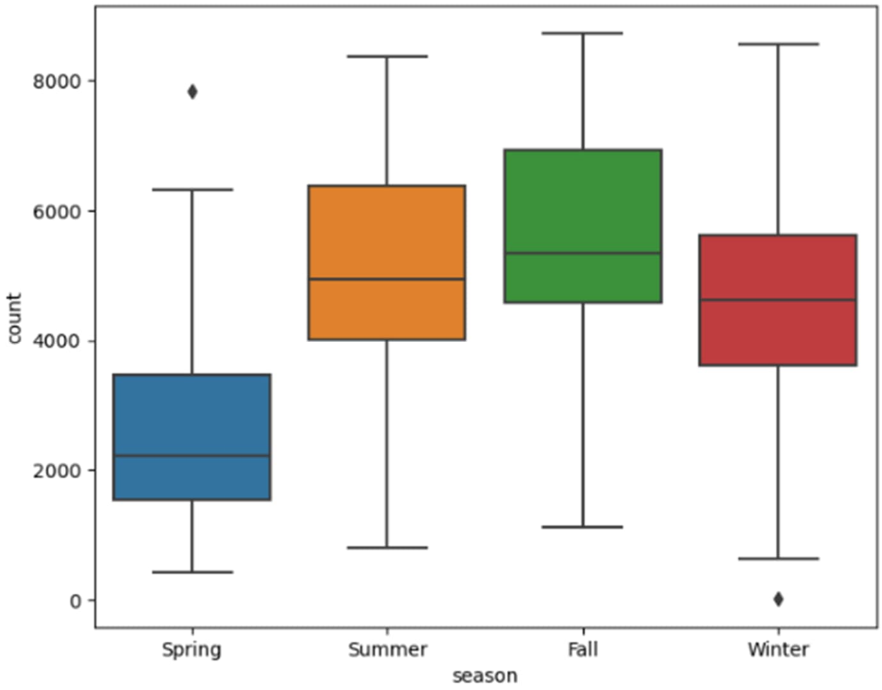
Linear Regression Assignment

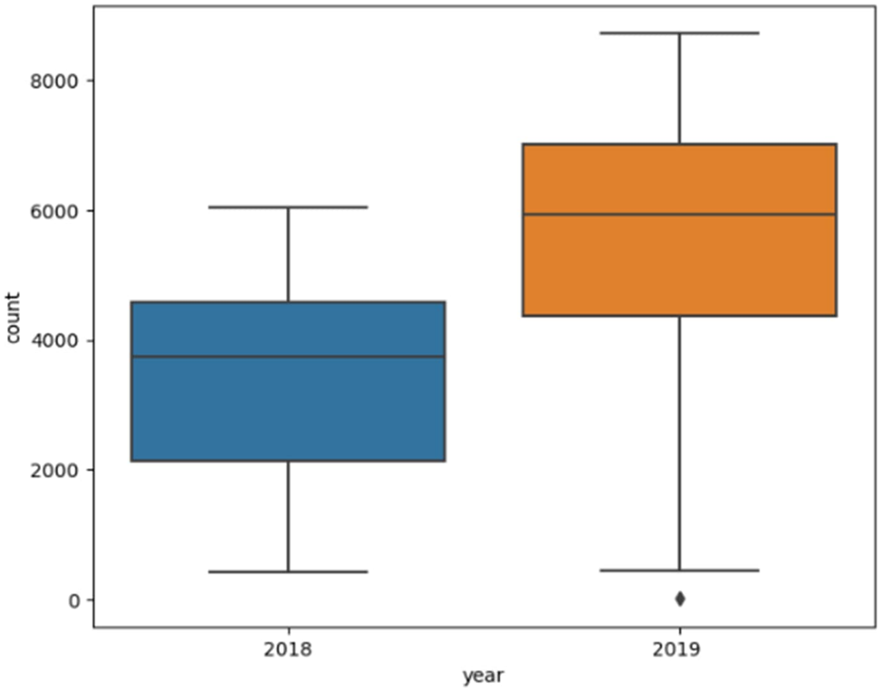
Assignment-based Subjective Questions

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

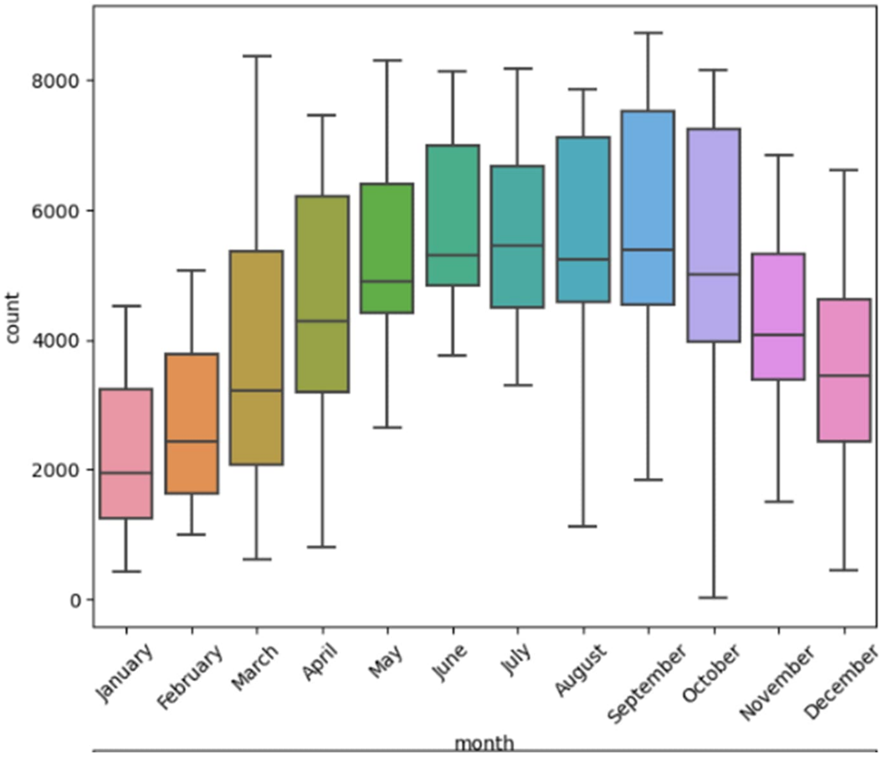
   After doing the data handling and cleaning, following were identified as the categorical variables in BoomBike dataset:

   - Season
   - Year
   - Month
   - Holiday
   - Weekday
   - Working day
   - Weather Situation

