Benefits of Union Membership Final Report

AUTHOR
Raunak Advani, Austin Barish, Sai Prerana

1 Introduction

The landscape of the American workforce has been shaped by numerous factors, among which labor unions have played a pivotal role. Historically, unions have been a cornerstone in advocating for workers' rights, influencing policies that govern wages, work hours, and working conditions. As students soon to be entering the workforce, we are particularly interested in understanding what we should look for in employers. In the context of this project, whether joining a unionized company would be beneficial to us. Beyond that, understanding this issue is key to being an informed adult as it is an issue that shapes the world around us and many of the people we will come into contact with.

Unionism in the United States boasts a rich history, originating from the industrial revolution's demands for fair labor practices. These organizations have been instrumental in creating and maintaining many of the labor rights we now consider standard, such as the eight-hour workday, minimum wage laws, and occupational safety standards. However, according to data from the Bureau of Labor Statistics (BLS) union membership has declined from over 20% in 1983 to 10% in 2022.

In addition to raising several questions relating to union membership, this statistic serves as a reflection of evolving work environments and the changing needs of workforce. In a labor market than now encompasses a wide array of industries and demographics, the question of whether unions still pay a critical role in advocating for workers' benefits becomes extremely relevant. This is especially pertinent given the growing diversity in the workforce, with varying implications for different genders, races, and ethnic origins.

Thus, this topic extends beyond the realm of labor economics; it touches upon deeper societal themes such as equity, fair treatment, and the quality of working life. The impact of union membership, or lack thereof, can be seen in the day-to-day life of workers, influencing their financial stability, job satisfaction, and overall well-being. In exploring these dimensions, we delve into not just the quantitative aspect of union membership, but also the qualitative experiences of those within and outside unionized environments.

The benefits (or drawbacks) of union membership is an extremely broad topic. Benefits could include things such as mental health or social benefits, neither of which we will touch upon here. To narrow down our analysis, we will be focusing on a Guardian article titled "How much does union membership benefit America's workers?". While an interesting article, it uses very overdone graphs and no statistical tests. We want to dive deeper into its claims to created more detailed graphs, analyses, and statistically verified conclusions.

This project aims to unravel these layers, offering insights into the significance of unions in the American workforce. By examining the intersection of union membership with factors like gender, race, and ethnicity, we seek to understand who benefits from unions in today's economy and in what ways. This exploration is ultimately a journey into understanding how the legacy of unions continues to shape the fabric of the American workplace.

2 Analysis

Dataset

Our data is from Data World and contains detailed information on “usual” weekly earnings in various industries and occupations in the United States. *Usual* is self-defined as this is all survey data; we can assume it to mean that it is the weekly earnings in the average 5 day work week. This dataset is instrumental for assessing the impact of unionization on wage levels across different sectors and demographic groups.

Research Questions

Through our dataset, the primary area we’d like to explore is **“How much does union membership benefit America’s workers?”;** the same as the Guardian article**.** With it, we seek to examine as many relationships to the benefits, measured by Median Weekly Earnings, as the data allows us to. We can then dispel any lingering questions about their analysis. This creates a series of sub-questions beyond our initial analysis:

- Does union membership benefit American workers and, if so, is that benefit equal across workers across different genders, races, and ethnic origins?
    - Does union membership disproportionately help Men vs Women?
        - If so, is this difference a product of union membership or structural inequalities that exist across union status?
    - Does union membership disproportionately help members of certain races?
        - If so, is this difference a product of union membership or structural inequalities that exist across union status
    - Does union membership disproportionately help members of certain ethnic origins?
        - If so, is this difference a product of union membership or structural inequalities that exist across union status?
    - Across these groups, who benefits most from joining a union, if at all?
        - Is there any group for whom union membership is disadvantageous?

We will see that while we are able to answer the majority of these questions, some of them are harder to answer due to the limited breakdowns within the data. Without any population sizes, we cannot further divide the data as will be explained in greater detail later.

```r
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.0     ✔ readr     2.1.4
✔ forcats   1.0.0     ✔ stringr   1.5.0
✔ ggplot2   3.4.2     ✔ tibble    3.1.8
✔ lubridate 1.9.2     ✔ tidyr     1.3.0
✔ purrr     1.0.1
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```r
library(ggplot2)

# Read Data
df <- read.csv('./union-dataset.csv')

# Display
head(df)
```

```
  Year Median.usual.weekly.earnings Union       Industry      Occupation
1 2000                          576   All All Industries All Occupations
2 2001                          596   All All Industries All Occupations
3 2002                          608   All All Industries All Occupations
4 2003                          620   All All Industries All Occupations
5 2004                          638   All All Industries All Occupations
6 2005                          651   All All Industries All Occupations
         Sex      Race Ethnic.Origin               Age
1 Both Sexes All Races   All Origins 16 years and over
2 Both Sexes All Races   All Origins 16 years and over
3 Both Sexes All Races   All Origins 16 years and over
4 Both Sexes All Races   All Origins 16 years and over
5 Both Sexes All Races   All Origins 16 years and over
6 Both Sexes All Races   All Origins 16 years and over
```

Data Checks and Cleaning

```r
cat("The shape of the dataframe is: ", length(df$Year), "x", length(names(df)))
```

```
The shape of the dataframe is:  532 x 9
```

```r
summary(df)
```

```
      Year      Median.usual.weekly.earnings    Union          
 Min.   :2000   Min.   : 381.0              Length:532        
 1st Qu.:2004   1st Qu.: 645.0              Class :character  
 Median :2009   Median : 763.0              Mode  :character  
 Mean   :2009   Mean   : 764.1                                
 3rd Qu.:2014   3rd Qu.: 884.2                                
 Max.   :2018   Max.   :1126.0                                
   Industry          Occupation            Sex                Race          
 Length:532         Length:532         Length:532         Length:532        
 Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Ethnic.Origin          Age           
 Length:532         Length:532        
 Class :character   Class :character  
 Mode  :character   Mode  :character  
```

1. Missing Values

```
cat("THERE ARE", sum(is.na(df)),  "MISSING OR NA VALUES.")
```

```
THERE ARE 0 MISSING OR NA VALUES.
```

```
cat("THERE ARE", sum(duplicated(df)), "DUPLICATE ROWS")
```

```
THERE ARE 0 DUPLICATE ROWS
```

There are no missing or incorrect values that we need to remove.

2. Incorrect Values

There is no reason to assume any of the values are incorrect.

3. Values with Improper Formatting

```
str(df)
```

```
'data.frame':	532 obs. of  9 variables:
 $ Year                         : int  2000 2001 2002 2003 2004 2005 2006 2007 2008
2009 ...
 $ Median.usual.weekly.earnings: int  576 596 608 620 638 651 671 695 722 739 ...
 $ Union                        : chr  "All" "All" "All" "All" ...
 $ Industry                     : chr  "All Industries" "All Industries" "All
Industries" "All Industries" ...
 $ Occupation                   : chr  "All Occupations" "All Occupations" "All
Occupations" "All Occupations" ...
 $ Sex                          : chr  "Both Sexes" "Both Sexes" "Both Sexes" "Both
Sexes" ...
 $ Race                         : chr  "All Races" "All Races" "All Races" "All Races"
...
 $ Ethnic.Origin                : chr  "All Origins" "All Origins" "All Origins" "All
Origins" ...
 $ Age                          : chr  "16 years and over" "16 years and over" "16
years and over" "16 years and over" ...
```

For now, all of the dates are fine as int, but we can also add a Date formatted column to have.

```
library(lubridate)
df$Date <- lubridate::ymd(df$Year, truncated = 2L)
```

```
str(df)
```

```
'data.frame':	532 obs. of  10 variables:
 $ Year                       : int  2000 2001 2002 2003 2004 2005 2006 2007 2008
2009 ...
 $ Median.usual.weekly.earnings: int  576 596 608 620 638 651 671 695 722 739 ...
 $ Union                      : chr  "All" "All" "All" "All" ...
 $ Industry                   : chr  "All Industries" "All Industries" "All
Industries" "All Industries" ...
 $ Occupation                 : chr  "All Occupations" "All Occupations" "All
Occupations" "All Occupations" ...
 $ Sex                        : chr  "Both Sexes" "Both Sexes" "Both Sexes" "Both
Sexes" ...
 $ Race                       : chr  "All Races" "All Races" "All Races" "All Races"
...
 $ Ethnic.Origin              : chr  "All Origins" "All Origins" "All Origins" "All
Origins" ...
 $ Age                        : chr  "16 years and over" "16 years and over" "16
years and over" "16 years and over" ...
 $ Date                       : Date, format: "2000-01-01" "2001-01-01" ...
```

4. General Data Consistency

All of the data is consistent, however, it is worth noting that it is difficult to measure cross categorical variables such as “Black Women” or “White Men” as the categories only split one by one. As we can see:

```
df %>% filter(Race == "White") %>% select("Sex") %>% unique()
```

```
         Sex
1 Both Sexes
```

```
df %>% filter(Sex == "Women") %>% select("Race") %>% unique()
```

```
       Race
1 All Races
```

```
df %>% filter(Union == "Non-union") %>% select("Sex", "Race") %>% unique()
```

```
          Sex                      Race
1  Both Sexes                 All Races
20        Men                 All Races
39      Women                 All Races
58 Both Sexes                     White
77 Both Sexes Black or African American
96 Both Sexes                     Asian
```

