BIKE SHARING PREDICTION PROJECT WITH MACHINE LEARNING

BY: CLARINDA PUSPITAJATI

INTRODUCTION: OBJECTIVES

1. Problem Statement:
    a. Bicycle utilization forecasting is essential for station management optimization, supporting a balanced supply, equitable distribution throughout the day, and sustaining a steady supply integration.
2. Stakeholders:
    a. Bike Sharing Operators
    b. CityPlanners
    c. Riders or Users
3. Purpose:
    a. To predict hourly bike rental demand to optimize fleet management, staffing, and operational.
    b. To predict bike rentals based on time, weather, user behavior.

INTRODUCTION: WORKFLOWS

01 **Data Import & Cleaning**
Importing libraries & Data Cleaning

02 **Exploratory Data Analysis (EDA)**
Exploring the Dataset with Analysis and Visualization

03 **Feature Engineering**
For date-time or time dataset

04 **Model Training**
Training Models

05 **Model Evaluation**
Evaluation for Model Types that are used in the jupyter file

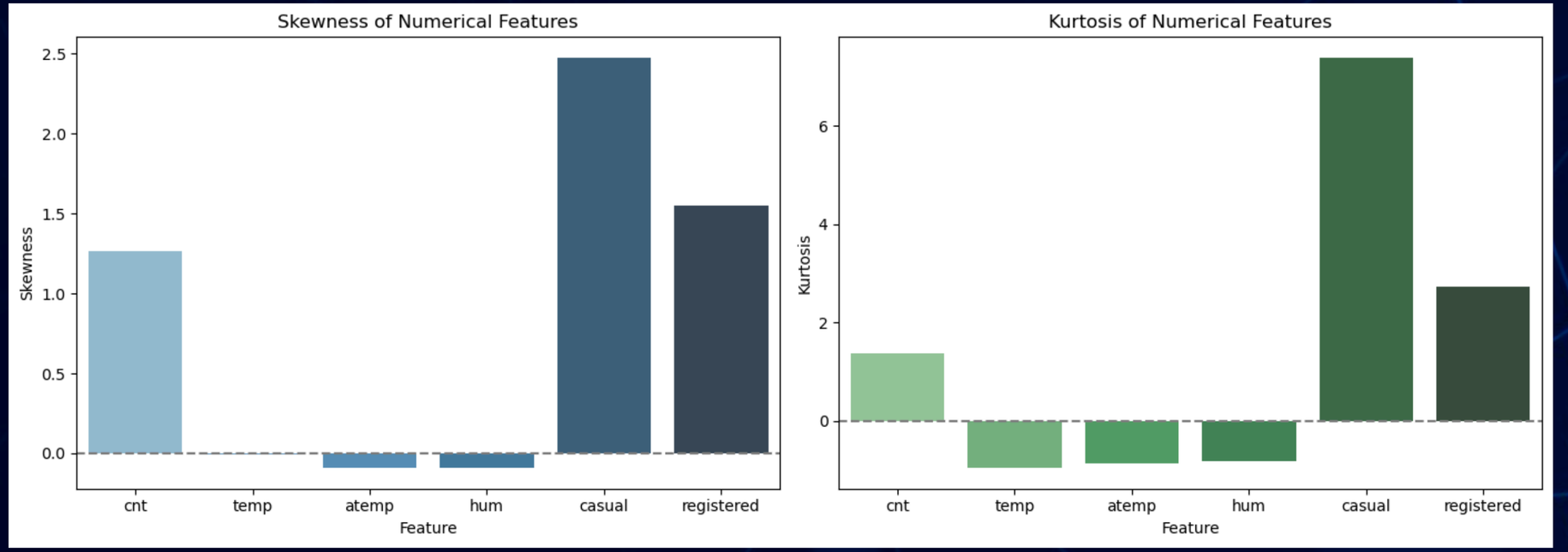
06 **Insights & Reccomendations**
Summary of the Jupyter file Project

