The Influence of Industry and Location on Data Analyst Salaries in Canada

DAMO 500 – PRINCIPLES OF ANALYTICS

INSTRUCTOR – Prof Mehdi (Matt) Mostofi

TEAM MEMBERS:

1. **Olajubu Emmanuel**
2. **Bushra Bushra**
3. **Vipul Khainar**
4. **Prawinder**

CHAPTER ONE

INTRODUCTION

1.0 Introduction

In an increasingly data-centric world, the significance of data analysts has become crucial. Organizations spanning various sectors depend on these professionals to convert raw data into valuable insights that influence strategic decision-making processes. In Canada, there is a rapid increase in the demand for data analysts, while salaries for such positions exhibit notable variations contingent upon industry and geographical location. Professionals in sectors such as finance and technology frequently receive higher salaries, whereas individuals employed in industries with less emphasis on data may encounter more modest compensation. Likewise, remuneration in areas such as Alberta may surpass that in provinces such as Nova Scotia or New Brunswick, indicative of the influence exerted by geographical determinants.

This project delves into the salary discrepancies to gain a deeper understanding of how industry and geographical location impact the remuneration of data analysts in Canada. Through an in-depth exploration of these discrepancies, our objective is to reveal patterns that can be advantageous to both individuals seeking employment and organizations looking to hire. For individuals seeking employment, this study offers a more defined understanding of areas to direct their professional ambitions towards in order to enhance remuneration and advancement prospects. For employers, it provides valuable insights for structuring competitive and equitable salary packages aimed at attracting high-caliber professionals. This study utilizes real-world data from Kaggle, capturing job market trends for data analysts across Canada.

1.2 Objectives

The project is centered around three primary objectives:

1. Examine Salary Trends across Industries: Gain insight into how sectors such as finance, healthcare, and technology influence the salary bands for data analysts.
2. Analyze Regional Disparities: Identify the provinces and cities that present the most competitive salaries for said positions.
3. Deliver Practical Insights: Offer actionable advice to job seekers to enhance their career opportunities and assist employers in remaining competitive in talent acquisition.

1.3 Scope

This project focuses exclusively on data analyst positions in Canada, examining salary variations across industries and geographical regions. The analysis is limited to understanding the Canadian job market and does not extend to international comparisons or roles outside data analytics. By concentrating on the nuances of the Canadian labour market, this study ensures its findings are relevant to job seekers and employers in this specific context.

1.4 Relevance

Comprehending the factors that impact salaries is imperative in the current competitive job market. For individuals seeking job opportunities, it can signify the distinction between obtaining a position with significant prospects for advancement and accepting a less favorable outcome. For employers, it underscores the significance of aligning salary frameworks with market trends in order to attract and retain proficient professionals. With the projected increase in demand for data analysts, this project addresses a crucial need by providing data-driven insights that are valuable to individuals and organizations navigating the evolving analytics landscape in Canada.

CHAPTER TWO

DATA DESCRIPTION

2.0 Dataset Overview

The dataset utilized for this project was sourced from Kaggle, a platform known for high-quality datasets across various domains. This dataset provides comprehensive information on data analyst job roles within Canada, including variables that are crucial for analyzing salary trends, industry impact, and geographic variation in job compensation. This data was originally scraped from job listing platforms like Indeed and Glassdoor, offering a real-world, raw view of the job market for data analysts in Canada.

2.1 Types of Data Collected

The dataset contains several key attributes that aligns closely with the objectives of the research, which is analyzing how Industry and location can determine salary trends for Data Analyst in Canada. The primary variable collected include:

1. Job Title: This variable lists the specific job titles within the data analytics domain, such as “Business System Analyst,” “Financial and Operational Analysts,” and “Senior Data Analyst.” These variations allow for granular role-based analysis and are relevant to understanding distinctions in job responsibilities and associated salaries.
2. Salary Information: Salary ranges, provided in annual figures in CAD (Canadian dollar), are included for each role such as “Minimum Salary,” “Maximum Salary,” and “Average Salary”. This is the central variable for analyzing how salary levels vary by industry and location, forming the basis of the study’s exploration of compensation trends.

3. Location (Province/Region): Each record includes geographic information at the province or city level, making it possible to analyze how salaries differ across the regions, such as Ontario, British Columbia, and Alberta.
4. Industry: This variable categorizes each job posting by industry, with sectors like technology, finance, and healthcare. Given that the demand for data analysts may vary across sectors, this attribute is critical for comparing salary differences by industry.
5. Skills/Experience Requirements: The dataset also contains fields detailing essential skills or qualifications for each position, such as programming languages (e.g., Python, SQL) and data analysis tools (e.g., Tableau, Power BI). This sheds light on the level of experience required for different positions and how it could affect pay.

A thorough examination of trends, comparisons, and correlations within the data can be made possible by the variety of data types that were gathered, which include both numerical (such as salary) and categorical (such as job title, location, and industry) variables.

2.2 Scale and Scope of Data

The dataset consists of approximately 1796 rows and 13 columns, which qualifies it as a robust sample of job advertisements from across Canada. This sizeable dataset enables statistically significant analysis, providing a broad view of salary distributions, industry presence, regional and local job employment availability for all forms of data analysts. The dataset's scope covers recent job listings from platforms like Indeed and Glassdoor, representing an accurate and precise picture of the Canadian labour market for data analysts.

2.3 Data Preprocessing

To prepare the dataset for analysis, several preprocessing steps were implemented to ensure data quality and consistency. These steps included:

2.3.1 Data Cleaning

1. Handling Missing Values: Missing values were present in some columns, especially in salary and job descriptions. We imputed values where possible, and records with critical missing fields were removed to ensure the dataset's integrity.
2. Standardizing Formats: Data inconsistencies were present, especially in job titles and salary information. Job titles were normalized to prevent duplication caused by minor title variations (e.g., "Data Analyst" vs. "Junior Data Analyst"). Salary information was standardized into a single unit (annual CAD) to maintain uniformity across records.

2.3.2 Normalization

Upon reviewing the dataset, it was determined that all salary data is already provided in a consistent format, specifically in annual CAD figures. As such, no further normalization or conversion of salary data was necessary. This ensured that the dataset was ready for analysis without requiring adjustments to salary units.

2.4 Feature Engineering

To enrich the dataset and derive meaningful insights, several new features were created based on existing columns:

2.4.1 Remote Work Indicator:

A new binary column, Remote Work, was introduced to classify job roles as either remote (1) or on-site (0). This was achieved by analyzing the city column and identifying entries such as "Remote in" or similar phrases.

1. Example: Remote in Montreal was classified as Remote while Montreal remained as an on-site job.

2.4.2 Province Standardization and Regional Grouping:

Provinces were grouped into broader geographic regions to enable regional-level analysis:

1. Western Canada: Alberta (AB), British Columbia (BC), Saskatchewan (SK).
2. Central Canada: Ontario (ON), Quebec (QC), Manitoba (MB).
3. Atlantic Canada: New Brunswick (NB), Nova Scotia (NS), Prince Edward Island (PE), Newfoundland and Labrador (NL).
4. Northern Canada: Yukon (YT), Northwest Territories (NT), Nunavut (NU).

This grouping helps in analyzing salary trends and job distributions by broader geographic zones.

2.4.3 Job Title Aggregation and Seniority Levels:

Job titles were aggregated to reduce redundancy and highlight meaningful distinctions. For example:

1. Titles such as “Junior Data Analyst” and “Entry-Level Data Analyst” were grouped under Junior Data Analyst.
2. A new column, Seniority Level, was added to classify job roles into Junior, Mid-Level, and Senior, based on keywords within the job titles.

