Analysis of Ontario wages based on Ontario Data Catalogue (1997-2019)

Borys Łangowicz (1010725967) Kian Dianati (1010205485)

2024-04-05

1 Description of the Data Set

1.1 Wages by education level

The `wages` data set includes the average weekly wages rates by education level and immigration status for Canada and Ontario in the years from 1997 to 2019. It includes the following columns:

```
## [1] "YEAR"            "Geography"       "Type.of.work"    "Wages"
## [5] "Education.level" "Age.group"       "Both.Sexes"      "Male"
## [9] "Female"
```

1. **`YEAR`**: Indicates the year in which the data was collected.
2. **`Geography`**: Indicates the region from which the data was collected. Its possible values include Canada as well as the Canadian provinces and territories.
3. **`Type.of.work`**: Indicates whether the data in the row is for full-time employees or part-time employees or both.
4. **`Wages`**:
    1. **`Total employees`**: The number of employees in the given age range, education level, and job status.
    2. **`Average hourly wage rate`**: The average hourly wage of the employees in the given age range, education level, and job status.
    3. And so on for **`Average weekly wage rate`**, **`Median hourly wage rate`**, and **`Median weekly wage rate`**.
5. **`Education.level`**: Indicates the level of education. It can include the following:

Education.level
Above bachelor's degree
Bachelor's degree
University certificate below bachelors degree
University degree
Community college, CEGEP
Trade certificate or diploma
Post-secondary certificate or diploma
Some post-secondary
High school graduate
Some high school
PSE (5,6,7,8,9))
No PSE (0,1,2,3,4)
0 - 8 years
Total, all education levels

6. `Age.group`: Indicates the age range of the individuals under consideration. It can include the following:

Age.group
25-64 years
25-54 years
25-34 years
20-34 years
15-24 years
55 years and over
25 years and over
15 years and over

7. `Both.sexes`: The data not seperated by gender.
8. `Male`: The data for males.
9. `Female`: The data for females.

1.2 Fuels price survey information

```
##  [1] "Date"                      "Ottawa"
##  [3] "Toronto.West"              "Toronto.East"
##  [5] "Windsor"                   "London"
##  [7] "Peterborough"              "St.Catharine"
##  [9] "Sudbury"                   "Sault.Saint.Marie"
## [11] "Thunder.Bay"               "North.Bay"
## [13] "Timmins"                   "Kenora"
## [15] "Parry.Sound"               "Ontario.Average"
## [17] "Southern.Average.Ontario" "Northern.Average.Ontario"
## [19] "Fuel.Type"
```

1. `Date`: Indicates the date on which the data was collected.
2. `Fuel Price`: Represents the price of fuel.
3. `Ottawa`, `Toronto.West`, `Toronto.East`, `Windsor`, `London`, `Peterborough`, `St.Catharine`, `Sudbury`, `Sault.Saint.Marie`, `Thunder.Bay`, `North.Bay`, `Timmins`, `Kenora`, `Parry.Sound`: Represents the fuel price in various locations in Ontario, Canada.
4. `Ontario.Average`: Indicates the average fuel price across different regions of Ontario.
5. `Southern.Average.Ontario`: Indicates the average fuel price across the southern regions of Ontario.
6. `Northern.Average.Ontario`: Indicates the average fuel price across the northern regions of Ontario.
7. `Fuel.Type`: Indicates the type of fuel associated with the data.

2 The Background of the Data

The labor and demographic dataset from the Ministry of Labour, Immigration, Training, and Skills Development provides insights into Ontario's workforce demographics, including age groups, employment types, educational levels, wages, and immigration statuses. It is annually updated and used by policymakers, researchers, and economists to inform decisions regarding education, training, workforce development, and immigration policies in the province.

Additionally, fuel price survey information from the Ministry of Energy offers weekly retail prices for gasoline, diesel, auto propane, and compressed natural gas across ten Ontario markets. This data aids in monitoring fuel price fluctuations and analyzing trends in the energy sector, supporting research and analysis efforts in economics, environmental studies, and energy policy.

3 Overall Research Question

Our research aims to explore wage dynamics in Ontario, considering economic factors. We're particularly interested in:

1. How has the average hourly wage rate changed over the years across different age groups?
2. How does the average hourly wage rate differ across various education levels?
3. Are there any trends in the employment rates based on different levels of education attainment?
4. How do wage rates vary across different age groups, and is there a trend in wage growth as individuals age?
5. Is there a significant gender wage gap, and how has it evolved over time?
6. How does educational attainment affect the gender wage gap?
7. How has the total number of employees changed over the years?

