Student ID Number: _____

Assignment 1

Data Exploration and Analysis

Semester 2 2024

Student Name: Fatema Hashmi
Student ID: 23195092

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

PAPER NAME: Data Analysis 1
Appendix 1 5
Appendix 2 5
Appendix 3 5
Dataset 6
Data Pre-processing 9
a) Handling Missing Values: 9
b) Handling Duplicates: 10
c) Handling Outliers: 11
Explore the Visualize Clean Dataset 15
Multivariate Analysis 18
a) Correlation Analysis: 18
b) Multivariate analysis 20
c) Aggregation analysis 21
d) Average Analysis 22
Conclusion: 25
APPENDIX-1 26
APPENDIX - 2 27
APPENDIX – 3 32

LIST OF TABLES

Table 1 - The first few rows of the dataset 7
Table 2- Datatypes of the dataset 8
Table 3 - Shape of the dataset 8
Table 4 - Missing Values in the columns 9
Table 5 - Duplicate rows in the dataset 10
Table 6 - The Outliers Table giving the Inter Quartile Range 13
Table 7 - Summary Statistics for Numerical Columns 15
Table 8- Table of Correlations between numerical features 18
Table 9 - Aggregation Analysis of mean and median of count by weather situation 21
Table 10 - Analysis of Count by season with Variation 22

LIST OF FIGURES

Figure 1- Outlier Table and Graph (Histogram) of Temperature 11
Figure 2- Outlier Table and graph (Histogram) of Humidity 12
Figure 3 - Scatterplot of windspeed with Outliers and Non-Outliers 13
Figure 4 - Scatterplot of temperature with Outliers and Non-Outliers 14
Figure 5 - Boxplot of Count by Season - for numerical distributions 16
Figure 6 - Pie Chart of Weather Situations- for Categorical data 17
Figure 7 - Correlation Heatmap of Numerical Features 19
Figure 8 - Average Count by Season and Weather Situation via Multivariate Analysis 20
Figure 9- Bar graph of mean and median of count by weather situation 21
Figure 10 - Plot of average count by season with variation 23

Appendix 1

Appendix Table 1- Summary Statistics of the dataset 26
Appendix Table 2 - Filled the values with Imputation 27

Appendix 2

Appendix Table 2- 1 - To check if the duplicates are removed it has to return 0 27
Appendix Table 2- 4 Outlier table and graph (histogram) of Windspeed 28
Appendix Table 2- 5 Outlier table and graph (histogram) of Count 29
Appendix Table 2- 6 Outlier table and graph (histogram) of Registered Users 29
Appendix Table 2- 7 Outlier table and graph (histogram) of Casual Users 30

Appendix 3

Appendix Figure 3- 1 Plot histogram for 'count' -comparison of numerical distributions 33
Appendix Figure 3- 2 Density plot of Count by Weather Situation - for numerical distribution 33
Appendix Figure 3- 4- Bar plot of Season - for categorical data 34
Appendix Figure 3- 5- Bar graph of Count plot of Holidays- for categorical data 34

Dataset

Purpose of the Report

This report aims to conduct a comprehensive exploratory data analysis (EDA) of the Bike Sharing Dataset to gain a deep understanding of the factors influencing bike rental demand in the Capital Bikeshare system in Washington, D.C. By examining various environmental conditions, seasonal variations, and temporal patterns, we seek to uncover valuable insights that can inform the optimization of bike-sharing operations. Through this analysis, we aim to identify key factors that drive or inhibit bike rental usage , analyse the impact of seasonal variations on rental demand , explore the relationship between environmental conditions (e.g., temperature, precipitation) and bike rentals , identify temporal trends and patterns in rental usage , provide data-driven recommendations to improve resource allocation, enhance user experience, and contribute to sustainable transportation initiatives.

Short Summary of the Dataset

The dataset under analysis is a comprehensive record of bike rentals from the Capital Bikeshare system, spanning the years 2011 and 2012. This dataset not only tracks the hourly and daily count of bike rentals but also incorporates detailed environmental and seasonal information that is closely linked to rental behaviour. The core attributes of the dataset include temperature, humidity, wind speed, and weather conditions, which were sourced from online weather services. Additionally, the dataset captures the day of the week, month, and season, all of which play a crucial role in understanding rental patterns.

This rich dataset allows for a multifaceted exploration of how bike-sharing usage is influenced by external factors such as weather and time of year. For instance, the dataset can be used to examine how temperature fluctuations or the presence of precipitation impact the number of rentals, or how rental behaviour changes across different seasons. By analysing these aspects, the report aims to reveal underlying trends and correlations that can be leveraged to improve the efficiency and sustainability of bike-sharing systems.

To load the dataset and perform data analysis, I used Pandas for data manipulation, NumPy for numerical operations, Seaborn for visualizations, and Matplotlib for basic plotting and chart generation.

First few rows of the dataset:

	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	count
0	1/01/2011	1	0	1	0	6	0	2	0.344167	0.363625	80.120000	0.160446	331	654	985
1	2/01/2011	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
2	3/01/2011	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
3	4/01/2011	1	0	1	0	2	1	1	0.200000	0.212122	0.590435	0.160296	108	1454	1562
4	5/01/2011	1	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.186900	82	1518	1600

Table 1 - The first few rows of the dataset

Observations:

- **Data Distribution:** The dataset appears to have a relatively balanced distribution of data points across different seasons, years, months, and weekdays.
- **Categorical Variables**: The attributes season, yr, mnth, holiday, weekday, workingday, and weathersit are likely categorical variables, possibly represented by numerical codes.
- **Numerical Variables:** The attributes temp, atemp, hum, and windspeed are numerical variables, representing continuous measurements.
- **Range of Values:** The summary statistics provide information about the minimum, maximum, quartiles, and mean values for each attribute, giving a sense of the data's range and distribution.
- **Correlation Analysis:** Further analysis, such as correlation analysis, can reveal relationships between different attributes like examining the correlation between temp, atemp, and count could show how temperature affects bike rentals, analyzing the impact of workingday and holiday on count can highlight how these factors influence rental patterns and understanding correlations between weathersit and count could provide insights into how weather conditions affect bike usage.

```
...
    Datatypes of each column:
    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
```

Table 2- Datatypes of the dataset

```
..
    Shape of the dataset:
    (734, 15)
    The dataset has 734 rows and 15 columns.
```

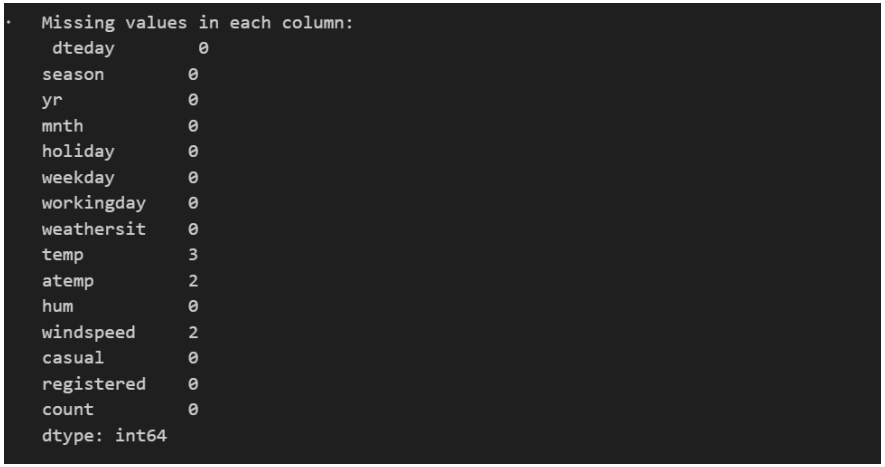
Table 3 - Shape of the dataset

Data Pre-processing

a) Handling Missing Values:

Given that the missing values in temp, atemp, and windspeed are relatively few compared to the total dataset size, **imputation** is a reasonable approach to handle them.
Reasons for Imputation: By imputing missing values, we can avoid losing valuable information that might be removed if we simply deleted rows with missing data. Imputation helps to ensure that the dataset remains consistent and can be used for further analysis without significant distortions.

