How Weather Affects Urban Bike-Sharing

Discovering the relationship between weather conditions and bike-sharing trends in the city.

2023-11-21

Team Members:

Sayantika Saha - T00731231

Shivani Tyagi - T00727866

Ahmad Khawaja -T00733331

---

1 TABLE OF CONTENT

1. Project Objective
2. About Dataset
3. Data Preprocessing
4. EDA
5. Conclusion
6. References

---

2 PROJECT OBJECTIVE

This study utilizes a two-year dataset from the Capital Bikeshare system in Washington D.C. for the years 2011 and 2012. The data includes time variables such as weather conditions, hour of day, etc. with the goal of investigating bike-sharing rental trends. Our hypothesis is that with worsening weather conditions like rainfall or snow, the number of daily active users on the bike sharing platform decreases significantly. Furthermore, our analysis will incorporate a range of statistical techniques to explore the relationships between different variables and bike rental patterns, enabling us to discover hidden trends.

3 ABOUT DATASET

Bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions, precipitation, day of week, season, hour of the day, etc. can affect the rental behaviors. The core data set is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which is publicly available at UCI machine learning repository. We have aggregated data on two categories, one is on hourly and other on daily basis and added according to the corresponding weather and seasonal information.

We are using **day.csv - bike sharing counts aggregated on daily basis.**

4 DATA PRE-PROCESSING

4.0.1 Required Package Imports

```
library(readr)   # For reading the data
library(dplyr)   # For data manipulation
library(knitr)
library(ggplot2)
library(dplyr)
library(gridExtra) # For arranging multiple plots
library(car)
```

4.1 Exploring Data-Sets

STEP 1 : Reading data files from local directory

```
dataset <- read_csv("day.csv")
```

```
head(dataset)
```

```
## # A tibble: 6 x 16
##   instant dteday     season    yr  mnth holiday weekday workingday weathersit
##     <dbl> <date>      <dbl> <dbl> <dbl>   <dbl>   <dbl>      <dbl>      <dbl>
## 1       1 2011-01-01      1     0     1       0       6          0          2
## 2       2 2011-01-02      1     0     1       0       0          0          2
## 3       3 2011-01-03      1     0     1       0       1          1          1
## 4       4 2011-01-04      1     0     1       0       2          1          1
## 5       5 2011-01-05      1     0     1       0       3          1          1
## 6       6 2011-01-06      1     0     1       0       4          1          1
## # i 7 more variables: temp <dbl>, atemp <dbl>, hum <dbl>, windspeed <dbl>,
## #   casual <dbl>, registered <dbl>, cnt <dbl>
```

```
colnames(dataset)
```

```
##  [1] "instant"    "dteday"     "season"     "yr"         "mnth"
##  [6] "holiday"    "weekday"    "workingday" "weathersit" "temp"
## [11] "atemp"      "hum"        "windspeed"  "casual"     "registered"
## [16] "cnt"
```

```
dim(dataset)
```

```
## [1] 731  16
```

```
sum(is.na(dataset))
```

```
## [1] 0
```

The study in question delves into a comprehensive dataset from the Capital Bikeshare system in Washington D.C., covering the years 2011 and 2012. This dataset, consisting of 731 rows and 16 columns, is aimed at analyzing bike-sharing rental trends in the city. It includes a variety of data types, with one column in datetime format and others being a mix of categorical and continuous numerical types. Key variables encapsulate time-related aspects such as the date ('dteday'), season (coded 1 for winter to 4 for fall), year ('yr' with 0 for 2011 and 1 for 2012), month ('mnth' from 1 to 12), and hour of the day ('hr' from 0 to 23).

Additionally, the dataset incorporates variables that describe the day's status and environmental conditions. These include 'holiday' (indicating whether the day is a holiday, based on, 'weekday' (day of the week), and 'workingday' (1 if the day is neither a weekend nor a holiday, otherwise 0). The 'weathersit' variable categorizes weather conditions, ranging from clear skies to heavy rain and snow. It also features normalized values for temperature ('temp'), feeling temperature ('atemp'), humidity ('hum'), and wind speed ('windspeed'). The dataset concludes with counts of casual ('casual'), registered ('registered'), and total ('cnt') bike rentals, providing a thorough insight into the patterns of bike-sharing use under various conditions. The dataset was pre-cleaned and contains no NULL values.

4.2 Reversing normalization

Normalization is performed using the formula:

$$\text{Normalized Value} = \frac{\text{Value} - \min}{\max - \min}$$

Where: - Value is the actual value. - min is the minimum value in the range. - max is the maximum value in the range.

To reverse the normalization and retrieve the actual values, we use the formula:

$$\text{Value} = (\text{Normalized Value} \times (\max - \min)) + \min$$

Where: - Normalized Value is the value after normalization. - min and max are as defined above.

The details of the normalization is mentioned in the dataset description and the abstract

```
# Denormalize temperature (temp)
# temp_min = -8, temp_max = 39
dataset$temp_actual = (dataset$temp * (39 - (-8))) + (-8)
# Denormalize feeling temperature (atemp)
# atemp_min = -16, atemp_max = 50
dataset$atemp_actual = (dataset$atemp * (50 - (-16))) + (-16)
# Denormalize humidity (hum)
# Since it's divided by 100, we just multiply by 100
dataset$hum_actual = dataset$hum * 100
# Denormalize wind speed (windspeed)
# windspeed_max = 67
dataset$windspeed_actual = dataset$windspeed * 67
# View the first few rows of the dataset to confirm the changes
head(dataset)
```

```
## # A tibble: 6 x 20
##   instant dteday     season    yr  mnth holiday weekday workingday weathersit
##     <dbl> <date>      <dbl> <dbl> <dbl>   <dbl>   <dbl>      <dbl>      <dbl>
## 1       1 2011-01-01      1     0     1       0       6          0          2
## 2       2 2011-01-02      1     0     1       0       0          0          2
```

```
## 3       3 2011-01-03      1     0     1       0       1          1          1
## 4       4 2011-01-04      1     0     1       0       2          1          1
## 5       5 2011-01-05      1     0     1       0       3          1          1
## 6       6 2011-01-06      1     0     1       0       4          1          1
## # i 11 more variables: temp <dbl>, atemp <dbl>, hum <dbl>, windspeed <dbl>,
## #   casual <dbl>, registered <dbl>, cnt <dbl>, temp_actual <dbl>,
## #   atemp_actual <dbl>, hum_actual <dbl>, windspeed_actual <dbl>
```