4 Analysis

4.1 How has the average hourly wage rate changed over the years across different age groups?

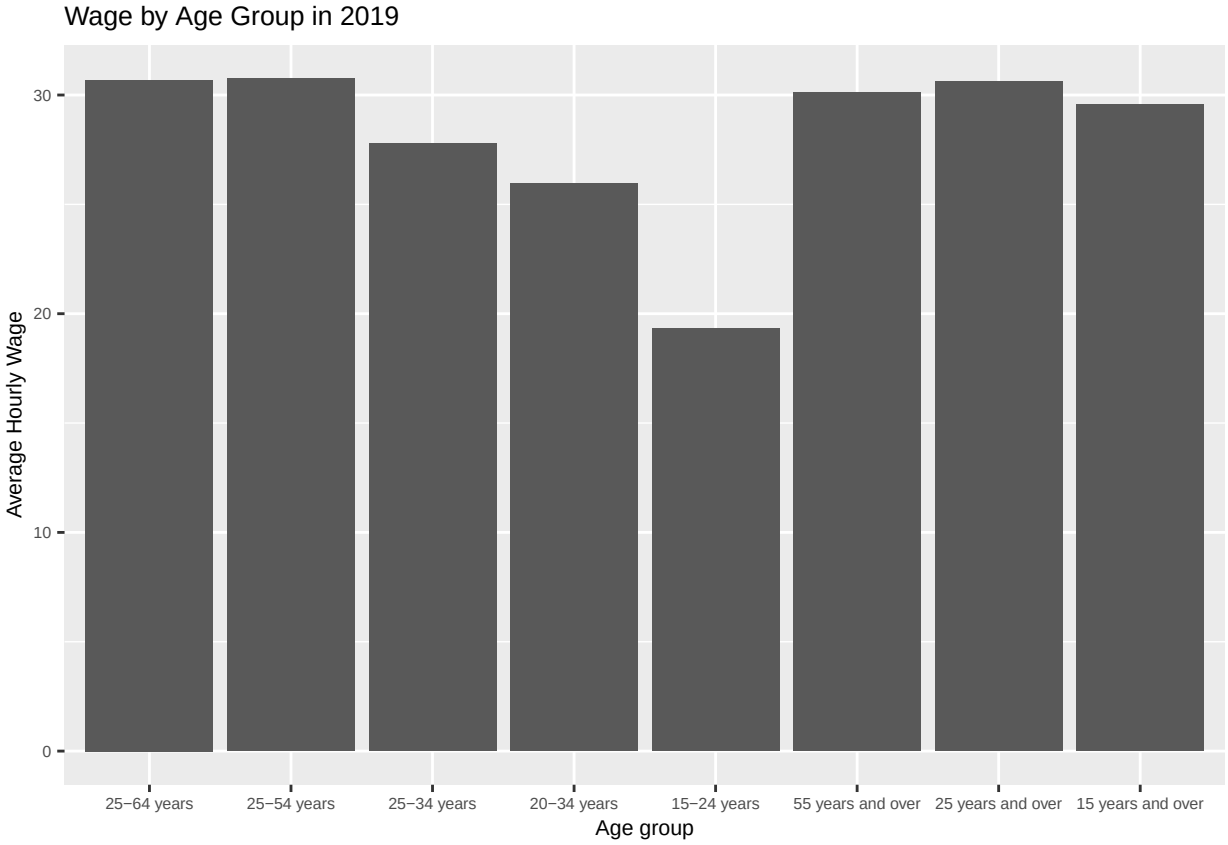
YEAR	25-64 years	25-54 years	25-34 years	20-34 years	15-24 years	55 years and over	25 years and over	15 years and over
1997	17.41	17.35	15.29	14.08	9.82	17.92	17.40	16.54
1998	17.64	17.57	15.66	14.41	10.06	18.15	17.62	16.76
1999	18.13	18.06	16.11	14.77	10.33	18.65	18.11	17.19
2000	18.69	18.63	16.72	15.31	10.81	19.04	18.67	17.72
2001	19.28	19.23	17.48	15.97	11.21	19.63	19.27	18.29
2002	19.87	19.80	17.94	16.34	11.37	20.19	19.84	18.83
2003	20.28	20.19	18.15	16.55	11.66	20.85	20.26	19.24
2004	20.80	20.70	18.54	16.88	11.77	21.31	20.77	19.71
2005	21.44	21.32	19.25	17.50	12.23	21.98	21.41	20.32
2006	22.12	22.03	19.94	18.15	12.84	22.56	22.10	20.99
2007	22.91	22.84	20.77	18.94	13.40	23.07	22.87	21.74
2008	23.86	23.77	21.69	19.81	14.06	24.16	23.83	22.67
2009	24.60	24.55	22.32	20.47	14.57	24.63	24.56	23.48
2010	25.09	24.98	22.70	20.86	14.76	25.39	25.04	23.97
2011	25.54	25.45	23.16	21.29	15.13	25.73	25.50	24.42
2012	26.27	26.20	23.92	21.98	15.50	26.31	26.22	25.13
2013	26.85	26.76	24.35	22.36	15.86	27.02	26.81	25.70
2014	27.33	27.29	24.89	22.87	16.20	27.25	27.28	26.17
2015	28.10	28.14	25.64	23.52	16.53	27.57	28.03	26.88
2016	28.69	28.66	26.18	24.05	16.81	28.48	28.62	27.47
2017	29.12	29.16	26.55	24.41	17.01	28.49	29.03	27.88
2018	29.75	29.75	27.10	25.07	17.96	29.38	29.68	28.56
2019	30.69	30.75	27.79	25.97	19.32	30.12	30.62	29.56

This table shows the average wage of each age group in dollars per hours for the given years. For instance, in the year 1997 those who were between 15 and 24 years of age made 9 dollars and 82 cents per hour of work while those who were 25 to 34 years old made 15 dollars and 29 cents.

