DATA607 - Final Project

Cycling Data

Donald Butler

12/08/2021

DATA 607 Final Project

Introduction

I serve on the board for my local cycling club and for the past 6 years I have been collecting the sign-in sheets from our club rides. We hold about 200 rides per year and regularly discuss falling membership and ride attendance, and often speculate that some rides are less popular for various reasons. I have thought about analyzing the sign-in sheet data that I have, but it's not currently in a format that would be conducive to reading digitally.

I was inspired by a dataset that I found which examined data from a bicycle ride sharing service in Washington DC and correlated historical weather data. I downloaded two years of data from https://www.capitalbikeshare.com/. Then acquired a corresponding weather dataset for the same time period.

Load required R Libraries

```
library(tidyverse)
library(lubridate)

# Federal holidays obtained from https://www.opm.gov/policy-data-oversight/pay-leave/federal-holidays
holidays <- as.Date(c('2018-01-01', '2018-01-15', '2018-02-19', '2018-05-28', '2018-07-04',
                      '2018-09-03', '2018-10-08', '2018-11-12', '2018-11-22', '2018-12-25',
                      '2019-01-01', '2019-01-21', '2019-02-18', '2019-05-27', '2019-07-04',
                      '2019-09-02', '2019-10-14', '2019-11-11', '2019-11-28', '2019-12-25'))

get_season <- function(date) {
  # 1 (Winter), 2 (Spring), 3 (Summer), 4 (Fall)
  return (((lubridate::month(date + 10) - 1) %/% 3) + 1)
}

is.holiday <- function(date) {
  return (ifelse(date %in% holidays, 1, 0))
}
```

Data

Download Bike Share Data

Capital Bike Share provides system data each month that contains complete trip history.

- Duration
- Start Date
- End Date
- Start Station
- End Station
- Bike Number
- Member Type - (Member, Casual)

Data is provided in monthly zip files. I downloaded the files for 2018 and 2019 then extract the csv file that is within the file.

```
for (year in 2018:2019) {
  for (month in 1:12) {
    zipfile <- paste0(year, str_pad(month, 2, "left", "0"), "-capitalbikeshare-tripdata.zip")
    if (!file.exists(paste0('./Data/Bike/Zips/', zipfile))) {
      download.file(paste0('https://s3.amazonaws.com/capitalbikeshare-data/', zipfile), paste0('./Data/
      unzip(paste0('./Data/Bike/Zips/',zipfile), exdir = 'Data/Bike')
    }
  }
}
```

Read Bike Share Data

Each data file contains an observation for each bike ride. For this project I wanted to compare the daily number of rides so I will be grouping the data to build a dataset with an observation for each day in my two years of data.

I created the same variables as the dataset I was using for my DATA 606 project so that I might be able to compare them later.

```
filenames <- list.files(path = './Data/Bike/', pattern = '.csv')

bike_df <- tibble()

for (i in 1:length(filenames)) {
  tmp <- read_csv(paste0('./Data/Bike/', filenames[i]), show_col_types = FALSE) %>%
    rename(start_date = `Start date`, member_type = `Member type`) %>%
    mutate(dteday = as.Date(start_date)) %>%
    group_by(dteday) %>%
    count(member_type, name = 'cnt') %>%
    mutate(yr = as.integer(lubridate::year(dteday)),
```

```
           mnth = as.integer(lubridate::month(dteday)),
           season = as.integer(get_season(dteday)),
           season_name = case_when(season == 1 ~ 'Winter', season == 2 ~ 'Spring',
                                   season == 3 ~ 'Summer', season == 4 ~ 'Fall'),
           holiday = as.integer(is.holiday(dteday)),
           weekday = as.integer(lubridate::wday(dteday) - 1),
           workingday = as.integer(ifelse(holiday == 0 & weekday %in% c(1:5), 1, 0)),
           member = as.integer(ifelse(member_type == 'Member', 1, 0))) %>%
    select(dteday,
           season,
           season_name,
           yr,
           mnth,
           holiday,
           weekday,
           workingday,
           member,
           cnt)

  bike_df <- bike_df %>% rbind(tmp)
}

glimpse(bike_df)
```

```
## Rows: 1,468
## Columns: 10
## Groups: dteday [730]
## $ dteday      <date> 2018-01-01, 2018-01-01, 2018-01-02, 2018-01-02, 2018-01-0~
## $ season      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ season_name <chr> "Winter", "Winter", "Winter", "Winter", "Winter", "Winter"~
## $ yr          <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018~
## $ mnth        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ holiday     <int> 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ weekday     <int> 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 0, 0, 1, 1, 2, 2, 3, 3~
## $ workingday  <int> 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1~
## $ member      <int> 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1~
## $ cnt         <int> 145, 1068, 179, 3613, 279, 4469, 66, 2302, 62, 2647, 52, 1~
```

