Income analysis

Introduction

Many factors effect a person's income over a period of years. In this analysis we will explore how income is affected by gender, education level, and BMI.The data set used contains data about income, years of education, and physical characteristics for respondents to NLSY '79.

Load libraries and data to be used

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## Warning: package 'tidyr' was built under R version 4.1.2

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(scales)
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor
```

```
load("/Users/disha/Documents/R\ Class/FinalProject-B/FinalProjectPartB/income_data_nlsy79.RData")
load("/Users/disha/Documents/R\ Class/FinalProject-B/FinalProjectPartB/education_data_nlsy79.RData")
load("/Users/disha/Documents/R\ Class/FinalProject-B/FinalProjectPartB/physical_data_nlsy79.RData")
```

Univariate Exploration

Exploring Income Data:

Income data is only available for years 1982-2014

```
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
```

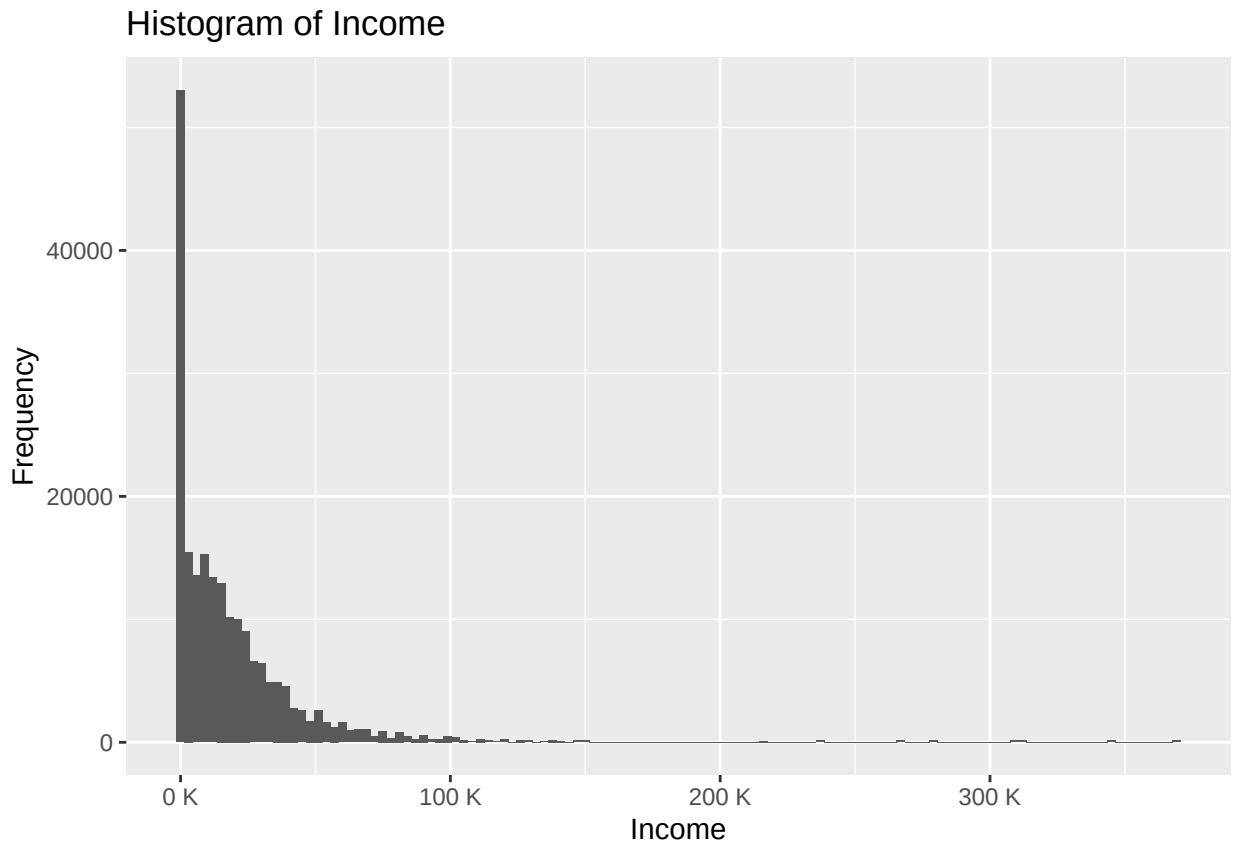
```
length(income_data_nlsy79$income)
```

```
## [1] 291778
```

Histogram of Income:

```
ggplot(data = income_data_nlsy79, aes(x= income)) + geom_histogram(binwidth = 3000) + scale_x_continuous
```

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



Observations: There are some extreme values of income

Summary statistics of income variable:

```
summary(income_data_nlsy79$income)
```

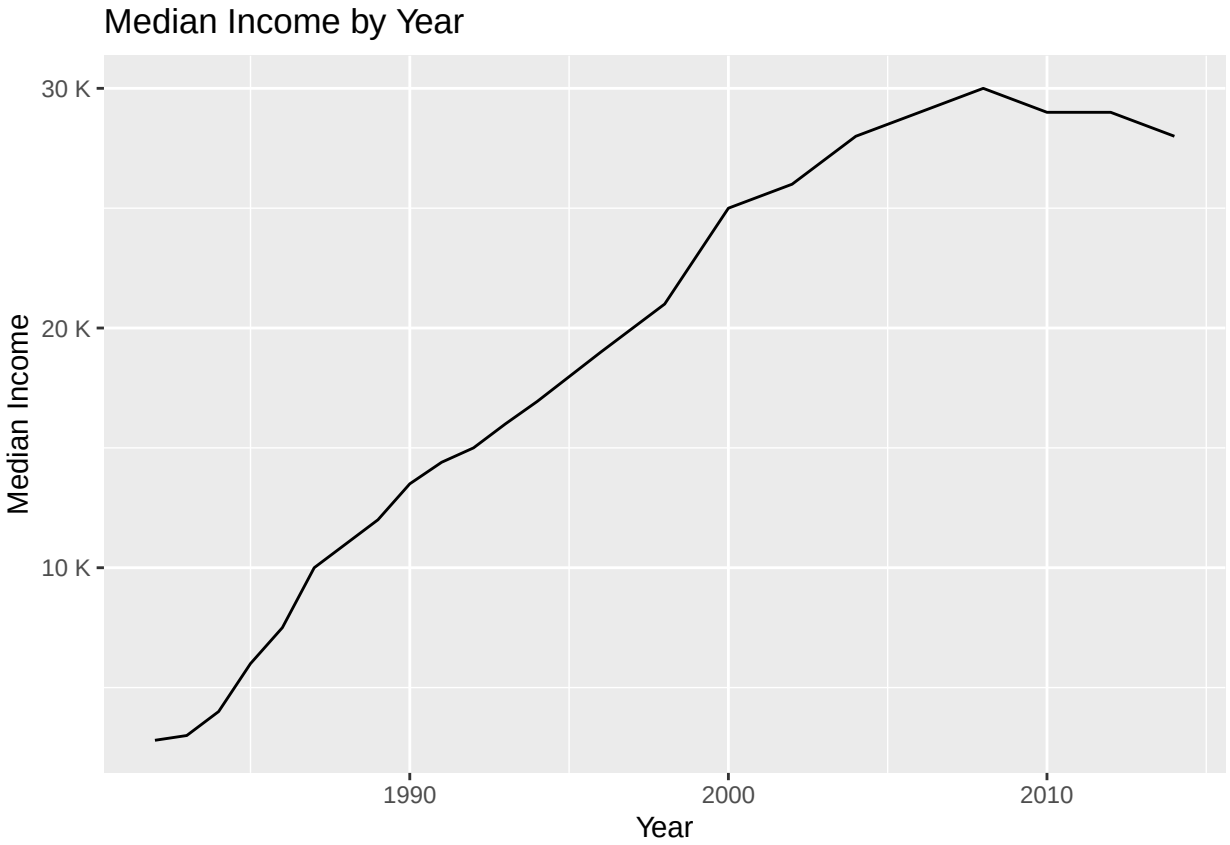
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0    1344   12000   19867   26000  370314   85628
```

```
quantile(income_data_nlsy79$income, probs = c(0.01, 0.05,0.10,0.50,0.75, 0.95, 0.99), na.rm = T)
```

```
##     1%     5%    10%    50%    75%    95%    99%
##      0      0      0  12000  26000  65000 130000
```

Looking at the year to year median income

```
median_income_by_year <- income_data_nlsy79 %>% group_by(year) %>% summarise(med_income = median(income
ggplot(median_income_by_year, aes(year,med_income)) + geom_line()+ scale_y_continuous(name="Median Inco
```



Found that there were a number of individuals with a max(income) of $370,314 which seems to be some sort of issue in the data, because it seems very unlikely that so many individuals would report such a high, yet very specific income.

```
income_by_ID <- income_data_nlsy79 %>% group_by(CASEID) %>% summarize (mean_income = as.integer(mean(in
```

Exploring Education Data

```
length(education_data_nlsy79$CASEID)
```

```
## [1] 329836
```

```
unique(education_data_nlsy79$year)
```

```
##  [1] 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993
## [16] 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014
```

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

```
##  [1]  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 95
```

