DATA 606 Data Project Proposal

David Simbandumwe

Research Question

You should phrase your research question in a way that matches up with the scope of inference your data set allows for.

I choose to study the results of the Financial well-being survey. The survey is part of the ongoing research from the Consumer Financial Protection Bureau. It was focused on understanding the factors that support consumer financial well-being in an effort to assist practitioners and policymakers empower more families to lead better financial lives to serve their own goals.

The research suggests that the following factors will influence adults well being:

- Income and employment
- Savings and safety nets
- Past financial experiences
- Financial behaviors, skills, and attitudes

The CFPB Financial Well-Being Scale contains the following 10 questions:

How well does this statement describe you or your situation?

1. I could handle a major unexpected expense
2. I am securing my financial future
3. Because of my money situation, I feel like I will never have the things I want in life*
4. I can enjoy life because of the way I'm managing my money
5. I am just getting by financially*
6. I am concerned that the money I have or will save won't last*

How often does this statement apply to you?

7. Giving a gift for a wedding, birthday or other occasion would put a strain on my finances for the month*
8. I have money left over at the end of the month
9. I am behind with my finances*
10. My finances control my life*

I planned to approach this project using the following three lines of inquiry:

1. Explore the relevance of the findings to the general population. Review the key financial wellness measure identified in the study and analyze their relevance to the general population. Using the t distribution I will look at the sample statistics for the FWBScore, LMScore and KJScore then calculate the corresponding confidence intervals for the population.

- FWBscore - Financial well-being scale score4 - IRT score
- FSscore - Financial skill scale score IRT score
- LMscore - Lusardi and Mitchell financial knowledge scale score - Summative scale score
- KHscore - Knoll and Houts financial knowledge scale score - IRT score

2. Impact of race, and gender on financial well being. Race and Gender are all categorical variables in the data set.

    - PPETHM - Race / Ethnicity - is there a statistical difference between the financial wellness of white non hispanics and the other ethnic groups surveyed.
    - PPGENDER - Gender - is there a statistical difference between financial wellness scores for male or female

Race

```
- H0 - being "White, Non Hispanic" does not impact the financial well being score (FWBscore) for an indi
- H1 - being "White, Non Hispanic" does impact the financial well being score (FWBscore) for an individu
```

Gender

```
- H0 - being a "Male" does not impact the financial well being score (FWBscore) for an individual in the
- H1 - being a "Male" does impact the financial well being score (FWBscore) for an individual in the pop
```

3. Impact of intersectionality on financial well being. Does race and gender have an impact on financial well being?

Intersectionality

```
- H0 - being a "White, Non Hispanic" "Male" does not impact the financial well being score (FWBscore) fo
- H1 - being a "White, Non Hispanic" "Male" does impact the financial well being score (FWBscore) for an
```

Cases

What are the cases, and how many are there?

The cases are the individual survey responses from 6394 survey participants.

Data Collection

Describe the method of data collection.

The data was collected as part of the Consumer Financial Protection Bureau's (CFPB) National Financial Well-Being Survey Public Use File (PUF). The PUF is a dataset containing

(1) data collected in the National Financial Well-Being Survey,
(2) data about members of the GfK KnowledgePanel collected prior to the survey, and
(3) data on poverty levels in respondents' counties of residence.

The National Financial Well-Being Survey was conducted in English and Spanish via web mode between October 27, 2016 and December 5, 2016. Overall, 6,394 surveys were completed: 5,395 from the general population sample and 999 from an oversample of adults aged 62 and older. The survey was designed to represent the adult population of the 50 U.S. states and the District of Columbia. The survey was fielded on the GfK KnowledgePanel®. The KnowledgePanel sample is recruited using address-based sampling and dual-frame landline and cell phone random digit dialing methods.

The PUF was published in 2017.

