Bike Rental Prediction Model

by:

Akhil Kumar Garg

Business Overview

A US bike-sharing provider **BoomBikes** has recently suffered considerable dips in their revenues due to the ongoing Corona pandemic. The company is finding it very difficult to sustain in the current market scenario. So, it has decided to come up with a mindful business plan to be able to accelerate its revenue as soon as the ongoing lockdown comes to an end, and the economy restores to a healthy state.

In such an attempt, BoomBikes aspires to understand the demand for shared bikes among the people after this ongoing quarantine situation ends across the nation due to Covid-19. They have planned this to prepare themselves to cater to the people's needs once the situation gets better all around and stand out from other service providers and make huge profits.

Problem Statement

They have contracted a consulting company to understand the factors on which the demand for these shared bikes depends. Specifically, they want to understand the factors affecting the demand for these shared bikes in the American market. The company wants to know:

1. Which variables are significant in predicting the demand for shared bikes.

2. How well those variables describe the bike demands

Our Approach

1. Go through different attributes in Bike data set, use reference of Data dictionary and understand the given data e.g numerical, categorical and null values
2. Import data into dataframe using python
3. Data cleaning:
    a. Identify and remove columns having all null values
    b. Identify percentage of null data in each columns
    c. Ensure there is no null data in final data set
4. Data Analysis: We performed univariate and bivariate analysis
5. Prepare the data for Modeling
    a. Dummy variables: Identified categorical columns and converted to dummy variable
    b. Drop columns which have been converted to dummy and extra columns
    c. Splitting data in train and test¶(70:30)
    d. Rescaling the Features (Min-Max scaling)
    e. Training the Model (Manual and RFE)
    f. Residual analysis
    g. Prediction and Evaluation on the Test Set

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?
Answer :

y= 0.0492*mnth_Sep - 0.1270*season_spring - 0.0806*holiday + 0.2406*yr - 0.1754*windspeed
- 0.0672*weathersit_mist + 0.4205*temp - 0.0748*mnth_July + 0.0264*mnth_Oct + const*0.2772

1. **Mnth** : In March, September and October high number of bike users. We can observe in linear regression equation July month negatively impacting the user count
2. **Seasons**: Summer and fall high number of bike users. Sprint has lowest number of users.We can observe in linear regression equation spring season negatively impacting the user count
3. **Holiday**: On non holiday high number of users. We can observe in linear regression equation holiday negatively impacting the user count
4. **Year**: Year 2019 have high number of users in comparison of 2018. We can observe in linear regression equation year is positively impacting the user count
5. **Weather** situation: On clear day we can observe high number of users. We can observe in linear regression equation mist weather is negatively impacting the user count

Assignment-based Subjective Questions

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

Answer : While creating dummy columns we get **n** dummy columns. We can drop one dummy column because one of the combination will be uniquely representing this redundant column. By dropping one dummy column we reduce one level of redundancy.

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

Answer

y= 0.0492*mnth_Sep - 0.1270*season_spring - 0.0806*holiday + 0.2406*yr - 0.1754*windspeed
- 0.0672*weathersit_mist + 0.4205*temp - 0.0748*mnth_July + 0.0264*mnth_Oct + const*0.2772

