Exploratory Data Analysis on the CO_County_x_2014_2018.xlxs

David Martinez (davelovesdata@gmail.com)

May 7, 2019

Purpose

To perform Exploratory Data Analysis on a dataset containing four years of Colorado county level medical and recreational marijuana sales as well as state revenue (taxes) collected. Two excel workbook sheets are imported to R and merged together, one for sales, the other for taxes.

1. Sales file description (CO_County_Sales_2014_2018.xlsx)

The sales files contains not only county level medical and recreational sales by year, but also population information and location information (State, County, Latitude, Longitude, Region). Additionally, medical and recreational sales for each county were applied against county population to determine an average of sales per county citizen for both medical and recreational sales.

Dataset fields: State - Currently only "COLORADO" County - Colorado County Name (e.g., "Adams" or "Yuma") Latitude - Latitude of County center Longitude - Longitude of County Center Region - An arbitrary assignment I made to quarter the state into geographic quadrants. Year - Collection Year Population - Estimated population between census reporting periods Med_Sales - County level sales of Medical Marijuana (see value explanation below) Rec_Sales - County level sales of Recreational Marijuana (see value explanation below) med_sales_per_citizen - a calculated value determined by dividing the "Med_Sales" value by the "Population" value. rec_sales_pre_citizen - a calculated value determined by dividing the "Rec_Sales" value by the "Population" value.

Med_Sales, Rec_sales, and the two calculated values have three possible values: **0** = No Sales of legal Marijuana occurred in that county. The original source material did not include counties that had no sales. This information was added to show a full statewide picture as well as county adoption over time. **NR** = Not releasable due to confidentiality requirements. The sum of all NR counties ("Not Reported" in the 'County' column) are captured as the last line for each year. **x** = A positive number representing sales at the dollar level.

2. Taxes file description (CO_County_Taxes_2014_2018.xlsx)

The taxes file contains taxes collected per county in three columns: Medical Sales Tax (2.9%), Retail Sales Tax (2.9%), Retail Marijuana Special Sales Tax.

Dataset fields: County - Colorado County Name (e.g., "Adams" or "Yuma") Year - Collection Year Medical Sales Tax (2.9%) - Sales tax applied to medical marijuana only. This is the only state tax paid. Retail Sales Tax (2.9%) - Sales tax applied to retail marijuana. Starting in 2018, this tax was no longer collected. Retail Marijuana Special Sales Tax - an additional tax on retail marijuana sales.

Medical Sales Tax (2.9%), Retail Sales Tax (2.9%), Retail Marijuana Special Sales Tax have three possible values: **0** = No taxes from legal Marijuana occurred in that county. The original source material did not include counties that had no tax information. This information was added to show a full statewide picture as well as county adoption over time. **NR** = Not releasable due to confidentiality requirements. The sum of all NR counties ("Not Reported" in the 'County' column) are captured as the last line for each year. **x** = A number representing taxes at the dollar level. Negative values indicate previous months overpayment of taxes being returned.

Data Collection and Merging steps

Dependencies

If needed, these packages can be installed using the install.packages() function

```
library("readxl")
library("formattable")
library("tidyverse")
library("tidyr")
library("ggplot2")
library("ggrepel")
```

Collect and Merge data

The two files are read into tibbles and then merged into a dataframe. Data is subsetted to remove 2018 values. A loop is performed to convert the sales and tax features to numeric and currency.

```
#gather sales and tax data into tibbles
sales_mj <- read_xlsx("CO_County_Sales_2014_2018.xlsx", sheet = "aggregate", range = NULL, col_names = T
taxes_mj <- read_xlsx("CO_County_Taxes_2014_2018.xlsx", sheet = "aggregate", range = NULL, col_names = T

#merge the two tibbles - {base} merge returns a dataframe
CCMDs <- merge(sales_mj, taxes_mj)

#remove 2018 values until both spreadsheets are fully populated
CCMDs <- subset(CCMDs, Year < "2018")

#create a list of column names for columns 8 through 14 - these are the columns related to sales and ta
cashcol <- colnames(CCMDs[8:14])

#loop to convert sales/tax columns to numeric/currency - this will introduce NAs for each of the 7 colu
for (i in cashcol) {
  CCMDs[[i]] <- as.numeric(CCMDs[[i]])   #character to numeric
  CCMDs[[i]] <- currency(CCMDs[[i]], digits = 0L) #numeric but with currency symbology
}
```

```
## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion
```

```r
#clean up unneeded files
rm(cashcol, i)
rm(sales_mj, taxes_mj)

#write dataframe to disk
write_excel_csv(CCMDs, "CCMDs.csv", na = "NA", append = FALSE)
```

