1.2 R Refresher Michelle Rice

Michelle Rice

6/10/2021

# Summary

For this study, I used the Bureau of Labor Statistics data on earnings of women and men. I will look at the wage gap to see how large the gap is, how it has changed over time, and how it differs among different ethnic groups.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Year | Total | Total White | Total Black or African American | Total Asian | Total Hispanic or Latino ethnicity | Group |
| 1979 | 4.44 | 4.51 | 4.11 | – | 4.08 | T |
| 1980 | 4.82 | 4.88 | 4.44 | – | 4.44 | T |
| 1981 | 5.15 | 5.18 | 4.90 | – | 4.81 | T |
| 1982 | 5.40 | 5.47 | 5.06 | – | 5.01 | T |
| 1983 | 5.59 | 5.66 | 5.15 | – | 5.09 | T |
| 1984 | 5.83 | 5.90 | 5.36 | – | 5.27 | T |

From the original data set, I have created three separate datasets, one for men, one for women and one for total subjects.

# Create separate datasets  
  
 men <- subset(earningsStudy, Group == 'M')  
 women <- subset(earningsStudy, Group == 'W')  
 total <- subset(earningsStudy, Group == 'T')  
  
# Show summary statistics of each  
 knitr::kable(summary(men[, 1:7]), "pipe")

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Year | Total | Total White | Total Black or African American | Total Asian | Total Hispanic or Latino ethnicity | Group |
|  | Length:41 | Min. : 5.65 | Min. : 5.79 | Min. : 4.890 | Length:41 | Min. : 4.790 | Length:41 |
|  | Class :character | 1st Qu.: 8.10 | 1st Qu.: 8.28 | 1st Qu.: 7.070 | Class :character | 1st Qu.: 6.660 | Class :character |
|  | Mode :character | Median :10.31 | Median :10.61 | Median : 9.770 | Mode :character | Median : 8.610 | Mode :character |
|  | NA | Mean :10.74 | Mean :10.98 | Mean : 9.482 | NA | Mean : 9.143 | NA |
|  | NA | 3rd Qu.:13.76 | 3rd Qu.:13.95 | 3rd Qu.:12.060 | NA | 3rd Qu.:11.810 | NA |
|  | NA | Max. :16.76 | Max. :17.10 | Max. :14.980 | NA | Max. :15.250 | NA |

knitr::kable(summary(women[, 1:7]), "pipe")

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Year | Total | Total White | Total Black or African American | Total Asian | Total Hispanic or Latino ethnicity | Group |
|  | Length:41 | Min. : 3.620 | Min. : 3.62 | Min. : 3.550 | Length:41 | Min. : 3.440 | Length:41 |
|  | Class :character | 1st Qu.: 6.110 | 1st Qu.: 6.13 | 1st Qu.: 5.880 | Class :character | 1st Qu.: 5.530 | Class :character |
|  | Mode :character | Median : 8.640 | Median : 8.73 | Median : 8.130 | Mode :character | Median : 7.460 | Mode :character |
|  | NA | Mean : 8.803 | Mean : 8.88 | Mean : 8.383 | NA | Mean : 7.818 | NA |
|  | NA | 3rd Qu.:11.760 | 3rd Qu.:11.83 | 3rd Qu.:11.010 | NA | 3rd Qu.:10.090 | NA |
|  | NA | Max. :14.850 | Max. :14.93 | Max. :14.000 | NA | Max. :13.740 | NA |

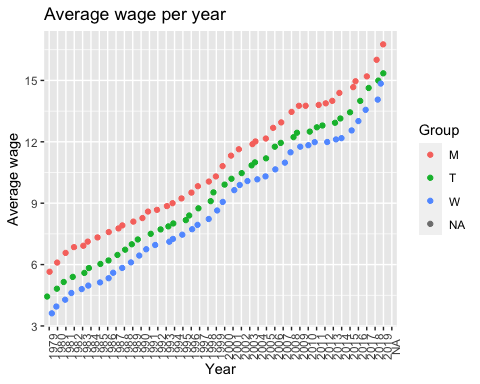
knitr::kable(summary(total[, 1:7]), "pipe")

