

Data Science

Capital Bikeshare Analysis



Shu Shang

April, 2019

Table of Contents

[Overview 3](#_Toc7099879)

[Where do capital Bikeshare riders go? 3](#_Toc7099880)

[Which station are most popular? 4](#_Toc7099881)

[When do they ride? 4](#_Toc7099882)

[What days of the week are most rides taken on? 6](#_Toc7099883)

[Machine learning 7](#_Toc7099884)

[Predicting bike using duration with decision trees 7](#_Toc7099885)

# Overview

This report is about capital bikeshare data analysis. The dataset comes from capital Bikeshare website. The data from 2010 to 2017, each csv file has 7 columns. Capital Bikeshare is metro DC’s bikeshare service, with 4,300 bikes and 500+ stations across 6 jurisdictions: Washington, DC.; Arlington, VA; Alexandria, VA; Montgomery, MD; Prince George’s County, MD; and Fairfax County, VA. Designed for quick trips with convenience in mind, it’s a fun and affordable way to get around. I used python and SQL to do data analysis and explained several questions about users’ behavior. And I applied machine learning methods (decision tree and random forest) to predict duration of bike using.

# Where do capital Bikeshare riders go?

**Year 2010 TOP 10 locations riders go**

1. 14th & V St NW
2. 14th & Rhode Island Ave NW
3. Adams Mill & Columbia Rd NW
4. 21st & I St NW
5. 17th & Corcoran St NW
6. Massachusetts Ave & Dupont Circle NW
7. 15th & P St NW
8. 20th St & Florida Ave NW
9. 17th & K St NW
10. 8th & H St NW

**Year 2011 TOP 10 Locations riders go**

1. Adams Mill & Columbia Rd NW
2. 14th & Rhode Island Ave NW
3. 15th & P St NW
4. 14th & V St NW
5. 21st & I St NW
6. Columbus Circle / Union Station
7. Massachusetts Ave & Dupont Circle NW
8. 17th & Corcoran St NW
9. 8th & H St NW
10. USDA / 12th & Independence Ave SW

# Which station are most popular?

**Most popular station in 2010 & 2011**

1. 14th & V St NW
2. 21st & I St NW
3. 15th & P St NW
4. Massachusetts Ave & Dupont Circle NW
5. Adams Mill & Columbia Rd NW
6. 14th & Rhode Island Ave NW
7. 20th St & Florida Ave NW
8. 17th & Corcoran St NW

# When do they ride?

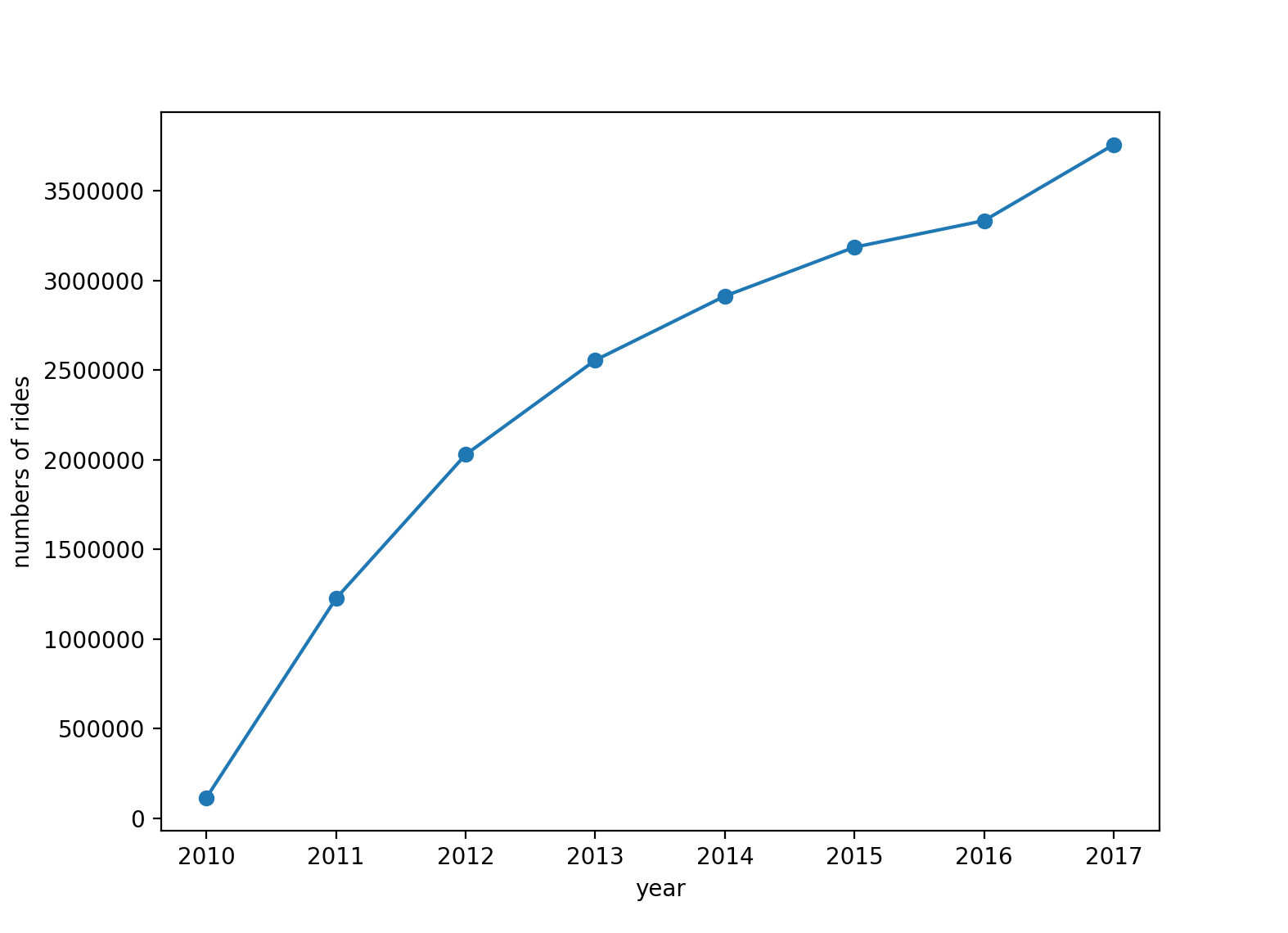


Fig 1. Number of rides from 2010 to 2017

From 2010 to 2017, the number of rides increasing dramatically, it is clear that riding a bike is a new fashion for city people. More and more people choose to take a bike to go to different places.

