Gender differences in the labor market in the 2010s

Data Analytics Capstone Course

Leonor Gutierrez 500925488

Ryerson University

Gender differences in the labor market in the 2010s

Table of Contents

Introduction **3**
Literature Review **3**
Conclusions 6
References 6
Dataset **6**
Labor Force Survey Sampling 7
Data set for this project 7
Attributes 8
Dataset descriptive statistics 14
Approach **19**
Step 1: Build data set 19
Step 2: Variable Analysis by Groups 19
Step 3: Hourly Wages by Sector, Industry and Occupation 20
Step 4: Multiple Regression Analysis 20
Step 5: Gender Proportions by Sector, Industry and Occupation 20
Results **21**
Step 1: Build data set 21
Step 2: Variable Analysis by Groups 21
Descriptive statistics 21
Outliers 21
Distribution of numeric variables 24
Correlation 30
Hourly Wages Relationship to Other Variables 32
Step 3: Hourly Wages by Sector, Industry and Occupation 41
Mann-Whitney U Test 41
Kruskal Wallis Test 42
Sector Analysis 43

Industry Analysis....46
Occupation Analysis....57
Step 4: Multiple Regression Analysis....66
Normalization....66
Training and Test Set....66
Variables excluded....67
Stepwise Regression....67
Step 5: Hypotheses Testing for Proportions....95
Two Proportions Z-Test....95
Sector Analysis....99
Industry Analysis....100
Occupation Analysis....103
Conclusions....106
Appendix....108
Post-Hoc Analysis....108
Industry....108

Gender differences in the labor market in the 2010s

Introduction

The main objective of this project is to explore differences by gender in the Canadian labor market with a focus on wages and participation by sector, industry, and job occupation at the beginning and ending of the 2010s decade. The specific research questions to solve are:

- What variables were relevant to explain hourly wages by gender when the decade started?
    - After 10 years, are these variables still important?
    - Are they the same for women and for men?
- What were the sectors, industries, and job occupations with highest hourly wages by gender at the beginning of the decade?
    - Are these the same by the end of 2019?
    - Which of these groups show more equality in wages by gender?
- What is the gender distribution by sector, industry, and job occupation?
    - What groups are more uneven?
    - Was there a change in the distribution over time?

The data source will be the “Labor Force Survey” records of September 2009 and September 2019 published by the government agency Statistics Canada. The analysis will require the application of statistical techniques such as multiple regression, and tests of hypothesis to compare distributions and proportions.

Literature Review

Differences in the labor market by gender in Canada and internationally have been studied extensively over the last decades. One of the more discussed topics within this subject is the gender pay gap, which is defined as the difference between women’s and men’s earnings from paid employment, expressed as a proportion of men’s earnings (Statistics Canada, 2019). Other variables that have been analyzed by researchers to describe differences by gender found in the labor market are educational attainment, marital status and motherhood, hours worked, job tenure, and participation by industry and occupations.

Regarding the gender pay gap, Statistics Canada (2019) states that there are several methods to calculate it. One of them consists of comparing the annual earnings of all employed women and men, which captures gender differences both in pay and in hours and weeks worked. In other words, it takes into consideration not only the net difference in wage, but also the impact in women’s earnings of having less disposable time to work due to home and family duties that have historically been covered by them. Measured this way, employed women aged 16 and older earned an average of $0.69 for every dollar earned by men in 2017.

Another way to estimate the gender pay gap consists of calculating the average hourly wages. It only captures the per unit (hour) price of labor, so it produces a smaller estimate than the method of comparing annual earnings. This approach is more appropriate to explore pay equity issues, that is, equal pay for work of equal or comparable value to the employer. In Canada, employed women aged 15 and older earned an average of $0.87 for every dollar earned by employed men on an hourly basis.

Moyser (2017) analyzed the relationship between wage inequality and occupations, and found that men out-earned women in every occupational group, except for managers and professionals in art, culture, recreation and sport (i.e., librarians, archivists, conservators and curators; writing, translating and related communications professionals; and creative and performing artists) and middle management occupations in production. Even in traditionally-female occupations like teaching, nursing, clerical, sales and services, the average hourly wages earned by men were greater than those earned by women. If women earned the same amount as men within occupations, the gender wage ratio would nearly reach equality at 0.97.

Motherhood has a persistent impact in wage disparity, although it has lessened in the last years. According to Statistics Canada (2017), in 1997, mothers with at least one child under the age of 18 earned $0.79 for every dollar earned by fathers, while women without children earned $0.88 for every dollar earned by men without children. By 2015, mothers earned $0.85 for every dollar earned by fathers, and women without children earned $0.90 for every dollar earned by men without children.

Drolet and Mumford (2012), with data from Statistics Canada Workplace and Employee Survey (WES) for private sector employees, discuss the relevance of age, fields of study and type of workplace on gender pay differentials.

Regarding age, they mention that is often assumed to be related to wages for several reasons, not least of which is the ability to acquire skills over time. Nevertheless, the results need to be interpreted with caution if women are more likely than men to have taken time out of the labor market; women's age may not accurately reflect the relationship between their actual work experience and earnings. At the same time, working part-time may be associated with less accumulated work experience. Of note, it has been identified that gender wage gap is seen to be smallest among young workers (those aged below 40 years) with little significant difference

About fields of study, many continue to be dominated by either men or women. While there have been notable changes in some discipline areas (such as biology, medicine, management and law), men and women tend to choose traditional disciplines, and this choice may also explain part of the gender differences in earnings.

Concerning the type of workplace, women are more commonly employed in low-paying workplaces than men. These results provide evidence that part of the variation in individual earnings attributable to, for example, differing levels of education may arise because high-wage workplaces disproportionately employ high-skilled workers.

Aside from gender pay gap, there are other aspects that show discrepancies and disparity between men and women at the workplace. Moyser (2017), mentions that over the last decades, it has become more likely for women to work. In 1950, only 22% of women in the core working ages of 25 to 54 years participated in the labour market, whereas in 2015 this number reached 82%. However, though women participation in the workforce have largely increased, their experiences of paid work still tend to differ from those of men, being shaped to a greater extent by their caregiving roles and/or their employers' presumptions of these roles. Based on Statistics Canada Labor Force Survey (LFS) data of 2015, she describes these differences from the optics of several variables:

- *Educational attainment.* The more educated, the more likely is that a woman works, though this likelihood is higher for a man with the same level of education. For example, in 2015, the employment rate of women with a high school diploma was 69% compared 83% for those with a university degree. The employment rate of men with a high school diploma was 82% compared to 90% for those with a university degree.

- *Hours worked.* Women generally perform fewer paid hours than men, as they spend more time on housework and childcare. They are also more likely than men to work part-time.

- *Long and short work absences.* Women's careers are interrupted more frequently than men's careers and for longer durations. These absences are frequently related to the role of women at home. In 2015, 22% of women who were absent attributed their absence to family or personal responsibilities, compared to only 9.3% of men who also cited those reasons.

- *Job tenure.* Women have nearly equivalent job tenure with current employer as men. Their average was 93.7 months, close to men's 94.9 months.

- *Self-employment.* Self-employment is less common among women than men, who represented close to two thirds of the self-employed population.

- *Industry.* There are clear differences in men and women participation across industries. Men are majority in the goods-producing sector, which includes industries related to the exploitation of natural resources, utilities, construction, and manufacturing. In contrast, women outnumber men in the services-producing sector, that consists of industries like wholesale and retail trade, finance, insurance, real estate, business, educational services, healthcare and social assistance, accommodation and food services, public administration, among others.

- *Occupation.* Women and men occupy distinct occupations, with women's typically being at lower levels than men's, even in industries dominated by women. For example, in the accommodation and food services industry, 59.7% of chefs and cooks were men, while 71.6% of food counter attendants, kitchen helpers and related support personnel were women.

- *Leadership and high-paying positions.* Women are underrepresented in leadership positions in the private sector, although not in the public sector. In 2015, 54% of

legislators and senior government managers and officials were women, whereas in the private sector they represented only 26% of senior managers.

Conclusions

There are many instances where differences by gender in the labor market have been identified. The relationship or possible impact of some variables to wage disparity has also been studied. When analyzing these attributes and their relevance to explain hourly wages by gender, we will see whether educational attainment, worked hours, job tenure, age, industry, and occupation have the same level of influence on the wages of men and women. Regarding wages disparity calculation, we will focus on pay equity, meaning that the method that compares average hourly wages will be applied. Furthermore, it will be explored whether some of the discrepancies by gender found in the past, have been persistent as of the end of the last decade.

References

Statistics Canada. 2019. “Measuring and analyzing the gender pay gap: A conceptual and methodological overview”. Studies on Gender and Intersecting Identities. Ottawa: Statistics Canada.
https://www150.statcan.gc.ca/n1/daily-quotidien/190830/dq190830d-eng.htm

Moyser, Melissa. 2017. “Women and Paid Work”. Women in Canada: A Gender-based Statistical Report. Ottawa: Statistics Canada. Catalogue no. 89-503-X. ISSN: 1719-4407.
https://www150.statcan.gc.ca/n1/pub/89-503-x/2015001/article/14694-eng.htm

Drolet, Marie and Karen Mumford. 2012. "The Gender Pay Gap for Private-Sector Employees in Canada and Britain." British Journal of Industrial Relations 50 (3): 529-553.

Dataset

A dataset will be created pulling records from the Public Use Microdata Files (PUMFs) of the Labor Force Survey (LFS) of September of 2009 and September 2019, published by Statistics Canada. The subjects covered in this survey are:

- Employment and unemployment
- Hours of work and work arrangements
- Industries
- Labour
- Occupations
- Unionization and industrial relations
- Wages, salaries, and other earnings

The survey is conducted nationwide, and the target population is the non-institutionalized population 15 years of age and over. Excluded groups represent less than 2% of target population: persons living on reserves and other Aboriginal settlements in the provinces, full-time members of the Canadian Armed Forces, Institutionalized population, and households in extremely remote areas with very low population density.

Deployment is performed monthly. The reference period to answer the questions is usually the week containing the 15th day of the month, and the collection period is the week following the reference period.

Labor Force Survey Sampling

This is a sample survey with a cross-sectional design (snapshot). It uses a probability sample that is based on a stratified multi-stage design. Each province is divided into large geographic stratum. The first stage of sampling consists of selecting smaller geographic areas, called clusters, from within each stratum. The second stage of sampling consists of selecting dwellings from within each selected cluster.

The LFS uses a rotating panel sample design so that selected dwellings remain in the LFS sample for six consecutive months. Each month about one-sixth of the LFS sampled dwellings are in their first month of the survey, one-sixth are in their second month of the survey, and so on. One feature of the LFS sample design is that each of the six rotation groups can be used as a representative sample by itself.

Within selected dwellings, basic demographic information is collected for all household members. Labor force information is collected for all civilian household members who are aged 15 and over.

Recently, the monthly LFS sample size has been approximately 56,000 households, resulting in the collection of labor market information for approximately 100,000 individuals. It should be noted that the LFS sample size is subject to change from time to time to meet data quality or budget requirements.

Data set for this project

Data sets from September 2009 and September 2019 were selected to represent the beginning and ending of the 2010s decade. Added together in a single table, the data from the two periods sums over 200,000 observations.

Attributes

Differences between LFS 2009 and 2019 variables

The LFS questionnaire is redesigned periodically. Variables out-of-date are removed, new variables can be introduced, and value sets of other variables can also be modified. LFS from September 2009 includes 82 attributes, whereas the one from 2019 consist of only 60 attributes. Here is a list of the attributes that were omitted by 2019, along with the reason of removal or change provided in the LFS documentation:

Attribute 2009	Description	Reason of Removal
ED76to89	Highest educational attainment	Out-of-date, replaced by EDUC90
NAICS_18	Industry of main job, 18 categories	Out-of-date, replaced by NAICS_21
NAICS_43 (Not available after 2015)	Industry of main job, 43 categories	Out-of-date
SOC80_49 (Not available after 2015)	Occupation at main job, 1980 Standard Occupational Classification, 49 categories	Out-of-date
SOC80_21 (Not available after 2015)	Occupation at main job, 1980 Standard Occupational Classification, 21 categories	Out-of-date
NOCS_06_25 (Not available after 2015)	Occupation at main job, 2001 National Occupational Classification for Statistics, 25 categories	Out-of-date, replaced by NOC_10
NOCS_06_47 (Not available after 2015)	Occupation at main job, 2001 National Occupational Classification for Statistics, 47 categories	Out-of-date, replaced by NOC_40
WHYPTOLD	Reason for part-time	Out-of-date, replaced by WHYPTNEW
YNOLKOLD	Reasons for not looking for work in the past four weeks, after looking for work in the last 6 months	Out-of-date, replaced with new variable
RELREFN	Relationship to reference person	Variables related to family and spouse/partner's labour force characteristics were removed
EFAMSIZE	Number of individuals in economic family, 1 to 5+	
EFAMEMPL	Number of employed persons	
EFAMUNEM	Number of unemployed persons	
SP_AGE	Age of spouse, if applicable.	
SP_LFSST	Labour force status of spouse, if applicable.	
SPED7689	Education of spouse, if applicable.	
SPED1990	Education of spouse, if applicable.	
SP_SOC80	Occupation of spouse, if applicable.	
SP_NOCS06	Occupation based on NOC-S2006	
SP_UHRSM	Spouse's usual hours at main job, employed,	
SP_UHRST	Spouse's usual hours at all jobs, employed	
SP_COWM	Spouse's class of worker at main job, employed	
SCH1624	At least one child age 16 to 24 in	

Besides the removals, there were variables which value sets were either expanded or collapsed:

- The **census metropolitan area (CMA)** variable expanded from the three largest CMAs in Canada (Montreal, Toronto, and Vancouver) to nine CMAs: Québec, Montreal, Ottawa, Toronto, Hamilton, Winnipeg, Calgary, Edmonton and Vancouver.
- Two **occupation** variables, based on the 2016 National Occupational Classification, were reintroduced: a generic, 10-category variable (NOC_10) and a detailed 40-category variable (NOC_40).
- The **labour force status** variable (LFSSTAT) was collapsed from six to four categories: employed, at work; employed, absent from work; unemployed; and not in the labour force.
- Two **industry** variables (NAICS_18 and NAICS_43) based on the 2012 North American Industry Classification System were replaced with one industry variable consisting of 21 categories, based on the same classification system.
- The **age of youngest child** variable (AGYOWNK) has been collapsed from six categories to four: under 6 years; 6 to 12 years; 13 to 17 years; and 18 to 24 years.
- The **school attendance** variable (SCHOOLN) has been collapsed from nine categories, which include information on the type of school attended, to three categories indicative of student status only: non-student; full-time student; and part-time student.

Finally, there is one new variable that was introduced in 2017:

- **Immigrant status** (IMMIG) distinguishes between recent immigrants, landed 10 or less years ago; established immigrant, landed more than 10 years ago; and non-immigrant.

Adjustments to attributes in project's data set

As noted, some variables value sets are not the same in 2019 as they were in 2009. The next tables show how those variable values will be matched to be comparable:

Census Metropolitan Area (CMA)

2009		2019	
CMA		CMA	
Code	Description	Code	Description
1	Montreal	2	Montréal
2	Toronto	4	Toronto
3	Vancouver	9	Vancouver
4	Other CMA or Non-CMA	1,3,5,6,7,8, 0	Quebec, Ottawa-Gatineau, Hamilton, Winnipeg, Calgary, Edmonton, Other

Labor Force Status (LFSSTAT)

2009		2019	
LFSSTAT		**LFSSTAT**	
Code	**Description**	**Code**	**Description**
1	Employed, at work	1	Employed, at work
2	Employed, absent from work	2	Employed, absent from work
3, 4, 5	3 = Unemployed, temporary layoff 4 = Unemployed, job searcher 5 = Unemployed, future start	3	Unemployed
6	Not in labour force	4	Not in labour force

Industry (NAICS)

2009		2019	
NAICS_18		**NAICS_21**	
Code	**Description**	**Code**	**Description**
1	Agriculture	1	Agriculture
2	Forestry, Fishing, Mining, Oil and Gas	2,3,4	2 = Forestry and logging and support activities for forestry 3 = Fishing, hunting and trapping 4 = Mining, quarrying, and oil and gas extraction
3	Utilities	5	Utilities
4	Construction	6	Construction
5	Manufacturing – durables	7	Manufacturing - durable goods
6	Manufacturing non-durables	8	Manufacturing - non-durable goods
7	Wholesale Trade	9	Wholesale trade
8	Retail Trade	10	Retail trade
9	Transportation and Warehousing	11	Transportation and warehousing
10	Finance, Insurance, Real Estate and Leasing	12,13	12 = Finance and insurance 13 = Real estate and rental and leasing
11	Professional, Scientific and Technical Services	14	Professional, scientific and technical services
12	Management, Administrative and Other Support	15	Business, building and other support services
13	Educational Services	16	Educational services
14	Health Care and Social Assistance	17	Health care and social assistance
15	Information, Culture and Recreation	18	Information, culture and recreation
16	Accommodation and Food Services	19	Accommodation and food services
17	Other Services	20	Other services (except public administration)
18	Public Administration	21	Public administration

Occupation (NOC)

2019		2009	
NOC_10		NOC_25	
Code	**Description**	**Code**	**Description**
1	Management occupations	1, 2	1 = Senior Management Occupations 2 = Other Management Occupations
2	Business, finance and administration occupations	3, 4, 5	3 = Professional Occupations in Business and Finance 4 = Financial, Secretarial and Administrative Occupations 5 = Clerical Occupations, Including Supervisors
3	Natural and applied sciences and related occupations	6	Natural and Applied Sciences and Related Occupations
4	Health occupations	7, 8	7 = Professional Occupations in Health, Nurse Supervisors and Registered Nurses 8 = Technical, Assisting and Related Occupations in Health
5	Occupations in education, law and social, community and government services	9, 10	9 = Occupations in Social Science, Government Service and Religion 10 = Teachers and Professors
6	Occupations in art, culture, recreation and sport	11	Occupations in Art, Culture, Recreation and Sport
7	Sales and service occupations	12, 13, 14, 15, 16, 17	12 = Wholesale, Technical, Insurance, Real Estate Sales Specialists, and Retail, Wholesale and Grain Buyers 13 = Retail Salespersons, Sales Clerks, Cashiers, Including Retail Trade, Supervisors 14 = Chefs and Cooks, and Occupations in Food and Beverage Service, Including Supervisors 15 = Occupation in Protective Services 16 = Childcare and Home Support Workers 17 = Sales and Service Occupations n.e.c. (not elsewhere classified), Including Occupations in Travel and Accommodation, Attendants in Recreation and Sport as well as Supervisors
8	Trades, transport and equipment operators and related occupations	18, 19, 20, 21, 22	18 = Contractors and Supervisors in Trades and Transportation 19 = Construction Trades 20 = Other Trades Occupations 21 = Transport and Equipment Operators 22 = Trades Helpers, Construction, and Transportation Labourers and Related Occupations
9	Natural resources, agriculture and related production occupations	23	Occupations Unique to Primary Industry
10	Occupations in manufacturing and utilities	24, 25	24 = Machine Operators and Assemblers in Manufacturing, Including Supervisors 25 = Labourer in Processing, Manufacturing and Utilities

Age of youngest child (AGYOWNKN)

2009		2019	
AGYOWNKN		AGYOWNKN	
Code	**Description**	**Code**	**Description**
1, 2	1 = Youngest child under 3 2 = Youngest child 3 to 5	1	Youngest child less than 6 years
3	Youngest child 6 to 12	2	Youngest child 6 to 12 years
4, 5	4 = Youngest child 13 to 15	3	Youngest child 13 to 17 years

	5 = Youngest child 16 to 17		
6	Not in labour force	4	Youngest child 18 to 24 years

Current Student Status (SCHOOLN)

2009		2019	
SCHOOLN		SCHOOLN	
Code	**Description**	**Code**	**Description**
1	Non-student	1	Non-student
2, 4, 6, 8	2 = Primary or secondary school, full-time 4 = University full-time 6 = Community college or CEGEP full-time 8 = Other full-time	2	Full-time student
3, 5, 7, 9	3 = Primary or secondary school, part-time 5 = University part-time 7 = Community college or CEGEP part-time 9 = Other part-time	3	Part-time student

Variables to be excluded from project's data set

LFS includes several questions about many aspects of unemployment, nevertheless the focus of this project is on the employed population. Thus, 18 variables related to this topic will be omitted.

Also, there are two variables that contain group age data. One of them includes information for all respondents (AGE_12), while the other only covers people between 15 and 29 years (AGE_6). The latter will be excluded.

Besides, there are 4 variables related to hours worked per week:

- UTOTHRS. Usual hours worked per week at all jobs
- ATOTHRS. Actual hours worked per week at all jobs
- UHRSMAIN. Usual hours worked per week at main job
- AHRSMAIN. Actual hours worked per week at main job

Only the first one (UTOTHRS) will be kept in the project's data set, since is the one that provides the usual hours considering all jobs, and not just actual hours worked on reference week or only at main job.

On the other hand, there are 2 attributes that give information about part-time employees. One of them identifies full time and part time workers (FTPTMAIN), and the other provides reasons for part-time work (WHYPT). Only the former will be kept, to explore if there is an impact in hourly wages.

There are 3 variables related to overtime worked in the reference week that will be excluded. There is no evidence to confirm whether the number of extra hours worked during this

timeframe is recurring or just a one-time event. And wages data focuses on “usual hourly wages”.

Lastly, 5 variables that provide details regarding reasons and time duration of current work absence will be excluded. This information is only available for people who declared being absent from work in the reference week. Furthermore, the impact of being absent from work will be captured by the labor force status variable (LFSSTAT), which distinguishes employees at work and employees absent from work.

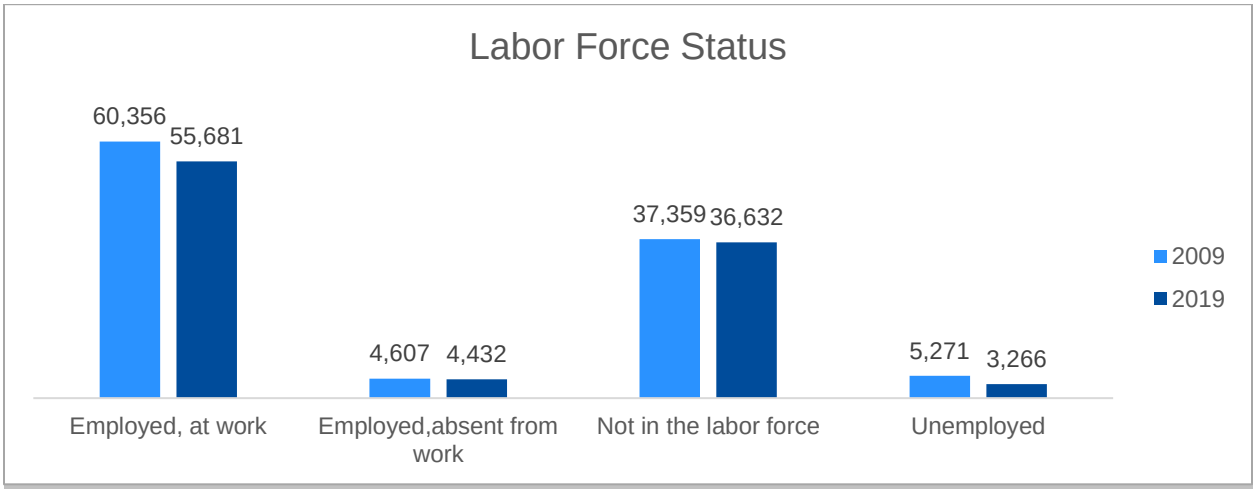
Complete list of attributes to be included in project’s data set

The new data set will consist of 27 attributes selected from the original microdata files:

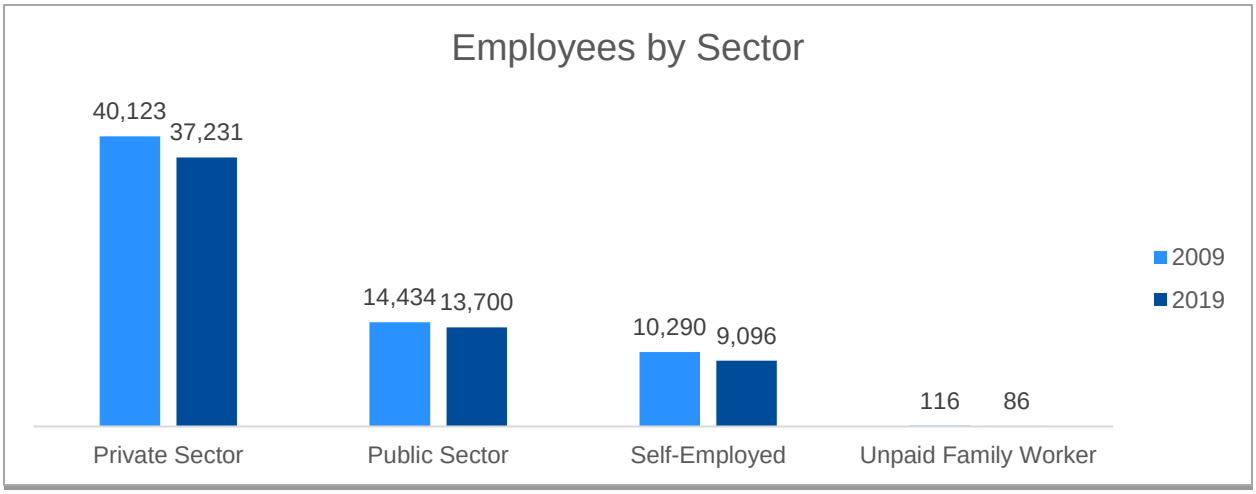
VARIABLE	DESCRIPTION	TYPE	COMMENTS
REC_NUM	Order of record in file	Nominal	
SURVYEAR	Survey year	Nominal	
SURVMNTH	Survey month	Nominal	
LFSSTAT	Labour force status	Nominal	2009 values adjusted to match 2019
PROV	Province	Nominal	
CMA	Three largest CMAs	Nominal	2019 values adjusted to match 2009
AGE_12	Five-year age group of respondent	Ordinal	
SEX	Sex of respondent	Nominal	
MARSTAT	Marital status of respondent	Nominal	
EDUC	Highest educational attainment	Ordinal	
MJH	Single or multiple jobholder	Nominal	
COWMAIN	Class of worker, main job	Nominal	
IMMIG	Immigrant status	Nominal	Only 2019 data
NAICS_18	Industry of main job	Nominal	2019 values adjusted to match 2009
NOC_10	Occupation at main job	Nominal	2009 values adjusted to match 2019
NOC_40	Occupation at main job	Nominal	Only 2019 data
FTPTMAIN	Full- or part-time status at main or only job	Nominal	
UTOTHRS	Usual hours worked per week at all jobs	Quantitative	
TENURE	Job tenure with current employer (months)	Quantitative	
HRLYEARN	Usual hourly wages, employees only	Quantitative	
UNION	Union status, employees only	Nominal	
PERMTEMP	Job permanency, employees only	Nominal	
ESTSIZE	Establishment size	Ordinal	
FIRMSIZE	Firm size	Ordinal	
SCHOOLN	Current student status	Nominal	2009 values adjusted to match 2019
EFAMTYPE	Type of economic family	Nominal	
AGYOWNK	Age of youngest child	Ordinal	2009 values adjusted to match 2019

Dataset descriptive statistics

The total number of records in LFS September 2009 and 2019 are 107,593 and 100,011, respectively. However, the focus of the research questions posed for this project is on the employed population. Looking at the "Labor force status" variable, it is possible to identify the respondents that declared to be employed:



Hourly wages data for self-employed workers are not shared on the public records of LFS. Also, there is a small portion of employees that self-identified as unpaid family worker. Because of the lack of wages data, both groups of workers will be excluded.



The total number of observations of only employed population from the private and public sector are 54,557 for 2009 and 50,931 for 2019, that sum up a global of 105,488 cases. For subsequent analysis, only these records will be included.

Missing Values

Here is the list of variables that contain missing values and how they will be handled:

Variable	Description	Number of cases with missing data	% of total cases (Out of 105,488 cases)	Solution
SCHOOLN	Current student status	3,002	3%	Replace with 4, label "Unknown"
AGYOWNK	Age of youngest child	65,144	62%	Replace with 5, label "Youngest > 24 years or no children"

Besides, only 2019 records have data for immigrant status (immig) and detailed occupation (noc_40) attributes. They should not be used in analysis where results will be compared to 2009, but they can provide insights when exploring the 2019 data set independently.

Categorical Attributes

The next tables show frequencies and proportions for the main categorical attributes in the dataset, by period:

Province

	Frequency (#)		Percentage (%)	
Province	**2009**	**2019**	**2009**	**2019**
AB	5,885	5,292	11%	10%
BC	5,996	6,096	11%	12%
MN	5,166	4,166	9%	8%
NB	2,827	2,484	5%	5%
NL	1,830	1,685	3%	3%
NS	2,632	2,565	5%	5%
ON	15,563	14,253	29%	28%
PEI	1,447	1,424	3%	3%
QC	9,370	9,294	17%	18%
SK	3,841	3,672	7%	7%

Age

	Frequency (#)		Percentage (%)	
Age Groups	**2009**	**2019**	**2009**	**2019**
15 to 19 years	3,572	2,576	7%	5%
20 to 24 years	5,055	4,215	9%	8%
25 to 29 years	5,649	5,288	10%	10%
30 to 34 years	5,495	5,550	10%	11%
35 to 39 years	5,870	5,605	11%	11%

40 to 44 years	6,305	5,477	12%	11%
45 to 49 years	7,491	5,461	14%	11%
50 to 54 years	6,835	5,545	13%	11%
55 to 59 years	4,784	5,450	9%	11%
60 to 64 years	2,509	3,754	5%	7%
65 to 69 years	736	1,384	1%	3%
70 and over	256	626	0%	1%

Sex

	Frequency (#)		Percentage (%)	
Sex	**2009**	**2019**	**2009**	**2019**
Female	27,615	25,432	51%	50%
Male	26,942	25,499	49%	50%

Marital Status

	Frequency (#)		Percentage (%)	
Marital Status	**2009**	**2019**	**2009**	**2019**
Married	26,641	23,434	49%	46%
Living in common-law	7,732	8,450	14%	17%
Widowed	635	540	1%	1%
Separated	1,576	1,410	3%	3%
Divorced	2,663	2,172	5%	4%
Single, never married	15,310	14,925	28%	29%

Education

	Frequency (#)		Percentage (%)	
Education Attainment	**2009**	**2019**	**2009**	**2019**
0 to 8 years	1,211	792	2%	2%
Some high school	5,892	3,891	11%	8%
High school graduate	11,614	9,980	21%	20%
Some postsecondary	4,454	2,946	8%	6%
Postsecondary certificate or diploma	20,090	19,724	37%	39%
Bachelor's degree	8,023	9,355	15%	18%
Above bachelor's degree	3,273	4,243	6%	8%

Sector

	Frequency (#)		Percentage (%)	
Class of Worker	**2009**	**2019**	**2009**	**2019**
Public sector employees	14,434	13,700	22%	23%
Private sector employees	40,123	37,231	62%	62%

Industry of main job

	Frequency (#)		Percentage (%)	
NAICS_18	**2009**	**2019**	**2009**	**2019**
Agriculture	718	709	1%	1%
Forestry, Fishing, Mining, Oil and Gas	1,566	1,481	3%	3%
Utilities	633	495	1%	1%
Construction	3,590	3,746	7%	7%
Manufacturing – durables	3,245	2,736	6%	5%
Manufacturing non-durables	2,783	2,308	5%	5%
Wholesale Trade	1,769	1,590	3%	3%
Retail Trade	7,149	6,125	13%	12%
Transportation and Warehousing	2,537	2,559	5%	5%
Finance, Insurance, Real Estate and Leasing	2,794	2,506	5%	5%
Professional, Scientific and Technical Services	2,160	2,494	4%	5%
Management, Administrative and Other Support	1,813	1,621	3%	3%
Educational Services	4,574	4,465	8%	9%
Health Care and Social Assistance	7,280	7,432	13%	15%
Information, Culture and Recreation	2,185	1,858	4%	4%
Accommodation and Food Services	3,947	3,540	7%	7%
Other Services	2,056	1,716	4%	3%
Public Administration	3,758	3,550	7%	7%

Occupation of main job

	Frequency (#)		Percentage (%)	
NOC_10	**2009**	**2019**	**2009**	**2019**
Management occupations	3,587	2,956	7%	6%
Business, finance, and administration occupations	9,822	8,085	18%	16%
Natural and applied sciences and related occupations	3,380	3,583	6%	7%
Health occupations	3,923	4,077	7%	8%
Occupations in education, law and social,	5,143	6,371	9%	13%

community and government services				
Occupations in art, culture, recreation, and sport	952	861	2%	2%
Sales and service occupations	14,578	12,782	27%	25%
Trades, transport and equipment operators and related occupations	8,766	8,055	16%	16%
Natural resources, agriculture, and related production occupations	1,581	1,572	3%	3%
Occupations in manufacturing and utilities	2,825	2,589	5%	5%

Quantitative Attributes

Real Values

Quantitative variables codes do not include decimal points and must be divided by 10 or 100 to get the actual numbers. Only “tenure” variable codes are equal to the real value.

VARIABLE	DESCRIPTION	CODE	REAL VALUE
UTOTHRS	Usual hours worked per week at all jobs	1-990	0.1-99.0 (one decimal implied)
TENURE	Job tenure with current employer (months)	1-240	1-240
HRLYEARN	Usual hourly wages, employees only	1-999999	$0.01-$9,999.99 (two decimals implied)

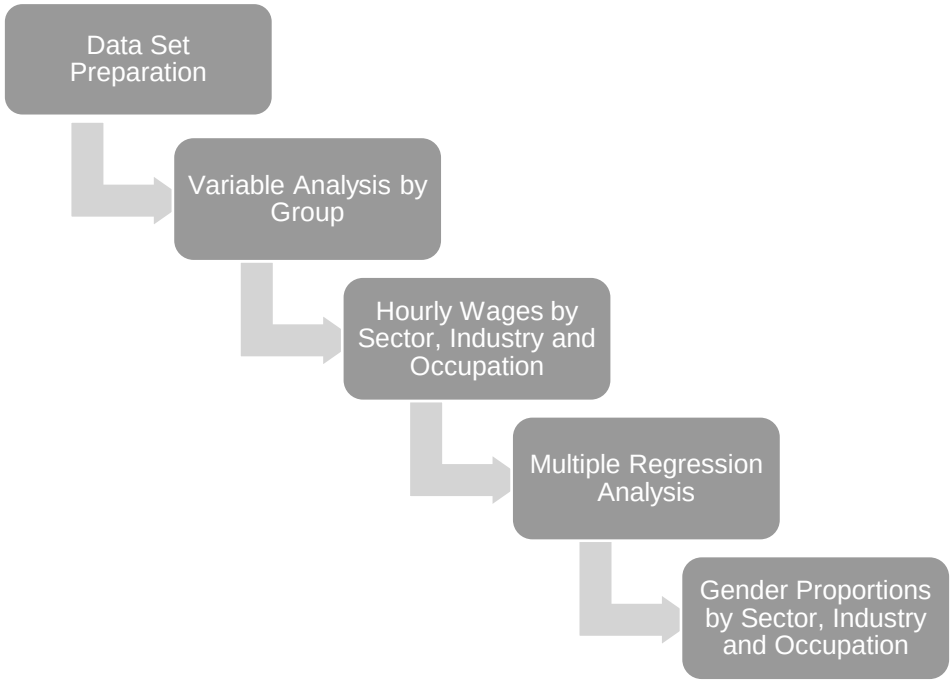
Statistics Summary

Here are the descriptive statistics for the main quantitative attributes:

VARIABLE	STATISTICS	2009	2019
Usual hours worked per week	MIN	0.4	1.0
	MAX	99.0	99.0
	AVERAGE	36.1	36.3
	STD DEV	11.2	11.2
Job tenure with current employer (months)	MIN	1	1
	MAX	240	240
	AVERAGE	85.3	86.8
	STD DEV	83.4	82.6
Usual hourly wages, employees only	MIN	2.0	3.0

	MAX	115.4	108.0
	AVERAGE	21.5	27.5
	STD DEV	11.3	13.7

Approach



Step 1: Data Set Preparation

In this step, records from data sets of Septembers 2009 and 2019 will be added into a single data set. Variables that are not needed will be removed, while the values of variables that have different value sets for each year will be matched. The right type will be assigned to each attribute. And, only the observations of people employed in public and private sector, as specified previously, will be included.

Step 2: Variable Analysis by Groups

First, the four groups by year and gender to be studied will be identified:

1. 2009, Males

2. 2019, Males
3. 2009, Females
4. 2009, Females

For each group, analysis of outliers using boxplots will be performed and numeric variables distribution will be verified. Correlation analysis will be run to explore relationships between numeric and ordinal variables. Also, relationships between hourly wages and the rest of the variables will be explored, except sector, industry, and occupations, which are included on the next step.

Step 3: Hourly Wages by Sector, Industry and Occupation

To answer the research question concerning comparison of hourly wages by sector, industries, and occupations, boxplot analysis will be run. Besides, parametric or non-parametric tests will be employed depending on the distribution of hourly wages variable. If assumptions to do parametric tests are met, analysis of variance will be applied to determine significant mean differences among the two periods studied and gender. If these assumptions are not met, non-parametric tests like Mann-Whitney U and Kruskal Wallis will be employed instead to compare observations ranks.

Step 4: Multiple Regression Analysis

To answer the research question regarding the relevant variables that explain hourly wages by gender at the start and end of the decade, multiple regression analysis will be applied. A model for each group will be defined. Then, these models will be compared to explore if they include the same independent variables and analyze which variables obtained the highest coefficients among the four groups by year and gender.

Step 5: Gender Proportions by Sector, Industry and Occupation

Finally, to answer the research question about gender distribution by sector, industry, and occupation, two proportions z-test will be applied to identify significant differences by gender through time.

Results

Step 1: Data Set Preparation

R code was employed to build the data set for this project, integrating observations from two CSV files that contain records for September 2009 and September 2019. The original data sets and code to build the one used in this project can be found at this location:

https://github.com/leonorgs/leo-capstone-.git

The names of the files are:

- pub0909 (CSV records September 2009)
- pub0919 (CSV records September 2019)
- Capstone_Project_RCode (R script)

Step 2: Variable Analysis by Groups

Descriptive statistics

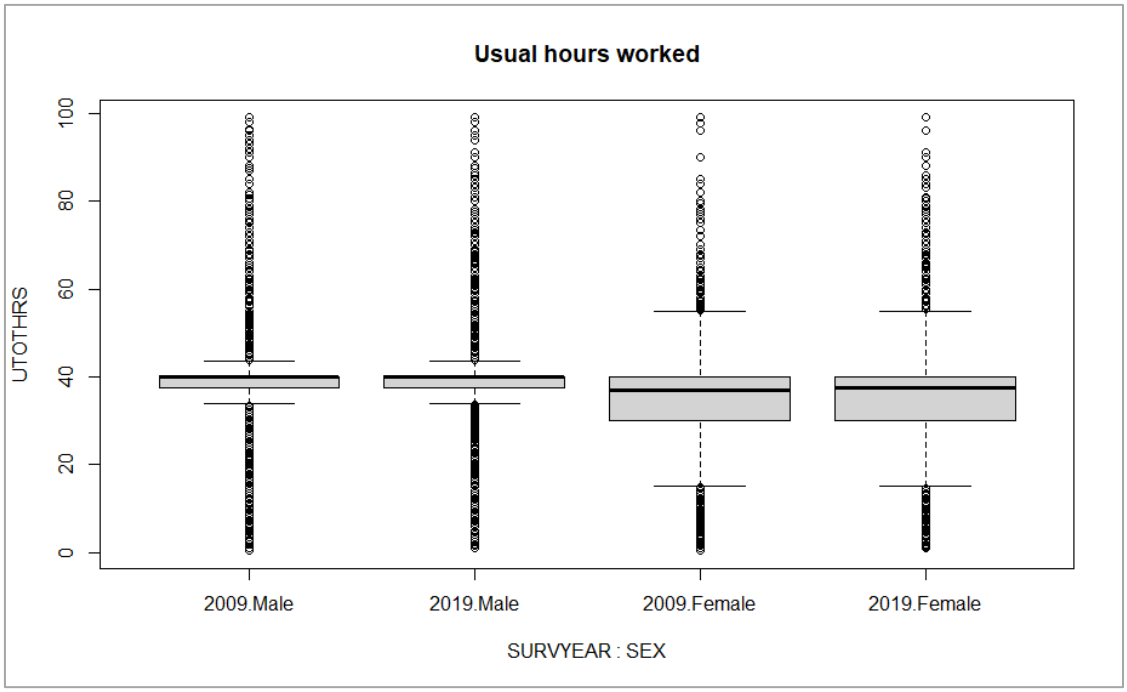
Next, we have the basic statistics of each group by gender:

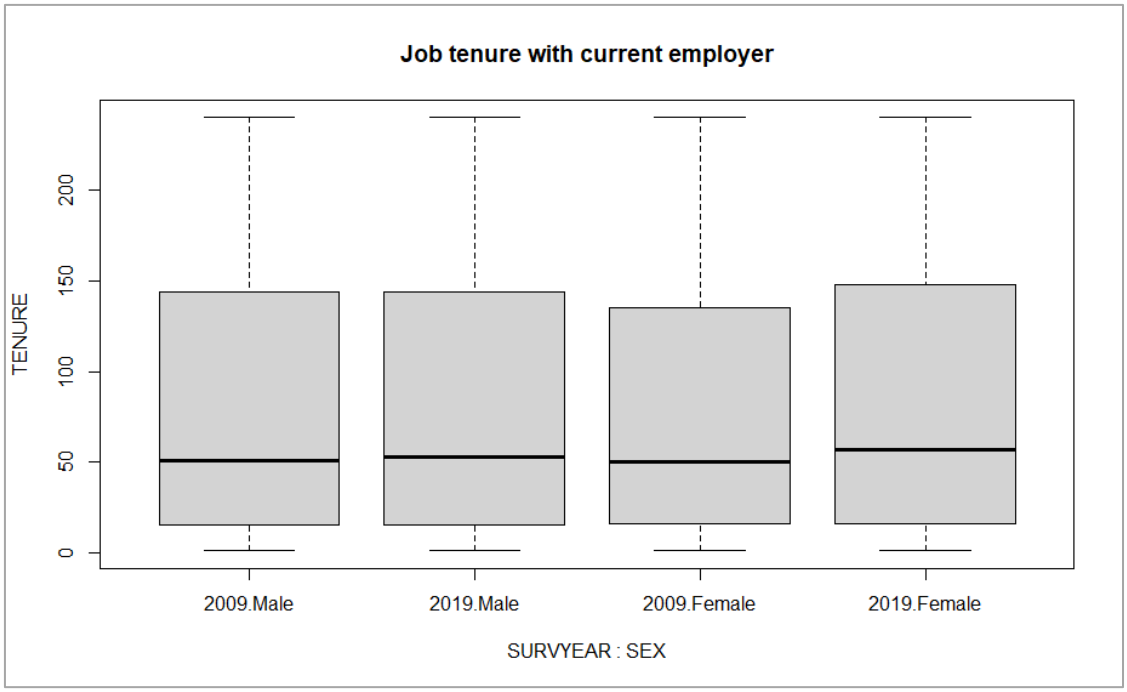
		Male		Female	
VARIABLE	STATISTICS	2009	2019	2009	2019
Usual hours worked per week	MIN	0.4	1.0	0.4	1.0
	MAX	99.0	99.0	99.0	99.0
	AVERAGE	39.0	38.8	33.3	33.9
	STD DEV	10.9	11.0	10.8	10.7
Job tenure with current employer (months)	MIN	1.0	1.0	1.0	1.0
	MAX	240.0	240.0	240.0	240.0
	AVERAGE	86.3	85.7	84.4	87.9
	STD DEV	84.7	82.5	82.2	82.7
Usual hourly wages, employees only	MIN	2.1	3.0	2.0	3.1
	MAX	115.4	108.0	89.7	106.7
	AVERAGE	23.3	29.3	19.7	25.7
	STD DEV	11.9	14.5	10.3	12.6

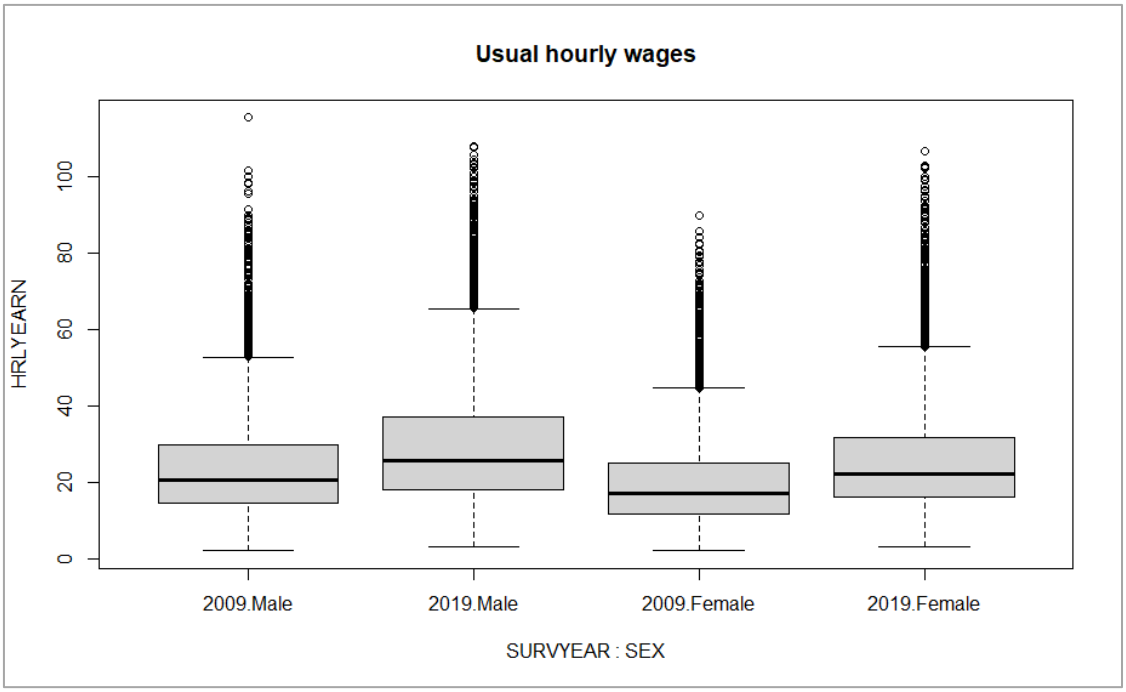
Outliers

To identify outliers in the numeric variables, boxplot analysis was applied to groups by year and gender. Next are the summary of boxplot statistics and the graphs:

		Male		Female	
VARIABLE	STATISTICS	2009	2019	2009	2019
	TOTAL CASES	**26,942**	**25,499**	**27,615**	**25,432**
Usual hours worked per week (UTOTHRS)	LOWER FENCE	33.8	33.8	15	15
	Q1	37.5	37.5	30	30
	MEDIAN	40	40	37	37.5
	Q3	40	40	40	40
	UPPER FENCE	43.5	43.7	55	55
	TOTAL OUTLIERS	8,453	8,080	2,324	2,096
	% OUTLIERS	31%	32%	8%	8%
Job tenure with current employer in months (TENURE)	LOWER FENCE	1	1	1	1
	Q1	15	15	16	16
	MEDIAN	51	53	50	57
	Q3	144	144	135	148
	UPPER FENCE	240	240	240	240
	TOTAL OUTLIERS	0	0	0	0
	% OUTLIERS	0%	0%	0%	0%
Usual hourly wages, employees only (HRLYEARN)	LOWER FENCE	2.14	3	2	3.07
	Q1	14.5	18	11.75	16
	MEDIAN	20.51	25.65	17.14	22
	Q3	29.8	37	24.91	31.79
	UPPER FENCE	52.69	65.38	44.62	55.38
	TOTAL OUTLIERS	675	618	708	681
	% OUTLIERS	3%	2%	3%	3%


Usual hours worked
UTOTHRS
0
20
40
60
80
100
2009.Male
2019.Male
2009.Female
2019.Female
SURVYEAR : SEX


Job tenure with current employer
TENURE
0
50
100
150
200
2009.Male
2019.Male
2009.Female
2019.Female
SURVYEAR : SEX



Distribution of numeric variables

To verify normality of the quantitative variables, histograms and q-q plots were built, and Kolmogorov-Smirnov tests were run. Shapiro-Wilk's method was not applied because is restricted to samples of size smaller than 5,000.

