PicnicHealth

Shaina Reji

I wanted to first explore what happened to the number of riders if it rained at all during the day. The below code manipulates the code so we can look at plots and t tests.

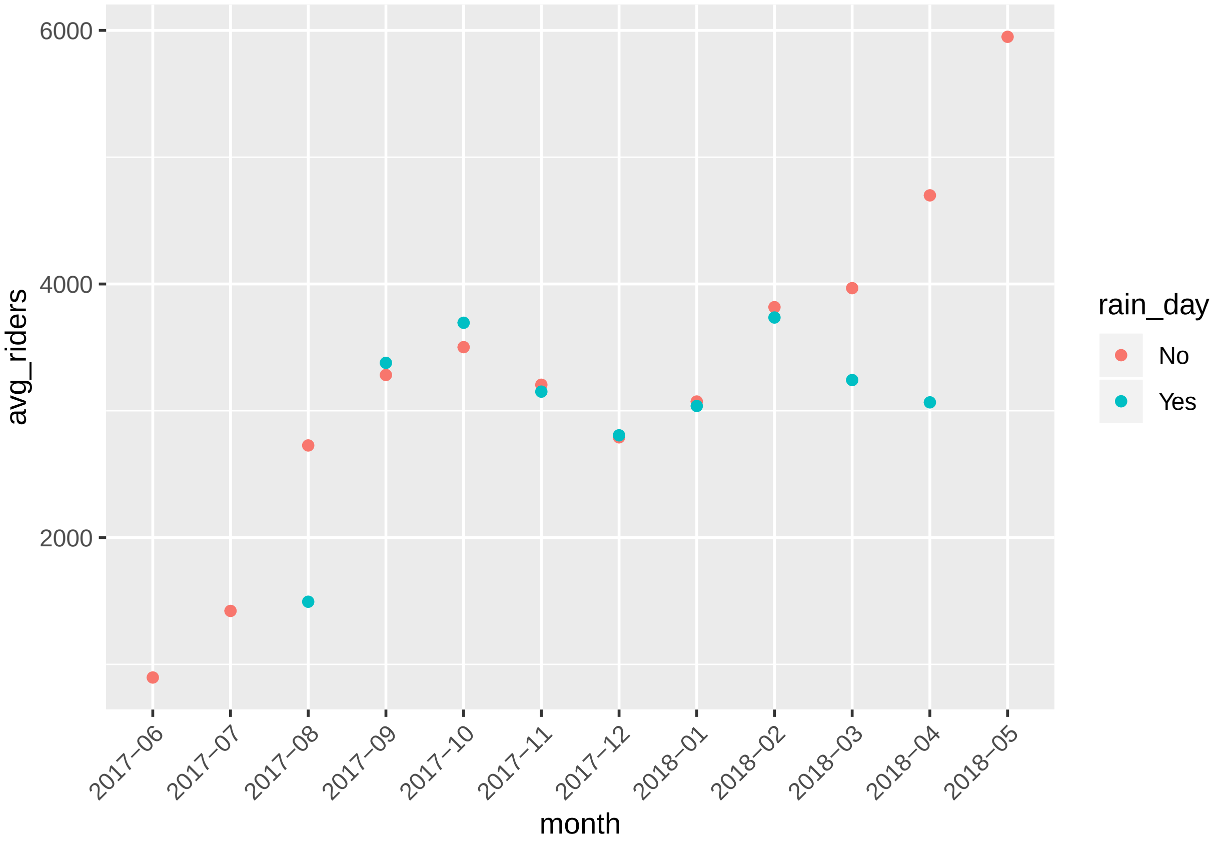
```
# Calling another file to read in the data, fix date time formats and then combine the bike data
source("/Users/sreji/Documents/PicnicHealth_Technical/setup.R")

# Gets the days that it rains at all
rain_days = weather_data %>%
  filter(!is.na(DAILYPrecip) & DAILYPrecip > 0) %>%
  distinct(ymd)
# Variable that denotes if a bike ride happens a day that it rains
bike_data = bike_data %>%
  mutate(rain_day = ifelse(start_ymd %in% rain_days$ymd, "Yes","No"))

# Aggreates the data to find riders per day and also groups by month and finds the average
# riders per month
rider_sum = bike_data %>%
  group_by(start_ymd, rain_day) %>%
  summarise(riders = n()) %>%
  ungroup() %>% mutate(month = (format(start_ymd, "%Y-%m"))) %>%
  group_by(month, rain_day) %>%
  mutate(avg_riders = mean(riders))
```

Looking at the plot of average riders, it seems as if the rain doesn't change the bike patterns. As a check, I ran a t-test on the whole data and again it showed that in general, there is no significant difference in number of riders and weather it rains or not.

There are three months that do look like it could use more analysis so for those three I ran a t-test to see if there was a significant difference in number of riders for those three months.