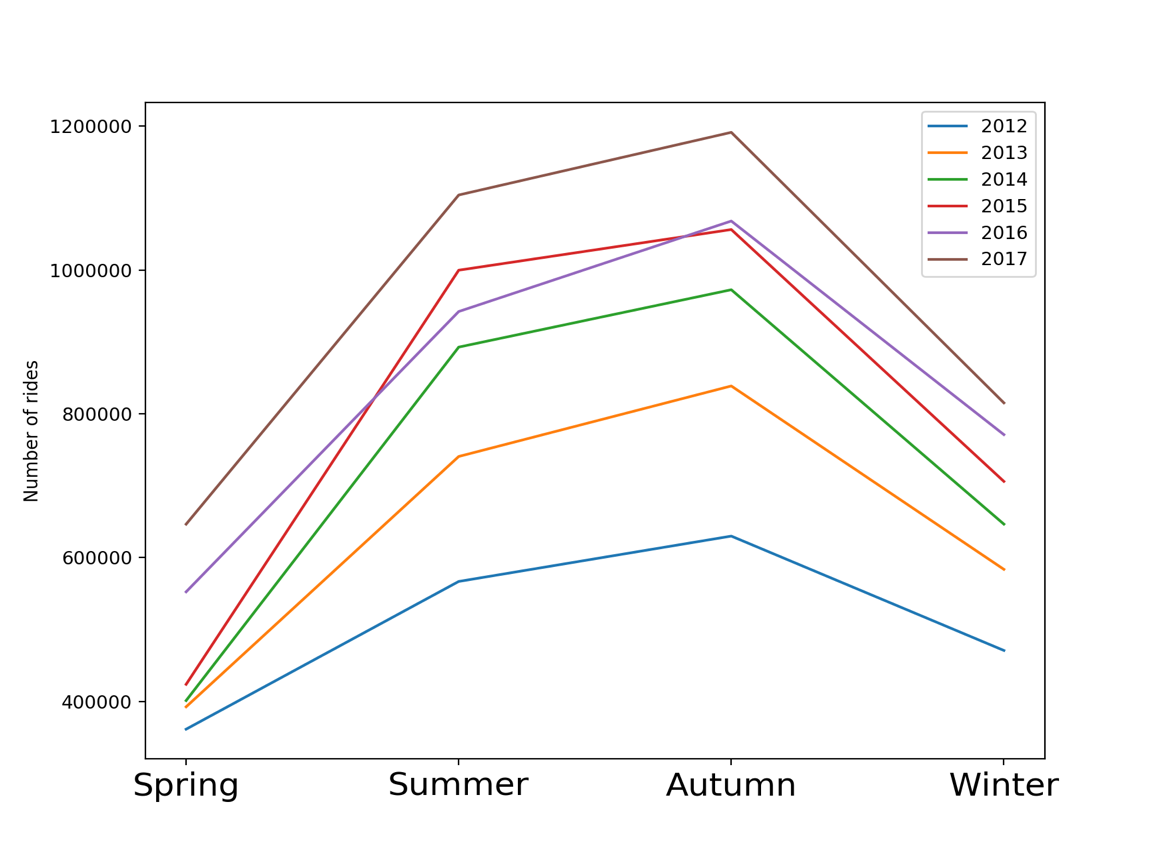


Fig 2. Number of rides in different seasons

Through the Fig 2, people prefer to ride bikes in Summer and Autumn than in Spring and Winter. The mainly reason is environmental temperature in summer and autumns are more suitable for people to go outside.

Comparing data from 2012 to 2017, Autumn is the most popular season for people riding bike. Why? In my opinion, people rode all summer long, Autumn is the cool weather making getting on your bike that much more enjoyable.

