STAT 471 Final Project Report

Tongchen He

2022-12-07

Section 1: Introduction

In this project, we use R to analyze the Bike Sharing Dataset from UCI Machine Learning Repository. The dataset contains the hourly and daily count of rental bikes between 2011 and 2012 in Capital bike-sharing system with the corresponding weather and seasonal information. Generally, we are curious about whether the company saw a growth in their user-base, what are the busiest hours of a day, and how the weather factors affect the rental bikes count?

Section 2: Questions of Interest

After looking at the dataset, we are interested in answering the following questions:

1. Is there a general growth in bike rentals from 2011 to 2012? Specifically for casual or registered users?

2. What hour of a day is the most popular time for bike rental? Specifically for casual and registered user?

3. How does temperature, based on the current weather situation, affect the number of bike rentals?

Section 3: Analysis

We have cleaned the datasets to make it more understandable and easier to use, so now we can take a look at the cleaned daily count dataset:

```
##    instant     dteday season   yr mnth holiday weekday workingday weathersit
## 1        1 2011-01-01      1 2011    1       0       6          0          2
## 2        2 2011-01-02      1 2011    1       0       0          0          2
## 3        3 2011-01-03      1 2011    1       0       1          1          1
## 4        4 2011-01-04      1 2011    1       0       2          1          1
## 5        5 2011-01-05      1 2011    1       0       3          1          1
## 6        6 2011-01-06      1 2011    1       0       4          1          1
## 7        7 2011-01-07      1 2011    1       0       5          1          2
## 8        8 2011-01-08      1 2011    1       0       6          0          2
## 9        9 2011-01-09      1 2011    1       0       0          0          1
## 10      10 2011-01-10      1 2011    1       0       1          1          1
##        temp    atemp      hum windspeed casual registered  cnt
## 1  0.344167 0.363625 0.805833 0.1604460    331        654  985
## 2  0.363478 0.353739 0.696087 0.2485390    131        670  801
## 3  0.196364 0.189405 0.437273 0.2483090    120       1229 1349
```

```
## 4  0.200000 0.212122 0.590435 0.1602960    108       1454 1562
## 5  0.226957 0.229270 0.436957 0.1869000     82       1518 1600
## 6  0.204348 0.233209 0.518261 0.0895652     88       1518 1606
## 7  0.196522 0.208839 0.498696 0.1687260    148       1362 1510
## 8  0.165000 0.162254 0.535833 0.2668040     68        891  959
## 9  0.138333 0.116175 0.434167 0.3619500     54        768  822
## 10 0.150833 0.150888 0.482917 0.2232670     41       1280 1321
```

The `instant` variable is the index of each record.

In `holiday` and `workingday`, if the day is a holiday/working day the value will be 1, otherwise it will be 0.

The `weathersit` column is the weather situation of the day: 1 means Clear, Few clouds, or Partly cloudy; 2 means Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, or Mist; 3 means Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds; 4 means Heavy Rain + Ice Pallets + Thunderstorm + Mist, or Snow + Fog.

Columns `temp`, `atemp`, `hum`, `windspeed` represents the temperature (°C), feeling temperature (°C), humidity, and windspeed, with all values normalized. The rest of the columns are self-explanatory.

The hourly count dataset is mostly the same except the records are specified by each hour instead of day.

Question 1: Is there a general growth in bike rentals from 2011 to 2012?

For the company, it might be interesting to evaluate their financial situation by looking at the growth in their user-base from 2011 to 2012.
First, we will calculate the mean of customers each day for the two years respectively:

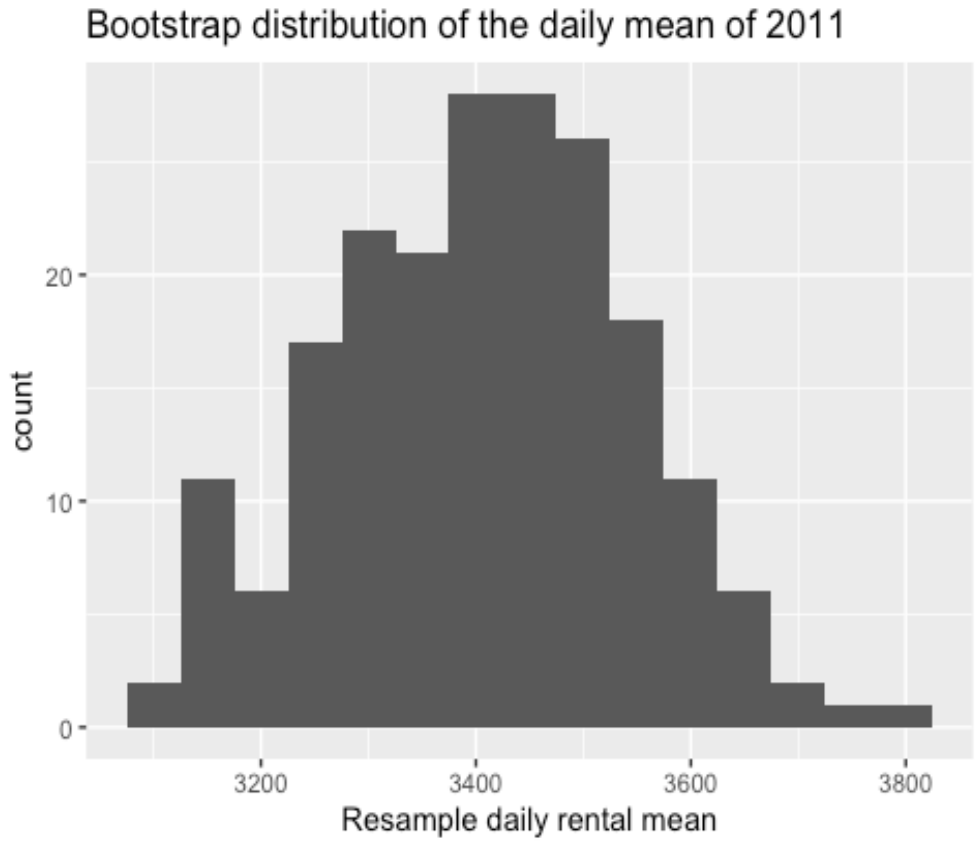
```
## # A tibble: 2 × 2
##   yr    daily_average
##   <fct>         <dbl>
## 1 2011          3406.
## 2 2012          5600.
```

