Ontario Covid_19 Cases Analysis

LANGXIN LI, 1008831374; Derrick Li,1009049959; Yiran Yu 1008838180

2025-03-21

1 Dataset and Variable Descriptions:

1.1 Background of Research and Data

The COVID–19 pandemic impacted public health systems worldwide. Ontario, Canada was no exception. In response to rising infections, the Government of Ontario released open-access datasets to track daily confirmed cases, vaccination rates, and other key metrics from 2021 through 2024. These data fueled research on the pandemic's spread, informed public health interventions, and guided future preparedness efforts.

This report analyzes COVID-19 case data, vaccination uptake, and demographic patterns across Ontario to identify key trends and assess the impact of public health measures at both provincial and municipal levels.

1.2 Dataset Sources

The main data come from **"Confirmed positive cases of COVID-19 in Ontario,"** published on the Ontario Data Catalogue. These raw files—`Cases2021.csv`, `Cases2022.csv`, `Cases2023.csv`, and `Cases2024.csv`—have been merged into a single dataset for analysis.

Additional datasets support various sections of the report. The **"City Analysis of COVID-19 Cases"** section uses `Ontario_Cities_Population_2021.csv` (federal government data) to compare case numbers with local populations. The **"Vaccination Analysis"** section draws on `covid19_vaccine_data.csv` and `cases_phu.csv` (Ontario Data Catalogue) to explore vaccination trends and active case details across different Public Health Units (called "PHU" in later analysis).

1.3 Description of Datasets

1.3.1 Main Dataset: `Cases2021.csv`, `Cases2022.csv`, `Cases2023.csv`, `Cases2024.csv`

This dataset contains confirmed COVID-19 case data from multiple Public Health Units (PHUs) across Ontario, recorded between January 2021 and June 2024. The variables are as follows:

- **Case_Reported_Date**: Date of reported case
- **Age_Group**: Age bracket of the individual (`<20`, `20s`, `30s`, `40s`, `50s`, `60s`, `70s`, `80+`, 'UNKNOWN')
- **Client_Gender**: Gender of the individual (`MALE`, `FEMALE`, `UNSPECIFIED`, 'GENDER_DIVERSE')
- **Reporting_PHU_ID**: Unique identifier for the reporting PHU
- **Reporting_PHU_City**: City in which the PHU is located
- **Reporting_PHU_Latitude**: PHU's latitude coordinate
- **Reporting_PHU_Longitude**: PHU's longitude coordinate

1.3.2 COVID-19 Vaccine Data: `covid19_vaccine_data.csv`

This dataset provides COVID-19 vaccination data by Public Health Unit (PHU) and age group in Ontario, from July 2021 to November 2024:

- **Date**: Exact date of vaccination data (yyyy-mm-dd)
- **PHU ID**: Unique PHU identifier
- **PHU name**: PHU name
- **Agegroup**: Age category of the vaccinated population (`12-17`, `18-29`, `30-39`, etc.)
- **At least one dose_cumulative**: Cumulative count of individuals with at least one dose
- **Second_dose_cumulative**: Cumulative count with second dose
- **fully_vaccinated_cumulative**: Cumulative count fully vaccinated
- **third_dose_cumulative**: Cumulative count with a third dose
- **Total population**: Population size for each date, age group, and PHU
- **Percent_at_least_one_dose**: Percentage of individuals with at least one dose
- **Percent_fully_vaccinated**: Percentage fully vaccinated
- **Percent_3doses**: Percentage with three doses

1.3.3 Status of COVID-19 Cases by PHU: `cases_phu.csv`

This dataset covers COVID-19 case data from various PHUs in Ontario, recorded between April 1, 2020, and May 9, 2020:

- **FILE_DATE**: Date of record (yyyy-mm-ddthh:mm:ss)
- **PHU_NAME**: Name of the PHU
- **PHU_NUM**: PHU's numerical identifier
- **ACTIVE_CASES**: Number of active cases on the given date
- **RESOLVED_CASES**: Number of resolved cases on the given date
- **DEATHS**: The number of deaths due to COVID-19 on the given date.

1.4 Overall Research Question

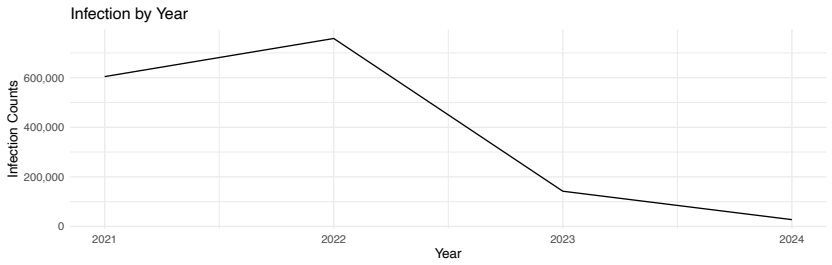
As researchers, our overarching goal is to investigate the temporal, demographic, and geographic patterns of COVID–19 in Ontario from 2021 to mid–2024. Specifically, this study examines how infection counts evolve over time, the impact of seasonal or monthly fluctuations, and the role of vaccination status and population size in shaping COVID–19 trends across different cities and age groups. Our key research questions are below:

- How did COVID–19 cases evolve from 2021 to mid–2024 in Ontario?
- Are there notable seasonal or monthly patterns in infection counts?
- Does population size or geographic location seem to influence infection rates in Ontario?
- What insights can be drawn from cities that experienced the greatest per-capita burdens?
- Is there a statistically significant difference in the average number of COVID-19 cases reported between FEMALE and MALE groups?
- Which age group is most affected by COVID-19 in Ontario, and how do these patterns change over time?
- How does vaccination coverage relate to COVID-19 case counts across Public Health Units (PHUs)?

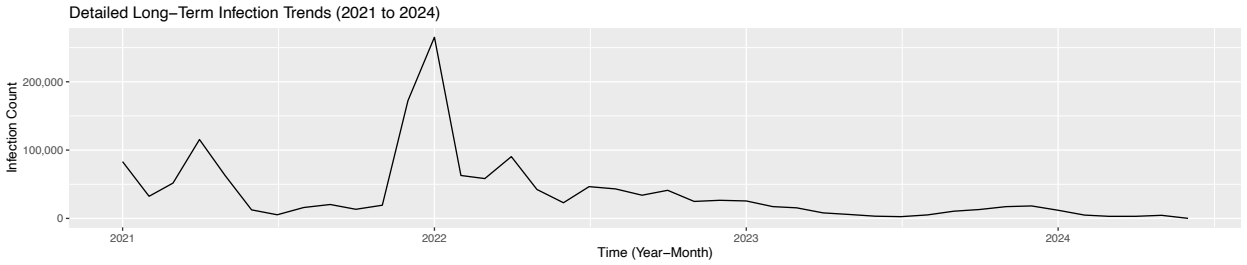
2. Time Series Analysis of COVID-19 Cases

