

Analysis of Ontario wages in relation to economic factors based on Ontario Data Catalogue (1997-2019)

Borys Łangowicz (1010725967)

Kian Dianati (1010205485)

2024-04-05

1 Loading the Data

We will use the following data sets:

```
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))
```

2 Description of the Data set

2.1 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:

```
names(wages)
```

```
## [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.

2.2 Fuels price survey information

```
names(fuel)
```

```
## [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.

3 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.

4 Overall Research Question

4.1 Trend Analysis

- How has the average hourly wage rate changed over the years across different age groups?
- Are there any noticeable trends in the median weekly wage rate for full-time employees over the past decade?
- What is the overall trend in the number of full-time employees versus part-time employees across different age groups?

4.2 Regional Disparities

- How do average hourly wage rates vary between different Canadian provinces and territories?
- Are there significant differences in the employment rates between urban and rural areas within a specific province?

4.3 Educational Attainment

- How does the average hourly wage rate differ across various education levels?
- Are there any trends in the employment rates based on different levels of education attainment?
- Is there a correlation between educational attainment and the likelihood of being employed full-time versus part-time?

4.4 Age Groups Analysis

- How do wage rates vary across different age groups, and is there a trend in wage growth as individuals age?
- Are there noticeable differences in employment rates between younger and older age groups?
- What is the distribution of educational attainment among different age groups, and how does it correlate with employment status and wage rates?

4.5 Gender Analysis

- Is there a significant gender wage gap, and how has it evolved over time?
- Are there differences in the distribution of employment types (full-time vs. part-time) between males and females?
- How does educational attainment affect the gender wage gap within specific age groups or regions?

4.6 Overall Employment Trends

- How has the total number of employees changed over the years?
- Are there seasonal variations in employment rates or wage rates within certain industries?
- What industries or sectors have shown the highest growth in employment rates, and how does this correlate with wage rates?

5 Summary

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

```
avg_wage_by_age <- wages %>%  
  filter(Wages == "Average hourly wage rate",  
         Geography == "Canada",  
         Type.of.work == "Full-time",
```

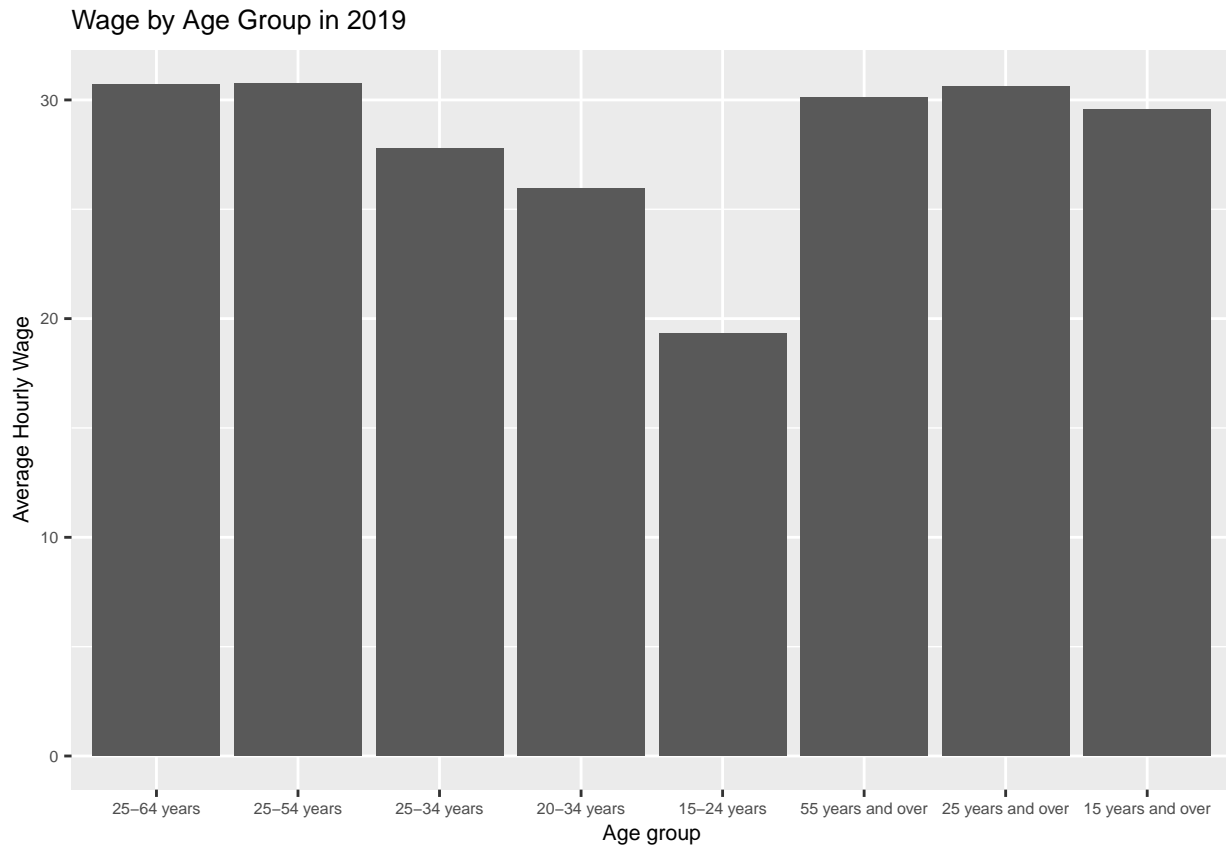
```
Education.level == "Total, all education levels") %>%
select(YEAR, Age.group, Both.Sexes) %>%
group_by(YEAR, Age.group) %>%
summarise(Avg_Hourly_Wage = mean(Both.Sexes))
```

