CER 3

Descriptive statistics

Part 3

Date: 24/09/2024

Prosit 2

Prosit Heading: Descriptive statistics

Roles:
Facilitator: Junior
Manager: Clement
Secretary: Jeremy
Scribe: Peïo

1) Understanding the Topic and Clarifying

1.1) Keywords:

- **Feasibility :**
  Feasibility refers to the practicality or possibility of implementing a plan, process, or solution, given constraints such as resources, time, and technology.

- **Competitors:**
  Competitors are individuals or organizations that vie for the same customers, market share, or resources in a business or competitive environment.

- **Exhaustive :**
  Exhaustive means including all possible elements, options, or outcomes in a given context.

- **Dataset :**

  A dataset is a collection of data points or observations, typically organized in rows and columns, which are used for analysis.

Correlation with Descriptive Statistics: Descriptive statistics are used to summarize and analyse datasets. Measures such as mean, median, variance, and frequency distributions are calculated from datasets to describe and understand the data.

1.2) Context:

Students offer bike service convince city representatives of feasibility.

2) Needs Analysis:

2.1) Issues:

- How to determine a feasibility of project?
- Convince?
- Finding right dates of opening?

2.2) Constraints:

- Limited dataset and context

- Competitors

2.3) Deliverable:

- Report/CER
- Excel? or Notebook (.ipynb, jupyter)

3) Generalization:

- Descriptive statistics

4) Solution Tracks:

- Should find the best opening day.
- Business should be opened on weekend.
- Business should be opened on weekday.
- Reorganizing data sets.
- Find mode and medians

5) Developing the Action Plan:

1. **Define keywords.**
2. **Exercise Basket / WS.**
3. **Resource Analysis.**
4. **Analyse datasets**
5. **Highlight important data + important cold**
6. **Graphs -> EDA**
7. **Feasibility analysis**
8. **Validation of assumptions.**
9. **Conclusion.**

6) Resource Analysis:

Descriptive statistics involves summarizing and organizing data to make it understandable and interpretable. ("Vocab, Definition, and Must Know Facts - Fiveable") It encompasses various measures and techniques that provide insights into the central tendency, dispersion, and overall distribution of data. Below are key keywords in descriptive statistics, along with their definitions, examples, and applications:

1. Mean (Average)

- **Definition**: The mean is the sum of all data points divided by the number of data points. ("Mean, Median, Mode, Range and Quartiles calculator") ("Mean, Median, Mode, Range and Quartiles calculator") It represents the central value of a dataset.
- **Example**: Consider the dataset [2, 4, 6, 8, 10]. The mean is (2 + 4 + 6 + 8 + 10) / 5 = 6.
- **Application**:
    - **Economics**: Calculating average income to assess economic well-being.
    - **Education**: Determining the average test score to evaluate student performance.

2. Median

- **Definition**: The median is the middle value in an ordered dataset. ("Median Definition and Uses - Statistics By Jim") ("Median Definition and Uses - Statistics By Jim") If the number of observations is even, it is the average of the two middle numbers.
- **Example**:
    - Odd dataset: [3, 5, 7] → Median = 5.
    - Even dataset: [3, 5, 7, 9] → Median = (5 + 7) / 2 = 6.
- **Application**:
    - **Real Estate**: Using median home prices to understand market trends, as it is less affected by extreme values.

- **Income Analysis**: Assessing median income to represent typical earnings without distortion from very high or low incomes.

3. Mode

- **Definition**: The mode is the value that appears most frequently in a dataset. (“Mode - Vocab, Definition, and Must Know Facts | Fiveable”) "A dataset can have one mode, multiple modes, or no mode." (“Univariate Analysis: Understanding Measures of Central Tendency and ...”)
- **Example**:
  - Single mode: [1, 2, 2, 3] → Mode = 2.
  - Multiple modes: [1, 1, 2, 2, 3] → Modes = 1 and 2.
- **Application**:
  - **Marketing**: Identifying the most popular product size or colour.
  - **Healthcare**: Determining the most common patient diagnosis.

4. Range

- **Definition**: The range is the difference between the maximum and minimum values in a dataset. (“Statistical Description Of Data - Medium”)
- **Example**: For [4, 8, 15, 16, 23, 42], Range = 42 - 4 = 38.
- **Application**:
  - **Quality Control**: Assessing the range of product dimensions to ensure consistency.
  - **Finance**: Evaluating the range of stock prices to understand volatility.

