PCA Data Analysis

Harper Schwab

October 27, 2020

Introduction and ADME

This study deals with a Principle Component Analysis (PCA) of 1270 drugs that are on the consumer market. The analysis of the results is to see what factors affect the commercial success of the drug. THe data was gathered throu StarDrop normally used for pharmacokinetics, and pharmacodynamics. They are essentially how the body engages with the drug, usually characterized by ADME. ADME is an acronym standing for Absorption, Distribution. Metabolism, and Excretion. I will now go into further detail on each word in context of drugs. The Absorption of a drug is usually is absorbed into the bloodstream through mucous surfaces such as intestines, for oral drugs. The ability for a drug to permeate these mucous surfaces can play a important role in weather or not a drug is successful. If a drug is not absorbed through the intestines, one must take the drug another way through IV or inhalation. Distribution is how the compound in the drug is taken to the location needed in the body. Usually though the bloodstream, the factors that affect this are the blood pressure and flow rate of the individual, the size of the compound and polarity, as well as the ability to pass through the blood brain barrier. the Metabolism is the breaking down of the compounds in the drugs usually carried out by the liver, parent compounds are broken apart if the drug is metabolized to severely it's effectiveness will decrease. Excretion is the process where the compounds are removed from the body through urine, bile, through the lungs, or through feces (ADME). Lipinski's rule of five is another way to refer to see if a drug is effective in humans. In this case it is to see if the drug is orally active in humans. The components of the rule are: That there shall be no more than 5 hydrogen bond donors and no more than 10 acceptors. It must have a molecular mass less than 500 daltons, and an octanol-water-partition coefficient that is less than of equal to five (Lipinski's Rule of Five). All of these factors may attribute to the success of the 1200+ drugs. A PCA will help to determine this.

PCA

A Principle Component Analysis or PCA allows for the dimension reduction of data with greater than 3 dimensions and make a 2 dimensional plot. To start you can plot data, and calculate the average measurements for each of the 2 variables you plotted against each-other to calculate the center of the data. This center point then becomes the origin. Then a line is plotted which shows the best fit of the data by seeing which line minimizes the distances from the data to the line, or maximizes the projected point's distance to the origin. Then all of these distances used to find the largest sum of the square distances. That line is Principle component 1. The slope of the principle complainants can tell you the scale between the 2 principle components. This linear combination of two principle component tells us which component is more important when describing the data. Principle component 2 is the perpendicular line to PC1 the goes through the origin. The loading scores of the components show the level of importance of each value is projected on the the principle components. For the final plot PC1 is rotated until it is horizontal. Then you have a completed PCA plot. Through Using R, this process is automatic and fairly simple the prcomp() command does this process and calculates all of the values for you, which made this project much easier than if we had to do it byu hand. (Starmer, 2018)

Steps Taken

The first thing I did was look at the origional data and compared it to transposed data. I decided that transposed data would be better for this PCA because the categories matched simillarly to the examples Stammer used in his videos.

```
sample_n(drugs, 2)
```

```
##                   logS logSpH7  logP  logD X2C9pKi hERGpIC50     BBB
## DIFE0XIN         0.9686  0.9395 4.511 1.198   5.402     5.097 -0.9641
## BROMPHENIRAMINE 2.6080  1.6160 3.699 1.618   4.840     5.703  0.9416
##                 Pgpcategory    MW HBD HBA  TPSA Flexibility RotatableBonds
## DIFE0XIN                  1 424.5   1   4 64.33      0.2286              8
## BROMPHENIRAMINE           0 319.2   0   2 16.13      0.2500              5
```

Above is a sample of the original data, as you can see the drugs are on the leftmost side while the variables are at the top. The transposed data flips those and has the drugs on the variables at the top and the drugs on the side. Another way to illustrate this is with the dim() command, which will follow.

```
dim(drugs)
```

```
## [1] 1270   14
```

```
dim(Tdrugs)
```

