Navigating the Gender Income Gap

Anurag Gandhi

October 17, 2016

Contents

Outline

Objective of this report is to answer the following question:

> Is there a significant difference in income between men and women? Does the difference vary depending on other factors?

Data:

We are going to use the NLSY97 (National Longitudinal Survey of Youth, 1997 cohort) data set. This data set contains survey responses on thousands of individuals who have been surveyed every one or two years starting in 1997.

Approach outline:

- **Exploring the data set:** First we are going to explore the different variables in the dataset, assign them meaningful names, map values to logical factors, derive more variables, and treat missing data points. We will also select a potential list of variable that might help us in answering the question about income gap.
- **Selecting variables:** In this section, we will look at 3 variables: `race`, `age`, and `industry`. I believe these might be some of the most important predictors of income gap. We will create *data summaries* for these variables, try to identify outliers and interesting patterns. We will also talk about how income gap various across different levels of each of these variables, whether the gap is statistically significant or not, etc. Finally, we will run a *regression model* to include the *interactions* of each of these variables with gender and how these interactions impact income difference between men and women. **Note:** We will build the model in a **cumulative** fashion, adding one interaction/variable at a time, and testing its usefulness.
- **Adding more variables:** Using the approach we build in previous section, we will try to explore some more variables and see their impact on income gap.
- **Building the model:** Once we have the variables we think are predictors for income gap, we will try to optimize our final model by again testing the significance of each of the interaction, but this time using a bottom-up approach, i.e., removing one interaction variable at a time. We will then perform some diagnostics to check how appropriate our model is, and if there are any outliers affecting it's performance.
- **Final comments:** In this section, we will summarize our findings, and also talk about the potentials pitfalls of this analysis.

Note: All analyses and outputs in this project were produced using the statistical software R.

Exploring the dataset

The NLSY data set contains survey responses on thousands of individuals who have been surveyed every one or two years starting in 1997. Let's take a look at the names of the variables.

```
##  [1] "E8043100" "E8043200" "E8043400" "R0000100" "R0069400" "R0070000"
##  [7] "R0323900" "R0513500" "R0514700" "R0514800" "R0514900" "R0515100"
## [13] "R0536300" "R0536401" "R0536402" "R0681300" "R0690800" "R0691200"
## [19] "R1200100" "R1200200" "R1201400" "R1204700" "R1235800" "R1302600"
## [25] "R1302700" "R1482600" "R1484900" "R1485000" "R1700500" "R1701100"
## [31] "R2191500" "R2600500" "R2600600" "R4908500" "S0920000" "S0920700"
## [37] "S2011600" "S2022700" "S4677200" "S4685800" "S6645400" "T1069100"
## [43] "T1069101" "T1069102" "T1069103" "T5211700" "T6650100" "T6656700"
## [49] "T6656900" "T6657000" "T6657100" "T6657300" "T6767000" "T7635600"
## [55] "T7635700" "T7635800" "T7638800" "T7639200" "T7639800" "T7640000"
## [61] "T7640300" "T7640400" "T7711600" "T7731100" "T8122500" "T8976500"
## [67] "T8976700" "T8976800" "T8977600" "T8978000" "T8978100" "Z9033700"
## [73] "Z9033900" "Z9034100" "Z9050100" "Z9050500" "Z9050600" "Z9050700"
## [79] "Z9122500"
```

These are a lot of variables. We probably do not need all of them for our analysis. Also, the names of these variables do not convey any meaningful information. So let's use the "Question Code" from survey to replace all of these variables.

```
##  [1] "INCARC_TOTNUM_XRND"             "INCARC_AGE_FIRST_XRND"
##  [3] "INCARC_LENGTH_LONGEST_XRND"     "PUBID_1997"
##  [5] "YSCH-36400_1997"                "YSCH-37000_1997"
##  [7] "YSAQ-010_1997"                  "YEXP-300_1997"
##  [9] "YEXP-1500_1997"                 "YEXP-1600_1997"
## [11] "YEXP-1800_1997"                 "YEXP-2000_1997"
## [13] "KEY_SEX_1997"                   "KEY_BDATE_M_1997"
## [15] "KEY_BDATE_Y_1997"               "PC9-002_1997"
## [17] "PC12-024_1997"                  "PC12-028_1997"
## [19] "CV_BIO_MOM_AGE_CHILD1_1997"     "CV_BIO_MOM_AGE_YOUTH_1997"
## [21] "CV_ENROLLSTAT_1997"             "CV_HH_NET_WORTH_P_1997"
## [23] "CV_SAMPLE_TYPE_1997"            "CV_HGC_RES_DAD_1997"
## [25] "CV_HGC_RES_MOM_1997"            "KEY_RACE_ETHNICITY_1997"
## [27] "FP_YMFRELAT_1997"               "FP_YFMRELAT_1997"
## [29] "YSCH-6800_1998"                 "YSCH-7300_1998"
## [31] "YSAQ-372B_1998"                 "FP_YMFRELAT_1998"
## [33] "FP_YFMRELAT_1998"               "YSAQ-371_2000"
## [35] "YSAQ-282J_2002"                 "YSAQ-282Q_2002"
## [37] "CV_HH_NET_WORTH_Y_2003"         "CV_BIO_CHILD_HH_2003"
## [39] "YSAQ-000B_2004"                 "YSAQ-373_2004"
## [41] "YSAQ2-292_2005"                 "YTEL-52~000001_2007"
## [43] "YTEL-52~000002_2007"            "YTEL-52~000003_2007"
## [45] "YTEL-52~000004_2007"            "CV_BIO_CHILD_HH_2010"
## [47] "CV_COLLEGE_TYPE.01_2011"        "CV_INCOME_FAMILY_2011"
## [49] "CV_HH_SIZE_2011"                "CV_HH_UNDER_18_2011"
## [51] "CV_HH_UNDER_6_2011"             "CV_HIGHEST_DEGREE_1112_2011"
## [53] "YSCH-3112_2011"                 "YSAQ-000A000001_2011"
```

```
## [55] "YSAQ-000A000002_2011"          "YSAQ-000B_2011"
## [57] "YSAQ-360C_2011"                "YSAQ-364D_2011"
## [59] "YSAQ-371_2011"                 "YSAQ-372CC_2011"
## [61] "YSAQ-373_2011"                 "YSAQ-374_2011"
## [63] "YHEA29-285_2011"               "YEMP_INDCODE-2002.01_2011"
## [65] "VERSION_R16_2013"              "YINC_1400_2013"
## [67] "YINC_1700_2013"                "YINC_1800_2013"
## [69] "YINC_2400_2013"                "YINC_2600_2013"
## [71] "YINC_2700_2013"                "CVC_SAT_MATH_SCORE_2007_XRND"
## [73] "CVC_SAT_VERBAL_SCORE_2007_XRND" "CVC_ACT_SCORE_2007_XRND"
## [75] "CVC_ASSETS_DEBTS_20_XRND"      "CVC_TTL_JOB_TEEN_XRND"
## [77] "CVC_TTL_JOB_ADULT_ET_XRND"     "CVC_TTL_JOB_ADULT_ALL_XRND"
## [79] "CVC_ASSETS_DEBTS_30_XRND"
```

OK. Now, we have something to work on. The next step is to reduce this list to only keep potentially important variables. For the scope of this project, I have chosen the following variables to proceed with:

- INCARC_TOTNUM_XRND: The total number of incarnations - might have some relation with respondent's criminal history
- KEY_SEX_1997: This is the gender variable
- KEY_BDATE_M_1997: The birth month of respondent
- KEY_BDATE_Y_1997: The birth year of respondent
- PC12-024_1997: If female is depressed/sad in 1997
- PC12-028_1997: If male is depressed/sad in 1997
- KEY_RACE_ETHNICITY_1997: Race of respondent
- YSAQ-372B_1998: Drug usage information in 1998
- YSAQ-371_2000: Question about marijuana consumption in 2000
- CV_HH_NET_WORTH_Y_2003: Total household net worth in 2003
- CV_HH_SIZE_2011: Total household size in 2011
- CV_HIGHEST_DEGREE_1112_2011: Highest degree obtained by respondent before 2013
- YSAQ-360C_2011: Whether the respondent smokes in 2011
- YSAQ-371_2011: Whether the respondent takes marijuana in 2011
- YSAQ-372CC_2011: Whether the respondent takes drugs in 2011
- YEMP_INDCODE-2002.01_2011: industry code where respondent is employed
- YINC_1700_2013: Total income and wages of respondent in 2013
- YINC_2400_2013: Whether spouse earned any income in last year
- YINC_2600_2013: Spouse's total income and wages
- CV_INCOME_FAMILY_2011: Gross family income in 2011

Transforming variables

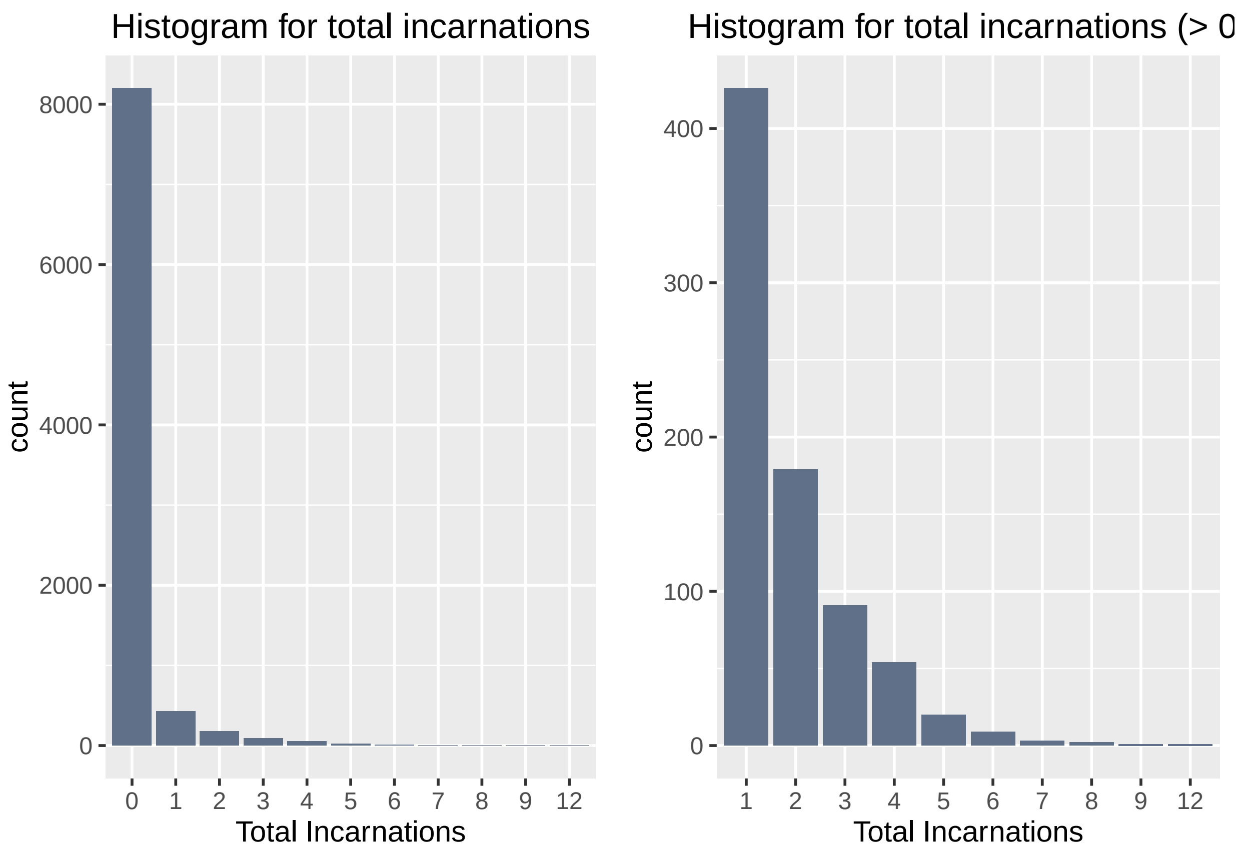
Let's give these variables some more meaningful names:

```
##  [1] "total.incarnations"     "gender"
##  [3] "birth.month"            "birth.year"
##  [5] "female.depression.1997" "male.depression.1997"
##  [7] "enrollstat.1997"        "hh.net.worth.1997"
##  [9] "race"                   "drug.use.1998"
## [11] "marijuana.2000"         "hh.net.worth.2003"
## [13] "hh.size.2011"           "highest.degree.2011"
```

```
## [15] "smoke.2011"             "marijuana.2011"
## [17] "drug_use.2011"          "industry.2011"
## [19] "income.2013"            "spouse.earned.2013"
## [21] "gross.family.income"    "spouse.income"
```

Now, let's transform some of these variables. We are going to perform the following transformations:

1. Change levels in `gender` to `male` and `female`
2. Modify `total.incarnations` from continuous to categorical: Look at the distribution of total incarnations variable:



Most of the respondents have 0 incarnations and only few of them have above 5. Treating incarnations as a continuous variable is thus probably not a good idea. So we are going to create 3 bins: `none`, `less than 2`, and `more than 2`

3. **Race:** Race is straightforward. We are going to map the respective numeric codes to that race's descriptive name.
4. **Drug use question:** We are going to decode the answers for `Yes`, `No` and `Refusal` as the three main levels and code the rest of the values (missing and non-interviews) as `unknown`. Refusal to share this information might have some significance for income, so we are considering this.
5. **Spouse factors:**

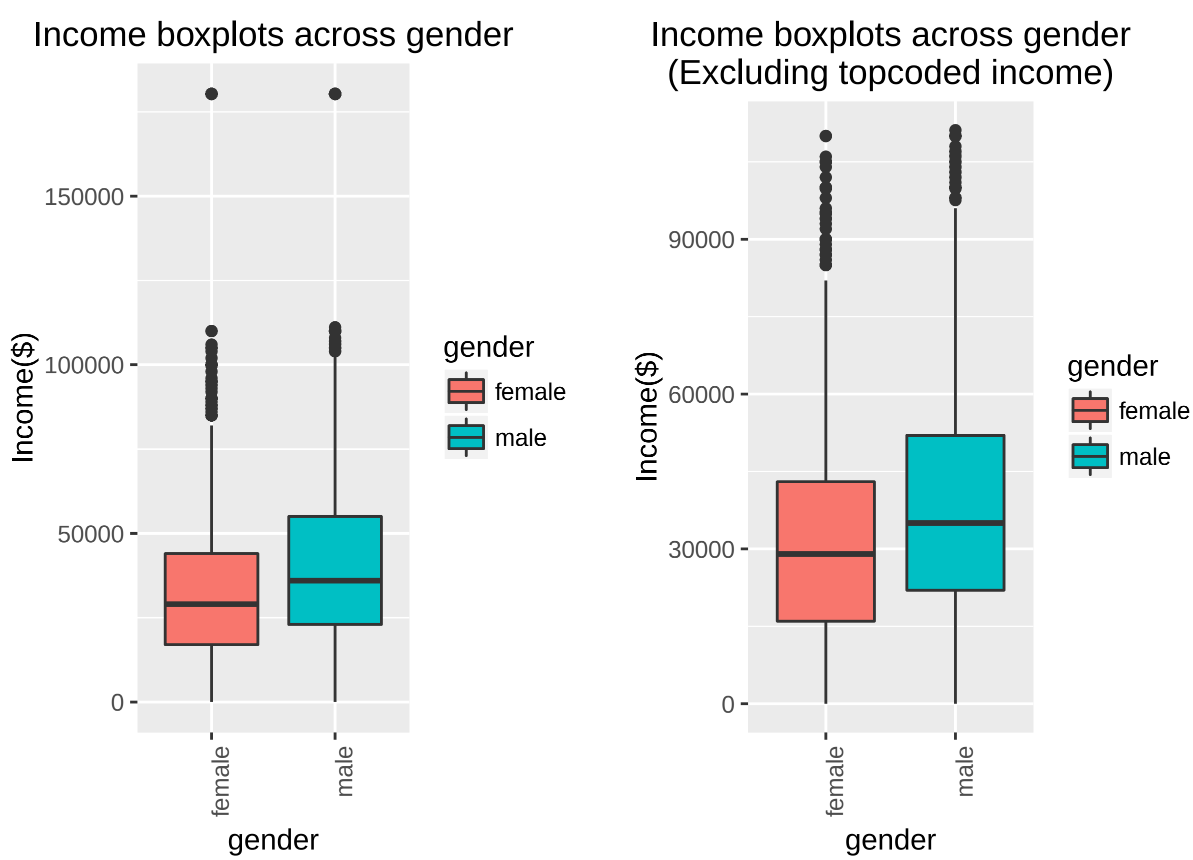
- spouse earned in last year?: For this variable, map the responses to standard `yes`, `no`, and `refusal`. The universe for this is married respondents. So, someone who *valid skips* this question is single. We will give these data points value: `no spouse`

- marital status: This can be derived from the previous question. This will have 2 levels: single and married
- spouse income in last year: The description of this variable says top 2% of the values are top coded. We will deal with this in the next session

6. **Highest Degree Earned:** This is also straightforward mapping of codes to their descriptive strings. We will code the missing values as `unknown` for now.
7. **Industry:** Proceed as usual. Code missing as `unknown`.
8. **Age:** Approximate age can be derived from birth year. We will assume the reference year to be 2013 and count years from birth year to get the respondent's age.
9. **Negative values in numeric variables:** For the scope of this project, I will code negative values in income variables as NA. Note that, however, these are not negative incomes. These are just skipped answers/non-interviews/refusals. This might not be the best approach, but income is a continuous variable and these values might affect the results of our model.
10. **Topcoded values in numeric variables:** Some of the variables have topcoded values. In this project, I have tried to show some summaries by including topcoded values as well. But we will later see how these values affect our regression, and then take a decision to either include or exclude them.

OK. We have the data in the right shape (for now). Now, let's look at our main variables - income and gender.

Income by gender



So, the male on average does earn more than a female. But is this difference significant? Let's do a t-test to test the significance of this gap.

```
##
##  Welch Two Sample t-test
##
## data:  income.2013 by gender
## t = -12.804, df = 4977.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12065.617  -8861.345
## sample estimates:
## mean in group female   mean in group male
##             32757.00             43220.48
```

The p-value is 0. Yes the gap is significant! Let's now look at some variables that might be affecting/related to this gap.

Selecting variables

---

1. Race

Data Summary

Let's first look at the variable `race`. I am choosing `race` because I feel there might be some variance in income gap across different races.

