# Capstone Project Statistical Report

*Tony Tushar Jr*

*January 23, 2018*

This is an exploratory data analysis of the 2017 Nice Ride MN bike share season and historical weather records. Much of the analysis is divided between weekdays, weekends, and account types; casual and member. The intended outcome of the EDA is to determine additional work necessary on the dataset prior regression and machine learning applications, and to spot patterns and trends in how casual and member bike use differs throughout the season.

### Load packages

```
library(readr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.2.1 --

## 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 ggplot2 2.2.1     v forcats 0.2.0

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

```
library(ggthemes)
library(ggmap)
```

### Load dataset ***TBD***

```
load("RidesWorkSpace.RData")
Rides <- Nice_Ride_Updated
```

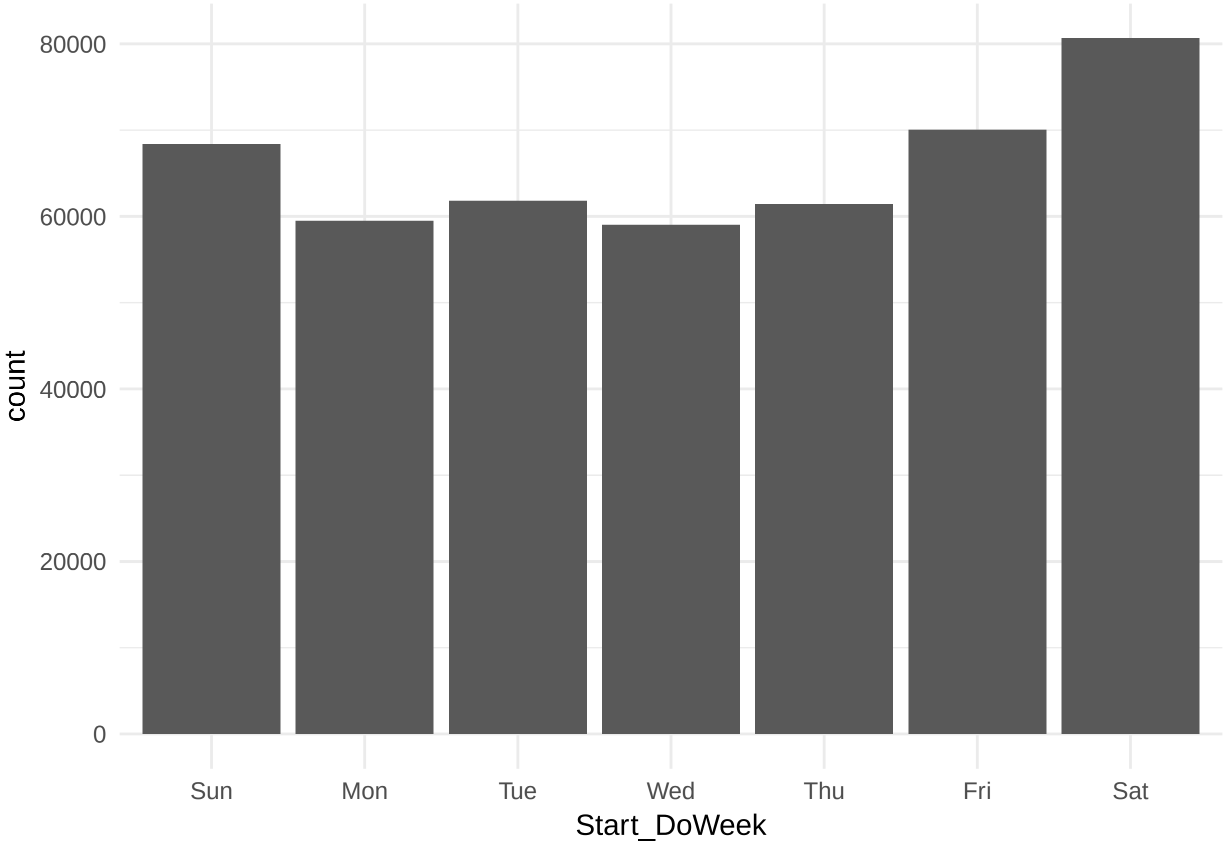
### Load theme

```
theme_set(theme_minimal())
```

## Bike share number of trips distributed by day of the week, hour of the day, and account type

### Distribution of trips by day of the week

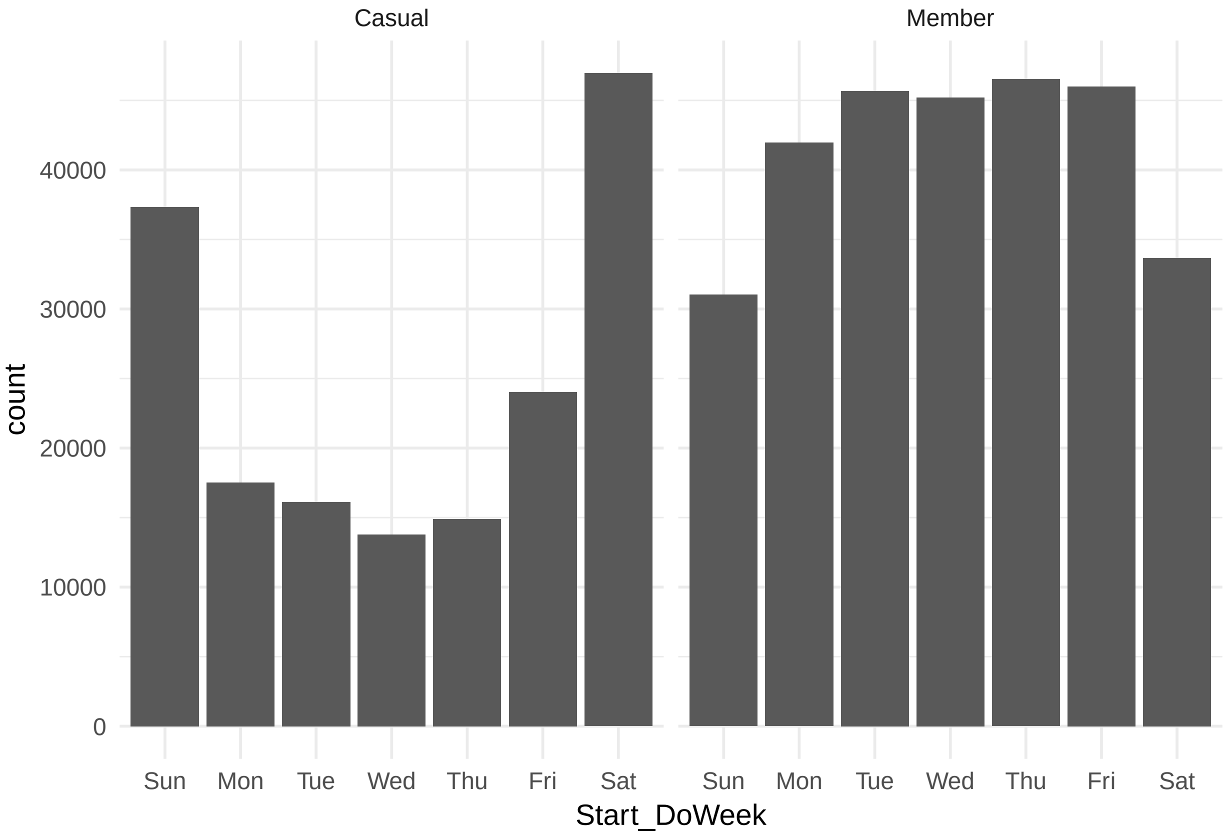
```
Rides %>% ggplot(aes(Start_DoWeek)) + geom_bar()
```



The greatest trip volume is on Saturday, followed by Friday and Sunday. The lowest trip volume is on Wednesday during the workweek, perhaps riders that use the bike share for commuting to and from work use other transportation in the middle of the work week?

### Distribution of trips by day of the week by account type

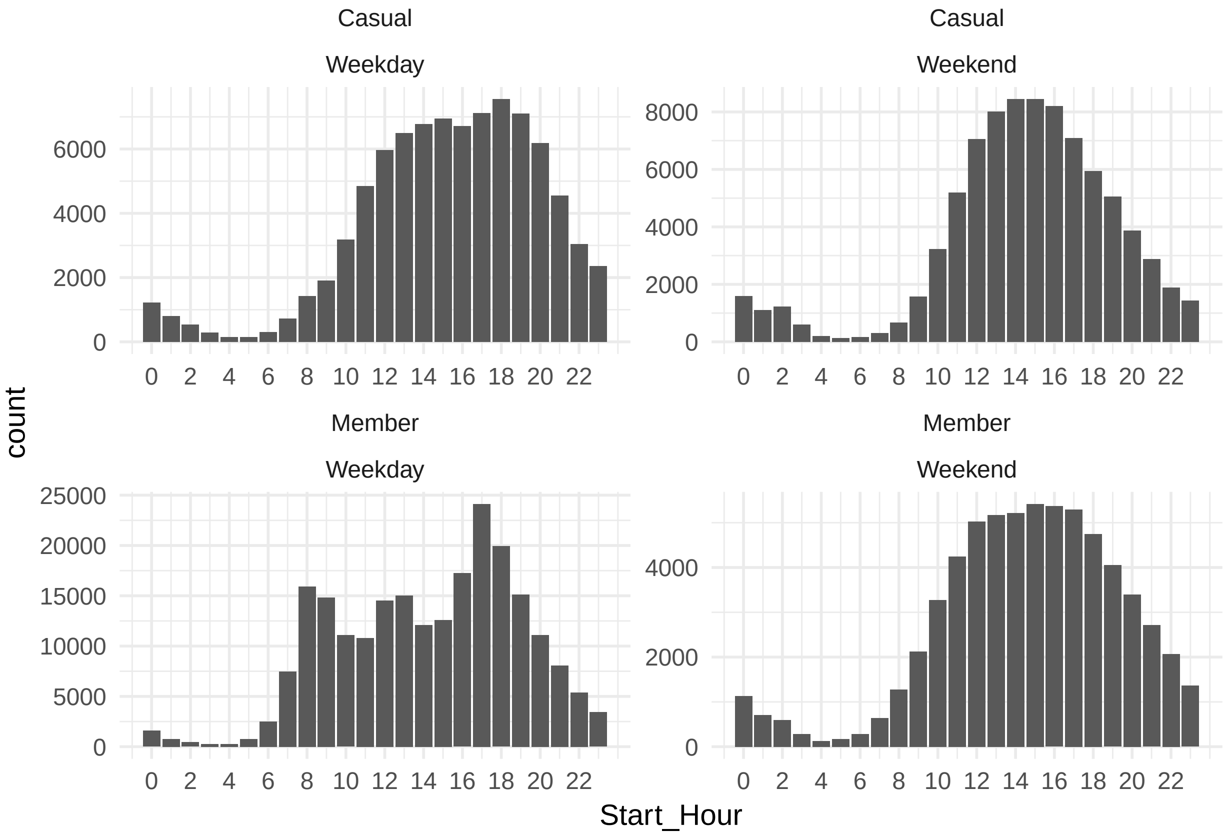
```
Rides %>% ggplot(aes(x = Start_DoWeek)) + geom_bar() + facet_grid(~ Account_Type)
```



Member trips occur more often on weekdays, likely due to work commutes, while casual trips occur on the weekends.

### Distribution of trips by hour of weekdays and weekends by account type