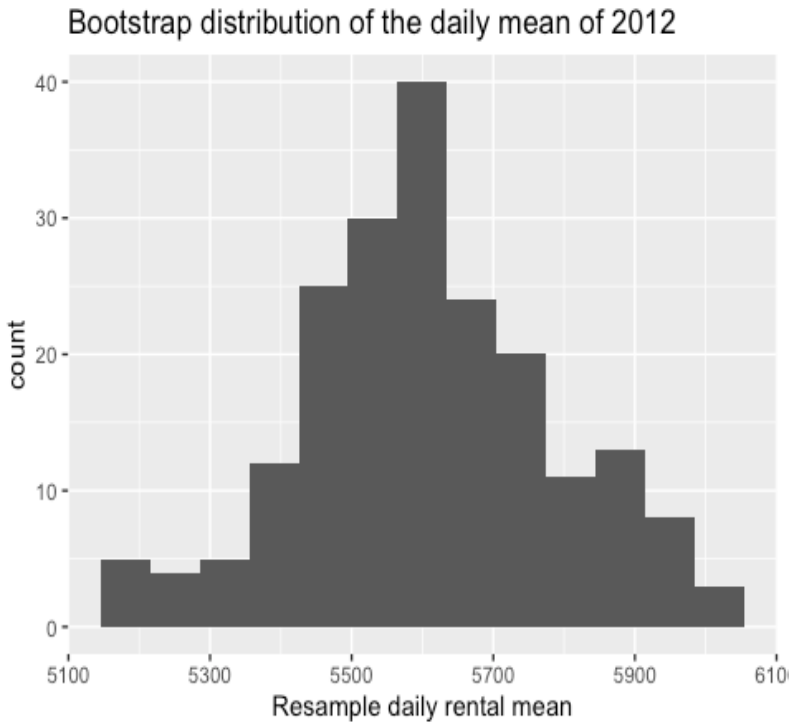
From the table above, we can see that the daily average was 3406 in 2011 and 5600 in 2012. It is seemingly apparent that there is a growth of bike rentals between the years, but just to make sure, we want to check if there is a significant statistical difference between

the daily menas of the two years. Therefore, we want to conduct the following hypothesis test:

$H_0: \mu_{2011} = \mu_{2012}$ versus $H_1: \mu_{2011} \neq \mu_{2012}$, where $\mu_{2011,2012}$ represents the mean of daily bike rentals of 2011 and 2012.

First, we use bootstrapping to resample the data of each year with size of 100 and 200 replicates. Then, we plotted the mean of each resampled data on a histogram:





Now, we conduct a two-sample t-test on the difference in sample mean of 2011 and 2012, and calculated the test statistic t:

```
##         t
## -138.3729
```

The p value turns out to be 2^{-323}. It is much lower than the significance value $\alpha = 0.05$, and therefore we can reject the null hypothesis and conclude that there is a significant difference between the average daily bike rental in 2011 and 2012.

We can also calculate the confidence interval. Since the distribution of daily mean in 2011 is slightly skewed, we need to use the standard error method to calculate the confidence interval, which turns out to be the following. And because the distribution of 2012 is mostly symmetric, we will use the percentile method to calculate the confidence interval.

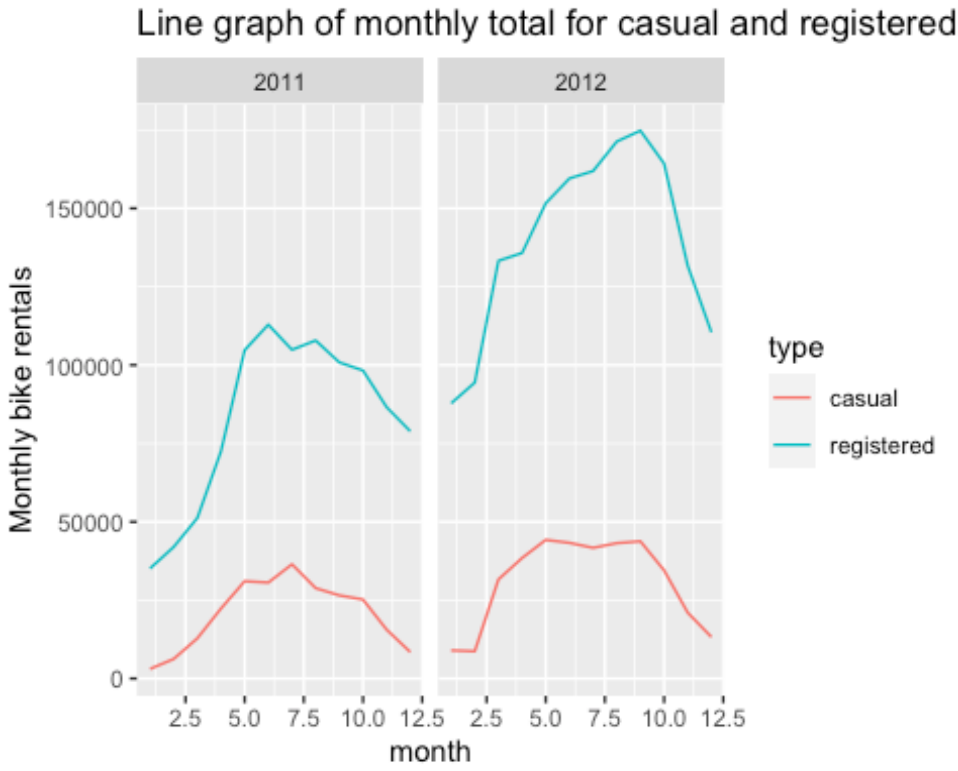
After our calculation, the 95% confidence interval of the difference between the two sample means is:

```
##   lower_ci upper_ci
## 1 2094.974  2278.28
```

As we can see, the two 95% confidence intervals does not contain 0. Therefore, we can conclude that there is a significant increase in average daily bike rentals from 2011 to 2012.

