ILLINOIS TECH

GENDER RELATED INCOME DISPARITY

Programming for Data Analytics (ITMD-514_05)
Final Project: ITMD 514 - Group 4
Prepared under supervision of Prof. Despina Stasi

Presented By:
Abhishek Anand
Apurva Anand
Surajit Patra

Working with raw data and Variable selection

Dataset: National Longitudinal Surveys (NLSY)
Data Rows: 8984
Variables: 95
Goal: To identify any difference in income between male and female
Raw data-frame: nlsy

Understanding the dataset:

- Top-coded income for 2% earner
- Missing/Negative values
- Numerical variable names
- Possible responses for each variable
- Selected Variables: 15

RAW Data - nlsy

```
## Rows: 8,984
## Columns: 95
## $ B0004600 <dbl> -4, -5, -4, -4, -5, 41, -5, -5, 251, -5, 280, -4, -5, -5, 334…
## $ E8043100 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0…
## $ E8043200 <dbl> -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -…
## $ E8043400 <dbl> -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -…
## $ R0000100 <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18…
## $ R0069400 <dbl> 1, 2, 2, 2, 2, 2, 2, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 4, 1, 2…
## $ R0070000 <dbl> 1, 2, 3, 3, 2, 2, 2, 1, 1, 1, 2, 2, 2, 2, 3, 2, 3, 4, 2, 2…
## $ R0323900 <dbl> -4, 3, 0, -4, 3, 4, 5, -4, 1, 1, 0, -4, 1, -4, 1, 3, -4, 0, 0…
## $ R0358900 <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1…
## $ R0513500 <dbl> 10, -4, -4, 46, -4, -4, -4, 25, -4, -4, -4, 50, -4, 90, -4, -…
## $ R0514700 <dbl> 100, -4, -4, 100, -4, -4, -4, 100, -4, -4, -4, 95, -4, 100, -…
## $ R0514800 <dbl> 10, -4, -4, 0, -4, -4, -4, 0, -4, -4, -4, 0, -4, 0, -4, -4, 0…
## $ R0514900 <dbl> 0, -4, -4, 0, -4, -4, -4, -2, -4, -4, -4, 0, -4, 0, -4, -4, 2…
## $ R0515100 <dbl> 100, -4, -4, 100, -4, -4, -4, 100, -4, -4, -4, 95, -4, 70, -4…
## $ R0536300 <dbl> 2, 1, 2, 2, 1, 2, 1, 2, 1, 1, 2, 1, 1, 1, 2, 1, 2, 1, 1, 1, 1…
## $ R0536401 <dbl> 9, 7, 9, 2, 10, 1, 4, 6, 10, 3, 6, 10, 11, 7, 1, 2, 11, 2, 4,…
## $ R0536402 <dbl> 1981, 1982, 1983, 1981, 1982, 1982, 1983, 1981, 1982, 1984, 1…
## $ R0648900 <dbl> 0, -4, 0, 0, 0, 0, 0, 0, 0, 0, 0, -4, 0, -4, 0, 0, 0, 0, 0…
## $ R0649100 <dbl> 0, -4, 1, 1, 0, 0, 0, 0, 0, 0, 0, -4, 0, -4, 0, 0, 0, 0, 0…
## $ R0681300 <dbl> 0, -4, 0, 1, 0, 0, 0, 0, 0, 0, 0, -4, 0, -4, 0, 0, 0, 0, 0…
## $ R0690800 <dbl> -4, -4, 1, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4, -4…
```

VARIABLES SELECTED		
gender	income	age
enrollment	citizenship	total income
race	industry code	marijuana
trustful	college type	incarcerations
birth yr	is sad	highest grade

NEGATIVE VALUE	
-1	Refusal
-2	Don't know
-3	Invalid skip
-4	Valid
-5	Non-interview

Data cleaning and EDA

New dataframe for selected variable: nlsy_df

Processing Applied to Raw Data:

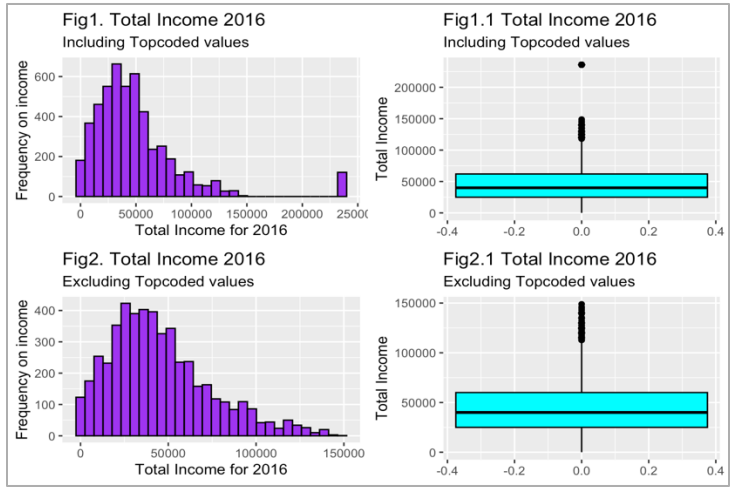
- Handled negatives by replacing with NA
- Omitted NA for some variables(income) and considered for some (current enrollment status)
- Converted numerical variables to categorical for better insights
- Factored variables with relevant factor levels (gender, race, industry code, marital status, etc)
- Truncated top-coded values for some observation
- Topcoded income count is 121
- Topcoded income was averaged to $235,884
- Factor levels were combined as 'Others' for variable like race (Hispanic, Black, Others combined for Mixed race and Non-black), trustful (trustful for values >=4 and untrustful for values 1 to 3)

Cleaned data – nlsy_df

```
## Rows: 8,984
## Columns: 15
## $ gender                    <fct> Female, Male, Female, Female, Male, Female,…
## $ birth_year                <dbl> 1981, 1982, 1983, 1981, 1982, 1982, 1983, 1…
## $ is_sad_depressed_unhappy_f <chr> NA, NA, "Sometimes true", NA, NA, NA, NA, N…
## $ current_enrollment_status <chr> "Enrolled in grades 1-12, not a high school…
## $ total_income_2016         <dbl> NA, 116000, NA, 45000, 125000, NA, NA, NA, …
## $ race                      <fct> Other, Hispanic, Hispanic, Hispanic, Hispan…
## $ highest_grade_completed   <chr> NA, NA, NA, NA, NA, NA, NA, NA, "6th Year C…
## $ age_1996                  <dbl> 15, 14, 13, 15, 14, 14, 13, 15, 14, 12, 14,…
## $ citizenship               <chr> "Non-Citizen", NA, "Citizen", "Non-Citizen"…
## $ used_marijuana            <fct> No, No, No, No, Yes, No, No, No, No, No, No…
## $ industry_code             <chr> "Wholesale Trade", "Public Administration",…
## $ marital_status            <fct> NA, Never-married, Married, Never-married, …
## $ total_num_incarcerations  <chr> "No incarcerations", "No incarcerations", "…
## $ college_type              <chr> NA, NA, NA, NA, NA, NA, NA, NA, "Private no…
## $ trustful_or_not           <fct> NA, Trustful, Trustful, NA, Trustful, Trust…
```

