Final Project - README FILE

Group 5
Annie (Mengqing) Wu, 1011429456
Belinda Kwan, 999206540
Nahel Sinan, 1007978093

Instructor: Dr. Maher Elshakankiri
Course code: INF1340
Course name: Programming for Data Science
Program: Master of Information
Faculty of Information
University of Toronto

Date Created: 2025-11-07
Date Modified: 2025-12-03

Understanding Wage Variation in Canada:

Personal and Situational Predictors of Higher Earnings

This project analyzes a large Public Use Microdata File (PUMF) from Statistics Canada's September 2025 Labour Force Survey to explore wage patterns and the characteristics associated with higher earnings. Before cleaning and transformation, the dataset contained 112,927 respondents and 59 variables capturing a wide range of information, including demographics, education, labour force status, employment conditions, hours worked, income measures, and overtime activity.

The primary aim of the project is to examine how personal and situational factors relate to hourly wages, visualize key labour market trends, and develop predictive models that classify workers above or below the national median wage. Through a combination of data processing, exploratory analysis, statistical diagnostics, and machine learning, the project offers insights into workforce dynamics and the attributes most strongly linked to higher pay among full-time workers in Canada.

The dataset is sourced from Statistics Canada and can be found here: https://www150.statcan.gc.ca/n1/pub/71m0001x/2021001/2025-09-CSV.zip

Installation

This project requires the following Python modules:

Core Libraries:

- pandas
- [numpy]
- [matplotlib.pyplot]
- [seaborn]

Statistical Libraries:

- [scipy]
- [statsmodels]

Machine Learning Libraries (scikit-learn)

- [scikit-learn] (includes KNeighborsClassifier, RandomForestClassifier, OneHotEncoder, ColumnTransformer, Pipeline, train_test_split, LogisticRegression, confusion_matrix, roc_curve, auc)

Install the packages on your computer or Python environment. If you are using colab, it is automatically installed.

Usage

1. To run the program, download the dataset `pub0925.csv` from https://www150.statcan.gc.ca/n1/pub/71m0001x/2021001/2025-09-CSV.zip. Move it to the same directory as the .py or Colab file, and **rename it** as `Final_Inputdata_Kwan_Sinan_Wu.csv`. Alternatively, the .csv file that comes with the assignment upload can be used as-is.

2. Run the program through Colab and follow the code chunk step by step.

Sample Output

SEPTEMBER 2025 LABOUR FORCE SURVEY
PUBLIC USE MICRODATA FILE (PUMF)
STATISTICS CANADA

OVERVIEW

# of Variables: 59
# of Observations: 112927

Dataset Variables & Their Types

	Code	Type	# Levels
0	AGE_6	float64	N/A
1	MJH	float64	N/A
2	EVERWORK	float64	N/A
3	FTPTLAST	float64	N/A
4	COWMAIN	float64	N/A
5	NAICS_21	float64	N/A
6	NOC_10	float64	N/A
7	NOC_43	float64	N/A
8	YABSENT	float64	N/A
9	WKSAWAY	float64	N/A
10	PAYAWAY	float64	N/A
11	UHRSMAIN	float64	N/A
12	AHRSMAIN	float64	N/A
13	FTPTMAIN	float64	N/A
14	UTOTHRS	float64	N/A
15	ATOTHRS	float64	N/A
16	HRSAWAY	float64	N/A
17	YAWAY	float64	N/A
18	PAIDOT	float64	N/A
19	UNPAIDOT	float64	N/A
20	XTRAHRS	float64	N/A
21	WHYPT	float64	N/A
22	TENURE	float64	N/A
23	PREVTEN	float64	N/A
24	HRLYEARN	float64	N/A
25	UNION	float64	N/A
26	PERMTEMP	float64	N/A
27	ESTSIZE	float64	N/A

28 FIRMSIZE float64 N/A
29 DURUNEMP float64 N/A
30 FLOWUNEM float64 N/A
31 UNEMFTPT float64 N/A
32 WHYLEFTO float64 N/A
33 WHYLEFTN float64 N/A
34 DURJLESS float64 N/A
35 AVAILABL float64 N/A
36 LKPUBAG float64 N/A
37 LKEMPLOY float64 N/A
38 LKRELS float64 N/A
39 LKATADS float64 N/A
40 LKANSADS float64 N/A
41 LKOTHERN float64 N/A
42 PRIORACT float64 N/A
43 YNOLOOK float64 N/A
44 TLOLOOK float64 N/A
45 SCHOOLN float64 N/A
46 AGYOWNK float64 N/A
47 SURVYEAR int64 N/A
48 SURVMNTH int64 N/A
49 LFSSTAT int64 N/A
50 PROV int64 N/A
51 CMA int64 N/A
52 AGE_12 int64 N/A
53 GENDER int64 N/A
54 MARSTAT int64 N/A
55 EDUC int64 N/A
56 IMMIG int64 N/A
57 EFAMTYPE int64 N/A
58 FINALWT int64 N/A

INITIAL VARIABLE SELECTION

For initial variable selection, we included measures that capture the main factors shaping hourly wages while ensuring a consistent group of workers.

Context variables define the analytic sample and help exclude respondents whose employment conditions do not inform our research question.

Meanwhile, hourly earnings represent as the outcome variable, with core predictors such as age group, hours worked, education, gender, and marital status. These are widely recognized influences on wage variation.

Additional controls related to job characteristics, overtime, tenure, and union

status help account for other sources of variation and reduce potential confounding in the model.

Preview of Post-Cut Dataset

REC_NUM	LFSSTAT	FTPTMAIN	COWMAIN	HRLYEARN	AGE_12	AHRSMAIN	EDUC	GENDER
1	1	1.0	6.0	NaN	9	360.0	4	2
2	1	1.0	2.0	1700.0	2	200.0	2	2
3	2	1.0	1.0	4087.0	4	0.0	5	2
4	4	NaN	NaN	NaN	11	NaN	5	2
5	4	NaN	NaN	NaN	10	NaN	5	1

REC_NUM	MARSTAT	MJH	PAIDOT	PERMTEMP	SCHOOLN	TENURE	UNION	UNPAIDOT
1	5	1.0	NaN	NaN	1.0	240.0	NaN	NaN
2	6	1.0	0.0	1.0	1.0	29.0	3.0	0.0
3	6	1.0	NaN	1.0	1.0	121.0	1.0	NaN
4	6	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	1	NaN	NaN	NaN	1.0	NaN	NaN	NaN

REC_NUM	XTRAHRS
1	NaN
2	0.0
3	NaN
4	NaN
5	NaN

FILTERING ROWS TO FIT RESEARCH SCOPE

We then filtered the dataset to include only people who were actively employed at work, working full-time in their main job, not attending school, and employed as wage or salary workers rather than self-employed.

These criteria create a consistent analysis sample representing typical full-time employees. After applying the filters, the variables used to screen the data (employment status, full-time status, and school attendance) contain only one remaining value for all cases, so they are removed to avoid redundancy and keep the dataset clean and focused.

After filtering for research scope, there are 40868 observations and 14 variables.

MISSING VALUES & DUPLICATE ROWS

Next, we checked for missing values and duplicate rows. Post-variable selection & row filtering, we found no missing values and 1263 duplicate rows.

Since this is population survey data and the record identifier was used as the index and not included in the duplicate check, the number of duplicate rows is acceptable.

We also checked for duplicates within the index row and found 0.

DATA TRANSFORMATION

The PUMF user guide said that several labour-market variables use implied decimals, so we adjusted
the raw values accordingly. We converted hourly wages from cents to dollars and usual hours from
tenths to full hours, and confirmed that the transformed fields had the correct data types.

