Final Project

Aaron Coates

6/8/2019

```r
setwd("/Users/aaroncoates/Documents/GitHub/MMSS_311_2")

library(tidytext)
library(tm)
```

```
## Loading required package: NLP
```

```r
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(stringr)
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:NLP':
##
##     annotate
```

```r
library(proxy)
```

```
##
## Attaching package: 'proxy'

## The following objects are masked from 'package:stats':
##
##     as.dist, dist

## The following object is masked from 'package:base':
##
##     as.matrix
```

```r
library(fields)
```

```
## Loading required package: spam

## Loading required package: dotCall64

## Loading required package: grid

## Spam version 2.2-2 (2019-03-07) is loaded.
## Type 'help( Spam)' or 'demo( spam)' for a short introduction
## and overview of this package.
## Help for individual functions is also obtained by adding the
## suffix '.spam' to the function name, e.g. 'help( chol.spam)'.

##
## Attaching package: 'spam'

## The following objects are masked from 'package:base':
##
##     backsolve, forwardsolve

## Loading required package: maps

## See https://github.com/NCAR/Fields for
##  an extensive vignette, other supplements and source code
```

```
library(mixtools)
```

```
## mixtools package, version 1.1.0, Released 2017-03-10
## This package is based upon work supported by the National Science Foundation under Grant No. SES-0518

##
## Attaching package: 'mixtools'

## The following object is masked from 'package:grid':
##
##     depth
```

```
library(xml2)
library(rvest)
```

```
##
## Attaching package: 'rvest'

## The following object is masked from 'package:readr':
##
##     guess_encoding
```

```
library(maps)
library(mapdata)
library(devtools)
