Final Project

Aaron Coates

6/10/2019

Often in American society, overall prosperity is measured by looking solely at economic measures such as household income. Yet, there are many additional, non-economic measures of prosperity that should be considered when looking at the overall prosperity of an area or region. These additional measures may or may not be correlated with economic measures such as income; yet, these additional measures will surely allow for useful inferences when they are included in prosperity calculations.

Thus, for this project, data will be combined from economic, environmental, health, safety, and educational measures of prosperity to look at relative prosperity levels throughout the United States. The project will look to investigate and display the qualitative differences in prosperity levels that exist throughout the United States at the county level while also investigating whether or not these differences in prosperity are tied to geographic location. K-means clustering will be applied to assist in the task of classifying 'quality of life' groupings that exist throughout the United States, and an elbow plot will be utilized to determine the appropriate number of clusters.

```r
setwd("/Users/aaroncoates/Documents/GitHub/MMSS_311_2/ML Final")
Packages <- c('tidytext', 'tm', 'readr', 'dplyr', 'stringr', 'ggplot2', 'proxy',
              'fields', 'mixtools', 'xml2', 'rvest', 'maps', 'mapdata', 'devtools',
              'ggmap', 'tidyr', 'RColorBrewer', 'usmap', 'scales')
lapply(Packages, require, character.only = TRUE)
```

Data USA is a collaborative effort between Deloitte, Datawheel, and MIT professor Cesar Hidalgo that houses comprehensive data about every county and state in the USA. For the purposes of this project, county level data on obesity, smoking, air pollution, and crime rates will be included. These data are 2018 estimates and are originally sourced from the Robert Wood Johnson Foundation's County Health Rankings, yet made available on Data USA's website.

Below, I show the data that I gathered from Data USA, where FIPS is a county-level identifying string. Obesity and smoking are recorded as percentages of the adult population, air pollution is recorded as a density per cubic meter, and the violent crime rate is the number of crimes per 100,000 people. All data are

2018 estimates.

These variables were chosen for a variety of reasons. Obesity and smoking rates can give a general picture of the overall physical wellbeing and health-related prosperity of citizens of the United States. Air pollution can serve as an environmental wellness indicator, or a predictor of overall pollution levels and sanitation in a given county. The crime rate can serve as an indicator of relative safety in a given county.

```r
USAData <- read_csv('DataUSA Health and Safety.csv')
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()
```

StatsAmerica is an online US datasource created and maintained by the Indiana Business Research Center. The site lists county-level data for every US county, and for this project, data on household income, high school graduation rates, white residents, family households, unemployment rates, and overall population have been pulled.

Below, data from the website is pulled and then combined into one large dataset. Household income is reported as the county median, high school graduation rate is a percent, and unemployment rate is a percent. Data on the white population and family households are reported as an absolute value, so in an attempt to standardize the data, values for these variables are divided by the overall county population. All data are 2017 or 2018 estimates.

These variables each serve as an indicator of a county's prosperity. Household income and unemployment rates serve as measures of economic prosperity in a given county, and high school graduation rates show the educational achievement and access available in a given county. Demographic information, such as the percentage of white people in a county, serves as an indicator of diversity within a county. The family household rate can serve as an indicator of whether a neighborhood is welcoming to children and families. This could perhaps serve as a qualitative indicator, measuring whether or not a county is a place where families want to 'settle down.'

```r
MedHouseInc <- read_csv('MedHousInc.csv')
MedHouseInc <- MedHouseInc[, 1:3]
HSDiploma <- read_csv('HSDiploma.csv')
HSDiploma <- HSDiploma[, 2:3]
WhitePop <- read_csv('WhitePop.csv')
WhitePop <- WhitePop[,2:3]
FamHouseholds <- read_csv('FamHouseholds.csv')
FamHouseholds <- FamHouseholds[,2:3]
UnempRate <- read_csv('UnempRate.csv')
UnempRate <- UnempRate[,2:3]
Pop <- read_csv('Pop.csv')
Pop <- Pop[,2:3]

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

Finally, the datasets from both websites are combined into one final dataset. Population is dropped from this dataset as it is not an indicator of a county's prosperity, and it was just used to standardize the white population and family household variables. This is the dataset that will be utilized for the k-means calculation.