Justification: These methods are straightforward and can provide reliable estimates for missing values, particularly when the data distribution is relatively balanced or skewed. Using the mean or median can help to mitigate the impact of outliers or extreme values that might distort the estimates.



```
Missing values in each column:
 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
```

Table 4 - Missing Values in the columns

To fill in missing values, use mean/median for numerical data and mode/random hot deck for categorical data. Consider data distribution and relationships between variables when choosing a method.

b) Handling Duplicates:

Duplicate rows:

	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	count
731	10/12/2012	4	1	12	0	1	1	2	0.435833	0.435575	0.925000	0.190308	329	4841	5170
732	9/01/2011	1	0	1	0	0	0	1	0.138333	0.116175	0.434167	0.361950	54	768	822
733	29/08/2012	3	1	8	0	3	1	1	0.685000	0.635733	0.552083	0.112562	1177	6520	7697

Table 5 - Duplicate rows in the dataset

Observations:

The dataset seems to monitor bike rentals over time, with each row representing data for a specific day. If each day should have only one record, duplicate rows might indicate data entry errors or redundancies. Keeping duplicates could skew results, especially when aggregating or averaging data. For example, duplicated rental counts might inflate daily, monthly, or seasonal totals, leading to incorrect conclusions. Removing duplicates ensures that each day has a single set of observations, providing a clearer and more accurate picture of trends over time.

Justification for Removing Duplicate Rows:

- Removing duplicates maintains the dataset's integrity by preventing artificial inflation or distortion of results caused by repeated data points.
- Ideally, each day should have only one entry, reflecting the unique conditions (like temperature, windspeed) and outcomes (like bike rental counts) for that day.

c) Handling Outliers:

Lower Bound: -0.1381255, Upper Bound: 1.1335425
Outliers based on the IQR method:

	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	count
322	19/11/2011	4	0	11	0	6	0	1	32.9167	0.324483	0.502083	0.224496	943	2720	3663
411	16/02/2012	1	1	2	0	4	1	2	31.6667	0.330162	0.752917	0.091425	74	2931	3005
717	18/12/2012	4	1	12	0	2	1	1	41.0833	0.409708	0.666250	0.221404	433	5124	5557

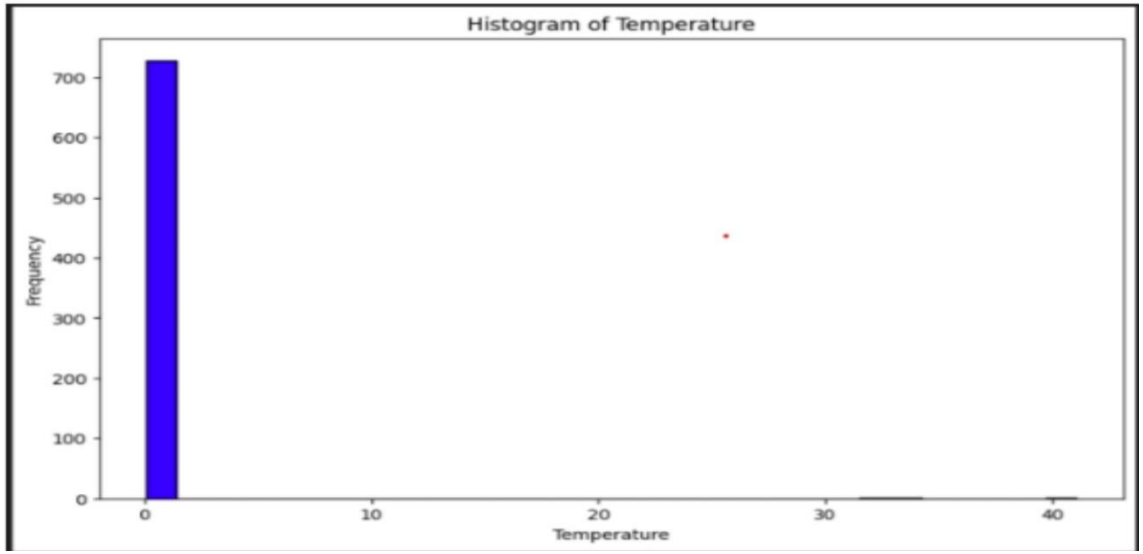


Figure 1- Outlier Table and Graph (Histogram) of Temperature

Lower Bound: 0.20343737500000014, Upper Bound: 1.0476043749999997
Outliers based on the IQR method:

	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	count
0	1/01/2011	1	0	1	0	6	0	2	0.344167	0.363625	80.120000	0.160446	331	654	985
49	19/02/2011	1	0	2	0	6	0	1	0.399167	0.391404	0.187917	0.507463	532	1103	1635
68	10/03/2011	1	0	3	0	4	1	3	0.389091	0.385668	0.000000	0.261877	46	577	623
91	2/04/2011	2	0	4	0	6	0	2	0.315000	0.315637	63.500000	0.197146	898	1354	2252
132	13/05/2011	2	0	5	0	5	1	2	0.512500	0.494300	88.660000	0.179725	692	3413	4105

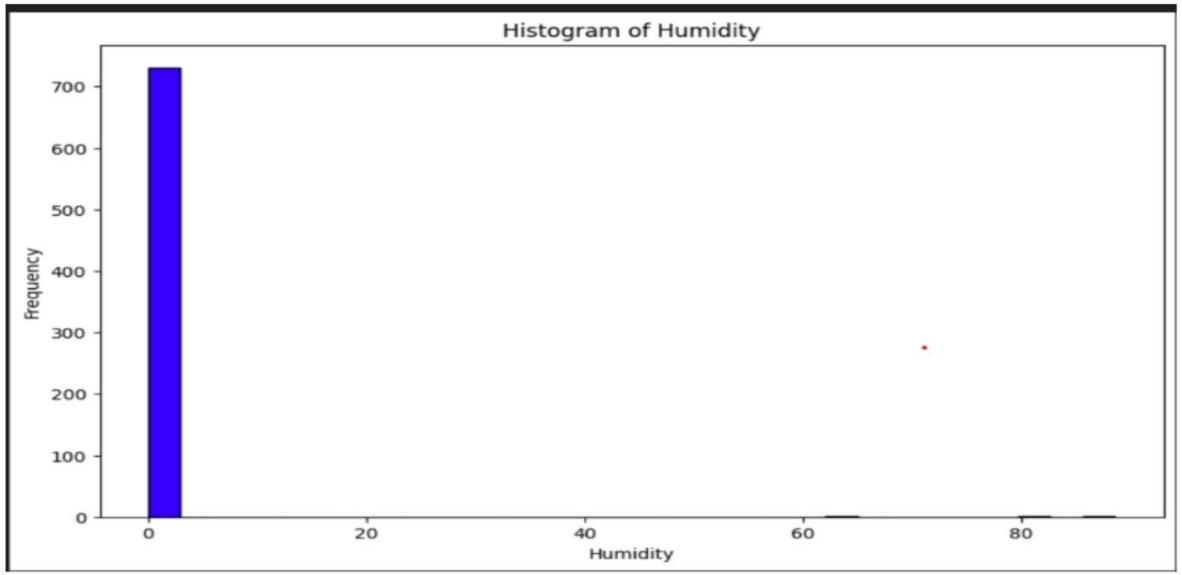


Figure 2- Outlier Table and graph (Histogram) of Humidity

```
windspeed - Lower Bound: -0.012441875000000047, Upper Bound: 0.38060312500000004
temp - Lower Bound: -0.1381255, Upper Bound: 1.1335425
hum - Lower Bound: 0.20343737500000014, Upper Bound: 1.0476043749999997
casual - Lower Bound: -856.625, Upper Bound: 2268.375
registered - Lower Bound: -959.875, Upper Bound: 8253.125
count - Lower Bound: -1082.75, Upper Bound: 10195.25
```