Comparison of Pre- and Post-Transformation Variables
---

	mean	median	std	min	max	count
HRLYEARN	3773.57	3300.0	1860.20	607.0	23791.0	40868.0
AHRSMAIN	396.02	400.0	92.88	0.0	990.0	40868.0
PAIDOT	11.08	0.0	43.30	0.0	720.0	40868.0
UNPAIDOT	6.22	0.0	27.67	0.0	660.0	40868.0
XTRAHRS	17.30	0.0	50.66	0.0	720.0	40868.0

	mean	median	std	min	max	count
HRLYEARN_T	37.74	33.0	18.60	6.07	237.91	40868.0
AHRSMAIN_T	39.60	40.0	9.29	0.00	99.00	40868.0
PAIDOT_T	1.11	0.0	4.33	0.00	72.00	40868.0
UNPAIDOT_T	0.62	0.0	2.77	0.00	66.00	40868.0
XTRAHRS_T	1.73	0.0	5.07	0.00	72.00	40868.0

Check Dataset Variables

COWMAIN float64
HRLYEARN float64
AGE_12 int64
AHRSMAIN float64
EDUC int64
GENDER int64
MARSTAT int64
MJH float64
PAIDOT float64

PERMTEMP float64
TENURE float64
UNION float64
UNPAIDOT float64
XTRAHRS float64
HRLYEARN_T float64
AHRSMAIN_T float64
PAIDOT_T float64
UNPAIDOT_T float64
XTRAHRS_T float64
dtype: object

Lastly, we dropped the original columns to keep the dataset streamlined.

CORRECTING INCONSISTENCIES

We checked for several types of inconsistencies to ensure the dataset was usable and analytically sound.

First, we verified that each variable had the correct data type, correcting fields that should be categorical or binary, or numeric so they would be interpreted properly in analysis. We also reviewed variable labels and coding schemes to confirm that categories matched the documentation
and that no unexpected or unclassified values were present.

Finally, we screened for impossible or illogical values, such as negative hours, wages equal to zero, or working hours exceeding realistic limits.

-- DATA TYPE CORRECTION

To correct the variable types in the dataset, We organized variables into categorical, numeric, and binary groups based on the codebook. Categorical and binary variables were converted to the "category" type to improve efficiency and support correct handling in analysis, while numeric variables were kept in their continuous form.

A table was generated to confirm that variable classifications were applied as intended.

Variable Summary Table

	Code	Type	# Levels
0	AGE_12	category	10
1	EDUC	category	7
2	MARSTAT	category	6
3	PERMTEMP	category	4

4 UNION category 3
5 COWMAIN category 2
6 GENDER category 2
7 MJH category 2
8 TENURE float64 N/A
9 HRLYEARN_T float64 N/A
10 AHRSMAIN_T float64 N/A
11 PAIDOT_T float64 N/A
12 UNPAIDOT_T float64 N/A
13 XTRAHRS_T float64 N/A

-- INVALID ENTRIES CHECK

For numeric variables, we checked whether values fell within possible labour-market limits. We flagged working hours above 168 (24 × 7) and those less than or equal to zero working hours. We
also treated PAIDOT_T, UNPAIDOT_T, and XTRAHRS_T values above 133 hours as implausible, given that
133 represents the maximum possible overtime in a week (24 × 7 − 35). Values beyond these thresholds likely reflect reporting errors rather than real labour behaviour.

Invalid working hrs > 168 hours: 0
Invalid working hrs <= to 0 hours: 19
Invalid extra hours worked > 133 hours: 0
Invalid paid overtime hrs > 133 hours: 0
Invalid unpaid overtime hrs > 133 hours: 0

For categorical variables, we looked at whether any values fell outside of the specified categories by checking for unique values in each category.

Invalid EDUC entries: 0
Invalid MARSTAT entries: 0
Invalid AGE_12 entries: 0
Invalid PERMTEMP entries: 0
Invalid UNION entries: 0
Invalid MJH entries: 0
Invalid GENDER entries: 0
Invalid COWMAIN entries: 0

-- RECODING LABELS

Next, we recoded labels for all categorical variables, to match the codebook (some labels have been
shortened for ease of visualization & reference).

EDUC: ['0-8 yrs', "> Bachelor's", "Bachelor's", 'HS Grad', 'Post-Sec Cert/Dipl', 'Some HS', 'Some Post-Sec']
MARSTAT: ['Common-Law', 'Divorced', 'Married', 'Separated', 'Single, never married', 'Widowed']
AGE_12: ['15-19 yrs', '20-24 yrs', '25-29 yrs', '30-34 yrs', '35-39 yrs', '40-44 yrs', '45-49 yrs', '50-54 yrs', '55-59 yrs', '60-64 yrs']
PERMTEMP: ['Permanent', 'Temp Other/Casual', 'Temp Seasonal', 'Temp Term/Contract']
UNION: ['Non-Unionized', 'Union Member', 'Unionized Non-Member']
MJH: ['Multiple jobs', 'Single job']
GENDER: ['Female', 'Male']
COWMAIN: ['Private sector', 'Public sector']

We also ensured ordinal categories were classified as ordinal.

AGE_12 (ordered):
['15-19 yrs', '20-24 yrs', '25-29 yrs', '30-34 yrs', '35-39 yrs', '40-44 yrs', '45-49 yrs', '50-54 yrs', '55-59 yrs', '60-64 yrs', '65-69 yrs', '70+']

EDUC (ordered):
['0-8 yrs', 'Some HS', 'HS Grad', 'Some Post-Sec', 'Post-Sec Cert/Dipl', "Bachelor's", "> Bachelor's"]

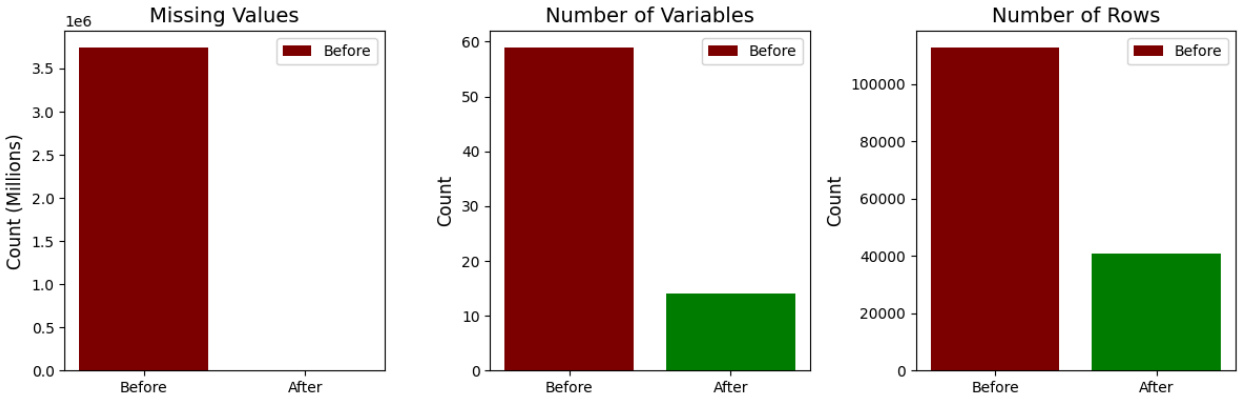
OUTLIER CHECK

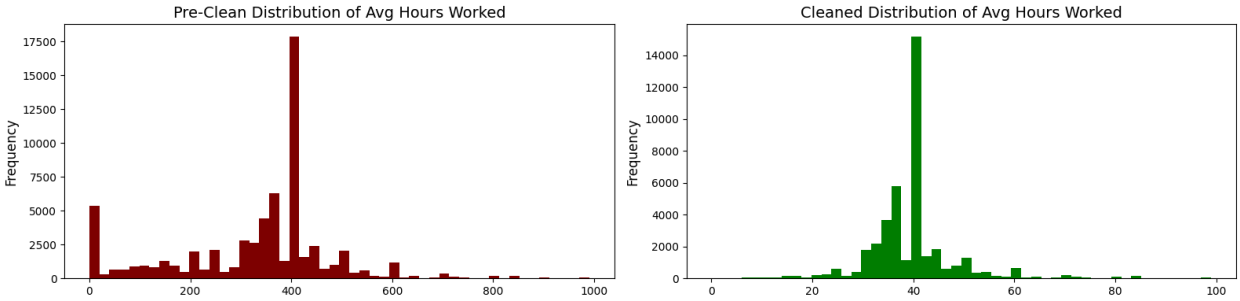
Using the 3 IQR rule, we checked for outliers in the numerical variables.

