Linear Regression – Bike Sharing Assignment

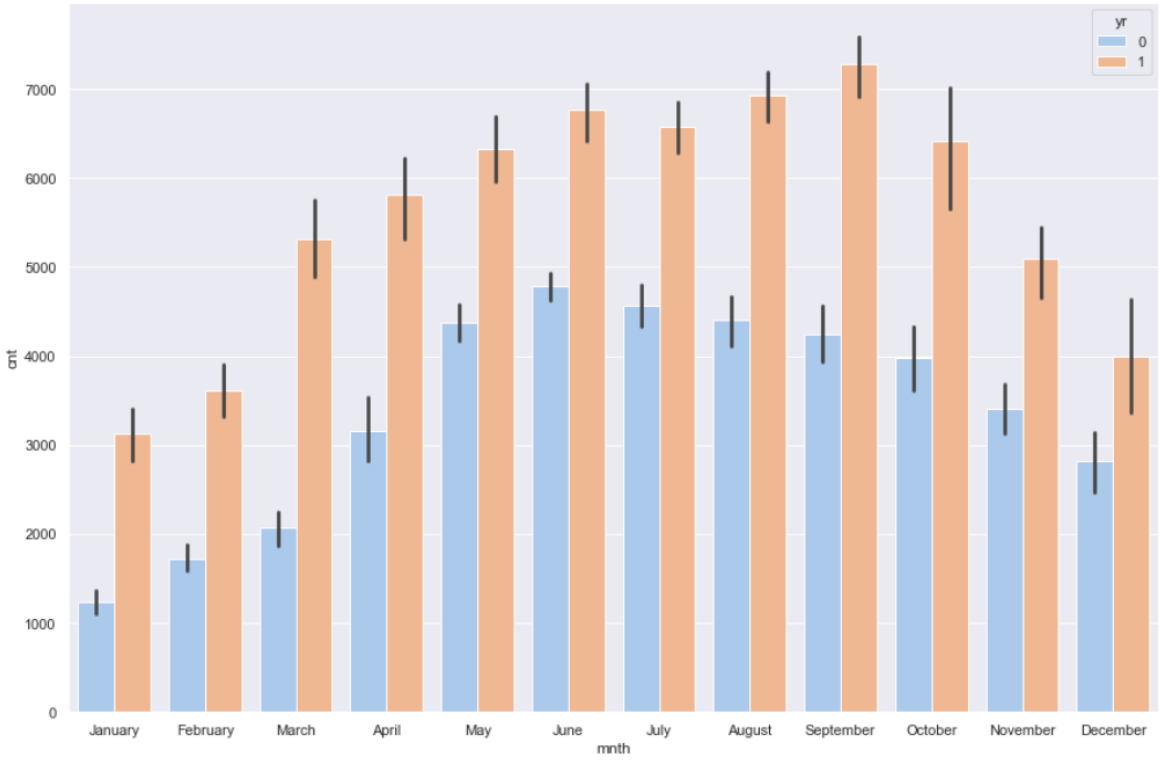
Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

