Data Project Proposal

Alexis Mekueko

10/24/2020

R Packages

```
library(tidyverse) #loading all library needed for this assignment
library(openintro)
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.0.3
```

```
#head(fastfood)
library(readxl)
library(data.table)
#library(DT)
library(knitr)

library(readr)
#library(plyr)
library(dplyr)
library(stringr)
#library(XML)
#library(RCurl)
#library(jsonlite)
#library(httr)

#library(maps)
#library(dice)
# #library(VennDiagram)
# #library(help = "dice")
#ibrary(DBI)
#library(dbplyr)

# library(rstudioapi)
# library(RJDBC)
# library(odbc)
# library(RSQLite)
# #library(rvest)

#library(readtext)
#library(ggpubr)
#library(fitdistrplus)
```

```
#library(ggplot2)
#library(moments)
#library(qualityTools)
#library(normalp)
#library(utils)
#library(MASS)
#library(qqplotr)
library(DATA606)
```

```
##
## Welcome to CUNY DATA606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 3rd Edition. You can read this by typing
## vignette('os3') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='DATA606') will list the demos that are available.
```

```
#library(knitLatex)
#library(knitr)
#library(markdown)
#library(rmarkdown)
#render("DATA606_Project_Proposal.Rmd", "pdf_document")
```

