# Project Outline

Jake Berberian

## Introduction

### Aim

We will use a dataset from UCI's Machine Learning Repository, called Bike Sharing Dataset. It contains data from Capital Bikeshare's official website, joined with weather data from I-weather and DC's holiday schedule. The project will aim to study the relationship between the number of users and the weather/holiday schedule. This particular research question is of interest as it is looking at DC data. Furthermore, the use of bike-share systems can tell you much more than Metro data may be able to.

### Research Questions

Are we able to predict the number of users in a given day based upon the weather? Are we able to cluster the number of riders given certain weather inputs? We'll look to use regression analysis and clasification techniques (LDA/QDA). Additionally, if we find multicolinearity (which is likely- temperature could depend on forecast/wind) we would carry out a ridge regression. If we can find some way to classify/predict the number of bikers based on weather, we could be able to suggest the best days for repairs/upgrades, as well as estimating a number of users in a given week- which could help keep track of things like time until repair. Also, being able to predict based off weather/holidays will be able to give us an insight on whether a "peak fare" could be charged or not for these bikes. Naturally, we suspect rush hours to have a higher number of users, but maybe less users are willing to bike for their commute if it's below a certain temperature. Finally, we'll want to try to predict the next year of bike users to create some sort of expected income. This can indicate to Capital Bike Share how much they should be spending on upgrades/bike stations and whether or not they should consider expanding with more stations in DC.

### Literature Review

The linked paper in the UCI ML repository discusses the idea that these bike-sharing services can be turned into a virtual sensor network used for sensing mobility in the city. This is a strength, as there's more than just business analytics being measured. However, only using two years worth of data seems like a weakness, as the second year had considerably more data points than year one. This is most likely due to the fact that 2011 was only a year after Capital Bikeshare was established. We'd lileky see larger growth over the following few years before stabilization, so the 2017-2020 datasets will also be important to factor in.

## Initial Hypotheses

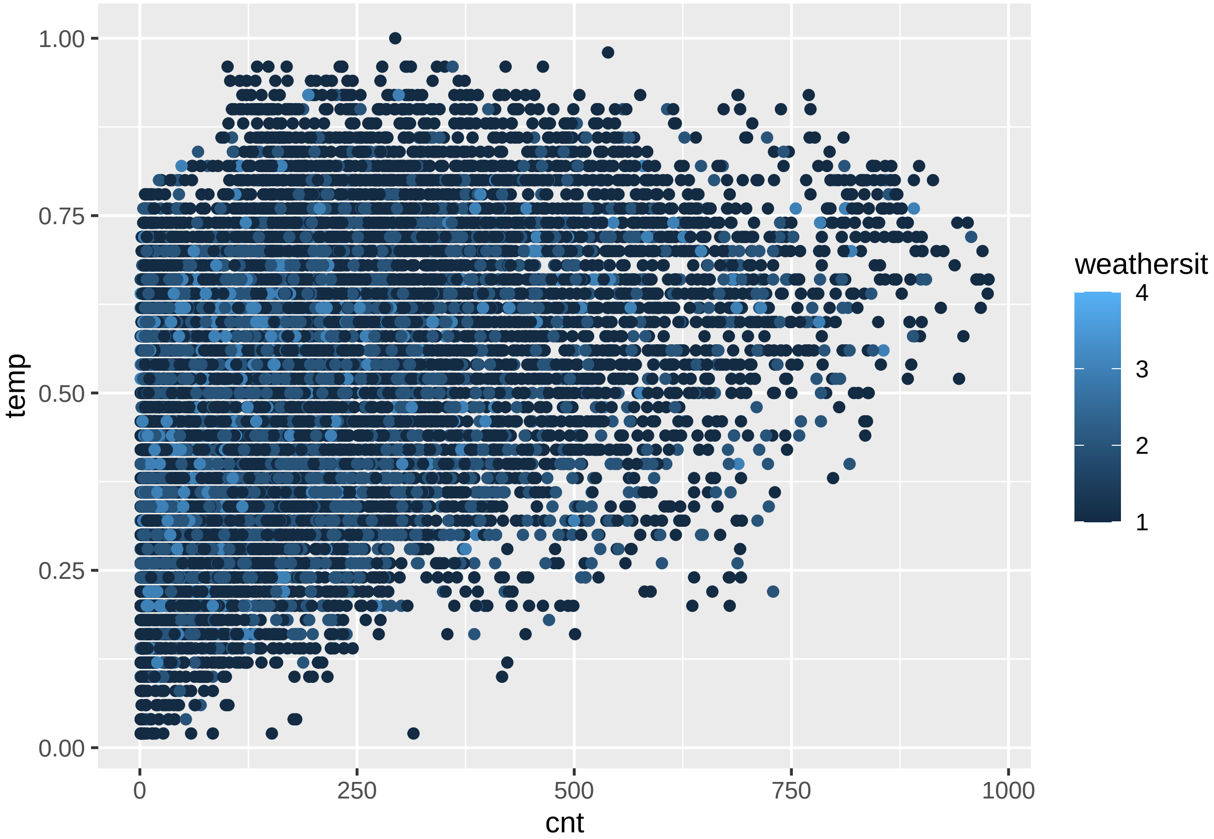
Our initial hypotheses are the following:

- Workdays/holidays and days with lower temperatures/worse weather will result in lower usage.
- Non-workdays will result in a much higher proportion of non-registered users using bikes.

