Canada Median Income

Minh Tran, Fahad Bhutta, Aktan Kenzhebay, Saba Abbasi

2025-11-19

Abstract

This statistical study aims to determine patterns and trends in the median income of people living in Canada at two time periods: 2015 and 2020. The data used in the study was retrieved from Statistics Canada.

Through this study, we determined the difference (or lack thereof) in household income in specific scenarios such like time periods and household sizes/situations. From these findings, we hope to be able to make predictions and recommendations to relevant parties.

Introduction

The dataset chosen is Household Income Statistics by household type, and it includes households from all provinces and territories in Canada. The data was collected through a census, and it has data from broad, general areas like whole provinces, and smaller, more specific areas like cities. The census in which data was derived from was done in 2016 and 2021, and the median incomes provided was from 2015 and 2020 respectively.

Link to dataset: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=9810005701

Link to cleaned dataset: https://drive.google.com/file/d/1kKOXK_0hBLu9F0nAYekCuXURS9noOP5v/view?usp=drive_link

Research questions

Question 1: Is there a difference in median after tax household income between 2015 and 2020?

Question 2: Is there a significant difference in income between households of 5 or more people and 4 persons (in 2020)? Do households with 5 or more people earn more than households with 4 people?

Question 3: Is there a difference in median income between Households with and without children? Do households with children earn more than households without children?

Variables selected:

Question 1: Median household after-tax income (2015), Median household after-tax income (2020)

Question 2: Median household total income (2020), Household Size: '4 persons' and '5 or more persons'

Question 3: Median household total income (2020), Household type including census family structure: 'With children' and 'Without children'

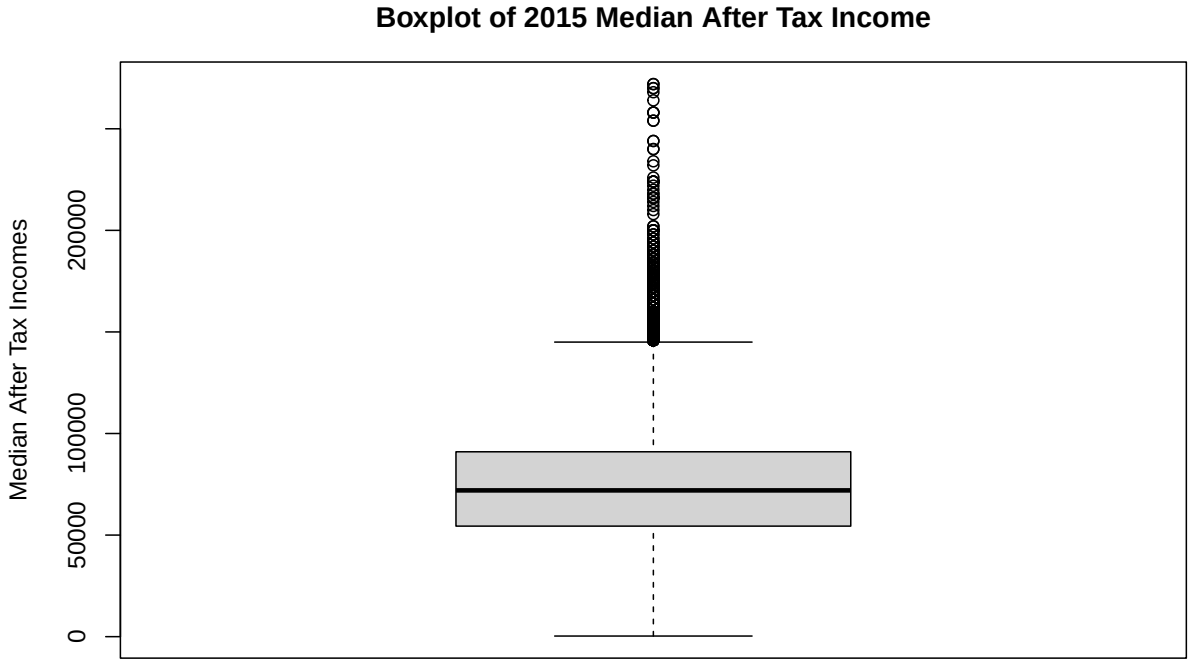
We excluded any rows that were had empty cells (missing values) or had a '0' in the Number of Households column. Since there are times when the column had data for 2016 and not 2021, or vice versa, we chose to completely remove any rows that had empty cells.

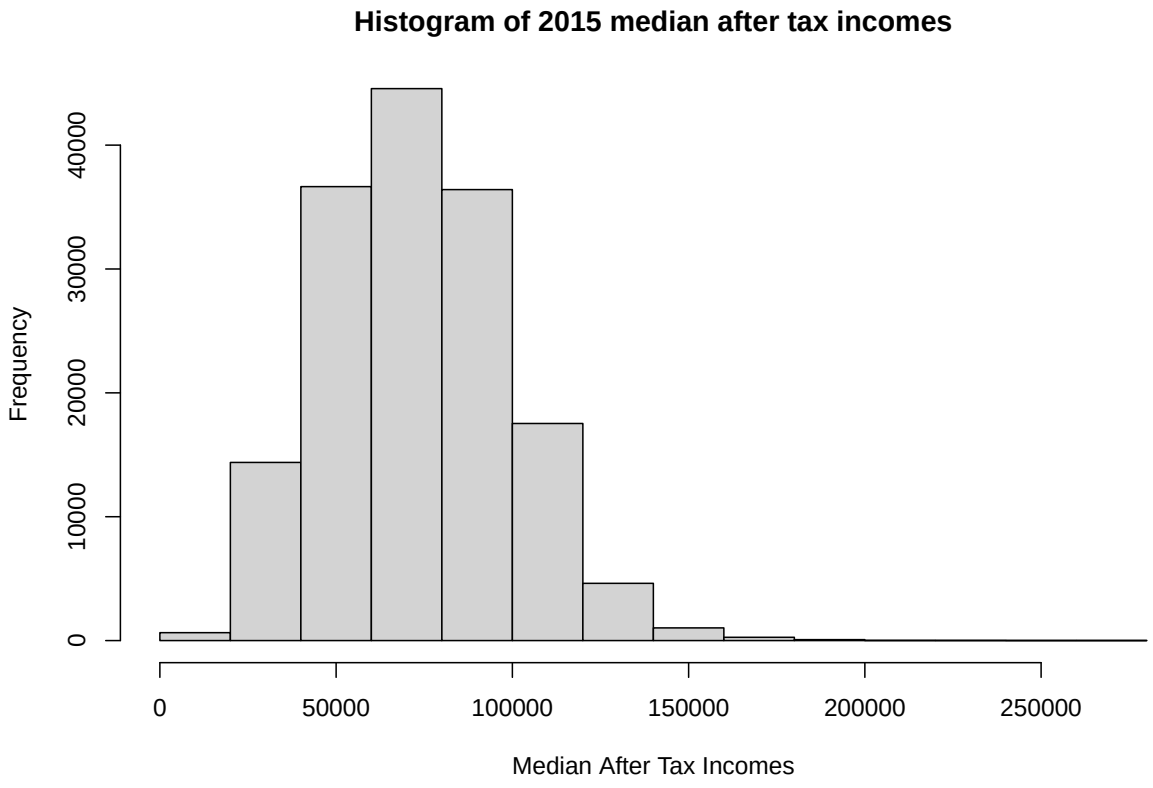
Statistical Results

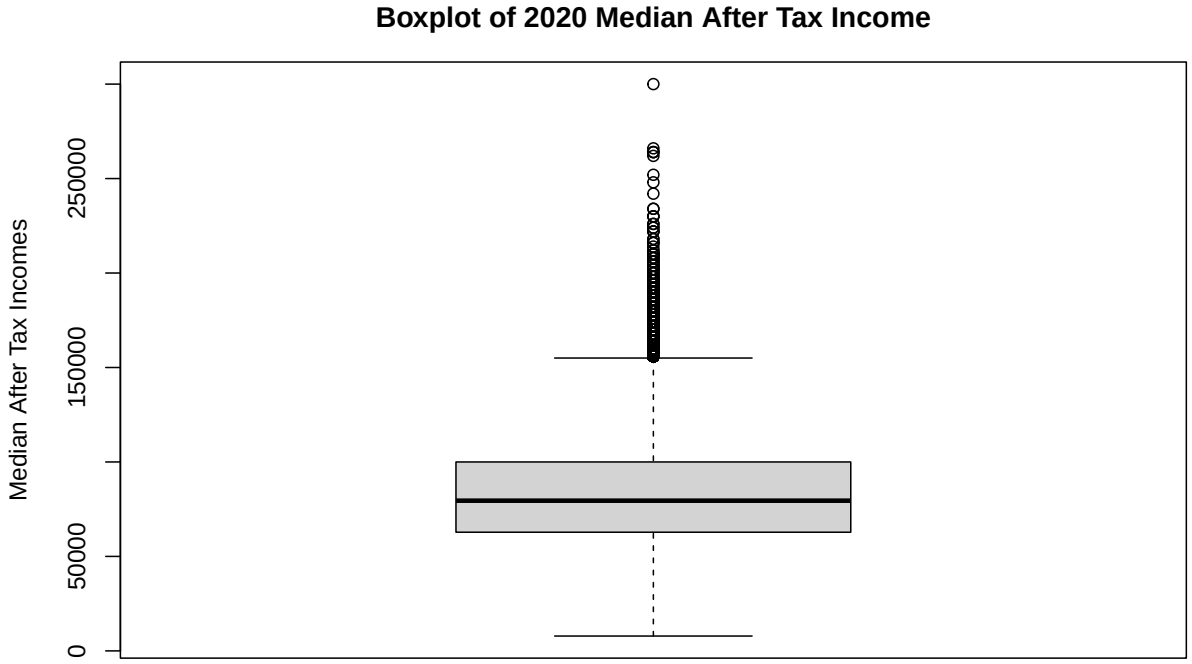
Question 1: Is there a difference in median after tax household income between 2015 and 2020?

