sample-analysis

Julian DeGroot-Lutzner & Adi Salwan

12/13/2017

f

```
library(readr)
randomsample <- read_csv("~/Documents/math154/ma154-project24-teambike/final_project/randomsample.csv")
```

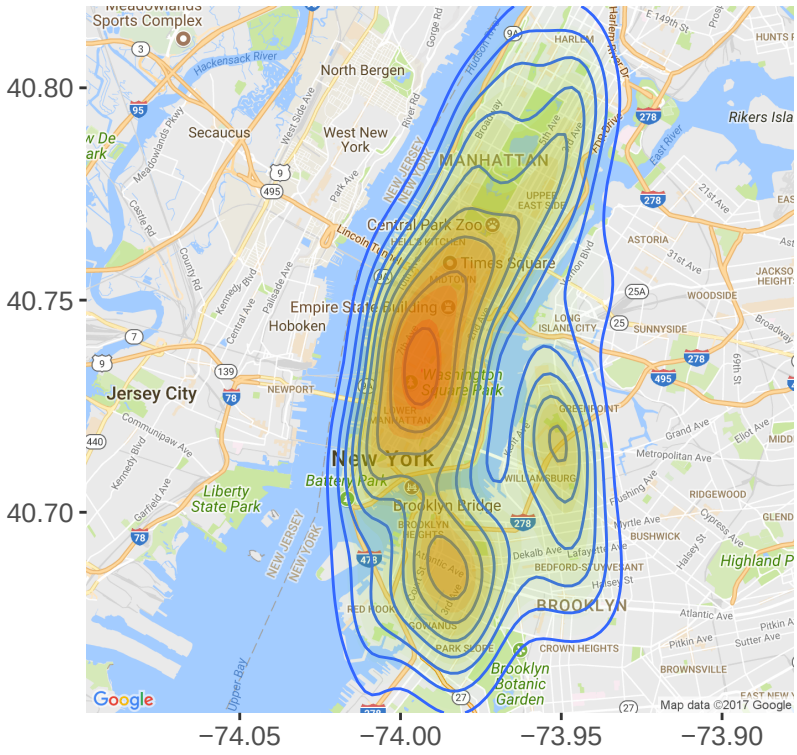
```
start_sums <- randomsample %>%
  group_by(start.station.id) %>%
  summarize(start.station.longitude = mean(start.station.longitude),
            start.station.latitude = mean(start.station.latitude),
            total.time.out = sum(tripduration),
            start.count = n()) %>%
  mutate(avg.time.out = total.time.out/start.count) %>%
  select(-total.time.out) %>%
  ungroup()
```

The Map Used

```
center.citibikes <- c(
  lon = mean(randomsample$start.station.longitude),
  lat = mean(randomsample$start.station.latitude))
mymap <- get_map(location = center.citibikes,
                 maptype = "roadmap",
                 zoom = 12)
```

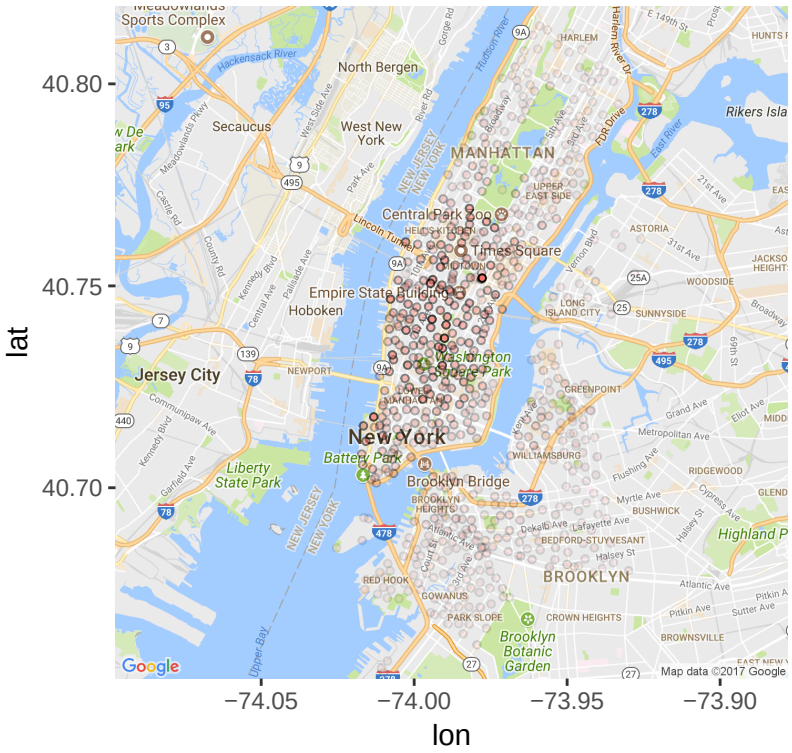
Location of Stations Map

```
 ggmap(mymap, extent = "panel", maprange=FALSE) +
  geom_density2d(data = start_sums,
                           aes(x = start.station.longitude,
                               y = start.station.latitude)) +
  stat_density2d(data = start_sums,
                           aes(x = start.station.longitude,
                               y = start.station.latitude, fill = ..level.., alpha = ..level..),
                 size = 0.01, bins = 16, geom = 'polygon') +
  scale_fill_gradient(low = "green", high = "red") +
  scale_alpha(range = c(0.00, 0.25), guide = FALSE) +
  theme(legend.position = "none", axis.title = element_blank(), text = element_text(size = 12))
```