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

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

| 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 |

```
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))
```



From the data, we can observe the following trends:

1. Across all age groups, there is a general trend of increasing average hourly wage rates over the years.
2. The wage rates tend to increase with age, with the highest rates typically observed in the 55 years and over age group.
3. The 15-24 years age group consistently has the lowest average hourly wage rates, which gradually increase as individuals move into older age groups.
4. There is a noticeable increase in wage rates during the late 1990s and early 2000s, followed by a slight dip during the economic downturn of 2008-2009, after which wages steadily increased again.

These observations provide an overview of how the average hourly wage rates have changed over the years across different age groups.

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

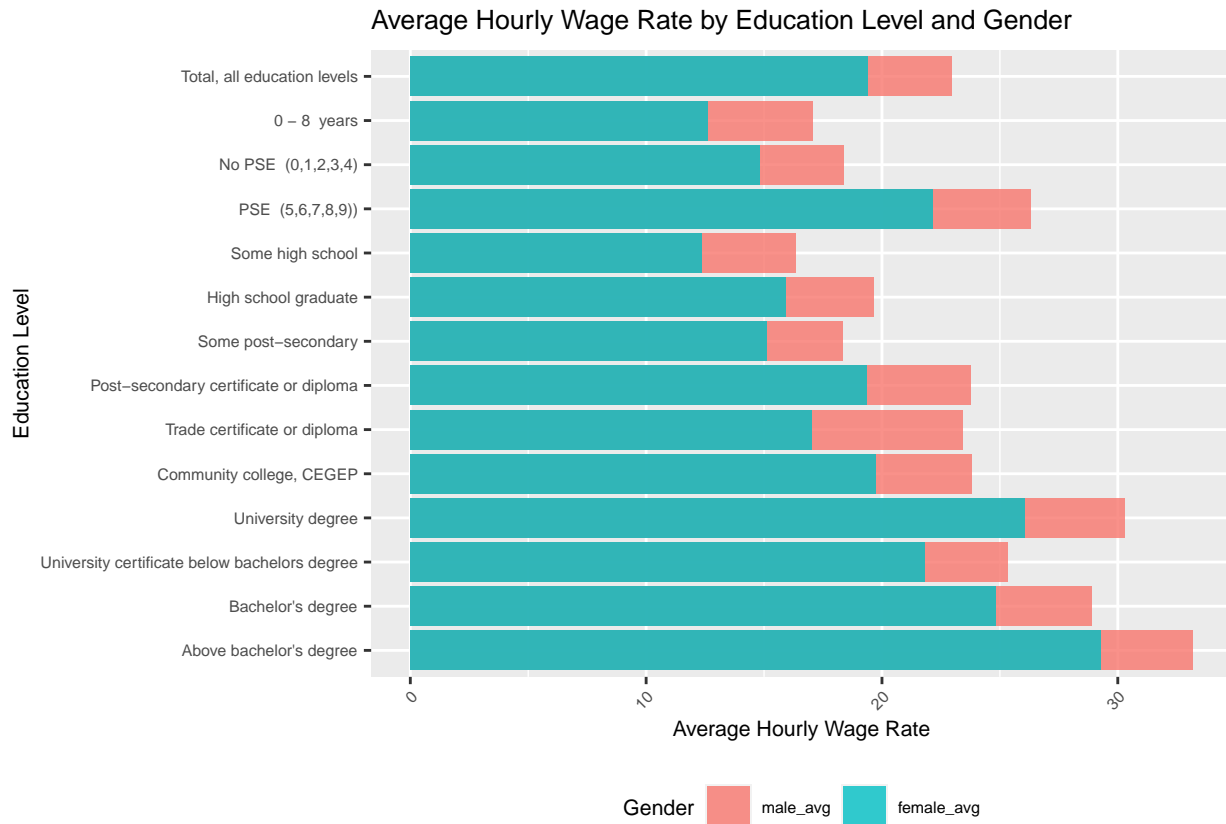
```
avg_wage_by_education <- wages %>%
  filter(Wages == "Average hourly wage rate",
         Geography == "Canada",
         Age.group == "15 years and over",
         Type.of.work == "Both full- and part-time") %>%
  select(Education.level, Male, Female) %>%
  group_by(Education.level) %>%
  summarise(male_avg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

kable(avg_wage_by_education)
```