Github Link: https://github.com/asmozo24/DATA606_Project_Proposal

Web link: https://rpubs.com/amekueko/682247

data source: https://archive.ics.uci.edu/ml/machine-learning-databases/00320/

Data Preparation

```
# load the text file which has the description of all the variable.
variable_details <- read.delim("https://raw.githubusercontent.com/asmozo24/DATA606_Project_Proposal/mai
variable_details
```

```
##                                                     X..Attributes.for.both.student.mat.csv..Math.cour
## 1                                                                           1-school: student's schoo
## 2
## 3
## 4                                                                                         4-address
## 5                                                                                5-famsize: family s
## 6                                                                               6-Pstatus: parent's
## 7  7-Medu: mother's education (numeric: 0 - none,  1 - primary education (4th grade), 2 â\200" 5th t
## 8  8-Fedu: father's education (numeric: 0 - none,  1 - primary education (4th grade), 2 â\200" 5th t
## 9                                   9-Mjob: mother's job (nominal: teacher, health care related, civ
## 10                                 10-Fjob: father's job (nominal: teacher, health care related, civ
## 11                                                 11-reason: reason to choose this school (nominal:
## 12                                                                                             12-g
## 13                                    13-traveltime: home to school travel time (numeric: 1 - <15 m
## 14                                                  14-studytime: weekly study time (numeric: 1 - <2
## 15                                                                                       15-failures
## 16
## 17
## 18                                                                 18-paid: extra paid classes withi
## 19                                                                                                 1
## 20
## 21
## 22
## 23
## 24                                                                       24-famrel: quality of famil
## 25                                                                              25-freetime: free ti
## 26                                                                                 26-goout: going o
## 27                                                                             27-Dalc: workday alco
## 28                                                                             28-Walc: weekend alco
## 29                                                                                 29-health: curren
## 30                                                                                                 3
## 31                                                                                        # these gr
## 32
## 33
## 34
## 35                                                                              Additional note: the
## 36                                                                                         These stu
## 37                                                                                                th
```

```r
student_math <- read.csv("https://raw.githubusercontent.com/asmozo24/DATA606_Project_Proposal/main/stude
glimpse(student_math)
```

```
## Rows: 395
## Columns: 33
## $ school     <chr> "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "G...
## $ sex        <chr> "F", "F", "F", "F", "F", "M", "M", "F", "M", "M", "F", "...
## $ age        <int> 18, 17, 15, 15, 16, 16, 16, 17, 15, 15, 15, 15, 15, 15, ...
## $ address    <chr> "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "...
## $ famsize    <chr> "GT3", "GT3", "LE3", "GT3", "GT3", "LE3", "LE3", "GT3", ...
## $ Pstatus    <chr> "A", "T", "T", "T", "T", "T", "T", "A", "A", "T", "T", "...
## $ Medu       <int> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4, 3, 3,...
