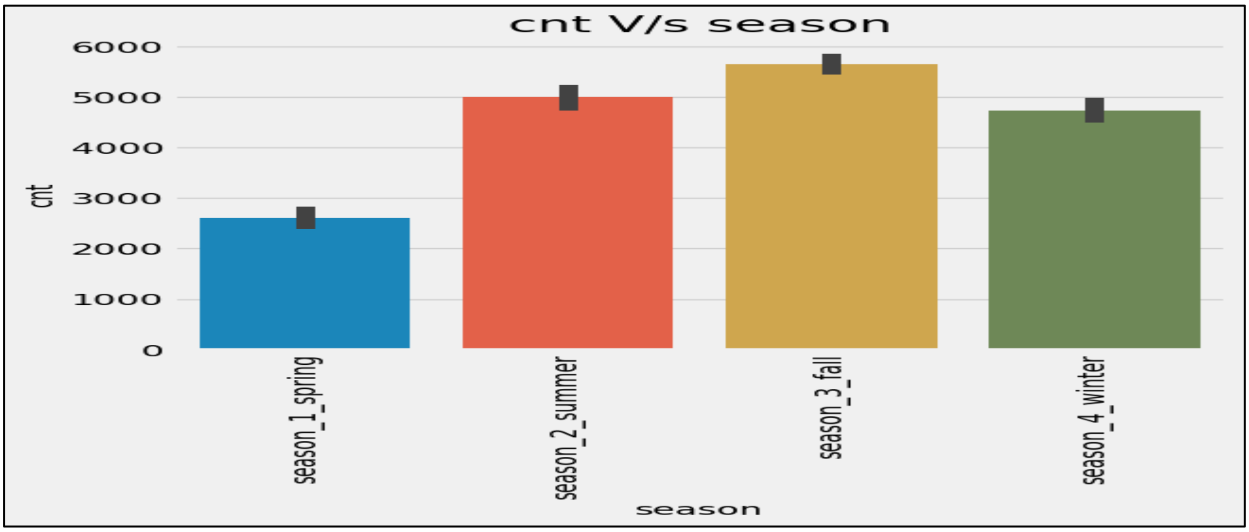
Assignment-based Subjective Questions

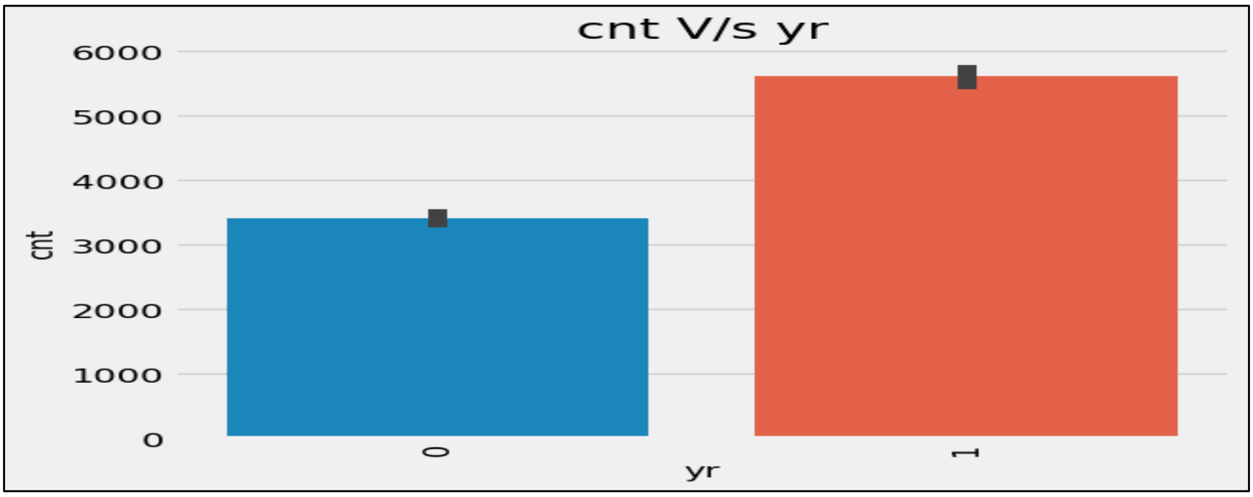
Question 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer 1:

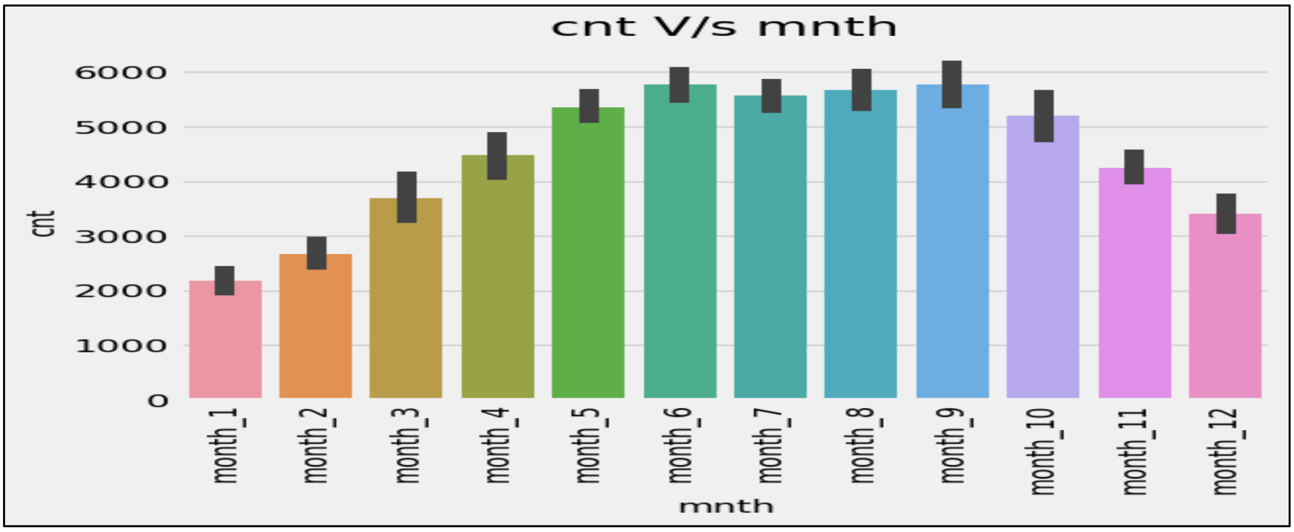
1. **Season:** Season 3/Fall has the highest count of shared bikes and season 1 /spring lowest.