This table indicates that there has been a slow but steady growth in wages accross all age groups over the years from 1997 to 2019. Moreover, it appears that older age groups consistently earn higher wages than those in younger ones. For instance, in 2019 the average wage for those who were older than 55 was 30 dollars and 12 cents while those who were between 15 and 24 made only 19 dollars and 32 cents. However,

the average wage of the oldest age group, 55+, was not much higher than the average wage of the total population—which in 2019 was 29.56 dollars per hour. This suggests that the number of employees between 15 and 24 years of age is insignificant relative to older age groups.

These results can also be seen in the following bar chart:



4.2 How does the average hourly wage rate differ across various education levels for different genders?

Table 4: Average wage by education level in male and female employees

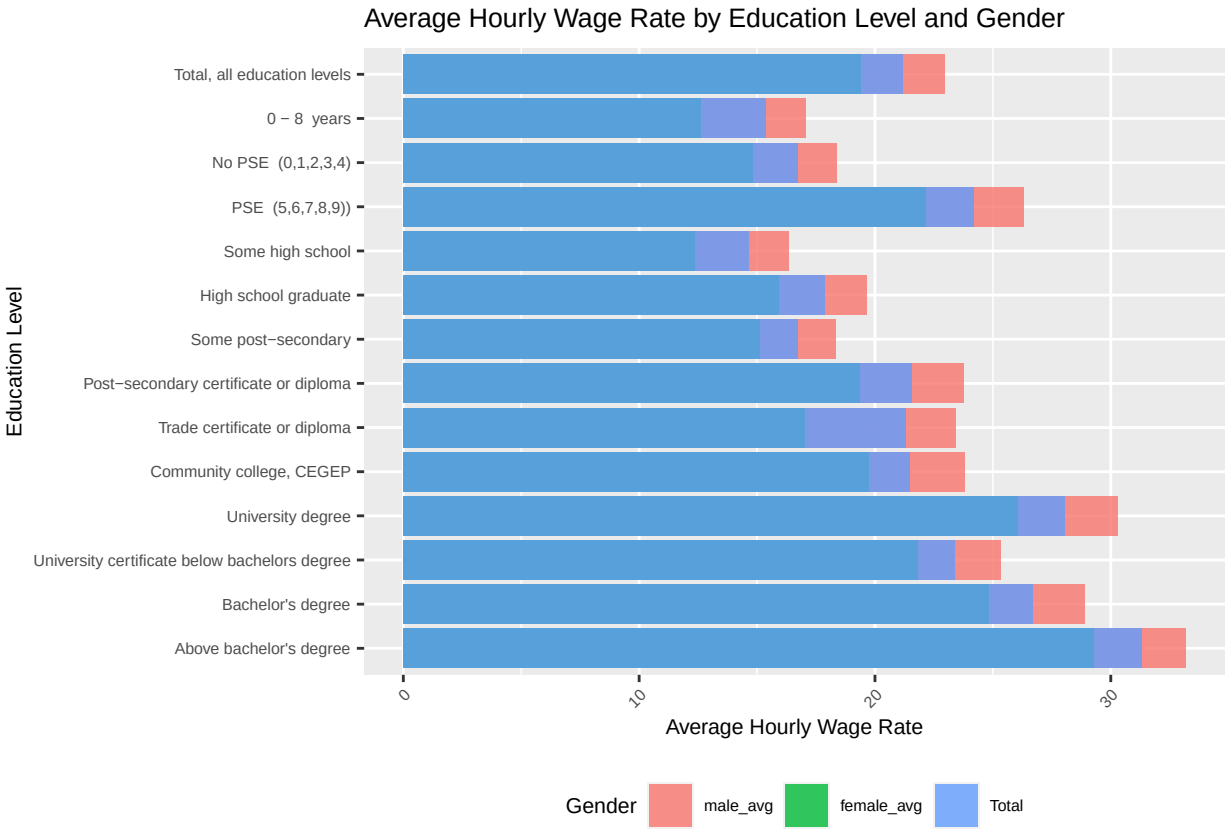
Education Level	Male	Female	Total	Pay gap
Above bachelor's degree	33.18565	29.29043	31.33478	3.895217
Bachelor's degree	28.89087	24.83000	26.68000	4.060870
University certificate below bachelors degree	25.33957	21.80435	23.37130	3.535217
University degree	30.29478	26.04565	28.06826	4.249130
Community college, CEGEP	23.81000	19.74957	21.46739	4.060435
Trade certificate or diploma	23.44783	17.01913	21.31000	6.428696
Post-secondary certificate or diploma	23.75652	19.35348	21.55739	4.403043
Some post-secondary	18.32478	15.12348	16.72391	3.201304
High school graduate	19.64739	15.93130	17.86696	3.716087
Some high school	16.34565	12.35565	14.66000	3.990000
PSE (5,6,7,8,9))	26.32174	22.16043	24.19913	4.161304
No PSE (0,1,2,3,4)	18.38304	14.83043	16.73391	3.552609
0 - 8 years	17.05087	12.62870	15.36783	4.422174

Education Level	Male	Female	Total	Pay gap
Total, all education levels	22.94130	19.38304	21.18957	3.558261