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(tidyr)
library(RColorBrewer)

USAData <- read_csv('/Users/aaroncoates/Desktop/ML Final/DataUSA Health and Safety.csv')
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   `ID Geography` = col_character(),
##   Geography = col_character()
## )

## See spec(...) for full column specifications.
```

```r
GoodVars <- c(1, 7, 12, 17, 22)
USAData <- USAData[ ,GoodVars]
usanames <- c("FIPS", "Obesity", "Smoking", "AirPol", "Crime")
names(USAData) <- usanames
USAData$FIPS <- gsub("^.*US", "", USAData$FIPS) %>%
  as.numeric()

MedHouseInc <- read_csv('/Users/aaroncoates/Desktop/ML Final/MedHousInc.csv')
```

```
## Parsed with column specification:
## cols(
##   County = col_character(),
##   `FIPS Code` = col_double(),
##   `Median Household Income 2017` = col_character(),
##   Rank = col_double()
## )

## Warning: 1 parsing failure.
##  row  col expected actual                                                 file
## 3142 Rank a double    N/A '/Users/aaroncoates/Desktop/ML Final/MedHousInc.csv'
```

```r
MedHouseInc <- MedHouseInc[, 1:3]
HSDiploma <- read_csv('/Users/aaroncoates/Desktop/ML Final/HSDiploma.csv')
```

```
## Warning: Missing column names filled in: 'X5' [5], 'X6' [6], 'X7' [7],
## 'X8' [8]

## Parsed with column specification:
## cols(
##   County = col_character(),
##   `FIPS Code` = col_double(),
##   `2017 ACS Population - High School Diploma or More as a percent of 2017 ACS Population - Populatio
##   Rank = col_double(),
##   X5 = col_logical(),
##   X6 = col_logical(),
##   X7 = col_logical(),
##   X8 = col_logical()
## )
```

```r
HSDiploma <- HSDiploma[, 2:3]
WhitePop <- read_csv('/Users/aaroncoates/Desktop/ML Final/WhitePop.csv')
```

```
