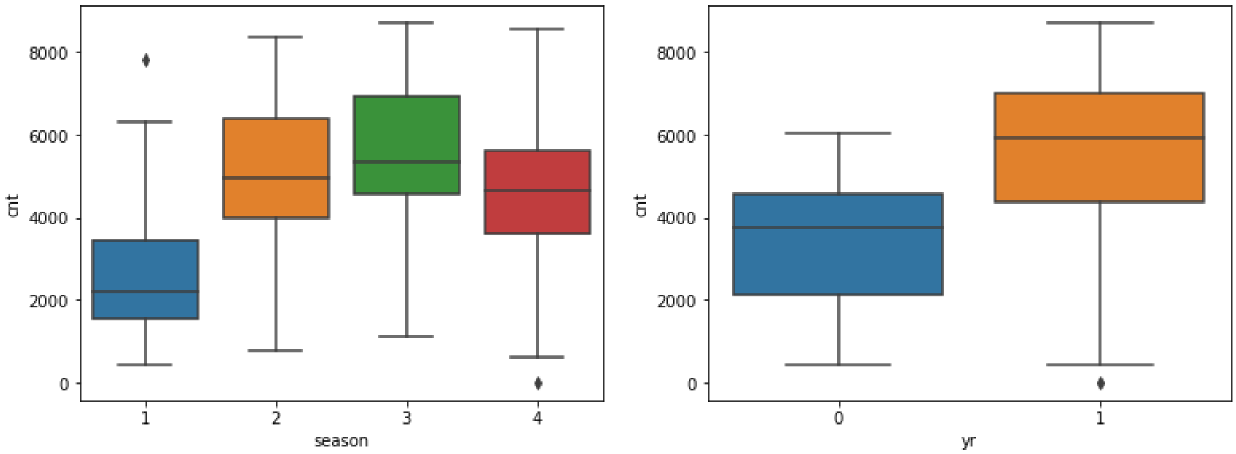
Assignment-based Subjective Questions

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

- As per the ‘Weathersit’ variable count of Bike Rentals is more during clear Weather (During Sunshine)
- As per ‘Season’ variable summer and winter are more favorable for bike rentals
- Year 2019 has high number of booking compared to year 2018

