Chicago Community Area Health Statistics: PCA and Spatial Clustering Analysis

Robert Hand

3/25/2022

OVERVIEW.

Find needed data and shape file here: Download csv from first link, download the zip file "shapefile" under "export" at second link. https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Selected-public-health-in/iqnk-2tcu

https://data.
cityofchicago.org/
Facilities-Geographic-
Boundaries/
Boundaries-
Community-Areas-
current-/cauq-8yn6
READ DATA
LOADING STEP
FOR DATA
LOADING GUIDE.

This project used the Selected Public Health Statistics data from the City of Chicago Data Portal.

The data includes 27 measures of overall health of a wide variety for each of Chicago's 77 community areas.

The purpose of this project was to see what insights about these neighborhoods could be gleaned through principal component analysis, since there are a large number of variables which may have some high correlations.

PCA of 23 correlated scaled variables resulted in only two principal components accounting for 75% of all the variance in the data, with the first principal component alone accounting for 57%. A screeplot illustrates the sharp drop in the variance explained by components.

Since the first two components account for 75% alone, a biplot was constructed to see what insights could be gained from visualizing these. The first biplot shows some insights from the data reduction.

Component 1 can be visually seen as almost two elements, income and overall health and safety, determining the component which accounts for over half of all variance alone. These have opposite directions, indicating an inverse relationship between health and safety, and income, for a community area.

Component 2 to illustrate an inverse relationship between income and another grouping of housing, education, and birthrates, within a community area.

Neighborhood clusters seem to emerge in the graph. color coding the neighborhoods in these clusters and adding 95% confidence ellipses for the groups to the biplot gives a visual illustration of these three groupings.

To further visualize this, the groupings were plotted onto a map of the Chicago community areas by color code. This shows a clear geographic pattern to these groups.

Following packages are needed to run this file.

Data Loading. Need the selected public health indicates by community area data file from the city of Chicago data portal and the shape file for the 77 official Chicago community areas, also easily publicly available.

```
data <- read.csv('Public_Health_Statistics_-_Selected_public_health_indicators_by_Chicago_community_area

#insert actual full fil epath to the location of the shapefile where it says file path.
#The way I did this is
#1. extracting all the files from the zipped download file
#2. insert the file path to the one that is a ".shp"
# make sure to "extract all" files from the zip file to and insert the file that is ".shp" in to the st

map <- st_read("C:/Users/Robert/Documents/Grad School/github_projects/projects/geo_export_cceb8ce1-c228
```

```
## Reading layer ‘geo_export_cceb8ce1-c228-4e57-a01b-9cea710b9da7’ from data source ‘C:\Users\Robert\Do
##   using driver ‘ESRI Shapefile’
## Simple feature collection with 77 features and 9 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -87.94011 ymin: 41.64454 xmax: -87.52414 ymax: 42.02304
## Geodetic CRS:  WGS84(DD)
```

Data Cleaning

The data had 12 communities with missing data for the Gonorrhea case counts in males and females. According to the data descriptions file these communities reported fewer than five cases total. Therefore, setting to zero.

```
data %>% select(Community.Area.Name, Gonorrhea.in.Males) %>% filter(Gonorrhea.in.Males == '.') %>% gt()
```

Community.Area.Name	Gonorrhea.in.Males

Edison Park	.
Norwood Park	.
Jefferson Park	.
Forest Glen	.
North Park	.
Montclaire	.
Hegewisch	.
Archer Heights	.
McKinley Park	.
Clearing	.
Mount Greenwood	.
O'Hare	.

```
#12 communities with "." reported 5 or fewer cases.


data %>% select(Community.Area.Name, Gonorrhea.in.Females) %>% filter(is.na(Gonorrhea.in.Females))  %>%
```

Community.Area.Name	Gonorrhea.in.Females
Edison Park	NA
Norwood Park	NA
Jefferson Park	NA
Forest Glen	NA
North Park	NA
Dunning	NA
Montclaire	NA
Hegewisch	NA
Archer Heights	NA
Clearing	NA
Mount Greenwood	NA
O'Hare	NA