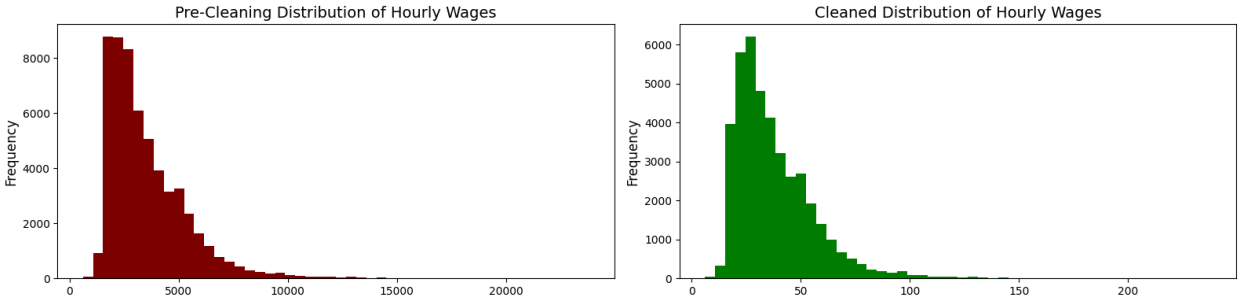
HRLYEARN_T Outliers (n): 264 (0.65%)
AHRSMAIN_T Outliers (n): 3612 (8.84%)
PAIDOT_T Outliers (n): 4830 (11.82%)
UNPAIDOT_T Outliers (n): 3423 (8.38%)
TENURE Outliers (n): 0 (0.00%)
XTRAHRS_T Outliers (n): 8037 (19.67%)

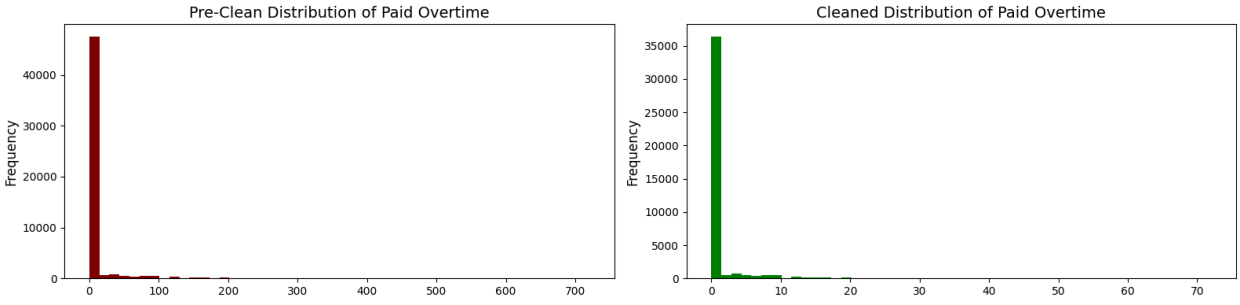
Outliers in HRLYEARN_T (hourly wages), AHRSMAIN_T (weekly hours worked), PAIDOT_T (paid overtime),
UNPAIDOT_T (unpaid overtime), and XTRAHRS_T (extra hours) reflect normal labour-market variation.

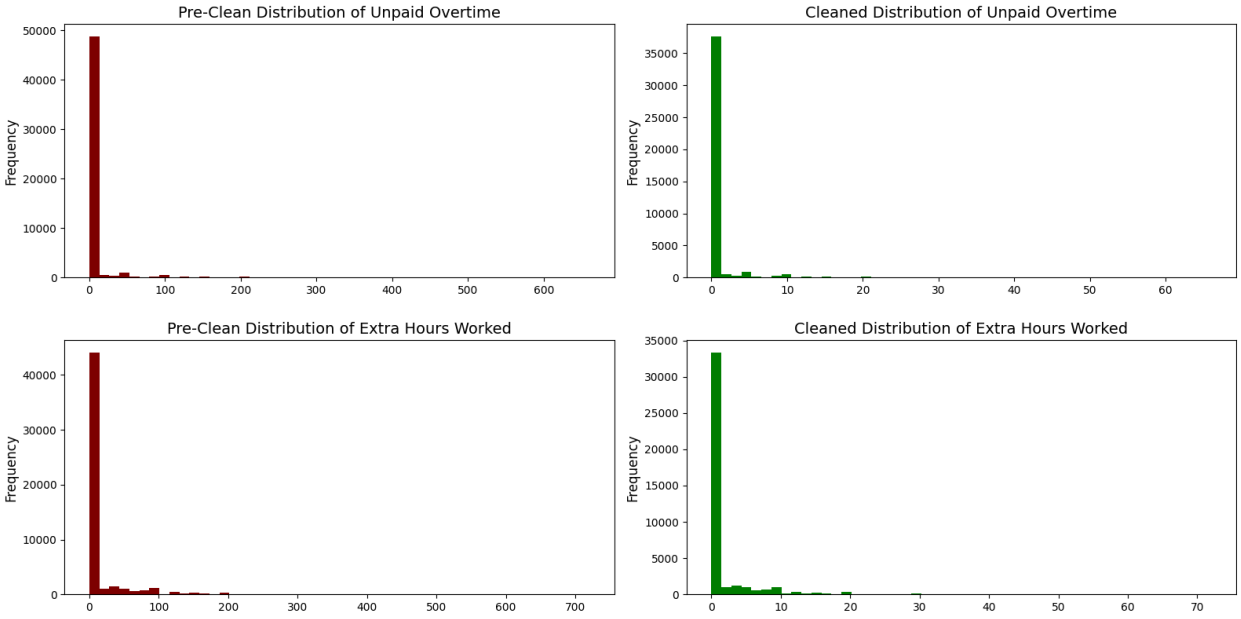
Hours and overtime typically show long right tails, so these values likely represent genuine behaviours worth retaining.


Missing Values
1e6
Count (Millions)
Before
After
Number of Variables
Count
Before
After
Number of Rows
Count
Before
After


Pre-Clean Distribution of Avg Hours Worked
Frequency
Cleaned Distribution of Avg Hours Worked
Frequency


Pre-Cleaning Distribution of Hourly Wages
Frequency
Cleaned Distribution of Hourly Wages
Frequency


Pre-Clean Distribution of Paid Overtime
Frequency
Cleaned Distribution of Paid Overtime
Frequency



MEASURES OF CENTRAL TENDENCY

	mean	median	std	min	max	count
TENURE	93.90	64.0	82.03	1.00	240.00	40868.0
HRLYEARN_T	37.74	33.0	18.60	6.07	237.91	40868.0
AHRSMAIN_T	39.60	40.0	9.29	0.00	99.00	40868.0
PAIDOT_T	1.11	0.0	4.33	0.00	72.00	40868.0
UNPAIDOT_T	0.62	0.0	2.77	0.00	66.00	40868.0
XTRAHRS_T	1.73	0.0	5.07	0.00	72.00	40868.0

The summary statistics indicate generally stable employment patterns. Tenure averages roughly 94
months, though the median of 64 months suggests many workers stay several years while fewer remain
for much longer.

Hourly earnings average about $38, with a median of $33, showing moderate dispersion.

Usual weekly hours centre tightly around 40. Median paid overtime, unpaid overtime, and extra hours
are all zero, meaning most workers report no additional hours. However, the higher means reflect a
smaller group that performs overtime or extra hours more regularly. Overall, the table shows
concentrated distributions with modest variability for most measures.