5. Variance

- **Definition**: Variance measures the average squared deviation of each data point from the mean, indicating data dispersion.
- **Example**: For the dataset [2, 4, 4, 4, 5, 5, 7, 9], the variance is calculated based on the deviations from the mean (5).
- **Application**:

- **Risk Assessment**: Measuring the variance of investment returns to assess financial risk.
- **Research**: Evaluating variability in experimental data to understand consistency.

6. Standard Deviation

- **Definition**: The standard deviation is the square root of the variance, providing a measure of dispersion in the same units as the data. ("Mean, Variance and Standard Deviation - GeeksforGeeks")
- **Example**: If variance is 16, the standard deviation is 4.
- **Application**:
  - **Psychology**: Measuring the variability in test scores to assess consistency.
  - **Engineering**: Determining the standard deviation of manufacturing measurements to maintain quality.

7. Quartiles

- **Definition**: Quartiles divide a dataset into four equal parts. The first quartile (Q1) is the 25th percentile, the second quartile (Q2) is the median (50th percentile), and the third quartile (Q3) is the 75th percentile. ("Quartiles and Their Importance in Statistical Analysis") ("Quartiles and Their Importance in Statistical Analysis")
- **Example**: For the dataset [1, 3, 5, 7, 9], Q1 = 3, Q2 = 5, Q3 = 7.
- **Application**:
  - **Education**: Analysing student performance by quartiles to identify top and bottom performers.
  - **Finance**: Evaluating income distribution by quartiles to understand economic inequality.

8. Percentiles

- **Definition**: Percentiles indicate the relative standing of a value within a dataset, showing the percentage of data points below that value.

- **Example**: The 90th percentile in a test score distribution means 90% of students scored below that value.
- **Application**:
    - **Standardized Testing**: Reporting student performance relative to peers.
    - **Health**: Assessing growth charts in pediatrics by percentiles.

9. Interquartile Range (IQR)

- **Definition**: IQR measures the middle 50% of a dataset by calculating the difference between the third quartile (Q3) and the first quartile (Q1). ("Quartiles - Vocab, Definition, and Must Know Facts | Fiveable") ("Quartiles - Vocab, Definition, and Must Know Facts | Fiveable")
- **Formula**: IQR=Q3−Q1\text{IQR} = Q3 - Q1IQR=Q3−Q1
- **Example**: If Q3 = 75 and Q1 = 25, then IQR = 50.
- **Application**:
    - **Data Analysis**: Identifying the spread of the central half of the data.
    - **Outlier Detection**: Using IQR to determine potential outliers (e.g., values beyond 1.5 * IQR from Q1 or Q3).

10. Skewness

- **Definition**: Skewness measures the asymmetry of a data distribution. Positive skewness indicates a longer right tail, while negative skewness indicates a longer left tail. ("Skewness - Vocab, Definition, and Must Know Facts | Fiveable")
- **Example**:
    - **Positive Skew**: Income distribution often has positive skewness, with most people earning below a high-income outlier.
    - **Negative Skew**: Exam scores where most students score high with a few low scores.
- **Application**:
    - **Finance**: Assessing the skewness of asset returns to understand the risk of extreme outcomes.

- **Quality Control**: Evaluating the distribution of product measurements to identify asymmetry.

11. Kurtosis

- **Definition**: Kurtosis measures the "tailedness" of a data distribution. High kurtosis indicates heavy tails and more outliers, while low kurtosis indicates light tails. (“Statistics For Data Science - GeeksforGeeks”)
- **Example**:
  - **High Kurtosis**: Financial returns often exhibit high kurtosis due to the presence of rare, extreme events.
  - **Low Kurtosis**: Uniform distributions have low kurtosis.
- **Application**:
  - **Risk Management**: Understanding the likelihood of extreme financial losses.
  - **Statistics**: Choosing appropriate models based on the distribution's kurtosis.

12. Frequency Distribution

- **Definition**: A frequency distribution organizes data into classes or intervals and displays the number of observations in each class.
- **Example**: A survey of ages with intervals 0-10, 11-20, etc., showing how many respondents fall into each age group.
- **Application**:
  - **Demographics**: Analysing population age groups.
  - **Market Research**: Categorizing customer ages to target marketing strategies.

