Technical Appendix - Clustering

Yueni Wang

11/8/2021

```
#Load the required libraries
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readxl)
library(FactoMineR)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(dplyr)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(grid)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(ggplotify)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(sf)
```

```
## Linking to GEOS 3.8.1, GDAL 3.2.1, PROJ 7.2.1
```

```
library(ggmap)
```

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.
```

```
## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```
library(maps)
```

```
##
## Attaching package: 'maps'
```

```
## The following object is masked from 'package:purrr':
##
##     map
```

```
library(mapdata)

#Registering google API key:
register_google(key="AIzaSyAhU33Q8tMCDcB1wAyKQCOXBV3TKc2kTBY")
```

```
#The transformed data:
#bikeshare <- read.csv("bike2019_transformed_new.csv")
#The raw data:
bikeshare_origin <- read.csv("bike_capacity_lon_lat.csv")
label1 <- read.csv("label_1.csv")
```

```
#Original Clustering using locations
bikeshare_dataframe1 <- bikeshare_origin %>%
  select(lon,lat) %>%
  data.frame()

clusters <- hclust(dist(bikeshare_dataframe1))

bikeshare_data1 <- bikeshare_dataframe1 %>%
  mutate(cluster_station = factor(cutree(clusters, 5)))

sum(bikeshare_data1$cluster_station==1)
```

```
## [1] 178
```

```
sum(bikeshare_data1$cluster_station==2)
```

```
## [1] 331
```

```
sum(bikeshare_data1$cluster_station==3)
```

```
## [1] 68
```

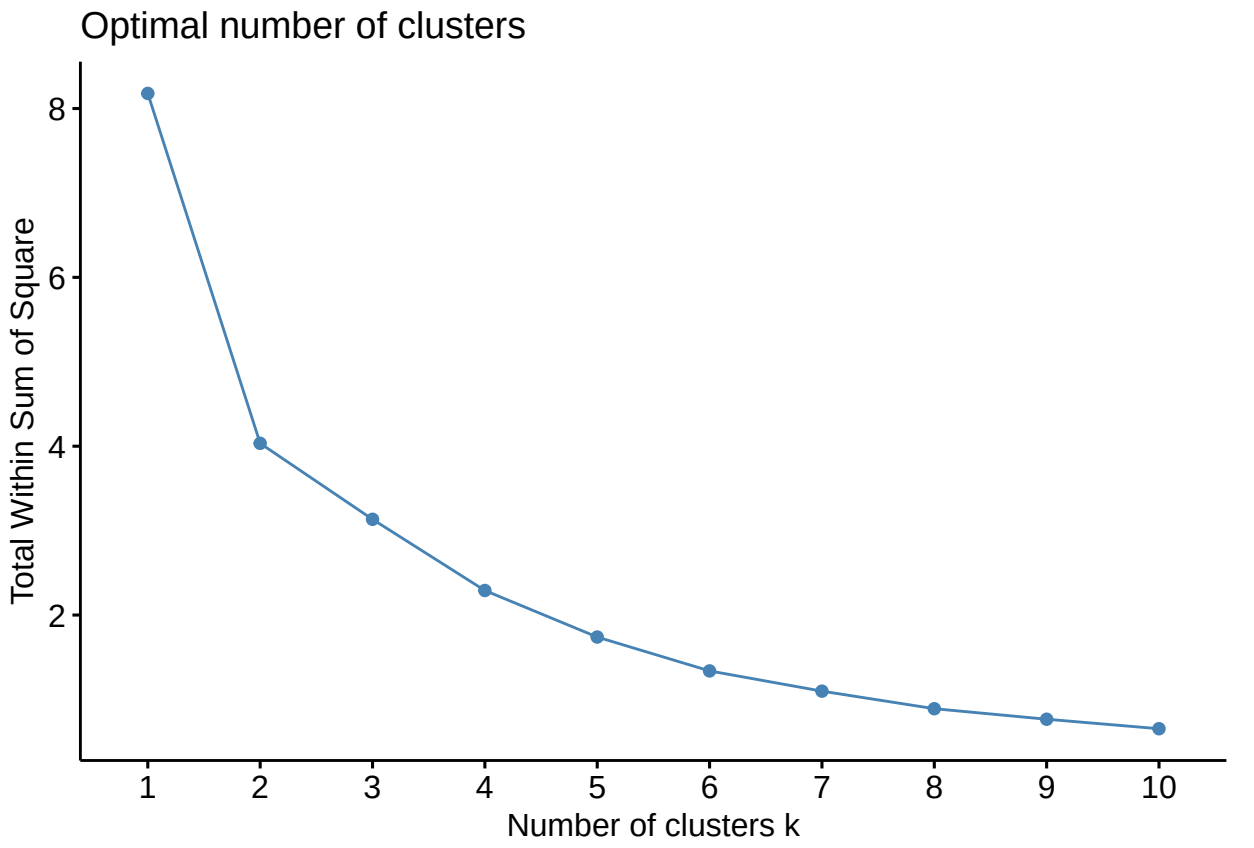
```
sum(bikeshare_data1$cluster_station==4)
```