| Education.level | male_avg | female_avg |
|---|----------|------------|
| Above bachelor's degree | 33.18565 | 29.29043 |
| Bachelor's degree | 28.89087 | 24.83000 |
| University certificate below bachelors degree | 25.33957 | 21.80435 |
| University degree | 30.29478 | 26.04565 |
| Community college, CEGEP | 23.81000 | 19.74957 |
| Trade certificate or diploma | 23.44783 | 17.01913 |
| Post-secondary certificate or diploma | 23.75652 | 19.35348 |
| Some post-secondary | 18.32478 | 15.12348 |
| High school graduate | 19.64739 | 15.93130 |
| Some high school | 16.34565 | 12.35565 |
| PSE (5,6,7,8,9)) | 26.32174 | 22.16043 |
| No PSE (0,1,2,3,4) | 18.38304 | 14.83043 |
| 0 - 8 years | 17.05087 | 12.62870 |
| Total, all education levels | 22.94130 | 19.38304 |

```
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()
```

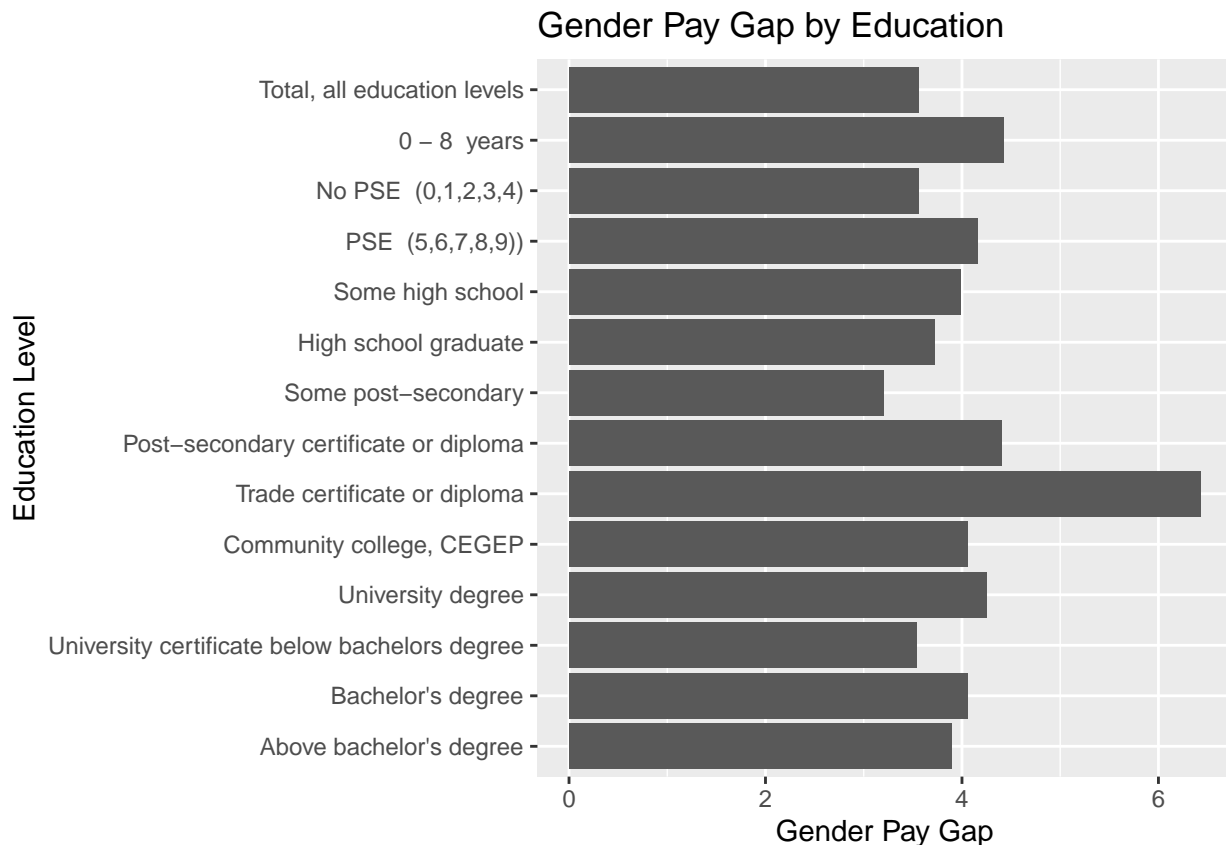
```
## Using Education.level as id variables
```



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

| Education.level | male_avg | female_avg | pay.gap |
|---|----------|------------|----------|
| Above bachelor's degree | 33.18565 | 29.29043 | 3.895217 |
| Bachelor's degree | 28.89087 | 24.83000 | 4.060870 |
