STAA57FinalGroupProject

Jay, Sohil, and Soham

2024-04-05

1. Introduction

```r
# Loading relevant libraries
  library(tidyverse)
  library(dplyr)
  library(ggplot2)
  library(randomForest)
  library(knitr)
# Reading the csv file that contains relevant data
wages_data <- read.csv("v0913_05.csv")
```

Background Information

The data has been collected by the government of Ontario since 1997 and was last updated in 2020. This data is available under Open Government Licence – Ontario, and includes the information about wages grouped by characteristics like education level, year, employment status. The data can be used to study the impact of variables like gender, education level on the wages of individuals. This also allows us to identify disparity between certain section of people in terms of wages.

Research Question

The overall research question of this report is to study and analyze personal and educational factors on the wages of individuals in Ontario. Most importantly, we aim to answer the following questions:

- What are the summary statistics of wages by gender?
- What is the impact of geography on wages?
- How does the average weekly wage compare by gender over the years?
- What's the yearly trend in wages by education level?
- Compare and test the wages of male and female.
- Examine the relationship of wages and wage type.
- Create a linear model to predict wages as a factor of other variables.
- Estimate the mean wage of male individuals.

2. Description of Variables and Data

```r
names(wages_data)
```

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

- YEAR: Starting from year 1997 and ending at 2019, this column contains the record of when the data for other respective columns were collected.
- Geography: List of all the provinces in Canada.
- Type.of.work: List of all the employment types. (For Ex. Part-time, Full-time)
- Wages: Data about the type of wage. (For Ex. Hourly hourly wage)
- Education.level: Includes various different Educational levels. (For Ex. High school graduate)
- Age.group: Data about the different age groups. (For Ex. 15 years and over)
- Both.Sexes: Contains the wages for both the sexes combines.
- Male: Contains just the male wages.
- Female: Contains just the female wages.

3. Tables

Average Total Wages by Provinces

Geography	Mean_Total_Wages
Alberta	1617.82609
British Columbia	1759.89565
Manitoba	506.62174
New Brunswick	304.09565
Newfoundland and Labrador	194.37391
Nova Scotia	378.87826
Ontario	5461.60870
Prince Edward Island	57.86957
Quebec	3280.10435
Saskatchewan	413.81304

Explanation: The table gives us insightful information on the distribution of total wages of male and female by provinces. It compares the mean of total income of male and female with “all education level”, of age group “15 years+”, who are “total employees” and work “Both full time and part-time”. The data suggests strong disparity in the distribution of wages by provinces with Ontario leading with mean total wage of 5461, while the data suggests that the mean total wage in Prince Edward Island is just about 57.