```
Rides %>% ggplot(aes(Start_Hour)) + geom_bar() + facet_wrap(Account_Type ~ StartWeek_Day_End, scales =
```

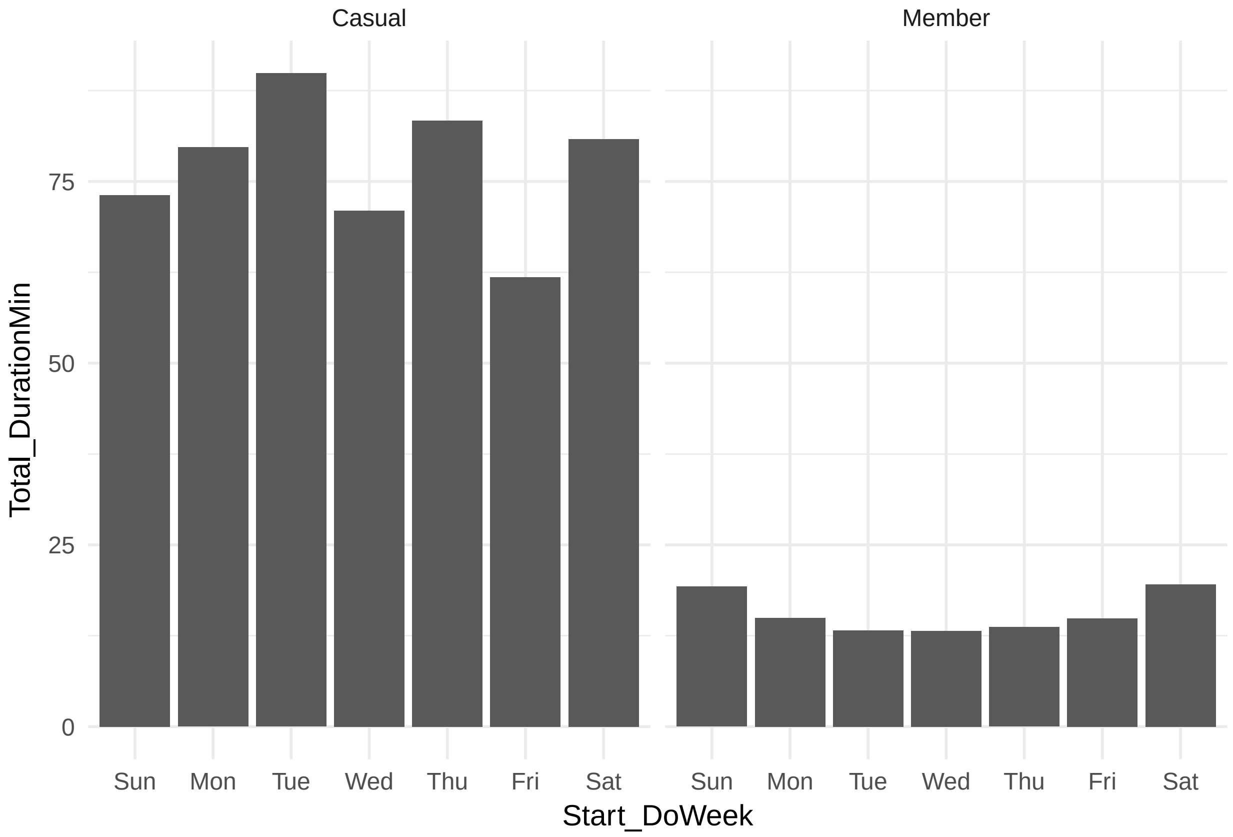


Separating weekdays and weekends confirms the surge in trip volume during the end of the workday between 5 - 6 PM. Factoring weekday and weekend also shows the surge in work commutes from 7 - 10 AM on the weekdays. The weekend shows a pattern of more trips taken from 12 - 7 PM.

## Observing bike trip duration and distance variables

### Mean duration of trip by day of week

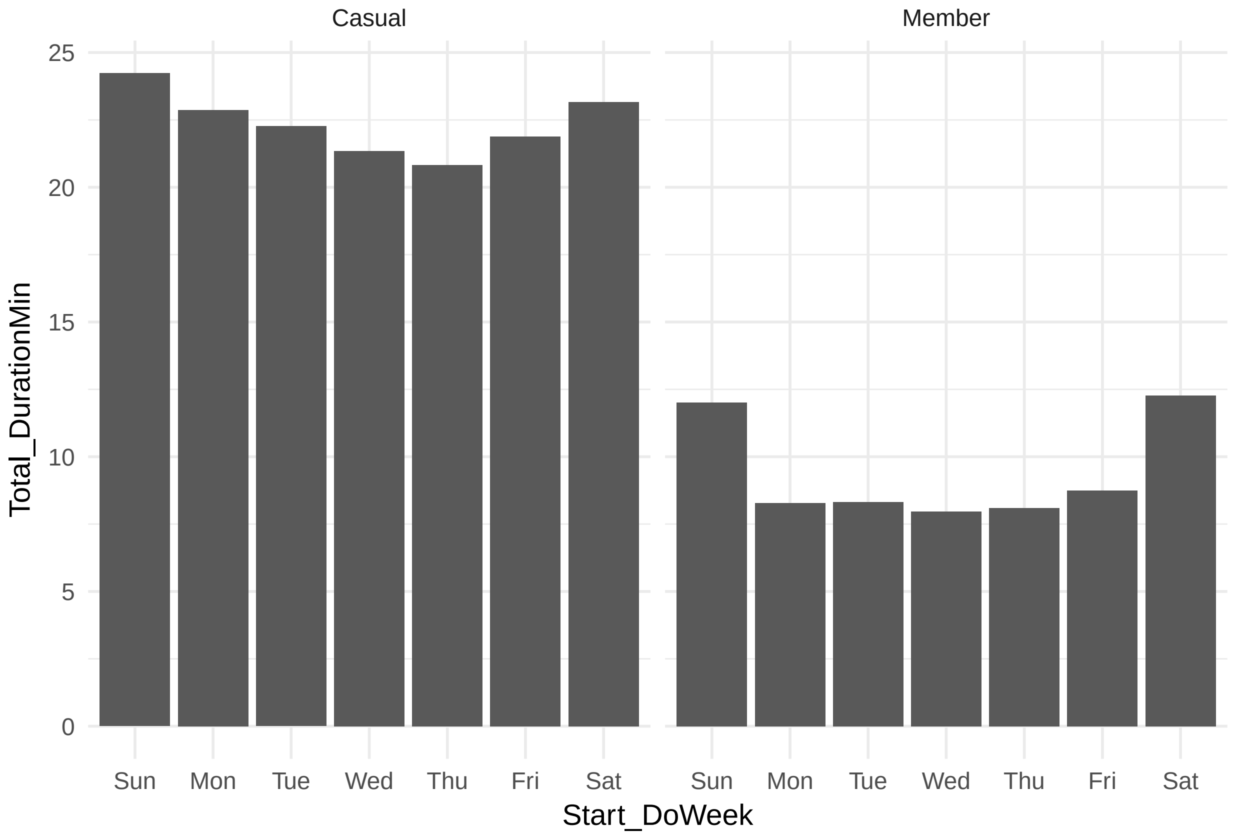
```
Rides %>% group_by(Start_DoWeek, Account_Type) %>% summarise(Total_DurationMin = mean(Total_DurationMin
```



Casual trips have a less uniform distribution of the mean trip duration per day, while member trips have a more bowl-shaped distribution from Sunday - Saturday. The mean for member trips is under 25 minutes on all seven days of the week, while the mean for casual trips is between 60-75+ minutes each day. *Should we question the integrity of our dataset from this perspective? Do we have outliers to address?*

### Median duration of trip by day of week

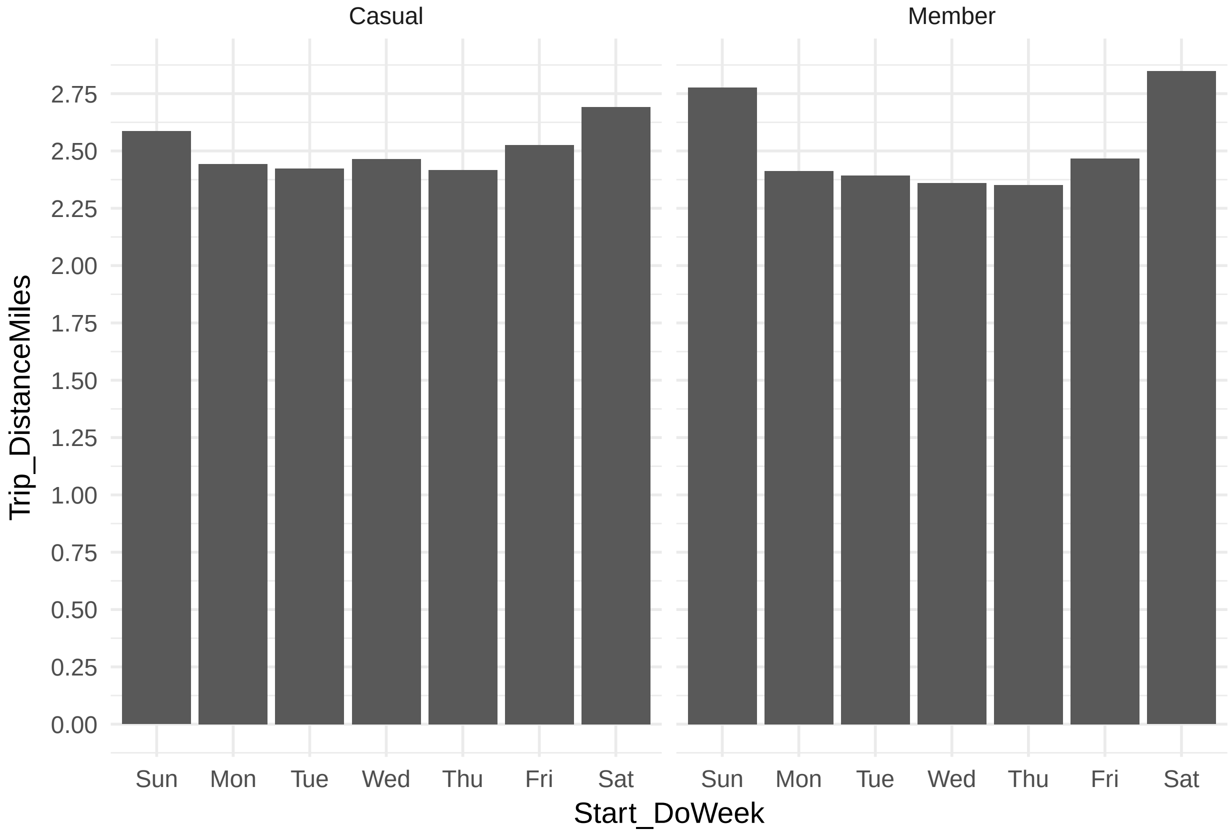
```
Rides %>% group_by(Start_DoWeek, Account_Type) %>% summarise(Total_DurationMin = median(Total_DurationMi
```



Visualizing the median trip duration by day for casual and member account types shows that we have outliers in our dataset to address. The median trip duration for a casual rider is between 20-25 minutes with a similar bowl-shaped distribution for the week. Member riders maintain a similar bowl-shaped distribution and have a median trip duration ranging from 7.5-12.5 minutes depending on the day. *This confirms we have outliers to address prior further analysis.*

### Mean distance of trip by day of week

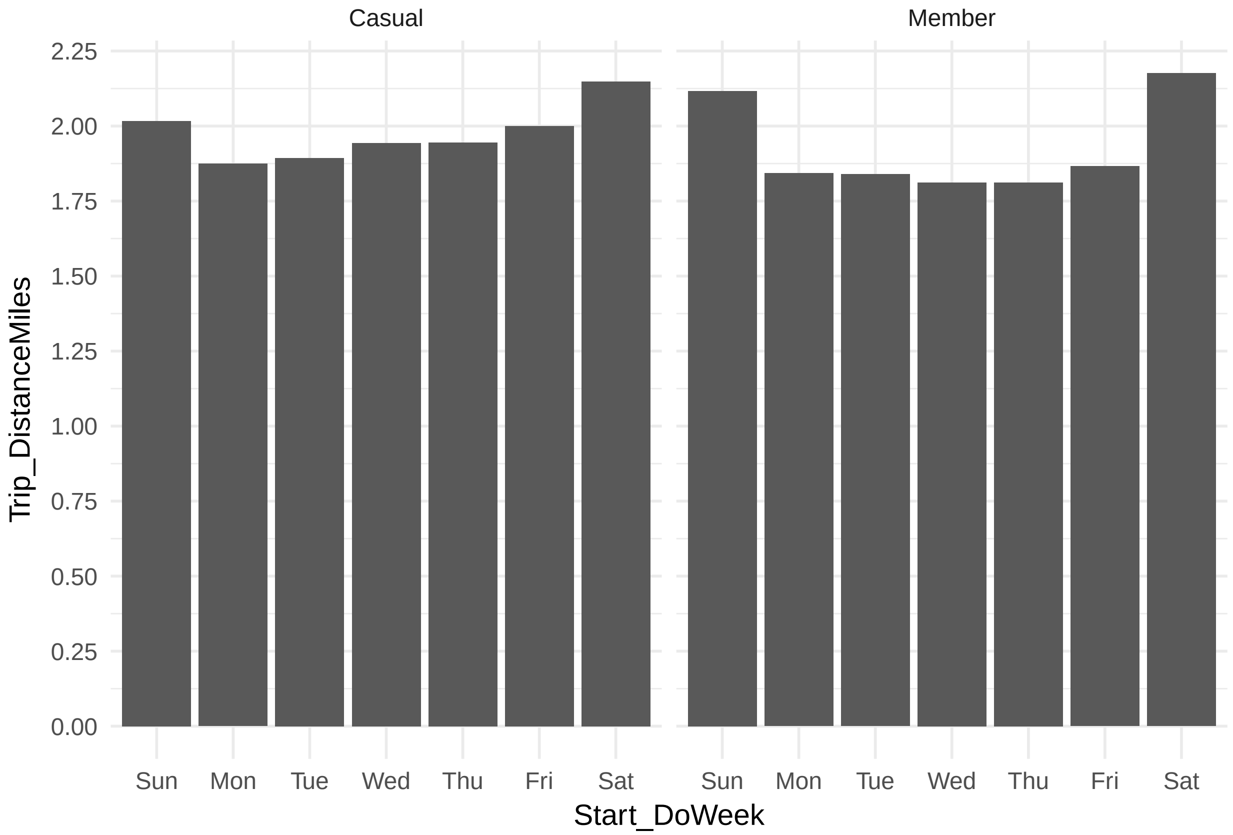
```
Rides %>% group_by(Start_DoWeek, Account_Type) %>% summarise(Trip_DistanceMiles = mean(Trip_DistanceMile
```



The mean trip distance in miles per day sorted by account type shows a similar distribution as noticed for trip duration, though the casual trip distances do not seem skewed by outliers as noted for duration. Casual rides have a mean range between 2.25 and 2.75 miles depending on the day, while member account types have a wider range from 2.25 to 3 miles.

### Median distance of trip by day of week

