How does Disability Different than Else from the respects of Employment and Income*

Analysis of Canadian health and disability survey, 1984 - adults

Meixuan Chen

30 April 2022

Abstract

There are about 6.2 million Canadians who have at least one disability. This paper analyzes how disabilities, education level, class of work, and other factors affect employment status and income using multiple linear regression and logistic models. The analysis finds that income and education level positively influence the weekly working hours while other factors like the job classes negatively impact the weekly working hours. It provides an insight into populations with disabilities to achieve equal rights in society.

Keywords: Disability, Health, Equal Rights, Employment, Working Hours, Labour Force

Contents

*Code and data are available at: https://github.com/chenme72/disability_analysis.git

1 Introduction

About 20 percent of the population in Canada is defined as disability (CCPA 2022a). Equal job opportunities and salary levels for the disabled population are among the most important fundamentals of human rights. According to the Employment Equity Act, disability is a long-term or recurring physical, mental, sensory, psychiatric, or learning impairment'. People with disability are those who consider themselves to be disadvantaged in employment or believe their employer consider them to be poor in career because of their impairment (CCPA 2022a).

This paper analyzes the labor force of disabled adults who were 15 years old or older in 1984. starting with a research question on how weekly working hours and income level are influenced by other factors such as disabilities, region, and education levels.

The data was collected in the Canadian Health and Disability Survey from 1983 to 1984 as a supplementary survey of the Canadian Labour Force Survey by the Canadian government, with almost 130,000 participants. This large sample size provides a unique and relatively accurate analysis for this paper.

By investigating the relationship between the weekly working hours and its influencing factors using a multiple linear regression model, the weekly working hours are positively influenced by gender and income and negatively affected by the class of work, province, and reasons that lose the work hours. By investigating the interactions between the variables, the average working hours per week of the disabled population is similar to others, which is around 40 hours per week. They have an equivalent income level as the population without disabled when they have university degrees. However, with an education level lower than a university degree, the disabled population generally has a lower income level than the general population. The disabilities are one of the most significant reasons reduce their working hours, which could also lead to a lower salary. The back musculoskeletal problem is one of the most common diseases that cause them to have lower productivity.

The rest of the paper is divided into six sections: the data section, simulation section, model section, results section, discussion section, and the appendix. The data section explains the variables that were analyzed, the reproducible steps of data cleaning and data simulation, as well as the methodology of the survey. The simulation section shows the comparison between the simulated data vs. origin data. The results section illustrates the results from the analysis with figures and tables. It mainly explores the relationship between the weekly working hours and the affecting factors found in the model section. The discussion section analyzes the findings from the results section from a more profound perspective. In this section, the weakness and the potential improvement in the future are also discussed.
The last section is the appendix, which shows the supplementary figures from the analysis. It contains the additional residual plots for the linear model diagnostics. Also includes the additional figures that illustrate the interactions between variables.

2 Data

2.1 Data Source

This paper uses the Canadian health and disability survey, 1983-1984 (D. Dolson 1884) from the Canadian Health and Disability Survey (CDHS), which was conducted as a supplement to the Labour Force Survey (LFS) in October 1983 and in June 1984. The origin dataset was composed of adults file and children file. This analysis uses the adult file intending to study the affecting factors of work hours of the disabled population.

The dataset was collected on the CDHS questionnaire and contained 8 sections, which are
- Section A: Screening
This section recorded the information on the types of conditions such as trouble walking 400 meters without resting, trouble moving from one room to another, difficulty getting in and out of bed, etc.
- Section B: Follow-Up
This section recorded the follow-up information from Section A, such as the main and the second main health problems that cause trouble walking 400 meters without resting.
- Section C: Nature of disability
This section recorded the natural disabilities caused by the conditions in Section A, such as mental disorders, hearing disorders, lower limbs arthritis, etc.
- Section D: Employment
This section describes the labor market-related information, such as conditions limiting working, special needs, employment status, etc.
- Section E: Education
This section recorded the education-related information for students and non-students.
- Section F: Transportation

This section recorded the transportation questions, such as if their transportation is private, special public services, local public services, etc.
- Section G: Accommodation
This section collected the information for special features usage, such as access ramps, wide doorways, street entrances, etc.
- Section H: Economic Characteristics
This section contains information about demographic, demographic, and socio-economic variables.

R studio (R Core Team 2020) is used to process the analysis. The package 'tidyverse' is used for gathering data (Wickham et al. 2019). The package 'dpyr' is used to process the data (Wickham et al. 2022) and the package 'knitr' is also used for processing the data and knitting the pdf(Xie 2022). The package 'ggplot'(Wickham 2016) and 'ggpubr' (Kassambara 2020) are used for marking statistical figures in the analysis. The package 'MASS' (Venables and Ripley 2002) is used to make tables for the analysis. The package 'car' (Fox and Weisberg 2019) is used to do the stepwise selection in the logistic and linear regression model. Lastly, the package 'pointblank' is used to do the pointblank check for the simulation data.

2.2 Methology

2.2.1 Canadian Health and Disability Survey Methology

The Canadian Health and Disability Survey 1984 contains 788 variables for the 15,854 samples who were aged 15 and older (CCPA 2022a). As a supplement to Labour Force Survey, it uses the same survey methodology as LFS. CHDS was collected in October 1983 and June 1984, and the response rate was around 90% (Statistic Canada 1884).

The target population was populations in Canada who have one or more non-behavioral disabilities, knowledge acquisition, or other educational disabilities and people who suffer from chronic disease and degenerative

nature that have a high chance of developing physical impairments. However, this dataset does not contain a population who are mentally ill.

The dataset was collected by using two-stage sampling. In the first stage, the dataset was collected by geographical regions, called primary sampling units. For each selected direct sampling unit, the statisticians draw up a list of dwellings. The screening section was processed in the first stage, which determines the target samples for the disability survey.

In the second stage, all of the residents of the dwellings make up the survey sample of persons. The samples selected by the screening section were followed up in the other survey, where section B (Follow-Up Section) was collected. During the second stage, the samples participated in personal interviews, and the rest of the data was collected by the second questionnaire.

In conclusion, there are four questionnaires in total. The first survey was the Labour Force Survey, and the Screening Survey was developed in the first stage. Two other surveys with follow-up questions and more detailed questionnaires that were designed for Disability Designed were used in the second stage.

2.2.2 Data Analysis Methology (Cleaning Steps)

In this analysis, 11 variables are selected from the raw dataset, which are clswrkr, ind, tenure, prov, sex, age, educ, tothrswk, d09b,h03 and whyloss. By using 'rename' function, some of the variable names are renamed with more representative names, for example, clswrkr are named as 'class_of_work', tothriswk are named as 'work_hrs_wk', 'h09b' are named as 'why_work_lim', 'h03' were named as 'income'. By using 'filter' function, the missing data and NA data were removed. By using 'mutate' and 'casewhen' function, the data were changed from a numeric code into its represented information. For example, the variable sex contains values of '1' and '2', then value '1' are changed to 'male', value '2' are changed to 'female'.

2.3 Data Description

Table 1: Clean Dataset

cls_of_wrk	ind	tenure	prov	sex	age	educ	wrk_hrs	limitations	income	whyloss
Paid worker, private	Manufacturing, non-durables	11-20 years	NL	Male	35-44 years	None/Elementary	40	Lower limbs muscu-loskele-tal	$15000-19999	Illness/disability/personal
Paid worker, govt non-business	Community services	11-20 years	NL	Male	25-34 years	University Degree	45	Hearing disor-ders	$25000-29999	Vacation
Paid worker, private	Retail trade	11-20 years	NL	Male	45-54 years	Post-secondary cer/diploma	40	Ischaemic heart disease	$25000-29999	Vacation
Paid worker, private	Finance, etc.	1-5 years	NL	Female	45-54 years	Post-secondary cer/diploma	40	Lower limbs muscu-loskele-tal	$10000-14999	Vacation
Paid worker, private	Construction	1-6 months	NL	Female	35-44 years	Post-secondary cer/diploma	40	Other/ns muscu-loskele-tal probs	$30000+	Vacation

As 1 shown, there are 11 variables in the cleaned dataset for this analysis.

- cls_of_wrk: Represents the classes of work of participants and it has 6 values excluding the missing values, which are private paid worker, government business paid work, government non-business paid worker, employer, own account and unpaid family worker.

- ind: Represents the industry that participants attended in 1984. It has 13 different industries, which are agriculture, other primary, non-durable manufacturing, durable manufacturing, construction transportation, wholesale trade, retail trade, finance,community services, personal services, business or misc. services and public administration.
- tenure: Represents the job tenures, which is separated in 6 month interval from 1 to 20 months, and over 20 years.
- prov: It describes the 10 provinces in Canada, however, it does not contains information about Yukon, Northwest Territories and the population living on Indian Reserves.
- sex: It describes the gender of the particpants, male and female.
- age: It represents age groups, which are separated into 11 groups in a five year interval from age 15 to 85+.
- educ: It describes the education level of the particpants, which are separated into five levels, which are none or elementary, some or completed high school, some post-secondary, post-secondary certificaion or diploma, and university degree.
- wrk_hrs: It is a numerical variable that describes the total usual hours worked per week.
- limitation: It describes the main health problem that causing act limitations at work and it is separated 20 different disorder, such as the sight disorder, endocrine, ischemic heart disease, etc.
- income: It describes the total income from all sources in the 1983 (past year) and it is separated into 8 groups in 5000 dollars interval from none to over 30,000 dollars.
- whyloss: It describes the reasons for time loss in the reference week. It is separated into 8 groups, such as bad weather, illness or disabilities, vacation, etc.

2.4 Strengths

The two stages of sampling can reduce the geographic spread of the sampled persons, which would increase the accuracy of the dataset. Additionally, collecting data from geographic balancing units avoids creating an extensive list of all addresses in Canada, which increases efficiency.

The sample size is large and unique because it is a supplementary survey of the Labour Force Survey, and it has the advantage that it covers people whether or not they are seeking medical care.

2.5 Weakness

The data might decrease accuracy since the survey is self-reported. Also, the survey mainly focuses on physical disabilities, such as being unable to walk over 400 meters. It lacks information about mental or psychological disabilities, such as necrosis or emotional disorders. Thus, the dataset could be biased by the definition of disabilities.

Additionally, each question in the survey has a significant amount of missing values due to the 'not application' issue or the sensitivity questions, which cause the non-response value, especially in sections on education and transportation. However, this analysis avoids using the data from these two sections to increase accuracy.

2.6 Key Features

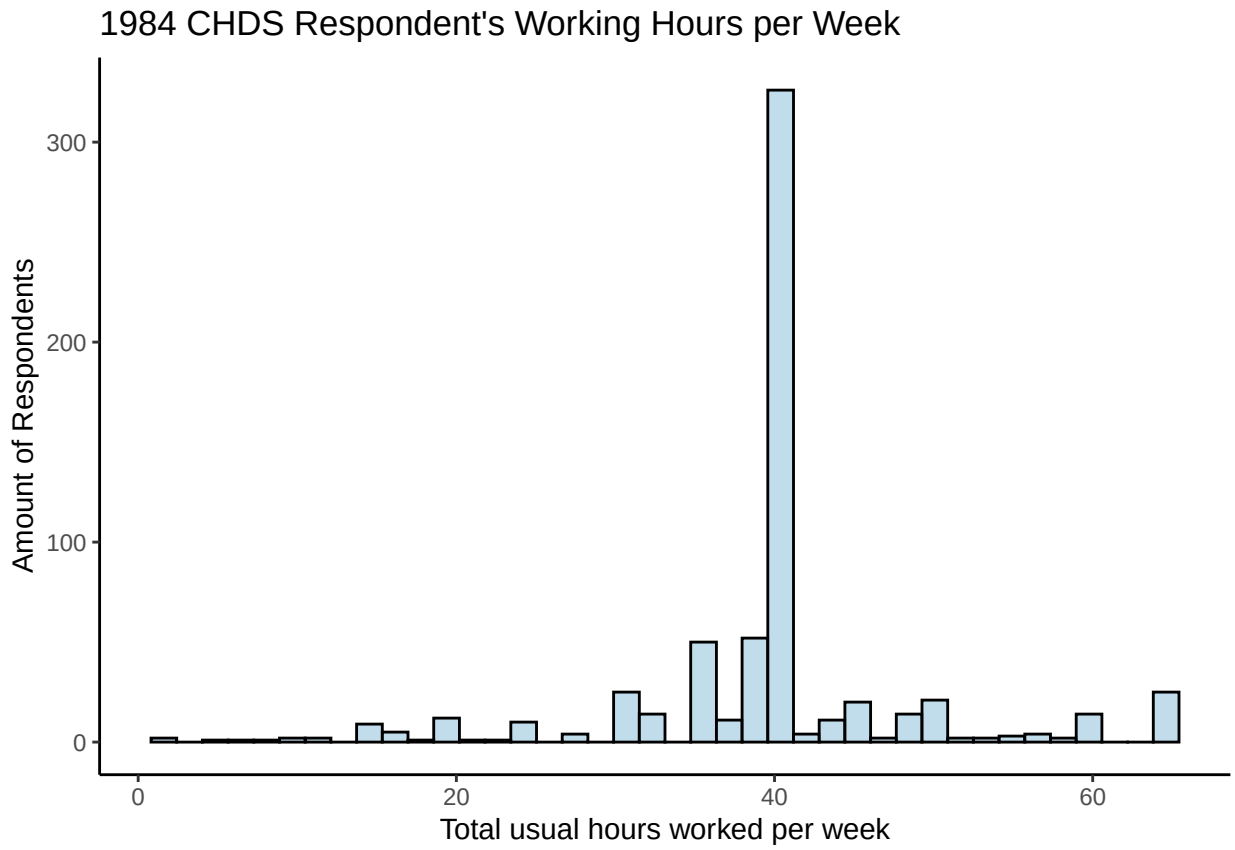


Figure 1: Time spent at work by the respondent is mainly concentrated around forty hours

Figure 1 depicts the distribution of the number of hours worked per week by the respondents. In the image, we find that the number of respondents who work 40 hours per week exceeds 300, and since this image shows an almost symmetrical distribution, we can speculate that the mean, median and plural of the respondents' weekly working hours are all close to 40. There are also a small proportion of respondents who work between 30 and 40 hours per week. However, the number of respondents with very short and very long working hours is very small. Overall, respondents work a relatively even number of hours, with the majority residing in the 40-hour range per week.

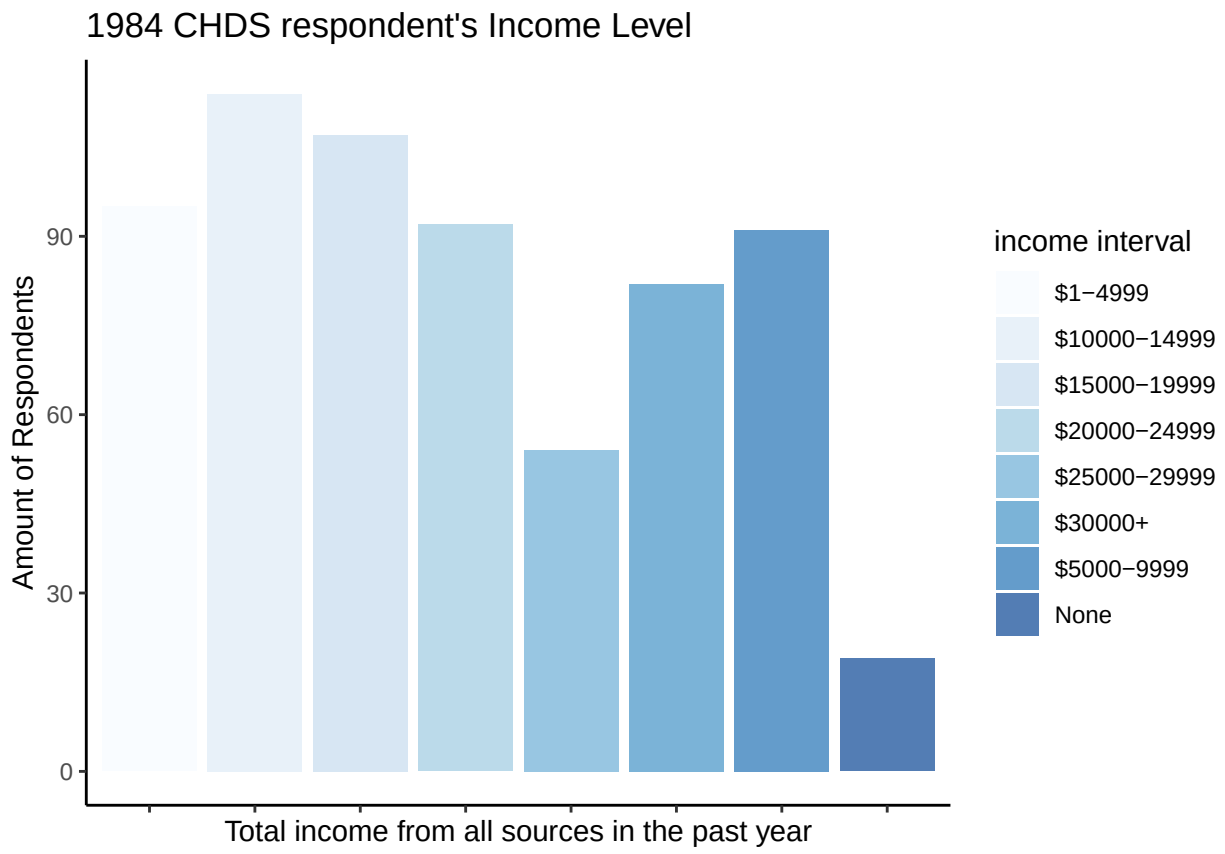


Figure 2: Relatively few respondents had no income last year

Figure 2 demonstrates the distribution of respondents' salaries. We find that the income of respondents is quite uneven, with a large number of respondents in the low-income group (less than $10,000). For the segment with salaries over $10,000 but less than $30,000, we observe that the number of respondents decreases as the salary range rises. However, it is also observed that a large proportion of respondents earn more than $30,000. Only a small percentage of respondents do not have any income. In general, the majority of respondents were able to earn salaries through their work, but the level of income varied considerably among individuals, with the low-income group accounting for a large portion of the total, but there were also many high-income individuals as well.