Table 6 - The Outliers Table giving the Inter Quartile Range

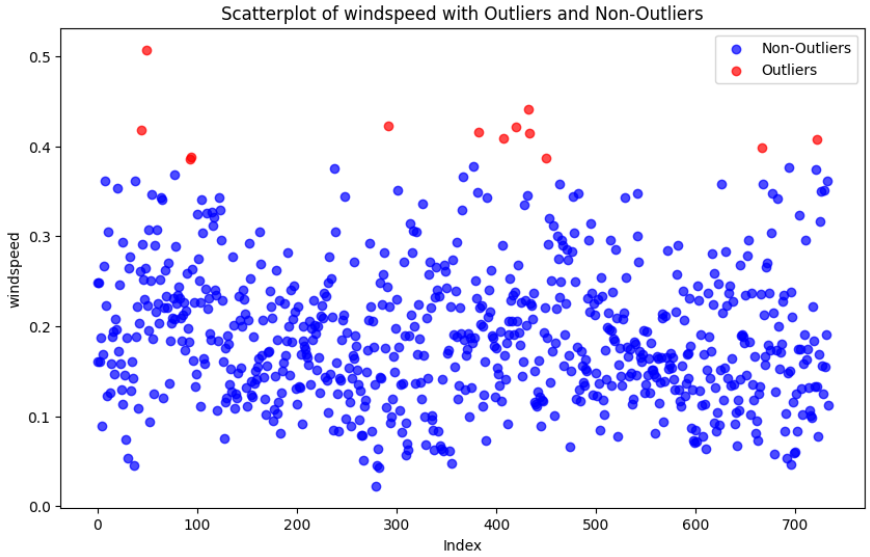


Figure 3 - Scatterplot of windspeed with Outliers and Non-Outliers

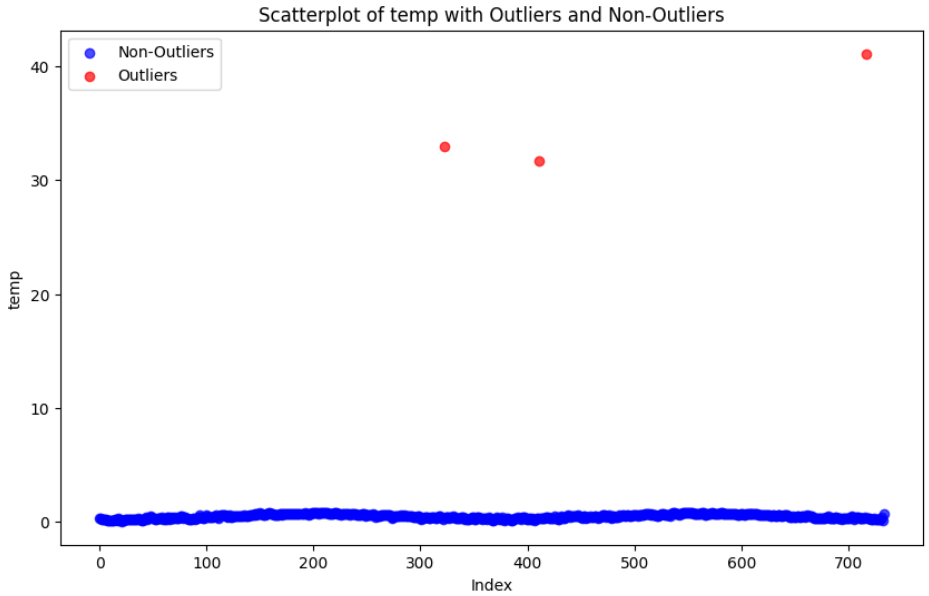


Figure 4 - Scatterplot of temperature with Outliers and Non-Outliers

Given the visual evidence of outliers and their potential impact on the analysis, it's advisable to remove or transform these outliers.
Justification: Outliers, while potentially valid, can introduce noise and bias. Removing them can improve data accuracy and model performance. Additionally, removing outliers can help maintain a clear data distribution.

The Z-score was chosen for normal distributions, while IQR was used for skewed data to accurately identify and handle outliers based on the data's characteristics.

Explore the Visualize Clean Dataset

	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual
count	734.000000	734.000000	734.000000	734.000000	734.000000	734.000000	734.000000	731.000000	732.000000	734.000000	732.000000	734.000000
mean	2.497275	0.501362	6.521798	0.028610	2.990463	0.683924	1.395095	0.637989	0.474873	0.941225	0.190555	846.835150
std	1.111453	0.500339	3.457233	0.166822	2.005087	0.465260	0.544628	2.245183	0.162644	4.950552	0.077610	686.217748
min	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	1.000000	0.059130	0.098839	0.000000	0.022392	2.000000
25%	2.000000	0.000000	4.000000	0.000000	1.000000	0.000000	1.000000	0.338750	0.338256	0.520000	0.134950	315.250000
50%	3.000000	1.000000	7.000000	0.000000	3.000000	1.000000	1.000000	0.502500	0.487364	0.626250	0.180975	711.500000
75%	3.000000	1.000000	10.000000	0.000000	5.000000	1.000000	2.000000	0.656667	0.609544	0.731042	0.233211	1096.500000
max	4.000000	1.000000	12.000000	1.000000	6.000000	1.000000	3.000000	41.083300	0.840896	88.660000	0.507463	3410.000000

Table 7 - Summary Statistics for Numerical Columns

Findings:

- The season variable's relatively low standard deviation suggests that data is evenly distributed across the seasons.
- With a mean of 0.68, this variable suggests that approximately 68% of the days are working days, while the rest are either weekends or holidays.
- The weathersit variable's mean suggests that the most common weather condition is between the first and second categories.
- The number of casual users shows considerable variability, with a high standard deviation (686.22) relative to the mean (846.84). The range of values (2 to 3410) indicates the presence of outliers or special events that significantly increased casual usage.
- After cleaning, the mean values for variables like temp, atemp, and windspeed shifted slightly, indicating that outliers or errors were addressed. The reduced standard deviation suggests a decrease in data variability, likely due to the removal of extreme values. Overall, these changes imply that the data is now more accurate and representative.

