Homework #8

Jonathan Beaubien
Math 381 - Discrete Modeling
University of Washington

March 14th, 2021

Multidimensional Scaling

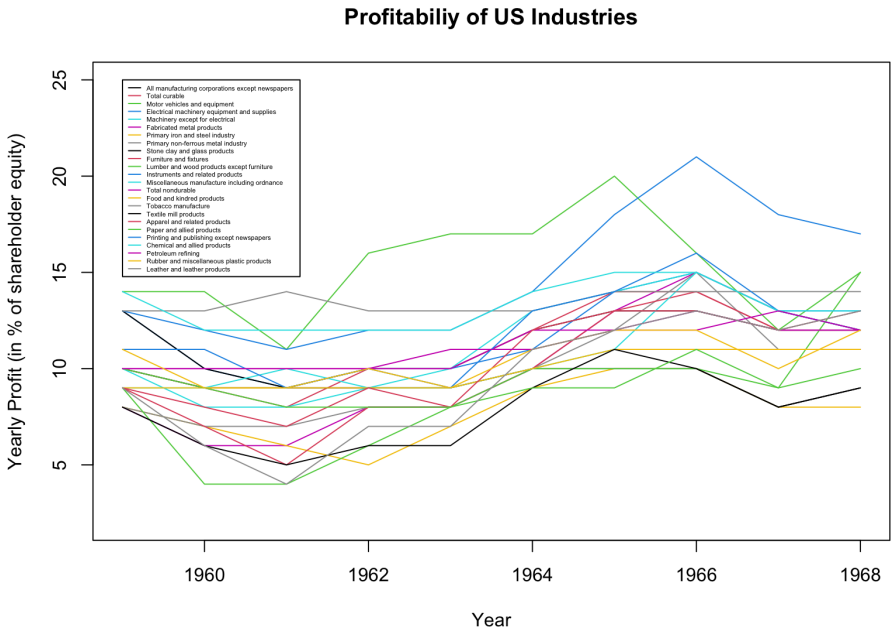
In this write-up I will analyze the profitability of various American industries over the period of 1959 to 1968 by using multidimensional scaling.
The data the I will be using is from the website: https://people.sc.fsu.edu/ jburkardt/-datasets/hartigan/hartigan.html under the file name file41.txt.
The industries that are included are various industries coded by the NAICS (North American Industry Classification System).
You can see all the industries that are tracked on the linked website.
The data is described by the author as this:

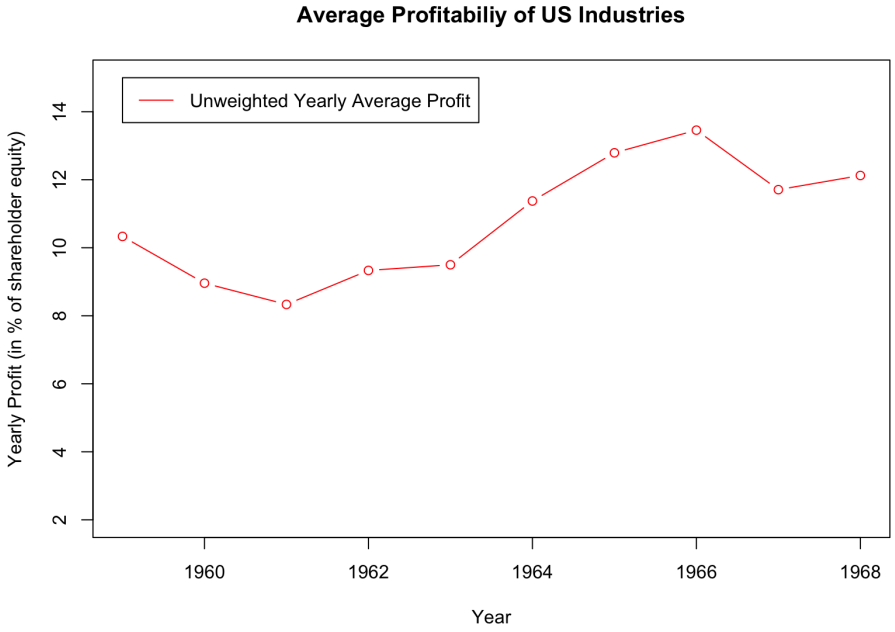
"Profit as a percentage of stockholder's equity is reported for various sectors of the US economy for the years 1959 through 1968."