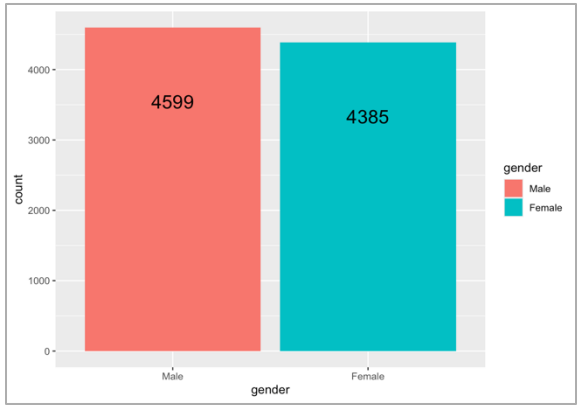
Selected Variables analysis

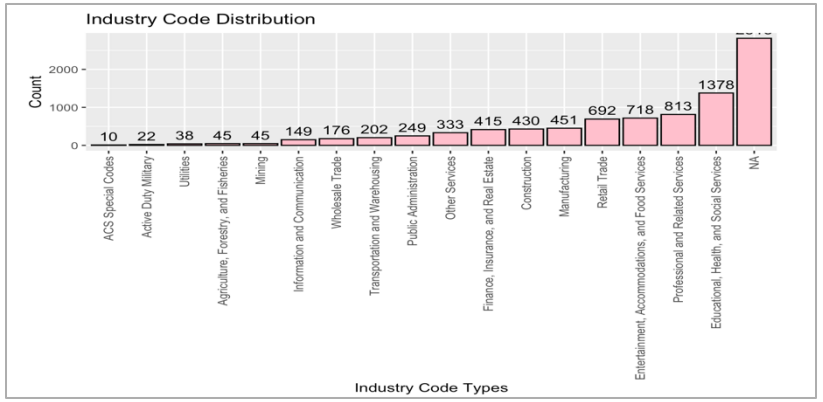
Total Income of 2016 Including/Excluding Top-coded values

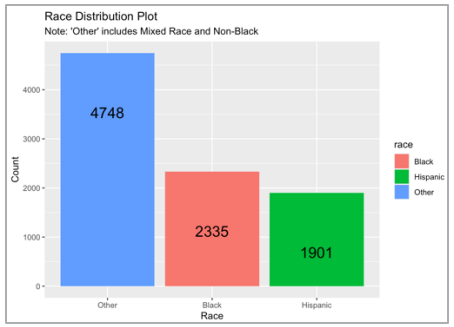


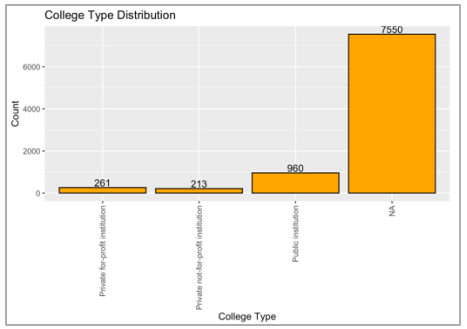
- Fig1 shows right skewed plot indicated very few people have high income of 235885(topcoded) and median is approx $40K.
- Fig2. shows skewed but less extreme distribution more accurate income and median $35K