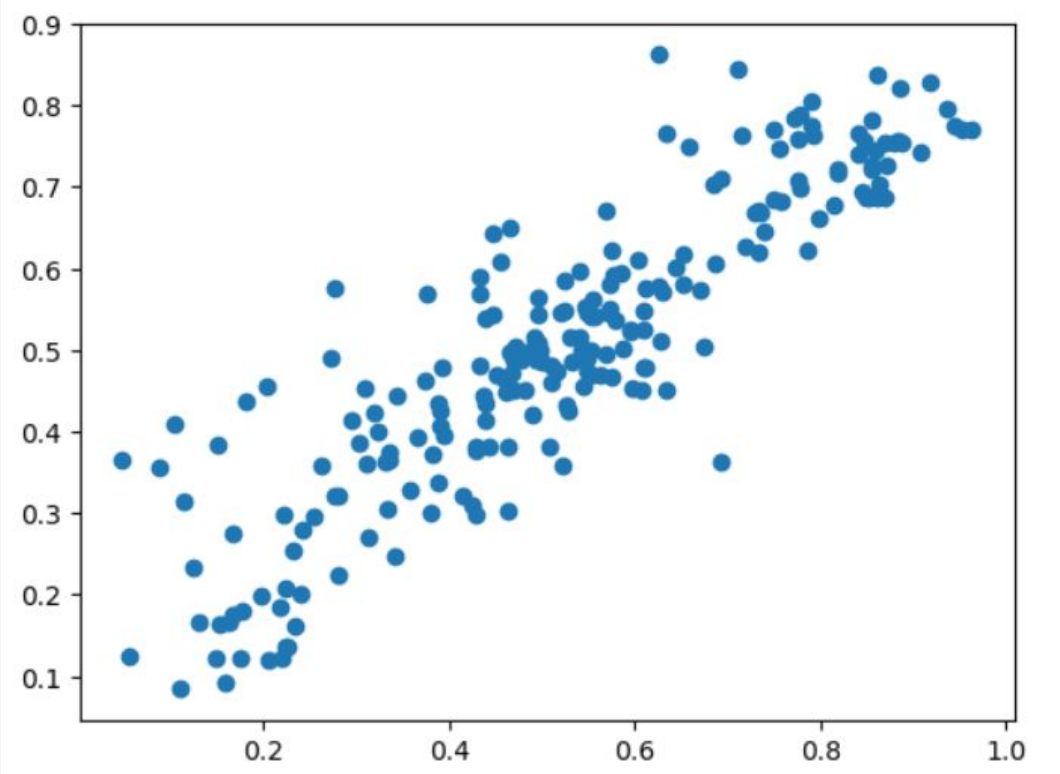
1. **Windspeed**: In low to moderate wind speed high number of user count. We can observe in linear regression equation wind speed is negatively impacting the user count. As the wind speed increase it will impact to decrease the user count
2. **Temp**:Medium to high temperature have high number of user count.As the temperature goes below 10 or above 30 users count start decreasing.

Assignment-based Subjective Questions

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

Answer : Assumptions of linear regression

- **Linear relationship between X and Y:** We can observe from the scatter plot predicted values are in straight line. Which shows X and Y has linear relationship

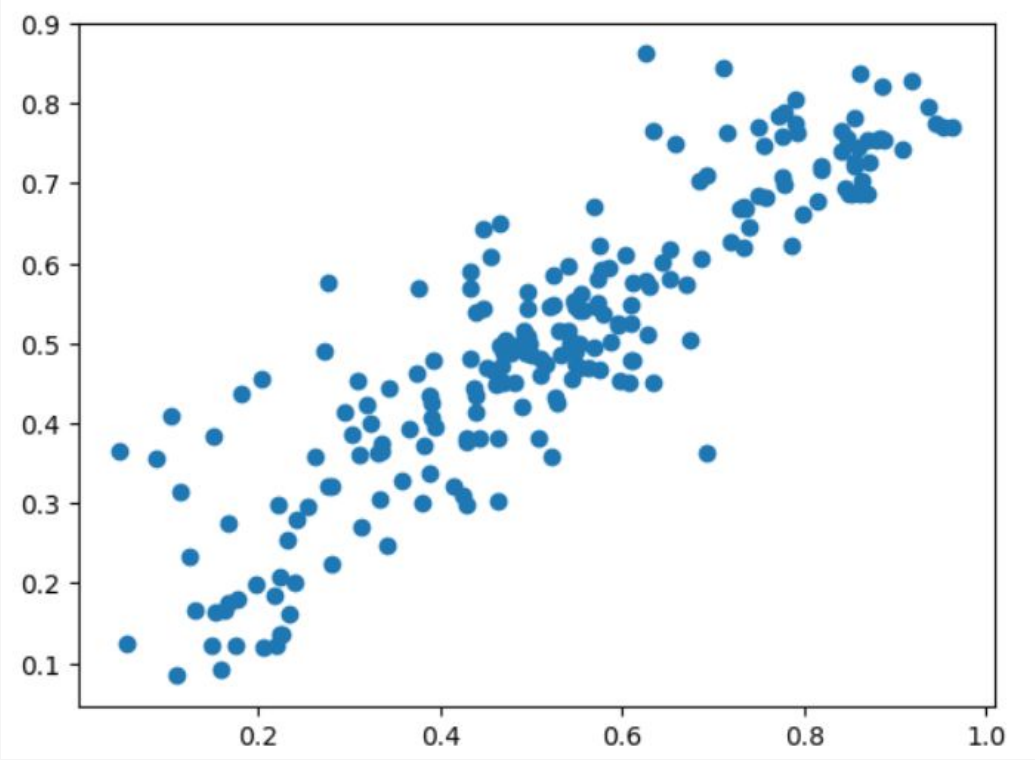


Assignment-based Subjective Questions

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

Answer : Assumptions of linear regression

- **Error terms are normally distributed around zero:** We generated distribution plot of residual and found error terms are normally distributed

