Nice Ride Stat Report

Tony Tushar Jr

February 2, 2018

This Exploratory Data Analysis (EDA) serves as a starting point for understanding our dataset prior regression and machine learning applications. A reminder of our two main questions and ultimate purpose of this project:

Two Main Questions: - What is the affect of weather on bikeshare volume? - How does the affect of weather on bikeshare volume differ between member and casual account types?

Ultimate Purpose: - To provide a predictive model for weekly bikeshare volume based on the input of weather forecast variables.

In this EDA we will perform the following steps:

1. General exploration of dataset:

- What is the distribution of trips per month, by year?
- What is the distribution of trips per day of the week, by account type?
- What is the distibution of trips by weekday and weekend for the season, by year, by account type?
- Summary of trip distance in miles
- Summary of trip duration in minutes
- Are there outliers to address? If so, how are they to be addressed?

2. Pearson's product-moment correlations for the affects of primary weather variables on bikeshare volume:

- Affect of average temperature
- Affect of average wind speed
- Affect of precipitation
- Affect of relative humidity
- Affect of heat index

3. T-tests for primary weather variables and relationship to bike count by account type, member and casual:

4. Conclusions, next steps:

Load packages

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------------------- tidyve

## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.3.4     v dplyr   0.7.4
## v tidyr   0.7.2     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.2.0

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

```r
library(readr)
library(ggthemes)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

Load dataset

```r
Rides1617 <- read_csv("Nice_Ride_1617.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   Start_DoWeek = col_character(),
##   Start_Station = col_character(),
##   Start_Latitude = col_double(),
##   Start_Longitude = col_double(),
##   End_Station = col_character(),
##   End_Latitude = col_double(),
##   End_Longitude = col_double(),
##   Total_DurationMin = col_double(),
##   Trip_DistanceMiles = col_double(),
##   Account_Type = col_character(),
##   Avg_Wind = col_double(),
##   Precip = col_double()
## )

## See spec(...) for full column specifications.

## Warning in rbind(names(probs), probs_f): number of columns of result is not
## a multiple of vector length (arg 1)

## Warning: 1974 parsing failures.
## row # A tibble: 5 x 5 col     row   col                expected actual                 file expected
## ... ................. ... ............................................................... ........
## See problems(...) for more details.
```