Data Preparation

Select Key Variables

```
# summary of data
dim(wellbeing_df)
```

```
## [1] 6394  217
```

```
df <- wellbeing_df %>%
    select(sample, fpl,
            FWBscore,FSscore,LMscore,KHscore,
           LIVINGARRANGEMENT,EARNERS, SAVINGSRANGES,
           HOUSING,VALUERANGES,MORTGAGE,
           agecat,generation,PPEDUC,PPETHM,PPGENDER,PPINCIMP,
           PPHHSIZE,PPMARIT,PPMSACAT,PPREG4,PPREG9
          )
```

Tiddy Factor Data

```
df$sample = revalue(factor(df$sample), c(
  `1` = "General population",
  `2` = "Age 62+ oversample",
  `3` = "Race/ethnicity and poverty oversample"
))
df$fpl = revalue(factor(df$fpl), c(
  `1` = "<100% FPL",
  `2` = "100%-199% FPL",
  `3` = "200%+ FPL"
))
df$LIVINGARRANGEMENT = revalue(factor(df$LIVINGARRANGEMENT), c(
  `-1` = "Refused",
  `1` = "I am the only adult in the household",
  `2` = "I live with my spouse/partner/significant other",
  `3` = "I live in my parents' home",
  `4` = "I live with other family, friends, or roommates",
  `5` = "Some other arrangement"
))
df$EARNERS = revalue(factor(df$EARNERS), c(
  `-1` = "Refused",
```

```
  `1` = "One",
  `2` = "Two",
  `3` = "More than two"
))
df$SAVINGSRANGES = revalue(factor(df$SAVINGSRANGES), c(
  `-1` = "Refused",
  `1` = "0",
  `2` = "$1-99",
  `3` = "$100-999",
  `4` = "$1,000-4,999",
  `5` = "$5,000-19,999",
  `6` = "$20,000-74,999",
  `7` = "$75,000 or more",
  `98` = "I don't know",
  `99` = "Prefer not to say"
))
df$HOUSING = revalue(factor(df$HOUSING), c(
  `-1` = "Refused",
  `1` = "I own my home",
  `2` = "I rent",
  `3` = "I do not currently own or rent"
))
df$VALUERANGES = revalue(factor(df$VALUERANGES), c(
  `-2` = "Question not asked because respondent not in item base",
  `-1` = "Refused",
  `1` = "Less than $150,000",
  `2` = "$150,000-249,999",
  `3` = "$250,000-399,999",
  `4` = "$400,000 or more",
  `98` = "I don't know",
  `99` = "Prefer not to say"
))
df$MORTGAGE = revalue(factor(df$MORTGAGE), c(
  `-2` = "Question not asked because respondent not in item base",
  `-1` = "Refused",
  `1` = "Less than $50,000",
  `2` = "$50,000-199,999",
  `3` = "$200,000 or more",
  `98` = "I don't know",
  `99` = "Prefer not to say"
))
df$SAVINGSRANGES = revalue(factor(df$SAVINGSRANGES), c(
  `-1` = "Refused",
  `1` = "0",
  `2` = "$1-99",
  `3` = "$100-999",
  `4` = "$1,000-4,999",
  `5` = "$5,000-19,999",
  `6` = "$20,000-74,999",
  `7` = "$75,000 or more",
  `98` = "I don't know",
  `99` = "Prefer not to say"
))
```

```
df$agecat = revalue(factor(df$agecat), c(
  `1` = "18-24",
  `2` = "25-34",
  `3` = "35-44",
  `4` = "45-54",
  `5` = "55-61",
  `6` = "62-69",
  `7` = "70-74",
  `8` = "75+"
))
df$generation = revalue(factor(df$generation), c(
  `1` = "Pre-Boomer",
  `2` = "Boomer",
  `3` = "Gen X",
  `4` = "Millennial"
))
df$PPEDUC = revalue(factor(df$PPEDUC), c(
  `1` = "Less than high school",
  `2` = "High school degree/GED",
  `3` = "Some college/Associate",
  `4` = "Bachelor's degree",
  `5` = "Graduate/professional degree"
))
df$PPETHM = revalue(factor(df$PPETHM), c(
  `1` = "White, Non-Hispanic",
  `2` = "Black, Non-Hispanic",
  `3` = "Other, Non-Hispanic",
  `4` = "Hispanic"