1. Comparing Numerical Distribution

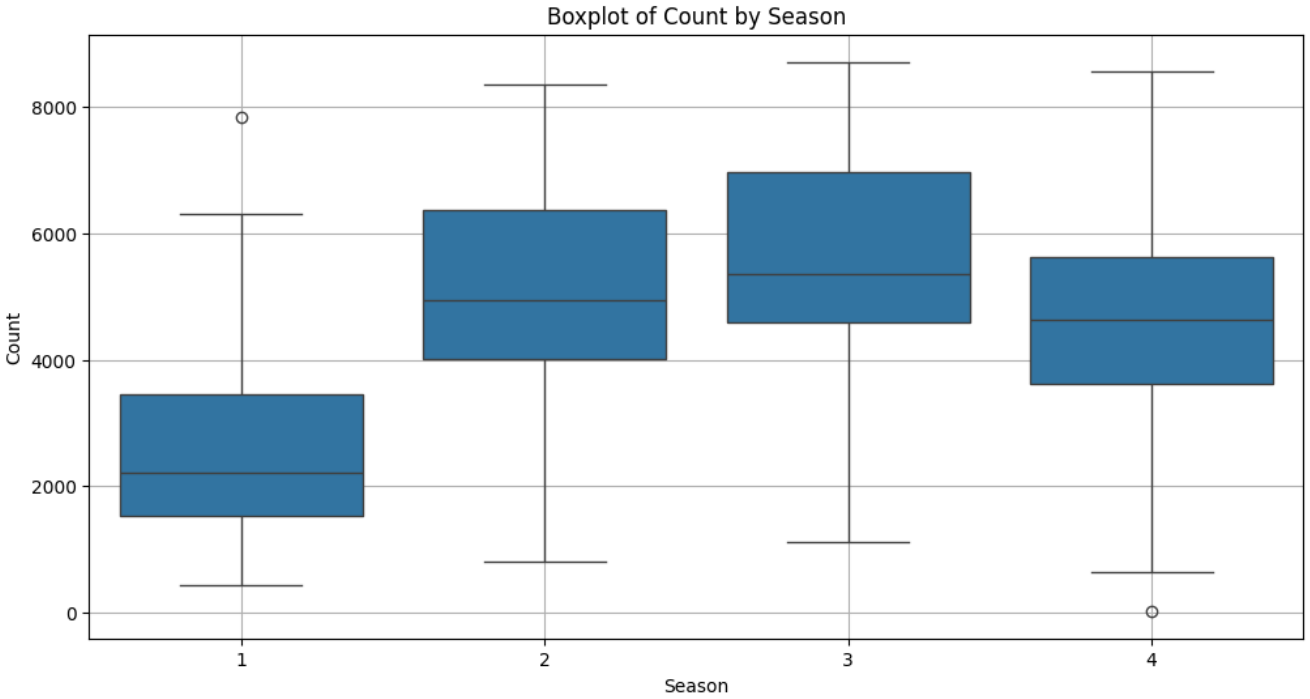


Figure 5 - Boxplot of Count by Season - for numerical distributions

Findings:

There is variation in the distribution of bike rental counts across seasons. Season 1 and 4 have slightly lower median counts compared to seasons 2 and 3. The IQRs for seasons 2 and 3 are slightly wider than for seasons 1 and 4, indicating a greater spread of rental counts in those seasons.A few outliers are observed in seasons 1 and 4, suggesting that there were a few days with unusually high or low rental counts in those seasons. Overall, the boxplot provides a clear visualization of the distribution of bike rental counts across seasons. The seasonal variations in median counts and IQRs suggest that factors related to seasons might influence bike rental demand.

2. Analysing Categorical Data

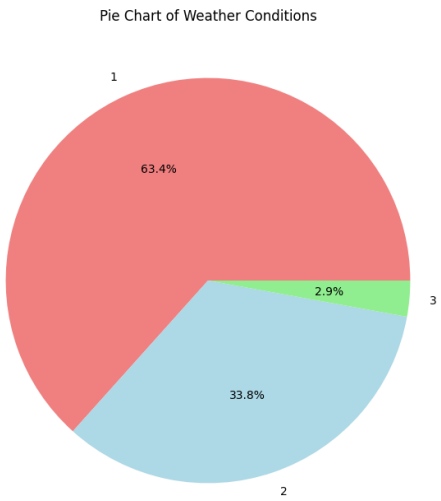


Figure 6 - Pie Chart of Weather Situations- for Categorical data

Findings:
Category 1 is the dominant category, representing 63.4% of the observations. Categories 2 and 3 account for 33.8% and 2.9% of the observations, respectively. The majority of observations fall into category 1, suggesting that this weather condition is the most common in the dataset. The distribution is skewed towards category 1. Categories 2 and 3 are relatively small, indicating that they occur less frequently. Overall, the pie chart effectively visualizes the distribution of the categorical variable, providing insights into the prevalence of different weather conditions in the dataset.

Multivariate Analysis

a) Correlation Analysis:

	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered
season	1.000000	0.002460	0.832405	-0.010616	-0.002154	0.016800	0.022065	0.062253	0.340506	-0.041962	-0.232201	0.210567	0.414402
yr	0.002460	1.000000	0.003138	0.007705	-0.004747	0.001852	-0.047036	0.027873	0.044986	-0.066557	-0.015422	0.248881	0.596103
mnth	0.832405	0.003138	1.000000	0.019023	0.010559	-0.000800	0.046863	0.058352	0.224723	-0.051630	-0.211254	0.123552	0.297696
holiday	-0.010616	0.007705	0.019023	1.000000	-0.101148	-0.252448	-0.034491	-0.013259	-0.033096	-0.011327	0.006127	0.054527	-0.108367
weekday	-0.002154	-0.004747	0.010559	-0.101148	1.000000	0.037712	0.030939	0.027688	0.002289	0.081552	0.009932	0.063086	0.059696
workingday	0.016800	0.001852	-0.000800	-0.252448	0.037712	1.000000	0.062788	0.005179	0.053728	-0.040334	-0.024620	-0.514135	0.307846
weathersit	0.022065	-0.047036	0.046863	-0.034491	0.030939	0.062788	1.000000	-0.020703	-0.121211	0.087472	0.040942	-0.247082	-0.257516
temp	0.062253	0.027873	0.058352	-0.013259	0.027688	0.005179	-0.020703	1.000000	0.035358	-0.006236	-0.019235	0.010920	0.046702
...													
windspeed	-0.232201	-0.015422	-0.211254	0.006127	0.009932	-0.024620	0.040942	-0.019235	-0.189440	-0.017514	1.000000	-0.170585	-0.223780
casual	0.210567	0.248881	0.123552	0.054527	0.063086	-0.514135	-0.247082	0.010920	0.543311	-0.022816	-0.170585	1.000000	0.396058
registered	0.414402	0.596103	0.297696	-0.108367	0.059696	0.307846	-0.257516	0.046702	0.543351	-0.073714	-0.223780	0.396058	1.000000
count	0.408388	0.568365	0.283582	-0.068065	0.070405	0.066404	-0.294855	0.041495	0.630481	-0.067470	-0.240839	0.672590	0.945886

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

Table 8- Table of Correlations between numerical features

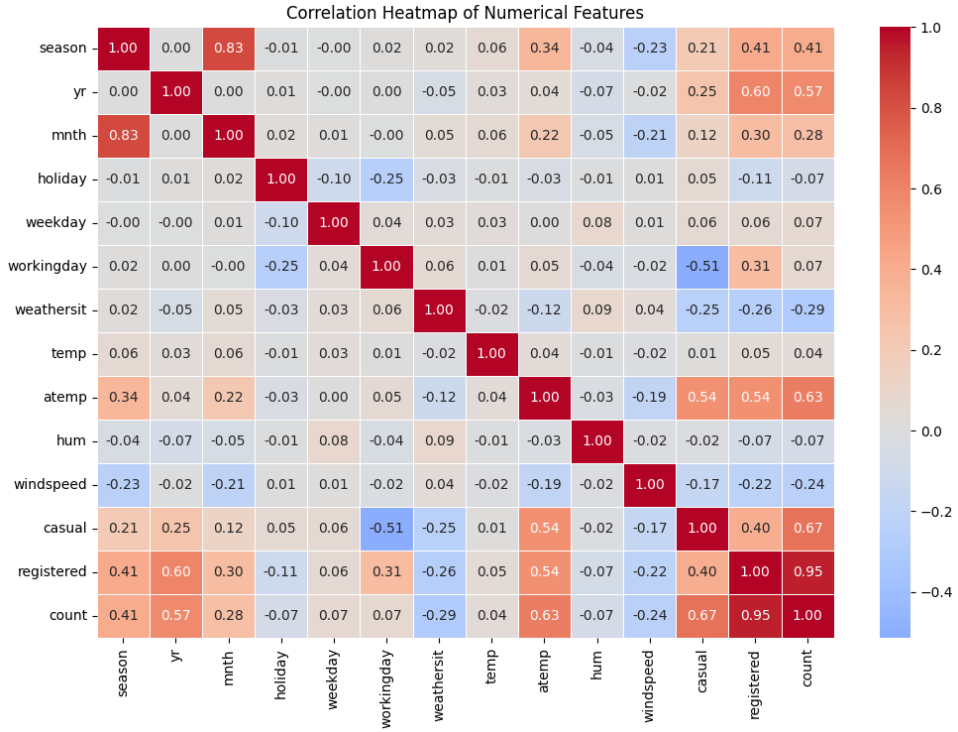


Figure 7 - Correlation Heatmap of Numerical Features

Analysing the Correlation Heatmap

The correlation heatmap visually represents the strength and direction of relationships between numerical variables in the dataset. The color scale indicates the correlation coefficient, ranging from -1 (strong negative correlation) to 1 (strong positive correlation).

1. **Positive Correlation Between registered and count:** The strong positive correlation between registered and count suggests that as the number of registered users increases, the overall bike rental count also tends to increase. This indicates a strong association between these two variables.

2. **Negative Correlation Between workingday and casual:** The negative correlation between workingday and casual suggests that on working days, the number of casual rentals tends to be lower compared to non-working days. This might imply that casual users are more likely to rent bikes on weekends or holidays.
3. **Weak Correlation Between temp and windspeed:** The near-zero correlation between temp and windspeed indicates a weak or negligible relationship between these two variables. This suggests that changes in temperature do not have a strong influence on wind speed in this dataset.

b) Multivariate analysis

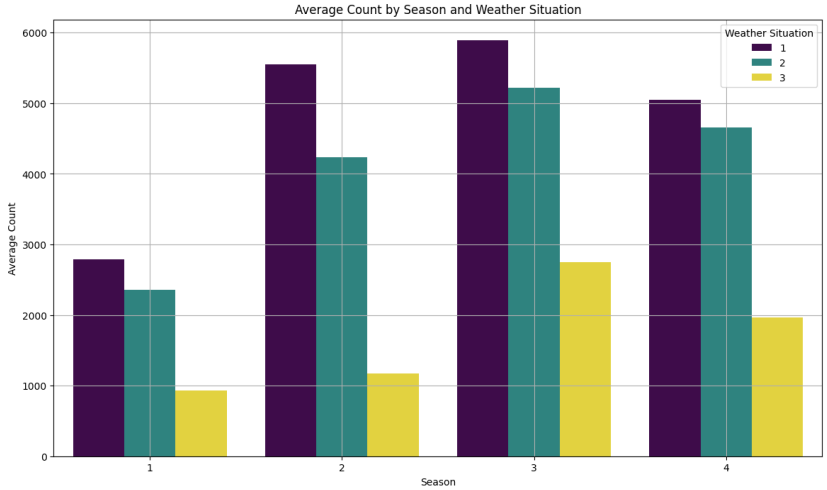


Figure 8 - Average Count by Season and Weather Situation via Multivariate Analysis

The choice of season and weathersit is rational because they are relevant factors influencing bike rental demand and are categorical variables suitable for multivariate analysis.

Findings:

- The average count varies across seasons, with Season 3 generally having the highest average count and Season 1 having the lowest.
- Within each season, the average count also varies depending on the weather situation. For example, in Season 1, the average count is highest for weathersit 1 and lowest for weathersit 3.

c) Aggregation analysis

	weathersit	mean	median
0	1	4874.131183	4844.0
1	2	4040.435484	4043.0
2	3	1803.285714	1817.0

Table 9 - Aggregation Analysis of mean and median of count by weather situation

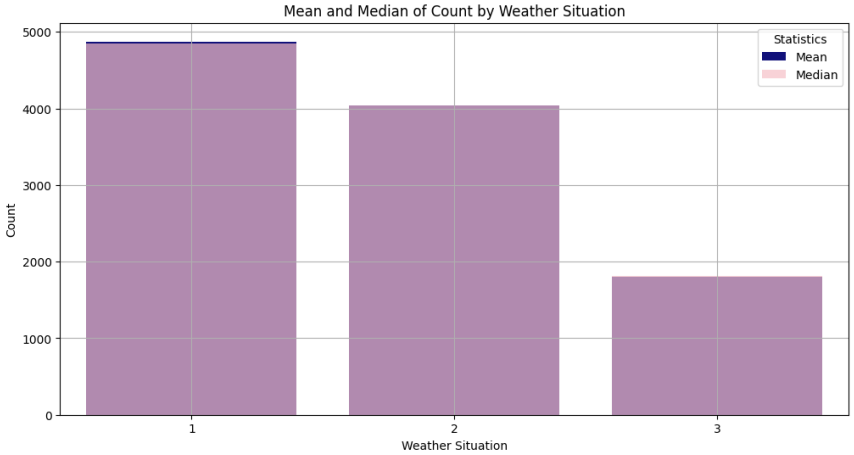


Figure 9- Bar graph of mean and median of count by weather situation

Findings:

- analysis confirms that weather situation has a significant impact on bike rental demand. Weather Situation 1 is associated with significantly higher average and median rental counts, suggesting that it is more favorable for bike usage.
- The relatively small differences between mean and median values for each weather situation indicate that the data distribution is not heavily skewed, suggesting a relatively balanced distribution of rental counts within each category.

  The findings suggest that Weather Situation 1 is more conducive to bike rentals, while Weather Situations 2 and 3 are associated with lower rental demand.

d) Average Analysis

	season	mean	std
0	1	2594.340659	1402.305471
1	2	4992.331522	1695.977235
2	3	5655.164021	1463.549093
3	4	4730.631285	1695.156053

Table 10 - Analysis of Count by season with Variation

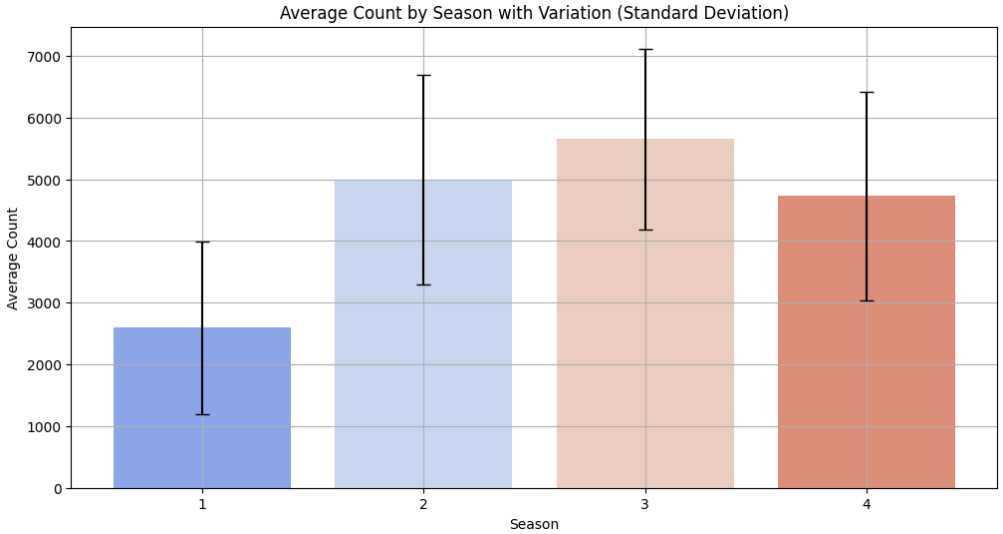


Figure 10 - Plot of average count by season with variation

The bar plot visually represents the average bike rental count (count) across different seasons, with error bars indicating the standard deviation.

Findings:

- There are clear differences in the average count across seasons. Season 3 has the highest average count, followed by Season 4, while Season 1 has the lowest.
- The error bars, representing the standard deviation, show that the variability in rental counts differs across seasons. Season 1 and 3 have higher variability compared to Season 2 and 4.

Discussion on the Significance of the Analysis:

1. This analysis helps identify seasonal trends in bike rental demand. Businesses can use this information to optimize their operations and resource allocation throughout the year. For example, they can increase bike availability during peak seasons and reduce it during slower periods.

