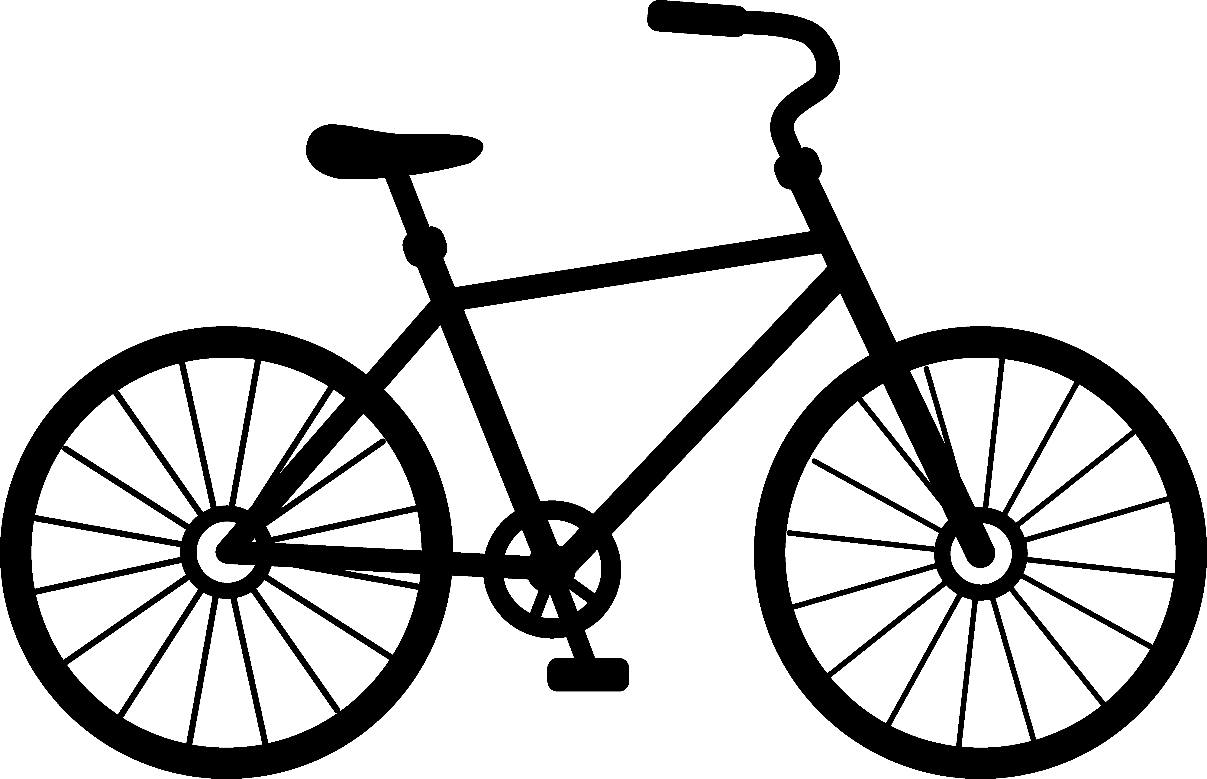
**Analysis of Casual vs. Registered Users of a Bike Share System in Washington DC**

April 12, 2023

UBC STAT 306

A report by Lab Group 2B1

|  |  |
| --- | --- |
| **Name** | **Student Number** |
| Alexander Proskiw | 27194166 |
| Youjung Kim | 38762639 |
| Mengfei Chen | 79921177 |
| Timothy Gao | 62490230 |



# Introduction

## Motivation

Bike sharing is where customers can rent a bike and return it to different places across a city, often meant to reduce vehicle dependency and extend transit capability in an urban area. Bike share services have been growing in popularity around Vancouver, with companies such as Mobi Bike Share, HOPR Bike Share, and Lime Electric Bike Sharing establishing themselves in recent years. With several cyclists in our group, we are accustomed to cycling all days of the week and in a variety of weather conditions in order to commute to school and/or work. In contrast, bicycle sharing services are often advertised as a fun and convenient way to enjoy a weekend while the weather is nice. But is this stereotype unfounded? In this report, we aim to explore the factors that influence the number of bike share users.