## Parsed with column specification:
## cols(
##   County = col_character(),
##   `FIPS Code` = col_double(),
##   `ACS Population - White Alone 2017` = col_number(),
##   Rank = col_double()
## )
```

```r
WhitePop <- WhitePop[,2:3]
FamHouseholds <- read_csv('/Users/aaroncoates/Desktop/ML Final/FamHouseholds.csv')
```

```
## Parsed with column specification:
## cols(
##   County = col_character(),
##   `FIPS Code` = col_double(),
##   `ACS Population - Family Households 2017` = col_number(),
##   Rank = col_double()
## )
```

```
FamHouseholds <- FamHouseholds[,2:3]
UnempRate <- read_csv('/Users/aaroncoates/Desktop/ML Final/UnempRate.csv')
```

```
## Parsed with column specification:
## cols(
##   County = col_character(),
##   `FIPS Code` = col_double(),
##   `Unemployment Rate 2018` = col_double(),
##   Rank = col_double()
## )
```

```
UnempRate <- UnempRate[,2:3]

StatsAmerica <- inner_join(MedHouseInc, HSDiploma, "FIPS Code") %>%
  inner_join(WhitePop, "FIPS Code") %>%
  inner_join(FamHouseholds, "FIPS Code") %>%
  inner_join(UnempRate, "FIPS Code")
statnames <- c("County", "FIPS", "MedHousInc", "HSDipl", "White", "FamHous", "UnempRate")
names(StatsAmerica) <- statnames

DataSet <- inner_join(StatsAmerica, USAData, "FIPS")

sum(is.na(DataSet$MedHousInc))
```

```
## [1] 0
```

```
sum(is.na(DataSet$HSDipl))
```

```
## [1] 0
```

```
sum(is.na(DataSet$White))
```

```
## [1] 0
```

```
sum(is.na(DataSet$FamHous))
```

```
## [1] 0
```

```
sum(is.na(DataSet$UnempRate))
```

```
## [1] 0
```

```
sum(is.na(DataSet$Obesity))
```

```
## [1] 0
```

```
sum(is.na(DataSet$Smoking))
```

```
## [1] 0
```

```
sum(is.na(DataSet$AirPol))
```

```
## [1] 33
```

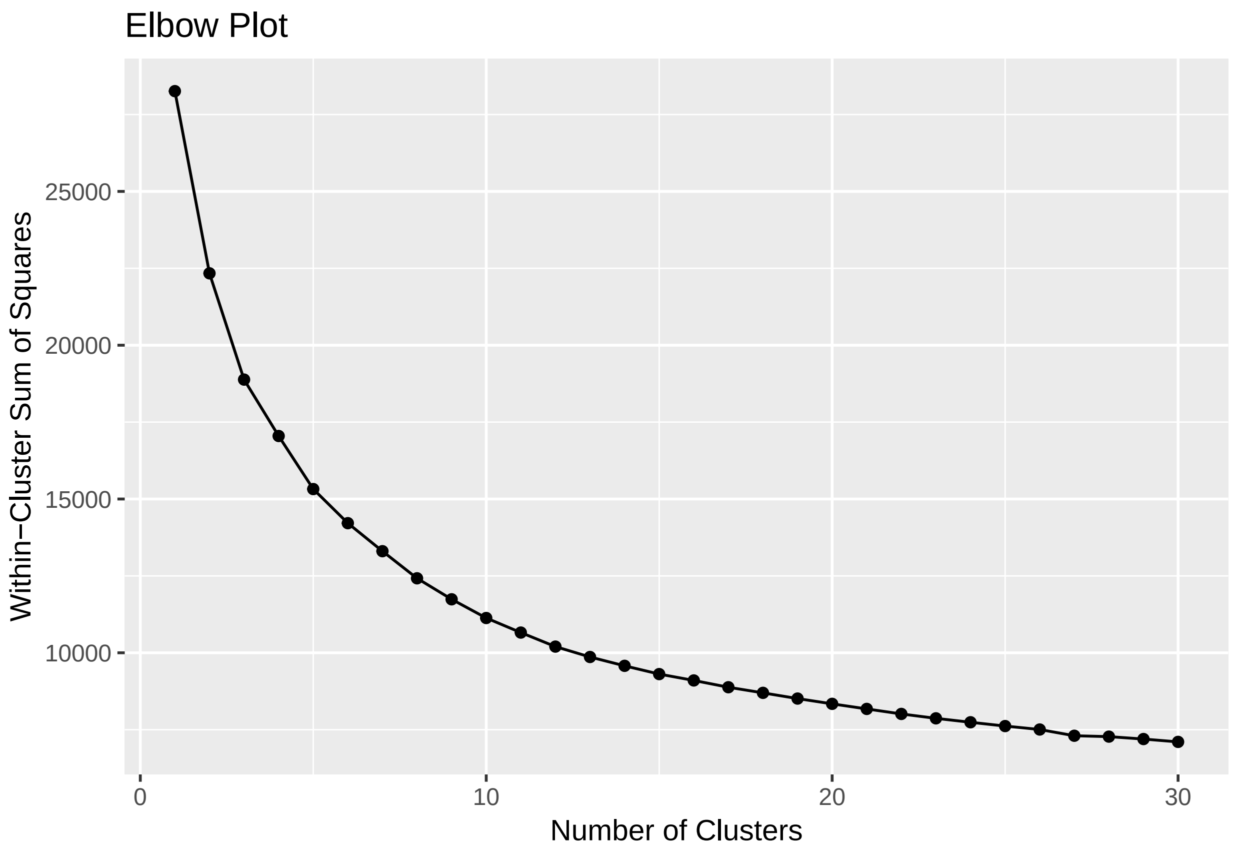
```
sum(is.na(DataSet$Crime))
