DDD Final Project: Higher Education IPUMS Data

My Data

The data used for this project is taken from the following IPUM's Higher Education Surveys (https://highered.ipums.org/highered/): the National Survey of College Graduates (NSCG), Survey of Doctorate Recipients (SDR), and the National Survey of Recent College Graduates (NSRCG). The variables extracted are related to demographic information, education and family information, and citizen satisfaction.

I chose this data to dive deeper into how demographics play a role in higher education, as well as how strongly degree level affects satisfaction in later life. The variables I chose in IPUMS enable me to compare data between different demographics to see potential correlations.

This data is from 2013, which is the most recent data collected from the above surveys. Formal Citation: Minnesota Population Center. IPUMS Higher Ed: Version 1.0 [dataset]. Minneapolis, MN: University of Minnesota, 2016. https://doi.org/10.18128/D100.V1.0 (https://doi.org/10.18128/D100.V1.0)

Downloading Data

Data was downloaded from IPUMS as a zipped CSV file. Data was downloaded on 1/5/2021.

```
library (readr)
originalData <- read_csv("/Users/divyanair/Desktop/DDD-I21/highered_00001.csv.gz")
```

```
##
## ── Column specification ────────────────────────────────────────────────────────
## cols(
##   PERSONID = col_double(),
##   YEAR = col_double(),
##   WEIGHT = col_double(),
##   SAMPLE = col_double(),
##   SURID = col_double(),
##   AGE = col_double(),
##   GENDER = col_double(),
##   RACETH = col_double(),
##   CTZUSIN = col_double(),
##   CHTOT = col_double(),
##   DGRDG = col_double(),
##   JOBSATIS = col_double(),
##   SATSAL = col_double(),
##   SATSOC = col_double()
## )
```

Questions this data can answer

Does level of education impact job satisfaction?

Are there correlations between job satisfaction, salary satisfaction, and satisfaction with how work contributes to society?

Are there racial and gender imbalances in higher education?

Data Key

All data from IPUMS automatically came in the form of numeric categorical variables. Though most variables have sensible categories based on the key below, the race/ ethnicity data is lacking in specificity. Since no further breakdown could be found for the race data, the below categories were used for the analysis. The specific questions asked for the satisfaction questions are:

- Are you satisfied with your job?
- Are you satisfied with your salary?
- Are you satisfied with your job's contribution to society?

Responses were given on a scale from 1 to 4, with 1 being most satisfied.
Key:

- Gender: 1 = Female, 2 = Male,
- Race/Ethnicity: 1 = Asian, 2 = White, 3 = Under-represented minority, 4 = other,
- Citizenship: 0 = yes, 1 = no,
- Degree: 1 = BA, 2 = MA, 3 = Doctorate, 4 = Professional,
- All Satisfaction Columns: 1 = very satisfied, 2 = satisfied, 3 = dissatisfied, 4 = very dissatisfied.

Data Cleaning Steps

1. Determining number of variables and observations in original data

```
print(dim(originalData))
```

```
## [1] 115152     14
```

There are 115,152 observations and 14 variables in the original dataframe.

2. Rename column names

```
highEdData <- originalData

#reassigns column names
colnames(highEdData) <- c("personID","year","studyWeight","sampleID","surveyID","age","g
ender","race","citizenStatus","numberOfChildren","highestDegree","jobSatisfaction","sala
rySatisfaction","socialSatisfaction")

print (colnames(highEdData))
```

```
##  [1] "personID"           "year"               "studyWeight"
##  [4] "sampleID"           "surveyID"           "age"
##  [7] "gender"             "race"               "citizenStatus"
## [10] "numberOfChildren"   "highestDegree"      "jobSatisfaction"
## [13] "salarySatisfaction" "socialSatisfaction"
```

3. Determine class and type of each column

```
colClass <- lapply (highEdData, class) #gets class for all columns
colType <- lapply (highEdData,typeof) #gets class for all columns

#binds the class and type into one dataframe
classAndType <- do.call(rbind, Map(data.frame, classOfData=colClass, typeOfData=colTyp
e))

print(classAndType)
```

```
##                   classOfData typeOfData
## personID              numeric     double
## year                  numeric     double
## studyWeight           numeric     double
## sampleID              numeric     double
## surveyID              numeric     double
## age                   numeric     double
## gender                numeric     double
## race                  numeric     double
## citizenStatus         numeric     double
## numberOfChildren      numeric     double
## highestDegree         numeric     double
## jobSatisfaction       numeric     double
## salarySatisfaction    numeric     double
## socialSatisfaction    numeric     double
```