EXPLORATORY DATA ANALYSIS

- No missing value
- No duplicated data

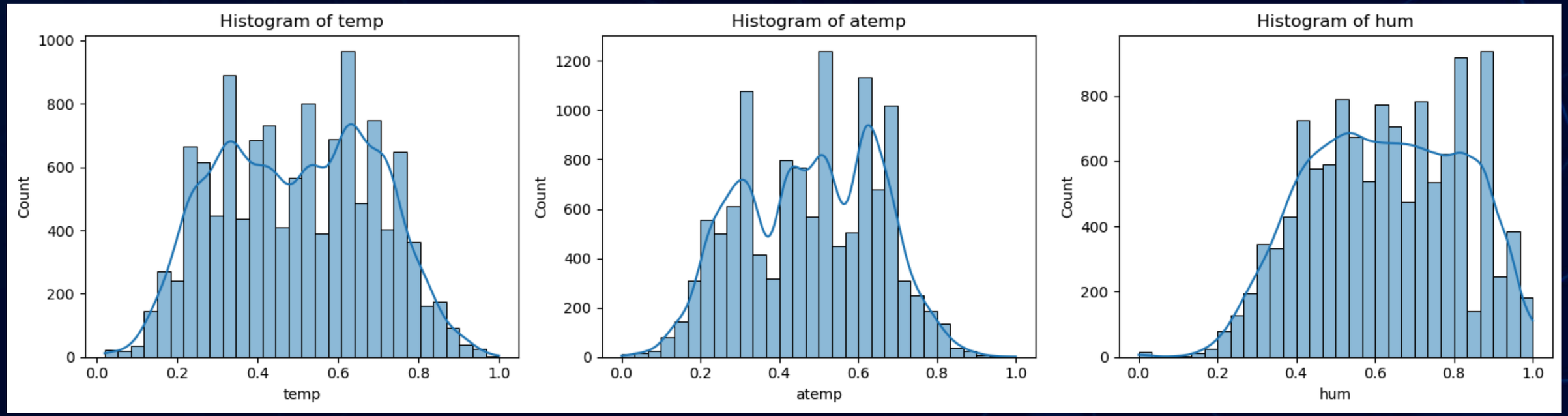
- *Bars above zero indicate right skewed and heavy tails.*

- *The distributions may need transformation for linear models to work effectively*