```
Rides %>% group_by(Start_DoWeek, Account_Type) %>% summarise(Trip_DistanceMiles = median(Trip_DistanceMi
```

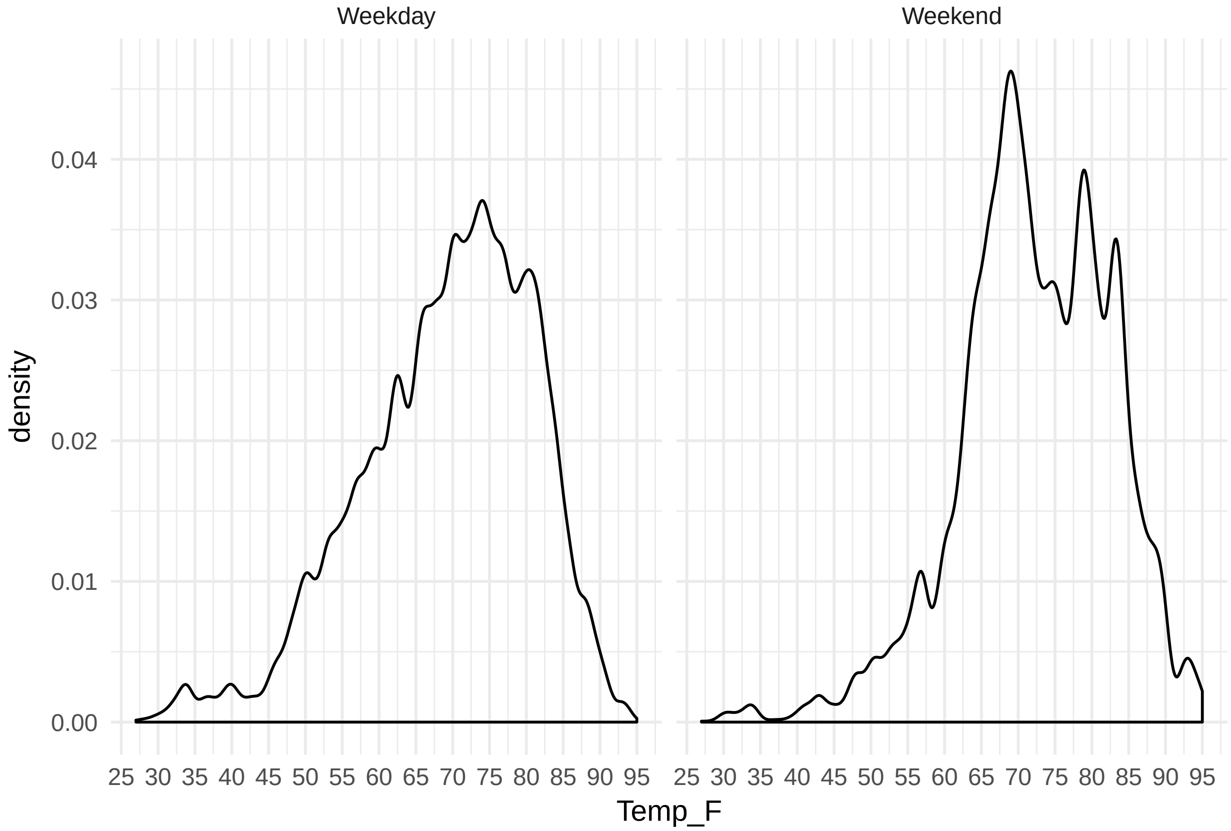


The median trip distance in miles per day holds a similar result to the mean distance, though the range per day is shorter with casual and member account falling in the same range of 1.75 to 2.25 miles.

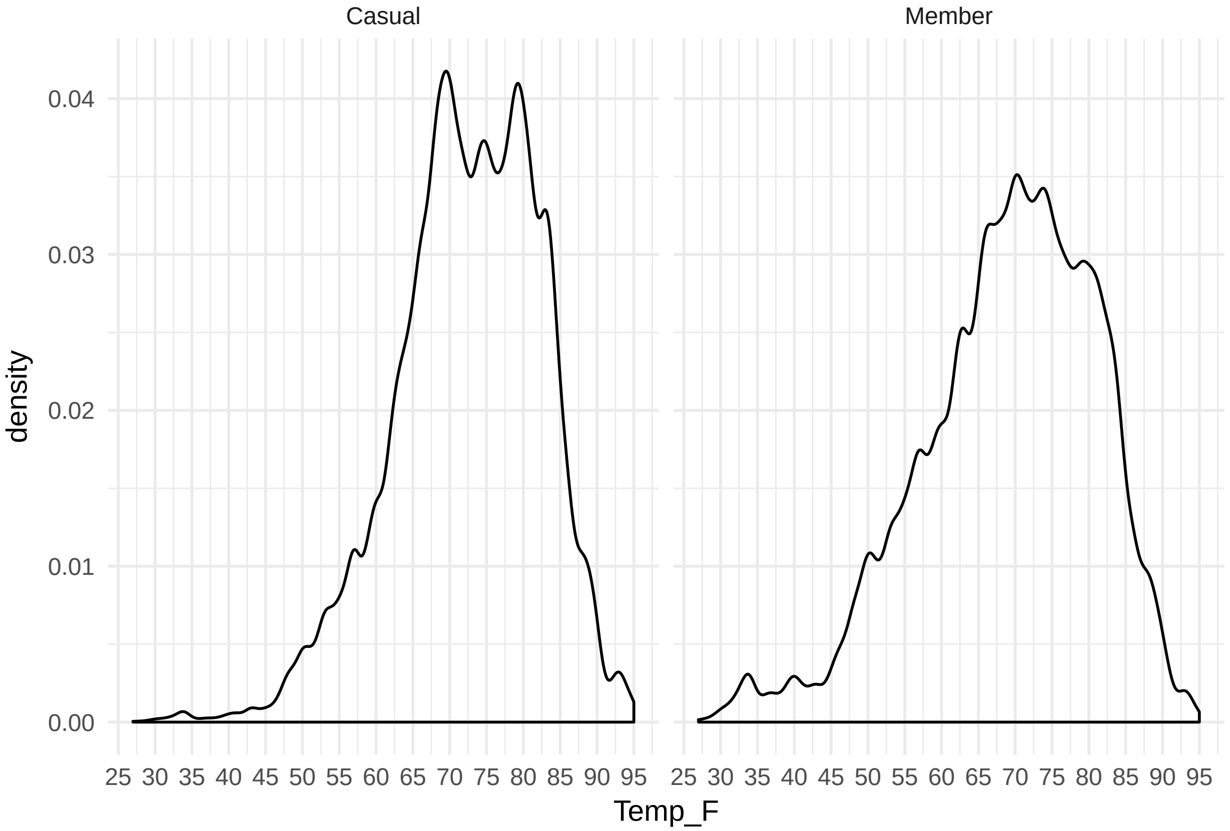
## Observing weather variable affects on bike use

### How does temperature affect bike use?

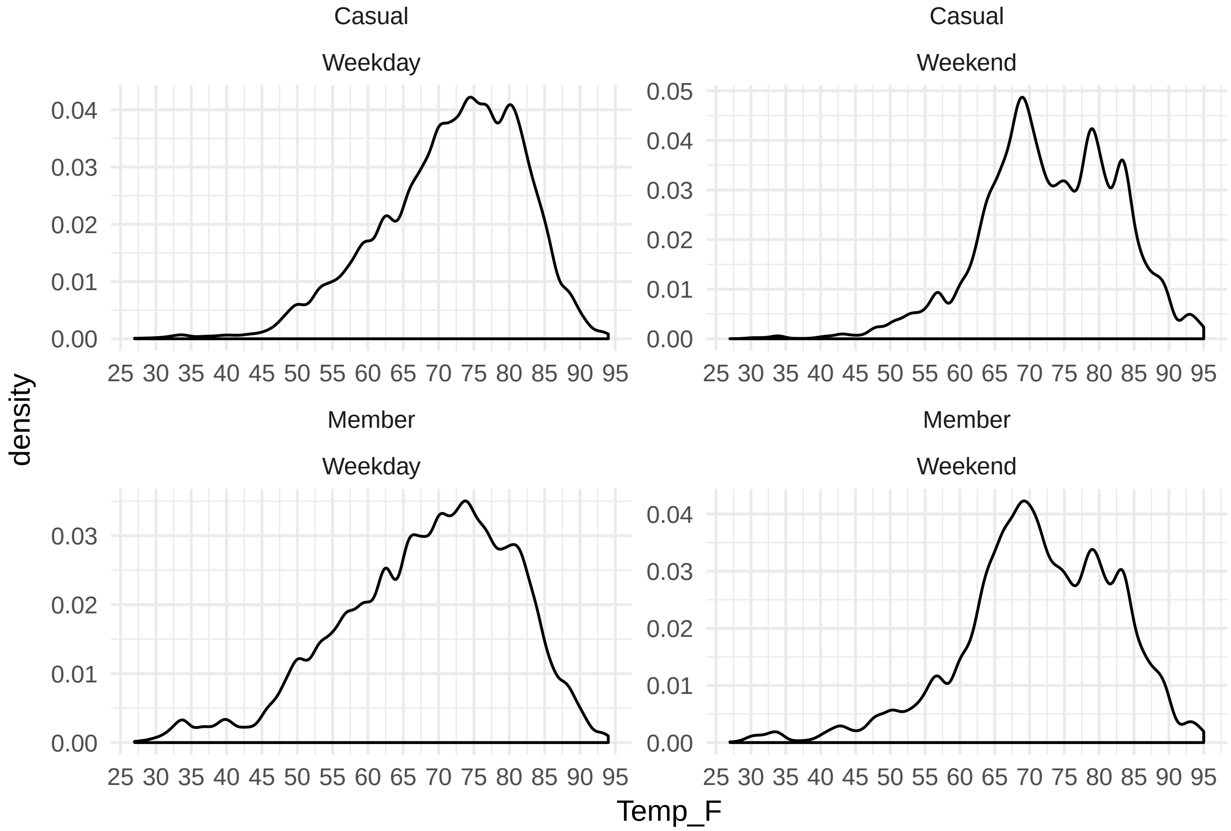
```
#Distribution of trips in relation to temperature by weekday and weekend
Rides %>% ggplot(aes(Temp_F)) + geom_density() + facet_wrap(~ StartWeek_Day_End) + scale_x_continuous(bi
```