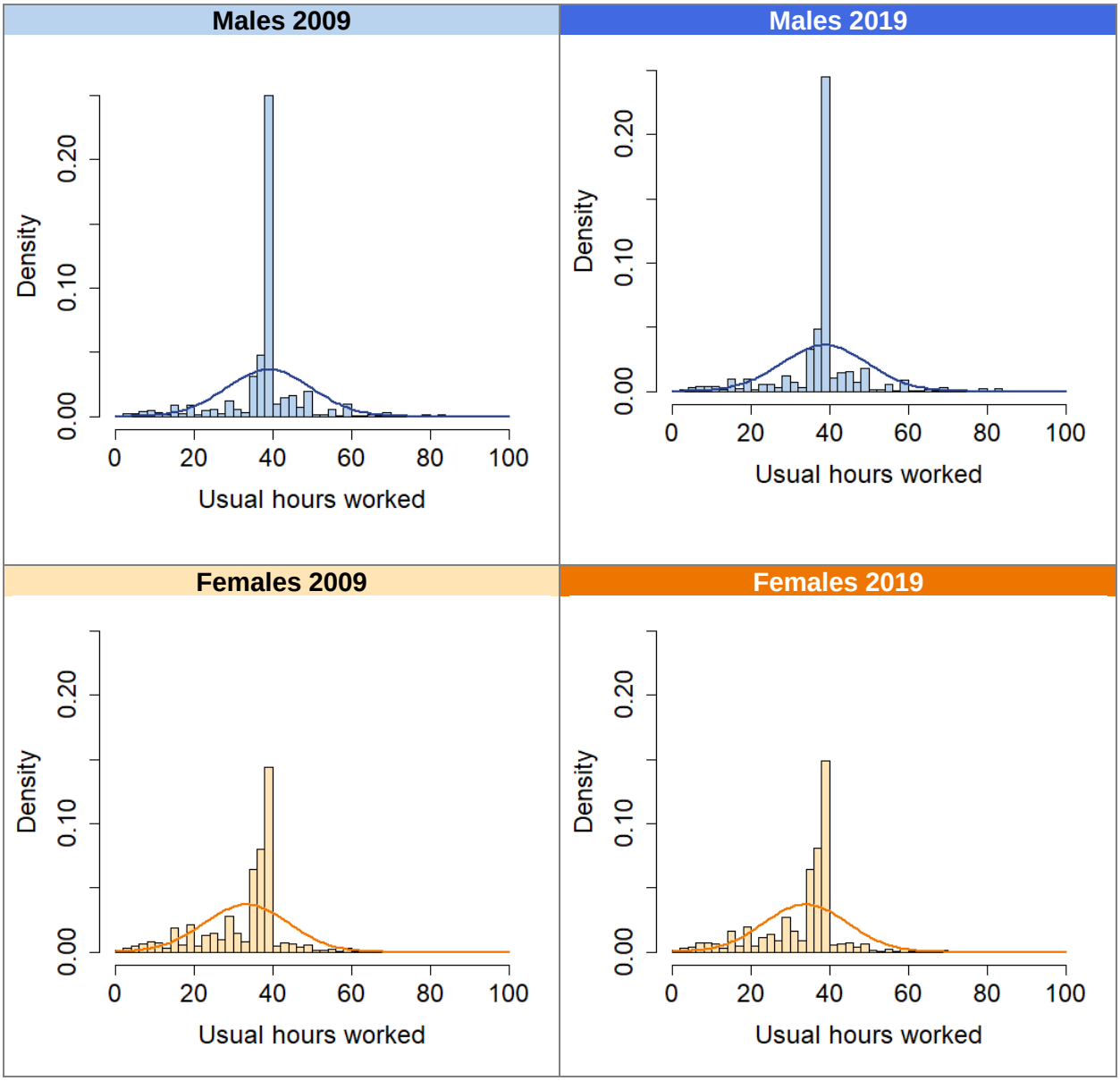
The histograms show the shape of the distribution and compare it to a curve that would follow a normal distribution with the same mean and standard deviation as the ones from the variable tested.

The q-q plot (or quantile-quantile plot) draws the correlation between a given sample and the normal distribution. A 45-degree reference line is also plotted. If both sets of quantiles come from the same distribution, we should see the points forming a line that is roughly straight. In contrast, if points form a curve instead of a straight line it usually means the sample data are skewed. Moreover, if the points fall along a line in the middle of the graph, but curve off in the extremities it is a sign that the data have more extreme values than would be expected if they truly came from a normal distribution.

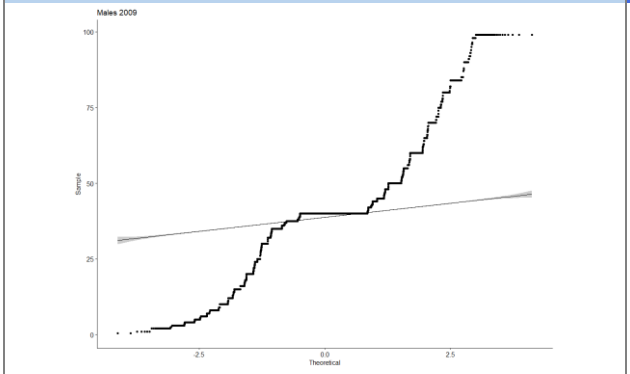
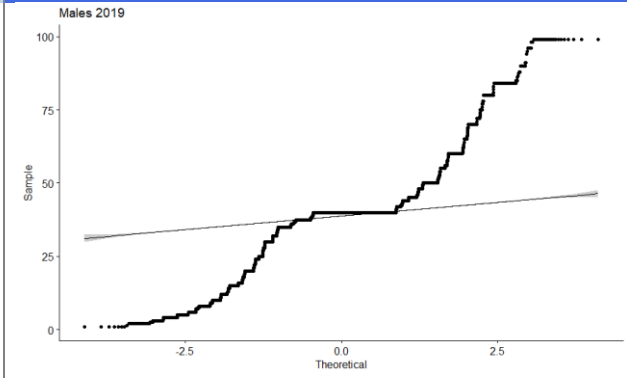
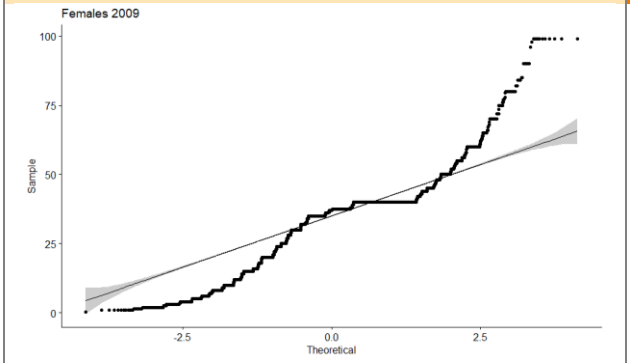
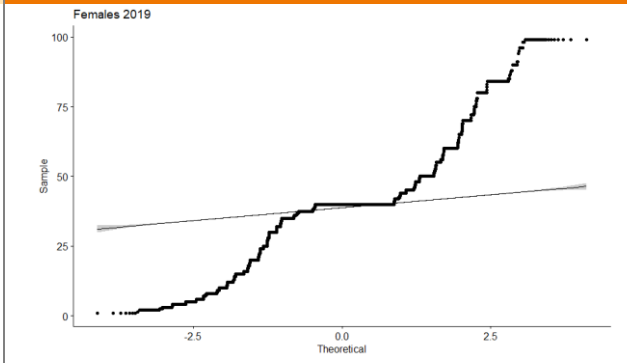
The Kolmogorov-Smirnov test compares the observed distribution with a theoretically specified distribution, in this case, the normal distribution. It is important that this distribution has identical mean and standard deviation as the sample evaluated. The null hypothesis of the K-S test is that the distribution is normal, therefore, if p-value of the test is >0.05, we do not reject the null hypothesis and conclude that the distribution in question is not statistically different from a normal distribution.

Usual hours worked per week (UTOTHRS)

Histograms



Q-Q Plots

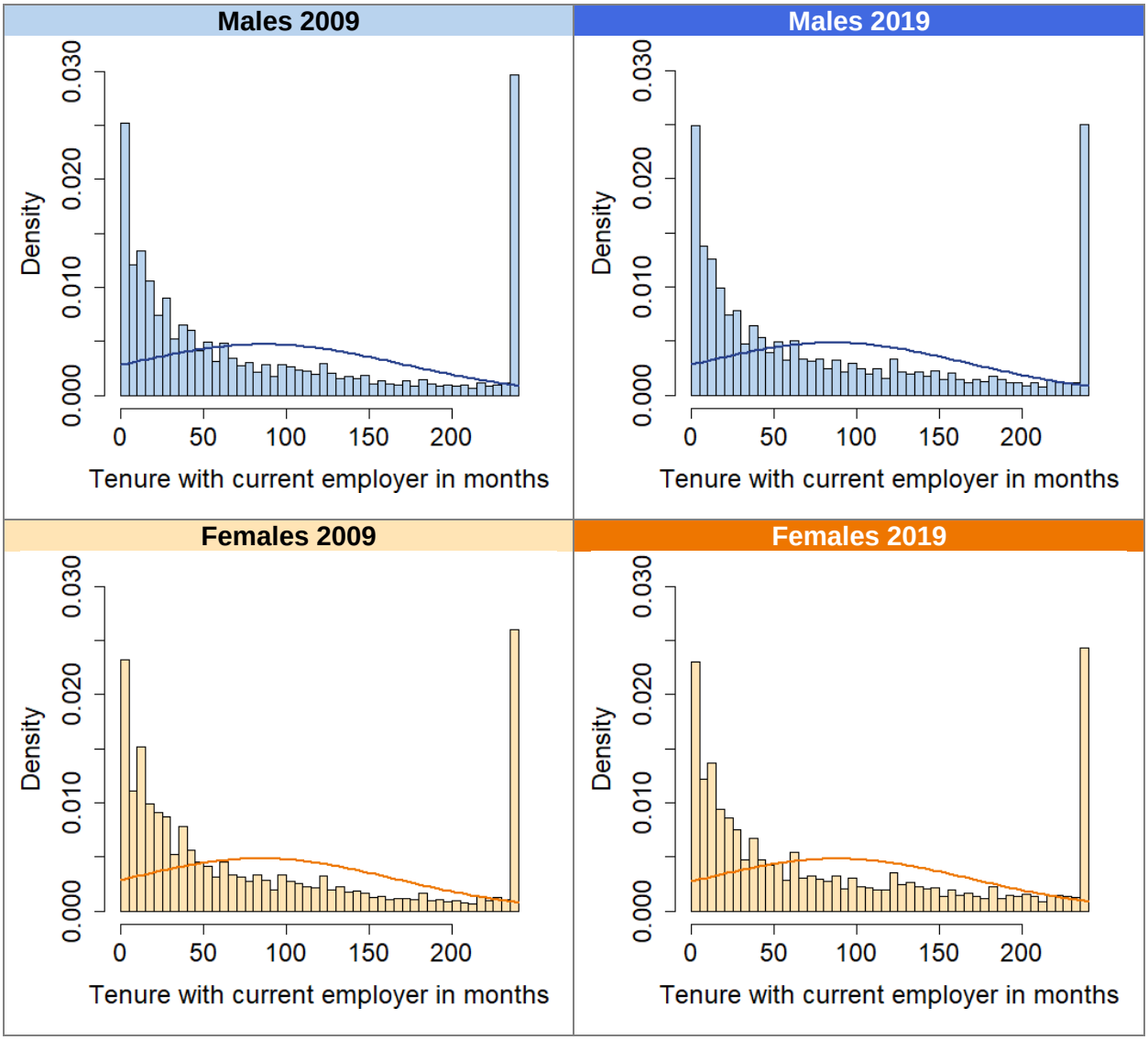
Males 2009	Males 2019
	
Females 2009	**Females 2019**
	

Kolmogorov-Smirnov Tests

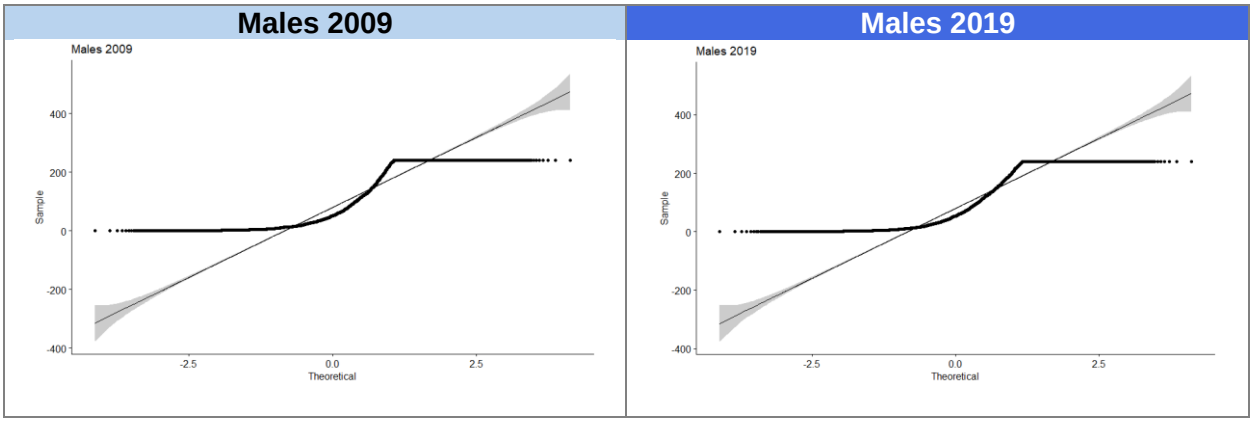
Males 2009	Males 2019
data: data.all.09.male$UTOTHRS D = 0.266, p-value < 2.2e-16 alternative hypothesis: two-sided	data: data.all.19.male$UTOTHRS D = 0.26344, p-value < 2.2e-16 alternative hypothesis: two-sided
Females 2009	**Females 2019**
data: data.all.09.fem$UTOTHRS D = 0.21438, p-value < 2.2e-16 alternative hypothesis: two-sided	data: data.all.19.fem$UTOTHRS D = 0.21143, p-value < 2.2e-16 alternative hypothesis: two-sided

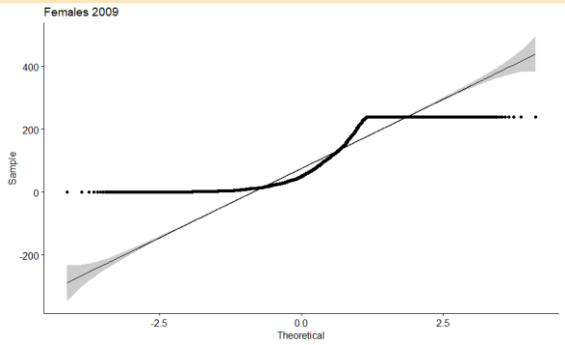
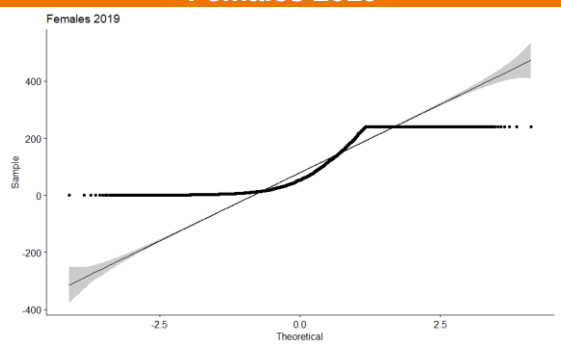
Job Tenure (TENURE)

Histograms


Males 2009
Density
0.000
0.010
0.020
0.030
0
50
100
150
200
Tenure with current employer in months
Males 2019
Density
0.000
0.010
0.020
0.030
0
50
100
150
200
Tenure with current employer in months
Females 2009
Density
0.000
0.010
0.020
0.030
0
50
100
150
200
Tenure with current employer in months
Females 2019
Density
0.000
0.010
0.020
0.030
0
50
100
150
200
Tenure with current employer in months

Q-Q Plots


Males 2009
Males 2009
Sample
400
200
0
-200
-400
-2.5
0.0
2.5
Theoretical
Males 2019
Males 2019
Sample
400
200
0
-200
-400
-2.5
0.0
2.5
Theoretical

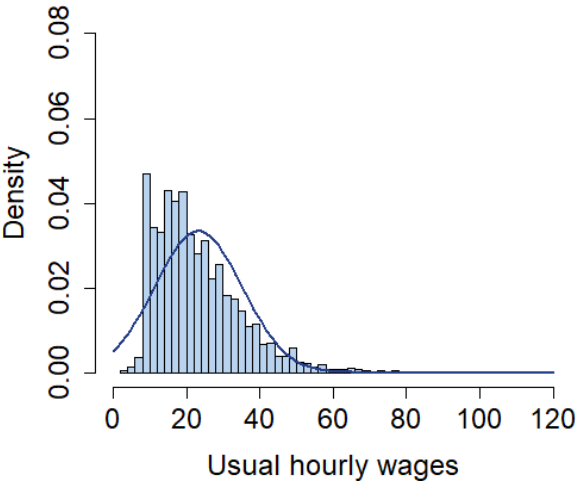
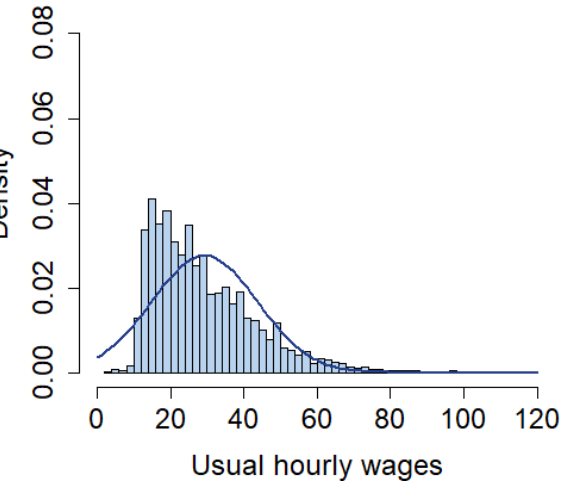
Females 2009	Females 2019
	

Kolmogorov-Smirnov Tests

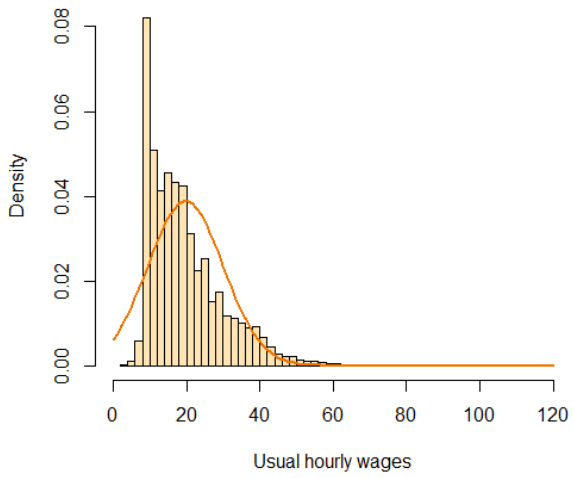
Males 2009	Males 2019
data: data.all.09.male$TENURE D = 0.16707, p-value < 2.2e-16 alternative hypothesis: two-sided	data: data.all.19.male$TENURE D = 0.15484, p-value < 2.2e-16 alternative hypothesis: two-sided
Females 2009	**Females 2019**
data: data.all.09.fem$TENURE D = 0.16497, p-value < 2.2e-16 alternative hypothesis: two-sided	data: data.all.19.fem$TENURE D = 0.1523, p-value < 2.2e-16 alternative hypothesis: two-sided

Usual Hourly Wages (HRLYEARN)

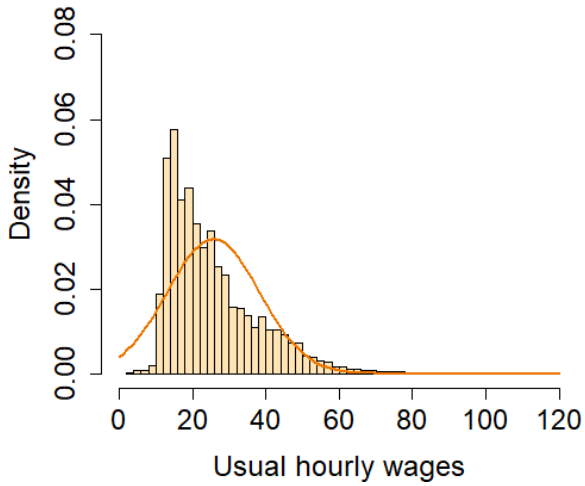
Histograms

Males 2009	Males 2019
	

Females 2009

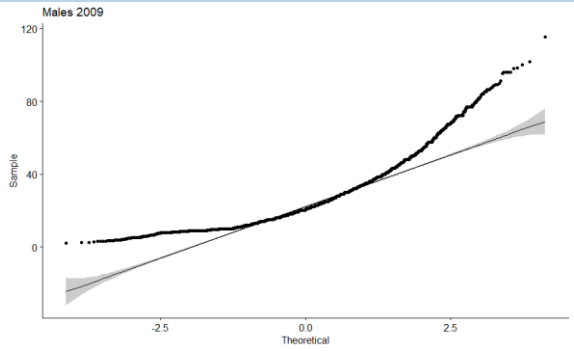


Females 2019

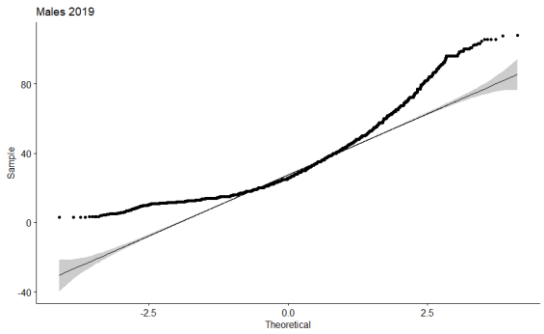


Q-Q Plots

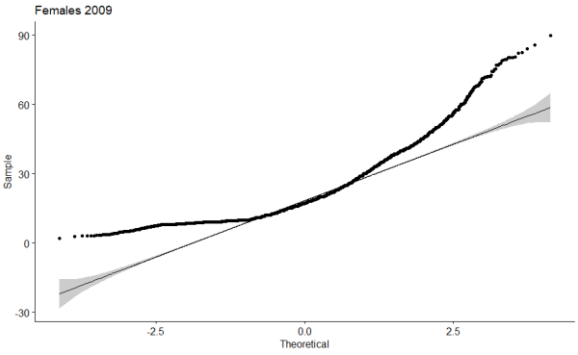
Males 2009



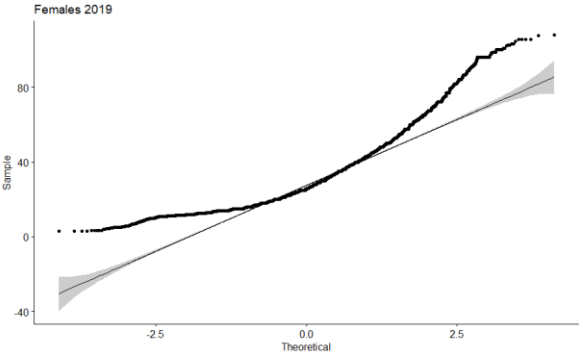
Males 2019



Females 2009



Females 2019



Kolmogorov-Smirnov Tests

Males 2009	Males 2019
data: data.all.09.male$HRLYEARN D = 0.098075, p-value < 2.2e-16 alternative hypothesis: two-sided	data: data.all.19.male$HRLYEARN D = 0.10788, p-value < 2.2e-16 alternative hypothesis: two-sided
Females 2009	**Females 2019**
data: data.all.09.fem$HRLYEARN D = 0.11895, p-value < 2.2e-16 alternative hypothesis: two-sided	data: data.all.19.fem$HRLYEARN D = 0.12557, p-value < 2.2e-16 alternative hypothesis: two-sided

Looking at the graphs and the results from Kolmogorov-Smirnov, we can conclude that none of the numeric variables are normally distributed.

Parametric tests usually required data to be normally distributed, though there are cases where they still can be applied when the sample size is large enough. Anyway, when parametric assumptions are violated non-parametric tests can be applied instead.

On the other hand, linear regression remains a statistically sound technique in studies of large sample sizes even when a dependent variable is not distributed normally. By the law of large numbers and the central limit theorem, the ordinary least squares (OLS) estimators in linear regression technique still will be approximately normally distributed around the true parameter values, which implies the estimated parameters and their confidence interval estimates remain robust. Hence, in a large sample, the use of a linear regression technique, even if the dependent variable violates the "normality assumption" rule, remains valid.

Correlation

Pearson correlation measures linear relationships between two numerical variables:

Males 2009

	UTOTHRS	TENURE	HRLYEARN
UTOTHRS	1.00	0.11	0.14
TENURE	0.11	1.00	0.30
HRLYEARN	0.14	0.30	1.00

Males 2019

	UTOTHRS	TENURE	HRLYEARN
UTOTHRS	1.00	0.12	0.17
TENURE	0.12	1.00	0.30
HRLYEARN	0.17	0.30	1.00

Females 2009

	UTOTHRS	TENURE	HRLYEARN
UTOTHRS	1.00	0.18	0.20
TENURE	0.18	1.00	0.35
HRLYEARN	0.20	0.35	1.00

Females 2019

	UTOTHRS	TENURE	HRLYEARN
UTOTHRS	1.00	0.14	0.19
TENURE	0.14	1.00	0.34
HRLYEARN	0.19	0.34	1.00

Overall, we can see that there is no evidence of relationship between the variables, except for a moderate positive correlation between tenure and hourly wages, that is a little more accentuated for women than men.

Besides, Spearman test allows to include ordinal variable to analyse the correlation:

Males 2009

	UTOTHRS	TENURE	HRLYEARN	AGE_12	EDUC	ESTSIZE	FIRMSIZE	AGYOWNK
UTOTHRS	1.00	0.10	0.14	0.13	-0.01	-0.04	-0.10	-0.13
TENURE	0.10	1.00	0.37	0.47	0.09	0.20	0.19	-0.14
HRLYEARN	0.14	0.37	1.00	0.31	0.37	0.27	0.24	-0.24
AGE_12	0.13	0.47	0.31	1.00	0.10	0.09	0.06	-0.04
EDUC	-0.01	0.09	0.37	0.10	1.00	0.16	0.15	-0.16
ESTSIZE	-0.04	0.20	0.27	0.09	0.16	1.00	0.60	-0.07
FIRMSIZE	-0.10	0.19	0.24	0.06	0.15	0.60	1.00	-0.05
AGYOWNK	-0.13	-0.14	-0.24	-0.04	-0.16	-0.07	-0.05	1.00

Males 2019

	UTOTHRS	TENURE	HRLYEARN	AGE_12	EDUC	ESTSIZE	FIRMSIZE	AGYOWNK
UTOTHRS	1.00	0.11	0.18	0.11	-0.02	0.02	-0.06	-0.12
TENURE	0.11	1.00	0.36	0.47	0.10	0.17	0.18	-0.12
HRLYEARN	0.18	0.36	1.00	0.26	0.39	0.28	0.25	-0.26
AGE_12	0.11	0.47	0.26	1.00	0.08	0.07	0.04	-0.02
EDUC	-0.02	0.10	0.39	0.08	1.00	0.18	0.18	-0.17
ESTSIZE	0.02	0.17	0.28	0.07	0.18	1.00	0.61	-0.09
FIRMSIZE	-0.06	0.18	0.25	0.04	0.18	0.61	1.00	-0.07
AGYOWNK	-0.12	-0.12	-0.26	-0.02	-0.17	-0.09	-0.07	1.00

Females 2009

	UTOTHRS	TENURE	HRLYEARN	AGE_12	EDUC	ESTSIZE	FIRMSIZE	AGYOWNK
UTOTHRS	1.00	0.16	0.22	0.12	0.14	0.10	0.04	-0.05
TENURE	0.16	1.00	0.42	0.51	0.11	0.20	0.16	-0.08
HRLYEARN	0.22	0.42	1.00	0.27	0.48	0.34	0.26	-0.17
AGE_12	0.12	0.51	0.27	1.00	0.05	0.09	0.04	0.05
EDUC	0.14	0.11	0.48	0.05	1.00	0.16	0.13	-0.15
ESTSIZE	0.10	0.20	0.34	0.09	0.16	1.00	0.55	-0.05
FIRMSIZE	0.04	0.16	0.26	0.04	0.13	0.55	1.00	-0.04
AGYOWNK	-0.05	-0.08	-0.17	0.05	-0.15	-0.05	-0.04	1.00

Females 2019

	UTOTHRS	TENURE	HRLYEARN	AGE_12	EDUC	ESTSIZE	FIRMSIZE	AGYOWNK
UTOTHRS	1.00	0.13	0.21	0.09	0.14	0.11	0.04	-0.05
TENURE	0.13	1.00	0.40	0.50	0.10	0.19	0.18	-0.07
HRLYEARN	0.21	0.40	1.00	0.20	0.48	0.33	0.27	-0.20
AGE_12	0.09	0.50	0.20	1.00	0.01	0.05	0.01	0.09
EDUC	0.14	0.10	0.48	0.01	1.00	0.18	0.14	-0.19
ESTSIZE	0.11	0.19	0.33	0.05	0.18	1.00	0.58	-0.06
FIRMSIZE	0.04	0.18	0.27	0.01	0.14	0.58	1.00	-0.05
AGYOWNK	-0.05	-0.07	-0.20	0.09	-0.19	-0.06	-0.05	1.00

There are moderate positive correlations between Hourly Wages, Education and Tenure. This means that as the value of education and tenure increase, hourly wage tend to increase as well.

There is also a moderate positive correlation between Tenure and Age, meaning that as people grow old, they accumulate more years of experience working for their employer.

Besides, there is a strong positive correlation between establishment and firm size, which is to be expected since large firms usually will have offices or other workplaces locations with many employees.

Hourly Wages Relationship to Other Variables

Numeric variables

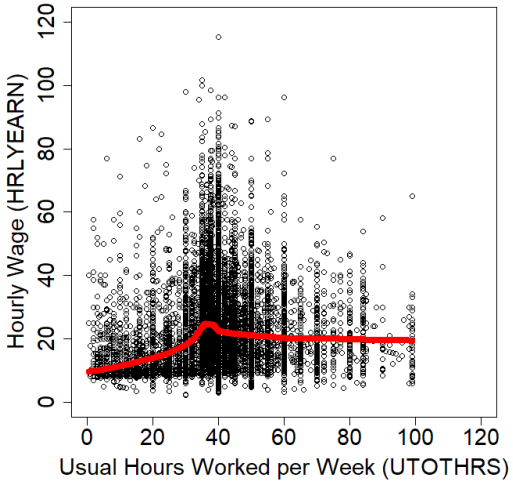
Scatterplots were built to explore further the relationship between hourly wages, tenure and usual number of hours worked weekly.

LOWESS (Locally Weighted Scatterplot Smoothing), was used to create a smooth line through the scatter plot to help distinguish the relationship between variables. This a non-parametric smoother to find a curve of best fit without assuming the data must fit some distribution shape.

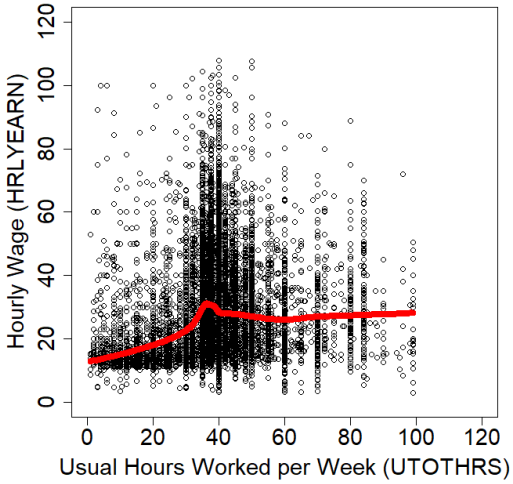
Usual Weekly Hours Worked

As we can see on the next graphs, hourly wages tend to increase as the hours worked per week do, but only while the hours worked are approximately less than 40. After the 40-hour mark, hourly wage would likely stabilize for men, and even decrease slightly for women.

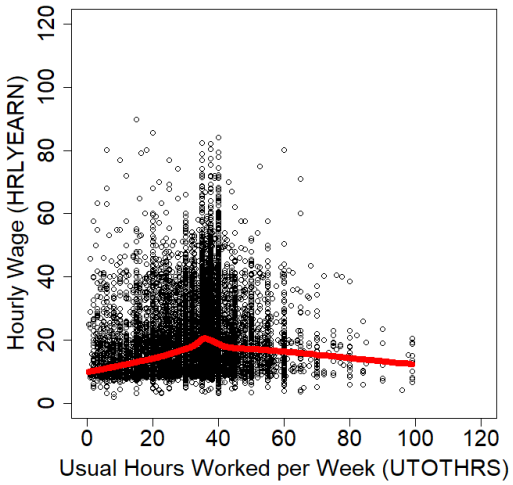
Males 2009



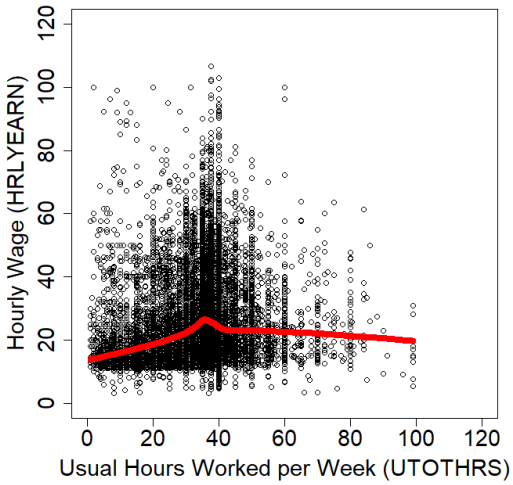
Males 2019



Females 2009



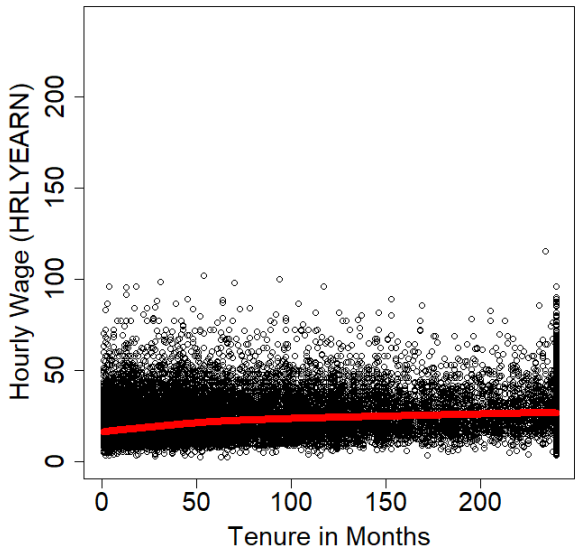
Females 2019



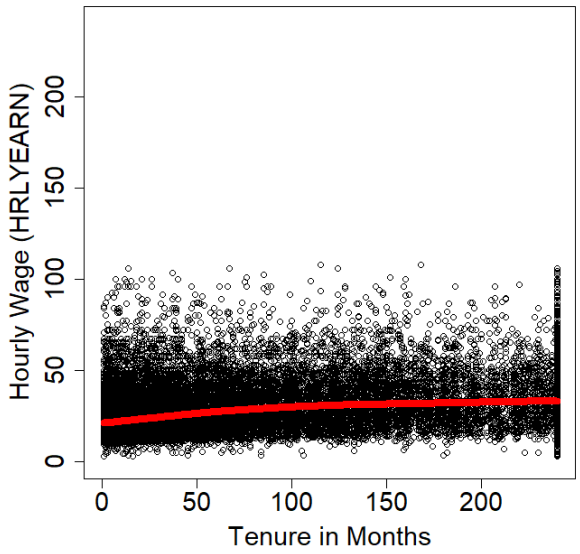
Tenure

Hourly wages increase just slightly as tenure does. This specially noted at the beginning, when the tenure is around less than 50 months (or 2 years), and it is seen in both males and females, in 2009 and 2019.

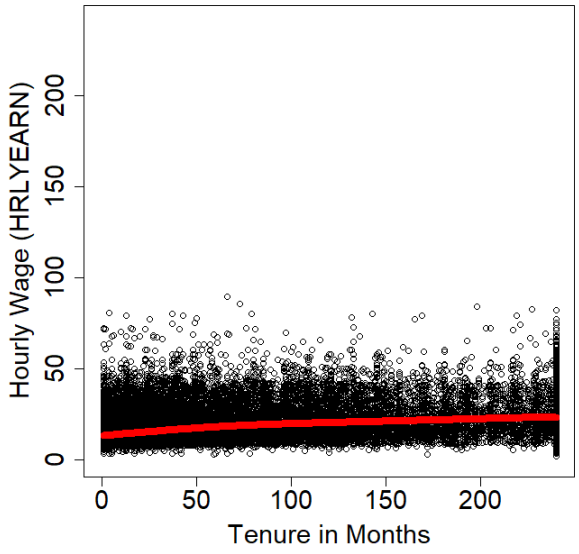
Males 2009



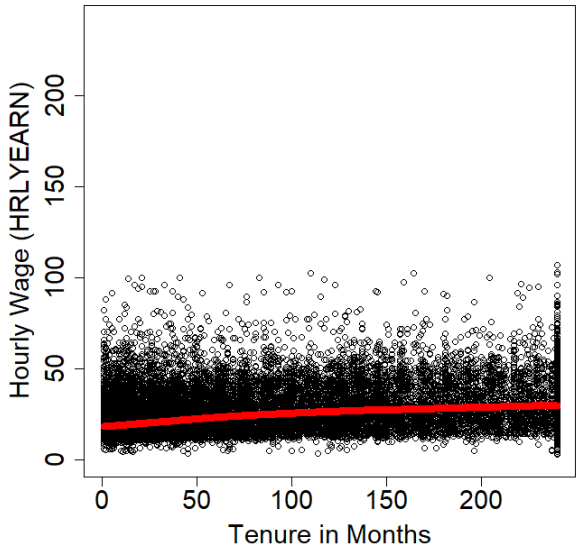
Males 2019



Females 2009



Females 2019



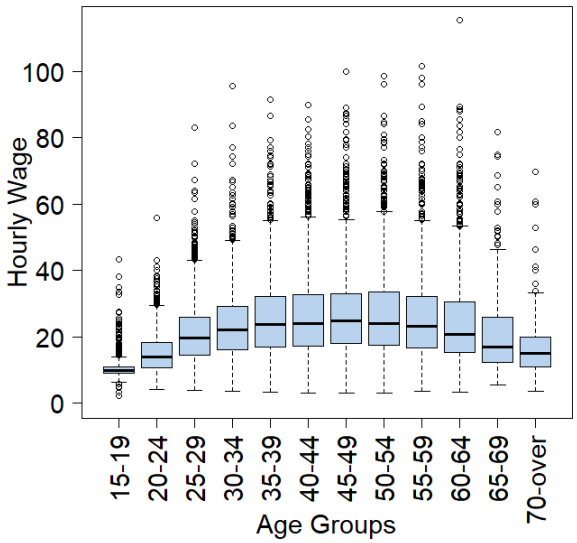
Ordinal Variables

Boxplot were built to get better understanding of the relationship between hourly wages and ordinal variables.

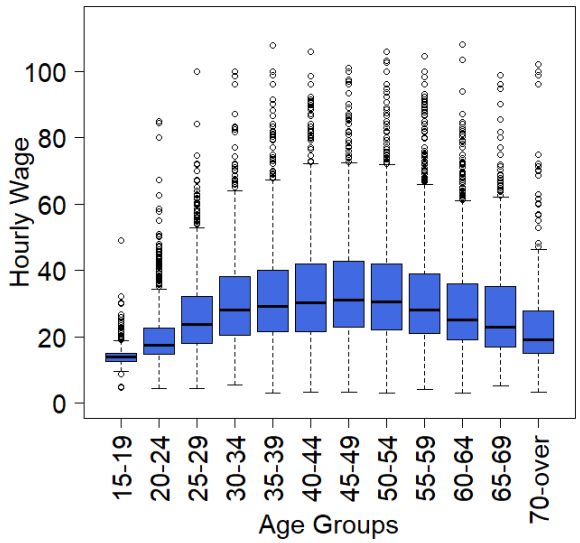
Age

There seems to be a quadratic relationship between age and hourly wage, and it is apparently more pronounced for men than women. Hourly wages increase as age does, then it tends to peak in the groups of people between 40 and 50 years old and decreases afterwards.

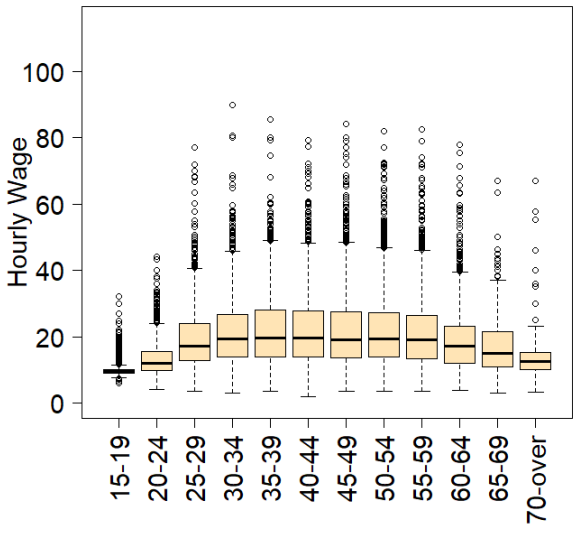
Males 2009



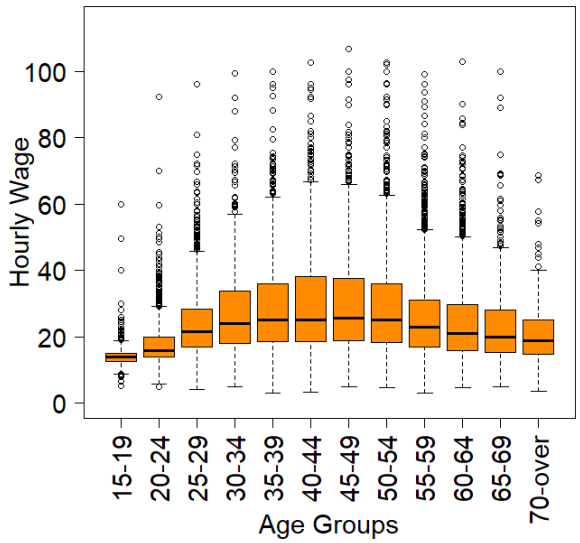
Males 2019



Females 2009



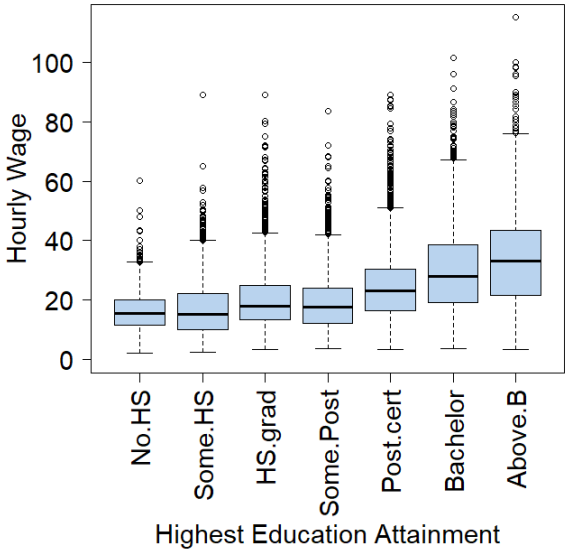
Females 2019



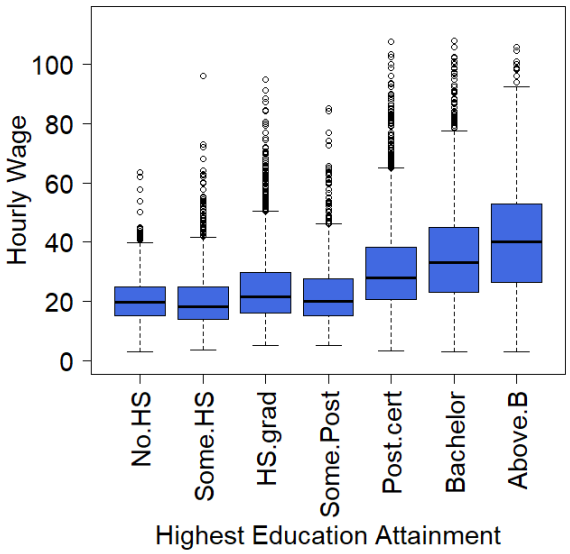
Education

Hourly wages tend to increase with more years of education. This trend is accentuated for the groups with highest education, “Postsecondary certificate or diploma”, “Bachelor´s degree” and “Above bachelor´s degree”.

