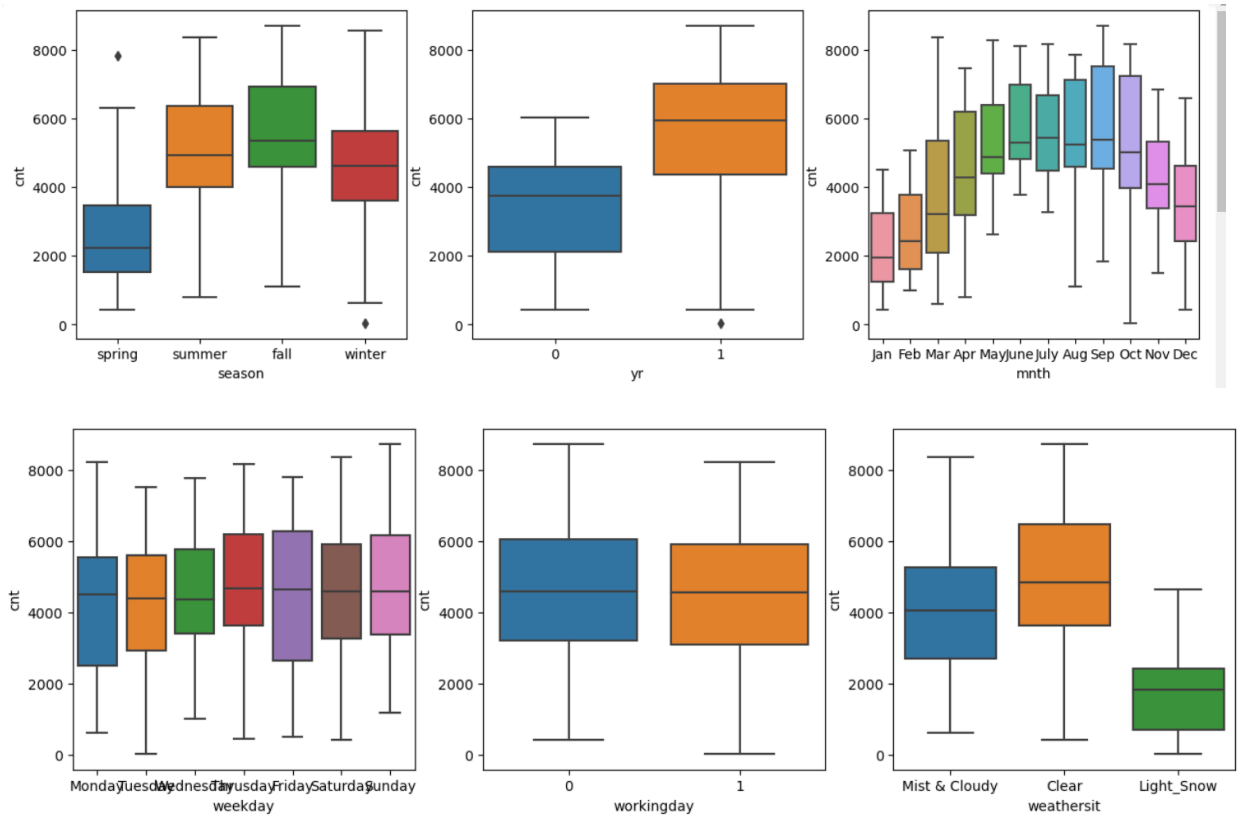
Assignment-based Subjective Questions

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

Ans. Season, Yr, Mnth, Holiday, Weekday, Weathersit are categorical variables in the dataset.

Bike demand in the fall is the highest.

- Bike demand takes a dip in spring.
- Bike demand in year 2019 is higher as compared to 2018.
- Bike demand is high in the months from May to October.
- The demand of bike is almost similar throughout the weekdays.
- Bike demand doesn't change whether day is working day or not



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

Ans. It is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables.

We use Syntax of dummy= pd.get_dummies Lets say we have 3 types of values in categorical column example as shown below:-

travel type	Train	Bike	Plane
Plane	0	0	1
Bike	0	1	0
Train	1	0	0

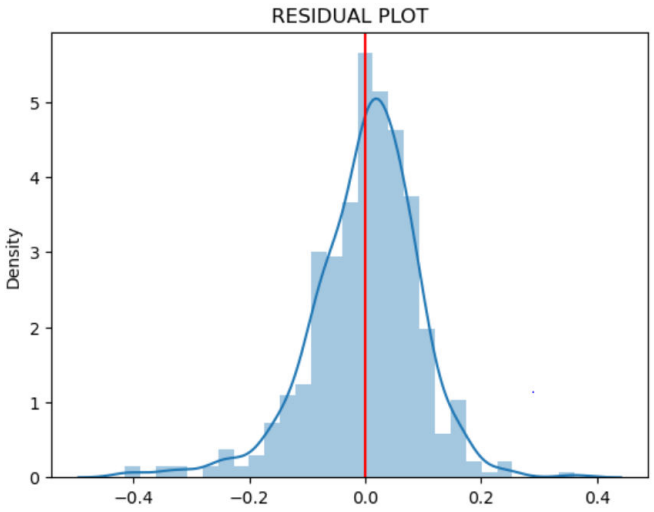
3.Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans. "atemp" and "temp" both have same correlation which is the highest among all numerical variables.

