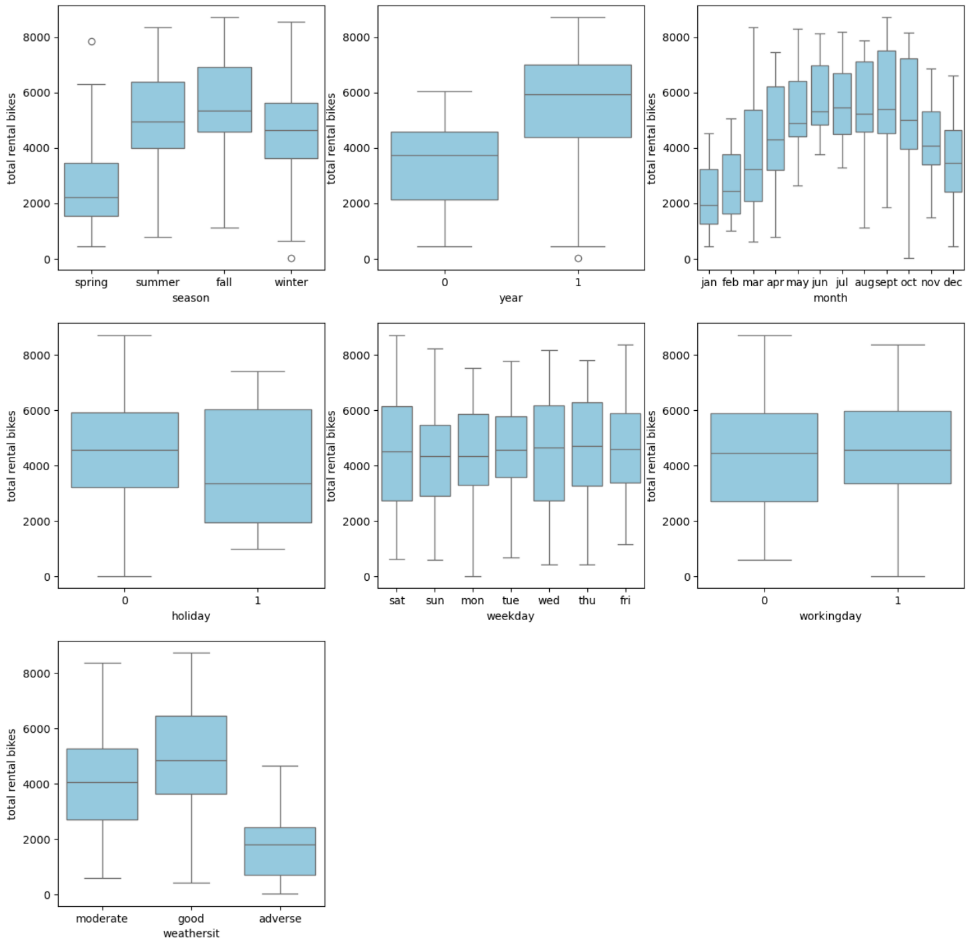
Assignment-based Subjective Questions

Question 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Answer: Please find the box plot for the categorical variable.
Observations:

- Season: Fall has the highest demand of bikes followed by summer & winter
- Year: over a period of time the demand increased in 2019, mostly post Covid
- Month: the demand constantly grown from Jan till June and then it became constant and started falling during oct to December
- Holiday: During Holidays demand has decreased
- Weekday: Weekday is having no impact on the demand
- Working day: The mean remains constant and it has no impact
- Weathershit: during good days the demand is really high followed by moderate