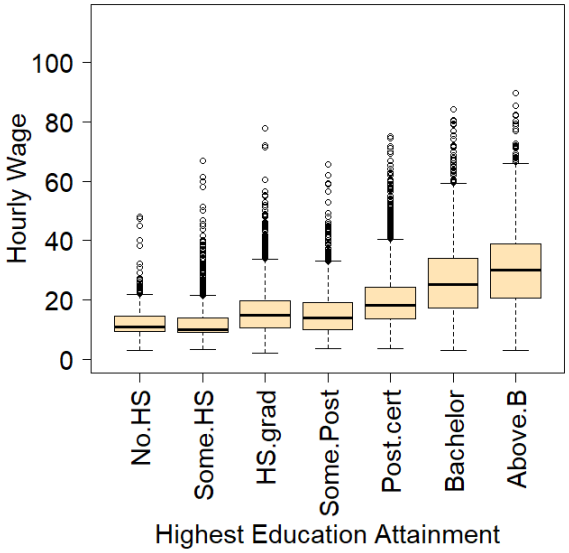
Males 2009



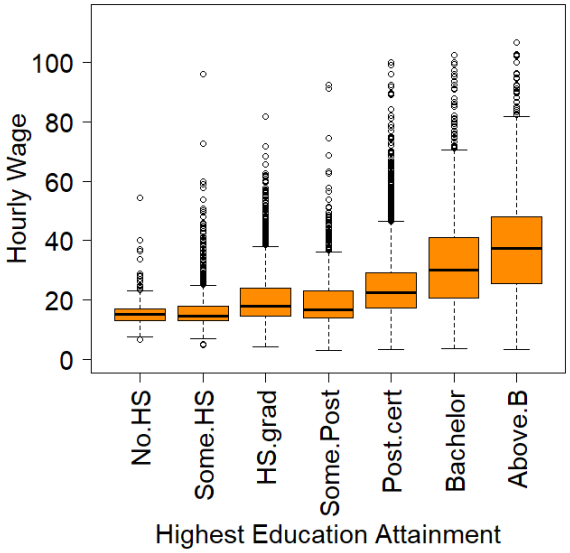
Males 2019



Females 2009

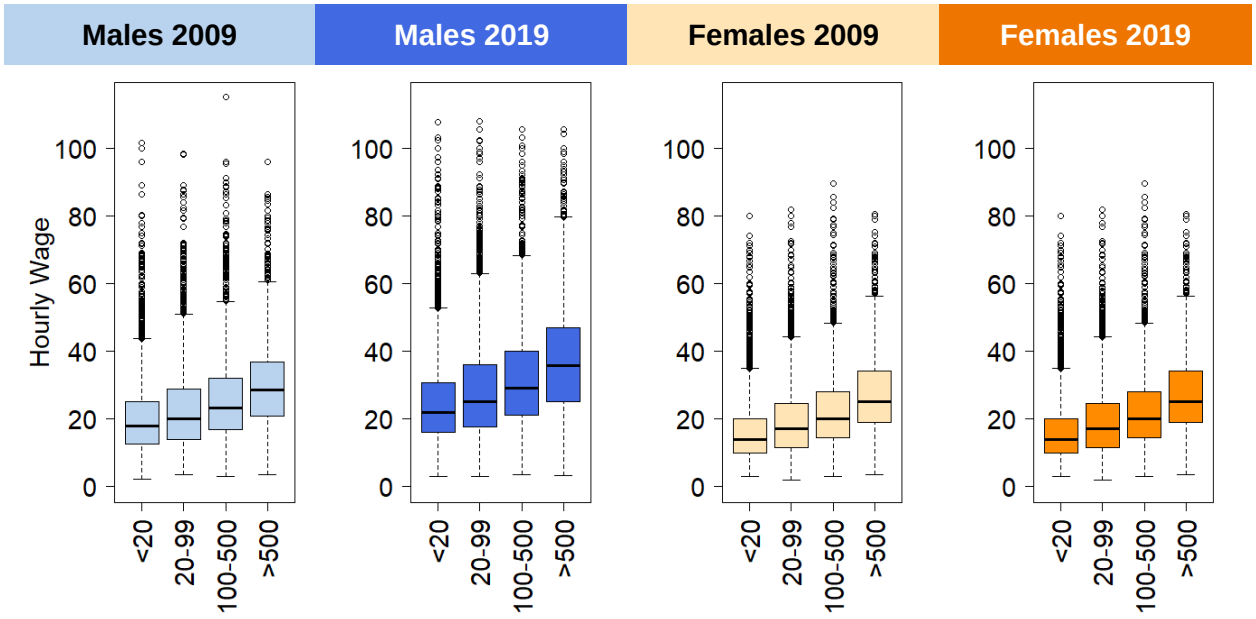


Females 2019



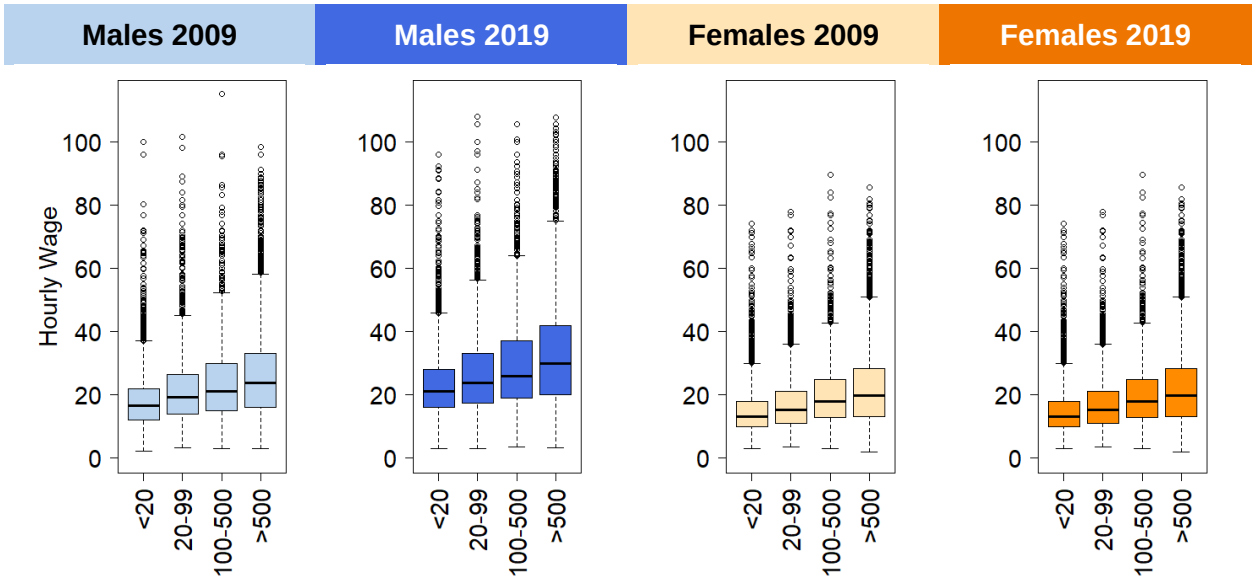
Establishment Size (Number of Employees)

Across all groups of year and gender the plots suggest that employees that work at locations with many coworkers might get higher hourly wages than the ones who work at smaller sites.



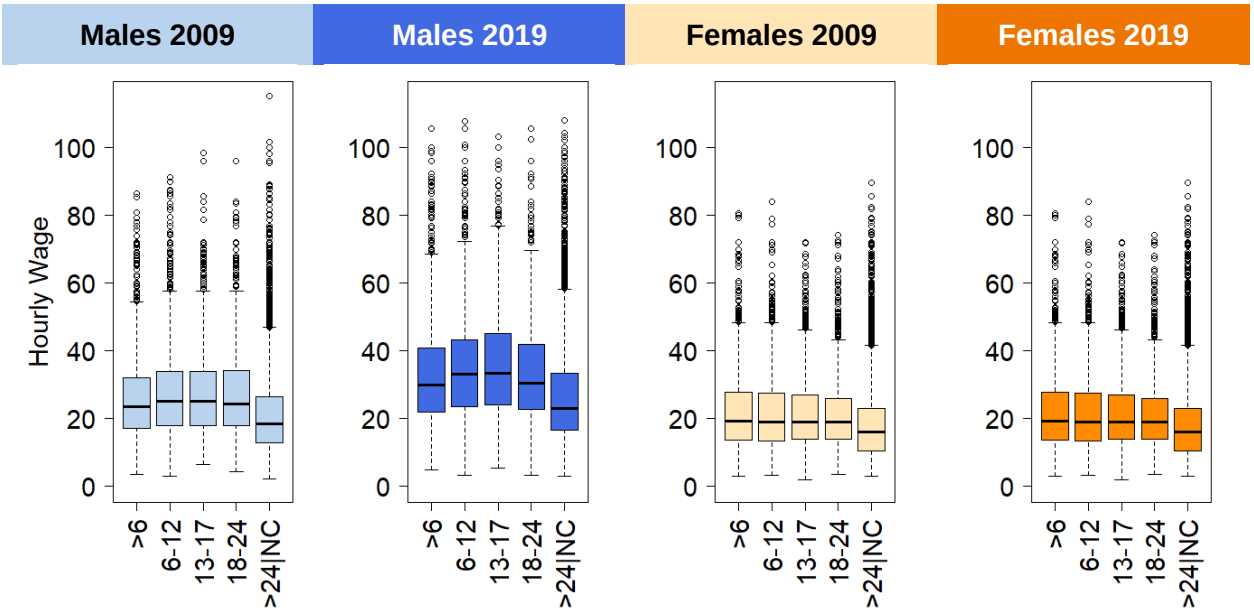
Firm Size (Number of Employees)

Employees of largest companies in terms of total number of employees tend to get higher hourly wages than the ones working at smaller firms.



Age of youngest child

For men, hourly wages by age of the youngest child in the household seem to follow a similar pattern as age groups. There seem to be a peak when the children are teenagers, and a decrement afterwards. However, women seem to have a stable hourly wage throughout the groups, until there is a decrease for the last group, when the youngest child is older than 24 or there are no children at all.



Nominal Variables

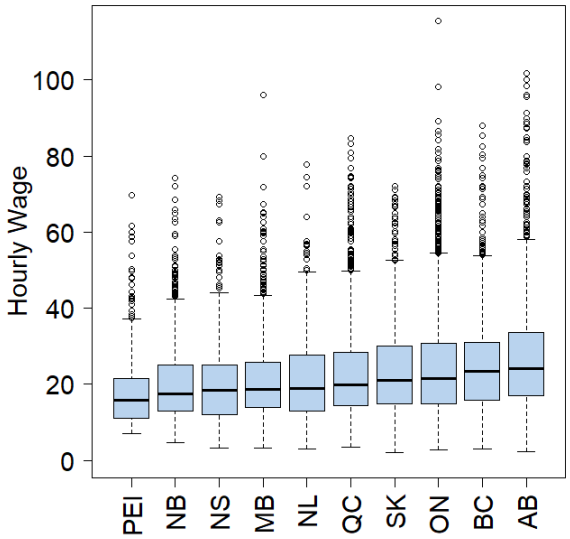
Next, an analysis for four nominal variables was done to explore if there is a possible relationship between them and hourly wages. The variables selected for this analysis were “Province”, “Marital status”, “Full/Part Time”, and “Job Permanency”. Note that all plots horizontal axis is ordered according to ascending median values of the categories.

Later, in Step 3 part of this document, there is a detailed analysis of the distribution of hourly wages by sector, industry, and occupations. Therefore, those nominal variables are not included in this section.

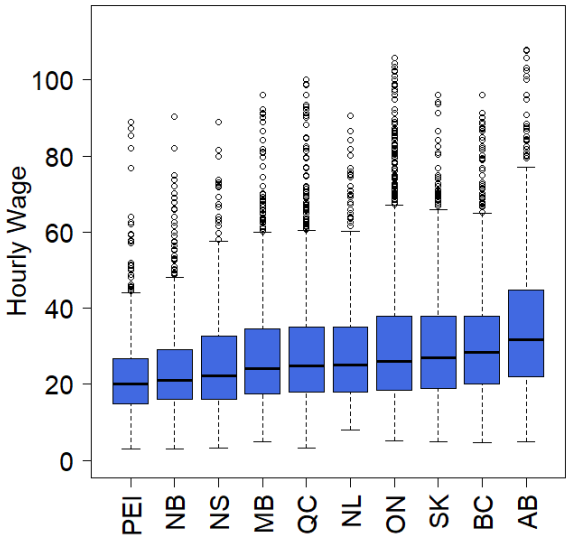
Province

From the plots and the medians comparison below it, we can say that the hourly wages vary across the provinces. For all the groups analyzed, the provinces that tend to have higher hourly wages are Alberta, British Columbia, Ontario, and Saskatchewan.

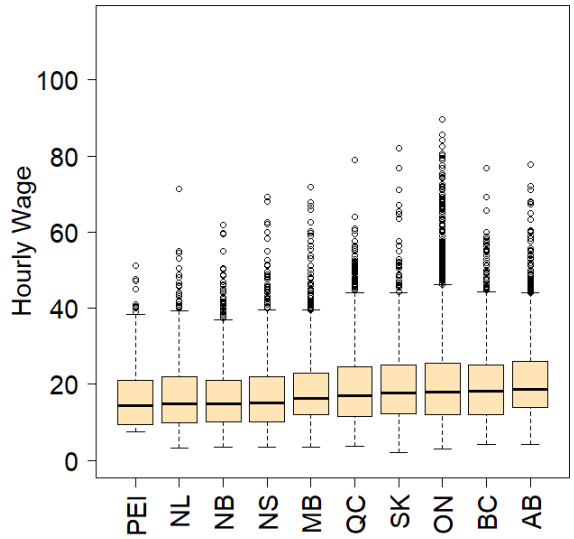
Males 2009



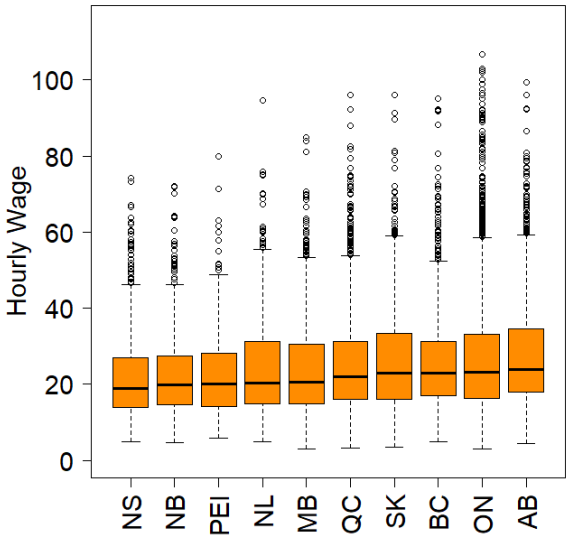
Males 2019



Females 2009



Females 2019



Medians per group

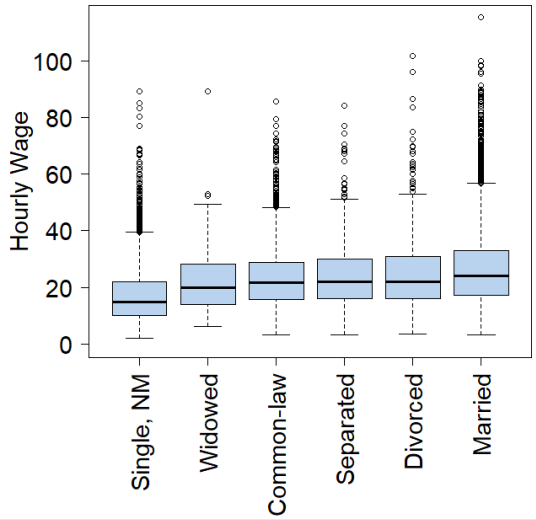
Top 3 provinces by median for each group are highlighted:

	NL	PEI	NS	NB	QC	ON	MB	SK	AB	BC
Males 2009	18.95	15.80	18.48	17.50	19.98	21.54	18.75	21.00	24.23	23.44
Males 2019	25.00	20.00	22.12	21.00	24.90	26.00	24.04	27.00	31.77	28.50
Females 2009	15.00	14.42	15.21	15.00	17.00	18.00	16.31	17.62	18.75	18.20
Females 2019	20.34	20.00	19.00	19.98	22.00	23.08	20.60	23.00	24.00	23.00

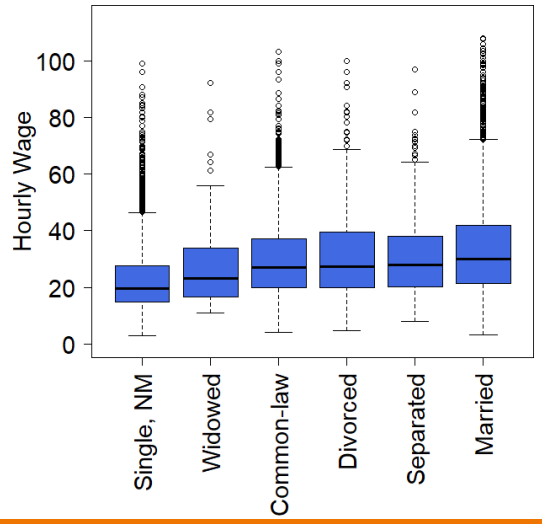
Marital Status

Married people tend to have higher wages compared to the other groups by marital status, whereas singles are the ones with lowest medians and their distributions by both year and gender are located at a slightly lower level compared to the rest.

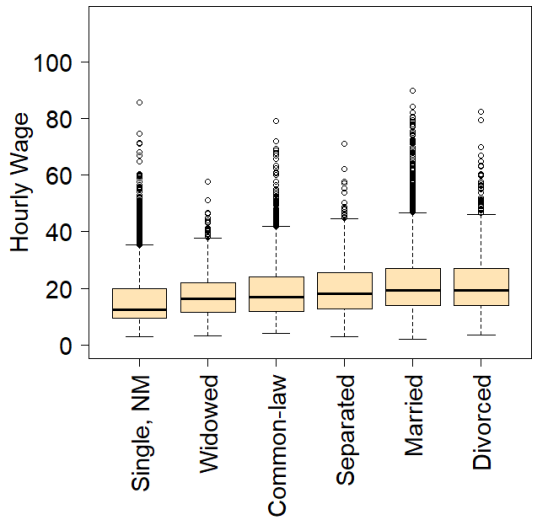
Males 2009



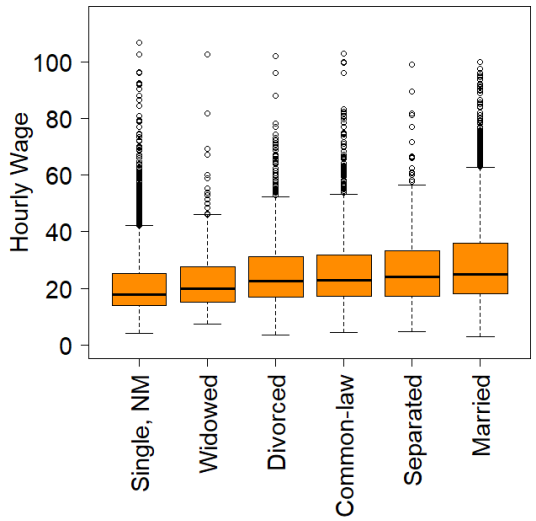
Males 2019



Females 2009



Females 2019



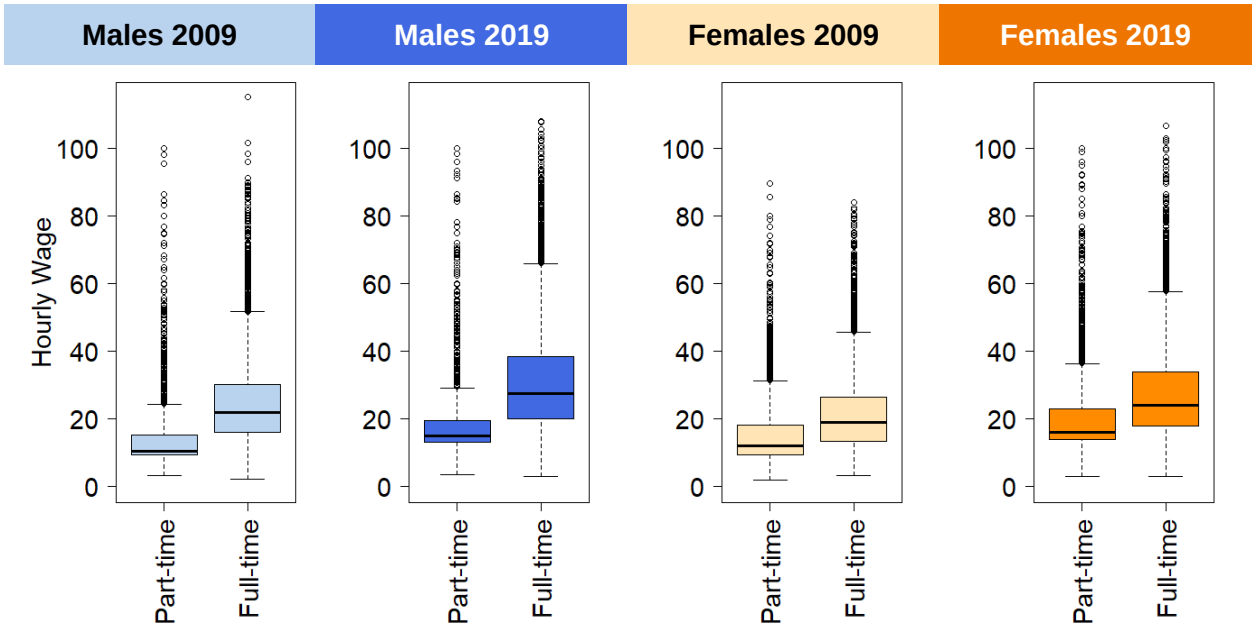
Medians per group

The highest medians for each group are highlighted:

	Married	Common-law	Widowed	Separated	Divorced	Single, NM
Males 2009	24.00	21.63	20.00	22.00	22.00	15.00
Males 2019	30.00	27.20	23.17	28.00	27.50	19.69
Females 2009	19.23	17.00	16.45	18.00	19.23	12.51
Females 2019	24.92	23.00	20.00	24.00	22.67	17.95

Full time or Part Time Status

Full time employees usually earn more per hour than part timers, as we can see from the boxplots and the median comparisons:



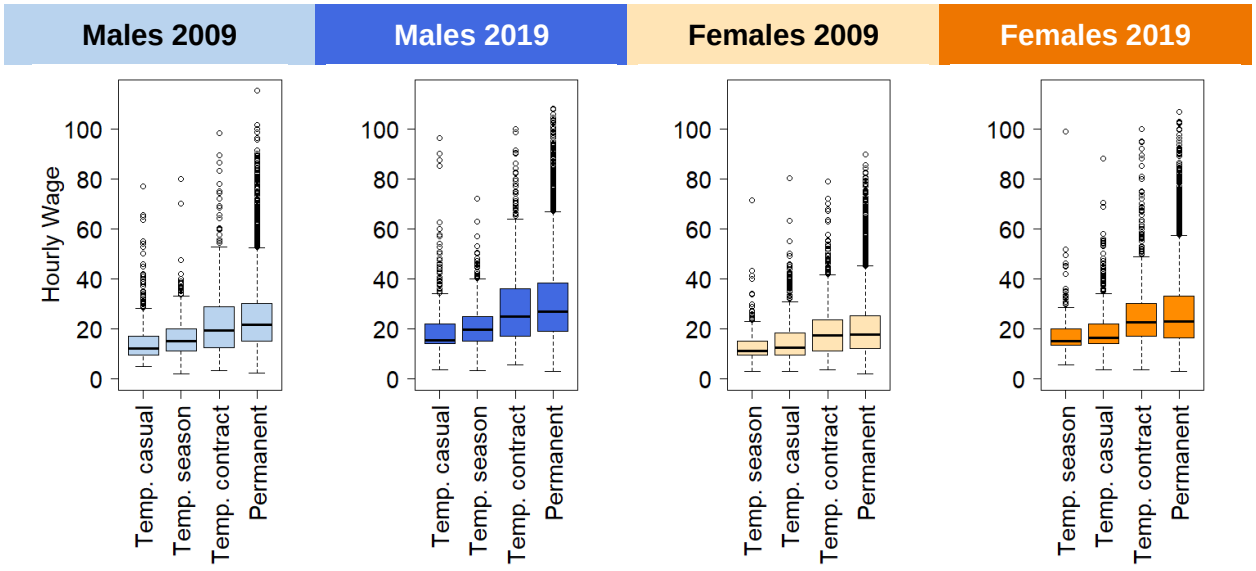
Medians per group

The highest medians for each group are highlighted:

	Full-time	Part-time
Males 2009	22.0	10.5
Males 2019	27.46	15.00
Females 2009	19	12
Females 2019	24.06	16.00

Job Permanency

Permanent and temporary terms or contracts are the categories that seem to be related to highest hourly wages, for men and women alike.



Medians per group

The highest medians for each group are highlighted:

	Permanent	Temp. season	Temp. contract	Temp. casual
Males 2009	21.63	15.00	19.23	12.00
Males 2019	26.92	19.50	24.93	15.54
Females 2009	17.85	11.00	17.21	12.50
Females 2019	23.00	15.00	22.44	16.25

Step 3: Hourly Wages by Sector, Industry and Occupation

To answer the second research question that inquiries about hourly wages evolution by gender and deep dive into the differences among sectors, industries and job occupations, a comparison between the distributions was performed. Given that the hourly wages variable is not normal, two non-parametric tests were applied, Mann-Whitney U test and Kruskal Wallis test.

Mann-Whitney U Test

Mann-Whitney U test is a non-parametric method appropriate for examining the difference in medians for 2 independent populations. It examines the relationship between a numeric outcome variable and a categorical explanatory variable (with 2 levels). The following assumptions must be met:

1) One dependent variable that is measured at the continuous or ordinal level.

2) One independent variable that consists of two categorical, independent groups (i.e., a dichotomous variable).

3) Observations are independent, which means that there is no relationship between the observations in each group of the independent variable or between the groups themselves.

The null and alternative hypotheses are:

- H0: the distribution of the dependent variable for the two groups are equal
- HA: the distribution of the dependent variable for the two groups are not equal

Another way to express the alternative hypothesis is as follows:

- HA: the mean ranks of the two groups are not equal

Kruskal Wallis Test

The Kruskal-Wallis is a non-parametric method for comparing 2 or more independent samples. It is roughly equivalent to a parametric one-way ANOVA with the data replaced by their ranks. It follows similar assumptions as Mann-Whitney U test:

1) One dependant variable at the continuous or ordinal scale

2) One independent variable with two or more levels (independent groups). The test is more commonly used when you have three or more levels.

3) Observations should be independent. In other words, there should be no relationship between the members in each group or between groups.

4) Provided the original observations are identically distributed this can be interpreted as testing for a difference between medians. But when observations represent quite different distributions Kruskal-Wallis is a test of dominance, much as the Wilcoxon-Mann-Whitney test is a test of dominance comparing just two samples. The test statistic is in fact identical to the Wilcoxon-Mann-Whitney statistic in the two-sample case.

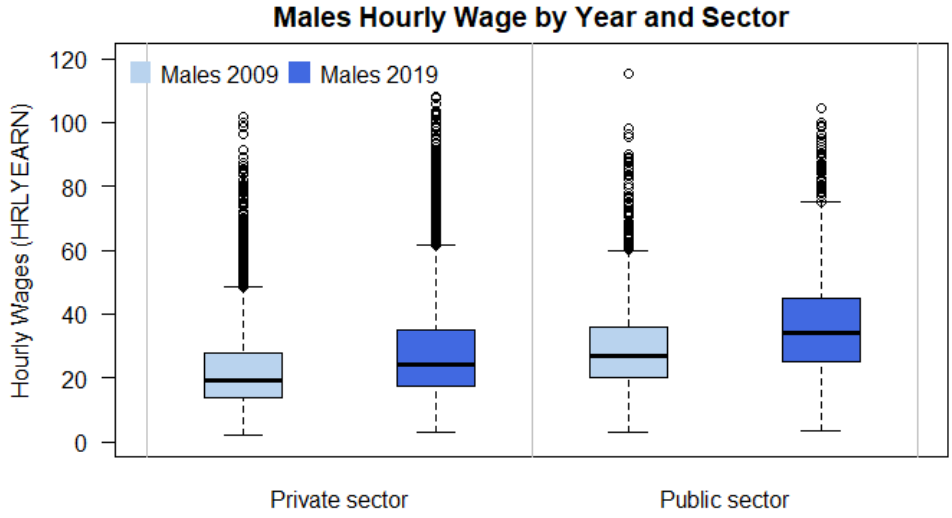
The null and alternative hypotheses are:

- H0: the distributions of the dependent variable in the groups are equal. It can also be said that each group is equally likely to obtain values above and below the common mean rank.
- HA: the distributions of the dependent variable in the groups are not equal. It can also be said that two or more groups differ in their mean rank.

Sector Analysis

There are two sectors, public and private. Next, the differences in the hourly wages over time for men and women are explored. Also, hourly wages of 2019 are compared by gender. To do these comparisons, we use the boxplot as well as the results of the Mann-Whitney U test.

Males



	2009.Private sector	2019.Private sector	2009.Public sector	2019.Public sec
Lower Fence	2.140	3.00	3.13	3.25
Q1	13.615	17.31	20.00	25.00
Median	19.290	24.04	27.00	34.00
Q3	27.640	35.00	36.00	45.00
Upper Fence	48.580	61.49	60.00	75.00

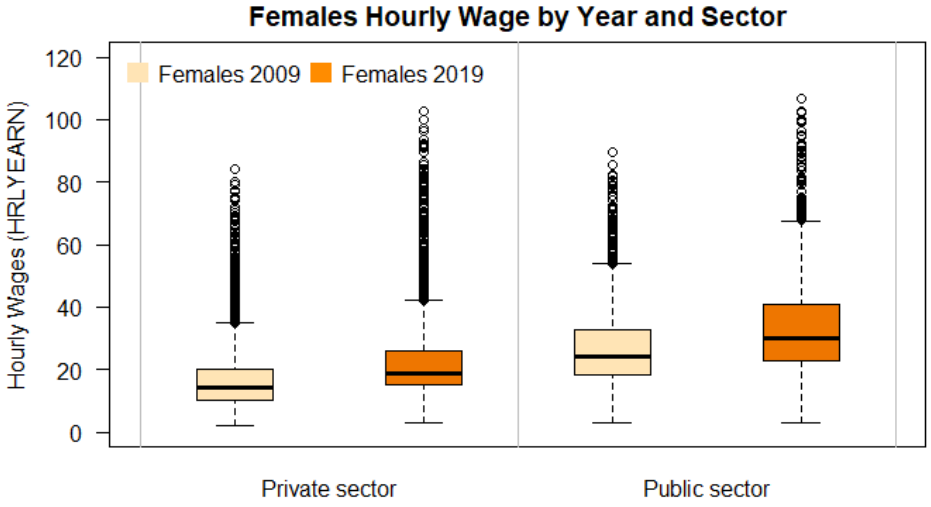
Males 2009 – Public v. Private

```
wilcox.test(HRLYEARN ~ COWMAIN,  data = data.all.09male, alt = "two.sided",
conf.int = T)

        Wilcoxon rank sum test with continuity correction

data:  HRLYEARN by COWMAIN
W = 78664610, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 6.700064 7.270075
sample estimates: difference in location 6.999958
```

Males 2019 – Public v. Private

```
wilcox.test(HRLYEARN ~ COWMAIN,  data = data.all.19male, alt = "two.sided",
conf.int = T)

        Wilcoxon rank sum test with continuity correction

data:  HRLYEARN by COWMAIN
W = 69385905, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 8.039925 8.999984
sample estimates: difference in location 8.500087
```

Females



	2009.Private sector	2019.Private sector	2009.Public sector	2019.Public sec
Lower Fence	2.00	3.13	3.070	3.07
Q1	10.00	15.00	18.500	23.00
Median	14.07	18.75	24.000	30.00
Q3	20.00	25.82	32.695	40.88
Upper Fence	35.00	42.05	53.850	67.31

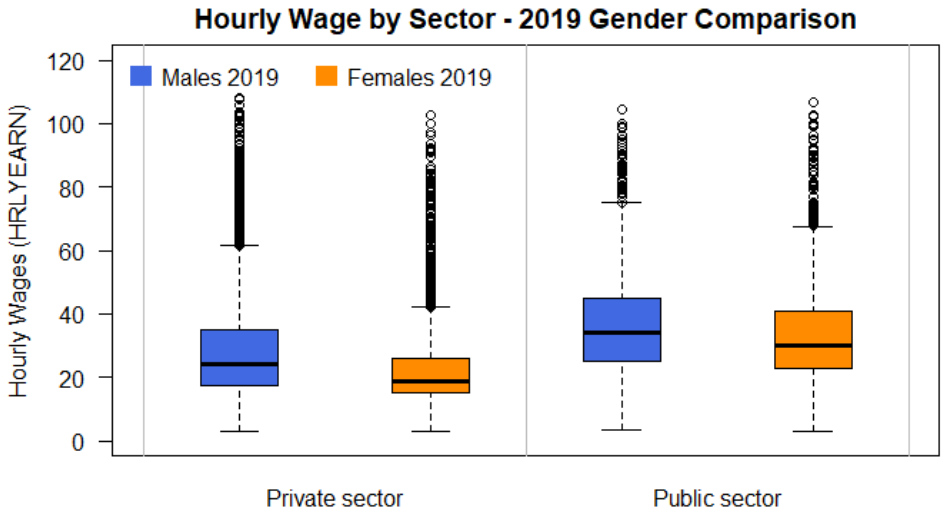
Females 2009 – Public v. Private

```
wilcox.test(HRLYEARN ~ COWMAIN,  data = data.all.09fem, alt = "two.sided", conf.int
= T)

        Wilcoxon rank sum test with continuity correction

data:  HRLYEARN by COWMAIN
W = 133145579, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 8.959938 9.269923
sample estimates: difference in location 9.039943
```

Females 2019 – Public v. Private

```
wilcox.test(HRLYEARN ~ COWMAIN,  data = data.all.19fem, alt = "two.sided", conf.int
= T)

        Wilcoxon rank sum test with continuity correction

data:  HRLYEARN by COWMAIN
W = 112929303, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:  9.710015 10.109941
sample estimates: difference in location 9.999988
```

Gender Comparison



Private Sector 2019 – Male v. Female

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 &
data.all$COWMAIN == "Private sector",], alt = "two.sided", conf.int = T)

        Wilcoxon rank sum test with continuity correction

data:  HRLYEARN by SEX
W = 218257664, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 4.000026 4.499960
sample estimates:
difference in location
              4.220031
```

Public Sector 2019 – Male v. Female

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 &
data.all$COWMAIN == "Public sector",], alt = "two.sided", conf.int = T)

        Wilcoxon rank sum test with continuity correction

data:  HRLYEARN by SEX
W = 24650974, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 2.729960 3.609973
sample estimates:
difference in location
              3.090046
```

Conclusions

The public sector tends to have higher hourly wages for both men and women, and this is consistent over the years. At the same time, men tend to have a higher hourly wage than women in both private and public sectors.

Industry Analysis

First, the hourly wages for men by industry in 2009 and 2019 are displayed in two separate boxplots, and the results of Kruskal Wallis tests are highlighted to see if there is statistical evidence of differences among the industries. Next, the top industries in hourly wages are identified. Then, a comparison between years is presented with another boxplot graph and Mann-Whitney U tests results to verify whether the male's wages in each industry changed over the years. Afterwards, same analysis is applied to women groups. Lastly, a comparison by gender in 2019 is shown alongside the result of Mann-Whitney U tests to confirm if there is evidence of differences in hourly wages for each industry.

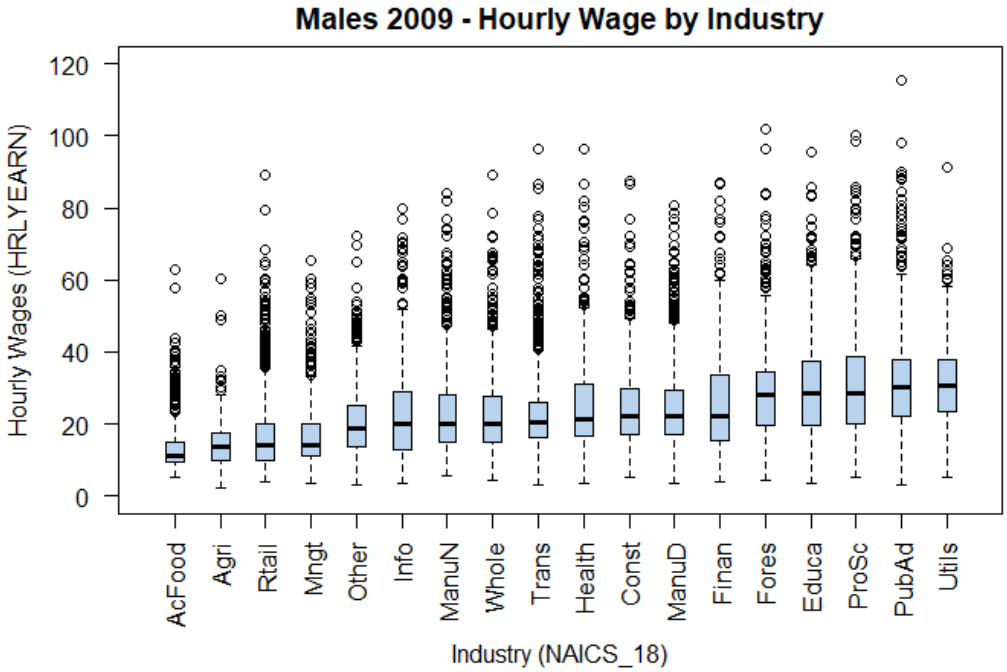
When differences were confirmed by the Kruskal Wallis tests, a post-hoc analysis was performed to identify the industries where hourly wages are significantly different when compared one by one against the others. To save space, the results of post-hoc comparisons of industries are listed on the Appendix section of this document.