4. Assigning missing data cells to NA. Missing data is denoted with 98, or a logical skip by the participant.

```
#assigns 98 columns to NA value
highEdData[highEdData == 98] <- NA
```

5. Create binary flags for missing data.

```
highEdData$naChildren <- ifelse(is.na(highEdData$numberOfChildren) == TRUE,1,0) #gives a
1 for people who didn't respond/have no children
highEdData$naJobSat <- ifelse(is.na(highEdData$jobSatisfaction) == TRUE,1,0) #gives a 1
 for people who didn't respond
highEdData$naCitizenship <- ifelse(is.na(highEdData$citizenStatus) == TRUE,1,0) #gives a
1 for people who didn't respond/ may be undocumented
highEdData$naSalarySat <- ifelse(is.na(highEdData$salarySatisfaction) == TRUE,1,0) #give
s a 1 for people who didn't respond
highEdData$naSocialSat <- ifelse(is.na(highEdData$socialSatisfaction) == TRUE,1,0) #give
s a 1 for people who didn't respond
```

6. Removing observations with no response for certain variables.

```
#Removes all rows with 1 (or no response) for each of the following columns
cleanedData <- highEdData[!(highEdData$naChildren == 1 |
                              highEdData$naJobSat == 1 |
                              highEdData$naCitizenship == 1 |
                              highEdData$naSalarySat == 1 |
                              highEdData$naSocialSat == 1),]

#Determining how many observations were dropped
print(length(rownames(originalData))-length(rownames(cleanedData)))
```

```
## [1] 71823
```

71,823 rows were dropped. Since this is a significant amount, the fully cleaned dataset will be used as sparingly as possible. Instead, dataframes of relevant extracted variables will be cleaned separately to include as much data as possible to draw more accurate conclusions.

Using data to determine relationships between variables

Since the necessary variables for the following two questions (gender, highest level of education, and race) had no missing values for observations, the non-cleaned data was used to incorporate as much data as possible.

How does gender impact higher education?

Data Used: gender and highestDegree columns with all observations from original data.

```
#extracting gender and highest degree data from overall dataframe

genderAndDegree <- highEdData %>%
  dplyr::select(gender , highestDegree) %>%
  dplyr::arrange(gender)

femaleDegree <- genderAndDegree[genderAndDegree$gender == 1, ] #subsects gender data for
only females
avgFemDegree <- mean(femaleDegree$highestDegree)

maleDegree <- genderAndDegree[genderAndDegree$gender == 2, ] #subsects gender data for o
nly males
avgMaleDegree <- mean(maleDegree$highestDegree)

#creating new dataframes with degree breakdown by gender
femDegreeBreakdown <- femaleDegree %>%
  dplyr::group_by(gender,highestDegree) %>%
  dplyr::summarise(breakdown = n()) %>%
  dplyr::mutate(percentDegree = (breakdown / sum(breakdown)*100))

maleDegreeBreakdown <- maleDegree %>%
  dplyr::group_by(gender,highestDegree) %>%
  dplyr::summarise(breakdown = n()) %>%
  dplyr::mutate(percentDegree = (breakdown / sum(breakdown)*100))
```

Figure 1.1: Degree Breakdown by Gender

Source: IPUMS 2013 Higher Education Surveys

	Female		Male	
Degree Name	# Degree	% Degree	# Degree	% Degree
Bachelors	18930	37.921433	25269	38.736529
Masters	16756	33.566378	16585	25.424248
Doctorate	12396	24.832228	21014	32.213757
Professional	1837	3.679962	2365	3.625466

The table above shows that on average, males have obtained a higher degree level than females. Males and females have similar percentage of professional degrees, males have a significantly higher percentage of doctorate Degrees, females have significantly higher percentage of masters, and both sexes have a relatively similar percentage of BAs. On average, females have a degree of 1.942707. Rounded to 2, females have a masters degree as their highest degree on average. Males have an overall average degree of 2.007282, which is higher than that of females. Still rounded to 2, males have a masters degree as highest degree on average.

