P4_Hershey

Sohrab Rahimi

January 26, 2017

Motivation

In this project I will explore a candy sale data set. This data includes infomation about the candy sale for 1253 markets in four major cities (i.e. Philadelphia, Boston, DC and Pittsburgh) for 31 types of chocolate. The geographic information about the markets are also included.

I will try to answer the following questions in this project:

1- Which Brands sell more and in which cities? 2- Which zipcodes consume more candy on average? 3- which markets sell more candies and what kinds of candies? 4- Which candies are more likely to be sold together? 5- How different markets compare together interms of their candy sales? 6- What parts of the city consume more chocolate? how different neighborhoods compare in terms of their chocolate consumption?

First we load all the required packages into R:

The data includes 1253 rows each representing one store. For each store the sales data for 31 brands of chocolate has been provided.

```
##  [1] "STORE"              "LONG"               "LAT"
##  [4] "ZIP.CODE"           "CITY"               "ALLAN"
##  [7] "BLISS"              "BREATHSAVERS"       "BROOKSIDE"
## [10] "BUBBLE.YUM"         "CADBURY"            "CANE.STK"
## [13] "CANISTER"           "GOOD.N.PLENTY"      "HEATH"
## [16] "HERSHEY.ARTISAN"    "HERSHEY.CHOC.ASST"  "HERSHEY.CHOC.REM"
## [19] "HERSHEY.GUM.REM"    "HERSHEY.NCHOC.REM"  "HERSHEY.SUGAREE.BFY"
## [22] "ICE.BREAKERS"       "JOLLY.RANCHER"      "KISSES"
## [25] "KIT.KAT"            "LANCASTER"          "LORENA"
## [28] "ALMOND.JOY"         "PAYDAY"             "PELON"
## [31] "POT.OF.GOLD"        "REESE"              "ROLO"
## [34] "SL.HARD.ROLL"       "SNACK.BITES"        "TAKE5"
## [37] "TOTAL"
```

```
## [1] 1253   37
```

STORE	LONG	LAT	ZIP.CODE	CITY
factor	numeric	numeric	integer	factor

the rest of 32 variables are 32 candy brands sales in USD (integer). The data is collected for four cities:

```
## [1] Boston       Washington   Pittsburgh   Philadelphia
## Levels: Boston Philadelphia Pittsburgh Washington
```

The data includes the candy sale data for 37 chain-markets:

```
##  [1] Big Kmart                 Cumberland Farms
##  [3] Dollar Express            Family Dollar
##  [5] Hannaford                 Market Basket
##  [7] Rite Aid                  Shaws Supermarket
##  [9] Star Market               Stop & Shop
## [11] Target Store              Walgreens
## [13] Wegmans Food Market       BJs Wholesale Club
## [15] Bethesda Mini Mart        Dollar General
## [17] Giant Food Store          Harris Teeter
## [19] Navy Bethesda Autoport    Navy Exchange
## [21] Safeway Store             ShopRite
## [23] Shoppers Food Warehouse   Super Giant
## [25] Target Express            Walgreens Rx Express
## [27] Giant Eagle Market        Circle K
## [29] AAFES Coraopolis Shoppette Giant Eagle Express
## [31] Pilot Travel Center       Super Kmart Center
## [33] Martins Food Market       Kmart
## [35] Weis Market               Acme Market
## [37] Browns Family ShopRite
## 37 Levels: AAFES Coraopolis Shoppette Acme Market ... Weis Market
```

for 30 candy brands:

```
##  [1] "ALLAN"               "BLISS"               "BREATHSAVERS"
##  [4] "BROOKSIDE"           "BUBBLE.YUM"          "CADBURY"
##  [7] "CANE.STK"            "CANISTER"            "GOOD.N.PLENTY"
## [10] "HEATH"               "HERSHEY.ARTISAN"     "HERSHEY.CHOC.ASST"
## [13] "HERSHEY.CHOC.REM"    "HERSHEY.GUM.REM"     "HERSHEY.NCHOC.REM"
## [16] "HERSHEY.SUGAREE.BFY" "ICE.BREAKERS"        "JOLLY.RANCHER"
## [19] "KISSES"              "KIT.KAT"             "LANCASTER"
## [22] "LORENA"              "ALMOND.JOY"          "PAYDAY"
## [25] "PELON"               "POT.OF.GOLD"         "REESE"
## [28] "ROLO"                "SL.HARD.ROLL"        "SNACK.BITES"
## [31] "TAKE5"
```

which markets sell more candies and what kinds of candies?

In order to get a sense of the candy sales in each city I will first create a “TOTAL” column where I add the sales for all types of candy for each market. Based on the plot below we can see that the target store sells highest in all except for Philadelphia. For Philadelphia alone, BJs wholesale club is slightly more than target on average.

```
data$TOTAL = rowSums(subset(data,
                            select = -c(STORE,LONG,LAT,ZIP.CODE,CITY)))

candy_by_city = data %>%
  group_by(CITY) %>%
  summarise (sales_mean = mean(TOTAL),
             sales_median = median(TOTAL),
             sales_sum = sum(TOTAL), n = n())
candy_by_city = as.data.frame(candy_by_city)
candy_by_city
```

```
##           CITY sales_mean sales_median sales_sum   n
## 1       Boston   103805.9        75528  10172974  98
## 2 Philadelphia   103710.9        58909  55174182 532
## 3   Pittsburgh    76240.9        38109  36290670 476
## 4   Washington   102167.1        82232  15018564 147
```

```
# convert to long format
candy_by_city_long = gather(candy_by_city, sales_mean,sales_median,sales_sum,
                            sales_mean:sales_median:sales_sum, factor_key=TRUE)
names(candy_by_city_long)[names(candy_by_city_long) == 'sales_mean'] <- "Measure"
names(candy_by_city_long)[names(candy_by_city_long) == 'sales_median'] <- "value"
```

We can do the same thing for stores to compare different types of stores in terms of their success in selling candies.

The top five successful brands and stores, respectively, are prtinted below:

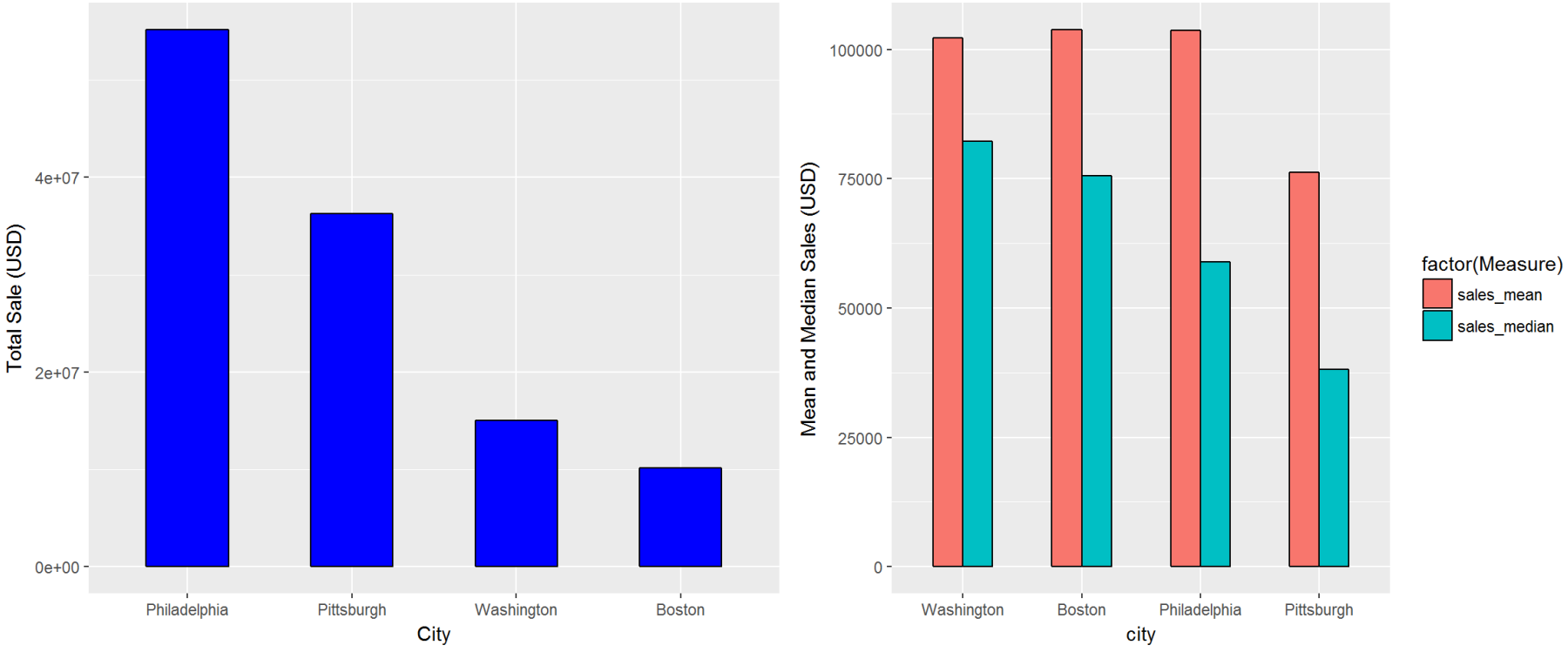
```
sales_by_candy[order(-sales_by_candy$sales_mean),][1:5,]
```

```
##               BRAND sales_mean sales_median sales_sum    n
## 27            REESE  14006.275         8187  17549863 1253
## 19           KISSES   6040.875         2639   7569217 1253
## 20          KIT.KAT   5606.303         3327   7024698 1253
## 17     ICE.BREAKERS   3486.207         1847   4368217 1253
## 13 HERSHEY.CHOC.REM   3145.659          998   3941511 1253
```