SKEWNESS & KURTOSIS CHECK

	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

The diagnostics show that TENURE is closest to a normal distribution, with only mild right skew and
light tails. HRLYEARN_T and AHRSMAIN_T are moderately right skewed with heavy tails, reflecting
small groups of very high earners and long hour workers. The overtime variables PAIDOT_T, UNPAIDOT_T, and XTRAHRS_T are extremely right skewed with very high kurtosis.

This indicates that most respondents report little or no overtime, while a small minority report unusually large amounts, creating long and heavy right tails. Overall, the patterns reflect typical labour market asymmetries where extreme values are concentrated among few individuals.

SUMMARIZING CATEGORICAL VARIABLES

Frequency tables helped us gain a better understanding of the distribution of categorical variables.

EDUC Frequency Table

Post-Sec Cert/Dipl	15584
Bachelor's	9535
HS Grad	6992
> Bachelor's	5079
Some HS	1791
Some Post-Sec	1445
0-8 yrs	442

MARSTAT Frequency Table

Married	19879
Single, never married	10636
Common-Law	7235
Divorced	1570
Separated	1259
Widowed	289

AGE_12 Frequency Table

40-44 yrs	5600
35-39 yrs	5489

45-49 yrs 5261
30-34 yrs 5145
50-54 yrs 4876
25-29 yrs 4461
55-59 yrs 4030
60-64 yrs 3124
20-24 yrs 2426
15-19 yrs 456
65-69 yrs 0
70+ 0

PERMTEMP Frequency Table

Permanent 37506
Temp Term/Contract 1907
Temp Seasonal 963
Temp Other/Casual 492

UNION Frequency Table

Non-Unionized 26649
Union Member 13315
Unionized Non-Member 904

MJH Frequency Table

Single job 39006
Multiple jobs 1862

GENDER Frequency Table

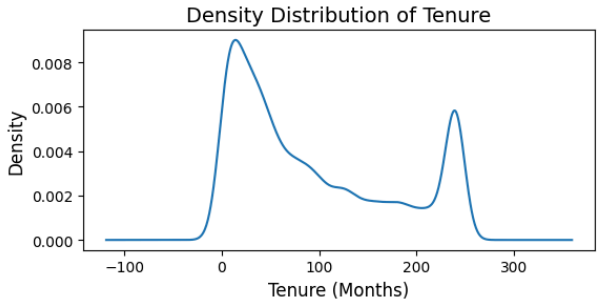
Male 22195
Female 18673

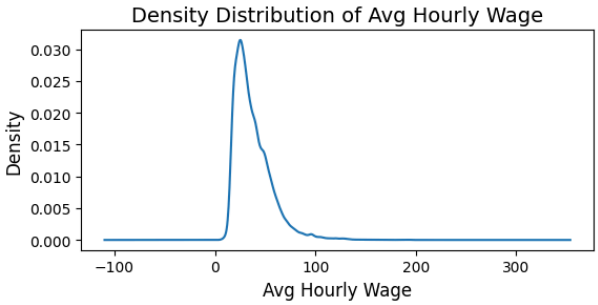
COWMAIN Frequency Table

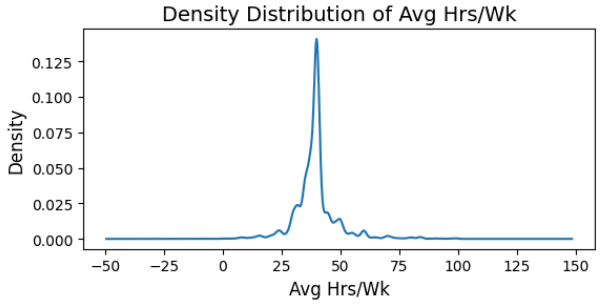
Private sector 29267
Public sector 11601

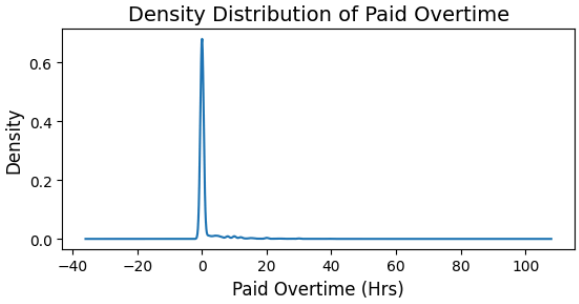
VISUALIZING NUMERICAL VARIABLES

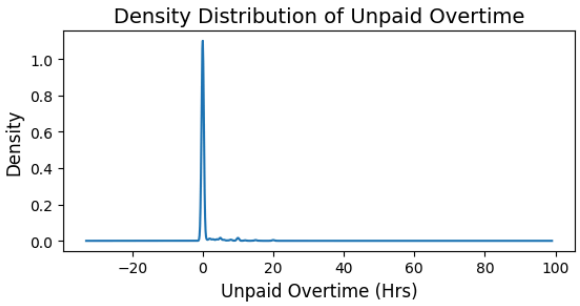
We then used density plots to visually complement the earlier numerical summaries, helping to confirm patterns in skewness, kurtosis, and overall central tendency.

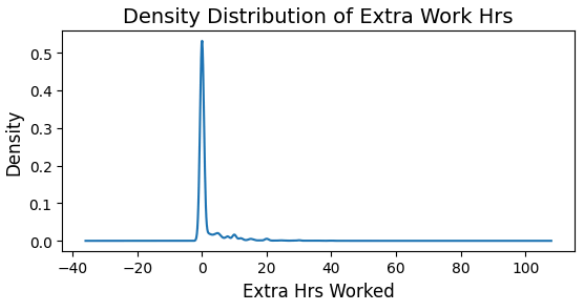








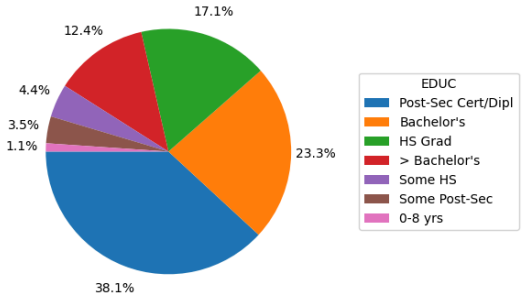




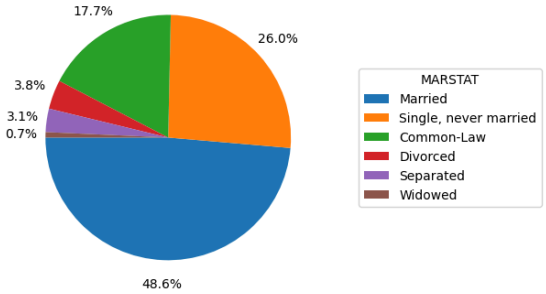
VISUALIZING CATEGORICAL VARIABLES

Using pie charts to complement our frequency tables, we visualized the proportion of the categorical variables.

