

DATA ACQUISITION AND DATA WRANGLING

PREPARED BY: SILIVIYA



WHAT IS DATA ANALYSIS?

Data analysis is the process of inspecting, cleaning, transforming, and interpreting data to extract meaningful insights. It helps in identifying patterns, trends, and relationships that support decision-making. By using statistical methods and visualization techniques, data analysis turns raw information into actionable knowledge.





THE DATA ANALYSIS PROCESS

Data Collection

Process of gathering information from various sources to analyze and make informed decisions. It can be done through surveys, observations, experiments, or digital tracking methods.

Cleaning


Process of identifying and correcting errors, inconsistencies, and missing values in a dataset. It ensures data accuracy, reliability, and readiness for analysis.

Data Analysis

Process of examining, organizing, and interpreting data to uncover patterns, trends, and insights. It helps in making informed decisions and solving problems effectively.

Interpretation, and Reporting

Involve explaining the meaning of analyzed data and presenting findings in a clear, concise manner. This helps stakeholders understand insights, make decisions, and take action based on the results.



TOOLS USED IN DATA ANALYSIS

Used Python for data processing, data manipulation, and numerical computations.

For visualization, utilized Matplotlib and Seaborn to create insightful charts and graphs.

These tools helped in understanding patterns, trends, and relationships in the data.

Python's versatility made the analysis efficient and effective.



DATA ACQUISITION AND DATA WRANGLING ON DATASET 1,2 AND DATASET 3

1. Data Acquisition

- Imported datasets into Python using Pandas.
- Checked the structure, dimensions, and data types of each dataset.

2. Data Cleaning

- Identified and handled missing values.
- Removed unnecessary or duplicate columns.
- Standardized column names for consistency.
- Replace zero windspeed values with the median (assuming 0 might indicate missing values)

3. Data Merging

- Merged the datasets 1 & 2 based on common attributes to ensure data completeness.(Merged dataset 1&2)
- Verified the merge by checking for duplicates or inconsistencies.
- Merged the clean dataset 3 with new merged dataset 1&2.

4. Data Transformation

- Converted data types where necessary (e.g., dates, categorical values).
- Created new meaningful variables for better analysis.

5. Data Exploration

- Performed statistical analysis to check central tendencies, outliers, and correlations.
- Used Matplotlib and Seaborn for visual insights into trends and patterns.

TOOLS USED IN DATA VISUALIZATION

1.Simple Plot (Line Chart)

- ❑ Used to show trends over time or comparisons between categories.
- ❑ Example: A line chart can display variations in total users per month, while a bar chart can compare seasonal trends.

2.Box Plot

- ❑ Helps in understanding data distribution, variability, and detecting outliers.
- ❑ Example: A box plot for temperature values can reveal extreme weather conditions affecting user activity.

