Student ID Number: 21151144

Assignment 1

Data Exploration and Analysis

Semester 2 2024

Student Name:
Student ID:
PAPER NAME: Data Analysis

PAPER CODE: COMP517

Due Date: Friday 30th August 2024 (midnight)

TOTAL MARKS: 100

INSTRUCTIONS:

1. **The following actions** may be deemed to constitute a breach of the General Academic **Regulations Part 7: Academic Discipline,**
   - Communicating with or collaborating with another person regarding the Assignment
   - Copying from any other student work for your Assignment
   - Copying from any third-party websites unless it is an open book Assignment
   - Uses any other unfair means
2. **Please email DCT.EXAM@AUT.AC.NZ if you have any technical issues with your Assessment/Assignment/Test submission on Canvas immediately**
3. **Attach your code for all the datasets in the appendix section**.

Table of Contents

1. Dataset

This report is based on the Bike Sharing dataset. This data set includes all the basic information about renting a bike. User is able to refer these bike sharing systems and rent a bike. Specifically, this dataset includes the characters that might affect the count of bikes. Any kind of details about a bike is recorded in these systems. Variables like weather conditions, precipitation, day of week, season can have an effect on the rental system. This dataset, compromises of the bike sharing information about two years.

The purpose of this dataset is to show which characteristics most affect the bike rentals. Characteristics like weather, temperature, seasons and day of the week are the main attributes we are testing to see how much of an effect they had on people renting bikes during the years 2011 and 2012.

We can use the Pandas library to load the Bike Sharing dataset into a DataFrame. From figure 1.1 you can see the first few rows of the data set. It has 15 variables. This includes 8 categorical data and 7 numerical data.

	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	\
0	1/01/2011	1	0	1	0	6	0	2	
1	2/01/2011	1	0	1	0	0	0	2	
2	3/01/2011	1	0	1	0	1	1	1	
3	4/01/2011	1	0	1	0	2	1	1	
4	5/01/2011	1	0	1	0	3	1	1	

	temp	atemp	hum	windspeed	casual	registered	count
0	0.344167	0.363625	80.120000	0.160446	331	654	985
1	0.363478	0.353739	0.696087	0.248539	131	670	801
2	0.196364	0.189405	0.437273	0.248309	120	1229	1349
3	0.200000	0.212122	0.590435	0.160296	108	1454	1562
4	0.226957	0.229270	0.436957	0.186900	82	1518	1600

Figure 1.1

```
dteday          object
season           int64
yr               int64
mnth             int64
holiday          int64
weekday          int64
workingday       int64
weathersit       int64
temp           float64
atemp          float64
hum            float64
windspeed      float64
casual           int64
registered       int64
count            int64
dtype: object
```

Figure 1.2

This dataset has 734 rows and 15 columns. If you look at figure 1.2, you can see the data types of the 15 variables. Most of them are integer and some are float. Normally, temperature, windspeed and humidity will have decimal values.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 734 entries, 0 to 733
Data columns (total 15 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   dteday      734 non-null    object
 1   season      734 non-null    int64
 2   yr          734 non-null    int64
 3   mnth        734 non-null    int64
 4   holiday     734 non-null    int64
 5   weekday     734 non-null    int64
 6   workingday  734 non-null    int64
 7   weathersit  734 non-null    int64
 8   temp        731 non-null    float64
 9   atemp       732 non-null    float64
 10  hum         734 non-null    float64
 11  windspeed   732 non-null    float64
 12  casual      734 non-null    int64
 13  registered  734 non-null    int64
 14  count       734 non-null    int64
dtypes: float64(4), int64(10), object(1)
memory usage: 86.1+ KB
None
```

Figure 1.3

General information about the variables is given figure 1.3.

2. Data Pre-processing

2a) Handling Missing Values

```
Missing values of the dataset:
dteday        0
season        0
yr            0
mnth          0
holiday       0
weekday       0
workingday    0
weathersit    0
temp          3
atemp         2
hum           0
windspeed     2
casual        0
registered    0
count         0
dtype: int64
```

Figure 2a.1

If you look at figure 2a.1, we can see that three variables have missing values. Let's look at the distribution of each of the variables with missing values and decide the best way to treat them.

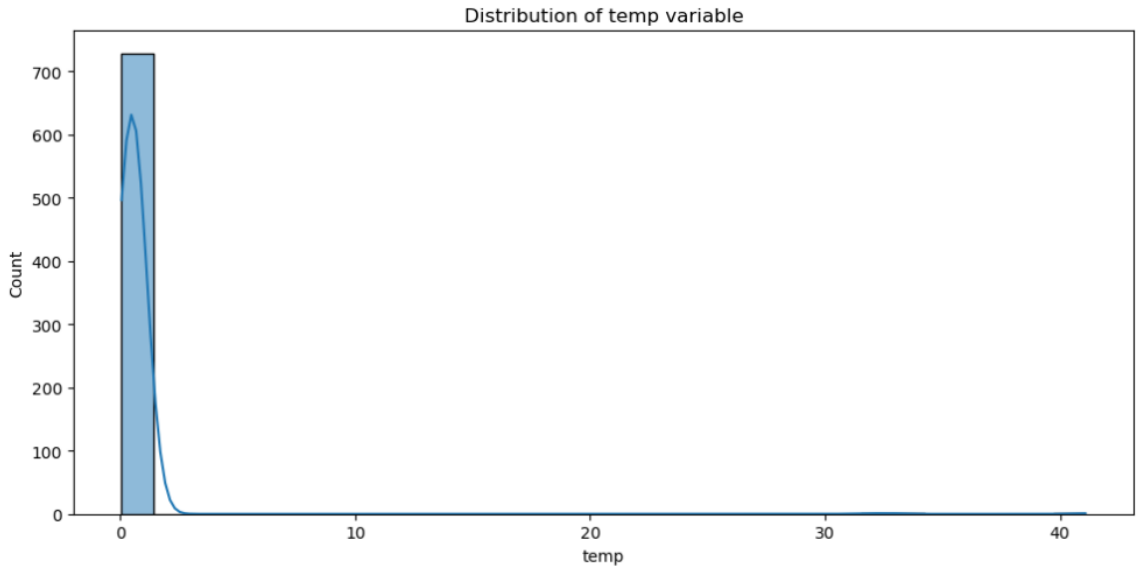


Figure 2a.2

Figure 2a.2 shows the distribution of the values in the temp variable. It is noted that there are few potential outliers in this variable. Due to that we cannot replace the missing values with the mean, because outliers pull the mean towards their extreme values. Therefore, the best way is to replace it with the median.

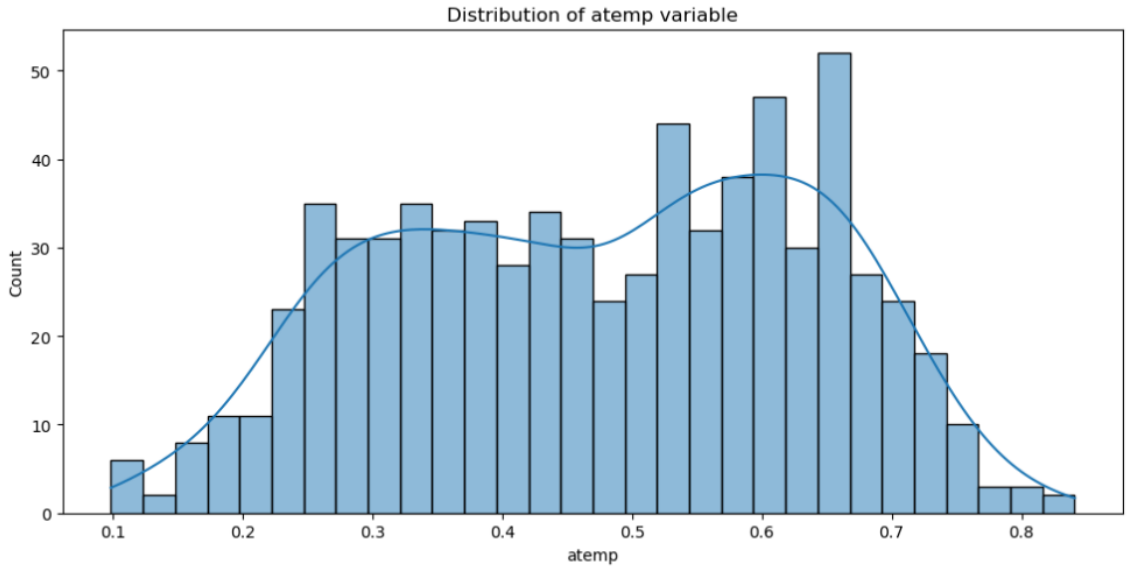


Figure 2a.3

Figure 2a.3, shows the distribution of the atemp variable This is normally distributed and there seems to be no extreme values. Therefore, we can easily use the mean to replace the missing values in it.

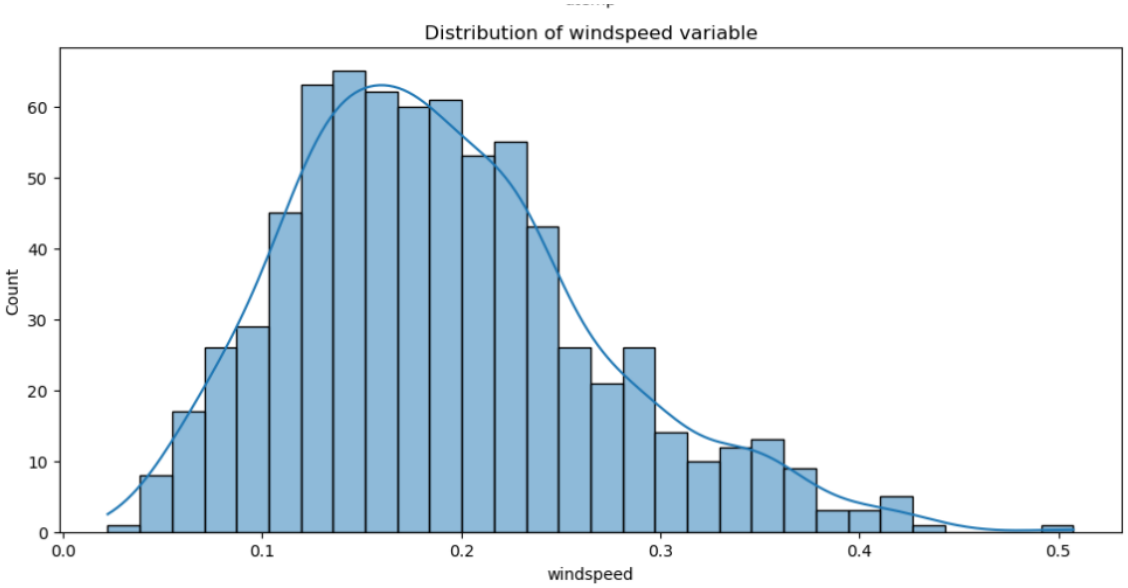


Figure 2a.4

Figure 2a.4, shows the distribution of the windspeed variable. This distribution is right Figure 2a.4, shows the distribution of the windspeed variable. This distribution is skewed and there are only two missing values. Therefore, we can impute the median for the missing data points.

After treating all the missing values, we can run the code to check for missing values again. Now you can see that there are no missing values. (Check for Appendix B)

2b) Handling Duplicates

