Data Analysis on bike share stations

Quoc Hoang Vu

2024-11-22

Import the `dplyr` library to better manipulate data:

```
if (!require("dplyr")) {
  install.packages("dplyr")
  library("dplyr")
}
```

```
## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Import the `ggplot2` library to draw plots:

```
if (!require("ggplot2")) {
  install.packages("ggplot2")
  library("ggplot2")
}
```

```
## Loading required package: ggplot2
```

Import the `igraph` library to work with graphs:

```
if (!require("igraph")) {
  install.packages("igraph")
  library("igraph")
}
```

```
## Loading required package: igraph

##
## Attaching package: 'igraph'
```

```
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

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

We read the following datasets:

- A file comprising a description of each bike share station
- A file containing trips between bike stations on Monday 7 July 2014 to Sunday 13 July 2014
- A file containing trips between bike stations on Monday 5 January 2015 to Sunday 11 January 2015
- A file containing trips between bike stations on Monday 6 July 2015 to Sunday 12 July 2015

```
stations <- read.csv("https://raw.githubusercontent.com/julien-arino/math-of-data-science/refs/heads/ma
trips_07072014_13072014 <- read.csv("https://raw.githubusercontent.com/julien-arino/math-of-data-scienc
trips_05012015_11012015 <- read.csv("https://raw.githubusercontent.com/julien-arino/math-of-data-scienc
trips_06072015_12072015 <- read.csv("https://raw.githubusercontent.com/julien-arino/math-of-data-scienc
```

We wish to get a concrete idea of what each dataset contains. We can use `head()` to get the first few entries of a dataset, and `dim()` to get the size of a dataset.

We examine the bike share station dataset, which we denote as `stations`:

```
head(stations)
```

```
##   id                              name      lat      long dock_count     city
## 1  2 San Jose Diridon Caltrain Station 37.32973 -121.9018         27 San Jose
## 2  3             San Jose Civic Center 37.33070 -121.8890         15 San Jose
## 3  4            Santa Clara at Almaden 37.33399 -121.8949         11 San Jose
## 4  5                  Adobe on Almaden 37.33141 -121.8932         19 San Jose
## 5  6                  San Pedro Square 37.33672 -121.8941         15 San Jose
## 6  7              Paseo de San Antonio 37.33380 -121.8869         15 San Jose
##   installation_date
## 1          8/6/2013
## 2          8/5/2013
## 3          8/6/2013
## 4          8/5/2013
## 5          8/7/2013
## 6          8/7/2013
```

```
dim(stations)
```

```
## [1] 70  7
```

We make the following observations:

- The dataset has 70 entries and 7 columns.
- The first column contains the `id` of the bike share station (interestingly, this doesn't start at 1).
- The second column contains the `name` of the bike share station.
- The third column contains the `latitude` (positionally) of the bike share station.
- The fourth column contains the `longtitude` (positionally) of the bike share station.
- The fifth column contains the `dock_count` of the bike share station (how many bikes are available).
- The sixth column contains the `city` the bike share station is located in.
- The seventh column contains the `installation_date` of the bike share station.

We examine the trips between bike stations on Monday 7 July 2014 to Sunday 13 July 2014 dataset, which we denote as `trips_07072014_13072014`:

```r
head(trips_07072014_13072014)
```

```
##   start_date_yyyymmdd                  start_station_name start_station_id
## 1          2014-07-13        Powell at Post (Union Square)               71
## 2          2014-07-13                        Market at 4th               76
## 3          2014-07-13                        Market at 4th               76
## 4          2014-07-13      Grant Avenue at Columbus Avenue               73
## 5          2014-07-13 Harry Bridges Plaza (Ferry Building)               50
## 6          2014-07-13    San Jose Diridon Caltrain Station                2
##   end_date_yyyymmdd              end_station_name end_station_id duration
## 1        2014-07-13         Embarcadero at Bryant             54      667
## 2        2014-07-13                 Market at 10th             67      401
## 3        2014-07-13                 Market at 10th             67      401
## 4        2014-07-13 Powell at Post (Union Square)             71      470
## 5        2014-07-13                  Howard at 2nd             63      421
## 6        2014-07-13        Santa Clara at Almaden              4      221
```

```r
dim(trips_07072014_13072014)
```

```
## [1] 6911    7
```

We make the following observations:

- The dataset has 6911 entries and 7 columns.
- The first column contains the `start_date_yyyymmdd` of the trip, or the starting date.
- The second column contains the `start_station_name` of the trip, or the name of the starting station.
- The third column contains the `start_station_id` of the trip, or the ID of the starting station.
- The fourth column contains the `end_date_yyyymmdd` of the trip, or the starting date.
- The fifth column contains the `end_station_name` of the trip, or the name of the ending station.
- The sixth column contains the `end_station_id` of the trip, or the ID of the ending station.
- The seventh column contains the `duration` of the trip in seconds.