Read Washington DC weather data

I had intend to use an API to obtain the weather data for my project, but had a hard time finding anything that would allow me to gather data from the time period of my cycling data. To move the project forward, I used a bulk download of data from https://openweathermap.org/. Recently I found a site that I could possibly scrape data from.

My weather data provides an observation for each hour of the two year period I selected. Using this data I constructed a daily summary which included average temperature, humidity, wind speed, precipitation, and cloud coverage.

```
if (!file.exists('./Data/Weather/WashingtonDC_Weather.csv')) {
  download.file('https://raw.githubusercontent.com/dab31415/DATA607/main/Projects/Project_Final/Data/Wea
}

weather_df <- read_csv('./Data/Weather/WashingtonDC_Weather.csv', show_col_types = FALSE) %>%
  filter(grepl('^201[8-9]',dt_iso)) %>%
  mutate(dteday = as.Date(substr(dt_iso,1,10)),
         precipitation = ifelse(is.na(rain_1h), 0, rain_1h) + ifelse(is.na(snow_1h), 0, snow_1h)) %>%
  select(dteday,
         temp,
         humidity,
         wind_speed,
         precipitation,
         clouds_all) %>%
  distinct() %>%
  group_by(dteday) %>%
  summarise(temp = mean(temp),
            humidity = mean(humidity),
            wind_speed = mean(wind_speed),
            precipitation = sum(precipitation),
            clouds = mean(clouds_all))

glimpse(weather_df)
```

```
## Rows: 730
## Columns: 6
## $ dteday        <date> 2018-01-01, 2018-01-02, 2018-01-03, 2018-01-04, 2018-01~
## $ temp          <dbl> 18.35000, 17.84500, 23.04542, 25.92667, 16.32000, 14.885~
## $ humidity      <dbl> 48.58333, 40.70833, 52.62500, 62.62500, 38.62500, 36.791~
## $ wind_speed    <dbl> 8.201250, 8.127083, 2.518750, 10.719583, 14.353750, 11.6~
## $ precipitation <dbl> 0.00, 0.00, 0.00, 6.22, 0.00, 0.00, 0.00, 0.59, 1.45, 0.~
## $ clouds        <dbl> 19.750000, 4.208333, 43.708333, 85.000000, 16.833333, 9.~
```

Join Bike and Weather data

I joined the two datasets by date to produce a single dataset that I could analyse.

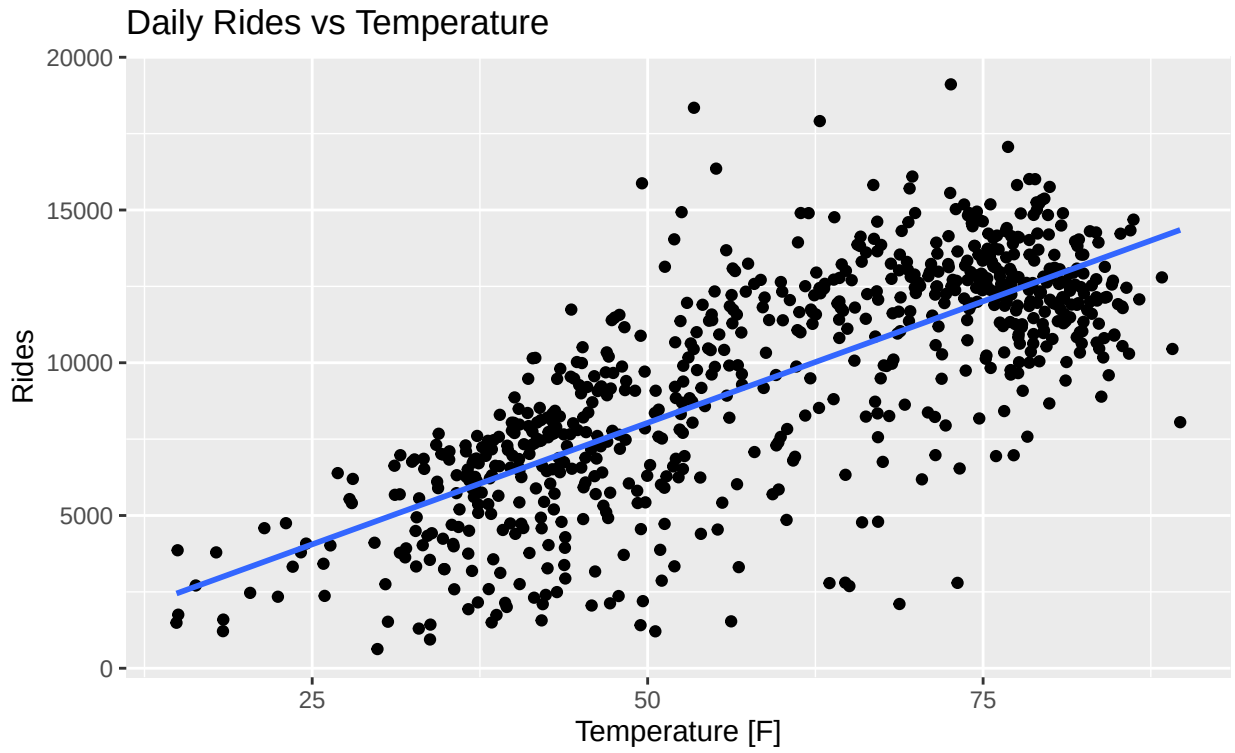
```
df <- bike_df %>%
  inner_join(weather_df, by = 'dteday')

glimpse(df)
```

```
## Rows: 1,468
## Columns: 15
## Groups: dteday [730]
## $ dteday        <date> 2018-01-01, 2018-01-01, 2018-01-02, 2018-01-02, 2018-01~
## $ season        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ season_name   <chr> "Winter", "Winter", "Winter", "Winter", "Winter", "Winte~
## $ yr            <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 20~
## $ mnth          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ holiday       <int> 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ weekday       <int> 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 0, 0, 1, 1, 2, 2, 3,~
## $ workingday    <int> 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1,~
## $ member        <int> 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0,~
## $ cnt           <int> 145, 1068, 179, 3613, 279, 4469, 66, 2302, 62, 2647, 52,~
## $ temp          <dbl> 18.35000, 18.35000, 17.84500, 17.84500, 23.04542, 23.045~
## $ humidity      <dbl> 48.58333, 48.58333, 40.70833, 40.70833, 52.62500, 52.625~
## $ wind_speed    <dbl> 8.201250, 8.201250, 8.127083, 8.127083, 2.518750, 2.5187~
## $ precipitation <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 6.22, 6.22, 0.00, 0.~
## $ clouds        <dbl> 19.750000, 19.750000, 4.208333, 4.208333, 43.708333, 43.~
```

Statistical Analysis

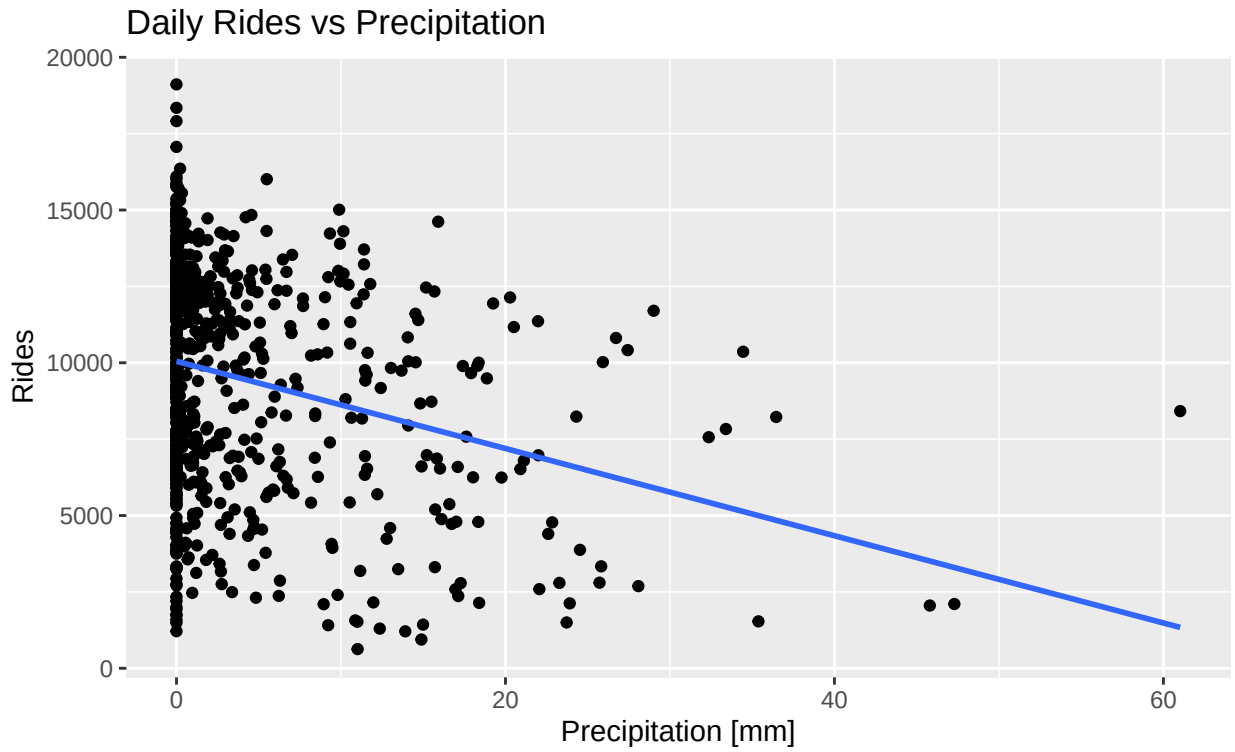
Impact of Temperature

```
df %>%
  group_by_at(vars(-member, -cnt)) %>%
  summarise(cnt = sum(cnt)) %>%
  ggplot(aes(x = temp, y = cnt)) +
  geom_point() + geom_smooth(method = 'lm', se = FALSE) +
  labs(title = 'Daily Rides vs Temperature', x = 'Temperature [F]', y = 'Rides')
```



Impact of Precipitation

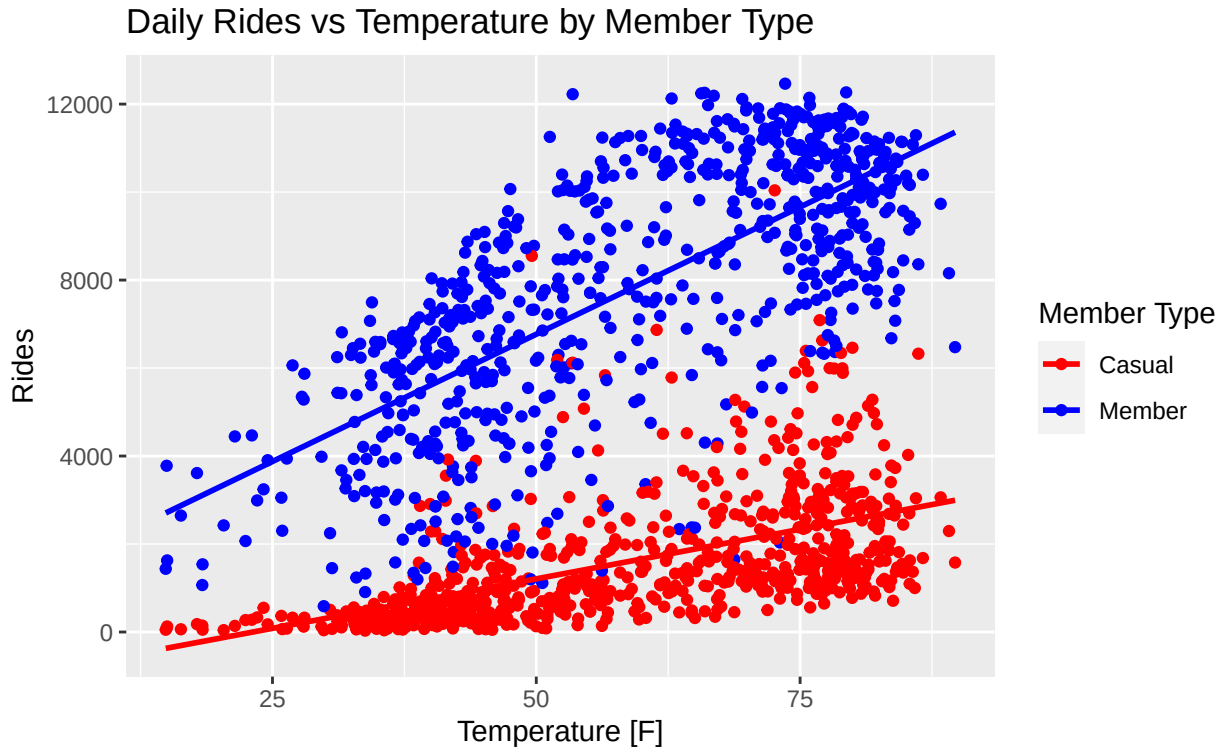
```
df %>%
  group_by_at(vars(-member, -cnt)) %>%
  summarise(cnt = sum(cnt)) %>%
  ggplot(aes(x = precipitation, y = cnt)) +
  geom_point() + geom_smooth(method = 'lm', se = FALSE) +
  labs(title = 'Daily Rides vs Precipitation', x = 'Precipitation [mm]', y = 'Rides')
```