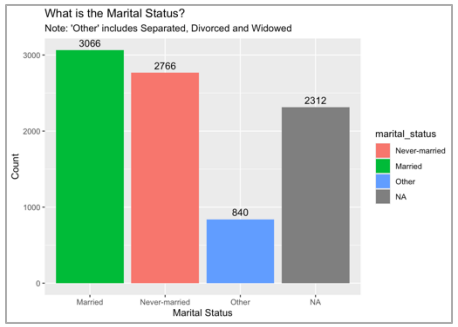
Analysis of other variables

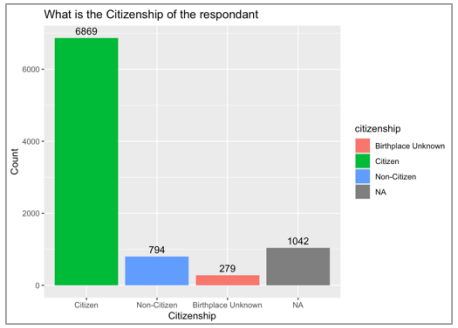


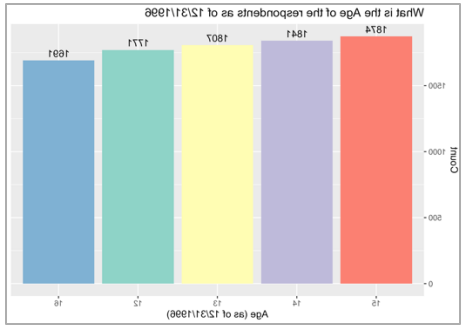


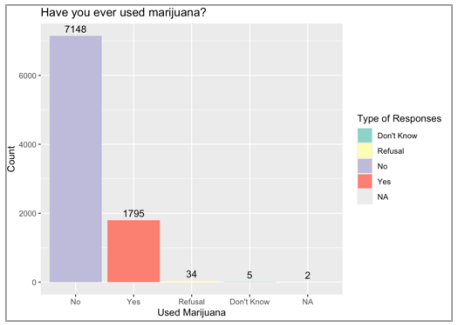










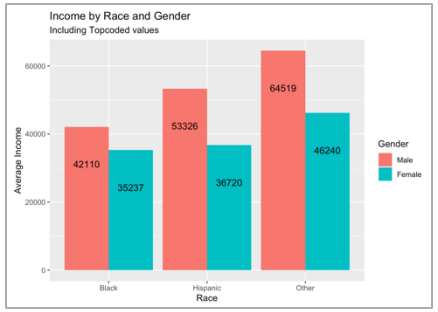


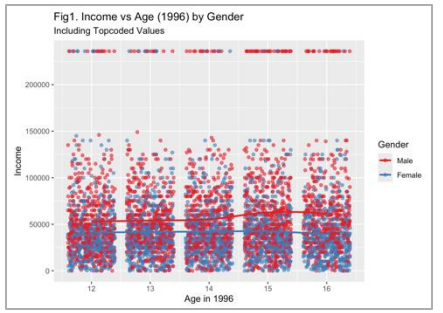
Relationship and Trend Analysis

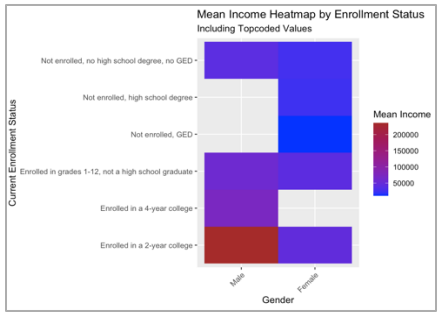
Income Distribution by Gender

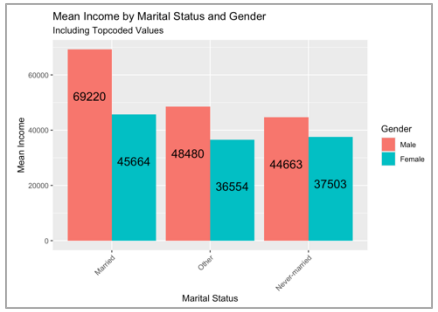
- Graph shows significant disparity in income for both groups with huge variability for males.
- Fig 1 shows income spreaded to $235,884 because of topcoded value while Fig 2 shows max income around $150K.

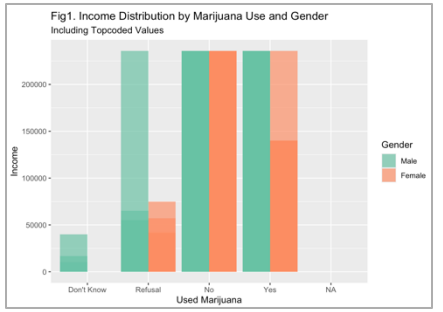
Income distribution by Gender and other variables

