Understanding Wage Variation in Canada:

Personal and Situational Predictors of Higher Earnings

Mengqing (Annie) Wu (1011429456), Belinda Kwan (999206540), Nahel Sinan (1007978093)

Group 5, Final Project

INF1340 Programming for Data Science

Prof. Maher Elshakankiri

November 26, 2025

Table of Contents

1.0 Introduction

Concerns about the Canadian job market have intensified in recent years, with many people finding it increasingly difficult to maintain financial stability. To start, unemployment has reached 7.1% overall (Rabinovitch, 2025) and 14.5% among youth (Statistics Canada, 2025a). At the same time, an RBC survey reports that half of Canadians can no longer maintain their standard of living and that many are relying on savings to cover essential expenses (The Canadian Press, 2025). As students preparing to return to the workforce, these conditions shaped how we approached our project.

Considering the broader economic pressures facing workers, we turned our attention to the role higher wages play in creating financial stability. Stronger earnings can help buffer against rising costs and periods of unemployment, which make it important to understand how personal context relates to higher wages. Our report therefore explores what personal and situational characteristics are associated with higher full-time wages in Canada. This focus is supported by recent findings from the Robert Half 2026 Salary Guide (Robert Half, 2025), which reports strong salary expectations among job seekers and an increasing emphasis on skills, qualifications, and individual circumstances among employers. Such trend reports highlight the growing relevance of personal context in shaping wage outcomes.

Our research focuses on active full-time employees not on leave and not in school: a group whose earnings, hours, and employment conditions tend to be more stable than part-time and self-employed workers. Concentrating on this group creates a more consistent basis for analyzing stable and higher base pay.

It should likewise be noted that we chose not to include detailed industry or occupational classifications as predictors or control variables. Adding these would shift the analysis toward sector-level wage differences rather than personal context, since industry often captures large structural pay gaps that can overshadow individual characteristics. We did retain broad contextual indicators such as public versus private sector employment and union status, as these provide meaningful labour market context without turning the analysis into a detailed sectoral comparison. Beyond these high-level distinctions, we avoided granular industry variables to keep

the focus on personal and situational factors. A sectoral analysis remains a valuable next step for future work.

To explore these relationships, we statistically analyze the Public Use Microdata File (PUMF) from Statistics Canada's September 2025 Labour Force Survey (Statistics Canada, 2025b). In addition to examining associations between variables, we also build predictive models that estimate the likelihood of earning a higher hourly wage. These models offer an initial view of how personal context may be associated with higher earnings and help highlight which characteristics appear most strongly linked to stronger wage outcomes. While our predictive study is only a starting point, the patterns it reveals could support future work that examines socioeconomic inequities in the labour market, including studies that explore the causal pathways through which personal and contextual factors shape access to higher-paying work.

2.0 Initial Data Exploration & Preparation

2.1 Overview of Dataset

The Labour Force Survey (LFS) is a monthly household survey measuring labour market indicators such as employment and unemployment rate, as well as employment estimates by industry, occupation, and more. Its demographic categories include gender, education, and geographic location. More detailed information, such as part-time status, union status, and reasons for respondents leaving their last job, are also available. The September 2025 data set includes **112,927** observations of **59** variables (not counting the record ID, `REC_NUM`, which serves as the index).

15 out of the 59 variables were **numerical**, while the rest were categorical. The numerical variables largely pertained to durations and dollar amounts, with the following key themes:

- Absence from Work
- Hours Worked (All Jobs vs. Main Job, Usual vs. Actual)
- Overtime (Paid vs. Unpaid)
- Unemployment
- Joblessness
- Job Tenure (Current vs. Previous)

2.2 Variable Selection

Our first step was to reduce the dataset by selecting only the columns relevant to our research question. This initial cut removed variables that did not contribute to understanding hourly wages and kept only the measures needed to define our analytical framework.

We included several context variables at this stage, such as labour force status, full-time or part-time classification, and worker class. These variables were not used as predictor or outcome variables but instead helped us identify the subset of respondents whose employment conditions matched the scope of our analysis. The usual hourly earnings variable was also retained as a potential outcome or transformable predictor depending on modelling needs.

We then selected the personal and situational characteristics most relevant to wage differences, including age group, education, gender, marital status, and hours worked, along with job-related controls such as overtime, job permanency, tenure, union status, and multiple-jobholding.

Table 1. Initial Variable Selection

Type of Variable	**Code**	**Description**
Context Variables	`LFSSTAT`	Labour force status
	`FTPTMAIN`	Full- or part-time status at main job
	`COWMAIN`	Class of worker, main job
	`SCHOOLN`	School attendance indicator
Potential Dependent Variable	`HRLYEARN`	Usual hourly wages ($)
Predictors	`AGE_12`	Age group
	`AHRSMAIN`	Actual hours worked per week
	`EDUC`	Highest education level
	`GENDER`	Gender
	`MARSTAT`	Marital status
Control Variables	`MJH`	Single or multiple jobholder

	PAIDOT	Paid overtime hours
	PERMTEMP	Job permanency
	TENURE	Job tenure with current employer (months)
	UNION	Union status
	UNPAIDOT	Unpaid overtime hours
	XTRAHRS	Number of extra hours worked

2.3 Filtering for Research Scope

Next, we filtered the dataset to include only respondents actively employed, not self-employed, working full-time in their main job, and not attending school. This criteria allowed us to focus on a consistent group of full-time employees whose earnings are stable and comparable.

Once these conditions were applied, most of the context variables used for screening (employment status, full-time status, and school attendance) contained a single value for all remaining cases. To avoid redundancy and maintain a clean dataset, we removed these variables after filtering. The one exception, worker class (COWMAIN), continued to distinguish between private and public sector employees, and was therefore retained.

After restricting the data set to our target population, the analytic sample consisted of **40,868** observations and **14** variables, which is adequate for producing stable and reliable statistical insights.

2.4 Missing & Duplicate Data

After variable selection and row filtering, **no missing values** and **1,236 duplicate rows** were found, the latter being an acceptable amount for population survey data given that the record identifier was indexed. We also reviewed the index for duplicate identifiers and found none.

2.5 Data Transformation

The user guide accompanying the dataset notes that several labour market variables use implied decimals (namely those with hour- and dollar-based values), so we scaled the raw values

accordingly. Hourly wages were converted from **cents to dollars** and usual hours from **tenths to whole hours**. After applying these transformations, we verified that the resulting fields had the correct numeric data types.

For clarity, we added an underscore and uppercase T (`_T`) to all transformed variable codes (Table 2). We also generated a summary statistics comparison table of pre- and post-transformation variables to confirm that the correct transformations occurred (Table 3).

The original variables were removed afterwards, as they were no longer needed.

Table 2. Summary of Transformed Variables

Before	After	Variable Description
HRLYEARN	HRLYEARN_T	Usual hourly wages ($)
AHRSMAIN	AHRSMAIN_T	Actual hours worked per week
PAIDOT	PAIDOT_T	Paid overtime hours
UNPAIDOT	UNPAIDOT_T	Unpaid overtime hours
XTRAHRS	XTRAHRS_T	Number of extra hours worked

Table 3. Checking for Proper Transformation

Variable	Mean	Median	Std Dev	Min	Max	Count
HRLYEARN	3773.57	3300	1860.2	607	23791	40868
HRLYEARN_T	37.74	33	18.6	6.07	237.91	40868
AHRSMAIN	396.02	400	92.88	0	990	40868
AHRSMAIN_T	39.6	40	9.29	0	99	40868
PAIDOT	11.08	0	43.3	0	720	40868
PAIDOT_T	1.11	0	4.33	0	72	40868
UNPAIDOT	6.22	0	27.67	0	660	40868
UNPAIDOT_T	0.62	0	2.77	0	66	40868

XTRAHRS	17.3	0	50.66	0	720	40868
XTRAHRS_T	1.73	0	5.07	0	72	40868

2.6 Inconsistencies: Data Type Correction & Invalid Entries Check

We checked for two major types of inconsistencies: data type issues and invalid entries.

