Bikes Analysis

Anisha

8/17/2021

ABOUT THE DATA

The Bay Area Bike Share enables quick, easy, and affordable bike trips around the San Francisco Bay Area. They make regular open data releases plus maintain a real-time API. There are 4 data sets available, out of which 3 are used for this paper. The data are as follows:

1. station - Contains data that represents a station where users can pickup or return bikes.
2. trip - Data about individual bike trips
3. weather - Data about the weather on a specific day for certain zip codes

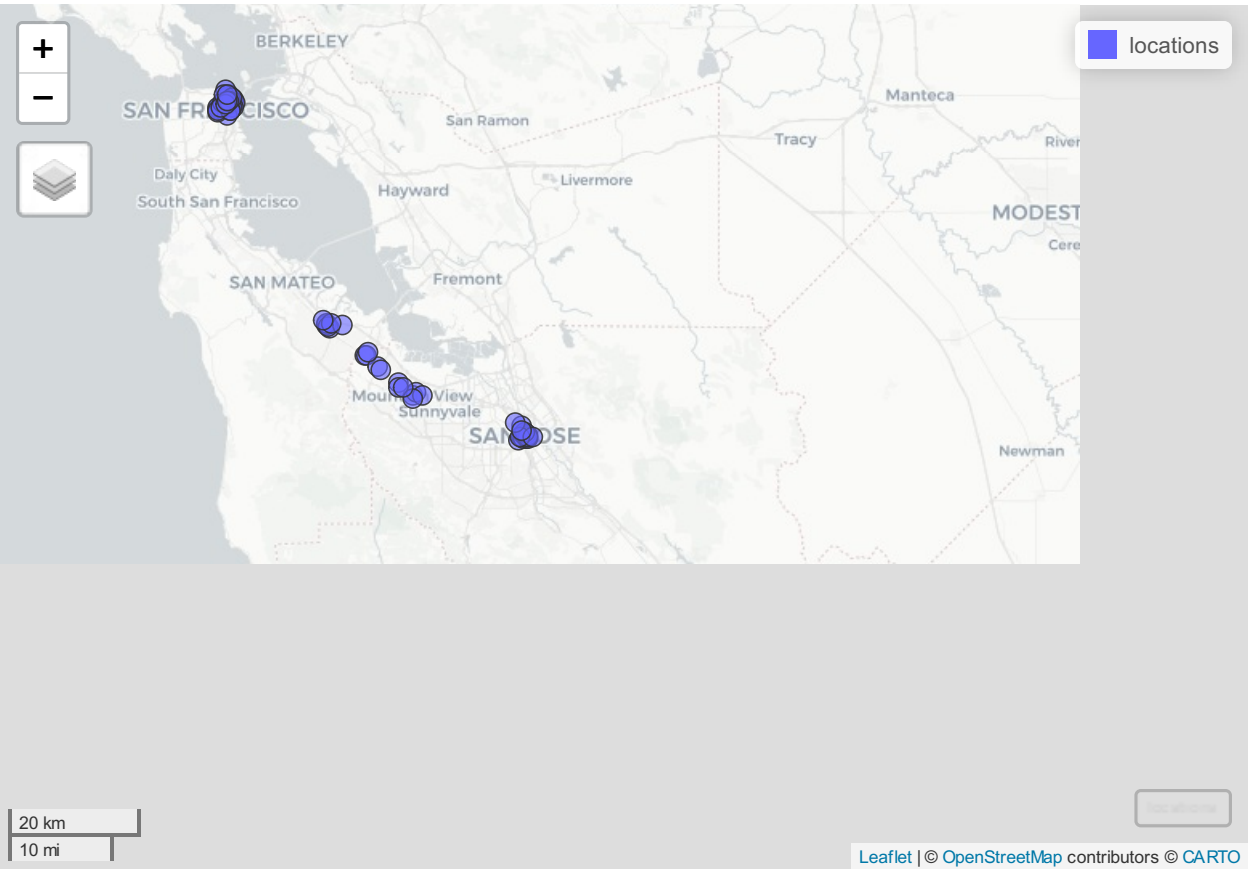
Assumption

As the data is of daily values from year 2013-2015, it is assumed the bikes are manual and not e-bikes.

OBJECTIVES

Based on the insights, where possible,suggest approach/es to optimize bike placement/reallocation schedule for the operator by considering usage patterns during the day and week, weather, and ride-distances.

```
# Transform station data as tibble
data <- as_tibble(station[1:4])
# Make coordinates as spatial
locations <- st_as_sf(data, coords = c("long", "lat"), crs = 4326)
# Plot stations on map
mapview(locations)
```



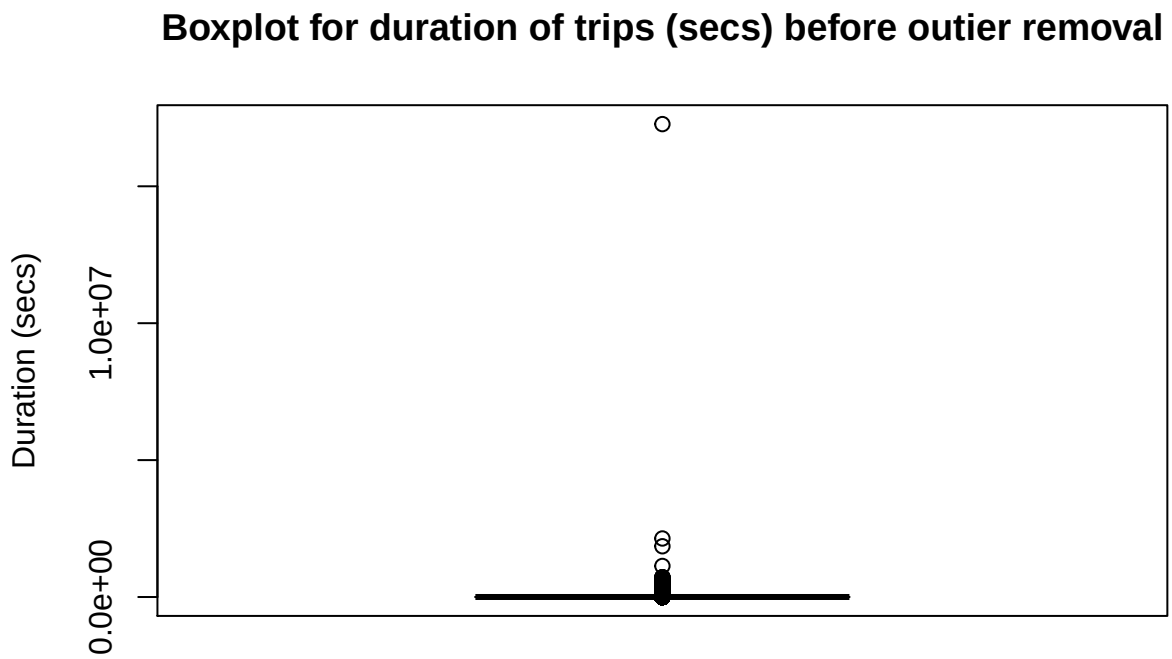
We can see that there are 5 cluster of plot. This suggests that there are 5 major cities where this data is recorded from, namely; San Jose, Redwood city, Mountain View, Palo Alto, and San Francisco.

