Commercially Succsessful Drugs PCA Analysis

Samuel Dummer

10/12/2021

Abstract

The use of medications is widespread and there are many different variables that are important in terms of the drugs commercial success. To combat this, we are using Principal Component Analysis (PCA) to easily analyze the dimensional reduction of the data. Within this data there are 14 different numerical variables that are being analyzed and the goal is to find the two most important variables in terms of commercial success. There are many steps taken to find these results, notably, scree plots and biplots, which help to visualize the variation. within these 14 variables, there are also 1270 different commercial available drugs that we are covering.

Introduction

Pharmacokinetics(PK) and Pharmacodynamics(PD)

Pharmacokinetics and pharmacodynamics are both parameters that help to define different drugs. Pharmacokinetics helps to focus on how the body drug is distributed, absorbed, metabolized, and excreted by the body. On the other hand, pharmacodynamics focuses on what effects the drug has on the body, or more simply, what the drug does to the body. This mainly has to do with the biochemical, physiologic, and molecular effects on the body, but also includes receptor binding (the technique that a specific compound identifies its receptor). When looking specifically at pharmacokinetics, we can see that it is able to find the effects of the drugs (a.k.a. its intensity and length) which are dependent on the the absorption, metabolism, etc.. Additionally, the variables that change the effect of the drug also vary between patients since everyone's body has different speeds of metabolism. These variables help to define the standard effects of drugs, but as stated earlier, each patient has a different reaction do the drugs so the standard effects may not be the only ones.

As we look at the pharmacokinetics of drugs, it is also important to remember AMDE. ADME also known as Absorbtion, Distribution, Metabolism, and Excretion, are the steps that a foreign substance, in this case the drugs, go through after entering the body. Firstly, absorbtion is the method in which the body absorbs the substance, generally absorbed in the lining of the stomach. Then, there is distribution which is a measure of how and where the blood is distributed throughout the body. There are many variables needed to be taken into account, for example, the blood-brain barrier (BBB) which is a membrane that helps prevent foreign substances from entering the brain. Next, we have metabolism which is a method in the body that changes the structure of the substance in the body so it is ready for excretion. The majority of this process is done by proteins. Lastly, there is excretion, which is the process in which the body removes the substance from the body. This whole process is the process that any foreign substance goes through.

In the dataset, there are many variables/descriptors that describe each drug. This includes logS, the measure of the drugs solubility. logsph7 which is very similar to logS. It measures the drugs solubility at a ph of 7. There is also pKi, the equilibrium dissociation constant. Additionally, we have HBA and HBD which the number of hydrogen blood acceptors and hydrogen blood donators. There is also MW and TPSA which are the molecular weight in Daltons and topological polar surface area which, seen in the name, measures polarity.

Principal Component Analysis (PCA)

PCA is a method in which one can measure the variation within a dataset while also finding patterns within this variation, which may not be visible without the PCA. This method of analysis takes multidimensional data and then simplifies it. It is simplified into multiple PCs or Principal Components (the number of PCs is based on the number of variables in the dataset or samples depending on the size of each). In doing so, we are able to maximize the variation, especially, in the first PC or PC1. This method of analysis is able to help detect which variable is causing the most variation, which, is this specific case, would cause the most variation in commercial success.

Cleaning Environment, Setting Up Directory, and Loading Libraries.

Before we are able to do any analysis of the dataset we must make sure that everything is set up in order for the script to work. This means we should clear the environment of any previous work and set the directory. Additionally, we must load in tidyverse since we will be doing some graphing which requires tidyverse.

```
rm(list=ls())
setwd("C:/Users/isabe/Desktop/RFLoder")
library(tidyverse)
```

Reading in the Data

Now that everything is set up, we are able to read in the all the data about the drugs. We will also use "row.names = 1" to make the first column become the row names.

```
#Sam Dummer
#WK8L2Dummer.R
#October 11, 2021
drug <- read.csv("drugs.csv", header=TRUE, row.names = 1)
```