```r
DataSet <- inner_join(StatsAmerica, USAData, "FIPS")
DataSet <- DataSet[ , c(1:7, 9:12)]
head(DataSet)
```

```
## # A tibble: 6 x 11
##   County  FIPS MedHousInc HSDipl  White FamHous UnempRate Obesity Smoking
##   <chr>  <dbl> <chr>      <chr>   <dbl>   <dbl>     <dbl>   <dbl>   <dbl>
## 1 Loudo~ 51107 $136,191   93.50%  0.617   0.232       2.5   0.215  0.113
## 2 Falls~ 51610 $123,923   98.20% NA      NA           2.1   0.281  0.128
## 3 Santa~  6085 $118,468   87.60%  0.449   0.234       2.6   0.197  0.0820
## 4 Los A~ 35028 $118,380   98.00%  0.817   0.260       3.4   0.204  0.1000
## 5 Fairf~ 51059 $117,989   92.00% NA      NA           2.4   0.235  0.104
## 6 San M~  6081 $115,908   88.90%  0.517   0.236       2.2   0.187  0.0908
## # ... with 2 more variables: AirPol <dbl>, Crime <dbl>
```

Before the k-means calculation is carried out, there are some missing data values in the above dataset that must be dealt with. Rather than removing counties with partially missing data from the k-means clustering, NA values will be replaced with the mean value from the rest of the observations in the dataset. Although this potentially falsifies some of the data and fails to capture true variation at the county-level, this adjustment should have minor effects. As can be seen below, only four of the nine covariates have any missing values, and the highest rate for missing data is just 175 out of over 3000 counties. Thus, this mean-substitution correction is most likely inconsequential.

```
for (i in 3:11) {
  print(sum(is.na(DataSet[i])))
}
```

```
## [1] 0
## [1] 0
## [1] 58
## [1] 58
## [1] 0
## [1] 0
## [1] 0
## [1] 33
## [1] 175
```

Above, it is shown that the variables for white population, family household rates, air pollution rates, and crime levels have missing data. These NA values are substituted with the mean values shown below.

```
FillWhite <- mean(DataSet$White, na.rm=TRUE)
FillWhite
```

```
## [1] 0.8306897
```

```
DataSet$White[is.na(DataSet$White)] <- FillWhite
FillFamHous <- mean(DataSet$FamHous, na.rm=TRUE)
FillFamHous
```

```
## [1] 0.2568315
```

```
DataSet$FamHous[is.na(DataSet$FamHous)] <- FillFamHous
FillAirPol <- mean(DataSet$AirPol, na.rm=TRUE)
FillAirPol
```

```
## [1] 8.95177
```

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

```
## [1] 245.458
```

```
DataSet$Crime[is.na(DataSet$Crime)] <- FillCrime
```

Finally, before the k-means clustering occurs, the variables must be standardized. For standardization, every observation has the mean value subtracted from it, and this number is then divided by the standard deviation of that variable. Scaling helps take into account differences in measuring. For instance, previously, household income was reported as an absolute number on a scale in thousands, white other variables were reported as a rate ranging from zero to one; thus, scaling helps to correct for these differences.

