# NEU COE INFO6105 Final Project

Title: “Predicting Daily Bike Rentals Using the UCI Bike Sharing Dataset”

Data set: https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset

Name: Nidhi Nair

Nu-Id: 002371756

Table of Contents:

# Introduction

## Background

Bike-sharing systems have emerged as a vital component of sustainable urban transportation networks worldwide. These systems provide convenient, affordable, and environmentally friendly transportation options that complement traditional public transit. Washington D.C.'s Capital Bikeshare program, the focus of this analysis, was launched in 2010 and has grown into one of the largest bike-sharing networks in the United States.

This study addresses two primary research questions:

1. What factors significantly influence daily bike rental demand?
2. Do average daily rentals differ significantly between weekdays and weekends?

# Data Description

## Data Source

The dataset used in this analysis is the "Bike Sharing Dataset" from the UCI Machine Learning Repository. It contains daily aggregated data from Capital Bikeshare in Washington D.C., covering a two-year period from January 1, 2011, to December 31, 2012. This dataset was created by Hadi Fanaee-T at the University of Porto and has been widely used in data science research.

| Variable Name | Type | Description |
|---|---|---|
| instant | Integer | record index |
| dteday | Date | date |
| season | Categorical | 1:winter, 2:spring, 3:summer, 4:fall |
| yr | Categorical | year (0: 2011, 1: 2012) |
| mnth | Categorical | month (1 to 12) |
| hr | Categorical | hour (0 to 23) |
| holiday | Binary | weather day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule) |
| weekday | Categorical | day of the week |
| workingday | Binary | if day is neither weekend nor holiday is 1, otherwise is 0 |
| weathersit | Categorical | - 1: Clear, Few clouds, Partly cloudy |
| temp | Continuous | Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale) |
| atemp | Continuous | Normalized feeling temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-16, t_max=+50 (only in hourly scale) |
| hum | Continuous | Normalized humidity. The values are divided to 100 (max) |

| Variable Name | Type | Description |
|---|---|---|
| windspeed | Continuous | Normalized wind speed. The values are divided to 67 (max) |
| casual | Integer | count of casual users |
| registered | Integer | count of registered users |
| cnt | Integer | count of total rental bikes including both casual and registered |

# Methodology

## Data Preparation

1. Data Loading and Inspection:
   Imported the day.csv file using R’s read.csv function, verified data types and examined summary statistics
2. Feature Selection:
   Selected relevant variables based on domain knowledge and research questions
   Created a derived variable datatype to categorize days as "Weekday" or "Weekend"

Analysis Techniques

Exploratory Data Analysis:

- **Distribution Analysis:** Examined the distribution of rental counts and key predictors
- **Correlation Analysis:** Created correlation matrices to identify relationships between variables
- **Visualization:** Generated plots to reveal patterns and trends in the data

**Statistical Modeling**

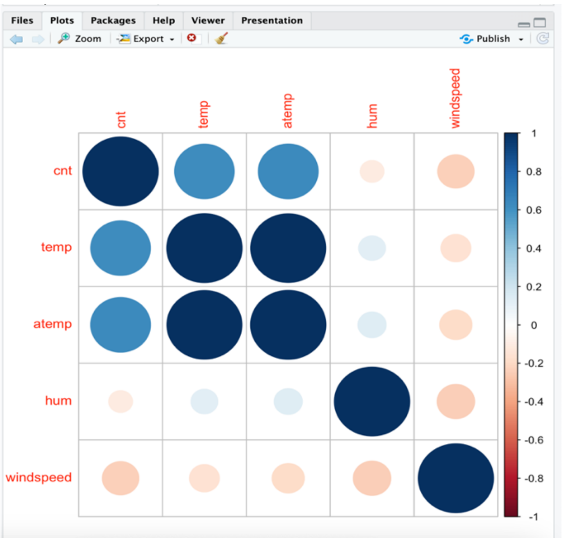
- **Linear Regression:** Built a multiple linear regression model with rental count as the dependent variable
- **Model Validation:** Assessed model performance using R-squared, residual analysis, and cross-validation