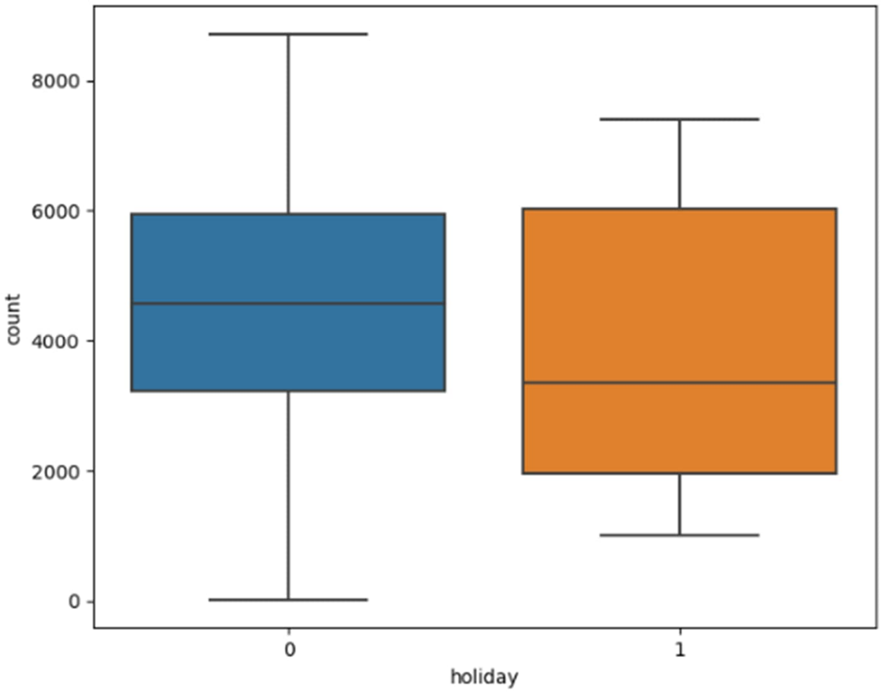




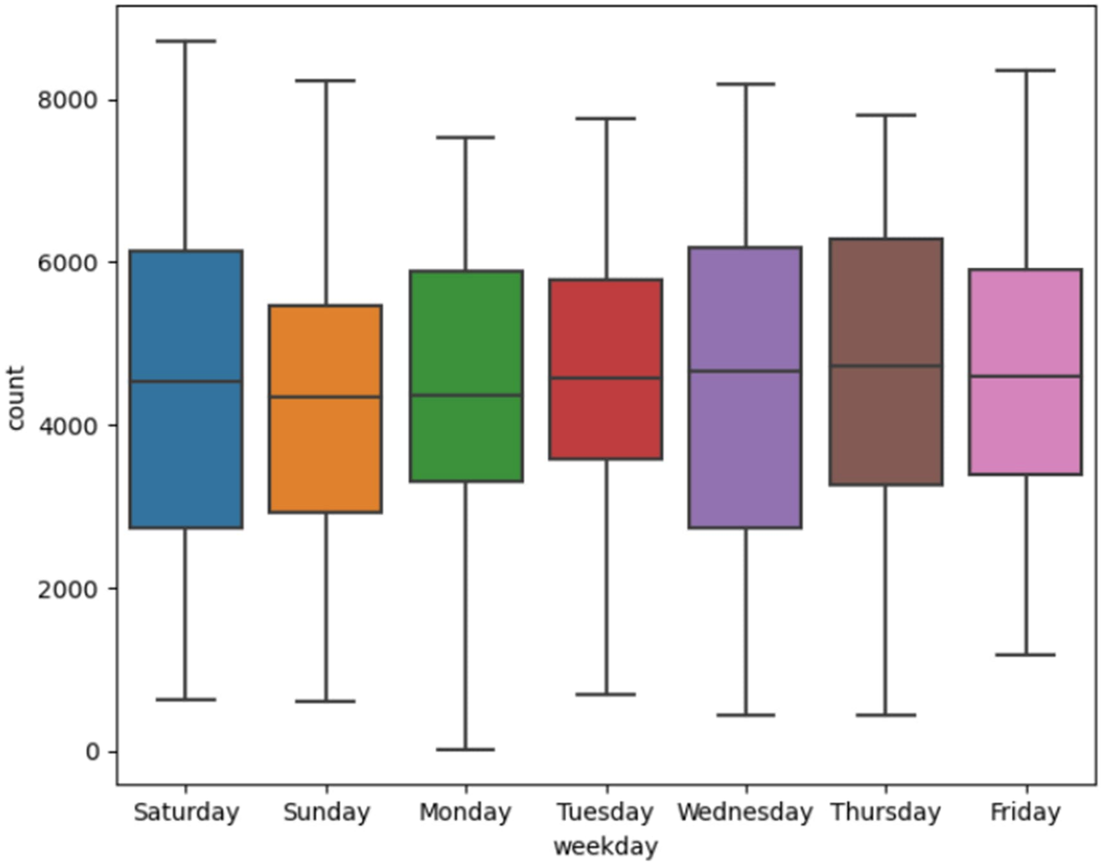
count
8000
6000
4000
2000
0
2018
2019
year



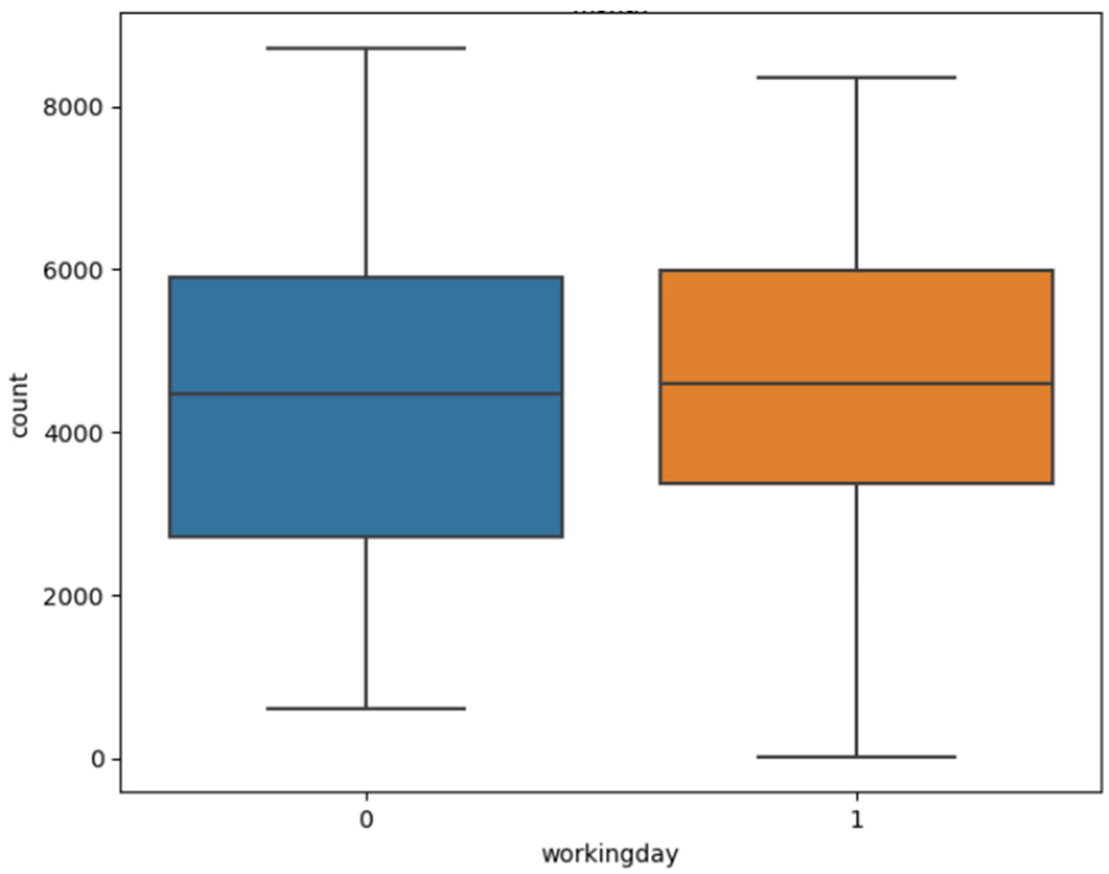
count
8000
6000
4000
2000
0
January
February
March
April
May
June
July
August
September
October
November
December
month

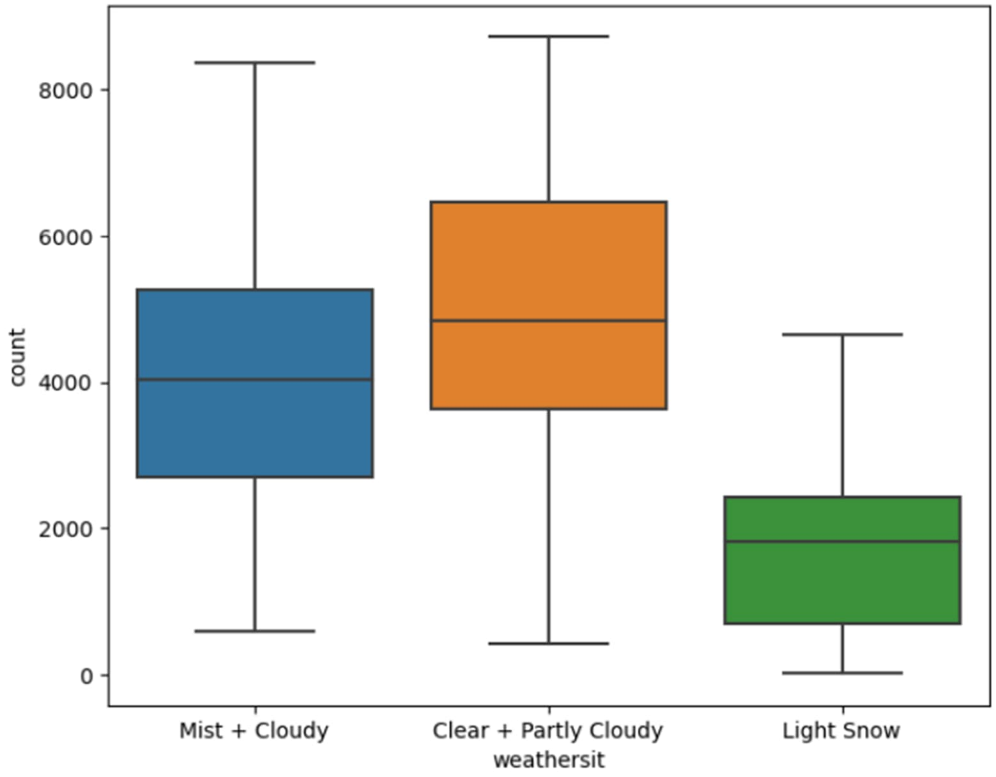


count
0
2000
4000
6000
8000
0
1
holiday



count
0
2000
4000
6000
8000
Saturday
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
weekday