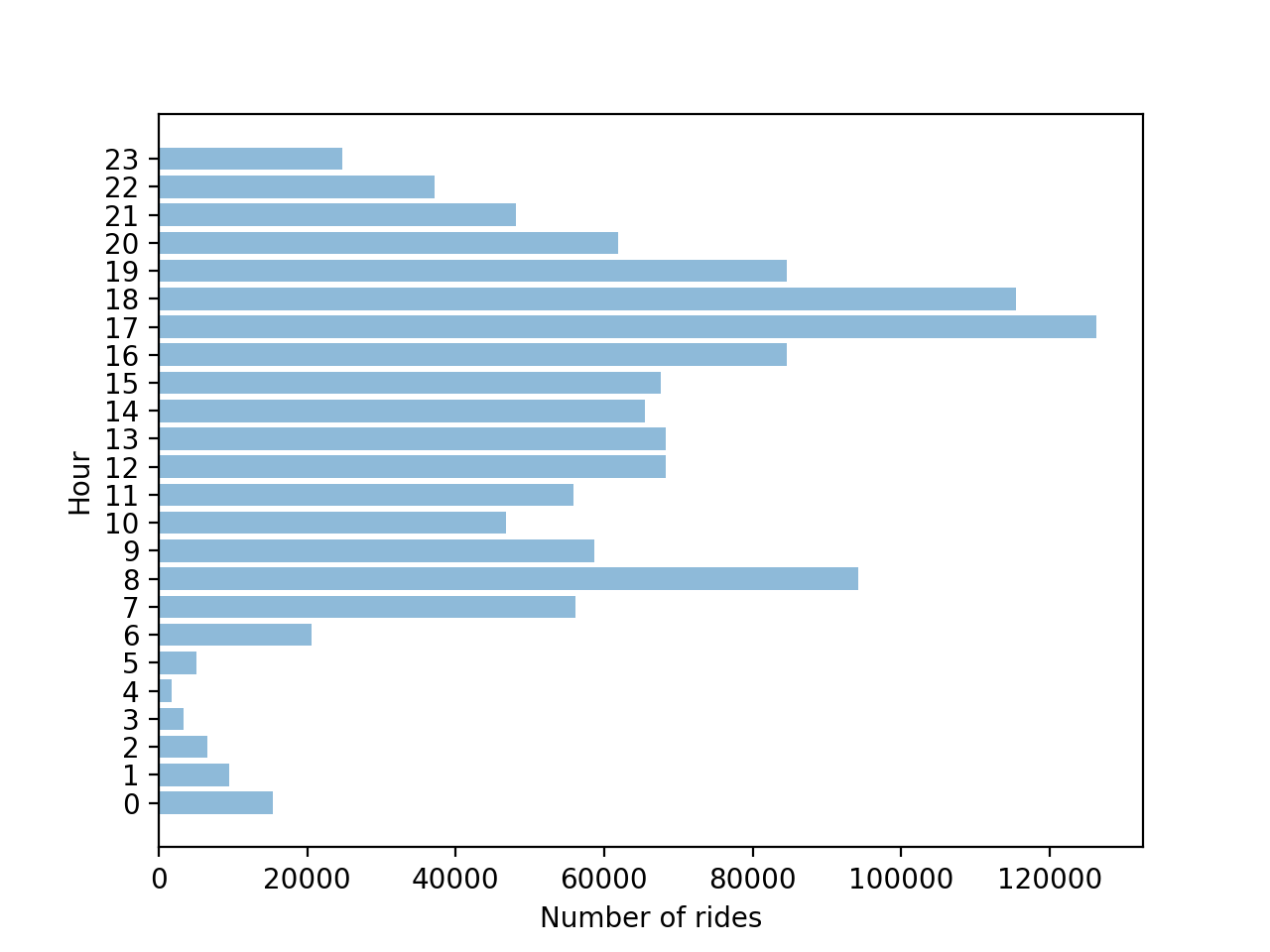
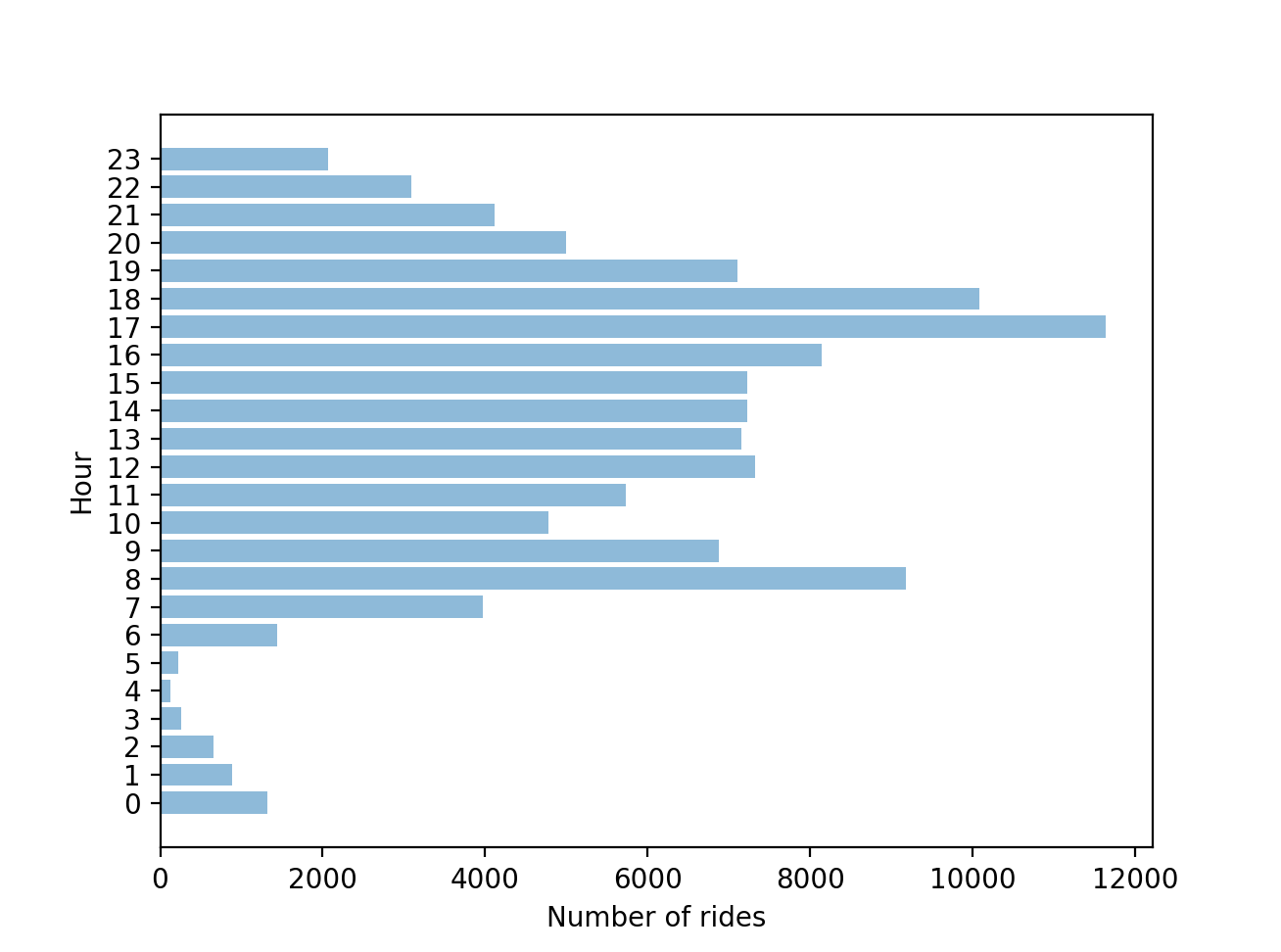


