Big data exercices

Albert Xavier Lopez Barrantes

18 de octubre de 2018

Big Data exercices

```
library(ggplot2)
library(tidyverse)
library(ggplot2)
library(RCurl)

data<-read.csv(text=getURL("https://raw.githubusercontent.com/fivethirtyeight/uber-tlc-foil-response/ma
```

Visualize where departures trips are located. The main goal of UBER is to know zones having more trips to increase the number of cars in those zones. NOTE: Investigate the function peel at bigvis package. Could this function help you in improving visualization?

First, having a first look to the dataset, we search in google the position of the first coordinates. As a s result, we notice the latitudes and longitudes are from New York, what is going to give us a context.

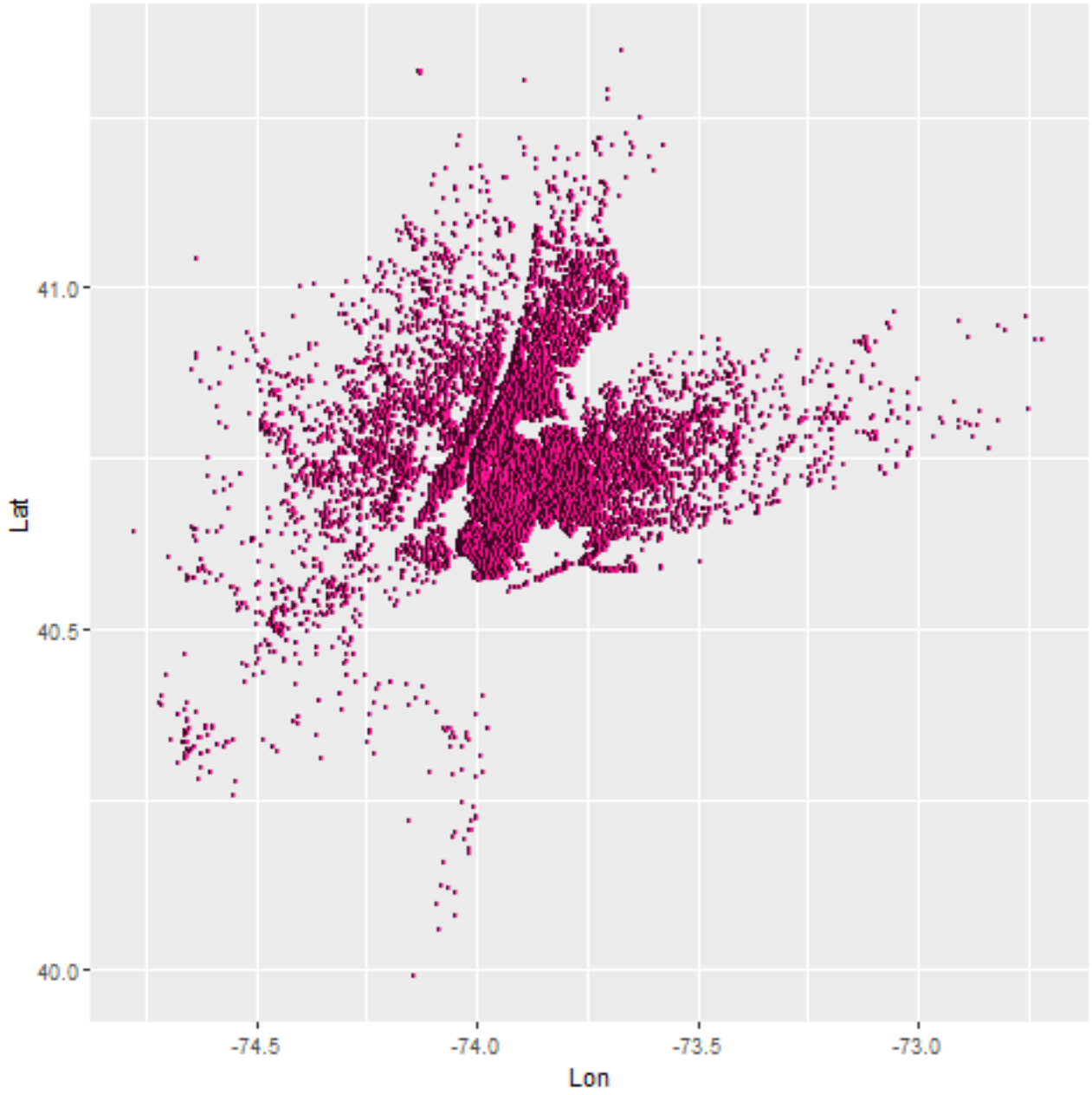
Once we are positioned we are going to visualize the data points. To do so, we are going to display a simple dotplot:

```
png("mapdata.png")
mapdata <- ggplot()+
  geom_point(data=data,
             aes(x=Lon, y=Lat),
             fill="Deep Pink",pch=21, size=.7, alpha=I(0.7))

print(mapdata)
dev.off()
```

```
## pdf
##   2
```

```
knitr::include_graphics("mapdata.png")
```