Exploratory Data Analysis

```r
summary(CCMDs)
```

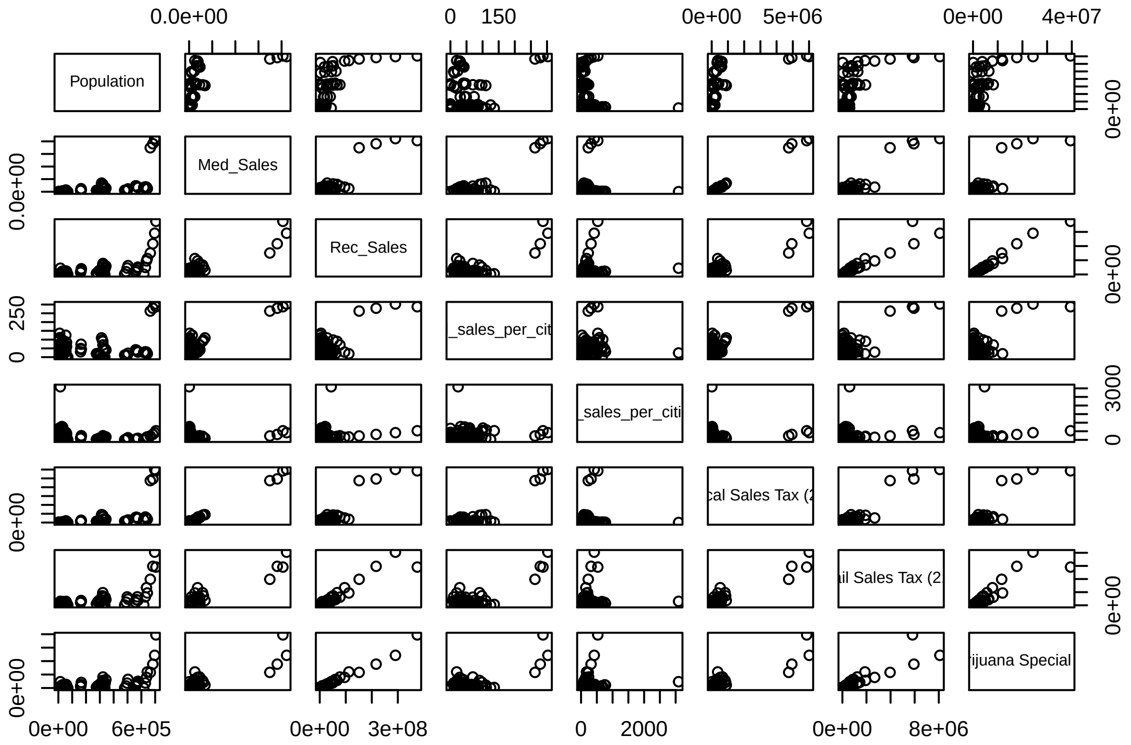
```
##     County               Year          State              Latitude    
##  Length:260         Min.   :2014   Length:260         Min.   :37.20  
##  Class :character   1st Qu.:2015   Class :character   1st Qu.:38.01  
##  Mode  :character   Median :2016   Mode  :character   Median :39.07  
##                     Mean   :2016                      Mean   :38.98  
##                     3rd Qu.:2016                      3rd Qu.:39.86  
##                     Max.   :2017                      Max.   :40.87  
##                                                       NA's   :4      
##    Longitude          Region            Population       Med_Sales        
##  Min.   :-108.6   Length:260         Min.   :   689   Min.   :        0  
##  1st Qu.:-106.9   Class :character   1st Qu.:  5719   1st Qu.:        0  
##  Median :-105.5   Mode  :character   Median : 14366   Median :        0  
##  Mean   :-105.4                      Mean   : 85681   Mean   :  8349120  
##  3rd Qu.:-103.8                      3rd Qu.: 42846   3rd Qu.:  2726766  
##  Max.   :-102.3                      Max.   :705651   Max.   :210860875  
##  NA's   :4                           NA's   :4        NA's   :61         
##    Rec_Sales          med_sales_per_citizen rec_sales_per_citizen
##  Min.   :        0   Min.   :  0.00        Min.   :   0.0       
##  1st Qu.:        0   1st Qu.:  0.00        1st Qu.:   0.0       
##  Median :        0   Median :  0.00        Median :   0.0       
##  Mean   : 12479031   Mean   : 32.53        Mean   : 149.6       
##  3rd Qu.:  7869992   3rd Qu.: 43.52        3rd Qu.: 206.1       
##  Max.   :374673239   Max.   :909.01        Max.   :3093.2       
##  NA's   :34          NA's   :65            NA's   :38           
##  Medical Sales Tax (2.9%) Retail Sales Tax (2.9%)
##  Min.   :      0          Min.   :      0        
##  1st Qu.:      0          1st Qu.:      0        
##  Median :      0          Median :      0        
##  Mean   : 230108          Mean   : 289108        
##  3rd Qu.:  81149          3rd Qu.: 188086        
##  Max.   :6011329          Max.   :8092488        
##  NA's   :62               NA's   :36             
##  Retail Marijuana Special Sales Tax
##  Min.   :       0                  
##  1st Qu.:       0                  
##  Median :       0                  
##  Mean   : 1135213                  
##  3rd Qu.:  659683                  
##  Max.   :39493849                  
##  NA's   :34                        
```

There are 260 observation with 14 variables. There are 346 NA's (~10%) that will need to be addressed.

```
#count the NAs
sum(is.na(CCMDs))
```

```
## [1] 346
```

```
#remove the NAs
CCMDs <- na.omit(CCMDs)

plot(CCMDs[,7:14])
```

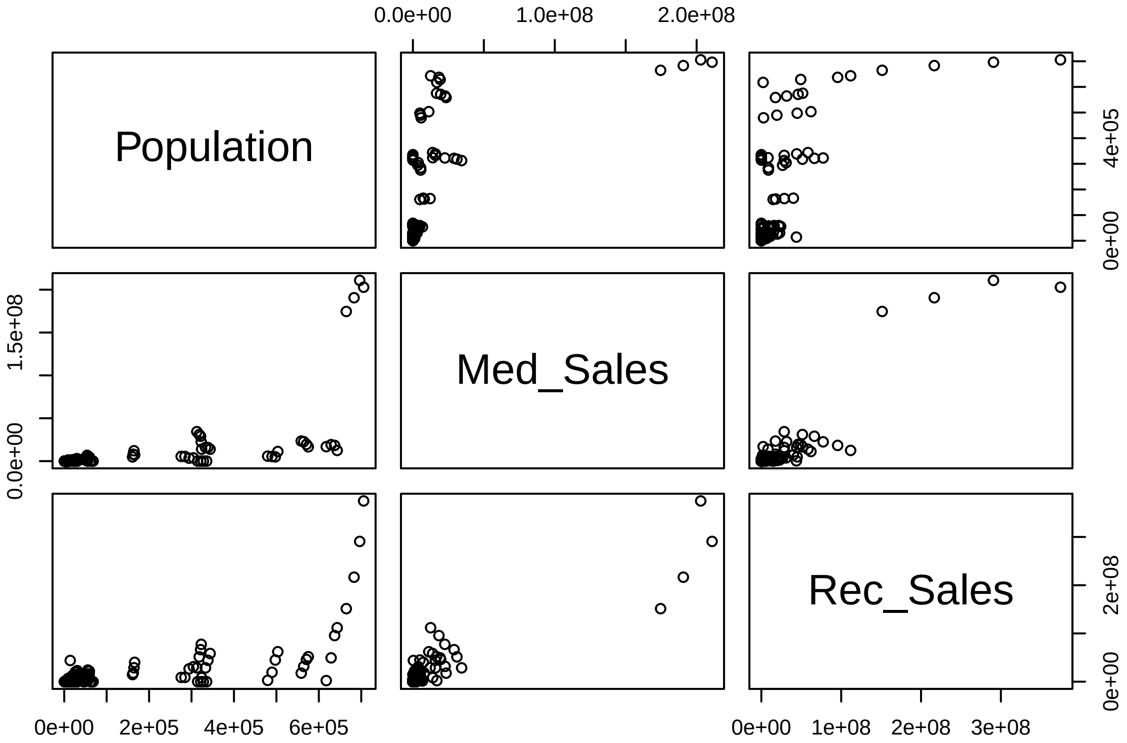


Right from the start, the data are redundant. Specifically, the 'med_sales_per_citizen' and 'rec_sales_per_citizen' variables which are calculated by dividing the county sales by the county population. Similarly, the tax data is also a function of the sales data. For now, I'm going to ignore that data.