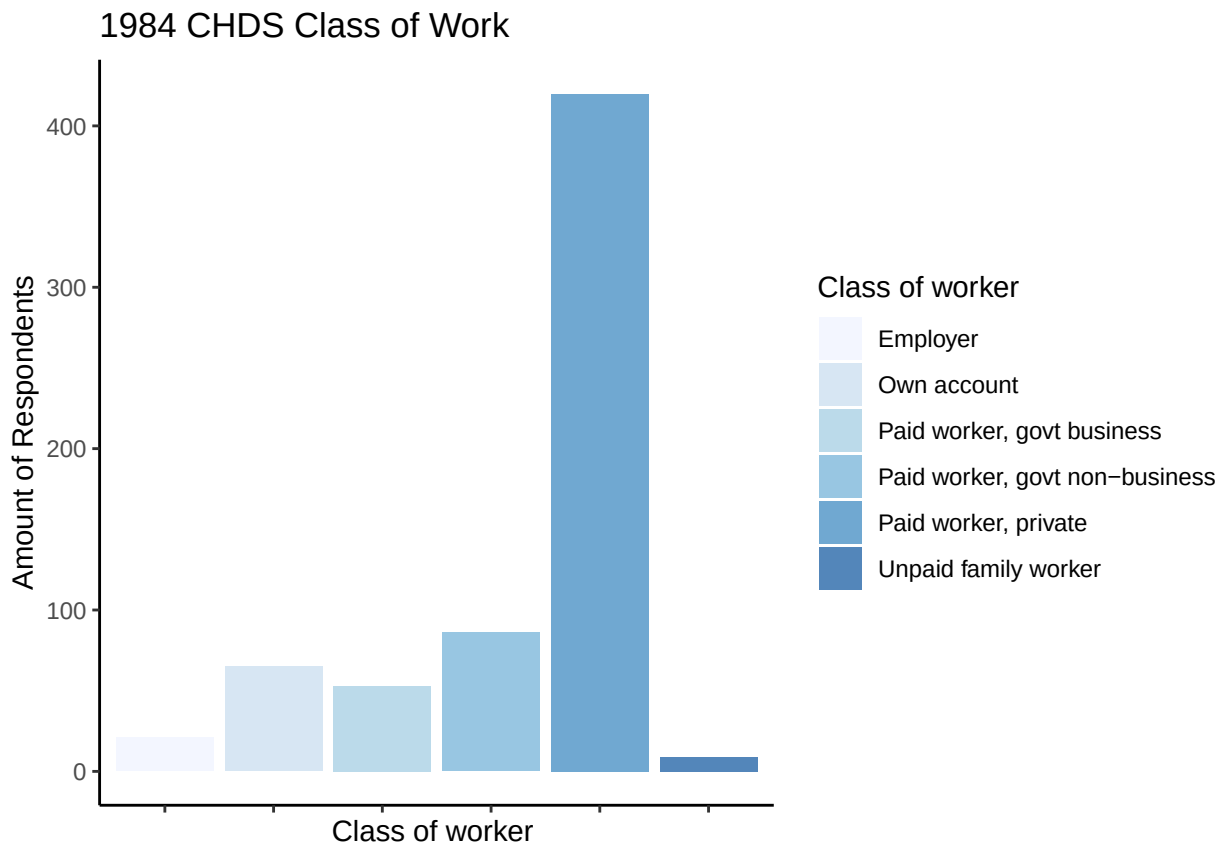


Figure 3: Most of the respondents are paid workers of private companies

Figure 3 shows the distribution of the respondents by their job categories. We find that most of the respondents are paid workers, but they work for different organizations. Over 400 of the respondents are paid workers for private companies, which is much higher than the number of paid workers for government. We believe that this is related to the properties of government departments and private companies. For government agencies, regardless of whether they are business-related or not, their main intention is to serve the people. And government sectors are funded by the state, rather than profit-making enterprises, so they do not need that many employees. In contrast, private companies are mainly profit-oriented and recruit more paid workers to help them expand their businesses and gain more benefits. Therefore, we observe that the number of paid workers for private companies is the greatest. In addition, we found that a small number of respondents are employers or worked for their own accounts, and only very few of the respondents are unpaid family members.

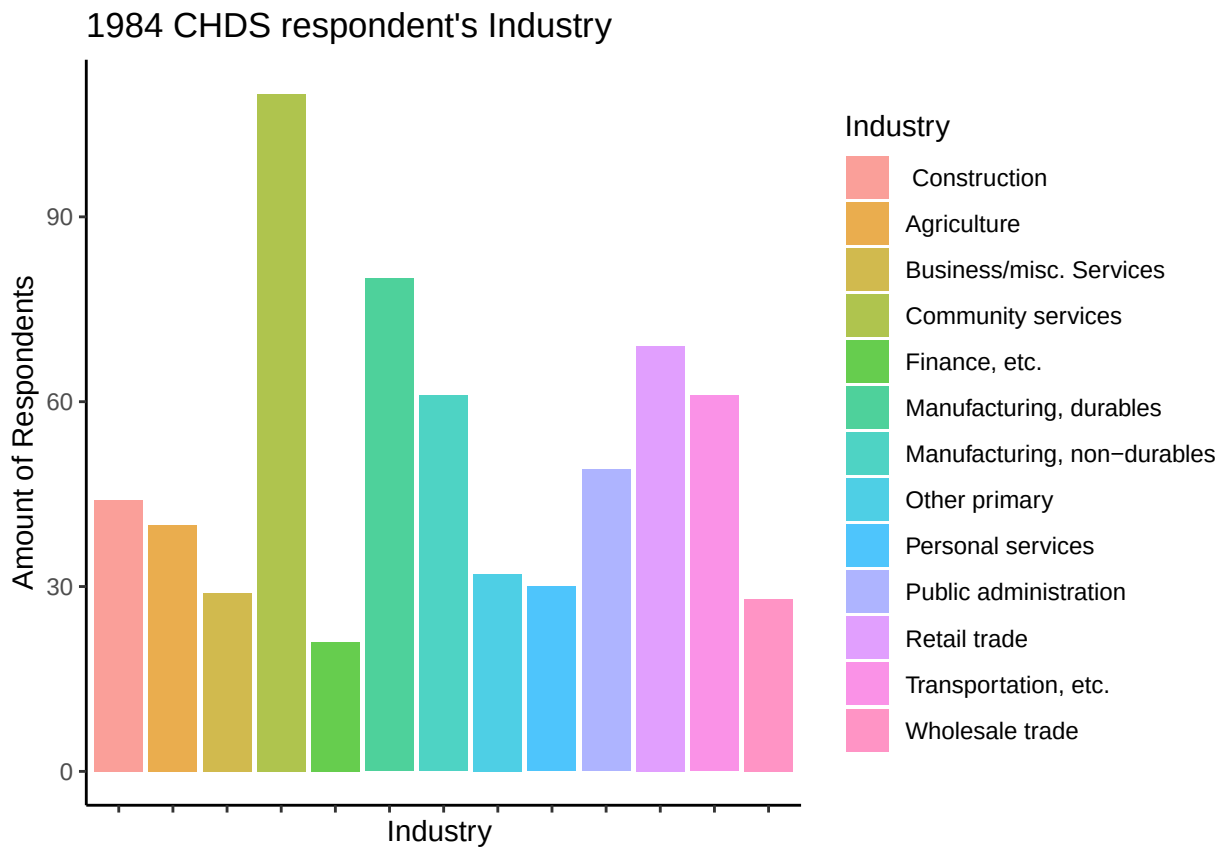


Figure 4: The industries in which people with disabilities work are very diverse

In Figure 4, we can view the distribution of the work industries of the respondents, and we find that the fields in which people with disabilities are working are very diverse. The largest number of respondents chose the community service type of work. The workload associated with community service is not very heavy and it meets the needs of people with disabilities. In addition, it has the advantages of flexibility and proximity to home, accommodating the potential special requirements of people with disabilities due to their physical condition. Therefore, this type of work was preferred by respondents. In addition, we found that manufacturing, retail trade and transportation are also popular choices for disabled people. However, the number of respondents who chose to work in the financial industry was very small, with only 1/4 of the number of community workers. This is likely due to the fact that the work in the financial industry is more stressful and disabled people have difficulties in balancing their physical condition with the huge work overload.

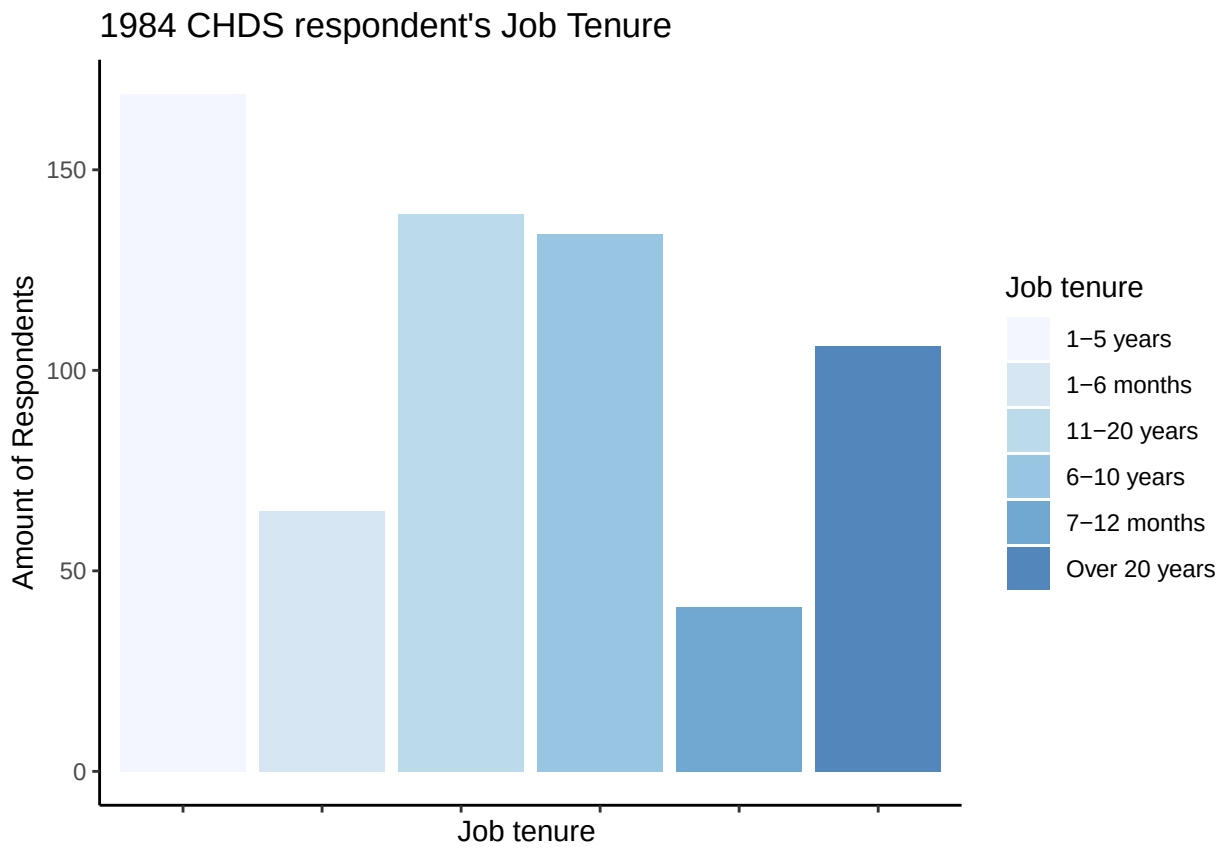


Figure 5: Very few respondents preferred short-term jobs

In Figure 5, the distribution of the different job tenures of the respondents is observed. Through this graph, we notice that most of the respondents tend to work in a job for a relatively long time. Only very few respondents have been in their current job for a couple of months. The number of respondents who have stuck in their current job for 1-5 years accounts for the largest proportion. The number of respondents with 11-20 years of job tenure and 6-10 years of job tenure also has a significant weight among all. In case of people with disabilities, it is not so easy to find a suitable job because of their physical conditions, so they are more willing to stick to their favorite job for a very long time rather than changing jobs frequently.

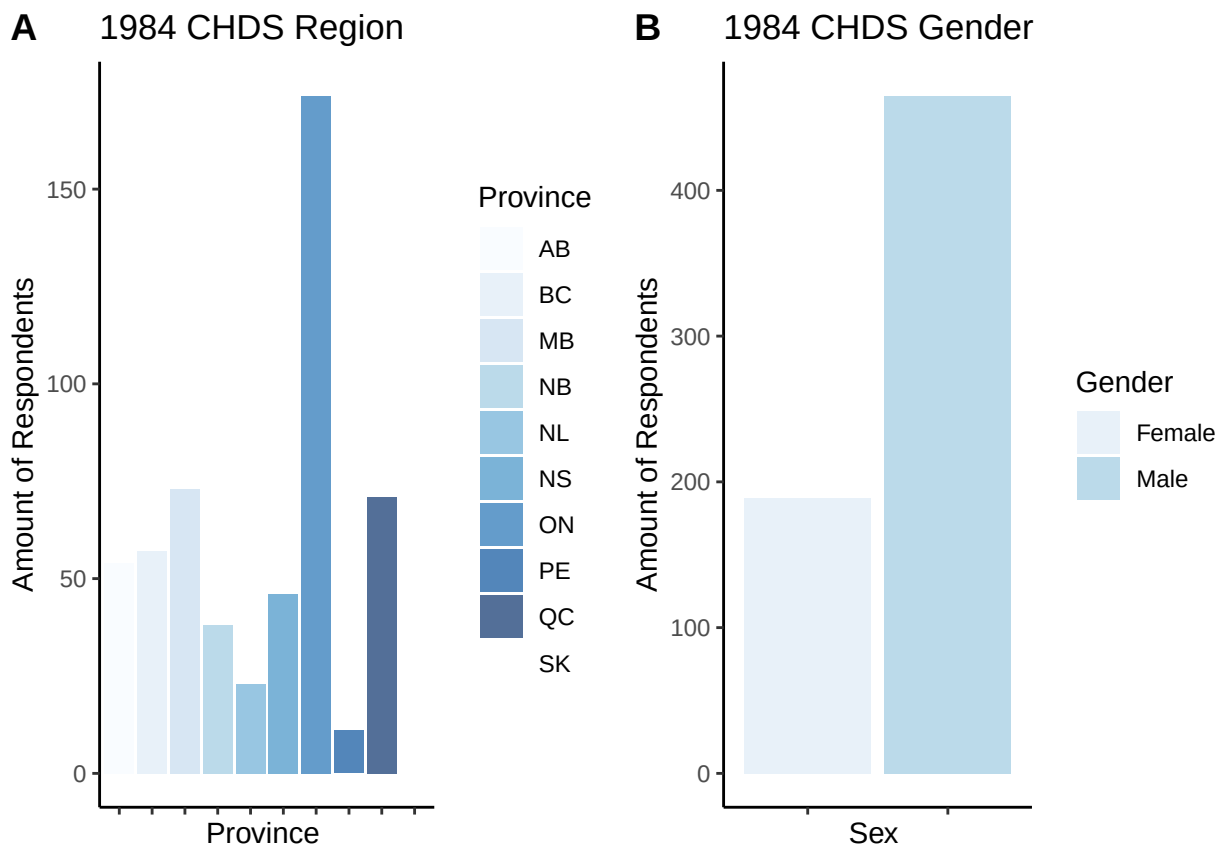


Figure 6: Respondents came from ten provinces in Canada, with varying numbers in each province. The amount of male respondents was much higher than that of the female respondents

In Figure 6 A, we identify the regions to which the different respondents belong. We know that there are ten provinces in Canada, and these respondents are from these ten different provinces. With over 150 respondents from Ontario, there were more respondents in Ontario than in any other province. The amount of participants from Newfoundland and Prince Edward Island was relatively low. Variations in the number of respondents are strongly related to the geographic distribution of the Canadian population. The total population of Canada is 35 million. The most populated province is Ontario, with 13 million people, followed by Quebec, with 8 million people. The provinces of New Brunswick, Newfoundland and Labrador, and Prince Edward Island, which account for one-third of Canada's landmass, are sparsely populated, which greatly reduces the population density of the country. We found that the number of respondents in each province is generally consistent with the distribution of the population in each province.

Figure 6 B organizes the distribution of the gender of the respondents. We find that the number of male respondents is more than 400, while the number of female respondents is less than 200. The number of males is much higher than the number of female respondents. However, according to our common sense, the number of males and females should exhibit the ratio of 1:1. The difference in the number of respondents by gender here leads us to infer that the number of males with disabilities is likely to be much higher than the number of females with disabilities.