Initially, all dataset variables were stored as either `Int64` or `Float64`, which did not reflect their intended structures. To correct this, we assigned each variable to categorical, numeric, or binary groups based on the codebook and converted them accordingly. Categorical and binary fields were assigned the `category` type to ensure correct handling while continuous measures stayed numeric. A summary table (Table 4) was then generated to confirm that each variable was classified as intended.

Table 4. Summarizing Data Type Correction

Variable	Raw	Corrected	Number of Levels (As Applicable)
AGE_12	int64	category	10
EDUC	int64	category	7
MARSTAT	int64	category	6
PERMTEMP	float64	category	4
UNION	float64	category	3
COWMAIN	float64	category	2
GENDER	int64	category	2
MJH	float64	category	2
TENURE	float64	float64	N/A
HRLYEARN_T	float64	float64	N/A
AHRSMAIN_T	float64	float64	N/A
PAIDOT_T	float64	float64	N/A

UNPAIDOT_T	float64	float64	N/A
XTRAHRS_T	float64	float64	N/A

To check invalid entries, we verified that numeric values fell within feasible labour market limits and that values in categorical variables only contained codes specified in the codebook. For average weekly working hours, we used **168 hours** as the absolute weekly maximum (24 hours x 7 days), seeking to flag any value above this threshold or **less than or equal to zero** (i.e. zero or negative hours worked).

For overtime-related variables, we used a maximum threshold of **133 hours**, which represents the most overtime a worker could log in a week after assuming a standard 35-hour workweek (168 possible hours - 35 regular hours). This served as the cut-off for paid overtime hours (PAIDOT_T), unpaid overtime (UNPAIDOT_T), and extra hours worked (XTRAHRS_T). Values above this limit were considered implausible and likely due to reporting or entry errors.

Ultimately, with these checks, we found **no invalid entries**.

2.7 Outliers

As a final step to data preparation, we used the **3-IQR rule** to check for any outliers that might represent entry errors and pose problems for our analysis. Several variables showed outliers, including hourly wages and the different measures of hours worked or overtime (Table 5).

Table 5. Outlier Count & Proportion

Code	Description	Outlier Count	Outlier %
XTRAHRS_T	Extra hours worked	8,037	19.67%
PAIDOT_T	Paid overtime hours	4,830	11.82%
AHRSMAIN_T	Actual hours worked per week	3,612	8.84%
UNPAIDOT_T	Unpaid overtime hours	3,423	8.38%
HRLYEARN_T	Usual hourly earnings	264	0.65%

TENURE	Job tenure with current employer	0	0.00%

Variables with a lower proportion of outliers (AHRSMAIN_T, UNPAIDOT_T, HRLYEARN_T) showed ranges consistent with normal labour market behaviour and were acceptable to retain.

Even variables with higher outlier percentages reflected typical patterns in real workplaces rather than data-entry errors. In most industries, employers aim to limit overtime and excess hours for cost, safety, and productivity reasons. However, extended hours and overtime still occur under recurring conditions such as staffing shortages, seasonal peaks, deadline-driven workloads, emergency coverage, or roles where shift extension is common. These situations can create right-tail distributions in measures like average weekly hours (AHRSMAIN_T), paid and unpaid overtime (PAIDOT_T, UNPAIDOT_T), and extra hours worked (XTRAHRS_T).

Because these patterns seem to mirror true employment circumstances rather than mistakes in reporting, **the observed outliers were preserved**. This ensures our analysis covers the full range of working conditions Canadian employees may experience.

2.8 Impact of Data Preparation

Our before-and-after comparison plots illustrate how data preparation changed the data.

Missing values were reduced from 3,746,983 to zero, the number of variables decreased from 59 to 14, and the row count narrowed from 112,927 to 40,868 rows after applying our inclusion criteria (Figure 1).

In the raw data, several variables displayed distorted ranges due to implied decimals and the presence of implausible values, including large values of zero hours worked (Figure 2). After correcting variable scales and removing records with zero reported weekly hours, the distributions of hourly wages, weekly hours, paid and unpaid overtime, and extra hours aligned with realistic labour market trends (Figures 2–6). The cleaned variables now exhibit meaningful ranges, providing a reliable foundation for modelling and interpretation.