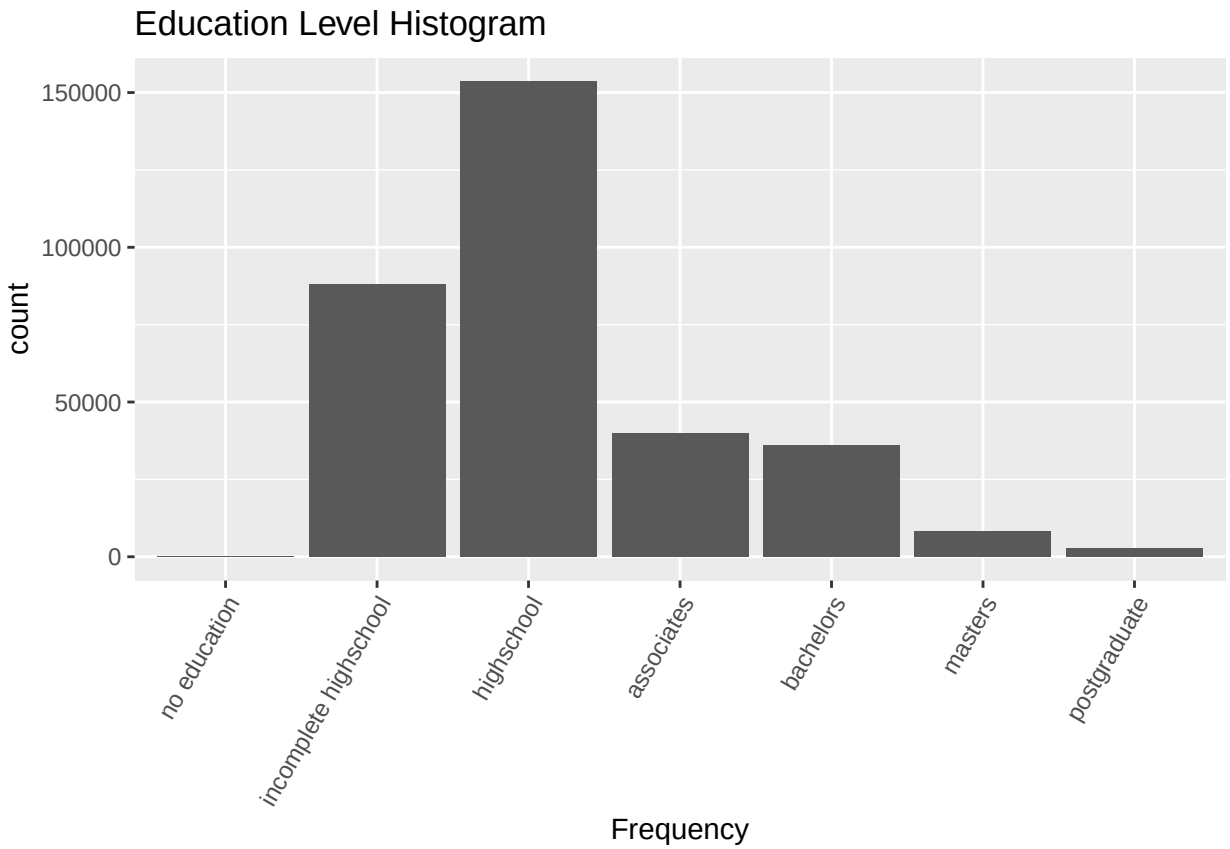
Education data is only available for 1979-2014. However, income is available for 1982-2014. Years 1979-1981 will be removed because of missing data. '

Unique values for education has number 95. Replacing the number 95 in education with NA. Created a for loop that will backfill in educations level for any as NAs such that if 1998 was NA and 1999 was 16, 1998 would get filled with 16 making it easier to look at education for any given year. Assigning a categorical variable for education for easier analysis. (For loop takes some time to run) Also removed NA's from education data.

```
education_95 <- education_data_nlsy79 %>% filter(education == 95)
education_data_nlsy79_1 <- education_data_nlsy79 %>% mutate(education = ifelse(education == 95, NA, edu
education_data_nlsy79_1 <-  arrange(education_data_nlsy79_1,CASEID, year)
for(i in 2:nrow(education_data_nlsy79_1)){
if(is.na(education_data_nlsy79_1$education[i]) & education_data_nlsy79_1$CASEID[i] == education_data_nls
  education_data_nlsy79_1$education[i] <- education_data_nlsy79_1$education[i-1]
  }
}
education_data_nlsy79_1 <- education_data_nlsy79_1%>% filter(!(is.na(education))) %>% mutate(education_
  education == 0 ~ "no education",
  education < 12 ~ "incomplete highschool",
  education > 11 & education < 14 ~ "highschool",
  education > 13 & education < 16 ~ "associates",
  education > 15 & education < 18 ~ "bachelors",
  education > 17 & education < 20 ~ "masters",
  education > 19 ~ "postgraduate"
))
```

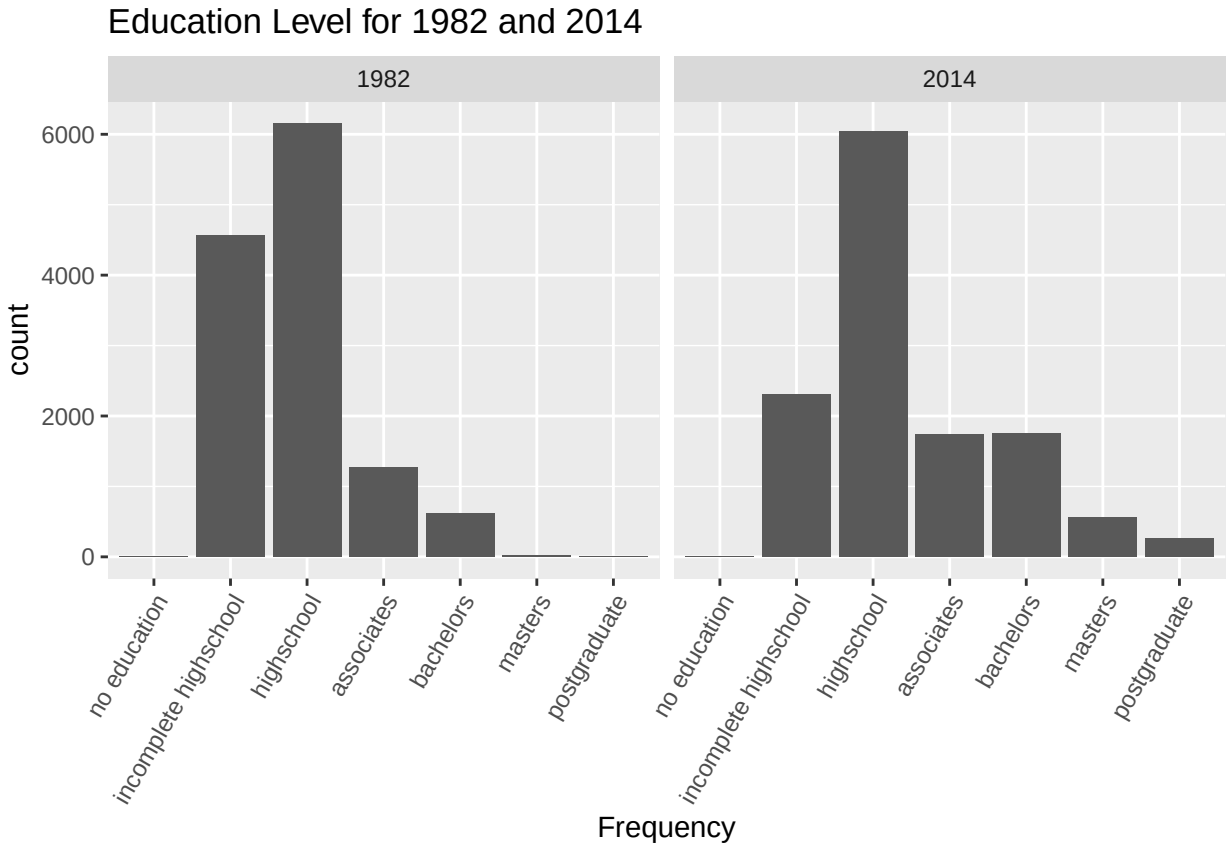
Distribution of education levels: Note:This histogram counts rows and is not counting per person.Additionally, the proportion of people with no education is very small and therefore does not show on the graph.

```
education_data_nlsy79_1$education_level <- factor(education_data_nlsy79_1$education_level, levels=c("no
ggplot(data = education_data_nlsy79_1, aes(x=education_level))+ geom_bar()+ theme(axis.text.x = element_
```



Compare education level for two extreme years- 1982 and 2014

```
education_data_nlsy79_1_fixedyear <- education_data_nlsy79_1%>% filter(year == 1982 | year == 2014)
education_data_nlsy79_1_fixedyear$education_level <- factor(education_data_nlsy79_1_fixedyear$education_
ggplot(education_data_nlsy79_1_fixedyear, aes(x= education_level)) + geom_bar() + facet_wrap(~year) + t
```