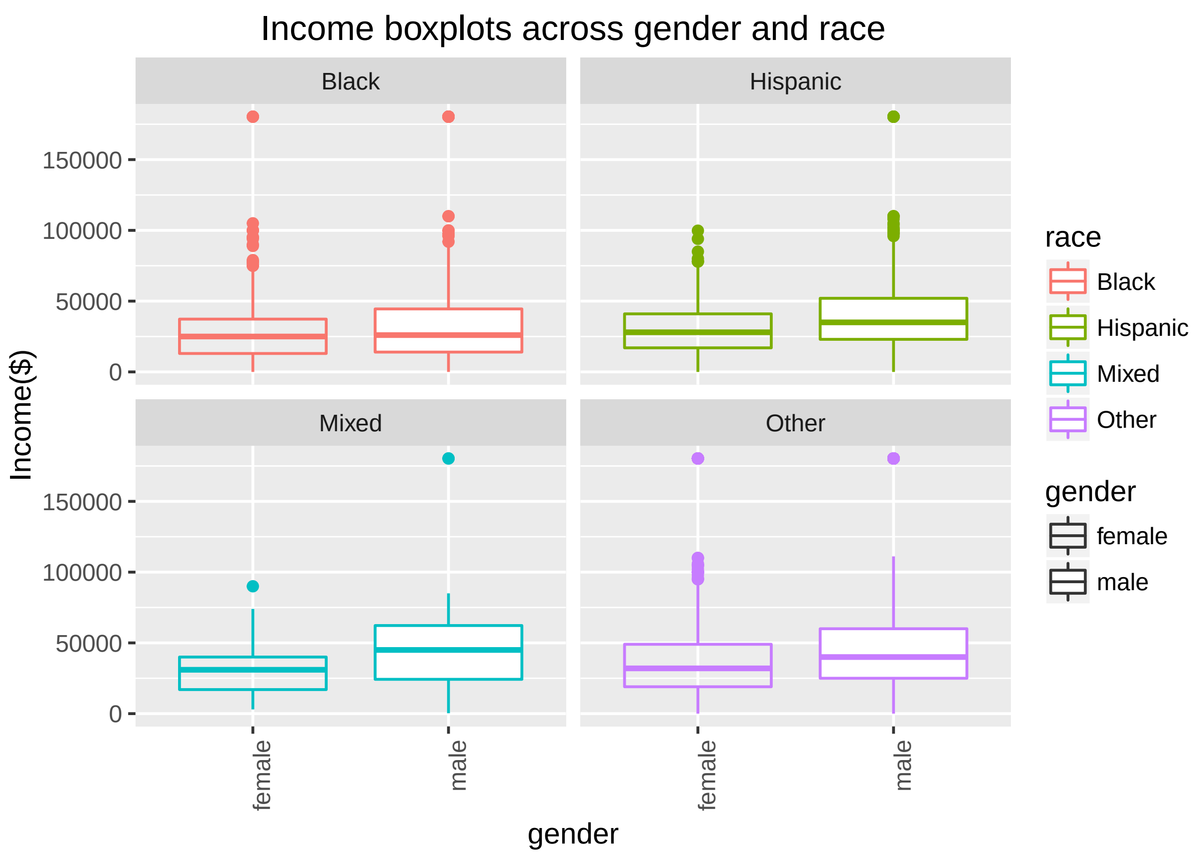
Below is a summary of how income gap, as well as average income varies across different types of `race`.

Race	Income Gap ($)	No. of respondents	% of males	% of females	Mean Income ($)	Std. Deviation: Income
Black	5526.10	1290	46.74%	53.26%	30356.84	26408
Hispanic	11249.24	1105	53.76%	46.24%	35731.97	25100
Mixed	15877.02	47	55.32%	44.68%	43644.94	38202
Other	11282.57	2783	54.40%	45.60%	42794.97	32939

The highest income gap is for `Mixed` race. However, notice that number of respondents are too low when compared to other `race` types. The standard deviation is also the highest for this group. So, we can not really hold this conclusion.

Respondents from `Black` race, however, have the lowest income gap. This might be something we would be interested to verify.

Let's plot the distribution of income across these `race` types.

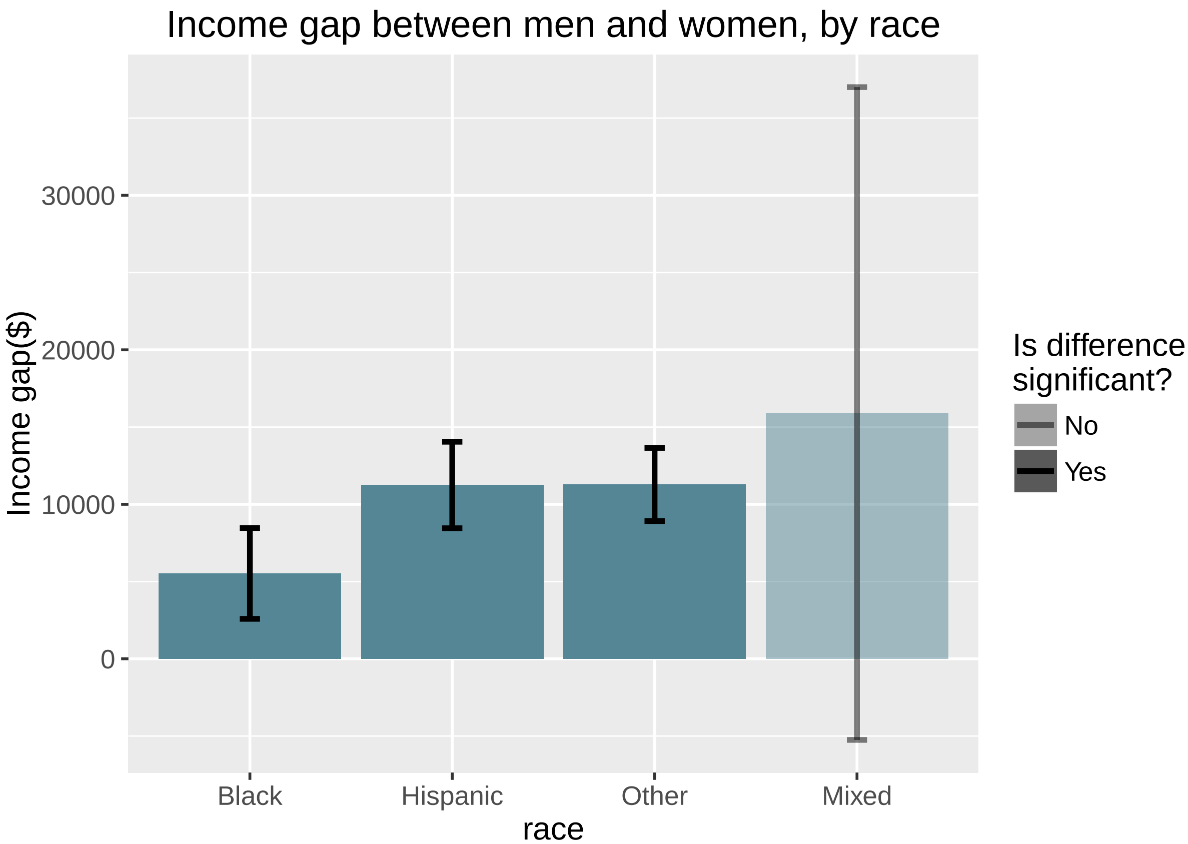


The above plot shows box plots of `income` distribution across race types. Notice the outliers for every `race` types. These are the top coded 2% data points for income. To get a clearer picture, let's ignore these data points for a while.

This tends to be in line with earlier observation for `Black` race. Also, notice that income gap seems to be higher for `Hispanic` and `Other` races.

Testing the difference in means

For each of the `race` group, we will perform a t-test (two sample hypothesis test) to test whether there is a statistically significant difference between males and females of a particular `race`. The variation of income gap by `race`, along with the test results is shown in bar plots below. The error bars indicate confidence intervals. **Note**: The significance results are tested at significance level = 0.05.



Observation: Note that for Mixed race, error bars are extremely wide. Again, this is because we do not have sufficient data to correctly estimate income gap between males and female for this race. The income gap is a significant difference for **Black, Other** and **Hispanic** races.

Building the model

Now, let's run a linear regression model with `income` as a linear function of `gender` and `race`, i.e.,

```
## lm(formula = income.2013 ~ gender + race, data = nlsy.subset)
```

Below is an output of coefficients from the model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25733	907	28.36	0.0000
gendermale	9892	820	12.07	0.0000
raceHispanic	4682	1212	3.86	0.0001
raceMixed	12440	4386	2.84	0.0046
raceOther	11681	997	11.72	0.0000

How to interpret this model? The coefficient for `gendermale` is the estimate of **income-gap** between males and females. It says that if you control for other variables(`race` in this case), males earn $9892 more than females. That is, there is an income gap of $9892 between respondents of the same race.

But this does not tell us how this **gap** varies across race. To get that, we need to model **interactions** between `gender` and `race`. Let's update our model to add an interaction term `gender*race`.

```
## lm(formula = income.2013 ~ gender + race + gender:race, data = nlsy.subset)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27774	1126	24.66	0.0000
gendermale	5526	1647	3.36	0.0008
raceHispanic	1911	1724	1.11	0.2677
raceMixed	7088	6539	1.08	0.2784
raceOther	8883	1398	6.35	0.0000
gendermale:raceHispanic	5723	2426	2.36	0.0183
gendermale:raceMixed	10351	8815	1.17	0.2404
gendermale:raceOther	5756	1994	2.89	0.0039

How to interpret this model? The coefficient for gender-race interaction terms is the estimate of how **difference in income-gap** varies across race. For instance, it says that the income gap between Hispanic males and females is about $5723 more than income gap between Black males and females.

Is the interaction term significant? To test this, let's perform an ANOVA test to test the difference between the updated and simple linear model.

```
## Analysis of Variance Table
##
## Model 1: income.2013 ~ 1
## Model 2: income.2013 ~ gender + race + gender:race
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   5224 4.8266e+12
## 2   5217 4.5450e+12  7 2.8161e+11 46.177 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is statistically significant, so we can reject the null that income gap is same across all race categories. That is, this shows that **the income gap between males and female does vary by race.**

Dealing with topcoded values

Does the topcoded value for top 2% earners affect our model in any way? To answer this, we will build the model again with topcoded income data points removed and then see the change in coefficients.

The coefficients from the updated model are shown below:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	26880	837	32.13	0.0000
gendermale	3180	1229	2.59	0.0097
raceHispanic	2805	1279	2.19	0.0283
raceMixed	7982	4844	1.65	0.0995
raceOther	7826	1040	7.52	0.0000
gendermale:raceHispanic	5925	1806	3.28	0.0010
gendermale:raceMixed	1898	6648	0.29	0.7753
gendermale:raceOther	4305	1490	2.89	0.0039

Testing the impact of topcoded datapoints on our model: Notice the change in coefficient for `gendermale:raceHispanic`. This changes our interpretation to: *The income gap between Hispanic males and females is about $5925 more than income gap between Black males and females.*

Inclusion of topcoded values increases this **variation in income gap** by $202 for `Hispanic` race. Also, see the impact on coefficients of `gendermale` and `raceOther`. For `gendermale`, the coefficients seem to have changed by a greater margin. Thus, to minimize this impact we are going to use income **without** the topcoded values as response.

2. Industry

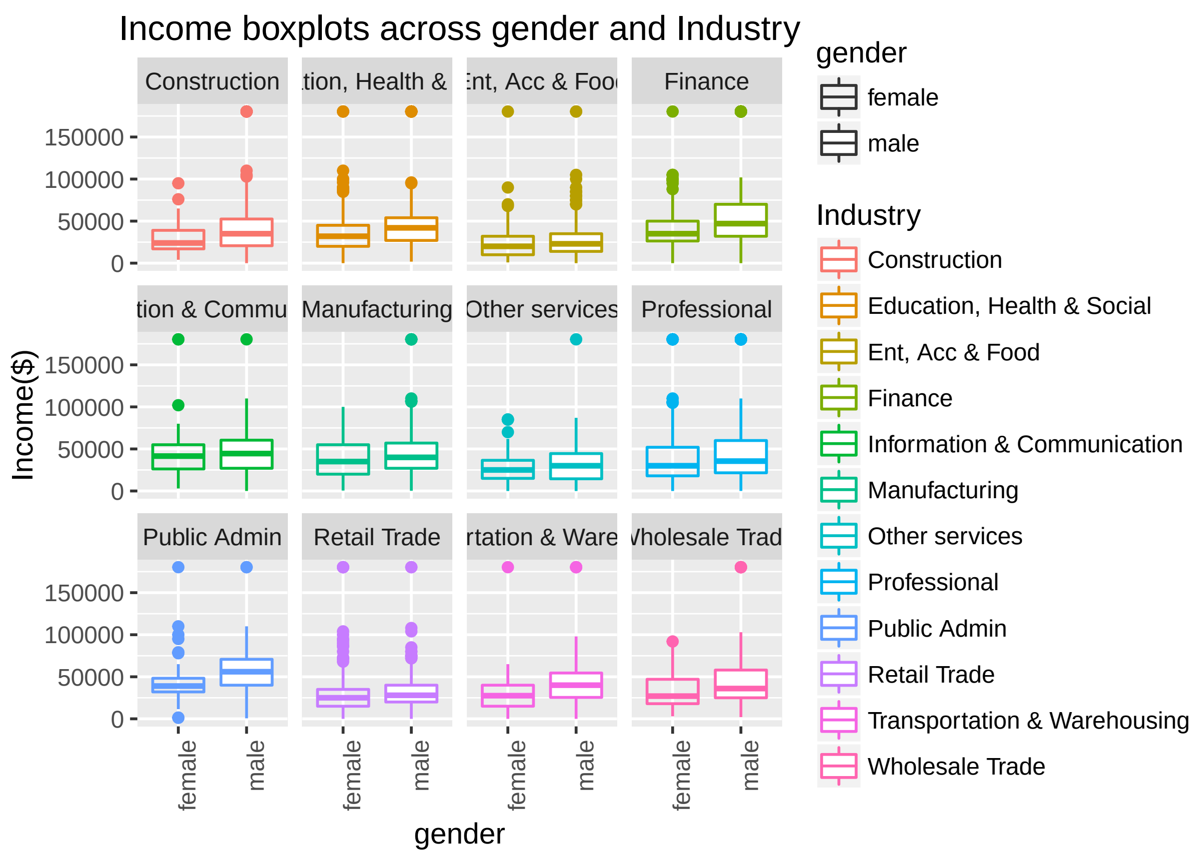
Let's look at the industry of respondent next. You would expect some relation between industry and income. Type of job does define salary. So let's look at the income gap across industries.

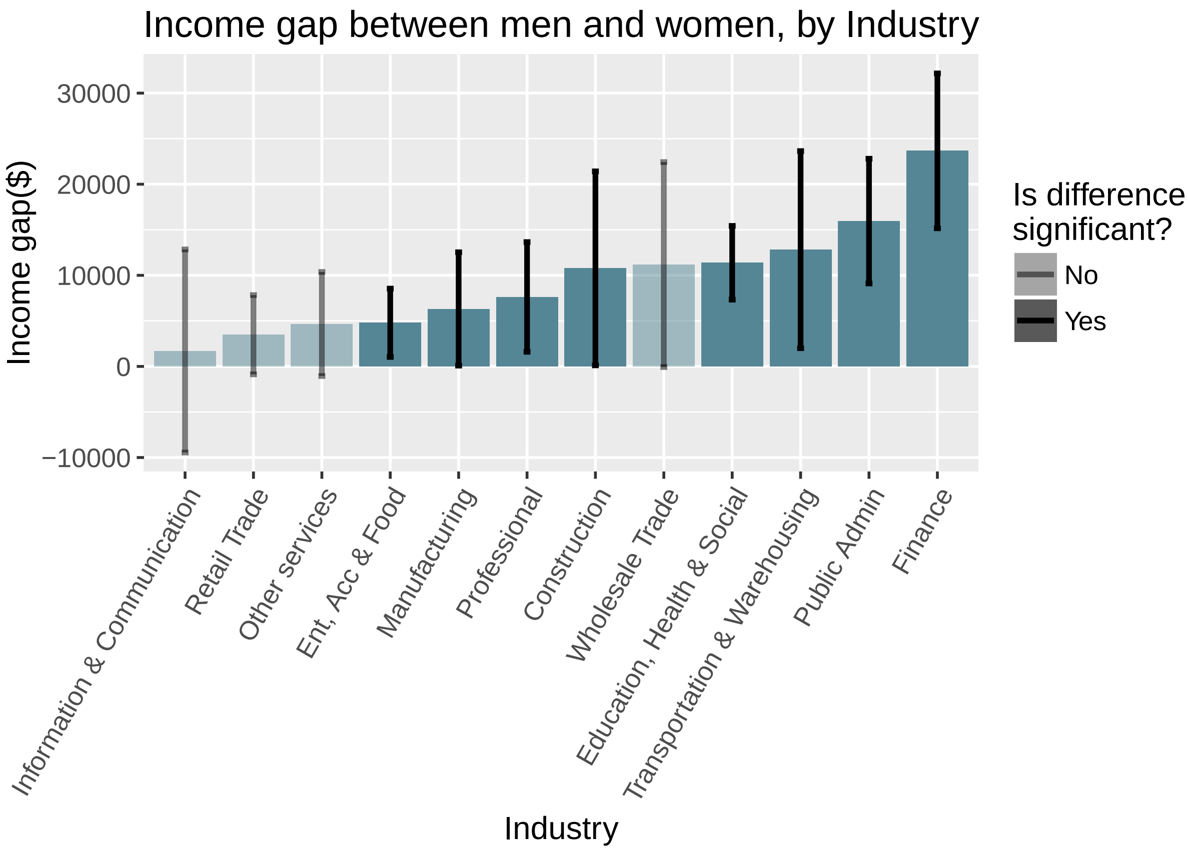
Again, let's summarize the income gap across industries.

