Forecasting Bikesharing Usage for DC's Capital Bikshare System

Table of Contents

Introduction

Bikesharing systems are an increasingly popular solution in major urban areas to increase trips taken by bike, which can help people get around without cars, thus improving the lives of both users, as well as non-users, as each bike trip potentially represents a trip that would otherwise have required a car. We hope to use data from the DC Capital Bikeshare in 2011 and 2012 to predict bikeshare usage system-wide.

The purpose of this analysis is to determine variables/factors that help estimate bikeshare useage and develop a model that predicts bikeshare useage based on certain predictor variables.
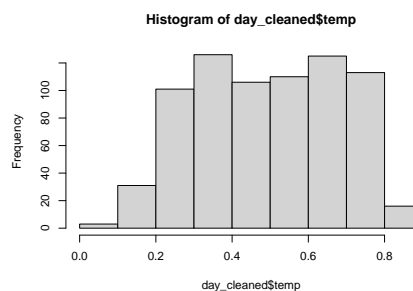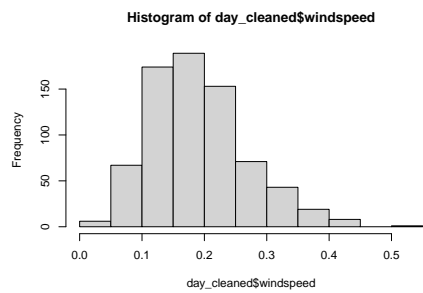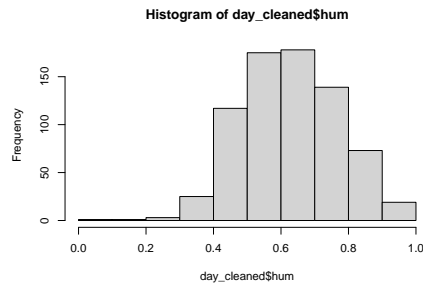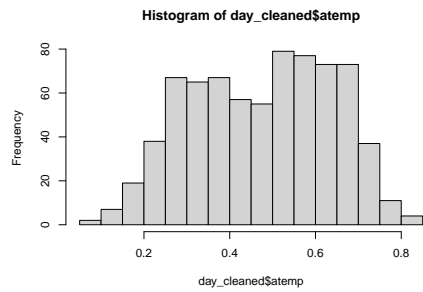
Current Status of the Project

Currently, the dataset has undergone exploratory data analysis and data cleaning. We've identified predictors that may impact a potential model's goodness of fit and converted some variables into categorical variables.

### Data Cleaning and Preprocessing

The dataset has required minimal cleaning. We had to convert several variables into factor variables (season, holiday, weekday, workingday, weather). Additionally we noted that the data key for our dataset mislabeled the season factor variable, which was trivial to correct. Fortunately, there was no missing data.

EDA / Visualization

Histogram of day_cleaned$temp

Abowath, Blakely,
Garrison, Loui, Peir

Forecasting Bikesharing Usage
Team 10 - Progress Report

MGT 6203

Fall 2023

Histogram of day_cleaned$atemp

Histogram of day_cleaned$hum

**Histogram of day_cleaned$windspeed**

Not all of our predictors or response variables are distributed normally. In particular, humidity exhibits leftward skew, and windspeed exhibits rightward skew.