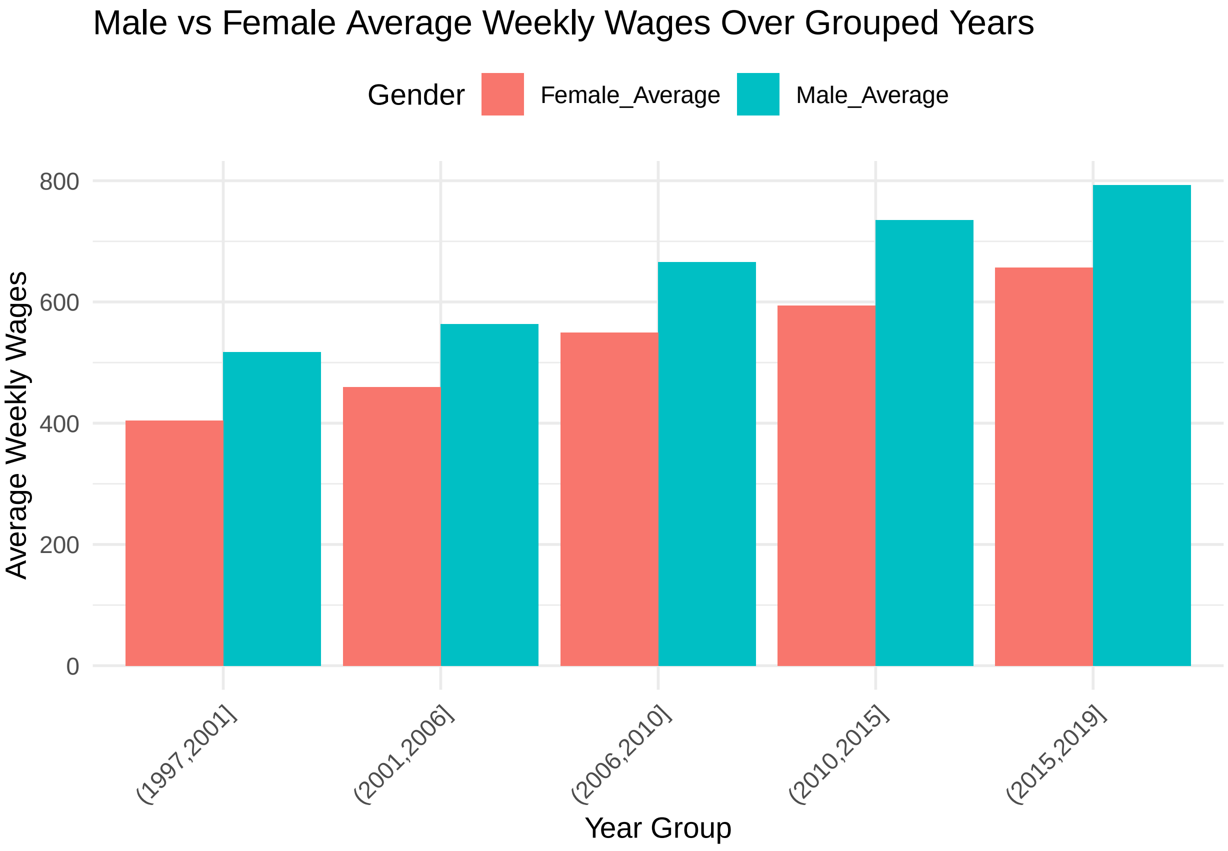
Wage Distribution between Men and Women over the Years

Gender	Mean_wage	Median_wage	Max_wage	Min_wage
Female	6884.135	7070.2	7986.1	5410.0
Male	7090.943	7107.2	8166.9	5954.5

Explanation: The table gives us information about the wages of the available genders, that is, Male and Female. It provides us with their mean wage, median wage, max wage, and min wage, over the years 1997-2019. One trend that seems to be prevalent is that the wages for men have always been higher as compared to the wages of women, whether its the mean or the median.