Figure 1. Overview of Data Preparation Before & After

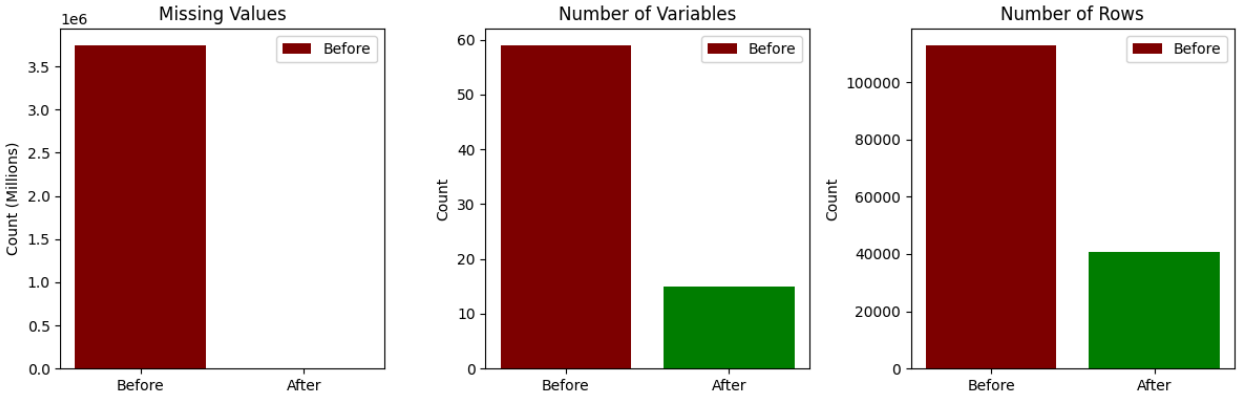


Figure 2. Before & After Distribution of Average Weekly Hours

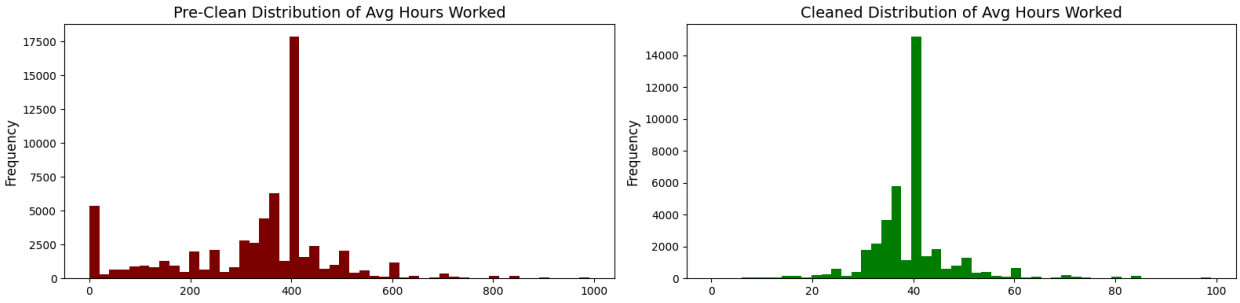


Figure 3. Before & After Distribution of Average Hourly Wages

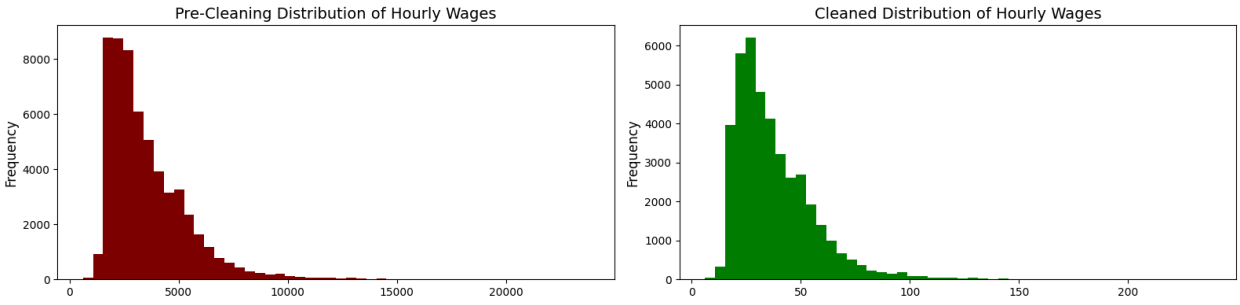


Figure 4. Before & After Distribution of Paid Overtime

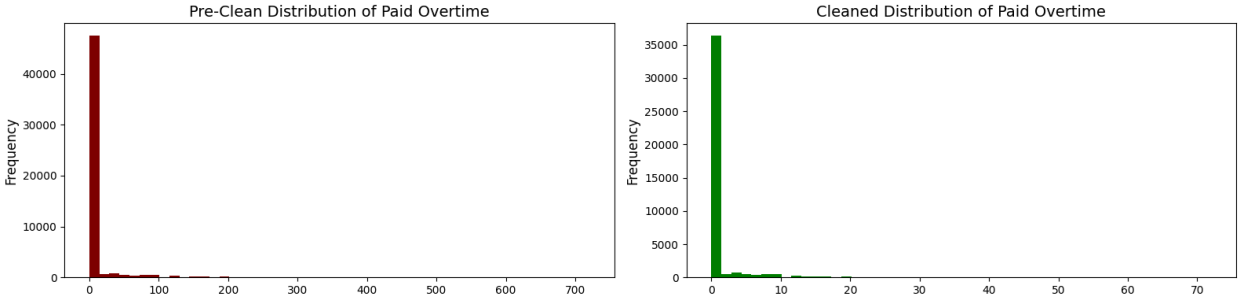


Figure 5. Before & After Distribution of Unpaid Overtime

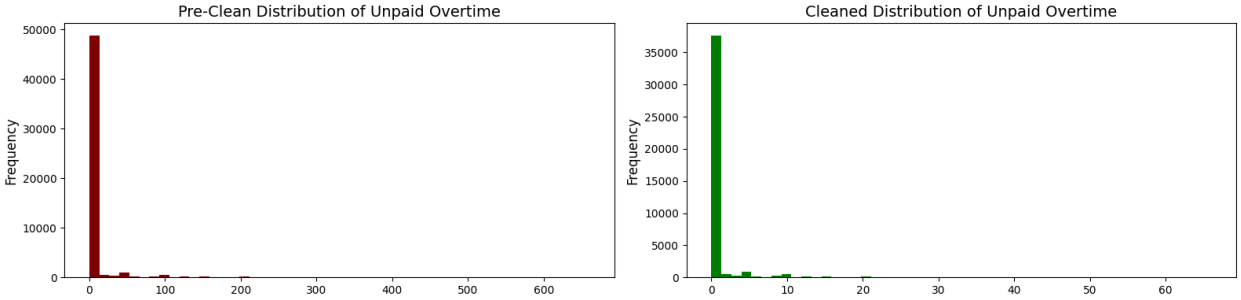
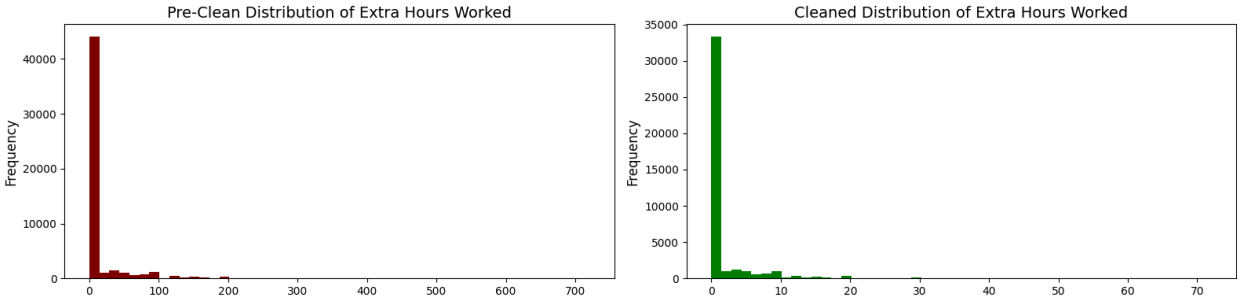


Figure 6. Before & After Distribution of Extra Hours Worked



3.0 Descriptive Analysis

Having established a clean and reliable dataset, we turn to descriptive analysis to understand the distribution of wages, hours, and worker attributes. This first analytical stage frames the deeper diagnostic and predictive work ahead.

3.1 Describing the Numeric Variables

Descriptive statistics for the numeric variables show generally stable patterns among full-time workers (Table 6).

Tenure averages about 94 months (approx. 8 years), with a lower median indicating that long job stays are less common. Hourly earnings average roughly $38.00, and weekly hours cluster tightly around 40, which makes sense given the typical full-time range of 35–40 hrs/week and the existence of overtime. Median paid, unpaid, and extra hours are all zero, reflecting that most workers do not work overtime, while higher means indicate a smaller group with substantial additional hours.