## $ Fedu       <int> 4, 1, 1, 2, 3, 3, 2, 4, 2, 4, 4, 1, 4, 3, 2, 4, 4, 3, 2,...
## $ Mjob       <chr> "at_home", "at_home", "at_home", "health", "other", "ser...
## $ Fjob       <chr> "teacher", "other", "other", "services", "other", "other...
## $ reason     <chr> "course", "course", "other", "home", "home", "reputation...
## $ guardian   <chr> "mother", "father", "mother", "mother", "father", "mothe...
## $ traveltime <int> 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 2, 1, 1, 1, 3, 1,...
## $ studytime  <int> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3, 2, 1,...
## $ failures   <int> 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3,...
## $ schoolsup  <chr> "yes", "no", "yes", "no", "no", "no", "no", "yes", "no",...
## $ famsup     <chr> "no", "yes", "no", "yes", "yes", "yes", "no", "yes", "ye...
## $ paid       <chr> "no", "no", "yes", "yes", "yes", "yes", "no", "no", "yes...
## $ activities <chr> "no", "no", "no", "yes", "no", "yes", "no", "no", "no", ...
## $ nursery    <chr> "yes", "no", "yes", "yes", "yes", "yes", "yes", "yes", "...
## $ higher     <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", ...
## $ internet   <chr> "no", "yes", "yes", "yes", "no", "yes", "yes", "no", "ye...
## $ romantic   <chr> "no", "no", "no", "yes", "no", "no", "no", "no", "no", "...
## $ famrel     <int> 4, 5, 4, 3, 4, 5, 4, 4, 4, 5, 3, 5, 4, 5, 4, 4, 3, 5, 5,...
## $ freetime   <int> 3, 3, 3, 2, 3, 4, 4, 1, 2, 5, 3, 2, 3, 4, 5, 4, 2, 3, 5,...
## $ goout      <int> 4, 3, 2, 2, 2, 2, 4, 4, 2, 1, 3, 2, 3, 3, 2, 4, 3, 2, 5,...
## $ Dalc       <int> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2,...
## $ Walc       <int> 1, 1, 3, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2, 1, 4,...
## $ health     <int> 3, 3, 3, 5, 5, 5, 3, 1, 1, 5, 2, 4, 5, 3, 3, 2, 2, 4, 5,...
## $ absences   <int> 6, 4, 10, 2, 4, 10, 0, 6, 0, 0, 0, 4, 2, 2, 0, 4, 6, 4, ...
## $ G1         <int> 5, 5, 7, 15, 6, 15, 12, 6, 16, 14, 10, 10, 14, 10, 14, 1...
## $ G2         <int> 6, 5, 8, 14, 10, 15, 12, 5, 18, 15, 8, 12, 14, 10, 16, 1...
## $ G3         <int> 6, 6, 10, 15, 10, 15, 11, 6, 19, 15, 9, 12, 14, 11, 16, ...
```

```
view(student_math)
#student_math
#summary(hfi)
#dim(hfi)

student_math0 <- student_math[,c( 'address','Pstatus','studytime','schoolsup', 'famsup','activities','h
student_math1 <- student_math[,c( 'studytime', 'G1', 'G2', 'G3')]

view(student_math1)
# data looks pretty clean, but let's check the missing data
sum(is.na(student_math1)) # 0 means no NA found
```

```
## [1] 0
```

```
student_math1 <- student_math1 %>%
  mutate(studyTime10 = ifelse(student_math1$studytime > 3, "yes", "no"))

study10plus <- student_math1 %>%
  filter(studyTime10 == "yes" )  # & G1 & G2 & G3

 study10below <- student_math1 %>%
  filter(studyTime10 == "no" )
```

Research question

There are some study out there suggesting that study time likely affects students performance. Let's verify that in this study. Do students studing at least 10hrs weekly do well in class than those with single parent? We could also explore the corelation between study time and students performance. Is there a linear relationship between study time and students performance? In another words, do students putting more hours in study their lecons get better grades than those with few hours of in study time? How does study time impact students grades?

Cases

Each case represents a student at one of the two schools ("GP" - Gabriel Pereira or "MS" - Mousinho da Silveira). There are 395 observations in the given dataset

Data collection

Data is collected or made available by archive.ics.uci.edu: The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. The archive was created as an ftp archive in 1987 by David Aha and fellow graduate students at UC Irvine. The current version of the web site was designed in 2007 by Arthur Asuncion and David Newman, and this project is in collaboration with Rexa.info at the University of Massachusetts Amherst. Funding support from the National Science Foundation is gratefully acknowledged.

