UNIVERSITY OF MICHIGAN
DATA MINING
STATS415

ASSIGNMENT 1

Author: Shu ZHOU
ID: 19342932
Lab Section: 002

September 7, 2020

1 Q1.

1.1

- Categorical Variable: Study Status (Remotely or In-Person)
- Ordinal Variable: Grade Year (Freshman, Sophomore, Junior or Senior)
- Continuous Variable: Age of Student

1.2

The grade year of student would be related to the age of student. Since we can infer that the larger the age of student, the higher the grade year of student.

1.3

We can not make inference on the study status with the age of student.

2 Q2.

$$f_{ij}^{*} = f_{ij} \ln \frac{n}{gj} \tag{1}$$

2.1

The purpose of this transformation is to determine the importance of a certain word in a set of documents or a text corpus. The importance of the word increases with the increment of the word frequency in *i-th* document and the decrement of the document frequency.

e.g. Considering the word "the" and "engine" in a set of documents related to cars. According to a document describing how the engine works, we can determine that f_{ij} of both "the" and "engine" is very high. However, since we can find the word "the" in almost every document in this set, (*i.e.* g_j of "the" is low, while g_j of "engine" is high), f_{ij}^{*} of "engine" would be much higher than "the", which means the term "engine" is more important.

2.2

- Supervised Learning: Given a set of documents related to cars, we want to study how the term "engine" is correlated to the term "Gas". We can calculate the f_{ij}^{*} of this two terms in each document and perform a linear regression.
- Unsupervised Learning: Given the same set of documents related to cars, we want to look for the three possibly key words of each document. Then, we can evaluate the f_{ij}^{*} of each word in each document.

3 Q3.

I performed some analysis with R

3.1

```
DATADIR <- "./extdata"
college_file_path <- file.path(DATADIR,"College.csv")
college_dt <-as.data.frame(read.csv(college_file_path))
head(college_dt)
summary (college_dt)
```

The result of the summary is in Figure. 1.

```
       X              Private              Apps           Accept          Enroll       Top10perc
 Length:777         Length:777         Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
 Class :character   Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
 Mode  :character   Mode  :character   Median : 1558   Median : 1110   Median : 434   Median :23.00
                                       Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
                                       3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
                                       Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
   Top25perc      F.Undergrad     P.Undergrad         Outstate       Room.Board       Books
 Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0
 1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0
 Median : 54.0   Median : 1707   Median :  353.0   Median : 9990   Median :4200   Median : 500.0
 Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4
 3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0
 Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0
    Personal         PhD           Terminal       S.F.Ratio      perc.alumni        Expend
 Min.   : 250   Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
 1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
 Median :1200   Median : 75.00   Median : 82.0   Median :13.60   Median :21.00   Median : 8377
 Mean   :1341   Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
 3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
 Max.   :6800   Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
   Grad.Rate
 Min.   : 10.00
 1st Qu.: 53.00
 Median : 65.00
 Mean   : 65.46
 3rd Qu.: 78.00
 Max.   :118.00
```