Exploring Gender Data

The data has no null values in the sex column and is pretty evenly distributed between male and female. 49.5% are female and 50.5% male. Data is available for 1981-2014

```
glimpse(physical_data_nlsy79)
```

```
## Rows: 253,720
## Columns: 9
## $ CASEID <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, ~
## $ weight <int> NA, 120, NA, 110, 130, 200, 131, 179, 145, 115, 155, 118, 180, ~
## $ year   <int> 1981, 1981, 1981, 1981, 1981, 1981, 1981, 1981, 1981, 1981, 198~
## $ eyes   <chr> NA, "hazel", "blue", "blue", NA, "brown", "brown", "hazel", "ha~
## $ hair   <chr> NA, "light brown", "blond", "light brown", NA, "brown", "brown"~
## $ race   <chr> "NBNH", "NBNH", "NBNH", "NBNH", "NBNH", "NBNH", "NBNH", "NBNH",~
## $ sex    <chr> "female", "female", "female", "female", "male", "male", "male",~
## $ height <int> 65, 62, NA, 67, 63, 64, 65, 65, 66, 66, 71, 66, 71, 67, 73, 63,~
## $ BMI    <dbl> NA, 21.94843, NA, 17.22855, 23.02862, 34.33015, 21.79968, 29.78~
```

```
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
```

```
table(physical_data_nlsy79$sex)
```

```
##
## female   male
## 125660 128060
```

```
proportion<-table(physical_data_nlsy79$sex)/ length(physical_data_nlsy79$sex)*100
format(round(proportion, 1), nsmall = 1)
```

```
##
## female   male
## "49.5" "50.5"
```

Exploring BMI data

Explore the BMI data

```
length(physical_data_nlsy79$CASEID)
```

```
## [1] 253720
```

```
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
```

Bivariate Exploration

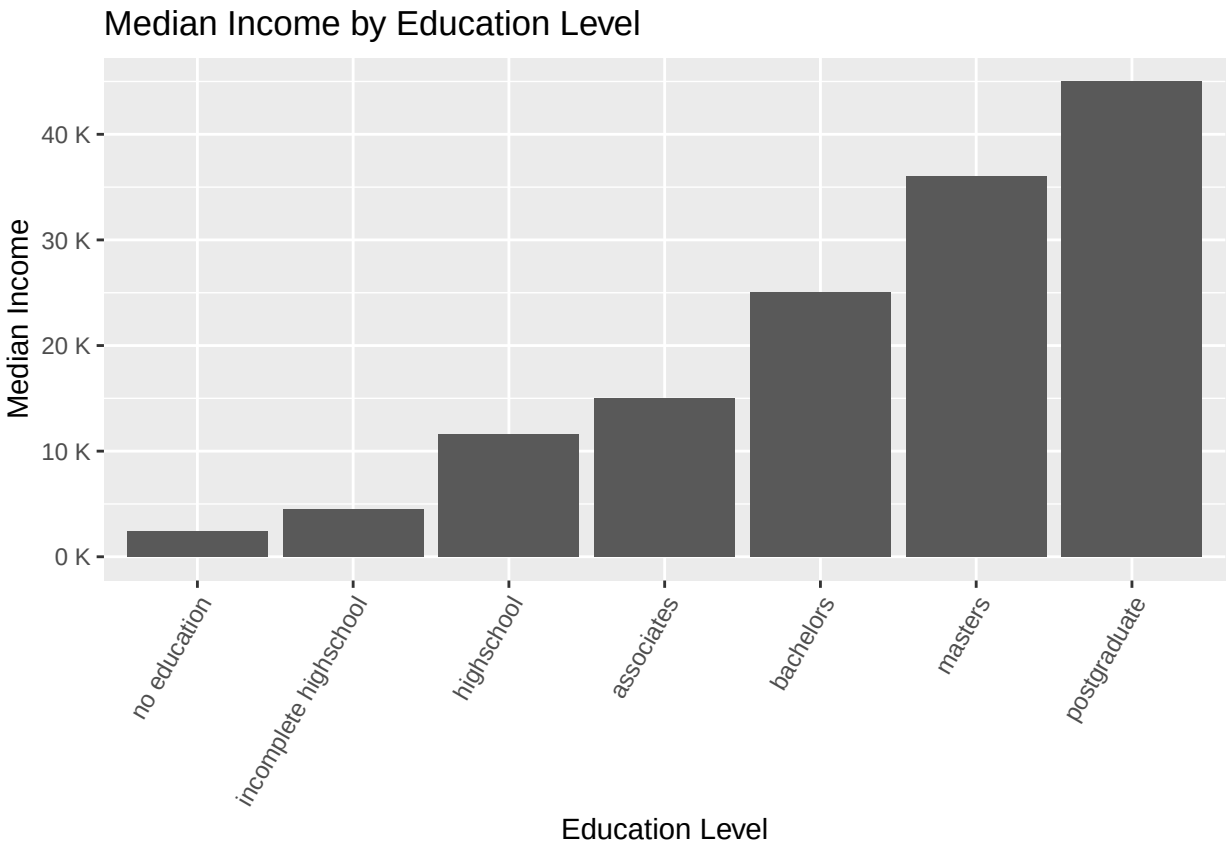
Education and Income

Merge the income data and education data by caseID and year. For education we have years 1979- 1981 which doesnt exist in income.

```
income_education_data <-  merge(education_data_nlsy79_1, income_data_nlsy79, by = c("CASEID", "year"))
income_education_data <- income_education_data %>% filter(!(is.na(income)))
```

Verify if higher education means higher income.

```
income_education_data1 <- income_education_data %>% group_by(education_level) %>% summarize(median_inco
income_education_data1$education_level <- factor(income_education_data1$education_level, levels=c("no ed
ggplot(income_education_data1, aes(education_level, median_income)) + geom_col() + theme(axis.text.x = e
```

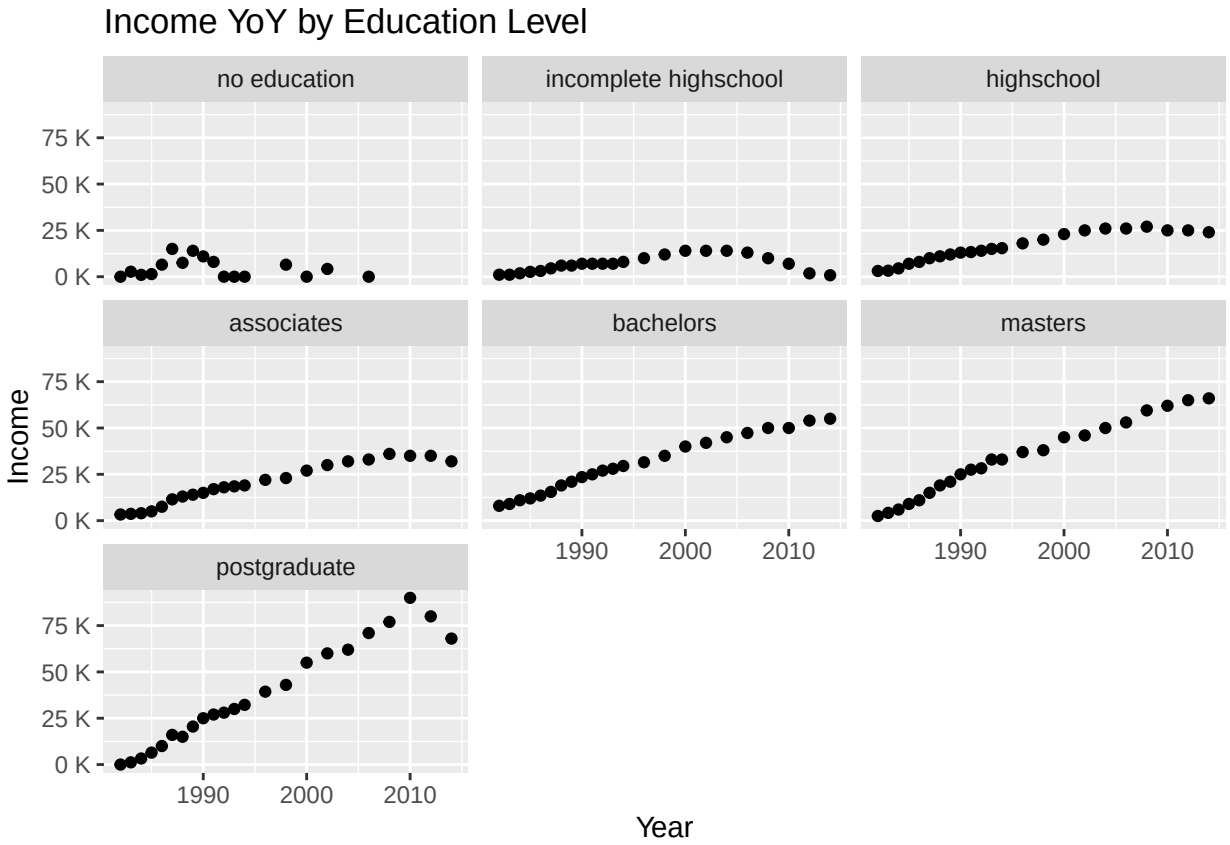


For same education level, compare income year to year

```
income_education_data2 <- income_education_data %>% group_by(education_level,year) %>% summarize(median_
```

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

```
income_education_data2$education_level <- factor(income_education_data2$education_level, levels=c("no ed
ggplot(income_education_data2, aes(year,median_income)) + geom_point() + facet_wrap(~education_level) +
```



Gender and Income

Merging Income and Gender

```
nrow(physical_data_nlsy79)
```

```
## [1] 253720
```

```
nrow(income_data_nlsy79)
```

```
## [1] 291778
```

```
sex_income <- merge(physical_data_nlsy79,  income_data_nlsy79, by= c("CASEID", "year"))%>% select(CASEI
summary(sex_income)
```

```
##      CASEID           year          sex                income
##  Min.   :    1   Min.   :1982   Length:241034      Min.   :     0
##  1st Qu.: 3172   1st Qu.:1989   Class :character   1st Qu.:  2000
##  Median : 6344   Median :1996   Mode  :character   Median : 14200
##  Mean   : 6344   Mean   :1997                      Mean   : 22642
##  3rd Qu.: 9515   3rd Qu.:2006                      3rd Qu.: 30000
##  Max.   :12686   Max.   :2014                      Max.   :370314
##                                                    NA's   :77963
```

