DATA 605 - Analyzing Healthcare Wait Times Across Canadian Provinces (2013–2023)

Joshua Ogunbo

SECTION A - INTRODUCTION:

Canada's publicly funded healthcare system is founded on universal access, yet wait times for medical procedures remain a critical issue affecting patient outcomes and satisfaction. Delays in receiving care, especially for procedures like hip replacements, cataract surgeries, MRIs, and CT scans, can reduce quality of life and strain the system's capacity.

Timely access to healthcare remains a pressing concern in Canada, and the main goal of this project is to explore how wait times have evolved across Canadian provinces over the past decade.

As provinces work to improve efficiency and equity within the system, analyzing past wait time trends can provide critical insights to guide future decisions and policy reform.

To guide the analysis, the project will pursue the following objectives:

1. Analyze the trend of average wait times for medical procedures in all Canadian provinces between 2013 and 2023.

2. Assess the impact of the COVID-19 pandemic (2020–2021) on average national wait times for medical procedures.

3. Identify the most frequently performed medical procedures across all Canadian provinces between 2013 and 2023.

4. Visualize the top three provinces with the most improvement and most decline in wait time performance across all measured procedures within the time frame.

SECTION B - ANALYSIS QUESTIONS:

The following study questions are designed to guide the exploratory data analysis and provide insight into the project objectives:

1. How have average wait times for medical procedures changed over time in all Canadian provinces between 2013 and 2023?

2. What was the effect of the COVID-19 pandemic (2020–2021) on national average wait times for these procedures?

3. Which procedures saw the highest number of cases performed across all provinces between 2013 and 2023?

4. Which top 3 provinces showed the most significant improvement in wait time performance between 2013 and 2023?

5. Which provinces experienced the greatest decline in wait time performance across all measured procedures during the same period?

SECTION C - DATA SOURCING AND JUSTIFICATION:

Dataset:

- **Title:** Wait times for priority procedures in Canada, 2024
- **Source:** Canadian Institute for Health Information (CIHI)
- **Link:** Data (https://www.cihi.ca/en/wait-times-for-priority-procedures-in-canada-2024)

1. Reason for Downloading this Dataset

This dataset was chosen because it is aligned with the project's objective, which is Healthcare Wait Times Across Canadian Provinces for key medical procedures. It is publicly available, well-documented, and originates from CIHI, a reputable federal health organization in Canada.

2. How the Dataset Supports the Project Objective

The dataset provides historical data on the average wait times and the number of procedures performed per year, broken down by province and procedure

3. Link to Dataset

https://www.cihi.ca/en/wait-times-for-priority-procedures-in-canada-2024

4. Use of Multiple Sources

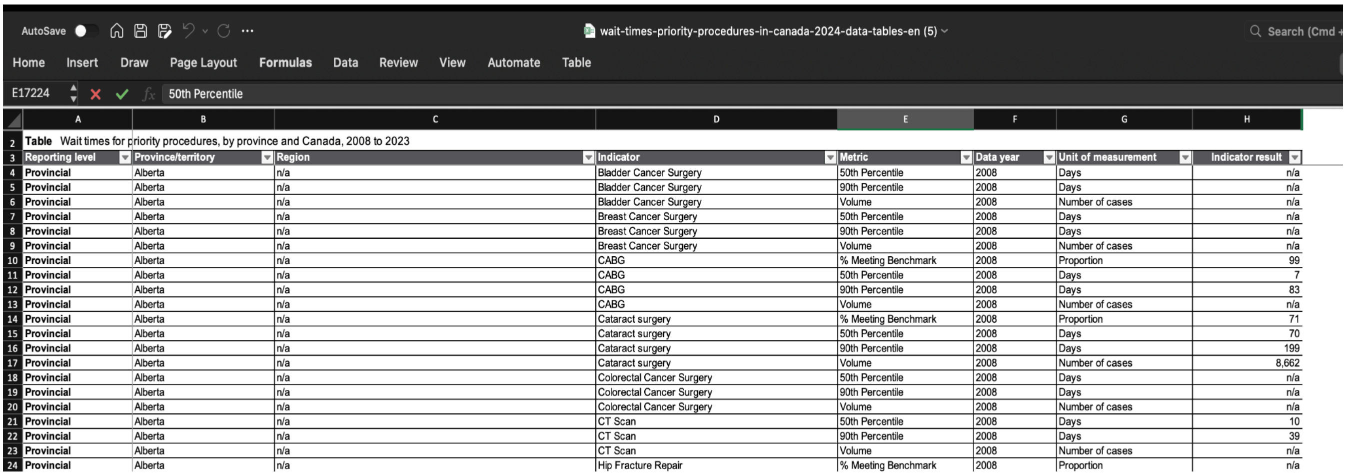
The selected dataset from CIHI is sufficient for this analysis, as it comprehensively supports the project objectives and addresses all the identified analysis questions.