count
8000
6000
4000
2000
0
0
1
workingday

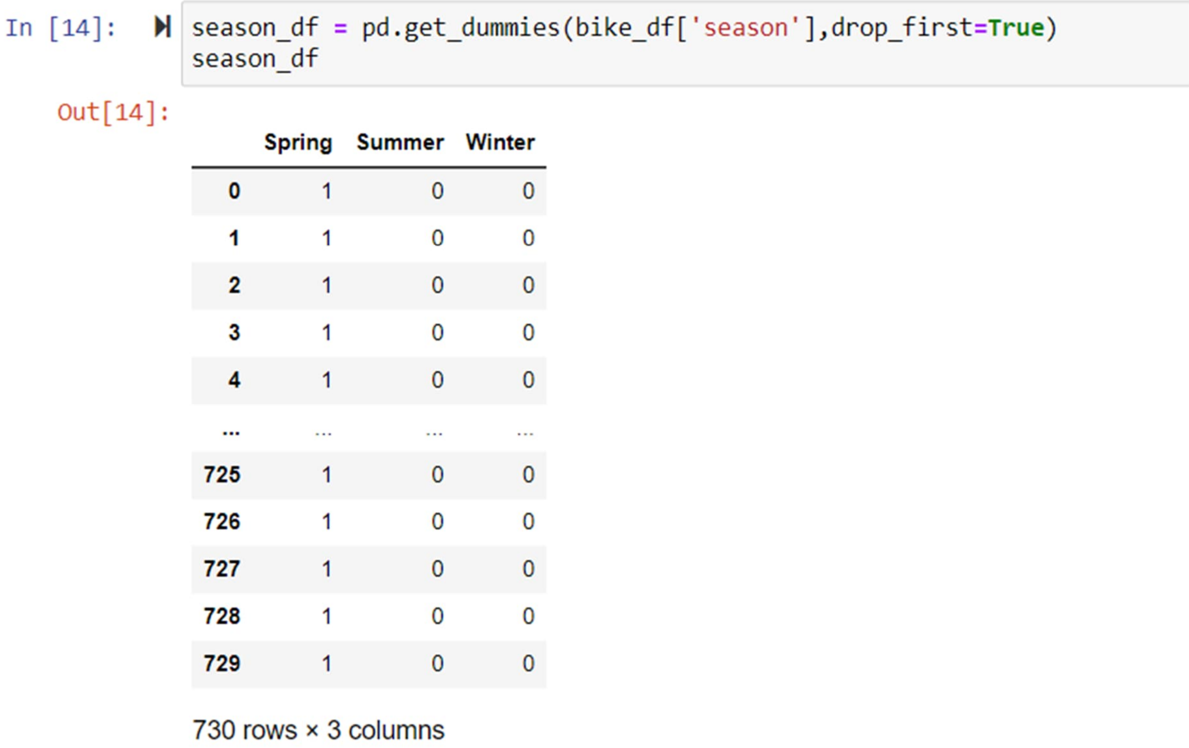

count
8000
6000
4000
2000
0
Mist + Cloudy
Clear + Partly Cloudy
Light Snow
weathersit

From the above plots we can infer that:

- In 'Fall' season, average bike rentals are highest, followed by summer and then winter.
- Year 2019 is having more bike rentals.
- Months from May to October are having average bike rentals ~ around 5k.
- There is not much difference in the mean bike rentals on weekday or weekend.
- Average bike rented on non-holiday days are more.
- There are more bikes rented when weather is clear and partly cloudy, followed by misty and cloudy weather.

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

drop_first = True drops the first dummy variable, it will give n-1 dummy variables out of n unique categorical levels.
for e.g., if we have a column Season with 4 values i.e., spring, summer, fall, winter. So, if we create dummy variables for this column season then it will create 3 columns instead of 4 as shown below.



```
In [14]: season_df = pd.get_dummies(bike_df['season'],drop_first=True)
         season_df
```

Out[14]:

	Spring	Summer	Winter
0	1	0	0
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0
...	...	...	...
725	1	0	0
726	1	0	0
727	1	0	0
728	1	0	0
729	1	0	0

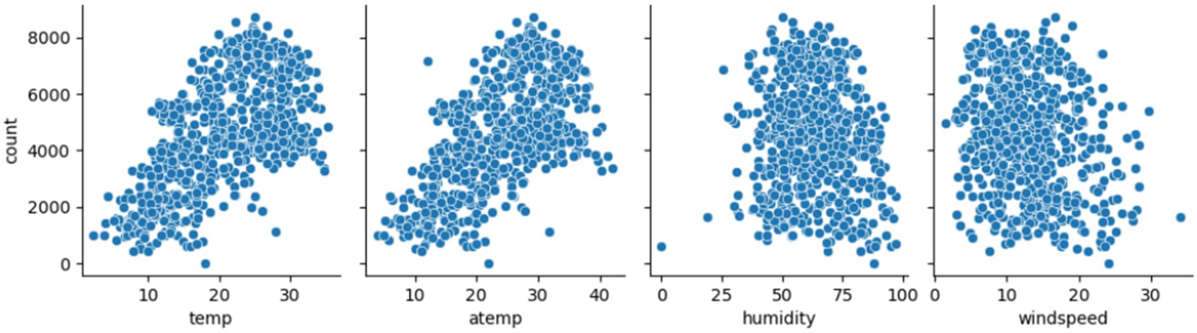
730 rows × 3 columns

Now when values of all these 3 are 0 then it means that season is ‘fall’.

It is important use drop_first= true as it helps in reducing the number of columns during the creation of dummy variable thus reducing the correlations among the dummies and avoiding the Dummy Variable Trap.

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

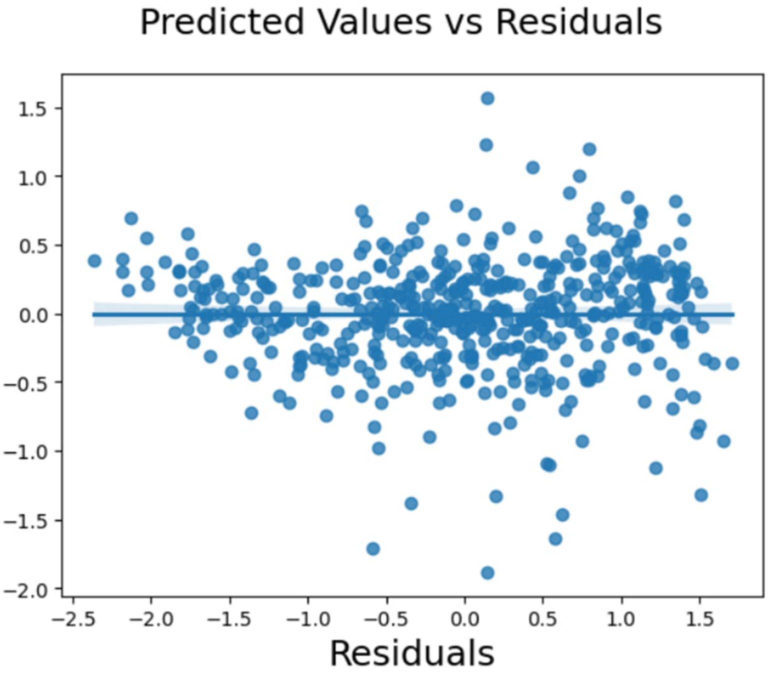
   ‘temp’ and ‘atemp’ has the highest correlation with target variable ‘count’ which can be clear from the below screenshot.



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

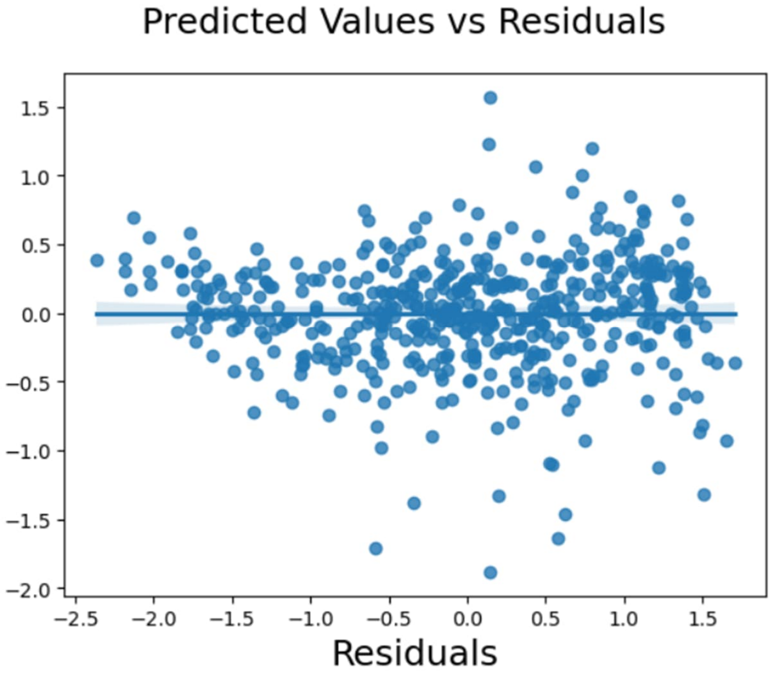
   Linear regression Assumptions: Linear regression relies on several assumptions, which are as follows:

   a. Linearity: The relationship between the variables is assumed to be linear. When the predicted values are plotted against the residuals.