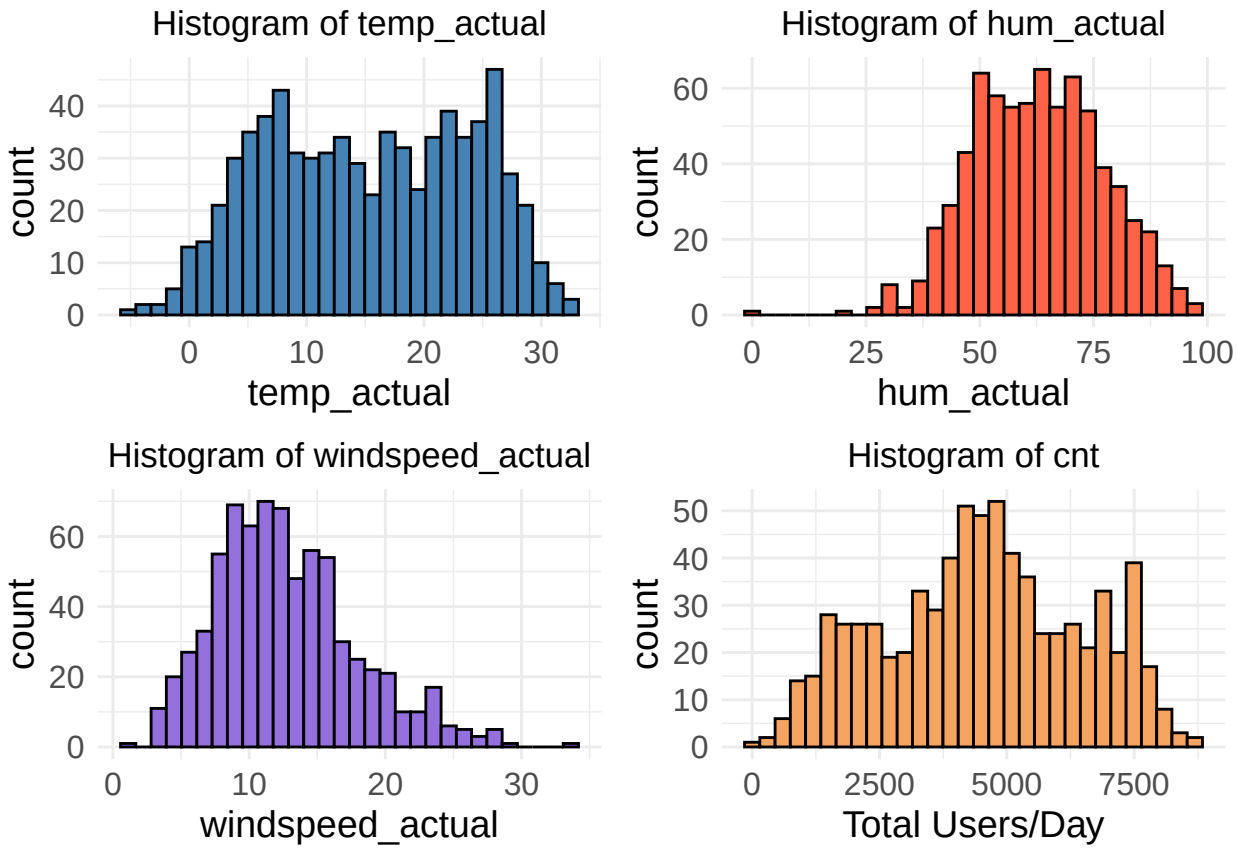
5 EDA

5.0.1 Summary of statistical features

```
##   temp_actual      atemp_actual       hum_actual    windspeed_actual
##  Min.   :-5.221   Min.   :-10.781   Min.   : 0.00   Min.   : 1.500
##  1st Qu.: 7.843   1st Qu.:  6.298   1st Qu.:52.00   1st Qu.: 9.042
##  Median :15.422   Median : 16.124   Median :62.67   Median :12.125
##  Mean   :15.283   Mean   : 15.307   Mean   :62.79   Mean   :12.763
##  3rd Qu.:22.805   3rd Qu.: 24.168   3rd Qu.:73.02   3rd Qu.:15.625
##  Max.   :32.498   Max.   : 39.499   Max.   :97.25   Max.   :34.000
##       cnt
##  Min.   :  22
##  1st Qu.:3152
##  Median :4548
##  Mean   :4504
##  3rd Qu.:5956
##  Max.   :8714
```

