Clustering Contractors with Major NAICS Codes

Nick Kallfa

February 10, 2018

Introduction

In this notebook I am going to take a step back by simplifying my analysis. Up until this point, the data I have been using included all companies and 6-digit NAICS codes used in FY2016 FPDS data. That's a lot of information when 2-digit NAICS codes could just as easily suffice. For example, a NAICS code that begins with "54" refers to any activity that falls under professional, scientific, and technical services. NAICS code "541214" falls under this broad category too, but refers more specifically to payroll services. On the other hand "541370" refers to activity that is classified as surveying and mapping (except geophysical) services. Rather than using the full 6 digit NAICS codes let's see if we can group companies using the major (i.e. 2 digit) NAICS codes. Before we perform the clustering, let's describe the data sets I have. They can also be found in the data folder on my GitHub repository.

The Data

I have saved a list of all 116,328 companies along with their global DUNS number in a CSV. Each row in the file is one company in the FY2016 FPDS data. The table below is a preview of the first 5 companies listed in this data set. We'll use this to match the names of the companies with the clusters we find in the data.

Table 1: First 5 companies (out of 116,328) in our data set

parentdunsnumber	vendorname
001000363	LOUIS M. GERSON CO., INC.
001001676	PRECISION SYSTEMS INC
001001767	ECKEL INDUSTRIES, INC.
001002054	BOMAS MACHINE SPECIALTIES INC
001002781	TRIANGLE ENGINEERING, INC.

Table 2 is just a description of what all the codes represent. There are 24 major NAICS codes in FY2016. Keep in mind that codes may change (new codes added, removed, modified) so this table represents a snapshot of the codes at one point in time.

Table 2: Major (2D) NAICS codes

Sector	Code	Column_Code
Agriculture; Forestry; Fishing; and Hunting	11	V1
Mining; Quarrying; Oil; and Gas Extraction	21	V2
Utilities	22	V3
Construction	23	V4
Manufacturing	31	V5
Manufacturing	32	V6
Manufacturing	33	V7
Wholesale Trade	42	V8
Retail Trade	44	V9
Retail Trade	45	V10
Transportation and Warehousing	48	V11
Transportation and Warehousing	49	V12
Information	51	V13
Finance and Insurance	52	V14
Real Estate; Rental; and Leasing	53	V15
Professional; Scientific; and Technical Services	54	V16
Management of Companies and Enterprises	55	V17
Administrative; Support; Waste Management and Remidiation Services	56	V18
Educational Services	61	V19
Health Care and Social Assistance	62	V20
Arts; Entertainment; and Recreation	71	V21
Accomodation and Food Services	72	V22
Other Services	81	V23
Public Administration	92	V24

In the data directory of my GitHub repo I have the raw data used prior to clustering. Since the data is sparse I have stored it in a special format. Each row of the file "indices_values_dollars_obligated_2d_naics.csv" gives the row and column number of every non-zero value. Table 3 shows the top 5 rows of the data.

Table 3: Preview of the data in its original form

row	col	values
49	1	3285.48
92	1	7469.17
262	1	40277.00
263	1	18840.00
273	1	15970.03
331	1	2900.35

However, before we perform the clustering I pre-process my data slightly. I have written a few custom functions that will do this automatically. You'll find these functions in the R directory.

First, we transform all values to be positive by taking their absolute value. Then, for each row, we divide every value by the sum of numbers in that row. Now each row represents a percentage of revenue a company generated under a major NAICS code. Next, we cap from below by setting any value less than 0.1 (10%) to zero. Lastly, we use the sparseMatrix function from the Matrix package to convert our data into an object we can actually work with. Then I convert the sparseMatrix to a matrix and then to a data frame. Now we have our data in a format that we can use for clustering.

```
# Create copy of original data set
df = df_original

# Step 1: Make all values positive
df = Transform_Negative_Values(df,
                               type = "absolute value")
# Step 2: Divide each value by the sum of numbers in its row
df = Transform_to_Percentage(df)
# Step 3: Set values below 0.1 to zero
df = Transform_by_Capping(df = df,
                          lower_bound = 0.1,
                          upper_bound = 0.9,
                          type = "Below")

# Step 4: Convert data to a data frame
matrix = sparseMatrix(i = df$row, j = df$col, x = df$values)
matrix = as.matrix(matrix)
df_for_clustering = as.data.frame(matrix)
```

Before we perform the clustering let's take a look at the data we are going to use. The table below shows the first 10 rows and columns of the data. Keep in mind, there are 24 columns, but I am only showing 10 so that it prints nicely.

Table 4: First 10 rows of our data prior to clustering. Each column corresponds to a major (2D) NAICS code. Only 10 of the 24 features are shown.

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
0	0	0	0	0.84	0.00	0.16	0	0	0
0	0	0	0	0.00	0.00	1.00	0	0	0
0	0	0	0	0.00	0.00	1.00	0	0	0
0	0	0	0	0.00	0.00	1.00	0	0	0
0	0	0	0	0.00	0.00	1.00	0	0	0
0	0	0	0	0.27	0.16	0.00	0	0	0
0	0	0	0	0.00	0.00	1.00	0	0	0
0	0	0	0	0.00	0.00	0.62	0	0	0
0	0	0	0	0.00	0.00	1.00	0	0	0
0	0	0	0	0.00	0.00	1.00	0	0	0

We can take a look at the dimensions of our data. We have 116,328 companies in our data set. These are the objects which we want to cluster. We will use 24 major NAICS codes as features to distinguish companies from one another. So we can see that our data for clustering has 116,328 rows and 24 columns. Rows represent companies and columns represent major NAICS codes.

Table 5: Number of instances and features of our data

Table	Rows	Columns
Companies	116328	2
Data for Clustering	116328	24
NAICS	24	3

Clustering

Finally, we're ready to do some clustering! I'll keep it simple and use K-Means clustering with 13 groups. Not included in this notebook was how I chose 13 groups. I performed K-Means clustering 29 times changing the number of groups each time from 2 to 30. Then using the elbow method I graphed the total within-cluster sum of squares output from base R's kmeans function for each iteration. K-Means with 13 groups was the point where the total within-cluster sum of squares decrease slowed. This is the point where adding additional groups only gives us marginal improvement.

```
# Set seed
set.seed(123)

# Perform K-Means clustering with 13 groups
kmeans13 = kmeans(x = df_for_clustering, centers = 13, nstart=50, iter.max = 15)
```

Cluster Size

Before we dive into each cluster to see what they are made of let's see how many companies are in each group. The largest group is cluster 3 with 24,449 companies followed by cluster 9 with 21,275 companies. Cluster 4 is the smallest with 2,168 companies. Let's take a look at cluster 4 and see if we can find out what type of companies are in that group.

