Research Replicability and Workflow Management

Yann Dorville
UQAM

September 20, 2024

1 Research Question

The wage gap has been widely studied and documented. In this analysis, we will explore it both visually and through a silly regression.

2 Program Overview

1. Load and clean data
2. Summary statistics
3. Plot
4. Regression

3 Detailed Execution

3.1 Load and clean data

Load the 2016 Census of Population.

Subsetting

Select the columns indicated in table 1 and recode them as described.
Variables Field of study, occupation, industry, and labor force status do not require recoding at this stage.

Filtering

Once the data is loaded, apply the following filters:
- Keep only workers who are employed and at work (LFACT = 1).
- Remove rows with missing values. This step should be done after recoding.
- For the industry, occupation, and field of study variables, exclude values 88 and 99.
- For the income variable, exclude values 88888888 and 99999999.

Transformation

Convert income in thousands
Then, compute the occupation×sex specific median.
Finally, create additional columns to label variables Field of Study, Industry, and Occupation, as per table 2, table 3, and table 4.

Table 1: Variable Recoding Specifications

Variable	**Variable ID**	**Recode**	**Value**
Age	AGEGRP	$x < 7$	NA
		$7 \leq x \leq 9$	"Young"
		$10 \leq x \leq 12$	"Middle"
		$13 \leq x \leq 16$	"Older"
		$x \geq 17$	NA
Sex	Sex	$x = 1$	"Female"
		$x = 2$	"Male"
Visible Minority	VisMin	$x < 13$	"Yes"
		$x = 13$	"No"
		$x > 13$	NA
Education	HDGREE	$x = 1$	"None"
		$2 \leq x \leq 8$	"Medium"
		$9 \leq x \leq 13$	"High"
		$x \geq 14$	NA
Income	EmpIn	-	-
Field of Study	CIP2011	-	-
Labor Status	LFACT	-	-
Occupation	NOCS	-	-
Industry	NAICS	-	-

Table 2: NOCS Classification

NOCS	Label
1	Management
2	Business & Finance
3	Sciences
4	Health
5	Social and Education
6	Arts
7	Sales and Services
8	Trades and Transport
9	Primary Industry
10	Manufacturing

3.2 Summary statistics

Compute the occupation×sex specific average income and standard deviation. Report the data in a table.

Table 5: Mean and Standard Deviation of Income by Occupation and Gender

Occupation	Income Statistics	
	Female	Male
Arts	32.4 (±33.0)	43.6 (±56.6)
Business & Finance	47.4 (±37.1)	75.4 (±111)
Health	54.2 (±40.7)	96.2 (±128.5)
Management	70.7 (±61.7)	101.6 (±130)
Manufacturing	33.8 (±25.7)	55.7 (±45.5)
Primary Industry	21.2 (±22.2)	50.6 (±58.5)
Sales and Services	25.1 (±25.6)	38.6 (±53.7)
Sciences	66.8 (±47.3)	83.1 (±77.5)
Social and Education	49.4 (±40.3)	83.2 (±94.0)
Trades and Transport	34.1 (±27.0)	53.5 (±40.7)

3.3 Plot

Create a facetted histogram displaying the distribution of female and male workers in each occupation. That is, for each occupation, overlap two histograms, each corresponding to the income distribution of a gender. Indicate the occupation×gender specific median with a vertical line in each panel.

Table 3: NAICS Classification

NAICS	Label
1	Agriculture
2	Mining & Extraction
3	Utilities
4	Construction
5	Manufacturing
6	Wholesale
7	Retail
8	Transportation
9	Information and Cultural Industries
10	Finance
11	Real Estate
12	Professional Services
13	Administrative and Support
14	Education
15	Health Care
16	Arts & Entertainment
17	Accommodation and Food
18	Other Services
19	Public Administration