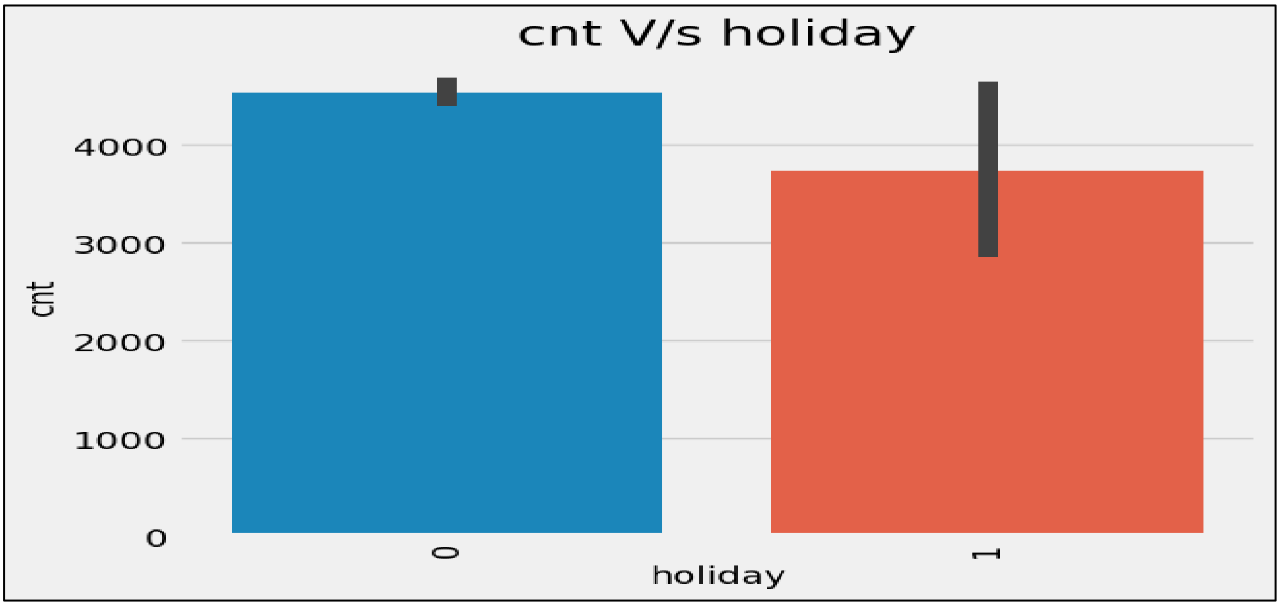
2. **Yr**: Year 2019 has more no. bikes hired compared to 2018. There seems to be an increasing trend of no. of bikes hired with each year and it shows increasing popularity of bike sharing amongst US population.



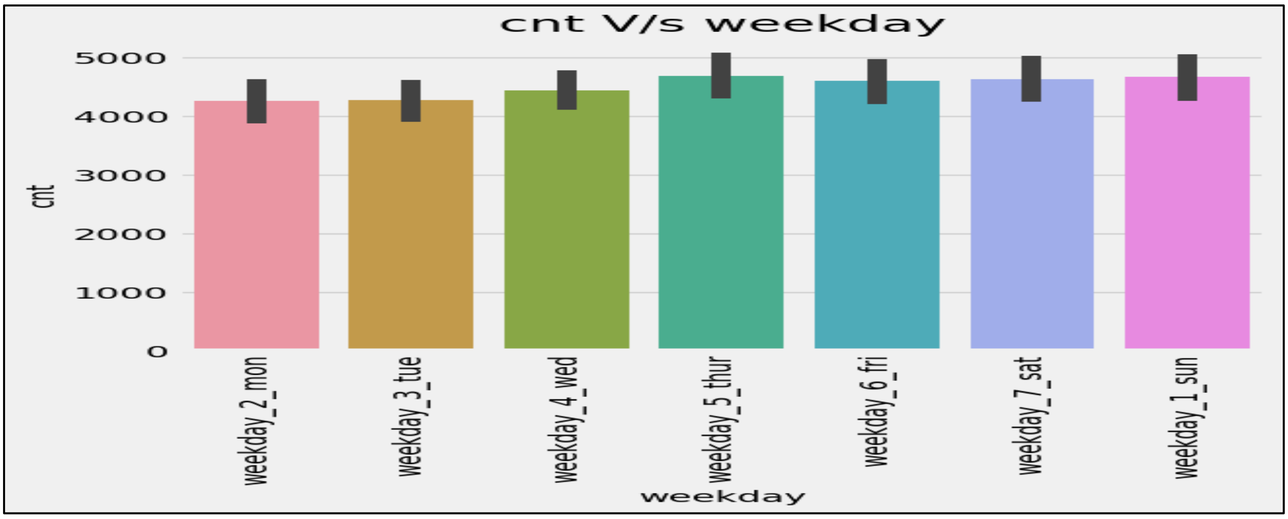
3. **Month:** Month-9 i.e. September month has highest demand of bikes and month-1 i.e. January has lowest. There is an increasing trend from month 1-/January, Most. no. of the bikes seem to be hired from month 5 to month 10 i.e. from may, to october and after September the trend seems to be declining.



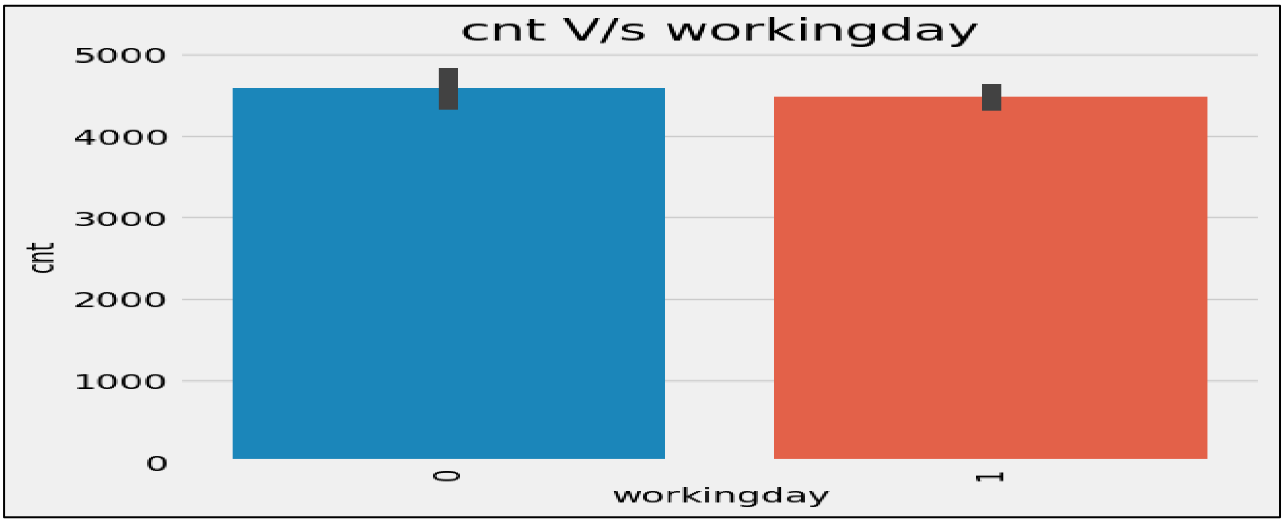
4. **Holiday:** More no. of bikes are hired in non-holidays . This may be because on holidays people prefer staying home and spend family time.



