Mapping the Boroughs

Landon C

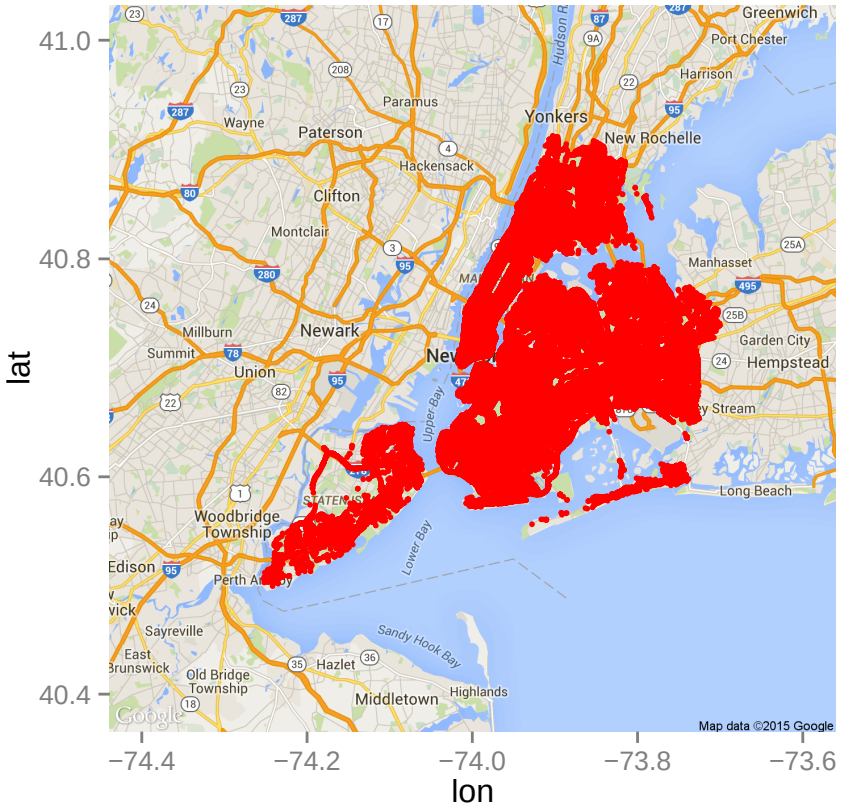
August 9, 2015

```
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)
```

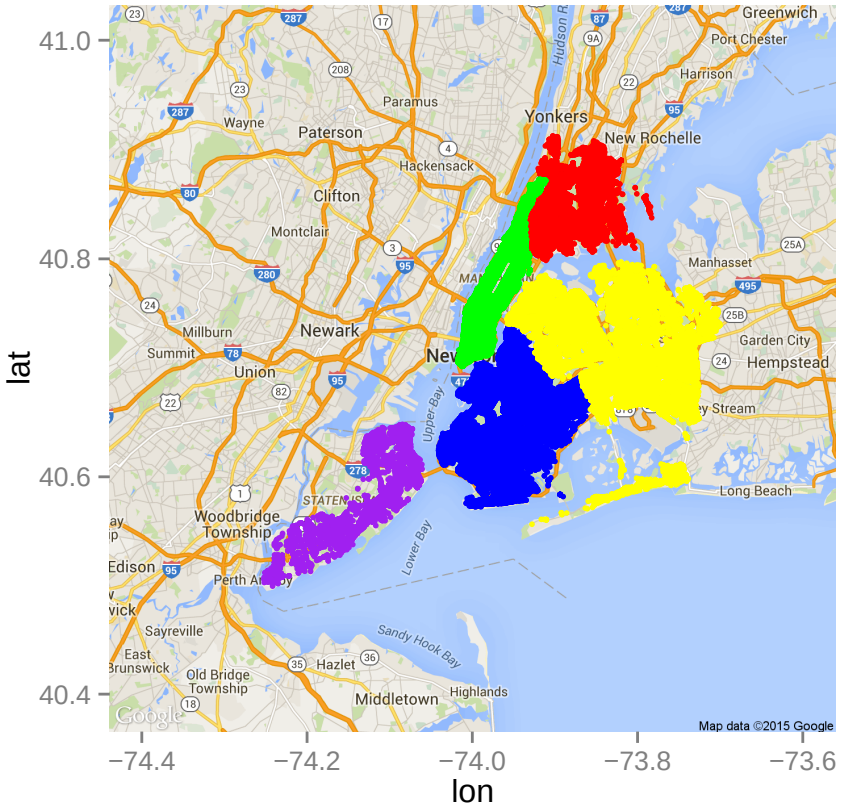
```
## Warning: Removed 20097 rows containing missing values (geom_point).
```



