# Data Exploration Report

**Student name: xinyu zhou**

**Student id: 30199719**

**Group id: Flex 02**

# 1. Introduction

Convenient bike sharing is very popular in megacities. But bike sharing is easily affected by many factors like weather,temperature,humidity,register... These factors can lead to lack or surplus of bike sharing demand. This report uses R and python to analyze and solve three problems.1.Register users and non-register users, which are the major users and how they affect the bike sharing demand of Washington, D.C. 2.What are the causes that affect bike sharing demand of Washington, D.C. 3.Find the correlation of variables. The motivation of this report is to rationalize the distribution of bike sharing resources to make it more convenient for citizens

# 2. Data Wrangling

The data sources includes two data sets, the first data set describes the 2011-2012 bike rental situation in Washington, D.C from the first day to the day 19 of each month. It includes weather data, bike demand,time and user structure. The table has 10887 rows and 12 columns. The second data set is from day 20 to the end of each month. This table is composed of 6494 rows and 9 columns. The link of data sources is https://www.kaggle.com/c/bike-sharing-demand/data

## 2.1 Data cleaning

**1. check data type**

I use python and pandas library to finish data wrangling.This table's data type includes float and int. They can be divided into three categories.

1.Time features: date time,season,holiday,working day

2.Weather features: weather,temp,atemp,humidity

3.Object features: casual,registered,count

**2. drop duplicate value,useless columns,rows and null value**

Every columns will be used in the analysis. And there are some null value in the casual, registered and count of test table. These null value should be filled in by analyst, our aim is just to find relationship between variables. So l can not drop these null value and useless rows or replace them by average value. I choose to make prediction of these null values after data cleaning and data checking.

## 2.2 Data transformations

**1. merge training dataset and test dataset**

To facilitate unified preprocessing of data.

**2. transform coordinate axis**

This table has good coordinate structure.

**3. feature transformation**

Transform quantified features into characters that are easy to identify and visualize. Weather and season has been quantified. And after searching and comparing with Washington, D.C weather and season online,I transform them into characters.

```
In [12]: datasets['season']=datasets['season'].map({1:'spring',2:'summer',3:'autumn','4':'winter'})
         datasets['weather']=datasets['weather'].map({1:'good',2:'normal',3:'bad',4:'awful'})
```

Fig1 feature transformation

**4. feature derive**

Datetime includes year,month,day and hour. We can use regular expression and timestamps to extract these features.

# 3.Data Checking

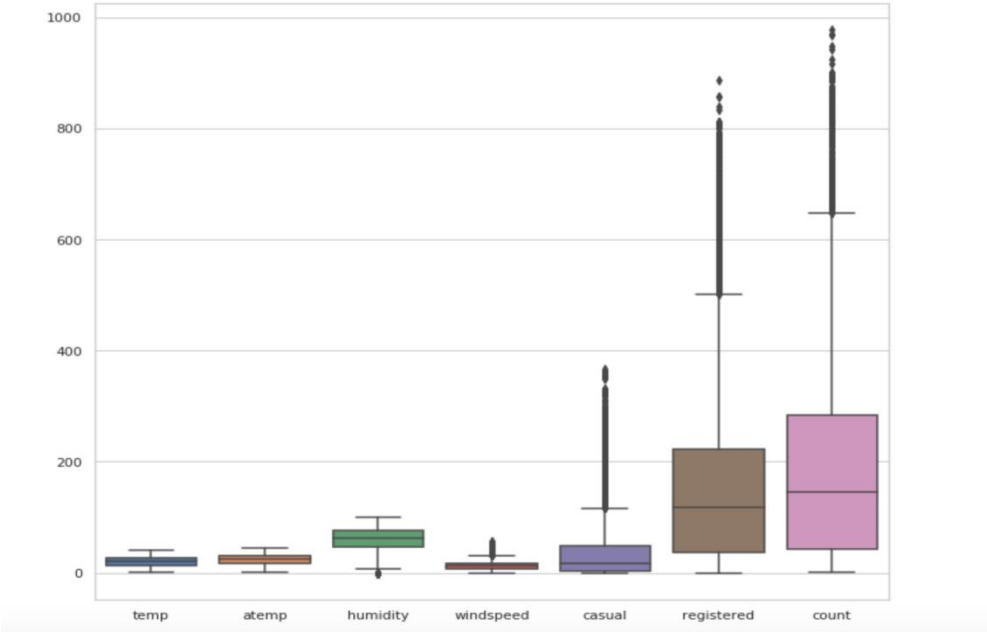
**1. check duplicate value,useless columns,rows and null value**