We were not able to find an income record for each of the sex records (77963). The NA's will be filled with the previous year income when available

```
nrow(sex_income)
```

```
## [1] 241034
```

```
for(i in 2:nrow(sex_income)){
if(is.na(sex_income$income[i]) & sex_income$CASEID[i] == sex_income$CASEID[i-1]){
  sex_income$income[i] <- sex_income$income[i-1]
  }
}
summary(sex_income)
```

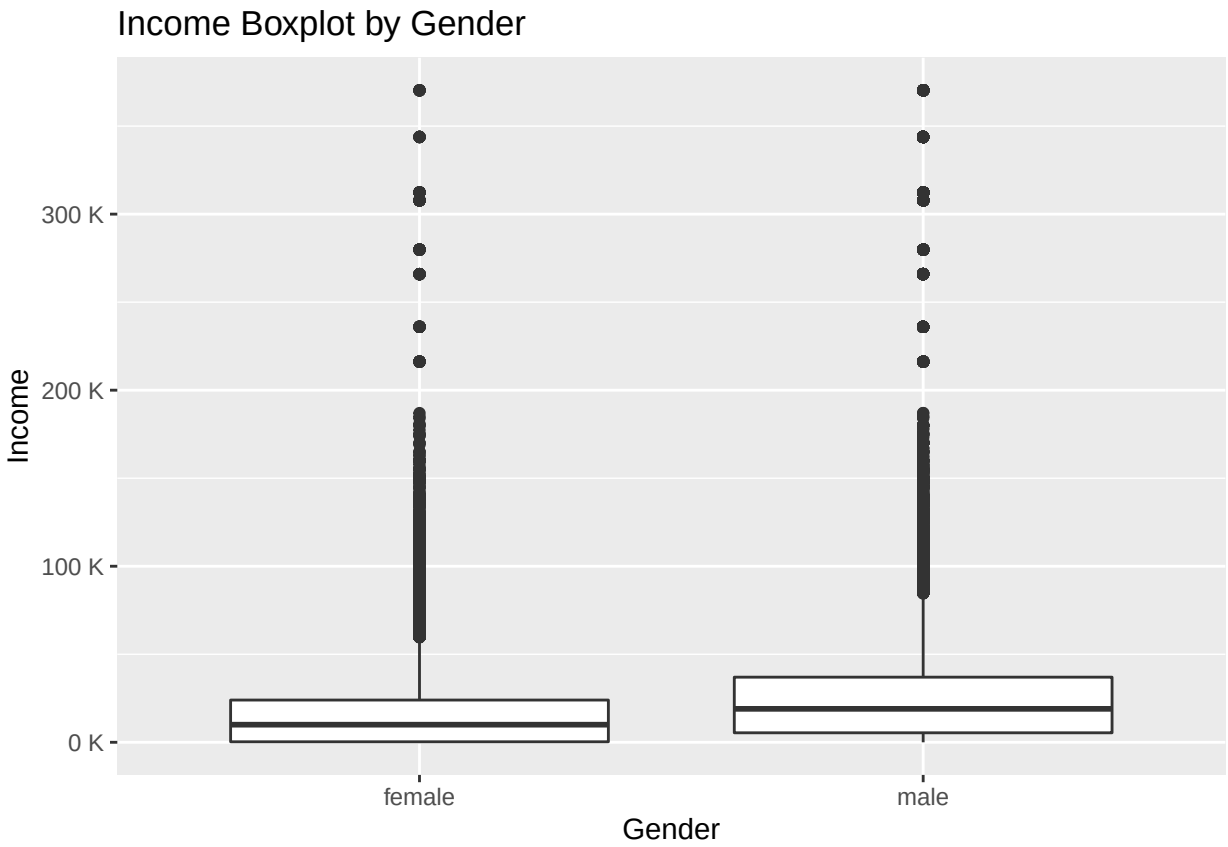
```
##      CASEID           year          sex                income
##  Min.   :    1   Min.   :1982   Length:241034      Min.   :     0
##  1st Qu.: 3172   1st Qu.:1989   Class :character   1st Qu.:   900
##  Median : 6344   Median :1996   Mode  :character   Median : 12000
##  Mean   : 6344   Mean   :1997                      Mean   : 20066
##  3rd Qu.: 9515   3rd Qu.:2006                      3rd Qu.: 27000
##  Max.   :12686   Max.   :2014                      Max.   :370314
##                                                    NA's   :5820
```

We were able to reduce the NA values to 5,820. These remaining records will be discarded since we don't have any income records for any year for these individuals and we cannot assume they received 0 income for that year.

Exploring how income differ according to gender. Based on the boxplot below we can see that the median salary for women is lower than the median salary for males.

```
sex_income <- merge(physical_data_nlsy79,  income_data_nlsy79, by= c("CASEID", "year"))%>% select(CASEI
ggplot(sex_income, aes(x=sex, y=income))+ geom_boxplot() + scale_y_continuous(name="Income",labels = la
```

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



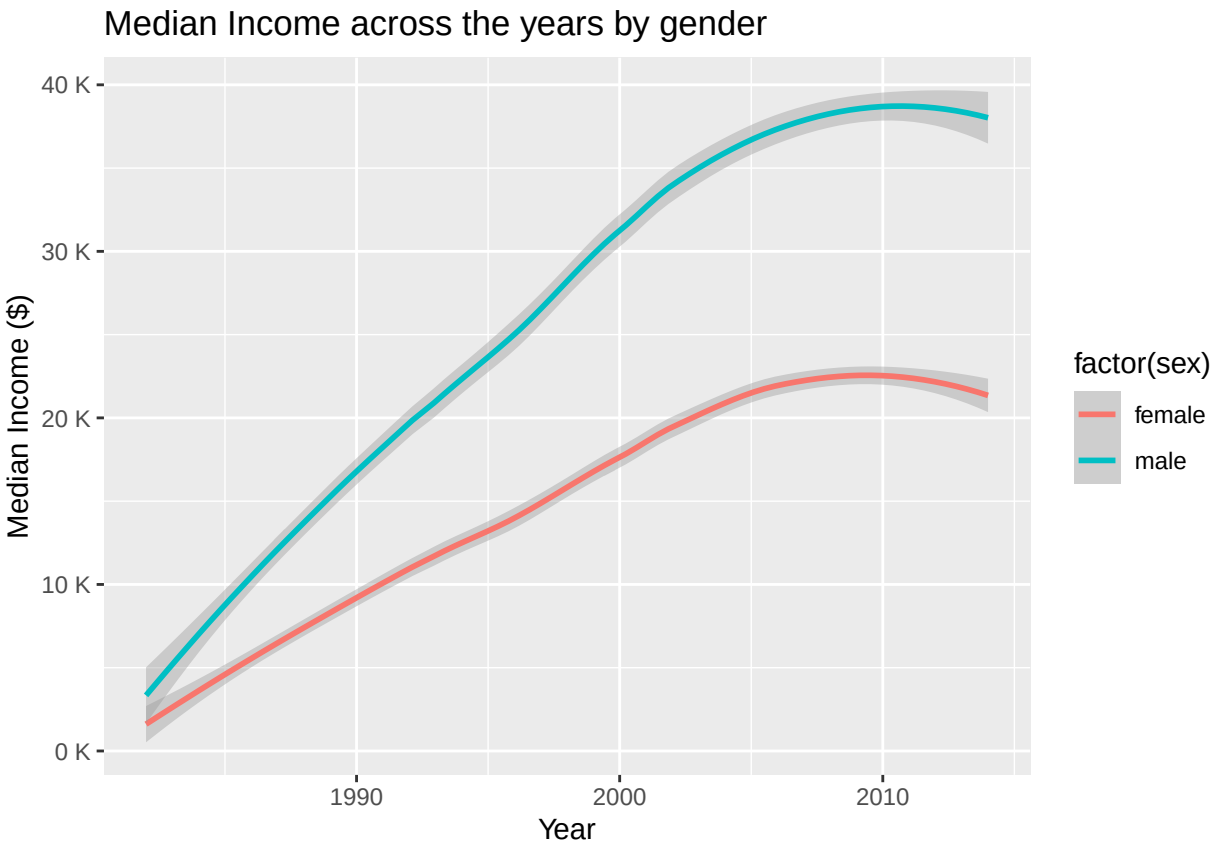
How has income change throughout the year for males and females? Based on the data below we can income has increased throughout the years, however it has increased at a faster rate for males.

```
years_income_sex<-sex_income%>%select(year, sex, income)%>%group_by(year,sex)%>% summarize(median_incom
```