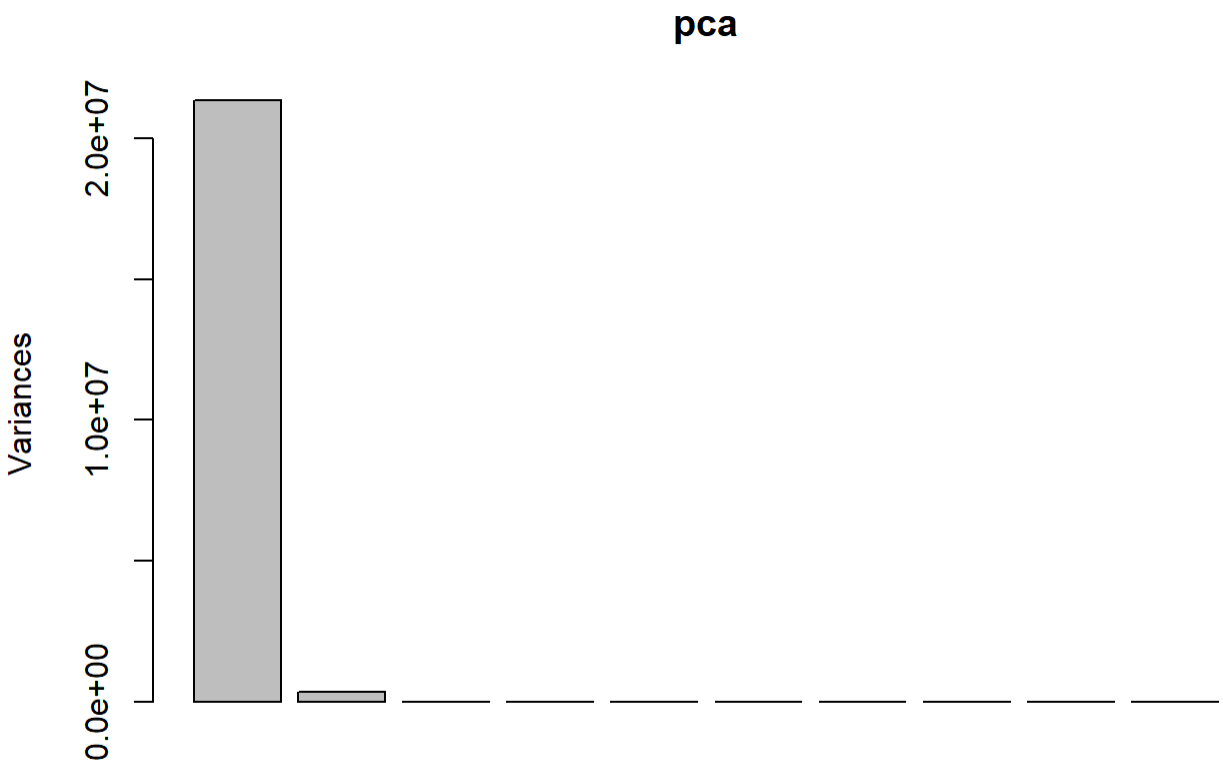
```
## [1]   14 1270
```

I performed a PCA on the transposed data with the following command.

```
Tpca <- prcomp(Tdrugs)
#I think this ones better
pca <- Tpca
```

I then plotted the 14 PCA's in a screeplot to see which PCA's had the most variance and therefore are better for analysis.