3.Scatterplot

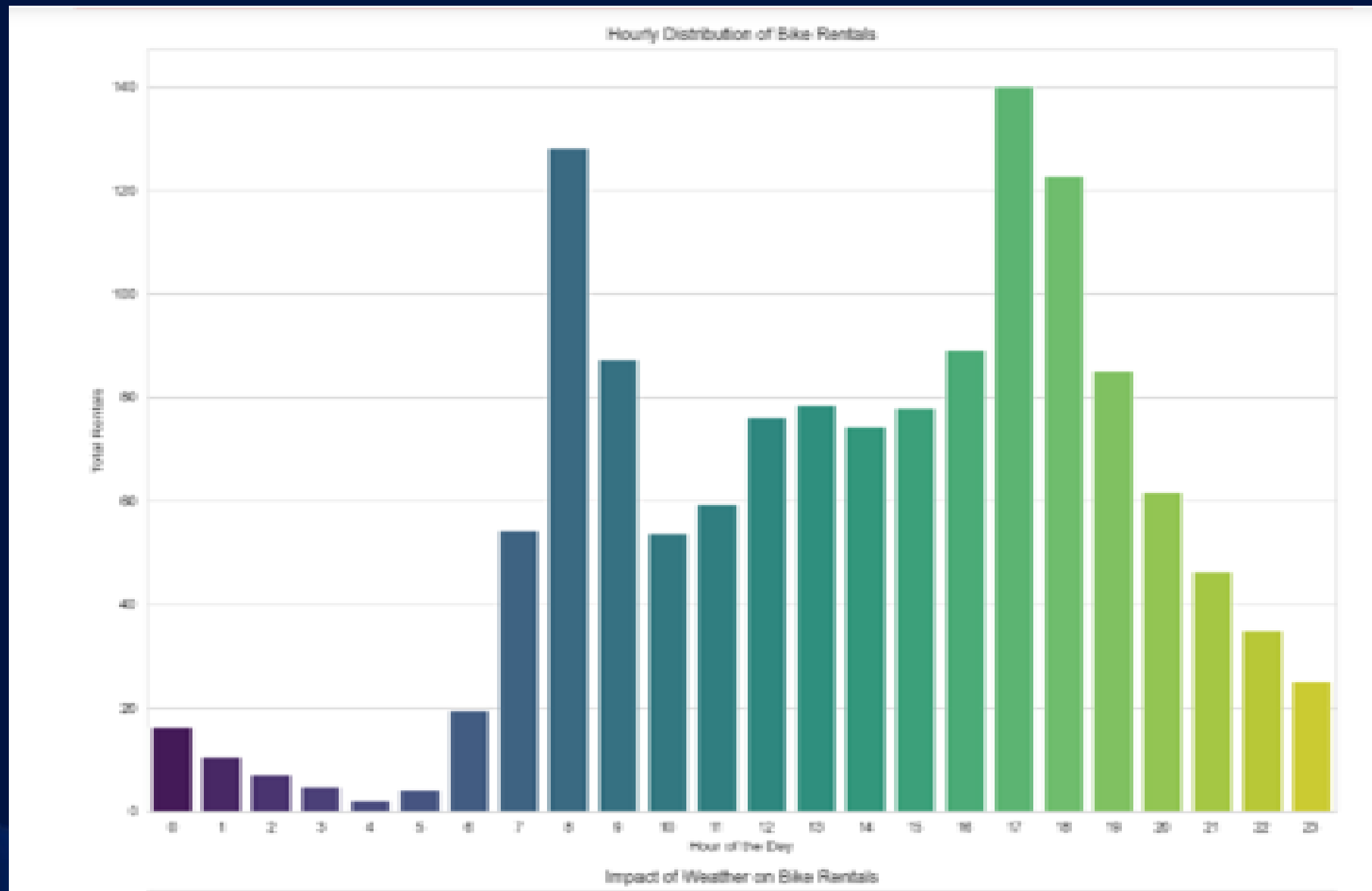
- ❑ Shows relationships between two numerical variables.
- ❑ Example: A scatterplot can help visualize how temperature impacts total user counts.

4.Correlation Heatmap

- ❑ Uses color intensity to show relationships between multiple variables.
- ❑ Example: A heatmap can highlight strong correlations, like how humidity and wind speed impact user behavior.

Continue...