## Data

The data being used for this project is [bike rental data](https://www.kaggle.com/datasets/prepinstaprime/bike-rental-data)[[1]](#footnote-1) found on Kaggle. The dataset contains information on bicycle rentals associated with bike-sharing systems in Washington, DC from 2011 to 2012. The Kaggle dataset doesn’t say where or how the data was obtained, although further research indicates it is likely from the Capital Bikeshare system.

## Variables

Response variables include:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| casual | The number of non-registered users renting bikes at that time |
| registered | The number of registered users renting bikes at that time |

Explanatory variables include:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| year | Categorical variable (2011, 2012) |
| time | Categorical variable representing 1-hour segments within a day (0, 1, 2, …, 23) |
| season | Categorical variable (January to March, April to June, July to September, October to December) |
| holiday | Categorical variable (not a holiday, holiday) |
| workingday | Categorical variable (weekend, weekday) |
| weather | Categorical variable (clear, mist/clouds, light rain/snow, heavy rain/snow) |
| atemp | Continuous variable representing the normalized feeling temperature (°C) |
| humidity | Continuous variable representing the humidity (normalized by 100) |
| windspeed | Continuous variable representing the wind speed (miles per hour normalized by 67) |

Several alterations were made to the raw dataset. Specifically:

* The raw dataset included a variable for datetime, representing a year-month-day-hour. We separated this into its components, and only kept year and hour as categorical variables. Month was dropped as the dataset already included a variable for season (representing a 3-month period). Day was dropped as the dataset already included a variable for weekday/weekends, which seems more applicable than whether the day is the 5th of the month vs the 17th of the month.
* The raw dataset included variables for both “temperature” and “apparent temperature”. These two variables were highly correlated (R2 = 0.985). As such, temperature was discarded in favour of apparent temperature, with the rationale being that the temperature a person “feels” would likely have a greater influence on their decision to bike.
* The raw dataset included variables for the casual, registered, and total number of users. The total number of users was not considered as our research questions focus on examining the differences between casual and registered users.

## Research Questions

One might expect that registered users use bike rentals more routinely, perhaps for commuting. In contrast, casual users might use bike rentals more regularly on weekends when the weather is nice.

We aim to use bike sharing data to investigate the difference in usage patterns between casual and registered users by exploring the following three research questions:

* Q1: Do casual users rent more bikes on weekends than weekdays?
* Q2: Do registered users rent more bikes at typical commuting times in the morning and evening?
* Q3: Are the number of casual and registered users that rent bikes influenced by weather related variables?

To answer these research questions, we will perform a combination of visualizations and linear regression modelling.

# Analysis

## Visualization

To gain an understanding of the influence of each explanatory variable on the response variables, plots were made depicting the number of casual/registered users for each explanatory variable, not accounting for any additions/interactions from other variables.

Figure 1 on the next page includes plots for the number of casual bike users, while Figure 2 on the following page includes plots for the number of registered bike users.

Note that the plots for year, holiday, and windspeed have been excluded for brevity’s sake, but can be found in the R code for this project. Despite their exclusion in the following figures, they are still summarized in Section 2.2.

Figure : Effect of Different Explanatory Variables on the Number of Casual Bike Users

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

Figure : Effect of Different Explanatory Variables on the Number of Registered Bike Users

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

## Visualization Summary

Common Among Both Casual and Registered Users:

* In general, there are more registered users than casual users.
* On average, there are more users in 2012 than 2011[[2]](#footnote-2).
* A day being a holiday does not appear to make a large difference on the number of users2.
* As weather gets worse, it appears the number of users decreases.
* As temperature increases, it appears the number of users increases.
* As humidity and wind speed increases, it appears the number of users decreases2.