##Type of study this is observational/experimental study

##Data source

I found some interesting dataset from -> data source: https://archive.ics.uci.edu/ml/machine-learning-databases/00320/. This data is about a study on students(395) taking math or/and portuguese language course. the data is pretty rich with a txt file that described all variables in the data. therefore there is no need to rename the column. The orignal data format is comma delimited and rendering from R was not easy. So, I used excel with one attemp to fix it. I am interested in the student taking Math course. with 33 variables. Data available –> https://github.com/asmozo24/DATA606_Project_Proposal

##Response the response variable is studytime and it is numerical.

##Explanatory The explanatory variable is student grade or the mean in student grade and it is numerical.

#Relevant summary statistics

```
describe(student_math1$studytime)
```

```
##    vars   n mean   sd median trimmed mad min max range skew kurtosis   se
## X1    1 395 2.04 0.84      2    1.96   0   1   4     3 0.63    -0.04 0.04
```

```
describe(student_math1$G1)
```

```
##    vars   n  mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 395 10.91 3.32     11    10.8 4.45   3  19    16 0.24    -0.71 0.17
```

```
describe(student_math1$G2)
```

```
##    vars   n  mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 395 10.71 3.76     11   10.84 2.97   0  19    19 -0.43     0.59 0.19
```

```
describe(student_math1$G3)
```

```
##    vars   n  mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 395 10.42 4.58     11   10.84 4.45   0  20    20 -0.73     0.37 0.23
```

```
describe(study10plus$G3)
```

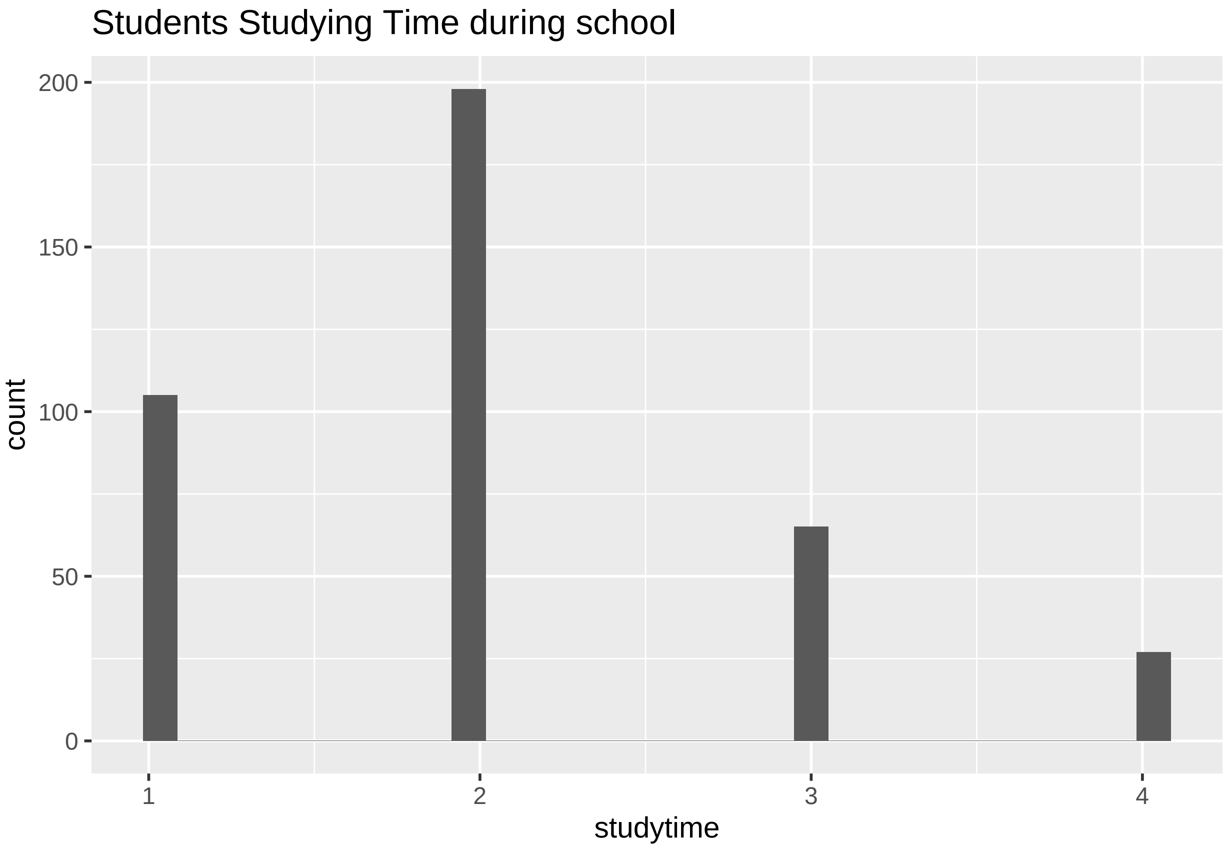
```
##    vars  n  mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 27 11.26 5.28     12   11.57 4.45   0  20    20 -0.7    -0.07 1.02
```

```
describe(study10below$G3)
```

```
##    vars   n  mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 368 10.35 4.53     11   10.78 4.45   0  19    19 -0.74     0.39 0.24
```