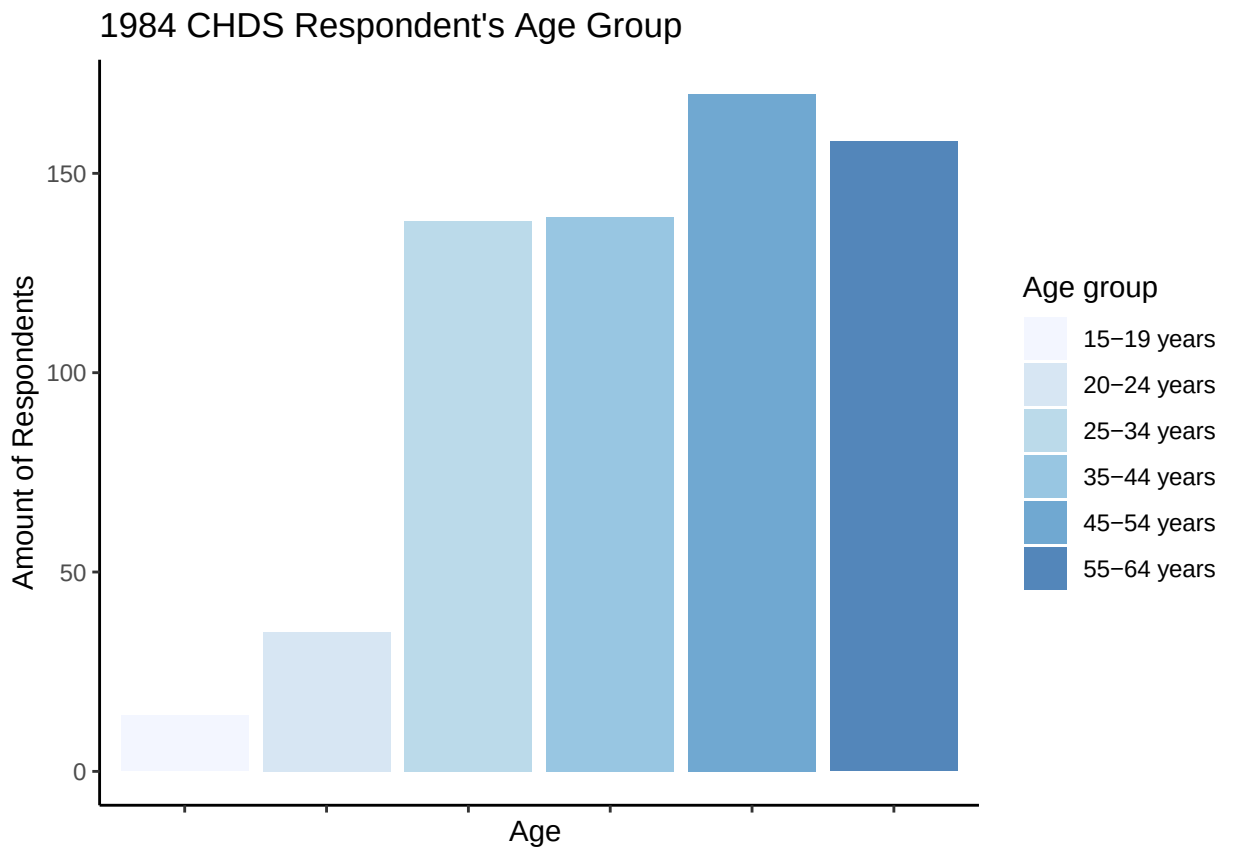


Figure 7: Very few of the respondents were minors

In Figure 7, we see the age distribution among the respondents. The age groups from oldest to youngest are arranged in light to dark colors. We find that the number of underage respondents is very low, and more than 95% of the respondents are adults. The middle-aged and old-aged groups make up a significant portion of the total sample, as the number of respondents between 45 and 54 years old even exceeds 150, and the number of respondents between 55 and 54 years old is also around 150. The number of respondents whose age bracket was between 25 and 34 years old was about the same as the number of respondents whose age range was between 35 and 44 years old.

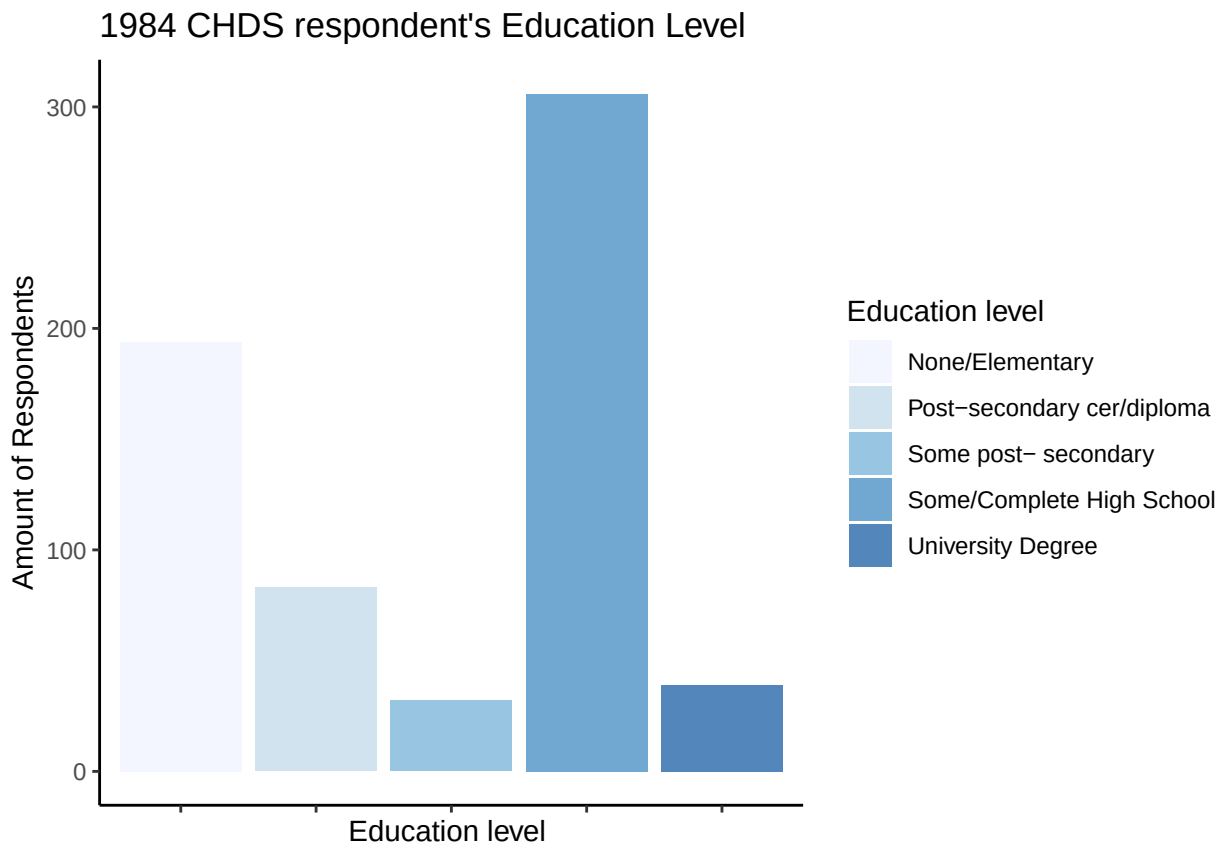


Figure 8: Respondents with a high school degree accounted for a very large percentage of all respondents

Figure 8 presents us with the distribution of the educational level of the respondents. We identified that close to 200 respondents had no education or only a relatively elementary education. More than 300 respondents have received some or completely high school education. But only a very small number of people have received a university education. From this image we find that the distribution of the respondents' education is uneven. The combination of the large number of people with limited education and the extremely large amount of people with high school education makes us aware of the bifurcation of the population's awareness of education. The fact that a minority of the participants have received university education also reflects the relatively weak consciousness of the people to receive higher education

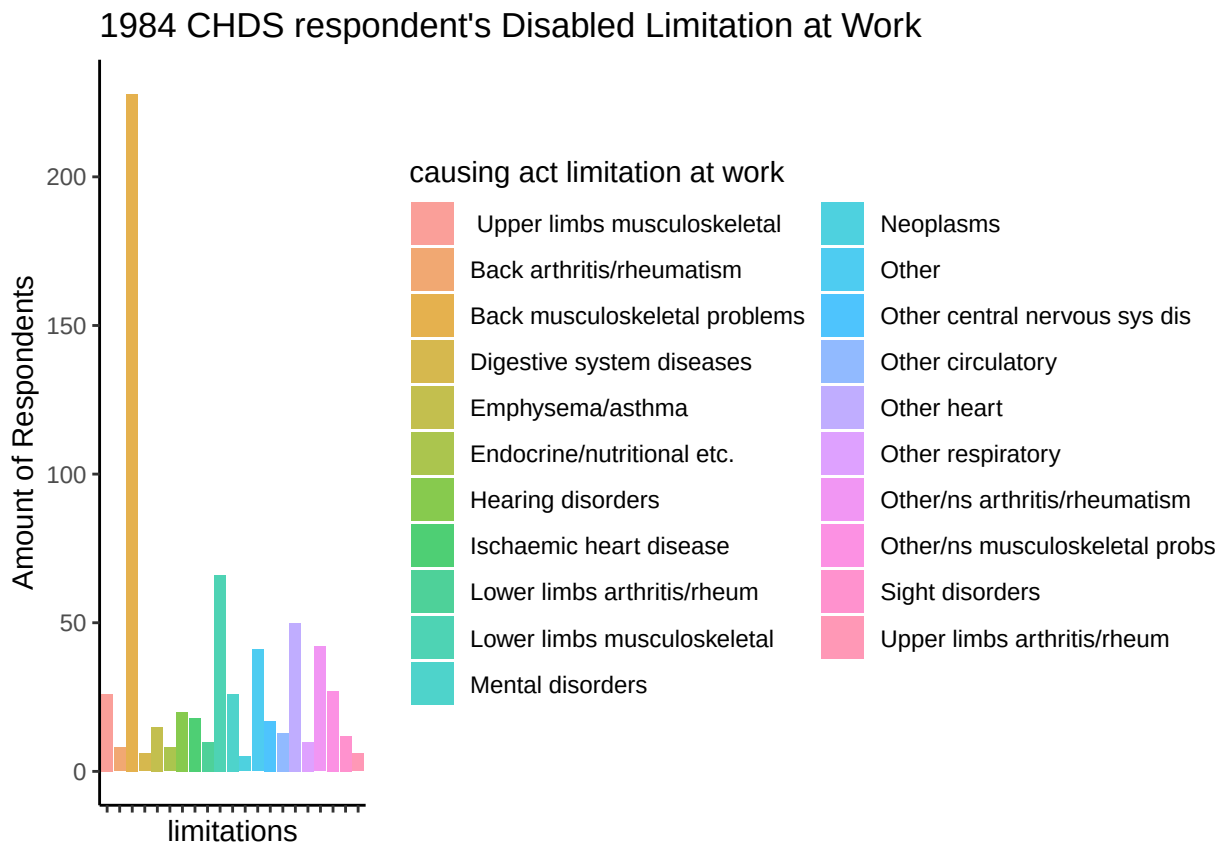


Figure 9: Various factors lead to work limitations for respondents

Figure 9 shows the distribution of respondents' factors that lead to behavioral limitations at work. We realized that there are many causes for respondents' behavioral limitations at work. Back musculoskeletal problems were definitely the most serious problem respondents encountered at work. More than two hundred respondents reported that back musculoskeletal problems seriously affect their behavior at work. In addition, issues related to upper limbs musculoskeletal, lower limbs musculoskeletal, other heart, other arthritis/rheumatism and other musculoskeletal probs also brought behavioral distress to the work of many respondents.

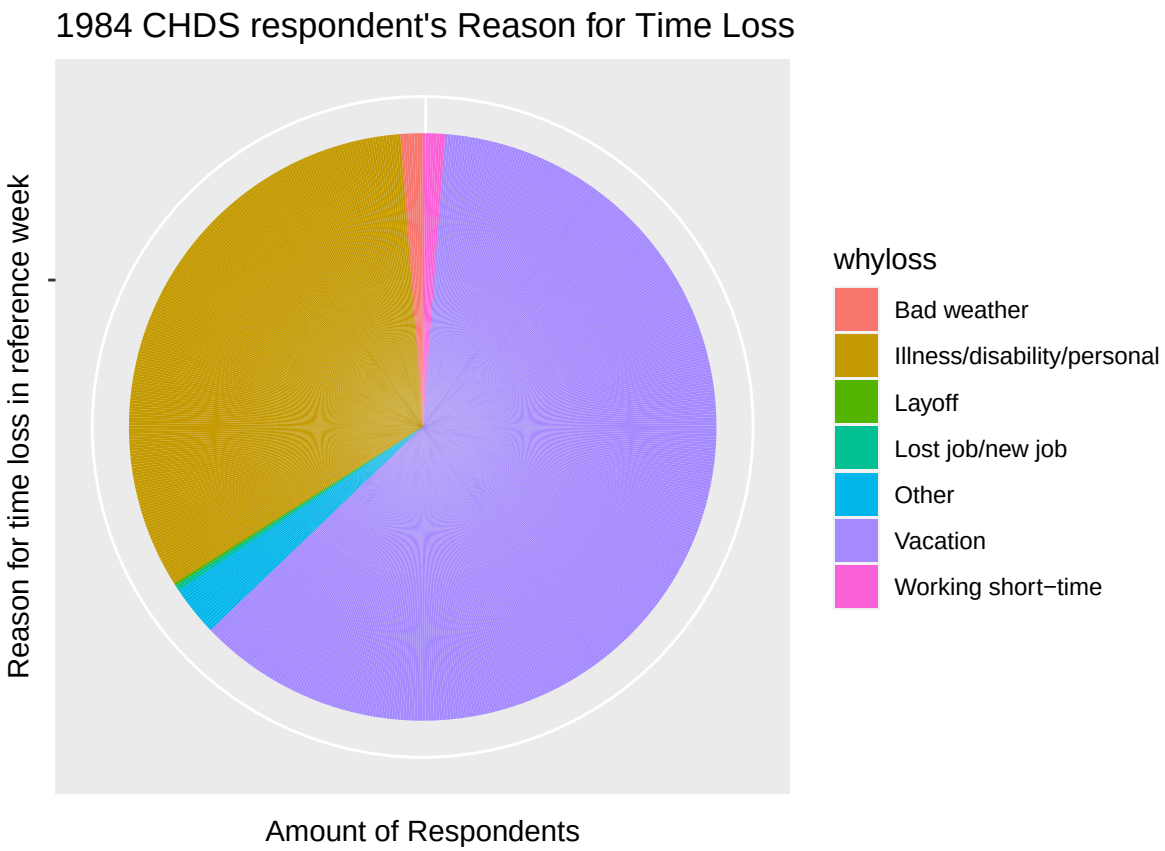


Figure 10: Illness and vacation are the most popular reasons for respondent's time loss

Figure 10 displays the reasons for people's time loss during the reference week. we can have an overview of seven reasons for people in terms of time loss, but the most popular one is vacation, followed by illness/disability/personal reasons. More than 400 people spend the longest time on vacation, which indicates that people are keen on leisure and recreational activities. The time loss of illness/disability/personal for many respondents indicates the inconvenience involved in the lives of people with disabilities, whose physical conditions can greatly affect their ability to work and live normally.

3 Simulation

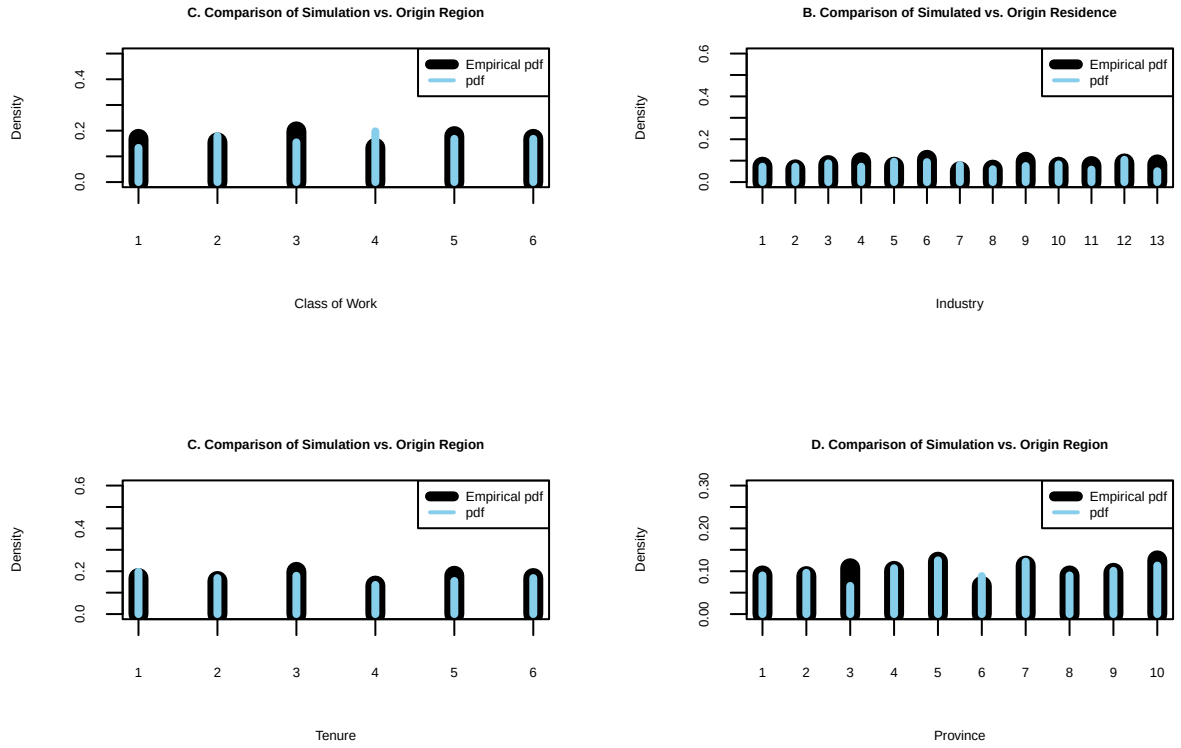


Figure 11: Simulation vs. Origin Data

Figure 11 and Figure 25 , Figure 26 in the appendix shows the comparison of simulated value and the actual values in the dataset for each of the variables.The blue line is the simulated data and the black line is the actual value for each of the variables. Figure 11 is shown above as an example of the simulation data and most of the simulated values are very similar to the actual values. Additionally, by the pointblank check, alll of the variables are indicated as correctly simulated, such that the simulation data is relatively accurate.

4 Model

4.1 Logistic regression model

Logistic regression is a regression model in which the response variable is a categorical variable. We usually evaluate the relationship between the categorical variable and the independent variable by using the logistic function as the probability estimation. In general, we apply logistic regression model to determine the likelihood of success of the dependent variable, as opposed to predicting the mean as in linear regression. In logistic regression, we set P to be the probability that an event will occur and 1-P to be the probability of the non-occurrence of that event. We are aware that the value of p is either 0 or 1, such as determining whether a person has a disease; if he or she has a disease, the probability that the patient has a disease is 1, and if not, the likelihood is 0. We use log(p/1-p) as the dependent variable to create a linear regression equation.

Here our response variable is work_hrs_wk, which represents the number of hours the respondents work per week. By observation, we know that this is a numerical variable. For the case where the response variable is a quantitative variable, we should use linear regression model to predict the relationship between the response variables and the predictor. However, linear regression requires a linear relationship between the independent variable and the dependent variable, while logistic regression does not require a linear relationship between the independent variable and the dependent variable. Since we found that most of our predictors are categorical variables, whereas there may not be a linear relationship between the response variable and these predictors. Therefore, we also decided to try to use logistic regression model to explore this dataset.