))
df$PPGENDER = revalue(factor(df$PPGENDER), c(
  `1` = "Male",
  `2` = "Female"
))
df$PPHHSIZE = revalue(factor(df$PPHHSIZE), c(
  `1` = "1",
  `2` = "2",
  `3` = "3",
  `4` = "4",
  `5` = "5+"
))
df$PPINCIMP = revalue(factor(df$PPINCIMP), c(
  `1` = "Less than $20,000",
  `2` = "$20,000 to $29,999",
  `3` = "$30,000 to $39,999",
  `4` = "$40,000 to $49,999",
  `5` = "$50,000 to $59,999",
  `6` = "$60,000 to $74,999",
  `7` = "$75,000 to $99,999",
  `8` = "$100,000 to $149,999",
  `9` = "$150,000 or more"
))
df$PPMARIT = revalue(factor(df$PPMARIT), c(
  `1` = "Married",
```

```
  `2` = "Widowed",
  `3` = "Divorced/Separated",
  `4` = "Never married",
  `5` = "Living with partner"
))
df$PPMSACAT = revalue(factor(df$PPMSACAT), c(
  `0` = "Non-Metro",
  `1` = "Metro"
))
df$PPREG4 = revalue(factor(df$PPREG4), c(
  `1` = "Northeast",
  `2` = "Midwest",
  `3` = "South",
  `4` = "West"
))
df$PPREG9 = revalue(factor(df$PPREG9), c(
  `1` = "New England",
  `2` = "Mid-Atlantic",
  `3` = "East-North Central",
  `4` = "West-North Central",
  `5` = "South Atlantic",
  `6` = "East-South Central",
  `7` = "West-South Central",
  `8` = "Mountain",
  `9` = "Pacific"
))
```

```
glimpse(df)
```

```
## Rows: 6,394
## Columns: 23
## $ sample            <fct> Age 62+ oversample, General population, General popu~
## $ fpl               <fct> 200%+ FPL, 200%+ FPL, 200%+ FPL, 200%+ FPL, 200%+ FP~
## $ FWBscore          <int> 55, 51, 49, 49, 49, 67, 51, 47, 43, 58, 78, 62, 50, ~
## $ FSscore           <int> 44, 43, 42, 42, 42, 57, 54, 35, 58, 42, 66, 57, 49, ~
## $ LMscore           <int> 3, 3, 3, 2, 1, 3, 3, 3, 2, 3, 2, 3, 2, 3, 3, 3, 1, 2~
## $ KHscore           <dbl> 1.267, -0.570, -0.188, -1.485, -1.900, 0.242, 1.267,~
## $ LIVINGARRANGEMENT <fct> "I am the only adult in the household", "I live with~
## $ EARNERS           <fct> One, Two, Two, Refused, Two, One, One, One, One, One~
## $ SAVINGSRANGES     <fct> "$20,000-74,999", "$1-99", "$1,000-4,999", "Refused"~
## $ HOUSING           <fct> I own my home, I own my home, I own my home, Refused~
## $ VALUERANGES       <fct> "$150,000-249,999", "$150,000-249,999", "$250,000-39~
## $ MORTGAGE          <fct> "$50,000-199,999", "$50,000-199,999", "$50,000-199,9~
## $ agecat            <fct> 75+, 35-44, 35-44, 35-44, 25-34, 25-34, 35-44, 25-34~
## $ generation        <fct> Pre-Boomer, Gen X, Gen X, Gen X, Millennial, Millenn~
## $ PPEDUC            <fct> Bachelor's degree, High school degree/GED, Some coll~
## $ PPETHM            <fct> "White, Non-Hispanic", "White, Non-Hispanic", "Black~
## $ PPGENDER          <fct> Male, Male, Male, Male, Male, Male, Female, Female, ~
## $ PPINCIMP          <fct> "$75,000 to $99,999", "$60,000 to $74,999", "$60,000~
## $ PPHHSIZE          <fct> 1, 2, 3, 1, 5+, 2, 5+, 3, 4, 3, 5+, 2, 4, 3, 3, 5+, ~
## $ PPMARIT           <fct> Divorced/Separated, Divorced/Separated, Divorced/Sep~
## $ PPMSACAT          <fct> Metro, Metro, Metro, Metro, Metro, Metro, Metro, Met~
## $ PPREG4            <fct> West, Midwest, West, South, Midwest, Midwest, Midwes~
```

```
## $ PPREG9            <fct> Mountain, East-North Central, Pacific, West-South Ce~
```