```
#Distribution of trips in relation temperature by member or casual
Rides %>% ggplot(aes(Temp_F)) + geom_density() + facet_wrap(~ Account_Type) + scale_x_continuous(breaks
```



```
#Distribution of trips in relation to temperature by weekday and weekend and member or casual
Rides %>% ggplot(aes(Temp_F)) + geom_density() + facet_wrap(Account_Type ~ StartWeek_Day_End, scale = "
```

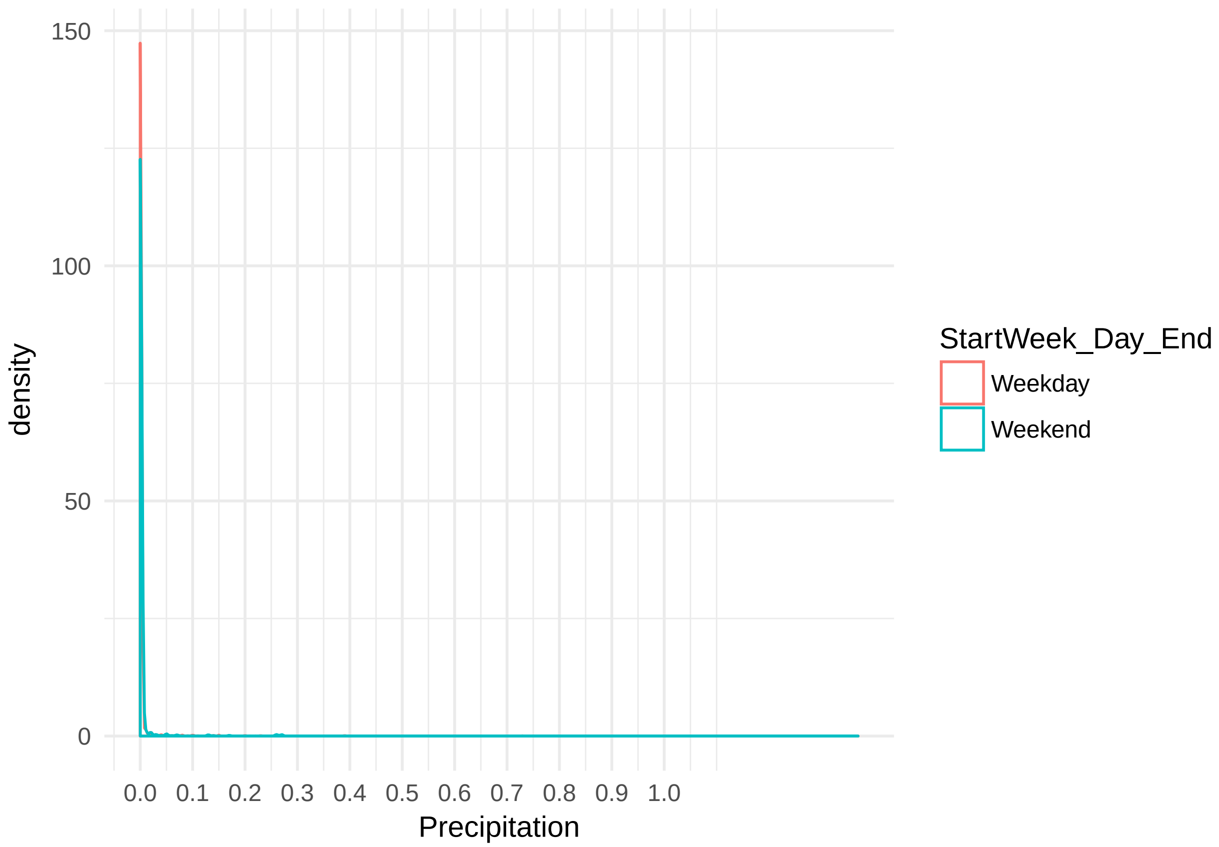


Visualizing weekday and weekend trips by account type in relation to the dry bulb temperature in Fahrenheit reading shows similarity between casual and member riders for weekdays and the same for weekends. Riders are willing to use the bike share at a higher range of temperatures during the weekdays than they are on the weekend. The majority of rides for weekdays occur between 70 and 82.5 degrees fahrenheit while rides on weekends peak at 68 degrees and steadily decline there afterward, though there are increases in volume in the high 70's and low 80's. For colder temperatures casual riders participate at minimal levels until temperatures approach the low 50's, while member riders show some bike share use in cold temperatures and a strong increase starting in the mid-40's.

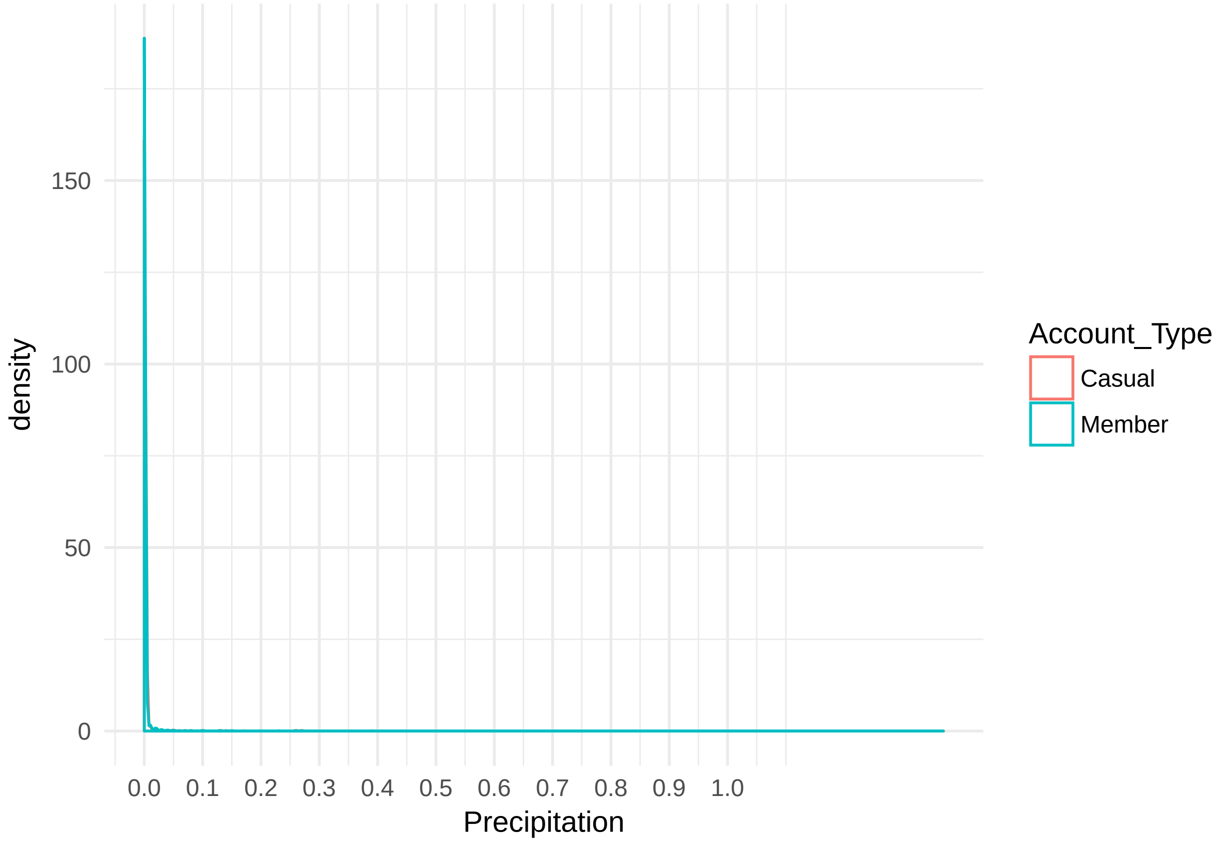
### How does precipitation affect bike use?

*This visualization doesn't quite capture it*

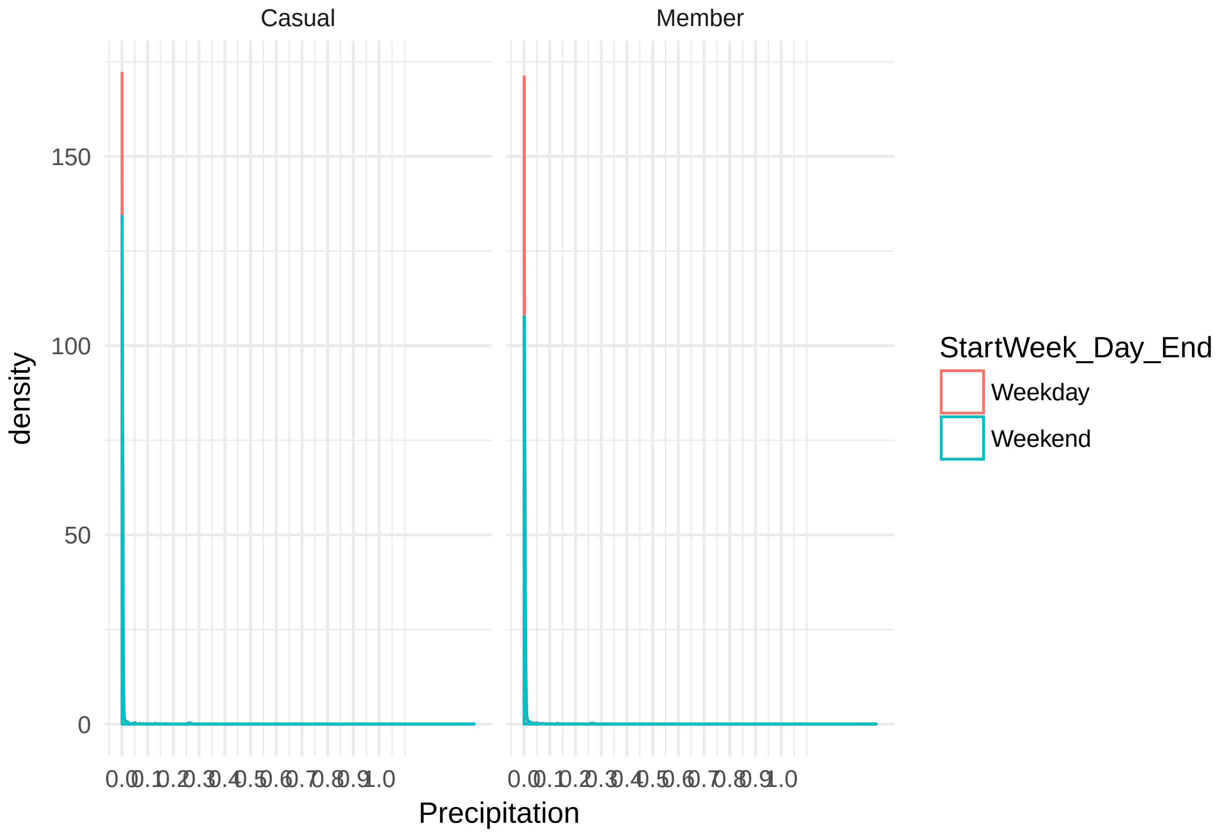
```
#Distribution of trips in relation to precipitation by weekday and weekend
Rides %>% ggplot(aes(Precipitation, color = StartWeek_Day_End)) + geom_density() + scale_x_continuous(bi
```