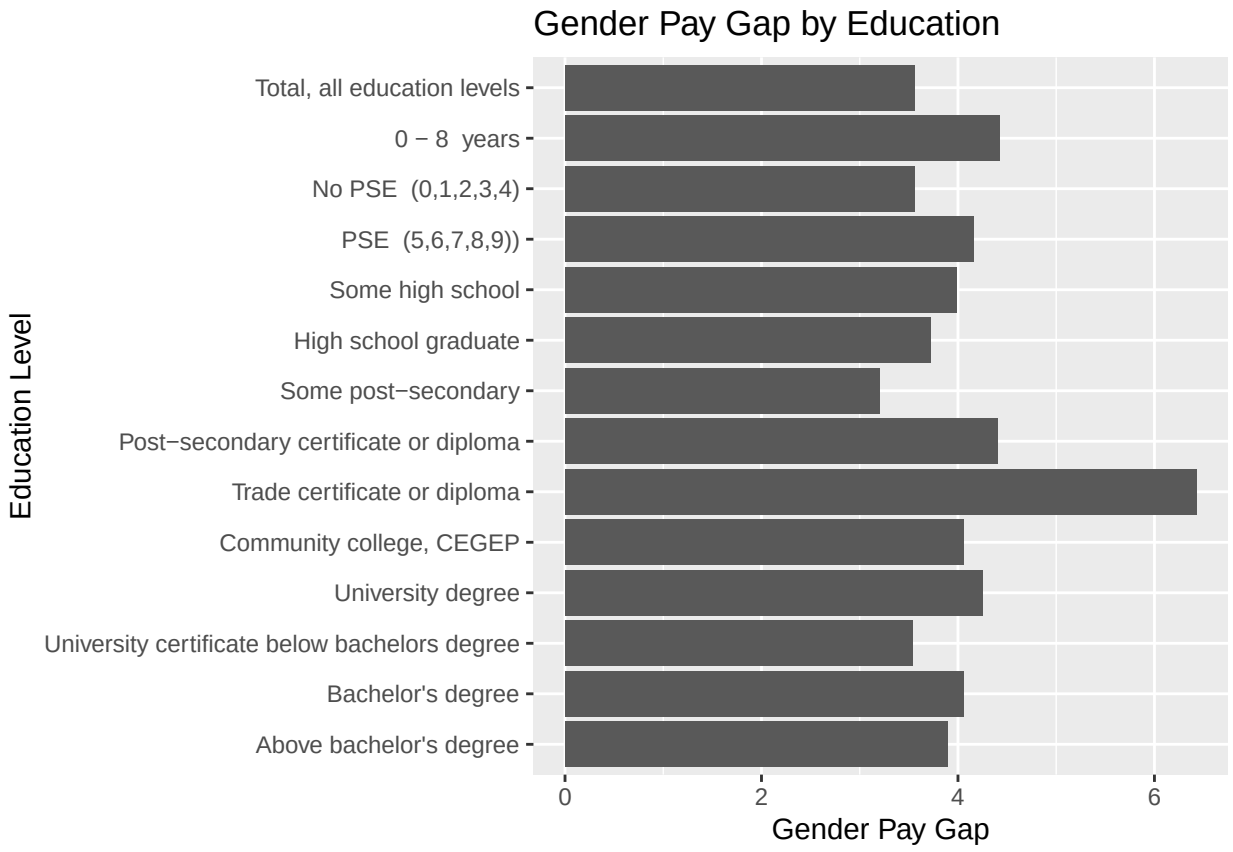
Conlusions: This table indicates the average hourly wage of employees based on their education level. It also indicates the pay based on the gender of the employee. For instance, a male with a bachelor's degree makes, on average, 28 dollars an hour while a female makes 24 dollars resulting in a 4 dollar pay gap.

From this table and the following graph, it is clear that with higher educational attainment the average wage rises. For example, an individual with a high school diploma makes about 14 dollars an hour while someone with a bachelor's degree makes 26 dollars.

```
## Using Education.level as id variables
```



Conlusions: A gender pay gap is observed, with males making about 3 dollars more on average than their female counterparts. An surprising result is that the gender pay gap doesn't seem to changewith education level. In the previous example, males with a high school degree made 3.99 dollars more than their female peers, while males with bachelor's degrees 4.06 dollars—only a 7 cent increase. This may also be observed in the following plot.