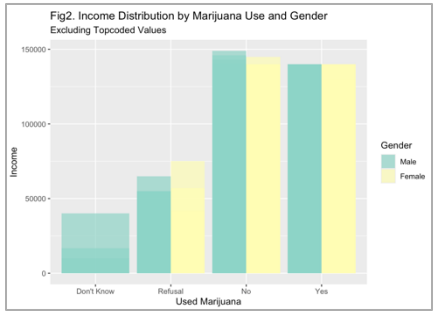


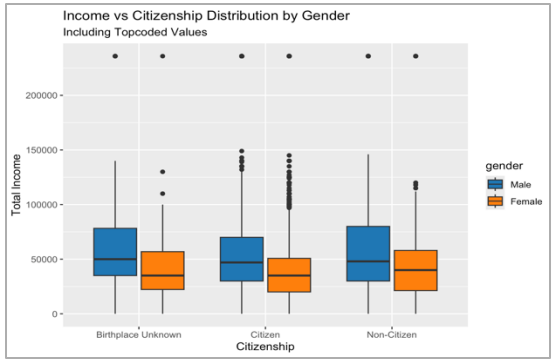


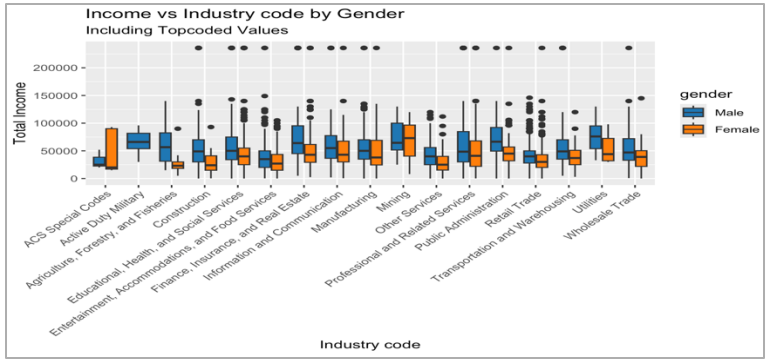












Normality check and TOBIT Testing

```
# To check for normality of total_income_2016 we have performed the Skewness and Kurtosis tests to check for norm
ality.

skewness(nlsy_df$total_income_2016, na.rm = TRUE)
```

```
## [1] 2.435323
```

```
kurtosis(nlsy_df$total_income_2016, na.rm = TRUE)
```

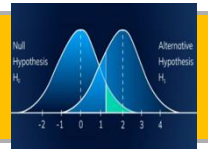
```
## [1] 11.26869
```

- Skewness 2.435 indicates total_income is positively skewed which means it has longer right tail. Majority of the income values are concentrated toward the lower end, with fewer values extending toward the higher end. Hence it can be determined that the data deviates from normality.
- Kurtosis 11.268 indicates it is leptokrutic distribution meaning heavy tailed distribution.
- Finally, we can say that our data deviates from normality

TOBIT REGRESSION

- Tested out Tobit to see how it handles topcoded values.
- Gender coefficient is statistically significant and indicates females earning (approximately $16180) less than males is not by chance but exists across the dataset.
- P-value < 2.2e-16 which is significantly smaller than 0.05, suggesting rejection of Null Hypothesis (H_o) in support of Alternate Hypothesis (H_a)
- 95% CI [13747, 18099] and mean income for males is $57202 while for females $41278

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept):1  5.740e+04  8.070e+02   71.13   <2e-16 ***
(Intercept):2  1.062e+01  1.087e-02  976.76   <2e-16 ***
genderFemale  -1.618e+04  1.164e+03  -13.89   <2e-16 ***
---
```



CI and Hypothesis testing

Top-coded CI