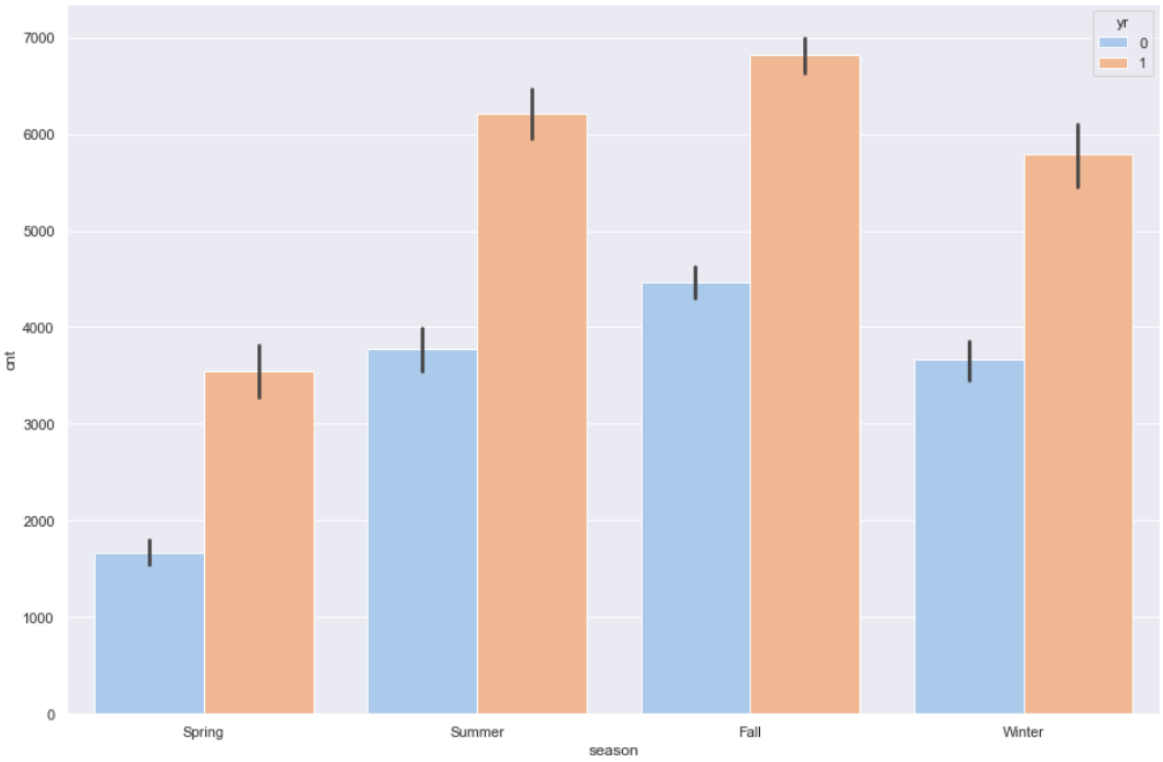
Ans. With the analysis conducted using the provided data, we can infer that:

- Below are the categorical variables identified
  - Mnth
  - Season
  - Weathersit
  - Weekday

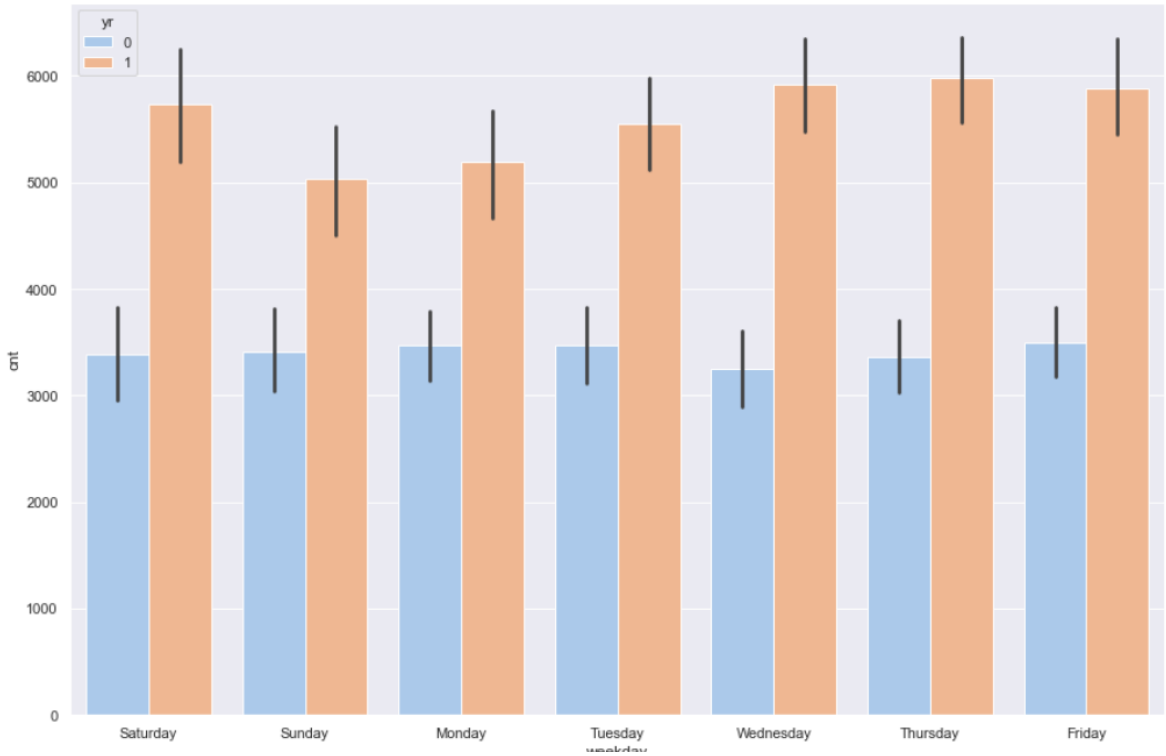
- Mnth variable analysis over the two years as below
  - Inference: In the year 2019, the rides are very high when compared to 2018
  - Rides count is higher during the period Apr-Oct



