Bike-share analysis

Luyang Shang

2/27/2022

Introduction

Bike sharing system is a shared transport service in which bicycles are made available for shared use to individuals, recently bike-sharing companies have become one of the hottest tech companies in the world especially in china, there are over 500 bike-sharing programs around the world.

The data about bike sharing is generated from a bike-sharing app in the Washington D.C, these apps generate a large amount of data on a daily basis and are important for studying the customer travel demand and their local road system. It includes following information:

- datetime - hourly date + timestamp
- season - 1 = spring, 2 = summer, 3 = fall, 4 = winter
- holiday - whether the day is considered a holiday
- workingday - whether the day is neither a weekend nor holiday
- weather
  1: Clear, Few clouds, Partly cloudy, Partly cloudy
  2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp - temperature in Celsius
- atemp - temperature in Celsius
- humidity - relative humidity
- windspeed - wind speed
- casual - number of non-registered user rentals initiated
- registered - number of registered user rentals initiated
- count - number of total rentals

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

I will make a deep analysis on bike sharing data to investigate which factors contribute to bike rental demand.

```
head(bike)
```

```
##              datetime season holiday workingday weather temp  atemp humidity
## 1 2011-01-01 00:00:00      1       0          0       1 9.84 14.395       81
## 2 2011-01-01 01:00:00      1       0          0       1 9.02 13.635       80
## 3 2011-01-01 02:00:00      1       0          0       1 9.02 13.635       80
## 4 2011-01-01 03:00:00      1       0          0       1 9.84 14.395       75
## 5 2011-01-01 04:00:00      1       0          0       1 9.84 14.395       75
## 6 2011-01-01 05:00:00      1       0          0       2 9.84 12.880       75
##   windspeed casual registered count
## 1    0.0000      3         13    16
## 2    0.0000      8         32    40
## 3    0.0000      5         27    32
## 4    0.0000      3         10    13
## 5    0.0000      0          1     1
## 6    6.0032      0          1     1
```

Methods

The data is retrieved from Kaggle, a famous data competition platform. You could observe the data from the following url: https://www.kaggle.com/c/bike-sharing-demand/data. Each row records relevant information within a 1 hour period. The data is pretty clear, it does not contain any missing value. To make the visualization more readable, I change the season column and weather column.

To make the visualization more readable, I change the season column and weather column. Meanwhile, to analyze customer riding time pattern, I extract month, day and hour from variable datetime.

```
bike$season <- recode(bike$season, '1'='spring', '2'='summer', '3'='fall', '4'='winter')
bike$weather <- recode(bike$weather, '1'='clear', '2'='cloudy', '3'='drizzle', '4'='rainstorm')
bike$datetime <- as.POSIXct(bike$datetime, format="%Y-%m-%d %H:%M:%S", tz="UTC")
bike$month <- months(bike$datetime)
bike$weekday <- weekdays(bike$datetime)
bike$hour <- format(bike$datetime, format = "%H")
bike$day <- format(bike$datetime, format = "%d")
```

I will use a R built-in package called ggplot to make visualization, and build a linear regression model to explore the relationship between multiple variable and customer rental demands