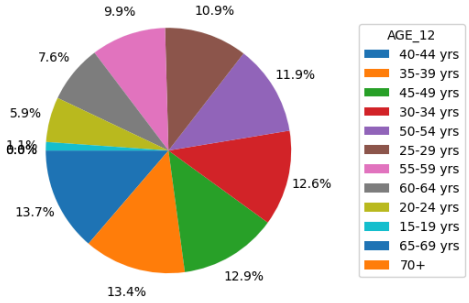
Education Proportion



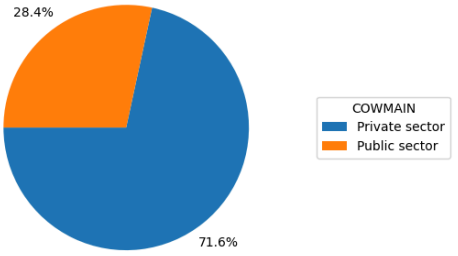
Marital Status Proportion



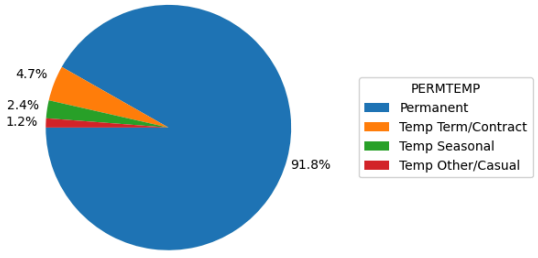
Age Proportion



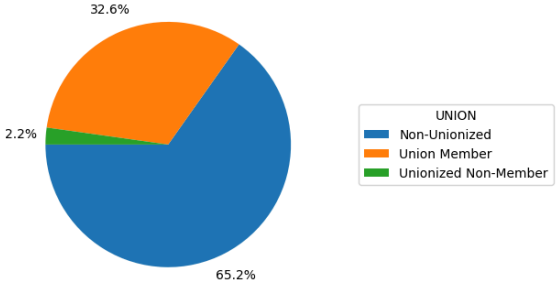
Worker Class Proportion



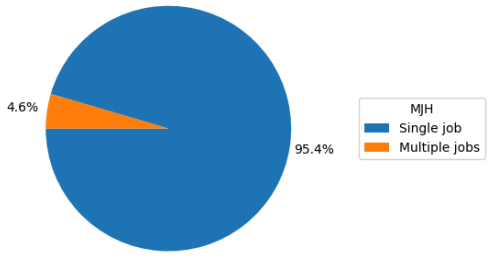
Job Permanency Proportion



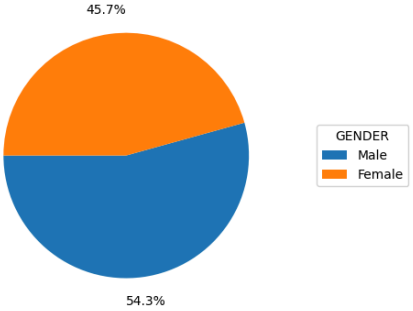
Unionization Proportion

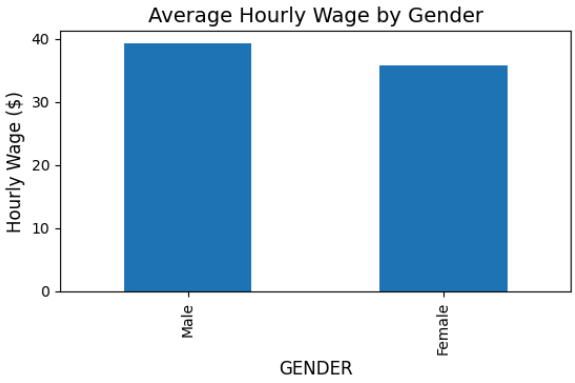


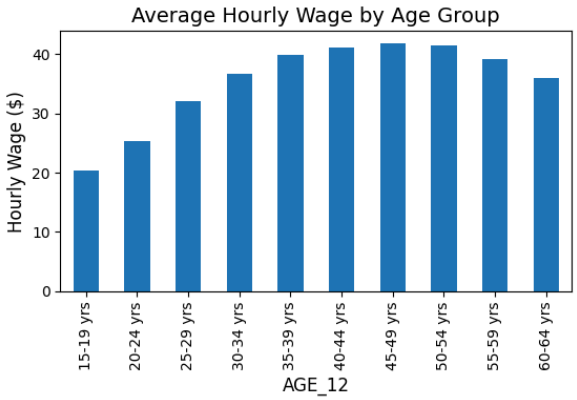
Job-Multiplicity Proportion

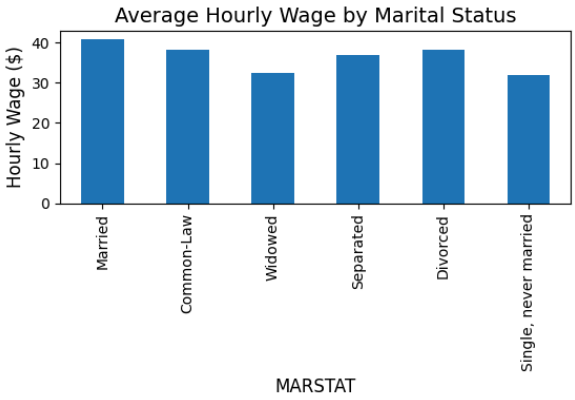


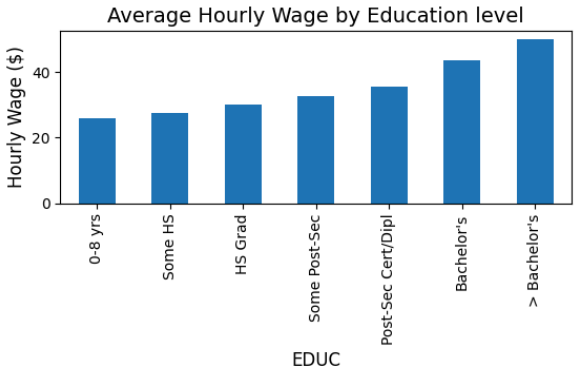
Gender Proportion

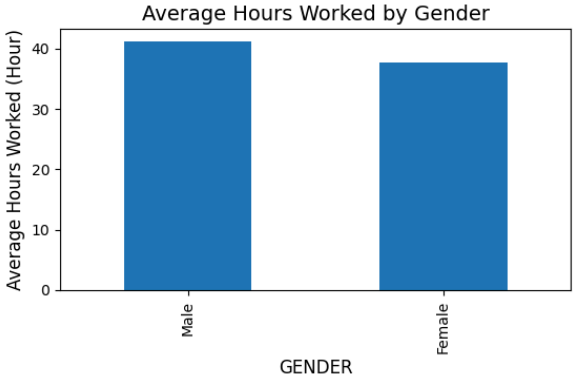


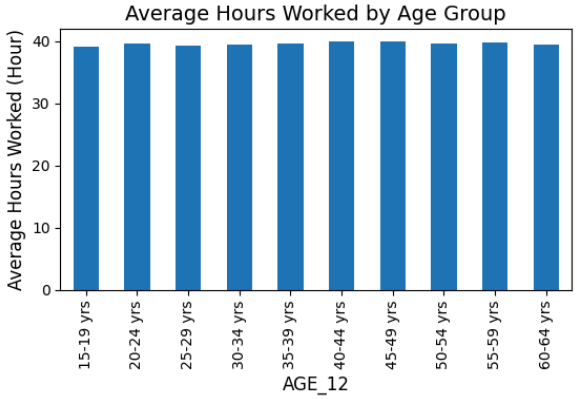

Average Hourly Wage by Gender
Hourly Wage ($)
0
10
20
30
40
Male
Female
GENDER

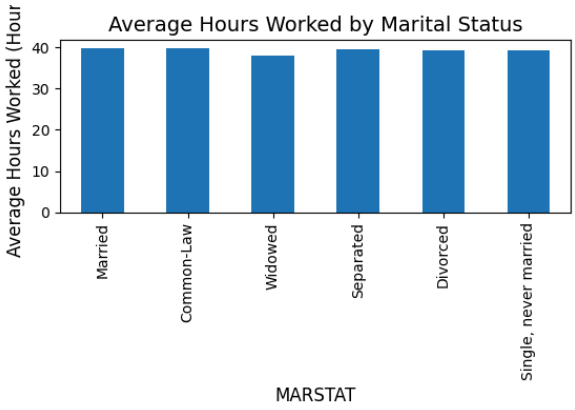

Average Hourly Wage by Age Group
Hourly Wage ($)
0
10
20
30
40
15-19 yrs
20-24 yrs
25-29 yrs
30-34 yrs
35-39 yrs
40-44 yrs
45-49 yrs
50-54 yrs
55-59 yrs
60-64 yrs
AGE_12