We examine the trips between bike stations on Monday 5 January 2015 to Sunday 11 January 2015 dataset, which we denote as `trips_05012015_11012015`:

```r
head(trips_05012015_11012015)
```

```
##   start_date_yyyymmdd                        start_station_name start_station_id
## 1          2015-01-11                      Embarcadero at Sansome               60
```

```
## 2          2015-01-11        San Jose Diridon Caltrain Station               2
## 3          2015-01-11 San Francisco Caltrain (Townsend at 4th)              70
## 4          2015-01-11                           5th at Howard              57
## 5          2015-01-11          Grant Avenue at Columbus Avenue              73
## 6          2015-01-11                           5th at Howard              57
##   end_date_yyyymmdd                 end_station_name end_station_id duration
## 1        2015-01-11 Civic Center BART (7th at Market)             72      962
## 2        2015-01-11             Santa Clara at Almaden              4      236
## 3        2015-01-11     Powell at Post (Union Square)             71      653
## 4        2015-01-11     Powell at Post (Union Square)             71      265
## 5        2015-01-11                Powell Street BART             39      608
## 6        2015-01-11            San Francisco City Hall             58      540
```

```r
dim(trips_05012015_11012015)
```

```
## [1] 6899    7
```

We can see the columns are the same as the prior dataset. There are 6899 entries in the dataset.

We examine the trips between bike stations on Monday 6 July 2015 to Sunday 12 July 2015 dataset, which we denote as `trips_06072015_12072015`:

```r
head(trips_06072015_12072015)
```

```
##   start_date_yyyymmdd                            start_station_name
## 1          2015-07-12                                 Howard at 2nd
## 2          2015-07-12 Temporary Transbay Terminal (Howard at Beale)
## 3          2015-07-12                            San Jose City Hall
## 4          2015-07-12                               Clay at Battery
## 5          2015-07-12                             Market at Sansome
## 6          2015-07-12                              Davis at Jackson
##   start_station_id end_date_yyyymmdd                end_station_name
## 1               63        2015-07-12               Market at Sansome
## 2               55        2015-07-12              Powell Street BART
## 3               10        2015-07-12      SJSU - San Salvador at 9th
## 4               41        2015-07-12             Washington at Kearny
## 5               77        2015-07-12 Grant Avenue at Columbus Avenue
## 6               42        2015-07-12                  Spear at Folsom
##   end_station_id duration
## 1             77      121
## 2             39      444
## 3             16      444
## 4             46      166
## 5             73      624
## 6             49      363
```

```r
dim(trips_06072015_12072015)
```

```
## [1] 7381    7
```

We can again see the columns are the same as the prior dataset. There are 7381 entries in the dataset.