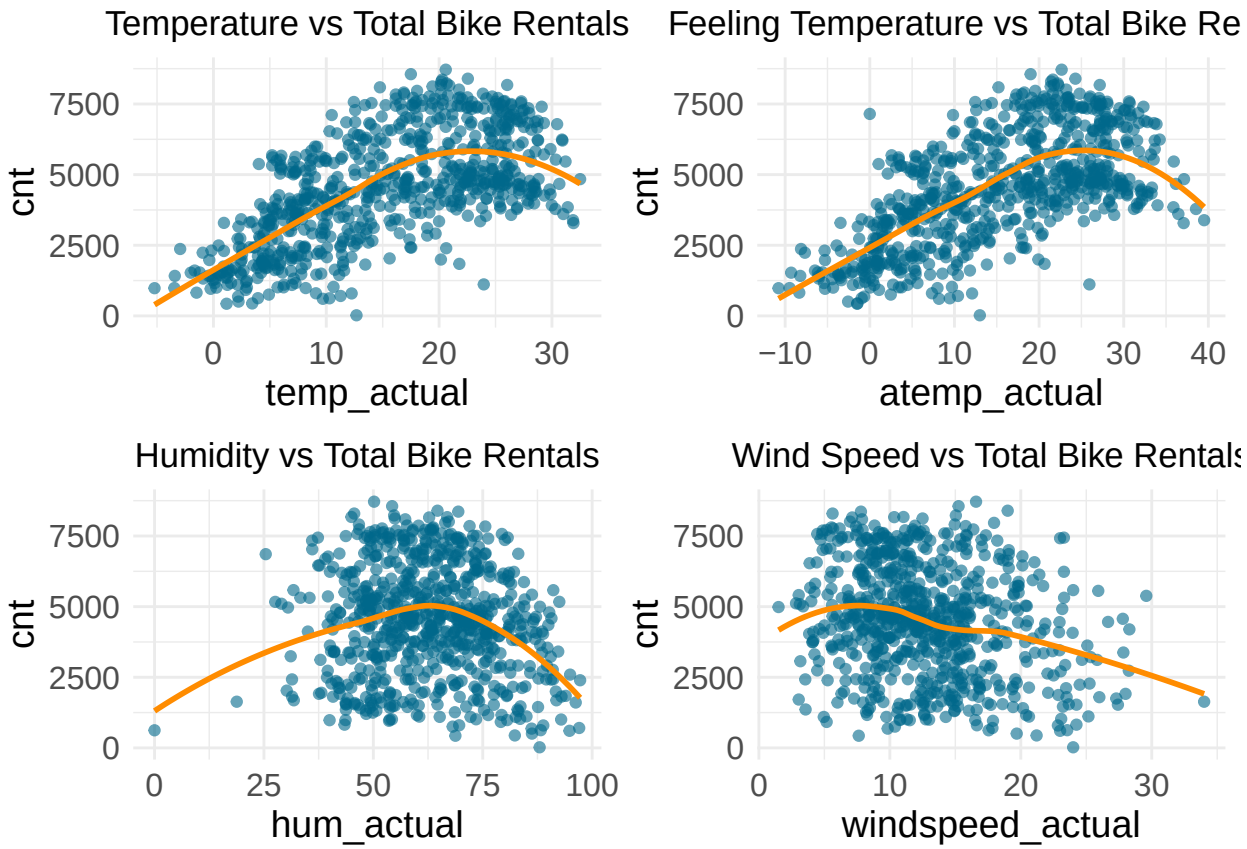
- **Temperature (°C)**: Mean = 15.28, Std Dev = 8.60, Min = -5.22, Max = 32.50
- **Feeling Temperature (°C)**: Mean = 15.31, Std Dev = 10.76, Min = -10.78, Max = 39.50
- **Humidity (%)**: Mean = 62.79, Std Dev = 14.24, Min = 0.00, Max = 97.25
- **Wind Speed**: Mean = 12.76, Std Dev = 5.19, Min = 1.50, Max = 34.00
- **Total Bike Rentals**: Mean = 4504.35, Std Dev = 1937.21, Min = 22, Max = 8714

5.0.2 Trends through graphical Illustrations:

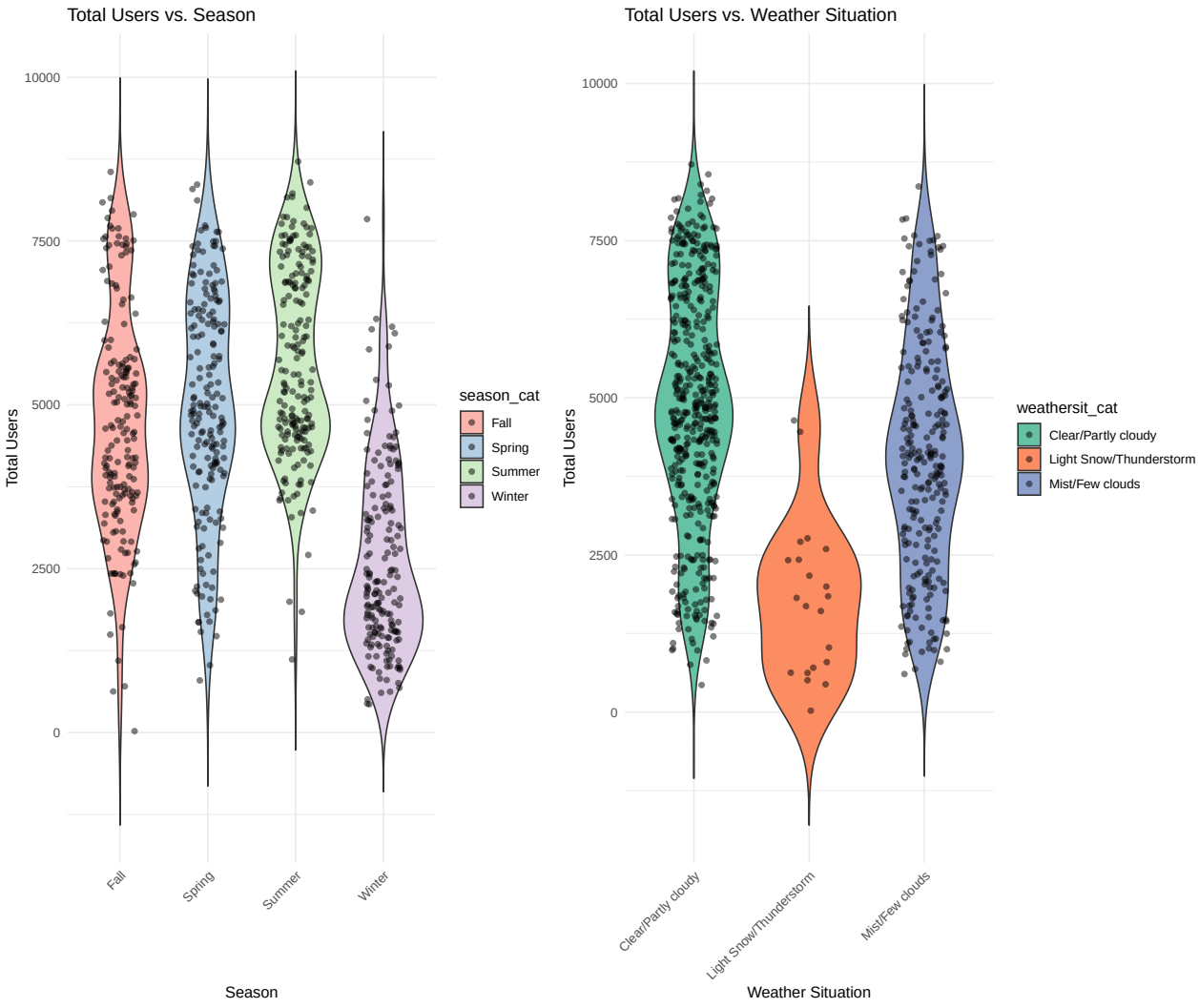


Distributions:

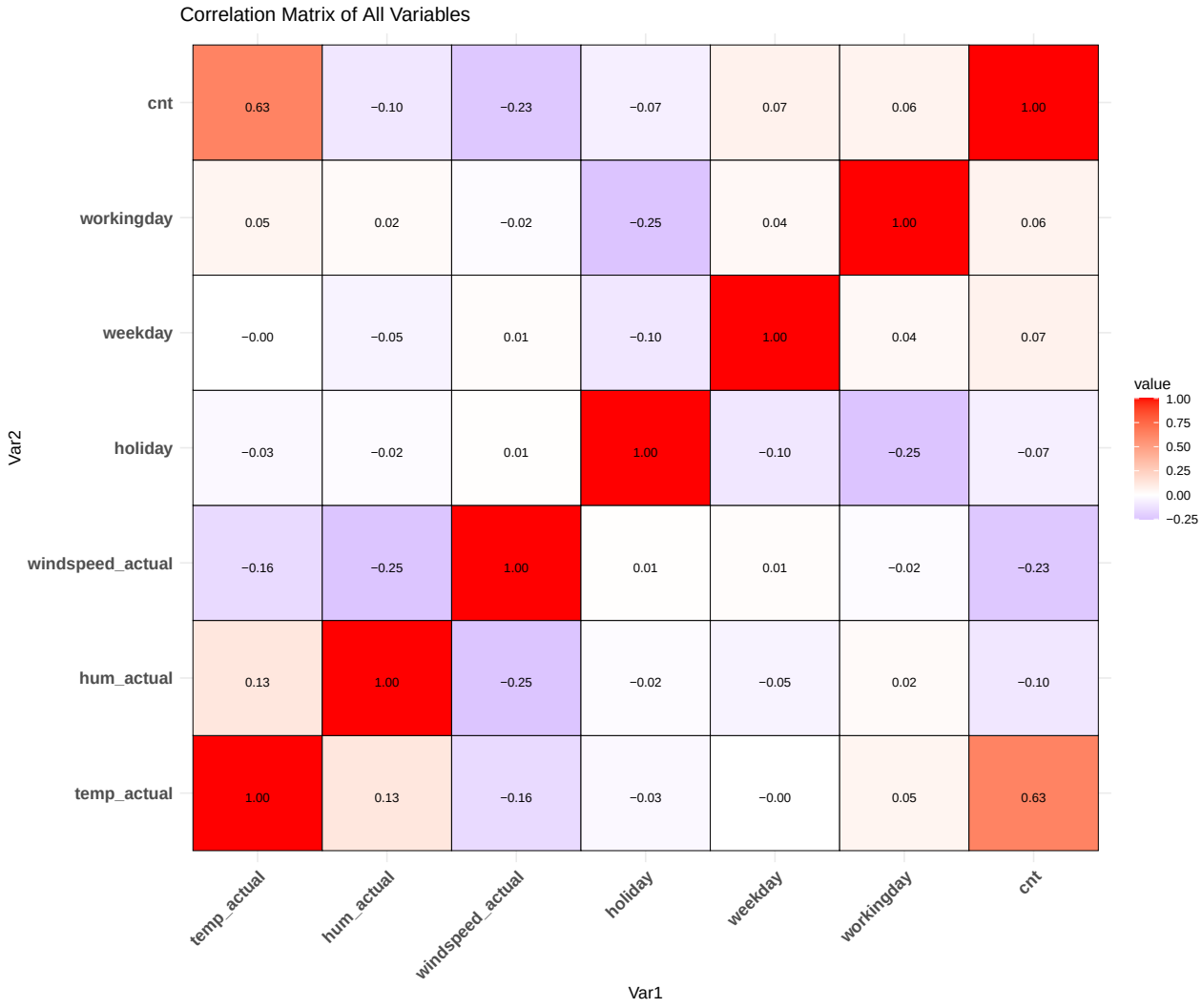
- **Temperature (temp_actual)**: The histogram of actual temperature shows a bimodal distribution with two peaks, which suggests that there are two different ranges of temperatures at which bike rentals occur more frequently. The most common temperature ranges for bike rentals are around 10-15 degrees Celsius and 25-30 degrees Celsius.
- **Humidity (hum_actual)**: The humidity histogram displays a distribution that is approximately normal, with a slight skew to the right. The most frequent humidity levels are clustered around 50-75%, indicating that these conditions are most common in the dataset. There are very few counts of extremely low or high humidity levels.
- **Wind Speed (windspeed_actual)**: The wind speed's histogram suggests a right-skewed distribution, with most of the bike rentals occurring at lower wind speeds. There is a peak at around 5-15, which implies that lower wind speeds are more common and possibly more conducive to bike rentals.
- **Total Users/Day (cnt)**: The histogram of the total daily users shows a distribution that is neither normal nor uniform. It appears to be right-skewed with a peak at around 2500-3000 users per day. This indicates that on most days, the bike rental counts are on the lower end of the scale, with fewer days experiencing very high rental counts.



- **Temperature vs. Total Bike Rentals (temp_actual vs. cnt)**: The scatter plot suggests a positive correlation between temperature and total bike rentals, with a non-linear relationship. As the temperature increases, the number of bike rentals also increases, peaking at moderate temperatures, and then slightly declines as temperatures become very high.
- **Feeling Temperature vs. Total Bike Rentals (atemp_actual vs. cnt)**: Similar to the actual temperature, the feeling temperature also shows a positive, non-linear correlation with bike rentals. This indicates that rentals are more frequent at comfortable perceived temperatures, with a peak at a certain point before decreasing slightly as the feeling temperature rises further.
- **Humidity vs. Total Bike Rentals (hum_actual vs. cnt)**: The relationship between humidity and bike rentals appears to be somewhat inverse and non-linear. Initially, as humidity increases, bike rentals increase but only up to a certain point. After this point, the number of rentals tends to decrease as humidity levels rise, suggesting that extremely high humidity might deter bike rentals.
- **Wind Speed vs. Total Bike Rentals (windspeed_actual vs. cnt)**: The scatter plot for wind speed against bike rentals shows a less clear trend compared to temperature and humidity. There appears to be a slight increase in rentals with an increase in wind speed up to a certain level, after which the trend flattens out. This suggests that while very low wind speeds do not encourage bike rentals, moderate wind speeds might be optimal, with higher wind speeds not having a significant impact on rental numbers.