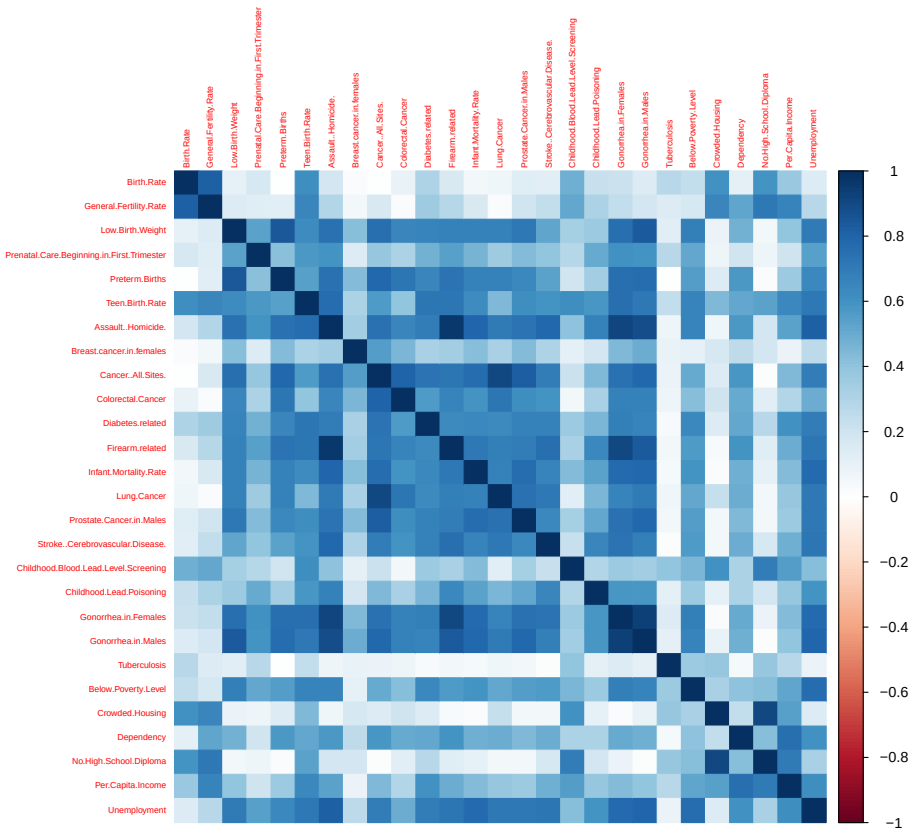
```
#12 communities with NA, reported 5 or fewer cases.

data %>% select(Community.Area.Name,Gonorrhea.in.Females,Gonorrhea.in.Males) %>% filter(is.na(Gonorrhea
```

Community.Area.Name	Gonorrhea.in.Females	Gonorrhea.in.Males
Edison Park	NA	.
Norwood Park	NA	.
Jefferson Park	NA	.
Forest Glen	NA	.
North Park	NA	.
Dunning	NA	70.1
Montclaire	NA	.
Hegewisch	NA	.
Archer Heights	NA	.
McKinley Park	141.4	.
Clearing	NA	.
Mount Greenwood	NA	.
O'Hare	NA	.

```
#fewer than five is for same community areas for males and females,
#except for Dunning and McKinley Park.

data[data == '.'] <- 0 #setting "missing" to zero since these mean fewer than five cases.
data[is.na(data)] <- 0
data <- data %>% mutate(Gonorrhea.in.Males = as.numeric(Gonorrhea.in.Males))
#making numeric. Was not when read in because of the '.' for missing data.
data <- data %>% mutate(Per.Capita.Income = as.numeric(Per.Capita.Income))
#read in as integer, just wanted to also make it numeric.
```

EDA. Now that the data is prepped for analysis, some initial assessment of the variable correlations. I expect there may be some high correlations between the variables.

```
#correlations for all

cor <- cor(data[ ,3:29])


#visualize

par(mar=(c(5,5,5,5)))

corrplot(abs(cor), method = "color", tl.cex=.5)
```