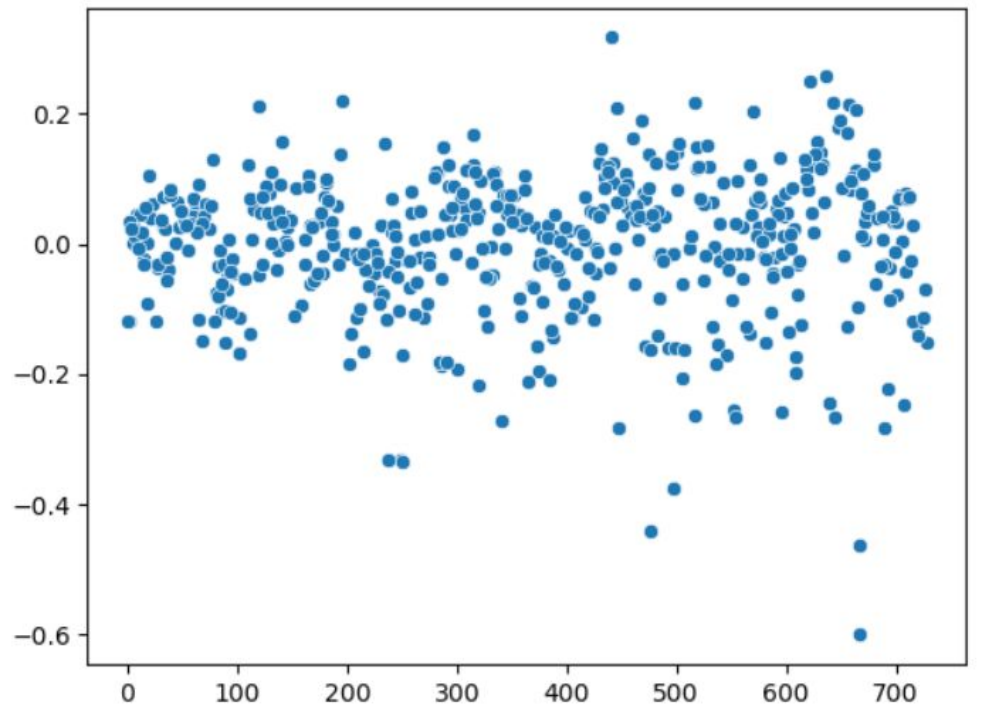


Assignment-based Subjective Questions

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

Answer : Assumptions of linear regression

- **Error terms are independent of each other:** We plotted scatter plot of error terms. We can observer from graph error terms are scattered and not following any pattern

