Imran Khan

STAT 515

Professor Daniel Carr

May 5, 2015

Final Paper: College Data

**Section 1: Introduction**

As a student planning on transferring schools next semester, I am forced to look at many factors before making the choice of which school will be the right choice for me. Students have access to hundreds of variables when choosing the right school – from number of majors offered to the quality of the food. I went in to this project not knowing what specific question I was trying to answer. Rather, I wanted to simply explore what each university has to offer in the United States and how different variables would relate to one another.

An increasing number of students across the United States are deciding to attend college to receive there bachelor’s degrees. So much so that it is commonly claimed that the bachelor’s degree is the new high school diploma because people who have one are now so easily available in such large supplies. People from all over the world travel to the United States to receive an education here not only because of the high quality of our higher education programs, but because of the wide range of choices they have to choose from.

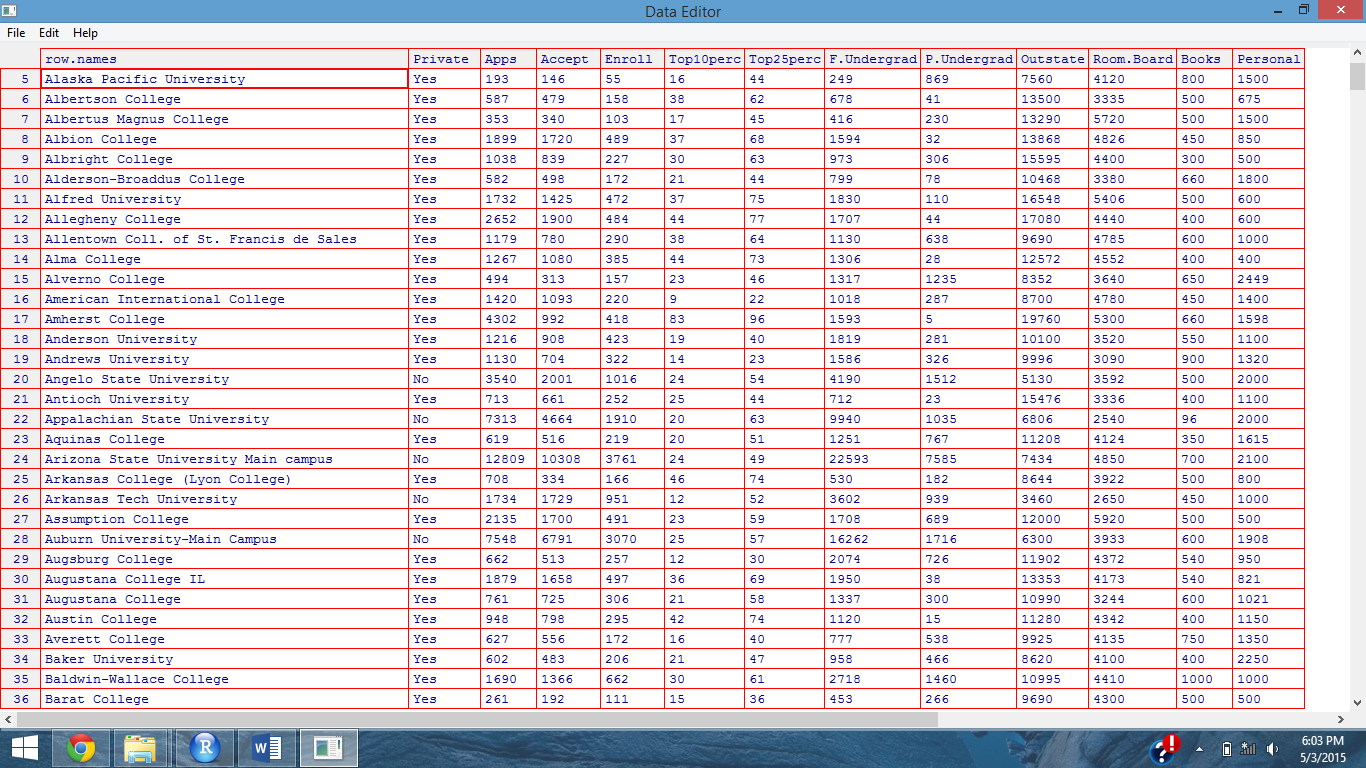
With this project, I – again – am not interested in answering any single question with my data. The goal I am trying to accomplish now is to visualize the data I’ve chosen in a way that makes it useful to the public. The graphs I choose to create may be worthwhile in the future for students looking for the perfect school for their college years. After all, college is supposed to be the best four years of your life. Why spend it anywhere but the perfect school for you?