In this data checking step,I use python and matplotlib. And after data wrangling, there are no duplicate value,useless columns,rows and null value.

**2. check outliers and distribution**

Check outliers and remove them.

Fig2 Boxplot of sharing bike

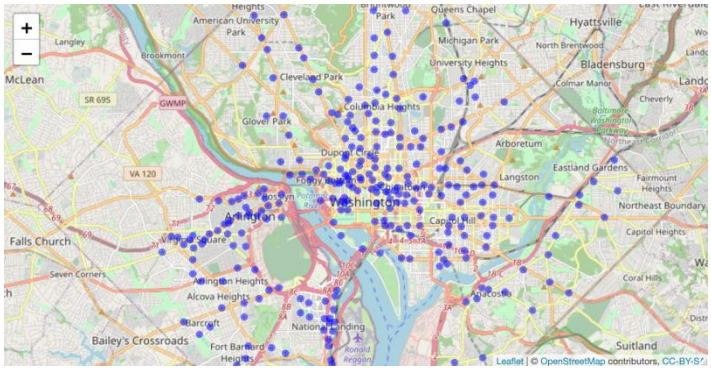


**3. check geographical position**

This table does not include longitude and latitude. I use python crawler to get the geographical position table of Washington, D.C sharing bike and make visualization. The geographical distribution is average and rational.

URL: http://tiny.cc/dcf/DC-Stations.cs

Fig3 bike sharing station map




After data wrangling and checking, Finally, I use Multiple Linear Regression, Support Vector Regression and Random Forest Regression to get the predicted test bike count. Then I get the appropriate table that can be load to R.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | season | holiday | workingda | weather | temp | atemp | humidity | windspeed | casual | registered | count | year | month | day | hour |
| 2 | 0 | 1 | 0 | 0 | 1 | 9.84 | 14.395 | 81 | 0 | 3 | 13 | 16 | 2011 | 1 | 1 | 0 |
| 3 | 1 | 1 | 0 | 0 | 1 | 9.02 | 13.635 | 80 | 0 | 8 | 32 | 40 | 2011 | 1 | 1 | 1 |
| 4 | 2 | 1 | 0 | 0 | 1 | 9.02 | 13.635 | 80 | 0 | 5 | 27 | 32 | 2011 | 1 | 1 | 2 |
| 5 | 3 | 1 | 0 | 0 | 1 | 9.84 | 14.395 | 75 | 0 | 3 | 10 | 13 | 2011 | 1 | 1 | 3 |
| 6 | 4 | 1 | 0 | 0 | 1 | 9.84 | 14.395 | 75 | 0 | 0 | 1 | 1 | 2011 | 1 | 1 | 4 |
| 7 | 5 | 1 | 0 | 0 | 2 | 9.84 | 12.88 | 75 | 6.0032 | 0 | 1 | 1 | 2011 | 1 | 1 | 5 |
| 8 | 6 | 1 | 0 | 0 | 1 | 9.02 | 13.635 | 80 | 0 | 2 | 0 | 2 | 2011 | 1 | 1 | 6 |
| 9 | 7 | 1 | 0 | 0 | 1 | 8.2 | 12.88 | 86 | 0 | 1 | 2 | 3 | 2011 | 1 | 1 | 7 |
| 10 | 8 | 1 | 0 | 0 | 1 | 9.84 | 14.395 | 75 | 0 | 1 | 7 | 8 | 2011 | 1 | 1 | 8 |
| 11 | 9 | 1 | 0 | 0 | 1 | 13.12 | 17.425 | 76 | 0 | 8 | 6 | 14 | 2011 | 1 | 1 | 9 |
| 12 | 10 | 1 | 0 | 0 | 1 | 15.58 | 19.695 | 76 | 16.9979 | 12 | 24 | 36 | 2011 | 1 | 1 | 10 |
| 13 | 11 | 1 | 0 | 0 | 1 | 14.76 | 16.665 | 81 | 19.0012 | 26 | 30 | 56 | 2011 | 1 | 1 | 11 |
| 14 | 12 | 1 | 0 | 0 | 1 | 17.22 | 21.21 | 77 | 19.0012 | 29 | 55 | 84 | 2011 | 1 | 1 | 12 |