| University certificate below bachelors degree | 25.33957 | 21.80435 | 3.535217 |
| University degree | 30.29478 | 26.04565 | 4.249130 |
| Community college, CEGEP | 23.81000 | 19.74957 | 4.060435 |
| Trade certificate or diploma | 23.44783 | 17.01913 | 6.428696 |
| Post-secondary certificate or diploma | 23.75652 | 19.35348 | 4.403043 |
| Some post-secondary | 18.32478 | 15.12348 | 3.201304 |
| High school graduate | 19.64739 | 15.93130 | 3.716087 |
| Some high school | 16.34565 | 12.35565 | 3.990000 |
| PSE (5,6,7,8,9)) | 26.32174 | 22.16043 | 4.161304 |
| No PSE (0,1,2,3,4) | 18.38304 | 14.83043 | 3.552609 |
| 0 - 8 years | 17.05087 | 12.62870 | 4.422174 |
| Total, all education levels | 22.94130 | 19.38304 | 3.558261 |

```
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()
```

Observations:

1. In almost all cases, the average hourly wage for males is higher than for females across various education levels.
2. The largest wage gaps are observed at “Trade certificate or diploma” where
3. At higher education levels, such as “Above bachelor’s degree” and “Bachelor’s degree”, the wage gap is relatively smaller compared to lower education levels but still exists.