13. Histogram

- **Definition**: A histogram is a graphical representation of a frequency distribution, using bars to show the number of data points within each interval. (“Histogram - Vocab, Definition, and Must Know Facts | Fiveable”)

- **Example**: A histogram displaying the distribution of exam scores with intervals 0-10, 11-20, etc.
- **Application**:
    - **Data Analysis**: Visualizing the distribution of data to identify patterns, skewness, and outliers.
    - **Quality Control**: Monitoring the distribution of product measurements over time.

14. Bar Chart

- **Definition**: A bar chart displays categorical data with rectangular bars representing the frequency or proportion of each category.
- **Example**: A bar chart showing the number of students enrolled in different majors.
- **Application**:
    - **Business**: Comparing sales across different product categories.
    - **Education**: Visualizing the number of students in various academic programs.

15. Box Plot (Box-and-Whisker Plot)

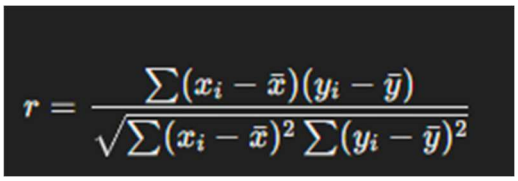
- **Definition**: A box plot graphically depicts the distribution of a dataset based on its quartiles, highlighting the median, IQR, and potential outliers.
- **Example**: A box plot showing the distribution of salaries within a company, indicating the median, quartiles, and any extreme salaries.
- **Application**:
    - **Comparative Analysis**: Comparing distributions across distinct groups or categories.
    - **Outlier Detection**: Identifying unusual data points that may require further investigation.

16. Scatter Plot

- **Definition**: A scatter plot displays pairs of numerical data points on a Cartesian plane, illustrating the relationship between two variables.
- **Example**: Plotting height versus weight for a group of individuals to examine correlation.
- **Application**:
    - **Economics**: Analysing the relationship between advertising spend and sales revenue.
    - **Science**: Investigating the correlation between temperature and reaction rates in experiments.

17. Correlation

- **"Definition**: Correlation measures the strength and direction of the linear relationship between two variables." ("Correlation and Covariance for GATE Exam - AlmaBetter")



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

- **Example**: A correlation coefficient of 0.8 between study hours and exam scores indicates a strong positive relationship.
- **Application**:
    - **Finance**: Assessing the correlation between different asset returns to build diversified portfolios.
    - **Healthcare**: Exploring the relationship between lifestyle factors and health outcomes.

18. Covariance

- **Definition**: Covariance indicates the direction of the linear relationship between two variables. Positive covariance means both variables increase together, while negative covariance means one increases as the other decreases. ("Scatter plots | PPT - SlideShare")

$$\text{Cov}(X,Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- **Example**: If higher temperatures are associated with increased ice cream sales, the covariance between temperature and ice cream sales is positive.
- **Application**:
    - **Portfolio Management**: Understanding how different investments move together to manage risk.
    - **Economics**: Analysing the relationship between GDP and unemployment rates.

19. Central Tendency

- **Definition**: Central tendency refers to the pivotal point or typical value of a dataset. Measures of central tendency include mean, median, and mode.
- **Example**: In [1, 2, 3, 4, 5], the mean is 3, the median is 3, and the mode does not exist.
- **Application**:
    - **Sociology**: Describing the central tendency of social indicators like income or education levels.
    - **Business**: Setting target performance levels based on average sales figures.

20. Dispersion (Variability)

- **Definition**: Dispersion measures the spread or variability of a dataset. Key measures include range, variance, standard deviation, and IQR.
- **Example**: Two datasets with the same mean but different variances indicate various levels of spread.
- **Application**:
    - **Engineering**: Controlling the variability of manufacturing processes to ensure quality.

- **Education**: Assessing the variability in student performance to identify areas needing support.

21. Outliers

- **Definition**: Outliers are data points that differ significantly from other observations in a dataset. ("Outliers - Vocab, Definition, and Must Know Facts | Fiveable") They can result from variability in the data or measurement errors.
- **Example**: In [10, 12, 12, 13, 12, 14, 100], the value 100 is an outlier.
- **Application**:
  - **Data Cleaning**: Identifying and addressing outliers to improve data quality.
  - **Finance**: Detecting fraudulent transactions that appear as outliers in transaction data.