```
         dteday  season  yr  mnth  holiday  weekday  workingday  weathersit
\
8     9/01/2011       1   0     1        0        0           0           1
606  29/08/2012       3   1     8        0        3           1           1
709  10/12/2012       4   1    12        0        1           1           2
731  10/12/2012       4   1    12        0        1           1           2
732   9/01/2011       1   0     1        0        0           0           1
733  29/08/2012       3   1     8        0        3           1           1

         temp     atemp       hum  windspeed  casual  registered  count
8    0.138333  0.116175  0.434167   0.361950      54         768    822
606  0.685000  0.635733  0.552083   0.112562    1177        6520   7697
709  0.435833  0.435575  0.925000   0.190308     329        4841   5170
731  0.435833  0.435575  0.925000   0.190308     329        4841   5170
732  0.138333  0.116175  0.434167   0.361950      54         768    822
733  0.685000  0.635733  0.552083   0.112562    1177        6520   7697
Number of duplicate rows: 6
```

Figure 2b.1

From figure 2b.1, we can see that three rows have been entered twice. This bike information is from the same date and has the same count. Therefore, we can think of them as errors occurred during data entry. These duplicated data can be removed by keeping one copy of each.

After removing the duplicates, if you run the same code that we used to get the duplicates, it will return as zero (check Appendix C).

2c) Handling Outliers

Z score of a data point tells us how many standard deviations a data point is from the mean of the distribution. Using the z score function from SciPy. Stats we can easily get the z scores of the numerical columns. To detect the outliers, we must set a threshold, usually any point which is greater than 3 standard deviations is considered as an outlier.

Using our threshold, we can filter the points with z scores greater than three and count them towards outliers. Figure 2c.1 shows how many outliers are present in this dataset.