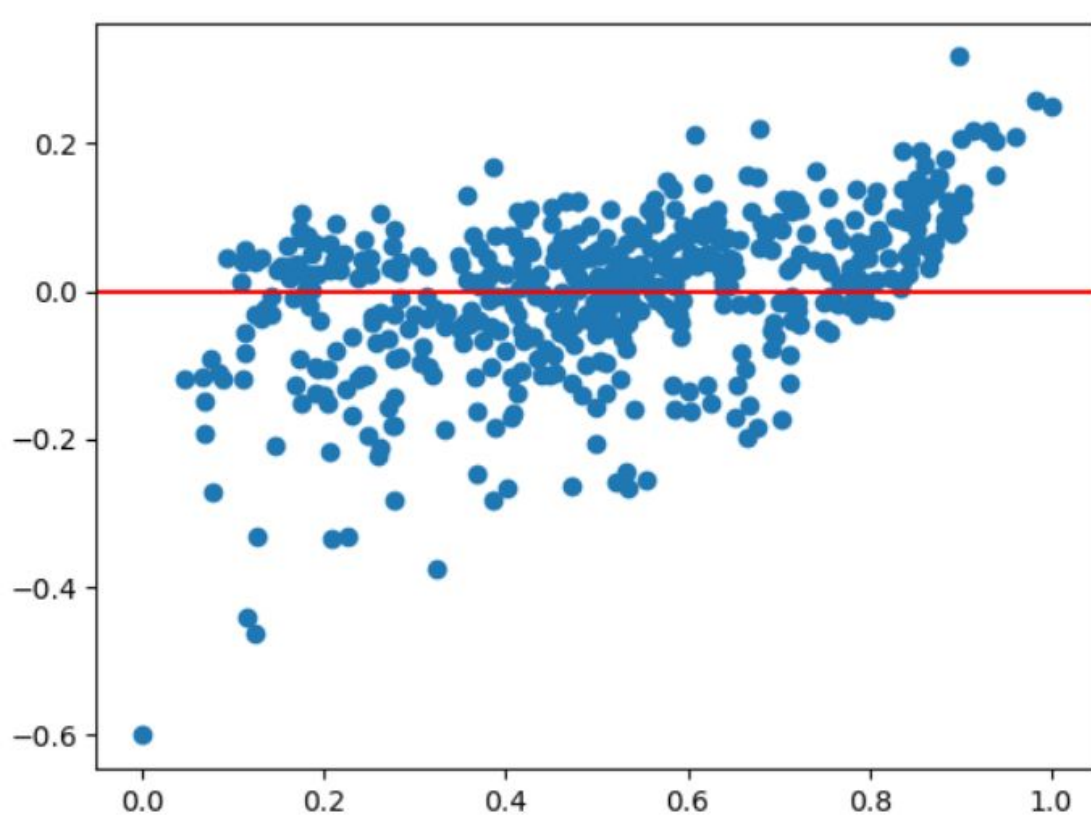


Assignment-based Subjective Questions

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

Answer : Assumptions of linear regression

- **Error terms have constant variance:** We plotted scatter plot of error terms and fit values. We can observer from graph the errors have constant variance, with the residuals scattered randomly around zero. The points on the plot above appear to be randomly scattered around zero, so assuming that the error terms have a mean of zero is reasonable. The vertical width of the scatter doesn't appear to increase or decrease across the fitted values, so we can assume that the variance in the error terms is constant.



Assignment-based Subjective Questions

Q4. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer :

y= 0.0492*mnth_Sep - 0.1270*season_spring - 0.0806*holiday + 0.2406*yr - 0.1754*windspeed

- 0.0672*weathersit_mist + 0.4205*temp - 0.0748*mnth_July + 0.0264*mnth_Oct + const*0.2772

Based on the final model top 3 features contributing significantly towards explaining the demand of the shared bikes are

1. **Temperature**
2. **Year**
3. **Windspeed**

General Subjective Questions

Q1. Explain the linear regression algorithm in detail?

Answer :

- Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.
- When there is only one independent feature, it is known as **Simple Linear Regression,** and when there are more than one feature, it is known as **Multiple Linear Regression**.
- Similarly, when there is only one dependent variable, it is considered **Univariate Linear Regression**, while when there are more than one dependent variables, it is known as **Multivariate Regression**.

Q2. Explain the Anscombe's quartet in detail.

Answer :

- **Anscombe's quartet** comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.
- The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.
- The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

General Subjective Questions

Q3. What is Pearson's R?

Answer :

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

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general 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

General Subjective Questions

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

Answer :

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

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.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

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

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

General Subjective Questions

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

Answer :