Summary of Data

Next, we quickly checked the summary of the data using head(), tail(), glimpse(), str(), and summary(). The head and tail commands help to show a quick sneak peek of the data and check if there are any mistakes. Next we use glimpse and str to observe the general structure of the dataset and the type of variables we are working with. Lastly we used summary to see a general summary of each variable which includes the minimum, 1st quartile, median, meann, 3rd quartile, and maximum.

```
#summary and stucture of data
head(drug)
```

```
##                logS logSpH7    logP     logD X2C9pKi hERGpIC50     BBB
## ABACAVIR      3.233   1.934  1.3870  0.40990   4.705     5.550 -0.4411
## ABARELIX      2.149   3.889  1.3900  4.28900   5.026     1.688 -1.0720
## ACAMPROSATE   6.357   3.807 -1.9190 -1.84400   3.866     3.551 -0.4696
## ACARBOSE      5.397   5.134 -2.4960 -0.91740   5.223     2.631 -1.5860
## ACEBUTOLOL    3.491   2.946  1.7100 -0.08998   4.396     4.696 -0.1500
## ACETAMIOPHEN 4.970   4.970  0.4711  0.47110   4.020     4.044 -0.3943
##              Pgpcategory     MW HBD HBA   TPSA Flexibility RotatableBonds
## ABACAVIR               1  286.3   3   7 101.90      0.1667             4
```

```
## ABARELIX               1 1416.0  13  28 425.00      0.4528             48
## ACAMPROSATE            0  181.2   2   5  83.47      0.5000              5
## ACARBOSE               1  645.6  14  19 321.20      0.1915              9
## ACEBUTOLOL             1  336.4   3   6  87.66      0.4583             11
## ACETAMIOPHEN           0  151.2   2   3  49.33      0.1818              2
```

```
tail(drug)
```

```
##                  logS logSpH7    logP    logD X2C9pKi hERGpIC50     BBB
## ZILEUTON        3.214  3.2140  2.1650  2.1650   3.984     4.689 -0.9100
## ZIPRASIDONE     2.072  0.5309  3.7200  3.5950   5.462     6.630 -0.1860
## ZOLEDRONIC ACID 5.230  4.7940 -0.1733 -3.6640   4.007     2.907 -1.3500
## ZOLMITRIPTAN    3.462  1.9010  2.1290  0.3214   4.367     4.864 -1.0400
## ZOLPIDEM        2.676  2.6760  2.9660  2.9660   5.321     5.258 -0.4607
## ZONISAMIDE      3.933  3.9330  0.9818  0.9818   3.778     4.884 -0.8108
##                 Pgpcategory    MW HBD HBA   TPSA Flexibility RotatableBonds
## ZILEUTON                  0 236.3   2   4  66.56      0.1765              3
## ZIPRASIDONE               1 412.9   1   5  48.47      0.1250              4
## ZOLEDRONIC ACID           0 272.1   5   9 153.10      0.2500              4
## ZOLMITRIPTAN              0 287.4   2   5  57.36      0.2174              5
## ZOLPIDEM                  0 307.4   0   4  37.61      0.1600              4
## ZONISAMIDE                0 212.2   1   5  86.19      0.1333              2
```

```
glimpse(drug)
```

```
## Rows: 1,270
## Columns: 14
## $ logS           <dbl> 3.2330, 2.1490, 6.3570, 5.3970, 3.4910, 4.9700, 4.9980,~
## $ logSpH7        <dbl> 1.9340, 3.8890, 3.8070, 5.1340, 2.9460, 4.9700, 4.9980,~
## $ logP           <dbl> 1.3870, 1.3900, -1.9190, -2.4960, 1.7100, 0.4711, -0.27~
## $ logD           <dbl> 0.40990, 4.28900, -1.84400, -0.91740, -0.08998, 0.47110~
## $ X2C9pKi        <dbl> 4.705, 5.026, 3.866, 5.223, 4.396, 4.020, 3.775, 4.104,~
## $ hERGpIC50      <dbl> 5.550, 1.688, 3.551, 2.631, 4.696, 4.044, 3.960, 2.853,~
## $ BBB            <dbl> -0.44110, -1.07200, -0.46960, -1.58600, -0.15000, -0.39~
