Eindopdracht Deel 1

Mohammed Al Hor

2022-09-25

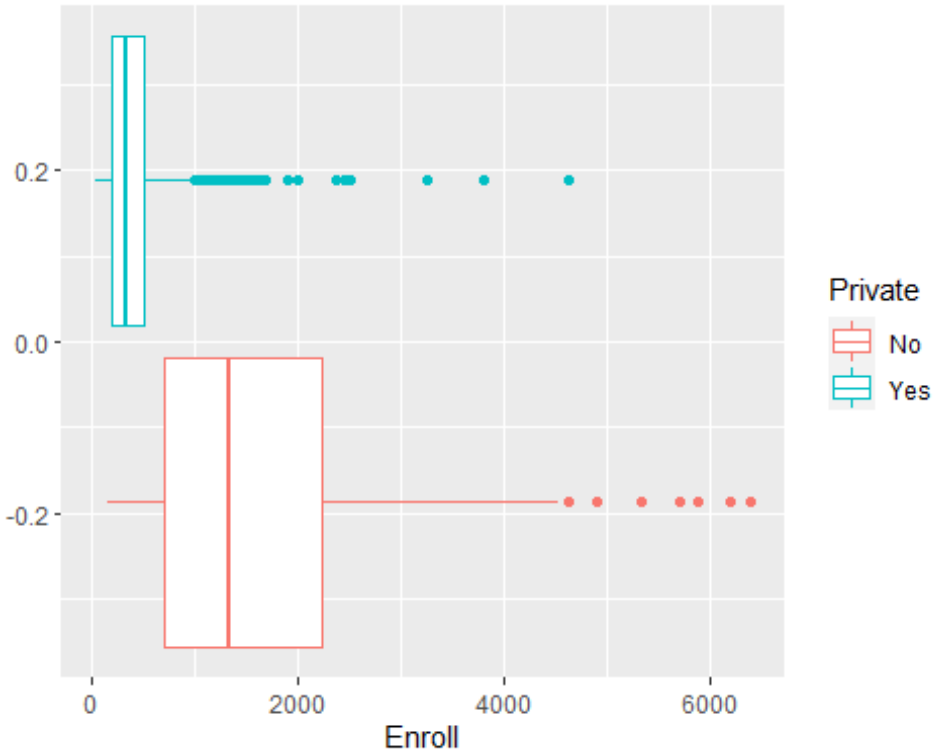
1. **Lees de data in als dataframe en zorg ervoor dat de eerste kolom als label voor de rijen wordt gebruikt. Hint: zie ?row.names.**

```
setwd("~/Data-Science-Business-Analytics/Data")
college_statistics <- read.csv("college_statistics.csv", header = TRUE)
rownames(college_statistics) <- college_statistics[,1]
college_statistics <- college_statistics[,-1]
```

The data is loaded in as a csv file with header, labels are set via rownames as the first column. Finally, the first column 'X' is dropped.

2. **Voer beschrijvende statistiek uit dmv. het maken van een aantal grafieken. Deze grafieken mag je via de "Base R graphics" of mbv. ggplot maken. Creeer geschikte grafieken om de volgende vragen te beantwoorden:**

(a) Zijn private universiteiten overwegend kleiner of groter dan publieke universiteiten? Je mag zelf een definitie voor groot/klein definieren.



The median or mid-point of the data for private schools is lower, so is the inter-quartile range (the middle box that represents 50% of the observations). These boxplots would