This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

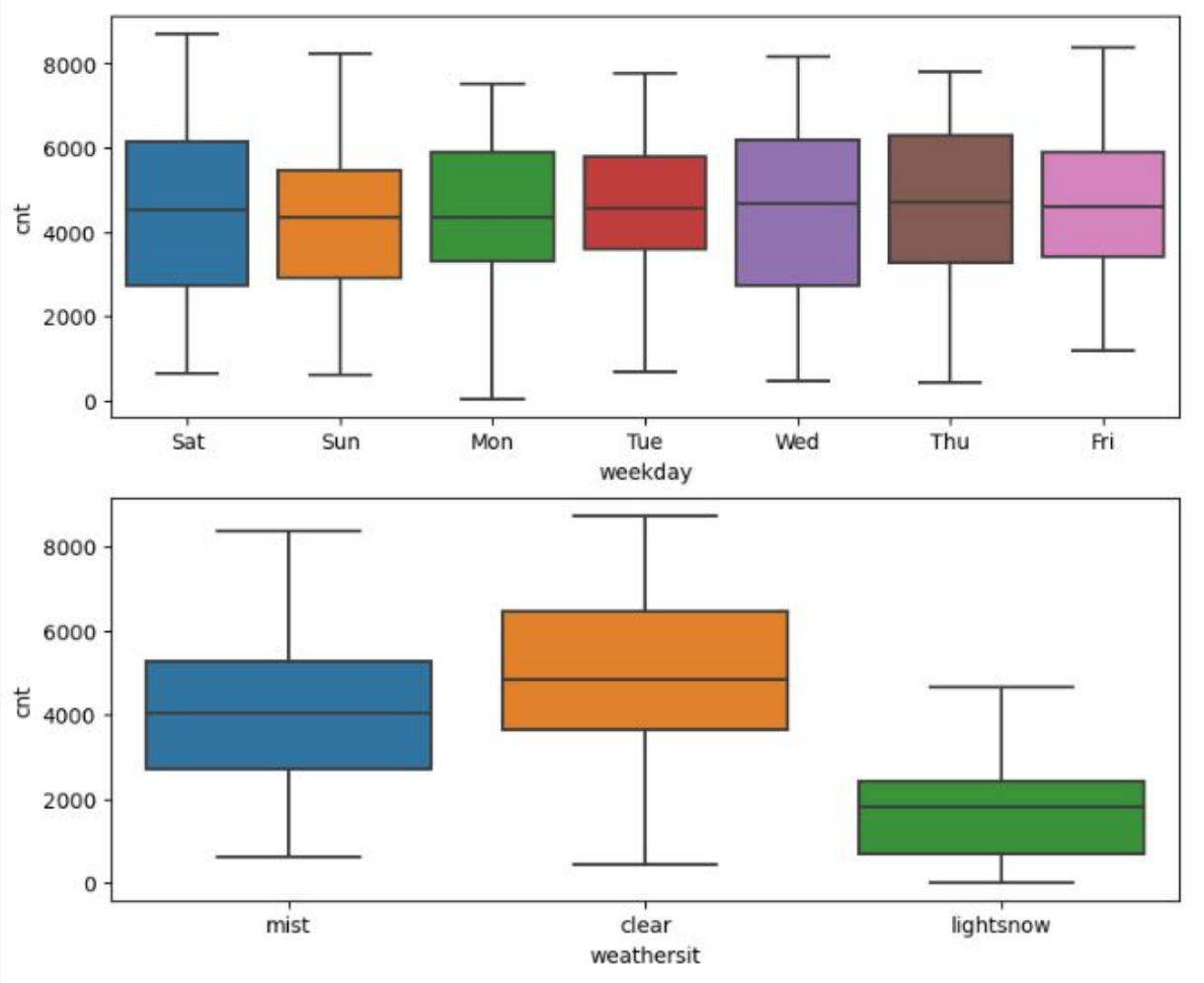
Answer :

The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.