SECTION D - DATA CLEANING

The dataset has a title in the worksheet and the following columns:

Reporting level, Province/territory, Region, Indicator, Metric, Data year, Unit of measurement, Indicator result, Unnamed: 8, Unnamed: 9, Unnamed: 10, and Unnamed: 11

SCREENSHOT OF ORIGINAL WORKSHEET



wait-times-priority-procedures-in-canada-2024-data-tables-en (5)

E17224 | 50th Percentile

Table Wait times for priority procedures, by province and Canada, 2008 to 2023

	Reporting level	Province/territory	Region	Indicator	Metric	Data year	Unit of measurement	Indicator result
4	Provincial	Alberta	n/a	Bladder Cancer Surgery	50th Percentile	2008	Days	n/a
5	Provincial	Alberta	n/a	Bladder Cancer Surgery	90th Percentile	2008	Days	n/a
6	Provincial	Alberta	n/a	Bladder Cancer Surgery	Volume	2008	Number of cases	n/a
7	Provincial	Alberta	n/a	Breast Cancer Surgery	50th Percentile	2008	Days	n/a
8	Provincial	Alberta	n/a	Breast Cancer Surgery	90th Percentile	2008	Days	n/a
9	Provincial	Alberta	n/a	Breast Cancer Surgery	Volume	2008	Number of cases	n/a
10	Provincial	Alberta	n/a	CABG	% Meeting Benchmark	2008	Proportion	99
11	Provincial	Alberta	n/a	CABG	50th Percentile	2008	Days	7
12	Provincial	Alberta	n/a	CABG	90th Percentile	2008	Days	83
13	Provincial	Alberta	n/a	CABG	Volume	2008	Number of cases	n/a
14	Provincial	Alberta	n/a	Cataract surgery	% Meeting Benchmark	2008	Proportion	71
15	Provincial	Alberta	n/a	Cataract surgery	50th Percentile	2008	Days	70
16	Provincial	Alberta	n/a	Cataract surgery	90th Percentile	2008	Days	199
17	Provincial	Alberta	n/a	Cataract surgery	Volume	2008	Number of cases	8,662
18	Provincial	Alberta	n/a	Colorectal Cancer Surgery	50th Percentile	2008	Days	n/a
19	Provincial	Alberta	n/a	Colorectal Cancer Surgery	90th Percentile	2008	Days	n/a
20	Provincial	Alberta	n/a	Colorectal Cancer Surgery	Volume	2008	Number of cases	n/a
21	Provincial	Alberta	n/a	CT Scan	50th Percentile	2008	Days	10
22	Provincial	Alberta	n/a	CT Scan	90th Percentile	2008	Days	39
23	Provincial	Alberta	n/a	CT Scan	Volume	2008	Number of cases	n/a
24	Provincial	Alberta	n/a	Hip Fracture Repair	% Meeting Benchmark	2008	Proportion	n/a

The relevant worksheet was loaded using pandas in Jupyter, and the following cleaning processes were performed:

```python
import pandas as pd

df = pd.read_excel("WaitTime.xlsx", sheet_name="Wait times 2008 to 2023", header=2)
df.head()
```

[129] ✓ 0.8s Python

	Reporting level	Province/territory	Region	Indicator	Metric	Data year	Unit of measurement	Indicator result	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11
0	Provincial	Alberta	NaN	Bladder Cancer Surgery	50th Percentile	2008	Days	NaN	NaN	NaN	NaN	NaN
1	Provincial	Alberta	NaN	Bladder Cancer Surgery	90th Percentile	2008	Days	NaN	NaN	NaN	NaN	NaN
2	Provincial	Alberta	NaN	Bladder Cancer Surgery	Volume	2008	Number of cases	NaN	NaN	NaN	NaN	NaN
3	Provincial	Alberta	NaN	Breast Cancer Surgery	50th Percentile	2008	Days	NaN	NaN	NaN	NaN	NaN
4	Provincial	Alberta	NaN	Breast Cancer Surgery	90th Percentile	2008	Days	NaN	NaN	NaN	NaN	NaN

Attribute: Unit of measurement
Issue:

This column had inconsistent values, such as "days" and "hours" for the 50th Percentile metric.

Why did it need cleaning?

Inconsistent time units made comparisons across provinces and procedures invalid. For example, 1 day vs 24 hours could appear as a significant difference if not converted properly.

How cleaning helped:

Standardizing all values to **hours** ensured that wait times were comparable across provinces and years.

Before Cleaning:

```
df_50th = df[df["Metric"] == "50th Percentile"]

unique_metrics_units = (
    df_50th[["Metric", "Unit of measurement"]]
    .dropna()
    .drop_duplicates()
    .reset_index(drop=True)
)

unique_metrics_units.head()
```

✓ 0.0s

	Metric	Unit of measurement
0	50th Percentile	Days
1	50th Percentile	Hours

Cleaning Technique:

All Indicator result values where the Unit of measurement is "Days" were converted to hours (value × 24), and the unit was converted to “Hours”.

```
def standardize_unit_of_measurement(row):
    if (
        row["Metric"] == "50th Percentile"
        and str(row["Unit of measurement"]).lower() == "days"
    ):
        return row["Indicator result"] * 24
    return row["Indicator result"]

def get_unit_of_measurement(row):
    if (
        row["Metric"] == "50th Percentile"
        and str(row["Unit of measurement"]).lower() == "days"
    ):
        return "Hours"
    return row["Unit of measurement"]

df_filtered["Indicator result"] = df_filtered.apply(
    standardize_unit_of_measurement, axis=1
)

df_filtered["Unit of measurement"] = df_filtered.apply(get_unit_of_measurement, axis=1)
```

Why It's Now Clean:

Every row now reports wait time in **the same unit (hours)**.

After Cleaning:

```python
df_50th = df_filtered[df_filtered["Metric"] == "50th Percentile"]

unique_metrics_units = (
    df_50th[["Metric", "Unit of measurement"]]
    .dropna()
    .drop_duplicates()
    .reset_index(drop=True)
)

unique_metrics_units.head()
```

1] ✓ 0.0s

	Metric	Unit of measurement
0	50th Percentile	Hours

Attribute: Indicator result

Issue:

Many rows had missing (NaN) values in the 50th Percentile and Volume metrics.

Why did it need cleaning?