Abbreviations for the industries names were used to accommodate them in the graphs and summary of results:

- ✓ Agri = "Agriculture"
- ✓ Fores = "Forestry, Fishing, Mining, Oil and Gas"
- ✓ Utils = "Utilities"
- ✓ Const = "Construction"
- ✓ ManuD = "Manufacturing durables"
- ✓ ManuN = "Manufacturing non-durables"
- ✓ Whole = "Wholesale Trade"
- ✓ Rtail = "Retail Trade"
- ✓ Trans = "Transportation and Warehousing"
- ✓ Finan = "Finance, Insurance, Real Estate and Leasing"
- ✓ ProSc = "Professional, Scientific and Technical Services"
- ✓ Mngt = "Management, Administrative and Other Support"
- ✓ Educa = "Educational Services"
- ✓ Health = "Health Care and Social Assistance"
- ✓ Info = "Information, Culture and Recreation"
- ✓ AcFood = "Accommodation and Food Services"
- ✓ Other = "Other Services"
- ✓ PubAd = "Public Administration"

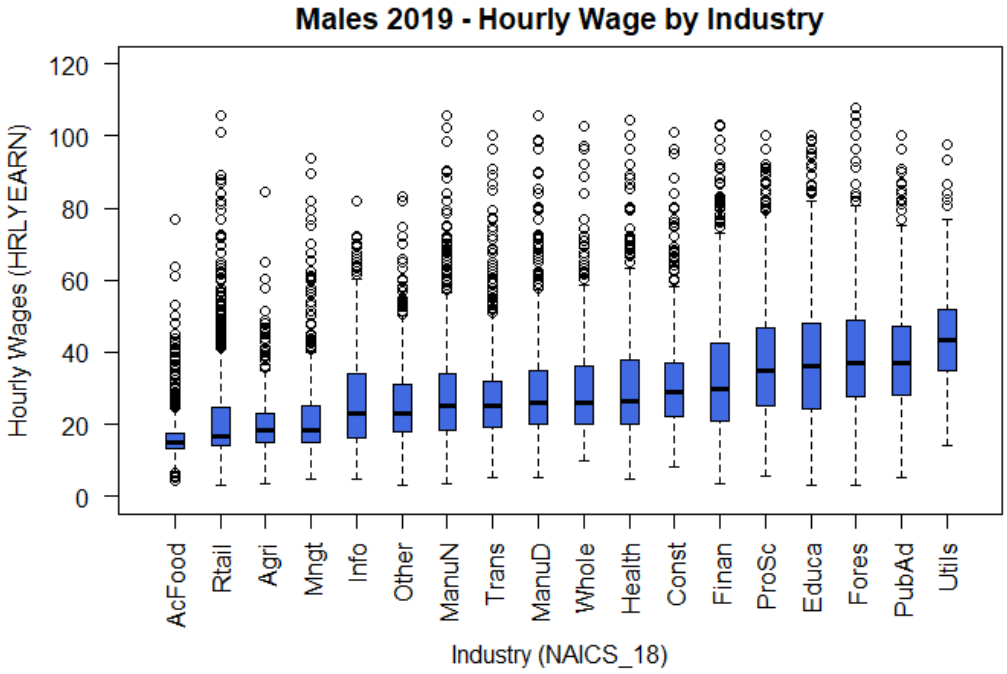
Males

Here are the boxplot results for males. Notice that the industries are ordered by ascending median in both the graph and the summary following it:




```
           AcFood  Agri Rtail  Mngt  Other  Info ManuN Whole  Trans Health
Lower Fence  5.00  2.14  3.79  3.33  3.13  3.380  5.45  4.25  3.210  3.300
Q1           9.35 10.00 10.00 11.00 13.70 13.000 15.00 15.00 16.000 16.465
Median      11.00 13.55 14.00 14.00 18.90 19.815 20.00 20.00 20.510 21.450
Q3          15.00 17.31 20.14 20.00 25.00 28.850 28.00 27.47 25.985 31.000
Upper Fence 23.08 28.00 35.20 33.00 41.83 51.920 47.16 46.15 40.870 52.200

            Const ManuD Finan Fores  Educa ProSc PubAd Utils
Lower Fence  5.10  3.48  3.75  4.44  3.610  5.00  3.13  5.13
Q1          17.00 17.00 15.34 19.78 19.700 20.19 22.03 23.50
Median      22.00 22.00 22.00 28.00 28.280 28.29 30.00 30.64
Q3          29.87 29.51 33.65 34.62 37.555 38.74 37.91 38.00
Upper Fence 49.15 48.08 60.00 55.77 64.100 65.93 61.54 58.24
```

```
kruskal.test(HRLYEARN ~ NAICS_18, data = data.all.09male) # Diff. among groups

        Kruskal-Wallis rank sum test

data:  HRLYEARN by NAICS_18
Kruskal-Wallis chi-squared = 6107.8, df = 17, p-value < 2.2e-16
```



	AcFood	Rtail	Agri	Mngt	Info	Other	ManuN	Trans	ManuD	Whole
Lower Fence	6.92	3.040	3.30	4.56	4.62	3.000	3.53	5.00	5.29	9.62
Q1	13.15	14.000	15.00	15.00	16.00	18.000	18.50	19.35	20.00	20.00
Median	15.00	16.750	18.47	18.50	23.05	23.080	25.00	25.00	26.00	26.00
Q3	17.61	24.855	23.08	25.00	34.07	30.965	34.00	32.00	35.00	36.00
Upper Fence	24.18	41.080	35.00	40.00	60.22	50.000	56.54	50.96	57.50	58.50

	Health	Const	Finan	ProSc	Educa	Fores	PubAd	Utils
Lower Fence	4.81	8.170	3.53	5.77	3.25	3.21	5.13	14.000
Q1	20.00	22.000	21.00	25.00	24.22	27.78	28.00	35.000
Median	26.25	28.745	29.77	35.00	36.06	37.00	37.00	43.475
Q3	37.90	37.000	42.31	46.67	48.08	49.04	47.00	52.000
Upper Fence	63.37	58.000	72.92	78.85	82.05	80.77	75.00	76.920

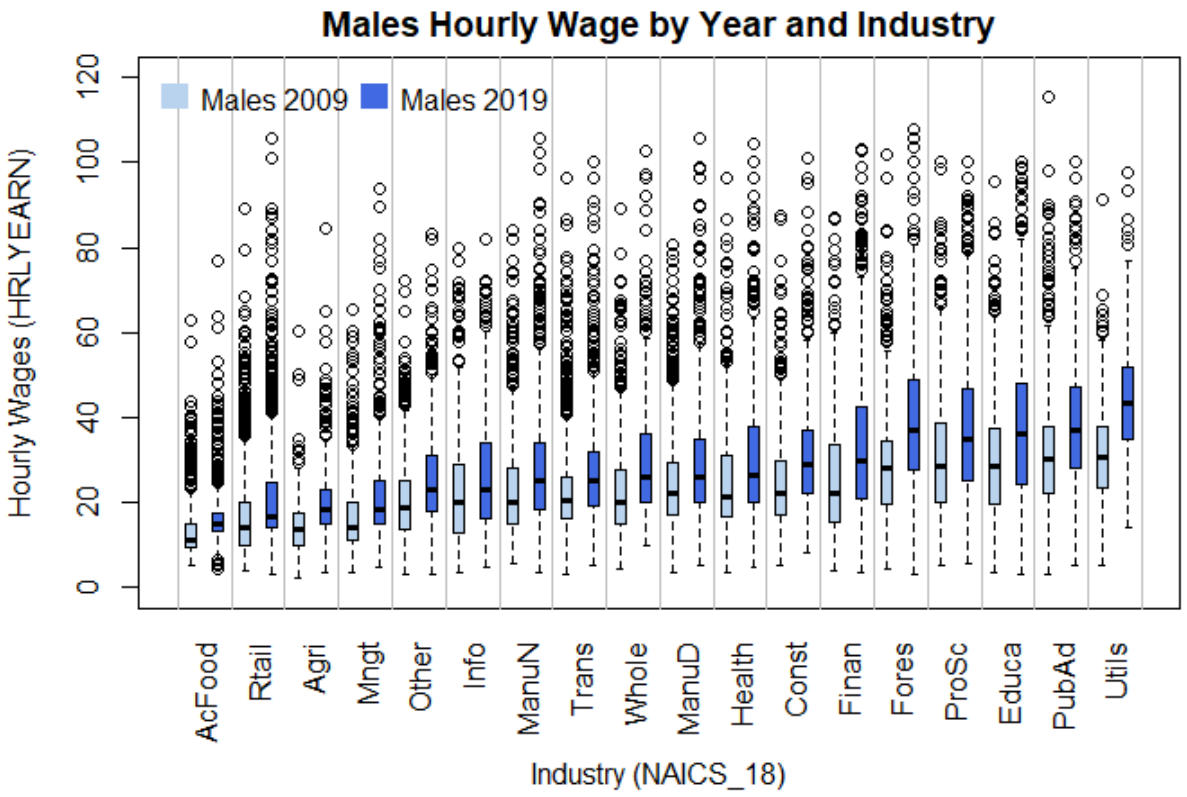
```
kruskal.test(HRLYEARN ~ NAICS_18short, data = data.all.19male) # Diff. among groups

        Kruskal-Wallis rank sum test

data:  HRLYEARN by NAICS_18short
Kruskal-Wallis chi-squared = 6604, df = 17, p-value < 2.2e-16
```

Top 5 Industries

In 2019, the industries with highest median for men were:

1. Utilities
2. Public Administration
3. Forestry, Fishing, Mining, Oil and Gas
4. Educational Services
5. Professional, Scientific and Technical Services

These are the same industries in the top 5 of 2009. To see if there was a change in these industries hourly wages with respect to 2009, Man Whitney U tests were performed, and a graph with year comparison was built:



Industry: Utilities - 2009 vs. 2019, Males

```
wilcox.test(HRLYEARN ~ SURVYEAR,  data = data.all[data.all$SEX == "Male" &
data.all$NAICS_18short == "Utils",], alt = "two.sided", conf.int = T) # Top1
data:  HRLYEARN by SURVYEAR
W = 44871, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: -14.07003 -10.89997
sample estimates: difference in location -12.52998
```

Industry: Public Administration – 2009 vs. 2019, Males

```
wilcox.test(HRLYEARN ~ SURVYEAR,  data = data.all[data.all$SEX == "Male" &
data.all$NAICS_18short == "PubAd",], alt = "two.sided", conf.int = T) # Top2
data:  HRLYEARN by SURVYEAR
W = 1116584, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: -8.049964 -6.410015
sample estimates: difference in location -7.230063
```

Industry: Forestry, Fishing, Mining, Oil and Gas – 2009 vs. 2019, Males

```
wilcox.test(HRLYEARN ~ SURVYEAR,  data = data.all[data.all$SEX == "Male" &
data.all$NAICS_18short == "Fores",], alt = "two.sided", conf.int = T) # Top3
data:  HRLYEARN by SURVYEAR
W = 502548, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: -10.849993  -8.589955
```

```
sample estimates: difference in location -9.739952
```

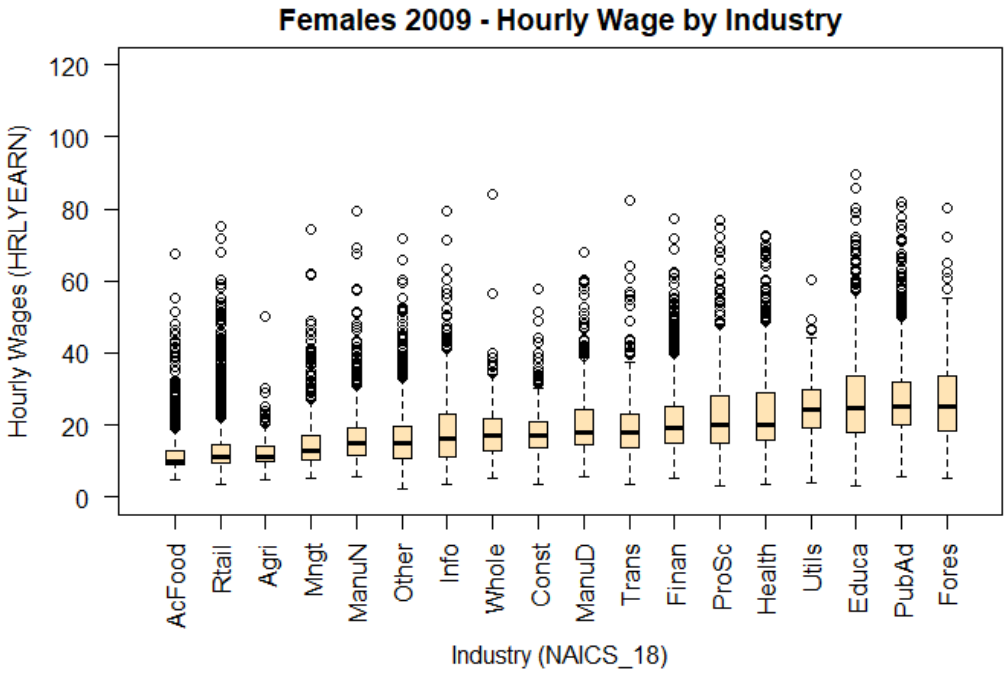
Industry: Educational Services – 2009 vs. 2019, Males

```
wilcox.test(HRLYEARN ~ SURVYEAR,  data = data.all[data.all$SEX == "Male" &
data.all$NAICS_18short == "Educa",], alt = "two.sided", conf.int = T) # Top4
data:  HRLYEARN by SURVYEAR
W = 677554, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: -8.579975 -6.310040
sample estimates: difference in location -7.450054
```

Industry: Professional, Scientific and Technical Services – 2009 vs. 2019, Males

```
wilcox.test(HRLYEARN ~ SURVYEAR,  data = data.all[data.all$SEX == "Male" &
data.all$NAICS_18short == "ProSc",], alt = "two.sided", conf.int = T) # Top5
data:  HRLYEARN by SURVYEAR
W = 554060, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: -6.899933 -4.619922
sample estimates: difference in location -5.719948
```

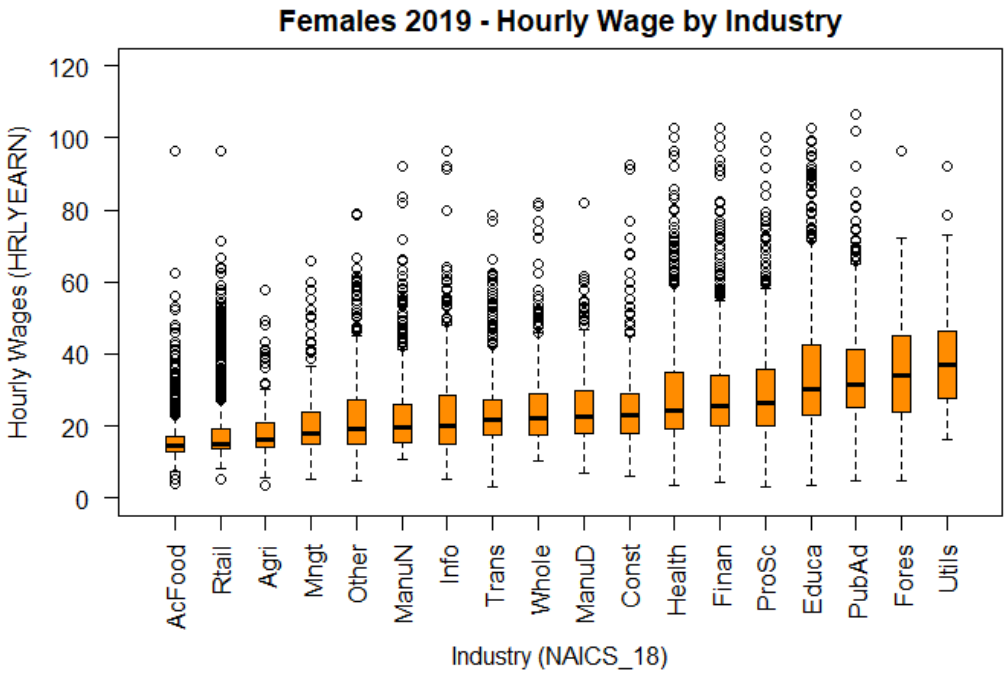
Females



	AcFood	Rtail	Agri	Mngt	ManuN	Other	Info	Whole	Const	ManuD	Trans
Lower Fence	4.77	3.66	4.81	5.03	5.77	2.00	3.55	5.13	3.50	5.490	3.55
Q1	9.00	9.50	9.70	10.30	11.50	10.53	11.00	13.00	13.85	14.420	13.50
Median	10.00	10.95	11.00	13.00	15.00	15.00	16.00	16.83	16.92	18.000	18.00
Q3	13.00	14.50	14.00	17.00	19.35	19.56	23.00	21.63	20.77	24.105	23.12
Upper Fence	19.00	22.00	20.00	27.00	31.00	32.97	40.87	34.00	30.29	38.460	37.33

	Finan	ProSc	Health	Utils	Educa	PubAd	Fores
Lower Fence	5.22	3.08	3.50	3.90	3.07	5.49	5.00
Q1	15.00	15.00	15.87	19.00	18.00	20.00	18.46
Median	19.23	20.00	20.00	24.16	24.62	25.00	25.25
Q3	25.00	28.21	29.07	29.78	33.65	31.87	33.65
Upper Fence	40.00	47.99	48.73	44.00	57.05	49.60	55.27

```
kruskal.test(HRLYEARN ~ NAICS_18short, data = data.all.09fem) # Diff. among groups

        Kruskal-Wallis rank sum test

data:  HRLYEARN by NAICS_18short
Kruskal-Wallis chi-squared = 9226.4, df = 17, p-value < 2.2e-16
```



	AcFood	Rtail	Agri	Mngt	Other	ManuN	Info	Trans	Whole	ManuD
Lower Fence	7.41	8.25	5.770	5.26	4.81	10.50	5.00	3.070	10.30	6.92
Q1	13.00	13.50	14.000	14.77	15.00	15.34	15.00	17.465	17.50	18.03
Median	14.50	15.00	16.000	18.00	19.17	19.75	20.00	21.720	22.09	22.50
Q3	16.86	19.00	20.875	23.67	27.40	26.00	28.35	27.295	28.85	29.81
Upper Fence	22.50	27.25	30.000	36.54	45.05	41.03	48.21	42.000	45.67	46.77

	Const	Health	Finan	ProSc	Educa	PubAd	Fores	Utils
Lower Fence	6.07	3.50	4.360	3.13	3.48	4.810	4.730	16.07
Q1	18.00	19.00	19.975	20.00	23.00	25.295	23.875	27.85
Median	23.00	24.04	25.640	26.44	30.00	31.370	34.000	37.00
Q3	28.90	35.00	34.055	35.49	42.31	41.325	45.095	46.15
Upper Fence	45.00	59.00	54.950	58.00	71.15	65.000	72.120	72.82

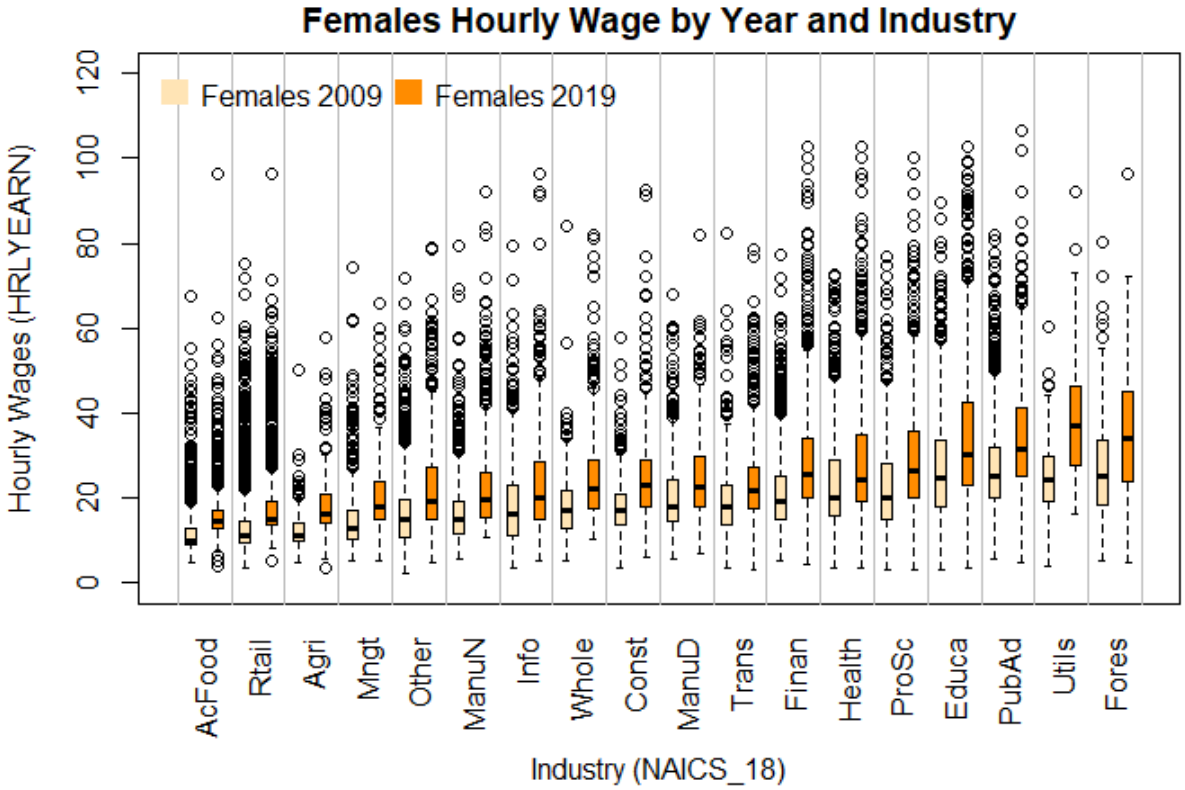
```
kruskal.test(HRLYEARN ~ NAICS_18short, data = data.all.19fem) # Diff. among
groups

        Kruskal-Wallis rank sum test

data:  HRLYEARN by NAICS_18short
Kruskal-Wallis chi-squared = 7815.6, df = 17, p-value < 2.2e-16
```

Top 5 Industries

In 2019, the industries with highest median for women were:

1. Utilities
2. Forestry, Fishing, Mining, Oil and Gas
3. Public Administration
4. Educational Services
5. Professional, Scientific and Technical Services

The only difference compared to the top 5 of 2009 for females is the inclusion of "Health Care and Social Assistance" instead of “Professional, Scientific and Technical Services” that we see in 2019. To verify if there were changes in the top industries hourly wages with respect to 2009, Man Whitney U tests were performed, and a graph with year comparison was built:



Industry: Utilities - 2009 vs. 2019, Females

```
wilcox.test(HRLYEARN ~ SURVYEAR,  data = data.all[data.all$SEX == "Female" & data.all$
NAICS_18short == "Utils",], alt = "two.sided", conf.int = T) # Top1
data:  HRLYEARN by SURVYEAR
W = 3277.5, p-value = 3.859e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: -14.999966  -9.439992
sample estimates: difference in location -12.13
```

Industry: Forestry, Fishing, Mining, Oil and Gas – 2009 vs. 2019, Females

```
wilcox.test(HRLYEARN ~ SURVYEAR,  data = data.all[data.all$SEX == "Female" & data.all$
NAICS_18short == "Fores",], alt = "two.sided", conf.int = T) # Top3
data:  HRLYEARN by SURVYEAR
W = 18717, p-value = 2.28e-10
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: -10.649967  -5.650075
sample estimates: difference in location -8.030013
```

Industry: Public Administration – 2009 vs. 2019, Females

```
wilcox.test(HRLYEARN ~ SURVYEAR,  data = data.all[data.all$SEX == "Female" & data.all$
NAICS_18short == "PubAd",], alt = "two.sided", conf.int = T) # Top2
data:  HRLYEARN by SURVYEAR
W = 1070692, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: -7.280006 -5.999961
sample estimates: difference in location -6.669972
```

Industry: Educational Services – 2009 vs. 2019, Females

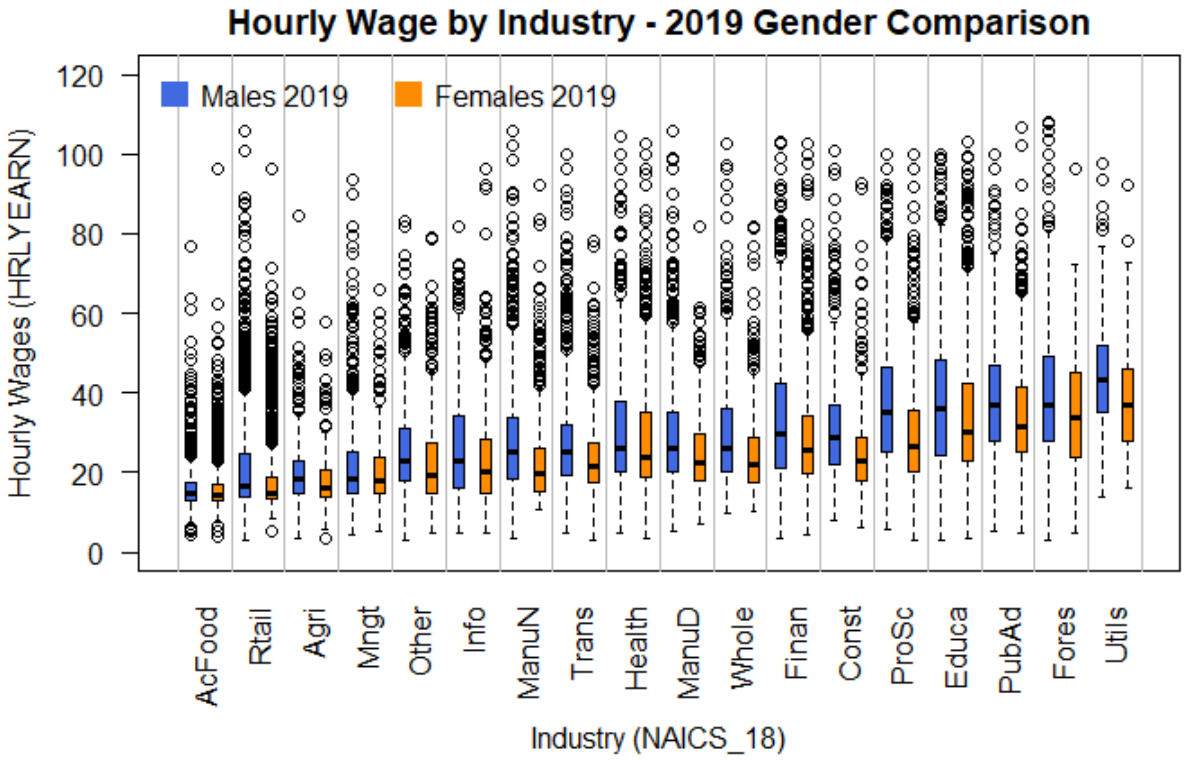
```
wilcox.test(HRLYEARN ~ SURVYEAR,  data = data.all[data.all$SEX == "Female" & data.all$
NAICS_18short == "Educa",], alt = "two.sided", conf.int = T) # Top4
data:  HRLYEARN by SURVYEAR
W = 3546959, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: -6.269986 -5.099947
sample estimates: difference in location -5.749971
```

Industry: Professional, Scientific and Technical Services – 2009 vs. 2019, Females

```
wilcox.test(HRLYEARN ~ SURVYEAR,  data = data.all[data.all$SEX == "Female" & data.all$
NAICS_18short == "ProSc",], alt = "two.sided", conf.int = T) # Top5
data:  HRLYEARN by SURVYEAR
W = 404999, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: -6.970001 -5.270070
sample estimates: difference in location -6.040043
```

Gender Comparison

To compare differences of hourly wages distribution in 2019, we look at the boxplot graph. Mann Whitney U tests were also ran and the ones with significant results are included after the visualization.



Industry: Utilities – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NAIC
S_18short == "Utils",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 27312, p-value = 4.826e-05
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 3.089948 8.900052
sample estimates: difference in location 6.000049
```

Industry: Forestry, Fishing, Mining, Oil and Gas – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NAIC
S_18short == "Fores",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 166042, p-value = 0.003606
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 1.000054 5.290070
sample estimates: difference in location 3.100048
```

Industry: Public Administration – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NAIC
S_18short == "PubAd",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 1873268, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 3.380016 5.049965
sample estimates: difference in location 4.219966
```

Industry: Educational Services – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NAIC
S_18short == "Educa",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 2383919, p-value = 2.516e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 3.000094 4.999943
sample estimates: difference in location 4.000059
```

Industry: Professional, Scientific and Technical Services – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NAIC
S_18short == "ProSc",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 1019833, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 6.169975 8.279945
sample estimates: difference in location 7.210066
```

Industry: Construction – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NAIC
S_18short == "Const",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 1020349, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 3.999967 5.829932
sample estimates: difference in location 4.999997
```

Industry: Finance, Insurance, Real Estate and Leasing – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NAIC
S_18short == "Finan",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 879092, p-value = 5.238e-12
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 2.469944 4.499994
sample estimates: difference in location 3.479976
```

Industry: Wholesale Trade – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NAIC
S_18short == "Whole",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 327178, p-value = 8.746e-13
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 2.500003 4.500087
sample estimates: difference in location 3.499906
```

Industry: Manufacturing durables – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NAIC
S_18short == "ManuD",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 699118, p-value = 1.544e-12
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 2.100034 3.950029
sample estimates: difference in location 3.000056
```

Industry: Health Care and Social Assistance – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NAIC
S_18short == "Health",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 4094719, p-value = 4.884e-07
```

```
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 0.9999984 2.1500867
sample estimates: difference in location 1.599926
```

Industry: Transportation and Warehousing – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NAICS_18short == "Trans",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 756647, p-value = 6.886e-12
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 1.929985 3.349944
sample estimates: difference in location 2.600061
```

Industry: Manufacturing non-durables – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NAICS_18short == "ManuN",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 790826, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 3.550011 5.000051
sample estimates: difference in location 4.24996
```

Industry: Information, Culture and Recreation – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NAICS_18short == "Info",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 500721, p-value = 9.825e-10
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 1.770008 3.499980
sample estimates: difference in location 2.549954
```

Industry: Management, Administrative and Other Support – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NAICS_18short == "Mngt",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 346817, p-value = 0.001885
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 0.250000 1.499922
sample estimates: difference in location 0.9999967
```

Industry: Agriculture – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NAICS_18short == "Agri",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 68024, p-value = 6.254e-05
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 0.9999976 2.5000029
sample estimates: difference in location 1.750036
```

Industry: Retail Trade – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NAICS_18short == "Rtail",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 5449884, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 1.000037 1.500014
sample estimates: difference in location 1.25008
```

Industry: Accommodations and Food Services – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NAICS_18short == "AcFood",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 1550054, p-value = 0.008176
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:  4.414496e-05 4.999908e-01
sample estimates: difference in location 0.14998
```

Conclusions

From the visualizations and non-parametric tests, we can conclude that:

- There are differences in the hourly wages among industries, for both men and women.
- These differences have been persistent over time, and we found them in 2009 and 2019.
- The top 5 industries in 2019 are the same for men and women. It has not changed over time; we observed the same top 5 industries in 2009:
  - ✓ Utilities
  - ✓ Public Administration
  - ✓ Forestry, Fishing, Mining, Oil and Gas
  - ✓ Educational Services
  - ✓ Professional, Scientific and Technical Services
- However, in the gender comparison we see that men tend to have higher hourly wages across all industries.
- The industries where this gap seems to be smallest are "Management, Administrative and Other Support" and "Accommodations and Food Services".

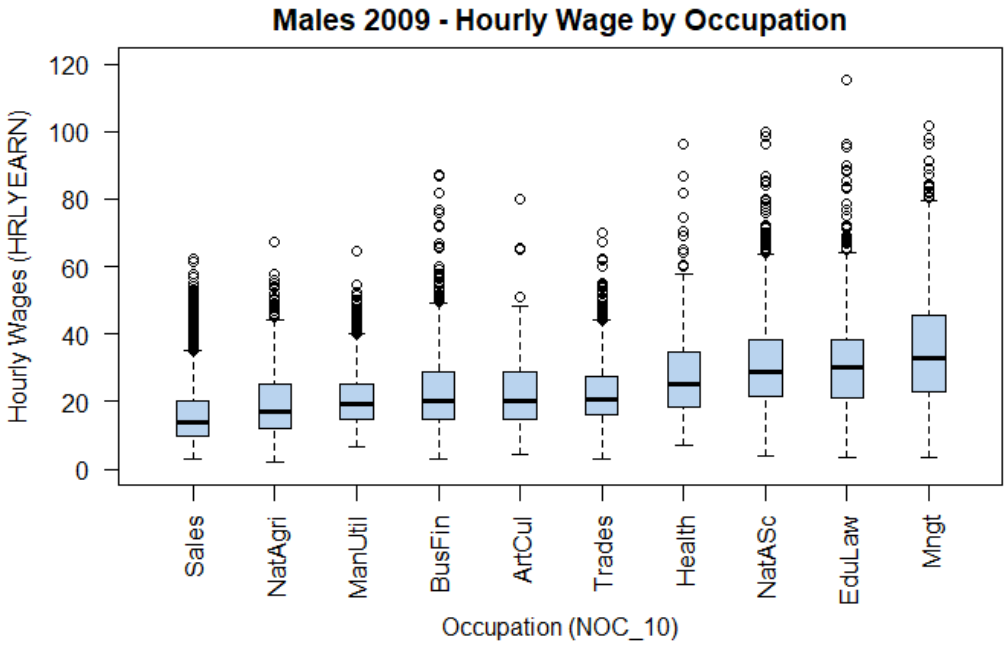
Occupation Analysis

There are 10 main categories of job occupations, and the variation of hourly wages among them is explored. We start with a breakdown by year for men with boxplots and Kruskal Wallis tests to confirm significant differences. Then we compare results over time to see if men's hourly wages changed from 2009 to 2019, using boxplots and Mann-Whitney U tests. Next, same sequence of analysis is presented for women. Afterwards, there is a comparison by gender for the 2019 data where boxplot and Mann-Whitney U tests are employed again to confirm if hourly wages are different for men and women performing same occupations.

Abbreviations for the occupations names were used to accommodate them in the graphs and summary of results:

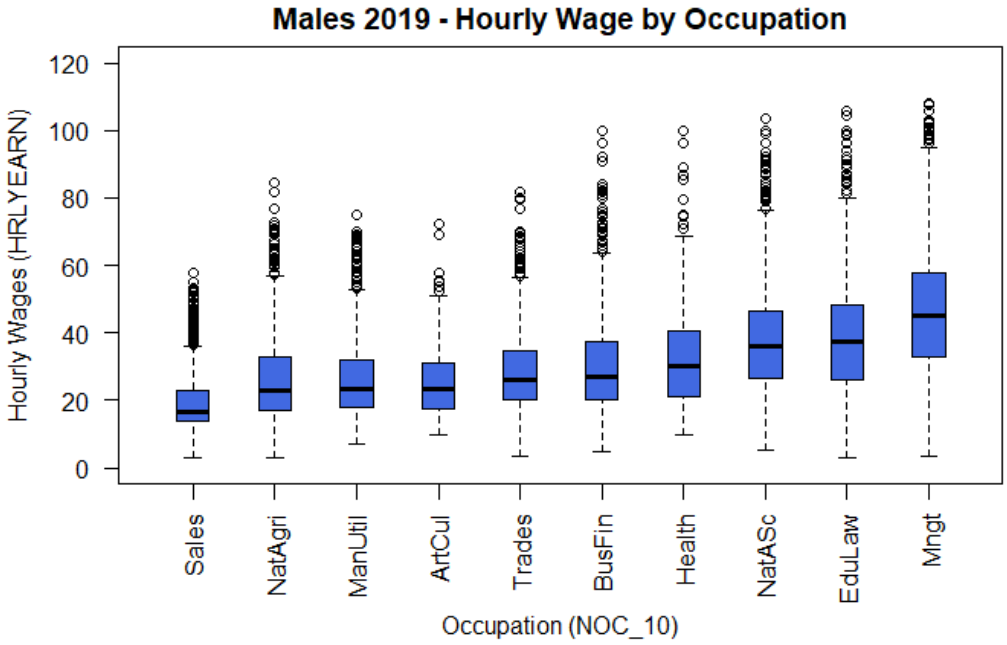
- ✓ Mngt = "Management"
- ✓ BusFin = "Business, finance and administration"
- ✓ NatASc = "Natural and applied sciences"
- ✓ Health ="Health"
- ✓ EduLaw = "Education, law, community and government services"
- ✓ ArtCul = "Art, culture, recreation and sport"
- ✓ Sales = "Sales and service"
- ✓ Trades = "Trades, transport and equipment operators"
- ✓ NatAgri = "Natural resources and agriculture"
- ✓ ManUtil = "Manufacturing and utilities

Males



	Sales	NatAgri	ManUtil	BusFin	ArtCul	Trades	Health	NatASc	EduLaw	Mngt
LowFence	3.13	2.14	6.67	3.13	4.210	3.210	7.00	3.79	3.610	3.48
Q1	10.00	12.00	15.00	15.00	15.000	16.000	18.50	21.63	20.975	23.08
Median	14.00	17.00	19.50	20.19	20.295	20.675	25.00	28.85	30.000	33.00
Q3	20.00	25.18	25.00	28.69	28.850	27.300	34.59	38.41	38.460	45.64
UpFence	35.00	44.41	40.00	49.04	48.210	44.230	57.69	63.57	64.100	79.47

```
kruskal.test(HRLYEARN ~ NOC_10, data = data.all.09male) # Diff. among groups

        Kruskal-Wallis rank sum test

data:  HRLYEARN by NOC_10
Kruskal-Wallis chi-squared = 5740.1, df = 9, p-value < 2.2e-16
```



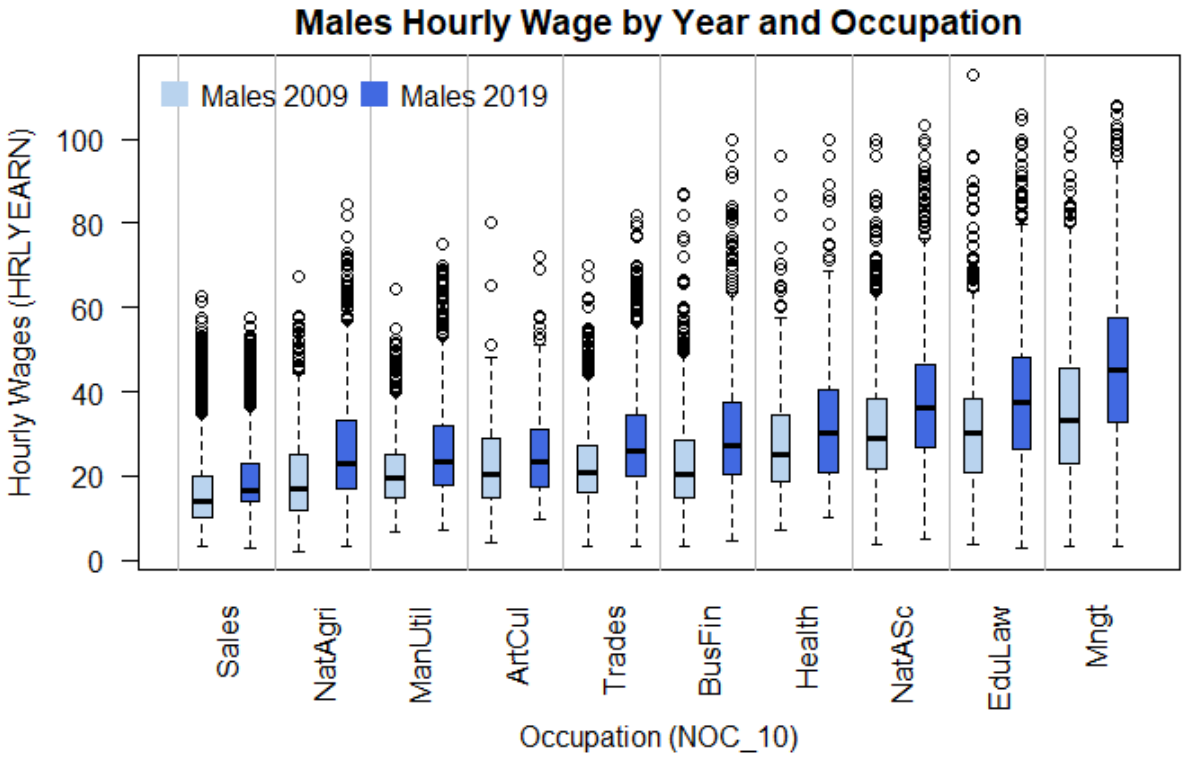
	Sales	NatAgri	ManUtil	ArtCul	Trades	BusFin	Health	NatASc	EduLaw	Mngt
LowFence	3.04	3.21	6.92	9.85	3.460	4.62	10.00	5.13	3.00	3.30
Q1	14.00	17.00	18.00	17.50	20.000	20.19	21.00	26.62	26.25	32.88
Median	16.50	23.00	23.36	23.50	26.000	27.00	30.00	36.06	37.50	45.00
Q3	23.00	33.00	32.00	31.25	34.625	37.50	40.49	46.63	48.08	57.69
UpFence	36.13	57.00	53.00	51.00	56.500	63.46	68.68	76.51	79.91	94.87

```
kruskal.test(HRLYEARN ~ NOC_10, data = data.all.19male) # Diff. among groups

        Kruskal-Wallis rank sum test

data:  HRLYEARN by NOC_10
Kruskal-Wallis chi-squared = 7220.2, df = 9, p-value < 2.2e-16
```

Top 3 Occupations

In 2009 and 2019, the occupations with highest median for men were:

1. Management
2. Education, law, community, and government services
3. Natural and applied sciences



Occupation: Management – 2009 vs. 2019, Males

```
wilcox.test(HRLYEARN ~ SURVYEAR,  data = data.all[data.all$SEX == "Male" & data.all$NO
C_10short == "Mngt",], alt = "two.sided", conf.int = T) # Top1
data:  HRLYEARN by SURVYEAR
W = 1120265, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: -12.25999 -10.00005
sample estimates: difference in location -11.17997
```

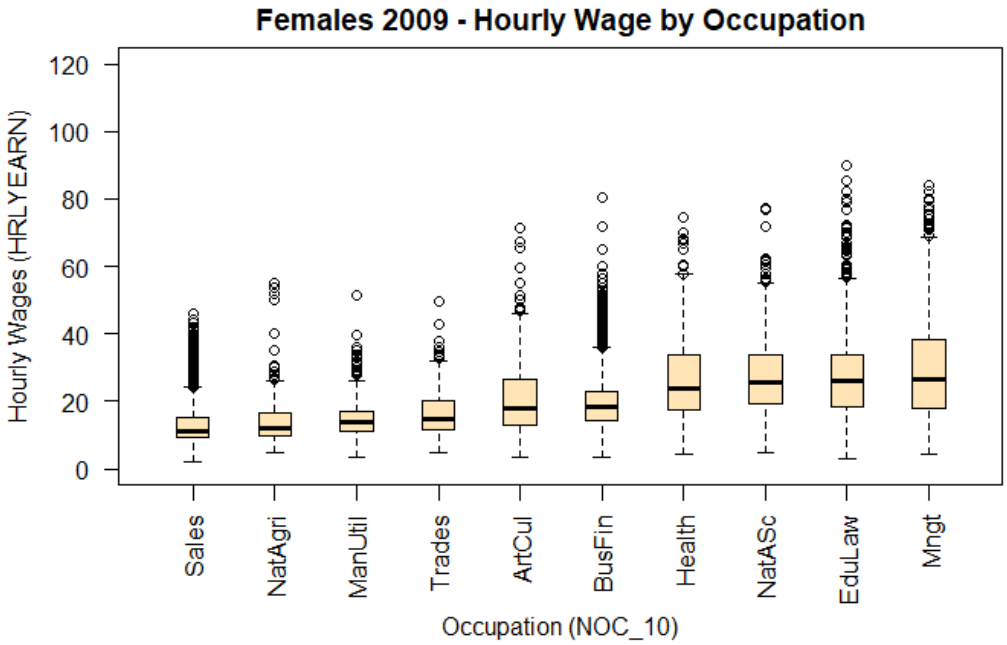
Occupation: Education, law, community, and government services – 2009 vs. 2019, Males

```
wilcox.test(HRLYEARN ~ SURVYEAR,  data = data.all[data.all$SEX == "Male" & data.all$NO
C_10short == "EduLaw",], alt = "two.sided", conf.int = T) # Top2
data:  HRLYEARN by SURVYEAR
W = 966244, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: -8.059976 -6.049982
sample estimates: difference in location -7.05005
```

Occupation: Natural and applied sciences – 2009 vs. 2019, Males

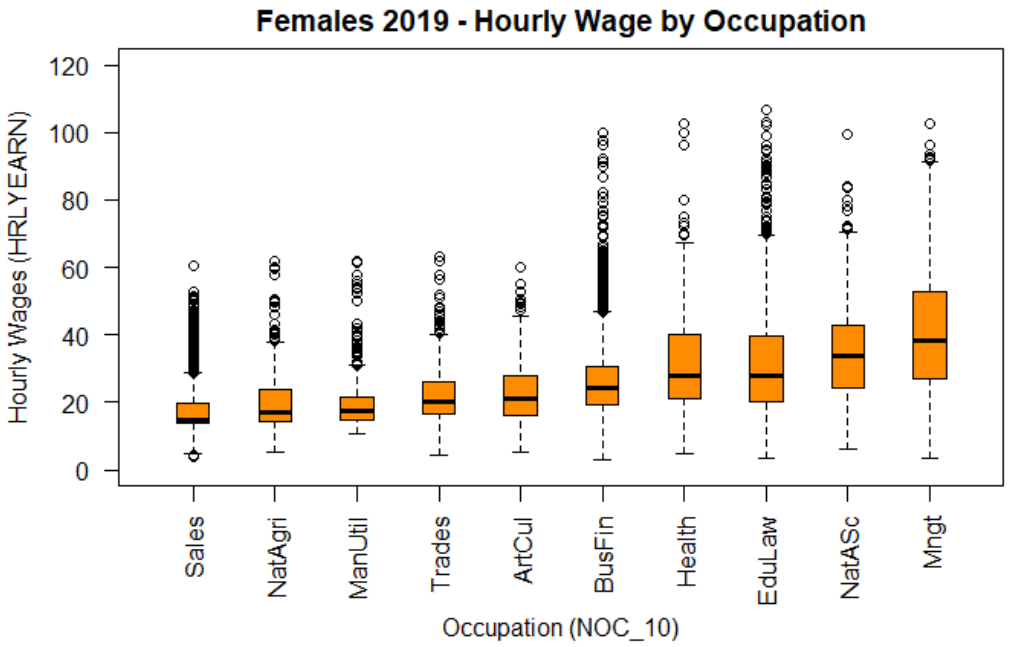
```
wilcox.test(HRLYEARN ~ SURVYEAR,  data = data.all[data.all$SEX == "Male" & data.all$NO
C_10short == "NatASc",], alt = "two.sided", conf.int = T) # Top3
data:  HRLYEARN by SURVYEAR
W = 2564647, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: -7.489955 -6.020037
sample estimates: difference in location -6.770048
```

Females




```
          Sales NatAgri ManUtil Trades ArtCul BusFin Health NatASc EduLaw   Mngt
LoFence   2.000    4.81   3.500   4.63   3.55   3.45   4.17  5.000  3.070  4.160
Q1        9.500   10.00  10.970  11.50  13.00  14.35  17.55 19.220 18.495 18.000
Median   11.000   12.00  13.685  15.00  18.00  18.50  23.90 25.640 26.000 26.555
Q3       15.365   16.50  17.150  20.00  26.39  23.08  33.65 33.685 33.750 38.460
UpFence  24.130   25.96  26.000  32.09  46.15  36.06  57.69 55.000 56.410 68.680
```

```
kruskal.test(HRLYEARN ~ NOC_10, data = data.all.09fem) # Diff. among groups

        Kruskal-Wallis rank sum test

data:  HRLYEARN by NOC_10
Kruskal-Wallis chi-squared = 9838.5, df = 9, p-value < 2.2e-16
```




```
         Sales NatAgri ManUtil Trades ArtCul BusFin Health EduLaw NatASc  Mngt
LoFence   5.00    5.26   10.50   4.55   5.42   3.07   5.05   3.48   6.25  3.30
Q1       13.75   14.50   15.00  16.54  16.00  19.49  21.00  20.00  24.52 26.92
Median   15.00   17.00   17.50  20.00  21.00  24.34  27.79  27.88  33.64 38.46
Q3       19.79   23.97   21.45  26.00  27.89  30.53  40.00  39.90  43.00 52.88
UpFence  28.72   38.00   31.00  40.15  45.64  47.00  67.31  69.71  70.62 91.35
```

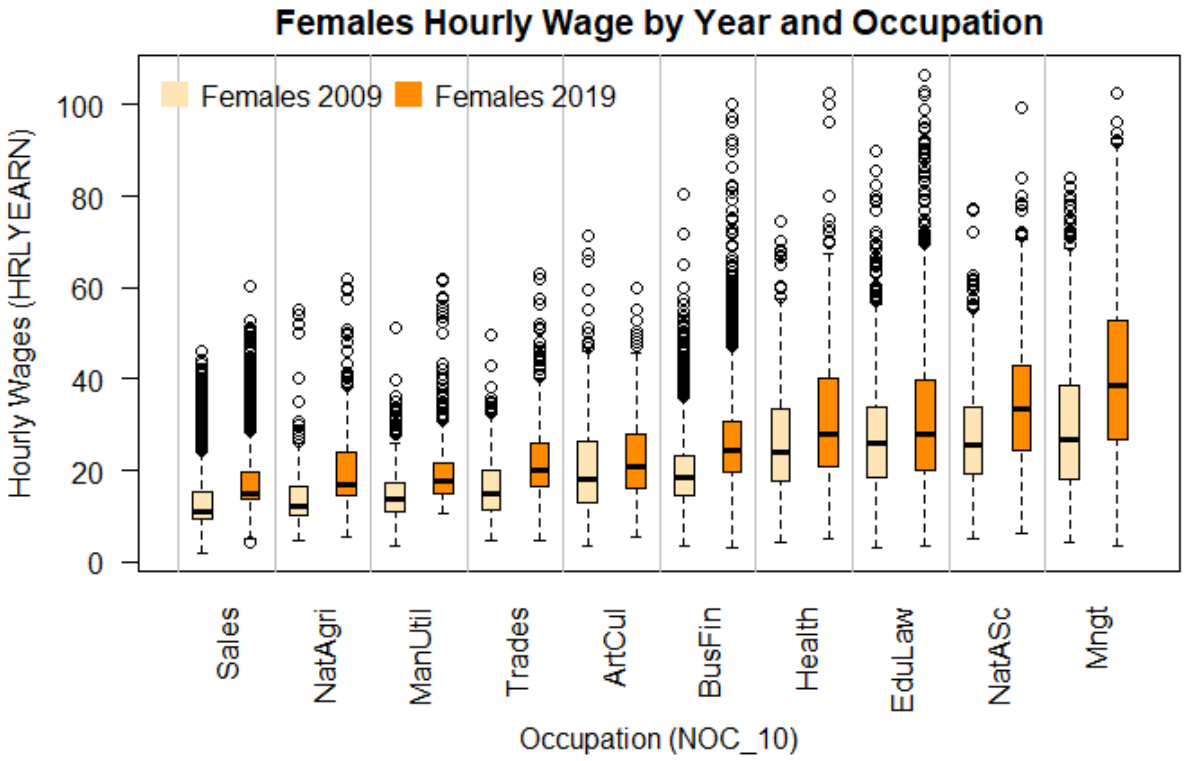
```
kruskal.test(HRLYEARN ~ NOC_10, data = data.all.19fem) # Diff. among groups

        Kruskal-Wallis rank sum test

data:  HRLYEARN by NOC_10
Kruskal-Wallis chi-squared = 8336.7, df = 9, p-value < 2.2e-16
```

Top 3 Occupations

In 2009 and 2019, the occupations with highest median for women were:

1. Management
2. Natural and applied sciences
3. Education, law, community, and government services

Only difference between years us the order, “Natural and applied sciences” jumped from the third to the second spot in 2019.



Occupation: Management – 2009 vs. 2019, Females

```
wilcox.test(HRLYEARN ~ SURVYEAR,  data = data.all[data.all$SEX == "Female" & data.all$
NOC_10short == "Mngt",], alt = "two.sided", conf.int = T) # Top1
data:  HRLYEARN by SURVYEAR
W = 593358, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: -12.019993  -9.619946
sample estimates: difference in location -10.81004
```

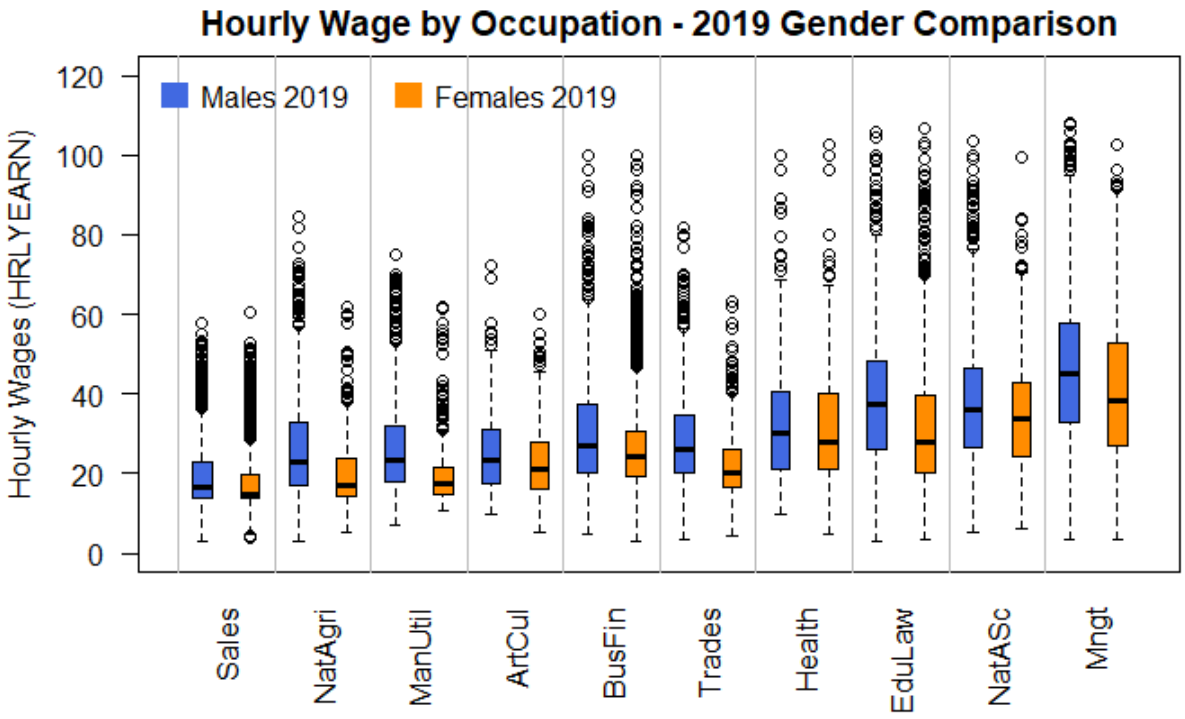
Occupation: Natural and applied sciences – 2009 vs. 2019, Females

```
wilcox.test(HRLYEARN ~ SURVYEAR,  data = data.all[data.all$SEX == "Female" & data.all$
NOC_10short == "NatASc",], alt = "two.sided", conf.int = T) # Top2
data:  HRLYEARN by SURVYEAR
W = 208380, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: -8.200015 -5.870033
sample estimates: difference in location -7.000035
```

Occupation: Education, law, community, and government services – 2009 vs. 2019, Females

```
wilcox.test(HRLYEARN ~ SURVYEAR,  data = data.all[data.all$SEX == "Female" & data.all$
NOC_10short == "EduLaw",], alt = "two.sided", conf.int = T) # Top3
data:  HRLYEARN by SURVYEAR
W = 7142727, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: -3.560001 -2.509974
sample estimates:ndifference in location -3.000021
```

Gender Comparison



Occupation: Management – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NOC_
10short == "Mngt",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 1268072, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 4.809938 7.350044
sample estimates: difference in location 6.080045
```

Occupation: Natural and applied sciences – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NOC_
10short == "NatASc",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 1290381, p-value = 2.365e-08
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 1.950024 4.000029
sample estimates: difference in location 2.999974
```

Occupation: Education, law, community and government services – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NOC_
10short == "EduLaw",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 5310381, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 6.639996 8.200009
sample estimates: difference in location 7.430066
```

Occupation: Health – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NOC_
10short == "Health",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 1140716, p-value = 0.01428
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 0.1500462 2.1200176
sample estimates: difference in location 1.150062
```

Occupation: Trades, transport and equipment operators – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NOC_
10short == "Trades",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 2803271, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 4.080045 5.749970
sample estimates: difference in location 4.999944
```

Occupation: Business, finance and administration – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NOC_
10short == "BusFin",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 7312384, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 2.130060 3.169963
sample estimates: difference in location 2.709973
```

Occupation: Art, culture, recreation and sport – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NOC_
10short == "ArtCul",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 104442, p-value = 0.0001424
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 1.000031 3.500068
sample estimates: difference in location 2.219955
```

Occupation: Manufacturing and utilities – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NOC_
10short == "ManUtil",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 903084, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 4.500042 5.949957
sample estimates: difference in location 5.050023
```

Occupation: Natural resources and agriculture – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NOC_
10short == "NatAgri",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 269920, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 3.500010 5.710072
sample estimates: difference in location 4.619989
```

Occupation: Sales and service – 2019 – Males vs. Females

```
wilcox.test(HRLYEARN ~ SEX,  data = data.all[data.all$SURVYEAR == 2019 & data.all$NOC_10short == "Sales",], alt = "two.sided", conf.int = T)
data:  HRLYEARN by SEX
W = 22827094, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval: 0.9999802 1.1500083
sample estimates: difference in location 1.000046
```

Conclusions

- There are differences in the hourly wages among occupations. This is true regardless the gender and the year.
- Both men and women share the same Top 3 occupations in terms of highest hourly wages: "Management", "Education, law, community, and government services", and "Natural and applied sciences".
- These 3 occupations remained at the top from 2009 to 2019.
- The occupations that seem to have the smallest gender gaps are "Health" and "Art, culture, recreation and sport".