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).
```

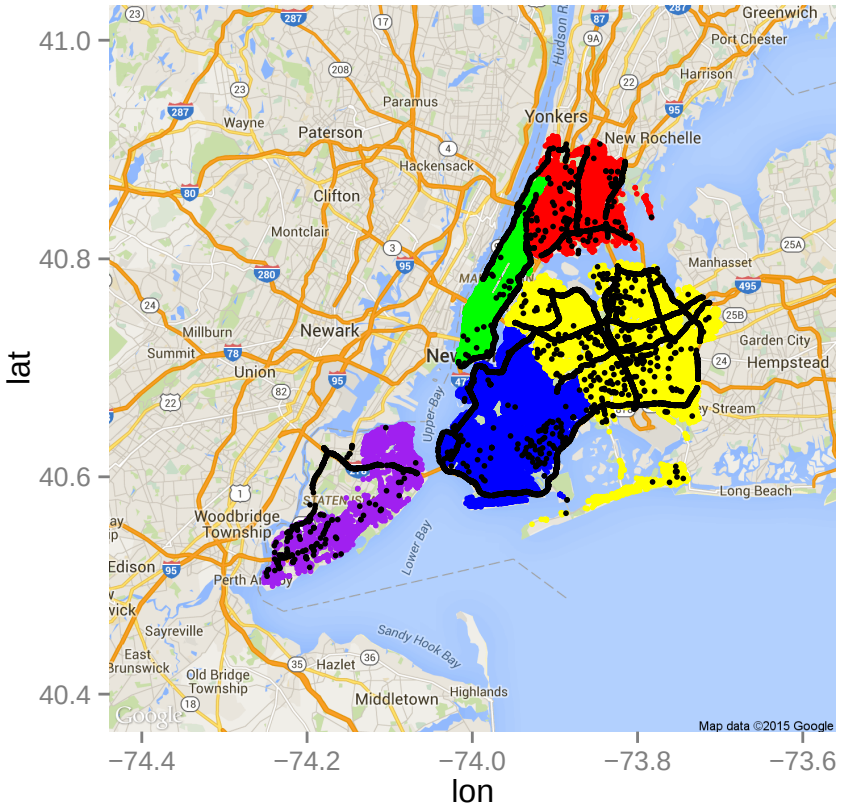


The map above shows all crashes in NYC by borough, where Bronx = red, Brooklyn = blue, Manhattan = green, Queens = yellow, Staten Island = purple.

```
#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).
```

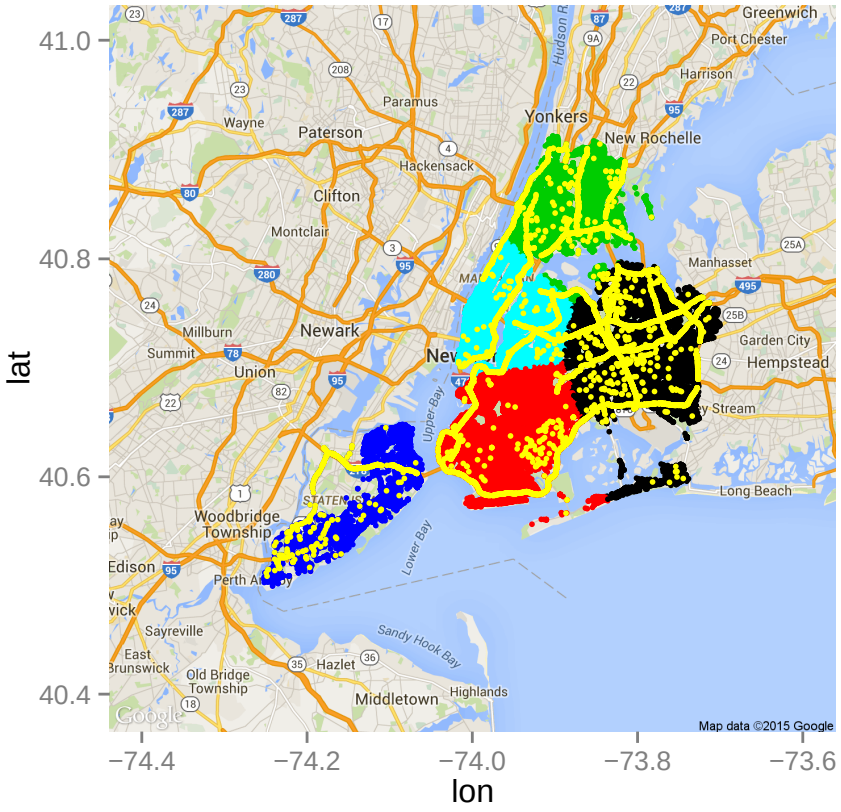
```
## Warning: Removed 20096 rows containing missing values (geom_point).
```



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
set.seed(1) #super important to get same clusters every time
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).
```




```
#gglocator()
#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
```