Data manipulation and combination

As installation date of a station is not much of a use, we can remove that column. We add a new column representing the zipcodes of cities the stations are located at. To see relation between stations and the frequency of trips, it is important to merge station and trip data. But before we merge the data sets, we will check for outliers in trip data. Since looking at the duration of trips might lead to interesting inferences, outliers from trip data are removed based on the boxplot statistics of duration of trips.

```
# Arrange stations in order and remove column installation date
station %<>% select(-installation_date) %>% arrange(id)
zip_codes <- unique(weather$zip_code)
city <- c("San Francisco","Redwood City","Polo Alto","Mountain View","San Jose")

# add column zip code based on city from weather data
station <- station %>% mutate(zip_code =
                    case_when(city == "San Jose" ~ 95113,
                              city == "Redwood City" ~ 94063,
                              city == "Mountain View" ~ 94041,
                              city == "Palo Alto" ~ 94301,
                              city == "San Francisco" ~ 94107))
```

```r
# Change names of column in trip data
colnames(trip)[1] <- "subscriber_id"
colnames(trip)[5] <- "id"

# plot of summary of trips duration
boxplot(trip$duration , col = "cyan4",
        main = "Boxplot for duration of trips (secs) before outier removal",
        ylab = "Duration (secs)")
```




```r
max_drtn<-which.max(trip$duration)
trip[max_drtn,]
```

```
##        subscriber_id duration      start_date       start_station_name id
## 573567        568474 17270400 12/6/2014 21:59 South Van Ness at Market 66
##                end_date end_station_name end_station_id bike_id
## 573567 6/24/2015 20:18    2nd at Folsom             62     535
##        subscription_type zip_code
## 573567          Customer    95531
```

```r
## removing outliers from boxplot statistics
index_output <- c() # empty vector for index output
out <- boxplot.stats(trip[,2])$out
index_output <- c(index_output,which(trip[,2] %in% c(out)))
trip <- trip[-unique(index_output),]
# 7% data removed
```

Looking at the boxplot, there is a trip which is 17270400 secs = approx 199 days long. when checked it was a customer's trip. There could be many reasons for this trip duration.

According to boxplot statistics of duration of trips, 7% of data was outside the whiskers of boxplot and was hence removed.

```
# merge station and trip data
final_data<-merge(station, trip, by ="id")
final_data[c(2,3,4,11,13,17)] <- NULL
# Change class and format of date
final_data$start_date<-as.POSIXct(strptime(final_data$start_date,
                                           format = "%m/%d/%Y %H:%M",
                                           tz= "US/Pacific"))
final_data$end_date<-as.POSIXct(strptime(final_data$end_date,
                                           format = "%m/%d/%Y %H:%M",
                                           tz= "US/Pacific"))


colnames(final_data) <- c("start_station_id","dock_count","city","zip_code",
                          "subscriber_id","trip_duration","start_date",
                          "end_data","end_station_id","bike_id",
                          "subscription_type")
```

Looking at the first 6 entries of the final merged data

```
head(final_data)
```

start_station_id	dock_count	city	zip_code	subscriber_id	trip_duration	start_date	end_data	end_station_id	bike_id	subscription_type
2	27	San Jose	95113	317748	927	2014-06-10 07:40:00	2014-06-10 07:56:00	10	77	Subscriber
2	27	San Jose	95113	475348	292	2014-09-30 07:42:00	2014-09-30 07:47:00	4	72	Subscriber
2	27	San Jose	95113	519965	919	2014-10-28 20:54:00	2014-10-28 21:09:00	9	93	Subscriber
2	27	San Jose	95113	21070	110	2013-09-14 08:49:00	2013-09-14 08:50:00	2	641	Subscriber
2	27	San Jose	95113	490739	459	2014-10-09 12:46:00	2014-10-09 12:54:00	11	305	Subscriber
2	27	San Jose	95113	226989	467	2014-03-25 18:45:00	2014-03-25 18:52:00	11	130	Subscriber

ANALYSIS

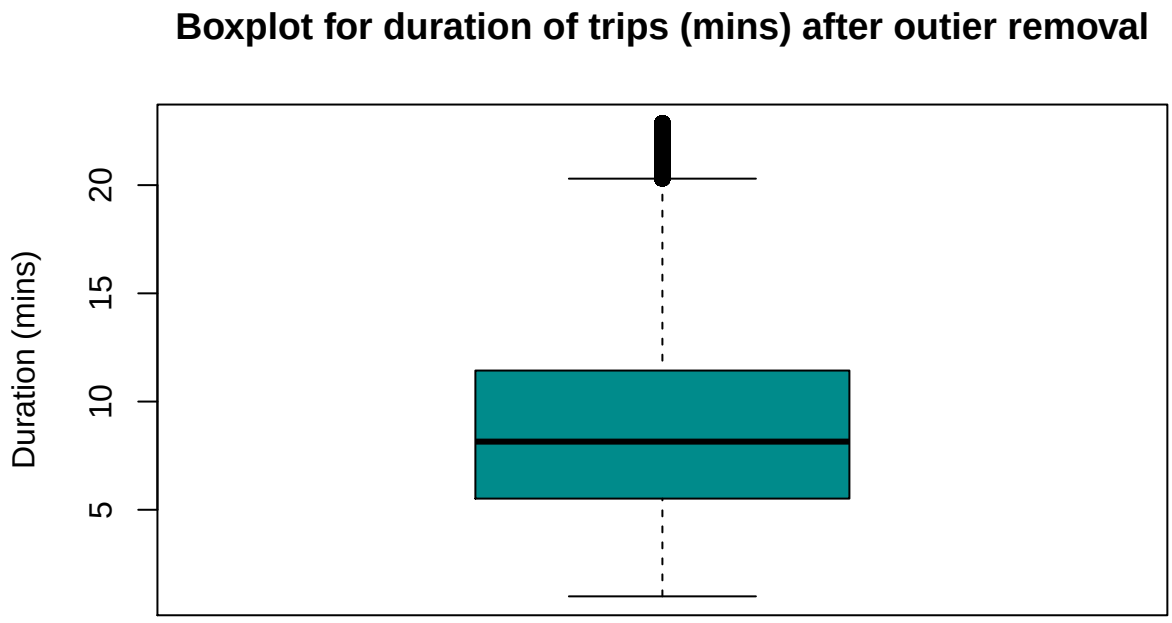
1. Duration of trips

As duration of trips is given in seconds, it is converted in minutes for ease in inferences.

