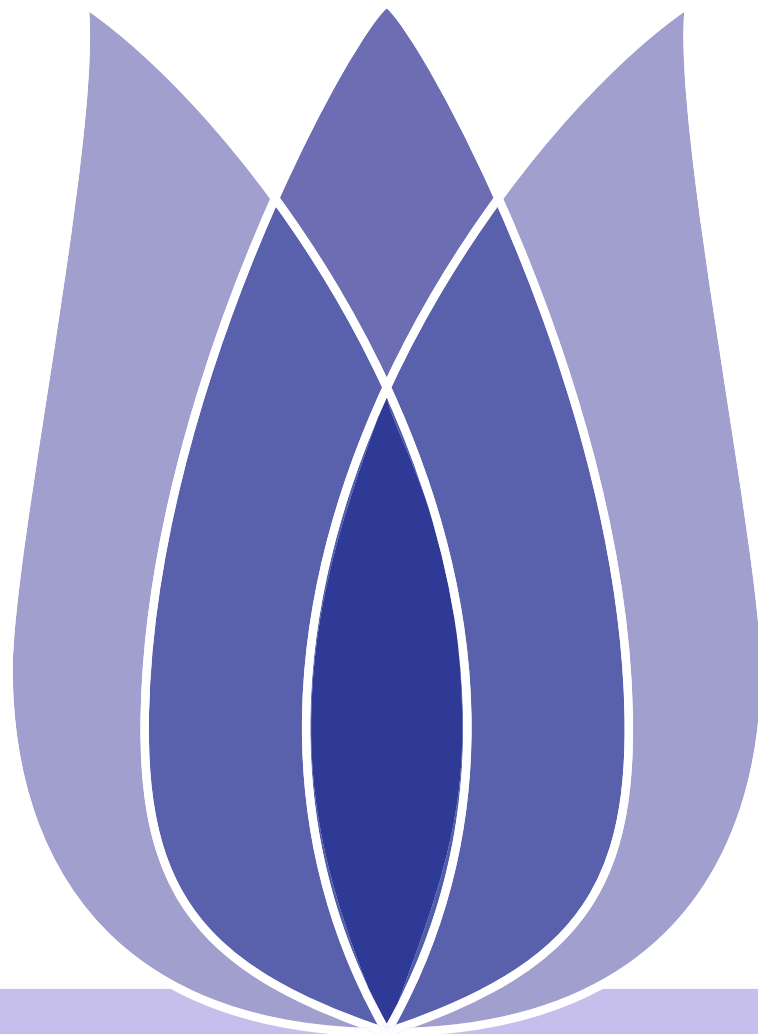


Bike Sharing Demand

Pratikshya Parajuli

Ministry of Finance
Government of Nepal

December 14, 2022





Overview

- [Problem Definition](#)
- [Exploratory Data Analysis](#)
- [Predictive Modelling](#)
- [Evaluation Results](#)
- [Conclusion](#)

- [Problem Definition](#)
- [Exploratory Data Analysis](#)
- [Predictive Modelling](#)
- [Evaluation Results](#)
- [Conclusion](#)



- Problem Definition
- Exploratory Data Analysis
- Predictive Modelling
- Evaluation Results
- Conclusion

Problem Definition



Bike Sharing Demand

- [Problem Definition](#)
- [Exploratory Data Analysis](#)
- [Predictive Modelling](#)
- [Evaluation Results](#)
- [Conclusion](#)

Defn

Bike sharing demand aims to forecast the use of the bikeshare system throughout the city.

- Dataset comprises of the hourly rental data spanning two year having data fields such as **datetime**, **seasontemp** etc.
- predict the **total count** of bikes rented during each hour



Data Summary

- Problem Definition
- Exploratory Data Analysis
- Predictive Modelling
- Evaluation Results
- Conclusion

- Training Set provides the data and usage of the first 19 days of each month
- Test Set provides the data from the 20th to the end of the month

```
train_df.head()
```

	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
0	2011-01-01 00:00:00	1	0	0	1	9.84	14.395	81	0.0	3	13	16
1	2011-01-01 01:00:00	1	0	0	1	9.02	13.635	80	0.0	8	32	40
2	2011-01-01 02:00:00	1	0	0	1	9.02	13.635	80	0.0	5	27	32
3	2011-01-01 03:00:00	1	0	0	1	9.84	14.395	75	0.0	3	10	13
4	2011-01-01 04:00:00	1	0	0	1	9.84	14.395	75	0.0	0	1	1

```
train_df.dtypes
datetime      object
season        int64
holiday       int64
workingday    int64
weather       int64
temp          float64
atemp         float64
humidity      int64
windspeed     float64
casual        int64
registered    int64
count         int64
dtype: object
```

