Mapping the Boroughs

Landon C

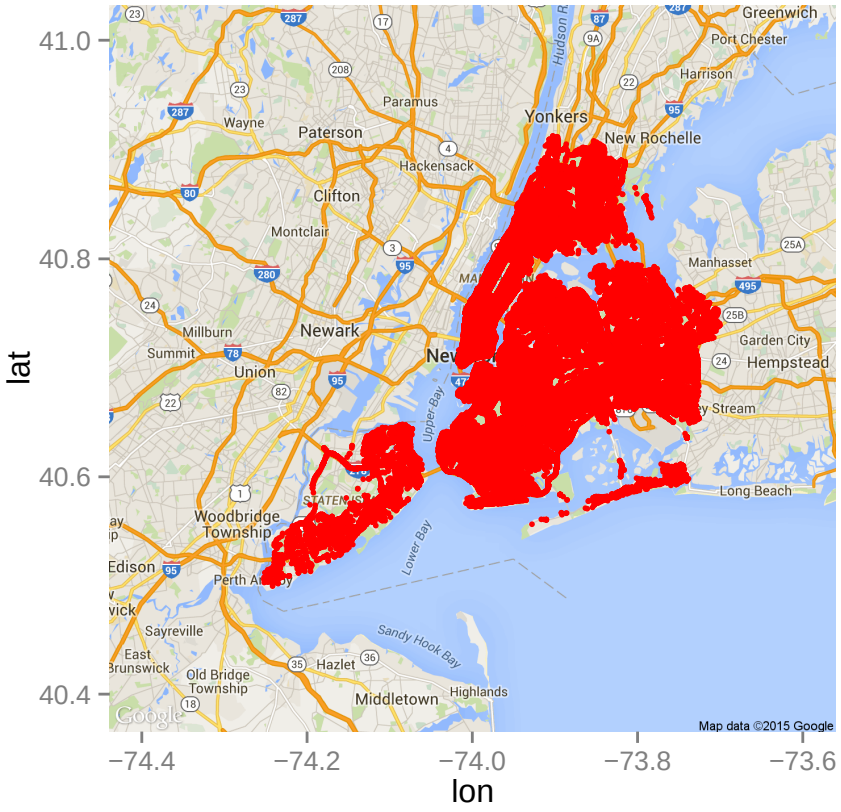
August 9, 2015

```
library(ggmap)
library(lubridate)
library(plyr)
library(geosphere) #calculating distances over lat. and long.

#Read in data and make the map
accident = read.csv("accident2015.csv", header=TRUE)
#slice up data by borough
accident_Bronx = subset(accident, BOROUGH == "BRONX")
accident_Brooklyn = subset(accident, BOROUGH == "BROOKLYN")
accident_Manhattan = subset(accident, BOROUGH == "MANHATTAN")
accident_Queens = subset(accident, BOROUGH == "QUEENS")
accident_StatenIsland = subset(accident, BOROUGH == "STATEN ISLAND")
accident_noborough = subset(accident, BOROUGH == "")

al1 = get_map(location = c(lon = -74., lat = 40.7), zoom = 10, maptype = 'roadmap')
```

```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=40.7,-74&zoom=10&size=640x640&sca
```

```
al1MAP = ggmap(al1)

#All crashes
al1MAP + geom_point(data = accident, aes(x = LONGITUDE, y = LATITUDE), colour = "red", size = 1)
```

