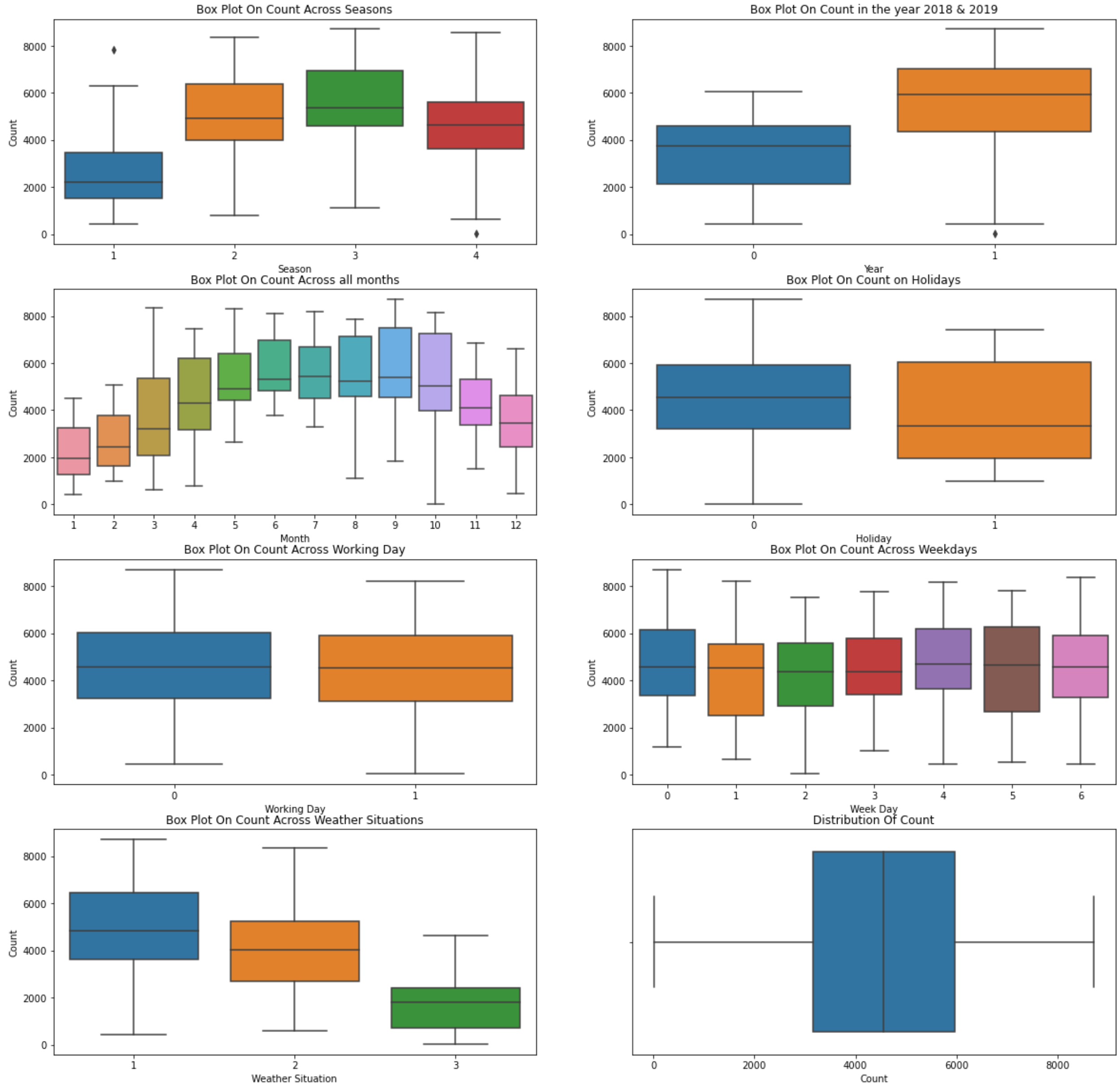
Assignment-based Subjective Questions

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

According to the analysis of the categorical variables from the dataset, we could infer a very strong effect on the dependent variable i.e., 'cnt'. From the below subplots prepared with categorical variables w.r.t 'cnt', we can see that variables- 'Season', 'Year', 'Month', 'Holiday' and 'Weathersit' strongly affect the dependent variable 'cnt'. Whereas, variables such as 'Weekday' and 'Workingday' do not show much variation in the number of total rental bikes including both casual and registered.