- No missing values in the dataset



- [Problem Definition](#)
- [Exploratory Data Analysis](#)
- [Predictive Modelling](#)
- [Evaluation Results](#)
- [Conclusion](#)

Exploratory Data Analysis



Dataset Description

Problem Definition
Exploratory Data Analysis
Predictive Modelling
Evaluation Results
Conclusion

■ Data Fields

- ◆ datetime - hourly date + timestamp
- ◆ season - 1 = spring, 2 = summer, 3 = fall, 4 = winter
- ◆ holiday - whether the day is considered a holiday
- ◆ workingday - whether the day is neither a weekend nor holiday
- ◆ weather - 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 - temperature in Celsius
- ◆ atemp - "feels like" temperature in Celsius
- ◆ humidity - relative humidity
- ◆ windspeed - wind speed
- ◆ casual - number of non-registered user rentals initiated
- ◆ registered - number of registered user rentals initiated
- ◆ count - number of total rentals



Data Preprocessing

Problem Definition
Exploratory Data Analysis
Predictive Modelling
Evaluation Results
Conclusion

- Season, holiday, workingday, and weather are int type, but category type is more suitable, so convert to category type
- split datetime into separate year, month, day, hour and dayofweek columns
- Analyse the missing values
- Remove datetime





Analysis of the features

- [Problem Definition](#)
- [Exploratory Data Analysis](#)
- [Predictive Modelling](#)
- [Evaluation Results](#)
- [Conclusion](#)

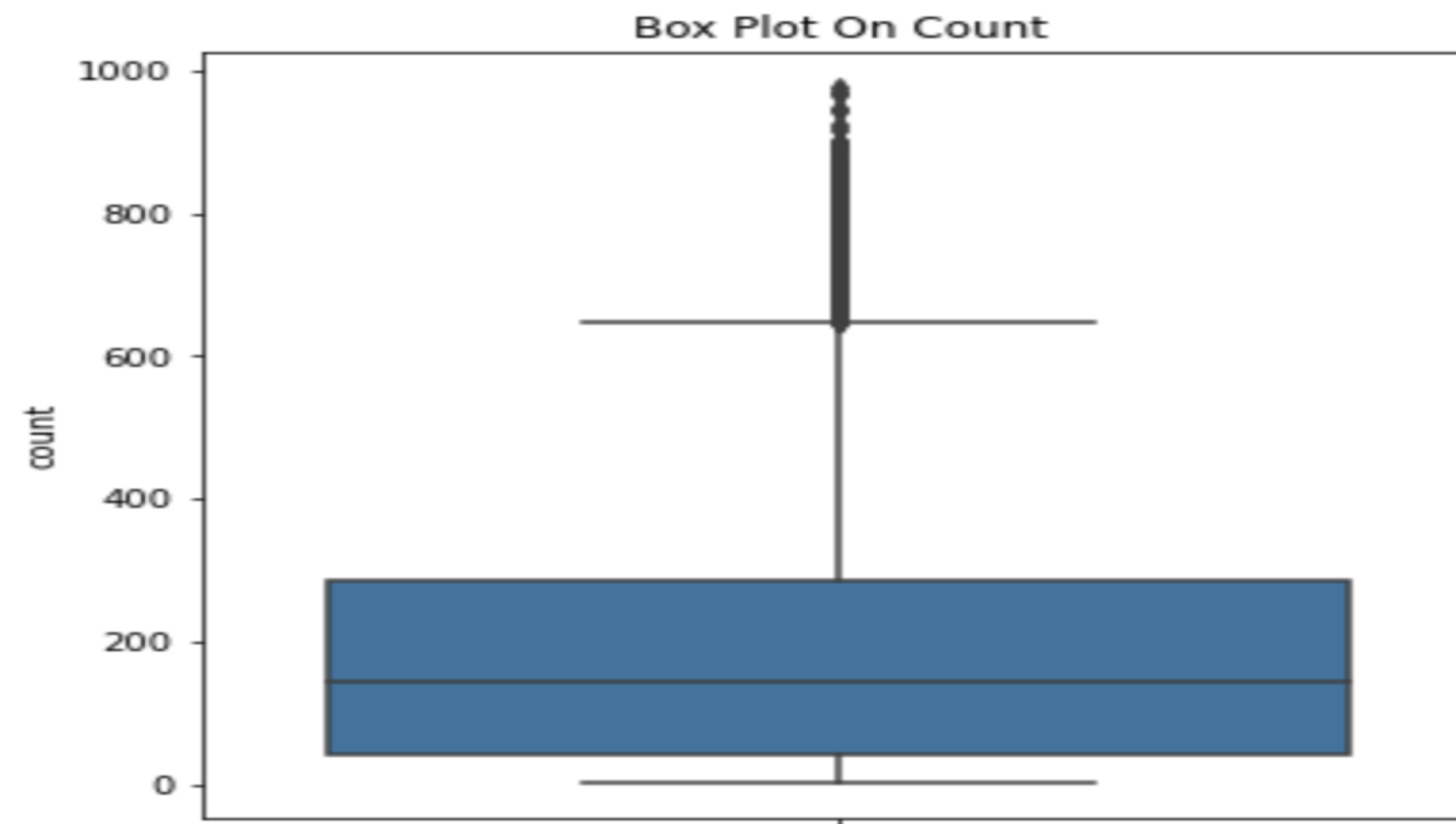
- **Category Features** : Hour, weekday, month, Season, holiday, Workingday, weather
- **Numerical Features**: temp, atemp, humidity, windspeed, registered and causal