```
#Distribution of trips in relation to precipitation by member or casual
Rides %>% ggplot(aes(Precipitation, color = Account_Type)) + geom_density() + scale_x_continuous(breaks
```



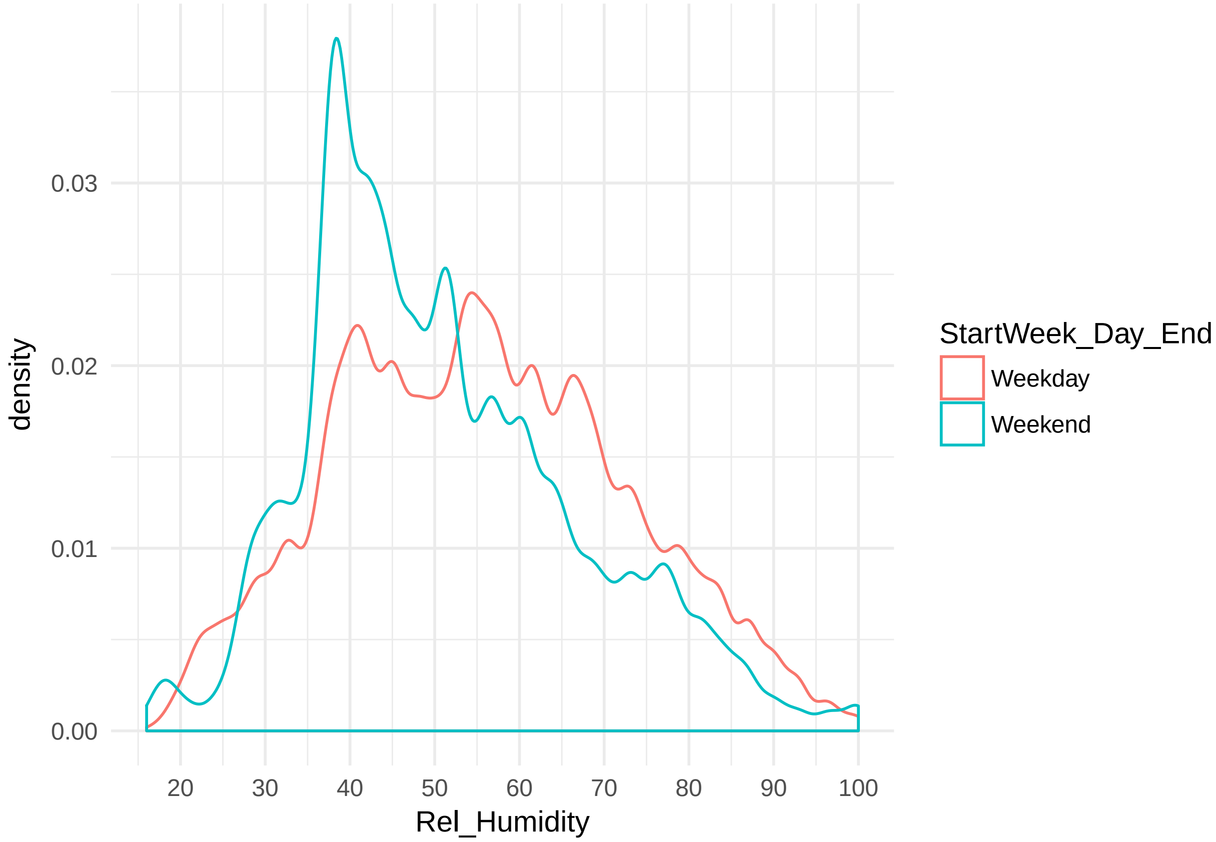
```
#Distribution of trips in relation to precipitation by weekday and weekend and member or casual
Rides %>% ggplot(aes(Precipitation, color = StartWeek_Day_End)) + geom_density() + facet_wrap(~ Account_
```



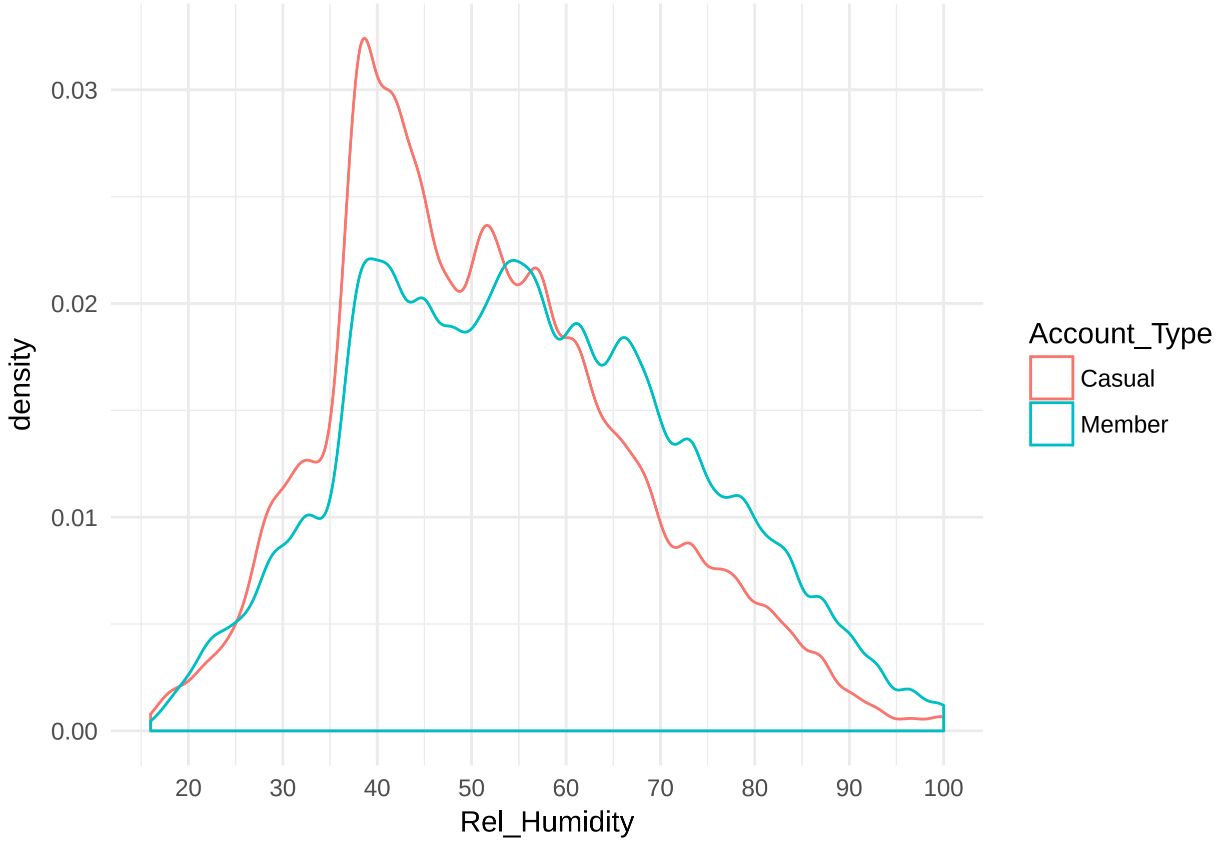
Precipitation play a large factor in bike use for both casual and member account types and greatly declines for precipitation levels beyond 0.1 inches.

### How does humidity affect bike use?

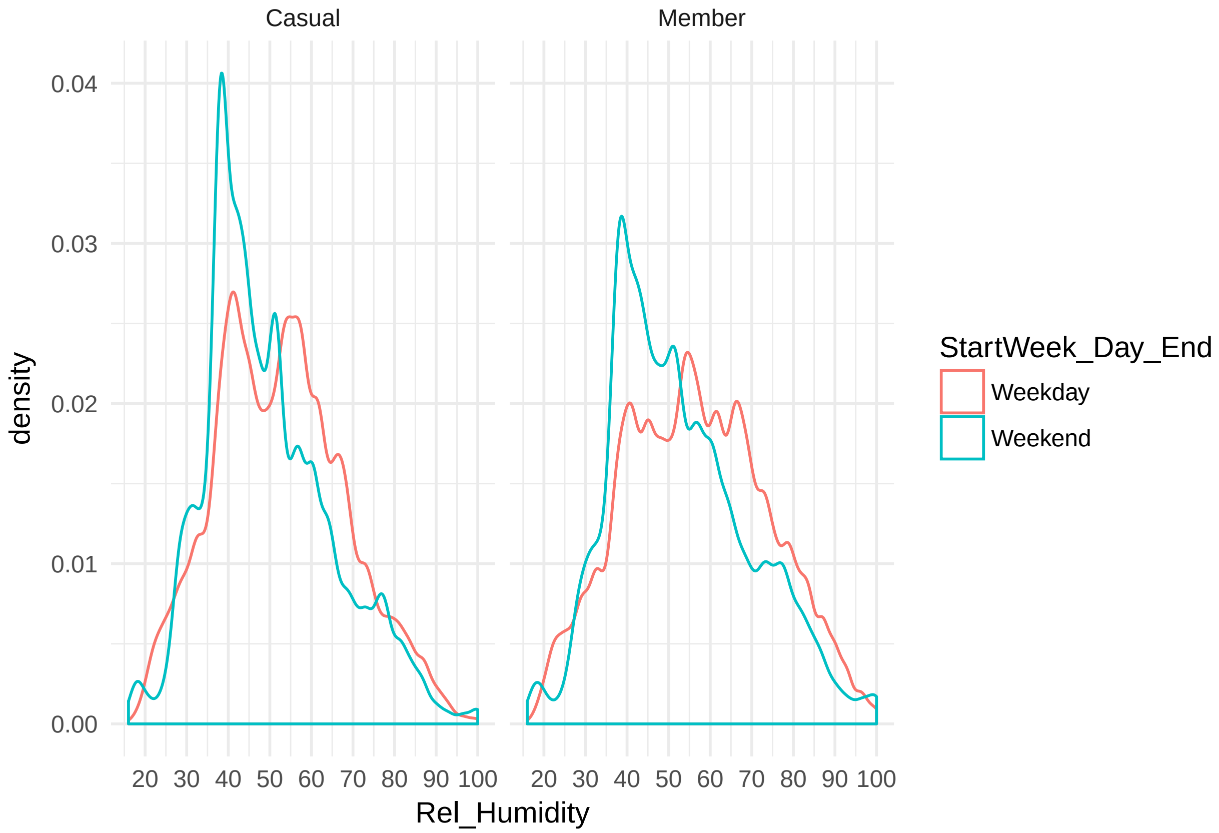
```
#Distribution of trips in relation to humidity by weekday and weekend
Rides %>% ggplot(aes(Rel_Humidity, color = StartWeek_Day_End)) + geom_density() + scale_x_continuous(bre
```



