Final Project Part C: Explore Factors Affecting Income in NLSY ’79 Data

Code

Team Members

Member 1: Gina O’Riordan
Member 2: Marisa Roman
Member 3: Justin Saslaw

Introduction

In this document, we survey the data from National Longitudinal Survey of Youth 1979 (NLSY79) to review the relationship of three variables to income:

- Years of Education
- Gender
- Race

Additionally, we explore the relationship between the combination of Years of Education & Gender as well as the combination of Years of Education & Race on income. This EDA analysis will conclude with hypothesis for further analysis.

We limit our analysis to income data from 2008 - 2014, as we want to focus on recent information. Note: We originally chose the range 2004 - 2014, but discovered that significant portions of education data were missing in 2004 and 2006, so we further restricted our range.

Description of the Data

We are using three datasets from the National Longitudinal Survey of Youth 1979 (NLSY79). The datasets have been provided to us in a tidied and cleaned form, though we will inspect the data in our chosen variables to determine whether further cleaning is necessary.

To start, we load the `tidyverse` library:

Hide

```
library(tidyverse)
```

```
Loading tidyverse: ggplot2
Loading tidyverse: tibble
Loading tidyverse: tidyr
Loading tidyverse: readr
Loading tidyverse: purrr
Loading tidyverse: dplyr
Conflicts with tidy packages ----------------------------------------------------------------------
filter(): dplyr, stats
lag():    dplyr, stats
```

Income

This analysis will review income as the dependent variable. First we will load the data and examine it.

Hide

```
load("income_data_nlsy79.RData")
```

Hide

```
glimpse(income_data_nlsy79)
```

```
Observations: 291,778
Variables: 3
$ CASEID <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22,...
$ income <int> NA, 10000, 7000, 1086, 2300, 3250, 4975, 7500, 5000, 9000, 4002, 9000, 18500, ...
$ year   <int> 1982, 1982, 1982, 1982, 1982, 1982, 1982, 1982, 1982, 1982, 1982, 1982, 1982, ...
```

Seen in the code below, the survey data ranges from 1982 to 2014. In this study, we will be reviewing data from 2008-2014.

Hide

```
unique(income_data_nlsy79$year)
```

```
 [1] 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1996 1998 2000 2002 2004
[19] 2006 2008 2010 2012 2014
```

From analyzing the unique year values above, it appears the study contains data from even years within the time frame to be studied. After truncating the income dataset to retain only years 2008 through 2014 (inclusive), there are 50,744 records available.

Hide

```
# Truncate income_data_nlsy79 so that only years 2008 - 2014 are retained
income_data_nlsy79 <- filter(income_data_nlsy79, year >= 2008)
# Review results
unique(income_data_nlsy79$year)
```

```
[1] 2008 2010 2012 2014
```

Hide

```
glimpse(income_data_nlsy79)
```

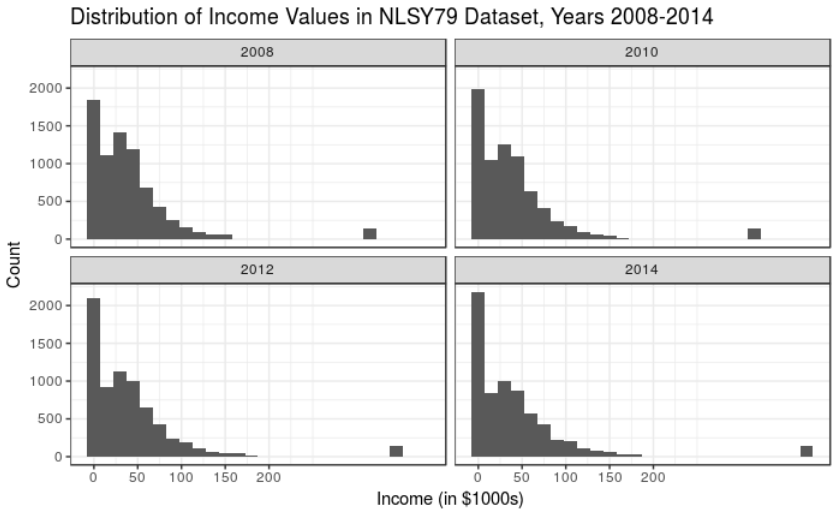
```
Observations: 50,744
Variables: 3
$ CASEID <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22,...
$ income <int> NA, 5000, 30000, NA, NA, 86000, 32500, 41000, 50000, NA, NA, NA, NA, 85000, 0,...
$ year   <int> 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, ...
```

Distribution of Income Data

A histogram of the income data reveals a right-skewed distribution, with values over $200K as candidates for outliers. For now, we will leave these values in place and we can filter them out if necessary later on when graphing and analyzing to see the full picture of the data.

Hide

```
# Generate histogram of income values for our income dataset to review the distribution
ggplot(
  data = income_data_nlsy79 %>%
    filter(!is.na(income)) %>%
    group_by(year),
  aes(x = income)
) + geom_histogram(binwidth = 15000) + facet_wrap(~year) +
  scale_x_continuous(breaks = c(0, 50000, 100000, 150000, 200000),
                     labels = c("0", "50", "100", "150", "200")) +
  theme_bw() + labs(title = "Distribution of Income Values in NLSY79 Dataset, Years 2008-2014",
                    x = "Income (in $1000s)", y = "Count")
```



Truncation of Income Data

The data dictionary truncates the income data by taking the top 2% of income each year and assigning it the minimum income value for the top 2% in that year. For example, the maximum income value for 2014 is `$370,314` . Repondents with incomes above that amount were recorded as having an income of `$370,314` . As noted above, we will keep this data in place in its truncated form as to not eliminate the top 2% of income values. We can choose whether to exclude these values if necessary later on when graphing and analyzing to see the full picture of the data.

The code below reviews the maximum income values by year and the count & percentage of data points with that maximum income value.

Hide

```
# Determine the maximum income amount by year, as well as the count and percentage of participants with their income recorde
d at the maximum value
income_data_nlsy79 %>%
  filter(!is.na(income)) %>%
  group_by(year) %>%
  summarize(max_income = max(income),
            count_with_max_income = sum(income == max_income),
            pct_with_max_income = (count_with_max_income*100)/n()) %>%
  rename(Year = year,
         "Max Income" = max_income,
         "Count with Max Income" = count_with_max_income,
         "% with Max Income" = pct_with_max_income)
```

Gender

The first independent variable studied is gender; once the data is loaded and cleaned, its relationship to income during the years 2008-2014 will be analyzed. Note that gender is represented by the ‘sex’ variable in this dataset.

Hide

```
load("physical_data_nlsy79.RData")
```

Hide

```
glimpse(physical_data_nlsy79)
```

```
Observations: 253,720
Variables: 9
$ CASEID <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22,...
$ weight <int> NA, 120, NA, 110, 130, 200, 131, 179, 145, 115, 155, 118, 180, 135, 185, 130, ...
$ year   <int> 1981, 1981, 1981, 1981, 1981, 1981, 1981, 1981, 1981, 1981, 1981, 1981, 1981, ...
$ eyes   <chr> NA, "hazel", "blue", "blue", NA, "brown", "brown", "hazel", "hazel", "hazel", ...
$ hair   <chr> NA, "light brown", "blond", "light brown", NA, "brown", "brown", "brown", "bro...
$ race   <chr> "NBNH", "NBNH", "NBNH", "NBNH", "NBNH", "NBNH", "NBNH", "NBNH", "NBNH", "NBNH"...
$ sex    <chr> "female", "female", "female", "female", "male", "male", "male", "female", "mal...
$ height <int> 65, 62, NA, 67, 63, 64, 65, 65, 66, 66, 71, 66, 71, 67, 73, 63, 69, 69, 64, 64...
$ BMI    <dbl> NA, 21.94843, NA, 17.22855, 23.02862, 34.33015, 21.79968, 29.78735, 23.40377, ...
```

Truncation of Gender Data

Then we truncate the data to view only the decade from 2008-2014 and view it.