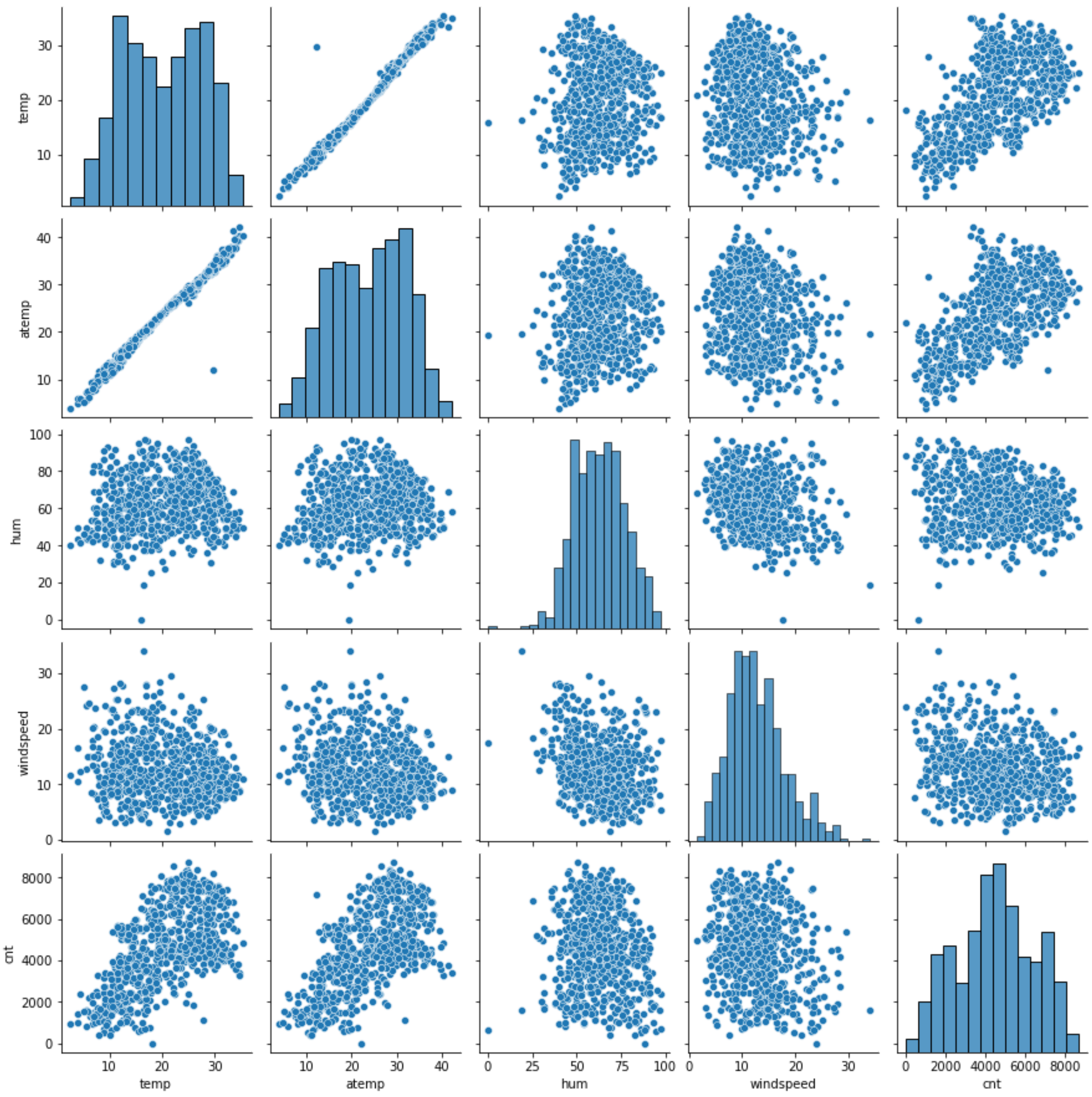
2. Why is it important to use drop_first=True during dummy variable creation?