Industry	Income Gap ($)	No. of respondents	% of males	% of females	Mean Income ($)	S
Acs Special	-23640.00	7	28.57%	71.43%	39385.71	
Agriculture	27302.76	29	58.62%	41.38%	35614.07	
Construction	10766.49	295	92.20%	7.80%	41375.67	
Education, Health & Social	11377.95	1081	24.14%	75.86%	36560.53	
Ent, Acc & Food	4790.99	499	48.90%	51.10%	25318.18	
Finance	23665.15	341	44.87%	55.13%	50005.28	
Information & Communication	1697.70	121	52.89%	47.11%	46547.17	
Manufacturing	6310.79	354	74.86%	25.14%	43442.90	
Military	16473.68	20	95.00%	5.00%	60650.00	
Mining	47105.41	30	96.67%	3.33%	81535.23	
Other services	4659.90	226	52.65%	47.35%	29760.87	
Professional	7628.01	604	60.60%	39.40%	44505.90	
Public Admin	15950.85	210	56.19%	43.81%	51791.20	
Retail Trade	3479.26	516	49.03%	50.97%	31117.65	
Transportation & Warehousing	12818.21	159	74.84%	25.16%	41111.58	
Unknown	10302.20	570	54.56%	45.44%	30913.15	
Utilities	45438.92	31	70.97%	29.03%	64013.65	
Wholesale Trade	11181.53	132	78.03%	21.97%	43332.74	

Note: Acs Special, Agriculture and Utilities have very low number of respondents. Also, the proportion of males and females in these industries is disproportionate based on the data for few respondents we have. For better presentation, I have removed these in subsequent plots. However, we will include these in our regression model.

For the rest of the industries, let's look at the gaps using box plots and histograms like we did for `race`.


Income boxplots across gender and Industry
Construction
tion, Health &
Ent, Acc & Food
Finance
tion & Commu
Manufacturing
Other services
Professional
Public Admin
Retail Trade
rtation & Ware
Wholesale Trad
150000
100000
50000
0
Income($)
female
male
gender
gender
female
male
Industry
Construction
Education, Health & Social
Ent, Acc & Food
Finance
Information & Communication
Manufacturing
Other services
Professional
Public Admin
Retail Trade
Transportation & Warehousing
Wholesale Trade



Observation: Notice, in both the plots, **Finance, Public Admin, Construction, Transportation & Warehousing, Construction, and Education, Health & social** are the industries where income gap seems to be most prevalent. Of course, this is just based on difference in means, and might be driven by other factors. We will explore this variable further in our model.

Building the model

Let's start by adding `industry` to our interaction model for `gender` and `race`. This will tell us how `industry` impacts income. **Note:** We are going to change the reference level to `Ent, Accs & Food`.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16813	1204	13.96	0.0000
gendermale	3589	1204	2.98	0.0029
raceHispanic	2604	1228	2.12	0.0340
raceMixed	10257	4643	2.21	0.0272
raceOther	7869	1000	7.87	0.0000
industryAcs Special	15870	7969	1.99	0.0465
industryAgriculture	8783	4003	2.19	0.0283
industryConstruction	10019	1579	6.35	0.0000
industryEducation, Health & Social	12006	1150	10.44	0.0000
industryFinance	17460	1502	11.62	0.0000
industryInformation & Communication	18089	2145	8.43	0.0000
industryManufacturing	14672	1473	9.96	0.0000
industryMilitary	32041	4786	6.70	0.0000
industryMining	33974	4301	7.90	0.0000

	Estimate	Std. Error	t value	Pr(>|t|)
industryOther services	4317	1683	2.57	0.0103
industryProfessional	12476	1285	9.71	0.0000
industryPublic Admin	24846	1736	14.32	0.0000
industryRetail Trade	4536	1323	3.43	0.0006
industryTransportation & Warehousing	12417	1936	6.42	0.0000
industryUnknown	4331	1291	3.35	0.0008
industryUtilities	30092	4007	7.51	0.0000
industryWholesale Trade	11053	2091	5.29	0.0000
gendermale:raceHispanic	5346	1732	3.09	0.0020
gendermale:raceMixed	453	6372	0.07	0.9433
gendermale:raceOther	2975	1433	2.08	0.0380

Notice that p-values for coefficients of most of the industries fall above significance level. Does that mean industry does not affect income? Not necessarily. But we can test the significance of this **relationship** in our model using ANOVA test.

```
## Analysis of Variance Table
##
## Model 1: income.exclude.topcode ~ gender + race + industry + gender:race
## Model 2: income.exclude.topcode ~ gender * race
##   Res.Df        RSS  Df   Sum of Sq      F    Pr(>F)
## 1   5092 2.2300e+12
## 2   5109 2.4423e+12 -17 -2.1231e+11 28.517 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for this test is below significance level (0.05) and so there does seem to be some association between industry and income. But, we are more interested in how industry impacts **income gap**. So let's add an interaction term instead.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17910	1479	12.11	0.0000
gendermale	1644	2111	0.78	0.4362
raceHispanic	2223	1231	1.81	0.0711
raceMixed	10301	4641	2.22	0.0265
raceOther	7452	1006	7.41	0.0000
industryAcs Special	24360	9434	2.58	0.0098
industryAgriculture	-4498	6180	-0.73	0.4667
industryConstruction	9202	4551	2.02	0.0432
industryEducation, Health & Social	11124	1501	7.41	0.0000
industryFinance	16152	2015	8.02	0.0000
industryInformation & Communication	18211	3108	5.86	0.0000
industryManufacturing	15410	2577	5.98	0.0000
industryMilitary	27090	20935	1.29	0.1957
industryMining	10638	20928	0.51	0.6113
industryOther services	4881	2407	2.03	0.0427
industryProfessional	13688	1898	7.21	0.0000
industryPublic Admin	19881	2555	7.78	0.0000
industryRetail Trade	4825	1846	2.61	0.0090
industryTransportation & Warehousing	6336	3594	1.76	0.0780
industryUnknown	2323	1849	1.26	0.2091

	Estimate	Std. Error	t value	Pr(>|t|)
industryUtilities	13363	7102	1.88	0.0599
industryWholesale Trade	11279	4098	2.75	0.0059
gendermale:raceHispanic	5714	1737	3.29	0.0010
gendermale:raceMixed	849	6371	0.13	0.8940
gendermale:raceOther	3370	1441	2.34	0.0194
gendermale:industryAcs Special	-30794	17579	-1.75	0.0799
gendermale:industryAgriculture	22985	8105	2.84	0.0046
gendermale:industryConstruction	1685	4919	0.34	0.7319
gendermale:industryEducation, Health & Social	1958	2408	0.81	0.4162
gendermale:industryFinance	2908	3022	0.96	0.3361
gendermale:industryInformation & Communication	-47	4289	-0.01	0.9912
gendermale:industryManufacturing	-392	3183	-0.12	0.9020
gendermale:industryMilitary	6020	21520	0.28	0.7797
gendermale:industryMining	25143	21400	1.17	0.2401
gendermale:industryOther services	-957	3361	-0.28	0.7758
gendermale:industryProfessional	-1693	2585	-0.65	0.5127
gendermale:industryPublic Admin	9058	3481	2.60	0.0093
gendermale:industryRetail Trade	-538	2641	-0.20	0.8385
gendermale:industryTransportation & Warehousing	8697	4301	2.02	0.0432
gendermale:industryUnknown	3849	2581	1.49	0.1359
gendermale:industryUtilities	24669	8609	2.87	0.0042
gendermale:industryWholesale Trade	354	4803	0.07	0.9412

We made some inferences from plots earlier about industries like Finance, Construction etc. **Are they in line with the coefficients?** For instance, let's take `Finance` industry. the interaction term coefficient is called `gendermale:industryFinance`, which estimates the gender gap in Finance industry. **How do you interpret this coefficient?** It says that the income gap between males and females working in Finance industry is about $2908 more than income gap between males and females working in `Ent, Acc & Food`.

Significance of interaction term

We will again do a ANOVA test for gender-finance interaction term to establish whether `industry` is indeed a significant predictor of income gap.

```
## Analysis of Variance Table
##
## Model 1: income.exclude.topcode ~ race + gender:race
## Model 2: income.exclude.topcode ~ gender * race + gender * industry
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   5109 2.4423e+12
## 2   5075 2.2132e+12 34 2.2911e+11 15.452 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is statistically significant, so we can reject the null that income gap is same across all industries. That is, this shows that **the income gap between males and female does vary by industry.**

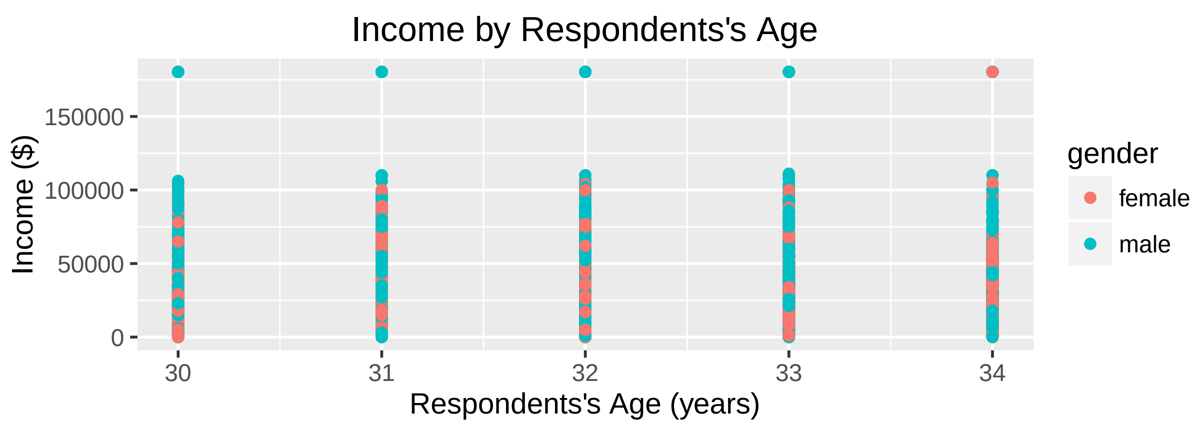
Based on the coefficients of this interaction term, let's sort the industries in descending order of income gap. **Note:** This list will exclude Military, Mining, Utilities, Unknown, Acs Special, and Agriculture for reasons highlighted before.

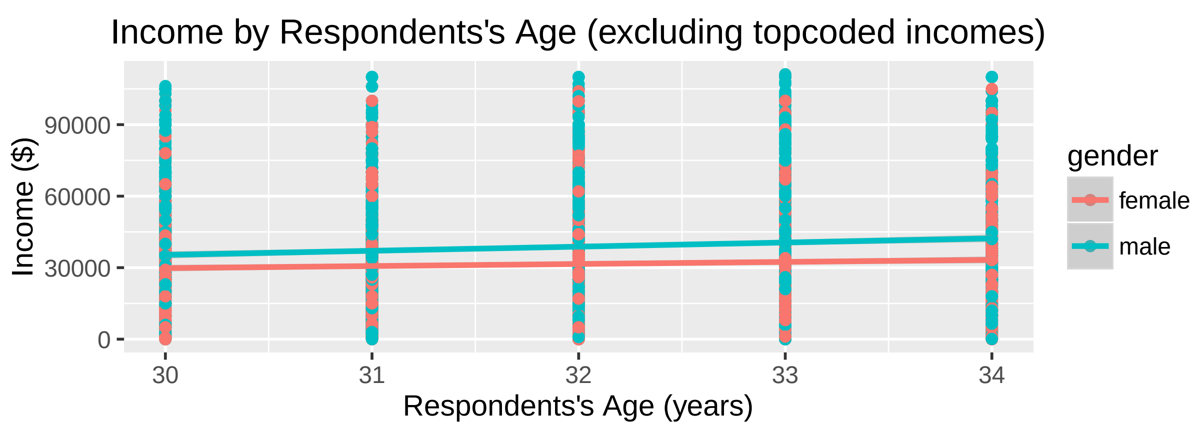
	Industry	Estimate
8	Public Admin	9057.63918
10	Transportation & Warehousing	8696.64951
3	Finance	2907.69636
2	Education, Health & Social	1958.02809
1	Construction	1684.97852
11	Wholesale Trade	354.31061
4	Information & Communication	-47.12472
5	Manufacturing	-391.88431
9	Retail Trade	-538.36336
6	Other services	-957.31926
7	Professional	-1692.76966

Observation: Finance, Public admin, Transportation & Warehousing, COnstruction validate our intial findings. This industries seem to have a larger income gap between men and women.

3. Age

First let's examine relationship between age and income. We are going to look at both versions, with and without topcoded data points.



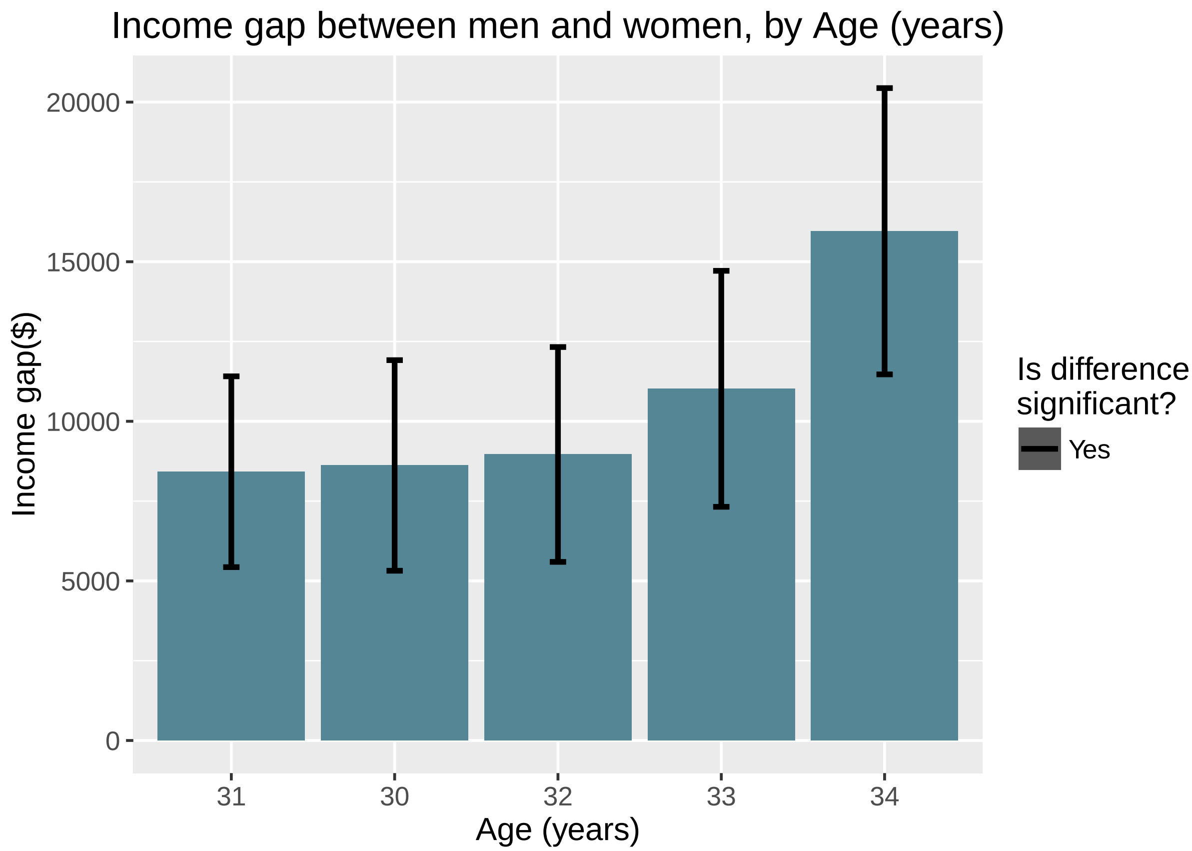


Well, that seems to be a very slight increasing trend. It's hard to tell really. There is not much variation in ages to begin with. All respondents are 30-34 years of age.

Difference in means

Since there are not a lot of values, we can even convert age to a categorical variable and see how income gap varies.

Below is a histogram showing the results:



There seems to be a statistically significant income gap (p-value < 0.05) for all ages. **But how does this gap varies by age?** There seems to be a It would be worth checking this aspect out in our model.

Building the model

Let's start by adding `age` to our interaction model. This will tell us how `age` impacts income. **Note:** We are going to treat age as a factor variable.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16014	1579	10.14	0.0000
gendermale	1619	2106	0.77	0.4420
raceHispanic	2182	1228	1.78	0.0757
raceMixed	10627	4630	2.30	0.0218
raceOther	7443	1003	7.42	0.0000
industryAcs Special	24187	9409	2.57	0.0102
industryAgriculture	-4679	6165	-0.76	0.4479
industryConstruction	9409	4539	2.07	0.0382
industryEducation, Health & Social	10824	1497	7.23	0.0000
industryFinance	15882	2010	7.90	0.0000
industryInformation & Communication	17864	3100	5.76	0.0000
industryManufacturing	15117	2572	5.88	0.0000
industryMilitary	25421	20882	1.22	0.2235

	Estimate	Std. Error	t value	Pr(>|t|)
industryMining	12459	20874	0.60	0.5506
industryOther services	5063	2402	2.11	0.0351
industryProfessional	13635	1893	7.20	0.0000
industryPublic Admin	19748	2548	7.75	0.0000
industryRetail Trade	4633	1841	2.52	0.0119
industryTransportation & Warehousing	6444	3585	1.80	0.0723
industryUnknown	2323	1844	1.26	0.2079
industryUtilities	12582	7084	1.78	0.0758
industryWholesale Trade	11036	4087	2.70	0.0069
age31	85	910	0.09	0.9256
age32	3566	920	3.88	0.0001
age33	3695	917	4.03	0.0001
age34	3146	951	3.31	0.0009
gendermale:raceHispanic	5784	1733	3.34	0.0008
gendermale:raceMixed	323	6357	0.05	0.9595
gendermale:raceOther	3494	1437	2.43	0.0151
gendermale:industryAcs Special	-28856	17534	-1.65	0.0999
gendermale:industryAgriculture	22453	8085	2.78	0.0055
gendermale:industryConstruction	1143	4907	0.23	0.8158
gendermale:industryEducation, Health & Social	2074	2401	0.86	0.3877
gendermale:industryFinance	2906	3014	0.96	0.3351
gendermale:industryInformation & Communication	99	4278	0.02	0.9816
gendermale:industryManufacturing	-424	3175	-0.13	0.8938
gendermale:industryMilitary	7099	21469	0.33	0.7409
gendermale:industryMining	23118	21346	1.08	0.2788
gendermale:industryOther services	-1097	3354	-0.33	0.7436
gendermale:industryProfessional	-1775	2578	-0.69	0.4912
gendermale:industryPublic Admin	8719	3473	2.51	0.0121
gendermale:industryRetail Trade	-270	2635	-0.10	0.9184
gendermale:industryTransportation & Warehousing	8240	4290	1.92	0.0548
gendermale:industryUnknown	3651	2575	1.42	0.1562
gendermale:industryUtilities	24809	8586	2.89	0.0039
gendermale:industryWholesale Trade	140	4791	0.03	0.9768

