Divvy Capstone: Wrangling Milestone

Introduction

The purpose of this first milestone was to properly wrangle my dataset in order to get an improved overview on the allocation of observations by day. The first sections of this analysis involved formatting the data properly, concentrating on date and time. The second section covers double checking for any NA values and making sure the number of observations in my post-wrangling datasets matched the original raw data. Finally, I visualized the data by year and month and was able to find a pattern of seasonality.

1. Tidy up name columns

- Date and time were originally combined, I separated the 2 and converted times to military time (this will be needed for later when using lubridate)

```
#separated the date and time in start and stop time columns order to make the analysis easier
library(tidyr)
FROM_sepstart <- separate(FROM, START.TIME, c("start.date", "start.time", "start.time.ampm"), sep = " ")
FROM <- separate(FROM_sepstart, STOP.TIME, c("start.date", "start.time", "start.time.ampm"), sep = " ")
View(FROM)

library(tidyr)
TO_sepstart <- separate(TO, START.TIME, c("stop.date", "stop.time", "stop.time.ampm"), sep = " ")
TO <- separate(TO_sepstart, STOP.TIME, c("stop.date", "stop.time", "stop.time.ampm"), sep = " ")
View(TO)

#combined the AM/PM in the time
library(tidyr)
FROM <- unite(FROM, "start.time", start.time, start.time.ampm, sep = " ")
TO <- unite(TO, "stop.time", stop.time, stop.time.ampm, sep = " ")

#converted AM/PM to military time for easy analysis
TO$stop.time<- (format(strptime(TO$stop.time, "%I:%M:%S %p"), "%H:%M:%S"))
FROM$start.time <- (format(strptime(FROM$start.time, "%I:%M:%S %p"), "%H:%M:%S"))
```

- Renamed & restructured the date columns as below. I separated month, day and year in order to make the analysis easier to group

```
#separated the dates into day, month, year for overview cvs file
library(tidyr)
FROM_overview <- separate(FROM, start.date, c("month", "day", "year"), sep = "/")
View(FROM_overview)

library(tidyr)
TO_overview <- separate(TO, stop.date, c("month", "day", "year"), sep = "/")
View(TO_overview)
```

2. Check for missing values

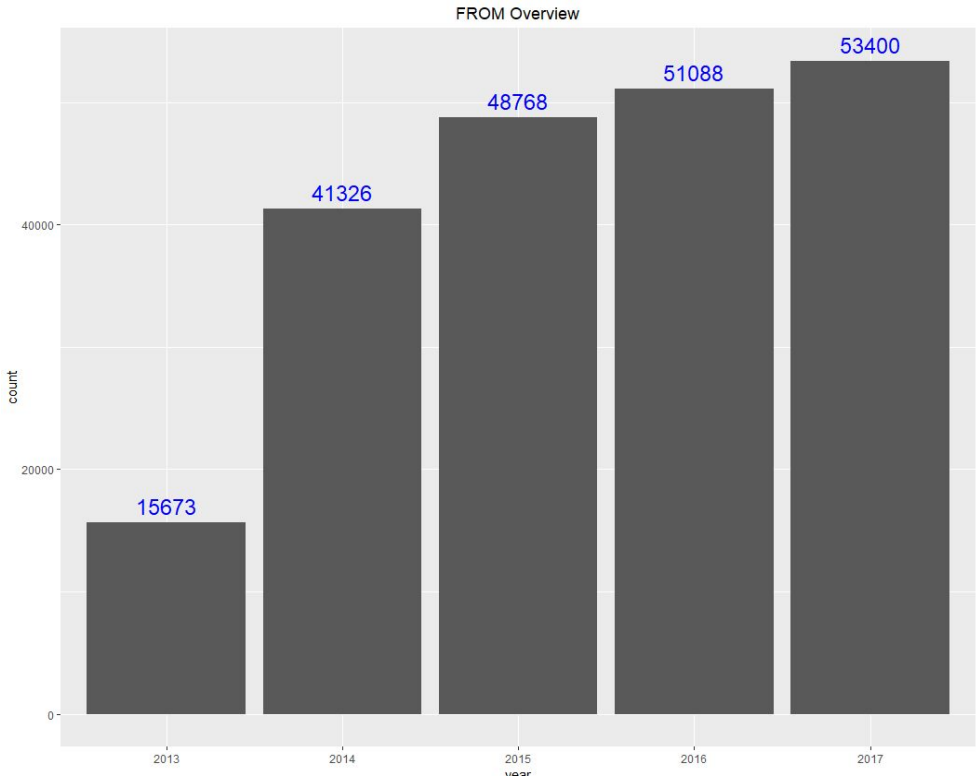
- Since the date columns are the most vital to my analysis I concentrated on this column to check for NA values as below in the TO and FROM datasets. A return of true would indicate an NA value. As per below there were no NA values found.