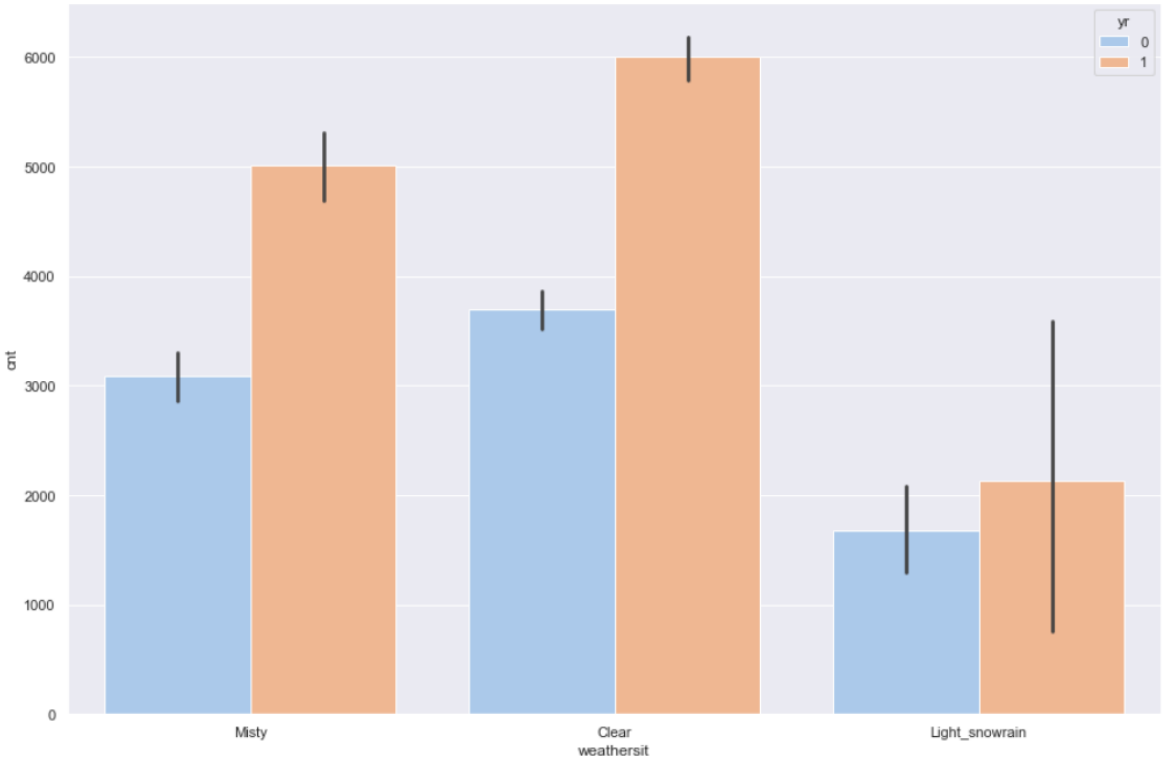
- Season variable analysis over the two years as below
  - Ride Count Seems to be in maximum in Fall (Autumn) followed by Summer, Spring & Winter respectively.



- Weekday variable analysis over the two years as below
    - Weekday Column seems to be scattered evenly across all the points. Unable to make out any pattern when predicting Ride Count from the weekday column alone.



