Preperation for Senior Thesis

Roma Wang

1/16/2021

Introduction

This file is a preperation for my senior thesis, where I study how sex ratio of a university moderates the effect personality traits have on students' performance, GPA.

The dataset I use for my thesis is the College Senior Survey (CSS) from Higher Education Research Institute (HERI) from 1994 to 2006. The College Senior Survey is designed as an exit survey for graduating seniors. The CSS focuses on a broad range of college outcomes and post-college goals and plans including:

academic achievement and engagement student-faculty interaction cognitive and affective development student goals and values satisfaction with the college experience degree aspirations and career plans post-college plans

From this dataset, I will ultimately conduct several linear regressions with personality and sex ratio being independent variables and GPA being the dependent variable. For now, I will examine how different sexes and races differ on their scores in personality traits.

Importing CSS Data "CSS.TRENDS.94.06.csv" and Variable Data "my variables.xlsx"

import: original data rawdata: a copy of original data vars: variables of interest data: original data with only variables of interest

```
library(dplyr)
library(readxl)
library(data.table)
library(gt)
library(psych)
library(ggplot2)
library(scales)
```

```
setwd(dir="C:\\Users\\lixia\\OneDrive\\JHU\\senior thesis\\DATA")
import <- read.csv("C:\\Users\\lixia\\OneDrive\\JHU\\senior thesis\\DATA\\CSS.TRENDS.94.06.csv")
rawdata <- import
names(rawdata) <- tolower(names(rawdata))
vars <- read_excel("C:\\Users\\lixia\\OneDrive\\JHU\\senior thesis\\DATA\\my variables.xlsx")
x <- c()
for (i in vars$var_name){
  x=c(x,i)
}
data <- rawdata %>% select(all_of(x))
```

Processing

My potnetial variables of interest is documented in the “my variables.xlsx” file. In this file, I selected the main variables of interests, the Big Five Inventory (extraversion, conscientiousness, agreeableness, openness, neuroticism), competitiveness, confidence, and locus of control. Due to time constraint, I only searched for variables in the 1994 codebook and only discovered survey questions that represent five traits: “conscientiousness”,“extraversion”,“neuroticism”,“competitiveness”,“confidence”.

Construction of Variables

I went through the codebook to select a series of survey questions that document respondents’ personality traits. I recoded the variables with higher values indicating lower scores to let them match with the other variables with the same direction. I then normalized the variables for better statistical outcomes and replaced the original column with the normalized columns. Then, I calculated the row mean of all varaibles that represent one personality traits. Eventually, I create a new data frame to only store the means.

personality: a list that stores the five traits of interest cons_vars, extr_vars, neur_vars, comp_vars, conf_vars: lists that store the survey questions corresponding to each personality trait normdata: a data frame that contain the row means of the normalized survey question responses demo_vars: lists that store the demographic information keep_vars: a list that stores all variables to be used, conssisting of personality variables, demographic variables, and college GPA keepdata: a data frame with only variables to be used

```
personality <- c("conscientiousness","extraversion","neuroticism","competitiveness","confidence"
)

cons_vars <- c("hpw12","hpw13","hpw14","act03","act04","act08","act10","act11","act12","act13",
"act14")
for (i in c("act08","act10","act11","act14")) {
  if (sum(is.na(data[[i]]))!=nrow(data)) {      #do not recode columns that are all na's
    data[[i]] <- recode(data[[i]],`2`=2,`1`=3,`3`=1)   #recode the rest of the columns
  }
}
normdata <- data %>% mutate_at(cons_vars,~(scale(.) %>% as.vector))
normdata$conscientiousness <- rowMeans(normdata[,cons_vars],na.rm=TRUE)

extr_vars <- c("rate09","rate10","hpw01","hpw05","hpw07")
normdata <- normdata %>% mutate_at(extr_vars,~(scale(.) %>% as.vector))
normdata$extraversion <- rowMeans(normdata[,extr_vars],na.rm=TRUE)

neur_vars <- "rate05"
normdata$rate05 <- recode(normdata$rate05,`1`=5,`2`=4,`3`=3,`4`=2,`5`=1)
normdata <- normdata %>% mutate_at(neur_vars,~(scale(.) %>% as.vector))
normdata$neuroticism <- normdata$rate05

comp_vars <- c("rate03","rate04")
normdata <- normdata %>% mutate_at(comp_vars,~(scale(.) %>% as.vector))
normdata$competitiveness <- rowMeans(normdata[,comp_vars],na.rm=TRUE)

conf_vars <- c("rate12","rate13")
normdata <- normdata %>% mutate_at(conf_vars,~(scale(.) %>% as.vector))
normdata$confidence <- rowMeans(normdata[,conf_vars],na.rm=TRUE)

demo_vars <- c("sex","loanamt","poliview","race1","race2","race3","race4","race5","race6","race
7","race8","natengsp")

keep_vars <- unique(c("conscientiousness","extraversion","neuroticism","competitiveness","confid
ence",demo_vars,"collgpa"))
keepdata <- normdata %>% select(all_of(keep_vars))
```