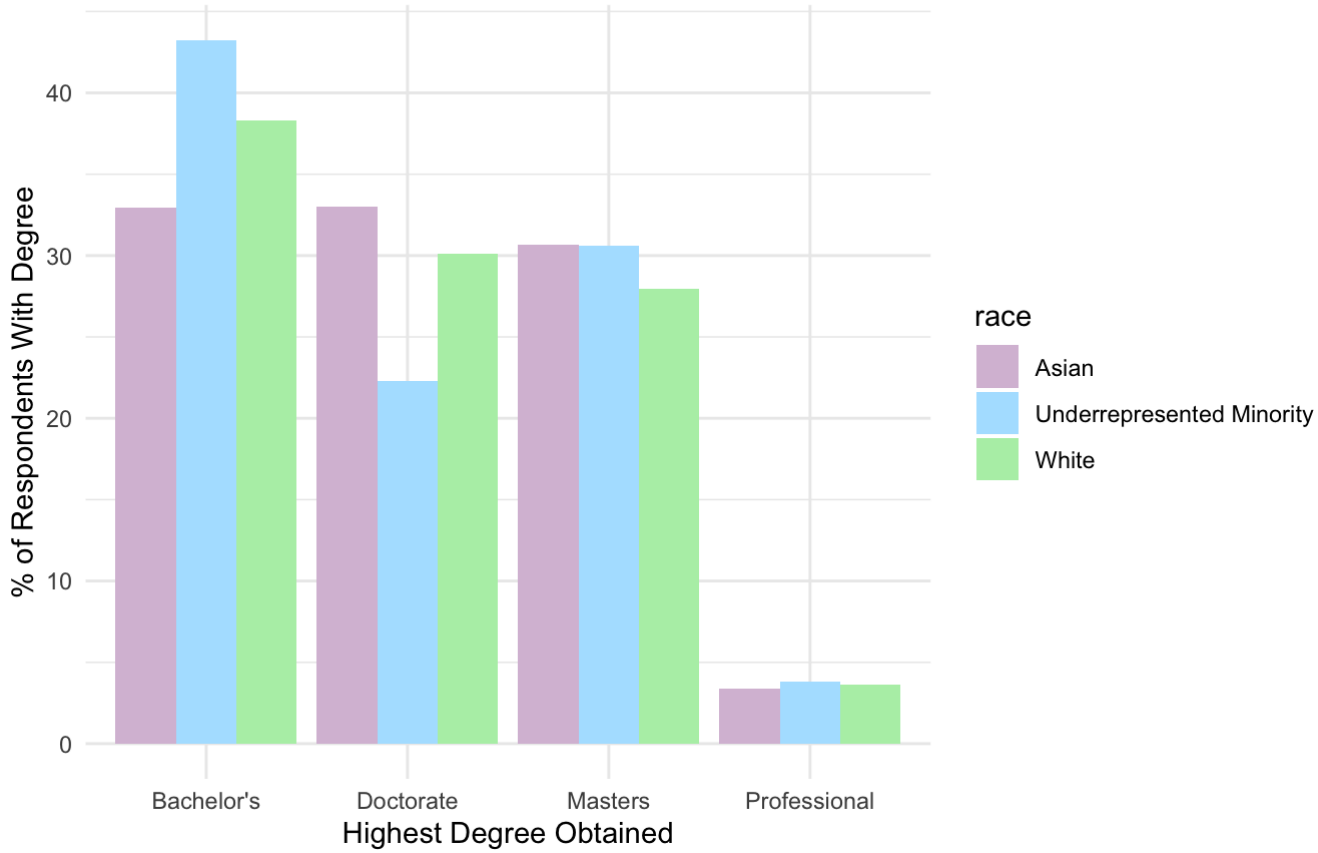
How does race impact higher education?