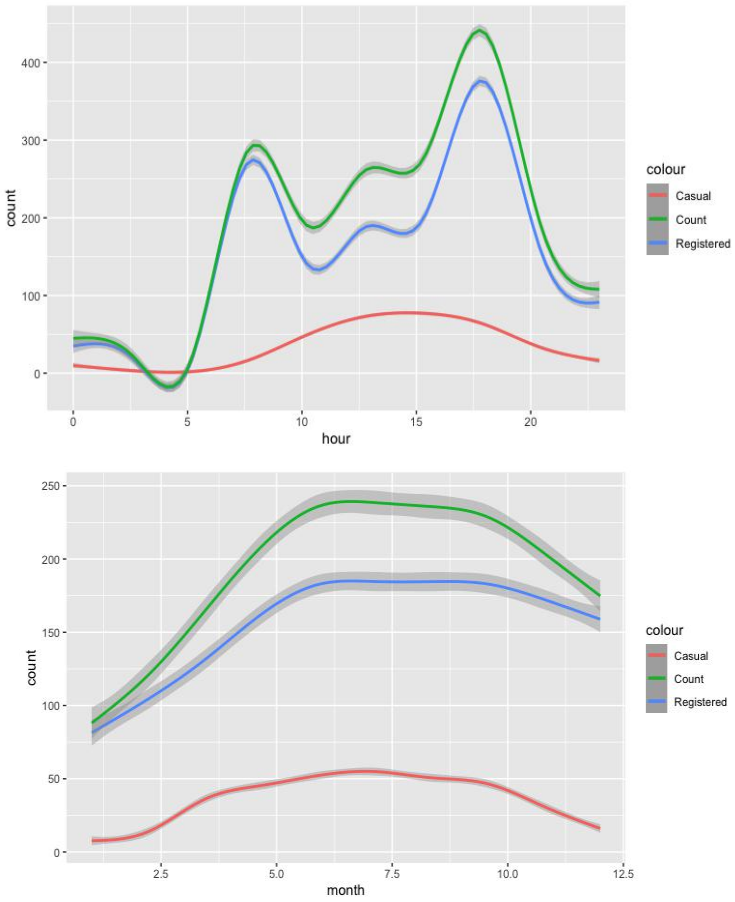
Fig4 Table after preprocessing

# 4. Data exploration

Q1: Register users and non-register users, which are the major users and how they affect the bike sharing demand of Washington, D.C.

In this part, I use R for data exploration.

Fig5 Casual and registered demand(hour)




**Casual VS Registered**

Fig5 and 6 show the liner relation function of the parameters. The curve of registered user and entire user have similar trend. It means most users are registered user and trend of casual users is different from other users.In fig5,there are two peaks in 7am and 18pm approximately. Because these two time ranges are commuter time and registered users ride bikes chronically. Casual users choose other tools at the commuter time. Fig6 shows bike sharing demand every month. Casual and registered have similar trend and figure shows in summer, demand is the highest and in winter, it reaches a low point. Fig7 and 8 are the working day and holiday demand distributions of casual and registered use. It shows in the working day, registered users choose bike as tools. And in holiday, casual and registered have similar trends. Because more causal users choose sharing bike as tools in holiday. In holiday, most users begin to ride bikes from afternoon. It is different from working day.