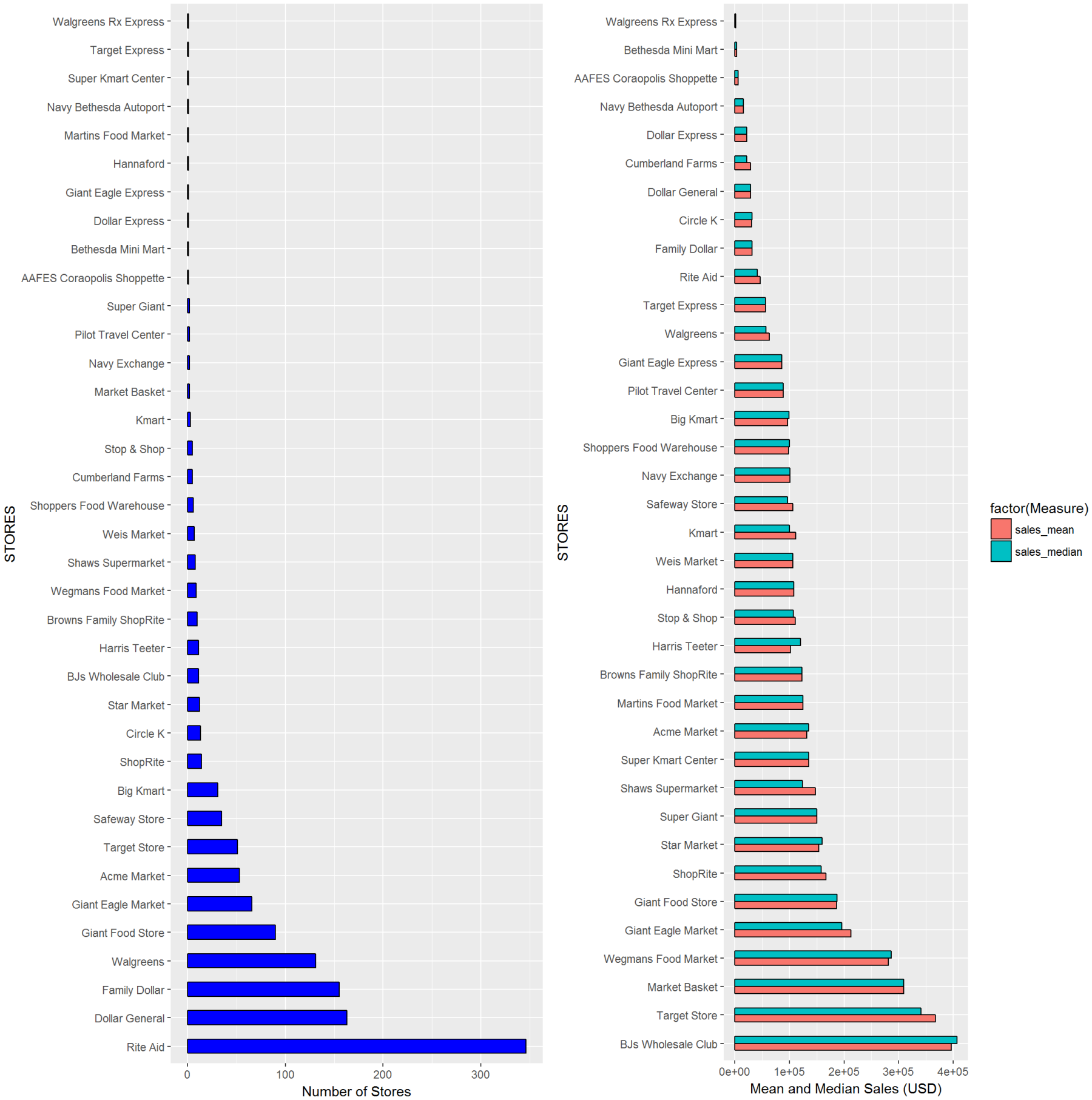
```
candy_by_store[order(-candy_by_store$sales_mean),][1:5,]
```

```
##                  STORE sales_mean sales_median sales_sum  n
## 5   BJs Wholesale Club   396078.2       407124   4356860 11
## 33        Target Store   367212.7       340962  18727850 51
## 18       Market Basket   309548.0       309548    619096  2
## 36 Wegmans Food Market   281409.1       286844   2532682  9
## 13  Giant Eagle Market   212501.4       196071  14025094 66
```

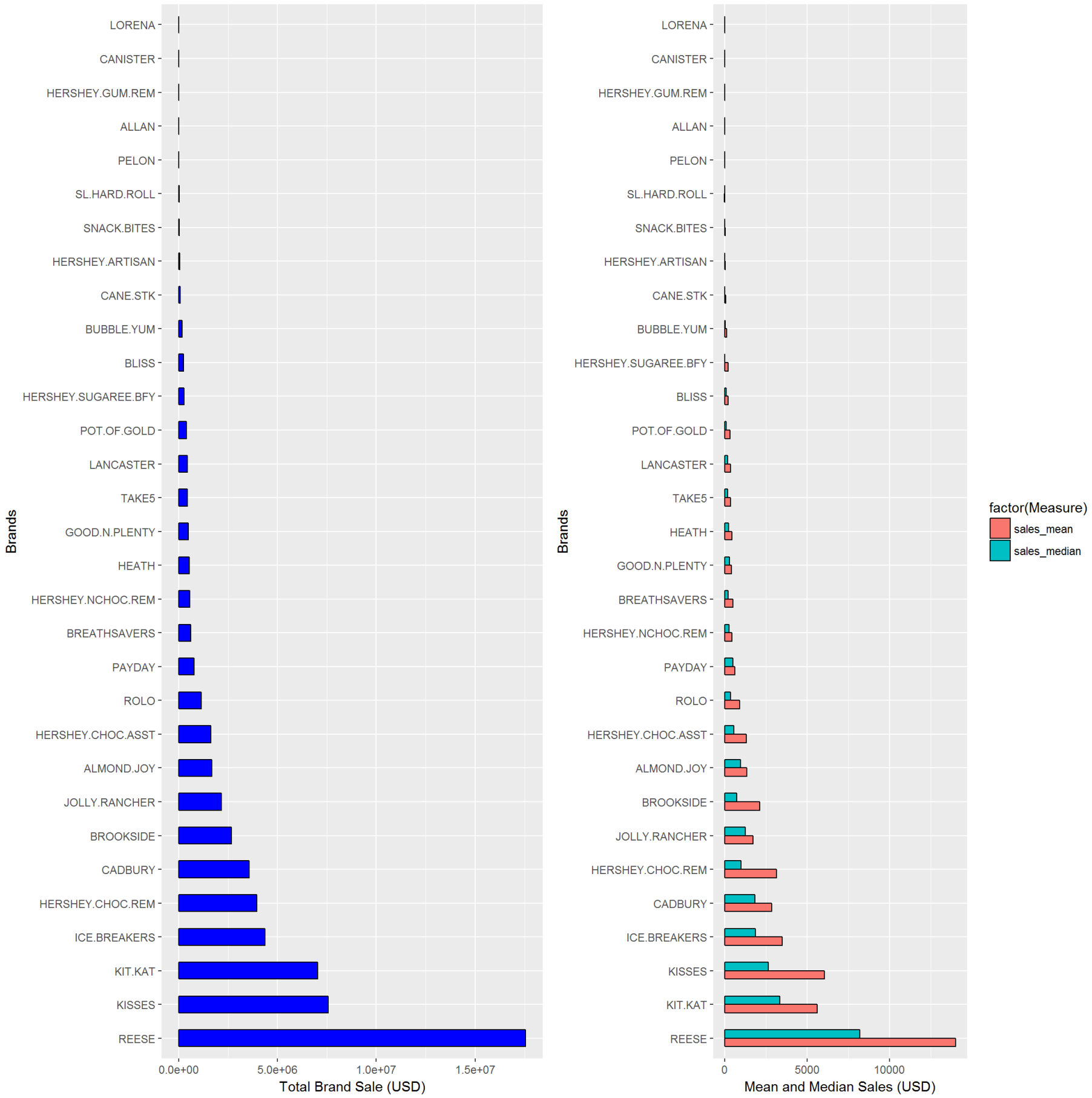
we can now go ahead and visualize some basic statistics in this data. In the two plots below we can see that Pittsburgh is significantly lower in candy sale from the three other cities.the median candy sale for stores in washington is highest than other cities, and the mean is almost the same as Boston and Philadelphia, meaning that Washington does well constantly in all stores.



We can now compare different Stores in their capacity for candy sale. In the plots below we can see that the number of Rite Aid stores is largesr, still, each Rite Aid Store perfoms bad on average. The best average sale belongs to BJs Wholesale Clubs and Target Stores. The number of Target Stores is 8th in rank in terms of frequency, meaning that Target stores are significant markets for candy sale in these four cities.

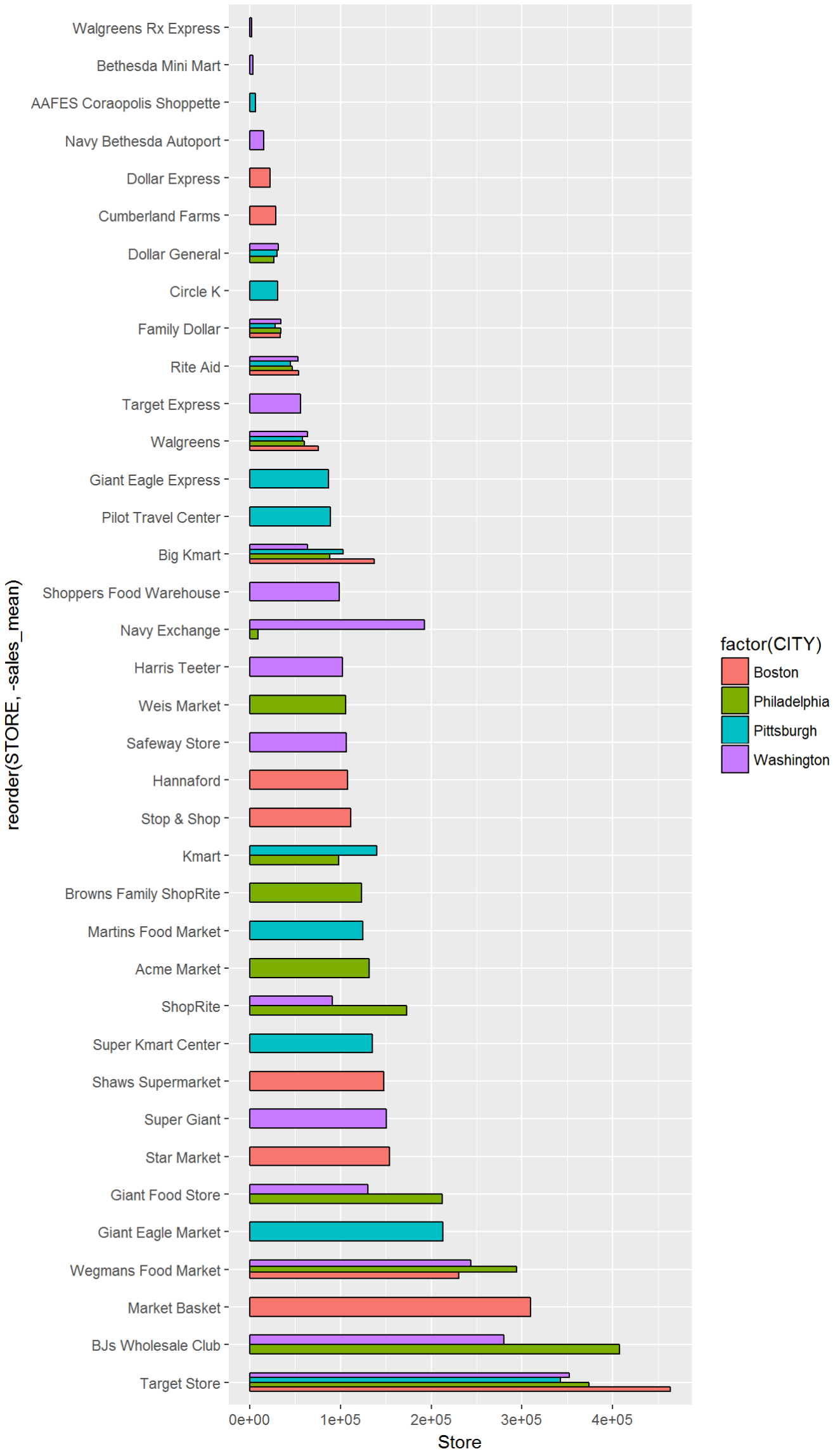


The brand sale plots below, indicate that Reese is the most successful overall. Kit-Kat and Kisses follow with a significant margin.



