Forecasting Bikesharing Usage for DC's Capital Bikshare System

Table of Contents

- Introduction
- Current Status of the Project
- Ongoing Work
- Literature Review Summary
- Works Cited

Introduction

Bikesharing systems are an increasingly popular solution in major urban areas to increase the usage of bicycles as a mode of transport. The riding of bikes helps to improve 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 bike usage and develop a model that predicts the usage based on certain predictor variables.

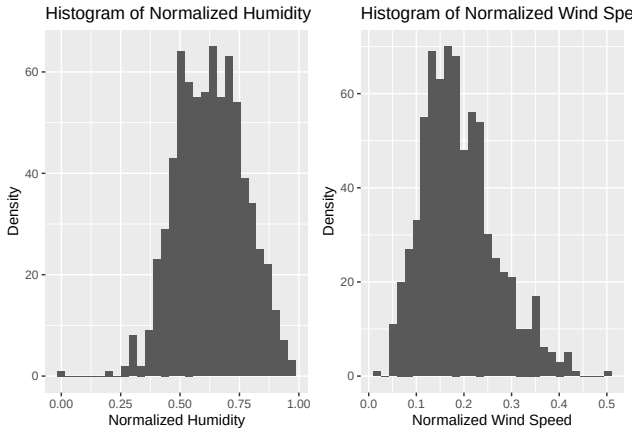
Current Status of the Project

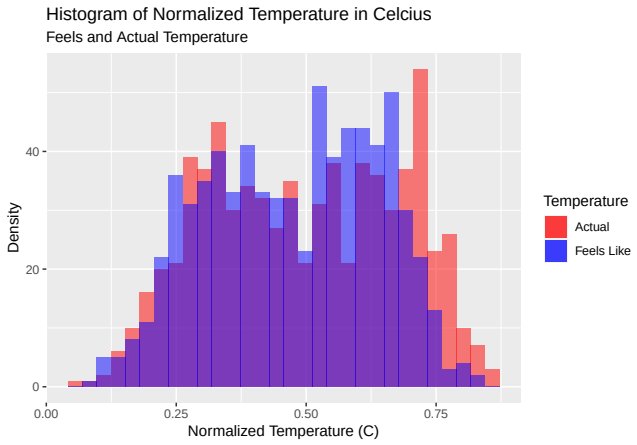
Currently, the dataset has undergone exploratory data analysis and data cleaning. We've identified predictors that may impact a potential model's goodness of fit and converted some variables into categorical variables. We have also applied very basic ARIMA and Linear Regression models to get an idea of the data. These models are very basic, but can be used to help guide the refinement of the model and the selection of the independent variables.

Data Cleaning and Preprocessing

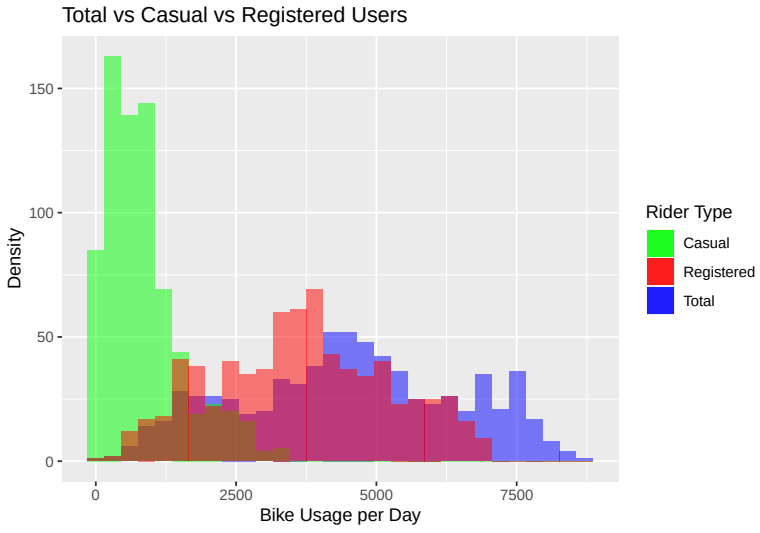
The dataset has required minimal cleaning. We had to convert several variables into factor variables (`season`, `holiday`, `weekday`, `workingday`, `weather`). Additionally we noted that the key for our dataset mislabeled the season variable, which was trivial to correct. Fortunately, there was no missing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

