CS2545 Mini-Project

Title: How Geographic Location and Lifestyle Factors Correlate with Health Outcomes in Canada

Student Name: Pierre Jaurel Ekotto Mebande

Student Number: 3725759

Course: CS2545

1. Introduction

Health outcomes differ between provinces and territories in Canada. These differences can be seen when looking at life expectancy, self-reported health status, obesity and death causes. The purpose of this project is to examine how physical activity, smoking and BMI connect with health outcomes and geographic location.

2. Related Work

The Canadian Government has recognized that various health determinants shape the general wellness of all citizens. According to the Public Health Agency of Canada, health factors affect Canadians because both genetic and biological elements, and lifestyle and socio-economic status and geographic areas play roles in health outcomes [Government of Canada, 2022].

Other variables such as income, social status, education, environment quality, health practices and access to healthcare services create health outcome disparities between Canadian regions. This study uses existing knowledge about lifestyle indicators to analyze how smoking rates, physical activity behaviour, and BMI levels affect life expectancy and personal health assessment results in Canadian geographical areas.

3. Problem Statement and Requirements

3.1 Problem Statement

Public health planners and researchers encounter a fundamental challenge to understand why Canadian provinces perform differently in terms of their health outcomes. It is a real challenge to analyze disparities because of inconsistent measurement and lack of understanding.

When factors including physical activity, smoking and body mass index remain obscure to regional health trend analysis, then creating effective intervention programs and strategic health resource management becomes complex. The absence of insight creates obstacles that prevent Canadian health organizations from eliminating health disparities and raising healthcare standards for all Canadian residents.

Our aim is to assess the relationship between lifestyle and local health patterns. Analyzing open data will produce useful findings for implementing equitable health policies across Canada.

3.2 Requirements

- All datasets in this project need to derive from publicly accessible sources
- Include geographic visualization with proper labelling
- Include statistical analysis and charts
- Implement with Jupyter Notebook using Python

4. Approach

This project uses a quantitative approach. We use data from Statistics Canada and the Canada Open Government Portal. Our datasets include life expectancy by province, body mass index and obesity by region, and a survey on perceived health and mortality statistics by condition and location.

All datasets are in CSV format and cleaned using Pandas to ensure naming conventions throughout the data sources. We merge appropriate datasets to analyze the relationships. We identify patterns through descriptive statistics, correlation analysis and visual analysis. We use Matplotlib and Seaborn to connect data.

This approach clearly examines health disparities in Canada. The analysis includes datasets from various years, but we try to match them chronologically and discuss limitations in the final sections.

5. Analysis

5.1 Load Libraries

Import Python Libraries required for the analysis

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

5.2 Load Datasets

Load csv dataset into clearly labeled dataframes (bmi, deaths_by_cause, life_expectancy, perceived_health) using Pandas.

```
bmi = pd.read_csv('BMI.csv') # BMI and Obesity by province
deaths_by_cause = pd.read_csv('Death_by_Cause.csv') # Deaths by cause
of death
life_expectancy =
pd.read_csv('Life_Expectancy_Canada_2000_to_2007.csv') # Life
expectancy by province
perceived_health = pd.read_csv('Perceived_Health_Canada_2015.csv') #
Perceived health by province
```

5.3 Data Cleaning and Exploratory Analysis

5.3.1 Body mass index and Obesity

Here we want three obesity groups: Normal, Overweight and Obese where both populations should be displayed. We want BMI percentage data breakdown in both provincial and yearly.

```
# Clean the column names for better usability
bmi.columns = bmi.columns.str.strip().str.replace(" ",
"_").str.replace("(", "").str.replace(")", "")
# Filter bmi for relevant data
bmi_filtered = bmi[
    (bmi["Sex"] == "Both sexes") &
    (bmi["Characteristics"] == "Percent") &
    (bmi["Body_mass_index_BMI"].isin([
        "Normal weight, body mass index 18.50 to 24.99",
        "Overweight, body mass index 25.00 to 29.99",
        "Obese, body mass index 30.00 or higher"
    ]))
]
# Rename columns for clarity
bmi_filtered = bmi_filtered.rename(columns={
    "REF_DATE": "Year",
    "GEO": "Province",
    "Body_mass_index_BMI": "BMI_Category",
    "VALUE": "Percentage"
})
# Most recent year
bmi_filtered = bmi_filtered[bmi_filtered["Year"] ==
bmi_filtered["Year"].max()]
# Save the cleaned data
bmi_filtered.to_csv('cleaned_bmi_data.csv', index=False)
# Average obesity rate by province
bmi_obese = bmi_filtered[bmi_filtered["BMI_Category"] == "Obese, body
mass index 30.00 or higher"]
avg_obese = bmi_obese.groupby("Province")
["Percentage"].mean().sort_values(ascending=False).reset_index()
# Plotting
```

```
plt.figure(figsize=(10, 6))
sns.barplot(x=avg_obese["Province"], y=avg_obese["Percentage"])
plt.title('Average Obesity Rate by Province (Most Recent Year)')
plt.xlabel('Province')
plt.ylabel('Obesity Rate (%)')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
# Save the plot
plt.savefig('average_obesity_rate_by_province.png')
```

Yukon stands as the most obese territory among Canadian provinces, where obesity exceeds 25% of its total population while Newfoundland and Labrador along with Northwest Territories follow closely behind. British Columbia represents the state with the lowest obesity rate since its health data reaches 11% above the preferred level.