Histogram of day_cleaned$casual

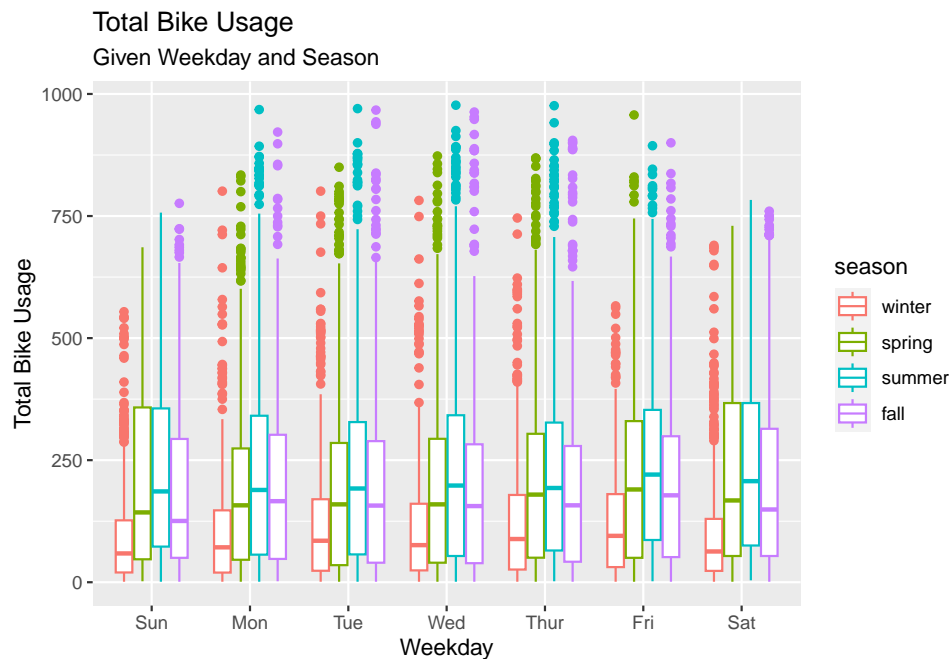**Histogram of day_cleaned$registered**
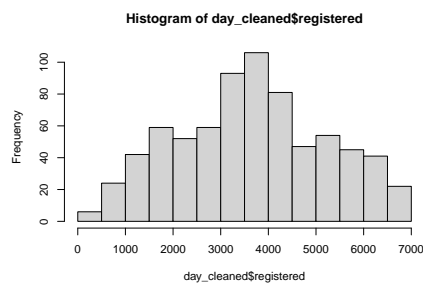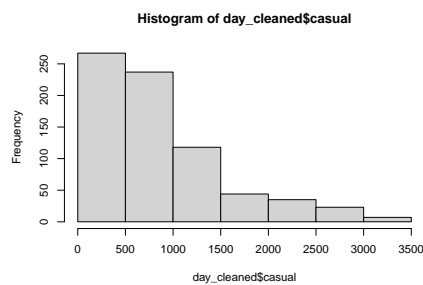
Casual users exhibit rightward skew, while registered users are not too far from the normal distribution.

Preliminary EDA reveals strong seasonality - overall usage on a daily basis is much higher in spring/summer than fall and much more than winter.

We wanted to look at the overall bike usage as it relates to seasonality. We had the hypothesis that we would have more users in the summer than in the winter. Looking at the plot below, we can see that not only do we have the max bike usage during the summer, but we also seem to have an increase to bike usage as the service becomes more mature.
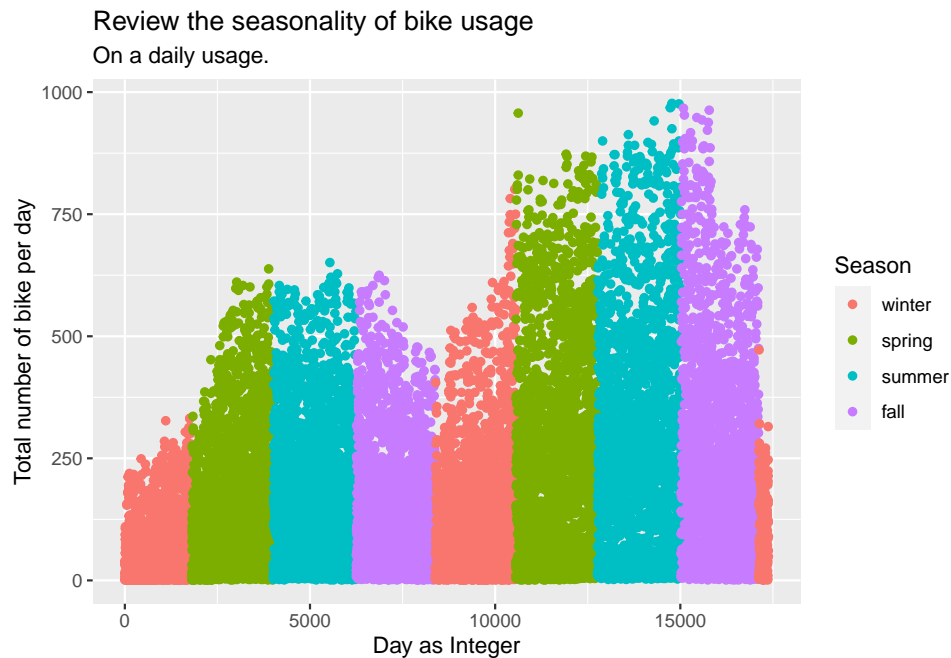
The next thing we wanted to take a look at is see whether the day of the week had a large impact on the total bike usage. This is important to see whether a specific day will have the majority of usage or if they are relatively evenly spread out. Ideally we want to see the data spread out. This would allow us to increase on the capitalization of renting our bikes every day.