Outliers Analysis

Problem Definition
Exploratory Data Analysis
Predictive Modelling
Evaluation Results
Conclusion

- From the Boxplot of count column, it is clearly visible that most of the data lies between 30-300 and a huge numbers of outliers are presesnt in the plot.
- Use the 3 sigma principle to remove outliers





Data Visualization between count vs. month and season

- Problem Definition
- Exploratory Data Analysis
- Predictive Modelling
- Evaluation Results
- Conclusion

- The use of shared bicycles from November to April will be a little less than in other months, which may be due to seasonal reasons.

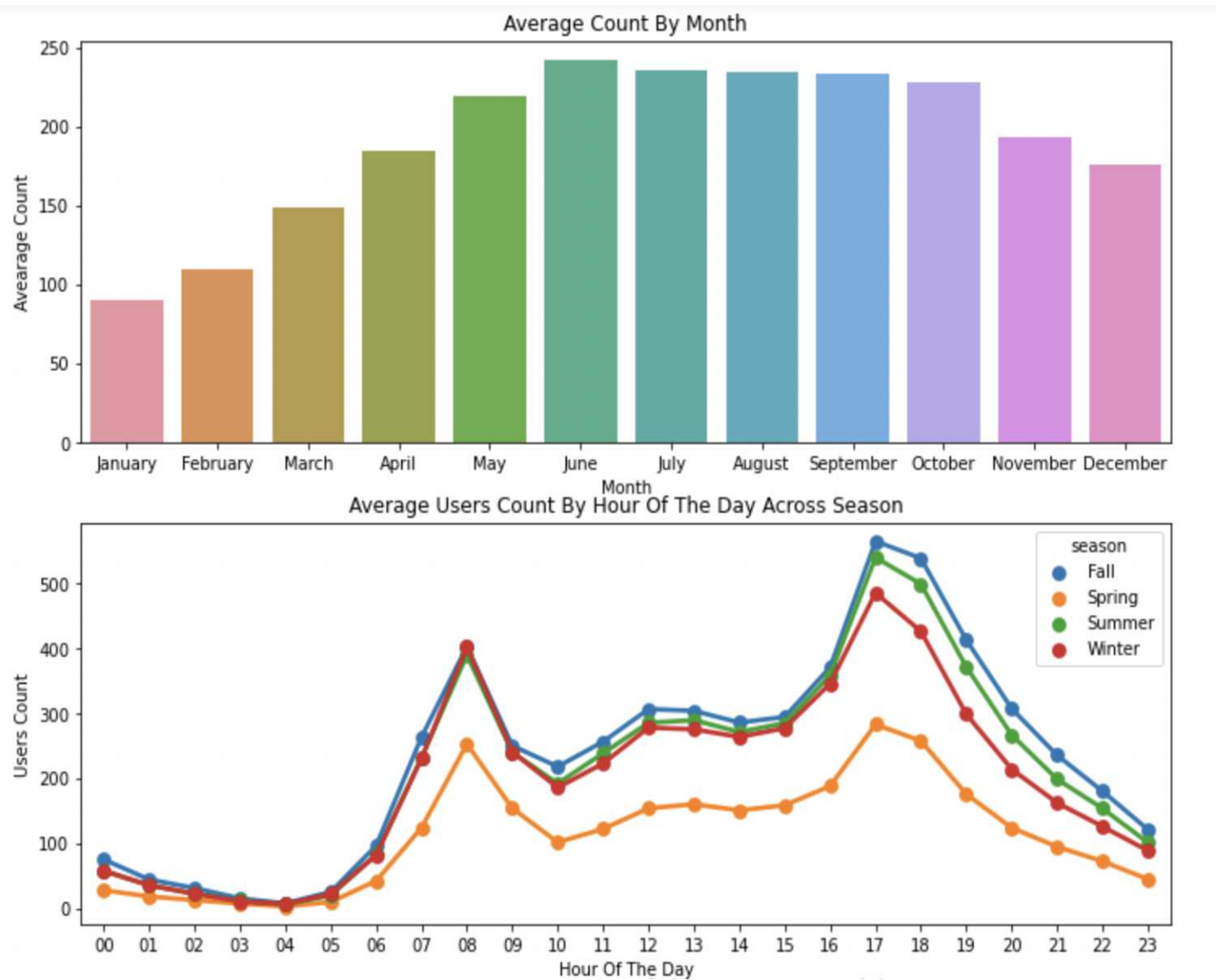


Figure 1: Data Visualization between count vs. month and season



Data Visualization between count vs. weekdays and usertype