Hide

```
physical_data_nlsy79 <- filter(physical_data_nlsy79, year >= 2008)
```

Hide

```
glimpse(physical_data_nlsy79)
```

```
Observations: 50,744
Variables: 9
$ CASEID <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22,...
$ weight <int> NA, 160, 189, NA, NA, 190, 165, 240, 190, NA, NA, NA, NA, 200, 230, 160, 183, ...
$ year   <int> 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, ...
$ eyes   <chr> NA, "hazel", "blue", "blue", NA, "brown", "brown", "hazel", "hazel", "hazel", ...
$ hair   <chr> NA, "light brown", "blond", "light brown", NA, "brown", "brown", "brown", "bro...
$ race   <chr> "NBNH", "NBNH", "NBNH", "NBNH", "NBNH", "NBNH", "NBNH", "NBNH", "NBNH", "NBNH"...
$ sex    <chr> "female", "female", "female", "female", "male", "male", "male", "female", "mal...
$ height <int> NA, 62, 70, NA, NA, 65, 69, 64, 67, NA, NA, NA, NA, 68, 74, 64, 69, 69, NA, 64...
$ BMI    <dbl> NA, 29.26457, 27.11885, NA, NA, 31.61786, 24.36640, 41.19618, 29.75840, NA, NA...
```

Distribution of Gender Data

There are two values for gender: female and male. There are no missing gender values. There are 480 more male data points than female in each year under consideration. We will keep this in mind when analyzing income by gender.

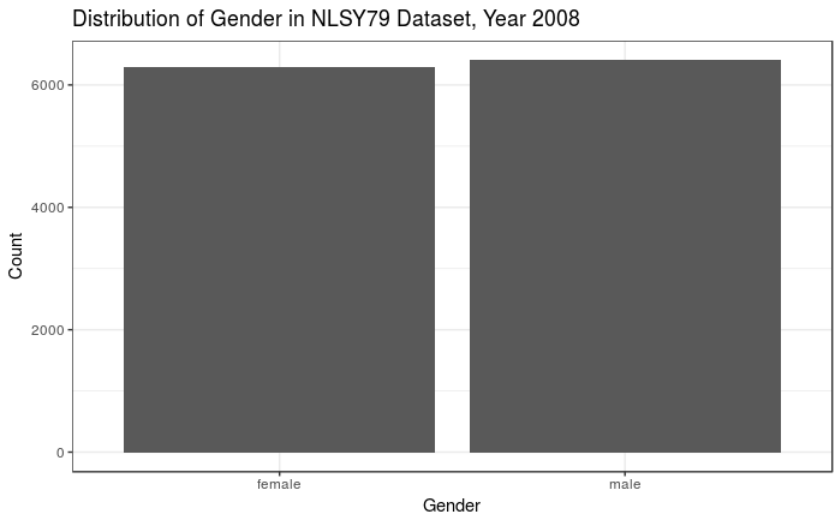
Hide

```
summary(as.factor(physical_data_nlsy79$sex))
```

```
female   male
 25132  25612
```

Hide

```
# Plot distribution of gender, restricting to the first year in the data set so as not to overstate the count of participant
s
ggplot(
  data = filter(physical_data_nlsy79, year == 2008),
  aes(x = sex)
) + geom_bar() + theme_bw() +
  xlab("Gender") +
  labs(title = "Distribution of Gender in NLSY79 Dataset, Year 2008", y = "Count")
```



Race

Race is available in the `physical_data_nlsy79` dataset previously loaded.

Truncation of Race Data

Note that the dataset has already been limited to years 2008-2014 for the Gender variable truncation above.

Distribution of Race Data

There are three possible values for race: Black, Hispanic, and NBNH (not black or Hispanic). There are no missing values, though the proportion of non-black/non-Hispanic subjects is much greater than that of black or Hispanic subjects. Again, we will keep this proportion in mind when performing our analyses.

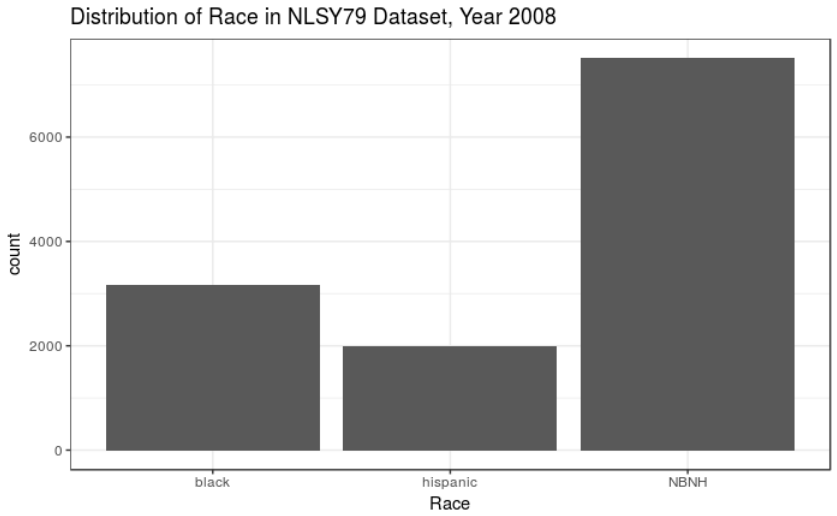
Hide

```
summary(as.factor(physical_data_nlsy79$race))
```

```
   black hispanic     NBNH
   12696     8008    30040
```

Hide

```
# Plot the distribution of race, restricting to the first year in the data set so as not to overstate the count of participa
nts
ggplot(
  data = filter(physical_data_nlsy79, year == 2008),
  aes(x = race)
) + geom_bar() + theme_bw() +
  xlab("Race") +
  labs(title = "Distribution of Race in NLSY79 Dataset, Year 2008")
```



Education

The variable that is reported here as education is the one on the survey described as “highest grade completed as of May 1 of survey year.” Therefore, this is a cumulative accounting of the achieved level of education.

Truncation of Education Data

First, we load the data -

Hide

```
load("education_data_nlsy79.RData")
```

then limit the education records to those from 2008 onward:

Hide

```
education_data_nlsy79 <- filter(education_data_nlsy79, year >= 2008)
```

Distribution of Education Data

Education is coded as a numeric value ranging from 0 to 20 inclusive, with an additional possible value of 95. A high school degree is 12 years of education, an associate college degree is 14 years, a bachelor degree is 16 years, and postgraduate education of 4 years or more is 20.

Hide

```
sort(unique(education_data_nlsy79$education))
```

```
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 95
```

The value of 95 corresponds to an “ungraded” education level, so we will replace ‘95’ values with `NA` then review the results in chart form and via histogram.

Hide

```
# Replace education values of 95 with NA since 95 corresponds to "ungraded"
education_data_nlsy79$education[education_data_nlsy79$education == 95] <- NA
# Review results
education_data_nlsy79 %>%
  group_by(year) %>%
  summarize("Count with Education Data" = sum(!is.na(education)),
            "Count without Education Data" = sum(is.na(education))) %>%
  rename("Year" = year)
```

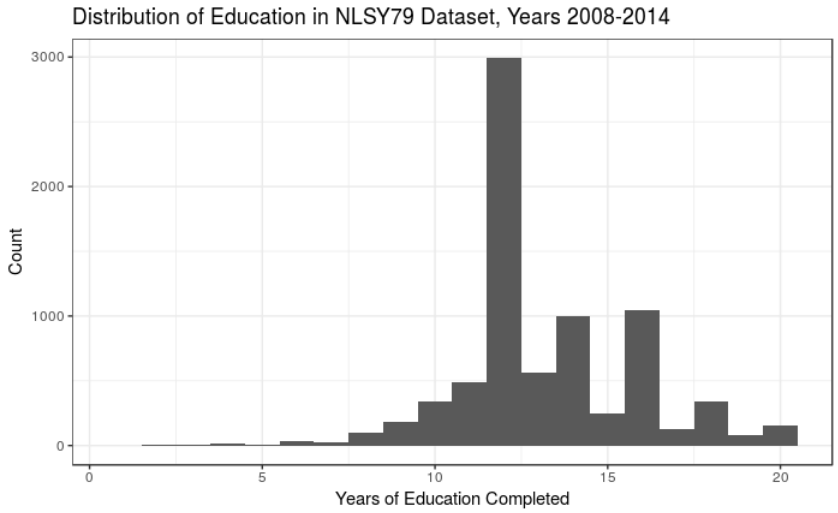
Year	Count with Education Data	Count without Education Data
<int>	<int>	<int>
2008	7738	4948
2010	7544	5142
2012	7290	5396
2014	7057	5629

4 rows

This plot will show there is a significant amount of respondents who have 12 years of education, signifying a high school education.