Rides vs Temperature by Member Type

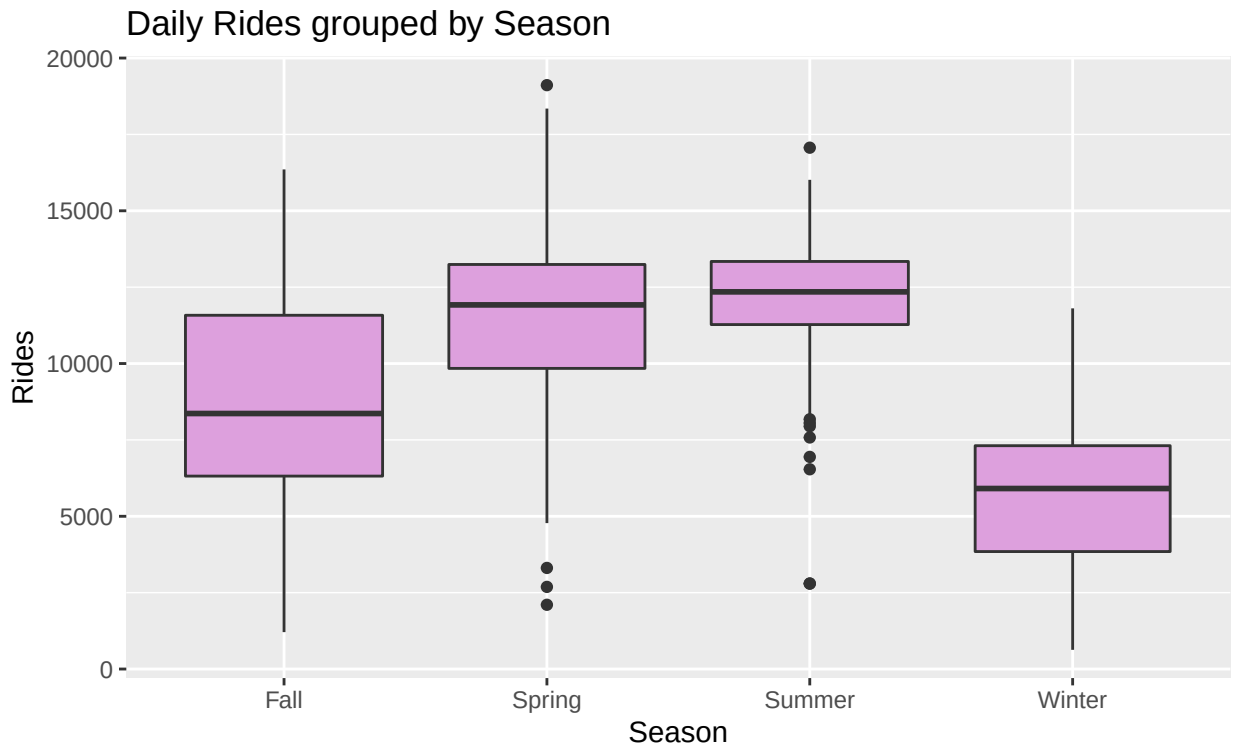
The bike share data classifies each rider based on their membership. A full member has an annual or monthly membership and a casual rider has a 1 or 5 day pass to the system. This seems to be important because full members are primarily commuters which may ride in colder conditions than the casual rider.

```
df %>%
  ggplot(aes(x = temp, y = cnt, color = factor(member))) +
  geom_point() + geom_smooth(method = 'lm', se = FALSE) +
  labs(title = 'Daily Rides vs Temperature by Member Type', x = 'Temperature [F]', y = 'Rides', color =
  scale_color_manual(labels = c('Casual','Member'), values = c('red', 'blue'))
```