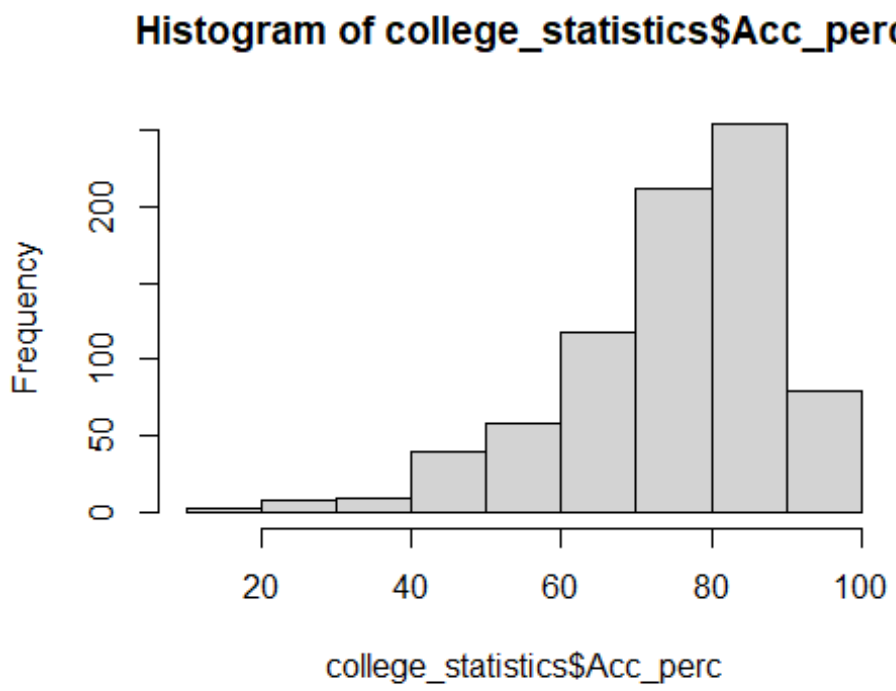
suggest that private schools have fewer enrolled students. However, to say this with reasonable certainty a statistical test on the mean must be done (private school observations obviously contain a significant amount of outliers).

(b) Hoe ziet de verdeling eruit van het acceptatiepercentage? Wat is het acceptatiepercentage voor de meest selectieve universiteit?

First off, we check for missing values in 'Accept' and 'Apps'. No missing values. To calculate the acceptance rate we must divide the number of accepted students by the number of applicants. To get a percentage we multiply by 100.

```
college_statistics$Acc_perc <- (college_statistics$Accept/college_statistics$
Apps)*100
```

To get a feeling for the distribution we generate a histogram for this new variable.



The distribution is negatively skewed (tail to the left).

To get the lowest acceptance rate for a school we run the following statement:

```
college_statistics[which.min(college_statistics$Acc_perc),]
```

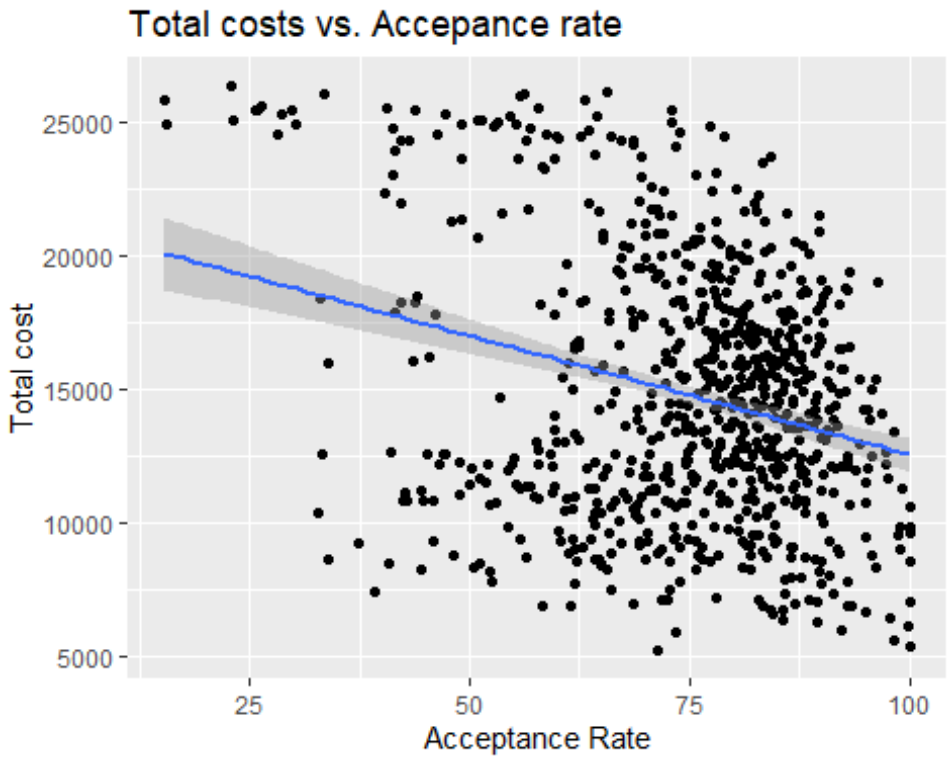
Princeton University is the school with the lowest acceptance rate.