- Missing 50th Percentile values would break trend analysis, COVID comparisons, and performance shifts.
- Missing Volume values would prevent answering Q3 (the relationship between procedure count and wait times).

How cleaning helped:

Enabled accurate, bias-free trendlines and statistical summaries. Also, preserved natural distribution without inflating values

Before Cleaning:

```
df_indicator_result = df[
    ["Province/territory", "Metric", "Indicator", "Indicator result"]
]

df_indicator_result.head()
```

✓ 0.0s

	Province/territory	Metric	Indicator	Indicator result
0	Alberta	50th Percentile	Bladder Cancer Surgery	NaN
1	Alberta	90th Percentile	Bladder Cancer Surgery	NaN
2	Alberta	Volume	Bladder Cancer Surgery	NaN
3	Alberta	50th Percentile	Breast Cancer Surgery	NaN
4	Alberta	90th Percentile	Breast Cancer Surgery	NaN

Cleaning Technique:

- Used **group-wise median imputation** by Province/territory, Indicator, and Metric.
- Applied **fallback imputation** using broader Indicator + Metric groups.
- For fully missing Volume groups, filled with 0 (representing no procedures).
- For unfixable missing 50th Percentile values, dropped the corresponding (Province, Indicator) group entirely.

```python
df_filtered["Indicator result"] = df_filtered.groupby(
    ["Province/territory", "Indicator", "Metric"]
)["Indicator result"].transform(lambda x: x.fillna(x.median()))

volume_all_null_groups = (
    df_filtered[df_filtered["Metric"] == "Volume"]
    .groupby(["Province/territory", "Indicator", "Metric"])["Indicator result"]
    .apply(lambda x: x.isna().all())
)

volume_fill_groups = volume_all_null_groups[volume_all_null_groups].index

df_filtered["Indicator result"] = df_filtered.apply(
    lambda row: 0
    if (
        row["Metric"] == "Volume"
        and (row["Province/territory"], row["Indicator"], row["Metric"])
        in volume_fill_groups
    )
    else row["Indicator result"],
    axis=1,
)


df_filtered["Indicator result"] = df_filtered.groupby(["Indicator", "Metric"])[
    "Indicator result"
].transform(lambda x: x.fillna(x.median()))


unfilled_50th = df_filtered[
    (df_filtered["Metric"] == "50th Percentile")
    & df_filtered["Indicator result"].isna()
][["Province/territory", "Indicator"]].drop_duplicates()


drop_keys = set([tuple(x) for x in unfilled_50th.values])

df_filtered = df_filtered[
    ~df_filtered[["Province/territory", "Indicator"]]
    .apply(tuple, axis=1)
    .isin(drop_keys)
].reset_index(drop=True)
```

Why It's Now Clean:

Every row has a valid, interpretable value. There are no NaNs, and values are contextually accurate due to smart group-based imputation.

After Cleaning:

```python
df_indicator_result = df_filtered[
    ["Province/territory", "Metric", "Indicator", "Indicator result"]
]

df_indicator_result.head()
```

✓ 0.0s

	Province/territory	Metric	Indicator	Indicator result
0	Alberta	50th Percentile	Bladder Cancer Surgery	672.0
1	Alberta	Volume	Bladder Cancer Surgery	477.0
2	Alberta	50th Percentile	Breast Cancer Surgery	360.0
3	Alberta	Volume	Breast Cancer Surgery	1169.0
4	Alberta	50th Percentile	CABG	288.0

Dirty Attribute: **Region and "Unnamed" Columns**

Issue:

Our analysis is based on the provincial level, so having Region in our working dataset is of no value, also, we have useless unnamed columns

Why did it need cleaning?

Not useful for our analysis, and it’s safe to be removed

How cleaning helped:

Kept the working data concise without ambiguity

Before Cleaning:

```python
import pandas as pd

df = pd.read_excel("WaitTime.xlsx", sheet_name="Wait times 2008 to 2023", header=2)
df.head()
```

[129] ✓ 0.8s Python

	Reporting level	Province/territory	Region	Indicator	Metric	Data year	Unit of measurement	Indicator result	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11
0	Provincial	Alberta	NaN	Bladder Cancer Surgery	50th Percentile	2008	Days	NaN	NaN	NaN	NaN	NaN
1	Provincial	Alberta	NaN	Bladder Cancer Surgery	90th Percentile	2008	Days	NaN	NaN	NaN	NaN	NaN
2	Provincial	Alberta	NaN	Bladder Cancer Surgery	Volume	2008	Number of cases	NaN	NaN	NaN	NaN	NaN
3	Provincial	Alberta	NaN	Breast Cancer Surgery	50th Percentile	2008	Days	NaN	NaN	NaN	NaN	NaN
4	Provincial	Alberta	NaN	Breast Cancer Surgery	90th Percentile	2008	Days	NaN	NaN	NaN	NaN	NaN

Cleaning Technique:

Dropped the region column

```python
df = df.loc[:, ~df.columns.str.contains("^Unnamed")]
df = df.drop(columns=["Region"], errors="ignore")
```

6] ✓ 0.0s

Why It's Now Clean:

Only Province/territory remains as the geographic identifier, and no unnamed columns

After Cleaning:

```
df_filtered.head()
```

✓ 0.0s

	Reporting level	Province/territory	Indicator	Metric	Data year	Unit of measurement	Indicator result
0	Provincial	Alberta	Bladder Cancer Surgery	50th Percentile	2013	Hours	672.0
1	Provincial	Alberta	Bladder Cancer Surgery	Volume	2013	Number of cases	477.0
2	Provincial	Alberta	Breast Cancer Surgery	50th Percentile	2013	Hours	360.0
3	Provincial	Alberta	Breast Cancer Surgery	Volume	2013	Number of cases	1169.0
4	Provincial	Alberta	CABG	50th Percentile	2013	Hours	288.0

Dirty Attribute: Reporting level

Issue:

Included values like "Provincial" and "Regional," which were irrelevant to your provincial-level study.