## $ Pgpcategory    <int> 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1~
## $ MW             <dbl> 286.30, 1416.00, 181.20, 645.60, 336.40, 151.20, 222.20~
## $ HBD            <int> 3, 13, 2, 14, 3, 2, 2, 1, 2, 2, 1, 2, 0, 2, 4, 1, 2, 1,~
## $ HBA            <int> 7, 28, 5, 19, 6, 3, 7, 2, 6, 3, 5, 4, 3, 4, 14, 3, 2, 4~
## $ TPSA           <dbl> 101.90, 425.00, 83.47, 321.20, 87.66, 49.33, 115.00, 37~
## $ Flexibility    <dbl> 0.16670, 0.45280, 0.50000, 0.19150, 0.45830, 0.18180, 0~
## $ RotatableBonds <int> 4, 48, 5, 9, 11, 2, 3, 0, 6, 1, 7, 3, 4, 4, 9, 6, 5, 6,~
```

```
str(drug)
```

```
## 'data.frame':    1270 obs. of  14 variables:
##  $ logS          : num  3.23 2.15 6.36 5.4 3.49 ...
##  $ logSpH7       : num  1.93 3.89 3.81 5.13 2.95 ...
##  $ logP          : num  1.39 1.39 -1.92 -2.5 1.71 ...
##  $ logD          : num  0.41 4.289 -1.844 -0.917 -0.09 ...
##  $ X2C9pKi       : num  4.71 5.03 3.87 5.22 4.4 ...
##  $ hERGpIC50     : num  5.55 1.69 3.55 2.63 4.7 ...
##  $ BBB           : num  -0.441 -1.072 -0.47 -1.586 -0.15 ...
```

```
##  $ Pgpcategory   : int  1 1 0 1 1 0 0 0 0 0 ...
##  $ MW            : num  286 1416 181 646 336 ...
##  $ HBD           : int  3 13 2 14 3 2 2 1 2 2 ...
##  $ HBA           : int  7 28 5 19 6 3 7 2 6 3 ...
##  $ TPSA          : num  101.9 425 83.5 321.2 87.7 ...
##  $ Flexibility   : num  0.167 0.453 0.5 0.192 0.458 ...
##  $ RotatableBonds: int  4 48 5 9 11 2 3 0 6 1 ...
```

```
summary(drug)
```

```
##       logS            logSpH7            logP              logD        
##  Min.   :-2.750   Min.   :-2.750   Min.   :-5.0810   Min.   :-5.4780  
##  1st Qu.: 1.770   1st Qu.: 1.665   1st Qu.: 0.6122   1st Qu.:-0.3665  
##  Median : 2.755   Median : 2.611   Median : 2.2770   Median : 1.1280  
##  Mean   : 2.902   Mean   : 2.759   Mean   : 2.0912   Mean   : 1.1240  
##  3rd Qu.: 3.920   3rd Qu.: 3.804   3rd Qu.: 3.5545   3rd Qu.: 2.5935  
##  Max.   : 9.765   Max.   :10.100   Max.   : 8.6360   Max.   :12.8500  
##     X2C9pKi         hERGpIC50            BBB             Pgpcategory     
##  Min.   :3.394   Min.   :-1.602   Min.   :-2.40000   Min.   :0.0000  
##  1st Qu.:4.276   1st Qu.: 3.744   1st Qu.:-1.07800   1st Qu.:0.0000  
##  Median :4.728   Median : 4.539   Median :-0.52290   Median :0.0000  
##  Mean   :4.694   Mean   : 4.440   Mean   :-0.49389   Mean   :0.4323  
##  3rd Qu.:5.043   3rd Qu.: 5.301   3rd Qu.: 0.06151   3rd Qu.:1.0000  
##  Max.   :6.374   Max.   : 7.977   Max.   : 1.44000   Max.   :1.0000  
##        MW               HBD              HBA              TPSA        
##  Min.   :  31.01   Min.   : 0.000   Min.   :  0.000   Min.   :   0.00  
##  1st Qu.: 254.32   1st Qu.: 1.000   1st Qu.:  3.000   1st Qu.:  42.72  
##  Median : 328.50   Median : 2.000   Median :  5.000   Median :  72.72  
##  Mean   : 387.33   Mean   : 2.451   Mean   :  6.514   Mean   :  95.55  
##  3rd Qu.: 428.60   3rd Qu.: 3.000   3rd Qu.:  7.000   3rd Qu.: 111.50  
##  Max.   :4492.00   Max.   :63.000   Max.   :115.000   Max.   :1903.00  
##   Flexibility     RotatableBonds  
##  Min.   :0.0000   Min.   :  0.000  
##  1st Qu.:0.1250   1st Qu.:  3.000  
##  Median :0.2064   Median :  5.000  
##  Mean   :0.2275   Mean   :  6.797  
##  3rd Qu.:0.3000   3rd Qu.:  8.000  
##  Max.   :0.9091   Max.   :187.000
```