Fig6 Casual and registered demand(month)

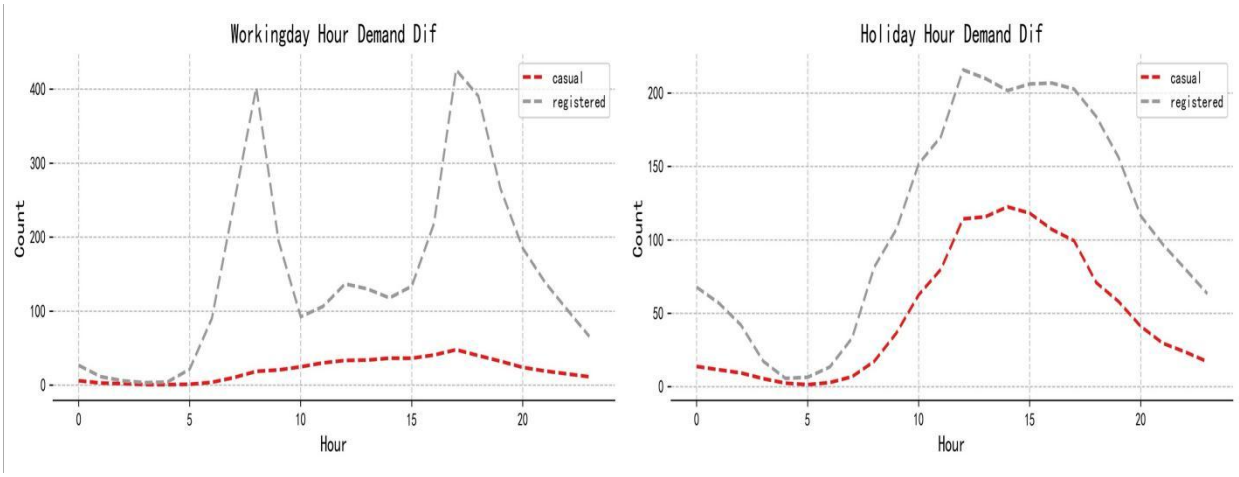



Fig7 and 8 working day and holiday demand distribution of casual and registered user

Q2.What are the causes that affect bike sharing demand of Washington, D.C?

**1.Weather**

Fig9 is the Polar area diagram of weather. Fig10 is the 3D chart of bike sharing demand in different seasons. Figure 9 shows that the weather is mostly good or normal from 2011 to 2012. Figure 10 shows that only a few people ride bikes in the awful and bad weather. In conclusion, bad weather reduces the use of sharing bicycle demand.

Fig10 3D chart of bike sharing demand in different seasons

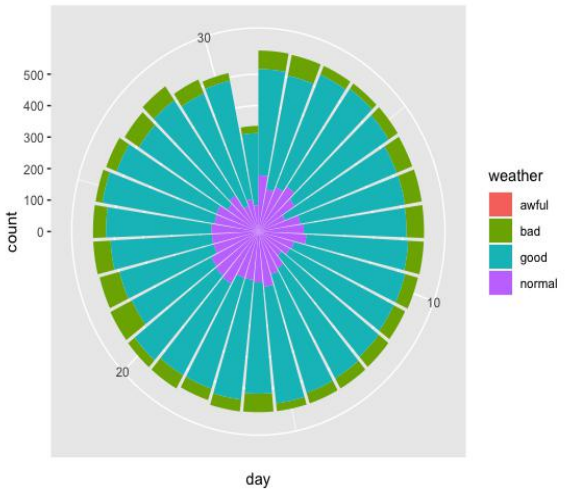
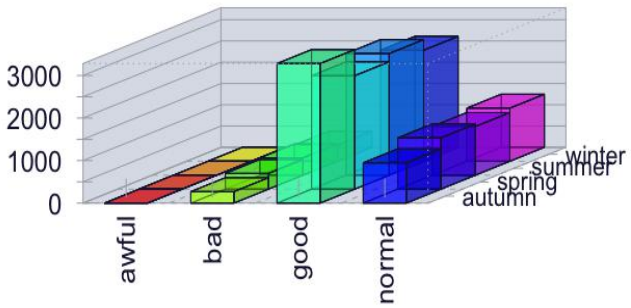



Fig9 Polar area diagram of weather




**2.Month**

Fig11 is the scatter plot of bike sharing demand in different month. From the figure, we can find that There are higher demand for use in the second half of the year than the first half.