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Year | Total | Total White | Total Black or African American | Total Asian | Total Hispanic or Latino ethnicity | Group |
|  | Length:41 | Min. : 4.44 | Min. : 4.510 | Min. : 4.110 | Length:41 | Min. : 4.080 | Length:41 |
|  | Class :character | 1st Qu.: 6.99 | 1st Qu.: 7.080 | 1st Qu.: 6.430 | Class :character | 1st Qu.: 6.070 | Class :character |
|  | Mode :character | Median : 9.53 | Median : 9.740 | Median : 8.850 | Mode :character | Median : 8.070 | Mode :character |
|  | NA | Mean : 9.62 | Mean : 9.782 | Mean : 8.873 | NA | Mean : 8.535 | NA |
|  | NA | 3rd Qu.:12.44 | 3rd Qu.:12.660 | 3rd Qu.:11.640 | NA | 3rd Qu.:10.970 | NA |
|  | NA | Max. :15.35 | Max. :15.850 | Max. :14.660 | NA | Max. :14.860 | NA |

The first plot that I will look at is a scatter plot with a different color representing the average wage each year for the three groups, men, women and total.

library(ggplot2)  
  
ggplot(data = earningsStudy, aes(y = Total, x = Year, color=Group)) + geom\_jitter() + labs(title="Average wage per year", x = "Year", y = "Average wage") +theme(axis.text.x = element\_text(angle = 90))

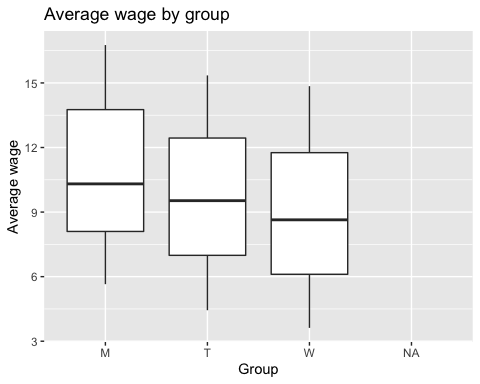
## Warning: Removed 1 rows containing missing values (geom\_point).



The scatter plot shows very clearly that men's wages were consistently higher than women. It also shows a very clear increase over time between all three groups indicating that wages rose consistently over time at about the same rate. I will also look at a boxplot of those same three categories.

library(ggplot2)  
  
ggplot(data = earningsStudy, aes(y = Total, x = Group)) + geom\_boxplot() + labs(title="Average wage by group", x = "Group", y = "Average wage")

## Warning: Removed 1 rows containing non-finite values (stat\_boxplot).



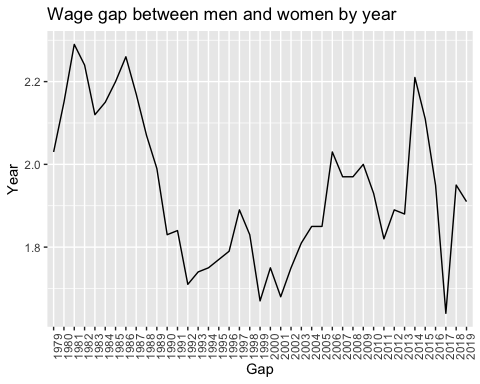
Again, the boxplot clearly shows a higher average wage for men. The distribution appears to be about equal for each group with men showing just a slightly higher distribution above the mean, indicating that there are more men that are high earners compared to other men, while women's wages are more evenly distributed.  
  
  
Now, in order to look at how the gap might differ based on ethnic group, I will calculate the gap between each subset of women - white, black and hispanic (I did not include Asian since there was a lot of missing date for that group) compared to the average of all men. I then created a new dataset that has a coloumn for the wage gap in each of those groups. Below is the summary of that dataset:

# Create variables to use to calculate and plot the wage gap  
year <- women$Year  
totalMen <- men$Total  
totalWomen <- women$Total  
whiteWomen <- women$'Total White'  
blackWomen <- women$'Total Black or African American'  
hispanicWomen <- women$'Total Hispanic or Latino ethnicity'  
  
gap <- totalMen - totalWomen  
gapWhite <- totalMen-whiteWomen  
gapBlack <- totalMen-blackWomen  
gapHispanic <- totalMen - hispanicWomen  
  
# Create a new dataframe for the wage gap information  
earningsGap <- data.frame(year, men, women, gap, gapWhite, gapBlack, gapHispanic)  
  
