EDA on what Charartеristics Correlate with Income

Contents

Introduction

Team Members:

Alex Pesantez

In this document, we survey the data from NLSY '79 and perform exploratory data analysis on the effects on income of combinations of other variables provided.

Our analysis will be to study the effects of education, race, gender and height on levels of income from 1981 to 2014. To begin, we will explore the effects of gender and height as it pertains to inocome.

Study of Income

When looking at the unique years, you can see that the years vary from 1982 to 2014, which means that we have income data for that time frame. When looking at the histogram of the income data, you can clearly see that the data is skewed to the right with some major outliers. While there are definitely some outliers, in some of the analyses these won't be removed because not everything is being modeled and we just want to see the general sense of the trends between the variables, and in real life there are people that have drastically more or less income than others which we want to demonstrate in the analysis.

```
options(scipen = 100)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

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

```
## Rows: 291,778
## Columns: 3
## $ CASEID <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, ~
## $ income <int> NA, 10000, 7000, 1086, 2300, 3250, 4975, 7500, 5000, 9000, 4002~
## $ year   <int> 1982, 1982, 1982, 1982, 1982, 1982, 1982, 1982, 1982, 1982, 198~
```

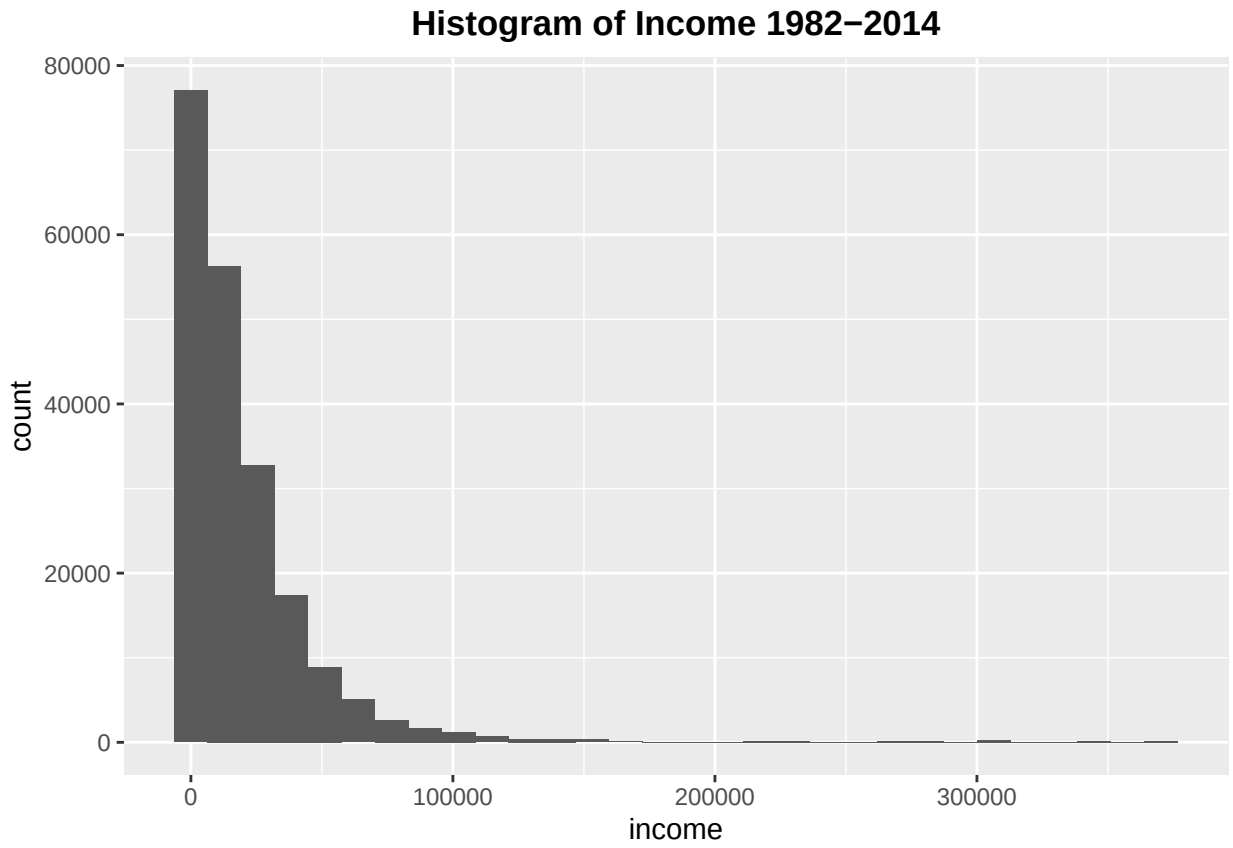
```
unique(income_data_nlsy79$year)
```

```
##  [1] 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1996 1998
## [16] 2000 2002 2004 2006 2008 2010 2012 2014
```

```
ggplot(income_data_nlsy79, aes(income)) + geom_histogram() + ggtitle("Histogram of Income 1982-2014") +
  theme(plot.title = element_text(hjust = 0.5,face = "bold"))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 85628 rows containing non-finite values (`stat_bin()`).
```




```
unique(filter(
  income_data_nlsy79,
  !is.na(income)
  )$year)
```

```
##  [1] 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1996 1998
## [16] 2000 2002 2004 2006 2008 2010 2012 2014
```

Study of Gender and Height

Physical Characteristics and Study of Gender and Height

```
load("physical_data_nlsy79.RData")
unique(physical_data_nlsy79$year)
```

```
##  [1] 1981 1982 1985 1986 1988 1989 1990 1992 1993 1994 1996 1998 2000 2002 2004
## [16] 2006 2008 2010 2012 2014
```

```
sort(unique(physical_data_nlsy79$height))
```

```
##  [1] 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72
## [26] 73 74 75 76 77 78 79 80 81 82 83 84 91
```

```
unique(physical_data_nlsy79$sex)
```

```
## [1] "female" "male"
```

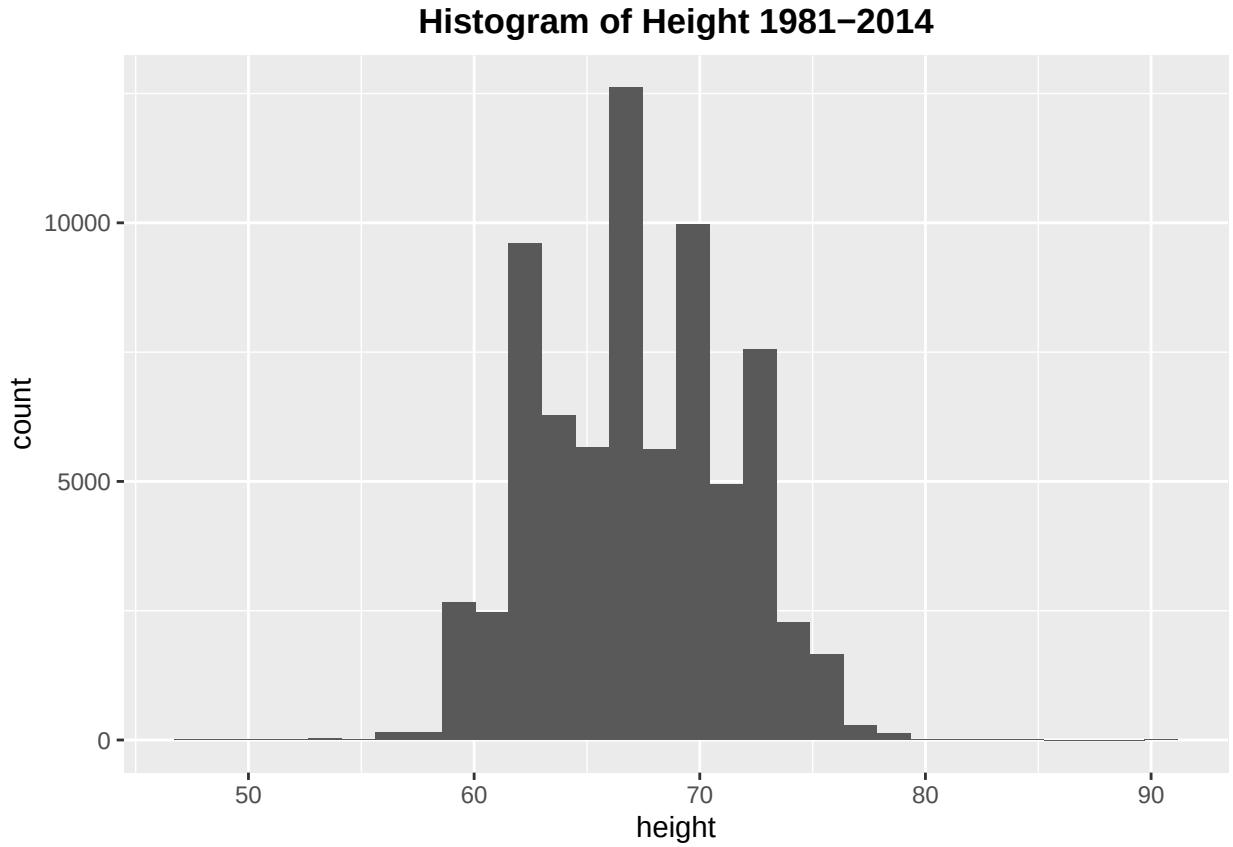
```
unique(filter(
  physical_data_nlsy79,
  !is.na(height)
  )$year)
```

```
## [1] 1981 1982 1985 2006 2008 2010 2012 2014
```

```
ggplot(physical_data_nlsy79, aes(height)) + geom_histogram() + ggtitle("Histogram of Height 1981-2014") +
  theme(plot.title = element_text(hjust = 0.5,face = "bold"))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 181608 rows containing non-finite values (`stat_bin()`).
```




```
unique(filter(
  physical_data_nlsy79,
  !is.na(sex)
  )$year)
```

```
##  [1] 1981 1982 1985 1986 1988 1989 1990 1992 1993 1994 1996 1998 2000 2002 2004
## [16] 2006 2008 2010 2012 2014
```

```
ggplot(physical_data_nlsy79, aes(sex)) + geom_bar() + ggtitle("Histogram of Gender 1981-2014") +
  theme(plot.title = element_text(hjust = 0.5,face = "bold"))
```


```
phys_income_all <- income_data_nlsy79 %>%
  inner_join(physical_data_nlsy79)
```

```
## Joining with `by = join_by(CASEID, year)`
```

```
ggplot(phys_income_all, aes(sex)) + geom_bar() + ggtitle("Histogram of Gender 1982-2014") +
  theme(plot.title = element_text(hjust = 0.5,face = "bold"))
```


```
unique((physical_data_nlsy79$height %in% phys_income_all$height))
```

```
## [1] TRUE
```

When looking at height and gender, we first wanted to see the range and different values there were for each variable. When looking at the unique values for each of the variables, you can see that there are only two genders observed, which are Male and Female, and that the height ranges from 48 inches to 91 inches tall. When looking at the years observed for gender, the values were from 1981 until 2014. When looking at the years for height, the data that was collected was only from 1981-1985 and 2006-2014. Since the years vary between all three variables, when we combine all three variables into a dataset, the only years that will be available are the years that contain the height data. When looking at the bar charts for the amount of each gender in the original data set, you can see that there were slightly more males than female. The bar chart for the combined data set shows the same story, with again more male than female. Then when observing the distribution for height you can see that there is a clear peak between 65 and 70 inches and that the range again varies from lower than 50 inches to greater than 80 inches.

Exploring the effects of Gender on Income