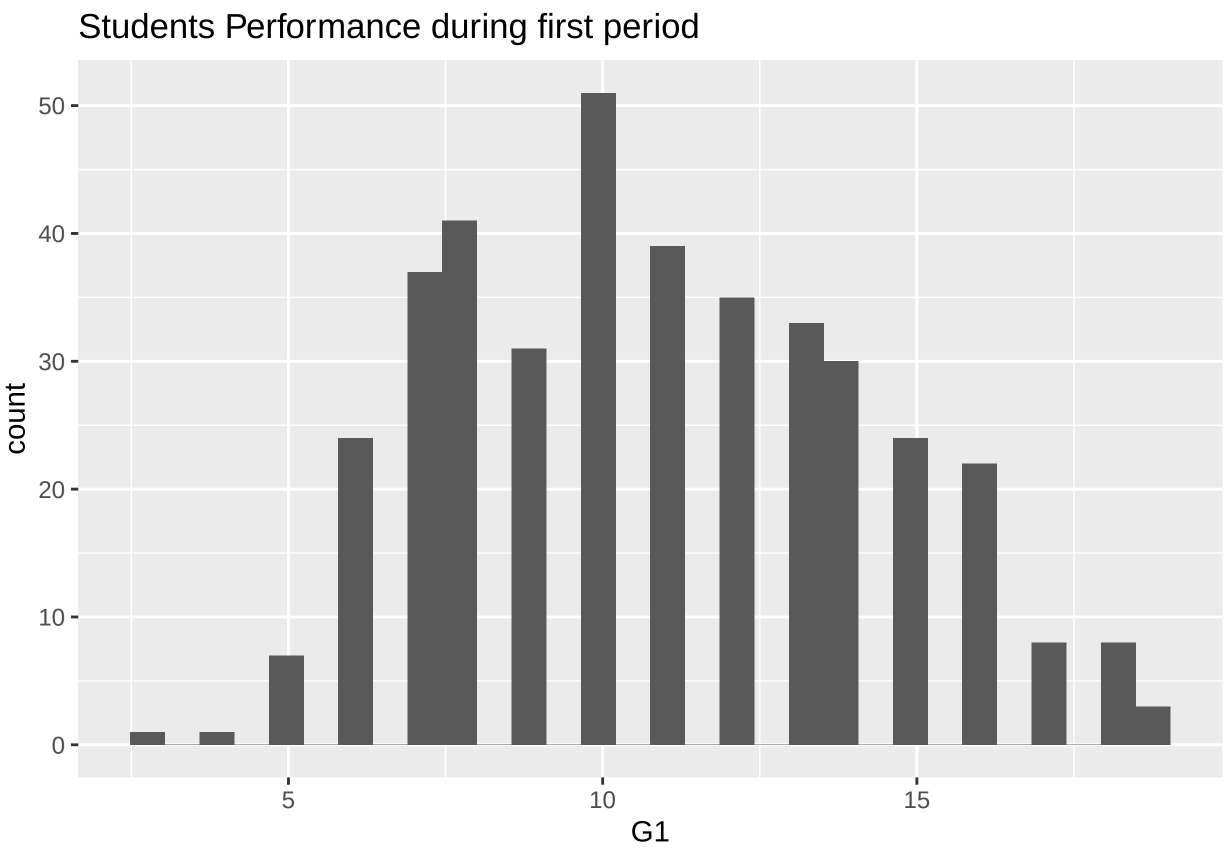
```
# Let's look at the distribution for each vration
ggplot(student_math1, aes(x=studytime)) + geom_histogram() + ggtitle("Students Studying Time during scho
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



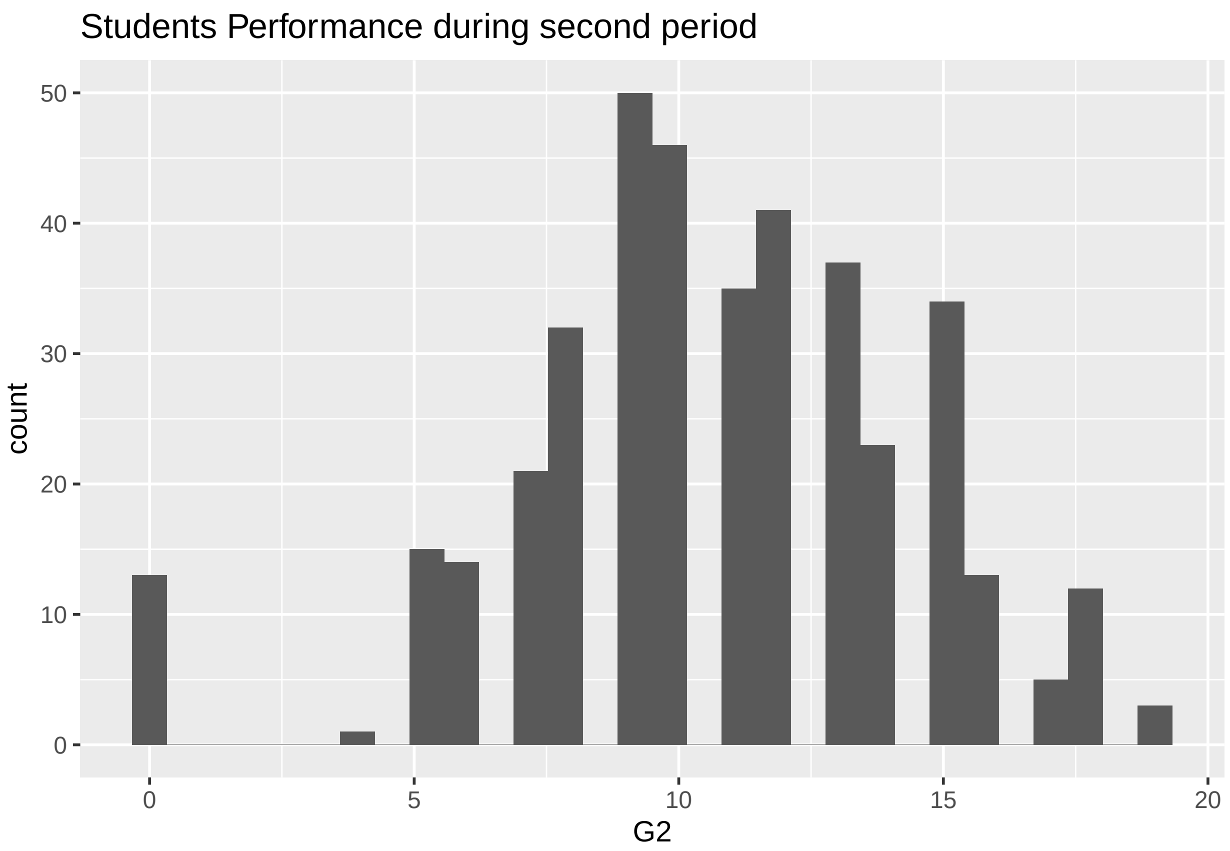
```
ggplot(student_math1, aes(x=G1)) + geom_histogram() + ggtitle("Students Performance during first period
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



