###Project 1

###Shivani Patel sjp2742

When searching for datasets to join for this project, I was looking for some that related to my current interests and curiosities. I came across one dataset that I really liked, and that related to what I am doing in my life right now- medical school applications. I found this particular dataset through the following website: https://vincentarelbundock.github.io/Rdatasets/datasets.html. This dataset contains information relating to MCAT scores, college GPAs, and medical school acceptances for about 55 separate students.

The Accept variable shows if a student was accepted or denied form medical school, and the Acceptance variable does this in numeric form. The Sex variable shows if a student was female or male, the Apps variable shows how many apps each student submitted, and the ID variable shows the student ID's that I made in the excel sheet prior to reading it into R Studio. The BCPM variable shows college science GPA's, and the GPA variable shows overall GPA. The MCAT variable shows individual MCAT scores, and the VR, PS, WS, and BS variables show the different individual section scores of the MCAT for each student. Overall, I expect all GPA types to be higher for individuals with higher MCAT and MCAT subsection scores. I also expect the scores, acceptances, apps, and GPAs to be similar for both genders.

1:

```
library(tidyverse)
library(ggplot2)
library(dplyr)

Comp_Bio_MedGPA <- read.csv("ShivCompBio.csv")

data("Comp_Bio_MedGPA")
glimpse(Comp_Bio_MedGPA)
```

```
## Observations: 54
## Variables: 12
## $ Accept     <fct> D, A, A, A, A, A, A, D, A, A, A, A, A, D, D, A, D, A, D,...
## $ Acceptance <int> 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0,...
## $ Sex        <fct> F, M, F, F, F, M, M, M, F, F, F, F, M, M, M, F, M, M, M,...
## $ BCPM       <dbl> 3.59, 3.75, 3.24, 3.74, 3.53, 3.59, 3.85, 3.26, 3.74, 3....
## $ GPA        <dbl> 3.62, 3.84, 3.23, 3.69, 3.38, 3.72, 3.89, 3.34, 3.71, 3....
## $ VR         <int> 11, 12, 9, 12, 9, 10, 11, 11, 8, 9, 11, 11, 8, 9, 11, 12...
## $ PS         <int> 9, 13, 10, 11, 11, 9, 12, 11, 10, 9, 9, 8, 10, 9, 8, 8, ...
## $ WS         <int> 9, 8, 5, 7, 4, 7, 6, 8, 6, 6, 8, 4, 7, 4, 6, 8, 8, 9, 5,...
## $ BS         <int> 9, 12, 9, 10, 11, 10, 11, 9, 11, 10, 11, 8, 10, 10, 7, 1...
## $ MCAT       <int> 38, 45, 33, 40, 35, 36, 40, 39, 35, 34, 39, 31, 35, 32, ...
## $ Apps       <int> 5, 3, 19, 5, 11, 5, 5, 7, 5, 11, 6, 9, 5, 8, 15, 6, 6, 1...
## $ ID         <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
```

```
Comp_Bio_MedGPA %>% pivot_wider(names_from = "Sex", values_from = "MCAT") %>%
    pivot_longer(c("M", "F"), names_to = "Sex", values_to = "MCAT") %>%
    na.omit() %>% glimpse()
```

```
## Observations: 54
## Variables: 12
## $ Accept     <fct> D, A, A, A, A, A, A, D, A, A, A, A, A, D, D, A, D, A, D,...
## $ Acceptance <int> 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0,...
## $ BCPM       <dbl> 3.59, 3.75, 3.24, 3.74, 3.53, 3.59, 3.85, 3.26, 3.74, 3....
## $ GPA        <dbl> 3.62, 3.84, 3.23, 3.69, 3.38, 3.72, 3.89, 3.34, 3.71, 3....
## $ VR         <int> 11, 12, 9, 12, 9, 10, 11, 11, 8, 9, 11, 11, 8, 9, 11, 12...
## $ PS         <int> 9, 13, 10, 11, 11, 9, 12, 11, 10, 9, 9, 8, 10, 9, 8, 8, ...
## $ WS         <int> 9, 8, 5, 7, 4, 7, 6, 8, 6, 6, 8, 4, 7, 4, 6, 8, 8, 9, 5,...
## $ BS         <int> 9, 12, 9, 10, 11, 10, 11, 9, 11, 10, 11, 8, 10, 10, 7, 1...
## $ Apps       <int> 5, 3, 19, 5, 11, 5, 5, 7, 5, 11, 6, 9, 5, 8, 15, 6, 6, 1...
## $ ID         <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
## $ Sex        <chr> "F", "M", "F", "F", "F", "M", "M", "M", "F", "F", "F", "...
## $ MCAT       <int> 38, 45, 33, 40, 35, 36, 40, 39, 35, 34, 39, 31, 35, 32, ...
```

```
compbiogpa <- Comp_Bio_MedGPA %>% select(ID, GPA, BCPM, Sex) %>%
    distinct()
glimpse(compbiogpa)
```

```
## Observations: 54
## Variables: 4
## $ ID   <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,...
## $ GPA  <dbl> 3.62, 3.84, 3.23, 3.69, 3.38, 3.72, 3.89, 3.34, 3.71, 3.89, 3....
## $ BCPM <dbl> 3.59, 3.75, 3.24, 3.74, 3.53, 3.59, 3.85, 3.26, 3.74, 3.86, 4....
## $ Sex  <fct> F, M, F, F, F, M, M, M, F, F, F, F, M, M, M, F, M, M, M, M, F,...
```

```
compbiomcat <- Comp_Bio_MedGPA %>% select(ID, Accept, Acceptance,
    VR, PS, WS, BS, MCAT, Apps) %>% distinct()
glimpse(compbiomcat)
```

```
## Observations: 54
## Variables: 9
## $ ID         <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
## $ Accept     <fct> D, A, A, A, A, A, A, D, A, A, A, A, A, D, D, A, D, A, D,...
## $ Acceptance <int> 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0,...
## $ VR         <int> 11, 12, 9, 12, 9, 10, 11, 11, 8, 9, 11, 11, 8, 9, 11, 12...
## $ PS         <int> 9, 13, 10, 11, 11, 9, 12, 11, 10, 9, 9, 8, 10, 9, 8, 8, ...
## $ WS         <int> 9, 8, 5, 7, 4, 7, 6, 8, 6, 6, 8, 4, 7, 4, 6, 8, 8, 9, 5,...
## $ BS         <int> 9, 12, 9, 10, 11, 10, 11, 9, 11, 10, 11, 8, 10, 10, 7, 1...
## $ MCAT       <int> 38, 45, 33, 40, 35, 36, 40, 39, 35, 34, 39, 31, 35, 32, ...
## $ Apps       <int> 5, 3, 19, 5, 11, 5, 5, 7, 5, 11, 6, 9, 5, 8, 15, 6, 6, 1...
```