```
#Distribution of trips in relation humidity by member or casual
Rides %>% ggplot(aes(Rel_Humidity, color = Account_Type)) + geom_density() + scale_x_continuous(breaks
```



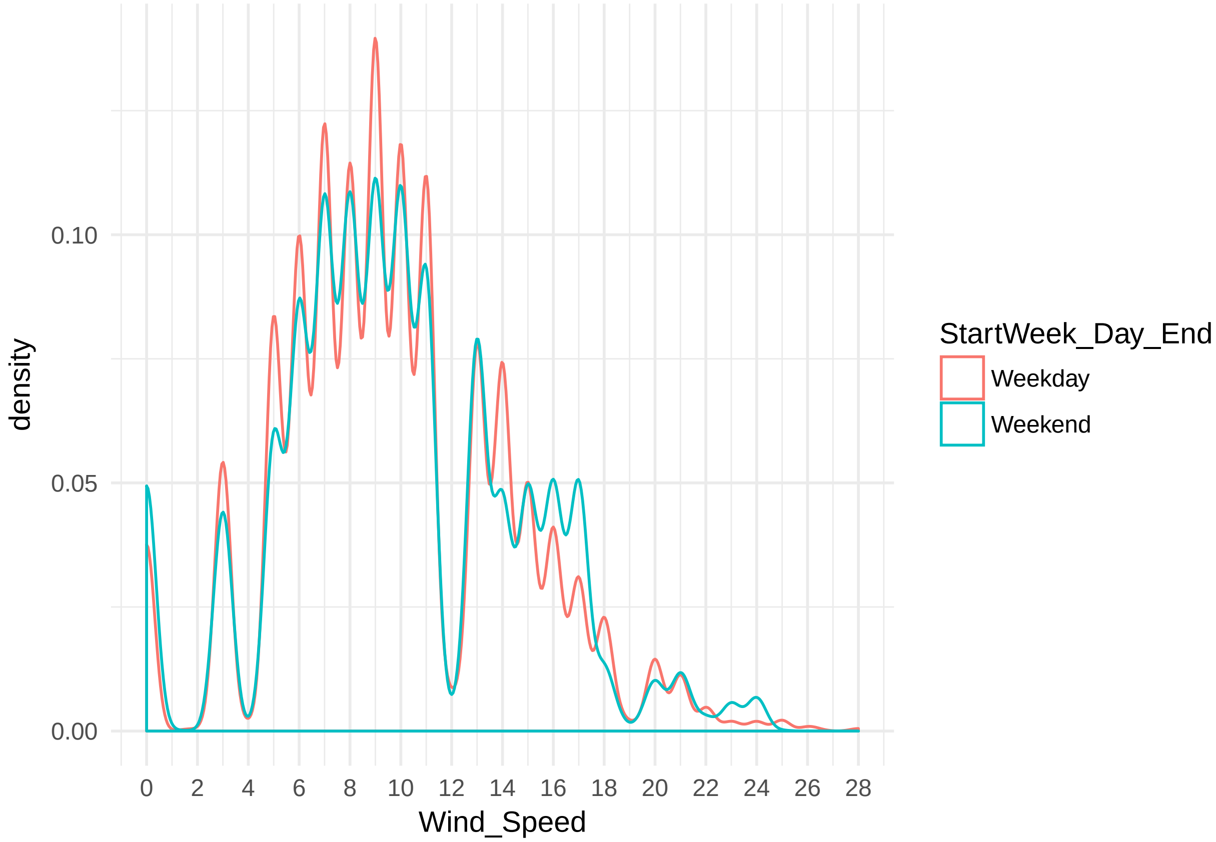
```
#Distribution of trips in relation to humidity by weekday and weekend and member or casual
Rides %>% ggplot(aes(Rel_Humidity, color = StartWeek_Day_End)) + geom_density() + facet_wrap(~ Account_
```



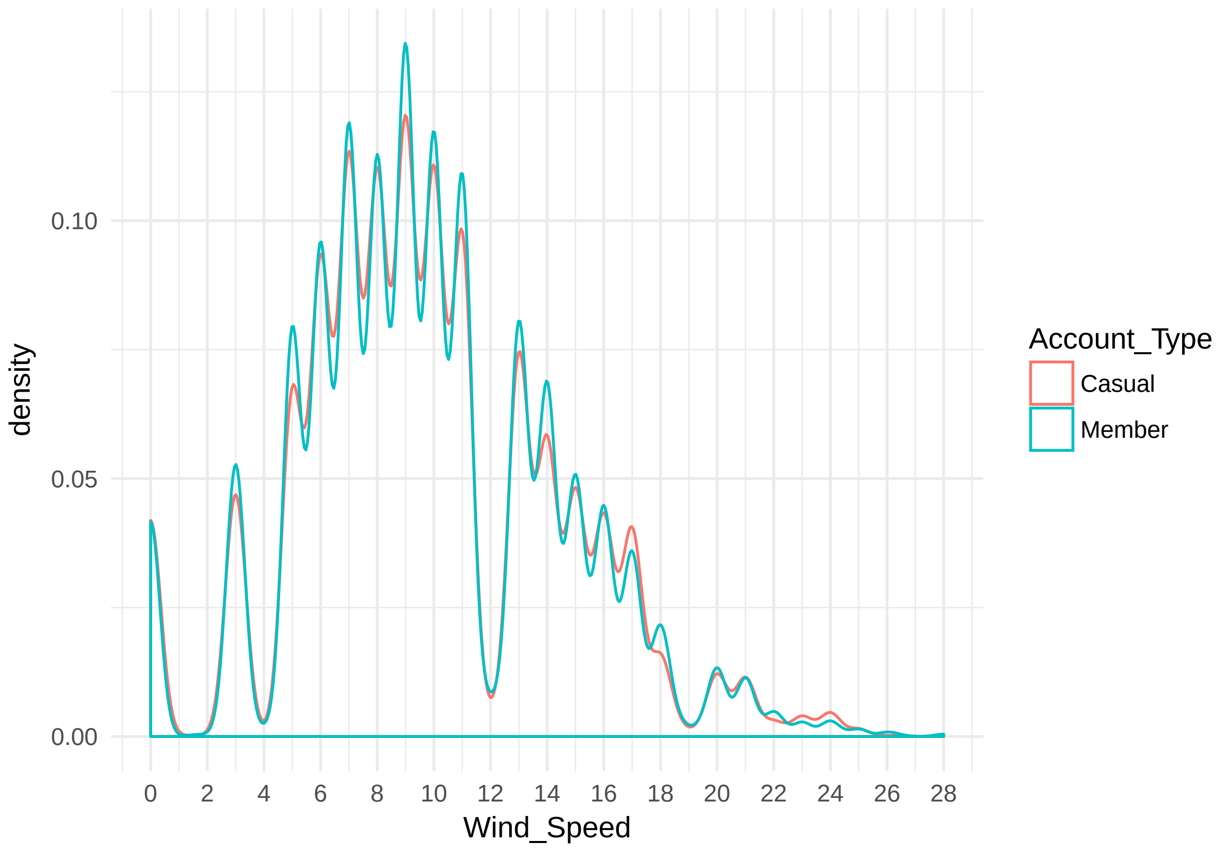
As relative humidity increases bike use declines for both account types. Both casual and member accounts show a comparable response to relative humidity during the weekday, though casual riders show a faster decline in bike usage as humidity rises. Weekend patterns show a similar response to humidity for casual and member riders.

### How does wind speed affect bike use?

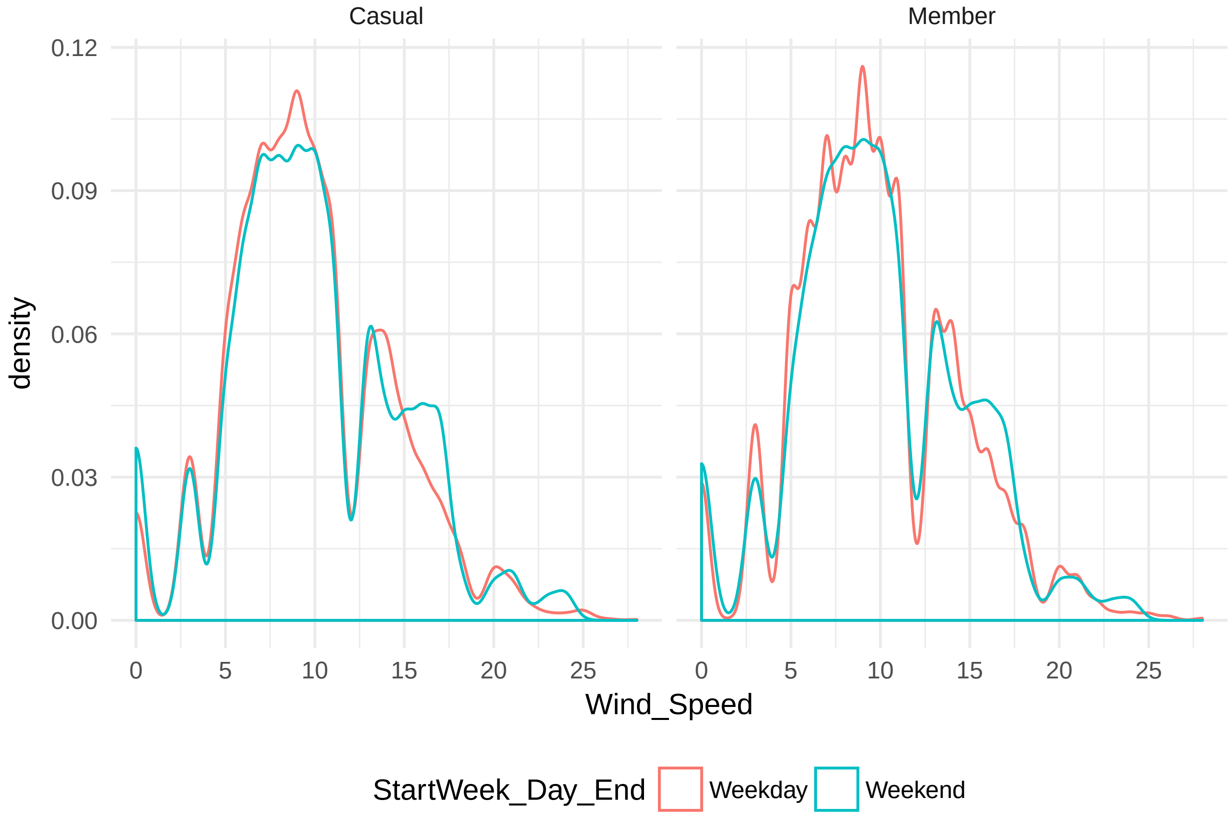
```
#Distribution of trips in relation to wind speed by weekday and weekend
Rides %>% ggplot(aes(Wind_Speed, color = StartWeek_Day_End)) + geom_density() + scale_x_continuous(brea
```