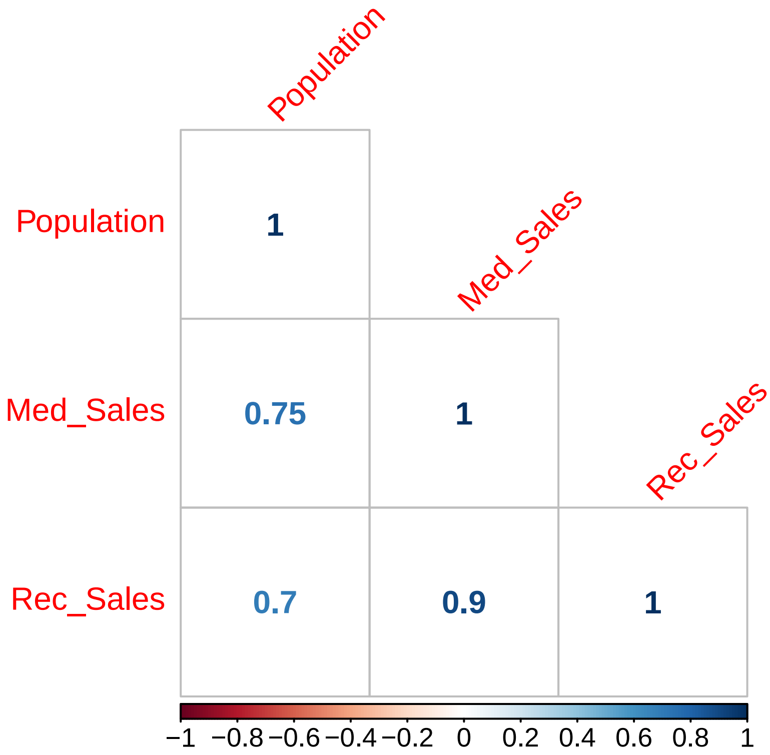
```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
#subset out unnecessary columns
plot(CCMDs[,7:9])
```

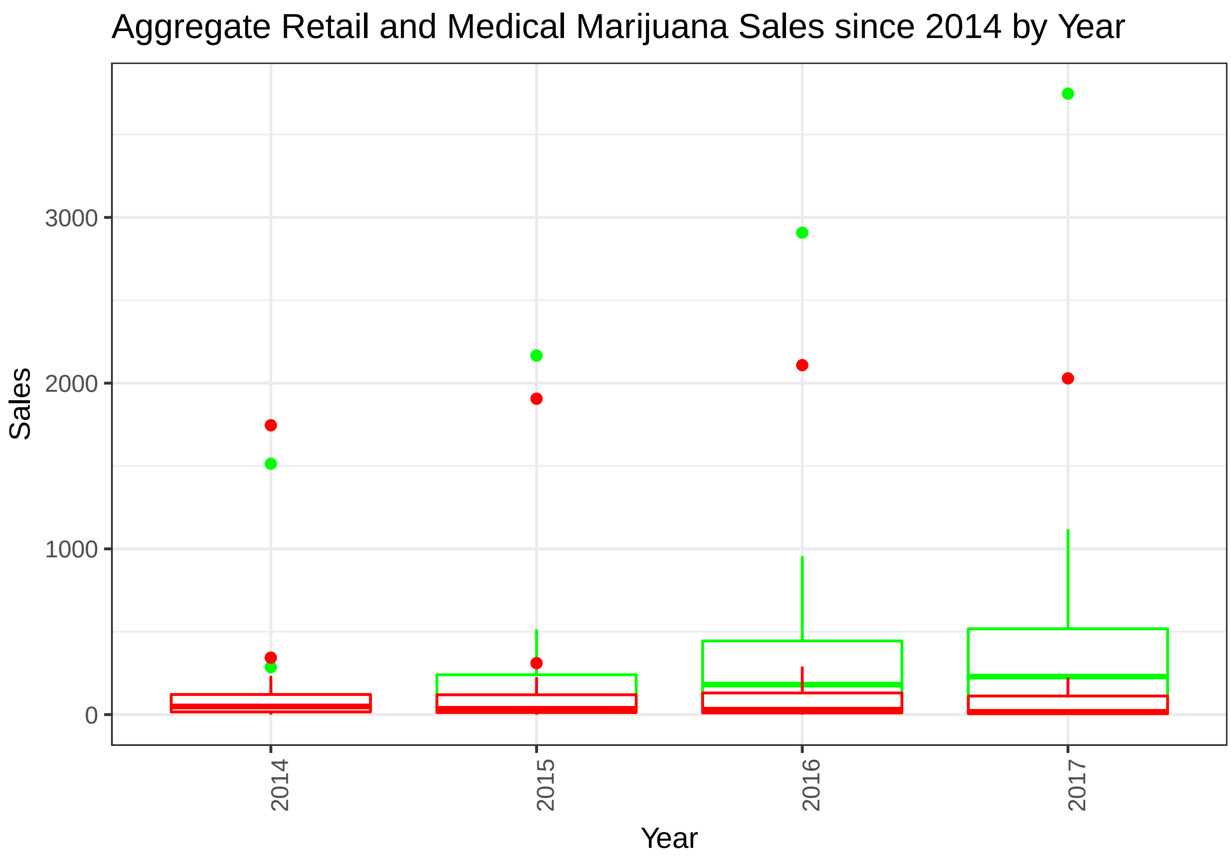


```
#corrplot the value features
m <- cor(CCMDs[, c(7:9)], use = "complete.obs", method = "spearman")
corrplot(m, method="number", type = "lower", order = "hclust", tl.srt = 45)
```