(c) Zijn de meer selectieve universiteiten ook overwegend duurder dan minder selec- tieve universiteiten? (Je mag zelf bepalen welke kosten je wel/niet mee neemt.)

To get this answer we must do some feature engineering. Which costs should we take into account? To do this, we take a look at histograms, summary stats and some quick plots of the different cost variables (Outstate, Room.Board, Personal and Books). We pick 'Outstate' and 'Room.Board' to use in our 'cost' variable. Mainly because we do not see any relationship bewteen Personal spending, the cost of books and the acceptance rate. This is explained in the inline comments in the added R script. Thus, to calculate the total cost we run the following statement:

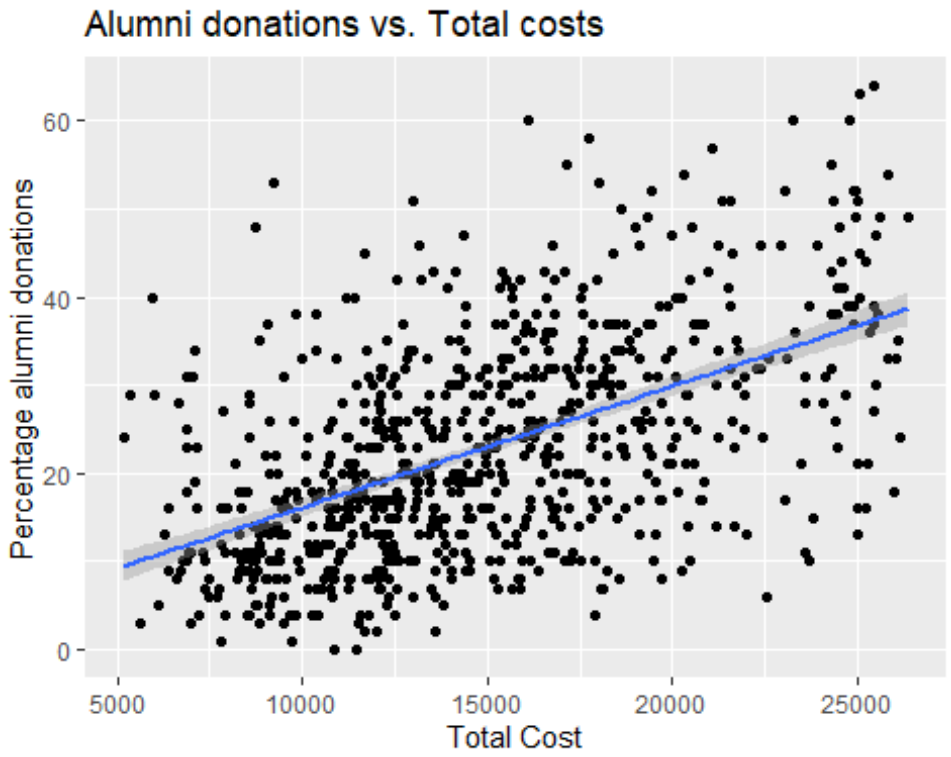
```
college_statistics$total_cost <- college_statistics$Outstate + college_statis
tics$Room.Board
```

To answer the question we plot the total costs and acceptance rate using the ggplot package.