2. Understanding seasonal variations can help businesses identify opportunities for growth and potential challenges. For instance, if bike rentals are significantly higher during certain seasons, businesses can focus on marketing and promotions to attract more customers during those times. On the other hand, they can prepare for potential challenges like increased maintenance or resource constraints during peak seasons.

Conclusion:

Key Findings and Insights from EDA

- **Distribution of Numerical Variables**: Histograms, box plots, and density plots revealed value ranges, outliers, and distribution patterns for count, temperature, and windspeed.
- **Relationships Between Variables**: Scatter plots and correlation analysis showed trends and correlations, such as how temperature influences bike-sharing counts.
- **Categorical Data Analysis**: Bar plots and pie charts highlighted frequencies and proportions of weather conditions and seasons, revealing their impact on bike-sharing activity.

Challenges Faced:

1. **Handling Missing Data**: Addressed missing entries with imputation and corrected data types.
2. **Visualization Accuracy**: Improved plots by adjusting parameters, data types, and formatting.
3. **Correlation Matrix Issues**: Pre-processed data to exclude non-numeric columns for accurate calculations.

Next Steps:

1. **Data Cleaning and Feature Engineering:** Improve dataset quality by handling missing values and creating new features for deeper insights.
2. **Advanced Analysis and Temporal Trends:** Conduct sophisticated statistical analyses and examine time-based trends to refine predictions and strategies.

APPENDIX-1

Summary statistics of the dataset:

	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	count
count	734.000000	734.000000	734.000000	734.000000	734.000000	734.000000	734.000000	731.000000	732.000000	734.000000	732.000000	734.000000	734.000000	734.000000
mean	2.497275	0.501362	6.521798	0.028610	2.990463	0.683924	1.395095	0.637989	0.474873	0.941225	0.190555	846.835150	3657.753406	4504.588556
std	1.111453	0.500339	3.457233	0.166822	2.005087	0.465260	0.544628	2.245183	0.162644	4.950552	0.077610	686.217748	1564.901961	1941.761607
min	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	1.000000	0.059130	0.098839	0.000000	0.022392	2.000000	20.000000	22.000000
25%	2.000000	0.000000	4.000000	0.000000	1.000000	0.000000	1.000000	0.338750	0.338256	0.520000	0.134950	315.250000	2495.000000	3146.500000
50%	3.000000	1.000000	7.000000	0.000000	3.000000	1.000000	1.000000	0.502500	0.487364	0.626250	0.180975	711.500000	3664.500000	4548.500000
75%	3.000000	1.000000	10.000000	0.000000	5.000000	1.000000	2.000000	0.656667	0.609544	0.731042	0.233211	1096.500000	4798.250000	5966.000000
max	4.000000	1.000000	12.000000	1.000000	6.000000	1.000000	3.000000	41.083300	0.840896	88.660000	0.507463	3410.000000	6946.000000	8714.000000

Appendix Table 1- Summary Statistics of the dataset

APPENDIX - 2

```
Missing values after imputation:
temp          0
atemp         0
windspeed     0
dtype: int64
```

Appendix Table 2 - Filled the values with Imputation

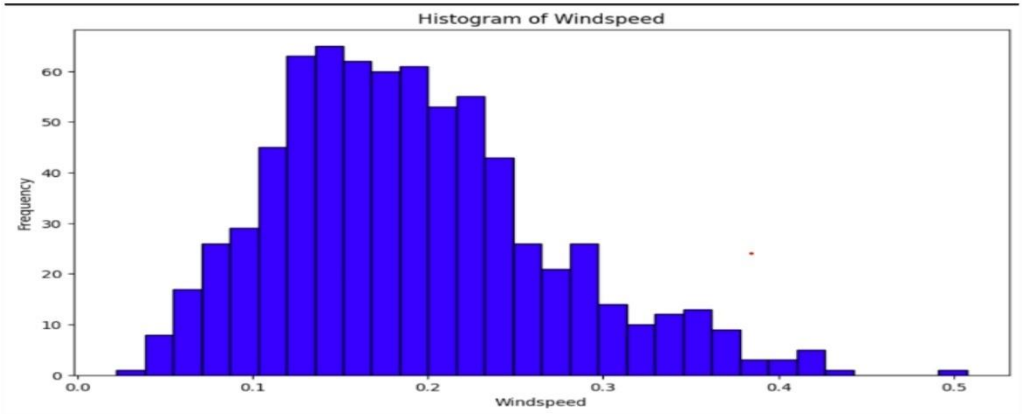
```
df_cleaned = df.drop_duplicates()

print(df_cleaned.duplicated().sum())

✓ 0.0s

0
```

Appendix Table 2- 1 - To check if the duplicates are removed it has to return 0

Student ID Number:_____

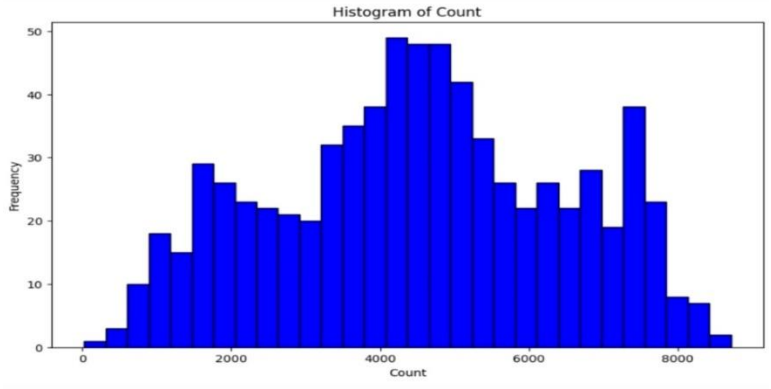
Lower Bound: -0.012441875000000047, Upper Bound: 0.3806031250000004

	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	count
44	14/02/2011	1	0	2	0	1	1	1	0.415000	0.398350	0.375833	0.417908	208	1705	1913
49	19/02/2011	1	0	2	0	6	0	1	0.399167	0.391404	0.187917	0.507463	532	1103	1635
93	4/04/2011	2	0	4	0	1	1	1	0.573333	0.542929	0.426250	0.385571	734	2381	3115
94	5/04/2011	2	0	4	0	2	1	2	0.414167	0.398350	0.642083	0.388067	167	1628	1795
292	20/10/2011	4	0	10	0	4	1	1	0.475833	0.466525	0.636250	0.422275	471	3724	4195
382	18/01/2012	1	1	1	0	3	1	1	0.303333	0.275254	0.443333	0.415429	109	3267	3376
407	12/02/2012	1	1	2	0	0	0	1	0.127500	0.101658	0.464583	0.409212	73	1456	1529
420	25/02/2012	1	1	2	0	6	0	1	0.290833	0.255675	0.395833	0.421642	317	2415	2732
432	8/03/2012	1	1	3	0	4	1	1	0.527500	0.524604	0.567500	0.441563	486	4896	5382
433	9/03/2012	1	1	3	0	5	1	2	0.410833	0.397083	0.407083	0.414800	447	4122	4569
450	26/03/2012	2	1	3	0	1	1	1	0.445833	0.438750	0.477917	0.386821	795	4763	5558
666	28/10/2012	4	1	10	0	0	0	2	0.477500	0.467771	0.694583	0.398008	998	3461	4459
721	22/12/2012	1	1	12	0	6	0	1	0.265833	0.236113	0.441250	0.407346	205	1544	1749