It is important to use drop_first=True after we create dummy variables for categorical variables as it helps in reducing the extra column. It therefore reduces the correlations created among dummy variables. Whenever we have ‘n categorical variables each with ‘m’ levels then number of dummy variables is (m-1)*n. For instance, in the assignment, we have created dummy variables for the column ‘weathersit’ using the same function to reduce one column.

```
Weathersit= pd.get_dummies(bikedata['weathersit'], drop_first = True)
```

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

According to the pair plot among the numerical variables, ‘temp’ and ‘atemp’ both visibly have the highest correlation with the target variable ‘cnt’. On the other hand, both ‘temp’ and ‘atemp’ have the highest correlation among themselves too. Therefore, we can infer that the ‘temp’ variable has the highest correlation with ‘cnt’.

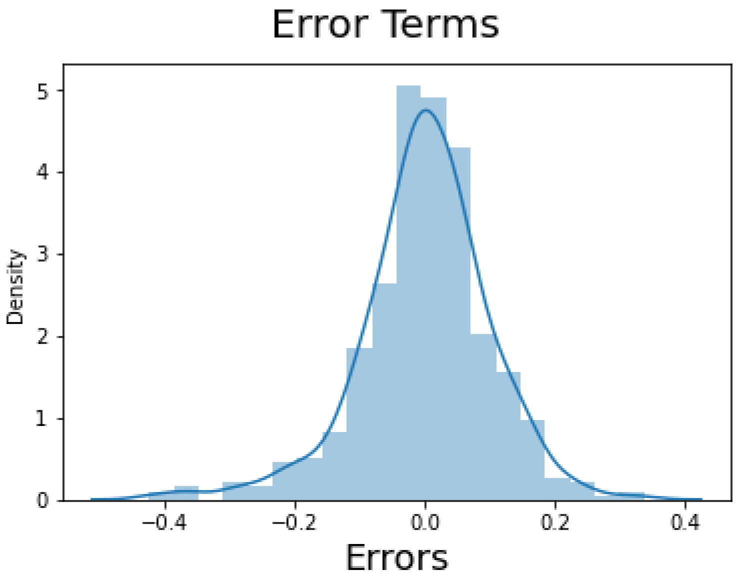


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

In the dataset given, I validated the assumptions of Linear Regression after building the model on the training set through the following steps.

1. I ensured that our X and Y follow a linear relationship.
2. I used visualisation of the data with a scatter plot and the fitted regression line to check for Homoscedasticity and if error terms are independent of each other.

3. I used heatmap to analyse correlation and rule out possibilities of Multicollinearity.
4. Lastly, I ensured that the error terms are normally distributed.



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

Based on the final model, the equation of our best fitted line is:

Count of total rental bikes = Season x 0.044 + Year x 0.234 + Temperature x 0.473 + Windspeed x (-0.137) + Mist x (-0.071) + Light Snow x (-0.277)

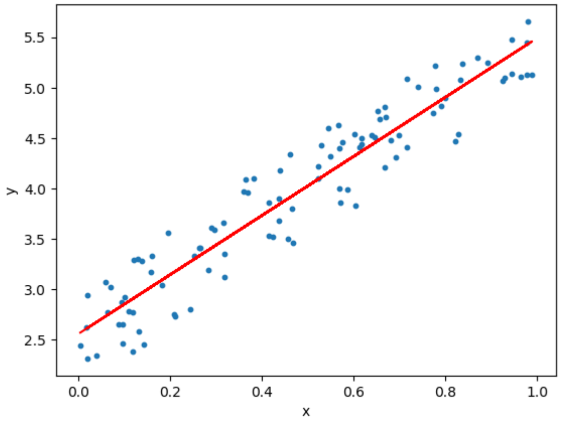
The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

- Temperature
- Light Snow (Weather situation)
- Year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a machine learning algorithm under the supervised learning model. This type of models are used for continuous variables for target prediction based on independent variables. It is useful in analysing the relationship between variables and forecasting.