```
nrow(phys_income_all %>% filter(sex == "male" & !is.na(income)))
```

```
## [1] 80012
```

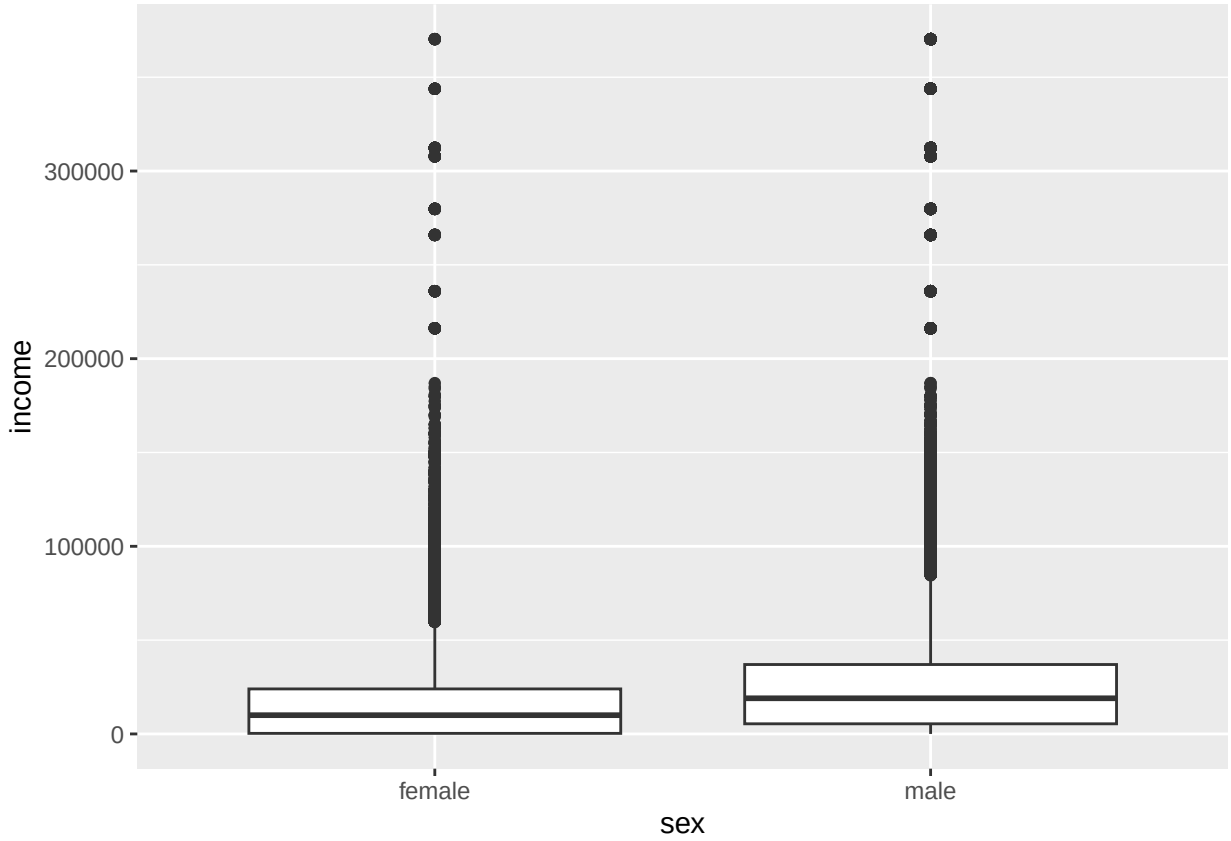
```
nrow(phys_income_all %>% filter(sex == "female" & !is.na(income)))
```

```
## [1] 83059
```

```
ggplot(
  data = phys_income_all,
  aes(x = sex, y = income)
```

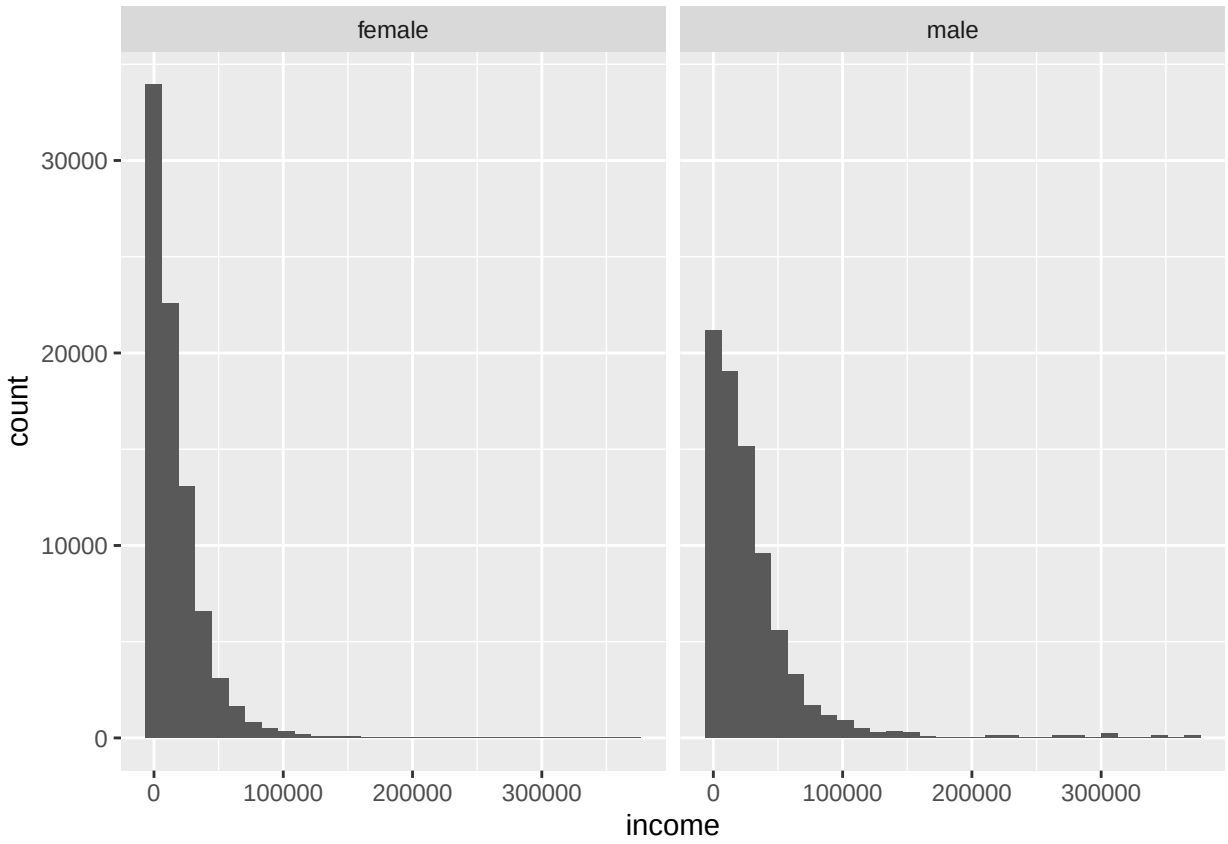
```
) + geom_boxplot()
```

```
## Warning: Removed 77963 rows containing non-finite values (`stat_boxplot()`).
```




```
ggplot(data = phys_income_all, aes(income)) + geom_histogram() + facet_wrap(~sex)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 77963 rows containing non-finite values (`stat_bin()`).
```



When observing the effect of gender on income with a box plot, you can see that there are again some clear outliers. Within this analysis, they will not be removed due to the fact that again we are just observing the trends to get a general sense of how income and gender have related to one another in the past 20 to 30 years. In the box plot, you can see that the mean for male's is greater than the mean for female's which means in the past 30 years in general, males have had more income than females. The next thing that we observed was the distributions of income for each gender. When looking at both of the histograms, you can see that there is again right skewness, however there are higher counts of higher incomes for males when compared to females, which makes sense given that they have a higher mean income within the data set. Another clear difference between the two histograms is that female's had a lot more observations of zero income when compared to the males. This could be due to the fact that in earlier years, women were stereotyped as "stay-at-home moms", and it could've been harder for them to get a well payed job back then.

Exploring the effects of Gender and Year on Income

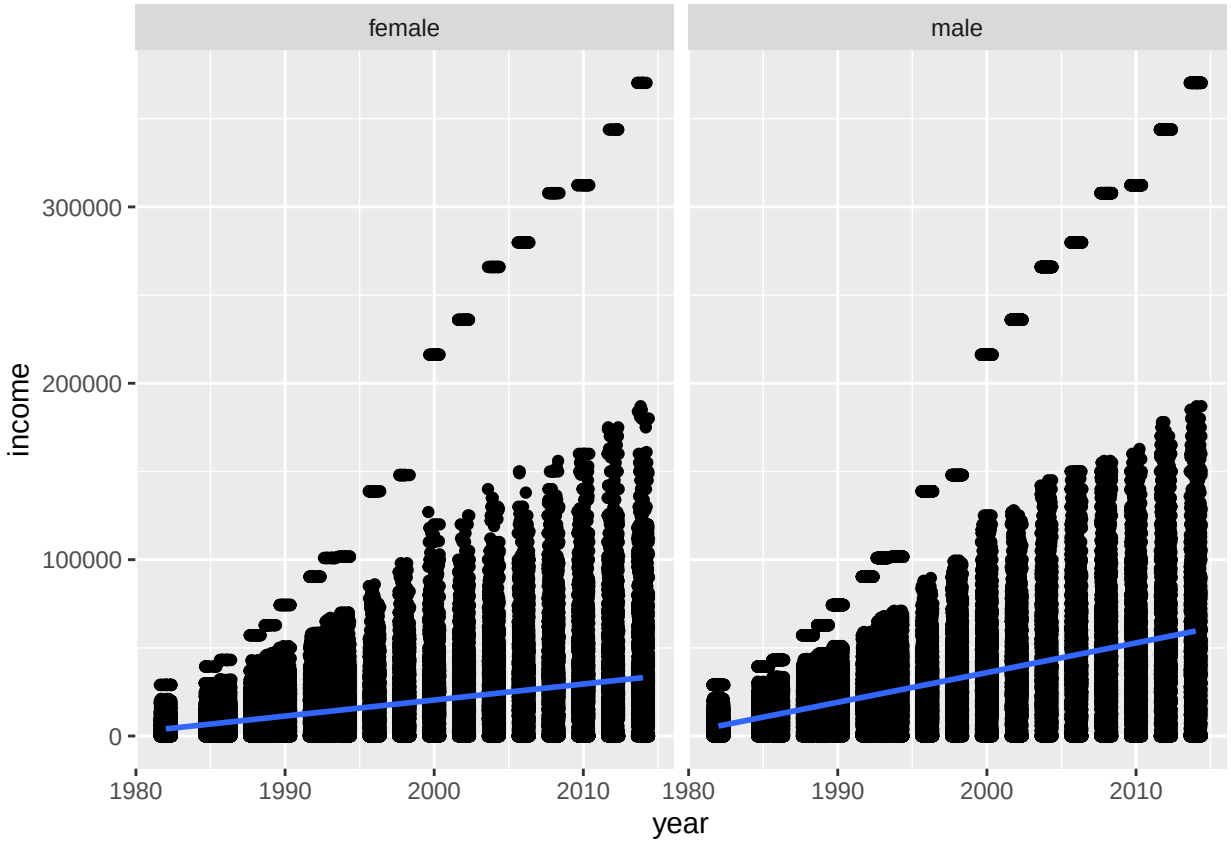
The next thing we wanted to look at was the effect of gender and year on income. In order to do this, we made a graph which was broken up between female and male with year on the x-axis and income on the y-axis. When looking at this graph, you can see that for both genders as time went on the higher the income there was for each gender. This could just be due to inflation but still interesting to theorize. Another observation that can be made from this, is that the slope for the male's is greater than the slope for female's. This could mean that male's have a higher ceiling or potential when searching for a job with a high income than a female.