Missing Data

The dataset is consisted mostly of numerical values with occasional string values. There are no open-ended questions. Only 15% of the observations have complete documentation on the five traits of interest.

```
personality_subset <- keepdata %>% select(all_of(personality))
percent_complete <- label_percent()(mean(complete.cases(personality_subset)))
percent_complete
```

```
## [1] "15%"
```

Reshaping Data

I reshaped the data frame from wide to long for future analysis becuase this is a panel dataset. GP is the measurement (dependent variable) and all else are conditions (independent variables). I added the identification column to the data frame and each identification (i.e. each row) describes a participant's response.

```
keepdata$id <- seq.int(nrow(keepdata))
keepdata <- keepdata %>% select(id, everything())

long <- tidyr::gather(keepdata,"condition","measurement","conscientiousness":"collgpa",factor_ke
y=FALSE)
```

Analysis

Descriptive Statistics of Personality by Sex and Race

I calculated the means of the personality traits for the entire sample, for only males and females, and for white people and non-white people.

overall_pers: a data frame that contains the mean values of personality traits for all respondents male, female, white, nonwhite: data frames that contain the mean values of personality traits for males, females, white people, and nonwhite people frame_pers: a data frame that contains the mean values of personality traits for all categories of people

Observation: Males score higher than females on conscientiousness, extraversion, competitiveness, and confidence. Females score higher on neuroticism. This is consistent with relevent literature. White people score higher on neuroticism, lower on conscientiousness and extraversion, and break-even on competitiveness and confidence.

```
overall_pers <- colMeans(keepdata[,personality],na.rm=TRUE)

male <- keepdata[keepdata$sex == 1,]
female <- keepdata[keepdata$sex == 2,]

male_pers <- colMeans(male[,personality],na.rm=TRUE)
female_pers <- colMeans(female[,personality],na.rm=TRUE)

white <- keepdata[keepdata$race1 == 2,]
nonwhite <- keepdata[keepdata$race1 != 2,]

white_pers <- colMeans(white[,personality],na.rm=TRUE)
nonwhite_pers <- colMeans(nonwhite[,personality],na.rm=TRUE)

names <- c("Overall","Male","Female","White","Non-White")
frame_pers <- transpose(data.frame(overall_pers,male_pers,female_pers,white_pers,nonwhite_pers))
frame_pers <- cbind(names,frame_pers)

table_pers <-
  frame_pers %>%
    gt(rowname_col = "names") %>%
    tab_header(
      title = "Personality by Sex and Race",
      subtitle = "Data from CSS 1994 to 2006"
    ) %>%
    tab_spanner(
      label = "Personality",
      columns = vars(V1,V2,V3,V4,V5)
    ) %>%
    cols_label(
      V1 = personality[[1]],
      V2 = personality[[2]],
      V3 = personality[[3]],
      V4 = personality[[4]],
      V5 = personality[[5]]
    ) %>%
    fmt_number(
      columns = vars(V1,V2,V3,V4,V5),
      decimals = 2
    ) %>%
    tab_footnote(
      footnote = "Data is standardized.",
      locations = cells_column_labels(
        columns = vars(V1,V2,V3,V4,V5)
      )
    )

table_pers
```

Personality by Sex and Race

Data from CSS 1994 to 2006

[1] Data is standardized.

Personality by Sex and Race

Data from CSS 1994 to 2006

	Personality				
	conscientiousness[1]	extraversion[1]	neuroticism[1]	competitiveness[1]	confidence[1]
Overall	−0.00	−0.00	0.00	0.00	−0.00
Male	0.04	0.05	−0.30	0.32	0.22
Female	−0.02	−0.03	0.17	−0.18	−0.13
White	−0.02	−0.02	0.02	0.00	−0.00
Non-White	0.08	0.09	−0.13	0.00	0.00

[1] Data is standardized.