Table 6: Size of each cluster

Cluster	Number of Companies
1	10602
2	14236
3	24449
4	2168
5	6290
6	12081
7	3218
8	4336
9	21275
10	3746
11	3897
12	7035
13	2995

Cluster 4

In the two tables below we see some information about the companies in cluster 4. The first table gives the first 5 rows of the clustering data with the name of the vendor attached as a column. Only two of the twenty-four features are shown, but we should noticed that the first feature has non-zero values whereas feature 2 is all zeroes.

The second table below gives the sum of the feature columns for companies in cluster 4. Most of the companies in cluster 4 are involved in agriculture, forestry, fishing and hunting.

Table 7: First 5 rows of companies in cluster 4. 2 of the 24 features are also shown.

vendorname	cluster	V1	V2
WHITE MOUNTAIN LUMBER COMPANY, INC.	4	1	0
WEYERHAEUSER NR COMPANY	4	1	0
MYERS EXCAVATION, INC	4	1	0
HERITAGE RANCH LLC	4	1	0
VIGIL, RAYMOND	4	1	0

Table 8: Primary industry associated with cluster 4.

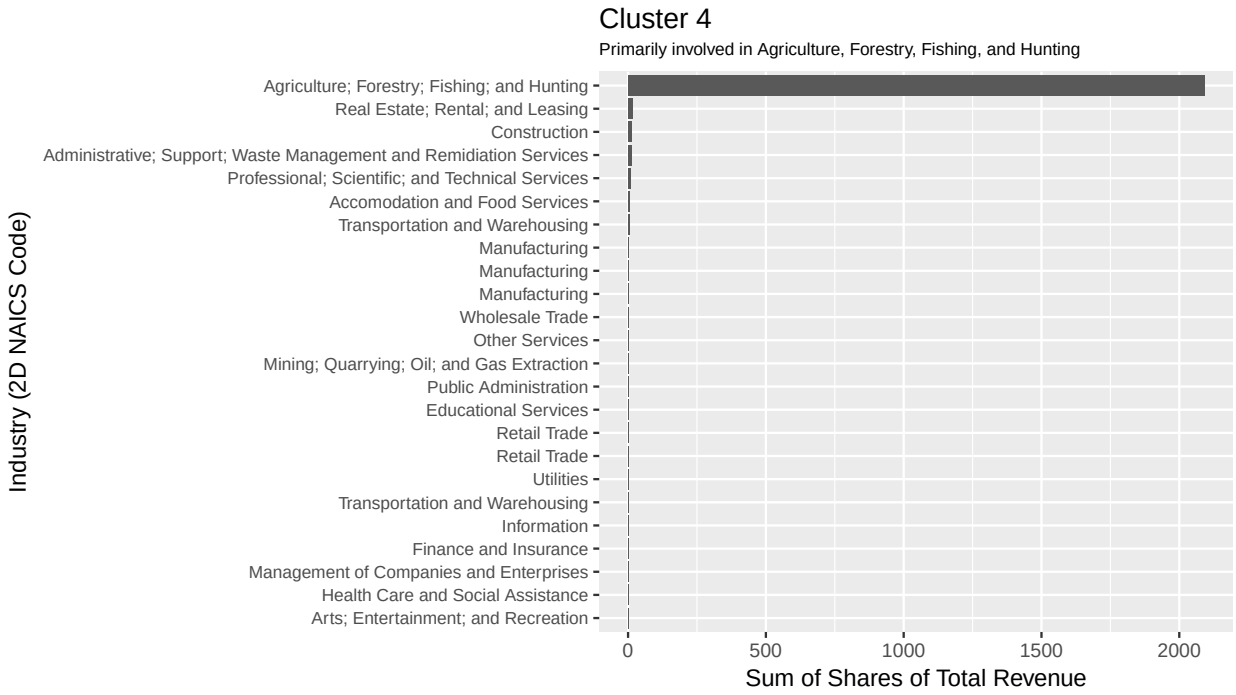
Sector	Column_Sums
Agriculture; Forestry; Fishing; and Hunting	2090.79
Real Estate; Rental; and Leasing	14.91
Construction	11.74
Administrative; Support; Waste Management and Remidiation Services	10.96
Professional; Scientific; and Technical Services	8.49
Accomodation and Food Services	5.62
Transportation and Warehousing	4.18
Manufacturing	3.34
Manufacturing	2.54
Manufacturing	2.31
Wholesale Trade	1.85
Other Services	0.98
Mining; Quarrying; Oil; and Gas Extraction	0.97
Public Administration	0.49
Educational Services	0.48
Retail Trade	0.31
Retail Trade	0.20
Utilities	0.14
Transportation and Warehousing	0.00
Information	0.00
Finance and Insurance	0.00
Management of Companies and Enterprises	0.00
Health Care and Social Assistance	0.00
Arts; Entertainment; and Recreation	0.00

We can plot the data in the table above. That plot is given below.

```
ColumnSums$Code = factor(ColumnSums$Code, levels = rev(ColumnSums$Code))
g4 = ggplot(data = ColumnSums,
       mapping = aes(x = Code,
                     y = Column_Sums)) +
         geom_bar(stat="identity") +
         coord_flip() +
         scale_x_discrete(breaks = ColumnSums$Code,
                          labels = c("11" = "Agriculture; Forestry; Fishing; and Hunting",
                                     "53" = "Real Estate; Rental; and Leasing" ,
                                     "23" = "Construction",
                                     "56" = "Administrative; Support; Waste Management and Remidiatio
```

```
                                        "54" = "Professional; Scientific; and Technical Services" ,
                                        "72" = "Accomodation and Food Services" ,
                                        "48" = "Transportation and Warehousing",
                                        "32" = "Manufacturing",
                                        "31" = "Manufacturing",
                                        "33" = "Manufacturing",
                                        "42" = "Wholesale Trade",
                                        "81" = "Other Services",
                                        "21" = "Mining; Quarrying; Oil; and Gas Extraction",
                                        "92" = "Public Administration",
                                        "61" = "Educational Services",
                                        "45" = "Retail Trade",
                                        "44" = "Retail Trade",
                                        "22" = "Utilities",
                                        "49" = "Transportation and Warehousing",
                                        "51" = "Information",
                                        "52" = "Finance and Insurance",
                                        "55" = "Management of Companies and Enterprises",
                                        "62" = "Health Care and Social Assistance",
                                        "71" = "Arts; Entertainment; and Recreation")
                            ) +
  labs(title = "Cluster 4",
       subtitle = paste("Primarily involved in ", gsub(pattern = ";",
                                                       replacement = ",",
                                                       x = ColumnSums$Sector[1],
                                                       fixed = TRUE),
                        sep = ""),
       x = "Industry (2D NAICS Code)",
       y = "Sum of Shares of Total Revenue") +
  theme(axis.text.y = element_text(size = 8),
        plot.subtitle = element_text(size = 8))

plot(g4)
```



Graphing it makes it clear that the companies in cluster 4 are almost exclusively involved in agriculture, forestry, fishing, and hunting. Now, let's produce this same plot for all the remaining clusters.

