Hi Folks! This is my first blog and I am super excited to share with you how I used R Programming to work upon a location based strategy in my E commerce organization.

Please check out [r-bloggers.com](https://www.google.com/url?q=http://r-bloggers.com&sa=D&ust=1561046396849000&usg=AFQjCNH1KB8Gz9uaGMg860SV23ZSfJlDEg) for more exciting stuff on R

***Just a little brief about the problem statement***

I work for an e-commerce organization (an online travel platform) for booking hotels and flights based out of India. This problem is concerned with the Hotel department.

Each locality in a city behaves differently based on certain features of the locality e.g. Airport Zone of a city would behave differently from a Central Zone in the vicinity of a famous Historical Site. Therefore separate strategies are required for different areas for monitoring and controlling parameters such as Inventory, Production and Demand.

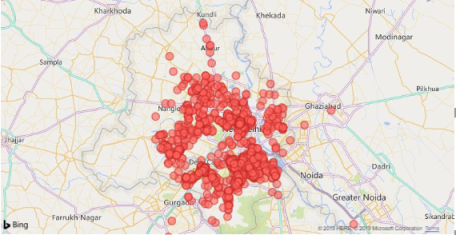
In the data I had latitude and longitude for each hotel and the task was to identify clusters of these hotels or what we call a hyperlocation.

**Let’s Get Started**

This is how the data looks



Let’s look at it Visually (I am using Power BI here)

Set of Hotels from Delhi

**Outlier Removal** In the image above we can see there are certain hotels outside the city that can create problems while forming clusters, let’s remove these outliers statistically.

library(geosphere)

#Mean of Lat Lon

MeanLat<- mean(HotelsCity$latitude, na.rm = TRUE)

MeanLon<- mean(HotelsCity$longitude, na.rm = TRUE)

#Distance of all hotels from mean lat lon

HotelsLatLon<- HotelsCity[,c(4,5)]

MeanLatLon<- data.frame(MeanLat,MeanLon)

Distance\_Mat<- distm(HotelsLatLon[2:1],MeanLatLon[2:1],fun = distHaversine)

Distance\_Mat<- as.data.frame(Distance\_Mat)

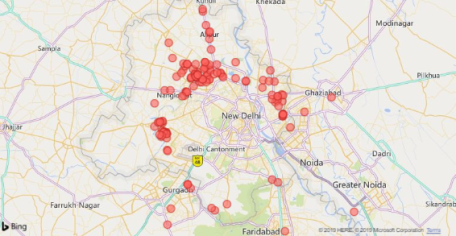
#Calculating Cutoff Distance for Outlier Removal

IQR<- IQR(as.numeric(Distance\_Mat[,1]),na.rm = TRUE)

Cutoff<- as.numeric(quantile(Distance\_Mat$V1,0.75,na.rm = TRUE)+IQR\*1.5)

HotelDetail$Flag<- ifelse(HotelDetail$V1>Cutoff,"Incorrect","Correct")

Outliers\_Final<- filter(HotelDetail,Flag=="Incorrect")

This is how the outliers look when plotted

**Clustering**

After cleaning the data (outlier removal) now let’s create a distance matrix i.e. distance of each hotel from every other hotel, I am doing this using the geosphere library in R.

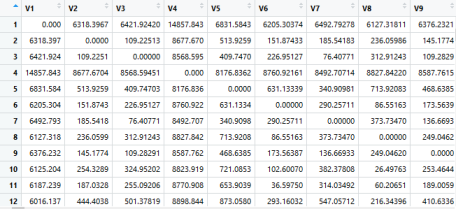
#Distance Matrix for city

Distance\_Mat<- distm(HotelsLatLon[2:1],HotelsLatLon[2:1],fun = distHaversine)

Distance\_Mat<- as.data.frame(Distance\_Mat)

Distance\_Mat[is.na(Distance\_Mat)]<-0

DMat<- as.dist(Distance\_Mat)

This is how the distance matrix looks like.

**Let’s Create Clusters now.**

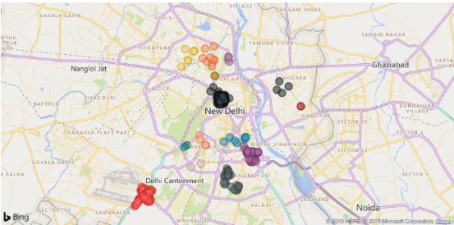
#Hierarchical Clustering

hc <- hclust(DMat, method="complete")

HotelCity\_Valid$Clusters<- cutree(hc, h=AvgDist2)

In the above code snippet in the cutree function I have used a different cutoff distance for different cities. How I arrived at that distance is a different science altogether, in this case the cutoff distance is around 2 KMs which means that each cluster would be roughly of a diameter of 2 KMs.

This is how these different clusters look like when plotted



How I named these localities? There was a system name tagged to each hotel’s locality, I used the most frequent name in that cluster as the Cluster Name

Please reach out to me at shubham.mpb95@gmail.com for any kind of queries regarding this.