- Problem Definition
- Exploratory Data Analysis
- Predictive Modelling
- Evaluation Results
- Conclusion

- The use of shared bicycles in winter and spring is relatively small compared to summer and autumn, which is mutually confirmed with the conclusions generated in the above months.

Histogram Plot of Count

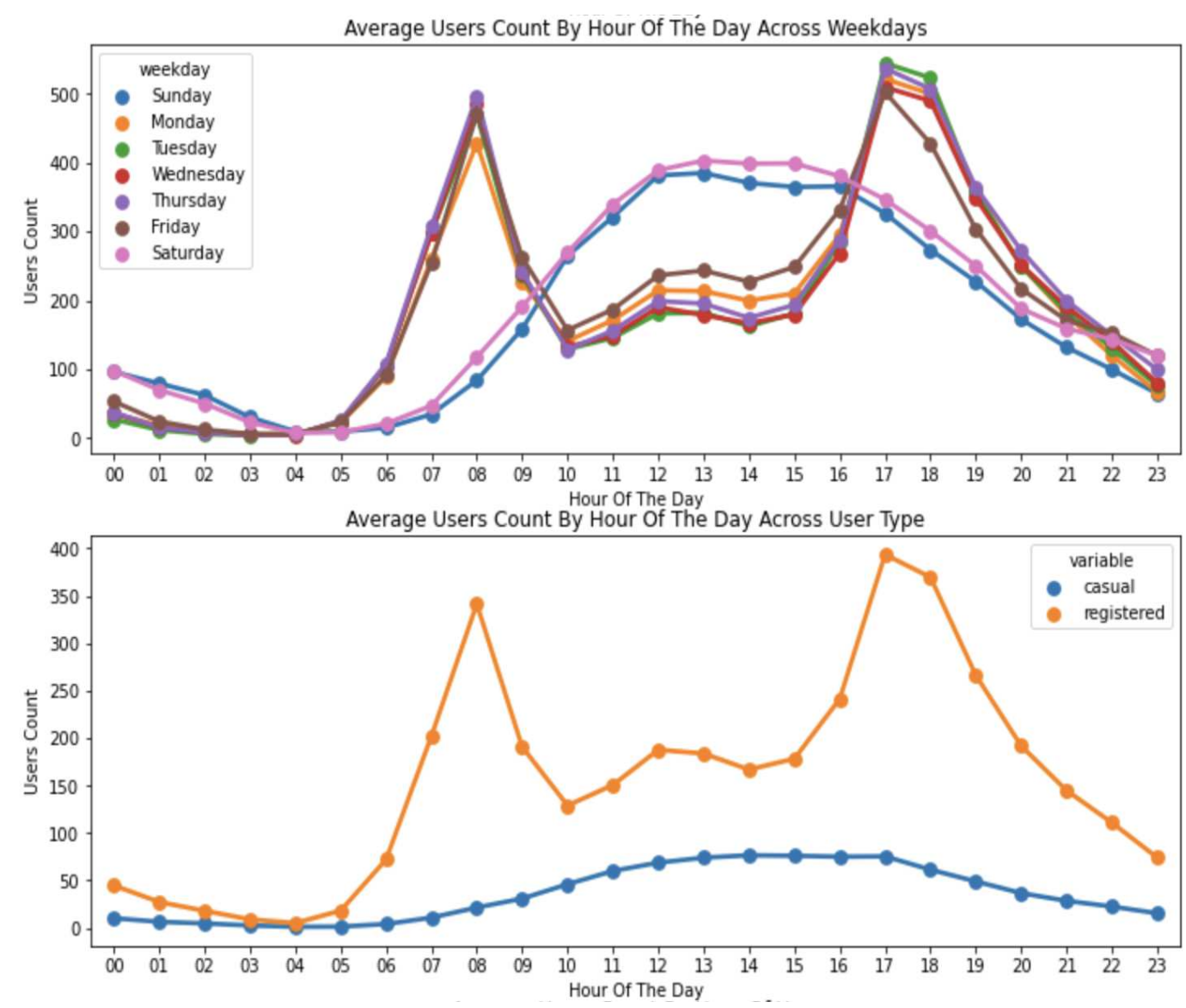


Figure 2: Data Visualization between count vs. weekdays and usertype



Data Visualization of count and [temp, windspeed, humidity]

- Problem Definition
- Exploratory Data Analysis
- Predictive Modelling
- Evaluation Results
- Conclusion

- From the scatter plot, windspeed has many values of 0 and is separated from other values, so it is predicted that this is not an actual measured value.