```
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()
head(scaledVars)
```

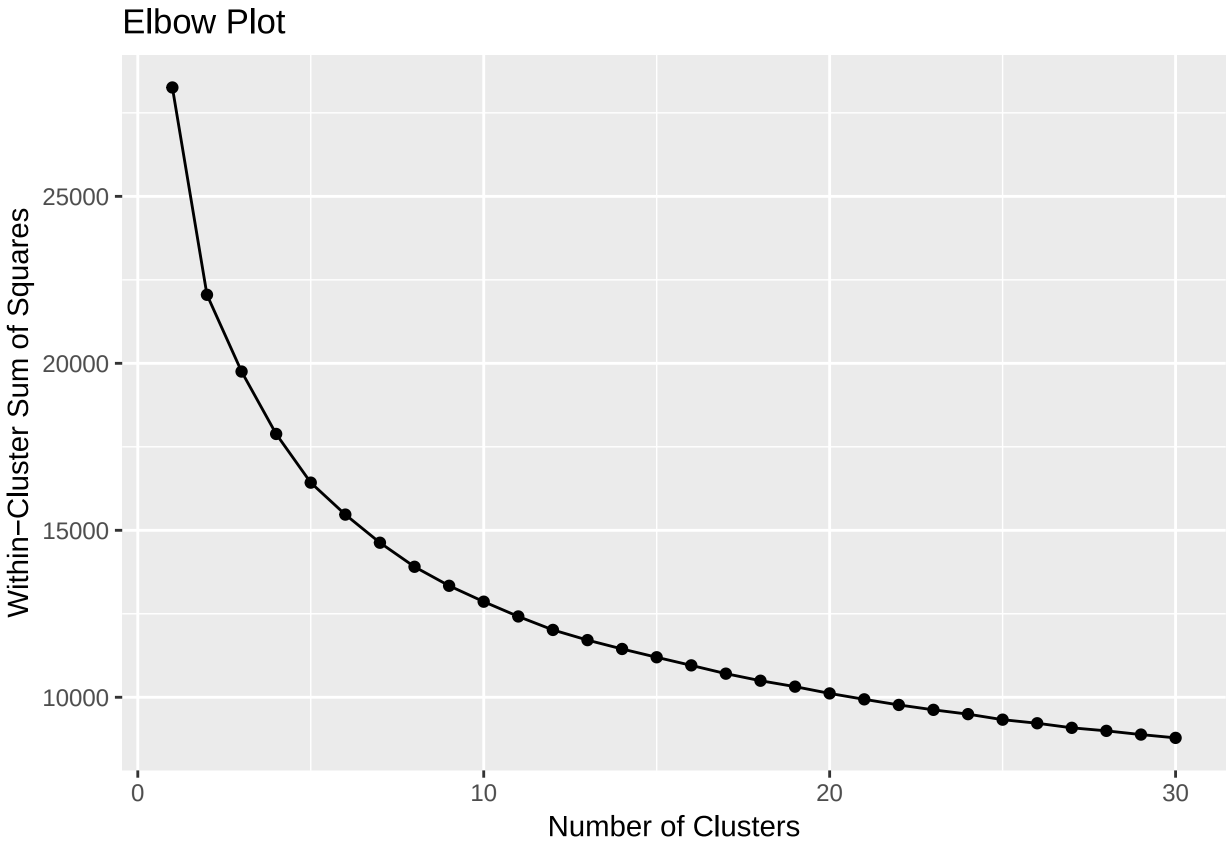
```
##   MedHousInc    HSDipl         White       FamHous  UnempRate    Obesity
## 1   6.304689 1.1280929 -1.274893e+00 -9.373492e-01 -1.0872711 -2.2115863
## 2   5.395811 1.8534946 -1.452201e-08 -2.175126e-07 -1.3535416 -0.7485624
## 3   4.991676 0.2174824 -2.276893e+00 -8.578471e-01 -1.0207035 -2.6105928
## 4   4.985156 1.8226264 -7.991424e-02  1.236453e-01 -0.4881625 -2.4554236
## 5   4.956189 0.8965818 -1.452201e-08 -2.175126e-07 -1.1538387 -1.7682457
## 6   4.802018 0.4181254 -1.871455e+00 -7.604284e-01 -1.2869740 -2.8322631
##     Smoking      AirPol       Crime
## 1 -1.784085  0.66556028 -0.86770419
## 2 -1.376462  0.72905398 -0.60091627
## 3 -2.643771 -0.09636416  0.04564432
## 4 -2.152629 -2.12816265  0.19170846
## 5 -2.041055 -0.09636416 -0.84358717
## 6 -2.401983 -0.54082008 -0.09679460
```

Next, the k-means calculation will be computed. K-means clustering takes a range of data inputs and hopes to classify data into groupings, or 'clusters,' based on the covariate inputs. Thus, for this application, k-means clustering is used in an attempt to capture variation in relative prosperity levels throughout the United States. The clustering algorithm will take each of the prosperity indicators in the dataset and attempt to form distinct groupings that exist throughout the US. For instance, perhaps the k-means clustering will reveal a cluster of counties that has relatively high income levels, but has low levels of diversity and family households. In addition, perhaps the k-means clustering will reveal a cluster of counties that is just the opposite: low income with high levels of family homes and diversity. In other words, the k-means clustering is being used to classify distinct 'quality of life' groupings that occur throughout the United States.

Obviously, there is not a distinct number of groupings that exists in the US in terms of prosperity. Thus, the k-means algorithm is carried out on any number of clusters between one and 30; this way, an elbow plot can be formed that illustrates an ideal number of clusters.

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