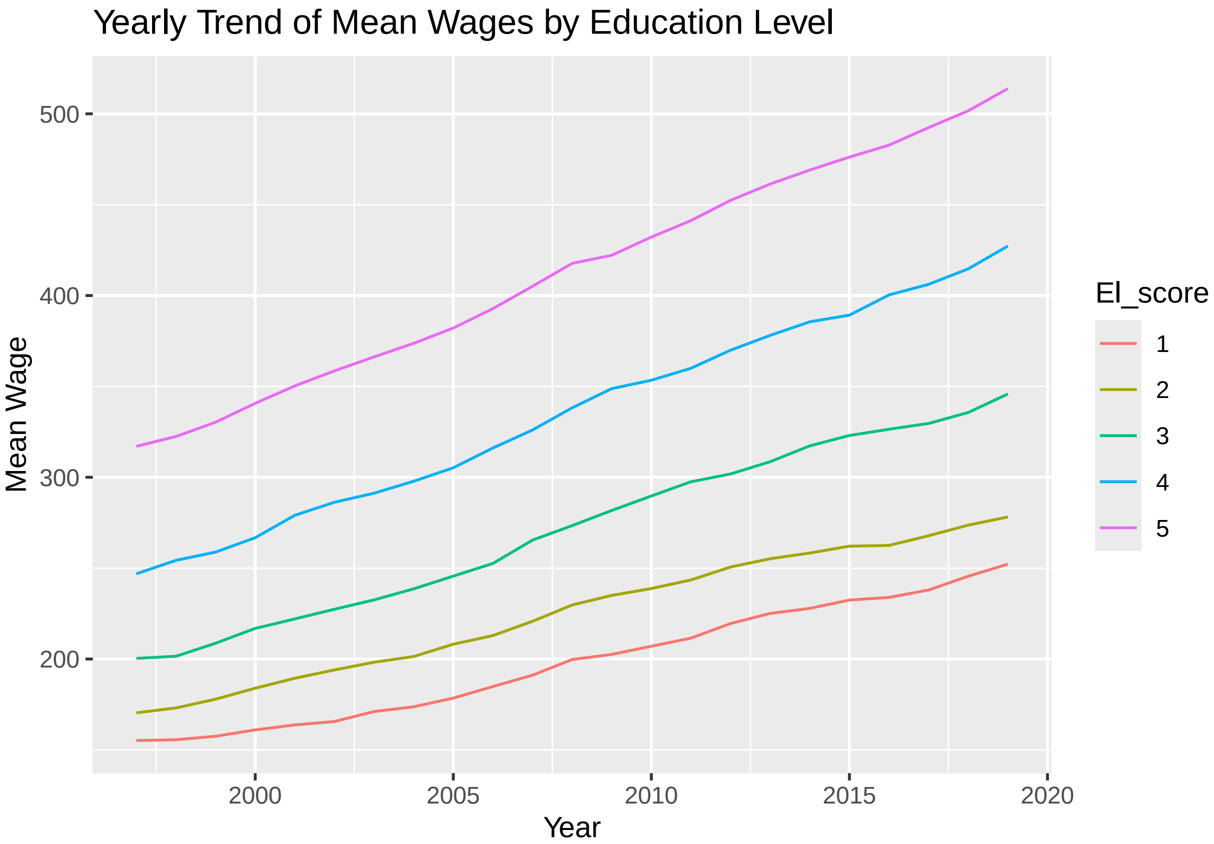
4. Graphs

Compare Male and Female Average Weekly Wages over Grouped Years



Description: In the graph, the columns YEAR, Male, and Female are used to create an illustration that compares the average weekly wage rate of men and women over the period of 1997 to 2019, where the years have been broken into sub-groups of 5 years. From the graph, we can see that there has been a steady increase in the weekly wage rate for both men and women over the given period of time, however, the weekly wage for men has always been higher as compared to the weekly wage for women.

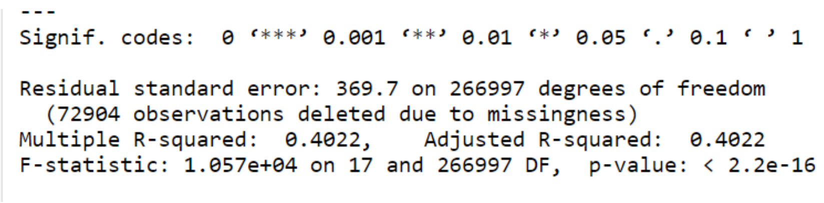
Yearly Trend of Mean Wages by Education Level



About Graph: Here, the legend El_score indicates the different education level with increasing El_score indicating higher education levels (5 being the highest education level and 1 being the lowest education level). The y axis represents the mean wages against the years on x-axis for different El_score i.e., different education levels. Inference: The plot shows that there's an increasing trend in the wages regardless of the education level, and the gap between the wages seems to be slightly increasing over the years. It can also be said that Education level has a direct and significant effect on mean wages.