- Weathesit variable analysis over the two years as below
    - Ride Count is more on plesant & moderate Days as compared to Light Snow / Rainfall

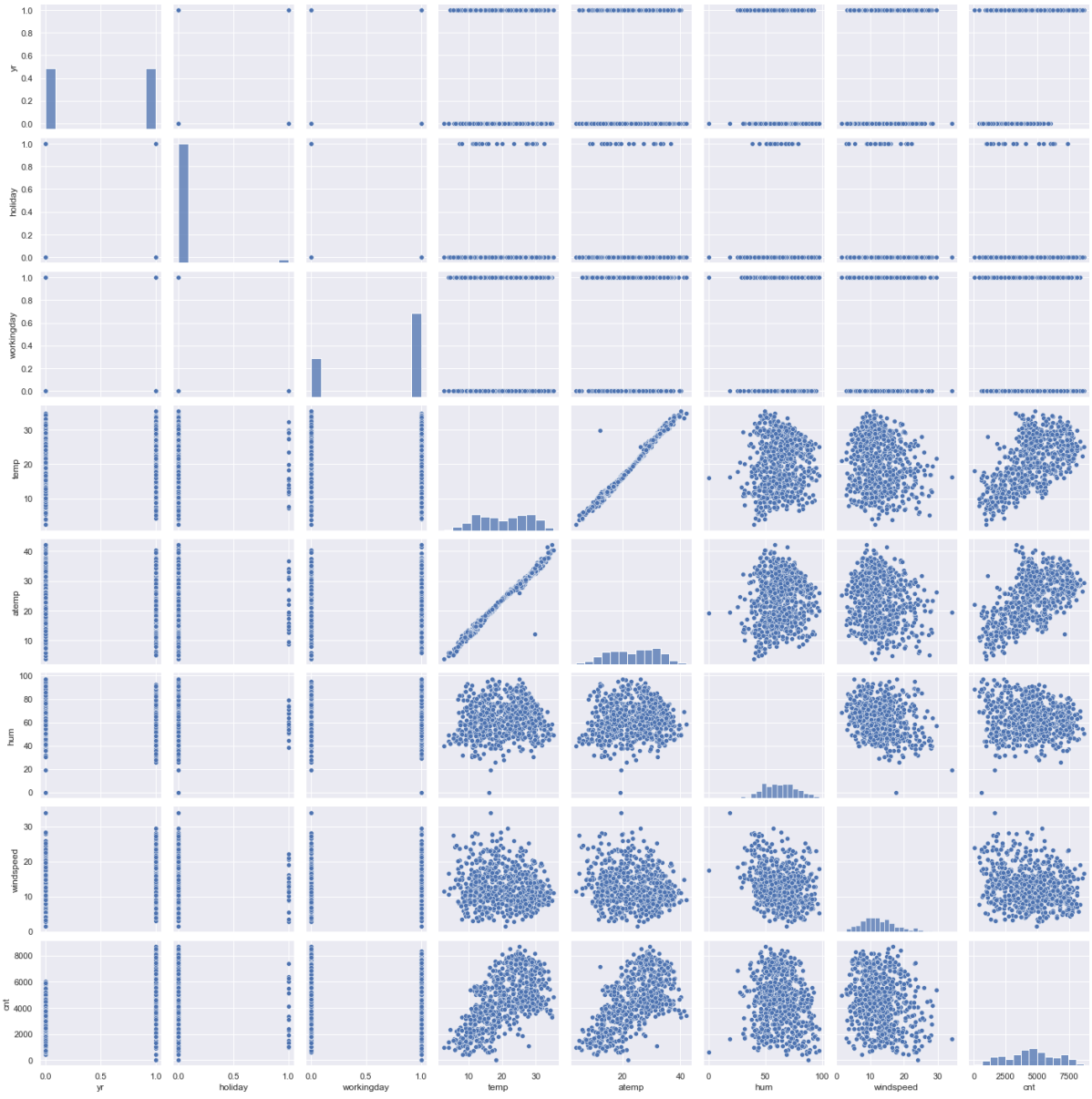


Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans. drop_first=true is necessary to remove the redundant variables. Drop first is essential when creating dummy variables because it helps to avoid the problem of data multicollinearity. If the first column is not removed, the model estimates will be unstable and biased due to perfect correlation with the other dummy variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans. For the pairplot below; we can easily infer that temp and atemp are having highest collinearity.