Step 4: Multiple Regression Analysis

Normalization

Numeric variables were normalized before applying the regression techniques to transform them to the same scale. This was the function used:

normalize <- function(x) {return ((x - min(x)) / (max(x) - min(x)))}

Training and Test Set

The dataset was divided into 4 groups, by year and gender:

1. Year 2009, Males
2. Year 2009, Females
3. Year 2019, Males
4. Year 2019, Females

Each of these subsets was split into 70% training and 30% testing data, selecting observations randomly.

Variables excluded

The next variables from the dataset were not included as predictors:

VARIABLE	DESCRIPTION	REASON OF EXCLUSION
rec_num	Order of record in file	Identification of record, unique
survyear	Survey year	One set value in each group (2009 or 2019)
survmnth	Survey month	One value (September)
Sex	Sex of respondent	One value in each group (Male or Female)
Immig	Immigrant status	Only 2019 records, no 2009 data
noc_40	Occupation at main job (40 categories)	Only 2019 records, no 2009 data (noc_10 was the variable included to analyse occupation)

Stepwise Regression

Stepwise regression (both directions) was the feature selection technique applied. Here is an example of the R code used to build the first model:

```
full <- lm(HRLYEARN ~ FSSTAT+PROV+CMA+AGE_12+MARSTAT+EDUC+MJH+COWMAIN
           +NAICS_18+NOC_10+FTPTMAIN+UTOTHRS+TENURE+UNION+PERMTEMP
           +ESTSIZE+FIRMSIZE+SCHOOLN+EFAMTYPE+AGYOWNK, data = train.09male)
null <- lm(HRLYEARN ~ 1, data = train.09male)
model <- stepAIC(null, scope=list(lower=null, upper=full), direction= "both", trace=TRUE)
```

To check the overall significance of the model built, we looked at the **F-statistic** and its p-value. The F-test follows these hypotheses:

- ✓ H0: Null hypothesis states that the model with no independent variables fits the data as well as your model.
- ✓ HA: Alternative hypothesis says that your model fits the data better than the intercept-only model, meaning that at least one of the independent variables in the model can explain some of the variance of the response variable.

If p-value is less or equal to 0.05, the null hypotheses is rejected to conclude that the regression model fits the data better than the model with no independent variables.

R-squared value was also checked. It indicates how well the model fits the data, measuring the strength of the relationship between it and the dependent variable.

Then, diagnostic plots were created to check regression assumptions:

- **Linearity of the data.** The relationship between the predictor (x) and the outcome (y) is assumed to be linear.
- **Normality of residuals.** The residual errors are assumed to be normally distributed.
- **Homogeneity of residuals variance.** The residuals are assumed to have a constant variance (homoscedasticity).
- **Independence of residuals error terms.**

The diagnostic plots employed to analyze model residuals were:

a) **Residuals vs Fitted.** Used to check the linear relationship assumptions. A horizontal line, without distinct patterns is an indication for a linear relationship.

b) **Normal Q-Q.** Used to examine whether the residuals are normally distributed. Residuals points should follow the straight dashed line.

c) **Scale-Location (or Spread-Location).** Used to check the homogeneity of variance of the residuals (homoscedasticity). Horizontal line with equally spread points is a good indication of homoscedasticity. Otherwise, we would say there is a heteroscedasticity problem.

d) **Residuals vs Leverage.** Used to identify influential cases, that is extreme values that might influence the regression results when included or excluded from the analysis. When data points are to the upper or lower right of the leverage plot, outside the Cook's distance lines they have leverage meaning they are influential to the regression results.

After analyzing the diagnostic plots, if required, Box Cox transformation technique was employed to correct for errors nonnormality and other assumptions violations.

Lastly, once a model was determined, it was used to predict the response variable (hourly wages) on the test set. A histogram of the errors was plotted, and both the Root Mean Squared Errors (RMSE) and the Mean Absolute Error (MAE) were calculated. The percentage of predictions that were close to the real value was calculated as well with the next formulas:

```
> diff.percent <- 100*(abs(errors)/(y_real_values))
> diff.25 <- length(diff.percent[diff.percent<=25])/total_observations
```

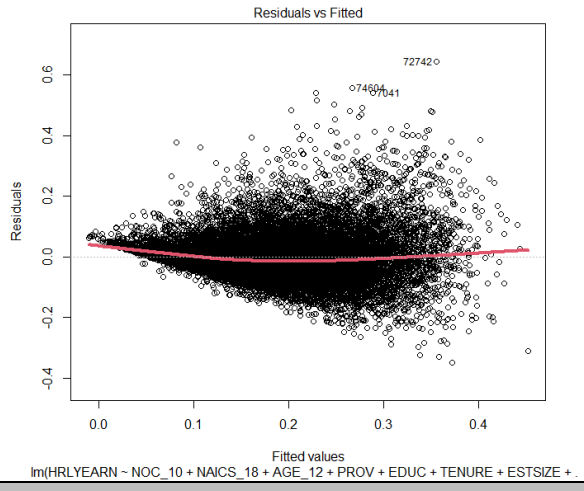
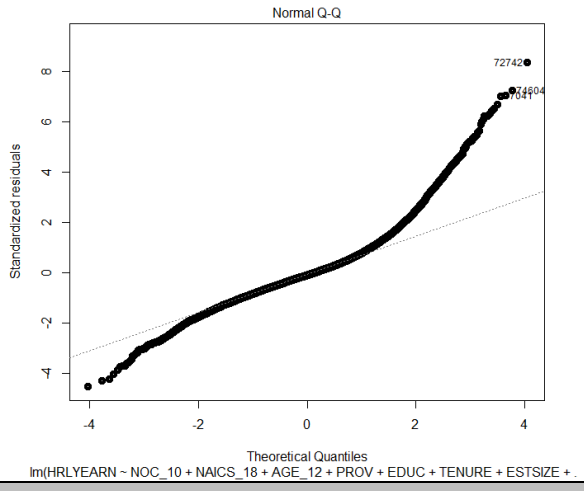
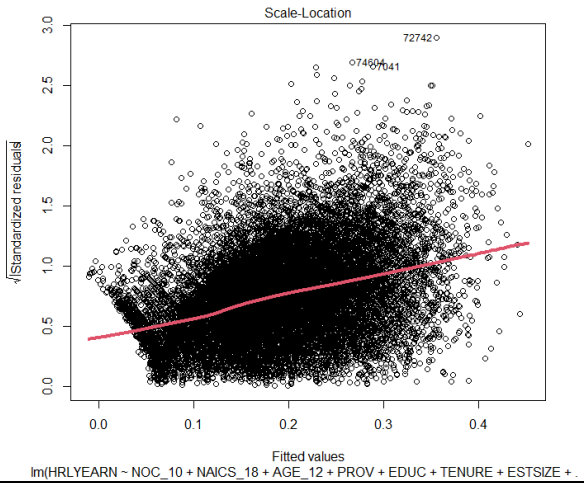
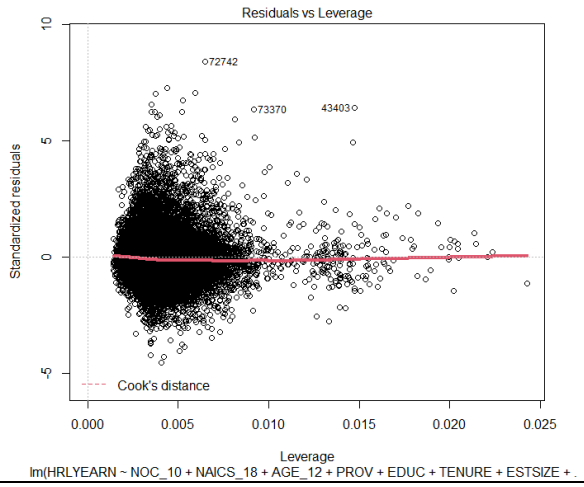
These are the results for each group:

Year 2009, Males
First Model

```
Call:
lm(formula = HRLYEARN ~ NOC_10 + NAICS_18 + AGE_12 + PROV + EDUC +
    TENURE + ESTSIZE + FIRMSIZE + AGYOWNK + PERMTEMP + FTPTMAIN +
    UTOTHRS + UNION + MARSTAT + LFSSTAT + COWMAIN, data = train.09male)
```

```
Multiple R-squared:  0.4733,   Adjusted R-squared:  0.4711
F-statistic: 219.1 on 77 and 18781 DF,  p-value: < 2.2e-16
```

Diagnostic Plots

Residuals vs Fitted (Linearity Test)	Normal Q-Q (Residuals Normality Test)
	
Scale-Location (Homoscedasticity Test)	**Residuals vs. Leverage (Outliers Test)**
	

The conclusions of analyzing the diagnostics plots are:

a) Residuals vs Fitted: The red line is approximately zero and looks only slightly curved. We could say that linearity result is good, although not perfect.

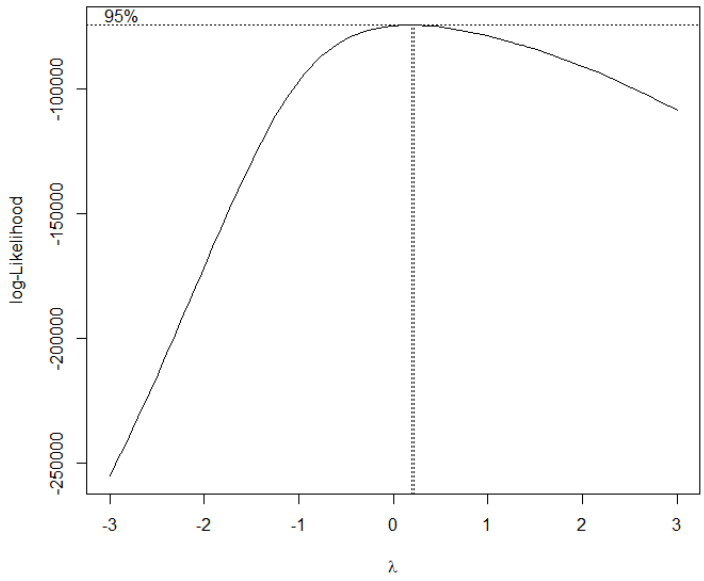
b) Normal Q-Q: normality of the errors does not hold, as we can specially see from the points at the upper right corner.

c) Scale-Location: The line is not straight, and the variability of the residual points increases with the value of the fitted outcome variable, suggesting non-constant variances in the residuals.

d) Residuals vs Leverage: Data do not present any influential points. Cook's distance lines (a red dashed line) are not shown on the Residuals vs Leverage plot because all points are within those limits.

Box Cox Transformation

Since the assumptions are not met, transformation of the dependant variable was applied. The best lambda estimation was 0.21. Here is the Maximum likelihood plot to visualize the best lambda estimation:



With the lambda value, a new model was built for the transformed dependent variable: (HRLYEARN)^0.21

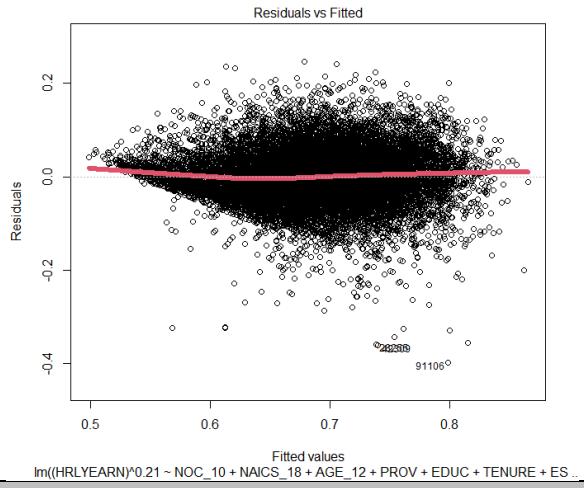
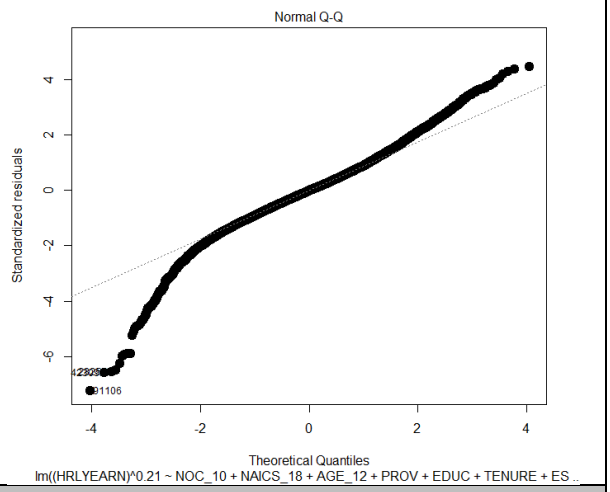
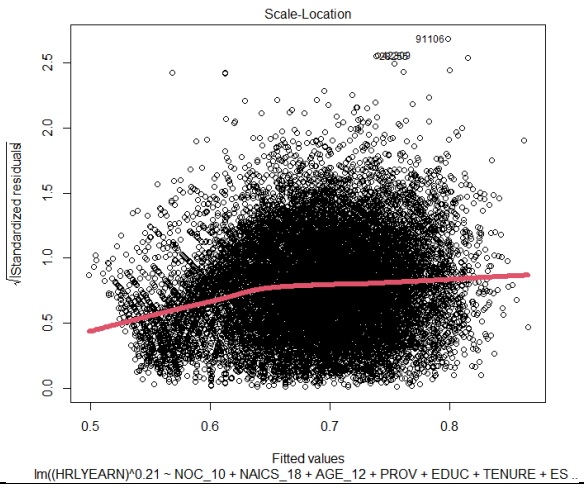
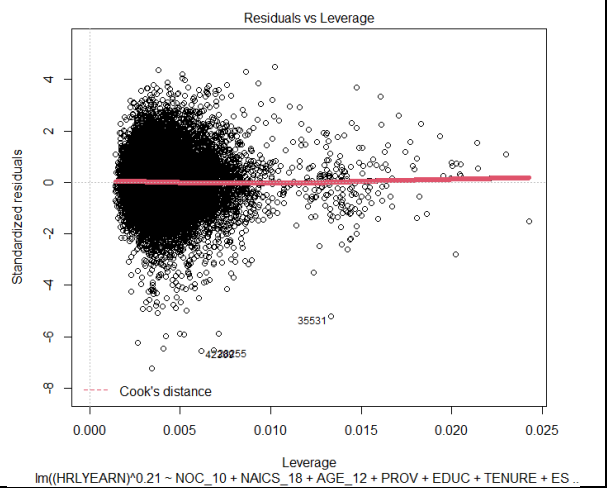
Final Model

```
Call:
lm(formula = (HRLYEARN)^0.21 ~ NOC_10 + NAICS_18 + AGE_12 + PROV +
    EDUC + TENURE + ESTSIZE + AGYOWNK + FIRMSIZE + PERMTEMP +
    FTPTMAIN + UTOTHRS + MARSTAT + UNION + LFSSTAT + MJH, data = train.09male
)
```

```
Multiple R-squared:  0.529,    Adjusted R-squared:  0.5271
F-statistic: 273.9 on 77 and 18781 DF,  p-value: < 2.2e-16
```

R-squared improved from 0.47 (First Model) to 0.53 (Final Model).

Diagnostic Plots

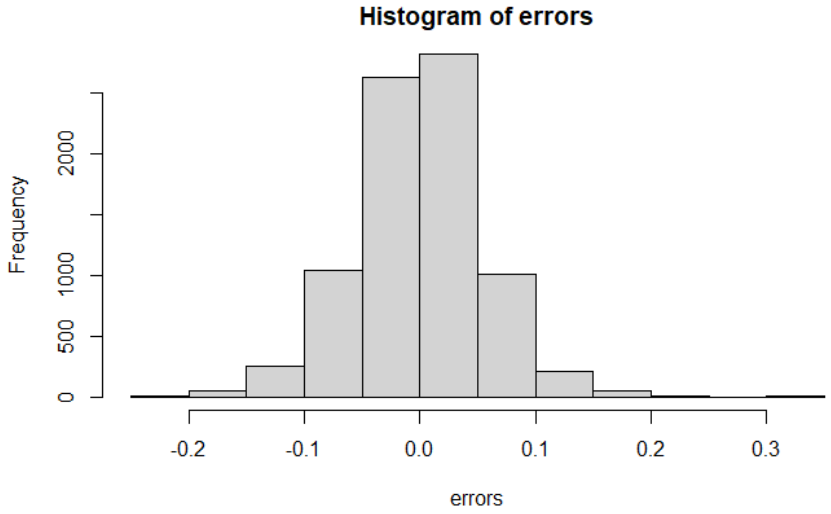
Residuals vs Fitted (Linearity Test)	Normal Q-Q (Residuals Normality Test)
	
Scale-Location (Homoscedasticity Test)	**Residuals vs. Leverage (Outliers Test)**
	

Comparing the diagnostics plots of the new model with the previous ones, we can see some improvement for linearity, normality, and homoscedasticity. The normality plot points still deviate from the dashed line on the extremes, specially as they approach the lower left corner. Nevertheless, there is an improvement in the level of skewness of the residuals:

```
> # Before Transformation:
> skewness(model$residuals)
[1] 1.105503
> # After Transformation
> skewness(new.model$residuals)
[1] -0.1655788
```

Prediction

The final model was applied to predict the hourly wages on the test data set. The next histogram shows the distribution of errors:



And here are the results for RMSE, MAE and proportion of predicted values close to real ones by 25% or less:

```
[1] "RMSE: 0.0555053839803959"
[1] "MAE: 0.0422321988304378"
[1] "Percentage of cases with less than 25% error: 99.0473833972535"
```

Variables, Coefficients and Significance

The final model equation with included variables, coefficients and significance is:

Independent Variables	Estimate	Std Error	t-value	p-value	Sig
(Intercept)	0.7177	0.0056	128.24	< 2e-16	***
NOC_10Business, finance & admin.	-0.0623	0.0020	-30.60	< 2e-16	***
NOC_10Natural & applied sciences	-0.0303	0.0020	-14.90	< 2e-16	***
NOC_10Health	-0.0205	0.0038	-5.42	0.0000	***
NOC_10Educ., law, community & gov. serv.	-0.0302	0.0027	-11.29	< 2e-16	***
NOC_10Art, culture, recreation & sport	-0.0436	0.0039	-11.20	< 2e-16	***
NOC_10Sales & service	-0.0691	0.0018	-37.41	< 2e-16	***
NOC_10Trades, transport & equip. operator	-0.0572	0.0018	-31.02	< 2e-16	***
NOC_10Natural resources & agriculture	-0.0584	0.0031	-18.95	< 2e-16	***

NOC_10Manufacturing & utilities	-0.0757	0.0024	-31.41	< 2e-16	***
NAICS_18Forestry, Fishing, Min., Oil &Gas	0.0492	0.0038	13.07	< 2e-16	***
NAICS_18Utilities	0.0531	0.0049	10.78	< 2e-16	***
NAICS_18Construction	0.0464	0.0040	11.51	< 2e-16	***
NAICS_18Manufacturing durables	0.0289	0.0041	7.05	0.0000	***
NAICS_18Manufacturing non-durables	0.0211	0.0042	5.03	0.0000	***
NAICS_18Wholesale Trade	0.0257	0.0043	6.04	0.0000	***
NAICS_18Retail Trade	-0.0023	0.0041	-0.56	0.5769	
NAICS_18Transportation & Warehousing	0.0150	0.0041	3.61	0.0003	***
NAICS_18Finance, Ins., Real Est. & Leas.	0.0268	0.0044	6.06	0.0000	***
NAICS_18Prof., Scientific & Technical Serv.	0.0456	0.0044	10.47	< 2e-16	***
NAICS_18Management, Admin. & Support	-0.0026	0.0042	-0.62	0.5343	
NAICS_18Educational Services	0.0178	0.0045	3.99	0.0001	***
NAICS_18Health Care & Social Assistance	0.0014	0.0046	0.31	0.7563	
NAICS_18Information, Culture & Recreation	0.0137	0.0043	3.18	0.0015	**
NAICS_18Accommodation & Food Services	-0.0138	0.0043	-3.20	0.0014	**
NAICS_18Other Services	0.0137	0.0044	3.13	0.0018	**
NAICS_18Public Administration	0.0438	0.0042	10.55	< 2e-16	***
AGE_12.L	0.0129	0.0034	3.80	0.0001	***
AGE_12.Q	-0.0528	0.0032	-16.45	< 2e-16	***
AGE_12.C	-0.0049	0.0028	-1.77	0.0773	.
AGE_12^4	-0.0110	0.0025	-4.32	0.0000	***
AGE_12^5	-0.0080	0.0023	-3.45	0.0006	***
AGE_12^6	0.0031	0.0021	1.47	0.1404	
AGE_12^7	-0.0042	0.0018	-2.33	0.0196	*
AGE_12^8	0.0039	0.0016	2.51	0.0122	*
AGE_12^9	0.0012	0.0014	0.83	0.4063	
AGE_12^10	0.0023	0.0013	1.80	0.0712	.
AGE_12^11	-0.0003	0.0012	-0.26	0.7972	
PROVPEI	-0.0152	0.0034	-4.45	0.0000	***
PROVNS	-0.0074	0.0029	-2.55	0.0107	*
PROVNB	-0.0069	0.0028	-2.42	0.0154	*
PROVQC	0.0078	0.0025	3.14	0.0017	**
PROVON	0.0177	0.0024	7.44	0.0000	***
PROVMB	0.0038	0.0026	1.48	0.1390	
PROVSK	0.0227	0.0027	8.39	< 2e-16	***
PROVAB	0.0393	0.0025	15.44	< 2e-16	***
PROVBC	0.0271	0.0026	10.56	< 2e-16	***
EDUC.L	0.0416	0.0020	20.34	< 2e-16	***
EDUC.Q	0.0044	0.0018	2.39	0.0170	*
EDUC.C	0.0025	0.0015	1.68	0.0929	.
EDUC^4	-0.0054	0.0013	-4.02	0.0001	***
EDUC^5	-0.0027	0.0010	-2.70	0.0069	**
EDUC^6	-0.0006	0.0011	-0.50	0.6174	
TENURE	0.0285	0.0014	20.04	< 2e-16	***
ESTSIZE.L	0.0150	0.0012	12.27	< 2e-16	***
ESTSIZE.Q	0.0026	0.0010	2.52	0.0118	*
ESTSIZE.C	0.0009	0.0009	1.02	0.3060	
AGYOWNK.L	-0.0054	0.0012	-4.38	0.0000	***
AGYOWNK.Q	-0.0021	0.0013	-1.58	0.1149	
AGYOWNK.C	-0.0007	0.0014	-0.51	0.6080	
AGYOWNK^4	-0.0035	0.0014	-2.44	0.0149	*

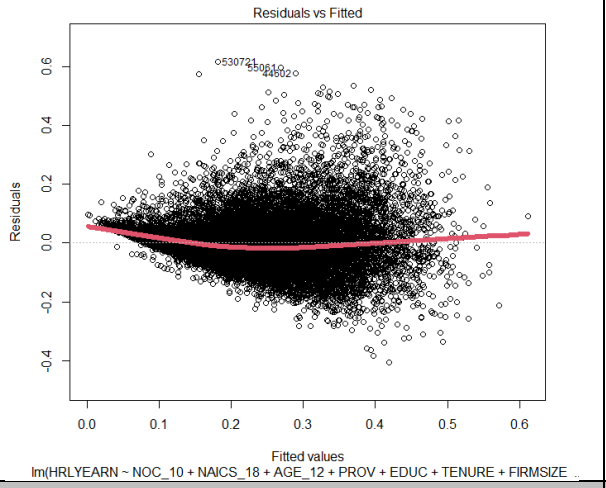
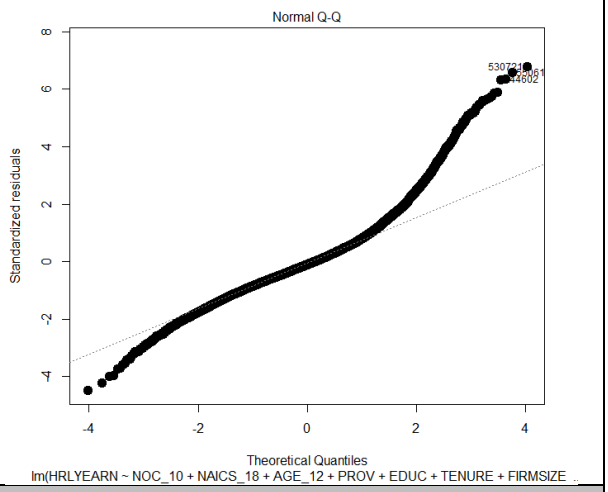
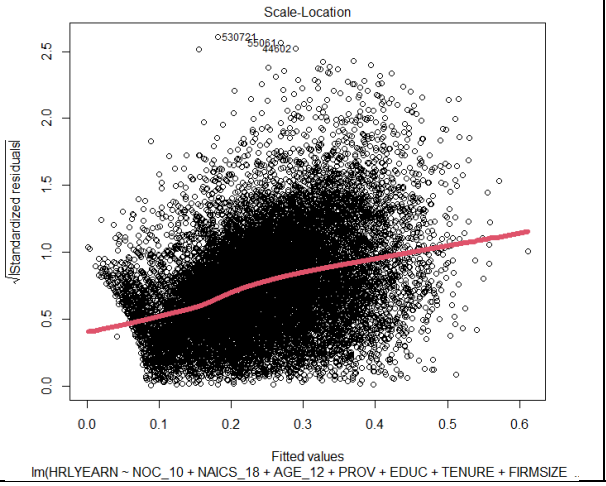
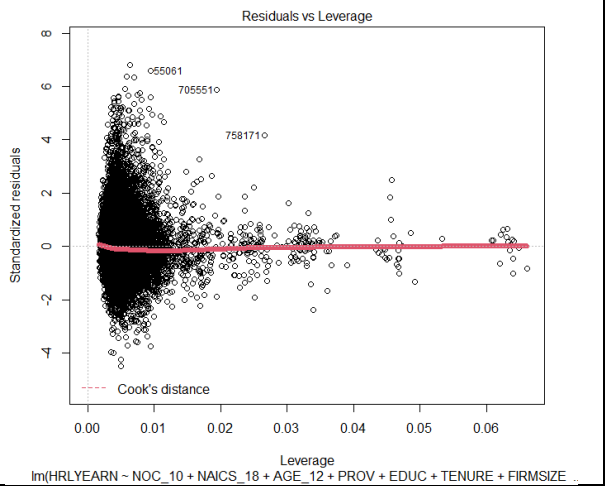
FIRMSIZE.L	0.0100	0.0011	9.04	< 2e-16	***
FIRMSIZE.Q	-0.0006	0.0011	-0.59	0.5561	
FIRMSIZE.C	0.0004	0.0011	0.36	0.7191	
PERMTEMPTemporary, seasonal	-0.0160	0.0019	-8.61	< 2e-16	***
PERMTEMPTemporary, term or contract	-0.0097	0.0018	-5.36	0.0000	***
PERMTEMPTemporary, casual or other	-0.0131	0.0027	-4.95	0.0000	***
FTPTMAINPart-time	-0.0257	0.0020	-12.71	< 2e-16	***
UTOTHRS	-0.0321	0.0056	-5.70	0.0000	***
MARSTATLiving in common-law	-0.0009	0.0013	-0.72	0.4743	
MARSTATWidowed	-0.0007	0.0058	-0.12	0.9045	
MARSTATSeparated	-0.0004	0.0027	-0.13	0.8933	
MARSTATDivorced	-0.0020	0.0023	-0.89	0.3736	
MARSTATSingle, never married	-0.0082	0.0014	-5.89	0.0000	***
UNIONNot a member but covered by a union contract	0.0004	0.0028	0.14	0.8895	
UNIONNon-unionized	-0.0108	0.0011	-10.14	< 2e-16	***
LFSSTATEmployed, absent from work	-0.0035	0.0018	-1.94	0.0525	.
MJHMultiple jobholder	-0.0029	0.0022	-1.32	0.1854	

Year 2019, Males

First Model

```
Call:
lm(formula = HRLYEARN ~ NOC_10 + NAICS_18 + AGE_12 + PROV + EDUC +
    TENURE + FIRMSIZE + MARSTAT + ESTSIZE + FTPTMAIN + UTOTHRS +
    PERMTEMP + EFAMTYPE + SCHOOLN + CMA + UNION + AGYOWNK + MJH +
    COWMAIN, data = train.1male)


Residual standard error: 0.09074 on 17749 degrees of freedom
Multiple R-squared:  0.4989,   Adjusted R-squared:  0.4961
F-statistic: 178.5 on 99 and 17749 DF,  p-value: < 2.2e-16
```

Diagnostic Plots

Residuals vs Fitted (Linearity Test)	Normal Q-Q (Residuals Normality Test)
	
Scale-Location (Homoscedasticity Test)	**Residuals vs. Leverage (Outliers Test)**
	

The conclusions of analyzing the diagnostics plots are:

e) Residuals vs Fitted: The red line is approximately zero, but looks slightly curved, specially towards the left side.

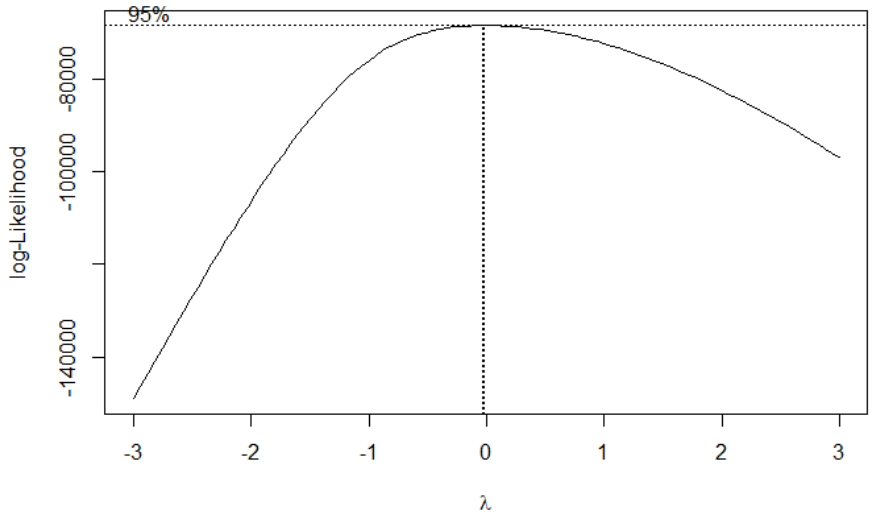
f) Normal Q-Q: normality of the errors does not hold, instead of a straight line we can see the pattern is curved, specially towards the top right corner.

g) Scale-Location: The line is not straight, and the variability of the residual points increases with the value of the fitted outcome variable, suggesting non-constant variances in the residuals.

h) Residuals vs Leverage: Data do not present any influential points. Cook's distance lines (a red dashed line) are not shown on the Residuals vs Leverage plot because all points are within those limits.

Box Cox Transformation

Since the assumptions are not met, transformation of the dependant variable was applied. The best lambda estimation was -0.03. Here is the Maximum likelihood plot to visualize the best lambda estimation:



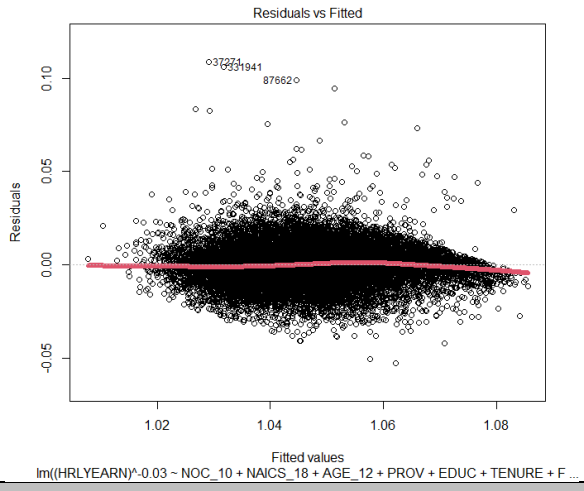
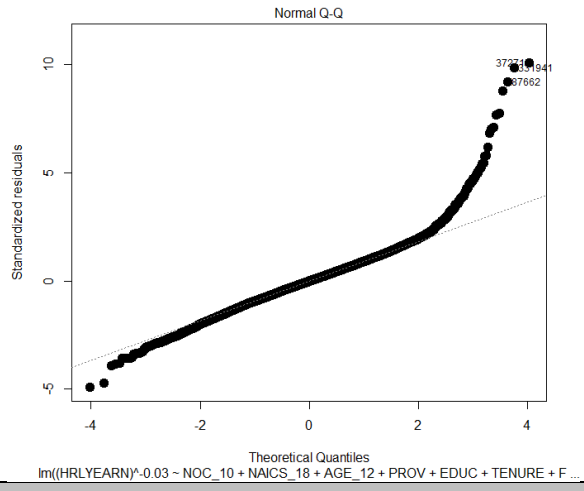
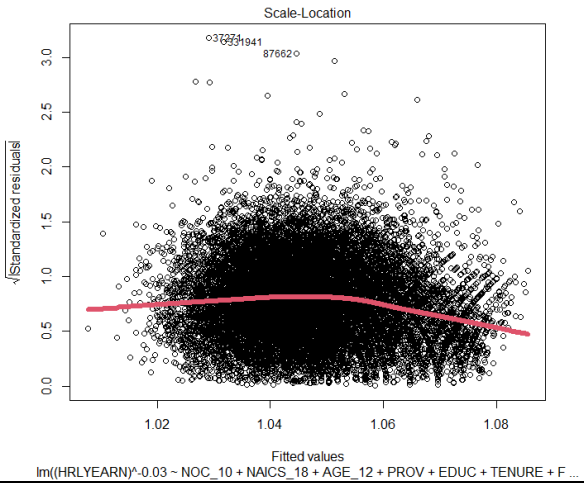
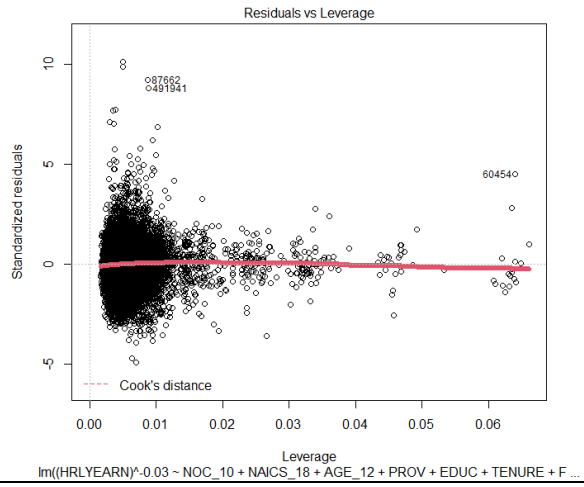
With the lambda value, a new model was built for the transformed dependent variable: (HRLYEARN)^-0.03

Final Model

```
Call:
lm(formula = (HRLYEARN)^-0.03 ~ NOC_10 + NAICS_18 + AGE_12 +
    PROV + EDUC + TENURE + FIRMSIZE + MARSTAT + ESTSIZE + FTPTMAIN +
    UTOTHRS + PERMTEMP + EFAMTYPE + SCHOOLN + CMA + UNION + AGYOWNK +
    MJH + COWMAIN, data = train.19male)

Residual standard error: 0.01079 on 17749 degrees of freedom
Multiple R-squared:  0.552,    Adjusted R-squared:  0.5495
F-statistic: 220.9 on 99 and 17749 DF,  p-value: < 2.2e-16
```

R-squared improved from 0.49 (First Model) to 0.55 (Final Model).