We applied the function summery to find out the distribution of the weekly working hours of the participants. We found that the mean and median of the weekly working hours of the respondents were around 40, therefore we set 40 hours as the cutoff, if the working hours of the respondents were less than or equal to 40, we defined their working hours as relatively short times, if the working hours of the respondents were is greater than 40 hours, we define his/her working time as a longer working period. We let p denote the probability of a longer working time of the respondent and created a logistic regression model based on this.

If the p-value for a variable is less than the significance level, we have enough evidence to reject the null hypothesis for the entire population. It indicates that there is a non-zero correlation between our response variable and the predictor. Changes in the independent variable are associated with changes in the dependent variable at the population level. This variable is statistically significant and probably a worthwhile addition to our regression model. Based on the summary of our model, it is observed that the predictors class_of_work, sex, ind, age, educ and whyloss have relative small p-value. Therefore, we would include these significant predictors to form our predicted model.

4.1.1 Table 2: Coefficients from the logistic regression model

Variable names	Estimate	Std. Error.	z value	Pr>|z|
(Intercept)	3.74017	1.55908	2.399	0.016442 *
as.factor(class_of_work)Own account	-1.31992	0.60401	-2.185	0.028870 *
as.factor(class_of_work)Paid worker, govt business	-3.35863	0.87384	-3.844	0.000121 ***
as.factor(class_of_work)Paid worker, govt non-business	-1.79158	0.78626	-2.279	0.022690 *
as.factor(class_of_work)Paid worker, private	-2.38876	0.58912	-4.055	5.02e-05 ***
as.factor(class_of_work)Unpaid family worker	-1.57160	0.97389	-1.614	0.106584
as.factor(sex)Male	1.11629	0.32549	3.430	0.000604 ***
as.factor(age)20-24 years	-1.52203	0.81218	-1.874	0.060929 .
as.factor(age)25-34 years	-1.66075	0.73152	-2.270	0.023192 *
as.factor(age)35-44 years	-1.32258	0.72861	-1.815	0.069493 .
as.factor(age)45-54 years	-1.55952	0.73833	-2.112	0.034666 *
as.factor(age)55-64 years	-1.89481	0.75640	-2.505	0.012244 *
as.factor(educ)Post-secondary cer/diploma	0.55063	0.40820	1.349	0.177361
as.factor(educ)Some post- secondary	0.94719	0.49780	1.903	0.057072 .
as.factor(educ)Some/Complete High School	0.11742	0.28492	0.412	0.680253
as.factor(educ)University Degree	1.32780	0.51085	2.599	0.009345 **
as.factor(whyloss)Illness/disability/personal	-1.96810	0.91059	-2.161	0.030668 *
as.factor(whyloss)Layoff	-15.17739	3956.18057	-0.004	0.996939
as.factor(whyloss)Lost job/new job	-15.52471	2786.63385	-0.006	0.995555
as.factor(whyloss)Other	-1.76389	1.06135	-1.662	0.096527 .
as.factor(whyloss)Vacation	-2.06838	0.91501	-2.260	0.023791 *
as.factor(whyloss)Working short-time	-1.49378	1.39963	-1.067	0.285851

4.1.2 Table 3: Stepwise Selection

Model	AIC
$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 X_{class\_of\_work} + \beta_2 X_{ind} + \beta_3 X_{sex} + \beta_4 X_{age} + \beta_5 X_{educ} + \beta_6 X_{whyloss}$	691.29
$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 X_{class\_of\_work} + \beta_2 X_{ind} + \beta_3 X_{sex} + \beta_4 X_{age} + \beta_5 X_{educ}$	684.47
$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 X_{class\_of\_work} + \beta_2 X_{ind} + \beta_3 X_{sex} + \beta_4 X_{educ}$	681.48

4.1.3 Table 4: Final Logistic Model

$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 X_{class\_of\_work} + \beta_2 X_{ind} + \beta_3 X_{sex} + \beta_4 X_{educ}$

After getting the predicted model, we derive all predictors that are relatively significant. In order to continuously optimize our existing model, we choose to calculate the AIC of every subset of the predictors to avoid the situation of overfitting and underfitting. We would like to choose the model with smaller AIC. At the same time, we expect to use fewer predictors but to capture more signal. By operating the function stepAIC, we conduct several reduced models as well as their AIC. Among these, we find the model $log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 X_{class\_of\_work} + \beta_2 X_{ind} + \beta_3 X_{sex} + \beta_4 X_{educ}$ has the smallest value of AIC and we would like to choose it as our final model.

4.2 Linear Regression Model

After analysis from the logistic model, the variable class of work, industry, sex and education level are found as significant factors of weekly working hours. To investigate a deeper relationship between the weekly working hours and other factors, in this section, we use a multiple linear regression model to check if there are linear relationship between variables, weekly working hours (wrk_hrs) is defined as response variable, and the rest of the variables are considered as potential predictors.

The multiple linear regression model is a model that has more than one explanatory variable that has the general linear equation $Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + ... + \beta_{ip} X_{ip} + e_i$ where p are the predictor variables, β is the coefficient of each of the predictor variable.

We use backward selecting to select the predictors based on the p-value. If the predictor has p-value < 0.05, then it is the potential preductor.

There are four assumptions which are linearity, independence, constant variacne and normal distribution that need to be satisfied, in order to conduct a more accurate linear regression model with unbiased β and minimum variance. These four assumptions can be checked by residual plots (the plot of estimation of population error) and quantile-quantile plot. There are also condition 1 and condition 2 to to be satisfied.
Condition 1 represents the conditional mean response is a single function of a linear combination of the predictors. It can be checked by Y vs. $\hat{Y}$ plot.
Condition 2 indicates the conditional mean of each predictor is a linear function with another predictor, which can be checked by pairs plots for numerical predictors.

4.2.1 Model Procedures

In order to conduct a linear regression model, it is necessary to follow the following steps to conduct an most accurate model.
Step 1: After summary the dataset and calculate the correlation, conduct QQ plot: It is for the assumption checks for linearity, independence, constant variance and normal distribution.
Step 2: If QQ plot fits, then check condition 1 and 2 If QQ plot does not fit, then transform variables and repeat step 1.
Step 3: If condition 1 and 2 satisfied, then check residual plots. If condition 1 and 2 are not satisfied, then back to step 2 to transform variables.
Step 4: If residual plots does not have patterns, then check outliers/leverage points/influential points If residual plots have patterns, then remove the predictor that doesn't have valid residual plot and regenerate model.

Step 5: Access multi-colinearity by testing VIF.
Step 6: If VIF $<$ 5, choose the bets fiited model as final model. If VIF $>$ or $=$ 5, then delete the predictor that has possible multi-colinearity, then regenerate model.
Step 7: Use variable selection to obtain the final model by checking AIC or partial F-test to obtain the final model.

4.2.2 Model Process

The initial model contains all of the variables from the cleaned dataset.So the initial linear model is

$Y = \beta_0 + \beta_1 X_{workclasses} + \beta_2 X_{ind} + \beta_3 X_{tenure} + \beta_4 X_{prov} + \beta_5 X_{educ} + \beta_6 X_{age} + \beta_7 X_{sex} + \beta_8 X_{limitation} + \beta_9 X_{whyloss} + \beta_1 0 X_{income} + e_i$

Since all of the potential predictors are categorical, the boxplots were conducted for each of the potential variables to check if there are obvious relationship between the response varaible and each potential predictor. Figure 12 shows two examples of the boxplots that were conducted.Figure 12 A shows the boxplot for time loss reasons vs. weekly working hours and Figure 12B shows the boxplot for class of work vs. weekly working hours. Notice that in Figure 12 B, the median of each variables have an obvious decreacing linear pattern, such that we can predict that the class of work propabaly is one of the predictor with the response variable. However, it is very hard to tell the correlation between the reponse variable with categorical predictors.

Figure 12: Examples EDA boxplots for potential categorical predictors

My model process is from
Full Model (initial model) -> Reduced model 1 -> Reduced model 2 (final model)
The assumptions were checked for each model following the procedure listed above.

4.2.2.1 Table 5: The significant variables from the FULL linear regression model

Variable names	Estimate	Std. Error.	t value	Pr>|z|
class of work: Own account	-10.19379		2.42558	-4.203
class of work:Paid worker,govt business	-11.08937	2.71583	-4.083	5.07e-05 ***
class of work:Paid worker,govt non-business	-9.44781	2.79592	-3.379	0.000776 ***
class of work:Paid worker,private	-9.62730	2.25051	-4.278	2.21e-05 ***
class of work:Unpaid family worker	-18.19073	3.88353	-4.684	3.51e-06 ***
ind:Agriculture	4.08346	2.25443	1.811	0.070612 .
prov:BC	0.34356	1.78627	0.192	0.847548
prov:MB	-3.78049	1.69850	-2.226	0.026413 *
prov:PE	-5.39687	3.08910	-1.747	0.081155 .
prov:QC	-3.51271	1.74001	-2.019	0.043970 *
prov:SK	-2.83705	1.58200	-1.793	0.073441 .
sex:Male	2.33678	0.98116	2.382	0.017557 *
limitation:Lower limbs arthritis/rheum	5.80432	3.49143	1.662	0.096964 .
limitation:Other circulatory	-7.23095	3.16754	-2.283	0.022802 *
income:$10000-14999	2.50182	1.43811	1.740	0.082450 .
income:$30000+	2.94451	1.61246	1.826	0.068350 .
whyloss:Other	-12.85149	4.29360	-2.993	0.002879 **

There are 10 variables in the model excluding the response variable, work_hrs_wk. By checking the p-value, there are 7 significant variables, which are class of work, province, industry, sex, limitations, income and time loss reason, as shown in Table 5.

4.2.2.2 Table 6: The significant variables from the REDUCED linear regression model 1

Variable names	Estimate	Std. Error.	t value	Pr>|z|
(Intercept)	401.328	46.914	8.555	< 2e-16 ***
class_of_workOwn account	-93.078	22.193	-4.194	3.16e-05 ***
class_of_workPaid worker, govt business	-107.325	24.716	-4.342	1.66e-05 ***
class_of_workPaid worker, govt non-business	-90.626	25.116	-3.608	0.000334 ***
class_of_workPaid worker, private	-93.727	20.485	-4.575	5.79e-06 ***
class_of_workUnpaid family worker	-159.107	35.673	-4.460	9.80e-06 ***
provMB	-36.882	15.510	-2.378	0.017729 *
provPE	-47.405	28.262	-1.677	0.094003 .
provQC	-32.967	15.805	-2.086	0.037417 *
provSK	-25.298	14.472	-1.748	0.080973 .
indAgriculture	51.782	20.248	2.557	0.010794 *
sexMale	23.364	8.665	2.696	0.007212 **
why_work_limLower limbs arthritis/rheum	55.690	31.864	1.748	0.081021 .
why_work_limOther circulatory	-65.633	28.882	-2.272	0.023417 *
why_work_limOther/ns arthritis/rheumatism	-41.876	21.277	-1.968	0.049516 *
income$30000+	23.929	13.907	1.721	0.085823 .
whylossOther	-123.388	38.945	-3.168	0.001612 **

From Table 6, there are 7 total variables in the reduced model 1, and all of the variables in reduced model 1 are significant, which is considered as a great redundancy in the reduced model 1. To obtain the best fitted model, AIC is used to select predictors by using the function 'stepAIC'

4.2.2.3 Table 7: REDUCED linear regression model 2

Model	AIC
$TransY = \beta_0 + \beta_1 X_{class\_of\_work} + \beta_2 X_{ind} + \beta_3 X_{sex} + \beta_4 X_{ind} + \beta_5 X_{income} + \beta_6 X_{whyloss} + \beta_7 X_{limitation}$	5834.05
$TransY = \beta_0 + \beta_1 X_{class\_of\_work} + \beta_2 X_{ind} + \beta_3 X_{sex} + \beta_4 X_{ind} + \beta_5 X_{income} + \beta_6 X_{whyloss}$	5815.75
$TransY = \beta_0 + \beta_1 X_{class\_of\_work} + \beta_2 X_{sex} + \beta_3 X_{prov} + \beta_4 X_{income} + \beta_5 X_{whyloss}$	5807.44

From Table 7, by comparing the AIC of each model, $TransY = \beta_0 + \beta_1 X_{class\_of\_work} + \beta_2 X_{sex} + \beta_3 X_{prov} + \beta_4 X_{income} + \beta_5 X_{whyloss}$ has the smallest AIC, which is 5807.44. Thus, the best-fitted model by AIC selection is $TransY = \beta_0 + \beta_1 X_{class\_of\_work} + \beta_2 X_{sex} + \beta_3 X_{prov} + \beta_4 X_{income} + \beta_5 X_{whyloss}$

4.3 Final Model

The following table shows the significant values of reduced model 2: $TransY = \beta_0 + \beta_1 X_{class\_of\_work} + \beta_2 X_{sex} + \beta_3 X_{prov} + \beta_4 X_{income} + \beta_5 X_{whyloss}$.

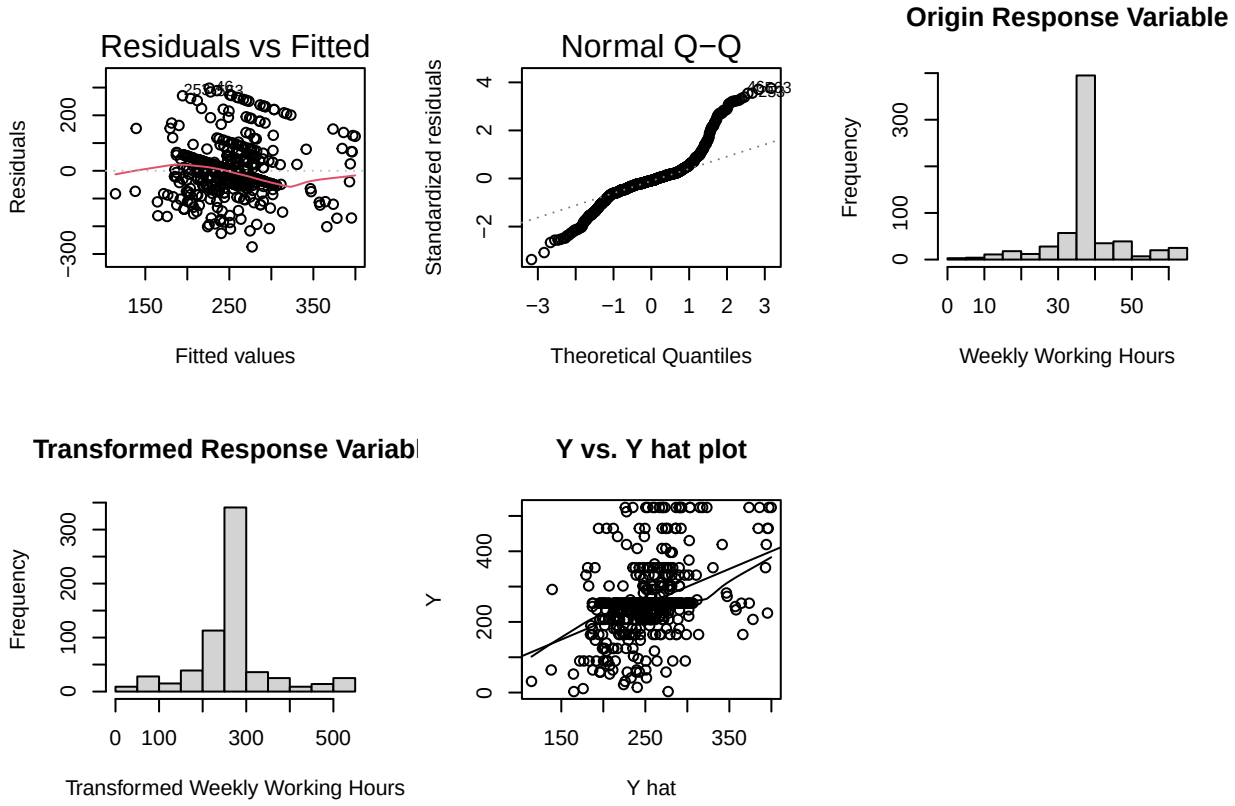
The last step is to use **partial F-test** evaluate the reduced model 2 is the better fitted model.
The null hypothesis is:
H0: $\beta = 0$ for all of the variables exist in the reduced model 1 and missing in the reduced model 2
By conducting the anova, the p-value of the partial F-test is 0.3381, which is significantly larger than 0.05. It indicates failing to reject H0 and we can conclude that reduced model 2 is better than reduced model 1.

Model	#non-significant predictors	Adjusted R^2	RSE	p-value
Reduced model 1	1	0.1339	82.76	
Reduced model 2	0	0.1299	82.95	0.3381

First of all, the p-value of the partial F-test is 0.3381, which is larger than 0.05 and it indicates that reduced model 2 is a better model. The adjusted R^2 indicates how does does the model explain the variance of the response variable (weekly working hours). The closer to 100%, the more appropriate the model is. From the above table, reduced model 1 has the adjusted R^2 of 13.39% and reduced model 2 has the adjusted R^2 of 12.99%, their adjusted R^2 are pretty close, however, the reduced model 1 slightly better explain the variance of the weekly working hours than reduced model 2.
Additionally, reduced model 2 has none of the non-significant variable, which also support that reduced model 2 is a better model than reduced model 1. Thus, reduced model 2 is the final model for the analysis and the following is the table of reduced model 2.
The smaller residual standard error (RSE), then more accuracy of the model is. From the above table, reduced model 2 has smaller residual standard error, such that reduced model 2 has higher accurat.
Thus, the final model is $TransY = \beta_0 + \beta_1 X_{class\_of\_work} + \beta_2 X_{sex} + \beta_3 X_{prov} + \beta_4 X_{income} + \beta_5 X_{whyloss}$
As the conclusion, the **reduced model 2 is the best fitted linear regression model**.

4.4 Final Model Diagnostics

First, the residual plot is used to check if assumptions hold for reduced model 2, since reduced model 2 is the best-fitted model. The assumptions hold if there is no discernible pattern seen in the residual plot.



The first graph is the residual vs. fitted value plot, which has no discernible patterns, which indicates that the popular errors have mean of zero, the errors of the fitted values are random and independent, also that the errors have constand spread around the conditional mean.