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

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

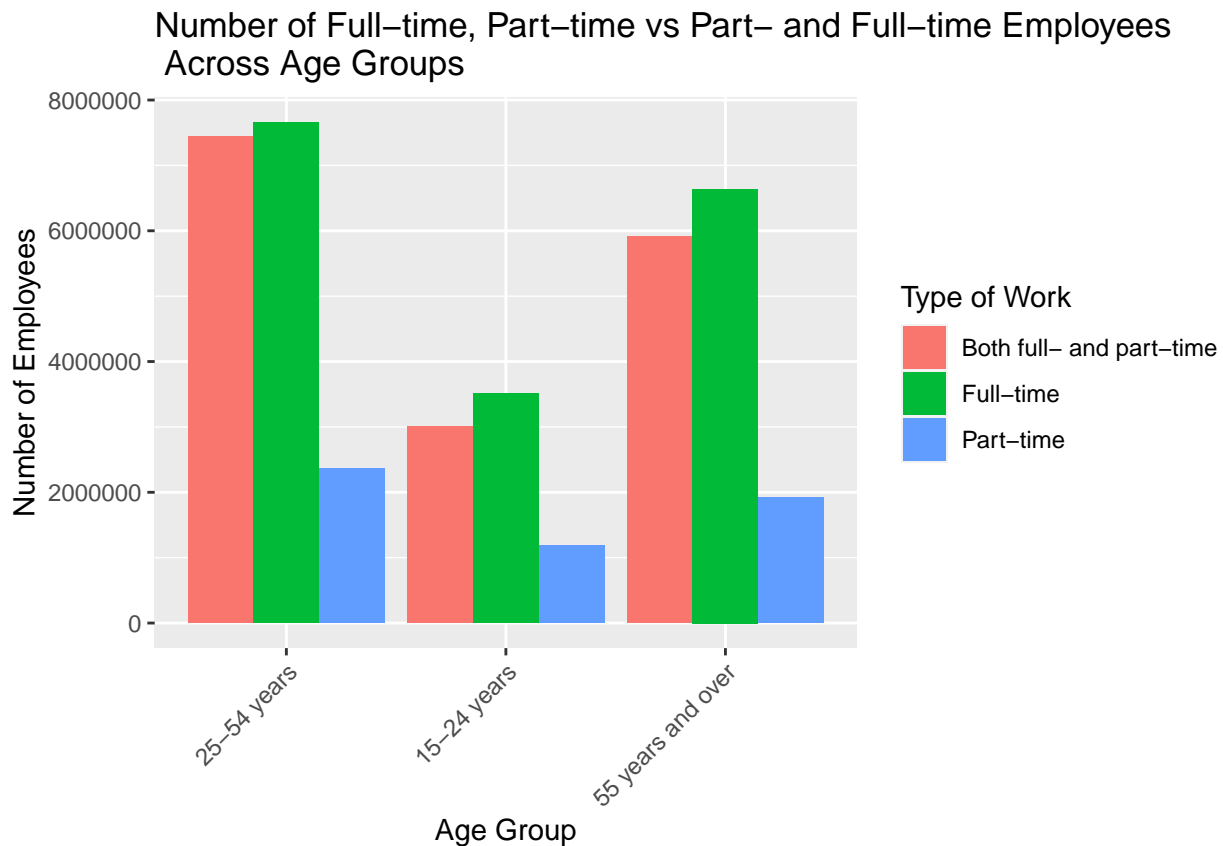
```
unique(wages$Type.of.work)
```

```
## [1] "Both full- and part-time" "Full-time"
## [3] "Part-time"
```

```
full_time_part_time <- wages %>%
  filter(Age.group %in% c("15-24 years",
                        "25-54 years",
                        "55 years and over")) %>%
  group_by(Age.group, Type.of.work) %>%
  summarise(Employees = sum(Both.Sexes))
```

```
## `summarise()` has grouped output by 'Age.group'. You can override using the
## `.groups` argument.
```

```
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))
```