```r
View(Rides1617)
```

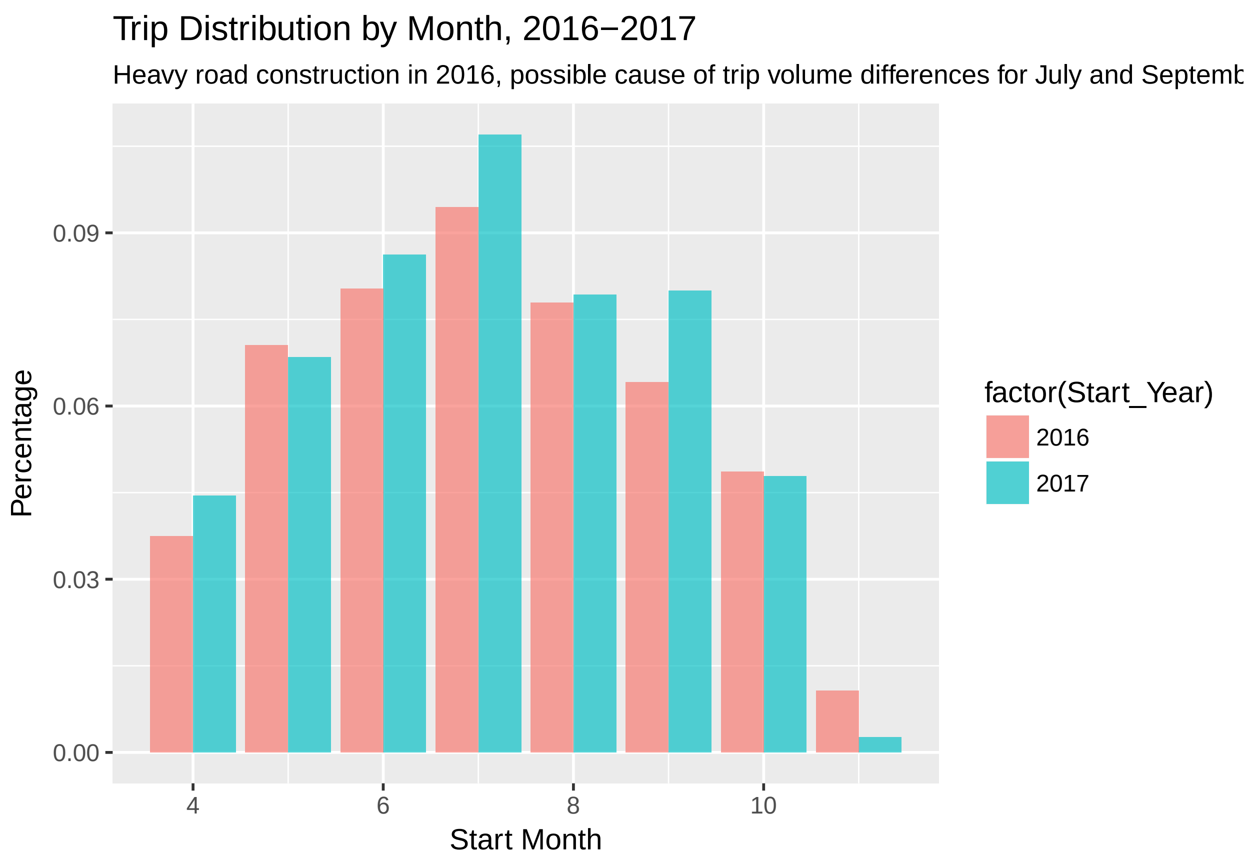
1. General exploration of dataset:

- What is the distribution of trips per month, for 2016-2017 seasons?

```r
#Create data subset
Trips_Month <- Rides1617 %>% count(Start_Year, Start_Month) %>% mutate(Percent_Year = prop.table(n))
head(Trips_Month, n=12)
```

```
## # A tibble: 12 x 4
##    Start_Year Start_Month     n Percent_Year
##         <int>       <int> <int>        <dbl>
##  1       2016           4 33447   0.03745498
##  2       2016           5 62985   0.07053255
##  3       2016           6 71716   0.08030979
##  4       2016           7 84324   0.09442862
##  5       2016           8 69542   0.07787528
##  6       2016           9 57301   0.06416743
##  7       2016          10 43430   0.04863425
##  8       2016          11  9529   0.01067087
##  9       2017           4 39693   0.04444945
## 10       2017           5 61158   0.06848662
## 11       2017           6 77005   0.08623258
## 12       2017           7 95575   0.10702783
```

```r
#Setup plot properties for possible reuse
gg_prop_Month <- ggplot(data = data.frame(), aes(x = Start_Month, y = Percent_Year, fill = factor(Start_
  geom_bar(stat = 'identity', position = 'dodge', alpha = 2/3) + labs(x = "Start Month", y = "Percentage