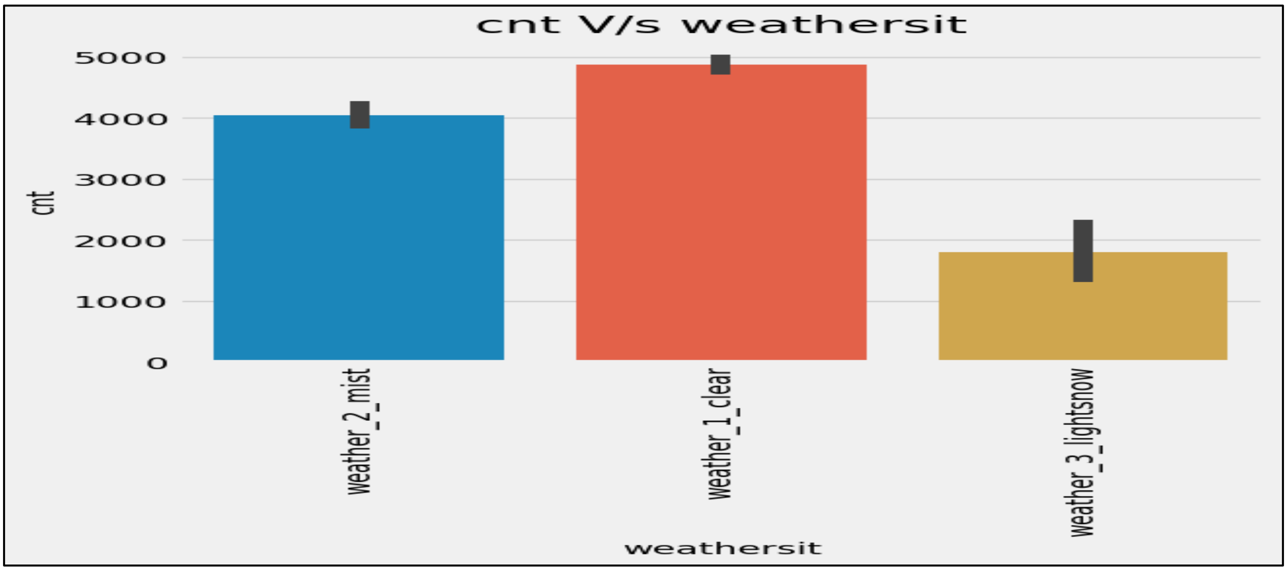
5. **Weekday:** More no. of bikes are hired on Thursday and Sunday and least on Monday and Tuesday



6. **Workingday:** Demand of bikes is approximately same for working and non-working day. Workingday seems to have not much impact on demand of shared bikes.



7. **Weathersit:** Weathersit 1 i.e. Clear, Few clouds, Partly cloudy, Partly cloudy has max. no. of bikes hired and weathersit 3 i.e. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds has lowest. It could be because for obvious reasons people will prefer riding a bike in clear weather.



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

Answer 2:

A dummy variable, is a binary (0 or 1) variable used in statistical analysis and regression modeling to represent categorical data. Dummy variables are particularly useful when working with categorical variables that don't have a natural numerical order or when we want to include categorical data in regression models.

While creating dummy variables we generally drop 1 column and hence it is said that for n no. of variables, n-1 variables are to be created. For eg., colour of a balls is red, blue and green we create 2 dummy variables blue and green

Ball colour	**blue**	**green**
Green	0	1
blue	1	0
red	0	0

In first case, green is 1 and blue is 0 , hence the colour of ball is green, In second case blue is 1 and green is 0, hence colour of ball is blue. In third case since both blue and green are 0 its obvious that ball will be red as there is no other possible outcome of colour of ball.

In python to drop one variable while creating dummy variables using pd.get_dummies we specify drop_first= TRUE to drop first column.