```r
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")
```

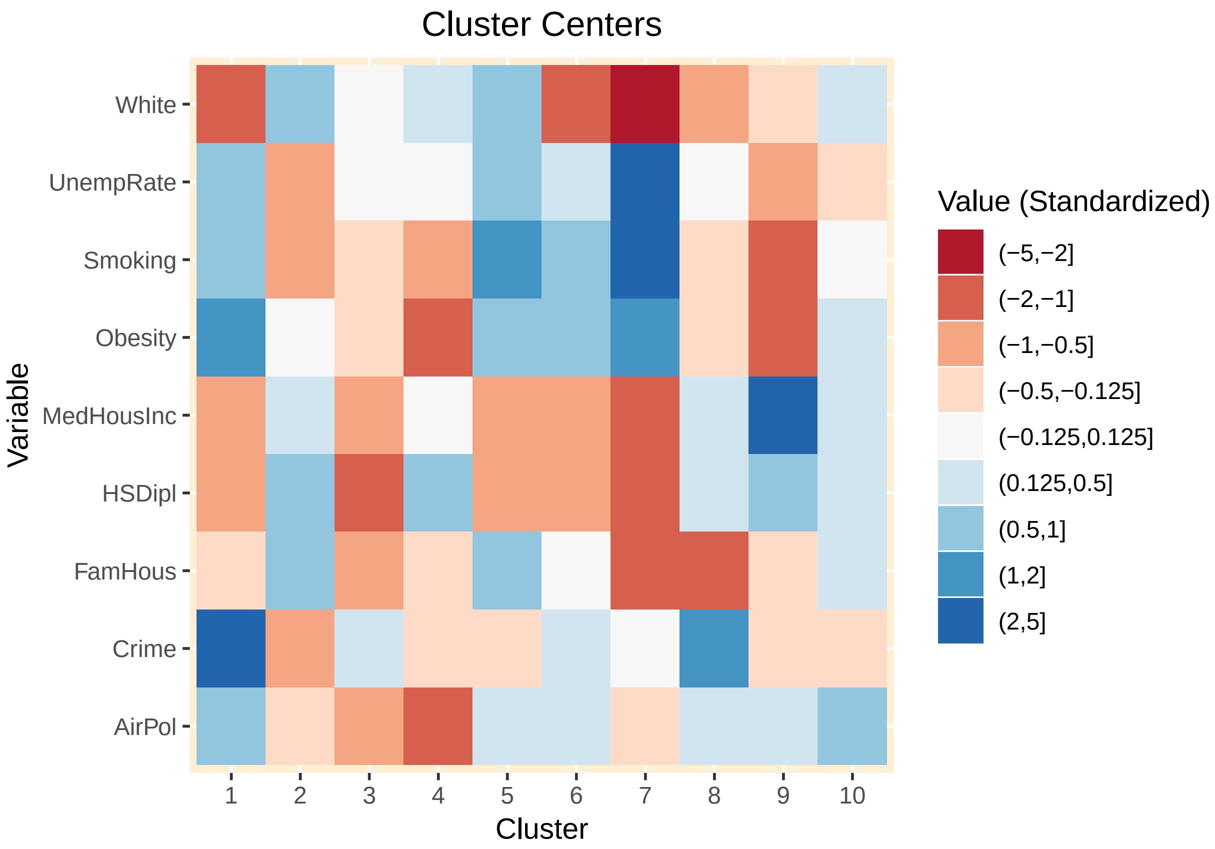


Looking at the above plot, ten clusters has been chosen as an effective number of clusters for this project. This makes sense: with less than ten clusters, the within-cluster sum of squares (or variation within clusters) can be decreased dramatically by increasing the number of clusters. After ten, this decrease in the sum of squares is not very large. It should be noted that the 'elbow' in the plot above is not very clear; thus, a number of clusters different that ten could have been chosen for this analysis. Yet, it seemed that choosing more than ten clusters would have just muddled the overall analysis by overfitting the model and making it harder to observe more general trends. Below, I complete the k-means calculation with ten clusters.

```
set.seed(650)
KMeans <- kmeans(scaledVars, 10, nstart=100, iter.max = 30)
```

Now, it is possible to investigate the clusters that arise from the k-means calculation by looking at the 'centers' of each cluster. The centers represent the average values of the observations in a cluster for each covariate. The centers are calculated and graphed below.

```r
Centers <- as.data.frame(KMeans$centers)

Centers$Cluster <- as.factor(c(1:10))

Centers <- gather(Centers, Variable, Value, -Cluster)

ggplot(Centers, aes(Cluster, Variable,

  fill=cut(Value, c(-5, -2, -1, -.5, -.125, .125, .5, 1, 2, 5)))) +

  geom_tile() + scale_fill_brewer(palette = "RdBu") +

  guides(fill=guide_legend(title="Value (Standardized)")) + ggtitle('Cluster Centers') +

  theme(plot.title = element_text(hjust = 0.5), panel.background = element_rect(fill =