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

Ans. to validate the assumptions of linear Regressions-:

- The relationship between the independent variables and the target variable should be linear
- The independent variables should not be highly correlated with each other
- The differences between the predicted and actual values have constant variance (homoscedasticity)
- The independent variables should not be highly correlated with each other
- Residual analysis should follow normal distribution.
  This is done by plotting x-y plot of residual and further visualizing them



5.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 based on coefficient are the variables are “yr “ : Positive Coef , “temp” : Positive Coef and “weathersit_Light_Snow”: Negative Coef are the three features contributing significantly towards explaining the demand of the shared bike.

General Subjective Questions

1. **Explain the linear regression algorithm in detail?**

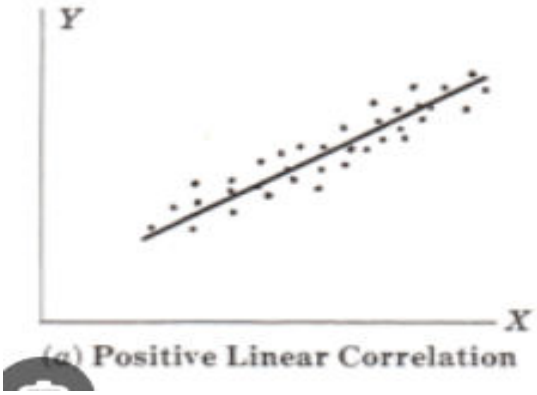
Ans. Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

The following is an example of a resulting linear regression equation:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots$$

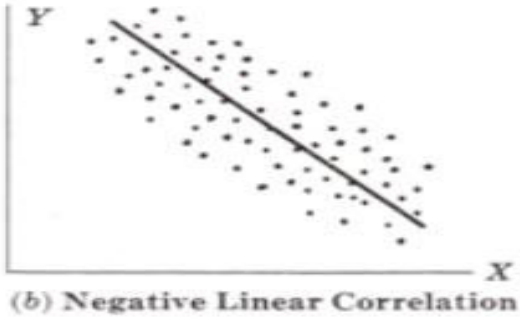
In the example above, y is the dependent variable, and x_1, x_2, and so on, are the explanatory variables. The coefficients (b_1, b_2, and so on) explain the correlation of the explanatory variables with the dependent variable. The sign of the coefficients (+/-) designates whether the variable is positively or negatively correlated. b_0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0.

After analyzing the data and identifying if there are any null values and cleaning the data, we can do appropriate analysis we will perform EDA on the data. After EDA we split the data into training data (which will be used to train a model) and test data (which will be used to check how close is our model to the actual output). As the model is prepared we will check for p value determine if the results are statistically significant or not and we will check the VIF for the magnitude of multicollinearity in the model, and dropping the variables accordingly till we get the perfect model. After that we do residual analysis check and check if the curve must be a normal curve. The conclusion drawn from the model will provide valuable insights/predictions of the data set.



(a) Positive Linear Correlation

In a positive linear relationship, as the value of one variable increases, the value of the other variable also increases in a linear manner. Let's consider an example where X represents the independent variable and Y represents the dependent variable.



(b) Negative Linear Correlation

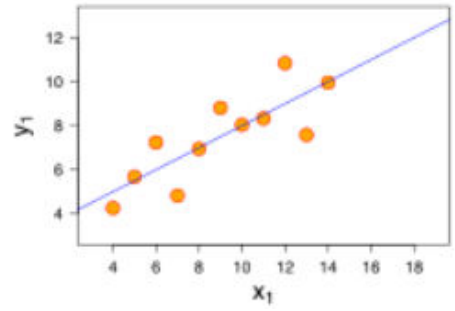
Negative Linear Relationship Graph: In a negative linear relationship, as the value of one variable increases, the value of the other variable decreases in a linear manner. Let's consider an example where X represents the independent variable and Y represents the dependent variable.

2.Explain the Anscombe’s quartet in detail.?