5. Cross Validation and Linear Regression

```
include_graphics("lm.png")
```



```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 369.7 on 266997 degrees of freedom
  (72904 observations deleted due to missingness)
Multiple R-squared:  0.4022,    Adjusted R-squared:  0.4022
F-statistic: 1.057e+04 on 17 and 266997 DF,  p-value: < 2.2e-16
```

Explanation: Interpreting model's R-squared value of 0.4022 suggests that the model explains about 40% of Both.Sexes column i.e., the total wages. The difference in this model and the later regression models is that this model is based only on the training data which is about 80% of the total data. Regression parameters: Intercept represents the value of Both.Sexes, i.e. the total wages of male and female when the value of all the other variables is 0, which isn't practical in this summary. P-value : Low p-values of all variables except Median hourly wage rate represent strong evidence against the null hypothesis(i.e., the coefficient is zero or there is no correlation).

```r
# Predictions for training data
y_hat_train <- predict(lm_model, new_data = wages_data %>% filter(group_ind == "train"))
y_train_mean = mean(y_hat_train)
cat("Estimated mean of test data: ", y_train_mean, "\n")
```

```
## Estimated mean of test data:  286.449
```

```r
# Calculate in-sample MSE
mse_train <- mean((wages_data$Both.Sexes[wages_data$group_ind == "train"] - y_hat_train)^2, na.rm = TRU
cat("In-sample Mean Squared Error (MSE):", mse_train, "\n")
```

```
## In-sample Mean Squared Error (MSE): 319006.3
```

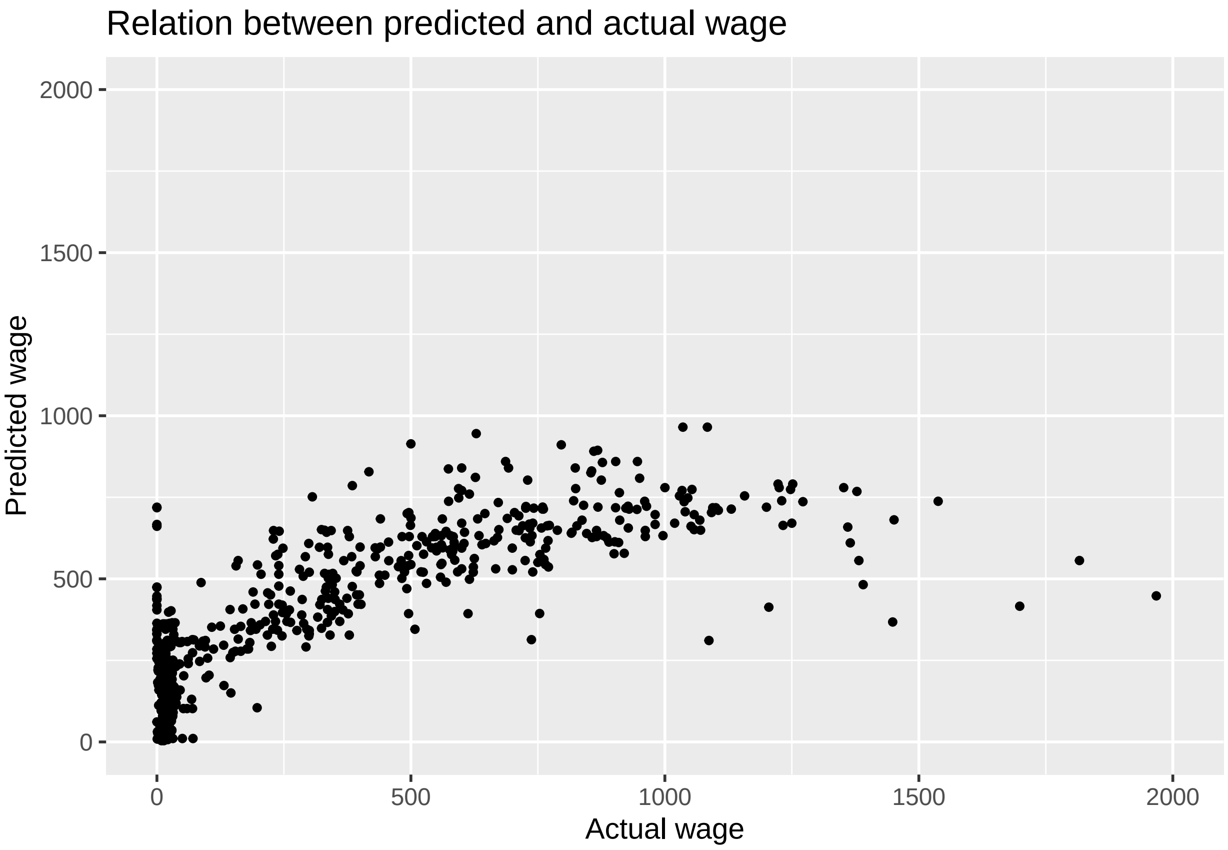
```r
# Predictions for test data
y_hat_test <- predict(lm_model, newdata = wages_data %>% filter(group_ind == "test"))
y_test_mean = mean(y_hat_test, na.rm = TRUE)
cat("Estimated mean of test data: ", y_test_mean, "\n")
```

```
## Estimated mean of test data:  285.6744
```

```r
# Calculate out-of-sample MSE
mse_test <- mean((wages_data$Both.Sexes[wages_data$group_ind == "test"] - y_hat_test)^2, na.rm = TRUE)
cat("Out-of-sample Mean Squared Error (MSE):", mse_test, "\n")
```

```
## Out-of-sample Mean Squared Error (MSE): 137256
```

```r
sample_data = data.frame(actual = wages_data$Both.Sexes[wages_data$group_ind == "test"], y_hat_test = y
correlation = ggplot(data = sample_data, aes(x = actual, y = y_hat_test)) +
geom_point(size = 1)+
xlim(0,2000) +
ylim(0,2000) +
labs(title = "Relation between predicted and actual wage",
x = "Actual wage",
y = "Predicted wage")
correlation
```



