

WageCAN

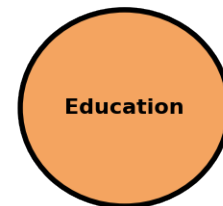
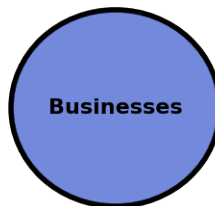
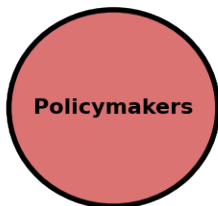
Analyzing Wages in Canada



Nikolay Shlepov | MS DS Student | CU Boulder
Project Progress Presentation, April 2024

Problem Context

- Wages are a key indicator of economic health, labor stability, and inequality
- COVID-19, inflation, and technological changes have reshaped labor markets
- Lack of structured, longitudinal, predictive wage studies in Canada
- Critical for policymakers, businesses, educators, and job seekers

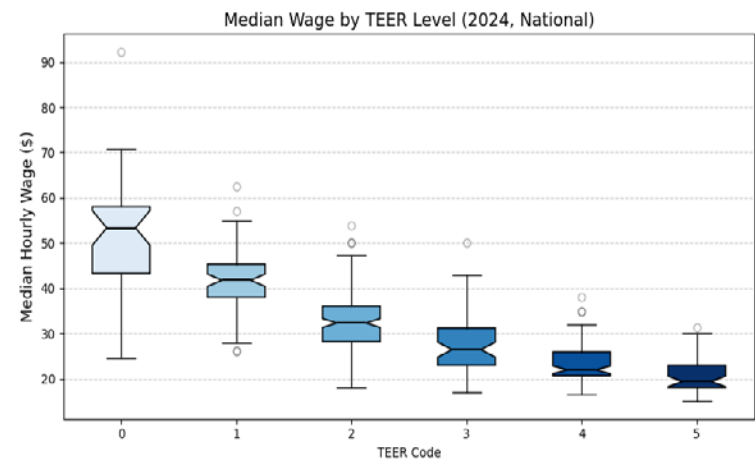


Project Summary

- Analyzed wages in Canada (2016–2024)
- Examined national, provincial, and regional disparities
- Identified **main wage drivers**
- Combined **descriptive, unsupervised, and supervised** methods
- Tested the stability of wage structures from 2016 to 2024

Key Findings

- **TEER level** (Training, Education, Experience, Responsibilities) is the strongest predictor of wages
- **Occupational category (BOC)** has secondary influence
- **Provincial effects** are modest
- **Regional outliers** exist but are not dominant drivers
- **Wage structures remained stable** from 2016 to 2024



Inside NOC 2021

5

Broad Category

Occupations in art, culture, recreation and sport

3

TEER

College diploma or ≥ 6 month training

53

Major Group

Occupations in art, culture and sport

532

Sub-Major Group

Occupations in sports

5320

Minor Group

Athletes, coaches, referees and related occupations

53201

Coaches





Government
of Canada

Gouvernement
du Canada



BANK OF CANADA
BANQUE DU CANADA

Methodology Overview

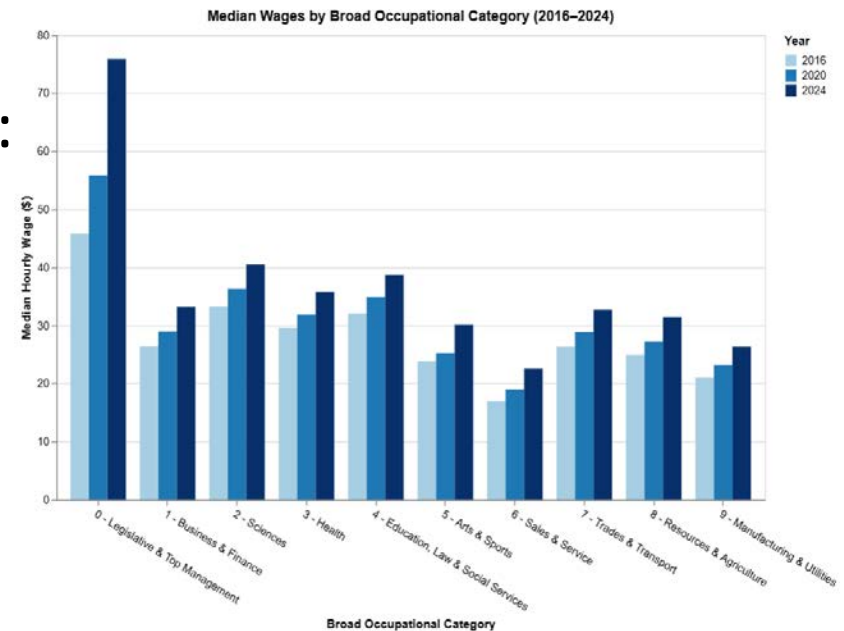
- Collected and unified Canadian wage datasets (2016, 2020, 2024)
- Mapped NOC 2016 codes to NOC 2021 codes
- Engineered features: TEER, Broad Category, Major Group
- Conducted Descriptive Analysis, Clustering and Predictive Modeling
- Evaluated correlation, clustering coherence, and model accuracy