Start of Principal Component Analysis

We are now able to start our PCA. This is done by using prcomp() to perform the analysis and simplify the data into the PCs. Then we use the summary() function and find that, in this case, we are given 14 PCs since there are 14 variables. We are also able to determine that PC1 and PC2 contain the most variation where PC1 had 97% of the variation and PC2 had 3% of the variation. The rest of the PCs have less that 1% variation and therefore don't matter as much. Additionally, we are able to look at the variation of each specific drug and can take a quick glimpse at the first 40.

```
#running analysis
pca <- prcomp(drug)
summary(pca)
```

```
## Importance of components:
##                             PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     301.6497 54.0552 4.93174 2.39577 1.41636 1.12415 0.86113
## Proportion of Variance   0.9685  0.0311 0.00026 0.00006 0.00002 0.00001 0.00001
## Cumulative Proportion    0.9685  0.9996 0.99988 0.99994 0.99996 0.99997 0.99998
##                            PC8    PC9 PC10   PC11   PC12   PC13    PC14
## Standard deviation     0.85307 0.6451 0.61 0.4498 0.3757 0.3011 0.09513
## Proportion of Variance 0.00001 0.0000 0.00 0.0000 0.0000 0.0000 0.00000
## Cumulative Proportion  0.99999 1.0000 1.00 1.0000 1.0000 1.0000 1.00000
```

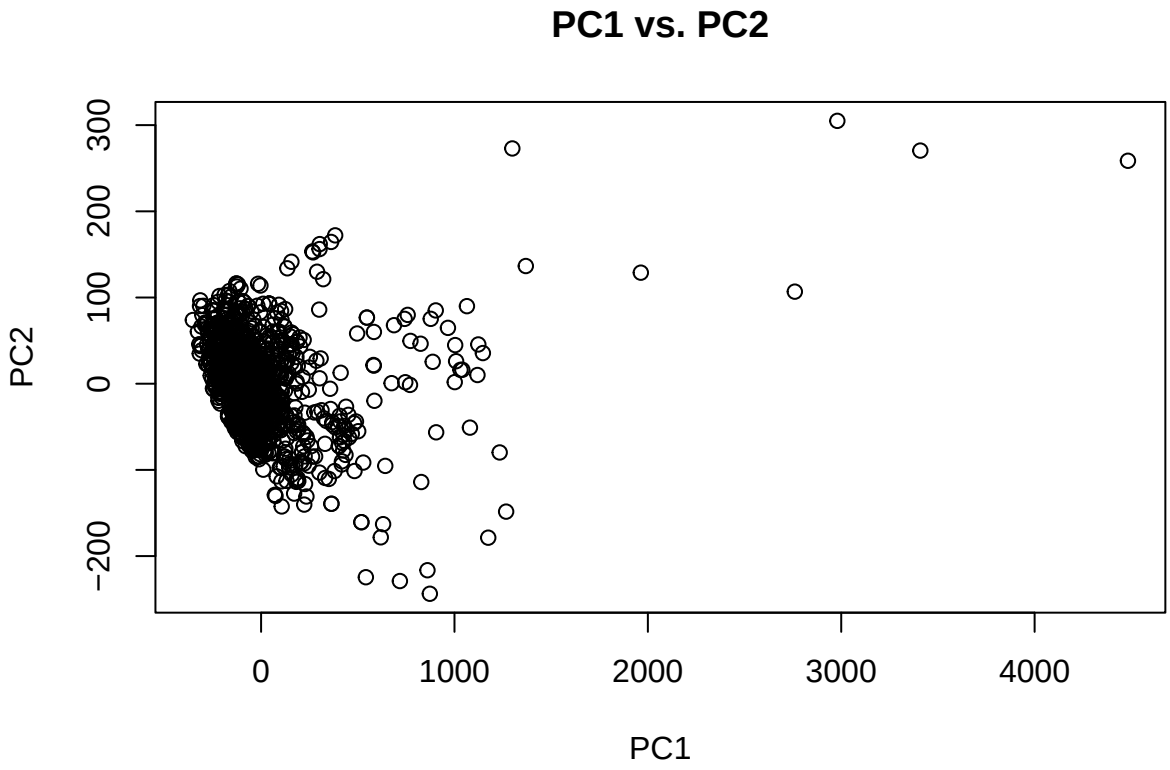
```
pca$x[1:39,1]
```

```
##                   ABACAVIR                   ABARELIX
##                 -92.421373                1079.947176
##                ACAMPROSATE                   ACARBOSE
##                -197.276297                 321.121901
##                 ACEBUTOLOL                ACETAMIOPHEN
##                 -50.312235                -237.421235
##               ACETAZOLAMIDE                ACETIC ACID
##                 -147.885864                -327.054692
##               ACETOHEXAMIDE       ACETOHYDROXAMIC ACID
##                  -60.069964                -308.722252
##              ACETOPHENAZINE                ACETRIZOATE
##                   5.715198                 148.372406
##               ACETYLCHOLINE             ACETYLCYSTEINE
##                 -250.123557                -220.144342
##             ACETYLDIGITOXIN                  ACITRETIN
##                  425.741715                 -74.286912
##                  ACRISORCIN                ACRIVASTINE
##                 -200.103854                 -51.259818
##                   ACYCLOVIR                   ADAPALENE
##                 -143.583228                   6.254410
##          ADEFOVIR DIPIVOXIL                   ADEOSINE
##                  132.229603                 -97.164589
##              ALATROFLOXACIN                ALBENDAZOLE
##                  182.811874                -124.264903
##                   ALBUTEROL ALCLOMETASONE DIPROPIONATE
##                 -146.644794                 129.170627
##                 ALENDRONATE                  ALFENTANIL
##                 -106.399738                  23.908452
##                   ALFUZOSIN                 ALITRETIOIN
##                    7.831018                -101.905719
##                  ALLOPURIOL                 ALMOTRIPTAN
##                 -242.708299                 -62.274225
##                   ALOSETRON            ALPHA-TOCOPHEROL
##                 -101.731571                  17.551603
##                  ALPRAZOLAM                 ALPROSTADIL
##                  -92.108594                 -30.824557
##                 ALTRETAMINE                  AMANTADINE
##                 -182.355088                -245.689950
##                   AMBEONIUM
##                  127.903286
```