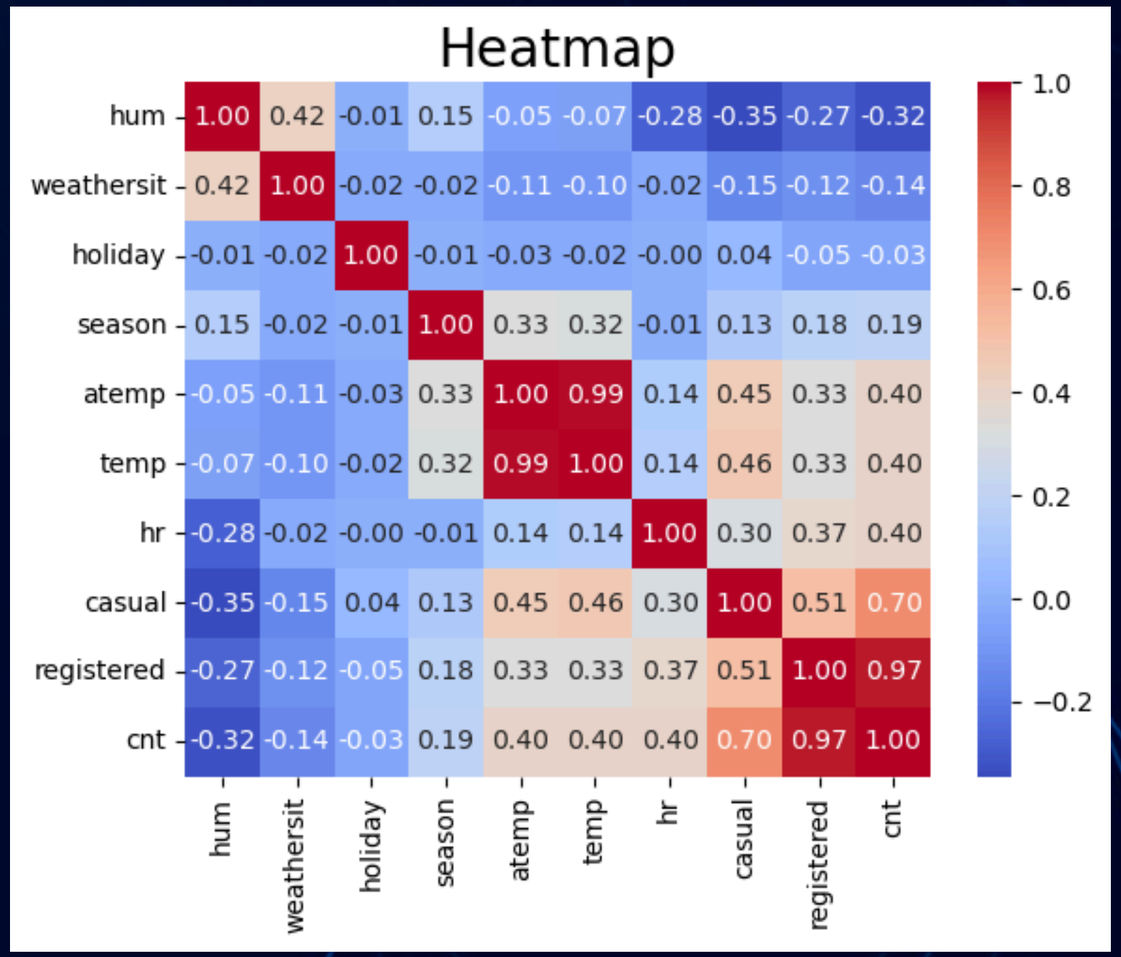
EXPLORATORY DATA ANALYSIS

HISTOGRAM OF:
TEMPT, ATEMPT, HUM



All three distributions appear continuous and likely normal but need checking for outliers.