```

```
## [1] 175
```

```r
FillAirPol <- mean(DataSet$AirPol, na.rm=TRUE)
DataSet$AirPol[is.na(DataSet$AirPol)] <- FillAirPol
FillCrime <- mean(DataSet$Crime, na.rm=TRUE)
DataSet$Crime[is.na(DataSet$Crime)] <- FillCrime

DataSet <- apply(DataSet, 2, function(x) gsub("[$,%]", "", x)) %>%
  as.data.frame()

VarsOnly <- DataSet[3:11] %>%
  apply(2, as.numeric)
scaledVars <- scale(VarsOnly) %>%
  as.data.frame()

set.seed(100)
WCSS <- numeric(30)
for (i in 1:30){
  km <- kmeans(scaledVars, i, nstart=100)
  WCSS[i] <- km$tot.withinss
}
```

```
## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations
```

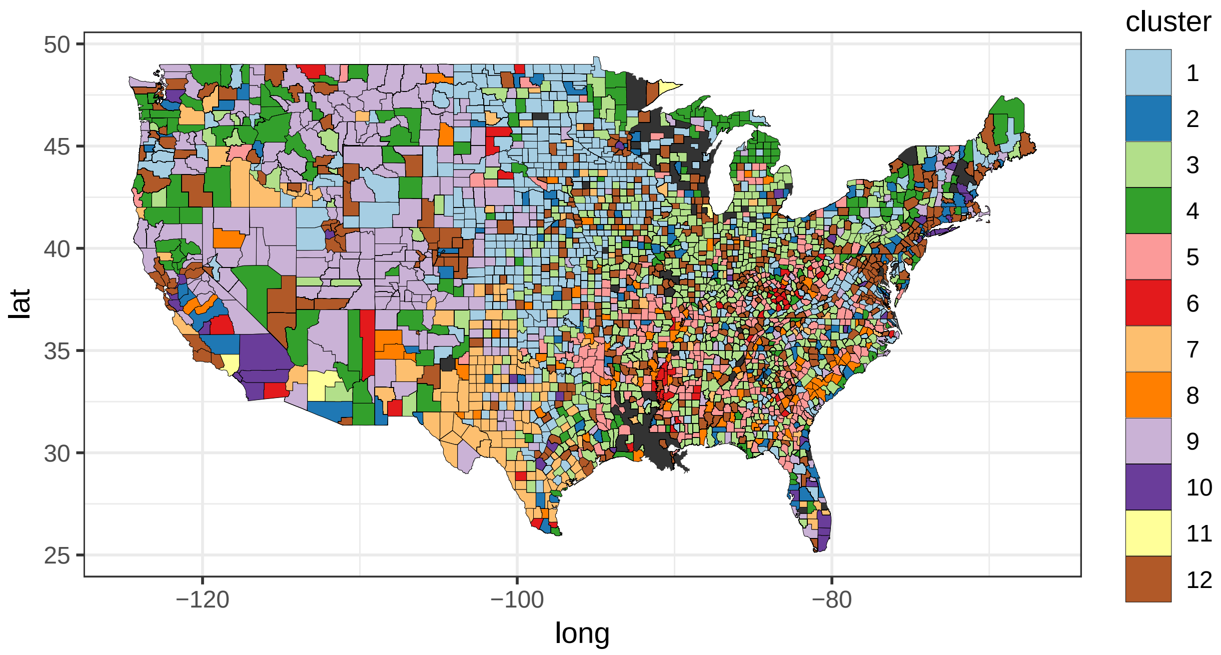
```
WCSS<- as.data.frame(WCSS)
WCSS$k <- c(1:30)
ggplot(WCSS, aes(k, WCSS)) + geom_line() + geom_point() +
  ggtitle("Elbow Plot") + xlab("Number of Clusters") +
  ylab("Within-Cluster Sum of Squares")
```



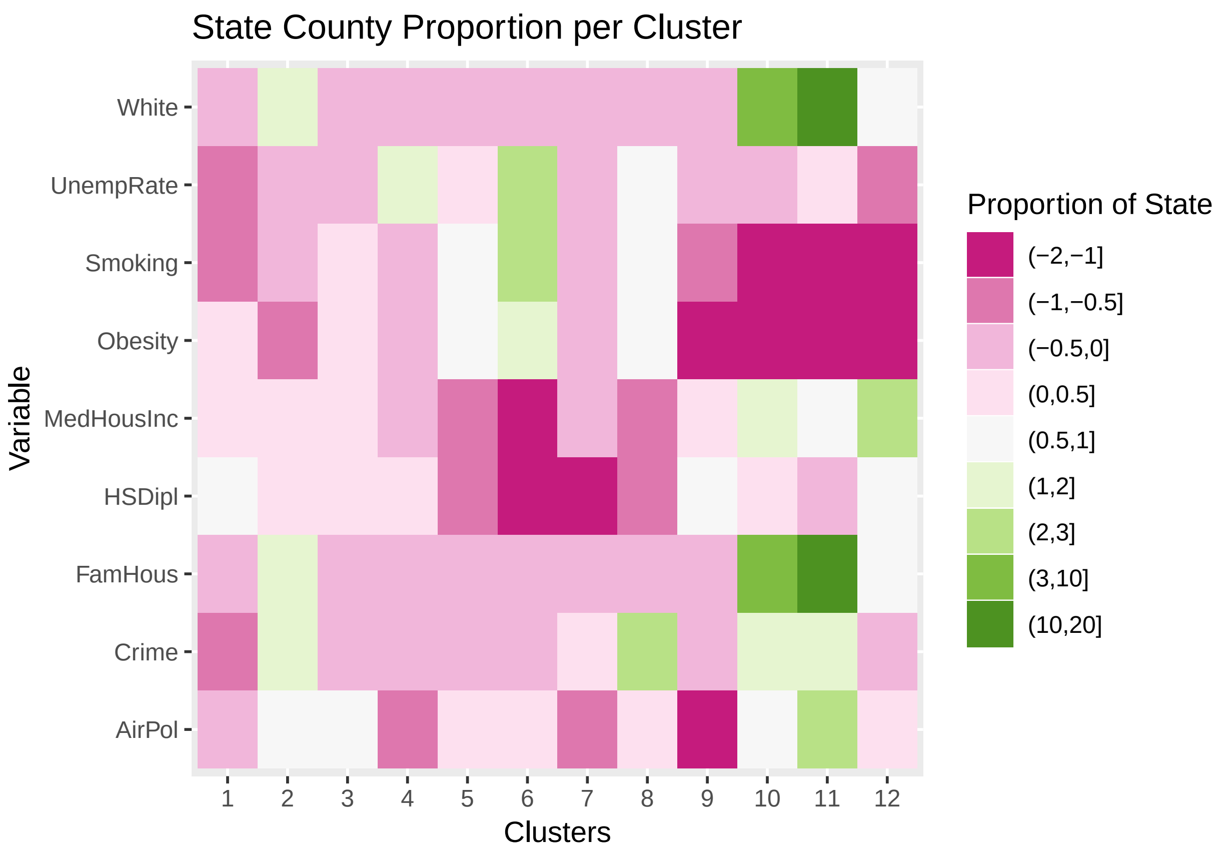
```
KMeans <- kmeans(scaledVars, 12, nstart=100)
```

```
## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations
```

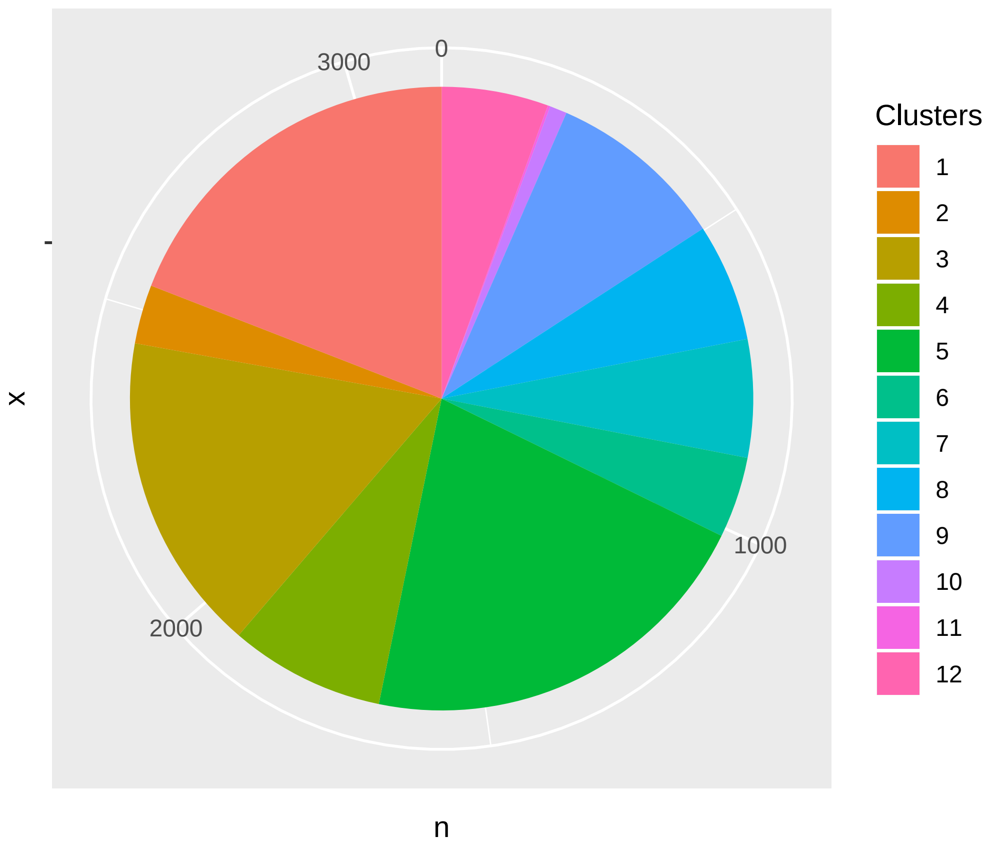
```
DataSet$Clusters <-KMeans$cluster %>%
  as.factor()


GraphData <- DataSet[ ,c(1, 12)]
GraphNames <- c("subregion", "cluster")
names(GraphData) <- GraphNames

USA <- map_data("usa")
Counties <- map_data("county")
GraphData$subregion <- as.character(tolower(gsub(" County ..$", "", GraphData$subregion)))
Counties$subregion <- as.character(Counties$subregion)

FinalGraph <- inner_join(GraphData, Counties, 'subregion')

ggplot(data = FinalGraph, mapping = aes(x = long, y = lat, group = group)) +
  coord_fixed(1.3) + geom_polygon(data = USA, aes(x=long, y = lat, group = group)) +
  geom_polygon(data = FinalGraph, aes(fill = cluster), color = "black", size=.05) +