Plotting PC1 vs. PC2

Here we are plotting the variation of of each drug PC1 vs. the variation of each drug in PC2. This graph helps us find patterns within the each variation. In this specific graph there aren't many major patters found, but we are able to observe that there are a lot of drugs in PC1 that are centered around zero, with a few other drugs stretching out all the way to 3000 to 4000. Additionally, this large variation is only in the positive direction meaning there aren't as many values lower than the calculated ones. In PC2, the variation is much more compact with the largest variation being around 200 is the positive and negative direction.

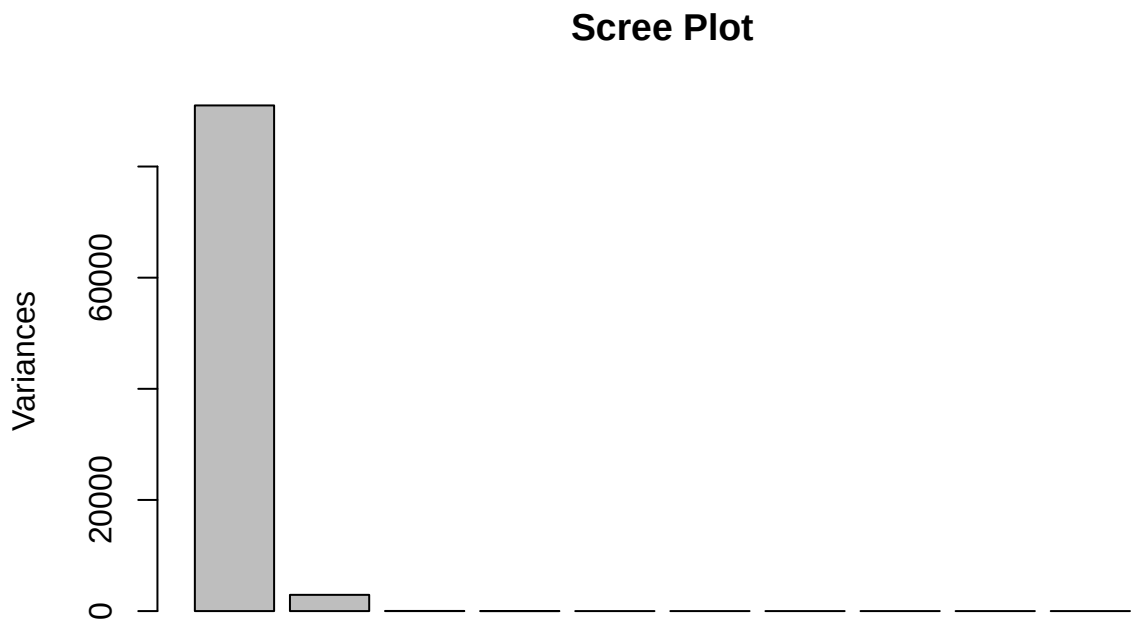
```
#plotting variance
plot(pca$x[,1], pca$x[,2], xlab = "PC1", ylab = "PC2", main = "PC1 vs. PC2")
```