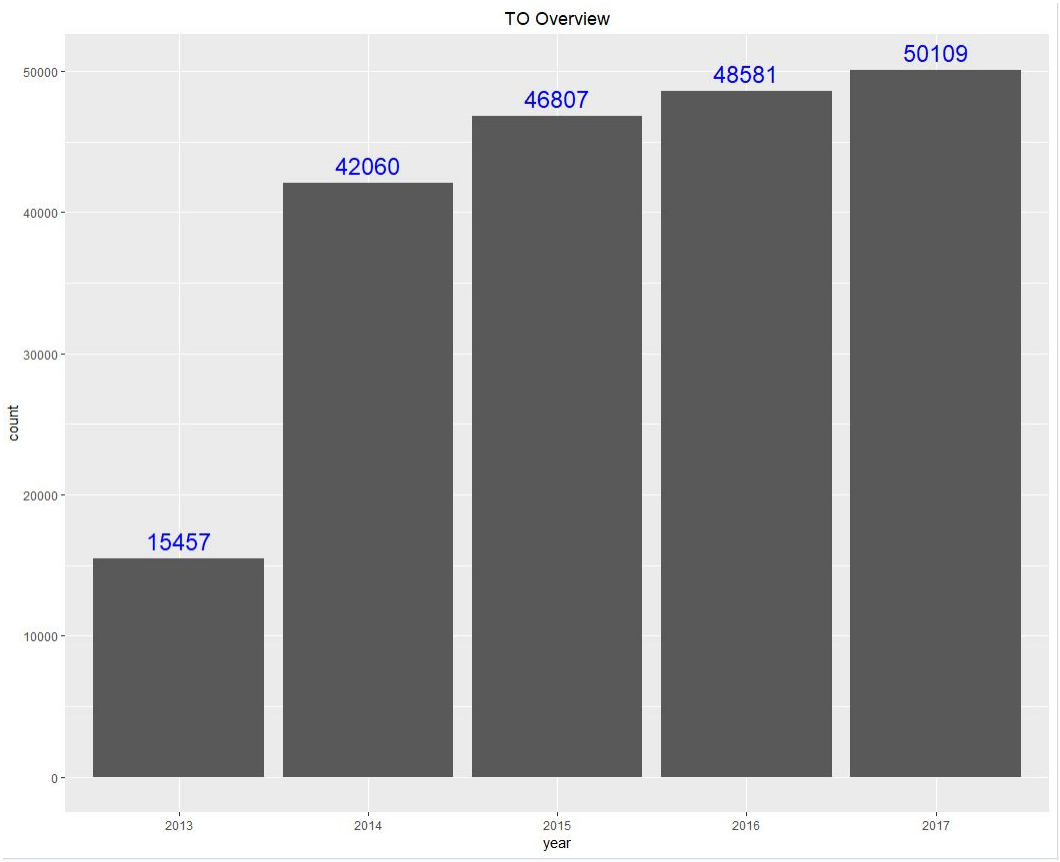
```
> summary(is.na(TO$year))
   Mode    FALSE
logical   203014
> summary(is.na(TO$month))
   Mode    FALSE
logical   203014
> summary(is.na(TO$day))
   Mode    FALSE
logical   203014
```

```
> summary(is.na(FROM$year))
   Mode    FALSE
logical   210255
> summary(is.na(FROM$month))
   Mode    FALSE
logical   210255
> summary(is.na(FROM$day))
   Mode    FALSE
logical   210255
>
```

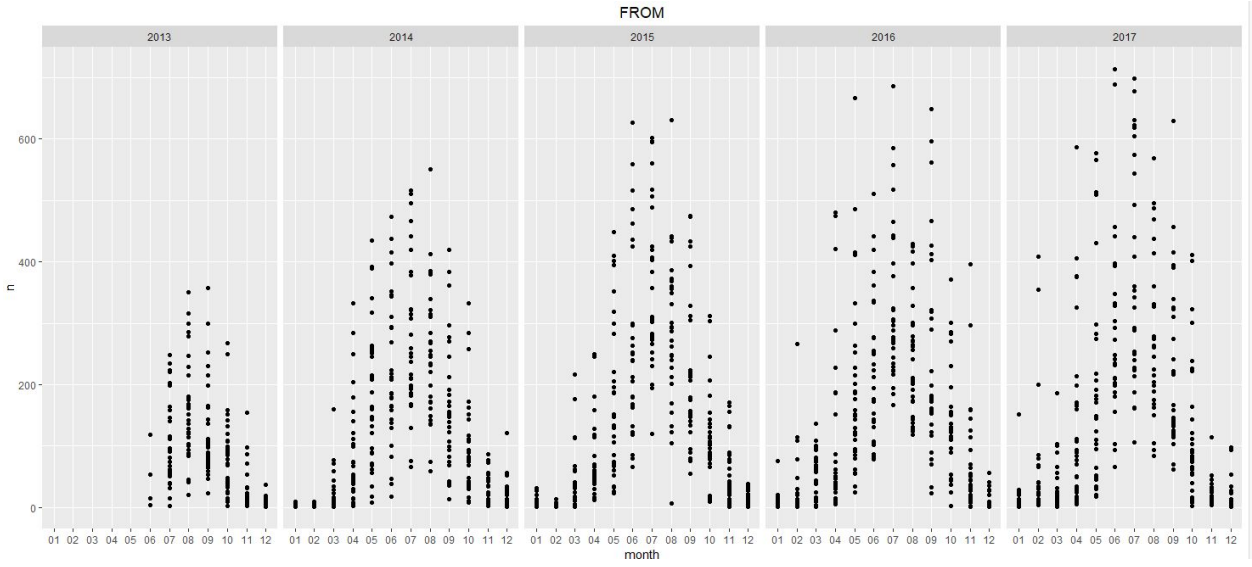
3. Overview on the number of observations per year, month & day