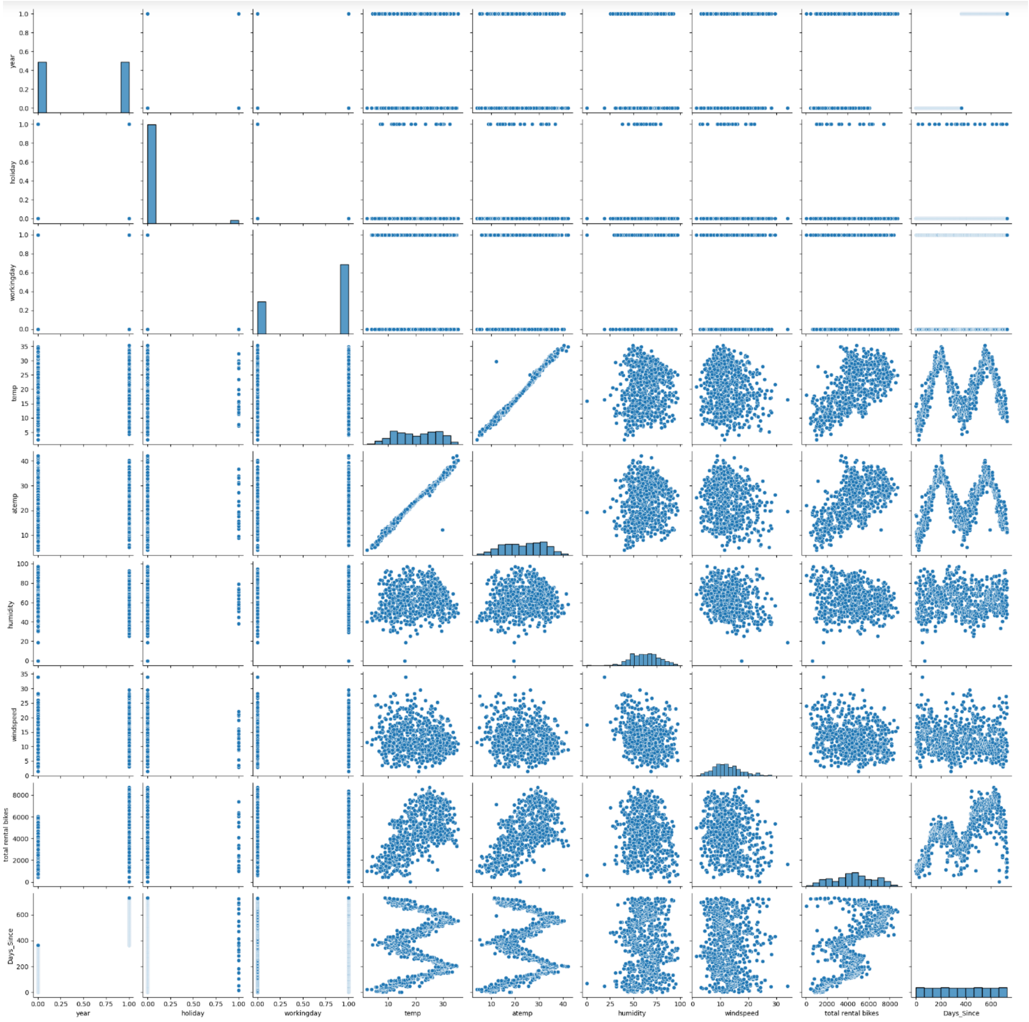
Question 2: Why is it important to use **drop_first=True** during **dummy variable** creation?
Answer: drop_first=True is used to avoid multicollinearity issues in regression models.

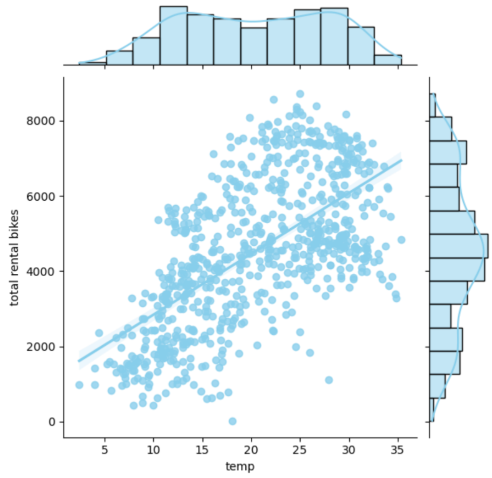
Example: Suppose we have a categorical variable "Day" with seven categories: Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday. When creating dummy variables, we aim to represent each day as a binary indicator variable.
If we create dummy variables without dropping the first category and encode Monday as "1" in the "Monday" dummy variable, Tuesday through Sunday would be represented by separate dummy variables. This means if all dummy variables for Tuesday through Sunday are "0", it automatically implies that the day is Monday.
However, including all dummy variables can lead to issues. For example, if all dummy variables for Tuesday through Sunday are "0", it's already clear that the day is Monday. Including dummy variables

for all days introduces redundant information, potentially causing problems in regression analysis due to multicollinearity.