- We will see a decrease in users in the summer months.
- Year-to-year bike data has probably now stabilized and is easier to predict.
- Predicting the number of users in a given day will be difficult- but like above, the number of users will increase on nice, 60-85 degree workdays.

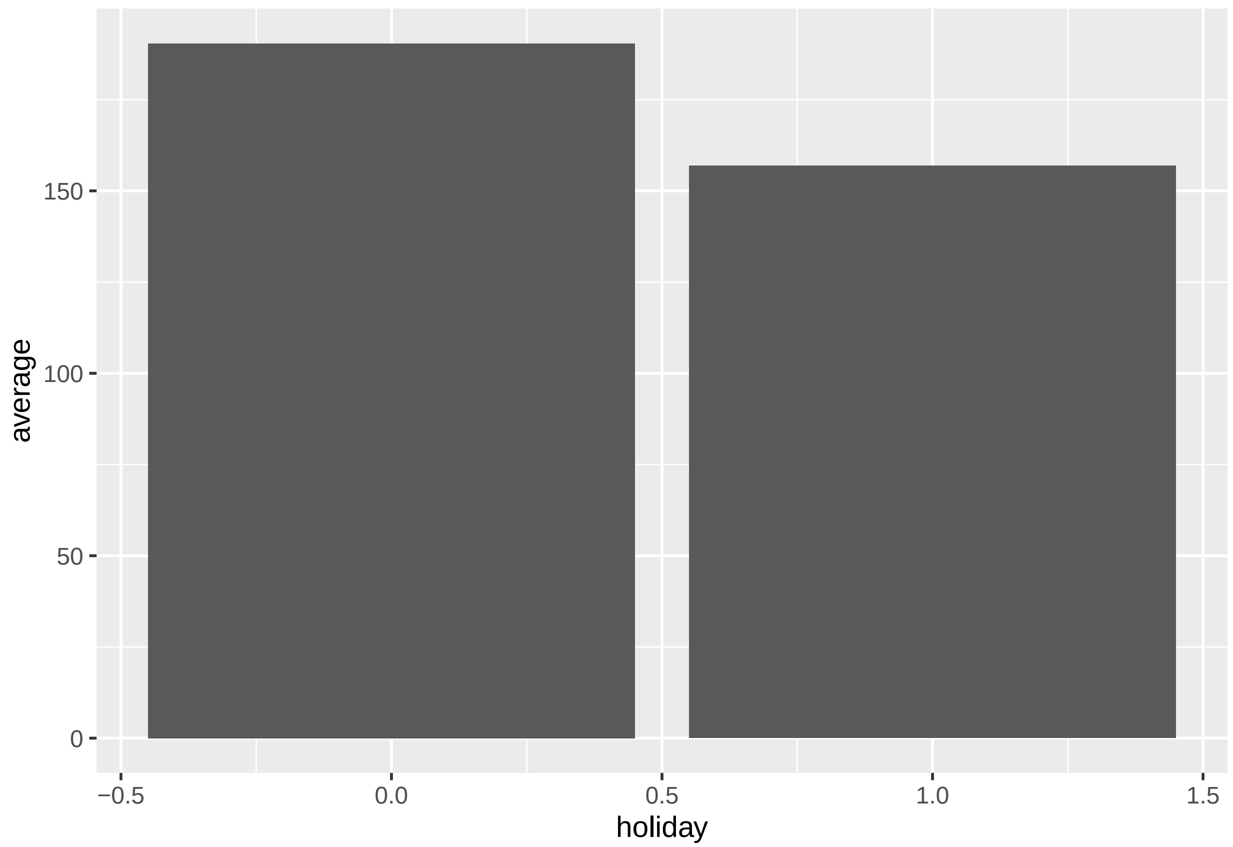
## Data-driven Hypotheses

We'll take a quick look at how count is impacted by the temperature and type of weather.