Hide

```
# Plot distribution of years of education, restricting to the first year in the data set so as not to overstate the count
ggplot(
  data = filter(education_data_nlsy79, year == 2008 & !is.na(education)),
  aes(x = education)
) + geom_histogram(binwidth = 1) + theme_bw() +
  xlab("Years of Education Completed") +
  labs(title = "Distribution of Education in NLSY79 Dataset, Years 2008-2014",
       y = "Count")
```



EDA: Factors Effecting Income

Joining the Data into a New Dataset

By joining the income, education, and physical data sets to produce a single data set, we have one set of data to reference going forward. The dataset will be called income_educ_phys_recent, and the records for which income is NA are excluded since income is the focus of the analysis.

Hide

```
# Create consolidated dataset for use in analysis
income_educ_phys_recent <- income_data_nlsy79 %>%
  inner_join(education_data_nlsy79, by = c("CASEID", "year")) %>%
  inner_join(physical_data_nlsy79, by = c("CASEID", "year")) %>%
  filter(!is.na(income) & !is.na(education) & !is.na(sex) & !is.na(race)) %>%
  select(CASEID, income, year, education, sex, race)
# Rename columns for nicer display
names(income_educ_phys_recent) = c("CASEID", "Income", "Year", "Years of Education Completed", "Gender", "Race")
# Review result
income_educ_phys_recent
```

CASEID <int>	Income <int>	Year <int>	Years of Education Completed <int>	Gender <chr>	Race <chr>
2	5000	2008	13	female	NBNH
3	30000	2008	12	female	NBNH
6	86000	2008	16	male	NBNH
7	32500	2008	12	male	NBNH
8	41000	2008	14	female	NBNH
9	50000	2008	16	male	NBNH
14	85000	2008	20	female	NBNH
15	0	2008	16	male	NBNH
16	66000	2008	13	female	NBNH
17	50000	2008	12	male	NBNH

1-10 of 28,370 rows Previous **1** 2 3 4 5 6 … 100 Next

Effect of Education on Income

For the first part of the analysis, we will review the effect of the independent variable education on income.

Mean and Median Income by Years of Education

First we will create an `income_by_education` tibble with a created field for mean income and one for median income. For both tibbles, rows are 'years of education completed' and one column for each year (2008, 2010, 2012, and 2014).

Hide

```
# Create income_by_education tibble, summarizing the mean and median income by year and years of education
income_by_education <- income_educ_phys_recent %>%
  group_by(Year, `Years of Education Completed`) %>%
  summarize(`Mean Income` = mean(Income, na.rm = TRUE),
            `Median Income` = median(Income, na.rm = TRUE))
# Display mean income by years of education completed in an easy-to-read format
spread(select(income_by_education, -`Median Income`), key = Year, value=`Mean Income`)
```

	Years of Education Completed <int>	2008 <dbl>	2010 <dbl>	2012 <dbl>	2014 <dbl>
1	1	0.000	25000.00	18000.000	*NA*
2	2	9333.333	10000.00	22550.000	43333.333
3	3	12514.400	14048.92	5000.000	5375.000
4	4	34081.818	66813.78	26394.000	20655.417
5	5	15957.143	11650.00	12833.333	4195.000

	Years of Education Completed	2008	2010	2012	2014
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
6	6	11533.906	11308.57	9317.857	11194.643
7	7	13630.931	12917.91	10406.250	4389.714
8	8	12919.065	11576.71	8297.707	14909.392
9	9	16842.416	15869.69	15617.939	13807.923
10	10	15963.675	15330.79	15409.268	16601.627

1-10 of 20 rows Previous **1** 2 Next

Hide

```
# Display median income by years of education completed in an easy-to-read format
spread(select(income_by_education, -`Mean Income`), key = Year, value=`Median Income`)
```

	Years of Education Completed	2008	2010	2012	2014
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	0	25000.0	18000	*NA*
2	2	0	10000.0	33000	42000.0
3	3	14000	12000.0	0	0.0
4	4	20000	22000.0	23250	15500.0
5	5	3300	6000.0	8000	0.0
6	6	7000	6500.0	0	3300.0
7	7	12000	1200.0	0	0.0
8	8	250	0.0	0	0.0
9	9	10000	6500.0	3320	0.0
10	10	9300	6000.0	300	3930.0

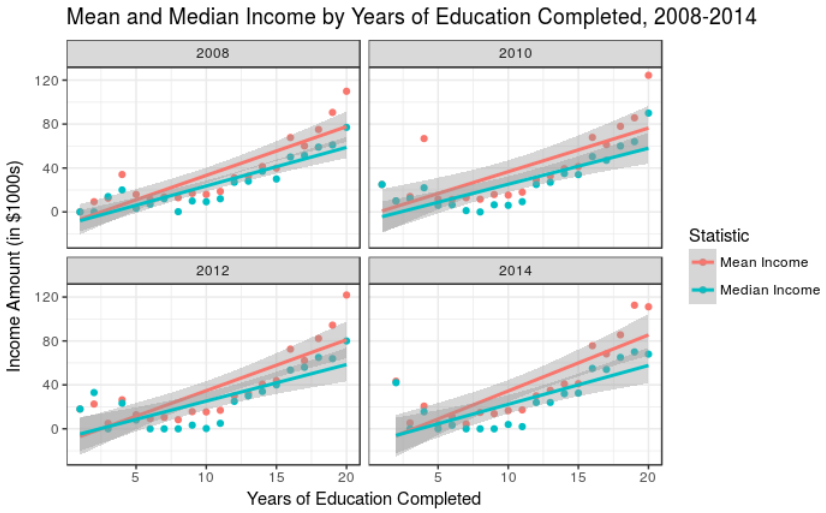
1-10 of 20 rows Previous **1** 2 Next

The plot below shows the mean and median income data in dollars by years of education. With the trend lines, it is clear that there is a positive correlation between income and education.

Additionally, the mean income is higher than the median (as expected for our right-skewed data), but since we want to keep the high income values for subsequent analysis, we will use only the mean going forward, since it is affected by outliers.

Hide

```
# Gather mean and median values for use in a plot
income_by_education <- income_by_education %>%
  gather(key = "Statistic", value = "Amount", -`Years of Education Completed`, -Year)
# Plot mean and median income values by year for each level of years of education completed
ggplot(
  data = income_by_education,
  aes(x = `Years of Education Completed`, y = Amount, color = Statistic)
) + geom_point() + geom_smooth(method = "lm") +
  scale_y_continuous(breaks = c(0, 40000, 80000, 120000, 160000),
                     labels = c(0, 40, 80, 120, 160)) +
  labs(x = "Years of Education Completed",
       y = "Income Amount (in $1000s)",
       title = "Mean and Median Income by Years of Education Completed, 2008-2014") +
  facet_wrap(~Year) +
  theme_bw()
```



Effect of Gender on Income

For the second part of the analysis, we will review the effect of the independent variable gender on income.

Mean Income by Gender

First we will create an `income_by_gender` tibble with a created field for mean income in the same way as above with the `income_by_education` dataset. The data is grouped by year and gender; it seems to demonstrate relatively no change year over year for female respondents versus more fluctuations year over year for male repsondents.