Table 6. Measures of Central Tendency

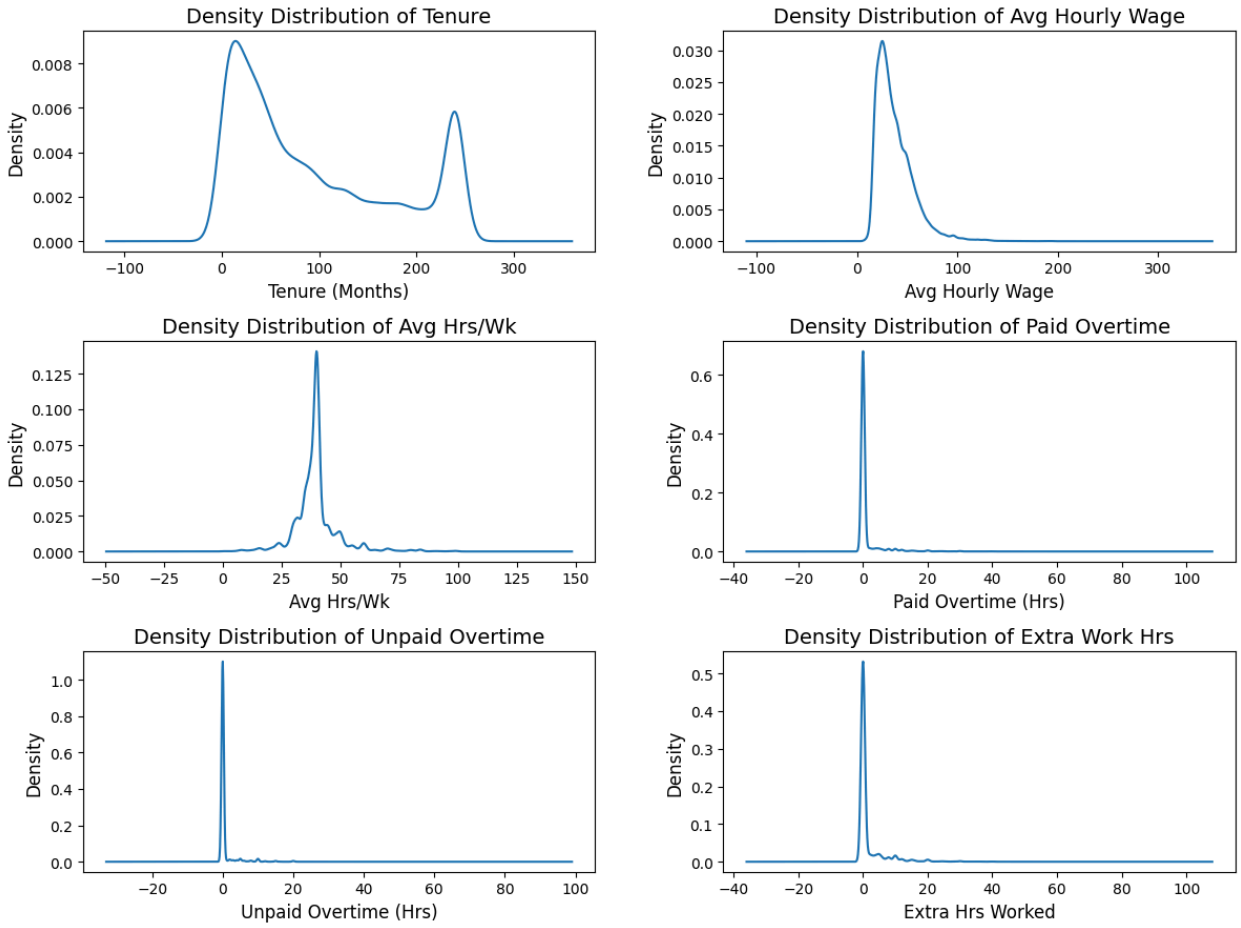
Variable	Mean	Median	Std Dev	Min	Max	Count
TENURE	93.9	64	82.03	1	240	40,868
HRLYEARN_T	37.74	33	18.6	6.07	237.91	40,868
AHRSMAIN_T	39.6	40	9.29	0	99	40,868
PAIDOT_T	1.11	0	4.33	0	72	40,868
UNPAIDOT_T	0.62	0	2.77	0	66	40,868
XTRAHRS_T	1.73	0	5.07	0	72	40,868

Distribution analyses reinforce these patterns (Table 7 & Figure 7). Tenure is bimodal, while wages and weekly hours who moderate right skew, likely driven by high earners and long-hour workers. Overtime variables exhibit extreme right skew and heavy tails, consistent with a labour market where only a minority regularly works extensive overtime.

Table 7. Skewness & Kurtosis Checks

Variable	Skewness	Kurtosis
TENURE	0.672	-0.987
HRLYEARN_T	1.867	6.333
AHRSMAIN_T	1.401	7.956
PAIDOT_T	5.999	46.385
UNPAIDOT_T	7.226	79.783
XTRAHRS_T	4.706	30.195

Figure 7. Density Plots of Numeric Variables



3.2 Describing the Categorical Summaries

For the categorical and binary variables (i.e. `EDUC`, `MARSTAT`, `AGE_12`, `PERMTEMP`, `UNION`, `MJH`, `GENDER`, `COWMAIN`), we used frequency tables and pie charts (Figure 8) to examine summary statistics. Due to the high number of categorical variables, this report only includes one of many frequency tables (Table 8); the others can be drawn from our codebase.

Table 8. Education (`EDUC`) Frequency Table — Example

Level	**Count**
Post-Secondary Certificate or Diploma	15,584
Bachelor's Degree	9,535
High School Graduate	6,992
Above Bachelor's Degree (Master's/Doctorate)	5,079
Some High School	1,791
Some Post-Secondary Education	1,445
0–8 Years of Non-Secondary Schooling	442

The frequency distributions show a full-time employed workforce that's generally well-educated, with most respondents holding post-secondary certificates or bachelor's degrees. Marital status is dominated by married individuals, followed by those who are single or in common-law relationships. The age profile is concentrated between 30 and 54, reflecting a predominantly mid-career workforce. The gender split is relatively balanced, with a slight majority being men.

Most full-time employees hold permanent positions, and non-unionized is more prevalent than unionized employment. The vast majority work a single job rather than multiple jobs. Finally, most respondents work in the private sector, though the public sector also represents a substantial share of full-time employment.

3.3 Segmentation

As the final step to our descriptive analysis, we segmented average hourly wage by gender, age group, marital status, and education (our main predictors, shown in Figure 9 & 10). From this we gathered the following insights.

Gender. Average hourly wages are slightly higher for men than women, reflecting a persistent but model gender wage gap. The difference is not extreme in this sample but aligns with broader Canadian (and global) labour market trends where men, on average, earn more per hour than women.

Age Group. Hourly wages increase steadily with age through the mid-career years. Earnings rise from early adulthood (15–24) into peak earning ages (40–54), then level off or slightly decline approaching the 60–64 range. This pattern reflects typical career progression, where experience, tenure, and seniority contribute to higher wages during prime working years.

Marital Status. Married and common-law respondents earn the highest hourly wages, while widowed, separated, and divorced individuals earn less on average. These patterns could possibly relate to age, job stability, and accumulated work experience, since married individuals tend to be older and more established in their careers. Such patterns may also reflect the impact of familial support, where this additional support results in improved career outcomes.

Figure 8. Visualizing Proportion of Categorical Variable Groups (Compiled Pie Charts)

Figure 9. Segmenting Avg Wage by Gender (GENDER) & Age Group (AGE_12)

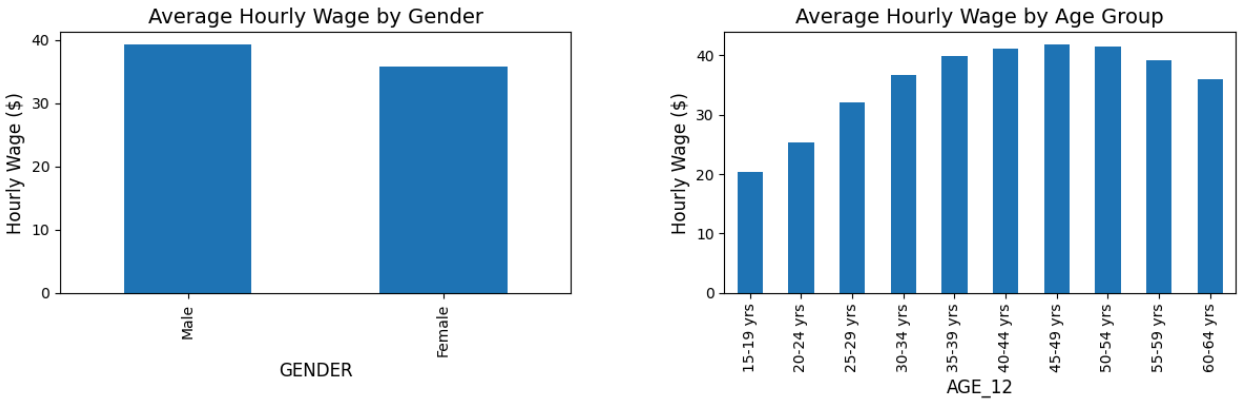
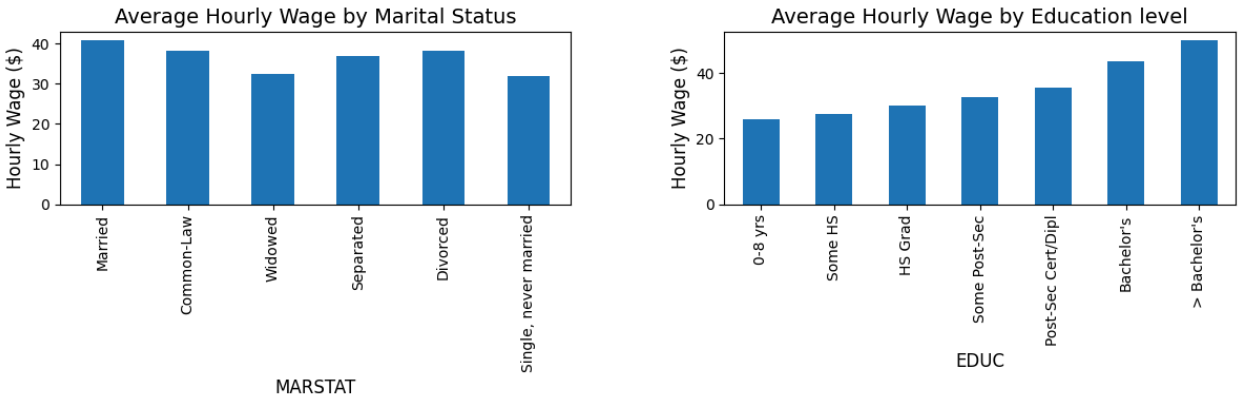