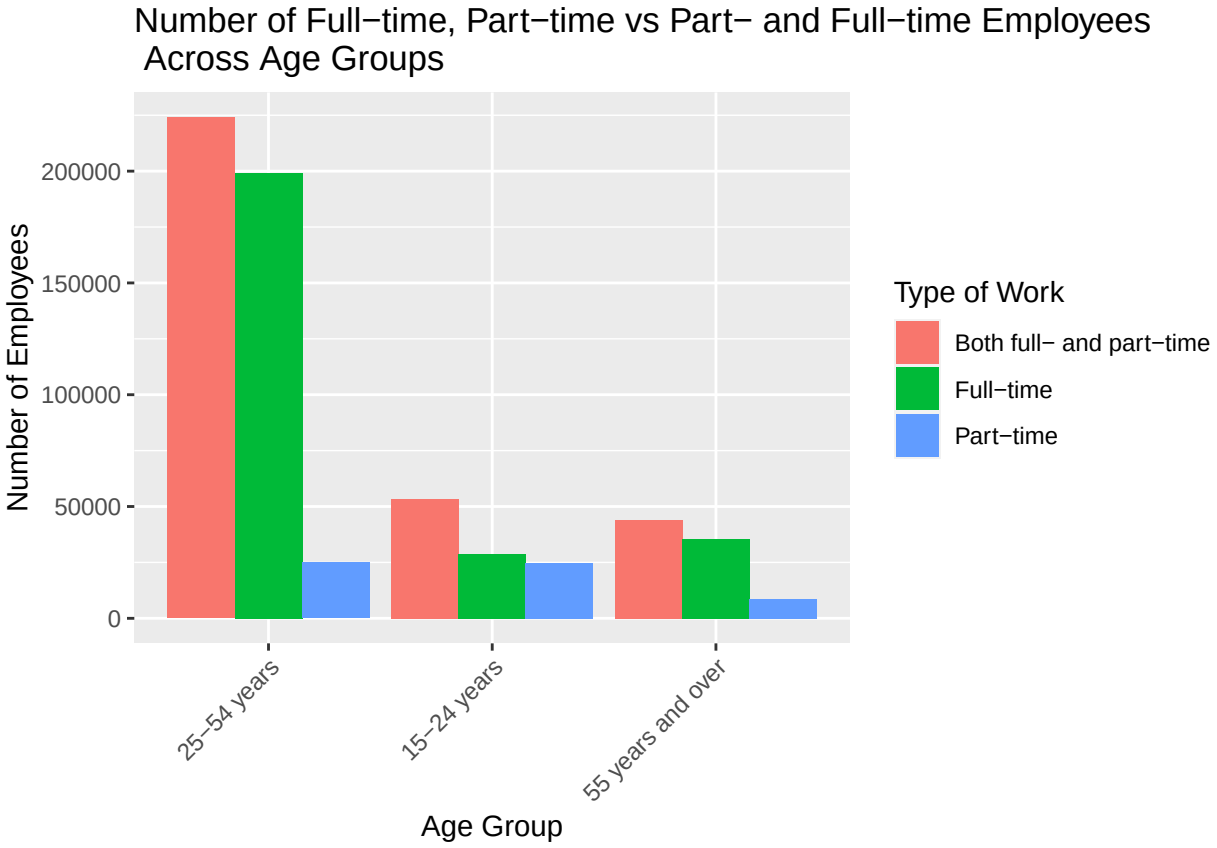


Conlusions:

Interestingly, the largest gender gap was observed in those with a trade certificate, where the pay gap is almost twice as large as in other education levels.

These observations highlight disparities in wages between genders across different education levels, indicating the presence of gender-based wage inequality.

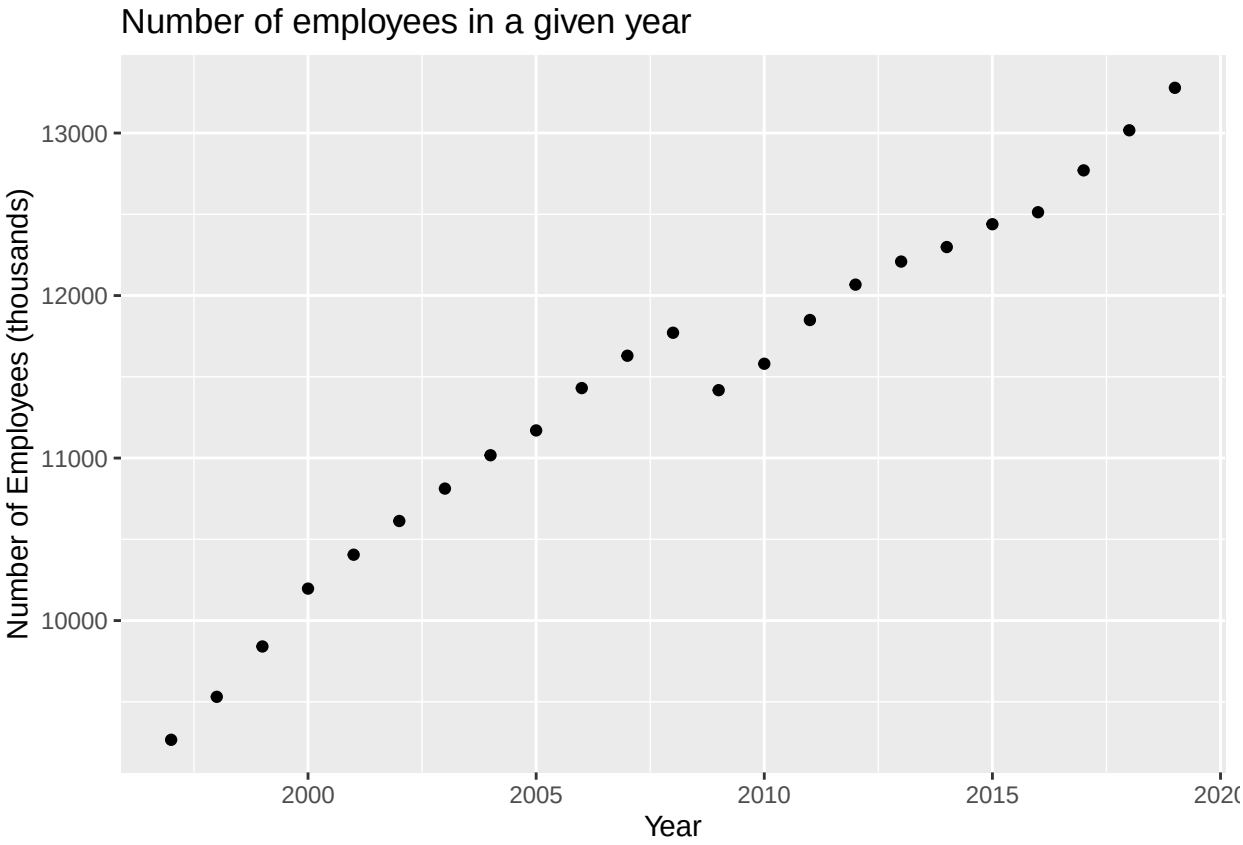
4.3 What is the overall trend in the number of full-time employees versus part-time employees across different age groups?