```
## [1] 67
```

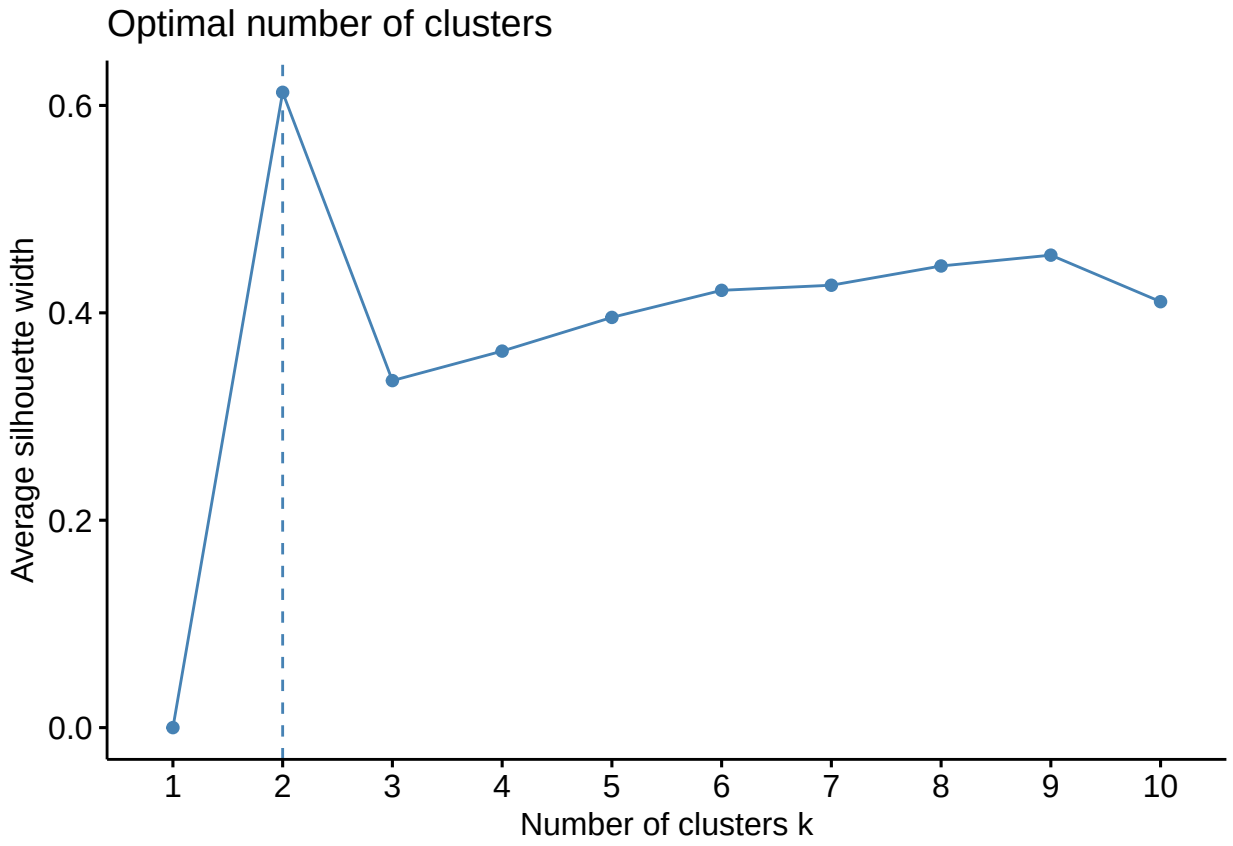
```
sum(bikeshare_data1$cluster_station==5)
```

```
## [1] 16
```

```
fviz_nbclust(bikeshare_dataframe1, FUN = hcut, method = "wss")
```




```
fviz_nbclust(bikeshare_dataframe1, FUN = hcut, method = "silhouette")
```




```
theme_set(theme_bw(16))
washingtonMap <- qmap("Washington DC", zoom = 11, color = "bw")
```