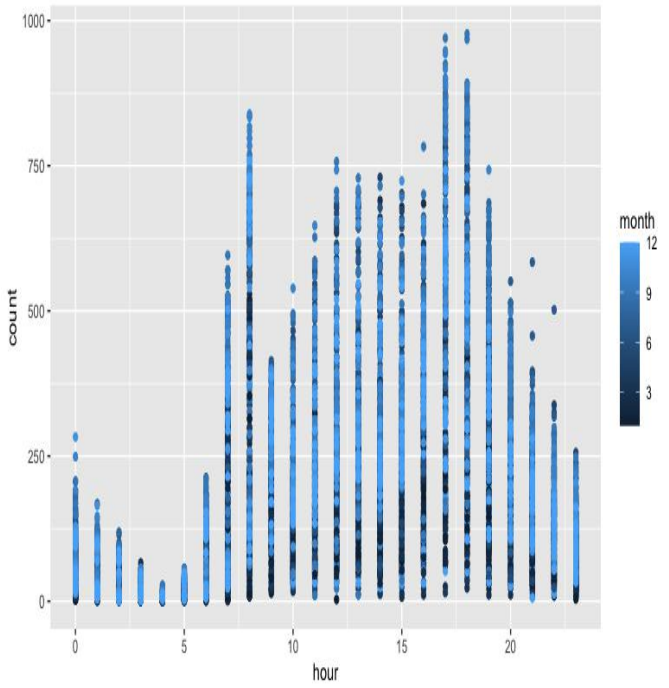



Fig11 scatter plot of bike sharing demand in different month

### 3.Windspeed, temperature and humidity

Fig12 shows how temperature affects bike sharing demand. The most suitable temperature is 20-30 ℃ and when temperature is too low,the demand is very small. Fig13 is about the relationship between temp,windspeed and count. From the figure, we find that people like to travel in low wind speed(10-25) and moderate temperature(20-30). Fig14 is the scatter plot that shows When humidity is 20-60, temperature is 20-30℃, demand is increasing.

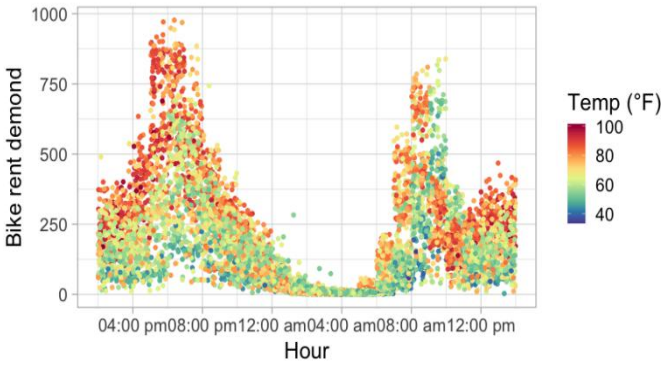



Fig12 scatter plot of how temp

Fig13 scatter plot(temp,humidity,count) affects demand

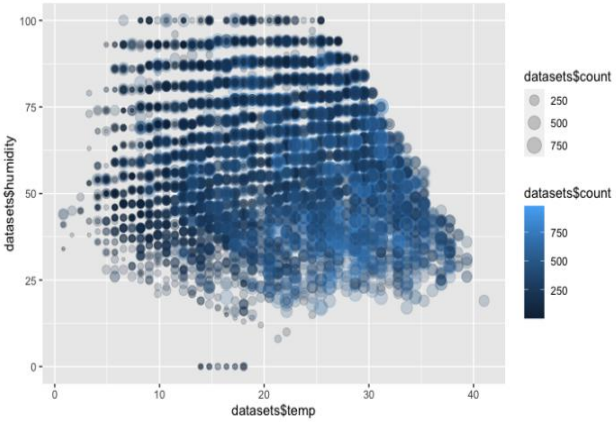



Fig14 scatter plot(wind speed,cont,temp)