Type of Study

What type of study is this (observational/experiment)?

This is an observational study based on a financial wellness survey conducted on 6394 participants

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

The data was collected as part of the Consumer Financial Protection Bureau's efforts to develop a rigorousness set of research activities designed to define and measure "success" for financial literacy initiatives.

The PUF survey results can be accessed as a csv file *Financial well-being survey data*

Dependent Variable

What is the response variable? Is it quantitative or qualitative?

The dependent variable for the analysis the Financial well-being scale score (FWBscore). This is a qualitative continuous variable.

Independent Variable

You should have two independent variables, one quantitative and one qualitative.

The independent variables for this analysis are Race (PPETHM) and Gender (PPGENDER) both are qualitative.

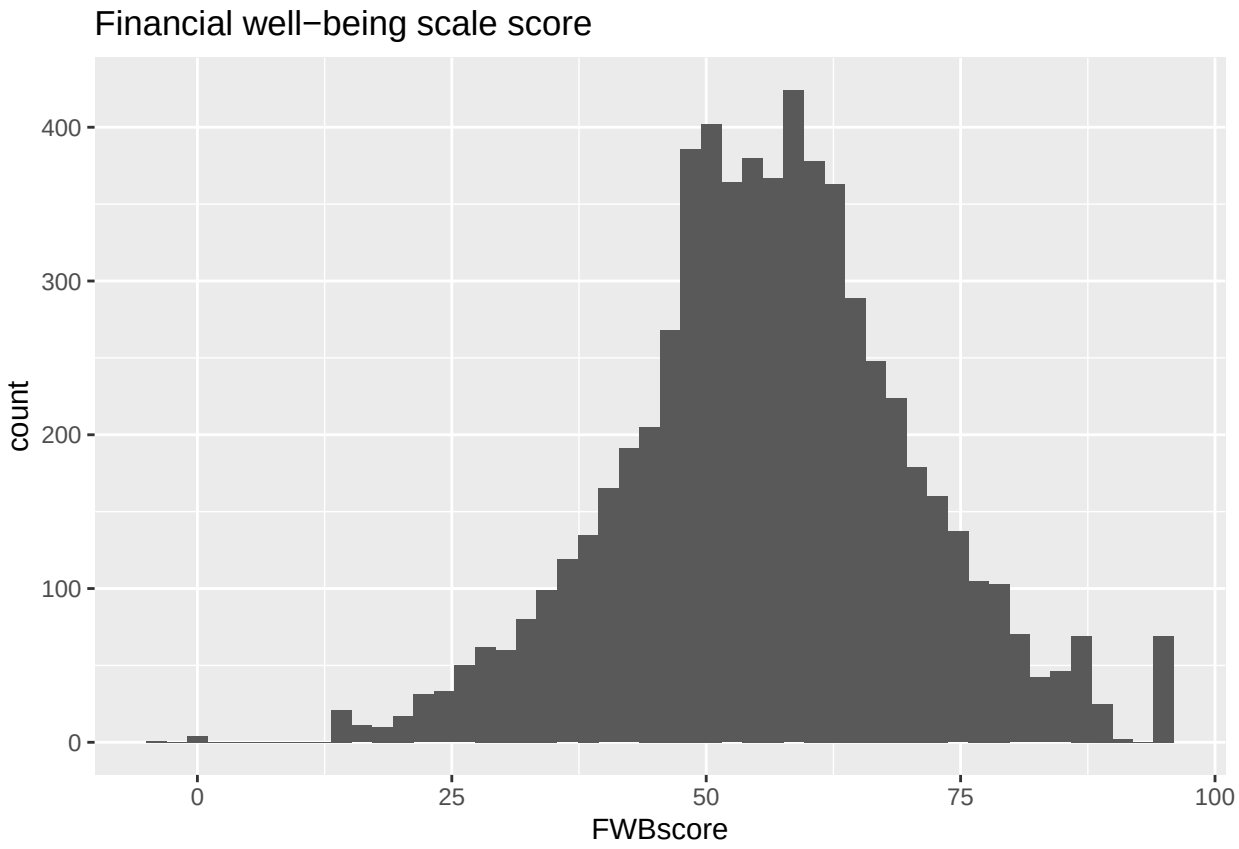
Relevant Summary Statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

Financial Well-Being Scale Score (FWBscore)

The financial well-being score appears normally distributed with mean of 56.03 and a median 56.00.

```
df %>%
ggplot(aes(x = FWBscore)) +
  geom_histogram(bins = 50) +
  labs(
   title = paste(
     "Financial well-being scale score")
  )
```




```
summary(df$FWBscore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -4.00   48.00   56.00   56.03   65.00   95.00
```

The financial skills score appears normally distributes however the LM and KH Financial Knowledge Score both seem right skewed but that could be factor or limited volumes.

```
summary(df$FSscore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -1.00   42.00   50.00   50.72   57.00   85.00
```

```
ggp1 <- df %>%
ggplot(aes(x = FSscore)) +
  geom_histogram(bins = 50) +
  labs(
   title = paste(
     "Financial skill scale score")
  )


summary(df$LMscore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   3.000   2.506   3.000   3.000
```