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

- Temp & ATemp is highly correlated
- hence, we can plan to drop one of the variable

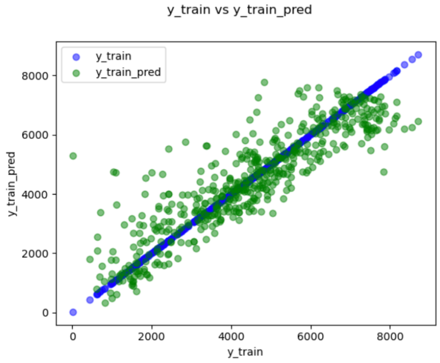




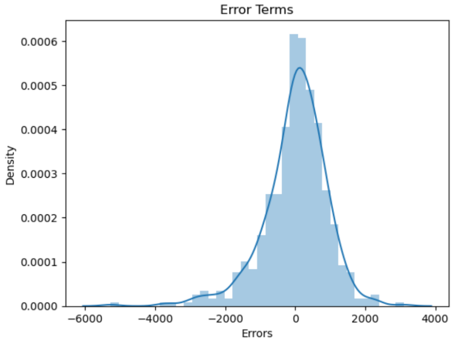
Question 4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: I have used 3 different criteria with help of visualization test it out

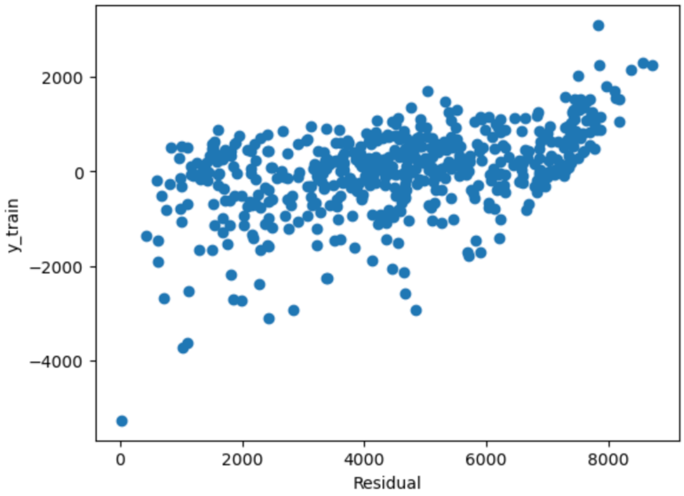
- **Linearity**: Check the linearity assumption by plotting the observed values against the predicted values. The plot should ideally form a straight line with no discernible pattern.



-
- **Normality of Residuals**: Examine the distribution of residuals using histograms. The residuals should ideally follow a normal distribution.



-
- **Homoscedasticity**: Assess the homoscedasticity assumption by plotting residuals against predicted values. The plot should ideally show constant variance across all predicted values. A cone-shaped or funnel-shaped pattern indicates heteroscedasticity.