My first order of business was creating a csv file to load into R Studio to to analysis.
After that, my first assumption I used to get a distance function for each entry. I will normalize the columns of this data set and therefore my assumption is that each year the profit percentage for all recorded industries is normally distributed.
Therefore, the distance between each entry is the equivalent to the difference in standard deviations away from the mean of the yearly average of all industries.
Here is a peak at the normalized data:
This data is not especially useful in being able to visualize the trends or anything about

	1	2	3	4	5	6	7
2	0.7218153						
3	6.8684300	6.6996491					
4	2.0258410	1.8957195	6.4042626				
5	1.3873905	1.0326703	6.8243961	2.2049082			
6	2.8432054	2.6240208	9.0487962	3.7099635	2.5376260		
7	4.8884535	5.0453056	11.3089047	5.9339896	5.3532492	4.2599127	
8	2.5921752	2.4951248	9.0601997	3.6387528	2.6505743	1.2988560	3.4386449
9	3.6628870	4.0024202	8.8630023	3.9604248	4.6192256	5.0826638	3.8574871
10	2.6036292	2.3600455	8.8383182	3.3891827	2.5048964	1.0161725	3.8663412
11	4.7012912	4.7920705	10.5379395	5.7461436	4.9388310	3.7839923	4.1572823
12	6.1447682	5.8731503	5.0763299	5.1988260	5.5986431	7.2333255	10.6924037

the data, really.





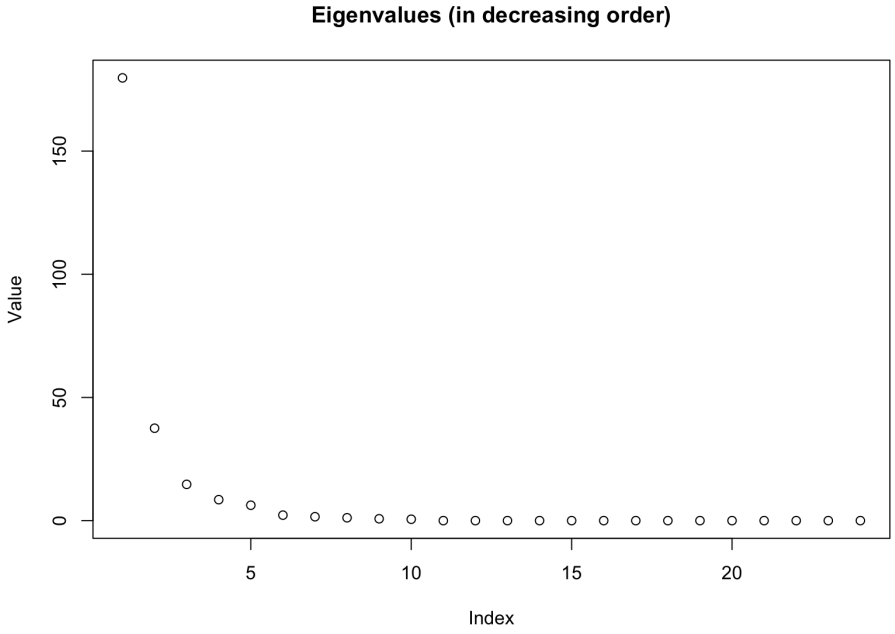
Here I have graphed a profit percentage vs. time graph for each industry. While this is also hard to dissect, it gives us some idea of what the data looks like. I have also inserted and graph of the unweighted average of each industry in order to see the broader trend of the overall economy in the 1960s.

We can see that these 24 industries had a slight dip in 1961 but rebounded and hit a 10-year high in 1966. This overall unweighted average is helpful in understanding a more broader trend, but part of this analysis is trying to identify characteristic factors of individual industries.

We will attempt to do that by using R's cmdscale() function with the previously mentioned normalized distances in order to scale our 10 dimensional data into lower dimensions. I will initially create models using 1-3 dimensions in order to see which will simultaneously be simple but also lead to a good fit for the data.
Here is a table of some findings:

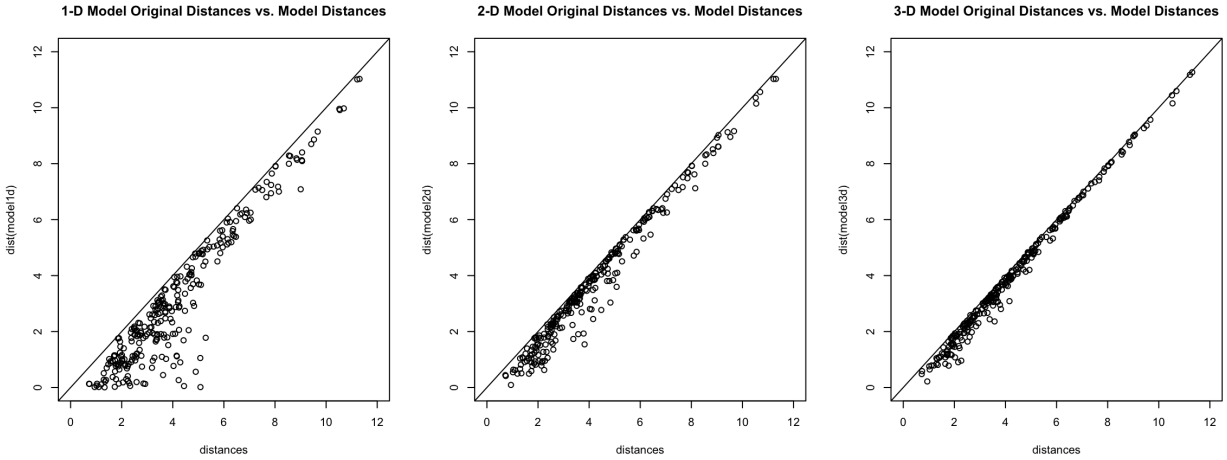
Model Dimension	GOF Value	Mean Absolute Error	Max Absolute Error	Mean Percentage Error	Max Percentage Error
1	0.7103349	1.043085	5.068391	33.62426%	99.71918%
2	0.8586168	0.4596764	2.280324	15.91977%	90.30906%
3	0.9168079	0.2828846	1.31296	10.39436%	76.9831%

As we can see from the table, the higher dimension we go, the better the model becomes. However, this increase is not linear with model dimension and we start to see diminishing returns.
The mean percentage error reduces by more than 50% from a 1-dimensional model to a 2-dimensional one. Similarly, the goodness of fit value increases by 0.14 while the increase from a 2-dimensional model to a 3-dimensional one is only 0.06.
Another (more visual) strategy we can use to analyze these different models is to graph the eigenvalues of the distance matrix:



As we can see, the second largest eigenvalue is much lower than the first, and the difference between each eigenvalue in sorted order decreases. This tells us, again, that a 2-dimensional model will be significantly better than a 1-dimensional one but a 3-dimensional one is not as significant a difference.

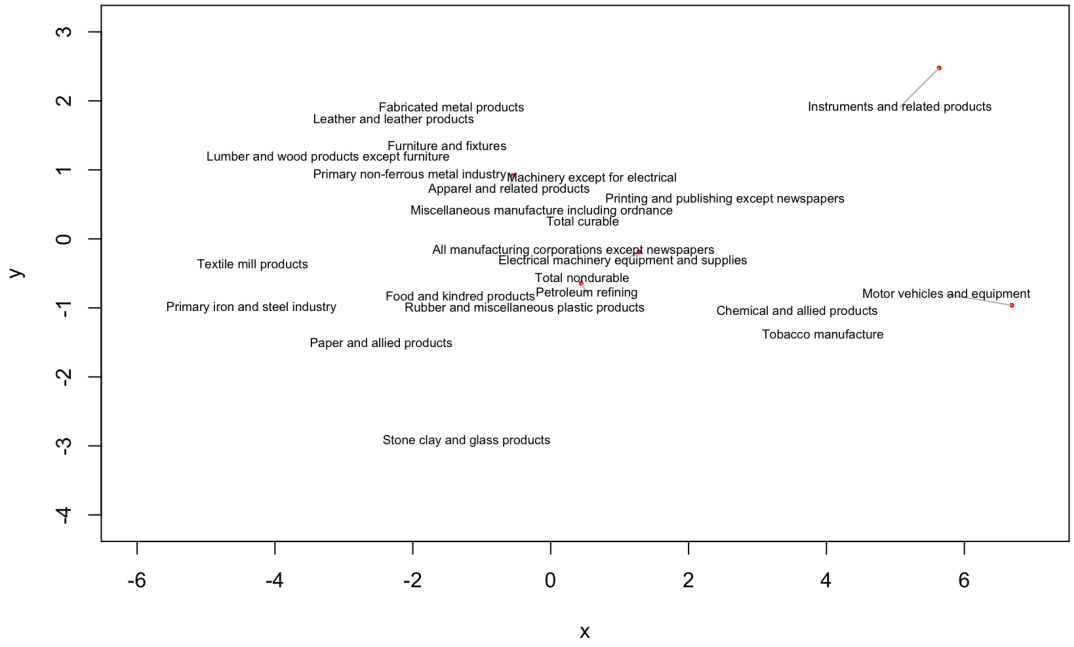
Lastly, we can graph the model's distances against the original distances we gave it. If the model is good, the resulting graph should have points that lie along the line $y = x$ since the model's distances should match those of the original distances.