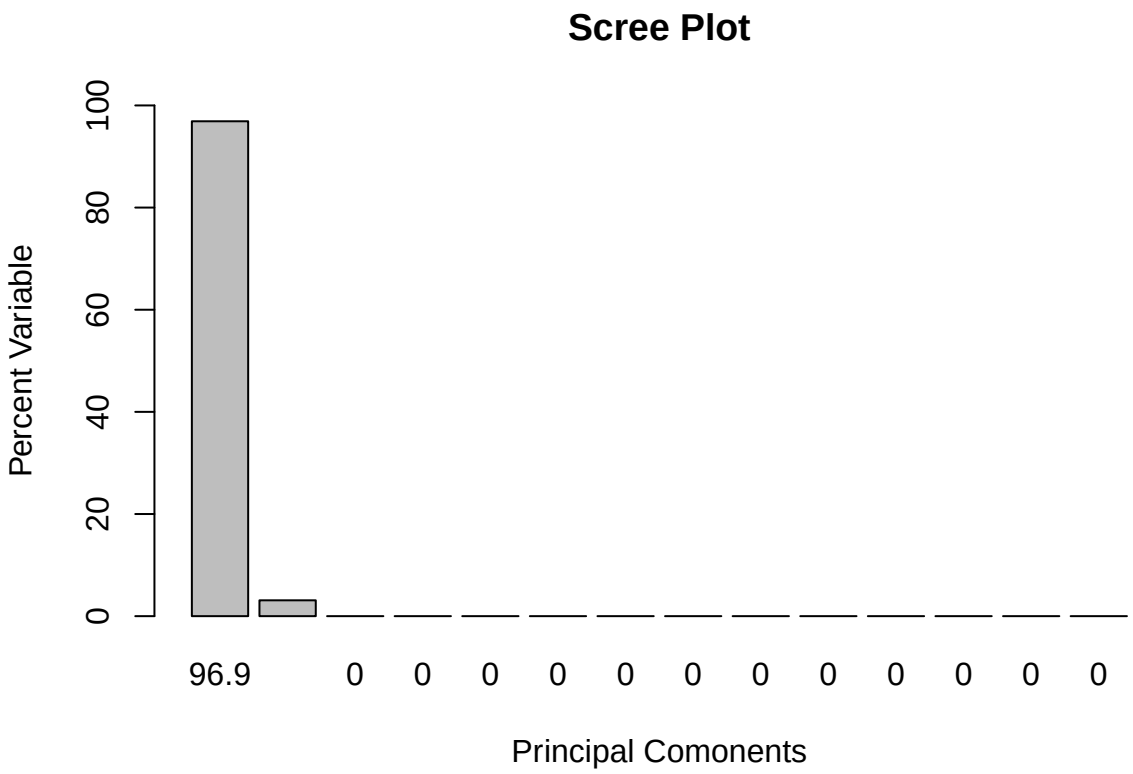
Scree Plot and Prettier Scree Plot

Following the plot of the variance, we decided to also create a scree plot to help visualize the variance of each PC. Firstly we use the screeplot() function, but the plot itself isn't labeled too well so we used ggplot2 to create a better plot. Firstly we used the data we had to create our own plots of the variance and then labeled the x and y axis and gave it a title. In general, this plot helps to visualize how little variance the other PCs have compared to PC1. Even PC2 is incredibly small compared to it, but PC2 still contains three percent which is still impactful of the data.

```
#screeplot
screeplot(pca, main = "Scree Plot")
```