5.4 What is the relationship between Wages and Fuel Price?

```
fuel.year = fuel %>%
  mutate(Year = year(as.Date(Date))) %>%
  group_by(Year) %>%
  summarise(Avg.Price = mean(Ontario.Average))

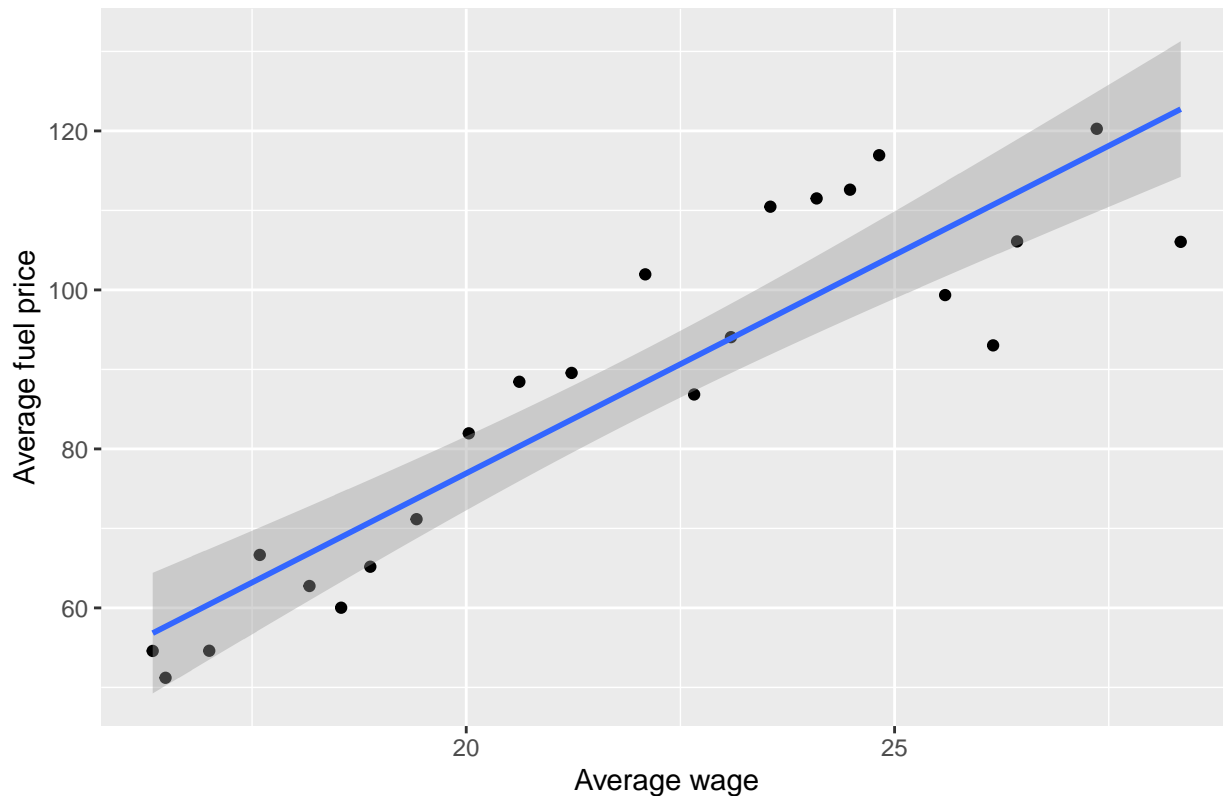
wages.year = wages %>%
  filter(Geography == 'Ontario',
         Wages == 'Average hourly wage rate',
         Type.of.work == 'Both full- and part-time',
         Education.level == 'Total, all education levels',
         Age.group == '15 years and over') %>%
  select(YEAR, Both.Sexes)

merged_data <- merge(wages.year, fuel.year, by.x = "YEAR", by.y = "Year")
```

```
ggplot(merged_data, aes(x = Both.Sexes, y = Avg.Price)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Average wage",
       y = "Average fuel price",
       title = "Relationship between Wages and Fuel Price")
```