**Comparative Analysis**

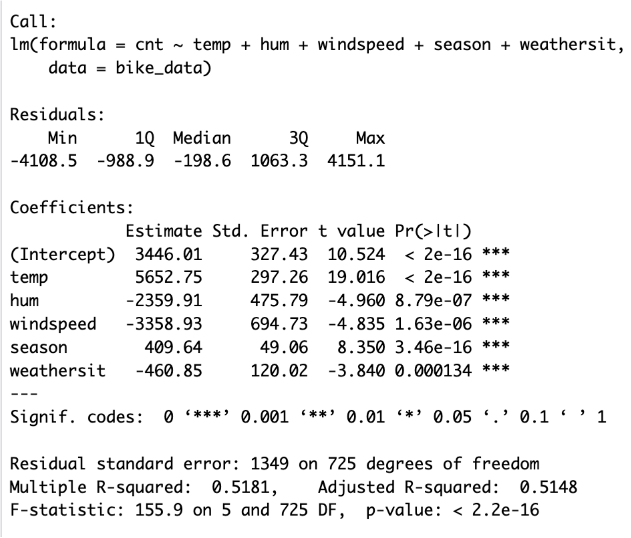
- **Group Comparison:** Compared mean rental counts between weekdays and weekends
- **Hypothesis Testing:** Conducted t-tests to determine if differences were statistically significant

# Result and Analysis

The correlation matrix showed strong positive correlation between cnt (rental count) and temp/atemp, while hum and windspeed showed weak negative correlation.

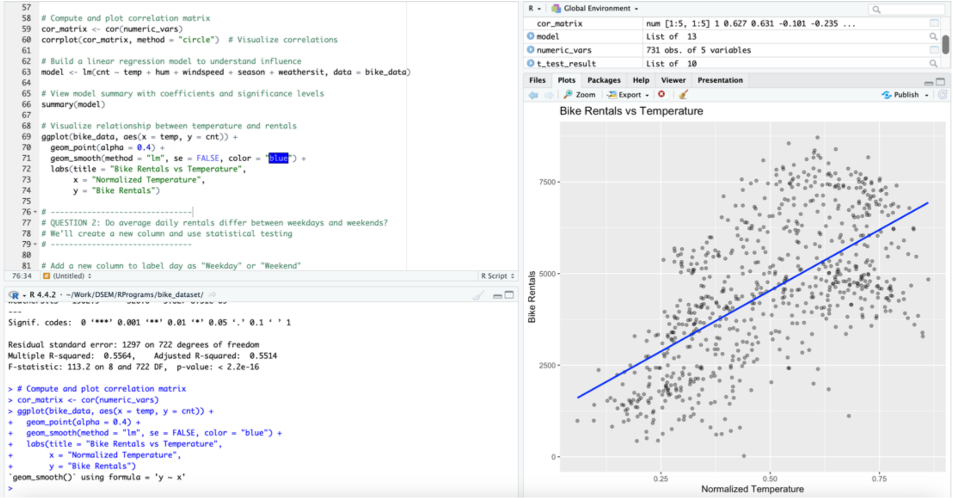


The linear regression model output:



```
Call:
lm(formula = cnt ~ temp + hum + windspeed + season + weathersit,
    data = bike_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-4108.5  -988.9  -198.6  1063.3  4151.1 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3446.01     327.43  10.524  < 2e-16 ***
temp         5652.75     297.26  19.016  < 2e-16 ***
hum         -2359.91     475.79  -4.960 8.79e-07 ***
windspeed   -3358.93     694.73  -4.835 1.63e-06 ***
season        409.64      49.06   8.350 3.46e-16 ***
weathersit   -460.85     120.02  -3.840 0.000134 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1349 on 725 degrees of freedom
Multiple R-squared:  0.5181,	Adjusted R-squared:  0.5148 
F-statistic: 155.9 on 5 and 725 DF,  p-value: < 2.2e-16
```