The second graph is the Normal Quantile-Quantile plot and it is used to test i the population errors are normally distributed. As shown above, the normal QQ plot is mostly on the line with minimal deviations at the ends.

On the second line, the first graph is the origin response varaible, which is not normally distributed, however, the final model has the trasformed Y, which is shown on the right graph on the second line. The histogram of the transformed Y is bell-shaped, symmetric and normally distributed, thus, the assumption of normal distribution is satified.

Ideally, Y should equals to $\hat{Y}$, the straight line on the Y vs. $\hat{Y}$ plot is $Y = \hat{Y}$, as the graph shown, the scatterplot mostly fits the line, such that condition 1 is satisfied, which indicates that there is single function of a linear combination of the predictor.

It's not applicable to check if this final linear regression model satisfies the condition 2, because condition 2 diagnostic requires pair comparison between numerical predictors, however, there is no numerical predictor in the final linear model.

In Figure 24 in the appendix, there are five the residual vs. predictor plots, which corresponds to the five predictors in the final linear model. There is no discernible patterns for the first four plolts. For the categorical predictors, they all have equal length of lines, thus the assumptions hold.

However, for the residual vs. whyloss plot, the length of the lines are not equal, such that the assumption of constant variance is not satisfied for the predictor whyloss. This indicates that there is a violation in the assumption and the model is not accurate for the whyloss predictor.

4.5 Final Model Interpretation

By the above analysis of logistic model and linear regression model, the final logistic model is $log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 X_{class\_of\_work} + \beta_2 X_{ind} + \beta_3 X_{sex} + \beta_4 X_{educ}$, which suggests that the weekly working hours is significantly affecting by class of work, industry, gender and education level.

By checking the residual plots, normal QQ plot Y vs. $\hat{Y}$ plot and condition 1 and condition 2, the final linear model mostly satisfies the assumptions, so I can conclude that the linear model mostly fits the dataset.

The final linear model is $TransY \sim \beta_0 + \beta_1 X_{class\_of\_work} + \beta_2 X_{sex} + \beta_3 X_{prov} + \beta_4 X_{income} + \beta_5 X_{whyloss}$.

Table of Final Linear Model

Variable names	Estimate	Std. Error.	t value	Pr>|z|
(Intercept)	412.136	40.100	10.278	< 2e-16 ***
class_of_workOwn account	-102.205	21.401	-4.776	2.23e-06 ***
class_of_workPaid worker, govt business	-126.840	22.074	-5.746	1.43e-08 ***
class_of_workPaid worker, govt non-business	-117.688	20.730	-5.677	2.10e-08 ***
class_of_workPaid worker, private	-113.940	19.074	-5.973	3.90e-09 ***
class_of_workUnpaid family worker	-151.251	34.668	-4.363	1.50e-05 ***
provMB	-41.883	15.176	-2.760	0.005951 **
provQC	-44.533	15.221	-2.926	0.003561 **
provSK	-31.260	14.103	-2.217	0.027013 *
sexMale	31.669	7.620	4.156	3.68e-05 ***
income$10000-14999	23.510	11.777	1.996	0.046333 *
income$30000+	26.320	12.926	2.036	0.042154 *
whylossOther	-128.874	37.113	-3.473	0.000551 ***

By the table of final linear model above, there are five predictors in the final model that has linear relationship with weekly working hours. The predictor class of work has negative correlation with the response variable. When all other predictors held constant, as the proportion of class of work increase, then the weekly working hours decrease. Unpaid Family worker has the most negative effect on the response variable, its estimate is -151.251, indicates that when all other predictors held constant, as the proportion of **unpaid family worker** increase by 1, then the weekly working hour decrease by 15125.1%.

The province also have negative correlation with the weekly working hour. Quebec has an estimate of -44.533, indicates that when all other predictor held constant, as the proportion of **Quebec** increase by 1, then the weekly working hours decrease by 4453.3%.

The other predictors could be interpreted by the same way. The gender **male** proportion has the positive correlation with the weekly working hours, the income level has mostly positive correlation with the response variable, while the income level of 5,000 dollars to 9,999 dollars has negative proportion with weekly working hours.

5 Results

5.1 Work Hour vs. Income

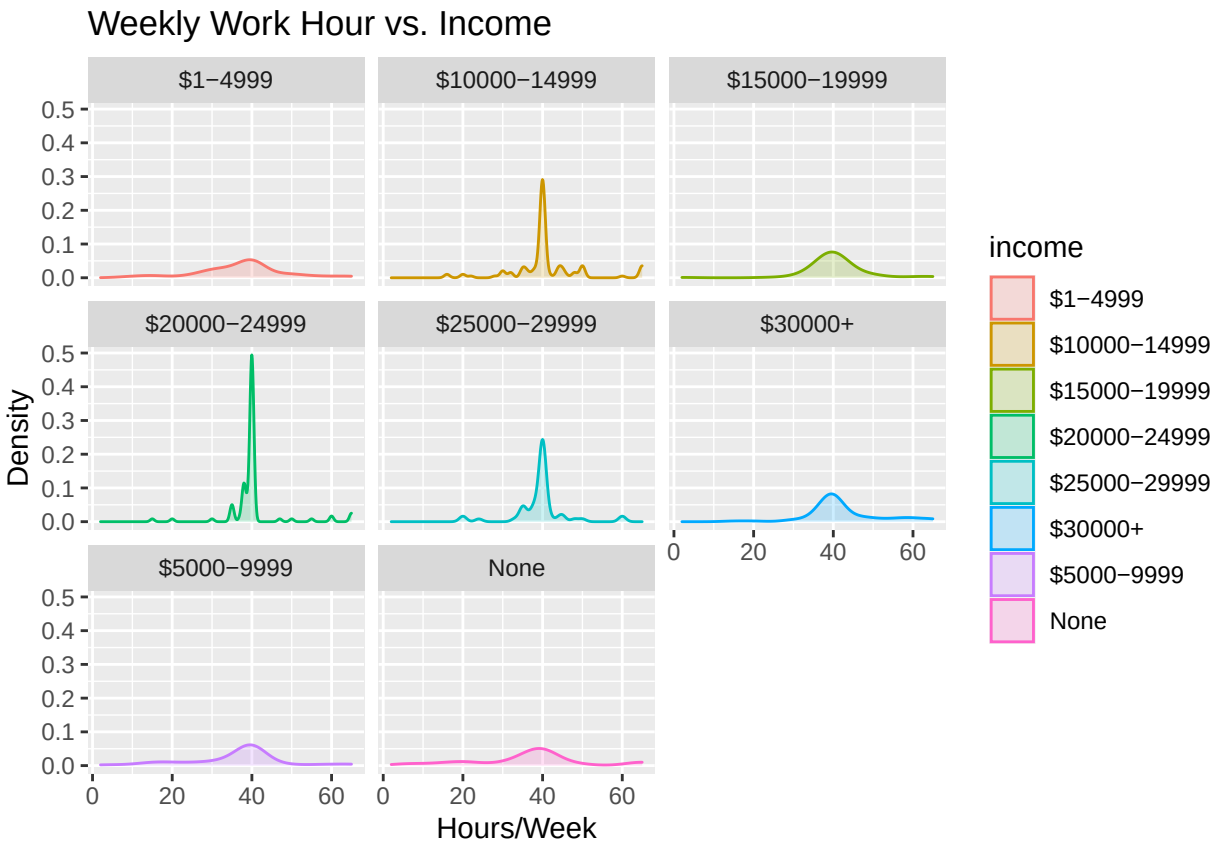


Figure 13: Work Hour vs. Income

Figure 13 indicates the distribution of the number of hours worked per week for respondents with different salary levels. Similarly, we find that the majority of distributions tend to be symmetric, with peaks at 40 h. However, for those with salaries below $10,000, we find that the proportion of people working 40 h per week is relatively low, with a significant number of respondents working longer or shorter than 40 h per week. For the high-income group earning more than $20,000, the proportion of respondents working 40 hours per week decreases as the salary level increases.

5.1.1 Income vs. Class of Work

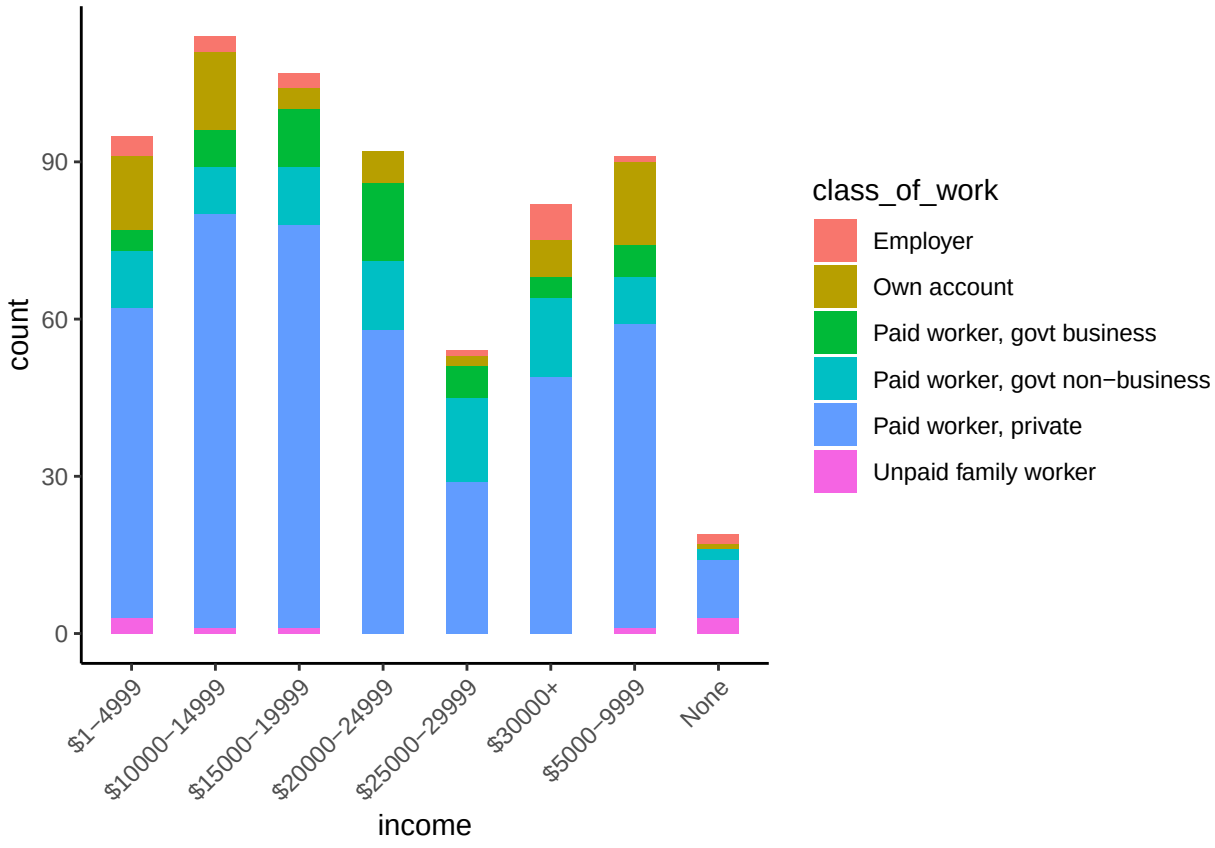


Figure 14: Income vs. Class of Work

Figure 14 shows that the public employees such as government paid workers have more proportion in the higher income level. However, private paid workers is the majority working type among all of the income levels.

5.1.2 Weekly Work Hours vs. Reasons Losing Working Hours

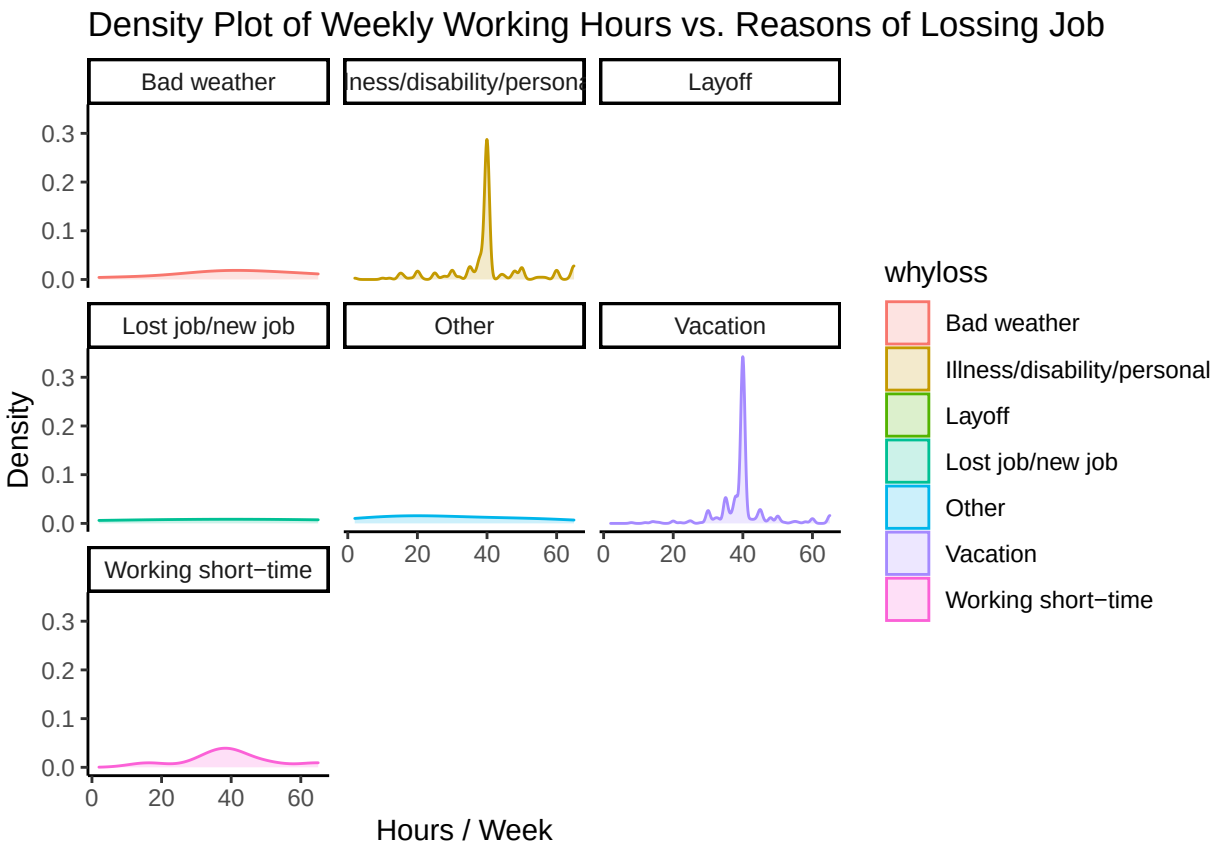


Figure 15: Work Hours vs. Reason Lossing Job

Figure 15 portrays the distribution of the weekly working hours of the respondents with different reasons for time loss during the reference week. As can be noticed from the image, most of the respondents lost time in the aspects of Illness/disability/personal issues and vacations. Only very few of them spend a lot of time on bad weather, lost job/new job or working short-time. For those who lost time in Illness/disability/personal issues and vacation, the distribution of their working hours was basically symmetrical, with a concentration of 40 h. While there were fewer respondents who worked less than or more than 40 hours per week.

5.1.3 Work Hours vs. Reasons Limit Work Activity

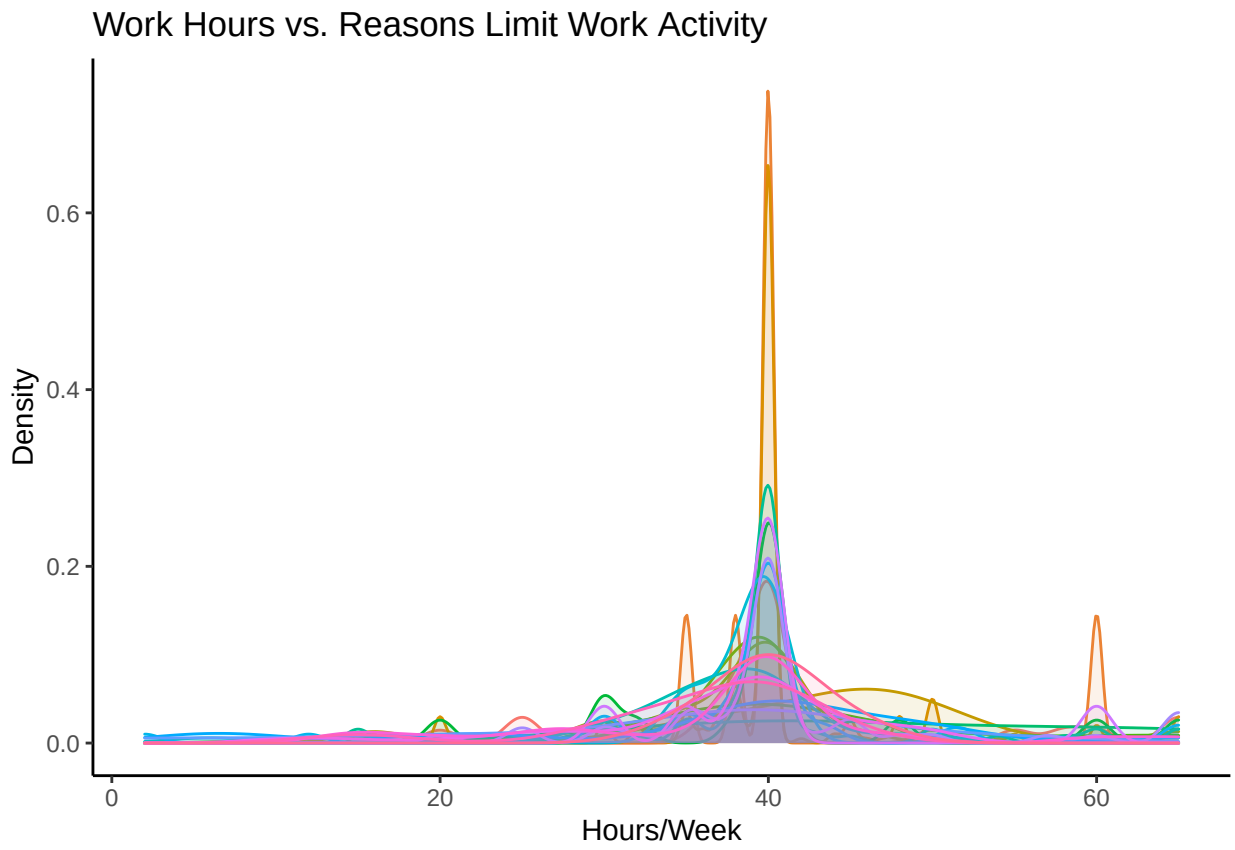


Figure 16: Work Hours vs. Reasons Limit Work Activity

Figure 16 delineates the distribution of the number of hours worked per week by respondents who had behavioral limitations at work for different reasons. First of all, we found that most of respondents reported that they had experienced serious back musculoskeletal problems at work. In addition to this, other heart problems, back arthritis/rheumatism and other musculoskeletal problems also caused great disturbance to the respondents' work status. We found that most of the distribution of working hours had a symmetrical shape, with the majority of respondents working approximately 40 hours per week. For the respondents who reported that Back arthritis/rheumatism affected their work status, we found two peaks in the distribution of hours worked, one at 35 hours and the other at 60 hours.