Data Source and Preprocessing

- **Source** - Government of Canada wage reports (open.canada.ca)
- Standardized **column** names and **formats**
- **Mapped** occupations from **NOC 2016 to NOC 2021**
- Cleaned **occupation** titles and **province** names
- **Filtered** outdated entries and **handled** missing data
- Created unified dataset (**11,448 records**) ready for analysis

EDA - National-Level Insights

- **TEER** is the strongest predictor of wages
- Broad Occupational Category (**BOC**) shows secondary but noticeable impact
- Consistent wage stratification:
 - TEER 0–1 → Highest wages
 - TEER 4–5 → Lowest wages
- Stable patterns across 2016, 2020, and 2024 datasets



Provincial and Regional Analysis

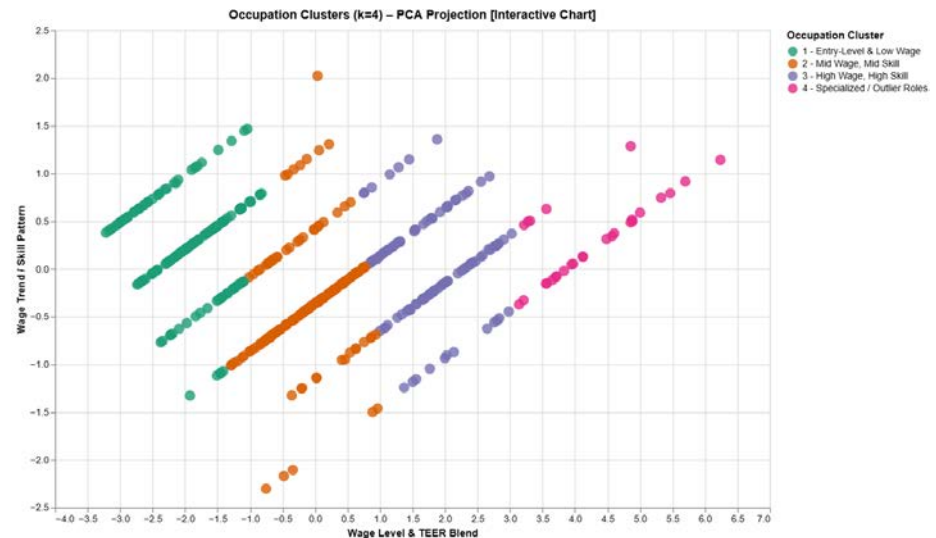
- Provinces have **minor impact** on wages compared to TEER
- **Alberta, BC, and Ontario** show **slightly higher** median wages
- **Regional outliers** identified in some economic zones
- Regional effects are **not significant** overall

Median Wage Distribution by Province (2016, 2020, 2024)



Pattern Detection via Clustering

- **K-Means and Hierarchical Clustering** on occupational wage data
- Clusters based on **TEER level** and **median wages** (2016–2024)
- **High TEER** and skill level → consistently **higher wages**
- Wage structures are stable across provinces
- PCA confirms strong structure (first PC >95% variance)



Supervised ML Wage Prediction

- **Built models using:**

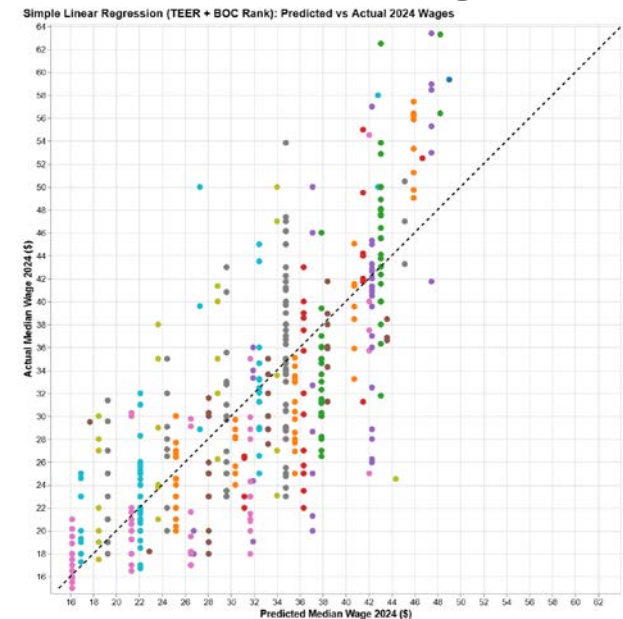
TEER, BOC, Province, Embedded Titles

- **Models:**

Linear Regression, Random Forest, Gradient Boosting

- **Key Insight:**

- Simple models (**TEER + BOC**) perform nearly as well as complex ones
- Embedded titles add little predictive power and risk overfitting