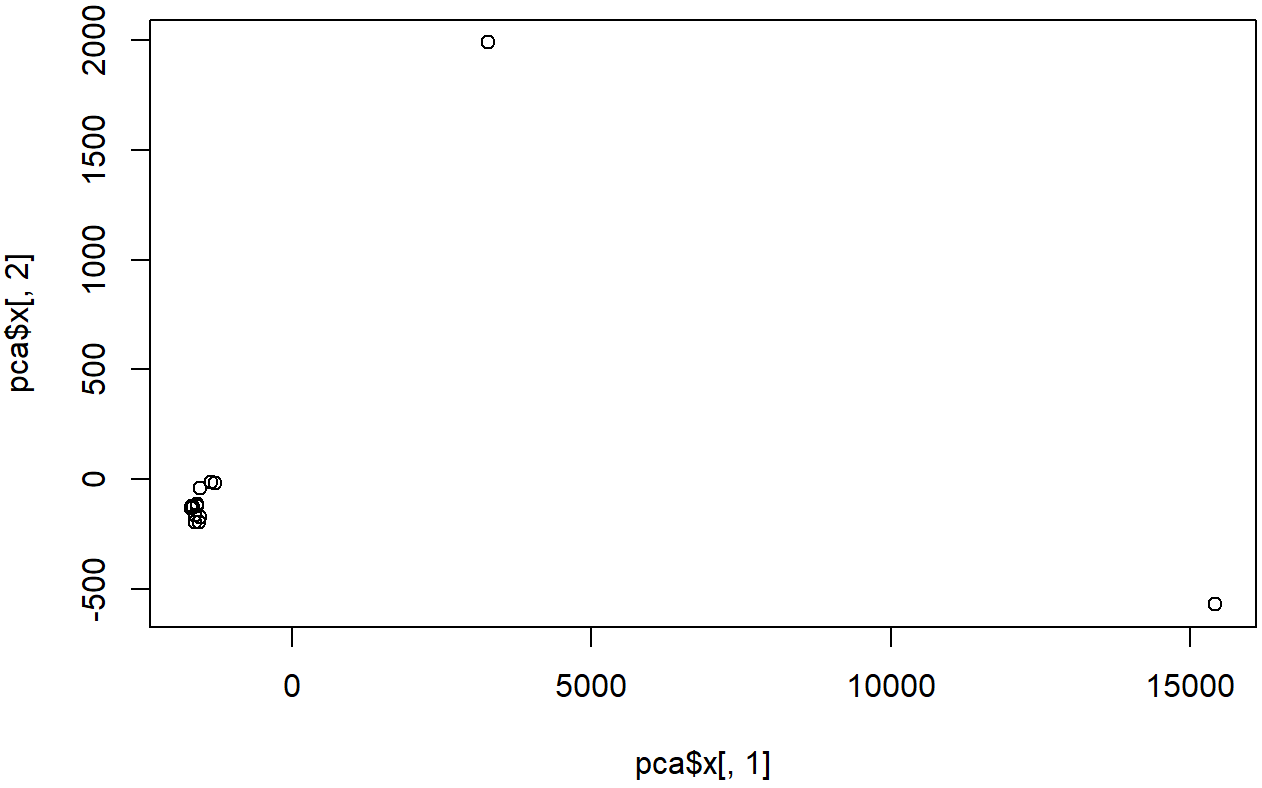
```
screeplot(pca)
```



Through the screeplot it was apparent that PCA 1 and 2 had the most variance and would become the axis in which I measured my variables on.

I also plotted the two PCA columns on a simple