Analyse_2

Georgie Knight

25 August 2016

```
library("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
```

```
library("tidyr")
library("lubridate")
```

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

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

```
library("readr")
library("ggplot2")
library("ggmap")
trip_read    <- read_csv("trip_full_updated.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   Start_trip = col_datetime(format = ""),
##   Start.Station = col_character(),
##   End_trip = col_datetime(format = ""),
##   End.Station = col_character(),
##   Subscriber.Type = col_character(),
##   Zip.Code = col_character(),
##   Date = col_datetime(format = ""),
##   landmark = col_character(),
##   start_lat = col_double(),
##   start_long = col_double(),
##   end_lat = col_double(),
##   end_long = col_double(),
##   Max.Sea.Level.PressureIn = col_double(),
```

```
##   Mean.Sea.Level.PressureIn = col_double(),
##   Min.Sea.Level.PressureIn = col_double(),
##   Events = col_character(),
##   Weekday = col_character()
## )

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

```
status_read  <- read_csv("status_full_updated.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   time = col_datetime(format = ""),
##   name = col_character(),
##   lat = col_double(),
##   long = col_double(),
##   landmark = col_character(),
##   installation = col_date(format = ""),
##   Date = col_date(format = ""),
##   Max.Sea.Level.PressureIn = col_double(),
##   Mean.Sea.Level.PressureIn = col_double(),
##   Min.Sea.Level.PressureIn = col_double(),
##   PrecipitationIn = col_character(),
##   Events = col_character()
## )
## See spec(...) for full column specifications.
```

```
trip         <- dplyr::tbl_df(trip_read)
status       <- dplyr::tbl_df(status_read)

glimpse(trip)
```

```
## Observations: 354,152
## Variables: 41
## $ Trip.ID                  <int> 913460, 913459, 913455, 913454, 9134...
## $ Duration                 <int> 765, 1036, 307, 409, 789, 293, 896, ...
## $ Start_trip               <time> 2015-08-31 23:26:00, 2015-08-31 23:...
## $ Start.Station            <chr> "Harry Bridges Plaza (Ferry Building...
## $ Start.Terminal           <int> 50, 31, 47, 10, 51, 68, 51, 60, 56, ...
## $ End_trip                 <time> 2015-08-31 23:39:00, 2015-08-31 23:...
## $ End.Station              <chr> "San Francisco Caltrain (Townsend at...
## $ End.Terminal             <int> 70, 27, 64, 8, 60, 70, 60, 74, 55, 6...
## $ Bike..                   <int> 288, 35, 468, 68, 487, 538, 363, 470...
## $ Subscriber.Type          <chr> "Subscriber", "Subscriber", "Subscri...
## $ Zip.Code                 <chr> "2139", "95032", "94107", "95113", "...
## $ Date                     <time> 2015-08-31, 2015-08-31, 2015-08-31,...
## $ landmark                 <chr> "San Francisco", "Mountain View", "S...
## $ start_lat                <dbl> 37.79539, 37.40044, 37.78898, 37.337...
## $ start_long               <dbl> -122.3942, -122.1083, -122.4035, -12...
## $ end_lat                  <dbl> 37.77662, 37.38922, 37.78226, 37.330...
## $ end_long                 <dbl> -122.3953, -122.0819, -122.3927, -12...
```

