## 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-
   1- Fall Season highest demend for rental bikes.
   2- Demand is continuously growing till June , then sudden hike in September then decrease in demand after September.
   3- Demand for next year has grown up.
   4- On Holiday , demand has decreased.
   5- The clear weather situation has highest demand.

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

Ans -Drop_first we use to remove the extra column that get created during dummy variable and it reduce the correlation between the dummy variables.Dropping the first columns as (p-1) dummies can explain the p categories .

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

Ans- temp and atemp has highest correlation (.63)with target variable (cnt).

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

Ans- Residual Analysis –

Error are normally distributed with mean of 0 .

Actual and predicted result follow same pattern.

The error terms are independent of each other .

R2 Value for test Prediction-

R2 value for predictions on test data is .819 is almost same as R2 value in train data . this is good R squared value.

Homoscedacity –

We can observe that variance of the residual is constant across prediction

Plot Test Vs Predicted value Test-

The Prediction for test data is very close to actuals

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- Top 3 Features are :

1- Yr (Positive Correlation )
2- Temp(Positive Correlation)
3- Weathersit_bad(Negative correlation)

## General Subjective Questions

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

Ans- Linear regression is a stats method used to predict analysis and show the relationship between continuous variables.Linear regression shows the linear relationship between independent variable(X) and the dependent variable (Y).If there is 1 input variable it is called linear regression .If there is more than 1 variable it is called multiple linear regression The Linear relationship model can be describe by