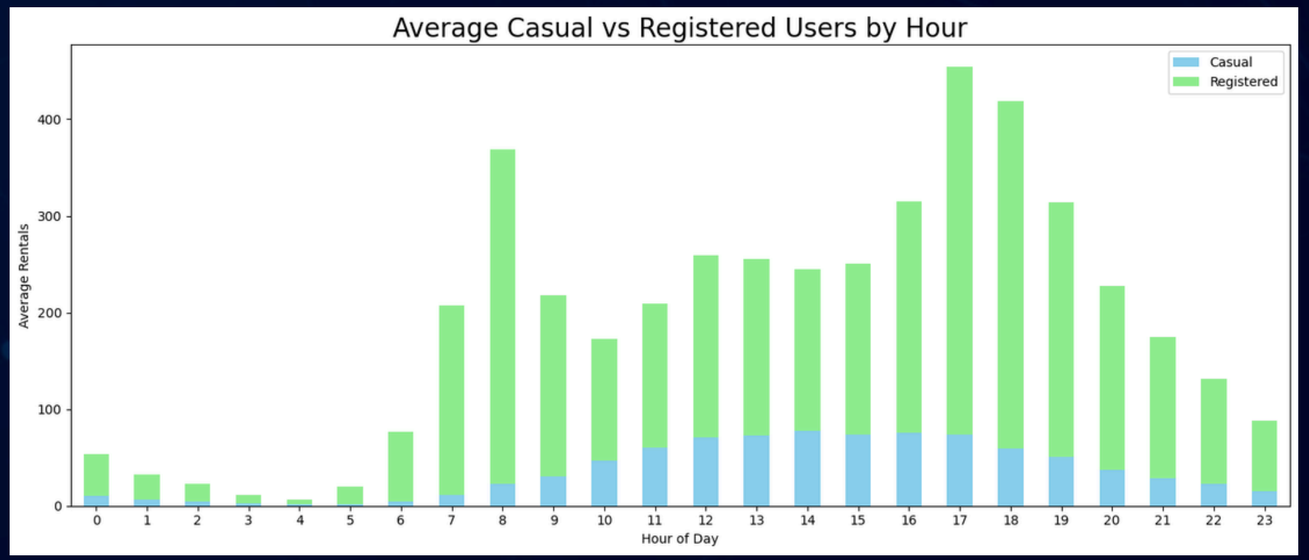
EXPLORATORY DATA ANALYSIS

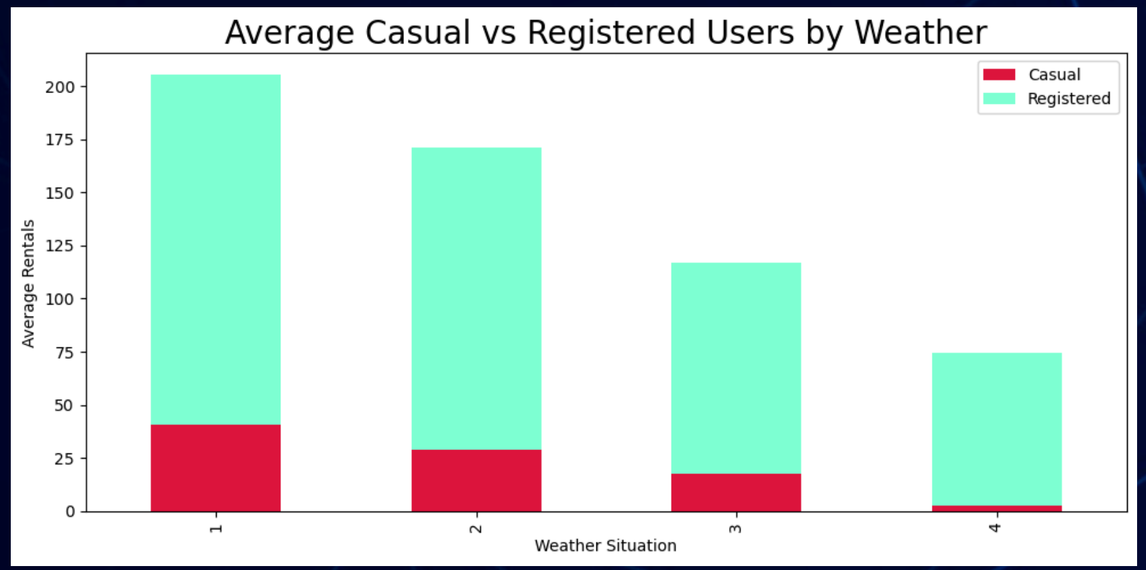


Heatmap:

- *Focusing on `season`, `weathersit`, `holiday`, `hr` columns only.*
- *High positive correlation between `registered` and `cnt`.*

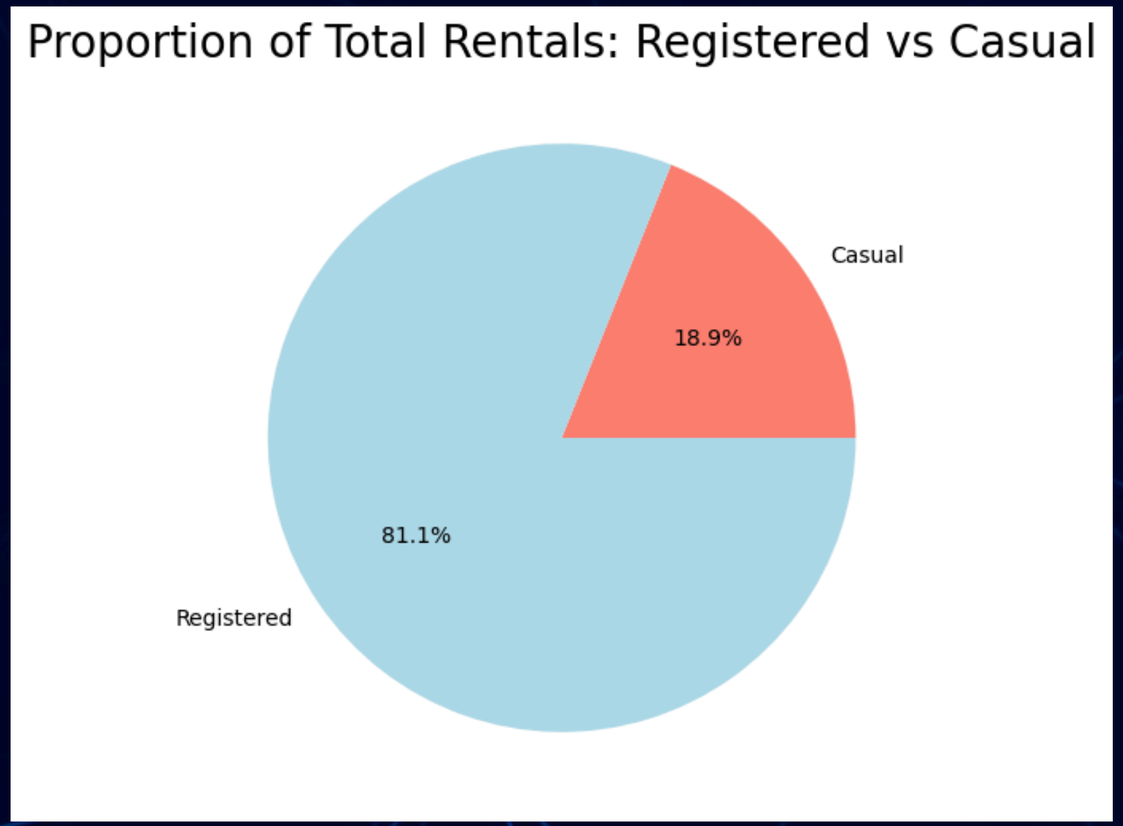
EXPLORATORY DATA ANALYSIS





- **HOURLY USAGE**: *Shows that registered users dominate peak commute hours (8AM, 5-6 PM)*
- **WEATHER USAGE**: *Casual usage drops significantly in worse weather.*