```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=Washington%20DC&zoom=11&size=640x640&s
```

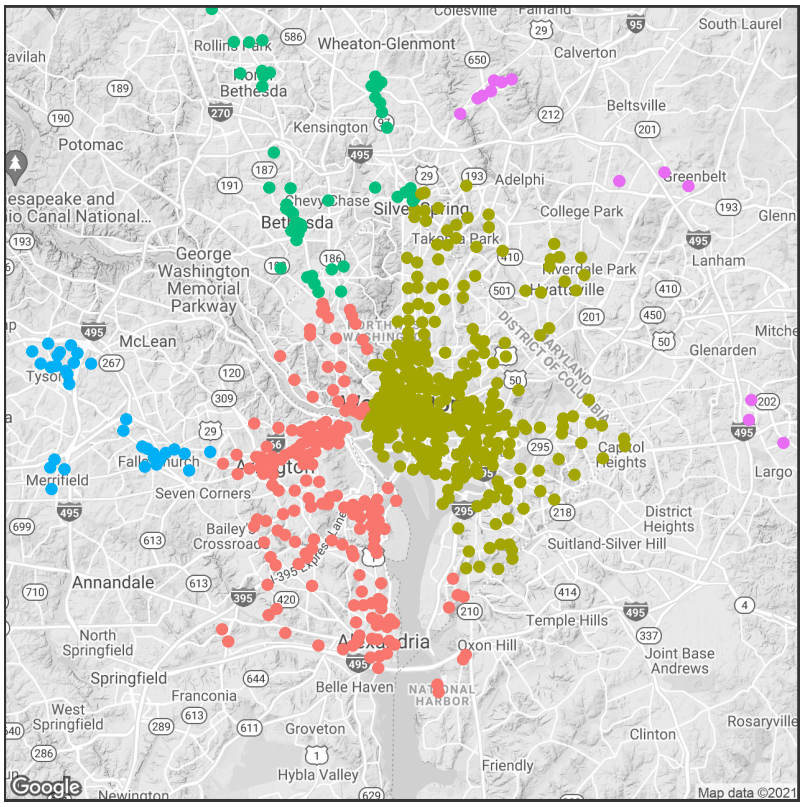
```
## Source : https://maps.googleapis.com/maps/api/geocode/json?address=Washington+DC&key=xxx
```

```
k <- washingtonMap +
geom_point(aes(x = lon, y = lat, colour=cluster_station),
data = bikeshare_data1)

k+ggtitle("Clusters on WashingtonDC Map")
```

```
## Warning: Removed 60 rows containing missing values (geom_point).
```

Clusters on WashingtonDC Map



cluster_station

- 1
- 2
- 3
- 4
- 5

```
#Scaled clustering (final clustering using scaled variance)
label1_standardized <-
  label1 %>%
  mutate(weighted_lon = 0.45*scale(lon), weighted_lat = 0.45*scale(lat),weighted_var=0.1*scale(variance)


bikeshare_dataframe <- label1_standardized %>%
  select(weighted_lon,weighted_lat,weighted_var) %>%
  data.frame()
```

```
clusters1 <- hclust(dist(bikeshare_dataframe))
```

```
bikeshare_data <- bikeshare_dataframe %>%
  mutate(cluster_station = factor(cutree(clusters1, 12)))