Conlusions:

From this plot, we can see that the number of full-time employees is consistently larger than the number of part-time employees across all age groups. In fact, this plot confirms our hypothesis from before, as we can see that the number of employed 15-24 year olds is much smaller than the number of employed 25-54 year-olds—who have the highest employment out of any age group. This data suggests that as individuals grow older, they tend to favor part-time employment. Further analysis, such as examining trends over time or considering demographic factors, may provide additional insights into the employment dynamics across different age groups.

4.4 How has the total number of employees changed over the years?



Conclusions:

This plot displays the total number of employees in a given year. For instance, in the year 2004 there were about 11 million employees in Canada. An interesting feature of this graph is how it captures the 2008 financial crisis, where almost 1 million Canadianas lost their jobs. What's even more interesting is that no decrease in wages was observed amongst those who kept their jobs, as was shown previously.

5 Hypothesis Testing

Null Hypothesis (H0): There is no statistically significant distinction in the average wages between male and female workers ($\mu_{\text{male}} = \mu_{\text{female}}$).

Alternative Hypothesis (H1): There exists a statistically significant disparity in average wages between male and female workers ($\mu_{\text{male}} \neq \mu_{\text{female}}$). To examine this hypothesis, we can employ a two-sample t-test to compare the wage distributions of male and female workers. This entails segregating the dataset into two distinct groups based on gender: male and female.

The resultant p-value derived from the selected statistical test indicates the probability of observing a wage difference as extreme as, or more extreme than, the observed difference, assuming the null hypothesis holds. If the obtained p-value falls below the predetermined significance level (typically 0.05), we reject the null hypothesis, indicating a significant discrepancy in wages between male and female workers.

Moreover, by computing a confidence interval for the disparity in mean wages, we can gauge the plausible range of values for the actual difference between male and female wages.

```
##
##  Welch Two Sample t-test
```

```
##
## data:  male_wages and female_wages
## t = 69.11, df = 155915, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  116.8124 123.6314
## sample estimates:
## mean of x mean of y
##  627.8701  507.6481
```

Conclusions

The outcomes derived from the Welch Two Sample t-test underscore a significant contrast in average earnings between male and female workers. With an extraordinarily minute p-value (< 0.00000000000000022), there's compelling evidence to dismiss the null hypothesis, indicating that the mean wages for men and women diverge significantly. The 95% confidence interval for the mean wage discrepancy extends from 116.8124 to 123.6314, suggesting that the actual difference in wages between male and female workers likely lies within this interval.

In summary, the findings overwhelmingly advocate for rejecting the null hypothesis, underscoring a noteworthy disparity in wages between male and female employees. Specifically, the average wage for male workers (627.8701) markedly exceeds that of female workers (507.6481).

6 Bootstrapping

We utilize bootstrapping to estimate the sampling distribution of the mean fuel price for Toronto West and to compute a confidence interval.

```
## Mean fuel price for Ontario: 80.06227
```

```
## 95% confidence interval: (  79.26425 - 80.78438  )
```

According to our findings, the mean fuel price for Ontario stands at 80.0622744. Our computed 95% confidence interval spans from 79.2642523 to 80.7843823‘. This suggests with 95% confidence that the genuine mean fuel price for Ontario lies within this interval.

7 Random Forest

```
##
## Call:
##  randomForest(formula = wage ~ Education.level + gender + Age.group,      data = train_data, ntree =
##                Type of random forest: regression
##                      Number of trees: 10
## No. of variables tried at each split: 1
##
##           Mean of squared residuals: 20.54053
##                     % Var explained: 52.07
```

Thus, this model has a RSS of 20.54053 dollars squared, which is quite large compared to the average of the numbers we are trying to predict. This means that the model is quite unreliable at correctly predicting the average income of a group based on demographic factors. This is also reflected in the low percentage of explained variablity. This model only explains 52.07 percent of the variability in this data, which is not particularly great, indicating that the relationship between income and other demographic factors is hard to explain.