Why did it need cleaning?

Our analysis is based on the provincial level, so having other reporting levels in our dataset is of no value

How cleaning helped:

Kept the working data concise without ambiguity

Before Cleaning:

```python
df_reporting_level = df["Reporting level"]

unique_reporting_levels = (
    df_reporting_level.dropna().drop_duplicates().reset_index(drop=True)
)

unique_reporting_levels.head()
```

✓ 0.0s

```
0                                          Provincial
1                                            National
2                                            Regional
3                                               Notes
4   n/a in the Region column means that the data r...
Name: Reporting level, dtype: object
```

Cleaning Technique:

Filtered the dataset to include only the provincial reporting level

```
df["Data year"] = pd.to_numeric(df["Data year"], errors="coerce")
df_filtered = df[(df["Data year"] >= 2013) & (df["Data year"] <= 2023)]

df_filtered = df_filtered[df_filtered["Reporting level"] == "Provincial"]
```

[189] ✓ 0.0s

Why It's Now Clean:

Every row now reflects province-level data only.

After Cleaning:

```
df_reporting_level = df_filtered["Reporting level"]

unique_reporting_levels = (
    df_reporting_level.dropna().drop_duplicates().reset_index(drop=True)
)

unique_reporting_levels.head()
```

04] ✓ 0.0s

```
0    Provincial
Name: Reporting level, dtype: object
```

Attribute: Data Year

Issue:

Our analysis is only interested in the data from 2013 to 2023

Why did it need cleaning?

Our dataset includes data from outside of our focus period

How cleaning helped:

Kept the working data concise and focused on the relevant year

Before Cleaning:

```python
df_reporting_level = df["Data year"].sort_values(ascending=True)

unique_reporting_levels = (
    df_reporting_level.dropna().drop_duplicates().reset_index(drop=True)
)

unique_reporting_levels.head()
```

[206] ✓ 0.0s

```
0    2008.0
1    2009.0
2    2010.0
3    2011.0
4    2012.0
Name: Data year, dtype: float64
```

Cleaning Technique:

Filtered the dataset to include only the provincial reporting level

```
df["Data year"] = pd.to_numeric(df["Data year"], errors="coerce")
df_filtered = df[(df["Data year"] >= 2013) & (df["Data year"] <= 2023)]

df_filtered = df_filtered[df_filtered["Reporting level"] == "Provincial"]
```

```
[189]  ✓ 0.0s
```

Why It's Now Clean:

Data before and after our focus period are no longer present in the dataset.

After Cleaning:

```
df_reporting_level = df_filtered["Data year"].sort_values(ascending=True)

unique_reporting_levels = (
    df_reporting_level.dropna().drop_duplicates().reset_index(drop=True)
)

unique_reporting_levels.head()
```

```
207]  ✓ 0.0s
```

```
...  0    2013
     1    2014
     2    2015
     3    2016
     4    2017
     Name: Data year, dtype: int64
```

Attribute: Metric (90th Percentile & % Meeting Benchmark)

Issue:

The dataset included multiple types of metrics under the Metric column, which are not relevant to our analysis: “90th Percentile” and “% Meeting Benchmark.”

Why did it need cleaning?

Our analysis needs only the 50th percentile and volume metric types

How cleaning helped:

Kept the working data concise and focused on the relevant metric types

Before Cleaning:

```
unique_metrics = df["Metric"].dropna().unique().tolist()
display(unique_metrics)
```

```
0.0s
```

```
['50th Percentile',
 '90th Percentile',
 'Volume',
 '% Meeting Benchmark',
 '50th percentile',
 '90th percentile',
 '% meeting benchmark']
```

Cleaning Technique:

Filtered the dataset to include only metrics that are needed for our analysis

```
df_filtered = df_filtered[df_filtered["Metric"] != "90th Percentile"]
df_filtered = df_filtered[df_filtered["Metric"] != "% Meeting Benchmark"]
df_filtered.head()
```

190] ✓ 0.0s

Why It's Now Clean:

The dataset contains only the 50th Percentile and Volume metrics

After Cleaning:

```
unique_metrics = df_filtered["Metric"].dropna().unique().tolist()
display(unique_metrics)
```