I used the code below to to do all of this in one for loop. First, I subset the data by collecting all features of companies in each cluster. Then I calculated the column sums for each of the 24 features and joined that to the NAICS code description table that we looked at earlier. In each iteration I saved the data as one element of a list.

```
# List for storing data on each cluster
output = list()

# Loop through each cluster
for (i in 1:13) {

  # Skip the 4th cluster since we did that already
  if (i == 4){
    output[[i]] = 0 # Need to put something in its place so just set the list element to zero
    next
  }
  # For each cluster calculate the sum of total revenue shares in each major NAICS code
  # to get an idea of what type of companies are in each cluster
  else {

    # Subset data for the given cluster
    cluster = clusters[clusters$cluster == i,]

    # Calculate the sum of each column from column 4 to column 27
    temp = data.frame(Column_Sums = colSums(cluster[,4:27]))

    # Get row names and then rename the rows
    temp$NAICS_2D = rownames(temp)
    rownames(temp) = 1:nrow(temp)
```

```
    # Join with NAICS code description table
    temp = left_join(x = temp,
                           y = naics_2D,
                           by = c("NAICS_2D" = "Column_Code"))
    # Only select columns we need for display
    temp = temp[,c(4,2,3,1)]
    # Arrange rows in decreasing order of colum sums
    temp = arrange(temp, desc(Column_Sums))
    # Store data in list for output
    output[[i]] = temp
  }
}
```

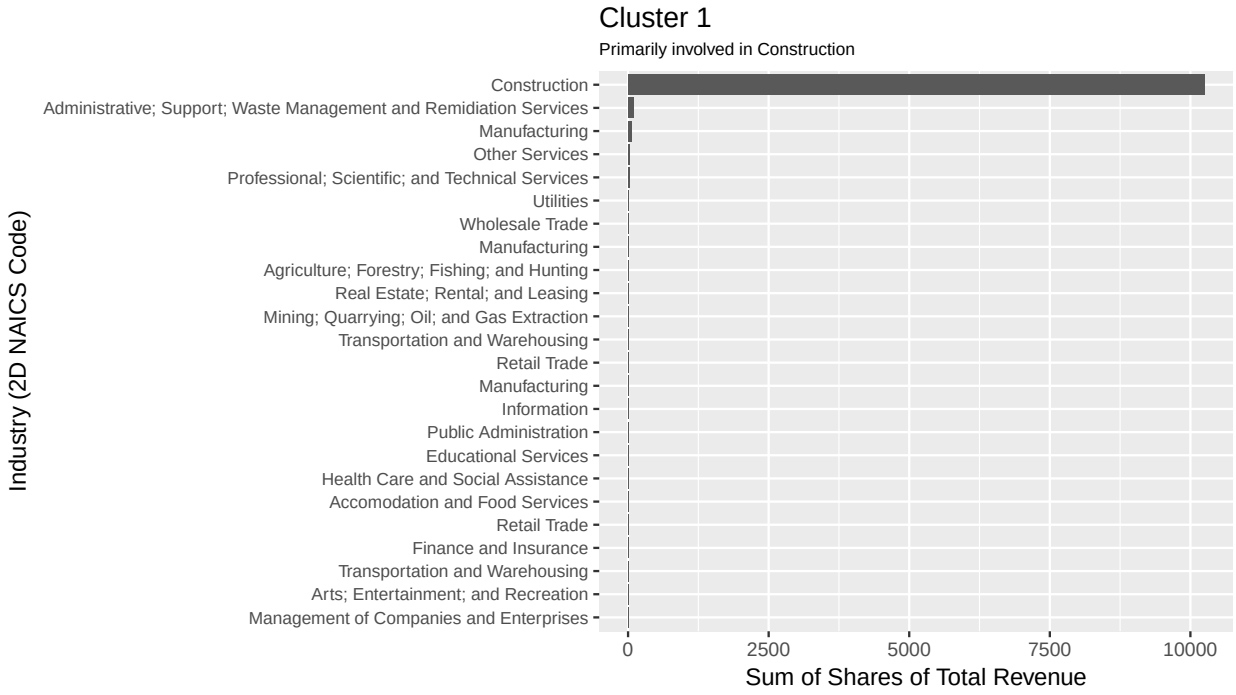
The rest of the plots in this notebook show the major NAICS codes associated with each cluster. We have already seen that cluster 4 is mostly companies involved in agriculture, forestry, fishing, and hunting. As we can see below, cluster 1 is made up of construction companies.