6. Bootstrap Sampling

```
## Average of bootstrap means: 271.4503
## 95% Confidence Interval for the mean wage: 235.7513 - 308.9974
```

7. Regression Analysis

```
cleaned_data <- na.omit(wages_data)
wages_data$Geography <- as.factor(wages_data$Geography)
wages_data$Age.group <- as.factor(wages_data$Age.group)
wages_data$Education.level <- as.factor(wages_data$Education.level)
wages_data$Type.of.work <- as.factor(wages_data$Type.of.work)
wages_data$Wages <- as.factor(wages_data$Wages)
wages_data <- wages_data %>%
filter(!is.na(Education.level), !is.na(Wages), !is.na(Both.Sexes))
long_wages_data <- wages_data %>%
pivot_longer(
cols = c(Male, Female),
names_to = "Gender",
values_to = "Wages_by_gender"
)
lm_model_1 <- lm(Both.Sexes ~ Wages, data = long_wages_data)
summary(lm_model_1)
```

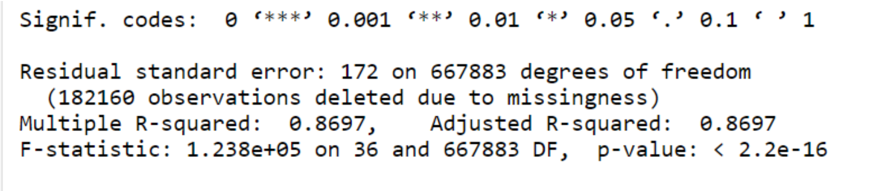
```
##
```

```
## Call:
## lm(formula = Both.Sexes ~ Wages, data = long_wages_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -597.4  -189.6    -3.0    11.6 15921.6
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     18.0296     0.9961  18.100   <2e-16 ***
## WagesAverage weekly wage rate 579.3306     1.4087 411.241   <2e-16 ***
## WagesMedian hourly wage rate   -1.6156     1.4087  -1.147    0.251
## WagesMedian weekly wage rate  528.6381     1.4087 375.256   <2e-16 ***
## WagesTotal employees          213.3390     1.4087 151.440   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 410.7 on 850075 degrees of freedom
## Multiple R-squared:  0.2702, Adjusted R-squared:  0.2702
## F-statistic: 7.868e+04 on 4 and 850075 DF,  p-value: < 2.2e-16
```