Hide

```
# Create income_by_gender tibble, summarizing the mean income by gender
income_by_gender <- income_educ_phys_recent %>%
  group_by(Year, Gender) %>%
  summarize(`Mean Income` = mean(Income, na.rm = TRUE))
# Display mean income by gender in an easy-to-read format
spread(income_by_gender, key = Year, value=`Mean Income`)
```

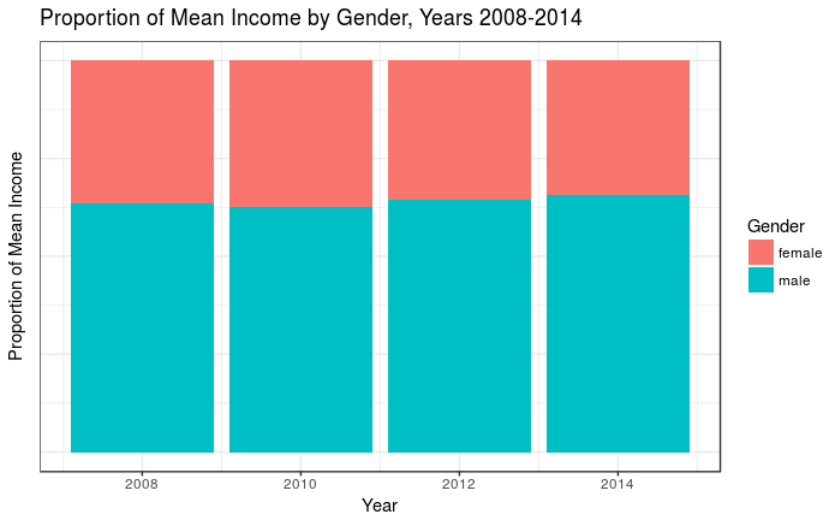
	Gender	2008	2010	2012	2014
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	female	29551.26	29836.60	29571.21	29493.59
2	male	51306.43	50054.71	53503.74	56170.20

2 rows

The plot below shows the proportion of mean income by gender for each of the four years of data. Though overall income rises over time for both genders, the incremental change is minimal; there is an initial dip from 2008 to 2010 and then an increase for 2012 and 2014. It should also be noted that the male mean income is greater than the female for each year. Let's explore this point further in the next plot.

Hide

```
# Plot the proportion of mean income by gender
ggplot(
  data = income_by_gender,
  aes(x = Year, y = `Mean Income`, fill = Gender)
) +  geom_col(position = "fill") + theme_bw() + scale_x_continuous(breaks = c(2008, 2010, 2012, 2014)) +
  # Remove y-axis labels and tick marks because the intent is to display the proportion of mean income
  theme(axis.ticks.y = element_blank(), axis.text.y = element_blank()) +
  ylab("Proportion of Mean Income") +
  labs(title = "Proportion of Mean Income by Gender, Years 2008-2014")
```



The plot below breaks down the gender variable's effect on mean income further by viewing the disparity by year making it easier to see the differences between the genders and trends year over year. The first point demonstrated in the graph is that males consistently make more than females, between 1.5 and 2x the female mean income. The second point is that the female mean income remains the same from 2008-2014, whereas the male income rises, particularly from 2012-2014. Later in this analysis we look at the combined effect of education and gender on income to see if the years of education for males & females can be attributed to this disparity in mean income.

Hide

```
# Plot mean income by gender per year
ggplot(
  data = income_by_gender,
  aes(Gender, `Mean Income`, fill = Gender)
) + geom_col() + facet_wrap(~Year) + theme_bw() +
  ylab("Mean Income (in $)") +
  labs(title = "Mean Income by Gender for Each Year, 2008-2014")
```

To further analyze the gender year over year change, the tibble below was created by finding the percent change for each year increment (2008 to 2010, 2010 to 2012, and 2012 to 2014) in the analysis. For females, there was about a 1% increase for the first year increment, followed by negative growth for the next two increments. On the other hand, male income decreased at first then grew for the following two increments.

Hide

```
# Calculate change in mean income by gender per year
income_by_gender_pct <- spread(income_by_gender, key = Year, value=`Mean Income`) %>%
  mutate(`% change 2008 to 2010` = (`2010`-`2008`)/`2008`*100,
         `% change 2010 to 2012` = (`2012`-`2010`)/`2010`*100,
         `% change 2012 to 2014` = (`2014`-`2012`)/`2012`*100,
         `% change 2008 to 2014` = (`2014`-`2008`)/`2008`*100) %>%
  select(Gender, `% change 2008 to 2010`, `% change 2010 to 2012`, `% change 2012 to 2014`, `% change 2008 to 2014`)
income_by_gender_pct
```

Gender	% change 2008 to 2010	% change 2010 to 2012	% change 2012 to 2014	% change 2008 to 2014
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
female	0.9655717	-0.8894627	-0.2624677	-0.1951243
male	-2.4397007	6.8905196	4.9836899	9.4798378

2 rows

Effect of Race on Income

For the third part of the analysis, we will review the effect of the independent variable race on income.

Mean Income by Race

First we create the mean_income_by_race tibble, that includes the mean income by race for the income_educ_phys_recent tibble. This tibble shows the large disparity between black, hispanic, and non black non hispanic respondent groups (listed here in ascending order).

Hide

```
#Show mean income by race
mean_income_by_race <-
  income_educ_phys_recent %>%
  group_by(Race) %>%
  summarise(`Mean Income` = mean(Income, na.rm = T))
mean_income_by_race
```

Race <chr>	Mean Income <dbl>
black	28976.22
hispanic	35853.11
NBNH	50028.29

3 rows

Then we will create an `income_by_race` tibble with a created field for mean income in the same way as the previous sections with the `income_by_education` & `income_by_gender` datasets. The data is grouped by year and race for the years 2008-2014 as opposed to the tibble above. At first glance, there seems to be relatively no change year over year for the black & hispanic respondents, while the non black non hispanic group's income rises each year, creating a larger overall mean.

Hide

```
# Create income_by_race to summarize the mean income by racial group per year reviewed
income_by_race <- income_educ_phys_recent %>%
  group_by(Year, Race) %>%
  summarize(`Mean Income` = mean(Income, na.rm = TRUE))
spread(income_by_race, key = Year, value=`Mean Income`)
```

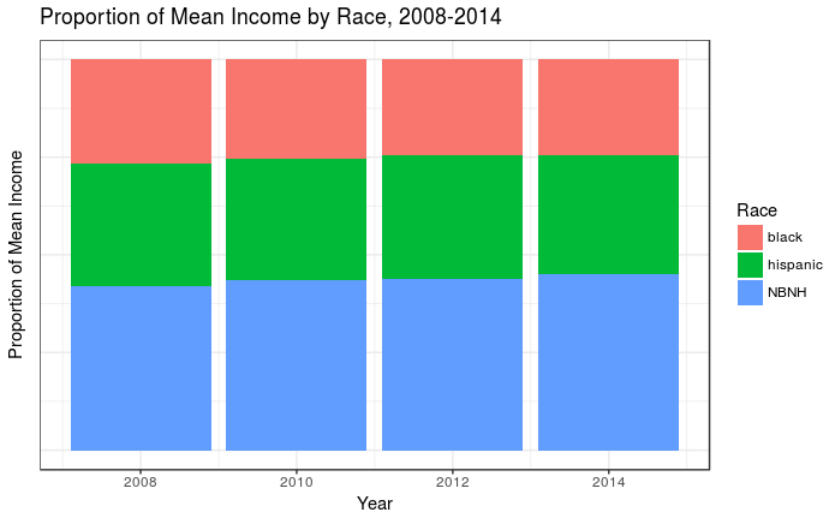
	Race <chr>	2008 <dbl>	2010 <dbl>	2012 <dbl>	2014 <dbl>
1	black	30268.35	28186.68	28364.36	29026.54
2	hispanic	36038.13	34935.00	36608.16	35840.37
3	NBNH	47876.25	48358.52	50876.11	53318.52

3 rows

The plot below shows the mean income by race for each of the four years of data. The income for non black non hispanic (NBNH) is overall highest per year, followed by hispanics, then blacks. Similarly to the gender trend, the incremental change for all three race groups is minimal per year and there is an initial dip in mean income from 2008 to 2010 and then an increase for 2012 and 2014.

Hide

```
# Plot proportion of mean income by race per year reviewed
ggplot(
  data = income_by_race,
  aes(x = Year, y = `Mean Income`, fill = Race)
) +  geom_col(position = "fill") + theme_bw() + ylab("Proportion of Mean Income") +
  scale_x_continuous(breaks = c(2008, 2010, 2012, 2014)) +
  theme(axis.ticks.y = element_blank(), axis.text.y = element_blank()) +
  labs(title = "Proportion of Mean Income by Race, 2008-2014")
```



To further analyze the race year over year change, the tibble below was created by finding the percent change for each year increment (2008 to 2010, 2010 to 2012, and 2012 to 2014) in the analysis. The black race group decreases then slightly increases for the next two increments; the hispanic group decreases increases then decreases again; and the non black non hispanic group increases over all three increments.