Summary Statistics of Persoinality Traits

I generated the summary statistics for all five personality traits for the entire sample.

cons_des, extr_des, neur_des, comp_des, conf_des: descriptive statistics (data frames) of the personality traits
frame_des: a data frame that contains all summary statistics

Observations: People generally score the highest on neuroticism and lowest on conscientiousness. The distriubtions of neuroticism and competitiveness is more spread out than the distributions of other personality traits.

```
cons_des <- describe(keepdata$conscientiousness)
extr_des <- describe(keepdata$extraversion)
neur_des <- describe(keepdata$neuroticism)
comp_des <- describe(keepdata$competitiveness)
conf_des <- describe(keepdata$confidence)

frame_des <- rbind(cons_des,extr_des,neur_des,comp_des,conf_des)
frame_des <- cbind(personality,frame_des)

table_des <-
  frame_des %>%
  gt(rowname_col = "personality") %>%
  tab_header(
    title = "Personality Summary Statistics",
    subtitle = "Data from CSS 1994 to 2006"
  ) %>%
  fmt_number(
    columns = vars(mean,sd,median,trimmed,mad,min,max,range,skew,kurtosis,se),
    decimals = 2
  )

table_des
```

Personality Summary Statistics

Data from CSS 1994 to 2006

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
conscientiousness	1	247348	−0.00	0.54	−0.06	−0.03	0.50	−2.05	4.74	6.79	0.69
extraversion	1	246130	−0.00	0.60	0.01	−0.01	0.59	−4.16	3.74	7.91	0.09
neuroticism	1	53926	0.00	1.00	0.56	0.02	1.95	−2.07	3.19	5.26	−0.09
competitiveness	1	228138	0.00	0.99	0.37	0.05	1.61	−2.89	1.81	4.70	−0.26
confidence	1	244127	−0.00	0.75	−0.01	0.00	0.76	−3.27	1.77	5.04	−0.10

Graphs of Personality Traits by Sex

I created five graphs for how males and females score differently on the five personality traits.

Observations:

Males and females are equally spread out on almost all personality traits, except that the means of the traits are slightly different. Further studies are needed to establish the statistic rigor of the patterns.

```
plot_cons <- ggplot2::ggplot(keepdata,ggplot2::aes(x = sex, y = conscientiousness)) +
  ggplot2::geom_point() +
  ggplot2::ggtitle("Conscientiousness by Sex") +
  ggplot2::xlab("Male-Female")
  ggplot2::ylab("Conscientiousness")
```

```
## $y
## [1] "Conscientiousness"
##
## attr(,"class")
## [1] "labels"
```

```
plot_extr <- ggplot2::ggplot(keepdata,ggplot2::aes(x = sex, y = extraversion)) +
  ggplot2::geom_point() +
  ggplot2::ggtitle("Extraversion by Sex") +
  ggplot2::xlab("Male-Female")
  ggplot2::ylab("Extraversion")
```

```
## $y
## [1] "Extraversion"
##
## attr(,"class")
## [1] "labels"
```

```
plot_neur <- ggplot2::ggplot(keepdata,ggplot2::aes(x = sex, y = neuroticism)) +
  ggplot2::geom_point() +
  ggplot2::ggtitle("Neuroticism by Sex") +
  ggplot2::xlab("Male-Female")
  ggplot2::ylab("Neuroticism")
```

```
## $y
## [1] "Neuroticism"
##
## attr(,"class")
## [1] "labels"
```

```
plot_comp <- ggplot2::ggplot(keepdata,ggplot2::aes(x = sex, y = competitiveness)) +
  ggplot2::geom_point() +
  ggplot2::ggtitle("Competitiveness by Sex") +
  ggplot2::xlab("Male-Female")
  ggplot2::ylab("Competitiveness")
```