  theme_bw() + scale_fill_brewer(palette = 'Paired')
```



```
Centers <- as.data.frame(KMeans$centers)
Centers$Clusters <- as.factor(c(1:12))
Centers <- gather(Centers, Variable, Value, -Clusters)

ggplot(Centers, aes(Clusters, Variable,
                    fill=cut(Value, c(-20, -10, -3, -2, -1, -.5, 0, .5, 1, 2, 3, 10, 20)))) +
  geom_raster() + scale_fill_brewer(palette = "PiYG") +
  guides(fill=guide_legend(title="Proportion of State")) + ggtitle('State County Proportion per Cluster
```



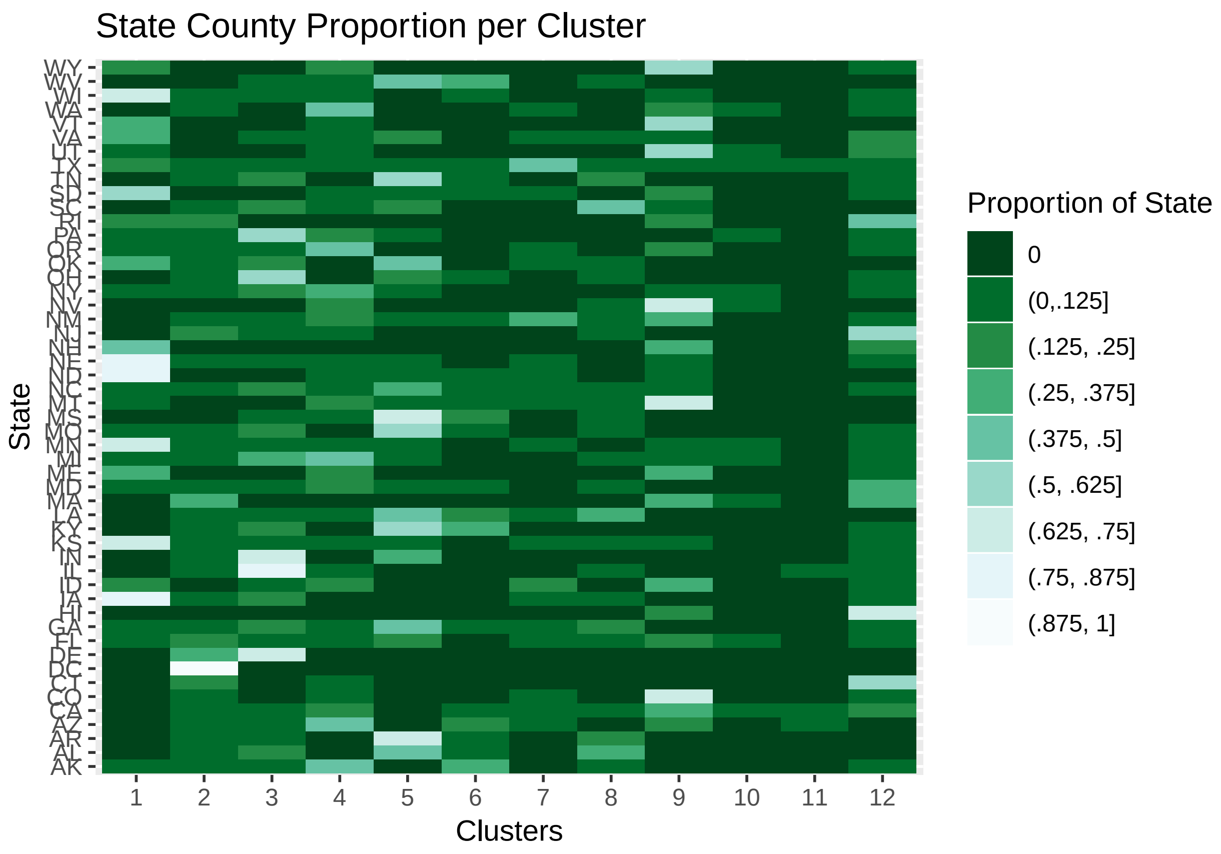
```r
ClusterTotals <- DataSet %>%
  group_by(Clusters) %>%
  count()

ggplot(ClusterTotals, aes("", n, fill=Clusters)) +
    geom_bar(width = 1, stat = "identity") + coord_polar("y", start=0)
```



```
DataSet$State <- str_sub(DataSet$County, -2, -1)
table(DataSet$State, DataSet$Clusters)
```

```
##
##         1   2   3   4   5   6   7   8   9  10  11  12
##   AK    3   1   3  11   0   9   0   1   0   0   0   1
##   AL    0   2  11   0  32   4   0  18   0   0   0   0
##   AR    0   1   9   0  48   2   0  15   0   0   0   0
##   AZ    0   1   1   6   0   2   1   0   3   0   1   0
##   CA    0   4   1   8   0   2   3   2  17   7   1  13
##   CO    0   3   0   2   0   0   3   0  48   0   0   8
##   CT    0   2   0   1   0   0   0   0   0   0   0   5
##   DC    0   1   0   0   0   0   0   0   0   0   0   0
##   DE    0   1   2   0   0   0   0   0   0   0   0   0
##   FL    3  11   2   6  14   0   6   6  14   4   0   1
##   GA    5   3  26   5  77   9   3  26   0   0   0   5
##   HI    0   0   0   0   0   0   0   0   1   0   0   3
##   IA   81   1  13   0   0   0   1   1   0   0   0   2
##   ID   10   0   1   8   0   0   8   0  16   0   0   1
##   IL    0   1  78   9   0   0   0   5   0   0   1   8