DATA VISUALIZATION



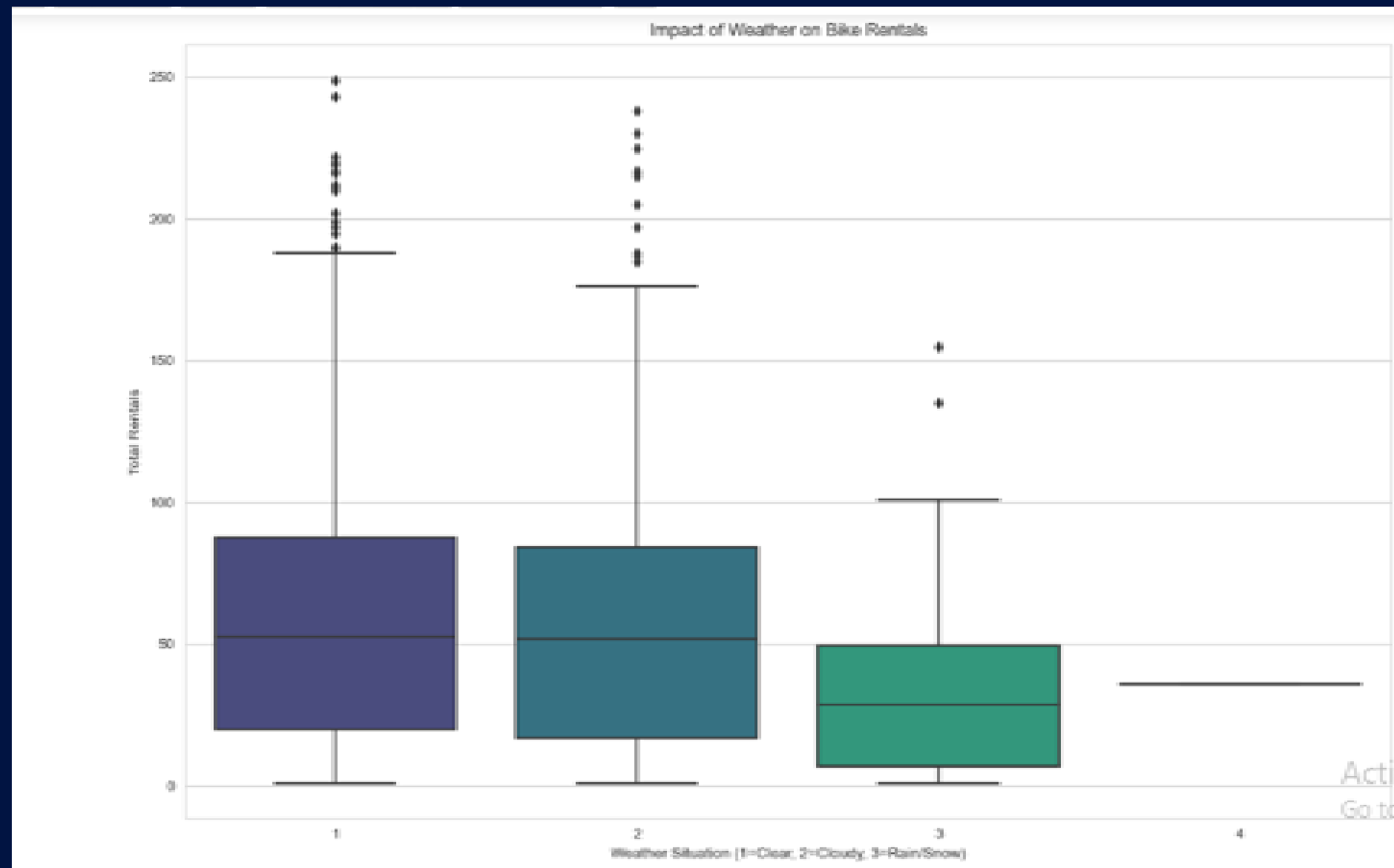
Insights:

- Bike rentals spike during morning (8–9 AM) and evening rush hours (5–7 PM). People use bikes for commuting.
- Day rentals are lower, but there's a slight increase in the evening (7–9 PM). This could be recreational users.
- Late-night rentals are minimal.

Actionable Strategy:

- Ensure enough bikes are available during peak commuting hours.
- Provide night-time biking incentives, such as safer routes or discounts.