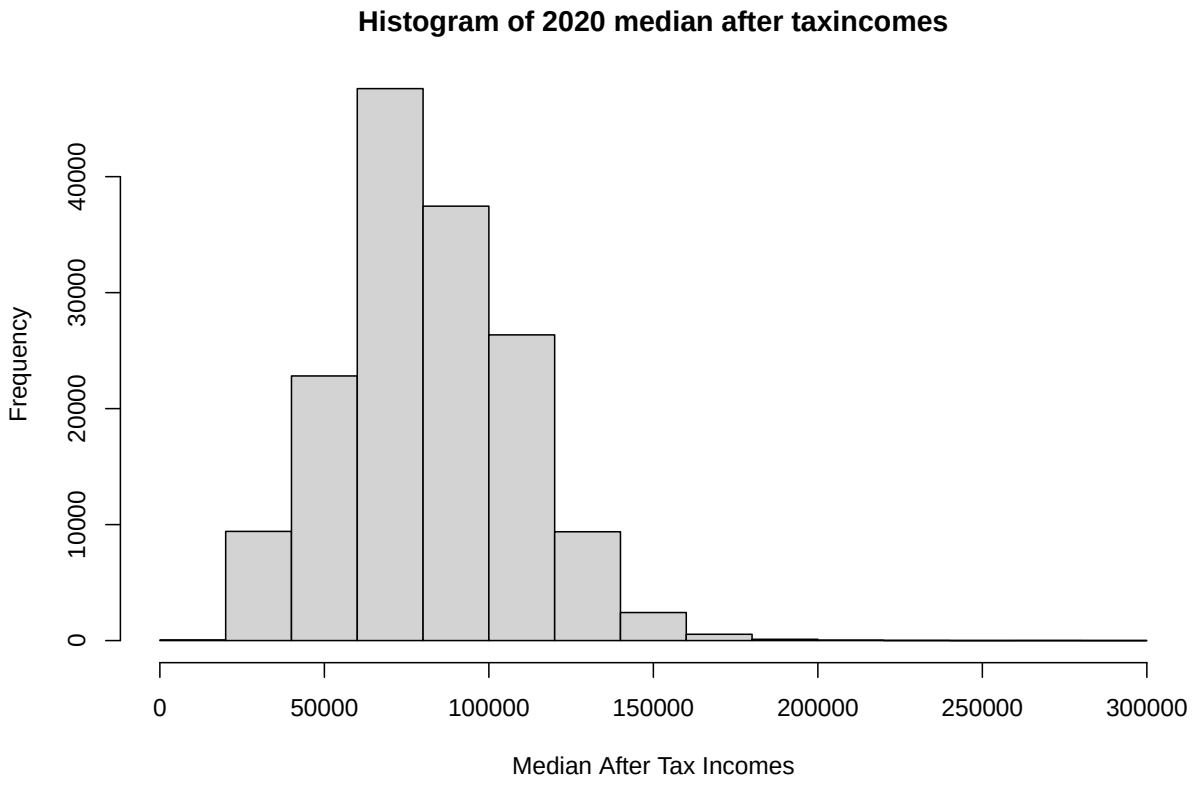
min	Q1	median	Q3	max	mean	sd	n	missing
7850	62800	79500	1e+05	3e+05	81984.76	26934.17	156189	0

min	Q1	median	Q3	max	mean	sd	n	missing
294	54400	72000	91000	272000	73357.16	26053.9	156189	0


Boxplot of 2015 Median After Tax Income
Median After Tax Incomes
0
50000
100000
200000


Histogram of 2015 median after tax incomes
Frequency
0
10000
20000
30000
40000
0
50000
100000
150000
200000
250000
Median After Tax Incomes


Boxplot of 2020 Median After Tax Income
Median After Tax Incomes
0
50000
150000
250000


Histogram of 2020 median after taxincomes
Frequency
0
10000
20000
30000
40000
0
50000
100000
150000
200000
250000
300000
Median After Tax Incomes

We performed a Welch's two sample t-test since the population variance is unknown, with Null Hypothesis (H0) being the median is equal between 2015 and 2020 (inc2020 = inc2015), and the Alternative Hypothesis (H1) being the medians are not equal (inc2020 != inc2015).

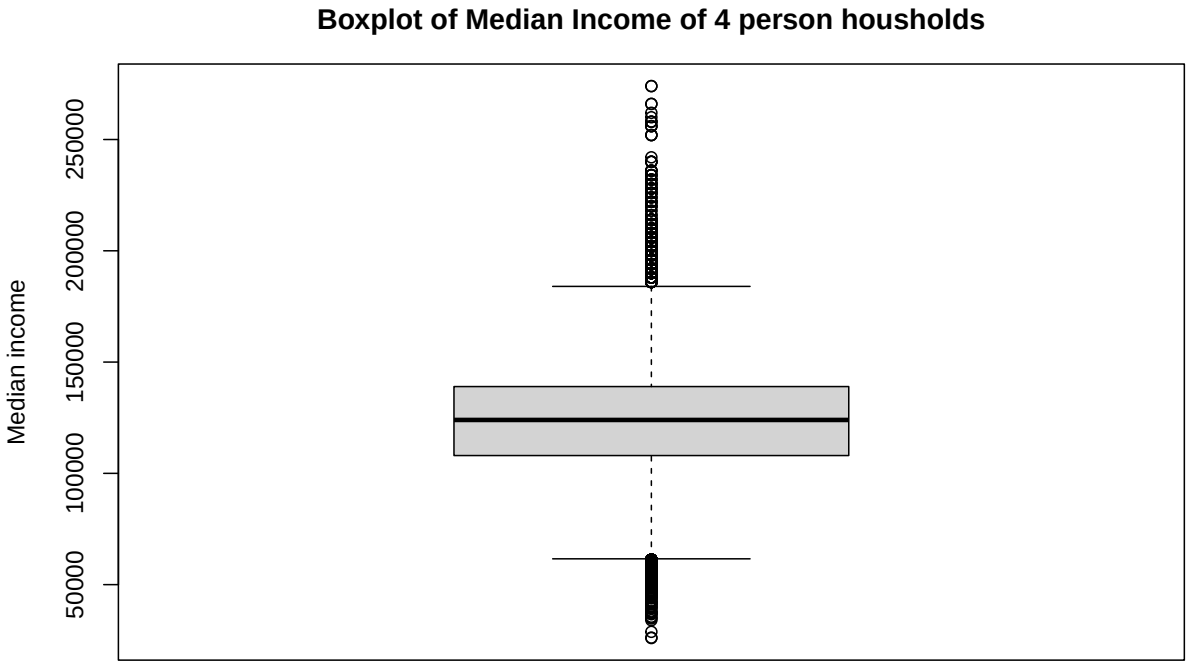
The p-value outputted from the t-test is less than 2.2e-16, which is extremely small, much smaller than our significance level of 5%. Thus, we conclude that there is a difference in the Median After Tax Household Income between 2015 and 2020.

