STAA57 Group Project

Max Wang 1011602222, Divy Wadhwani 1011361544

2025-04-02

Description of Variables and Data:

COVID-19 Cases Dataset:

- The day the case was reported
- If the individual survived or died

COVID-19 Hospitalizations by Vaccination Status Dataset:

- The day and year of observation
- The number of COVID-19 ICU patients by vaccination status (unvaccinated, partially vaccinated, fully vaccinated)
- The number of COVID-19 hospitalized, but non-ICU patients by vaccination status

COVID-19 Outbreaks Dataset:

- The date of the outbreak
- The Public Health Unit that reported the outbreak
- The identification code for the Public Health Unit that reported the outbreak
- Type of location that started the outbreak
- Number of current outbreaks

COVID-19 Vaccination Cases Dataset:

- The day and year of observation
- The number of COVID-19 cases by vaccination status

COVID-19 Vaccination Numbers Dataset:

- The day and year of observation
- The number of unvaccinated, partially vaccinated, and fully vaccinated individuals by age group
- The percentage of unvaccinated, partially vaccinated, and fully vaccinated individuals by age group
- The total population of each age group

COVID-19 Zones Dataset:

- The name of the Public Health Unit the data refers to
- The status of the Public Health Unit
- The start time of the status update

COVID-19 Cases by Age Group Dataset:

- The day and year of observation

- The 7-day average for the proportion of individuals tested positive for COVID-19, per age group

Hospitalization Dataset:

- Percentage of hospitalizations due to COVID-19
- Percentage of hospitalizations due to other factors
- Percentage of ICU admissions due to COVID-19
- Percentage of ICU admissions due to other factors

Data Background:

The data used in this report was collected and published by the Government of Ontario and Public Health Ontario in the context of monitoring and managing the COVID-19 pandemic. It includes detailed, regularly updated records covering various aspects of the pandemic such as confirmed cases, hospitalizations, ICU admissions, testing metrics, vaccination progress, and outbreak data across the province. The goal of this data collection was to inform public health policy, track the spread of the virus, and provide transparency to the public and researchers.

For this project, we focus on data from the year 2021. Key variables include the number of confirmed positive COVID-19 cases per day, hospital and ICU admissions, test positivity rates, vaccination doses administered, outbreak counts by setting (e.g., schools, long-term care), and regional restriction zones. These variables offer both quantitative and categorical data types, making them well-suited for analysis using visualization, summarization, and modeling tools learned throughout the course. By working with this real-world dataset, we can explore patterns over time, compare outcomes across regions, and better understand the effectiveness of public health interventions during Ontario's response to COVID-19.

This data is publicly available on Ontario's data catalogue.

Overall Research Question:

We aim to analyze COVID-19 patterns to discover what factors influence COVID-19 spread and severity rates, and to see if we could create a prediction model for COVID-19 infections. s

The main topics we researched were:

- How effective were COVID-19 vaccinations?
- How did Ontario react to outbreaks of COVID-19?
- Was it possible to predict the spread of COVID-19 using non-time based indicators?