This again confirms our claim. The points get increasingly closer to the line $y = x$ as the model dimension increases.

For the ease of graphing, and the diminishing returns of including a 3rd dimension, I will be using the 2-dimensional model to analyze.

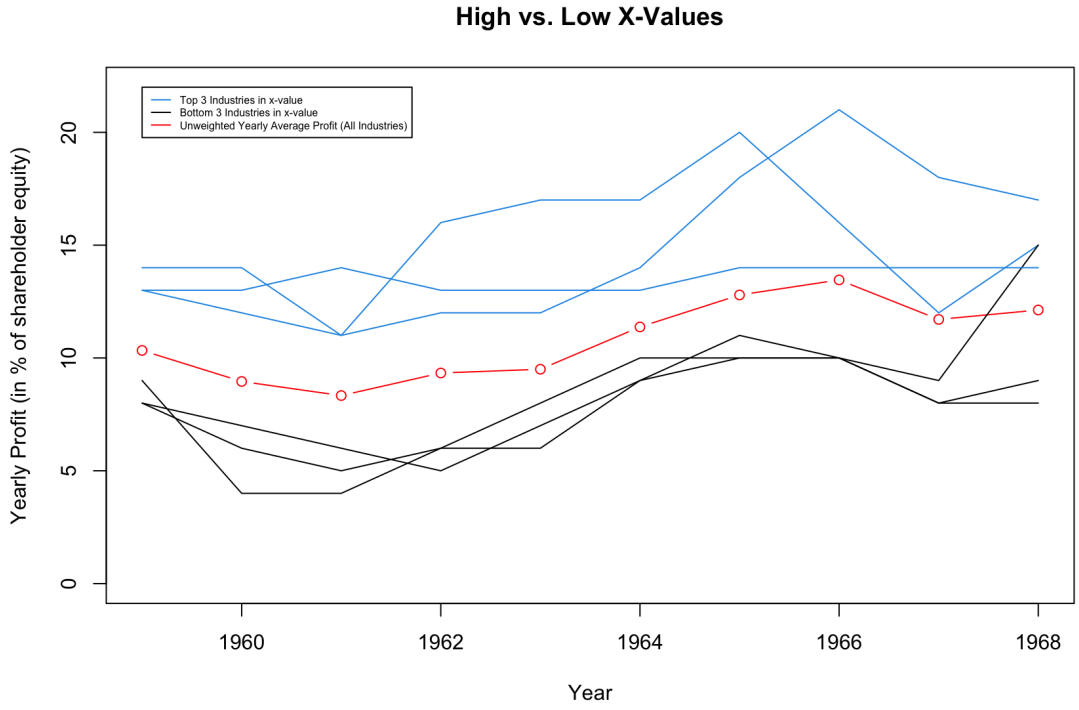
Here is the graph of the 2-dimensional model given by cmdscale() with labeled points:



At first glance, we can see some outliers. Instruments and Related Products has both large x and y values, Motor vehicles and equipment has a very large x value, etc. I have decided to pick the maximum and minimum 3 industries by x and y value. Here is a table of those industries:

3 Highest x-values	3 Lowest x-values
Motor vehicles and equipment	Primary iron and steel industry
Instruments and related products	Textile mill products
Tobacco manufacture	Lumber and wood products except furniture