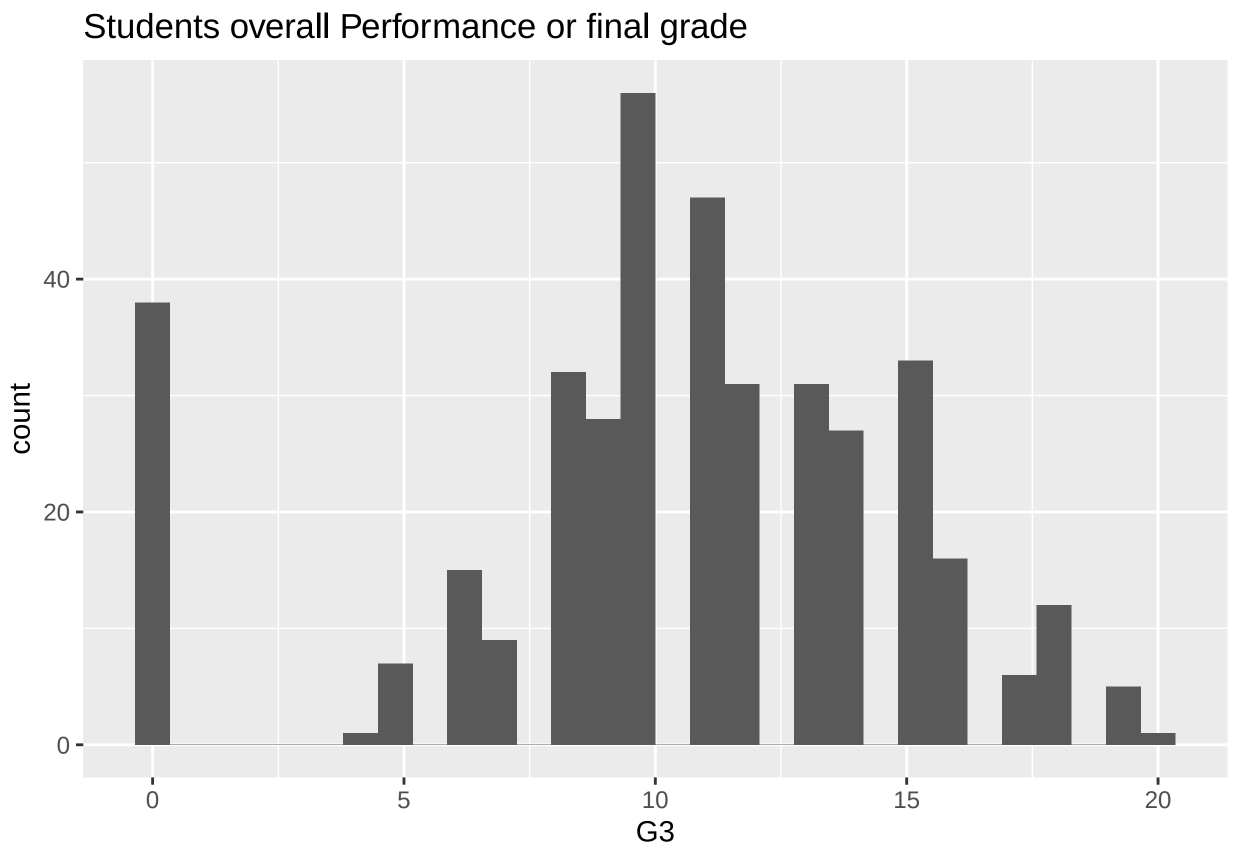
```
ggplot(student_math1, aes(x=G2)) + geom_histogram() + ggtitle("Students Performance during second perio
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