  "papayawhip"))
```



From the above plot, it is clear that there are indeed distinct 'quality of life' groupings throughout the United States at the county level. Next, the distinguishing characteristics of each cluster will be discussed.

Several clusters exhibit similar overall trends. Cluster 1 represents counties that have relatively high nonwhite, diverse populations. Yet, these counties are plagued by high levels of unemployment, smoking, obesity, and crime. There are also low levels of family households, high school graduation rates, and household income. Clusters 6 and 7 are similar to cluster 1, yet cluster 6's values are centered more closely around the mean of zero, while cluster 7's values are much higher in magnitude and extreme. In addition, Cluster 5 is similar to Clusters 1, 6, and 7 but is characterized by a mostly white population, rather than the higher nonwhite populations that are visible in Clusters 1, 6, and 7. Cluster 2 is interesting because it exhibits the opposite tendencies as Clusters 1, 6, and 7.

Each of the other clusters are more distinct in nature. Cluster 3 captures the least-educated populations in the United States in terms of high school graduation rates, and this low education rate is correlated with low levels of household income, and, interestingly, air pollution. Cluster 4 seems to be a 'healthy' cluster, characterized by low levels of smoking, obesity, and air pollution, with high levels of high school graduation. Cluster 9 is what could perhaps be called the 'ideal' cluster. Cluster 9 is characterized by slightly more diverse populations; low unemployment, smoking, crime, and obesity; and high income and high school graduation rates.

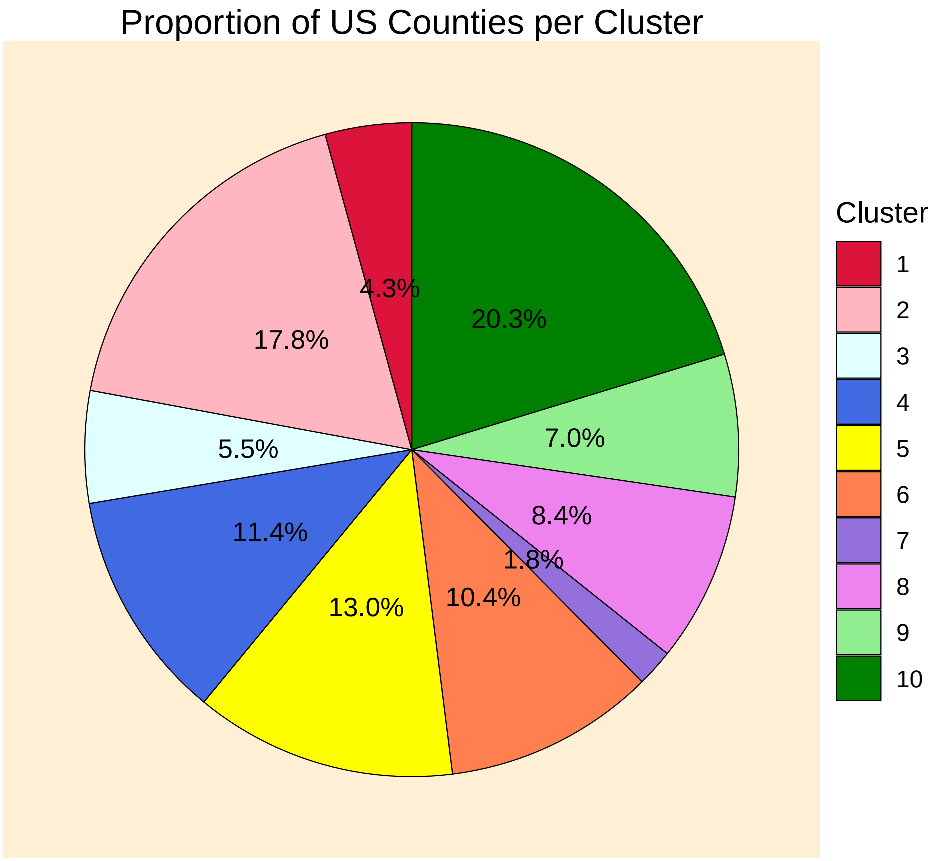
Finally, Cluster 10 represents the 'average' county in the United States. Looking at the graph, none of the covariate levels deviate far from the overall mean, except for a slightly elevated air pollution level. Cluster 8 is similar to Cluster 10 but is distinguished by its slightly higher nonwhite populations, higher crime levels, and lower levels of family housing.

Thus, it is possible to make conclusions about relative prosperity levels throughout the United States. For instance, it is clear in Clusters 1, 6, and 7 that we can see higher levels of minority populations being plagued by negative characteristics, such as high levels of smoking and obesity and low levels of high school graduation and household income.