Cluster 1

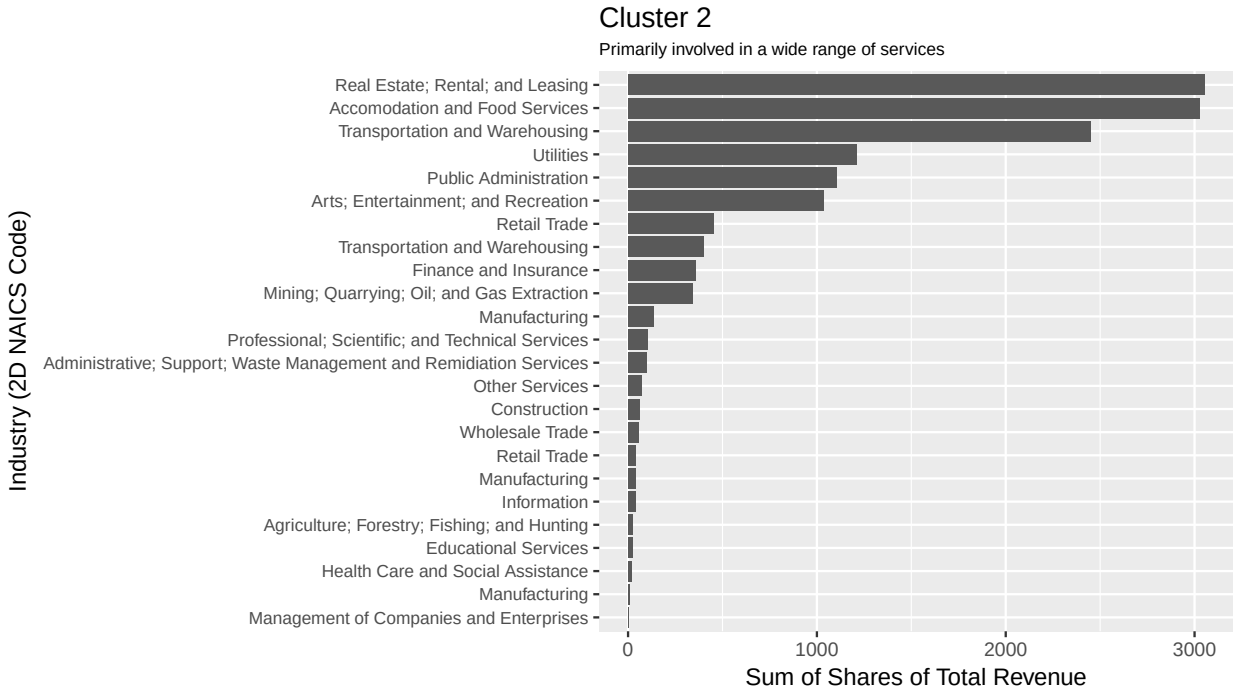
```
output[[1]]$Code = factor(output[[1]]$Code, levels = rev(output[[1]]$Code))
g1 = ggplot(data = output[[1]],
       mapping = aes(x = Code,
                     y = Column_Sums)) +
         geom_bar(stat="identity") +
         coord_flip() +
         scale_x_discrete(breaks = output[[1]]$Code,
                          labels = c("11" = "Agriculture; Forestry; Fishing; and Hunting",
                                     "53" = "Real Estate; Rental; and Leasing" ,
                                     "23" = "Construction",
                                     "56" = "Administrative; Support; Waste Management and Remidiatio
                                     "54" = "Professional; Scientific; and Technical Services" ,
                                     "72" = "Accomodation and Food Services" ,
                                     "48" = "Transportation and Warehousing",
                                     "32" = "Manufacturing",
                                     "31" = "Manufacturing",
                                     "33" = "Manufacturing",
                                     "42" = "Wholesale Trade",
                                     "81" = "Other Services",
                                     "21" = "Mining; Quarrying; Oil; and Gas Extraction",
                                     "92" = "Public Administration",
                                     "61" = "Educational Services",
                                     "45" = "Retail Trade",
                                     "44" = "Retail Trade",
                                     "22" = "Utilities",
                                     "49" = "Transportation and Warehousing",
                                     "51" = "Information",
                                     "52" = "Finance and Insurance",
                                     "55" = "Management of Companies and Enterprises",
                                     "62" = "Health Care and Social Assistance",
                                     "71" = "Arts; Entertainment; and Recreation")
                         ) +
  labs(title = "Cluster 1",
```

```
      subtitle = paste("Primarily involved in ", gsub(pattern = ";",
                                                      replacement = ",",
                                                      x = output[[1]]$Sector[1],
                                                      fixed = TRUE),
                       sep = ""),
      x = "Industry (2D NAICS Code)",
      y = "Sum of Shares of Total Revenue")  +
  theme(axis.text.y = element_text(size = 8),
        plot.subtitle = element_text(size = 8))

plot(g1)
```

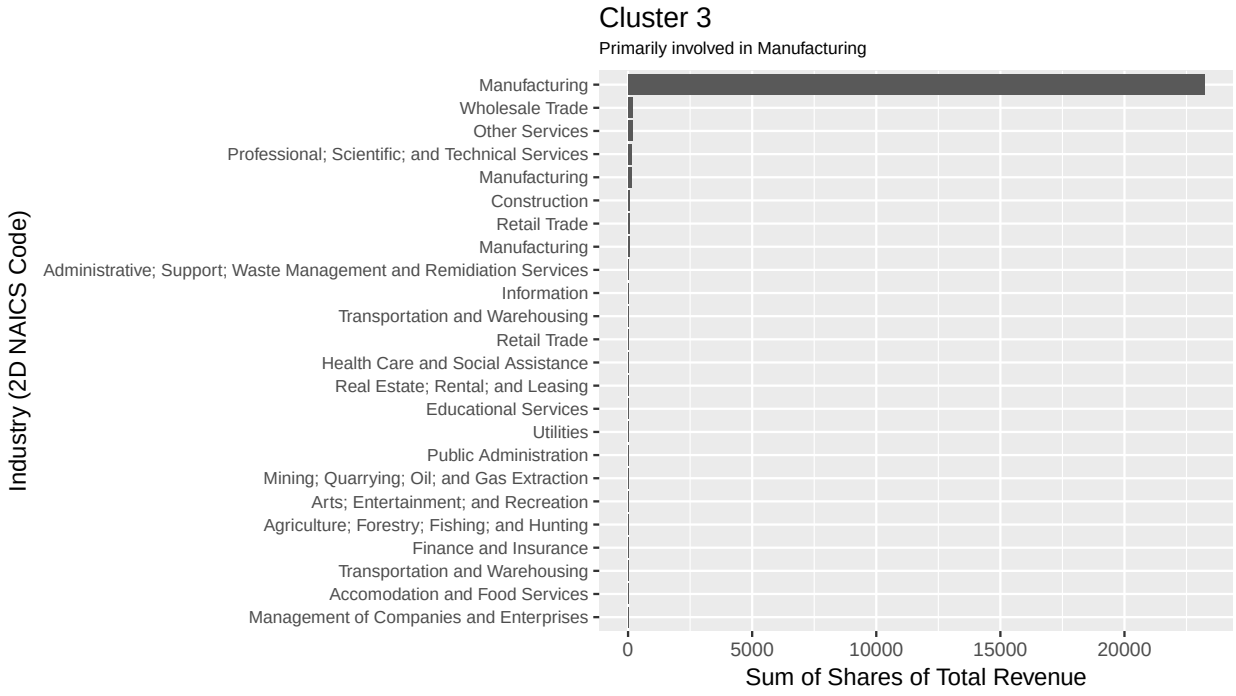


Cluster 2



Cluster 2 is interesting because it seems to be one of the only clusters that included companies that are involved in a wide range of major NAICS codes.