Question 1.5: Do casual or registered users contribute to the overall growth from 2011 to 2012?

To answer this question, we decided to use a simpler method than the previous one. First we calculated the monthly total of casual and registered members:



From the line graph above, we can't conclude an increase in casual bike rentals from 2011 to 2012. However, there seems to be a noticeable increase in the number of registered users. Therefore, a table of the average monthly casual and registered users are created as the following:

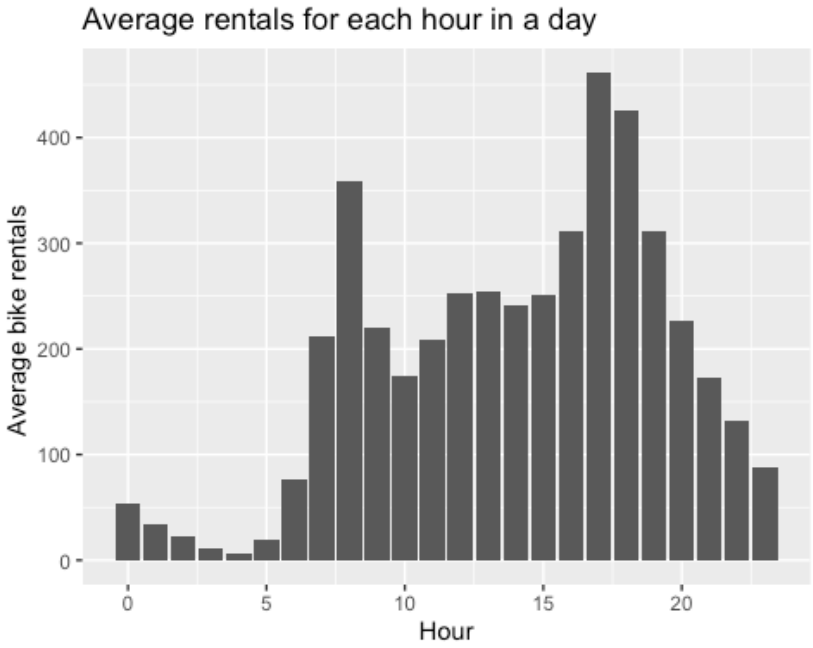
```
## # A tibble: 4 × 3
## # Groups:   yr [2]
##   yr    type          mean
##   <fct> <chr>        <dbl>
```

```
## 1 2011  casual      20604.
## 2 2012  casual      31064.
## 3 2011  registered  82988.
## 4 2012  registered 139734.
```

In fact, there is a 50.76 percent increase in casual users, and 68.38 percent increase in registered users. Therefore, both casual and registered users contribute to the overall growth of bike rentals from 2011 to 2012.

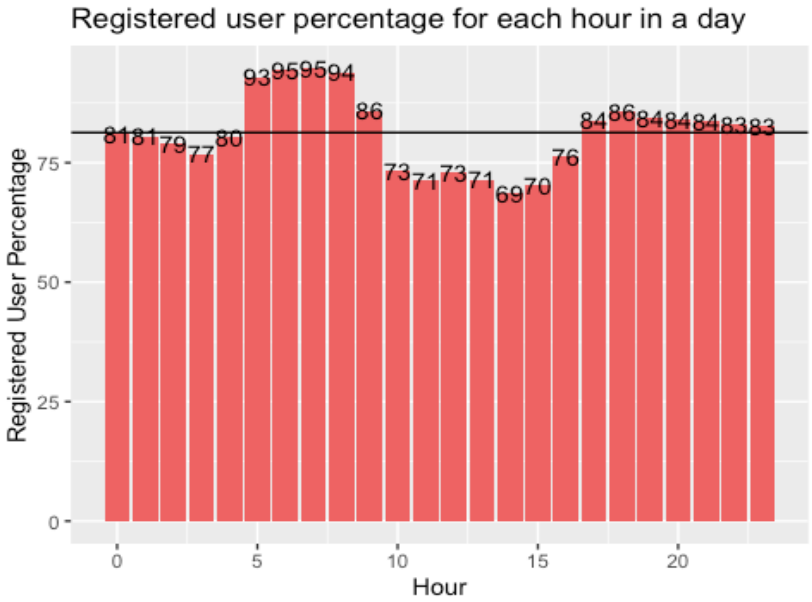
Question 2: What hour of a day is the most popular time for bike rental? Specifically for registered and casual users?

We are curious about this question because this will help the bike sharing company redistribute their bike supply every day. We will use the `hour.csv` dataset for this question. Now, let us look at plot of the average bike rental for every hour of a day.