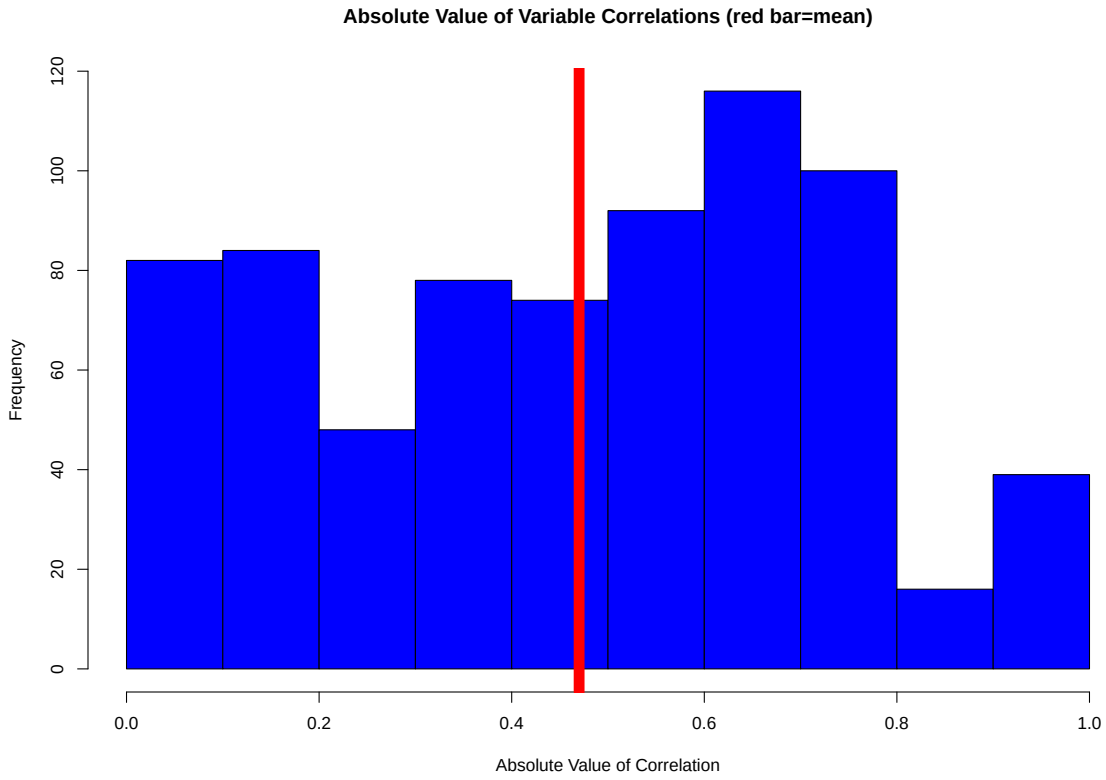

```
#took absolute value just to make the overall color scheme easier to interpret in terms of magnitude of

#Just looking at the color scheme of this plot overall, you can see there are a number of large correla


#doing absolute values just to more easily look at magnitude.

mean(abs(cor))
```

```
## [1] 0.4695837
```

```r
#mean correlation of 0.47.
median(abs(cor))
```

```
## [1] 0.4968592
```

```r
#median of 0.49.
hist(abs(cor), main = "Absolute Value of Variable Correlations (red bar=mean)", xlab = "Absolute Value
abline(v=0.47,col="red",lwd=10)
```




```r
#there are quite a few large ones.

cor_reduced <- cor
#check just the correlations with absolute values above 0.7, strong correlations.
cor_reduced[abs(cor_reduced) < .7 | cor==1] <- NA #set to NA if not in that range.
cor_reduced <- as.data.frame(as.table(cor_reduced)) #make into a table format.
cor_reduced <- cor_reduced %>% filter(!is.na(cor_reduced$Freq)) #drop the NAs.
cor_reduced <- cor_reduced[!duplicated(cor_reduced$Freq),] #drop duplicates - showing same correlations
```

```r
colnames(cor_reduced) <- c('Variable_1', 'Variable_2', 'Correlation') #update colnames.

dim(cor_reduced) #have 64 correlations after removing double counted duplicates and self-correlations t
```

```
## [1] 64  3
```

```r
#counts of variables with correlations over .7
cor_reduced %>% select(Variable_1) %>% group_by(Variable_1) %>% tally(sort = T) %>% kable()
```