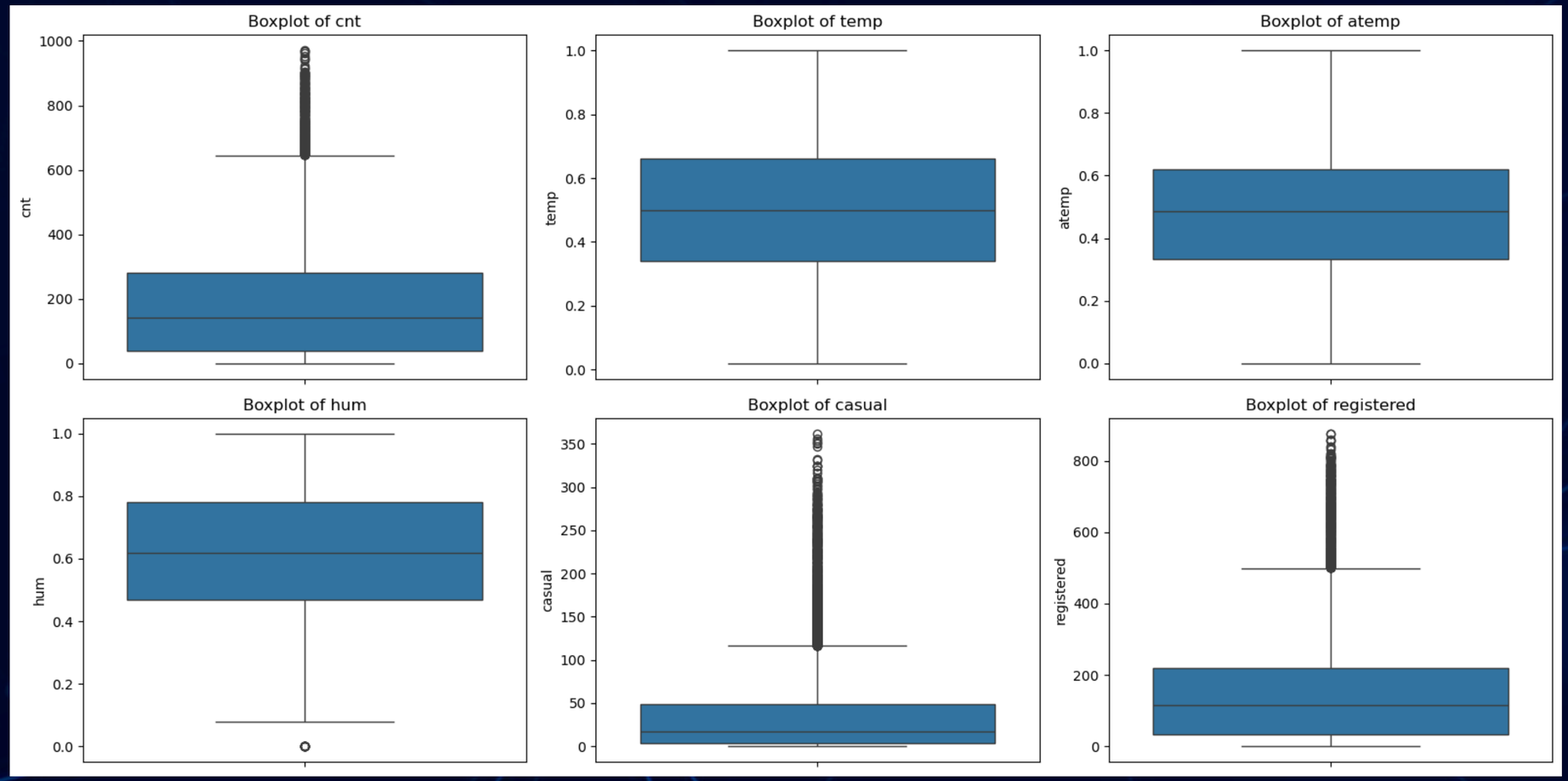
EXPLORATORY DATA ANALYSIS



Registered users make up a larger portion of total rentals (81.1 %)