I chose to break the dataset that I found and really liked into two separate datasets, and then rejoin them. My dataset was already tidy when I found it, so I made it untidy using pivot_wider, and then tidy again using pivot_longer, followed by na.omit to get rid of NAs. I then broke my dataset into two separate ones, both with the "ID" column in common.

2:

```
Comp_Bio_MedGPA_Edits <- compbiogpa %>% full_join(compbiomcat) %>%
    glimpse()
```

```
## Observations: 54
## Variables: 12
## $ ID         <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
## $ GPA        <dbl> 3.62, 3.84, 3.23, 3.69, 3.38, 3.72, 3.89, 3.34, 3.71, 3....
## $ BCPM       <dbl> 3.59, 3.75, 3.24, 3.74, 3.53, 3.59, 3.85, 3.26, 3.74, 3....
## $ Sex        <fct> F, M, F, F, F, M, M, M, F, F, F, F, M, M, M, F, M, M, M,...
## $ Accept     <fct> D, A, A, A, A, A, A, D, A, A, A, A, A, D, D, A, D, A, D,...
## $ Acceptance <int> 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0,...
## $ VR         <int> 11, 12, 9, 12, 9, 10, 11, 11, 8, 9, 11, 11, 8, 9, 11, 12...
## $ PS         <int> 9, 13, 10, 11, 11, 9, 12, 11, 10, 9, 9, 8, 10, 9, 8, 8, ...
## $ WS         <int> 9, 8, 5, 7, 4, 7, 6, 8, 6, 6, 8, 4, 7, 4, 6, 8, 8, 9, 5,...
## $ BS         <int> 9, 12, 9, 10, 11, 10, 11, 9, 11, 10, 11, 8, 10, 10, 7, 1...
## $ MCAT       <int> 38, 45, 33, 40, 35, 36, 40, 39, 35, 34, 39, 31, 35, 32, ...
## $ Apps       <int> 5, 3, 19, 5, 11, 5, 5, 7, 5, 11, 6, 9, 5, 8, 15, 6, 6, 1...
```

I used a full_join to join the two separated datasets by "ID", as I did not want to lose any columns or add any NAs.

No cases were dropped in either dataset.

3:

```
Comp_Bio_MedGPA_Edits %>% select(Accept, GPA) %>% filter(Accept ==
    "A", GPA == max(GPA)) %>% glimpse()
```

```
## Observations: 2
## Variables: 2
## $ Accept <fct> A, A
## $ GPA    <dbl> 3.97, 3.97
```

```
Comp_Bio_MedGPA_Edits <- Comp_Bio_MedGPA_Edits %>% arrange(desc(BS)) %>%
    mutate(BS_Percent = BS/MCAT * 100) %>% glimpse()
```

```
## Observations: 54
## Variables: 13
## $ ID         <int> 47, 20, 2, 23, 24, 46, 5, 7, 9, 11, 17, 18, 31, 36, 40, ...
## $ GPA        <dbl> 3.97, 3.89, 3.84, 3.91, 3.88, 3.36, 3.38, 3.89, 3.71, 3....
## $ BCPM       <dbl> 3.98, 3.95, 3.75, 4.00, 3.98, 3.25, 3.53, 3.85, 3.74, 4....
## $ Sex        <fct> F, M, M, M, M, F, F, M, F, F, M, M, M, M, M, F, F, M, F,...
## $ Accept     <fct> A, A, A, A, A, D, A, A, A, A, D, A, A, A, A, A, A, A, A,...
## $ Acceptance <int> 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ VR         <int> 11, 13, 12, 10, 9, 11, 9, 11, 8, 11, 8, 9, 10, 10, 10, 1...
## $ PS         <int> 13, 14, 13, 13, 10, 8, 11, 12, 10, 9, 8, 10, 12, 11, 9, ...
## $ WS         <int> 10, 8, 8, 6, 8, 9, 4, 6, 6, 8, 8, 9, 6, 8, 8, 8, 7, 7, 6...
## $ BS         <int> 14, 13, 12, 12, 12, 12, 11, 11, 11, 11, 11, 11, 11, 11, ...
## $ MCAT       <int> 48, 48, 45, 41, 39, 40, 35, 40, 35, 39, 35, 39, 39, 40, ...
## $ Apps       <int> 6, 5, 3, 17, 17, 12, 11, 5, 5, 6, 6, 1, 7, 8, 12, 5, 5, ...
## $ BS_Percent <dbl> 29.16667, 27.08333, 26.66667, 29.26829, 30.76923, 30.000...
```

```
Comp_Bio_MedGPA_Edits %>% summarize(mean(WS), sd(WS))
```

```
##   mean(WS)   sd(WS)
## 1 7.148148 1.606677
```

```
Comp_Bio_MedGPA_Edits %>% summarize(n_distinct(MCAT))
```

```
##   n_distinct(MCAT)
## 1               14
```

```
Comp_Bio_MedGPA_Edits %>% summarize(mean(VR), sd(VR), IQR(VR))
```

```
##   mean(VR)    sd(VR) IQR(VR)
## 1 9.814815 1.759868    2.75
```

```
Comp_Bio_MedGPA_Edits %>% summarize(mad(PS), median(PS))
```

```
##   mad(PS) median(PS)
## 1  1.4826         10
```

```
Comp_Bio_MedGPA_Edits %>% group_by(Accept) %>% summarize(max_Apps = max(Apps)) %>%
    glimpse()
```

```
## Observations: 2
## Variables: 2
## $ Accept   <fct> A, D
## $ max_Apps <int> 19, 24
```

```
Comp_Bio_MedGPA_Edits %>% group_by(Sex) %>% summarize(min_BS = min(BS)) %>%
    glimpse()
```

```
## Observations: 2
## Variables: 2
## $ Sex    <fct> F, M
## $ min_BS <int> 8, 6
```

```
Comp_Bio_MedGPA_Edits %>% group_by(Accept, Sex) %>% summarize(n = n())
```

```
## # A tibble: 4 x 3
## # Groups:   Accept [2]
##   Accept Sex       n
##   <fct>  <fct> <int>
## 1 A      F        18
## 2 A      M        12
## 3 D      F        10
## 4 D      M        14
```

```
Comp_Bio_MedGPA_Edits %>% group_by(Accept) %>% summarize(var_GPA_BCPM = var(GPA,
    BCPM)) %>% glimpse()
```

```
## Observations: 2
## Variables: 2
## $ Accept       <fct> A, D
## $ var_GPA_BCPM <dbl> 0.05627356, 0.07175978
```

```
Correlation_Matrix <- Comp_Bio_MedGPA_Edits %>% select_if(is.numeric)

cor(Correlation_Matrix) %>% glimpse()
```