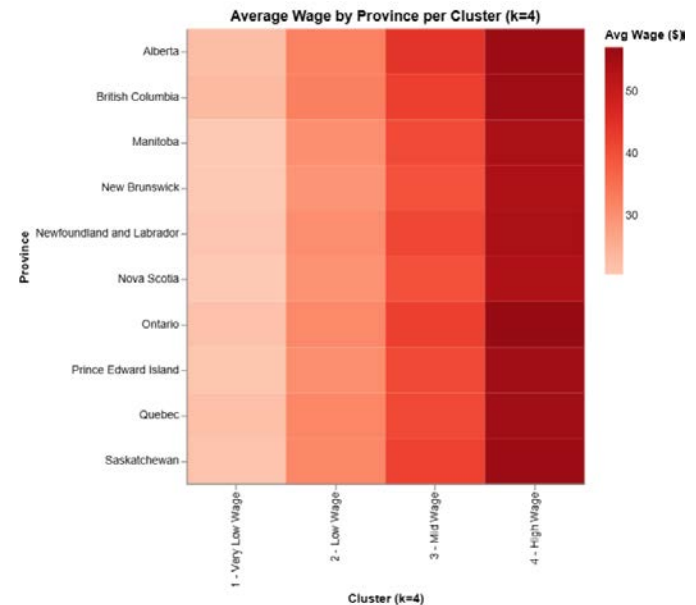


Wage Stability Modelling

- **Goal:**
Predict 2024 wages using 2016 and 2020 data only
- **Models:**
Linear Regression, Random Forest, Gradient Boosting
- **Results:**
 - **High accuracy** ($R^2 > 0.90$)
 - **Low RMSE** across all models
- **Conclusion:**
Wage growth patterns are highly stable and predictable, even through global disruptions

Evaluation - Correlation and Clustering

- Strong negative **TEER-wage** correlation:
Pearson -0.75 , Spearman -0.77
- Moderate negative **BOC-wage** correlation:
Pearson -0.34 , Spearman -0.31
- **Clustering validated:**
 - Clear **TEER-wage** clusters
 - PCA $>95\%$ variance explained
 - Highest-paid clusters showed higher dispersion

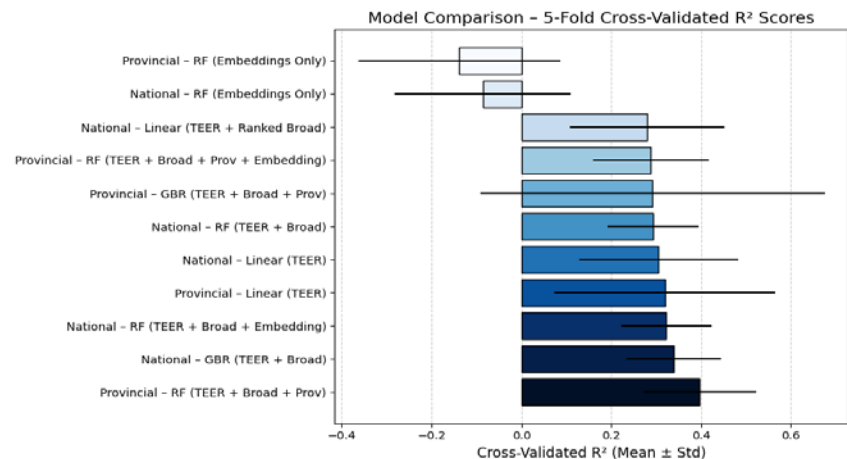


Evaluation – Model Performance

- Compared models using
 - R^2 (explained variance) and **RMSE** (prediction error)
 - 5-fold-cross-validation to assess generalization

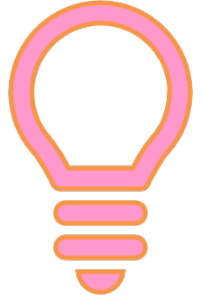
- **Results:**

- TEER is the strongest predictor of wages
- Adding BOC and Province gives only minor gains
- Embeddings caused clear overfitting (negative R^2)





Challenges & Solutions



- **NOC 2016 → NOC 2021 mapping:**
Resolved with concordance tables + LLM assisted refinement
- **Missing Values:**
Imputed using national medians or removed
- **Data integrity:**
Maintained across merged datasets
- **Pipeline design:**
Modular pipeline built for efficient EDA and modelling

Potential Next Steps

- Expand **longitudinal modelling** to future datasets (2026, 2028)
- Communicate findings to Statistics Canada and stakeholders
- Explore advanced modelling (e.g., neural networks with TensorFlow, PyTorch)
- Monitor for structural shifts in labor markets over time



Biases and Limitations

- **Data Structure Bias:** Dataset differences across years
- **Mapping Uncertainty:** Potential manual errors during NOC transitions
- **Selection Bias:** Excluded low-data territories
- **Overfitting Risk:** Complex models showed optimism bias
- **External Factors:** Future wage shifts (e.g., policy, technology) may break historical trends

Conclusion

- **TEER level** remains the strongest predictor of wages
- Wage structures were **stable and predictable** from 2016–2024
- **Provincial and regional** effects are minor compared to training and education
- **GitHub Repository:**
[https://github.com/nickshlepov/WageCAN Project](https://github.com/nickshlepov/WageCAN)

WageCAN

Analyzing Wages in Canada