Figure 10. Segmenting Avg Wage by Marital Status (MARSTAT) & Education Level (EDUC)



Education Level. Education shows a clear upward trajectory: higher levels of schooling correspond to higher hourly wages. Workers with 0–8 years of schooling or some high school earn the least, and wages increase progressively through post-secondary credentials. Bachelor's and above-bachelor's degrees show the highest returns, consistent with strong labour market premiums for advanced education.

Although we also segmented based on average hours worked, we found this less helpful due to the normal range of 35–40 hrs/week in full time workers. As a result, we have omitted segmentation by average weekly hours from this report.

4.0 Diagnostic Analysis

Following the descriptive overview, we conducted diagnostic testing to understand how characteristics interact and to verify data readiness for modelling. Using cross-tabulations, regression diagnostics, statistical tests, and pair plots, we examined relationships between predictors, checked for structural issues such as multicollinearity, and identified which variables carried the strongest explanatory signals.

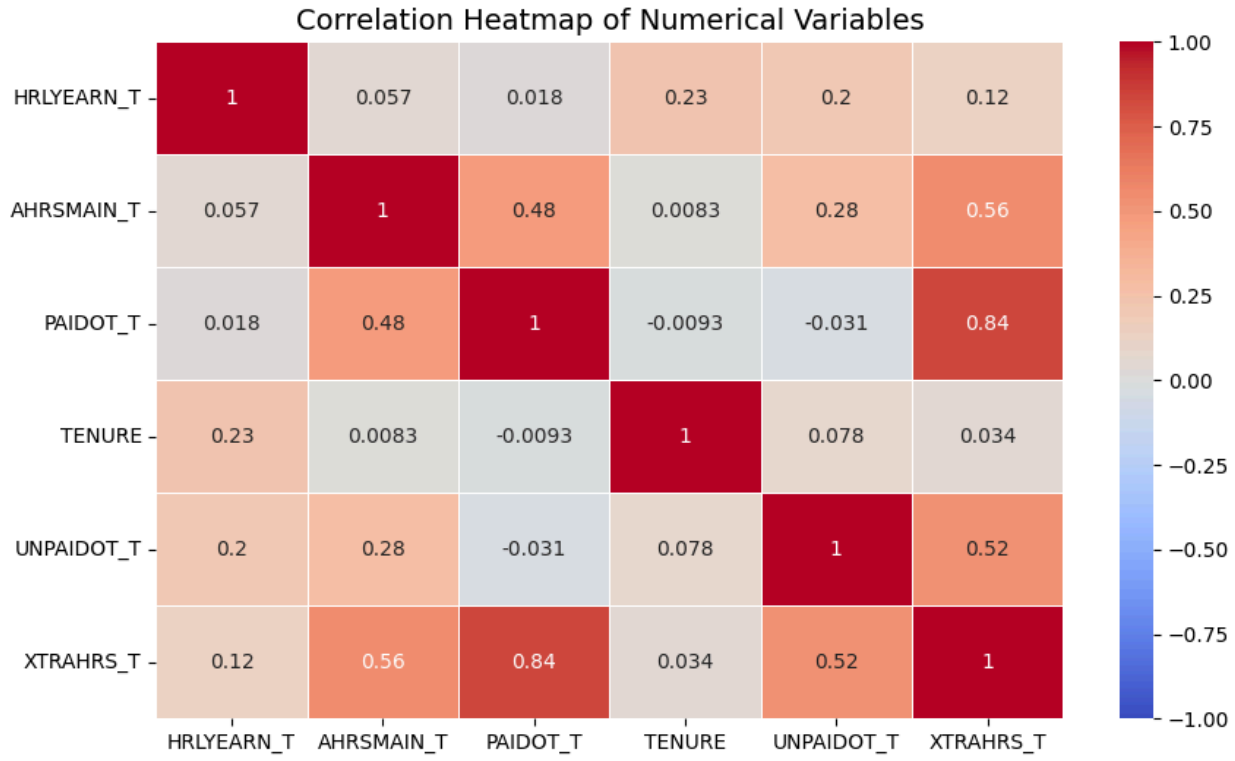
4.1 Correlation Analysis

We produced a text-labelled correlation heatmap (Figure 11), which showed most numeric variables having weak or modest relationships with hourly wages, indicating no single continuous predictor overwhelmingly drives earnings.

That said, the diagnostic shows a strong cluster among the overtime-related variables. Paid overtime (`PAIDOT_T`), unpaid overtime (`UNPAIDOT_T`), and extra hours worked (`XTRAHRS_T`) are highly correlated with one another, but real-life context tells us that these strong internal correlations are likely the result of multicollinearity rather than meaningful relationships. Of course, those who work overtime will also report extra hours as overtime constitutes extra hours.

To avoid redundancy and preserve model stability, we removed extra hours (`XTRAHRS_T`) from subsequent statistical testing and predictive modelling.