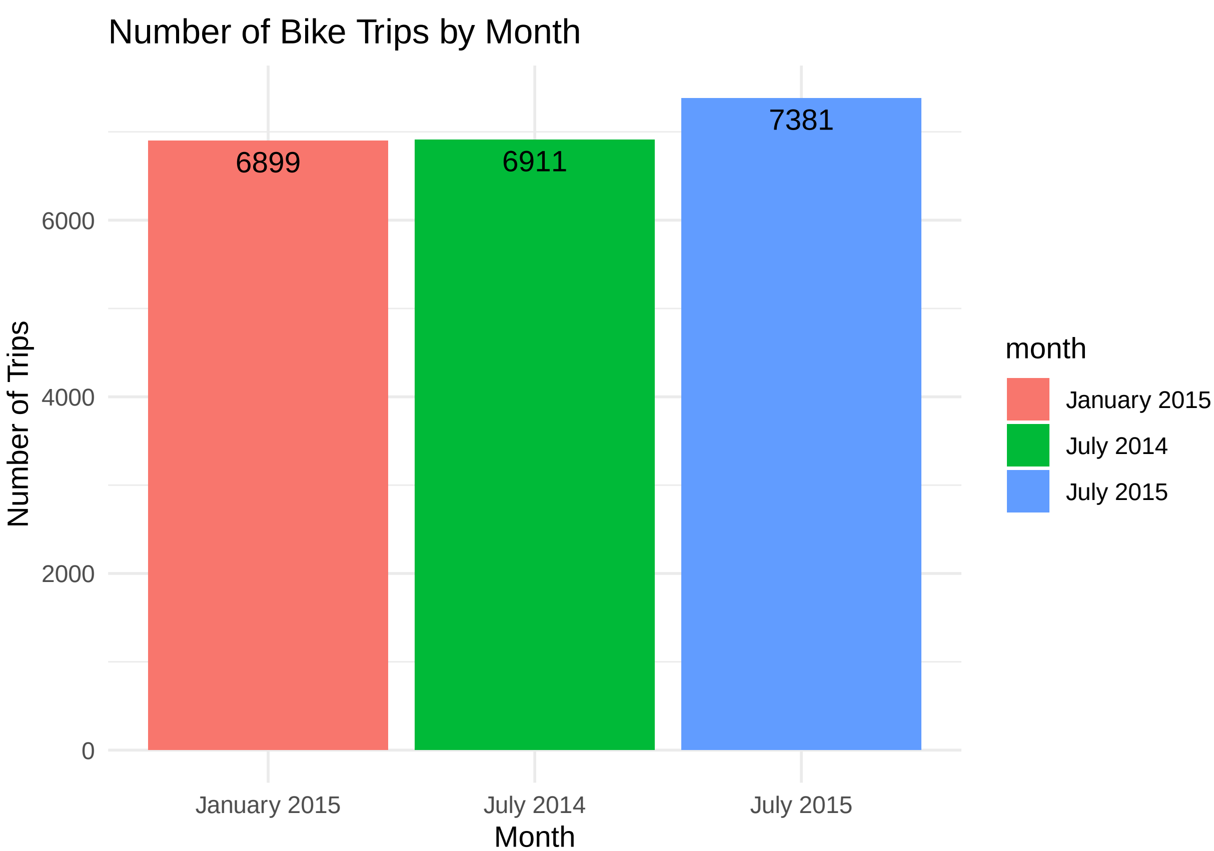
To further the analysis, we wish to merge the datasets into one big dataset. To better differentiate which data item belongs to which initial dataset, we add an extra column to each dataset called `month` describing the month of each dataset:

- `trips_07072014_13072014` will be `July 2014`
- `trips_05012015_11012015` will be `January 2015`
- `trips_06072015_12072015` will be `July 2015`

```
trips_07072014_13072014$month <- "July 2014"
trips_05012015_11012015$month <- "January 2015"
trips_06072015_12072015$month <- "July 2015"

trips <- rbind(trips_07072014_13072014, trips_05012015_11012015, trips_06072015_12072015)
```

Now, we plot respectively how many times there was a bike trip in the first full week of each recorded month:

```
ggplot(trips, aes(x = month, y = after_stat(count),  fill = month)) +
  geom_bar() +
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = 1.5) +
  labs(title = "Number of Bike Trips by Month",
       x = "Month",
       y = "Number of Trips") +
  theme_minimal()
```



We can observe that July 2015 had the highest number of trips (7381), followed by July 2014 (6911), and January 2015 had the least (6899). This suggests a seasonal pattern, with more trips occurring in summer months compared to winter.

Regardless, this doesn't tell us much: we only have three data points compromising of how many total trips are made. This would not suggest, for example, the number of trips made in January 2016 would be less than July 2015, or the number of trips made in July 2016 would increase from prior months.

The data we *do* have an abundance of is the individual trips themselves. Using `dim()` on the combined dataset:

```r
dim(trips)
```

```
## [1] 21191     8
```

There are 21191 total rows across the three datasets, and each row is an individual trip. We can, for example, identify the most popular starting and ending stations for each month. We will generate data frames for the top 5 most popular starting stations and ending stations for each dataset, with their corresponding frequency in the dataset.