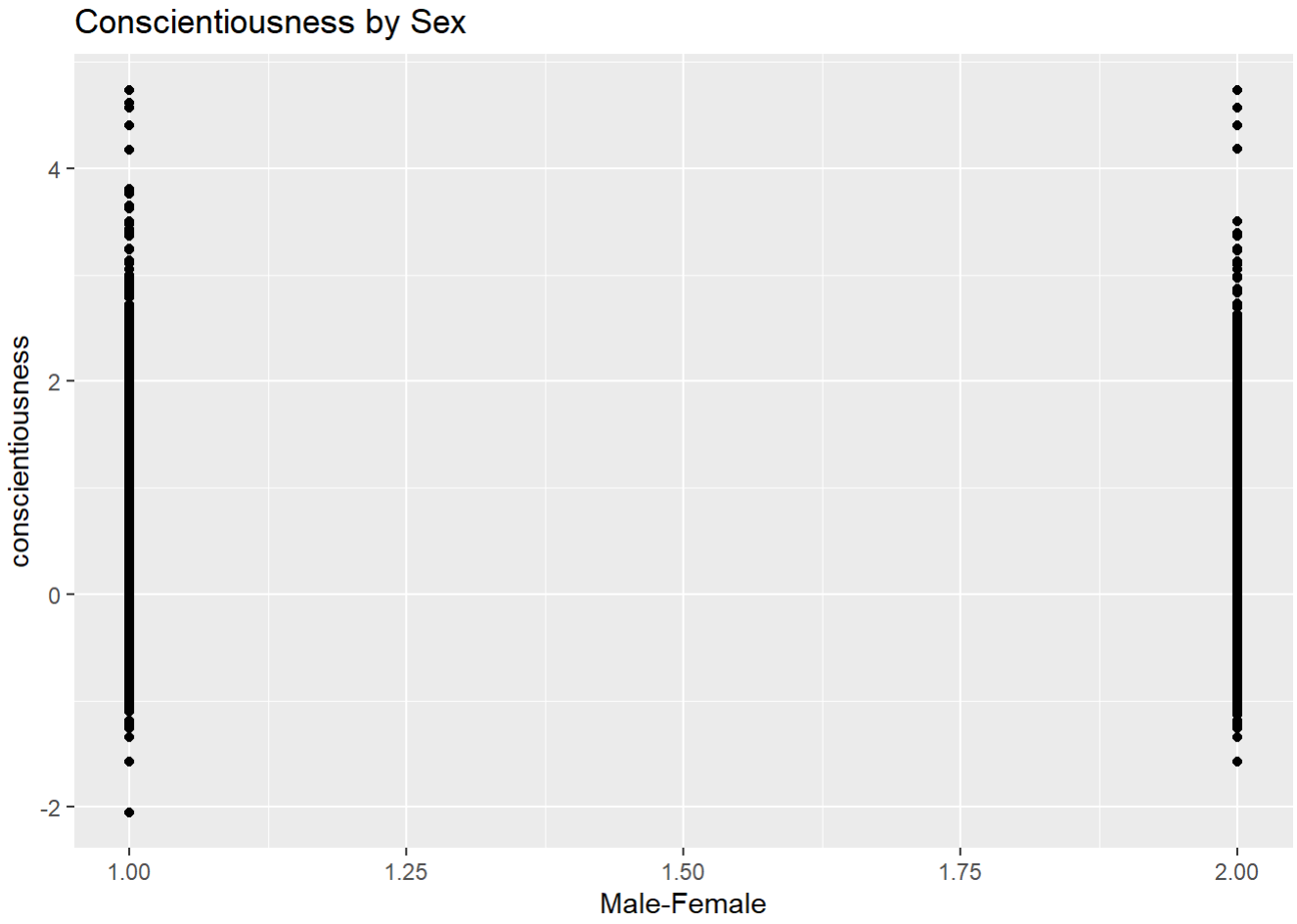
```
## $y
## [1] "Competitiveness"
##
## attr(,"class")
## [1] "labels"
```

```
plot_conf <- ggplot2::ggplot(keepdata,ggplot2::aes(x = sex, y = confidence)) +
  ggplot2::geom_point() +
  ggplot2::ggtitle("Confidence by Sex") +
  ggplot2::xlab("Male-Female")
  ggplot2::ylab("Confidence")
```

```
## $y
## [1] "Confidence"
##
## attr(,"class")
## [1] "labels"
```