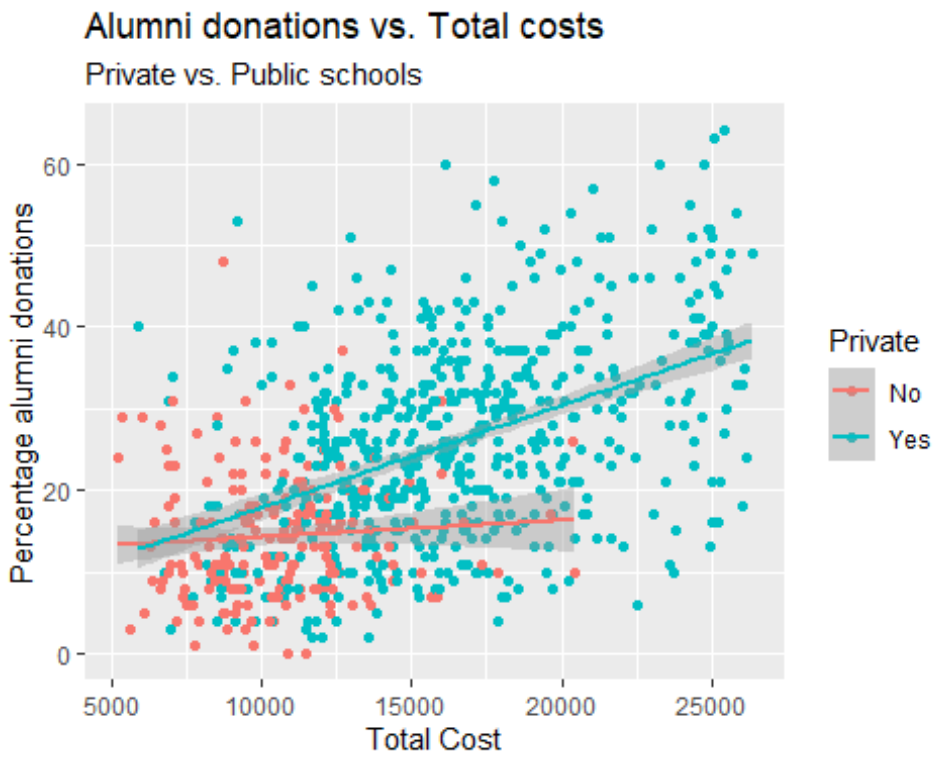
Looking at this plot we see a negative relationship between the cost and the acceptance rate. The more selective schools seem to be more expensive than the less selective schools.

(d) Bedenk zelf een extra vraag en creeer een geschikte figuur om deze vraag mee te beantwoorden. Hebben duurdere universeiten een hoger percentage aan afgestudeerden die geld doneren? Bekijk ook het verschil tussen private en publieke universiteiten.



This plot shows a positive relationship between costs and the percentage of alumnis who donate to their alma mater. This makes sense, the more affluent students who go to these schools probably have more disposable income to donate.

Let's take a look at this relationship with taking into account whether it's a private or public school.



This plot shows that alums who when to private schools, which were more expensive, tend to donate more (positive relationship between costs and perc of alums who donate). We don't see the same 'strong' positive relationship for students who went to public schools. In short, students who went to private schools tend to donate more often, the more expensive a school is.

3. **Voer hypothesetoetsen uit om de volgende vragen te beantwoorden. Geef telkens duidelijk aan wat de exacte nul- en alternatieve hypothese is die je toetst, motiveer de keuze van de specifieke toets en verwoord duidelijk de conclusie.**