Fig 3. Number of rides in different time in a day (2010, 2011)

From Fig 3, the most popular time of riding is 8-9 am and 5-6 pm. More and more people choose to ride a bike to go to work/school. Comparing data from 2010 and 2011, it is clear that the data trend looks similar, that means riders have fixed habit to ride bikes and the capital bikeshare company has a large number of fixed users which guarantee the business operation stably over years.

# What days of the week are most rides taken on?

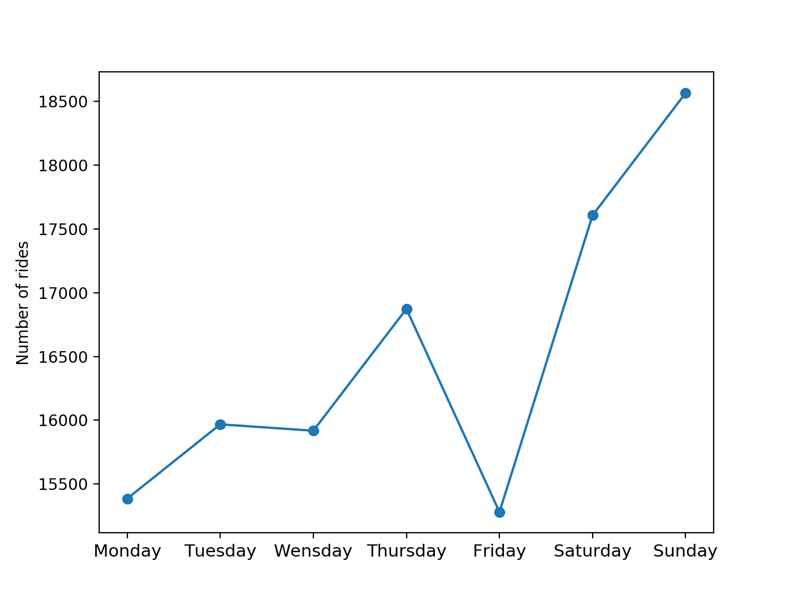
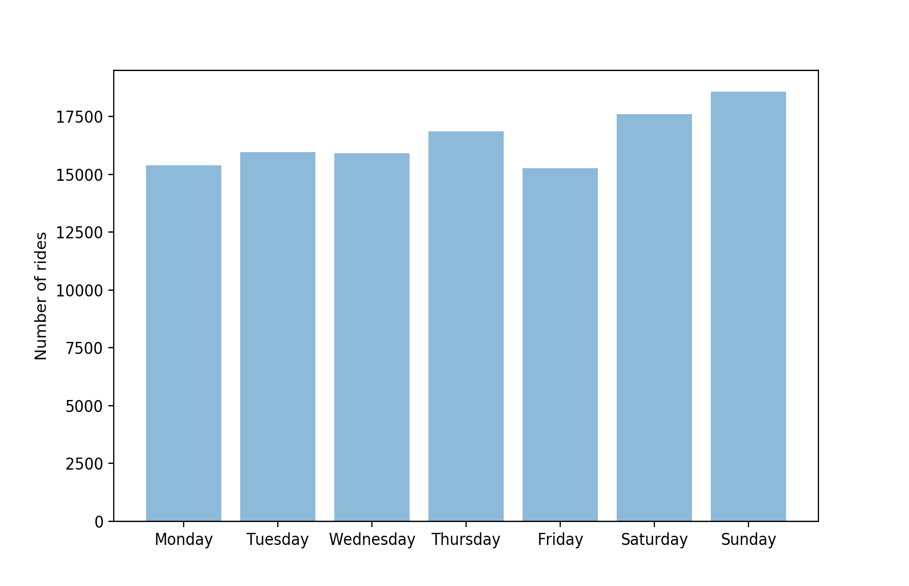


Fig 4. Number of rides in different day of week (2010)

In 2010, most rides happened on Saturday and Sunday. At 2010, people just started to use capital bike App, they prefer to ride a bike on weekend.

