Data Evaluation for Final Project

STAT 244

Anya Weaver

Setup file

Load dataset of interest:

```
#commented out because I made a new CSV
#field = read_csv("Field.csv")
```

Rename CIPDESC to major for clarity:

```
#field = rename(field, Major = CIPDESC)
```

Asses contents of Dataset:

```
#head(field)
```

Sort by Major to asses options:

```
# commented out b/c it takes a while, makes a lot of data, and I don't need
 ↪  to do it again
#table(field$major)
```

Mutate Dataset to only include Biology, General and Biochemistry, Biophysics and Molecular Biology Majors & rename those majors to biology and biochem+:

```
#field = field  %>% mutate(Major = ifelse(Major == 'Biochemistry, Biophysics
 ↪  and Molecular Biology.', "biochem+", Major))

#field = field  %>% mutate(Major = ifelse(Major == 'Biology, General.',
 ↪  "biology", Major))

#bio <- field %>% filter((Major == 'biochem+') | (Major == 'biology'))

#bio <- bio %>% mutate(Major = as.factor(Major))
```

Turn new dataset into its own CSV so I don't have to load field again:

```
#commented out because I only need the one CSV
#write_csv(bio, "bio.csv")
#bio = read_csv("bio.csv")
```

Mutate Dataset to only include some variables:

```
#figure out what column names translate to the columns DEBTMEDIAN, DEBTMEAN,
 ↪  and MD_EARN_WNE. Figure out how gender comes into it.

# bio <- bio %>% select(Major, INSTNM, CONTROL, CREDDESC,
 ↪  DEBT_ALL_PP_EVAL_MDN, DEBT_ALL_STGP_EVAL_MDN, EARN_MDN_HI_1YR, DISTANCE,
 ↪  EARN_MDN_5YR)
# head(bio)
```

Filter out non responses:

```
# bio = bio %>% filter(!is.na(DEBT_ALL_PP_EVAL_MDN) & (DEBT_ALL_PP_EVAL_MDN
 ↪  !="PS"))
```

Add STGPand PP together to make one value for debt:

```
# bio$DEBT_ALL_PP_EVAL_MDN = as.numeric(bio$DEBT_ALL_PP_EVAL_MDN)
# bio$DEBT_ALL_STGP_EVAL_MDN = as.numeric(bio$DEBT_ALL_STGP_EVAL_MDN)
# bio$Debt_median <- bio$DEBT_ALL_PP_EVAL_MDN + bio$DEBT_ALL_STGP_EVAL_MDN
```

Rename included variables for clarity:

```
# bio = rename(bio, Institution = INSTNM)
# bio = rename(bio, School_type = CONTROL)
# bio = rename(bio, Degree = CREDDESC)
# bio = rename(bio, Distance = DISTANCE)
# bio = rename(bio, Earnings_median_1yr = EARN_MDN_HI_1YR)
# bio = rename(bio, Earnings_median_5yr = EARN_MDN_5YR)
```

Remove old columns:

```
# bio <- bio %>% select(Major, Institution,
#                            School_type, Degree, Debt_median, Distance,
 ↪  Earnings_median_1yr, Earnings_median_5yr)
# head(bio)
```

Filter out NAs:

```
# bio$Earnings_median_1yr = as.numeric(bio$Earnings_median_1yr)
# bio$Earnings_median_5yr = as.numeric(bio$Earnings_median_5yr)
```

Make sure quantitative variables are recorded as numbers:

```
# bio = bio %>% filter(!is.na(Earnings_median_1yr) & (Earnings_median_5yr
 ↪  !="PS"))
# bio$Earnings_median_1yr = as.numeric(bio$Earnings_median_1yr)
# bio$Earnings_median_5yr = as.numeric(bio$Earnings_median_5yr)
```

Read Bio:

```
# write_csv(bio, "bio.csv")
bio = read_csv("bio.csv")
```

Analyze Quantitative Variables:

```
mean(bio$Debt_median)
```

```
[1] 56783.04
```

```
sd(bio$Debt_median)
```

```
[1] 19378.69
```

```
mean(bio$Earnings_median_1yr)
```

```
[1] 27300.29
```

```
sd(bio$Earnings_median_1yr)
```

```
[1] 4747.215
```

```
mean(bio$Earnings_median_5yr)
```

```
[1] 62974.57
```

```
sd(bio$Earnings_median_5yr)
```

```
[1] 10156.37
```

Analyze Qualitative Variables:

```
count(bio$Institution)
```

```
n_Alabama A & M University
                         1
```

```
count(bio$School_type)
```

```
n_Private, for-profit
                    1
```

```
count(bio$Degree)
```