EDA / Visualization



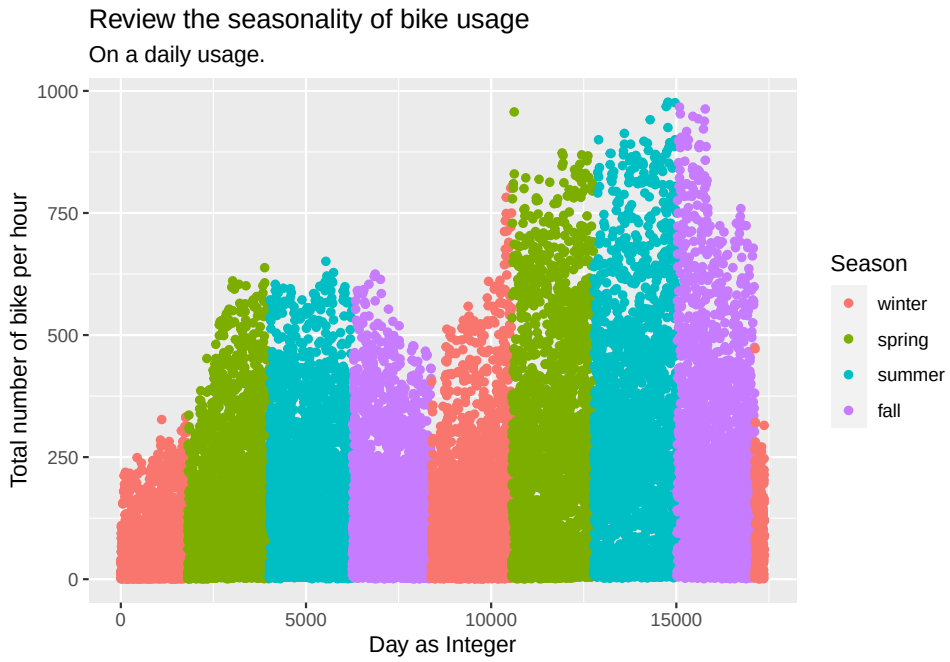


Not all of our predictors or response variables are distributed normally. In particular, `humidity` exhibits leftward skew, and `windspeed` exhibits rightward skew.

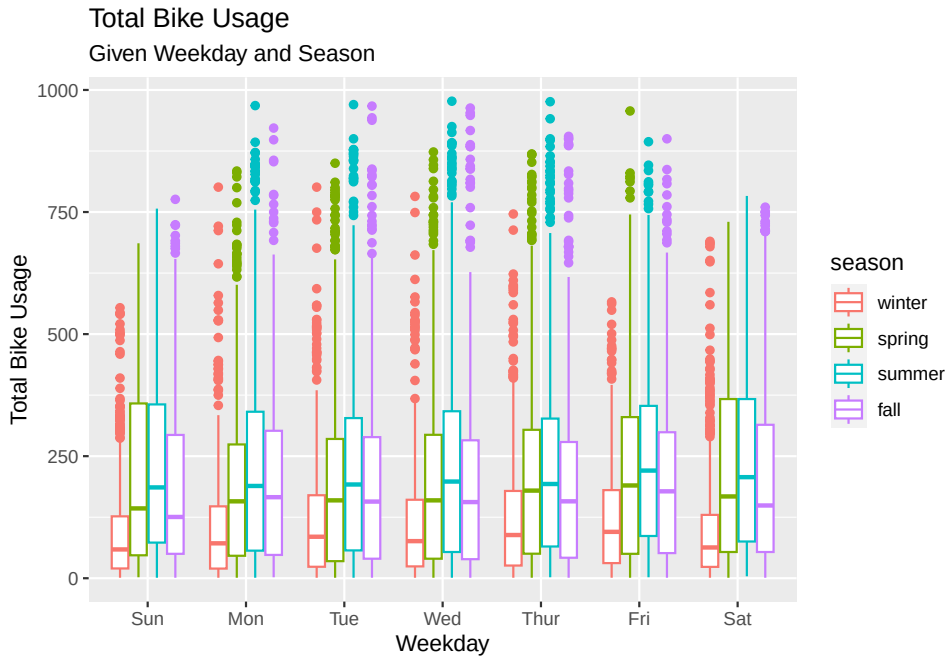


Casual users exhibit rightward skew, while registered users are not too far from the normal distribution.