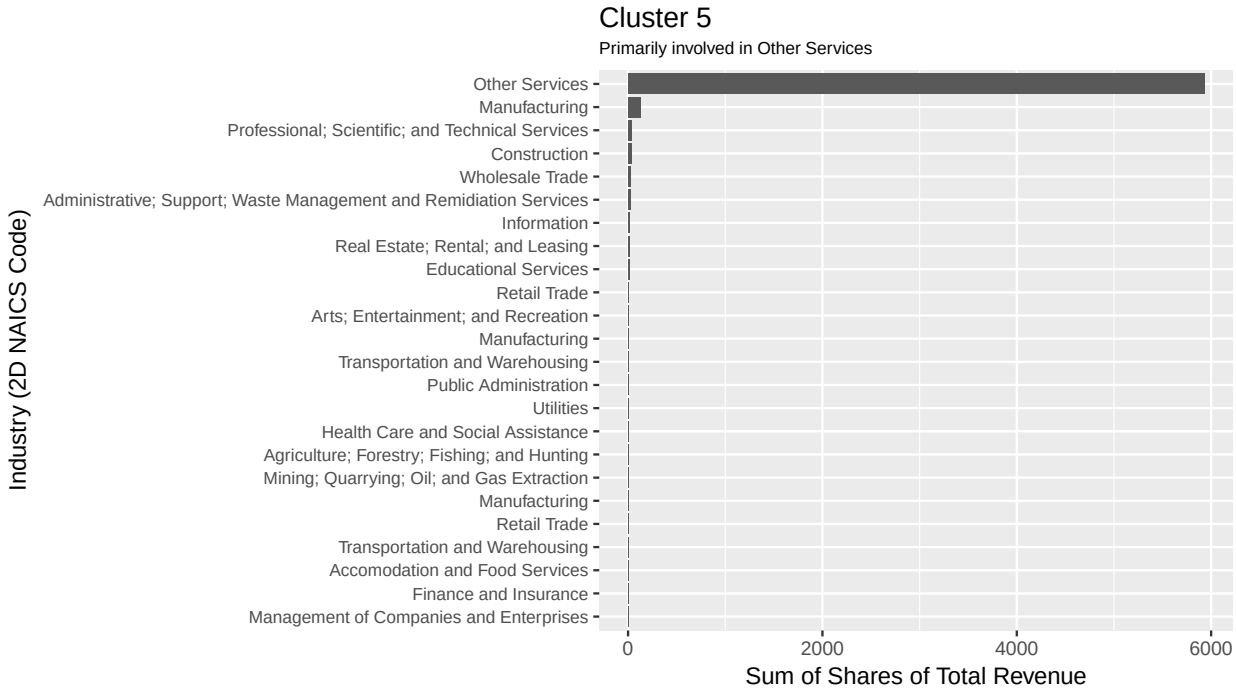
Cluster 3



Cluster 3 contains companies involved in manufacturing. Keep in mind there are three different major NAICS

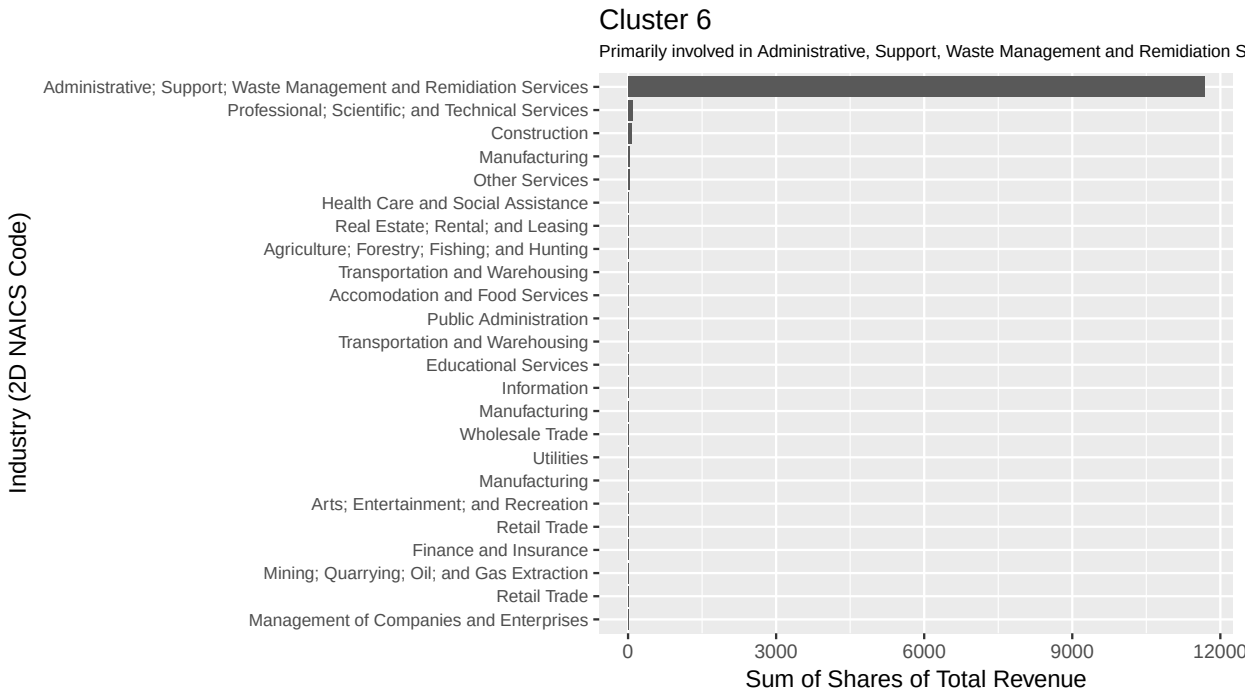
codes for manufacturing (31,32,33) so we will see later that there are other clusters primarily involved in manufacturing. I'm curious if there are any major differences between these different codes for manufacturing.

Cluster 5



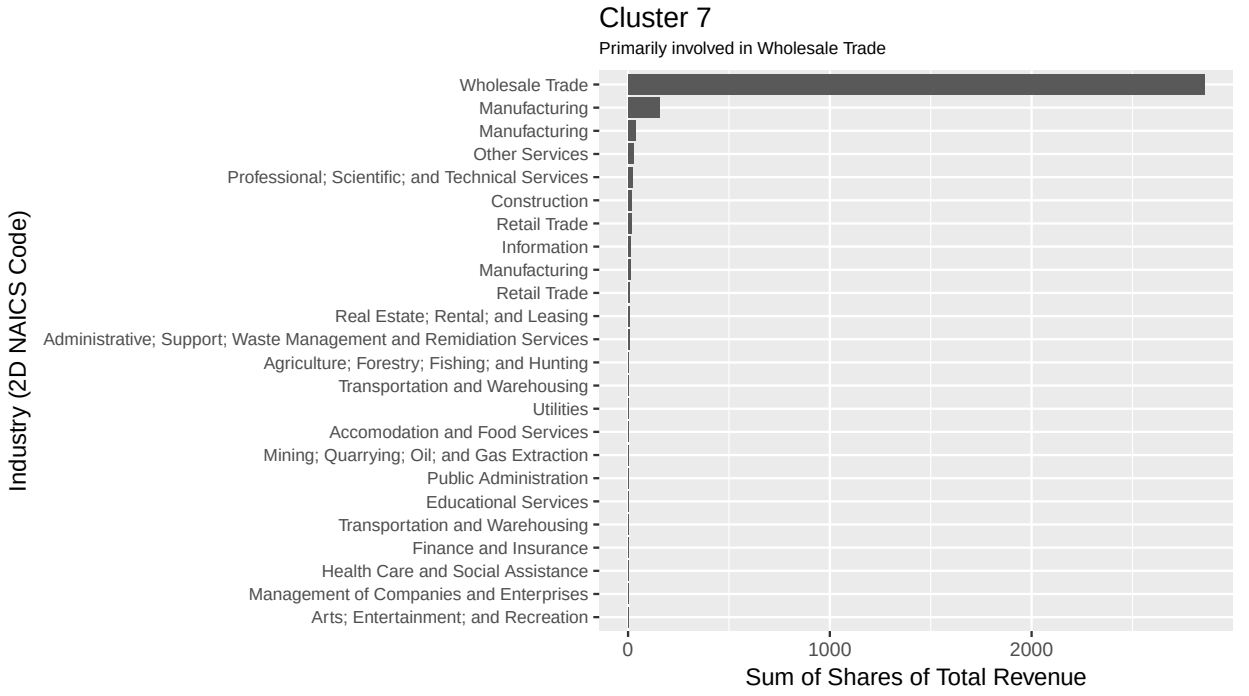
Cluster 5 consists of companies in the mysterious other services category.