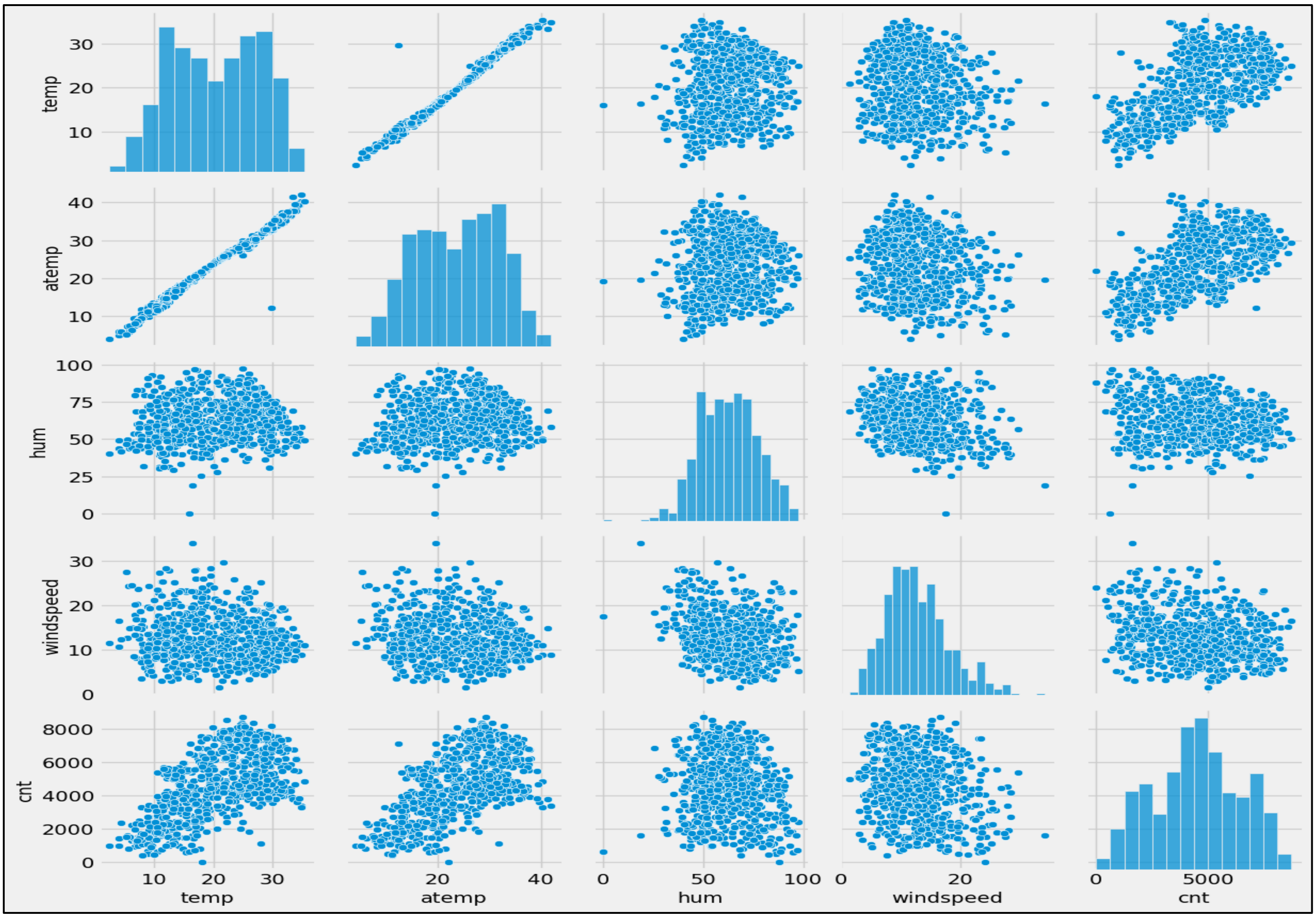
This deletion is required to avoid **dummy variable trap.** The dummy variable trap is a specific instance of multicollinearity that occurs in regression analysis when we include dummy variables for all categories of a categorical variable without excluding one category. This situation creates a **perfect multicollinearity problem**. Perfect multicollinearity is a situation that can occur in multiple linear regression when two or more independent variables in the model are highly correlated with each other, to the extent that they are perfectly correlated or perfectly predictable from each other making it impossible for the regression model to produce reliable results and coefficients.

The presence of this perfect linear relationship among the dummy variables is the dummy variable trap. It leads to multicollinearity, which can have several negative consequences, including:

1. **Unreliable Coefficients:** With perfect multicollinearity, the regression model cannot estimate unique coefficients for each dummy variable. Instead, it produces coefficients that are linear combinations of the dummy variables, making them unreliable for interpretation.
2. **Loss of Statistical Significance:** Coefficients associated with the dummy variables may not be statistically significant due to the multicollinearity issue.

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

Answer 3:



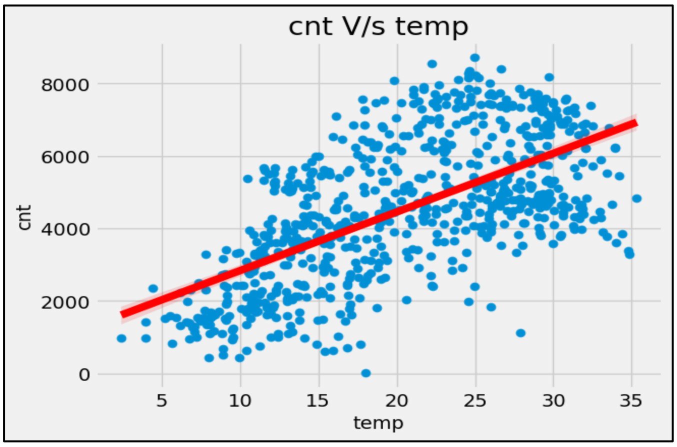
The temp and atemp variables seems to have highest co-relation with our target variable cnt.

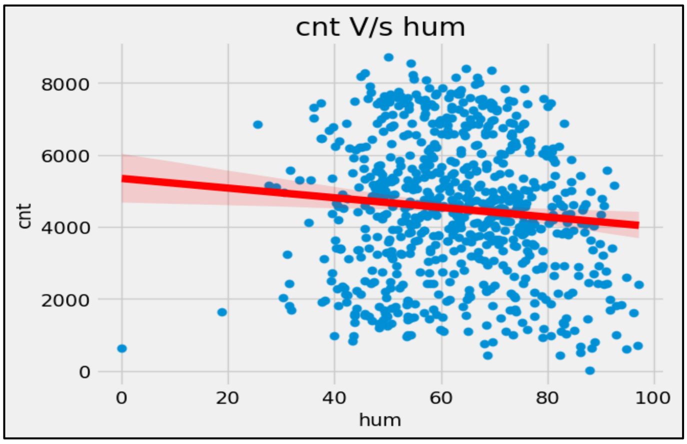
Question 4: How did you validate the assumptions of Linear Regression after building the model on the training set?

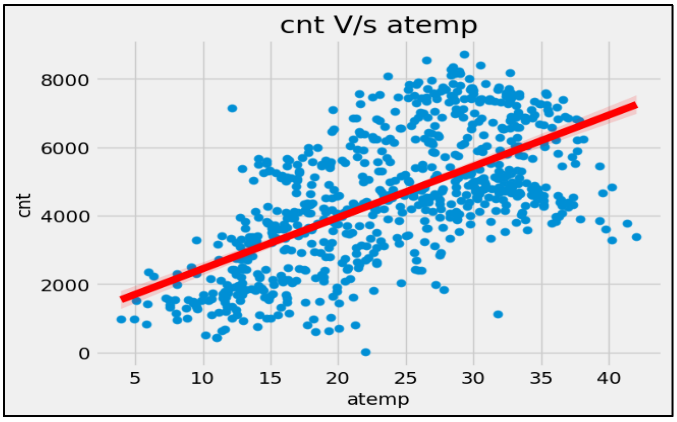
Answer 4:

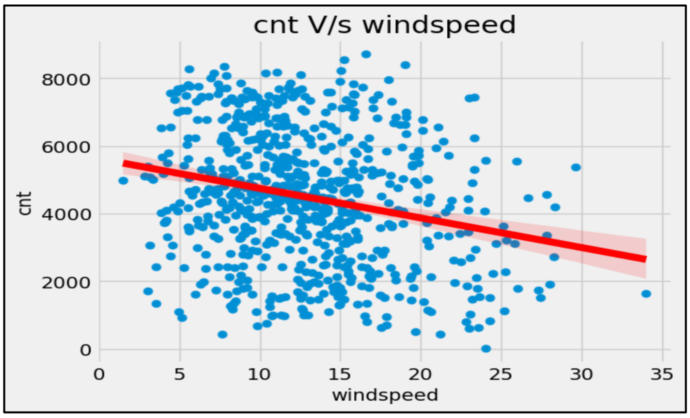
Assumption 1: Linear relationship between dependent and independent variables

To check this assumption we plot scatter plots between dependent variable and continuous independent variables





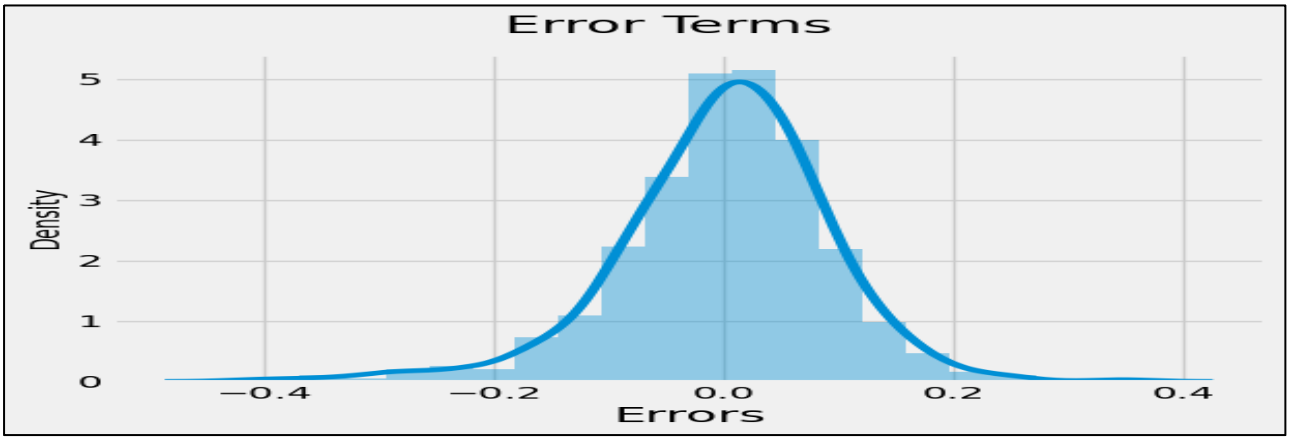




From above diagrams we can observe that there is a linear relationship between dependent and independent variables and assumption of linearity is verified.

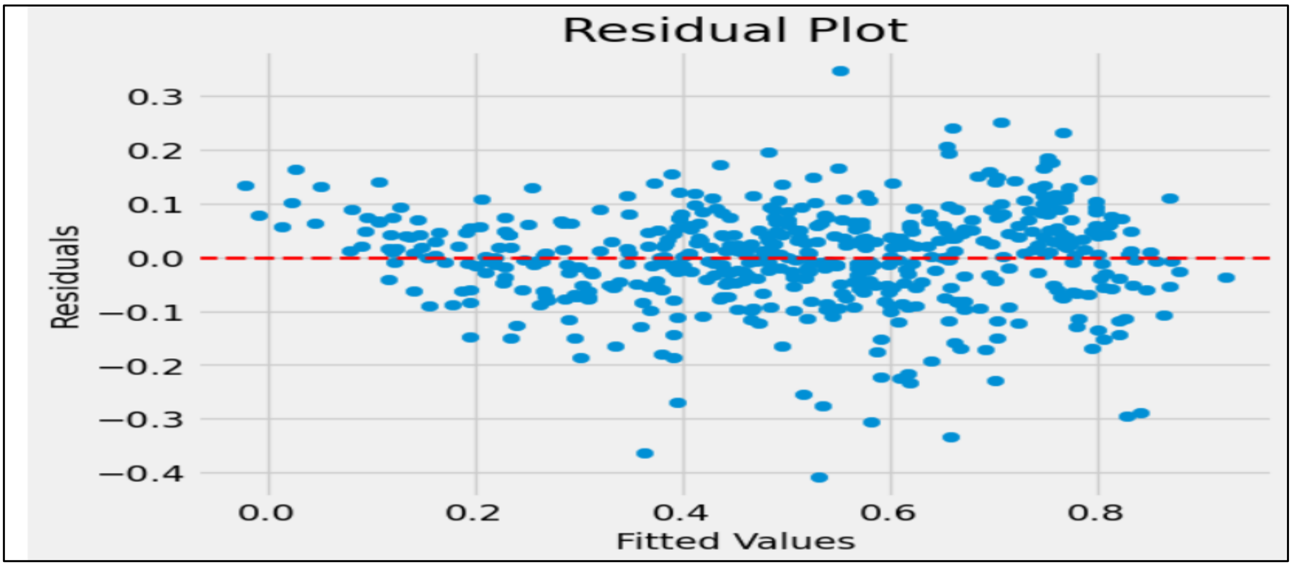
Assumption 2: Error terms are normally distributed with mean zero

We performed residual analysis by plotting a distribution plot for errors and checked if the curve obtained is a bell curve and if the errors/residuals are normally distributed.



The above curve was thus obtained which is a bell curve and errors are normally distributed Hence the assumption of normal distribution of errors is verified.

Assumption 3: homoscedasticity (constant variance) of the errors



To verify this assumption we plotted residual plot and observed the error terms. Since there is a constant (approx.) spread of residuals across the range of predicted values hence homoscedasticity is verified.

Assumption 4: statistical independence of the errors

From residual plot above, since error terms do not follow any pattern, statistical independence of the error term is confirmed.

Also we performed Durbin-watson test whose value was observed to be 2.050 and hence no autocorrelation or independence of error terms was verified.

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

Answer 5:

Based on my final model, top 3 features obtained are:

1. Temp- Coefficient= 0.5174
2. weather_3_lightsnow/Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds- Coefficient= - 0.2828
3. Year- Coefficient= 0.2325

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

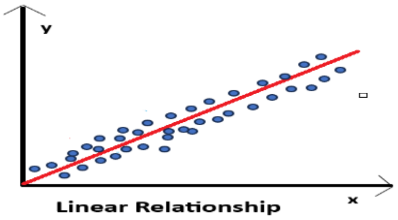
Answer 1:

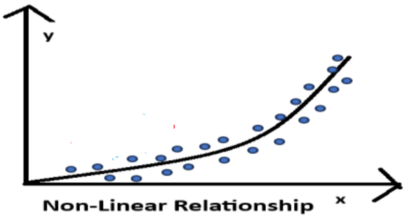
Linear regression is a linear approach to modelling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the independent variable and the dependent variable, and aims to find the best-fitting line that describes the relationship. The line is determined by minimizing the sum of the squared differences between the predicted values and the actual values.