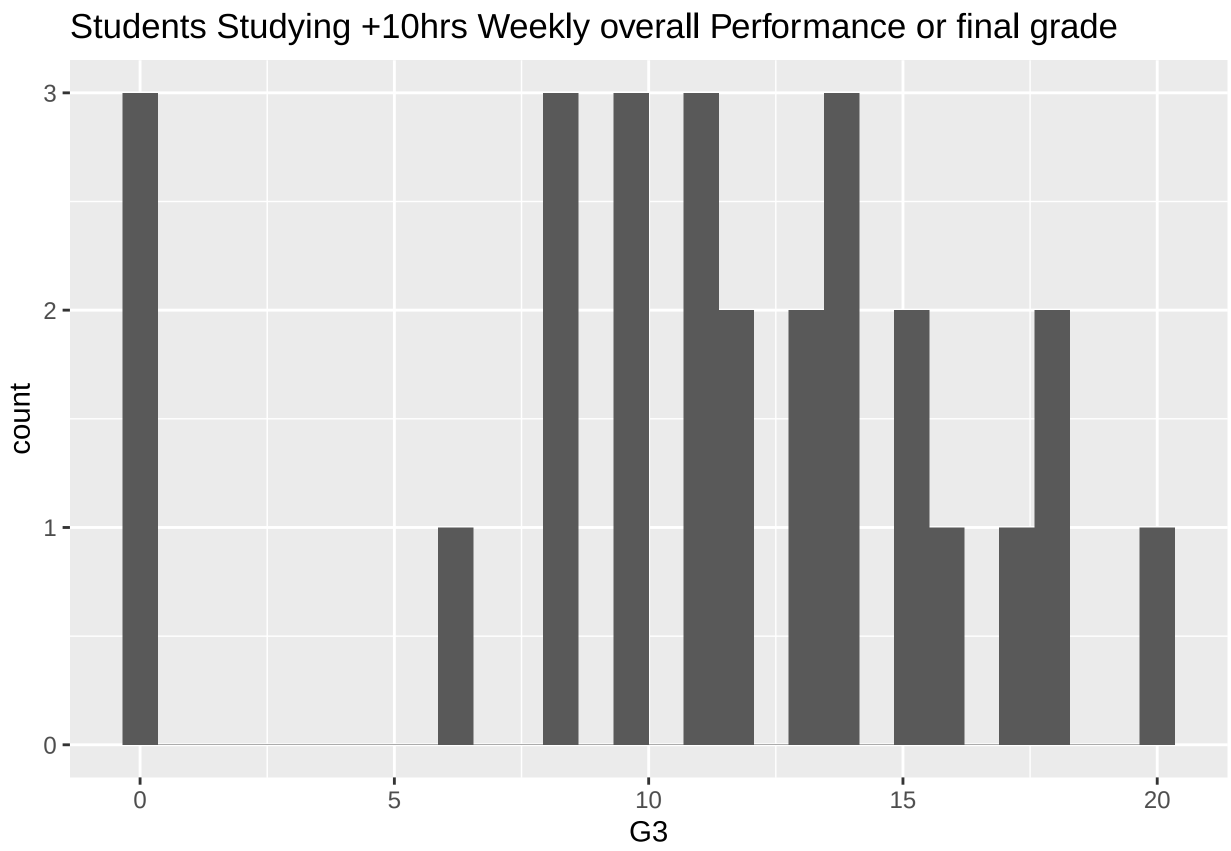
```
ggplot(student_math1, aes(x=G3)) + geom_histogram() + ggtitle("Students overall Performance or final gra
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