**Section 2: The Data**

I first set out to find the data I was looking for by doing a quick google search for it. However, the best data I could find was from US News and World Report rankings of schools. I didn’t feel like it was a good dataset for my project, because it was centered about the top rated schools in the country. I wanted my project to become one day useful to anyone who was applying to colleges – not just the country’s brightest students who had the perfect grades and high SAT scores enough to get into one of the top rated schools. I wanted a dataset that had a better look at all of the available colleges.

Another dataset I had found had all of the information I was looking for. However, the price tag for the dataset was at $42. I figured it was out of the scope of the project to have to pay the money for the data, so I kept looking. I also figured it was out of the scope of the project to collect and create the data myself.

Finally, I came across a dataset called “College” from the ISLR R Package which I found on google. This dataset had all the information I wanted, but the data collected was from 1995. I decided to use this data. The data came from the StatLib Library which is maintained by Carnegie Mellon University. Original data was collected from US News and World Report. For more information about the data, you can visit <http://cran.r-project.org/web/packages/ISLR/ISLR.pdf>. The data itself consists of 777 colleges and looks at 18 variables between them, including the number of applications received, the number of full time undergraduates enrolled, the percent of alumni who choose to donate back to the school, and others.

Here is a snapshot of what the data looks like once originally read in:

After reading in the data, I exported it to Microsoft Excel, so that I could do a few quick things to fix it up how I’d like it. I added a few more columns in which I did some simple calculations in order to make working with data easier in R. I saved the file as a CSV and read it back into R in a new file called “colleges.csv”.

*R Code: Section 2*

*#Section 2: reading in the data*

*#http://cran.r-project.org/web/packages/ISLR/ISLR.pdf*

*#data taken from CMU, taken from US News & World Report*

*library(ISLR)*

*College*

*summary(College)*

*library(xlsx)*

*write.xlsx(College, "College.xlsx")*

*colleges <- read.csv("College.csv")*

*colleges*

*summary(colleges)*

**Section 3: Setup**

In setting up the R workspace for the project, I had to import a few libraries so that I could work on the project with methods already written. Some of the important ones that we went over in class include lattice and ggplot2.

Also included in the setup was the adding of the hw function from the earlier homework. The hw function is useful in making the grids produced more easy to see. It (i) removes tick marks from the axes of the grids (ii) makes the labels on the outside of the plot more easily read by making them darker and (iii) gets rid of some of the lighter lines inside of the plot, which again makes it easier to read the contents of the plot. Once the hw function is read into the workspace at the beginning of the project, it can easily be implemented into plots later on by simply adding +hw to the end of the plotting method.