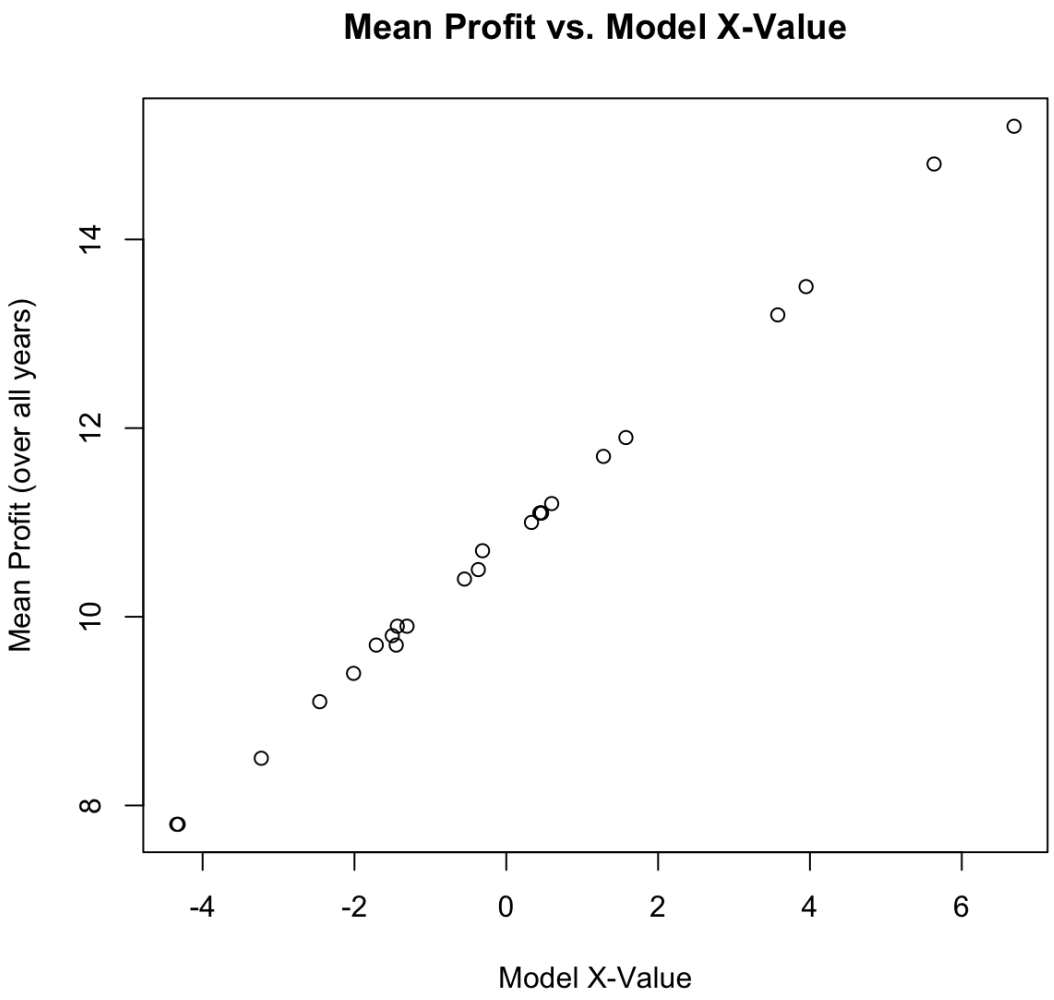
To do some exploratory analysis of these industries with more extreme x-values, I will graph their yearly profitability over the entire period versus the average.



As we can see, it is very clear that the top 3 industries ranked by x-value have average profitability that is considerably higher than that of the average profitability of all industries within each year. Similarly, the bottom 3 industries ranked by x-value have a lower profitability.