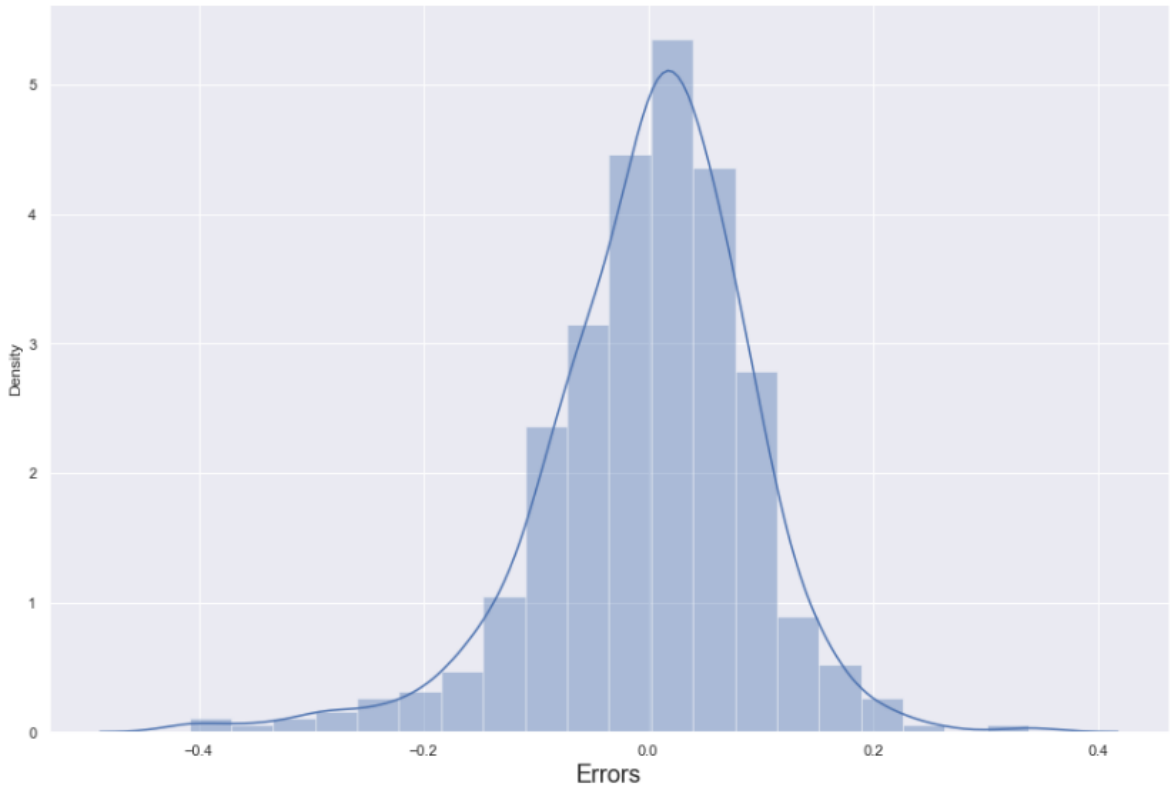
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans. We can validate the assumptions of linear regression by analysing the pairplots, understanding the relationship between dependent and independent variables

Validate multicollinearity through VIF values

	Features	VIF
1	temp	5.12
2	windspeed	4.62
6	season_Summer	2.23
5	season_Spring	2.08
0	yr	2.07
7	season_Winter	1.78
3	mnth_July	1.58
10	weathersit_Misty	1.55
4	mnth_September	1.33
8	weekday_Saturday	1.18
9	weathersit_Light_snowrain	1.08

From the below graph, we can infer that error distribution is normallly distributed across 0, which indicates that our model has handled the assumption of error normal distribution properly.