Explanation: The summary of the model suggests the relationship between total wages and wage type. R squared value: 0.27 suggests 27% of the wages can be explained by wage type which indicates that wage type is a decent indicator to predict wages.
Regression Parameters: Intercept - this tells us the value of Both.Sexes i.e., the total wages when all the variables of the model is 0. This isn't a practical value here.
P-value represents the probability of the null hypothesis i.e., the value of the coefficient = 0. From the above summary, it can be interpreted that all the parameters except Median Hourly wage having negligible p-value are significant in our model.

```
lm_model_2 <- lm(Both.Sexes ~ . , data = long_wages_data)
summary(lm_model_2)
```

```
include_graphics("Lm_group_project.png")
```



```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 172 on 667883 degrees of freedom
  (182160 observations deleted due to missingness)
Multiple R-squared:  0.8697,	Adjusted R-squared:  0.8697 
F-statistic: 1.238e+05 on 36 and 667883 DF,  p-value: < 2.2e-16
```

Explanation: The values most concerning to us here is the R-squared value i.e., 0.8697 which suggests that 86% of the total wages can be explained just by the variables in our dataset. Looking back at the linear model of total wages by wage type which explained about 27% of the total wages, it looks like a fair indicator which could explain 27% of 86% of the total wages.

8. Test of Hypothesis

```
var.test(Wages_by_gender ~ Gender, data = long_wages_data,
alternative = "two.sided",
conf.level = 0.95)
```

```
##
##  F test to compare two variances
```

```
##
## data:  Wages_by_gender by Gender
## F = 0.65261, num df = 425039, denom df = 425039, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6498475 0.6553968
## sample estimates:
## ratio of variances
##          0.6526132
```

Explanation: Low p-value suggests strong evidence agains the null hypothesis, i.e., the ratio of variances of wages of male and female is 1, or the variance for wages of male and female are equal. Furthermore, the 95% confidence interval suggests that the ratio of variances of wages of male and female lie between 0.649 - 0.655, further suggesting that the variances are not equal. Hence, we can use the t.test under the assumption that variances of the two columns are not equal.

```r
t.test(Wages_by_gender ~ Gender, data = long_wages_data,
alternative = "two.sided",
conf.level = 0.95,
var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Wages_by_gender by Gender
## t = -57.03, df = 814106, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -48.36436 -45.15049
## sample estimates:
## mean in group Female   mean in group Male
##             224.7834             271.5408
```

Explanation: An extremely small p-value suggests strong evidence against the null-hypothesis i.e., the mean wages of male and female are equal. The 95% confidence interval suggests that the difference in the mean wages of female and male lie between -48.36 to -45.15 further suggesting evidence against the hypothesis that mean wages of male and female are equal. At last, t.test gives the mean wages of femal and male which are 224.78 and 271.54 respectively.

9. Findings Summary

- The wages tend to increase for all education levels over the years, with a sustained and slightly increasing gap between different education levels, where higher education levels tend to have significantly more wages than lower education levels.
- According to the Geography table, the province with the highest mean wage is Ontario, second is Quebec, and the third is British Columbia
- According to the wage distribution table, Men have always had higher wages on average as compared to women.
- Wages type turns out to be a significant factor in determining the total wages of male and female for any given category explaining about 27% of the wage.
- Average weekly wage of male is consistently greater than average wage of females
- There's a 95% chance that our mean wage of male lies within [231.01, 308.55].