```
ggplot(data = phys_income_all, aes(year,income)) +
  geom_jitter() + geom_smooth(method = "lm") + facet_wrap(~sex)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 77963 rows containing non-finite values (`stat_smooth()`).
```

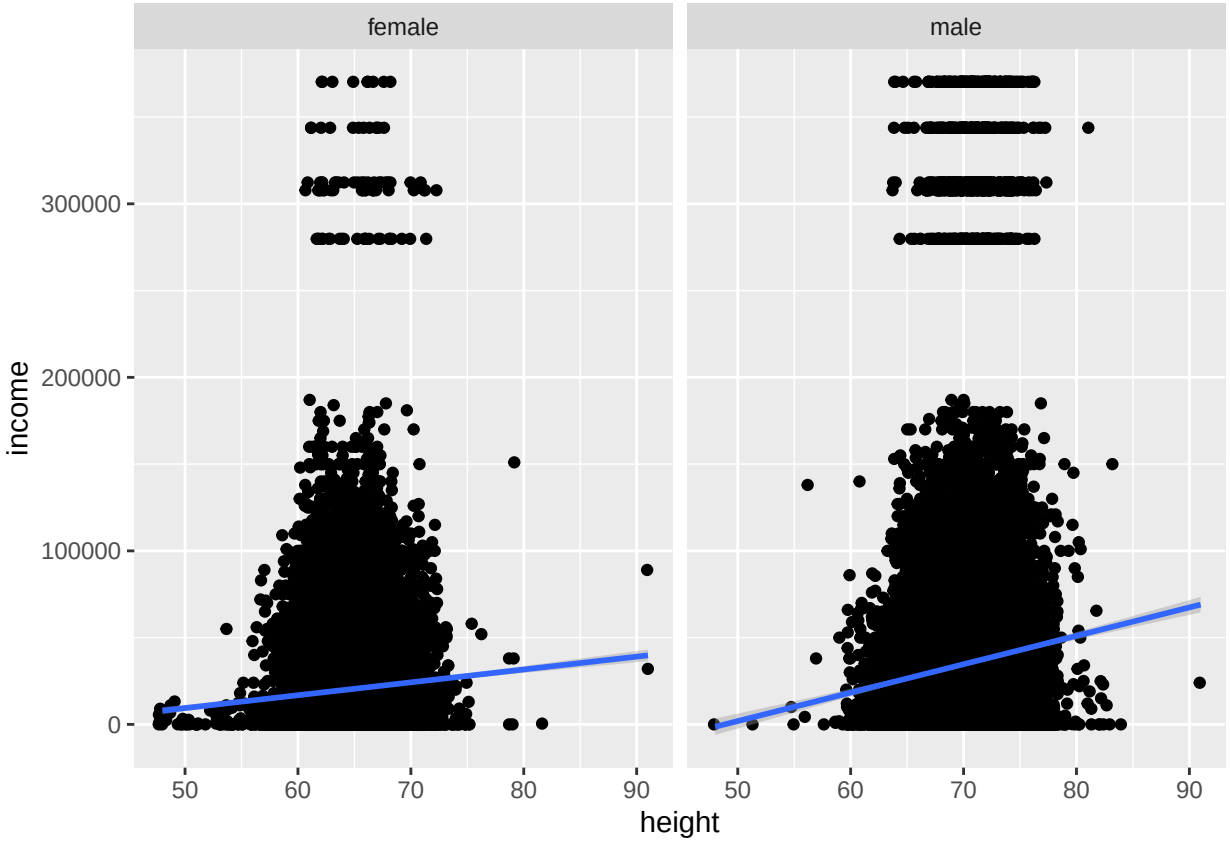
```
## Warning: Removed 77963 rows containing missing values (`geom_point()`).
```



Exploring the combined effects of Gender and Height on Income

```
ggplot(data = phys_income_all, aes(height,income)) + geom_jitter() +
  geom_smooth(method = "lm") + facet_wrap(~sex)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## Warning: Removed 183806 rows containing non-finite values (`stat_smooth()`).
## Warning: Removed 183806 rows containing missing values (`geom_point()`).
```



The next thing explored was the combined effects of gender and height on income. Both of these graphs show a general trend that no matter what gender you are, the taller you the more income you will get. You can see this by the positive slopes for both male's and female's. Another takeaway you can see when looking at both graphs is that the slope for the male's is clearly greater than the slope for the female's, this means that being taller as a man is more impactful on income than it is being a taller female. This could be due to the fact people could find taller male's more intimidating than taller females which leads to them getting a higher income. This could also just be due to the fact that Male's have just made a lot more income in the data set skewing the slope to be greater just due to the fact of them having a greater income in general.

Mean income by gender

Here you can see that male's on average made 12189.38 more than female's in the data observed.

```
filter(phys_income_all) %>%
  group_by(sex) %>%
  summarise(mean_income = mean(income, na.rm = T))
```

```
## # A tibble: 2 x 2
##   sex    mean_income
##   <chr>        <dbl>
## 1 female      16661.
## 2 male        28850.
```

Mean income by height

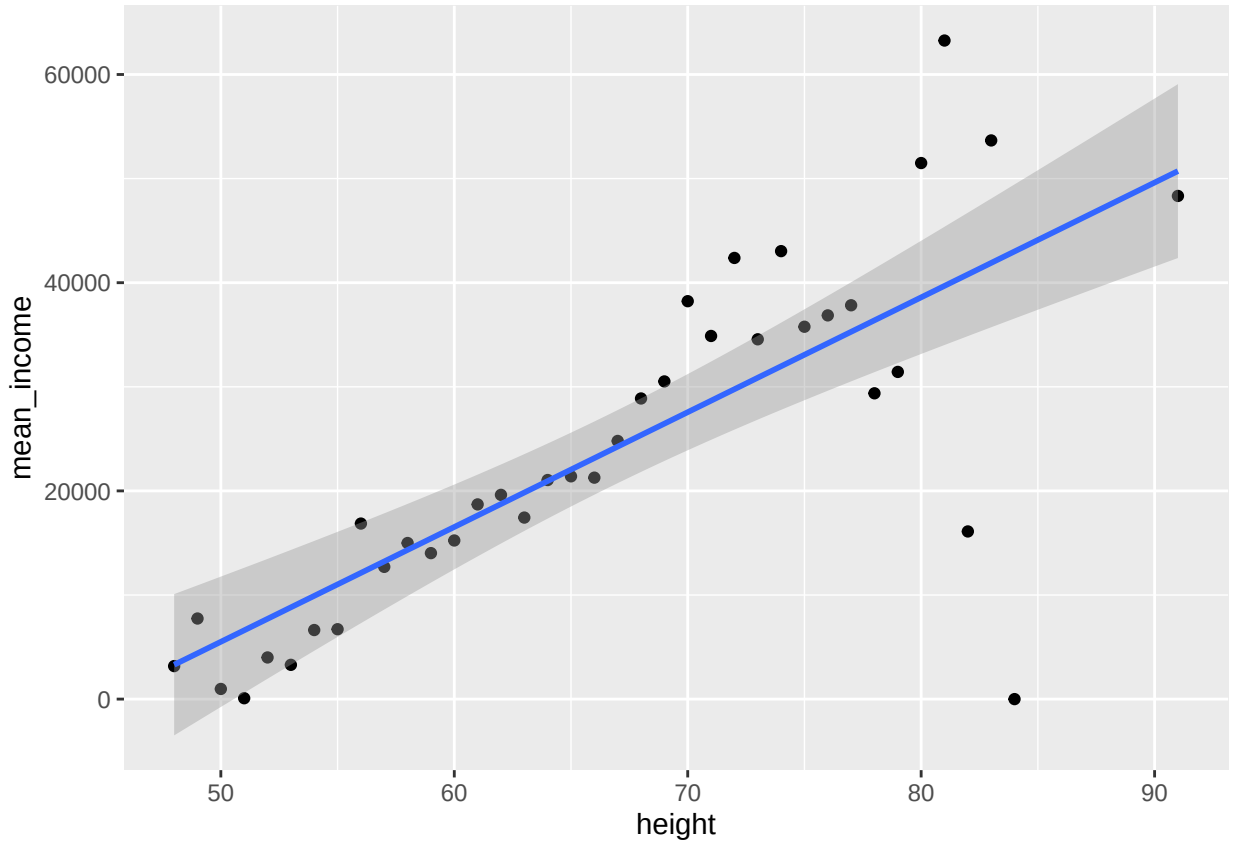
Here you can see that on average, the taller you are, the more income you will get. This seems to be true for both female and male which was observed earlier in this analysis.

```r
height_income <- filter(phys_income_all) %>%
 group_by(height) %>%
 summarise(mean_income = mean(income, na.rm = T))
ggplot(height_income, aes(height,mean_income)) + geom_point() + geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



Mean income by height and gender

```r
filter(phys_income_all) %>%
  group_by(height, sex) %>%
  summarise(mean_income = mean(income, na.rm = T))
```

```
## `summarise()` has grouped output by 'height'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 68 x 3
## # Groups:   height [39]
##    height sex    mean_income
##     <int> <chr>        <dbl>
##  1     48 female       3623.
##  2     48 male            0
##  3     49 female       7740
##  4     50 female        976.
```

```
##  5     51 female        100
##  6     51 male            0
##  7     52 female       4000
##  8     52 male          NaN
##  9     53 female       3291.
## 10     54 female       6634.
## # i 58 more rows
```

Here you can observe the difference of income between male and female at each height. You can generally see that as the heights get bigger, so does the income. It's also interesting to see the difference between incomes when a male and female are the same height.

Following our study of gender and height, we will begin to dive into our exploratory data analysis involving education and race.

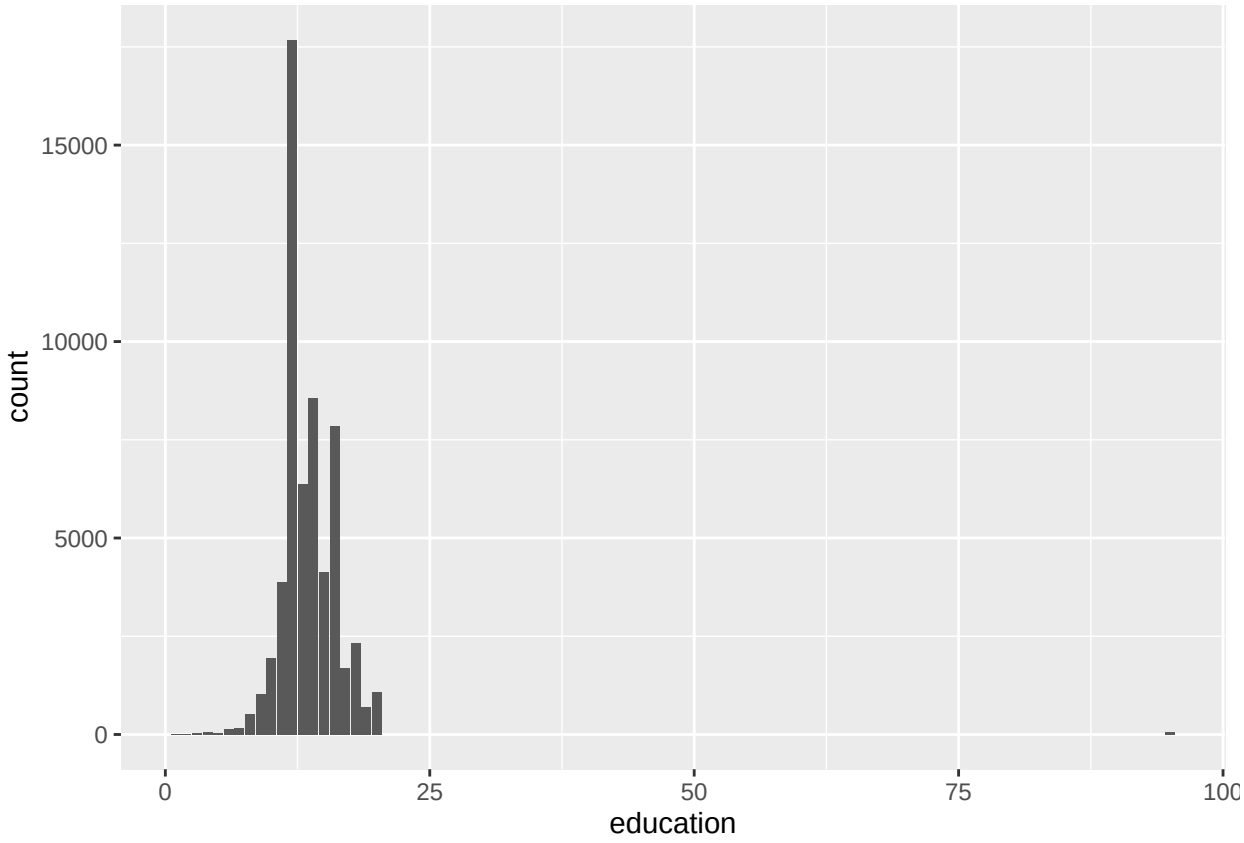
Study of Education

```
library(tidyverse)
load("education_data_nlsy79.RData")
load("income_data_nlsy79.RData")
load("physical_data_nlsy79.RData")
educ_income <- education_data_nlsy79 %>% inner_join(income_data_nlsy79)
```

```
## Joining with `by = join_by(CASEID, year)`
```