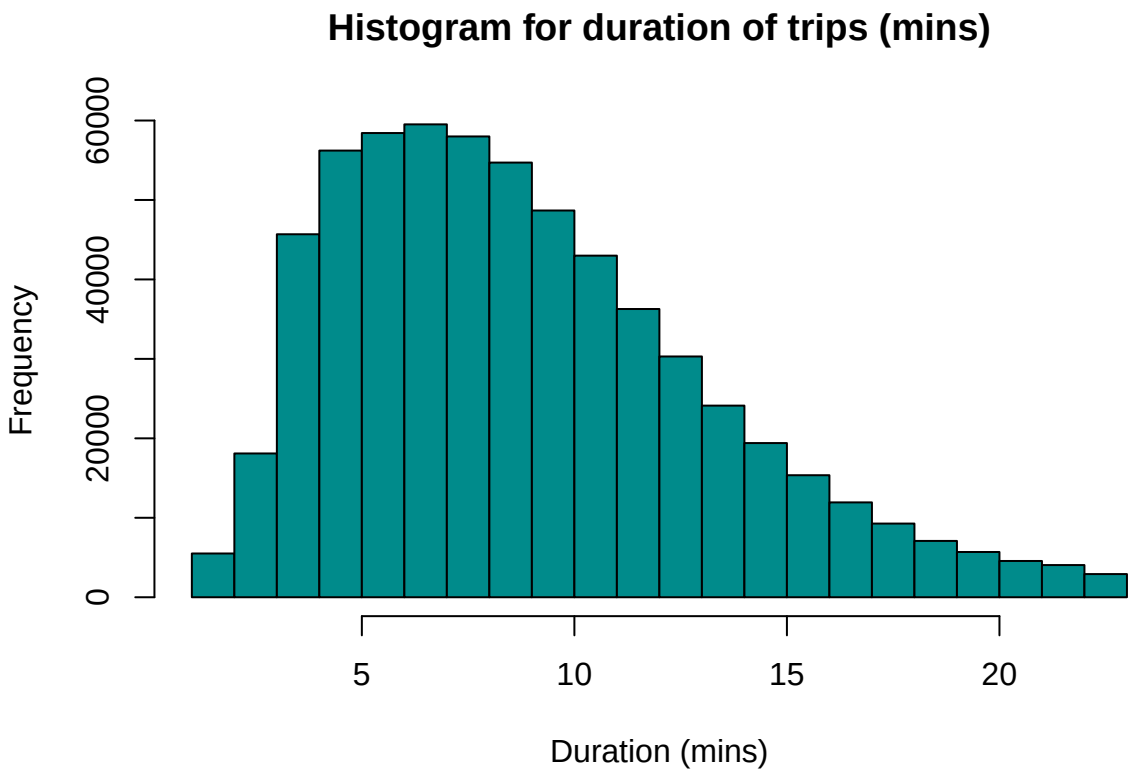
```
# convert trips duration in mins
final_data$trip_duration <- (final_data$trip_duration)/60
# Summary of trips duration
summary(final_data$trip_duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   5.517   8.150   8.860  11.433  22.850
```

```
boxplot(final_data$trip_duration , col = "cyan4",
        main = "Boxplot for duration of trips (mins) after outier removal",
        ylab = "Duration (mins)")
```




```
hist(final_data$trip_duration, col = "cyan4",
     main = "Histogram for duration of trips (mins)", xlab = "Duration (mins)")
```




```
table(final_data$subscription_type)
```

```
##
##   Customer Subscriber
##      61295     557533
```

Inferences

Statistical

We notice that the maximum duration of a trip is approximately 23 mins long and the average duration of a trip is only 8 mins long. We can also see that the average and median duration of trips is almost equal at 8 mins, which suggests that the duration of trips is normally distributed, Which is also suggested by the box plot. The box plot also shows some trips have duration greater than the maximum value. It is also seen that most trips are taken by the subscribers.

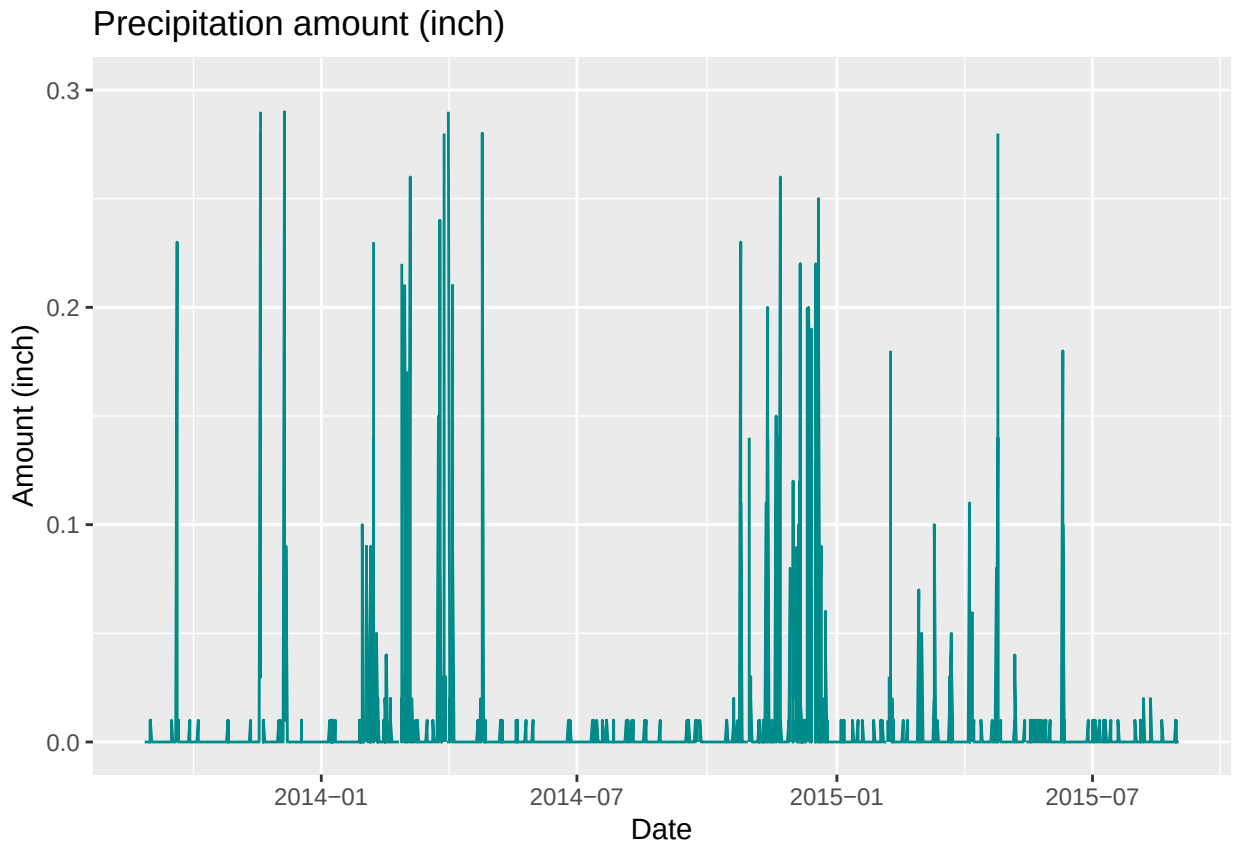
What is means for stakeholders

When checked the price of each 30min trip is USD 3 for a customer and USD 0 for a subscriber, but additional 15mins costs USD 3 and USD 2 extra respectively. The summary and the cost suggests that most subscribers don't want to pay extra 2 dollars for riding a bike for more than 30 mins. Looking at the histogram we can see that people can not ride a manual bike for more than 8-10 mins. Only very few exceptional people, who enjoy riding a bike would want to use a bike for a longer period of time. This also suggests that people use these bikes mostly for only short trips and not for leisure.