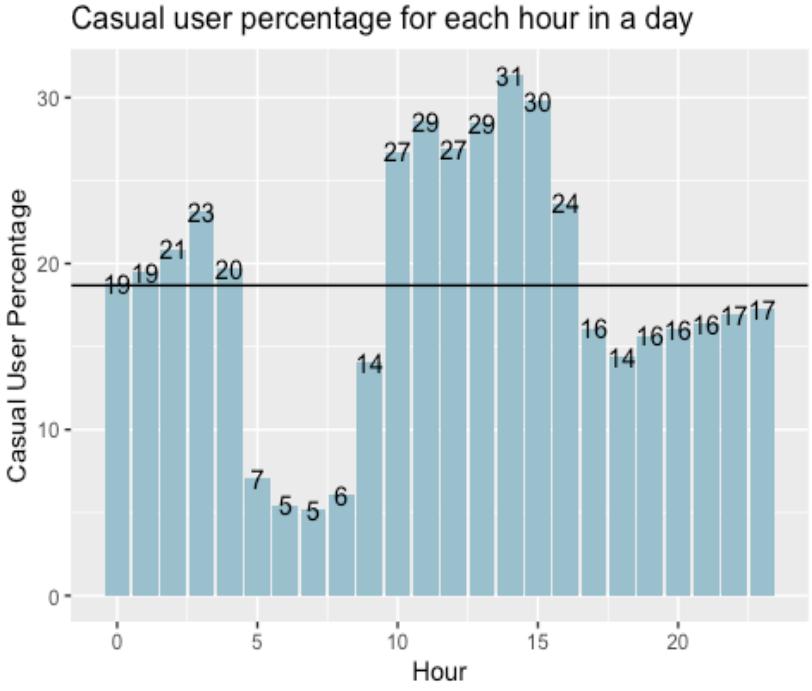
We can see that 8am, 5pm, and 6pm are the most popular hours of a day, averaging over 350 rentals in each of these hours. We suspect that this is mainly due to people going to / coming back from work in these hours. Therefore, the majority of users should be

registered. We made the following chart for the registered user percentage:



As expected, we can see from the bar chart that during 5-8am, more than 90% of the bike rentals came from registered users, while it is around 85% during 5-9 pm. This conforms with our prediction that most users during the most popular hours are registered, because they use the bike rental system regularly for commuting purposes.

Now, we can also take a look at the percentage of casual users throughout a day:



The chart shows that there are few casual users in the morning hours (5am-9am), averaging only around 5 percent of total users during that time. However, most casual users appear during noon and afternoon (11am-3pm), accounting for over 27 percent of all users. This is most likely because casual users use the bike sharing system for recreational purposes rather than the regular every-day commuting.

Question 3: How does temperature, based on the current weather situation, affect the number of bike rentals?

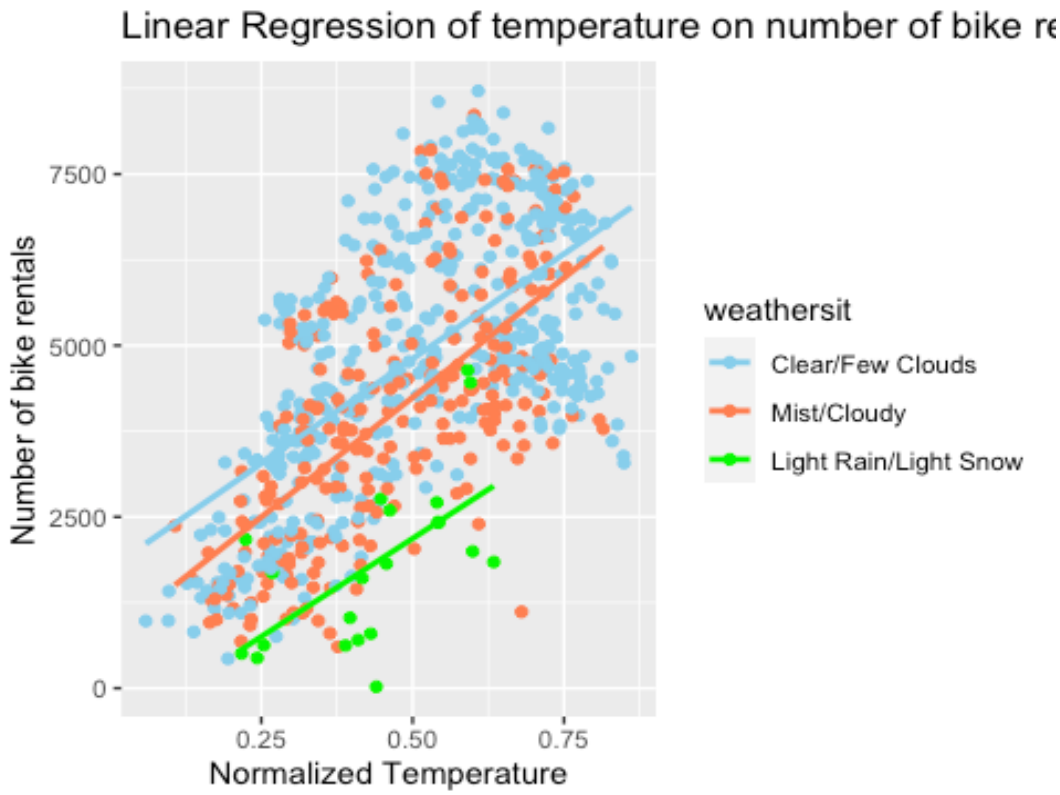
First, let us compute the correlation coefficient between the two numerical variables, temperature and number of bike rentals per day:

```
##        cor
## 1 0.627494
```

The correlation coefficient is 0.627, which means that there is a fairly strong correlation between temperature and the number of people renting bike.

Before we continue, only 3 weather situations appeared in the daily count dataset, even though there are four in total (Heavy Rain + Ice Pallets + Thunderstorm + Mist, or Snow + Fog), which implies that the fourth condition did not happen consistently throughout any day from 2011 to 2012.

We will now use a scatter plot to visualize their relationship, and we will also add `weathersit` as a color factor. Note that even though `weathersit` is labeled numerically, it is a categorical variable of integers from 1 to 4.



We notice that, first, all three regression lines are positively sloped, which means that higher temperatures tend to lead to more bike rentals, and second, Clear/Few Clouds weathers account for the most bike rentals, followed by Mist/Cloudy weathers, while Light Rain/Light Snow accounts for the lowest numbers.

