Final Project Step 2

Logan Quandt

2/19/2022

Libraries

```
library(ggplot2)
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(readxl)
library(purrr)
```

Loading and Cleaning Data

For my data there are multiple different data sources that I am using to analyse multiple different facets of income inequity. As I did not find it practical to try and combine the datasets I have loaded them all in individually. For the census and Glassdoor data I also narrowed down the columns I will be using in my project to eliminate some unnecessary columns in the datasets. I ended up having to edit some datasets (Personal Inc (pinc) data) outside of R to convert them from Xlsx to Csv format to fix an error in the column names.

```
setwd("/Users/logan/Documents/GitHub/DSC520LQ/Final Project")

#Dataset one
census_df <- read.csv("data/census.csv")
head(census_df)
```

```
##   age         workclass  education       maritalstatus        occupation
## 1  39         State-gov  Bachelors       Never-married      Adm-clerical
## 2  50  Self-emp-not-inc  Bachelors  Married-civ-spouse   Exec-managerial
## 3  38           Private    HS-grad            Divorced Handlers-cleaners
## 4  53           Private       11th  Married-civ-spouse Handlers-cleaners
## 5  28           Private  Bachelors  Married-civ-spouse    Prof-specialty
## 6  37           Private    Masters  Married-civ-spouse   Exec-managerial
##    relationship   race     sex capitalgain capitalloss hoursperweek
## 1 Not-in-family  White    Male        2174           0           40
## 2       Husband  White    Male           0           0           13
```

```
## 3  Not-in-family  White    Male           0           0           40
## 4        Husband  Black    Male           0           0           40
## 5           Wife  Black  Female           0           0           40
## 6           Wife  White  Female           0           0           40
##    nativecountry over50k
## 1  United-States   <=50K
## 2  United-States   <=50K
## 3  United-States   <=50K
## 4  United-States   <=50K
## 5           Cuba   <=50K
## 6  United-States   <=50K
```

```
census_df_final <- select(census_df, age, sex, race, education, occupation, over50k)

#Dataset two
pinc_df_whites <- read.csv('data/pinc_white_test_csv.csv')

pinc_df_afr_amer <- read.csv('data/pinc_african_amer_test_csv.csv')

#Dataset three
glassdoor_pay_df <- read.csv('data/Glassdoor Gender Pay Gap.csv')
head(glassdoor_pay_df)
```

```
##             JobTitle Gender Age PerfEval Education           Dept Seniority
## 1    Graphic Designer Female  18        5   College     Operations         2
## 2   Software Engineer   Male  21        5   College     Management         5
## 3 Warehouse Associate Female  19        4       PhD Administration         5
## 4   Software Engineer   Male  20        5   Masters          Sales         4
## 5    Graphic Designer   Male  26        5   Masters    Engineering         5
## 6                  IT Female  20        5       PhD     Operations         4
##   BasePay Bonus
## 1   42363  9938
## 2  108476 11128
## 3   90208  9268
## 4  108080 10154
## 5   99464  9319
## 6   70890 10126
```

```
glassdoor_pay_df_final <- select(glassdoor_pay_df, JobTitle, Gender, Age, Education,Seniority, BasePay,

#Dataset four
college_df_wages <- read.csv('data/wages.csv')
college_df_unemployment <- read.csv('data/Unemployment_rate.csv')
college_df_underemployment <- read.csv('data/under_employment_college_grads.csv')
college_df_labor_market <- read.csv('data/labor_market_college_grads.csv')
```

What does the final dataset look like.

As I am uncertain on the practicality of combining this many datasets I currently have 8 different dataframes that I have created. This seems to be alot but I will be using them to analyze different things. I will use the Glassdoor data set to analyze Gender pay inequity. I plan on using the Personal Income (pinc) datasets to analyze some racial inequity. The additional college and census dataframes will be used as secondary datasets to help analyze causes along with age inequity.

##What Information is not self evident

It isn't entirely clear if the Glassdoor pay data is from just the US seeing as Glassdoor maintains data for companies worldwide. The sources for all the other data made it clear that it is from the US but I am going to work under the assumption that the Glassdoor data on Gender inequity is applicable to the US as well as it is a strong debate in this country. The census data also only shows if someone makes under or over 50k which is a different measurement then the other dataframes I am using.–>

##What are different ways you can look at the data

Different ways I can look at the data are to compare different variables such as race to income, age to income, gender to income. I could then verify the significane of other variables such as education to income, education between different genders/races. Experience, Regions and Occupation are other areas I could look at for comparison.