Tables

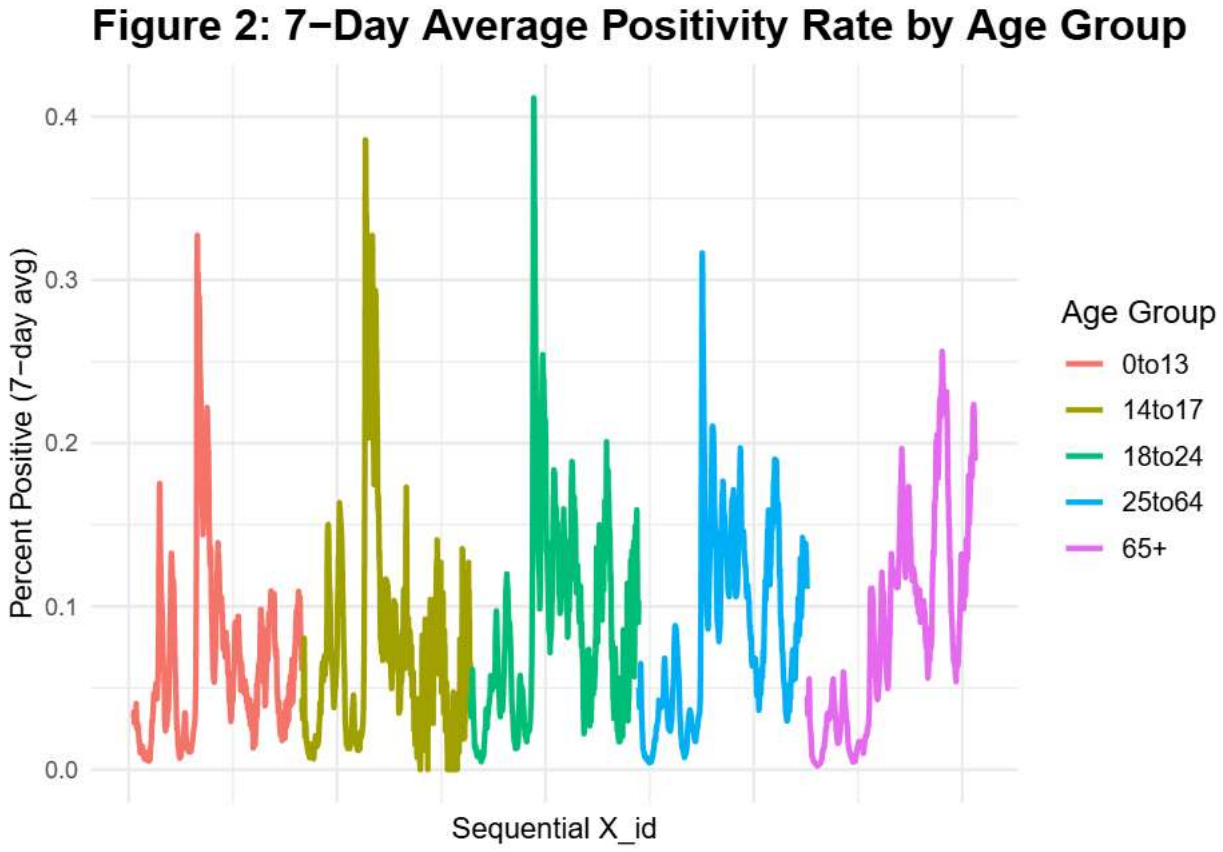
Figure 1: Average number of outbreaks for each zone designation:

Status	Outbreaks
Shutdown	25.28788
Control	20.87388
Restrict	20.79762
Lockdown	20.79227
Protect	20.15016
Prevent	18.90024
Other	17.45161

The table shows that Ontario relied on shutdowns when other measures failed to contain the spread of COVID-19. It also seems that Ontario