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

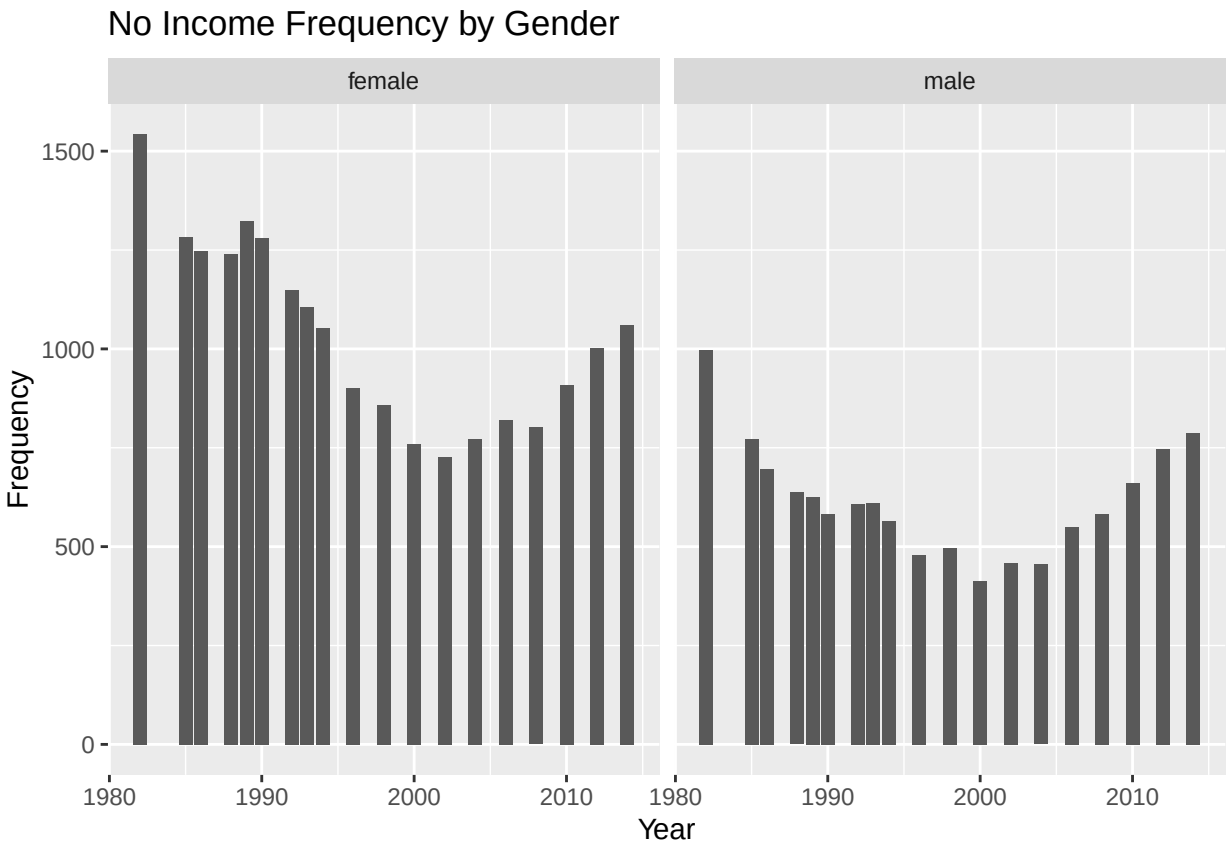
```
ggplot(years_income_sex, aes(x=year, y=median_income))+geom_smooth(aes(colour = factor(sex)))+ scale_y_
```

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



One reason why median income is lower for females than males could be because more females than male report not having any income at all

```
years_noincome_sex<-sex_income%>%select(year, sex, income)%>%group_by(year,sex)%>%filter(income==0)
ggplot(years_noincome_sex, aes(x=year))+geom_bar()+facet_wrap(~sex)+ggtitle("No Income Frequency by Gen
```



BMI and Income

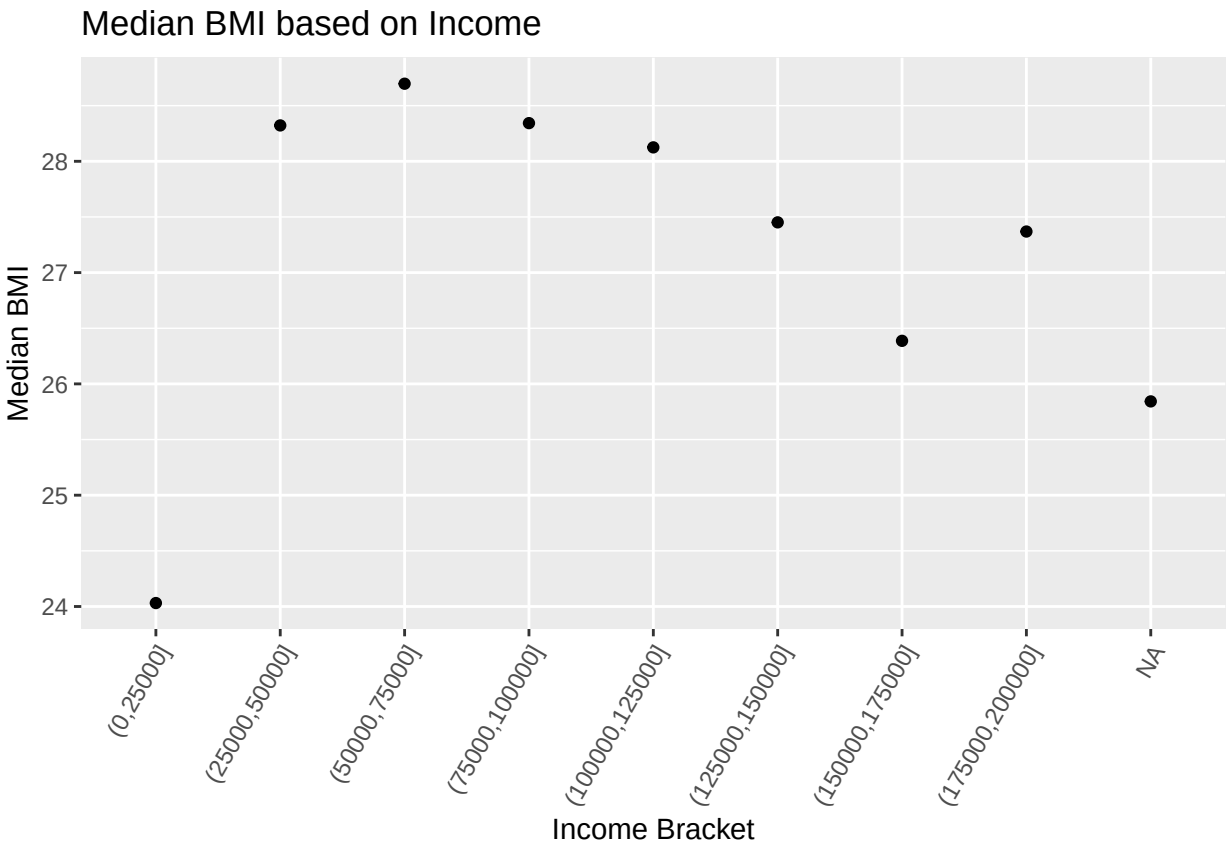
Limit the BMI data to only the last year

```
BMI_phys <- physical_data_nlsy79 %>% filter(!is.na(BMI)) %>% select(CASEID,year,BMI,sex)
BMI_phys <- BMI_phys %>% group_by(CASEID) %>% filter(year == max(year))
```

Merge the BMI and Income data

```
BMI_Income <- merge(BMI_phys,income_data_nlsy79, by = c("CASEID","year"), all.x = TRUE) %>% filter(incor
```

```
point_plot <- BMI_Income %>% group_by(inc_bracket = cut(income, breaks = seq(0,250000, by = 25000), dig
ggplot(point_plot, aes(inc_bracket,med_BMI)) + geom_point() + theme(axis.text.x = element_text(angle = 
```