Cluster 6



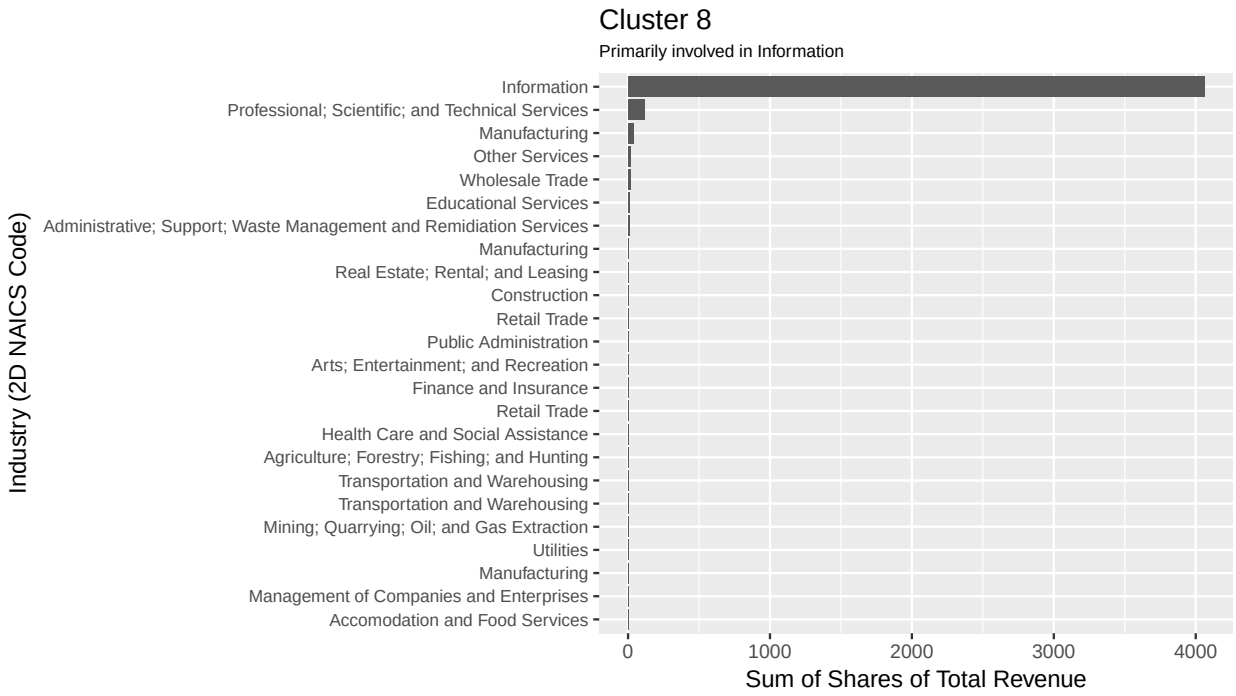
Cluster 6 includes companies involved in administrative, support, waste management, and remidiation services.

Cluster 7



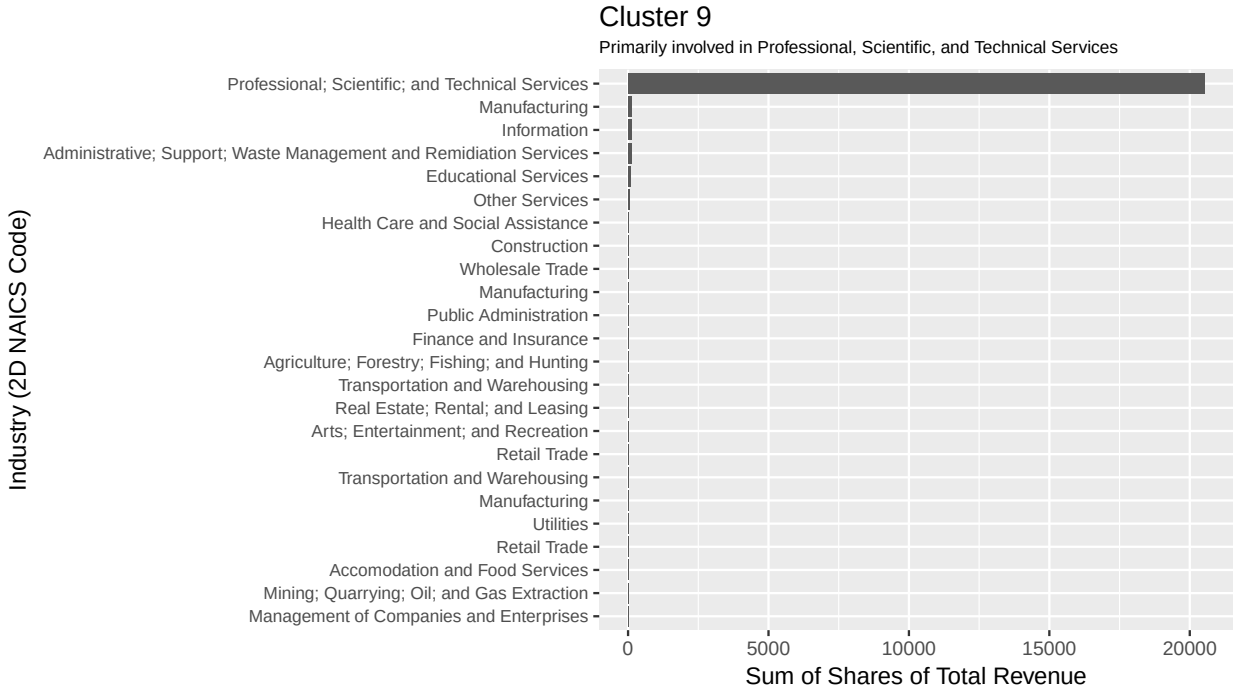
Cluster 7 includes companies involved in wholesale trade.