For August 2017, there was only 1 day that it rained, which accounts for the difference and it was insufficient data to conduct a t-test. For March 2018 and April 2018 the p-values show that with an alpha of 0.10, it is statistically significant that there is a change is bike riding patterns when it rains. (We do see that if alpha is 0.05 only the April difference has a significance.)

```
# Looking at the average number of riders per month given if it rains or not
ggplot(rider_sum %>% distinct(month, rain_day, avg_riders), aes(x = month, y = avg_riders, group_by(rai
  geom_point(aes(color = rain_day))  + theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```r
# Looking at the months that look different
#t.test(riders ~ rain_day, data = rider_sum %>% filter(month == '2017-08'))
t.test(riders ~ rain_day, data = rider_sum %>% filter(month == '2018-03'))
```

```
##
##  Welch Two Sample t-test
##
## data:  riders by rain_day
## t = 1.7388, df = 25.195, p-value = 0.09427
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -133.3484 1582.9817
## sample estimates:
##  mean in group No mean in group Yes
##          3967.067          3242.250
```

```r
t.test(riders ~ rain_day, data = rider_sum %>% filter(month == '2018-04'))
```

```
##
##  Welch Two Sample t-test
##
## data:  riders by rain_day
## t = 2.6843, df = 7.3421, p-value = 0.02999
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   207.8793 3057.0373
## sample estimates:
##  mean in group No mean in group Yes
##          4698.792          3066.333
```

```
# T-test to look at the different number of riders in total
t.test(riders ~ rain_day, data = rider_sum)
```

```
##
##  Welch Two Sample t-test
##
## data:  riders by rain_day
## t = 0.29888, df = 98.452, p-value = 0.7657
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -276.0223  373.9165
## sample estimates:
##  mean in group No mean in group Yes
##          3192.664          3143.717
```

So, if we only look at rain days between months, we see that there is a decrease in riders in March and April but not the other months.

The next thing I looked into was if there were differences in bike and rain patterns based on the stations the rider started from. The following code manipulates the data so that we get the average riders on a rainy and non rainy day per month per station. I took the average to normalize the days in case there was a disproportionate difference between the number of rain and no rain days. Then I filtered out all the stations were there was only one data point because t-tests cannot be performed on that.

```
# Data manipuation to get average riders on rain and no rain days per station
diff_stations = bike_data %>%
  group_by(start_station_name,rain_day, start_ymd) %>%
  summarise(riders = n()) %>%
  ungroup() %>%
  mutate(month = (format(start_ymd, "%Y-%m"))) %>%
  group_by(start_station_name, month, rain_day) %>%
  mutate(avg_riders = mean(riders)) %>%
  distinct(start_station_name, rain_day, month, avg_riders) %>%
  spread(., rain_day, avg_riders) %>%
  mutate(Yes = ifelse(is.na(Yes), 0, Yes))

# Stations with only one observation
one_obs = diff_stations %>% group_by(start_station_name) %>%
  summarise(count = n()) %>%
  arrange(count)  %>% filter(count <= 1)

# Taking all stations with more than one observation
diff_stations = diff_stations %>% filter(!(start_station_name %in% one_obs$start_station_name))
```

Once I ran a t-test on the stations, I found 81 stations with alpha as 0.05 where there is a significant difference in the average number of riders between a rainy and non rainy day. When alpha is 0.01 there is only 9 stations with a significant difference. Below I showed 5 stations from each of these alphas.

```
# T-test on all stations
sig_stations = diff_stations %>%
  group_by(start_station_name) %>%
  do(tidy(t.test(.$No,
                 .$Yes,
                 mu = 0,
                 alt = "two.sided",
                 paired = TRUE,
                 conf.level = 0.99)))

# Looking at significance with alpha = 0.05
alpha_05 = sig_stations %>%
  filter(p.value < 0.05) %>% select(start_station_name, p.value)
head(alpha_05)
```