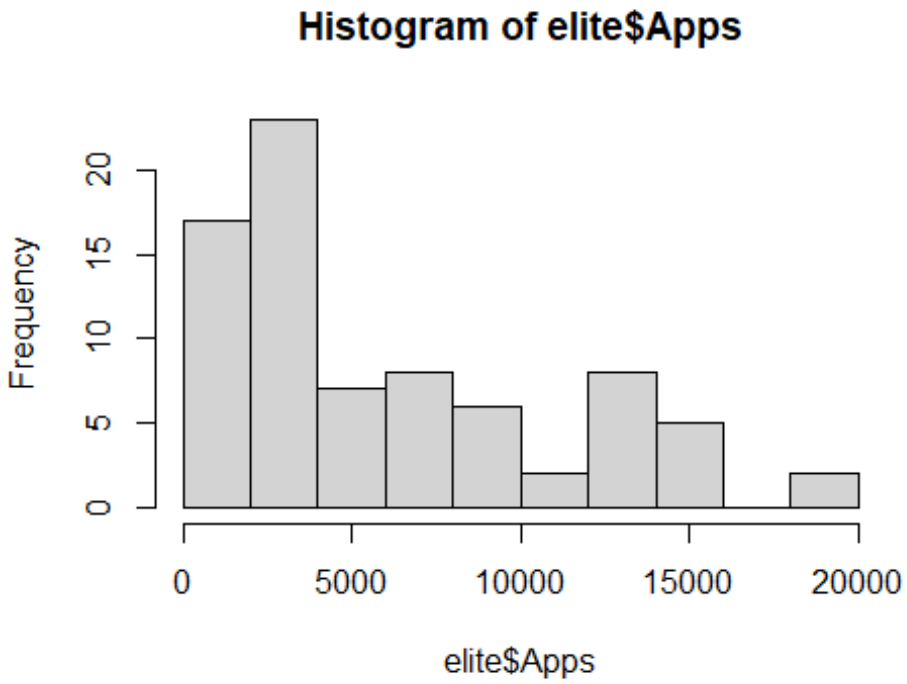
(a) Ontvangen elite scholen een ander aantal aanmeldingen in vergelijking met niet- elite scholen? Definieer "elite-school" als scholen waarvoor geldt dat meer dan 50% van de studenten tot de top 10% van hun high school behoort.

To answer this question, we first get a subset of the data, schools with > 50% in the top10 and schools <= 50% in the top10. We name them elite and not_elite, respectively.

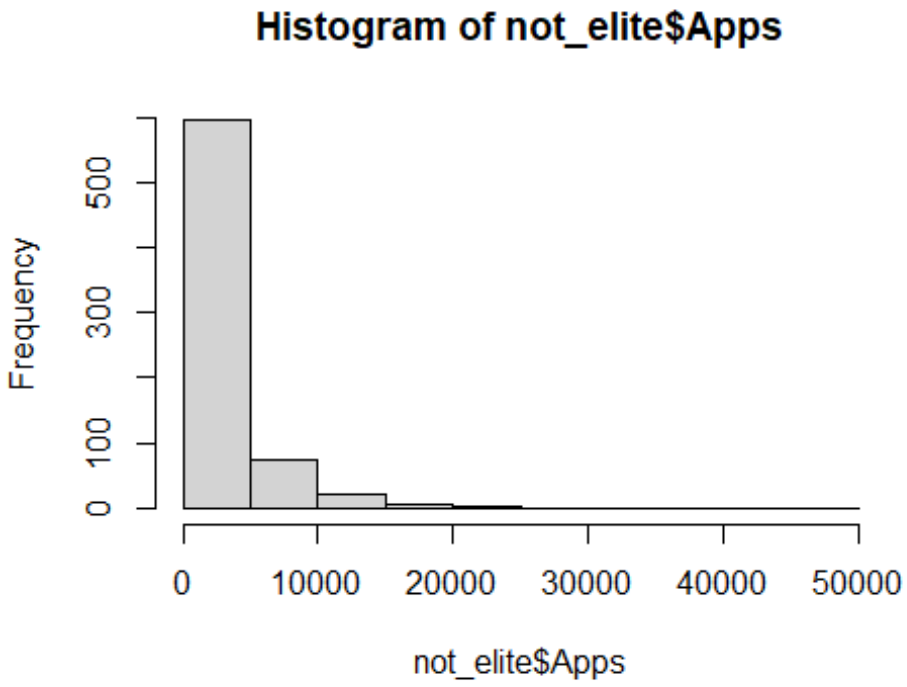
```
elite <- subset(college_statistics, Top10perc > 50)
not_elite <- subset(college_statistics, Top10perc <= 50)
```