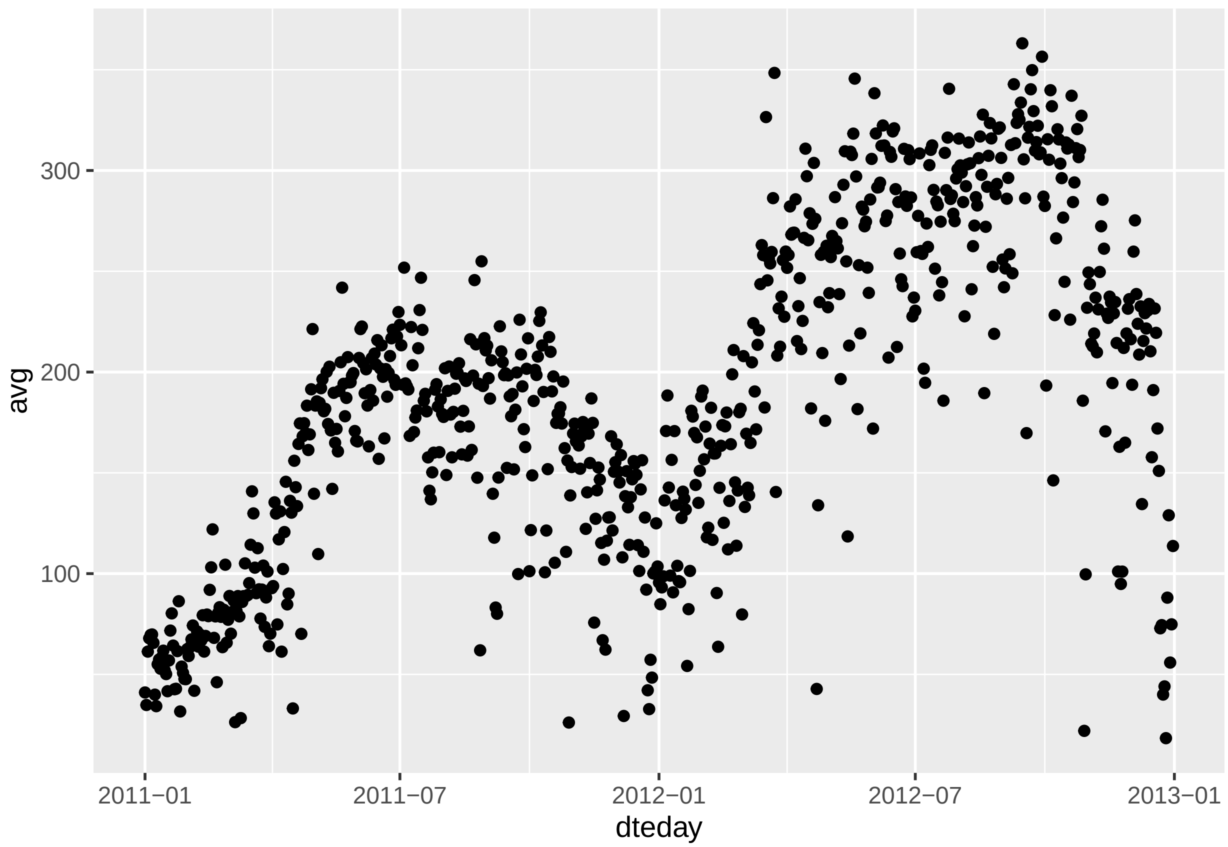
There seems to be a funnel shape, which would make sense. As the temperature converges to "optimal" bike-riding temperature, there will be more riders. Furthermore, we see a large amount of darker points, indicating that more people are riding in nicer weather. This is to be expected from our initial hypotheses.

Now, let's take a look at how holidays affect bike-usage.



We'll adjust to take the average number of riders on holidays and non-holidays. As expected, there are more bikers using Capital Bikeshare on non-holidays.

Finally, we'll take a glance at the number of users in the span of the year.



Again, we can see our initial hypothesis seem to be supported: that the winters of 2011 and 2012 had less people biking. However, it seems that more people than expected are biking in the summer, perhaps regardless of temperature.

## Proposed Work and Discussion

Since this data is from 2011 and 2012, we'll attempt to model the number of users and compare it to the publically available data from more recent on Capital Bikeshare's website. Being able to accurately model the number of users will allow optimization of the operation and maximization of profits. Additionally, we may carry out some possible time-series analysis to determine if the number of users has stabilized over the years. Being able to correctly model the following year then that'll help inform future decisions. This is important because it can apply to most "business" models that rely on outside factors. For example, this can help create models/inform decisions in the electric shooter-share industry, as well as other businesses like AirBnb (predicting how many bookings, service fees, income, etc.) and places in the food industry like McDonalds (predicting number of customers in order to stock foods appropiately.) Many places probably have a general idea of this, but not a concrete model/process for making these decisions.

## References

Fanaee-T, Hadi, and Gama, Joao, 'Event labeling combining ensemble detectors and background knowledge', Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset.

More references will be included for the final presentation.