```
## $ Max.TemperatureF          <int> 78, 82, 78, 85, 78, 78, 78, 78, 78, ...
## $ Mean.TemperatureF         <int> 69, 72, 69, 72, 69, 69, 69, 69, 69, ...
## $ Min.TemperatureF          <int> 60, 61, 60, 59, 60, 60, 60, 60, 60, ...
## $ Max.Dew.PointF            <int> 58, 62, 58, 59, 58, 58, 58, 58, 58, ...
## $ MeanDew.PointF            <int> 57, 56, 57, 55, 57, 57, 57, 57, 57, ...
## $ Min.DewpointF             <int> 54, 52, 54, 51, 54, 54, 54, 54, 54, ...
## $ Max.Humidity              <int> 84, 84, 84, 84, 84, 84, 84, 84, 84, ...
## $ Mean.Humidity             <int> 67, 63, 67, 58, 67, 67, 67, 67, 67, ...
## $ Min.Humidity              <int> 50, 42, 50, 32, 50, 50, 50, 50, 50, ...
## $ Max.Sea.Level.PressureIn  <dbl> 29.95, 29.97, 29.95, 29.95, 29.95, 2...
## $ Mean.Sea.Level.PressureIn <dbl> 29.91, 29.92, 29.91, 29.90, 29.91, 2...
## $ Min.Sea.Level.PressureIn  <dbl> 29.87, 29.86, 29.87, 29.85, 29.87, 2...
## $ Max.VisibilityMiles       <int> 10, 10, 10, 10, 10, 10, 10, 10, 10, ...
## $ Mean.VisibilityMiles      <int> 10, 10, 10, 10, 10, 10, 10, 10, 10, ...
## $ Min.VisibilityMiles       <int> 9, 10, 9, 10, 9, 9, 9, 9, 9, 9, 9, 9...
## $ Max.Wind.SpeedMPH         <int> 18, 22, 18, 20, 18, 18, 18, 18, 18, ...
## $ Mean.Wind.SpeedMPH        <int> 9, 6, 9, 6, 9, 9, 9, 9, 9, 9, 9, 9, ...
## $ Max.Gust.SpeedMPH         <int> 21, 25, 21, 24, 21, 21, 21, 21, 21, ...
## $ PrecipitationIn           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ CloudCover                <int> 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ Events                    <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ WindDirDegrees            <int> 246, 6, 246, 308, 246, 246, 246, 246...
## $ Zip                       <int> 94107, 94041, 94107, 95113, 94107, 9...
## $ Weekday                   <chr> "Monday", "Monday", "Monday", "Monda...
```

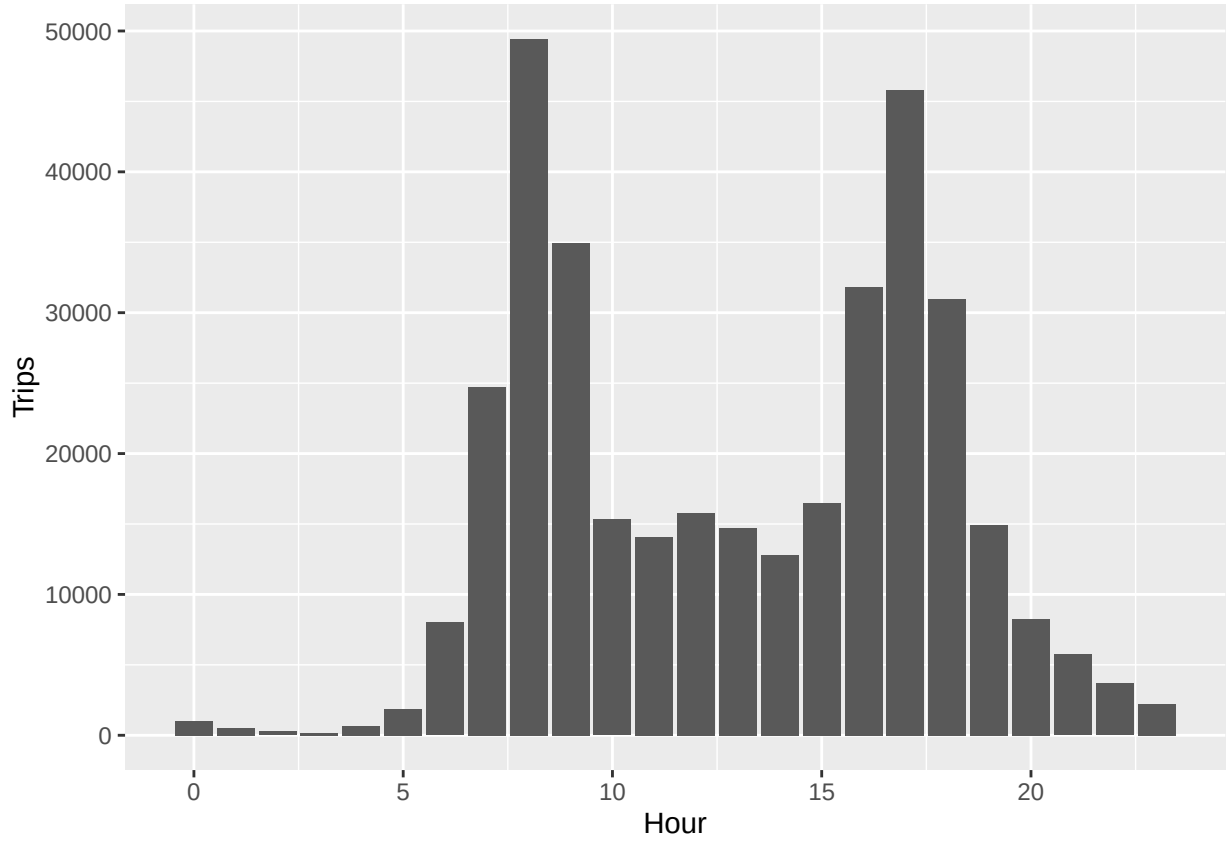
```
glimpse(status)
```