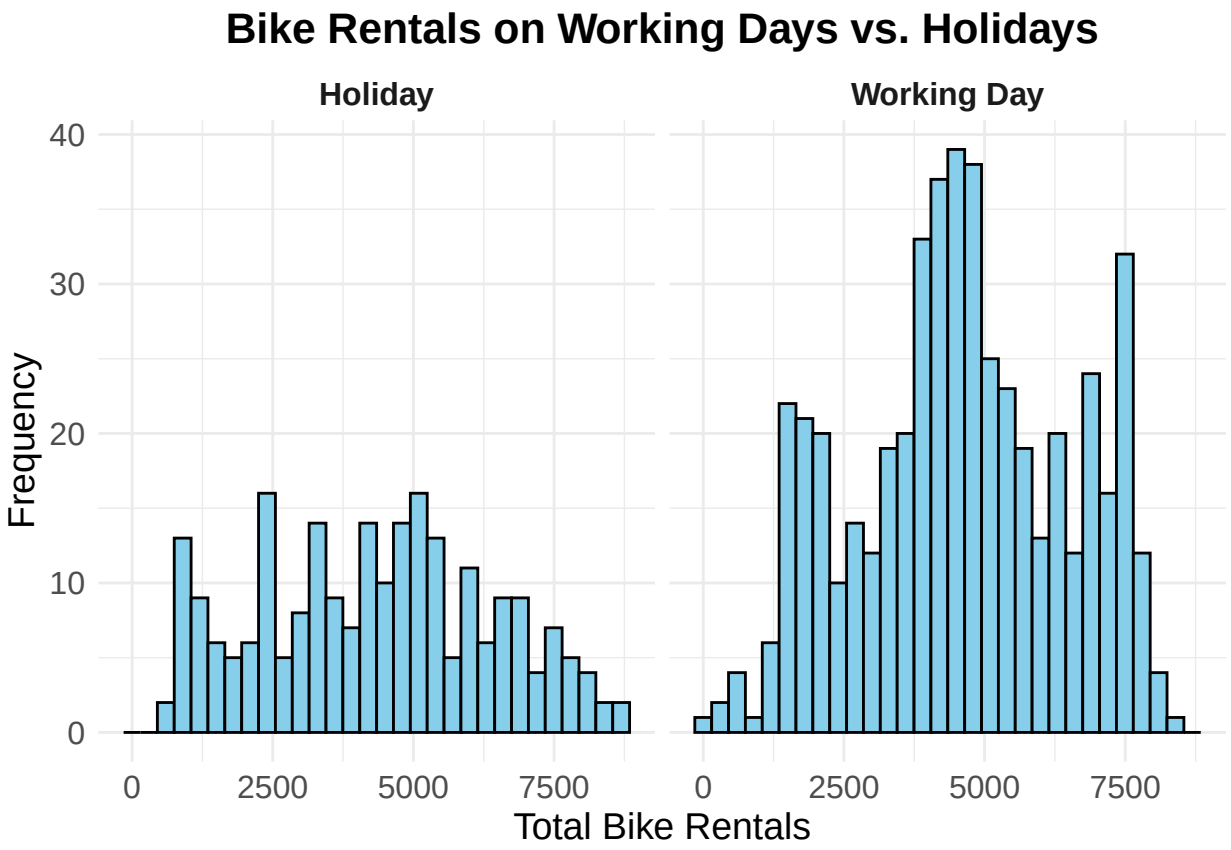
- **Total Users vs. Season**: The violin plot for seasons shows that bike rental counts vary significantly with the seasons. The widest parts of the violins for spring, summer, and fall indicate that there are a higher number of days with medium to high rental counts in these seasons, with summer having the broadest distribution, suggesting it's the most popular season for bike rentals. Conversely, the winter season shows a narrower distribution, implying lower overall rental counts, which could be due to the colder weather making biking less appealing.
- **Total Users vs. Weather Situation**: The violin plot for weather situations shows that clear or partly cloudy days have a wider distribution and higher median rental counts, suggesting that bike rentals are more popular during good weather conditions. Days with light snow or thunderstorms have the lowest median rental count and a narrower distribution, indicating fewer rentals. The distribution for misty or few clouds conditions is intermediate between the clear and adverse weather conditions, with a moderate median rental count.



The correlation matrix provided shows the relationship between various variables and the target variable "**cnt**", which represents the total bike rentals.

- The variable "**temp_actual**" has a strong positive correlation with "cnt" (0.63), suggesting that higher temperatures are associated with an increased number of bike rentals.
- "**hum_actual**" (humidity) has a slight positive correlation (0.13) with "cnt", indicating a very weak linear relationship between humidity and the total number of bike rentals.
- "**windspeed_actual**" shows a weak negative correlation with "cnt" (-0.16), which could mean that higher wind speeds slightly discourage bike rentals.
- "**workingday**" seems to have a very weak correlation with "cnt" (0.05), implying that whether a day is a working day or not does not significantly affect bike rentals.
- "weekday" and "holiday" have negligible correlations with "cnt" (0.00 and -0.03, respectively), indicating no linear relationship with the number of bike rentals.