Unique to Casual Users:

* The average number of casual users is lowest around 12am to 6am, and then gradually increases throughout the day, reaching a peak around 2pm to 6pm, and then gradually decreases.
* On average, the season with the most casual users is July to September, while the season with the least casual users is January to March.
* It appears the number of casual users is larger on a weekend day compared to on a weekday.

Unique to Registered Users:

* The average number of registered users is highest around 8am, and 5pm to 6pm.
* On average, January to March has the least number of registered users, while the other 3 seasons all have a similar number of registered users.
* It appears the number of registered users is larger on a weekday compared to on a weekend day.

## Linear Regression Modelling

Separate models for both the number of casual bike users and the number of registered bike users were fitted using varying numbers of parameters. The 9 different explanatory variables (year, time, season, holiday, weekday, weather, temperature, humidity, and windspeed) result in 35 possible different parameters due to the categorical nature of several of the variables. The regsubsets command was used to determine the best linear model for each number of parameters. Interaction terms, quadratic terms, and log transformations were not applied.

For each of the casual and registered users, 3 different models are presented. First is a “simple to understand” model which evaluate some of the most important parameters. Next is the full 35 parameter model, as it had almost the best adjusted R2 and Mallows' Cp value[[3]](#footnote-3). Finally, due to the nature of the number of bicycle riders being a “count”, a Poisson model was also fitted. These models were also compared based on residual standard error, BIC values, and training/testing error.

Note that for casual users, a 3-parameter model was considered as the “simple to understand” model. In contrast, when examining registered users, the best 3-parameter model included terms for 8am, 5pm, and 6pm. Rather than separating out these terms in isolation, the “simple to understand” model for registered users was constructed just with the time variable (a categorical variable with 24 possibilities).

### Models For Casual Users

A summary of the models for predicting the number of casual users is provided below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | Simple Model | Full Model | Poisson Model |
| **Number of Parameters** | 3 | 35 | 35 |
| **Adjusted R2** | 0.43 | 0.58 | - |
| **BIC** | 109,921 | 106,854 | 146,444 |
| **Testing Error** | 20,111,694 | 20,899,291 | 20,047,149 |
| **Base Case** | Weekend | 2011, January to March, 12am, not a holiday, weekend, clear weather | |
| **Parameters For** | Weekday, temperature, humidity | All | |
| **Parameter Significance** | All p values < 0.05 | p values > 0.05 for 1am, 2am, 5am, 6am, and heavy rain/snow | All p values < 0.05 |

The simple model has a lower adjusted R2, larger Mallows’ Cp value, as well as a larger residual standard error and BIC in comparison to the full model. Thus, the full model is preferred over the simple model. Interestingly the simple model has a lower testing error, indicating the full model may potentially be overfitting the data. The simple model is also helpful by indicating that that weekday, temperature, and humidity are the 3 most important variables for predicting the number of casual users.

Between the full model and the Poisson model, the Poisson model has a larger BIC. However, its parameters are all significant, it has a lower testing error, and its plots of predicted vs actual values, residuals, and Normal Q-Q plots appear better than those of the full model (see Figure 3 below). Furthermore, the full model can predict a negative number of users, which doesn’t make sense in the context of counting bicycle riders. For these reasons, we would pick the Poisson as the “best” model for predicting the number of casual users.

Figure : Casual Users - Fitted Values (left), Residuals (middle) and Normal Q-Q Plot (right)

|  |  |  |  |
| --- | --- | --- | --- |
| Full Model |  |  |  |
| Poisson Model |  |  |  |

### Models For Registered Users

A summary of the models for predicting the number of registered users is provided below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | Simple Model | Full Model | Poisson Model |
| **Number of Parameters** | 23 | 35 | 35 |
| **Adjusted R2** | 0.53 | 0.69 | - |
| **BIC** | 132,172 | 127,806 | 348,622 |
| **Testing Error** | 188,487,268 | 207,878,572 | 249,438,369 |
| **Base Case** | 12am | 2011, January to March, 12am, not a holiday, weekend, clear weather | |
| **Parameters For** | 1am to 11pm | All | |
| **Parameter Significance** | All p values < 0.05 | p values > 0.05 for holiday and windspeed | All p values < 0.05 |

The simple model has a lower adjusted R2, as well as a larger residual standard error and BIC in comparison to the full model. Thus, the full model is preferred over the simple model. Interestingly the simple model has a lower testing error, indicating the full model may potentially be overfitting the data. The simple model is also helpful by indicating that that time of day is the most important variable for predicting the number of registered users.