For Monday 7 July 2014 to Sunday 13 July 2014:

```r
top_stations_j14 <- trips_07072014_13072014 %>%
  reframe(
    top_start = names(sort(table(start_station_name), decreasing = TRUE)[1:5]),
    top_start_freq = sort(table(start_station_name), decreasing = TRUE)[1:5],
    top_end = names(sort(table(end_station_name), decreasing = TRUE)[1:5]),
    top_end_freq = sort(table(end_station_name), decreasing = TRUE)[1:5]
  )

top_stations_j14
```

```
##                                       top_start top_start_freq
## 1      San Francisco Caltrain (Townsend at 4th)            546
## 2          Harry Bridges Plaza (Ferry Building)            315
## 3       San Francisco Caltrain 2 (330 Townsend)            314
## 4                              Market at Sansome            302
## 5 Temporary Transbay Terminal (Howard at Beale)            275
##                                  top_end top_end_freq
## 1 San Francisco Caltrain (Townsend at 4th)          716
## 2                        Market at Sansome          341
## 3  San Francisco Caltrain 2 (330 Townsend)          339
## 4     Harry Bridges Plaza (Ferry Building)          333
## 5                    Embarcadero at Sansome          325
```

For Monday 5 January 2015 to Sunday 11 January 2015:

```r
top_stations_a15 <- trips_05012015_11012015 %>%
  reframe(
    top_start = names(sort(table(start_station_name), decreasing = TRUE)[1:5]),
    top_start_freq = sort(table(start_station_name), decreasing = TRUE)[1:5],
    top_end = names(sort(table(end_station_name), decreasing = TRUE)[1:5]),
    top_end_freq = sort(table(end_station_name), decreasing = TRUE)[1:5]
  )

top_stations_a15
```

```
##                                       top_start top_start_freq
## 1      San Francisco Caltrain (Townsend at 4th)            643
## 2       San Francisco Caltrain 2 (330 Townsend)            372
```

```
## 3          Harry Bridges Plaza (Ferry Building)           323
## 4                                Townsend at 7th           296
## 5 Temporary Transbay Terminal (Howard at Beale)           291
##                                 top_end top_end_freq
## 1 San Francisco Caltrain (Townsend at 4th)          809
## 2  San Francisco Caltrain 2 (330 Townsend)          401
## 3     Harry Bridges Plaza (Ferry Building)          324
## 4                          Townsend at 7th          320
## 5                          2nd at Townsend          286
```

For 6 July 2015 to Sunday 12 July 2015:

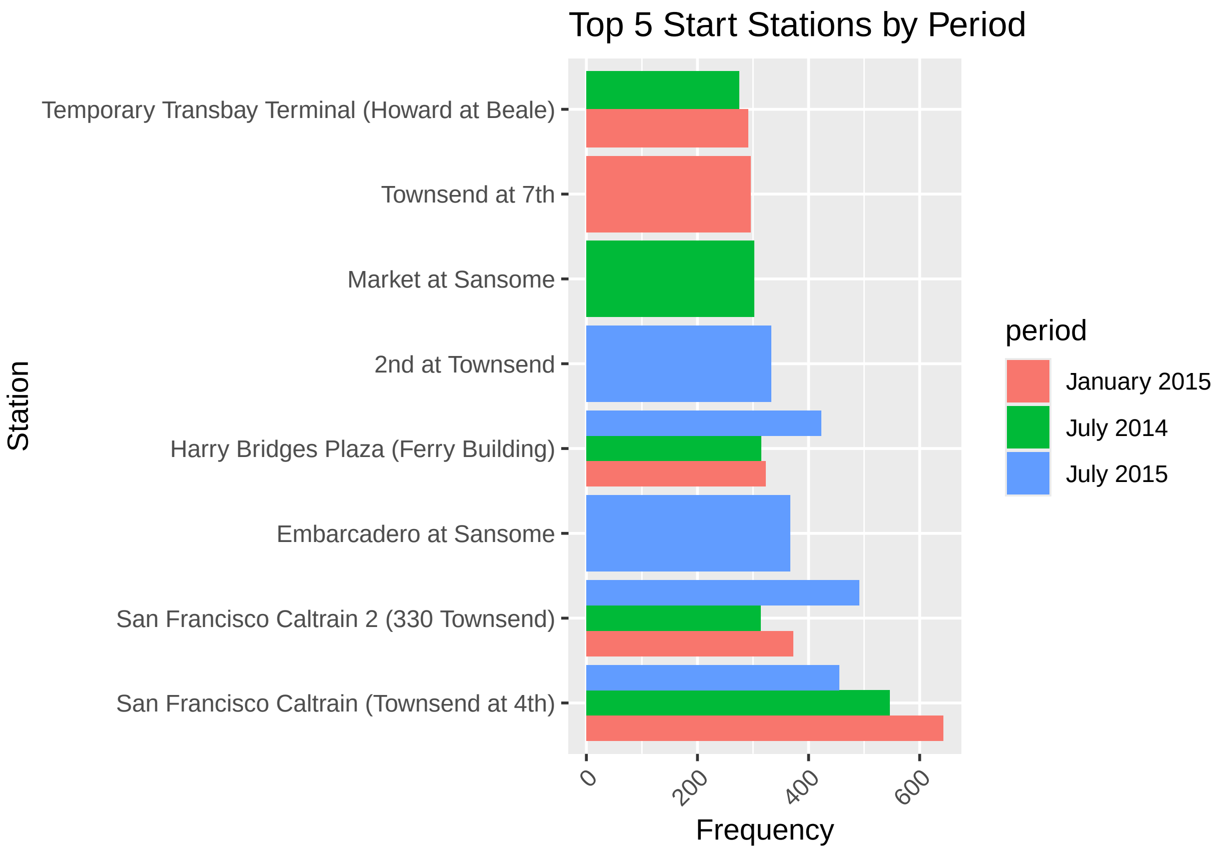
```
top_stations_j15 <- trips_06072015_12072015 %>%
  reframe(
    top_start = names(sort(table(start_station_name), decreasing = TRUE)[1:5]),
    top_start_freq = sort(table(start_station_name), decreasing = TRUE)[1:5],
    top_end = names(sort(table(end_station_name), decreasing = TRUE)[1:5]),
    top_end_freq = sort(table(end_station_name), decreasing = TRUE)[1:5]
  )

top_stations_j15
```

```
##                                top_start top_start_freq
## 1  San Francisco Caltrain 2 (330 Townsend)            491
## 2 San Francisco Caltrain (Townsend at 4th)            455
## 3     Harry Bridges Plaza (Ferry Building)            423
## 4                   Embarcadero at Sansome            367
## 5                          2nd at Townsend            333
##                                  top_end top_end_freq
## 1 San Francisco Caltrain (Townsend at 4th)          639
## 2  San Francisco Caltrain 2 (330 Townsend)          533
## 3     Harry Bridges Plaza (Ferry Building)          411
## 4                   Embarcadero at Sansome          403
## 5                          2nd at Townsend          327
```