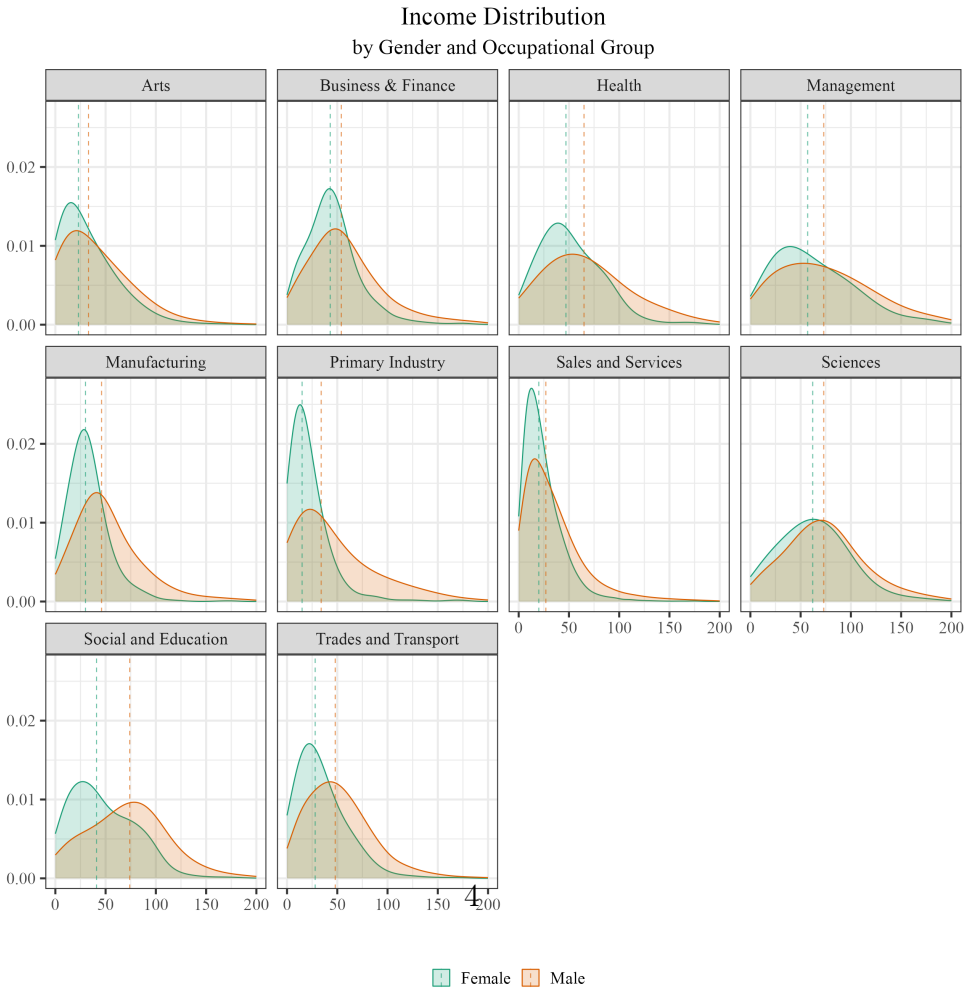

Table 4: Field of Study Classification

Field of Study	**Label**
1	Education
2	Communications
3	Humanities
4	Social Sciences & Law
5	Business & Management
6	Life Sciences
7	Mathematics & Computer Sciences
8	Architecture & Engineering
9	Agriculture
10	Health
11	Transportation Services
12	Other
13	No Degree

3.4 Regression

Estimate the following model:

$$income = \mu + \beta Sex + \gamma X + \varepsilon$$

Where X contains controls listed in 6

Table 6: Control variables

Age
Education
Field of Study
Visible Minority
Occupation
Industry

The results indicate that the factors controlled for in this analysis are not sufficient to explain the gender differences in income.

Table 7: Regression Summary (1/2)

Variable	**Estimate**	**Std. Error**	**P-value**
Intercept	32.0870	1.2665	< 2e-16 ***
SEX			
Male	20.1871	0.2416	< 2e-16 ***
OCCUPATION			
Business and Finance	11.5574	0.7273	< 2e-16 ***
Health	29.5928	0.8958	< 2e-16 ***
Management	40.7642	0.7493	< 2e-16 ***
Manufacturing	4.0053	0.8869	6.30e-06 ***
Primary Industry	10.0129	1.0850	< 2e-16 ***
Sales and Services	7.5037	0.7306	< 2e-16 ***
Sciences	17.0535	0.7911	< 2e-16 ***
Social and Education	20.2722	0.7678	< 2e-16 ***
Trades and Transport	5.6019	0.7743	4.66e-13 ***
FIELD OF STUDY			
Architecture and Engineering	12.1840	0.9577	< 2e-16 ***
Business and Management	11.6885	0.9606	< 2e-16 ***
Communications	-3.3413	1.1357	0.00326 **
Education	0.8645	1.1049	0.43398
Health	10.9971	1.0307	< 2e-16 ***
Humanities	-6.2693	1.0837	7.26e-09 ***
Life Sciences	2.4814	1.1307	0.02820 *
Mathematics and Computer Sciences	5.4791	1.1022	6.66e-07 ***
No Degree	3.7068	0.9404	8.09e-05 ***
Social Sciences and Law	5.7321	0.9988	9.52e-09 ***
Transportation Services	7.7630	1.0477	1.27e-13 ***
AGE			
Older	10.0387	0.2318	< 2e-16 ***
Young	-18.5711	0.2822	< 2e-16 ***
MINORITY			
Yes	-14.5972	0.2609	< 2e-16 ***
EDUCATION			
Medium	-25.8503	0.2921	< 2e-16 ***
None	-32.0685	0.5012	< 2e-16 ***

Table 8: Regression Summary (2/2)

Variable	**Estimate**	**Std. Error**	**P-value**
INDUSTRY			
Administrative and Support	1.5159	0.6514	0.01996 *
Agriculture	-12.4117	0.9247	< 2e-16 ***
Arts and Entertainment	1.0118	0.9028	0.26240
Construction	14.2683	0.6270	< 2e-16 ***
Education	4.4944	0.6636	1.27e-11 ***
Finance	39.7469	0.6393	< 2e-16 ***
Health Care	3.7464	0.6116	9.04e-10 ***
Information and Cultural Industries	25.5525	0.8192	< 2e-16 ***
Manufacturing	19.3766	0.6005	< 2e-16 ***
Mining and Extraction	85.0540	1.0045	< 2e-16 ***
Other Services	1.9258	0.6536	0.00321 **
Professional Services	22.5766	0.6155	< 2e-16 ***
Public Administration	20.9234	0.6221	< 2e-16 ***
Real Estate	15.8009	0.9062	< 2e-16 ***
Retail	2.9646	0.5001	3.06e-09 ***
Transportation	14.2057	0.6653	< 2e-16 ***
Utilities	53.4601	1.2731	< 2e-16 ***
Wholesale	23.5877	0.6761	< 2e-16 ***