At first glance we can see huge concentration in the Manhattan Island and surroundings. If we wanna have a clear picture on the biggest concentration of trips in New York we will have to adjust some parameters. To do so, we will use the outputs from the function distribution provided by the function "summary"

```
summary(data$Lat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   39.99   40.72   40.74   40.74   40.76   41.35
```

```
summary(data$Lon)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -74.77  -74.00  -73.98  -73.97  -73.96  -72.72
```

Now that we know the distributions, we are going to adjust the parameters:

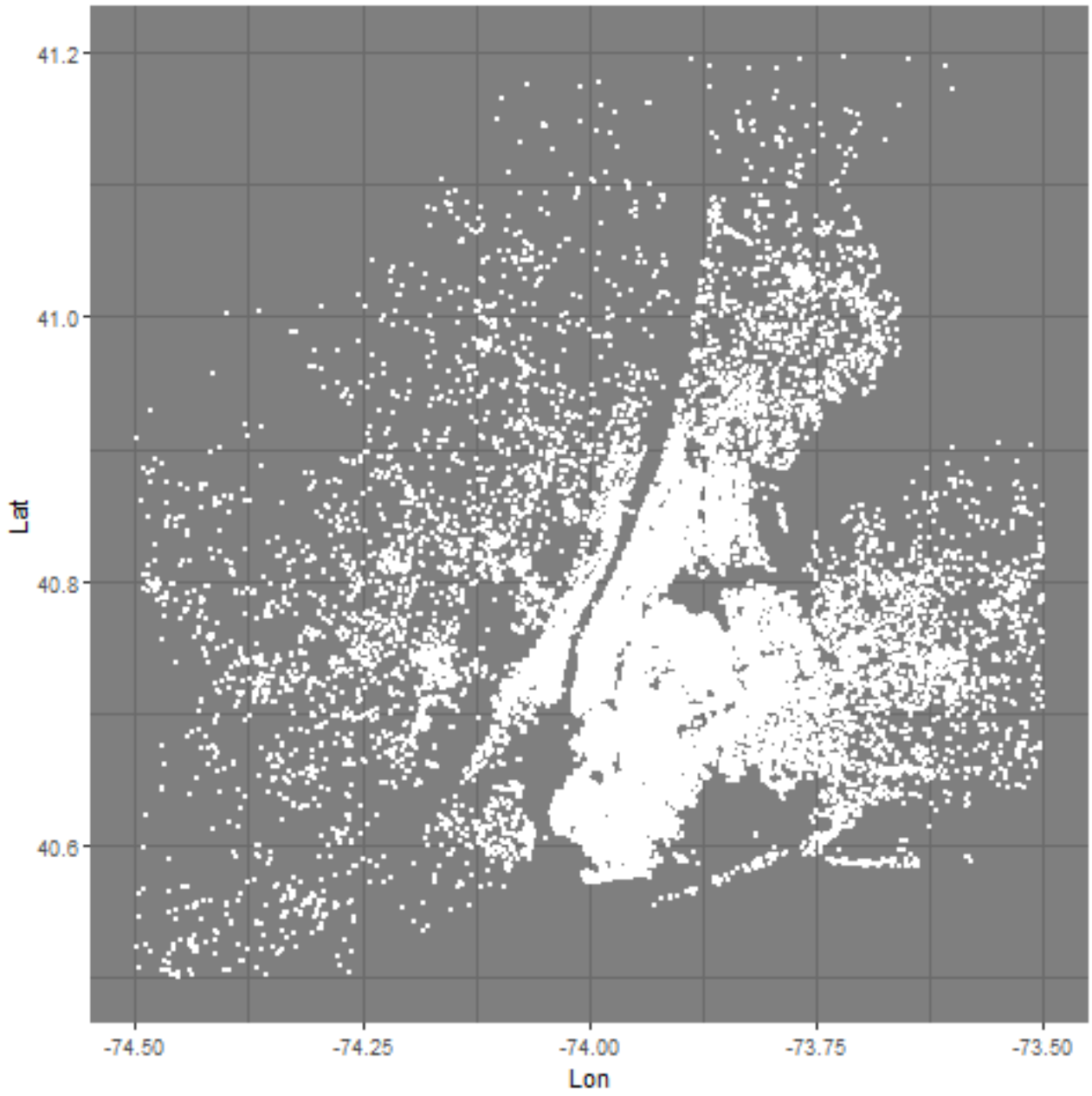
```
png("mapdata2.png")
mapdata2 <- ggplot(data=data, aes(x=Lon, y=Lat))+
  geom_point(color="white", size=0.03) +
            scale_x_continuous(limits=c(-74.5, -73.5)) +
            scale_y_continuous(limits=c(40.5, 41.2)) +
            theme_dark()
print(mapdata2)
```

```
## Warning: Removed 1273 rows containing missing values (geom_point).
```

```
dev.off()
```

```
## pdf
##   2
```

```
knitr::include_graphics("mapdata2.png")
```