*R Code: Section 3*

*library(ggplot2)*

*library(grid)*

*library(lattice)*

*library(MASS)*

*library(hexbin)*

*library(randomForest)*

*library(MASS)*

*library(car)*

*library(ellipse)*

*hw <- theme\_gray() + theme(*

*strip.background=element\_rect(fill=rgb(.9,.95,1),*

*colour=gray(.5), size=.2),*

*panel.border=element\_rect(fill=FALSE,colour=gray(.50)),*

*axis.text=element\_text(colour="black"),*

*axis.ticks=element\_blank(),*

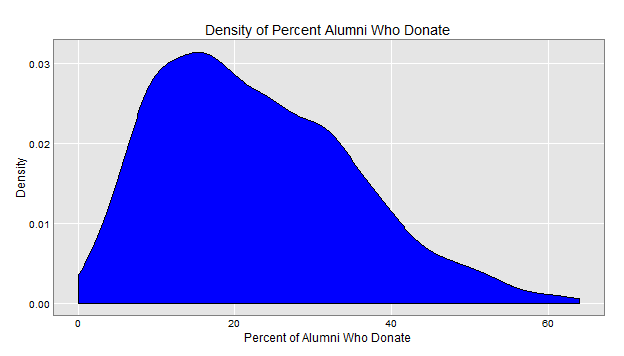
*axis.ticks.margin=unit(-0.05,"cm"),*

*panel.grid.minor = element\_blank()*

*)*

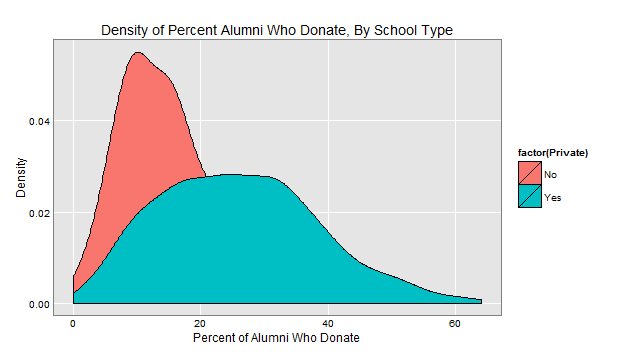
**Section 4: Density Graphs**

One measure of how happy a student is at his/her university is by the percentage of alumni who donate back to the university. Of course, that is not the only thing it measures. It can also be an indicator of how annoying the school is with requesting donations from their graduates, or how successful a graduate from that school becomes later on in life. Regardless, information on the percentage of alumni who donate is given in the dataset I chose, so I decided to explore a little bit on the distribution of alumni who have decided to donate back to their schools. The graph of the distribution can be seen below.



From the above graph, we can easily determine that the mode of the schools percentage alumni who donate is around seventeen percent. The average amount a school gets in donations is about twenty-two percent. There are quite a few schools who get no money from any of their alumni, and there is even a small amount of schools who receive donations from over sixty percent of their alumni.

I wanted to explore this data a little bit further. I decided to split the percent alumni up over the factor “private.” The Private column in the data just mentions a Yes or No whether or not the school is Private. If a school has a value No for this, this means it is a public school – it is funded mainly by the state rather than existing on its own. I predicted that a public school would not receive as much funding from its alumni as would a private school. Below is the plot of what I found.



As can be seen from the graph, when looking at just public schools, the mode of the data is only about fifteen percent. For private schools, the mode jumps a little bit higher. There is an obvious trend that can be observed from this plot – private schools tend to receive more alumni donations than do public schools. It can also be seen that private schools have a wider standard deviation. We will look a little bit further into this in the next section.

Why might a private school receive more money than public school? One possible explanation is that the school spirit at most private schools is relatively high compared to a non-private school. It interests me that more people would donate to their private schools, because (assuming most public school attendees pay instate tuition) private school attendees pay more than their public school counterparts just to attend school. We will look into a bit more of tuition costs in a few sections, too.

In order to create this graph, all that had to be changed from the previous one was changing the aes method of the geom\_density function to contain a fill that was differentiated between the two factors of the Private column. The first graph was a simple use of ggplot2’s geom\_density component.

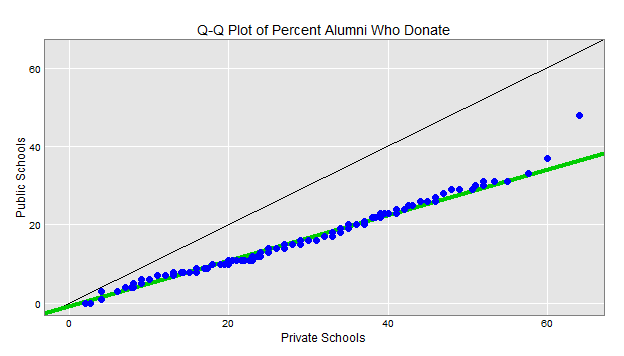
*R Code: Section 4*

*#Section 4: densities of percent alumni who donate*

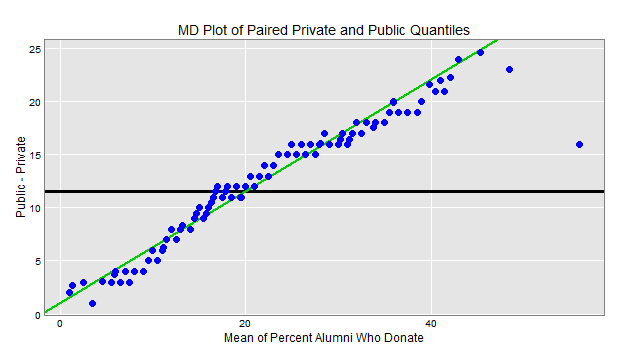
*ggplot(colleges, aes(x = perc.alumni)) + geom\_density(fill = "blue") + hw + labs(y="Density",x="Percent of Alumni Who Donate", title="Density of Percent Alumni Who Donate")*