2.1 Annual Trends

	Year	Total_Cases
1	2021	604515
2	2022	758269
3	2023	142134
4	2024	27529



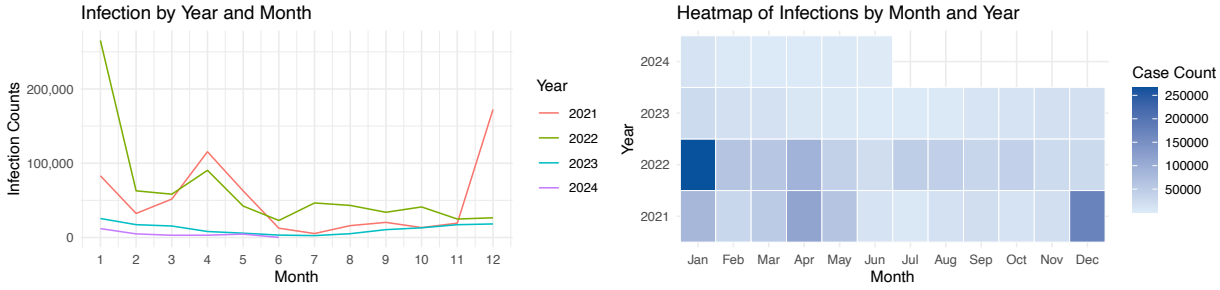
The plot reveals a sharp increase in COVID-19 cases from 2021 (604,515) to 2022 (758,269), followed by a dramatic drop in 2023—with only 142,134 cases, representing just **18.74% of the previous year**'s total. The 2024 figure appears much lower (27,529) because the dataset only includes cases from January to June; this limitation should be considered when interpreting the overall trend



The time series plot shows a **clear spike in January 2022**, which stands out as the **highest monthly** case count in the dataset. Before this peak, case numbers in 2021 fluctuated without a consistent pattern. After January 2022, infections declined sharply and continued to decrease, resulting in a relatively **stable and low trend throughout 2023 and the first half of 2024**.

This pattern helps explain why 2022 recorded the highest annual case count. A significant portion of that total appears to have occurred in January alone. The next section explores this further by comparing monthly trends within each year.

2.2 Monthly Trends by Year



In both the line plot and the heatmap, while **January 2022** stands out as the month with the highest case count, **December 2021** also should be highlighted. It shows a sharp rise, marked by dark shading in the heatmap, and appears to be the **second-highest month** overall. This suggests that the **major surge likely began in late 2021 and continued rising into January 2022**, where it reached its peak.

Furthermore, although January accounts for a large share of the total cases in 2022, the other months that year also remained high (as indicated by the green line on the graph). In fact, with the exception of April

and December, every month in 2022 recorded more cases than the corresponding months in other years. This indicates that case levels stayed elevated throughout most of 2022, not just during the early spike.

It is also worth noting that **April** had relatively high case counts in both 2021 and 2022, as shown in both graphs. In 2021, cases peaked in April, then declined through the summer, before rising again in December. While December 2021 has already been discussed, the earlier rise in April is another key moment in the year's case trend. The darker shading in the heatmap for both April and December highlights these shifts. These patterns suggest that **certain months may have had impact on trends**, which will be explored further in the next section.

In contrast, 2023 displays much lower and more stable monthly case counts. The data from 2024, which only includes the first six months, shows the lowest overall values, with little variation from month to month.

2.3 Seasonal Distribution of COVID-19 Cases by Month

Year	Minimum Month	Minimum Cases	Maximum Month	Maximum Cases	SD
2021	July	5324	December	172347	50942.397
2022	June	22935	January	265270	66396.232
2023	July	2560	January	25553	7133.434
2024	June	95	January	11939	3975.545

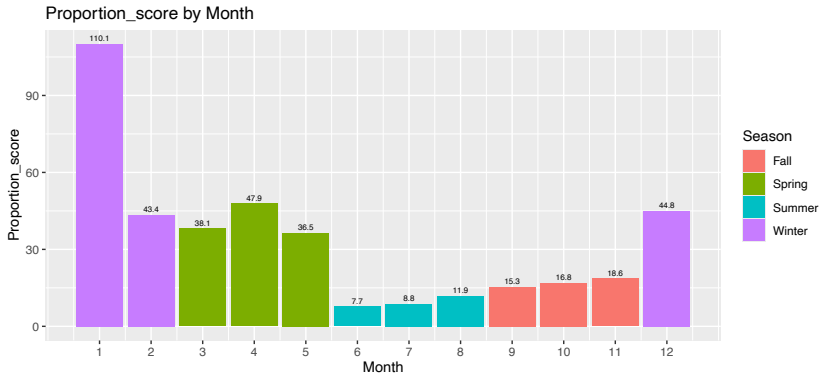
The **standard deviation values** reflect differences in case distribution. Both 2021 (SD $\approx$ **50,942**) and 2022 (SD $\approx$ **66,369**) show high variability across months, consistent with periods of sharp increases and declines. In contrast, 2023 (SD $\approx$ **7,133**) displays a more stable pattern, and 2024 (SD $\approx$ **3,976**) appears to continue that trend. However, it is important to note that **2024 data only includes January to June**, and thus may not reflect the full yearly trend.

This table reinforces earlier observations by confirming that **December 2021** (with **172,347** cases) and **January 2022** (with **265,270** cases) had the highest counts in the dataset. It also highlights a clear seasonal pattern: **June** and **July** consistently saw the **lowest case numbers**, while **December** (2021) and **January** (2022, 2023, and 2024) repeatedly recorded the **highest counts** across all years. Taken together, it suggests that **seasonal factors may have played a role** in shaping the overall case patterns, which will be explored further in the next section.

2.4 Month-Level Proportion Score

Note: The `Proportion_score` shows how much each month contributed to yearly case totals, averaged across all years. Instead of using raw case counts, it adds up the monthly percentages from each year. This way, every year is treated equally, no matter how many total cases it had. It helps avoid letting high-case years like 2022 overshadow the rest. Range of `Proportion_score` is 0-400.

	Month	Proportion_score
1	1	110.07
2	2	43.45
3	3	38.11
4	4	47.9
5	5	36.51
6	6	7.71
7	7	8.81
8	8	11.93
9	9	15.28
10	10	16.8
11	11	18.59
12	12	44.84



This chart reinforces earlier findings that **certain months consistently contribute more** to the total annual case counts, as measured by their `Proportion_score`. In particular, **January** stands out with a score of **110.07**, indicating that it accounts for a substantial share of the cases each year based on the 2021–2024 data, which aligns with the previous observation.

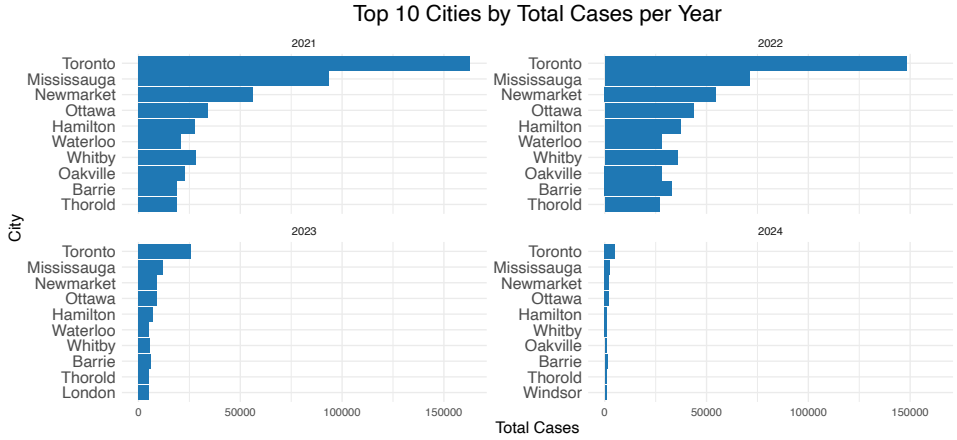
Other months also show noteworthy contributions—for example, **February** (**43.45**), **April** (**47.9**), and **December** (**44.84**). In contrast, **June** (**7.71**) and **July** (**8.81**) have much lower scores, supporting the observation that the summer months tend to have fewer cases.