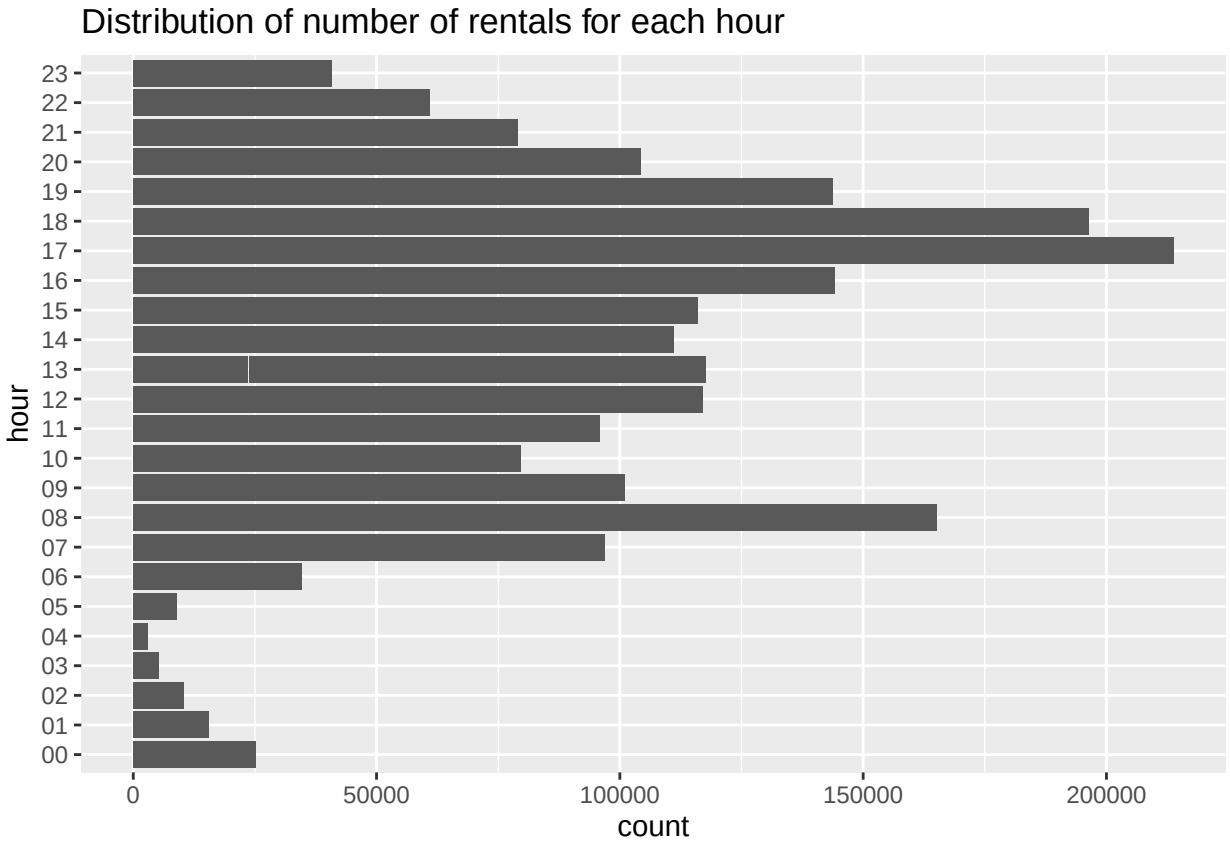
Preliminary

Let's start to explore which factors contribute to bicycle rental demands. The following plot shows summary statistics for the number of total rentals in each hour period. We could observe that on average 191.6 people rent a bike each hour. However, the maximum number of customers for each hour is 977, while the minimum is only 1. This suggests time might be a factor contributing to the amount of rentals.