We will now get the regression table of the linear interaction model (Note: `temp` is the normalized temperature between 0 and 1, which is why the slopes are such large numbers):

```
## # A tibble: 6 × 7
##   term             estimate std_error statistic p_value lower_ci upper_ci
##   <chr>               <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept           1748.      190.     9.17    0        1374.    2122.
## 2 temp                6117.      349.    17.5     0        5432.    6802.
## 3 weathersit: 2       -994.      333.    -2.99    0.003   -1648.    -341.
## 4 weathersit: 3      -2426.     1118.    -2.17    0.03    -4622.    -230.
## 5 temp:weathersit2     862.      649.     1.33    0.184    -412.    2135.
## 6 temp:weathersit3    -392.     2463.    -0.159   0.873   -5228.    4443.
```

Using the regression table, we can write out the equation for the regression lines:

$\hat{Y} = 1747.58 + 6117.22temp - 994.3I_{weathersit:2}(x) - 2425.71I_{weathersit:3}(x) + 861.58tempI_{weathersit:2}(x) - 392.28tempI_{weathersit:3}(x)$, where $I(x)$ is the indicator function for weathersit 2 and 3.

To specify:

- for Clear/Partly Cloudy weathers, $\hat{Y} = 1747.58 + 6117.22temp$;
- for Mist/Cloudy weathers, $\hat{Y} = 753.273 + 6978.801temp$;
- for light rain/light snow weathers, $\hat{Y} = -678.132 + 5724.936temp$.

Overall, the linear regression model shows us that the daily bike rental count increases by roughly the same amount for every unit of increase in temperature for different weather situations. And as expected, given the same temperature, daily bike rental numbers in light rain/light snow weathers will be about 2400 lower than clear/few cloud weathers and about 1500 lower than cloudy/mist weathers.

Section 4: Conclusion

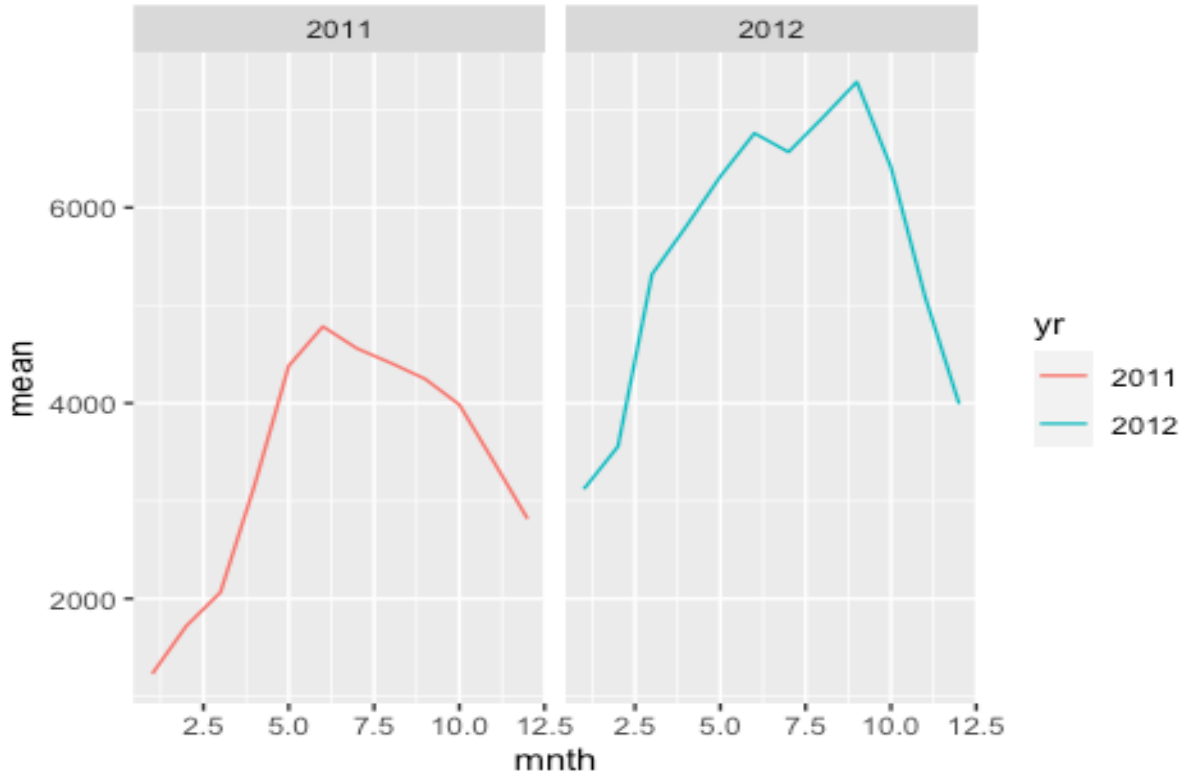
In conclusion, we found that there is a significant growth in the bike rental user-base for both casual and registered users from 2011 to 2012. We also discovered that 8am, 5pm, and 6pm are the most popular hours, mainly because registered users are commuting to / off work in these hours. As for casual users, they like to use rental bikes in the noon and afternoon hours, possibly for recreational purposes. One thing the company can use this information for is to redistribute their supply of bikes between stations. Finally, we showed that the higher the temperature is, the more people will use rental bikes. Cloudy/mist and light rain/light snow weathers negatively influences on bike rentals, but they do not have a significant impact on how temperature affects bike rentals.