```
Number of outliers in each variable:
temp: 3
atemp: 0
hum: 3
windspeed: 2
casual: 8
registered: 0
count: 0
```

Figure 2c.1

Now we will use boxplots to visualize the outliers in the dataset. Below you can see the variables with the outliers. All other variables show no outliers. (See Appendix D)

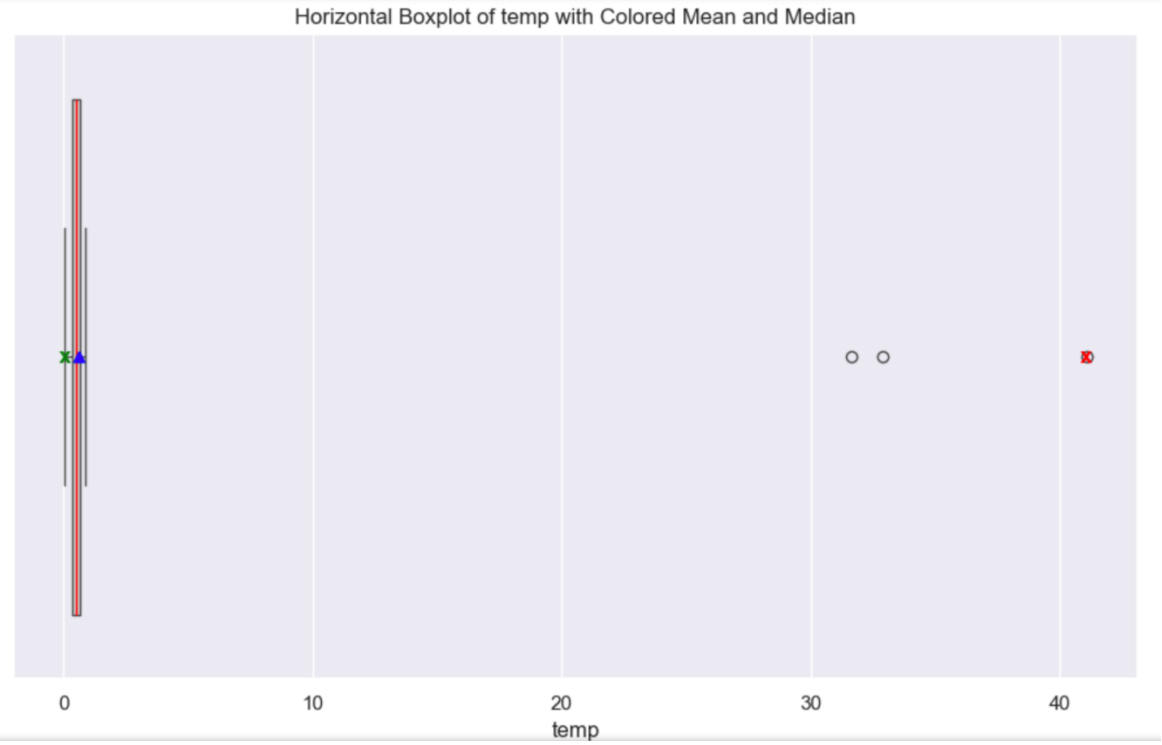


Figure 2c.2

Figure 2c.2 shows that there are three potential outliers in the temp dataset. From the z-score calculations also it showed as 3 outliers.

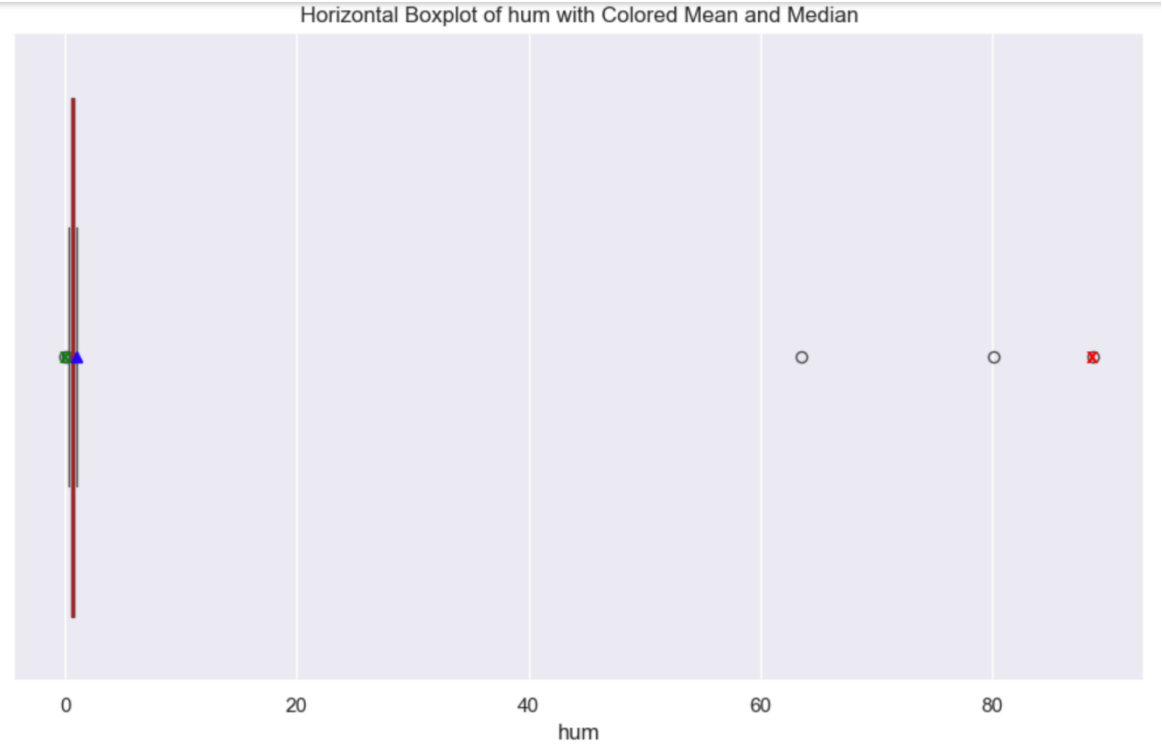


Figure 2c.3

Figure 2c.3 shows that humidity has 3 outliers as shown from the z-score calculations.

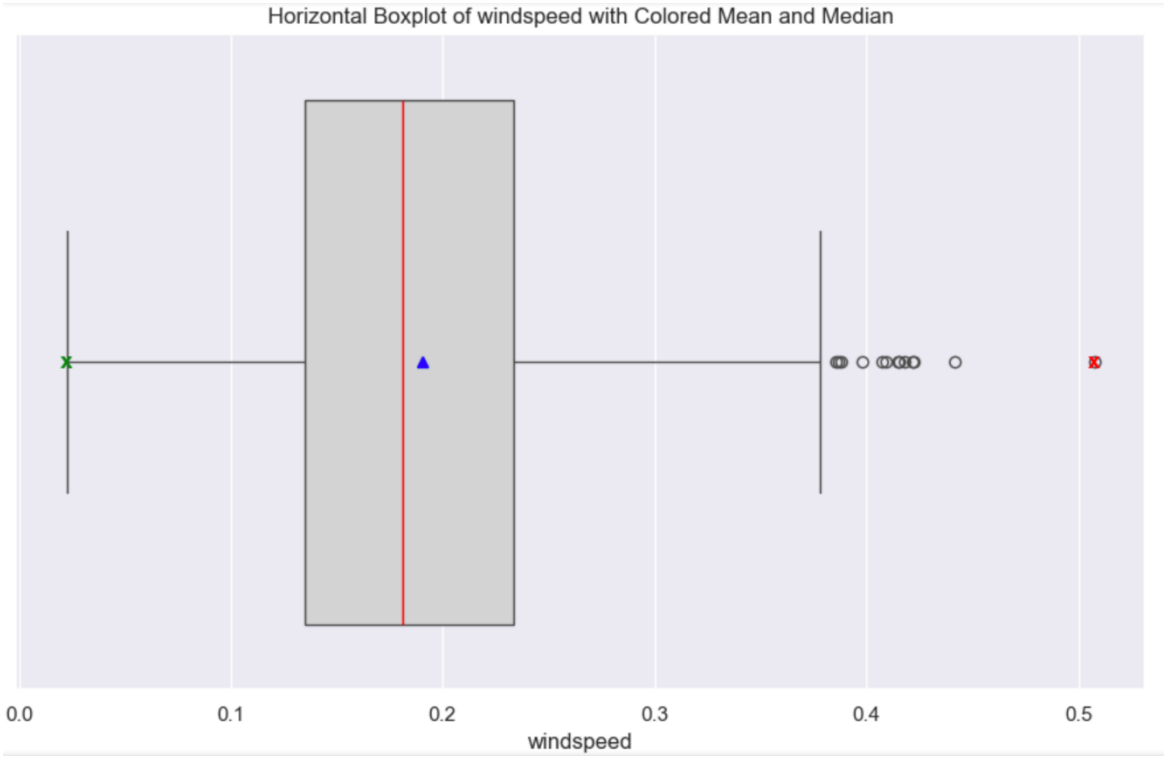


Figure 2c.4

Figure 2c.4 shows that there are several outliers in the windspeed variable. The z-score generated 2 outliers, but the boxplot shows more than two. If we carefully look at the boxplot, we can notice that some values are closer to the IQR, and cannot be considered exactly as outliers. Therefore, the small circles that are at the end of the box plot can only be considered as outliers.

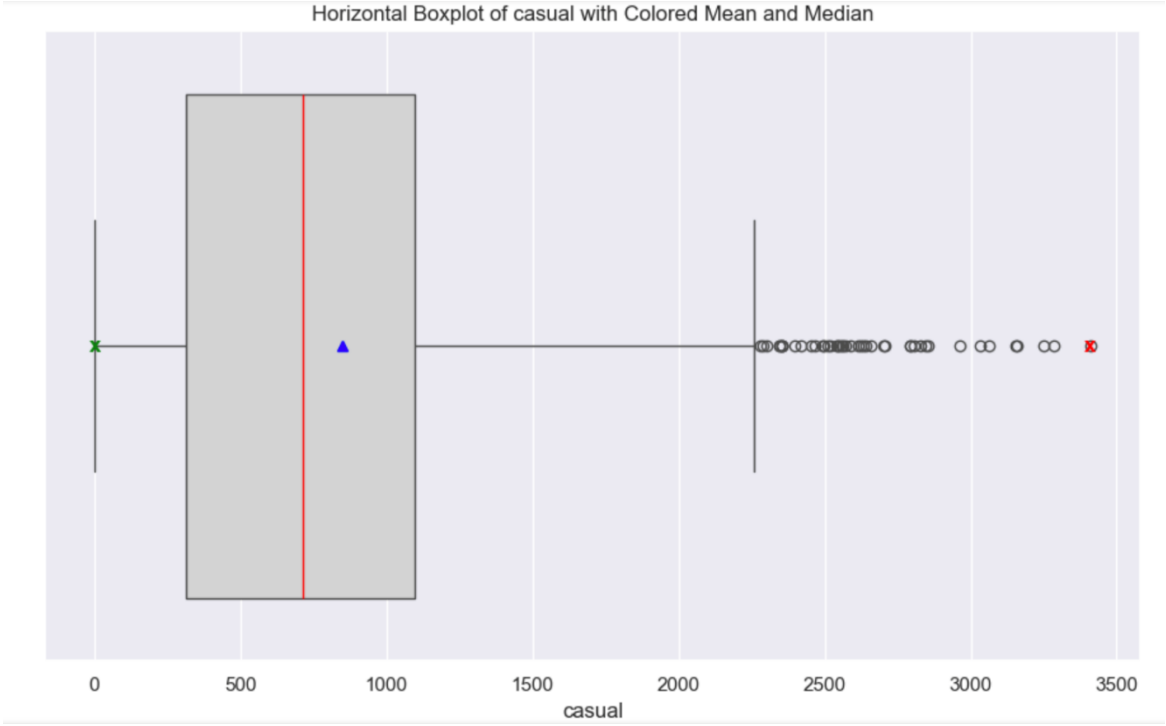


Figure 2c.5

Figure 2c.5 depicts that the casual variable too has some outliers. But if we take the casual variable, it shows the count of casual bike users. Therefore, the outliers shown in the plot does not mean that they are outliers. These points are simply high number of bike users on several days. Therefore, it is best to not remove the outliers in the casual variable.

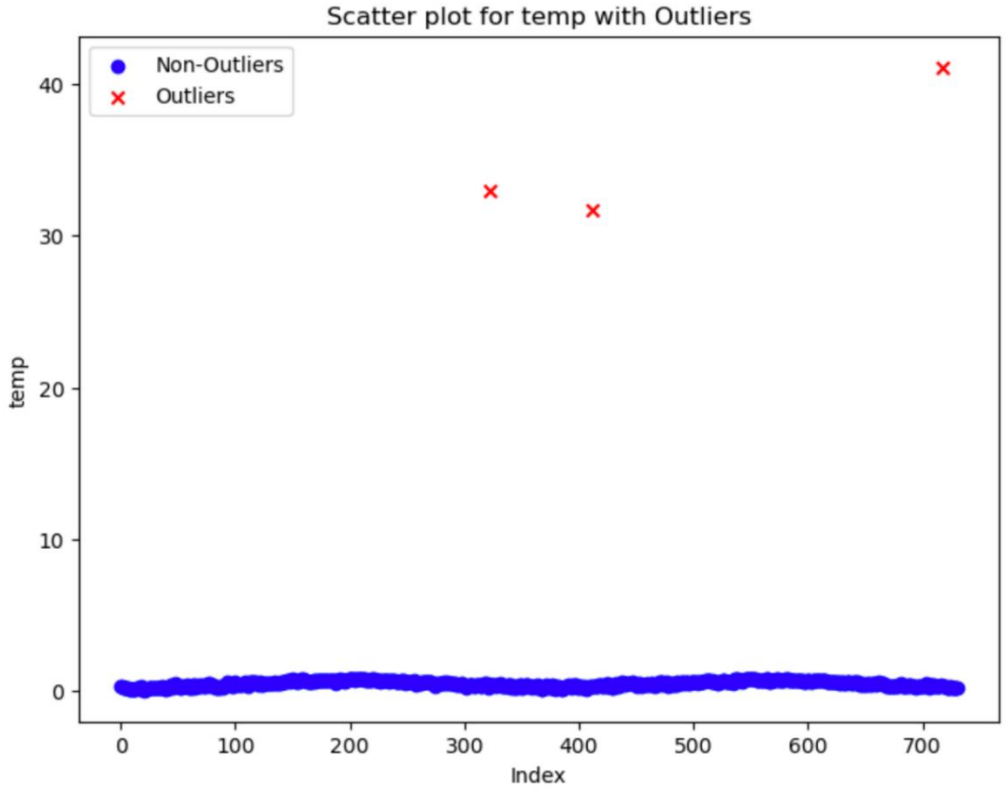


Figure 2c.6

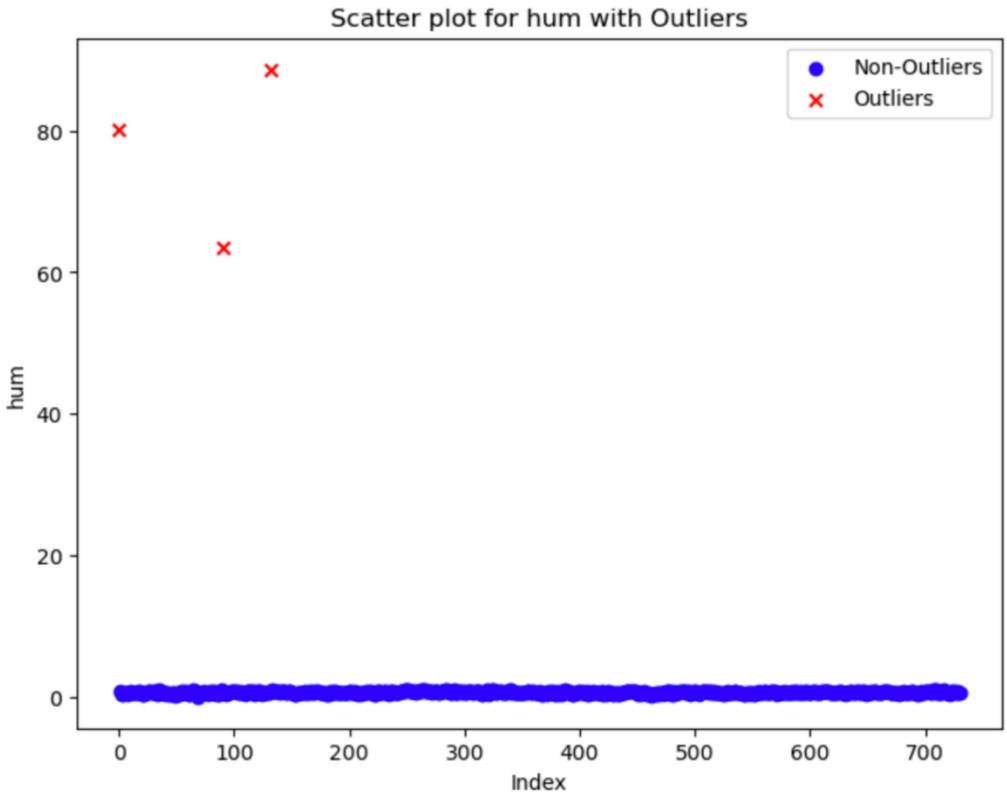


Figure 2c.7

Figure 2c.6 and 7, shows the scatter plots of temperature and humidity. The ‘x’ mark depicts the outliers.

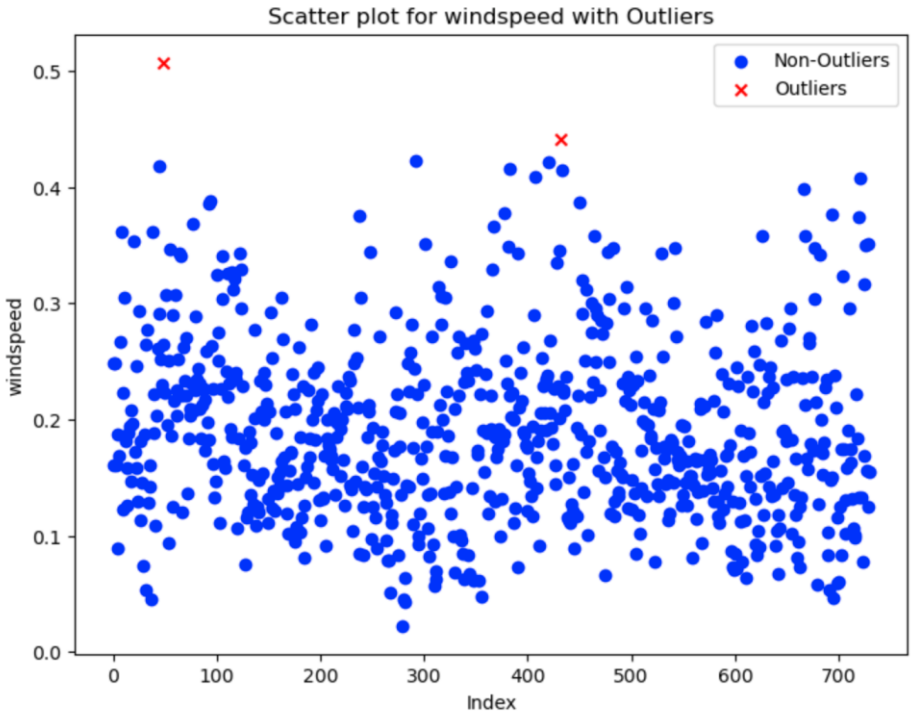


Figure 2c.8

As I said earlier, for the windspeed variable, only two data points are said to be 3 standard deviations away from the mean, which means all other data points that were shown as outliers in the box plot are simply high values.

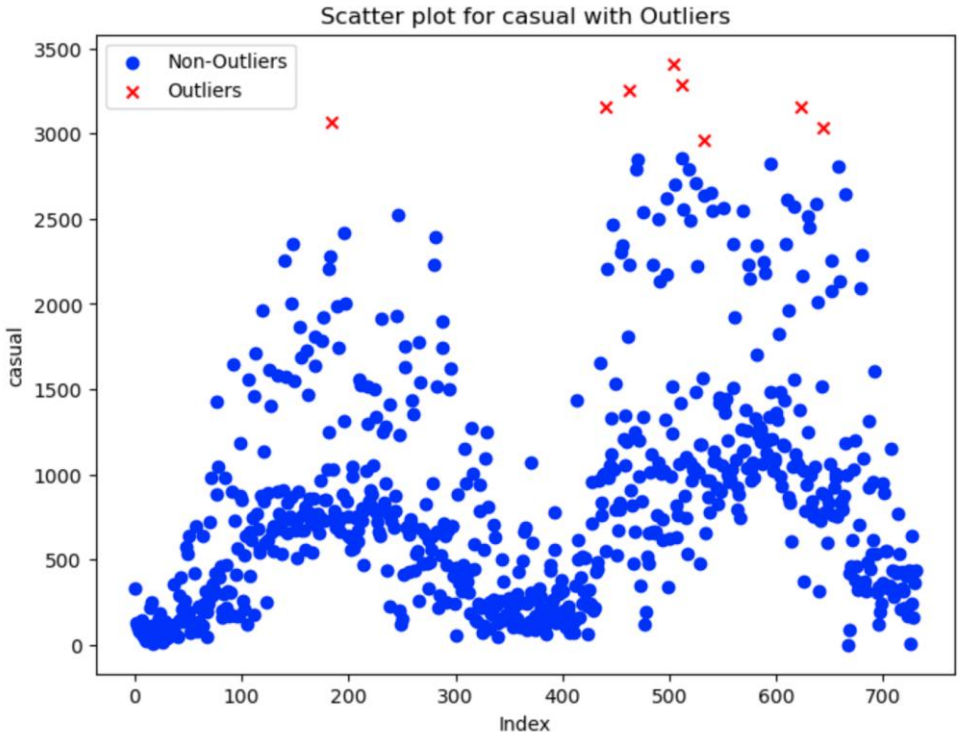


Figure 2c.9

Figure 2c.6,7,8 and 9 shows us the scatter plots of the four variables with outliers. From both boxplots and scatterplots, we can see that there are four variables that have outliers. They are temp, hum, windspeed and casual.

When we look at the scatter plots and the z-scores, for temp and hum variables we can see that there is a huge effect of outliers, this will affect the statistical calculations (mean, median, etc) therefore, its best to remove them, you can see in appendix E that no transformation will remove the outliers. The best way is to replace them with the median.

For the windspeed variable, we will use a square root transformation to transform the outliers. After replacing the outliers, the scatter plots are as below.

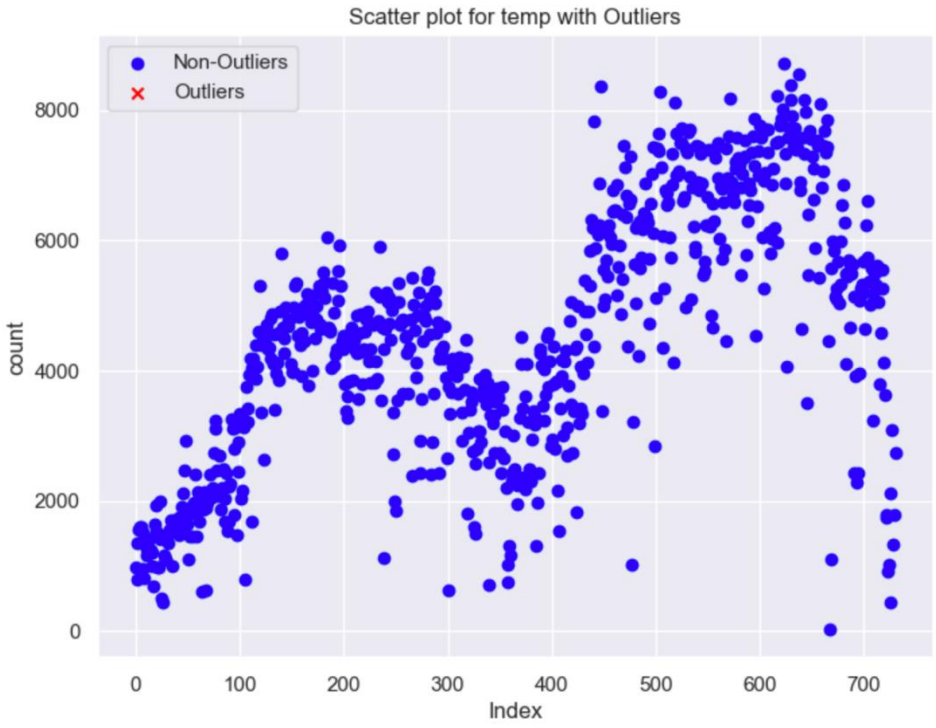


Figure 2c.10

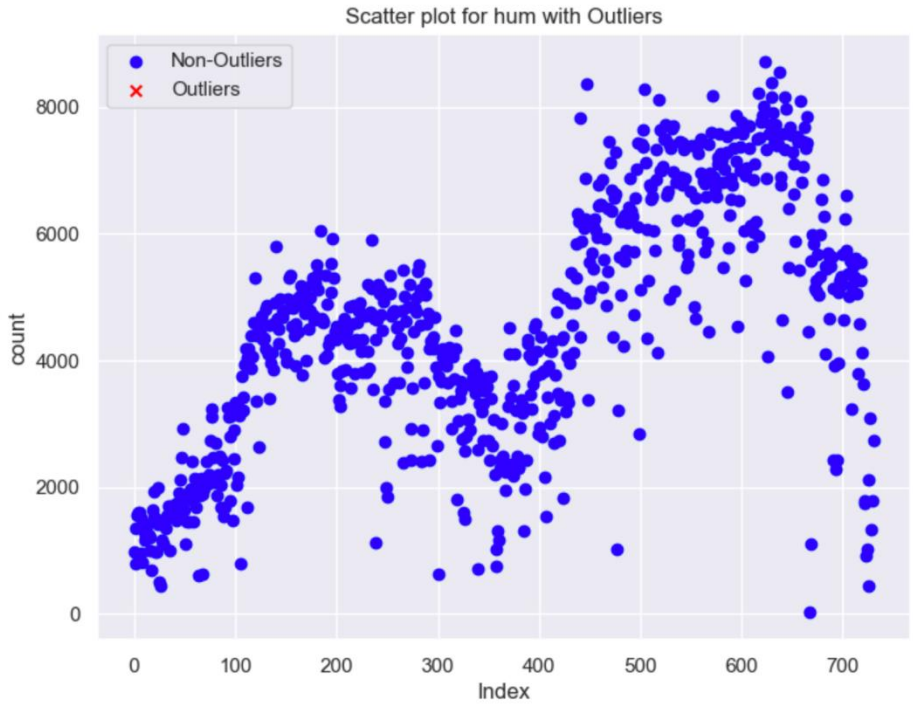


Figure 2c.11

Figure 2c.12

Now, all duplicates have been removed, missing values have been replaced and outliers have been treated. This dataset is said to be clean and ready to analysed.