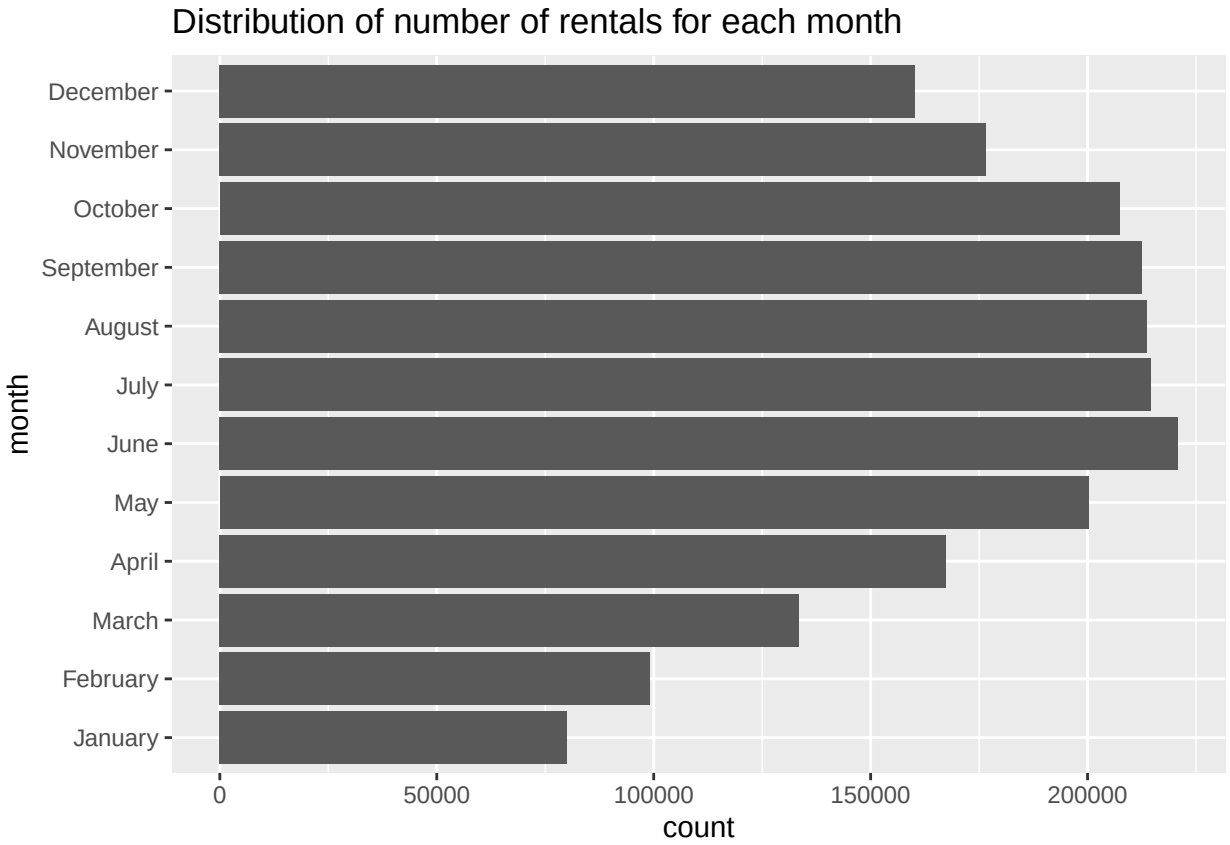
Table 1: Number of total rentals

mean	median	min	max
191.6	145	1	977

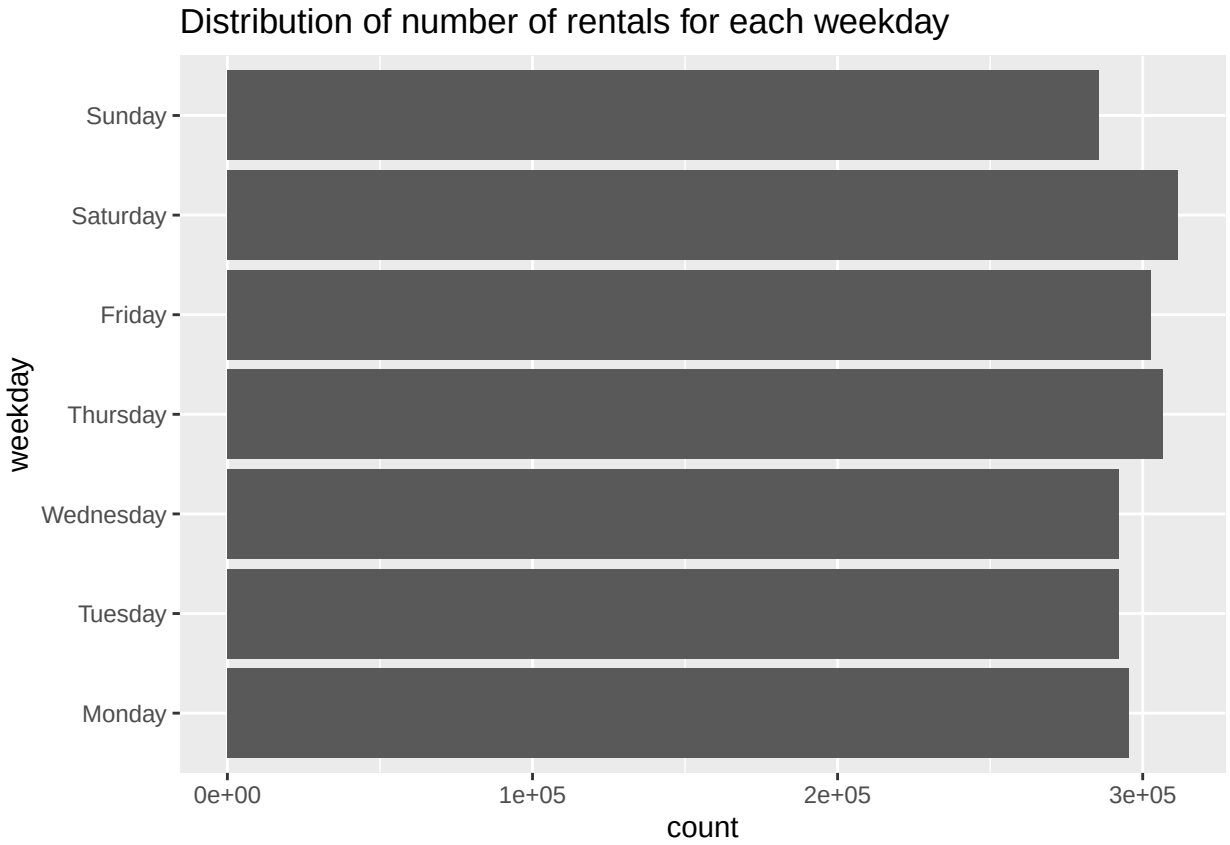
The following bar plot shows the distribution of the number of rentals for each hour. There is a huge fluctuation among the number of riders per day, ranging from less than 5000 to more than 200000. It is clear that most customers rent bicycles during day time, meanwhile, There are two peaks each day on commuting hours.