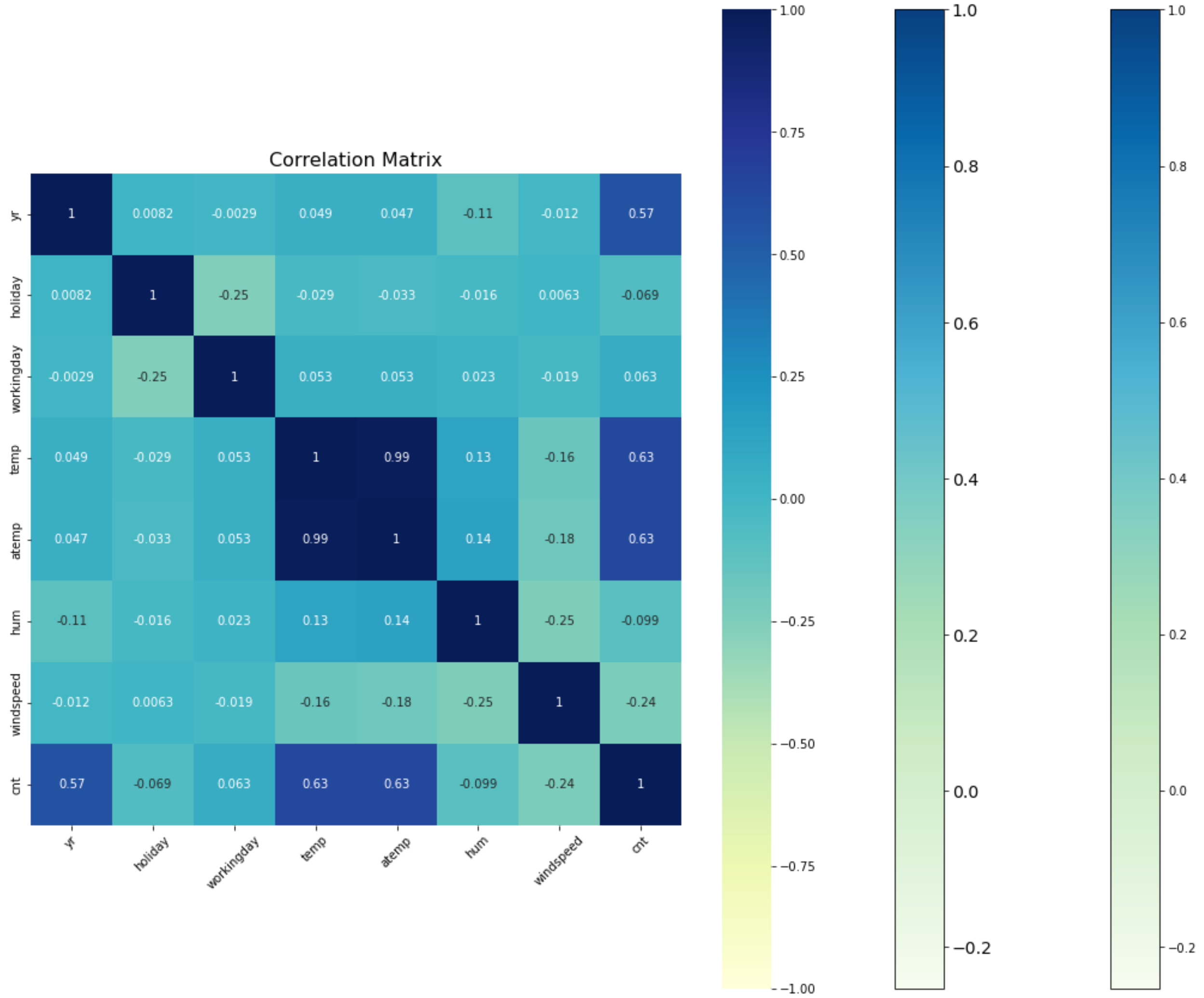


2. **Why is it important to use drop_first=True during dummy variable creation?**

- We used drop_first = True during dummy variable creation to avoid redundant features.
- And if we don’t use drop, then dummy variables will be correlated and it will affect the model.

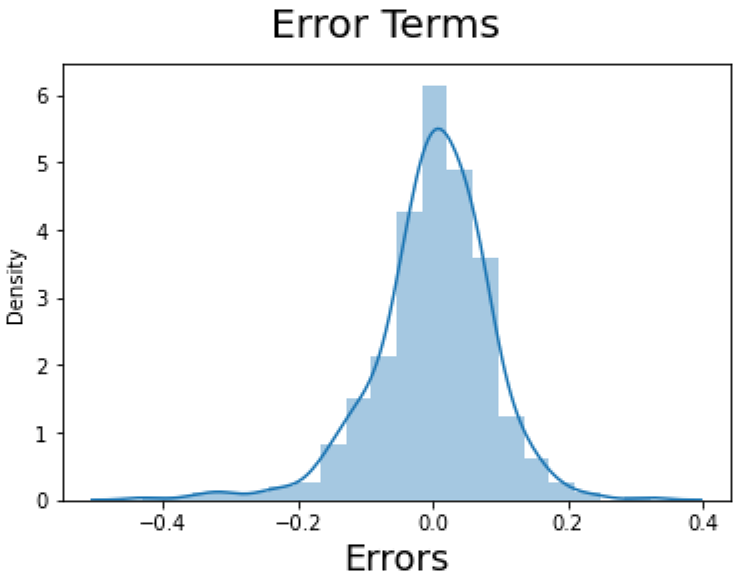
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

- Count(Cnt) (Target Variable) has significantly High Correlation With temperature (Temp) variable



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

- Residual Analysis of the train data



- The R² value for the test data and train data:-
    - The R² value for the test data = 0.8281625450407033,
    - The R² value for the train data = 0.853;

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

- Temperature (0.4014)
- Weather Situation :- Sunshine (0.0611)
- Year (0.2321)

General Subjective Questions

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

   Linear regression is one of the basic methods of machine learning in which we train a model to predict the behavior of your data based on variables. In the linear regression case two variables on the x-axis and the y-axis must be linearly correlated.