3.2

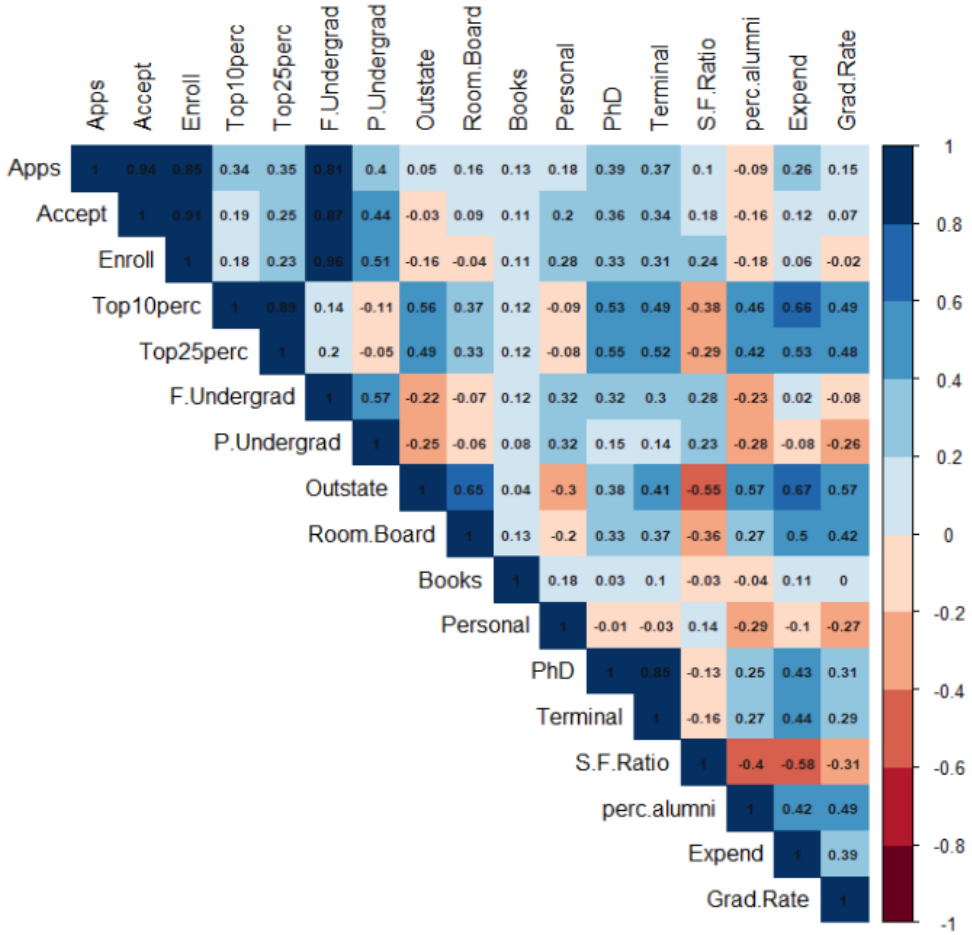
Then I try to study the correlation between each variable. Also, I changed the type of Private into a factor(categorical variable)

```
#Change the Private Type into Factor
college_dt$Private <- as.factor(college_dt$Private)
summary (college_dt)


#Study the correlation
cor(college_dt[,c("Apps","Accept","Enroll","Top10perc","Top25perc","F.Undergrad","P.
    Undergrad","Outstate","Room.Board","Books","Personal","PhD","Terminal","S.F.Ratio","perc
    .alumni","Expend","Grad.Rate")])

corrplot(cor(college_dt[,c("Apps","Accept","Enroll","Top10perc","Top25perc","F.Undergrad","P
    .Undergrad","Outstate","Room.Board","Books","Personal","PhD","Terminal","S.F.Ratio","
    perc.alumni","Expend","Grad.Rate")]), method="color", type = "upper", col=brewer.pal(n
    =10, name="RdBu"),
         tl.col="black",tl.srt=90, addCoef.col = "gray8", diag = T, number.cex = 0.65)
```

The result is shown in the heat map.



3.3

Then I try to conduct some graphical summaries, from the numerical summary, I find that the variable "Top25perc" is interesting. Since numerically this variable is likely to follow a normal distribution.

```
top25perc<-college_dt[,7]
hist(top25perc)
boxplot(top25perc)
```