3. Explore the Visualize Clean Dataset

	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual
count	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000	731.000000
mean	2.496580	0.500684	6.519836	0.028728	2.997264	0.683995	1.395349	0.495840	0.475197	0.628739	0.190401	848.176471
std	1.110807	0.500342	3.451913	0.167155	2.004787	0.465233	0.544894	0.182620	0.162097	0.139140	0.077351	686.622488
min	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	1.000000	0.059130	0.098839	0.254167	0.022392	2.000000
25%	2.000000	0.000000	4.000000	0.000000	1.000000	0.000000	1.000000	0.340000	0.338363	0.522291	0.134952	315.500000
50%	3.000000	1.000000	7.000000	0.000000	3.000000	1.000000	1.000000	0.502500	0.486733	0.625625	0.180975	713.000000
75%	3.000000	1.000000	10.000000	0.000000	5.000000	1.000000	2.000000	0.655000	0.608602	0.729583	0.233206	1096.000000
max	4.000000	1.000000	12.000000	1.000000	6.000000	1.000000	3.000000	0.861667	0.840896	0.972500	0.507463	3410.000000

registered	count
731.000000	731.000000
3656.172367	4504.348837
1560.256377	1937.211452
20.000000	22.000000
2497.000000	3152.000000
3662.000000	4548.000000
4776.500000	5956.000000
6946.000000	8714.000000

Figure 3.1

Figure 3.1, gives an overview of the statistical measures of the variables in the dataset. Below, you can see a detailed explanation of each measure.

- Count: The count of data points helps identify missing values. In this dataset there seems to be no missing values as they were being treated.

- Mean: the mean gives a central measure of the data. It shows the average of the data points.

- Standard Deviation: this shows us the spread of the data points. A high standard deviation means the values are more spread out, while a low standard deviation means they are closer to the mean. As you can see, count variable has the highest standard deviation, which means it is more spread out.

- Min and Max: Show the range of your data, indicating the minimum and maximum values present in the set of data values.

- Quartiles: These values divide the data into four parts. The 50th percentile is the median which provides the middle value, and the 25th and 75th percentiles show the spread of the middle 50% of the data. These values help you identify the skewness and outliers of the data.

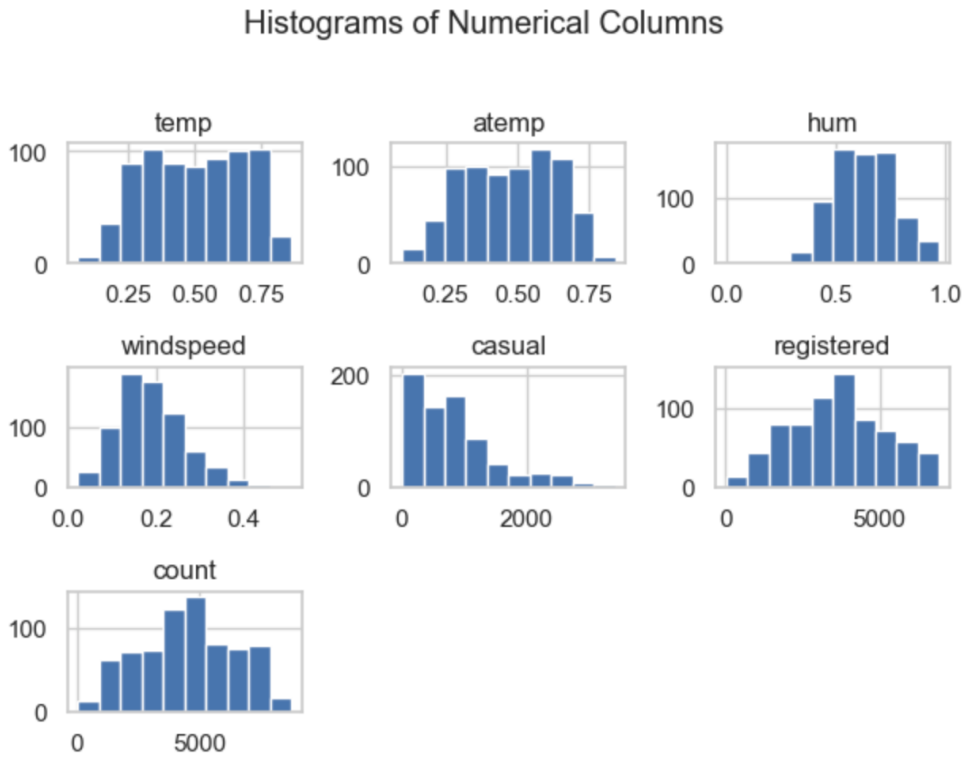


Figure 3.2

Figure 3.2 shows the distribution of the numerical variables in the dataset. Most columns seem to have a normal distribution while, casual and windspeed has a bit of a skewed distribution. Now let's see the distribution of some categorical data.

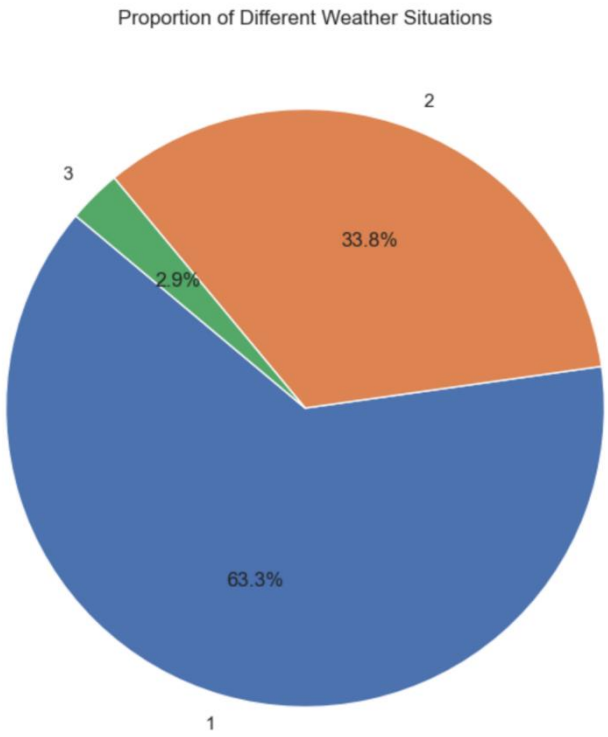


Figure 3.3

Figure 3.3 shows a pie chart of the weather situations.

1- Clear, Few clouds, partly cloudy, partly cloudy
2- Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3- Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

Weather type one has been prevailing for most of the days which has a percentage of 63.3% which is more than half of the time.

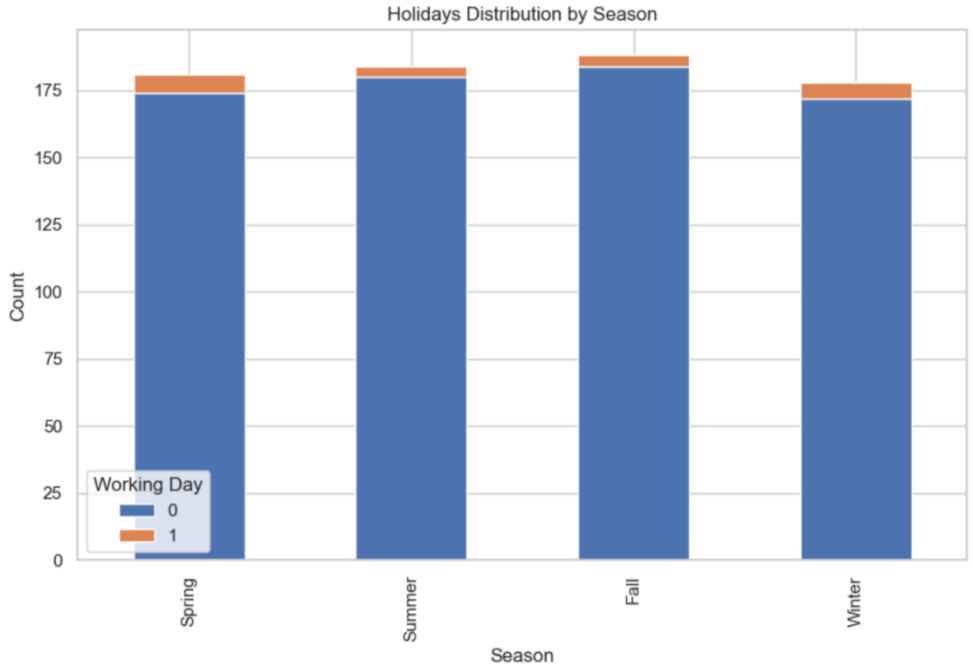


Figure 3.4