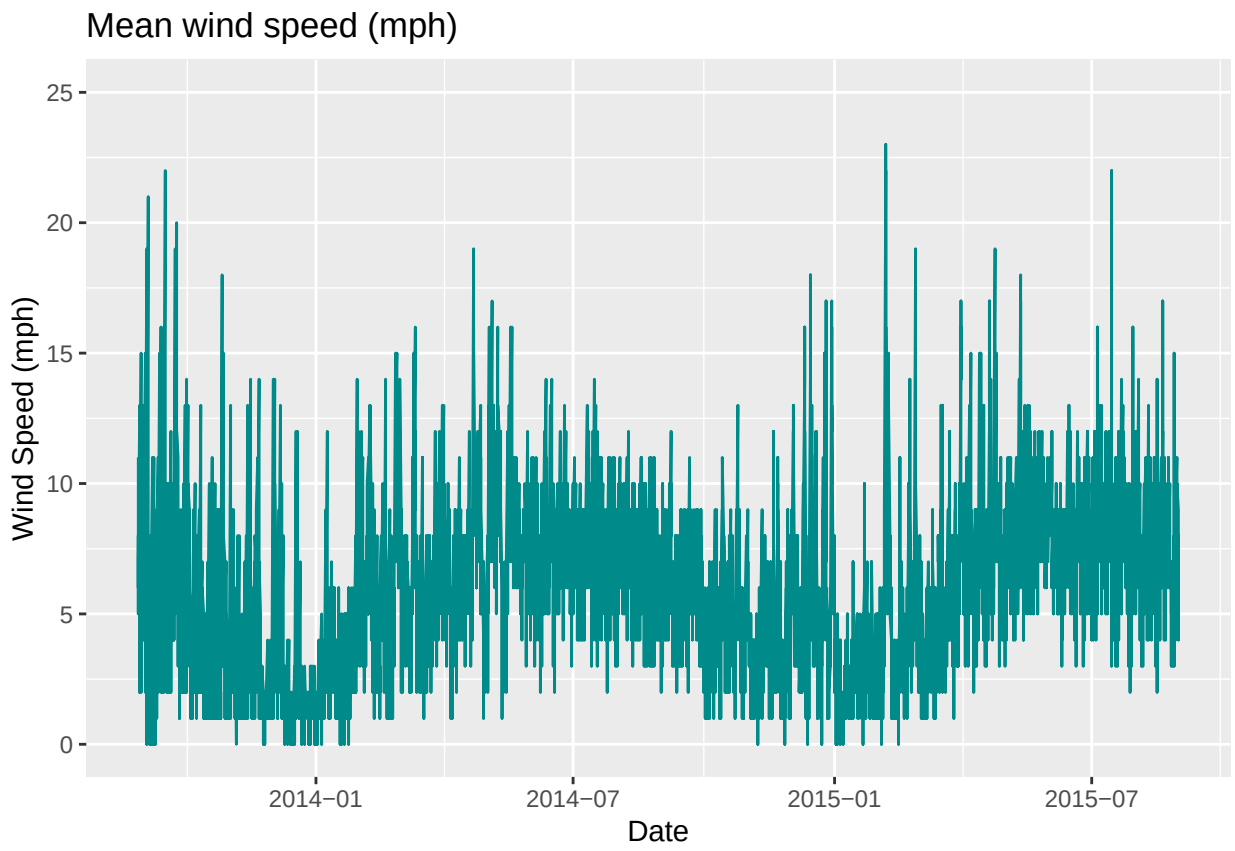
2. Weather during the given duration of data

Riding a manual bike in itself is difficult but it is tougher when the weather does not support either. Lets see how the plots of precipitation amount, mean wind speed, and mean temperature looks like from 2013/06 to 2015/08 in the five cities combined.

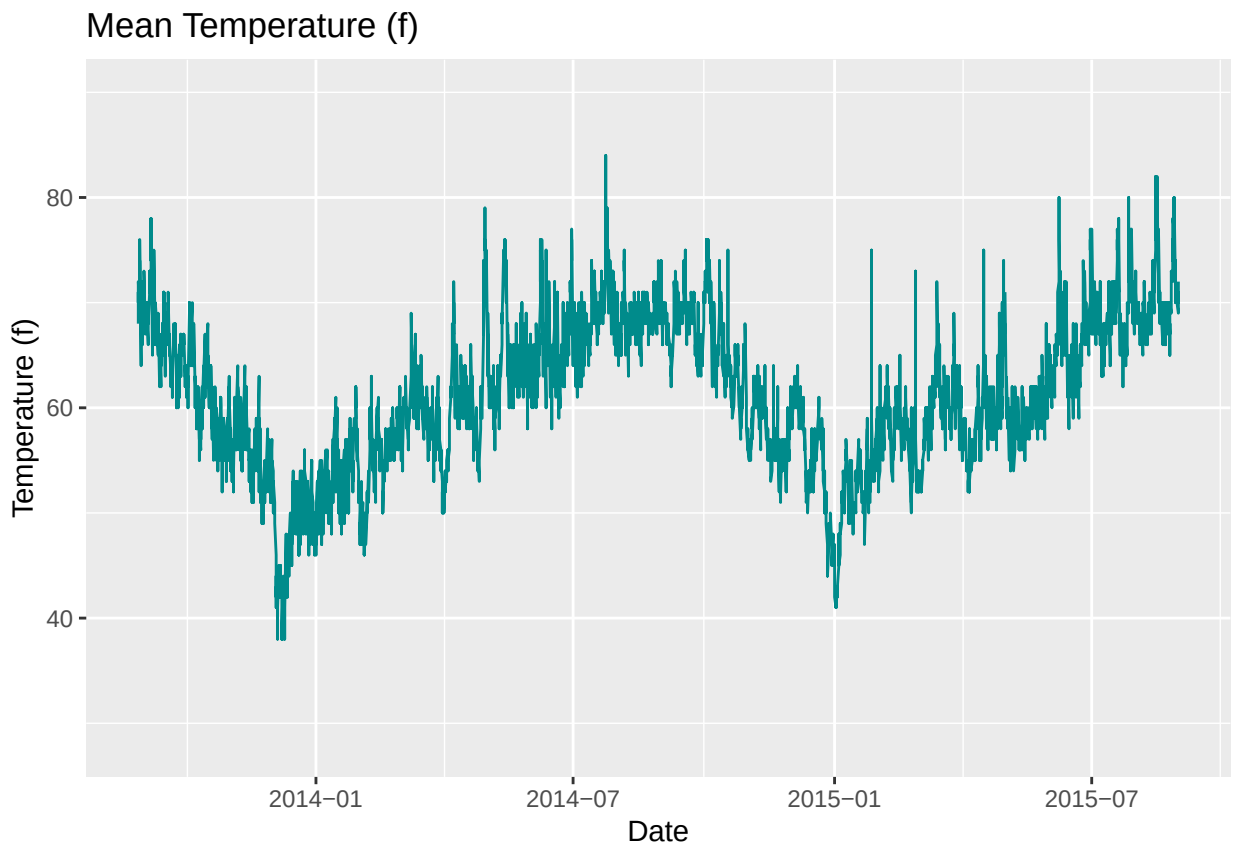
```
#### Weather data
# Replace T with 0.01 in precipitation amount
weather$precipitation_inches[weather$precipitation_inches=="T"] <- 0.01
# rain as numeric
weather$precipitation_inches <- as.numeric(weather$precipitation_inches)
# date as class Posixct for easy plotting
weather$date <- as.POSIXct(strptime(weather$date,
                                    format = "%m/%d/%Y",
                                    tz= "US/Pacific"))
# Plot amount of rainfall
ggplot(weather, aes(date,precipitation_inches)) + geom_line(color="cyan4") +
  ylim(0,0.3) + labs(title = "Precipitation amount (inch)", x = "Date",
                     y = "Amount (inch)")
```




```
# Plot Mean wind speed
ggplot(weather, aes(date,mean_wind_speed_mph)) + geom_line(color="cyan4") +
  ylim(0,25) + labs(title = "Mean wind speed (mph)", x = "Date",
                     y = "Wind Speed (mph)")
```




```
# Plot Mean Temperature
ggplot(weather, aes(date,mean_temperature_f)) + geom_line(color="cyan4") +
  ylim(28,90) + labs(title = "Mean Temperature (f)", x = "Date",
                     y = "Temperature (f)")
```