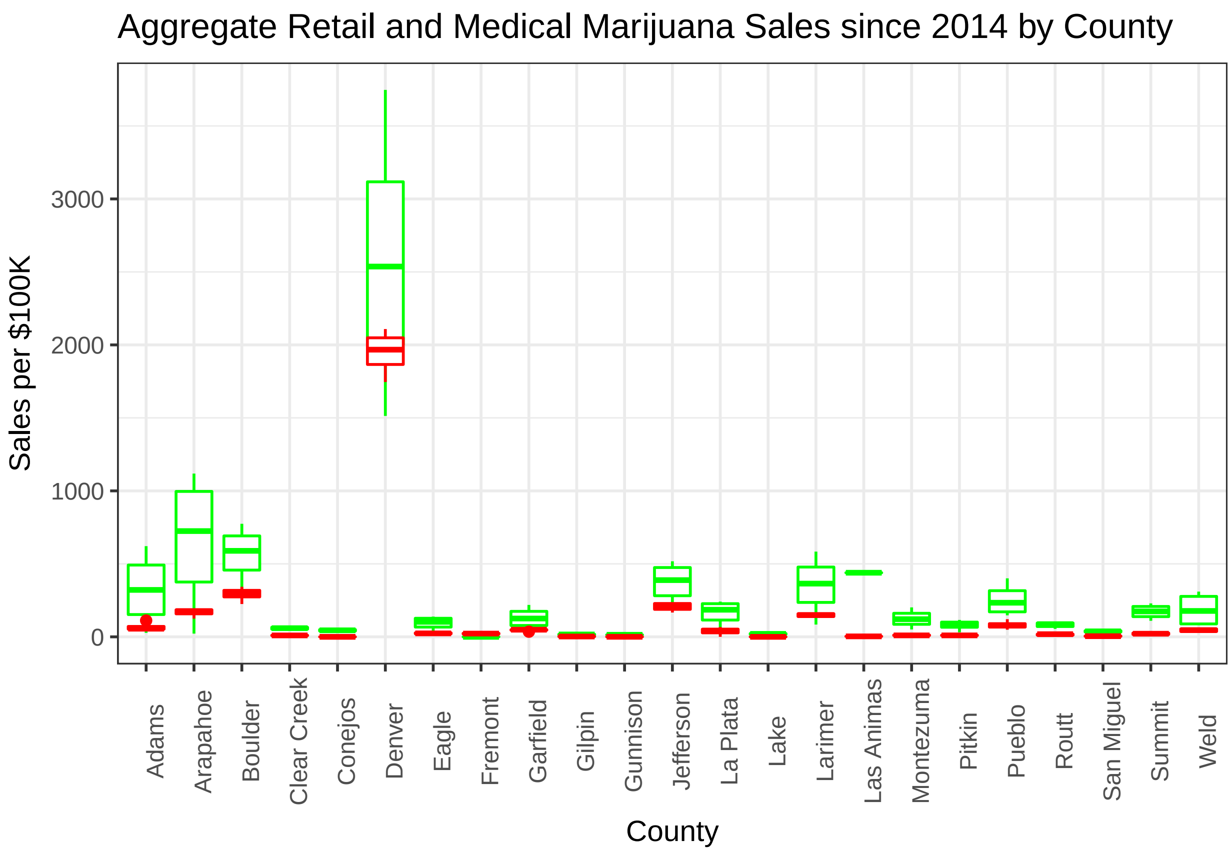


Of course, it makes sense that population would correlate to sales and that med/rec sales would correlate so highly to each other.

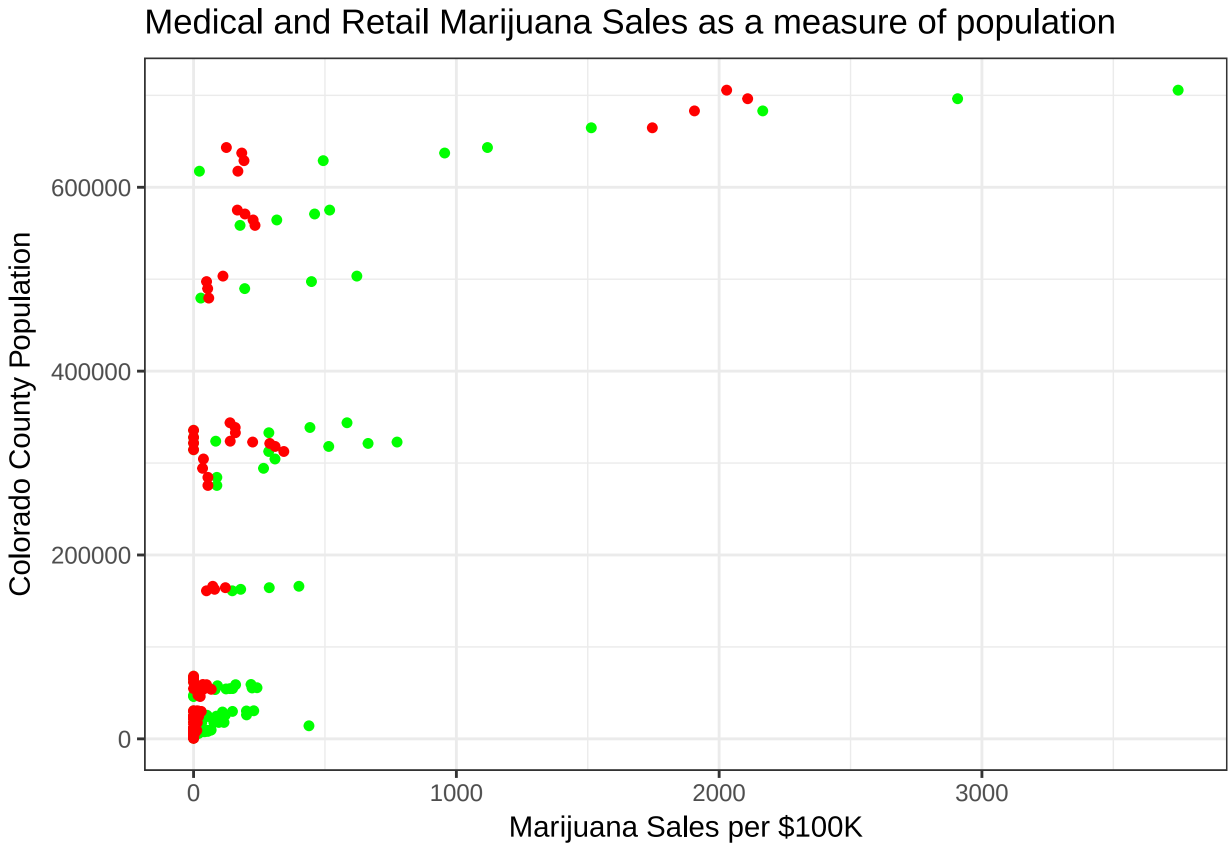
```
CCMDs_f1 <- filter(CCMDs, Med_Sales != 0 | Rec_Sales != 0)
CCMDs_f2 <- filter(CCMDs, Med_Sales == 0 & Rec_Sales == 0)


#CCMDs_f1 <- CCMDs_f1 %>% arrange(Population)

#View(CCMDs_f1)

ggplot(data=CCMDs_f1, aes(x=as.factor(Year)))+
  geom_boxplot(aes(y=Rec_Sales/100000), color="green", show.legend=TRUE)+
  geom_boxplot(aes(y=Med_Sales/100000), color="red", show.legend=TRUE)+
  labs(title="Aggregate Retail and Medical Marijuana Sales since 2014 by Year", x= "Year", y= "Sales")+
  theme_bw()+
  theme(axis.text.x = element_text(angle = 90))
```



```
ggplot(data=CCMDs_f1, aes(x=as.factor(County)))+
  geom_boxplot(aes(y=Rec_Sales/100000), color="green", show.legend=TRUE)+
  geom_boxplot(aes(y=Med_Sales/100000), color="red", show.legend=TRUE)+
  labs(title="Aggregate Retail and Medical Marijuana Sales since 2014 by County", x= "County", y= "Sale
  theme_bw()+
  theme(axis.text.x = element_text(angle = 90))
```