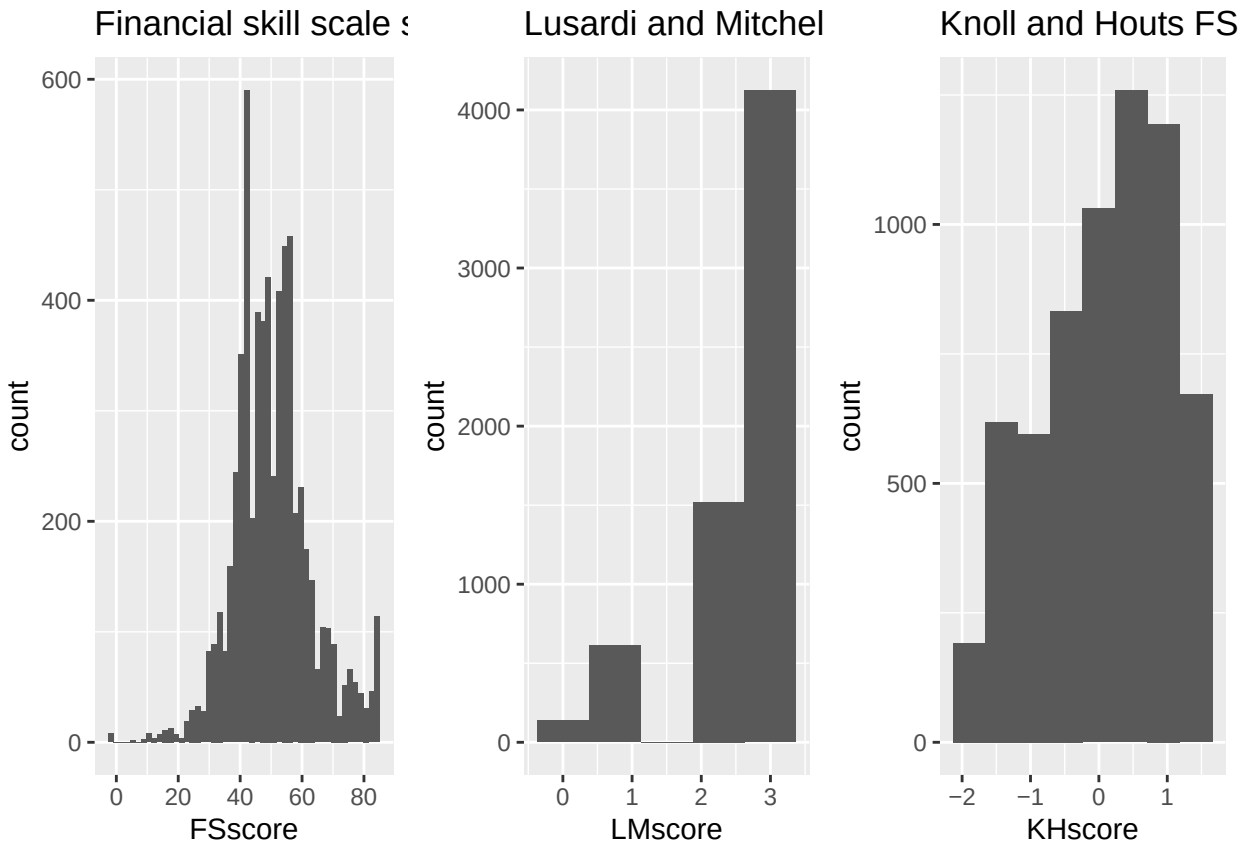
```
ggp2 <- df %>%
ggplot(aes(x =  LMscore)) +
  geom_histogram(bins = 5) +
  labs(
   title = paste(
     "Lusardi and Mitchell FS Knowledge")
  )


summary(df$KHscore)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.05300 -0.57000 -0.18800 -0.05694  0.71200  1.26700
```

```
ggp3 <- df %>%
ggplot(aes(x =  KHscore)) +
  geom_histogram(bins = 8) +
  labs(
   title = paste(
     "Knoll and Houts FS Knowledge")
  )

grid.arrange(ggp1, ggp2, ggp3, ncol = 3)
```