```
##
##  Welch Two Sample t-test
##
## data:  total_income_2016 by gender
## t = 14.346, df = 4876.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Male and group Female is not equal to 0
## 95 percent confidence interval:
##  13747.84 18099.96
## sample estimates:
##   mean in group Male mean in group Female
##             57202.82             41278.92
```

- 95% CI [13747, 18099]
- The mean income of male is $57,202 and female is $41,278. The wider interval shows greater variability in the mean income.
- The higher mean income is due to topcoded extreme values.

Non- Top-coded CI

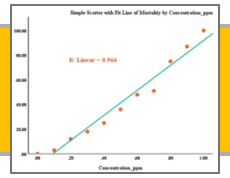
```
##
##  Welch Two Sample t-test
##
## data:  total_income_2016 by gender
## t = 14.91, df = 4938.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Male and group Female is not equal to 0
## 95 percent confidence interval:
##  10326.42 13453.00
## sample estimates:
##   mean in group Male mean in group Female
##             50775.95             38886.23
```

- 95% CI [10326, 13453] and p-value suggests rejection of H_o
- The mean income of male is $50,775 and female is $38,886. This shows narrower interval shows smaller variability in the mean income.
- The lower mean income is beacuse of truncation of topcoded extreme values.
- T-statistic implies stronger difference in mean income

Hypothesis Tests for Top-coded and Non-Top-coded