# Show summary stats of the new dataset  
knitr::kable(summary(earningsGap[, 1:10]), "pipe")

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | year | Year | Total | Total.White | Total.Black.or.African.American | Total.Asian | Total.Hispanic.or.Latino.ethnicity | Group | Year.1 | Total.1 |
|  | Length:41 | Length:41 | Min. : 5.65 | Min. : 5.79 | Min. : 4.890 | Length:41 | Min. : 4.790 | Length:41 | Length:41 | Min. : 3.620 |
|  | Class :character | Class :character | 1st Qu.: 8.10 | 1st Qu.: 8.28 | 1st Qu.: 7.070 | Class :character | 1st Qu.: 6.660 | Class :character | Class :character | 1st Qu.: 6.110 |
|  | Mode :character | Mode :character | Median :10.31 | Median :10.61 | Median : 9.770 | Mode :character | Median : 8.610 | Mode :character | Mode :character | Median : 8.640 |
|  | NA | NA | Mean :10.74 | Mean :10.98 | Mean : 9.482 | NA | Mean : 9.143 | NA | NA | Mean : 8.803 |
|  | NA | NA | 3rd Qu.:13.76 | 3rd Qu.:13.95 | 3rd Qu.:12.060 | NA | 3rd Qu.:11.810 | NA | NA | 3rd Qu.:11.760 |
|  | NA | NA | Max. :16.76 | Max. :17.10 | Max. :14.980 | NA | Max. :15.250 | NA | NA | Max. :14.850 |

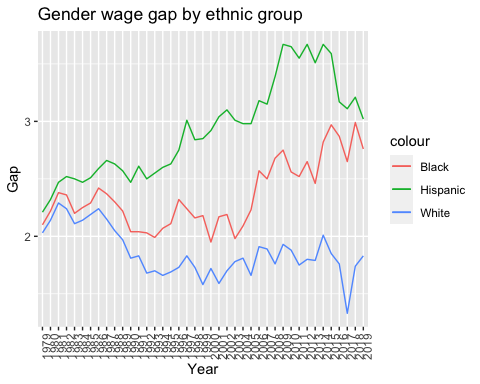
First I will look at a line graph of the gap between total men and women over time:

library (ggplot2)  
ggplot(data = earningsGap, aes(year, gap, group = 1)) + geom\_line() + labs(title="Wage gap between men and women by year", x = "Gap", y = "Year") + theme(axis.text.x = element\_text(angle = 90))



The graph shows us that the gap went down in the late 80's and early 90's but then had perios of increaseing and decreasing throughout the 90's and early 2000's. The gap hit a relatively high point around 2014, but then had it's lowest point in 2017.  
  
Next I will look at this same plot with the women divided into ethnic groups:

ggplot() + geom\_line(data = earningsGap, aes(year, gapWhite, group = 1, color = 'White')) + geom\_line(data = earningsGap, aes(year, gapBlack, group = 3, color = 'Black')) + geom\_line(data = earningsGap, aes(year, gapHispanic, group = 4, color = 'Hispanic')) + labs(title="Gender wage gap by ethnic group", x = "Year", y = "Gap") + theme(axis.text.x = element\_text(angle = 90))



This graph is interesting as it shows that the wage gap for white women has had some significant decreases while the gap for black and hispanic women has actually had larger increases, especially hispanic women.

# Save dataframes to csv  
  
write.csv(earningsStudy,"all\_earnings.csv")   
write.csv(men,"men\_earnings.csv")   
write.csv(women,"women\_earnings.csv")   
write.csv(total,"total\_earnings.csv")   
write.csv(earningsGap,"earnings\_gap.csv")

# Summary of findings

From this data we can see a couple of different things:  
  
As a whole, wages for both men and women have risen, and we can see a linear relationship between years and wage increase.  
  
Average wages for women have remained consistently lower than men, but there have been minor fluctuations in that gap that we can not explain with this data alone. It would be interesting to study employment trends in these years to see if there is a correlation with other things happening in the workforce.  
  
The wage gap seems to have improved slightly for white women, while getting worse for black and hispanic women. Again, this would be interesting to study other employement trends to see if there are any other factors that we could relate to this.  
  
Looking at the summary statistics for both men and women, we can see that the mean and max wages are higher for white individuals than other races, however, this gap is slightly larger for white men.