Figure 11. Correlation Heatmap of Numerical Variables

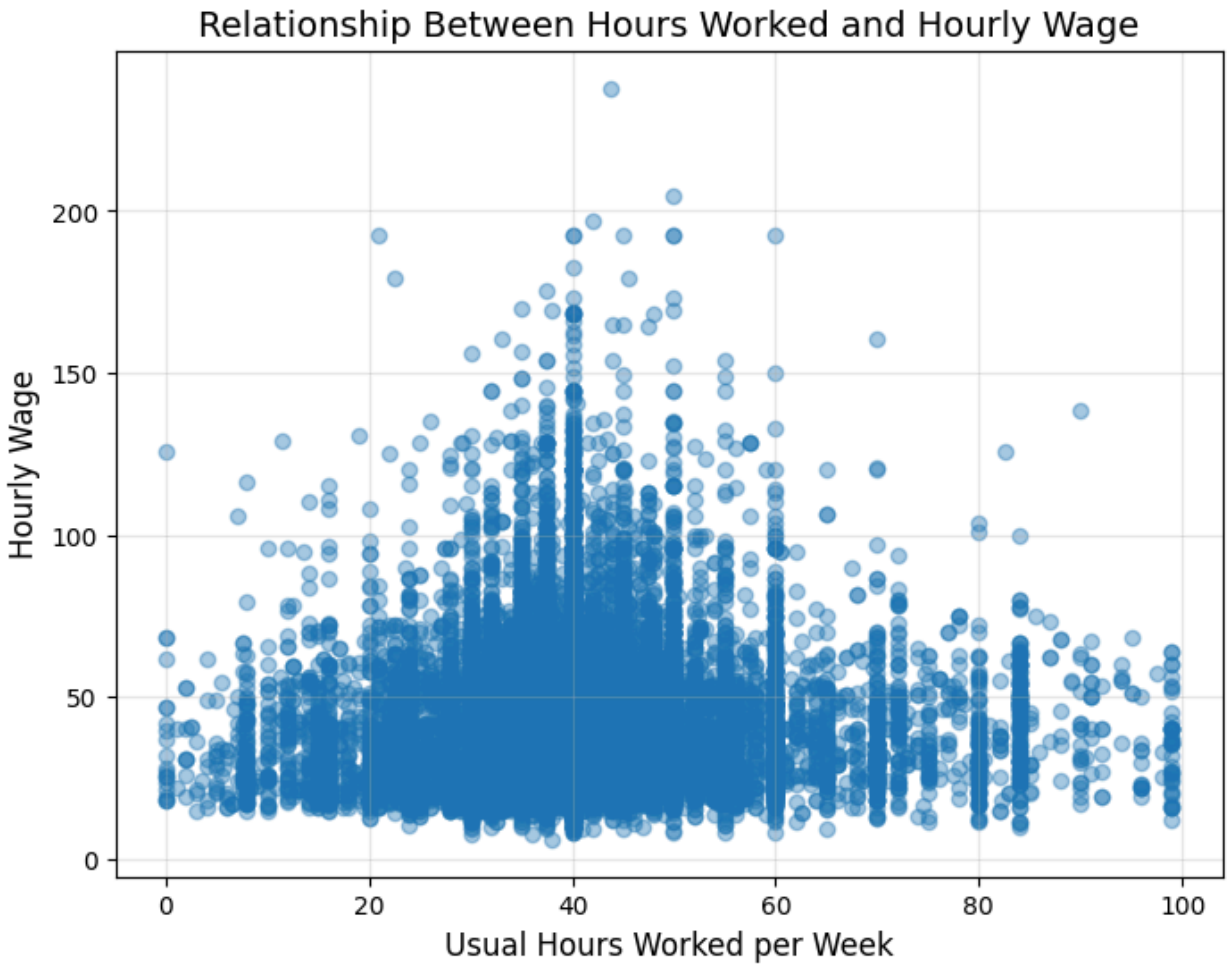


4.2 Scatter & Pair Plots

To further understand how variables interact, we used both bivariate scatterplots and pair plots. These visual tools allow us to see how relationships behave across the distribution rather than relying only on linear associations. We focused on hourly wages, usual weekly hours, and tenure because these are central labour market indicators that frequently explain variation in earnings and interact with many of the other predictors.

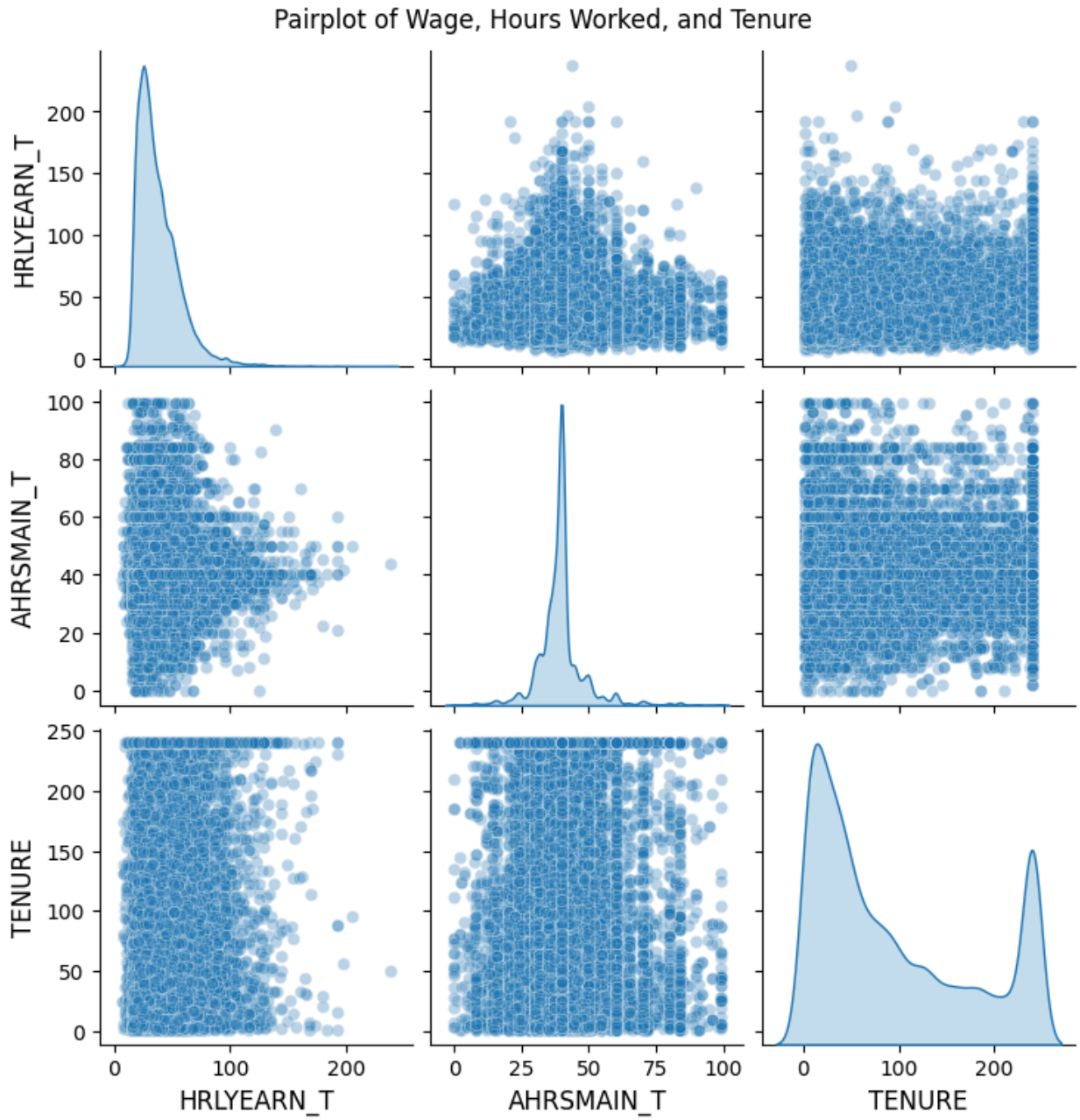
The scatterplot of hours worked and hourly wage (Figure 12) shows that the highest wages tend to occur among people working a typical full-time schedule of about 35-45 hrs/week. Very short or very long workweeks are linked to lower wages, suggesting that hours alone are not a strong driver of pay and that wage patterns are more complex than a simple upward or downward trend.

Figure 12. Bivariate Scatterplot of Usual Weekly Hours & Hourly Wage



The pair plot visualization (Figure 13) adds another layer to our analysis but shows how wages, hours, and tenure move together. We see that, in addition to our standalone bivariate scatterplot, workers with longer tenure often earn more, although there is considerable variation.

Figure 13. Pair Plot Visualization of Relationships between Wage, Hours Worked, Tenure



4.3 Cross-Tabulation

To explore wage differences across groups, we used cross-tabulations to examine how earnings are distributed within each category of the main variables. This method adds a categorical perspective to the patterns already suggested in earlier descriptive statistics, distributions, and diagnostic checks.

We temporarily grouped hourly earnings into five wage bands ranging from $0–20/hr to $80+/hr to enable clearer comparisons, and then generated percentage-based cross-tabs against all categorical and binary variables. Two results stood out as especially informative.