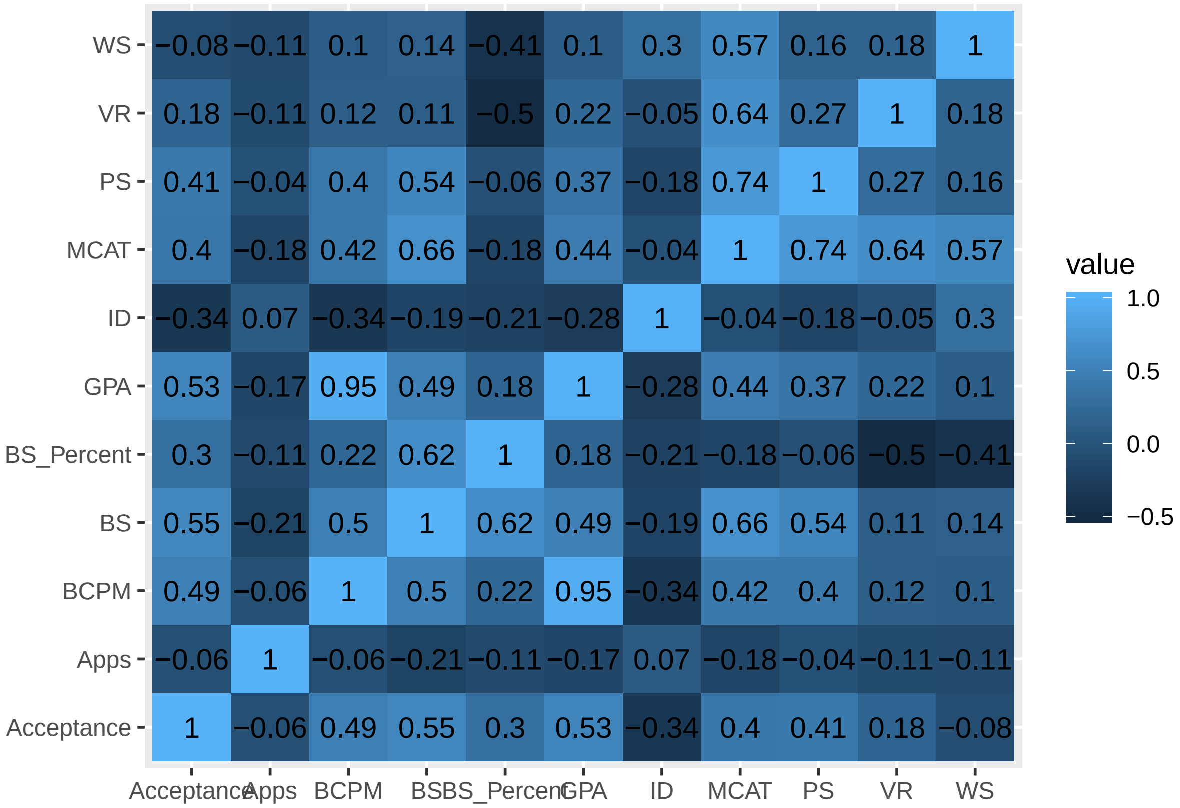
```
##  num [1:11, 1:11] 1 -0.2838 -0.339 -0.3402 -0.0495 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:11] "ID" "GPA" "BCPM" "Acceptance" ...
##   ..$ : chr [1:11] "ID" "GPA" "BCPM" "Acceptance" ...
```

When selecting Accept (categorical) and GPA, and filtering by only accepted students and max GPA, I found that the two students with 3.97 max GPAs were both accepted to medical schools. When arranging by descending BS scores, and using mutate to create the variable BS_Percent which showed BS score percentage of MCAT score, I found that the BS Percentage overall decreased as BS decreased. When summarizing the mean and standard deviation of WS scores, I found that the mean was 7.15 and standard deviation was 1.61. When summarizing the number of distinct MCAT scores, I found that it was 14 scores. When summarizing the mean, standard deviation, and inter quartile range of VR scores, I found them to be 9.81, 1.76, and 2.75 respectively. When summarizing the mad and median of PS scores, I found them to be 1.48 and 10 respectively.