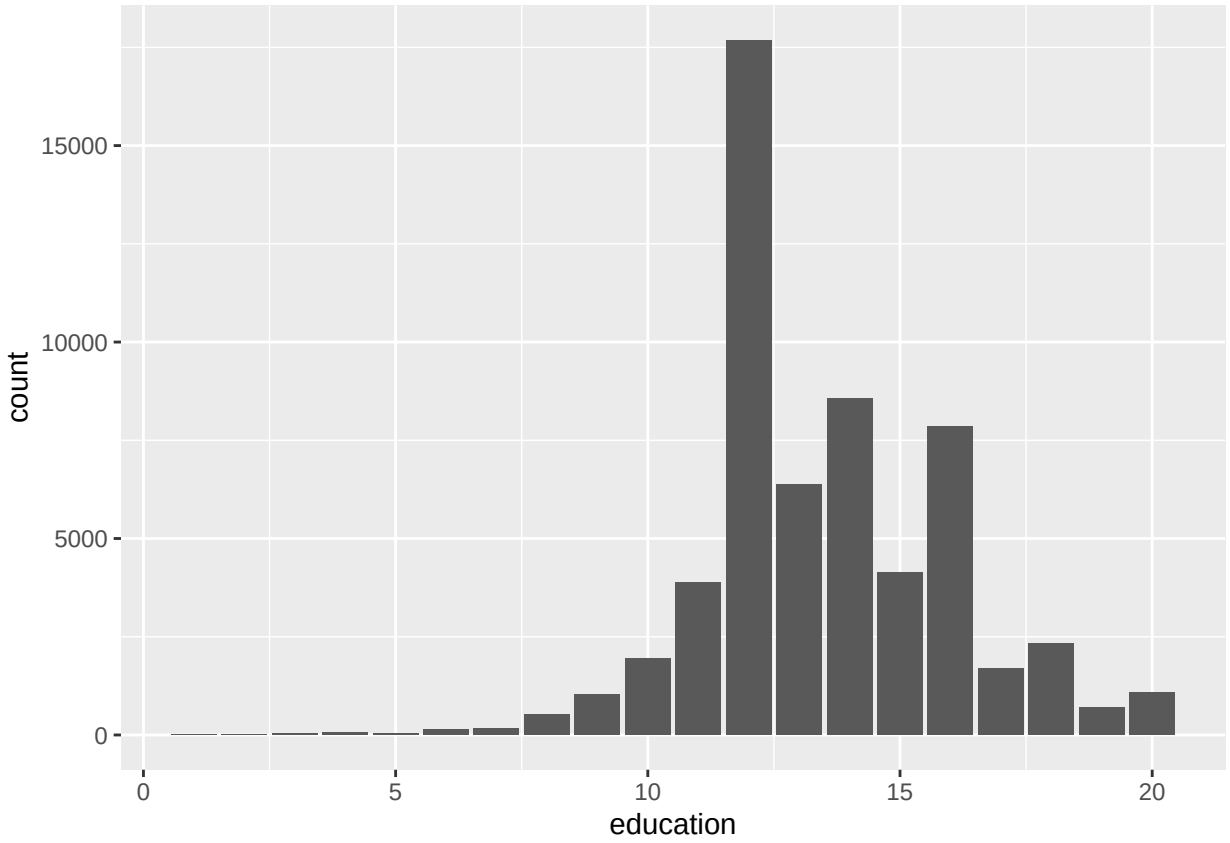
```
ggplot(data = educ_income, aes(x=education)) + geom_bar()
```

```
## Warning: Removed 233424 rows containing non-finite values (`stat_count()`).
```



Let's exclude the value 95, which means there are some ungraded education, from our analysis so that we can clearly see the histogram of education values between 0 and 25.

```
ggplot(data = filter(educ_income, education < 95), aes(x=education)) + geom_bar()
```



There are local peaks in the distribution at 12 and 16 years for high school and college degress, 14 years for associate degrees, and 18 years for an MBA or other 2-year master's degree.

Between 1981 and 2014 there are 58K cases of educational data.

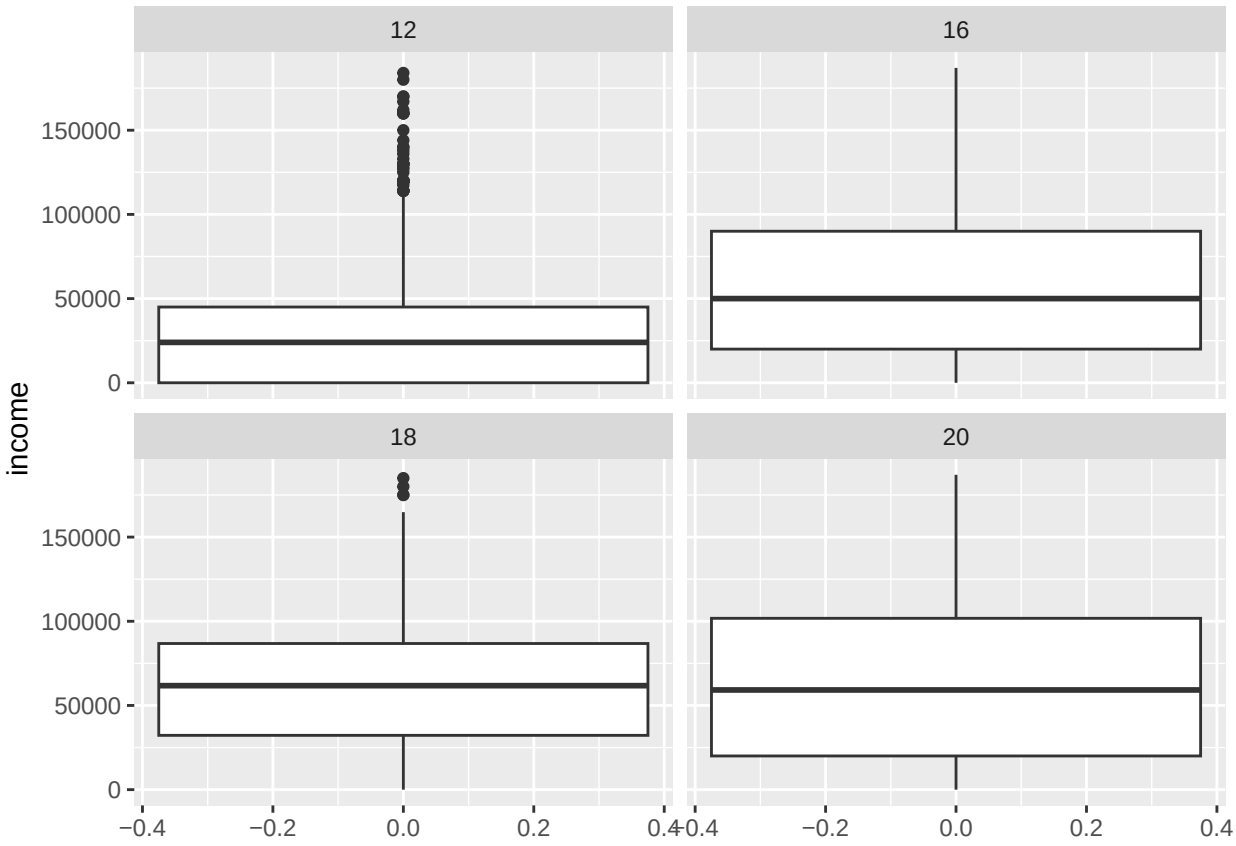
```
sum(!is.na(educ_income$education))
```

```
## [1] 58354
```

Explore the effect of Education on Income over time.

We'll first look at the year 2014 and the distribution of income levels based on 4 education levels: High School, College, 2-year Masters, and 4-year post-undergraduate degree. There are some outliers above the $300K income level so in order to clearly see the differences, we removed the outliers.

```
ggplot(
  data = filter(educ_income, year==2014, education==12 | education==16 | education==18 | education==20,
  aes(y=income)
) + geom_boxplot() + facet_wrap(~education)
```



It's clear that a college degree gives a significant boost in income over a high school degree with an increase in the median income of close to 25K. The upper quartile also saw a significant increase of nearly double in income. Masters and Post-graduate income levels are not that significantly different from college degree income levels.

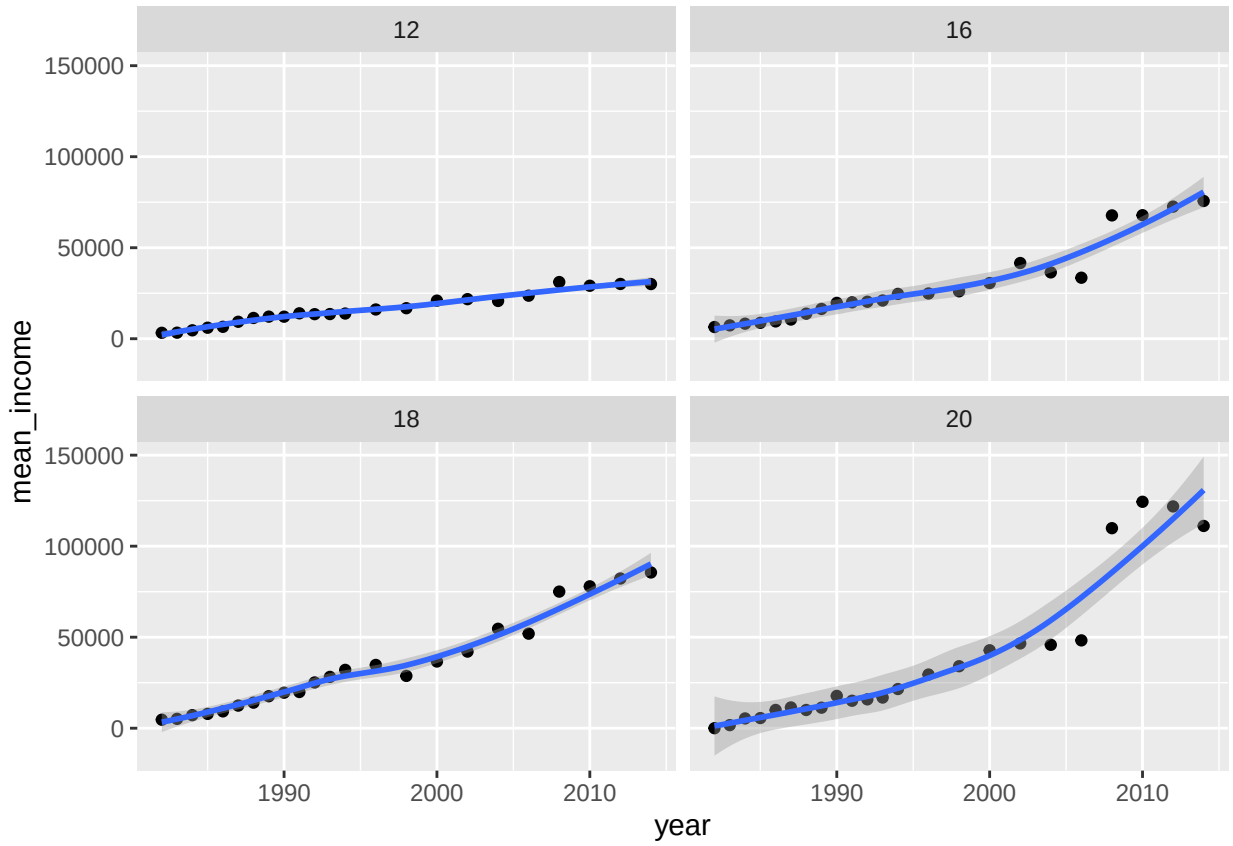
To understand how income levels change over time based on the education level the person achieved, we will be looking at how the mean income level per year for each education level behaves.

```
mean_inc_educ_timeseries <- educ_income %>%
  filter(education == 12|education==16|education==18|education==20) %>%
  group_by(education, year) %>%
  summarise(mean_income = mean(income, na.rm=T))
```

```
## `summarise()` has grouped output by 'education'. You can override using the
## `.groups` argument.
```

```
ggplot(
  mean_inc_educ_timeseries,
  aes(x=year,y=mean_income))+geom_point()+geom_smooth()+facet_wrap(~education)
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



This plot clearly demonstrates that with better education your income potential over time is much higher. We can see that with education 20 (4-year post-graduate degrees) in more recent times income has spiked, because presumably the degree is much more valuable and therefore can demand a higher income. Whereas with a high-school degrees, income levels stay relatively stagnant over time.

Mean Income by Education