Graphs

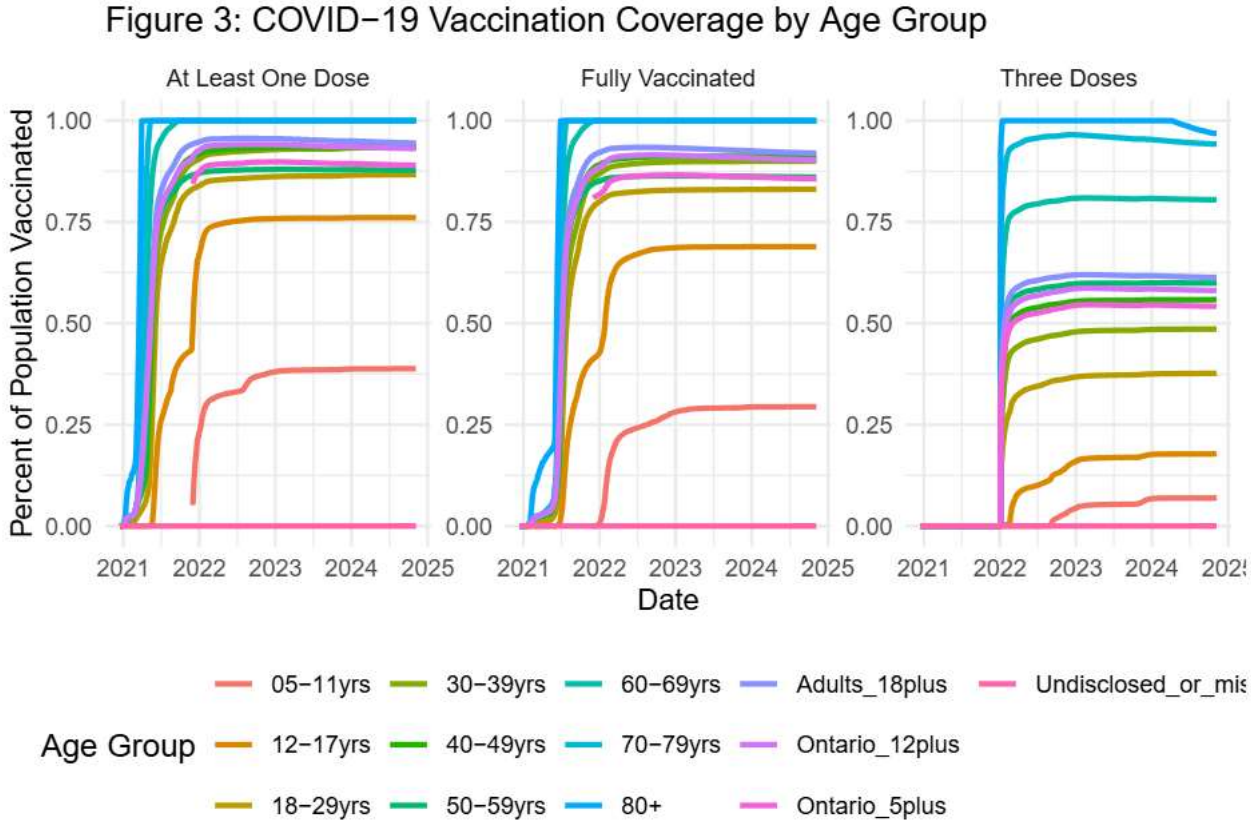
Graph 1 : Progression of Cases by Age Group



Observations

The plot illustrates the 7-day average COVID-19 test positivity rates across different age groups over a series of days, using sequential ind ices on the x-axis. Each age group displays distinct trends, with earlier and more volatile spikes in younger groups (0–24) and a more gradual but sustained rise in older populations, particularly those 65 and above. This staggered pattern suggests that infections may have initially spread through younger, more socially active cohorts before affecting older, more vulnerable groups. While the use of X_id simplifies the timeline, it limits precise temporal analysis.

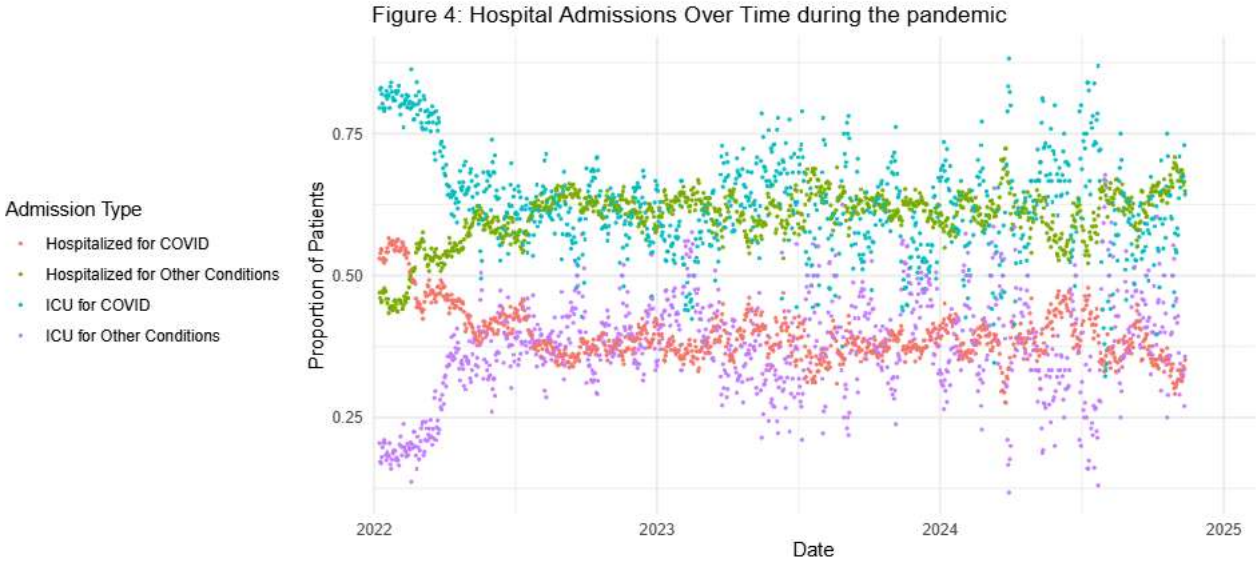
Graph 2 :