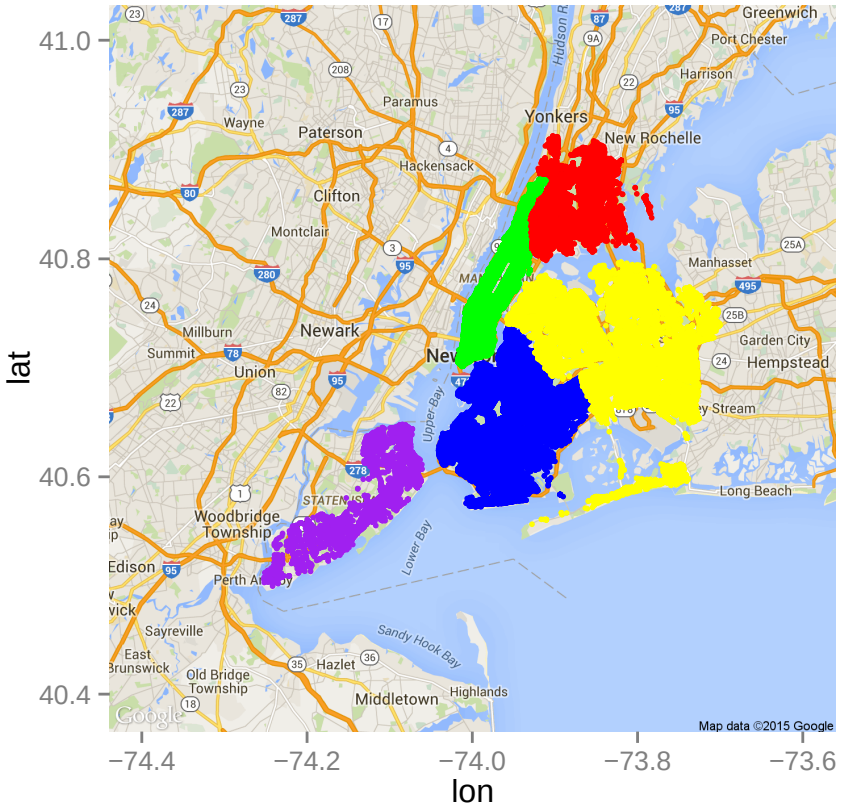
```
## Warning: Removed 20097 rows containing missing values (geom_point).
```



The map above shows location of all crashes in NYC.

```
al1MAP + geom_point(data = accident_Bronx, aes(x = LONGITUDE, y = LATITUDE), colour = "red", size = 1) +
geom_point(data = accident_Brooklyn, aes(x = LONGITUDE, y = LATITUDE), colour = "blue", size = 1) +
geom_point(data = accident_Manhattan, aes(x = LONGITUDE, y = LATITUDE), colour = "green", size = 1) +
geom_point(data = accident_Queens, aes(x = LONGITUDE, y = LATITUDE), colour = "yellow", size = 1) +
geom_point(data = accident_StatenIsland, aes(x = LONGITUDE, y = LATITUDE), colour = "purple", size = 1)
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

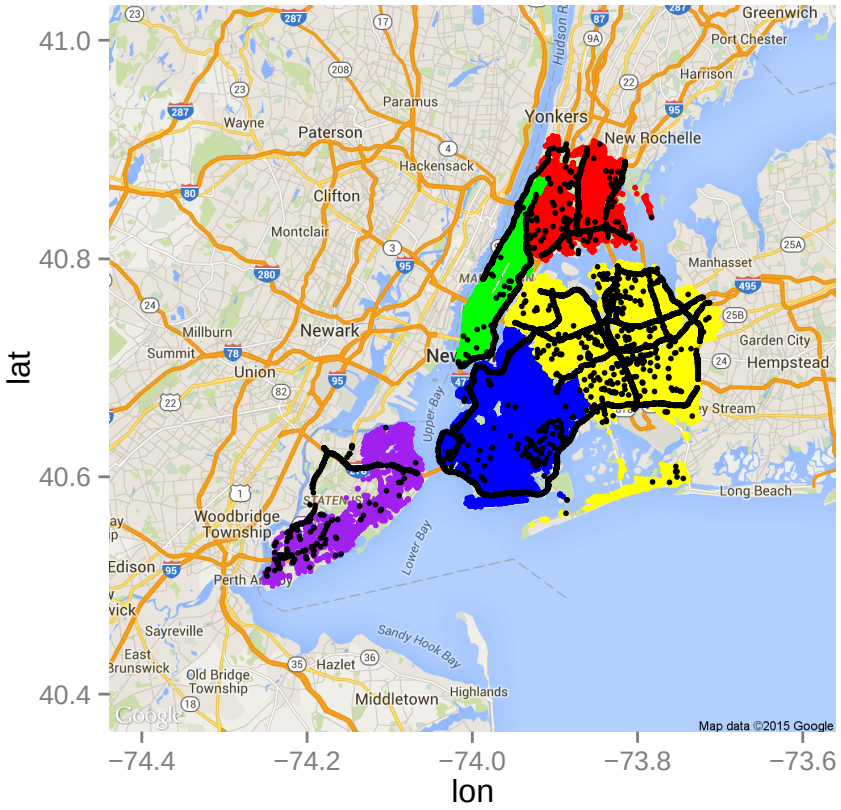


The map above shows all crashes in NYC by borough, where Bronx = red, Brooklyn = blue, Manhattan = green, Queens = yellow, Staten Island = purple.