#Plot data
gg_prop_Month %+% Trips_Month
```



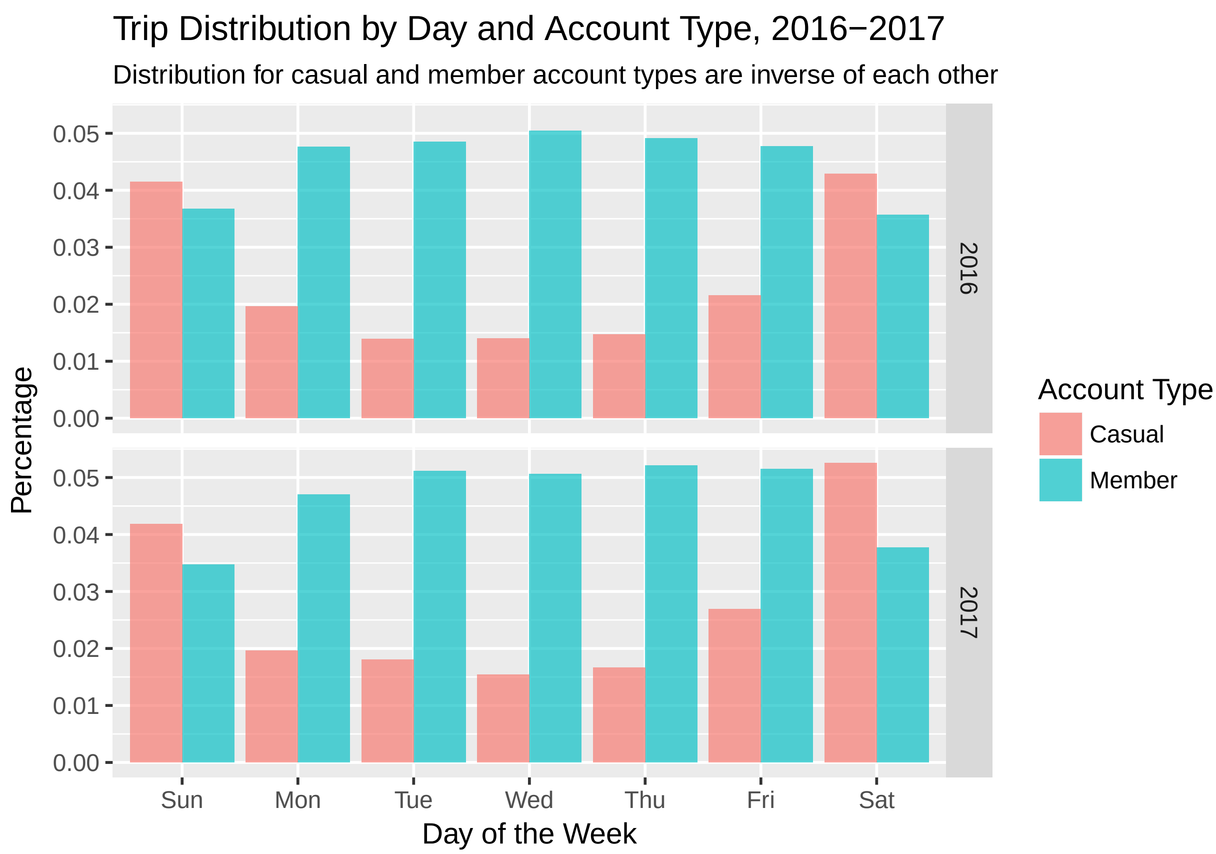
Trip distrubution is fairly even for a year-over-year comparison, however, July and September show larger variation. Research shows that 2016 was a particularly dense season of road construction, as Minneapolis installed biking lanes throughout the metro. This factor is likely a contributor to the variance between 2016 and 2017, however, we cannot control for construction data other than to note here.

- What is the distribution of trips per day of the week, by account type?

```
#Reorder days of the week
Rides1617$Start_DoWeek <- ordered(Rides1617$Start_DoWeek, levels=c("Sun", "Mon", "Tue", "Wed", "Thu",
"Fri", "Sat"))

#Create data subset
Trips_Week <- Rides1617 %>% count(Start_Year, Start_DoWeek, Account_Type) %>% mutate(Percent_Year = prop
head(Trips_Week, n=12)
```