To see how each store performs in each city we should group by both store and city. Target store is Significantly high in all four cities although more successful in Boston. in philadelphia only BJs Wholesale club is nmore successful than Target stores.

```
candy_by_store_city = data %>%
  group_by(STORE,CITY) %>%
  summarise (sales_mean = mean(TOTAL),
             sales_median = median(TOTAL),
             sales_sum = sum(TOTAL), n = n())

ggplot(aes(x = reorder(STORE,-sales_mean),  y= sales_mean),
       data = candy_by_store_city)+
  geom_bar(aes(fill = factor(CITY)),stat="identity",
           position = "dodge",width=.5,color=I('black')) +
  coord_flip()+ylab("Store")
```



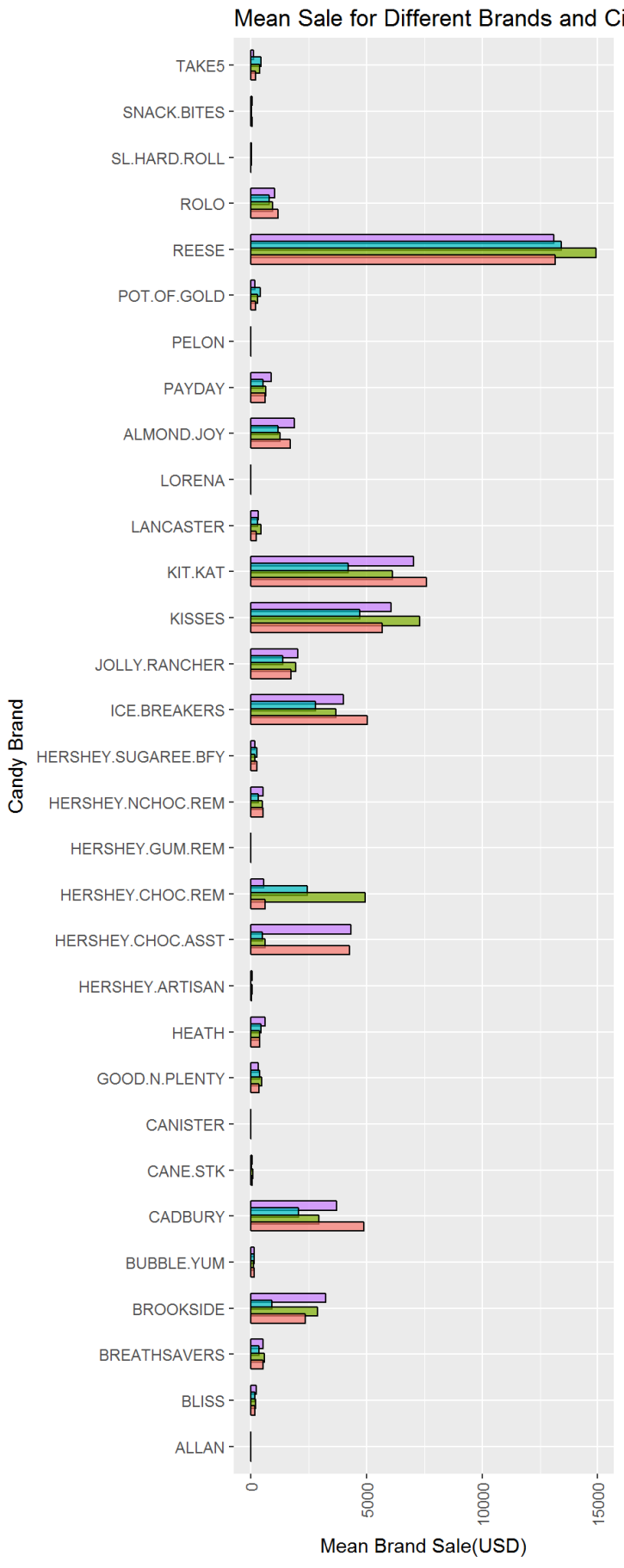
Which Brands sell more and in which cities?

Now let's convert the data to long format so that we can analyse each brand same as we did for each store. the plot below indicates how different brands compare in different cities. We can see that "Reese" is successful in all 4 cities.Some cities are specifically higher than others in some brands. For example, Hershey's Chocolate Assortment is significantly higher in Boston and DC. KitKat and Cadbury are highest in Boston while Philadelphia is more successful in Reese and Kisses.

```
data_long = melt(data,id.vars = c("STORE","LONG","LAT","ZIP.CODE","CITY","TOTAL"))
# change the name of the newly created variables
names(data_long)[names(data_long) == 'variable'] <- "BRAND"
names(data_long)[names(data_long) == 'value'] <- "BRAND.SALE"
names(data_long)
```

```
## [1] "STORE"      "LONG"       "LAT"        "ZIP.CODE"   "CITY"
## [6] "TOTAL"      "BRAND"      "BRAND.SALE"
```

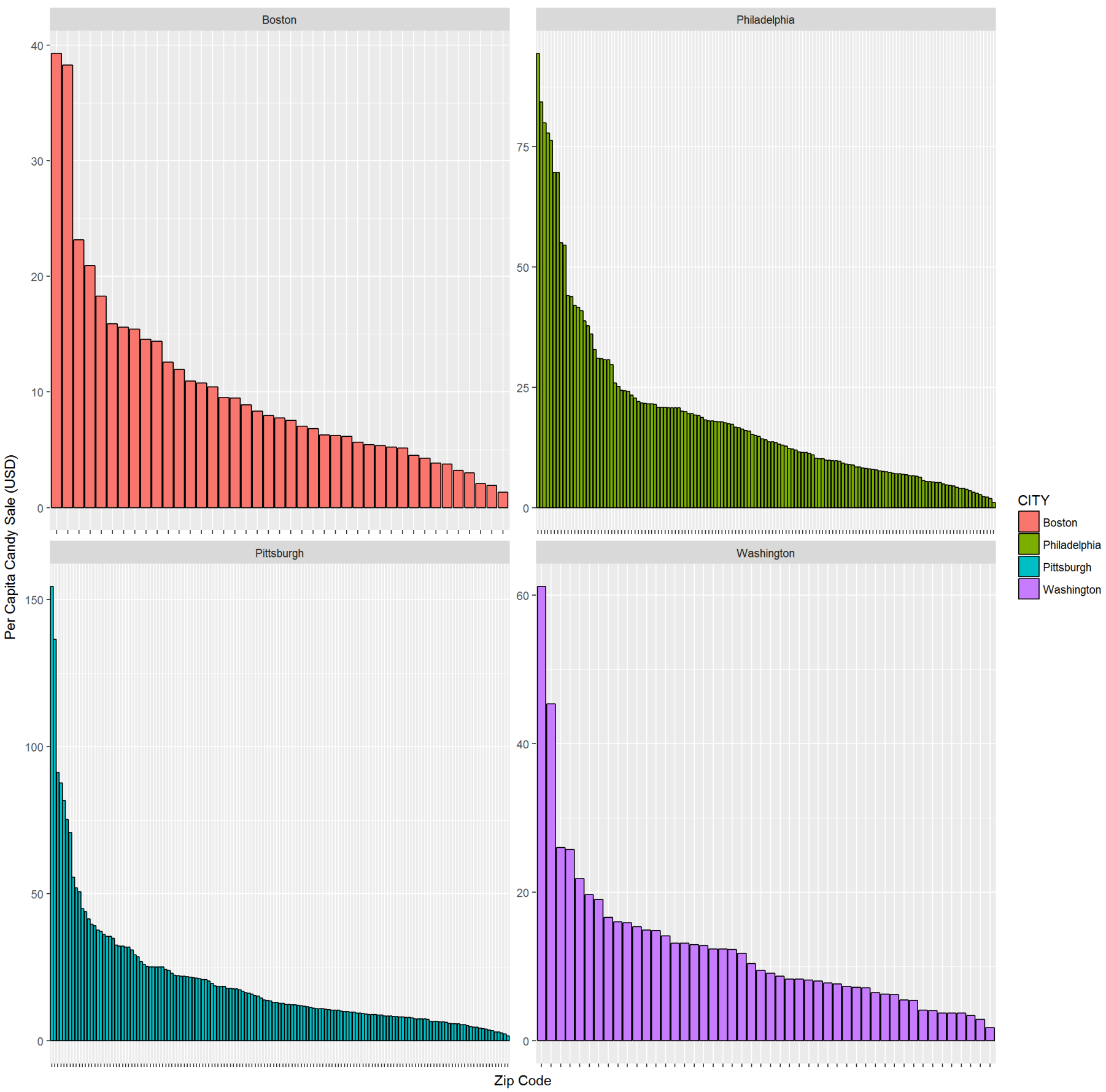
```
p1<- ggplot(aes(x = BRAND,  y= BRAND.SALE), data = data_long) +
  stat_summary(aes(fill = factor(CITY)),colour = "black",fun.y=mean, geom="bar",
               position=position_dodge(0.6), width = 0.8, alpha = 0.7)+
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))+
  xlab("Candy Brand")+ylab("Mean Brand Sale(USD)")+
  labs(title = "Mean Sale for Different Brands and Cities")+
  scale_fill_discrete(guide = guide_legend(title = "CITY"))+ coord_flip()

p2<- ggplot(aes(x = BRAND,  y= BRAND.SALE), data = data_long) +
  stat_summary(aes(fill = factor(CITY)),colour = "black",fun.y=median, geom="bar",
               position=position_dodge(0.6), width = 0.8, alpha = 0.7)+
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))+
  xlab("Candy Brand")+ylab("Median Brand Sale (USD)")+
  labs(title = "Median Sale for Different Brands and Cities")+
  scale_fill_discrete(guide = guide_legend(title = "CITY"))+ coord_flip()

grid.arrange(p1,p2,ncol = 2)
```



Which zip codes consume more candy on average?

