DSO545 HW03

Xu Zhang

February 11, 2018

```
Data=read.csv("college_recent_grads.csv")
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggplot2)
```

(1) Create a dataframe (call it female_engineering) for all engineering majors who has 50% or more female students. Sort the dataframe in decreasing order based on the % of female students.

```
female_engineering=
  Data %>%
  filter(major_category=="Engineering", sharewomen>=0.5) %>%
  arrange(-sharewomen)
```

(2) What is the total number (sum) of both engineering and business majors?

```
Q2=Data %>%
  filter(major_category %in% c("Engineering","Business"))
  nrow(Q2)

## [1] 42
```

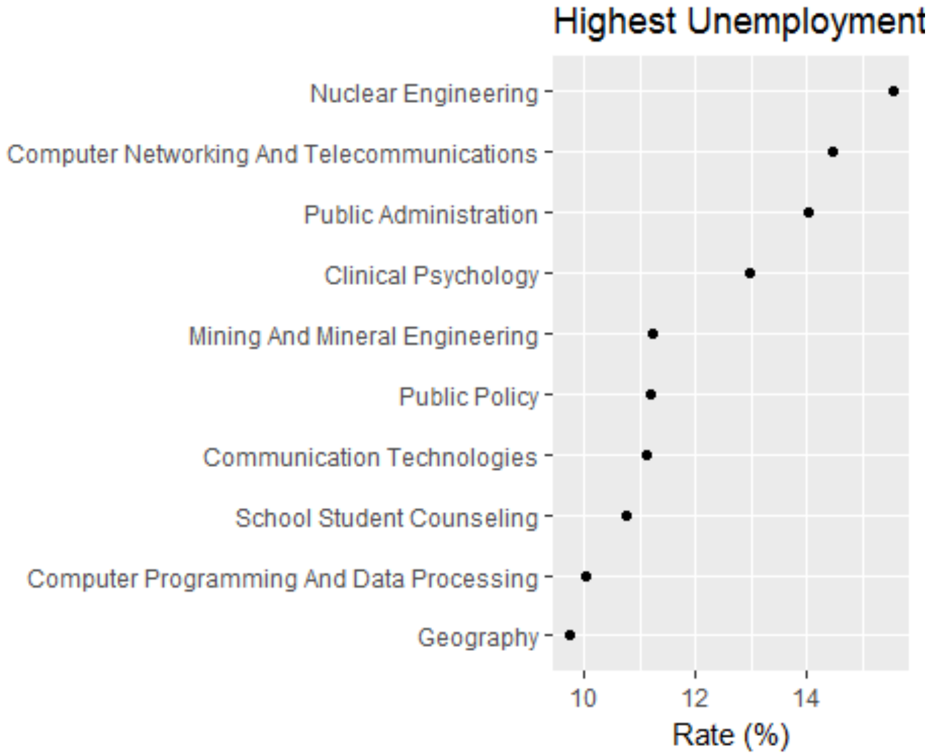
(3) Create a new variable unemployment_rate (defined as the the number of people unemployed divided by the total), and add it to the original dataset. Then, return a table (call it top_10) of the 10 majors with the highest rates as well as those corresponding rates.

```
Top= Data %>%
       mutate(unemployment_rate= unemployed/total) %>%
       select(major,unemployment_rate) %>%
       arrange(-unemployment_rate)
Top_10=Top[1:10,]
Top_10
```

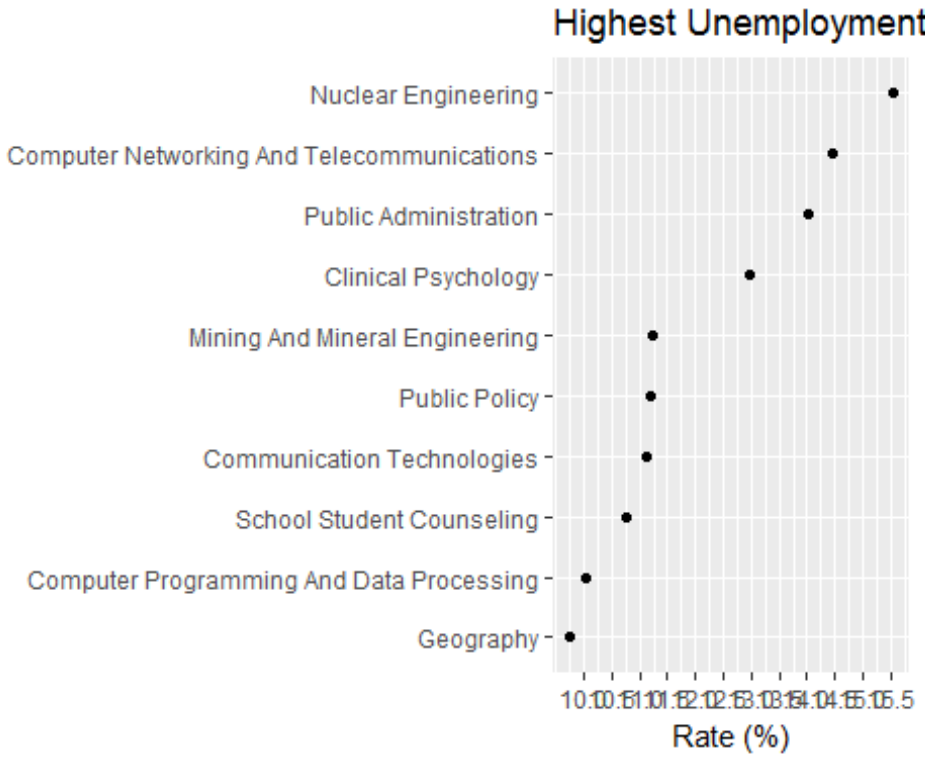
```
##                                          major unemployment_rate
## 1                          Nuclear Engineering        0.15546055
## 2  Computer Networking And Telecommunications        0.14448969
## 3                        Public Administration        0.14016699
## 4                          Clinical Psychology        0.12966878
## 5               Mining And Mineral Engineering        0.11243386
## 6                                Public Policy        0.11207762
## 7                   Communication Technologies        0.11122817
## 8                    School Student Counseling        0.10757946
## 9     Computer Programming And Data Processing        0.10052783
## 10                                   Geography        0.09734848
```

