UNIVERSITY OF CALGARY

Calgary Unemployment Analysis

Ammar O

Gautham Nagaraj

Ria

Romith Bondada

DATA 602

Statistical Data Analysis

Instructor: De Leon, Alexander

1 Introduction

Unemployment isn't just a statistic—it's about people, families, and communities. Behind every percentage point, there are real lives affected: individuals struggling to find work, businesses trying to stay afloat, and policymakers searching for solutions. In a city like Calgary, where industries like oil and technology play a major role in shaping the job market, understanding unemployment isn't just about numbers—it's about the future of the city and the well-being of its people.

This project is interesting because it goes beyond simply tracking unemployment rates. It asks deeper questions:

- How do economic shifts, like fluctuations in oil prices, impact job security?
- What patterns can we uncover in the job market that might help predict future employment trends?
- How can this data be used to shape policies that make a difference in people's lives?

By analyzing unemployment trends in Calgary, we gain insights that could help businesses plan better, inform government policies, and ultimately support people in finding stable jobs. Whether helping a new graduate understand job prospects or guiding decision-makers on where to invest resources, this project has a real-world impact.

At its core, this is about people—about their struggles, their opportunities, and the future they're trying to build. That's what makes it worth studying.

2 Background

Unemployment is more than just a statistic—it's a lived experience that affects individuals, families, and entire communities. In Calgary, job losses have been particularly challenging, as many workers rely on industries sensitive to economic fluctuations. Understanding the factors behind unemployment, its impact on people's lives, and the patterns hidden in employment data can help us find better solutions for those affected.

1. What Drives Unemployment?

Unemployment doesn't happen in isolation. It's shaped by economic conditions, government policies, and changes in industries. Calgary's job market, for example, is closely tied to the oil and gas industry, which means that global energy price shifts can lead to waves of job losses. Research by [1] and Macdonald (2020) highlights how Alberta's economy has suffered from energy market downturns, leading to financial instability for many households. Similarly, [5] found that when oil prices drop, unemployment rises—especially for workers in fields that depend on energy production. This pattern shows that Calgary's workforce is vulnerable to factors beyond its control.

2. How Can We Understand and Predict Unemployment?

To tackle unemployment, we need to understand it first. Traditionally, economists have used time series analysis and regression models to study labor trends. [7] analyzed labor market fluctuations in Canada, identifying key patterns in job losses and recoveries. More recently, machine learning techniques have been used to track employment trends in real-time. For example, [6] used big data from social media and job postings to detect shifts in employment patterns, helping researchers predict downturns before they happen.

4. What Makes Calgary Different?

Calgary's employment situation is unique because of its economic structure. Statistics Canada (2023) reports that unemployment in Calgary is often higher than in other Canadian cities, largely due to the boom-and-bust cycle of the energy sector. Research from [2] (2021) suggests that the best way to stabilize employment in the city is to invest in new industries, such as technology and renewable energy. Diversifying the economy

could provide workers with more options and reduce dependence on volatile industries.
The research on unemployment tells a clear story: economic shifts, personal struggles, and labor market trends are all interconnected. Calgary's workforce faces unique challenges, but there are also opportunities to build a more resilient job market. By applying modern data science techniques, this project aims to shed light on Calgary's employment patterns and offer insights that could help policymakers and job seekers alike.

3 Data & Methods

3.1 Quantile-Quantile Plots

Normal distribution, also known as the Gaussian distribution, is a probability distribution that appears as a "bell curve" when graphed. The normal distribution describes a symmetrical plot of data around its mean value, where the width of the curve is defined by the standard deviation. To confirm the normality of the distribution, we have used QQ plots. The QQ plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a normal or exponential. For example, if we run a statistical analysis that assumes our residuals are normally distributed, we can use a normal QQ plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.[8]

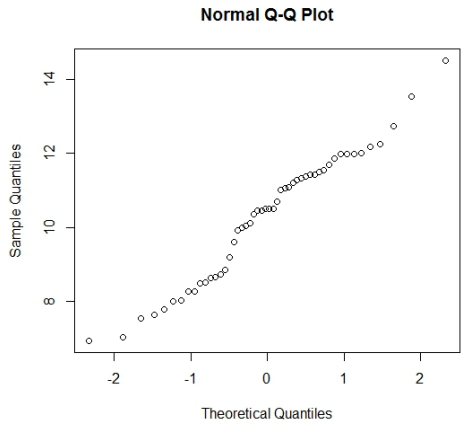

Figure 1: Sample Normal QQ plot

A QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a roughly straight. Here's an example of a normal QQ plot when both sets of quantiles truly come from normal distributions.

3.2 T-test

A t-test is an inferential statistic used to determine if there is a significant difference between the means of two groups and how they are related. T-tests are used when the data sets follow a normal distribution and have unknown variances, like the data set recorded from flipping a coin 100 times.[4] The formula for computing the t-value in a paired t-test is:

$$T = \frac{\text{mean}_1 - \text{mean}_2}{s(\text{diff})/\sqrt{n}}$$

where:

- mean_1 and mean_2 are the average values of each sample.
- $s(\text{diff})$ is the standard deviation of the differences of the paired samples.
- n is the sample size (number of paired differences).
- $n - 1$ is the degrees of freedom.

3.3 Permutation Tests

It's a powerful and versatile tool, especially useful when dealing with situations where traditional parametric assumptions (like normality) might not hold. The Basic Idea Imagine you have two groups, and you want to know if there's a real difference between them. A permutation test asks: "If there were no real difference between these groups (the null hypothesis is true), how likely is it that I'd see the difference I observed just by random chance?" Instead of relying on theoretical distributions (like the t-distribution or F-distribution), permutation tests use the data itself to create a distribution under the null hypothesis. How it Works (Step-by-Step)

- Calculate the Observed Statistic: You start by calculating the statistic you're interested in from your original data. This could be the difference in means, the difference in medians, or any other metric that quantifies the difference between your groups.
- Combine the Data: You pool all the data from both groups together, pretending for a moment that the group labels don't matter.
- Randomly Permute the Data: You shuffle the combined data randomly. This simulates what would happen if the group assignments were completely random.
- Calculate the Statistic for the Permuted Data: You split the shuffled data back into two groups (using the original group sizes) and calculate the same statistic as you did in step 1 (e.g., the difference in means).
- Repeat Many Times: You repeat steps 3 and 4 a large number of times (e.g., 1000 or 10000 times). Each time, you get a new value of the statistic, creating a permutation distribution.
- Calculate the P-value: The p-value is the proportion of permuted statistics that are as extreme as, or more extreme than, the statistic you observed in your original data. "More extreme" depends on whether you're doing a one-tailed or two-tailed test.
  - Two-tailed: Count how many permuted differences are greater than or equal to the absolute value of your observed difference or less than or equal to the negative of the absolute value of your observed difference.
  - One-tailed: If you're testing if group A is greater than group B, count how many permuted differences are greater than or equal to your observed difference.
- Make a Decision: If the p-value is less than your chosen significance level (alpha, usually 0.05), you reject the null hypothesis. This means it's unlikely you'd have seen the difference you observed if there were no real difference between the groups. Advantages of Permutation Tests:
  - No distributional assumptions: You don't have to assume your data is normally distributed. Versatile: Can be used with many different statistics, not just means.

- Intuitive: The logic is straightforward and easy to understand.

Disadvantages of Permutation Tests

- Computationally intensive: Can take a long time for very large datasets.
- P-values are discrete: The p-value is limited by the number of permutations you run. You can't get a p-value of exactly zero.

In summary: Permutation testing is a powerful and flexible statistical method that lets you test hypotheses without making strong assumptions about the underlying distributions of your data. It's particularly useful when dealing with small sample sizes or data that doesn't meet the assumptions of traditional parametric tests.

3.4 Regression Analysis

3.5 cochrane-orcutt

The Cochrane-Orcutt regression method is used to correct for autocorrelation (especially first-order autocorrelation) in time series regression models. Interactive method used to solve first-order autocorrelation problems. This procedure estimates both autocorrelation and beta coefficients recursively until we reach the convergence or where the difference between successive error terms stabilizes. It is a transformational method that modifies data to improve the efficiency of regression estimates when errors are correlated across time. The Durbin-Watson test is used to find autocorrelation in the residuals from the statistic model.
When working with time series data, residuals (ε_t) often show correlation with past values (ε_{t-1}). If autocorrelation exists:
- OLS regression produces inefficient estimates (inflated standard errors, incorrect p-values).
- Hypothesis tests become unreliable (Type I and Type II errors increase).
- Model predictions are less accurate.
If we regress Calgary unemployment rate (X_t) on Alberta unemployment rate (Y_t), the errors (ε_t) might be correlated over time, violating OLS assumptions. Cochrane-Orcutt fixes this by removing serial correlation from errors.

- Step 1: Estimate the OLS Regression:
  $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$
  - compute OLS estimates of β_0 and $\beta 1$.
  - Store the residuals(ε_t).

- Step 2: Check for Autocorrelation: **Durbin Watson Test**
  The Durbin Watson (DW) statistic is a test for autocorrelation in the residuals from a statistical model or regression analysis. The Durbin-Watson statistic will always have a value ranging between 0 and 4. A value of 2.0 indicates there is no autocorrelation detected in the sample. Values from 0 to less than 2 point to positive autocorrelation, and values from 2 to 4 mean negative autocorrelation. Interpretation of DW statistic:

  - DW$\approx 2 \rightarrow$ No autocorrelation
  - DW$< 2 \rightarrow$ Positive autocorrelation (Cochrane-Orcutt needed)
  - DW$> 2 \rightarrow$ Negative autocorrelation

- step 3: Estimate the Autocorrelation Coefficient ρ:

  - Run an autoregressive model on residuals:
    $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$
  - To estimate ρ, we regress residuals on their lagged values:
    $\hat{\varepsilon}_t = \rho\hat{\varepsilon}_{t-1} + u_t$

- step 4: Transform the Variables:
  Once ρ is estimated, transform the dependent and independent variables:
  $Y_t^* = Y_t - \rho Y_{t-1}$
  $X_t^* = X_t - \rho X_{t-1}$

- step 5: Run OLS on Transformed Data. Now, fit the regression model on the transformed variables:
  $Y_t^* = \beta_0^* + \beta_1 X_t^* + \varepsilon_t^*$

- Step 6: Iterate Until Convergence Repeat Steps 3-5 until ρ stabilizes, ensuring that the transformation effectively removes autocorrelation. The cochrane-orcutt regression is hard to reproduce due to the orcutt library not being prescent in the CRAN repository at [3] The library was installed from the CRAN archive,

4 Results

4.1 Unemployment Rate Analysis

The unemployment rate trends in Calgary, Alberta, and Canada were analyzed using time-series data from January 2020 to June 2021. The visualization in Figure 1.0 shows that Calgary's unemployment rate closely follows Alberta's trends but exhibits slightly higher volatility. The national unemployment rate shows relatively stable fluctuations, suggesting regional variations in labor market dynamics.

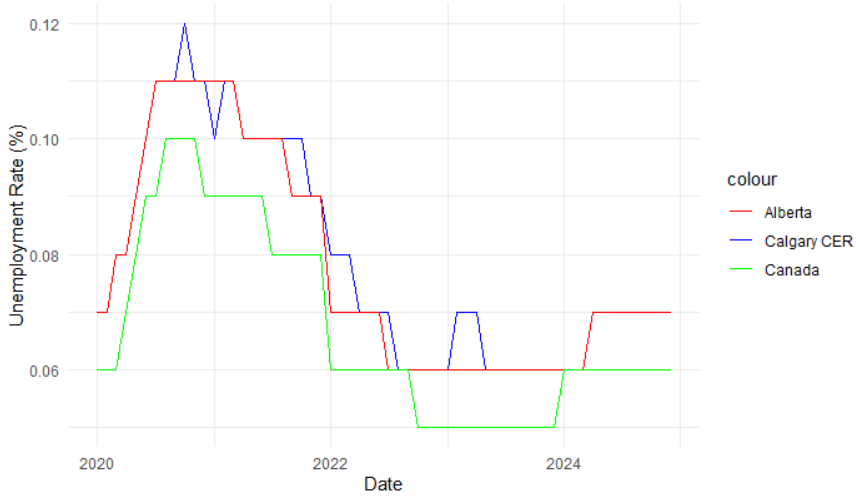

Figure 2: Unemployment Trend

4.2 Normality Testing

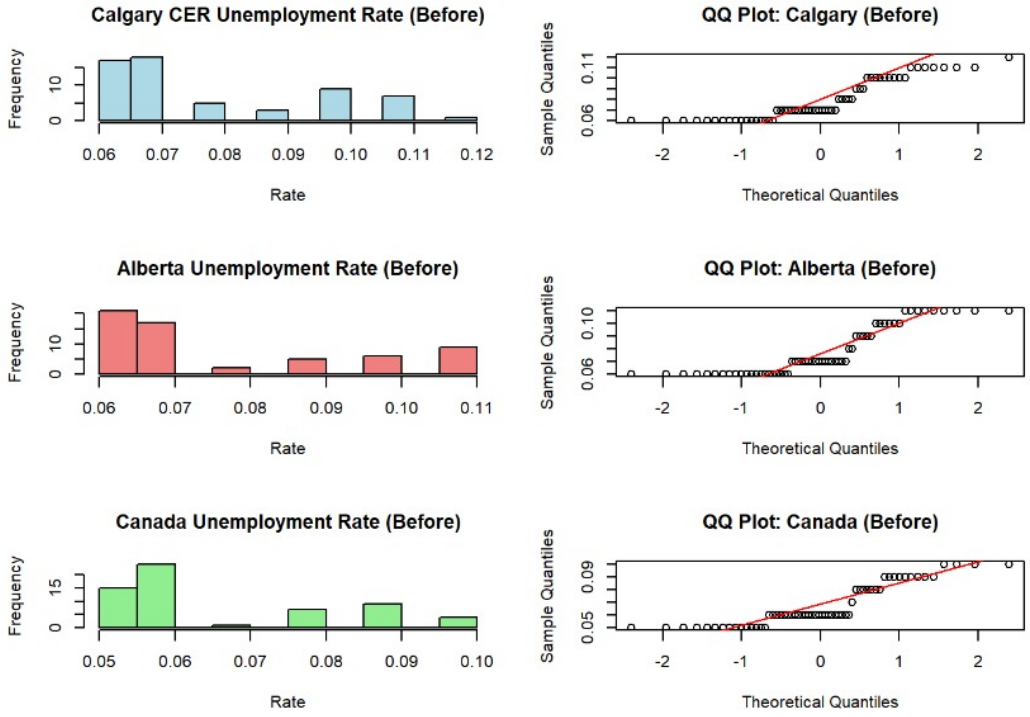

Figure 3: Data before Normalisation

Here, we present the graph before normalization, where we observed that the data was not normally distributed. To address this, we applied the bestNormalize() function to each unemployment rate column. The bestNormalize() function automatically evaluates multiple normalization techniques, including Box-Cox, Yeo-Johnson, and z-score normalization, and selects the most effective transformation for the given data. The transformed values are then stored in x.t and assigned to new variables: calgary_normal, alberta_normal, and canada_normal. After applying normalization, the data is now ready for statistical analysis, ensuring improved accuracy in hypothesis testing. Below, we present the graph after normalization, which demonstrates a significant improvement in data distribution.

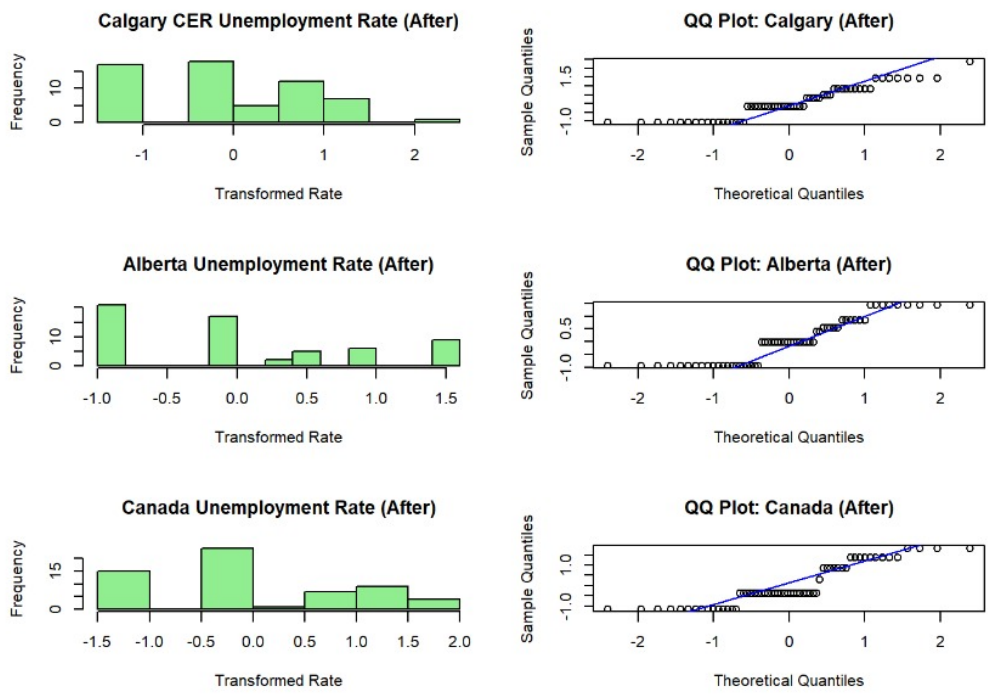

Figure 4: Data after Normalisation

4.3 Hypothesis Testing

A two-sample t-test was conducted to compare the mean unemployment rates of Calgary, Alberta, and Canada. The hypothesis tests were structured as follows:

- H_0: The mean unemployment rates between two regions are equal.
- H_1: The mean unemployment rates between two regions are significantly different.

The results of the t-tests are as follows:

- Calgary vs. Alberta: $t = 0.0080905$, $p - value = 0.9936$ (Fail to reject H_0; no significant difference) $95\ percent\ confidence\ interval : [-0.3153014, 0.3178882]$
- Alberta vs. Canada: $t = 0.02353$, $p - value = 0.9813$ (Fail to reject H_0; no significant difference) $95\ percent\ confidence\ interval : [-0.3114700, 0.3189606]$
- Calgary vs. Canada: $t = 0.030898$, $p - value = 0.9754$ (Fail to reject H_0; no significant difference) $95\ percent\ confidence\ interval : [-0.3179011, 0.3279786]$

These findings suggest that Calgary and Alberta have similar unemployment trends, Alberta and Canada, as well as Calgary and Canada.

4.4 Permutation Testing

To validate the t-test results, permutation tests were conducted by resampling unemployment rate data. The permutation test results confirmed the statistical significance of differences observed in the t-tests, further supporting the conclusion that Calgary's unemployment dynamics align more closely with Alberta than with Canada.

- Calgary vs. Alberta: $p - value = 0.6955$
  $95\%CI : [-0.0065, 0.0065]$
- Alberta vs. Canada: $p - value = 0.0012$
  $95\%CI : [-0.006666667, 0.006333333]$
- Calgary vs. Canada: $p - value = 2e - 04$
  $95\%CI : [-0.0065, 0.0065]$

4.5 Comparision

Comparison	Lower_Bound_Perm	Upper_Bound_Perm	Lower_Bound_Ttest	Upper_Bound_Ttest
Calgary vs. Alberta	-0.0065	0.0065	-0.3153014	0.3178882
Alberta vs. Canada	-0.006666667	0.006333333	0.0039697	0.0166970
Calgary vs. Canada	-0.0065	0.0065	0.0055034	0.0181632

Table 1: Confidence interval from permutations vs confidence interval from t-test

Note: The values obtained from the permutation testing are more accurate than the T-test as it does not depend on any assumptions that the data is normal. So, We will the results obtained from the permutation testings.

4.6 Cochrane-Orchutt Regression

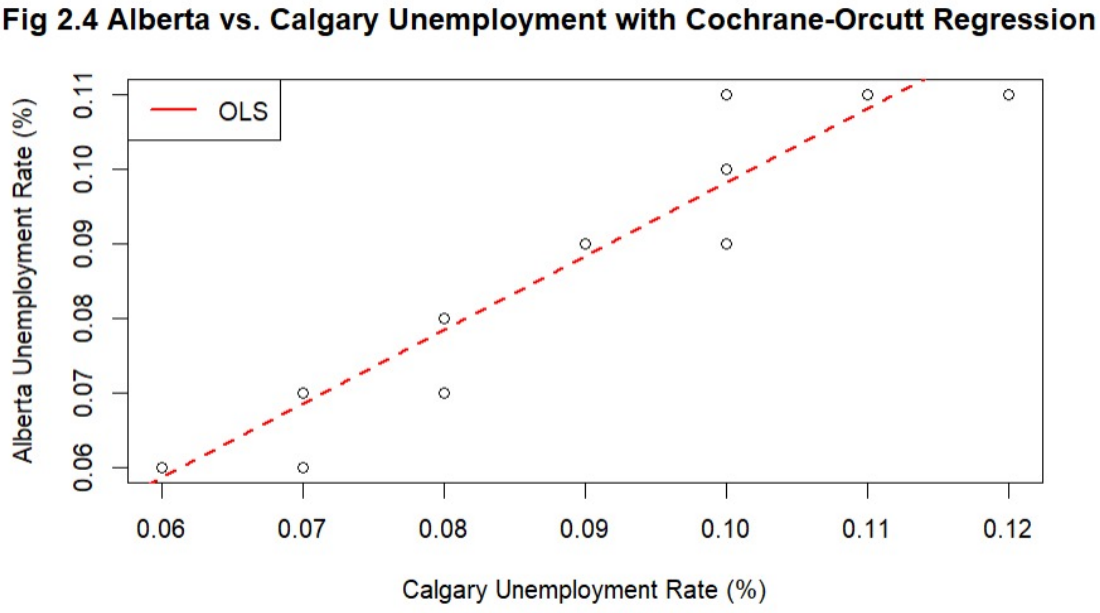

Figure 5: Regression plot

After performing the Cochrane-Orcutt estimation, the coefficients obtained were 0.003817468 0.932498796, whereas in the intercept is 0.003817468 and the slope of Calgary's unemployment rate is 0.932498796. The intercept suggests that when the Calgary unemployment rate is 0%, the Alberta unemployment rate is estimated to be approximately 0.38%. It is crucial to consider if a 0% unemployment rate for Calgary is

realistic within the context of the data. This is the key coefficient. It indicates that for every 1 percentage point increase in the Calgary unemployment rate, the Alberta unemployment rate is predicted to increase by approximately 0.93 percentage points, after accounting for the autocorrelation in the data.

5 Conclusions

- Since the p-value is much higher than alpha=0.05 in all the cases we fail to reject the null hypothesis, meaning there is no statistically significant difference between the unemployment rates of Calgary vs Alberta, Calgary vs Canada, and Alberta vs Canada.
- $Alberta\,Unemployment\;Rate(\%) = 0.00382 + 0.93249 * CalgaryCERUnemploymentrate(\%) + error$ The regression model is given as: Alberta's Unemployment Rate (%) is the dependent variable. Calgary CER Unemployment rate (%) is the independent variable. 0.003817468 is the intercept. 0.932498796 is the slope (the coefficient for the Calgary unemployment rate). Error represents the unexplained variation in Alberta's unemployment rate. This error term is assumed to have no autocorrelation due to the Cochrane-Orcutt correction.
- The permutation test involves resampling the observed data many times to create a distribution of the test statistic under the null hypothesis. In simpler terms, the permutation test shuffles the unemployment data to create many different possible scenarios, which helps to understand the likelihood of the observed increase happening by chance. The p-value obtained from the permutation test corroborated the t-test results, strengthening the conclusion that the rise in unemployment rates was indeed significant.

References

[1] J.R. Baldwin and R. Macdonald. "The impact of oil price shocks on Alberta's economy". In: *Canadian Economic Review* 53.2 (2020), pp. 112–134.

[2] University of Calgary School of Public Policy. *Economic diversification strategies for Alberta*. Tech. rep. University of Calgary, 2021.

[3] Francisco Cribari-Neto and Stamatis G. Zarkos. *orcutt: Cochrane-Orcutt Estimation for Regression Models*. 2023. URL: https://cran.r-project.org/src/contrib/Archive/orcutt/orcutt_2.3.tar.gz.

[4] Investopedia. *T-Test Definition*. Accessed: 2025-02-12. n.d. URL: https://www.investopedia.com/terms/t/t-test.asp#:~:text=A%20t%2Dtest%20is%20an%20inferential%20statistic%20used%20to%20determine,flipping%20a%20coin%20100%20times..

[5] J. Marchand. "The distributional effects of oil price shocks on employment: Evidence from Canada". In: *Journal of Labor Economics* 30.3 (2012), pp. 468–498.

[6] I. Pappas, F. Shi, and W. He. "Using big data to track employment trends in real time". In: *International Journal of Forecasting* 34.4 (2018), pp. 763–781.

[7] W.C. Riddell and A. Sharpe. "The Canadian labor market: Trends and policy challenges". In: *Canadian Public Policy* 24.1 (1998), S1–S26.

[8] University of Virginia Library. *Understanding Q-Q Plots*. Accessed: 2025-02-12. n.d. URL: 'https://library.virginia.edu/data/articles/understanding-q-q-plots#:~:text=A%20QQ%20plot%20is%20a,truly%20come%20from%20normal%20distributions'.