To compare the means of two groups a widely used method is the t-test. The t-test does make some assumptions about the data (independent, normality, and similar amount of variance). Let's check for this. Because the data concerns separate schools we can assume independency. For normality we first look at the histograms.

```
hist(elite$Apps)
```




```
hist(not_elite$Apps)
```



It is obvious from the histograms that the data is not normally distributed (I don't want to overload you with every plot I make in R, so I chose not to show these). To make sure we can do a Shapiro-Wilk test for normality.

```
shapiro.test(elite$Apps)

##
##  Shapiro-Wilk normality test
##
## data:  elite$Apps
## W = 0.87149, p-value = 1.19e-06

shapiro.test(not_elite$Apps)

##
##  Shapiro-Wilk normality test
##
## data:  not_elite$Apps
## W = 0.61192, p-value < 2.2e-16
```

H0: elite$Apps/not_elite$Apps is normally distributed

H0: elite$Apps/not_elite$Apps is not normally distributed

The results from both Shaprio-Wilk tests confirm our suscpicions. The null hypothesis can be rejected (p value is extremely small), thus the data for elite and non elite schools in not normally distributed. Now let's take a look at the variance.

```
var(elite$Apps)

## [1] 25257257

var(not_elite$Apps)

## [1] 12763705

sd(elite$Apps)

## [1] 5025.66

sd(not_elite$Apps)

## [1] 3572.633
```

At first glance the variances and standard deviations of elite vs. non elite schools' number of applications are quite different. Elite schools showing almost twice as much variance. Let's put this to the test, using a Fischer F test for variance.

```
var.test(elite$Apps,not_elite$Apps, alternative=c("two.sided","less","greater
"))

##
##  F test to compare two variances
```

```
##
## data:  elite$Apps and not_elite$Apps
## F = 1.9788, num df = 77, denom df = 698, p-value = 8.922e-06
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.448177 2.829268
## sample estimates:
## ratio of variances
##           1.978834
```

$H_0 : \sigma_{elite} = \sigma_{not\_elite}$

$H_0 : \sigma_{elite} \neq \sigma_{not\_elite}$

Our assumption about the variance seems to be true. We can reject the null hypothesis (very small p-value), and say the variance in number of applications for elite schools is statistically different from that of non elite schools.

Finally, now that we know the data is not normally distributed and the variances are not similar we can pick the correct test. Note that in the beginning we mentioned the t-test to compare the means of two groups. Because of the normality and similar variance conditions of this test are not met, this might not be the the best method. Considering we have a decent sample size, we could get away with using the two sample t-test. However, I feel more comfortable using a non parametric test. The Wilcoxon Rank-Sum test is such a method.

```
wilcox.test(elite$Apps, not_elite$Apps, alternative=c("two.sided","less","greater"))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  elite$Apps and not_elite$Apps
## W = 40896, p-value = 4.108e-13
## alternative hypothesis: true location shift is not equal to 0
```

H0: the 2 groups are equal in terms of number of applications Ha: the 2 groups are different in terms of number of applications

The p-value is extremely small. Therefore, we can reject the null hypothesis at 1% level and conclude that the number of applications are significantly different for elite and non elite schools.

(b) Is er een verband tussen acceptance rate en graduation rate?

First, let's take a quick look at both variables.

```
summary(college_statistics$Acc_perc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.45   67.56   77.88   74.69   84.85  100.00
```