22. Summary Statistics

- **Definition**: Summary statistics provide a concise overview of the major features of a dataset, including measures of central tendency, dispersion, and shape.
- **Example**: A summary of exam scores might include the mean, median, standard deviation, minimum, and maximum values.
- **Application**:
  - **Reporting**: Creating executive summaries in business reports.
  - **Research**: Summarizing experimental data to highlight key findings.

7) Data set analysis:

Data types:

```
RangeIndex: 731 entries, 0 to 730
Data columns (total 16 columns):
 #    Column       Non-Null Count  Dtype
---   ------       --------------  -----
 0    instant      731 non-null    int64
 1    dteday       731 non-null    object
 2    season       731 non-null    int64
 3    yr           731 non-null    int64
 4    mnth         731 non-null    int64
 5    holiday      731 non-null    int64
 6    weekday      731 non-null    int64
 7    workingday   731 non-null    int64
 8    weathersit   731 non-null    int64
 9    temp         731 non-null    float64
 10   atemp        731 non-null    float64
 11   hum          731 non-null    float64
 12   windspeed    731 non-null    float64
 13   casual       731 non-null    int64
 14   registered   731 non-null    int64
 15   cnt          731 non-null    int64
dtypes: float64(4), int64(11), object(1)
memory usage: 91.5+ KB
```

Sample of data:

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
4	2011-01-04	1	0	1	0	2	1	1	0.200000	0.212122	0.590435	0.160296	108	1454	1562
5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.186900	82	1518	1600
6	2011-01-06	1	0	1	0	4	1	1	0.204348	0.233209	0.518261	0.089565	88	1518	1606
7	2011-01-07	1	0	1	0	5	1	2	0.196522	0.208839	0.498696	0.168726	148	1362	1510
8	2011-01-08	1	0	1	0	6	0	2	0.165000	0.162254	0.535833	0.266804	68	891	959
9	2011-01-09	1	0	1	0	0	0	1	0.138333	0.116175	0.434167	0.361950	54	768	822
10	2011-01-10	1	0	1	0	1	1	1	0.150833	0.150888	0.482917	0.223267	41	1280	1321
11	2011-01-11	1	0	1	0	2	1	2	0.169091	0.191464	0.686364	0.122132	43	1220	1263

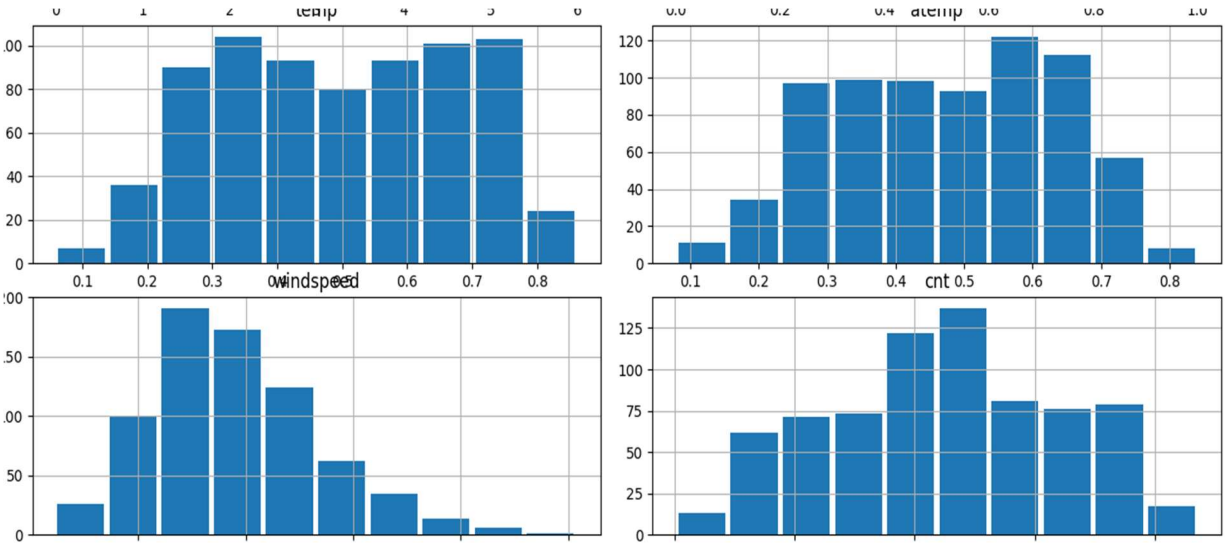
Key points about each data:

	count	mean	std	min	25%	\
instant	731.0	366.000000	211.165812	1.000000	183.500000	
season	731.0	2.496580	1.110807	1.000000	2.000000	
yr	731.0	0.500684	0.500342	0.000000	0.000000	
mnth	731.0	6.519836	3.451913	1.000000	4.000000	
holiday	731.0	0.028728	0.167155	0.000000	0.000000	
weekday	731.0	2.997264	2.004787	0.000000	1.000000	
workingday	731.0	0.683995	0.465233	0.000000	0.000000	
weathersit	731.0	1.395349	0.544894	1.000000	1.000000	
temp	731.0	0.495385	0.183051	0.059130	0.337083	
atemp	731.0	0.474354	0.162961	0.079070	0.337842	
hum	731.0	0.627894	0.142429	0.000000	0.520000	
windspeed	731.0	0.190486	0.077498	0.022392	0.134950	
casual	731.0	848.176471	686.622488	2.000000	315.500000	
registered	731.0	3656.172367	1560.256377	20.000000	2497.000000	
cnt	731.0	4504.348837	1937.211452	22.000000	3152.000000	

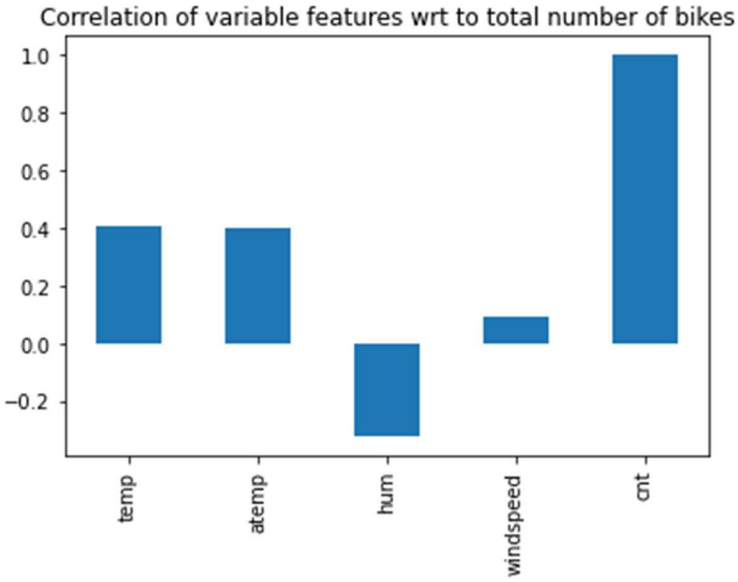
	50%	75%	max
instant	366.000000	548.500000	731.000000
season	3.000000	3.000000	4.000000
yr	1.000000	1.000000	1.000000
mnth	7.000000	10.000000	12.000000
holiday	0.000000	0.000000	1.000000
weekday	3.000000	5.000000	6.000000
workingday	1.000000	1.000000	1.000000
weathersit	1.000000	2.000000	3.000000
temp	0.498333	0.655417	0.861667
atemp	0.486733	0.608602	0.840896
hum	0.626667	0.730209	0.972500
windspeed	0.180975	0.233214	0.507463
casual	713.000000	1096.000000	3410.000000
registered	3662.000000	4776.500000	6946.000000
cnt	4548.000000	5956.000000	8714.000000

8) Graphs:

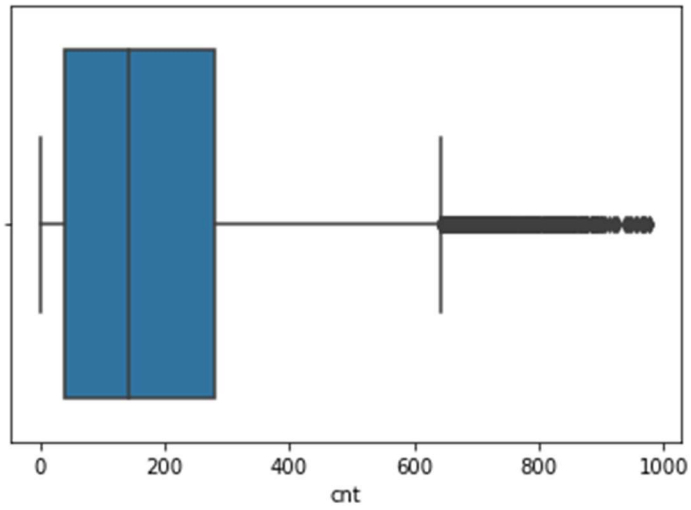
Bar chart for each data set i.e. column.