The stacked bar plot in figure 3.4 shows the distribution of holidays throughout each season of the year.

4. Multivariate Analysis

4a) Correlation Analysis

	temp	atemp	hum	windspeed	casual	registered	\
temp	1.000000	0.984251	0.129302	-0.158106	0.543044	0.540153	
atemp	0.984251	1.000000	0.138916	-0.181329	0.540979	0.538530	
hum	0.129302	0.138916	1.000000	-0.245192	-0.075426	-0.087258	
windspeed	-0.158106	-0.181329	-0.245192	1.000000	-0.166628	-0.216398	
casual	0.543044	0.540979	-0.075426	-0.166628	1.000000	0.395282	
registered	0.540153	0.538530	-0.087258	-0.216398	0.395282	1.000000	
count	0.627522	0.625483	-0.097013	-0.233349	0.672804	0.945517	

	count
temp	0.627522
atemp	0.625483
hum	-0.097013
windspeed	-0.233349
casual	0.672804
registered	0.945517
count	1.000000

Figure 4a.1

Figure 4a.1 gives us the correlation coefficients of each numerical variable. To get a clearer overview, we can create a heatmap.

Figure 4a.2

From the above heat map, you can see the correlation between each variable. The variables which have a coefficient close to 1, seems to have a strong linear relationship. The variables that have a coefficient close to 0, have a no correlation. You can see it well by looking at the coloured squares. If it is more towards red, that means it has a strong linear relationship.

For example, 'temp' and 'atemp' indicates a strong positive correlation as their coefficient is 0.98 which is very close to 1. This suggests that higher temperatures are strongly related with higher feeling temperatures. This can be expected as the feeling temperature is normally generated from the normal temperature.

'hum' and 'windspeed' have a negative correlation as their coefficient is -0.7 which is close to -1. This means as one variable increases, the other decreases. The literal meaning of this is low windspeed is related with high humidity.

Key insights:

- As observed strong correlation between temperature and apparent temperature means these two are related to each other. This makes sense as the apparent temperature is derived from the actual temperature.

- When considering humidity and windspeed, it is useful to understand their correlation as these patterns will have an effect on bike usage.

- We can also see that count has a strong positive correlation with temperature. This means that higher temperature leads to more bike usage.

4b) Perform multivariate analysis

The two categorical variables chosen for the multivariate analysis are 'season' and 'weathersit'.

	season	weathersit	count
0	1	1	2811.135135
1	1	2	2357.166667
2	1	3	934.750000
3	2	1	5548.548673
4	2	2	4236.705882

Figure 4b.1

Above figure shows the counts of bike rentals related to season and weather situations. Let's see the distribution of these factors in a bar plot.

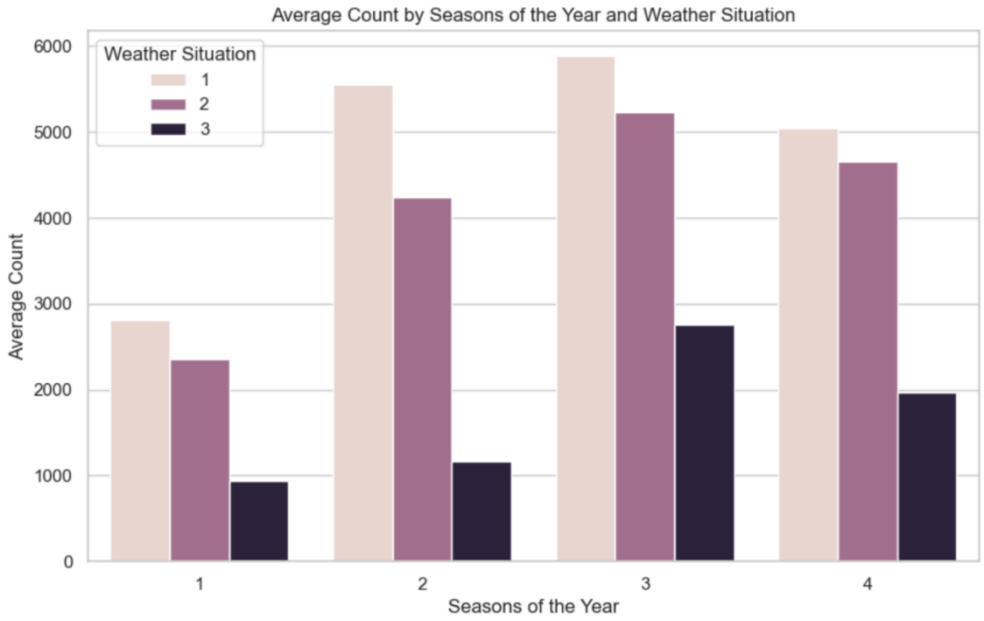


Figure 4b.2

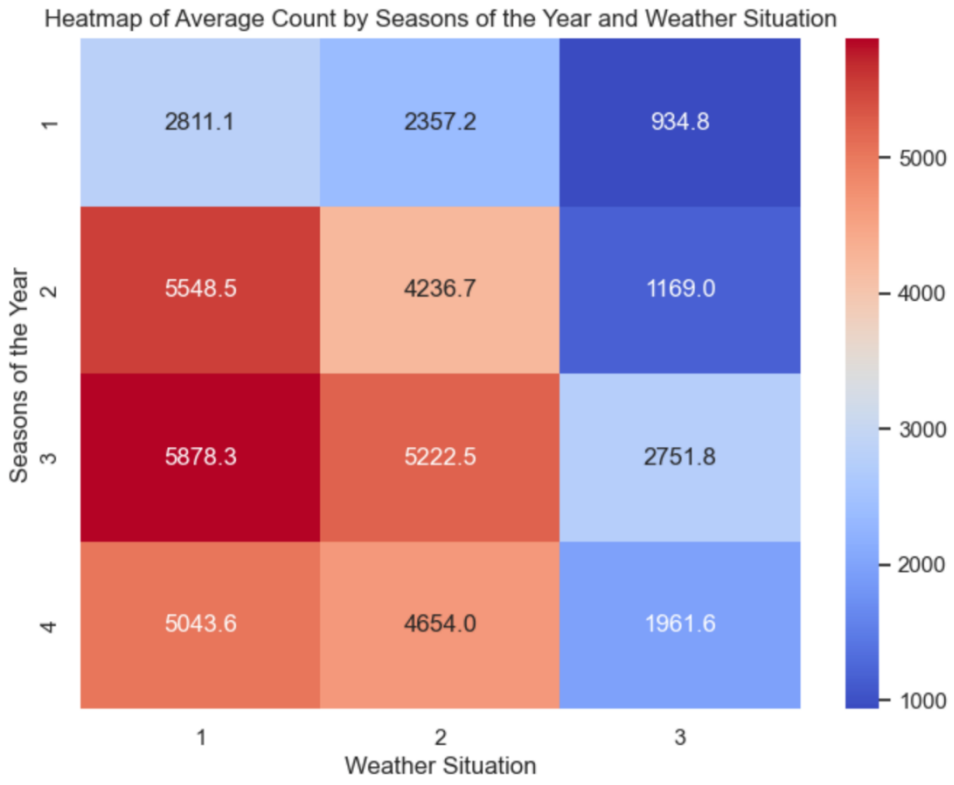


Figure 4b.3

From figure 4b.2 and 3, we can see the relationship between these two variables with count. As you can see the average count varies with different seasons. It is clear that during weather situation one which is clear and partly cloudy the average mean is high. This can be because during warmer climate people tend to rent more bikes. As well as during cold climates they show a less count of bike rentals.