```
## # A tibble: 12 x 5
##    Start_Year Start_DoWeek Account_Type     n Percent_Year
##         <int>        <ord>        <chr> <int>        <dbl>
##  1       2016          Sun       Casual 37048   0.04148749
##  2       2016          Sun       Member 32820   0.03675285
##  3       2016          Mon       Casual 17486   0.01958136
##  4       2016          Mon       Member 42558   0.04765776
##  5       2016          Tue       Casual 12422   0.01391054
##  6       2016          Tue       Member 43336   0.04852899
##  7       2016          Wed       Casual 12531   0.01403260
##  8       2016          Wed       Member 45031   0.05042710
##  9       2016          Thu       Casual 13120   0.01469218
## 10       2016          Thu       Member 43890   0.04914938
## 11       2016          Fri       Casual 19216   0.02151867
## 12       2016          Fri       Member 42644   0.04775407
```

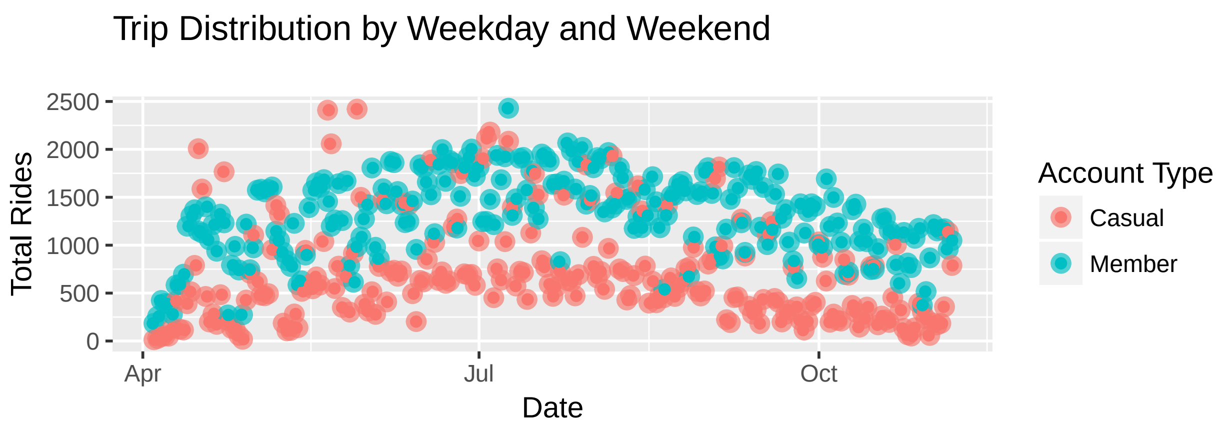
```
#Setup plot properties for possible reuse
gg_prop_Week <- ggplot(data = data.frame(), aes(x = Start_DoWeek, y = Percent_Year, fill = factor(Accou
  geom_bar(stat = 'identity', position = 'dodge', alpha = 2/3) + facet_grid(Start_Year ~ .) + labs(x =

#Plot data
gg_prop_Week %+% Trips_Week
```

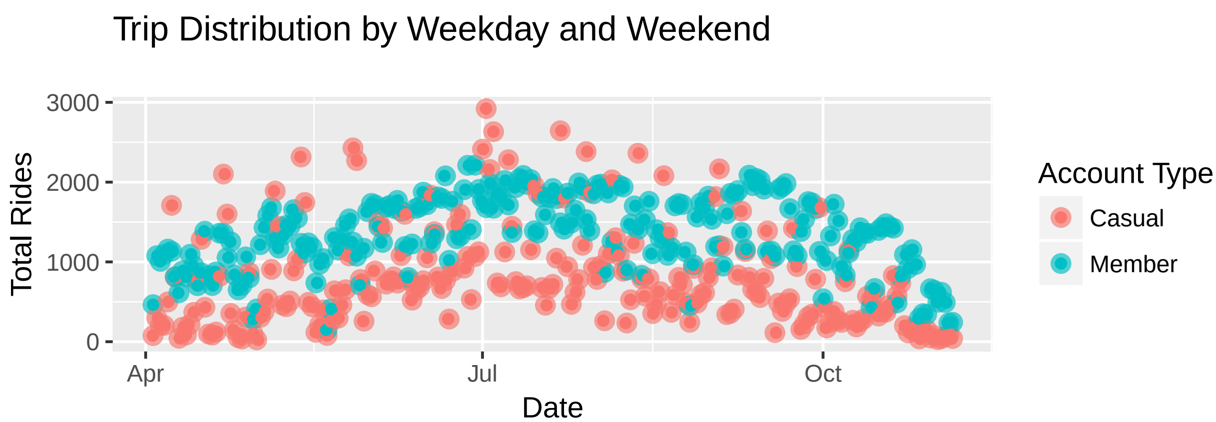


Daily distribution by year confirms consistency year-over-year, member riders are utilizing bikeshare for commuting purposes while casual riders utilize bikeshare for leisure.

- What is the distibution of trips per year by account type?