Diagnostic Plots

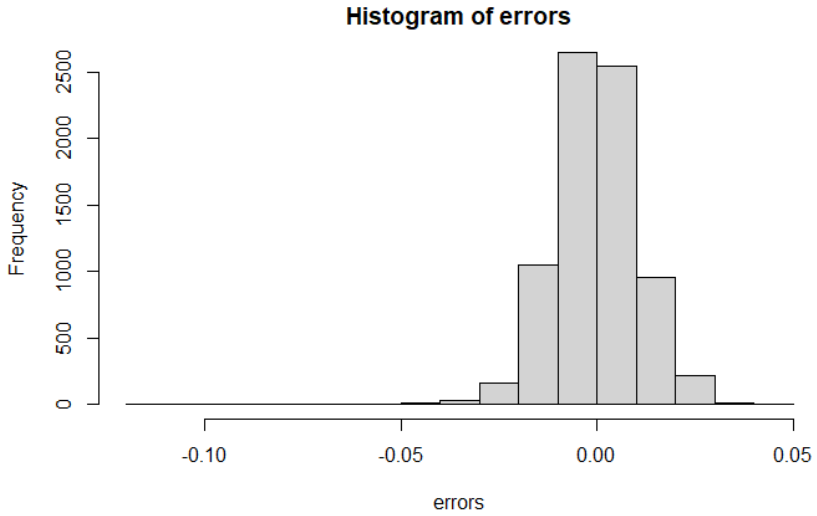
Residuals vs Fitted (Linearity Test)	Normal Q-Q (Residuals Normality Test)
	
Scale-Location (Homoscedasticity Test)	**Residuals vs. Leverage (Outliers Test)**
	

Looking at the diagnostics plots of the new model we confirm that there are some improvements for linearity, normality, and homoscedasticity. The normality plot points still deviate from the dashed line on the top right corner. However, there is an improvement in the level of skewness of the residuals:

```
> # Before Transformation:
> skewness(model$residuals)
[1] 1.032538
> # After Transformation
> skewness(new.model$residuals)
[1] 0.4713438
```

Prediction

The final model was applied to predict the hourly wages on the test data set. The next histogram shows the distribution of errors:



And here are the results for RMSE, MAE and proportion of predicted values close to real ones by 25% or less:

```
[1] "RMSE: 0.0110021049366726"
[1] "MAE: 0.00830950408353929"
[1] "Percentage of cases with less than 25% error: 100"
```

Variables, Coefficients and Significance

The final model equation with included variables, coefficients and significance is:

Independent Variables	Estimate	Std. Error	t value	Pr(>|t|)	Sig
(Intercept)	1.04E+00	1.34E-03	772.60	< 2e-16	***
NOC_10Business, finance & admin.	1.17E-02	4.30E-04	27.14	< 2e-16	***
NOC_10Natural & applied sciences	6.93E-03	4.14E-04	16.73	< 2e-16	***
NOC_10Health	6.63E-03	7.33E-04	9.04	< 2e-16	***
NOC_10Educ., law, community & gov. serv.	6.69E-03	4.90E-04	13.67	< 2e-16	***
NOC_10Art, culture, rec. & sport	1.26E-02	8.06E-04	15.65	< 2e-16	***
NOC_10Sales & service	1.68E-02	4.02E-04	41.90	< 2e-16	***
NOC_10Trades, transport & equipm. Operat.	1.22E-02	3.86E-04	31.54	< 2e-16	***
NOC_10Natural resources & agriculture	1.42E-02	6.20E-04	22.96	< 2e-16	***
NOC_10Manufacturing & utilities	1.50E-02	5.02E-04	29.83	< 2e-16	***
NAICS_18Forestry, Fish., Min., Oil, Gas	-1.24E-02	7.51E-04	-16.53	< 2e-16	***
NAICS_18Utilities	-1.31E-02	1.07E-03	-12.30	< 2e-16	***

NAICS_18Construction	-8.59E-03	8.12E-04	-10.58	< 2e-16	***
NAICS_18Manufacturing durables	-4.16E-03	8.36E-04	-4.98	6.51E-07	***
NAICS_18Manufacturing non-durables	-3.29E-03	8.59E-04	-3.83	0.00013	***
NAICS_18Wholesale Trade	-6.06E-03	8.68E-04	-6.98	3.02E-12	***
NAICS_18Retail Trade	1.04E-03	8.21E-04	1.27	0.203213	
NAICS_18Transportation & Warehousing	-2.59E-03	8.34E-04	-3.11	0.001888	**
NAICS_18Finance, Ins., Real E.,Leas.	-6.01E-03	8.76E-04	-6.86	7.07E-12	***
NAICS_18Prof., Scientific & Technical Services	-7.57E-03	8.59E-04	-8.82	< 2e-16	***
NAICS_18Management, Admin. & Support	-6.33E-04	8.49E-04	-0.75	0.456062	
NAICS_18Educational Services	-3.40E-03	9.49E-04	-3.59	0.000335	***
NAICS_18Health Care & Social Assist.	6.96E-04	9.35E-04	0.75	0.456521	
NAICS_18Information, Culture & Rec.	-2.47E-03	8.70E-04	-2.83	0.0046	**
NAICS_18Accommodation & Food Serv.	2.42E-03	8.74E-04	2.77	0.005627	**
NAICS_18Other Services	-3.38E-03	8.94E-04	-3.78	0.000155	***
NAICS_18Public Administration	-5.90E-03	9.10E-04	-6.49	9.08E-11	***
AGE_12.L	3.41E-04	6.27E-04	0.54	0.586753	
AGE_12.Q	7.59E-03	5.64E-04	13.46	< 2e-16	***
AGE_12.C	1.31E-03	4.60E-04	2.84	0.004563	**
AGE_12^4	5.06E-04	4.10E-04	1.23	0.217689	
AGE_12^5	1.66E-03	3.77E-04	4.40	1.08E-05	***
AGE_12^6	-1.54E-04	3.46E-04	-0.45	0.655704	
AGE_12^7	9.28E-04	3.13E-04	2.97	0.002991	**
AGE_12^8	3.33E-04	2.82E-04	1.18	0.237998	
AGE_12^9	5.23E-05	2.61E-04	0.20	0.84117	
AGE_12^10	-6.96E-05	2.51E-04	-0.28	0.781179	
AGE_12^11	-1.91E-04	2.49E-04	-0.77	0.4419	
PROVPEI	3.53E-03	6.79E-04	5.20	1.98E-07	***
PROVNS	2.37E-03	5.85E-04	4.05	5.17E-05	***
PROVNB	3.30E-03	5.86E-04	5.63	1.84E-08	***
PROVQC	-1.54E-04	5.10E-04	-0.30	0.763376	
PROVON	-2.43E-03	4.89E-04	-4.98	6.45E-07	***
PROVMB	-9.36E-05	5.40E-04	-0.17	0.862393	
PROVSK	-2.85E-03	5.48E-04	-5.20	2.05E-07	***
PROVAB	-7.27E-03	5.18E-04	-14.03	< 2e-16	***
PROVBC	-4.70E-03	5.34E-04	-8.79	< 2e-16	***
EDUC.L	-7.63E-03	4.28E-04	-17.84	< 2e-16	***
EDUC.Q	-1.31E-03	3.95E-04	-3.32	0.000905	***
EDUC.C	-3.13E-04	3.12E-04	-1.01	0.314844	
EDUC^4	4.85E-04	3.00E-04	1.62	0.105728	
EDUC^5	-1.21E-04	2.08E-04	-0.58	0.560434	
EDUC^6	-2.35E-04	2.68E-04	-0.88	0.379268	
TENURE	-6.34E-03	2.89E-04	-21.95	< 2e-16	***
FIRMSIZE.L	-2.47E-03	2.30E-04	-10.73	< 2e-16	***
FIRMSIZE.Q	3.82E-04	2.13E-04	1.79	0.073443	.
FIRMSIZE.C	-3.35E-04	2.16E-04	-1.55	0.120931	
MARSTATLiving in common-law	2.67E-04	2.52E-04	1.06	0.288969	
MARSTATWidowed	2.20E-03	1.20E-03	1.84	0.066055	.
MARSTATSeparated	-4.21E-04	6.67E-04	-0.63	0.527849	
MARSTATDivorced	-9.90E-04	5.78E-04	-1.71	0.086697	.
MARSTATSingle, never married	1.71E-03	4.14E-04	4.12	3.76E-05	***
ESTSIZE.L	-2.40E-03	2.47E-04	-9.75	< 2e-16	***
ESTSIZE.Q	-8.07E-04	2.04E-04	-3.95	7.75E-05	***

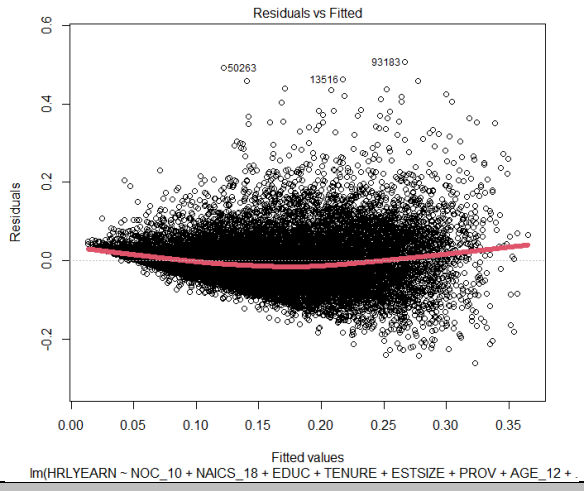
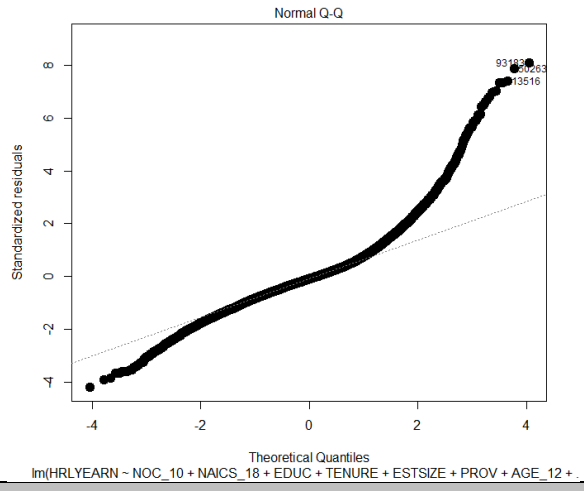
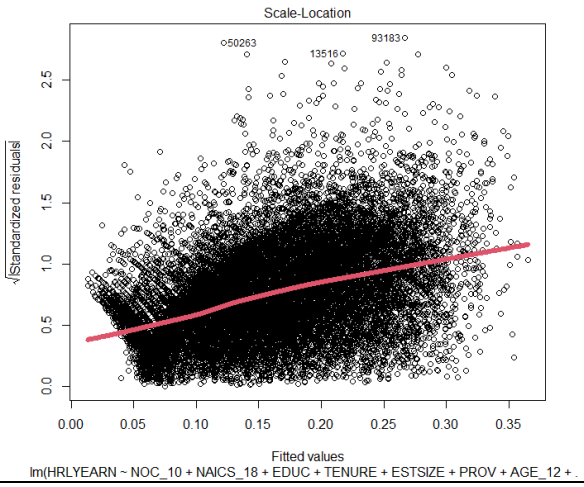
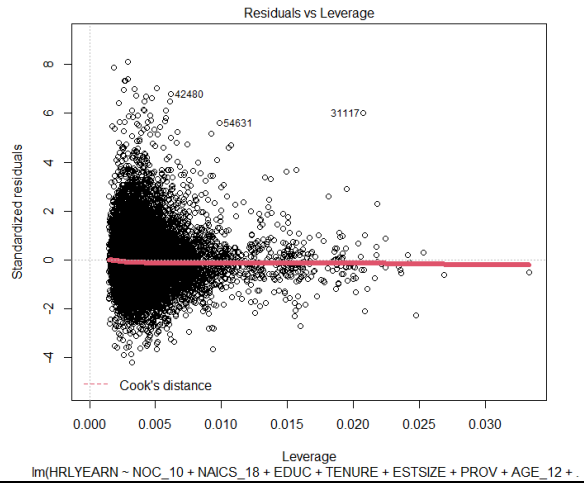
ESTSIZE.C	2.60E-04	1.81E-04	1.44	0.151296	
FTPTMAINPart-time	5.36E-03	4.21E-04	12.74	< 2e-16	***
UTOTHRS	5.10E-03	1.14E-03	4.45	8.51E-06	***
PERMTEMPTemporary, seasonal	2.89E-03	4.07E-04	7.09	1.36E-12	***
PERMTEMPTemporary, term or contract	7.42E-04	3.66E-04	2.03	0.042358	*
PERMTEMPTemporary, casual or other	2.26E-03	5.10E-04	4.44	9.10E-06	***
EFAMTYPEHWDENC	-7.46E-04	4.36E-04	-1.71	0.087013	.
EFAMTYPEHWDE17	7.81E-04	5.81E-04	1.34	0.179368	
EFAMTYPEHWDE24	1.25E-03	5.46E-04	2.30	0.021695	*
EFAMTYPEHWSHNC	-1.10E-03	5.01E-04	-2.19	0.028471	*
EFAMTYPEHWSH17	9.80E-04	6.58E-04	1.49	0.136104	
EFAMTYPEHWSH24	4.95E-04	8.67E-04	0.57	0.568034	
EFAMTYPEHWSWNC	1.69E-03	1.20E-03	1.41	0.158891	
EFAMTYPEHWSW17	1.46E-03	1.91E-03	0.77	0.443709	
EFAMTYPEHWSW24	-5.55E-04	1.50E-03	-0.37	0.710996	
EFAMTYPEHWNENC	2.46E-03	9.11E-04	2.71	0.006812	**
EFAMTYPEHWNE17	7.11E-03	2.65E-03	2.68	0.007294	**
EFAMTYPEHWNE24	-1.89E-04	2.23E-03	-0.09	0.932427	
EFAMTYPESPE17	-1.59E-04	6.98E-04	-0.23	0.820329	
EFAMTYPESPE24	1.49E-03	7.60E-04	1.96	0.049558	*
EFAMTYPESPN17	-2.65E-04	1.61E-03	-0.17	0.86908	
EFAMTYPESPN24	2.18E-03	1.86E-03	1.18	0.240192	
EFAMTYPEOther	1.73E-03	3.80E-04	4.56	5.19E-06	***
SCHOOLNFull-time student	8.40E-04	5.23E-04	1.61	0.108026	
SCHOOLNPart-time student	3.04E-03	6.86E-04	4.43	9.66E-06	***
SCHOOLNUnknown	NA	NA	NA	NA	
CMAToronto	8.38E-04	5.96E-04	1.41	0.159738	
CMAVancouver	2.16E-03	6.78E-04	3.19	0.001423	**
CMAOther	5.70E-04	4.61E-04	1.24	0.216587	
UNIONNot a member but covered by a union contract	-7.65E-04	6.06E-04	-1.26	0.206887	
UNIONNon-unionized	1.92E-03	2.26E-04	8.49	< 2e-16	***
AGYOWNK.L	1.26E-03	5.53E-04	2.28	0.022569	*
AGYOWNK.Q	1.27E-03	3.76E-04	3.38	0.000721	***
AGYOWNK.C	4.74E-04	4.22E-04	1.12	0.260905	
AGYOWNK^4	1.67E-04	3.98E-04	0.42	0.674359	
MJHMultiple jobholder	7.44E-04	4.25E-04	1.75	0.079817	.
COWMAINPrivate sector	3.34E-04	3.97E-04	0.84	0.400269	

Year 2009, Females

First Model

```
Call:
lm(formula = HRLYEARN ~ NOC_10 + NAICS_18 + EDUC + TENURE + ESTSIZE +
    PROV + AGE_12 + COWMAIN + PERMTEMP + FIRMSIZE + FTPTMAIN +
    UTOTHRS + UNION + AGYOWNK + SCHOOLN + CMA, data = train.09fem)


Residual standard error: 0.06268 on 19253 degrees of freedom
Multiple R-squared:  0.5226,   Adjusted R-squared:  0.5207
F-statistic: 277.3 on 76 and 19253 DF,  p-value: < 2.2e-16
```

Diagnostic Plots

Residuals vs Fitted (Linearity Test)	Normal Q-Q (Residuals Normality Test)
	
Scale-Location (Homoscedasticity Test)	**Residuals vs. Leverage (Outliers Test)**
	

The conclusions of analyzing the diagnostics plots are:

a) Residuals vs Fitted: The red line is approximately zero but looks curved.

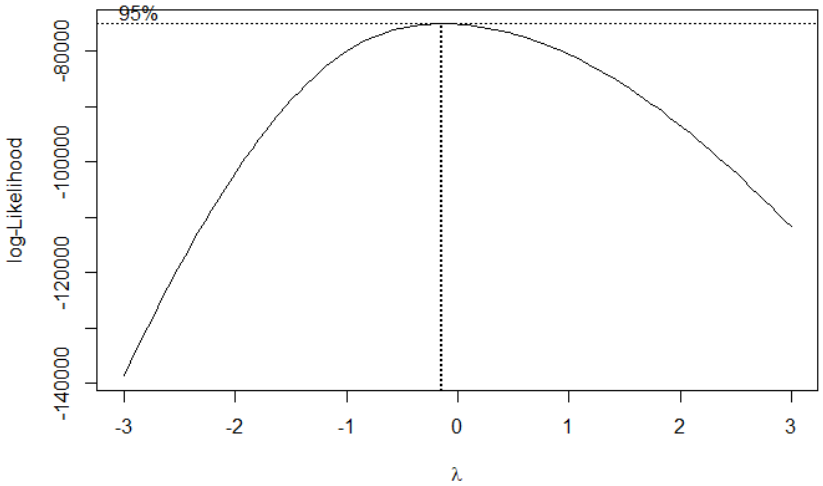
b) Normal Q-Q: normality of the errors does not hold, we can see the marks do not follow a straight diagonal line, specially towards the top right corner.

c) Scale-Location: The line is not completely horizontal, it points upward. The variability of the residual points increases with the value of the fitted outcome variable, suggesting non-constant variances in the residuals.

d) Residuals vs Leverage: Data do not present any influential points. Cook's distance lines (a red dashed line) are not shown on the Residuals vs Leverage plot because all points are within those limits.

Box Cox Transformation

Since the assumptions are not met, transformation of the dependant variable was applied. The best lambda estimation was -0.15. Here is the Maximum likelihood plot to visualize the best lambda estimation:



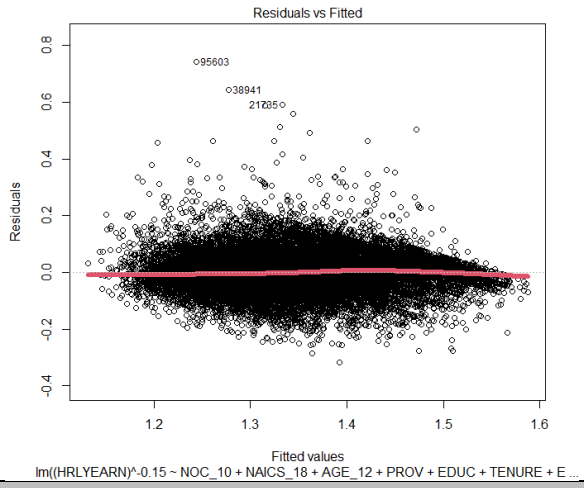
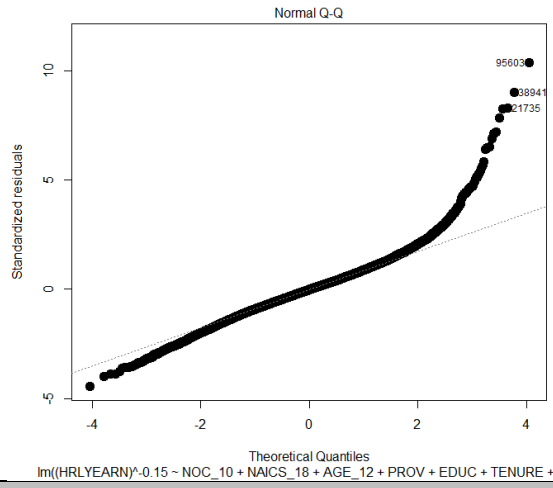
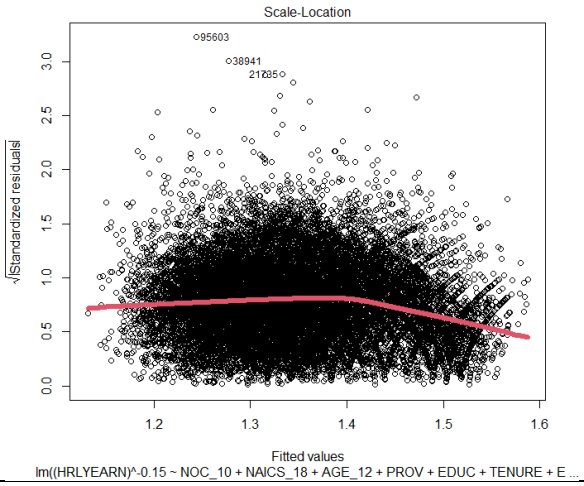
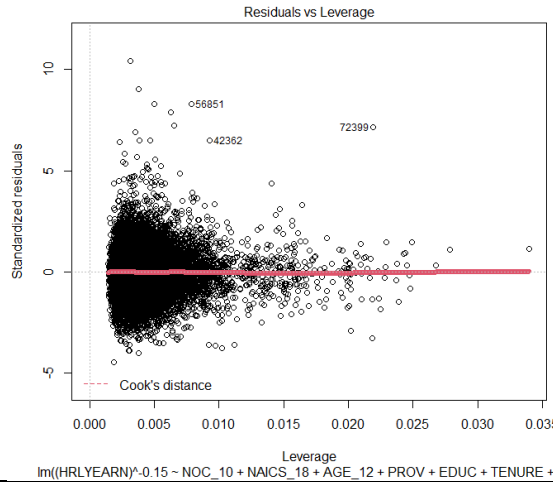
With the lambda value, a new model was built for the transformed dependent variable: (HRLYEARN)^-0.15

Final Model

```
Call:
lm(formula = (HRLYEARN)^-0.15 ~ NOC_10 + NAICS_18 + AGE_12 +
    PROV + EDUC + TENURE + ESTSIZE + AGYOWNK + FIRMSIZE + PERMTEMP +
    FTPTMAIN + UTOTHRS + MARSTAT + UNION + LFSSTAT + MJH, data = train.09fem)

Residual standard error: 0.07137 on 19252 degrees of freedom
Multiple R-squared:  0.5881,   Adjusted R-squared:  0.5864
F-statistic: 356.9 on 77 and 19252 DF,  p-value: < 2.2e-16
```

R-squared improved from 0.52 (First Model) to 0.58 (Final Model).

Diagnostic Plots

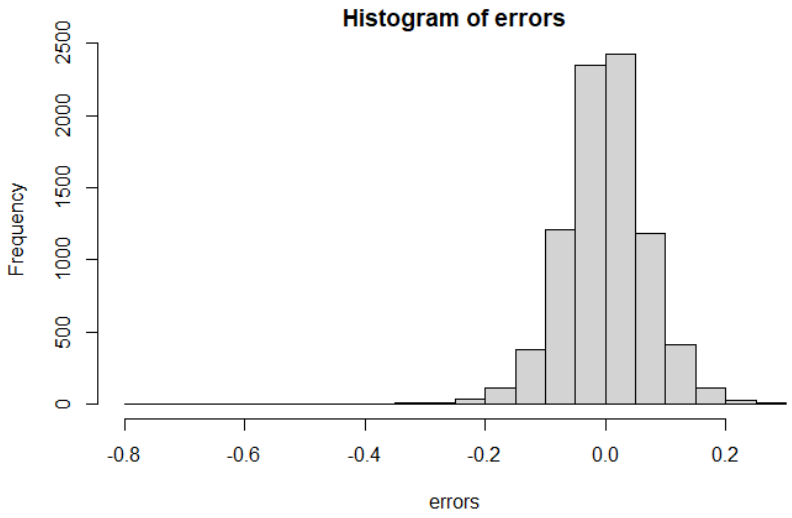
Residuals vs Fitted (Linearity Test)	Normal Q-Q (Residuals Normality Test)
	
Scale-Location (Homoscedasticity Test)	**Residuals vs. Leverage (Outliers Test)**
	

The transformation of the dependent variable helped to improve linearity, and we can see that the line on the plot is now fairly straight. There are some improvements for normality, and homoscedasticity. The normality plot points still deviate from the dashed line on the extremes, specially as they approach the top right corner. Still, there is an improvement in the level of skewness of the residuals:

```
> # Before Transformation:
> skewness(model$residuals)
[1] 1.181981
> # After Transformation
> skewness(new.model$residuals)
[1] 0.5286001
```

Prediction

The final model was applied to predict the hourly wages on the test data set. The next histogram shows the distribution of errors:



And here are the results for RMSE, MAE and proportion of predicted values close to real ones by 25% or less:

```
[1] "RMSE: 0.0727054261959815"
[1] "MAE: 0.0534999014295685"
[1] "Percentage of cases with less than 25% error: 99.8430709802028"
```

Variables, Coefficients and Significance

The final model equation with included variables, coefficients and significance is:

Independent Variables	Estimate	Std. Error	t value	Pr(>|t|)	Sig
(Intercept)	1.3541	0.0099	137.05	< 2e-16	***
NOC_10Business, finance & administration	0.0654	0.0025	26.19	< 2e-16	***
NOC_10Natural & applied sciences	0.0336	0.0040	8.46	< 2e-16	***
NOC_10Health	0.0159	0.0030	5.32	1.07E-07	***
NOC_10Educ., law, community & gov. serv.	0.0313	0.0029	10.98	< 2e-16	***
NOC_10Art, culture, recreation & sport	0.0488	0.0044	11.09	< 2e-16	***
NOC_10Sales & service	0.0921	0.0026	35.67	< 2e-16	***
NOC_10Trades, transport & equip. operator	0.0885	0.0046	19.38	< 2e-16	***
NOC_10Natural resources & agriculture	0.0730	0.0079	9.29	< 2e-16	***
NOC_10Manufacturing & utilities	0.1105	0.0046	24.19	< 2e-16	***
NAICS_18Forestry, Fishing, Min., Oil & Gas	-0.0910	0.0093	-9.78	< 2e-16	***
NAICS_18Utilities	-0.0805	0.0110	-7.31	2.70E-13	***
NAICS_18Construction	-0.0625	0.0094	-6.68	2.45E-11	***
NAICS_18Manufacturing durables	-0.0694	0.0090	-7.73	1.14E-14	***
NAICS_18Manufacturing non-durables	-0.0402	0.0088	-4.56	5.19E-06	***

NAICS_18Wholesale Trade	-0.0547	0.0090	-6.06	1.41E-09	***
NAICS_18Retail Trade	-0.0022	0.0083	-0.26	0.79342	
NAICS_18Transportation & Warehousing	-0.0493	0.0089	-5.56	2.68E-08	***
NAICS_18Finance, Ins., Real E., Leas.	-0.0585	0.0084	-6.94	4.17E-12	***
NAICS_18Prof., Scientific & Technical Services	-0.0630	0.0086	-7.33	2.35E-13	***
NAICS_18Management, Admin. & Support	-0.0182	0.0086	-2.11	0.03463	*
NAICS_18Educational Services	-0.0580	0.0085	-6.84	8.12E-12	***
NAICS_18Health Care & Social Assistance	-0.0365	0.0084	-4.37	1.23E-05	***
NAICS_18Information, Culture & Recreation	-0.0384	0.0086	-4.49	7.18E-06	***
NAICS_18Accommodation & Food Services	-0.0055	0.0084	-0.65	0.51345	
NAICS_18Other Services	-0.0396	0.0086	-4.62	3.88E-06	***
NAICS_18Public Administration	-0.0780	0.0085	-9.22	< 2e-16	***
AGE_12.L	-0.0126	0.0055	-2.30	0.02175	*
AGE_12.Q	0.0483	0.0054	8.91	< 2e-16	***
AGE_12.C	-0.0123	0.0048	-2.57	0.01008	*
AGE_12^4	0.0019	0.0042	0.45	0.65124	
AGE_12^5	0.0069	0.0036	1.89	0.05873	.
AGE_12^6	-0.0073	0.0031	-2.34	0.01934	*
AGE_12^7	0.0065	0.0026	2.49	0.01274	*
AGE_12^8	-0.0035	0.0022	-1.64	0.10203	
AGE_12^9	-0.0006	0.0018	-0.34	0.73106	
AGE_12^10	-0.0007	0.0016	-0.45	0.65218	
AGE_12^11	0.0017	0.0015	1.17	0.2421	
PROVPEI	0.0011	0.0042	0.26	0.79482	
PROVNS	-0.0051	0.0036	-1.42	0.1562	
PROVNB	0.0004	0.0036	0.10	0.91714	
PROVQC	-0.0209	0.0031	-6.75	1.54E-11	***
PROVON	-0.0333	0.0030	-11.18	< 2e-16	***
PROVMB	-0.0183	0.0033	-5.56	2.80E-08	***
PROVSK	-0.0340	0.0034	-9.98	< 2e-16	***
PROVAB	-0.0603	0.0033	-18.49	< 2e-16	***
PROVBC	-0.0410	0.0032	-12.77	< 2e-16	***
EDUC.L	-0.0686	0.0030	-23.21	< 2e-16	***
EDUC.Q	-0.0140	0.0027	-5.27	1.37E-07	***
EDUC.C	-0.0056	0.0021	-2.64	0.00828	**
EDUC^4	0.0047	0.0018	2.53	0.01147	*
EDUC^5	0.0071	0.0013	5.34	9.26E-08	***
EDUC^6	0.0029	0.0015	1.95	0.05069	.
TENURE	-0.0496	0.0019	-26.27	< 2e-16	***
ESTSIZE.L	-0.0251	0.0015	-16.44	< 2e-16	***
ESTSIZE.Q	-0.0007	0.0013	-0.53	0.59692	
ESTSIZE.C	0.0003	0.0012	0.24	0.81171	
AGYOWNK.L	0.0017	0.0016	1.08	0.27855	
AGYOWNK.Q	-0.0025	0.0016	-1.51	0.13118	
AGYOWNK.C	0.0002	0.0017	0.13	0.89598	
AGYOWNK^4	-0.0029	0.0017	-1.72	0.08529	.
FIRMSIZE.L	-0.0108	0.0014	-7.97	1.66E-15	***
FIRMSIZE.Q	0.0009	0.0014	0.65	0.51598	
FIRMSIZE.C	-0.0021	0.0015	-1.47	0.14102	
PERMTEMPTemporary, seasonal	0.0178	0.0033	5.43	5.72E-08	***
PERMTEMPTemporary, term or contract	0.0089	0.0022	4.13	3.68E-05	***
PERMTEMPTemporary, casual or other	0.0115	0.0026	4.38	1.22E-05	***

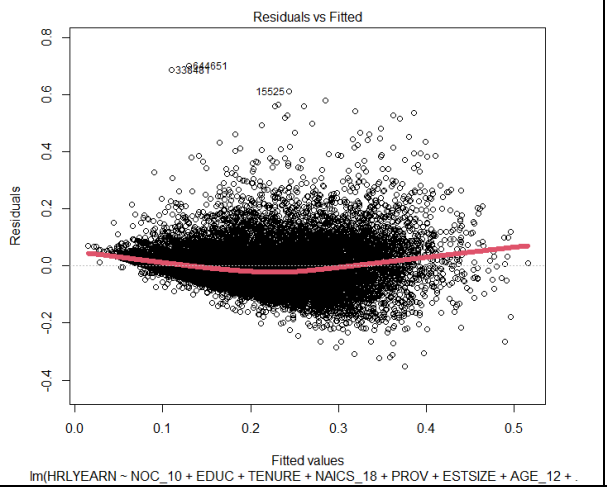
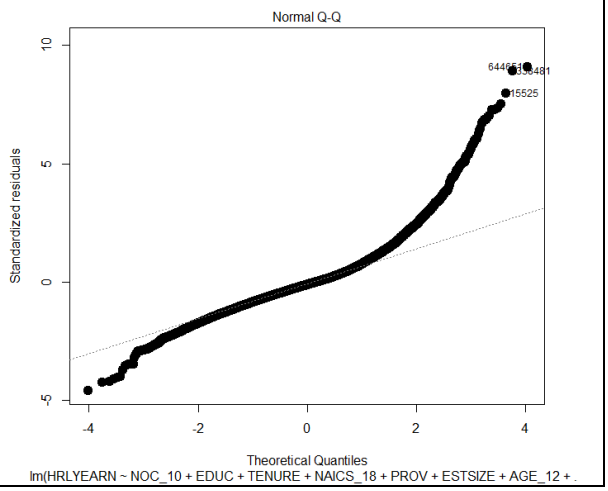
FTPTMAINPart-time	0.0187	0.0022	8.53	< 2e-16	***
UTOTHRS	0.0203	0.0091	2.24	0.02544	*
MARSTATLiving in common-law	0.0011	0.0017	0.62	0.53455	
MARSTATWidowed	0.0092	0.0042	2.21	0.02689	*
MARSTATSeparated	0.0012	0.0029	0.43	0.66826	
MARSTATDivorced	0.0003	0.0022	0.12	0.90279	
MARSTATSingle, never married	0.0035	0.0017	2.08	0.03718	*
UNIONNot a member but covered by a union contract	-0.0002	0.0039	-0.04	0.96695	
UNIONNon-unionized	0.0208	0.0015	14.09	< 2e-16	***
LFSSTATEmployed, absent from work	0.0022	0.0019	1.14	0.25286	
MJHMultiple jobholder	0.0003	0.0024	0.11	0.91514	

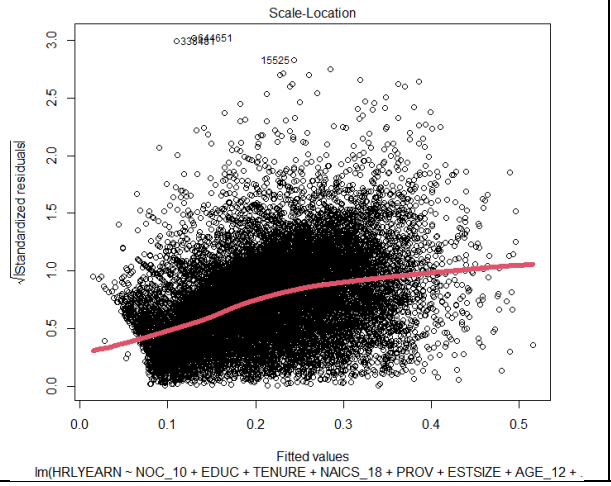
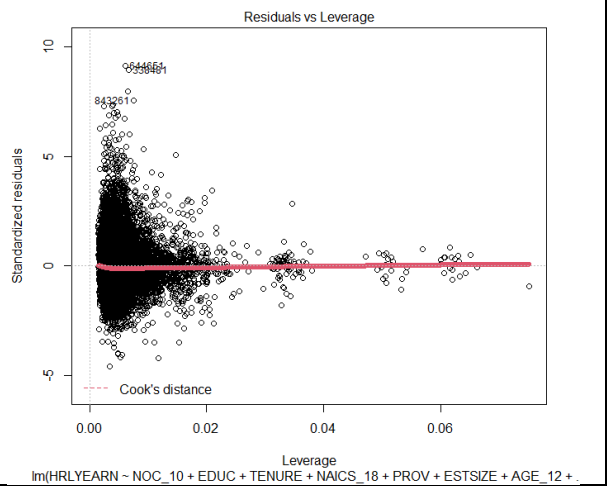
Year 2019, Females

First Model

```
Call:
lm(formula = HRLYEARN ~ NOC_10 + EDUC + TENURE + NAICS_18 + PROV +
    ESTSIZE + AGE_12 + COWMAIN + PERMTEMP + FIRMSIZE + FTPTMAIN +
    UTOTHRS + EFAMTYPE + LFSSTAT + SCHOOLN + MARSTAT, data = train.19fem)


Residual standard error: 0.077 on 17711 degrees of freedom
Multiple R-squared:  0.516,    Adjusted R-squared:  0.5135
F-statistic: 209.8 on 90 and 17711 DF,  p-value: < 2.2e-16
```

Diagnostic Plots

Residuals vs Fitted (Linearity Test)	Normal Q-Q (Residuals Normality Test)
	

Scale-Location (Homoscedasticity Test)	Residuals vs. Leverage (Outliers Test)
	

The conclusions of analyzing the diagnostics plots are:

a) Residuals vs Fitted: The red line is approximately zero, although is curved to the sides, which is not favorable for the linearity assumption.

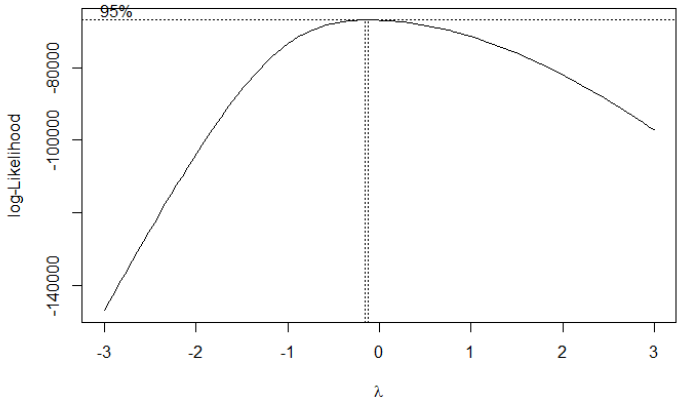
b) Normal Q-Q: the errors do not seem to be normally distributed. The marks do not follow a straight line.

c) Scale-Location: The line is not completely horizontal, it points upward, which implies non-constant variances in the residuals.

d) Residuals vs Leverage: Data do not present any influential points.

Box Cox Transformation

Since the assumptions are not met, transformation of the dependant variable was applied. The best lambda estimation was -0.15. Here is the Maximum likelihood plot to visualize the best lambda estimation:



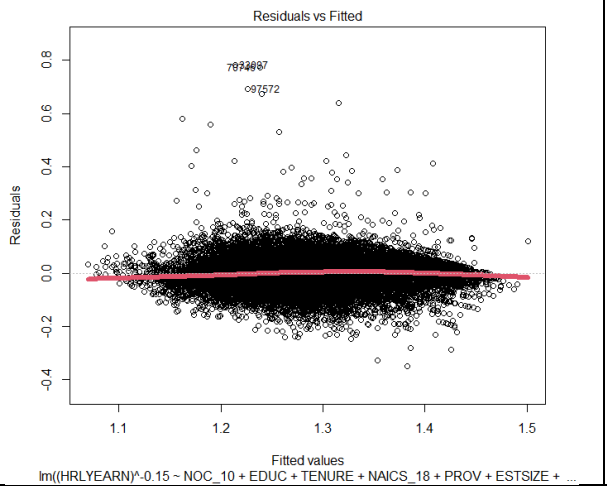
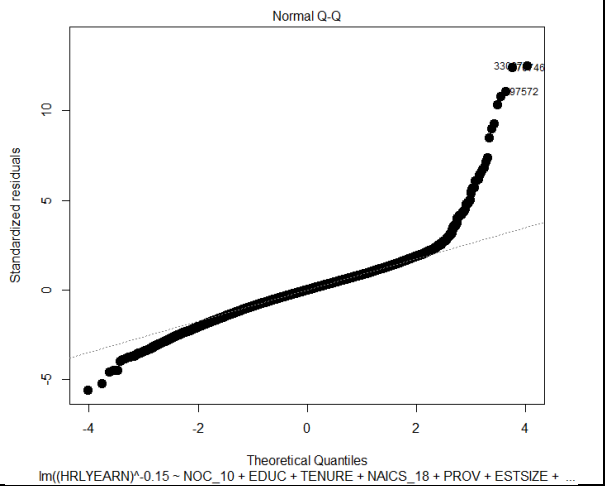
With the lambda value, a new model was built for the transformed dependent variable: (HRLYEARN)^-0.15

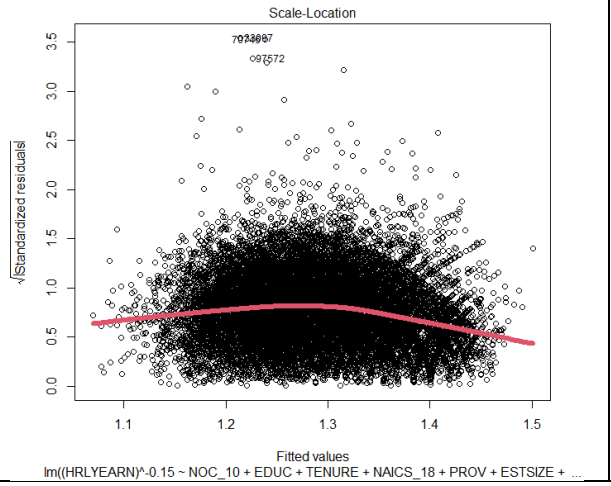
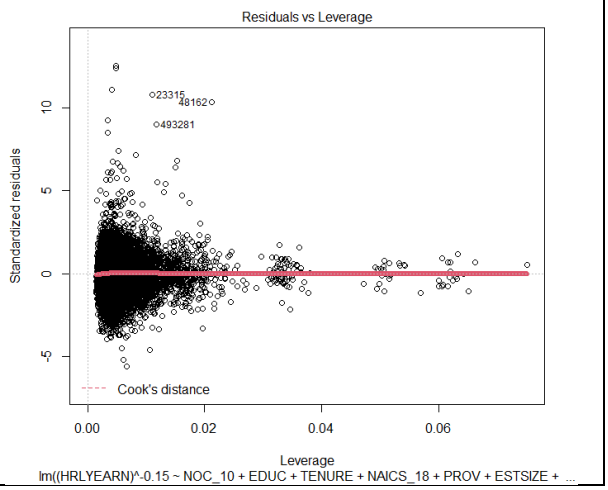
Final Model

```
Call:
lm(formula = (HRLYEARN)^-0.15 ~ NOC_10 + EDUC + TENURE + NAICS_18 +
    PROV + ESTSIZE + AGE_12 + COWMAIN + PERMTEMP + FIRMSIZE +
    FTPTMAIN + UTOTHRS + EFAMTYPE + LFSSTAT + SCHOOLN + MARSTAT,
    data = train.19fem)

Residual standard error: 0.06251 on 17711 degrees of freedom
Multiple R-squared:  0.5654,  Adjusted R-squared:  0.5632
F-statistic: 256.1 on 90 and 17711 DF,  p-value: < 2.2e-16
```

R-squared improved from 0.51 (First Model) to 0.56 (Final Model).

Diagnostic Plots

Residuals vs Fitted (Linearity Test)	Normal Q-Q (Residuals Normality Test)
	

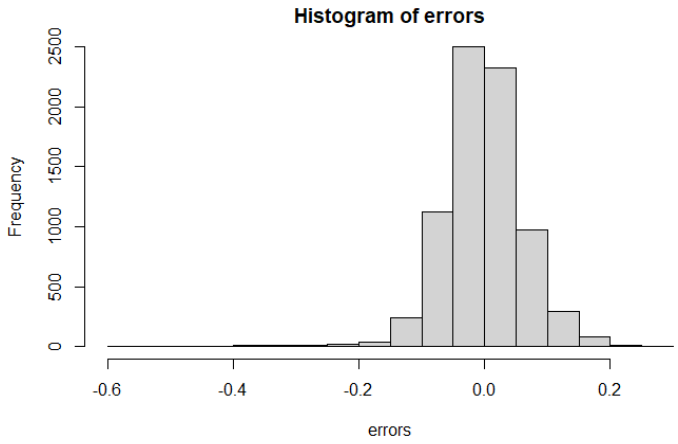
Scale-Location (Homoscedasticity Test)	Residuals vs. Leverage (Outliers Test)
	

Looking at the first plot, we can see that linearity has improved substantially. Normality and homoscedasticity have improved as well, though not completely. The QQ plot still shows that the marks are not following a straight line on the extremes. Nevertheless, there is an improvement in the level of skewness of the residuals:

```
> # Before Transformation:
> skewness(model$residuals)
[1] 1.290823
> # After Transformation
> skewness(new.model$residuals)
[1] 0.7025932
```

Prediction

The final model was applied to predict the hourly wages on the test data set. The next histogram shows the distribution of errors:



And here are the results for RMSE, MAE and proportion of predicted values close to real ones by 25% or less:

```
[1] "RMSE: 0.06424580615293 13"
[1] "MAE: 0.0470490092790988"
[1] "Percentage of cases with less than 25% error: 99.8951507208388"
```

Variables, Coefficients and Significance

The final model equation with included variables, coefficients and significance is:

Independent Variables	Estimate	Std. Error	t value	Pr(>|t|)	Sig
(Intercept)	1.2645	0.0097	130.38	< 2e-16	***
NOC_10Business, finance & administration	0.0632	0.0024	26.39	< 2e-16	***
NOC_10Natural & applied sciences	0.0278	0.0035	7.97	1.63E-15	***
NOC_10Health	0.0266	0.0028	9.61	< 2e-16	***
NOC_10Educ., law, community & gov. serv.	0.0488	0.0026	18.81	< 2e-16	***
NOC_10Art, culture, recreation & sport	0.0663	0.0042	15.79	< 2e-16	***
NOC_10Sales & service	0.0944	0.0025	37.70	< 2e-16	***
NOC_10Trades, transport & equip. operator	0.0732	0.0041	17.81	< 2e-16	***
NOC_10Natural resources & agriculture	0.0629	0.0074	8.50	< 2e-16	***
NOC_10Manufacturing & utilities	0.1068	0.0045	23.96	< 2e-16	***
EDUC.L	-0.0658	0.0032	-20.37	< 2e-16	***
EDUC.Q	-0.0156	0.0030	-5.22	1.80E-07	***
EDUC.C	-0.0143	0.0024	-6.06	1.41E-09	***
EDUC^4	0.0082	0.0020	4.10	4.19E-05	***
EDUC^5	0.0049	0.0014	3.63	0.000283	***
EDUC^6	0.0005	0.0016	0.30	0.761548	
TENURE	-0.0512	0.0017	-29.88	< 2e-16	***
NAICS_18Forestry, Fishing, Min., Oil & Gas	-0.0990	0.0085	-11.67	< 2e-16	***
NAICS_18Utilities	-0.0949	0.0107	-8.84	< 2e-16	***
NAICS_18Construction	-0.0598	0.0089	-6.75	1.56E-11	***
NAICS_18Manufacturing durables	-0.0649	0.0089	-7.27	3.84E-13	***
NAICS_18Manufacturing non-durables	-0.0492	0.0088	-5.60	2.14E-08	***
NAICS_18Wholesale Trade	-0.0645	0.0089	-7.24	4.61E-13	***
NAICS_18Retail Trade	-0.0187	0.0083	-2.24	0.025049	*
NAICS_18Transportation & Warehousing	-0.0442	0.0087	-5.06	4.20E-07	***
NAICS_18Finance, Ins., Real E. & Leas.	-0.0696	0.0084	-8.27	< 2e-16	***
NAICS_18Prof., Scientific & Technical Services	-0.0635	0.0085	-7.48	7.75E-14	***
NAICS_18Management, Admin. & Support	-0.0341	0.0085	-4.00	6.39E-05	***
NAICS_18Educational Services	-0.0446	0.0085	-5.23	1.68E-07	***
NAICS_18Health Care & Social Assistance	-0.0312	0.0083	-3.74	0.000183	***
NAICS_18Information, Culture & Recreation	-0.0365	0.0086	-4.24	2.29E-05	***
NAICS_18Accommodation & Food Services	-0.0229	0.0084	-2.72	0.006637	**
NAICS_18Other Services	-0.0395	0.0086	-4.61	3.98E-06	***
NAICS_18Public Administration	-0.0600	0.0085	-7.06	1.68E-12	***
PROVPEI	0.0082	0.0038	2.17	0.02989	*
PROVNS	0.0202	0.0033	6.11	1.01E-09	***
PROVNB	0.0172	0.0034	5.10	3.47E-07	***
PROVQC	-0.0040	0.0029	-1.39	0.164415	

PROVON	-0.0147	0.0028	-5.29	1.26E-07	***
PROVMB	0.0067	0.0031	2.17	0.03025	*
PROVSK	-0.0067	0.0031	-2.13	0.033308	*
PROVAB	-0.0299	0.0030	-9.88	< 2e-16	***
PROVBC	-0.0184	0.0029	-6.26	3.95E-10	***
ESTSIZE.L	-0.0142	0.0014	-10.34	< 2e-16	***
ESTSIZE.Q	-0.0019	0.0011	-1.62	0.106111	
ESTSIZE.C	0.0015	0.0011	1.40	0.162469	
AGE_12.L	0.0112	0.0039	2.87	0.004056	**
AGE_12.Q	0.0348	0.0034	10.12	< 2e-16	***
AGE_12.C	0.0090	0.0029	3.06	0.002238	**
AGE_12^4	-0.0003	0.0026	-0.11	0.910013	
AGE_12^5	0.0083	0.0024	3.50	0.000475	***
AGE_12^6	0.0003	0.0021	0.13	0.893695	
AGE_12^7	0.0052	0.0019	2.78	0.005525	**
AGE_12^8	0.0021	0.0017	1.26	0.206423	
AGE_12^9	-0.0009	0.0015	-0.61	0.540125	
AGE_12^10	-0.0003	0.0014	-0.18	0.858006	
AGE_12^11	-0.0011	0.0014	-0.76	0.448722	
COWMAINPrivate sector	0.0276	0.0017	15.94	< 2e-16	***
PERMTEMPTemporary, seasonal	0.0185	0.0032	5.82	6.04E-09	***
PERMTEMPTemporary, term or contract	0.0085	0.0020	4.34	1.41E-05	***
PERMTEMPTemporary, casual or other	0.0160	0.0024	6.80	1.07E-11	***
FIRMSIZE.L	-0.0099	0.0013	-7.68	1.69E-14	***
FIRMSIZE.Q	0.0007	0.0012	0.55	0.582668	
FIRMSIZE.C	-0.0005	0.0013	-0.39	0.693809	
FTPTMAINPart-time	0.0194	0.0018	10.72	< 2e-16	***
UTOTHRS	0.0367	0.0072	5.11	3.31E-07	***
EFAMTYPEHWDENC	0.0009	0.0024	0.36	0.716525	
EFAMTYPEHWDE17	0.0006	0.0024	0.25	0.806441	
EFAMTYPEHWDE24	0.0041	0.0026	1.58	0.114401	
EFAMTYPEHWSHNC	0.0085	0.0071	1.20	0.229702	
EFAMTYPEHWSH17	0.0102	0.0062	1.65	0.099388	.
EFAMTYPEHWSH24	-0.0132	0.0083	-1.58	0.113947	
EFAMTYPEHWSWNC	0.0048	0.0030	1.61	0.106624	
EFAMTYPEHWSW17	0.0107	0.0041	2.61	0.009197	**
EFAMTYPEHWSW24	0.0144	0.0055	2.60	0.009464	**
EFAMTYPEHWNENC	0.0131	0.0059	2.21	0.027448	*
EFAMTYPEHWNE17	0.0186	0.0153	1.21	0.225868	
EFAMTYPEHWNE24	0.0038	0.0138	0.28	0.780926	
EFAMTYPESPE17	0.0059	0.0023	2.59	0.009608	**
EFAMTYPESPE24	0.0087	0.0035	2.48	0.01334	*
EFAMTYPESPN17	0.0055	0.0109	0.51	0.611402	
EFAMTYPESPN24	0.0134	0.0110	1.21	0.22626	
EFAMTYPEOther	0.0083	0.0022	3.74	0.000186	***
LFSSTATEmployed, absent from work	0.0029	0.0017	1.72	0.086171	.
SCHOOLNFull-time student	0.0087	0.0025	3.40	0.000685	***
SCHOOLNPart-time student	0.0011	0.0030	0.37	0.710844	
SCHOOLNUnknown	NA	NA	NA	NA	
MARSTATLiving in common-law	-0.0009	0.0015	-0.61	0.539187	
MARSTATWidowed	0.0030	0.0043	0.69	0.492362	
MARSTATSeparated	-0.0093	0.0033	-2.79	0.005228	**
MARSTATDivorced	0.0003	0.0029	0.12	0.905203	

MARSTATSingle, never married	0.0027	0.0022	1.23	0.219413	

Conclusions

Calls Comparison

Here are the formulas to create the model of each group:

Year 2009, Males	Year 2019, Males	Year 2009, Females	Year 2019, Females
lm(formula = (HRLYEARN)^0.21 ~ NOC_10 + NAICS_18 + AGE_12 + PROV + EDUC + TENURE + ESTSIZE + AGYOWNK + FIRMSIZE + PERMTEMP + FTPTMAIN + UTOTHRS + MARSTAT + UNION + LFSSTAT + MJH , data = train.09male)	lm(formula = (HRLYEARN)^-0.03 ~ NOC_10 + NAICS_18 + AGE_12 + PROV + EDUC + TENURE + FIRMSIZE + MARSTAT + ESTSIZE + FTPTMAIN + UTOTHRS + PERMTEMP + EFAMTYPE + SCHOOLN + CMA + UNION + AGYOWNK + MJH + COWMAIN , data=train.19male)	lm(formula = (HRLYEARN)^-0.15 ~ NOC_10 + NAICS_18 + AGE_12 + PROV + EDUC + TENURE + ESTSIZE + AGYOWNK + FIRMSIZE + PERMTEMP + FTPTMAIN + UTOTHRS + MARSTAT + UNION + LFSSTAT + MJH , data = train.09fem)	lm(formula = (HRLYEARN)^-0.15 ~ NOC_10 + EDUC + TENURE + NAICS_18 + PROV + ESTSIZE + AGE_12 + COWMAIN + PERMTEMP + FIRMSIZE + FTPTMAIN + UTOTHRS + EFAMTYPE + LFSSTAT + SCHOOLN + MARSTAT , data = train.19fem)

If we compare the four different calls, we can see that:

- Occupation (NOC_10), industry (NAICS_18), age (AGE_12), and province (PROV) are the variables that appear to have more impact
- Education (EDUC) and tenure are also important, especially for women in 2019.