Beyond this, we can see that some counties focus on health outcomes the most (Cluster 4), some counties have healthy environmental indicators (Cluster 3), and some counties have seemed to find a happy medium (Cluster 9). These results show that there is indeed a wide variation in relative prosperity throughout the United States at the county level.

One idea that is important to consider but has not yet been mentioned is the relative size of each cluster. It is important to note that these differences in relative prosperity are most important to consider in the context of the size of each cluster. For instance, if there were ten clusters, and one cluster occupied 90% of the counties in the USA, we could not safely assume that there are major, distinct, quality of life differences throughout the US. Thus, below, the proportion of counties per cluster is graphed.

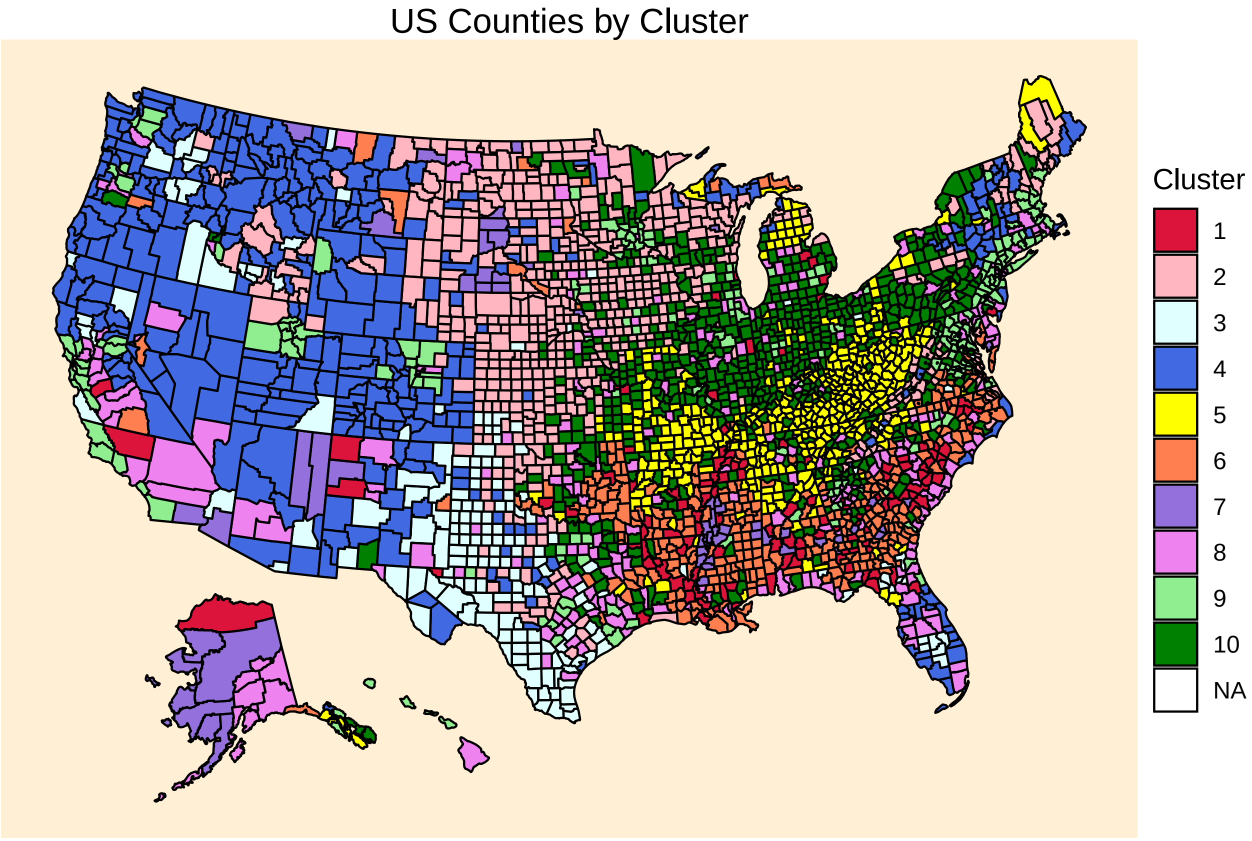
```r
DataSet$Cluster <-KMeans$cluster %>%
  as.factor()
ClusterTotals <- DataSet %>%
  group_by(Cluster) %>%
  count()
AbsTotal <- sum(ClusterTotals$n)
ClusterTotals$Prop <- percent(ClusterTotals$n / AbsTotal)
ggplot(ClusterTotals, aes("", n, fill=Cluster)) +
    geom_bar(width = 3, stat = "identity", color='black', size=.2) + coord_polar("y",
    start=0) + scale_fill_manual(values = c('#DC143C', 'lightpink', 'lightcyan', 'royalblue',
  'yellow', 'coral', 'mediumpurple', 'violet', 'lightgreen', '#008000')) +
    geom_text(aes(label=Prop), position = position_stack(vjust=.5), size=3.5) +
    ggtitle("Proportion of US Counties per Cluster") + theme_void() + theme(plot.title =
    element_text(hjust = 0.5), panel.background = element_rect(fill = "papayawhip"))
```



The above plot shows that none of the generated clusters enveloped more than 20% of the US counties, while every cluster contained at least 1.8% of the counties in the US. This is promising, as it reveals that each of the ten clusters represents a significant portion of the counties in the United States. This points toward the conclusion that there are in fact major variations in relative prosperity levels at the county level throughout the United States.

Next, it may be interesting to consider whether or not the clusters generated in the process above exhibit any geographic groupings or correlation. Below, I map the clusters onto the United States geographically.