DATA VISUALIZATION



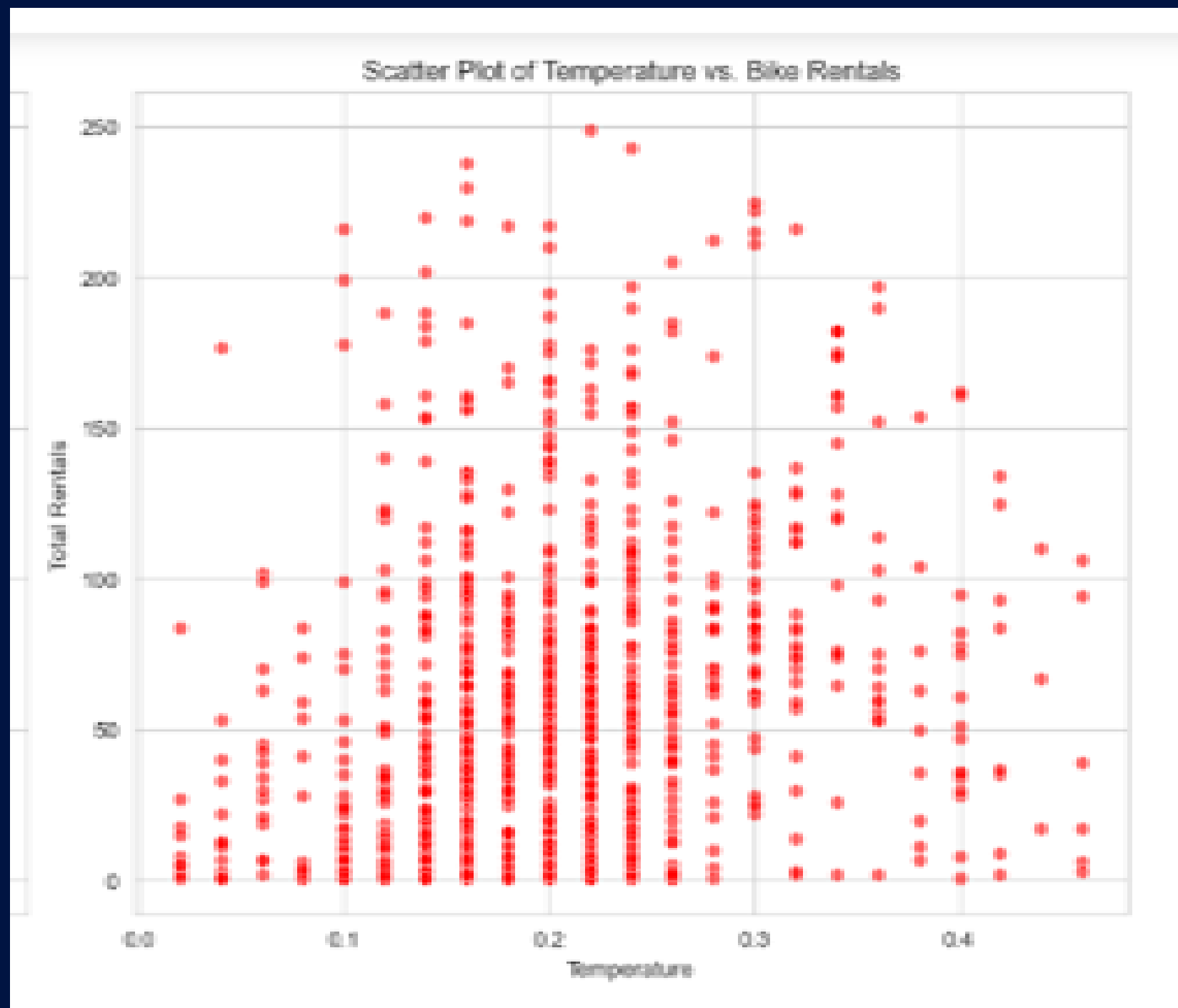
Insights:

- Some days have extremely high rentals. These could be due to promotions, public events, or good weather.
- There are also days with unusually low rentals. Possibly due to bad weather, major holidays, or transportation strikes.

Actionable Strategy:

- Investigate reasons for high-rental days and replicate those conditions (e.g., events, discounts).
- Offer alternative services during low-rental days (e.g., delivery services).