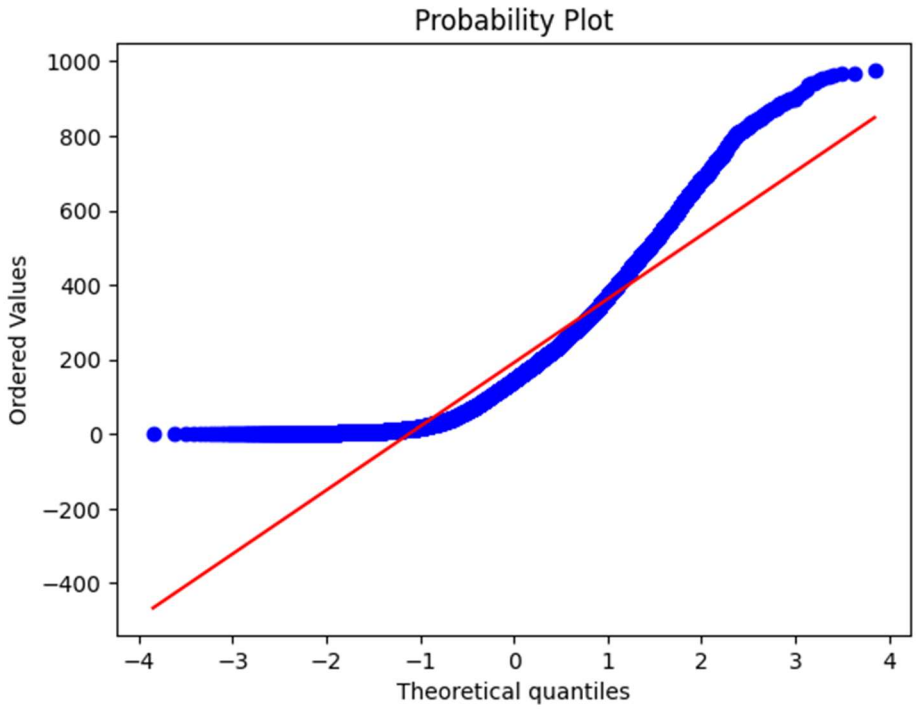
Correlation between different various variables against bike count