```
#Create data subset
Trips_Account <- Rides1617 %>% count(Start_Year, Start_Month, Start_Day, Weekend, Account_Type)
Trips_Account16 <- Trips_Account %>% mutate(Dates = as.Date(paste(Start_Year, Start_Month, Start_Day, se
Trips_Account17 <- Trips_Account %>% mutate(Dates = as.Date(paste(Start_Year, Start_Month, Start_Day, se

#Setup plot properties for possible reuse
gg_prop_Account <- ggplot(data = data.frame(), aes(x = Dates, y = n, color=factor(Account_Type))) +
  geom_point(size=3, alpha = 2/3) + geom_jitter() + labs(x = "Date", y = "Total Rides", title="Trip Dis

#Plot data
multiplot(gg_prop_Account %+% Trips_Account16, gg_prop_Account %+% Trips_Account17)
```





- Summary of trip distance

```
summary(Rides1617$Trip_DistanceMiles)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.047   1.944   2.523   3.324  29.046
```

Median trip distance is 1.944 miles while the max is 29 miles, is this logical in comparison to trip duration?

- Summary of trip duration in minutes

```
summary(Rides1617$Total_DurationMin)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##      1.00      6.80     12.72     34.81     24.50 280043.68
```

It is odd to see a max ride observation of over 280,000 minutes, this equates to 194 days! How many bike rides are greater than one day?

- Are there outliers to address? If so, how are they to be addressed?

```
# What percentage of bike rides are over one day in length?
# Trips equal to or less than one day
```

```
Rides_Day <- Rides1617 %>% mutate(Total_DurationDay = (Total_DurationMin <= 24*60))

Day_Length <- nrow(Rides_Day)

sum(Rides_Day$Total_DurationDay/Day_Length)
```

```
## [1] 0.9983225
```

There are 1,498 observations in which ride duration is greater than one day, this accounts for less than 0.02% of our dataset. As our intended outcome of this study is to create a predictive model for daily trips as related to daily weather, we will consider these observations as outliers and remove them.

Remove outliers from dataset

