Report on TASK 1a

group 27

21 April 2018

Data Collection, Cleaning and Exploration

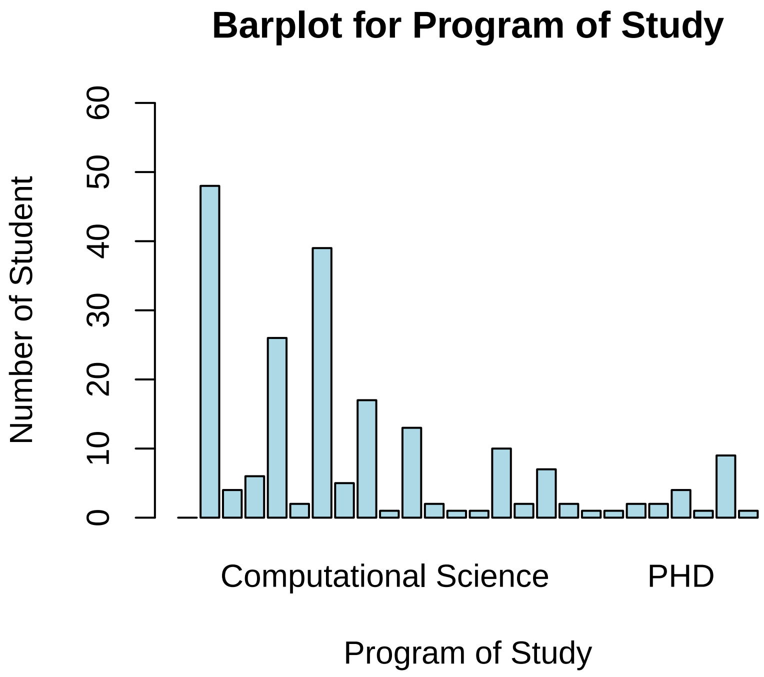
The dataset: *ODI-2018.csv* contained students response that has been gathered during the first lecture of this course. Our findings about the records of the raw data is given below: * Total number of records : 218 * Real record : 217 * 1st record : empty (maybe for header!) * Total number of attributes: 16 * The attributes related to Gender, Machine Learning course, Information Retrieval course, Database course, Statistics course and Chocolates are categorial types. * The study program is string type attribute, good day

The study program feature was cleaned using the since there are numerous input for same study program. The original 'birthday' feature was cleaned and splitted into a day, month and year feature. Unfortunately, the 'bedtime' feature cleaning was problematic and was manually formatted. Finally, the 'good day' features were cleaned using exact string matching and the levenstein distance coefficient. We removed some columns that we found not much interesting to analyse.

After finishing data cleaning we generated our clean dataset : *ODI-2018_clean.csv* which we are going to use for further data exploration.

Various Plots for Dataset

We generated the barplot to identify some interesting facts about study program. Looking at the barplot we can say that such a diverse group of students are taking the data mining course(26 different program). Among them, maximum number of students(48) are from Artifitial Intelligence background. We found one invalid entry for this attribute(i.e 12-05-1995).