Assumptions of Linear Regression

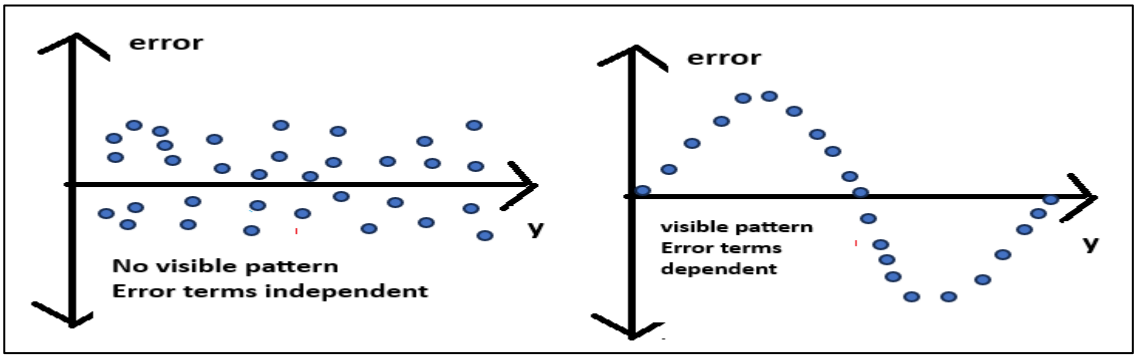
Regression is a parametric approach, which means that it makes assumptions about the data for the purpose of analysis. For successful regression analysis, it's essential to validate the following assumptions.

1. **Linearity of residuals**: There needs to be a linear relationship between the dependent variable and independent variable(s).

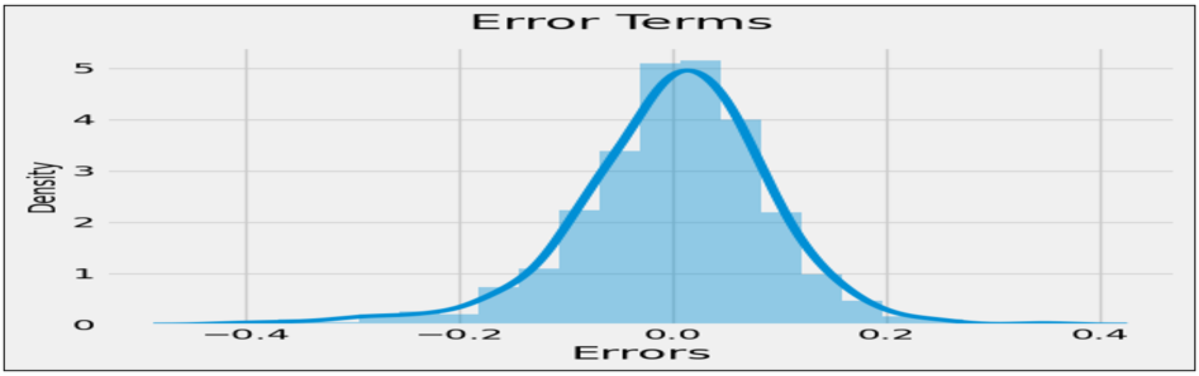




2. **Independence of residuals:** The error terms should not be dependent on one another (like in time-series data wherein the next value is dependent on the previous one). There should be no correlation between the residual terms. The absence of this phenomenon is known as **Autocorrelation.** There should not be any visible patterns in the error terms.

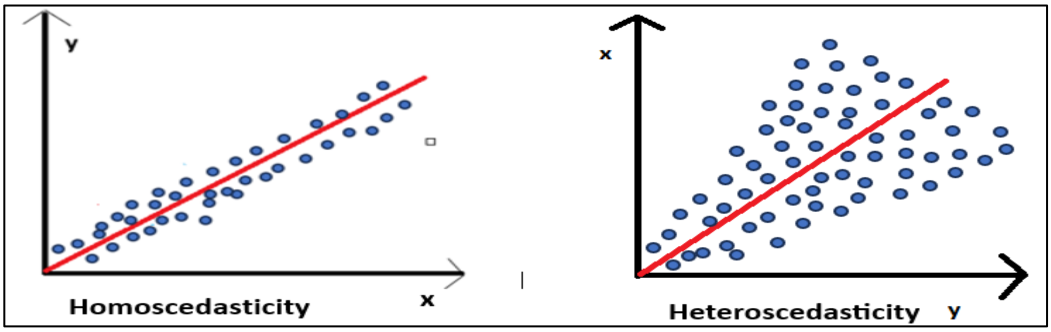


3. **Normal distribution of residuals:** The mean of residuals should follow a normal distribution with a mean equal to zero or close to zero. This is done in order to check whether the selected line is actually the line of best fit or not. If the error terms are non-normally distributed, suggests that there are a few unusual data points that must be studied closely to make a better model.



Normal distribution of error

4. **The equal variance of residuals:** The error terms must have constant variance. This phenomenon is known as **Homoscedasticity**.The presence of non-constant variance in the error terms is referred to as **Heteroscedasticity**. Generally, non-constant variance arises in the presence of outliers or extreme leverage values.



Linear regression is classified into following two types:

1. Simple Linear Regression
2. Multiple Linear Regression

Simple Linear Regression: In a simple linear regression, 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.

best-fit line is a straight line that best represents the relationship between two variables in a dataset. To calculate best-fit line linear regression uses a traditional slope-intercept form which is given below:

$Y_i = \beta_0 + \beta_1 X_i$
where $Y_i =$ Dependent variable, $\beta_0 =$ constant/Intercept, $\beta_1 =$ Slope/Intercept, $X_i =$ Independent variable.

This algorithm explains the linear relationship between the dependent(output) variable y and the independent(predictor) variable X using a straight line $Y = B_0 + B_1 X$. The goal of the linear regression algorithm is to get the best values for B0 and B1 to find the best fit line. The best fit line is a line that has the least error which means the error between predicted values and actual values should be minimum.