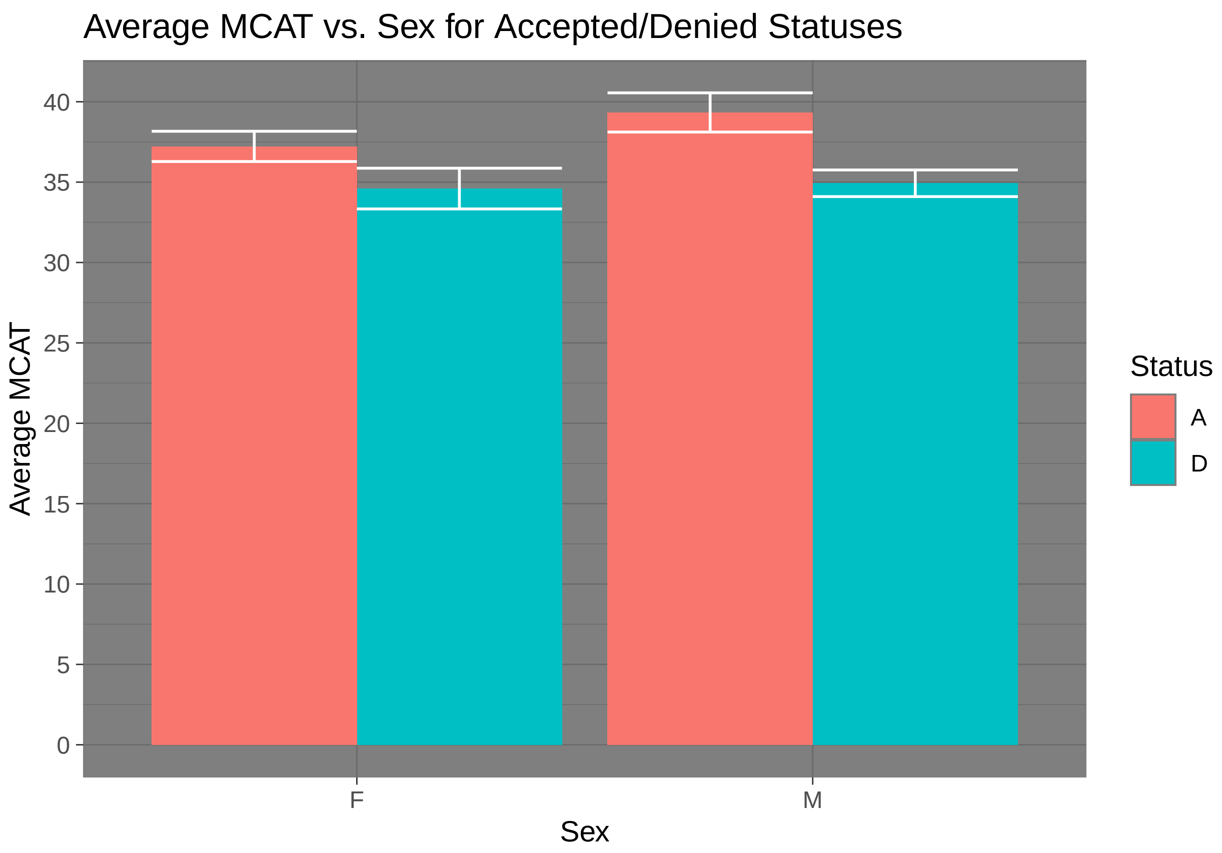
When grouping by Accept (categorical) and summarizing the max apps submitted, I found that for accepted students the maximum number of apps submitted was 19, and for denied students the number was 24. When grouping by Sex (categorical) and summarizing the minimum BS score, I found that for females, the minimum BS score was 8, and for males it was 6. When grouping by Accept (categorical) and Sex (categorical), and summarizing the number of each (n), I found that there were 18 accepted females, 12 accepted males, 10 denied females, and 14 denied males. When grouping by Accept (categorical) and summarizing the variance of GPA and BCPM, I found the variance to be around 0.056 for accepted individuals, and 0.072 for denied individuals. I further went on to make a correlation matrix of my numerical variables. One interesting thing that I noticed was that the number of apps had a negative correlation with all other numeric variables besides itself.

4:

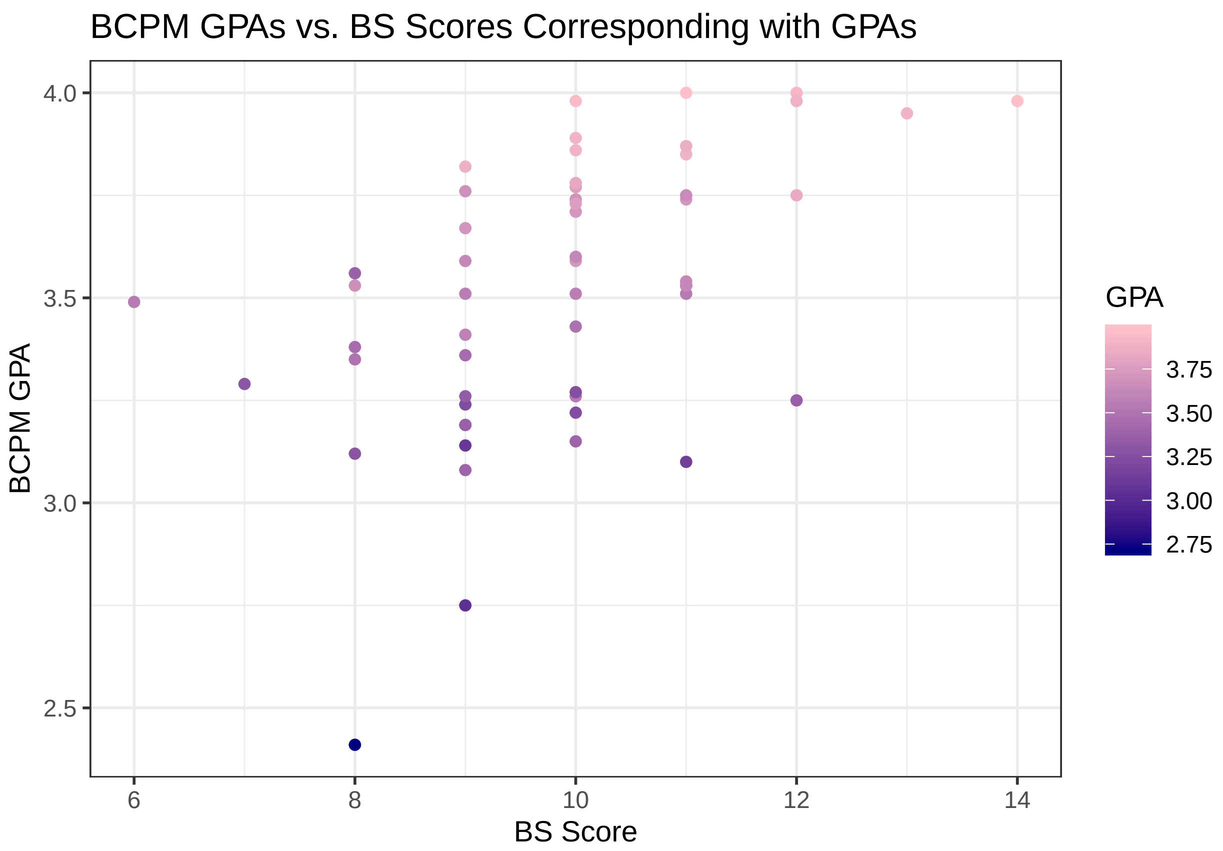
```
Comp_Bio_MedGPA_Edits %>% select_if(is.numeric) %>% cor %>% as.data.frame %>%
    rownames_to_column %>% pivot_longer(-1) %>% ggplot(aes(rowname,
    name, fill = value)) + geom_tile() + geom_text(aes(label = round(value,
    2))) + xlab("") + ylab("")
```



```r
ggplot(Comp_Bio_MedGPA_Edits, aes(x = Sex, y = MCAT, fill = Accept)) +
    geom_bar(stat = "summary", fun.y = "mean", position = "dodge") +
    scale_y_continuous(breaks = seq(0, 40, 5)) + geom_errorbar(stat = "summary",
    position = "dodge", col = c("white")) + ggtitle("Average MCAT vs. Sex for Accepted/Denied Statuses")
    ylab("Average MCAT") + labs(fill = "Status") + theme_dark()
```



```r
ggplot(Comp_Bio_MedGPA_Edits, aes(BS, BCPM, color = GPA)) + geom_point() +
    ggtitle("BCPM GPAs vs. BS Scores Corresponding with GPAs") +
    ylab("BCPM GPA") + xlab("BS Score") + theme_bw() + scale_color_gradient(low = "navy blue",
    high = "pink")
```



I made a correlation heatmap of all of my numeric variables. This showed the relative correlations of all of the variables with each other, using a lighter shade of blue for stronger correlations, and a darker shade for weaker correlations. It seems as though there is a relatively strong correlation between MCAT scores and MCAT sub-section scores, as well as between GPA and Science GPA (BCPM).

I made a side-by-side bar chat with standard error bars, which showed relationship between average MCAT scores and sex for accepted/denied students. It showed that on average, accepted males performed better on the MCAT than accepted females did. However, it also showed that denied males and females performed relatively similarly on the MCAT as each other.

I made a scatterplot, which showed the relationship between BCPM GPAs of students and their BS component scores, in relation to their overall GPAs. It showed that the higher the overall GPAs, the higher the BS scores were, and the higher the BCPM GPAs were. All three of the variables seemed to have a direct positive correlation.

5:

```
Comp_Bio_MedGPA_Edits <- Comp_Bio_MedGPA_Edits %>% select(-ID)
Comp_Bio_Nums <- Comp_Bio_MedGPA_Edits %>% select_if(is.numeric) %>%
    scale
Comp_Bio_PCA <- princomp(Comp_Bio_Nums)
names(Comp_Bio_PCA)
```

```
## [1] "sdev"     "loadings" "center"   "scale"    "n.obs"    "scores"   "call"
```