Cluster 8



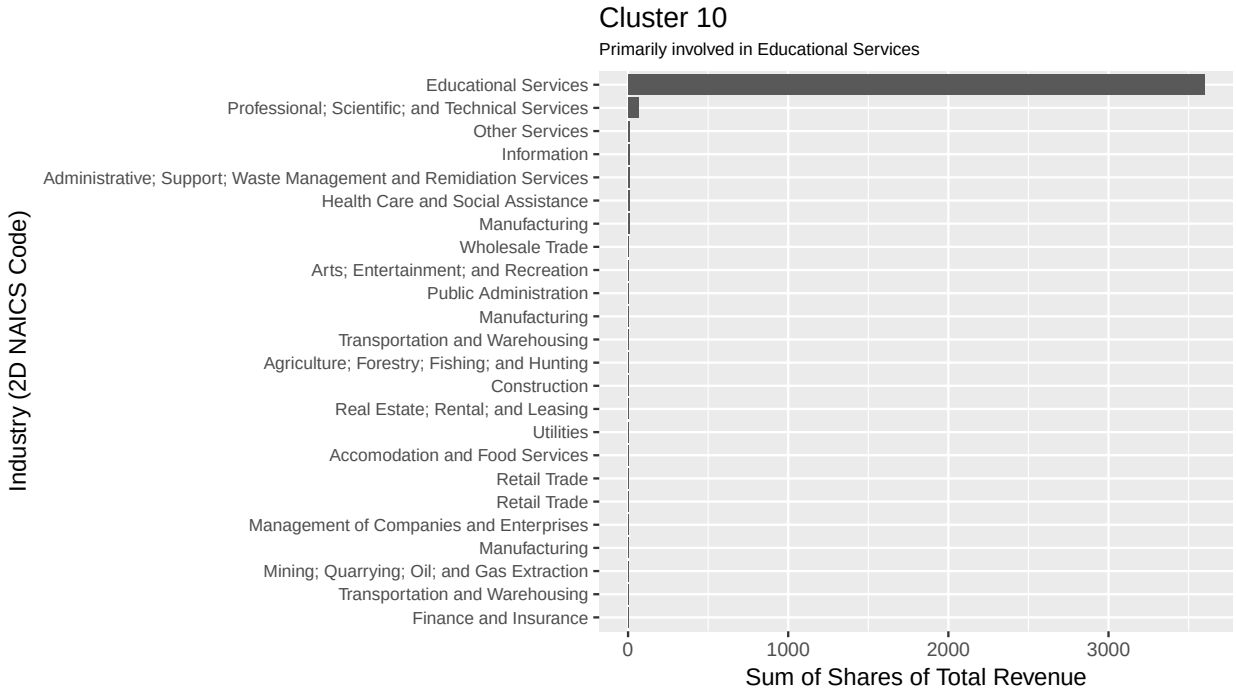
Cluster 8 includes companies involved in information services.

Cluster 9



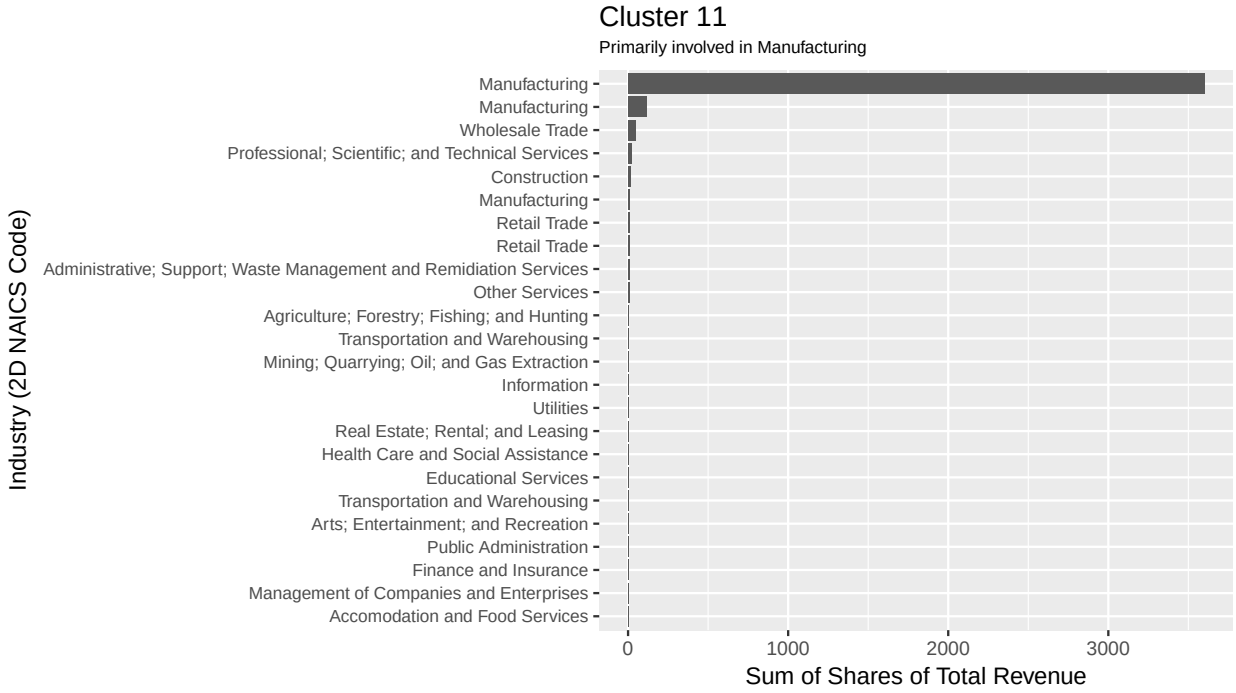
Cluster 9 includes companies involved in professional, scientific, and technical services.