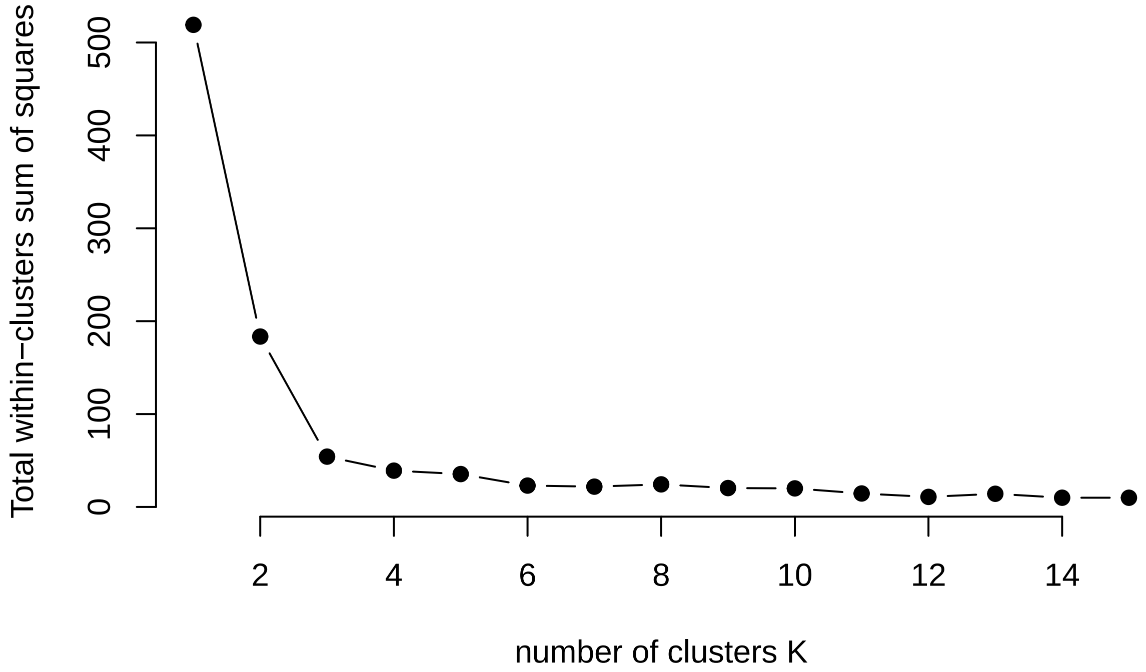


```
#disable scientific notation
options(scipen = 999)

#table(CCMDs[, 1,8:9])

#filter by year

ggplot(data=CCMDs, aes(x = Population))+
  geom_point(aes(y = Rec_Sales/100000), color="green", show.legend = TRUE, size = 1.25)+
  geom_point(aes(y = Med_Sales/100000), color="red", show.legend = TRUE, size = 1.25)+
  labs(title="Medical and Retail Marijuana Sales as a measure of population", x= "Colorado County Popul
  theme_bw()+
  coord_flip()
```



Four clusters seem to be immediately evident (low pop/low sales, med pop/low sales, high pop/low sales, and high pop/higher sales starting around 1000K). Cluster analysis will need to be performed to validate optimal cluster size and groupings.

```
#https://www.r-bloggers.com/finding-optimal-number-of-clusters/


#first scale and sequester the data fields of interest into a matrix.
CCMDs_scale <- scale(CCMDs[,7:9])

#use elbow method to determine optimal number of clusters
set.seed(12345)

#set the max number of clusters
k.max <- 15

#generate within-cluster sum of squares
wss <- sapply(1:k.max, function(k){kmeans(CCMDs_scale, k, nstart=50, iter.max=15)$tot.withinss})

#generate elbow plot
plot(1:k.max, wss, type="b", pch=19, frame=FALSE, xlab="number of clusters K", ylab="Total within-clust
```

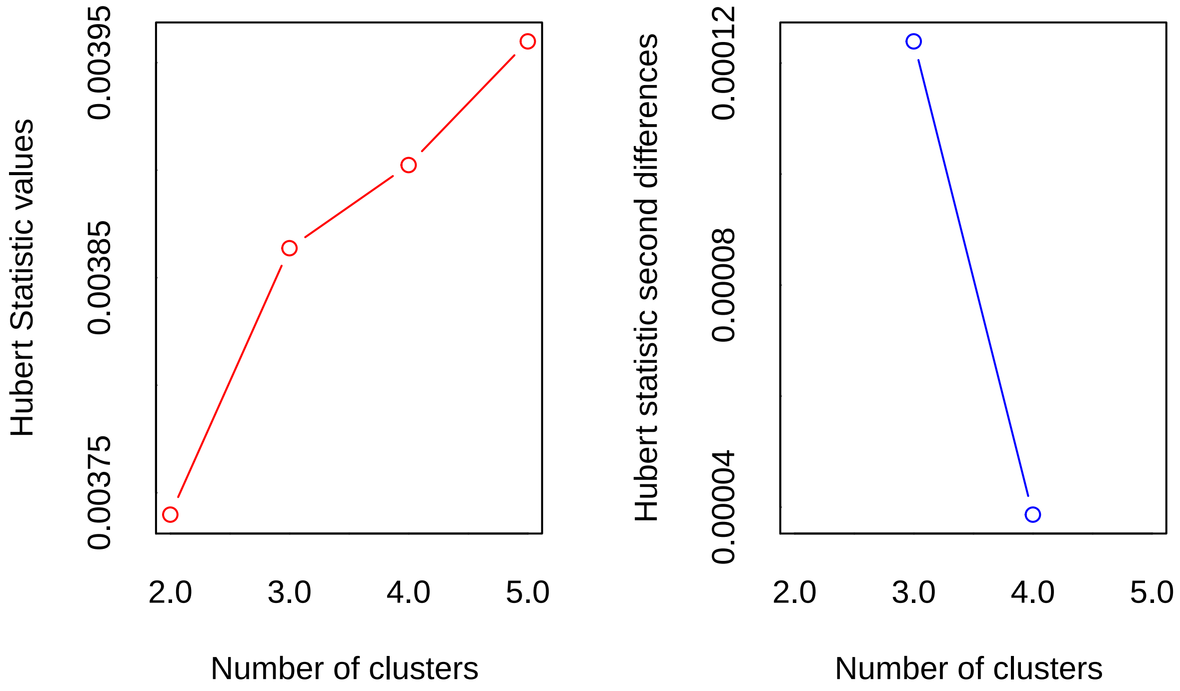


It seems that three clusters was the magic number. I'll also use NbClust to additionally validate the number of clusters. NbClust uses multiple indices to determine the number of clusters and most optimal clustering scheme. Optimal clusters are by determining the amount of variation between the data points.