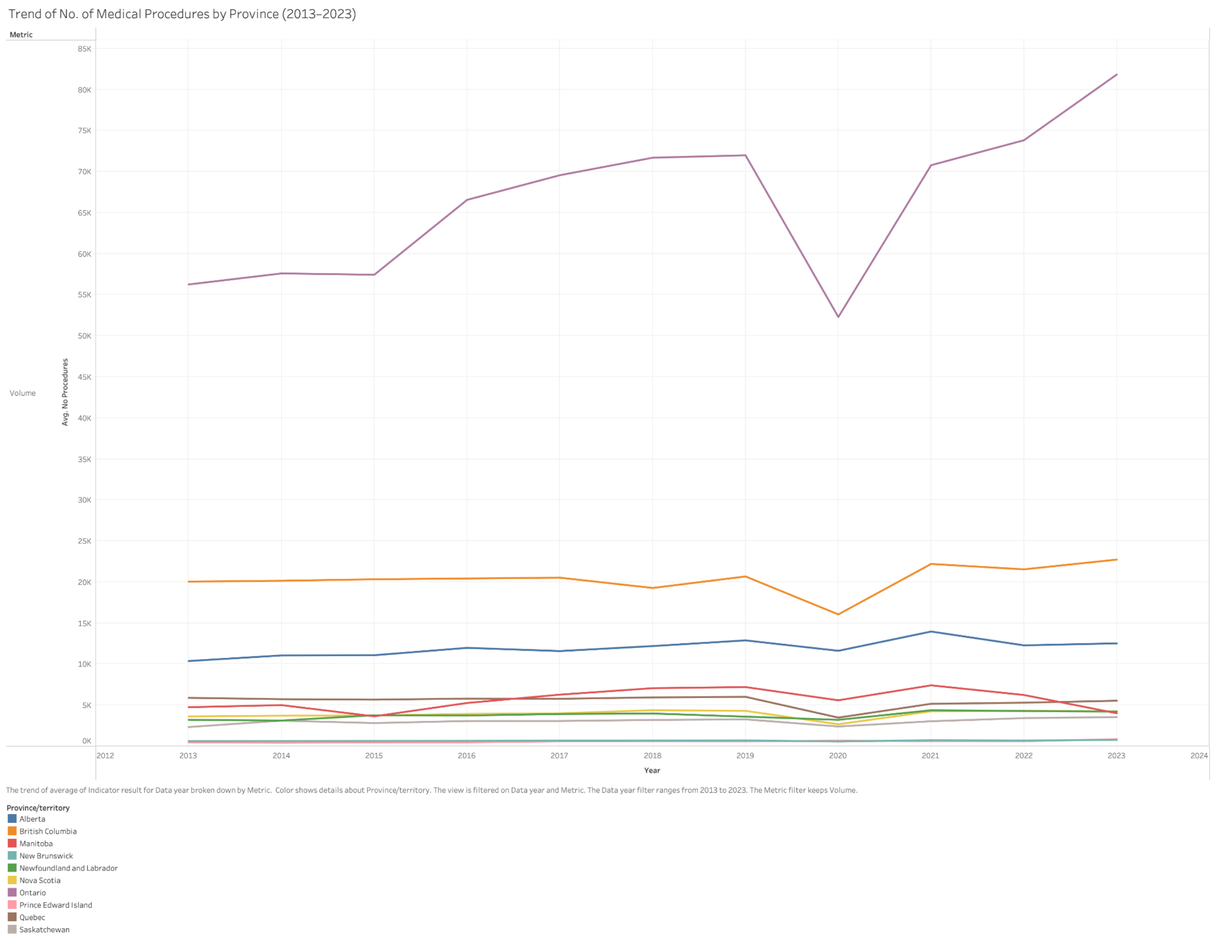
11] ✓ 0.0s

```
['50th Percentile', 'Volume']
```

SECTION E - VISUALIZATIONS AND FINDINGS

Question 1a:

Chart Used: Line Chart



Why This Chart Was Chosen:

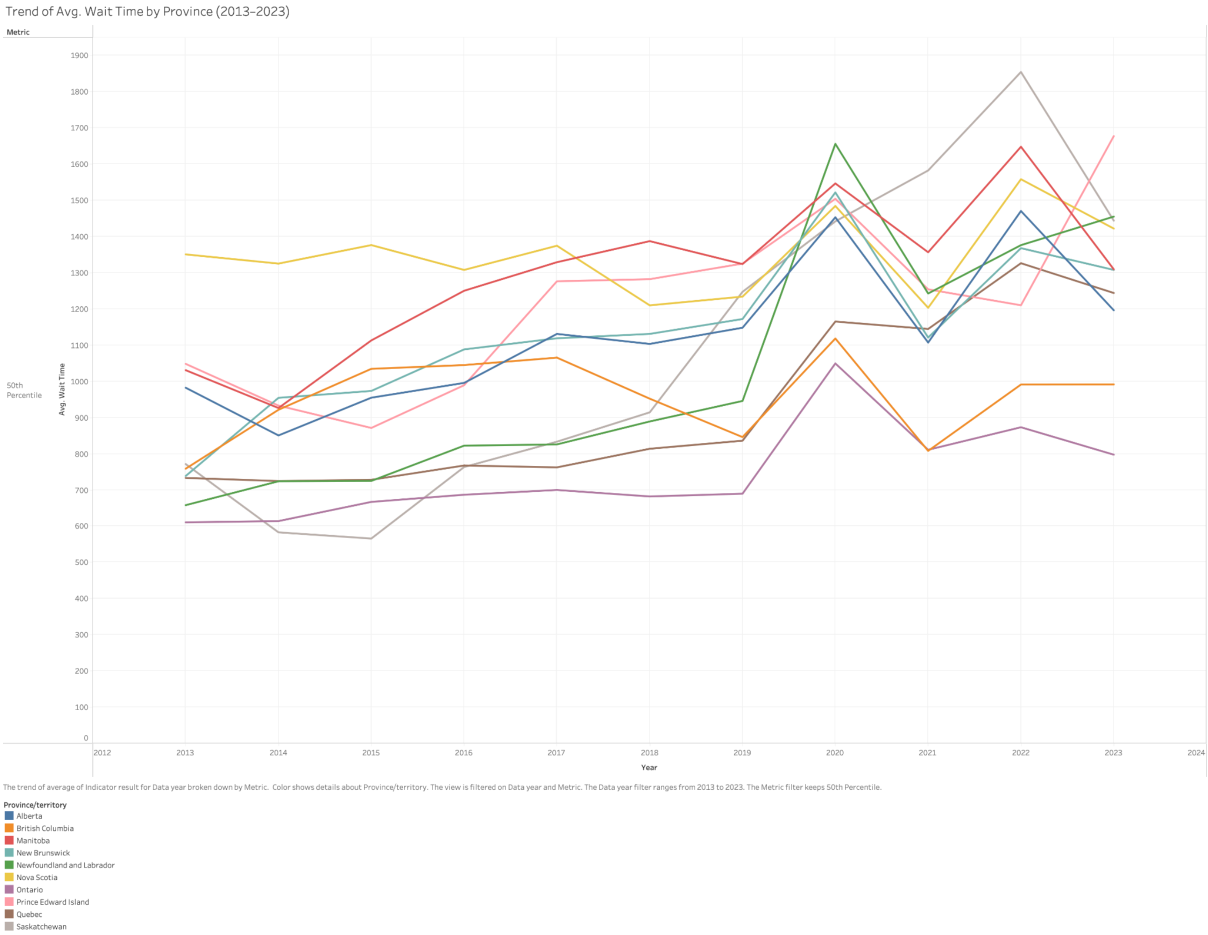
A line chart is the best choice for analyzing time series data across multiple categories. It allows us to observe long-term trends in procedural volume and compare the trajectory of each province's healthcare activity over the past decade.

Discussion of Findings:

The chart shows that Ontario consistently had the highest number of medical procedures each year. There was a clear upward trend in procedure volume from 2016 to 2019, followed by a sharp decline in 2020, which reflects the disruption caused by COVID-19. After that drop, Ontario's numbers began to recover quickly, with steady growth through 2021 and continuing into 2023.

Other provinces, including British Columbia, Alberta, and Quebec, maintained relatively stable volumes throughout most of the period. Like Ontario, they experienced a noticeable drop in 2020, but the impact varied in how quickly they bounced back. In contrast, smaller provinces such as Prince Edward Island, Newfoundland and Labrador, and New Brunswick recorded lower volumes overall, which is expected due to their smaller populations. However, their recovery patterns after 2020 were more varied and less consistent

Question 1b:
Chart Used: Line Chart



Why This Chart Was Chosen:

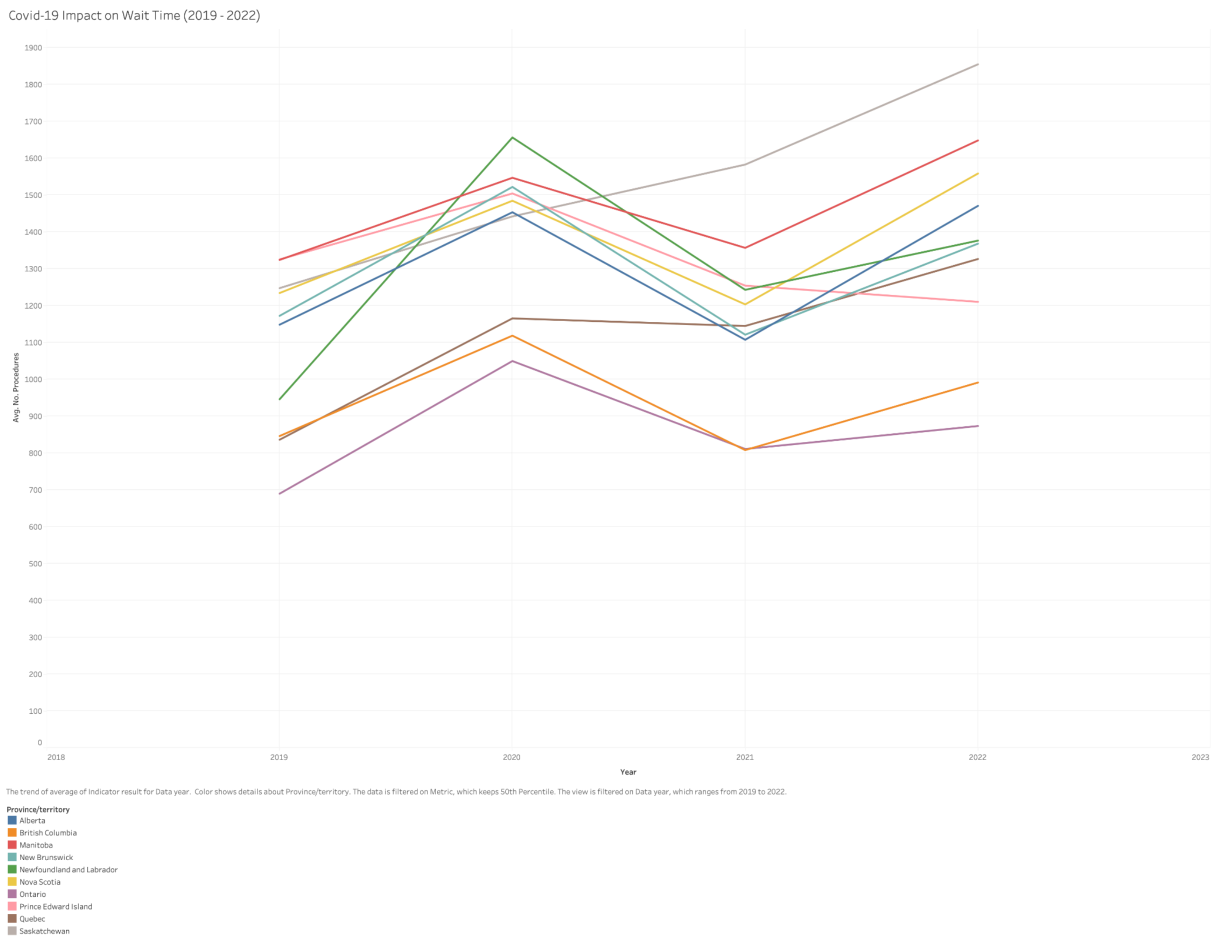
A line chart is the best choice for analyzing time series data across multiple categories. It allows us to observe long-term trends in average wait time and compare the trajectory of each province's healthcare activity over the past decade.