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 a year-over-year increase to the usage as the service matures.



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. Total bike usage is the combination of the `registered` and `casual` users. 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.



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.

We can also verify these findings by conducting an ANOVA and pairwise analysis.

Terms	DoF	Sum Sq	Mean Sq	F-Value	Pr(>F)
df_h$season	3	37729358	12576452.56	409.181	0
Residuals	17375	534032233	30735.67	NA	NA

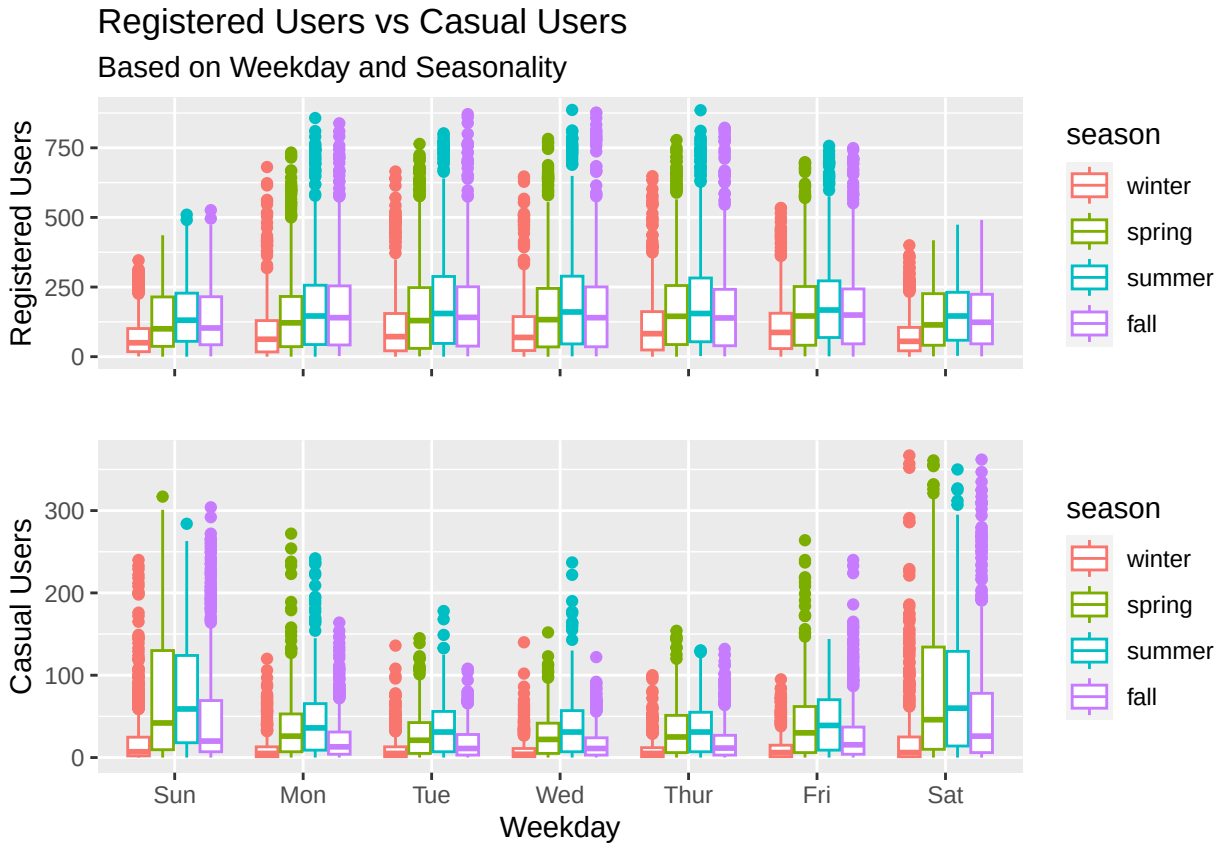
The p-value of the F-statistic for season is very small and statistically significant. At least one group mean is different from the rest.