Section 5: Appendix

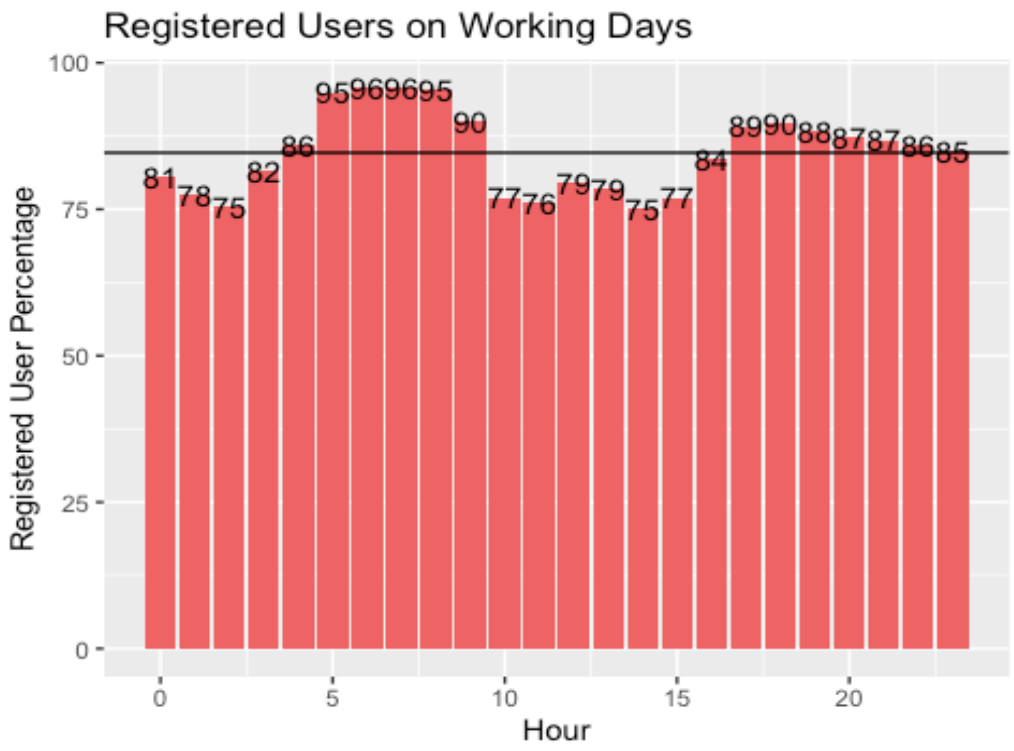
5.1 Some Interesting Graphs

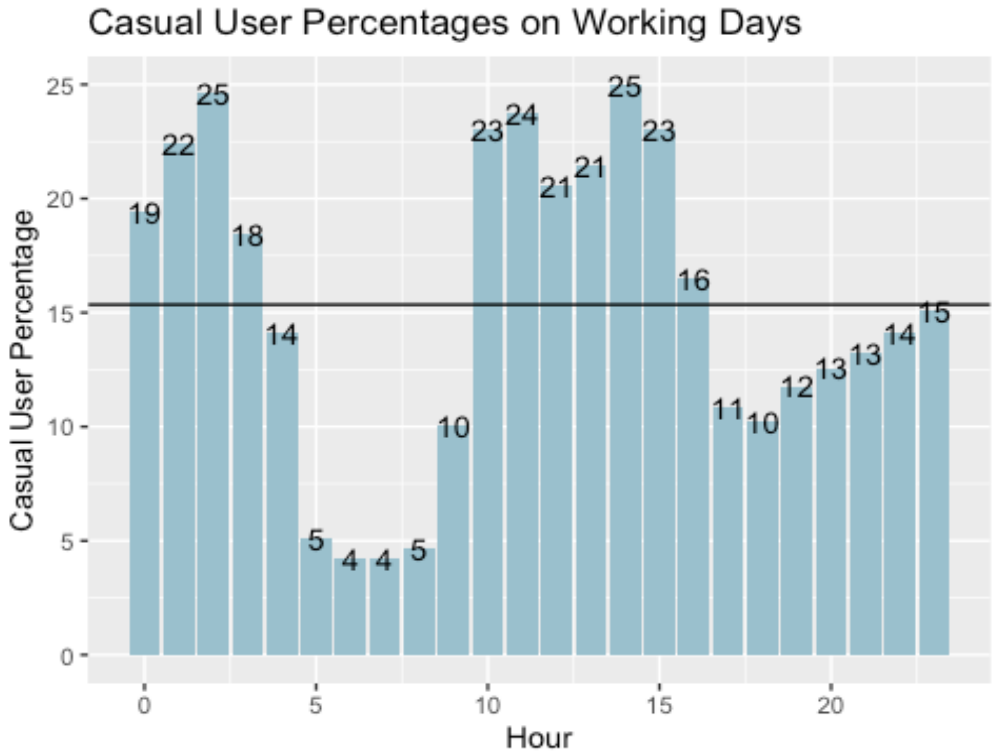
I initially planned to include these in the analysis, but they might make the report over the 10-page limit, so I will put them here.