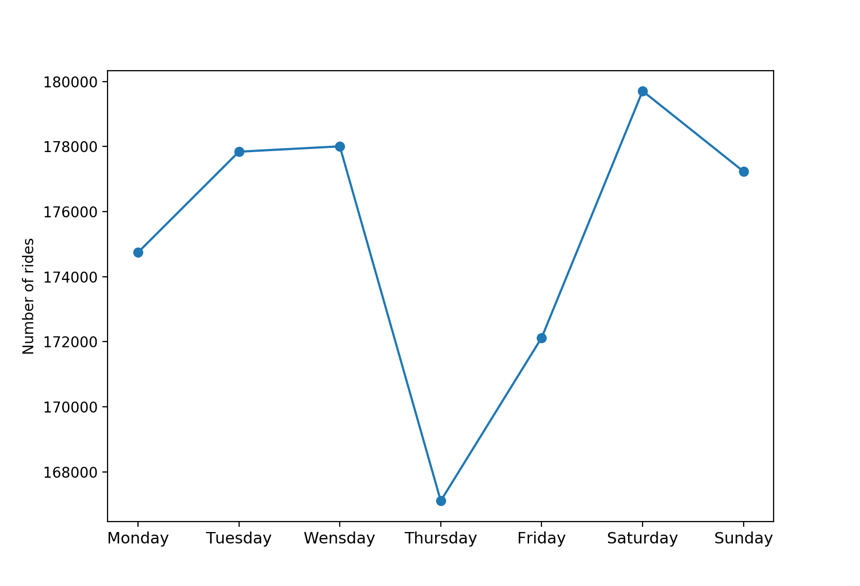
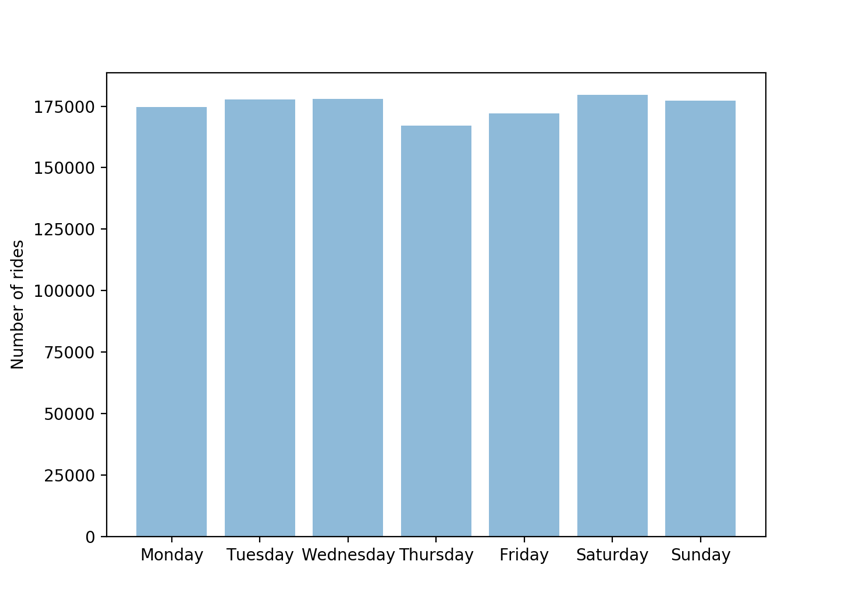


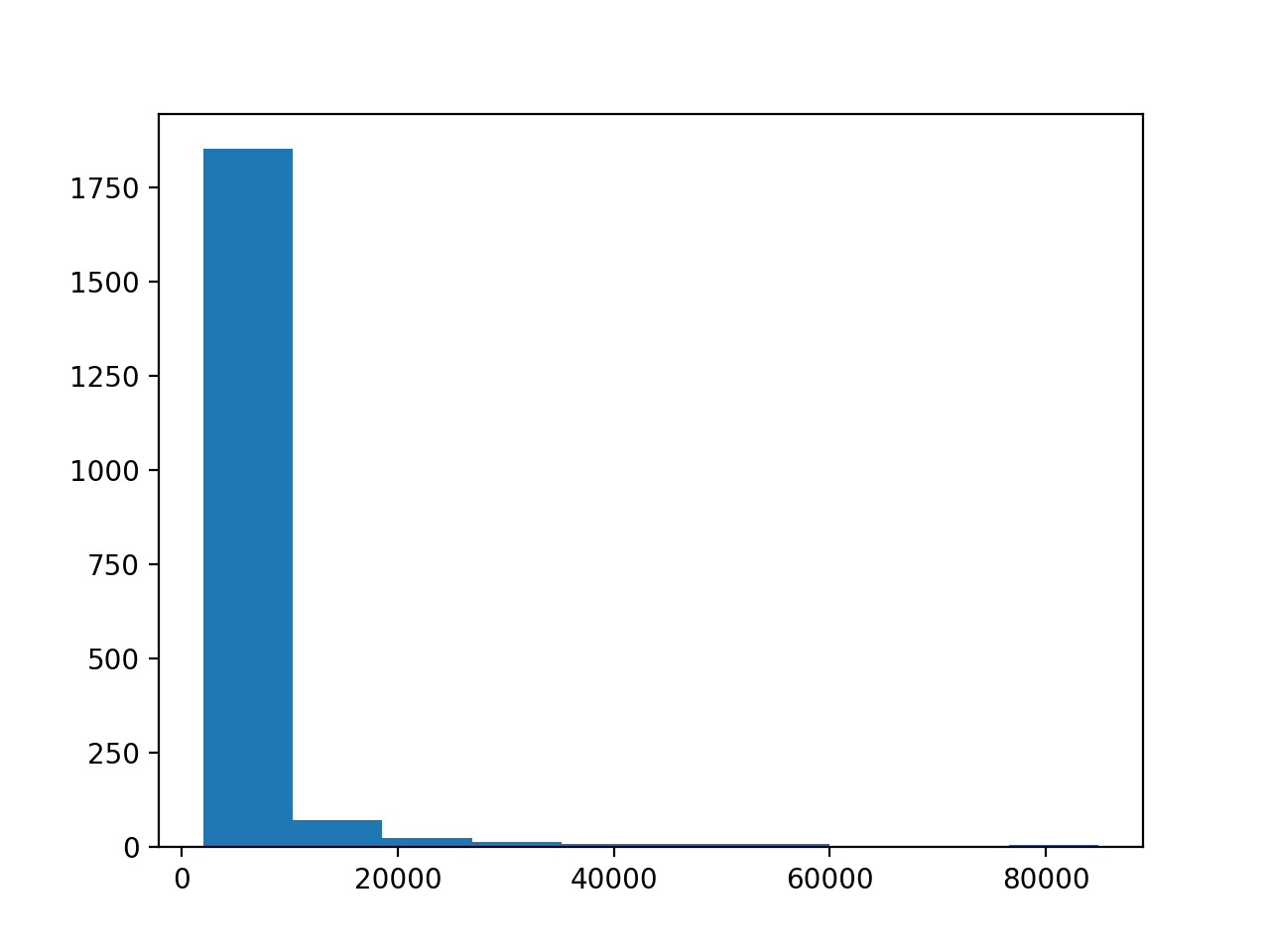
Fig 5. Number of rides in different day of week (2011)

In 2011, most rides happened on Tuesday, Wednesday, Saturday and Sunday. At 2011, the number of rides increased dramatically, more and more people used capital bike on both weekday and weekends. It’s a popular style for people riding to work.