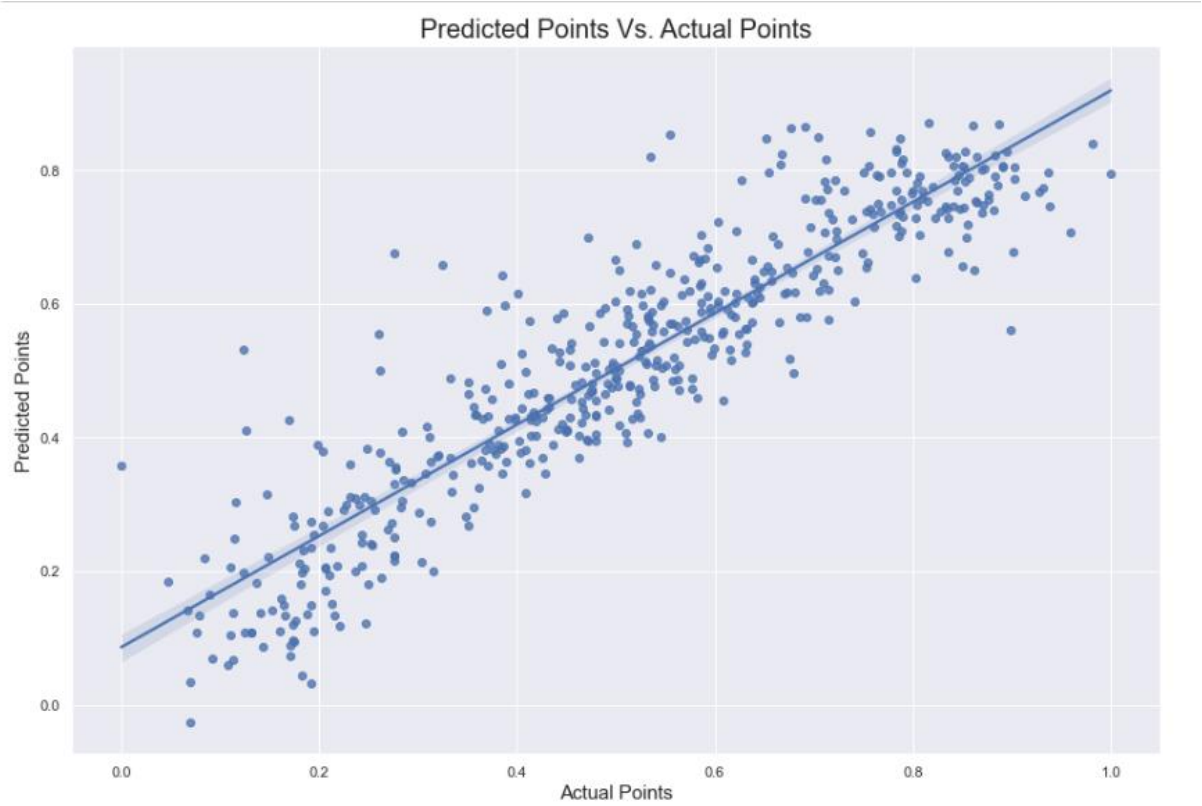


From the below graph,

- We can infer that residuals are equal distributed across predicted value.

- It implies, we see equal variance and we do not observe high concentration of data points in certain region & low concentration in certain regions.

This proves Homoscedasticity of Error Terms



Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans. Based on the final model, the top three features contributing significantly towards explain the demand for shared bikes are: temp, year, and September

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans. Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation –