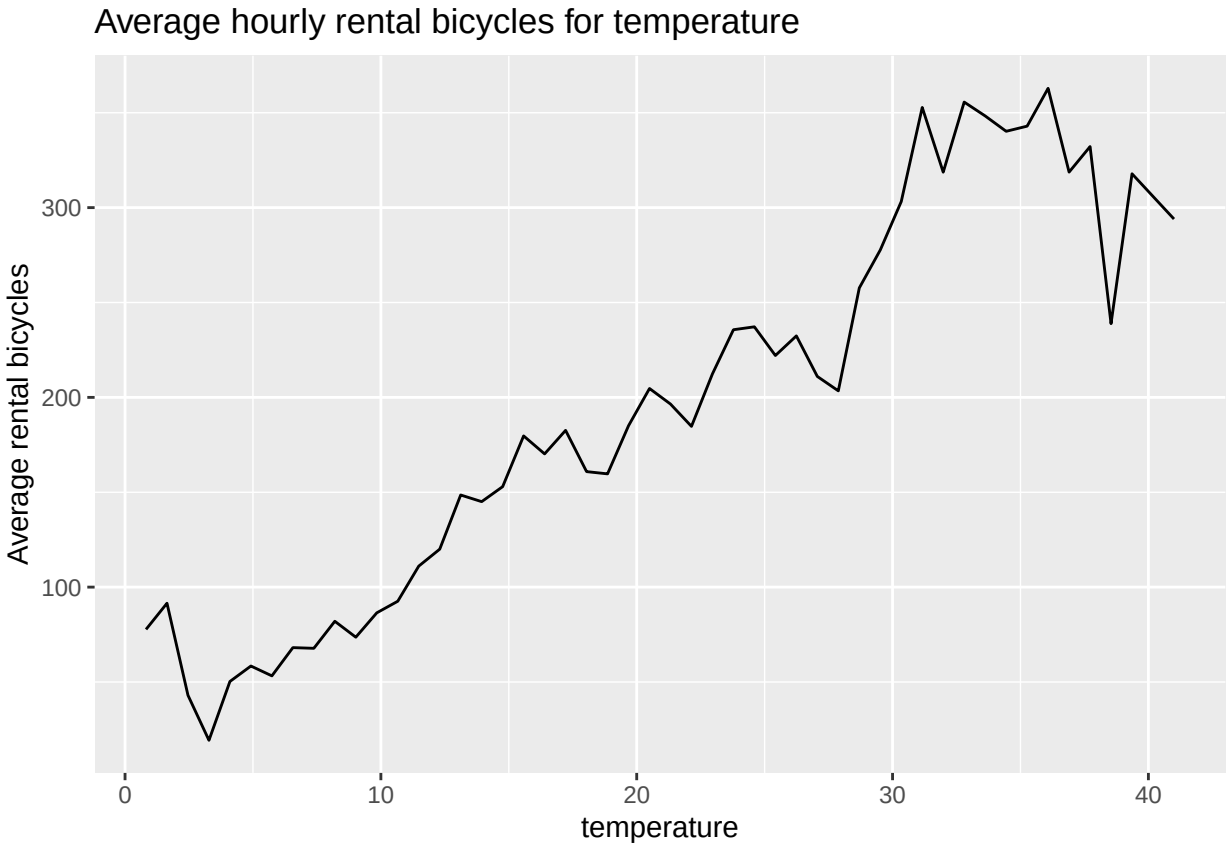
In terms of the month dimension, the demand for orders appears in an annual cycle: the demand for bikes starts to rise in the spring of each year, reaches and remains at a peak level in the summer and fall, and then starts to fall back in the winter, which suggests the daily temperature and weather could be potential factors contributing to demand for rental.