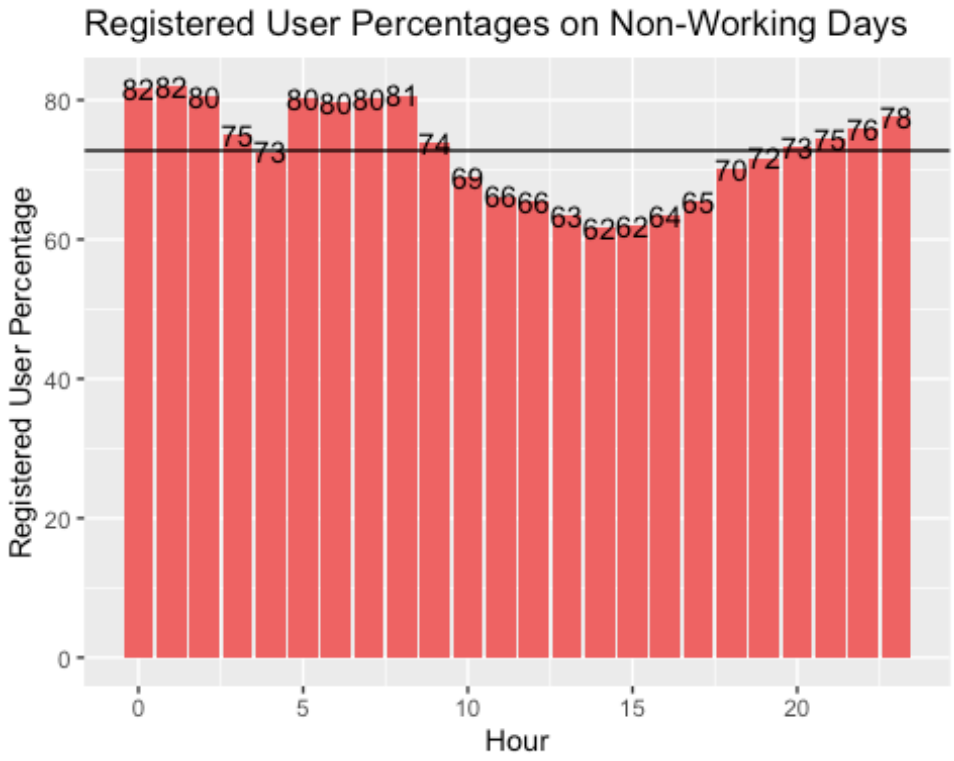
5.1.1 Monthly Average

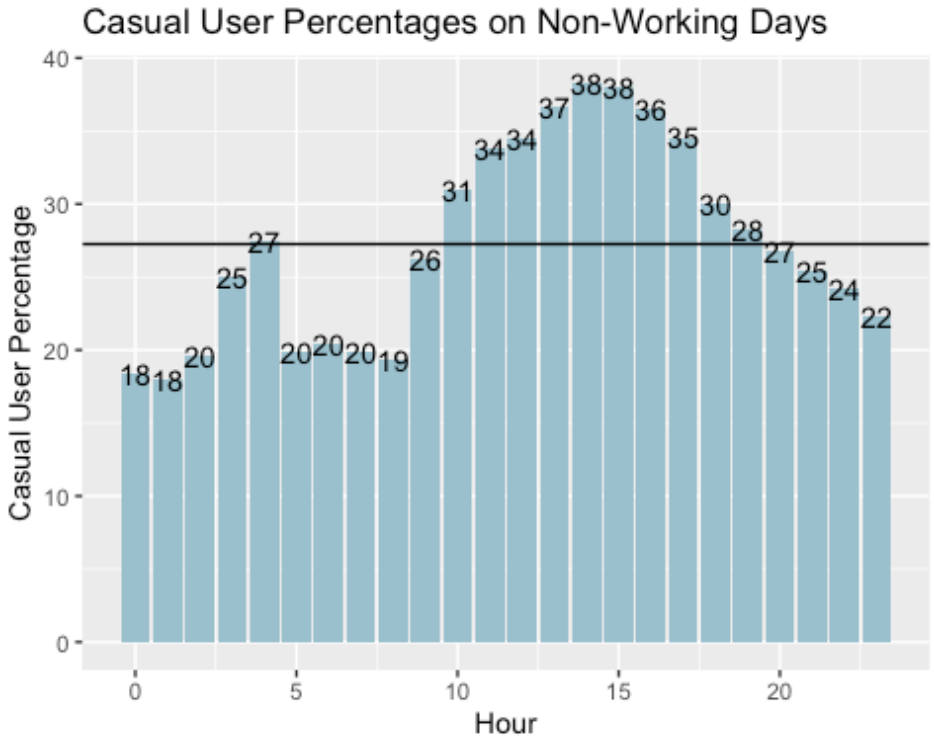


5.1.2 Popular hours on working days vs. non-working days (weekends and holidays), casual vs registered






Registered User Percentages on Non-Working Days
Registered User Percentage
Hour
0
20
40
60
80
0
5
10
15
20
82 82 80 75 73 80 80 80 81 74 69 66 66 63 62 62 64 65 70 72 73 75 76 78


Casual User Percentages on Non-Working Days
Casual User Percentage
Hour
0
10
20
30
40
0
5
10
15
20
18 18 20 25 27 20 20 20 19 26 31 34 34 37 38 38 36 35 30 28 27 25 24 22