Hide

```
#Percentage Change from 2008-2014.
income_by_race_pct <- spread(income_by_race, key = Year, value=`Mean Income`) %>%
  mutate(`% change 2008 to 2010` = (`2010`-`2008`)/`2008`*100,
         `% change 2010 to 2012` = (`2012`-`2010`)/`2010`*100,
         `% change 2012 to 2014` = (`2014`-`2012`)/`2012`*100,
         `% change 2008 to 2014` = (`2014`-`2008`)/`2008`*100) %>%
  select(Race, `% change 2008 to 2010`, `% change 2010 to 2012`, `% change 2012 to 2014`, `% change 2008 to 2014`)
income_by_race_pct
```

Race <chr>	% change 2008 to 2010 <dbl>	% change 2010 to 2012 <dbl>	% change 2012 to 2014 <dbl>	% change 2008 to 2014 <dbl>
black	-6.877369	0.6303825	2.334524	-4.102666
hispanic	-3.061000	4.7893444	-2.097313	-0.548744
NBNH	1.007327	5.2060904	4.800708	11.367373

3 rows

Effect of Education & Gender on Income

Mean Income by Education & Gender

First we will create an `income_by_gender_educ` tibble by filtering the `income_educ_phys_recent` tibble by the year 2014 (to avoid data replication) with a created field for mean income as done with previous analysis datasets.

Hide

```
# Create income_by_gender_educ to summarize the mean income by gender & years of education for 2014
income_by_gender_educ <- income_educ_phys_recent %>%
  filter(Year == 2014) %>%
  group_by(Gender, `Years of Education Completed`) %>%
  summarize(`Mean Income` = mean(Income, na.rm = TRUE))
# Display income by gender and years of education completed in an easy-to-read format
spread(income_by_gender_educ, key = Gender, value = `Mean Income`)
```

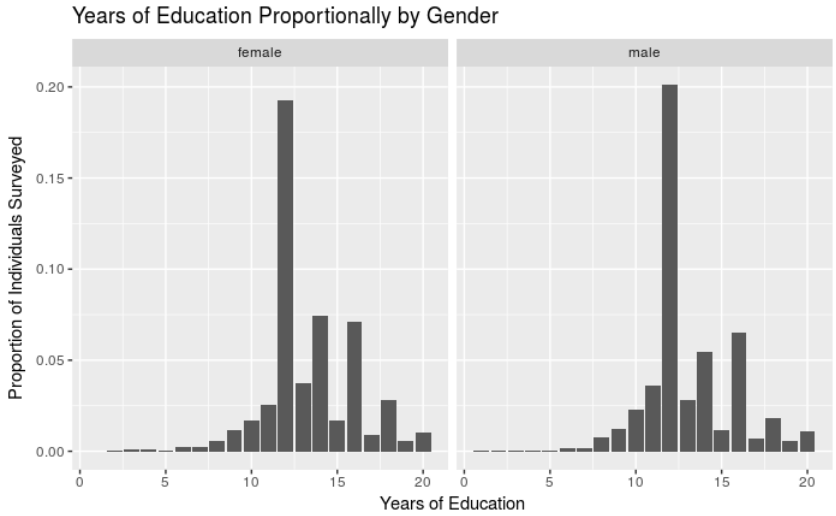
	Years of Education Completed <int>	female <dbl>	male <dbl>
1	2	0.000	65000.00
2	3	3750.000	7000.00
3	4	9644.167	31666.67
4	5	0.000	7341.25
5	6	5655.263	22888.89
6	7	1337.800	12019.50
7	8	5027.927	22144.04
8	9	9345.688	18212.95
9	10	11186.037	20749.74
10	11	11786.374	21362.44

1-10 of 19 rows Previous **1** 2 Next

The plot below shows the proportion of repondents by years of education, faceted by gender. This shows a significant amount of respondents received a high school education, with spikes at associate degree level and bachelor degree level. There are more males with a high school education, but there are more females with a college and higher education proportionally.

Hide

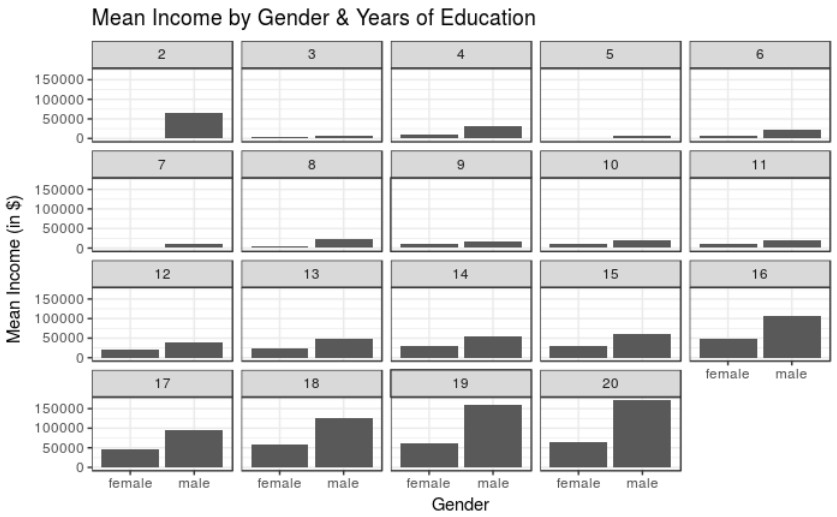
```
ggplot(
  data = income_educ_phys_recent,
  aes(x = `Years of Education Completed`)
) + geom_bar(aes(y = ..count../sum(..count..))) +
  facet_wrap(~Gender) +
  ylab("Proportion of Individuals Surveyed") + xlab("Years of Education") +
  labs(title = "Years of Education Proportionally by Gender")
```



Now let's analyze the mean income for each gender by faceting the data by year of education completed. Besides respondents with 2 years of education (where there are no female corresondents to compare male data against), the male respondents with 3 or more years of education make more than the female respondents for each year of education. The disparity is not as drastic until the 16 years of education graph, where male mean income is twice as high as females and then triple for 19 and 20+ years of education. Let's review this segment of data further.

Hide

```
ggplot(
  data = income_by_gender_educ,
  aes(x = Gender, y = `Mean Income`)
) + geom_col() + facet_wrap(~`Years of Education Completed`) + theme_bw() +
  ylab('Mean Income (in $)') +
  labs(title = "Mean Income by Gender & Years of Education")
```



First let's review the total proportion of respondents by gender who have higher education (since 16 years is bachelor degree and anything higher is graduate level).

Hide

```
#reviews the proportion of respondents with higher education, by gender
prop_educ_gender <-
  income_by_gender_educ %>%
  group_by(Gender) %>%
  summarise(higher_edu = mean(`Years of Education Completed` >= 16))
prop_educ_gender
```

Gender <chr>	higher_edu <dbl>
female	0.2631579
male	0.2631579

2 rows

Then let's create the `income_by_gender_educ` tibble to only include repsondents with higher education (>= 16) for plot analysis.

Hide

```
#filter data by >=16 years of education
income_by_gender_educ_16plus <- filter(income_by_gender_educ, `Years of Education Completed` >= 16)
# Display income by gender/years of education in an easy-to-read format
spread(income_by_gender_educ_16plus, key = Gender, value= `Mean Income`)
```

	Years of Education Completed <int>	female <dbl>	male <dbl>
1	16	47835.05	106351.59
2	17	46622.70	96251.49

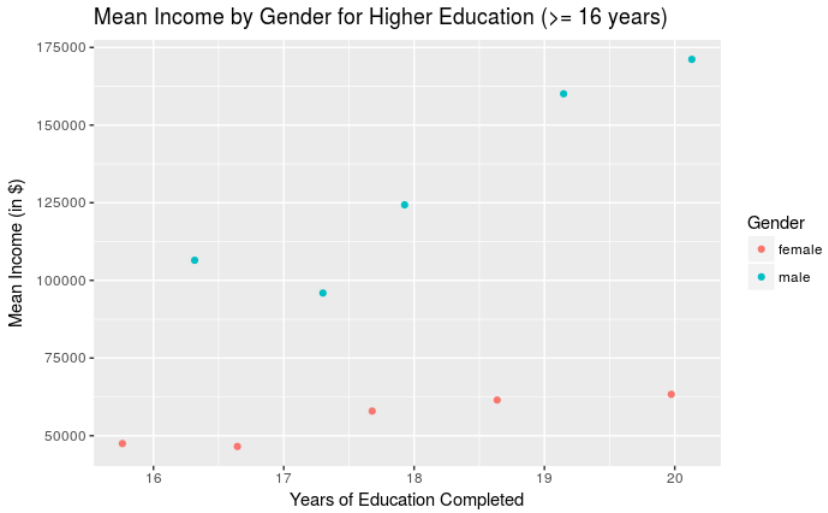
	Years of Education Completed <int>	female <dbl>	male <dbl>
3	18	58104.20	124759.29
4	19	61002.28	159675.83
5	20	63059.76	171176.85

5 rows

The jitter plot below gives a further analysis of the higher education segment. It demonstrates the gender gap in mean income, showing higher-educated females making less than higher-educated males.