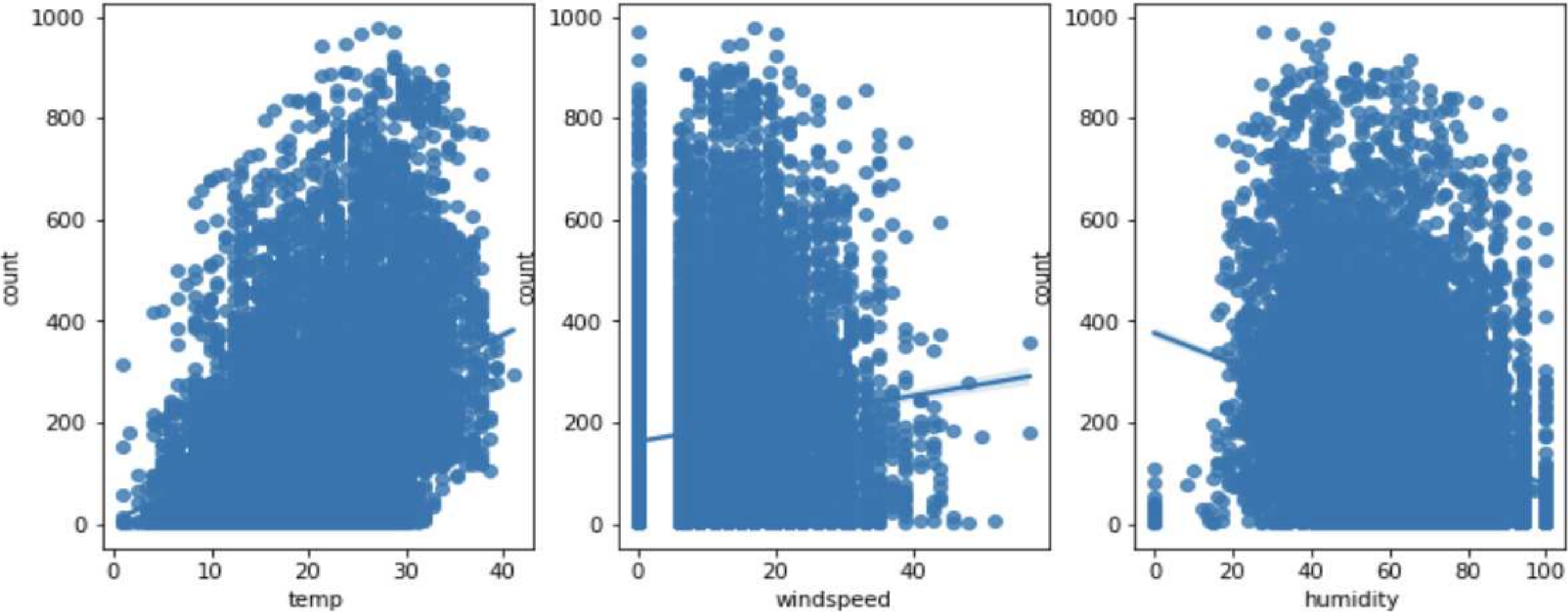


Figure 3: Scatter Plot



Correlation Matrix

- [Problem Definition](#)
- [Exploratory Data Analysis](#)
- [Predictive Modelling](#)
- [Evaluation Results](#)
- [Conclusion](#)

- Heatmap of the correlation matrix between count and [temp, atemp, humidity, windspeed, casual, registered].