5.2 All R Codes

```
# read and clean
# read the datasets
day_df_raw = read.csv("datasets/day.csv")
hour_df_raw = read.csv("datasets/hour.csv")

# function to clean the datasets
clean_df = function(df){
  df$season = as.factor(df$season)
  df$holiday = as.factor(df$holiday)
  df$weekday = as.factor(df$weekday)
  df$workingday = as.factor(df$workingday)
  df$weathersit = as.factor(df$weathersit)
  # change 0 to 2011 and 1 to 2012 in the yr column
  df = df |>
    mutate(yr = replace(yr, yr == 0, 2011)) |>
    mutate(yr = replace(yr, yr == 1, 2012))
  df$yr = as.factor(df$yr)
  return(df)
}

# run the function on both datasets
day_df = clean_df(day_df_raw)
hour_df = clean_df(hour_df_raw)

day_df |> head(10)


#Q1
day_df |>
  group_by(yr) |>
  summarize(daily_average = mean(cnt))

set.seed(40)
# bootstrap resampling
resampled_2011 = day_df |>
  select(yr, cnt) |>
  filter(yr==2011) |>
  rep_sample_n(size = 100, replace = TRUE, reps = 200)
resampled_2011_mean = resampled_2011 |>
  group_by(replicate) |>
  summarize(daily_mean = mean(cnt))
resampled_2012 = day_df |>
  select(yr, cnt) |>
  filter(yr==2012) |>
  rep_sample_n(size = 100, replace = TRUE, reps = 200)
resampled_2012_mean = resampled_2012 |>
  group_by(replicate) |>
```

```
  summarize(daily_mean = mean(cnt))

ggplot(resampled_2011_mean, aes(x = daily_mean)) +
  geom_histogram(binwidth =50) +
  labs(x = "Resample daily rental mean", title = "Bootstrap distribution of t
he daily mean of 2011")

# t test statistic and p value
t.test(resampled_2011_mean$daily_mean, resampled_2012_mean$daily_mean)$statis
tic

t.test(resampled_2011_mean$daily_mean, resampled_2012_mean$daily_mean)$p.valu
e


# population mean of 2011 distribution
x_bar_2011 = resampled_2011_mean$daily_mean |> mean()
# calculate the confidence interval using standard error method
ci_2011 <- resampled_2011_mean |>
  get_confidence_interval(level = 0.95, type = "se", point_estimate = x_bar_2
011)
ci_2012 <- resampled_2012_mean |>
  get_confidence_interval(level = 0.95, type = "percentile")
ci_2012 - ci_2011


monthly_total = day_df |>
  group_by(yr, mnth) |>
  summarise(casual = sum(casual),registered = sum(registered)) |>
  pivot_longer(c(casual, registered),
               names_to = "type", values_to = "monthly_total")
monthly_total |>
  ggplot(aes(color=type))+
  geom_line(aes(x=mnth, y=monthly_total)) +
  facet_wrap(~yr) +
  labs(title = "Line graph of monthly total for casual and registered users i
n 2011 and 2012", x="month", y="Monthly bike rentals")

avg_monthly_total = monthly_total |>
  group_by(yr, type) |>
  summarize(mean = mean(monthly_total)) |>
  arrange(type)
avg_monthly_total

#Q2
hour_grouped = hour_df |>
  group_by(hr) |>
  summarize(hourly_avg = mean(cnt), casual_pct = mean(casual)/hourly_avg, reg
istered_pct = mean(registered)/hourly_avg)
```

```
hour_grouped |> ggplot() +
  geom_bar(aes(x=hr, weight = hourly_avg)) +
  labs(title = "Average rentals for each hour in a day", x="Hour", y="Average
bike rentals ")

hour_grouped  |>
  ggplot(aes(x=hr, weight = registered_pct * 100)) +
  geom_bar(fill="indianred2") +
  geom_text(aes(label = round(registered_pct * 100, 0), y = registered_pct *
100)) +
  geom_hline(aes(yintercept = mean(registered_pct * 100))) +
  labs(title = "Registered user percentage for each hour in a day", x="Hour",
y="Registered User Percentage")

hour_grouped  |>
  ggplot(aes(x=hr, weight = casual_pct * 100)) +
  geom_bar(fill="lightblue3") +
  geom_text(aes(label = round(casual_pct * 100, 0), y = casual_pct * 100)) +
  geom_hline(aes(yintercept = mean(casual_pct * 100))) +
  labs(title = "Casual user percentage for each hour in a day", x="Hour", y="
Casual User Percentage")

#Q3
day_df |>
  get_correlation(formula = cnt ~ temp)

day_df |>
  ggplot(aes(x = temp, y = cnt, color = weathersit)) +
  geom_point() +
  geom_smooth(method ="lm", formula = y ~ x, se = FALSE) +
  labs(x="Normalized Temperature", y="Number of bike rentals", title = "Linea
r Regression of temperature on number of bike rentals based on weather") +
  scale_color_manual(values = c("skyblue", "coral", "green"), labels = c("Cle
ar/Few Clouds", "Mist/Cloudy", "Light Rain/Light Snow"))
score_model_interaction <- lm(cnt ~ temp * weathersit, data = day_df)
get_regression_table(score_model_interaction)

# Extra plots
# 5.1.1
monthly_avg = day_df |>
  group_by(yr, mnth) |>
  summarise(mean = mean(cnt))
monthly_avg |>
  ggplot(aes(x=mnth, y=mean, color=yr))+
  geom_line() +
  facet_wrap(~yr)

# 5.2.2
working_days_grouped = hour_df |>
```

```
  filter(holiday == 0 & workingday == 1)  |>
  group_by(hr) |>
  summarize(hourly_avg = mean(cnt), casual_pct = mean(casual)/hourly_avg, reg
istered_pct = mean(registered)/hourly_avg)
non_working_days_grouped = hour_df |>
  filter(holiday == 1 | workingday == 0) |>
  group_by(hr) |>
  summarize(hourly_avg = mean(cnt), casual_pct = mean(casual)/hourly_avg, reg
istered_pct = mean(registered)/hourly_avg)
working_days_grouped |>
  ggplot(aes(x=hr, weight = registered_pct * 100)) +
  geom_bar(fill="indianred2") +
  geom_text(aes(label = round(registered_pct * 100, 0), y = registered_pct *
100)) +
  geom_hline(aes(yintercept = mean(registered_pct * 100))) +
  labs(title = "Registered Users on Working Days", x="Hour", y="Registered Us
er Percentage")

working_days_grouped  |>
  ggplot(aes(x=hr, weight = casual_pct * 100)) +
  geom_bar(fill="lightblue3") +
  geom_text(aes(label = round(casual_pct * 100, 0), y = casual_pct * 100)) +
  geom_hline(aes(yintercept = mean(casual_pct * 100))) +
  labs(title = "Casual User Percentages on Working Days", x="Hour", y="Casual
User Percentage")

non_working_days_grouped |>
  ggplot(aes(x=hr, weight = registered_pct * 100)) +
  geom_bar(fill="indianred2") +
  geom_text(aes(label = round(registered_pct * 100, 0), y = registered_pct *
100)) +
  geom_hline(aes(yintercept = mean(registered_pct * 100))) +
  labs(title = "Registered User Percentages on Non-Working Days", x="Hour", y
="Registered User Percentage")


non_working_days_grouped  |>
  ggplot(aes(x=hr, weight = casual_pct * 100)) +
  geom_bar(fill="lightblue3") +
  geom_text(aes(label = round(casual_pct * 100, 0), y = casual_pct * 100)) +
  geom_hline(aes(yintercept = mean(casual_pct * 100))) +
  labs(title = "Casual User Percentages on Non-Working Days", x="Hour", y="Ca
sual User Percentage")
```