Abowath, Blakely,
Garrison, Loui, Peir

**Forecasting Bikesharing Usage
Team 10 - Progress Report**

MGT 6203

Fall 2023

Total Bike Usage
Given Weekday and Season

Next, let's verify these findings by conducting an ANOVA and pairwise analysis.

```
##                  Df    Sum Sq   Mean Sq F value Pr(>F)
## df_h$season       3  37729358  12576453   409.2 <2e-16 ***
## Residuals     17375 534032233     30736
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of the F-statistic for season is very small and statistically significant. At least one group mean is different from the rest.

Next, we compare the means of each pair of seasons.

```
##    Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = df_h$total_count ~ df_h$season, data = df_h)
##
## $`df_h$season`
##                    diff        lwr        upr     p adj
## spring-winter  97.229500   87.54202 106.9169764 0.0000000
## summer-winter 124.901668  115.26026 134.5430748 0.0000000
## fall-winter    87.754288   77.96798  97.5405970 0.0000000
## summer-spring  27.672168   18.12517  37.2191613 0.0000000
## fall-spring    -9.475213  -19.16852   0.2180949 0.0581801
## fall-summer   -37.147380  -46.79465 -27.5001142 0.0000000
```

From the Tukey method, all the pairs are statistically different except the fall-spring pair. This confirms the visual in the box plot, where we see similar distributions in fall and spring.

We can see by the above given boxplot, that the data appears to be evenly spread out between each weekday and across all the seasons. This shows that in general each day will yield approximately the same number of riders, which will help increase our profit. The question though is do we see a difference between casual and registered users.

Registered Users vs Casual Users
Based on Weekday and Seasonality
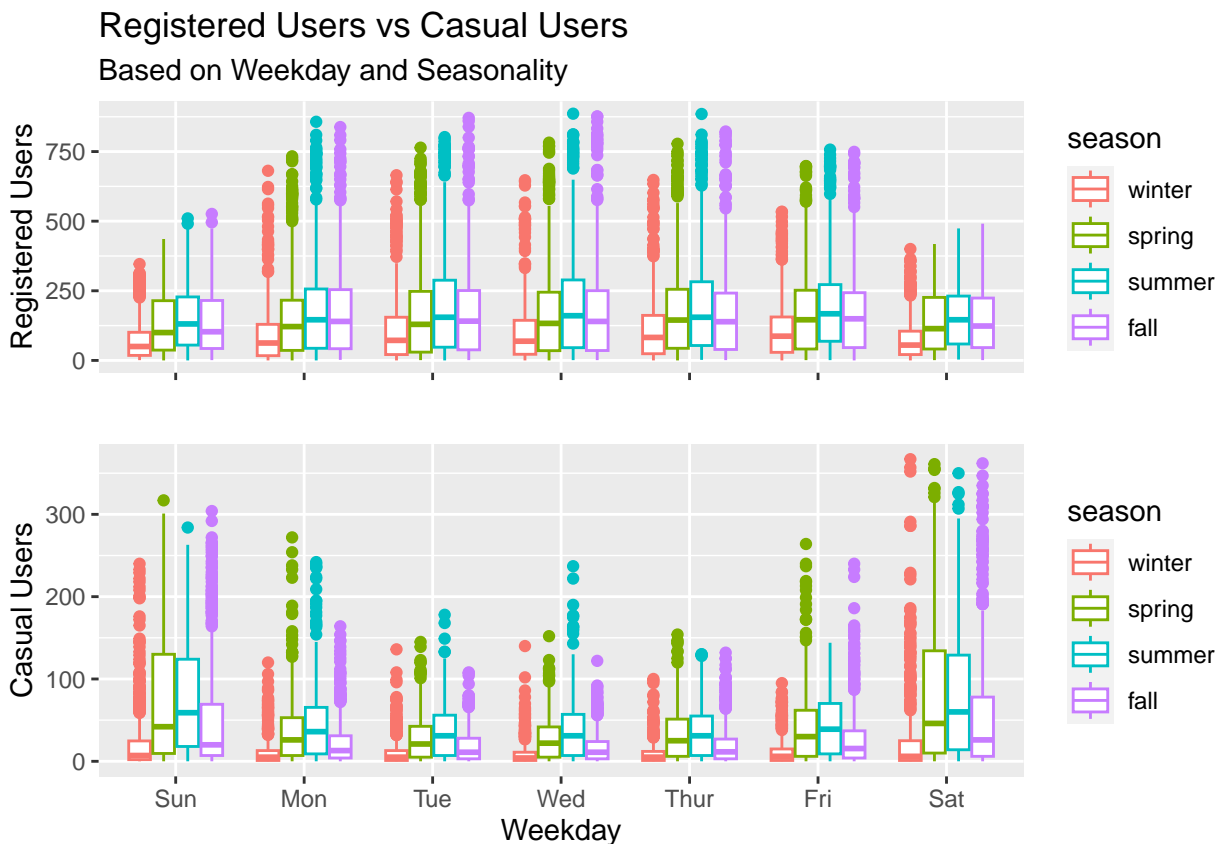
We can see above that the majority of the `casual` user's usage is on the weekend, and the registered user uses them consistently throughout the week. This indicates that we might need to look at different models depending on the user type. We might see a little more non-linearity with casual users than with the registered users.

Preliminary Time Series Forecasting

We have started with a simple time series forecasting model using the day wise usage data. We have used ARIMA to forecast for on the last 51 days of the year. We understand from the data analysis that we have conducted above that seasonal effects do have an effect on the usage.