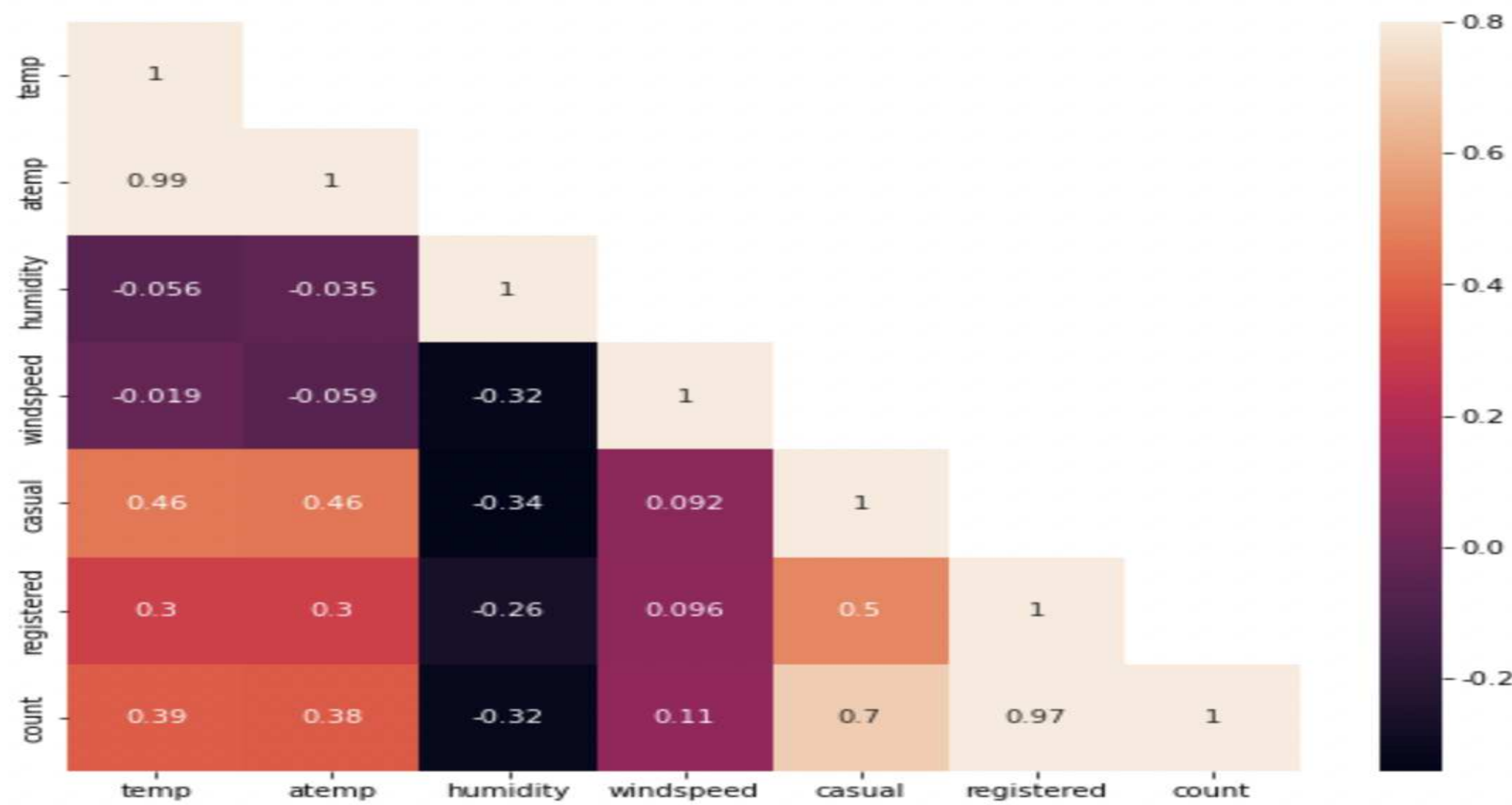


Figure 4: Heatmap



Feature Engineering

Problem Definition
Exploratory Data Analysis
Predictive Modelling
Evaluation Results
Conclusion

Based on the above heatmap, we can see that some of the features have no relation with the response variable. we can drop those columns.

- **humidity, temp** are negatively correlated with count
- There is a strong correlation between **temp** and **atemp**, if both are included in the model, it will cause multicollinearity problems, so one of the features must be deleted. We remove the atemp feature because it has a weaker correlation with count than temp.
- **Casual, Registered** are not considered and removed during model building
- **humidity, temp** and **windspeed** features are considered during future modelling
- fill in the zero values in the windspeed feature: usage is high when the wind speed is 0, which may be caused by null filling. Therefore, a random forest model is used here to fill in the zero values.



Skewness in Data Distribution

- Problem Definition
- Exploratory Data Analysis
- Predictive Modelling
- Evaluation Results
- Conclusion

- From the histogram plot, we can say that count data is skewed (concentrated on the one side) and the data is not equally distributed.
- Solution is to log-transform the count variable after removing outlier data points.