# Machine learning

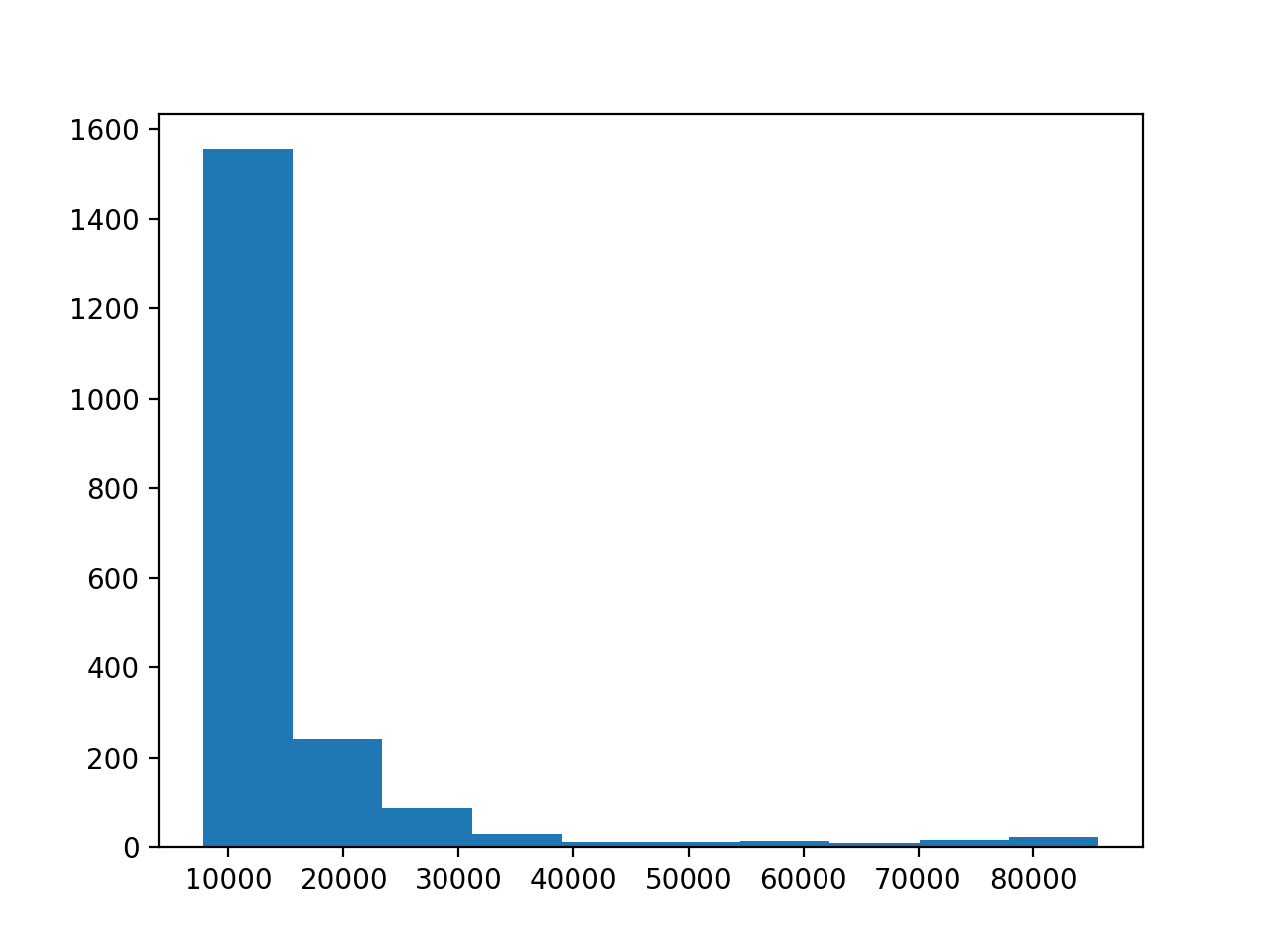
# Predicting bike using duration with decision trees

In this project, we are going to look at dataset that contains duration of rental bikes. From the dataset, we are going to apply various machine learning algorithms to generate a model that can predict the duration of bike using.