EXPLORATORY DATA ANALYSIS

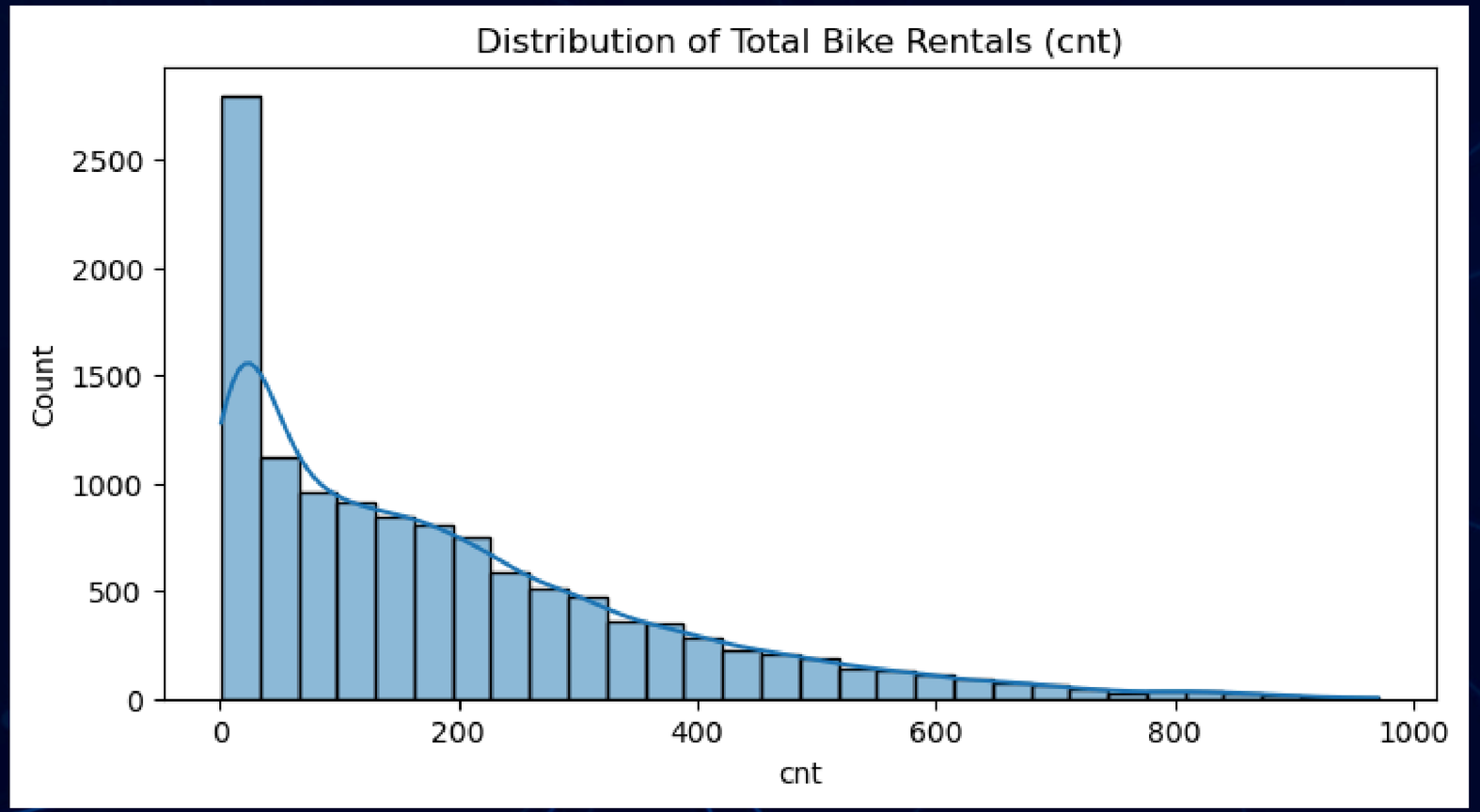
Boxplot of:
cnt,
temp,
atemp,
hum,
casual,
registered



- cnt (total rentals) is moderately right skewed with a few high-value outliers suggesting spikes in demand.
- temp, atemp and hum have stable measurements:
  - Because these features exhibit slight skewness alongside low outlier values.