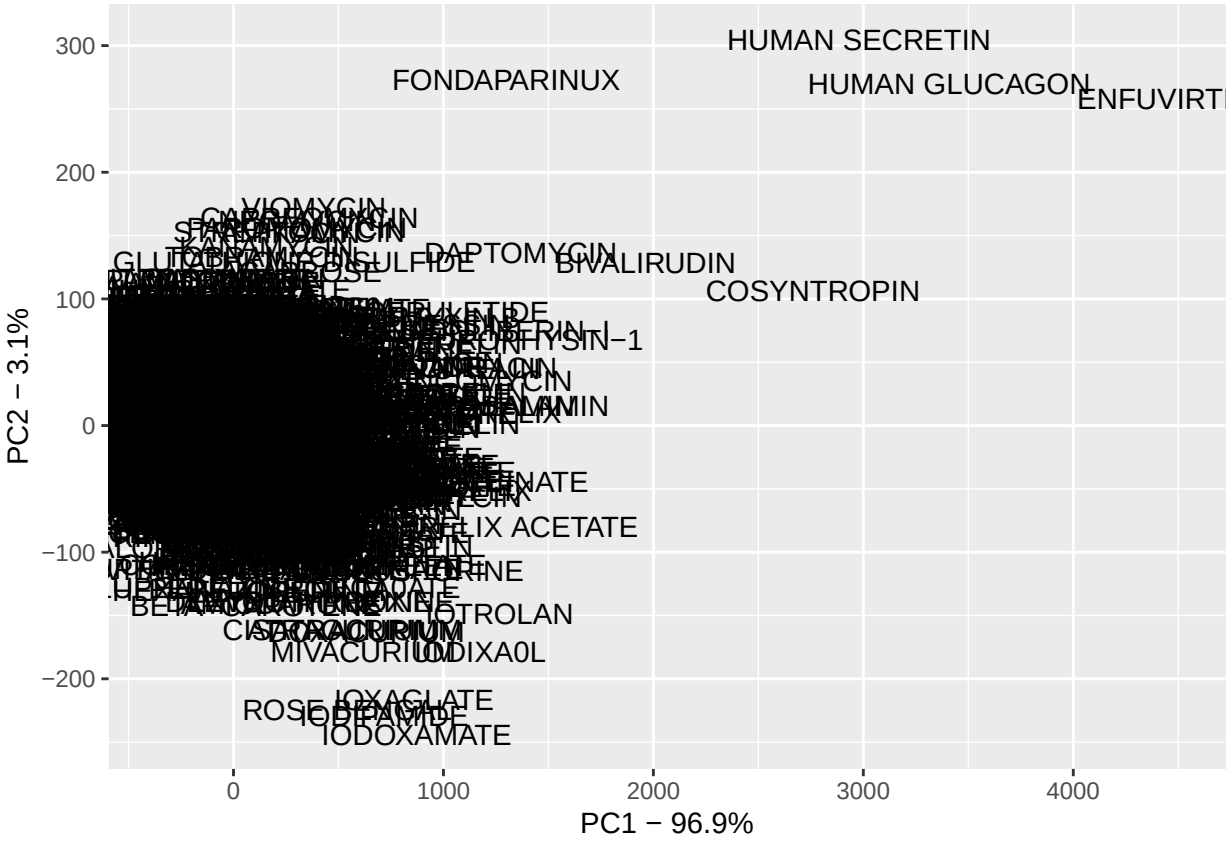

```
#revising the screeplot
drug.variance <- pca$sdev^2
drug.variance.per <- round(drug.variance/sum(drug.variance)*100,1)
barplot(drug.variance.per, ylim = c(0, 100), names.arg = drug.variance.per,
        main = "Scree Plot",
        xlab = "Principal Comonents",
        ylab = "Percent Variable")
```



Text Plot

Another very helpful plot in terms of understanding PCA is a text plot. This can be done by using ggplot2 where we also created specific titles that displayed the amount of variance in each PC. Additionally, we decided to use this becuase it helps to label each data point meaning we can see the drugs that vary the most from the other. When we create the plot we can observe that the drugs that have the most variance in PC1 are enfuvirtide, human glucagon, human secretin, cosyntropin.

```
#plotting
drug.data <- data.frame(Sample = rownames(pca$x), X = pca$x[,1],
                        Y = pca$x[,2])
ggplot(data = drug.data, aes(x = X, y =Y, label =Sample)) +
  geom_text() +
  xlab(paste("PC1 - ", drug.variance.per[1], "%", sep = ""))  +
  ylab(paste("PC2 - ", drug.variance.per[2], "%", sep = ""))
```