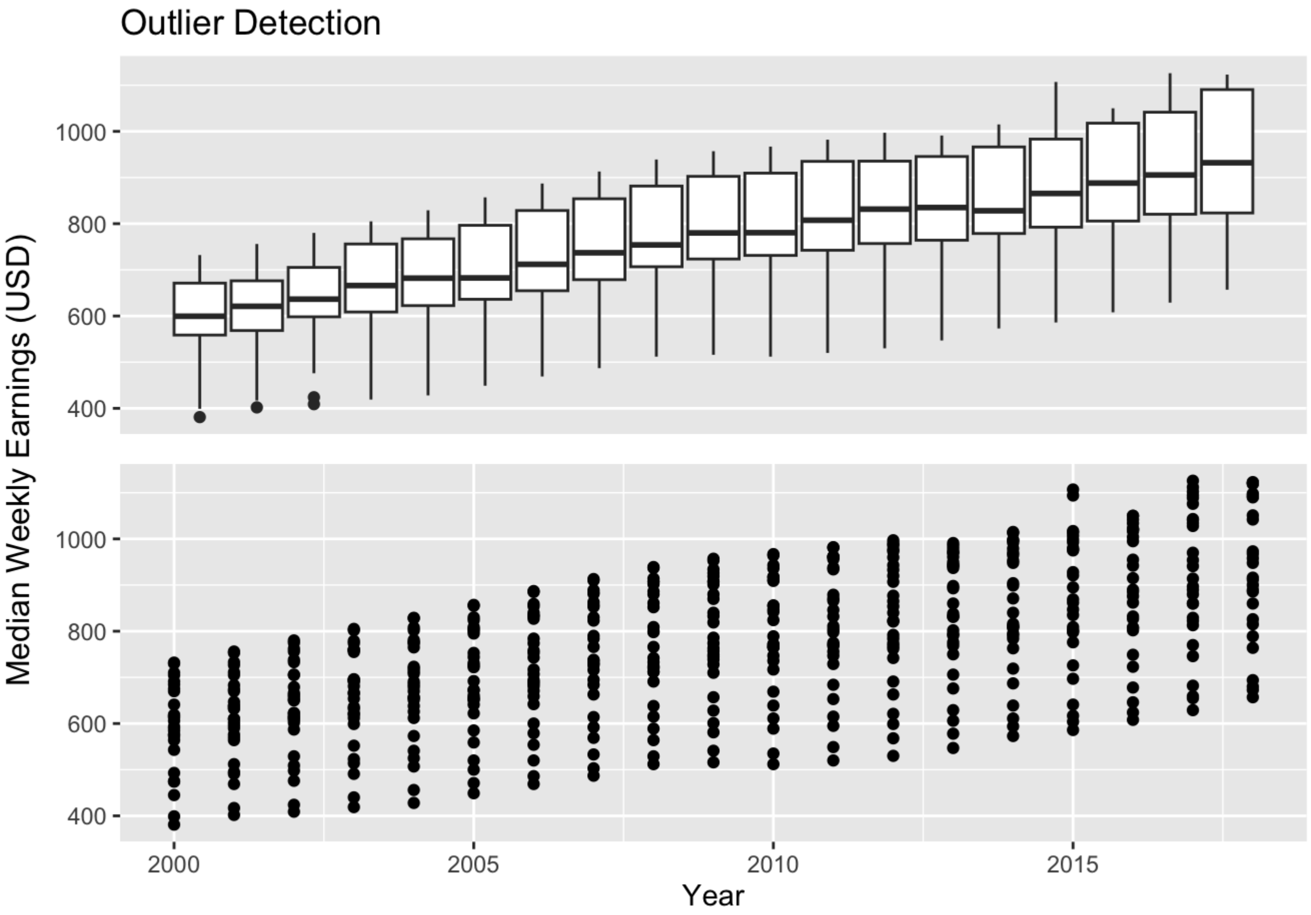
There is no intersection between the breakdowns of Sex, Race, or Ethnic Origin. Without population sizes, we cannot further divide the data as we do not know how many black women were included in the sample, as an example. Fortunately, we can still view these all by specific union status, allowing us to answer the majority of our questions.

5. Data Formatting

All of the data is in appropriate formats.

6. Outliers

```
library(ggplot2)
library(patchwork)

# Create Scatter Plot
scatter_plot <- ggplot(df, aes(x = Year, y = Median.usual.weekly.earnings)) +
  geom_point() +
  labs(y = "Median Weekly Earnings (USD)")

# Create Box Plot
box_plot <- ggplot(df, aes(group = Year, y = Median.usual.weekly.earnings)) +
  geom_boxplot() +
  labs( title = "Outlier Detection") + scale_x_continuous(breaks=unique(df$Year))

scatter_plot$labels$y <- box_plot$labels$y <- " "

# Combine
combined_plot <-  box_plot / scatter_plot +
  plot_layout(guides = "collect", heights = c(3,3))

combined_plot
# Help from https://stackoverflow.com/questions/65291723/merging-two-y-axes-titles-in-
grid::grid.draw(grid::textGrob("Median Weekly Earnings (USD)", x = 0.02, rot = 90))
```



None of the data appears to be an obvious outlier. Perhaps in 2000 or 2015 there are two outliers but they do not appear to be distant enough to be removed. We will later see that these points seem consistent with trends across the data.

7. Normalization

Since there is a clear upward trend in the data due to inflation, we can normalize using the CPI provided by FRED through R’s Quantmod library. Our data is only annual weekly earnings therefore we have to pick a specific month to chain our CPI to; we have chosen the first CPI of the year (January). Any other month would work just as well as an annual average. Additionally, any year could work to chain the dollar to but we have chosen the final year of the dataset (2018) for simplicity and most similarity to the current dollar. As an example:

$$\text{Adjusted 2002 Dollars} = \text{2002 Dollars} \times \frac{CPI_{2018}}{CPI_{2002}}$$

The second term here is the “inflation factor”, which varies year by year.

```r
# install.packages("quantmod")
library(quantmod)
```

```
Loading required package: xts

Loading required package: zoo


Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric


######################### Warning from 'xts' package ##########################
#                                                                             #
# The dplyr lag() function breaks how base R's lag() function is supposed to  #
# work, which breaks lag(my_xts). Calls to lag(my_xts) that you type or       #
# source() into this session won't work correctly.                            #
#                                                                             #
# Use stats::lag() to make sure you're not using dplyr::lag(), or you can add #
# conflictRules('dplyr', exclude = 'lag') to your .Rprofile to stop           #
# dplyr from breaking base R's lag() function.                                #
#                                                                             #
# Code in packages is not affected. It's protected by R's namespace mechanism #
# Set `options(xts.warn_dplyr_breaks_lag = FALSE)` to suppress this warning.  #
#                                                                             #
###############################################################################


Attaching package: 'xts'

The following objects are masked from 'package:dplyr':

    first, last
```

```
Loading required package: TTR

Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo
```

```
library(dplyr)

# Download CPI data
getSymbols("CPIAUCNS", src = "FRED", return.class="data.frame")
```

```
[1] "CPIAUCNS"
```

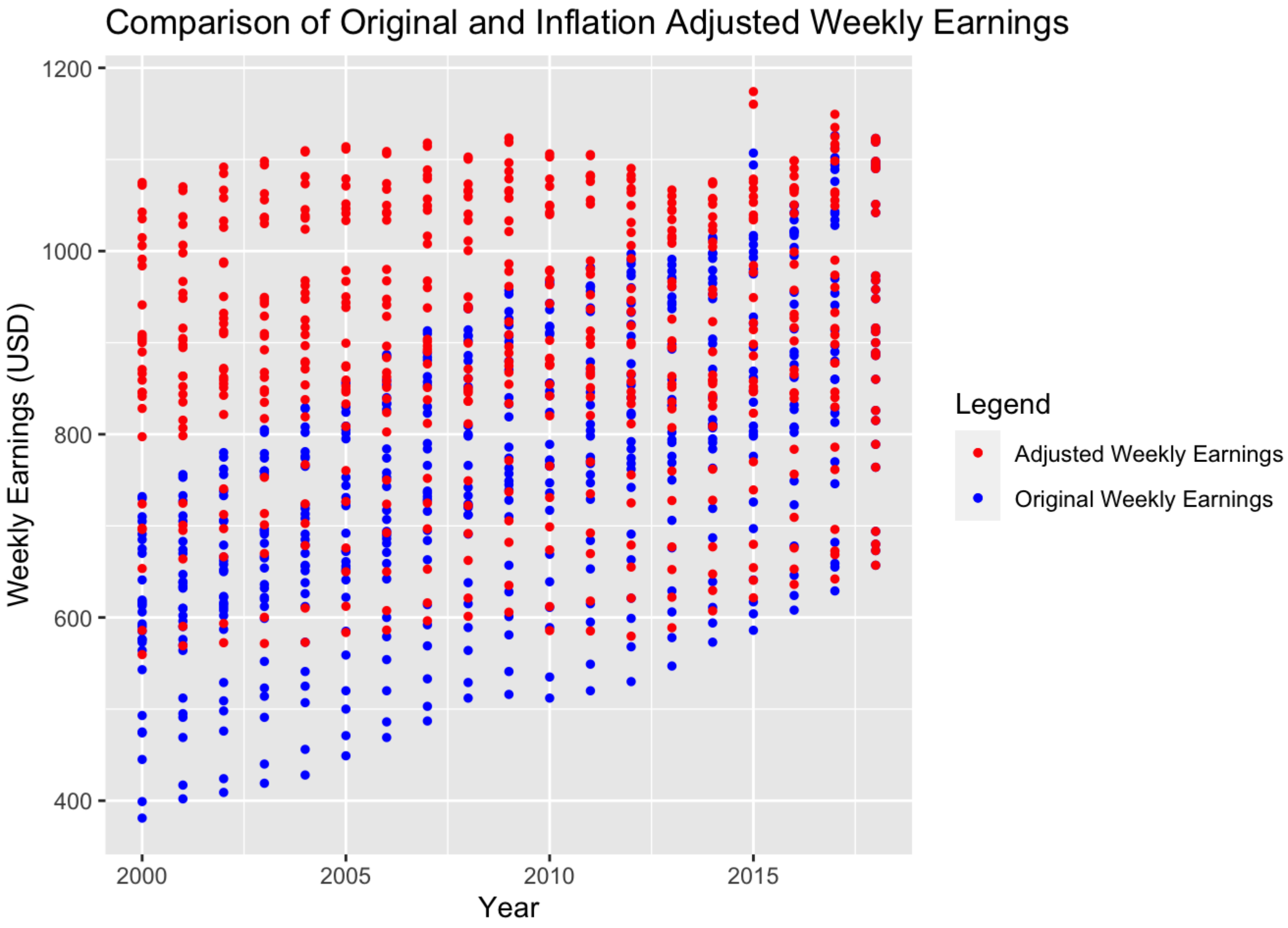
```
# Get dates as a column
CPIAUCNS <- rownames_to_column(CPIAUCNS, var = "Date")

# Use the January CPI as the annual CPI, could also use the average
CPIAUCNS <- CPIAUCNS %>%
  filter(month(Date) == "1") %>%
  mutate(Year = as.integer(year(Date))) %>% select("Year", "CPIAUCNS")

# Merge on the Year column
df <- merge(df, CPIAUCNS, by = "Year")

# Find inflation as a factor of the oldest year in the dataset
inflation_factor <- df$CPIAUCNS / df$CPIAUCNS[which.max(df$Year)][1]

# Find adjusted weekly earnings
df$Adjusted.Weekly.Earnings <- df$Median.usual.weekly.earnings / inflation_factor

# Display
df %>%
  select(Year, Median.usual.weekly.earnings, Adjusted.Weekly.Earnings) %>%
  group_by(Year) %>%
  summarise(
    Median_Weekly_Earnings = median(Median.usual.weekly.earnings, na.rm = TRUE),
    Adjusted_Median_Weekly_Earnings = median(Adjusted.Weekly.Earnings, na.rm = TRUE)
  )
```

```
# A tibble: 19 × 3
    Year Median_Weekly_Earnings Adjusted_Median_Weekly_Earnings
   <int>                  <dbl>                           <dbl>
 1  2000                   600.                            880.
 2  2001                   621                             879.
 3  2002                   636.                            891.
 4  2003                   666                             909.
 5  2004                   682                             913.
 6  2005                   682.                            887.
 7  2006                   712                             890.
 8  2007                   737                             902.
 9  2008                   754                             885.
10  2009                   780                             916.
11  2010                   780.                            893.
```

```
12  2011                      808.                             909.
13  2012                      832.                             909.
14  2013                      835                              899.
15  2014                      828                              877.
16  2015                      866.                             918.
17  2016                      888                              929.
18  2017                      906.                             924.
19  2018                      932                              932
```

As hoped, it is now all in 2018 dollars. As a sanity check, we see that in 2018, the adjusted earnings are the same while the years are all similar high or low but just adjusted now.