5.3.2 Deaths by Cause

This data presents age-specific mortality rate per 100,000 population for both sexes and provincial divisions and yearly. We want only five major fatal conditions: cancer, heart disease, stroke, chronic respiratory diseases, and diabetes.

```
# Clean column names
deaths_by_cause.columns =
deaths_by_cause.columns.str.strip().str.replace(" ",
"_").str.replace("(", "").str.replace(")", "")
# Filter by both sexes and mortality rate
deaths_by_cause_filtered = deaths_by_cause[
    (deaths_by_cause["Sex"] == "Both sexes") &
    (deaths_by_cause["Characteristics"] == "Age-specific mortality
rate per 100,000 population")
]
# Filter for selected causes
selected_causes = [
    "Malignant neoplasms [C00-C97]",
    "Diseases of heart [I00-I09, I11, I13, I20-I51]",
    "Cerebrovascular diseases [I60-I69]",
    "Chronic lower respiratory diseases [J40-J47]",
    "Diabetes mellitus [E10-E14]"
]
deaths_by_cause_filtered = deaths_by_cause_filtered[
    deaths_by_cause_filtered["Leading_causes_of_death_ICD-
10"].isin(selected_causes)
]
# Rename for clarity
deaths_by_cause_filtered = deaths_by_cause_filtered.rename(columns={
    "REF_DATE": "Year",
    "GEO": "Country",
    "Leading_causes_of_death_ICD-10": "Cause_of_Death",
    "VALUE": "Mortality_Rate"
})[["Year", "Country", "Cause_of_Death", "Mortality_Rate"]]
# Pivot to wide format
deaths_by_cause_clean = deaths_by_cause_filtered.pivot_table(
    index=["Country", "Year"],
    columns="Cause_of_Death",
    values="Mortality_Rate",
).reset_index()
# Get most recent year per province
latest_deaths =
deaths_by_cause_clean.sort_values("Year").drop_duplicates("Country",
keep="last").set_index("Country")
# Drop the Year column
latest_deaths = latest_deaths.drop(columns=["Year"])
# Plotting
plt.figure(figsize=(12, 8))
sns.heatmap(latest_deaths, annot=True, cmap="Reds", fmt=".1f")
plt.title("Age-Specific Mortality Rates by Cause")
plt.xlabel("Cause of Death")
plt.ylabel("Country")
plt.tight_layout()
plt.show()
# Save the plot
plt.savefig('mortality_rates_heatmap.png')
```

The chart displays national mortality rates by cause of death. Unsurprisingly, cancer (430.1) and heart disease (425.1) are the leading causes, together accounting for the vast majority of age-specific deaths. Other causes like stroke, chronic respiratory illness, and diabetes show lower mortality but are still significant — especially as they are closely tied to modifiable lifestyle risks such as obesity, smoking, and physical inactivity.

5.3.3 Life Expectancy

The focus here is on birth expectancy between male and female groups across provincial regions with respective reference timeframes.

```
# Clean column names
life_expectancy.columns =
life_expectancy.columns.str.strip().str.replace(" ",
"_").str.replace("(", "").str.replace(")", "")
# Filter for at birth, both sexes, and life expectancy values
life_expectancy_filtered = life_expectancy[
    (life_expectancy["Age_group"] == "At birth") &
    (life_expectancy["Sex"] == "Both sexes") &
    (life_expectancy["Characteristics"] == "Life expectancy") &
    (life_expectancy["GEO"] != "Canada")
]
# Define list of valid Canadian provinces and territories
canadian_provinces = [
    "British Columbia", "Alberta", "Saskatchewan", "Manitoba",
"Ontario", "Quebec",
    "New Brunswick", "Nova Scotia", "Prince Edward Island",
"Newfoundland and Labrador",
    "Yukon", "Northwest Territories", "Nunavut"
]
# Keep only actual provinces (exclude health regions)
life_expectancy_provinces =
life_expectancy_filtered[life_expectancy_filtered["GEO"].isin(canadian
_provinces)]
# Keep latest entry per province
latest_life_expectancy =
life_expectancy_provinces.sort_values("REF_DATE").drop_duplicates("GEO
", keep="last")
# Rename columns
latest_life_expectancy = latest_life_expectancy.rename(columns={
    "REF_DATE": "Year",
    "GEO": "Province",
    "VALUE": "Life_Expectancy"
})[["Year", "Province", "Life_Expectancy"]]
# Plotting
plt.figure(figsize=(10,6))
sns.barplot(data=latest_life_expectancy.sort_values("Life_Expectancy")
,
            x="Life_Expectancy", y="Province")
plt.title("Life Expectancy at Birth by Province")
plt.xlabel("Life Expectancy (Years)")
plt.ylabel("Province")
plt.tight_layout()
plt.show()
# Save the plot
plt.savefig('life_expectancy_by_province.png')
```

The province of British Columbia presents the highest life expectancy rates compared to Nunavut, Yukon and the Northwest Territories.

5.3.4 Perceived Health

Here we want the percentage of individuals who rate their health status as “very good or excellent” while providing sex and province-based data.

```
# Clean the column names for better usability
perceived_health.columns =
perceived_health.columns.str.strip().str.replace(" ",
"_").str.replace("(", "").str.replace(")", "")
# Filter for oth sexes and precent format
perceived_health_filtered = perceived_health[
    (perceived_health["Sex"] == "Both sexes") &
    (perceived_health["Characteristics"] == "Percent") &
    (perceived_health["Indicators"] == "Perceived health, very good or
excellent")
]
# Exclude Canada as a whole
perceived_health_filtered =
perceived_health_filtered[perceived_health_filtered["GEO"] != "Canada
(excluding territories)"]
# Rename columns for clarity
perceived_health_filtered = perceived_health_filtered.rename(columns={
    "REF_DATE": "Year",
    "GEO": "Province",
    "VALUE": "Percentage"
})
# Keep only the most recent year per province
latest_perceived_health =
perceived_health_filtered.sort_values("Year").drop_duplicates("Provinc
e", keep="last")
# Plotting
plt.figure(figsize=(10,6))
sns.barplot(data=latest_perceived_health.sort_values("Percentage"),
x="Percentage", y="Province", palette="Greens")
plt.title("Self-Reported 'Very Good or Excellent' Health by Province")
plt.xlabel("Percentage of Respondents")
plt.ylabel("Province")
plt.tight_layout()
plt.show()
# Save the plot
plt.savefig('perceived_health_by_province.png')
```

Ontario demonstrates the highest rate of resident health with sixty percent who consider themselves very healthy among other provinces which include Saskatchewan and Quebec. The people of Nova Scotia along with Prince Edward Island show the most minimal levels of perceived health according to investigation results.

5.4 Correlation and Insights

Here we investigate he interactions between obesity rates, perceived helth and life expectancy by merging cleaned data. A merged data table and calculation of Pearson correlation coefficients enable us to measure the power of connections between the fundamental health-related variables.

We are looking at:

- Life Expectancy
- Obesity Rate
- Perceived Health (percentage reporting "very good or excellent" health)

```
# Rename avg_obese column for clarity
avg_obese = avg_obese.rename(columns={"Percentage": "Obesity_Rate"})

# Merge datasets on Province
merged = latest_life_expectancy.merge(
    latest_perceived_health[["Province",
"Percentage"]].rename(columns={"Percentage": "Perceived_Health"}),
    on="Province"
).merge(
    avg_obese,
    on="Province"
)
# Correlation matrix
correlation_matrix = merged[["Life_Expectancy", "Perceived_Health",
"Obesity_Rate"]].corr()
# Plotting
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm",
fmt=".2f")
plt.title("Correlation Matrix: Health Indicators by Province")
plt.tight_layout()
plt.show()
# Save the plot
plt.savefig('correlation_matrix_health_indicators.png')
```

As shown:

This analysis demonstrates that life expectancy is directly proportional to obesity rate (–0.83).

Self-reported health does not create a strong connection to life expectancy based on the present data (–0.06).

The perception of health among residents decreases slightly as obesity rates increases throughout the provinces (–0.09).

6. Conclusion and Discussion

This project was conducted to establish relationships between geographic positions and lifestyle influences on Canadian health results using open data resources. Our analysis concentrated on obesity statistics alongside perceived health data and life expectancy figures which we supported with national mortality records.

The analysis demonstrated that obesity rate shows the strongest negative correlation against life expectancy (–0.83) which establishes it as the crucial determinant in this study.

The relationship between perceived health assessment and life expectancy duration remained weak because subjective measures might fail to represent actual health risks.

Cancer and heart disease are the most deathly and are linked to lifestyle factors such as obesity.

Trends:

British Columbia demonstrated low obesity rates and was one of the provinces with the highest life expectancy.

The eastern Canadian provinces demonstrated both higher rates of obesity along with reduced life expectancy rates.

Limitations:

Our analysis had restricted value because it used different time periods across datasets and the research lacked specific death reasons at the provincial level. Future research needs to include the analysis of smoking behavior together with economic status and healthcare access as this information would enhance the existing studies.

Conclusion:

Population health status depends remarkably on factors that people can control such as obesity rates. The analysis would benefit public health planning through additional evaluation of lifestyle and structural variables.

References

- Statistics Canada. *Life expectancy and death statistics*. Retrieved from https://www.statcan.gc.ca/en/subjects-start/health/life_expectancy_and_deaths

- Government of Canada. (2022). What makes Canadians healthy or unhealthy? Public Health Agency of Canada. https://www.canada.ca/en/public-health/services/health-promotion/population-health/what-determines-health/what-makes-canadians-healthy-unhealthy.html

- Statistics Canada. *Body mass index and health*. Retrieved from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310009601

- Statistics Canada. *Perceived health, by province and sex*. Retrieved from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310009602

- Statistics Canada. *Leading causes of death, total population, by age group*. Retrieved from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310039401