Data used: race and highestDegree columns with all observations from original data.

```
#Race and Highest Degree
raceAndDegree <- highEdData %>%
  dplyr::select(race , highestDegree) %>%
  dplyr::arrange(race)

#Creates new dataframe with only data for Asians
asianDegree <- raceAndDegree[raceAndDegree$race == 1, ]
avgAsianDegree <- mean(asianDegree$highestDegree)

#Creates new dataframe with only data for White respondents
whiteDegree <- raceAndDegree[raceAndDegree$race == 2, ]
avgWhiteDegree <- mean(whiteDegree$highestDegree)

#Creates new dataframe with only data for respondents who are an underrepresented minori
ty
underRepMinDegree <- raceAndDegree[raceAndDegree$race == 3, ]
avgUnderRepMinDegree <- mean(underRepMinDegree$highestDegree)

#Creating dataframes with summaries of how many/ percentage of individuals obtained whic
h degree by race:
asianBreakdown <- asianDegree %>%
  dplyr::group_by(race,highestDegree) %>%
  dplyr::summarise(breakdown = n()) %>%
  dplyr::mutate(percentDegree = (breakdown / sum(breakdown)*100))

whiteBreakdown <- whiteDegree %>%
  dplyr::group_by(race,highestDegree) %>%
  dplyr::summarise(breakdown = n()) %>%
  dplyr::mutate(percentDegree = (breakdown / sum(breakdown)*100))

underRepMinBreakdown <- underRepMinDegree %>%
  dplyr::group_by(race,highestDegree) %>%
  dplyr::summarise(breakdown = n()) %>%
  dplyr::mutate(percentDegree = (breakdown / sum(breakdown)*100))
```

On average, Asian respondents have a degree of 2.069207. White respondents have a degree of 1.991291 on average, and respondents of an underrepresented minority have an average highest degree of 1.86711. While Asians have the highest degree on average, all races' average degree is rounded to 2, or a masters degree.

Figure 1.2: Breakdown of Degree by Race
Source: IPUMS 2013 Higher Education Surveys



When looking at the graphs above, the breakdown by race for highest degree varies. For Asian respondents, the most common highest degree was a doctorate. For white respondents and those of an underrepresented minority, this was a bachelor's degree. All three groups had comparable percentages of those with professional degrees. Looking at each degree separately, those of an underrepresented minority have the highest percentage of respondents who had a bachelor's degree as their highest degree. Asian respondents had the highest percentage of doctorates. White respondents and those of an underrepresented minority both had the highest percent of those with a master's degree, and underrepresented minorities had the highest level of professional degrees.

How does highest degree impact satisfaction?

Data used: highestDegree, jobSatisfaction, salarySatisfaction, and socialSatisfaction columns for observations that had answers to all three satisfaction questions. 17,101 observations were dropped.

```
# Cleans the full dataframe with all observations to only include highestDegree, jobSati
sfaction, salarySatisfaction, and socialSatisfaction columns. Removes observations with
 at least one NA to the three questions.

degAndSat <- highEdData %>%
  dplyr::select(
    highestDegree, jobSatisfaction, salarySatisfaction, socialSatisfaction, naJobSat, na
SalarySat, naSocialSat) %>%
  dplyr::arrange(highestDegree)

degAndSatCleaned <- degAndSat[!(degAndSat$naJobSat == 1 |
                                degAndSat$naSalarySat == 1 |
                                degAndSat$naSocialSat == 1),]

#gets the averages of all satisfaction responses
degAndSatCleaned$avgSat <- ((degAndSatCleaned$jobSatisfaction + degAndSatCleaned$salaryS
atisfaction + degAndSatCleaned$socialSatisfaction)/3)

#Creates a dataframe with all satisfaction data for those with bachelor's degrees.
BASat <- degAndSatCleaned[degAndSatCleaned$highestDegree == 1, ]
#Calculates average satisfaction amongst those with BAs.
BAAvgSat <- mean(BASat$avgSat)

#Creates a dataframe with all satisfaction data for those with master's degrees.
MaSat <- degAndSatCleaned[degAndSatCleaned$highestDegree == 2, ]
#Calculates average satisfaction amongst those with MAs.
MaAvgSat <- mean(MaSat$avgSat)

#Creates a dataframe with all satisfaction data for those with doctorate degrees.
DocSat <- degAndSatCleaned[degAndSatCleaned$highestDegree == 3, ]
#Calculates average satisfaction amongst those with doctorate degrees.
DocAvgSat <- mean(DocSat$avgSat)

#Creates a dataframe with all satisfaction data for those with professional degrees.
ProfSat <- degAndSatCleaned[degAndSatCleaned$highestDegree == 4, ]
#Calculates average satisfaction amongst those with professional degrees.
ProfAvgSat <- mean(ProfSat$avgSat)
```

When looking at the average satisfaction scores across education level, respondents with bachelor's degrees were the most satisfied, with an average score of 1.84. Those with masters had an average score of 1.77, and those with doctorate degrees had an average score of 1.72. Respondents with professional degrees had the lowest average satisfaction score of 1.63. Though the average scores varied numerically, all the scores rounded to 2 or "Satisfied."