sum(bikeshare_data$cluster_station==1)
```

```
## [1] 63
```

```
sum(bikeshare_data$cluster_station==2)
```

```
## [1] 90
```

```
sum(bikeshare_data$cluster_station==3)
```

```
## [1] 45
```

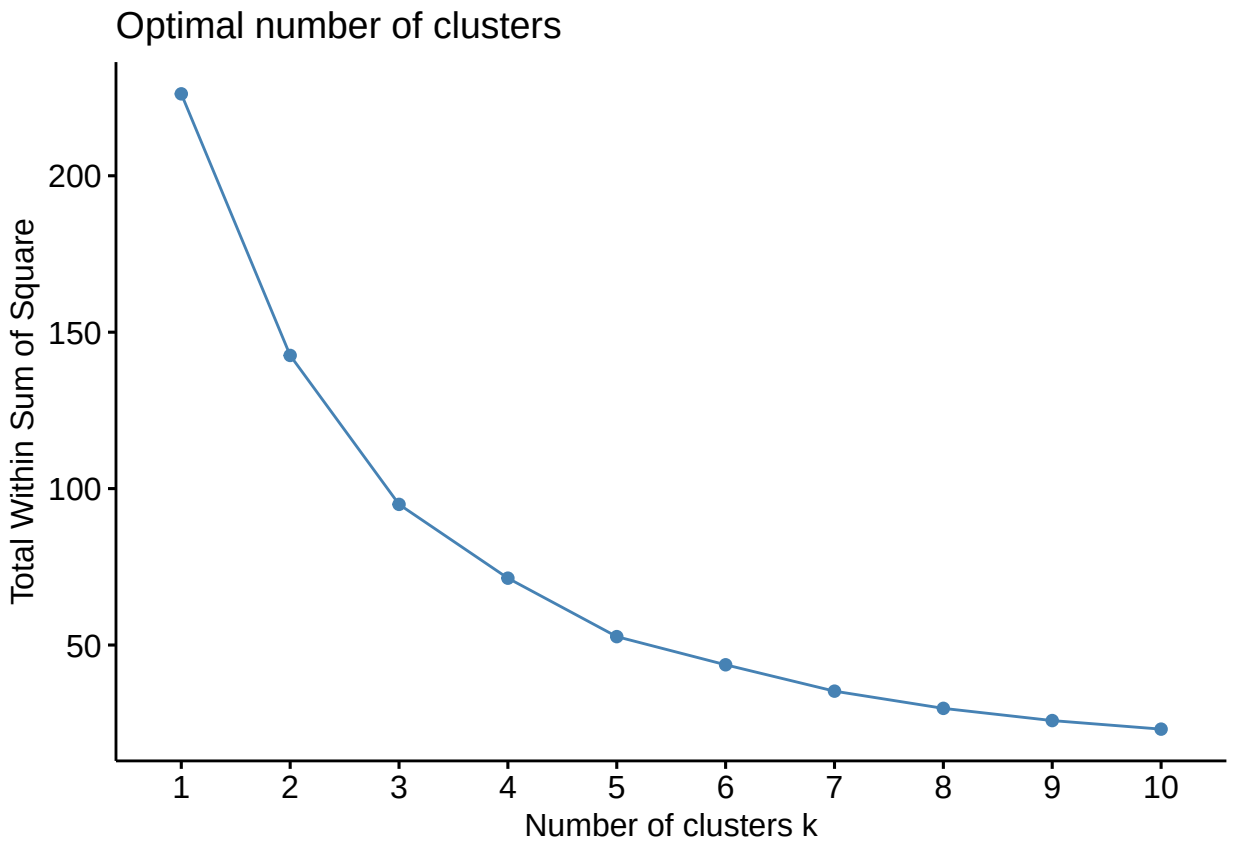
```
sum(bikeshare_data$cluster_station==4)
```