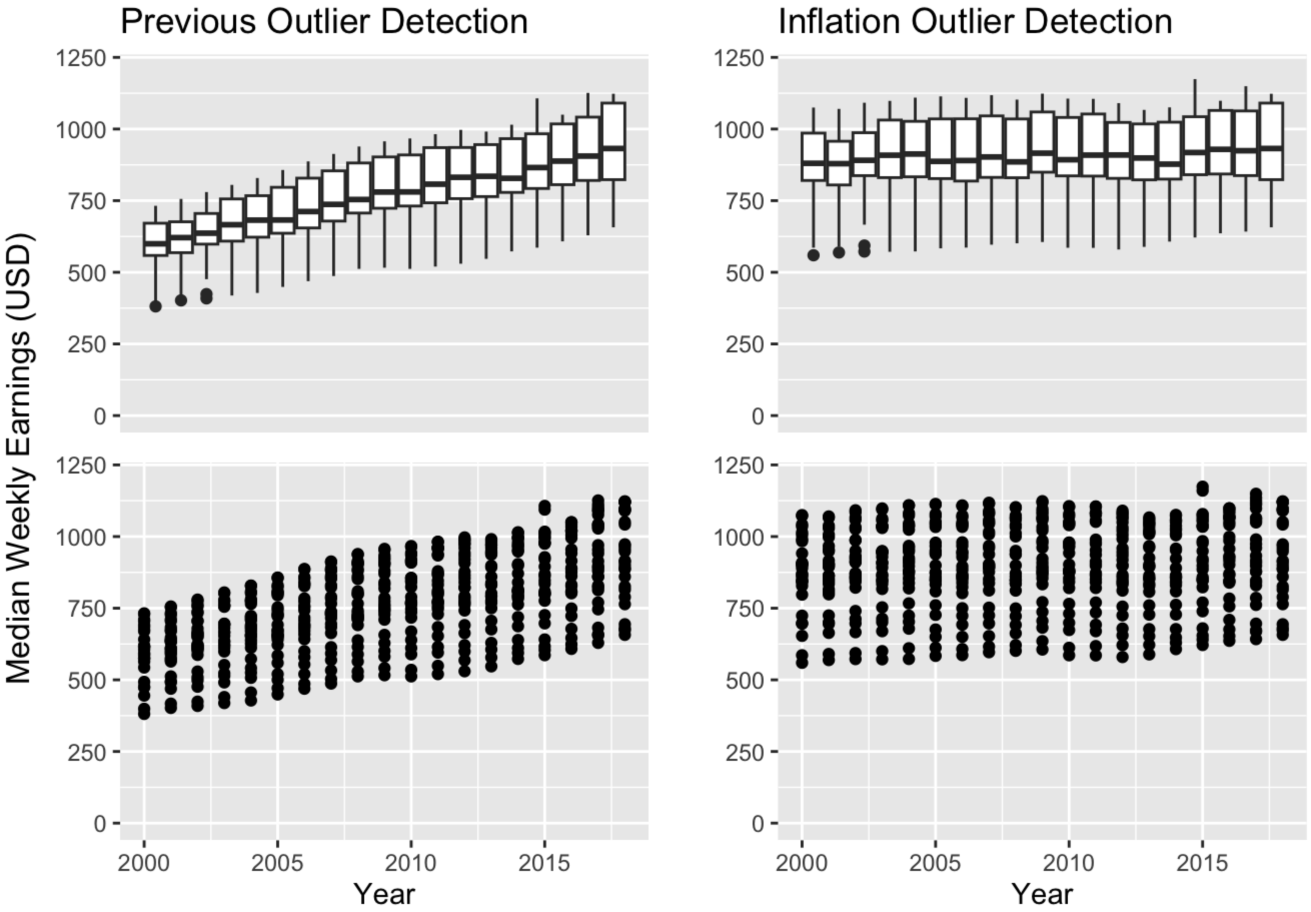
```
ggplot(df, aes(x = Year)) +
  geom_point(aes(y = Median.usual.weekly.earnings, color = "Original Weekly Earnings")
  geom_point(aes(y = Adjusted.Weekly.Earnings, color = "Adjusted Weekly Earnings"), si
  labs(
    title = "Comparison of Original and Inflation Adjusted Weekly Earnings",
    x = "Year",
    y = "Weekly Earnings (USD)",
    color = "Legend"
  ) +
  scale_color_manual(values = c("Original Weekly Earnings" = "blue", "Adjusted Weekly
```



Similarly, we see the dollars that were most effected are in the earliest years while years closer to 2018 are virtually unchanged.

```
library(ggplot2)
library(patchwork)
```

```
# Create Scatter Plot
previous_scatter_plot <- ggplot(df, aes(x = Year, y = Median.usual.weekly.earnings)) +
  geom_point() +
  labs(y = "Median Weekly Earnings (USD)") + ylim(0,1200)

# Create Box Plot
previous_box_plot <- ggplot(df, aes(group = Year, y = Median.usual.weekly.earnings)) +
  geom_boxplot() +
  labs( title = "Previous Outlier Detection") + scale_x_continuous(breaks=unique(df$Ye

# Create Scatter Plot
scatter_plot <- ggplot(df, aes(x = Year, y = Adjusted.Weekly.Earnings)) +
  geom_point()+ ylim(0,1200)

# Create Box Plot
box_plot <- ggplot(df, aes(group = Year, y = Adjusted.Weekly.Earnings)) +
  geom_boxplot() +
  labs( title = "Inflation Outlier Detection") + scale_x_continuous(breaks=unique(df$Y

scatter_plot$labels$y <- box_plot$labels$y <- previous_scatter_plot$labels$y <- previo

# Combine
combined_plot <-  (previous_box_plot / previous_scatter_plot)|(box_plot / scatter_plot
  plot_layout(guides = "collect", heights = c(3,3))

combined_plot
# Help from https://stackoverflow.com/questions/65291723/merging-two-y-axes-titles-in-
grid::grid.draw(grid::textGrob("Median Weekly Earnings (USD)", x = 0.02, rot = 90))
```



We can then compare our outlier detection and see that even with adjustments for inflation, there are no notable outliers. The remainder of our analysis will use Adjusted Weekly Earnings to account for inflation in our tests.

8. Transformation

The data does not require any further transformations.

9. Removing Unnecessary Columns

We can check to see if there are more than 1 unique value in the following columns:

```
cat("UNIQUE AGE VALUES: ", unique(df$Age))
```

```
UNIQUE AGE VALUES:  16 years and over
```

```
cat("UNIQUE INDUSTRY VALUES: ", unique(df$Idustry))
```

```
UNIQUE INDUSTRY VALUES:
```

```
cat("UNIQUE OCCUPATION VALUES:", unique(df$Occupation))
```

```
UNIQUE OCCUPATION VALUES: All Occupations
```

There is only one category in the age, industry, and occupation category, therefore we cannot test anything in relation to them so we can remove them. This data can be collected, however, we are seeking to dive deeper in to the original Guardian article from the dataset, and therefore we do not want to expand the data beyond what they used. We will keep Median Usual Weekly Earnings just in case we need it, even though we are not planning to use it anymore and could recover it from Adjusted Weekly Earnings. Leaving the following columns:

```
df <- df[ , !names(df) %in%
    c("Age", "Industry", "Occupation")]

print(names(df))
```

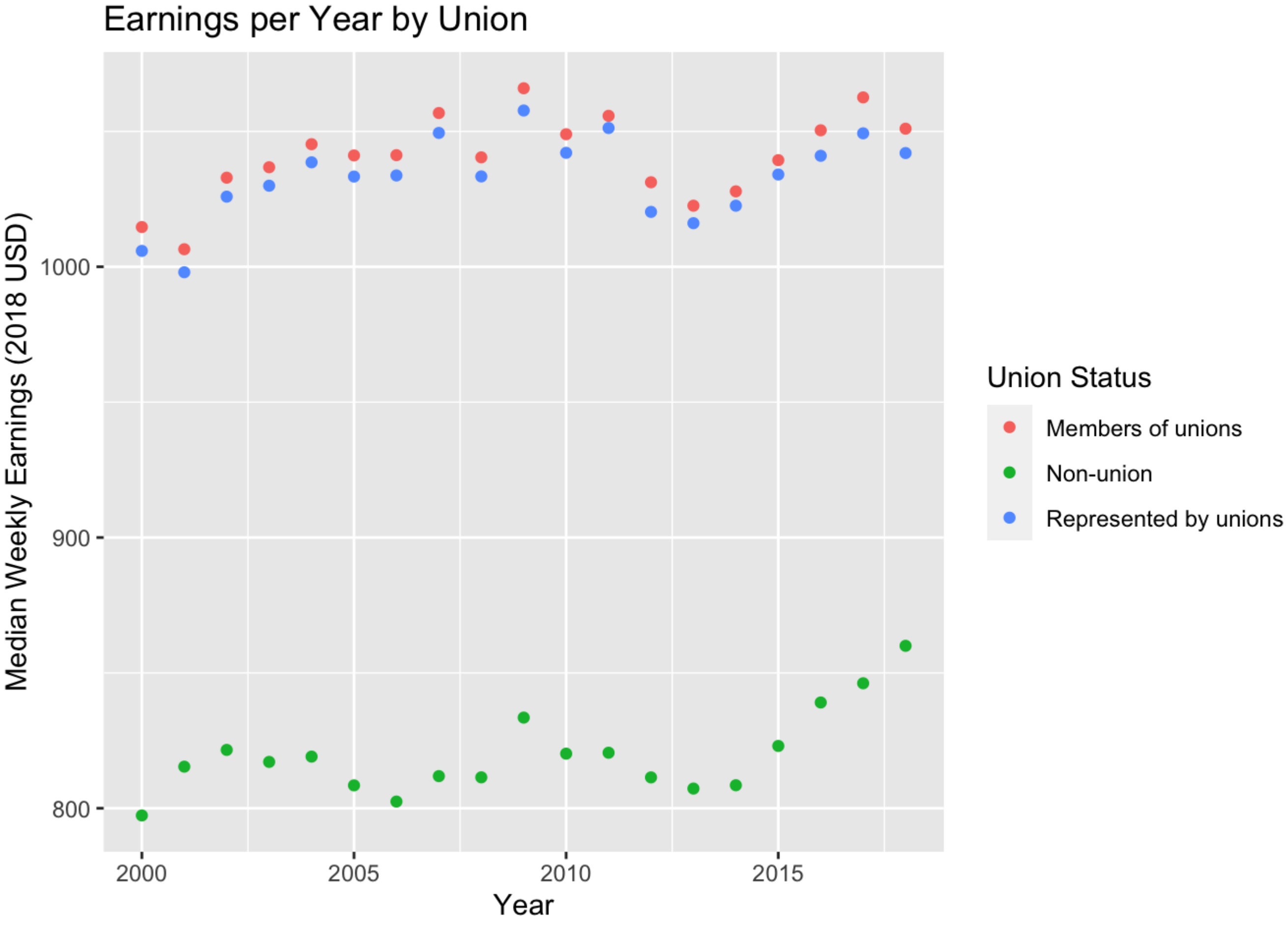
```
[1] "Year"                        "Median.usual.weekly.earnings"
[3] "Union"                       "Sex"
[5] "Race"                        "Ethnic.Origin"
[7] "Date"                        "CPIAUCNS"
[9] "Adjusted.Weekly.Earnings"
```

EDA

We can now compare each categorical variables earnings on both a year basis and in the final year, to gain a better understanding of the data.