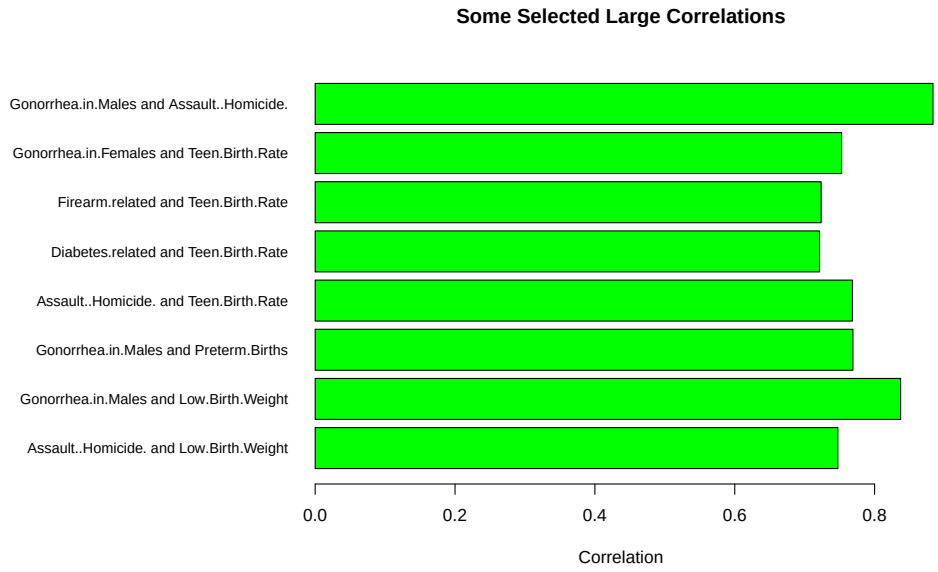
Variable_1	n
Unemployment	11
Gonorrhea.in.Males	10
Gonorrhea.in.Females	9
Prostate.Cancer.in.Males	5
Firearm.related	4
Assault..Homicide.	3
Cancer..All.Sites.	3
Infant.Mortality.Rate	3
Lung.Cancer	3
Stroke..Cerebrovascular.Disease.	3
Colorectal.Cancer	2
Diabetes.related	2
No.High.School.Diploma	2
Per.Capita.Income	2
General.Fertility.Rate	1
Preterm.Births	1

```r
cor_reduced %>% select(Variable_2) %>% group_by(Variable_2) %>% tally(sort = T) %>% kable()
```

Variable_2	n
Assault..Homicide.	9
Cancer..All.Sites.	8

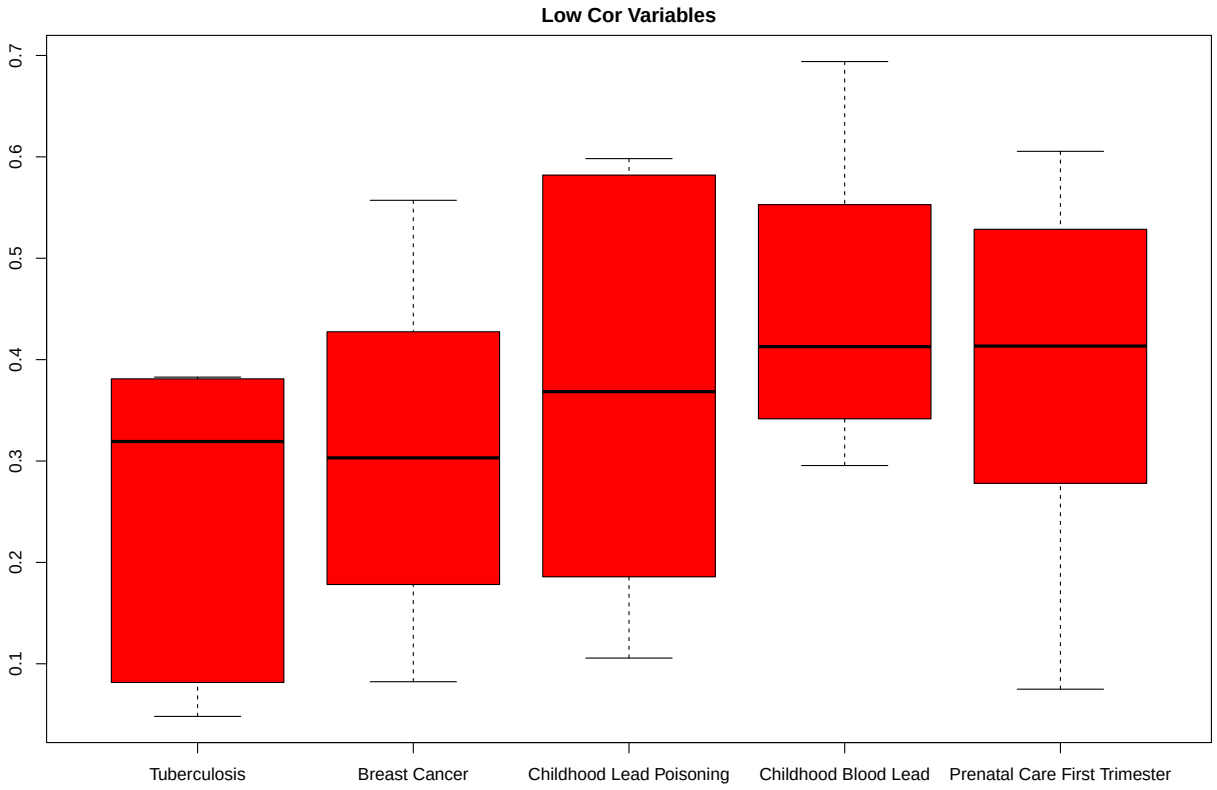
Variable_2	n
Low.Birth.Weight	7
Preterm.Births	6
Teen.Birth.Rate	6
Firearm.related	5
Infant.Mortality.Rate	4
Lung.Cancer	4
Prostate.Cancer.in.Males	3
Stroke..Cerebrovascular.Disease.	2
Gonorrhea.in.Females	2
Birth.Rate	1
General.Fertility.Rate	1
Colorectal.Cancer	1
Gonorrhea.in.Males	1
Below.Poverty.Level	1
Crowded.Housing	1
Dependency	1
No.High.School.Diploma	1

```
#plot of some more interesting ones.
par(mar = c(15,15,5,15))
barplot(abs(cor_reduced$Correlation[c(4,8,15,16,17,18,19,29)]), main = "Some Selected Large Correlations
```




```
#interestingly the above .7s included at least one instance of every one of the variables except for
#childhood blood lead level, childhood lead poisoning, tuberculosis, prenatal care beginning in first t
#breast cancer in females.

#visualize overall correlation level for these ones:
cor2 <- cor
cor2 <- as.data.frame(as.table(cor2))
cor2 <- cor2[!duplicated(cor2$Freq),]
tb <- cor2 %>% filter(Var2=="Tuberculosis")
bc <- cor2 %>% filter(Var2=="Breast.cancer.in.females")
cl <- cor2 %>% filter(Var2=="Childhood.Lead.Poisoning")
cb <- cor2 %>% filter(Var2=="Childhood.Blood.Lead.Level.Screening")
pn <- cor2 %>% filter(Var2=="Prenatal.Care.Beginning.in.First.Trimester")
par(mar = c(4,2,2,2))
boxplot(abs(tb$Freq),abs(bc$Freq),abs(cl$Freq),abs(cb$Freq),abs(pn$Freq),names = c("Tuberculosis","Breas
```

```
## Warning in (function (z, notch = FALSE, width = NULL, varwidth = FALSE, :
## Duplicated argument ylab = "Correlation" is disregarded
```




```
#they are pretty low as a whole, so  going to leave them out of the PCA and focus on the other variable
```