Organizing the data by **season** further highlights this trend: **winter** (January and February) and **spring** (March, April, and May) months consistently show higher contributions, while **summer** (June, July, August) and **fall** (September, October, November) display lower contributions. This seasonal variation may suggest a relationship between temperature and transmission, though further investigation is needed to confirm this.

3. City Analysis of COVID-19 Cases

Note: This section is grouped by `Reporting_PHU_ID` and `Reporting_PHU_City`, which we rename it to `City` directly since there is only one Public Health Unit (PHU) per city. Also, 2021 census population is used for comparison. Although the reference year for the population data does not align exactly with the case years, the slight fluctuations in population should not significantly impact the burden ratio.

3.1 General City Analysis



	Top 12 Most Populous Ontario Cities (2021)	Population
1	Toronto	2794356
2	Ottawa	1017449
3	Mississauga	717961
4	Hamilton	569353
5	London	422324
6	Windsor	229660
7	Oakville	213759
8	Sudbury	166004
9	Barrie	147829
10	Guelph	143740
11	Whitby	138501
12	Kingston	132485

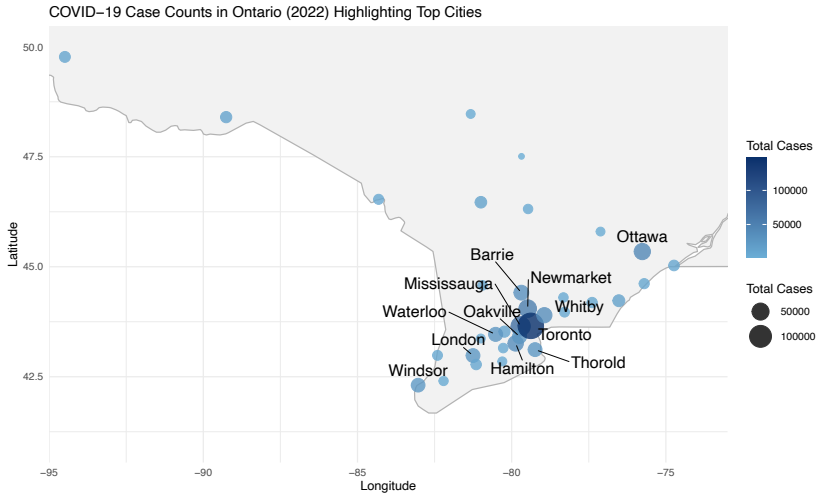
During 2021–2024, **Toronto, Mississauga, Newmarket, Ottawa, and Hamilton** consistently record the highest case counts. Although the remaining five cities in each year's "Top 10" shift slightly, they remain fairly stable overall. The bar chart suggests that **larger municipalities** like Toronto, Ottawa, and Mississauga usually have higher numbers, as confirmed by **2021 Census** data.

From all the cities that appeared in the Top 10 during these years, we identified **12** municipalities: **Toronto, Mississauga, Newmarket, Ottawa, Whitby, Hamilton, Oakville, Waterloo, Thorold, Barrie, London,** and **Windsor**.

These same 12 cities are plotted on the map below using **2022** data; notably, they cluster in **southern Ontario**, indicating that **geographical proximity** may also play a role in COVID-19 transmission patterns.

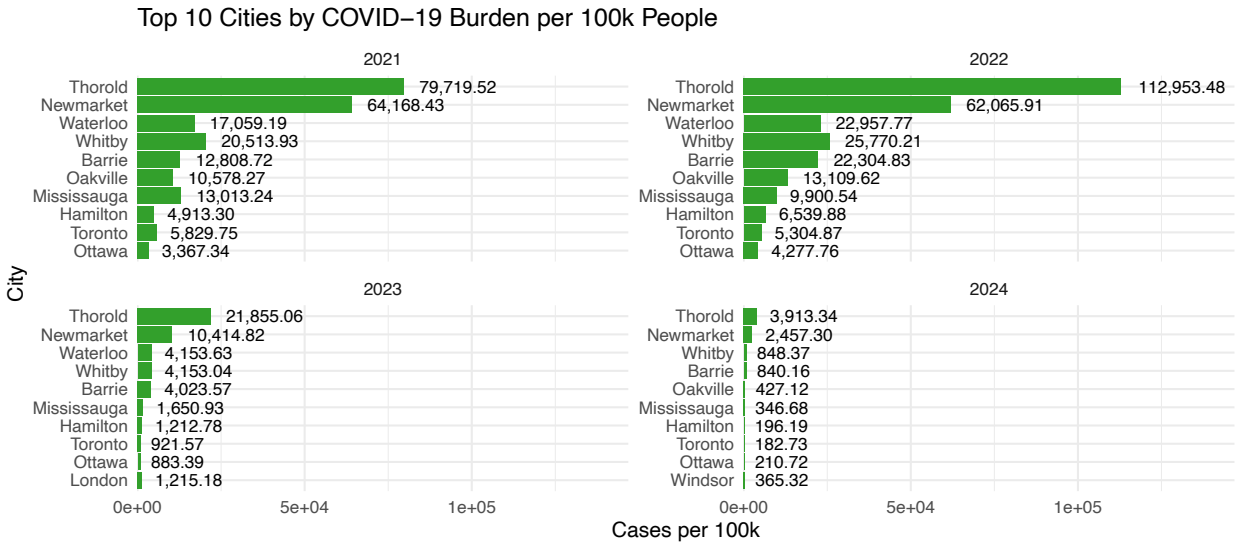
When compared to the **top 12 most populous Ontario cities**, most of the high-case-count municipalities also boast large populations (for example, Toronto leads in both cases and population, with 2,794,356 residents). Notably, **Newmarket, Waterloo, and Thorold** are not among the top 12 by population, yet they consistently rank in the Top 10 for total cases, suggesting elevated infection ratios.

Finally, **Newmarket, Waterloo, and Thorold** appear to have higher totals than expected for their populations, indicating **elevated infection ratios**. The next section examines these cities' per capita rates.



3.2 Infection Ratio by City

Note: The "COVID–19 Burden per 100k People" ratio is calculated by dividing each city's total cases by its 2021 census population. Although the reference year for the population data does not align exactly with the case years, the slight fluctuations in population should not significantly impact the burden ratio.



The chart displays the top 10 cities in Ontario ranked by their COVID–19 burden per 100,000 people for the years 2021 through 2024. **Thorold** consistently **topping the list each year**, while **Newmarket** also remains near the top, along with **Waterloo, Whitby, and Barrie**, which frequently appear in the upper half of the top 10. These findings align with the earlier observation that Thorold, Newmarket, and Waterloo—despite not being among the most populous cities—still reported high absolute case counts.

In contrast, larger cities such as **Oakville, Mississauga, Hamilton, Toronto,** and **Ottawa**—which typically report high overall case numbers—tend to have lower per-capita rates. This suggests that when population size is taken into account, their relative burden is mitigated compared to that of smaller cities. In 2024, overall rates appear lower for every municipality, since **2024 data only includes January to June**.