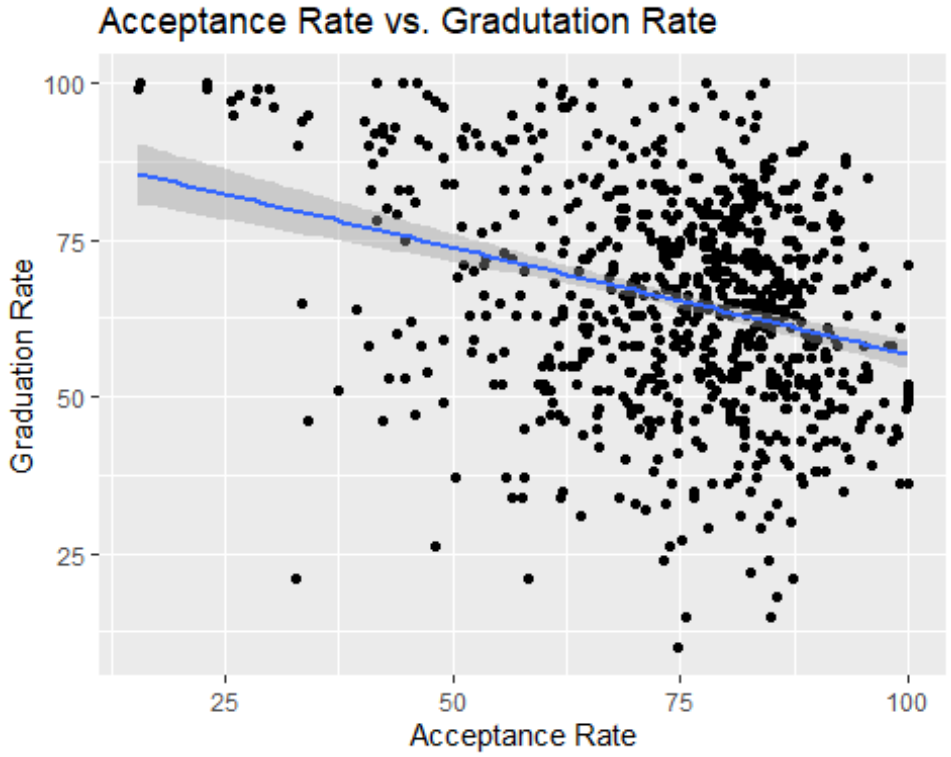
```
summary(college_statistics$Grad.Rate)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   53.00   65.00   65.46   78.00  118.00
```

We immediately see that the variable 'Grad.Rate' has a max value of 118.00. This cannot be right, the graduation rate cannot be more than 100. We drop the observations with graduation rate larger than 100.

```
college_statistics_b <- college_statistics[college_statistics$Grad.Rate <= 100,]
```

Now we can delve into the data further by making a quick plot using ggplot.



The plots shows a negative relation between the acceptance rate and graduation rate. In other words, the more selective schools have higher graduation rates.

Now, let's take a look at the distribution of the data. Do they follow a normal distribution? To do this, we use the Shapiro-Wilk test for normality.

```
shapiro.test(college_statistics_b$Acc_perc)

##
##  Shapiro-Wilk normality test
##
## data:  college_statistics_b$Acc_perc
## W = 0.93383, p-value < 2.2e-16
```

```
shapiro.test(college_statistics_b$Grad.Rate)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  college_statistics_b$Grad.Rate
## W = 0.99259, p-value = 0.0006568
```

H0: Acc_perc/Grad.Rate is normally distributed

Ha: Acc_perc/Grad.Rate is not normally distributed

The p-values are very small. We can reject the null hypothesis at the 1% level and conclude that the data is not normally distributed. This affects the method we choose to test for correlation (we cannot use Pearson correlation, because of its assumption of normally distributed data). Instead, we use a non parametric test for correlation, the Kendall Rank correlation test or Kendall's tau.