2.4.4 Salary Bands:

To categorize salaries into manageable ranges, a Salary Band column was introduced using quantiles:

1. Low Salary Band: 0th–25th percentile.
2. Medium Salary Band: 26th–75th percentile.
3. High Salary Band: 76th–100th percentile.

This helps in understanding salary distributions across industries and locations.

2.5 Filtering

Focus on Relevant Roles:

The dataset was filtered to include only roles directly relevant to the scope of this research—data analyst job roles within Canada. Irrelevant or out-of-scope job titles and postings (e.g., "Programmer, Trainer") were excluded to maintain focus on the influence of industry and location on salaries.

2.5.1 Handling Non-Standard and Duplicated Entries:

1. Non-standard entries, such as misspelled cities (Montr√E ▵ √ÇR@al) or industries, were corrected to their accurate forms (Montreal).
2. Duplicate records were identified and removed to ensure data integrity, reducing noise in the analysis.

2.6 Descriptive Analysis and Visualization

To better understand the dataset and provide a foundation for further analysis, descriptive statistics and frequency counts were performed:

2.6.1 Numerical Variables:

Descriptive statistics for salary-related columns (Min Salary, Max Salary, Avg Salary) revealed outliers and variations across industries and regions. These were addressed through the outlier detection and feature engineering steps described earlier.

2.6.2 Categorical Variables:

Frequency counts were calculated for key categorical variables such as City, Province, Industry, and Position:

1. City: The dataset included 1,798 job listings across cities. Cities with terms like “Remote in” were carefully handled to distinguish between remote and on-site roles.
2. Province: A total of 13 provinces and territories were represented. Ontario (ON) accounted for the largest share of job listings, with 949 roles.
3. Industry: Listings spanned multiple industries, with the highest concentration in technology, finance, and healthcare.
4. Position: The dataset featured a wide range of roles, with Data Analyst being the most common, followed by Business Analyst and System Analyst.

2.6.3 Visualizations:

Several visualizations were generated to explore and summarize the data:

1. Job Distribution by Province: A bar chart (*Figure 2.1*) illustrates the distribution of job postings across provinces. As shown, Ontario (ON) accounts for most of data analyst roles, followed by British Columbia (BC) and Alberta (AB). Provinces like Yukon (YT) and Newfoundland and Labrador (NFL) had the least number of postings, reflecting the concentration of opportunities in urban and economically active regions.
2. Job Distribution by Industry: A second bar chart (*Figure 2.2*) demonstrates the distribution of job postings across industries. The technology and finance sectors dominate the demand for data analysts, reflecting the high data-dependency in these fields. Industries like public services and healthcare had fewer postings, consistent with their slower adoption of analytics roles.
3. Remote vs. On-Site Roles: A pie chart (*Figure 2.3*) visualizes the proportion of remote vs. on-site roles in the dataset. Approximately 92% of roles are on-site, while 8% are

classified as remote. This indicates that the majority of employers still favor in-office roles, though remote opportunities are growing.

4. Salary Distribution Across Regions: A box plot (Figure 5.3) in chapter 5 depicts the distribution of salaries across regions, including Western, Central, Atlantic, and Northern Canada. The plot highlights significant disparities, with Central Canada (Ontario and Quebec) displaying the highest median salaries. Atlantic Canada shows a wider range of salaries but with a lower median, reflecting the economic conditions of the region.
5. Salary Distribution Across Industries: Another box plot (Figure 5.1) captures the variation in salaries across industries. The technology sector emerges as the highest-paying field, while industries like healthcare and education tend to offer lower compensation. This visualization underscores the importance of industry in determining salary levels for data analyst roles.

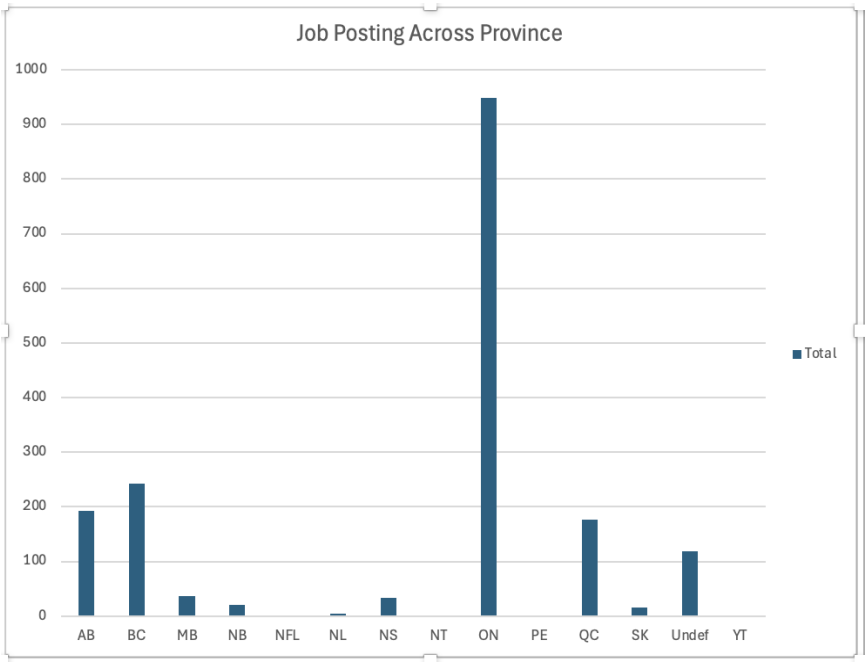

Figure 2.1: Bar Chart of Job Counts by Province

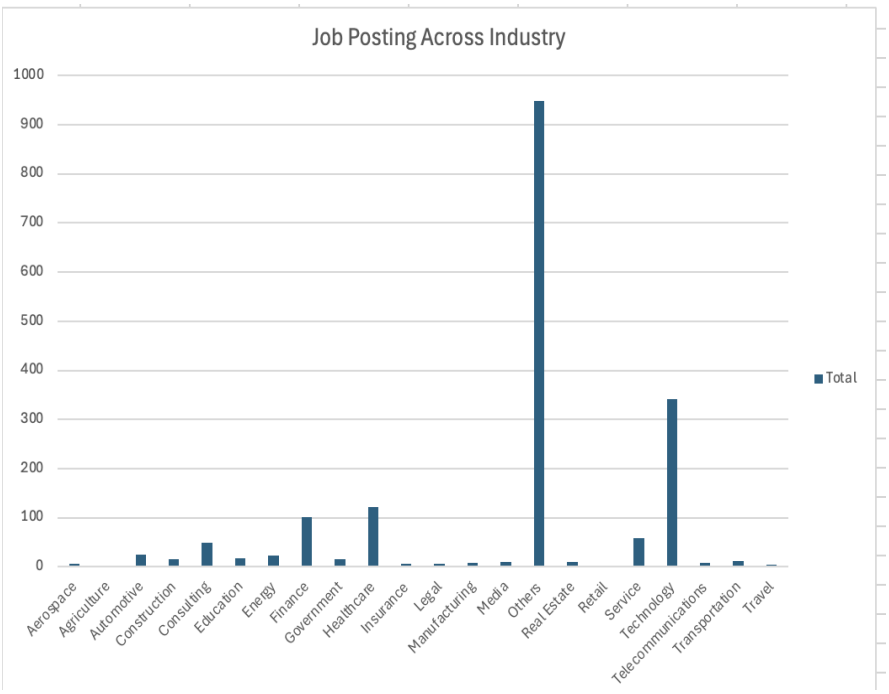

Figure 2.2: Bar Chart of Job Counts by Industry

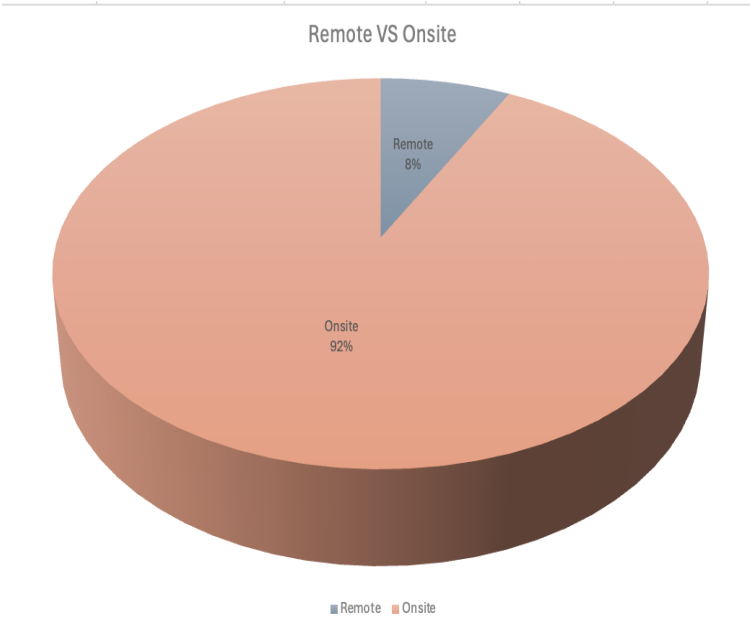

Figure 2.3: Pie Chart of Remote vs. On-Site Roles

CHAPTER THREE

HYPOTHESIS AND RESEARCH QUESTIONS

3.0 General Hypothesis

The salaries of data analysts in Canada are influenced by both industry type, and geographic location, with additional variations based on work mode (remote vs. on-site) and positions. These factors collectively shape salary trends and provide actionable insights for job seekers and employers.

3.1 Null and Alternative Hypotheses

To ensure statistical rigor and clarity, the hypotheses are divided into **Null (H_0)** and **Alternative (H_1)** forms:

H1: Industry Type Influences Salaries of Data Analysts in Canada.

1. Null Hypothesis (H_0): There is no significant difference in data analyst salaries across industries.
2. Alternative Hypothesis (H_1)**:** Salaries for data analysts differ significantly between industries, with certain sectors like technology and finance offering higher compensation.

H2: Geographic Location Significantly Impacts Salaries.

1. Null Hypothesis (H_0): There is no significant variation in data analyst salaries across provinces and regions in Canada.
2. Alternative Hypothesis (H_1)**:** Salaries for data analysts vary significantly by geographic location, with provinces like Alberta and Ontario offering higher average compensation.

H3: The Interaction Between Industry Type and Position Affects Salaries.

1. Null Hypothesis (H_0): There is no significant interaction between industry type and position in determining salaries. Salaries for positions (e.g., Data Analyst, Business Analyst) do not significantly differ across industries

2. Alternative Hypothesis (H_1): There is a significant interaction between industry type and position in determining salaries. Salaries for positions (e.g., Data Analyst, Business Analyst) vary significantly depending on the industry, with certain industries offering higher compensation for specific positions.

3.2 Research Questions

The hypotheses are guided by the following specific, measurable research questions:

1. Industry Influence: How do salaries vary across different industries for data analyst roles in Canada?
2. Geographic Disparities: What are the salary differences between provinces and regions in Canada?
3. Positions Effect: What is the combined influence of industry type and position on salary levels?

These hypotheses will be tested through statistical and exploratory analysis in subsequent chapters, using descriptive statistics, visualizations, and hypothesis testing methods (e.g., ANOVA, t-tests) to validate or reject them.

CHAPTER FOUR

METHODOLOGY

4.0 Methodology and Statistical Analysis Framework

4.1 Statistical Methods Overview for Hypothesis One

1. Descriptive statistics were used to summarize the dataset and identify the central tendencies, and variability of salary across the 23 industries. This included calculating:
    a. Metrics: Mean, Standard Deviation, Minimum, Maximum, Count.
    b. Purpose: Summarize salary trends across industries.
    c. Tools Used: Python's pandas library for data aggregation and statistical summaries.
2. One-Way ANOVA:
    a. Purpose: Test for statistically significant differences in average salaries across industries.
    b. Tools Used: Python's scipy.stats.f_oneway.
3. Chi-Square Test:
    a. Purpose: Test the association between salary bands and industries.
    b. Tools Used: Python's scipy.stats.chi2_contingency.

4.1.2 Visualization Techniques

1. Box Plot: Used to represent the salary distribution across industries visually. This highlighted median salaries, interquartile ranges, and outliers within the data as shown in *(Figure 5.1)*.
2. Bar Chart: Employed to display the average salary for each industry, enabling an easy comparison across sectors depicted in *(Figure 5.2).*

4.1.3 Justification

1. Excel: Ideal for organizing and preprocessing large datasets.
2. Google Colab (Python): Suitable for advanced statistical analysis and hypothesis testing and use of Python provides reproducibility, scalability, and precise statistical computations.
3. One-Way ANOVA: Examined average salary differences across industries.
4. Chi-Square Test: Analyzed the relationship between industries and salary bands.

4.2 Statistical Methods Overview for Hypothesis Two

1. Descriptive Statistics: Summarize salary trends across regions (Atlantic Canada, Central Canada, and Northern Canada) to identify patterns in average salaries and their variability.
    a. Metrics: Mean, median, standard deviation, minimum, maximum, and count.
    b. Role: Provides a foundational understanding of salary variations between regions and helps to visualize general trends.
    c. Tools:
        i. Python's pandas library for data grouping and aggregation.
        ii. groupby() and statistical functions such as .mean(), .std(), etc.
2. One-Way ANOVA:
    a. Purpose: Test for significant differences in salaries across regions.

b. Tools Used: Python’s scipy.stats.f_oneway.

3. Chi-Square Test:

   a. Purpose: Analyze the association between regions and salary bands.

4.2.2 Visualization Techniques

   a. Box Plot: Show the distribution of salaries (range, quartiles, outliers) within each region to illustrate disparities visually shown in *(Figure 5.3)*.

   b. Tools:

      i. Python's seaborn library for generating box plots.

   c. Role: Offers a visual representation of salary distribution and highlights median differences and the presence of outliers in the data.

4.2.3 Justification

1. Descriptive Statistics and Box Plots:

   a. These techniques effectively summarize and visualize salary trends, offering both numerical and graphical insights into regional differences.

   b. They set the stage for hypothesis testing by highlighting areas of potential significance.

2. One-Way ANOVA:

   a. Suitable for comparing means across three or more groups (regions in this case).

   b. Its ability to determine statistical significance provides clarity on whether regional salary disparities are due to random variation or actual differences.

3. Chi-Square Test:

   a. Appropriate for examining relationships between two categorical variables (region and salary band).

b. Ensures that the impact of categorical variables is not overlooked when analyzing salary structures.

4.3 Statistical Methods Overview for Hypothesis Three

1. Two-Way ANOVA:
    a. Purpose: Examine the interaction between two categorical variables—Industry Type and Merged Position—on the dependent variable Avg_Salary.
    b. Tools Used: Python's statsmodels.api.ols.
2. Kruskal-Wallis Test:
    a. Purpose: Test for differences in salary distributions across groups when normality assumptions are violated.
    b. Role: Provides a non-parametric alternative to ANOVA for analyzing salary variations within specific industries or positions.
    c. Tool Used: Python (scipy.stats).

4.3.2 Visualization Techniques

1. Bar Plot: Display mean salaries for different positions within each industry to provide a visual overview of trends.
2. Box Plot: Illustrate salary distributions across regions and positions to identify patterns and outliers.
3. Tool Used: Python (matplotlib and seaborn).

4.3.3 Justification

1. Two-Way ANOVA: Allows simultaneous testing of two independent variables and their interaction effect, offering a comprehensive view of the factors influencing salaries.
2. Kruskal-Wallis Test: Addresses situations where salary data deviate from normality, making the results robust against such violations.

3. Visualization: Charts and graphs enhance understanding by presenting complex data in an easily interpretable format.

CHAPTER FIVE
RESULTS

5.0 Results Overview

The hypothesis that industry type significantly influences salaries for data analysts was tested using descriptive statistics, One-Way ANOVA, and Chi-Square analysis. The statistical analyses and visualizations provided insights into the relationships between salaries and various factors, including industry type, region, position, work type, and seniority.

5.1 Results for Hypothesis One: Industry Type Influences Salaries

5.1.2 Descriptive Analysis

1. Technology had the highest average salary: $27,350,022.15.
2. Aerospace and Insurance sectors exhibited the lowest average salaries ($609,520 and $326,207.5, respectively).
3. Standard deviation was highest for Technology ($18,421.49), indicating greater salary variation.

5.1.3 Hypothesis Testing

Two statistical tests were performed to validate the hypothesis:

1. One-Way ANOVA:
    a. Purpose: To test whether there are statistically significant differences in the average salaries across different industries.
    b. Test Statistic: F-value = 1.2923, p-value = 0.2755.
    c. Outcome: No significant difference in average salaries across industries, as $p > 0.05$.

2. Chi-Square Test:
    a. Purpose: To examine the association between industries and salary bands (high, medium, low).
    b. Test Statistic: Chi-Square = 49.6151, p-value = 0.1957.
    c. Outcome: No significant association between salary bands and industries, as p > 0.05.
3. Visualizations: Figure 5.1 ,5.2 and Figure 5.3 shows
    a. Bar Plot: Highlights Technology as the highest-paying industry.
    b. Box Plot: Shows variability in salaries within industries.
    c. Stacked bar chart shows distribution based on salary bands