```
## Observations: 1,135,974
## Variables: 33
## $ station_id                <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...
## $ bikes_available           <int> 15, 14, 15, 14, 15, 14, 15, 14, 15, ...
## $ docks_available           <int> 12, 13, 12, 13, 12, 13, 12, 13, 12, ...
## $ time                      <time> 2014-09-01 00:00:03, 2014-09-01 02:...
## $ name                      <chr> "San Jose Diridon Caltrain Station",...
## $ lat                       <dbl> 37.32973, 37.32973, 37.32973, 37.329...
## $ long                      <dbl> -121.9018, -121.9018, -121.9018, -12...
## $ landmark                  <chr> "San Jose", "San Jose", "San Jose", ...
## $ installation              <date> 2013-08-29, 2013-08-29, 2013-08-29,...
## $ Date                      <date> 2014-09-01, 2014-09-01, 2014-09-01,...
## $ Max.TemperatureF          <int> 86, 86, 86, 86, 86, 86, 86, 86, 86, ...
## $ Mean.TemperatureF         <int> 72, 72, 72, 72, 72, 72, 72, 72, 72, ...
## $ Min.TemperatureF          <int> 58, 58, 58, 58, 58, 58, 58, 58, 58, ...
## $ Max.Dew.PointF            <int> 60, 60, 60, 60, 60, 60, 60, 60, 60, ...
## $ MeanDew.PointF            <int> 54, 54, 54, 54, 54, 54, 54, 54, 54, ...
## $ Min.DewpointF             <int> 50, 50, 50, 50, 50, 50, 50, 50, 50, ...
## $ Max.Humidity              <int> 86, 86, 86, 86, 86, 86, 86, 86, 86, ...
## $ Mean.Humidity             <int> 59, 59, 59, 59, 59, 59, 59, 59, 59, ...
## $ Min.Humidity              <int> 31, 31, 31, 31, 31, 31, 31, 31, 31, ...
## $ Max.Sea.Level.PressureIn  <dbl> 29.86, 29.86, 29.86, 29.86, 29.86, 2...
## $ Mean.Sea.Level.PressureIn <dbl> 29.81, 29.81, 29.81, 29.81, 29.81, 2...
## $ Min.Sea.Level.PressureIn  <dbl> 29.75, 29.75, 29.75, 29.75, 29.75, 2...
## $ Max.VisibilityMiles       <int> 10, 10, 10, 10, 10, 10, 10, 10, 10, ...
## $ Mean.VisibilityMiles      <int> 10, 10, 10, 10, 10, 10, 10, 10, 10, ...
## $ Min.VisibilityMiles       <int> 10, 10, 10, 10, 10, 10, 10, 10, 10, ...
```