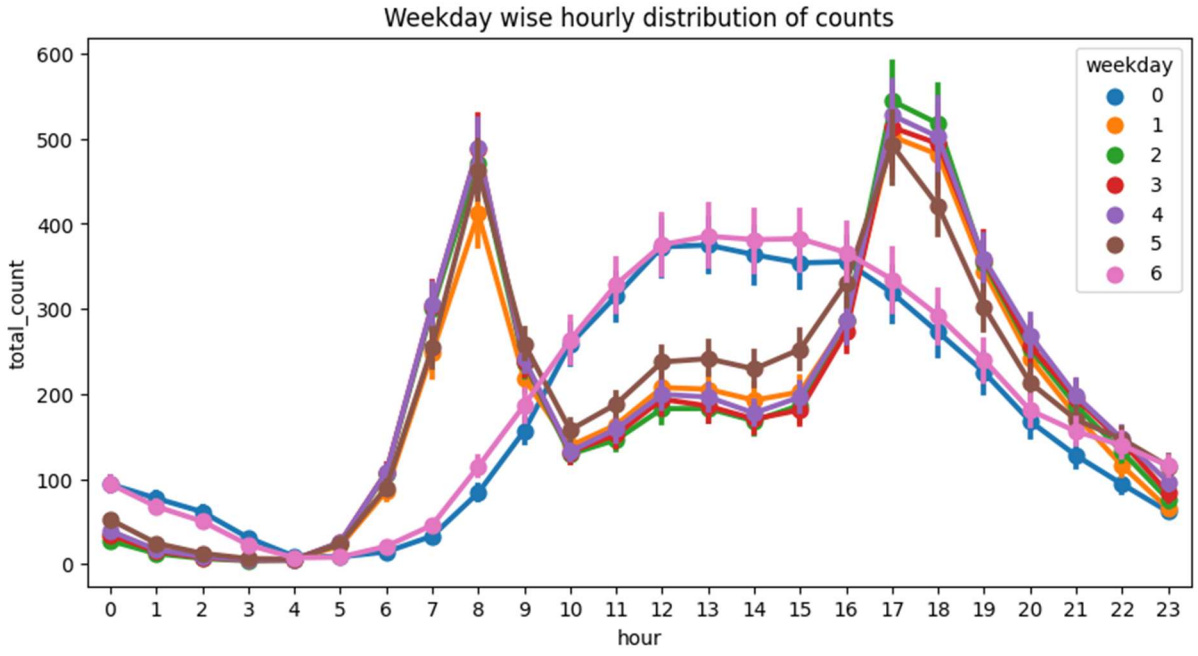
Box plot graph



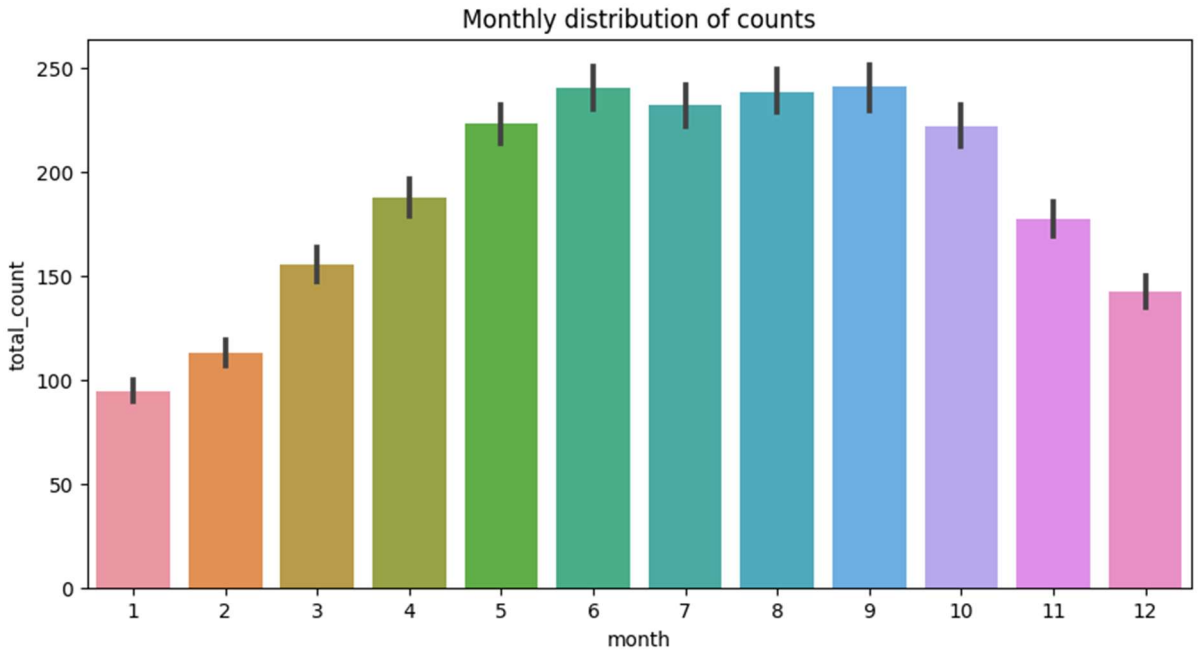
Linear regression



Any deviation from a straight line in a normal probability plot or signs of skewness or multi-modality in a histogram indicate that the data does not meet the assumptions of normality.



Ridership patterns show interesting trends, with higher usage during afternoon hours on weekends, while weekdays experience increased ridership in the mornings and evenings.



As shown in the graph above, the best opening days would be in the months of June through September as we observe the highest ridership, suggesting that fall is a favourable season for bike-sharing programs in Washington, D.C.

Based on available data, there are two groups of variables that affect ridership in a bike sharing system:

1. **Weather Condition Variables**: These include factors such as temperature, humidity, and other weather-related metrics.
2. **Holiday and Weekend Effects**: These variables have implications for ridership patterns on working and non-working days.

9) Conclusion:

In conclusion, various facts i.e. graphs shown above, provide solutions for the best opening days, the correlation between the datasets, etc. which should be enough to convince the newly elected representative.