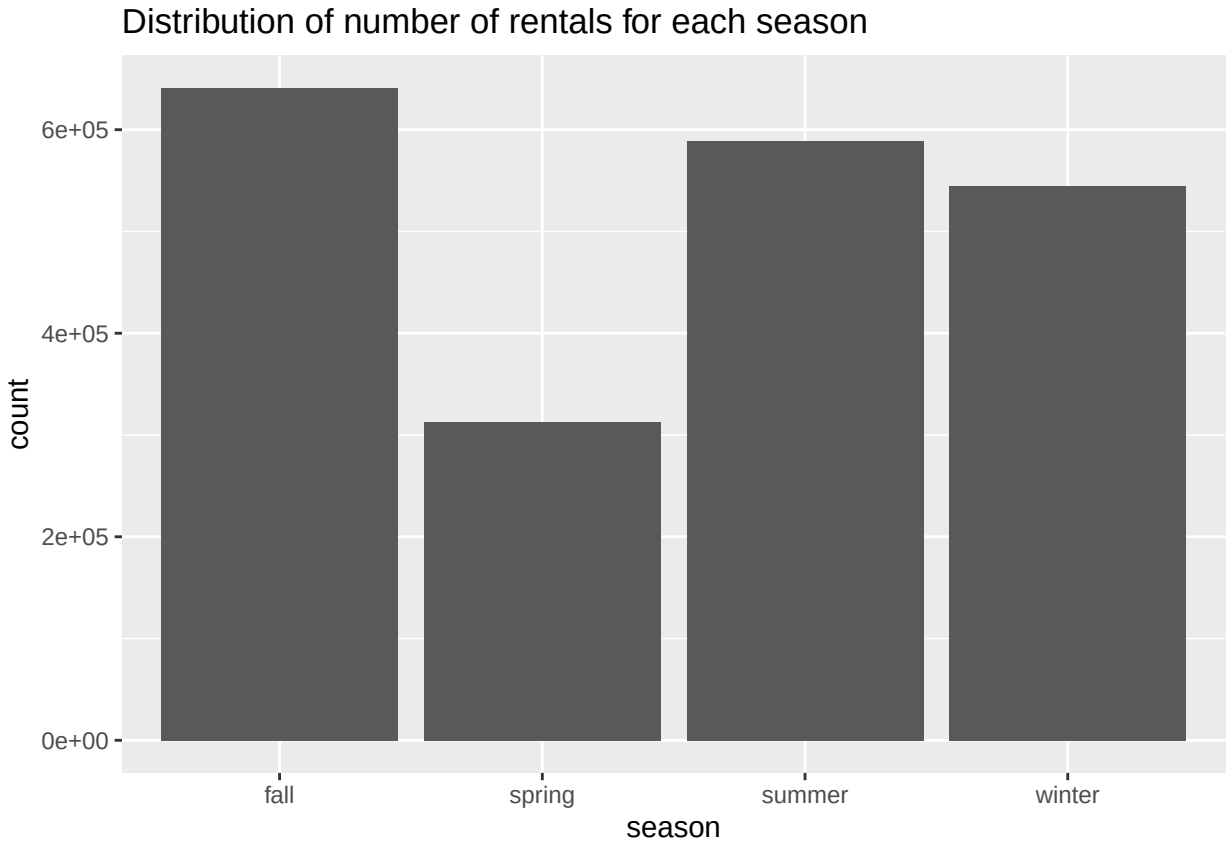
The following plot shows the distribution of the number of rentals for each weekday. Surprisingly, there is little fluctuation among weekdays, which indicates most customers not only use the bicycle for transportation to work, but also rent bicycles for daily commuting.