```
Rides1617_Mod <- Rides1617 %>% filter(Total_DurationMin<=24*60)
```

2. Correlation testing for the affects of primary weather variables on bikeshare volume:

- Affect of average temperature

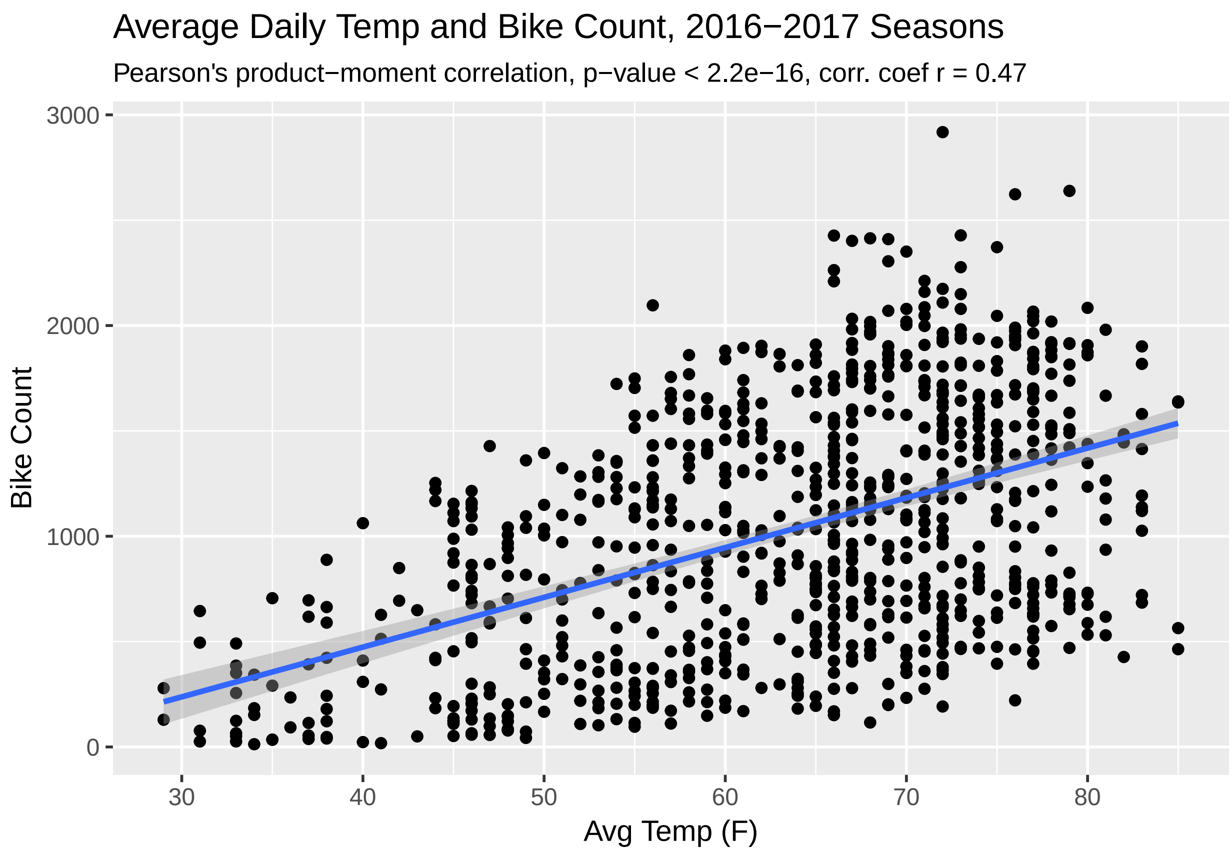
```
# Data subset
Temp <- Rides1617_Mod %>% select(Start_Year:Start_Day, Account_Type, Avg_Temp) %>% group_by(Start_Year,

#Correlation test between daily average temperature and bike count
Cor_Temp <- cor.test(x = Temp$Avg_Temp, y = Temp$n)

Cor_Temp
```

```
##
##  Pearson's product-moment correlation
##
## data:  Temp$Avg_Temp and Temp$n
## t = 15.861, df = 866, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4212174 0.5244458
## sample estimates:
##       cor
## 0.4744612
```

Null hypothesis: the correlation of average daily temperature to daily bike count is 0 Correlation test results show we reject the null hypothesis for 95% confidence in correleation between average daily temperature and bike count

Plot correlation test results

```
plot_cor_temp <- ggplot(Temp, aes(Avg_Temp, n))
plot_cor_temp + geom_point() + geom_smooth(method="lm") + labs(x = "Avg Temp (F)", y = "Bike Count", ti
```



- Affect of average wind speed

```
# Data subset
Wind <- Rides1617_Mod %>% select(Start_Year:Start_Day, Account_Type, Avg_Wind) %>% group_by(Start_Year,

#Correlation test between daily average wind speed and bike count
Cor_Wind <- cor.test(x = Wind$Avg_Wind, y = Wind$n)