We are analyzing wage data from Canada, specifically focusing on the average hourly wage rates for different demographic groups. After preprocessing the data to filter out irrelevant categories and reshape it into a

suitable format, we split the dataset into training and testing sets using a stratified sampling approach.

Next, we employ a random forest regression model to predict wages based on several predictor variables, including education level, gender, and age group. The random forest algorithm is chosen for its ability to handle complex relationships between predictors and response variables, as well as its robustness to overfitting.

Explanation:

The model's accuracy is assessed using metrics such as the mean squared residuals and the percentage of variation explained (Var explained). The mean squared residuals measure the average discrepancy between the observed and predicted wage rates, with a lower value indicating a better fit of the model to the data. In this case, the mean of squared residuals is 18.7288.

The percentage of variation explained (Var explained) provides an indication of how well the model accounts for the variability in the wage rates. A higher percentage suggests that the predictor variables included in the model collectively explain a larger proportion of the variability in the data. Here, the model explains 56.15% of the variation in the wage rates, indicating a moderate level of explanatory power.

8 Cross Validation

```
## [1] 17.70338
```

**The MSE of around 17.7033755 shows the average difference between actual and predicted wages. Such a value is considered significant, suggesting that the model may not generalize well to new data. This indicates that the relationship between education level, gender, age group, and wage is not straightforward or obvious.

9 Summary of Research

Based on the research conducted, the key findings are:

9.1 Wage Trends Across Age Groups:

1. Wages have steadily increased from 1997 to 2019.
2. Older age groups earn higher wages, but the disparity isn't significant compared to the overall average.

9.2 Education Level and Gender Wage Differences:

1. Higher education correlates with higher wages.
2. A consistent gender wage gap exists across all education levels, with males earning more.
3. The gap remains consistent, indicating persistent gender-based wage inequality.

9.3 Employment Trends Across Age Groups:

1. Full-time employment is predominant across age groups.
2. Part-time employment increases with age.

9.4 Total Employee Numbers Over Time:

1. Employee numbers fluctuate over time, with significant drops during events like the 2008 financial crisis.
2. Wages remain stable for those employed despite fluctuations.

9.5 Gender Wage Disparity Analysis:

1. Welch Two Sample t-test shows a significant difference in average earnings between genders.
2. The test rejects the null hypothesis, confirming a notable wage gap.
3. The 95% confidence interval suggests a substantial range, indicating gender-based wage inequality.

In summary, the research highlights persistent wage disparities across age, education, and gender, underscoring the need for interventions to address these inequalities in the workforce.

10 Appendix

1. **Importing libraries**

```
library(tidyverse)
library(ggplot2)
library(reshape2)
library(knitr)
library(randomForest)
library(dplyr, warn.conflicts = FALSE)
options(dplyr.summarise.inform = FALSE)
```

2. **Loading the data**

```
wages = read.csv("wages.csv") %>% mutate_if(is.character, str_trim)

wages$Education.level =
  factor(wages$Education.level,
         levels = c("Above bachelor's degree",
                    "Bachelor's degree",
                    "University certificate below bachelors degree",
                    "University degree",
                    "Community college, CEGEP",
                    "Trade certificate or diploma",
                    "Post-secondary certificate or diploma",
                    "Some post-secondary",
                    "High school graduate",
                    "Some high school",
                    "PSE  (5,6,7,8,9))",
                    "No PSE  (0,1,2,3,4)",
                    "0 - 8  years",
                    "Total, all education levels"),
         ordered = TRUE)

wages$Age.group =
  factor(wages$Age.group,
         levels = c("25-64 years",
                    "25-54 years",
                    "25-34 years",
                    "20-34 years",
                    "15-24 years",
                    "55 years and over",
                    "25 years and over",
                    "15 years and over"),
         ordered = TRUE)

# Fuel
fuel = read.csv("fuel.csv") %>% mutate_if(is.character, str_trim)

fuel <- fuel %>% rename(
  Toronto.West = Toronto.West.Ouest,
  Toronto.East = Toronto.East.Est,
  St.Catharine = St..Catharine.s,
  Ontario.Average = Ontario.Average.Moyenne.provinciale,
  Southern.Average.Ontario = Southern.Average.Moyenne.du.sud.de.l.Ontario,
```

```
  Northern.Average.Ontario = Northern.Average.Moyenne.du.nord.de.l.Ontario
) %>% select(!(Type.de.carburant))
```

3. Description of the Data Set

```
names(wages)
```

```
kable(wages %>%
  group_by(Education.level) %>%
  reframe(Education.level) %>%
  unique())
```

```
kable(wages %>%
  group_by(Age.group) %>%
  reframe(Age.group) %>%
  unique())
```

```
names(fuel)
```