*ggplot(colleges, aes(x = perc.alumni)) + geom\_density(aes(fill = factor(Private))) + hw+ labs(y="Density",x="Percent of Alumni Who Donate", title="Density of Percent Alumni Who Donate, By School Type")*

**Section 5: Pairing Quantiles of Data**

As seen from the last section, the standard deviations between the percent of alumni that donated to private schools versus the percent of alumni that donated to public schools was very different. This implies that the quantiles between the two must not be the same. Judging based on the second diagram from the last section, neither distribution seemed to have a perfect normal curve. Although not plotted, it didn’t seem that they had a perfect normal curve with respect to each other as well. A simple QQ plot would help to see the relationship the data had with each other. It is shown below.

In the QQ plot, paired quantiles of percent of alumni who donate to private schools and percent of alumni who donate to public schools are plotted matching up with the x and y axis for each. A QQ Plot with a matched normal distribution should line up evenly with the black line who’s equation is given by y=x. However, the line of best fit for the QQ plot has a slope of less than one. It can also be seen by the two points furthest to the right of the graph that the plot has a thick tail. The slope tells us the skewedness of the data. We can see that the data may be more right skewed from this (meaning there is more area to the right of the center of the original density plot than to the left). The thick tail means that towards the end, if we possibly had more data points to look at, the QQ plot might approach normal values.

Below this is the mean-difference plot for the same data. 

This mean difference plot shows a distinct pattern in the amount who donate. The difference gets linearly higher between public and private schools for the amount who donate as can be seen from the robust line fit. On average, about eleven percent alumni more donate to private schools then to public schools.

*R Code: Section 5*

*#Section 5: pairing quantiles of the data*

*x <- colleges$perc.alumni[colleges$Private=="Yes"]*

*y <- colleges$perc.alumni[colleges$Private=="No"]*

*qqplot(x,y,las=1, xlab="Private",ylab="Public",*

*main="Q-Q Plot for Percent Alumni")*

*n <- min(length(x),length(y))*

*probs <- seq(0,1,length=n)*

*qx <- quantile(x,probs=probs)*

*qy <- quantile(y,probs=probs)*

*df <- data.frame(qx,qy)*

*ab <- coef(rlm(qy~qx))*

*rx <- range(colleges$perc.alumni)*

*ggplot(df,aes(x=qx,y=qy))+*

*geom\_abline(intercept=0,slope=1,col="black")+*

*geom\_abline(int=ab[1],sl=ab[2],*

*size=1.5,col=rgb(0,.8,0))+*

*geom\_point(size=3.2,col="blue")+*

*xlim(rx)+ylim(rx)+*

*labs(x="Private Schools",y="Public Schools",*

*title="Q-Q Plot of Percent Alumni Who Donate")+*

*hw*

*qMean <- (qx + qy)/2*

*qDiff <- qx - qy*

*df <- data.frame(qMean,qDiff)*

*rFit <- rlm(qDiff~qMean)$coef*

*ggplot(df,aes(x=qMean,y=qDiff))+*

*geom\_hline(yint=mean(qDiff),size=1.2)+*

*geom\_abline( int=rFit[1],sl=rFit[2],*

*size=1,col=rgb(0,.8,0))+*

*geom\_point(col="blue",size=3.2)+*

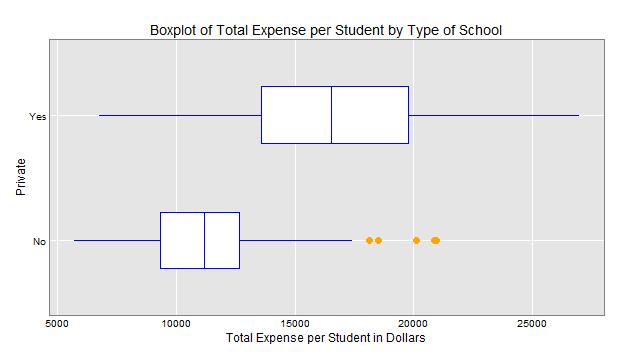
*labs(x="Mean of Percent Alumni Who Donate",*

*y="Public - Private",*

*title="MD Plot of Paired Private and Public Quantiles")+*

*hw*

**Section 6: Boxplots for Total Expenses**

In this section, I wanted to shift gears and look at some other variable from my dataset. I decided to question whether or not a private school was really more expensive than a public school. I knew students usually paid less to attend public schools because most public schools attendees try to remain in state. However, the data from my dataset allowed me to view all schools out of state tuitions. Public school tuitions for out of state students can reach very high sums of money. I decided to do total expenses as the variable I looked at, so I manually added a new column in excel to the data which summed out of state tuitions, book fees, and room and board fees. I read the new table back in with the added column called Tot.Exp. One thing to take note of was that the information from the dataset is about twenty years old, so the conclusion from this may not still remain true today. My hypothesis was that they fees would be approximately the same. The boxplot for the information from the dataset is pictured below.