Cor_Wind
```

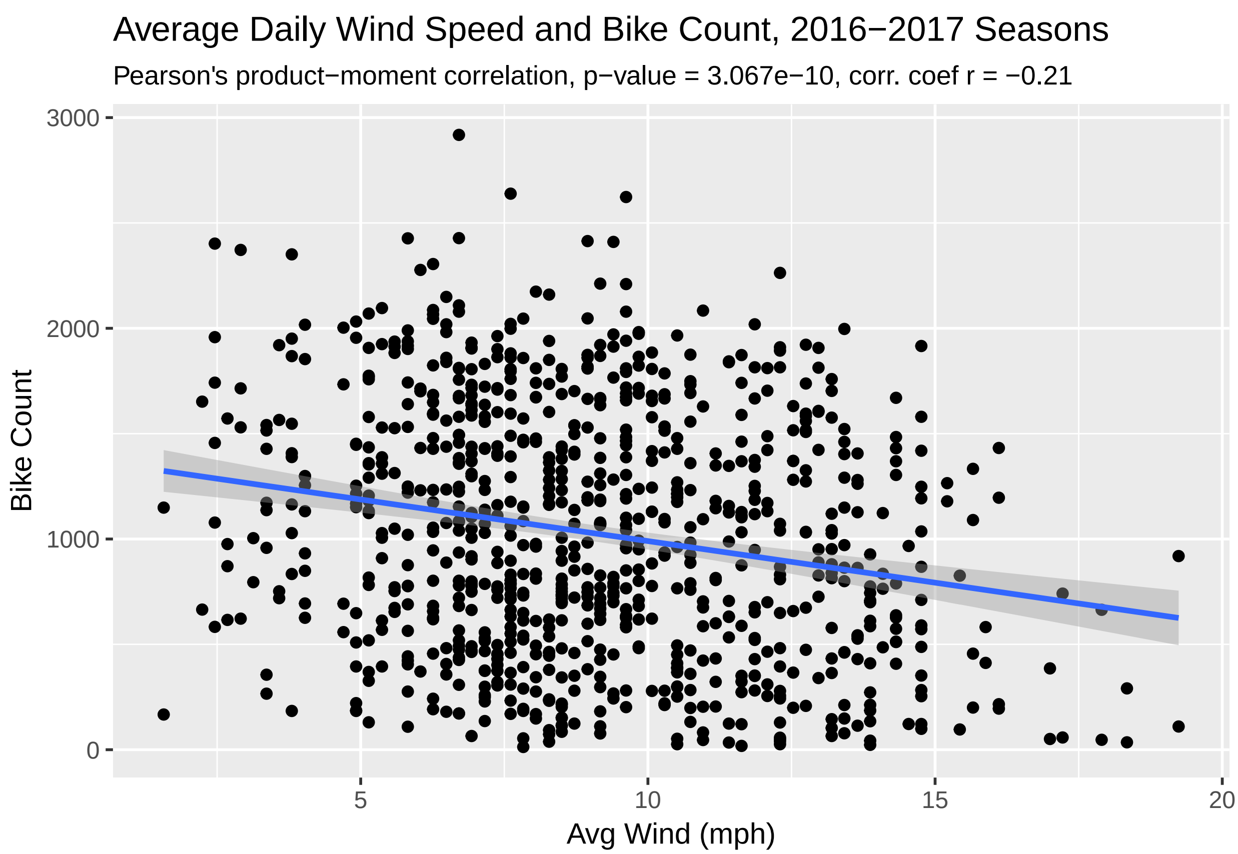
```
##
##  Pearson's product-moment correlation
##
## data:  Wind$Avg_Wind and Wind$n
## t = -6.3699, df = 866, p-value = 3.067e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2742415 -0.1470882
## sample estimates:
```

```
##         cor
## -0.2115599
```

Null hypothesis: the correlation of average daily wind to daily bike count is 0 Correlation test results show we reject the null hypothesis for 95% confidence in a negative correlation between average daily wind speed and bike count

Plot correlation test results

```
plot_cor_wind <- ggplot(Wind, aes(Avg_Wind, n))
plot_cor_wind + geom_point() + geom_smooth(method="lm") + labs(x = "Avg Wind (mph)", y = "Bike Count", t
```



- Affect of precipitation

```
# Data subset
Precip <- Rides1617_Mod %>% select(Start_Year:Start_Day, Account_Type, Precip) %>% group_by(Start_Year,

#Correlation test between daily precipitation and bike count
Cor_Precip <- cor.test(x = Precip$Precip, y = Precip$n)