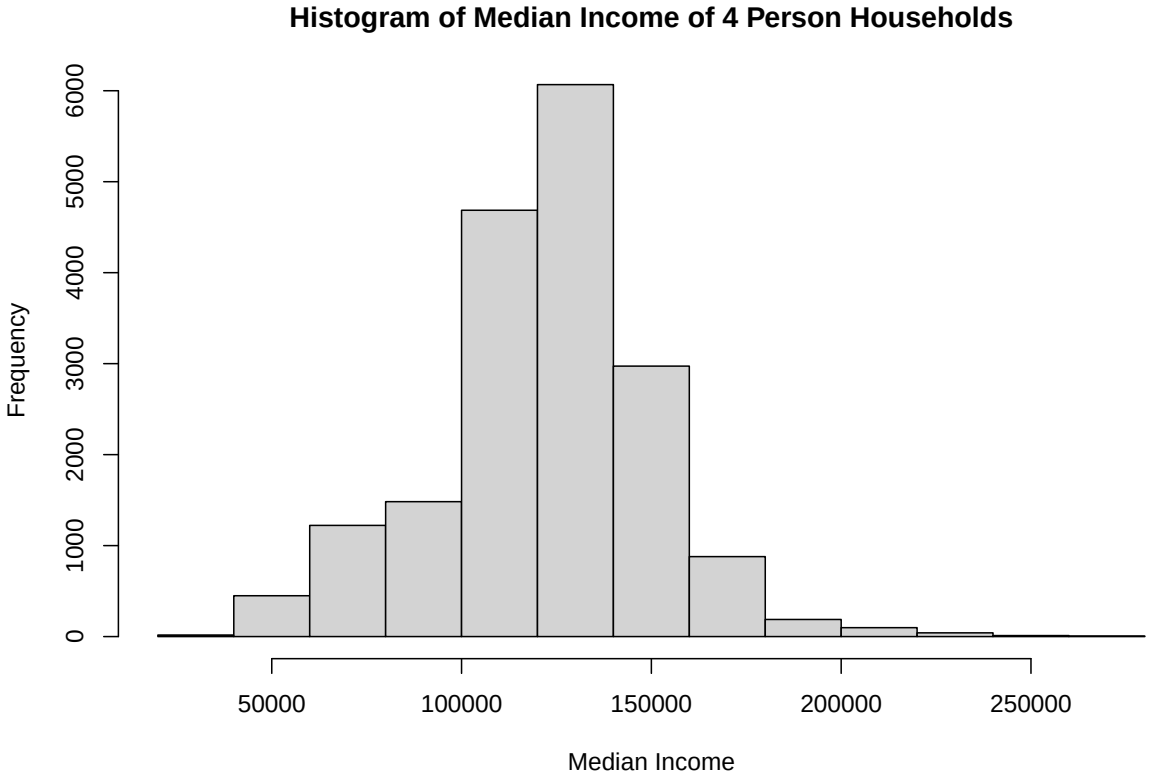
Additionally, the confidence interval generated for the difference between 2020 and 2015 median after tax income (inc2020 - inc2015) is (8441.755,8813.442), which has positive upper and lower confidence limits. This suggests that we can be 95% confident in concluding that the Median Household After Tax Income is higher in 2020 than 2015.

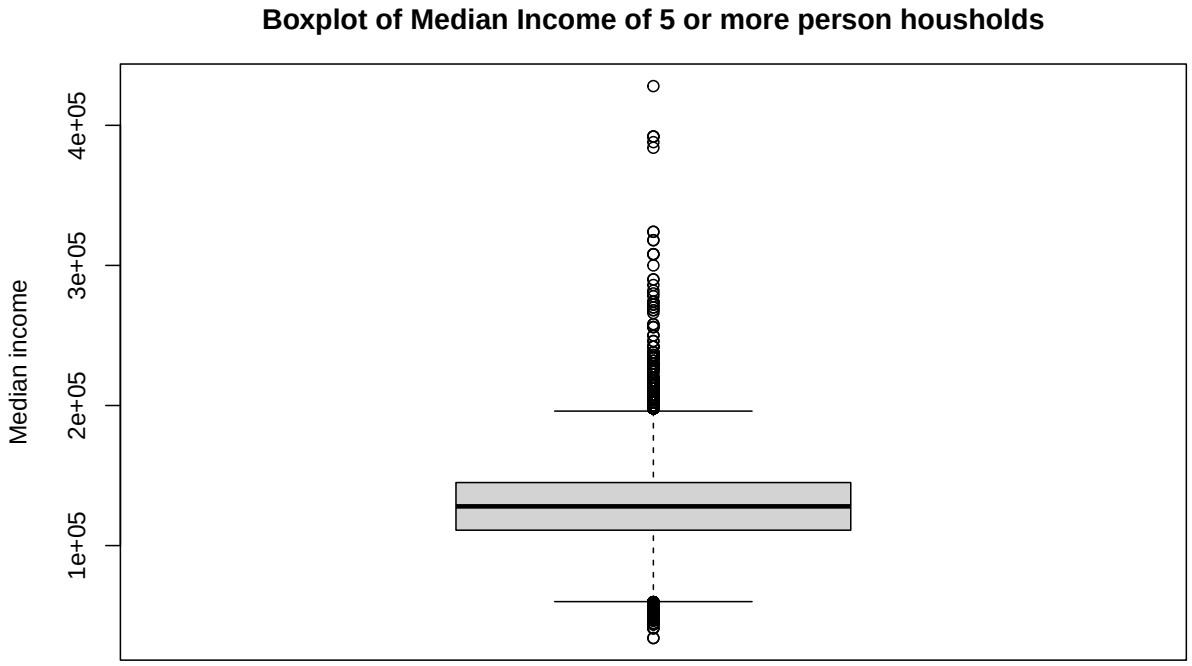
Question 2: Is there a significant difference in income between households of 5 or more people and 4 persons (in 2020)?

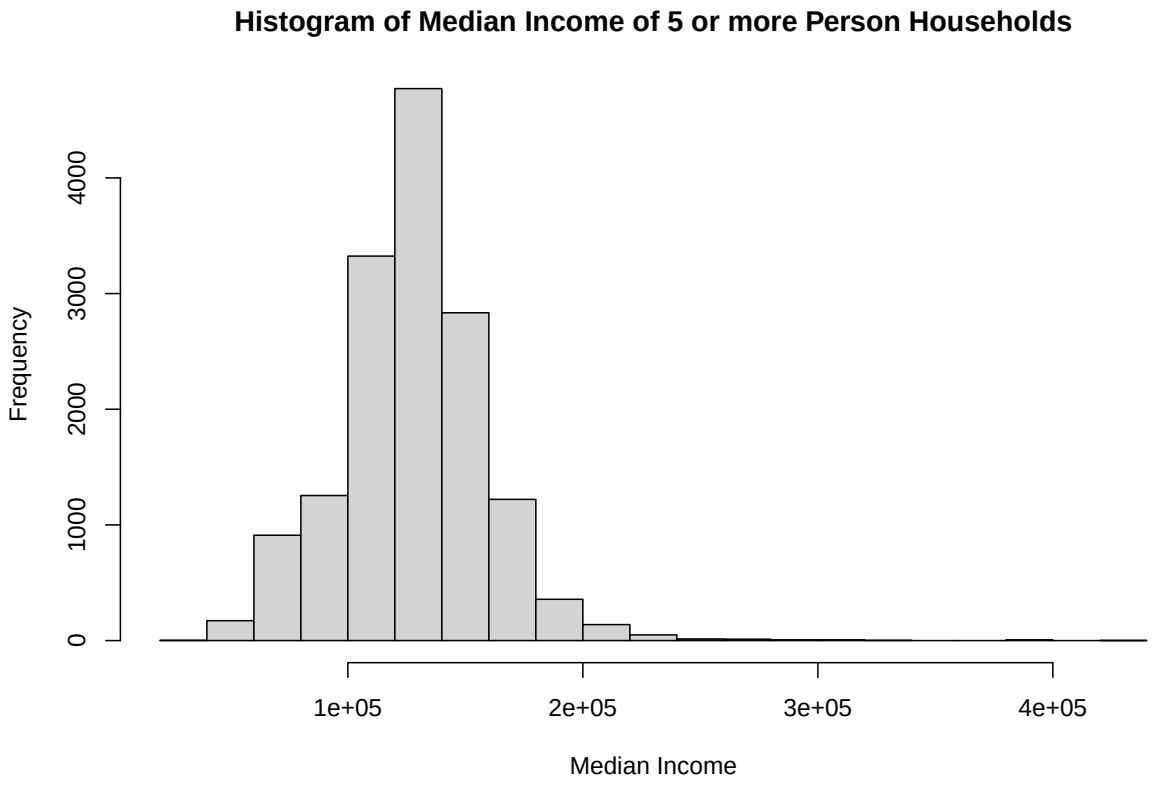
min	Q1	median	Q3	max	mean	sd	n	missing
26000	108000	124000	139000	274000	122603.8	28411.19	18119	0

min	Q1	median	Q3	max	mean	sd	n	missing
34000	111000	128000	145000	428000	128213.6	30803.87	15078	0


Boxplot of Median Income of 4 person housholds
Median income
50000
100000
150000
200000
250000


Histogram of Median Income of 4 Person Households
Frequency
0
1000
2000
3000
4000
5000
6000
50000
100000
150000
200000
250000
Median Income


Boxplot of Median Income of 5 or more person housholds
Median income
1e+05
2e+05
3e+05
4e+05


Histogram of Median Income of 5 or more Person Households
Frequency
0
1000
2000
3000
4000
1e+05
2e+05
3e+05
4e+05
Median Income

We performed a Welch's two sample t-test since the population variance is unknown. The Null Hypothesis (H0) is The median income between households of 4 persons and 5 or more persons are the same (inc4 = inc5p), and the Alternative Hypothesis (H1) is The median income are different.