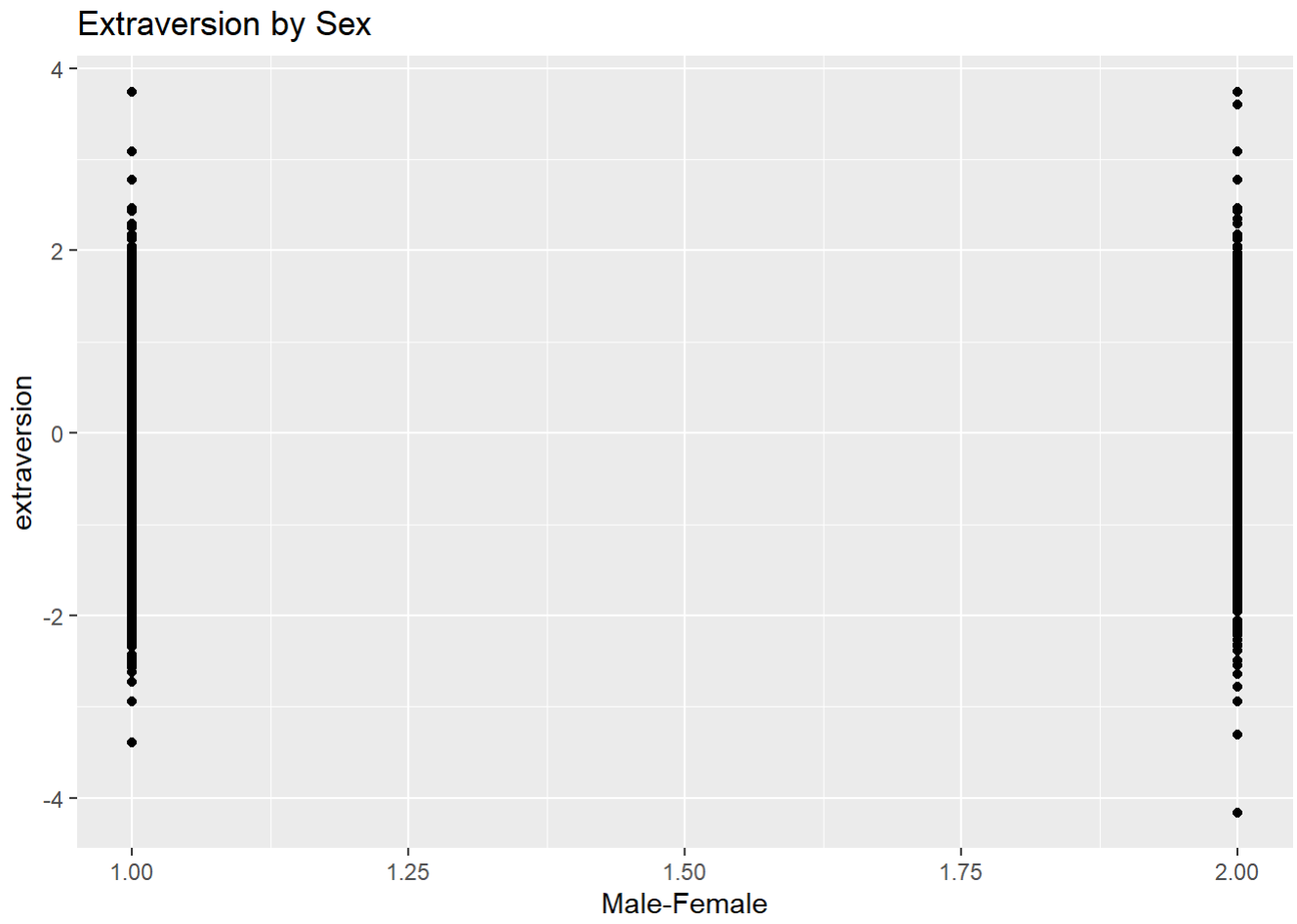
```
plot_cons
```

```
## Warning: Removed 1087 rows containing missing values (geom_point).
```