X=duration, Y=times

Fig 6. Member users’ duration distribution



X=duration, Y=times

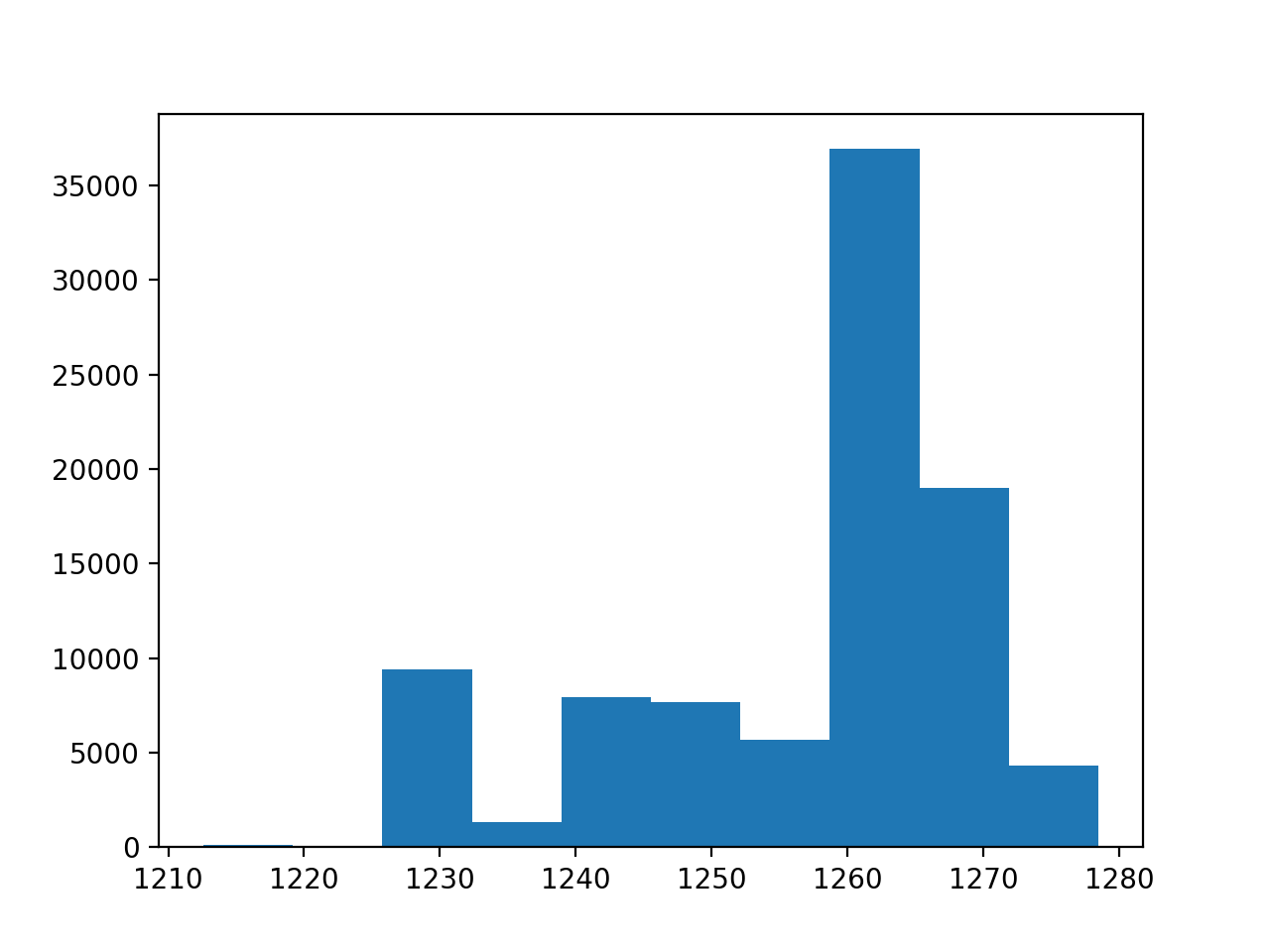
Fig 7. Casual users’ duration distribution

Figure 6 and 7 shows the duration distribution of different member types. It is obvious that more casual member using bike longer time.

Duration statistics (2010)

|  |  |
| --- | --- |
| count | 115597 |
| mean | 1254 |
| std | 2914 |
| min | 60 |
| 25% | 403 |
| 50% | 665 |
| 75% | 1120 |
| max | 85644 |