We can now check the zip codes to see which ones sell more candies. I use the Census 2010 population data to normalize the candy sales on population. I found that three zipcodes are extremely higher than others so I removed them (i.e. “19112”,“19372”,“2199”).

```
zipcode_pop = read.csv("C:/Users/sur216/Box Sync/school stuff/Udacity (sur216@psu.edu)/Data Analyst/p4_hershey/Zipcode-pop.c
sv",header = T)

colnames(zipcode_pop)[1]<- "ZIP.CODE"
candy_by_zipcode = data %>%
  group_by(ZIP.CODE,CITY) %>%
  summarise (sales_mean = mean(TOTAL),
             sales_median = median(TOTAL),
             sales_sum = sum(TOTAL), n = n())


candy_by_zipcode=merge(candy_by_zipcode,zipcode_pop,by = "ZIP.CODE")
candy_by_zipcode = candy_by_zipcode[,-8]


ggplot(aes(x =reorder(as.character(ZIP.CODE),-sales_sum/X2010.Population),
           y= sales_sum/X2010.Population),
data =candy_by_zipcode[!candy_by_zipcode$ZIP.CODE%in%c("19112","19372","2199"),]) +
geom_bar(aes(fill = factor(CITY)),colour = 'black', stat = "identity") +
theme(axis.text.x=element_blank()) +
facet_wrap(~CITY, scales = "free", ncol = 2)+
scale_fill_discrete(guide = guide_legend(title = "CITY"))+
  xlab("Zip Code")+ylab("Per Capita Candy Sale (USD)")+
  labs(title = "Per Capita Candy Sale in Different Zip Codes")
```

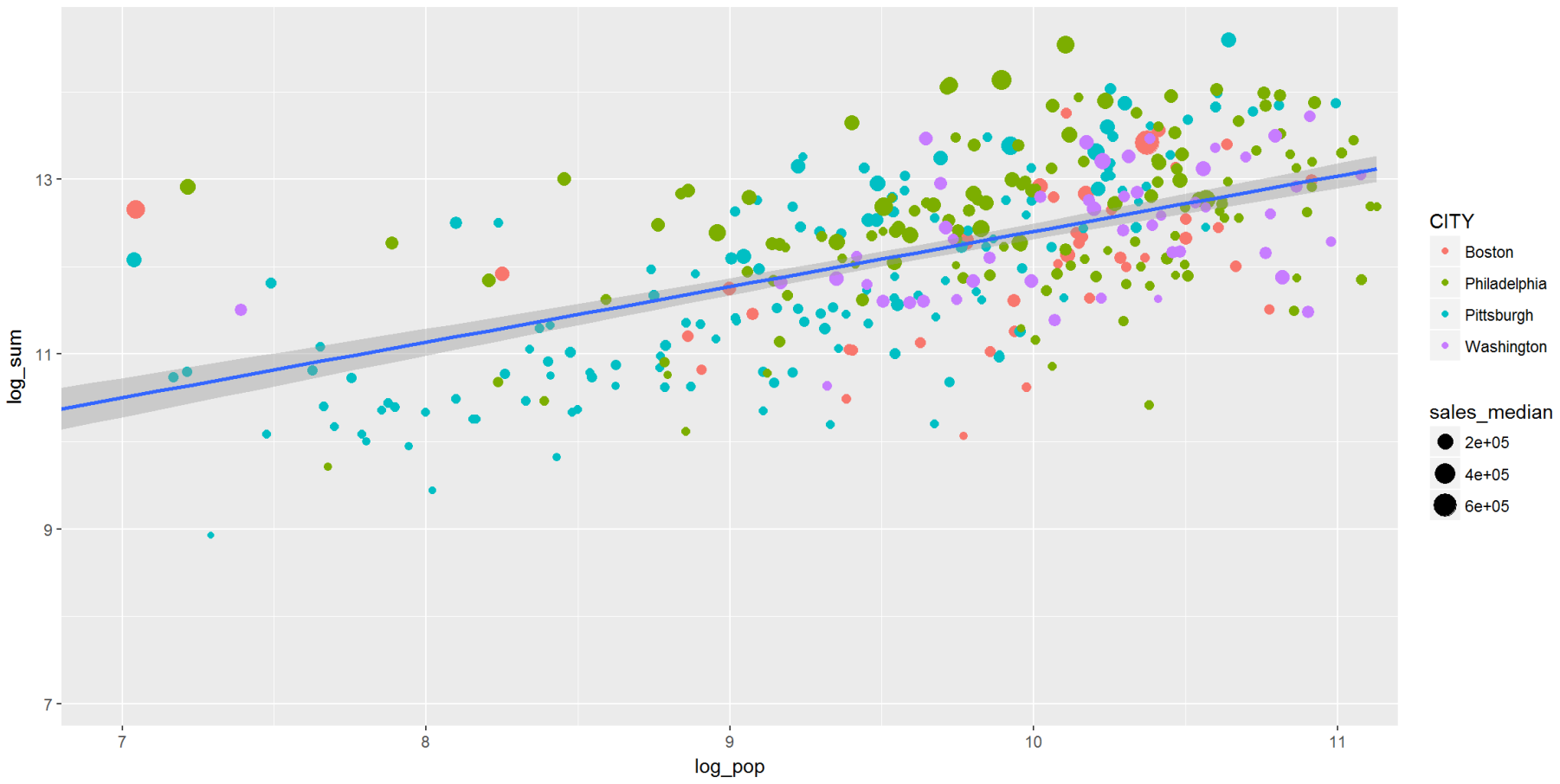


Now we can run a regression to examin the relationship between zipcode population and candy sale. It make sense to have higher candy consumption in zipcodes with higher populations. I will use the log-log transformation since the raw data will give us a fan-shaped scatter plot.

```
candy_fit = lm(sales_sum~X2010.Population,
               data = candy_by_zipcode)
summary(candy_fit)
```

```
##
## Call:
## lm(formula = sales_sum ~ X2010.Population, data = candy_by_zipcode)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -600292 -138972  -73282   80215 1721017
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.793e+04  2.429e+04   4.032 6.69e-05 ***
## X2010.Population 9.908e+00  9.352e-01  10.595  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 272600 on 376 degrees of freedom
## Multiple R-squared:  0.2299, Adjusted R-squared:  0.2279
## F-statistic: 112.3 on 1 and 376 DF,  p-value: < 2.2e-16
```

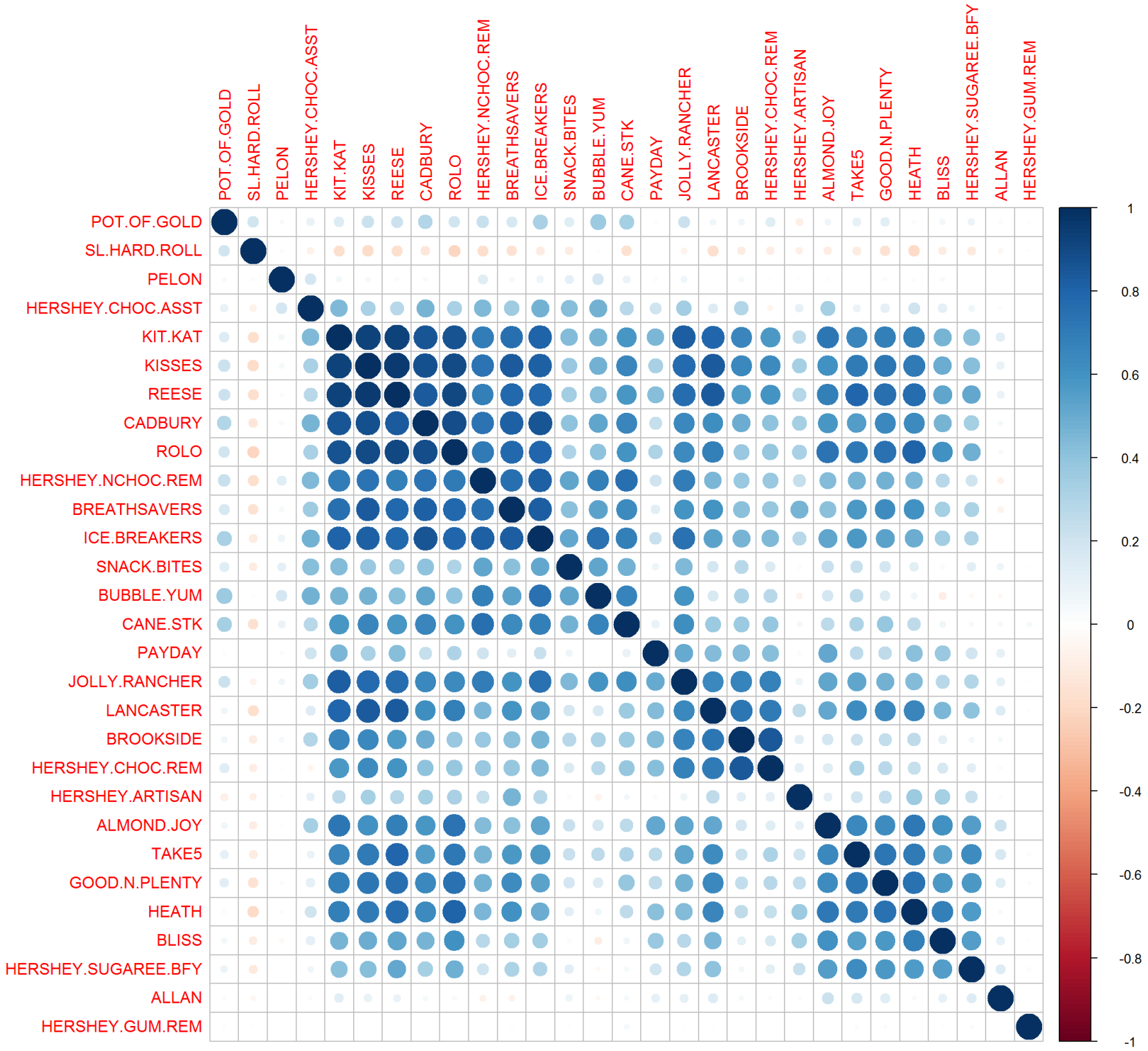
```
candy_zip_trans = transform(candy_by_zipcode,
                            log_pop = log(X2010.Population),
                            log_sum = log(sales_sum))

ggplot(aes(x=log_pop, y =log_sum), data = candy_zip_trans)+
  geom_point(aes(colour = CITY, size = sales_median))+
  geom_smooth(method = 'lm', formula = y~x)+coord_cartesian(xlim = c(7, 11))
```



Which brands sell together?

e first look into the simple correlations between different brands. we should use the wide format for correlations. As the correlation suggests, KitKat, Kisses, Reese, Cadbury, and Rolo are strongly correlated. The scatter plot is ordered through a hierarchial clustering and shows a number of interesting clusters.

```
# isolate all the candiy sales and ignore every other data
candies = data[,c(6:12,14:26,28:36)]
# create a correlation matrix
corr_candies = cor(candies)
# visualize the correlation matrix
corrplot(corr_candies, order = "hclust")
```