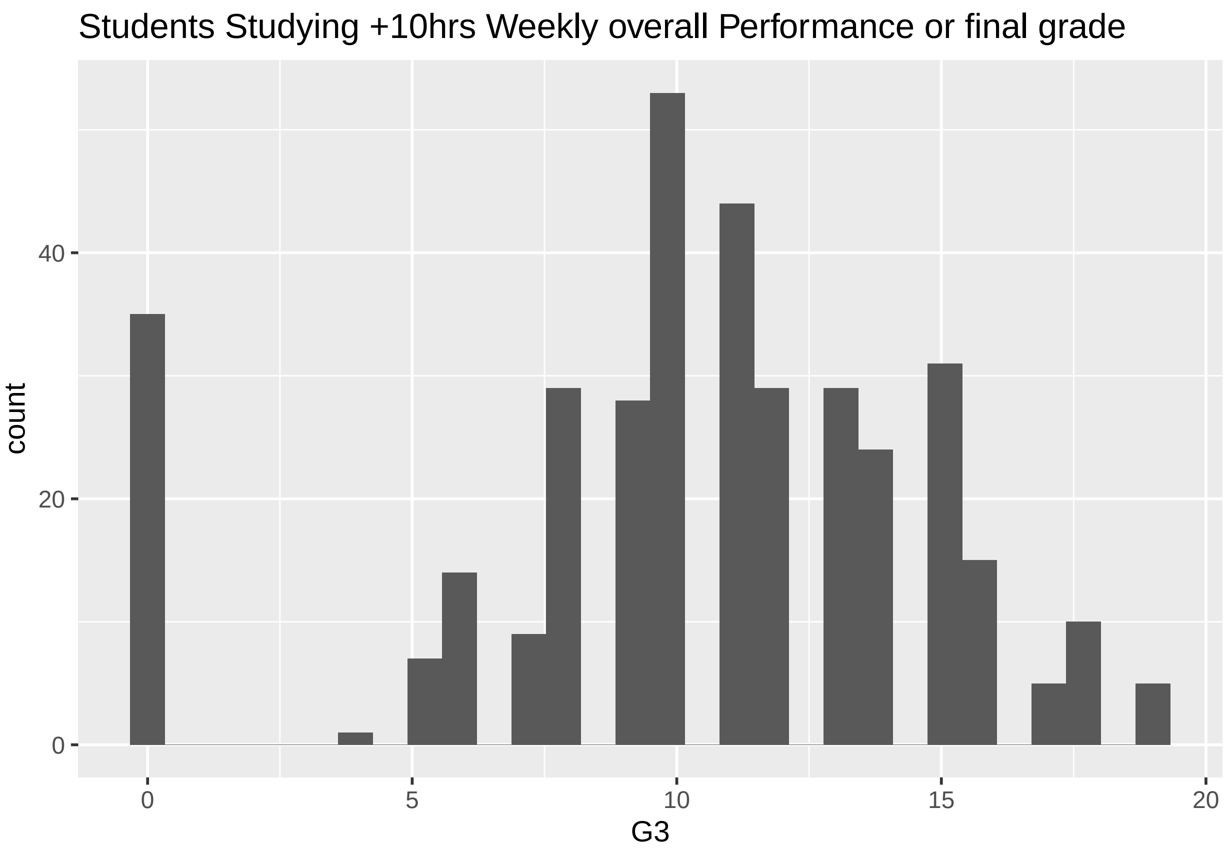
```r
ggplot(study10plus, aes(x=G3)) + geom_histogram() + ggtitle("Students Studying +10hrs Weekly overall Pe
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

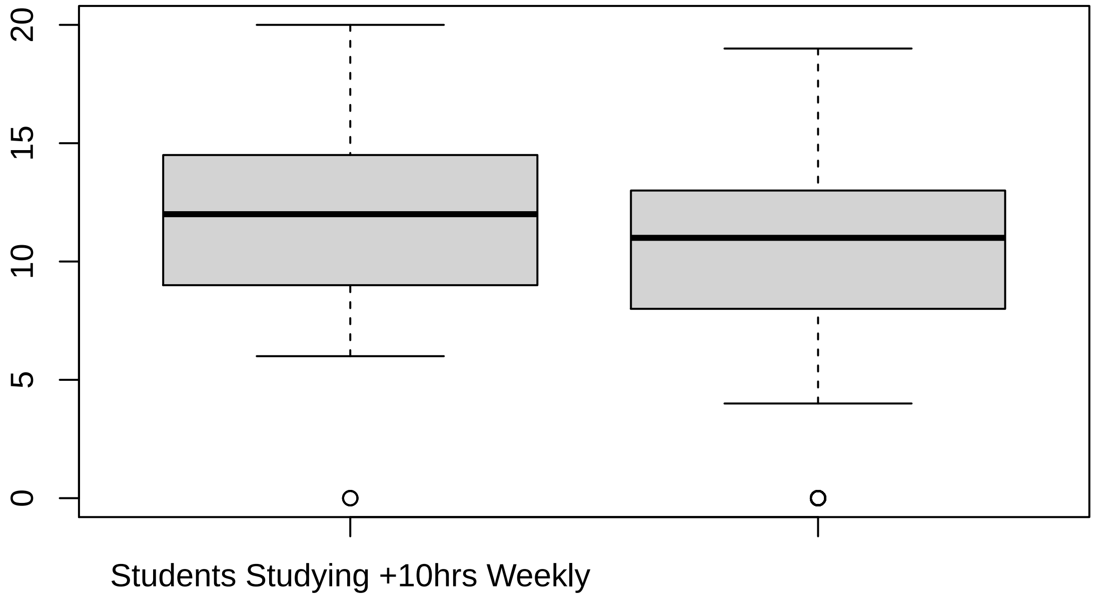


```
ggplot(study10below, aes(x=G3)) + geom_histogram() + ggtitle("Students Studying +10hrs Weekly overall P
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



```
boxplot(study10plus$G3, study10below$G3, names = c("Students Studying +10hrs Weekly", "Students Studyin
```



#Conclusion The New York Times API can be easy to use in scraping articles published on their website. However, I think the website it is pretty nested and need a better understanding of the New York Times website structure.