##How do you plan to slice and dice the data

I plan on slicing the data by the following:

1. Age
2. Race
3. Gender
4. Income
5. Education
6. Experience
7. Occupation

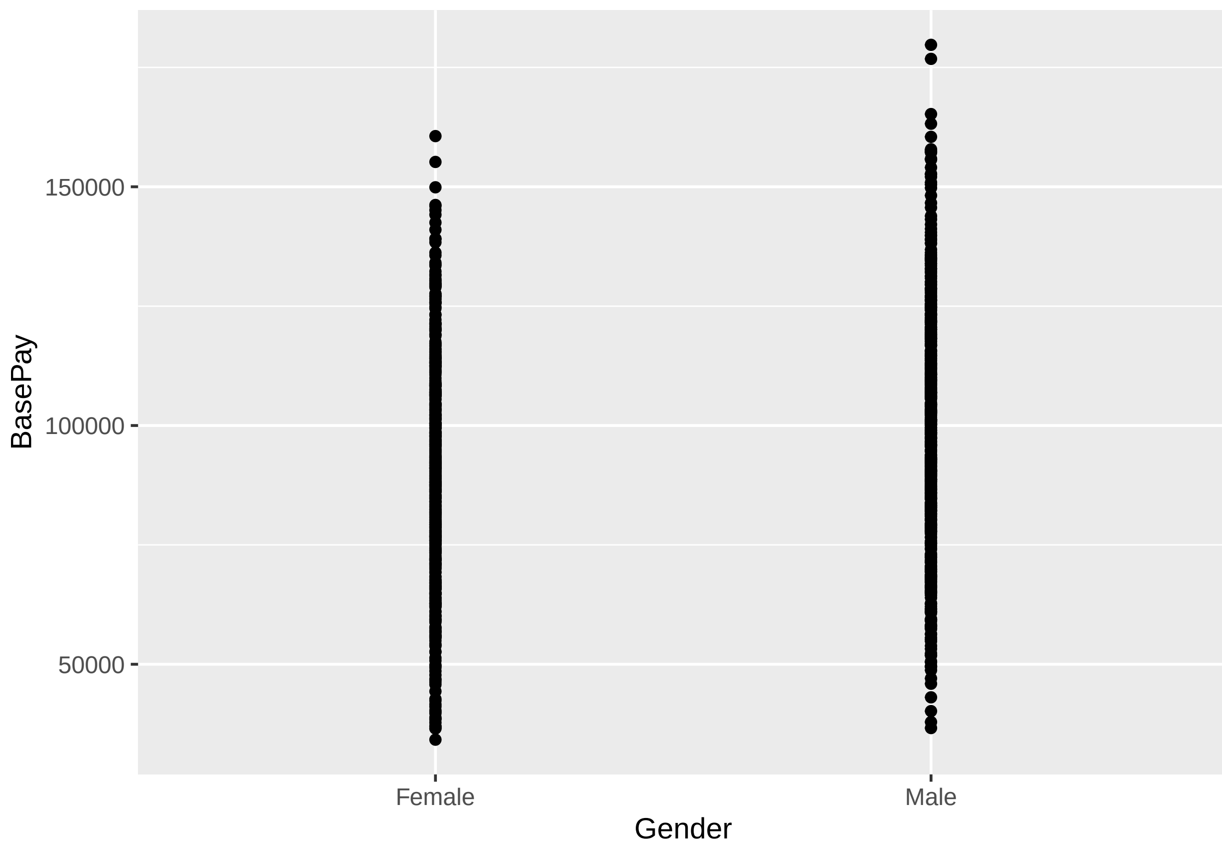
##How could you summarize your data to answer key questions

Mostly I plan on looking at the relationships between the variables such as the correlation and covariance. I plan on looking at plots and tables to help visualize my data. Descriptive statistics such as mean, median, mode, standard dev and variance will also be very useful. Summary statistics such as ranges can be helpful in my analysis as well.

##What plots and tables will help you illustrate your findings

Scatterplots will be very useful in helping to explore the relationships between variables. Histograms can help me check for normality and Box plots can help me identify any outliers. Tables may also be useful in understanding some variables such as median income by occupation.

```r
#As an example here is a scatterplot of Gender vs Income
ggplot(glassdoor_pay_df_final, aes(x=Gender, y=BasePay)) + geom_point()
```



Do you plan on incorporating any machine learning techniques to answer your questions.

I do not plan on incorporating any ML techinques to answer my questions. If I gain more experience in using these techniques I may incorporate them into future projects.

##Questions for Future Steps

A question I have on future steps is how I can help narrow down the Pinc datasets that I have created. It has over 50 columns but I believe they all provide good data on income inequity by showing different levels of income. I may end up removing the lower and upper income outliers to help condense those datasets