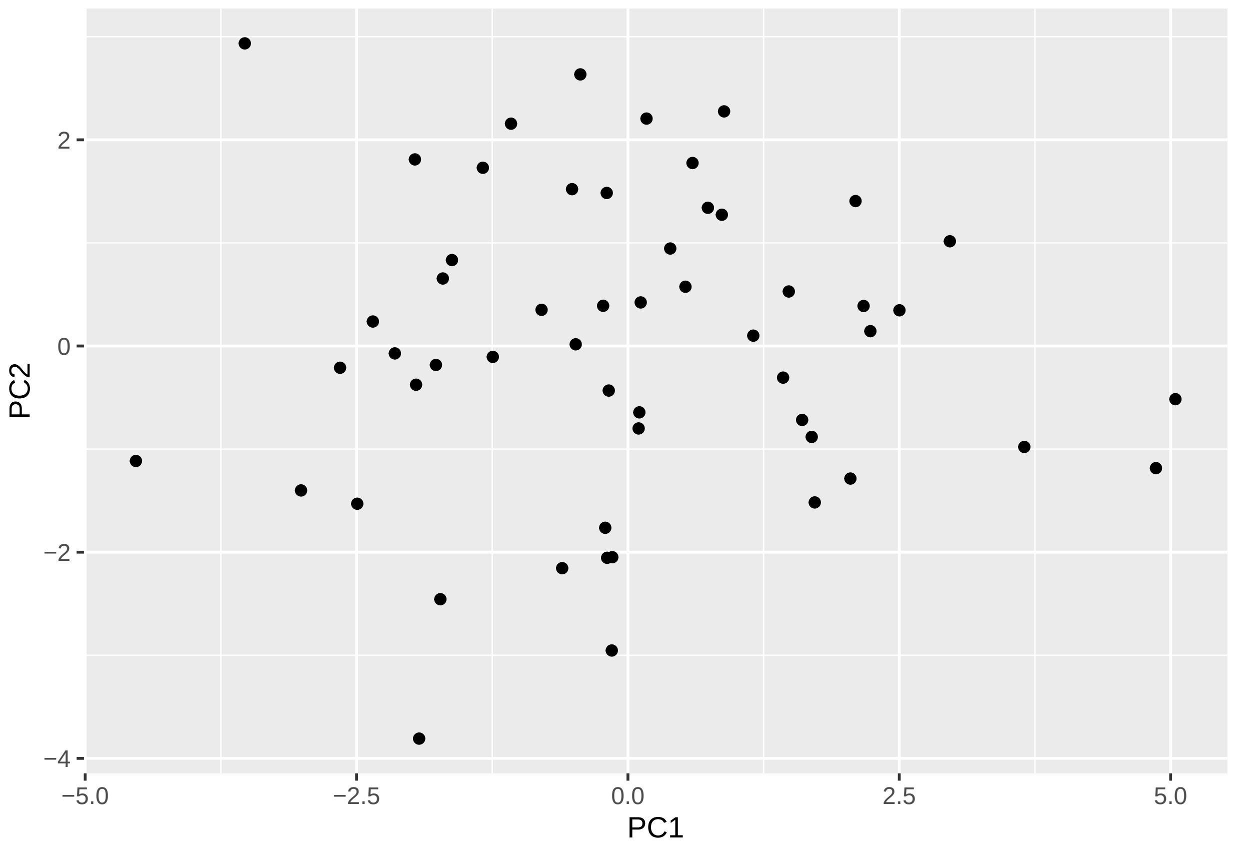
```
summary(Comp_Bio_PCA, loadings = T)
```

```
## Importance of components:
##                           Comp.1   Comp.2    Comp.3    Comp.4     Comp.5
## Standard deviation     1.960720 1.456019 1.0319632 0.9745004 0.92600166
## Proportion of Variance 0.391696 0.215999 0.1085041 0.0967569 0.08736579
## Cumulative Proportion  0.391696 0.607695 0.7161991 0.8129560 0.90032183
##                            Comp.6     Comp.7      Comp.8       Comp.9
## Standard deviation     0.70982982 0.66163708 0.182378250 0.0586425608
## Proportion of Variance 0.05133651 0.04460233 0.003388941 0.0003503836
## Cumulative Proportion  0.95165834 0.99626068 0.999649616 1.0000000000
##                             Comp.10
## Standard deviation     2.356080e-08
## Proportion of Variance 5.655853e-17
## Cumulative Proportion  1.000000e+00
##
## Loadings:
##            Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## GPA         0.403  0.130  0.291  0.469                       0.708
## BCPM        0.393  0.163  0.335  0.400  0.183  0.135  0.126 -0.691
## Acceptance  0.346  0.201  0.161 -0.161 -0.263 -0.390 -0.751
## VR          0.209 -0.439  0.134        -0.561 -0.322  0.382        -0.257
## PS          0.364 -0.124        -0.428         0.711 -0.143        -0.220
## WS          0.144 -0.422 -0.286  0.161  0.638 -0.301 -0.206        -0.243
## BS          0.406  0.219 -0.314 -0.281        -0.190  0.292         0.632
## MCAT        0.424 -0.319 -0.156 -0.202                0.131
## Apps       -0.116         0.696 -0.496  0.398 -0.243  0.182
## BS_Percent         0.617 -0.264 -0.142        -0.172  0.260        -0.646
##            Comp.10
## GPA
## BCPM
## Acceptance
## VR         -0.336
## PS         -0.297
## WS         -0.307
## BS         -0.278
## MCAT        0.792
## Apps
## BS_Percent
```

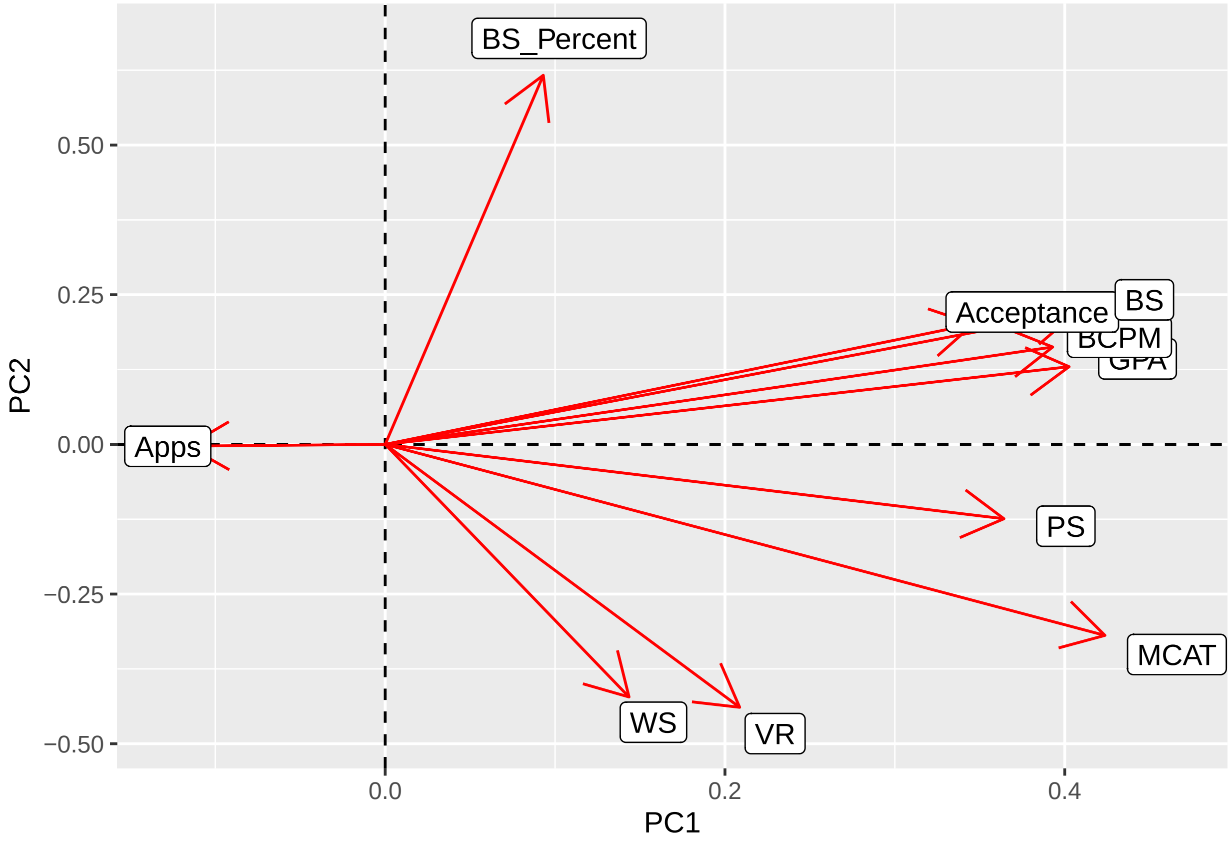
```
EigVal <- Comp_Bio_PCA$sdev^2
EigVal
```

```
##       Comp.1       Comp.2       Comp.3       Comp.4       Comp.5       Comp.6
## 3.844423e+00 2.119991e+00 1.064948e+00 9.496511e-01 8.574791e-01 5.038584e-01
##       Comp.7       Comp.8       Comp.9      Comp.10
## 4.377636e-01 3.326183e-02 3.438950e-03 5.551115e-16
```