```
##                   ME      RMSE      MAE       MPE      MAPE
## Test set  -915.7404  1846.462  1295.75  -70.53495  76.82173
```

Daily Bike–Sharing Demand Forecast
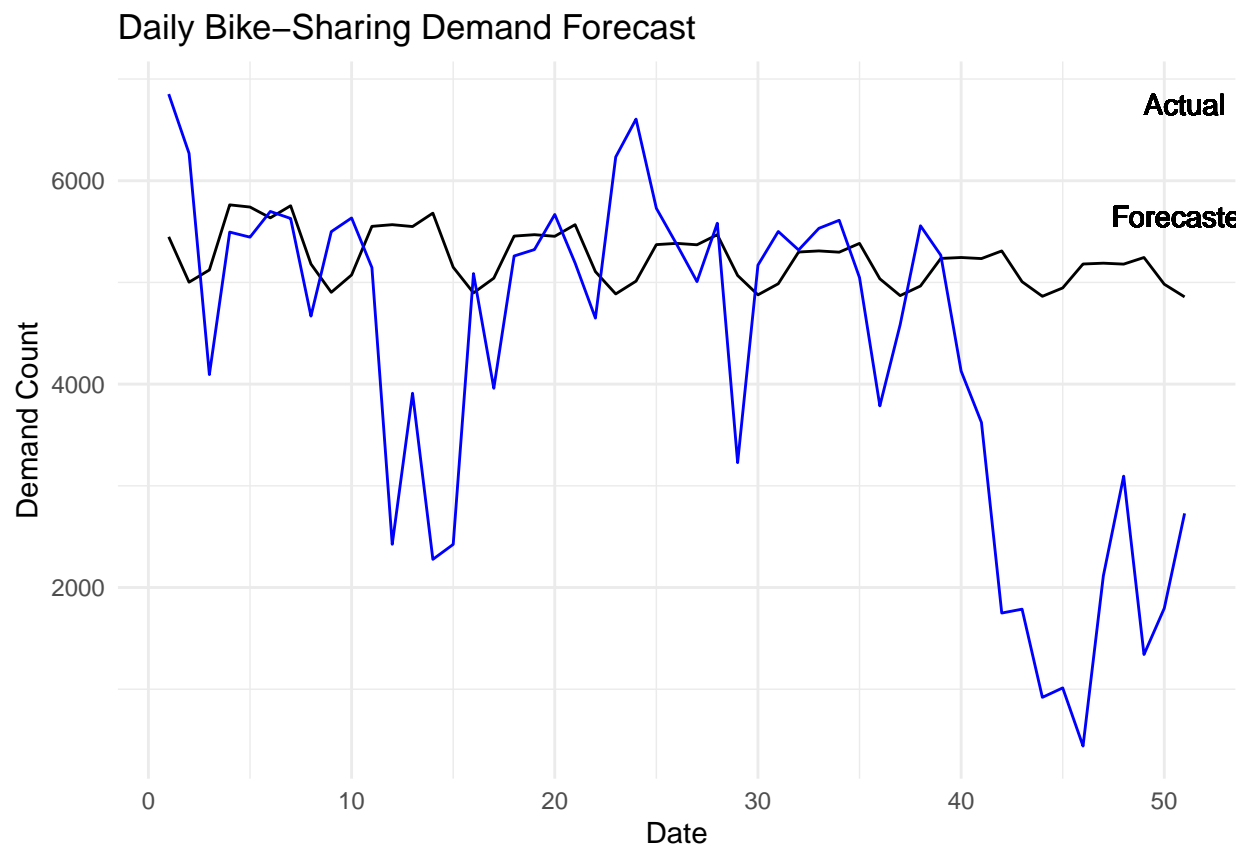
Ongoing Work

Anticipated challenges:

- Distinguishing between the influences of weather and the influence of the season, particularly on casual usage. DC has many visitors in Spring and Summer who would show up under the casual response variable.

- Accounting for the overall increase in usage over the two years spanning our dataset. The dataset was collected towards the beginning of the Capital Bikeshare program, so findings we draw from the overall increasing usage would not necessarily translate directly to mature systems.

Future Modeling

The initial results of a time series forecasting model on a day level depicts that our accuracy is pretty low and especially it is unable to gauge the seasonal effects in the data especially the sudden drop of usage in winters. In order to improve our models we will now explore different modeling techniques such as Decision tree/Ensemble trees and use train and test data set that has all the components of seasons and other time level factors. In addition to this, we will also try to improve on the time series model by either exploring the hourly patters and forecasting for usage for a

particular hours in a day. Lastly, we will consider getting more years of data which will allow us to capture the seasonal and time based patterns of usage in the model.