Inferences

Statistical

1. Precipitation amount: There was most rainfall recorded during the months of October, November, December, February, and March over the years.

2. Mean wind speed: The mean wind speed curve looks like a sine curve and hence suggests periodicity. We can see that the wind starts to increase from the month of march and reaches its peak by June. It starts decreasing from July and reaches its lowest in December. We can see that in December 2014 there is an increase in the wind speed. This is due to the winter storm called "Storm of the Decade" that the west coast of US experienced during that time.

3. Mean Temperature: The mean temperature plot suggests that the west coast of USA experiences peak summer during the months of June, July, and August. While winters are experienced during November, December, and January.

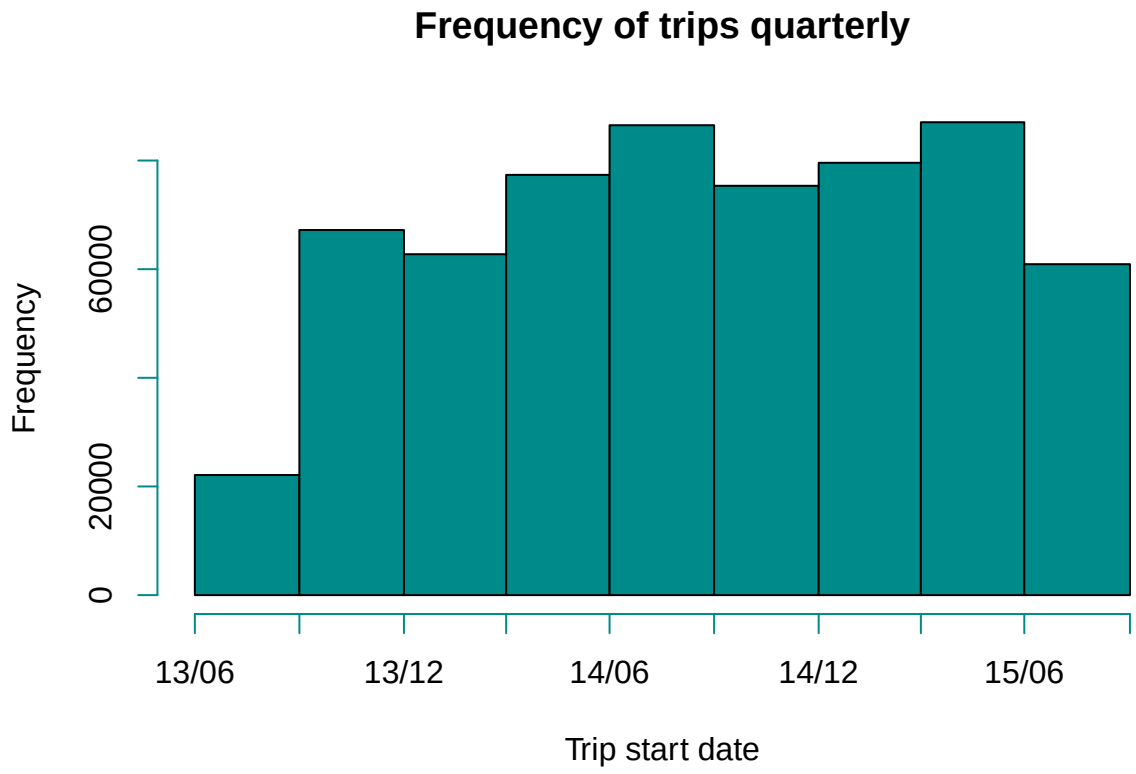
What is means for stakeholders

Weather is one of the most crucial external variable that is to be considered by any person before riding a bike. From these inferences, the stakeholder can come up with proper marketing strategies. The company can come up offers that would interests people according to the weather.