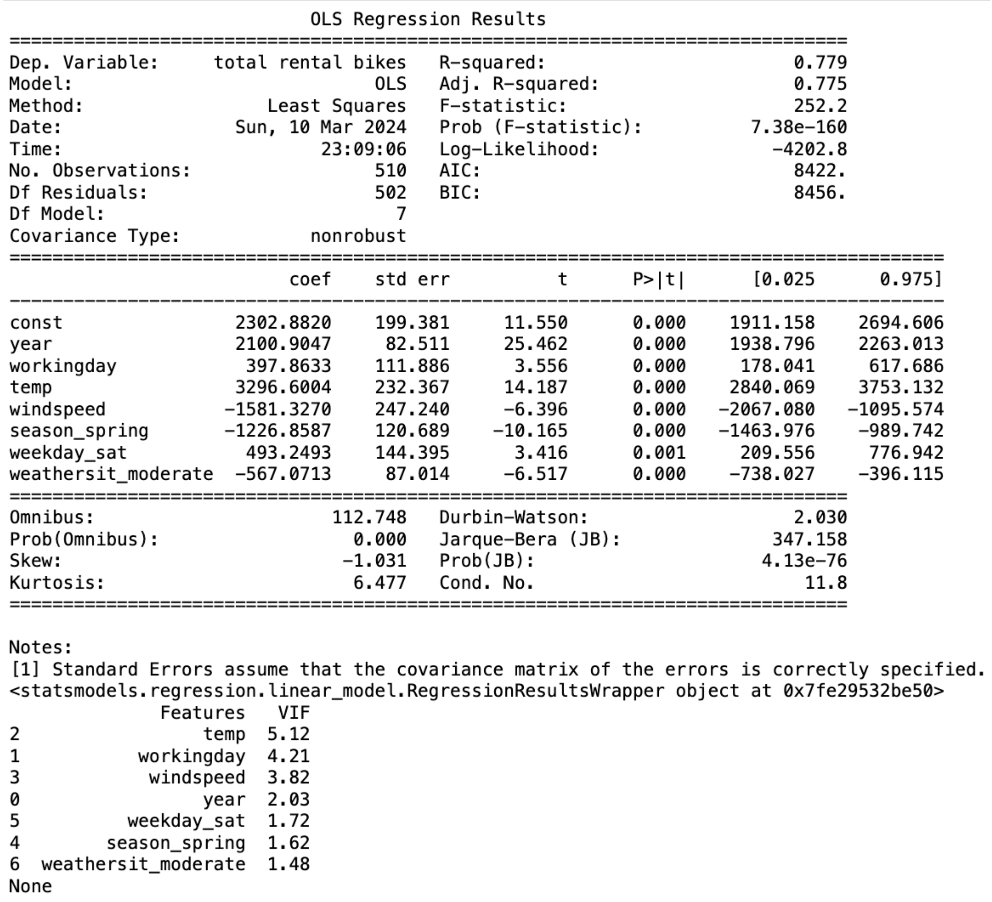
-

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

Answer:

Top 3 features are

1. Temp (Coef: 3296)
2. Year (Coef: 2100)
3. Windspeed (Coef: -1581)



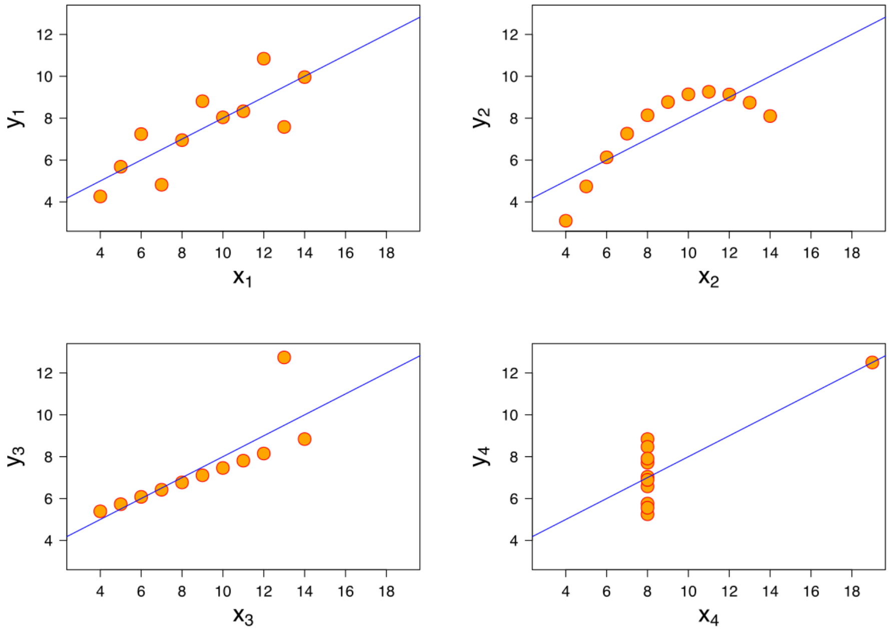
```
                            OLS Regression Results
==============================================================================
Dep. Variable:     total rental bikes   R-squared:                       0.779
Model:                            OLS   Adj. R-squared:                  0.775
Method:                 Least Squares   F-statistic:                     252.2
Date:                Sun, 10 Mar 2024   Prob (F-statistic):          7.38e-160
Time:                        23:09:06   Log-Likelihood:                -4202.8
No. Observations:                 510   AIC:                             8422.
Df Residuals:                     502   BIC:                             8456.
Df Model:                           7
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                2302.8820    199.381     11.550      0.000    1911.158    2694.606
year                 2100.9047     82.511     25.462      0.000    1938.796    2263.013
workingday            397.8633    111.886      3.556      0.000     178.041     617.686
temp                 3296.6004    232.367     14.187      0.000    2840.069    3753.132
windspeed           -1581.3270    247.240     -6.396      0.000   -2067.080   -1095.574
season_spring       -1226.8587    120.689    -10.165      0.000   -1463.976    -989.742
weekday_sat           493.2493    144.395      3.416      0.001     209.556     776.942
weathersit_moderate  -567.0713     87.014     -6.517      0.000    -738.027    -396.115
==============================================================================
Omnibus:                      112.748   Durbin-Watson:                   2.030
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              347.158
Skew:                          -1.031   Prob(JB):                     4.13e-76
Kurtosis:                       6.477   Cond. No.                         11.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
<statsmodels.regression.linear_model.RegressionResultsWrapper object at 0x7fe29532be50>
              Features   VIF
2                 temp  5.12
1           workingday  4.21
3            windspeed  3.82
0                 year  2.03
5          weekday_sat  1.72
4        season_spring  1.62
6  weathersit_moderate  1.48
None
```

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.
Answer:
Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model. Linear regression is based on the popular equation "y = mx + c". It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error. In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term. Regression is broadly divided into simple linear regression and multiple linear regression.

1. **Simple Linear Regression** : SLR is used when the dependent variable is predicted using only one independent variable.
2. **Multiple Linear Regression** :MLR is used when the dependent variable is predicted using multiple independent variables.

Question 2: Explain the Anscombe's quartet in detail.
Answer:
Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties



● The first scatter plot (top left) appears to be a simple linear relationship.
● The second graph (top right) is not distributed normally, while there is a relation between them, it's not linear.
● In the third graph (bottom left), the distribution is linear, but should have a different regression line the calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
● Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not. indicate any relationship between the variables.

Question 3: What is Pearson’s R?
Answer :
The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

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**.

Question 4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Answer:
Scaling is the process of transforming numerical features of a dataset to a standard range or distribution. It involves adjusting the values of the features so that they have a similar scale, typically between 0 and 1 or with a mean of 0 and a standard deviation of 1. Scaling is performed to ensure that the features contribute equally to the analysis and to improve the performance of certain machine learning algorithms.

The main reasons for performing scaling are:

Equalize Feature Influence: Most of the times, collected 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.
Convergence Speed: Scaling can help algorithms converge faster during the training process, especially for optimization algorithms like gradient descent.
Numerical Stability: Scaling can improve numerical stability and prevent overflow or underflow issues in calculations.

There are two common methods of scaling: normalized scaling and standardized scaling.

Normalized Scaling:
- Also known as min-max scaling.
- Transforms features to a range between 0 and 1.
- Preserves the original distribution of the data but compresses it into the specified range.

Standardized Scaling:
- Also known as z-score scaling or standardization.
- Transforms features to have a mean of 0 and a standard deviation of 1
- Shifts the distribution of the data so that it has a mean of 0 and a standard deviation of 1.
- Results in data with approximately normal distribution.

Question 5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer: The VIF (Variance Inflation Factor) measures the extent of multicollinearity in a regression analysis, specifically quantifying how much the variance of the estimated regression coefficients is inflated due to multicollinearity among the independent variables.