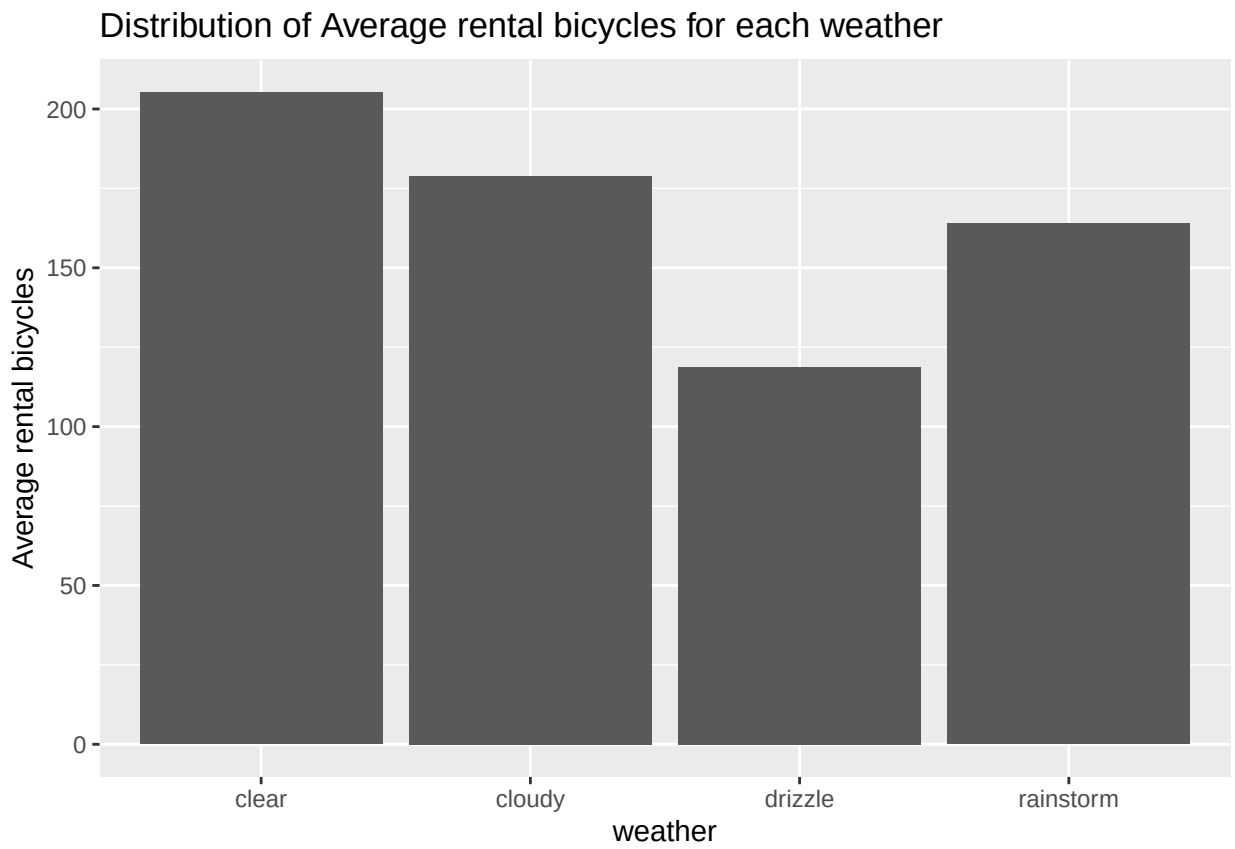
Previous analysis suggests temperature and weather could be a potential factor for bicycles rental demands, the following plot shows a strong linear relationship between number of rental bicycles and temperature, as the temperature increases, the demands for bicycles rentals increases, but when temperatures is above 36 degrees, the situation changes a bit. The trend begins to move in the opposite direction.



The following plot verify my previous assumption that from spring to autumn, the order volume is in a growing trend, and starts to fall after reaching the peak in autumn, and reaches a low point in spring.