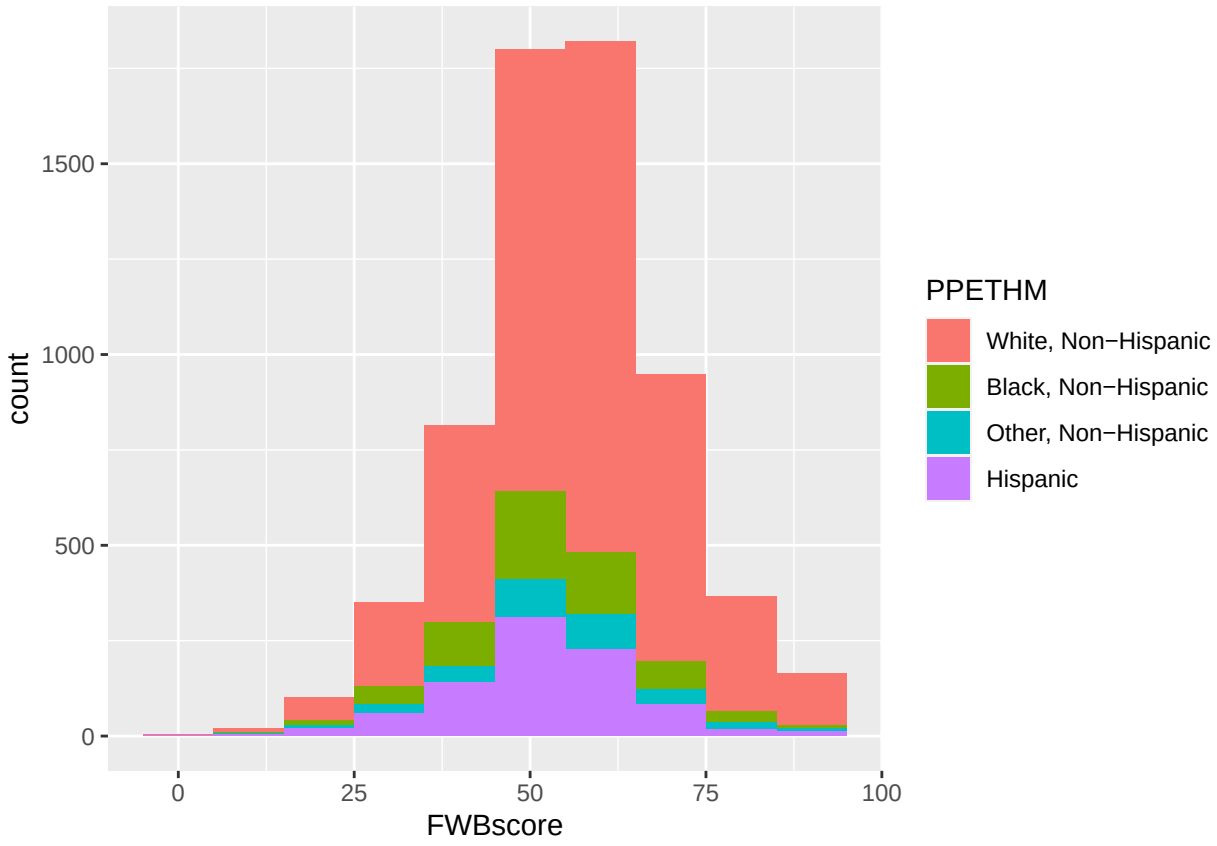
Independent Variable - Race (PPETHM)

The means financial wellness scores for all non white ethnic groups are lower with Blacks and Hispanics showing a smaller standard deviation. The histogram's across ethnic group are normally distributed an show a substantial overlap.

```
df %>%
    group_by(PPETHM) %>%
    dplyr::summarize(n = n(), mean=mean(FWBscore), median(FWBscore), sd(FWBscore))
```

```
## # A tibble: 4 x 5
##   PPETHM                 n  mean `median(FWBscore)` `sd(FWBscore)`
##   <fct>              <int> <dbl>              <dbl>          <dbl>
## 1 White, Non-Hispanic  4498  57.4                 58           14.2
## 2 Black, Non-Hispanic   685  52.9                 52           13.7
## 3 Other, Non-Hispanic   336  54.5                 55           14.5
## 4 Hispanic              875  52.2                 52           12.9
```

```
df %>%
    ggplot(aes(x=FWBscore, fill=PPETHM)) +
    geom_histogram(binwidth=10)
```