```
## [1] 87
```

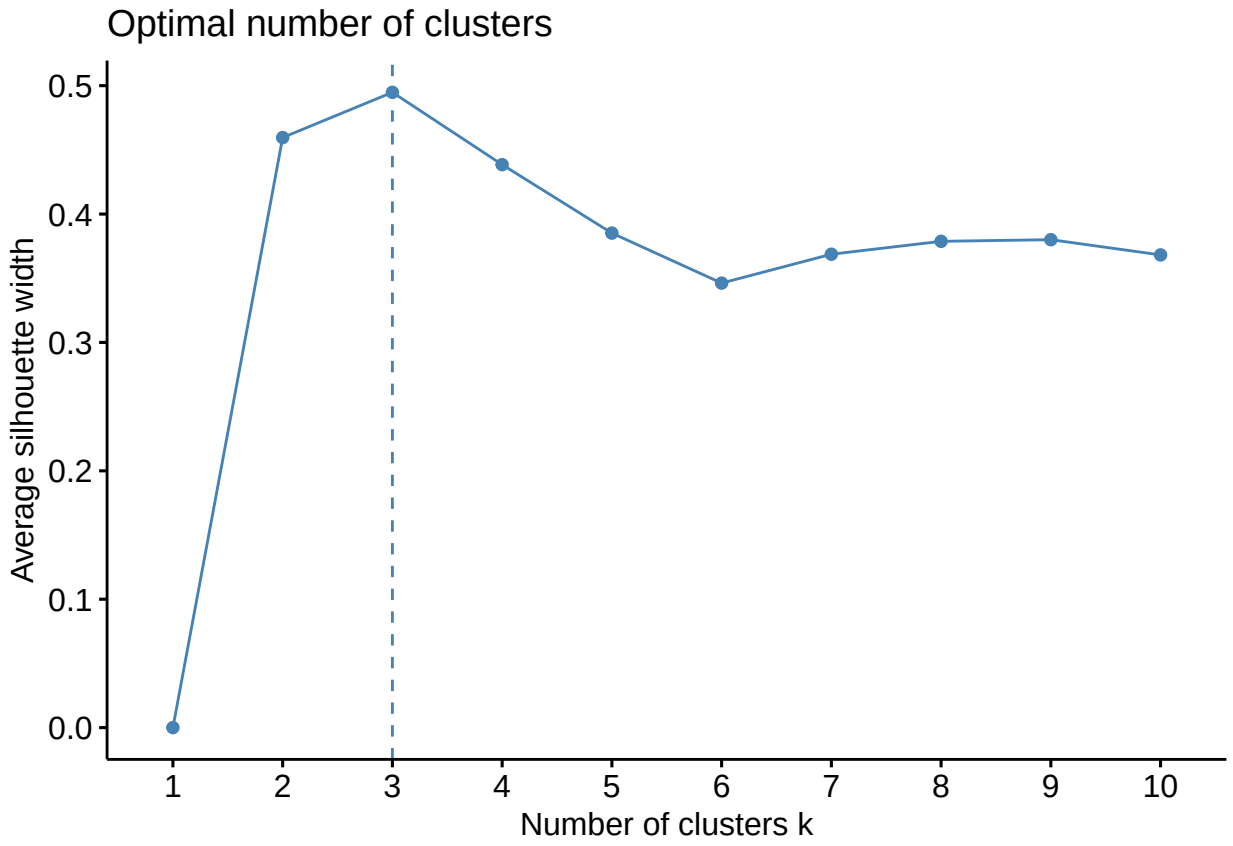
```
sum(bikeshare_data$cluster_station==5)
```

```
## [1] 95
```

```
fviz_nbclust(bikeshare_dataframe, FUN = hcut, method = "wss")
```




```
fviz_nbclust(bikeshare_dataframe, FUN = hcut, method = "silhouette")
```




```
theme_set(theme_bw(16))
washingtonMap <- qmap("Washington DC", zoom = 11, color = "bw")
```

```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=Washington%20DC&zoom=11&size=640x640&s
```

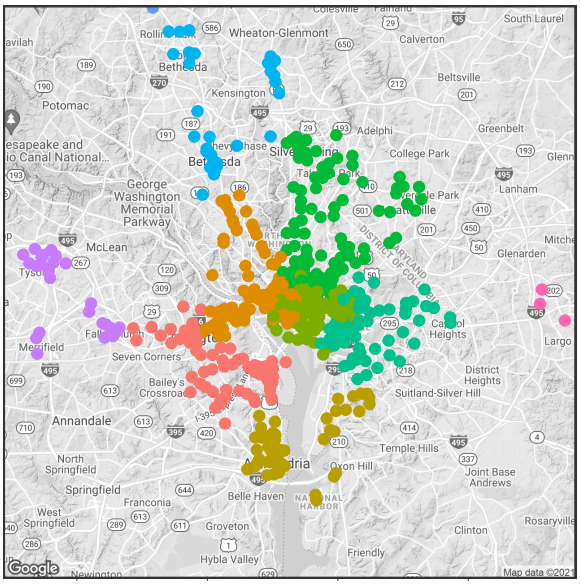
```
## Source : https://maps.googleapis.com/maps/api/geocode/json?address=Washington+DC&key=xxx
```

```
k1 <- washingtonMap +
geom_point(aes(x = lon, y = lat, colour=bikeshare_data$cluster_station),
data = label1)