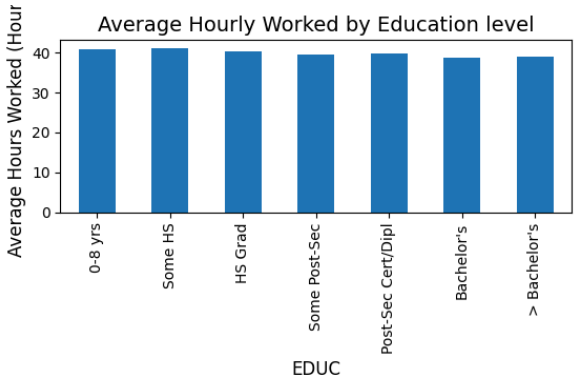

Average Hourly Wage by Marital Status
Hourly Wage ($)
0
10
20
30
40
Married
Common-Law
Widowed
Separated
Divorced
Single, never married
MARSTAT

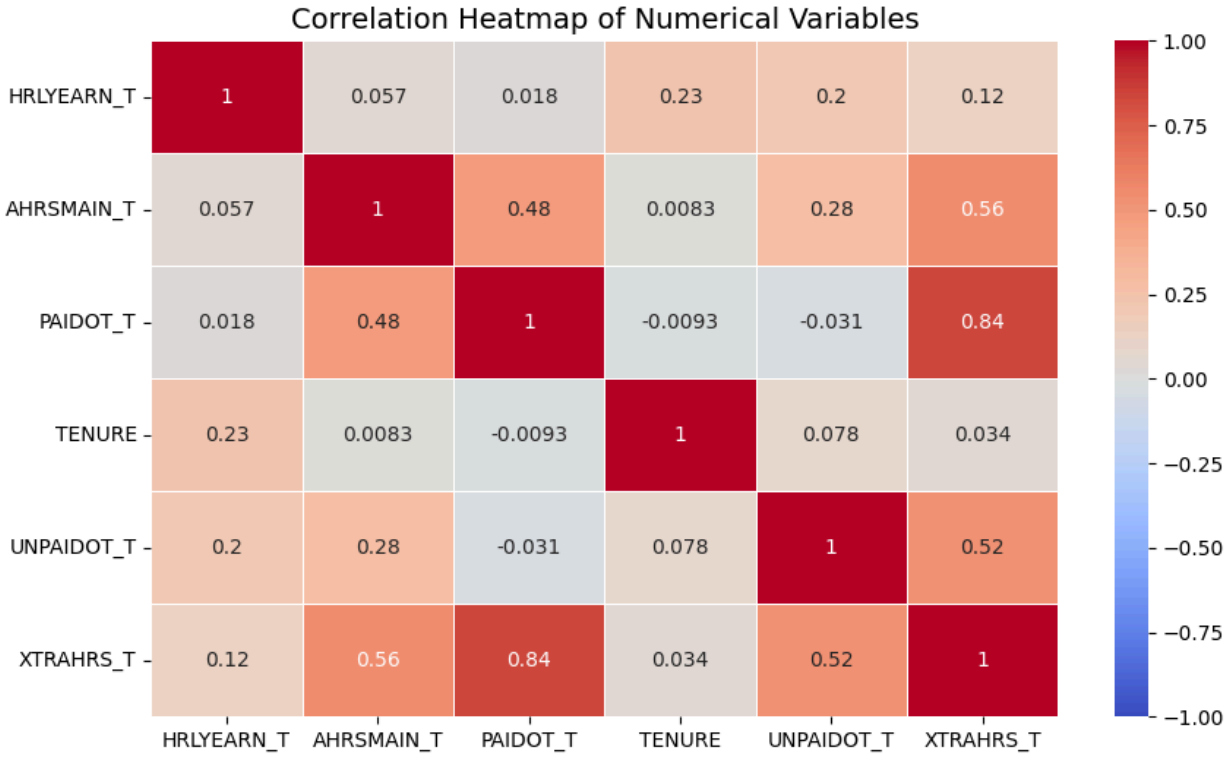

Average Hourly Wage by Education level
Hourly Wage ($)
0
20
40
0-8 yrs
Some HS
HS Grad
Some Post-Sec
Post-Sec Cert/Dipl
Bachelor's
> Bachelor's
EDUC

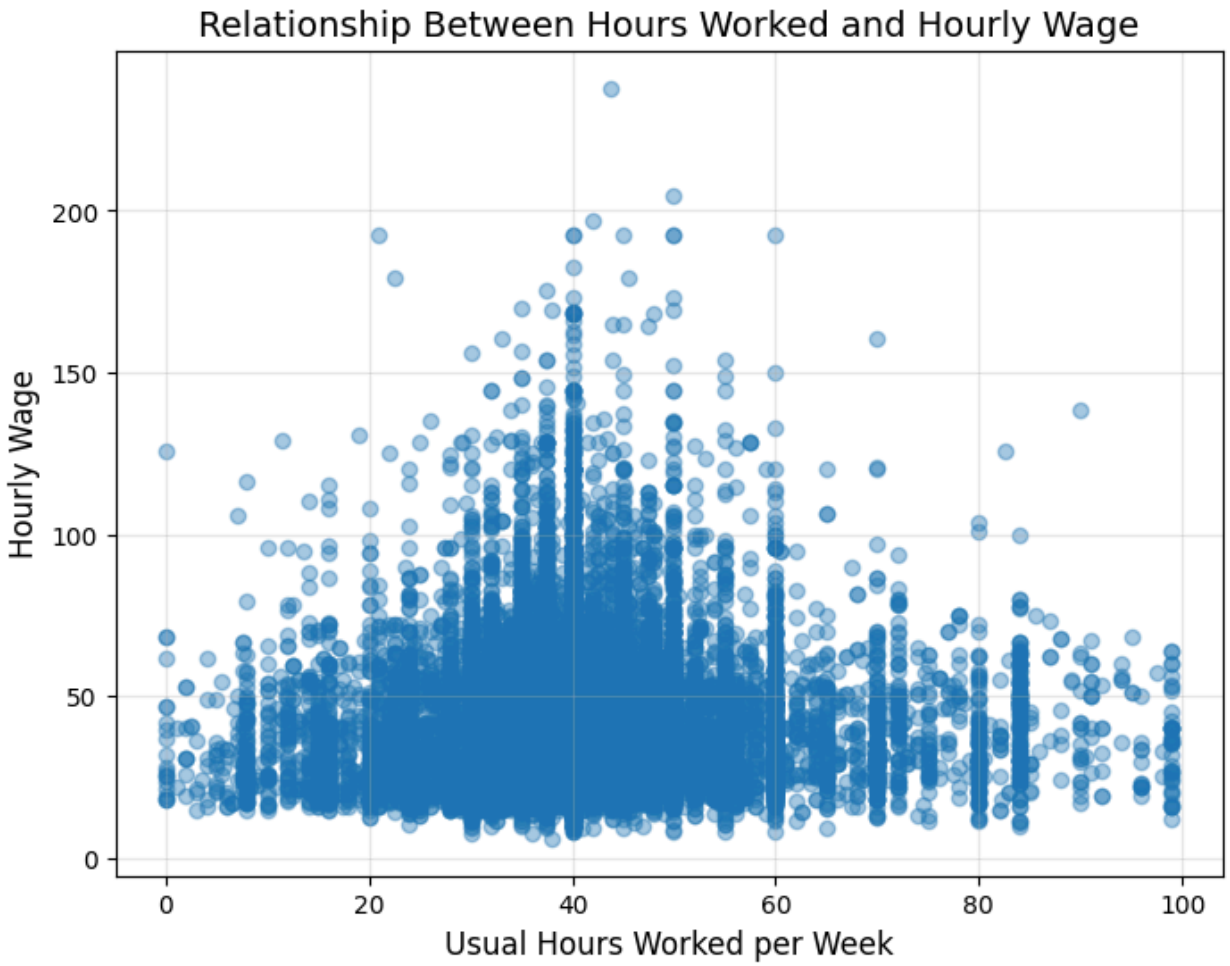

Average Hours Worked by Gender
Average Hours Worked (Hour)
0
10
20
30
40
Male
Female
GENDER