Cluster 10



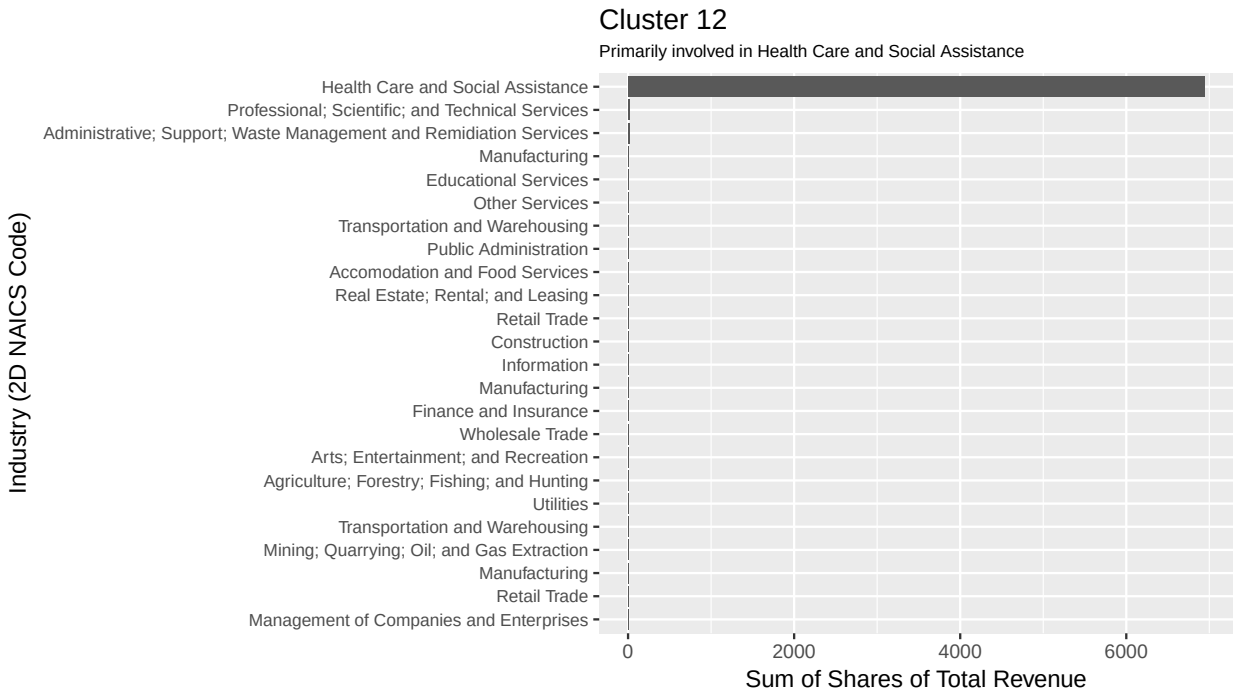
Cluster 10 includes companies involved in educational services.

Cluster 11



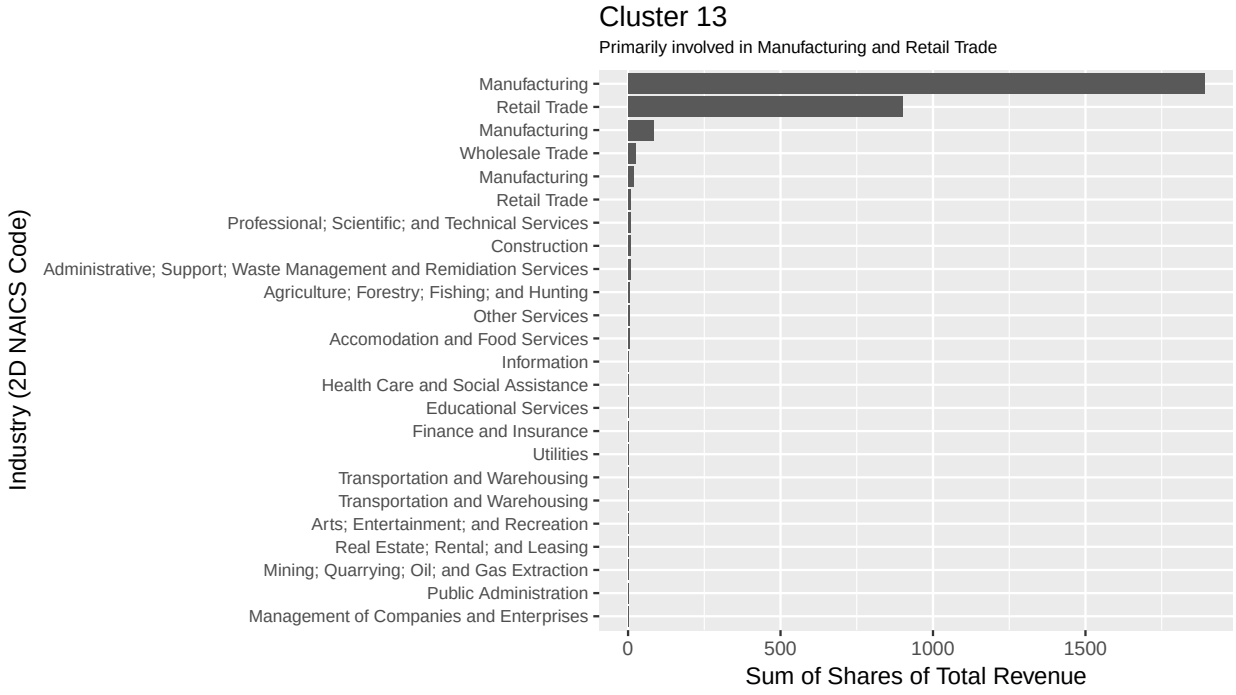
Cluster 11 is a second cluster that includes companies involved in manufacturing.

Cluster 12



Cluster 12 includes companies involved in health care and social assistance.

Cluster 13



Lastly, cluster 13 includes companies involved in manufacturing (again!) and retail trade.

Next Steps

It's pretty clear that k-means found groups among the companies using the 24 major NAICS codes. I benefited by looking at only a small set of features this time rather than all 6 digit NAICS codes which gives me 1,204 features. I was surprised to see that most companies receieved revenue from just one major industry. Aside from cluster 13 (manufacturing & retail trade) and cluster 2 (real estate, accomodation & food services, transportation & warehousing, utilities, public administration, arts) most clusters consisted of companies receiving a large share of their revenue almost exclusively from one industry. Prior to this I thought we'd see more overlap. For example, I thought that companies engaged in professional services might also offer information services. This doesn't usually seem to be the case.

Up until now I've been clustering on the entire data set using all 6 digit NAICS codes. This has made things difficult because (1) there are a lot of features and (2) a lot of companies making some clustering techniques break down when I use them. It's just too computationally expensive to do. One idea that this analysis has spawned is that it might be easier to first cluster companies at a high level using the major NAICS codes. Then, within each of these clusters, I perform a separate cluster analysis to determine more detailed characteristics for distinguishing between companies. Doing this top-down clustering would have a couple advantages:

1. Reduce the number of companies I am clustering
2. Reduce the number of features I am using

For example, suppose we find about 20,000 companies are broadly involved in professional, scientific, and technical services. However, maybe some of those companies specialize only in legal services whereas others specialize in marketing & PR services. Maybe some companies are a jack of all trades and do a little of everything. I don't really know what I expect to see, but this approach could be worth pursuing.

Before I finish a couple other low-hanging fruit things I'd like to take a look at are:

1. What is the average number of major NAICS codes a company uses? And for each major NAICS codes, what is the average number of 6-digit NAICS code a company uses.
2. What are other cluster evaluation metrics? Up until now I've been relying on the elbow method, but it's probably better to choose the number of clusters using multiple criteria.