```
educ_income %>%
  filter(education < 95, education == 12|education==16|education==18|education==20) %>%
  group_by(education) %>%
  summarise(mean_income = mean(income, na.rm=T))
```

```
## # A tibble: 4 x 2
##   education mean_income
##       <int>       <dbl>
## 1        12      22161.
## 2        16      43038.
## 3        18      56266.
## 4        20      78044.
```

Linear Model of Mean Income over time for High-School vs College

```
summary(lm(year~mean_income, data=subset(mean_inc_educ_timeseries,education==12)))
```

```
##
## Call:
## lm(formula = year ~ mean_income, data = subset(mean_inc_educ_timeseries,
```

```
##     education == 12))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3452 -1.1128  0.3419  0.9405  3.1449
##
## Coefficients:
##                  Estimate    Std. Error t value            Pr(>|t|)
## (Intercept) 1977.87384239    0.80428099 2459.18 <0.0000000000000002 ***
## mean_income    0.00110704    0.00004468   24.77 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.838 on 21 degrees of freedom
## Multiple R-squared:  0.9669, Adjusted R-squared:  0.9653
## F-statistic: 613.8 on 1 and 21 DF,  p-value: < 0.00000000000000022
```

```
summary(lm(year~mean_income, data=subset(mean_inc_educ_timeseries,education==16)))
```

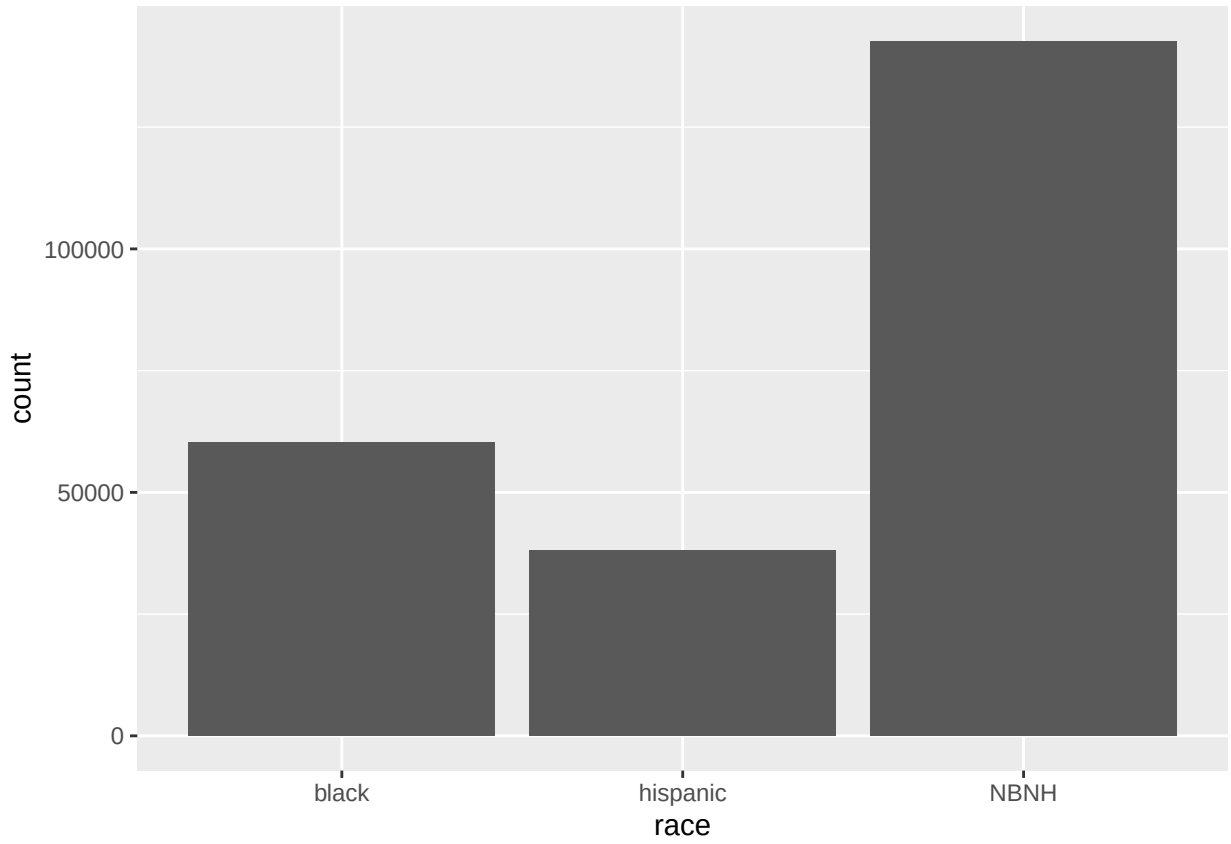
```
##
## Call:
## lm(formula = year ~ mean_income, data = subset(mean_inc_educ_timeseries,
##     education == 16))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0151 -1.9287 -0.9444  1.0620  8.6141
##
## Coefficients:
##                  Estimate    Std. Error t value             Pr(>|t|)
## (Intercept) 1983.06168162    1.09833124    1806 < 0.0000000000000002 ***
## mean_income    0.00042736    0.00003054      14     0.00000000000407 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.145 on 21 degrees of freedom
## Multiple R-squared:  0.9032, Adjusted R-squared:  0.8986
## F-statistic: 195.9 on 1 and 21 DF,  p-value: 0.000000000004067
```

Study of Race

```
library(tidyverse)
race_income <- physical_data_nlsy79 %>% inner_join(income_data_nlsy79) %>% subset(select=c("CASEID","yea
```

```
## Joining with `by = join_by(CASEID, year)`
```

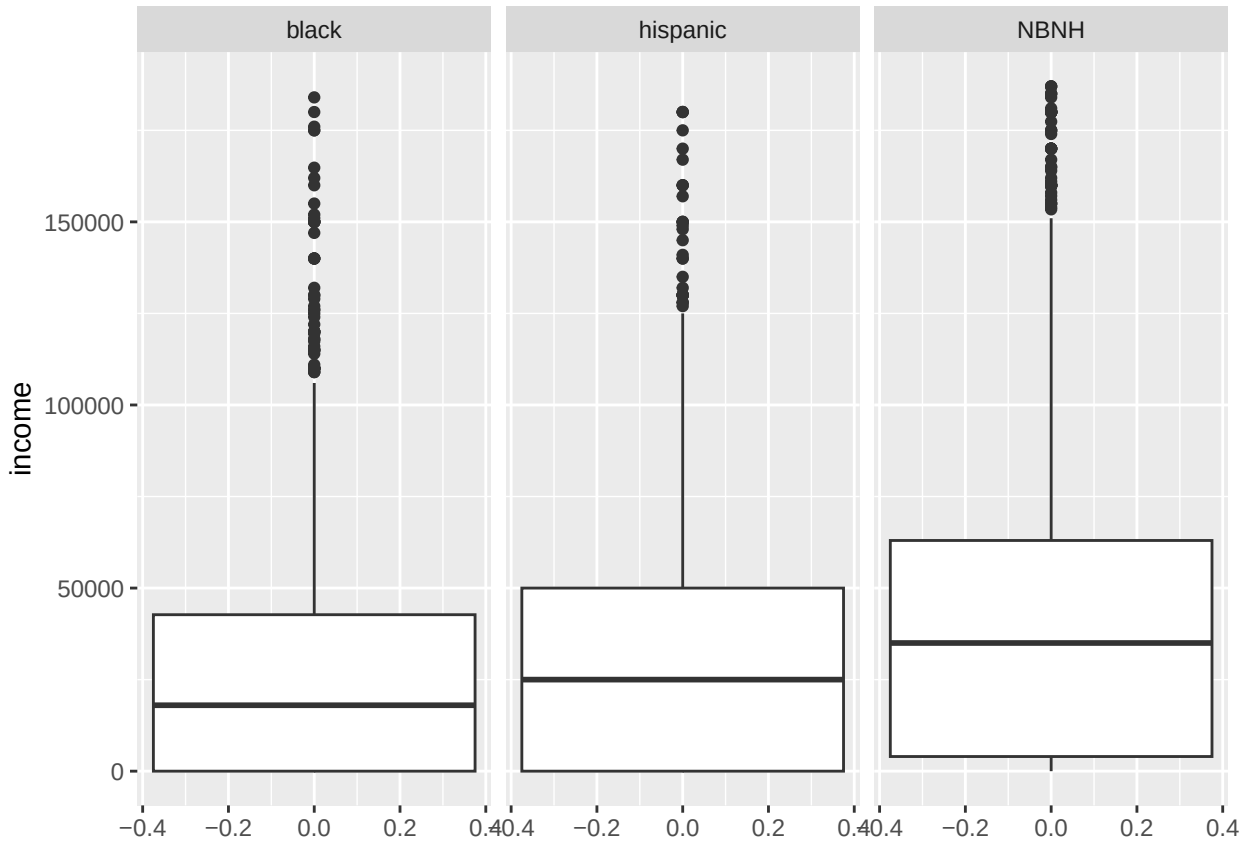
```
ggplot(
  data = race_income,
  aes(x=race)
) +geom_bar()
```



Explore the effect of Race on Income over time.

We'll first look at the year 2014 to initially determine how race influences income levels, removing the outliers of $300k income levels.

```
ggplot(
  data = filter(race_income, year==2014, income < 300000),
  aes(y=income)
) + geom_boxplot() + facet_wrap(~race)
```



This plot indicates to us that in the year 2014, NBNH population earns higher incomes than Black/Hispanics. We can tell since the median and upper quartile of NBNH are noticeably higher than that of the other 2 races.

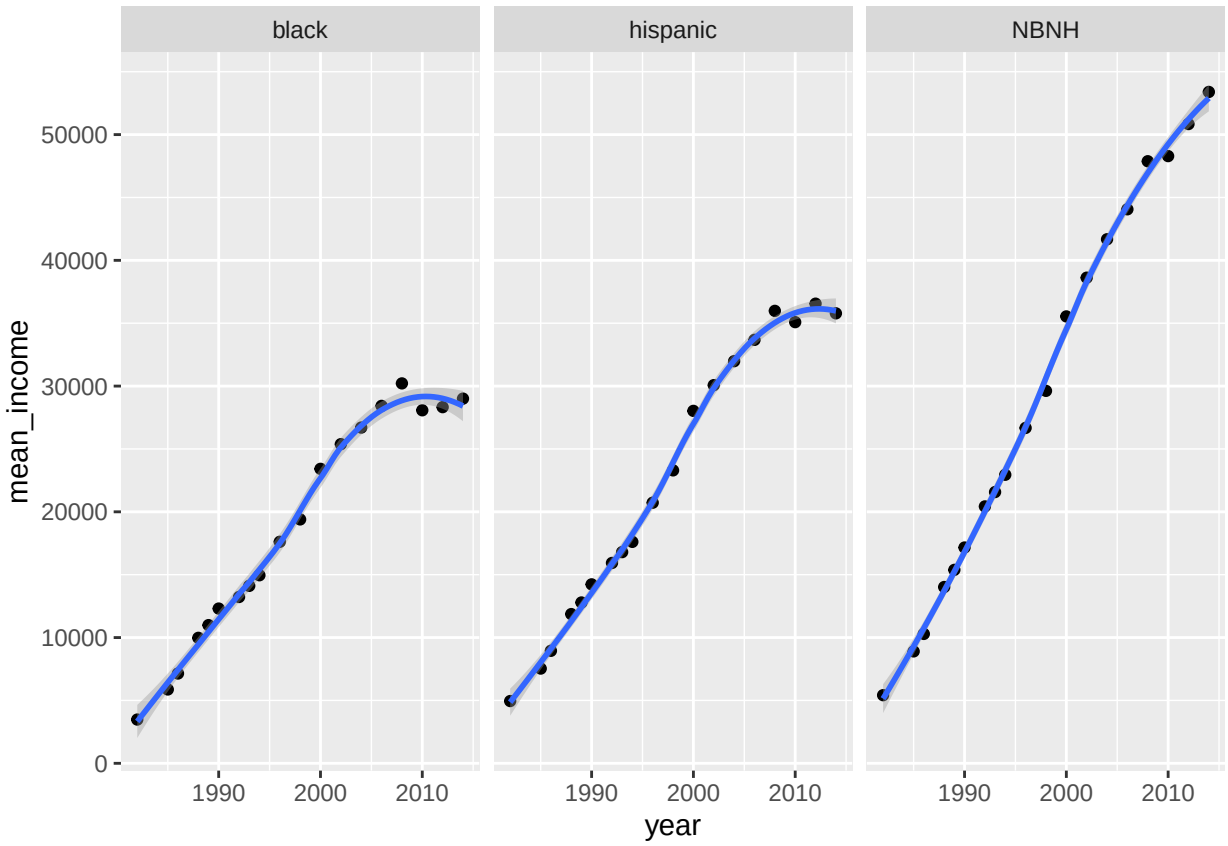
Now we will study if being a particular race affect your income potential over time, by doing a time-series analysis of mean income levels by race.