There are equally spread residuals around a horizontal line without distinct patterns are a good indication of having the linear relationships thus satisfying the assumption of linearity.

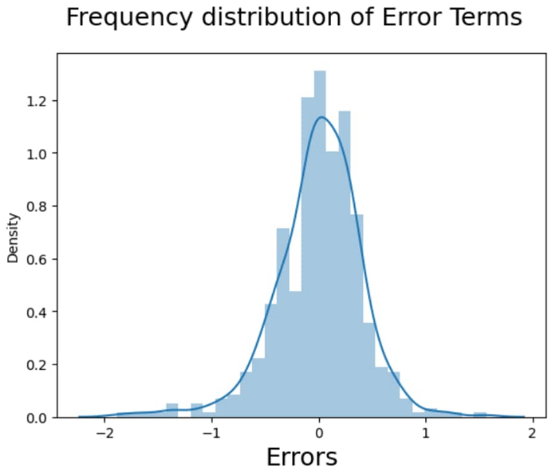
b. Independence of residuals: Observations are assumed to be independent of each other.



Same plot shows that residual errors are independent of each other since the Predicted values vs Residuals plot doesn't show any trend. This satisfies the second assumption.

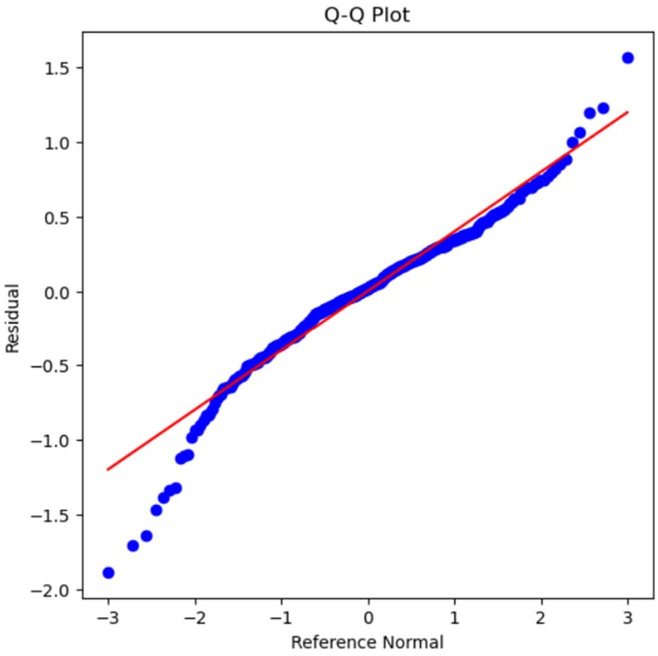
c. Normality: The error terms should follow a normal distribution.

On plotting the histogram for residuals or error terms, following is observed.



Here error terms are normally distributed.

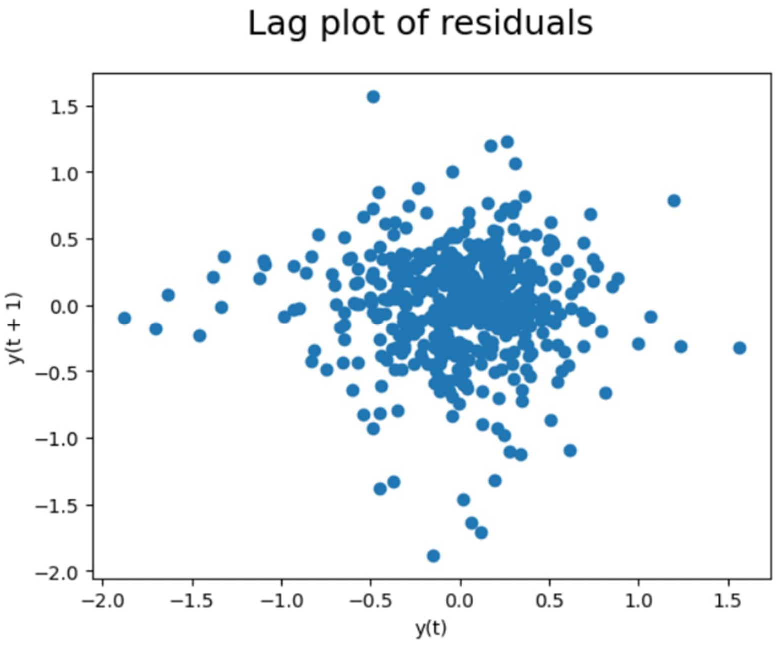
Another way, it was proved using Q-Q plot.



This plot further shows that the residual distribution is approximately normal for all test data with values within range of training data.

d. Homoscedasticity: The variability of the error terms should be constant across all levels of the independent variable.

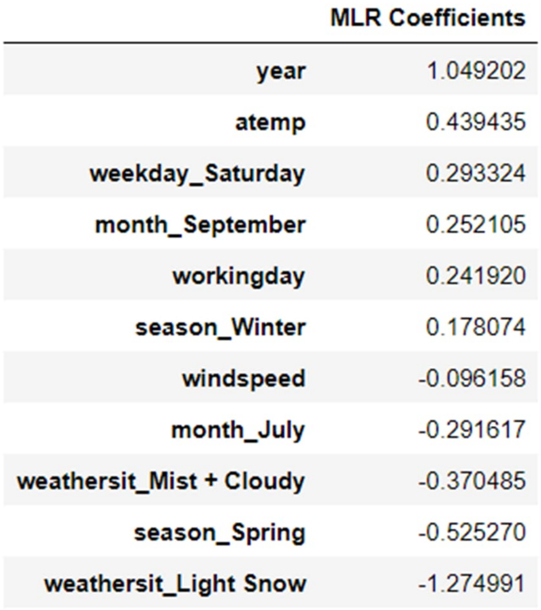
This is proved by Lag plot.



The Residual Lag Plot constructed by plotting residual y(t) against residual y(t+1), is useful for examining the dependency of the error terms. Any non-random pattern in a lag plot suggests that the variance is not random.
Thus satisfying assumption of Linear Regression i.e., Homoscedasticity

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

Following are the list to variables and their corresponding coefficients.



	MLR Coefficients
year	1.049202
atemp	0.439435
weekday_Saturday	0.293324
month_September	0.252105
workingday	0.241920
season_Winter	0.178074
windspeed	-0.096158
month_July	-0.291617
weathersit_Mist + Cloudy	-0.370485
season_Spring	-0.525270
weathersit_Light Snow	-1.274991

From this it is clear, top 3 influencing factors are:

- weathersit_Light Snow
- year
- season_spring

General Subjective Questions

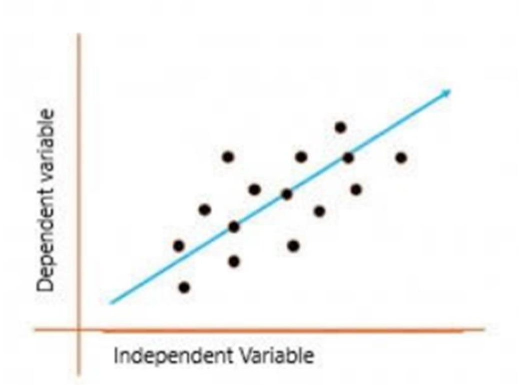
1. Explain the linear regression algorithm in detail.

Linear regression is a fundamental machine learning algorithm used for predicting a continuous outcome variable (also called the dependent variable) based on one or more input features (also called independent variables or predictors). It assumes a linear relationship between the input features and the target variable. In simpler terms, linear regression tries to find the best-fitting straight line through the data points to make predictions.

Here's a detailed explanation of how linear regression works:

Simple Linear Regression: In this case, we have one input feature (X) and one target variable (Y) i.e., there is one independent variable and one dependent variable. The model estimates the slope and intercept of the line of best fit, which represents the relationship between the variables.