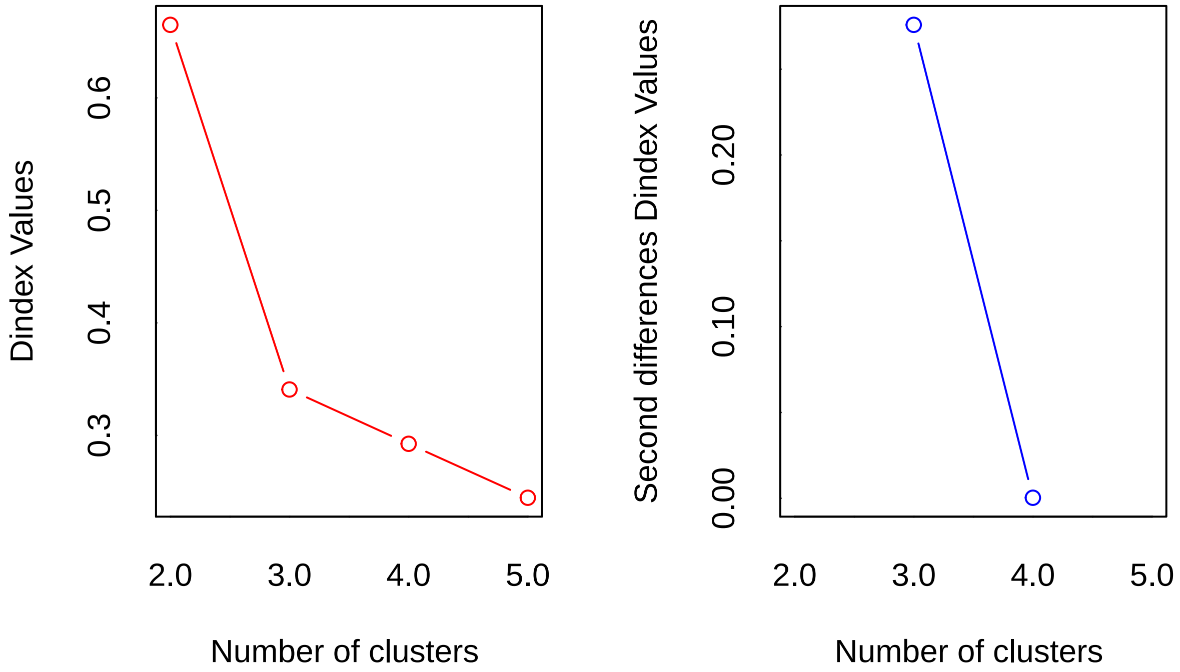
```
#install.packages("NbClust", dependencies=TRUE)
library(NbClust)
```

```
## Warning: package 'NbClust' was built under R version 3.5.2
```

```
nb <- NbClust(CCMDs_scale, distance="euclidean", min.nc=2, max.nc = 5, method="kmeans", index="all", al
```

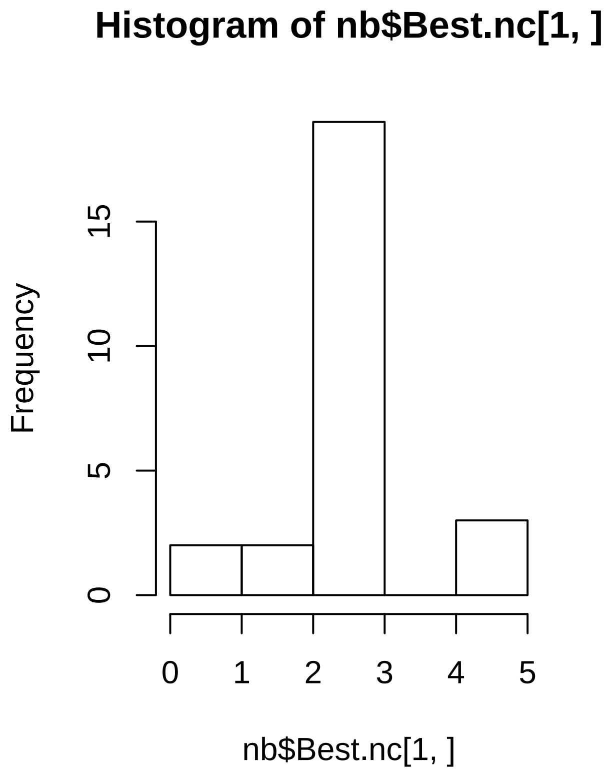


```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that corresponds to a
##                 significant increase of the value of the measure i.e the significant peak in Hubert
##                 index second differences plot.
##
```



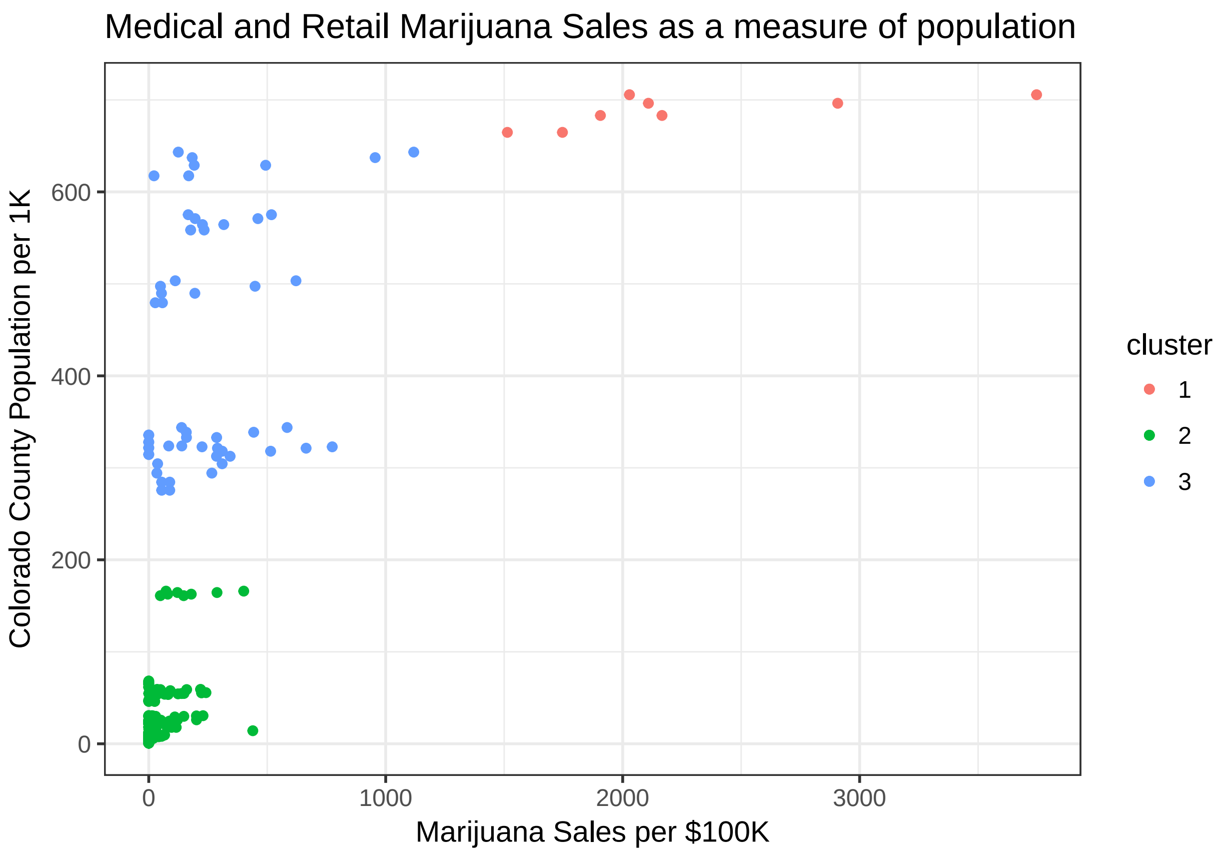
```
## *** : The D index is a graphical method of determining the number of clusters.
##                  In the plot of D index, we seek a significant knee (the significant peak in Dindex
##                  second differences plot) that corresponds to a significant increase of the value of
##                  the measure.
##
## *******************************************************************
## * Among all indices:
## * 2 proposed 2 as the best number of clusters
## * 19 proposed 3 as the best number of clusters
## * 3 proposed 5 as the best number of clusters
##
##                    ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  3
##
##
## *******************************************************************
```

```
hist(nb$Best.nc[1,], breaks = max(na.omit(nb$Best.nc[1,])))
```