**$Y = \beta 0 + \beta.X + \epsilon$**

Y- Dependent Variable

X-Independent Variable

**β0**- Intercept of the line

**β** - Linear regression Coefficient

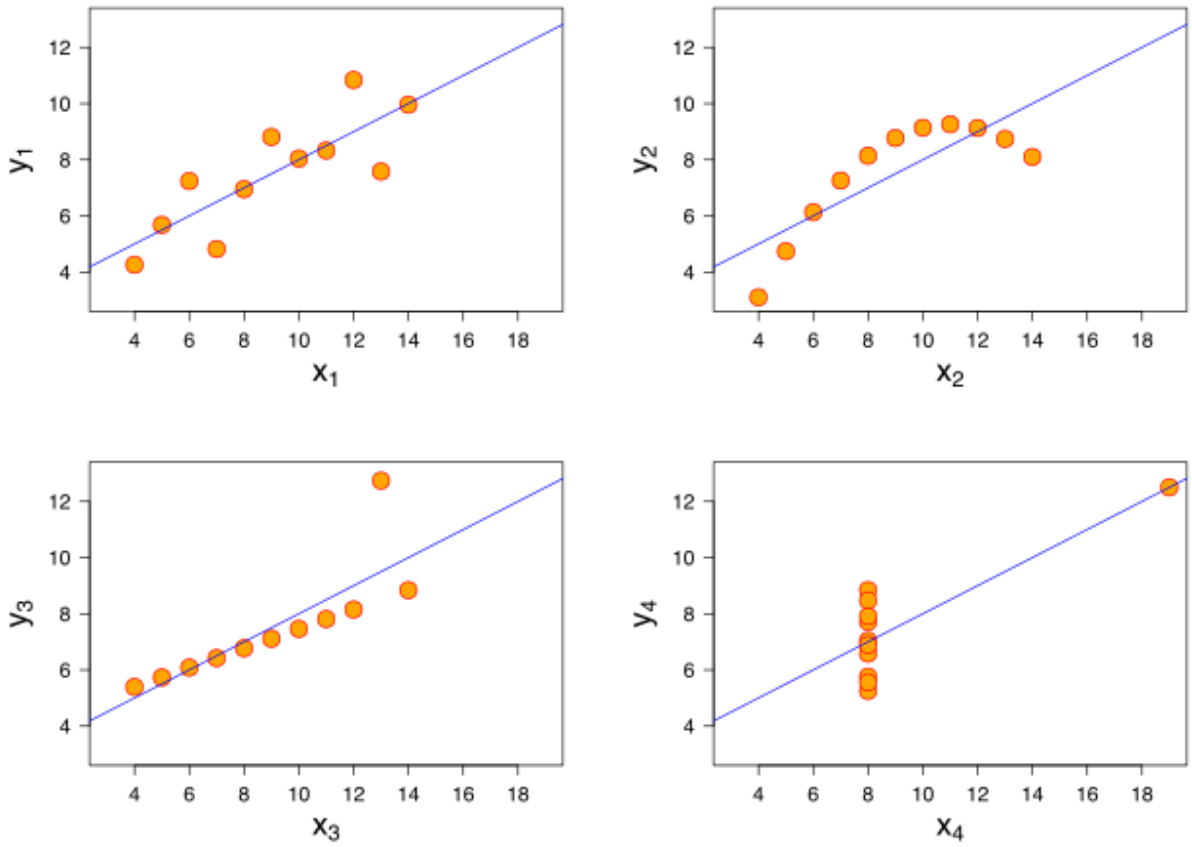
**ϵ** -random error

The goal of linear regression algorithm is to get the best value for **β0 β.X and** find the best fit line. The best fit line should have least error means the error between predicted value and actual value should be minimum.

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

Ans- **Anscombe's quartet** comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.

This tells us about the importance of visualizing the data before applying various algorithm out there to build models out of them which suggests that the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers .Also , the Linear regression can only be considered a fit for the data with linear relationship and is capable of any other kind.




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

   Ans- Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

   The Pearson's correlation coefficient varies between -1 and +1 where:

   r = 1 means the data is perfectly linear with a positive slope ( i.e., both variables tend to change in the same direction)

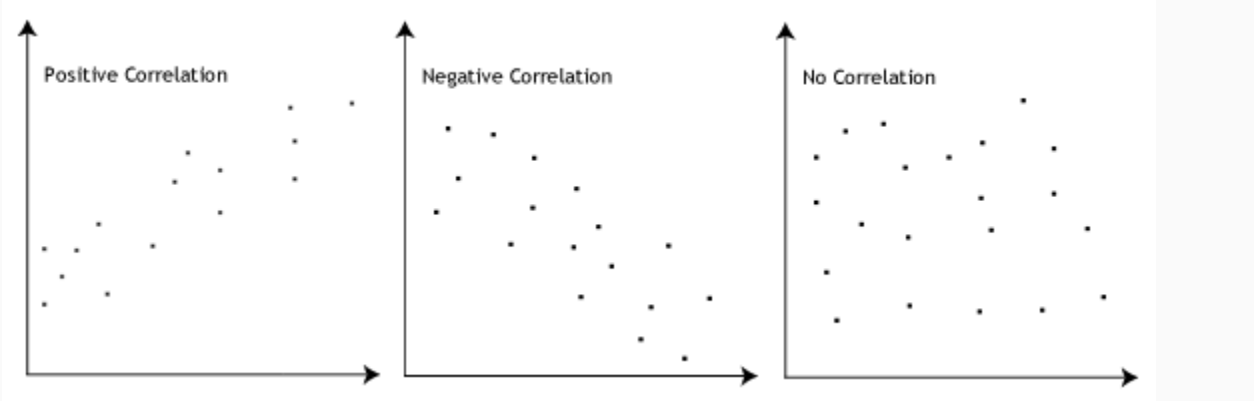
   r = -1 means the data is perfectly linear with a negative slope ( i.e., both variables tend to change in different directions)

   r = 0 means there is no linear association

   r > 0 < 5 means there is a weak association

   r > 5 < 8 means there is a moderate association

   r > 8 means there is a strong association




4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   Ans- It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

When we collect data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

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

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

- sklearn.preprocessing.scale helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means ofestimating

parameters in a location-scale family of distributions.A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.