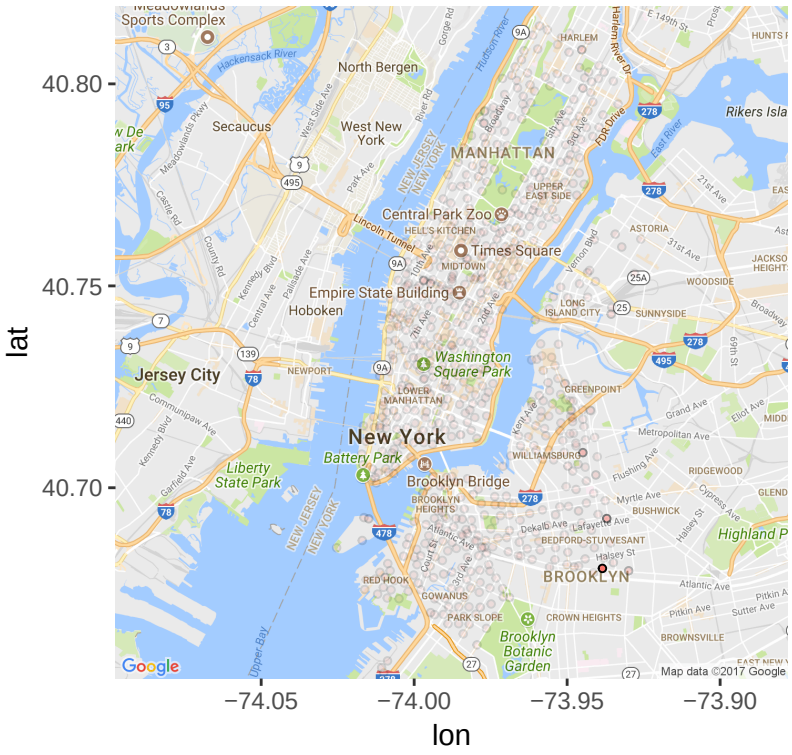
Count visualization Map

```
ggmap(mymap) +  geom_point(data = start_sums,
                           aes(x = start.station.longitude,
                               y = start.station.latitude,
                               fill = "red", alpha = start.count),
                           size = 1, shape = 21) +
  guides(fill=FALSE, alpha=FALSE, size=FALSE)
```