When the VIF value is infinite, it indicates perfect multicollinearity among the independent variables. Perfect multicollinearity occurs when one or more independent variables can be exactly predicted from the others. This situation leads to a singularity in the matrix of independent variables, causing the calculation of the inverse matrix (required for computing VIF) to fail, resulting in an infinite VIF value.

consider an example where we're analysing the factors influencing students' exam scores. Suppose we have three independent variables: "Study Hours," "Homework Completion Percentage," and "Class Attendance Percentage."
Example Data:

- Study Hours: [4, 5, 6, 7, 8] (hours)
- Homework Completion Percentage: [80, 85, 90, 95, 100] (%)
- Class Attendance Percentage: [90, 92, 94, 96, 98] (%)
- Exam Scores: [80, 85, 88, 92, 95] (%)

Now, let's say we want to predict students' exam scores using linear regression. We scale the independent variables (Study Hours, Homework Completion Percentage, and Class Attendance Percentage) to ensure that they have a similar range of values, which can help stabilize the model's performance.

We calculate the VIF to check for multicollinearity among the independent variables:

1. **Study Hours**:
   - VIF = 1 / (1 - R^2) = 1 / (1 - 1) = 1 (No multicollinearity)
2. **Homework Completion Percentage**:
   - VIF = 1 / (1 - R^2) = 1 / (1 - 1) = Infinite (Perfect multicollinearity with Study Hours)
3. **Class Attendance Percentage**:
   - VIF = 1 / (1 - R^2) = 1 / (1 - 1) = Infinite (Perfect multicollinearity with Study Hours)

In this example, the VIF for Homework Completion Percentage and Class Attendance Percentage is infinite because each of these variables can be perfectly predicted from the Study Hours variable. This indicates that including both of these variables in the regression model would result in redundancy and numerical instability.

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

Answer: Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential

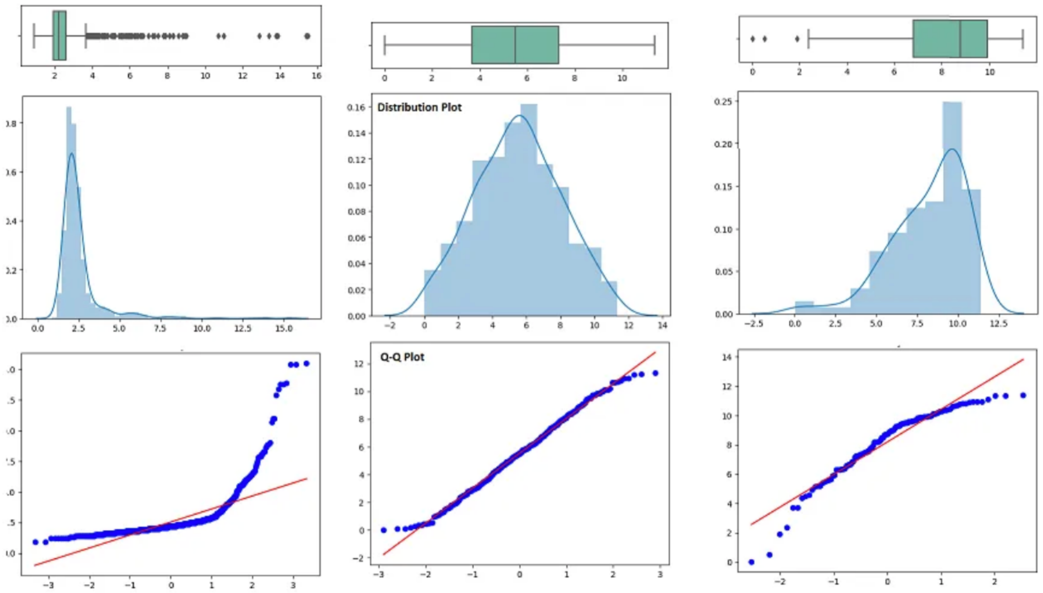
The power of Q-Q plots lies in their ability to summarize any distribution visually.

QQ plots is very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
- Skewness of distribution

In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.

If the datasets we are comparing are of the same type of distribution type, we would get a roughly straight line. Here is an example of normal distribution.


Distribution Plot
Q-Q Plot