```
#Crashes with no borough in black
al1MAP + geom_point(data = accident_Bronx, aes(x = LONGITUDE, y = LATITUDE), colour = "red", size = 1) +
  geom_point(data = accident_Brooklyn, aes(x = LONGITUDE, y = LATITUDE), colour = "blue", size = 1) +
  geom_point(data = accident_Manhattan, aes(x = LONGITUDE, y = LATITUDE), colour = "green", size = 1) +
  geom_point(data = accident_Queens, aes(x = LONGITUDE, y = LATITUDE), colour = "yellow", size = 1) +
  geom_point(data = accident_StatenIsland, aes(x = LONGITUDE, y = LATITUDE), colour = "purple", size =
  geom_point(data = accident_noborough, aes(x = LONGITUDE, y = LATITUDE), colour = "black", size = 1)
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

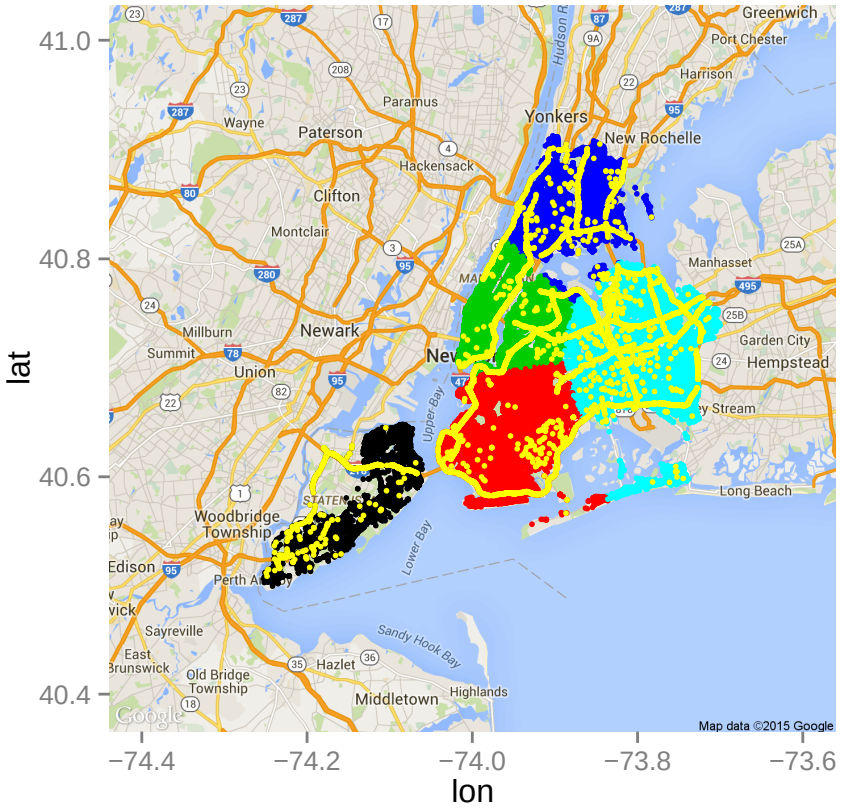
```
## Warning: Removed 20096 rows containing missing values (geom_point).
```