And if we adjust the zoom a little bit more, we change the plot parameters, and we get:

```
png("mapdata3.png")
mapdata3 <- ggplot(data=data, aes(x=Lon, y=Lat))+
  geom_point(color="white", size=0.02, alpha=I(0.7)) +
             scale_x_continuous(limits=c(-74.05, -73.85)) +
             scale_y_continuous(limits=c(40.6, 40.9)) +
             theme_dark()

mapdata3
```

```
## Warning: Removed 57216 rows containing missing values (geom_point).
```

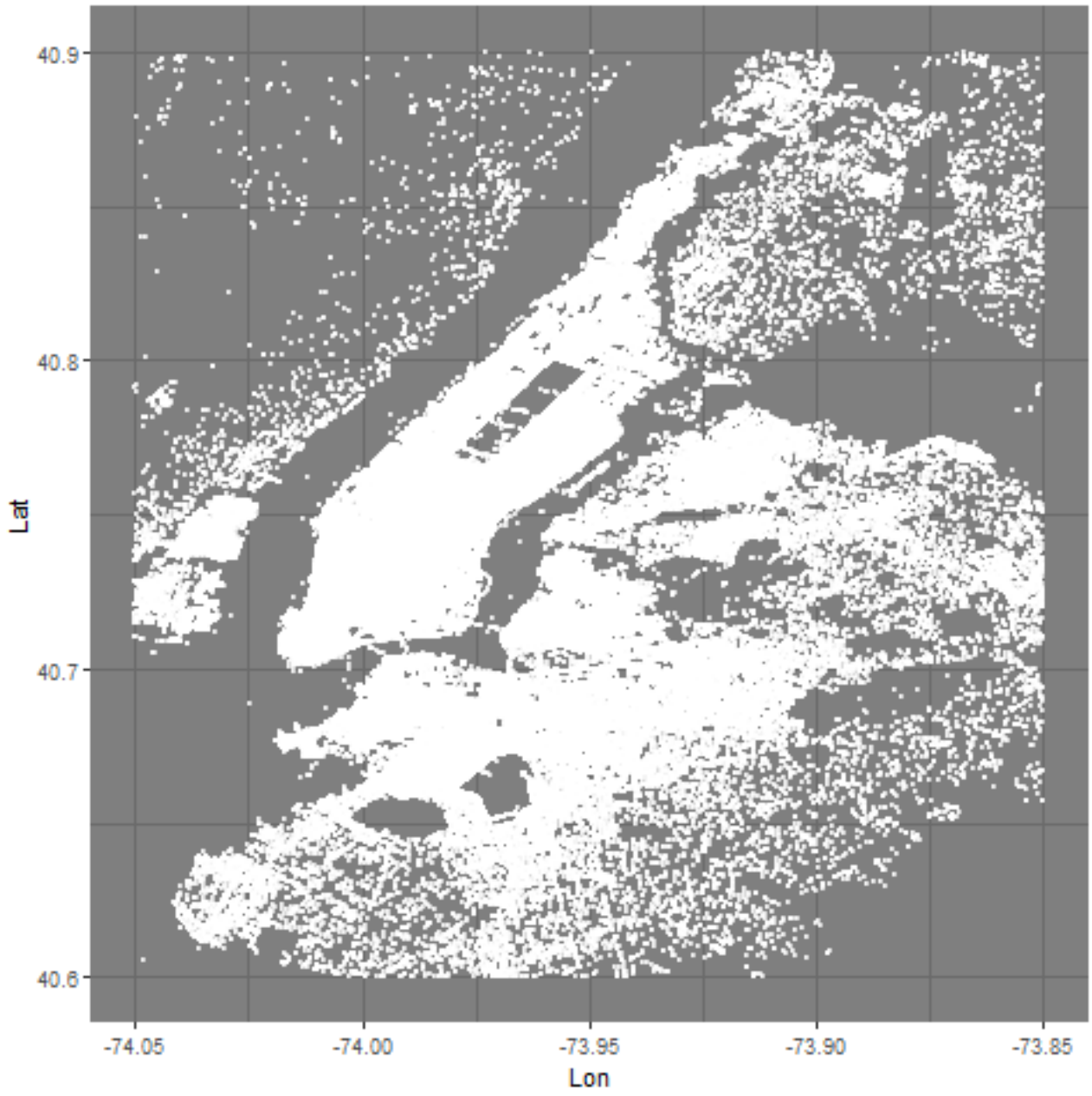
```
print(mapdata3)
```

```
## Warning: Removed 57216 rows containing missing values (geom_point).
```

```
dev.off()
```

```
## pdf
##   2
```

```
knitr::include_graphics("mapdata3.png")
```



My final recommendation for Uber would be to concentrate drivers in the island of Manhattan, Brooklyn, and the riverside of New Jersey.