```
mean_inc_race_timeseries <- race_income %>%
  group_by(race, year) %>%
  summarise(mean_income = mean(income, na.rm=T))
```

```
## `summarise()` has grouped output by 'race'. You can override using the
## `.groups` argument.
```

```
ggplot(
  mean_inc_race_timeseries,
  aes(x=year,y=mean_income))+geom_point()+geom_smooth()+facet_wrap(~race)
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



This plot shows us that over time being Black or Hispanic, your income potential plateaus between 30k and 40k. Whereas for NBNH, income potential doesn't really seem to be limited at all.

Mean Income by Race

```
race_income %>% group_by(race) %>% summarise(mean_income = mean(income, na.rm=T))
```

```
## # A tibble: 3 x 2
##   race     mean_income
##   <chr>          <dbl>
## 1 NBNH          25855.
## 2 black         17528.
## 3 hispanic      21183.
```

Mean Income by Race and Year

```
race_income %>% group_by(race, year) %>% summarise(mean_income = mean(income, na.rm=T))
```

```
## `summarise()` has grouped output by 'race'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 57 x 3
## # Groups:   race [3]
##    race   year mean_income
##    <chr> <int>       <dbl>
##  1 NBNH   1982       5418.
##  2 NBNH   1985       8900.
##  3 NBNH   1986      10290.
```

```
##  4 NBNH   1988      14019.
##  5 NBNH   1989      15383.
##  6 NBNH   1990      17161.
##  7 NBNH   1992      20428.
##  8 NBNH   1993      21572.
##  9 NBNH   1994      22946.
## 10 NBNH   1996      26672.
## # i 47 more rows
```

Study of Race and Education Combined

Setting up the data

The next variable that we hope to explore is the relationship between education and race. As studied above, the variables of education and race are shown how they act independently but how do they effect one another. To begin, we must join the data sets of education and race to be able to study the impact of the two variables.

```
educ_race <- educ_income %>% inner_join(race_income)
```

```
## Joining with `by = join_by(CASEID, year, income)`
```

This data frame will provide us with one main data frame to study the relationship between the two variables.

Similar to the prior dataset, it is necessary to get rid of education values that are 95 or above as those values are clearly outliers.

```
educ_race <- educ_race %>% filter(education <95)
```

Similarly, we will exclude all NA values for race as we cannot set a median value or any other estimate for a persons race in this data set.

```
 sum(is.na(educ_race$race))
```

```
## [1] 0
```

Using the function above, we can see that our data does not have any NA values for race which allows us to proceed.

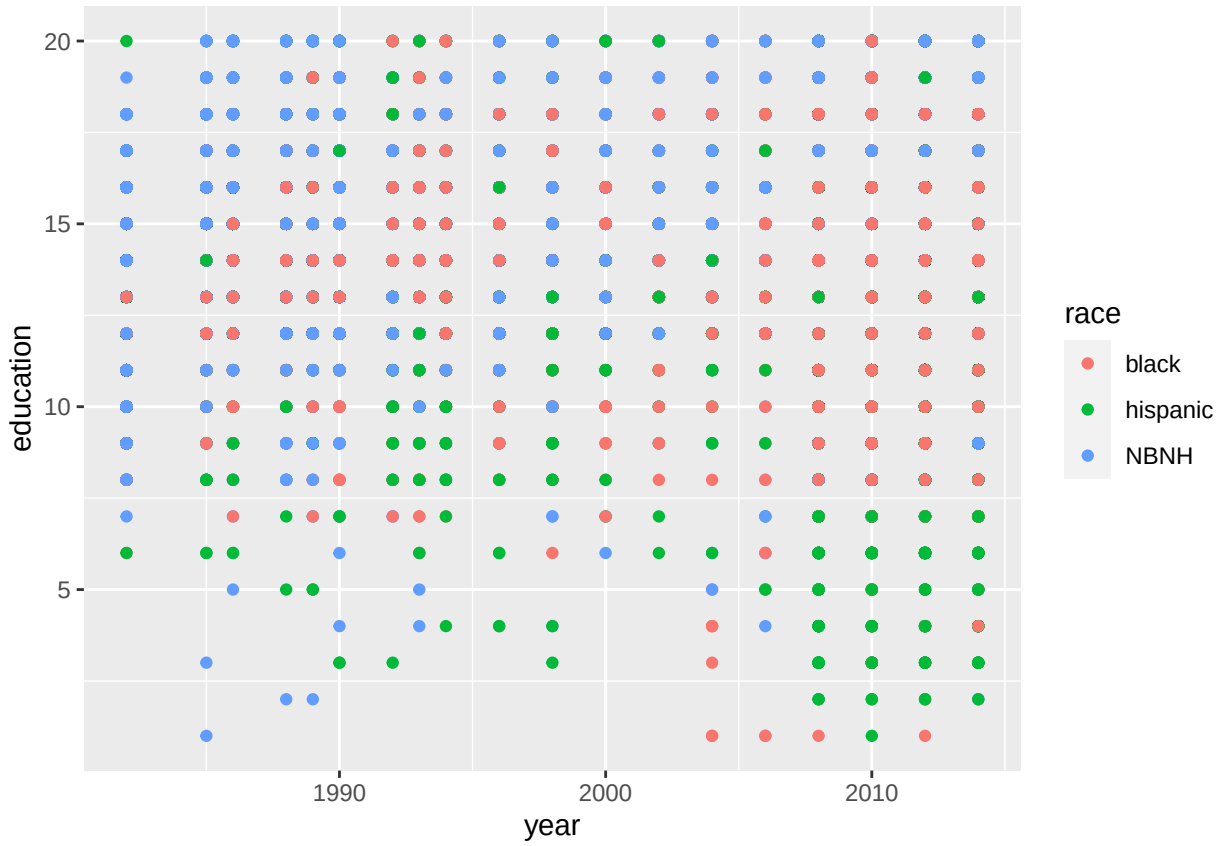
```
sum(is.na(educ_race$education))
```

```
## [1] 0
```

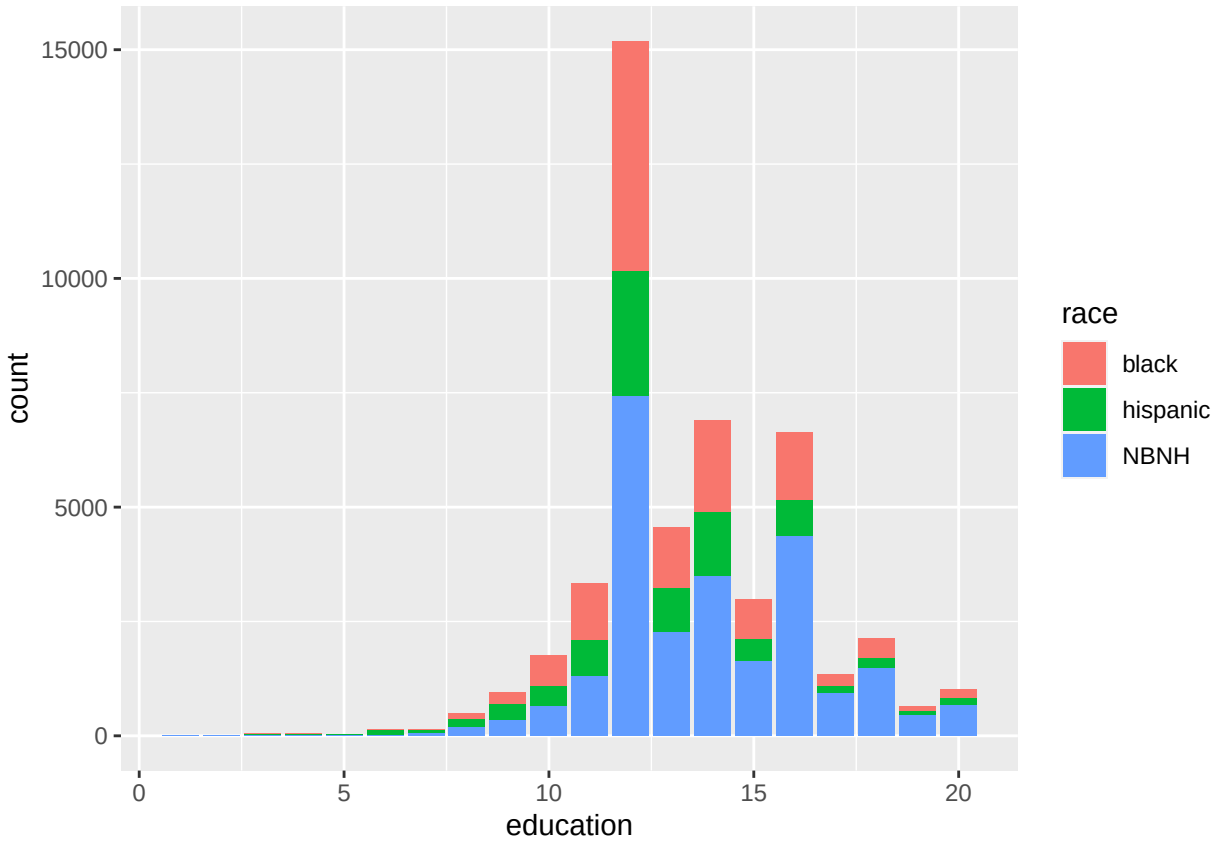
Similarly, we can see that for this new dataset, we have a corresponding race for each variable of education with no missing values present.

Visuals of the present data

```
ggplot(data=educ_race, mapping=aes(x=year,y=education, color=race ))+geom_point()
```




```
ggplot(data=educ_race, mapping=aes(x=education, fill=race))+geom_bar()
```



According to the data, it appears that NBNH participants of the NLSY79 survey make up the majority of the present values for all education levels. This can clearly be seen in the histogram above as the percentage of blue in comparison to the other two colors is much greater.