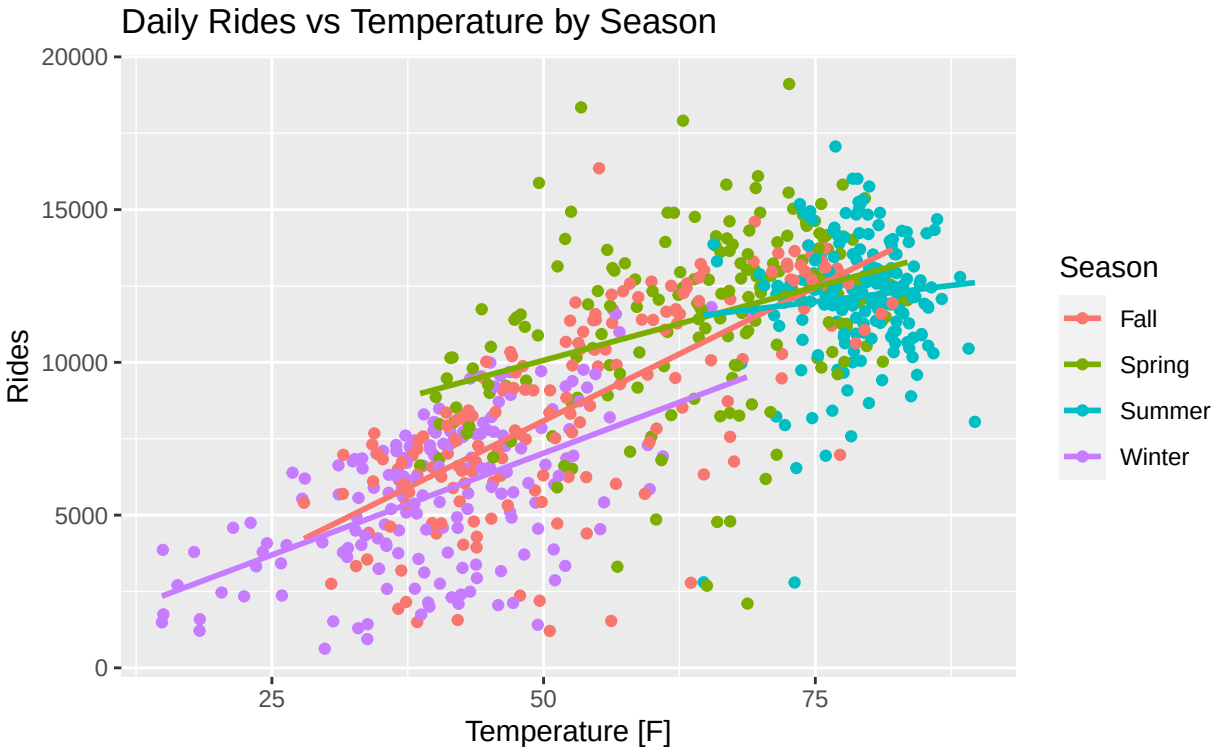


Impact of the Season

```
df %>%
  group_by_at(vars(-member, -cnt)) %>%
  summarise(cnt = sum(cnt)) %>%
  ggplot(aes(x = season_name, y = cnt)) +
  geom_boxplot(varwidth=T, fill='plum') +
  labs(title = 'Daily Rides grouped by Season',
       x = 'Season',
       y = 'Rides')
```