Independent Variables Coefficients

The dependant variable was transformed to improve the models for all four groups. In all cases, the transformation required to power the hourly wages to a fraction, meaning that the value changed to a root of the original one. Also, is important to notice that only the model for the Males 2009 population employed a positive exponent, while the rest of the models used negative exponents. This has an effect in the way to interpret the sign of the coefficients.

The sign of a regression coefficient tells whether there is a positive or negative correlation between each independent variable the dependent variable. Usually, a positive coefficient indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase. However, when the dependant variable is transformed applying a negative power, the relationship with the coefficients is inversed. That is, a coefficient with a

positive sign decreases the value of the dependant variable, and a coefficient with a negative sign increases the value of the dependant variable.

Next, we can see the list of top 10 variable coefficients that help to increment or reduce hourly wages. Only variables that are significant according to the t test are included. Besides, different colors were applied to cells to help identify variables and easily distinguish if they appear in several models and compare their position.

Top variable coefficients that increase Hourly Wages

Year 2009, Males		Year 2019, Males		Year 2009, Females		Year 2019, Females	
Ind. Var.	Est.	Ind. Var.	Est.	Ind. Var.	Est.	Ind. Var.	Est.
NAICS_18Utils	0.053	NAICS_18Utils	-0.013	NAICS_18Fores	-0.091	NAICS_18Fores	-0.099
NAICS_18Fores	0.049	NAICS_18Fores	-0.012	NAICS_18Utils	-0.081	NAICS_18Utils	-0.095
NAICS_18Const	0.046	NAICS_18Const	-0.009	NAICS_18PubAd	-0.078	NAICS_18Finan	-0.070
NAICS_18ProSc	0.046	EDUC.L	-0.008	NAICS_18ManuD	-0.069	EDUC.L	-0.066
NAICS_18PubAd	0.044	NAICS_18ProSc	-0.008	EDUC.L	-0.069	NAICS_18ManuD	-0.065
EDUC.L	0.042	PROVAB	-0.007	NAICS_18ProSc	-0.063	NAICS_18Whole	-0.064
PROVAB	0.039	TENURE	-0.006	NAICS_18Const	-0.062	NAICS_18ProSc	-0.064
NAICS_18ManuD	0.029	NAICS_18Whole	-0.006	PROVAB	-0.060	NAICS_18PubAd	-0.060
TENURE	0.028	NAICS_18Finan	-0.006	NAICS_18Finan	-0.059	NAICS_18Const	-0.060
PROVBC	0.027	NAICS_18PubAd	-0.006	NAICS_18Educa	-0.058	TENURE	-0.051

From the list above, we can see that:

- Overall, industries provide the biggest increments across all four groups:
  - “Utilities” and “Forestry, fishing, mining, oil and gas” are the top industries that increase hourly wages for both men and women.
  - “Construction” is also relevant, mainly for men.
  - “Finance, Insurance, Real estate and Leasing” gain relevance among women in 2019.
  - “Professional, Scientific and Technical Services” also rises hourly wages for both genders.
  - “Manufacturing durables” is also relevant, though it lost impact in the Males 2019 group, whereas “Wholesale Trade” appear in the top for 2019 groups.
- Education has a positive linear relationship with hourly wages in all four groups, as we expected after seeing the boxplot analysis from Step 2.
- Alberta is the province that seems to provide the largest boost to wages for all groups except Females 2019.

Top variable coefficients that decrease Hourly Wages

Year 2009, Males		Year 2019, Males		Year 2009, Females		Year 2019, Females	
Ind. Var.	Est.	Ind. Var.	Est.	Ind. Var.	Est.	Ind. Var.	Est.
NOC_10ManUtil	-0.076	NOC_10Sales	0.017	NOC_10ManUtil	0.111	NOC_10ManUtil	0.107
NOC_10Sales	-0.069	NOC_10ManUtil	0.015	NOC_10Sales	0.092	NOC_10Sales	0.094
NOC_10BusFin	-0.062	NOC_10NatAgri	0.014	NOC_10Trades	0.088	NOC_10Trades	0.073
NOC_10NatAgri	-0.058	NOC_10ArtCul	0.013	NOC_10NatAgri	0.073	NOC_10ArtCul	0.066
NOC_10Trades	-0.057	NOC_10Trades	0.012	NOC_10BusFin	0.065	NOC_10BusFin	0.063
AGE_12.Q	-0.053	NOC_10BusFin	0.012	NOC_10ArtCul	0.049	NOC_10NatAgri	0.063
NOC_10ArtCul	-0.044	AGE_12.Q	0.008	AGE_12.Q	0.048	NOC_10EduLaw	0.049
UTOTHRS	-0.032	EFAMTYPE HWNE17	0.007	NOC_10NatASc	0.034	UTOTHRS	0.037
NOC_10NatASc	-0.030	NOC_10NatASc	0.007	NOC_10EduLaw	0.031	AGE_12.Q	0.035
NOC_10EduLaw	-0.030	NOC_10EduLaw	0.007	UNIONNon-uni.	0.021	NOC_10NatASc	0.028

From the table, we observe that:

- Many of the variables that can have a negative impact in hourly wages are occupations:
    - "Manufacturing and utilities" are the occupations with highest coefficients that decreases hourly wages. Of note, "Utilities" was previously identified as the one of the industries that provides a positive impact in hourly wages. These results tell us that, probably, within that industry other type of occupations are the ones receiving the highest wages.
    - "Sales and services" occupations also tend to point to lower hourly wages.
    - "Natural resources and agriculture" as well as "Trades, transport and equipment operators" also appear to decrease the dependent variable.
    - The only occupations that are not listed are "Management" and "Health".
- There is negative quadratic relationship between age and hourly wages, as we could see in the boxplot analysis on Step 2, where hourly wages tend to increase with age at the beginning, then there is a peak and a subsequent decrement as the employees get older. These seem to be especially noticeable in male groups.

Step 5: Gender Proportions by Sector, Industry and Occupation

Last research questions require an analysis and comparison of the proportions of men and women by sector, industry, and job occupation. To solve these inquiries, two proportions z-test were performed.

Two Proportions Z-Test

This test is used to compare two observed proportions. The null hypothesis is defined as follow:

- H0: Two proportions are the same, equal

And there are two possible alternative hypotheses:

- HA: Two proportions are not the same, unequal (two-tailed test), or
- HA: One proportion is greater or less than the other (one-tailed test)

The test statistic (z-test) it is calculated using the proportions from each group or sample, and the overall proportions p and q, where:

- p = sum of successes in both samples / sum of both samples (n1 + n2)
- q = 1 – p

The significance level alpha can be set at 0.05 or 5%, to get 95% confidence level. If the p-value of the test is less than the significance level alpha, we can conclude that the proportions of the two compared groups are significantly different.

This z-test is valid only when the sample sizes are large enough, meaning that all next calculations should result in a number greater or equal to 5:

- Sample n1 * p - Sample n1 * q - Sample n2 * p - Sample n2 * q	>= 5

Other conditions to perform this test are:

- Randomness: Randomly selected observations
- Normality: number of successes and failures in each of the sample must be at least 10
- Independence: sampling with replacement or sample size is no longer than 10% of the population.

Large Sample Assumption Test

To verify that the sample size of each subset is large enough to do the z-test, the required calculations were performed considering the proportion of women in each group. All results were greater than 5, as can be seen in the table below.

Column Names:

- n1 = Sample 2009
- n2 = Sample 2019
- p = Overall proportion of women (sum of women cases in 2009 and 2019, divided by sum of samples)
- q = 1 – p

	Samples		Overall Proportions		>= 5 Check			
Variable	**n1**	**n2**	**p**	**q**	**n1*p**	**n1*q**	**n2*p**	**n2*q**
SECTOR								
Private sector	40,123	37,231	0.45	0.55	18,236	21,887	16,922	20,309
Public sector	14,434	13,700	0.64	0.36	9,178	5,256	8,711	4,989
INDUSTRY								
Accommodation & Food Serv.	3,947	3,540	0.63	0.37	2,489	1,458	2,233	1,307
Agriculture	718	709	0.33	0.67	238	480	235	474
Construction	3,590	3,746	0.12	0.88	422	3,168	441	3,305
Educational Services	4,574	4,465	0.69	0.31	3,174	1,400	3,099	1,366
Finance, Ins., Real Est. & Leas.	2,794	2,506	0.63	0.37	1,748	1,046	1,567	939
Forest., Fishing, Min., Oil & Gas	1,566	1,481	0.16	0.84	244	1,322	231	1,250
Health Care & Social Assist.	7,280	7,432	0.85	0.15	6,156	1,124	6,284	1,148
Information, Culture & Rec.	2,185	1,858	0.48	0.52	1,057	1,128	899	959
Management, Admin. & Support	1,813	1,621	0.43	0.57	785	1,028	702	919
Manufacturing durables	3,245	2,736	0.20	0.80	634	2,611	535	2,201
Manufacturing non-durables	2,783	2,308	0.36	0.64	998	1,785	828	1,480
Other Services	2,056	1,716	0.54	0.46	1,103	953	921	795
Prof., Scientific & Technical S.	2,160	2,494	0.48	0.52	1,041	1,119	1,201	1,293
Public Administration	3,758	3,550	0.50	0.50	1,893	1,865	1,789	1,761
Retail Trade	7,149	6,125	0.56	0.44	4,033	3,116	3,455	2,670
Transportation & Warehousing	2,537	2,559	0.27	0.73	679	1,858	685	1,874
Utilities	633	495	0.23	0.77	143	490	112	383
Wholesale Trade	1,769	1,590	0.30	0.70	523	1,246	470	1,120
JOB OCCUPATION								
Art, culture, recreation & sport	952	861	0.59	0.41	561	391	508	353
Business, finance & admin.	9,822	8,085	0.75	0.25	7,320	2,502	6,026	2,059
Educ., law, community & gov. s.	5,143	6,371	0.72	0.28	3,677	1,466	4,556	1,815
Health	3,923	4,077	0.86	0.14	3,361	562	3,492	585
Management	3,587	2,956	0.43	0.57	1,529	2,058	1,260	1,696
Manufacturing & utilities	2,825	2,589	0.27	0.73	756	2,069	693	1,896
Natural & applied sciences	3,380	3,583	0.23	0.77	767	2,613	813	2,770
Natural resources & agriculture	1,581	1,572	0.19	0.81	298	1,283	297	1,275
Sales & service	14,578	12,782	0.59	0.41	8,544	6,034	7,492	5,290
Trades, transport & equipm. op.	8,766	8,055	0.07	0.93	572	8,194	525	7,530

Normality Assumption Test

Both total number of observations for men and women are greater than 10 in all groups:

	Female		Male	
Value	**2009**	**2019**	**2009**	**2019**
SECTOR				
Private sector	18,508	16,650	21,615	20,581
Public sector	9,107	8,782	5,327	4,918
INDUSTRY				
Accommodation & Food Services	2,518	2,204	1,429	1,336
Agriculture	221	252	497	457
Construction	381	482	3,209	3,264
Educational Services	3,114	3,159	1,460	1,306
Finance, Insurance, Real Est. & Leas.	1,823	1,492	971	1,014
Forestry, Fishing, Min., Oil & Gas	236	239	1,330	1,242
Health Care & Social Assistance	6,213	6,227	1,067	1,205
Information, Culture & Recreation	1,079	877	1,106	981
Management, Admin. & Support	821	666	992	955
Manufacturing durables	640	529	2,605	2,207
Manufacturing non-durables	1,007	819	1,776	1,489
Other Services	1,127	897	929	819
Prof., Scientific & Technical Services	1,074	1,168	1,086	1,326
Public Administration	1,878	1,804	1,880	1,746
Retail Trade	4,155	3,333	2,994	2,792
Transportation & Warehousing	677	687	1,860	1,872
Utilities	140	115	493	380
Wholesale Trade	511	482	1,258	1,108
JOB OCCUPATION				
Art, culture, recreation & sport	576	493	376	368
Business, finance & administration	7,388	5,958	2,434	2,127
Educ., law, community & gov. services	3,664	4,569	1,479	1,802
Health	3,398	3,455	525	622
Management	1,546	1,243	2,041	1,713
Manufacturing & utilities	800	649	2,025	1,940
Natural & applied sciences	748	832	2,632	2,751
Natural resources & agriculture	266	329	1,315	1,243
Sales & service	8,695	7,341	5,883	5,441
Trades, transport & equipment operators	534	563	8,232	7,492

Independence Assumption Test

The population of employees was estimated by multiplying the total population of Canada of age 15 or older by the percentages of employed people in the data set (LFSSTAT = "Employed, at work", or "Employed, absent from work").

	2009	2019
Total Canadian Population 15 years or older	28,008,741	31,574,973
Employed Percentage (LFSSTAT = "Employed, at work", or "Employed,absent from work")	60%	56%
Total Employed Estimate	**16,911,247**	**17,641,170**

Then, population estimates for each group were calculated multiplying the different proportions observed by sector, industry, and job occupation. All sub-sample sizes were less than 10% of the estimated total population:

	% of Employed		Population Estimate		10% of Estimate		Sample Size		Sample Size > 10%	
	'09	'19	'09	'19	'09	'19	'09	'19	'09	'19
Total Employed Estimate	**100%**	**100%**	**16,911,247**	**17,641,170**	**1,691,125**	**1,764,117**	**54,557**	**50,931**	**F**	**F**
SECTOR										
Private sector	74%	73%	12,437,084	12,895,847	1,243,708	1,289,585	40,123	37,231	F	F
Public sector	26%	27%	4,474,164	4,745,323	447,416	474,532	14,434	13,700	F	F
INDUSTRY										
Accommodation & Food Serv.	7%	7%	1,223,467	1,226,164	122,347	122,616	3,947	3,540	F	F
Agriculture	1%	1%	222,561	245,579	22,256	24,558	718	709	F	F
Construction	7%	7%	1,112,806	1,297,517	111,281	129,752	3,590	3,746	F	F
Educational Services	8%	9%	1,417,821	1,546,560	141,782	154,656	4,574	4,465	F	F
Finance, Ins., Real Est. & Leas.	5%	5%	866,067	868,013	86,607	86,801	2,794	2,506	F	F
Forest., Fishing, Min., Oil & Gas	3%	3%	485,419	512,980	48,542	51,298	1,566	1,481	F	F
Health Care & Social Assist.	13%	15%	2,256,610	2,574,251	225,661	257,425	7,280	7,432	F	F
Information, Culture & Rec.	4%	4%	677,293	643,563	67,729	64,356	2,185	1,858	F	F
Management, Admin. & Support	3%	3%	561,983	561,472	56,198	56,147	1,813	1,621	F	F
Manufacturing durables	6%	5%	1,005,865	947,679	100,587	94,768	3,245	2,736	F	F
Manufacturing non-durables	5%	5%	862,657	799,431	86,266	79,943	2,783	2,308	F	F
Other Services	4%	3%	637,306	594,378	63,731	59,438	2,056	1,716	F	F
Prof., Scientific & Technical Se.	4%	5%	669,544	863,857	66,954	86,386	2,160	2,494	F	F
Public Administration	7%	7%	1,164,882	1,229,627	116,488	122,963	3,758	3,550	F	F
Retail Trade	13%	12%	2,216,004	2,121,540	221,600	212,154	7,149	6,125	F	F
Transportation & Warehousing	5%	5%	786,404	886,371	78,640	88,637	2,537	2,559	F	F
Utilities	1%	1%	196,213	171,455	19,621	17,146	633	495	F	F

Wholesale Trade	3%	3%	548,344	550,735	54,834	55,073	1,769	1,590	F	F
JOB OCCUPATION										
Art, culture, recreation & sport	2%	2%	295,095	298,228	29,510	29,823	952	861	F	F
Business, finance & admin.	18%	16%	3,044,564	2,800,433	304,456	280,043	9,822	8,085	F	F
Educ., law, community & gov. s.	9%	13%	1,594,196	2,206,748	159,420	220,675	5,143	6,371	F	F
Health	7%	8%	1,216,028	1,412,166	121,603	141,217	3,923	4,077	F	F
Management	7%	6%	1,111,876	1,023,881	111,188	102,388	3,587	2,956	F	F
Manufacturing & utilities	5%	5%	875,676	896,762	87,568	89,676	2,825	2,589	F	F
Natural & applied sciences	6%	7%	1,047,712	1,241,058	104,771	124,106	3,380	3,583	F	F
Natural resources & agriculture	3%	3%	490,069	544,500	49,007	54,450	1,581	1,572	F	F
Sales & service	27%	25%	4,518,800	4,427,351	451,880	442,735	14,578	12,782	F	F
Trades, transport & equipm. op.	16%	16%	2,717,231	2,790,042	271,723	279,004	8,766	8,055	F	F

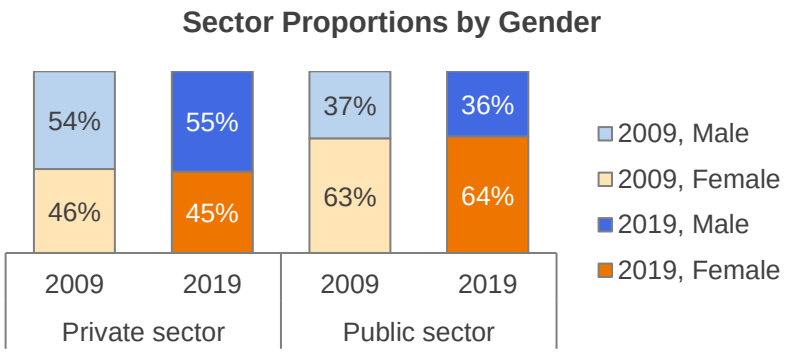
Note: F = FALSE

Next, the results of analysing gender distribution by Sector, Industry, and Occupation. Each section includes a summary of findings, graphs to show the proportions by group and the z-test results.

Note that all the graphs and tables are sorted according to female proportion, in ascending order. Also, the alternative hypotheses chosen depended on whether women participation increased or decreased from 2009 to 2019.

Sector Analysis

The following breakdown shows that there is a larger female presence in the public sector than in the private sector. In the public sector 6 out of 10 employees are women. Besides, the proportion of women in this sector increased by 1% between years. In contrast, the participation of women in the private sector decrease slightly from 46% in 2009 to 45% in 2019.



Sector	Test Results
Private	X-squared = 15.371, df = 1, p-value = 4.417e-05 alternative hypothesis: greater 95 percent confidence interval: 0.008155203 1.000000000 sample estimates: prop 1 prop 2 0.4612816 0.4472080
Public	X-squared = 3.0416, df = 1, p-value = 0.04058 alternative hypothesis: less 95 percent confidence interval: -1.0000000000 -0.0005711443 sample estimates: prop 1 prop 2 0.6309408 0.6410219

Industry Analysis

The industries mostly dominated by men are: “Construction”, "Forestry, Fishing, Mining, Oil and Gas", and "Manufacturing durables". In these industries, 8 to 9 out of 10 employees are men. Other examples of industries were male presence is prominent are “Utilities”, "Transportation and Warehousing", "Wholesale Trade", "Manufacturing non-durables" and "Agriculture".

On the opposite side, the industries with large proportion of female employees are: "Health Care and Social Assistance", "Educational Services", and "Accommodation and Food Services".

From 2009 to 2019, these are the industries were women participation rose:

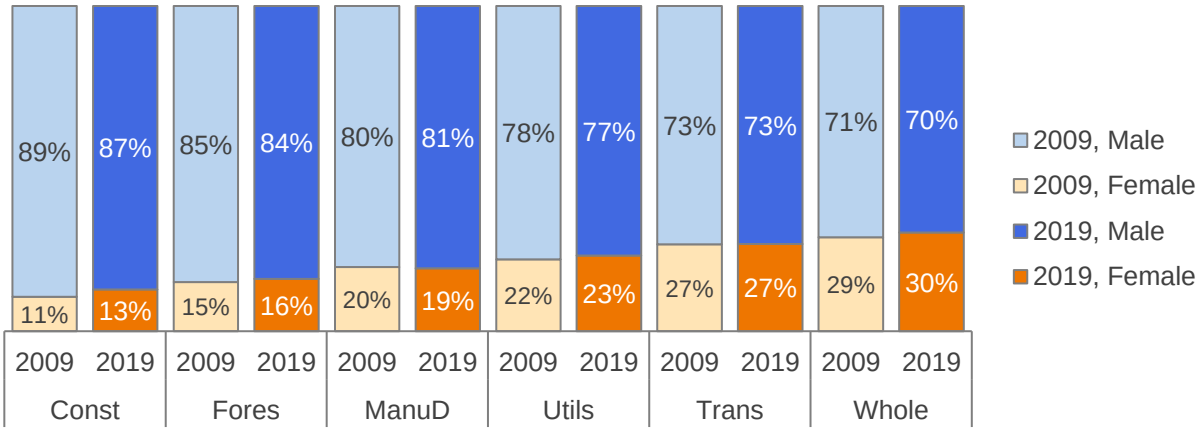
- Agriculture, from 31% to 36%
- Educational services, from 68% to 71%
- Construction, from 11% to 13%

Meanwhile, these were the ones where female presence decreased significantly:

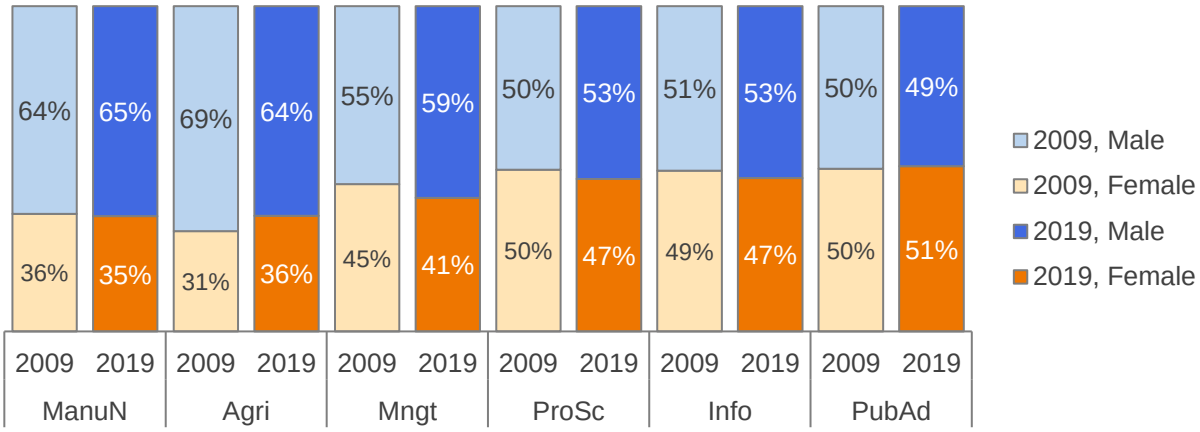
- Finance, Insurance, Real Estate and Leasing, from 65% to 60%
- Management, Administrative and Other Support, from 45% to 41%
- Retail Trade, from 58% to 54%
- Professional, Scientific and Technical Services, from 50% to 47%
- Health Care and Social Assistance, from 85% to 84%

More details are provided in the next graphs and tests results. The visual analysis of industries was divided in three graphs to accommodate text.

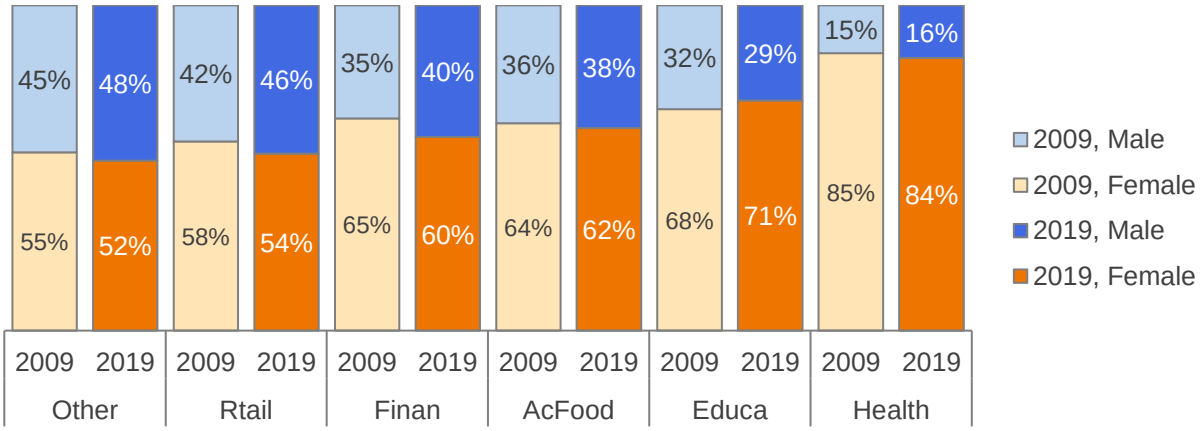
Industry Proportions by Gender (Part 1)



Industry Proportions by Gender (Part 2)



Industry Proportions by Gender (Part 3)



Industry	Test Results
Construction	X-squared = 8.7586, df = 1, p-value = 0.001541 alternative hypothesis: less 95 percent confidence interval: -1.000000000 -0.009921907 sample estimates: prop 1 prop 2 0.1061281 0.1286706
Forestry, Fishing, Mining, Oil and Gas	X-squared = 0.58053, df = 1, p-value = 0.2231 alternative hypothesis: less 95 percent confidence interval: -1.00000000 0.01162354 sample estimates: prop 1 prop 2 0.1507024 0.1613774
Manufacturing Durables	X-squared = 0.11841, df = 1, p-value = 0.3654 alternative hypothesis: greater 95 percent confidence interval: -0.01337678 1.00000000 sample estimates: prop 1 prop 2 0.1972265 0.1933480
Utilities	X-squared = 0.13892, df = 1, p-value = 0.3547 alternative hypothesis: less 95 percent confidence interval: -1.00000000 0.03201062 sample estimates: prop 1 prop 2 0.2211690 0.2323232
Transportation and Warehousing	X-squared = 0.0096919, df = 1, p-value = 0.4608 alternative hypothesis: less 95 percent confidence interval: -1.00000000 0.01918169 sample estimates: prop 1 prop 2 0.2668506 0.2684642
Wholesale Trade	X-squared = 0.75298, df = 1, p-value = 0.1928 alternative hypothesis: less 95 percent confidence interval: -1.0000000 0.0122707 sample estimates: prop 1 prop 2 0.2888638 0.3031447
Manufacturing non-durables	X-squared = 0.23825, df = 1, p-value = 0.3127 alternative hypothesis: greater 95 percent confidence interval: -0.01560947 1.00000000 sample estimates: prop 1 prop 2 0.3618397 0.3548527
Agriculture	X-squared = 3.4405, df = 1, p-value = 0.03181 alternative hypothesis: less 95 percent confidence interval: -1.000000000 -0.005276911 sample estimates: prop 1 prop 2 0.3077994 0.3554302
Management, Administration and Support	X-squared = 5.9742, df = 1, p-value = 0.007258 alternative hypothesis: greater 95 percent confidence interval: 0.01358235 1.00000000 sample estimates: prop 1 prop 2 0.4528406 0.4108575
Professional, Scientific and Technical Services	X-squared = 3.7569, df = 1, p-value = 0.0263 alternative hypothesis: greater 95 percent confidence interval: 0.00431574 1.00000000 sample estimates:

	prop 1 prop 2 0.4972222 0.4683240
Information, Culture and Recreation	X-squared = 1.826, df = 1, p-value = 0.0883 alternative hypothesis: greater 95 percent confidence interval: -0.004620229 1.000000000 sample estimates: prop 1 prop 2 0.4938215 0.4720129
Public Administration	X-squared = 0.48639, df = 1, p-value = 0.2428 alternative hypothesis: less 95 percent confidence interval: -1.00000000 0.01108627 sample estimates: prop 1 prop 2 0.4997339 0.5081690
Other Services	X-squared = 2.3303, df = 1, p-value = 0.06344 alternative hypothesis: greater 95 percent confidence interval: -0.001929488 1.000000000 sample estimates: prop 1 prop 2 0.5481518 0.5227273
Retail Trade	X-squared = 18.252, df = 1, p-value = 9.675e-06 alternative hypothesis: greater 95 percent confidence interval: 0.0226837 1.0000000 sample estimates: prop 1 prop 2 0.5812002 0.5441633
Finance, Insurance, Real Estate and Leasing	X-squared = 18.143, df = 1, p-value = 1.025e-05 alternative hypothesis: greater 95 percent confidence interval: 0.03481883 1.00000000 sample estimates: prop 1 prop 2 0.6524696 0.5953711
Accommodation and Food Services	X-squared = 1.8235, df = 1, p-value = 0.08845 alternative hypothesis: greater 95 percent confidence interval: -0.003296079 1.000000000 sample estimates: prop 1 prop 2 0.6379529 0.6225989
Educational Services	X-squared = 7.4584, df = 1, p-value = 0.003157 alternative hypothesis: less 95 percent confidence interval: -1.0000000 -0.0105416 sample estimates: prop 1 prop 2 0.6808045 0.7075028
Health Care and Social Assistance	X-squared = 6.7094, df = 1, p-value = 0.004795 alternative hypothesis: greater 95 percent confidence interval: 0.005639885 1.000000000 sample estimates: prop 1 prop 2 0.8534341 0.8378633

Occupation Analysis

The job occupations dominated by men are “Trades, transport, and equipment operators”, “Natural resources and agriculture”, “Natural and applied sciences”, and “Manufacturing and utilities”.

On the other hand, the occupations were women have their largest participation are “Health”, “Business, finance, and administration”, “Education, law, community, and government services”.

Over the course of the decade, the presence of women grew significantly in:

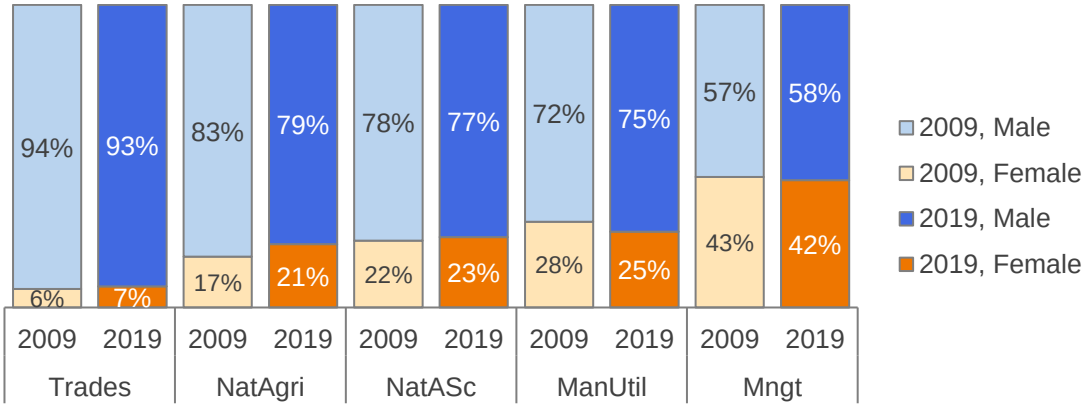
- Natural resources and agriculture from 17% to 21%
- Trades, transport, and equipment operators from 6% to 7%

At the same time, the female proportion decreased significantly in the next occupations:

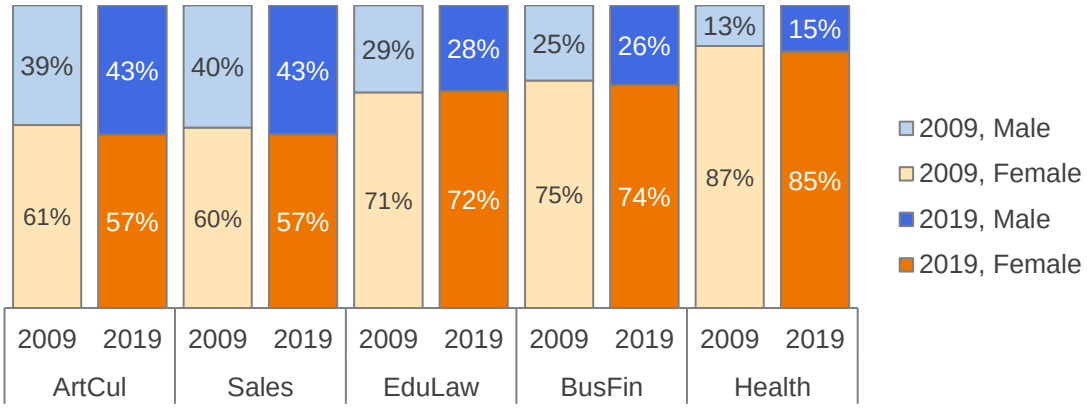
- Manufacturing and utilities from 28% to 25%
- Sales and services, from 60% to 57%
- Business, finance, and administration, from 75% to 74%
- Health from 87% to 85%

All results are provided in the next graphs and tables. Occupations are shown in two separate graphs to accommodate text.

Occupations Proportions by Gender (Part 1)



Occupations Proportions by Gender (Part 2)



Job Occupation	Test Results
Trades, transport, and equipment operators	X-squared = 5.403, df = 1, p-value = 0.01005 alternative hypothesis: less 95 percent confidence interval: -1.000000000 -0.002573939 sample estimates: prop 1 prop 2 0.06091718 0.06989448
Natural resources and agriculture	X-squared = 8.4056, df = 1, p-value = 0.00187 alternative hypothesis: less 95 percent confidence interval: -1.0000000 -0.0175078 sample estimates: prop 1 prop 2 0.1682479 0.2092875
Natural and Applied Sciences	X-squared = 1.1179, df = 1, p-value = 0.1452 alternative hypothesis: less 95 percent confidence interval: -1.000000000 0.005891185 sample estimates: prop 1 prop 2 0.2213018 0.2322076
Manufacturing and utilities	X-squared = 7.1194, df = 1, p-value = 0.003813 alternative hypothesis: greater 95 percent confidence interval: 0.01237363 1.00000000 sample estimates: prop 1 prop 2 0.2831858 0.2506759
Management	X-squared = 0.68827, df = 1, p-value = 0.2034 alternative hypothesis: greater 95 percent confidence interval: -0.01000759 1.00000000 sample estimates: prop 1 prop 2 0.4310008 0.4205007
Art, culture, recreation, and sport	X-squared = 1.8359, df = 1, p-value = 0.08771 alternative hypothesis: greater 95 percent confidence interval: -0.006708564 1.000000000 sample estimates: prop 1 prop 2 0.605042 0.572590
Sales and service	X-squared = 13.65, df = 1, p-value = 0.0001101 alternative hypothesis: greater 95 percent confidence interval: 0.01223069 1.00000000 sample estimates: prop 1 prop 2 0.5964467 0.5743233
Education, law, community, and government services	X-squared = 0.28984, df = 1, p-value = 0.2952 alternative hypothesis: less 95 percent confidence interval: -1.000000000 0.009369943 sample estimates: prop 1 prop 2 0.7124247 0.7171559
Business, finance, and administration	X-squared = 5.3661, df = 1, p-value = 0.01027 alternative hypothesis: greater 95 percent confidence interval: 0.004375397 1.000000000 sample estimates: prop 1 prop 2 0.7521890 0.7369202
Health	X-squared = 5.5633, df = 1, p-value = 0.00917 alternative hypothesis: greater 95 percent confidence interval: 0.005612905 1.000000000 sample estimates: prop 1 prop 2 0.8661738 0.8474368

Conclusions

Here are the answers to the research questions.

What variables were relevant to explain hourly wages by gender when the decade started?
After 10 years, are these variables still important?
Are they the same for women and for men?

According to the regression analysis performed and the subsequent models obtained, the variables that have more impact on hourly wages regardless year and gender are occupation (NOC_10), industry (NAICS_18), age (AGE_12), and province (PROV). Besides, education (EDUC) and tenure are important, especially for women in 2019.

Particularly, industries like "Utilities", "Forestry, Fishing, Mining, Oil and Gas", show a positive influence on hourly wages for both genders throughout the years. Also, "Construction" was mainly beneficial for men, and "Finance, Insurance, Real Estate and Leasing" for women.

Education has a linear relationship with hourly wages, which tend to increase as the former does.
Meanwhile, age has a quadratic relationship, where wages tend to increase and peak when individuals are 40 to 50 years old and begin to decrease afterwards. Also, among provinces, Alberta is the one that seems to have the most positive effect.

In contrast, some occupations tend to reflect negatively on hourly wages, especially “Manufacturing and utilities” and “Sales and Services”. Since “Utilities” was previously identified as one of the industries with positive impact, these results might imply that in this industry other type of occupations are the ones receiving the highest wages. Of note, the only occupations that did not appear among the variables with biggest diminishing effect on hourly wages were "Management" and “Health”.

What were the sectors, industries, and job occupations with highest hourly wages by gender at the beginning of the decade?
Are these the same by the end of 2019?
Which of these groups show more equality in wages by gender?

There are two sectors, public and private. Between them the one that tend to offer the highest hourly wages is the public sector for both genders over time. Anyway, men out earn women in both sectors.

The industries with distributions at the higher end of hourly wages are "Utilities", "Public Administration", "Forestry, Fishing, Mining, Oil and Gas", "Educational Services" and "Professional, Scientific and Technical Services". These remained the same through time. Men tend to get higher hourly wages than women overall, and the smallest gaps are found in "Management, Administrative and Other Support" and "Accommodation and Food Services".

Among occupations, the top ones by hourly wages are “Management”, “Educational Services”, and “Natural and Applied Sciences”. Again, these were consistent over the course of the decade. Males tend to have higher hourly wages, and the smallest gaps by gender are observed on occupations in “Health” and "Art, culture, recreation and sport".

What is the gender distribution by sector, industry, and job occupation?
What groups are more uneven?
Was there a change in the distribution over time?

Public sector has most female employees, around 6 out of 10 employees are women, and their participation grew from 2009 to 2019. On the other hand, private sector is almost even, and female proportion decreased slightly through the decade.

Industries that are dominated by men are “Construction”, "Forestry, Fishing, Mining, Oil and Gas", and "Manufacturing durables", “Utilities”, "Transportation and Warehousing", and "Wholesale Trade". In contrast, the industries where women are majority are: "Health Care and Social Assistance", "Educational Services", and "Accommodation and Food Services".

The industries where women participation rose over time are “Agriculture”, “Educational services”, and “Construction”. And the ones were the proportion decreased the most were “Finance, Insurance, Real Estate and Leasing”, “Management, Administrative and Other Support”, and “Retail Trade”.

Lastly, regarding job occupation, the ones dominated by men are “Trades, transport, and equipment operators”, “Natural resources and agriculture”, “Natural and applied sciences”, and “Manufacturing and utilities”. Meanwhile the occupations with more women are “Health”, “Business, finance, and administration”, “Education, law, community, and government services”.

Female participation became more common in “Natural resources and agriculture” and “Trades, transport, and equipment operators” occupations from 2009 to 2019. On the other hand, the female proportion decreased the most in “Manufacturing and utilities” and “Sales and services”.

Appendix

Post-Hoc Analysis

Comparisons that are NOT significant were highlighted in gray color.