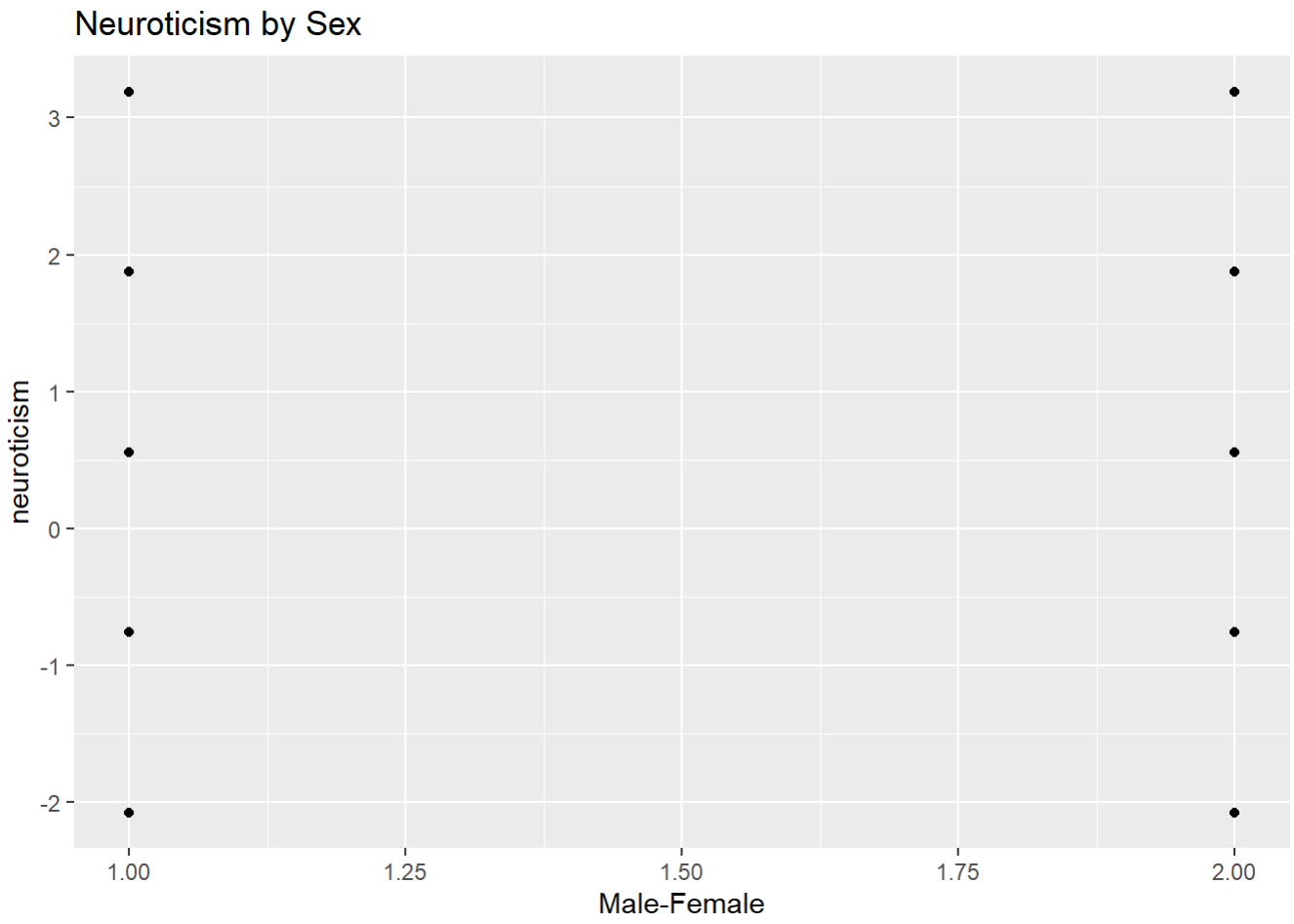

```
plot_extr
```

```
## Warning: Removed 2298 rows containing missing values (geom_point).
```