scatter plot and had these results.

```
plot(pca$x[,1],pca$x[,2])
```



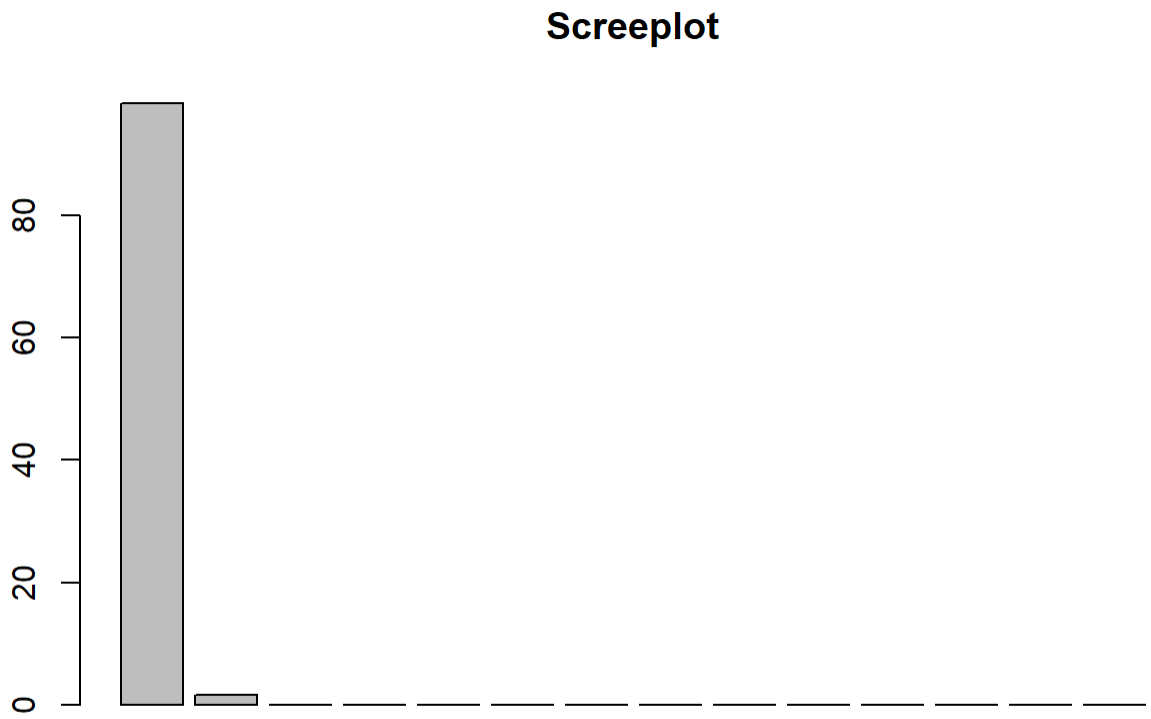
We can tell, with those two outlines and the cluster that these PCA components explain a large amount of the variance, but how much?

