interpret all the features in same scale we perform scaling.

c. **Normalized** scaling, scales the data between 0 and 1. This is also known as min max scaling.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardized scaling centers data around the mean and scales to a standard deviation of 1, Equation: (x – mean)/standard deviation.

5.You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Ans – When the variable can completely be represented by other variables, the vif comes out to be infinite.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Ans – A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Q-Q plot can help to know whether the two data sets come from populations having similar distribution or not, have same scale or not.

In case of linear regression we can see whether the test and train data are following the same distribution or not, also shifts in scales and outliers can also be detected. Moreover, we can check whether residuals follows normal distribution or not.