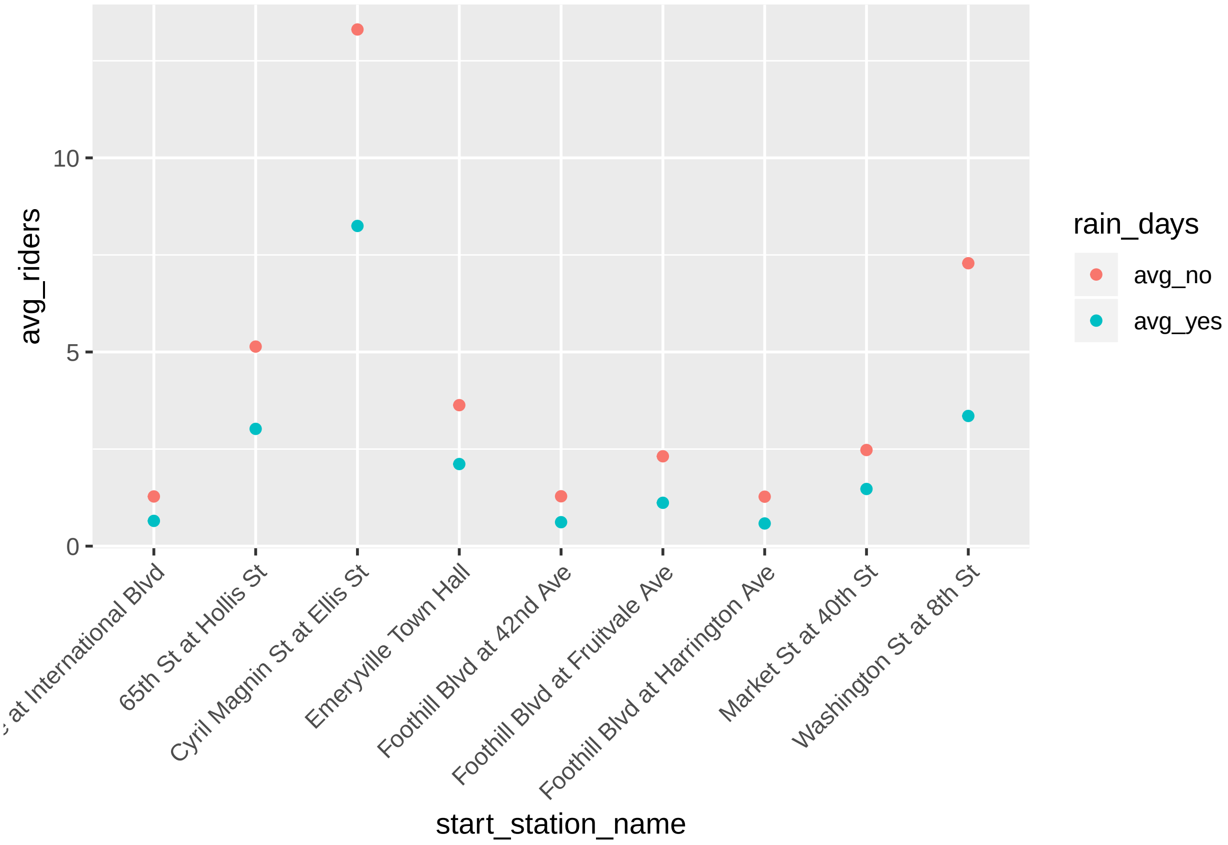
```
## # A tibble: 6 x 2
## # Groups:   start_station_name [6]
##   start_station_name                                    p.value
##   <chr>                                                   <dbl>
## 1 10th Ave at E 15th St                                  0.0497
## 2 10th St at Fallon St                                   0.0326
## 3 12th St at 4th Ave                                     0.0428
## 4 14th St at Mission St                                  0.0445
## 5 17th & Folsom Street Park (17th St at Folsom St)       0.0314
## 6 18th St at Noe St                                      0.0180
```

```
# Looking at significance with alpha = 0.01
alpha_01 = sig_stations %>%
  filter(p.value < 0.01) %>% select(start_station_name, p.value)
head(alpha_01)
```

```
## # A tibble: 6 x 2
## # Groups:   start_station_name [6]
##   start_station_name             p.value
##   <chr>                            <dbl>
## 1 26th Ave at International Blvd 0.00727
## 2 65th St at Hollis St           0.00775
## 3 Cyril Magnin St at Ellis St    0.00403
## 4 Emeryville Town Hall           0.00581
## 5 Foothill Blvd at 42nd Ave      0.00587
## 6 Foothill Blvd at Fruitvale Ave 0.00153
```