To find that out I found the variance through squaring the standard deviation and then changed that value into a percent that was rounded. I found that PCA 1 accounted for 98.3% of the Variance while PCA 2 only accounted for 1.6%. This relationship is visualized through the screeplot below.

```
#
pca.varience <- pca$sdev^2
pca.varience.per <- round(pca.varience/sum(pca.varience)*100,1)
pca.varience.per
```

```
##  [1] 98.4  1.6  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
```

```
#PCA 1 accounts for 98.4% of the varience in the data, while PCA 2 accounts for 1.6%
barplot(pca.varience.per,main = 'Screeplot')
```



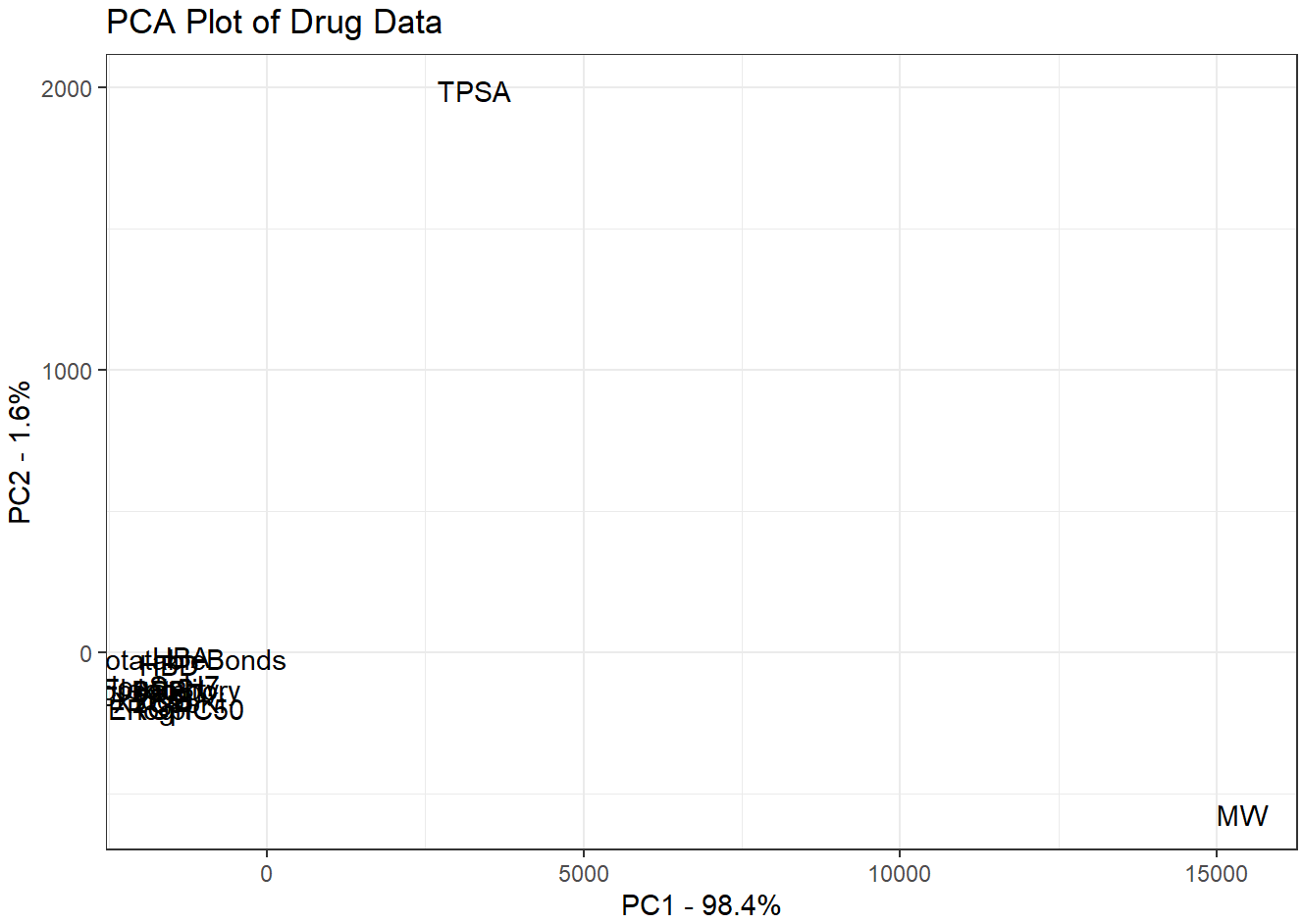
I then went our to create a better PCA plot. First thing I did was sample the data from the two columns I was using Using the following code.

```
#A little nicer plot
pca.data <- data.frame(Sample = rownames(pca$x), X=pca$x[,1], Y=pca$x[,2])
pca.data
```

```
##                         Sample         X          Y
## logS                       logS -1595.874 -122.65631
## logSpH7                 logSpH7 -1587.401 -115.04553
## logP                       logP -1615.724 -196.58613
## logD                       logD -1624.835 -163.57657
## X2C9pKi                 X2C9pKi -1534.006 -172.28378
## hERGpIC50             hERGpIC50 -1552.506 -197.79457
## BBB                         BBB -1691.092 -129.84905
## Pgpcategory         Pgpcategory -1655.722 -126.00836
## MW                           MW 15414.245 -570.24446
## HBD                         HBD -1530.992  -39.98425
## HBA                         HBA -1347.245  -14.65678
## TPSA                       TPSA  3273.529 1991.03211
## Flexibility         Flexibility -1665.673 -123.23515
## RotatableBonds RotatableBonds -1286.704  -19.11118
```

Using the data I just created I used ggplot to create a PCA plot.

```
pcaggplot <- ggplot(data= pca.data, aes(x=X,y=Y, label = Sample)) +
  geom_text()+
  xlab(paste("PC1 - ", pca.varience.per[1], "%", sep = ""))+
  ylab(paste("PC2 - ", pca.varience.per[2], "%", sep = ""))+ theme_bw()+ggtitle("PCA Plot of Drug
 Data")
pcaggplot
```




```
#
```