Avg. Trip Time Map

```
ggmap(mymap) +  geom_point(data = start_sums,
                           aes(x = start.station.longitude,
                               y = start.station.latitude,
                               fill = "red", alpha = avg.time.out),
                           size = 1, shape = 21) +
  guides(fill=FALSE, alpha=FALSE, size=FALSE)
```




```
start_sums %>%
  arrange(desc(start.count)) %>%
  select(-start.station.latitude, -start.station.longitude) %>%
  head()
```

```
## # A tibble: 6 x 3
##   start.station.id start.count avg.time.out
##              <int>       <int>        <dbl>
## 1              519        5367        838.2
## 2              497        4057        749.4
## 3              435        3948        658.4
## 4              426        3691       1221.8
## 5              402        3482        743.2
## 6              293        3461        720.4
```

```
start_sums %>%
  arrange(desc(avg.time.out)) %>%
  select(-start.station.latitude, -start.station.longitude) %>%
  head()
```

```
## # A tibble: 6 x 3
##   start.station.id start.count avg.time.out
##              <int>       <int>        <dbl>
## 1             3044          29        61691
## 2             3058          91        28080
## 3             3076         125        22146
## 4             3042         118        12161
## 5             3518           7        11179
## 6             3342          26         7160
```

```
end_sums <- randomsample %>%
  group_by(end.station.id) %>%
```

```
  summarize(end.station.longitude = mean(end.station.longitude),
            end.station.latitude = mean(end.station.latitude),
            total.time.in = sum(tripduration),
            end.count = n()) %>%
  mutate(avg.time.in = total.time.in/end.count) %>%
  select(-total.time.in) %>%
  ungroup()
```

```
colnames(start_sums)[1]<- c("id")
colnames(end_sums)[1] <- c("id")
joined_data<-inner_join(start_sums,end_sums,by="id")
joined_data <- joined_data %>%
mutate(difference = start.count - end.count,
       station.latitude =
         (start.station.latitude+ end.station.latitude)/2,
       station.longitude =
         (start.station.longitude + end.station.longitude)/2,
       total.ride.count = start.count + end.count) %>%
  mutate(normalized.difference =
           difference/(start.count+end.count),
         positive.difference = difference>0) %>%
  select(-start.station.latitude,
         -start.station.longitude,
         -end.station.longitude,
         -end.station.latitude)
```

```
ggplot(data = start_sums, aes(x=total.ride.count)) +
  geom_histogram()
```

```
top_ten_perc <- joined_data$total.ride.count %>%
  quantile(0.90)
```

```
biggest_differences <- joined_data %>%
  filter(total.ride.count >= top_ten_perc) %>%
  arrange(desc(normalized.difference)) %>%
head(10)
biggest_differences %>%
    select(id, start.count, end.count,
           difference, normalized.difference)
```

```
## # A tibble: 10 x 5
##       id start.count end.count difference normalized.difference
##    <int>       <int>     <int>      <int>                 <dbl>
##  1   521        2649      2380        269               0.05349
##  2   519        5367      4835        532               0.05215
##  3   281        2493      2279        214               0.04484
##  4   490        3264      3064        200               0.03161
##  5   479        2026      1904        122               0.03104
##  6   457        2287      2160        127               0.02856
##  7   523        2631      2496        135               0.02633
##  8  2006        2706      2574        132               0.02500
##  9   380        2199      2105         94               0.02184
## 10   528        1889      1811         78               0.02108
```

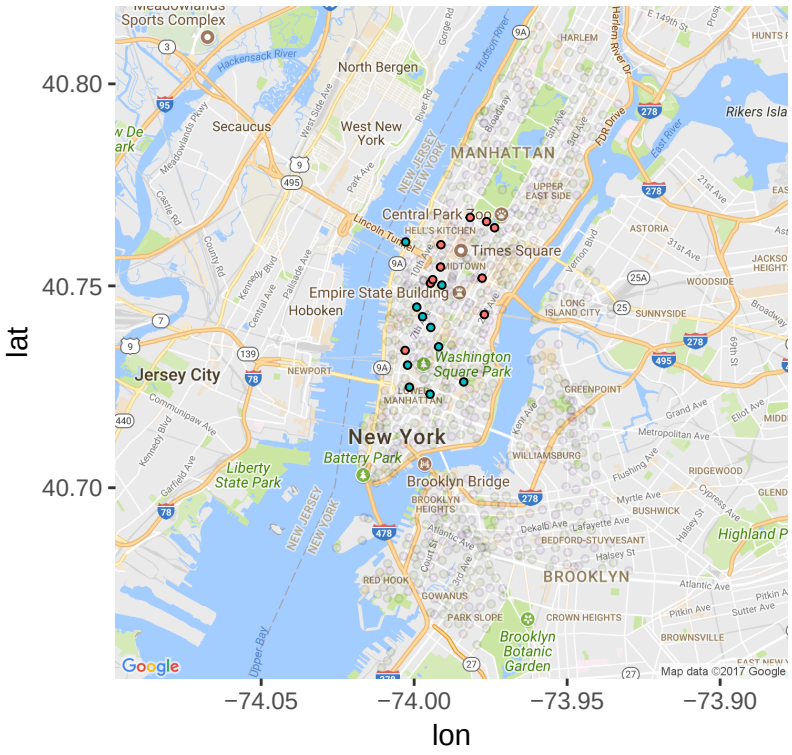
```
smallest_differences <- joined_data %>%
```

```
  filter(total.ride.count >= top_ten_perc) %>%
  arrange(normalized.difference) %>%
head(10)
smallest_differences %>%
  select(id, start.count, end.count,
         difference, normalized.difference)
```