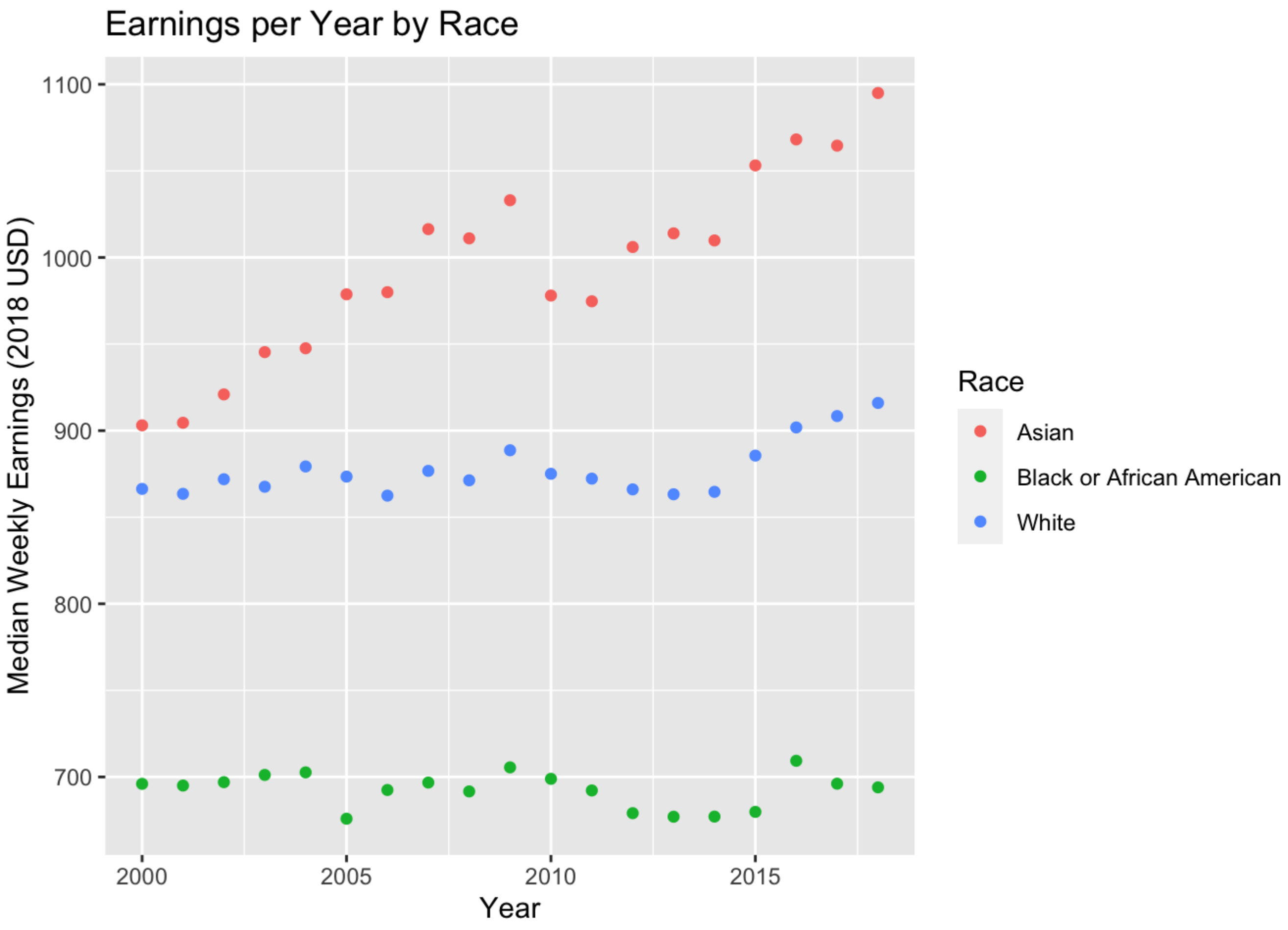
Yearly Earnings by Categorical Variables

```
# Dataframe that only compares Median Earnings, Year, and Union Membership
union_df <- df %>% filter(Sex == "Both Sexes", Race=="All Races", Ethnic.Origin=="All

# Plotting
ggplot(union_df, aes(x = Year, y = Adjusted.Weekly.Earnings, color=Union)) +
  geom_point() +
  labs(title = "Earnings per Year by Union", x = "Year", y = "Median Weekly Earnings (
```



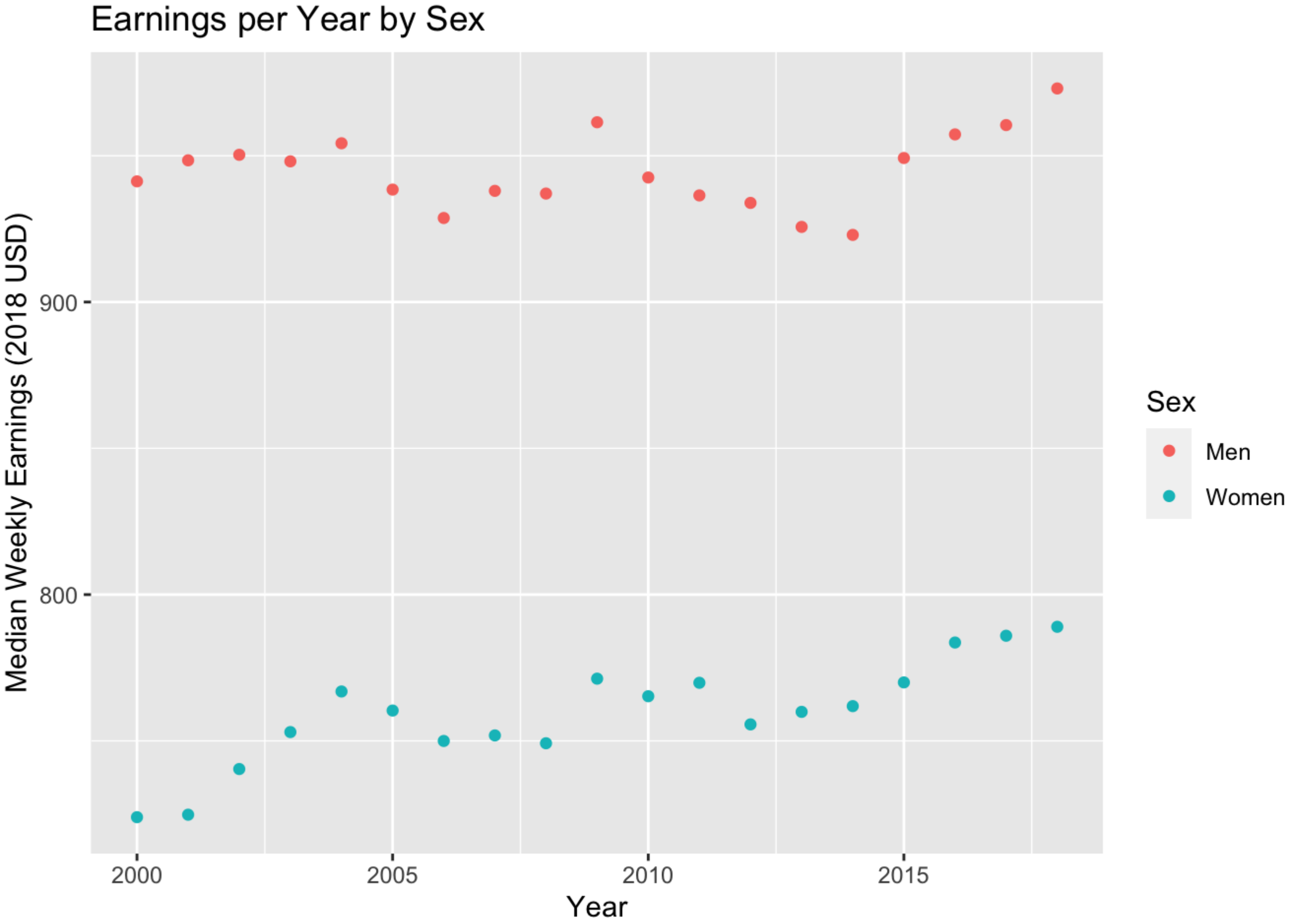
We can see that Members of Unions seem to earn the most, shortly followed by people Represented by Unions with a significant gap between non-union workers. Finally, we can see that earnings of all three groups seem to move together, with some minor differences such as in 2001 when Non-Union earnings increased while Union earnings decreased.

```
# Dataframe that just focuses on Race, only looking at the grouped other columns
race_df <- df %>% filter(Sex == "Both Sexes", Union=="All", Ethnic.Origin=="All Origin

# Plotting
ggplot(race_df, aes(x = Year, y =Adjusted.Weekly.Earnings, color=Race)) +
  geom_point() +
  labs(title = "Earnings per Year by Race", x = "Year", y = "Median Weekly Earnings (2
```



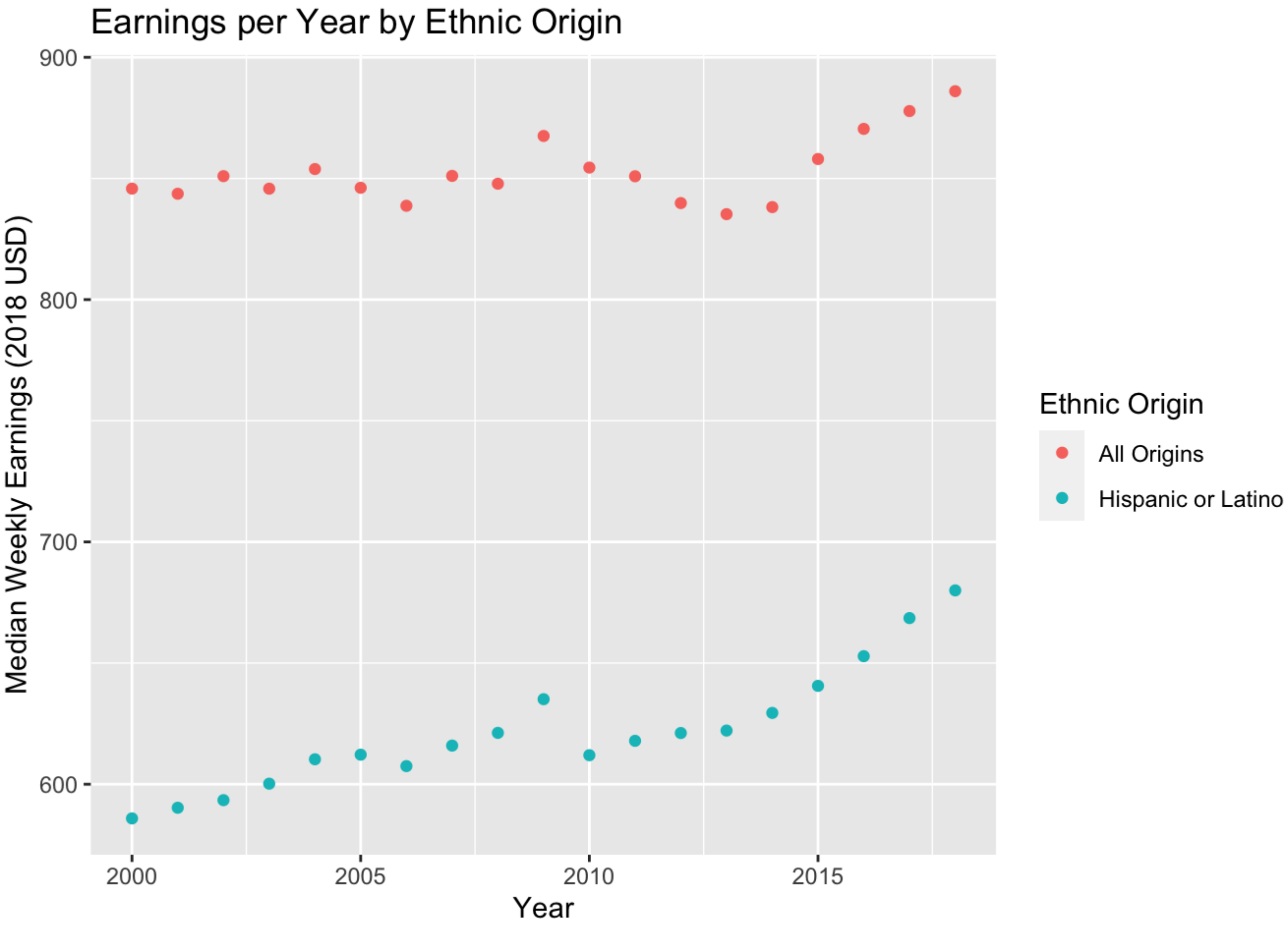
Here we see that Asian Americans seemed to earn the most year over year, followed by White Americans, then Black or African Americans. Furthermore, we seem to see a growth in Asian American Earnings not found in other groups. Compared to in the previous graph, there is not nearly the same unison in growth and decline.

```
# Dataframe that compares differences between the sexes
sex_df <- df %>% filter(Race == "All Races", Union=="All", Ethnic.Origin=="All Origins

# Plotting
ggplot(sex_df, aes(x = Year, y = Adjusted.Weekly.Earnings, color=Sex)) +
  geom_point() +
  labs(title = "Earnings per Year by Sex", x = "Year", y = "Median Weekly Earnings (20
```