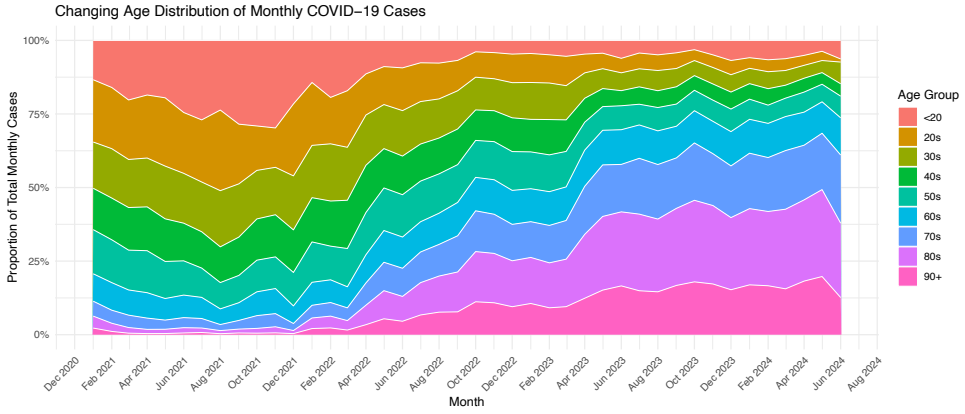
4 Gender and Age Analysis of COVID-19 Cases

4.1 Infection cases by Age Group

Note: There are only 542 cases of UNKNOWN in Age_Group. These cases are ignored in the analysis.

Total Cases by Age Group

	Age_Group	Total_Cases
1	20s	266058
2	30s	240078
3	40s	206551
4	50s	194341
5	60s	134651
6	70s	104602
7	80s	105027
8	90+	61440
9	<20	219157



The summary table shows that **people in their 20s had the highest total number of COVID–19 cases (266,058)**, followed by those in their 30s and under-20s. Overall, **individuals under 40 made up the majority of cases** throughout the study period.

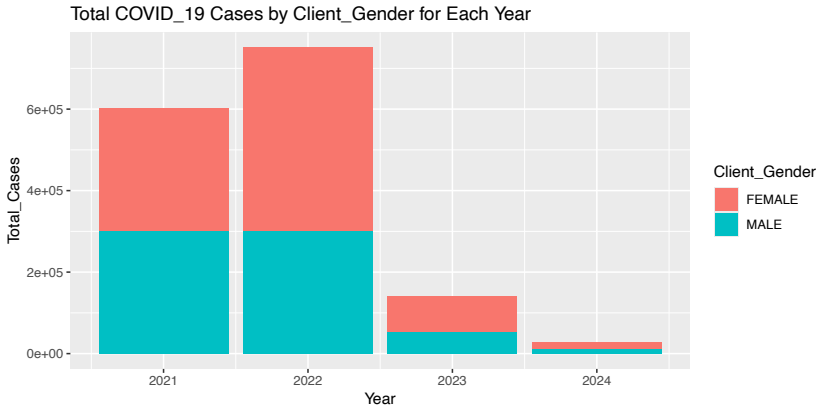
The area chart illustrates how this distribution changed over time. From July to November 2021, the under-20 age group had the largest monthly share of cases. However, during the spike in infections from December 2021 to January 2022, people in their **20s** became the **most affected group**.

At the start of the pandemic, those aged 80 and above comprised only a small portion of cases—just 2.53% in 2021, increasing slightly to 7.42% in the first half of 2022. Beginning in July 2022, older adults accounted for a growing share of infections. **By October 2022, individuals in their 80s and 90s made up 28.2% of monthly cases; this rose to 34.1% in April 2023, 45.6% in October 2023, and nearly half (49.3%) of all cases by May 2024.**

These trends suggest that **while younger people were more affected early on, the impact of COVID–19 gradually shifted toward older adults, especially those over 80, as the pandemic progressed.**

4.2 Infection cases by Client_Gender for Each Year

Note: In this analysis, records where `Client_Gender` is neither MALE nor FEMALE were excluded due to their low frequency.



Total Cases by Client Gender

	Year	Client_Gender	Total_Cases
1	2021	FEMALE	300885
2	2021	MALE	301168
3	2022	FEMALE	449596
4	2022	MALE	301050
5	2023	FEMALE	88254
6	2023	MALE	53143
7	2024	FEMALE	15983
8	2024	MALE	11481

The number of reported COVID-19 cases was consistently **higher among females** than males. Over the period, a total of **854,718** cases were reported for females, compared to **666,842** for males. This **gender-based difference** suggests that females were more likely to be diagnosed and reported as COVID-19 cases during the observed timeframe.

5. Hypothese testing and Boostrap

5.1 Hypothese testing

We want to determine whether there is a statistically significant difference in the average total COVID-19 cases between FEMALE and MALE groups per year.

5.1.1 Hypothesis:

H_0 (Null Hypothesis): There is no difference in the mean number of total cases per year between male and female clients.

H_a (Alternative Hypothesis): There is a difference in the mean number of total cases per year between male and female clients. ie: $\mu_{FEMALE} \neq \mu_{MALE}$

```
##
##  Welch Two Sample t-test
##
## data:  Total_Cases by Client_Gender
## t = 0.3721, df = 5.6856, p-value = 0.7233
## alternative hypothesis: true difference in means between group FEMALE and group MALE is not equal to
## 95 percent confidence interval:
##  -266082.4  360020.4
## sample estimates:
## mean in group FEMALE   mean in group MALE
##             213679.5             166710.5
```

5.1.2 Interpretation and Conclusion:

The **p-value (0.7233)** is greater than the **significance level of 0.05**. Therefore, we **fail to reject the null hypothesis**, indicating that there is not enough evidence to conclude a significant difference in the mean total cases per year between FEMALE and MALE groups.

The **95% confidence interval for the difference in means (−266,082.4 to 360,020.4)** includes **zero**, indicating that there is no statistically significant difference between the two group means.

Although the sample mean for **females (213,679.5)** appears higher than that for **males (166,710.5)**, this difference is not statistically significant based on the test results.

5.2 Bootstrap

Estimating the 95% confidence interval for the difference in mean total annual cases between male and female clients

We repeatedly (1,000 times) drew samples with replacement from the total COVID-19 cases per year in our dataset.For each bootstrap sample, we computed the mean. The distribution of these bootstrap means

provides an empirical approximation of the sampling distribution of the mean. We then used the 2.5th and 97.5th percentiles of this distribution to form a 95% confidence interval.

```
##      2.5%     97.5%
## -263057.6  158625.5
```

5.2.1 Interpretation and Conclusion

Based on the bootstrap distribution, the **95% confidence interval** for the mean total COVID–19 cases per year is approximately **-263057.6 to 158625.5**. This means that we are **95% confident** that the true mean of total cases lies within this interval.

If we were to repeat this data collection process many times, **95% of the similarly constructed bootstrap intervals** would contain the true mean total cases.

6 Regression Analysis and Cross Validation

6.1 Decision Choices

We selected the **random forest model** for our analysis due to its effectiveness in statistical modeling and machine learning, as it enhances predictive performance by reducing overfitting and capturing variable interactions.

We did not use **logistic regression**, as it is more suited for **binary outcomes**, whereas our outcome variable (new COVID-19 cases per population) is continuous. Additionally, linearity assumptions do not hold due to the observed non-linear relationships between vaccination rates and COVID-19 cases.

6.2 Data Preprocessing

Before conducting regression analysis, we preprocess the data to ensure consistency and compatibility between the two datasets. Given that these datasets originate from different sources and have distinct structures, we align them based on common variables such as PHU ID and time periods (Year and Month).

6.2.1 Preprocessing Steps:

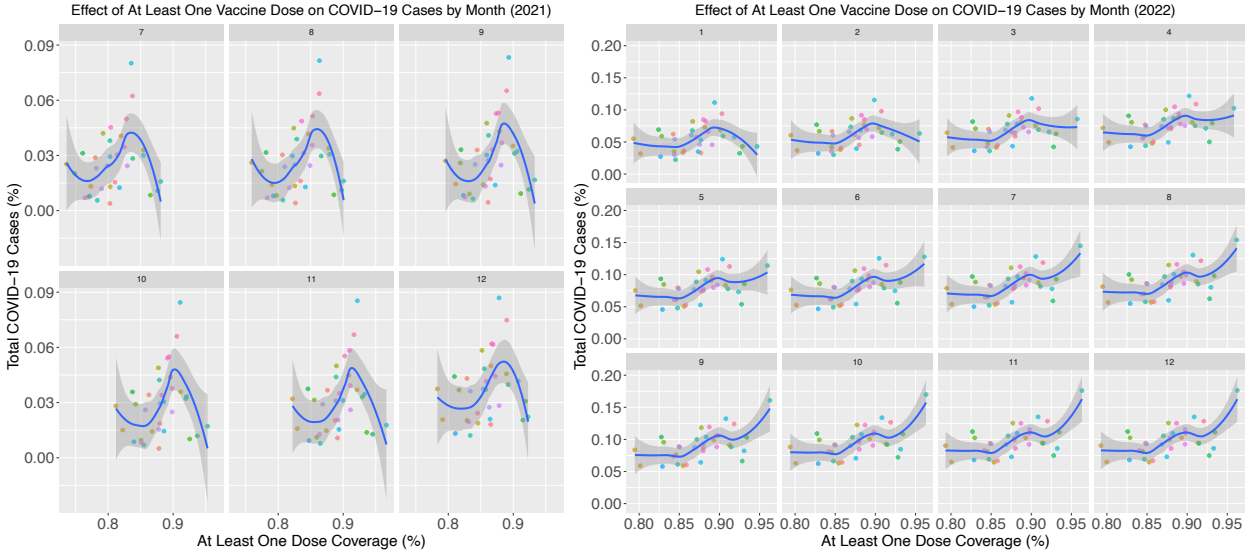
1. Handling Missing Values:
    - Remove NA values from the COVID-19 case data.
    - Convert dates in both datasets to year, month, and day formats for uniformity.
2. Aggregating Vaccine Data:
    - Group the COVID-19 vaccine data by PHU and month.
    - Calculate total vaccine doses administered, population size, and vaccination percentages for each PHU.
3. Summarizing COVID-19 Case Data:
    - Compute the number of new cases per day.
    - Aggregate the COVID-19 case data by PHU and month.
4. Merging Datasets:
    - Join the COVID-19 vaccine and case data using PHU ID, PHU name, year, and month as key columns.

6.3 Data Visualization

To examine the relationship between vaccination coverage and COVID-19 cases in Ontario, we generate visualizations using the preprocessed data. These plots help identify trends, patterns, and correlations across different Public Health Units (PHUs) and time periods. We focus on data from 2021 and 2022 for our analysis.

6.3.1 Rate of COVID-19 Cases vs. First-Dose Vaccination Rate

The following plots illustrate the relationship between first-dose vaccination rate and COVID-19 case rate per population for each PHU and month in 2021 and 2022.

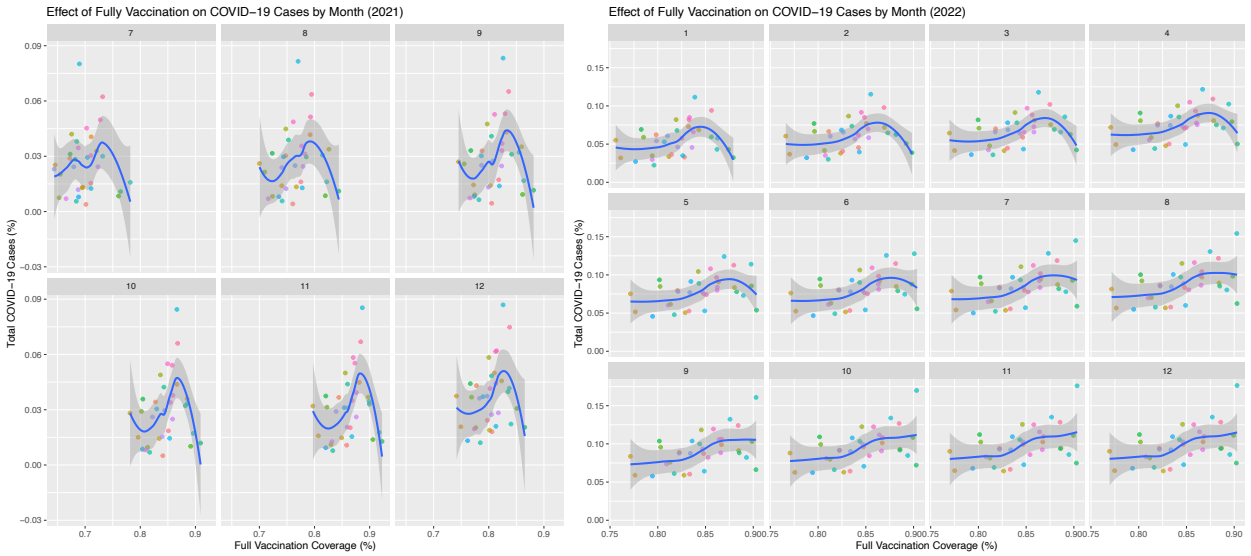


Findings:

- In 2021, a **cubic** relationship is observed between case percentage and vaccination coverage, suggesting an initial protective effect.
- In 2022, **fluctuations decrease, and an increasing trend emerges**, likely influenced by waning immunity, new variants, and public health measures.

6.3.2 Rate of COVID-19 Cases vs. Fully Vaccinated Rate

The following plots examine the relationship between fully vaccinated rate and COVID-19 case rate per population for each PHU and month in 2021 and 2022

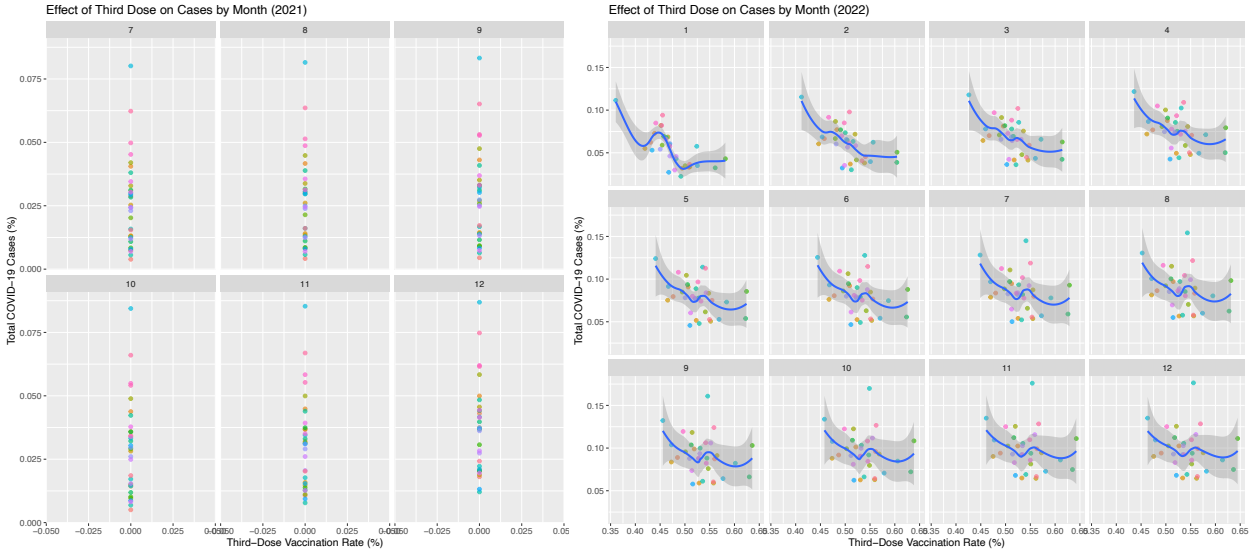


Findings:

- A complex interplay exists between vaccination and case rates.
- The trend stabilizes over time but shows a slight increase, indicating that **higher vaccination levels alone may not prevent infections**, possibly due to vaccine effectiveness and population immunity dynamics.