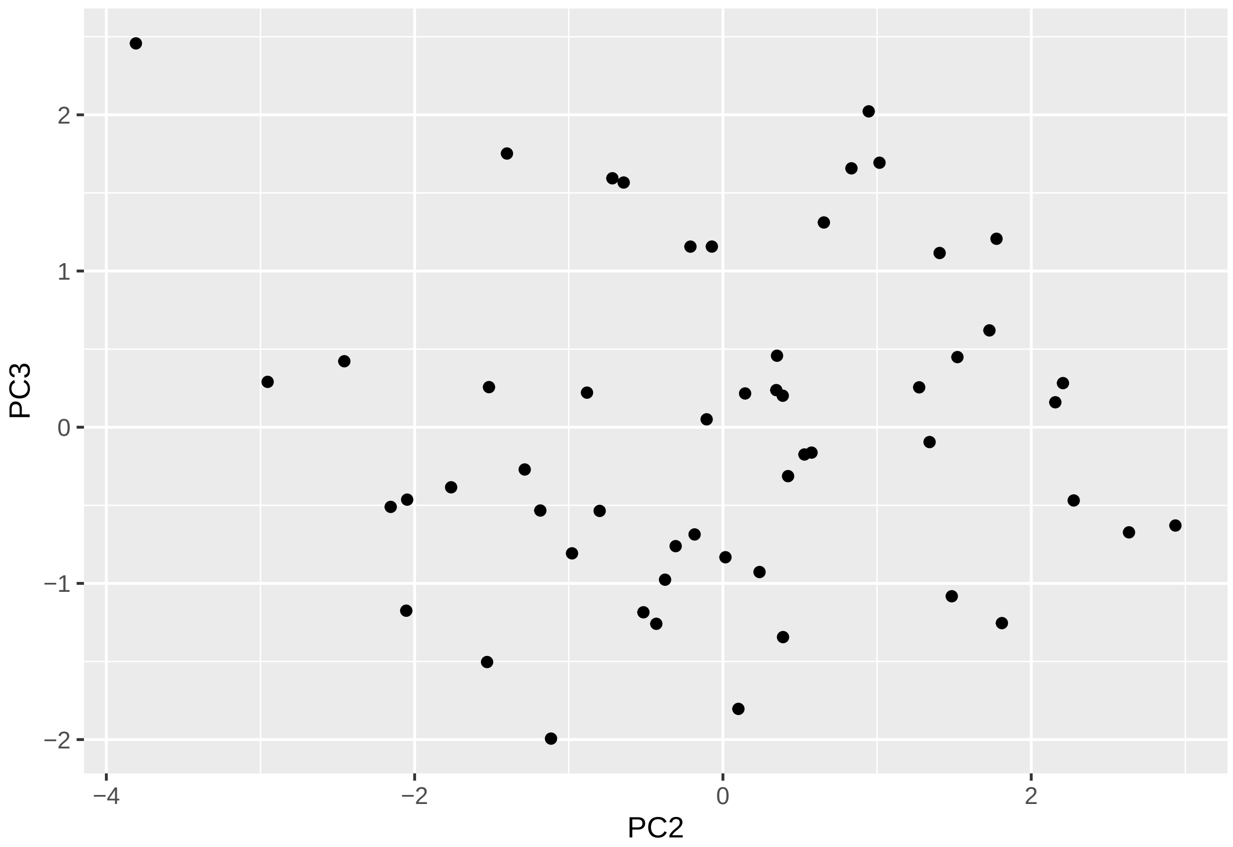
```
Comp_Bio_df1 <- data.frame(PC1 = Comp_Bio_PCA$scores[, 1], PC2 = Comp_Bio_PCA$scores[,
    2])
ggplot(Comp_Bio_df1, aes(PC1, PC2)) + geom_point()
```