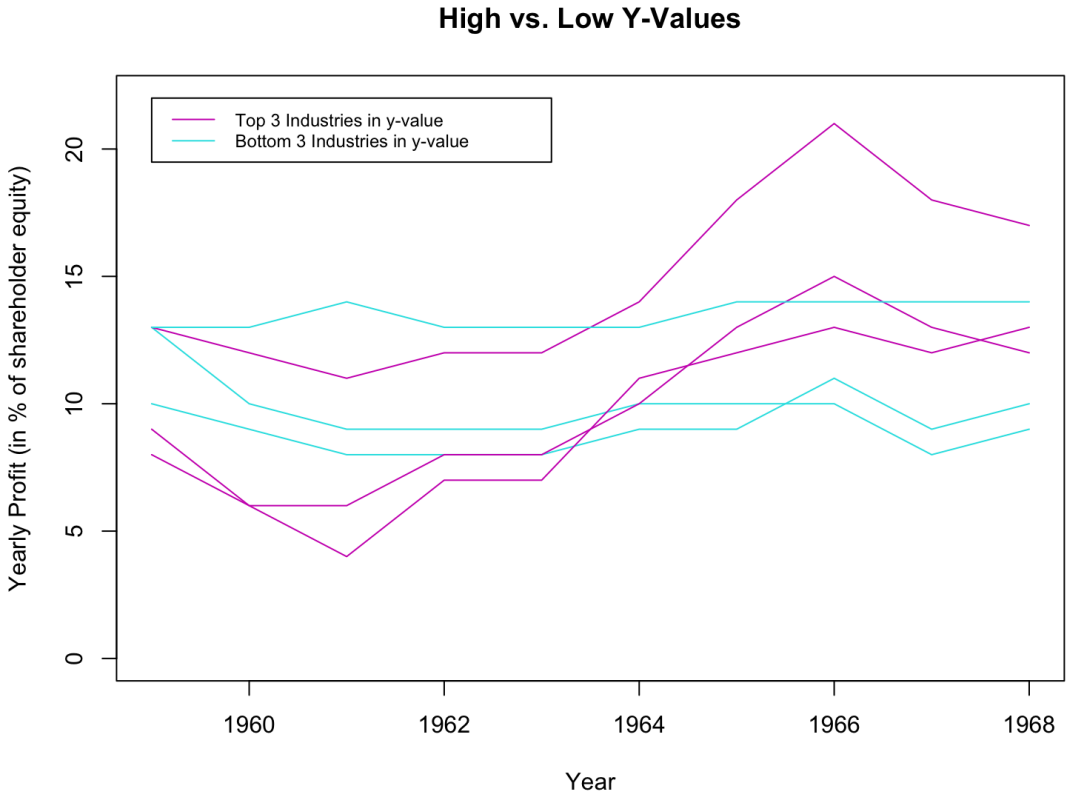
While this is not necessarily proof of this correlation, we can get some more hard data on all industries within our data set.

In this next graph, I have plotted each industry's average profitability (over all 10 years) versus the industry's x-value within our 2-dimensional model:

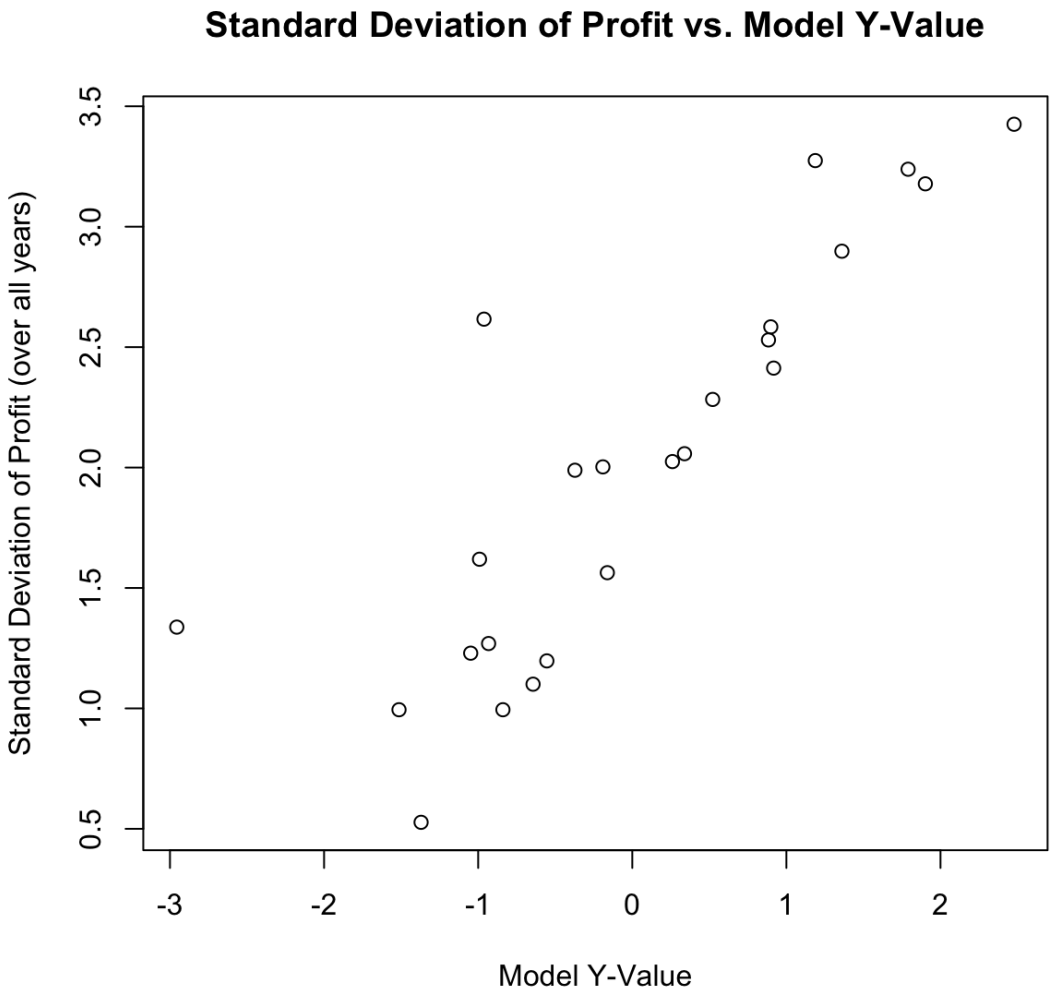


This graph makes it incredibly clear that there is a distinct correlation between mean profitability of an industry and it's x-value within our model. In fact, the Pearson correlation coefficient is $r = 0.9993642$ which shows an extreme correlation.
Next, I will analyze the industries with extreme y-values.
In the same way, I have chosen the highest and lowest 3 industries regarding y-value to do a sort of visual exploration of what makes these industries different.

3 Highest y-values	3 Lowest y-values
Instruments and related products	Stone clay and glass products
Fabricated metal products	Paper and allied products
Leather and leather products	Tobacco manufacture



At first glance, it seems like the lower y-valued industries stay constant while the higher y-valued industries are more variable. This may correlate with some measure of "volatility" in each industry or how much an industry is susceptible to large swings in its profit. To confirm this, I have used standard deviation as a typical measure of variability and graphed it versus the y-values of each industry.



While this graph doesn't have as strong a correlation as the x-values do to average profit, it is still relatively strong with a Pearson correlation coefficient of $r = 0.8516654$.

These findings suggest that the 2-dimensional model that was created may capture some characteristics of these industries regarding their average profitability and their volatility within this 10 year period.

What I find interesting is that we can now make some types of categorization of these industries. When I graphed the 3 min/max y/x value industries, I noticed that there were 2 industries that were in both graphs: Tobacco manufacture and Instruments and related products. We can now confidently say that tobacco manufacture (having a small y-value and large x-value) is a highly profitable industry with low volatility whereas instruments and related products (having a large y-value and large x-value) is a highly profitable industry with high volatility.