The map above shows the points with no borough specified in black.

```
####Kmeans clustering
lat_lon = accident[,5:6]
lat_lon = lat_lon[is.finite(lat_lon$LATITUDE) & is.finite(lat_lon$LONGITUDE), ] # take out nan values

# Run k-means with 5 clusters and 50 re-starts
clust5 = kmeans(lat_lon, 5, nstart=50)
al1MAP + geom_point(data = lat_lon, aes(x = LONGITUDE, y = LATITUDE), colour = factor(clust5$cluster), s
  geom_point(data = accident_noborough, aes(x = LONGITUDE, y = LATITUDE), colour = "yellow", size = 1)
```

```
## Warning: Removed 20096 rows containing missing values (geom_point).
```



Kmeans does pretty good, especially for Staten Island and Brooklyn. We will insert the boroughs in for the data points with an empty borough field but only for the Staten Island and Brooklyn clusters. Then, we will make custom cuts on top of the clusters to improve the clustering, basically, fine tuning the clustering process.

```
#update accident with cluster number
#first remove the same points as removed for kmeans
accident = accident[is.finite(accident$LATITUDE) & is.finite(accident$LONGITUDE), ] # take out nan valu
accident$CLUSTER = clust5$cluster

count(accident, c("BOROUGH","CLUSTER"))
```

```
##          BOROUGH CLUSTER  freq
## 1                      1   652
## 2                      2  1687
## 3                      3  2324
## 4                      4  1541
## 5                      5  4616
## 6          BRONX       4 12004
## 7       BROOKLYN       2 24549
## 8       BROOKLYN       3  3331
## 9       BROOKLYN       5   752
## 10     MANHATTAN       3 20625
## 11     MANHATTAN       4  3216
## 12        QUEENS       2   509
## 13        QUEENS       3  5311
## 14        QUEENS       4   536
## 15        QUEENS       5 17596
```

```
## 16 STATEN ISLAND       1  3620
```

```
#Bronx = 5, Brooklyn = 4 Manhattan = 2, Queens = 3, Staten Island = 1

#update the obvious blank boroughs
accident$BOROUGH[accident$BOROUGH == "" & accident$CLUSTER == 1] = "STATEN ISLAND"
accident$BOROUGH[accident$BOROUGH == "" & accident$CLUSTER == 4] = "BROOKLYN"
count(accident, c("BOROUGH","CLUSTER"))
```

```
##           BOROUGH CLUSTER  freq
## 1                       2  1687
## 2                       3  2324
## 3                       5  4616
## 4           BRONX       4 12004
## 5        BROOKLYN       2 24549
## 6        BROOKLYN       3  3331
## 7        BROOKLYN       4  1541
## 8        BROOKLYN       5   752
## 9       MANHATTAN       3 20625
## 10      MANHATTAN       4  3216
## 11         QUEENS       2   509
## 12         QUEENS       3  5311
## 13         QUEENS       4   536
## 14         QUEENS       5 17596
## 15 STATEN ISLAND       1  4272
```

```
# make custom cuts to improve on kmeans and update blank boroughs
#Manhattan(2) -> Queens(3) cut lon(-73.93 to -73.8) lat(40.7 to 40.75)
accident$BOROUGH[accident$BOROUGH == "" & accident$CLUSTER == 2 &
                   accident$LONGITUDE >= -73.93 & accident$LONGITUDE <= -73.8 &
                   accident$LATITUDE >= 40.7 &
                   accident$LATITUDE <= 40.75] = "QUEENS"
#now put the rest of 3 cluster into queens
accident$BOROUGH[accident$BOROUGH == "" & accident$CLUSTER == 3] = "QUEENS"
count(accident, c("BOROUGH","CLUSTER"))
```

```
##           BOROUGH CLUSTER  freq
## 1                       2  1687
## 2                       5  4616
## 3           BRONX       4 12004
## 4        BROOKLYN       2 24549
## 5        BROOKLYN       3  3331
## 6        BROOKLYN       4  1541
## 7        BROOKLYN       5   752
## 8       MANHATTAN       3 20625
## 9       MANHATTAN       4  3216
## 10         QUEENS       2   509
## 11         QUEENS       3  7635
## 12         QUEENS       4   536
## 13         QUEENS       5 17596
## 14 STATEN ISLAND       1  4272
```