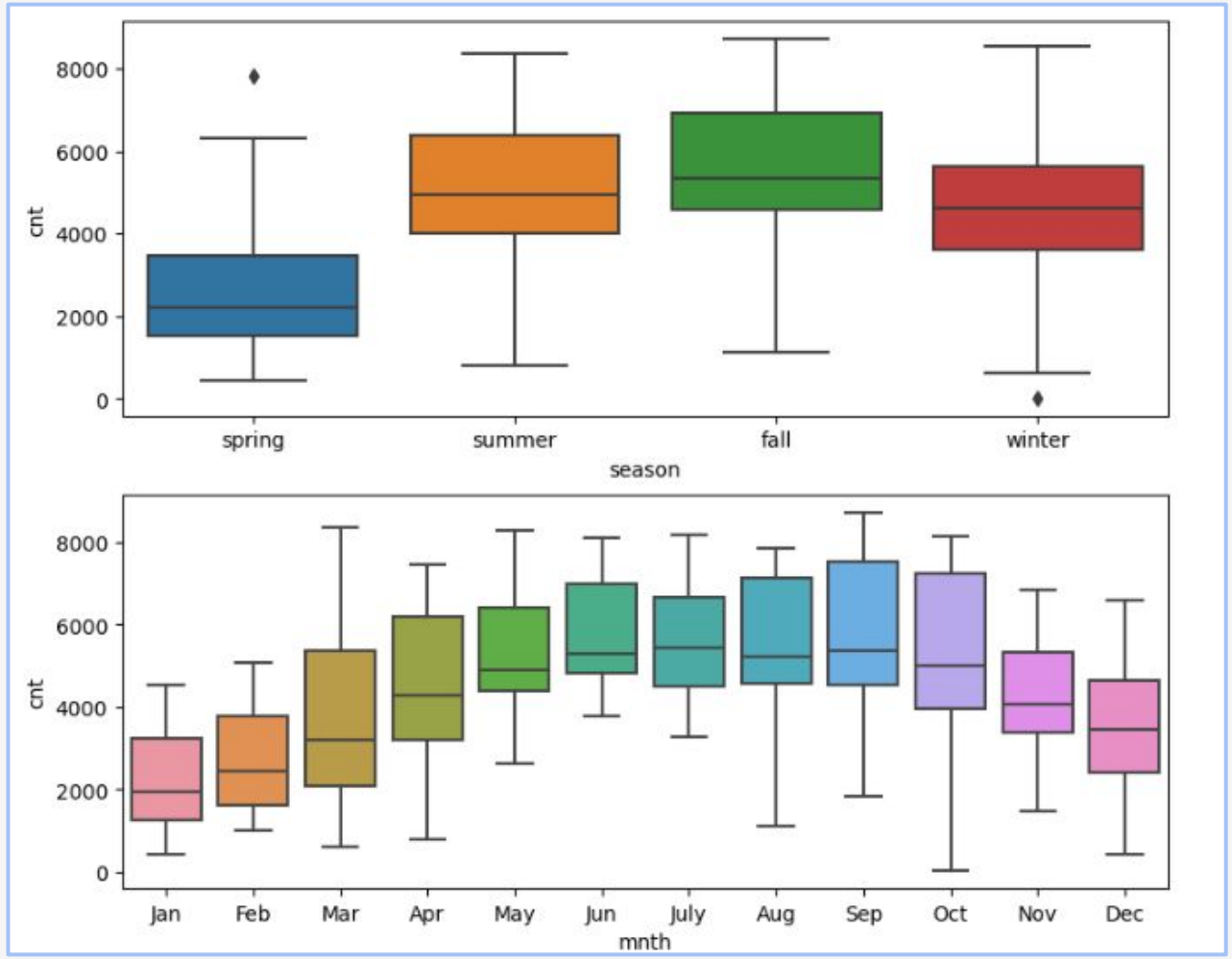
Interpretation of Q-Q plot:

- If the points on the plot fall approximately along a straight line, it suggests that your dataset follows the assumed distribution.
- Deviations from the straight line indicate departures from the assumed distribution, requiring further investigation.