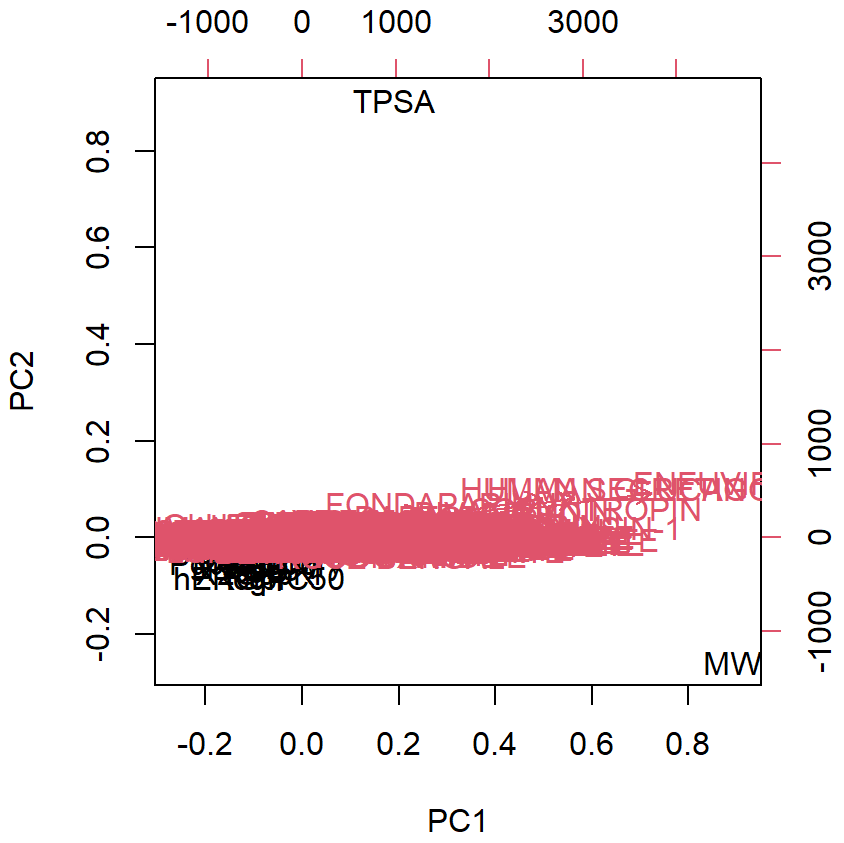
Through this I see that the molecular weight and the topographical polar surface area are the farthest away from the cluster, maybe those are the best factors.

I continued to try to figure out what factor contributed to PCA 1 and 2 the most using loading scores. Loading scores show which specific ones affect the PCA.

```
#Which factors are contributing to pca 1 and pca 2
loading_scores <- pca$rotation[,1]
gene_scores <- abs(loading_scores)
gene_scores_ranked <- sort(gene_scores, decreasing =T)
top10<- names(gene_scores_ranked[1:10])
top10
```

```
##  [1] "ENFUVIRTIDE"       "HUMAN GLUCAGON"    "HUMAN SECRETIN"
##  [4] "COSYNTROPIN"       "BIVALIRUDIN"       "DAPTOMYCIN"
##  [7] "IOTROLAN"          "FONDAPARINUX"      "GANIRELIX ACETATE"
## [10] "IODIXA0L"
```

```
biplot(pca)
```



The top ten scores of the Transposed data show that the factors of human glucagon and Secretion have the most affect on the drug.

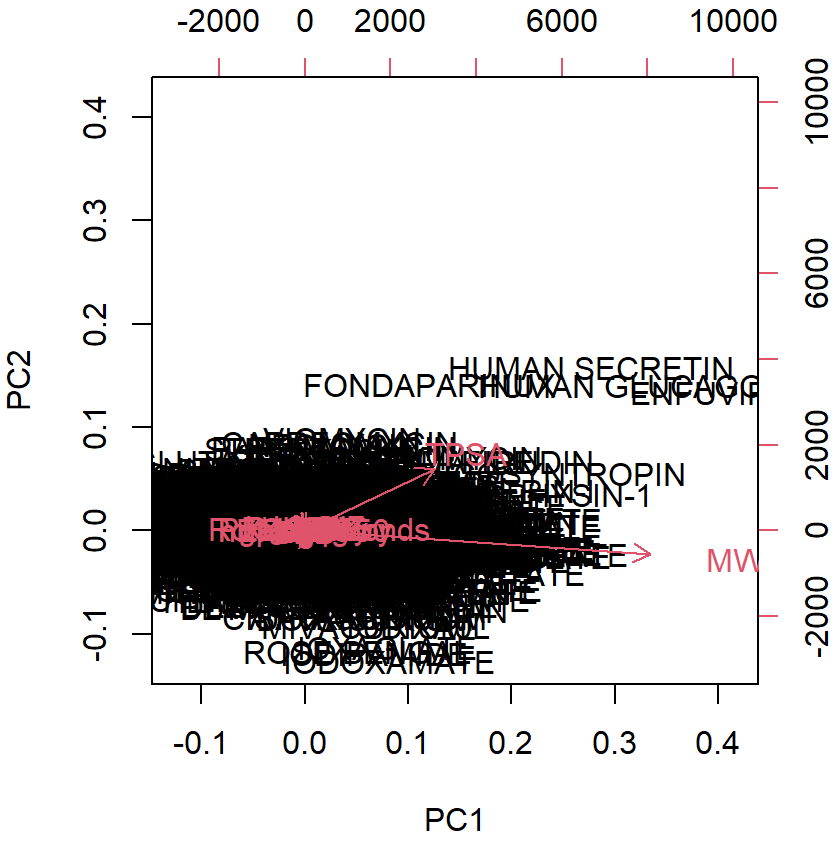
I then wanted to see if there were any differeneces with non-transposed data so I did a similar analysis, but on the original, non-transposed, data

```
pca <- prcomp(drugs)
#Which factors are contributing to pca 1 and pca 2
loading_scores <- pca$rotation[,1]
gene_scores <- abs(loading_scores)
gene_scores_ranked <- sort(gene_scores, decreasing =T)
top10<- names(gene_scores_ranked[1:10])
top10
```

```
##  [1] "MW"             "TPSA"           "RotatableBonds" "HBA"
##  [5] "HBD"            "logD"           "logSpH7"        "hERGpIC50"
##  [9] "logS"           "X2C9pKi"
```