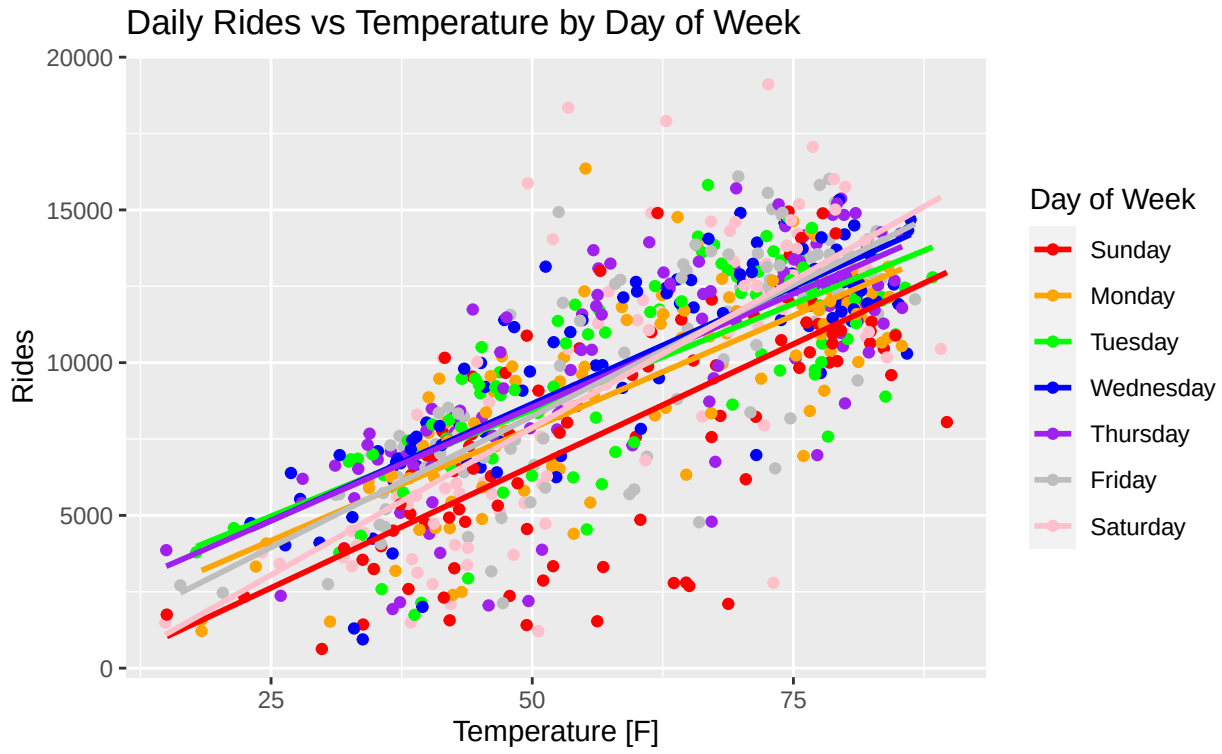



```
df %>%
  group_by_at(vars(-member, -cnt)) %>%
  summarise(cnt = sum(cnt)) %>%
  ggplot(aes(x = temp, y = cnt, color = season_name)) +
  geom_point() + geom_smooth(method = 'lm', se = FALSE) +
  labs(title = 'Daily Rides vs Temperature by Season', x = 'Temperature [F]', y = 'Rides', color = 'Seas
```