5.2 Weekly Work Hours vs. Class of Work with Different Genders

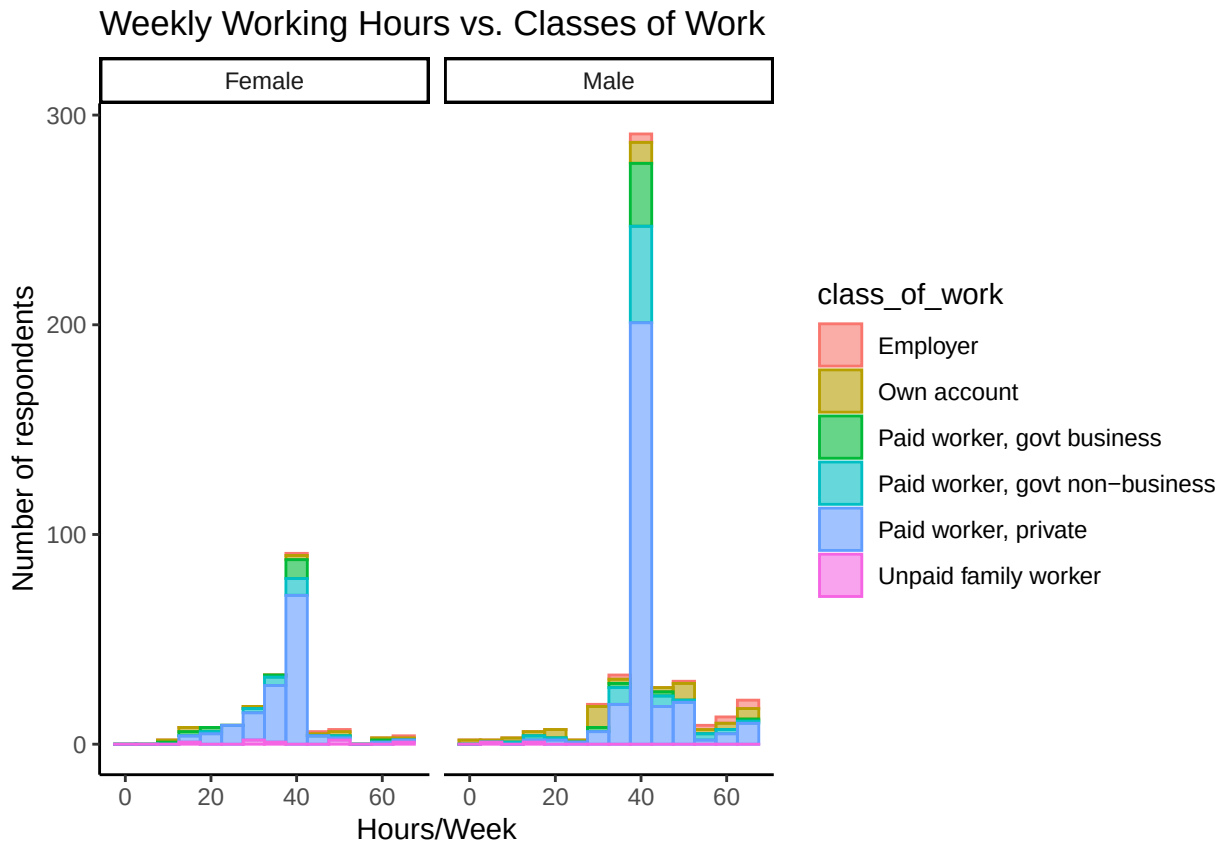


Figure 17: Weekly Work Hours vs. Class of Work with Different Genders

Figure 17 depicts the working hours of respondents of different genders, and we use different colors to classify the types of work of these respondents. First, by the difference in the height of the bar chart, we discovered that there is a huge difference in the total number of males and females, but most male and female respondents work around 40 hours per week. The amount of female respondents working less than 40 hours is larger than that of male respondents. Among the male respondents, a fraction of them works more than 40 hours, which is not seen in the trend of female respondents' working hours. Overall, although the majority of male respondents and female respondents work 40 hours per week, the average number of hours worked by men is longer than that of women, men seem to prefer to work for longer periods of time. A slightly larger proportion of female respondents work in the private sector than males, while more male respondents work in government than females. Both male and female work hours are close to symmetrical distribution.

5.2.1 Work Hours Per Week vs. Industry

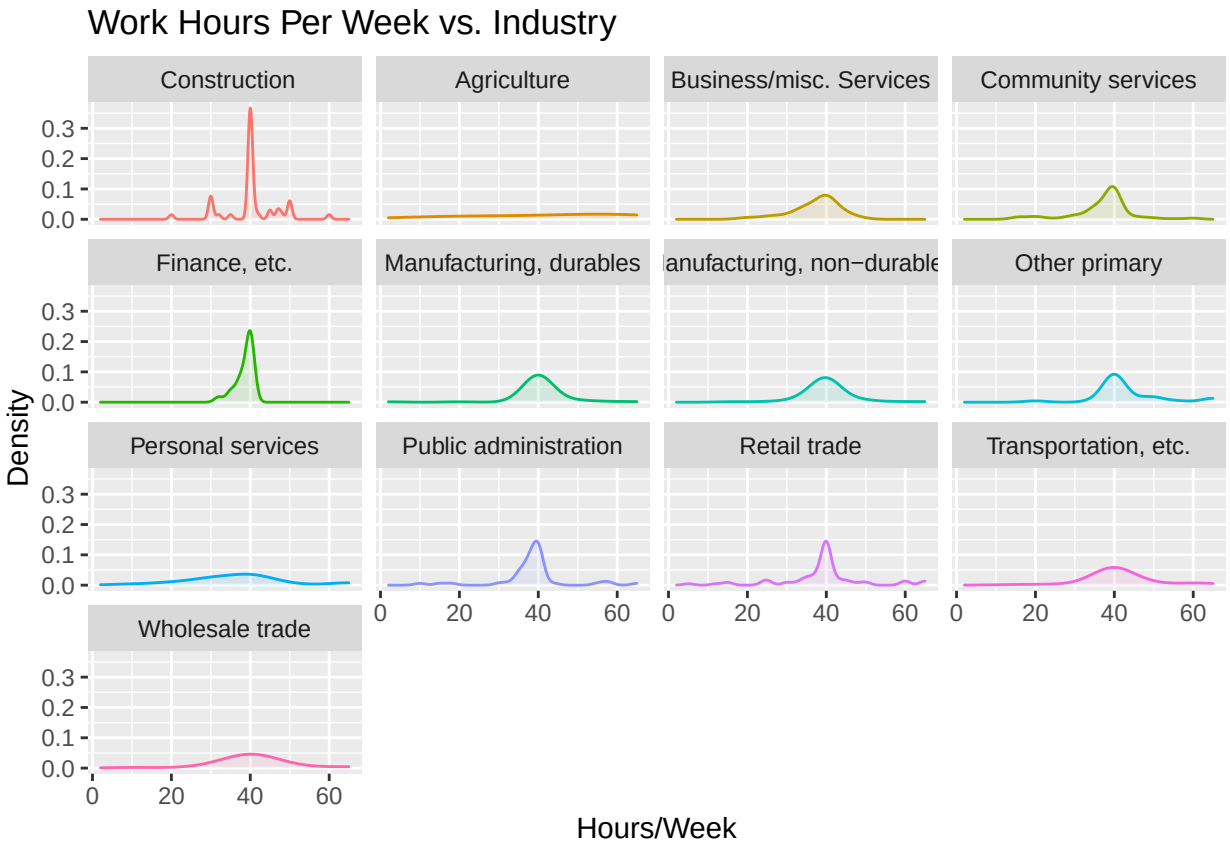


Figure 18: Work Hours Per Week vs. Industry

Figure 18 exhibits to us the weekly working hours of the respondents from different industries, and we reveal a symmetrical shape of working hours for almost all industries. For respondents in most industries, the working hours are mainly concentrated around 40 hours. The proportion of respondents working less than 40 hours is slightly higher in manufacturing than that in other industries. In each industry, very few respondents chose to work more than 40 hours or less than 40 hours. Moreover, a straight pattern was found for those working in agriculture, indicating that the working hours of respondents working in agriculture are more even.

5.2.2 Income vs. Tenure

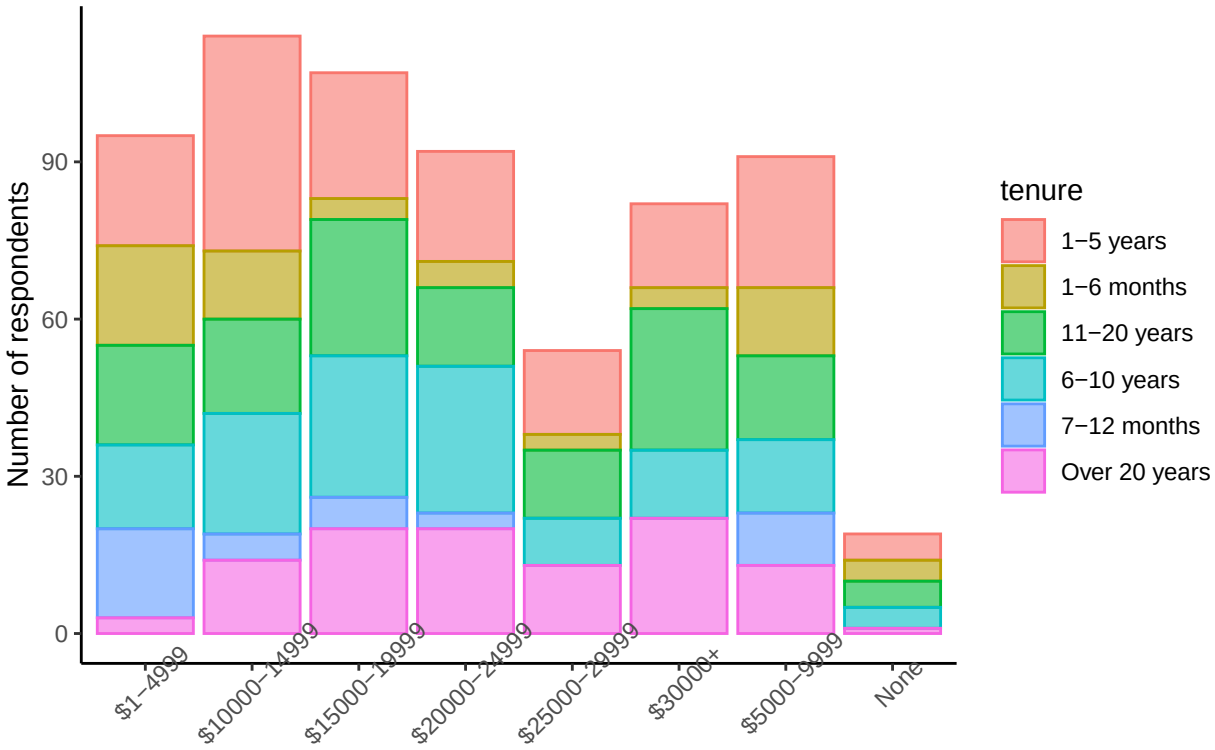


Figure 19: tenure and income

In Figure 19, there is an increasing of over 20 years tenure and decreasing trend of tenure that less than a year with higher income level. This could represents that participants in CHDS have higher salary with longer job tenures and the higher job satisfaction, verse vice.

5.3 Weekly Work Hours vs. Education

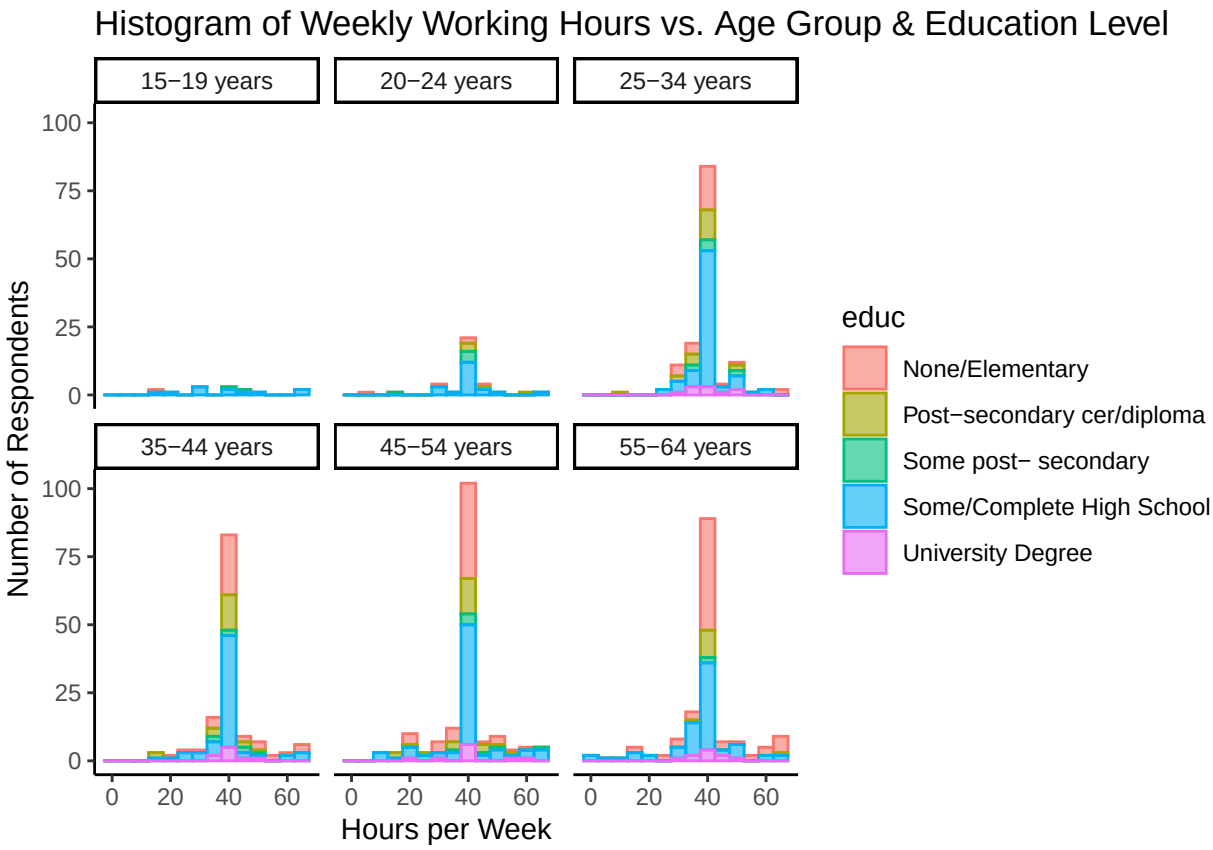


Figure 20: Weekly Work Hours vs. Education Level

Figure 20 depicts the working hours of respondents in different age groups, and we use different colored sections to identify the percentage of respondents in each education level. The first thing we can remark is that only a very small number of respondents are minors. The proportion of relatively young individuals who received high school education is significantly higher than that of the older residents. Younger people are more inclined to receive higher level of education. Among the older age group, the share of those who had only primary education was very high. The number of hours worked by respondents of different age groups is generally concentrated around 40 hours. The majority of people who work very long hours each week have not received any education or have received only primary education.

5.3.1 Income vs. Education Level

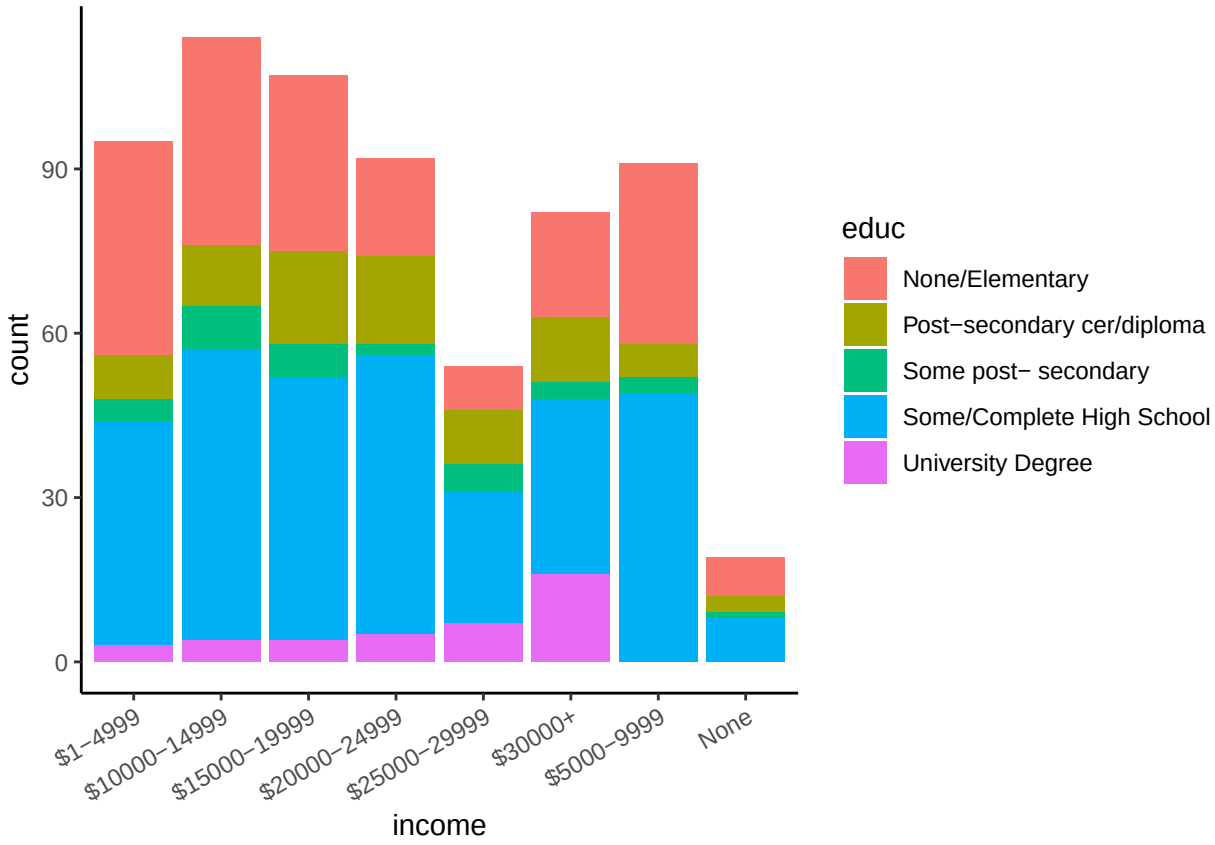


Figure 21: Higher education level lead to a higher income level

In Figure 21, the purple section indicates paritipants who complete university degree, as Figure 21 shown, there are more proportion of participants who complete university degrees in a higher income level. In the income level of 5,000 to 9,999 dollars, there doesn't have participants who complete university degree.