The slope represents the change in the dependent variable for each unit change in the independent variable, while the intercept represents the predicted value of the dependent variable when the independent variable is zero.



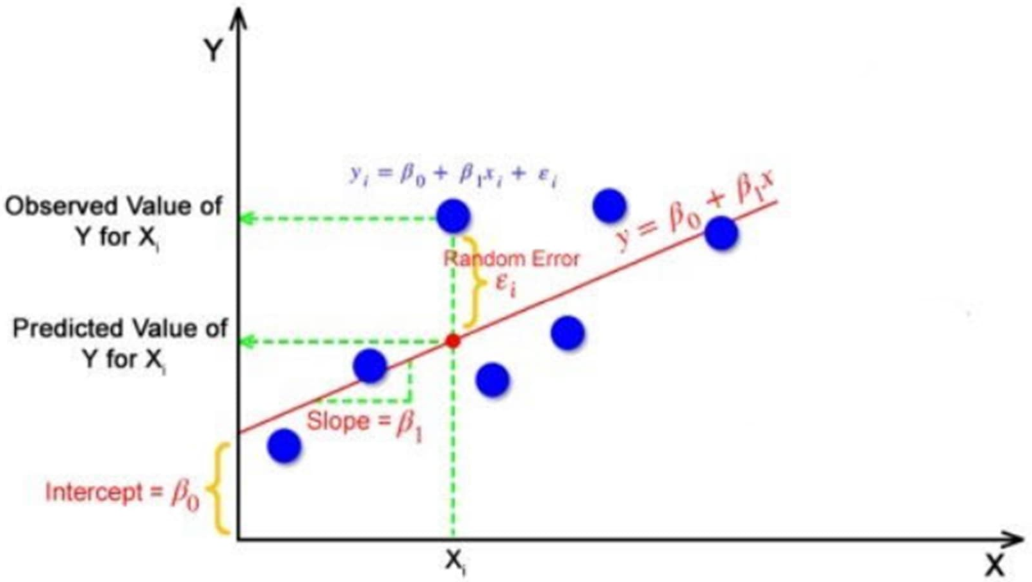
The above graph presents the linear relationship between the output(y) variable and predictor(X) variables. The blue line is referred to as the best fit straight line. Based on the given data points, a line is plotted that fits the points the best.

The equation of a straight line is generally given by: **Yi = Bo +B1Xi**

Where:

- Yi = Dependent variable,
- β0 = constant/Intercept,
- β1 = Slope/Intercept,

- Xi = Independent variable.


Y
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
Observed Value of Y for X_i
$y = \beta_0 + \beta_1 x$
Random Error ε_i
Predicted Value of Y for X_i
Slope = β_1
Intercept = β_0
X_i
X

The goal of linear regression is to find the best values for B0 and B1 that minimize the difference between the actual target values and the predicted values (error or residual).

Random Error (Residuals): In regression, the difference between the observed value of the dependent variable (Yi) and the predicted value (Ypred) is called the residuals.

Ei = Ypred – Yi

Where Ypred = B0 + B1 Xi

Multiple Linear Regression: In many real-world scenarios, there are multiple input features that can influence the target variable. In multiple linear regression, the equation becomes a bit more complex:

Y = b0 + b1X1 + b2X2 + ... + bn*Xn

Where:

- Y is the predicted value (target variable).
- X1, X2, ..., Xn are the individual input features.
- b0 is the y-intercept.
- b1, b2, ..., bn are the coefficients associated with each input feature.

Again, the goal is to find the best values for the coefficients that minimize the overall difference between the actual target values and the predicted values.

Training the Model: To train a linear regression model, we need a dataset containing both input features and target values. The model adjusts the coefficients during training to minimize a certain error metric, typically the Mean Squared Error (MSE) or Mean Absolute Error (MAE). These metrics measure the difference between the predicted values and the actual values.

Gradient Descent (Optional): Gradient descent is an optimization technique often used to find the optimal coefficients for linear regression. It iteratively adjusts the coefficients in the direction that reduces the error. The algorithm calculates the gradient of the error function with respect to the coefficients and updates them accordingly.

Model Evaluation Metrics: The strength of any linear regression model can be assessed using various evaluation metrics. These evaluation metrics usually provide a measure of how well the observed outputs are being generated by the model.

The most used metrics are,

- Coefficient of Determination or R-Squared (R2): It measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

  Mathematically it can be represented as

$$R^2 = 1 - (RSS/TSS)$$

- Root Mean Squared Error (RSME): It calculates the average of the squared differences between the predicted and actual values. A lower MSE signifies better model performance.

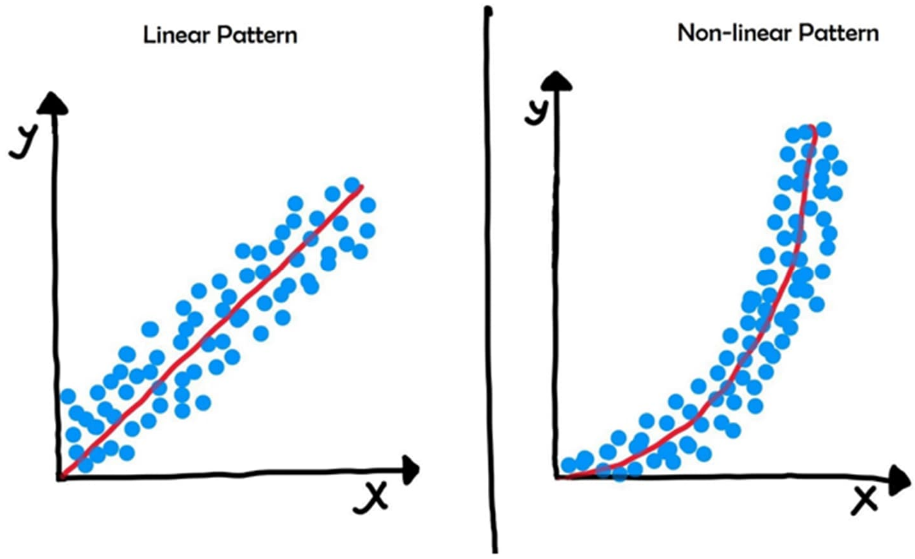
  It is represented as follows:

$$RMSE = \sqrt{\frac{RSS}{n}} = \sqrt{\sum_{i=1}^{n}\left(y_i^{Actual} - y_i^{Predicted}\right)^2 \Big/ n}$$

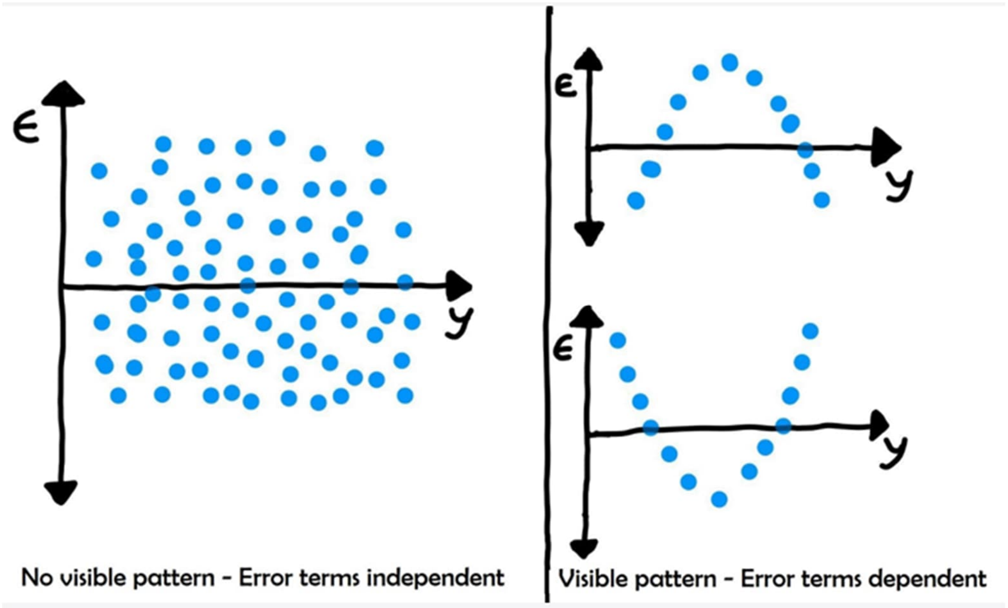
Making Predictions: Once the model is trained, we can use it to make predictions on unseen data or test data. Simply plug in the input feature values into the equation, and the model will provide a predicted value for the target variable.