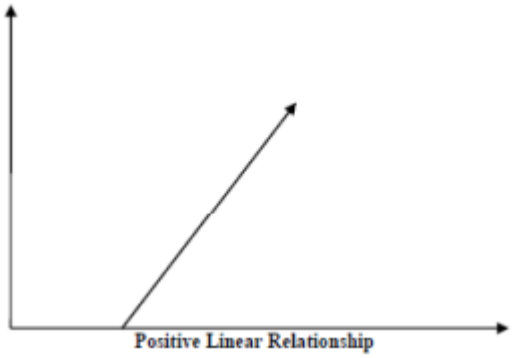
$Y = mX + c$

Here,

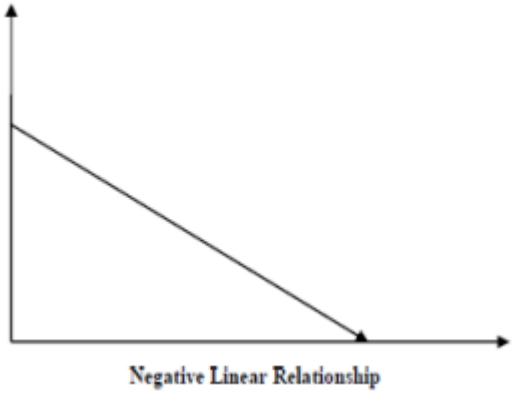
- Y is the dependent variable we are trying to predict.
- X is the independent variable we are using to make predictions.
- m is the slope of the regression line which represents the effect X has on Y
- c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

Furthermore, the linear relationship can be positive or negative in nature as explained below−

- Positive Linear Relationship: A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph –



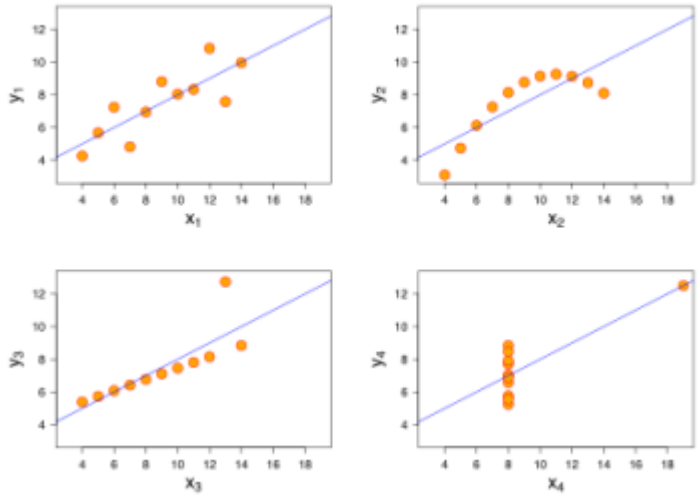
- Negative Linear relationship: A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph –



- Linear regression is of the following two types –
    - Simple Linear Regression
    - Multiple Linear Regression

Q2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough



Q3. What is Pearson's R?

Ans. The Pearson correlation coefficient is a descriptive statistic, which means it summarizes a dataset's characteristics.

The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:

- Pearson's r
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient.

It describes the strength and direction of a linear relationship between two quantitative variables in particular. Although different disciplines have different interpretations of relationship strength (also known as effect size). In addition, the Pearson correlation coefficient is an inferential statistic, which means it can be used to test statistical hypotheses. We can specifically test for a significant relationship between two variables.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans. It is a data Pre-Processing step that is applied to independent variables in order to normalize the data within a specific range. It also aids in the speeding up of algorithm calculations.

Most of the time, the collected data set contains features with widely disparate magnitudes, units, and ranges. If scaling is not performed, the algorithm only considers magnitude rather than units, resulting in incorrect modelling.

To solve this problem, we must scale all of the variables to the same magnitude level. Normalized Scaling - It gathers all data between 0 and 1. sklearn.preprocessing. MinMaxScaler aids in the

implementation of normalization in Python. Values are replaced by their Z scores after standardization.

Standardized Scaling - It transforms the data into a standard normal distribution with a mean () of zero and a standard deviation of one (). sklearn.preprocessing. Python's scale aids in the implementation of standardization. One disadvantage of normalization over standardization is that it removes some data information, particularly about outliers.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans. The value of VIF is infinite when there is perfect multicollinearity in the data, meaning that the predictor variables are perfectly correlated with each other, making it impossible to determine the unique effect of each variable on the outcome.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans. A Q-Q plot (Quantile-Quantile Plot) is a graphical method to check if a set of data is approximately normally distributed. It plots the sample quantiles against the theoretical quantiles of a normal distribution.

In linear regression, a Q-Q plot is used to check the assumption of normality of residuals, which is important for valid inference. If the residuals are not normally distributed, it can affect the validity of statistical tests and confidence intervals for the regression coefficients. A Q-Q plot helps to visually inspect if the residuals are close to a straight line, indicating normality, or if there are deviations, suggesting non-normality