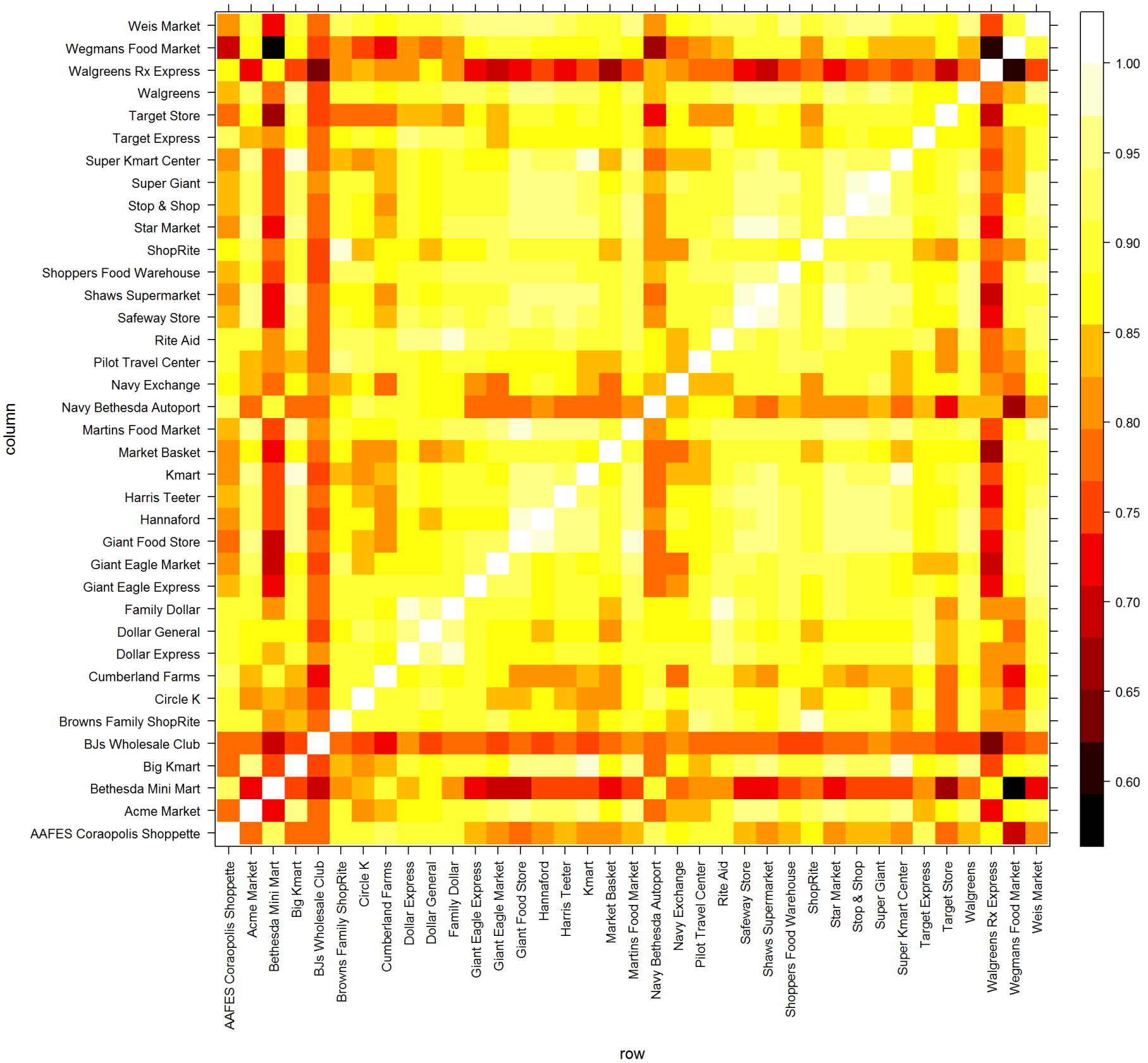
How different markets compare together in terms of their candy sales?

To compare different markets in terms of their candy sale, I will first define a function to calculate the cosine similarity for us. The function below takes the data as matrix and returns a cosine similarity matrix.

```
# define the cosine similarity function
cosine <- function( x, y=NULL ) {

  if ( is.matrix(x) && is.null(y) ) {

    co = array(0,c(ncol(x),ncol(x)))
    f = colnames( x )
    dimnames(co) = list(f,f)

    for (i in 2:ncol(x)) {
      for (j in 1:(i-1)) {
        co[i,j] = cosine(x[,i], x[,j])
      }
    }
    co = co + t(co)
    diag(co) = 1

    return (as.matrix(co))

  } else if ( is.vector(x) && is.vector(y) ) {
    return ( crossprod(x,y) / sqrt( crossprod(x)*crossprod(y) ) )
  } else {
    stop("Error: input should be either a matrix or two vectors")
  }

}
```

Now we will apply this function to find the pairwise similarity matrix for markets. Using the levelplot() function we can visualiza this matrix in form of a heat map. we can now see interesting patterns in the heat map. for example, Target stores are distant from the BJs Wholesale Clubs and similar to Giant stores in terms of their candy sale.

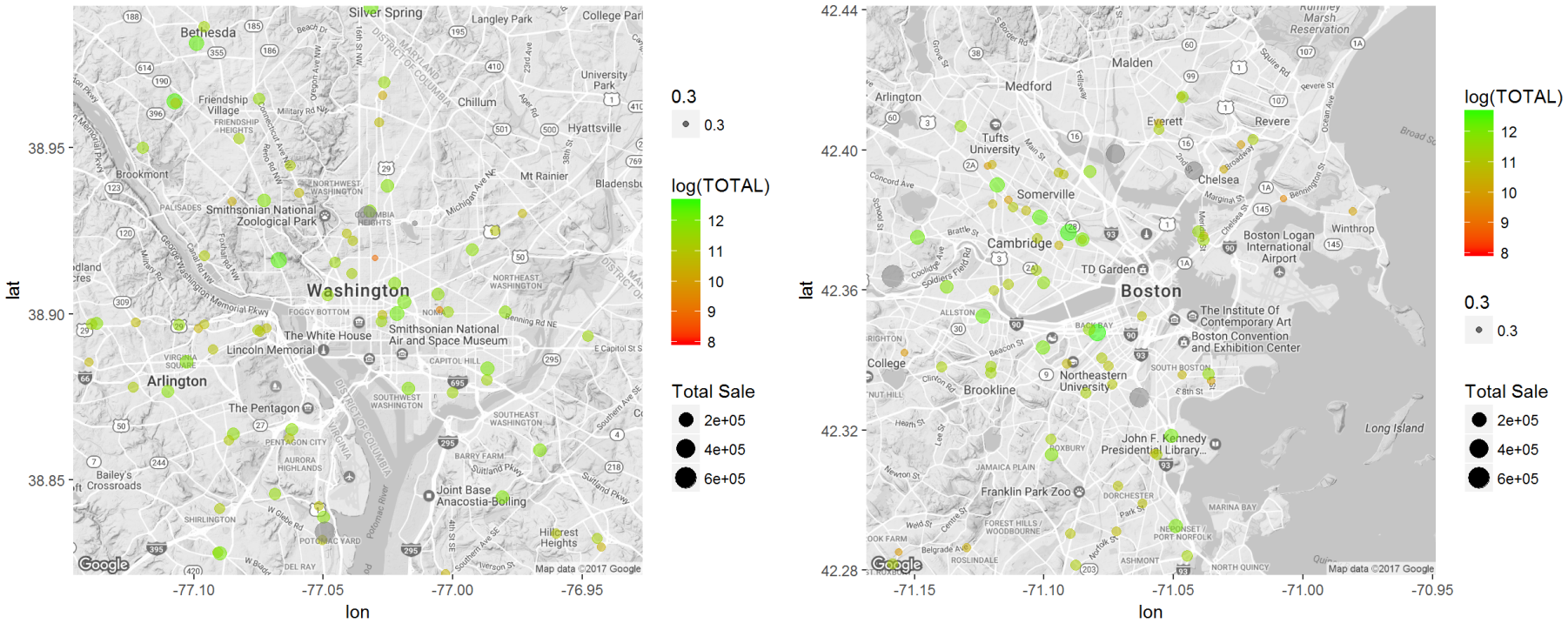
```
# concatenate the stores with their candy sale
store_candies = cbind(data$STORE,candies)
names(store_candies)[names(store_candies)=="data$STORE"] <- "STORE"

#aggregate the candy sales for each store type
stores <- store_candies %>%
group_by(STORE) %>%
  summarise_each(funs(mean))

stores = as.data.frame(stores)
rownames(stores)<- stores[,1]
stores_t = as.data.frame(t(stores)[-1,])
stores_t[is.na(stores_t)] <- 0
rownames(stores_t)<-c()


# create a matrix of the resulting dataframe
mat_stores = data.matrix(stores_t)
# create cosine similarity matrix
sim_stores = cosine(mat_stores,y=NULL)



# visualize the cosine similarity matrix
new.palette=colorRampPalette(c("black","red","yellow","white"),
                             space="rgb")

levelplot(sim_stores,col.regions=new.palette(20),
          scales=list(y=list(rot=0), x=list(rot=90)))
```

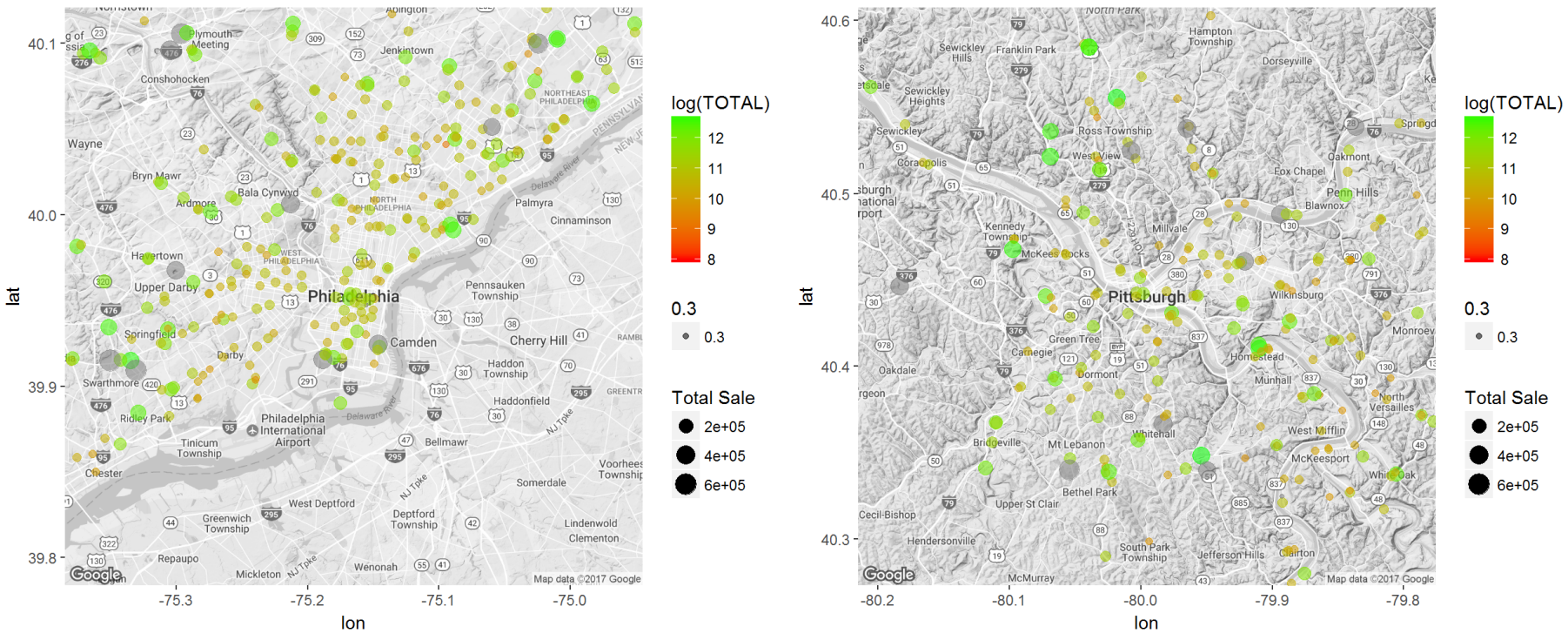


How different neighborhoods compare in terms of candy consumption?