All ages above 31 show a positive association with income when compared to age 30. With age, income increases? Not necessarily but this does imply some positive correlation.

```
## Analysis of Variance Table
##
## Model 1: income.exclude.topcode ~ gender * race + gender * industry
## Model 2: income.exclude.topcode ~ gender + race + industry + age + gender:race +
##     gender:industry
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   5075 2.2132e+12
## 2   5071 2.1985e+12  4 1.4695e+10 8.4736 8.235e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for this test is below significance level (0.05) and so there does seem to be some association between age and income. But, like before, we are more interested in how age impacts **income gap**. So let's add an interaction term instead:

```
## lm(formula = income.exclude.topcode ~ gender + race + industry +
##     age + gender:race + gender:industry + gender:age, data = nlsy.subset)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16677	1682	9.92	0.0000
gendermale	339	2394	0.14	0.8873
raceHispanic	2167	1229	1.76	0.0779
raceMixed	10439	4632	2.25	0.0243
raceOther	7432	1003	7.41	0.0000
industryAcs Special	24318	9414	2.58	0.0098
industryAgriculture	-4688	6168	-0.76	0.4473
industryConstruction	9317	4539	2.05	0.0402
industryEducation, Health & Social	10921	1498	7.29	0.0000
industryFinance	15995	2012	7.95	0.0000
industryInformation & Communication	17936	3101	5.78	0.0000
industryManufacturing	15272	2574	5.93	0.0000
industryMilitary	25729	20892	1.23	0.2182
industryMining	12005	20883	0.57	0.5654
industryOther services	5028	2404	2.09	0.0366
industryProfessional	13649	1893	7.21	0.0000
industryPublic Admin	19809	2548	7.77	0.0000
industryRetail Trade	4683	1842	2.54	0.0110
industryTransportation & Warehousing	6406	3585	1.79	0.0741
industryUnknown	2339	1845	1.27	0.2049
industryUtilities	12881	7087	1.82	0.0692
industryWholesale Trade	11145	4088	2.73	0.0064
age31	-115	1316	-0.09	0.9306
age32	2594	1310	1.98	0.0478
age33	2788	1323	2.11	0.0351
age34	1569	1355	1.16	0.2472
gendermale:raceHispanic	5794	1733	3.34	0.0008
gendermale:raceMixed	387	6359	0.06	0.9515
gendermale:raceOther	3535	1438	2.46	0.0140
gendermale:industryAcs Special	-28623	17548	-1.63	0.1029
gendermale:industryAgriculture	22348	8090	2.76	0.0058
gendermale:industryConstruction	1142	4907	0.23	0.8159
gendermale:industryEducation, Health & Social	1903	2404	0.79	0.4285
gendermale:industryFinance	2705	3018	0.90	0.3701
gendermale:industryInformation & Communication	-53	4279	-0.01	0.9900
gendermale:industryManufacturing	-692	3180	-0.22	0.8278
gendermale:industryMilitary	6601	21477	0.31	0.7586
gendermale:industryMining	23506	21354	1.10	0.2710
gendermale:industryOther services	-1098	3356	-0.33	0.7435
gendermale:industryProfessional	-1872	2579	-0.73	0.4681
gendermale:industryPublic Admin	8466	3477	2.44	0.0149
gendermale:industryRetail Trade	-358	2636	-0.14	0.8919
gendermale:industryTransportation & Warehousing	8165	4291	1.90	0.0571
gendermale:industryUnknown	3548	2576	1.38	0.1684
gendermale:industryUtilities	24338	8590	2.83	0.0046
gendermale:industryWholesale Trade	-164	4794	-0.03	0.9727
gendermale:age31	440	1822	0.24	0.8094
gendermale:age32	1914	1840	1.04	0.2983
gendermale:age33	1766	1836	0.96	0.3362

	Estimate	Std. Error	t value	Pr(>|t|)
gendermale:age34	3110	1903	1.63	0.1022

Are the inferences from hypothesis tests in line with the coefficients? For instance, let's take age 32. the interaction term coefficient is called `gendermale:age32`, which estimates the income gap between men and women **within** respondents who are 32 years old in 2013. **How do you interpret this coefficient?** It says that the income gap between males and females working who are 32 years old is about $1914 more than income gap between males and females who are 30.

Significance of interaction term

We will again do a ANOVA test for gender-finance interaction term to establish whether `age` is indeed a significant predictor of income gap.

```
## Analysis of Variance Table
##
## Model 1: income.exclude.topcode ~ race + industry + gender:race + gender:industry
## Model 2: income.exclude.topcode ~ gender + race + industry + age + gender:race +
##     gender:industry + gender:age
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   5075 2.2132e+12
## 2   5067 2.1970e+12  8 1.6202e+10 4.6709 1.034e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value < 0.05 (significance level). We can hence reject the null and conclude that interaction between gender and age is a significant predictor in our model.

Implications: Although our model says that with age, income gap seems to increase, we need to remember that we have chosen a very small sample of ages to begin with. All respondents are 30-35 age. So although there might be some correlation between age and income difference, it would not necessarily hold true for other ages as well.

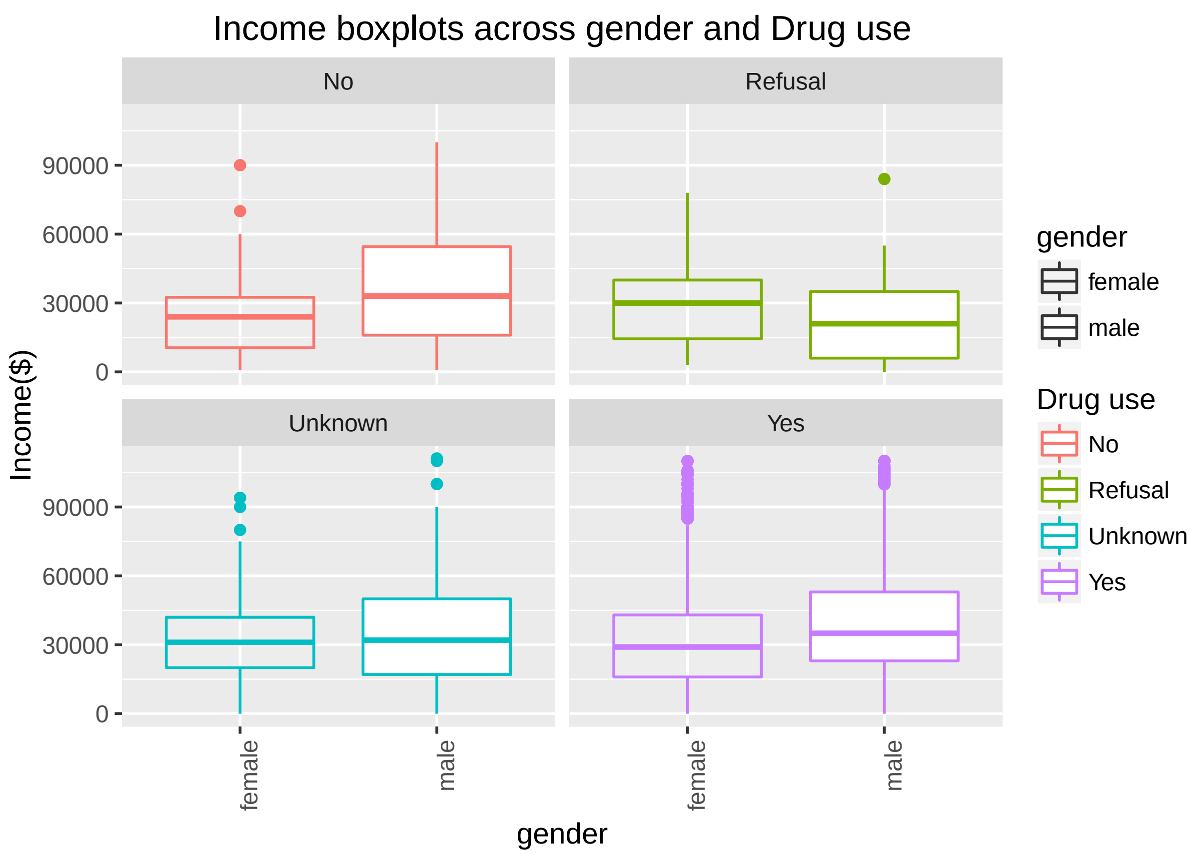
Adding more variables

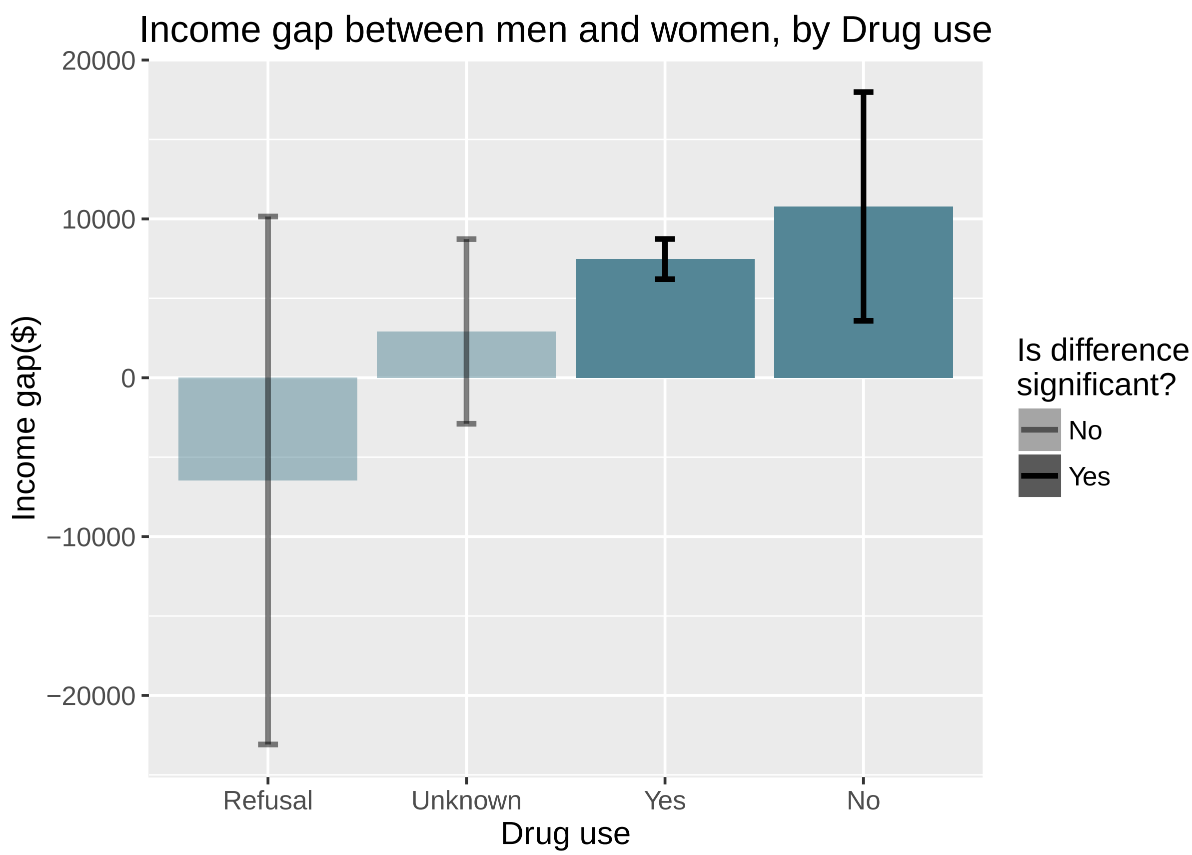
Till now, we have defined income and income gap as a function of race, industry, and age. Let's look at some other variables and add those to our model. The variables we are going to look at are:

- Drug Use
- Spouse factors
- Education (highest degree earned)
- Total Incarnations
- Household net worth
- Gross Family Income

4. Drug Use

The variable we are interested in is drug use since last survey. We will look at three responses to this question: "Yes", "No" and refusal to answer.


Income boxplots across gender and Drug use
No
Refusal
Unknown
Yes
90000
60000
30000
0
Income($)
female
male
gender
gender
female
male
Drug use
No
Refusal
Unknown
Yes



Observation: The difference is significant and larger for those who responded no. Does that mean drug users have less income gap?

Adding drug use-gender interaction term

Let's see the impact of drug use-gender interaction term on our regression model

```
## Analysis of Variance Table
##
## Model 1: income.exclude.topcode ~ race + industry + age + gender:race +
##     gender:industry + gender:age
## Model 2: income.exclude.topcode ~ gender + race + industry + age + drug_use.2011 +
##     gender:race + gender:industry + gender:age + gender:drug_use.2011
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   5067 2.1970e+12
## 2   5061 2.1814e+12  6 1.5606e+10 6.0346 2.644e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significant! The p-value is < 0.05. So the answer to this question seems to have some relationship with income gap.Let's look at the coefficients to see exactly what the model is saying.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12415	3188	3.89	0.0001

	Estimate	Std. Error	t value	Pr(>|t|)
gendermale	2286	4242	0.54	0.5900
raceHispanic	2073	1227	1.69	0.0911
raceMixed	10450	4620	2.26	0.0238
raceOther	7106	1006	7.06	0.0000
industryAcs Special	24980	9395	2.66	0.0079
industryAgriculture	-4918	6153	-0.80	0.4242
industryConstruction	9035	4530	1.99	0.0462
industryEducation, Health & Social	10702	1500	7.13	0.0000
industryFinance	15878	2008	7.91	0.0000
industryInformation & Communication	17708	3097	5.72	0.0000
industryManufacturing	15107	2571	5.88	0.0000
industryMilitary	25313	20831	1.22	0.2244
industryMining	11778	20822	0.57	0.5716
industryOther services	4891	2400	2.04	0.0416
industryProfessional	13579	1889	7.19	0.0000
industryPublic Admin	19601	2543	7.71	0.0000
industryRetail Trade	4442	1841	2.41	0.0159
industryTransportation & Warehousing	6094	3579	1.70	0.0887
industryUnknown	-2642	2105	-1.25	0.2095
industryUtilities	12500	7068	1.77	0.0770
industryWholesale Trade	10946	4081	2.68	0.0073
age31	-116	1312	-0.09	0.9294
age32	2454	1307	1.88	0.0605
age33	2519	1320	1.91	0.0564
age34	1424	1352	1.05	0.2925
drug_use.2011Refusal	6261	6423	0.97	0.3297
drug_use.2011Unknown	17214	3815	4.51	0.0000
drug_use.2011Yes	4817	2859	1.68	0.0921
gendermale:raceHispanic	5840	1731	3.37	0.0007
gendermale:raceMixed	671	6347	0.11	0.9159
gendermale:raceOther	3881	1439	2.70	0.0070
gendermale:industryAcs Special	-29466	17502	-1.68	0.0923
gendermale:industryAgriculture	22557	8068	2.80	0.0052
gendermale:industryConstruction	1365	4897	0.28	0.7805
gendermale:industryEducation, Health & Social	2129	2401	0.89	0.3754
gendermale:industryFinance	2683	3012	0.89	0.3731
gendermale:industryInformation & Communication	41	4271	0.01	0.9924
gendermale:industryManufacturing	-551	3175	-0.17	0.8622
gendermale:industryMilitary	6843	21414	0.32	0.7493
gendermale:industryMining	23569	21292	1.11	0.2684
gendermale:industryOther services	-1031	3348	-0.31	0.7581
gendermale:industryProfessional	-1819	2573	-0.71	0.4797
gendermale:industryPublic Admin	8547	3469	2.46	0.0138
gendermale:industryRetail Trade	-199	2632	-0.08	0.9397
gendermale:industryTransportation & Warehousing	8571	4282	2.00	0.0454
gendermale:industryUnknown	6989	2905	2.41	0.0162
gendermale:industryUtilities	24542	8568	2.86	0.0042
gendermale:industryWholesale Trade	112	4784	0.02	0.9814
gendermale:age31	501	1817	0.28	0.7827
gendermale:age32	2100	1835	1.14	0.2526
gendermale:age33	2042	1831	1.11	0.2649
gendermale:age34	3311	1898	1.74	0.0811

	Estimate	Std. Error	t value	Pr(>|t|)
gendermale:drug_use.2011Refusal	-15939	8480	-1.88	0.0602
gendermale:drug_use.2011Unknown	-10682	4982	-2.14	0.0321
gendermale:drug_use.2011Yes	-2367	3658	-0.65	0.5175

Implication: This is in line with what we saw in box plots. People who responded **yes** to this question seem to have less income difference (the coefficient is 2100) than those who responded **no**. This does not hold with our expectations. Maybe we are missing some variables that might explain this? Let's keep this variable for a while, regardless of contradiction.

5. Spouse

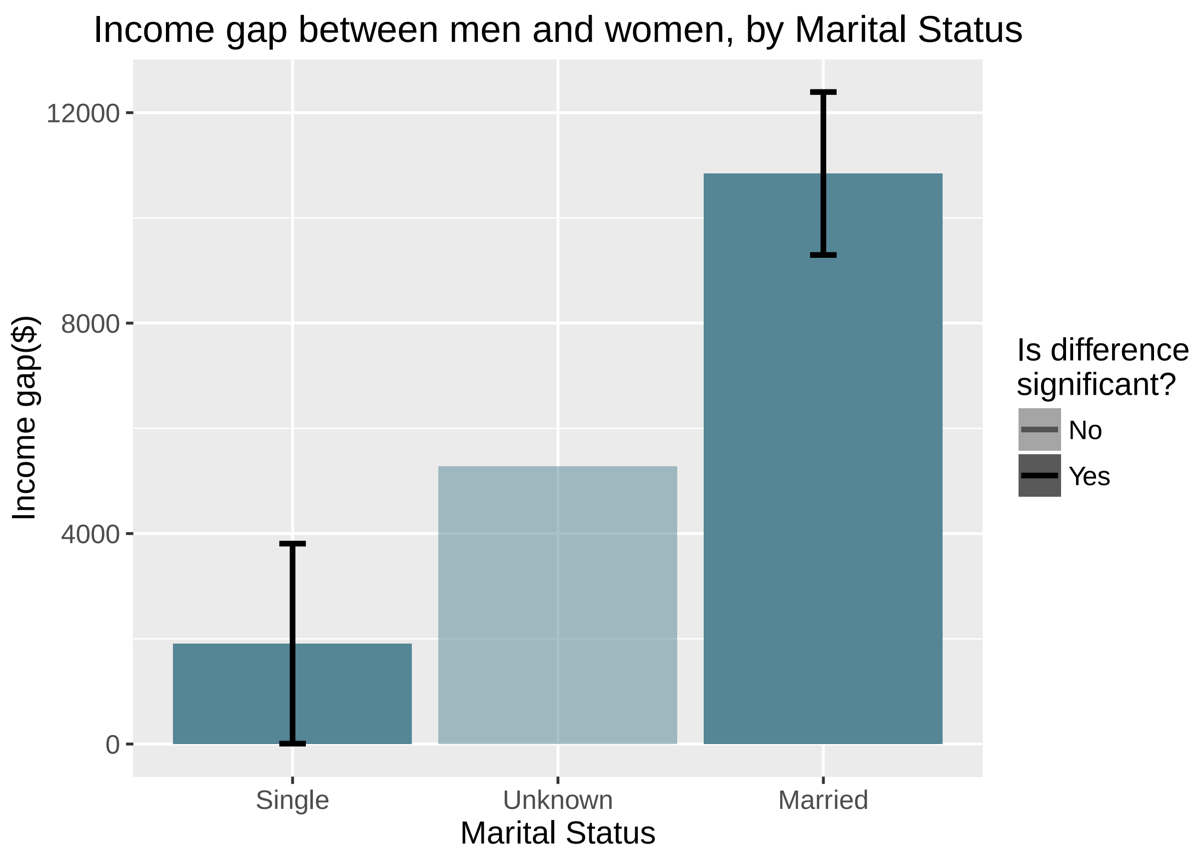
5a. Marital Status

Does married sample of respondents have a larger/smaller income gap than the single sample? In other words, does the income gap depend on marital status?