6.3.3 Rate of COVID-19 Cases vs. Third-Dose Vaccination Rate

The following plots examine the relationship between fully vaccinated rate and COVID-19 case rate per population for each PHU and month in 2021 and 2022.



Findings:

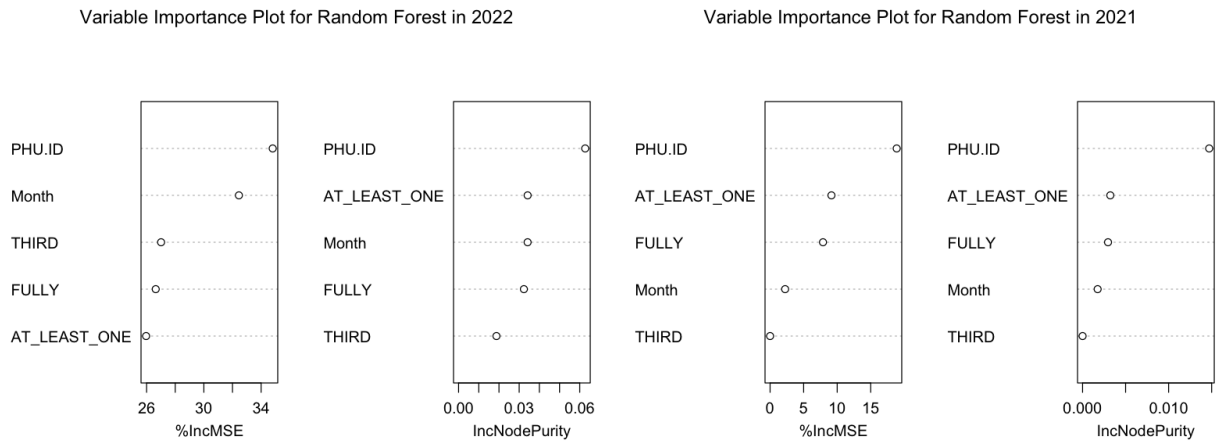
- In 2021, third-dose coverage is minimal and shows **no clear relationship** with case rates.
- In 2022, a **non-linear decreasing trend emerges**, suggesting **increased uptake of third doses correlates with fewer confirmed**. This indicates the importance of booster doses in reducing COVID-19 cases.

6.4 Random Forest Regression and Cross Validation Analysis

Finally, to enhance our predictive analysis, we apply random forest regression with 10-fold cross-validation to identify key predictors of COVID-19 cases.

6.4.1 Random Forest Model Feature Importance Plot

The following plots display the variable importance for the random forest models in 2021 and 2022:



6.4.2 10-fold Random Forest Cross Validation Analysis

The head of training data for each year is displayed to illustrate cross-validation data splitting. The mean squared errors (MSE) for the random forest models in 2021 and 2022 are calculated below.

Table 2: Training Data for Random Forest Model in 2021

PHU.ID	PHU.name	Year	Month	CASES	MIN_ONE	FULLY	THIRD	group_ind
2226	ALGOMA DISTRICT	2021	7	0.0038804	0.8027057	0.7015387	0	3
2226	ALGOMA DISTRICT	2021	8	0.0041582	0.8265531	0.7606829	0	3
2226	ALGOMA DISTRICT	2021	10	0.0050397	0.8781857	0.8408864	0	2
2226	ALGOMA DISTRICT	2021	11	0.0107404	0.8910819	0.8582666	0	6
2226	ALGOMA DISTRICT	2021	12	0.0180707	0.8653107	0.8057869	0	5
2227	BRANT COUNTY	2021	7	0.0288357	0.7802456	0.6722841	0	4

Table 3: Training Data for Random Forest Model in 2022

PHU.ID	PHU.name	Year	Month	CASES	MIN_ONE	FULLY	THIRD	group_ind
2226	ALGOMA DISTRICT	2022	1	0.0329383	0.8768952	0.8279860	0.5078920	6
2226	ALGOMA DISTRICT	2022	2	0.0454481	0.8799292	0.8423731	0.5407858	1
2226	ALGOMA DISTRICT	2022	3	0.0589455	0.8800537	0.8463325	0.5476991	8
2226	ALGOMA DISTRICT	2022	5	0.0751388	0.8798847	0.8482009	0.5547459	3
2226	ALGOMA DISTRICT	2022	6	0.0766870	0.8796000	0.8484145	0.5556623	6
2226	ALGOMA DISTRICT	2022	7	0.0797833	0.8791284	0.8481476	0.5570948	2

```
## [1] "The Mean of all MSE for the Random Forest Model in 2021: 0.00016727492879897"
```

```
## [1] "The Mean of all MSE for the Random Forest Model in 2022: 7.65529358005519e-05"
```

Findings:

- The variable importance plots highlight the significance of **first-dose percentage** and **fully vaccinated percentage** variables in predicting COVID-19 cases in Ontario for both years, while the **third-dose variable** becomes more important in 2022.
- The random forest models exhibit **strong predictive power**, with mean squared errors (MSE) ranging from 10^{-6} to 10^{-5}, indicating close alignment between predicted and actual values.
- The random forest model underscores the **critical role of vaccination status** in shaping COVID-19 outcomes and the **evolving importance of booster doses in 2022**.

7. Conclusion

The following are the key findings:

- COVID-19 cases in Ontario **peaked sharply in early 2022—particularly in January** before declining throughout 2023. Infections remained relatively low into mid-2024, although the absence of data for the second half of 2024 means final trends may still shift.
- **Seasonal patterns were evident**: case counts were highest in **winter months (especially December and January)** and lowest during **summer months (notably June and July)**, suggesting strong seasonal influences on transmission.
- **Major urban centers such as Toronto, Ottawa, Mississauga, and Hamilton recorded the highest total case counts**, largely reflecting their population size. However, several smaller cities—especially in southern Ontario—had disproportionately high per-capita infection rates.
- **Cities like Thorold, Newmarket, and Waterloo consistently experienced some of the highest relative infection burdens**, despite having smaller populations. This suggests that local conditions such as demographics or community behavior can significantly shape infection rates.
- **Females reported more confirmed COVID-19 cases than males** throughout the study period, with 666,842 cases among females versus 854,718 among males. This difference was statistically significant and may be influenced by differences in exposure, occupation, or healthcare-seeking behavior.

- **Shifts in case distribution by age group were observed over time**. While younger adults were more affected earlier in the pandemic, older adults—**especially those over 80**—represented a growing proportion of cases in later periods.
- **Vaccination coverage was strongly associated with lower case counts**. In 2021, first-dose uptake was the strongest predictor of reduced infections, while in 2022, full vaccination and third-dose rates also became significant. PHUs with higher vaccination rates consistently reported fewer cases.
- **Predictive modeling confirmed these findings, with both regression and random forest models achieving high accuracy**. Higher-order polynomial regression further improved prediction in 2022, reflecting the increasing complexity of factors influencing transmission over time.

This analysis highlights the multifaceted nature of COVID-19 transmission across Ontario, shaped by time, seasonality, geography, demographic variables, and vaccination efforts. These findings underscore the value of **targeted public health interventions**—particularly in high-burden communities—and support the continued role of vaccination and local health data in guiding future pandemic responses.