Union Status x Hourly Wage (Table 9). Unionized workers appear less often in the lowest wage category and more often in mid-range wage bands compared with non-union workers. This distribution suggests that union membership may be associated with stronger wage floors or more structured compensation.

Table 9. Union Status × Hourly Wage Category (%)

Union Status	**$0–20/hr**	**$20–40/hr**	**$40–60/hr**	**$60–80/hr**	**$80+/hr**
Union Member	4.7	51.8	35.5	6.8	1.2
Unionized Non-Member	9.2	50.9	28	8.8	3.1
Non-Unionized	17.2	52.1	19.4	7.3	4

Public vs. Private Sector x Hourly Wage (Table 10). public sector workers show lower representation in the lowest wage band and higher representation in the $40–60/hr and $60–80/hr wage bands. Meanwhile, private-sector workers appear more concentrated in the lower wage ranges. This pattern runs counter to the common assumption that private-sector jobs always offer higher pay, suggesting instead that the greater consistency and standardized compensation structures in the public sector may place more workers safely above the lowest wage tiers.

Table 10. Worker Class × Hourly Wage Category (%)

Worker Class	**$0–20/hr**	**$20–40/hr**	**$40–60/hr**	**$60–80/hr**	**$80+/hr**
Public sector	2.4	46.5	38.1	10.4	2.6
Private sector	17.1	54.1	19.6	5.9	3.3

4.4 Initial Regression Analysis & Statistical Testing

To close off our diagnostic analysis, we performed an initial linear regression analysis and statistical testing, treating `HRLYEARN_T` as our outcome variable and `COWMAIN`, `AGE_12`, `EDUC`, `GENDER`, `MARSTAT`, `MJH`, `PERMTEMP`, `UNION`, `TENURE`, `AHRSMAIN_T`, `PAIDOT_T`, and `UNPAIDOT_T` as our additive predictor variables. We found that the model explains about **25.3%** of wage variation (**$R^2 = 0.263$**), which is reasonable for cross-sectional labour market data where unobserved firm- and occupation-level factors typically play significant roles. Our key findings can be summarized as follows.

1. Sector effects are large and highly significant.

Workers in the **private sector** earn about **$4.41 less per hour** compared with public sector workers, holding all else constant. This sizable, persistent gap highlights structural compensation differences independent of personal characteristics.

2. Age patterns show a clear mid-career earnings peak.

Coefficients rise steadily through the 30s and 40s, with the **45–49 age group** earning about **$7.53 more per hour** relative to the 15–19 reference group. Earnings plateau after the mid-50s and decline slightly near retirement ages, reflecting typical productivity and seniority trajectories.

3. Education is one of the strongest predictors of higher wages.

Post-secondary pathways yield substantial wage increases. A bachelor's degree is associated with a **$7.79/hour** increase in wage compared to someone with a post-secondary diploma or certificate, and a **$21.88/hour** increase compared to someone with no high school experience.

4. Gender inequality appears clearly in the model.

The coefficient for **women** is **-$6.37/hour**, a statistically significant decrease even after controlling for job tenure, overtime, marital status, and education. This indicates persistent gender wage disparities not explained by the other predictive factors.

5. Marital status has nuanced associations.

Most categories differ significantly from the married reference group. **Widowed, separated, divorced,** and **single** respondents show hourly wages at **$0.18–3.41 lower per hour**, while common-law workers are paid slightly higher (**+$0.46**), suggesting interpersonal circumstances and household structure may interact with career opportunities or constraints.

6. Temporary or non-standard employment types reduce earnings.

Temporary contract, seasonal, and casual roles all predict lower ranges ranging from **$2.16 to $5.70 less per hour** than permanent employees.

7. Unionization is positively associated with wages.

Union members earn about **$1.89 more per hour** and **non-member unionized workers** about **$1.46** more than employees at non-unionized workplaces, hinting at a positive influence of collective bargaining environments on wage rates.

8. Overtime explains meaningful wage variation.

Meanwhile, overtime measures also show positive associations with hourly wages. Each additional hour of **paid overtime** is linked to an estimated increase of **$0.15**, while each hour of **unpaid overtime** corresponds to an estimated increase of **$0.82**. The relatively larger unpaid overtime coefficient pay reflects that higher earning roles often carry expectations of performance -driven extra work that is not formally compensated.

9. Tenure shows a small but consistent positive effect.

Tenure shows a small but consistent positive relationship with wages. The coefficient of approximately **$0.037/month** translates to about **$0.44/year**, suggesting a gradual increase in earnings as workers accumulate firm-specific experience.

5.0 Predictive Analysis

To understand how personal and situational characteristics relate to higher earnings, we built predictive models based on a derived variable from `HRLYEARN_T`, called `HIGH_WAGE`, which classified workers as earning above or below the **median hourly wage** of **$33.00/hour**. Converting wages into a binary outcome allows us to compare individuals to the broader population and provides a practical way to evaluate model performance through confusion matrices, ROC curves, and accuracy measures.

Before modelling, we created a train-test split, holding out 25% of the data for testing. We confirmed that the training data remained balanced, with about 50.5% of workers above the median and 49.5% below it. This helped to ensure that predictions were not biased toward either class. We then ran and evaluated three different modelling approaches using the same outcome and predictor variables:

1. **Logistic regression**
2. **K-Nearest Neighbour** (KNN) with **k=7**, and
3. **Random Forest**.

Logistic regression served as a clear and interpretable baseline, regressing the binary outcome on linear additive relationships. KNN allowed us to test whether workers who share similar characteristics tend to fall on the same side of the wage distribution. Meanwhile, Random Forest provided a more flexible method capable of capturing complex interactions among variables.

Across all evaluation metrics, logistic regression (Model 1, Table 11) produced the strongest results, with an accuracy of 70.85% and an AUC of 0.783. Random Forest (Model 3, Table 13) followed closely with an accuracy of 70.47 % and an AUC of 0.77. KNN (Model 2, Table 12)

performed less effectively but still moderately well, with an accuracy of 66.29% and an AUC of 0.727.

A direct accuracy comparison is outlined in Table 14, and the ROC curves in Figure 14 visualize these differences, with all models showing a strong separation from the random guess line.

In plain language, these results mean that logistic regression correctly classified about seven out of ten workers and showed the strongest ability to tell higher and lower earners apart. Random Forest performed at almost the same level, while KNN captured some useful patterns but had more difficulty with the complex relationships in the data. All three models demonstrated real predictive value, suggesting that the characteristics included in our analysis contain enough information to estimate whether a full-time worker is likely to earn above the median wage in Canada.

Table 11. Model 1 — Logistic Confusion Matrix

Model 1	**Predicted At or Above Median**	**Predicted Below Median**
Actual High	TP = 3724 (71.64%)	FN = 1474 (28.36%)
Actual Not High	FP = 1505 (29.99%)	TN = 3514 (70.01%)
	Model 1 Accuracy	70.84
	Model 1 Precision	71.22
	Model 1 Negative Predictive Value	70.45

Table 12. Model 2 — KNN (k = 7) Confusion Matrix

Model 2	**Predicted At or Above Median**	**Predicted Below Median**
Actual High	TP = 3527 (67.85%)	FN: 1671 (32.15%)
Actual Not High	FP = 1773 (35.33%)	TN = 3246 (64.67%)
	Model 2 Accuracy	66.29
	Model 2 Precision	66.55
	Model 2 Negative Predictive Value	66.02

Table 13. Model 3 — Random Forest Confusion Matrix

Model 3	Predicted At or Above Median	Predicted Below Median
Actual High	TP = 3668 (70.57%)	FN = 1530 (29.43%)
Actual Not High	FP = 1488 (29.65%)	TN = 3531 (70.35%)
Model 3 Accuracy		70.46
Model 3 Precision		71.14
Model 3 Negative Predictive Value		69.77

