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)

1. 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

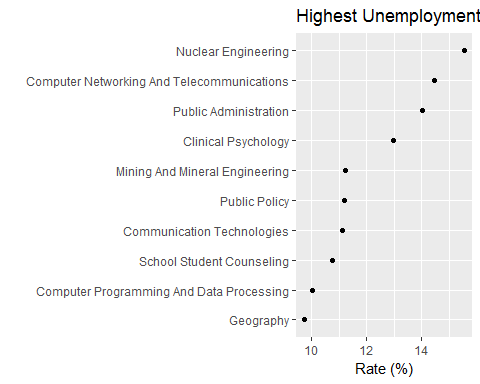
1. Create a new variable unemployment\_rate (defined as the the number of people unemployed divided by the total), and add it to the original dataset. Then, return a table (call it top\_10) of the 10 majors with the highest rates as well as those corresponding rates.

Top= Data %>%  
 mutate(unemployment\_rate= unemployed/total) %>%  
 select(major,unemployment\_rate) %>%  
 arrange(-unemployment\_rate)  
Top\_10=Top[1:10,]  
Top\_10

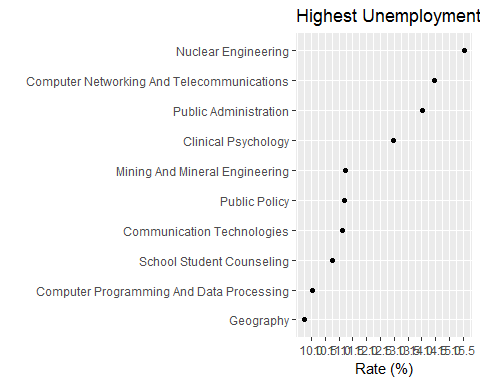
## major unemployment\_rate  
## 1 Nuclear Engineering 0.15546055  
## 2 Computer Networking And Telecommunications 0.14448969  
## 3 Public Administration 0.14016699  
## 4 Clinical Psychology 0.12966878  
## 5 Mining And Mineral Engineering 0.11243386  
## 6 Public Policy 0.11207762  
## 7 Communication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

1. Create a dot plot chart to show the highest unemployment rates for the differnt majors.

ggplot(Top\_10, aes(x=unemployment\_rate\*100,y = reorder(major, unemployment\_rate)))+  
 xlab("Rate (%)")+  
 ylab("")+  
 ggtitle("Highest Unemployment Rates by Major")+  
 geom\_point()



ggplot(Top\_10, aes(x=unemployment\_rate\*100,y = reorder(major, unemployment\_rate)))+  
 scale\_x\_continuous(breaks = seq(10,15.5,0.5))+  
 xlab("Rate (%)")+  
 ylab("")+  
 ggtitle("Highest Unemployment Rates by Major")+  
 geom\_point()



1. 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?

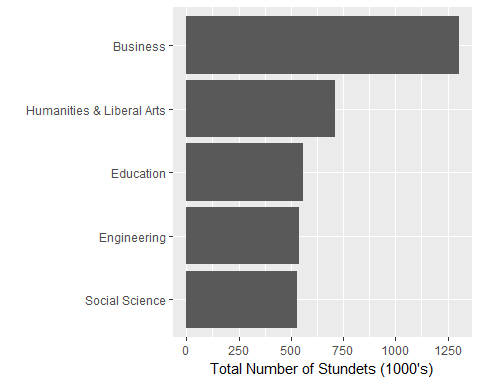
majors\_total=Data %>%  
 select(major\_category,total) %>%  
 group\_by(major\_category) %>%  
 summarise(Sum\_of\_total=sum(total)) %>%  
 arrange(-Sum\_of\_total)  
majors\_total=as.data.frame(majors\_total)  
majors\_total

## 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.

1. 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=Sum\_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()



1. 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

1. 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\_total/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"))