Also, higher income level generally have less proportion of particpants who have none or elementary degrees, which indicates that higher education level probably have higher salary in the disabled population.

5.3.2 Weekly Work Hours vs. Provinces

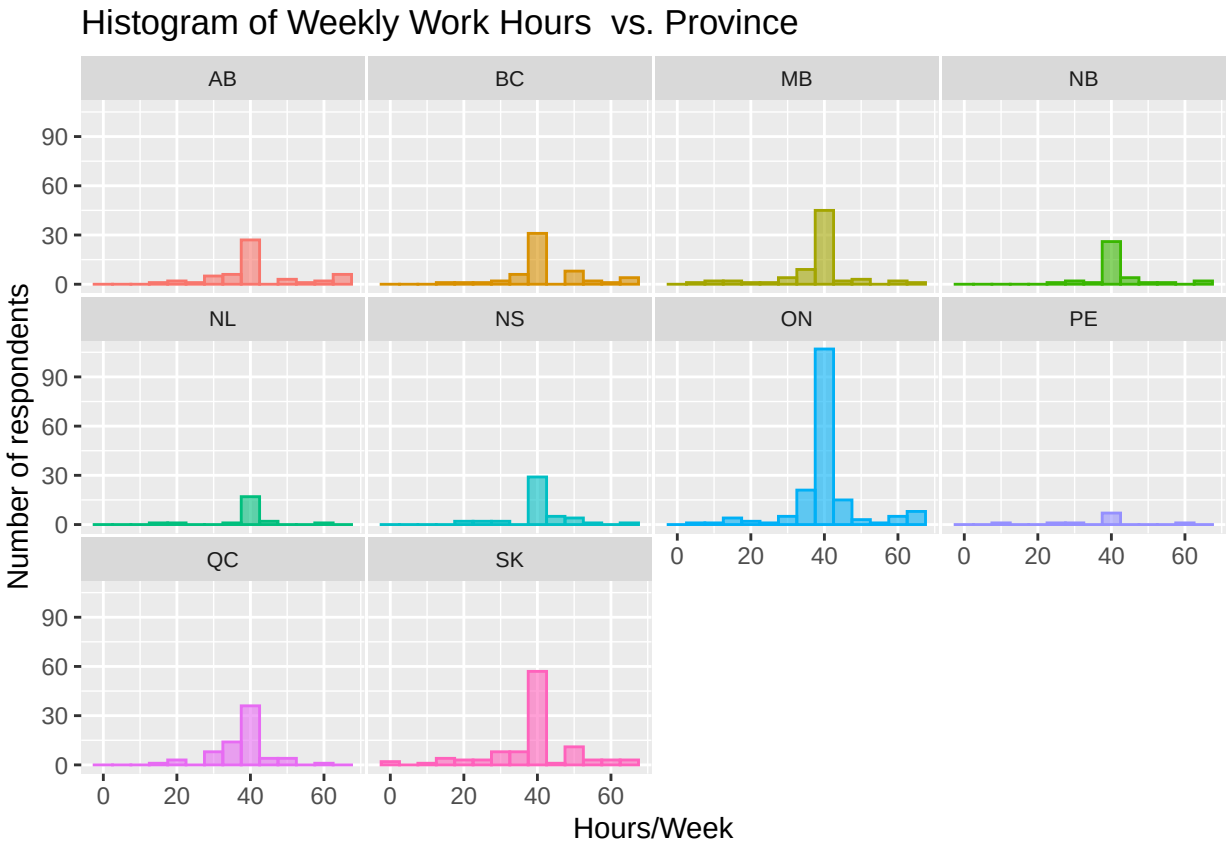


Figure 22: Weekly Work Hours vs. Province

Figure 22 portrays the hours worked by people in different provinces. Based on the height of the bins, we realize that there were the most respondents in Ontario, while Prince Edward Island had the fewest respondents. The number of hours worked by respondents in each province tends to be symmetrically distributed, with most people working 40 hours per week. However, we found that Ontario had the highest number of respondents working more than 40 hours. We hypothesize that this may be related to the fact that Ontario itself has a larger population, and the denser population brings more competition for work, forcing people to work longer hours to boost their work output. In New Brunswick, Newfoundland and Labrador, and Nova Scotia, the distribution of hours worked is more concentrated, with few people working more than 40 hours or less than 40 hours.

5.3.3 Weekly Work Hours vs. Tenure

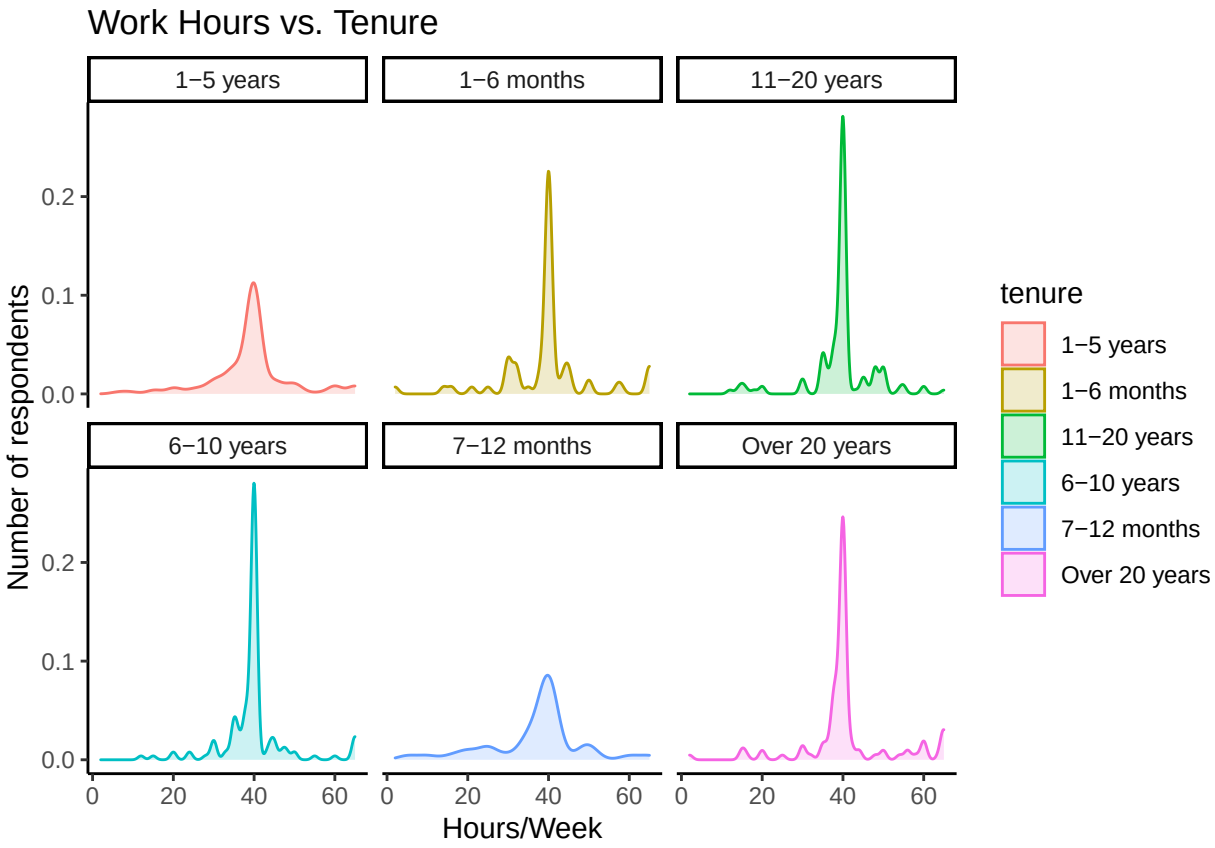


Figure 23: Work Hours vs. Tenure

Figure 23 indicates the weekly working hours of the respondents with different tenures. We find that for the respondents with different tenures, their working hours are mainly concentrated in 40 hours, and the distribution of working hours is generally symmetrical. Especially for respondents who have worked for 6-10 years, almost none of them work much more than 40 hours or much less than 40 hours. However, for respondents who have only been working for 1-6 months, the distribution of working hours is more volatile, with a significant proportion working less than 40 hours. We speculate that this may be because employees who have just joined the company still need a period of adaptation to accept normal intensity of work. We found that when tenure exceeded 6 years, the proportion of respondents who worked 40 hours per week tended to stabilize at about 0.2.

6 Discussion

6.1 Weekly Work Hours, Income and disability

The working hours does not have linearly effects on the income level. In the linear regression model, income level from 10,000 dollars to 14,999 dollars and income level over 30,000 dollars have positive linear correlation with the weekly working hours, which indicates that the proportion of disabled population who earn over 10,000 dollars per year would have higher weekly working hours. By Figure 13, people who earned 20,000 dollars to 24,999 dollars have highest density around 40 hours per week. Most of the participants have weekly working hours around 40 and does not deviate a lot from that, thus the difference in salary can be resulted by different hourly wage.

People with higher hourly wage could earn more salary with less working hours. The hourly wage can be influenced by the types of disabilities, class of work and education levels (Statistics Canada 2022), which will be discussed in the following few sections. In this section, how does disability affect weekly hours and income level is discussed.

To analyze how does the disability affect weekly hours, by Figure 10, disability problem is the second largest reason why participants lose working hours, which can cause a decrease in their salary. Due to the different types of the disabilities, such that their marginal productivity of labor decreases, which could lead to the lower salary than the normal population. By the Figure 9, most of the disabled population were affected by back musculoskeletal problems. Back musculoskeletal problems produce neck and back pain that can cause people having trouble sitting for a long time or moving heavy objects. However, these specific diseases are not one of the predictors in the linear regression model nor logistic model, so it is reasonable to predict that back musculoskeletal problem does not significantly affect working hours.

6.2 Weekly Work Hours, Class of Work

Weekly work hours and income level are closely related to class of work. There are five different kinds of class of work in the dataset, which are own account, paid worker in government business, paid worker in government non-business, unpaid family worker and paid private worker. By Figure 17, the proportion of private paid worker is the highest, and the proportion of employers is the lowest.

By the 2017 Labour Force Survey (Statistics Canada 2022), there were about 20% public sector employees and 64.7% of private sector employees, 15.1% of self employed workers and 1.35% of unpaid family workers. The Labour Force Survey aims for the whole Canadian Employee Population, which is very similar to the results with the disability population in 1984, so the disability does not make people have higher chance to be a housewife or a stay-at-home-dad. The average working hour per week for full time employment in Canada is 40 hours per week (Statistics Canada 2022), it also indicates that the Canadian population has similar working hours with the disabled population. The results reflect that the disabled population have very similar working hours and classes of work to the general population.

By Canadian Center for Policy Alternatives, people earn around 5% more in the public-sector jobs than working in the private sector (CCPA 2022b), such that, people who were paid government workers could have higher wage than the private workers when they have the same weekly working hours. By Figure 14, the proportion of private paid worker is the highest, however, the income level higher than 25,000 dollars has generally lower proportion of private paid worker and the proportion of paid government business worker increase.

By the results in both of the logistic model and linear regression model, class of work is an important predictor that affects weekly working hours of disability population in Canada in 1984. The linear regression model shows that class of work has negative linear correlation between the weekly working hours, in the other words, when all other predictors hold constant, the smaller proportion of the class of work is, the larger of the weekly working hour is. This can be caused by the needs to complete the same amount of work, but with less human resources, which also leads to the increase of average working hours.

6.3 Weekly Work Hours, Education Level and Tenure

Figure 21 shows that the higher income level comes with higher proportion of university degree. By the previous article from Statistics Canada, disabilities would have similar employment rate to the general population, however, the disabilities who have education level lower than university degree would have lower employment rate than the population without disability (Martin Turcotte 2015). The analysis results from Statistic Canada also support the finding from this analysis. Thus, a disabled individual can have higher income level after they complete a university degree.

The job tenure generally reflects the job satisfaction (S.Priya 2020). By Figure 23, population who have tenure of 6-20 years have the highest proportion working around 40 hours per week, in the contrary, the participants have less working hours when their tenure is less than 6 years. This indicates that disabled population tend to work more hours when they have a stable job, i.e. When their job tenure is over 5 years.

By Figure 19, as the income level gets higher, the longer tenure tends to take a larger proportion, the most obvious is over 20 years tenure. On the contrary, the shorter tenure tends to take a shorter proportion, and the most obvious is the 1-6 month tenure. Thus, the disabled population generally could have higher income with higher job satisfaction.

6.4 Limitation

The limitation of linear regression model: The predictor whyloss has some violation of assumption checking the residual plot, shown by the residual plot, the length of the line in the whyloss vs. residual plot are not equal length, then one of linearity, constant variance or independence might not hold. Thus, the linear regression model might not be the most fitted model.

Since all of the predictors are categorical variables, it is hard to use statistical models to analysis the relationship between weekly working hours and its affecting factors, the use of the linear model is lack of accuracy, the residual standard error (RSE) of the final model is very large, which indicates that the accuracy of the final linear regression model is moderate and it is not the best model that represents the dataset.

6.5 Future Improvement

Due to the limitation of the learning, this analysis is limited by two models, logistic model and linear regression model. The scientists can develop another model that can fits the dataset better, such that increasing the accuracy of the model to analysis the relationship between the weekly working hours of disability population and its other affecting factors, such as binomial model, etc.

The survey could also contains questions that determine the serverity of the disabilities, such that to make the analysis of the severity of the disabilities vs. labour force much easier.