Finding the Top Two Variables with the most Variance

One of the most important actions that can be done once the Principal Component Analysis is done is finding the variables in the dataset that contributed to the most variance and therefore the most success of the drugs. We do this by finding the absolute values of the loading values for each variable and then ranking them from highes to lowest. When this is done, we can find that "MW" or molecular weight and "TPSA" or the topological polar surface area. Molecular weight affects the absorbtion of the drug, meaning that the higher the molecular weight, the lower the absorbtion. This means it would be better for drugs to have a lower molecular weight so the absorbtion is higher. TPSA has to do with the drugs ability to permeate cells which is important since that is how the drug takes effect on the body.

```
drug_loading <- pca$rotation[,1]
drug_loading
```

```
##           logS        logSpH7           logP           logD        X2C9pKi
##  -0.0008123991   0.0011237261  -0.0005458531   0.0023574450   0.0007378755
##      hERGpIC50            BBB    Pgpcategory             MW            HBD
##  -0.0010615570  -0.0006790201   0.0007279316   0.9361005757   0.0113384415
##            HBA           TPSA    Flexibility RotatableBonds
##   0.0219351913   0.3495631958   0.0001008758   0.0300186721
```

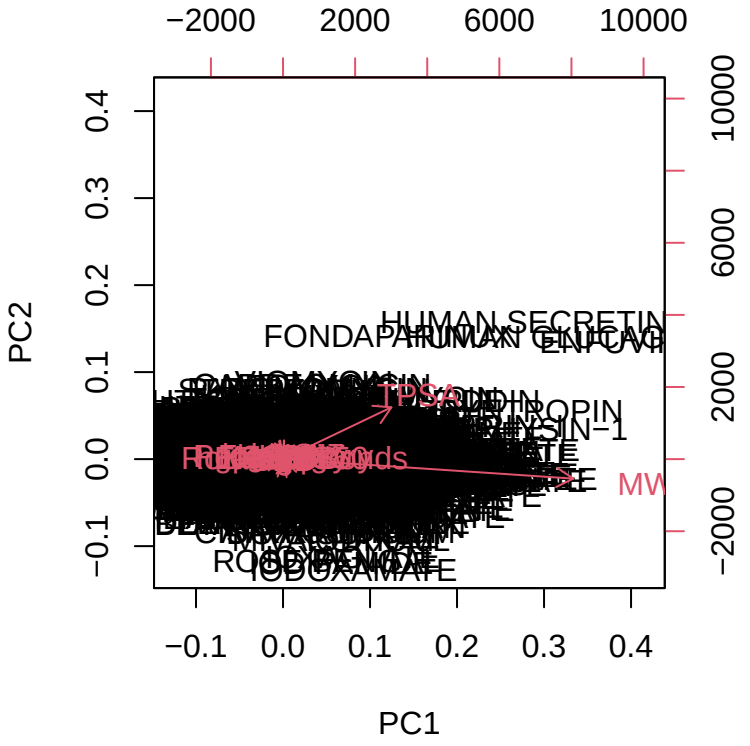
```
drug_score <- abs(drug_loading)
ranked <- sort(drug_score, decreasing = T)
top10 <- names(ranked[1:2])
top10
```

```
## [1] "MW"   "TPSA"
```

Creating a Biplot

Lastly, we created a biplot which is able to show the variation of both the variables affecting the drugs and the drugs themselves. This biplot helps to further visualize how molecular weight and topological polar surface area have more variance than the other variables.

```
biplot(pca)
```



Conclusion

To sum up everything that has been stated, we read in data describing different commercially available drugs. Then, we used prcomp() to simplify the multidimensional data into PCs. Then we found that PC1 had 97% of the variance and PC2 had 3% of the variance while all the other variables have very little variance. Next we created many different plots to help visualize the variance and show which variables and drugs had the most variance. This was done using scree plot, scatter plots, and biplots. Lastly, we used the loading data to find the highest two variables that cause the most commercial success in drugs. We found that molecular weight and topoligical polar surface area.