(4) Create a dot plot chart to show the highest unemployment rates for the diffrent majors.

```
ggplot(Top_10, aes(x=unemployment_rate*100,y = reorder(major, unemployment_ra
te)))+
    xlab("Rate (%)")+
    ylab("")+
    ggtitle("Highest Unemployment Rates by Major")+
    geom_point()
```



```
ggplot(Top_10, aes(x=unemployment_rate*100,y = reorder(major, unemployment_ra
te)))+
    scale_x_continuous(breaks = seq(10,15.5,0.5))+
    xlab("Rate (%)")+
    ylab("")+
    ggtitle("Highest Unemployment Rates by Major")+
    geom_point()
```



(6) Create a dataframe (call it majors_total) that shows the total number of students in each of the major categories. Which major category had the highest number of students?

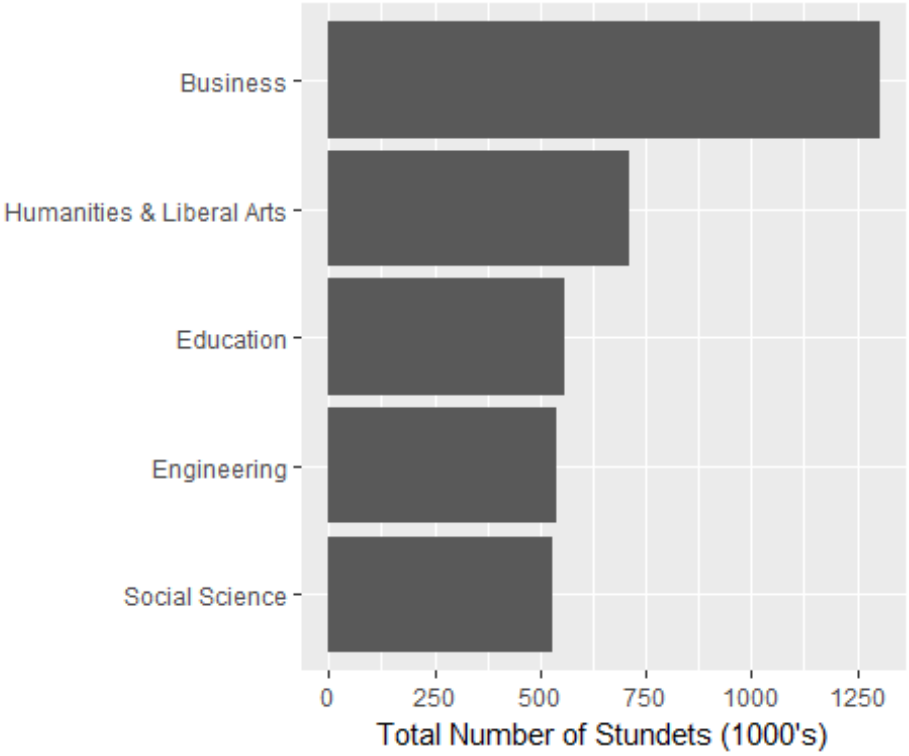
```
majors_total=Data %>%
       select(major_category,total) %>%
       group_by(major_category) %>%
       summarise(Sum_of_total=sum(total)) %>%
       arrange(-Sum_of_total)
majors_total=as.data.frame(majors_total)
majors_total

##                          major_category Sum_of_total
## 1                              Business      1302376
## 2             Humanities & Liberal Arts       713468
## 3                             Education       559129
## 4                           Engineering       537583
## 5                        Social Science       529966
## 6              Psychology & Social Work       481007
## 7                                Health       463230
## 8               Biology & Life Science        453862
## 9            Communications & Journalism      392601
## 10                                 Arts       357130
## 11              Computers & Mathematics       299008
## 12 Industrial Arts & Consumer Services       229792
## 13                     Physical Sciences       185479
## 14                   Law & Public Policy       179107
## 15     Agriculture & Natural Resources          79981
## 16                     Interdisciplinary        12296
```

Business category has the highest number of students.