10. Appendix

```
# Loading relevant libraries
  library(tidyverse)
  library(dplyr)
  library(ggplot2)
  library(randomForest)
  library(knitr)
# Reading the csv file that contains relevant data
wages_data <- read.csv("v0913_05.csv")
names(wages_data)
filtered_data <- wages_data %>%
filter(Geography != "Canada",
Education.level == "Total, all education levels",
Age.group == "15 years and over ",
Wages == "Total employees",
Type.of.work == "Both full- and part-time")
# Calculate mean total wages for males and females, distinguished by provinces
mean_wages_by_gender <- filtered_data %>%
group_by(Geography) %>%
summarise(Mean_Total_Wages = mean(Both.Sexes, na.rm = TRUE))
kable(mean_wages_by_gender)
wages_data_long <- wages_data %>%
  filter(Geography == "Canada",
         Education.level == "Total, all education levels",
         Age.group == "15 years and over ",
         Wages == "Total employees",
         Type.of.work == "Both full- and part-time") %>%
  pivot_longer(cols = c(Male, Female), names_to = "Gender", values_to = "Wage") %>%
  group_by(Gender) %>%
  summarise(Mean_wage = mean(Wage),
            Median_wage = median(Wage),
            Max_wage = max(Wage),
            Min_wage = min(Wage))

kable(wages_data_long)
average_wages <- wages_data %>% filter(Wages == "Average weekly wage rate") %>%
  group_by(YEAR) %>%
  summarize(Male_Average = mean(Male, na.rm = TRUE),
            Female_Average = mean(Female, na.rm = TRUE))

# Create a new column to group years
average_wages$Year_Group <- cut(average_wages$YEAR, breaks = 5)

# Reshape the data into long format
average_wages_long <- average_wages %>%
  pivot_longer(cols = c(Male_Average, Female_Average),
               names_to = "Gender",
               values_to = "Average_Wages")

# Create a bar graph to compare male and female average weekly wages over grouped years
ggplot(average_wages_long, aes(x = Year_Group, y = Average_Wages, fill = Gender)) +
  geom_bar(position = "dodge", stat = "identity") +
  labs(title="Male vs Female Average Weekly Wages Over Grouped Years",x = "Year Group", y = "Average We
  scale_x_discrete(labels = function(x) gsub("\\.", "-", x)) +  # Tiled x-axis labels
  theme_minimal() +
  theme(legend.position = "top", axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis la
wages_data$Education.level <- trimws(wages_data$Education.level)
wages_data <- wages_data %>%
mutate(
El_score = case_when(
Education.level %in% c("0 - 8 years", "Some high school", "No PSE (0,1,2,3,4)") ~ 1,
Education.level %in% c("High school graduate", "Some post-secondary") ~ 2,
Education.level %in% c("Post-secondary certificate or diploma",
"Trade certificate or diploma",
"Community college, CEGEP",
"University certificate below bachelors degree") ~ 3,

Education.level %in% c("University degree", "Bachelor's degree", "Above bachelor's degree") ~ 4,
Education.level %in% c("Total, all education levels") ~ 5
)
)
summary_data <- wages_data %>%
group_by(El_score, YEAR) %>%
summarise(mean_wage = mean(Both.Sexes, na.rm = TRUE))
summary_data <- summary_data %>%
filter(!is.na(El_score))
# Create yearly trend plot
g <- ggplot(summary_data, aes(x = YEAR, y = mean_wage, color = as.factor(El_score))) +
geom_line() +
labs(title = "Yearly Trend of Mean Wages by Education Level",x = "Year",
y = "Mean Wage",