We better visualize the previous data using `ggplot`:

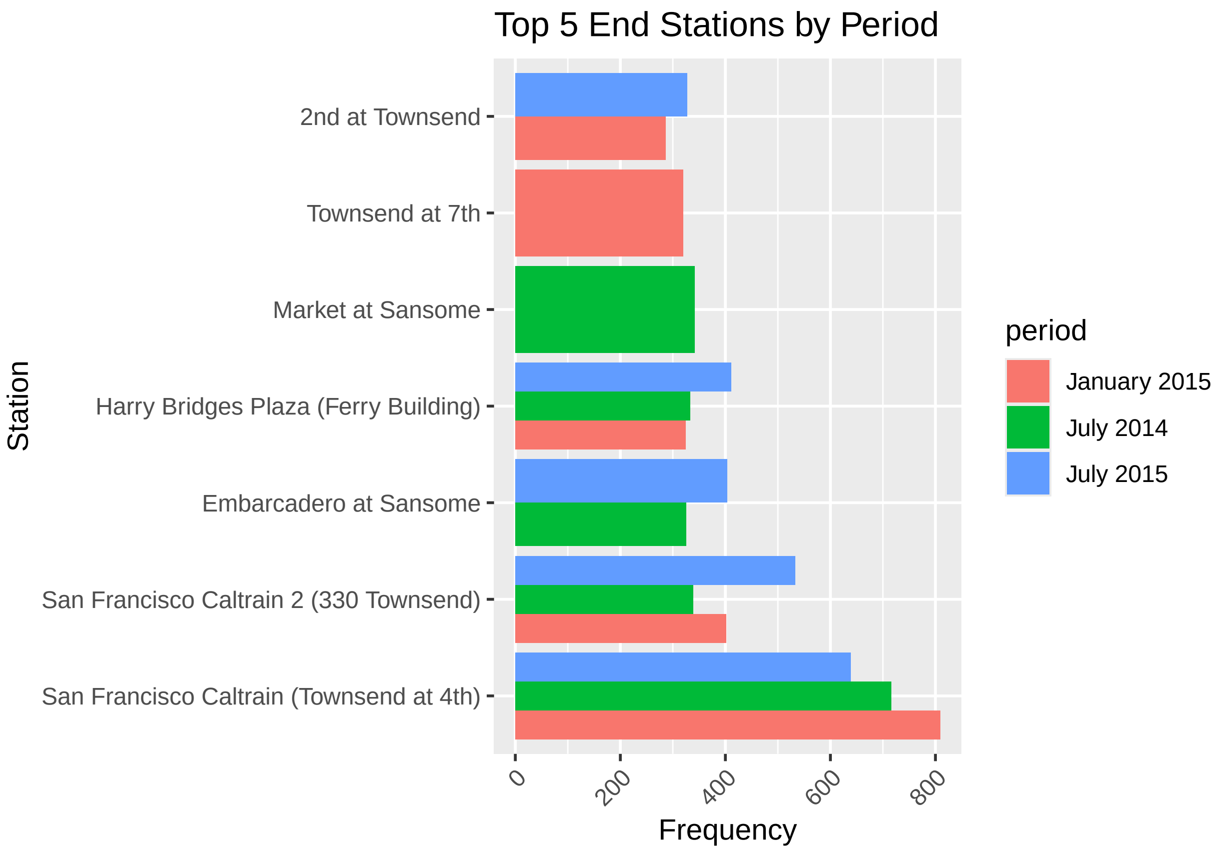
```
top_stations <- as.data.frame(rbind(
  mutate(top_stations_j14, period = "July 2014"),
  mutate(top_stations_a15, period = "January 2015"),
  mutate(top_stations_j15, period = "July 2015")
))

ggplot(top_stations, aes(x = reorder(top_start, -top_start_freq), y = top_start_freq, fill = period)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Top 5 Start Stations by Period", x = "Station", y = "Frequency") +
  coord_flip()
```

```
## Don't know how to automatically pick scale for object of type <table>.
## Defaulting to continuous.
```




```r
ggplot(top_stations, aes(x = reorder(top_end, -top_end_freq), y = top_end_freq, fill = period)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Top 5 End Stations by Period", x = "Station", y = "Frequency") +
  coord_flip()
```

```
## Don't know how to automatically pick scale for object of type <table>.
## Defaulting to continuous.
```



Now we start to notice some interesting observations:

- The most popular stations remain relatively consistent across the three time periods, with some variations in ranking.
- San Francisco Caltrain stations (both at Townsend St and 4th St) consistently appear as top starting and ending points. This suggests these stations are major transportation hubs, likely due to their proximity to the Caltrain station.