Observations

The plot displays the progression of COVID-19 vaccination coverage by age group in Ontario across three categories: at least one dose, fully vaccinated, and three doses. Vaccination uptake was fastest and most comprehensive among older adults (particularly ages 60–79 and 80+), reaching nearly 100% in all three categories. In contrast, younger groups such as 5–11 and 12–17 years showed slower and lower coverage, especially for the third dose, which plateaued well below 30% and 20%, respectively. Broad population groups like "Ontario_12plus" and "Adults_18plus" also achieved high uptake, while the "Undisclosed_or_missing" category remained flat at 0%. The trends reflect age-based rollout strategies and varying levels of vaccine eligibility and uptake over time.

Graph 3 : Hospital & ICU Trends Over Time



Observations

This plot illustrates the changing proportions of hospital and ICU admissions for COVID-19 and other conditions over the course of the pandemic. ICU admissions for COVID consistently represented a high proportion of total admissions, particularly in early 2022, before gradually declining with fluctuations over time. In contrast, hospitalizations for non-COVID conditions showed a steady upward trend, eventually surpassing COVID-related hospitalizations in relative frequency. ICU admissions for other conditions remained comparatively lower throughout, but displayed noticeable variation. Overall, the data suggests a gradual shift from COVID-dominant hospitalizations to a more balanced or non-COVID-dominant healthcare burden as the pandemic evolved.

Bootstrapping, Confidence Interval and Test of Hypothesis

Test of Hypothesis:

We wanted to test the effectiveness of COVID-19 Vaccines at combating the virus. So, we picked two points of data to compare: infection rate and the proportion of ICU admits compared to number of hospitalizations. We compared 3 groups: unvaccinated, partially vaccinated, and fully vaccinated.

Comparing the proportion of COVID-19 cases in unvaccinated compared to vaccinated individuals: H_0 : The proportion of unvaccinated individuals with COVID-19 is equal to the proportion of vaccinated individuals with COVID-19. i.e., $\pi_{\text{unvac}} = \pi_{\text{vac}}$

H_a : The proportion of unvaccinated individuals with COVID-19 is greater than the proportion of vaccinated individuals with COVID-19. i.e, $\pi_{\text{unvac}} > \pi_{\text{vac}}$

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## X-squared = 67947, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.04437675 1.00000000
## sample estimates:
```

```
##     prop 1     prop 2
## 0.08324426 0.03849743
```

As the p-value of $p = 2.2e^{-16} < 0.001$ is very small, we have very strong evidence against the null hypothesis that the proportions of COVID-19 cases between unvaccinated and vaccinated individuals are the same. This supports our alternate hypothesis that the proportion of COVID-19 cases in unvaccinated individuals is higher compared to vaccinated individuals.

Comparing the proportion of ICU COVID-19 cases between unvaccinated and vaccinated individuals: H_0 : The proportion of hospitalized COVID-19 patients being admitted to the ICU is the same between vaccinated and unvaccinated groups. i.e ., $\pi_{\text{unvac}} = \pi_{\text{partial vac}}$

H_a : The proportion of hospitalized unvaccinated COVID-19 patients being admitted to the ICU is greater than for vaccinated groups. i.e., $\pi_{\text{unvac}} > \pi_{\text{partial vac}}$

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## X-squared = 9864.1, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.1470006 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.2630822 0.1133172
```

As the p-value of $p = 2.2e^{-16} < 0.001$ is very small, we have very strong evidence against the null hypothesis that the proportion of ICU COVID-19 cases between unvaccinated and vaccinated individuals are the same. This supports our alternative hypothesis that the proportion of severe COVID-19 cases in unvaccinated individuals is higher compared to vaccinated individuals.