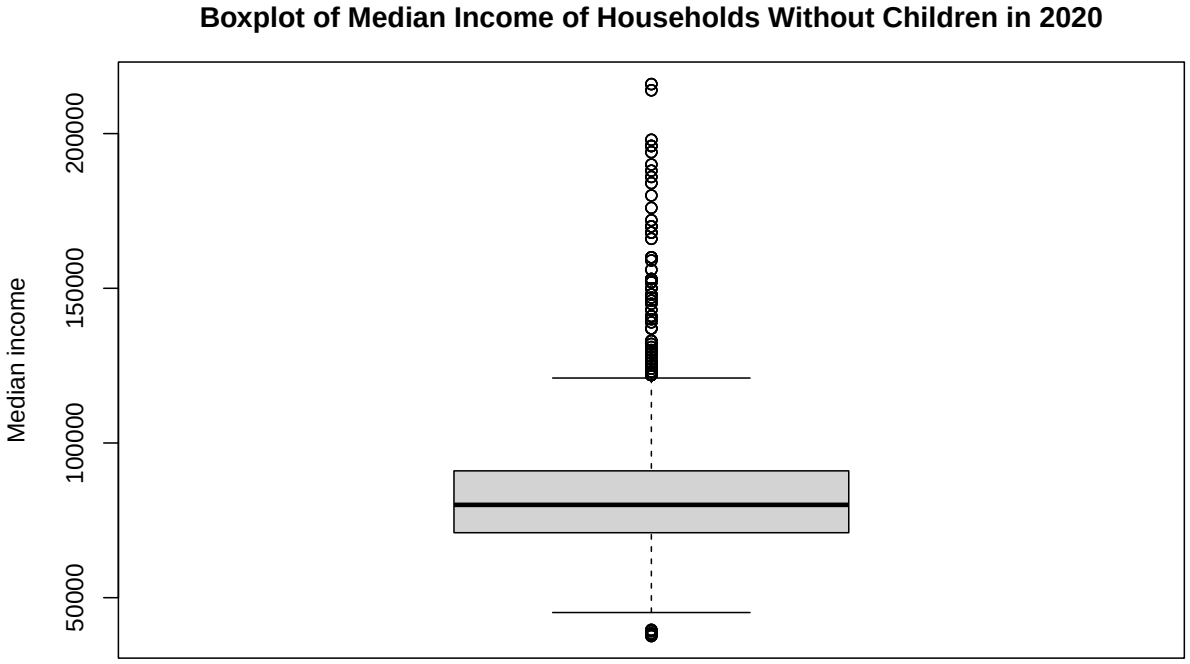
The p-value we got was less than 2.2e-16, which is extremely small, much smaller than our significance level of 5%. Thus, we reject H0, and conclude that the Median Income between Households with 4 persons and 5 or more persons is different.

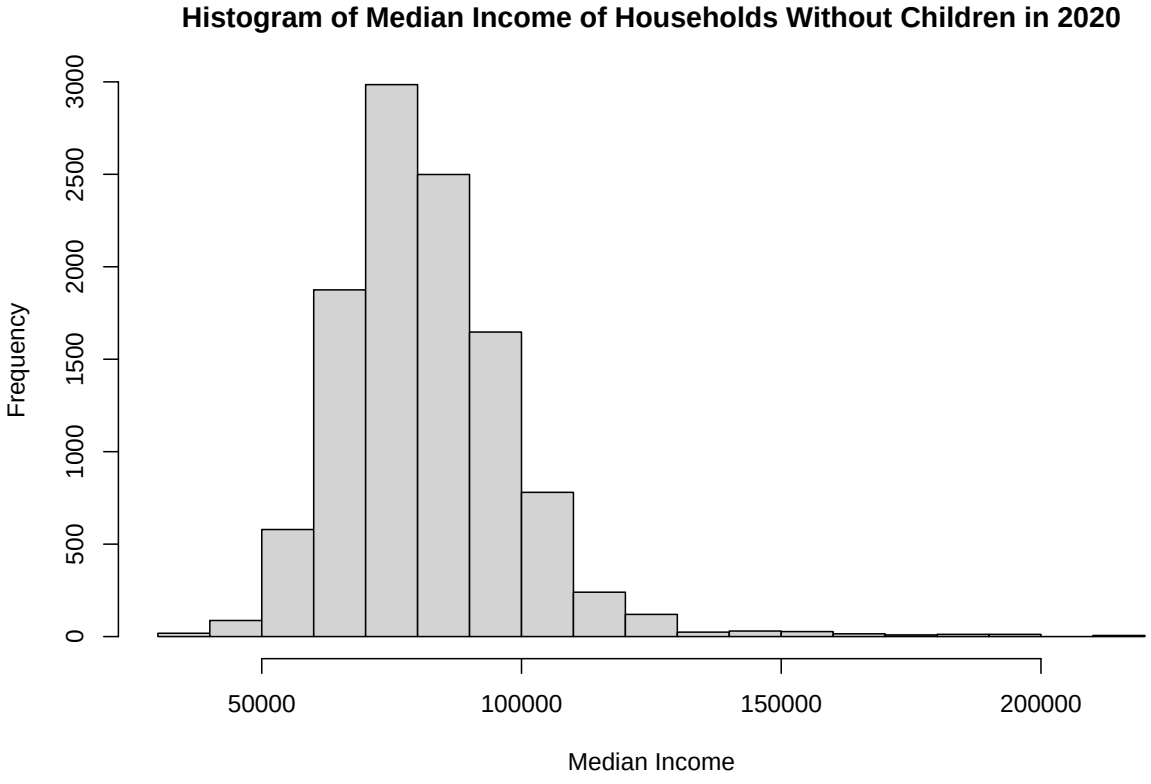
Additionally, the 95% confidence interval for the difference between Median Income of 4 person and 5 plus person households (inc4 - inc5p) is (-6,252.363, -4,967.192). This indicates that we can be 95% confident that households with 4 persons earn less than household with 5 or more persons.

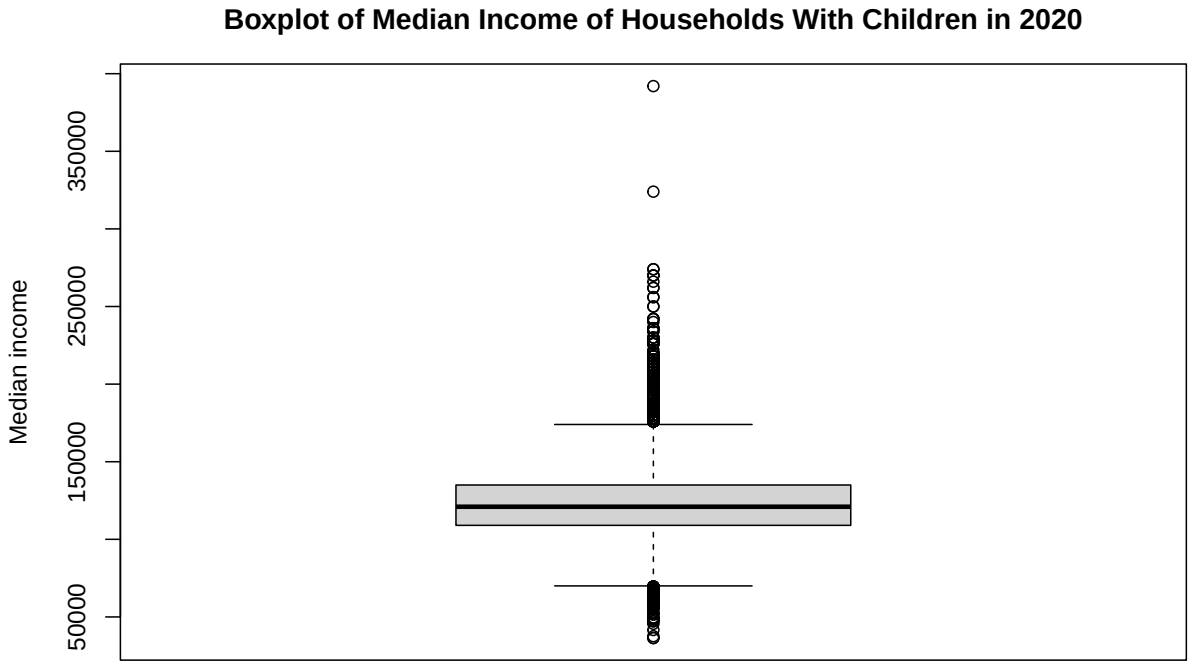
Question 3: Is there a difference in median income between Households with and without children?