If we look at the interactions between seasons and weather, we can see that despite of cold seasons people tend to rent more bikes as the weather is good. This means that weather conditions outweigh the seasonality.

4c) Perform aggregation analysis

	weathersit	mean	median
0	1	4876.786177	4844.0
1	2	4035.862348	4040.0
2	3	1803.285714	1817.0

Figure 4c.1

In figure 4c.1, you can see two common statistics; mean and median. When considering these two values, both are closer to each other, which means, that the data is symmetrically distributed with no skewness or outliers.

Weather Situation 1: The mean count is significantly higher than the median, this means that on some days the high rentals are making the average high making it skewed.

Weather Situation 2: The mean and median are very close; the bike rentals seem to be consistent.

Weather Situation 3: In this case, the mean is lower than the median, this suggests that there are some days with low rentals that is making the average lower but actually most values are relatively higher.

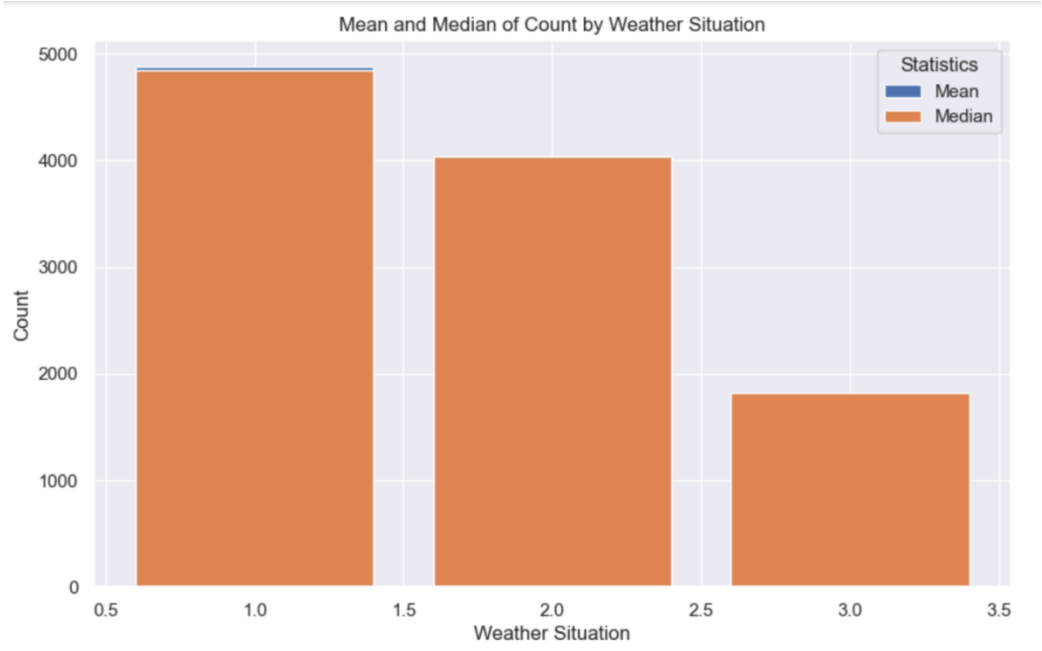


Figure 4c.2

Similar to what we saw from the table earlier, the graph confirms that the mean and the median are very close to each other.

From this, we can see that people are more favourable to rent bikes where the weather is warmer.

4d) Perform analysis to show how different 'season' affect the 'count' on average and how much variation exists using an appropriate plot.

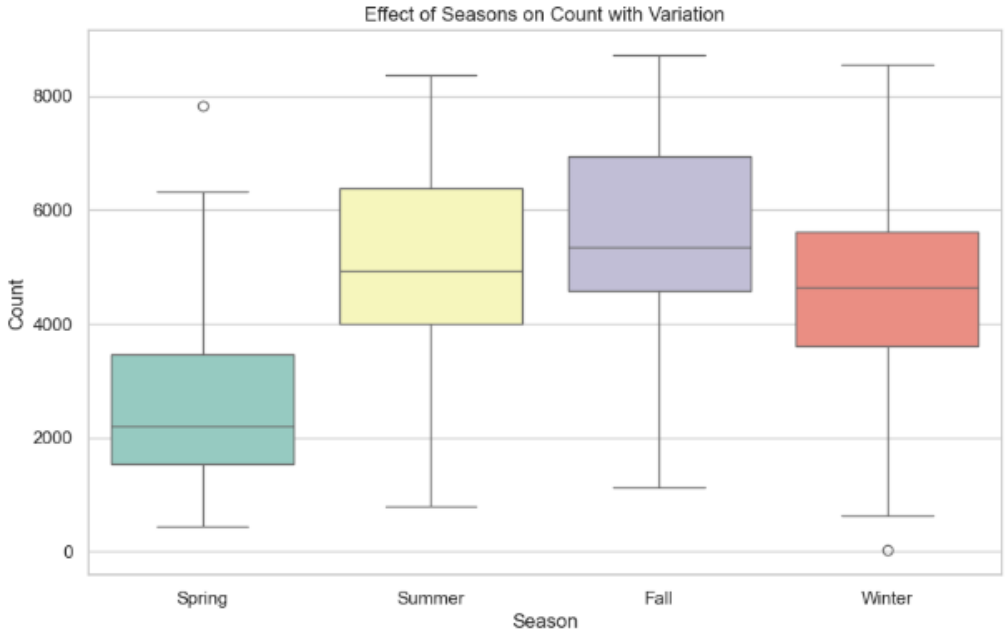


Figure 4d.1

The above figure shows the effect of different seasons on the count of bike rentals. As said earlier, despite of the seasons people tend to rent bikes. Generally, it is shown that the peak of bike rentals is in fall and the lowest in spring.

The reason why it is important to understand seasonal patterns with the count of bike rentals is that we know the amount of bike allocations needed for each season and staff we need during those peak times. You can see that in the winter there is a high interquartile range which means the bike rental count has a wider spread.

Another significance of performing this analysis is that, we can identify the variability of the data. As you can see a wide box with high variability suggests extreme scenarios that can be affected by weather conditions.

5) Conclusion

We will use a time series plot to identify the seasonal patterns over the two years. In the appendix, you will be able to see the adjusted dataframe according to the datetime system.

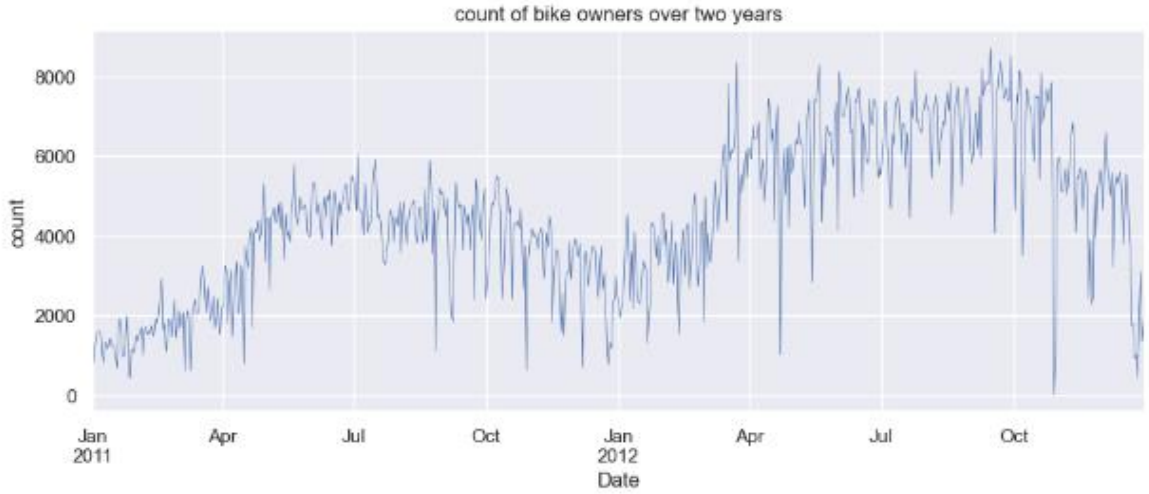


Figure 5.1

In figure 5.1, you can see the trend of bike rental count. We can see strong seasonal patterns in the plot above. There is an upward trend throughout the beginning of the year and a downward trend at the end of the year. Therefore, we can confirm that there are more bike rentals in the beginning of the year.