```r
GraphData <- DataSet[ ,c(2, 12)]
GraphNames <- c("fips", "cluster")
names(GraphData) <- GraphNames
GraphData$fips <- as.character(GraphData$fips)
GraphData$fips <- str_pad(as.numeric(GraphData$fips), 5, side = "left", pad = 0)
plot_usmap(data=GraphData, values='cluster', lines='black') + theme_bw() +
  scale_fill_manual(values = c('#DC143C', 'lightpink', 'lightcyan', 'royalblue',
  'yellow', 'coral', 'mediumpurple', 'violet', 'lightgreen', '#008000')) +
  labs(title ="US Counties by Cluster") + theme_void() + theme(plot.title =
  element_text(hjust = 0.5), panel.background = element_rect(fill = "papayawhip")) +
  guides(fill=guide_legend(title="Cluster"))
```



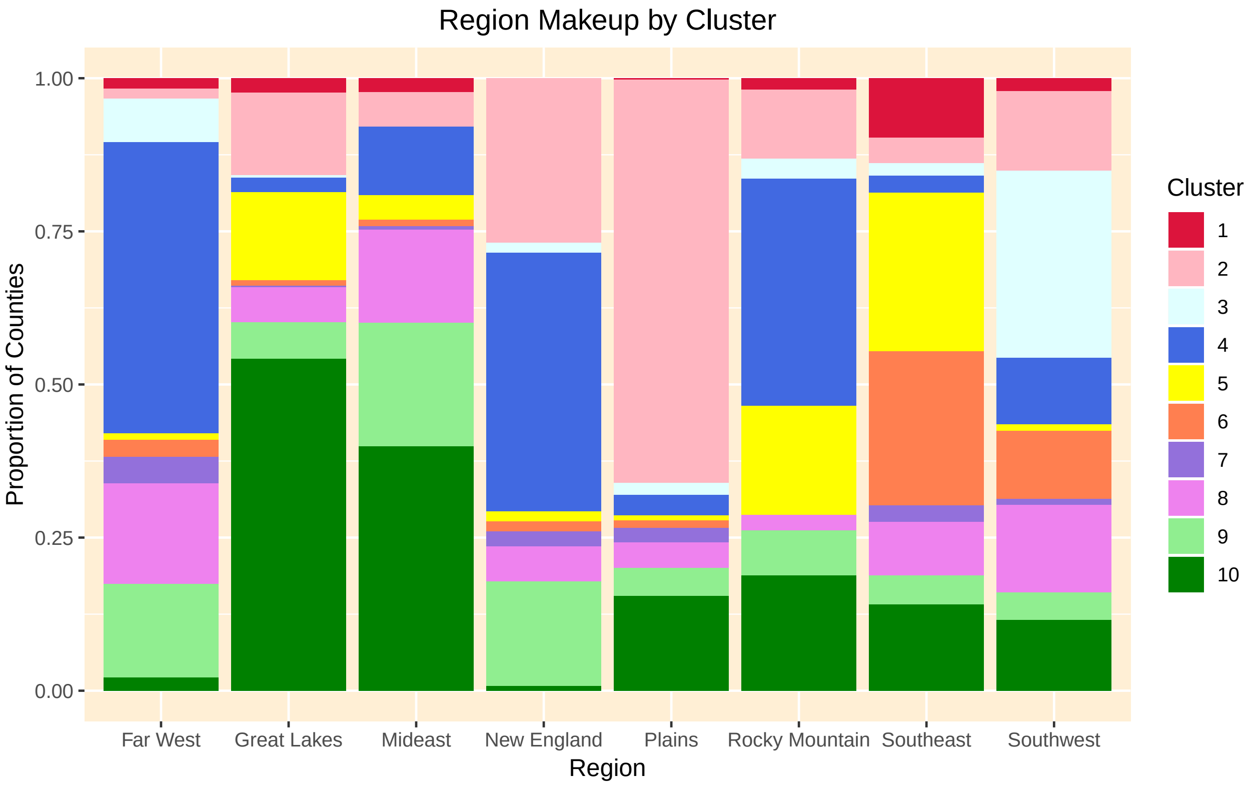
Above, it is clear that some of the clusters are overrepresented in certain geographic regions throughout the United States. However, there are also many 'outliers' that are in clusters that are completely different from the surrounding counties. A further geographic analysis by "Region" occurs below. First, a table is created listing the total number of counties per cluster per state, and then, a "Region" factor is added based on the state. Regions are based on official classifications from the Bureau of Economic Analysis.