```
#Distribution of trips in relation to wind speed by member or casual
Rides %>% ggplot(aes(Wind_Speed, color = Account_Type)) + geom_density() + scale_x_continuous(breaks =
```



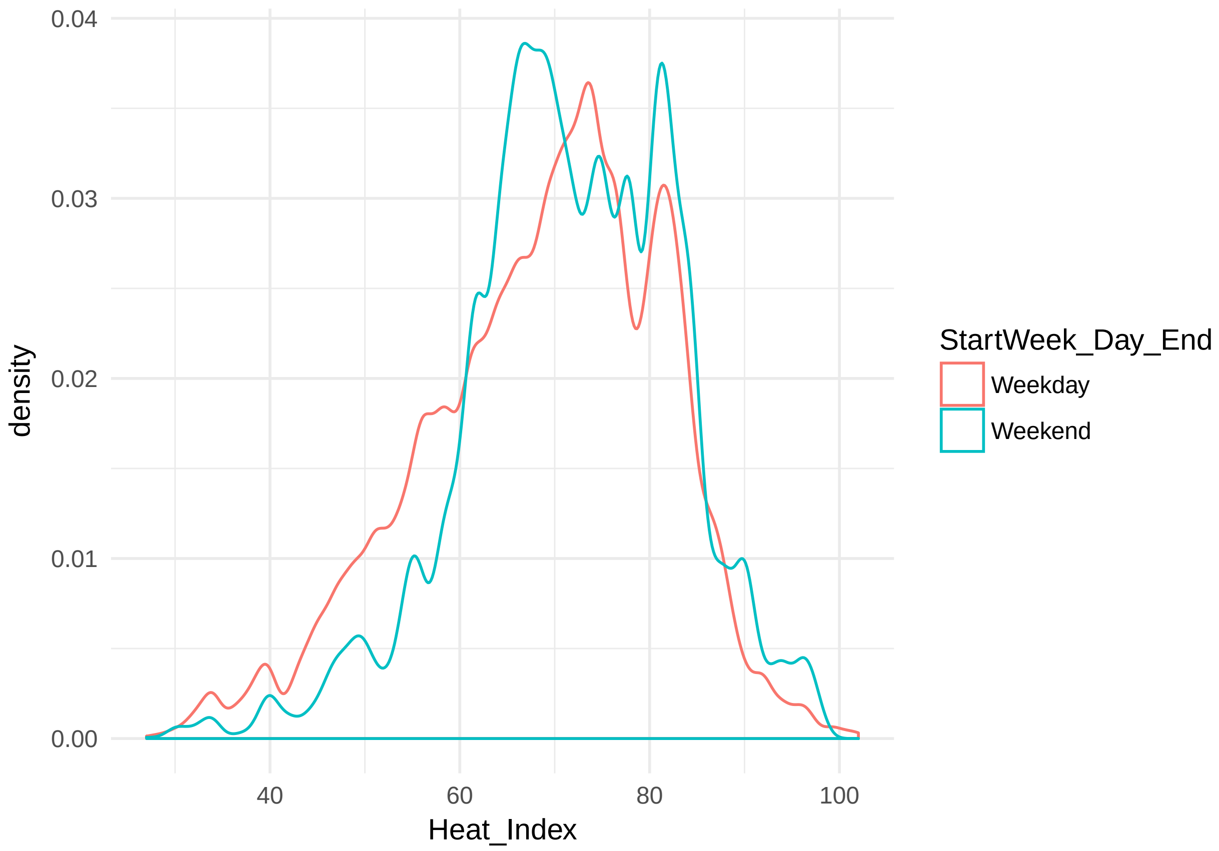
```
#Distribution of trips in relation to wind speed by weekday and weekend and member or casual
Rides %>% ggplot(aes(Wind_Speed, color = StartWeek_Day_End)) + geom_density(adjust = 1.25) + facet_wrap
```



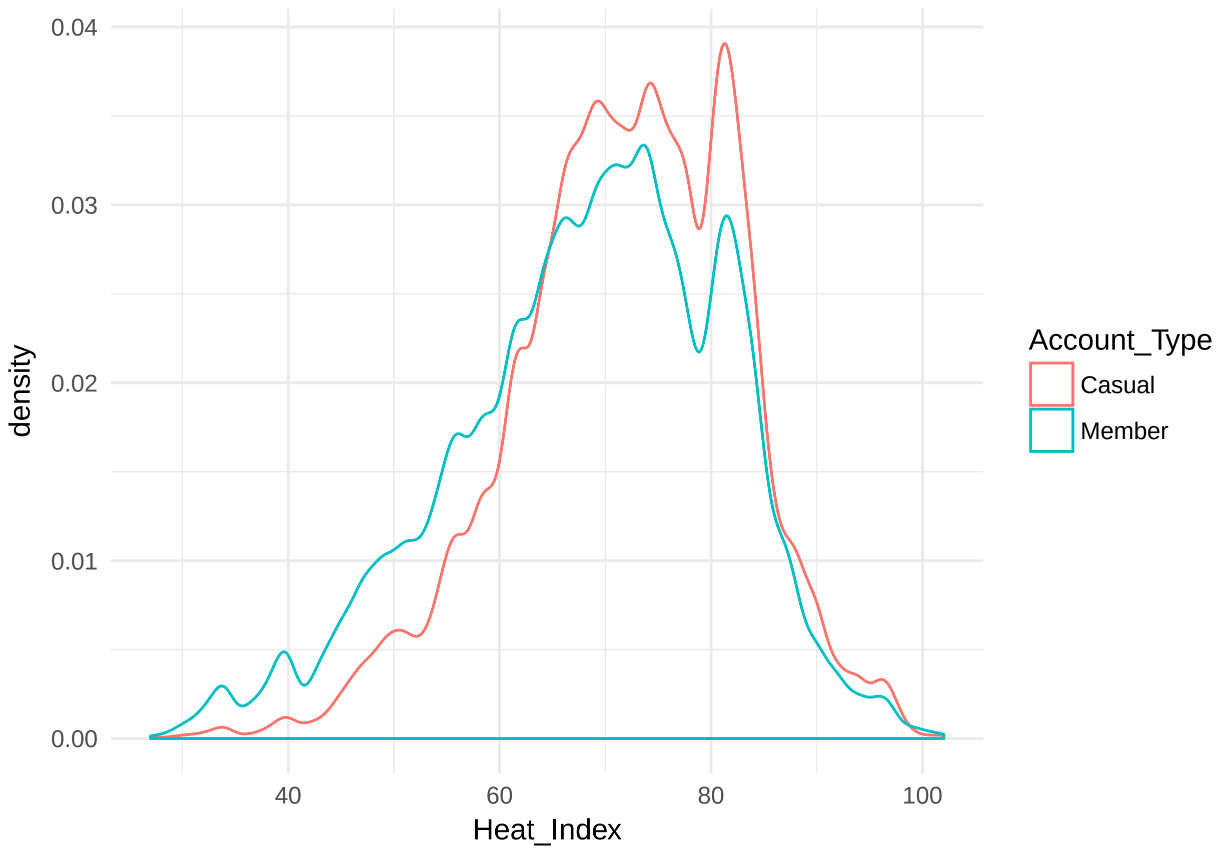
Casual rides show similar responses to an increase in wind speed for weekday and weekend trips, though there is slightly more tolerance of riders for wind speeds of 15-17.5 miles per hour on the weekends. Member riders show a very similar response.

### How does heat index affect bike use?

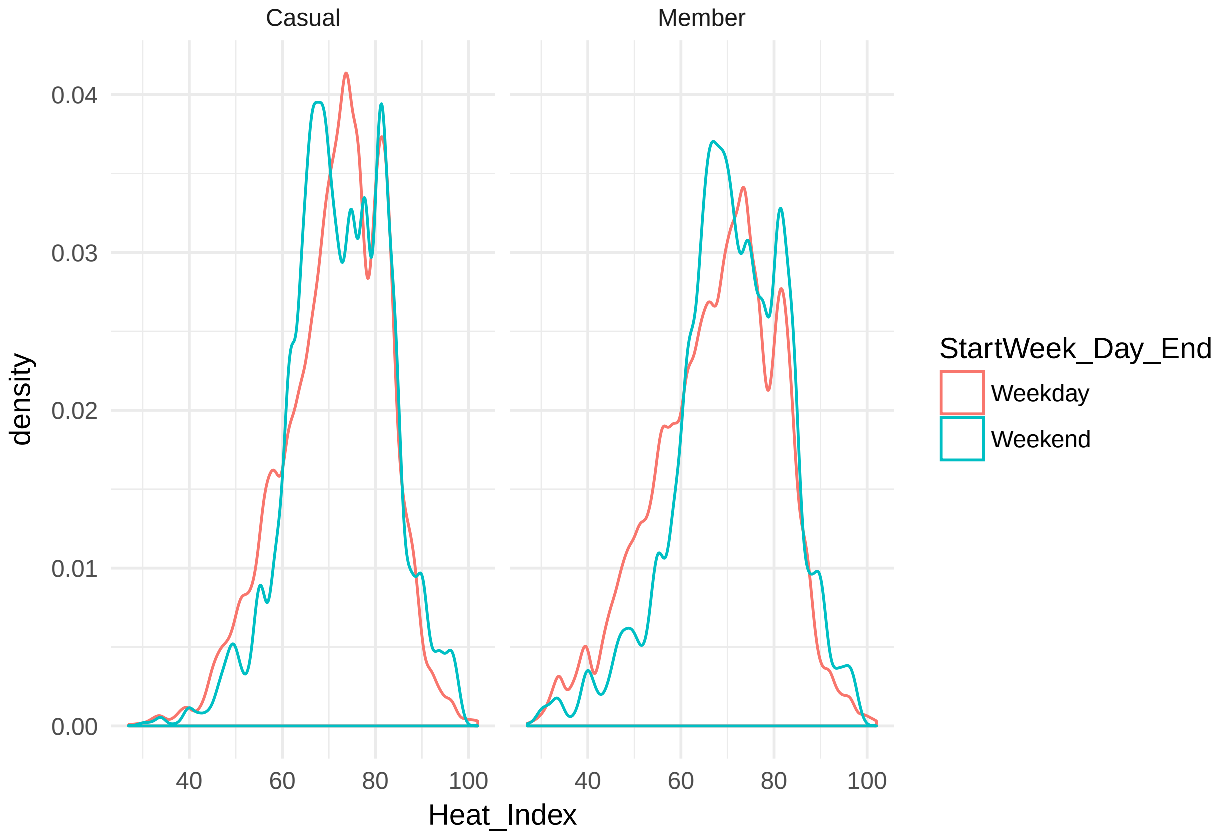
```
#Distribution of trips in relation to heat index by weekday and weekend
ggplot(Rides, aes(Heat_Index, color = StartWeek_Day_End)) + geom_density()
```