Multiple Linear Regression:

Multiple linear regression is a technique to understand the relationship between *a* **single dependent variable and multiple independent variables**. The formula for multiple linear regression is also similar to simple linear regression with the small change that instead of having one beta variable, we will now have betas for all the variables used. The formula is given as:

$Y = B_0 + B_1X_1 + B_2X_2 + \ldots + B_pX_p + \varepsilon$

Random Error(Residuals)
In regression, the difference between the observed value of the dependent variable(yi) and the predicted value(predicted) is called the residuals.
$\varepsilon_i = y_{predicted} - y_i$
where $y_{predicted} = B_0 + B_1 X_i$

Question 2: Explain the Anscombe's quartet in detail.

Answer 2:

Anscombe's quartet is a collection of four small datasets that were created by the British statistician Francis Anscombe in 1973. These datasets were designed to have nearly identical summary statistics (e.g., means, variances, correlations, and linear regression parameters) but to exhibit very different patterns when visualized graphically. The purpose of Anscombe's quartet is to **demonstrate the importance of data visualization** and to highlight the **limitations of relying solely on summary statistics** to understand data.

Dataset 1:

- x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
- y-values: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68

Graph of Dataset 1 consists of a set of (x,y) points that represent a linear relationship with some variance.

Dataset 2:

- x-values: Same as Dataset I
- y-values: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74

Graph of Dataset 2 shows a curve shape but doesn’t show a linear relationship.

Dataset 3:

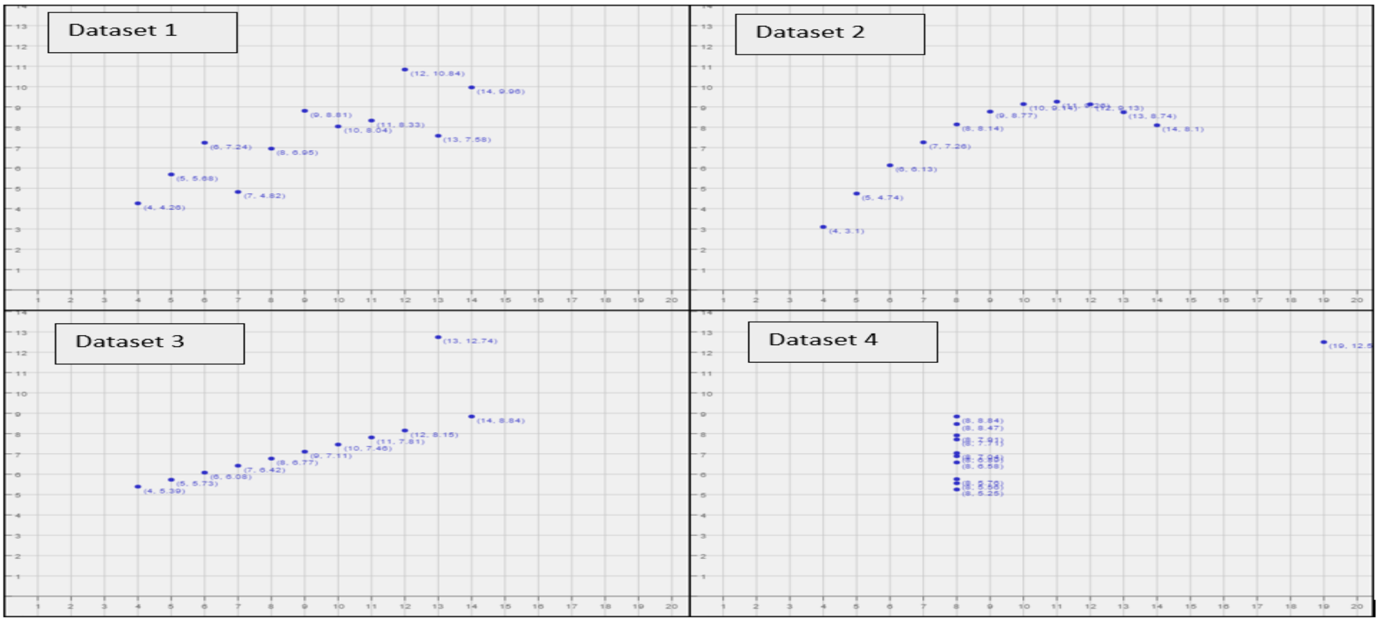
- x-values: Same as Dataset I
- y-values: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73

Graph of Dataset 3 looks like a tight linear relationship between *x* and *y*, except for one large outlier.

Dataset 4:

- x-values: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
- y-values: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89

Graph of Dataset 4 looks like the value of *x* remains constant, except for one outlier as well.



The key takeaway from Anscombe's quartet is that summary statistics alone can be deceiving. All four datasets have virtually the same mean, variance, correlation coefficient, and linear regression parameters when calculated. However, when we visualize the data, we can see that they have very different underlying structures and relationships. This emphasizes the importance of data visualization as a critical step in understanding and interpreting data correctly.

Question 3: What is Pearson's R?

Answer 3:

Pearson's correlation coefficient, often denoted as "r," is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It was developed by Karl Pearson in the early 20th century and is one of the most widely used methods for assessing the degree of association between two variables in statistics.

Pearson's R can take values between -1 and 1 and is interpreted as follows:

1. **Positive Correlation (r > 0)**: When r is positive, it indicates a positive linear relationship between the two variables. This means that as one variable increases, the other tends to increase as well, and vice versa. The closer the value of r is to 1, the stronger the positive correlation.
2. **No Correlation (r ≈ 0)**: When r is close to zero, it suggests little to no linear relationship between the variables. In other words, changes in one variable are not associated with predictable changes in the other.
3. **Negative Correlation (r < 0)**: When r is negative, it indicates a negative linear relationship between the two variables. This means that as one variable increases, the other tends to decrease, and vice versa. The closer the value of r is to -1, the stronger the negative correlation.

   The formula for calculating Pearson's correlation coefficient is as follows:

   $r= n(\sum xy)-(\sum x)(\sum y)/\sqrt{[n\sum x^2 – (\sum x)^2][n\sum y^2-(\sum y)^2]}$

Where :

- r is the Pearson correlation coefficient.
- n is the number of data points.
- ∑xy is the sum of the products of the paired values of the two variables.
- ∑x and ∑y are the sums of the individual values of the two variables.
- $\sum x^2$ and $\sum y^2$ are the sums of the squares of the individual values of the two variables.

Pearson's correlation coefficient is a valuable tool in statistics and data analysis for exploring relationships between variables. It helps researchers and analysts understand whether there is a linear association between two variables and the strength and direction of that association. However, it's important to note that Pearson's R only measures linear relationships and may not capture more complex or nonlinear associations between variables.

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

Answer 4:

Scaling refers to the process of transforming the values of variables or features in a dataset to bring them to a common scale or range. Scaling is performed for various reasons, primarily to make the data more suitable for analysis and modeling, especially when working with machine learning algorithms or statistical techniques that are sensitive to the scale of variables. The two common types of scaling are **normalized scaling** and **standardized scaling** (also known as z-score scaling or standardization).