k1+ggtitle("Clusters on WashingtonDC Map")
```

```
## Warning: Removed 38 rows containing missing values (geom_point).
```

Clusters on WashingtonDC Map



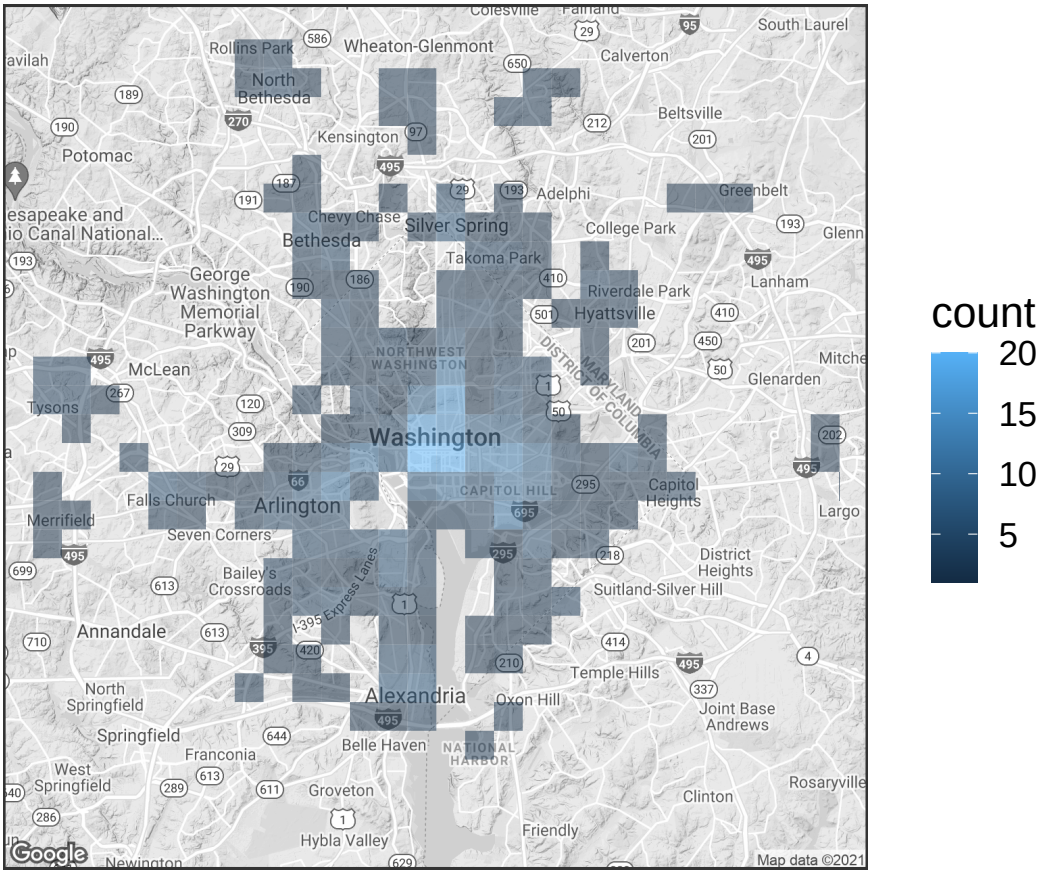
bikeshare_data$cluster_station

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12

```
#Some compensation Plots
washingtonMap +
stat_bin2d(
aes(x = lon, y = lat, colour = station_id, size = station_id),
size = .5, bins = 30, alpha = 1/2,
data = bikeshare_origin)
```

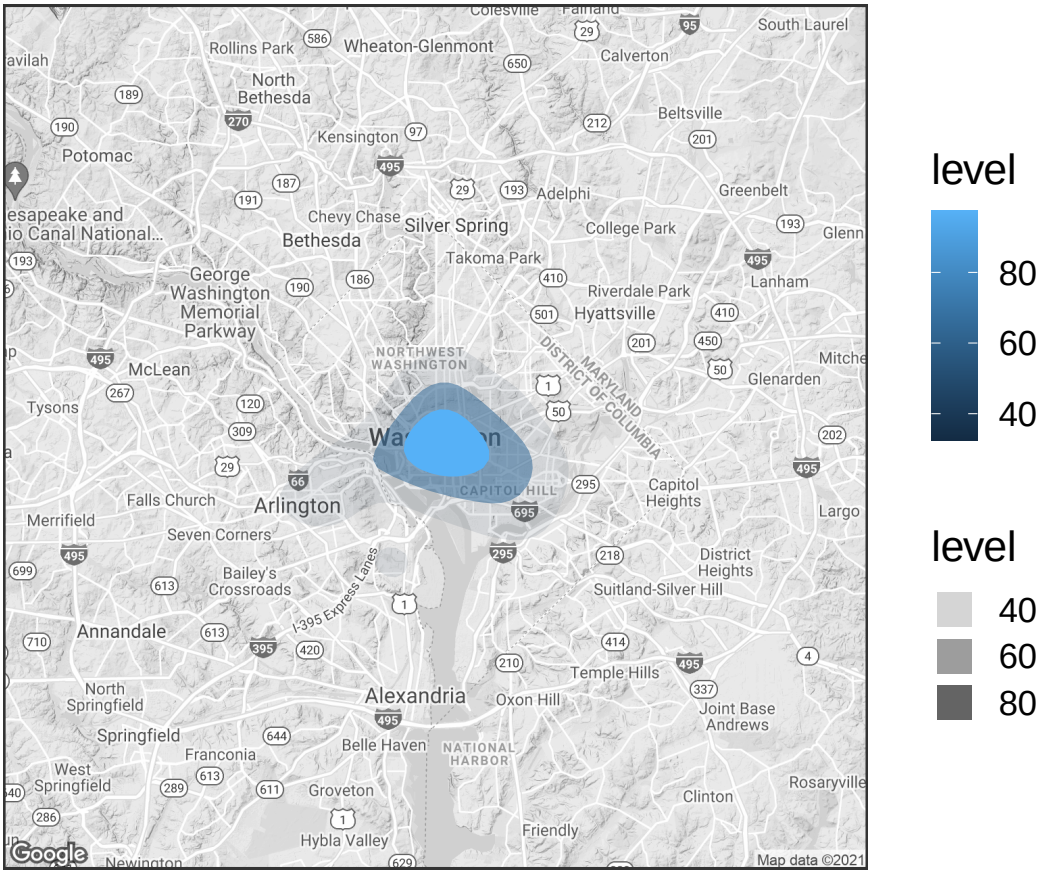
```
## Warning: Removed 60 rows containing non-finite values (stat_bin2d).
```

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