You can see lots of fluctuations as there are many factors affecting bike rentals. Using the rolling average, we can easily see the trend overtime.

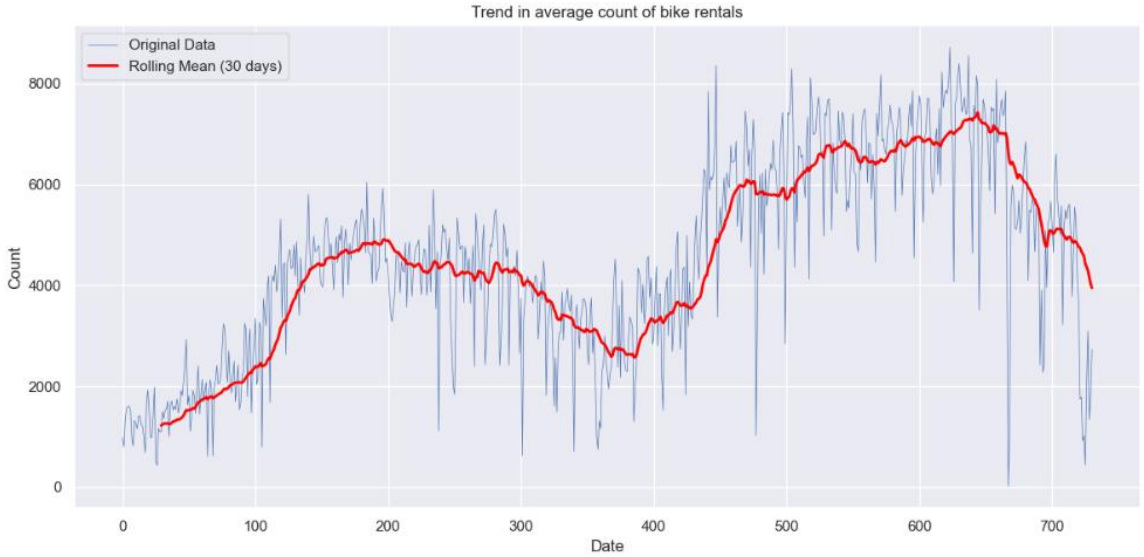


Figure 5.2

You can see as the mean is increasing overtime, there is a rising trend in bike rentals. As shown in the previous analysis, during the middle of the year, the highest bike rentals can be seen.

It is recommended that during the middle of the year bike sellers should focus on increasing staff and more stock of bikes. It should be taken into consideration that weather plays a crucial role in rental patterns.

The correlation matrix shows a strong positive correlation between temperature and bike rentals, this highlights the impact of climate in the bike rental process.

Bicycle rentals show significant seasonal patterns, with peak rentals in the beginning of the year and a notable decline at the end, according to the exploratory data analysis. The number of rentals is also significantly influenced by the weather; clear skies are positively correlated with higher rental numbers. The heatmap and box plot analysis brought attention to the distribution and variability of bike rentals, emphasizing the necessity for focused approaches to deal with variations brought on by the seasons and the weather.

Few challenges faced during the data analysis process are, missing values and outliers. Some of the variables where skewed and had many potential outliers. Considering the type of variable and its impact some data were replaced and some were transformed. These steps helped to maintain the accuracy and consistency of our data.

Further analysis might be required to build forecasting models. A time series analysis has already been done and a development of predictive models to forecast the bike rentals based on suggested factors is needed. The accuracy of these models can be improved by ensuring the data quality is enhanced. This can be done by referring to external data sources.

6) Appendix

Appendix A

Detailed overview of the variables in the Bike Sharing datatset is given below.

- **dteday** : date

- **season** : season (1:springer, 2:summer, 3:fall, 4:winter)

- **yr** : year (0: 2011, 1:2012)

- **mnth** : month (1 to 12)

- **holiday** : weather day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)

- **weekday** : day of the week

- **workingday** : if day is neither weekend nor holiday is 1, otherwise is 0.

+ **weathersit** :
- 1: Clear, Few clouds, Partly cloudy, Partly cloudy
- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

- **temp** : Normalized temperature in Celsius. The values are divided to 41 (max)

- **atemp**: Normalized feeling temperature in Celsius. The values are divided to 50 (max)

- **hum**: Normalized humidity. The values are divided to 100 (max)

- **windspeed**: Normalized wind speed. The values are divided to 67 (max)

- **casual**: count of casual users

- **registered**: count of registered users

- **count**: count of total rental bikes including both casual and registered

Appendix B

Output of the code to show the missing values have been treated.