Additionally, as mentioned above, the two different response variables `reigstered` and `casual` exhibit significant differences and likely warrant building two separate models. Due to the correlation of our temperature variables with seasons (logically) we may needs to consider variable selection techniques to reduce the number of factors down.

Literature Review Summary

Existing literature around bikeshare usage generally emphasizes the following:

- Time of day is typically the most important predictor, but different days of the week have different trends based on time of day
- Specifically, usage is often bimodal on weekdays reflecting commuter patterns
- Usage is not bimodal on weekends, typically with the highest value in mid-afternoon.
- Usage increases as temperature increases, then starts to decrease as temperatures go into the 90s (Fahrenheit), which can be too hot
- Precipitation of any amount discourages cycling
- High humidity has a negative effect on cycling
- High winds can have a negative effect on cycling
- Usage is often higher in spring and summer, and lowest in winter

Works Cited

Bean, R., Pojani, D., & Corcoran, J. (2021). How does weather affect bikeshare use? A comparative analysis of forty cities across climate zones. *Journal of Transport Geography*, 95. https://doi.org/10.1016/j.jtrangeo.2021.103155.

Eren, E., & Uz, V. E. (2020). A review on bike-sharing: The factors affecting bike-sharing demand. *Sustainable Cities and Society*, 54. https://doi.org/10.1016/j.scs.2019.101882

Ashgar, H. I., Elhenawy, M., & Rakha, H. A. (2019). Modeling bike counts in a bike-sharing system considering the effect of weather conditions. *Case Studies on Transport Policy*, 7(2), 261-268. https://doi.org/10.1016/j.cstp.2019.02.011