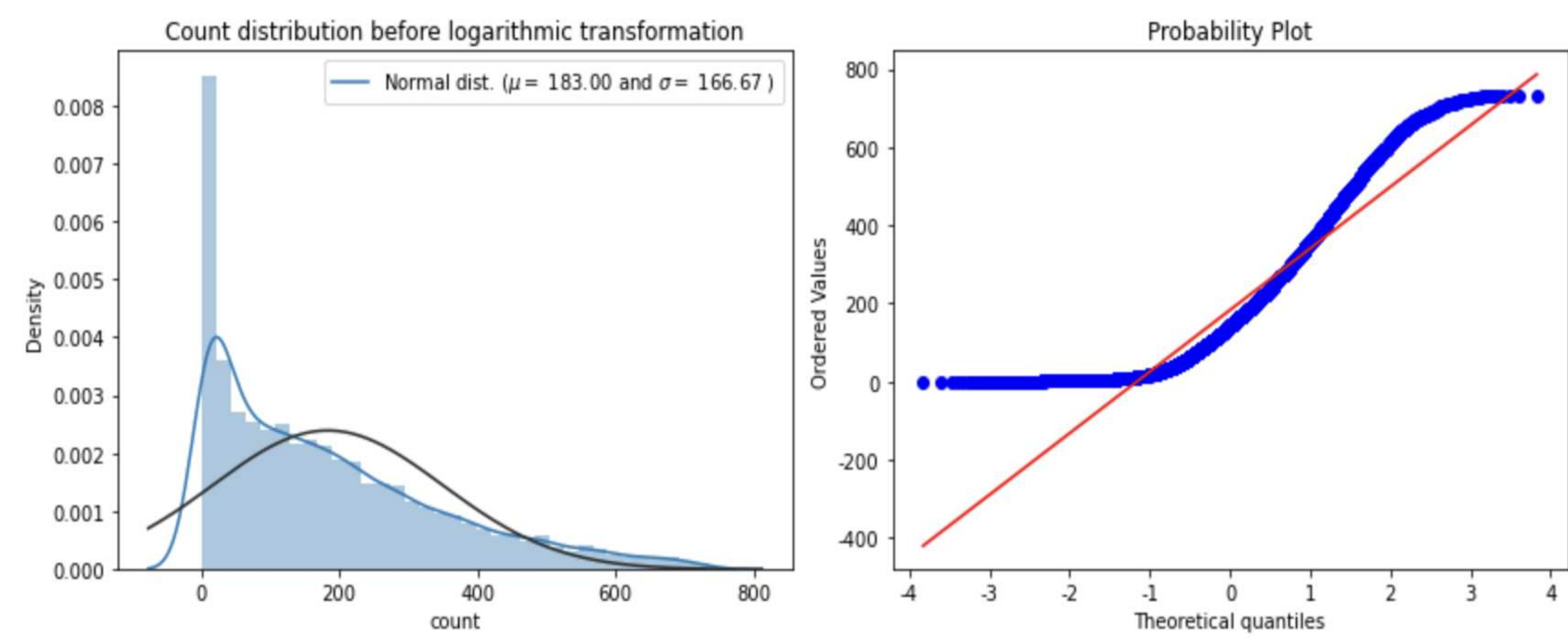


Figure 5: Count distribution before logarithmic transformation



Solution to Skewness in Data Distribution

- Problem Definition
- Exploratory Data Analysis
- Predictive Modelling
- Evaluation Results
- Conclusion

- Solution is to log-transform the count variable after removing outlier data points.