```
## $ Max.Wind.SpeedMPH         <int> 17, 17, 17, 17, 17, 17, 17, 17, 17, ...
## $ Mean.Wind.SpeedMPH        <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ...
## $ Max.Gust.SpeedMPH         <int> 22, 22, 22, 22, 22, 22, 22, 22, 22, ...
## $ PrecipitationIn           <chr> "0", "0", "0", "0", "0", "0", "0", "...
## $ CloudCover                <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ Events                    <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ WindDirDegrees            <int> 296, 296, 296, 296, 296, 296, 296, 2...
## $ Zip                       <int> 95113, 95113, 95113, 95113, 95113, 9...
```

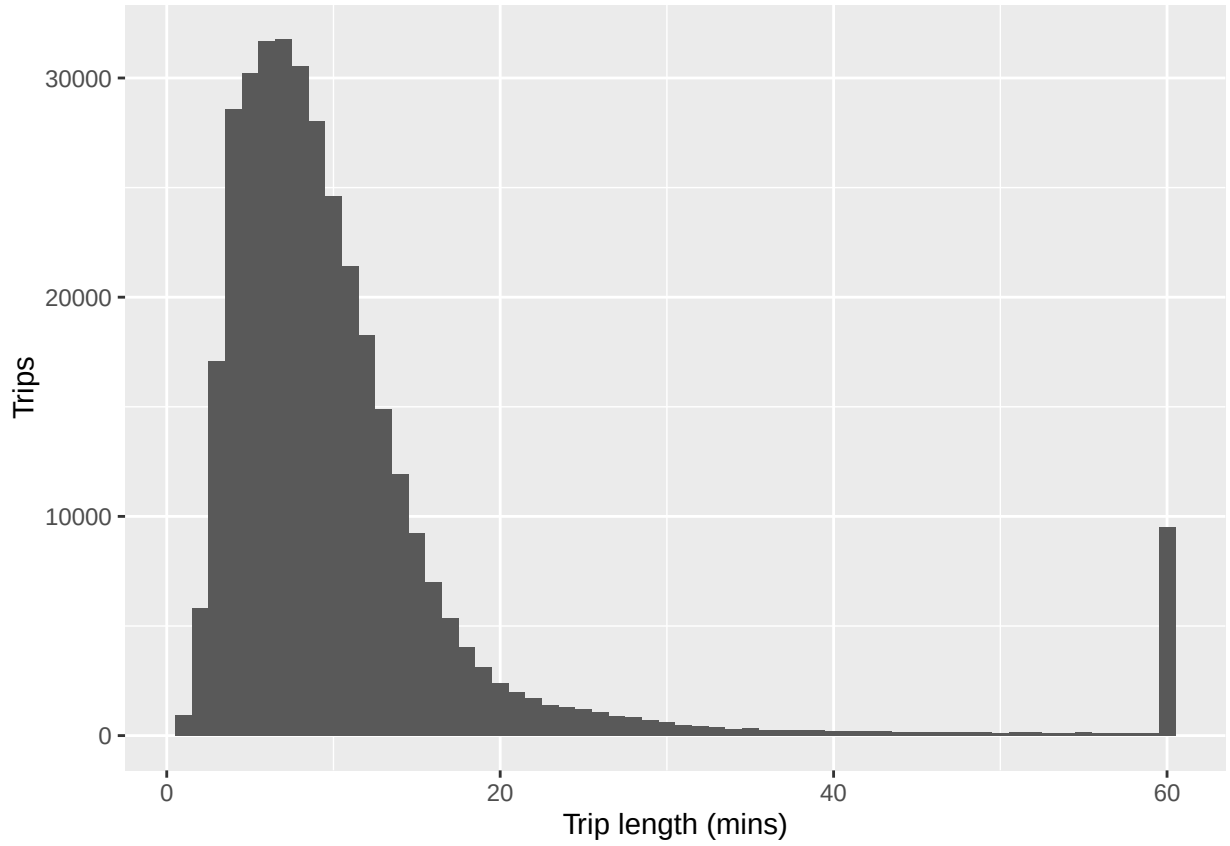
Let's have a look at how the trips vary by time of day:

```
trip_per_hour <- trip %>%
  mutate(Hour = hour(Start_trip))

ggplot(trip_per_hour, aes(x = Hour))+
  geom_bar()+
  labs(x= "Hour", y="Trips")
```



How about the length of the trip

```
trip_l <- trip %>%
  mutate(Duration = Duration/60) %>%
  mutate(Duration = ifelse(Duration >60,60, Duration))
ggplot(trip_l, aes(x= Duration))+
  geom_histogram(binwidth = 1)+
  labs(x= "Trip length (mins)", y = "Trips")
```




```
trip  %>%  filter(Duration > 7* 24* 60 * 60)  %>%
  select(Start.Station, End.Station, Duration, Date, Subscriber.Type, Bike..)
```

```
## # A tibble: 8 x 6
##                   Start.Station
##                           <chr>
## 1        University and Emerson
## 2             Market at Sansome
## 3    Palo Alto Caltrain Station
## 4   San Antonio Shopping Center
## 5        University and Emerson
## 6         San Jose Civic Center
## 7      South Van Ness at Market
## 8 Redwood City Caltrain Station
## # ... with 5 more variables: End.Station <chr>, Duration <int>,
## #   Date <time>, Subscriber.Type <chr>, Bike.. <int>
```

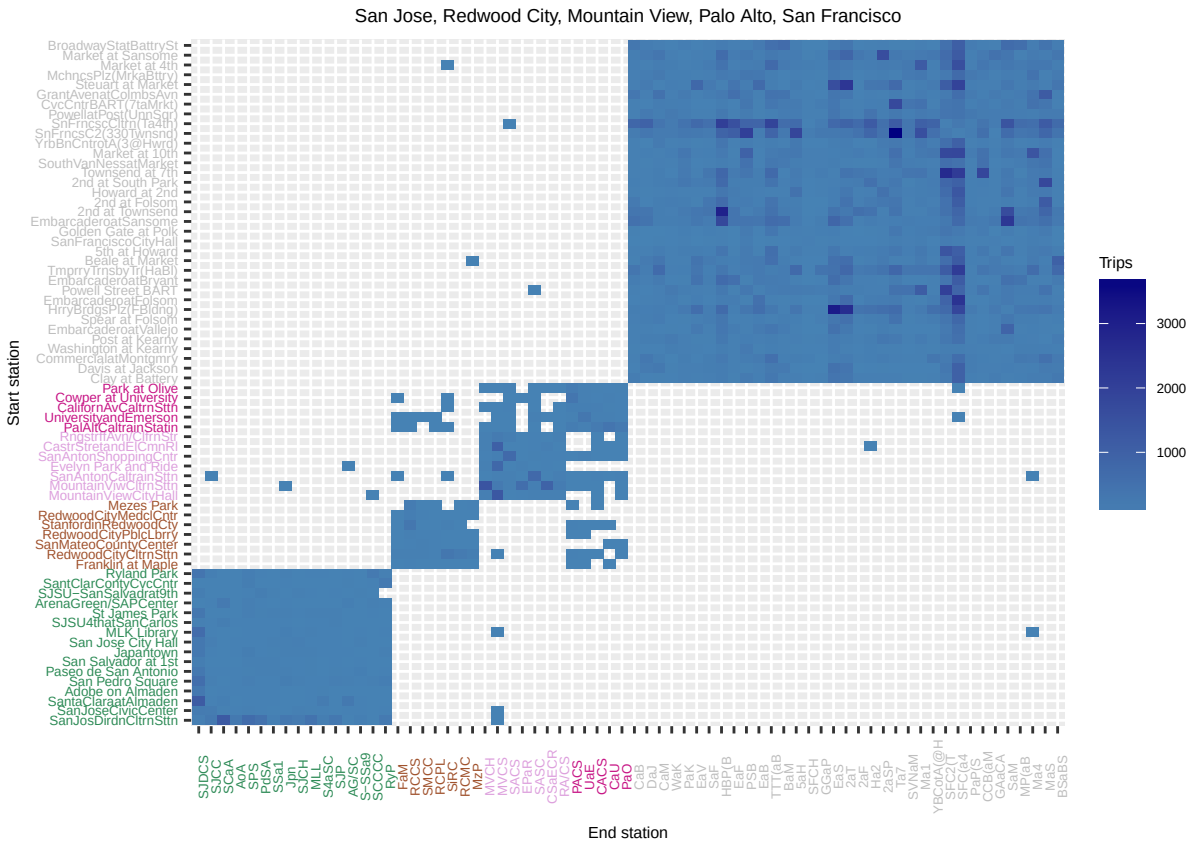
Let's plot a heat map of all the trips made.