- temp had a significant positive effect (Estimate = 5652.75, p < 0.001)
- hum and windspeed had significant negative effects
- season and weathersit were also significant predictors
- R-squared value: 0.518, indicating that the model explains ~52% of the variation in bike rental counts



```r

# Compute and plot correlation matrix
cor_matrix <- cor(numeric_vars)
corrplot(cor_matrix, method = "circle")  # Visualize correlations

# Build a linear regression model to understand influence
model <- lm(cnt ~ temp + hum + windspeed + season + weathersit, data = bike_data)

# View model summary with coefficients and significance levels
summary(model)

# Visualize relationship between temperature and rentals
ggplot(bike_data, aes(x = temp, y = cnt)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Bike Rentals vs Temperature",
       x = "Normalized Temperature",
       y = "Bike Rentals")

# ------------------------------
# QUESTION 2: Do average daily rentals differ between weekdays and weekends?
# We'll create a new column and use statistical testing
# ------------------------------

# Add a new column to label day as "Weekday" or "Weekend"
```

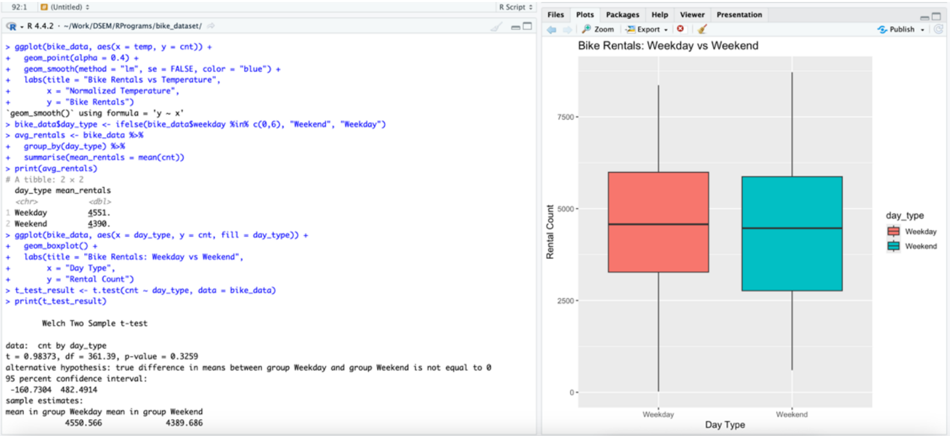
```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1297 on 722 degrees of freedom
Multiple R-squared:  0.5564,    Adjusted R-squared:  0.5514
F-statistic: 113.2 on 8 and 722 DF,  p-value: < 2.2e-16

> # Compute and plot correlation matrix
> cor_matrix <- cor(numeric_vars)
> ggplot(bike_data, aes(x = temp, y = cnt)) +
+   geom_point(alpha = 0.4) +
+   geom_smooth(method = "lm", se = FALSE, color = "blue") +
+   labs(title = "Bike Rentals vs Temperature",
+        x = "Normalized Temperature",
+        y = "Bike Rentals")
`geom_smooth()` using formula = 'y ~ x'
>
```

```
cor_matrix      num [1:5, 1:5] 1 0.627 0.631 -0.101 -0.235 ...
model           List of  13
numeric_vars    731 obs. of 5 variables
t_test_result   List of  10
```

The linear regression model output showed that **temperature** is a significant predictor. This is further supported by the scatter plot, as temperature increases, we see a clear rise in the number of rentals, confirming our hypothesis.