##   IN    0   2  64   0  24   0   0   0   0   0   0   2
##   KS   73   1  12   1   5   0   8   3   1   0   0   1
##   KY    0   1  20   0  65  33   0   0   0   0   0   1
##   LA    0   2   5   1  25   9   1  21   0   0   0   0
##   MA    0   5   0   0   0   0   0   0   4   1   0   4
##   MD    2   2   3   4   1   1   0   2   0   0   0   9
##   ME    6   0   0   4   0   0   0   0   5   0   0   1
##   MI   4   4  21  38   7   0   0   3   1   2   0   3
##   MN  55   2   6   9   1   0   1   0   2   1   0  10
##   MO   8   2  25   0  71   1   0   7   0   0   0   1
##   MS   0   0   5   1  52  19   0   5   0   0   0   0
##   MT   1   0   0  10   1   2   2   1  39   0   0   0
##   NC  11   4  19   9  32   1   4  12   5   0   0   3
##   ND  42   0   0   4   1   2   1   0   3   0   0   0
##   NE  72   1   3   2   2   0   2   0   9   0   0   2
##   NH   5   0   0   0   0   0   0   0   3   0   0   2
##   NJ   0   4   1   2   0   0   0   1   0   0   0  13
##   NM   0   1   1   7   1   1   9   3   9   0   0   1
##   NV   0   0   0   3   0   0   0   1  12   1   0   0
##   NY   4   5  10  23   2   0   0   0   6   5   0   7
##   OH   0   5  53   0  21   5   0   1   0   0   0   3
##   OK  21   2  13   0  33   0   4   4   0   0   0   0
##   OR   3   1   1  18   0   0   2   0   9   0   0   2
##   PA   4   4  40  11   2   0   0   0   0   1   0   5
##   RI   1   1   0   0   0   0   0   0   1   0   0   2
##   SC   0   4  11   1   9   0   0  20   1   0   0   0
##   SD  41   0   0   2   4   7   1   0   9   0   0   2
##   TN   0   4  15   0  57   2   0  16   0   0   0   1
##   TX  33   4  25  11  18   3 120   6  18   4   1  11
##   UT   2   0   0   2   0   0   0   0  18   1   0   6
##   VA  48   0   8   3  30   0   8   6   5   0   0  25
##   VT   