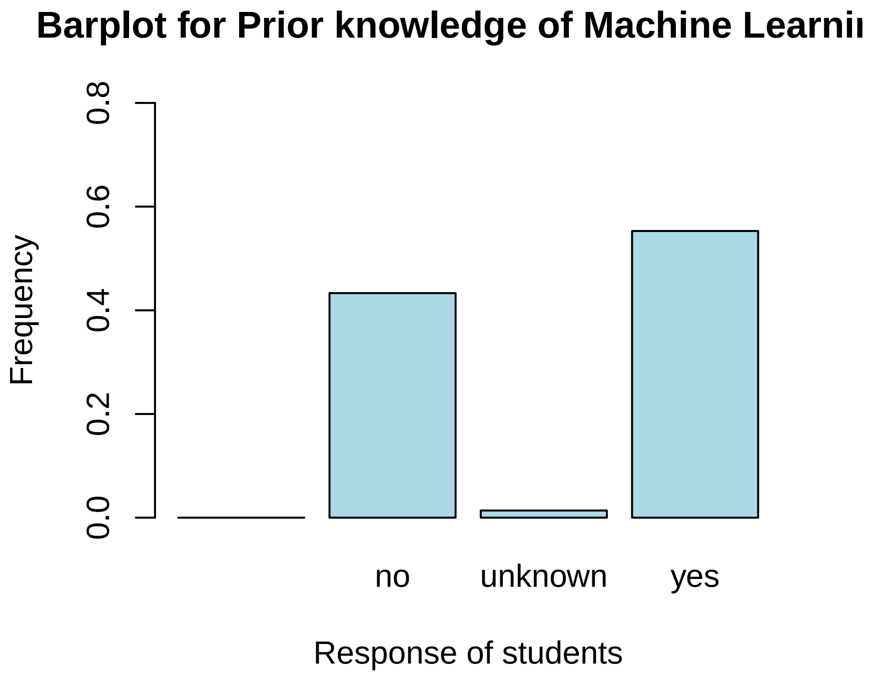


Import pie chart for gender

From the Pie chart we can see that more male students(150) are taking this course compared to female students(63). And there is even some entries(4) for Unknown gender type.

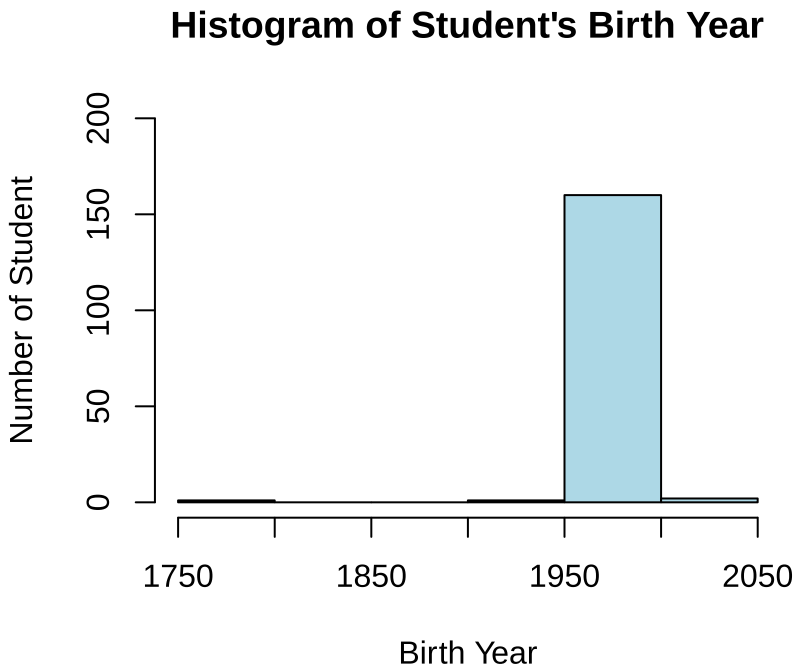
Barplot for Have you taken Machine Learning

We wanted to check the how many students have taken Machine Learning course prior to take this course. For this we plotted a barplot with the Machine learning course attribute. If we look into the barplot we see that highest frequency = yes that is, approximately 0.6. And in total 120 students have taken Machine learning while on the other hand 94 students didn't take it and 3 answers are unknown.