DATA VISUALIZATION



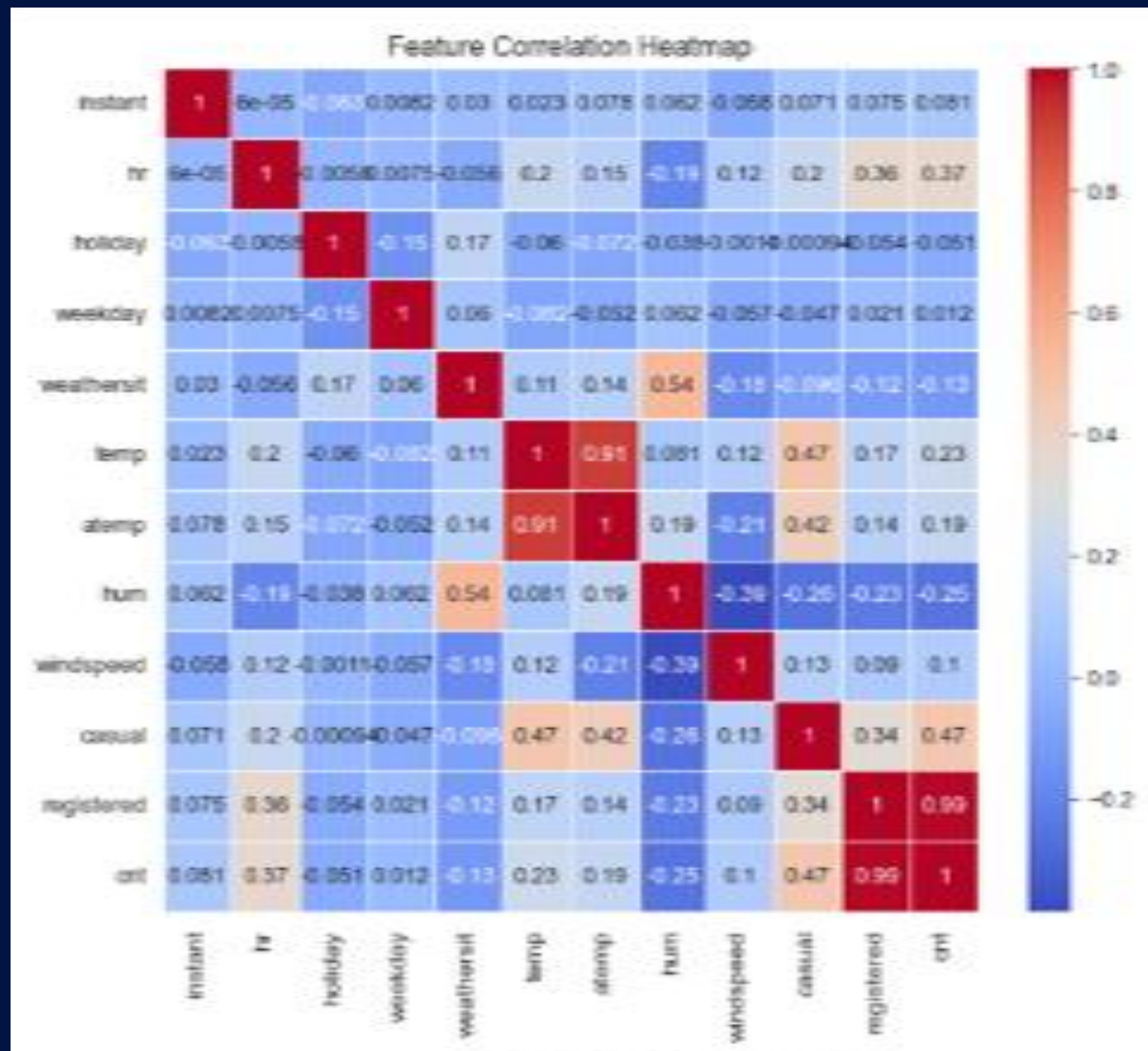
Insights:

- As temperature rises, rentals increase—up to a certain point.
- However, on extremely hot days, rentals decline. People avoid biking in extreme heat.
- Mild weather (15°C–25°C) is the sweet spot for biking.
- Actionable Strategy:

Actionable Strategy:

- Promote biking during pleasant weather periods.
- Offer shaded parking areas and water stations during hot weather.

DATA VISUALIZATION



Insights:

- Higher temperatures lead to more bike rentals. People prefer biking in pleasant weather.
- Humidity and wind speed negatively impact rentals. Bad weather discourages cycling.
- Time-related factors (season, month, and year) are strongly correlated with rentals. Rentals follow a clear seasonal trend.
- Casual users and registered users behave differently. Casual users are more affected by weather, while registered users rent bikes more consistently.

Actionable Strategy:

- Offer weather-based discounts (e.g., lower prices on hot days or windy days).
- Ensure more bikes are available in peak seasons.



CHALLENGES IN DATA ANALYSIS

Data Merging Issues:

No common values in the instant column, making merging difficult.
Used an inner join on relevant columns after verification.

Missing & Inconsistent Data:

Season, yr, mnth had missing values affecting analysis.
Dropped highly incomplete columns and filled numeric gaps with mean/median.

Outliers & Data Distribution Issues:

Extreme rental values detected in box plots.
Verified distribution before deciding to keep or remove outliers.

Visualization Challenges:

Blank charts due to missing values and incorrect data types.
Fixed by converting dteday to datetime and adjusting figure sizes.

Identifying Key Insights:

Rentals peak at specific hours & months, drop at night.
Weather significantly impacts demand, with temperature showing strong correlation.



ACTIONABLE RECOMMENDATIONS

Optimize Bike Availability-

- Increase bike supply during peak commuting hours.
- Reduce idle bikes at night to minimize operational costs.

Weather-Based Demand Forecasting-

- Provide more bikes on warm, clear days when rentals are high.
- Offer discounts or incentives on rainy/cold days to boost usage.

Seasonal Strategy

- Increase fleet size in summer and spring to meet demand.
- Launch winter promotions to encourage rentals during the off-season.

Improve Infrastructure

- Add docking stations in high-demand areas, especially near offices and transit hubs.
- Improve bike lanes and safety measures to encourage ridership.

Targeted Marketing & Pricing

- Implement dynamic pricing (higher during peak, discounts in off-peak hours).
- Special promotions for commuters during weekday rush hours





THANK YOU