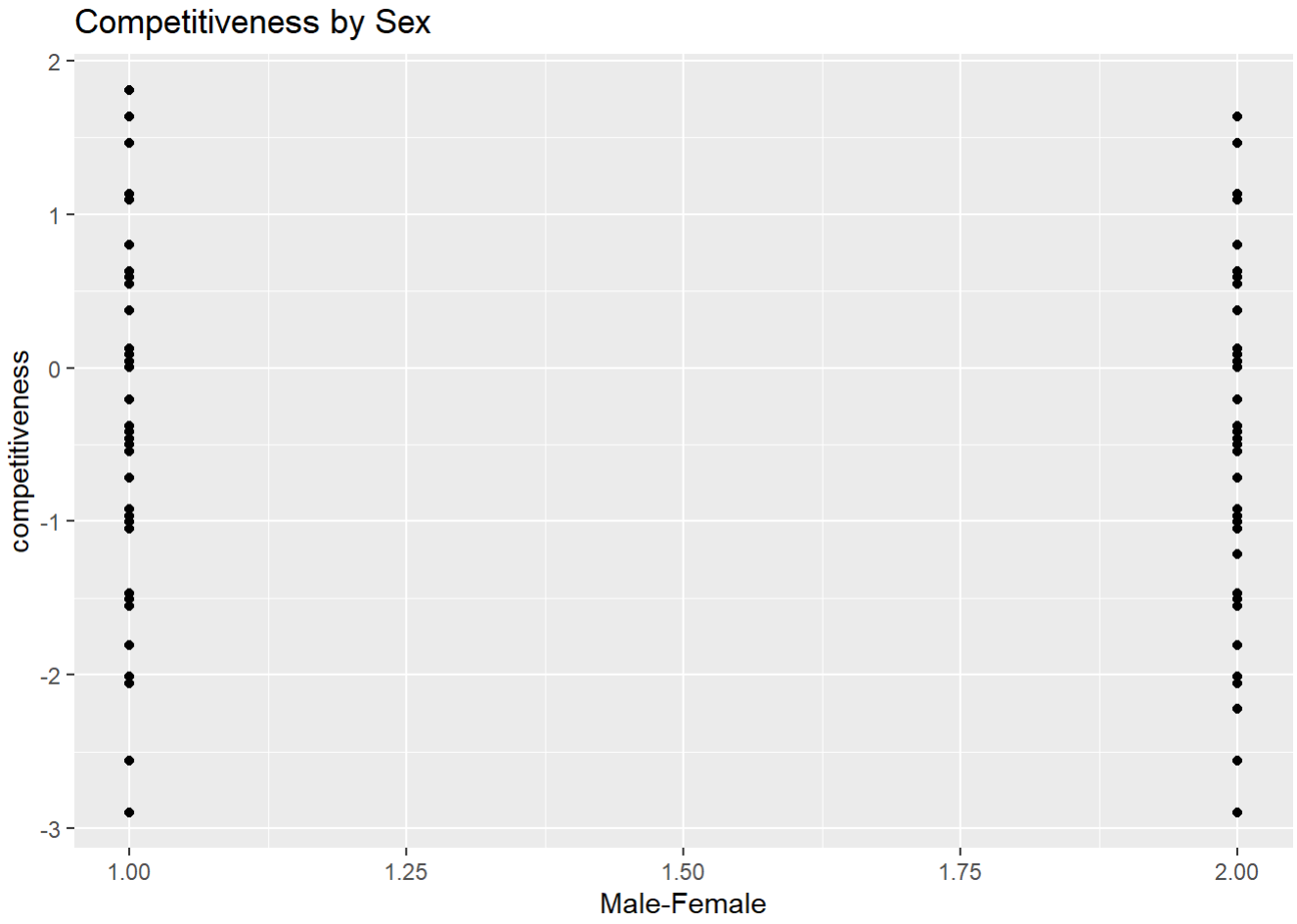

```
plot_neur
```

```
## Warning: Removed 194192 rows containing missing values (geom_point).
```