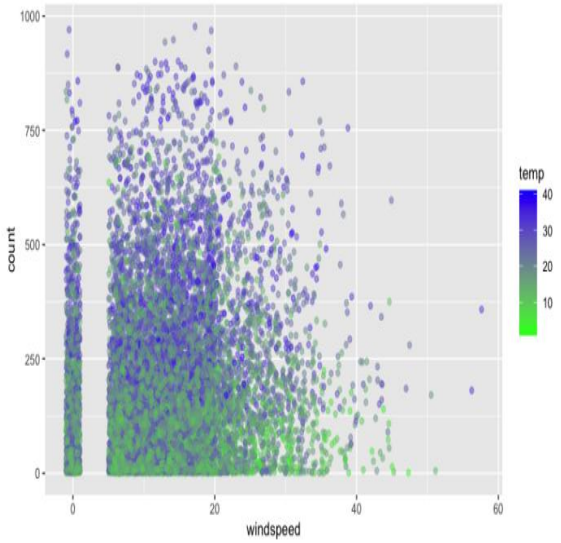
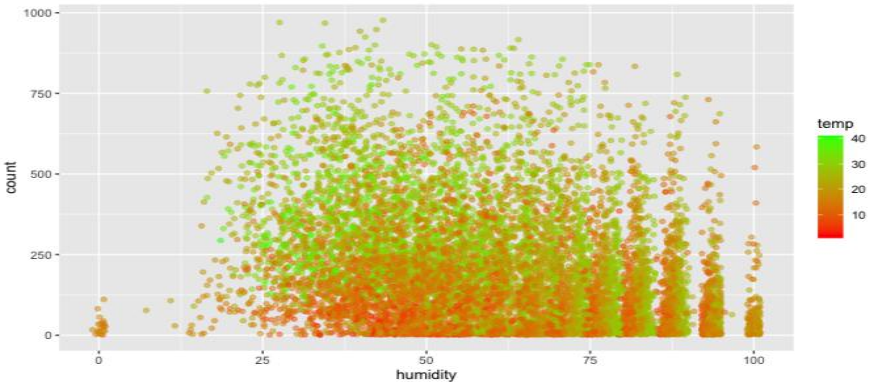



Fig15 is about the relationship between temp, humidity and count. From the figure, we can see that when temp is 20-30℃ and humidity is 40-80, demand is increasing. In conclusion,high and low wind speed, temperature and humidity will decrease the sharing bike demand. The suitable interval is temperature:20-30℃, humidity:40-60, wind speed: 10-25.

Fig15 scatter plot that how humidity and temp affect demand



## Q3. Find the correlation of variables

I choose corrplot to represent their correlation.