Hide

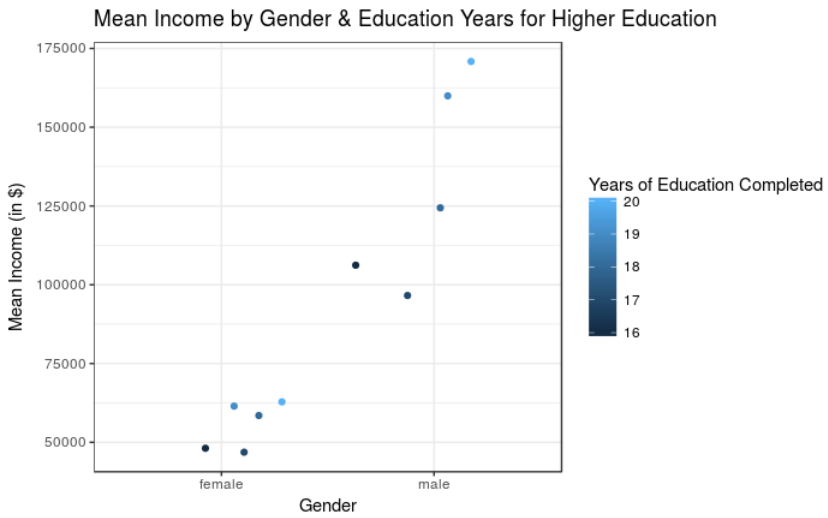
```
ggplot(
  data = income_by_gender_educ_16plus,
  aes(x = `Years of Education Completed`, y = `Mean Income`, color = Gender)
) + geom_jitter() +
  ylab("Mean Income (in $)") +
  labs(title = "Mean Income by Gender for Higher Education (>= 16 years)")
```



To dig into this trend further, the jitter plot below shows the gender gap based on the specific number of years educated. For females, the mean income stays relatively stable within $50-60k showing no positive correlation between years of education and mean income for this segment. For males, not only are the mean incomes higher overall for the same total years of education but also there is a positive correlation between years of higher education and mean income. As males spend more years in grad school, there is an increase in their mean income.

Hide

```
#plot gender & years of education (>=16) showing mean income
ggplot(
  data = income_by_gender_educ_16plus,
  aes(x = Gender, y = `Mean Income`, color = `Years of Education Completed`)
) + geom_jitter() + theme_bw() +
  ylab("Mean Income (in $)") +
  labs(title = "Mean Income by Gender & Education Years for Higher Education")
```



Effect of Education & Race on Income

Mean Income by Education & Race

First we will create a `mean_educ_race` tibble by filtering the income_educ_phys_recent tibble by excluding the values of 95 and creating the field for mean years of education by race. This shows that the black and hispanic respondents have an average of less than 13 years of education, while the non black and non hispanic respondents have almost 14 years of education.

Hide

```
#Show average education by race
mean_educ_race <- filter(income_educ_phys_recent, `Years of Education Completed` < 95) %>%
  group_by(Race) %>%
  summarise(mean_educ_years = mean(`Years of Education Completed`, na.rm = T))
mean_educ_race
```

Race <chr>	**mean_educ_years** <dbl>
black	12.90480
hispanic	12.41504
NBNH	13.71187

3 rows

The next tibble, `income_by_race_educ`, looks at the data from year 2014 (to avoid data replication) listing the respondents by unique race & years of education with corresponding mean income. Let's plot this data to find trends.

Hide

```
##Mean Income by Race & Education
income_by_race_educ <- income_educ_phys_recent %>%
  filter(Year == 2014) %>%
  group_by(Race, `Years of Education Completed`) %>%
  summarize(`Mean Income` = mean(Income, na.rm = TRUE))
# Display income by race/years of education in an easy-to-read format
spread(income_by_race_educ, key = Race, value= `Mean Income`)
```

Years of Education Completed <int>	**black** <dbl>	**hispanic** <dbl>	**NBNH** <dbl>

	Years of Education Completed <int>	black <dbl>	hispanic <dbl>	NBNH <dbl>
1	2	*NA*	42000.000	44000.00000
2	3	*NA*	5285.714	6000.00000
3	4	46000.00	18351.364	*NA*
4	5	0.00	7341.250	0.00000
5	6	7175.00	13004.348	0.00000
6	7	0.00	9442.769	14.18182
7	8	5720.00	17565.333	18634.97222
8	9	5581.35	17834.569	15483.66667
9	10	11478.47	17538.828	22612.98795
10	11	11631.87	21228.180	21223.73984

1-10 of 19 rows Previous **1** 2 Next

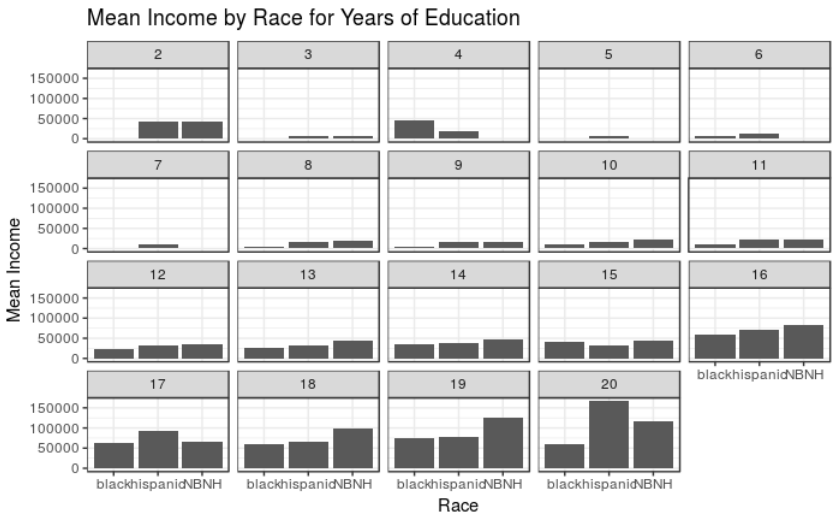
The plot below shows the mean income by years of education for each race group. For 2-7 years of education, there is inconsistent data in that not every racial grouping is represented, so we started our analysis at 8 years of education (a graduate from elementary and junior high school).

For 8-14 years of education, blacks are making the smallest mean income, followed by hispanics, and then non black and non hispanics are making the highest mean income. A low overall mean income would make sense since jobs not requiring high school or college degrees pay less. However, this data seems to demonstrate a racial disparity even with a low overall mean income based on years of education.

Then in year 15, blacks have a higher mean income than hispanics, before the trend continues in year 16. The overall means increases at 16 years of education, which makes sense since that represents a completed bachelors degree. At that point, however, the trends between race groups changes, with hispanics making more than the other two groups with 17 years of education. This is the same trend seen for 20+ years of education. The reason for this highest mean income for the hispanic race group would need further research to explain. For the >= 16 years of education that do not have this trend (16, 18, and 19), the non black non hispanic race group make the highest mean income.This could be attributed to the samel demonstrated racial disparity in years 8-14.