EXPLORATORY DATA ANALYSIS

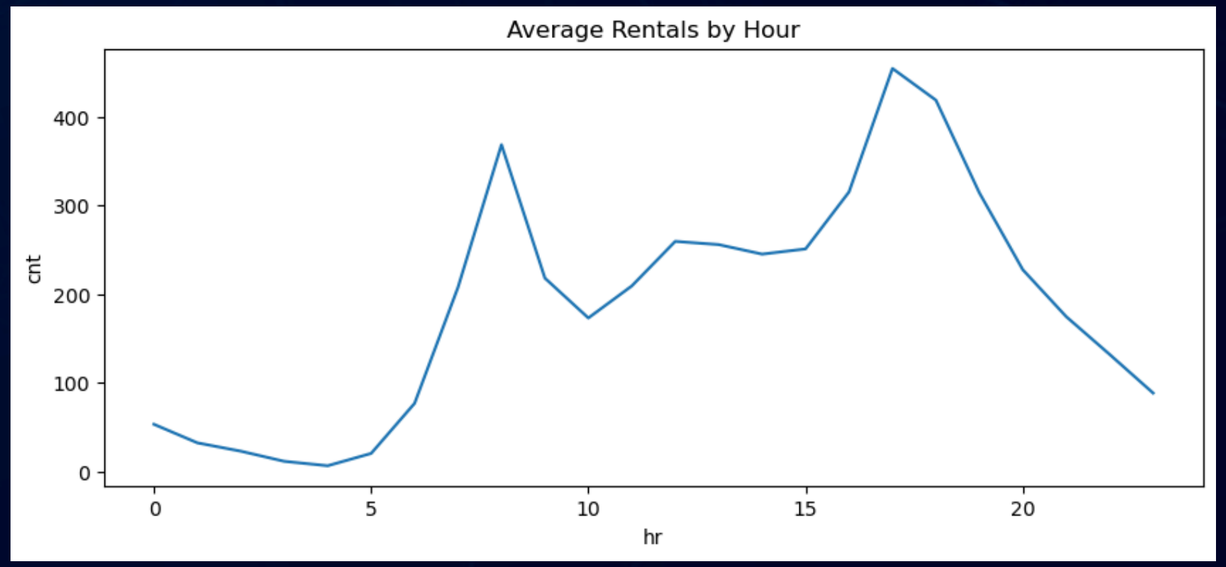
TOTAL BIKE RENTALS (cnt)

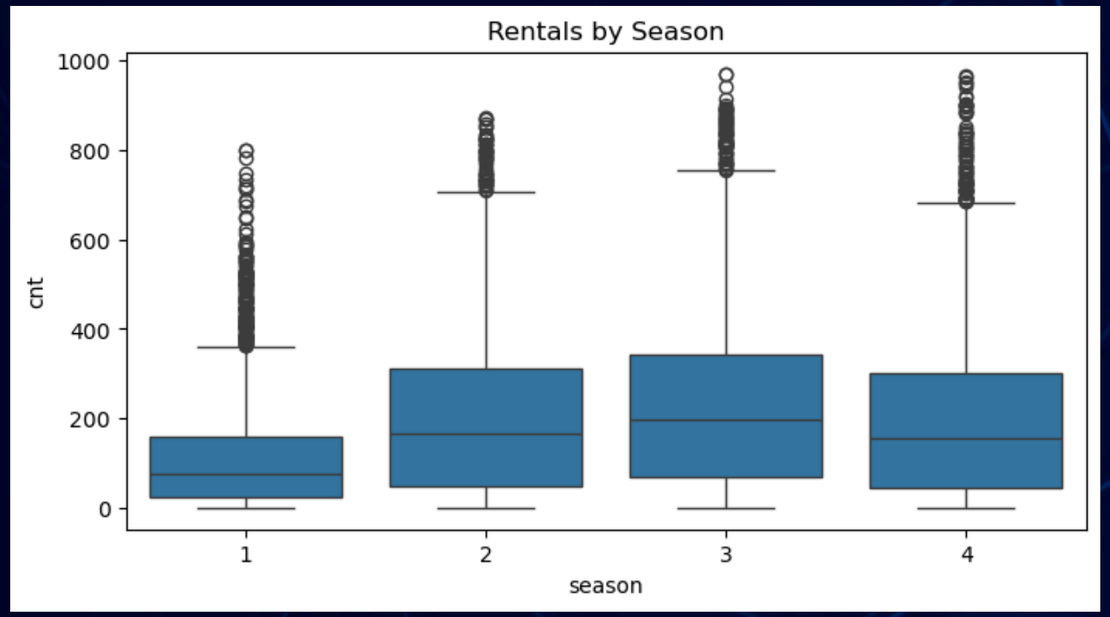


- cnt (total rentals) is moderately right skewed with a few high-value outliers suggesting spikes in demand.

AVG BY HOUR & SEASON

EXPLORATORY DATA ANALYSIS





- Hourly Rentals:
  - Peak at 8 AM and 5-6 PM — commuting hours.
- Seasonal:
  - Highest rentals in season 3 (Fall/Autumn).

FEATURE ENGINEERING

Understand each variable's role when choosing features for modeling.