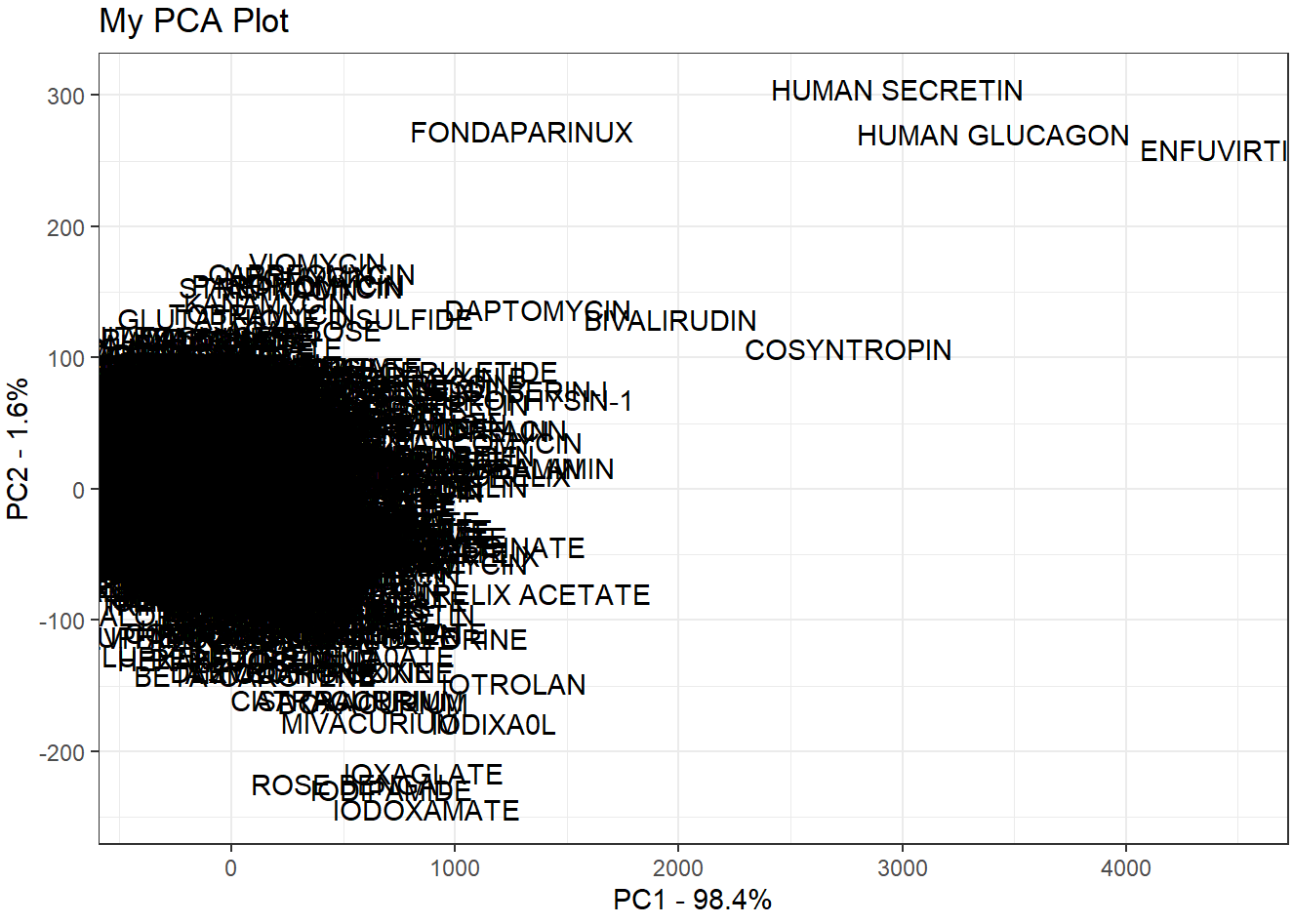
```
biplot(pca)
```

```
## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped
```