Another interesting question to answer would be the geographical aspect of the markets. For each market we have the latitude and longitude data. We will use the ggmap() function to and assign colors to the stores to see which ones sell more. It looks like stores in the periphery are more successful in general.

```
# get  background maps for the four cities from google
phil <- get_map(location = "philadelphia",
                zoom = 11, source = "google", color = c("bw"))
bos  <- get_map(location = "boston",
                zoom = 12, source = "google", color = c("bw"))
dc  <- get_map(location = "washington dc",
               zoom = 12, source = "google", color = c("bw"))
pit  <- get_map(location = "pittsburgh",
                zoom = 11, source = "google", color = c("bw"))

p1<- ggmap(dc)+ geom_point(data = data,
     aes(x=LONG, y=LAT, colour = log(TOTAL), size =TOTAL , alpha = 0.3))+
  scale_colour_gradient(limits=c(8, 12.72), low="red", high="green")+
  scale_size_continuous (name = c("Total Sale"))

p2<- ggmap(bos)+ geom_point(data = data,
     aes(x=LONG, y=LAT, colour = log(TOTAL), size =TOTAL , alpha = 0.3))+
  scale_colour_gradient(limits=c(8, 12.73), low="red", high="green")+
  scale_size_continuous (name = c("Total Sale"))

p3<- ggmap(phil)+ geom_point(data = data,
     aes(x=LONG, y=LAT, colour = log(TOTAL), size =TOTAL, alpha = 0.3))+
  scale_colour_gradient(limits=c(8, 12.73), low="red", high="green")+
  scale_size_continuous (name = c("Total Sale"))

p4<- ggmap(pit)+ geom_point(data = data,
     aes(x=LONG, y=LAT, colour = log(TOTAL), size =TOTAL, alpha = 0.3))+
  scale_colour_gradient(limits=c(8, 12.73), low="red", high="green")+
  scale_size_continuous (name = c("Total Sale"))

grid.arrange(p1,p2,p3,p4,ncol = 2)
```

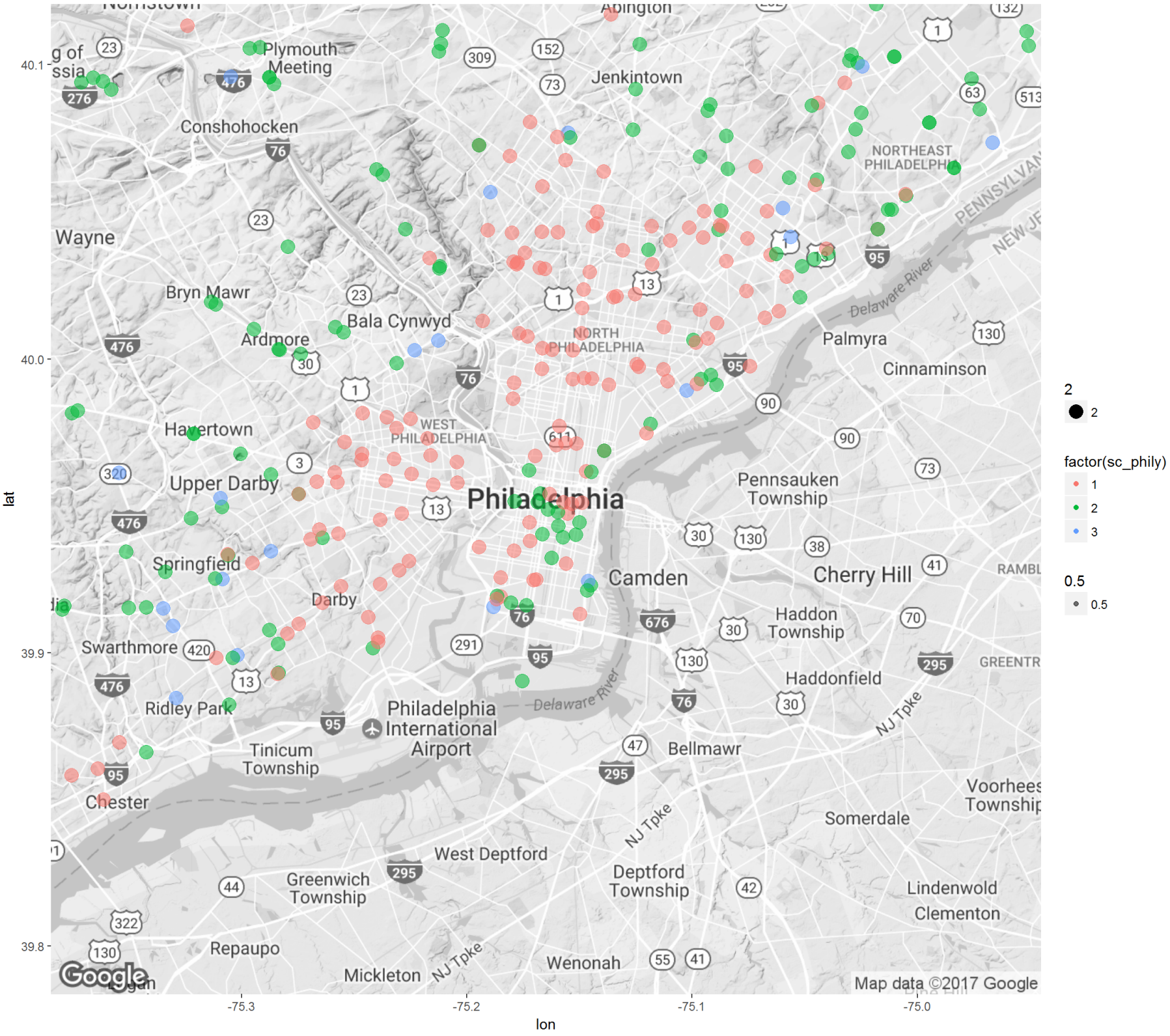




A more advanced way of going about the geographic aspect of the candy consumption is to find clusters of similar stores. That is, we first calculate a similarity matrix with the function that I previously define and then find clusters using spectral clustering. Every store in philadelphia will be compared to all other stores (pairwise) based on the candy sales for each store.

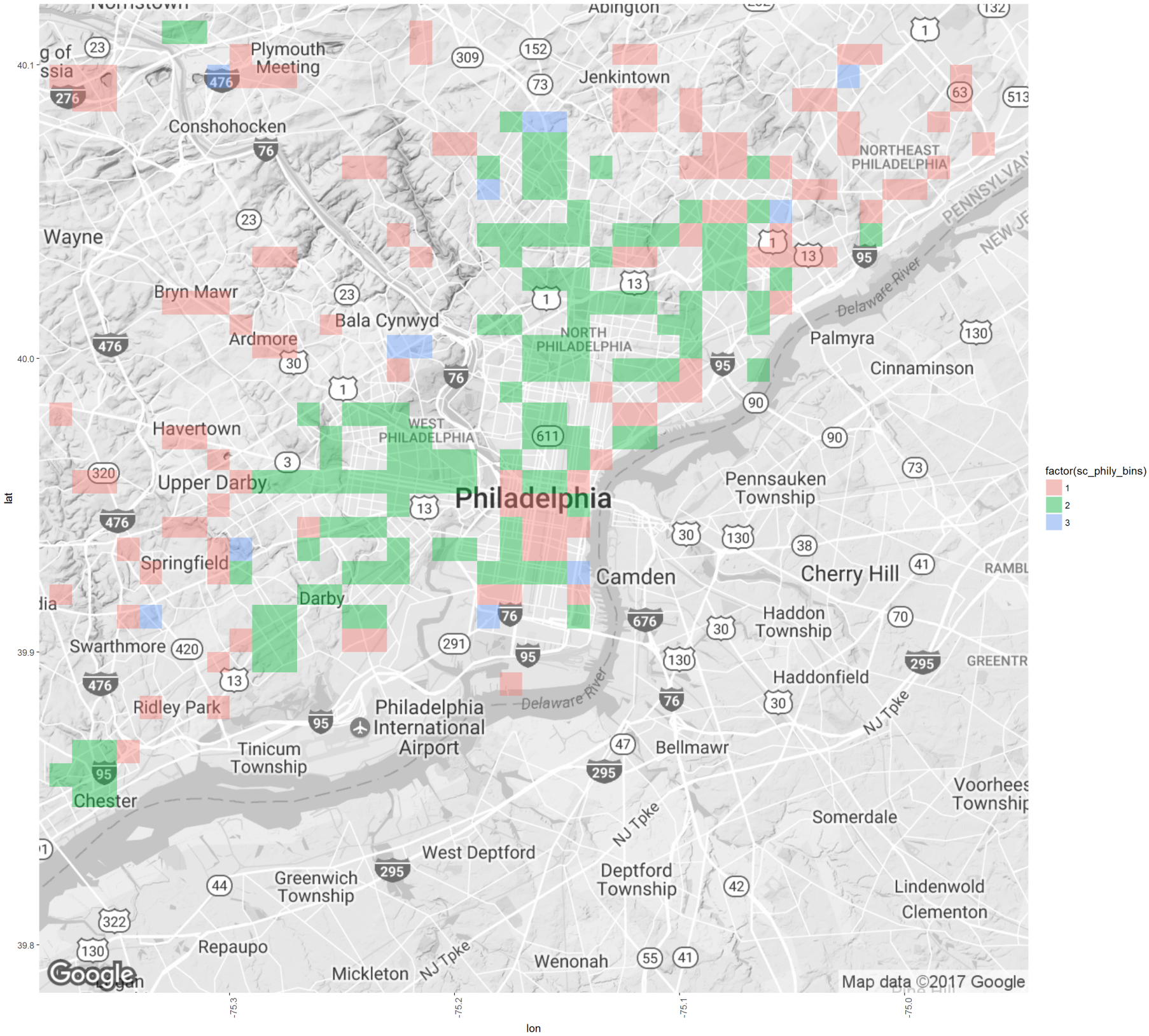
Doing so for Philadelphia, we can clearly see that there is a clear spatial pattern with downtown and suburbia being almost the same and the neighborhoods between them following another candy consumption patter. This shows that geography is correlated with candy taste.

```
#spectral clustering for Phily

candies_phily = data.matrix(t(candies[data$CITY == "Philadelphia",]))
sim_phily = cosine(candies_phily, y=NULL)
sc_phily<- spectralClustering(sim_phily, 3, type = 3)
phily_markets_SC = cbind(sc_phily,
                         data[data$CITY == "Philadelphia",c("LONG","LAT")])

ggmap(phil) +
  geom_point(data = phily_markets_SC,
             aes(x=LONG, y=LAT, colour = factor(sc_phily),alpha= 0.5, size = 2))
```