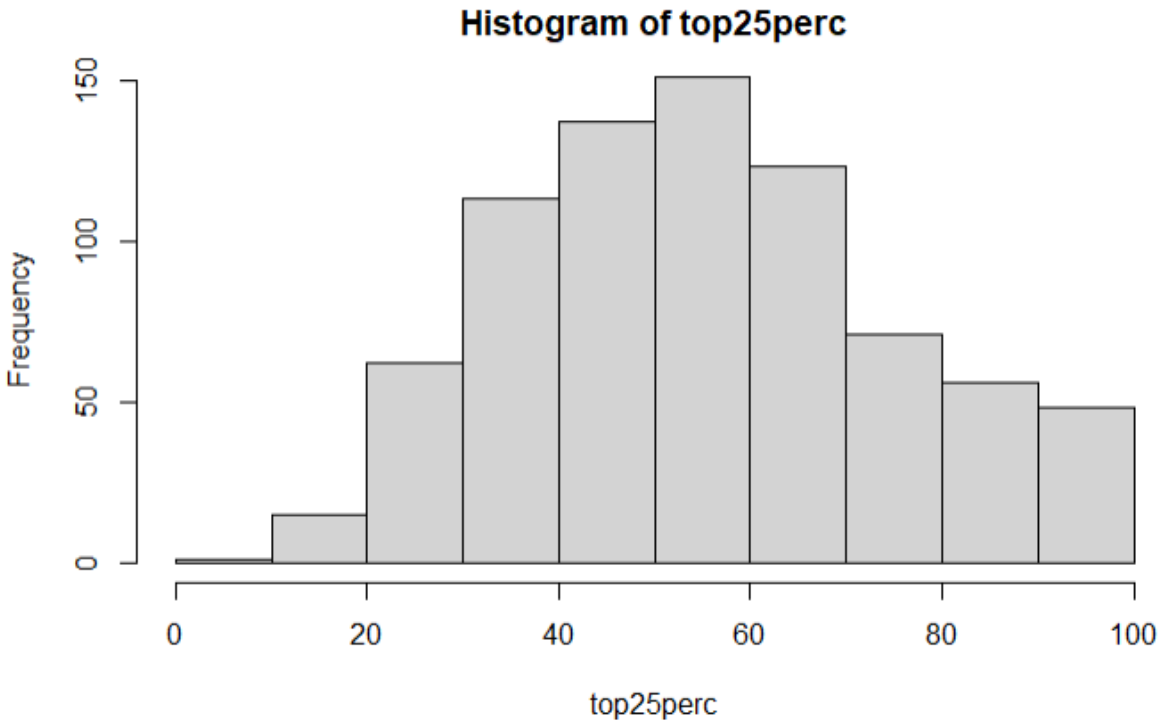


Figure 1: Histogram of Top25perc

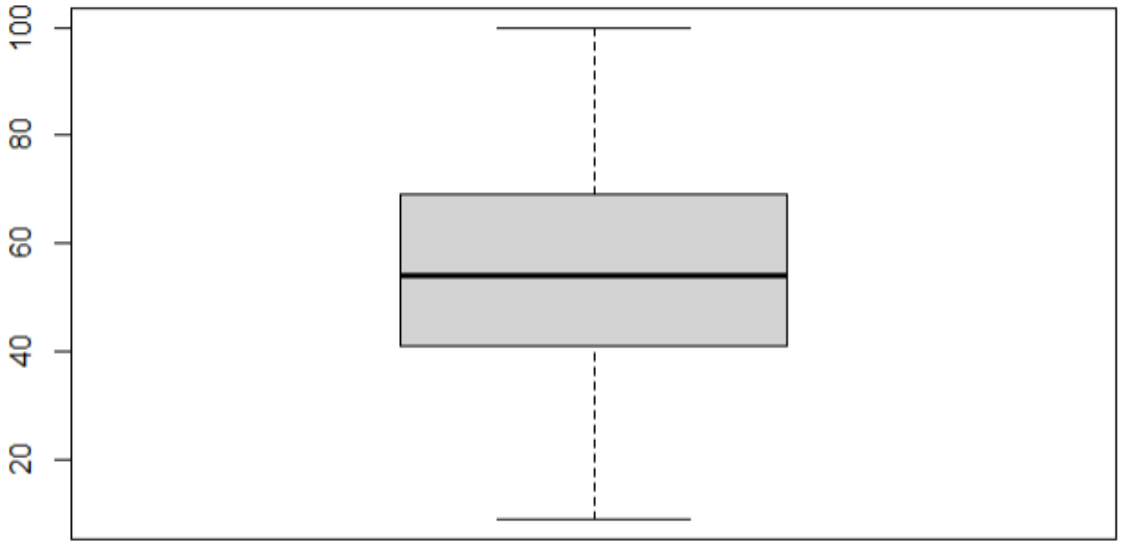


Figure 2: Boxplot of Top25perc

And I test whether the variable follows a normal distribution.

```
ks.test(top25perc, "pnorm", mean = mean(top25perc), sd =  sqrt(var(top25perc)))
```

```
Kolmogorov - Smirnov□□□□□□Ö□□□□□□□□□
        One-sample Kolmogorov-Smirnov test

data:  top25perc
D = 0.046288, p-value = 0.07162
alternative hypothesis: two-sided
```

Figure 3: Result of the K-S test

Since p-value is greater than 0.05, we cannot reject that the variable follows a normal distribution.

3.4

From the heatmap, we can observe some interesting high correlations which cannot be comprehended directly. (*e.g.* The correlation between Top10Perc and Expend)

```
##correlation plot of top10perc and Expenditure
perc_and_Expend<-college_dt[,c(6,18)]
ggplot(college_dt, aes(Top10perc, Expend))+ geom_point(colour = "black", shape = 21, size =
    3, aes(fill = factor(Private)))+
  scale_fill_brewer(palette = "OrRd")+
  geom_smooth(method = "loess", col = "dodgerblue4", fill = "lightsteelblue3", size = 1.2)+
  annotate("text", x = 0.8, y = 0.52, label = "italic(r) == 0.42", parse = T, size = 6, col
    = "gray20")+
  labs(x = "top10perc", y = "Expenditure")+
  theme_economist()+
  theme(axis.text.x = element_text(size=10), axis.text.y = element_text(size=10), legend.
    position = "right")+
  guides(fill = guide_legend(title = "Private or not"))