5   0   0   1   0   0   0   0   8   0   0   0
##   WA   0   2   0  19   0   0   4   0   9   1   0   4
##   WI  52   1   6   6   0   1   0   0   1   0   0   5
##   WV   0   0   6   2  24  17   0   6   0   0   0   0
##   WY   5   0   0   3   0   0   0   0  14   0   0   1
```

```r
StateTable <- DataSet %>%
  group_by(State, Clusters) %>%
  count()
StateTable$Clusters <- as.integer(StateTable$Clusters)
StateTable <- StateTable %>%
  group_by(State) %>%
  complete(Clusters = seq(1, 12), fill=list(n=0))
StateTable$Clusters <- as.factor(StateTable$Clusters)

StateMatrix <- spread(StateTable, key = "State", value = "n") %>%
  as.data.frame()
StateMatrix <- StateMatrix[ ,2:52]
StateMatrix[is.na(StateMatrix)] <- 0

CountyTotals <- colSums(StateMatrix) %>%
  as.data.frame()
colnames(CountyTotals) <- 'Total'
CountyTotals$State <- rownames(CountyTotals)

StateTable <- inner_join(StateTable, CountyTotals, 'State')
StateTable$Prop <- StateTable$n / StateTable$Total

ggplot(StateTable, aes(Clusters, State,
  fill=cut(Prop, c(1, .875, .75, .625, .5, .375, .25, .125, .0001, 0),
         labels = c("0", "(0,.125]", "(.125, .25]", "(.25, .375]", "(.375, .5]",
                    "(.5, .625]", "(.625, .75]", "(.75, .875]", "(.875, 1]"), include.lowest = TRUE))
  geom_tile() + scale_fill_brewer(palette = "BuGn", direction=-1) +
  guides(fill=guide_legend(title="Proportion of State")) + ggtitle('State County Proportion per Cluster
```


State County Proportion per Cluster
State
WY
WV
WI
WA
VT
VA
UT
TX
TN
SD
SC
RI
PA
OR
OK
OH
NY
NV
NM
NJ
NH
NE
ND
NC
MT
MS
MO
MN
MI
ME
MD
MA
LA
KY
KS
IN
IL
ID
IA
HI
GA
FL
DE
DC
CT
CO
CA
AZ
AR
AL
AK
Clusters
Proportion of State
(0,.125]
(.125, .25]
(.25, .375]
(.375, .5]
(.5, .625]
(.625, .75]
(.75, .875]
(.875, 1]