In the above plot, the top boxplot represents private schools while the bottom boxplot represents public schools. You can see that the median public school expenses were around $11,500 while the median private school expenses were at about $16,000. Obviously, the private schools expenses were far more than the public schools. It can also be seen from the plots that the private schools had a far greater range in their expenses than public schools, stretching from $6,500 to $27,500. Although the range was greater for private schools, only public schools had any outliers. The outliers for the public schools expenses are shown in the diagram as orange dots. These dots are show schools that are much higher than the expected range of a normal public schools expenses. Again, the conclusions from this data about tuition is probably not totally accurate because of the age of the data.

*R Code: Section 6*

*#Section 6: Boxplots*

*ggplot(data=colleges, aes(x=Private,y=Tot.Exp)) +*

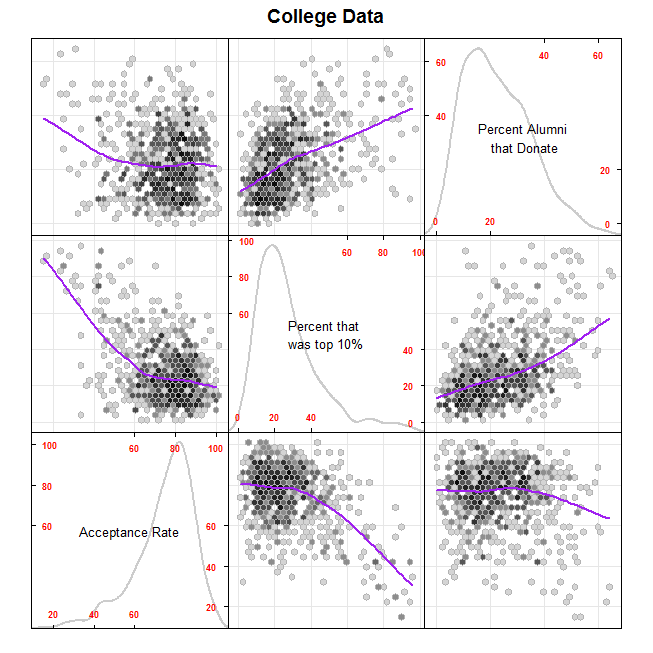
*geom\_boxplot(width=.6,col="blue",*

*outlier.colour='orange',outlier.size=3.5, lwd=.1)+*

*coord\_flip() + hw + labs(x="Private",y="Total Expense per Student in Dollars",*

*title="Boxplot of Total Expense per Student by Type of School")*

**Section 7: Scatterplot Matrix**

For this section and the next few, I decided to try and compare multiple variables from the dataset at a time. After looking at some of the factors that caused an alumni to donate back to a school, I was curious about what kind of effects the competitiveness of a school had on the amount that an alumni would donate. I felt that people who graduated from a more competitive school might become more successful and thus have more money they could give back to their alma mater. If not that, then possibly those who went to a competitive school may have had a higher sense of school spirit which could have pushed them to give back to their school. Whatever the cause may be, the graph of what I found is displayed below. 

As can be seen from above, there were some very strong relations seen between the three variables I chose for this scatterplot matrix. It seems that a lower acceptance rate corresponded with a higher percentage of alumni that donated and a higher percent of students that were top ten in high school fairly well. The acceptance rate and percent that was top ten in high school proved fairly accurate predictors of the percent of alumni that donated. This could explain why schools who have more money to spend tend to have lower acceptance rates.

For this plot, I used hexagon binning. Hexagon binning is useful in showing where the denser parts of each scatter plot is, especially in a case like this with almost 800 points in each box. The scatterplot matrix also shoes the density for each variable in the main box. Finally, it gives a line sketched in each box which shows the relation between the data.