```

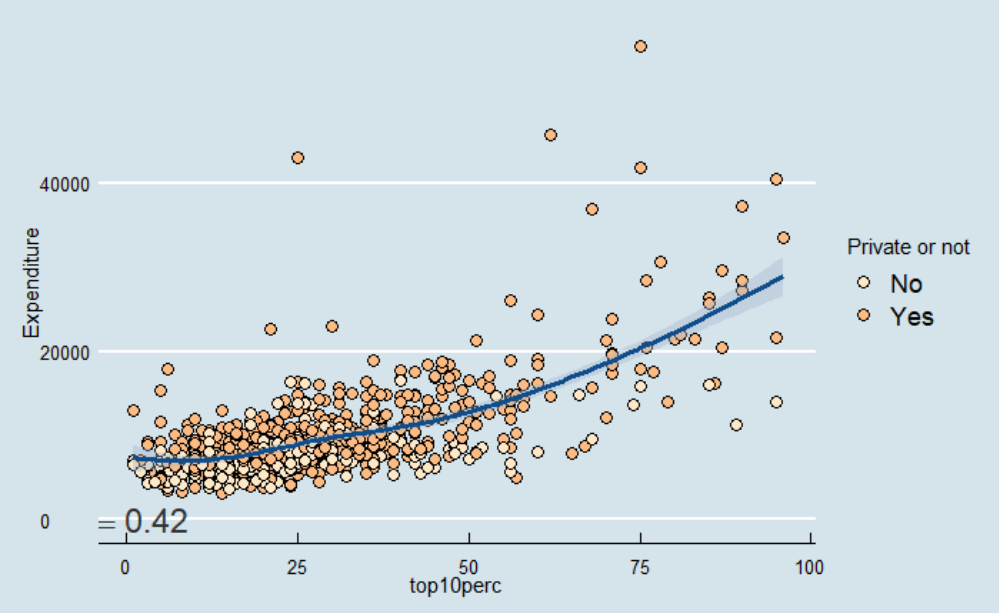


Figure 4: Result of the correlation investigation ('geom_smooth()' using formula 'y - x')

From this graph, we can roughly determine that the higher the top10perc rate the higher the Instructional expenditure per student

For some obvious relationship (*e.g.* The relationship between number of applicants accepted and number of new students enrolled), we use a side-by-side boxplot to investigate.

First, we convert the number of acceptance and enrollment into rate by dividing the corresponding number with the number of application.

```
college_dt$accept_rate<-college_dt[,4]/college_dt[,3]
college_dt$enroll_rate<-college_dt[,5]/college_dt[,3]
boxplot(college_dt[,c(20,21)])
```

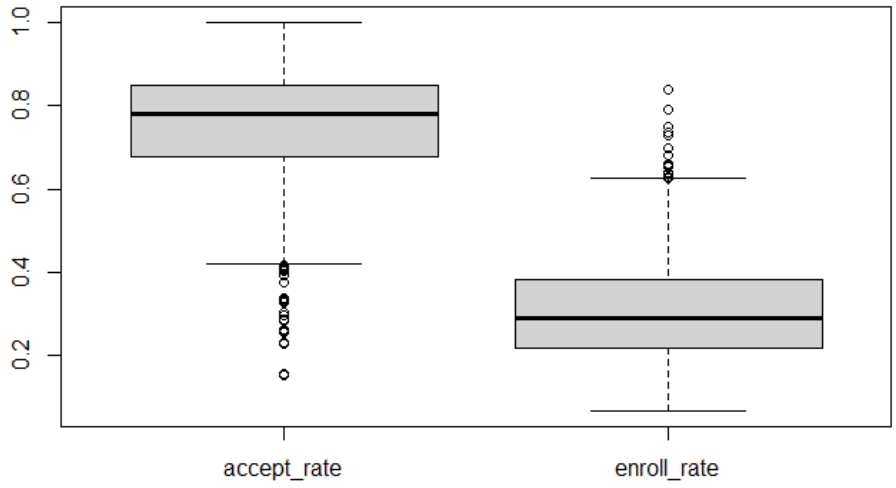


Figure 5: Result of the side-by-side boxplot

We can see that the accept rate has many lower outliers and enrolled rate has many upper outliers.