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.

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

```
##           BOROUGH CLUSTER  freq
## 1                       1  4616
## 2                       2  1687
## 3                       3  1541
## 4                       4   652
## 5                       5  2324
## 6           BRONX       3 12004
## 7        BROOKLYN       1   754
## 8        BROOKLYN       2 24547
## 9        BROOKLYN       5  3331
## 10      MANHATTAN       3  3216
## 11      MANHATTAN       5 20625
## 12         QUEENS       1 17596
```

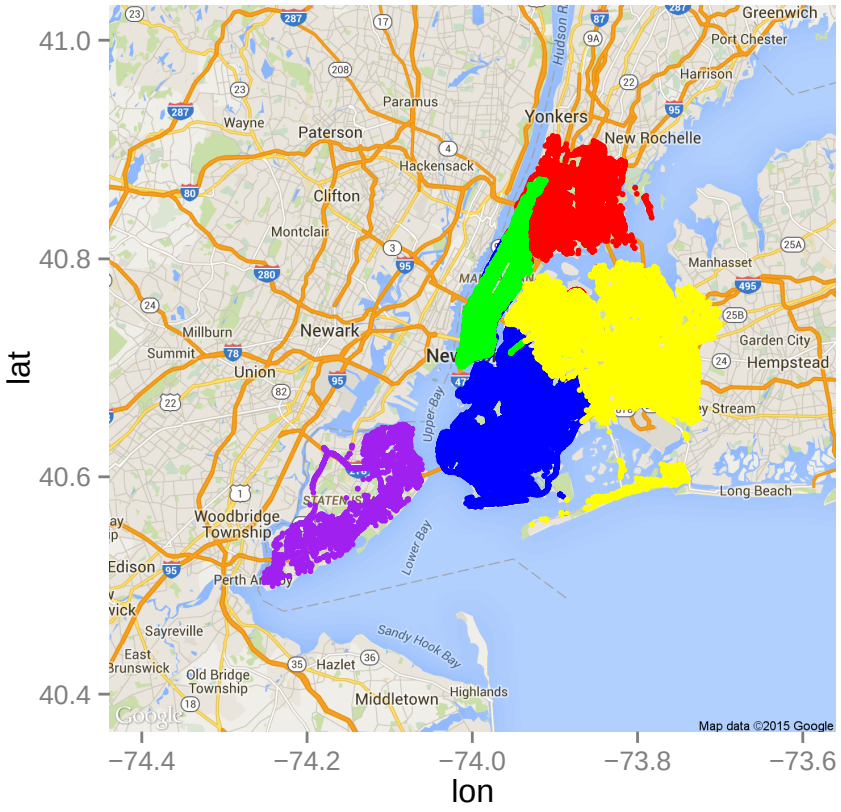
```
## 13        QUEENS       2   509
## 14        QUEENS       3   536
## 15        QUEENS       5  5311
## 16 STATEN ISLAND       4  3620
```

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


###Slice up chunks of clusters that are in the wrong borough, and assign these chunks to the correct bo
#Manhattan(5) -> Brooklyn(2) cut lon(-74.002, -73.95366) lat(40.69084, 50.70843)
accident$BOROUGH[accident$BOROUGH == "" & accident$CLUSTER == 5 &
                   accident$LONGITUDE >= -74.002 & accident$LONGITUDE <= -73.95366 &
                   accident$LATITUDE >= 40.69084 &
                   accident$LATITUDE <= 50.70843] = "BROOKLYN"