Here, we see that Men seem to earn more year over year than Women. The groups seem to move similarly, with little visible change in the size of the gap between the two. However, we cannot claim to know for certain without more rigorous statistical tests.

```
# Dataframe that compares differences between the ethnic origin
ethnic_df <- df %>% filter(Race == "All Races", Union=="All", Sex =="Both Sexes")  %>%

# Plotting
ggplot(ethnic_df, aes(x = Year, y =Adjusted.Weekly.Earnings, color=Ethnic.Origin)) +
  geom_point() +
  labs(title = "Earnings per Year by Ethnic Origin", x = "Year", y = "Median Weekly Ea
```

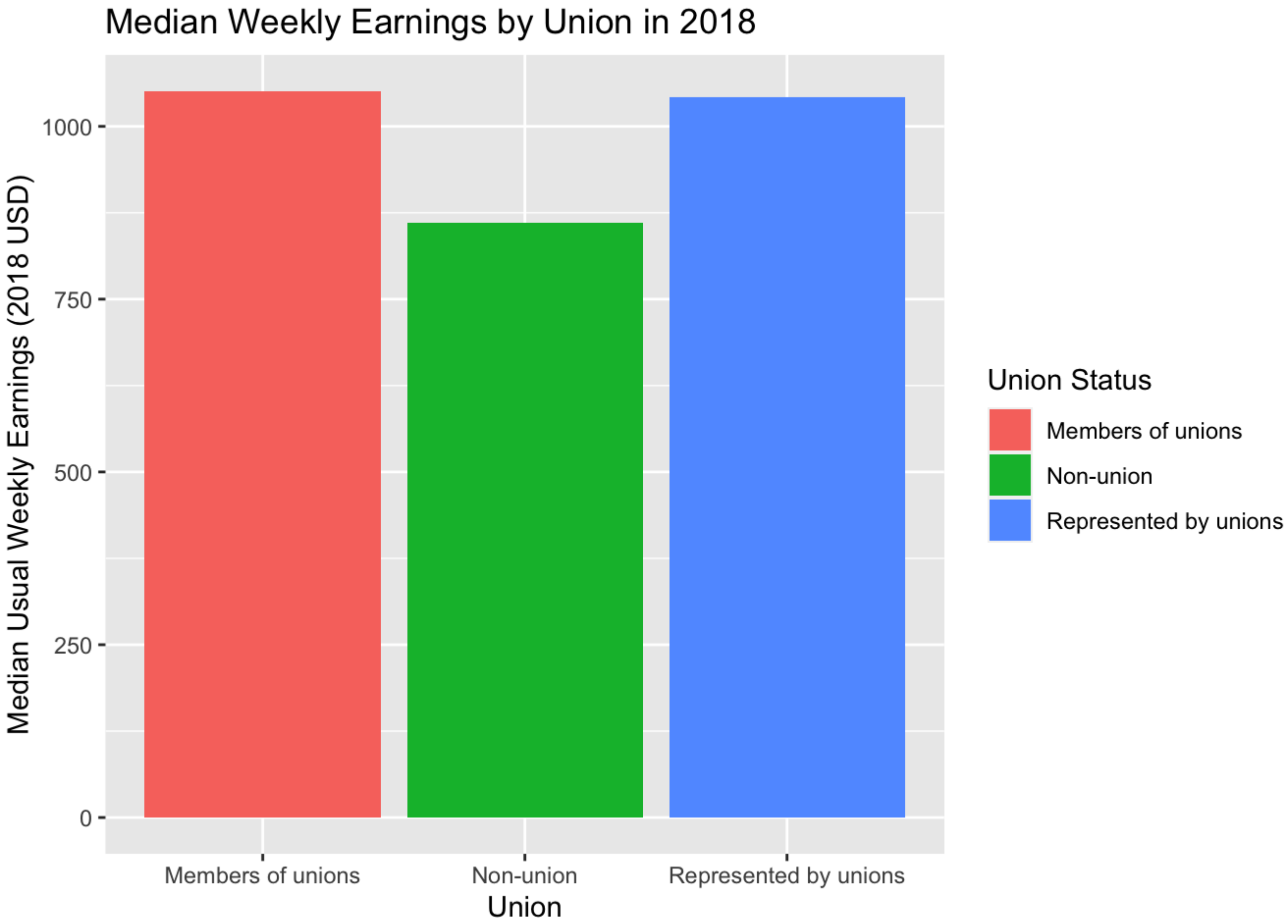


Finally, we see that people of ‘All Origins’ seem to earn more that people of Hispanic or Latino origin. However, we found ‘All Origins’ to be too vague of a grouping to conduct any further analysis on it as it is unclear whether it contains truly *all* origins or is focused on people who are not of Hispanic or Latino origin.

Final Year Histograms

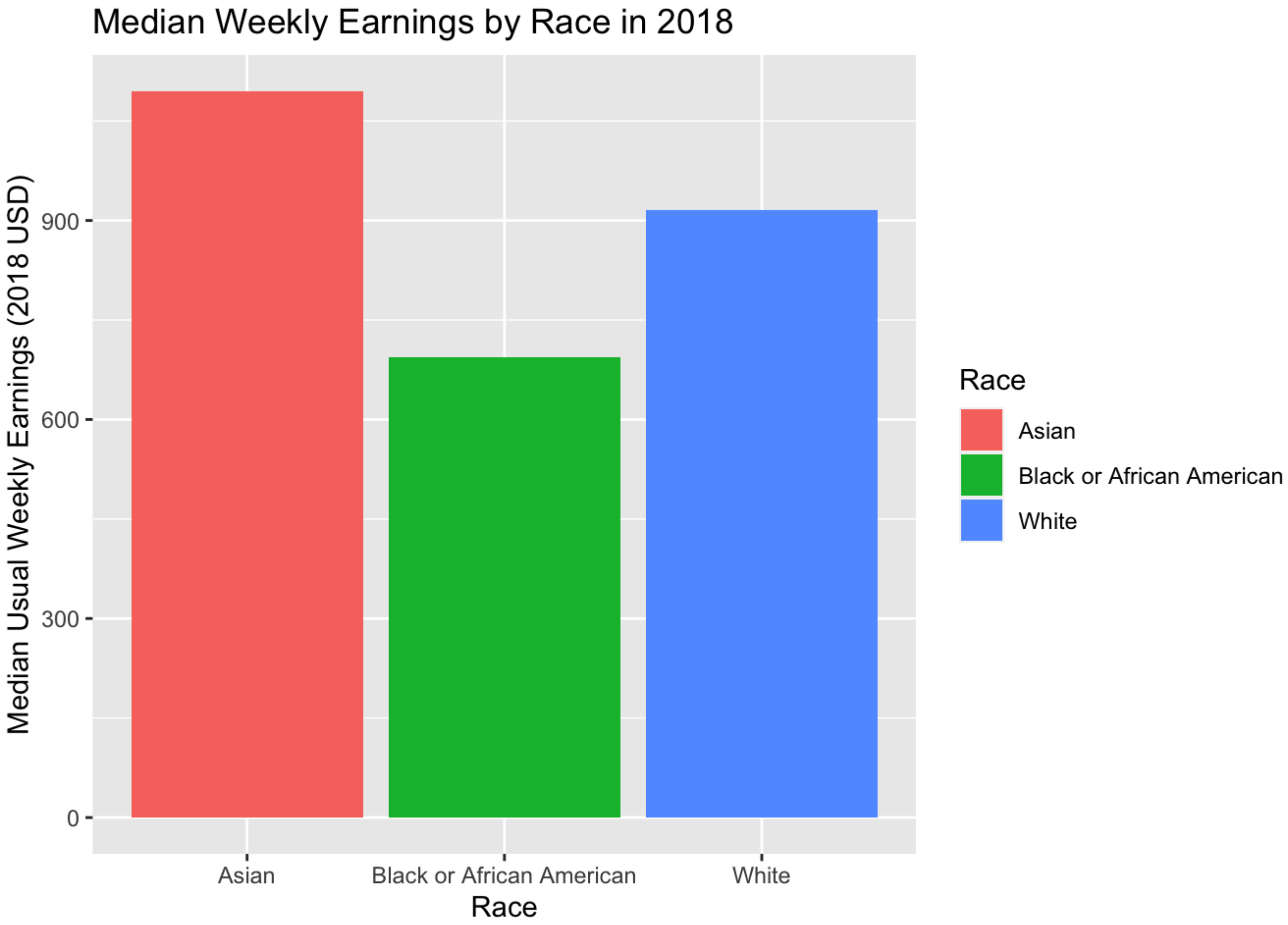
Here, we can more easily view the difference between each categorical group in the final year of the dataset:

```
ggplot(union_df %>% filter(Year == max(df$Year), Union != "All"), aes(x = Union, y = A
  geom_col() +
  labs(title = "Median Weekly Earnings by Union in 2018", x = "Union", y = "Median Usu
```