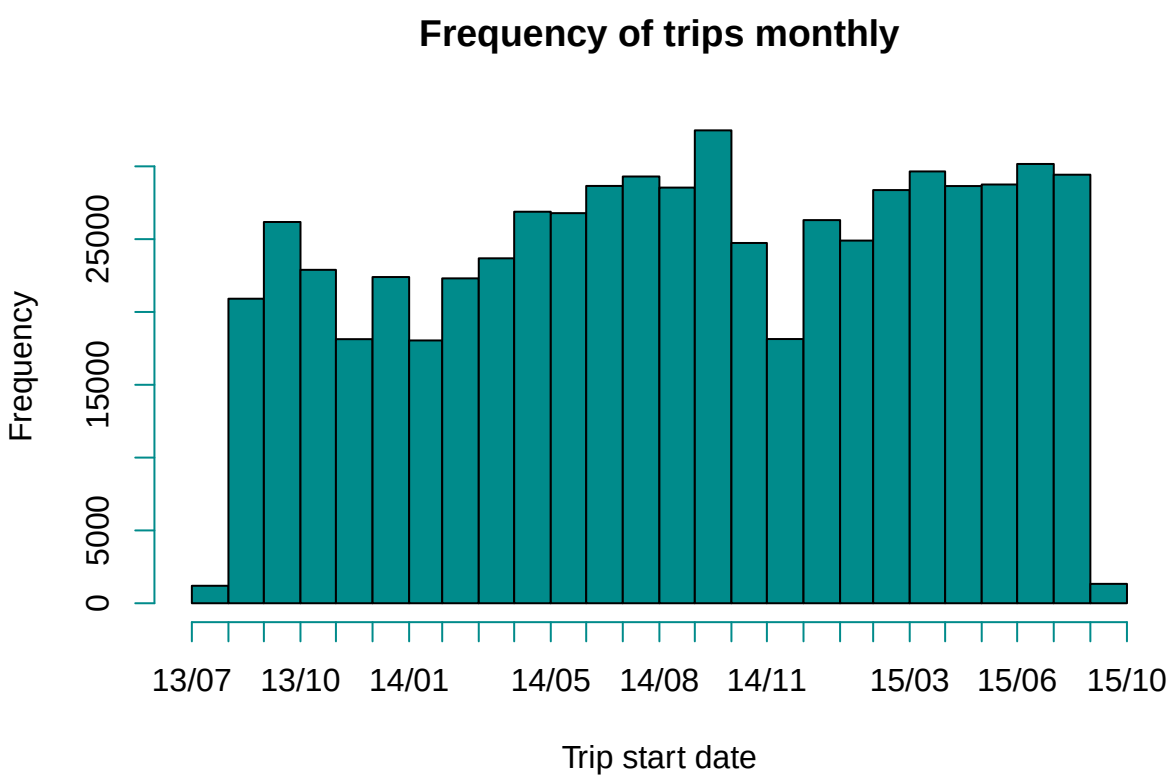
3. Frequency of trips

Now, that we have looked at weather patterns for the given 5 cities, let us take a look at the frequency of trips quarterly and monthly. There might me some relation between weather and frequency of trips.

```
# frequency of trips Quarterly
hist(final_data$start_date, breaks = "quarters", freq = TRUE, format = "%y/%m",
     col = "cyan4", main = "Frequency of trips quarterly",
     xlab = "Trip start date")
```




```
# frequency of trips Monthly
hist(final_data$start_date, breaks = "months", freq = TRUE, format = "%y/%m",
     col = "cyan4", main = "Frequency of trips monthly",
     xlab = "Trip start date")
```



Inferences

Statistical

From the quarterly plot, We see that during 2nd and 3rd quarters the frequency of trips are higher. On comparing this with the above weather results, 2nd and 3rd quarters constitutes of April to September. The time where there is least or no amount of rainfall, the mean temperature is between 60-70 degree Fahrenheit and means spring and summer, and the wind is between 3-7 mph. From the monthly plot, we see that the maximum number of trips were recorded in October'2014. Else the number of trip are most as suggested by the quarterly plots.

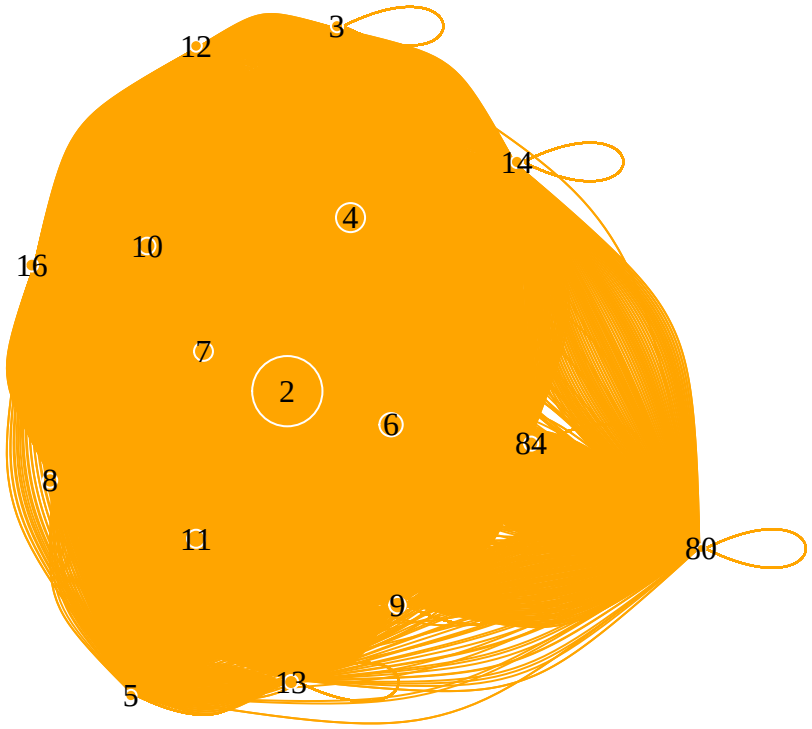
What is means for stakeholders

This suggests that the stakeholders can come up with a scheme where people who already own bikes can put their bikes on share, increasing the number of available bikes. People who are going on vacations and/or not using their personal bikes much can put their bikes on share on commission basis with SF Bay bikes. This is would increase in number of available bikes and also would not cost much to the stakeholders. A cost that the company would incur would be to increase the number of docks. But, that would be a one time investment. The next part can find efficient solution on where to increase/decrease the number of stations.

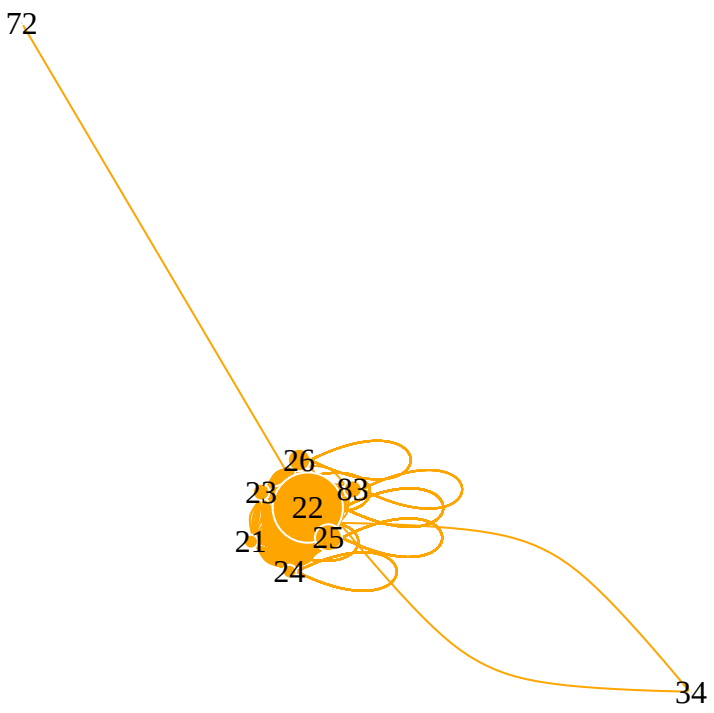
4. Network between stations for each city

Lets look at the city wise stations that has been accessed the most by looking at their network plots.

```r
## network plots of trips in each city

# Filter data city wise using zipcodes
# Make data as matrix and graph
# This will help to interpret the graph. convert as adjacency matrix
# visualization using degrees

# san Jose
zip.sj_trip <- final_data %>% filter(zip_code == 95113) %>%
  select(c(start_station_id,end_station_id))

zip.sj_trip_mat <- as.matrix(get.adjacency(graph.data.frame(zip.sj_trip)))
gra.sj <- graph_from_adjacency_matrix(zip.sj_trip_mat)

deg.sj <- degree(gra.sj)
node_sizes <- 1 + 20*deg.sj / max(deg.sj)
igraph.options(vertex.label = colnames(zip.sj_trip_mat),
               vertex.color = "orange",
               vertex.frame.color = "#ffffff",
               vertex.label.color = "black",
               edge.color = "orange",
               edge.arrow.size = 0.01)
plot(gra.sj, layout = layout.fruchterman.reingold(gra.sj),
     vertex.size = node_sizes, main = "Network of trips in San Jose")
```