Linear regression Assumptions: Linear regression relies on several assumptions, which are:

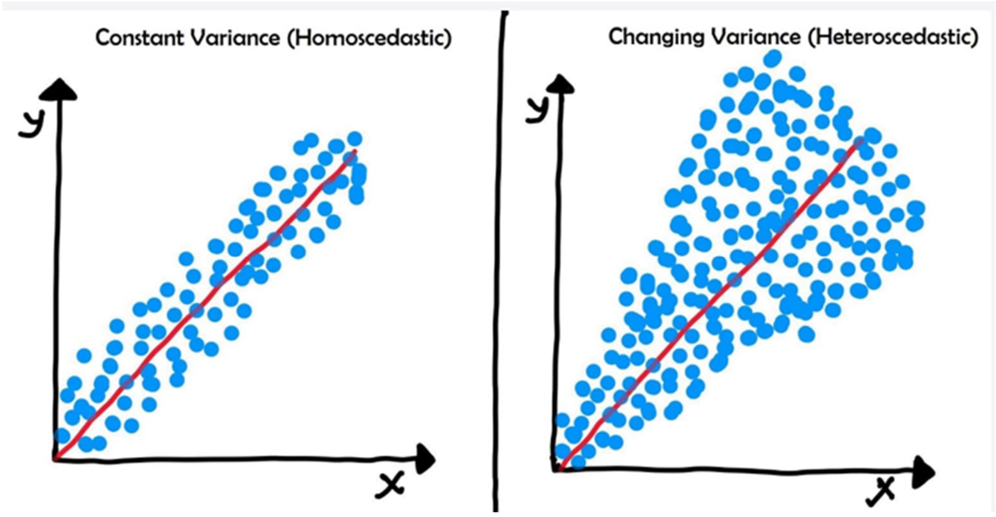
- Linearity: The relationship between the variables is assumed to be linear.



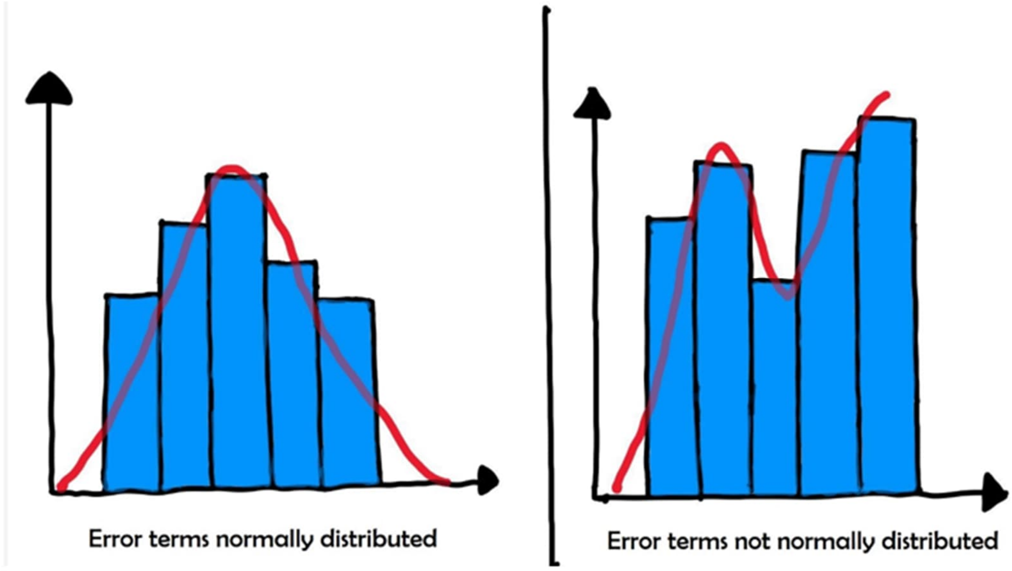
- Independence of residuals: Observations are assumed to be independent of each other.



- Homoscedasticity: The variability of the error terms should be constant across all levels of the independent variable.



- Normality: The error terms should follow a normal distribution.



2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a set of four small datasets that have nearly identical statistical properties, yet they exhibit vastly different graphical representations and relationships between variables.
The quartet was created by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and not relying solely on summary statistics. It serves as a powerful illustration of the limitations of using only statistical measures to describe a dataset.

Here are the details of the four datasets in Anscombe's quartet:

Dataset I:

- X values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
- Y values: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82

Dataset II:

- X values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
- Y values: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26

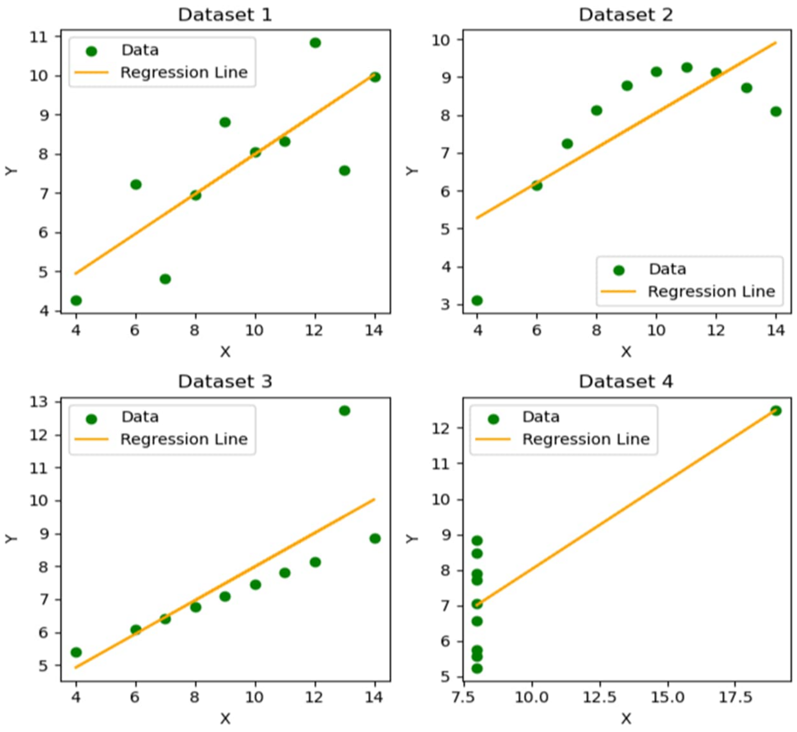
Dataset III:

- X values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
- Y values: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42

Dataset IV:

- X values: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8
- Y values: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91

Despite having nearly identical summary statistics (means, variances, correlations), these datasets have very different patterns when plotted:



- Dataset I resemble a linear relationship.
- Dataset II has a linear relationship but with an outlier.

- Dataset III is non-linear and has a strong influence from one outlier.
- Dataset IV has a weak correlation but is heavily influenced by an outlier.

Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

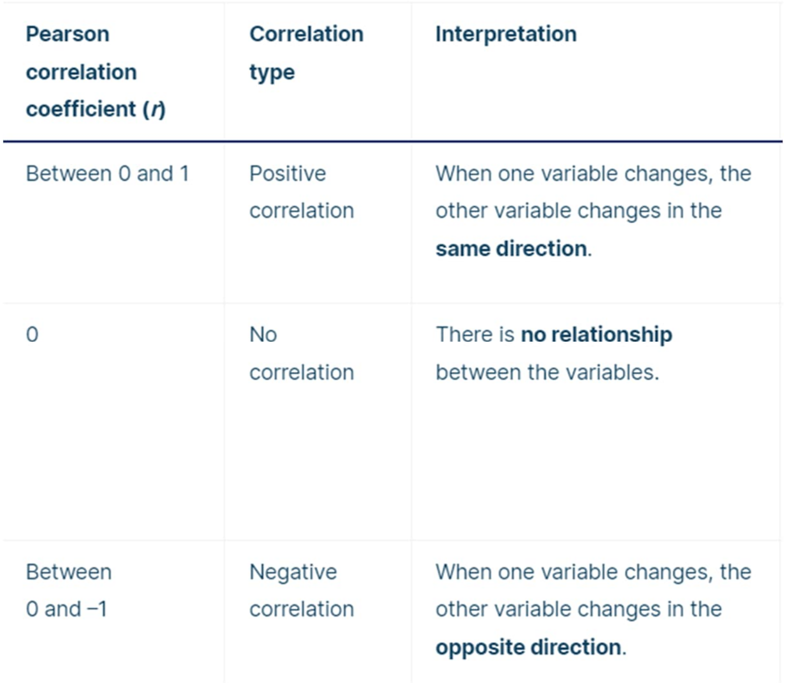
This illustration serves as a reminder to researchers, analysts, and anyone working with data that a comprehensive analysis involves both statistical techniques and data visualization to gain a deeper understanding of the underlying patterns and relationships in the data.

3. What is Pearson’s R?

Pearson's correlation coefficient, often denoted as r or Pearson's r. It is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables.