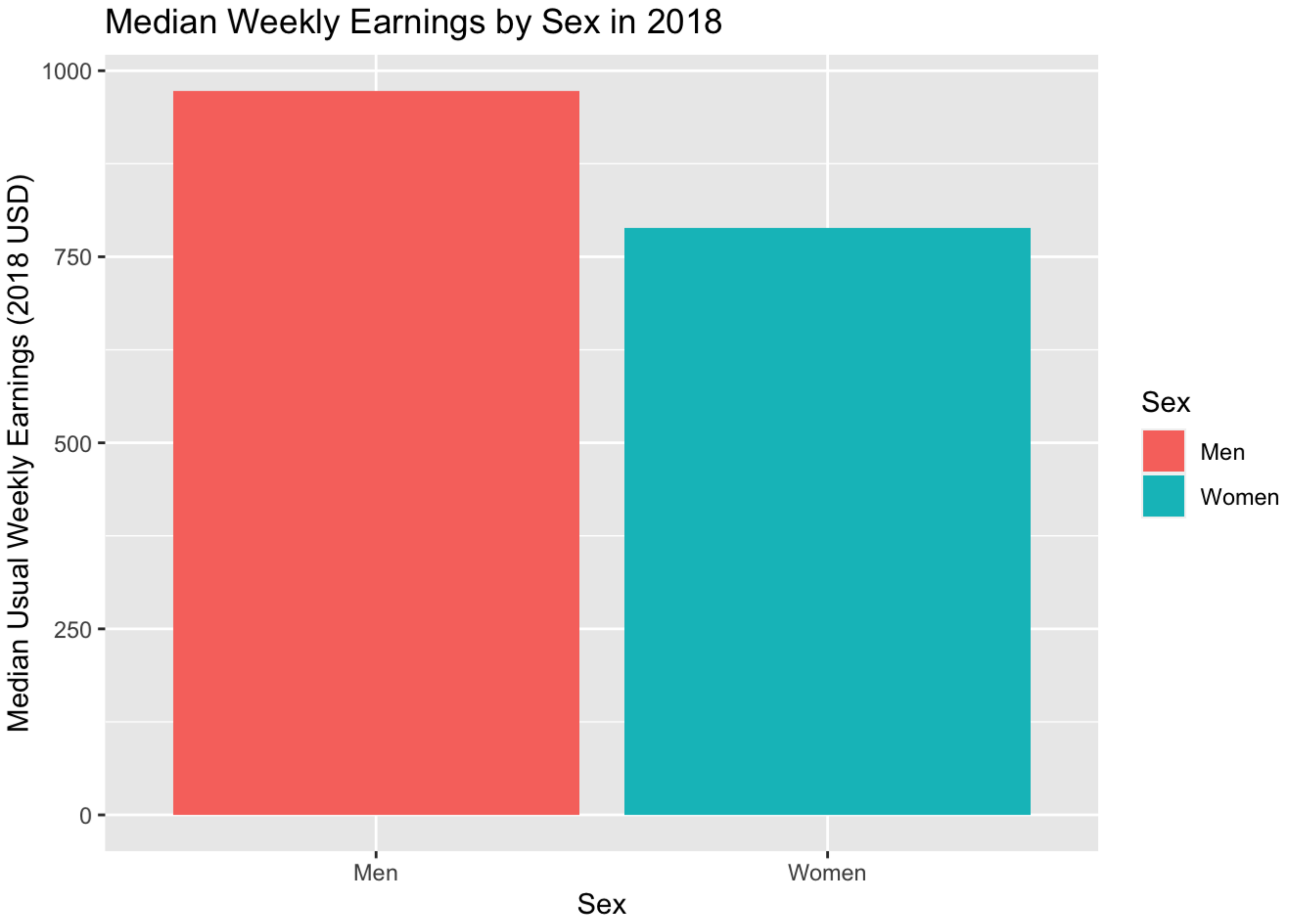
Just as above, it appears that Members of Unions earn the most, closely followed by people represented by unions. They are then both significantly larger than non-union workers, which is roughly 75% of the other two from what we can see.

```
ggplot(race_df %>% filter(Year == max(df$Year), Race!="All Races"), aes(x = Race, y =
  geom_col() +
  labs(title = "Median Weekly Earnings by Race in 2018", x = "Race", y = "Median Usual
```



Earnings in 2018 appear to be in line with what we observed from the scatter plot.

```
ggplot(sex_df %>% filter(Year == max(df$Year), Sex!="Both Sexes"), aes(x = Sex, y = Ad
  geom_col() +
  labs(title = "Median Weekly Earnings by Sex in 2018", x = "Sex", y = "Median Usual W
```



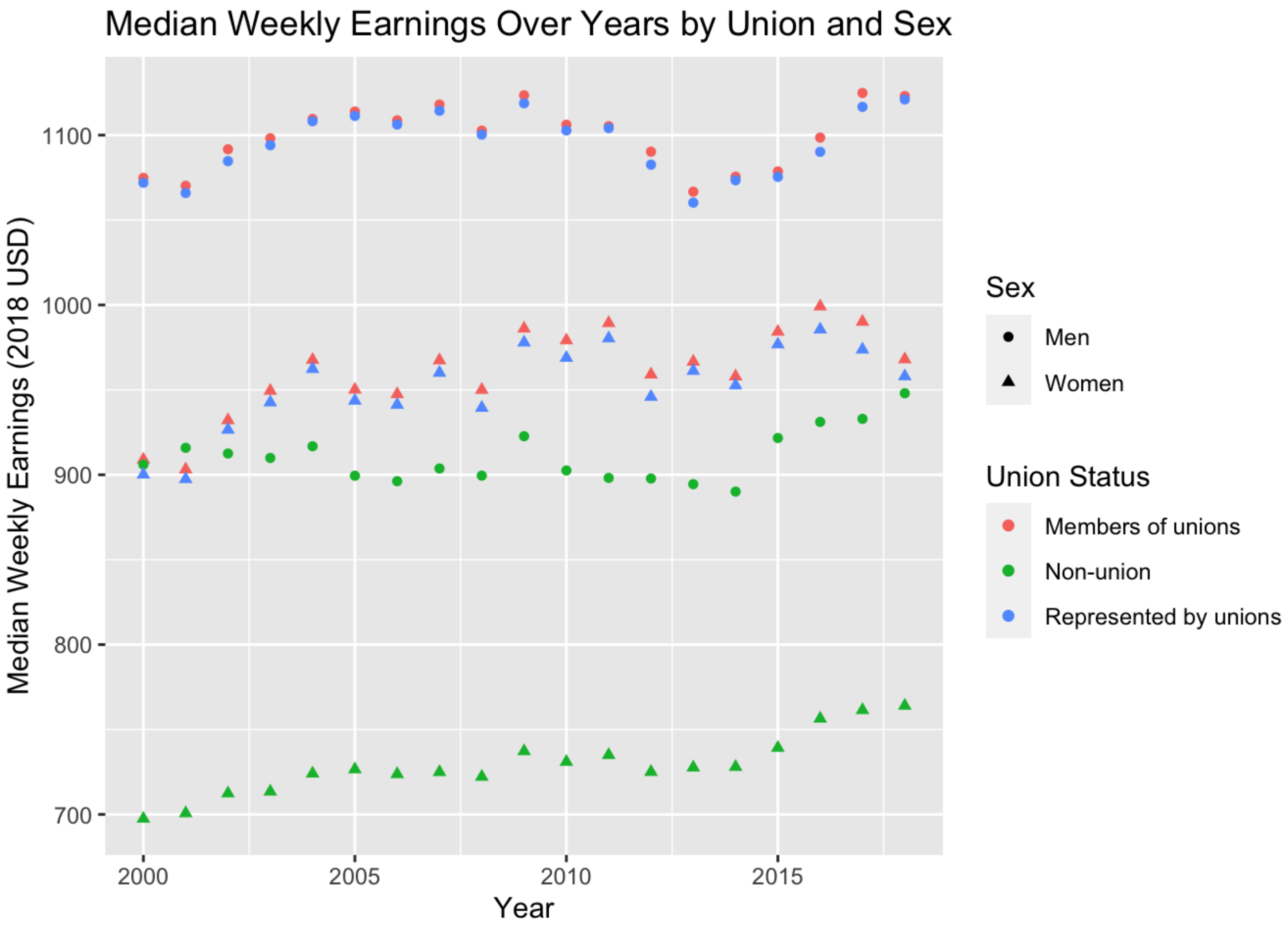
Finally, here we see the infamous gender wage gap around the often discussed value around 76 cents on the dollar between women and men in 2018.

Combined Plots

Next, we can combine the plots to attempt to observe relationships between people of differing union status and demographics. However, the data does not allow us to compare across demographics. For example, we cannot look at “White Men” vs “Asian Women” as there are no population sizes nor earnings values for those groups.

```
ggplot(df %>% filter(Sex!="Both Sexes", Union !="All") %>% group_by(Year, Union, Sex)
  geom_point() +
  labs(
    title = "Median Weekly Earnings Over Years by Union and Sex",
    x = "Year",
    y = "Median Weekly Earnings (2018 USD)",
    color = "Union Status",
    shape = "Sex"
  )
```

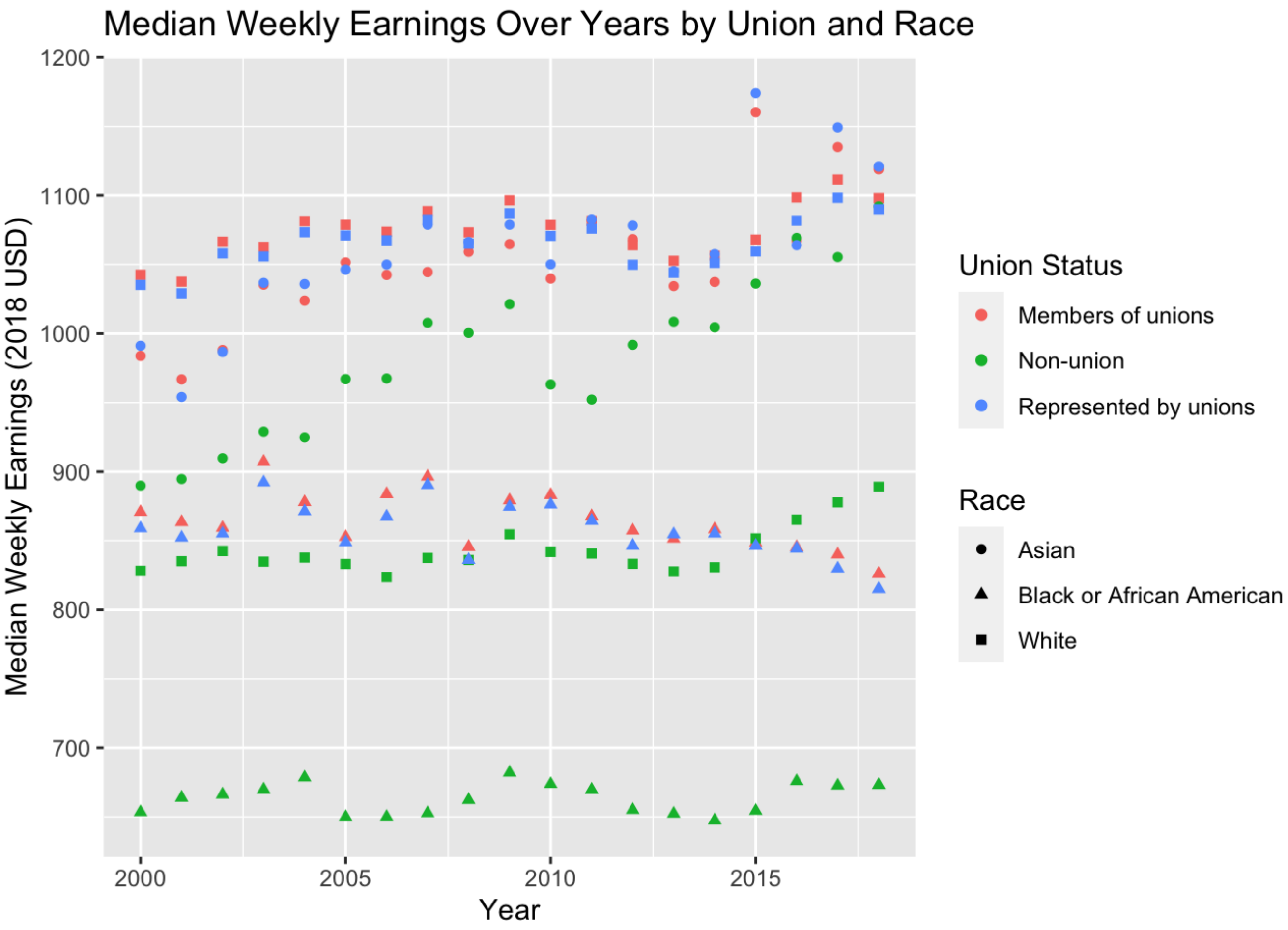
```
`summarise()` has grouped output by 'Year', 'Union'. You can override using the
`.groups` argument.
```