```
station <-read_csv("201508_station_data.csv")
```

```
## Parsed with column specification:
## cols(
##   station_id = col_integer(),
```

```
##   name = col_character(),
##   lat = col_double(),
##   long = col_double(),
##   dockcount = col_integer(),
##   landmark = col_character(),
##   installation = col_character()
## )
```

```
station <- dplyr::tbl_df(station)
station <- station %>%
  select(station_id, name, landmark) %>%
  mutate(landmark = factor(landmark, levels = c("San Jose", "Redwood City", "Mountain View", "Palo Alto
    arrange(landmark)

stationLabels <- station$name
stationLabels[39] = "Washington at Kearny"
stationLabels[40] = "Post at Kearny"
stationidsy <- abbreviate(stationLabels,20)
stationidsx <- abbreviate(stationLabels,3)

trip_numbers <- trip  %>%
  group_by(Start.Station, End.Station)  %>%
  summarise(trips = n())

# We'll colour the axis labels by region
myPalette <- c("SeaGreen", "Sienna", "Plum", "MediumVioletRed", "Grey")
names(myPalette) <- levels(station$landmark)



ggplot(trip_numbers, aes(y=Start.Station, x=End.Station))+
    geom_tile(aes(fill = trips))+
    scale_fill_gradient2(low ="PowderBlue",mid = "SteelBlue", high = "Navy")+
  theme(text = element_text(size=6), axis.text.x = element_text(angle=90, vjust=1))+
  scale_x_discrete(limits=stationLabels, labels = stationidsx)+
  scale_y_discrete(limits=stationLabels, labels = stationidsy)+
  labs(x="End station", y= "Start station", fill = "Trips", title = "San Jose, Redwood City, Mountain Vi
  theme(axis.text.x = element_text(colour=myPalette[station$landmark]),
        axis.text.y = element_text(colour=myPalette[station$landmark]))
```



Second way to do it. We'll rewrite the station ids so they match areas

```
station <-read_csv("201508_station_data.csv")
```

```
## Parsed with column specification:
## cols(
##   station_id = col_integer(),
##   name = col_character(),
##   lat = col_double(),
##   long = col_double(),
##   dockcount = col_integer(),
##   landmark = col_character(),
##   installation = col_character()
## )
```

```
station <- dplyr::tbl_df(station)
station <- station %>%
  select(station_id, landmark) %>%
  arrange(landmark) %>%
  mutate(myStationCode = c(1:70))

trip <- trip %>%
  select(-landmark) %>%
  left_join(station, by = c("Start.Terminal"= "station_id")) %>%
  rename(StartStationCode= myStationCode, startLandmark =landmark) %>%
```

```
  left_join(station, by = c("End.Terminal"= "station_id")) %>%
  rename(EndStationCode= myStationCode, endLandmark = landmark)

status <- left_join(status, station)
```

```
## Joining, by = c("station_id", "landmark")
```

```
#write_csv(trip,   file="trip_full_updated2.csv")
#write_csv(status, file="status_full_updated2.csv")
```

myStationCode 1-7 are in Mountain view, 8-12 are in Palo Alto, 13-19 are in Redwood City, 20-54 are in San Francisco and 55-70 are in San Jose.

We can now draw a heatmap

```
trip_numbers <- trip  %>%
  group_by(StartStationCode, EndStationCode)  %>%
  summarise(trips = n())

ggplot(trip_numbers, aes(x=StartStationCode, y =EndStationCode))+
  geom_tile(aes(fill = trips))+
  scale_fill_gradient(low = "yellow",high = "red")
```