*R Code: Section 7*

*c <- data.frame(colleges$perc.accept, colleges$Top10perc, colleges$perc.alumni)*

*myPanel <- function(x,y,...){*

*panel.grid(h=-1,v=-1,...)*

*panel.points(x,y,...,pch=21,fill=rgb(0,.9,0),col="black")*

*panel.loess(x , y, ..., lwd=3,col='purple')*

*}*

*windows(width=8,height=8)*

*splom(c,*

*varnames=c("Acceptance Rate",*

*"Percent that \nwas top 10%",*

*"Percent Alumni\n that Donate"),*

*xlab='',main="College Data",*

*pscale=4, varname.cex=0.8,axis.text.cex=0.6,*

*axis.text.col="red",axis.text.font=2,*

*axis.line.tck=.5,*

*panel=function(x,y,...){*

*panel.grid(h=-1,v=-1,...)*

*panel.hexbinplot(x,y,...,border=gray(.7),trans=function(x)x^.2)*

*panel.loess(x , y, ..., lwd=2,col='purple')*

*},*

*diag.panel = function(x, ...){*

*yrng <- current.panel.limits()$ylim*

*d <- density(x, na.rm=TRUE)*

*d$y <- with(d, yrng[1] + 0.95 \* diff(yrng) \* y / max(y) )*

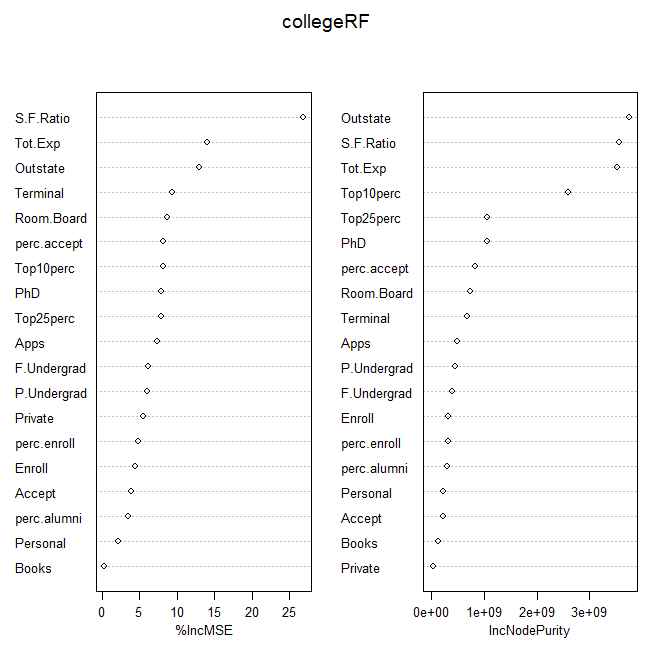
*panel.lines(d,col=gray(.8),lwd=2)*

*diag.panel.splom(x, ...)*

*}*

*)*

**Section 8: Importance of Variables**

In this section, I again want to look at how each variable affects the others. To do this, I used the importance function. The importance function takes in a random forest argument, so I had to create a random forest in order to make this chart. The chart is shown below.

The importance of each variable for %IncMSE and IncNodePurity can be seen above. The information presented in the graph is fairly self-explanatory.