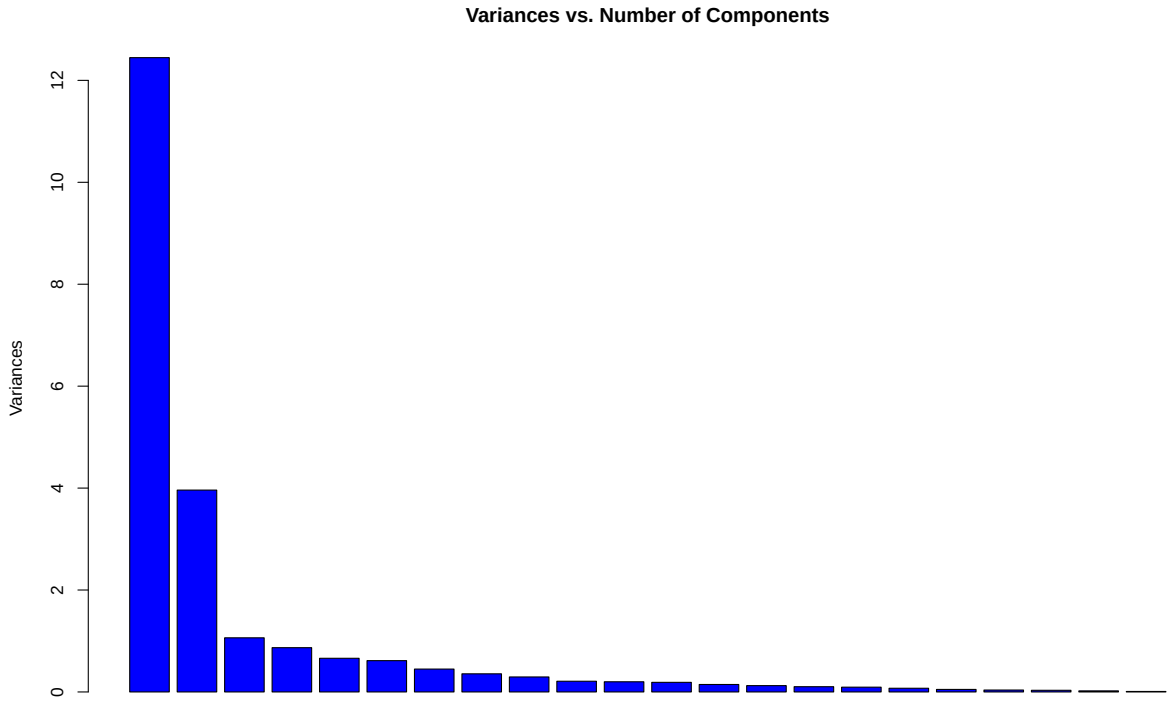
Principal Component Analysis

```
pca <- prcomp(data[,c(3:5,7:9,11:18,21:22,24:29)],scale. = TRUE)
summary(pca)
```