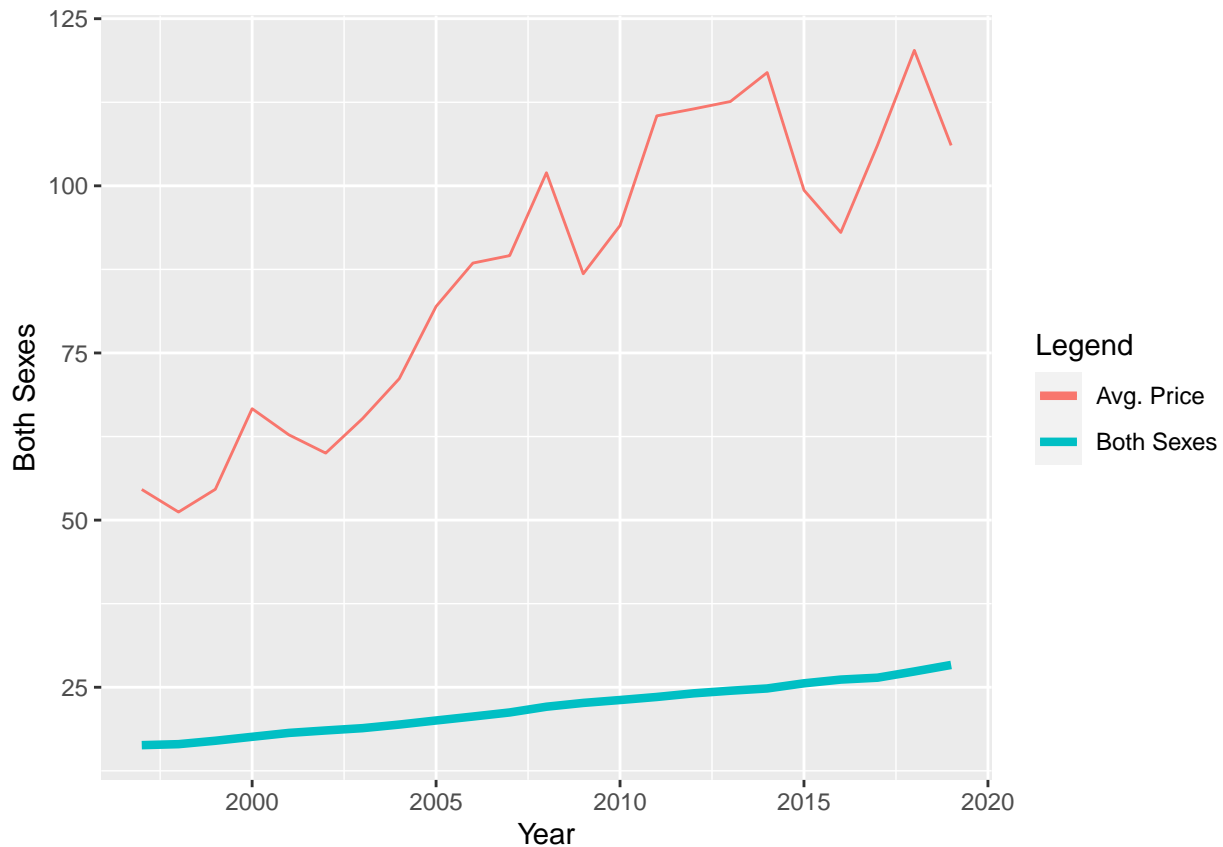
```
## `geom_smooth()` using formula = 'y ~ x'
```

Relationship between Wages and Fuel Price



```
# fuel and wage vs year
ggplot(merged_data, aes(x = YEAR)) +
  geom_line(aes(y = Both.Sexes, color = "Both Sexes"), size = 1.5) +
  geom_line(aes(y = Avg.Price, color = "Avg. Price")) +
  # scale_y_continuous(sec.axis = sec_axis(~./0.1, name = "Avg. Price")) +
  labs(x = "Year", y = "Both Sexes", color = "Legend")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



```
kable(avg_wage_by_education %>%
  mutate(pay_gap = male_avg - female_avg))
```

| Education.level | male_avg | female_avg | pay_gap |
|---|----------|------------|----------|
| Above bachelor's degree | 33.18565 | 29.29043 | 3.895217 |
| Bachelor's degree | 28.89087 | 24.83000 | 4.060870 |
| University certificate below bachelors degree | 25.33957 | 21.80435 | 3.535217 |
| University degree | 30.29478 | 26.04565 | 4.249130 |
| Community college, CEGEP | 23.81000 | 19.74957 | 4.060435 |
| Trade certificate or diploma | 23.44783 | 17.01913 | 6.428696 |
| Post-secondary certificate or diploma | 23.75652 | 19.35348 | 4.403043 |
| Some post-secondary | 18.32478 | 15.12348 | 3.201304 |
| High school graduate | 19.64739 | 15.93130 | 3.716087 |
| Some high school | 16.34565 | 12.35565 | 3.990000 |
| PSE (5,6,7,8,9)) | 26.32174 | 22.16043 | 4.161304 |
| No PSE (0,1,2,3,4) | 18.38304 | 14.83043 | 3.552609 |
| 0 - 8 years | 17.05087 | 12.62870 | 4.422174 |
| Total, all education levels | 22.94130 | 19.38304 | 3.558261 |

6 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.

```
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)
```

```
##
##  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).

7 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 )
```

8 Non-linear Regression Analysis

```
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)
```

9 Cross Validation

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

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

```
## [1] 16.248
```

The MSE of around 16.2480022 shows the average difference between actual and predicted wages.

10 Summary of Research

11 Appendix