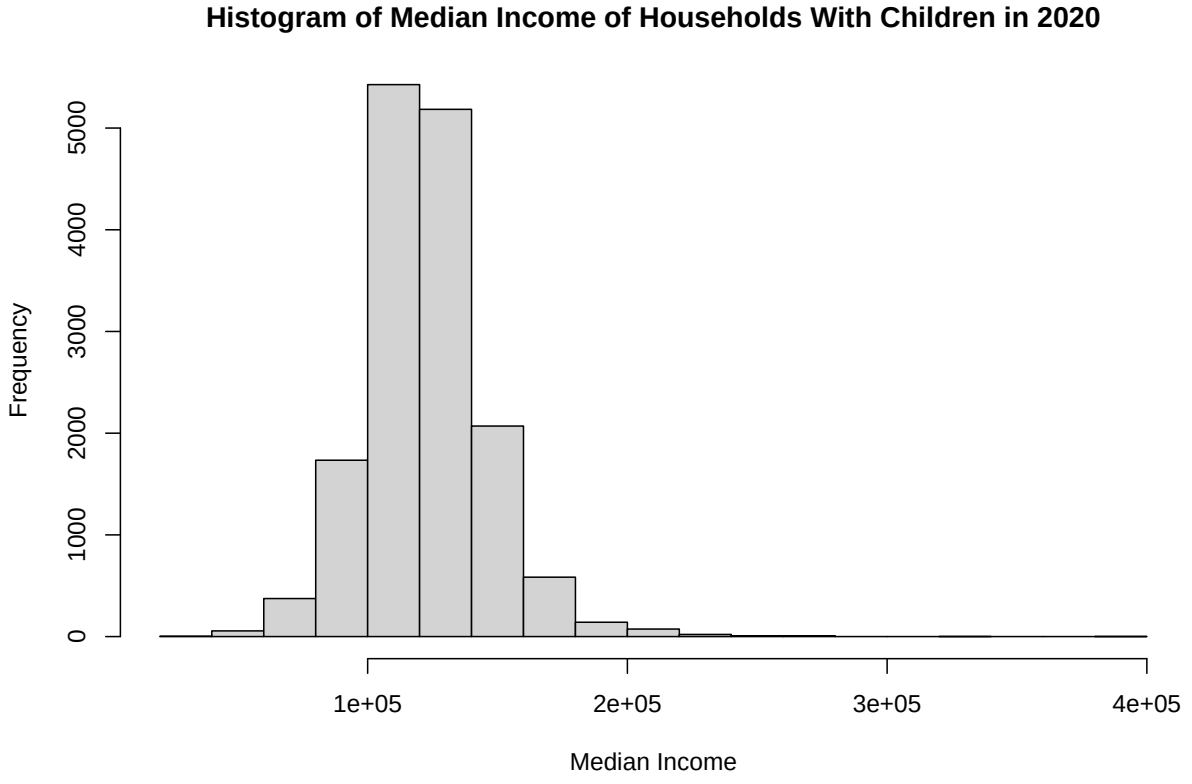
min	Q1	median	Q3	max	mean	sd	n	missing
37600	71000	80000	91000	216000	82239.62	17383.4	10965	0

min	Q1	median	Q3	max	mean	sd	n	missing
36400	109000	121000	135000	392000	122754.9	23401.19	15685	0


Boxplot of Median Income of Households Without Children in 2020
Median income
50000
100000
150000
200000


Histogram of Median Income of Households Without Children in 2020
Frequency
0
500
1000
1500
2000
2500
3000
50000
100000
150000
200000
Median Income


Boxplot of Median Income of Households With Children in 2020
Median income
50000
150000
250000
350000


Histogram of Median Income of Households With Children in 2020
Frequency
0
1000
2000
3000
4000
5000
1e+05
2e+05
3e+05
4e+05
Median Income

We performed a Welch's two sample t-test since the population variance is unknown. The Null Hypothesis (H0) is there is no difference in median income between households with and without children (incW0 = incW) and the Alternative Hypothesis (H1) is the median household incomes are different (incW0 != incW).

The p-value is less than 2.2-e16, which is smaller than our significance level at 5%. Thus, we can conclude that the Median Household Income between Families with children and without children is different.

The 95% confidence interval for the difference between the median income (incW0 - incW) is (-41,005.23, -40025.42). Since both confidence limits are negative, we can conclude with 95% confidence that the median income of families who have children are higher than those without.

This makes sense since households who have children are typically able to handle the financial burden that comes with raising a child.

Conclusion and Future Direction

Summary

1. **2015 vs 2020 after-tax household income**

   - The t-test shows a highly significant difference between 2015 and 2020 ($p < 2.2e\text{-}16$).
   - The 95% CI for (2020 - 2015) is (8441.755, 8813.442), entirely positive.
   - **Conclusion:** Median after-tax household income is higher in 2020 than in 2015.

2. **Household size: 4 persons vs 5+ persons (2020)**

   - The t-test again shows a highly significant difference ($p < 2.2e\text{-}16$).
   - The 95% CI for (4-person - 5+-person) income is (-6,252.363, -4,967.192), entirely negative.
   - **Conclusion:** Households with 5 or more people earn more on average than 4-person households.

3. **Households with vs without children**

   - The t-test finds a highly significant difference in income ($p < 2.2e\text{-}16$).
   - The 95% CI for (without children - with children) is (-41,005.23, -40025.42), all negative.
   - **Conclusion:** Households with children earn substantially more than households without children.

Limitations

1. One limitation is the dataset gives median incomes by region and household category, not individual households. Performing a t-test on median data would mean we are comparing the mean of medians, not the actual median. Hence, p-values and confidence intervals provide an approximation of the difference, not the true difference in population medians.

2. Another limitation is the dataset didn't have information for every single geographical location, which led to some rows either having 0 for Number of Households, and missing values for median incomes. One problem this could lead to is misrepresentation, since lower income regions could be missing, making my data more biased towards higher income areas

3. Another limiation is this dataset did not have information on key characteristics that affect earning like education, age, employment status, number of earners. These characteristics could potentially improve the precision of the analysis.

Possible extensions/improvements.

1. Get raw, individual data instead of grouped data. This will allow us to perform a much more accurate test, and the result will represent the actual mean better.

2. Unrealistic, but to get data from every region, and from every household. This will allow us to get accurate information on means and medians of each groups, which will provide an unskewed result that is representative of the whole population.

3. Get more characteristics from each sampled household, so we can better determine trends and correlations.

Appendix (R code)

Note that each question is done in a seperate R script, will be provided upon request.

Question 1: Is there a difference in median after tax household income between 2015 and 2020?

Load libraries

```
library(dplyr)
library(ggplot2)
```

Read data

```
income <- read.csv("incomes_cleaned.csv", check.names = FALSE)
```

filter by the first 11 rows of each geographic region that is not Canada

11 since these 11 rows have data of all household sizes

The difference between these rows are household types, e.g. Couple with children, without children, one parent, two parent, etc

```
income_sub <- income%>%
  filter(GEO != 'Canada') %>%
  filter(!is.na(`Household type including census family structure  (11)`))
```

exclude any rows where income is 0

```
income_sub <- income%>%
  filter(
    !is.na(`Median household after-tax income (2020)`),
    !is.na(`Median household after-tax income (2015)`),
    `Median household after-tax income (2020)` != 0,
    `Median household after-tax income (2015)` != 0
  )
```

getting the data from their columns

```
inc2020 <- income_sub$`Median household after-tax income (2020)`
inc2015 <- income_sub$`Median household after-tax income (2015)`
```

Summary Statistics & variances

```
var_median_2020  <- var(inc2020)
var_median_2015  <- var(inc2015)

summary(inc2020)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7850   62800   79500   81985  100000  300000
```

```
summary(inc2015)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     294   54400   72000   73357   91000  272000
```

```
var_median_2020
```

```
## [1] 725449693
```

```
var_median_2015
```

```
## [1] 678805957
```

Two sample t-test (unequal variances)

```
test <- t.test(inc2020, inc2015, alternative = "two.sided", var.equal = FALSE)
test
```

T-test since we do not know the population variance

```
##
##  Welch Two Sample t-test
##
## data:  inc2020 and inc2015
## t = 90.99, df = 312032, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  8441.755 8813.442
## sample estimates:
## mean of x mean of y
##  81984.76  73357.16
```

```
p_val <- test$p.value
p_val
```

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

Since P-value is smaller than 0.05, we reject H0 at 5% significance, and conclude that there is a difference in median after tax household income between 2020 and 2015.

Assumptions

Data is continuous:

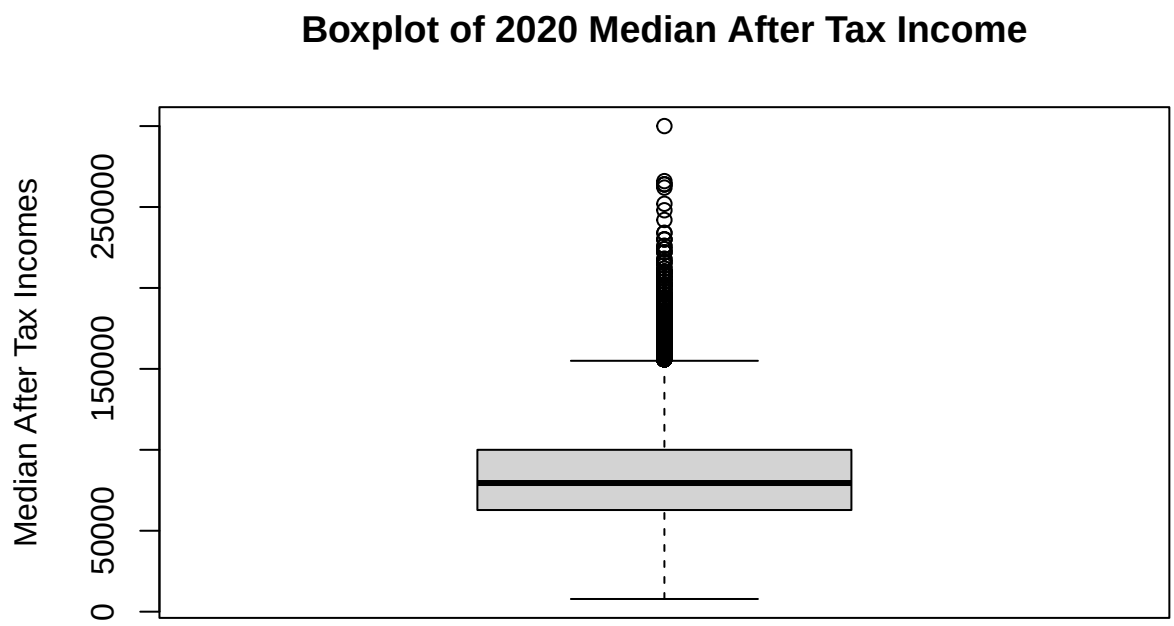
Data is independent:

This one is probably true since the data are from two different time periods

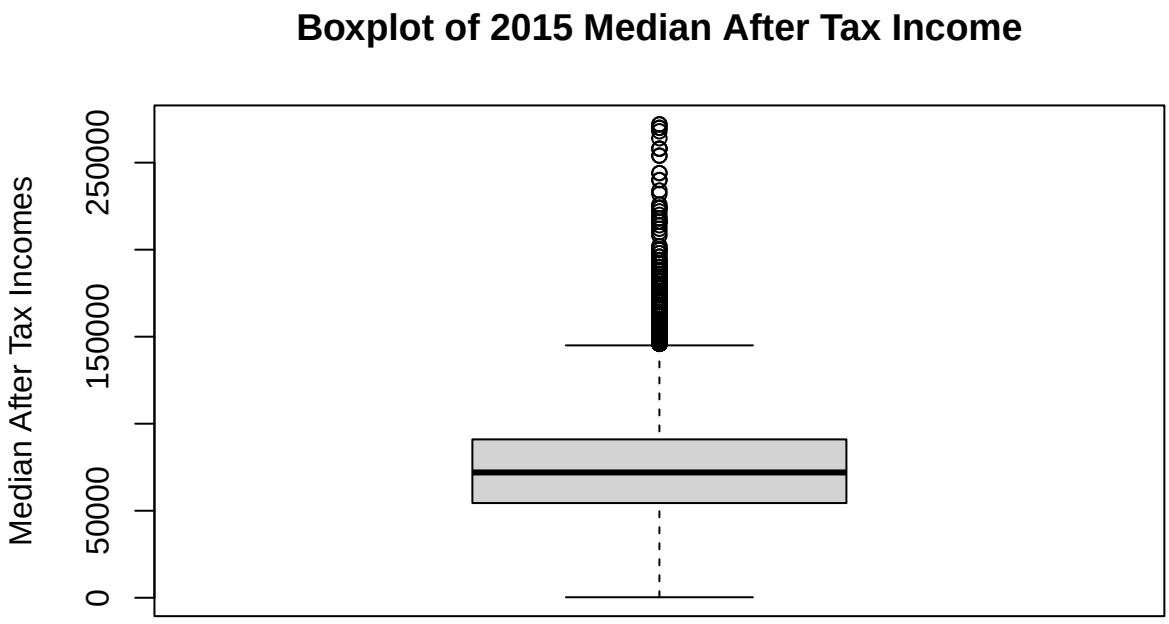
Distribution of data is approximately normal:

```
# Boxplots
box2020 = boxplot(inc2020, ylab = "Median After Tax Incomes", main = "Boxplot of 2020 Median After Tax I
```

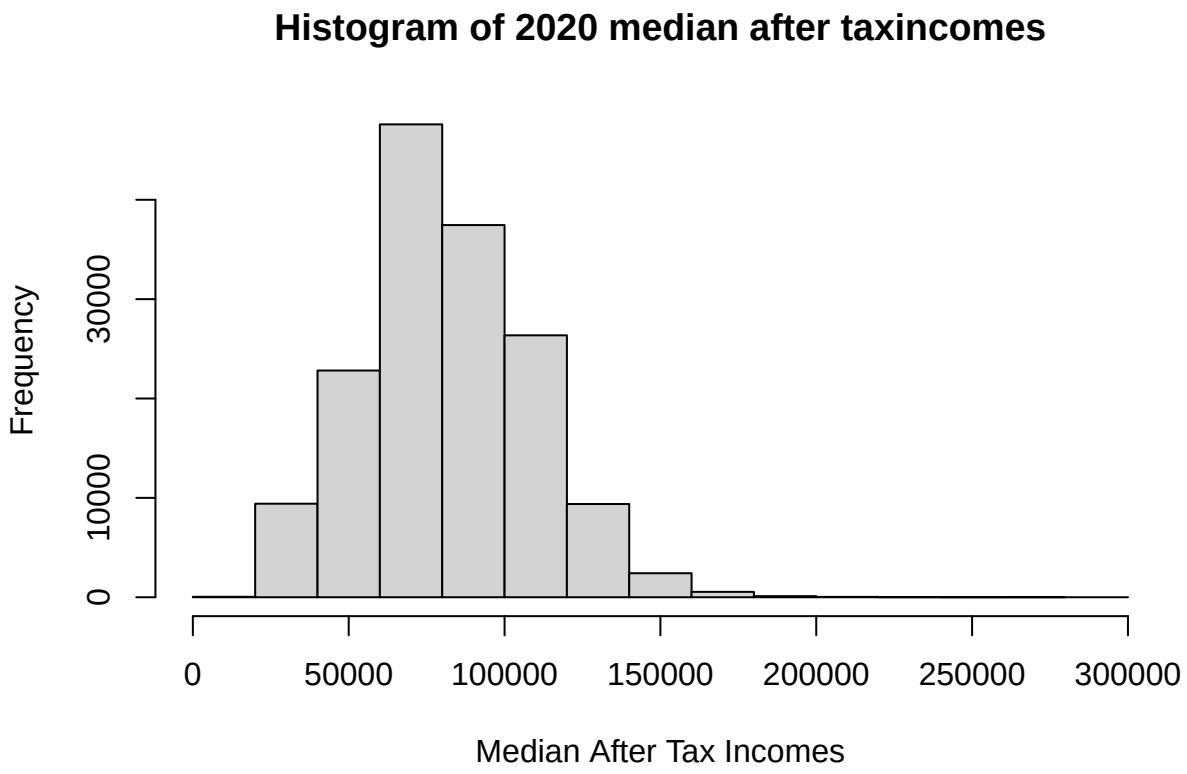
We use boxplot, histogram and QQ plot for this



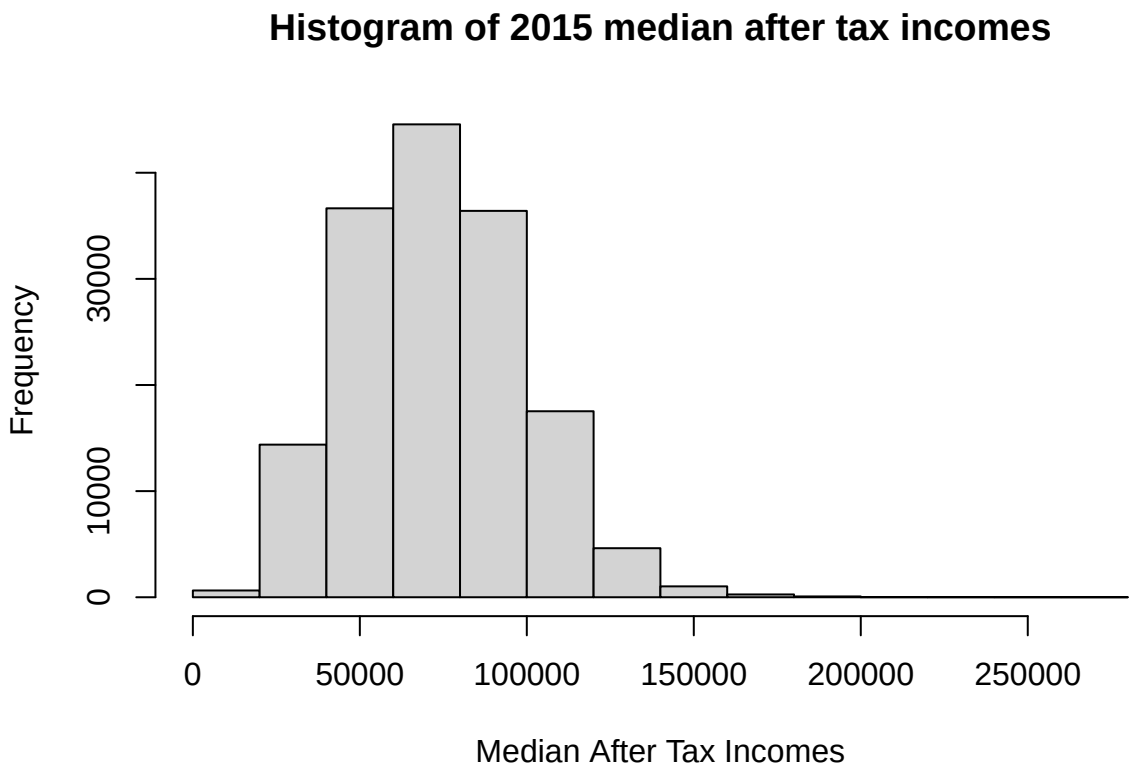

```
box2015 = boxplot(inc2015, ylab = "Median After Tax Incomes", main = "Boxplot of 2015 Median After Tax I
```