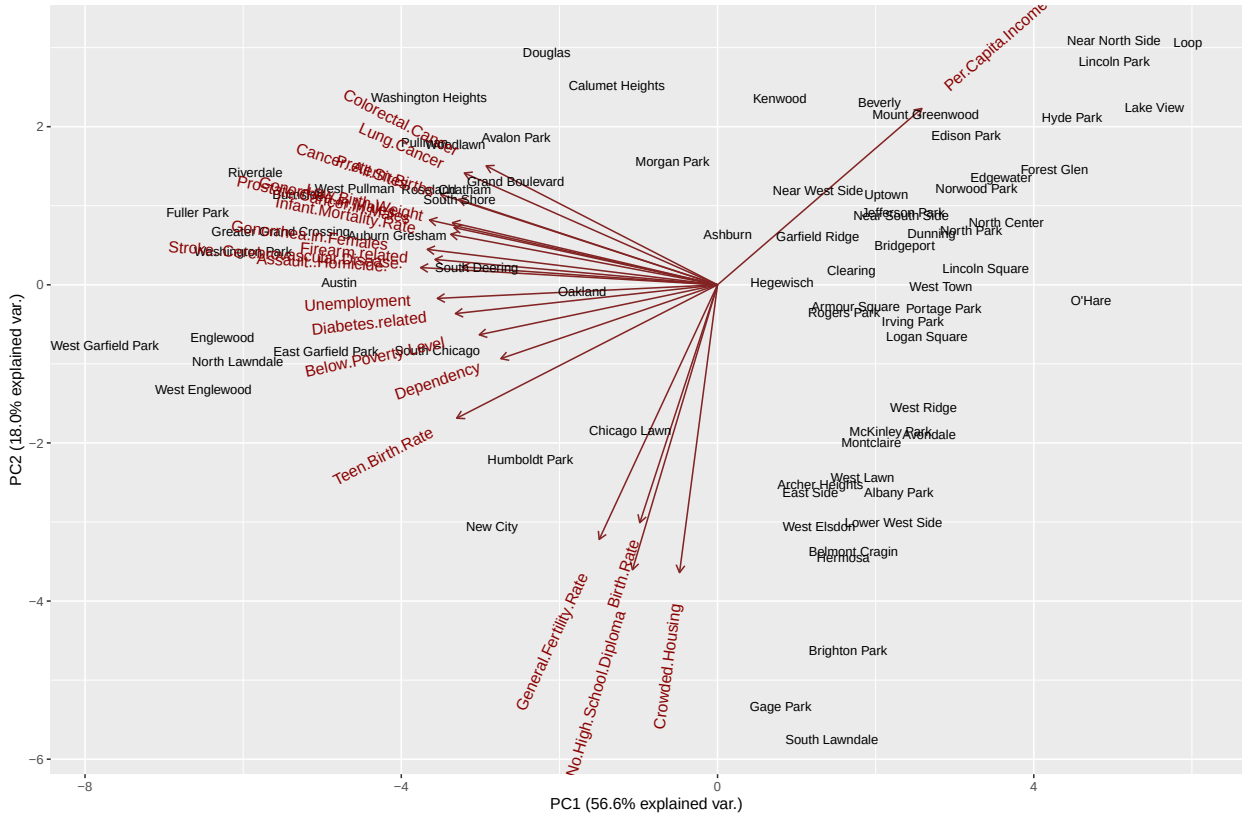
```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.5279 1.9902 1.03068 0.93200 0.81282 0.78370 0.67002
## Proportion of Variance 0.5657 0.1800 0.04829 0.03948 0.03003 0.02792 0.02041
## Cumulative Proportion  0.5657 0.7458 0.79404 0.83353 0.86356 0.89147 0.91188
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.59639 0.54235 0.45856 0.44671 0.43410 0.38280 0.35240
```

```
## Proportion of Variance 0.01617 0.01337 0.00956 0.00907 0.00857 0.00666 0.00564
## Cumulative Proportion  0.92805 0.94142 0.95097 0.96005 0.96861 0.97527 0.98092
##                           PC15    PC16   PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.32068 0.30650 0.2695 0.22291 0.19402 0.18070 0.14450
## Proportion of Variance 0.00467 0.00427 0.0033 0.00226 0.00171 0.00148 0.00095
## Cumulative Proportion  0.98559 0.98986 0.9932 0.99542 0.99713 0.99862 0.99956
##                           PC22
## Standard deviation     0.09784
## Proportion of Variance 0.00044
## Cumulative Proportion  1.00000
```

```r
#first two components alone explain about 75% of the variance.

#One on its own accounts for about 57%.

screeplot(pca, npcs = 22,col="blue",main = "Variances vs. Number of Components")
```