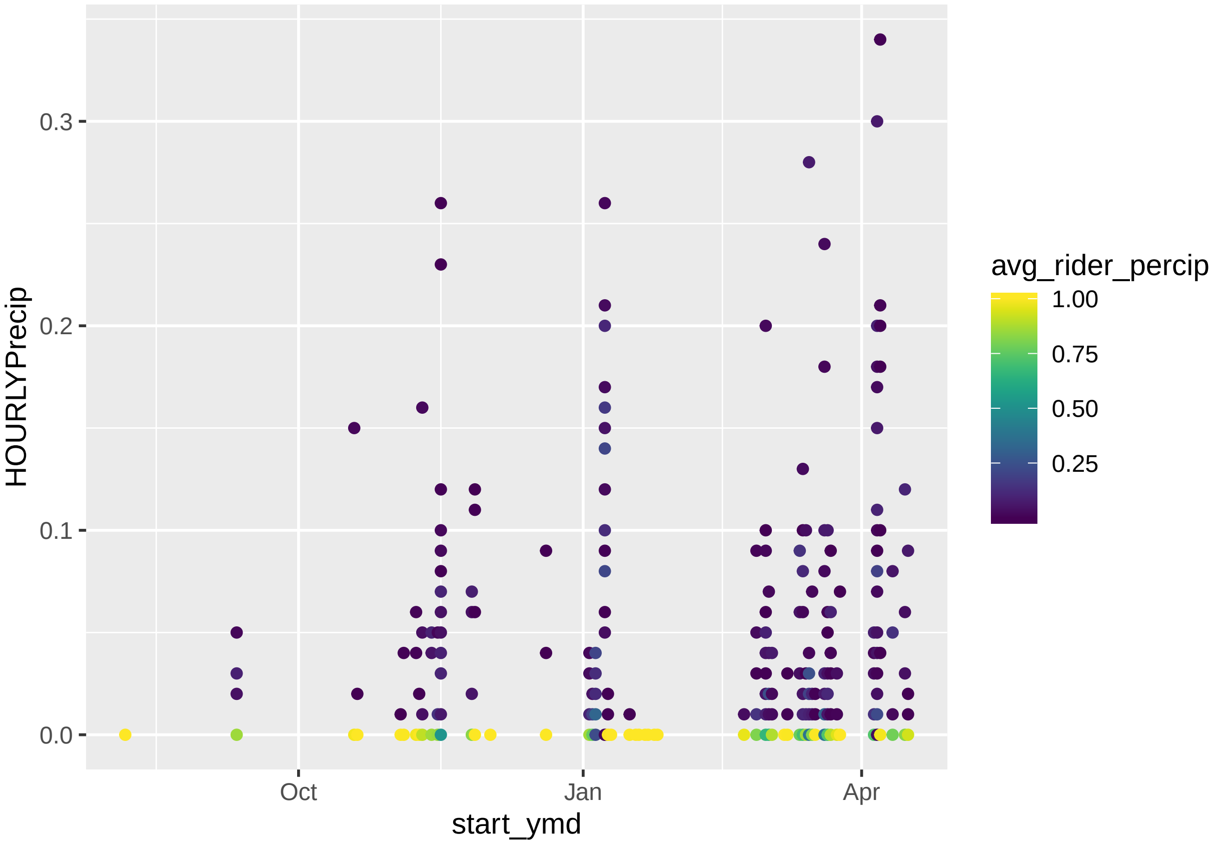
For visualization sake I produced the difference of the 9 stations that are statistically significant with alpha as 0.01. It would be interesting to see if these stations were located near eachother! With more time I would be able to compare the latitude and longitude of these locations and check that. At overview shows that three of them are on Foothill Blvd, which might indicate that they actually are near.

```
# Looking at the average number of riders per month given if it rains or not
avg_diff_stations = diff_stations %>% filter(start_station_name %in% alpha_01$start_station_name) %>%
  group_by(start_station_name) %>%
  summarise(avg_no = mean(No), avg_yes = mean(Yes)) %>%
  gather(., rain_days, avg_riders, avg_no, avg_yes)

# Plotting differences of the 9 most significant station differences
ggplot(avg_diff_stations, aes(x = start_station_name, y = avg_riders, group_by(rain_days))) +
  geom_point(aes(color = rain_days))  + theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Another interesting exploration is whether or not the percentage of riders during a rainly day changes as the percipitation changes. To do this I looked at just the rainy days and then calculated what percentage of the total riders that day were riding given a certain level of percipitation. And as expected the graph does show that as perciptation increases the amount of riders from that day decrease.

```
# Manipulating the data to find the percentage of riders during different levels of percipitation
percip_data = bike_data %>%
  filter(start_ymd %in% rain_days$ymd) %>%
  left_join(., (weather_data %>%
  filter(!is.na(HOURLYPrecip)) %>%
  select(ymd, hour, min ,sec, HOURLYPrecip)), by = c('start_ymd' = 'ymd', 'start_hour' = 'hour')) %>%
  mutate(HOURLYPrecip = ifelse(is.na(HOURLYPrecip), 0, HOURLYPrecip)) %>%
  group_by(start_ymd) %>%
  mutate(count = n()) %>%
  ungroup() %>% group_by(start_ymd, HOURLYPrecip, count) %>%
  summarise(rider_per_percip = n()) %>%
  mutate(avg_rider_percip = rider_per_percip/count)

# Plotting the change when it rains
ggplot(percip_data, aes(start_ymd, HOURLYPrecip)) +
  geom_point(aes(colour = avg_rider_percip)) +
  scale_color_viridis(discrete = FALSE) +
  scale_fill_viridis(discrete = FALSE)
```


HOURLYPrecip
0.3
0.2
0.1
0.0
Oct
Jan
Apr
start_ymd
avg_rider_percip
1.00
0.75
0.50
0.25