Between the full model and the Poisson Model, the Poisson Model has a much larger BIC and larger testing error. However, its parameters are all significant, and its plots of predicted vs actual values, residuals, and Normal Q-Q plots appear better than those of the full model (see Figure 4 below). Furthermore, the full model can predict a negative number of users, which doesn’t make sense in the context of counting bicycle riders. For these reasons, we would pick the Poisson as the “best” model for predicting the number of registered users.

Figure : Registered Users - Fitted Values (left), Residuals (middle) and Normal Q-Q Plot (right)

|  |  |  |  |
| --- | --- | --- | --- |
| Full Model |  |  |  |
| Poisson Model |  |  |  |

# Conclusion

## Key Findings

With respect to our research questions, the key findings are as follows:

Q1: Do casual users rent more bikes on weekends than weekdays?

* Based on visualizations, yes, it appears that casual users rent more bikes on weekends than weekdays.
* Based on modelling, yes. The full model for casual users includes an adjustment term for weekday. All other parameters being equal, there were approximately 36 fewer casual users per hour on a weekday than a weekend day.

Q2: Do registered users rent more bikes at typical commuting times in the morning and evening?

* Based on visualizations, yes, the peak times for registered users to rent bikes appears to be at 8am, and 5pm to 6pm.
* Based on modelling, yes. The full model for registered users includes adjustment terms for different hours of the day. In comparison to 12am with all other parameters being equal, the largest of these were 8am (approximately 300 more users per hour), 5pm (approximately 334 more users per hour), and 6pm (approximately 313 more users per hour).

Q3: Are the number of casual and registered users that rent bikes influenced by weather related variables?

* Based on visualizations, yes. Across both casual and registered users, it appears that more users rent bikes when the temperature is higher, and fewer users rent bikes when humidity and windspeed are higher. Worse weather conditions result in fewer users across both groups. However, there is some slight variation in season. Both groups prefer renting in the summer, but registered users appear more willing to rent bikes in the spring and fall compared to casual users.
* Based on modelling, yes. The full models for both casual and registered users include significant parameters for season, weather, temperature, humidity, and windspeed. However, in the simple model for casual users, two of the three parameters were weather related (temperature and humidity) while in contrast, the simple model for registered users was based solely on time of day. This indicates that perhaps weather is slightly more of an important consideration for casual users.

## Limitations and Future Improvements

Several limitations were identified throughout this report. Potential methods of improvement include:

* Explore additional models with interaction terms, quadratic terms, or logarithmic transformations. Note that this would make the model even more difficult to interpret.
* Try to find data for additional years. Two years of data is not the largest when commenting on seasonal effects. If 2011 and 2012 had unusually nice weather the findings might may be skewed.
* Achieve further refinement by using month instead of season (e.g., 12 months vs 4 seasons). This would come at the cost of further model complexity however.
* Achieve further refinement by using day of the week (e.g., Mon, Tues, etc.) rather than just weekday vs weekend. This would come at the cost of further model complexity however.

1. Dataset accessed March 15, 2023 from <https://www.kaggle.com/datasets/prepinstaprime/bike-rental-data> [↑](#footnote-ref-1)
2. Note that the plots for year, holiday, and windspeed were excluded from Figure 1 and figure 2 for brevity’s sake. Please see the R code if interested in viewing these plots. [↑](#footnote-ref-2)
3. A slightly more optimal model could have been chosen that had 32 parameters for casual users and 34 parameters for registered users. However, the difference in Adjusted R2 and Mallows’ Cp value was very small, and hence we decided to keep the full models. [↑](#footnote-ref-3)