Cor_Precip
```

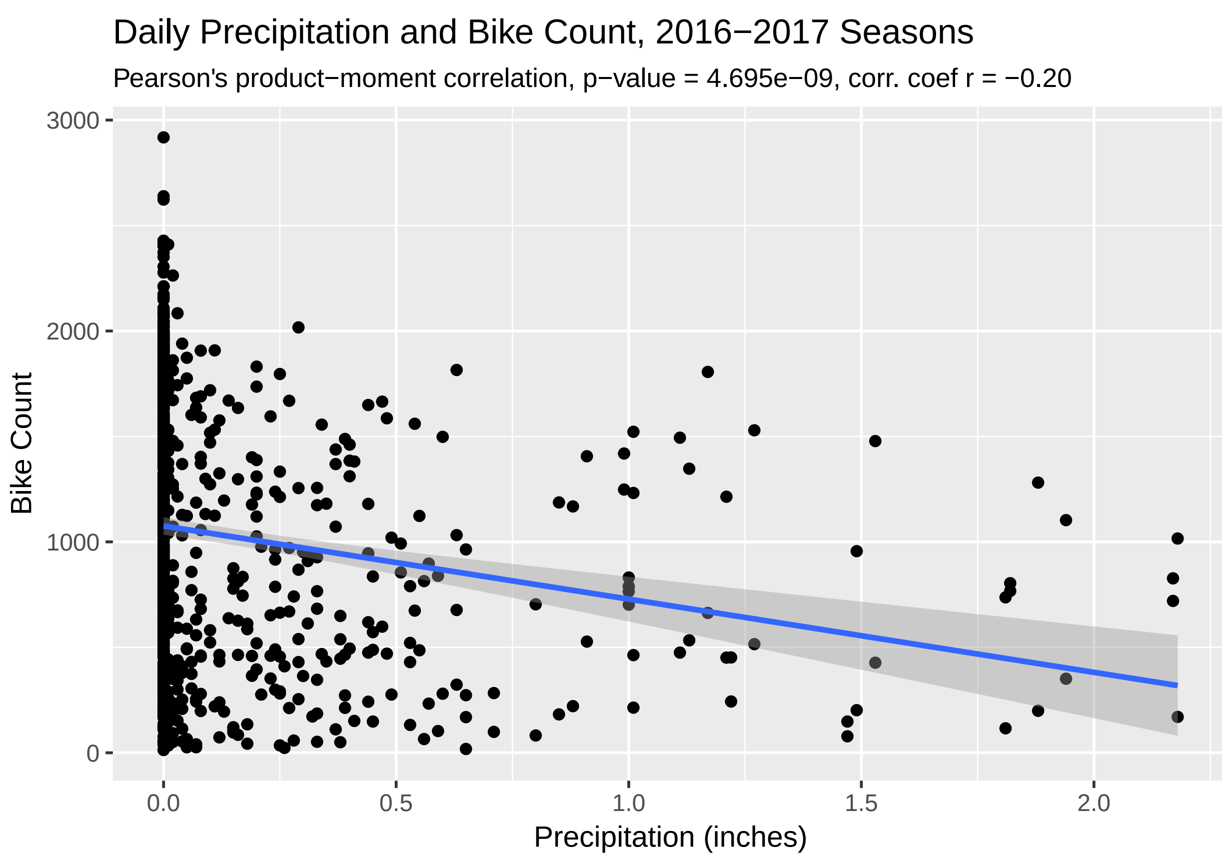
```
##
```

```
##  Pearson's product-moment correlation
##
## data:  Precip$Precip and Precip$n
## t = -5.9179, df = 866, p-value = 4.695e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2602785 -0.1323447
## sample estimates:
##        cor
## -0.1971508
```

Null hypothesis: the correlation of daily precipitation to bike count is 0 Correlation test results show we reject the null hypothesis for 95% confidence in a negative correlation between daily precipitation and bike count.

Plot correlation test results

```r
plot_cor_precip <- ggplot(Precip, aes(Precip, n))
plot_cor_precip + geom_point() + geom_smooth(method="lm") + labs(x = "Precipitation (inches)", y = "Bik
```



- Affect of Relative Humidity
- Affect of Heat Index

Multiplot function