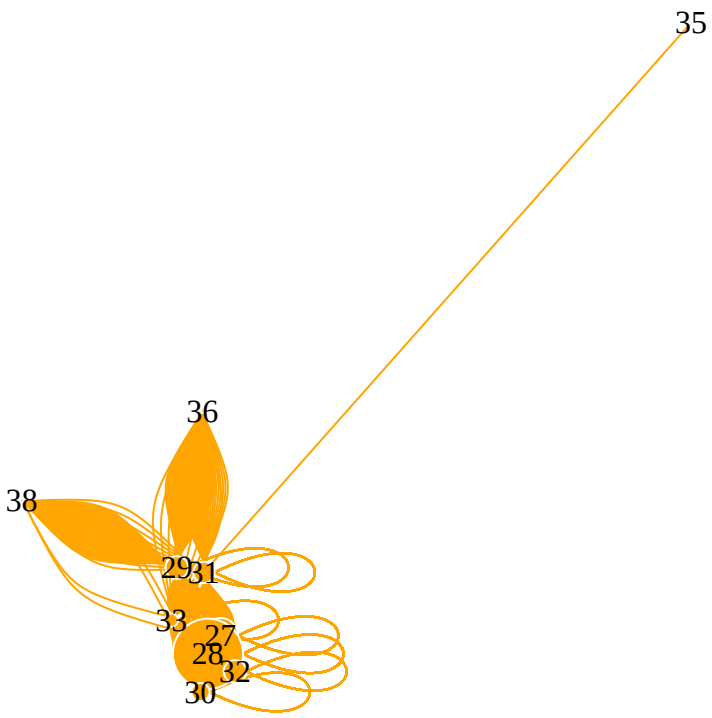



```
# Redwood city
zip.rc_trip <- final_data %>% filter(zip_code == 94063) %>%
  select(c(start_station_id,end_station_id))

zip.rc_trip_mat <- as.matrix(get.adjacency(graph.data.frame(zip.rc_trip)))
gra.rc <- graph_from_adjacency_matrix(zip.rc_trip_mat)

deg.rc <- degree(gra.rc)
node_sizes <- 1 + 20*deg.rc / max(deg.rc)
igraph.options(vertex.label = colnames(zip.rc_trip_mat),
               vertex.color = "orange",
               vertex.frame.color = "#ffffff",
               vertex.label.color = "black",
               edge.color = "orange",
               edge.arrow.size = 0.01)
plot(gra.rc, layout = layout.fruchterman.reingold(gra.rc),
     vertex.size = node_sizes,main = "Network of trips in Redwood City")
```



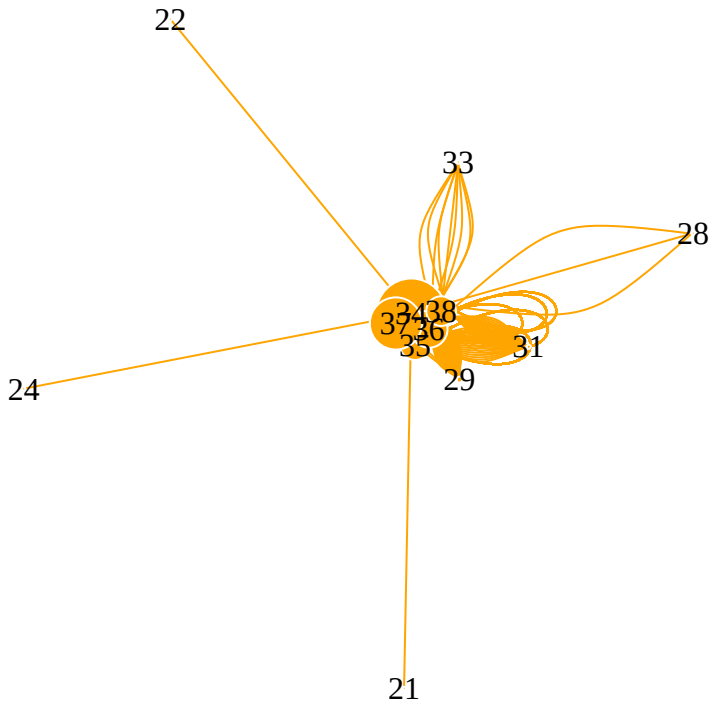

```
# Mountain View
zip.mv_trip <- final_data %>% filter(zip_code == 94041) %>%
  select(c(start_station_id,end_station_id))

zip.mv_trip_mat <- as.matrix(get.adjacency(graph.data.frame(zip.mv_trip)))
gra.mv <- graph_from_adjacency_matrix(zip.mv_trip_mat)

deg.mv <- degree(gra.mv)
node_sizes <- 1 + 20*deg.mv / max(deg.mv)
igraph.options(vertex.label = colnames(zip.mv_trip_mat),
               vertex.color = "orange",
               vertex.frame.color = "#ffffff",
               vertex.label.color = "black",
               edge.color = "orange",
               edge.arrow.size = 0.01)
```

```
plot(gra.mv, layout = layout.fruchterman.reingold(gra.mv),
     vertex.size = node_sizes,main = "Network of trips in Mountain View")
```




```
# Palo Alto
zip.pa_trip <- final_data %>% filter(zip_code == 94301) %>%
  select(c(start_station_id,end_station_id))

zip.pa_trip_mat <- as.matrix(get.adjacency(graph.data.frame(zip.pa_trip)))
gra.pa <- graph_from_adjacency_matrix(zip.pa_trip_mat)

deg.pa <- degree(gra.pa)
node_sizes <- 1 + 20*deg.pa / max(deg.pa)
igraph.options(vertex.label = colnames(zip.pa_trip_mat),
               vertex.color = "orange",
               vertex.frame.color = "#ffffff",
               vertex.label.color = "black",
```

```
               edge.color = "orange",
               edge.arrow.size = 0.01)
plot(gra.pa, layout = layout.fruchterman.reingold(gra.pa),
     vertex.size = node_sizes,main = "Network of trips in Palo Alto")
```




```
# San Francisco
zip.sf_trip <- final_data %>% filter(zip_code == 94107) %>%
  select(c(start_station_id,end_station_id))

zip.sf_trip_mat <- as.matrix(get.adjacency(graph.data.frame(zip.sf_trip)))
gra.sf <- graph_from_adjacency_matrix(zip.sf_trip_mat)

deg.sf <- degree(gra.sf)
node_sizes <- 1 + 20*deg.sf / max(deg.sf)
igraph.options(vertex.label = colnames(zip.sf_trip_mat),
               vertex.color = "orange",
```

```
               vertex.frame.color = "#ffffff",
               vertex.label.color = "black",
               edge.color = "orange",
               edge.arrow.size = 0.01)
# plot(gra.sf, layout = layout.fruchterman.reingold(gra.sf),
#      vertex.size = node_sizes,main = "Network of trips in San Francisco")
```

```
dist <- st_distance(locations,locations, by_element = FALSE)
```

Inferences

Statistical

1. San Jose and San Francisco has a very busy network of trips. People have traveled equally to and fro from each station in these two cities.
2. Mountain View has one trip to a station and multiple trips to two station in Palo Alto .
3. Redwood City has one trip to a station in San Francisco and two trips to a station in Palo Alto. All the stations here are busy.
4. Palo Alto has single trips to three station in Redwood city and three trips to a station and also multiple to another station in Mountain View. Most of the busy stations here are station id: 38,37,35,34.