Industry

Males 2009

```
> dunnTest(HRLYEARN ~ NAICS_18short, data = data.all.09male, method = "bonferroni")
Dunn (1964) Kruskal-Wallis multiple comparison
  p-values adjusted with the Bonferroni method.
```

	Comparison	Z	P.unadj	P.adj
1	AcFood - Agri	-4.5099351	6.484745e-06	9.921660e-04
2	AcFood - Const	-37.7674843	0.000000e+00	0.000000e+00
3	Agri - Const	-20.0447500	2.243056e-89	3.431876e-87
4	AcFood - Educa	-42.3702745	0.000000e+00	0.000000e+00
5	Agri - Educa	-25.8376551	3.349067e-147	5.124072e-145
6	Const - Educa	-11.8969838	1.227023e-32	1.877344e-30
7	AcFood - Finan	-29.2143292	1.275422e-187	1.951396e-185
8	Agri - Finan	-17.7710667	1.184175e-70	1.811787e-68
9	Const - Finan	-0.3792241	7.045215e-01	1.000000e+00
10	Educa - Finan	8.7340372	2.457401e-18	3.759824e-16
11	AcFood - Fores	-41.0747855	0.000000e+00	0.000000e+00
12	Agri - Fores	-25.3003738	3.164115e-141	4.841096e-139
13	Const - Fores	-11.1578304	6.557231e-29	1.003256e-26
14	Educa - Fores	0.3085226	7.576847e-01	1.000000e+00
15	Finan - Fores	-8.2913240	1.119947e-16	1.713520e-14
16	AcFood - Health	-29.7847894	6.149310e-195	9.408445e-193
17	Agri - Health	-17.8655628	2.187150e-71	3.346340e-69
18	Const - Health	-0.1125725	9.103695e-01	1.000000e+00
19	Educa - Health	9.2261127	2.806318e-20	4.293667e-18
20	Finan - Health	0.2234733	8.231672e-01	1.000000e+00
21	Fores - Health	8.7568708	2.007455e-18	3.071407e-16
22	AcFood - Info	-23.4295596	2.136269e-121	3.268491e-119
23	Agri - Info	-13.0269075	8.602409e-39	1.316169e-36
24	Const - Info	7.5361437	4.840733e-14	7.406321e-12
25	Educa - Info	16.0131327	1.034666e-57	1.583038e-55
26	Finan - Info	6.2909461	3.155369e-10	4.827714e-08
27	Fores - Info	15.3987377	1.668955e-53	2.553501e-51
28	Health - Info	6.2163107	5.089797e-10	7.787389e-08
29	AcFood - ManuD	-37.4270624	1.418558e-306	2.170394e-304
30	Agri - ManuD	-20.3726198	2.925987e-92	4.476760e-90
31	Const - ManuD	-1.1738112	2.404707e-01	1.000000e+00
32	Educa - ManuD	10.5409332	5.594102e-26	8.558977e-24
33	Finan - ManuD	-0.4539006	6.499004e-01	1.000000e+00
34	Fores - ManuD	9.8785308	5.158362e-23	7.892294e-21
35	Health - ManuD	-0.7422397	4.579421e-01	1.000000e+00
36	Info - ManuD	-8.1842671	2.739659e-16	4.191678e-14
37	AcFood - ManuN	-29.2603684	3.314235e-188	5.070780e-186
38	Agri - ManuN	-15.8625561	1.151119e-56	1.761211e-54
39	Const - ManuN	5.4537016	4.933197e-08	7.547791e-06
40	Educa - ManuN	15.1969136	3.706788e-52	5.671385e-50
41	Finan - ManuN	4.3892938	1.137194e-05	1.739906e-03
42	Fores - ManuN	14.4824941	1.563205e-47	2.391703e-45
43	Health - ManuN	4.2669192	1.981909e-05	3.032321e-03
44	Info - ManuN	-2.6492314	8.067508e-03	1.000000e+00
45	ManuD - ManuN	6.2474806	4.171262e-10	6.382030e-08
46	AcFood - Mngt	-10.7344702	7.012001e-27	1.072836e-24
47	Agri - Mngt	-3.7986455	1.454890e-04	2.225982e-02

48	Const - Mngt	20.8518073	1.467729e-96	2.245626e-94
49	Educa - Mngt	27.5375599	6.236695e-167	9.542143e-165
50	Finan - Mngt	17.0873946	1.842227e-65	2.818608e-63
51	Fores - Mngt	26.7299537	2.112084e-157	3.231489e-155
52	Health - Mngt	17.2648998	8.645375e-67	1.322742e-64
53	Info - Mngt	11.3134000	1.126384e-29	1.723367e-27
54	ManuD - Mngt	21.1331322	3.944908e-99	6.035710e-97
55	ManuN - Mngt	15.0413425	3.934883e-51	6.020372e-49
56	AcFood - Other	-19.9172536	2.883802e-88	4.412217e-86
57	Agri - Other	-10.8783977	1.461086e-27	2.235462e-25
58	Const - Other	9.7081699	2.782886e-22	4.257816e-20
59	Educa - Other	17.5669688	4.411211e-69	6.749153e-67
60	Finan - Other	8.1836290	2.754212e-16	4.213944e-14
61	Fores - Other	16.9688359	1.396735e-64	2.137004e-62
62	Health - Other	8.1489250	3.671735e-16	5.617754e-14
63	Info - Other	2.2227781	2.623077e-02	1.000000e+00
64	ManuD - Other	10.2750259	9.131989e-25	1.397194e-22
65	ManuN - Other	4.9492725	7.449140e-07	1.139718e-04
66	Mngt - Other	-8.6691671	4.352926e-18	6.659976e-16
67	AcFood - ProSc	-41.5224239	0.000000e+00	0.000000e+00
68	Agri - ProSc	-26.5288259	4.508432e-155	6.897902e-153
69	Const - ProSc	-13.4007115	5.988628e-41	9.162601e-39
70	Educa - ProSc	-2.3677253	1.789782e-02	1.000000e+00
71	Finan - ProSc	-10.3371712	4.784527e-25	7.320326e-23
72	Fores - ProSc	-2.6058010	9.165971e-03	1.000000e+00
73	Health - ProSc	-10.8217299	2.716020e-27	4.155510e-25
74	Info - ProSc	-17.1634486	4.986293e-66	7.629029e-64
75	ManuD - ProSc	-12.1673354	4.639828e-34	7.098937e-32
76	ManuN - ProSc	-16.3998441	1.917295e-60	2.933461e-58
77	Mngt - ProSc	-27.9592280	5.091681e-172	7.790273e-170
78	Other - ProSc	-18.6200554	2.209853e-77	3.381075e-75
79	AcFood - PubAd	-49.3334070	0.000000e+00	0.000000e+00
80	Agri - PubAd	-29.6707283	1.832678e-193	2.803997e-191
81	Const - PubAd	-18.2579525	1.788743e-74	2.736777e-72
82	Educa - PubAd	-4.4351282	9.201750e-06	1.407868e-03
83	Finan - PubAd	-13.0667428	5.100305e-39	7.803467e-37
84	Fores - PubAd	-4.6443285	3.411843e-06	5.220119e-04
85	Health - PubAd	-13.7310960	6.612562e-43	1.011722e-40
86	Info - PubAd	-20.9272479	3.024630e-97	4.627684e-95
87	ManuD - PubAd	-16.4999676	3.670896e-61	5.616470e-59
88	ManuN - PubAd	-20.8994697	5.414311e-97	8.283895e-95
89	Mngt - PubAd	-32.8157234	3.513473e-236	5.375614e-234
90	Other - PubAd	-22.2413990	1.366295e-109	2.090432e-107
91	ProSc - PubAd	-1.5698229	1.164563e-01	1.000000e+00
92	AcFood - Rtail	-12.6163940	1.714959e-36	2.623887e-34
93	Agri - Rtail	-3.5261348	4.216723e-04	6.451586e-02
94	Const - Rtail	31.3060019	3.866308e-215	5.915452e-213
95	Educa - Rtail	36.6854413	1.246455e-294	1.907076e-292
96	Finan - Rtail	21.9154038	1.852445e-106	2.834241e-104
97	Fores - Rtail	35.1816700	3.813268e-271	5.834301e-269
98	Health - Rtail	22.4220798	2.397189e-111	3.667699e-109
99	Info - Rtail	15.1385577	9.016963e-52	1.379595e-49
100	ManuD - Rtail	30.8441456	6.711616e-209	1.026877e-206
101	ManuN - Rtail	21.1733968	1.680014e-99	2.570422e-97
102	Mngt - Rtail	1.0363160	3.000547e-01	1.000000e+00
103	Other - Rtail	11.5499518	7.386185e-31	1.130086e-28
104	ProSc - Rtail	35.7364794	1.073060e-279	1.641782e-277
105	PubAd - Rtail	45.0525901	0.000000e+00	0.000000e+00
106	AcFood - Trans	-29.4477419	1.345623e-190	2.058803e-188
107	Agri - Trans	-15.8636141	1.131884e-56	1.731783e-54
108	Const - Trans	5.6696311	1.431053e-08	2.189512e-06
109	Educa - Trans	15.4665636	5.833927e-54	8.925908e-52
110	Finan - Trans	4.5240362	6.067134e-06	9.282714e-04
111	Fores - Trans	14.7340540	3.896151e-49	5.961110e-47
112	Health - Trans	4.4059008	1.053452e-05	1.611781e-03
113	Info - Trans	-2.5689704	1.020012e-02	1.000000e+00
114	ManuD - Trans	6.4625799	1.029327e-10	1.574871e-08
115	ManuN - Trans	0.1184802	9.056872e-01	1.000000e+00
116	Mngt - Trans	-15.0645812	2.769148e-51	4.236796e-49

```
117   Other - Trans  -4.8902660  1.006998e-06  1.540707e-04
118   ProSc - Trans  16.6451239  3.282858e-62  5.022773e-60
119   PubAd - Trans  21.2666469 2.312302e-100  3.537822e-98
120   Rtail - Trans -21.3468864 4.167997e-101  6.377036e-99
121  AcFood - Utils -34.5189151 4.173790e-261 6.385898e-259
122    Agri - Utils -24.6699310 2.249349e-134 3.441504e-132
123   Const - Utils -12.4423806  1.538492e-35  2.353893e-33
124   Educa - Utils  -4.3447751  1.394184e-05  2.133101e-03
125   Finan - Utils -10.6325226  2.103446e-26  3.218273e-24
126   Fores - Utils  -4.5139364  6.363531e-06  9.736202e-04
127  Health - Utils -10.9793525  4.803543e-28  7.349421e-26
128    Info - Utils -15.9668762  2.174045e-57  3.326288e-55
129   ManuD - Utils -11.6243437  3.099662e-31  4.742483e-29
130   ManuN - Utils -14.9918027  8.306759e-51  1.270934e-48
131    Mngt - Utils -24.6692896 2.285284e-134 3.496485e-132
132   Other - Utils -17.2929157  5.319241e-67  8.138439e-65
133   ProSc - Utils  -2.4203115  1.550722e-02  1.000000e+00
134   PubAd - Utils  -1.4151453  1.570259e-01  1.000000e+00
135   Rtail - Utils -28.7492567 9.253465e-182 1.415780e-179
136   Trans - Utils -15.1434995  8.364188e-52  1.279721e-49
137  AcFood - Whole -27.1432118 3.045119e-162 4.659033e-160
138    Agri - Whole -15.3741342  2.440846e-53  3.734494e-51
139   Const - Whole   4.5608481  5.094743e-06  7.794957e-04
140   Educa - Whole  13.7067661  9.248910e-43  1.415083e-40
141   Finan - Whole   3.8767445  1.058634e-04  1.619710e-02
142   Fores - Whole  13.1095053  2.904871e-39  4.444453e-37
143  Health - Whole   3.7409380  1.833347e-04  2.805021e-02
144    Info - Whole  -2.6942361  7.055016e-03  1.000000e+00
145   ManuD - Whole   5.3204910  1.034876e-07  1.583360e-05
146   ManuN - Whole  -0.2599357  7.949134e-01  1.000000e+00
147    Mngt - Whole -14.2665621  3.535849e-46  5.409849e-44
148   Other - Whole  -4.8539725  1.210125e-06  1.851492e-04
149   ProSc - Whole  15.0203171  5.404853e-51  8.269425e-49
150   PubAd - Whole  18.7229109  3.220606e-78  4.927528e-76
151   Rtail - Whole -19.1594368  8.074212e-82  1.235354e-79
152   Trans - Whole  -0.3700849  7.113192e-01  1.000000e+00
153   Utils - Whole  14.1827829  1.171088e-45  1.791765e-43
```

Males 2019

```
dunnTest(HRLYEARN ~ NAICS_18short, data = data.all.19male, method = "bonferroni") # Po
st Hoc
Dunn (1964) Kruskal-Wallis multiple comparison
  p-values adjusted with the Bonferroni method.

        Comparison            Z       P.unadj         P.adj
1     AcFood - Agri  -8.28835953  1.148211e-16  1.756762e-14
2    AcFood - Const -40.58853323  0.000000e+00  0.000000e+00
3      Agri - Const -17.40120727  8.078227e-68  1.235969e-65
4    AcFood - Educa -43.05676910  0.000000e+00  0.000000e+00
5      Agri - Educa -22.56317389 9.970615e-113 1.525504e-110
6     Const - Educa -10.90898810  1.044121e-27  1.597505e-25
7    AcFood - Finan -33.60189100 1.574159e-247 2.408463e-245
8      Agri - Finan -16.86767516  7.779861e-64  1.190319e-61
9     Const - Finan  -2.25964582  2.384324e-02  1.000000e+00
10    Educa - Finan   6.59285511  4.314476e-11  6.601148e-09
11   AcFood - Fores -45.84930897  0.000000e+00  0.000000e+00
12     Agri - Fores -24.82225983 5.155853e-136 7.888455e-134
13    Const - Fores -14.66570181  1.069061e-48  1.635663e-46
14    Educa - Fores  -3.32443977  8.859637e-04  1.355524e-01
15    Finan - Fores  -9.63296487  5.803046e-22  8.878661e-20
16  AcFood - Health -30.73637062 1.860008e-207 2.845813e-205
17    Agri - Health -14.05182606  7.507299e-45  1.148617e-42
18   Const - Health   2.88206029  3.950841e-03  6.044787e-01
19   Educa - Health  11.37416148  5.624077e-30  8.604837e-28
20   Finan - Health   4.18601878  2.838899e-05  4.343515e-03
21   Fores - Health  14.49462440  1.310171e-47  2.004561e-45
22    AcFood - Info -22.22851216 1.820694e-109 2.785662e-107
```

23	Agri - Info	-8.57178333	1.018947e-17	1.558988e-15
24	Const - Info	10.53666269	5.853952e-26	8.956546e-24
25	Educa - Info	17.53453771	7.808202e-69	1.194655e-66
26	Finan - Info	10.38076401	3.033378e-25	4.641069e-23
27	Fores - Info	20.42859365	9.314622e-93	1.425137e-90
28	Health - Info	6.66230978	2.695571e-11	4.124224e-09
29	AcFood - ManuD	-34.38131524	4.796908e-259	7.339269e-257
30	Agri - ManuD	-14.45017730	2.500414e-47	3.825633e-45
31	Const - ManuD	4.58905794	4.452508e-06	6.812338e-04
32	Educa - ManuD	13.85382049	1.206220e-43	1.845517e-41
33	Finan - ManuD	5.47490712	4.377412e-08	6.697440e-06
34	Fores - ManuD	17.34938442	1.993691e-67	3.050348e-65
35	Health - ManuD	0.81853081	4.130542e-01	1.000000e+00
36	Info - ManuD	-6.70212956	2.054038e-11	3.142678e-09
37	AcFood - ManuN	-28.73533937	1.381142e-181	2.113148e-179
38	Agri - ManuN	-11.85017875	2.147323e-32	3.285404e-30
39	Const - ManuN	7.52746795	5.173370e-14	7.915256e-12
40	Educa - ManuN	15.63082401	4.489595e-55	6.869080e-53
41	Finan - ManuN	7.77685113	7.435202e-15	1.137586e-12
42	Fores - ManuN	18.84932372	2.976228e-79	4.553628e-77
43	Health - ManuN	3.56792855	3.598146e-04	5.505163e-02
44	Info - ManuN	-3.60507603	3.120616e-04	4.774542e-02
45	ManuD - ManuN	3.24822810	1.161261e-03	1.776730e-01
46	AcFood - Mngt	-12.68110854	7.525988e-37	1.151476e-34
47	Agri - Mngt	-1.55069642	1.209745e-01	1.000000e+00
48	Const - Mngt	21.22621904	5.468948e-100	8.367490e-98
49	Educa - Mngt	26.73019453	2.098515e-157	3.210727e-155
50	Finan - Mngt	19.11965392	1.732533e-81	2.650775e-79
51	Fores - Mngt	29.50545833	2.450287e-191	3.748938e-189
52	Health - Mngt	15.78236506	4.114766e-56	6.295592e-54
53	Info - Mngt	8.73897385	2.352373e-18	3.599131e-16
54	ManuD - Mngt	16.89634135	4.787026e-64	7.324150e-62
55	ManuN - Mngt	13.15827510	1.525283e-39	2.333683e-37
56	AcFood - Other	-20.82190322	2.740806e-96	4.193433e-94
57	Agri - Other	-8.13350708	4.170457e-16	6.380799e-14
58	Const - Other	10.08687550	6.314756e-24	9.661577e-22
59	Educa - Other	16.85794133	9.172882e-64	1.403451e-61
60	Finan - Other	10.12011387	4.498919e-24	6.883346e-22
61	Fores - Other	19.62017409	1.039964e-85	1.591146e-83
62	Health - Other	6.55960474	5.395059e-11	8.254440e-09
63	Info - Other	0.22318801	8.233892e-01	1.000000e+00
64	ManuD - Other	6.54376419	5.998937e-11	9.178373e-09
65	ManuN - Other	3.65045998	2.617711e-04	4.005098e-02
66	Mngt - Other	-8.11967399	4.674376e-16	7.151796e-14
67	AcFood - ProSc	-43.48585293	0.000000e+00	0.000000e+00
68	Agri - ProSc	-22.79603167	5.020164e-115	7.680851e-113
69	Const - ProSc	-11.28240245	1.603073e-29	2.452701e-27
70	Educa - ProSc	-0.26244195	7.929807e-01	1.000000e+00
71	Finan - ProSc	-6.85994790	6.888566e-12	1.053951e-09
72	Fores - ProSc	3.07762801	2.086552e-03	3.192425e-01
73	Health - ProSc	-11.67262056	1.759194e-31	2.691567e-29
74	Info - ProSc	-17.83448583	3.815434e-71	5.837613e-69
75	ManuD - ProSc	-14.21439308	7.459535e-46	1.141309e-43
76	ManuN - ProSc	-15.96496852	2.241548e-57	3.429568e-55
77	Mngt - ProSc	-27.05681925	3.175212e-161	4.858074e-159
78	Other - ProSc	-17.13736948	7.810512e-66	1.195008e-63
79	AcFood - PubAd	-50.05247341	0.000000e+00	0.000000e+00
80	Agri - PubAd	-26.07686093	6.673344e-150	1.021022e-147
81	Const - PubAd	-16.90002649	4.497045e-64	6.880479e-62
82	Educa - PubAd	-3.93321219	8.381818e-05	1.282418e-02
83	Finan - PubAd	-10.63342957	2.083084e-26	3.187119e-24
84	Fores - PubAd	-0.32689716	7.437457e-01	1.000000e+00
85	Health - PubAd	-15.97351064	1.954676e-57	2.990655e-55
86	Info - PubAd	-22.17301208	6.257944e-109	9.574654e-107
87	ManuD - PubAd	-19.59333730	1.762495e-85	2.696618e-83
88	ManuN - PubAd	-20.87833653	8.427511e-97	1.289409e-94
89	Mngt - PubAd	-31.85273431	1.206413e-222	1.845812e-220
90	Other - PubAd	-21.13907470	3.478303e-99	5.321804e-97
91	ProSc - PubAd	-3.66941512	2.431060e-04	3.719522e-02

```
92   AcFood - Rtail -12.09650690  1.102014e-33  1.686081e-31
93     Agri - Rtail   0.92637367  3.542518e-01  1.000000e+00
94    Const - Rtail  35.52779969 1.829994e-276 2.799891e-274
95    Educa - Rtail  37.97409013  0.000000e+00  0.000000e+00
96    Finan - Rtail  27.19444312 7.556062e-163 1.156077e-160
97    Fores - Rtail  41.18765667  0.000000e+00  0.000000e+00
98   Health - Rtail  23.75275139 1.029114e-124 1.574544e-122
99     Info - Rtail  14.33946062  1.240155e-46  1.897438e-44
100   ManuD - Rtail  27.71464342 4.651086e-169 7.116162e-167
101   ManuN - Rtail  21.20473421 8.635704e-100  1.321263e-97
102    Mngt - Rtail   3.59988096  3.183629e-04  4.870952e-02
103   Other - Rtail  13.12691720  2.308651e-39  3.532236e-37
104   ProSc - Rtail  38.47749480  0.000000e+00  0.000000e+00
105   PubAd - Rtail  46.44076198  0.000000e+00  0.000000e+00
106  AcFood - Trans -29.23520649 6.924158e-188 1.059396e-185
107    Agri - Trans -11.45908476  2.117405e-30  3.239629e-28
108   Const - Trans   9.35483503  8.373028e-21  1.281073e-18
109   Educa - Trans  17.42963392  4.915969e-68  7.521432e-66
110   Finan - Trans   9.03924198  1.577620e-19  2.413758e-17
111   Fores - Trans  20.77129398  7.870915e-96  1.204250e-93
112  Health - Trans   4.71310081  2.439754e-06  3.732824e-04
113    Info - Trans  -2.85240756  4.338943e-03  6.638583e-01
114   ManuD - Trans   4.60680472  4.089037e-06  6.256226e-04
115   ManuN - Trans   1.03147241  3.023193e-01  1.000000e+00
116    Mngt - Trans -12.81733875  1.311242e-37  2.006200e-35
117   Other - Trans  -2.93572647  3.327676e-03  5.091344e-01
118   ProSc - Trans  17.79262966  8.060670e-71  1.233282e-68
119   PubAd - Trans  23.21280955 3.380486e-119 5.172144e-117
120   Rtail - Trans -21.57983720 2.778444e-103 4.251019e-101
121  AcFood - Utils -36.52497457 4.452330e-292 6.812065e-290
122    Agri - Utils -24.11756293 1.635629e-128 2.502512e-126
123   Const - Utils -14.85599154  6.361963e-50  9.733803e-48
124   Educa - Utils  -7.68707808  1.505333e-14  2.303159e-12
125   Finan - Utils -12.03695878  2.271772e-33  3.475811e-31
126   Fores - Utils  -5.39525541  6.842615e-08  1.046920e-05
127  Health - Utils -15.33781111  4.273641e-53  6.538670e-51
128    Info - Utils -19.67601083  3.462043e-86  5.296926e-84
129   ManuD - Utils -16.77541906  3.692449e-63  5.649448e-61
130   ManuN - Utils -18.10651340  2.831220e-73  4.331766e-71
131    Mngt - Utils -26.15147778 9.481096e-151 1.450608e-148
132   Other - Utils -19.32439609  3.348796e-83  5.123658e-81
133   ProSc - Utils  -7.52434189  5.298653e-14  8.106939e-12
134   PubAd - Utils  -5.37313123  7.738097e-08  1.183929e-05
135   Rtail - Utils -31.47661611 1.815283e-217 2.777383e-215
136   Trans - Utils -19.13185665  1.371031e-81  2.097678e-79
137  AcFood - Whole -29.35351872 2.154785e-189 3.296821e-187
138    Agri - Whole -13.37481241  8.486753e-41  1.298473e-38
139   Const - Whole   3.61098251  3.050392e-04  4.667100e-02
140   Educa - Whole  11.81910528  3.109788e-32  4.757976e-30
141   Finan - Whole   4.75824437  1.952840e-06  2.987845e-04
142   Fores - Whole  14.87023778  5.143035e-50  7.868844e-48
143  Health - Whole   0.68237248  4.950035e-01  1.000000e+00
144    Info - Whole  -5.88731236  3.925264e-09  6.005654e-07
145   ManuD - Whole  -0.02490314  9.801322e-01  1.000000e+00
146   ManuN - Whole  -2.76876074  5.626994e-03  8.609301e-01
147    Mngt - Whole -14.84228444  7.805201e-50  1.194196e-47
148   Other - Whole  -5.83008235  5.540003e-09  8.476204e-07
149   ProSc - Whole  12.11160070  9.168720e-34  1.402814e-31
150   PubAd - Whole  16.31467944  7.761254e-60  1.187472e-57
151   Rtail - Whole -22.25827122 9.379515e-110 1.435066e-107
152   Trans - Whole  -3.84308098  1.214994e-04  1.858940e-02
153   Utils - Whole  15.65709670  2.971561e-55  4.546489e-53
```

Females 2009

```
dunnTest(HRLYEARN ~ NAICS_18short, data = data.all.09fem, method = "bonferroni") # Pos
t Hoc
Dunn (1964) Kruskal-Wallis multiple comparison
```

p-values adjusted with the Bonferroni method.

	Comparison	Z	P.unadj	P.adj
1	AcFood - Agri	-1.985620744	4.707545e-02	1.000000e+00
2	AcFood - Const	-16.896256590	4.793910e-64	7.334683e-62
3	Agri - Const	-9.337097328	9.901124e-21	1.514872e-18
4	AcFood - Educa	-61.060837033	0.000000e+00	0.000000e+00
5	Agri - Educa	-21.506799824	1.344671e-102	2.057346e-100
6	Const - Educa	-13.038381009	7.401198e-39	1.132383e-36
7	AcFood - Finan	-39.411403383	0.000000e+00	0.000000e+00
8	Agri - Finan	-15.059728267	2.980094e-51	4.559544e-49
9	Const - Finan	-5.026972405	4.982840e-07	7.623746e-05
10	Educa - Finan	14.394074379	5.637458e-47	8.625310e-45
11	AcFood - Fores	-24.157639898	6.206493e-129	9.495934e-127
12	Agri - Fores	-16.080780293	3.479684e-58	5.323916e-56
13	Const - Fores	-8.640667679	5.588545e-18	8.550474e-16
14	Educa - Fores	-0.120048644	9.044446e-01	1.000000e+00
15	Finan - Fores	-6.253140174	4.022808e-10	6.154896e-08
16	AcFood - Health	-57.107896437	0.000000e+00	0.000000e+00
17	Agri - Health	-17.673558052	6.703248e-70	1.025597e-67
18	Const - Health	-7.963631139	1.670628e-15	2.556061e-13
19	Educa - Health	13.087244905	3.894806e-39	5.959053e-37
20	Finan - Health	-5.148461423	2.626318e-07	4.018266e-05
21	Fores - Health	4.455025911	8.388306e-06	1.283411e-03
22	AcFood - Info	-23.564888366	8.834307e-123	1.351649e-120
23	Agri - Info	-9.725970343	2.336668e-22	3.575102e-20
24	Const - Info	1.197730151	2.310221e-01	1.000000e+00
25	Educa - Info	22.052954487	8.948676e-108	1.369147e-105
26	Finan - Info	9.230757231	2.687254e-20	4.111498e-18
27	Fores - Info	10.953631840	6.383568e-28	9.766860e-26
28	Health - Info	14.908353697	2.908353e-50	4.449780e-48
29	AcFood - ManuD	-25.543938342	6.411814e-144	9.810075e-142
30	Agri - ManuD	-12.707606357	5.365165e-37	8.208703e-35
31	Const - ManuD	-3.121239490	1.800915e-03	2.755400e-01
32	Educa - ManuD	11.651649662	2.250593e-31	3.443408e-29
33	Finan - ManuD	1.767411970	7.715926e-02	1.000000e+00
34	Fores - ManuD	6.746598558	1.513513e-11	2.315674e-09
35	Health - ManuD	5.259424958	1.445066e-07	2.210951e-05
36	Info - ManuD	-5.478713176	4.284302e-08	6.554982e-06
37	AcFood - ManuN	-18.843057870	3.350383e-79	5.126086e-77
38	Agri - ManuN	-7.582675801	3.384999e-14	5.179049e-12
39	Const - ManuN	3.761360768	1.689915e-04	2.585570e-02
40	Educa - ManuN	25.761554033	2.392682e-146	3.660803e-144
41	Finan - ManuN	12.974313569	1.711225e-38	2.618174e-36
42	Fores - ManuN	13.025291074	8.786538e-39	1.344340e-36
43	Health - ManuN	19.032607263	9.158076e-81	1.401186e-78
44	Info - ManuN	3.534305638	4.088480e-04	6.255375e-02
45	ManuD - ManuN	8.470545525	2.442464e-17	3.736970e-15
46	AcFood - Mngt	-12.334117586	5.933538e-35	9.078314e-33
47	Agri - Mngt	-4.702858954	2.565436e-06	3.925118e-04
48	Const - Mngt	6.986775199	2.812764e-12	4.303529e-10
49	Educa - Mngt	29.077420051	6.929006e-186	1.060138e-183
50	Finan - Mngt	17.041926314	4.012935e-65	6.139790e-63
51	Fores - Mngt	15.554720619	1.477988e-54	2.261321e-52
52	Health - Mngt	22.981774973	7.092856e-117	1.085207e-114
53	Info - Mngt	7.810681691	5.687950e-15	8.702563e-13
54	ManuD - Mngt	12.043781265	2.091446e-33	3.199912e-31
55	ManuN - Mngt	4.399430035	1.085356e-05	1.660594e-03
56	AcFood - Other	-18.525675772	1.281837e-76	1.961210e-74
57	Agri - Other	-7.131407581	9.934765e-13	1.520019e-10
58	Const - Other	4.469272880	7.848595e-06	1.200835e-03
59	Educa - Other	27.975983928	3.184789e-172	4.872728e-170
60	Finan - Other	14.462780811	2.082127e-47	3.185655e-45
61	Fores - Other	13.698492970	1.036534e-42	1.585896e-40
62	Health - Other	21.162323338	2.124901e-99	3.251099e-97
63	Info - Other	4.542624196	5.555824e-06	8.500411e-04
64	ManuD - Other	9.431718667	4.034237e-21	6.172382e-19
65	ManuN - Other	0.890655284	3.731141e-01	1.000000e+00
66	Mngt - Other	-3.666830756	2.455752e-04	3.757301e-02

67	AcFood - ProSc	-35.257291866	2.652663e-272	4.058574e-270
68	Agri - ProSc	-15.510087612	2.964741e-54	4.536053e-52
69	Const - ProSc	-5.972606846	2.334921e-09	3.572429e-07
70	Educa - ProSc	9.933438805	2.978042e-23	4.556405e-21
71	Finan - ProSc	-1.897049885	5.782135e-02	1.000000e+00
72	Fores - ProSc	5.002237294	5.666878e-07	8.670323e-05
73	Health - ProSc	1.941657192	5.217861e-02	1.000000e+00
74	Info - ProSc	-9.918665741	3.453416e-23	5.283726e-21
75	ManuD - ProSc	-3.087514247	2.018381e-03	3.088123e-01
76	ManuN - ProSc	-13.276718628	3.159383e-40	4.833856e-38
77	Mngt - ProSc	-17.024973214	5.361700e-65	8.203402e-63
78	Other - ProSc	-14.562972805	4.831039e-48	7.391490e-46
79	AcFood - PubAd	-56.367605614	0.000000e+00	0.000000e+00
80	Agri - PubAd	-22.208002465	2.874351e-109	4.397757e-107
81	Const - PubAd	-14.056727326	7.005158e-45	1.071789e-42
82	Educa - PubAd	-2.812295477	4.918929e-03	7.525962e-01
83	Finan - PubAd	-15.409572636	1.411422e-53	2.159475e-51
84	Fores - PubAd	-1.072354845	2.835607e-01	1.000000e+00
85	Health - PubAd	-14.032300186	9.889129e-45	1.513037e-42
86	Info - PubAd	-22.544431494	1.522871e-112	2.329993e-110
87	ManuD - PubAd	-12.843444189	9.361507e-38	1.432311e-35
88	ManuN - PubAd	-26.014201105	3.421035e-149	5.234183e-147
89	Mngt - PubAd	-29.229467676	8.190471e-188	1.253142e-185
90	Other - PubAd	-27.990479751	2.121725e-172	3.246240e-170
91	ProSc - PubAd	-11.336005774	8.702654e-30	1.331506e-27
92	AcFood - Rtail	-7.113867123	1.128358e-12	1.726388e-10
93	Agri - Rtail	-0.584576872	5.588323e-01	1.000000e+00
94	Const - Rtail	13.995120227	1.669471e-44	2.554291e-42
95	Educa - Rtail	61.462317073	0.000000e+00	0.000000e+00
96	Finan - Rtail	36.746428376	1.325821e-295	2.028506e-293
97	Fores - Rtail	21.891263318	3.146718e-106	4.814479e-104
98	Health - Rtail	58.354280854	0.000000e+00	0.000000e+00
99	Info - Rtail	19.836233795	1.449169e-87	2.217229e-85
100	ManuD - Rtail	22.398234300	4.094884e-111	6.265173e-109
101	ManuN - Rtail	14.887301332	3.985338e-50	6.097567e-48
102	Mngt - Rtail	8.274749282	1.287262e-16	1.969511e-14
103	Other - Rtail	14.419491780	3.902052e-47	5.970139e-45
104	ProSc - Rtail	32.289056166	9.963572e-229	1.524427e-226
105	PubAd - Rtail	55.347377569	0.000000e+00	0.000000e+00
106	AcFood - Trans	-23.589754785	4.910168e-123	7.512557e-121
107	Agri - Trans	-11.384104313	5.018119e-30	7.677722e-28
108	Const - Trans	-1.443613907	1.488476e-01	1.000000e+00
109	Educa - Trans	14.507632366	1.083979e-47	1.658488e-45
110	Finan - Trans	4.237532037	2.259902e-05	3.457650e-03
111	Fores - Trans	8.245540493	1.644153e-16	2.515555e-14
112	Health - Trans	8.100630424	5.467511e-16	8.365292e-14
113	Info - Trans	-3.341542576	8.331424e-04	1.274708e-01
114	ManuD - Trans	1.986371906	4.699204e-02	1.000000e+00
115	ManuN - Trans	-6.412248098	1.433892e-10	2.193855e-08
116	Mngt - Trans	-10.123598590	4.341533e-24	6.642545e-22
117	Other - Trans	-7.348328609	2.007006e-13	3.070719e-11
118	ProSc - Trans	5.373410904	7.726099e-08	1.182093e-05
119	PubAd - Trans	15.556446045	1.438688e-54	2.201192e-52
120	Rtail - Trans	-20.305897725	1.140411e-91	1.744829e-89
121	AcFood - Utils	-18.935697807	5.794283e-80	8.865253e-78
122	Agri - Utils	-13.932402676	4.025811e-44	6.159491e-42
123	Const - Utils	-7.239059407	4.518071e-13	6.912649e-11
124	Educa - Utils	-0.090061123	9.282386e-01	1.000000e+00
125	Finan - Utils	-4.928881381	8.270176e-07	1.265337e-04
126	Fores - Utils	0.003041285	9.975734e-01	1.000000e+00
127	Health - Utils	-3.453328170	5.537149e-04	8.471839e-02
128	Info - Utils	-8.758871192	1.972153e-18	3.017394e-16
129	ManuD - Utils	-5.503301605	3.727439e-08	5.702982e-06
130	ManuN - Utils	-10.439980964	1.628434e-25	2.491503e-23
131	Mngt - Utils	-12.560975566	3.460396e-36	5.294406e-34
132	Other - Utils	-10.939481014	7.462587e-28	1.141776e-25
133	ProSc - Utils	-3.998587698	6.372157e-05	9.749400e-03
134	PubAd - Utils	0.849056918	3.958496e-01	1.000000e+00
135	Rtail - Utils	-17.044441720	3.843970e-65	5.881274e-63

```
136   Trans - Utils  -6.710044392  1.945652e-11  2.976848e-09
137  AcFood - Whole -18.842558230  3.382164e-79  5.174710e-77
138    Agri - Whole  -9.625102889  6.264420e-22  9.584563e-20
139   Const - Whole   0.215451142  8.294156e-01  1.000000e+00
140   Educa - Whole  15.132132029  9.942144e-52  1.521148e-49
141   Finan - Whole   5.948655855  2.703534e-09  4.136407e-07
142   Fores - Whole   9.279768480  1.698472e-20  2.598663e-18
143  Health - Whole   9.450019472  3.387677e-21  5.183145e-19
144    Info - Whole  -1.057613656  2.902316e-01  1.000000e+00
145   ManuD - Whole   3.650298082  2.619362e-04  4.007623e-02
146   ManuN - Whole  -3.896850841  9.745160e-05  1.491009e-02
147    Mngt - Whole  -7.427618872  1.105700e-13  1.691721e-11
148   Other - Whole  -4.692802384  2.694878e-06  4.123163e-04
149   ProSc - Whole   6.898546308  5.253736e-12  8.038217e-10
150   PubAd - Whole  16.122333990  1.777666e-58  2.719828e-56
151   Rtail - Whole -15.669353674  2.450587e-55  3.749398e-53
152   Trans - Whole   1.826593310  6.776094e-02  1.000000e+00
153   Utils - Whole   7.652828804  1.966056e-14  3.008065e-12
```

Females 2019

```
> dunnTest(HRLYEARN ~ NAICS_18short, data = data.all.19fem, method = "bonferroni") # P
ost Hoc
Dunn (1964) Kruskal-Wallis multiple comparison
  p-values adjusted with the Bonferroni method.

        Comparison            Z       P.unadj         P.adj
1    AcFood - Agri  -5.14494721  2.675960e-07  4.094219e-05
2   AcFood - Const -21.42172806 8.382100e-102  1.282461e-99
3     Agri - Const  -9.45537683  3.218605e-21  4.924466e-19
4   AcFood - Educa -57.63233126  0.000000e+00  0.000000e+00
5     Agri - Educa -19.20899094  3.112572e-82  4.762235e-80
6    Const - Educa -10.68221908  1.232887e-26  1.886318e-24
7   AcFood - Finan -39.05486409  0.000000e+00  0.000000e+00
8     Agri - Finan -14.20139142  8.981129e-46  1.374113e-43
9    Const - Finan  -4.43159366  9.353916e-06  1.431149e-03
10   Educa - Finan   9.23760420  2.520780e-20  3.856794e-18
11  AcFood - Fores -25.30349906 2.923191e-141 4.472483e-139
12    Agri - Fores -15.29597062  8.133906e-53  1.244488e-50
13   Const - Fores  -8.16624953  3.181249e-16  4.867311e-14
14   Educa - Fores  -1.84370034  6.522682e-02  1.000000e+00
15   Finan - Fores  -5.94019857  2.846770e-09  4.355559e-07
16 AcFood - Health -50.12363773  0.000000e+00  0.000000e+00
17   Agri - Health -14.00958999  1.361889e-44  2.083691e-42
18  Const - Health  -3.49363784  4.764869e-04  7.290250e-02
19  Educa - Health  16.35206621  4.204771e-60  6.433299e-58
20  Finan - Health   2.32463304  2.009159e-02  1.000000e+00
21  Fores - Health   7.29550157  2.975485e-13  4.552492e-11
22   AcFood - Info -19.42144023  5.084388e-84  7.779114e-82
23     Agri - Info  -6.06186191  1.345547e-09  2.058687e-07
24    Const - Info   5.32204545  1.026069e-07  1.569886e-05
25    Educa - Info  21.59200471 2.135462e-103 3.267256e-101
26    Finan - Info  12.54863093  4.044422e-36  6.187966e-34
27    Fores - Info  12.98943131  1.404669e-38  2.149144e-36
28   Health - Info  12.94631456  2.465072e-38  3.771560e-36
29  AcFood - ManuD -22.35663975 1.040586e-110 1.592097e-108
30    Agri - ManuD  -9.67165790  3.978794e-22  6.087556e-20
31   Const - ManuD  -0.08349097  9.334612e-01  1.000000e+00
32   Educa - ManuD  11.00747225  3.517339e-28  5.381529e-26
33   Finan - ManuD   4.48443566  7.310716e-06  1.118540e-03
34   Fores - ManuD   8.22177175  2.005185e-16  3.067934e-14
35  Health - ManuD   3.53116854  4.137280e-04  6.330039e-02
36    Info - ManuD  -5.57700586  2.446936e-08  3.743812e-06
37  AcFood - ManuN -18.87458996  1.845506e-79  2.823625e-77
38    Agri - ManuN  -5.97313133  2.327424e-09  3.560959e-07
39   Const - ManuN   5.30836685  1.106118e-07  1.692361e-05
40   Educa - ManuN  21.09338908  9.146823e-99  1.399464e-96
41   Finan - ManuN  12.34636009  5.096560e-35  7.797738e-33
```

42	Fores - ManuN	12.93258380	2.947443e-38	4.509588e-36
43	Health - ManuN	12.64247815	1.231116e-36	1.883607e-34
44	Info - ManuN	0.06134563	9.510840e-01	1.000000e+00
45	ManuD - ManuN	5.55760060	2.735083e-08	4.184678e-06
46	AcFood - Mngt	-12.30549576	8.461527e-35	1.294614e-32
47	Agri - Mngt	-2.73121658	6.310099e-03	9.654451e-01
48	Const - Mngt	8.91338940	4.949577e-19	7.572853e-17
49	Educa - Mngt	24.75208163	2.944351e-135	4.504858e-133
50	Finan - Mngt	16.42017372	1.371734e-60	2.098754e-58
51	Fores - Mngt	15.63711618	4.067349e-55	6.223044e-53
52	Health - Mngt	17.12601899	9.493258e-66	1.452468e-63
53	Info - Mngt	4.49958500	6.808625e-06	1.041720e-03
54	ManuD - Mngt	9.24265333	2.404590e-20	3.679023e-18
55	ManuN - Mngt	4.37522225	1.213088e-05	1.856025e-03
56	AcFood - Other	-17.50080079	1.412689e-68	2.161415e-66
57	Agri - Other	-4.92300233	8.522644e-07	1.303964e-04
58	Const - Other	6.80003818	1.045914e-11	1.600249e-09
59	Educa - Other	23.95757558	7.704329e-127	1.178762e-124
60	Finan - Other	14.58501715	3.498341e-48	5.352461e-46
61	Fores - Other	14.15080730	1.846256e-45	2.824771e-43
62	Health - Other	15.37850230	2.281646e-53	3.490918e-51
63	Info - Other	1.73256138	8.317366e-02	1.000000e+00
64	ManuD - Other	7.10138435	1.235133e-12	1.889754e-10
65	ManuN - Other	1.64067180	1.008656e-01	1.000000e+00
66	Mngt - Other	-2.91288590	3.581054e-03	5.479013e-01
67	AcFood - ProSc	-37.65046582	0.000000e+00	0.000000e+00
68	Agri - ProSc	-14.69275448	7.173738e-49	1.097582e-46
69	Const - ProSc	-5.27368178	1.337137e-07	2.045819e-05
70	Educa - ProSc	6.91666213	4.624092e-12	7.074860e-10
71	Finan - ProSc	-1.36482736	1.723073e-01	1.000000e+00
72	Fores - ProSc	5.07852971	3.803669e-07	5.819614e-05
73	Health - ProSc	-3.77363412	1.608867e-04	2.461566e-02
74	Info - ProSc	-13.14343985	1.855941e-39	2.839590e-37
75	ManuD - ProSc	-5.34746015	8.919704e-08	1.364715e-05
76	ManuN - ProSc	-12.95083633	2.324082e-38	3.555846e-36
77	Mngt - ProSc	-16.85761617	9.223477e-64	1.411192e-61
78	Other - ProSc	-15.08115654	2.154614e-51	3.296560e-49
79	AcFood - PubAd	-54.77966723	0.000000e+00	0.000000e+00
80	Agri - PubAd	-20.77484012	7.310708e-96	1.118538e-93
81	Const - PubAd	-12.91268538	3.817516e-38	5.840800e-36
82	Educa - PubAd	-4.73453857	2.195540e-06	3.359176e-04
83	Finan - PubAd	-12.28513251	1.088665e-34	1.665657e-32
84	Fores - PubAd	-0.23288912	8.158475e-01	1.000000e+00
85	Health - PubAd	-18.58444733	4.294004e-77	6.569827e-75
86	Info - PubAd	-23.41409141	3.071029e-121	4.698675e-119
87	ManuD - PubAd	-13.28431993	2.854400e-40	4.367231e-38
88	ManuN - PubAd	-22.94614978	1.609855e-116	2.463078e-114
89	Mngt - PubAd	-26.35826474	4.126591e-153	6.313684e-151
90	Other - PubAd	-25.60555774	1.322918e-144	2.024064e-142
91	ProSc - PubAd	-10.02704896	1.159297e-23	1.773724e-21
92	AcFood - Rtail	-7.03843286	1.944139e-12	2.974533e-10
93	Agri - Rtail	2.27899522	2.266735e-02	1.000000e+00
94	Const - Rtail	18.13869260	1.577390e-73	2.413407e-71
95	Educa - Rtail	56.63381089	0.000000e+00	0.000000e+00
96	Finan - Rtail	35.83075647	3.667394e-281	5.611113e-279
97	Fores - Rtail	22.84777930	1.537603e-115	2.352532e-113
98	Health - Rtail	48.88123145	0.000000e+00	0.000000e+00
99	Info - Rtail	15.33965892	4.153717e-53	6.355187e-51
100	ManuD - Rtail	18.99882977	1.743895e-80	2.668159e-78
101	ManuN - Rtail	14.85048710	6.906500e-50	1.056695e-47
102	Mngt - Rtail	8.26695353	1.374255e-16	2.102610e-14
103	Other - Rtail	13.28954094	2.662064e-40	4.072958e-38
104	ProSc - Rtail	34.39185505	3.337551e-259	5.106453e-257
105	PubAd - Rtail	52.89214308	0.000000e+00	0.000000e+00
106	AcFood - Trans	-22.09996952	3.162804e-108	4.839091e-106
107	Agri - Trans	-8.46672954	2.523792e-17	3.861402e-15
108	Const - Trans	1.87623214	6.062341e-02	1.000000e+00
109	Educa - Trans	15.05671560	3.119019e-51	4.772099e-49
110	Finan - Trans	7.45353065	9.087488e-14	1.390386e-11

111	Fores - Trans	10.08724002	6.291357e-24	9.625776e-22
112	Health - Trans	6.88159837	5.918468e-12	9.055257e-10
113	Info - Trans	-3.73474661	1.879043e-04	2.874936e-02
114	ManuD - Trans	2.01810756	4.358006e-02	1.000000e+00
115	ManuN - Trans	-3.73559708	1.872703e-04	2.865236e-02
116	Mngt - Trans	-7.75209259	9.039045e-15	1.382974e-12
117	Other - Trans	-5.37597632	7.616880e-08	1.165383e-05
118	ProSc - Trans	8.25654839	1.499454e-16	2.294165e-14
119	PubAd - Trans	17.25460932	1.033182e-66	1.580768e-64
120	Rtail - Trans	-18.43518357	6.858557e-76	1.049359e-73
121	AcFood - Utils	-20.57540441	4.559119e-94	6.975452e-92
122	Agri - Utils	-14.44858620	2.558844e-47	3.915032e-45
123	Const - Utils	-8.58478224	9.100865e-18	1.392432e-15
124	Educa - Utils	-3.88240058	1.034303e-04	1.582484e-02
125	Finan - Utils	-6.80685231	9.975725e-12	1.526286e-09
126	Fores - Utils	-2.15772792	3.094899e-02	1.000000e+00
127	Health - Utils	-7.71198020	1.238804e-14	1.895369e-12
128	Info - Utils	-12.02601158	2.593916e-33	3.968692e-31
129	ManuD - Utils	-8.60810398	7.427887e-18	1.136467e-15
130	ManuN - Utils	-12.00688268	3.269381e-33	5.002153e-31
131	Mngt - Utils	-14.10129994	3.728379e-45	5.704420e-43
132	Other - Utils	-12.87224600	6.449892e-38	9.868334e-36
133	ProSc - Utils	-6.19468598	5.840134e-10	8.935406e-08
134	PubAd - Utils	-2.37944272	1.733884e-02	1.000000e+00
135	Rtail - Utils	-18.71304134	3.876108e-78	5.930445e-76
136	Trans - Utils	-9.94913482	2.543851e-23	3.892092e-21
137	AcFood - Whole	-20.45485120	5.438795e-93	8.321357e-91
138	Agri - Whole	-8.82995848	1.047146e-18	1.602133e-16
139	Const - Whole	0.75475045	4.503987e-01	1.000000e+00
140	Educa - Whole	11.67643989	1.681934e-31	2.573359e-29
141	Finan - Whole	5.35955357	8.342785e-08	1.276446e-05
142	Fores - Whole	8.78078888	1.623320e-18	2.483680e-16
143	Health - Whole	4.52195934	6.126983e-06	9.374284e-04
144	Info - Whole	-4.46459616	8.021991e-06	1.227365e-03
145	ManuD - Whole	0.85558579	3.922269e-01	1.000000e+00
146	ManuN - Whole	-4.46148723	8.139280e-06	1.245310e-03
147	Mngt - Whole	-8.10040063	5.477849e-16	8.381109e-14
148	Other - Whole	-5.93917830	2.864542e-09	4.382749e-07
149	ProSc - Whole	6.17172622	6.754839e-10	1.033490e-07
150	PubAd - Whole	13.86088228	1.093222e-43	1.672630e-41
151	Rtail - Whole	-17.14101835	7.335462e-66	1.122326e-63
152	Trans - Whole	-1.05797559	2.900666e-01	1.000000e+00
153	Utils - Whole	9.05325046	1.387745e-19	2.123250e-17