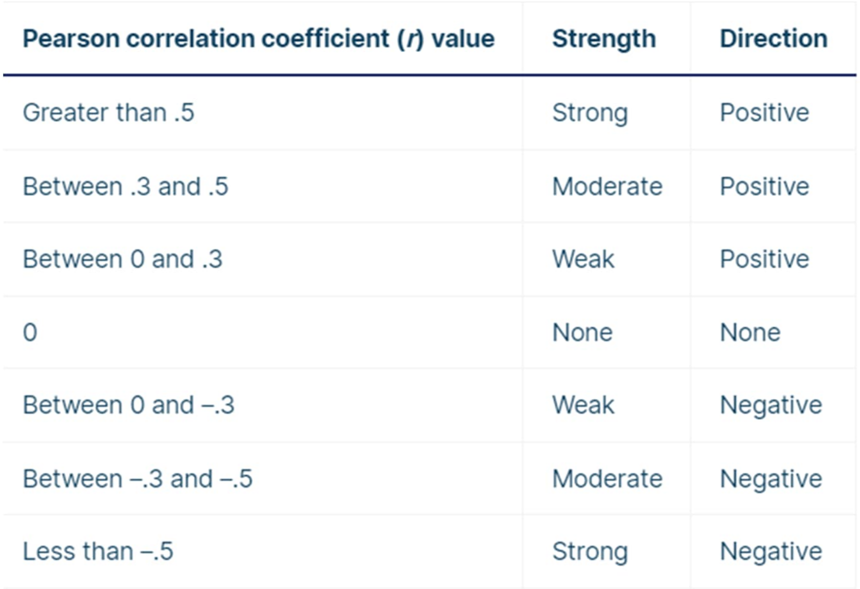
It is used to assess how closely the data points of two variables cluster around a straight line, indicating the degree of correlation between them.

The Pearson correlation coefficient ranges from -1 to 1:

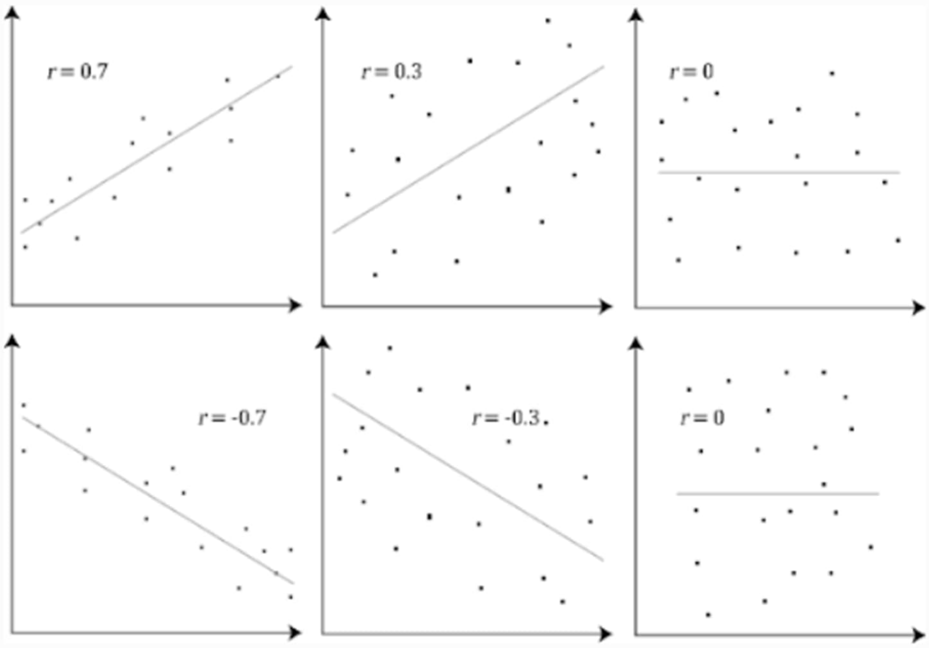


Pearson correlation coefficient (*r*)	**Correlation type**	**Interpretation**
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.
0	No correlation	There is **no relationship** between the variables.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.

Although interpretations of the relationship strength vary between disciplines, the table below gives the rules of thumb:

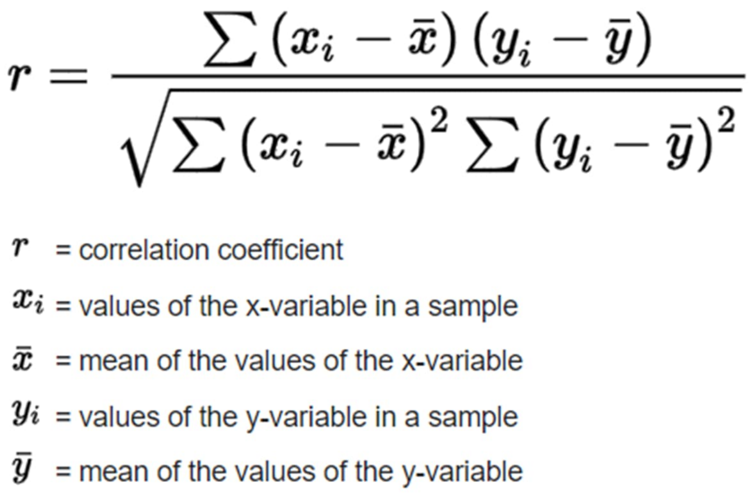


Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative



Different correlation values

The formula to calculate Pearson's correlation coefficient r between two variables X and Y is as follows:



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Pearson's correlation coefficient has a few important characteristics:

1. Symmetry: The correlation between X and Y is the same as the correlation between Y and X, i.e., corr(X, Y) = corr(Y,X)
2. No Units: r is a unitless value, which means it doesn't depend on the units of measurement of the variables.
3. Sensitive to Linear Relationships: r specifically measures linear relationships. If the relationship between variables is nonlinear, Pearson's correlation might not fully capture the association.
4. Outliers: Outliers can heavily influence the correlation coefficient, potentially making it appear stronger or weaker than it is.
5. Range: The range of the correlation coefficient is -1 to 1, with 0 indicating no linear correlation. However, the strength of the correlation depends on how close r is to -1 or 1.

Pearson's correlation coefficient is widely used in various fields, including economics, social sciences, biology, and more, to quantify the relationship between variables and assess the degree of correlation.

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

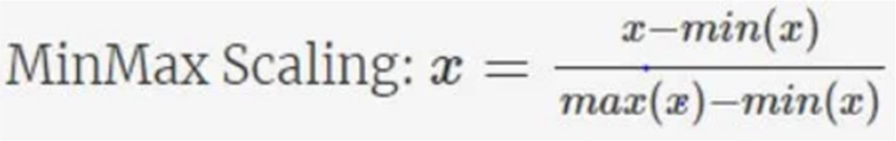
Scaling is a pre-processing technique used in data analysis and machine learning to transform the features (variables) of a dataset so that they are on a similar scale. Scaling is performed to ensure that all features contribute equally to the analysis and to prevent certain algorithms from being dominated by features with larger values.

Scaling is necessary for several reasons:

- Algorithm Sensitivity: Many machine learning algorithms are sensitive to the scale of features. If one feature has a much larger scale than others, it can disproportionately influence the outcome of algorithms.
- Distance Metrics: Algorithms that rely on distance metrics, like k-nearest neighbours and hierarchical clustering, can be affected by features with different scales.
- Gradient Descent: Algorithms that use gradient descent for optimization, like neural networks and some regression techniques, can converge faster with scaled features.
- Regularization: Some regularization techniques assume features are on similar scales, so scaling aids in proper regularization.
- Interpretability: Scaling ensures that features are comparable, aiding in interpretation.

Normalized Scaling: Normalized scaling (also known as min-max scaling) transforms the features to a specific range, usually between 0 and 1.

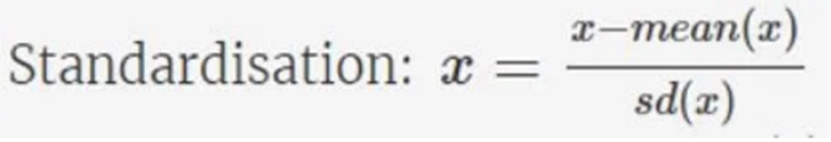
The formula for normalized scaling is:



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

Standardized Scaling: Standardized scaling (also known as z-score normalization) transforms features, so they have a mean of 0 and a standard deviation of 1.