These plots get a bit more difficult to read as they become increasingly complex. A few points of note on this plot can be drawn. First, we again see members of unions and people represented by unions moving in near parallel with members of unions slightly above. Second, it appears union associated people earn more than their counterparts. Non-union women, as we would have expected from our prior EDA, earn what looks like significantly the least. We cannot observe whether the gap between union workers and non-union workers is any larger for either sex. Finally, it appears that unionized women and non-union men earned similar amounts in both 2000 and 2018.

```
ggplot(df %>% filter(Union!="All", Race !="All Races") %>% group_by(Year, Union, Race)
  geom_point() +
  labs(
    title = "Median Weekly Earnings Over Years by Union and Race",
    x = "Year",
    y = "Median Weekly Earnings (2018 USD)",
    color = "Union Status",
    shape = "Race"
  )
```

```
`summarise()` has grouped output by 'Year', 'Union'. You can override using the
`.groups` argument.
```



Black or African American non-union members appear to be significantly below all other groups. Similarly, non-union members across the board appear below their counterparts. T We can see some interesting trends such as rapid growth in non-union Asian workers pay that does not seem to be followed by their unionized counterparts. Furthermore, non-union Asian workers appear to earn more than unionized Black or African American workers; in recent years, it seems that white workers are increasingly doing the same. here is much more overlap between groups here, making it difficult to assess the scale of much of the gaps. Additionally, this difficulty further demonstrates the need for statistical testing.

3 Results

Hypothesis Testing

Independent Two-Sample T-Test

From our EDA plots, we see a notable gap between the wages of workers that are represented by unions, and those that are not. While Members of Unions and people Represented by Unions are not exactly the same, they are nearly identical. We are primarily concerned with differences between using a union at all and a total lack of unionization. Therefore, to test whether there is a significant difference between the wages of these 2 groups of workers, we can conduct an **Independent two-sample t-test**.

```
# add a new column combining union members and represented by unions
union_df <- union_df %>%
  mutate(union_combined = ifelse(Union == "Non-union", "non-union", "union"))
df <- df %>%
  mutate(union_combined = ifelse(Union == "Non-union", "non-union", "union"))
```

This test takes 2 assumptions: **normality** and **equality of variance**, both of which we will first test for.

First, checking for normality, we use the Shapiro test.

- Null Hypothesis (Ho): The distribution of wages is normal
- Alternative Hypothesis (Ha): The distribution of wages is not normal

```
shapiro.test(union_df$Adjusted.Weekly.Earnings[union_df$union_combined == "union"])
```

```

	Shapiro-Wilk normality test

data:  union_df$Adjusted.Weekly.Earnings[union_df$union_combined == "union"]
W = 0.97677, p-value = 0.6033
```

```
shapiro.test(union_df$Adjusted.Weekly.Earnings[union_df$union_combined == "non-union"]
```

```

	Shapiro-Wilk normality test

data:  union_df$Adjusted.Weekly.Earnings[union_df$union_combined == "non-union"]
W = 0.90687, p-value = 0.06489
```

Looking at our results, we see that for both distributions, the p value is greater than 0.05. This indicates that we fail to reject our null hypothesis, passing the test for normality for both our samples.

Next, we check for equal variances between the two groups using Levene’s test.

- Null Hypothesis (Ho): The variances in wages are equal.
- Alternative Hypothesis (Ha): The variances in wages are not equal.

```
library(car)
leveneTest(Adjusted.Weekly.Earnings ~ union_combined, data = df)
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  0.0085 0.9265
      530               
```

Since our p value is greater than 0.05, we fail to reject our null hypothesis that the distribution’s variances are equal, passing the test for equality of variances.

Given that we have satisfied our assumptions, we can go ahead and conduct the 2 sample t-test

Null Hypothesis: The 2 groups’ mean salaries are the same.

Alternate Hypothesis: The 2 groups’ mean salaries are not the same.

```
t_test_result <- t.test(Adjusted.Weekly.Earnings ~ union_combined, data = df)

print(t_test_result)
```

```

	Welch Two Sample t-test

data:  Adjusted.Weekly.Earnings by union_combined
t = -10.965, df = 231.05, p-value < 2.2e-16
alternative hypothesis: true difference in means between group non-union and group
union is not equal to 0
95 percent confidence interval:
 -169.9045 -118.1451
sample estimates:
mean in group non-union     mean in group union 
               792.1493                936.1741 
```

From our test, our extremely low p-value (< 0.05) indicates very strong evidence against the null hypothesis, so we can confidently reject the null hypothesis that the two groups’ mean salaries are equal. Furthermore, looking at our 95% confidence interval of [-169.90,-118.14], we can further support our hypothesis that the average salaries for non-union workers are significantly lower than that of union workers. It is also important to understand what that confidence interval represents. That range indicates that non-union workers are earning between $170 and $118 less than their equal counterparts who are represented by a union. Using the most conservative estimate (the bottom end of the interval), this projects to over $5000 less annually, a massive difference.

Bootstrapping

We will now test the same hypothesis using another method, bootstrapping. Bootstrapping is a statistical procedure that involves re-sampling a single dataset to create many simulated samples. This process allows us to calculate standard errors, construct confidence intervals, and even perform hypothesis tests for numerous types of sample statistics. Bootstrap methods are alternative approaches to traditional hypothesis testing, and are also notable for being easier to understand and valid for more conditions.

```
in_union = df[df$union_combined == "union",]
out_union = df[df$union_combined == "non-union",]
```

Now, we will create 100,000 bootstrap samples of difference of sample means between our two groups, union and non-union workers:

```
N = 100000

earnings_diff = numeric(N)


for(i in 1:N){
  boot_i = mean(sample(in_union$Adjusted.Weekly.Earnings, length(in_union$Adjusted.Wee
  boot_o = mean(sample(out_union$Adjusted.Weekly.Earnings, length(out_union$Adjusted.W
  earnings_diff[i] = boot_i - boot_o
}

head(earnings_diff)
```

```
[1] 163.2710 122.2067 162.4763 139.2057 161.4729 164.8315
```

We will use our bootstrapping to make a 95% confidence interval for the difference estimate, and using this confidence interval, draw a conclusion on our hypothesis.

```
conf = quantile(earnings_diff, c(0.025, 0.975))

conf
```

```
    2.5%    97.5% 
118.0882 169.6415 
```

Based on the outputted confidence interval, we can reject our null hypothesis that there is no significant difference between the mean salaries of union vs. non-union workers. We can say this because according to our confidence interval, we see that a union worker's median weekly earnings are greater than that of their non-union counterparts by $118 - $170, 95% of the time.

Bootstrapping Ratio of Means

We conducted another t-test to explore if the median weekly earnings between between men and women are significantly different.

Testing for normality and equality of variance before performing T-test on the samples:

```
# Filter data for men and women
data_men <- filter(df, Sex == "Men")
data_women <- filter(df, Sex == "Women")
```

Test for Normality using Shapiro's test

Null Hypothesis (Ho): The distribution of wages is normal

Alternative Hypothesis (Ha): The distribution of wages is not normal

```
shapiro_test_men <- shapiro.test(data_men$Adjusted.Weekly.Earnings)
shapiro_test_women <- shapiro.test(data_women$Adjusted.Weekly.Earnings)

print("Men's Shapiro Test:")
```

```
[1] "Men's Shapiro Test:"
```

```
print(shapiro_test_men)
```

```

	Shapiro-Wilk normality test
```

```
data:  data_men$Adjusted.Weekly.Earnings
W = 0.82155, p-value = 3.684e-08
```

```
print("Women's Shapiro Test:")
```

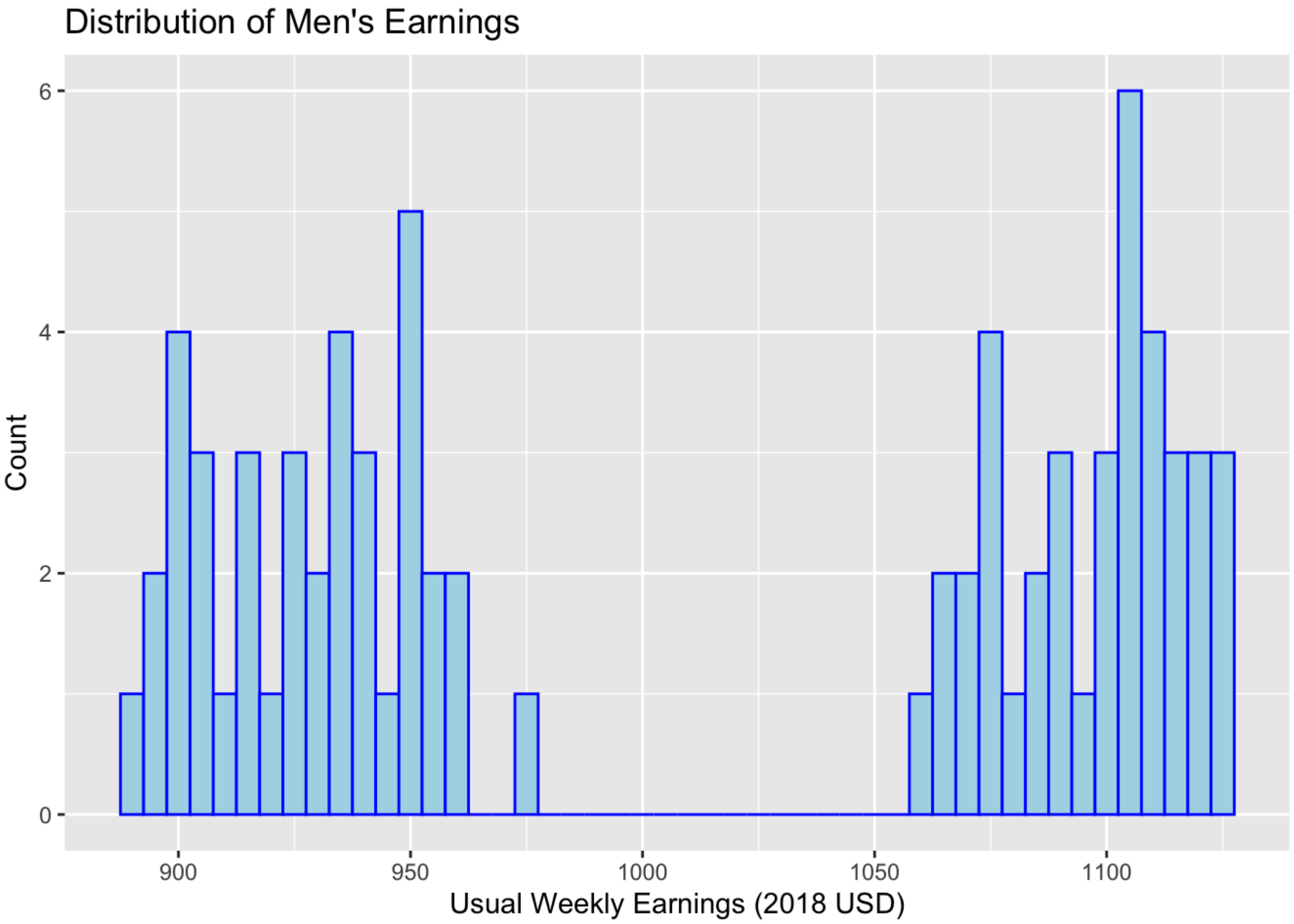
```
[1] "Women's Shapiro Test:"
```

```
print(shapiro_test_women)
```