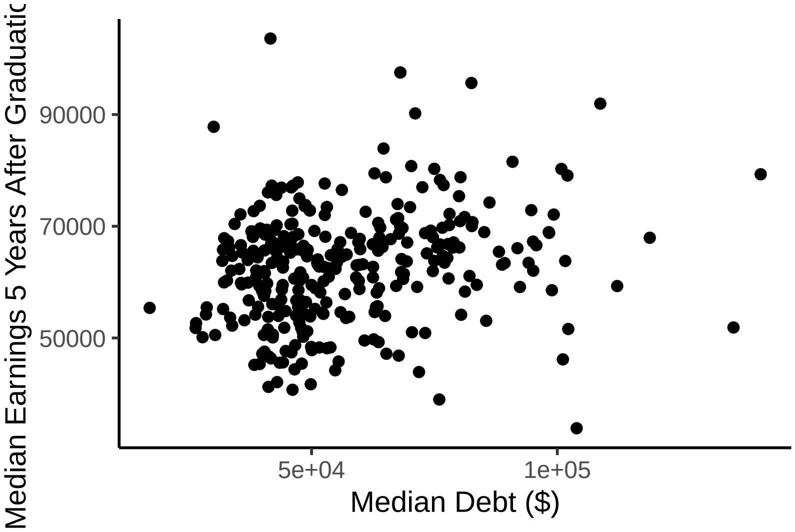
```
n_Bachelor's Degree
                329
```

```
count(bio$Distance)
```

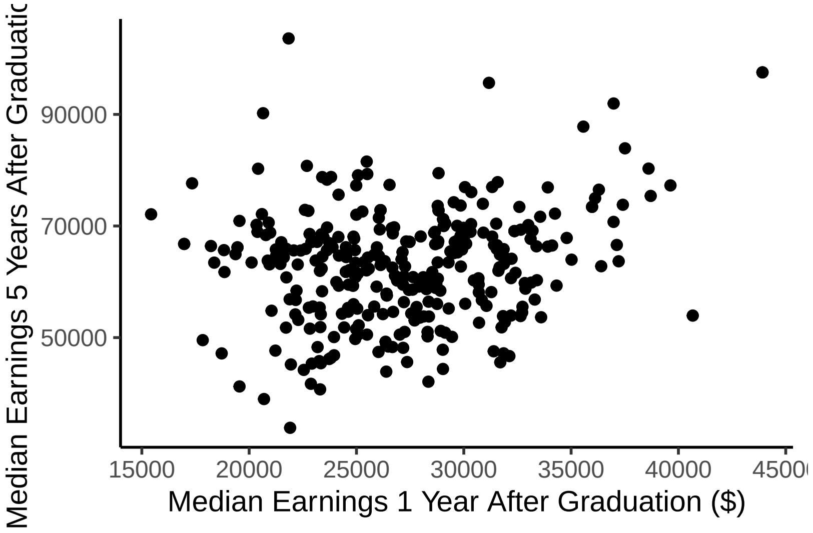
```
n_1
318
```

Produce data visualizations

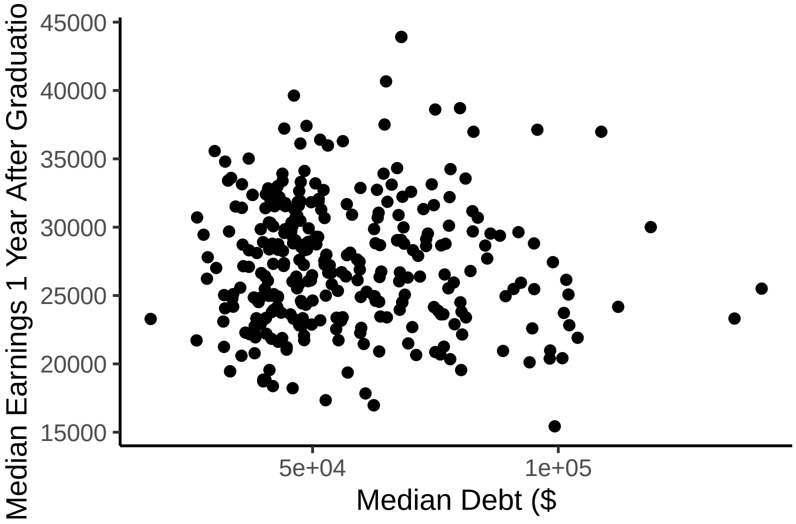
```
bio %>%
ggplot(aes(x = Debt_median, y = Earnings_median_5yr)) +
geom_point() +
labs(x = 'Median Debt ($)', y = 'Median Earnings 5 Years After Graduation
 ↪  ($)') +
theme_classic()
```



```
bio %>%
ggplot(aes(x = Earnings_median_1yr, y = Earnings_median_5yr)) +
geom_point() +
labs(x = 'Median Earnings 1 Year After Graduation ($)', y = 'Median Earnings
 ↪  5 Years After Graduation ($)') +
theme_classic()
```



```
bio %>%
ggplot(aes(x = Debt_median, y = Earnings_median_1yr)) +
geom_point() +
labs(x = 'Median Debt ($', y = 'Median Earnings 1 Year After Graduation ($)')
 ↪  +
theme_classic()
```



Histograms for dataset: