WageCAN: Analyzing Wages in Canada (2013-2023)

Nikolay Shlepov
MS DS Student, CU Boulder
Boulder CO
Nikolay.Shlepov@colorado.edu

ABSTRACT

Intent:

This study, WageCAN: Analyzing Wage Trends in Canada (2013–2023), examines national and provincial wage patterns using datasets from the 2016, 2020, and 2024 Canadian Government wage reports. The project combines descriptive analysis and supervised learning models to identify wage disparities, assess post-pandemic shifts, and evaluate the predictive stability of wage trends.

Key Findings:

The analysis confirms that **TEER level** - reflecting education and training requirements - is the primary factor of wage variation across Canada. Occupational category adds minor additional explanatory power, while regional differences are comparatively modest. Supervised learning models reinforced these findings, validating that simple models based on TEER already achieve strong explanatory performance. More complex models, including title embeddings and multi-feature approaches, provided limited additional predictive accuracy and exhibited signs of overfitting. Wage patterns have become increasingly stable between 2016 and 2024, supporting cautious optimism for forecasting future wage trends.

INTRODUCTION

Wages are a fundamental indicator of economic health, labor market stability, and income inequality. Wage trends are shaped by a complex interplay of factors, including inflation, technological change, labor demand, public policy, and major disruptions such as the COVID-19 pandemic. Despite the availability of wage data, there remains a limited number of structured, longitudinal analyses examining how wages have evolved across industries, regions, and occupational categories over the past decade. Furthermore, predicting future wage trends remains challenging due to the dynamic and multifactorial nature of labor markets.

Understanding wage trends is critical for a range of stakeholders:

- **Policymakers** need insights to address wage inequality, set minimum wage policies, and respond to economic shifts.
- **Businesses** rely on wage analysis for workforce planning, salary benchmarking, and labor cost management.
- **Job seekers and employees** benefit from wage transparency when making career decisions.
- **Educational institutions** can align curriculum development with high-growth, high-wage sectors to better prepare students for evolving labor demands.

Additionally, the COVID-19 pandemic significantly disrupted the labor market, making it crucial to analyze pre-pandemic, pandemic, and post-pandemic wage changes.

Limitations of Existing Solutions

- **Lack of predictive analysis**: Many wage studies focus on historical trends but do not provide data-driven wage forecasting.
- **Limited integration of machine learning**: Most wage reports rely on descriptive statistics, rather than leveraging supervised learning for predictive insights.
- **Industry and regional disparities**: National-level wage analyses often overlook sector-specific and regional variations, leading to generalized conclusions that may not reflect local labor market conditions.

WageCAN was designed to address these gaps.

This project conducted a structured analysis of Canadian wage trends across three key reference points (2016, 2020, and 2024), incorporating descriptive statistics, clustering analysis, and supervised learning models. Its primary contributions are:

- **Identifying TEER levels** (training, education, and experience requirements) **as by far the most dominant factor** influencing median wages.
- **Confirming a stable occupational wage stratification**, with regional and sectoral differences playing a secondary role.
- Demonstrating that **machine learning models, while useful for confirming relationships**, offered limited additional predictive power due to consistent overfitting risks in a relatively small and complex dataset.
- Providing an **open, reproducible, and data-driven framework** to analyze wage trends and predict future developments.

By combining traditional statistical methods with modern machine learning techniques, WageCAN offers a comprehensive and transparent contribution to understanding the evolving dynamics of wages in Canada.

RELATED WORK

Wage analysis is an area where significant research has been conducted historically and recently. Below are key references and their relevance to the current study.

1. Tools and Data Sources

Statistics Canada's Labour Force Survey (LFS):

The LFS is the timeliest source of data on Canadian wages. The survey uses a rotating panel design where each responding household is surveyed monthly for six consecutive months. In the first month, all employed respondents are asked about their wages. However, in subsequent months, respondents are asked about their wages only if they report specific changes to their employment information [1].

Job Vacancy and Wage Survey (JVWS):

Conducted by Statistics Canada, the JVWS collects data directly from respondents using electronic questionnaires. It provides insights into job vacancies and wage statistics across various sectors [2].

Employment and Social Development Canada (ESDC) and Statistics Canada Methodology:

The ESDC–Statistics Canada partnership provides wage benchmarks (low, median, and high) used by resources like Job Bank [3].

2. Methods and Approaches

Machine Learning Integration:

Recent studies have examined the impact of machine learning (ML) on the labor market. For instance, research by the Institute for Work & Health used novel analytical approaches to assess the extent to which different jobs in Canada may be exposed to ML, highlighting potential inequities [4].

AI Exposure Assessment:

Experimental analyses by Statistics Canada estimate the proportion of jobs susceptible to AI-driven transformation [5].

Wage Growth Measurement Techniques:

The Bank of Canada has developed methods that move beyond averages, offering a more sophisticated view of wage growth trends [1].

3. Results and Findings

Impact of ML on Employment:

Approximately 1.9 million Canadians work in occupations with high exposure to machine learning, raising concerns about demographic disparities [6].

Wage Growth Trends:

Post-pandemic job postings showed accelerated wage growth initially, but growth slowed starting mid-2022, indicating labor market rebalancing [7].

WageCAN's Contribution

WageCAN builds on these foundations by conducting a structured, multi-year analysis of Canadian wages, focusing explicitly on both historical patterns and predictive modeling.

Specifically, **WageCAN**:

- **Utilized, cleaned and integrated datasets** from the LFS, JVWS, and ESDC methodologies (2016, 2020, and 2024 cycles).
- **Applied unsupervised clustering** to reveal hidden structure in occupational and regional wage distributions.
- **Implemented supervised learning models** (Random Forest, Gradient Boosting, Embedding-based models) to test predictive relationships.
- **Identified TEER level** (training, education, and experience requirements) **as by far the strongest wage determinant**, exceeding both occupational category and provincial differences.
- Demonstrated that **machine learning models confirmed known relationships** but faced consistent overfitting risks, highlighting limits of predictability given available data.
- Provided an open and **reproducible framework**, allowing policymakers, businesses, educators, and job seekers to better interpret wage dynamics across sectors and provinces [8].

Thus, WageCAN extends beyond descriptive reporting by offering a data-driven, technically rigorous view of Canadian wage evolution, grounded in modern machine learning practices but maintaining critical awareness of model limitations.

METHODOLODY

1. Data Collection and Preprocessing

This study uses three primary datasets from the Government of Canada’s wage reports: **2016**, **2020**, and **2024**. These datasets provide wage records by occupation, province/region, and economic region code (ER_Code), and were downloaded from open.canada.ca [8]. No modifications were made to the raw files before processing.

Each dataset contains approximately **43,000 to 44,000** records and follows a similar schema, including low, median, and high wage values, NOC codes, occupation titles, region identifiers, and province. One major structural challenge encountered during initial data understanding was the **transition from NOC 2016 (used in 2016 and 2020 datasets) to NOC 2021 (used in 2024).** This shift introduced mismatches due to merged, split, or reclassified occupations and required careful reconciliation.

Key Preprocessing Steps:

- **Filtering:** Removed rows with outdated reference periods (e.g., 2011) or missing wage values.
- **Standardization**: Harmonized province names, cleaned occupation titles, and filled missing national-level data.
- **Occupation Code Mapping:** Developed a refined 1-to-1 mapping from NOC 2016 to NOC 2021 using official

concordance tables [9], supplemented by manual review and **semantic similarity analysis** with a language model.

- **Merging:** Combined all datasets using NOC and ER_Code, aligning everything to the NOC 2021 structure.
- **Feature Engineering**: Extracted Broad Category, TEER level, and Major Group metadata, adding descriptive names.

The final **merged dataset** contains **11,448 complete records** with occupation, region, and wage information across all three years. The unified dataset is stored as a local warehouse table to support efficient downstream analysis.

2. Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) was performed at three levels — national, provincial, and regional — to explore wage distributions, disparities, and structural patterns across Canada between 2016 and 2024.

National-Level Analysis

At the national level, key analyses included:

- **Wage distribution over time**: Boxplots revealed rising median wages from 2016 to 2024. The highest-paid national occupation in 2024 earned over $184,000 annually.
- **TEER and wage correlation**: A clear negative correlation was observed between TEER level and median wage, confirmed by both Pearson and Spearman correlation coefficients.

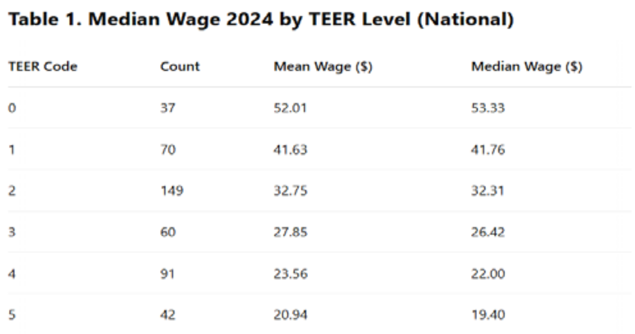


Table 1. Median Wage 2024 by TEER Level (National)

TEER Code	Count	Mean Wage ($)	Median Wage ($)
0	37	52.01	53.33
1	70	41.63	41.76
2	149	32.75	32.31
3	60	27.85	26.42
4	91	23.56	22.00
5	42	20.94	19.40

Figure 1. National Median Wage Distribution by TEER Level

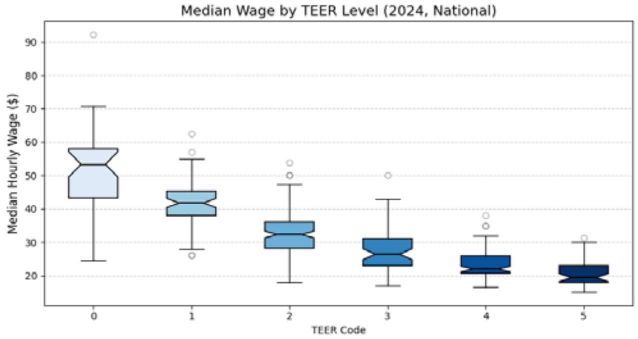


Figure 1. This boxplot displays the spread and median of 2024 national-level median hourly wages by TEER level. TEER 0 (Management) and TEER 1 (University degree) show the highest median wages, with a consistent downward trend toward TEER 5 (entry-level, low-skill occupations).

- **Broad occupational categories**: Some categories (e.g., management, sciences, healthcare) consistently reported higher wages. Interactive Altair heatmaps, KDE plots, and bubble charts helped visualize these patterns by TEER and occupational category over time.

Figure 2. Median Wage Heatmap by Broad Occupational Category and TEER Level (2016–2024, National)

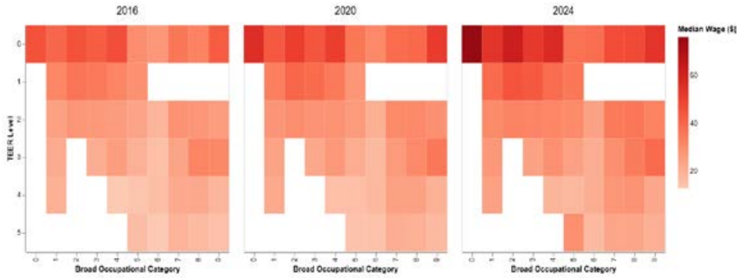


Figure2. This heatmap shows average national median wages by Broad Occupational Category and TEER level (2016–2024). It highlights consistent wage stratification—higher wages in TEER 0–1 roles (university-level) and lower wages in TEER 4–5 (high school or short-term training). A link to interactive visuals will be included in the final report.

Provincial-Level Analysis

- **Disparities** across provinces were analyzed using side-by-side visualizations (2016, 2020, 2024), showing wage growth and sustained gaps between top provinces (Alberta, Ontario, BC) and lower-wage provinces (e.g., PEI, Newfoundland and Labrador (NL)).
- **Top 3 vs Bottom 3** wage provinces were highlighted using faceted Altair heatmaps by occupation and TEER level.

Regional-Level Analysis

- Regional outliers were identified by grouping economic regions with sufficient data (≥10 records) and analyzing mean and distribution of 2024 wages.
- A **horizontal boxplot** showed the top and bottom 10 regions, with provinces overlaid using color-coded highlights for visibility.

Distributional Insights

- **Kernel density estimates (KDE) plots** showed shifts in wage density across categories over time.

Interactive Altair charts and static Matplotlib plots were used for all visualizations.

3. Pattern Detection and Clustering

To move beyond descriptive statistics, the project applied unsupervised learning techniques to uncover latent patterns in wage structures. These clustering methods helped identify groupings of occupations or regions that share similar wage characteristics, independent of pre-defined labels.

3.1 Occupational Clustering (NOC-Level Analysis)

K-Means clustering was applied to occupation-level data (NOC_2021) using four primary features: **median wages in 2016,**

2020, and 2024, and TEER level. The goal was to detect wage-related clusters across all occupations.

- With **k=4**, four distinct clusters were identified:
  - 1 – Entry-Level & Low Wage
  - 2 – Mid Wage, Mid Skill
  - 3 – High Wage, High Skill
  - 4 – Specialized / Outlier Roles
- With **k=7**, additional resolution revealed subgroups such as:
  - Very Low Wage, Minimal Skill
  - Low Wage, Mid Skill
  - Specialized High-Skill Outliers

These clusters were projected into 2D space using **Principal Component Analysis (PCA)** for visualization. Interactive **Altair scatter plots** allowed exploration of occupation-level clusters by wage level, category, and TEER.

3.2 Provincial Wage Profile Clustering

To evaluate whether the **same occupations have consistent wages across provinces**, K-Means clustering was performed on a pivoted dataset of **Median_Wage_2024 per NOC and Province**. Missing values were imputed using national medians.

- PCA revealed that **over 95% of variance was explained by a single principal component**, indicating that **occupational wage patterns are largely stable across provinces**.

Figure 3. Average Wage by Province per Cluster (k=4)

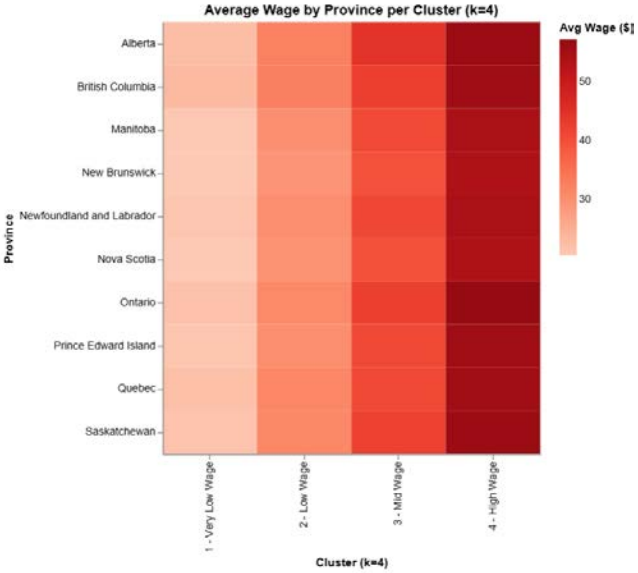


Figure 3. This heatmap shows average median wages by province across four occupation clusters from K-Means. Wages are mainly shaped by cluster membership (i.e., training and skill), with minor provincial differences. Alberta and Ontario often report slightly higher wages within the same cluster, reflecting broader economic patterns.

Clustering identified occupations with **similar wage consistency profiles**, ranging from:

- 1 – Very Low Wage NOCs
- to 4 – High Wage NOCs with minimal regional fluctuation

3.3 Hierarchical Clustering

A **hierarchical clustering** approach using Ward's method was applied to the same provincial wage matrix. At a distance threshold of t = 10, **7 occupation clusters** were identified. These were grouped and labeled by average wage and regional spread.

This method provided additional validation of K-Means clusters and helped detect **distinct regional wage patterns**, such as clusters concentrated in provinces like Alberta or Ontario.

3.4 Cluster Interpretation

- Clusters were evaluated by **TEER composition**, **average wages**, and **intra-cluster wage dispersion**, validating the economic meaning of groupings.
- Additional heatmaps were created to show:
  - Average wage per **province per cluster**
  - **Share of occupations per cluster** within each province

Clustering confirmed that wages are strongly shaped by TEER level and occupational category, with consistent patterns across provinces.

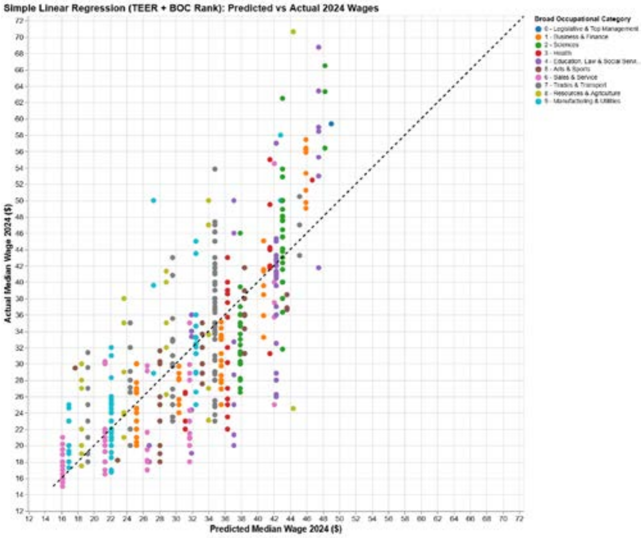
4. Machine Learning Approach

To complement descriptive and clustering analysis, supervised learning models were developed to predict wage levels based on occupation and structural features.

4.1 Wage Prediction using TEER and Broad Category

- A **Simple Linear Regression** model using **TEER Code** and **Broad Category Rank** (CustomIndex = TEER × 10 + BOC Rank) was trained.
- It achieved a substantial fit with $R^2 \approx 0.59$, confirming the strong predictive power of these two structural factors alone.

Figure 4. Actual vs Predicted Wages using Custom Index (TEER × BOC Rank)



4.2 Extended Modeling: Testing Additional Predictive Features

To investigate whether additional information could significantly improve wage prediction, several models were trained using progressively richer feature sets, including **Broad Category, Province, and embedded NOC Titles**.

While the inclusion of more features slightly increased predictive accuracy, **the gains were modest and often came at the cost of generalization**. Simpler models based primarily on TEER level maintained better cross-validated performance, while more complex models tended to overfit.

4.3 Wage Stability Prediction

Linear Regression, Random Forest, and **Gradient Boosting** models were trained to predict 2024 wages using **only historical wage data from 2016 and 2020**, without relying on occupational or regional features.

This tested whether wage growth patterns were stable enough for accurate forecasting based purely on past wages.

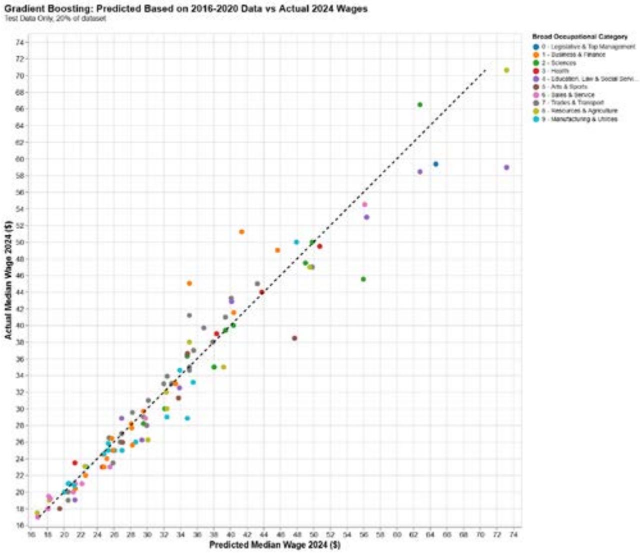
All three models achieved **strong predictive accuracy**, with R² scores consistently above 0.90 and low RMSE values.

These results confirm that historical wage trends alone provide a highly reliable basis for predicting future wages across occupations.

Importantly, **despite global events such as the COVID-19 pandemic, economic disruptions, and advances in AI, no major breaks in wage growth patterns were detected** in Canada up to 2024.

Wage evolution during this period remained stable, predictable, and resilient to external shocks.

Figure 5. Gradient Boosting: Predicted vs Actual Wages Based on 2016–2020 Data



To ensure the reliability of the results, all models were evaluated using an **80–20 train-test split** as well as **5-fold cross-validation**.

5. Evaluation

5.1 Evaluation Methods and Metrics

The project employed a multi-tiered evaluation strategy, combining exploratory validation, unsupervised learning assessments, and supervised model evaluation.

EDA-Based Validation

Exploratory validation was conducted at the national level using correlation analysis between key features and wages. TEER level showed a strong negative correlation with median wage (Pearson: -0.75, Spearman: -0.77), while Broad Occupational Category (BOC) showed a moderate correlation (Pearson: -0.34, Spearman: -0.31). These early findings confirmed that training and education levels (proxied by TEER and BOC) are major determinants of wage variation in Canada.

Clustering Evaluation

Clustering effectiveness was assessed via:

- **Visual separability** of clusters through PCA projection
- **Internal consistenc**y by comparing average wage and TEER level within each cluster
- **Descriptive interpretability**

A key finding from cluster centroid evaluation was that clusters containing highly paid occupations showed significantly higher internal dispersion. This reflects the diverse nature of specialized or senior management roles. Table 1 (from 2-1-clustering_summary_k=n) illustrates the average, standard deviation, and maximum distance to centroid for each occupational cluster (k=4).

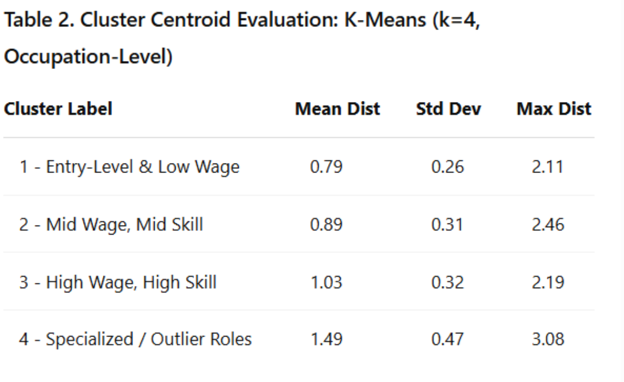


Table 2. Cluster Centroid Evaluation: K-Means (k=4, Occupation-Level)

Cluster Label	Mean Dist	Std Dev	Max Dist
1 - Entry-Level & Low Wage	0.79	0.26	2.11
2 - Mid Wage, Mid Skill	0.89	0.31	2.46
3 - High Wage, High Skill	1.03	0.32	2.19
4 - Specialized / Outlier Roles	1.49	0.47	3.08

This table confirms that specialized occupations form the most dispersed cluster, aligning with their diverse titles and salary scales.

Supervised Model Evaluation

Model performance was assessed using:

- **R²** (coefficient of determination) for explained variance
- **RMSE** (Root Mean Squared Error) for average prediction error
- **5-fold Cross-Validation** for estimating generalization reliability

All models used an 80/20 train-test split and were evaluated on the test set and across 5 folds.

Table 3 summarizes the cross-validated R² scores for all tested models.

Table 3. Cross-Validated R² Scores (Mean and Std)

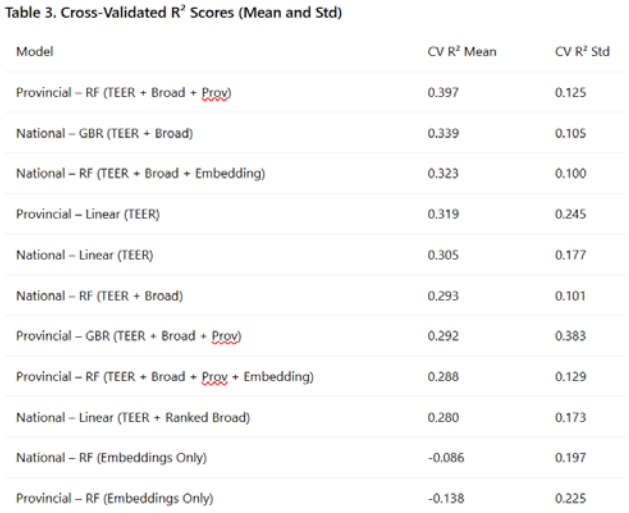


Table 3. Cross-Validated R² Scores (Mean and Std)

Model	CV R² Mean	CV R² Std
Provincial – RF (TEER + Broad + Prov)	0.397	0.125
National – GBR (TEER + Broad)	0.339	0.105
National – RF (TEER + Broad + Embedding)	0.323	0.100
Provincial – Linear (TEER)	0.319	0.245
National – Linear (TEER)	0.305	0.177
National – RF (TEER + Broad)	0.293	0.101
Provincial – GBR (TEER + Broad + Prov)	0.292	0.383
Provincial – RF (TEER + Broad + Prov + Embedding)	0.288	0.129
National – Linear (TEER + Ranked Broad)	0.280	0.173
National – RF (Embeddings Only)	-0.086	0.197
Provincial – RF (Embeddings Only)	-0.138	0.225

These results confirm that TEER and BOC are the strongest predictors of wage, while embeddings introduce overfitting and instability. Simpler models often generalize better.

5.2 Effectiveness and Efficiency of the Approach

Effectiveness:

- EDA and clustering techniques proved highly effective in detecting wage structure and groupings.
- TEER and BOC consistently emerged as dominant explanatory variables.
- Predictive modeling confirmed these patterns and provided insights into generalization capacity.

Efficiency:

- All analysis was conducted efficiently on a standard machine using Python (pandas, scikit-learn, Matplotlib, Altair).
- Data pipelines were modular and reusable, allowing rapid experimentation and evaluation.

5.3 Experimental Setup

The modeling phase used a structured and reproducible experimental setup:

- **80/20 train-test split**
- **5-fold cross-validation** for generalization assessment
- **Standard metrics (R², RMSE)** used across all models
- **Baseline models** (e.g., TEER-only) compared against complex feature sets

This setup ensured both comparability and reliability across model types.

DISCUSSION

Project Timeline

The project has been successfully completed on schedule, with all core components delivered by the April 28, 2025 deadline. Exploratory analysis, clustering, supervised learning, evaluation, and documentation were fully implemented. Final deliverables include this report, supporting technical files, and an accompanying presentation

Current Status

At this stage, the project has completed the following milestones:

- **Data collection:** Wage reports from 2016, 2020, and 2024 were retrieved from the Government of Canada's open data portal.
- **Data Understanding and Preprocessing**: Comprehensive cleaning and merging of datasets, normalization of occupational codes, and handling of missing values were conducted.
- **NOC Mapping:** A custom mapping between NOC 2016 and NOC 2021 was created using both official mappings and refined matches to ensure consistency.
- **Exploratory Data Analysis (EDA):** National, provincial, and regional wage trends were explored through summary statistics, boxplots, KDE plots, and Altair-based heatmaps.
- **Clustering:** K-Means and Hierarchical Clustering revealed structural groupings based on TEER levels, occupations, and wage profiles.
- **Supervised Learning Pipeline:** A comprehensive modeling pipeline was developed, including linear models, Random Forests, and Gradient Boosting for wage prediction at national and provincial levels.
- **Evaluation:** Extensive evaluation was performed using R² scores, RMSE, and cross-validation to assess model generalization.
- **Communication**: A dedicated GitHub repository was created to organize and share the project's data, code, models, and documentation, supporting transparency and stakeholder communication.

Key Challenges and Solutions

Two main challenges were encountered and successfully addressed during the project:

- **NOC 2016 → NOC 2021 Transition:** The unification of two classification systems involved resolving complex code splits, merges, and inconsistencies. This was mitigated by integrating both official mappings and refined matches (including LLM-assisted review), leading to a robust and accurate final mapping.
- **Handling Missing and Incomplete Data:** Wage datasets varied in structure and completeness. NaNs in key columns were either removed or imputed using national-level averages, preserving data integrity for modeling and EDA.

Path Forward

With the core project objectives fully implemented, several meaningful avenues for future work have been identified:

- **Continuity of Analysis:**

  Expand the WageCAN framework into a long-term analytical platform by applying the developed methodologies to future Canadian wage datasets (e.g., 2026, 2028). This would help monitor evolving labor market trends and detect structural changes over time.

- **Stakeholder Communication:**

  Share the project's findings and technical considerations with Statistics Canada and related organizations. In particular, advocate for enhanced synchronization between datasets collected across different years, which would improve the reliability of longitudinal wage analysis and predictive modeling.

- **Advanced Modeling Exploration:**

  Explore deep learning approaches using frameworks such as TensorFlow and PyTorch. Comparing neural network-based regressors to traditional ensemble methods (Random Forest, Gradient Boosting) could provide insights into potential performance gains for wage prediction tasks.

  These steps would not only strengthen the project's long-term relevance but also contribute to the broader discussion of wage analytics in Canada

CONCLUSION

Final Analytical Summary

The analysis confirms that **TEER** (Training, Education, Experience, and Responsibilities) remains by far the strongest predictor of median wages in Canada. Wage levels are primarily shaped by the level of education and training associated with each occupation — a relationship that holds consistently across all provinces and industries.

While wages also vary across **Broad Occupational Categories (BOCs)**, this variation is notably smaller. The only major exception is **BOC 0** (Legislative and Senior Management), which belongs to the highest TEER level and consistently exhibits higher wages.

Provincial differences in wages are relatively minor, with only a few notable outliers. While some regional variation exists, it is not a dominant factor in wage prediction.

Machine learning experiments **revealed moderate overfitting tendencies**, especially when combining TEER, BOC, and Province features. More complex models such as Random Forest and Gradient Boosting did not substantially outperform simple linear regression based solely on TEER.

Using embedded **NOC titles** (via sentence-transformers) did not add predictive value. In fact, models trained exclusively on semantic embeddings often performed worse than a naive mean predictor, confirming that similar-sounding job titles are not reliable indicators of wage levels.

However, modeling wage stability between 2016, 2020, and 2024 showed outstanding predictability: previous wages proved to be excellent predictors of future wages, even though major global disruptions such as the COVID-19 pandemic.

This strongly suggests that occupational wage structures in Canada have remained highly stable over time.

Main Insights

- **Education and training (TEER) are the primary determinants** of wage levels.
- **Wage patterns remained remarkably stable** from 2016 to 2024, despite global economic shocks.
- **Industrial category and regional factors matter**, but significantly less than training level.
- **Semantic similarity between job titles is not predictive** of wages.

Next Steps

- **Extend longitudinal modeling** by incorporating historical data (2016 and earlier) to track wage dynamics over longer periods and identify evolving occupational trends.
- **Explore advanced modeling techniques** (e.g., TensorFlow, PyTorch) to benchmark performance against traditional machine learning models.

Suggestions for Statistics Canada

- **Incorporate NOC 2021 codes into historical datasets** to ensure continuity and improve the accuracy of longitudinal labor market research.

Notes on Model Interpretations

- **"National – GBR (TEER + Broad)"** achieved the best balance between performance and generalization at the national level.
- **Embedding-only models** exhibited negative cross-validated R^2 scores — often performing worse than naive predictions.
- **Simple TEER-based models** performed well, reinforcing the central role of education and training in wage determination.
- **Adding many features (e.g., embeddings, provinces)** introduced overfitting without meaningful improvement, even with regularization techniques**.**

POTENTIAL BIASES

Several sources of potential bias were identified and considered during the WageCAN analysis:

- **Data Structure Bias:** Wage datasets from 2016, 2020, and 2024 differed in structure and completeness. Efforts were made to harmonize them, but subtle inconsistencies may remain due to mapping between NOC 2016 and NOC2021.
- **Mapping Uncertainty:** The manual and semi-automated mapping between NOC versions may introduce classification

errors, particularly for occupations that were merged, split, or redefined.

- **Selection Bias:** Territories and economic regions with insufficient data were excluded from clustering and modeling to maintain quality. This exclusion may limit generalizability to less populous regions.
- **Model Bias and Overfitting:** More complex models showed tendencies toward overfitting, especially when many features (e.g., embedded NOC titles) were included. Careful cross-validation was used to mitigate this risk, but slight optimism bias may still affect results.
- **Unmeasured External Factors:** While historical wage patterns proved highly predictive, external influences such as labor market policy changes, technological disruptions (e.g., AI adoption), or fundamental shifts in the international state of affairs (e.g., implementation of trade tariffs) may impact future wages beyond the historical trend captured here.

Efforts were made throughout the project to detect, mitigate, and transparently communicate these limitations.

DISCLOSURE

This project was primarily designed, developed, and executed by the author.

The author is proficient in Python and key data science libraries (pandas, scikit-learn, Altair, Matplotlib) and conducted all technical work independently.

Large Language Models (LLMs), specifically ChatGPT, were used selectively to assist with:

- Brainstorming and refining ideas
- Improving the clarity and style of documentation
- Assisting with bug checking and code optimization for efficiency

All core analytical work — including data preprocessing, exploratory analysis, clustering, modeling, and evaluation — reflects the author's original design and decision-making.

LLM assistance was strictly supportive and did not replace the author's work in developing the project's methodology, findings, and conclusions.

Project Repository

GitHub Repository for WageCAN Project:

https://github.com/nickshlepov/WageCAN_Project

REFERENCES

The following sources were consulted during the course of this project to ensure methodological rigor, accuracy, and consistency.

[1] https://www.bankofcanada.ca/2024/10/staff-analytical-note-2024-23/

[2] https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5217

[3] https://www.jobbank.gc.ca/trend-analysis/search-wages/wage-methodology

[4] https://www.iwh.on.ca/scientific-reports/machine-learning-and-labour-market-portrait-of-occupational-and-worker-inequities-in-canada

[5] https://www150.statcan.gc.ca/n1/pub/11f0019m/11f0019m2024005-eng.htm

[6] https://www.researchgate.net/publication/386042325_Machine_learning_and_the_labour_market_A_portrait_of_occupational_and_worker_inequities_in_Canada

[7] https://www.hiringlab.org/en-ca/2023/06/27/canadian-indeed-wage-tracker/

[8] https://github.com/nickshlepov/WageCAN_Project

[9] https://open.canada.ca/data/en/dataset/adad580f-76b0-4502-bd05- 20c125de9116/resource/d16e10ea-77bf-4db8-bdb5-adc709e6cada

[10] https://noc.esdc.gc.ca/Versions/NOCConcordanceTables