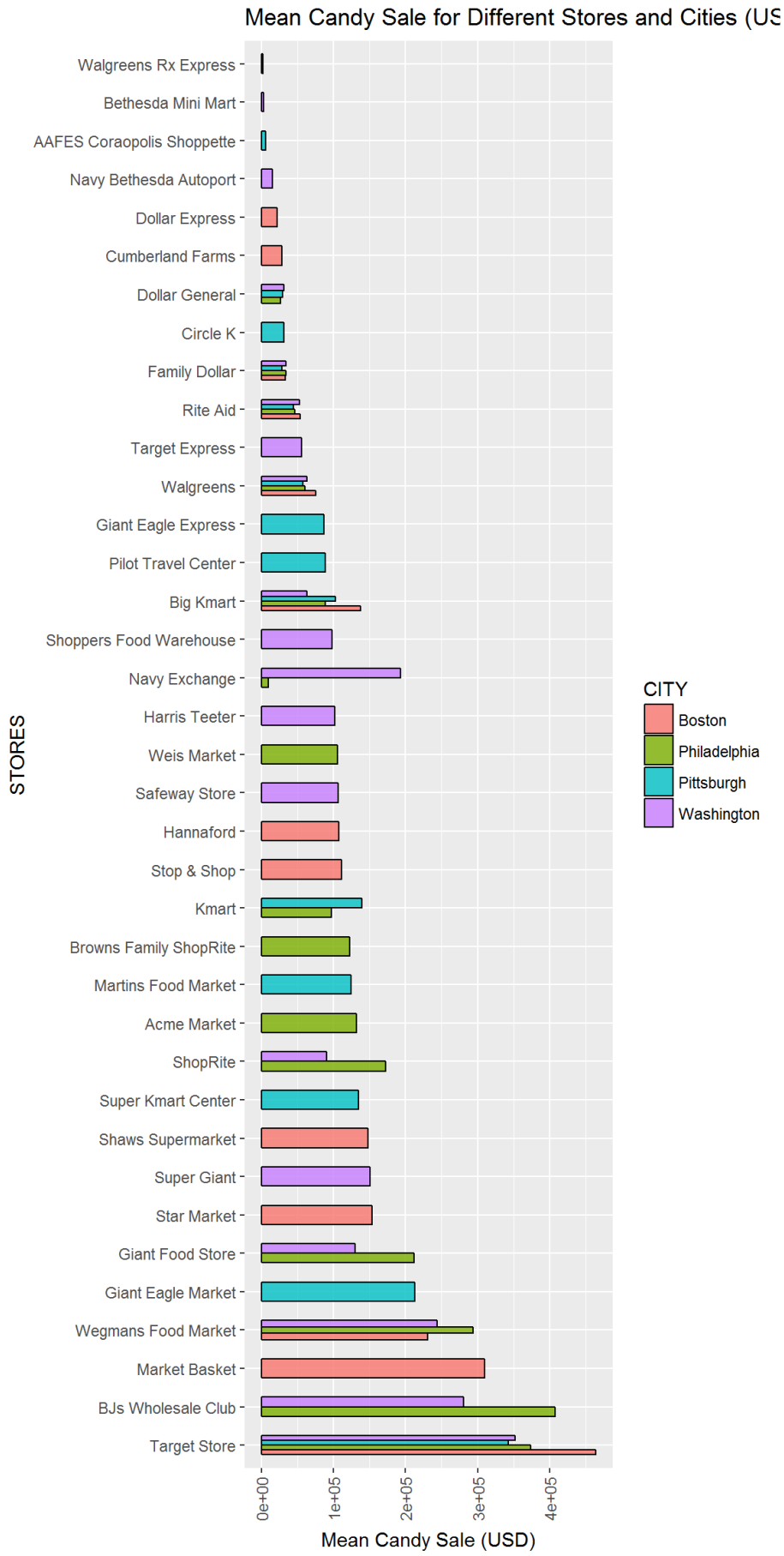
although the plot above is informative, if we intend to see the neighborhoods' behavior in terms of candy consumption, it would be a better idea to divide the city to a number of spatial bins, i.e. geographic areas with similar dimensions. After doing this, we can average teh candy consumption in each area and run the spectral clustering algorithm once again for these spacial bins.