```
## [1] "Reject the null hypothesis: Income significantly differs between males and females."
```

- p-value suggests rejection of H_o as the value 2.2e-16 is less than 0.05



Linear Regression With and Without Top-coded

With Top-coded

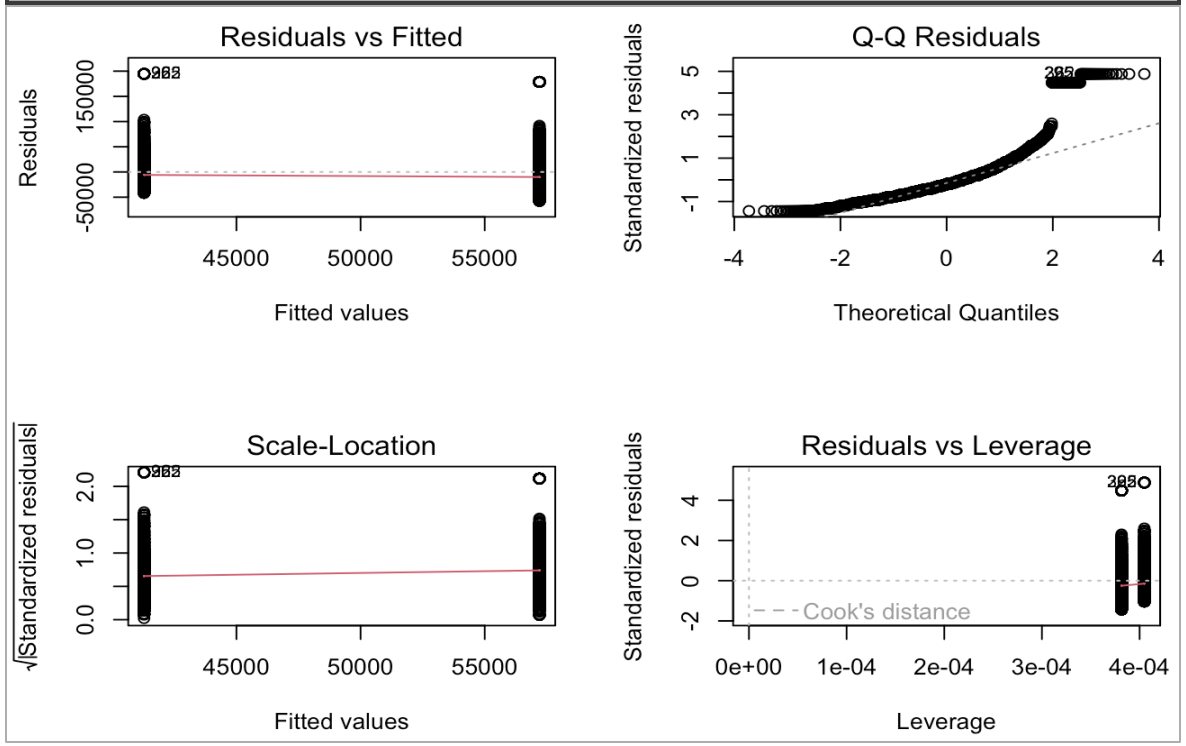
```
##
## Call:
## lm(formula = total_income_2016 ~ gender, data = filtered_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -57203 -24203  -8203  12797 194605
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   57202.8      779.3   73.41   <2e-16 ***
## genderFemale -15923.9     1118.8  -14.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39900 on 5089 degrees of freedom
## Multiple R-squared:  0.03829,    Adjusted R-squared:  0.0381
## F-statistic: 202.6 on 1 and 5089 DF,  p-value: < 2.2e-16
```

Without Top-coded

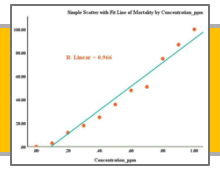
```
##
## Call:
## lm(formula = total_income_2016 ~ gender, data = notop_filtered_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50776 -20776  -4776  14224 106114
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50775.9      559.9   90.69   <2e-16 ***
## genderFemale -11889.7      799.1  -14.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28160 on 4968 degrees of freedom
## Multiple R-squared:  0.04267,    Adjusted R-squared:  0.04247
## F-statistic: 221.4 on 1 and 4968 DF,  p-value: < 2.2e-16
```

- Intercept 57202 is the mean income for males
- Gender Female coefficient indicates females earn 15923 less than males on average.
- Significantly small p-value confirms gender is a strong predictor of income
- R2 = 0.0381 explains only 3.8% variability in income suggesting other factors contribute to income differences.

Total Income and Gender

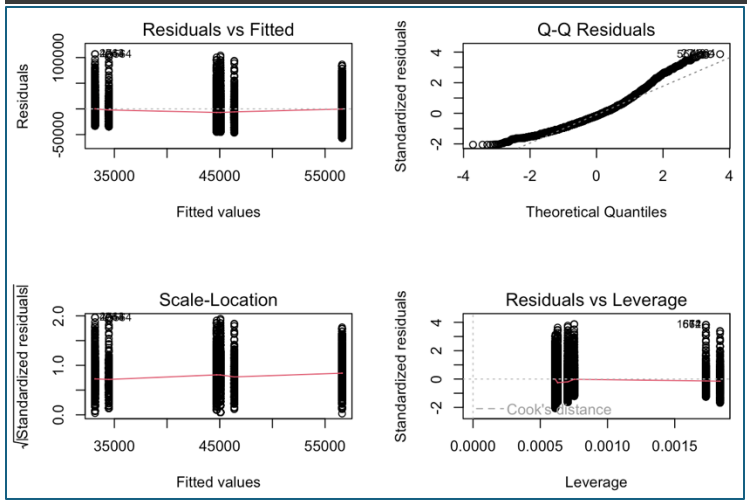


- The residuals appear scattered but exhibit some clustering at extreme values, indicating possible deviations from linearity.
- The residuals deviate significantly from the diagonal line suggesting the residuals are not normally distributed.
- The residuals' spread appears uneven, with greater variation at extreme fitted values, indicating heteroscedasticity (constant variance of residuals).
- Several observations near the top right corner with high leverage suggest some data points may disproportionately impact the model.



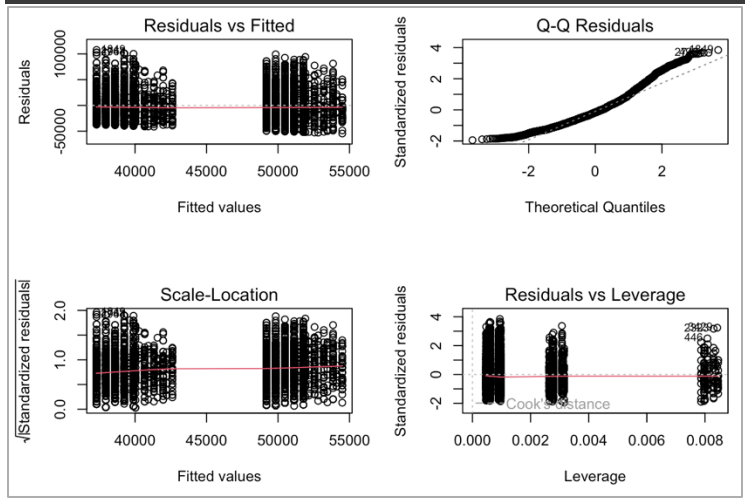
Linear Regression Top-coded