```
washingtonMap +
stat_density2d(
aes(x = lon, y = lat, fill = ..level.., alpha = ..level..),
size = 2, bins = 4, data = bikeshare_origin,
geom = "polygon"
)
```

```
## Warning: Removed 60 rows containing non-finite values (stat_density2d).
```

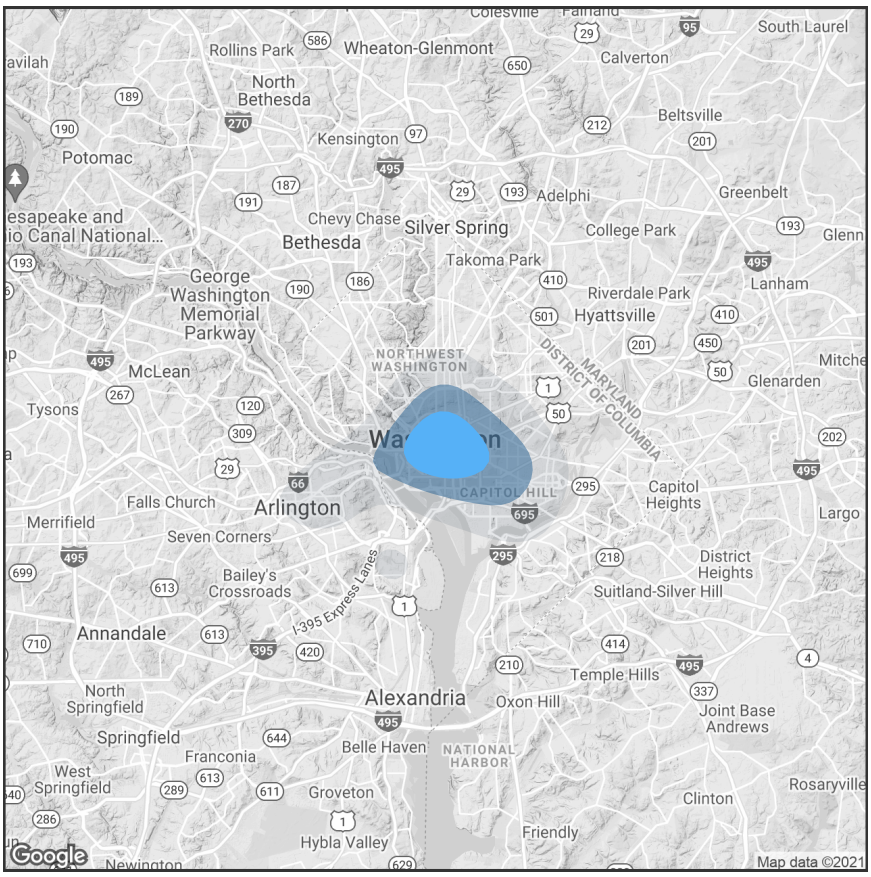


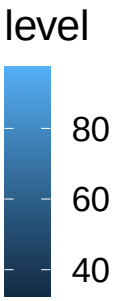