#Manhattan(5) -> Queens(1) cut lon(-73.93621,-73.89635) lat(40.70533, 40.71567 )
accident$BOROUGH[accident$BOROUGH == "" & accident$CLUSTER == 5 &
                   accident$LONGITUDE >= -73.93621 & accident$LONGITUDE <= -73.89635 &
                   accident$LATITUDE >= 40.70533 &
                   accident$LATITUDE <= 40.71567] = "QUEENS"

#Manhattan(5) -> Queens(1) cut lon(-73.93592 to -73.86769) lat(40.72395 to 40.76843)
accident$BOROUGH[accident$BOROUGH == "" & accident$CLUSTER == 5 &
                   accident$LONGITUDE >= -73.93592 & accident$LONGITUDE <= -73.86769 &
                   accident$LATITUDE >= 40.72395 &
                   accident$LATITUDE <= 40.76843] = "QUEENS"

#Bronx(3) -> Manhattan(5) cut lon(-73.96321,-73.93043) lat(40.8274 , 40.84499)
accident$BOROUGH[accident$BOROUGH == "" & accident$CLUSTER == 3 &
                   accident$LONGITUDE >= -73.96321 & accident$LONGITUDE <= -73.93043 &
                   accident$LATITUDE >= 40.8274 &
                   accident$LATITUDE <= 40.84499] = "MANHATTAN"

#Bronx(3) -> Queens(1) cut lon(-73.87588,-73.85268) lat( 40.77257,40.77981)
accident$BOROUGH[accident$BOROUGH == "" & accident$CLUSTER == 3 &
                   accident$LONGITUDE >= -73.87588 & accident$LONGITUDE <= -73.85268 &
                   accident$LATITUDE >= 40.77257 &
                   accident$LATITUDE <= 40.77981] = "QUEENS"

#Bronx(3) -> Queens(1) cut lon(-73.84313,-73.82676) lat( 40.79222,40.80257)
accident$BOROUGH[accident$BOROUGH == "" & accident$CLUSTER == 3 &
                   accident$LONGITUDE >= -73.84313 & accident$LONGITUDE <= -73.82676 &
                   accident$LATITUDE >= 40.79222 &
                   accident$LATITUDE <= 40.80257] = "QUEENS"

#Brooklyn() -> Queens(1) cut lon(-73.88543,-73.8213) lat(40.53567,40.57912)
accident$BOROUGH[accident$BOROUGH == "" & accident$CLUSTER == 2 &
                   accident$LONGITUDE >= -73.88543 & accident$LONGITUDE <= -73.8213 &
                   accident$LATITUDE >= 40.53567 &
                   accident$LATITUDE <= 40.57912] = "QUEENS"

#now put the remaining pieces of clusters in their correct borough
accident$BOROUGH[accident$BOROUGH == "" & accident$CLUSTER == 1] = "QUEENS"
accident$BOROUGH[accident$BOROUGH == "" & accident$CLUSTER == 2] = "BROOKLYN"
```

```
accident$BOROUGH[accident$BOROUGH == "" & accident$CLUSTER == 3] = "BRONX"
accident$BOROUGH[accident$BOROUGH == "" & accident$CLUSTER == 4] = "STATEN ISLAND"
accident$BOROUGH[accident$BOROUGH == "" & accident$CLUSTER == 5] = "MANHATTAN"
count(accident, c("BOROUGH","CLUSTER")) #all blank boroughs should be gone!
```

```
##           BOROUGH CLUSTER  freq
## 1           BRONX       3 13387
## 2        BROOKLYN       1   754
## 3        BROOKLYN       2 26232
## 4        BROOKLYN       5  4300
## 5       MANHATTAN       3  3374
## 6       MANHATTAN       5 21272
## 7          QUEENS       1 22212
## 8          QUEENS       2   511
## 9          QUEENS       3   536
## 10         QUEENS       5  6019
## 11 STATEN ISLAND       4  4272
```

```
#test it!
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 == "")

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)
```



Above is the map after assignming the missing borough data points to clusters using kmeans and seven different custom partitioning. We have picked all 10,000 points with no boroughs and assigned almost all of them to the correct borough using kmeans and custom partitioning!