A linear regression model finds the best estimate of Y for each X (Predictor variable) using the following equation:

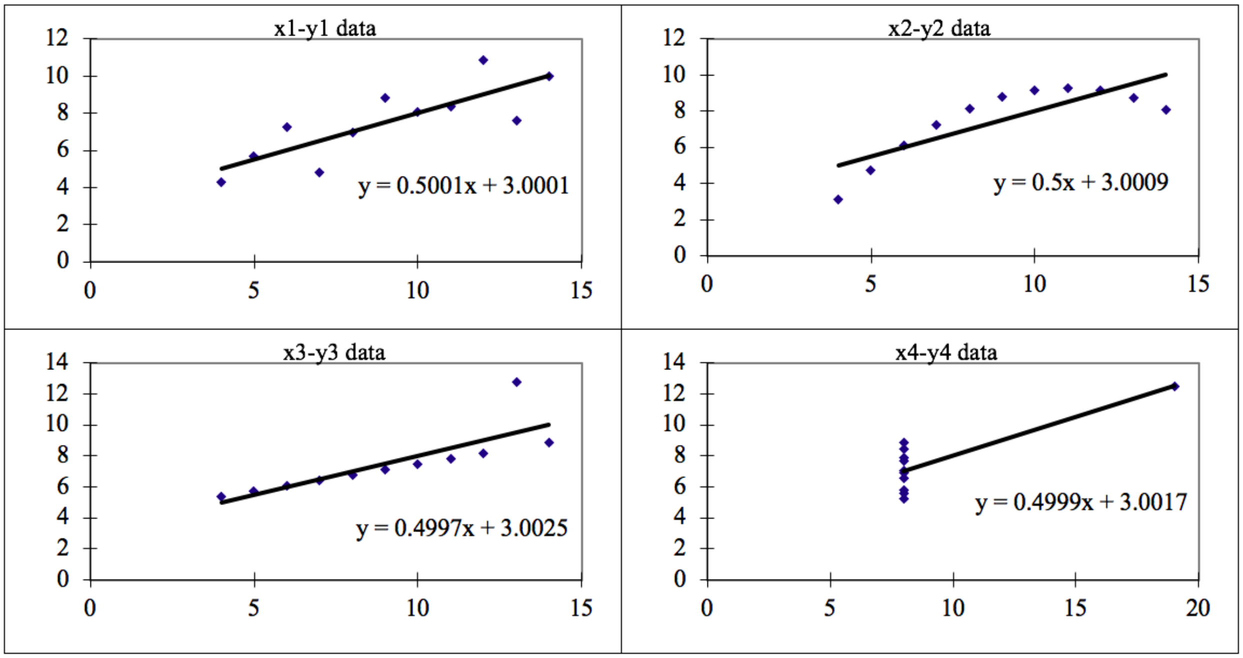
$$y = \theta_1 + \theta_2.x$$

While predicting the values for Y, it creates a distribution of error terms. Therefore, there are some assumptions for a linear regression model.

- Linear Relationship between X and Y
- Error terms have constant variance (Homoscedasticity)
- Error terms are independent of each other
- Absence of Multicollinearity
- Error terms are normally distributed with a mean value of zero

2. Explain the Anscombe's quartet in detail

Constructed in 1973 by the statistician Francis Anscombe; Anscombe's quartet comprises of four datasets that have nearly identical simple statistical properties, yet appear very different when graphed using scatter plots. This method describes the importance of data visualisation and how any regression algorithm can be fooled by the same. For example:
This data has similar means and standard deviations for all four datasets but the scatter plots are extremely different from each other.