The need for scaling arises because different variables in a dataset may have different units of measurement or different orders of magnitude. If these variables are not on a common scale, it can lead to issues when performing certain types of analyses or building predictive models. Scaling helps address these issues by making the data more comparable and ensuring that the contribution of each variable to the analysis or model is appropriate.

Criteria	**Normalization**	**Standardization**
Range of Values:	Normalization scales data to a specific range, typically between 0 and 1, but it can be any user-defined range [a, b].	Standardization scales data to have a mean of 0 and a standard deviation of 1.
Formula:	The formula for normalized scaling of a variable X in the range [a,b] is: $X_{normalized}$=(X-min(X))*((b-a)+a) /(max(X)-min(X))	The formula for standardized scaling of a variable X is: $X_{standardized}$=X-μ/σ Here, μ is the mean of the variable, and σ is the standard deviation.
Purpose:	The primary purpose of normalization is to ensure that all variables have values within the same range.	The primary purpose of standardization is to make variables comparable, especially when they have different units or different orders of magnitude.
Preservation of Interpretability:	Normalization preserves the original range of the data, making it easier to interpret in the context of the original values.	Standardization creates values with no inherent meaning or direct interpretation in the original scale.
Sensitivity to Outliers:	Normalization can be sensitive to outliers since it relies on the minimum and maximum values.	Standardization is less sensitive to outliers because it centers the data

		around the mean and scales by the standard deviation.

In summary, normalization and standardization are both techniques for scaling data, but they serve different purposes. Normalization is suitable when we want to bring data into a specific range, preserving the interpretability of the original scale. Standardization is useful when we want to make variables comparable, remove the influence of scale, and deal with data that may not follow a normal distribution.

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

Answer 5:

The Variance Inflation Factor (VIF) is a measure used in regression analysis to assess multicollinearity, which is the phenomenon where two or more predictor variables in a regression model are highly correlated with each other. VIF quantifies how much the variance of the estimated regression coefficients is increased due to multicollinearity. A high VIF indicates that a predictor variable is highly correlated with other predictors, making it challenging to interpret the individual effects of these variables.

A VIF value can become infinite (or extremely large) when there is **perfect multicollinearity** in the regression model. In a perfect correlation situation, the R-squared will be equal to 1, meaning that the independent variable(s) can completely explain the variability in the dependent variable. Hence according to **formula of VIF (VIF =1/1-R2)** if R-squared is 1, denominator will be 0 and 1/0= infinity.

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

Answer 6:

A Quantile-Quantile plot, commonly referred to as a Q-Q plot, is a graphical tool used in statistics and data analysis to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles (ordered values) of the observed data to the quantiles of the expected theoretical distribution. The main purpose of a Q-Q plot is to visually inspect whether the observed data deviates from the expected distribution.

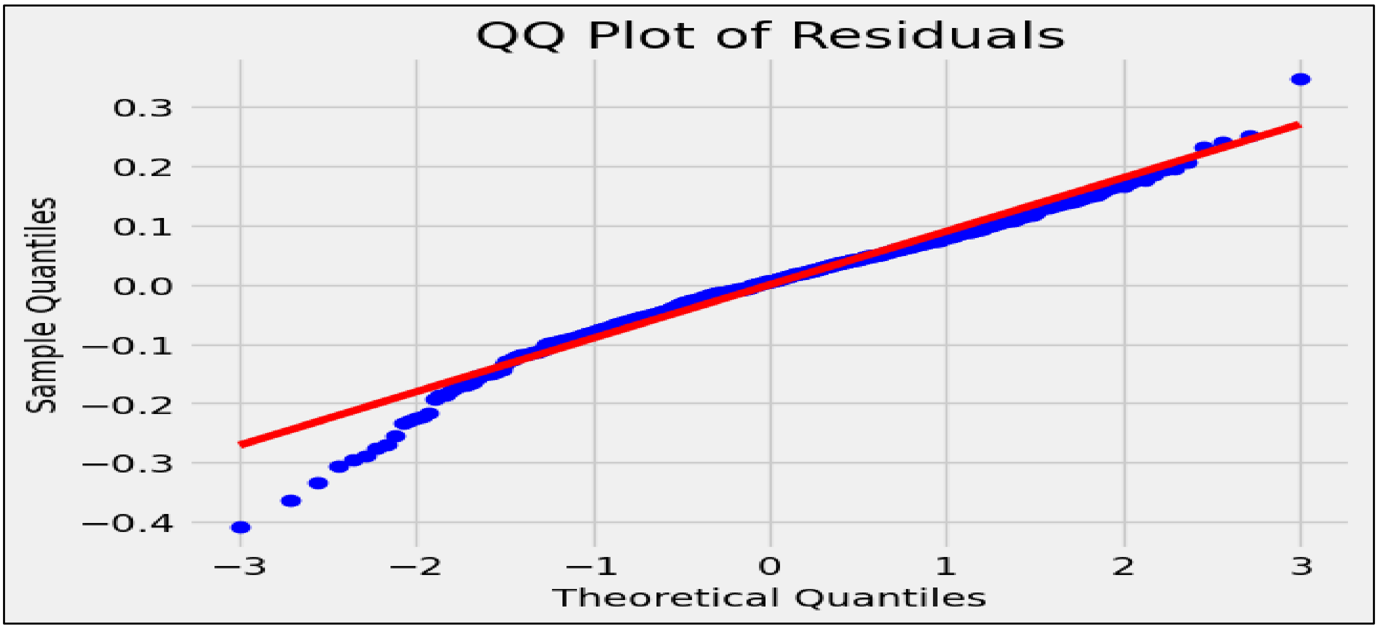


Fig: Q-Q Plot of a normal distribution

Use and Importance of a Q-Q Plot in Linear Regression:

1. **Assumption Checking:** Q-Q plots are used to assess whether the residuals from a linear regression model follow a normal distribution. If the residuals do not follow a normal distribution, it can indicate that the linear regression assumptions are violated.
2. **Identification of Outliers:** Q-Q plots can reveal outliers in the data. Outliers may appear as data points that deviate significantly from the theoretical quantiles, leading to a departure from the expected straight-line pattern in the Q-Q plot.
3. **Model Validity and Interpretability:**
   - Ensuring that the residuals are normally distributed is important for the validity of hypothesis tests and confidence intervals in linear regression.
   - Normally distributed residuals also make the interpretation of coefficients and the assessment of statistical significance more straightforward..