- Harry Bridges Plaza (Ferry Building) is another consistently popular location, likely due to its central location and connection to ferry services. Harry Bridges Plaza (Ferry Building) is more popular in summer months (July 2014 and July 2015) compared to winter (January 2015). This could be due to increased tourism and better weather conditions in summer.
- Stations near major transit hubs (Caltrain, Ferry Building) are consistently popular. This suggests many users are using bike-sharing for the "last mile" of their commute.

To further understand these hot spots, we wish to model the network of bike stations as a graph. We will focus on only the top stations, so we want to define a function that generate a graph given the dataset and the stations we're interested in (so the graph will only include trips to and from said stations).

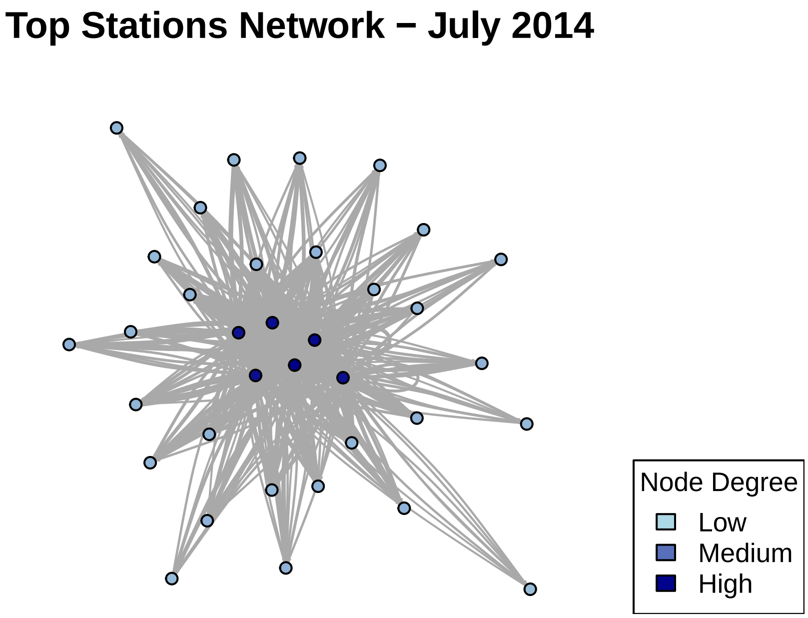
```
create_focused_graph <- function(trips_data, top_stations) {
  top_station_ids <- unique(c(top_stations$top_start, top_stations$top_end))
  filtered_trips <- trips_data %>%
    filter(start_station_name %in% top_station_ids | end_station_name %in% top_station_ids)

  edges <- filtered_trips %>%
    group_by(start_station_name, end_station_name) %>%
    summarise(weight = n(), .groups = "drop")
```

```
  graph <- graph_from_data_frame(edges, directed = TRUE)
  return(graph)
}
```