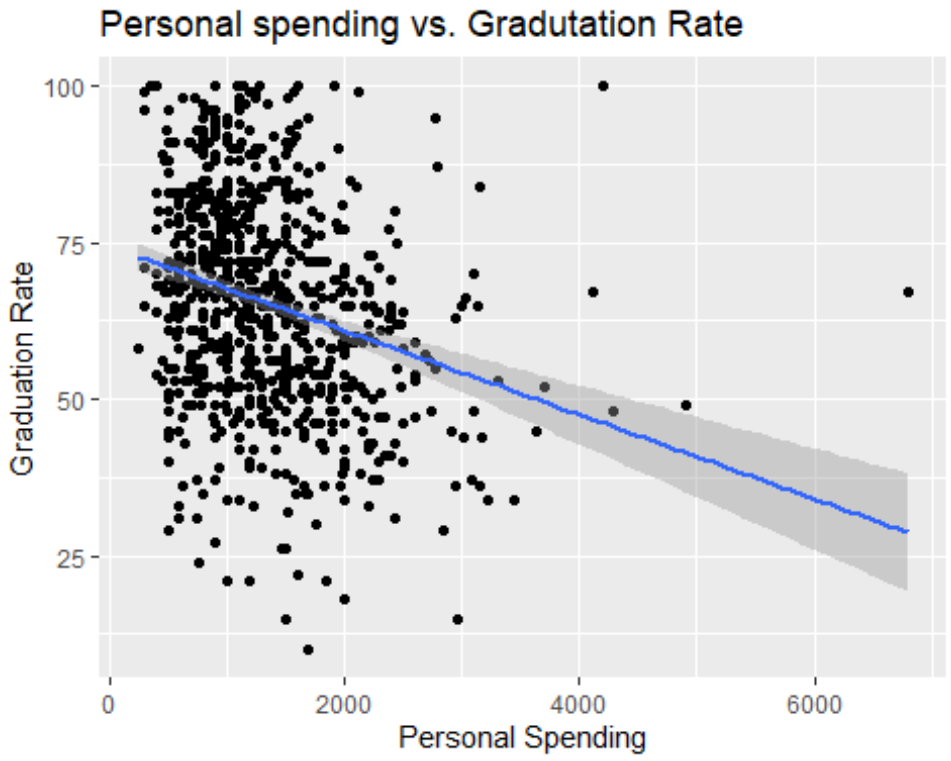
```
cor.test(college_statistics_b$Acc_perc, college_statistics_b$Grad.Rate, metho
d = 'kendall')
```

```
##
##  Kendall's rank correlation tau
##
## data:  college_statistics_b$Acc_perc and college_statistics_b$Grad.Rate
## z = -6.6432, p-value = 3.07e-11
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## -0.1606286
```

H0: correlation is equal to 0 Ha: correlation is not equal to 0

The p-value is very small, we can reject the null hypothesis that correlation is equal to 0 and conclude that Acc_perc and Grad.Rate are negatively correlated with a correlation coefficient equal to -0.161. So, we find a correlation between the variables. However, we cannot say much about a causal relationship. We do not know if Acc_perc causes Grad.Rate and vice versa. More extensive modelling and replication is required to prove a causal relationship.

(c) Bedenk zelf ook een extra hypothese om te toetsen en voer de hypothesetoets uit.
Is er een verband tussen persoonlijke uitgaven van studenten en de graduation rate?



The plots shows a negative relation between the graduation rate and personal spending. In other words, the schools with higher estimated spending have lower graduation rates. Lets put this to the test, but first let's check for normality.

```
shapiro.test(college_statistics_b$Personal)

##
##  Shapiro-Wilk normality test
##
## data:  college_statistics_b$Personal
## W = 0.89295, p-value < 2.2e-16

shapiro.test(college_statistics_b$Grad.Rate)

##
##  Shapiro-Wilk normality test
##
## data:  college_statistics_b$Grad.Rate
## W = 0.99259, p-value = 0.0006568
```

H0: Personal/Grad.Rate is normally distributed

Ha: Personal/Grad.Rate is not normally distributed

We can reject the null hypothesis at the 1% level for bot variables. Therefore, we can conclude both variables are not normally distributed. Correlation between non normally

distributed variables are best tested using the non parametric method (Kendall Rank correlation test)

```
cor.test(college_statistics_b$Personal, college_statistics_b$Grad.Rate, metho
d = 'kendall')
```

```
##
##  Kendall's rank correlation tau
##
## data:  college_statistics_b$Personal and college_statistics_b$Grad.Rate
## z = -8.002, p-value = 1.225e-15
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## -0.1949937
```

H0: correlation is equal to 0

Ha: correlation is not equal to 0

The p-value is very small, we can reject the null hypothesis that correlation is equal to 0 and conclude that personal speninding and the graduation rate are negatively correlated with a correlation coefficient equal to -0.195. Again, we cannot say much about a causal relationship, but there's definitely correlation between the variables.