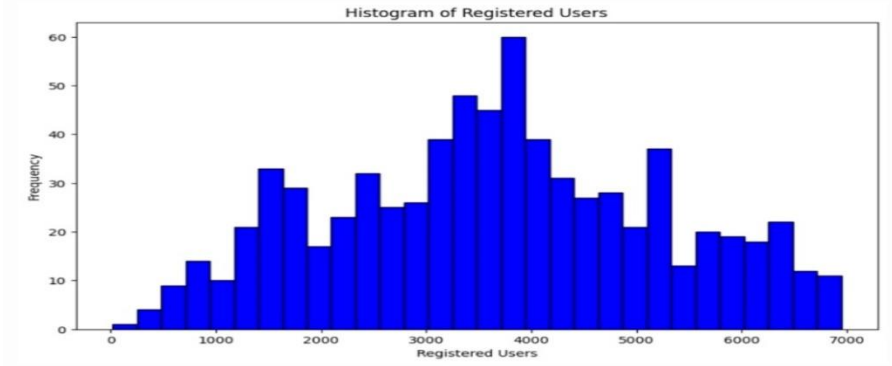
Appendix Table 2- 2 Outlier table and graph (histogram) of Windspeed

```
Lower Bound: -1082.75, Upper Bound: 10195.25
Outliers based on the IQR method:
 Empty DataFrame
```



Appendix Table 2- 3 Outlier table and graph (histogram) of Count

```
Lower Bound: -959.875, Upper Bound: 8253.125
Outliers based on the IQR method:
 Empty DataFrame
```

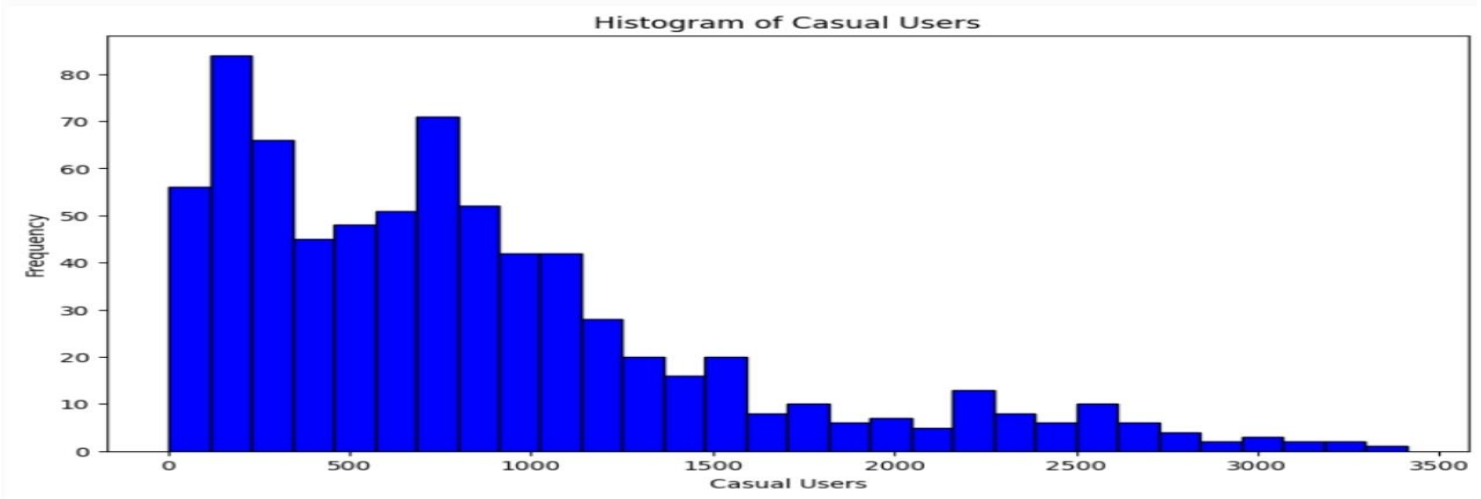


Appendix Table 2- 4 Outlier table and graph (histogram) of Registered Users

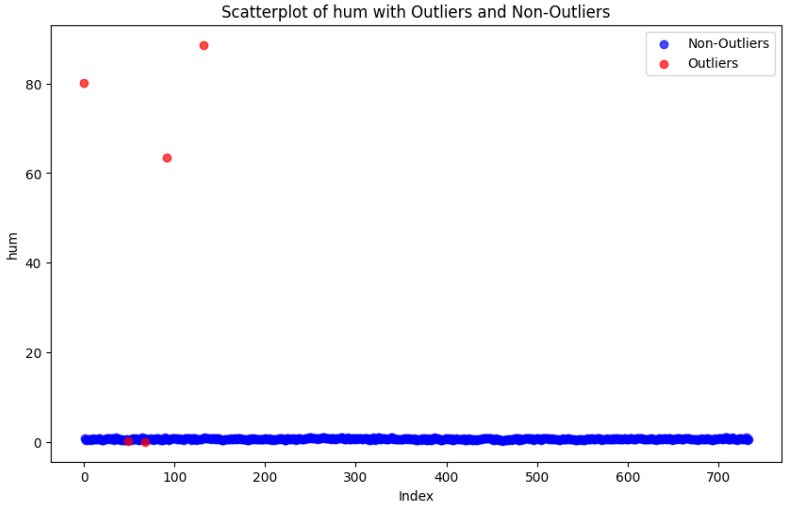
Student ID Number:_____

Lower Bound: -856.625, Upper Bound: 2268.375

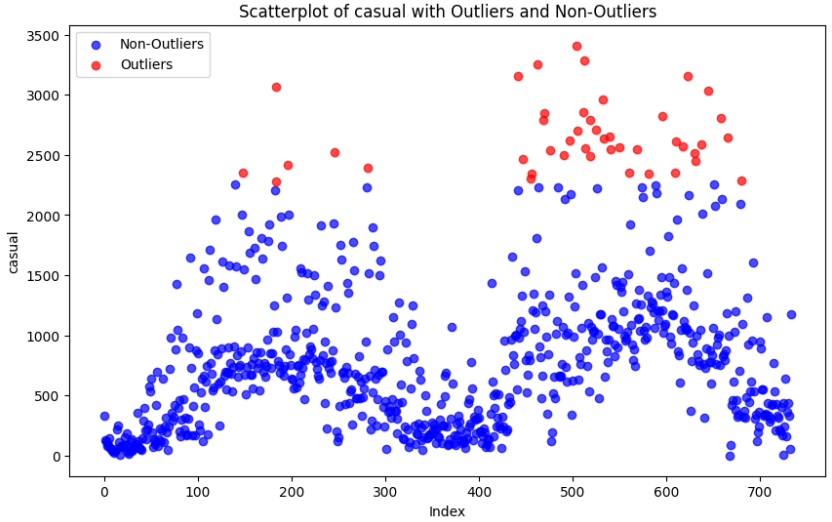
	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	count
148	29/05/2011	2	0	5	0	0	0	1	0.667500	0.615550	0.818750	0.213938	2355	2433	4788
183	3/07/2011	3	0	7	0	0	0	2	0.716667	0.668575	0.682500	0.228858	2282	2367	4649
184	4/07/2011	3	0	7	1	1	0	2	0.726667	0.665417	0.637917	0.081479	3065	2978	6043
196	16/07/2011	3	0	7	0	6	0	1	0.686667	0.638263	0.585000	0.208342	2418	3505	5923
246	4/09/2011	3	0	9	0	0	0	1	0.709167	0.665429	0.742083	0.206467	2521	2419	4940
281	9/10/2011	4	0	10	0	0	0	1	0.540833	0.523983	0.727500	0.063450	2397	3114	5511
441	17/03/2012	1	1	3	0	6	0	2	0.514167	0.505046	0.755833	0.110704	3155	4681	7836
447	23/03/2012	2	1	3	0	5	1	2	0.601667	0.570067	0.694167	0.116300	2469	5893	8362
455	31/03/2012	2	1	3	0	6	0	2	0.424167	0.421708	0.738333	0.250617	2301	3934	6235
456	1/04/2012	2	1	4	0	0	0	2	0.425833	0.417287	0.676250	0.172267	2347	3694	6041
462	7/04/2012	2	1	4	0	6	0	1	0.437500	0.426129	0.254167	0.274871	3252	3605	6857
469	14/04/2012	2	1	4	0	6	0	1	0.495000	0.487996	0.502917	0.190917	2795	4665	7460
470	15/04/2012	2	1	4	0	0	0	1	0.606667	0.573875	0.507917	0.225129	2846	4286	7132
476	21/04/2012	2	1	4	0	6	0	1	0.570000	0.542921	0.682917	0.283587	2541	4083	6624
490	5/05/2012	2	1	5	0	6	0	2	0.621667	0.584608	0.756667	0.152992	2496	4387	6883
497	12/05/2012	2	1	5	0	6	0	1	0.564167	0.544817	0.480417	0.123133	2622	4807	7429
504	19/05/2012	2	1	5	0	6	0	1	0.600000	0.566908	0.456250	0.083975	3410	4884	8294
505	20/05/2012	2	1	5	0	0	0	1	0.620833	0.583967	0.530417	0.254367	2704	4425	7129
511	26/05/2012	2	1	5	0	6	0	1	0.692500	0.642696	0.732500	0.198992	2855	3681	6536
512	27/05/2012	2	1	5	0	0	0	1	0.690000	0.641425	0.697083	0.215171	3283	3308	6591
513	28/05/2012	2	1	5	1	1	0	1	0.712500	0.679300	0.676250	0.196521	2557	3486	6043
518	2/06/2012	2	1	6	0	6	0	1	0.583333	0.566288	0.549167	0.186562	2795	5325	8120
519	3/06/2012	2	1	6	0	0	0	1	0.602500	0.575133	0.493333	0.184087	2494	5147	7641
...															
644	6/10/2012	4	1	10	0	6	0	1	0.554167	0.538521	0.664167	0.268025	3031	4934	7965
658	20/10/2012	4	1	10	0	6	0	1	0.484167	0.472842	0.572917	0.117537	2806	5284	8090
665	27/10/2012	4	1	10	0	6	0	2	0.530000	0.515133	0.720000	0.235692	2643	5209	7852
680	11/11/2012	4	1	11	0	0	0	1	0.420833	0.421713	0.659167	0.127500	2290	4562	6852