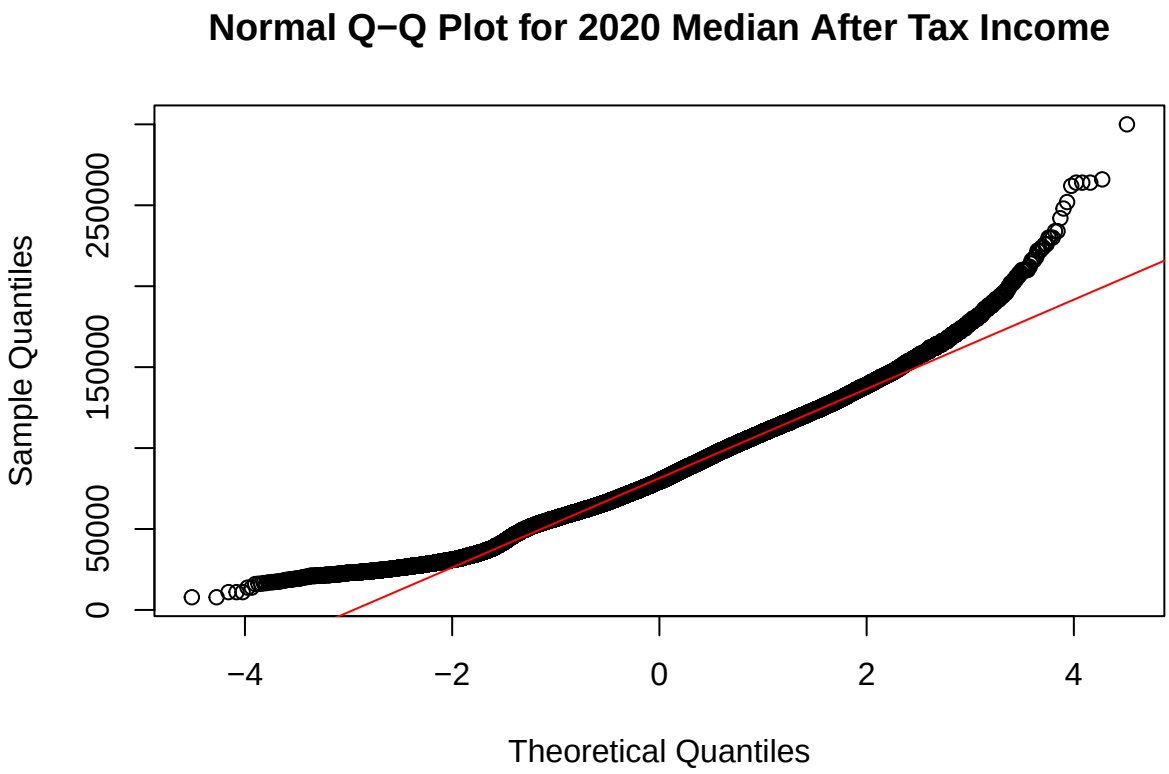

```
# Histograms
hist2020 = hist(inc2020, xlab = "Median After Tax Incomes", main = "Histogram of 2020 median after taxi
```



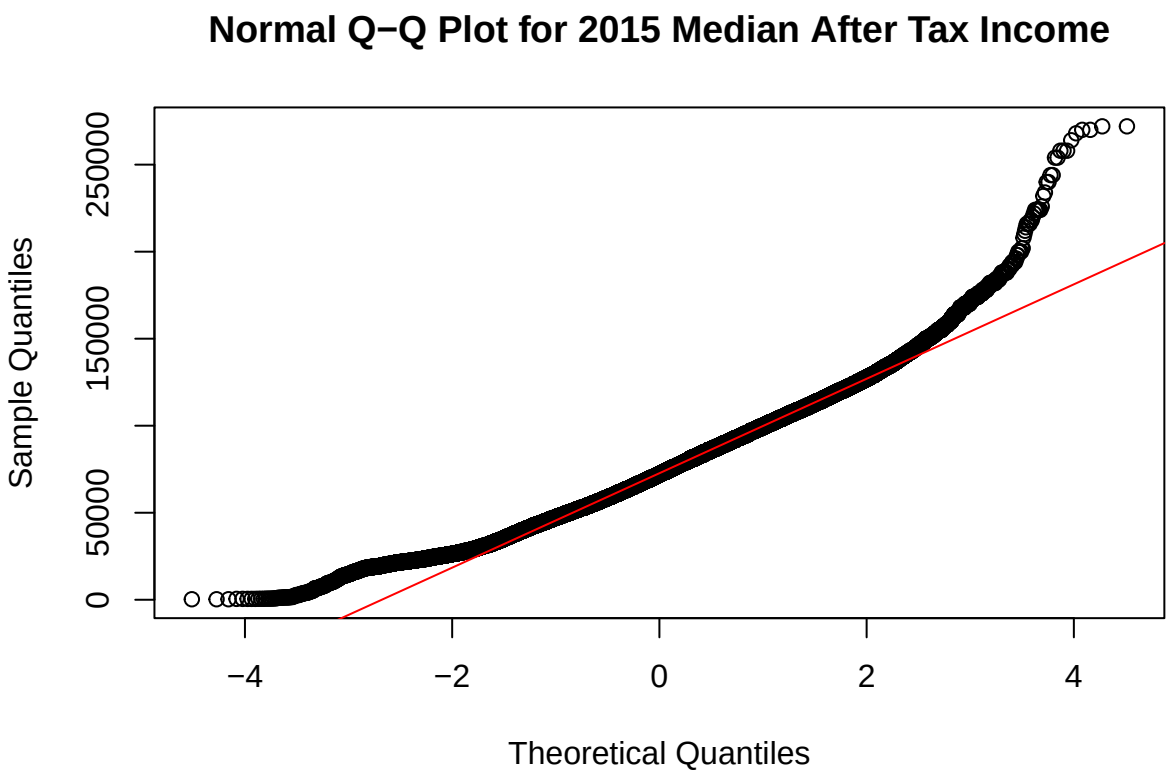

```
hist2015 = hist(inc2015, xlab = "Median After Tax Incomes", main = "Histogram of 2015 median after tax i
```




```r
# QQ Plots
qq2020 <- qqnorm(inc2020, main = "Normal Q-Q Plot for 2020 Median After Tax Income")
qqline2020 <- qqline(inc2020,col = "red")
```




```
qq2015 <- qqnorm(inc2015, main = "Normal Q-Q Plot for 2015 Median After Tax Income")
qqline2015 <- qqline(inc2015,col = "red")
```


Normal Q−Q Plot for 2015 Median After Tax Income
Sample Quantiles
0
50000
150000
250000
−4
−2
0
2
4
Theoretical Quantiles

Question 2: Is there a significant difference in income between households of 5 or more people and 4 persons (in 2020)?

Load Libraries

```
library(dplyr)
library(ggplot2)
```

Getting data from file

```
income <- read.csv("incomes_cleaned.csv", check.names = FALSE)
```

Data from relevant rows only (rows where household size is '4 persons' and '5 or more persons')

```
income_sub <- income %>%
  filter(GEO != 'Canada') %>%
  filter(
    `Household size (7)` %in% c("4 persons", "5 or more persons"),
    !is.na(`Median household total income (2020)`),
    `Median household total income (2020)` != 0
  )

# Attaching 4 persons household rows
inc4 <- income_sub %>%
  filter(`Household size (7)` == "4 persons") %>%
  pull(`Median household total income (2020)`)

# Attaching 5 or more persons household rows
inc5p <- income_sub%>%
  filter(`Household size (7)` == "5 or more persons") %>%
  pull(`Median household total income (2020)`)
```

Summary statistics, variances

```
var4 <- var(inc4)
var5p <- var(inc5p)

summary(inc4)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26000  108000  124000  122604  139000  274000
```

```
summary(inc5p)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34000  111000  128000  128214  145000  428000
```

```
var4
```

```
## [1] 807195987
```

```
var5p
```

```
## [1] 948878690
```

Two Sample t-test (unequal variances)

T-test since we do not know the population variance

```
test <- t.test(inc4, inc5p, alternative = "two.sided", var.equal = FALSE)
test
```

```
##
##  Welch Two Sample t-test
##
## data:  inc4 and inc5p
## t = -17.111, df = 31036, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6252.363 -4967.192
## sample estimates:
## mean of x mean of y
##  122603.8  128213.6
```

```
p_val <- test$p.value
p_val
```

```
## [1] 2.444757e-65
```

Since the p-value is less than 0.05, we reject H0, and conclude that there is a difference in income between household of 4 persons and households of 5 or more persons