To find the marital status of a respondent we use the question: **SPOUSE RECEIVE ANY INCOME FROM A JOB IN PAST YEAR?**

Note: The universe for this question is our married sample, and the valid skips are assumed to be single.
Data Summary:

Marital Status	Income Gap ($)	No. of respondents	% of males	% of females	Mean Income ($)	Std. Deviation: I
Married	10843.90	3119	50.75%	49.25%	37665.12	
Single	1908.11	1980	53.43%	46.57%	31530.34	
Unknown	5273.33	18	50.00%	50.00%	22807.78	



That looks like a huge difference in income gap between married and single respondents! Could the reason be that in married couple, the male on average earns more salary? Could be. But there might be other variables at play here, for ex- household factors. Let's directly add the interaction term for this variable.

```
## Analysis of Variance Table
##
## Model 1: income.exclude.topcode ~ race + industry + age + drug_use.2011 +
##     gender:race + gender:industry + gender:age + gender:drug_use.2011
## Model 2: income.exclude.topcode ~ gender + race + industry + age + drug_use.2011 +
##     married + gender:race + gender:industry + gender:age + gender:drug_use.2011 +
##     gender:married
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   5061 2.1814e+12
## 2   5057 2.1437e+12  4 3.7697e+10 22.232 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As expected, marital status is a significant predictor of income gap

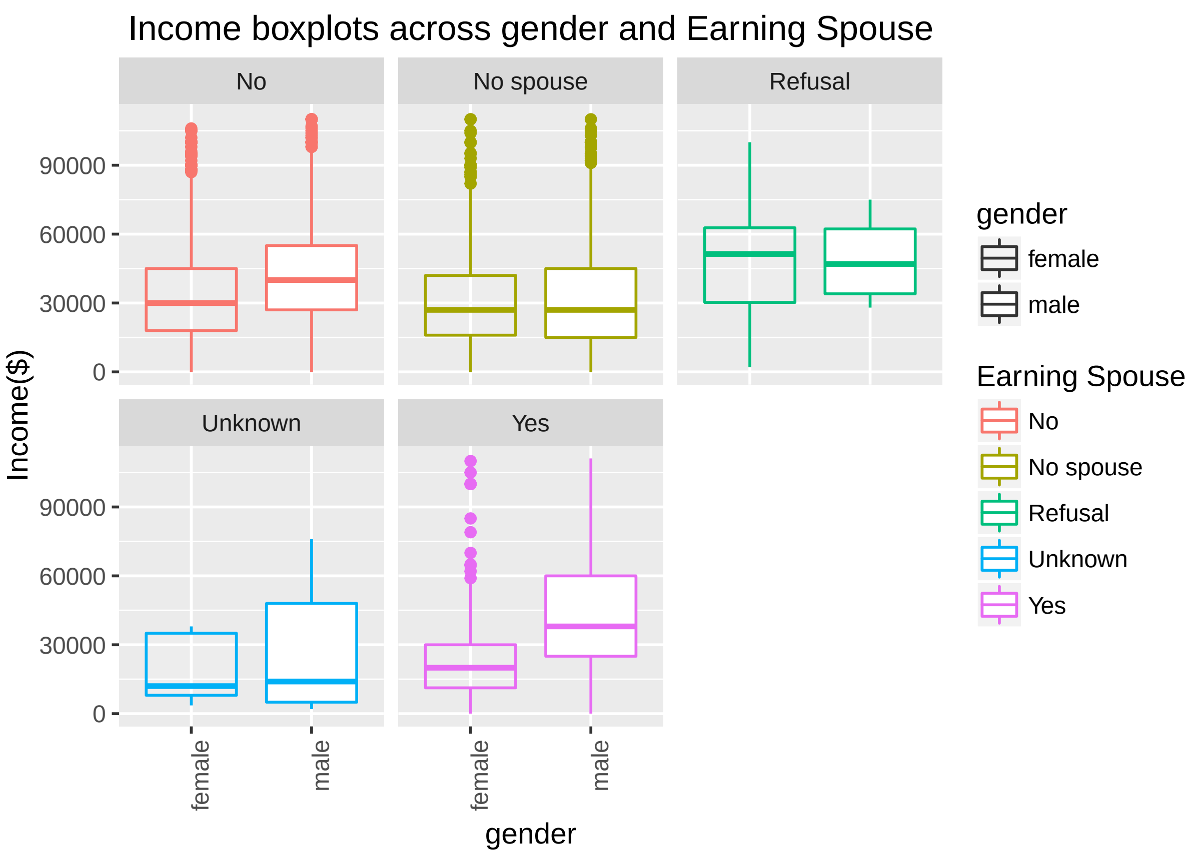
5b. Earning Spouse

Now, let's find if an earning spouse has an effect on income gap. **Data Summary**

Earning Spouse	Income Gap ($)	No. of respondents	% of males	% of females	Mean Income ($)	Std. Deviation: I
No	10302.05	2566	45.25%	54.75%	37272.08	

Earning Spouse	Income Gap ($)	No. of respondents	% of males	% of females	Mean Income ($)	Std. Deviation:
No spouse	1908.11	1980	53.43%	46.57%	31530.34	
Refusal	412.50	12	33.33%	66.67%	48975.00	
Unknown	5273.33	18	50.00%	50.00%	22807.78	
Yes	17258.96	541	77.26%	22.74%	39278.44	

No. of respondents who refused to answer are too low for a hypothesis test to make sense.



Note: Incomes are plotted by ignoring topcoded values for this variable for the purpose of presentation. The box plots are easier to interpret without outliers.

> **Observation:** From bar plots, you can see that not having a spouse, having a non-earning spouse, and having an earning spouse - all three categories have statistically significant income gap. However, this income gap seems to be a lot higher for the sample of respondents who had an earning spouse. We will further investigate the validity of this hypothesis in our model.

Notice, we have a level called `no spouse`. These are the same respondents who were **single** in previous variable `married`. We are essentially fitting the model on the same information for single respondents in both the variables. > Does the presence of two same levels increase the **collinearlity** between the two factor variables? We will answer this question in a while.

Following the protocol, let's test the significance of interaction term.

```
## Analysis of Variance Table
##
## Model 1: income.exclude.topcode ~ race + industry + age + drug_use.2011 +
##     married + gender:race + gender:industry + gender:age + gender:drug_use.2011 +
##     gender:married
## Model 2: income.exclude.topcode ~ gender + race + industry + age + drug_use.2011 +
##     married + spouse.earned.2013 + gender:race + gender:industry +
##     gender:age + gender:drug_use.2011 + gender:married + gender:spouse.earned.2013
##   Res.Df        RSS Df  Sum of Sq      F  Pr(>F)
## 1   5057 2.1437e+12
## 2   5053 2.1381e+12  4 5594152006 3.3051 0.01031 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significant! So not just marital status, but an earning spouse also does seem to be associated with difference in income between men and women.

Let's look at the coefficients only for this variable.

	Estimate	Std. Error	t value	Pr(>|t|)
gendermale:spouse.earned.2013Refusal	-17881	12683	-1.41	0.1586
gendermale:spouse.earned.2013Yes	5089	2284	2.23	0.0259

Notice anything strange? There are no coefficients for the `No spouse`. (The reference level is `No`). Why is that? Because it is already modeled under the variable `married`. This variable does give some more information like the difference in income gap for those have an earning spouse relative to those who don't. So let's keep this variable in our model.

5c. Spouse's Income

We have seen the effects of marital status and if married, if spouse earned or not. It would be interesting to see the impact of spouse's salary on this gap.

But, that would mean looking at only the sample of married respondents. Do we want to ignore all the single respondents from our model? Probably not. So, for that reason, I am not going to consider this variable for our model. However, let's do a quick scatter plot to see how income of men and women varies with spouse's income. I am going to remove topcoded values in both incomes for better visualization.

The average income of males increases at a higher rate than that for females. Does that mean the gap is increasing with spouse's income? Would be an interesting question to explore outside the scope of this project.

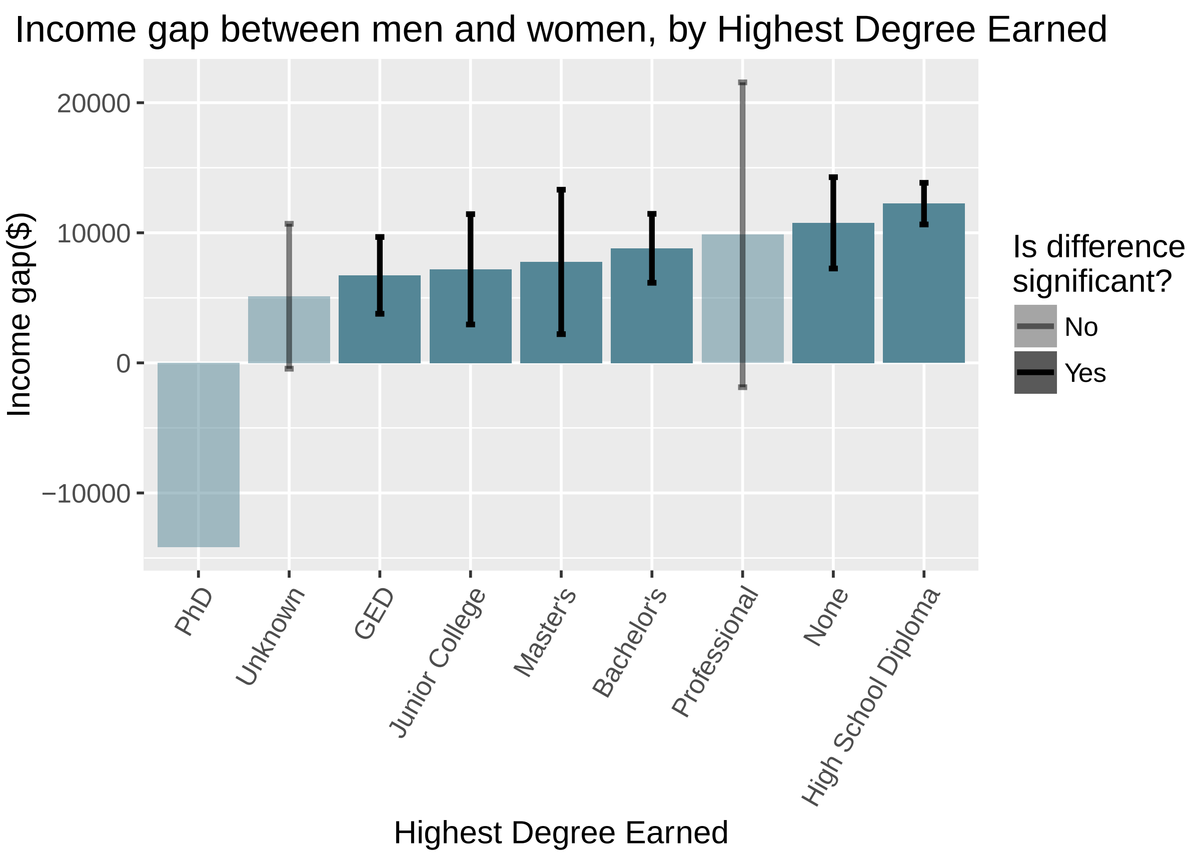
6. Education

This one is obvious. Education can arguably impact income to a very large degree. For education, we are going to consider the answer to the question: **HIGHEST DEGREE RECEIVED PRIOR TO THE 11/12 ACAD YEAR**.

A quick summary:

Highest Degree Earned	Income Gap ($)	No. of respondents	% of males	% of females	Mean Income ($)	Std. Dev
Bachelor's	8807.25	1152	44.18%	55.82%	45124.32	
GED	6727.16	515	60.39%	39.61%	24853.15	
High School Diploma	12242.17	2154	54.92%	45.08%	31385.62	
Junior College	7192.86	377	48.01%	51.99%	36583.49	
Master's	7760.81	276	34.78%	65.22%	52054.39	
None	10766.64	338	63.61%	36.39%	22196.27	
PhD	-14166.67	9	33.33%	66.67%	65111.11	
Professional	9848.68	44	45.45%	54.55%	63330.84	
Unknown	5120.35	252	52.38%	47.62%	35305.54	

Notice the disproportion. PhD and professional degree holders are a lot less than people with other degrees.



Except Professional and PhD degrees, for all other degrees, the income gap is statistically significant. Samples of respondents with a high school diploma and those with no degree seem to have high income gaps. Does the highest degree earned affect income gap?

Let's add this variable (interaction term) to our model and check it's significance:

```
## lm(formula = income.exclude.topcode ~ gender + race + industry +
##     age + drug_use.2011 + married + spouse.earned.2013 + highest.degree.2011 +
##     gender:race + gender:industry + gender:age + gender:drug_use.2011 +
##     gender:married + gender:spouse.earned.2013 + gender:highest.degree.2011,
##     data = nlsy.subset)
```

```
## Analysis of Variance Table
##
## Model 1: income.exclude.topcode ~ race + industry + age + drug_use.2011 +
##     married + spouse.earned.2013 + gender:race + gender:industry +
##     gender:age + gender:drug_use.2011 + gender:married + gender:spouse.earned.2013
## Model 2: income.exclude.topcode ~ gender + race + industry + age + drug_use.2011 +
##     married + spouse.earned.2013 + highest.degree.2011 + gender:race +
##     gender:industry + gender:age + gender:drug_use.2011 + gender:married +
##     gender:spouse.earned.2013 + gender:highest.degree.2011
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   5053 2.1381e+12
## 2   5037 1.8198e+12 16 3.1835e+11 55.073 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As expected, this is a significant interaction. Let's examine only the coefficients for this interaction term:

	Estimate	Std. Error	t value	Pr(>|t|)
gendermale:highest.degree.2011Bachelor's	1123	2548	0.44	0.6593
gendermale:highest.degree.2011GED	-1258	2777	-0.45	0.6505
gendermale:highest.degree.2011High School Diploma	3525	2342	1.50	0.1324
gendermale:highest.degree.2011Junior College	-1131	2974	-0.38	0.7038
gendermale:highest.degree.2011Master's	-732	3378	-0.22	0.8284
gendermale:highest.degree.2011PhD	-21091	13676	-1.54	0.1231
gendermale:highest.degree.2011Professional	4704	6258	0.75	0.4523
gendermale:highest.degree.2011Unknown	1407	5535	0.25	0.7994

Observation: Compared to respondents with no degrees, income gap between men and women seems to be higher for respondents with Bachelor's, Master's and High School Diploma.

7. Total Incarnations

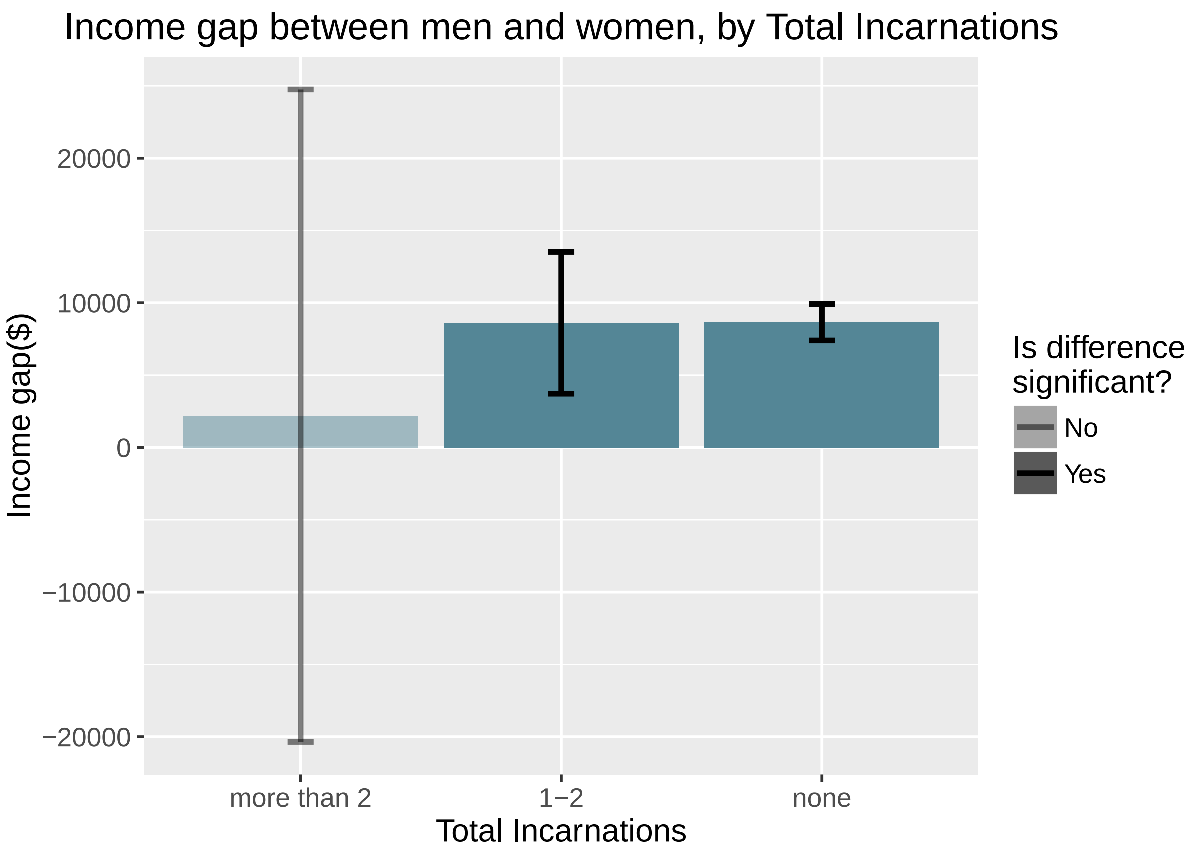
As pointed out before, we have converted this variable to a factor variable. This probably will give us an estimate of how crime history is related to income gap. Do people who have had more incarnations in past have a greater difference in income between men and women?

Let's find out:

Total Incarnations	Income Gap ($)	No. of respondents	% of males	% of females	Mean Income ($)	Std. Deviatio
1-2	8618.57	288	82.64%	17.36%	25207.25	

Total Incarnations	Income Gap ($)	No. of respondents	% of males	% of females	Mean Income ($)	Std. Deviatio
more than 2	2197.49	75	92.00%	8.00%	20541.69	
none	8658.47	4754	49.28%	50.72%	36078.63	

Most of the population has no incarnations, which makes sense. So let's take this level as a reference level.



Income gap is not significant for people with more than 2 incarnations. Why would that be the case? May be because there are only 75 respondents in this bin to begin with. On top of that only 8% of these are female. So, this looks more of a case of insufficient data to form the conclusion.

Let's add this interaction to our regression model and test the significance:

```
## Analysis of Variance Table
##
## Model 1: income.exclude.topcode ~ race + industry + age + drug_use.2011 +
##     married + spouse.earned.2013 + highest.degree.2011 + gender:race +
##     gender:industry + gender:age + gender:drug_use.2011 + gender:married +
##     gender:spouse.earned.2013 + gender:highest.degree.2011
## Model 2: income.exclude.topcode ~ gender + race + industry + age + drug_use.2011 +
##     married + spouse.earned.2013 + highest.degree.2011 + total.incarnations +
##     gender:race + gender:industry + gender:age + gender:drug_use.2011 +
##     gender:married + gender:spouse.earned.2013 + gender:highest.degree.2011 +
##     gender:total.incarnations
##   Res.Df        RSS Df  Sum of Sq     F    Pr(>F)
## 1   5037 1.8198e+12
## 2   5033 1.8006e+12  4 1.9219e+10 13.43 6.868e-11 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is significant! So, there is some additional value in adding this variable. Let's look at the coefficients for only this interaction:

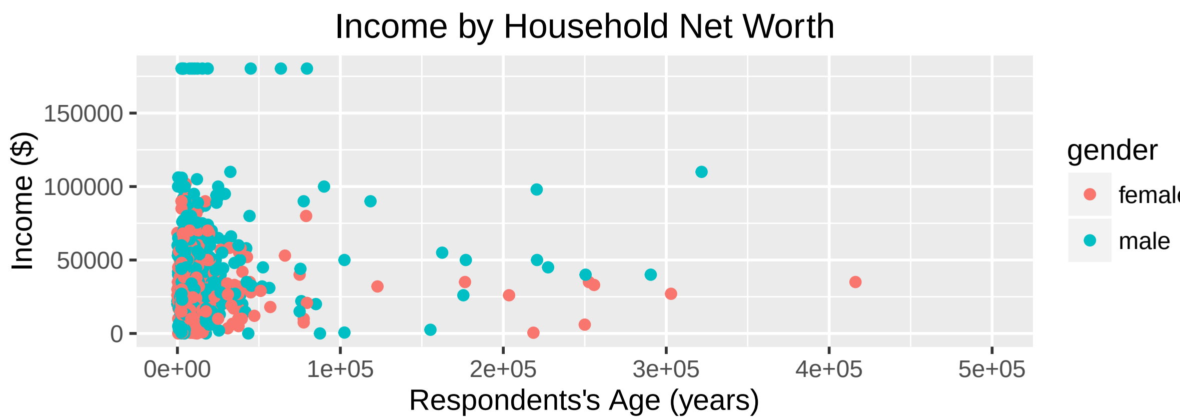
	Estimate	Std. Error	t value	Pr(>|t|)
gendermale:total.incarnations1-2	-2236	3069	-0.73	0.4663
gendermale:total.incarnationsmore than 2	1032	8195	0.13	0.8998

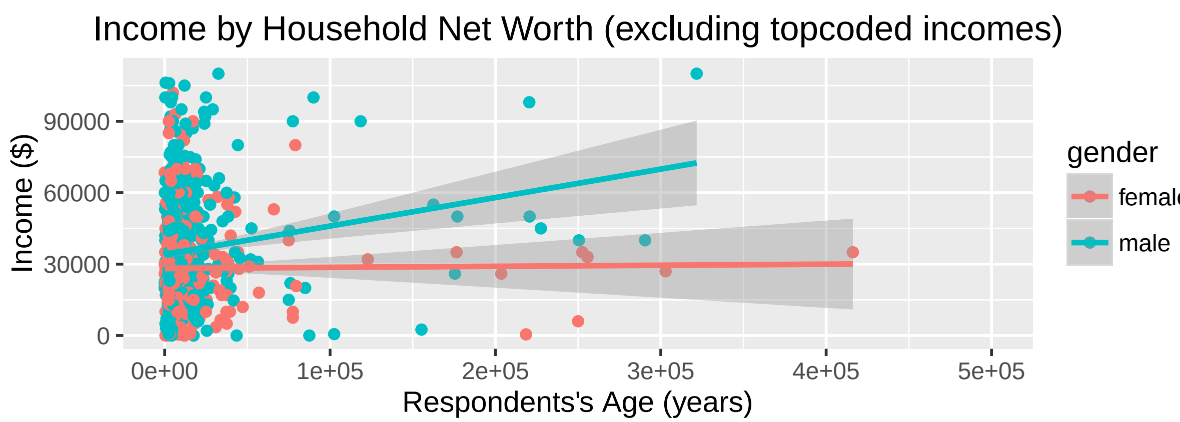
This is not in line with our observation from bar plots.

Interpretation: People with more than 2 incarnations question seem to have a difference of income between males and female $1032 more than those with no incarnations.

8. Household net Worth

This is a continuous variable. We are interested in this variable because there might be some relation between household factors and income earned by a respondent. Household net worth is a good proxy for economic condition of the household.





I can definitely see an increasing difference between average male and female incomes. So, let's add this variable to our model and test its significance:

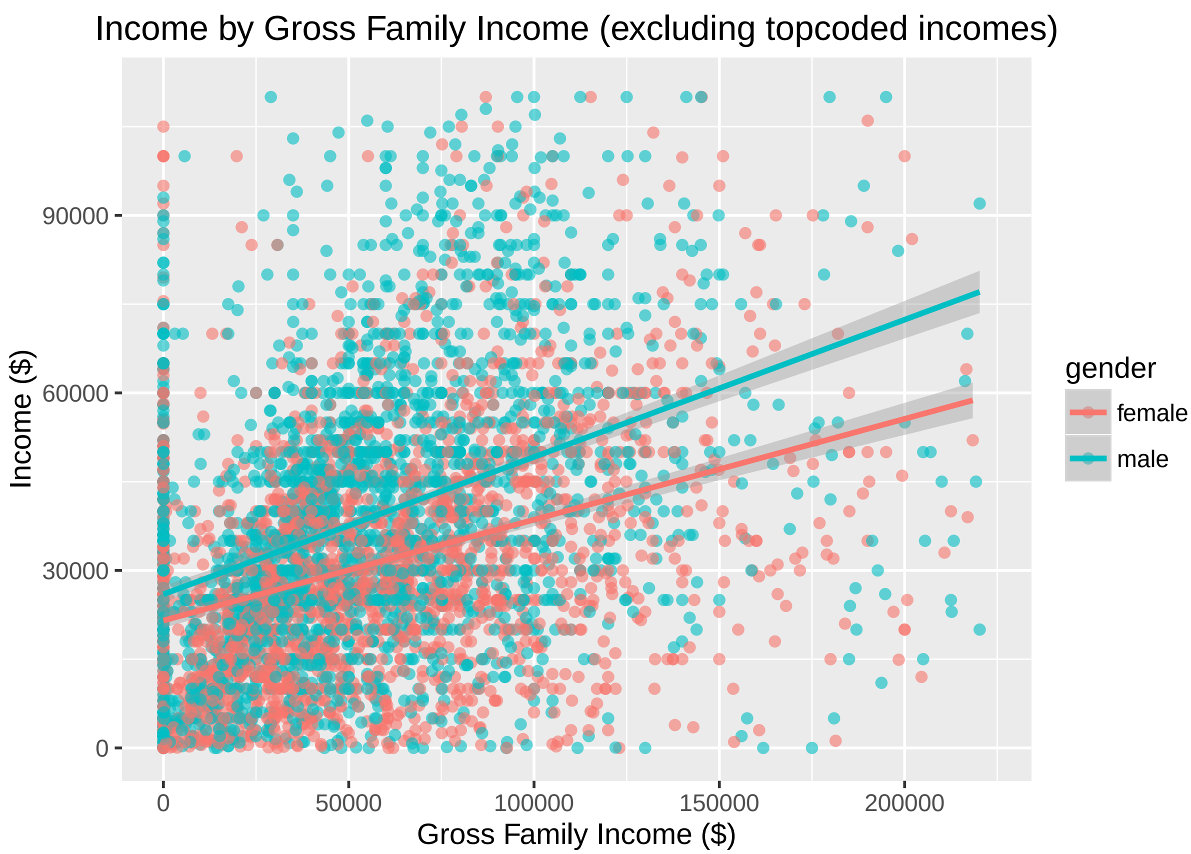
```
## Analysis of Variance Table
##
## Model 1: income.exclude.topcode ~ race + industry + age + drug_use.2011 +
```

```
##     married + spouse.earned.2013 + highest.degree.2011 + total.incarnations +
##     gender:race + gender:industry + gender:age + gender:drug_use.2011 +
##     gender:married + gender:spouse.earned.2013 + gender:highest.degree.2011 +
##     gender:total.incarnations
## Model 2: income.exclude.topcode ~ gender + race + industry + age + drug_use.2011 +
##     married + spouse.earned.2013 + highest.degree.2011 + total.incarnations +
##     hh.net.worth.2003 + gender:race + gender:industry + gender:age +
##     gender:drug_use.2011 + gender:married + gender:spouse.earned.2013 +
##     gender:highest.degree.2011 + gender:total.incarnations +
##     gender:hh.net.worth.2003
##   Res.Df        RSS Df  Sum of Sq     F Pr(>F)
## 1   1056 3.1635e+11
## 2   1054 3.1364e+11  2 2712072578 4.557 0.0107 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is < 0.05. So the interaction seems to be a significant predictor of income. Let's look at one more continuous variable:

9. Gross Family Income

Gross family income, again is a proxy for economic condition of household's economic situation. I am interested to see if this has a different effect from household net worth, or if there is some collinearity. For this variable, I am going to use just the `gross.family.income` term (without interaction) because I feel there might be high correlation between respondent's and family's income.



Again, you can see an increasing trend for both average male and female incomes. So, let's add this variable to our model and test its significance:

```
## Analysis of Variance Table
##
## Model 1: income.exclude.topcode ~ gender + race + industry + age + drug_use.2011 +
##     married + spouse.earned.2013 + highest.degree.2011 + total.incarnations +
##     hh.net.worth.2003 + gender:race + gender:industry + gender:age +
##     gender:drug_use.2011 + gender:married + gender:spouse.earned.2013 +
##     gender:highest.degree.2011 + gender:total.incarnations +
##     gender:hh.net.worth.2003
## Model 2: income.exclude.topcode ~ gender + race + industry + age + drug_use.2011 +
##     married + spouse.earned.2013 + highest.degree.2011 + total.incarnations +
##     hh.net.worth.2003 + gross.family.income + gender:race + gender:industry +
##     gender:age + gender:drug_use.2011 + gender:married + gender:spouse.earned.2013 +
##     gender:highest.degree.2011 + gender:total.incarnations +
##     gender:hh.net.worth.2003
##   Res.Df        RSS Df  Sum of Sq     F    Pr(>F)
## 1    981 2.8653e+11
## 2    980 2.7302e+11  1 1.3509e+10 48.49 6.076e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significant! So this term does add some useful information to the model. Let's look at the coefficient of this term:

	Estimate	Std. Error	t value	Pr(>|t|)
gross.family.income	0.1	0	6.96	0

> **Interpretation:** This term is highly significant! So, every 1$ increase in gross family income is associated with $`round(coef(nlsy.lm)['gross.family.income'],3)` increase in respondent's income.

Building the model

Ok, we have a regression model in place, that we built as an additive process- adding one variable at a time. Below is the coefficient table so far:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2205.80	5993	0.37	0.7129
gendermale	-776.64	8067	-0.10	0.9233
raceHispanic	2148.75	2422	0.89	0.3753
raceMixed	25398.47	7780	3.26	0.0011
raceOther	1529.42	2051	0.75	0.4561
industryAcs Special	-29992.60	16905	-1.77	0.0763
industryAgriculture	-1596.25	8869	-0.18	0.8572
industryConstruction	11758.87	6734	1.75	0.0811
industryEducation, Health & Social	1964.95	2575	0.76	0.4455
industryFinance	8268.65	3837	2.15	0.0314

	Estimate	Std. Error	t value	Pr(>|t|)
industryInformation & Communication	14211.20	6388	2.22	0.0263
industryManufacturing	17370.53	5749	3.02	0.0026
industryMilitary	34689.41	7465	4.65	0.0000
industryMining	25133.10	7843	3.20	0.0014
industryOther services	2290.62	4215	0.54	0.5870
industryProfessional	3409.13	3219	1.06	0.2898
industryPublic Admin	7613.90	4663	1.63	0.1028
industryRetail Trade	5036.12	3127	1.61	0.1077
industryTransportation & Warehousing	3761.93	6766	0.56	0.5783
industryUnknown	-903.07	3455	-0.26	0.7938
industryUtilities	29976.30	12093	2.48	0.0133
industryWholesale Trade	104.22	8663	0.01	0.9904
age31	1019.45	1800	0.57	0.5712
age32	5101.40	4790	1.07	0.2871
age33	-1119.64	8591	-0.13	0.8963
age34	4908.82	10132	0.48	0.6282
drug_use.2011Refusal	6489.69	10834	0.60	0.5493
drug_use.2011Unknown	-3355.02	9275	-0.36	0.7176
drug_use.2011Yes	5354.40	4735	1.13	0.2584
marriedMarried	-1947.43	1718	-1.13	0.2573
marriedUnknown	-4770.95	9852	-0.48	0.6283
spouse.earned.2013Refusal	25468.45	17995	1.42	0.1573
spouse.earned.2013Yes	-1386.75	3435	-0.40	0.6865
highest.degree.2011Bachelor's	20345.05	3705	5.49	0.0000
highest.degree.2011GED	1919.06	4074	0.47	0.6377
highest.degree.2011High School Diploma	5913.53	3399	1.74	0.0822
highest.degree.2011Junior College	9958.13	4241	2.35	0.0191
highest.degree.2011Master's	28990.76	5010	5.79	0.0000
highest.degree.2011PhD	33033.28	17127	1.93	0.0541
highest.degree.2011Professional	43534.73	7218	6.03	0.0000
highest.degree.2011Unknown	4590.40	7728	0.59	0.5526
total.incarnations1-2	1969.41	6034	0.33	0.7442
total.incarnationsmore than 2	-5882.75	12219	-0.48	0.6303
hh.net.worth.2003	0.02	0	0.69	0.4914
gross.family.income	0.10	0	6.96	0.0000
gendermale:raceHispanic	3674.68	3275	1.12	0.2621
gendermale:raceMixed	-19432.66	14483	-1.34	0.1800
gendermale:raceOther	6285.91	2793	2.25	0.0246
gendermale:industryAgriculture	24977.94	12510	2.00	0.0461
gendermale:industryConstruction	-846.48	7387	-0.11	0.9088
gendermale:industryEducation, Health & Social	2039.96	4137	0.49	0.6220
gendermale:industryFinance	4099.02	5498	0.75	0.4561
gendermale:industryInformation & Communication	-4243.73	8324	-0.51	0.6103
gendermale:industryManufacturing	-4205.29	6507	-0.65	0.5182
gendermale:industryOther services	5896.64	5773	1.02	0.3073
gendermale:industryProfessional	5641.08	4378	1.29	0.1979
gendermale:industryPublic Admin	11888.94	6456	1.84	0.0659
gendermale:industryRetail Trade	-4167.21	4333	-0.96	0.3364
gendermale:industryTransportation & Warehousing	2387.68	7977	0.30	0.7648
gendermale:industryUnknown	7293.91	4675	1.56	0.1191
gendermale:industryWholesale Trade	9556.01	9654	0.99	0.3225
gendermale:age31	-2389.81	2381	-1.00	0.3157

	Estimate	Std. Error	t value	Pr(>|t|)
gendermale:age32	-5919.79	6278	-0.94	0.3459
gendermale:age33	-9987.51	10828	-0.92	0.3566
gendermale:drug_use.2011Refusal	-17255.09	20769	-0.83	0.4063
gendermale:drug_use.2011Yes	1659.36	6049	0.27	0.7839
gendermale:marriedMarried	5364.12	2322	2.31	0.0211
gendermale:marriedUnknown	-4755.21	14120	-0.34	0.7364
gendermale:spouse.earned.2013Yes	7830.08	4132	1.90	0.0584
gendermale:highest.degree.2011Bachelor's	-7247.19	4858	-1.49	0.1361
gendermale:highest.degree.2011GED	-1298.34	5178	-0.25	0.8021
gendermale:highest.degree.2011High School Diploma	-964.23	4327	-0.22	0.8237
gendermale:highest.degree.2011Junior College	99.43	5550	0.02	0.9857
gendermale:highest.degree.2011Master's	1914.52	7236	0.26	0.7914
gendermale:highest.degree.2011Professional	-6180.69	10545	-0.59	0.5579
gendermale:highest.degree.2011Unknown	-2315.08	12680	-0.18	0.8552
gendermale:total.incarnations1-2	-9885.58	6561	-1.51	0.1322
gendermale:total.incarnationsmore than 2	1627.14	13474	0.12	0.9039
gendermale:hh.net.worth.2003	0.04	0	1.04	0.3008

But, what if the significance of some variables might have changed in the process of adding more variables. There can be several reasons for this: multicollinearity is one. Let's look at some variables we thought are not so important for income gap.