Hide

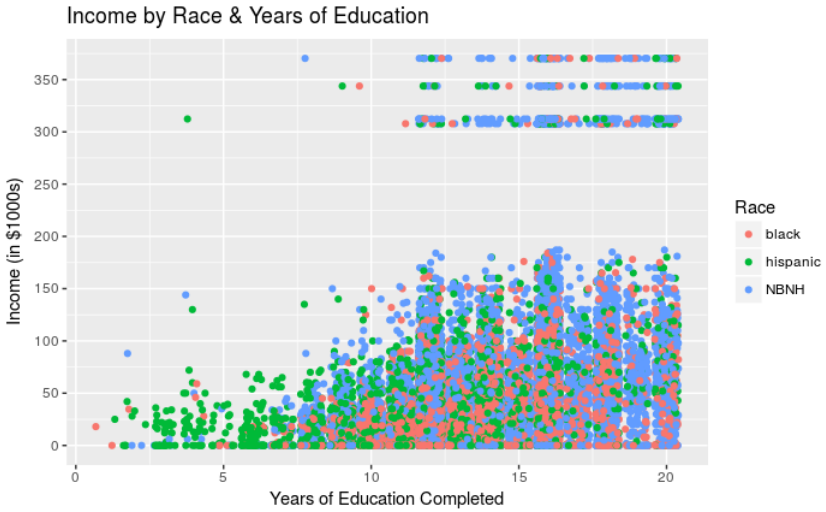
```
ggplot(
  data = income_by_race_educ,
  aes(x = Race, y = `Mean Income`)
) + geom_col() + facet_wrap(~`Years of Education Completed`) + theme_bw() +
   labs(title = "Mean Income by Race for Years of Education")
```



The jitter plot below gives a further analysis of education and race. It illustrates that a higher number of hispanic and blacks have less than a high school education and that correlates with lower income. There is a general uwpard bias for income and education regardless of race although the degree is worth exploring further.

Hide

```
ggplot(
  data = income_educ_phys_recent,
  aes(x = `Years of Education Completed`, y = Income, color = Race)
) + geom_jitter() +
  ylab("Income (in $1000s)") +
  scale_y_continuous(breaks = c(0, 50000, 100000, 150000, 200000, 250000, 300000, 350000),
                     labels = c(0, 50, 100, 150, 200, 250, 300, 350)) +
  labs(title = "Income by Race & Years of Education")
```



Mean Income by Race, Distinguishing for 12 & 16 Education Levels

The tibble below demonstrates the proportion of individual in each race cohort with at least a high school education. This illustrates that 52.1% of non black non hispanic respondents have a high school degree compared to just 38.9% of blacks and 38.5% of hispanics over the course of our truncated data set.

Hide

```
prop_educ_race <-
  income_educ_phys_recent %>%
  group_by(Race) %>%
  summarise(prop_over_12_pct = mean(`Years of Education Completed` > 12)*100)
prop_educ_race
```

Race <chr>	**prop_over_12_pct** <dbl>
black	38.97354
hispanic	38.48546
NBNH	52.16594

3 rows

To drill down a bit further into a higher education subset by race, the below tibble demonstrates the proportion of respndents in each race cohort with at least a bachelors degree. This illustrates that 13.0% of non black non hispanic respondents have a bachelors compared to just 6.3% of blacks and 5.5% of hispanics over the course of our truncated data set.

Hide

```
prop_grad_degree_race <-
  income_educ_phys_recent %>%
  group_by(Race) %>%
  summarise(prop_over_16_pct = mean(`Years of Education Completed` > 16)*100)
prop_grad_degree_race
```

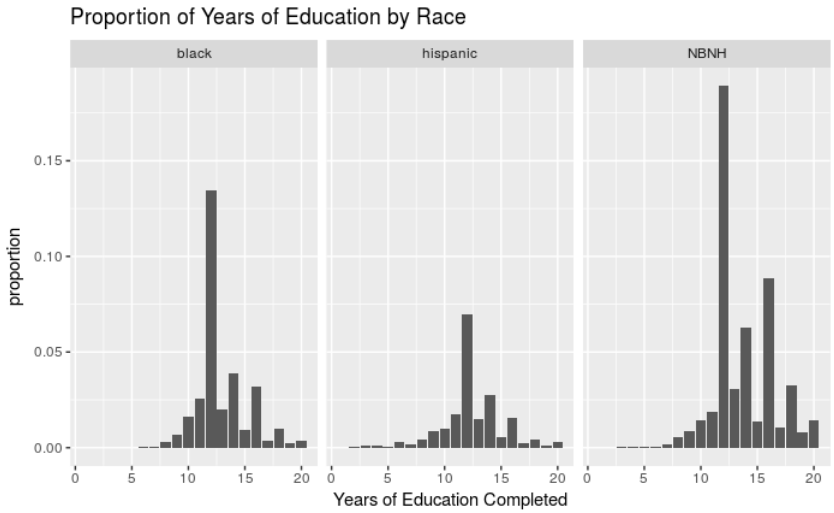
Race <chr>	prop_over_16_pct <dbl>
black	6.369573
hispanic	5.359429
NBNH	12.974460

3 rows

The below plot is a graphical representation of the tibble above. This illustrates the discrepancy in education amongst the different race cohorts. There is a significant spike for all three groups at 12 to signify a high school degree, but there is an almost right tail skew in the non black non hispanic category suggesting a higher proportion of high school, College, and graduate level work. Note this data & trends are consistent with the "Distribution of Education in NLSY79 Dataset, Years 2008-2014" plot above.

Hide

```
ggplot(data = income_educ_phys_recent,
                              aes(x = `Years of Education Completed`)) +
  geom_bar(aes(y = ..count../sum(..count..))) + facet_wrap(~Race) +
  ylab("proportion") +
  labs(title = "Proportion of Years of Education by Race")
```



Let's explore the differences in mean income for each race, based on the fact that 12 years of education seems to be a natural point of reflection. Below there are 3 tibbles. The first 2 explore the difference in mean income by race cohort for those with less than a high school degree and greater than a high school degree. The final tibble combines the first 2 to see the disparity between respondents for each race, for < and >12 years of education.

Hide

```
#Explore income by race for education <12
mean_income_by_race_educ_12_less <-
  filter(income_educ_phys_recent, `Years of Education Completed` <12) %>%
  group_by(Race) %>%
  summarise(`Mean Income` = mean(Income, na.rm = T))
colnames(mean_income_by_race_educ_12_less)[2]  <- "Mean Income < High School"
#Explore income by race for education <12
mean_income_by_race_educ_12_more <-
  filter(income_educ_phys_recent, `Years of Education Completed` >12) %>%
  group_by(Race) %>% summarise(`Mean Income` = mean(Income, na.rm = T))
colnames(mean_income_by_race_educ_12_more)[2]  <- "Mean Income High School+"
#Join the two tibbles and create a new column for percent change
mean_income_by_race_diff <-
  inner_join(mean_income_by_race_educ_12_less, mean_income_by_race_educ_12_more, by = "Race")
mean_income_by_race_diff
```

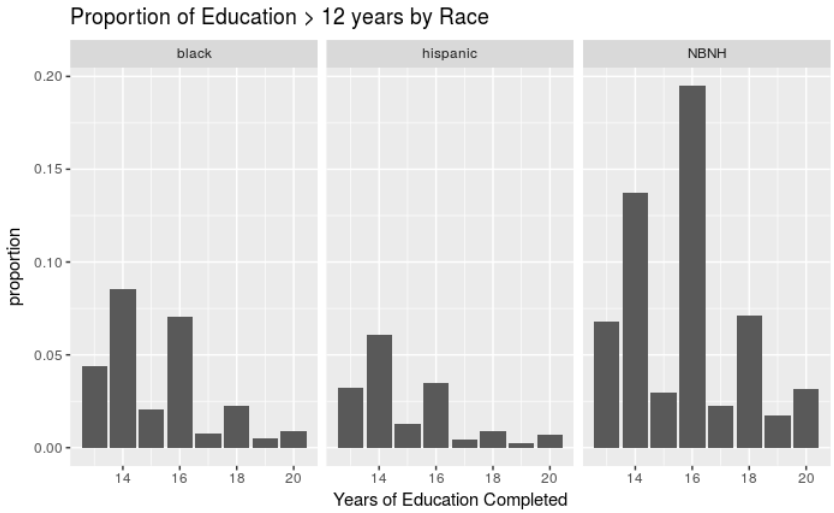
Race <chr>	Mean Income < High School <dbl>	Mean Income High School+ <dbl>
black	11248.40	43750.64
hispanic	17601.97	51839.11
NBNH	19690.49	66911.81

3 rows

The following 2 plots are yet another graphical representation of the proportion of individuals in each race cohort that have a high school degree, less than a high school degree, and a college degree, respectively. As illustrated there is a signifantly higher proportion of NBNH with a college degree, and a signficanlty lower amount of NBNH with less than a high school degree than black or hispanic in the data set. The first plot shows a righ-skew with fewer respondents with greater than 16 years of education, while the second plot shows a left-skew with more respondents completing 12 years of education.