Categorical Features:

- **Season: Captures seasonal trends affecting consumer behavior.**
- **Weathersit: Categorizes weather conditions impacting transportation.**
- **Holiday: Indicates holidays, influencing activity levels and traffic.**
- **Hr: Represents the hour of the day, capturing patterns.**

Numerical Features:

- **Temp: Influences behaviors like energy usage.**
- **Atemp: Represents apparent temperature, offering nuanced weather understanding.**
- **Hum: Affects outcomes such as comfort and machinery performance.**

TYPES OF MODELS AND MODEL EVALUATIONS

MODEL USED AND MODEL EVALUATION THAT ARE USED IN THIS PROJECT

- Linear Regression
- Ridge & Lasso Regression
- Decision Tree
- Random Forest
- Gradient Boosting

- MAPE
- RMSE
- MAE
- R2 SCORE

MODEL TYPES & EVALUATIONS

Performance Metrics:

Best Performer:

- **RMSE:** Gradient Boosting = 106.40
- **MAE:** Random Forest = 71.95
- **R^2:** Gradient Boosting = 0.64
- **MAPE:** Random Forest = 0.95

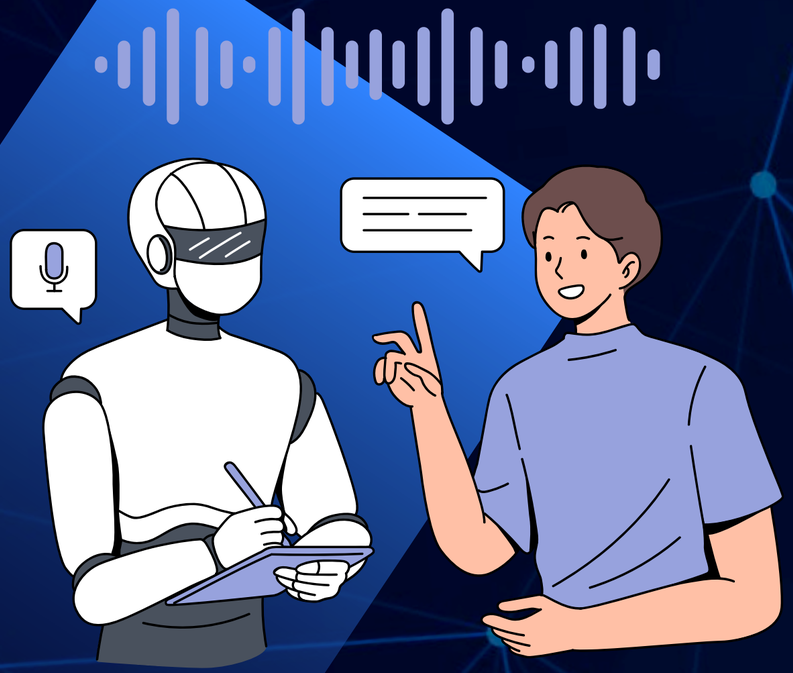
	RMSE	MAE	R2	MAPE
Linear Regression	109.601227	79.112957	0.614560	2.417644
Ridge	109.598880	79.074595	0.614576	2.413524
Lasso	114.321824	83.401041	0.580642	2.393122
Decision Tree	137.205304	88.825983	0.395956	1.077178
Random Forest	109.529622	71.954142	0.615063	0.947500
Gradient Boosting	106.404237	74.863830	0.636718	1.544035

CONCLUSION

DATA EXPLORATION	1. The target cnt has a right skew and thick tails which indicates the very high spike in demand	2. Registered members dominates the rentals , especially during commuting hours. Season: Rentals peak in the fall (season 3: Autumn/Fall)	3. The poor holidays and poor weather decrease rentals a lot.
PREPROCESSING STRATEGY	4. Categorical Features: Onehot encoded categories	5. Numerical Features: Normalized contributions	6. Target: cnt (total rentals)
EVALUATING THE MODEL	7. Best Model: Gradient Boosting with 106.40	8. Best Balance (Bias & Variance). Random Forest performed comparably, (MAPE)	9. Linear Mode (Linear, Ridge, Lasso) did poorly (due to non-linear relationship)
RECOMMENDATIONS	10. More Feature Engineering: Include Weekday, Workingday for better outcomes	11. Use RobustScaler insted of just StandardScaler (for outliers)	12. Ensemble Methods: Explore stacking Random Forest with Gradient Boosting

FINAL PAGE

THANK YOU
FOR YOUR ATTENTION



CREATED BY:
CLARINDA PUSPITAJATI