The formula for standardized scaling is:



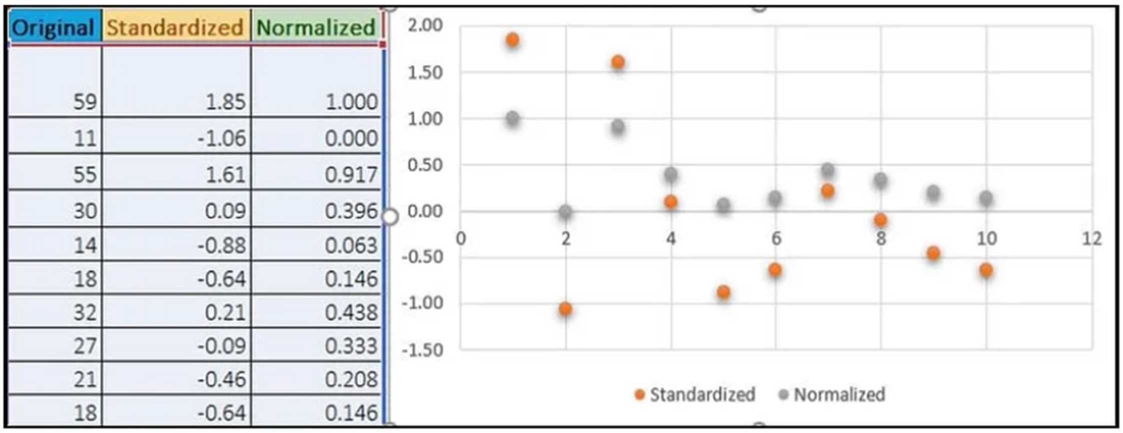
$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

Difference between Normalized Scaling and Standardized Scaling:

Parameters	Normalized Scaling	Standardized Scaling
Range	Scales feature to a specific range (e.g., 0 to 1)	Centres feature around zero with a standard deviation of 1.

Sensitive to Outliers	Can be sensitive to outliers, as the range is influenced by extreme values	More robust to outliers.
Interpretability	Retains the original distribution's shape	May change the shape due to centering and scaling
Usage	when you have a reason to believe the data should be constrained within a certain range	where the distribution's shape matters less, and you want features with similar mean and variance
Another name	It is also known as Scaling Normalization	It is also known as Z-Score
Data distribution	It is applied when we are not sure about the data distribution	It is used when the data is Gaussian or normally distributed

Below shows example of Standardized and Normalized scaling on original values.



Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146

In summary, scaling is crucial to ensure features are comparable and that machine learning algorithms work effectively. The choice between normalized and standardized scaling depends on the specific characteristics of your data and the requirements of the algorithm you're using.

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

The Variance Inflation Factor (VIF) is a metric used to assess multicollinearity in regression analysis. It quantifies how much the variance of the estimated regression coefficients is increased due to multicollinearity among the predictor variables.

A high VIF value suggests a high degree of multicollinearity, which can lead to unstable and unreliable regression coefficient estimates.

A situation where the VIF value becomes infinite can occur when perfect multicollinearity is present among the predictor variables. Perfect multicollinearity happens when one predictor variable is a perfect linear combination of one or more other predictor variables.

In other words, one variable can be exactly predicted using a combination of other variables.

Example:

Suppose we have a dataset with three variables: A1, A2, and B. However, A2 is a perfect linear combination of A1, which means A2 can be exactly predicted using A1.

Here's the data:

A1: 1, 2, 3, 4, 5
A2: 2, 4, 6, 8, 10
Y: 3, 5, 7, 9, 11

In this case, A2 = 2 x A1, which means the relationship between A2 and A1 is perfectly linear.

Let's calculate the VIF for A2 to see what happens:

The formula for VIF is:

$$\text{VIF} = \frac{1}{1 - R^2}$$

For A2:

- R =1, because A2 is perfectly correlated with 1X1.
- 1 - 1^2 = 1−1=0

So, the VIF calculation for A2 becomes: 1/0 = Infinity

Dividing by zero results in an infinite value. This indicates that there is perfect multicollinearity between A2 and A1.

In practice, software tools like Python's stats models or scikit-learn will often handle these situations gracefully, providing warnings about the multicollinearity issue and suggesting solutions such as removing one of the perfectly correlated variables before fitting a regression model.

Infinite VIF values are a clear indicator of a severe multicollinearity issue due to perfect linear relationships between predictor variables. This situation is problematic because it can cause numerical instability in regression calculations, making it difficult to obtain reliable coefficient estimates.

It's important to identify and address perfect multicollinearity by either removing one of the perfectly correlated variables or rethinking the model structure.

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

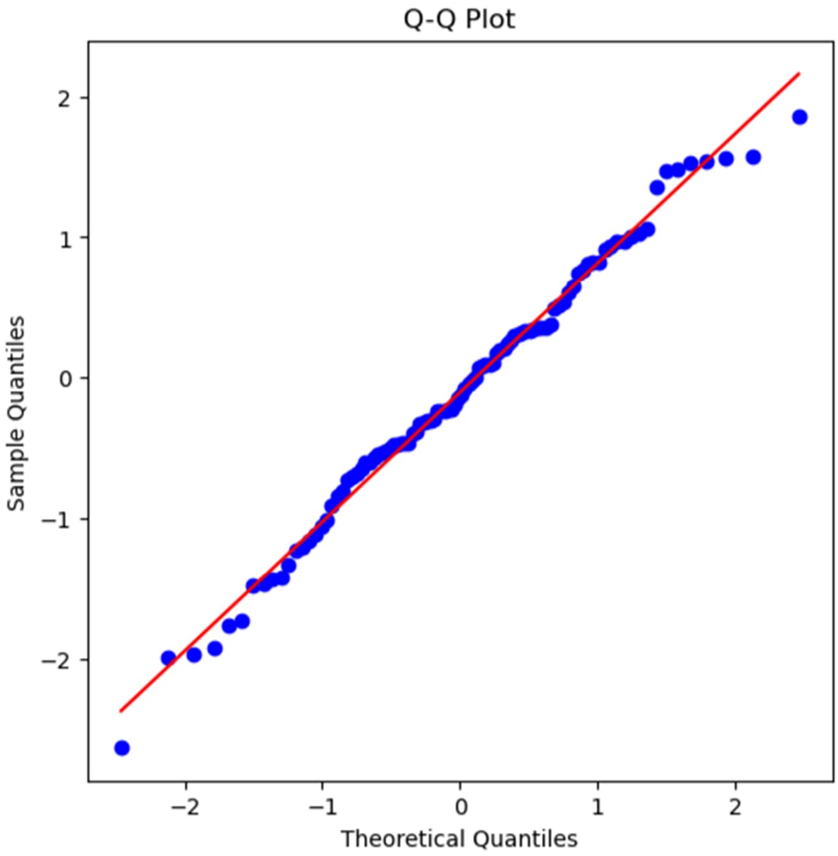
Q-Q (Quantile-Quantile) plot is a graphical tool used in statistics to assess whether a given dataset follows a particular theoretical distribution. It's especially useful for checking the normality assumption of a dataset, which is important in various statistical analyses, including regression.

How Q-Q Plot Works: A Q-Q plot compares the quantiles of the observed data against the quantiles of a theoretical distribution (often the normal distribution). If the dataset follows the theoretical distribution closely, the points on the Q-Q plot will fall approximately along a straight line. Deviations from a straight line indicate departures from the assumed distribution.

Why Q-Q Plot is Important in Regression:

- Checking Normality Assumption: In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed. If the residuals are not normally distributed, it can affect the validity of statistical tests, confidence intervals, and hypothesis tests associated with the regression model. A Q-Q plot of the residuals can help you assess whether they follow a normal distribution.
- Validating Regression Assumptions: Linear regression relies on several assumptions, including linearity, homoscedasticity (constant variance of residuals), and normality of residuals. A Q-Q plot can help you evaluate the assumption of normality, which is essential for reliable inference and hypothesis testing.
- Identifying Outliers: Outliers can distort the normality assumption and influence regression results. A Q-Q plot can reveal whether there are outliers in the data that deviate significantly from the expected distribution, helping you identify potential influential points.
- Model Improvement: If the Q-Q plot indicates departures from normality, it might suggest the need for data transformation or the use of more advanced regression techniques that can handle non-normal data.
- Interpreting Regression Results: Normality of residuals is important when interpreting the significance of coefficients, p-values, and confidence intervals in

a regression model. Departures from normality can lead to incorrect conclusions about the relationship between variables.



In the above example:

- 100 random data points generated from a normal distribution with a mean (loc) of 0 and standard deviation (scale) of 1.
- Here plotted data is compared against theoretical normal distribution.
- The Q-Q plot is telling how well the data's quantiles match the quantiles of the normal distribution.

If the data points fall approximately along a straight line, it suggests that the data is normally distributed. If the points deviate significantly from a straight line, it indicates departures from normality.