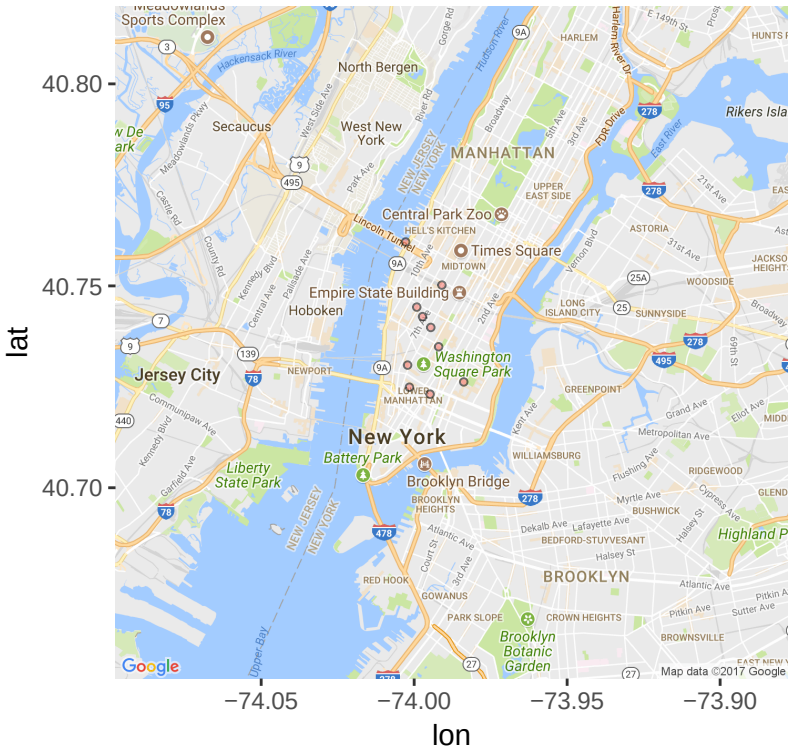
```
## # A tibble: 10 x 5
##       id start.count end.count difference normalized.difference
##    <int>       <int>     <int>      <int>                  <dbl>
##  1   432        2114      2330       -216                -0.04860
##  2   492        2384      2602       -218                -0.04372
##  3   382        2728      2966       -238                -0.04180
##  4   514        2632      2849       -217                -0.03959
##  5   348        1768      1910       -142                -0.03861
##  6   453        1847      1971       -124                -0.03248
##  7   251        1973      2105       -132                -0.03237
##  8   168        2558      2724       -166                -0.03143
##  9   368        2921      3110       -189                -0.03134
## 10   334        1886      1995       -109                -0.02809
```