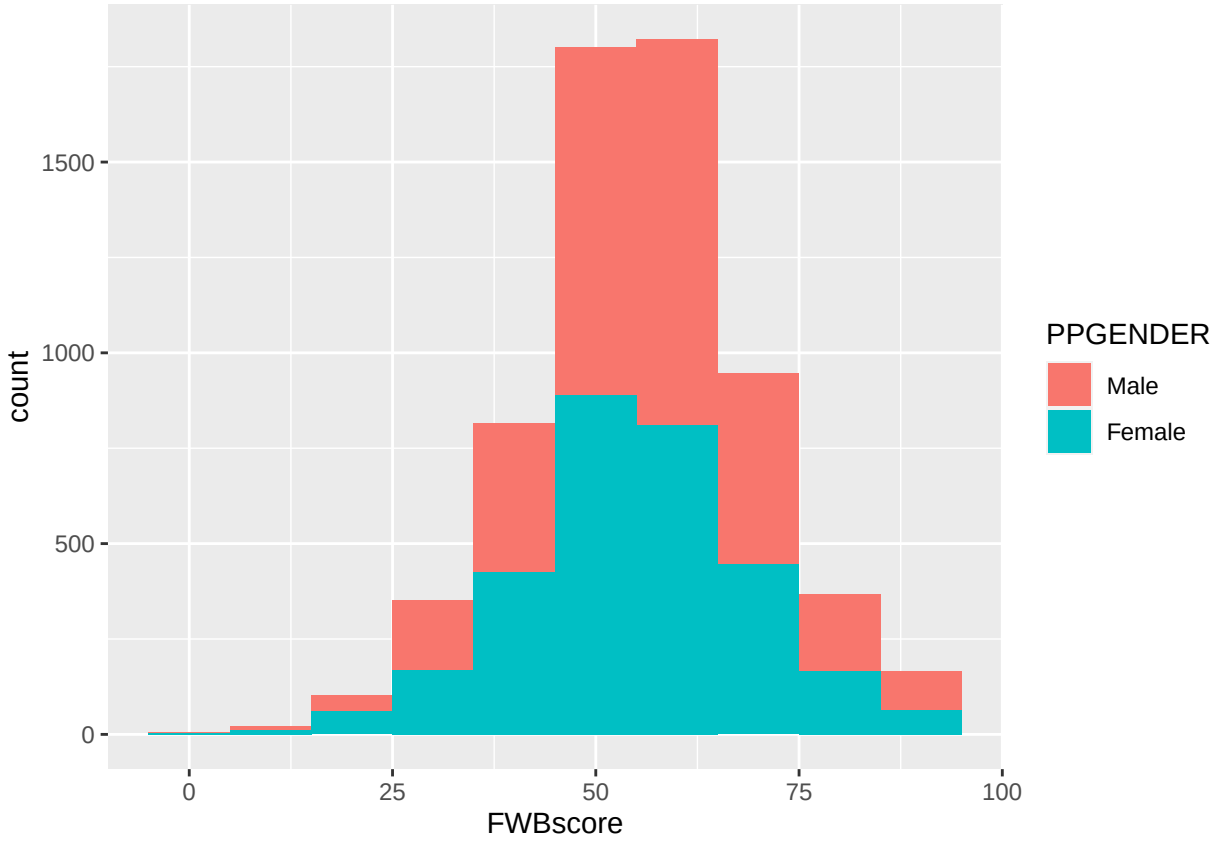
Independent Variable - Gender (PPGENDER)

The means financial wellness scores for all non white ethnic groups are lower with Blacks and Hispanics showing a smaller standard deviation. The histogram's across ethnic group are normally distributed an show a substantial overlap.

```
df %>%
    group_by(PPGENDER) %>%
    dplyr::summarize(n = n(), mean=mean(FWBscore), median(FWBscore), sd(FWBscore))
```

```
## # A tibble: 2 x 5
##   PPGENDER     n  mean `median(FWBscore)` `sd(FWBscore)`
##   <fct>    <int> <dbl>              <dbl>          <dbl>
## 1 Male      3352  56.7                 57           14.1
## 2 Female    3042  55.3                 55           14.1
```

```
df %>%
    ggplot(aes(x=FWBscore, fill=PPGENDER)) +
    geom_histogram(binwidth=10)
```



Final Thoughts

The value of this data set is that I will be able to conduct the initial analysis and look for additional confounders or additional variables that could assist in explaining the difference in financial wellness scores

across the surveys.