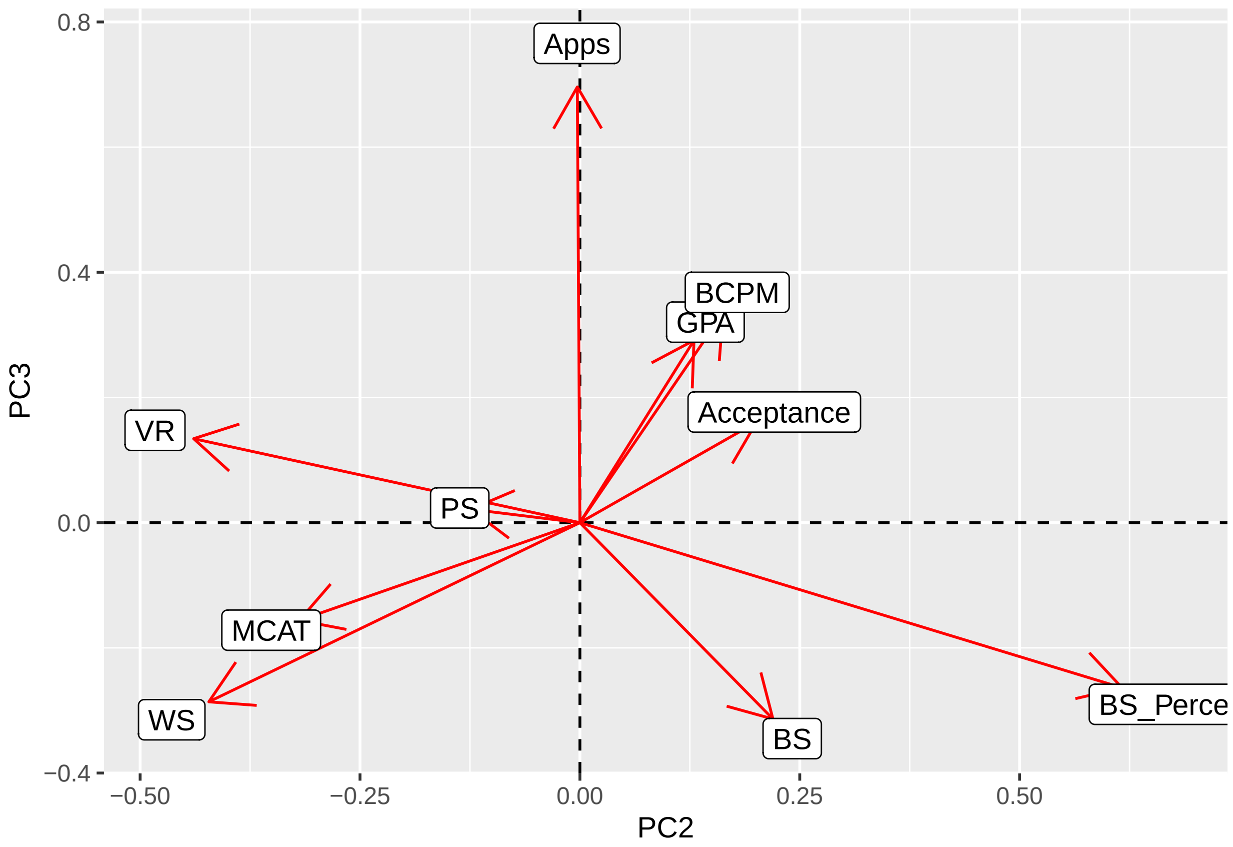
```
Comp_Bio_PCA$loadings[1:10, 1:2] %>% as.data.frame %>% rownames_to_column %>%
    ggplot() + geom_hline(aes(yintercept = 0), lty = 2) + geom_vline(aes(xintercept = 0),
    lty = 2) + ylab("PC2") + xlab("PC1") + geom_segment(aes(x = 0,
    y = 0, xend = Comp.1, yend = Comp.2), arrow = arrow(), col = "red") +
    geom_label(aes(x = Comp.1 * 1.1, y = Comp.2 * 1.1, label = rowname))
```



```
Comp_Bio_df2 <- data.frame(PC2 = Comp_Bio_PCA$scores[, 2], PC3 = Comp_Bio_PCA$scores[,
    3])
ggplot(Comp_Bio_df2, aes(PC2, PC3)) + geom_point()
```



```
Comp_Bio_PCA$loadings[1:10, 2:3] %>% as.data.frame %>% rownames_to_column %>%
    ggplot() + geom_hline(aes(yintercept = 0), lty = 2) + geom_vline(aes(xintercept = 0),
    lty = 2) + ylab("PC3") + xlab("PC2") + geom_segment(aes(x = 0,
    y = 0, xend = Comp.2, yend = Comp.3), arrow = arrow(), col = "red") +
    geom_label(aes(x = Comp.2 * 1.1, y = Comp.3 * 1.1, label = rowname))
```



In this dimensionality reduction step, I cleaned my data by removing the ID variable, normalized my data by scaling it, ran princomp, and then summarized my results. After squaring standard deviations to obtain eigenvalues, I decided on how many principal components to keep. I used Kaiser's rule, and since the 4th component would have had a value less than 1 when squared, I used 3 principal components. These 3 components counted for 71.6% of the variance. A high score on PC1 corresponded with more apps submitted, and lower MCAT score, GPAs, and acceptances. A high score on PC2 corresponded with higher GPAs and acceptances, and overall lower MCAT scores and subscores not including BS. A high score on PC3 corresponded with higher GPAs, acceptances, and apps submitted, and lower MCAT scores and subscores in general.

I then went on to make scatterplots and loadings plots for PC1 and PC2, as well as PC2 and PC3. For the PC1 and PC2 plots, this showed a strong correlation as seen by smaller angles between both types of GPAs and BS/acceptance, and a weak correlation between both types of GPAs and BS percentage. It also showed a stronger correlation between MCAT and PS vs. between MCAT and VR and WS. For the PC2 and PC3 plots, this showed a strong correlation between both types of GPAs, and between MCAT and WS scores. It also showed a weak correlation between apps submitted and VR/PS scores, and between BS scores and BS percentage. All of these results made sense in context of medical school acceptance statistics.