```
#Distribution of trips in relation to wind speed by member or casual
ggplot(Rides, aes(Heat_Index, color = Account_Type)) + geom_density()
```



```
#Distribution of trips in relation to wind speed by weekday and weekend and member or casual
ggplot(Rides, aes(Heat_Index, color = StartWeek_Day_End)) + geom_density() + facet_wrap(~ Account_Type)
```

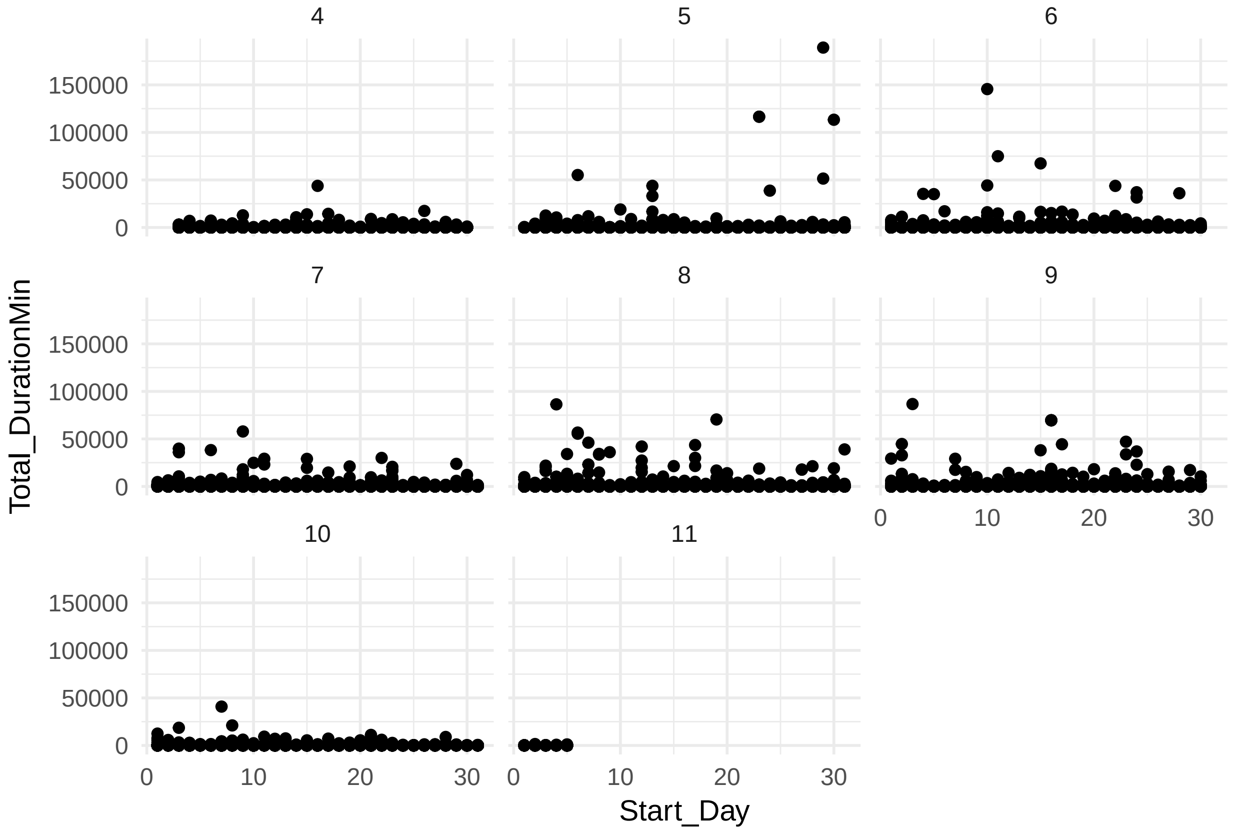


The heat index - variable combining temperature and relative humidity - holds comparable responses in rider volumes for casual and member account types. Riders of both account types are more tolerant of a low heat index on the weekdays and a higher heat index on the weekends. Both casual and member rides drop severely as the heat index increases above the low 80's, as to be expected.

## Time Series Plots

### DOES NOT WORK CORRECTLY DUE TO DISCRETE OBSERVATIONS AND THE SEPARATION OF DATE/TIME VARIABLES? HOW TO ADDRESS THIS?

```
Rides %>% ggplot(aes(Start_Day, Total_DurationMin)) + geom_point() + facet_wrap(~ Start_Month)
```



### Geospatial exploration - busiest starting stations in Minneapolis

```
#Load map
Map_TwinCities <- get_map(c(lon = -93.25576, lat = 44.97394), zoom = 12, maptype = "roadmap", source =
```

```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=44.97394,-93.25576&zoom=12&size=
```

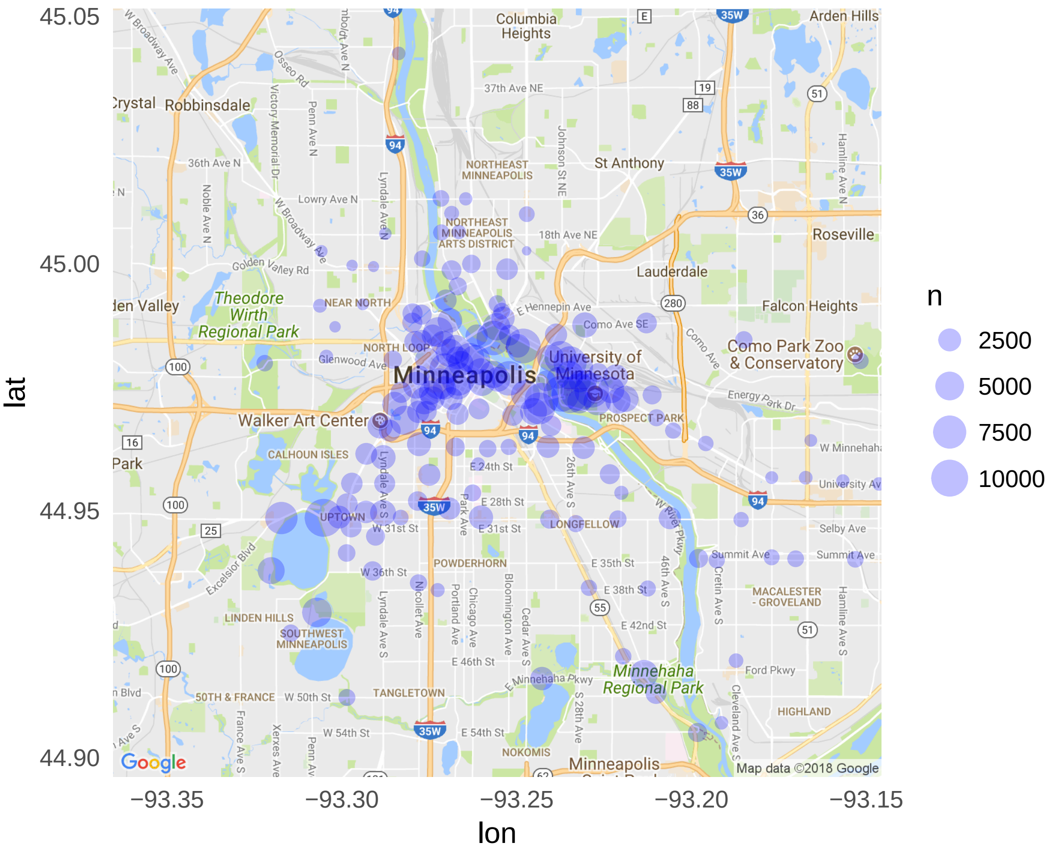
```
#Set map data subset of Rides
Geo_Rides <- Rides %>% group_by(Start_Longitude, Start_Latitude) %>% count(Start_Station) %>% arrange(d

attach(Geo_Rides)
```

#### Observe bike stations by number of trips for the season

```
ggmap(Map_TwinCities) + geom_point(aes(x = Start_Longitude, y = Start_Latitude, size = n), data = Geo_R
```

```
## Warning: Removed 34 rows containing missing values (geom_point).
```



The stations with the highest volume are in the downtown metro, University of Minnesota west bank campus and the lake of the Calhoun Isles area. *Mapping the station volume brought 34 rows containing missing values to our attention.*

## Next steps, changes in research purpose and intended outcome

Next steps are:

- Address outliers and dataset concerns brought forth in the EDA process before moving on to next phase
- Evaluate intended application of the project. The goal of this project remains focused on exploring how weather impacts bike share volume, but we must also consider other variables. *Should we account for holidays as they alter weekday ridership into patterns more closely tied to weekend use? What other variables affect ridership besides weather? City events?*
- The ultimate goal is to build a predictive model that shows how an altering of weather variables affects ride volume for a given day. *We have not explored the impact of ridership from our weather type categorical variable.*