```r
DataSet$State <- str_sub(DataSet$County, -2, -1)
StateTable <- DataSet %>%
  group_by(State, Cluster) %>%
  count()
StateTable$Region <- as.character(612L)
StateTable <- within(StateTable, Region <-
    ifelse(State %in% c('AK','CA','HI','NV','OR','WA'), 'Far West',
    ifelse(State %in% c('CO','ID','MO','UT','WY'), 'Rocky Mountain',
    ifelse(State %in% c('AZ', 'NM', 'OK', 'TX'), 'Southwest',
    ifelse(State %in% c('AL', 'AR', 'FL', 'GA', 'KY', 'LA', 'MS', 'NC',
                        'SC', 'TN', 'VA', 'WV'), 'Southeast',
    ifelse(State %in% c('IA', 'KS', 'MN', 'MO', 'NE', 'ND', 'SD'), 'Plains',
    ifelse(State %in% c('IL', 'IN', 'MI', 'OH', 'WI'), 'Great Lakes',
    ifelse(State %in% c('DE', 'DC', 'MD', 'NJ', 'NY', 'PA'),
           'Mideast', 'New England'))))))))
head(StateTable)
```

```
## # A tibble: 6 x 4
## # Groups:   State, Cluster [6]
##   State Cluster     n Region
##   <chr> <fct>   <int> <chr>
## 1 AK    1           1 Far West
## 2 AK    4           2 Far West
## 3 AK    5           2 Far West
## 4 AK    6           2 Far West
## 5 AK    7           7 Far West
## 6 AK    8          10 Far West
```

Then, below, I create a bar graph listing each region of the United States. The bars represent the total proportion of each region that is comprised of a certain cluster, and each cluster is color-coded.

```r
ggplot(StateTable, aes(Region, n, fill=Cluster)) + geom_bar(stat="Identity",
  position="fill") + scale_fill_manual(values = c('#DC143C', 'lightpink', 'lightcyan', 'royalblue',
  'yellow', 'coral', 'mediumpurple', 'violet', 'lightgreen', '#008000')) + theme(plot.title =
  element_text(hjust = 0.5), panel.background = element_rect(fill = "papayawhip")) +
  labs(title ="Region Makeup by Cluster", x = "Region", y = "Proportion of Counties")
```



Thus, we can see from the above graph that a sizeable proportion of the variation we see between counties can be observed at the regional level. For instance, we see that the Far West, New England, and the Rocky Mountain region each have a large representation from Cluster 4, which was dubbed the ‘healthy’ cluster, exhibiting low levels of smoking, obesity, and air pollution. The Great Lakes and the Mideast have large stakes in Cluster 10, the ‘average’ cluster with no significant or striking characteristics. The Great Plains can be characterized by Cluster 2, exhibiting low levels of diversity, unemployment, smoking, obesity, and crime, with high levels of family households, high school graduation rates, and household income. The Southeast has large representations from Clusters 5 and 6, which displayed trends that were opposite those in Cluster 2. Finally, the Southwest engulfs Cluster 3, which was marked by lower income levels and the lowest levels of high school graduation.

Yet, interestingly, for almost every region excluding the Plains, the largest cluster representation is only about half of the total region. Thus, it is difficult to make generalizations such as those above for entire regions at a time. Rather, it is clear that there are indeed significant changes in quality of life at the county level that are not determined by region. This is powerful because it shows that the results of the k-means calculation, which argued that quality of life varies widely from county to county throughout the United States, is not something that can be easily determined by looking at the geographic location of a county.

Thus, this project has been effectively able to investigate and illustrate relative differences in prosperity that exist throughout the United States, and these results have greater implications regarding inequality and inequal access to positive opportunities within the United States. For instance, is it fair that people who live in Clusters 1, 6, and 7 have almost objectively worse qualities of life in this model than those who live in Cluster 9?

Of course, when considering these implications, there are some limitations to the model that must be taken into account. For instance, missing data for some counties was accounted for by substituting overall population means, potentially obscuring data for these counties. In addition, the graphs in this project do not take into account the relative populations of each county in the United States. Thus, the actual number of citizens residing in each cluster may be different than what appears in this project. Finally, the measures of prosperity used in this project are not all-encompassing, as data that was readily available was utilized for this paper; thus, using different measures could have yielded different results and different clusters. Taking all of this into account, however, this project is still able to provide useful insights about differences in quality of life that exist throughout the United States at the county and regional level.