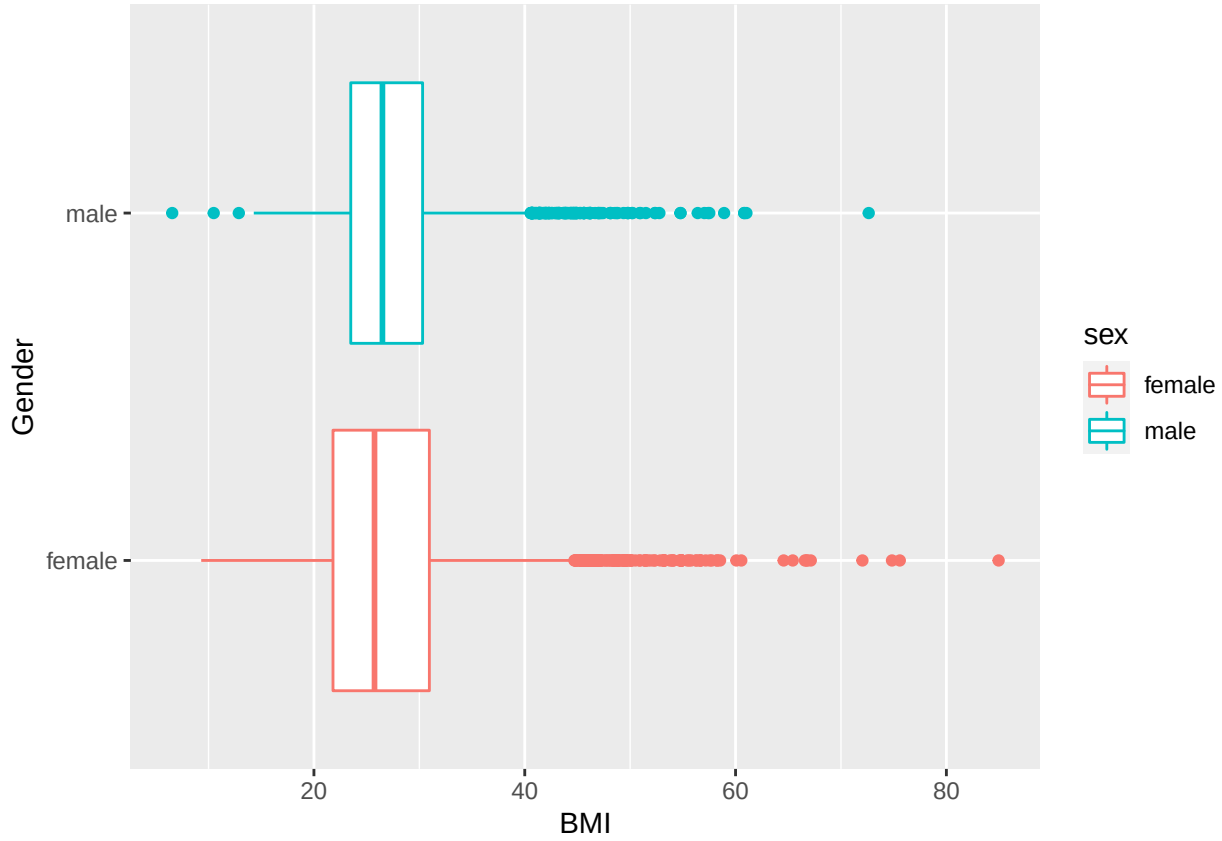


BMI and Gender

Median BMI seems lower for females than males. Also females have a larger spread of BMI than males

```
BMI_Sex_BoxPlot <- BMI_Income %>% filter(income <= 250000)

ggplot(BMI_Sex_BoxPlot, aes(x=BMI, y=sex)) + geom_boxplot(aes(color = factor(sex)))+ylab("Gender")+scal
```




```
boxplot1 <- BMI_Income %>% group_by(inc_bracket = cut(income, breaks = seq(0,250000, by = 25000), dig.la
boxplot1
```

```
## # A tibble: 8,588 x 6
## # Groups:   inc_bracket [8]
##    CASEID  year   BMI sex    income inc_bracket
##     <int> <int> <dbl> <chr>   <int> <fct>
##  1      2  2014  29.5 female  21000 (0,25000]
##  2      3  2014  23.0 female  40000 (25000,50000]
##  3      5  1982  23.9 male     2300 (0,25000]
##  4      6  2014  32.4 male   112000 (100000,125000]
##  5      8  2014  34.0 female  47000 (25000,50000]
##  6      9  2014  30.7 male    80000 (75000,100000]
##  7     10  1985  21.8 female   6000 (0,25000]
##  8     11  1985  21.6 male    20000 (0,25000]
##  9     12  1985  17.6 female  25000 (0,25000]
## 10     13  2006  31.9 male     8000 (0,25000]
## # ... with 8,578 more rows
```

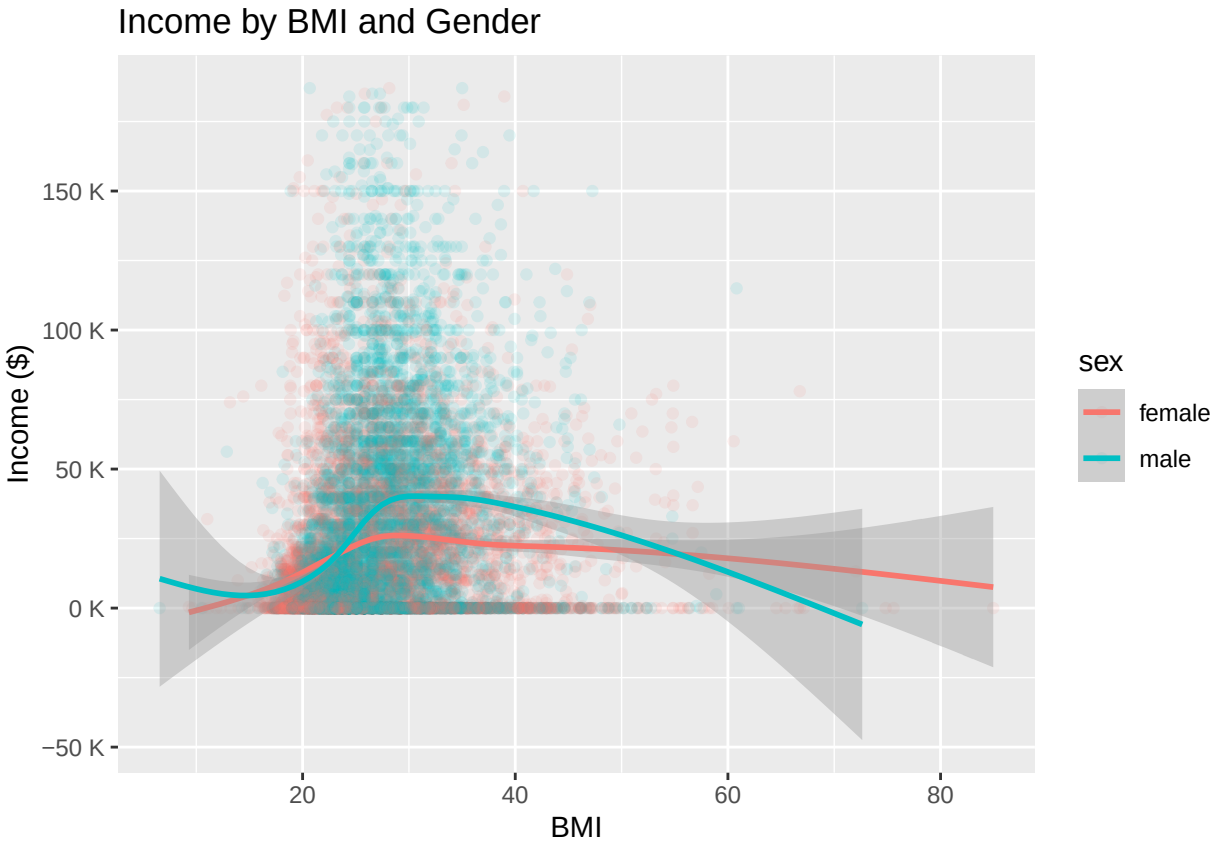
Multivariate Exploration

Income, Gender, BMI

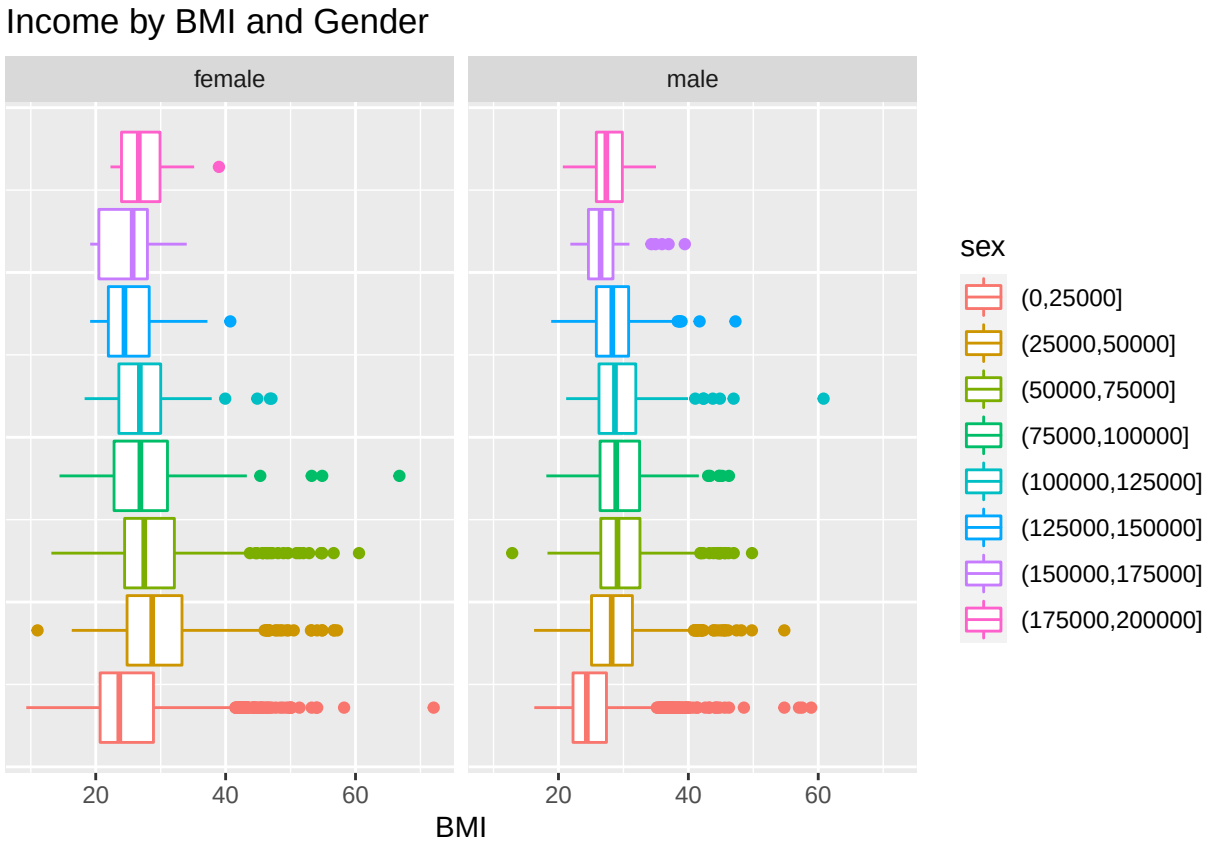
Income, BMI, and gender were graphed in different ways to explore their multivariate relationship in order to see which one would give us the most information to continue our investigation.

```
ggplot(BMI_Income,aes(BMI,income)) + geom_point(alpha = .1, aes(colour = factor(sex))) + geom_smooth(aes
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
ggplot(boxplot1, aes(BMI)) + geom_boxplot(aes(color = factor(inc_bracket))) + facet_grid(~sex) + theme(a
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank()) + ggtitle("Income by BMI and Gender")+scale_colour_discrete("sex")
```




```
point_plot2 <- BMI_Income %>% group_by(inc_bracket = cut(income, breaks = seq(0,250000, by = 25000), di
```