Appendix

```r
#Load Data
library(tidyverse); library(lubridate); library(knitr); library(dplyr);
library(ggplot2);library(gridExtra); library(patchwork)
vaccine = read_csv("COVID-Vaccine.csv"); case2021 = read_csv("Cases2021.csv")
case2022 = read_csv("Cases2022.csv"); case2023 = read_csv("Cases2023.csv")
case2024 = read_csv("Cases2024.csv")

#Combining the raw datasets.
Raw <- bind_rows(case2021, case2022, case2023, case2024)
Raw <- Raw %>% mutate(Year = year(as.Date(Case_Reported_Date)),
                      Month = month(as.Date(Case_Reported_Date)))
Raw <- Raw%>% rename(id = `_id`)

# 2.1 Annual Trends
YearInfection <- Raw %>% group_by(Year) %>%
  summarise(Total_Cases = n(), .groups = "drop", na.rm = FALSE)
#table_grob <- tableGrob(YearInfection)
p1 <- ggplot(YearInfection, aes(y = Total_Cases, x = Year)) + geom_line() +
  labs(title = "Infection by Year", y = "Infection Counts") +
  scale_y_continuous(labels = scales::comma) + theme_minimal()

#grid.arrange(table_grob, p1, ncol = 2, widths = c(1, 2))

# Line plot by month, colored by year
p_line <- ggplot(YearMonthInfection, aes(y = count, x = Month, color = factor(
  Year))) + geom_line() + labs(title = "Infection by Year and Month",
                               y="Infection Counts", color = "Year") +
  scale_x_continuous(breaks = 1:12) + scale_y_continuous(
    labels = scales::comma) +theme_minimal()

# Plot 2: Heatmap
p_heat <- YearMonthInfection %>%
  mutate(Month = factor(Month, levels = 1:12, labels = month.abb),
         Year = factor(Year)) %>% ggplot(aes(x = Month, y = Year, fill = count))
+ geom_tile(color = "white") + scale_fill_gradient(
  low = "#deebf7", high = "#08519c") +
  labs(title = "Heatmap of Infections by Month and Year", fill = "Case Count") +
  theme_minimal()

YearMonthInfection$Month <- month.name[YearMonthInfection$Month]
# Get minimum months
min_months <- YearMonthInfection %>% group_by(Year) %>%
  filter(count == min(count)) %>% slice(1) %>% ungroup()
```

```r
# Get maximum months and their proportions
max_months <- YearMonthInfection %>% group_by(Year) %>%
  filter(count == max(count)) %>%slice(1) %>% ungroup()

# Generate summary table
TimeStatTable <- YearMonthInfection %>% group_by(Year) %>% summarise(
    `Minimum Cases` = min(count),`Maximum Cases` = max(count), SD = sd(count),
    .groups = "drop") %>%
  left_join(min_months %>% select(Year, `Minimum Month` = Month), by = "Year")
%>% left_join(max_months %>% select(Year, `Maximum Month` = Month), by = "Year")
%>% relocate(`Minimum Month`, .before = `Minimum Cases`) %>%
  relocate(`Maximum Month`, .before = `Maximum Cases`)

#Proportion score
YearMonthInfection <- Raw %>% group_by(Year, Month) %>% summarise(
  count = n(), .groups = "drop")
Month_case_score <- YearMonthInfection %>% left_join(
  YearInfection, by = "Year") %>% mutate(Proportion = round(
    count / Total_Cases * 100, 2)) %>% group_by(Month)%>% summarise(
      Proportion_score = sum(Proportion))
table_prop <- tableGrob(Month_case_score)
Month_case_score <- Month_case_score %>% mutate(Season = case_when(
  Month %in% c(12, 1, 2) ~ "Winter", Month %in% c(3, 4, 5) ~ "Spring",
```

```r
  Month %in% c(6, 7, 8) ~ "Summer",Month %in% c(9, 10, 11) ~ "Fall"))

p2<- ggplot(Month_case_score, aes(y=Proportion_score, x=Month, fill=Season))+
  geom_col() + labs(title = "Proportion_score by Month", y="Proportion_score")+
  scale_x_continuous(breaks = 1:12) + geom_text(aes(label =scales::comma(
    Proportion_score)), vjust = -0.5, size = 2)

#grid.arrange(table_prop, p2, ncol = 2, widths = c(1, 2))

## 3.1 General City Analysis
Ontario_pop <- read.csv("Ontario_Cities_Population_2021.csv")
Top10_Cities_Per_Year <- Raw %>% group_by(Year, Reporting_PHU_City) %>%
  summarise(Total_Cases = n(), .groups = "drop_last") %>%
  slice_max(order_by = Total_Cases, n = 10, with_ties = FALSE) %>%
  arrange(Year, desc(Total_Cases))

Top10_Cities_Per_Year <- Top10_Cities_Per_Year %>% left_join(
  Ontario_pop, by = c("Reporting_PHU_City" = "City")) %>%
  mutate(cases_per_100k = (Total_Cases / Population_2021) * 100000)

# 1. Total cases plot
p <- ggplot(Top10_Cities_Per_Year, aes(y = Total_Cases, x = reorder(
  Reporting_PHU_City, Total_Cases))) + geom_col(fill = "#1f78b4") + coord_flip()
+ facet_wrap(~Year, scales = "free_y") + labs(
  x = "City", y = "Total Cases", title = "Top 10 Cities by Total Cases per Year")
+ theme_minimal()+ theme(axis.text.y = element_text(size = 16),
    plot.title = element_text(size = 18, hjust = 0.5))

# Prepare Table 1: Top 12 Populous Cities
top12pop <- Ontario_pop %>% arrange(desc(Population_2021)) %>% select(City) %>%
  head(12)
common12 <- Top10_Cities_Per_Year %>% ungroup() %>% distinct(Reporting_PHU_City)
%>% rename(City2 = Reporting_PHU_City)
combined_table <- bind_cols(top12pop, common12)
names(combined_table) <- c(
  "Top 12\nMost Populous\nOntario Cities (2021)",
  "Common Cities\nin Top 10\n(All Years)")
t1<- tableGrob(combined_table)
#grid.arrange(p, t1,ncol = 2, widths = c(2, 1.2))

City_per_Year <- Raw %>%
  group_by(Year, Reporting_PHU_City) %>%
```

```r
  left_join(Ontario_pop, by = c("Reporting_PHU_City" = "City")) %>%
  mutate(cases_per_100k = (Total_Cases / Population_2021) * 100000)

library(ggrepel)
city_coords <- Raw %>% distinct(Reporting_PHU_City, Reporting_PHU_Latitude,
                                Reporting_PHU_Longitude)%>%
  rename(lat =Reporting_PHU_Latitude, lon = Reporting_PHU_Longitude)

# Step 1: Get top 12 cities from previous step
label_cities <- Top10_Cities_Per_Year %>%
  ungroup() %>% distinct(Reporting_PHU_City) %>% pull(Reporting_PHU_City)

# Step 2: Prepare 2022 data with coordinates
map_data <- City_per_Year %>% filter(Year == 2022) %>%
  left_join(city_coords, by = "Reporting_PHU_City")

# Step 3: Filter only the cities to label
label_data <- map_data %>% filter(Reporting_PHU_City %in% label_cities)

# Step 4: Plot
#ggplot(map_data, aes(x = lon, y = lat)) +
#  borders("world", regions = "Canada", fill = "gray95", colour = "gray70") +
#  geom_point(aes(size = Total_Cases, color = Total_Cases), alpha = 0.8) +
#  geom_text_repel(data = label_data, aes(label = Reporting_PHU_City),
#    size = 5, max.overlaps = Inf, force = 2, force_pull = 2, nudge_y = 0.2,
```