4. Analysis

```
avg_wage_by_age <- wages %>%
  filter(Wages == "Average hourly wage rate",
         Geography == "Canada",
         Type.of.work == "Full-time",
         Education.level == "Total, all education levels") %>%
  select(YEAR, Age.group, Both.Sexes) %>%
  group_by(YEAR, Age.group) %>%
  summarise(Avg_Hourly_Wage = mean(Both.Sexes))

kable(avg_wage_by_age %>%
  pivot_wider(names_from = Age.group, values_from = Avg_Hourly_Wage))
```

```
ggplot(avg_wage_by_age %>%
         filter(YEAR == "2019"),
       aes(x = Age.group, Avg_Hourly_Wage)) +
  geom_bar(stat = "identity") +
  labs(x = "Age group",
       y = "Average Hourly Wage",
       title = "Wage by Age Group in 2019") +
  theme(text = element_text(size = 8))
```

```
avg_wage_by_education <- wages %>%
  filter(Wages == "Average hourly wage rate",
         Geography == "Canada",
         Age.group == "15 years and over",
         # YEAR == '2019',
         Type.of.work == "Both full- and part-time") %>%
  select(Education.level, Male, Female, Both.Sexes) %>%
  group_by(Education.level) %>%
  summarise(male_avg = mean(Male),
            female_avg = mean(Female),
            Total = mean(Both.Sexes))

kable(avg_wage_by_education %>%
  mutate(pay.gap = male_avg - female_avg),
```

```
  col.names = c("Education Level", "Male", "Female", "Total", "Pay gap"),
  caption = "Average wage by education level in male and female employees")
```

```
ggplot(avg_wage_by_education %>% melt(),
       aes(x = Education.level, fill = variable)) +
  geom_bar(aes(y = value),
           stat = "identity",
           alpha = 0.8,
           show.legend = TRUE,
           position = position_identity()) +
  labs(title = "Average Hourly Wage Rate by Education Level and Gender",
       x = "Education Level",
       y = "Average Hourly Wage Rate",
       fill = "Gender") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        text = element_text(size = 8),
        legend.position = "bottom") +
  coord_flip()
```

```
## Using Education.level as id variables
```

```
ggplot(avg_wage_by_education %>% mutate(pay.gap = male_avg - female_avg),
       aes(x = Education.level, y = pay.gap)) +
  geom_bar(stat = "identity") +
  labs(y = "Gender Pay Gap",
       x = "Education Level",
       title = "Gender Pay Gap by Education") +
  coord_flip()
```

```
full_time_part_time <- wages %>%
  filter(Age.group %in% c("15-24 years",
                          "25-54 years",
                          "55 years and over"),
         Geography == 'Canada',
         Education.level == 'Total, all education levels',
         Wages == "Total employees") %>%
  group_by(Age.group, Type.of.work) %>%
  summarise(Employees = sum(Both.Sexes))

options(scipen = 999)
ggplot(full_time_part_time,
       aes(x = Age.group, y = Employees, fill = Type.of.work)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Number of Full-time, Part-time vs Part- and Full-time Employees\n Across Age Groups",
       x = "Age Group",
       y = "Number of Employees",
       fill = "Type of Work") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
employee.year = wages %>%
  filter(Wages == "Total employees",
         Geography == "Canada",
         Type.of.work == "Full-time",
         Education.level == "Total, all education levels",
         Age.group == '15 years and over') %>%
```

```
  select(YEAR, Both.Sexes)

ggplot(employee.year, aes(x = YEAR, y = Both.Sexes)) +
  geom_point() +
  labs(title = "Number of employees in a given year",
       x = "Year",
       y = "Number of Employees (thousands)")
```

5. **Hypothesis Testing**

```
male_wages <- wages %>% filter(Wages == "Average weekly wage rate") %>% select(Male)
female_wages <- wages %>% filter(Wages == "Average weekly wage rate") %>% select(Female)

t.test(male_wages, female_wages)
```

6. **Bootstrapping**

```
library(boot)

mean_fuel_price <- function(data) {
  mean(data[["Toronto.West"]])
}

set.seed(123)
n_boot <- 1000
bootstrap_means <- replicate(n_boot, {
  sample_data <- fuel[sample(1:nrow(fuel), replace = TRUE), ]
  mean_fuel_price(sample_data)
})

ci <- quantile(bootstrap_means, c(0.025, 0.975))

# Print the results
cat("Mean fuel price for Ontario:", mean(fuel$Toronto.West), "\n")
```

```
## Mean fuel price for Ontario: 80.06227
```

```
cat("95% confidence interval:", "( ", ci[1], "-", ci[2], " )", "\n")
```

```
## 95% confidence interval: (  79.26425 - 80.78438  )
```

7. **Random Forest**

```
d = wages %>%
  select(!Both.Sexes) %>%
  filter(Wages == 'Average hourly wage rate',
         Type.of.work == 'Both full- and part-time',
         Geography == 'Canada',
         Education.level != 'Total, all education levels',
         Age.group != '15 years and over') %>%
  pivot_longer(c(Male, Female),
               names_to = "gender",
               values_to = "wage")

d = d %>%
  mutate(group = sample(c('train', 'test'),
                        size = nrow(d),
```

```
                        prob = c(0.9, 0.1),
                        replace=TRUE))

train_data = d %>% filter(group == 'train')
test_data = d %>% filter(group == 'test')

model = randomForest(wage ~ Education.level + gender + Age.group,
                     data = train_data,
                     ntree = 10,
                     importance = TRUE)
```

8. **Cross Validation**

```
test_data$predicted = predict(model, newdata = test_data)

mean((test_data$wage - test_data$predicted)^2)
```