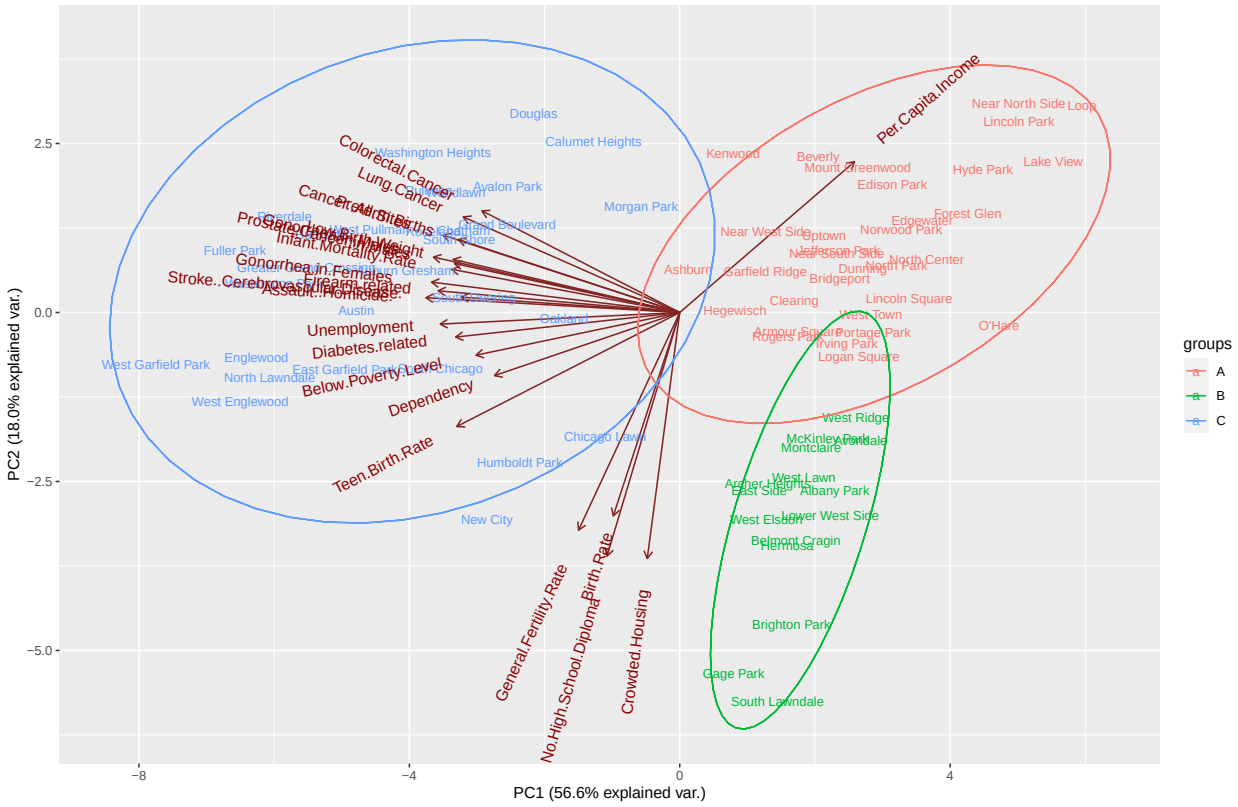


```
#scree plot visualization.

#See a really steep drop from one to two and then from two onward.


#biplot since the first two are very large to visualize this.

ggbiplot(pca, labels = data$Community.Area.Name, varname.size = 4, obs.scale = 1, varname.adjust = 1.5)
```




```
#Can see the first component showing health and safety together versus income.


#second component mainly showing housing, birthrate and education versus income.


#can see some potential grouping of neighborhoods.


#creating codings. Was tedious so put in C for a bunch of them, once they
#had color it was easier to see, and used case when to update to appropriate labels
#as needed.
```

```
z <- c("A","B","A","A","A","A","A","A","A","A","A","A","A","A","B","A","A","A","B","B","B","B","A","C","A",
groupings <- cbind(data$Community.Area.Name,z)
groupings <- as.data.frame(groupings)
groupings <- groupings %>% mutate(
  z = case_when(
              groupings$V1 %in% c("Austin","North Lawndale","Humboldt Park")~"C",
              groupings$V1 %in% c("Kenwood","Beverly","Mount Greenwood","Hyde Park","Edgewater","Near
              groupings$V1 %in% c("West Lawn","McKinley Park","Archer Heights","East Side","West Elsd
              TRUE ~ z
              )
)

ggbiplot(pca, labels = data$Community.Area.Name, varname.size = 4, obs.scale = 1, varname.adjust = 1.5,
```




```
#with color coding and 95% probability ellipses for each grouping,
#you can see three pretty well delineated neighborhood groupings.
```