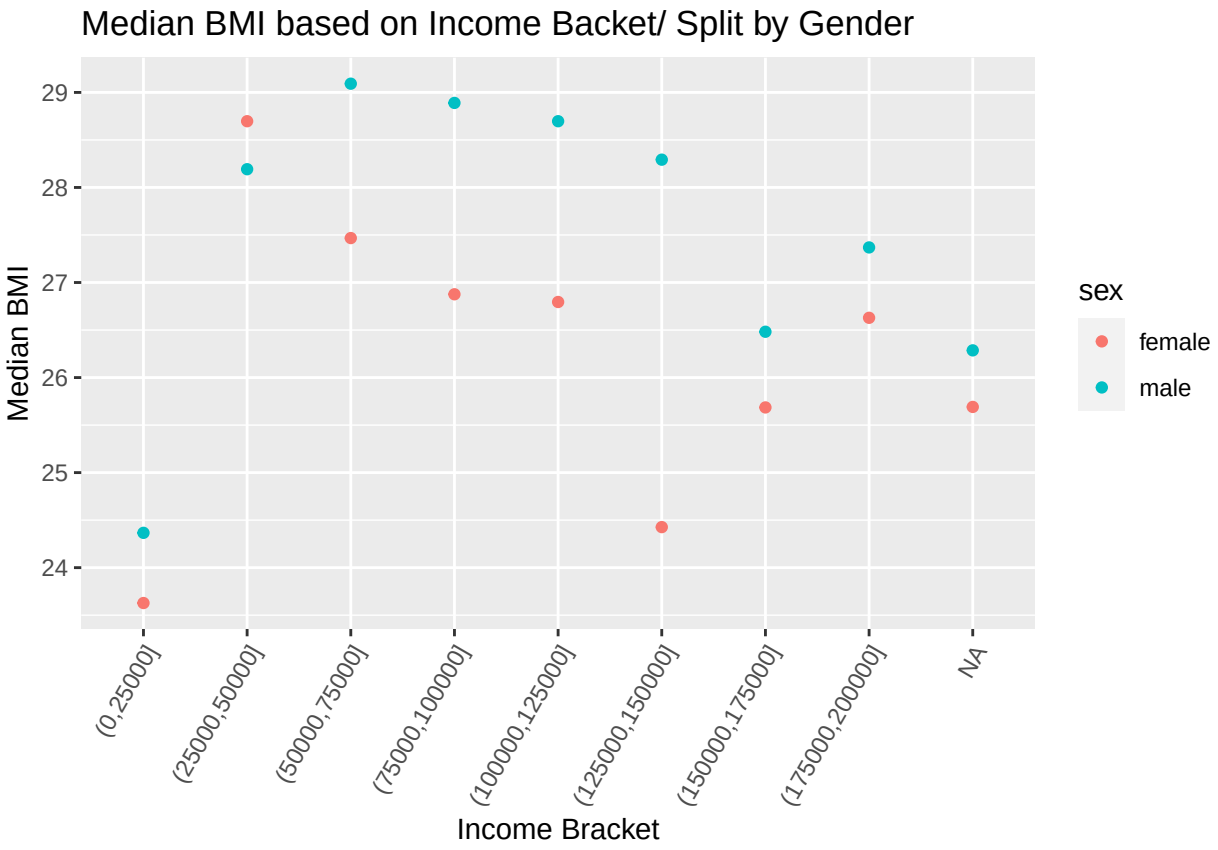
```
## `summarise()` has grouped output by 'inc_bracket'. You can override using the `.groups` argument.
```

```
point_plot2
```

```
## # A tibble: 18 x 3
## # Groups:   inc_bracket [9]
##    inc_bracket     sex    med_BMI
##    <fct>           <chr>    <dbl>
##  1 (0,25000]       female    23.6
##  2 (0,25000]       male      24.4
##  3 (25000,50000]   female    28.7
##  4 (25000,50000]   male      28.2
##  5 (50000,75000]   female    27.5
##  6 (50000,75000]   male      29.1
##  7 (75000,100000]  female    26.9
##  8 (75000,100000]  male      28.9
##  9 (100000,125000] female    26.8
## 10 (100000,125000] male      28.7
## 11 (125000,150000] female    24.4
## 12 (125000,150000] male      28.3
## 13 (150000,175000] female    25.7
## 14 (150000,175000] male      26.5
## 15 (175000,200000] female    26.6
```

```
## 16 (175000,200000] male      27.4
## 17 <NA>             female    25.7
## 18 <NA>             male      26.3
```

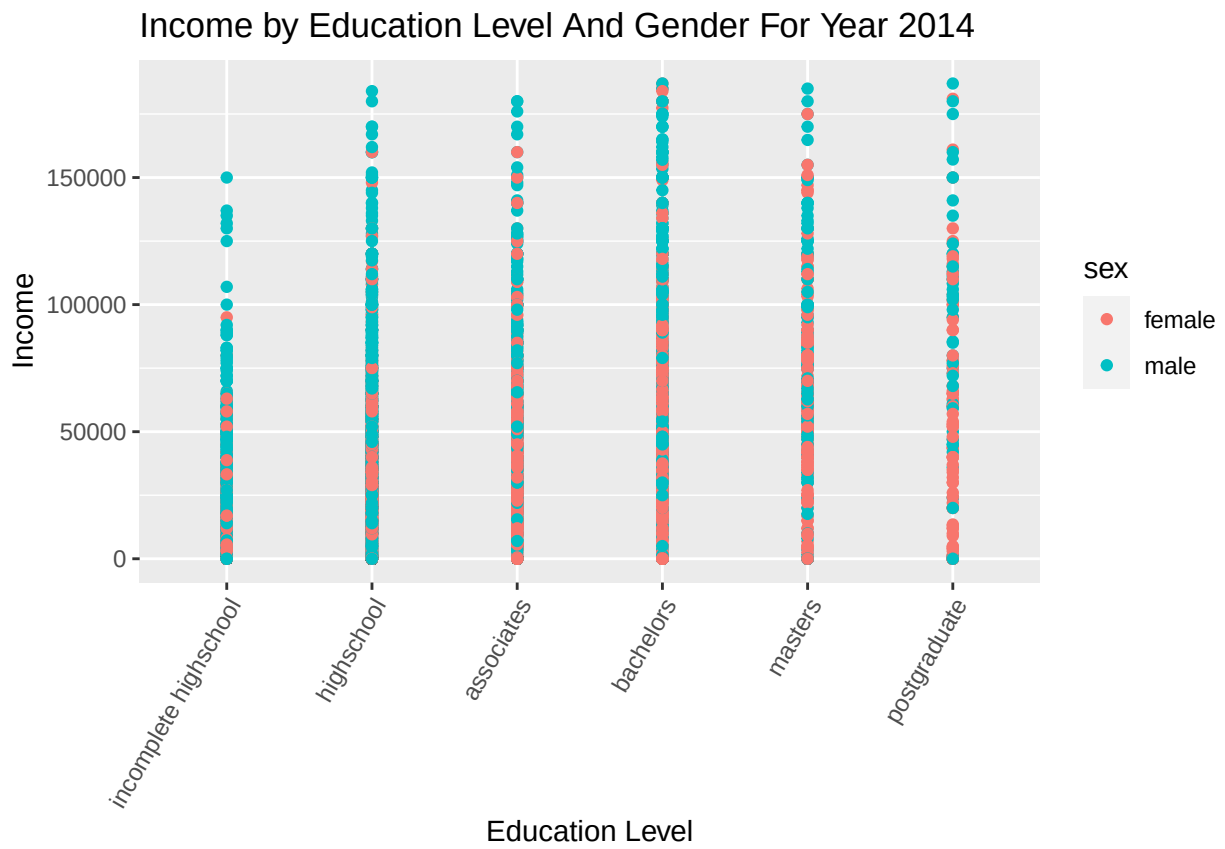
```
ggplot(point_plot2, aes(inc_bracket,med_BMI)) + geom_point(aes(colour = factor(sex))) + theme(axis.text
```



Gender, BMI, Education Level for the year 2014

Join the education, BMI data frames

```
all_factors_df <- merge(income_education_data,physical_data_nlsy79, by = c("CASEID","year"))
all_factors_df <- all_factors_df %>% select(CASEID, year, education, education_level, income, sex, BMI)%
all_factors_df$education_level <- factor(all_factors_df$education_level, levels=c("no education","incom
ggplot(data = all_factors_df, aes(x = education_level, y = income)) + geom_point(aes(color = factor(sex)
```