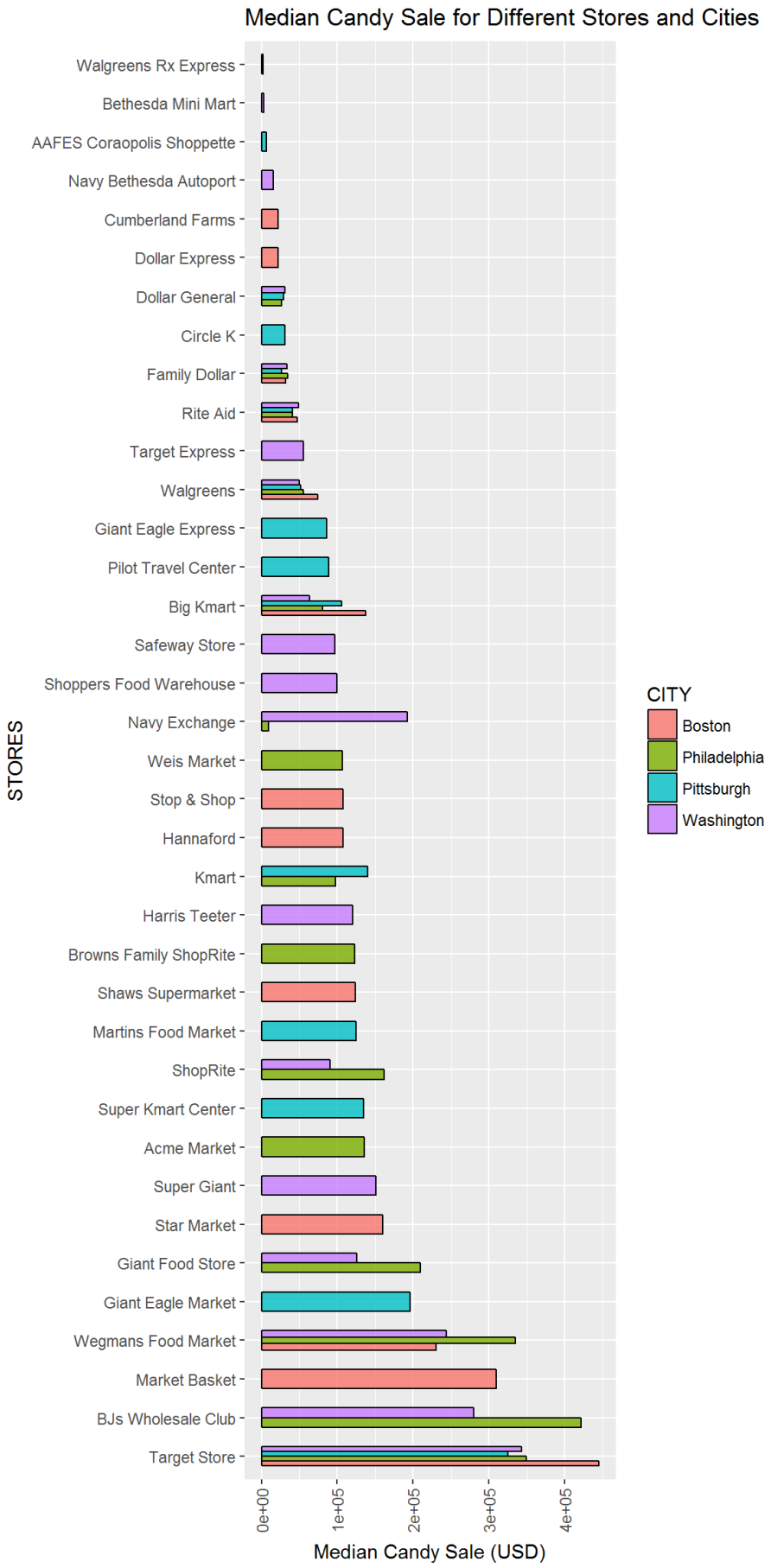
```
phily_data = cbind(candies[data$CITY == "Philadelphia",],
                   data[data$CITY == "Philadelphia",c("LONG","LAT")])

# divide the philly area to a 90*90 grid of squares
phily_data$lat_bins = cut(phily_data$LAT,breaks = 90)
phily_data$lon_bins = cut(phily_data$LONG,breaks = 90)

# The candy sale in each square will
# be averaged nd assigned to that square
phily_bins<- phily_data%>%
  group_by(lat_bins,lon_bins) %>%
  summarise_each(funs(mean))


#separate the candy rows and standardize
phily_bins_cands=t(apply(as.data.frame(phily_bins)[3:31], 1,
                         function(x)(x-min(x))/(max(x)-min(x))))

#create similarity matrix
phily_bins_cands_t = data.matrix(t(phily_bins_cands))
sim_phily_bins = cosine(phily_bins_cands_t, y=NULL)
sc_phily_bins<- spectralClustering(sim_phily_bins, 3, type = 3)


phily_clusters = data.frame(phily_bins,sc_phily_bins)

# extract the minimum and maximum lat and long for
#each square (so that we can draw them as rectangles later on)

maxlon = sapply(str_extract_all
                (as.character(phily_clusters[,2]),
                "\\d+\\.*\\d*"), "[[", 1)
minlon =  sapply(str_extract_all
                 (as.character(phily_clusters[,2]),
                 "\\d+\\.*\\d*"), "[[", 2)
maxlat = as.numeric(sapply(str_extract_all
                           (as.character(phily_clusters[,1])
                           , "\\d+\\.*\\d*"), "[[", 1))
minlat =  as.numeric(sapply(str_extract_all
                            (as.character(phily_clusters[,1]),
                            "\\d+\\.*\\d*"), "[[", 2))
minlon=  as.numeric(minlon)*-1
maxlon = as.numeric(maxlon)*-1

phily_clusters = data.frame(phily_clusters,minlat,maxlat,minlon,maxlon)

# draw the resulting rectangles from the previous step
ggmap(phil) +
  geom_rect(data=phily_clusters,
            mapping=aes(xmin=minlon, xmax=maxlon, ymin=minlat, ymax=maxlat,
            fill=factor(sc_phily_bins)), alpha=0.4,inherit.aes=FALSE)+
theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Final plots and summary

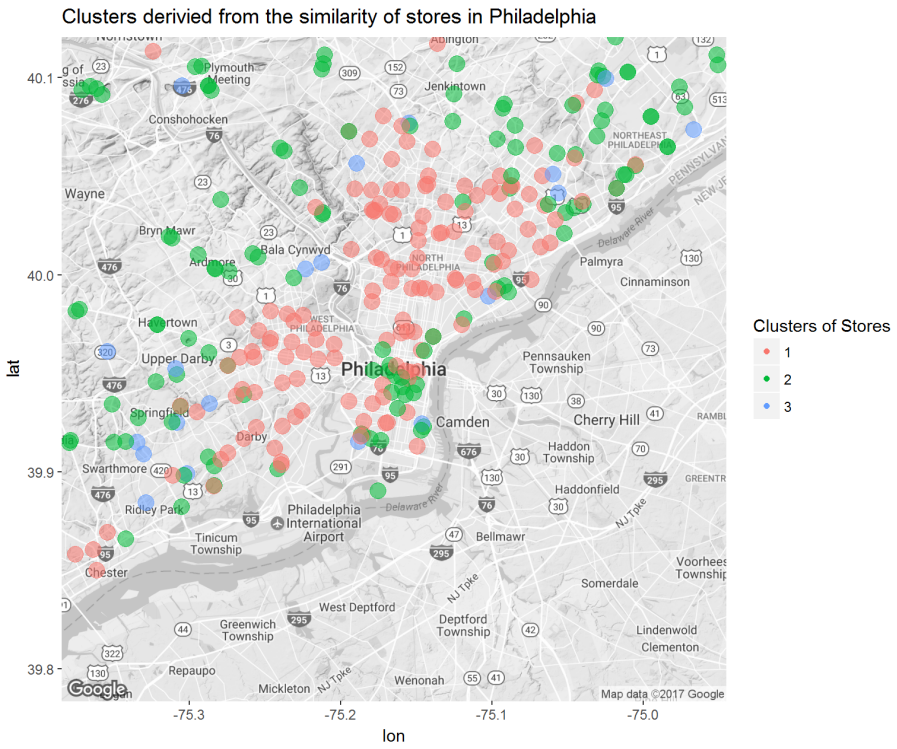
In this section I will select the three most informative plots that I hjave made through out the course of exploratory data analysis on the candy data. First, the bar plot for each market plotted below indicates that the target is highest on average in both median values and mean values for all 4 citiesExcept for Philadelphia, as explained earlier.

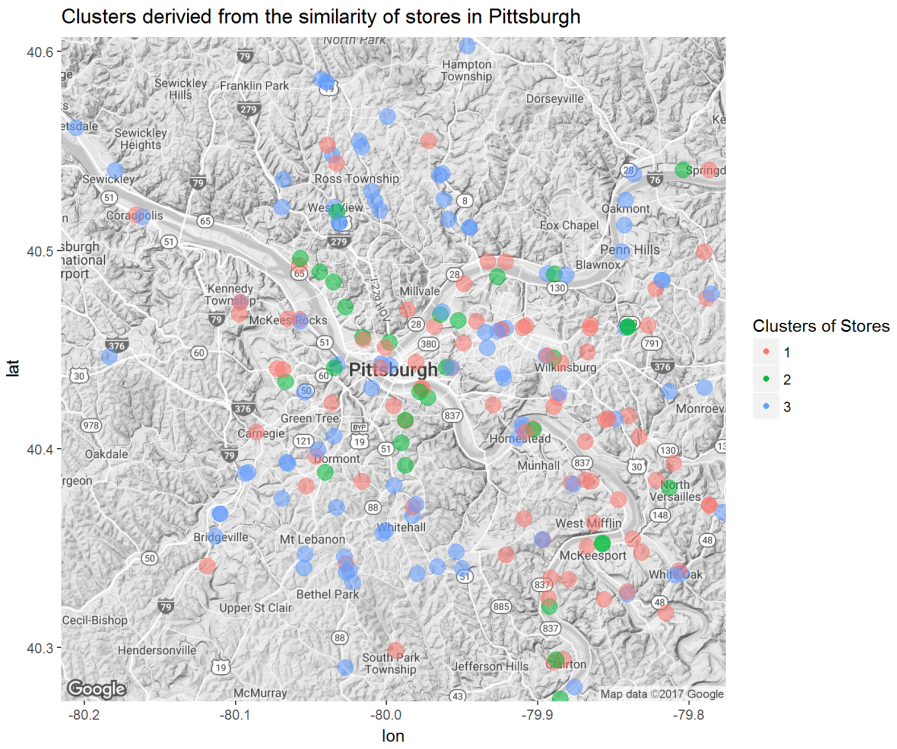


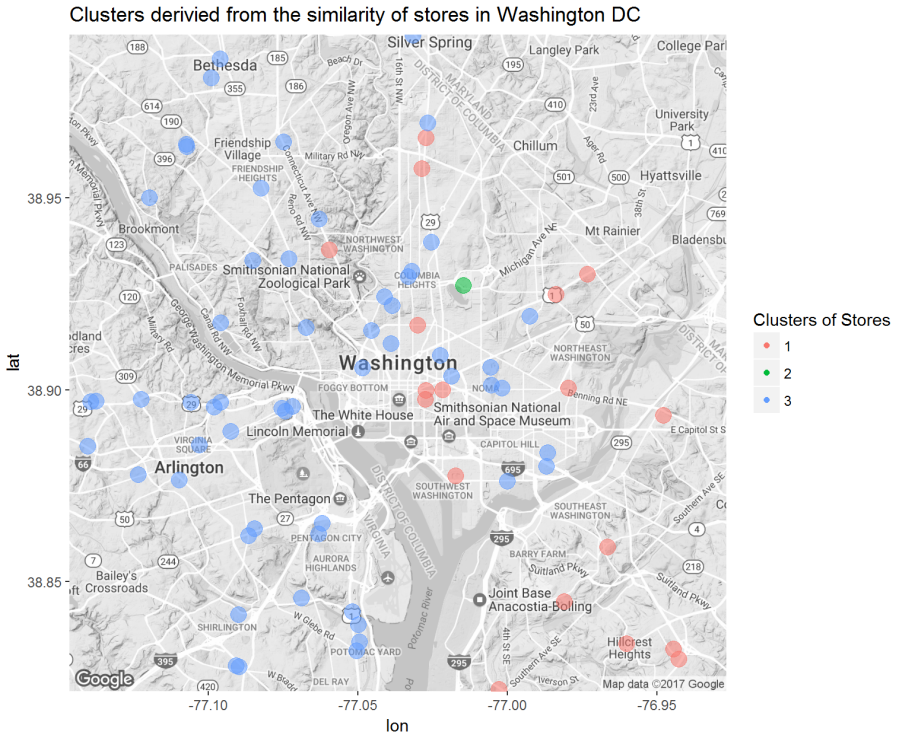


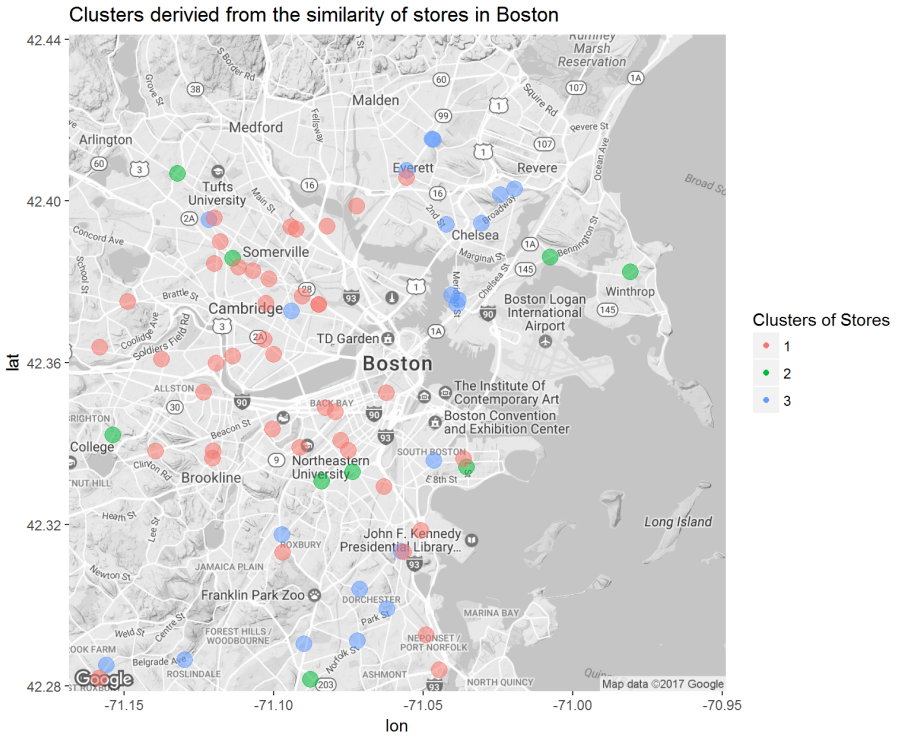
For the regression, we limit the x axis to see the bulk of the data more clearly. we assign node size to the points which corresponds with the median sales in stores. We can now clearly see that there is a positive linear association between the log population and log candy sales in zipcodes.

At last, we apply the spectral clustering method to all four cities to investigate patterns of chocolate use in the city. We can see that washington can be divided into two part, the east and west. Recall that washington is also racially divided in this manner. Philadelphia also has this clear spatial clusters. For Pittsburgh, it seems that the neighborhoods are not as distinct, however some clear green points can be seen closer to the center. Boston as well shows some similarity in candy consumption in south and west.









Reflection

Our data analysis revewled some interesting patterns in the candy consumption data set. We found out which markets are more likely to sell more candy, We calos learnt how different brands compare in different cities. More importantly, we found out about cities and neighborhoods in each city. We found that there are clear candy consumption patterns in the four cities that we focused on.

At the same time, thhere were a number of limitations that cannot be neglected. First, the number of data points for DC and Boston were significantly lower than the two other. This made it harder for a comprehensive comparison between the four cities. Also, There were a number of candies that were not included in the dataset that coule inform more about the tastes in different cities and neighborhoods and improve our clusterings. Future work can take the zip code demographic data into consideration and investigate the associations between different factors such as income, racial composition and etc.

Although the data was not large (i.e. 1253 entities and 37 variables) There were a numbe rof challenges associated with working with this dataset. First and foremost, the multitude of categorical variables (i.e. city, zip codes,stores) reuired one to create a large number of visualizations to understand the dynamics of the dataset. The second challenge was the geographic component of the data. Finding geographical patterns within the dataset, although was ultimately achieved, required a wide range of techniques including calculating cosine similarities, clustering, spatial bins, and a number of visualization techniques. My personal take from this data, was the interesting association that I found between the neighborhoods and the taste of their residents.