Visualization on Chicago community areas map

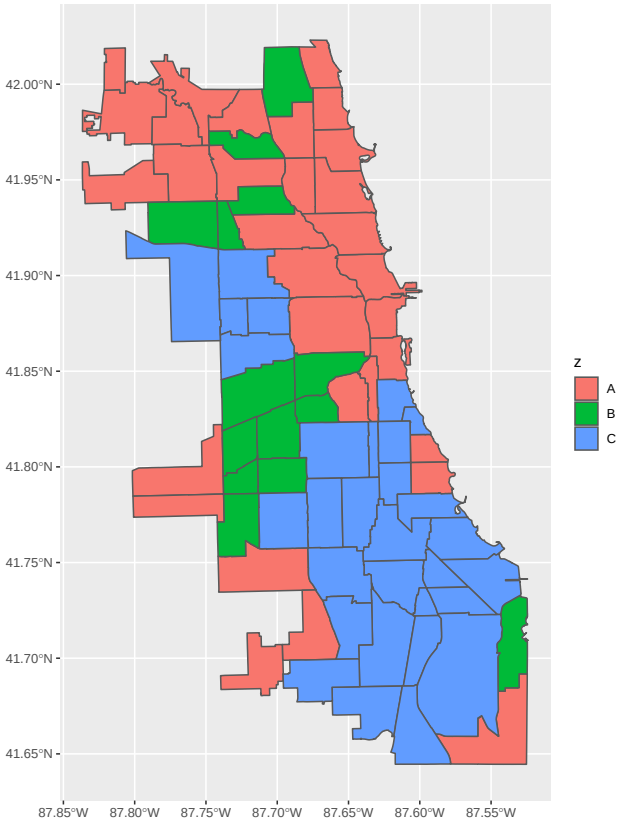
```
#mapping these color codes onto a Chicago neighborhood map.
#shows this aligns with clear geographic clustering,
#with a few exceptions, most of which fall in multiple ellipses, such as Hegeswich.

str(map)
```

```
## Classes 'sf' and 'data.frame':   77 obs. of  10 variables:
##  $ area      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ area_num_1: chr  "35" "36" "37" "38" ...
##  $ area_numbe: chr  "35" "36" "37" "38" ...
##  $ comarea   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ comarea_id: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ community : chr  "DOUGLAS" "OAKLAND" "FULLER PARK" "GRAND BOULEVARD" ...
##  $ perimeter : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ shape_area: num  46004621 16913961 19916705 48492503 29071742 ...
##  $ shape_len : num  31027 19566 25339 28197 23325 ...
##  $ geometry  :sfc_MULTIPOLYGON of length 77; first list element: List of 1
##   ..$ :List of 1
##   .. ..$ : num [1:352, 1:2] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##   ..- attr(*, "class")= chr [1:3] "XY" "MULTIPOLYGON" "sfg"
##  - attr(*, "sf_column")= chr "geometry"
##  - attr(*, "agr")= Factor w/ 3 levels "constant","aggregate",..: NA NA NA NA NA NA NA NA NA
##   ..- attr(*, "names")= chr [1:9] "area" "area_num_1" "area_numbe" "comarea" ...
```

```
groupings <- groupings %>% mutate(community = toupper(V1))
together <- inner_join(map,groupings)
```

```
## Joining, by = "community"
```

```
ggplot(together) + geom_sf(aes(fill = z))
```


42.00°N
41.95°N
41.90°N
41.85°N
41.80°N
41.75°N
41.70°N
41.65°N
87.85°W
87.80°W
87.75°W
87.70°W
87.65°W
87.60°W
87.55°W
z
A
B
C