   Mathematically, we can write a linear regression equation as:

   $$y = a + bx$$

   Where a and b given by the formulas:

   $$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

   $$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

   Here, x and y are two variables on the regression line.

   b = Slope of the line

   a = y-intercept of the line

   x = Independent variable from dataset

   y = Dependent variable from dataset

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

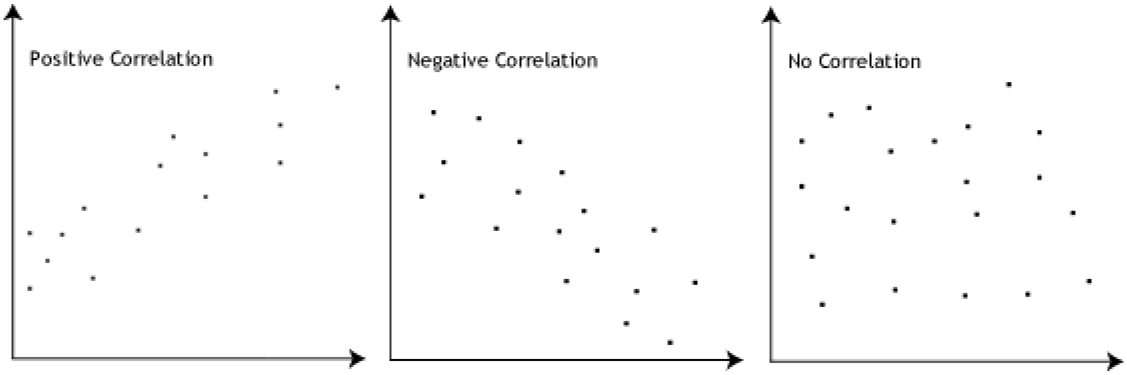
   Anscombe Quartet includes four sets of data with similar mathematical features, but they appear to be very different when inserted into a graph. Each data set has eleven points (x, y). They were developed in 1973 by mathematician Francis Anscombe to illustrate both the importance of graph data before analyzing it and the external impact on mathematical structures.

3. What is Pearson’s R?

In the statistics, the Pearson coefficient coefficient (PCC), also known as Pearson's r, Pearson product-moment coefficient (PPMCC), or bivariate coefficient, is the measure of the linear correlation between two sets of data. It is the compatibility of two variables, which are distinguished by the product of their standard deviation; therefore a fairly standardized estimate of covariance, so that the result remains a value between − 1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Pearson r Formula

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

Here,

- *r* = Pearson correlation coefficient
- *x* = Values in the first set of data
- *y* = Values in the second set of data
- *n* = Total number of values.

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

- It is a step-by-step data Pre-Processing variable to make data more general within a given range. It also helps to speed up the calculations in the algorithm.

- Most of the time, a set of data collected contains features that vary greatly in size, units, and widths. If the scaling is not done the algorithm only takes the size of the account and not the units which is why making the model wrong. To solve this problem, we have to do the scaling to bring all the variables to the same magnitude.

- It is important to note that the scaling only affects the coefficients and there are no other parameters that are affected such as t-statistic, F-statistic, p-values, R-squared, etc.
- Normalization usually measures values in the range [0,1]. Standardization usually means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

S. N O.	Normalisation	Standardisation
1.	Minimum and maximum value of	Mean and standard deviation is used for scaling.

	features are used for scaling	
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bound to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.

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

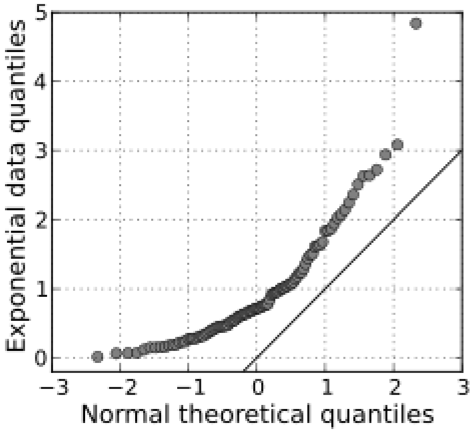
   If there is a complete correlation, then VIF = infinity. This shows a complete correlation between two independent variables. In the case of absolute correlation, we get R2 = 1, which results in 1 / (1-R2) infinity. To solve this problem we need to dump one separate element in the database that results in this complete multicollinearity.

   The infinite value of VIF indicates that the corresponding variables can be displayed exactly with the line combination of other variables (also showing the infinite VIF).

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

Q-Q Plots (Quantile-Quantile Sites) are sections of two quantiles opposite each other. A quantile is a fraction where certain values fall below that quantile. For example, a median is a quantile where 50% of the data falls below that point and 50% lies above it. The purpose of Q Q sites is to determine whether the two data sets are from the same distribution. A 45 degree angle is set on the Q Q plot; if two sets of data come from a standard distribution, points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:


Exponential data quantiles
Normal theoretical quantiles