It's important to note that while "temp_actual" shows a notable positive correlation, the other variables do not seem to have a strong linear relationship with the target variable "cnt" on their own



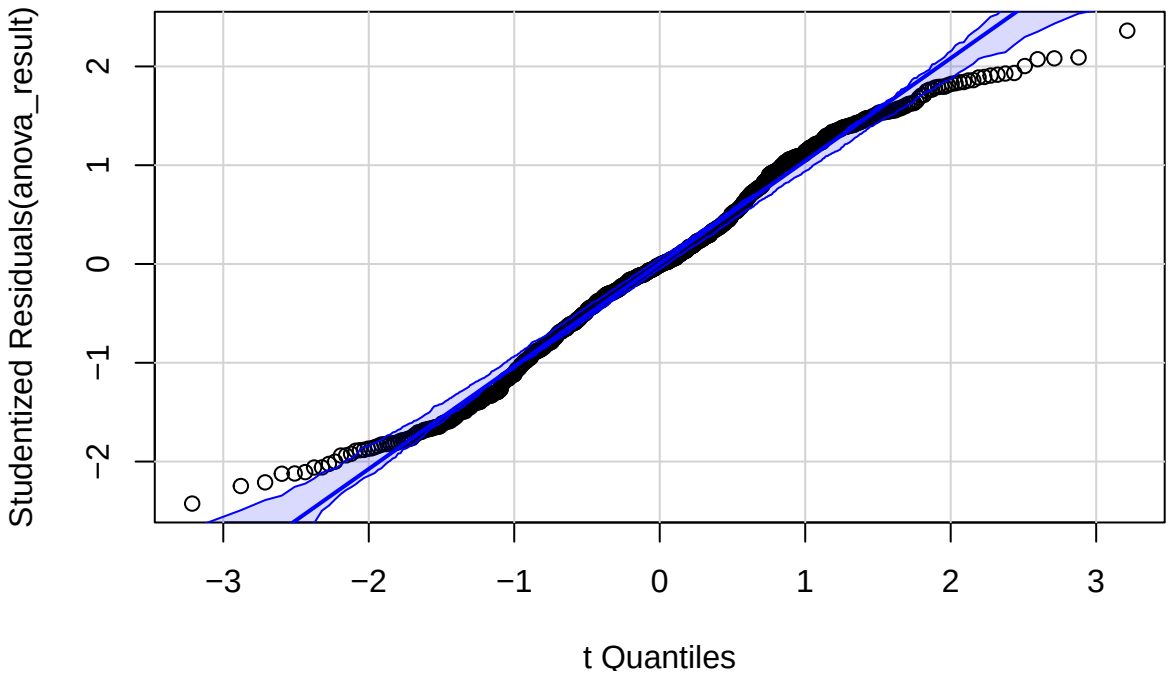
The histogram compares the frequency of total bike rentals on holidays versus working days. For holidays, the distribution of bike rentals shows that the highest frequencies are in the lower range of total rentals, suggesting that there are fewer bike rentals on holidays overall. In contrast, the histogram for working days shows a higher frequency of days with a moderate number of bike rentals, with the most common values being around the middle of the range. There's also a broader distribution of rental frequencies on working days, indicating more variation in the number of rentals. This could imply that bike usage is more consistent and possibly higher on working days compared to holidays, which may be due to commuters using the bike share system to travel to work.

6 Hypothesis Testing

6.0.1 ANOVA test

Linearity Test for ANOVA test

```
anova_result <- aov(cnt ~ factor(weathersit), data=dataset)
qqPlot(anova_result, id=FALSE)
```




```
leveneTest(cnt ~ factor(weathersit), data=dataset)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   2  2.9675 0.05205 .
##       728
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.0.2 Analysis and Interpretation:

6.0.2.1 1. Boxplot of Bike Rentals for Different Weather Conditions The boxplot visualizes the distribution of daily bike rentals (cnt) across different weather conditions. It can be observed that the median and spread of rentals vary with the weather, suggesting a potential impact of weather on bike rentals.

6.0.2.2 2. QQ Plot of Residuals The QQ Plot (Quantile-Quantile Plot) of the residuals from the ANOVA model is used to check the normality of residuals, which is an assumption of ANOVA. The plot shows how closely the residuals follow a normal distribution. In this case, the residuals mostly follow the line, indicating a reasonable approximation to normality, though there are some deviations at the tails.

6.0.2.3 3. Levene's Test for Equality of Variances Levene's Test result:

- Statistic: 2.9675
- P-value: 0.05205

Levene's Test is used to assess the equality of variances for a variable calculated for two or more groups. In this case, it tests whether the variance in daily bike rentals is the same across different weather conditions.

6.0.2.4 Interpretation:

- The P-value is marginally above 0.05, suggesting a borderline result regarding the equality of variances assumption.
- While it's not a clear violation, this borderline result warrants cautious interpretation of the ANOVA results, as ANOVA assumes equal variances across groups.

```
# ANOVA test
anova_result <- aov(cnt ~ factor(weathersit), data=dataset)
summary(anova_result)
```

```
##                      Df    Sum Sq   Mean Sq F value Pr(>F)
## factor(weathersit)    2 2.716e+08 135822286   40.07 <2e-16 ***
## Residuals           728 2.468e+09   3389960
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Sum of Squares (sum_sq)**: The between-groups sum of squares (for weathersit) is approximately 2.716×10^8, and the within-groups sum of squares (Residual) is about 2.468×10^9
- **Degrees of Freedom (df)**: There are 2 degrees of freedom for weathersit (since it has 3 categories: 1, 2, and 3) and 728 degrees of freedom for the residuals.
- **F-statistic (F)**: The calculated F-statistic value is approximately 40.07.
- **P-value (PR(>F))**: The P-value is extremely low (3.106×10^−17).

6.0.3 Interpretation:

- The low P-value (much less than 0.05) suggests that there are statistically significant differences in the average number of bike rentals (cnt) among different weather situations (clear/few clouds, mist/cloudy, light snow/rain).
- This result supports the hypothesis that weather conditions have a significant impact on the number of daily active users on the bike-sharing platform. Specifically, it indicates that the number of users varies significantly with different weather conditions.