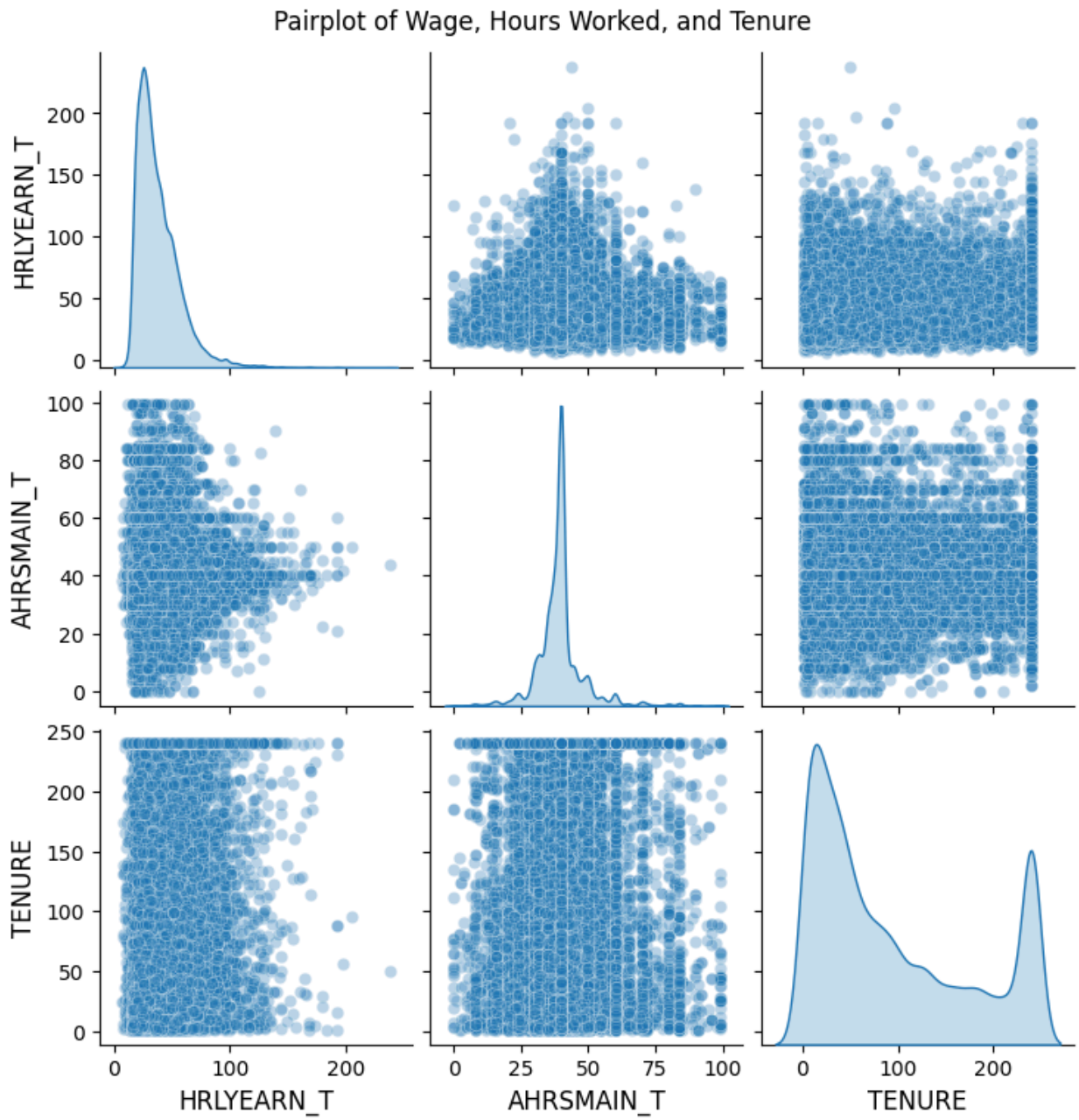

Average Hours Worked by Age Group
Average Hours Worked (Hour)
0
10
20
30
40
15-19 yrs
20-24 yrs
25-29 yrs
30-34 yrs
35-39 yrs
40-44 yrs
45-49 yrs
50-54 yrs
55-59 yrs
60-64 yrs
AGE_12


Average Hours Worked by Marital Status
Average Hours Worked (Hour
0
10
20
30
40
Married
Common-Law
Widowed
Separated
Divorced
Single, never married
MARSTAT


Average Hourly Worked by Education level
Average Hours Worked (Hour
0
10
20
30
40
0-8 yrs
Some HS
HS Grad
Some Post-Sec
Post-Sec Cert/Dipl
Bachelor's
> Bachelor's
EDUC


Correlation Heatmap of Numerical Variables
HRLYEARN_T
AHRSMAIN_T
PAIDOT_T
TENURE
UNPAIDOT_T
XTRAHRS_T
1
0.057
0.018
0.23
0.2
0.12
0.057
1
0.48
0.0083
0.28
0.56
0.018
0.48
1
-0.0093
-0.031
0.84
0.23
0.0083
-0.0093
1
0.078
0.034
0.2
0.28
-0.031
0.078
1
0.52
0.12
0.56
0.84
0.034
0.52
1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Relationship Between Hours Worked and Hourly Wage
Hourly Wage
0
50
100
150
200
0
20
40
60
80
100
Usual Hours Worked per Week



Crosstab: EDUC × Hourly Wage Category (%)

HRLYEARN_CAT	$0–20/hr	$20–40/hr	$40–60/hr	$60–80/hr	$80+/hr
0-8 yrs	32.8	60.2	5.7	1.4	0.0
Some HS	28.4	61.1	9.2	0.9	0.4
HS Grad	22.5	61.0	13.1	2.5	0.8
Some Post-Sec	17.6	60.3	17.1	3.5	1.5
Post-Sec Cert/Dipl	10.3	60.0	23.4	4.7	1.5
Bachelor's	9.1	40.0	34.9	10.9	5.1
> Bachelor's	6.2	31.0	35.8	18.2	8.9

Crosstab: MARSTAT × Hourly Wage Category (%)

HRLYEARN_CAT	$0–20/hr	$20–40/hr	$40–60/hr	$60–80/hr	$80+/hr
Married	10.2	47.8	28.0	9.6	4.4
Common-Law	8.9	54.7	27.6	6.3	2.4
Widowed	20.8	54.7	18.7	4.5	1.4
Separated	11.4	55.9	24.9	5.4	2.5
Divorced	10.7	52.5	26.1	7.8	2.8
Single, never married	21.0	57.2	17.1	3.6	1.2

Crosstab: AGE_12 × Hourly Wage Category (%)

HRLYEARN_CAT	$0–20/hr	$20–40/hr	$40–60/hr	$60–80/hr	$80+/hr
15-19 yrs	62.3	37.3	0.4	0.0	0.0
20-24 yrs	32.0	62.1	5.6	0.3	0.0
25-29 yrs	17.2	60.2	19.6	2.5	0.5
30-34 yrs	11.3	54.7	27.1	5.2	1.8
35-39 yrs	9.8	49.8	28.8	8.2	3.4
40-44 yrs	9.3	47.2	29.9	9.8	3.9
45-49 yrs	8.5	46.7	30.0	10.7	4.1
50-54 yrs	9.7	47.8	27.7	9.8	4.9
55-59 yrs	11.0	51.8	24.7	8.3	4.2
60-64 yrs	14.4	57.6	18.4	6.1	3.6