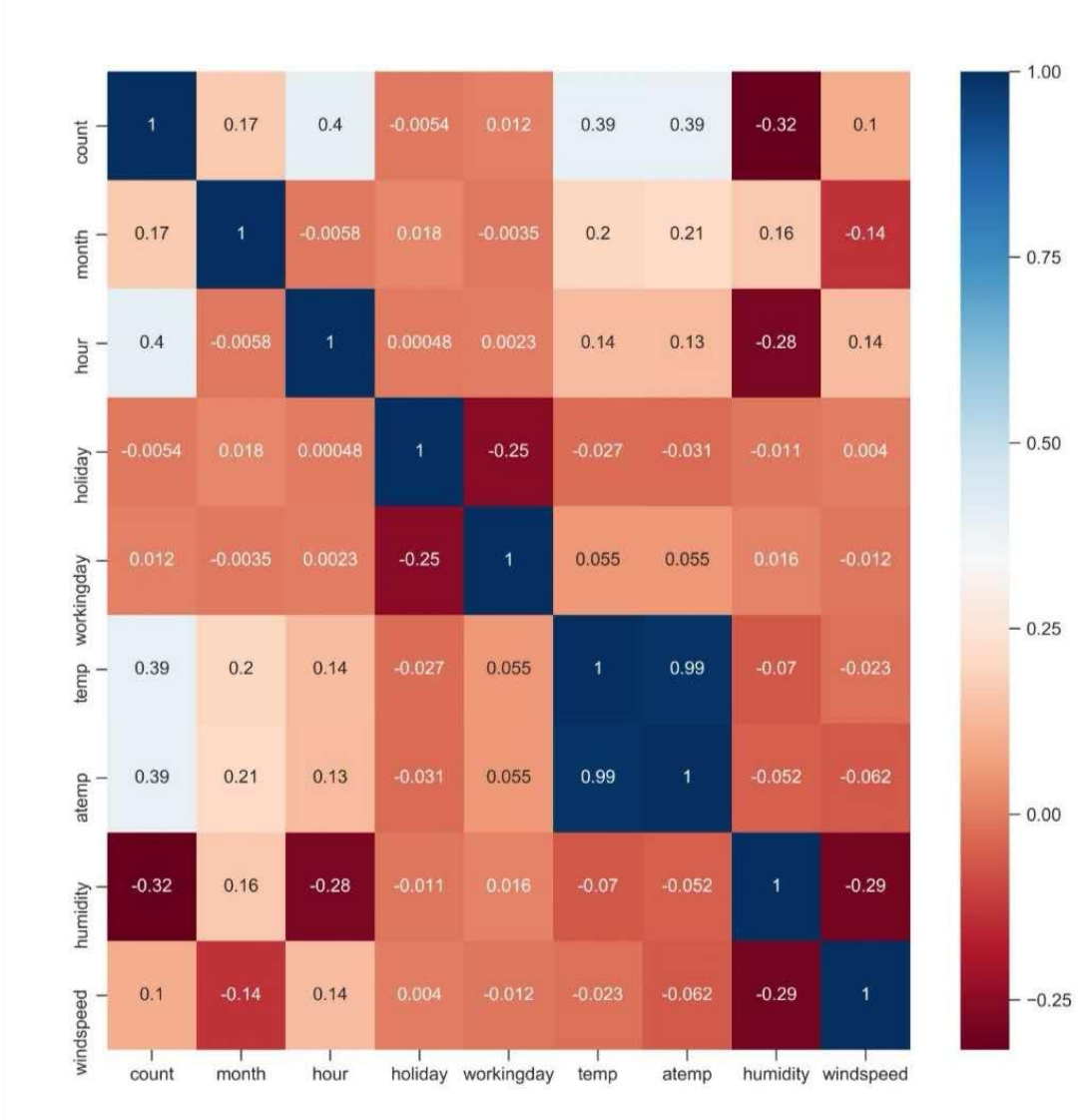



Fig16 variable correlation

In conclusion, correlation coefficient between hour and count is he maximum. The temperature is the second. And humidity has the smallest correlation coefficient.

## 5. Conclusion

In conclusion, registered users make up the majority of bike sharing users. Registered users have their fixed mode of action.They choose bike shring as tools at the commuter time. Compared with working day, sharing bike is more popular for casual users in holiday. So commuter time of working day and afternoon of holiday are two main periods. high and low Wind speed, temperature and humidity will decrease the bike sharing demand. The suitable interval is temperature:20-30℃, humidity: 40-60,wind speed: 10-25. The hour and temperature have the greatest effect on

demand. So company can decide if bike count need to be increased under certain circumstances to make it more convenient for citizen.

## 6. Reflection

In this project,I learned how to make 3D chart and apply different types of graphs in different situations. In hindsight I might have chosen different analysis method and algorithm.

## 7. Bibliography

Kaggle(2015) bike sharing demand
https://www.kaggle.com/c/bike-sharing-demand/data

DC-Station(2015) Washington, D.C bike sharing station
http://tiny.cc/dcf/DC-Stations.cs