Clearly there is some correlation between these two variable but what is the extent. Since race is a non numeric variable, let's use a linear model to understand the data better.

```
model <-lm(education~race, data=educ_race)
summary(model)
```

```
##
## Call:
## lm(formula = education ~ race, data = educ_race)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.917  -1.917  -0.783   2.083   7.217
##
## Coefficients:
##              Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  13.11730    0.02100 624.767 <0.0000000000000002 ***
## racehispanic -0.33428    0.03376  -9.903 <0.0000000000000002 ***
## raceNBNH      0.80012    0.02624  30.496 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.501 on 48419 degrees of freedom
## Multiple R-squared:  0.03519,    Adjusted R-squared:  0.03515
```

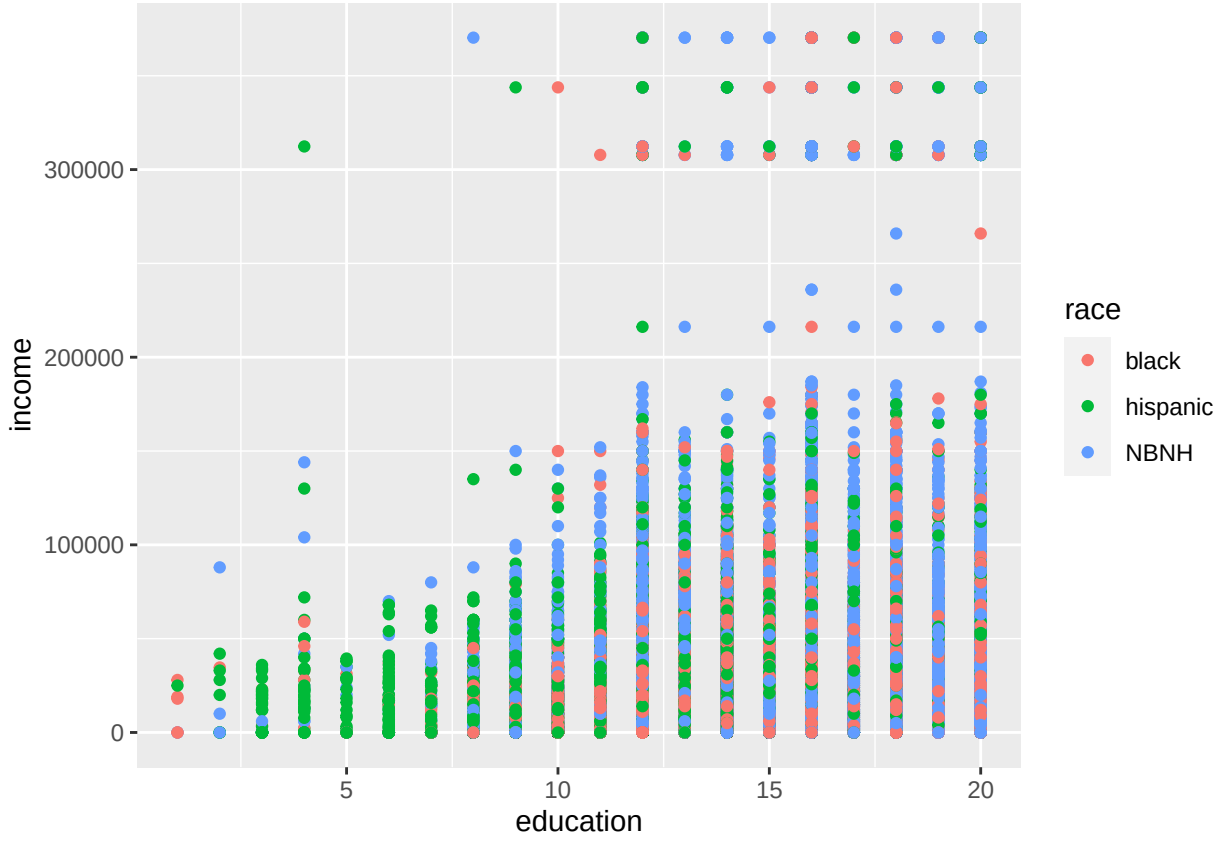
```
## F-statistic: 883.1 on 2 and 48419 DF,  p-value: < 0.00000000000000022
```

According to the T-values there is a significant correlation between race and education. However, we want to better understand the relationship between income and race and eduation.

Let's begin by plotting out data.

```
ggplot(data=educ_race, mapping=aes(x=education,y=income, color=race))+geom_point()
```

```
## Warning: Removed 2424 rows containing missing values (`geom_point()`).
```



As shown in the data above, as education levels increase it appears that income increases as well. In terms of race, it appears that NBNH is currently found to have more years of education as opposed to those who are black or hispanic. Those who participated in the survey who are black do appear to have the second most frequent responses in terms of years of education greater than 12. These higher education levels directly correlate to greater income and are seen to be possessed more frequently by those who are NBNH and black as opposed to the hispanic survey participants.

```
model2<- lm(income~race + education, data=educ_race)
summary(model2)
```

```
##
## Call:
## lm(formula = income ~ race + education, data = educ_race)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -64018 -24166 -10176  12519 365033
##
## Coefficients:
```

```
##                Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  -41676.61    1148.48  -36.29 <0.0000000000000002 ***
## racehispanic   6315.70     610.38   10.35 <0.0000000000000002 ***
## raceNBNH       7799.91     479.49   16.27 <0.0000000000000002 ***
## education      4894.75      82.32   59.46 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43950 on 45994 degrees of freedom
##   (2424 observations deleted due to missingness)
## Multiple R-squared:  0.08344,    Adjusted R-squared:  0.08338
## F-statistic:  1396 on 3 and 45994 DF,  p-value: < 0.00000000000000022
```

```
library(car)
```

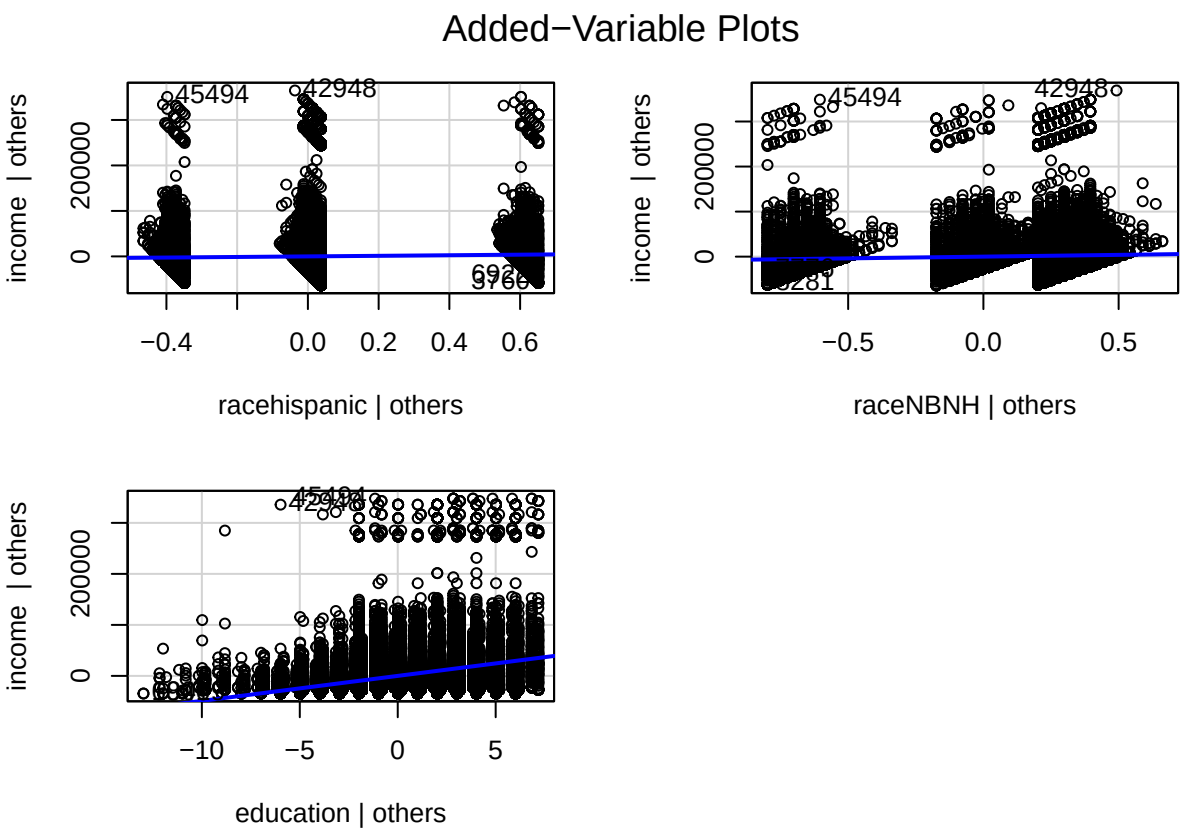
```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
avPlots(model2)
```



These graphs simply show the predictor variable on the x-axis and the y-variable of income as the response

variable. The blue lines shows the assocation between them. Here we can see that income is fairly consistent across races but has a slight postive correlation in terms of education.

Mean Income of the Data

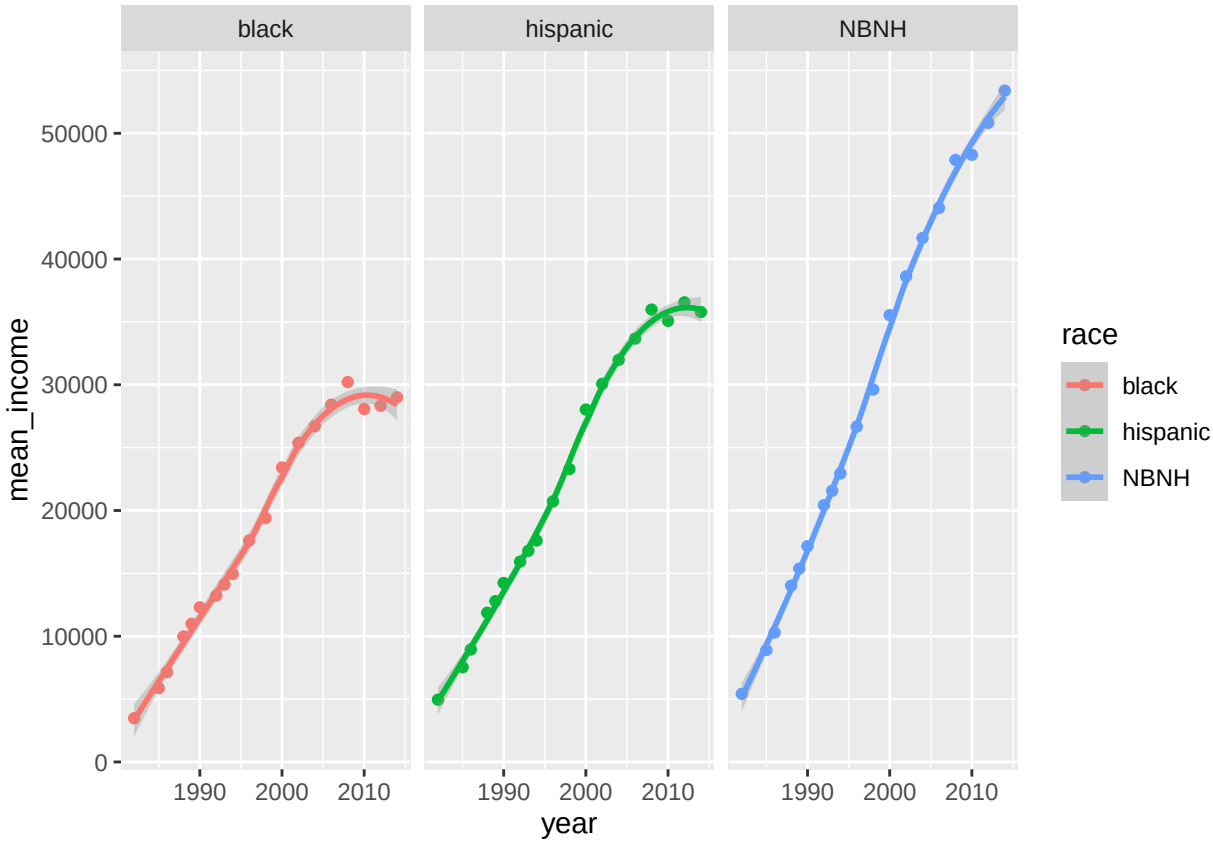
Over the years, the mean income for all races has steadily increased and can be seen to do so in the graphs below.

```
mean_inc_race_timeseries <- race_income %>%
  group_by(race, year) %>%
  summarise(mean_income = mean(income, na.rm=T))
```

```
## `summarise()` has grouped output by 'race'. You can override using the
## `.groups` argument.
```

```
ggplot(
  mean_inc_race_timeseries,
  aes(x=year,y=mean_income,color=race))+geom_point()+geom_smooth()+facet_wrap(~race)
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

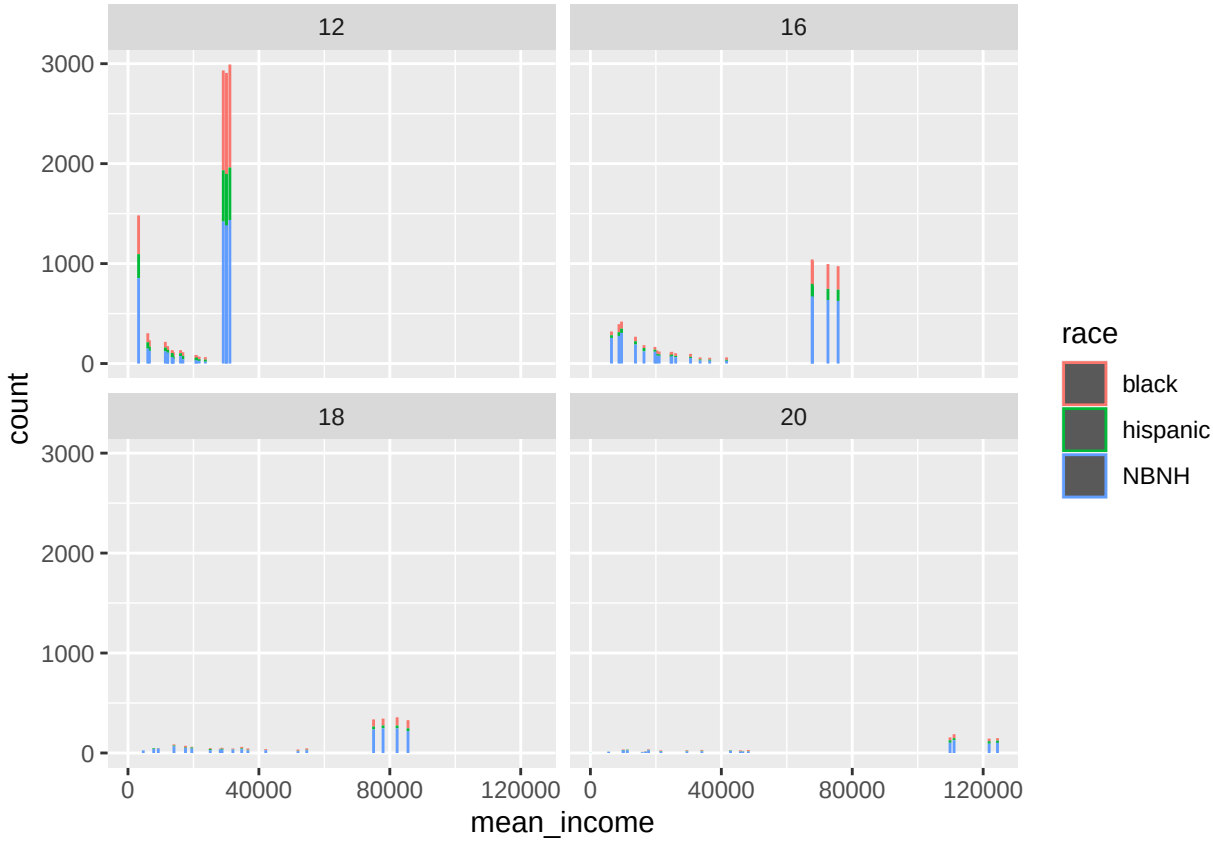