What is means for stakeholders

This suggests that there some people who use bikes for longer trips. SF Bay bikes were initially started for short trips, so the company can come up discount facilities that would encourage people to hire bikes for longer period of time. Also, we see that all the stations are very busy, but we don't know the dock availability yet. Another data set is available which records the availability of bikes at SF Bay bike stations for every minute of day. That file is too heavy for R language and would hence require to be analysed on some other language.

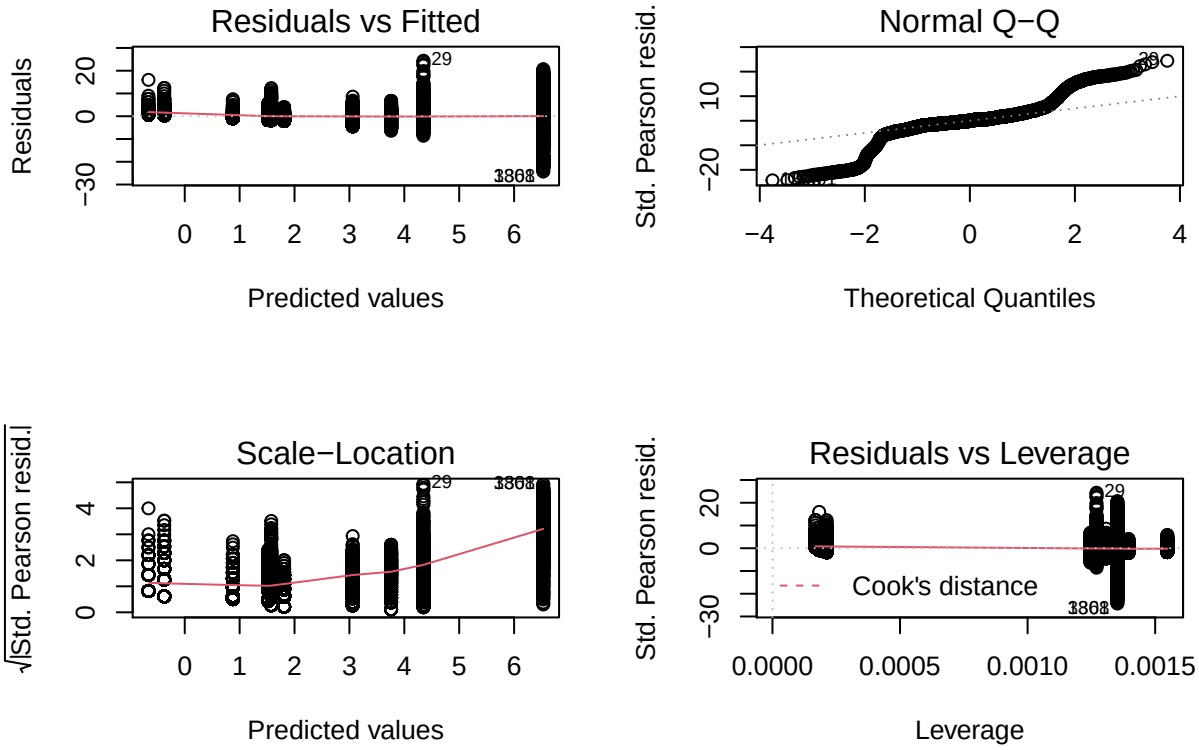
5. Predictive Model

Let us now develop a model that could predict the number of trips based on the zip code and subscription type of the user. As number of trips is a discrete data therefore we will fit a poisson regression model on our data.

```
## data for model
cnt_trip <- final_data %>%
  group_by(as.Date(start_date),subscription_type, zip_code) %>%
            tally()
colnames(cnt_trip)[4] <- "count"
cnt_trip <- as.data.frame(cnt_trip)
cnt_trip$subscription_type <- as.factor(cnt_trip$subscription_type)
cnt_trip$zip_code <- as.factor(cnt_trip$zip_code)

# Fit model
fit <- glm(count ~  subscription_type + zip_code, data = cnt_trip,
           family = "poisson")
summary(fit)
```

```
##
## Call:
## glm(formula = count ~ subscription_type + zip_code, family = "poisson",
##     data = cnt_trip)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -31.711   -1.868   -0.046    1.565   18.615
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  0.876493   0.008778   99.85   <2e-16 ***
## subscription_typeSubscriber  2.183159   0.004257  512.83   <2e-16 ***
## zip_code94063               -1.537248   0.019972  -76.97   <2e-16 ***
## zip_code94107                3.471751   0.007941  437.19   <2e-16 ***
## zip_code94301               -1.249354   0.017014  -73.43   <2e-16 ***
## zip_code95113                0.699616   0.009496   73.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1856694  on 5852  degrees of freedom
## Residual deviance:  185494  on 5847  degrees of freedom
## AIC: 212064
##
## Number of Fisher Scoring iterations: 5
```

```r
# Model Diagnostics
par(mfrow = c(2,2))
plot(fit)
```



Inference

Statistical

1. We see that deviance residuals are approximately normally distributed as it shows some skewness as and median of deviance residuals is not quite zero.
2. All the explanatory variables are significant in modeling the response variable. The can be interpreted as follows:
    a. The estimated co-effcient for subscribers is 2.18, which suggests that the expected log count for a one unit increase in number of subscribers is 2.18.
    b. The variable zipcode94063 compared zipcode 94041 has an expected log count to decrese by 1.5.
    c. The variable zipcode94107 compared zipcode 94041 has an expected log count to increase by 3.4.
    d. The variable zipcode94301 compared zipcode 94041 has an expected log count to decrese by 1.2.
    e. The variable zipcode95113 compared zipcode 94041 has an expected log count to increase by 0.6.
3. AIC is measure of goodness of fit test, since we did not try fitting any other model to our data, we will ignore the high value of AIC.
4. The residual plots suggests normality of the residuals as the central limit theorem kicks in.

What is means for stakeholders

This model suggests that the expectations of trips to increase is very high for subscribers and the company should come up with attractive offers that would increase their subscribers. The model also suggests that increase in stations in San Francisco and San Jose will lead to a great increase in the number of trips as compare to stations in Mountain view. Also, that decrease in stations in Redwood city and Palo Alto might lead to a some increase in the number of trips as compare to stations in Mountain view. Based on this stakeholders can change the number of station in an optimal manner.

Suggestions

The analysis suggests that SF Bay Bikes should:

1. Create a plan which would allow all private bike owners to share their bikes to reduce cost of new bikes. This plan could be only when the number of trips are the highest to avoid crowding at stations.
2. E-bikes can be a motivating factor to increase the trip duration.
3. Counting the distance and giving points which could be reimbursed at some food joint. SF Bay Bike would have to collaborate with most famous food joint in US. This would also motivate people to bike for longer duration.

Future Work

There are a lot of subsets to this data set which could be analysed on a language and computer which does not have memory problems. I was thinking an interactive plot suggesting the number of available bikes on a station at all times can be created using the status data set. Also, a more optimized predictive model can be created which would would take into account the effects of weather on the number of trips.