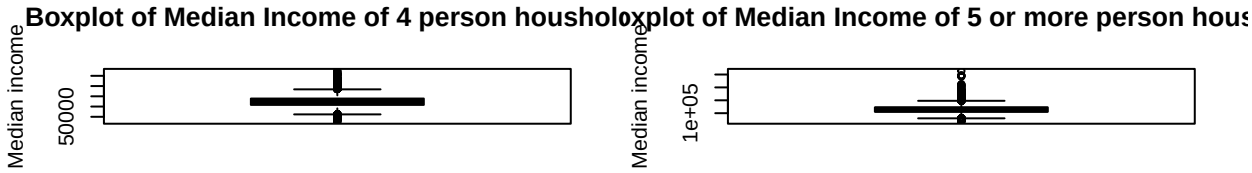
Assumptions

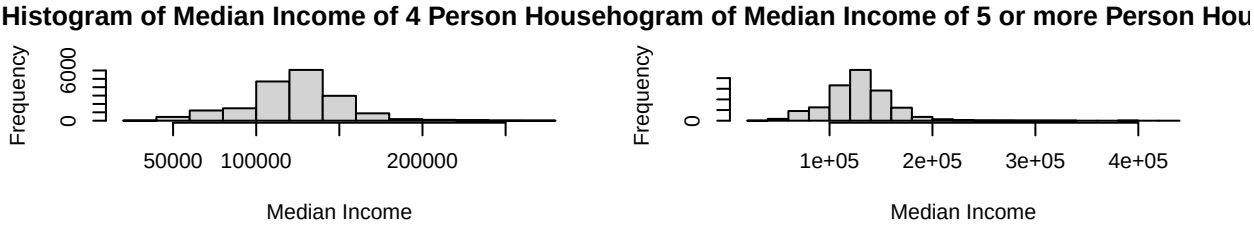
Data is continuous:

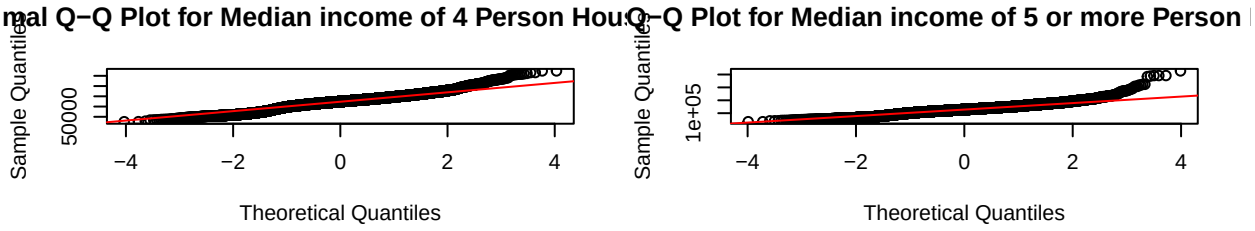
Data is independent: Probably true since the two groups are very distinct (4 and 5 or more persons household)

Data is normally distributed: We use boxplot, histogram, and qq plot

```r
par(mfrow=c(3,2))

# Boxplots
box4 = boxplot(inc4, ylab = "Median income", main = "Boxplot of Median Income of 4 person housholds")
box5p = boxplot(inc5p, ylab = "Median income", main = "Boxplot of Median Income of 5 or more person hous

# Histograms
hist4 = hist(inc4, xlab = "Median Income", main = "Histogram of Median Income of 4 Person Households")
hist5p = hist(inc5p, xlab = "Median Income", main = "Histogram of Median Income of 5 or more Person Hous

# QQ plots
qq4 <- qqnorm(inc4, main = "Normal Q-Q Plot for Median income of 4 Person Households")
qqline4 <- qqline(inc4,col = "red")

qq5p <- qqnorm(inc5p, main = "Normal Q-Q Plot for Median income of 5 or more Person Households")
qqline5p <- qqline(inc5p, col = "red")
```







Question 3: Is there a difference in median income between Households with and without children?

Do households with children earn more than households without children?

Load libraries

```
library(dplyr)
library(ggplot2)
```

Getting data

```
income <- read.csv("incomes_cleaned.csv", check.names = FALSE)
```

Filtering data

```
# We check data by the Household type... column
income_sub <- income %>%
  filter(GEO != 'Canada') %>%
  filter(
    # Include the data if the row is 'Without children' or 'With children'
    `Household type including census family structure  (11)` %in% c("Without children", "With children")
    !is.na(`Median household total income (2020)`),
    `Median household total income (2020)` != 0
  )

# Attaching the data
without_children <- income_sub %>%
  filter(`Household type including census family structure  (11)` == "Without children") %>%
  pull(`Median household total income (2020)`)

with_children <- income_sub %>%
  filter(`Household type including census family structure  (11)` == "With children") %>%
  pull(`Median household total income (2020)`)
```

Summary statistics & variances

```
var_without <- var(without_children)
var_with <- var(with_children)

summary(without_children)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   37600   71000   80000   82240   91000  216000
```

```r
summary(with_children)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   36400  109000  121000  122755  135000  392000
```

```r
var_without
```

```
## [1] 302182460
```

```r
var_with
```

```
## [1] 547615912
```

Two Sample t-test (unequal variances)

```r
test <- t.test(without_children, with_children, alternative = "two.sided", var.equal = FALSE)
test
```

T-test since we do not know the population variance

```
##
##  Welch Two Sample t-test
##
## data:  without_children and with_children
## t = -162.1, df = 26551, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -41005.23 -40025.42
## sample estimates:
## mean of x mean of y
##  82239.62 122754.94
```

```r
p_val <- test$p.value
p_val
```

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

Since the p-value of our t-test is less than 0.05, we reject H0, and conclude that there is a difference between the income of households with children and without children

Assumptions

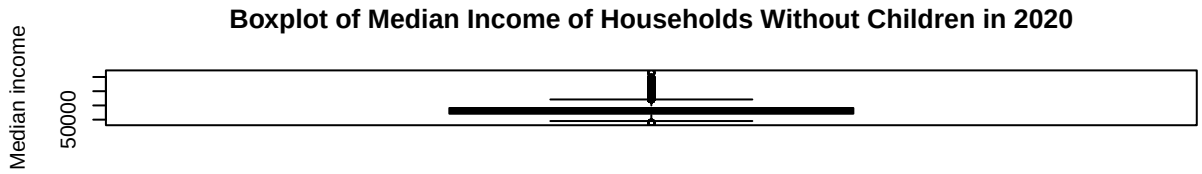
Data is continuous:

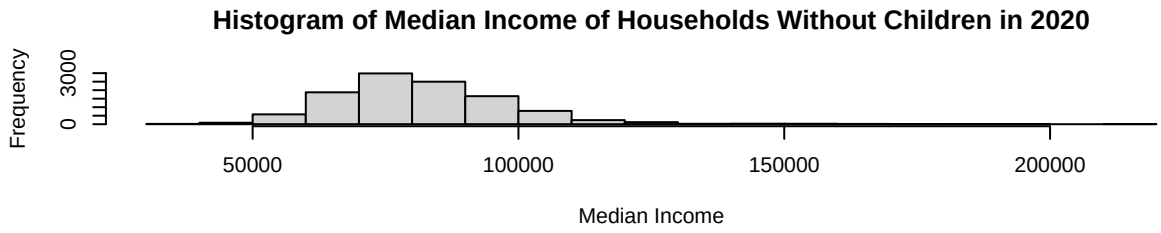
Data is independent: This is probably true since people with children wouldn't report that they don't have children when replying to a census, and vice versa

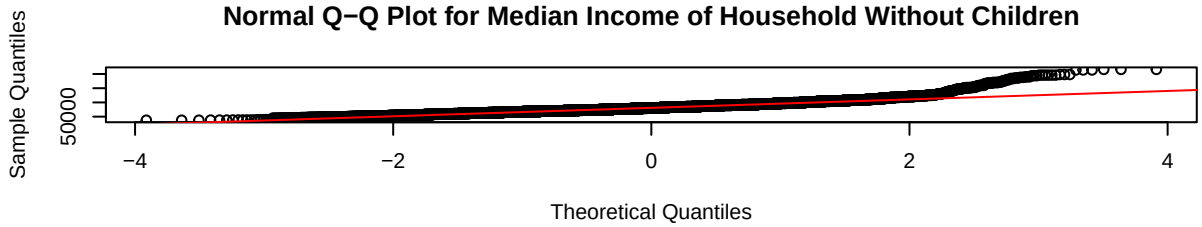
Data is Normally distributed: We will use boxplots, histograms, and qq-plots to check this assumption

Families without children

```
par(mfrow=c(3,1))

# Boxplot
withoutbox <- boxplot(without_children, ylab = "Median income", main = "Boxplot of Median Income of Hous

# Histogram
withoutHist <- hist(without_children, xlab = "Median Income", main = "Histogram of Median Income of Hous

# QQ plot
qqwithout <- qqnorm(without_children, main = "Normal Q-Q Plot for Median Income of Household Without Chi
qqlinewithout <- qqline(without_children, col = "red")
```







Families With Children

```
par(mfrow=c(3,1))
# Boxplot
withbox <- boxplot(with_children, ylab = "Median income", main = "Boxplot of Median Income of Household

# Histogram
withHist <- hist(with_children, xlab = "Median Income", main = "Histogram of Median Income of Household

# QQ plot
qqwith <- qqnorm(with_children, main = "Normal Q-Q Plot for Median Income of Household Without Children
qqlinewith <- qqline(with_children, col = "red")
```