Crosstab: PERMTEMP × Hourly Wage Category (%)

HRLYEARN_CAT	$0–20/hr	$20–40/hr	$40–60/hr	$60–80/hr	$80+/hr
Permanent	12.1	51.3	25.6	7.6	3.3
Temp Seasonal	30.7	62.6	6.3	0.2	0.1
Temp Term/Contract	14.5	58.0	22.3	3.7	1.4
Temp Other/Casual	31.5	52.8	13.6	1.4	0.6

Crosstab: UNION × Hourly Wage Category (%)

HRLYEARN_CAT	$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	28.0	8.8	3.1
Non-Unionized	17.2	52.1	19.4	7.3	4.0

Crosstab: MJH × Hourly Wage Category (%)

HRLYEARN_CAT	$0–20/hr	$20–40/hr	$40–60/hr	$60–80/hr	$80+/hr
Single job	12.7	51.8	25.1	7.3	3.1
Multiple jobs	17.5	55.4	19.8	4.5	2.8

Crosstab: GENDER × Hourly Wage Category (%)

HRLYEARN_CAT $0–20/hr $20–40/hr $40–60/hr $60–80/hr $80+/hr
Male 11.3 50.9 25.7 8.0 4.1
Female 14.8 53.2 23.9 6.2 1.9

Crosstab: COWMAIN × Hourly Wage Category (%)

HRLYEARN_CAT $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

OLS Regression Results
==========

Dep. Variable: HRLYEARN_T R-squared: 0.263
Model: OLS Adj. R-squared: 0.263
Method: Least Squares F-statistic: 455.9
Date: Wed, 03 Dec 2025 Prob (F-statistic): 0.00
Time: 15:53:24 Log-Likelihood: -1.7121e+05
No. Observations: 40868 AIC: 3.425e+05
Df Residuals: 40835 BIC: 3.428e+05
Df Model: 32
Covariance Type: nonrobust
==========

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	23.1070	1.181	19.572	0.000	20.793	25.421
C(COWMAIN)[T.Private sector]	-4.4129	0.234	-18.884	0.000	-4.871	-3.955
C(AGE_12)[T.20-24 yrs]	0.3946	0.821	0.481	0.631	-1.215	2.004
C(AGE_12)[T.25-29 yrs]	2.2848	0.801	2.852	0.004	0.714	3.855
C(AGE_12)[T.30-34 yrs]	4.9889	0.803	6.211	0.000	3.415	6.563
C(AGE_12)[T.35-39 yrs]	6.8342	0.807	8.466	0.000	5.252	8.416
C(AGE_12)[T.40-44 yrs]	7.3288	0.812	9.026	0.000	5.737	8.920
C(AGE_12)[T.45-49 yrs]	7.5340	0.817	9.220	0.000	5.932	9.136
C(AGE_12)[T.50-54 yrs]	7.2049	0.822	8.769	0.000	5.595	8.815
C(AGE_12)[T.55-59 yrs]	5.8318	0.828	7.043	0.000	4.209	7.455
C(AGE_12)[T.60-64 yrs]	3.9025	0.839	4.649	0.000	2.257	5.548
C(AGE_12)[T.65-69 yrs]	-4.47e-15	7.7e-16	-5.803	0.000	-5.98e-15	-2.96e-15
C(AGE_12)[T.70+]	-7.153e-16	1.04e-15	-0.685	0.493	-2.76e-15	1.33e-15
C(EDUC)[T.Some HS]	1.6741	0.849	1.971	0.049	0.009	3.339
C(EDUC)[T.HS Grad]	4.3290	0.786	5.506	0.000	2.788	5.870
C(EDUC)[T.Some Post-Sec]	7.3775	0.872	8.458	0.000	5.668	9.087
C(EDUC)[T.Post-Sec Cert/Dipl]	8.7472	0.775	11.284	0.000	7.228	10.267
C(EDUC)[T.Bachelor's]	16.5325	0.786	21.023	0.000	14.991	18.074
C(EDUC)[T.> Bachelor's]	21.8849	0.803	27.246	0.000	20.311	23.459

```
C(GENDER)[T.Female]                -6.3741      0.167    -38.249      0.000      -6.701      -6.047
C(MARSTAT)[T.Common-Law]            0.4621      0.225      2.051      0.040       0.020       0.904
C(MARSTAT)[T.Widowed]              -3.4051      0.953     -3.572      0.000      -5.274      -1.536
C(MARSTAT)[T.Separated]            -0.9344      0.466     -2.005      0.045      -1.848      -0.021
C(MARSTAT)[T.Divorced]             -0.1795      0.423     -0.424      0.672      -1.009       0.650
C(MARSTAT)[T.Single, never married]  -2.5039      0.224    -11.202      0.000      -2.942      -2.066
C(MJH)[T.Multiple jobs]            -3.6316      0.380     -9.552      0.000      -4.377      -2.886
C(PERMTEMP)[T.Temp Seasonal]       -5.7307      0.530    -10.813      0.000      -6.769      -4.692
C(PERMTEMP)[T.Temp Term/Contract]    -2.1648      0.383     -5.649      0.000      -2.916      -1.414
C(PERMTEMP)[T.Temp Other/Casual]     -5.5176      0.730     -7.563      0.000      -6.948      -4.088
C(UNION)[T.Unionized Non-Member]      1.8887      0.552      3.419      0.001       0.806       2.971
C(UNION)[T.Non-Unionized]          1.4623      0.216      6.763      0.000       1.039       1.886
TENURE                             0.0367      0.001     32.120      0.000       0.034       0.039
AHRSMAIN_T                         0.0040      0.011      0.378      0.705      -0.017       0.025
PAIDOT_T                           0.1541      0.021      7.218      0.000       0.112       0.196
UNPAIDOT_T                         0.8227      0.031     26.566      0.000       0.762       0.883
==============================================================================
Omnibus:                    17089.372   Durbin-Watson:                   2.007
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           132286.453
Skew:                           1.829   Prob(JB):                         0.00
Kurtosis:                      11.019   Cond. No.                     1.26e+16
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 4.24e-24. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Median hourly wage (cutoff): $33.00

```
HIGH_WAGE
Proportion high wage        0.505922
Proportion not-high wage    0.494078
Name: proportion, dtype: float64
```

TRAIN/TEST SPLIT

Training set size: 30651 rows

Testing set size: 10217 rows

HIGH WAGE PROPORTION IN TRAINING SET (%)

HIGH_WAGE
1   50.5
0   49.5
Name: proportion, dtype: float64

Model 1 — Logistic Regression Confusion Matrix

TP: 3724 TP%: 71.64293959215082
FP: 1505 FP%: 29.986052998605288
FN: 1474 FN%: 28.357060407849175
TN: 3514 TN%: 70.01394700139471
Accuracy: 70.84271312518352
Precision: 71.2182061579652
Negative Predictive Value: 70.44907778668805

Model 2 — KNN (k=7) Confusion Matrix

TP: 3527 TP%: 67.85302039245865
FP: 1773 FP%: 35.32576210400478
FN: 1671 FN%: 32.146979607541354
TN: 3246 TN%: 64.67423789599522
Accuracy: 66.29147499265929
Precision: 66.54716981132076
Negative Predictive Value: 66.01586333129957

Model 3 — Random Forest Confusion Matrix

TP: 3668 TP%: 70.56560215467488
FP: 1488 FP%: 29.647340107591162
FN: 1530 FN%: 29.434397845325122
TN: 3531 TN%: 70.35265989240884
Accuracy: 70.46099637858472
Precision: 71.14041892940264
Negative Predictive Value: 69.76882039122702

MODEL ACCURACY COMPARISON

Logistic Regression: 0.708
KNN (k=7): 0.663

Random Forest: 0.705