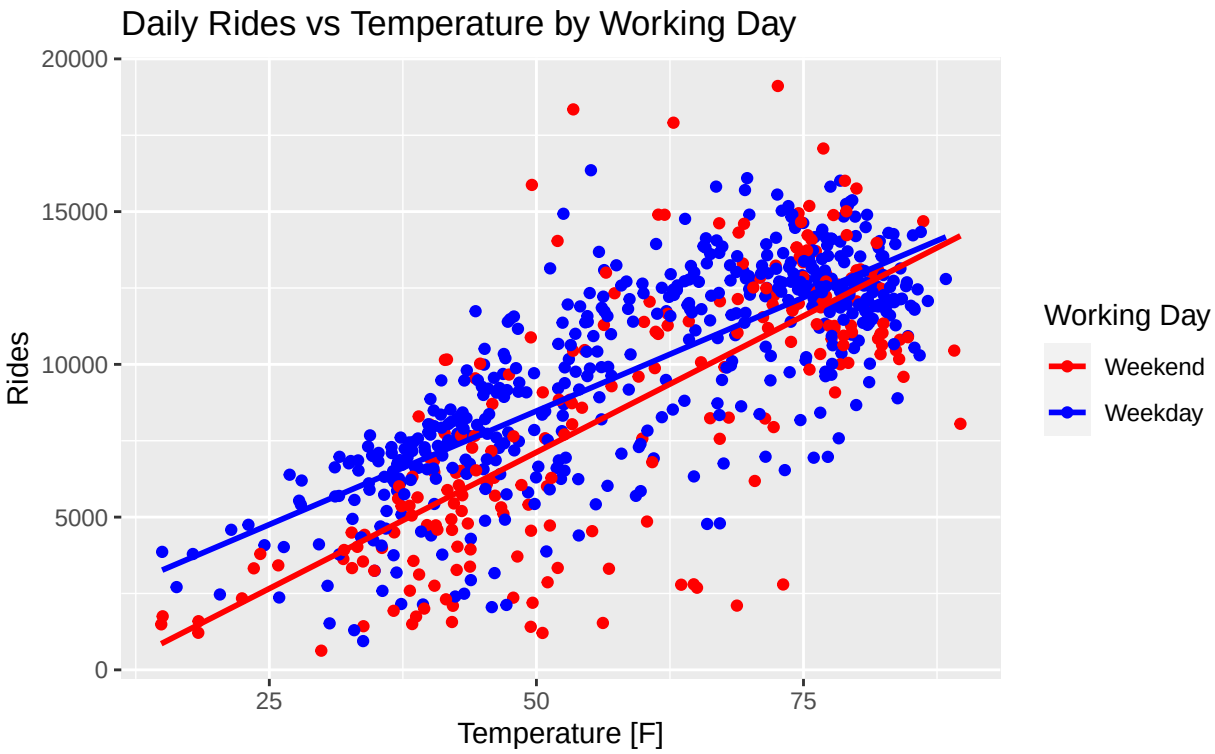

Daily Rides vs Temperature by Season
Rides
20000
15000
10000
5000
0
25
50
75
Temperature [F]
Season
Fall
Spring
Summer
Winter

Rides by Day of Week

```
df %>%
  group_by_at(vars(-member, -cnt)) %>%
  summarise(cnt = sum(cnt)) %>%
  ggplot(aes(x = temp, y = cnt, color = factor(weekday))) +
  geom_point() + geom_smooth(method = 'lm', se = FALSE) +
  labs(title = 'Daily Rides vs Temperature by Day of Week', x = 'Temperature [F]', y = 'Rides', color =
  scale_color_manual(labels = c('Sunday','Monday','Tuesday','Wednesday','Thursday','Friday','Saturday')
```




```
df %>%
  group_by_at(vars(-member, -cnt)) %>%
  summarise(cnt = sum(cnt)) %>%
  ggplot(aes(x = temp, y = cnt, color = factor(workingday))) +
  geom_point() + geom_smooth(method = 'lm', se = FALSE) +
  labs(title = 'Daily Rides vs Temperature by Working Day', x = 'Temperature [F]', y = 'Rides', color =
  scale_color_manual(labels = c('Weekend','Weekday'), values = c('red', 'blue'))
```


Daily Rides vs Temperature by Working Day
Rides
20000
15000
10000
5000
0
25
50
75
Temperature [F]
Working Day
Weekend
Weekday

Linear Regression

```
m_best <- df %>%
  lm(cnt ~ temp + workingday + + season_name + member + precipitation + humidity, data = .)
summary(m_best)
```

```
##
## Call:
## lm(formula = cnt ~ temp + workingday + +season_name + member +
##     precipitation + humidity, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5472.2 -1056.8    28.5  1086.0  6762.9
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1724.892    297.692  -5.794 8.40e-09 ***
## temp                80.369      4.357  18.446  < 2e-16 ***
## workingday         471.100     89.745   5.249 1.75e-07 ***
## season_nameSpring   498.663    130.899   3.810 0.000145 ***
## season_nameSummer  -174.653    161.301  -1.083 0.279086
## season_nameWinter  -459.376    128.986  -3.561 0.000381 ***
## member             6210.557     83.081  74.753  < 2e-16 ***
## precipitation       -66.900      6.986  -9.576  < 2e-16 ***
## humidity            -22.034      3.851  -5.722 1.28e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1592 on 1459 degrees of freedom
## Multiple R-squared:  0.8294, Adjusted R-squared:  0.8284
## F-statistic: 886.4 on 8 and 1459 DF,  p-value: < 2.2e-16
```

The predictors of temperature, season, member type, precipitation, humidity, and working day all shown to be statistically significant and were able to account for 82% of the variability in the daily number of rides.

Conclusion

I would like to expand this work with my own club's cycling data. I may be able to add additional predictors for the ride (distance, hilliness, etc) and also include member attributes (age, sex, employment status, etc).