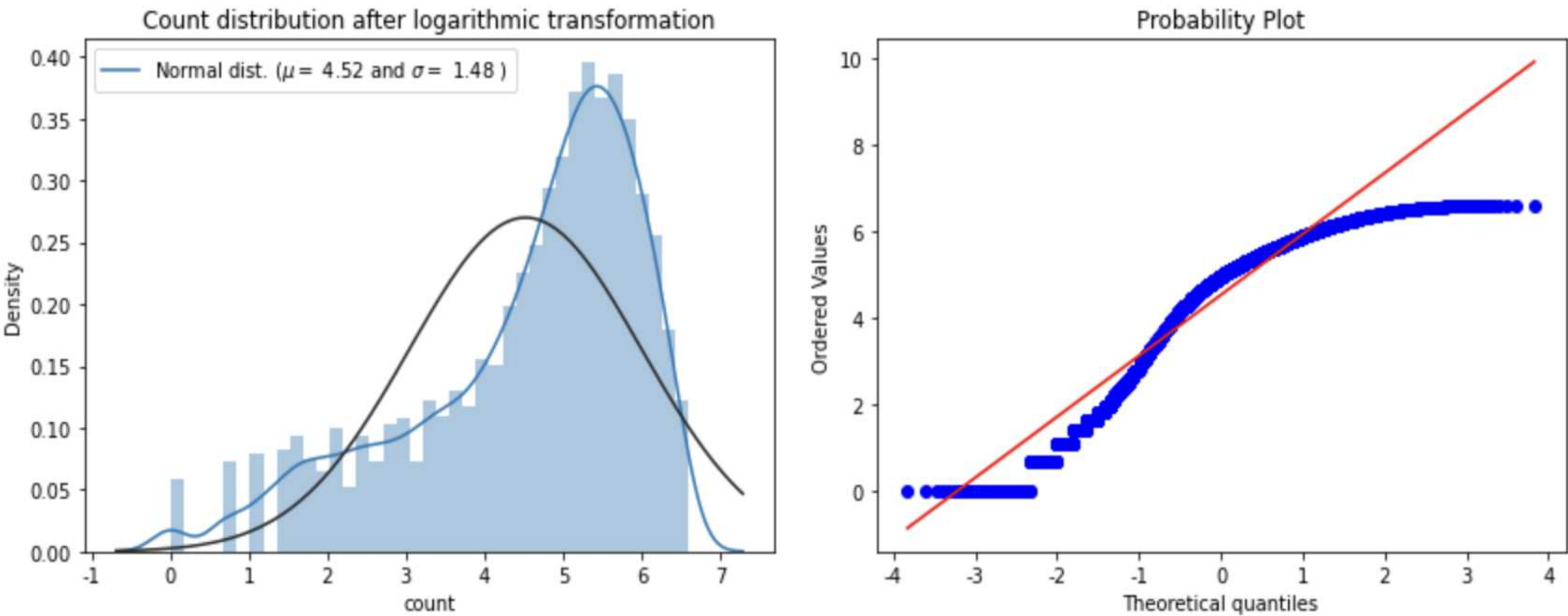


Figure 6: Count distribution after logarithmic transformation



- [Problem Definition](#)
- [Exploratory Data Analysis](#)
- [Predictive Modelling](#)
- [Evaluation Results](#)
- [Conclusion](#)

Predictive Modelling



Data Preparation

- [Problem Definition](#)
- [Exploratory Data Analysis](#)
- [Predictive Modelling](#)
- [Evaluation Results](#)
- [Conclusion](#)

- Split the dataset into train set and test set
- Train Data size : 0.7
- Test Data size : 0.3



Predictions with linear Model

- [Problem Definition](#)
- [Exploratory Data Analysis](#)
- [Predictive Modelling](#)
- [Evaluation Results](#)
- [Conclusion](#)

- Linear Regression
- Ridge Regression
- Lasso Regression
- Logistic Regression
- ElasticNet



Predictions with Ensemble learning Models

- [Problem Definition](#)
- [Exploratory Data Analysis](#)
- [Predictive Modelling](#)
- [Evaluation Results](#)
- [Conclusion](#)

- Bagging Regressor
- Random Forest Regressor
- Gradient Boosting Regressor
- AdaBoost Regressor



- [Problem Definition](#)
- [Exploratory Data Analysis](#)
- [Predictive Modelling](#)
- [Evaluation Results](#)**
- [Conclusion](#)

Evaluation Results



Evaluation

- [Problem Definition](#)
- [Exploratory Data Analysis](#)
- [Predictive Modelling](#)
- [Evaluation Results](#)
- [Conclusion](#)

- Evaluation Indicators: root mean square error is required (Root Mean Squared Logarithmic Error, RMSLE) to evaluate the quality of the model.

$$RMSLE = \sqrt{\frac{1}{n} \sum_{i=1}^n [\log(p_i + 1) - \log(a_i + 1)]^2}$$

Among them, n is the number of samples in the test set, p_i is the test value, and a_i is the actual value. The smaller the root mean square error, the better the fitting effect of the data, and the closer the test value is to the actual value.





Model Evaluation Results

- [Problem Definition](#)
- [Exploratory Data Analysis](#)
- [Predictive Modelling](#)
- [Evaluation Results](#)
- [Conclusion](#)

Table 1: Model Evaluation Results

Model	Accuracy
Random Forest Regression	0.376319
Bagging Regression	0.395248
GBRT	0.430378
AdaBoost Regression	0.703528
Ridge Regression	1.045335
Lasso Regression	1.045453
ElasticNet Regression	1.045489
Linear Regression	1.046341
Logistic Regression	1.131105



- [Problem Definition](#)
- [Exploratory Data Analysis](#)
- [Predictive Modelling](#)
- [Evaluation Results](#)
- [Conclusion](#)**

Conclusion



Conclusion

- [Problem Definition](#)
- [Exploratory Data Analysis](#)
- [Predictive Modelling](#)
- [Evaluation Results](#)
- [Conclusion](#)

- Basic modelling of the data
- Results can be further enhanced



Questions?

- [Problem Definition](#)
- [Exploratory Data Analysis](#)
- [Predictive Modelling](#)
- [Evaluation Results](#)
- [Conclusion](#)



Contact Information

Pratikshya Parajuli
Ministry of Finance
Government of Nepal



MISSPRATIKSHYA@GMAIL.COM