```
dteday        0
season        0
yr            0
mnth          0
holiday       0
weekday       0
workingday    0
weathersit    0
temp          0
atemp         0
hum           0
windspeed     0
casual        0
registered    0
count         0
dtype: int64
```

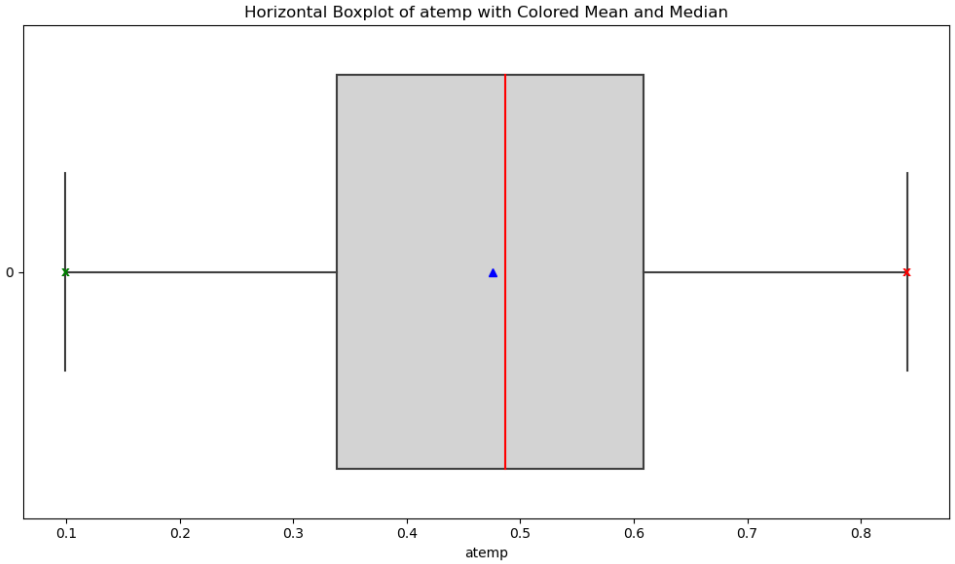
Appendix C

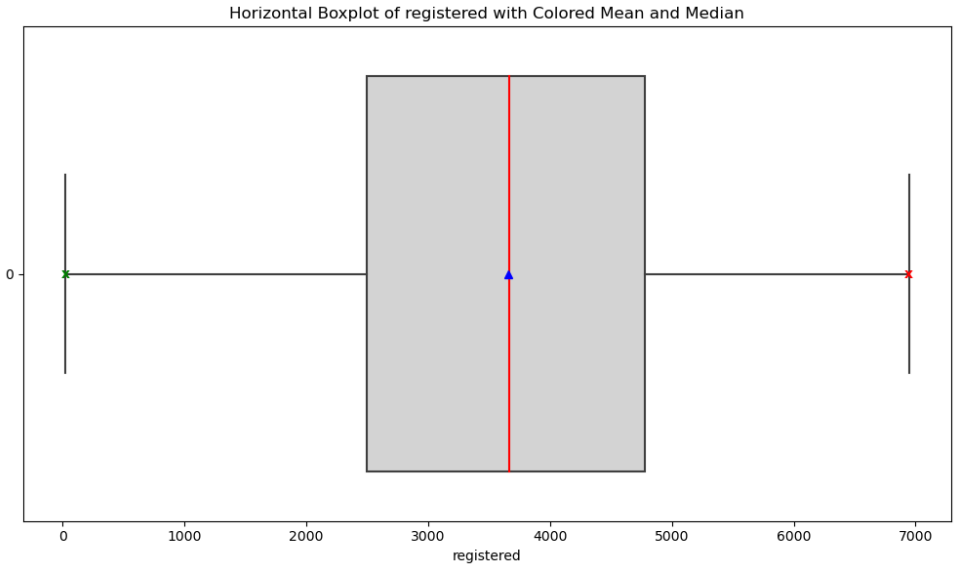
Output of the code to show the duplicated rows have been removed.

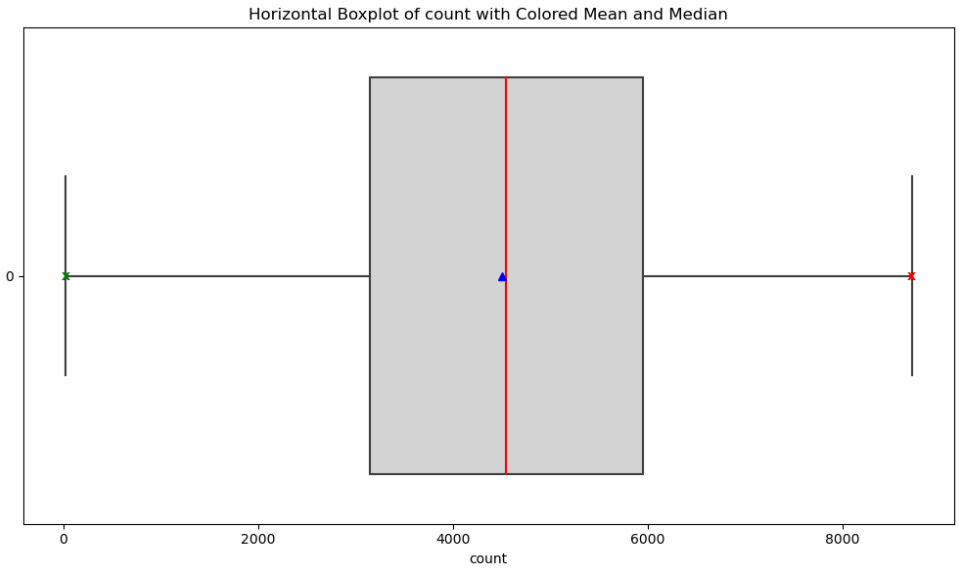
```
0      False
1      False
2      False
3      False
4      False
       ...
726    False
727    False
728    False
729    False
730    False
Length: 731, dtype: bool
Empty DataFrame
Columns: [dteday, season, yr, mnth, holiday, weekday, workingday, weathersit, t
emp, atemp, hum, windspeed, casual, registered, count]
Index: []
Number of duplicate rows: 0
```

Appendix D

Below shows the box plots of the variables with no outliers.

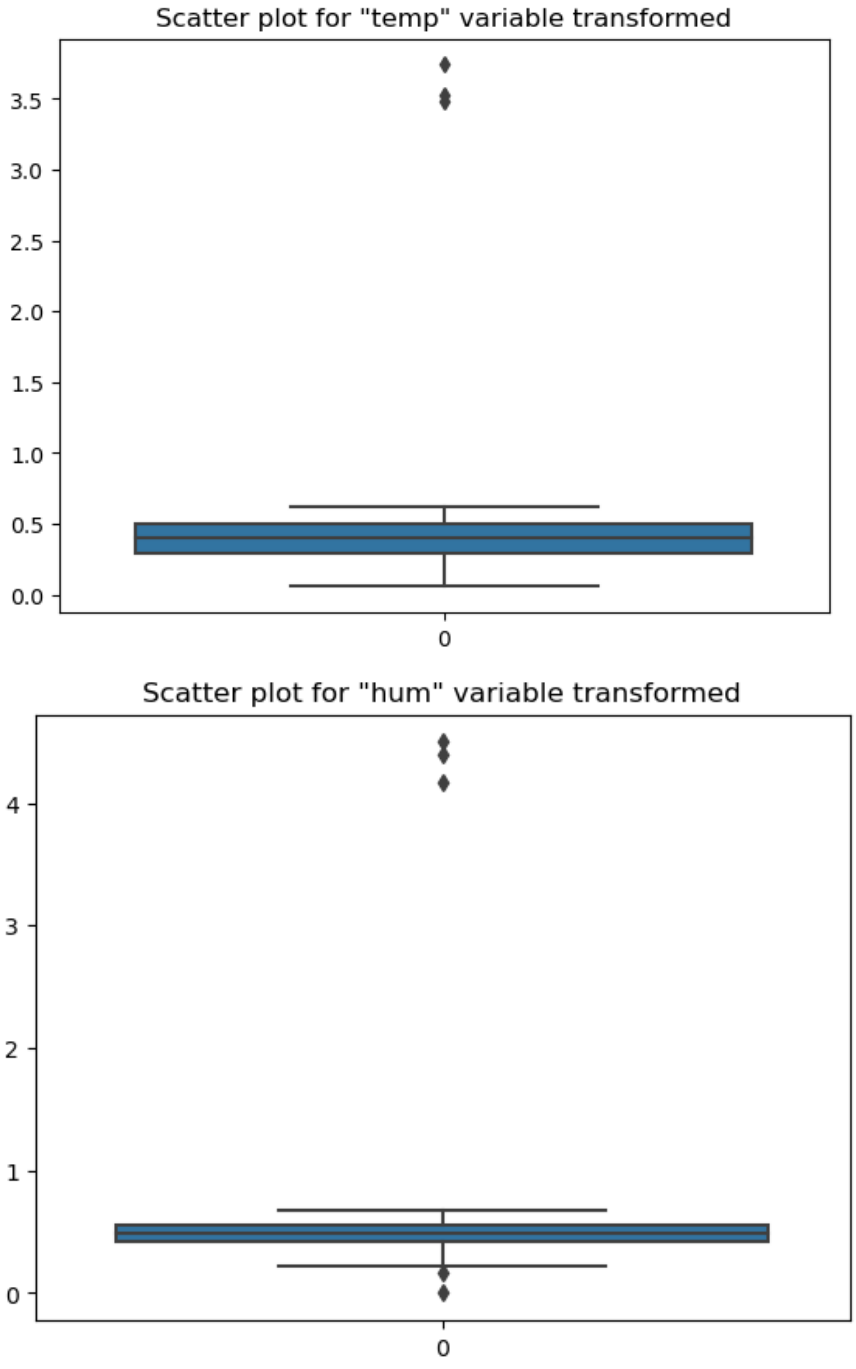






Appendix E

Below plot shows that transformations do not work for the ‘temp’ and ‘hum’ variable. Log transformations have been applied to both.



Appendix F

In the below figure, the 'dteday' column has been changed into the datetime format.

dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	count
2011-01-01	Spring	0	1	0	6	0	2	0.344167	0.363625	80.120000	0.160446	331	654	985
2011-01-02	Spring	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
2011-01-03	Spring	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
2011-01-04	Spring	0	1	0	2	1	1	0.200000	0.212122	0.590435	0.160296	108	1454	1562
2011-01-05	Spring	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.186900	82	1518	1600
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2012-12-27	Spring	1	12	0	4	1	2	0.254167	0.226642	0.652917	0.350133	247	1867	2114
2012-12-28	Spring	1	12	0	5	1	2	0.253333	0.255046	0.590000	0.155471	644	2451	3095
2012-12-29	Spring	1	12	0	6	0	2	0.253333	0.242400	0.752917	0.124383	159	1182	1341
2012-12-30	Spring	1	12	0	0	0	1	0.255833	0.231700	0.483333	0.350754	364	1432	1796
2012-12-31	Spring	1	12	0	1	1	2	0.215833	0.223487	0.577500	0.154846	439	2290	2729

731 rows × 14 columns