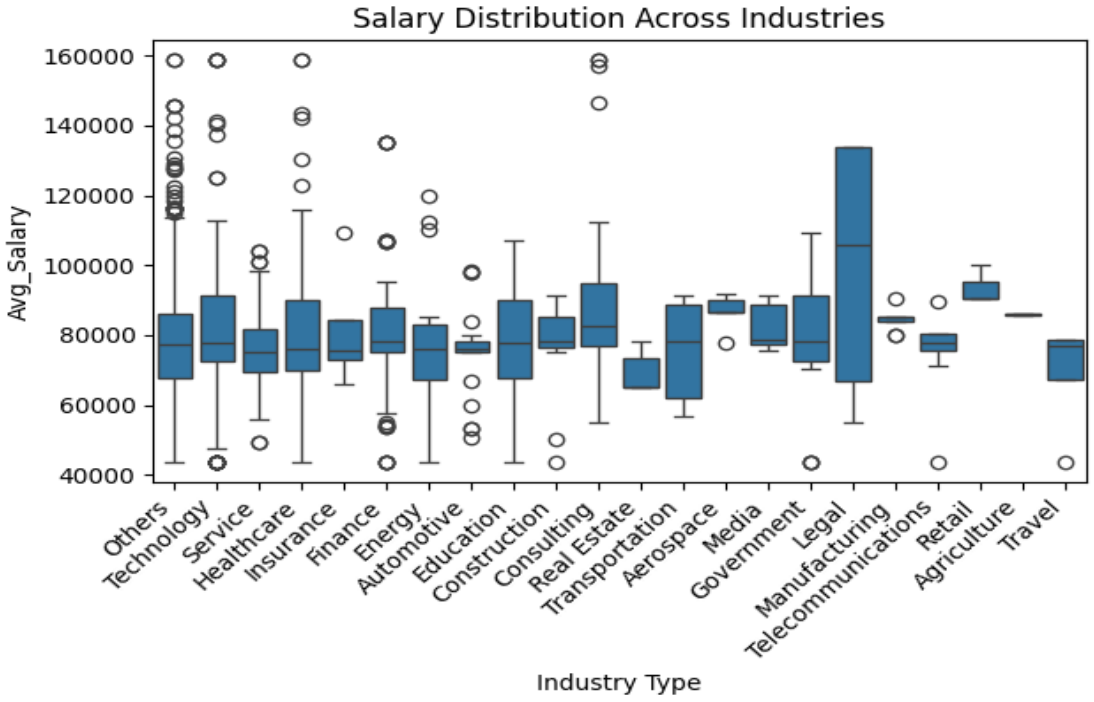

Figure 5.1 Boxplot Salary distribution across Industries

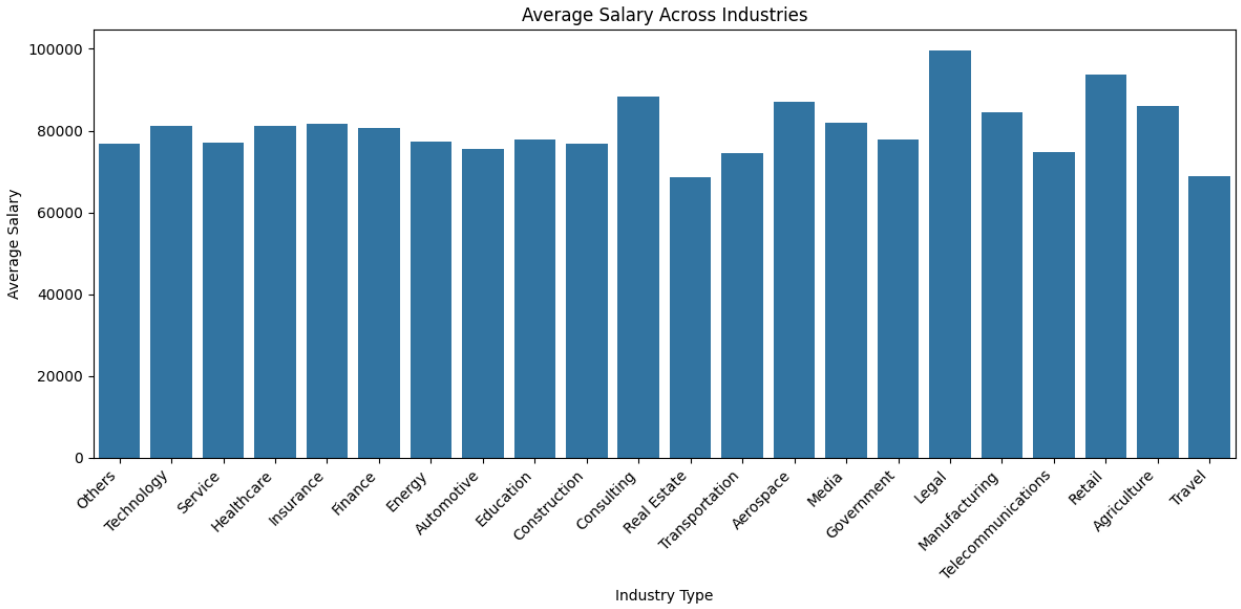

Figure 5.2 Bar chart Salary distribution across Industries

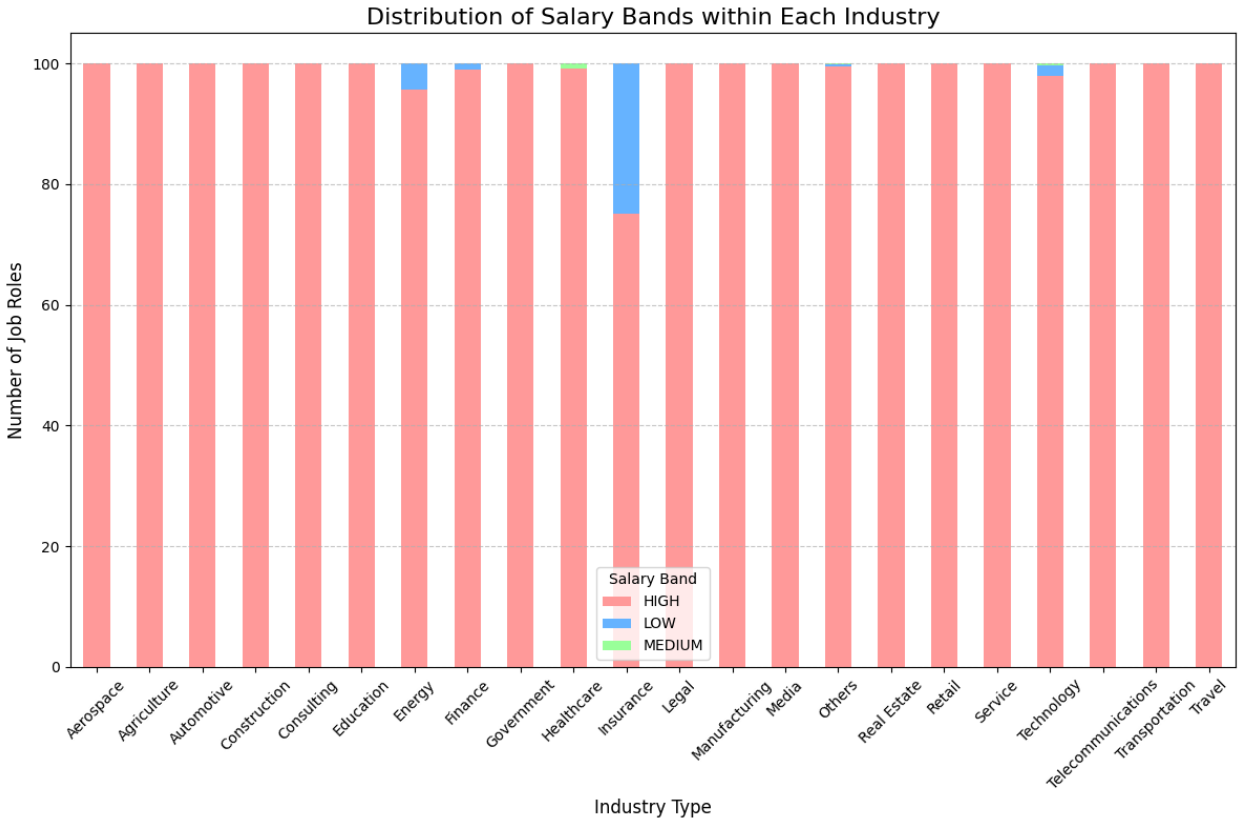

Figure 5.1 Boxplot Salary distribution across Industries

For Hypothesis One, based on the results of the statistical tests:

1. One-Way ANOVA:
    a. The p-value is 0.2755 (greater than 0.05).
    b. This indicates that there is no statistically significant difference in average salaries across industries.
    c. Conclusion: We fail to reject the null hypothesis.
2. Chi-Square Test:
    a. The p-value is 0.1957 (greater than 0.05).
    b. This indicates no statistically significant association between industries and salary bands (high, medium, low).
    c. Conclusion: We fail to reject the null hypothesis.

Since neither test provides evidence to reject the null hypothesis, the alternative hypothesis is not supported. This means that, statistically, industry type does not significantly influence salary levels or distribution, despite observable trends in the descriptive statistics.

5.2 Results for Hypothesis Two: Geographic Location Impacts Salaries

5.2.2 Descriptive Analysis:

1. Northern Canada showed the highest mean salary ($81,055.09).
2. Atlantic Canada had the lowest ($72,052.03).

5.2.3 Hypothesis Testing

Two statistical tests were performed to validate the hypothesis:

1. One-Way ANOVA Results:
    a. F-statistic = 10.89, $p < 0.05$ (2.001659758711854e-05).
    b. Conclusion: Significant salary differences across regions.
2. Chi-Square Test Results:
    a. Chi-Square = 1.51, $p = 0.82$.
    b. Conclusion: No significant association between salary bands and regions
3. Visualizations: Figure 5.4
    a. Box Plot: Highlights regional salary disparities

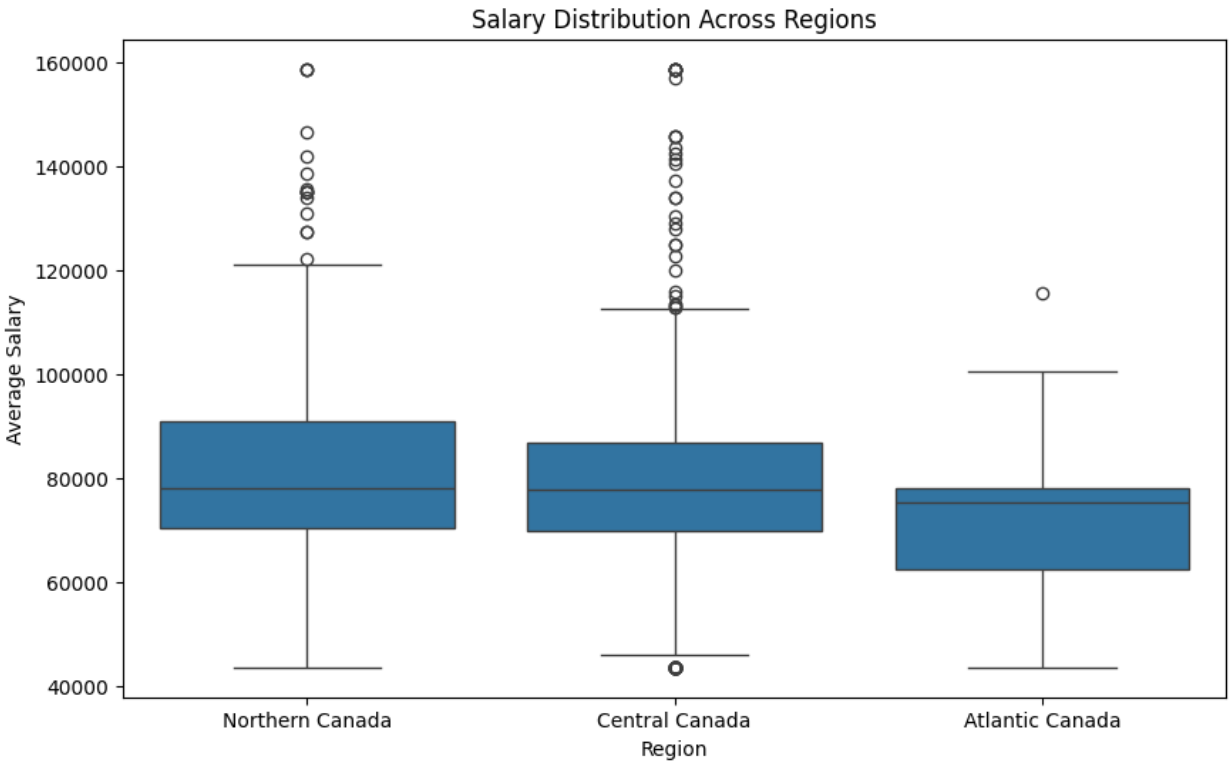

Figure 5.3 Boxplot Salary distribution across regions

For Hypothesis Two, based on the results of the statistical tests:

1. One-Way ANOVA Results:

a. F-Statistic: *10.8910.89*10.89
b. P-Value: *2*.001659758711854e-05
c. Interpretation:
    i. Since p<0.05, we reject the null hypothesis.
    ii. Conclusion: Significant differences exist between regional salaries.

2. Chi-Square Test Results:
    a. Chi-Square: *1.511.51*1.51
    b. P-Value: *0.820.82*0.82
    c. Interpretation:
        i. Since p>0.05, we fail to reject the null hypothesis.
        ii. Conclusion: No significant association exists between regions and salary bands.

These findings underscore the influence of geography on salary trends, affirming the hypothesis that regional differences significantly impact average salaries but not necessarily the overall salary band distribution.

5.3 Results for Hypothesis Three: Interaction Between Industry Type and Position Affects Salaries

5.3.2 Statical Analysis

Two-Way ANOVA Results:

a. Industry Type Main Effect (F=1.72, p=0.064):

    Marginally insignificant, suggesting weak evidence that industries impact salaries.

b. Merged Position Main Effect (F=0.63, p=0.535):

    No significant difference in salaries across positions when industry effects are ignored.

c. Interaction Effect ($F=2.17$, $p<0.001$):
    a. Significant interaction indicates that the relationship between salaries and positions depends on the industry.

Kruskal-Wallis Test Results:

For industries and positions where normality assumptions were violated, the Kruskal-Wallis Test provided additional insights:

a. Technology ($p=0.0036$): Salaries differ significantly across positions, highlighting the influence of the role within the tech sector.
b. Finance ($p=0.0429$): Salary variations are observed among positions in finance.
c. Others ($p<0.0001$): Indicates highly significant differences across positions in miscellaneous industries, the results are shown in Appendix C.

Visualization

1. Bar Plot of Average Salaries by Industry and Position shown in *(Figure 5.4)* and *(Figure 5.5)*:
    a. Technology offers the highest salaries across most positions, particularly for Data Engineers.
    b. Lower-paying positions, such as Risk Analyst, are less common in sectors like education and healthcare.
2. Box Plot of Salary Distributions by Region and Position:
    a. Central Canada demonstrates broader salary variability, particularly for Data Analysts and System Analysts.
    b. Atlantic Canada has more consistent salary ranges, with narrower distributions.

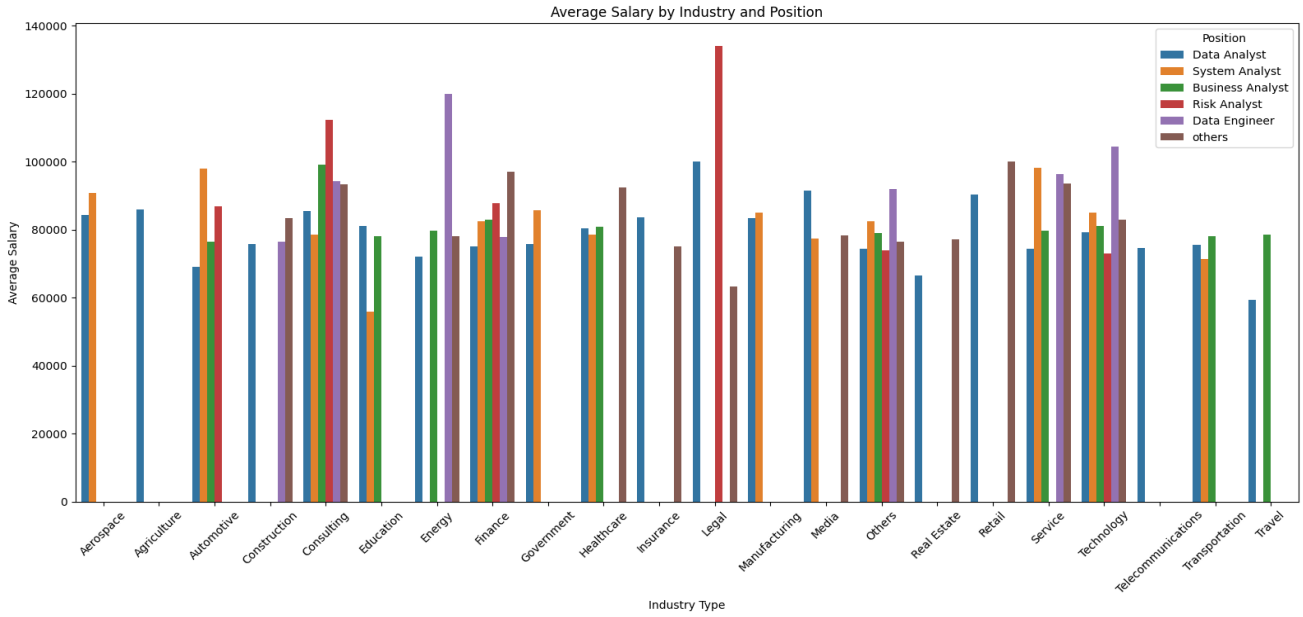

Figure 5.4: Bar Chart: Salary Distribution by Region and Position

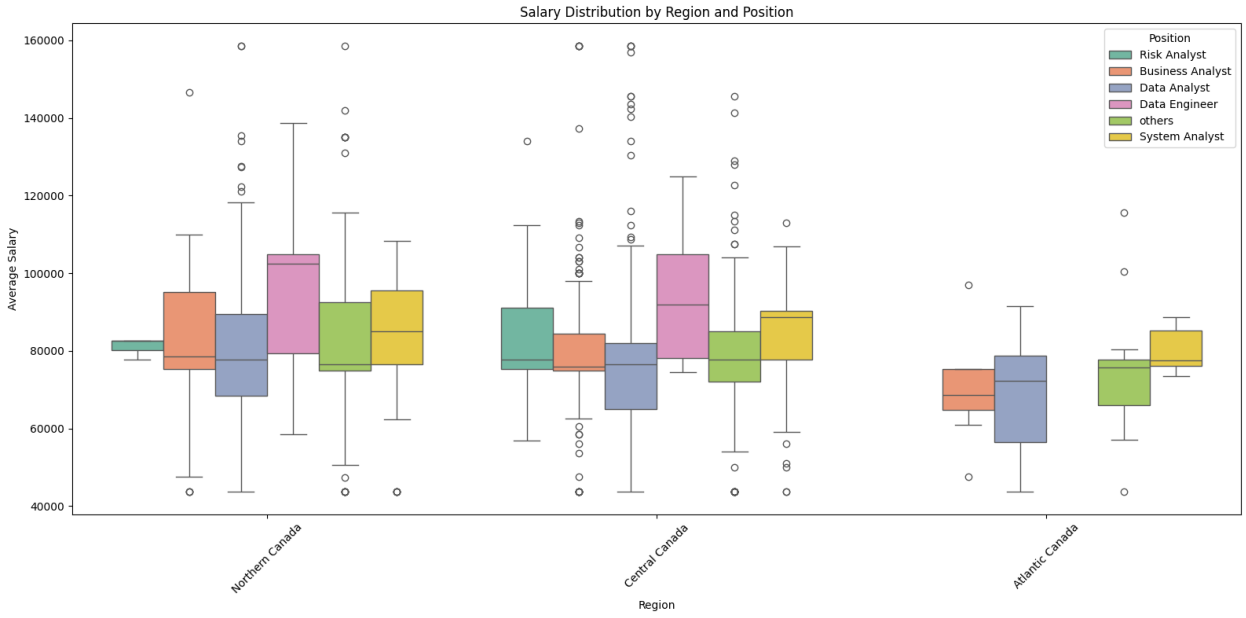

Figure 5.5: Box Plot: Salary Distribution by Region and Position

The results validate that the combined influence of industry and position significantly impacts salary trends. This insight underscores the need for tailored career planning strategies based on specific sectors and roles.

CHAPTER SIX

DISCUSSION

The findings from this study align with previous research that highlights significant disparities in salaries across different industries and regions. For instance, the observed variation in salaries between regions like Northern Canada and Atlantic Canada mirrors the broader geographic disparities highlighted by Reichert (2020). This suggests that regional economic conditions and workforce demand play pivotal roles in salary structuring.

The role of industry type in influencing salaries is consistent with studies that identify technology and finance as high-paying sectors (Gupta, Gulliver, & Singh, 2023). The significant interaction effect observed between industry type and position aligns with Southam and Sapp's (2010) findings on the importance of role-specific compensation across industries.

Additionally, the information regarding regional differences in salaries emphasises how crucial spatial economic analysis is. This reinforces the study's contribution to understanding the nuanced factors affecting salaries in Canada, particularly for data analysts.

While this research successfully identifies trends, certain limitations persist. For example, insufficient data in some industries and regions, such as Agriculture and Telecommunications, restricts broader generalization. Future research should address this gap by incorporating larger datasets and focusing on emerging industries.

CHAPTER SEVEN

RECOMMENDATIONS

The findings underscore several actionable recommendations for stakeholders:

1. Industry-Specific Compensation Strategies: Employers in technology and finance sectors should continue leveraging competitive compensation as a talent attraction strategy. These insights align with the recommendations by Wu and Hewage (2024) on equitable pay policies across high-demand roles.
2. Regional Pay Adjustments: Employers in lower-paying regions, such as Atlantic Canada, should consider introducing regional pay adjustments to attract skilled professionals. This strategy has been supported by Reichert (2020), who emphasized the role of regional incentives in improving workforce distribution.
3. Tailored Career Planning: Universities and career advisors should guide data analysts toward industries and regions with higher salary prospects. For instance, sectors like technology not only offer competitive salaries but also opportunities for role specialization (Gupta, Balcom, & Singh, 2022).
4. Policy Advocacy for Equitable Pay: Policymakers should address income inequality by promoting industry-specific wage regulations, as discussed in Kakwani's (1980) framework for equitable compensation.

These recommendations provide a roadmap for stakeholders to address salary disparities effectively and ensure equitable pay practices across industries and regions.

CHAPTER EIGHT

CONCLUSION

This study demonstrates that salaries for data analysts in Canada are influenced by a combination of industry type, geographic location, and the interaction between industry and position. Consistent with findings by Gupta, Gulliver, and Singh (2023), the technology and finance industries were identified as high-paying sectors, while regions like Northern Canada exhibited the highest average salaries.

The significant interaction effect between industry type and position, as highlighted by Southam and Sapp (2010), emphasizes the importance of role-specific compensation strategies. This insight is crucial for both employers and policymakers seeking to address salary disparities.

Despite these contributions, certain limitations exist. The study's findings are constrained by the availability of data for certain industries and regions, as noted by Wu and Hewage (2024). Future research should expand on this work by exploring emerging sectors and incorporating longitudinal salary data to better capture trends over time.

This project adds value to the existing body of knowledge by providing actionable insights into salary structures for data analysts in Canada. It offers a foundation for further research on equitable pay practices and highlights the need for industry and region-specific strategies to ensure fair compensation for skilled professionals.

REFERECNCES

Gupta, N., Balcom, S. A., & Singh, P. (2022). Gender composition and wage gaps in the Canadian health policy research workforce. *Human Resources for Health, 20*(1), 1–12. https://doi.org/10.1186/s12960-022-00774-5

Gupta, N., Gulliver, A., & Singh, P. (2023). Relative remoteness and wage differentials in the Canadian allied health professional workforce. *Informit Journal.* https://search.informit.org/doi/abs/10.3316/informit.197554040496878

Kakwani, N. C. (1980). *Income inequality and poverty: Methods of estimation and policy applications.* Oxford University Press. https://doi.org/10.1093/0195200976.001.0001

Reichert, P. (2020). Internationalization & career-focused programming for international students. *University of Calgary.* https://prism.ucalgary.ca/server/api/core/bitstreams/308e4993-fd2c-4316-ae74-4b34026e10d9/content

Southam, C., & Sapp, S. (2010). Compensation across executive labor markets. *Journal of International Business Studies, 41*(8), 1398–1418. https://doi.org/10.1057/jibs.2009.34

Wu, L., & Hewage, W. (2024). Investigating equity in remote salaries in the data science field. *Rere Awhio Journal.* https://online.op.ac.nz/assets/Uploads/AIC096-Rere-Awhio-Journal-2024_WEB.pdf#page=34

APPENDICES

Appendix A: Hypothesis one python code

```
import pandas as pd
from scipy.stats import f_oneway
# Load dataset
data = pd.read_csv("Book3.csv")
# Extract salaries grouped by Industry Type
industries = data.groupby('Industry Type')['Avg_Salary']
# Separate salary data by industry
technology = industries.get_group('Technology')
healthcare = industries.get_group('Healthcare')
service = industries.get_group('Service')
# Perform One-Way ANOVA
f_stat, p_value = f_oneway(technology, healthcare, service)
print(f"F-Statistic: {f_stat}, P-Value: {p_value}")
```

F-Statistic: 1.292310794941507, P-Value: 0.27553849588364837

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.boxplot(x='Industry Type', y='Avg_Salary', data=data)
plt.title('Salary Distribution Across Industries')
plt.show()
from scipy.stats import chi2_contingency
# Create contingency table
contingency_table = pd.crosstab(data['Industry Type'], data['Salary Bound'])
# Perform Chi-Square Test
chi2, p_value, dof, expected = chi2_contingency(contingency_table)
print(f"Chi-Square: {chi2}, P-Value: {p_value}, Degrees of Freedom: {dof}")
```

Chi-Square: 49.615144895999876, P-Value: 0.19567534778256082, Degrees of Freedom: 42

```
import pandas as pd
import matplotlib.pyplot as plt
# Load your dataset
df = pd.read_csv("Book3.csv")
# Group data by Industry Type and Salary Bound
grouped_data = df.groupby(['Industry Type', 'Salary Bound']).size().unstack(fill_value=0)
# Display grouped data (for verification)
print(grouped_data)
# Convert counts to percentages
```

```
grouped_data_percentage = grouped_data.div(grouped_data.sum(axis=1), axis=0) * 100
# Plot the stacked bar chart
grouped_data_percentage.plot(kind='bar', stacked=True, figsize=(12, 8),
color=['#ff9999','#66b3ff','#99ff99'])
# Add labels and title
plt.title("Distribution of Salary Bands within Each Industry", fontsize=16)
plt.xlabel("Industry Type", fontsize=12)
plt.ylabel("Number of Job Roles", fontsize=12)
plt.legend(title="Salary Band", fontsize=10)
plt.xticks(rotation=45)
plt.grid(axis='y', linestyle='--', alpha=0.7)
# Show the plot
plt.tight_layout()
plt.show()
```

Salary Bound HIGH LOW MEDIUM Industry Type Aerospace 7 0 0 Agriculture 2 0 0

Automotive 26 0 0 Construction 15 0 0 Consulting 49 0 0 Education 18 0 0 Energy 22 1 0

Finance 99 1 0 Government 15 0 0 Healthcare 115 0 1 Insurance 3 1 0 Legal 6 0 0

Manufacturing 8 0 0 Media 10 0 0 Others 924 4 1 Real Estate 10 0 0 Retail 3 0 0 Service 57 00

Technology 330 6 1 Telecommunications 8 0 0 Transportation 12 0 0 Travel 4 0 0

Appendix B: Hypothesis two python code

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import f_oneway, chi2_contingency
# Load your dataset
data = pd.read_csv("Book3.csv")
# Descriptive Statistics by Region
region_summary = data.groupby('Region')['Avg_Salary'].agg(['mean', 'median', 'std', 'min', 'max',
'count'])
print(region_summary)
```

	mean	median	std	min	max	\
Region						
Atlantic Canada	72052.032131	75250.0	14247.150620	43720.28	115689.6	
Central Canada	77736.802578	77750.0	17471.940393	43720.28	158640.0	
Northern Canada	81055.089145	78000.0	18780.843065	43720.28	158640.0	

	count
Region	
Atlantic Canada	61
Central Canada	1125
Northern Canada	573

```
# Box Plot
plt.figure(figsize=(10, 6))
sns.boxplot(x='Region', y='Avg_Salary', data=data)
plt.title("Salary Distribution Across Regions")
plt.xlabel("Region")
plt.ylabel("Average Salary")
plt.show()
# One-Way ANOVA
regions = [data[data['Region'] == region]['Avg_Salary'] for region in data['Region'].unique()]
f_stat, p_value = f_oneway(*regions)
print(f"F-Statistic: {f_stat}, P-Value: {p_value}")

# Interpretation
if p_value < 0.05:
print("Reject the null hypothesis: Significant differences exist between regions.")
else:
print("Fail to reject the null hypothesis: No significant differences between regions.")
```

F-Statistic: 10.885880360276799, P-Value: 2.001659758711854e-05 Reject the null hypothesis: Significant differences exist between regions.

```
# Contingency Table
contingency_table = pd.crosstab(data['Region'], data['Salary Bound'])
chi2, p_value, dof, expected = chi2_contingency(contingency_table)
print(f"Chi-Square: {chi2}, P-Value: {p_value}, Degrees of Freedom: {dof}")

# Interpretation
if p_value < 0.05:
print("Reject the null hypothesis: Region and Salary Band are associated.")
else:
```

```
print("Fail to reject the null hypothesis: Region and Salary Band are independent.")
```

Chi-Square: 1.5145061857361106, P-Value: 0.824068809623743, Degrees of Freedom: 4 Fail to reject the null hypothesis: Region and Salary Band are independent

Appendix C: Hypothesis three python code

```
# Import necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import f_oneway, kruskal
# Load the dataset (assuming you have already loaded it into 'df')
# df = pd.read_csv("your_dataset.csv")
# Step 1: Group by Industry and Position
industry_position_group = df.groupby(['Industry Type', 'Merged_Position']).agg(
Avg_Salary=('Avg_Salary', 'mean'),
Min_Salary=('Avg_Salary', 'min'),
Max_Salary=('Avg_Salary', 'max'),
Count=('Avg_Salary', 'count')
).reset_index()
print(industry_position_group)
# Step 2: Visualize Trends (Bar Plot of Average Salary by Industry and Position)
plt.figure(figsize=(16, 8))
sns.barplot(data=industry_position_group, x='Industry Type', y='Avg_Salary', hue='Merged_Position')
plt.title('Average Salary by Industry and Position')
plt.xticks(rotation=45)
plt.ylabel('Average Salary')
plt.xlabel('Industry Type')
plt.legend(title='Position')
plt.tight_layout()
plt.show()
# Step 3: Visualize Salary Distributions by Position within Regions
plt.figure(figsize=(16, 8))
sns.boxplot(data=df, x='Region', y='Avg_Salary', hue='Merged_Position', palette='Set2')
plt.title('Salary Distribution by Region and Position')
plt.xticks(rotation=45)
plt.ylabel('Average Salary')
plt.xlabel('Region')
plt.legend(title='Position')
plt.tight_layout()
plt.show()
```

```
# Step 4: Statistical Testing (Two-Way ANOVA)
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
# Prepare the data for Two-Way ANOVA
# Step 4: Rename columns to remove spaces or special characters
anova_data = df[['Avg_Salary', 'Industry Type', 'Merged_Position']].copy()
anova_data.rename(columns=lambda x: x.strip().replace(' ', '_').replace('-', '_'), inplace=True)
# Fit the ANOVA model
anova_model = ols('Avg_Salary ~ C(Industry_Type) + C(Merged_Position) +
C(Industry_Type):C(Merged_Position)',
data=anova_data).fit()
anova_results = anova_lm(anova_model, typ=2)
print("Two-Way ANOVA Results:")
print(anova_results)
# Step 5: Kruskal-Wallis Test (Non-Parametric Test for Position within Industries)
industry_groups = anova_data.groupby('Industry_Type')
kruskal_results = {}
for industry, group_data in industry_groups:
# Check if there are at least 2 groups with sufficient data
position_groups = [group['Avg_Salary'].values for name, group in
group_data.groupby('Merged_Position') if len(group) > 1]
if len(position_groups) > 1: # At least 2 groups needed for Kruskal-Wallis
stat, p = kruskal(*position_groups)
kruskal_results[industry] = (stat, p)
else:
kruskal_results[industry] = (None, "Insufficient data")

print("\nKruskal-Wallis Test Results:")
for industry, result in kruskal_results.items():
print(f"{industry}: Statistic={result[0]}, P-Value={result[1]}")
```

Two-Way ANOVA Results: sum_sq df F PR(>F) C(Industry_Type) 1.086345e+10 21.0 1.719807 0.063603 C(Merged_Position) 9.415950e+08 5.0 0.626073 0.534811 C(Industry_Type):C(Merged_Position) 6.838418e+10 105.0 2.165197 0.000001 Residual 5.074392e+11 1687.0 NaN NaN Kruskal-Wallis Test Results: Aerospace: Statistic=4.9411764705882355, P-Value=0.026224181348737993 Agriculture: Statistic=None, P-Value=Insufficient data Automotive: Statistic=5.937317954095323, P-Value=0.11469998762686731 Construction: Statistic=0.3145161290322551, P-Value=0.5749220401700352 Consulting: Statistic=3.442619153358599, P-

Value=0.4866566914285323 Education: Statistic=2.950967161493466, P-Value=0.22866812080324125 Energy: Statistic=1.3912614980289093, P-Value=0.23819217832018205 Finance: Statistic=11.467820290851229, P-Value=0.0428541608180724 Government: Statistic=0.3381642512077391, P-Value=0.5608907723295907 Healthcare: Statistic=4.14422257333445, P-Value=0.24630593122428085 Insurance: Statistic=None, P-Value=Insufficient data Legal: Statistic=None, P-Value=Insufficient data Manufacturing: Statistic=0.7407407407407428, P-Value=0.38942369573502533 Media: Statistic=6.579999999999995, P-Value=0.0372538493962159 Others: Statistic=55.651342395529, P-Value=9.587189236063813e-11 Real Estate: Statistic=4.401041666666659, P-Value=0.03591698606632877 Retail: Statistic=None, P-Value=Insufficient data Service: Statistic=7.872716696917937, P-Value=0.04871683154403238 Technology: Statistic=17.549212012223048, P-Value=0.0035678056094995635 Telecommunications: Statistic=None, P-Value=Insufficient data Transportation: Statistic=0.5182567726737392, P-Value=0.7717239378737903 Travel: Statistic=2.3999999999999986, P-Value=0.12133525035848367