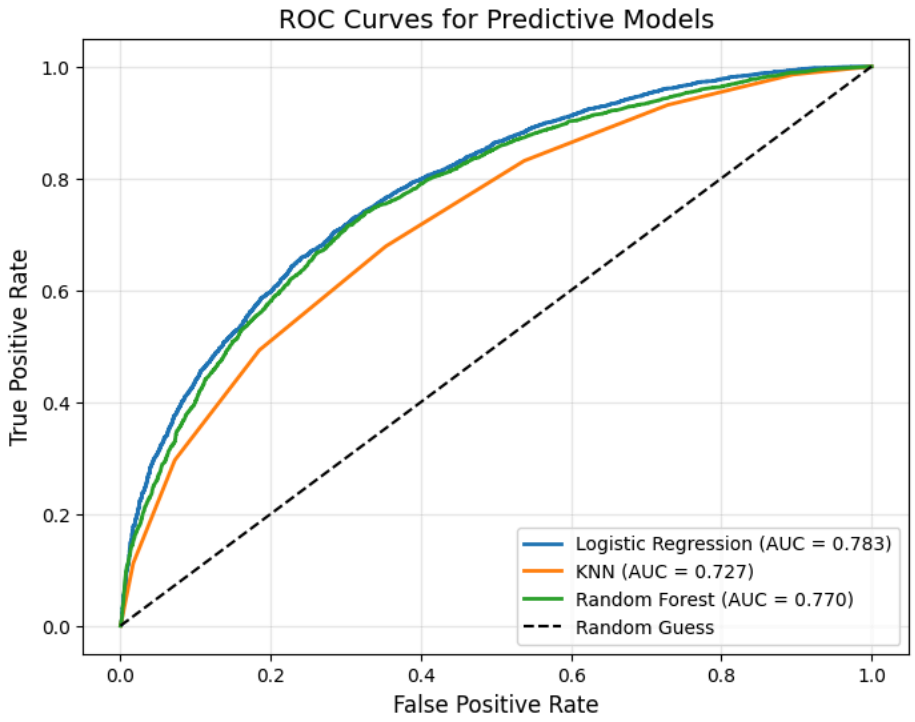
Table 14. Model Accuracy Comparison

Model	Type	Accuracy
1	Logistic Regression	0.708
2	KNN (k = 7)	0.663
3	Random Forest	0.705

Our resulting model (logistic regression, the most accurate one) ultimately draws from a worker's personal and situational characteristics to estimate the probability of earning above or below the median wage, providing a practical tool for applying our findings in real life settings. For example, an employment agency could enter a person's education level, age group, gender, marital status, job permanency, union environment, tenure, and over time patterns to get a prediction of their wage bracket. This could help counsellors identify people who may need additional training or support.

Organizations planning workforce programs could run the model on entire groups of employees to see which characteristics are linked to higher or lower earnings in their own workforce. Policy makers could apply the model to labour-force microdata to spot regions or demographic groups that are consistently predicted to fall below the median wage and then design interventions to address those gaps.

Figure 14. Line Plot of ROC Curves for Predictive Models



6.0 Conclusion

This project set out to understand which personal and situational characteristics are associated with higher hourly wages among full-time workers in Canada. Using the September 2025 Labour Force Survey, we applied descriptive, diagnostic, and predictive analysis to examine how age education, gender, marital status, job permanency, union status, tenure, and overtime patterns relate to earning above the national median wage. Several consistent themes emerged from the results.

Education proved to be one of the strongest predictors of higher earnings, with bachelor's degrees and higher credentials linked to substantially greater hourly wages. Age patterns followed a familiar trajectory, with earnings rising into the mid-career years before stabilizing. Gender remained a significant factor even after accounting for other characteristics, which suggests that substantial wage differences between men and women continue to persist. Immediate employment conditions also played an important role. Permanent jobs and unionized

workplaces were associated with higher earnings, while temporary roles were linked to lower wage levels. Tenure contributed a modest but steady effect over time.

Beyond these direct findings, the results revealed deeper patterns that gesture toward broader labour market structures. The size of public sector wage premiums, the concentration of peak earnings in mid-career, and the positive relationship between unpaid overtime and higher wages point to institutional and cultural factors that influence compensation. These observations raise further questions about job quality, workplace norms, and organizational wage-setting practices beyond what individual characteristics can explain.

With the predictive models, logistic regression performed most effectively, correctly classifying about seven in ten workers relative to the median wage threshold. A secondary statistical insight here is that more advanced forms of machine learning such as KNN and Random forest might not always outperform simpler methods like logistic regression. While the accuracy of the resulting model and the others show that they are not suited for high-stakes decision-making, they show that basic worker information can be used to identify broad wage patterns and support planning or screening in certain settings. Further work could apply these models to time-series data to see whether the predictors retain their strength over time and how the relationships among factors evolve as labour market conditions change—the stability and recurring nature of the Labour Force Survey makes this especially possible.

Our analysis was shaped by certain constraints that create clear opportunities for future development. As students who are still building skills in programming and statistics, we focused on methods we could apply with confidence and accuracy. This meant leaving aside more advanced approaches such as interaction modelling, regularization techniques, hierarchical models, and causal inference. We also chose to exclude occupation and industry in order to concentrate on personal characteristics and immediate circumstances. Additionally, the PUMF limits the depth of available detail for confidentiality purposes, which could have been addressed with bootstrapping but was beyond the scope of our project. These constraints point toward natural next steps, including incorporating richer modelling techniques, expanding the scope of variables, and exploring more complex data sources as our technical capacity grows.

7.0 Bibliography

Government of Canada, S. C. (2025, May 26). *Survey of Financial Security: Public Use Microdata File*. https://www150.statcan.gc.ca/n1/en/catalogue/13M0006X

Nguyen, T. (2025, November 8). Rising cost of living pushing more Atlantic Canadians into debt cycle, experts say. *CBC News*. https://www.cbc.ca/news/canada/prince-edward-island/pei-affordability-credit-debt-9.6964313

Rabinovitch, A. (2025, October 10). *Unemployment rate held steady at 7.1% in September, StatCan says—National | Globalnews.ca* [News Platform]. Global News. https://globalnews.ca/news/11473168/canada-unemployment-jobs-report-september-2025/

Robert Half. (2025, September 29). *2026 Canada Salary Guide: Insights and Compensation Trends from Robert Half* [Organizational Website]. Robert Half Canada. https://www.roberthalf.com/ca/en/insights/research/2026-canada-salary-guide-insights-and-trends

Statistics Canada. (2021). *Labour Force Survey: Public Use Microdata File* (Public Use Microdata File No. 71M0001X; Version September 2025) [CSV]. Statistics Canada. https://www150.statcan.gc.ca/n1/pub/71m0001x/71m0001x2021001-eng.htm

Statistics Canada. (2025a, September 5). The Daily: Labour Force Survey, August 2025 [Government Department Website]. *Statistics Canada*. https://www150.statcan.gc.ca/n1/daily-quotidien/250905/dq250905a-eng.htm

Statistics Canada. (2025b, October 10). *Labour Force Survey: Public Use Microdata File*. https://www150.statcan.gc.ca/n1/en/catalogue/71M0001X

Statistics Canada. (2025c, November 17). *The Daily—Consumer Price Index, October 2025* [Government Department Website]. Statistics Canada. https://www150.statcan.gc.ca/n1/daily-quotidien/251117/dq251117a-eng.htm

The Canadian Press. (2025, January 23). *Canadians continue to struggle financially due to higher costs: RBC poll*. CTVNews. https://www.ctvnews.ca/business/article/canadians-continue-to-struggle-financially-due-to-higher-costs-rbc-poll/