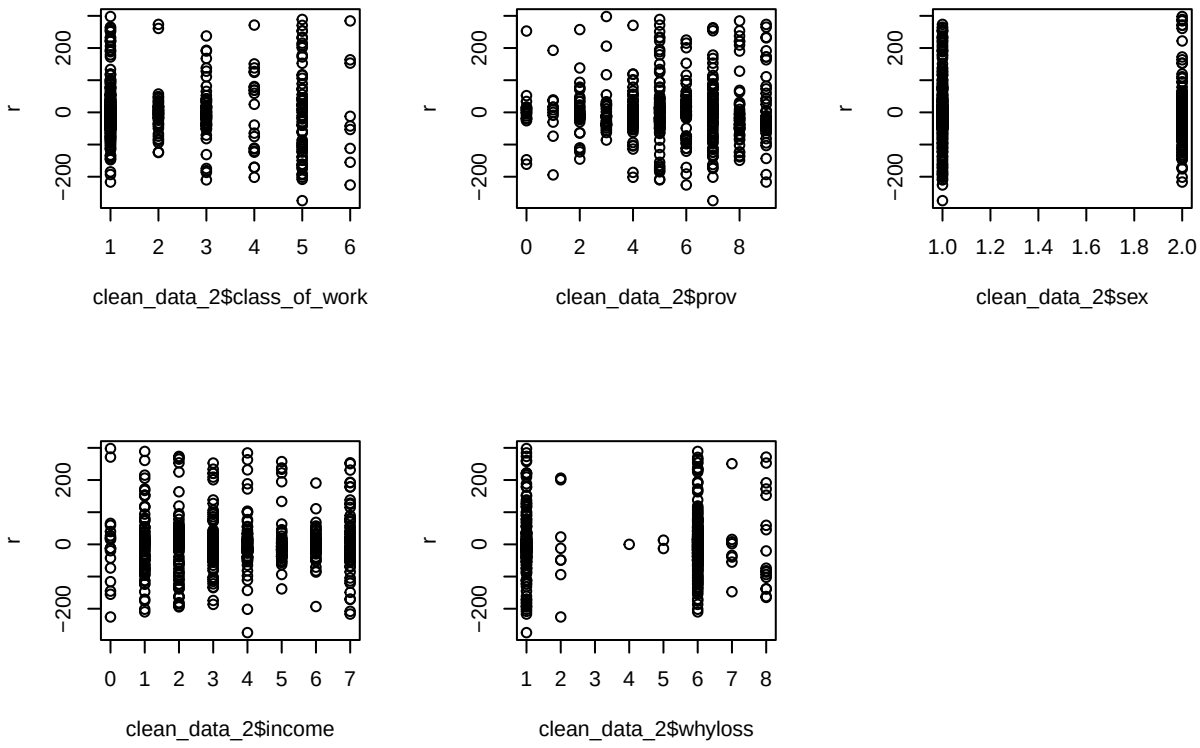


Figure 24: Predictor vs. Residuals Plots for Linear Regression Model

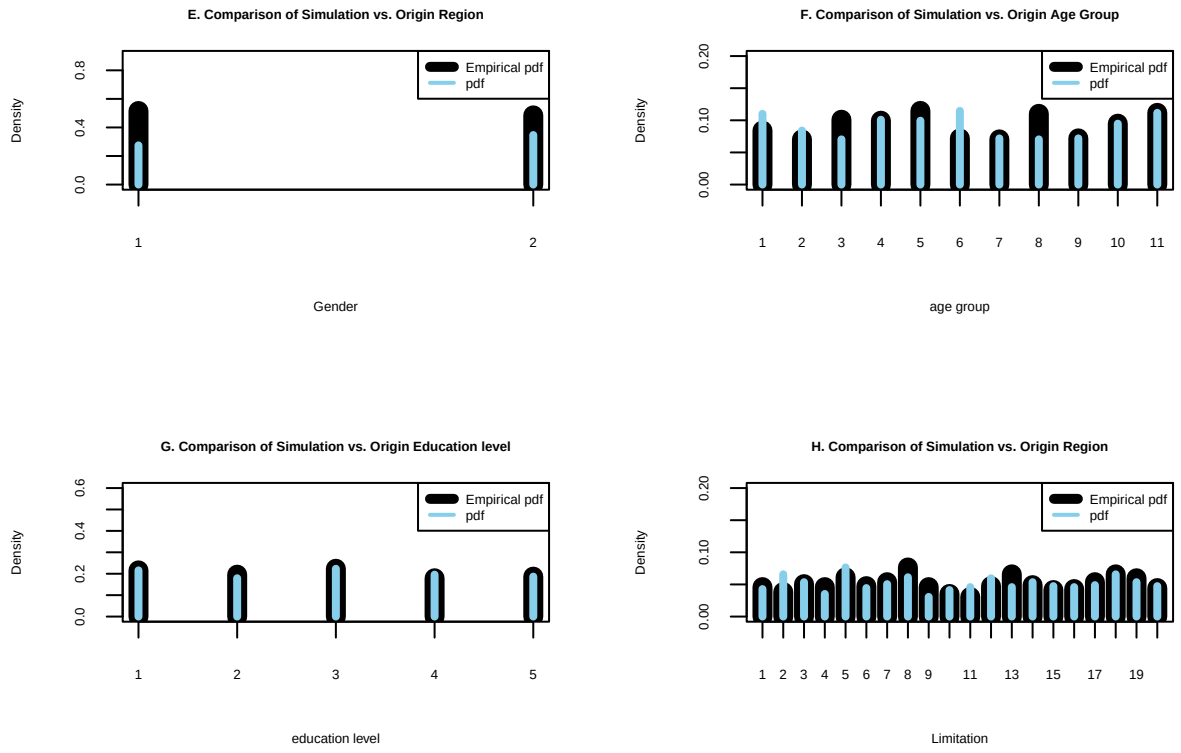


Figure 25: Simulation vs. Origin Data

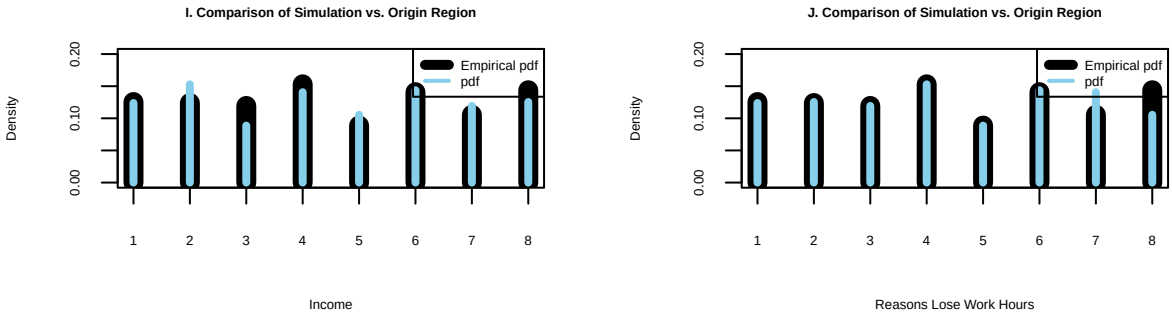


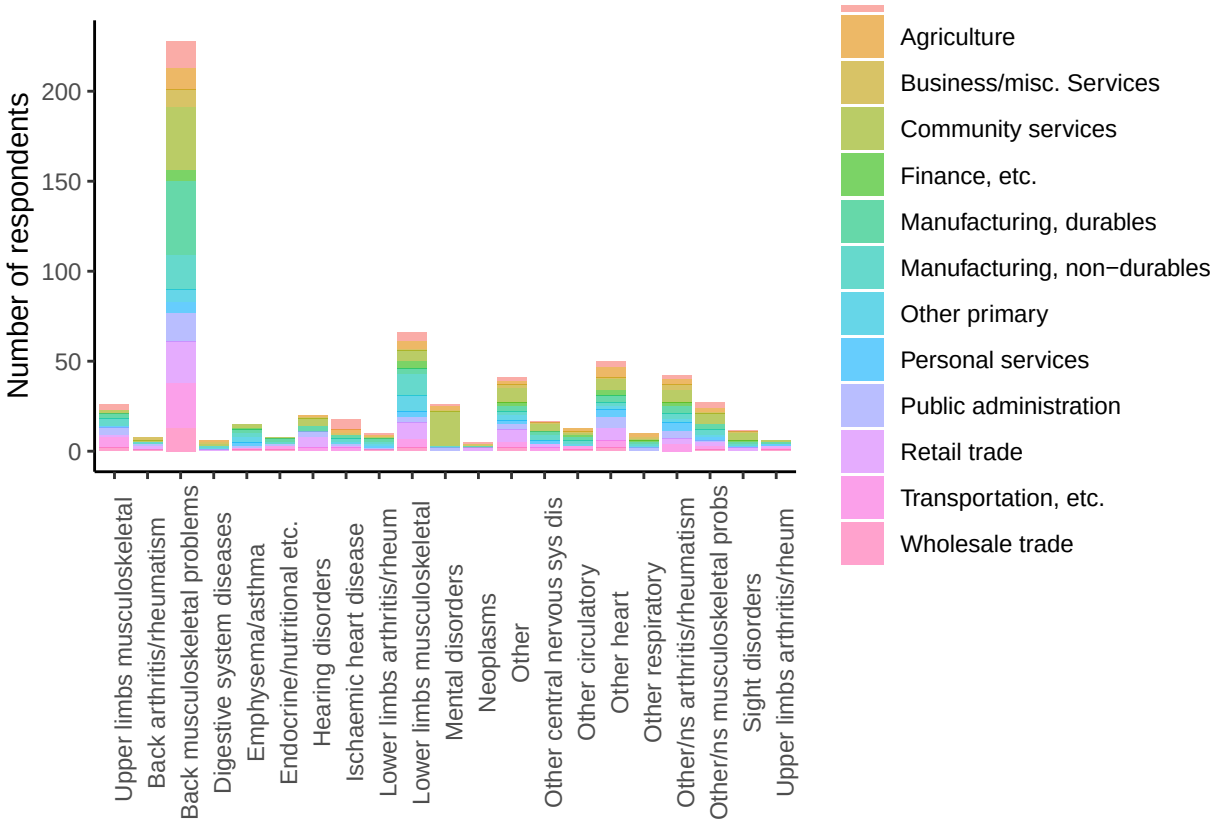
Figure 26: Simulation vs. Origin Data

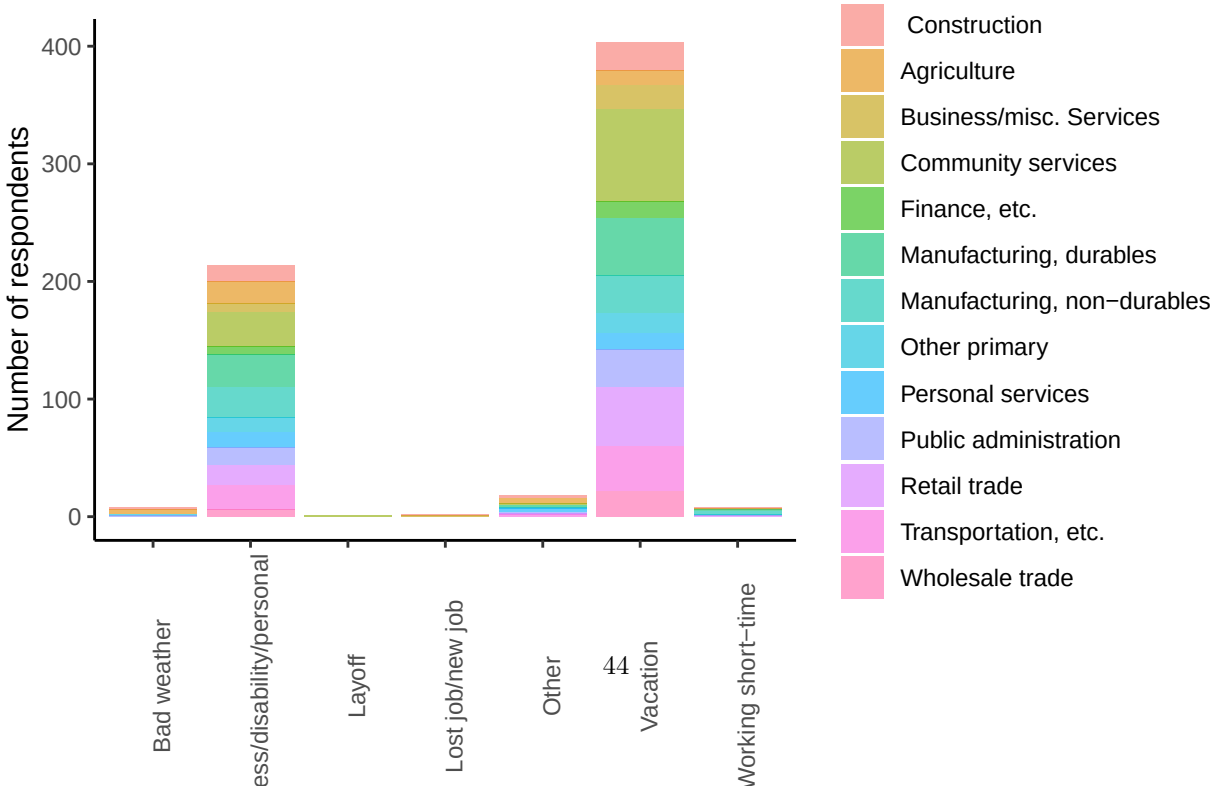
Appendix

.1 Linear regression model Residual Plot

.2 Simulation vs. Origin Data

.3 Other Variable Interactions





.3.1 income and work limitation reason

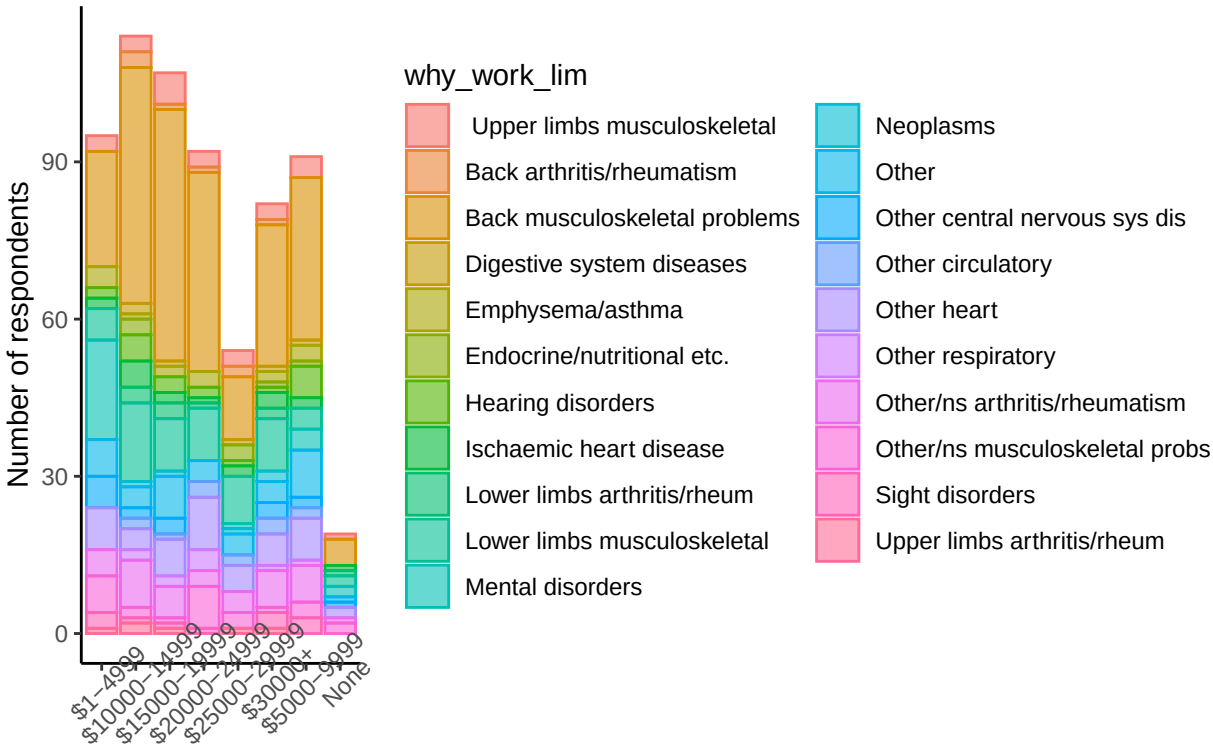
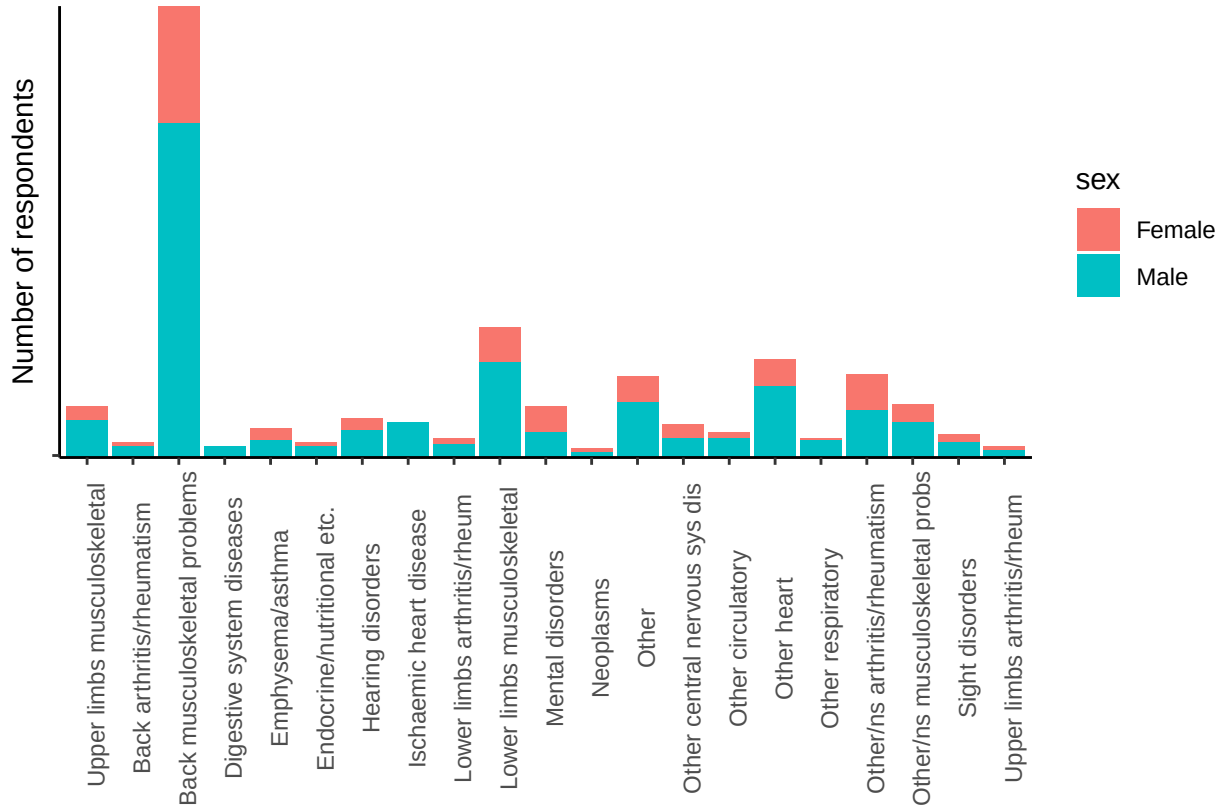


Figure 27: income and work limitation reason

.3.2 Sex vs. Why work lim



References

Canada, Statistic. 1884. "Report of the Canadian Health and Disability Survey." *Statistic Canada*. https://publications.gc.ca/collections/collection_2017/statcan/CS82-555-1986-eng.pdf.

Canada, Statistics. 2022. "Employement by Class of Worker." https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410002701.

CCPA. 2022b. "Public and Private Secpr Pay Differences." https://policyalternatives.ca/newsroom/updates/public-and-private-sector-pay-differences.

———. 2022a. "Public and Private Secpr Pay Differences." https://www150.statcan.gc.ca/n1/en/catalogue/89-654-X.

D. Dolson, J. P. Morin, P. Giles. 1884. "A Methodology for Survey Disabled Persons Using a Supplement to the Labour Force Survey." *Statistic Canada* 10 (2). https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1984002/article/14358-eng.pdf?st=tRyhfAf5.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Kassambara, Alboukadel. 2020. *Ggpubr: 'Ggplot2' Based Publication Ready Plots*. https://CRAN.R-project.org/package=ggpubr.

Martin Turcotte, Statistics Canada. 2015. "Persons with Disabilities and Emloyment." https://www150.statcan.gc.ca/n1/pub/75-006-x/2014001/article/14115-eng.htm.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

S.Priya, L.Sreejith. 2020. "Exploring the Relationship Between Job Tenure and Salary: An Empirical Analysis Related Paper the Effect of Emotional Labour on Organizational Commitment Among Call Ceniame." *International Journal of Management of Economics*. https://www.researchgate.net/publication/343679063_Exploring_the_Relationship_between_Job_Tenure_and_Salary_An_Empirical_Analysis.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2022. *Knitr:a General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.