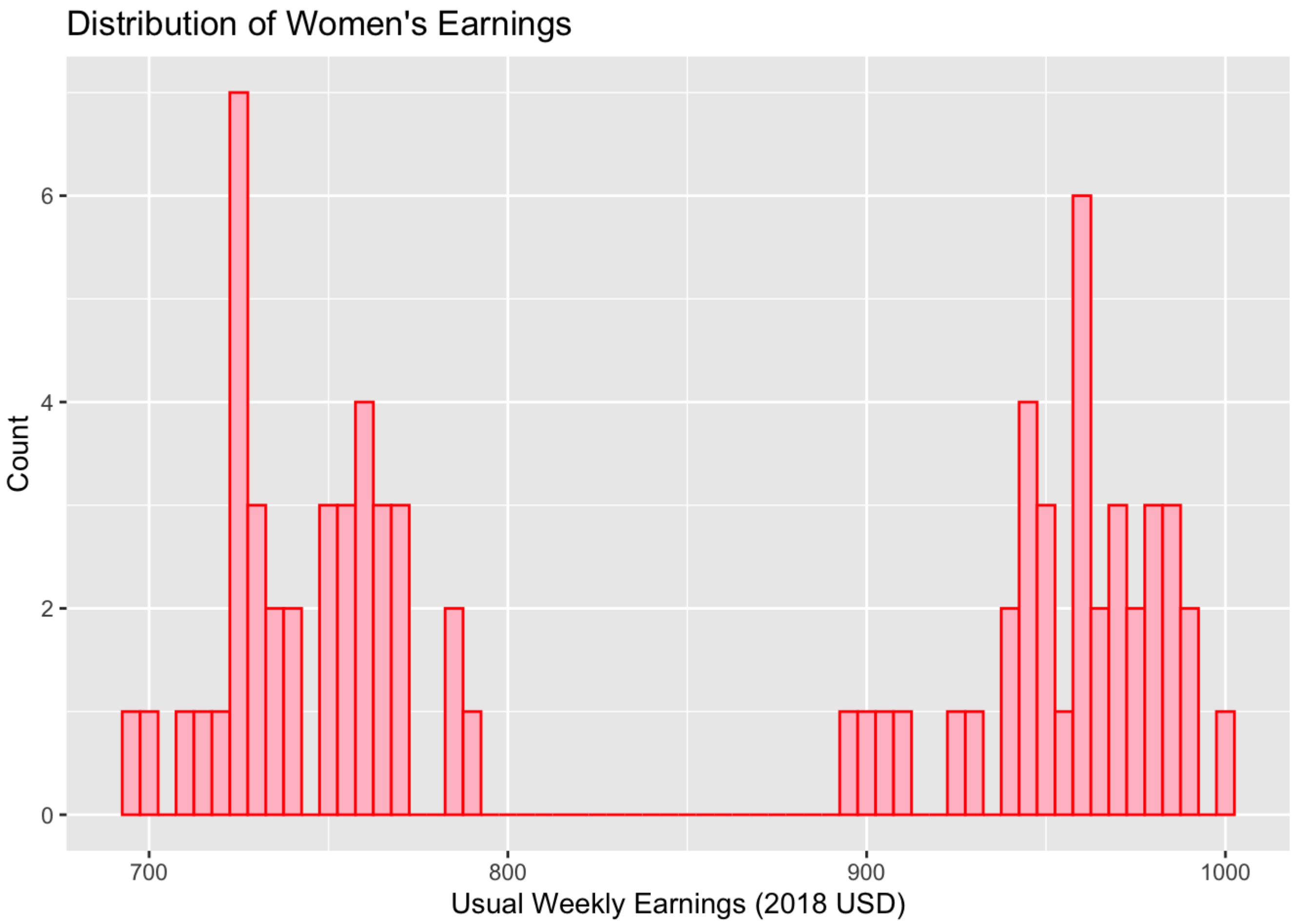
```

	Shapiro-Wilk normality test

data:  data_women$Adjusted.Weekly.Earnings
W = 0.8115, p-value = 1.881e-08
```

Both of our p-values are significantly below 0.05, demonstrating that neither of the distributions are normal. We can double check to see if this is matched by an eye test by plotting the data:

```
ggplot(data_men, aes(x=Adjusted.Weekly.Earnings)) + geom_histogram(binwidth = 5, color
```



```
ggplot(data_women, aes(x=Adjusted.Weekly.Earnings)) + geom_histogram(binwidth = 5, col
```


Distribution of Women's Earnings

We can very clearly see that neither of the two is normally distributed, breaking one of the assumptions of a t-test. Therefore, we will have to use something else. We can instead, bootstrap and compare the ratio of means between Men and Women on their Usual Weekly Earnings:

```r
data_filtered <- subset(df, Sex %in% c("Men", "Women"))
```

```r
library(boot)
ratio_of_means <- function(data, indices) {
  men <- data$Adjusted.Weekly.Earnings[data$Sex == "Men"][indices]
  men <- na.omit(men)
  women <- data$Adjusted.Weekly.Earnings[data$Sex == "Women"][indices]
  women <- na.omit(women)
  mean_ratio <- mean(women) / mean(men)
  return(mean_ratio)
}

# Perform bootstrap
bootstrap_results <- boot(data = data_filtered, statistic = ratio_of_means, R = 10000)

# Get Confidence Interval
boot.ci(bootstrap_results, conf = 0.95)
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 10000 bootstrap replicates

CALL :
boot.ci(boot.out = bootstrap_results, conf = 0.95)
```

```
Intervals :
Level      Normal              Basic
95%   ( 0.8125,  0.8673 )   ( 0.8115,  0.8671 )

Level     Percentile            BCa
95%   ( 0.8132,  0.8688 )   ( 0.8126,  0.8682 )
Calculations and Intervals on Original Scale
```

The bootstrapping results in a confidence interval range between 0.81 and 0.87 on the ratio of means between women and men. In plainer terms, on average, women earning between 81% and 87% of what their male counterparts are earning, a massive discrepancy. Therefore, we can conclude with 95% confidence that women are earning at least 10% less than their male counterparts.

One-way Anova Test

The ANOVA test is a statistical method used to compare the means of three or more groups to determine if there is a statistically significant difference among them. A One-way ANOVA tests differences between groups that are based on one independent variable. In this case, we would like to compare the mean usual weekly earnings.

Unions aim to reduce racial disparities by advocating for equal pay and treatment for workers of all racial backgrounds. Therefore, we would like to answer the data science question if there is a significant pay disparity between the racial groups i.e Asian Americans, Whites Americans, and Blacks or African Americans who belong to a Union.

Null Hypothesis (Ho): There is no significant difference in the mean weekly earnings between the three racial groups.

Alternative Hypothesis (Ha): There is a significant difference in the mean weekly earnings between the three racial groups.

```
data_subset <- subset(df, Union %in% c("Members of unions","Represented by unions") &

anova_result <- aov(Adjusted.Weekly.Earnings ~ Race, data = data_subset)
print(summary(anova_result))
```

```
             Df  Sum Sq Mean Sq F value Pr(>F)    
Race          2 1046454  523227   479.6 <2e-16 ***
Residuals   111  121099    1091                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The very low p-value suggests that the difference in the mean weekly earnings between the groups is significant. The large F-value of 479 gives more evidence against the null hypothesis and suggests that a large portion of the variance in earnings is explained by racial inequalities.

From this test, we can reject the Null hypothesis to conclude there is a significant difference in the mean weekly earnings between the three racial groups. Extending our results, these would suggest that, despite improving pay in general for its members, unions could be doing more to promote racial pay equality for those who are members and who they represent.

4 Conclusions

In our analyses, the first hypothesis we sought to test is whether there is a significant difference between the weekly earnings for union vs non-union members. To do this, we first conducted a 2-sample t-test, with the null hypothesis being that there is no significant difference between the salaries of Union and Non-Union workers. However, our t-test yielded a very small p-value, giving us significant evidence to reject our null hypothesis and claim that **union membership correlates to statistically significantly higher level of earnings**. We then tested this same hypothesis by creating 100,000 bootstrap samples. Our bootstrap test also gave us the same results, meaning that we could confidently conclude that there's a significant difference between the median weekly earnings for union vs. non-union members, with union members earning more, on average, than their non-union counterparts.

The outcome of the bootstrap ratio of means tests suggests a notable pay gap between men and women, with women earning between 81% and 87% of their male counterparts. We then sought to observe how well unions represent racial equality. From the ANOVA test we can see that the union members of different the racial groups have significantly different weekly earnings. Therefore, we can conclude that Unions can do better in advocating for all racial groups in terms of achieving equitable pay.

Finally, our analysis could be continued and expanded in a variety of ways. First, with population data or a deeper dataset, it would be interesting to compare how well unions are doing at representing more specific marginalize groups such as non-union black women or also people who are in some groups that benefit from inequality but others that are not such as unionized white women. Furthermore, other metrics of the benefits of unions could be considered such as hours worked and health care benefits. Finally, other deliminators considered by the BLS could be included such as Industry to see if women in finance are better represented than women in education, for example.

5 References

- Bureau of Labor Statistics. "Current Population Survey Design and Methodology Technical Paper 77." Census.gov, 2019, https://www2.census.gov/programs-surveys/cps/methodology/CPS-Tech-Paper-77.pdf. Accessed 29 November 2023.

- Chalabi, Mona. "How much does union membership benefit America's workers? | Mona Chalabi." The Guardian, 24 November 2019, https://www.theguardian.com/news/datablog/2019/nov/24/how-much-does-union-membership-benefit-americas-workers. Accessed 29 November 2023.

- FRED. "Consumer Price Index for All Urban Consumers: All Items in U.S. City Average." FRED, https://fred.stlouisfed.org/series/CPIAUCNS. Accessed 29 November 2023.

- "Table 2. Median weekly earnings of full-time wage and salary workers by union affiliation and selected characteristics." Bureau of Labor Statistics, 16 September 2015, https://www.bls.gov/webapps/legacy/cpslutab2.htm#. Accessed 29 November 2023.

- Van Green, Ted. "Majorities of adults see decline of union membership as bad for the U.S. and working people." Pew Research Center, https://www.pewresearch.org/shortreads/2023/04/19/majorities-of-adults-see-decline-of-union-membership-as-bad-for-the-u-s-and-working-people/. Accessed 29 November 2023.