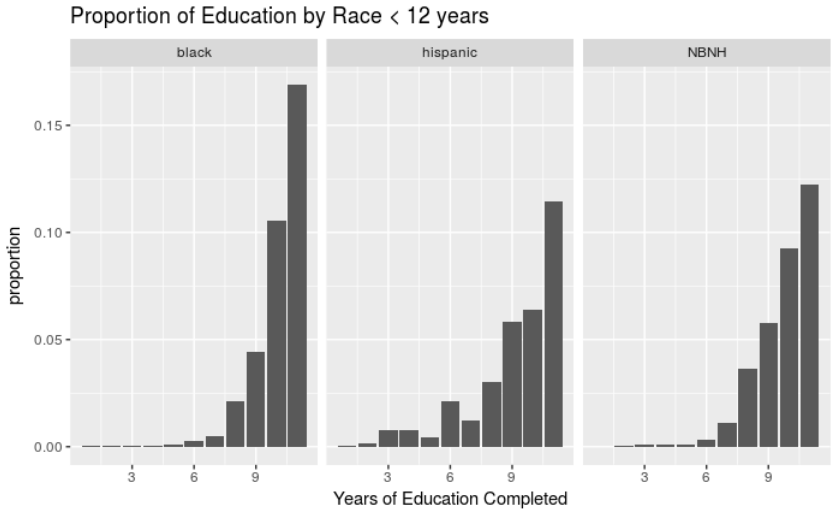
Hide

```
ggplot(
  data = filter(income_educ_phys_recent,
                `Years of Education Completed` > 12),
  aes(x = `Years of Education Completed`)
) + geom_bar(aes(y = ..count../sum(..count..))) + facet_wrap(~Race) +
  ylab("proportion") +
  labs(title = "Proportion of Education > 12 years by Race")
```



Hide

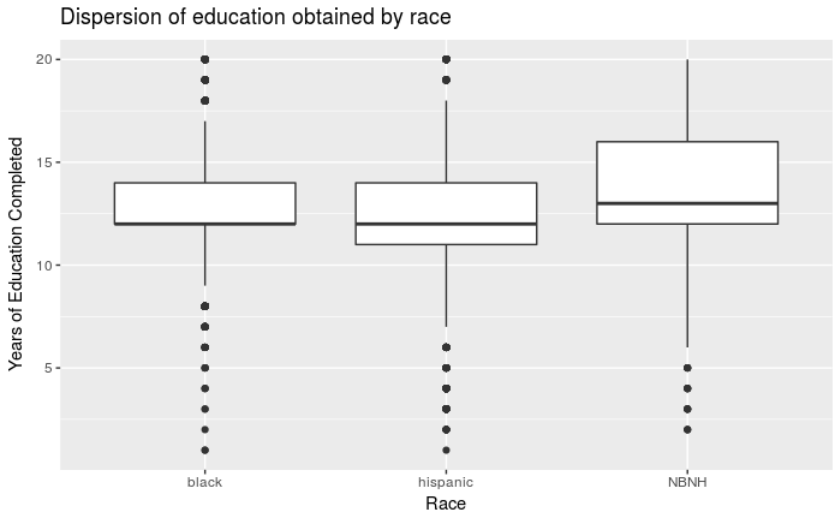
```
ggplot(
  data = filter(income_educ_phys_recent,
                `Years of Education Completed` < 12),
  aes(x = `Years of Education Completed`)
) + geom_bar(aes(y = ..count../sum(..count..))) + facet_wrap(~Race) + ylab("proportion") +
  labs(title = "Proportion of Education by Race < 12 years")
```



The following box plot is an illustration of that skew discussed above between race and education completed. You can seee NBNH is skewed towards greater education in each quartile versus hispanic and / or black.

Hide

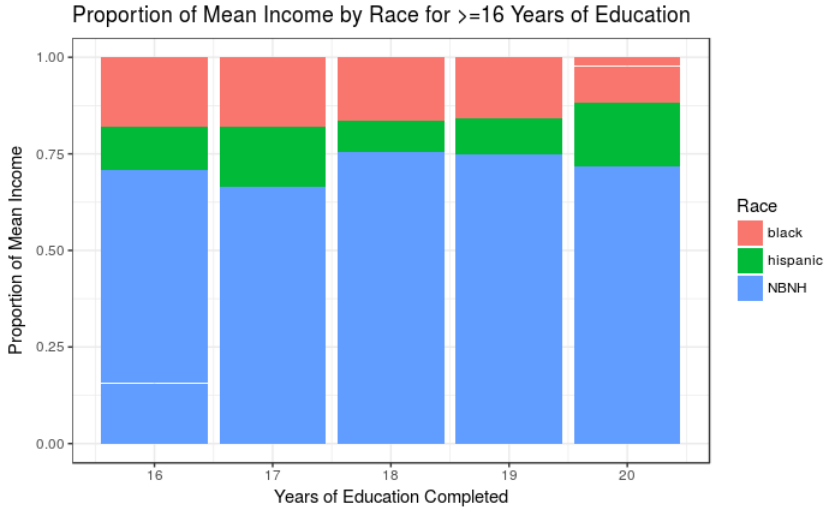
```
#Box Plot
race_box <- ggplot(
  data = filter(income_educ_phys_recent, `Years of Education Completed` < 95),
  aes(x = Race, y = `Years of Education Completed`)
) + geom_boxplot() +
  ggtitle("Dispersion of education obtained by race")
race_box
```



The final stacked bar graph reflects the proportion of mean income by race for those repondents with graduate educations. It is clear that majority of mean income for each set of education years is for the non black non hispanic race group.

Hide

```
# Plot the proportion of mean income by race for respondents with >= 16 years of education completed
ggplot(
  data = filter(income_educ_phys_recent,
                `Years of Education Completed` > 15),
  aes(x = `Years of Education Completed`, y = `Income`, fill = `Race`)) +
  geom_col(position = "fill") + theme_bw() +
  ylab("Proportion of Mean Income") +
  labs(title = "Proportion of Mean Income by Race for >=16 Years of Education")
```



Hypotheses for Further Analysis

Based on the plots and the grouped summary analyses we state the following hypotheses.

1. Mean income is positively correlated with years of education completed.
2. Mean income shows a gender-based disparity, with men having a higher mean income than women, despite more females with college and graduate degrees.
3. NBNH respondents are overall more educated, as more Hispanics & Blacks have less than a high school education and lower income.
4. Mean income shows racial disparity, with NBNH respondents having the highest (and majority) mean income, followed by Hispanics, followed by Blacks. Though the overall respondent mean income rose, only the NBNH segment had consistent mean income growth during the 2008-2014 time period.
5. For respondents with at least 16 years of education, the mean income of men is approximately twice that of women. For that same cohort, the mean income of the NBNH racial group is approximately three times that of Blacks, and approximately four times that of Hispanics.
6. Within the higher educated segment, female mean income has remained stagnant versus an increasing male mean income for the 2008-2014 time period.

Testing the statistical significance of these hypotheses requires statistical analyses to be learned later. Our team would also be interested to analyze other factors such as profession and industry.

Further Data Points for Analysis

During this analysis, it came to our attention that more detail from the respondents on the following topics would create clearer hypotheses:

- Type of college and graduate degrees (law, medical, humanities, etc.)
- GED versus 18-year-old high school graduate for the 12 year educated respondents
- Type of location (urban, suburban, versus country)
- Physical location (city, state)
- Industry (technology, pharmaceuticals, etc.)