Figure 1.3: Bachelor's Satisfaction Breakdown

Source: IPUMS 2013 Higher Education Surveys

	Bachelors					
	# of	% of	# of	% of	# of Respondents	% of Responden

Figure 1.3: Bachelor's Satisfaction Breakdown

Source: IPUMS 2013 Higher Education Surveys

Satisfaction	# of Respondents Satisfied with Job	% of Respondents Satisfied with Job	# of Respondents Satisfied with Salary	% of Respondents Satisfied with Salary	# of Respondents Satisfied with Social Contribution	% of Respondents Satisfied wi Social Contributio
1	14903	40.736387	10200	27.881041	16039	43.84157
2	17348	47.419637	18050	49.338509	14810	40.48217
3	3309	9.044938	5660	15.471244	4005	10.94740
4	1024	2.799038	2674	7.309206	1730	4.72884

Figure 1.4: Master's Satisfaction Breakdown

Source: IPUMS 2013 Higher Education Surveys

Master's					
# of Respondents Satisfied with Job	% of Respondents Satisfied with Job	# of Respondents Satisfied with Salary	% of Respondents Satisfied with Salary	# of Respondents Satisfied with Social Contribution	% of Respondents Satisfied with Social Contribution
12524	43.604206	8172	28.452058	14994	52.203886
13222	46.034399	14066	48.972913	10345	36.017687
2383	8.296776	4432	15.430680	2492	8.676276
593	2.064619	2052	7.144349	891	3.102152

Figure 1.5: Doctorate Satisfaction Breakdown

Source: IPUMS 2013 Higher Education Surveys

Doctorate					
# of Respondents Satisfied with Job	% of Respondents Satisfied with Job	# of Respondents Satisfied with Salary	% of Respondents Satisfied with Salary	# of Respondents Satisfied with Social Contribution	% of Respondents Satisfied with Social Contribution
13712	47.165658	8724	30.008255	16008	55.063291
12593	43.316593	13849	47.636901	10579	36.388965
2246	7.725647	4549	15.647358	1994	6.858833
521	1.792102	1950	6.707485	491	1.688910

Figure 1.6: Professional Degree Satisfaction Breakdown

Source: IPUMS 2013 Higher Education Surveys

Professional Degree					
# of Respondents Satisfied with Job	% of Respondents Satisfied with Job	# of Respondents Satisfied with Salary	% of Respondents Satisfied with Salary	# of Respondents Satisfied with Social Contribution	% of Respondents Satisfied with Social Contribution
2008	54.669208	1302	35.447863	2439	66.403485
1345	36.618568	1671	45.494146	907	24.693711
240	6.534168	463	12.605500	230	6.261911
80	2.178056	237	6.452491	97	2.640893

Though the average satisfaction score places those with BAs to be most satisfied, examining the satisfaction breakdown by type (job, salary, social) is more realistic. The table above shows the number and percentage of respondents and their satisfaction level by highest degree obtained. Scrolling to the right, the number of respondents and their satisfaction levels by education level can be observed. Those with professional degrees had the highest percentage of individuals “Very Satisfied” with their job, and those with bachelor’s degrees had the least. A similar trend is seen for salary satisfaction, with respondents with professional degrees having the highest satisfaction and bachelor degree recipients being the least. The same trend can be seen for social satisfaction.

Conclusions

In conclusion, the results from this data analysis aligned with previous knowledge on how socioeconomic status factors into education level and career satisfaction. Females had a slightly lower highest degree on average, which is to be expected. Similarly, minority races had the lowest degree on average as well. This shows that there are disparities in race and gender within higher education. Interestingly, degree had a clear effect on salary, job, and social satisfaction as well with professional degrees having the most satisfied respondents for each question. It is important to note that no test for statistical significance has been done for this analysis, so the differences observed may be due to chance and randomness in the dataset.