Discussion of Findings:
Most provinces saw gradual increases in wait times from 2013 through 2019, followed by a dramatic spike in 2020. This sharp rise reflects the COVID-19 pandemic's disruption to regular healthcare services.

After 2020, the chart shows some provinces recovering more quickly than others. For example, Nova Scotia and Prince Edward Island show a noticeable drop in wait times between 2020 and 2021, while others like Saskatchewan and Manitoba remained elevated well into 2022. Newfoundland and Labrador appears to have had one of the highest peaks in 2021, suggesting a severe pandemic impact on access to care.

Question 2:

Chart Used: Line Chart (per province, 2019–2022)



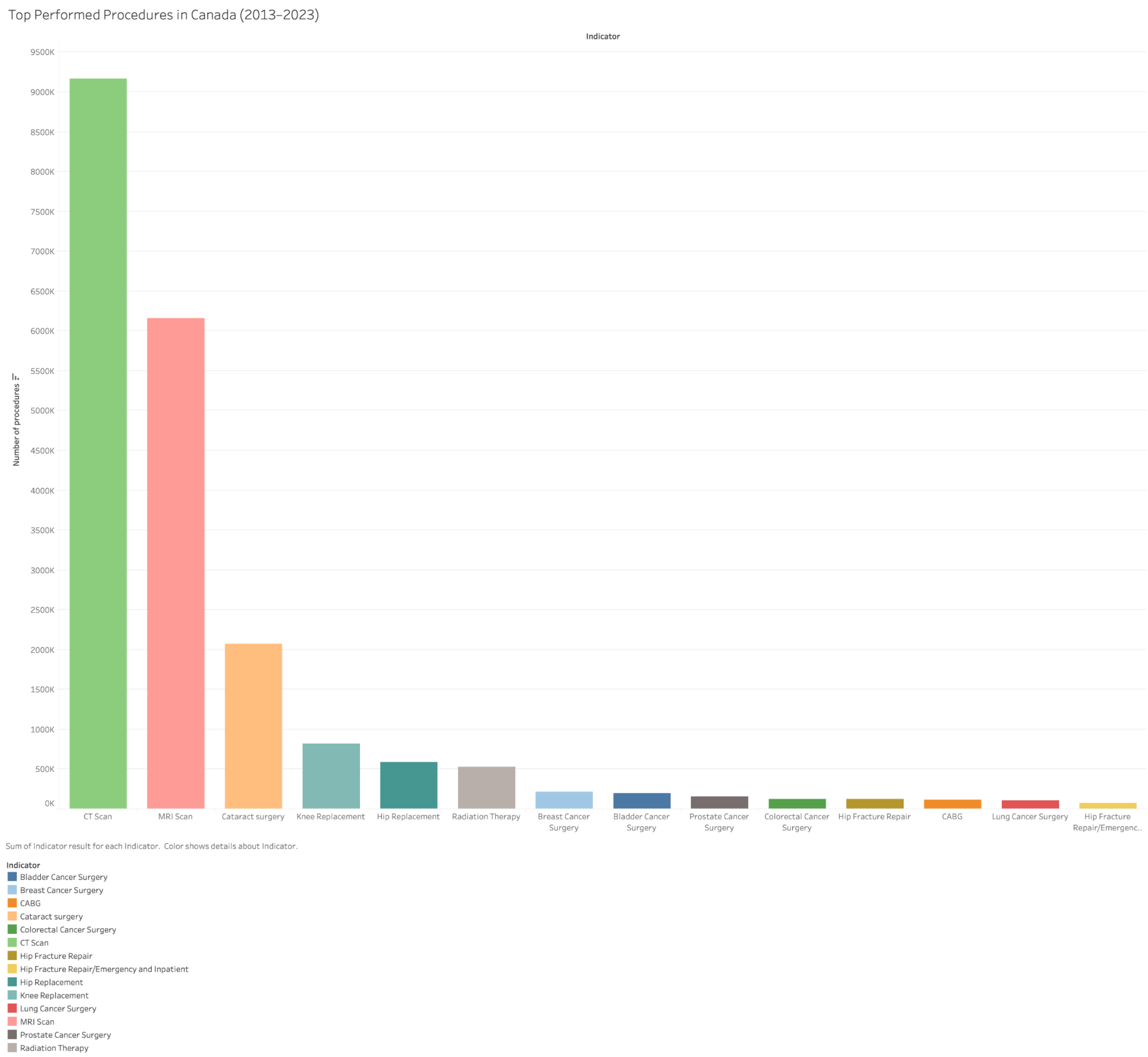
Why This Chart Was Chosen:

A line chart is effective for identifying patterns across sequential periods. In this context, it enables us to observe whether the pandemic years (2020–2021) led to measurable changes in healthcare wait times and how different provinces responded.

Discussion of Findings:
The visualization reveals that most provinces experienced a sharp rise in wait times in 2020, coinciding with the initial outbreak of COVID-19. This reflects the widespread deferral of elective and non-urgent procedures to prioritize emergency and COVID-related care. For instance, provinces such as Saskatchewan and Newfoundland and Labrador showed particularly steep increases, indicating a potentially higher level of healthcare system strain or disruption.