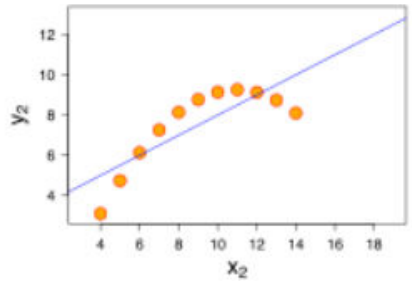
Ans. Anscombe's quartet is a fascinating example in statistics that demonstrates the importance of data visualization and the dangers of relying solely on summary statistics. It consists of four datasets, each containing 11 data points, which have nearly identical summary statistics but very different patterns when plotted. These datasets were created by the British statistician Francis Anscombe in 1973 to emphasize the need for visual exploration of data before drawing conclusions or making decisions based on summary statistics.

The four datasets can be described as:

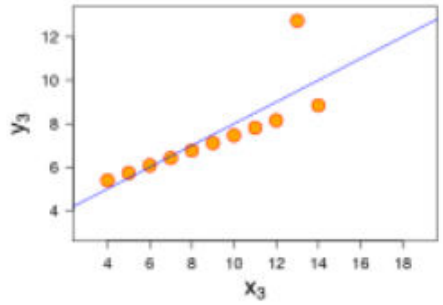
Dataset 1: This fits the linear regression model pretty well.



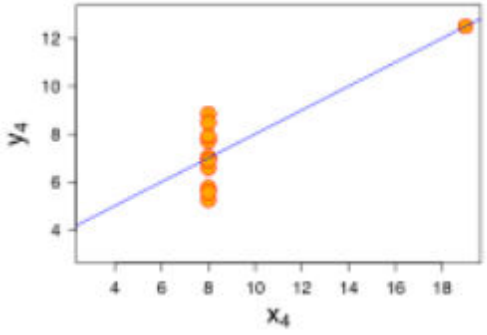
Dataset 2: This could not fit linear regression model on the data quite well as the data is nonlinear.



Dataset 3: Shows the outliers involved in the dataset which cannot be handled by linear regression model



Dataset 4: Shows the outliers involved in the dataset which cannot be handled by linear regression



3.What is Pearson's R?

Ans. Pearson's correlation coefficient, often denoted as Pearson's R or simply R, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It was developed by Karl Pearson and is widely used in statistics and data analysis.

Pearson's R is a value between -1 and 1. The interpretation of the coefficient is as follows:

- If R is close to +1, it indicates a strong positive linear relationship. As one variable increases, the other tends to increase as well.
- If R is close to -1, it indicates a strong negative linear relationship. As one variable increases, the other tends to decrease.
- If R is close to 0, it indicates a weak or no linear relationship. There is little to no linear association between the two variables.

$$R=\sum_{i=1n}(X_i-\bar{X})_2\sum_{i=1n}(Y_i-\bar{Y})_2\sum_{i=1n}(X_i-\bar{X})(Y_i-\bar{Y})$$

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

Ans. Scaling is a preprocessing step in data preparation that involves transforming the features of a dataset to a specific range or distribution. The purpose of scaling is to bring all the features on a similar scale, which can be important for certain machine learning algorithms and statistical techniques. Scaling ensures that no single feature dominates the model's learning process due to its larger magnitude or range.

Scaling is performed for:-

- Better convergence
- Dimensionality reduction
- Improved accuracy
- Regularization

There are 2 types of scaling NORMALIZED SCALING AND STANDARDIZATION SCALING

STANDARDIZATION SCALING:- In standardized scaling, the values of the features are transformed to have zero mean and unit variance. Standardized scaling is less sensitive to outliers since it is based on the mean and standard deviation, which are less affected by extreme values.

The formula for standardized scaling is - $X\text{standardized}=\sigma X-\mu$

NORMALIZED SCALING: Normalized scaling is sensitive to outliers, as it compresses the range of the data to a fixed interval, the values of the features are transformed to a specific range, usually between 0 and 1.

The formula for standardized scaling is - scaling is - $X\text{normalized}=X\text{max}-X\text{min}X-X\text{min}$

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

Ans. VIF stands for Variance Inflation Factor, VIF is infinity, because a large value of VIF indicates that there is correlation between variables, if there is perfect correlation then VIF = INFINITY. This variable must be dropped in order to successfully run the model. The VIF is calculated as VIF = 1 / (1 - R1^2)

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

Ans. A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It is a powerful visualization technique to compare the quantiles of the data to the quantiles of a specified theoretical distribution.

The use and importance are as follows:

Linear regression assumes that the errors (residuals) of the model are normally distributed with a mean of zero.
Checking the normality assumption is crucial because if the errors are not normally distributed, it may affect the validity of statistical inference and prediction in linear regression.
Outliers are data points that deviate significantly this helps in identifying outliers • This plot provides visual tool to evaluate the adequacy of the model by examining the residual distribution properties • This plot serves as a diagnostic tool in linear regression analysis .