Total Income by Marital status and Gender



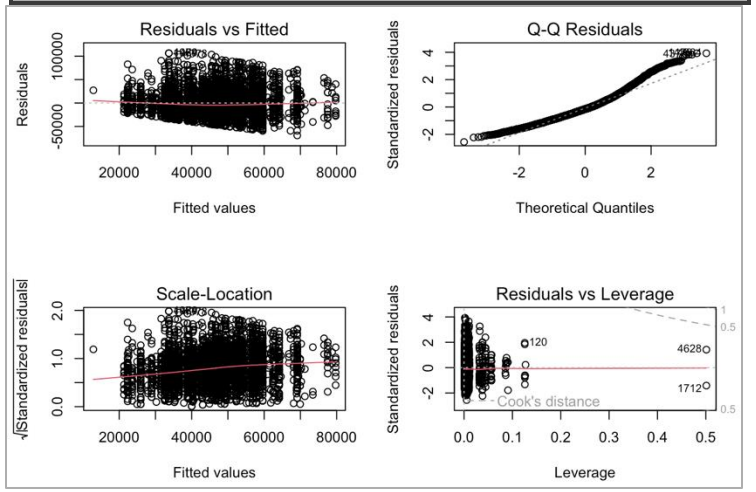
- Uneven residual patterns, with clustering and potential non-linearity issues. | Possible violations of **homoscedasticity** (equal variance)
- Residuals deviate significantly in the tails | indicates **strong non-normality**.
- Upward trend in the red line, shows increasing residual spread with larger fitted values | Indicates **mild heteroscedasticity**.
- Points 1672 and 1674 have high leverage but do not cross Cook's threshold | **Outliers exist** but are not highly influential.

Total Income by age, citizenship and Gender



- Residuals are scattered around zero but show some mild patterns | Slight linearity or homoscedasticity issues.
- Residuals deviate in the tails but align well in the middle | Suggests **mild non-normality**.
- Mostly flat red line, with a slight increase in spread at higher fitted values | Suggests **minimal heteroscedasticity**.
- Points 244, 1200, and 3220 have high leverage but are within Cook's threshold | Indicates **mild influence from these points**.

Total Income by Race, Used marijuana, Industry code and Gender



- Residuals are scattered but show increasing variance with fitted values | **Mild heteroscedasticity** is observed.
- Deviations at the tails, but less deviation than in the Marital Status model | Indicates **moderate non-normality**.
- Flat red line with some variation in spread. | Confirms **mild heteroscedasticity**.
- Points 120, 4628, and 1712 have high leverage but remain within acceptable influence limits | Indicates **low risk of undue influence**.

CONCLUSION

Many factors impact the income disparity for both groups. We cleaned data and truncated top-coded values for getting different and accurate insights.

Plotting graphs for 15 variables against income and gender, gave consistent insight. Some exception like industry code ASC and for use of marijuana refusal shows females earn more.

However, across dataset females earn significantly less than males which is illustrated by all the plot we covered earlier.

The best fit linear regression model is **Age + Citizenship** model has the fewest violations of assumptions, showing only mild deviations in normality and variance.

We are confident in our analysis and findings that there is significant income gap between males and females.