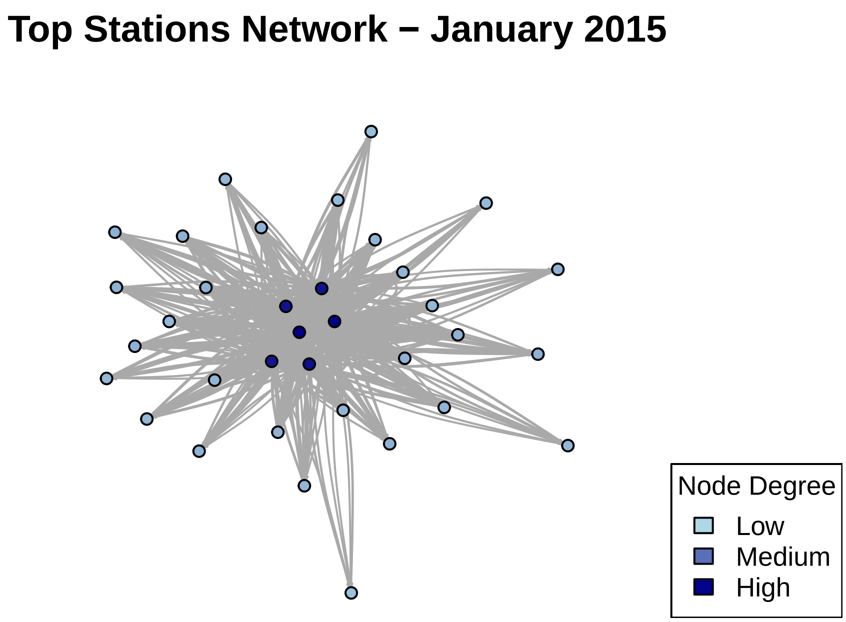
We now create and plot graphs for each dataset:

```
graph_j14 <- create_focused_graph(trips_07072014_13072014, top_stations_j14)
graph_a15 <- create_focused_graph(trips_05012015_11012015, top_stations_a15)
graph_j15 <- create_focused_graph(trips_06072015_12072015, top_stations_j15)

plot_focused_graph <- function(graph, title) {
  set.seed(123)  # For reproducibility

  # Calculate node degrees
  node_degrees <- degree(graph, mode = "total")

  # Create a color palette based on node degrees
  color_palette <- colorRampPalette(c("lightblue", "darkblue"))(max(node_degrees) + 1)

  # Set node colors based on degree
  V(graph)$color <- color_palette[node_degrees + 1]

  # Calculate edge weights
  edge_weights <- E(graph)$weight

  # Normalize edge weights for visualization
  normalized_weights <- (edge_weights - min(edge_weights)) / (max(edge_weights) - min(edge_weights))

  # Set edge width based on normalized weights
  E(graph)$width <- 1 + 5 * normalized_weights

  # Use Fruchterman-Reingold layout for better spacing
  layout <- layout_with_fr(graph)

  # Plot the graph
  plot(graph,
       layout = layout,
       vertex.size = 5,  # Adjust node size based on degree
       vertex.label = "",
       vertex.label.cex = 0.6,
       vertex.label.color = "black",
       vertex.label.dist = 0.5,
       edge.arrow.size = 0.1,
       edge.curved = 0.1,
       main = title)

  # Add a legend for node degrees
  legend("bottomright",
         legend = c("Low", "Medium", "High"),
         fill = color_palette[c(1, floor(max(node_degrees)/2), max(node_degrees))],
         title = "Node Degree",
         cex = 0.8)
}
```

```r
plot_focused_graph(graph_j14, "Top Stations Network - July 2014")
```