```
# 2. Per capita cases plot
#ggplot(Top10_Cities_Per_Year, aes(y = cases_per_100k, x = reorder(
#Reporting_PHU_City, cases_per_100k))) + geom_col(fill = "#33a02c") +
#coord_flip() + facet_wrap(#~Year, scales = "free_y") + labs(x = "City",
#y = "Cases per 100k", #title = "Top 10 Cities by COVID-19 Burden per 100k Peopl
#) + theme_minimal()
```

```
knitr::opts_chunk$set(echo = TRUE)
required_packages <- c("tidyverse", "knitr", "caret", "rpart", "rpart.plot",
                       "randomForest")
for (pkg in required_packages) {
  if (!require(pkg, character.only = TRUE)) {
    install.packages(pkg)
    library(pkg, character.only = TRUE)
  }
}
# Read the dataset
Raw$Client_Gender[Raw$Client_Gender == "GENDER DIVERSE"] <- "UNSPECIFIED"

# 4.1 Age_Group
agg_data <- Raw %>%
  filter(Age_Group!="UNKNOWN")%>%
  group_by(Year, Month, Age_Group) %>%
  summarise(Total_Cases = n(), .groups = "drop", na.rm=TRUE)

# Create table grobs
table1_data <- Raw %>%
  filter(Age_Group!="UNKNOWN")%>%
  group_by(Age_Group) %>%
  summarise(Total_Cases = n()) %>%
  tableGrob()

title1 <- textGrob("Total Cases by Age Group", gp = gpar(
  fontsize = 14, fontface = "bold"))
table1 <- arrangeGrob(title1, table1_data, ncol = 1, heights = c(0.2, 1))

# Create a full date for each year-month combination
agg_data_clean <- agg_data %>%
  mutate(Date = as.Date(paste(Year, Month, "01", sep = "-"))) %>%
  group_by(Date) %>%
  mutate(Prop = Total_Cases / sum(Total_Cases)) %>%
  ungroup()

p <- ggplot(agg_data_clean, aes(x = Date, y = Prop, fill = Age_Group)) +
  geom_area(position = "stack", color = "white", size = 0.2) +
  scale_y_continuous(labels = scales::percent_format()) +
  scale_x_date(date_labels = "%b %Y", date_breaks = "2 months") +
  labs(
    title = "Changing Age Distribution of Monthly COVID-19 Cases",
    x = "Month",
    y = "Proportion of Total Monthly Cases",
    fill = "Age Group"
  ) +
  theme_minimal() +
  theme(
    legend.position = "right",
    axis.text.x = element_text(angle = 45, hjust = 1)
  )

grid.arrange(table1, p,ncol = 2, widths = c(1, 3.5))

# 4.2 Client_Gender
Gender_data <- Raw %>% filter(Client_Gender != "UNSPECIFIED") %>% group_by(
  Client_Gender, Year)%>% summarise(Total_Cases = n())

p <- ggplot(Gender_data, aes(x = Year, y = Total_Cases, fill = Client_Gender)) +
  geom_col() +
  labs(
    title = "Total COVID_19 Cases by Client_Gender for Each Year"
  )

table2_data <- Raw %>%
  group_by(Year, Client_Gender) %>%
  filter(Client_Gender != "UNSPECIFIED") %>%
  summarise(Total_Cases = n())%>%
  tableGrob()
title2 <- textGrob("Total Cases by Client Gender", gp = gpar(
  fontsize = 14, fontface = "bold"))
table2 <- arrangeGrob(title2, table2_data, ncol = 1, heights = c(0.2, 1))
grid.arrange(p, table2, ncol = 2, widths = c(2, 1))

# T-test
gender_test_data <- Gender_data %>%
  filter(Client_Gender %in% c("MALE", "FEMALE"))
t_test_gender <- t.test(Total_Cases ~ Client_Gender, data = gender_test_data)
t_test_gender

# Bootstrape
Gender_data <- Gender_data %>%
  filter(Client_Gender %in% c("MALE", "FEMALE"))

set.seed(123)
```

```
n_boot <- 1000
boot_diffs <- replicate(n_boot, {
  sample_data <- Gender_data %>% sample_frac(replace = TRUE)
  means <- sample_data %>% group_by(Client_Gender) %>%
    summarise(mean_cases = mean(Total_Cases)) %>% pull(mean_cases)
    diff(means)})
quantile(boot_diffs, c(0.025, 0.975))
```



```
# Random Forest Cross Validation for the Year 2021
d2021_cv = doses_cases2021 %>% mutate(PHU.ID = as.factor(PHU.ID),
                                      Year = as.factor(Year),
                                      Month = as.factor(Month),
                                      group_ind = sample(c(1:10), size = n(),
                                                         replace = T))
mse_vec2021 = vector()
for (i in 1:10) {
  train_data2021_cv = d2021_cv %>% filter(group_ind != i)
  test_data2021_cv = d2021_cv %>% filter(group_ind == i)
  rf_cv2021 = randomForest(PERCENT_CASES ~ PERCENT_AT_LEAST_ONE_DOSE +
              PERCENT_FULLY_VACCINATED + PERCENT_THIRD_DOSE + PHU.ID + Month,
              data = train_data2021_cv, importance = T)
  rf_predict2021_cv = predict(rf_cv2021, newdata = test_data2021_cv)
  mse_vec2021[i] = mean((rf_predict2021_cv - test_data2021_cv$PERCENT_CASES)^2)
}
# Display the head of the training data for 2021
kable(head(train_data2021_cv %>%
             select(PHU.ID, PHU.name, Year, Month, CASES = PERCENT_CASES,
                    MIN_ONE = PERCENT_AT_LEAST_ONE_DOSE,
                    FULLY = PERCENT_FULLY_VACCINATED,
                    THIRD = PERCENT_THIRD_DOSE, group_ind)),
      caption = "Training Data for Random Forest Model in 2021")
# Random Forest Cross Validation for the Year 2022
d2022_cv = doses_cases2022 %>% mutate(PHU.ID = as.factor(PHU.ID),
                                      Year = as.factor(Year),
                                      Month = as.factor(Month),
                                      group_ind = sample(c(1:10), size = n(),
                                                         replace = T))
mse_vec2022 = vector()
for (i in 1:10) {
  train_data2022_cv = d2022_cv %>% filter(group_ind != i)
  test_data2022_cv = d2022_cv %>% filter(group_ind == i)
  rf_cv2022 = randomForest(PERCENT_CASES ~ PERCENT_AT_LEAST_ONE_DOSE +
              PERCENT_FULLY_VACCINATED + PERCENT_THIRD_DOSE + PHU.ID + Month,
              data = train_data2022_cv, importance = T)
  rf_predict2022_cv = predict(rf_cv2022, newdata = test_data2022_cv)
  mse_vec2022[i] = mean((rf_predict2022_cv - test_data2022_cv$PERCENT_CASES)^2)
}
# Display the head of the training data for 2022
kable(head(train_data2022_cv %>%
             select(PHU.ID, PHU.name, Year, Month, CASES = PERCENT_CASES,
                    MIN_ONE = PERCENT_AT_LEAST_ONE_DOSE,
                    FULLY = PERCENT_FULLY_VACCINATED,
                    THIRD = PERCENT_THIRD_DOSE, group_ind)),
      caption = "Training Data for Random Forest Model in 2022")
print(paste("The Mean of all MSE for the Random Forest Model in 2021:", mean(mse_vec2021)))
print(paste("The Mean of all MSE for the Random Forest Model in 2022:", mean(mse_vec2022)))
```