*R Code: Section 8*

*#Section 8: importance (see Week 10)*

*set.seed(4543)*

*collegeRF <- randomForest(x = colleges[ , 2:20], y=colleges[, 21],*

*importance=TRUE, proximity=FALSE, ntree=500,*

*keepForest=TRUE)*

*collegeRF*

*dep <- colleges[, 21]*

*sc <- diff(range(dep))*

*sc*

*scaledRes <- (dep - predict(collegeRF))/sc*

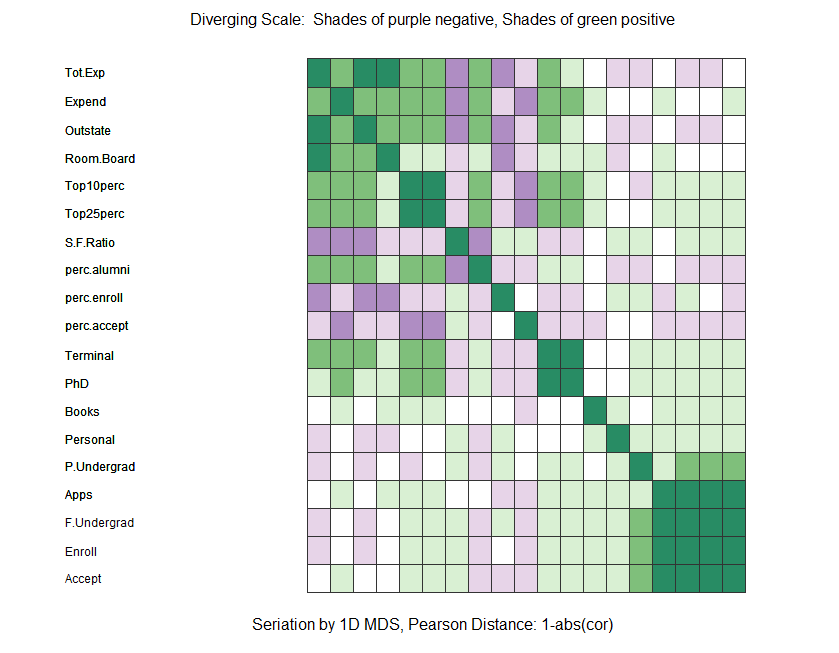
*mean(abs(scaledRes)) # 0.026*

*imp <- importance(collegeRF)*

*imp*

*varImpPlot(collegeRF,cex=.8)*

**Section 9: Color Correlation Matrix for the Data**

For the final section, I wanted to be able to look at all of the variables and see how they correlated with each other. Many of the variables seemed to be easily relatable. For example, I predicted that the percent accepted would have a strong correlation with expenditures the school has, because the students that are accepted to more competitive schools would likely want to attend a school with enough money to provide them opportunities to get the most of their education. Below is the color correlation matrix for the dataset.

The matrix is displayed using a diverging scale, with darker purple indicating a strong negative correlation between variables and darker green indicating a strong positive correlation between variables. Obviously, each variable has a perfect correlation with itself, which explains the diagonal line of dark green. Some of the trends in the correlation values can be easily explained. For example, the square of dark green in the bottom right. Schools with high volume will easily get many applications, full time undergraduates, enrolled students, and accepted students whereas the opposite is true for schools with low volume.

*R Code: Section 9*

*#Section 9: color correlation (see Week 10)*

*subs <- c(3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21)*

*collegedat <- colleges[,subs]*

*corMat <- cor(collegedat) # get correlations*

*x <- cmdscale(1-abs(corMat), k=1) # seriation*

*ord <- rev(order(x))*

*corMatOrd <- corMat[ord, ord]*

*round(corMatOrd, 2)*

*# define colors*

*mat <- matrix(c(*

*120, 60, 180,*

*175, 141, 195,*

*231, 212, 232,*

*255, 255, 255,*

*217, 240, 211,*

*127, 191, 123,*

*40, 140, 100), ncol=3, byrow=TRUE)*

*colors=rgb(mat[, 1], mat[, 2], mat[, 3], max=255)*

*# assign color classes*

*tmp <- as.vector(corMatOrd)*

*brk <- c(-1.01, -.70 , -.40, -.10, .10, .40, .70, 1.01)*

*colorSub <- cut(tmp, brk, label=FALSE)*

*# plot*

*windows(w=9, h=7)*

*par(mai=c(.4, .5, .3, .2))*

*nr <- nrow(corMatOrd)*

*x <- 1:nr-.5*

*cen <- expand.grid(list(y=rev(x), x=x))*

*plot(c(-10, nr+2), c(0, nr+.3), type='n', axes="FALSE", ylab='')*

*mtext(side=3,line=0,*

*'Diverging Scale: Shades of purple negative, Shades of green positive')*

*mtext(side=1, line=0,*

*'Seriation by 1D MDS, Pearson Distance: 1-abs(cor)')*

*# could use image()*

*rect(cen$x-.5, cen$y-.5, cen$x+.5, cen$y+.5, col=colors[colorSub],*

*border=rgb(.2, .2, .2))*

*text(rep(-10.5, 35), y=rev(x), rownames(corMatOrd), adj=0, cex=.75)*

**Section 10: Conclusion**

Through this paper, I was able to explore the relationship of the dataset using some of the methods I learned throughout this semester in class. Had I had more time and space, I may have tried some new things on the data, such as mapping variables in 3D with a convex hull. If I could change something about the way I did the project for the future, I would find a better dataset. The one I have here is from 1995. There were better datasets available, but with a price tag involved. I enjoyed learning what I could in this class, and it was nice to be able to put it all together into one culminating paper.