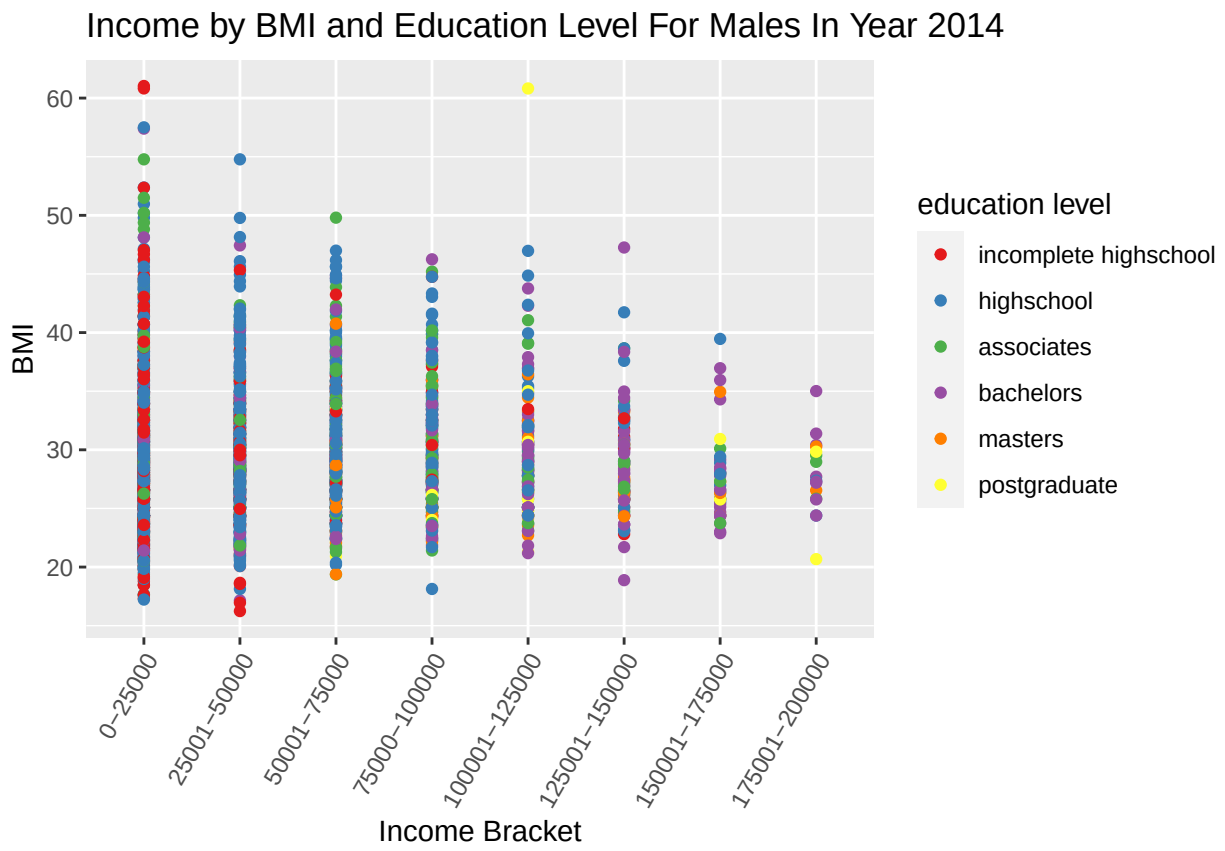
Income, BMI, Education Level for Males for the year 2014 The analysis below is restricted to a single year which is the last year of information available, 2014.

```
all_factors_df_male <- all_factors_df %>% filter(sex == "male") %>% mutate(income_bracket = case_when(
  income >= 0 & income <= 25000 ~ "0-25000",
  income >25000 & income <= 50000 ~ "25001-50000",
  income >50000 & income <= 75000 ~ "50001-75000",
  income >75000 & income <= 100000 ~ "75000-100000",
  income >100000 & income <= 125000 ~ "100001-125000",
  income > 125000 & income <= 150000 ~"125001-150000",
  income > 150000 & income <= 175000 ~"150001-175000",
  income > 175000 & income <= 200000 ~"175001-200000",
)) %>% filter(!(is.na(education_level)))

all_factors_df_male$income_bracket <- factor(all_factors_df_male$income_bracket, levels = c("0-25000","

all_factors_df_male$education_level <- factor(all_factors_df_male$education_level, levels=c("no educati


ggplot(data = all_factors_df_male, aes(y = BMI, x = income_bracket))+geom_point(aes(color = factor(educa
```



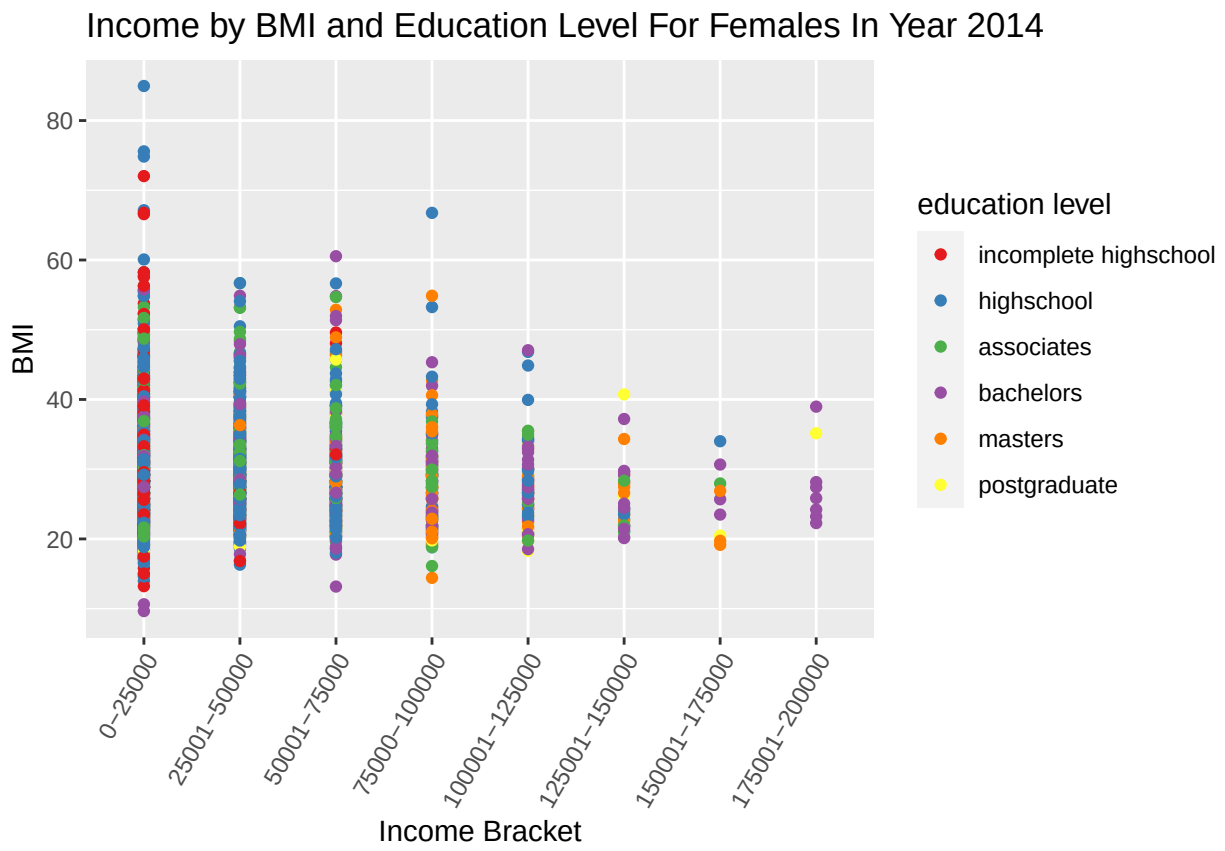
Income, BMI, Education Level for Females for the year 2014

```
all_factors_df_female <- all_factors_df %>% filter(sex == "female") %>% mutate(income_bracket = case_wh
  income >= 0 & income <= 25000 ~ "0-25000",
  income >25000 & income <= 50000 ~ "25001-50000",
  income >50000 & income <= 75000 ~ "50001-75000",
  income >75000 & income <= 100000 ~ "75000-100000",
  income >100000 & income <= 125000 ~ "100001-125000",
  income > 125000 & income <= 150000 ~"125001-150000",
  income > 150000 & income <= 175000 ~"150001-175000",
  income > 175000 & income <= 200000 ~"175001-200000"
)) %>% filter(!(is.na(education_level)))


all_factors_df_female$education_level <- factor(all_factors_df_female$education_level, levels=c("no edu


all_factors_df_female$income_bracket <- factor(all_factors_df_female$income_bracket, levels = c("0-2500

ggplot(data = all_factors_df_female, aes(y = BMI, x = income_bracket))+geom_point(aes(color = factor(edu
```



Observations and Hypothesis to test:

1. There are more people with bachelors, masters and post-graduate level education in year 2014 than in the year 1982.
2. Women make less income than men. This may be because more female report that they make no income.
3. The general trend is that BMI decreases as the income increases.
4. Females have a wider spread of BMI than males.
5. For the same level of education, males make more income than females.
6. For the most recent year, 2014, for both male and female, as the income and education level increases, the BMI decreases.
7. For both male and female, the highest income bracket (income>175000), there are more people with bachelors degree.

Further statistical analysis is needed to test the above hypothesis.