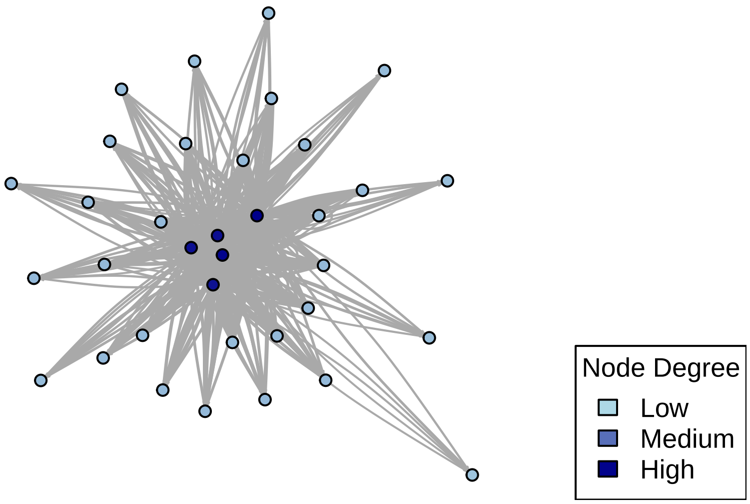



```r
plot_focused_graph(graph_a15, "Top Stations Network - January 2015")
```




```r
plot_focused_graph(graph_j15, "Top Stations Network - July 2015")
```

Top Stations Network – July 2015



Though the graph's too cluttered… we note that there are only 6 central stations (5 in the case of July 2015) that connect to and from other stations. We may want to calculate some graph measures of these graphs, such as order, size, density, average degree, diameter, and average path length:

```
calculate_graph_measures <- function(graph) {
  list(
    nodes = vcount(graph),
    edges = ecount(graph),
    density = edge_density(graph),
    avg_degree = mean(degree(graph)),
    diameter = diameter(graph),
    avg_path_length = average.path.length(graph)
  )
}

measures_j14 <- calculate_graph_measures(graph_j14)
```

```
## Warning: 'average.path.length()' was deprecated in igraph 2.0.0.
## i Please use 'mean_distance()' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

```
measures_a15 <- calculate_graph_measures(graph_a15)
measures_j15 <- calculate_graph_measures(graph_j15)

measures_df <- data.frame(
  Period = c("July 2014", "January 2015", "July 2015"),
  Nodes = c(measures_j14$nodes, measures_a15$nodes, measures_j15$nodes),
  Edges = c(measures_j14$edges, measures_a15$edges, measures_j15$edges),
  Density = c(measures_j14$density, measures_a15$density, measures_j15$density),
  Avg_Degree = c(measures_j14$avg_degree, measures_a15$avg_degree, measures_j15$avg_degree),
  Diameter = c(measures_j14$diameter, measures_a15$diameter, measures_j15$diameter),
  Avg_Path_Length = c(measures_j14$avg_path_length, measures_a15$avg_path_length, measures_j15$avg_path_
```

```
)

measures_df
```

```
##          Period Nodes Edges   Density Avg_Degree Diameter Avg_Path_Length
## 1    July 2014    35   361 0.3033613   20.62857       19        5.523529
## 2 January 2015    35   349 0.2932773   19.94286       20        5.480672
## 3    July 2015    35   308 0.2588235   17.60000       26        7.969748
```

Now this is interesting: in this subgraph concerning these central stations, the size, density, and average degree decrease over time and the diameter and average path length increase over time. The decreases suggest that the network is becoming less interconnected among these key stations. As well, the increases indicate that trips between these central stations are becoming more indirect or require more intermediate stops.

These trends, combined with the overall increase in trip volume by July 2015, suggest several possibilities:

- **Decentralization**: The bike-sharing network may be expanding beyond the initial core stations, with new popular routes emerging in other areas of the city.
- **Changed Usage Patterns**: Riders might be using the bikes for longer, more diverse trips rather than just shuttling between major transit hubs.
- **System Growth**: The increase in total trips despite decreased connectivity among central stations implies that the system is growing in other areas, possibly with new stations being added or becoming more popular.

Some further analysis we can perform to better understand this include checking if new stations were added, comparing average trip durations, and mapping the stations by their geographical location.