The following bar plot shows distribution of Average rental bicycles for each weather, the plot indicates the demand for renting bicycles is highest when weather is clear, however lots of people choose to rent bicycles during rainstorm, we need to analyse the data further to find reason for that



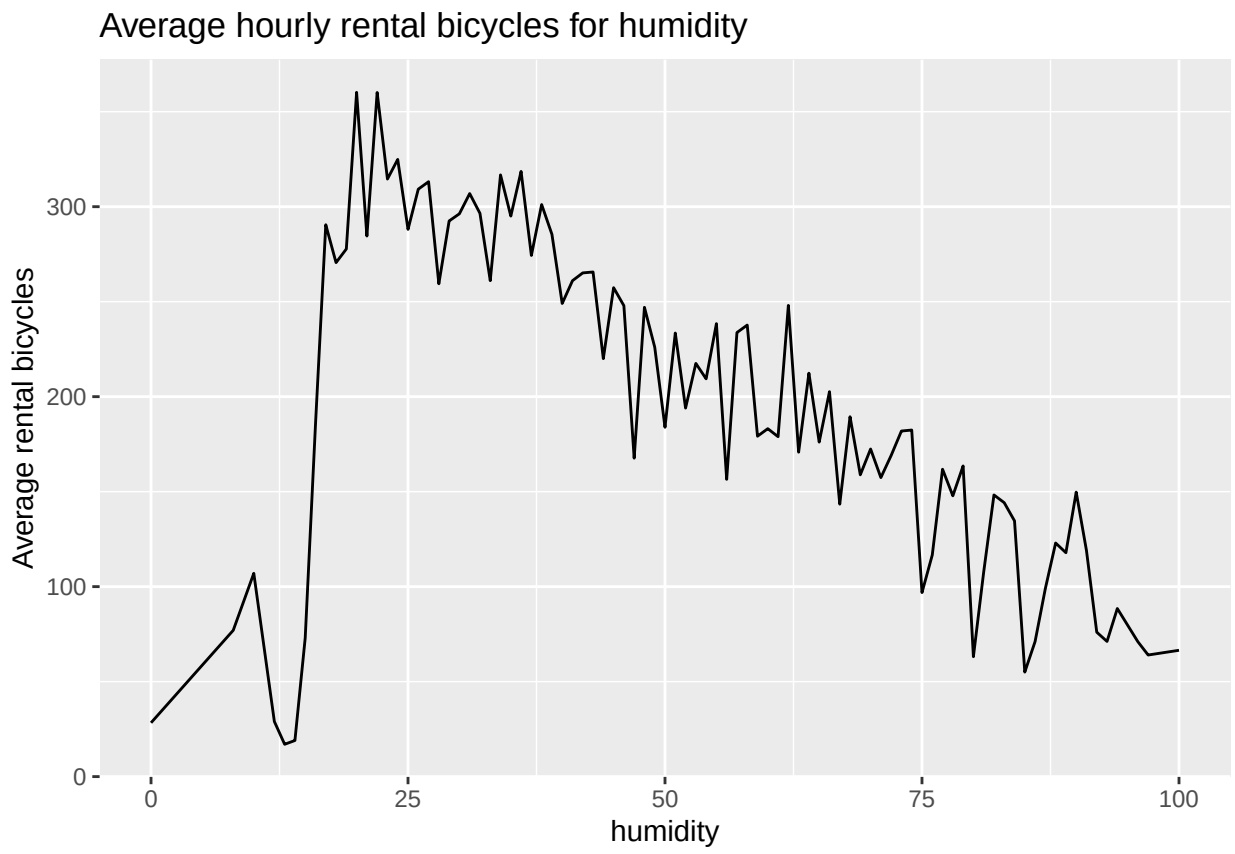
Considering the low probability of extreme weather, it is very likely that the demand is not representative of the real demand in rainstorm. Combining our common sense with other available data, we can assume that the worse the weather, the lower the demand for bicycles.

Table 2: Weather Distribution

weather	n
clear	7192
cloudy	2834
drizzle	859
rainstorm	1

The following plot suggests humidity is one of factors influencing bicycle rental demand, however it is possible that humidity has a strong correlation with weather. However, higher humidity indicates a higher possibility of raining, so humidity could be a confounding variable.

```
bike %>%
  group_by(humidity) %>%
  summarise(avg_bike = mean(count)) %>%
  ggplot(mapping=aes(x = humidity, y=avg_bike)) +
  geom_line() +
  ggtitle("Average hourly rental bicycles for humidity") +
  xlab("humidity") +
  ylab("Average rental bicycles")
```




```
humidity_data <- bike %>%
  group_by(weather) %>%
  summarise(humidity_mean = mean(humidity))

knitr::kable(humidity_data, caption="Mean humidity for each weather")
```