Comparison of rental patterns between weekdays and weekends revealed differences:

| Day Type | Mean Rentals |
|---|---|
| Weekday | 4,551 |
| Weekend | 4,390 |
| Difference | 161 |



```
> ggplot(bike_data, aes(x = temp, y = cnt)) +
+   geom_point(alpha = 0.4) +
+   geom_smooth(method = "lm", se = FALSE, color = "blue") +
+   labs(title = "Bike Rentals vs Temperature",
+        x = "Normalized Temperature",
+        y = "Bike Rentals")
`geom_smooth()` using formula = 'y ~ x'
> bike_data$day_type <- ifelse(bike_data$weekday %in% c(0,6), "Weekend", "Weekday")
> avg_rentals <- bike_data %>%
+   group_by(day_type) %>%
+   summarise(mean_rentals = mean(cnt))
> print(avg_rentals)
# A tibble: 2 × 2
  day_type mean_rentals
  <chr>           <dbl>
1 Weekday         4551.
2 Weekend         4390.
> ggplot(bike_data, aes(x = day_type, y = cnt, fill = day_type)) +
+   geom_boxplot() +
+   labs(title = "Bike Rentals: Weekday vs Weekend",
+        x = "Day Type",
+        y = "Rental Count")
> t_test_result <- t.test(cnt ~ day_type, data = bike_data)
> print(t_test_result)

        Welch Two Sample t-test

data:  cnt by day_type
t = 0.98373, df = 361.39, p-value = 0.3259
alternative hypothesis: true difference in means between group Weekday and group Weekend is not equal to 0
95 percent confidence interval:
 -160.7304  482.4914
sample estimates:
mean in group Weekday mean in group Weekend
             4550.566              4389.686
```

According to the t-test results shown in the R output:

- t = 0.9837, df = 361.39, p-value = 0.3259

- 95% confidence interval for the difference: [-160.7304, 482.4914]
- Mean in weekday group: 4550.566; Mean in weekend group: 4389.686

The p-value of 0.3259 suggests that the difference between weekday and weekend rental averages is not statistically significant at the conventional alpha level of 0.05. This indicates that while there is a numerical difference of approximately 161 more rentals on weekdays, this difference could be due to random variation rather than representing a true population difference.

# Discussion

**Weather Factors:**

The strong positive coefficient for temperature (5652.75) confirms that it is the most influential weather-related predictor of bike rentals. For each unit increase in normalized temperature, we can expect an increase of approximately 5,653 bike rentals, highlighting temperature as the dominant factor influencing ridership.

The substantial negative coefficients for humidity (-2359.91) and wind speed (-3358.93) demonstrate that uncomfortable weather conditions significantly deter riders. Poor weather situations (weathersit coefficient = -460.85) also show significant negative impact on rental demand.

**Temporal Patterns:**

While the numerical difference between weekday and weekend rentals (161 bikes, or about 3.7% higher on weekdays) suggests a potential trend toward commuting usage, the lack of statistical significance (p = 0.3259) means we cannot conclusively state that weekday usage differs from weekend usage in the population.

The significant positive coefficient for season (409.64) confirms that seasonal variation is an important factor in bike rental patterns, with higher rentals expected during warmer seasons.

# Conclusion and Recommendations

This analysis provides several important insights about bike-sharing usage patterns:

1. Weather conditions strongly influence daily bike rental demand, with temperature being the most powerful predictor, Humidity and wind speed have significant negative effects on ridership

2. Seasonal variations are substantial and statistically significant

3. The numerical difference between weekday and weekend rentals was not statistically significant in this dataset

Recommendations:

1. Develop targeted promotions during periods of expected low demand.
2. Consider covered bike options for harsh weather
3. Strengthen bike-transit connections to maintain ridership during adverse weather

# Video Walkthrough:

Recording-20250418_201051.webm

# References:

1. Fanaee-T, Hadi, and Gama, João. (2013). *Event labeling combining ensemble detectors and background knowledge*. UCI Machine Learning Repository - Bike Sharing Dataset
2. Napkin.ai. (2025). *Data visualization/image generated using AI tool*. https://www.napkin.ai