Appendix Table 2- 5 Outlier table and graph (histogram) of Casual Users



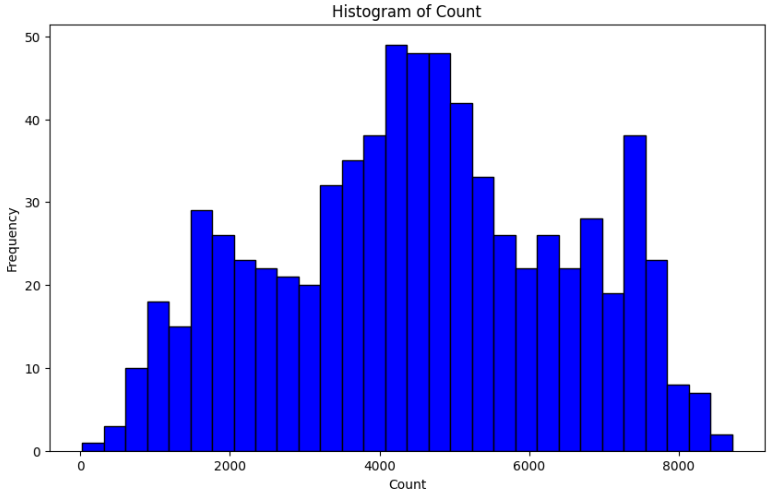
Appendix Figure 1 - Scatterplot of humidity with Outliers and Non-Outliers



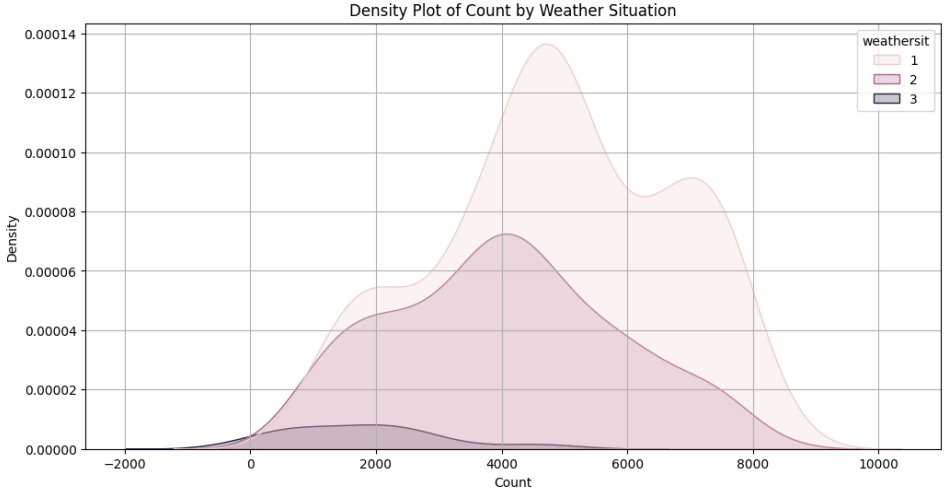
Appendix Figure 2 - Scatterplot of Casual Users with Outliers and Non-Outliers

Student ID Number:_____

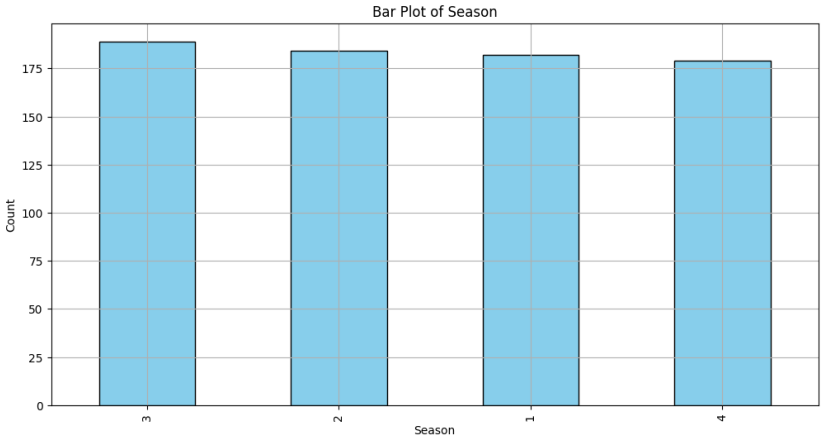
APPENDIX – 3



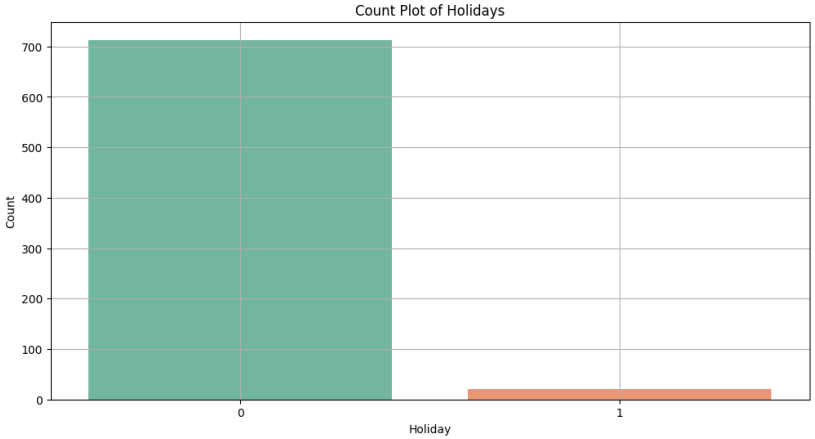
Appendix Figure 3- 1 Plot histogram for 'count' -comparison of numerical distributions



Appendix Figure 3- 2 Density plot of Count by Weather Situation - for numerical distribution



Appendix Figure 3- 3- Bar plot of Season - for categorical data



Appendix Figure 3- 4- Bar graph of Count plot of Holidays- for categorical data