Age was, if you recall, between 30-34 for all respondents. Can this small a range really impact the income difference?

Let's try removing age and testing for significance:

```
## Analysis of Variance Table
##
## Model 1: income.exclude.topcode ~ race + industry + drug_use.2011 + married +
##     spouse.earned.2013 + highest.degree.2011 + total.incarnations +
##     hh.net.worth.2003 + gross.family.income + gender:race + gender:industry +
##     gender:drug_use.2011 + gender:married + gender:spouse.earned.2013 +
##     gender:highest.degree.2011 + gender:total.incarnations +
##     gender:hh.net.worth.2003
## Model 2: income.exclude.topcode ~ gender + race + industry + age + drug_use.2011 +
##     married + spouse.earned.2013 + highest.degree.2011 + total.incarnations +
##     hh.net.worth.2003 + gross.family.income + gender:race + gender:industry +
##     gender:age + gender:drug_use.2011 + gender:married + gender:spouse.earned.2013 +
##     gender:highest.degree.2011 + gender:total.incarnations +
##     gender:hh.net.worth.2003
##   Res.Df        RSS Df  Sum of Sq      F Pr(>F)
## 1    987 2.7431e+11
## 2    980 2.7302e+11  7 1291762508 0.6624 0.7041
```

Not a significant difference! This sits in line with our expectations. So, let's remove this variable and update our model.

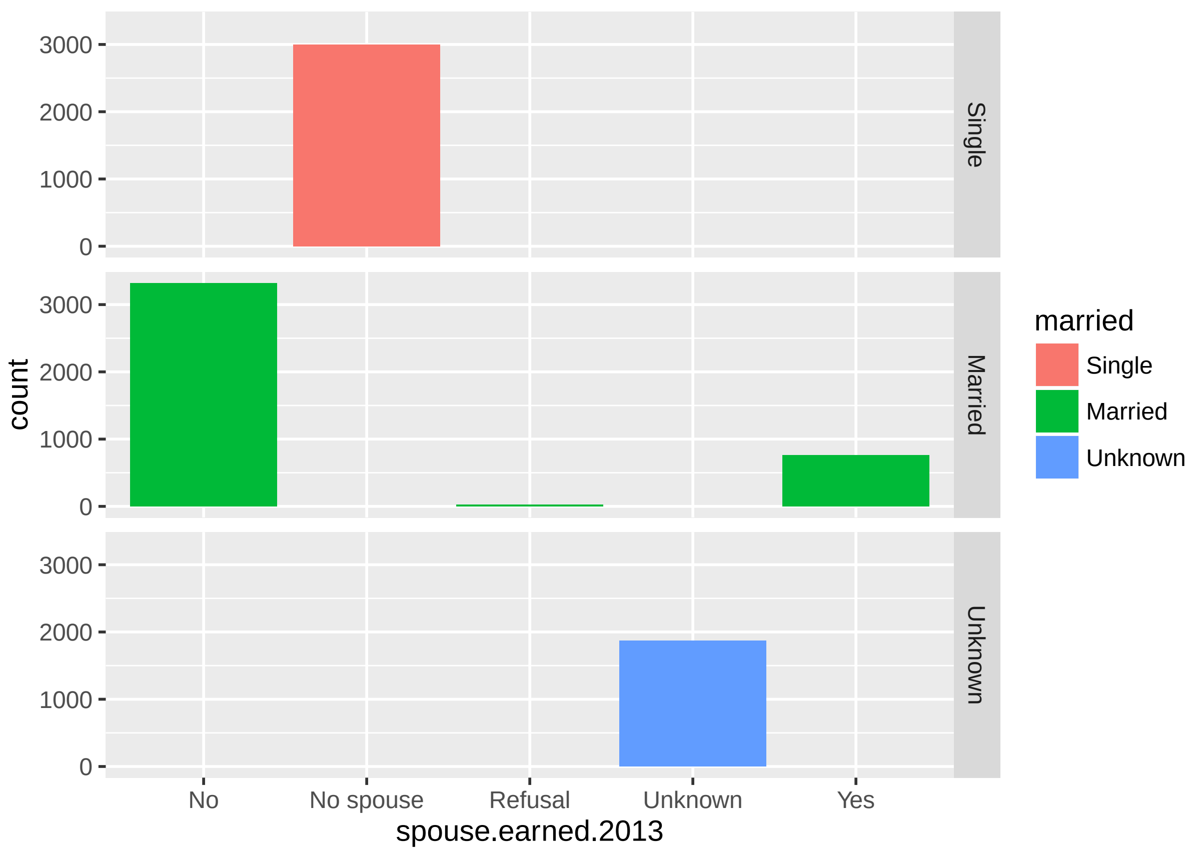
We also discussed in previous section how one level of `married` and `spouse.earned.2013` had same effects. Now, since `spouse.earned.2013` gives more information than just marital status, we probably want to try removing `married`

```
## Analysis of Variance Table
```

```
##
## Model 1: income.exclude.topcode ~ race + industry + drug_use.2011 + spouse.earned.2013 +
##     highest.degree.2011 + total.incarnations + hh.net.worth.2003 +
##     gross.family.income + gender:race + gender:industry + gender:drug_use.2011 +
##     gender:spouse.earned.2013 + gender:highest.degree.2011 +
##     gender:total.incarnations + gender:hh.net.worth.2003
## Model 2: income.exclude.topcode ~ gender + race + industry + drug_use.2011 +
##     married + spouse.earned.2013 + highest.degree.2011 + total.incarnations +
##     hh.net.worth.2003 + gross.family.income + gender:race + gender:industry +
##     gender:drug_use.2011 + gender:married + gender:spouse.earned.2013 +
##     gender:highest.degree.2011 + gender:total.incarnations +
##     gender:hh.net.worth.2003
##   Res.Df        RSS Df  Sum of Sq F Pr(>F)
## 1    987 2.7431e+11
## 2    987 2.7431e+11  0 9.1553e-05
```

There is no impact! The two models are essentially the same (Look at Df = 0). Let's remove marital status and update our model.

The above relationship between the two variables is clearer using the following plot:



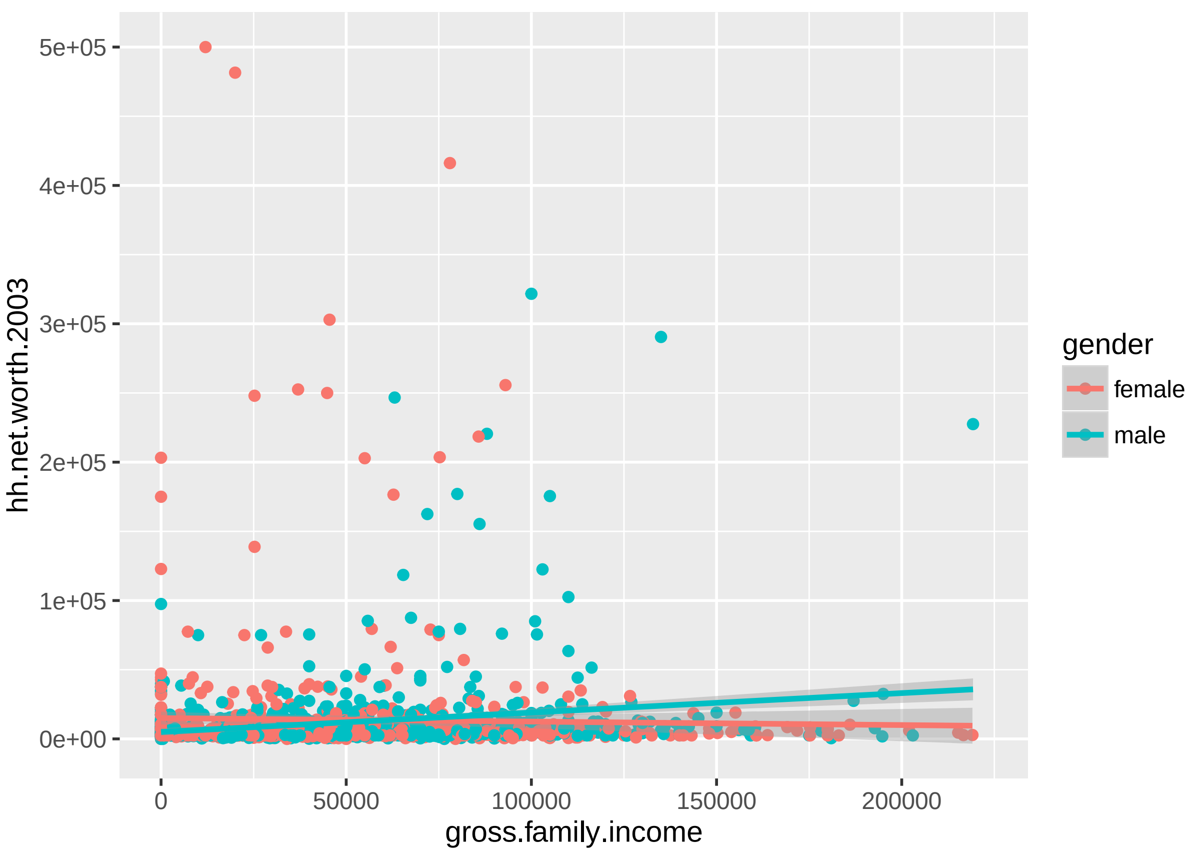
Notice the 100% overlap between `married="Single"` and `spouse.earned.2013="No spouse"`.

In the previous section, we also talked about a potentially contradicting result. Non-drug users had a larger average income gap. Let's try removing the variable `drug_use.2011`

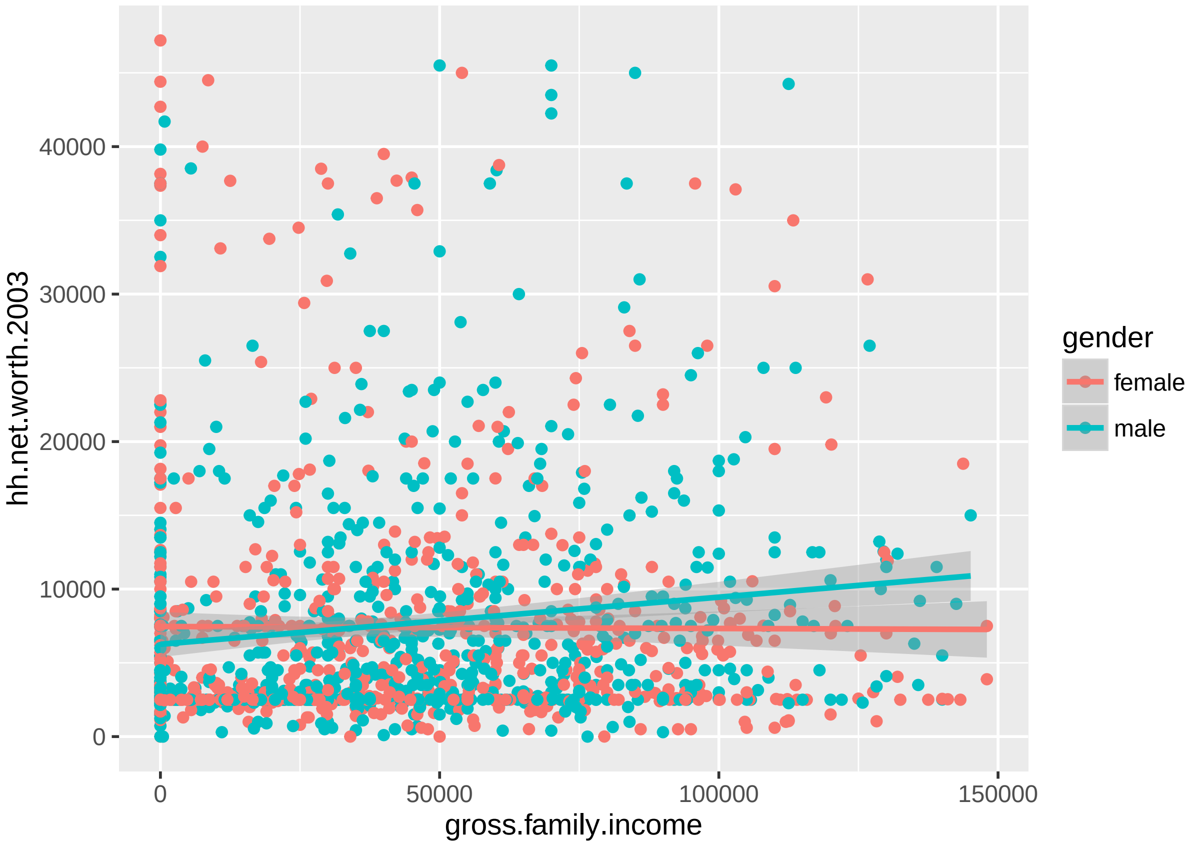
```
## Analysis of Variance Table
##
## Model 1: income.exclude.topcode ~ race + industry + spouse.earned.2013 +
##     highest.degree.2011 + total.incarnations + hh.net.worth.2003 +
##     gross.family.income + gender:race + gender:industry + gender:spouse.earned.2013 +
##     gender:highest.degree.2011 + gender:total.incarnations +
##     gender:hh.net.worth.2003
## Model 2: income.exclude.topcode ~ gender + race + industry + drug_use.2011 +
##     spouse.earned.2013 + highest.degree.2011 + total.incarnations +
##     hh.net.worth.2003 + gross.family.income + gender:race + gender:industry +
##     gender:drug_use.2011 + gender:spouse.earned.2013 + gender:highest.degree.2011 +
##     gender:total.incarnations + gender:hh.net.worth.2003
##   Res.Df        RSS Df  Sum of Sq      F Pr(>F)
## 1    992 2.7636e+11
## 2    987 2.7431e+11  5 2048381443 1.4741 0.1956
```

Not a significant difference! So, we were probably right. The sample of non-drug users is probably large enough to have estimated the effects of some other variables. Let's remove this from our model as well.

We also talked about potential collinearity between **household net worth** and **gross family income**. Let's try removing household net worth.



There seems to be some increasing trend for men, but you cannot really tell due to outliers. Let's focus on the bottom left portion:



Still can't tell? Let's just try removing household net worth from the model and see what happens.

```
## Analysis of Variance Table
##
## Model 1: income.exclude.topcode ~ race + industry + spouse.earned.2013 +
##     highest.degree.2011 + total.incarnations + gross.family.income +
##     gender:race + gender:industry + gender:spouse.earned.2013 +
##     gender:highest.degree.2011 + gender:total.incarnations
## Model 2: income.exclude.topcode ~ gender + race + industry + spouse.earned.2013 +
##     highest.degree.2011 + total.incarnations + hh.net.worth.2003 +
##     gross.family.income + gender:race + gender:industry + gender:spouse.earned.2013 +
##     gender:highest.degree.2011 + gender:total.incarnations +
##     gender:hh.net.worth.2003
##   Res.Df        RSS Df  Sum of Sq      F Pr(>F)
## 1    994 2.7748e+11
## 2    992 2.7636e+11  2 1117306769 2.0053 0.1352
```

Not significant! So, maybe because of collinearity or maybe because of another reason, our `hh.net.worth.2003` interaction has rendered not a significant predictor. Let's remove this too from the model.