color = "El_score")
g
wages_data = wages_data %>% mutate(group_ind = sample(c("train","test"),

size = nrow(wages_data),
prob = c(0.8,0.2),
replace = T))

wages_data = wages_data %>% mutate(group_ind = sample(c("train","test"),

size = nrow(wages_data),
prob = c(0.8,0.2),
replace = T))

lm_model <- lm(Both.Sexes ~ Wages + Geography + El_score + Type.of.work,
data = wages_data %>% filter(group_ind == "train"))
include_graphics("lm.png")
# Predictions for training data
y_hat_train <- predict(lm_model, new_data = wages_data %>% filter(group_ind == "train"))
y_train_mean = mean(y_hat_train)
cat("Estimated mean of test data: ", y_train_mean, "\n")
# Calculate in-sample MSE
mse_train <- mean((wages_data$Both.Sexes[wages_data$group_ind == "train"] - y_hat_train)^2, na.rm = TRU
cat("In-sample Mean Squared Error (MSE):", mse_train, "\n")
# Predictions for test data
```

```r
y_hat_test <- predict(lm_model, newdata = wages_data %>% filter(group_ind == "test"))
y_test_mean = mean(y_hat_test, na.rm = TRUE)
cat("Estimated mean of test data: ", y_test_mean, "\n")
# Calculate out-of-sample MSE
mse_test <- mean((wages_data$Both.Sexes[wages_data$group_ind == "test"] - y_hat_test)^2, na.rm = TRUE)
cat("Out-of-sample Mean Squared Error (MSE):", mse_test, "\n")
sample_data = data.frame(actual = wages_data$Both.Sexes[wages_data$group_ind == "test"], y_hat_test = y_
correlation = ggplot(data = sample_data, aes(x = actual, y = y_hat_test)) +
geom_point(size = 1)+
xlim(0,2000) +
ylim(0,2000) +
labs(title = "Relation between predicted and actual wage",
x = "Actual wage",
y = "Predicted wage")
correlation
boot_function = function(){
# Sample from the 'Male' column with replacement
s = sample(wages_data$Male, size = 500, replace = TRUE)
# Calculate and return the mean of the sample, ignoring NA values
return(mean(s, na.rm = TRUE))
}
# Replicate the bootstrapping process 10,000 times
boot_mean = replicate(10000, boot_function())
# Calculate the mean of the bootstrap means
boot_mean_avg <- mean(boot_mean)
# Calculate a 95% confidence interval for the bootstrap means
CI_lower <- quantile(boot_mean, probs = 0.025)
CI_upper <- quantile(boot_mean, probs = 0.975)
# Display the results

cat("Average of bootstrap means:", boot_mean_avg, "\n")
cat("95% Confidence Interval for the mean wage:", CI_lower, "-", CI_upper, "\n")
cleaned_data <- na.omit(wages_data)
wages_data$Geography <- as.factor(wages_data$Geography)
wages_data$Age.group <- as.factor(wages_data$Age.group)
wages_data$Education.level <- as.factor(wages_data$Education.level)
wages_data$Type.of.work <- as.factor(wages_data$Type.of.work)
wages_data$Wages <- as.factor(wages_data$Wages)
wages_data <- wages_data %>%
filter(!is.na(Education.level), !is.na(Wages), !is.na(Both.Sexes))
long_wages_data <- wages_data %>%
pivot_longer(
cols = c(Male, Female),
names_to = "Gender",
values_to = "Wages_by_gender"
)
lm_model_1 <- lm(Both.Sexes ~ Wages, data = long_wages_data)
summary(lm_model_1)
lm_model_2 <- lm(Both.Sexes ~ . , data = long_wages_data)
summary(lm_model_2)
include_graphics("lm_group_project.png")
var.test(Wages_by_gender ~ Gender, data = long_wages_data,
alternative = "two.sided",
conf.level = 0.95)
t.test(Wages_by_gender ~ Gender, data = long_wages_data,
alternative = "two.sided",
conf.level = 0.95,
var.equal = FALSE)
```