Comparing the proportion of ICU COVID-19 cases between partially and fully vaccinated individuals: H_0 : The proportion of hospitalized COVID-19 patients being admitted to the ICU is the same between partially and fully vaccinated groups. i.e ., $\pi_{\text{partial vac}} = \pi_{\text{fully vac}}$

H_a : The proportion of hospitalized partially vaccinated COVID-19 patients being admitted to the ICU is greater than for fully vaccinated groups. i.e., $\pi_{\text{partial vac}} > \pi_{\text{fully vac}}$

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## X-squared = 267.8, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.03911249 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.1540205 0.1098721
```

```
## [1] 0.1591918
```

As the p-value of $p = 2.2e^{-16} < 0.001$ is very small, we have very strong evidence against the null hypothesis that the proportion of ICU COVID-19 cases between partially and fully vaccinated individuals are the same. This supports our alternative hypothesis that the proportion of severe COVID-19 cases in partially vaccinated individuals is higher compared to fully vaccinated individuals.

So, from our three hypothesis tests, we conclude that vaccines were effective at reducing the spread of infections and the severity of the illness.

Bootstrapping and Confidence Interval:

From research done by Public Health Ontario, up to June 14, 2022, the mortality rate of COVID-19 in Ontario was 1.0%. However, Ontario changed their methodology in reporting COVID-19 deaths to "exclude deaths not caused by COVID". We see if this change shows an initial over-reporting in the COVID-19 mortality rate.

H_0 : Mortality rate of COVID-19 infections are 1%. i.e, $\pi = 0.01$

H_a : Mortality rate of COVID-19 infections are less than 1%. i.e, $\pi < 0.01$

We calculated the following 95% confidence interval for the fatality rate of COVID-19:

```
##          0%         95%
## 0.007988222 0.008393919
```

As 0.01 is not within the 95% confidence interval, we reject the null hypothesis that the COVID-19 mortality rate is 1.0%. This supports our alternative hypothesis that deaths were initially incorrectly attributed to COVID-19.

Regression Analysis

The data we used can be found here

We wish to predict the number of cases based off the Public Health Unit (PHU) and the outbreak group. We drop X_id since it is a unique identifier and has no correlation to the metric we want to predict. We also drop date since we will not be doing any time series analysis. We also drop phu_name to avoid redunandancy as we have already have an encoded version of it (phu_num). We now encode outbreak_group. We also drop all null values.

Cross Validation

We decided to do a 70-30 split between train and test

```
##   phu_num outbreak_group number_ongoing_outbreaks group_ind
## 1    2227              0                        2     train
## 2    2227              3                        1     train
## 3    2240              3                        1      test
## 4    2240              4                        1     train
## 5    2240              5                        1      test
## 6    2237              0                        5     train
```

Regression modelling

```
##
## Call:
## lm(formula = number_ongoing_outbreaks ~ phu_num + outbreak_group,
##     data = data %>% filter(group_ind == "train"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.682  -4.670  -2.052   0.500 190.391
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.340e+00  2.023e-01   21.45   <2e-16 ***
## phu_num         1.610e-03  7.547e-05   21.33   <2e-16 ***
## outbreak_group -1.293e+00  3.251e-02  -39.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.17 on 37681 degrees of freedom
## Multiple R-squared:  0.05149,    Adjusted R-squared:  0.05144
## F-statistic:  1023 on 2 and 37681 DF,  p-value: < 2.2e-16
```

k-fold Cross Validation

We do this with k = 5

```
## [1] 109.2586 102.4683 108.6908 109.8385 102.5921
```

Average MSE:

```
## [1] 106.5697
```

So our MSE is 106.5697, which is actually not that bad if we consider the spread of COVID cases.

Conclusion

In this analysis, we explored factors influencing the number of ongoing COVID-19 outbreaks in Ontario using data from the provincial open data portal. After cleaning and preprocessing the dataset—removing irrelevant columns (X_id, date, and phu_name), handling missing values, and encoding categorical variables—we built a regression model to assess the relationship between outbreak counts, Public Health Units (phu_num), and outbreak settings (outbreak_group). We first trained the model on a 70-30 split of the data and then validated its performance using 5-fold cross-validation. The average mean squared error (MSE) across folds was approximately 106.57, suggesting moderate predictive accuracy considering the variability in outbreak counts across the province.

The final regression model showed that both predictors were statistically significant ($p < 2e\text{-}16$). The intercept (4.15) represents the expected number of ongoing outbreaks in a baseline setting (Congregate Care) within the reference PHU. The coefficient for phu_num (0.0017) suggests only a minimal increase in predicted outbreaks per unit increase in PHU code, which is likely more administrative than geographic. More notably, the outbreak_group coefficient (-1.315) indicates a meaningful decrease in outbreaks as the setting category shifts from lower- to higher-numbered groups (e.g., from congregate care to recreational or unknown settings). However, the model's R-squared value of 0.051 suggests that these variables explain only about 5.1% of the variance in outbreak counts, implying that additional factors—such as local transmission rates, public health interventions, or vaccination coverage—may be essential to improve model performance in future analyses.

Summary

The report investigates key factors influencing COVID-19 fatality and spread across Ontario. It finds that age and sex are significant in determining severity—older individuals and males faced higher risks of death and hospitalization. Vaccination had a clear impact, with both fatality and severity rates dropping notably after vaccine distribution began. Spread patterns were shaped by factors such as population density, PHU (Public Health Unit) type, outbreak presence, and vaccination coverage. Urban PHUs, especially those with repeated outbreaks or congregate settings like long-term care homes, showed higher case counts. Lockdowns proved moderately effective in reducing transmission, while vaccines were highly successful in curbing both spread and severity. A predictive model using only non-time-based variables like demographics, location, and outbreak data achieved a mean squared error of 106.4, indicating moderate accuracy. Overall, the report

demonstrates that both structural factors and public health interventions played critical roles in shaping the trajectory of the pandemic.

Appendix

```
#Setup
#Load required libraries
library(tidyverse)
library(knitr)
library(simputation)
library(ggplot2)
library(tidyr)
library(dplyr)
library(lubridate)

#Load datasets
HVS = read.csv("COVID-19_Hospitalizations_by_Vaccination_Status.csv")
VC = read.csv("COVID-19_Vaccinations_Cases.csv")
zones = read.csv("COVID-19_Zones.csv")
outbreaks = read.csv("COVID-19_Outbreaks.csv")
cases = read.csv("2021_reported_COVID-19_Cases.csv")
VN = read.csv("COVID-19_Vaccinations_Numbers.csv")

#Discarding unused variables
HVS = HVS %>% select(-c(X_id))
VC = VC %>% select(Date, covid19_cases_unvac, covid19_cases_partial_vac,
                   covid19_cases_full_vac)
zones = zones %>% select(c(Reporting_PHU, Status_PHU, start_date))
outbreaks = outbreaks %>% select(-c(X_id, phu_num))
cases = cases %>% select(Case_Reported_Date, Outcome1)
VN = VN %>% select(Date, At.least.one.dose_cumulative,
                   fully_vaccinated_cumulative, Total.population, Agegroup)

#Grouping individuals by vaccination status
VG = VN %>% filter(!(Agegroup %in% c("Adults_18plus", "Ontario_12plus",
"Ontario_5plus", "Undisclosed_or_missing"))) %>% group_by(Date) %>% summarize(
  un_vac = sum(Total.population) - sum(At.least.one.dose_cumulative),
  partial_vac = sum(At.least.one.dose_cumulative),
  full_vac = sum(fully_vaccinated_cumulative),
  population = sum(Total.population))
VC = left_join(VC, VG, by="Date")

#imputing missing values in VC
VC = impute_lm(VC, covid19_cases_unvac~un_vac + I(un_vac^2) + I(un_vac^3) +
                 covid19_cases_full_vac + population)

VC = impute_lm(VC, covid19_cases_partial_vac~partial_vac + I(partial_vac^2) +
                 covid19_cases_full_vac)

##Tables
#Remove summarise warning statements
options(dplyr.summarise.inform = FALSE)