The factors on Molecular Weight, Topographical surface area and Rotatable Bonds seem to affect it the most.

I will also show some of the previous plots but with the non-transposed data.



Analysis

Through the PCA plots and top ten loading scores I can see the top three from each data-set, transposed/non-transposed. The transposed data, had multiple drugs that highly affected the varience of the principle components, but I am looking for the factors. Together, the 5 factors are: Molecular weight, the topological polar surface area (TPSA), Rotatable Bonds, Human glucagon, human secretion. In accordance to AMDE we can tell that the molecular weight and the polar surface area directly affect the distribution of the drug. The molecular size, which should attribute to weight and surface area, can stop drugs from permeating through mucous surfaces and being delivered through the blood stream and or breaking the blood-bran barrier. It also aligns with Lipinnski's Rule of five. The molecular mass needs to be less than 500 daltons to have a more certain success in the human body. The rotatable bond level, is in conjunction with Lipinski's rule of five where there shall be no more than 5 hydrogen bond donors and no more than 10 acceptors. Human glucagon is a hormone that helps the matabolism process. Weather or not a drug contains a certain amount of this hormone could affect its effectiveness through the metabolism process. And finally human secretion which is a important process in ADME. THe principle component analysis of the drug data illustrated the factors which contribute the most to commercial success of drugs.

References

ADME. (2020, April 28). Retrieved October 27, 2020, from https://en.wikipedia.org/wiki/ADME (https://en.wikipedia.org/wiki/ADME)

S.K. Balani; V.S.Devishree; G.T. Miwa; L.S. Gan; J.T. Wu; F.W. Lee (2005). "Strategy of utilizing in vitro and in vivo ADME tools for lead optimization and drug candidate selection". Curr Top Med Chem. 5 (11): 1033–8. doi:10.2174/156802605774297038 (doi:10.2174/156802605774297038). PMID 16181128.

Singh S.S. (2006). "Preclinical pharmacokinetics: an approach towards safer and efficacious drugs". Curr Drug Metab. 7 (2): 165–82. doi:10.2174/138920006775541552 (doi:10.2174/138920006775541552). PMID 16472106.

Tetko IV, Bruneau P, Mewes HW, Rohrer DC, Poda GI (2006). "Can we estimate the accuracy of ADME-Tox predictions?," (pre-print). Drug Discov Today. 11 (15–16): 700–7. doi:10.1016/j.drudis.2006.06.013 (doi:10.1016/j.drudis.2006.06.013). PMID 16846797.

Lipinski's Rule of Five. (2012, December 10). Retrieved October 27, 2020, from https://en.wikipedia.org/wiki/Lipinski's_Rule_of_Five (https://en.wikipedia.org/wiki/Lipinski's_Rule_of_Five)

Starmer, J. (Writer). (2018, April 2). StatQuest: Principal Component Analysis (PCA), Step-by-Step [Video file]. Retrieved October 27, 2020, from https://www.youtube.com/watch?v=FgakZw6K1QQ&t=3s (https://www.youtube.com/watch?v=FgakZw6K1QQ&t=3s)