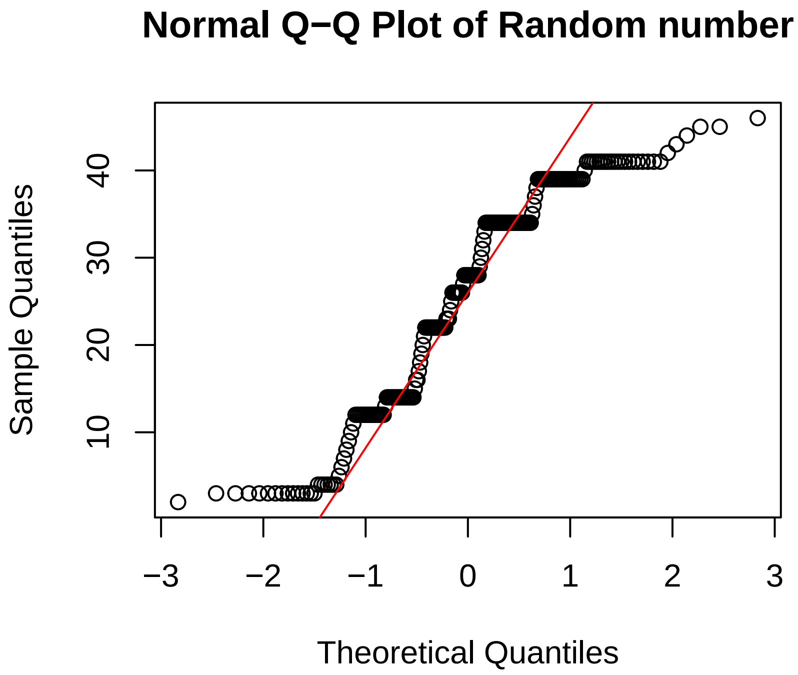


Histogram of Birth Year

Observing the above histogram of Student's birth year we figured out some ranges. For instance majority of students have birth year within (1950 - 2000). There are also three ranges that was valid but cannot be possible as birth year for a masters students. Those ranges are *(1750-1800) = 1 student, (1900-1950) = 1 student, (2000-2050) = 3 students*. Maximum students gave $NA(53)$ answer for this particular question. There are also some interesting entries for birth year i.e. 1768, 1931, 2000, 2018.



QQ-Plot of given random number between 0 to 100



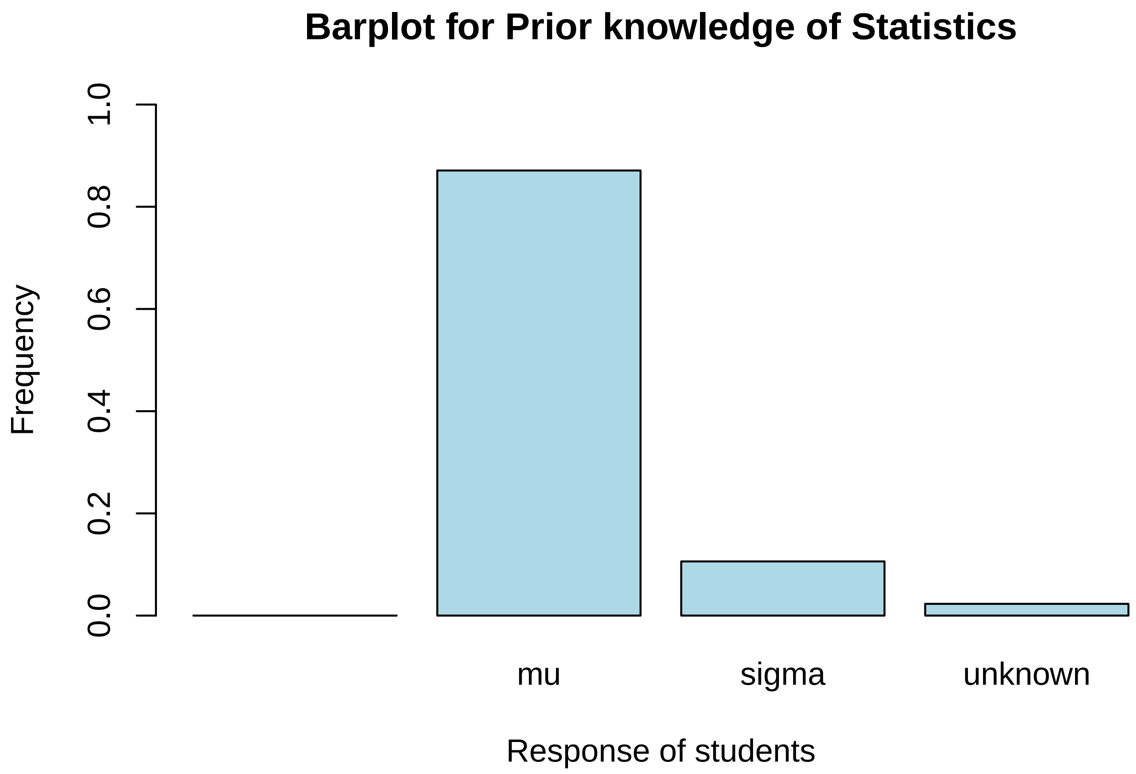
The QQ-Plot doesn't seem normal rather it seems like stepped! We could see the highest frequency

= 7 from the barplot. Although maximun limit was 100 but there are some strange/invalid input for example : *rnorm(n=1,mu=12,sigma=1)*

Have you taken statistics

We also wanted to investigate the amount of students who have prior knowledge of statistics since this course has lots of statistical data to analyse.

```r
stat = as.factor(Have.you.taken.a.course.on.statistics.)
barplot(prop.table(table(stat)),  main="Barplot for Prior knowledge of Statistics",
        xlab= "Response of students",
        ylab="Frequency",
        col="lightblue",
        ylim = c(0, 1))
```



From the barplot we found highest frequency = Yes (~0.85) And from pie chart we found the total number of response that is: yes = 189, No = 23 and Unknown = 5