```
overlay <- stat_density2d(
aes(x = lon, y = lat, fill = ..level.., alpha = ..level..),
bins = 4, geom = "polygon",
data = bikeshare_origin
)
washingtonMap + overlay + inset(
grob = ggplotGrob(ggplot() + overlay + theme_inset()),
xmin = -95.35836, xmax = Inf, ymin = -Inf, ymax = 29.75062
)
```

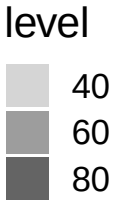
```
## Warning: Removed 60 rows containing non-finite values (stat_density2d).
```

```
## Warning in out$x[is.infinite(data$x)] <- squish_infinite(data$x): number of
## items to replace is not a multiple of replacement length
```

```
## Warning in out$y[is.infinite(data$y)] <- squish_infinite(data$y): number of
## items to replace is not a multiple of replacement length
```








```
#Station Selection using variance:
#label1$cluster_station <- bikeshare_data$cluster_station
#data1 <- label1_standardized %>%
  #filter(cluster_station == 4) %>%
  #arrange(variance)

#print(data1$station_id[67:87])
```

```
#Station Selection without using variance:
bikeshare_origin$cluster_station <- bikeshare_data1$cluster_station
data2 <- bikeshare_origin %>%
  filter(cluster_station == 2)

print(data2$station_id[1:20])
```

```
##  [1] 31100 31101 31102 31103 31104 31105 31106 31107 31108 31201 31202 31203
## [13] 31204 31205 31400 31401 31502 31600 31601 31602
```

```
#Final 20 stations:
a = c(31623, 31209, 31233, 31230, 31243, 31205, 31277, 31200, 31101, 31217, 31248, 31272, 31227, 31268,

plotting <- bikeshare_origin %>%
  filter(station_id %in% a)

theme_set(theme_bw(16))
washingtonMap <- qmap("Washington DC", zoom = 14, color = "bw")
```

```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=Washington%20DC&zoom=14&size=640x640&s
```

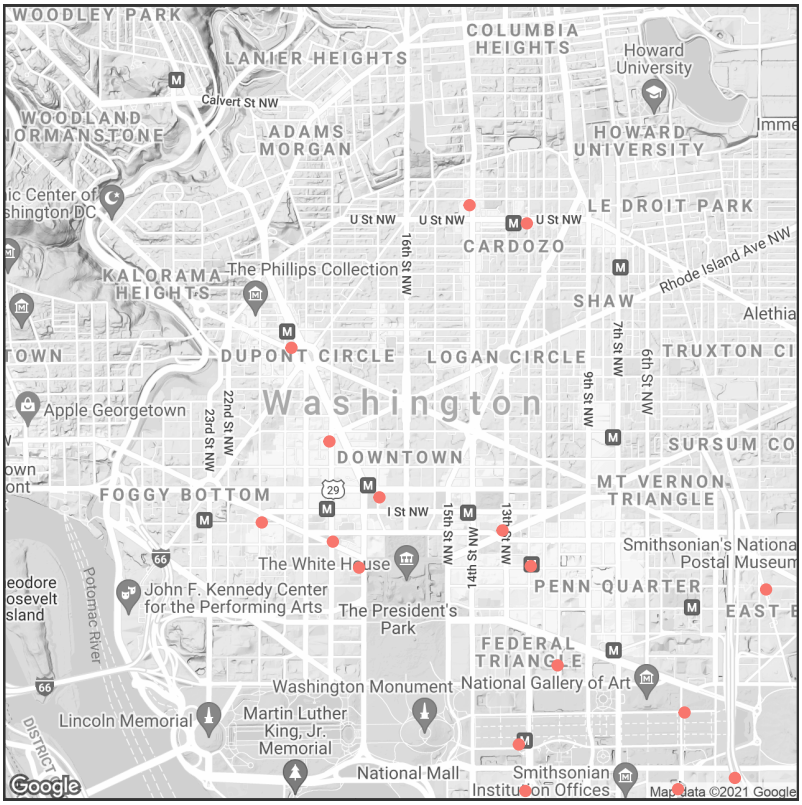
```
## Source : https://maps.googleapis.com/maps/api/geocode/json?address=Washington+DC&key=xxx
```

```
k2 <- washingtonMap +
geom_point(aes(x = lon, y = lat,colour="blue"),
data = plotting)

k2+ggtitle("Selected 20 Stations")
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```

Selected 20 Stations



colour

- blue