Difference visualization Map

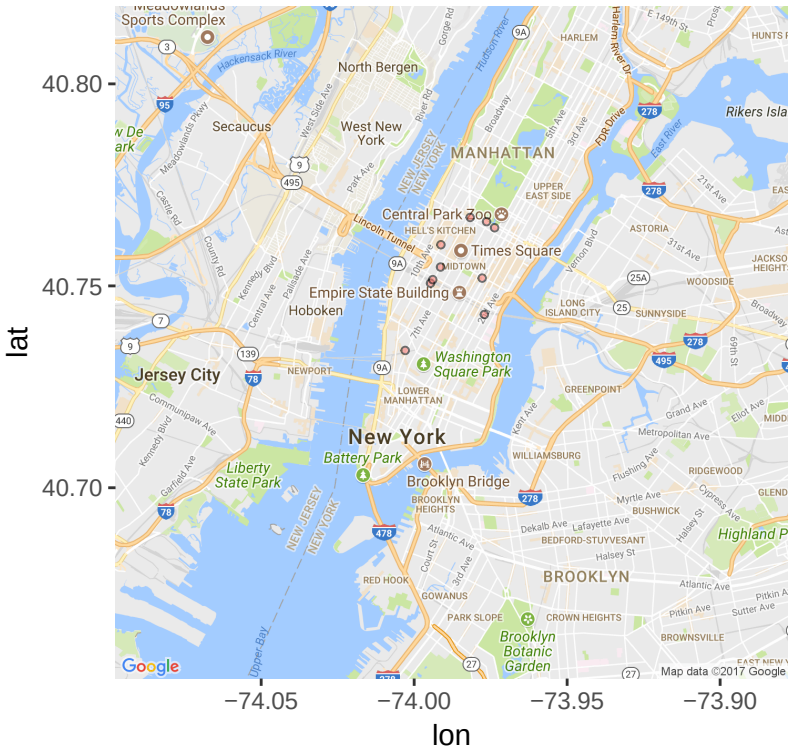
```
ggmap(mymap) +  geom_point(data = joined_data,
                           aes(x = station.longitude,
                               y = station.latitude,
                               fill = positive.difference,
                               alpha = 0.7),
                           size = 1, shape = 21) +
  guides(fill=FALSE, alpha=FALSE, size=FALSE) +
  geom_point(data=biggest_differences,
             aes(x = station.longitude,
                 y = station.latitude,
                 fill = "blue",
                 alpha = 1.0),
             size = 1, shape = 21) +
    geom_point(data=smallest_differences,
             aes(x = station.longitude,
                 y = station.latitude,
                 fill = "red",
                 alpha = 1.0),
             size = 1, shape = 21)
```




```
ggmap(mymap) +
    geom_point(data=smallest_differences,
               aes(x = station.longitude,
                   y = station.latitude,
                   fill = "blue",
                   alpha = 1.0),
               size = 1, shape = 21) +
  guides(fill=FALSE, alpha=FALSE, size=FALSE)
```




```
ggmap(mymap) +
  geom_point(data=biggest_differences,
             aes(x = station.longitude,
                 y = station.latitude,
                 fill = "red",
                 alpha = 1.0),
             size = 1, shape = 21) +
  guides(fill=FALSE, alpha=FALSE, size=FALSE)
```




```
head(joined_data)
```

```
## # A tibble: 6 x 11
##      id start.count avg.time.out end.count avg.time.in difference
##   <int>       <int>        <dbl>     <int>       <dbl>      <int>
## 1    72        1419       1069.6      1383      1071.5         36
## 2    79        1093        874.8      1145       995.0        -52
## 3    82         471       1078.0       441       889.1         30
## 4    83         567       1271.3       589      1565.7        -22
## 5   116        2124        712.5      2153       656.6        -29
## 6   119          98        727.5        98       741.8          0
## # ... with 5 more variables: station.latitude <dbl>,
## #   station.longitude <dbl>, total.ride.count <int>,
## #   normalized.difference <dbl>, positive.difference <lgl>
```

```
stations_of_interest <- joined_data %>%
  filter(id == 521 | id == 432)
stations_of_interest
```

```
## # A tibble: 2 x 11
##      id start.count avg.time.out end.count avg.time.in difference
##   <int>       <int>        <dbl>     <int>       <dbl>      <int>
## 1   432        2114        777.9      2330       849.2       -216
## 2   521        2649        800.2      2380       764.2        269
## # ... with 5 more variables: station.latitude <dbl>,
## #   station.longitude <dbl>, total.ride.count <int>,
## #   normalized.difference <dbl>, positive.difference <lgl>
```

```
ggmap(mymap) +
  geom_point(data=stations_of_interest,
             aes(x = station.longitude,
```

```
              y = station.latitude,
              fill = positive.difference,
              alpha = 1.0),
          size = 4, shape = 21) +
guides(fill=FALSE, alpha=FALSE, size=FALSE)
```