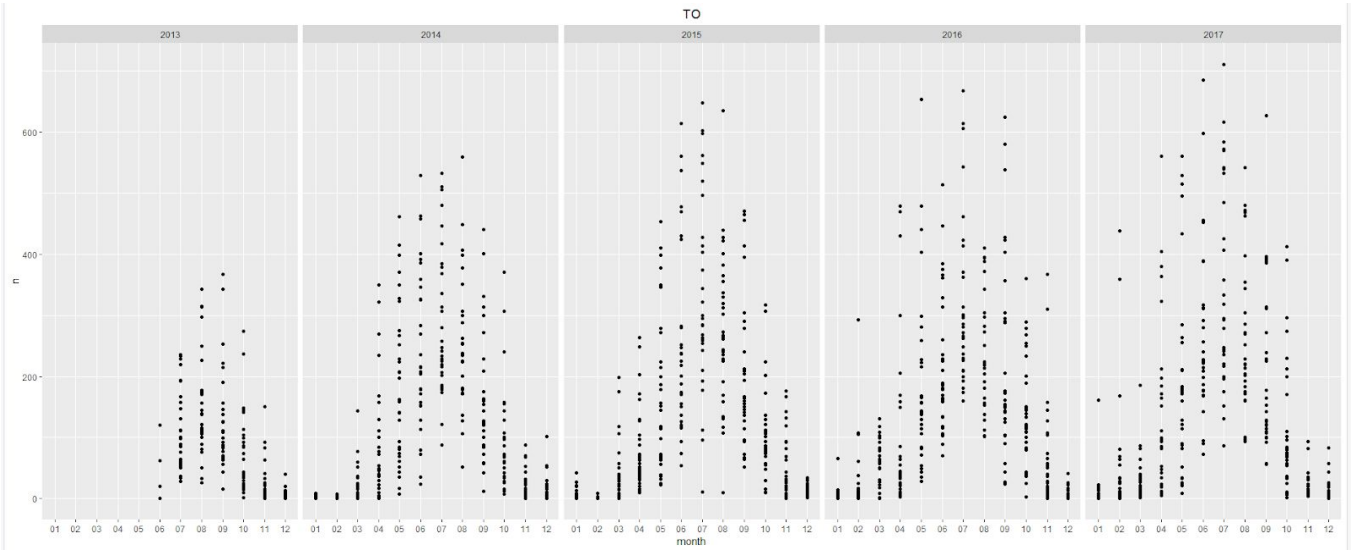
- First I wanted to get an overview of the number of observations by year to see if any years should be removed from the analysis





- Observations broken down by month and year





- For an even more detailed analysis I Grouped observations but the number of observations per day and exported them to excel for an easy analysis. The excels for TO and FROM may be found here.

- I noticed within the excel that certain days were skipped on forecasting. However, since the total of the observations in the excel matches the total observations in the TO (203014) and FROM (210255) these missing dates were probably not caused by my data extraction, rather they were not included / observed in the original data set

206	205	2014	2	24	2
207	206	2014	2	28	5
208	207	2014	3	1	1
209	208	2014	3	4	1
210	209	2014	3	7	17
211	210	2014	3	9	9
212	211	2014	3	10	44
213	212	2014	3	11	15
214	213	2014	3	12	1

Conclusion

After wrangling the data a few patterns emerged:

- Over the period of 2013 - 2017 the number of Divvy Bikes brought to and taken from the Divvy station increased - this makes sense as Divvy bikes began operations in 2013 so as brand awareness increased more people took advantage of using this service
- There is a clear pattern of seasonality as demand increases in the summer months and drops off during the winter month as seen in the final 4 visualizations