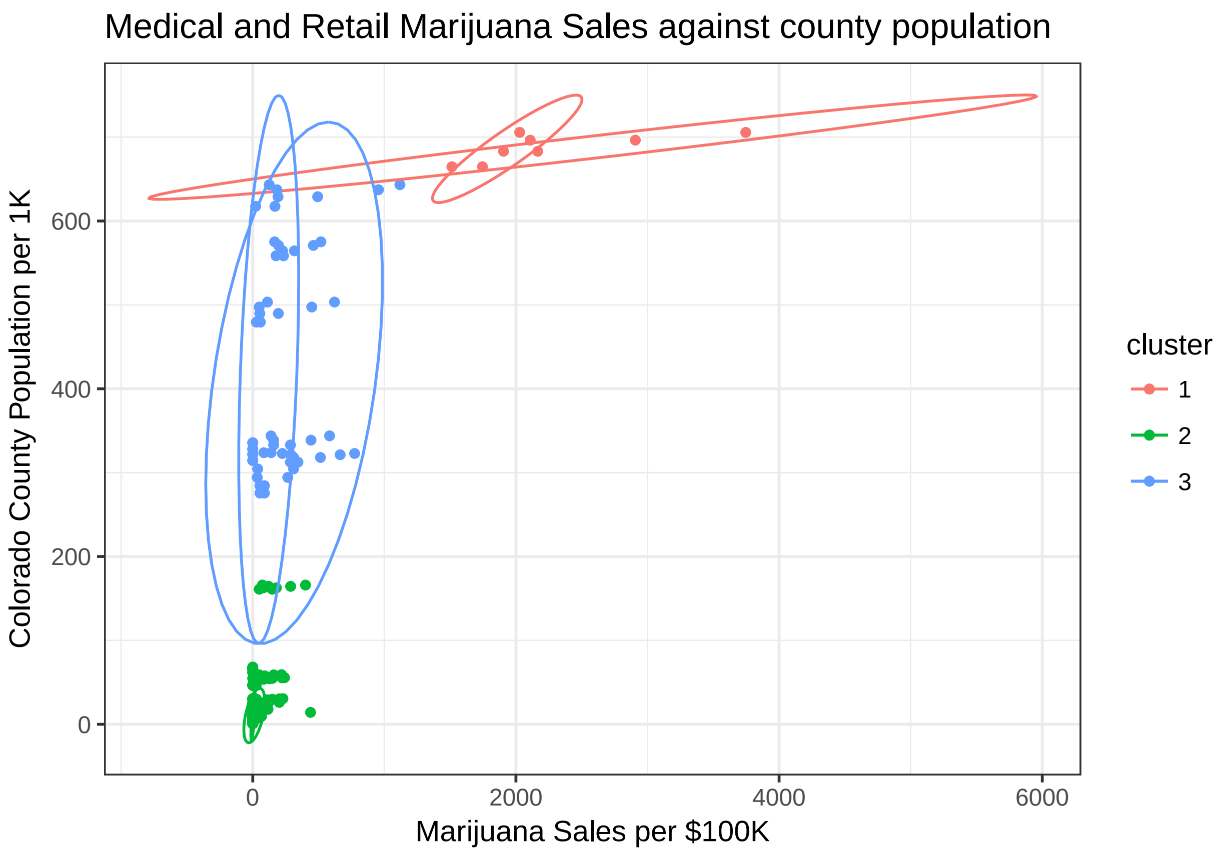
three clusters it is.

```
#use kmeans to perform cluster identification
cluster <- kmeans(CCMDs_scale, centers=3, iter.max=50, nstart=5)

#bind the cluster field to to the values from scale
cluster <- cluster$cluster

#convert cluster number to factor
cluster <- as.factor(cluster)

ggplot(data=CCMDs, aes(x = Population/1000, color=cluster))+
  geom_point(aes(y = Rec_Sales/100000), show.legend = TRUE, size = 1.25)+
  geom_point(aes(y = Med_Sales/100000), show.legend = TRUE, size = 1.25)+
  labs(title="Medical and Retail Marijuana Sales as a measure of population", x= "Colorado County Popula
  theme_bw()+
  coord_flip()
```