(7) Create an EXACT copy of the following graph of the 5 major categories with the most total students.

```
Top5major_catagory=as.data.frame(majors_total[1:5,])
ggplot(Top5major_catagory, aes(x = reorder(major_category, Sum_of_total),y=Su
m_of_total/1000))+
  scale_y_continuous(breaks = seq(0,1250,250)) +
  geom_bar(stat = "identity")+
  xlab("") +
  ylab("Total Number of Stundets (1000's)") +
  coord_flip()
```



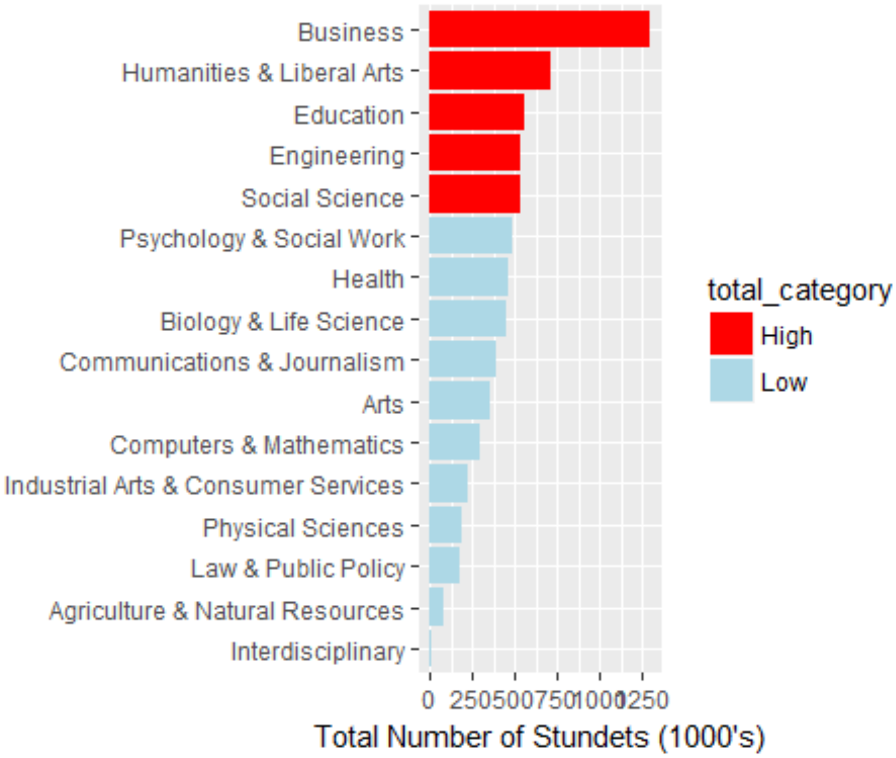
(8) Using the majors_total table you created earlier, create a new variable called total_category such that if total number of students is less than or equal 500,000, then the category is "Low", otherwise, it is "High". (You can use and if() or ifelse() statement to create the categories based on the specified condition)

```
majors_total$total_category=NA
for (i in 1:dim(majors_total)[1]){
    if(majors_total$Sum_of_total[i]>=500000){
      majors_total$total_category[i]="High"
      }
   else{
    majors_total$total_category[i]="Low"
   }
}
majors_total
```

```
##                          major_category Sum_of_total total_category
## 1                              Business      1302376           High
## 2             Humanities & Liberal Arts       713468           High
## 3                             Education       559129           High
## 4                           Engineering       537583           High
## 5                        Social Science       529966           High
## 6              Psychology & Social Work       481007            Low
## 7                                Health       463230            Low
## 8                Biology & Life Science       453862            Low
## 9           Communications & Journalism       392601            Low
## 10                                 Arts       357130            Low
## 11              Computers & Mathematics       299008            Low
## 12  Industrial Arts & Consumer Services       229792            Low
## 13                    Physical Sciences       185479            Low
## 14                  Law & Public Policy       179107            Low
## 15      Agriculture & Natural Resources        79981            Low
## 16                    Interdisciplinary        12296            Low
```

(9) Use the majors_total to create a copy of the following barchart (High =“red”, Low = “lightblue”):

```
ggplot(majors_total, aes(x = reorder(major_category,Sum_of_total),y=Sum_of_to
tal/1000,fill=total_category))+
  scale_y_continuous(breaks = seq(0,1250,250)) +
  geom_bar(stat = "identity")+
  xlab("") +
  ylab("Total Number of Stundets (1000's)") +
  coord_flip()+
  scale_fill_manual(values = c("red","lightblue"))
```


Business
Humanities & Liberal Arts
Education
Engineering
Social Science
Psychology & Social Work
Health
Biology & Life Science
Communications & Journalism
Arts
Computers & Mathematics
Industrial Arts & Consumer Services
Physical Sciences
Law & Public Policy
Agriculture & Natural Resources
Interdisciplinary
0
250
500
750
1000
1250
Total Number of Stundets (1000's)
total_category
High
Low