Code

```
library(wordcloud)
library(BBmisc)

normalized <- normalize(profit,method="standardize",margin=2)
distances <- dist(normalized)
print(distances)
model1d <- cmdscale(distances, k = 1)
model2d <- cmdscale(distances, k = 2)
model3d <- cmdscale(distances, k = 3)

print(profit[1,])
plot(c(1959:1968),profit[1,c(2:11)], col=1, type="l", ylim=c(2,25),
   main="Profitabiliy of US Industries", xlab="Year", ylab="Yearly
   Profit (in % of shareholder equity)")
for (i in c(2:24)) {
  lines(c(1959:1968),profit[i,c(2:11)], col = i)
}
legend(1959, 25, legend=profit$Industry, col=1:24, lty=1, cex=0.35)
textplot(model2d[,1],model2d[,2],gsub("(  [  \\s]+\\s{1})","",profit
   $Industry,perl=TRUE),asp=1,ylim=c(-3,2),xlim=c(-6,7),cex=0.6)

model2d_eig <- cmdscale(distances, k = 2, eig=TRUE)
plot(model2d_eig$eig, main="Eigenvalues (in decreasing order)", ylab
   ="Value")
means <- c(1:10)
for (i in c(2:11)) {
  means[i - 1] <- mean(profit[,i])
}
print(means)
plot(c(1959:1968),means, col="red", type="b", ylim=c(2,15), main="
   Average Profitabiliy of US Industries", xlab="Year", ylab="Yearly
    Profit (in % of shareholder equity)")
legend(1959, 15, legend=c("Unweighted Yearly Average Profit"), col="
   red", lty=1, cex=1)

print(model2d)
x_low <- c(7,17,11)
x_high <- c(3,12,16)
plot(c(1959:1968), means, col="red", type="b", ylim=c(0,22), main="
   High vs. Low X-Values", xlab="Year", ylab="Yearly Profit (in % of
    shareholder equity)")
for (i in c(1:3)) {
  lines(c(1959:1968),profit[x_low[i],c(2:11)], col = 1)
  lines(c(1959:1968),profit[x_high[i],c(2:11)], col = 4)
}
```

```
legend(1959, 22, legend=c("Top 3 Industries in x-value", "Bottom 3 Industries in x-value", "Unweighted Yearly Average Profit (All Industries)"), col=c(4, 1, "red"), lty=1, cex=0.5)

y_low <- c(9,19,16)
y_high <- c(12,6,24)
for(i in c(1:3)) {
  print(profit$Industry[x_high[i]])
  print(profit$Industry[x_low[i]])
  print(profit$Industry[y_high[i]])
  print(profit$Industry[y_low[i]])
}
y_low_means <- c(1:10)
y_high_means <- c(1:10)
for (j in c(1:10)) {
  y_low_total <- 0
  y_high_total <- 0
  for (i in c(1:3)) {
    y_low_total <- y_low_total + profit[y_low[i], j + 1]
    y_high_total <- y_high_total + profit[y_high[i], j + 1]
  }
  y_low_means[j] <- y_low_total / 3
  y_high_means[j] <- y_high_total / 3
}

#plot(c(1959:1968), means, col="red", type="b", ylim=c(0,22), main="High vs. Low Y-Values", xlab="Year", ylab="Yearly Profit (in % of shareholder equity)")
#lines(c(1959:1968), y_low_means, col ="green", type="b")
#lines(c(1959:1968), y_high_means, col ="orange", type="b")
plot(c(1959:1968),profit[y_low[1],c(2:11)], type="l", col = 5, ylim=c(0,22), main="High vs. Low Y-Values", xlab="Year", ylab="Yearly Profit (in % of shareholder equity)")
lines(c(1959:1968),profit[y_high[1],c(2:11)], col = 6)
for (i in c(2:3)) {
  lines(c(1959:1968),profit[y_low[i],c(2:11)], col = 5)
  lines(c(1959:1968),profit[y_high[i],c(2:11)], col = 6)
}
legend(1959, 22, legend=c("Top 3 Industries in y-value", "Bottom 3 Industries in y-value"), col=c(6, 5), lty=1, cex=0.75)

plot(c(1959:1968), profit[16,c(2:11)], col="blue", type="b", ylim=c(0,22), main="High vs. Low X-Values", xlab="Year", ylab="Yearly Profit (in % of shareholder equity)")
```

```r
diff1 <- distances - dist(model1d)
diff2 <- distances - dist(model2d)
diff3 <- distances - dist(model3d)
print(mean(abs(diff1)))
print(mean(abs(diff2)))
print(mean(abs(diff3)))
print(max(abs(diff1)))
print(max(abs(diff2)))
print(max(abs(diff3)))
print(mean(100 * abs(diff1) / distances))
print(mean(100 * abs(diff2) / distances))
print(mean(100 * abs(diff3) / distances))
print(max(100 * abs(diff1) / distances))
print(max(100 * abs(diff2) / distances))
print(max(100 * abs(diff3) / distances))
hist(abs(diff))
par(mfrow=c(1,3))
plot(distances, dist(model1d), xlim=c(0,12), ylim=c(0,12), main="1-D
    Model Original Distances vs. Model Distances")
abline(0, 1)
plot(distances, dist(model2d), xlim=c(0,12), ylim=c(0,12), main="2-D
    Model Original Distances vs. Model Distances")
abline(0, 1)
plot(distances, dist(model3d), xlim=c(0,12), ylim=c(0,12), main="3-D
    Model Original Distances vs. Model Distances")
abline(0, 1)

par(mfrow=c(1,1))
data <- data.frame(profit)
df <- subset(data, select = -c(Industry))
mean_industry <- rowMeans(df)
sd_industry <- c(1:24)
for (i in c(1:24)) {
  sd_industry[i] <- sd(df[i,])
}
print(range_industry)
mean_cor <- cor(model2d[,1], mean_industry)
sd_cor <- cor(model2d[,2], sd_industry)
plot(model2d[,1], mean_industry, xlab="Model X-Value", ylab="Mean
   Profit (over all years)", main="Mean Profit vs. Model X-Value")
plot(model2d[,2], sd_industry, , xlab="Model Y-Value", ylab="
   Standard Deviation of Profit (over all years)", main="Standard
   Deviation of Profit vs. Model Y-Value")

print(mean_cor)
print(sd_cor)
```