6.0.4 T test between Group 1 (weather situations 1 and 2) and Group 2 (weather situation 3) clear weather vs. rainy/snowy

```
# Grouping weather situations
group_1_2 <- subset(dataset, weathersit %in% c(1, 2))$cnt
group_3 <- subset(dataset, weathersit == 3)$cnt
# Conducting the t-test
# Assuming unequal variances (Welch Two Sample t-test)
t_test_result <- t.test(group_1_2, group_3, var.equal = FALSE)
# Displaying the results
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  group_1_2 and group_3
## t = 9.937, df = 22.86, p-value = 9.166e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2201.828 3360.080
## sample estimates:
## mean of x mean of y
##  4584.239  1803.286
```

The t-test comparing the number of bike rentals (cnt) between two groups—Group 1 (weather situations 1 and 2) and Group 2 (weather situation 3)—yields the following results:

- **T-statistic**: 9.937
- **P-value**: $9.166 \times 10\hat{}-10$

6.0.5 Interpretation:

- The T-statistic is significantly high, and the P-value is extremely low (much less than 0.05), indicating a statistically significant difference in the number of bike rentals between the two groups.
- This result supports the hypothesis that the number of daily active users on the bike-sharing platform decreases significantly with worsening weather conditions. Specifically, it shows that the number of rentals is significantly lower in weather situation 3 (light snow, light rain + thunderstorm + scattered clouds, light rain + scattered clouds) compared to weather situations 1 and 2 (clear, few clouds, partly cloudy, misty conditions).

6.0.6

6.1 Good weather vs Bad weather

Group 1: Good weather (season 2 and 3 - spring and summer)

Group 2: Bad weather (season 1 - winter

```
# Subset data for good weather (seasons 2 and 3)
good_weather <- subset(dataset, season %in% c(2, 3))$cnt
# Subset data for bad weather (season 1)
bad_weather <- subset(dataset, season == 1)$cnt
# Perform the t-test
t_test_result <- t.test(good_weather, bad_weather, var.equal = FALSE)
# Print the result
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  good_weather and bad_weather
## t = 20.361, df = 405.38, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2455.30 2980.08
## sample estimates:
## mean of x mean of y
##  5321.823  2604.133
```

The t-test comparing the number of bike rentals (cnt) between two groups—Group 1 (good weather: spring and summer) and Group 2 (bad weather: winter)—yields the following results:

- **T-statistic**: 20.361
- **P-value**: 2.2e-16

6.1.1 Interpretation:

- The T-statistic is notably high, and the P-value is extremely low (far less than 0.05), indicating a statistically significant difference in the number of bike rentals between the two groups.
- This result strongly supports the hypothesis that the number of daily active users on the bike-sharing platform is significantly higher in good weather conditions (spring and summer) compared to bad weather conditions (winter).

This analysis reaffirms that seasonal weather variations have a substantial impact on bike-sharing usage patterns, with more favorable weather conditions (like in spring and summer) leading to increased bike rentals.

6.2 Regression

```
library(dplyr)
# Dropping unnecessary columns and normalized columns
bike_data_for_regression <- dataset %>%
  select(-c(registered, casual, instant, dteday, temp, atemp,yr,mnth, hum, windspeed,weathersit,weathers
# Splitting the data into X (predictors) and y (response)
X <- bike_data_for_regression %>% select(-cnt)
y <- bike_data_for_regression$cnt
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
set.seed(42)  # For reproducibility
trainIndex <- createDataPartition(y, p = 0.8, list = FALSE)
X_train <- X[trainIndex, ]
X_test <- X[-trainIndex, ]
y_train <- y[trainIndex]
y_test <- y[-trainIndex]
```

```r
# Fitting the model
regression_model <- lm(y_train ~ ., data = X_train)
# Summary of the model
summary(regression_model)
```

```
##
## Call:
## lm(formula = y_train ~ ., data = X_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4828.2 -1056.3   -42.4  1036.9  3651.9
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4922.888    383.678  12.831  < 2e-16 ***
## holiday          -332.068    345.982  -0.960    0.338
## weekday            28.161     29.421   0.957    0.339
## workingday        124.108    130.554   0.951    0.342
## temp_actual       141.679      6.938  20.421  < 2e-16 ***
## hum_actual        -30.473      4.301  -7.085 4.04e-12 ***
## windspeed_actual  -66.449     11.726  -5.667 2.30e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1421 on 580 degrees of freedom
## Multiple R-squared:  0.4639, Adjusted R-squared:  0.4583
## F-statistic: 83.65 on 6 and 580 DF,  p-value: < 2.2e-16
```

```r
# Getting coefficients
feature_weights <- coef(regression_model)
# Displaying the coefficients
feature_weights
```

```
##      (Intercept)          holiday          weekday       workingday
##       4922.88795       -332.06812         28.16136        124.10755
##      temp_actual       hum_actual windspeed_actual
##        141.67899        -30.47337        -66.44892
```

6.2.1 Regression Analysis Summary

- **Residuals**: The spread of residuals, ranging from -4828.2 to 3651.9, indicates the differences between observed and predicted values of bike rentals. The median near -42.4 suggests a slight bias in underestimation.
- **Coefficients**:
  - **(Intercept) 4922.89**: This is the expected value of cnt when all other predictors are 0. A high t-value and a very low p-value (< 2e-16) indicate it's highly significant.
  - **holiday -332.07**: Suggests bike rentals decrease by 332 units on holidays, but it's not statistically significant (p-value 0.338).

- **weekday 28.16**: Indicates a slight increase in bike rentals depending on the day of the week, but not significant (p-value 0.339).
- **workingday 124.11**: Suggests an increase in bike rentals on working days, but again, not statistically significant (p-value 0.342).
- **temp_actual 141.68**: Shows a significant positive relationship between temperature and bike rentals, with an increase in temperature leading to more rentals.
- **hum_actual -30.47**: Indicates that higher humidity is associated with a decrease in bike rentals, and it's statistically significant.
- **windspeed_actual -66.45**: Shows that higher wind speeds are associated with fewer bike rentals, and it's also significant.