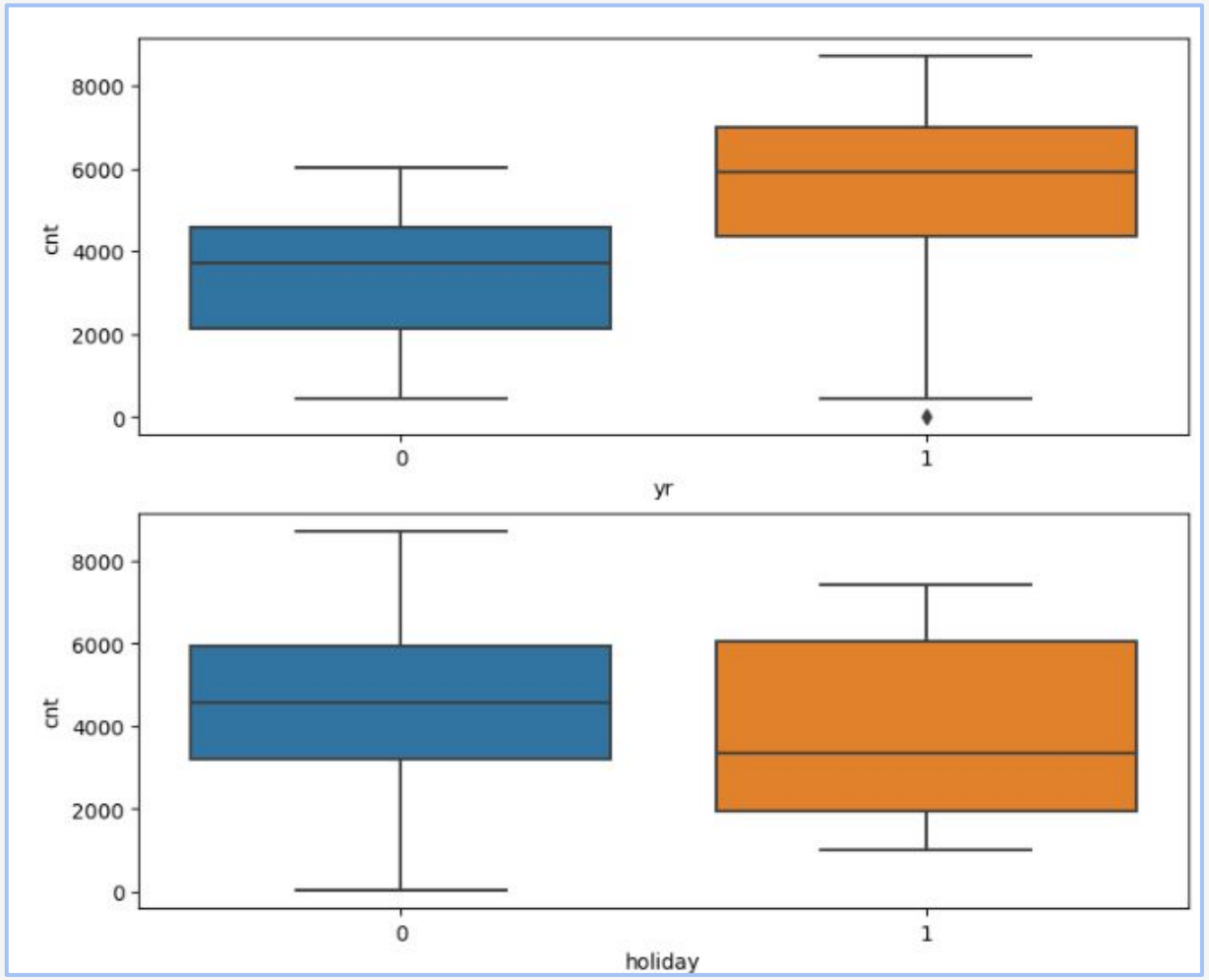
Visualisation Categorical



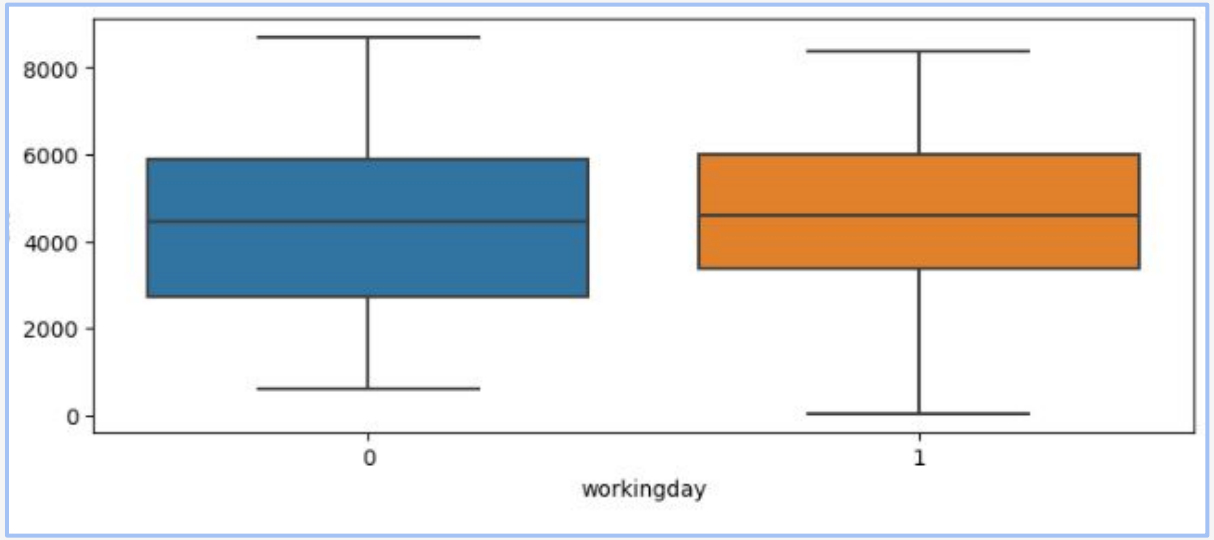
Visualisation Categorical



Visualisation Categorical



Visualisation Categorical



Visualisation Numerical variables