By 2021, there were early signs of stabilization or recovery in some provinces. However, others continued to experience elevated wait times, suggesting that service backlogs or operational challenges persisted well beyond the initial shock of the pandemic.

Question 3:
Chart Used: Bar Chart



Why This Chart Was Chosen:

A bar chart was selected because it offers a clear visual ranking of procedures based on their total volume and also maintains clarity and readability even with a wide range of procedure types.

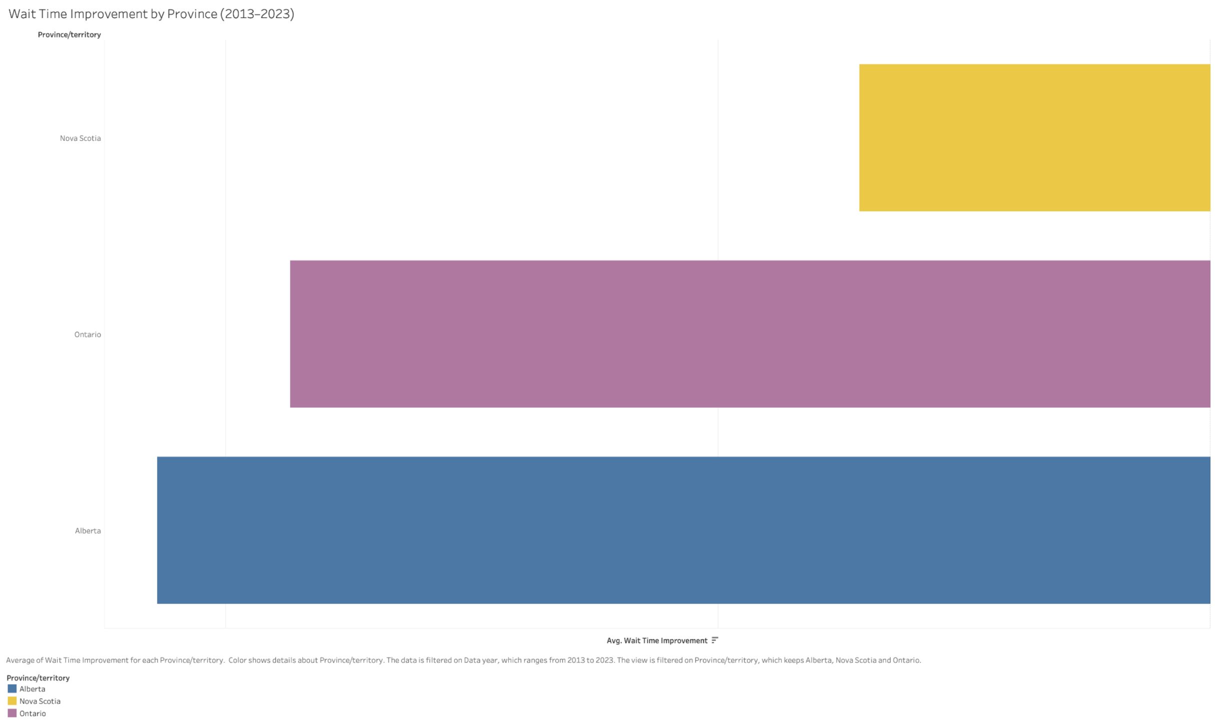
Discussion of Findings:

The chart shows that CT scans were the most commonly performed medical procedure in Canada over the last decade, with more than 9 million reported cases. This is followed by MRI scans, which exceeded 6 million cases. Cataract surgery came third, with just over 2 million procedures, reflecting the high demand for vision-related interventions, particularly among aging populations.

The chart also shows a steep drop-off in volume after the top three, with procedures like knee replacement, hip replacement, and radiation therapy occurring at much lower frequencies but still representing significant healthcare activities

Cancer-related surgeries, such as breast, prostate, and colorectal cancer surgeries, also appeared in the distribution, though at considerably lower volumes.

Question 4:
Chart Used: Horizontal Bar Chart (Top 3 Filtered)



Why This Chart Was Chosen:

A horizontal bar chart was selected because it makes it easy to compare the top performers side by side. The bars clearly show which provinces made the most progress in reducing delays

Discussion of Findings:

According to the visualization, Nova Scotia achieved the greatest overall improvement in wait time performance between 2013 and 2023. This suggests that the province made notable strides in reducing delays across a variety of medical procedures.

Ontario and Alberta also showed substantial improvements, though not as pronounced as Nova Scotia's. Their performance still reflects meaningful progress in addressing wait time challenges.

Question 5:

Chart Used: Horizontal Bar Chart (Top 3 Filtered)

Why This Chart Was Chosen:

The horizontal bar chart effectively showcases negative performance outcomes. By sorting provinces from worst to least severe declines and filtering for only the bottom three, the chart draws attention to the most concerning cases.

Discussion of Findings:

According to the chart, Newfoundland and Labrador experienced the greatest decline in wait time performance, with the largest average increase in median wait times between 2013 and 2023. Saskatchewan also saw a substantial decline, indicating persistent or worsening delays in accessing care. Prince Edward Island, while showing the least decline among the three, still experienced a meaningful increase in average wait times over the decade.

References:

Canadian Institute for Health Information. (2024). Wait times for priority procedures in Canada, 2024. Ottawa, ON: CIHI. https://www.cihi.ca/en/wait-times-for-priority-procedures-in-canada-2024 [Accessed: May 2025]

Ogunbo, J. (2025). Healthcare Wait Time Analysis Canada. GitHub repository. https://github.com/frontdevguy/healthcare-wait-time-analysis-canada