3. What is Pearson’s R?

Pearson’s R or Pearson’s correlation coefficient is defined as the measurement of the strength of the relationship between two variables. It also measures their association with each other by analyzing the effect of change in one variable when the other variable changes. This relationship between two variables can be positive or negative. The Pearson's correlation coefficient varies between -1 and +1.

That is, if the correlation between two variables is 1, it means the data is perfectly linear with a positive slope i.e., both variables tend to change in the same direction. On the other hand, if the correlation is -1 means the data is perfectly linear with a negative slope i.e., both variables tend to change in different directions whereas, 0 means there is no linear association.

Pearson’s Correlation Coefficient equation:

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

Where;
r= correlation coefficient
xi= values of the x-variable in a sample
x= mean of the values of the x-variable

yi= values of the y-variable in a sample
y= mean of the values of the y-variable

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

Scaling is the procedure of measuring and normalizing the range of independent variables or features of data. It is also known as data normalization and it helps ensure that the gradient descent moves smoothly towards the minima. There are two types of scaling i.e., Normalization / MinMax Scaling and Standardization method.

Normalization is a scaling technique that shifts and rescales values so that they end up ranging between 0 and 1.

Formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

In python:

```
scaler = MinMaxScaler()
num_vars = ['Variables']

df_train[num_vars] = scaler.fit_transform(df_train[num_vars])
```

Standardization is another scaling technique where the values are equated using mean and a unit standard deviation. This means that the mean of the variable becomes zero and the resultant distribution has a unit standard deviation.

Formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

In python:

```
scaler = StandardScaler()
num_vars = ['Variables']
```

df_train[num_vars] = scaler.fit_transform(df_train[num_vars])

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

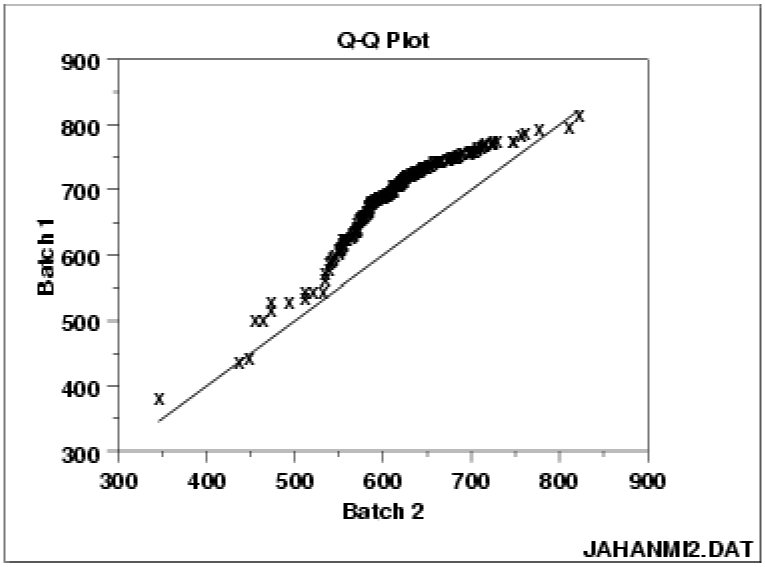
The variance inflation factor (VIF) quantifies the correlation between one predictor variable and the other predictor variables in a model. This method is used for checking if the model follows collinearity/multicollinearity between variables. The VIF for a variable is computed as:

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

Higher values of VIF signifies a greater correlation between variables. Specifically, values of more than 4 or 5 are regarded as being moderate to high and values of 10 or more being regarded as very high. If the VIF is infinite, it infers that the variables are perfectly correlated which makes it statistically insignificant.

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

Quantile- Quantile plot or a Q-Q plot is a graphical tool that helps in linear regression; specifically when we have separate training and test dataset. Using Q-Q plot, we can confirm that both the data sets are from populations with the same distributions.



This plot is used to check scenarios such as:

- If two data sets come from populations with a common distribution

- If two data sets have common location and scale
- If two data sets have similar distributional shapes
- If two data sets have similar tail behavior

A Q-Q plot is a plot depicting quantiles of the first data set against the quantiles of the second data set.