```
educ_race2 <- educ_race %>% inner_join(mean_inc_educ_timeseries)
```

```
## Joining with `by = join_by(education, year)`
```

When studying the correlation of the mean income and education and race, we can see again that as education levels increase that mean income also increases. Similarly, the NBNH make up the largest percentage of the larger mean incomes seen below in each category of education from 12 years on.

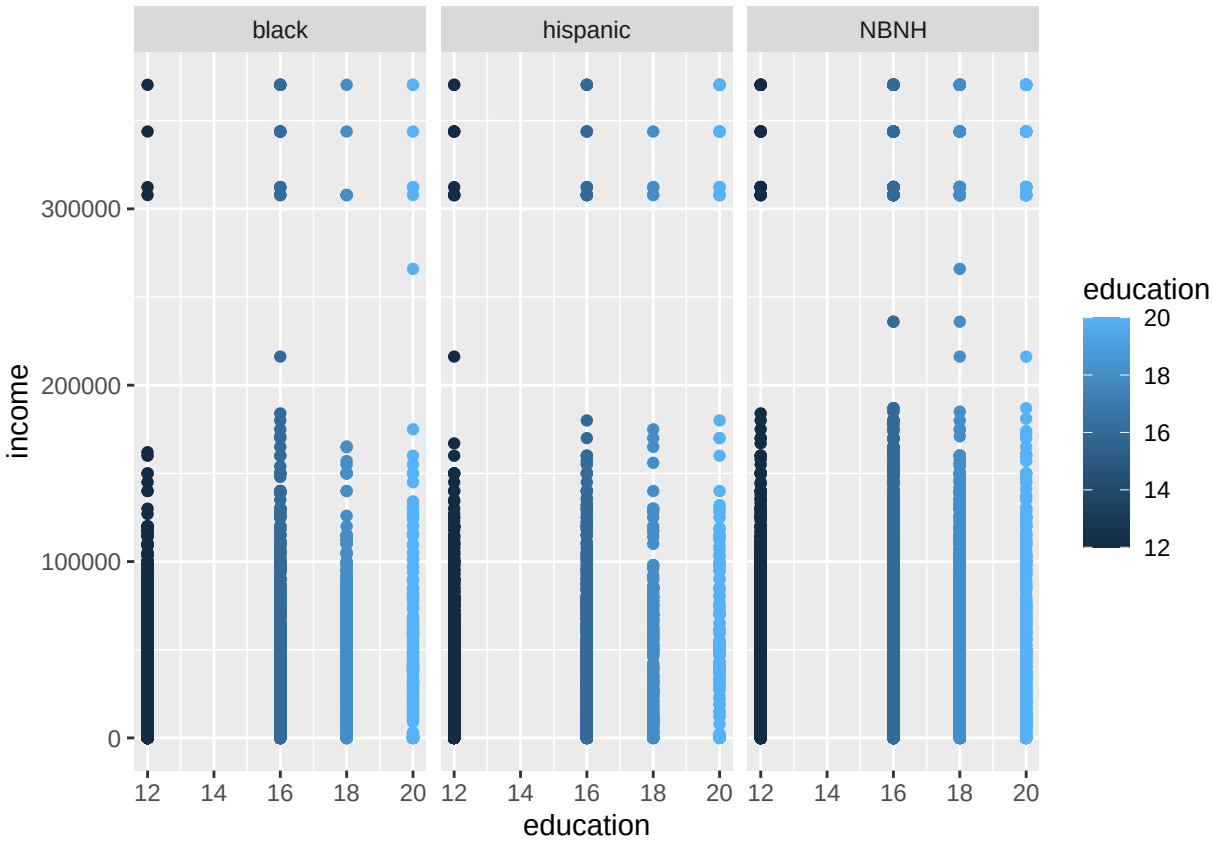
```
ggplot(educ_race2, aes(x=mean_income, color=race))+geom_bar()+facet_wrap(~education)
```



For another viewpoint of the data, we can separate by race to study the density of mean income as related to education and compare the three graphs to make conclusions.

```
ggplot(educ_race2, aes(x=education,y=income, color=education))+geom_point()+facet_wrap(~race)
```

```
## Warning: Removed 948 rows containing missing values (`geom_point()`).
```



Similarly, when looking at the previous data, we can see an increase in mean income with an increase in education as well as an increase in the NBNH community as opposed to the Hispanic and black survey participants.

Mean Income and the effects of Education and Race in 2014

Lastly, let's check our most recent data one more time in order to make some final conclusions.

```
educ_race2014 <- educ_race %>% filter(year==2014)
educ_race2014.5 <- educ_race2014 %>% group_by(education,race) %>% summarise(mean_income = mean(income,na
```

```
## `summarise()` has grouped output by 'education'. You can override using the
## `.groups` argument.
```

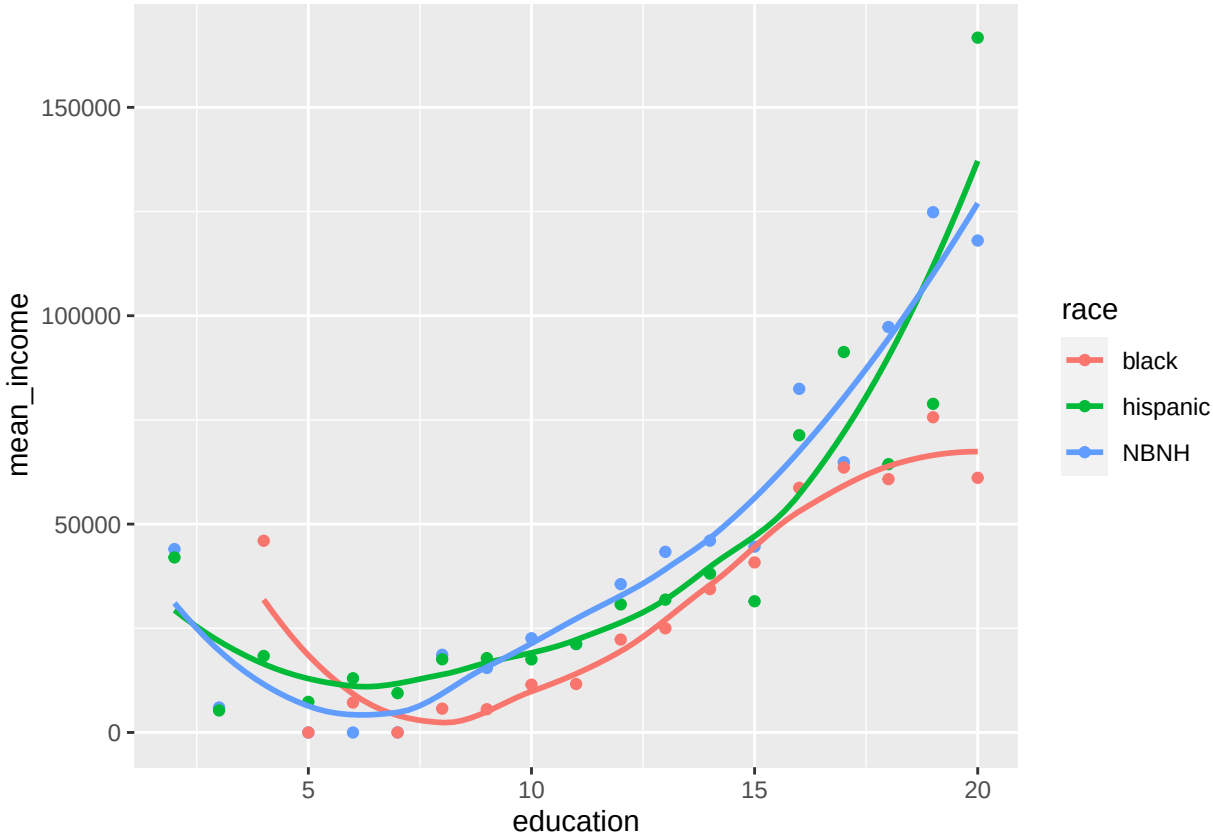
```
educ_race2014.5
```

```
## # A tibble: 55 x 3
## # Groups:   education [19]
##    education race     mean_income
##        <int> <chr>          <dbl>
##  1         2 NBNH          44000
##  2         2 hispanic      42000
##  3         3 NBNH           6000
##  4         3 black           NaN
##  5         3 hispanic       5286.
##  6         4 black         46000
##  7         4 hispanic      18351.
##  8         5 NBNH              0
##  9         5 black             0
```

```
## 10         5 hispanic       7341.
## # i 45 more rows
```

```
ggplot(data=educ_race2014.5, mapping=aes(x=education,y=mean_income, color=race))+geom_point()+geom_smoo
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## Warning: Removed 1 rows containing non-finite values (`stat_smooth()`).
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



In the most recent years of the survey, we can see that there is clearly a increase in mean income as education levels increase. In 2014, the trend shows that Hispanic individuals of higher levels of education are trending to having a higher mean income than those who are NBNH and black. Those who are NBNH who make it past 8th grade do sharply increase their mean income as their level of education increases. In the black survery participants, it appears that for those who continue past an 8th grade education, they do increase their mean income but at a much less sharp rate as compared to their NBNH and hispanic survey counterparts. They also seem to plateau as the other two races increase as education continues.

Conclusion

In conclusion, income can be seen to be effected by many factors according to the nlsy79 survey. The first variables studied were the height of the participants as well as their gender. According to the data gathered, there was a slight increase in mean income as height increased in both male and female survey participants. Male mean income was also significantly higher than their female counterparts.

The next variables that were explored were education and race and the effects that the variables had on each other. As imagined, as education levels increased the mean income of the survey participants increased. This makes sense as with greater knowledge comes more responsibilities within the workplace and greater

pay. In regards to race, it appeared that the NBNH consistently had the highest records of mean income as compared to the black and Hispanic survey participants. However in recent years, it appears that the mean income for the Hispanic survey participants is sharply growing. This could signal good things for the hispanic community in regard to securing high paying jobs going forward.