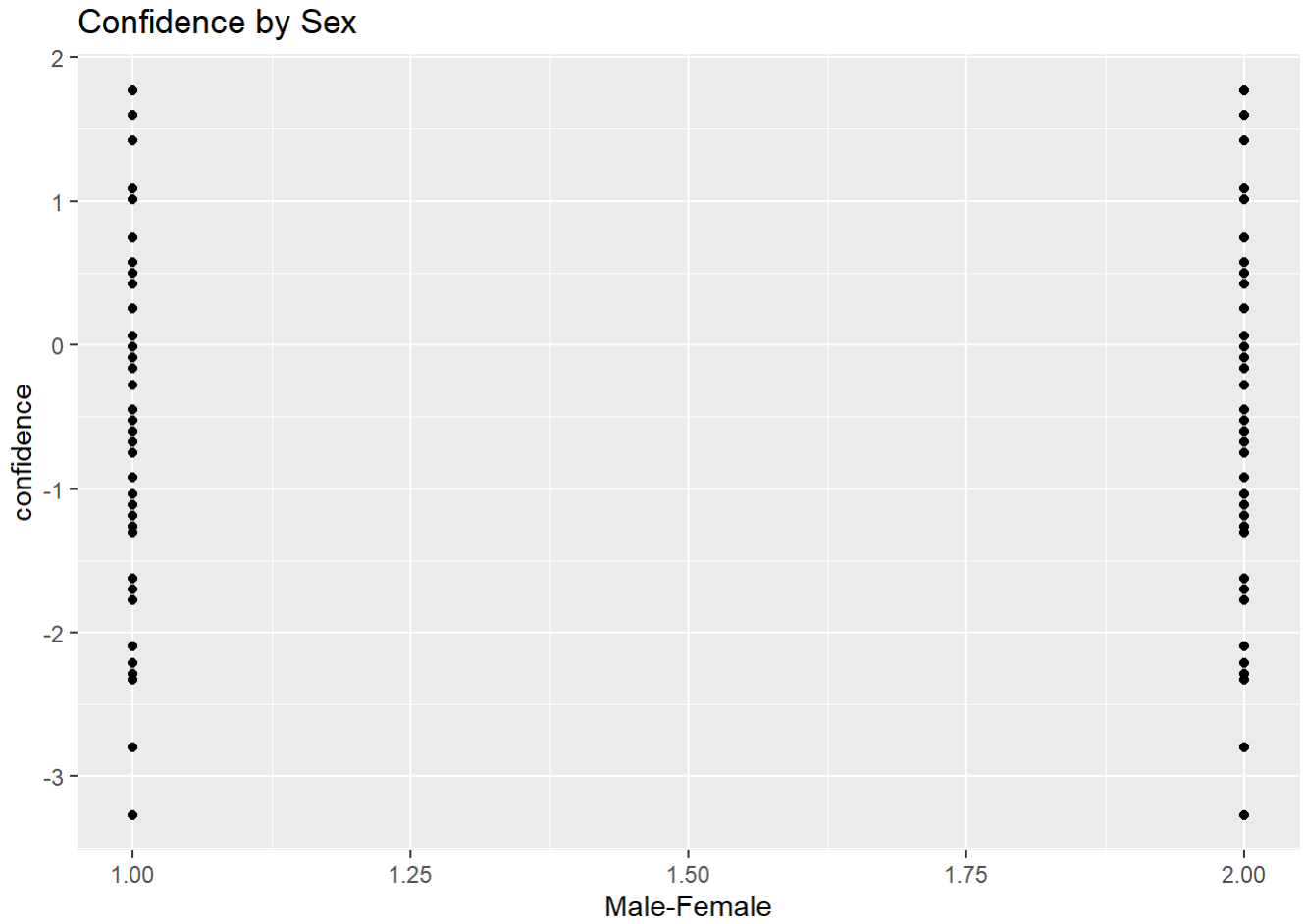

```
plot_comp
```

```
## Warning: Removed 20283 rows containing missing values (geom_point).
```




```
plot_conf
```

```
## Warning: Removed 4293 rows containing missing values (geom_point).
```



Conclusion

This project prepares me to analyze the data further with statistical methods in the future. For now, I reached several qualitative conclusions from the tables and graphs. First, male students generally score higher on conscientiousness, which is the most important trait for academic success. Female students' higher score on neuroticism is foudn to be negtively correlated with academic success. Moreover,

My key comparison is I confirmed the previous studies that have consistently found that neuroticism is the only trait females score higher on. Although I selected the variables that represent trait based on my previous expereience with psychological traits and the survey questions do not match perfectly well with the established psychological constructs, my study nonetheless confirm that there is indeed a difference in how males and females score differently in those traits.

In the future, I will include variables to be found in all other codebooks spanning from 1994 to 2006 becuase the survey is slightly different in each year. Also, I noticed that there is discrepency between the GPA variable in the codebook and the real dataset, so I need to rename the GPA variable. Then, I will collect how colleges' sex ratio vaied over time. I will finally conduct multiple linear regressions on traits
and sex ratio plus some rigorous tests. I included the hypotheses here:

a. Female college students score higher on agreeableness, higher on extraversion, higher on conscientiousness, lower on emotional stability, higher on openness, lower on confidence, lower on competitiveness, higher on fear of failure, and more internal locus of control.
b. Better academic success (i.e. higher GPA) is correlated with higher conscientiousness, higher confidence, higher competitiveness, lower fear of failure, and more external locus of control. The sign of the other four Big Five Inventories remains unknown due to contradicting scholarly findings.

c. Female students do better academically in colleges with more female students, while male students' performances are not affected by the gender composition.
d. As the ratio of female or male students increases, the increase in academic performance of female students with lower confidence and lower competitiveness will is larger than female students with higher confidence and higher competitiveness. Other traits and male students remain unknown and will be explored as well.