Table 3: Mean humidity for each weather

weather	humidity_mean
clear	56.71677
cloudy	69.10056
drizzle	81.34109
rainstorm	86.00000

Then I build a linear regression by considering temperature, hour, season and weather, the p value for each variable is less than the significance level (0.05), indicating all those variables are statistically significant. Meanwhile, the r-square for this model is 0.6252, indicating the model has a relatively good performance

```
## Loading required package: nlme
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse

## This is mgcv 1.8-38. For overview type 'help("mgcv-package")'.

##
## Call:
## lm(formula = count ~ temp + as.factor(hour) + as.factor(season) +
##     as.factor(weather), data = bike)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -385.12  -61.91   -9.62   51.47  516.99
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -78.0234     8.7413  -8.926  < 2e-16 ***
## temp                            6.7296     0.2441  27.565  < 2e-16 ***
## as.factor(hour)01             -16.9342     7.3577  -2.302 0.021379 *
## as.factor(hour)02             -28.1739     7.3831  -3.816 0.000136 ***
## as.factor(hour)03             -39.7510     7.4488  -5.337 9.66e-08 ***
## as.factor(hour)04             -41.1506     7.4126  -5.551 2.90e-08 ***
## as.factor(hour)05             -24.9687     7.3735  -3.386 0.000711 ***
## as.factor(hour)06              34.0875     7.3629   4.630 3.71e-06 ***
## as.factor(hour)07             169.4943     7.3600  23.029  < 2e-16 ***
## as.factor(hour)08             313.6667     7.3548  42.648  < 2e-16 ***
## as.factor(hour)09             166.1914     7.3551  22.595  < 2e-16 ***
## as.factor(hour)10             112.6986     7.3606  15.311  < 2e-16 ***
## as.factor(hour)11             140.8423     7.3741  19.100  < 2e-16 ***
## as.factor(hour)12             182.2048     7.3860  24.669  < 2e-16 ***
## as.factor(hour)13             178.2700     7.4042  24.077  < 2e-16 ***
## as.factor(hour)14             161.6493     7.4186  21.790  < 2e-16 ***
## as.factor(hour)15             172.5136     7.4246  23.235  < 2e-16 ***
## as.factor(hour)16             234.8845     7.4195  31.658  < 2e-16 ***
## as.factor(hour)17             392.0016     7.4070  52.923  < 2e-16 ***
## as.factor(hour)18             357.9655     7.3954  48.404  < 2e-16 ***
## as.factor(hour)19             244.8855     7.3730  33.214  < 2e-16 ***
## as.factor(hour)20             162.4264     7.3635  22.058  < 2e-16 ***
## as.factor(hour)21             110.3722     7.3552  15.006  < 2e-16 ***
## as.factor(hour)22              73.6974     7.3517  10.024  < 2e-16 ***
## as.factor(hour)23              34.3459     7.3508   4.672 3.01e-06 ***
## as.factor(season)spring        -9.7699     4.9899  -1.958 0.050262 .
## as.factor(season)summer        22.4354     3.3356   6.726 1.83e-11 ***
## as.factor(season)winter        48.4365     4.2189  11.481  < 2e-16 ***
## as.factor(weather)cloudy      -17.2154     2.4744  -6.957 3.66e-12 ***
## as.factor(weather)drizzle     -88.8342     4.0111 -22.147  < 2e-16 ***
## as.factor(weather)rainstorm -161.3552   111.0510  -1.453 0.146258
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 110.9 on 10855 degrees of freedom
## Multiple R-squared:  0.6262, Adjusted R-squared:  0.6252
## F-statistic: 606.2 on 30 and 10855 DF,  p-value: < 2.2e-16
```

Conclusion

This data analysis aim to find the factors contribute to number of bicycle rental. We find both temperature, season, time and weather are key factors influencing bicycle rental demand. We found customer tent to rent bike when temperature is relatively high, but they avoid extremely hot weather. Meanwhile, bicycle use on weekdays is concentrated between 7-10 a.m. and 18-20 p.m., coinciding with traffic rush hour. For season, we found the demand for bikes starts to rise in the spring of each year, reaches and remains at a peak level in the summer and fall, and then starts to fall back in the winter, this founding coincide with the founding for temperature. For the weather, most customer avoid using bike in raining day, and they tend to ride bike when weather is clear.