So, our final model and coefficients looks like this:

```
##
## Call:
## lm(formula = income.exclude.topcode ~ gender + race + industry +
##     spouse.earned.2013 + highest.degree.2011 + total.incarnations +
```

```
##     gross.family.income + gender:race + gender:industry + gender:spouse.earned.2013 +
##     gender:highest.degree.2011 + gender:total.incarnations, data = nlsy.subset)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67002 -11603  -1467   9521  91903
##
## Coefficients:
##                                                   Estimate
## (Intercept)                                    6897.591919
## gendermale                                     6825.090158
## raceHispanic                                   1697.943115
## raceMixed                                      9989.343286
## raceOther                                      1241.656542
## industryAcs Special                           15858.313369
## industryAgriculture                           -3441.580681
## industryConstruction                           8178.954654
## industryEducation, Health & Social             3885.276658
## industryFinance                                9747.629298
## industryInformation & Communication           10102.045048
## industryManufacturing                         10635.805410
## industryMilitary                              25182.700801
## industryMining                                 9247.618336
## industryOther services                         1311.655127
## industryProfessional                           7196.072115
## industryPublic Admin                           8872.063595
## industryRetail Trade                           3600.031979
## industryTransportation & Warehousing           2639.722190
## industryUnknown                               -3206.017113
## industryUtilities                              4555.072534
## industryWholesale Trade                       10515.309920
## spouse.earned.2013No spouse                    1999.935318
## spouse.earned.2013Refusal                     20818.546729
## spouse.earned.2013Unknown                     -4192.108804
## spouse.earned.2013Yes                          2561.348396
## highest.degree.2011Bachelor's                 18613.516903
## highest.degree.2011GED                         2823.200492
## highest.degree.2011High School Diploma         5157.887661
## highest.degree.2011Junior College             12590.464562
## highest.degree.2011Master's                   24910.994377
## highest.degree.2011PhD                        43415.793970
## highest.degree.2011Professional               35186.893969
## highest.degree.2011Unknown                    10592.922175
## total.incarnations1-2                         -3677.921611
## total.incarnationsmore than 2                 -8477.852513
## gross.family.income                               0.127900
## gendermale:raceHispanic                        3897.464537
## gendermale:raceMixed                          -7539.139189
## gendermale:raceOther                           3238.158242
## gendermale:industryAcs Special               -19705.057723
## gendermale:industryAgriculture                18856.478590
## gendermale:industryConstruction                2224.529977
## gendermale:industryEducation, Health & Social -1584.539884
## gendermale:industryFinance                      224.823821
```

```
## gendermale:industryInformation & Communication     -1771.529262
## gendermale:industryManufacturing                     625.381105
## gendermale:industryMilitary                         -866.651829
## gendermale:industryMining                          20723.265828
## gendermale:industryOther services                    890.563568
## gendermale:industryProfessional                      -44.730936
## gendermale:industryPublic Admin                     9576.931389
## gendermale:industryRetail Trade                     -812.073469
## gendermale:industryTransportation & Warehousing     9060.640210
## gendermale:industryUnknown                          5947.773636
## gendermale:industryUtilities                       22022.229359
## gendermale:industryWholesale Trade                 -1398.158774
## gendermale:spouse.earned.2013No spouse             -8225.401812
## gendermale:spouse.earned.2013Refusal              -10018.212581
## gendermale:spouse.earned.2013Unknown               -2976.155166
## gendermale:spouse.earned.2013Yes                    1508.395607
## gendermale:highest.degree.2011Bachelor's             123.443002
## gendermale:highest.degree.2011GED                   -691.885595
## gendermale:highest.degree.2011High School Diploma   2295.804913
## gendermale:highest.degree.2011Junior College       -2711.130992
## gendermale:highest.degree.2011Master's              -395.891691
## gendermale:highest.degree.2011PhD                 -12113.314582
## gendermale:highest.degree.2011Professional          4361.090181
## gendermale:highest.degree.2011Unknown              -6282.946323
## gendermale:total.incarnations1-2                   -2049.329316
## gendermale:total.incarnationsmore than 2             922.190115
##                                                      Std. Error t value
## (Intercept)                                         2104.484274   3.278
## gendermale                                          2809.861541   2.429
## raceHispanic                                        1116.114941   1.521
## raceMixed                                           4110.874310   2.430
## raceOther                                            966.800466   1.284
## industryAcs Special                                 8142.381090   1.948
## industryAgriculture                                 5560.239302  -0.619
## industryConstruction                                4012.869177   2.038
## industryEducation, Health & Social                  1343.496598   2.892
## industryFinance                                     1799.294319   5.417
## industryInformation & Communication                 2749.004628   3.675
## industryManufacturing                               2278.287179   4.668
## industryMilitary                                   18042.982474   1.396
## industryMining                                     18050.388042   0.512
## industryOther services                              2109.235326   0.622
## industryProfessional                                1677.073643   4.291
## industryPublic Admin                                2268.300748   3.911
## industryRetail Trade                                1608.500911   2.238
## industryTransportation & Warehousing                3146.403492   0.839
## industryUnknown                                     1857.464565  -1.726
## industryUtilities                                   6140.927841   0.742
## industryWholesale Trade                             3541.067753   2.970
## spouse.earned.2013No spouse                          841.273340   2.377
## spouse.earned.2013Refusal                           6405.993430   3.250
## spouse.earned.2013Unknown                           6070.274303  -0.691
## spouse.earned.2013Yes                               1772.529513   1.445
## highest.degree.2011Bachelor's                       1863.922326   9.986
```

```
## highest.degree.2011GED                           2084.108863   1.355
## highest.degree.2011High School Diploma           1754.096985   2.940
## highest.degree.2011Junior College                2134.337705   5.899
## highest.degree.2011Master's                      2228.591588  11.178
## highest.degree.2011PhD                           7576.481979   5.730
## highest.degree.2011Professional                  4221.189444   8.336
## highest.degree.2011Unknown                       4182.865290   2.532
## total.incarnations1-2                            2666.857924  -1.379
## total.incarnationsmore than 2                    8109.303674  -1.045
## gross.family.income                                 0.006731  19.003
## gendermale:raceHispanic                          1563.578608   2.493
## gendermale:raceMixed                             5587.716834  -1.349
## gendermale:raceOther                             1346.018085   2.406
## gendermale:industryAcs Special                  15204.528492  -1.296
## gendermale:industryAgriculture                   7176.140963   2.628
## gendermale:industryConstruction                  4333.304221   0.513
## gendermale:industryEducation, Health & Social    2165.449417  -0.732
## gendermale:industryFinance                       2687.168334   0.084
## gendermale:industryInformation & Communication   3806.856556  -0.465
## gendermale:industryManufacturing                 2800.154663   0.223
## gendermale:industryMilitary                     18552.103566  -0.047
## gendermale:industryMining                       18481.126824   1.121
## gendermale:industryOther services                2927.682508   0.304
## gendermale:industryProfessional                  2273.918442  -0.020
## gendermale:industryPublic Admin                  3078.144801   3.111
## gendermale:industryRetail Trade                  2304.686912  -0.352
## gendermale:industryTransportation & Warehousing  3762.061755   2.408
## gendermale:industryUnknown                       2551.196088   2.331
## gendermale:industryUtilities                     7564.103932   2.911
## gendermale:industryWholesale Trade               4153.209036  -0.337
## gendermale:spouse.earned.2013No spouse           1159.154721  -7.096
## gendermale:spouse.earned.2013Refusal            11079.286817  -0.904
## gendermale:spouse.earned.2013Unknown             8584.000009  -0.347
## gendermale:spouse.earned.2013Yes                 2068.838518   0.729
## gendermale:highest.degree.2011Bachelor's         2444.045363   0.051
## gendermale:highest.degree.2011GED                2633.946924  -0.263
## gendermale:highest.degree.2011High School Diploma   2230.659890   1.029
## gendermale:highest.degree.2011Junior College     2849.626467  -0.951
## gendermale:highest.degree.2011Master's           3252.966964  -0.122
## gendermale:highest.degree.2011PhD               12943.021049  -0.936
## gendermale:highest.degree.2011Professional       6095.927866   0.715
## gendermale:highest.degree.2011Unknown            5949.364917  -1.056
## gendermale:total.incarnations1-2                 2965.800165  -0.691
## gendermale:total.incarnationsmore than 2         8440.539590   0.109
##                                                 Pr(>|t|)
## (Intercept)                                     0.001055 **
## gendermale                                      0.015179 *
## raceHispanic                                    0.128252
## raceMixed                                       0.015137 *
## raceOther                                       0.199102
## industryAcs Special                             0.051519 .
## industryAgriculture                             0.535971
## industryConstruction                            0.041587 *
## industryEducation, Health & Social              0.003846 **
```

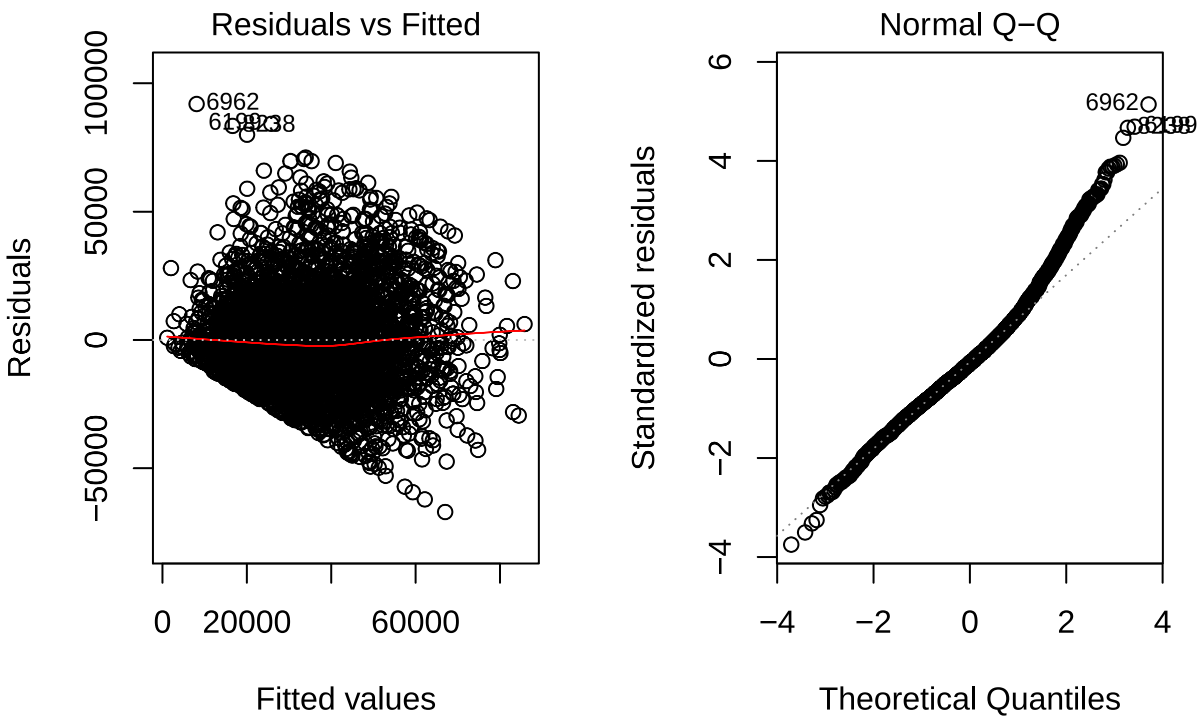
```
## industryFinance                                      6.34e-08 ***
## industryInformation & Communication                  0.000241 ***
## industryManufacturing                                3.12e-06 ***
## industryMilitary                                     0.162869
## industryMining                                       0.608449
## industryOther services                               0.534062
## industryProfessional                                 1.82e-05 ***
## industryPublic Admin                                 9.31e-05 ***
## industryRetail Trade                                 0.025259 *
## industryTransportation & Warehousing                 0.401531
## industryUnknown                                      0.084410 .
## industryUtilities                                    0.458272
## industryWholesale Trade                              0.002998 **
## spouse.earned.2013No spouse                          0.017481 *
## spouse.earned.2013Refusal                            0.001163 **
## spouse.earned.2013Unknown                            0.489853
## spouse.earned.2013Yes                                0.148517
## highest.degree.2011Bachelor's                         < 2e-16 ***
## highest.degree.2011GED                               0.175600
## highest.degree.2011High School Diploma               0.003293 **
## highest.degree.2011Junior College                    3.91e-09 ***
## highest.degree.2011Master's                           < 2e-16 ***
## highest.degree.2011PhD                               1.06e-08 ***
## highest.degree.2011Professional                       < 2e-16 ***
## highest.degree.2011Unknown                           0.011359 *
## total.incarnations1-2                                0.167922
## total.incarnationsmore than 2                        0.295870
## gross.family.income                                   < 2e-16 ***
## gendermale:raceHispanic                              0.012713 *
## gendermale:raceMixed                                 0.177326
## gendermale:raceOther                                 0.016178 *
## gendermale:industryAcs Special                       0.195039
## gendermale:industryAgriculture                       0.008625 **
## gendermale:industryConstruction                      0.607726
## gendermale:industryEducation, Health & Social        0.464365
## gendermale:industryFinance                           0.933326
## gendermale:industryInformation & Communication       0.641701
## gendermale:industryManufacturing                     0.823282
## gendermale:industryMilitary                          0.962743
## gendermale:industryMining                            0.262208
## gendermale:industryOther services                    0.760999
## gendermale:industryProfessional                      0.984306
## gendermale:industryPublic Admin                      0.001874 **
## gendermale:industryRetail Trade                      0.724586
## gendermale:industryTransportation & Warehousing      0.016060 *
## gendermale:industryUnknown                           0.019776 *
## gendermale:industryUtilities                         0.003615 **
## gendermale:industryWholesale Trade                   0.736399
## gendermale:spouse.earned.2013No spouse               1.47e-12 ***
## gendermale:spouse.earned.2013Refusal                 0.365920
## gendermale:spouse.earned.2013Unknown                 0.728825
## gendermale:spouse.earned.2013Yes                     0.465975
## gendermale:highest.degree.2011Bachelor's             0.959720
## gendermale:highest.degree.2011GED                    0.792809
```

```
## gendermale:highest.degree.2011High School Diploma 0.303436
## gendermale:highest.degree.2011Junior College      0.341451
## gendermale:highest.degree.2011Master's            0.903140
## gendermale:highest.degree.2011PhD                 0.349375
## gendermale:highest.degree.2011Professional        0.474391
## gendermale:highest.degree.2011Unknown             0.290990
## gendermale:total.incarnations1-2                  0.489608
## gendermale:total.incarnationsmore than 2          0.913003
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17990 on 4727 degrees of freedom
##   (4186 observations deleted due to missingness)
## Multiple R-squared:  0.3633, Adjusted R-squared:  0.3539
## F-statistic: 38.53 on 70 and 4727 DF,  p-value: < 2.2e-16
```

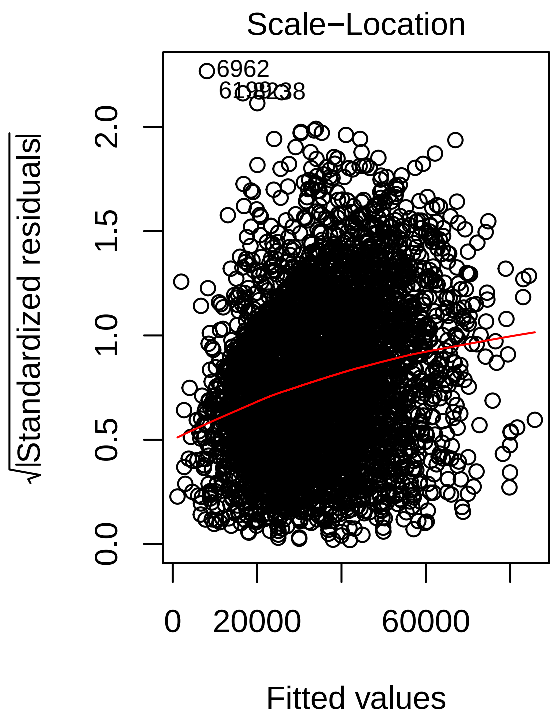
gendermale coefficient: The coefficient for term `gendermale` is now reduced to 6825. This implies if you control for other factors and interactions, a male respondent earns $6825 more than a female respondent. Recall when we started with only `gender` and `race` as a variable, this gap was 9892. We have identified certain factors that, on the whole, increased this gap and now when we are controlling for those factors, we see a reduction in this coefficient.

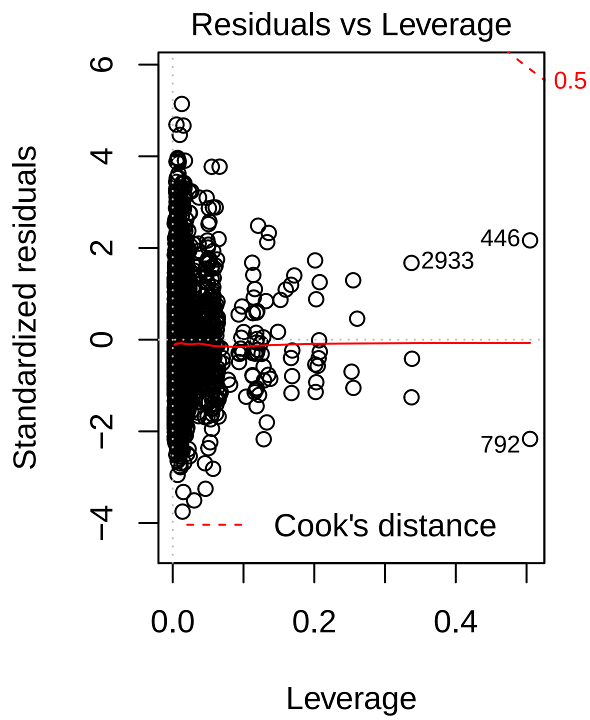
Diagnostics

Let's plot some diagnostic plots for our final regression model.



Residuals vs Fitted
Residuals
Fitted values
6962
0
20000
60000
-50000
50000
100000
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−4
−2
0
2
4
6





1. **Residuals vs Fitted Plot:** The red line shows the average value of the residuals at each fitted value. This indicates that, on average, there is no trend to the residuals. However notice the lower portion of the plot. There is some kind of an increasing (in negative direction) trend to the residuals. So for data points with negative residuals, there is an indication of **increasing variance**.
2. **Normal QQ Plot:** This plot tells us whether the residuals from our model are normally distributed. There seems to be a clear **heavy tail** (residuals at tail tend to have larger values than expected in a normal distribution).
3. **Residuals vs Leverage:** There are some points (2933, 446, 793) with high leverage and high residuals. These might be potential outliers affecting the model fit.

Final Comments

Approach Summary and Insights

In this project, we have looked at the income difference between men and women, and the factors that potentially aggravate or reduce this gap. To start the analysis, we selected some variables based on intuition, general understanding of gender stereotypes, and prior knowledge/biases. We then visualized the difference in income across each of these variable with means of plots and summaries to validate/contradict our assumptions. To get a more concrete conclusion we performed hypothesis tests and built a regression model cumulatively, adding one term at a time and testing its significance. We interpreted the implications of adding each interaction term, and whether it contradicts with our findings.

Finally, we went back and forth to remove some variables from our final model and ended up selecting the following variables:

- Gender
- Race
- Industry
- Whether spouse earns or not?
- Highest degree earned
- Total incarnations
- Gross family income

The variables that we initially thought would be good predictors, but later rejected were:

- age
- Whether respondent takes drugs?
- Household net worth
- Marital Status
- Spouse's income

Once we finalized our model, we built some diagnostics plots to assess the performance of our model.

Some of the key insights highlighted throughout this project are:

1. `Black` race has a low association with income gap, i.e., there seems to be lesser difference in incomes of men and women, than other race groups.

2. Respondents were within 30-34 years of age. The variation in their age does not seem to have a significant impact on income difference. However, the small range might be a contributing factor for this conclusion.

3. `Drug use` analysis initially presented some contradictory results. However this just might be due to a large sample not taking drugs, and hence subjected to other factors affecting this gap, and not drug use per se. Adding more terms in our model eventually made this variable insignificant.

4. `Spouse` factors seem to be important when it comes to estimating income gap between genders. Sample of respondents with an earning spouse seem to have a significantly higher income gap between genders than those who are single or do not have an earning spouse.

5. `Industry` and `education` are expected to impact income, and they do show association. Different industries also show a variation in income difference between men and women. Some industries like Finance, construction, admin. staff etc. tend to have males earning more than females on average. For education, we considered highest degree earned. Our regression coefficients convey that compared to respondents with no degrees, income gap between men and women seems to be higher for respondents with Bachelor's, Master's and High School Diploma.

6. Household net worth in 2003 did not seem to be a good predictor when gross family income was also added. This might be due to multicollinearity between the two variables.

Confidence in final model

I think some of the variables that I selected performed really well in predicting income gap. For example, whether spouse earned or not: Plots, hypothesis tests, and regression results all show a high association of an earning spouse and income difference between men and women. Industry, education and race also gave some believable insights.

The model building approach - adding variables and then removing from final model, did validate most of my earlier contradictions.

I am not so confident about the `total.incarnations` variable though. It provides contradictory results. Part of the reason might be the decision to convert it into a categorical variable. The bins I created were disproportionate.

Also, I was aiming to focus mostly on the "income gap" aspect, and not income. This might be one of the reasons why the diagnostic plots show a potential contradiction for linear model assumptions.

In the initial selection of variables, I overlooked most of the childhood factors that might have an impact on income difference. These include emotional factors, household's economic status in childhood etc. Some other variables that might have been worth looking were household size, schooling, college type et.c

I am comfortable giving a few recommendations to policy makers based on this model (spouse, industry, race). But, to present the entire model to them would take some more iterations to correct and discover potential flaws.