6.2.2 Model Fit

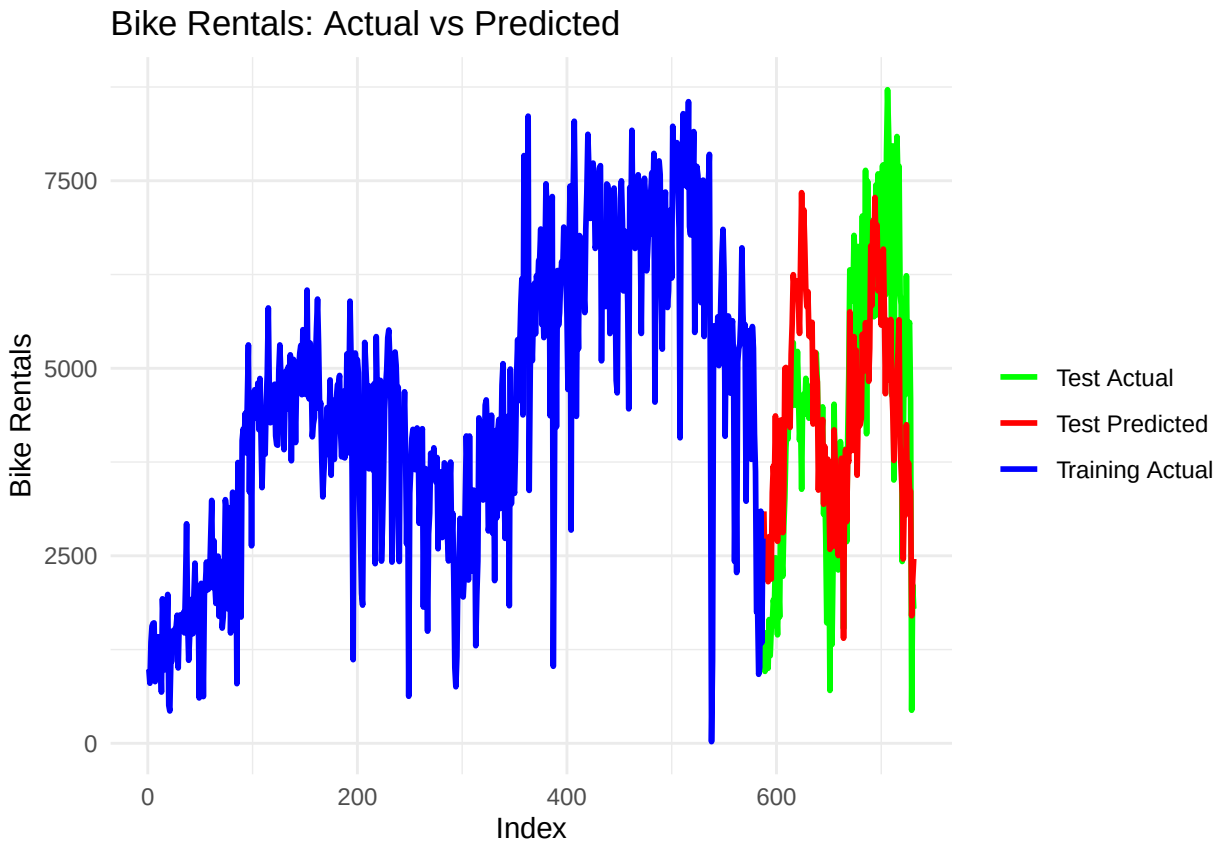
- **Residual Standard Error**: 1421 on 580 degrees of freedom. This value measures the typical size of the residuals.
- **Multiple R-squared**: 0.4639. About 46.39% of the variance in bike rental counts is explained by the model, which is a moderate fit.
- **Adjusted R-squared**: 0.4583. This is a slight adjustment to the R-squared value, accounting for the number of predictors.
- **F-statistic**: 83.65 on 6 and 580 DF, with a p-value $<$ 2.2e-16. This suggests the model as a whole is statistically significant.

6.2.3 Interpretation

- The model suggests that temperature has the most significant positive impact on bike rentals, followed by negative impacts from humidity and wind speed.
- Variables like holidays, weekdays, and working days are not statistically significant in predicting bike rentals in this model.
- The model's moderate R-squared value implies there is room for improvement, possibly by including other relevant variables or interactions not considered in this model.
- The significance of temperature, humidity, and wind speed aligns with intuitive expectations about outdoor activities like bike-sharing.

```
library(ggplot2)

# Creating index vectors for train and test sets
index_train <- 1:length(y_train)
index_test <- (length(y_train) + 1):(length(y_train) + length(y_test))
y_pred_test <- predict(regression_model, newdata = X_test)

# Creating a data frame for the training set (actual values)
data_train <- data.frame(Index = index_train,Value = y_train,Type = "Training Actual"
)

# Creating a data frame for the test set (actual values)
data_test_actual <- data.frame(Index = index_test,Value = y_test,Type = "Test Actual"
)
```

```r
# Creating a data frame for the test set (predicted values)
data_test_pred <- data.frame(Index = index_test,Value = y_pred_test,Type = "Test Predicted"
)

# Combining all data into one data frame
all_data <- rbind(data_train, data_test_actual, data_test_pred)

# Plotting
ggplot(all_data, aes(x = Index, y = Value, color = Type)) +
  geom_line(size = 1) +
  scale_color_manual(values = c("Training Actual" = "blue", "Test Actual" = "green", "Test Predicted" =
  ggtitle("Bike Rentals: Actual vs Predicted") +
  ylab("Bike Rentals") +
  theme_minimal() +
  theme(legend.title = element_blank())
```



---

7 CONCLUSION

The extensive analysis performed on migration trends usage data in Capital Bikeshare system across different weather conditions, offers important insights for examining whether with worsening weather conditions like rainfall or snow, the number of daily active users on the bike sharing platform decreases significantly.

1. Weather

- Temperature: Significant positive effect
- Wind: Significant negative effect.
- Humidity: Weak negative effect.

2. Climate : Clear and light cloudy days had much higher users as compared to rainy/snowy days.
3. Season : Winters had a lower number of users. Fairly consist user distribution in summer, spring and fall with summers having a slightly higher average

In conclusion, these findings imply that users prefer moderate temperatures, lower wind speeds, and clear weather conditions, such as those found in spring and summer, which tend to attract higher engagement. Conversely, higher wind speeds and extreme weather conditions, like those experienced in winter or during rainy/snowy days, negatively impact user activity.

---

8 REFERENCES

1. Dataset Sources

1.1 Fanaee-T,Hadi. (2013). Bike Sharing Dataset. UCI Machine Learning Repository. https://doi.org/10.1

2. Code References

2.1 ADSC1000_01 - Statistical Data Analysis/Lectures Slides

2.2 Linear Regression in R - https://www.codecademy.com/learn/learn-linear-regression-in-r/modules/lin

2.3 Hypothesis Testing : https://www.r-bloggers.com/2022/12/hypothesis-testing-in-r/