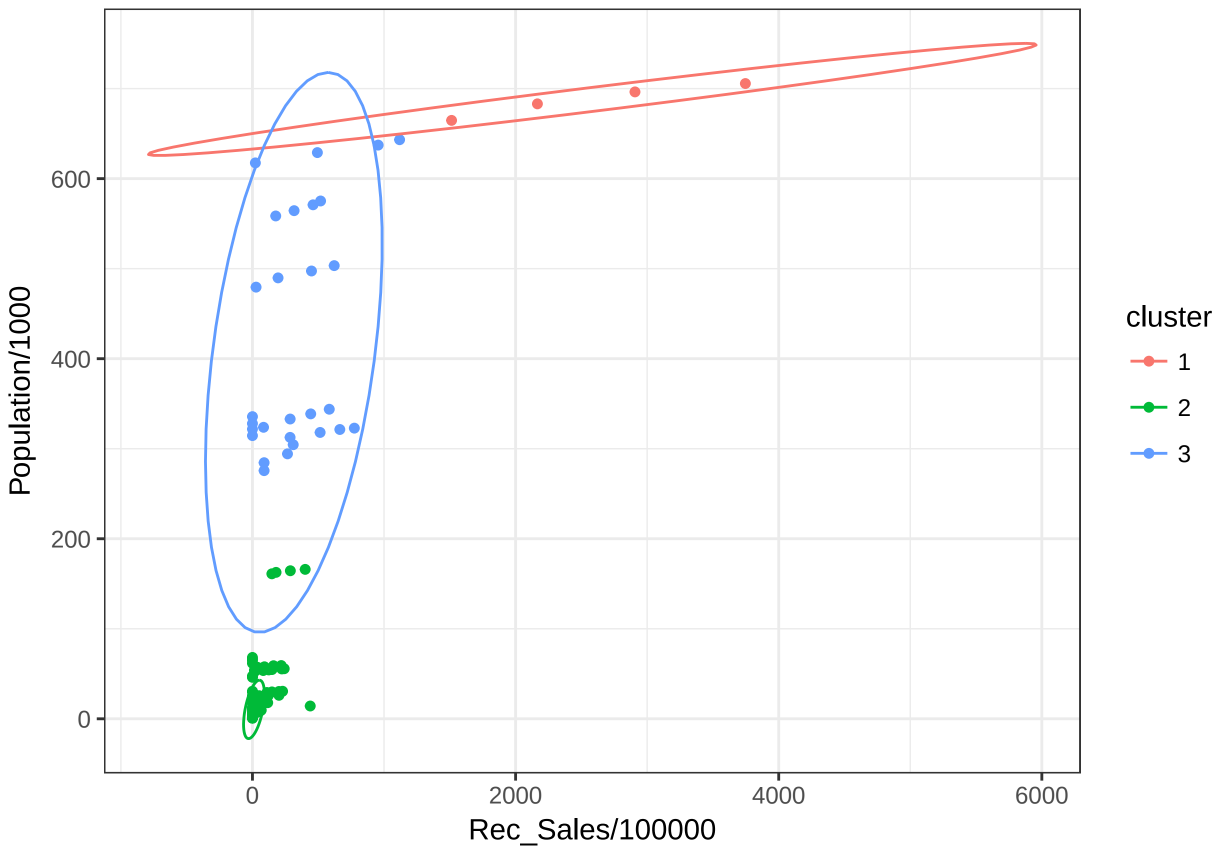


It seems that three clusters was the magic number. One question becomes immediately obvious. Why are there areas with a high population (>600K) but low sales? This will need to be further investigated.

```
CCMDs_f3 <- filter(CCMDs, Population > 600000)
View(CCMDs_f3)

ggplot(data=CCMDs, aes(x = Population/1000, color=cluster))+
  geom_point(aes(y = Rec_Sales/100000), show.legend = TRUE, size = 1.25)+
  stat_ellipse(aes(y = Rec_Sales/100000))+
  geom_point(aes(y = Med_Sales/100000), show.legend = TRUE, size = 1.25)+
  stat_ellipse(aes(y = Med_Sales/100000))+
  labs(title="Medical and Retail Marijuana Sales against county population", x= "Colorado County Popula
  theme_bw()+
  coord_flip()
```



```
ggplot(data=CCMDs, aes(x = Population/1000, color=cluster))+
  geom_point(aes(y = Rec_Sales/100000), show.legend = TRUE, size = 1.25)+
  stat_ellipse(aes(y = Rec_Sales/100000))+
  theme_bw()+
  coord_flip()
```



ref: http://rpubs.com/sinhrks/plot_pca

spare or inwork stuff…

ggplot(data=CCMDs, aes(x = Population))+ geom_point(aes(y = Rec_Sales/100000), color="green", show.legend = TRUE, size = 1.25)+ labs(title="Retail Marijuana Sales as a measure of population", x= "Colorado County Population", y= "Marijuana Sales per $100K")+ theme_bw()+ coord_flip()

ggplot(data=CCMDs, aes(x = Population))+ geom_point(aes(y = Med_Sales/100000), color="red", show.legend = TRUE, size = 1.25)+ labs(title="Medical Marijuana Sales as a measure of population", x= "Colorado County Population", y= "Marijuana Sales per $100K")+ theme_bw()+ coord_flip()

ggplot(data=CCMDs, mapping=aes(x=Year))+ geom_col(mapping=aes(y=Rec_Sales/100000), position=position_dodge(), fill="green")+ geom_col(mapping=aes(y=Med_Sales/100000), position=position_dodge() , fill="red")+ labs(title="Colorado Medical and Retail Marijuana Sales 2014-2018", x= "Years of Recreational Legalization", y= "Sales per $100K")+ theme_bw()

ggplot(data=CCMDs, aes(x=as.factor(CCMDs$County)))+ geom_boxplot(aes(y=Rec_Sales), color="green")+ geom_boxplot(aes(y=Med_Sales), color="red")+ labs(title="Colorado Retail and Medical Marijuana Sales since 2014", x= "Colorado County", y= "Sales")+ theme_bw()+ theme(axis.text.x = element_text(angle = 90, hjust = 1))

ggplot(data=CCMDs, mapping=aes(x=Year))+ geom_jitter(aes(y=Rec_Sales), color="green")+ geom_jitter(aes(y=Med_Sales), color="red")+ labs(title="Colorado Medical and Retail Marijuana Sales 2014-2018", x= "Years of Recreational Legalization", y= "Sales per $100K")+ theme_bw()

CCMDs_pop <- CCMDs[, c(1,2,7)] CCMDs_pop$County < -as.factor(CCMDs_popCounty$ County) CCMDs_pop <- CCMDs_pop[order(CCMDs_pop$Population),]

str(CCMDs_pop)

#CCMDs_pop <- CCMDs_pop[order(CCMDs_pop$Population),]

View(CCMDs_pop)

#ggplot(data=CCMDs, mapping=aes(x=Year, y=Population, group=Region))+ # geom_line(color="red")+ # geom_point(color="red")+ # scale_y_continuous(limits = c(min(CCMDs$Population), max(CCMDs$Population)))+ # geom_label_repel(aes(label = Population), nudge_x = 1)+ # labs(title="Colorado Estimated Population Growth since 2010 Census")+ # theme_bw()

#CCMDs_pop <- data.frame(order(CCMDs_pop$Population))

ggplot(data=CCMDs_pop, mapping=aes(x=County, y=Population))+ geom_bar(stat="identity", position="dodge", fill="blue")+ labs(title="Colorado Population by County", x= "County", y= "Population")+ theme_bw()+ theme(axis.text.x = element_text(angle = 90))

ggplot(data=CCMDs_pop, aes(x=as.factor(CCMDs_pop$County)))+ geom_boxplot(aes(y=Population), color="blue")+ labs(title="Colorado Population by County", x= "County", y= "Population")+ theme_bw()+ theme(axis.text.x = element_text(angle = 90))

#corrplot the value features #m <- cor(CCMDs[, c(8:14)], use = "complete.obs", method = "spearman") #require("corrplot") #corrplot(m, type = "upper", order = "hclust", tl.srt = 45)