1. linear regression

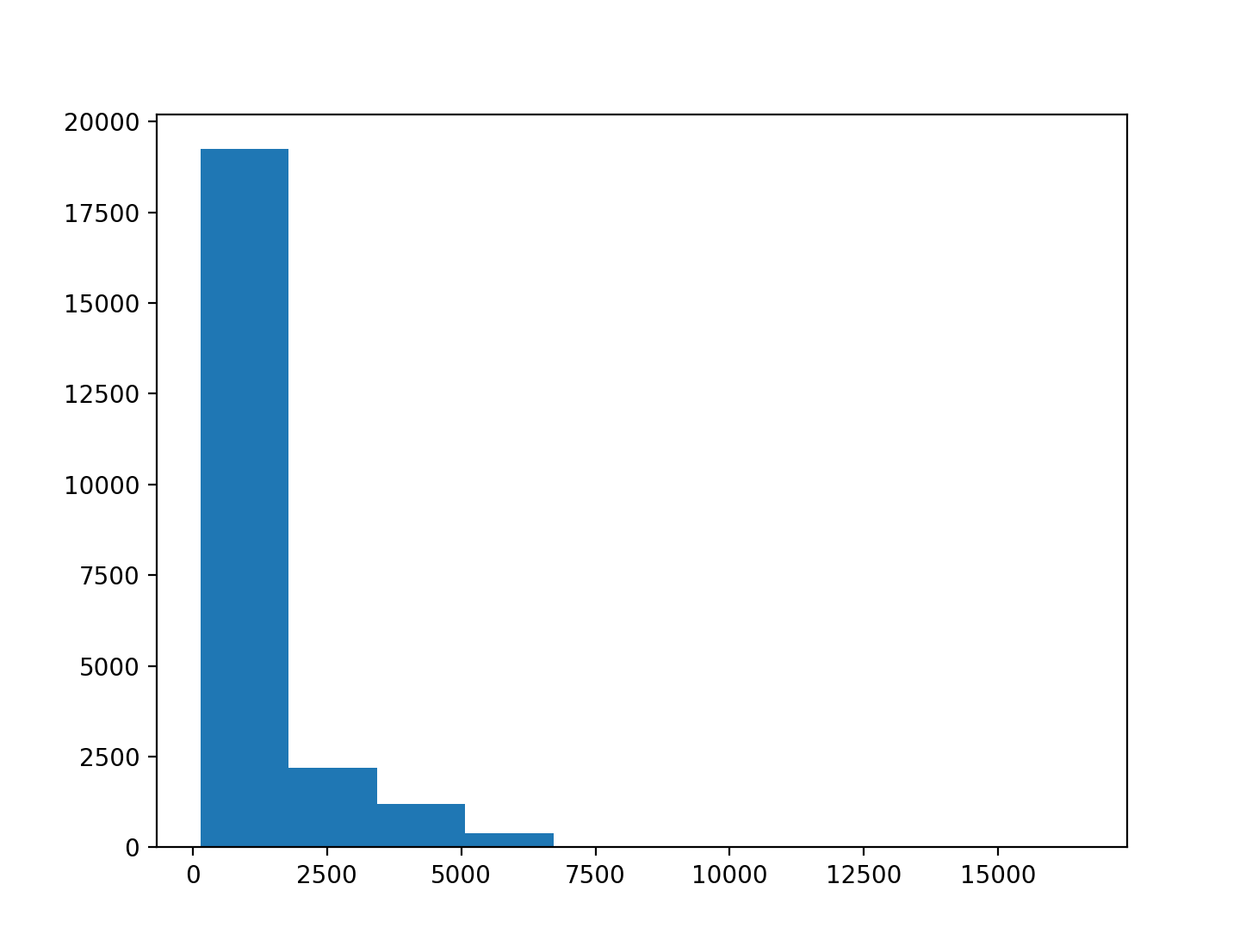


X=duration, Y=times

MSE = 7539610.0724807475

Using linear regression to predict train data, the result shows high MSE. Most of predict value focus on 1260-1270, while the mean value of duration is 1254. So, the predict value is close to mean value of the duration.

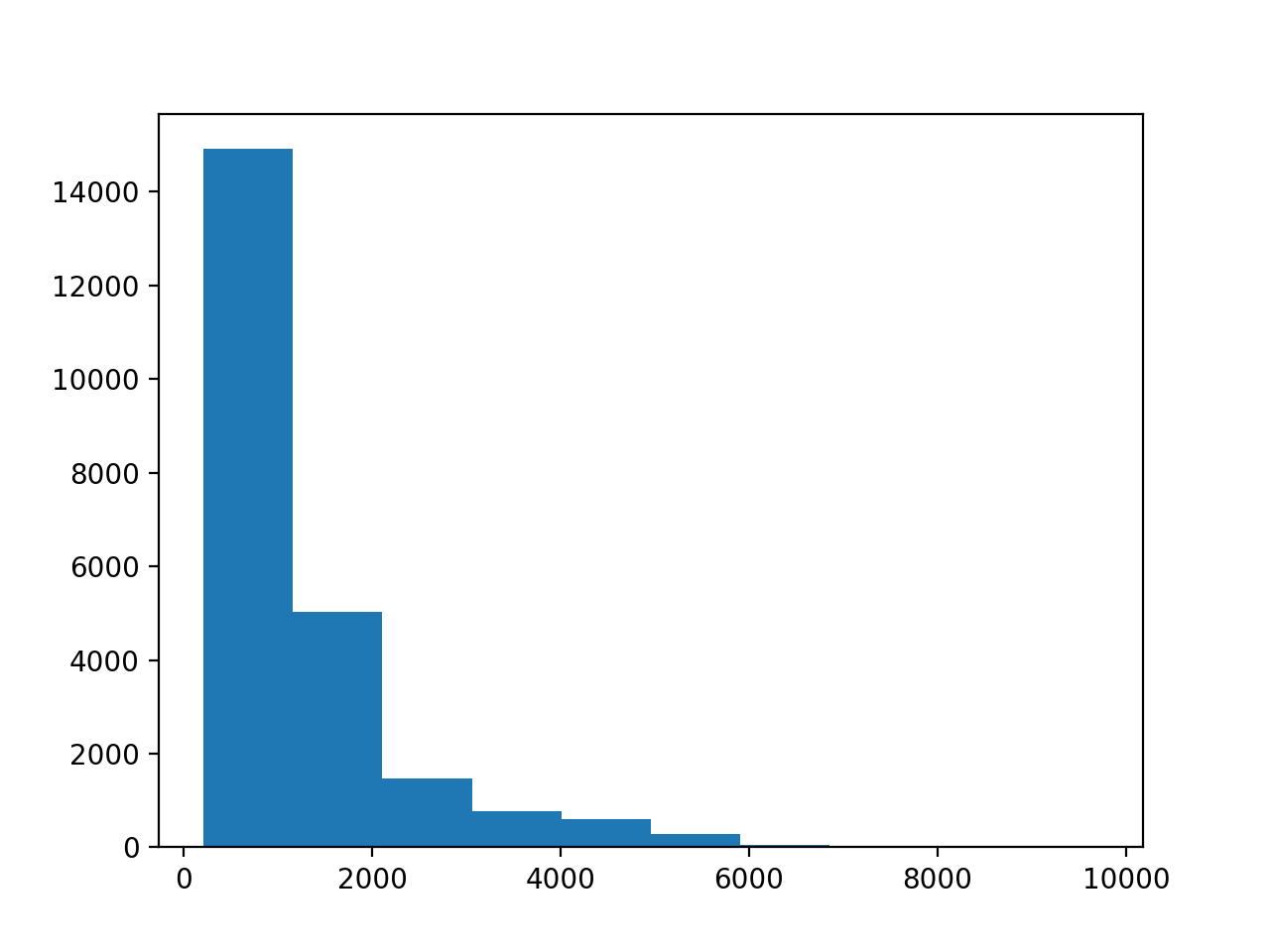
1. Decision tree



MSE = 6766705.10485867

As we can see, the decision tree model reduced our error significantly. We can further improve our result using a forest of decision trees to reduce overfitting.

1. Random forest



MSE = 6618142.8323392

I specified the hyperparameter values ‘min\_samples\_leaf’ and ‘n\_estimators’, I optimized these values by using a for loop. Using 500 trees and 13 min\_samples\_leaf, MSE lower to 6618142.

Machine Learning summary:

Concepts explored: pandas, matplotlib, features engineering, linear regression, decision trees, random forest, MSE.