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

term	contrast	null.value	estimate	conf.low	conf.high	adj.p.value
df_h$season	spring-winter	0	97.229500	87.54202	106.9169764	0.0000000
df_h$season	summer-winter	0	124.901668	115.26026	134.5430748	0.0000000
df_h$season	fall-winter	0	87.754288	77.96798	97.5405970	0.0000000
df_h$season	summer-spring	0	27.672168	18.12517	37.2191613	0.0000000
df_h$season	fall-spring	0	-9.475213	-19.16852	0.2180949	0.0581801
df_h$season	fall-summer	0	-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.

The next question though is do we see a difference between `casual` and `registered` users.

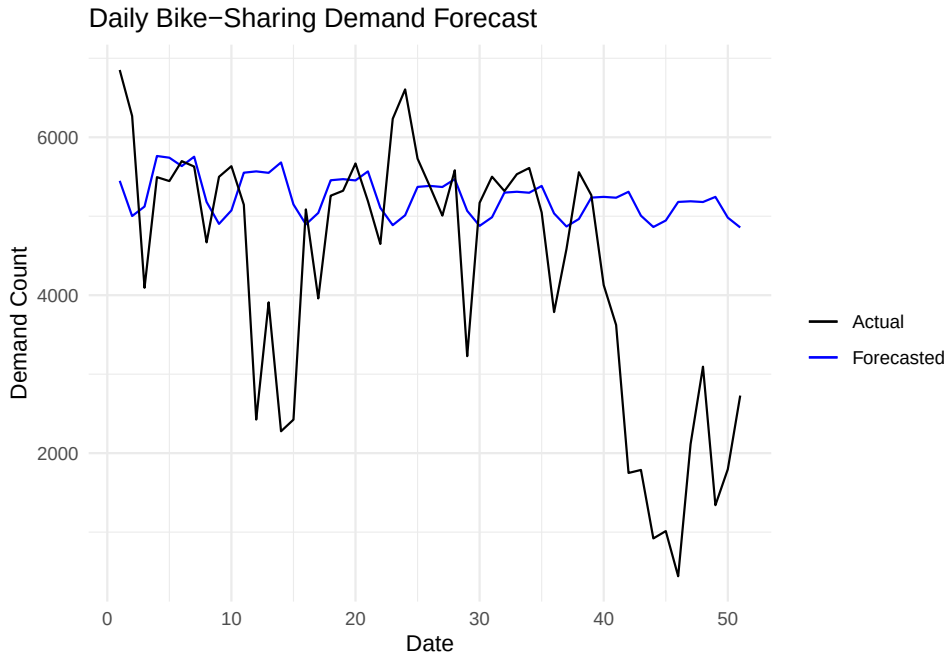


We can see above that the majority of the `casual` user's usage is on the weekend, and the `registered` user use the bikes consistently throughout the week. This indicates that we might need to look at different models depending on the user type. We might see a little more non-linearity with casual users 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 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
```



There is a lot of room for improvement with building and fine tuning our models.

Ongoing Work

Anticipated challenges:

- Distinguishing between the influences of `weather` and the influence of the seasons, 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. It is also unable to determine the seasonal effects in the data, e.g. the sudden drop of usage in winters. In order to improve our models we will explore different modeling techniques such as Decision tree/Ensemble trees. The models will have a 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 patterns and forecasting the 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. We have found more data for the DC Capital Bike Share program, but it doesn't include any weather forecasts. Therefore, depending on the functionality of the model we may or may not use that data (can find the extra

data here: capitalbikeshare-data). This will require the pulling of additional weather data to be used along with the bike data.

Additionally, as mentioned above, the two different response variables `registered` 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