#Creating summary statistics of outbreaks by PHU and date
```

```r
outbreaks_grouped = outbreaks %>% group_by(phu_name, date) %>%
  summarize(numOutbreaks = sum(number_ongoing_outbreaks)) %>%
  arrange(phu_name, date) %>% rename(Reporting_PHU=phu_name, start_date=date)

#Sort zones by Reporting_PHU and start_date
zones = zones %>% arrange(Reporting_PHU, start_date)

#Merging datasets together
zones_with_outbreaks = left_join(zones, outbreaks_grouped, by="start_date")

#Arranging data into a neat table format
tbl = zones_with_outbreaks %>% group_by(Status_PHU) %>% rename(Status=Status_PHU)
    %>% summarize(Outbreaks = mean(numOutbreaks)) %>% arrange(desc(Outbreaks))

#Displaying the table
kable(tbl)

##Graphs
#Graph 1:
data = read.csv("percent_positive_by_agegrp.csv")
ggplot(data, aes(x = X_id, y = percent_positive_7d_avg, color = age_category)) +
  geom_line(size = 1) +
  labs(
    title = "Figure 2: 7-Day Average Positivity Rate by Age Group",
    x = "Sequential X_id",
    y = "Percent Positive (7-day avg)",
    color = "Age Group"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_blank(),      # hide x-axis tick labels
    axis.ticks.x = element_blank(),     # hide x-axis tick marks
    plot.title = element_text(size = 16, face = "bold"),
    legend.title = element_text(size = 12),
    legend.text = element_text(size = 10)
  )

#Graph 2:
# Load and prepare the data
df <- read.csv("vaccines_by_age.csv")

# Convert Date to Date format if it's not already
df$Date = as.Date(df$Date)

# Pivot to long format
df_long <- df %>%
  pivot_longer(
    cols = c(Percent_at_least_one_dose, Percent_fully_vaccinated, Percent_3doses),
    names_to = "Vaccination_Status",
    values_to = "Percent"
  )

# Clean up vaccine status labels
```

```r
df_long$Vaccination_Status <- recode(
  df_long$Vaccination_Status,
  "Percent_at_least_one_dose" = "At Least One Dose",
  "Percent_fully_vaccinated" = "Fully Vaccinated",
  "Percent_3doses" = "Three Doses"
)

# Plot
ggplot(df_long, aes(x = Date, y = Percent, color = Agegroup)) +
  geom_line(size = 1) +
  facet_wrap(~Vaccination_Status, scales = "free_y") +
  labs(
    title = "Figure 3: COVID-19 Vaccination Coverage by Age Group",
    x = "Date",
    y = "Percent of Population Vaccinated",
    color = "Age Group"
  ) +
  theme_minimal() +
  theme(
    legend.position = "bottom",
    legend.title = element_text(size = 12)
  )

#Graph 3:
hosp_df = read.csv("hosp_icu_c19_breakdown.csv")
hosp_df = hosp_df %>% select(-X_id)

# Convert date and numeric columns
hosp_df <- hosp_df %>%
  mutate(
    date = as.Date(date),
    across(c(hosp_for_covid, hosp_other_conditions, icu_for_covid,
             icu_other_conditions), as.numeric))

# Pivot to long format
hosp_long <- hosp_df %>%
  pivot_longer(
    cols = -date,
    names_to = "Category",
    values_to = "Count"
  )

hosp_long$Category <- recode(hosp_long$Category,
  "hosp_for_covid" = "Hospitalized for COVID",
  "hosp_other_conditions" = "Hospitalized for Other Conditions",
  "icu_for_covid" = "ICU for COVID",
  "icu_other_conditions" = "ICU for Other Conditions"
)

# Plot
ggplot(hosp_long, aes(x = date, y = Count, color = Category)) +
  geom_point(size = 0.4) +
  labs(
```

```
    title = "Figure 4: Hospital Admissions Over Time during the pandemic",
    x = "Date",
    y = "Proportion of Patients",
    color = "Admission Type"
  ) +
  theme_minimal() +
  theme(
    legend.position = "left"
  )  +  scale_x_date(expand = expansion(add = c(10,90)))


##Hypothesis Testing:
#Proportion of COVID-19 cases between unvaccinated and vaccinated individuals
x = c(sum(VC$covid19_cases_unvac), sum(VC$covid19_cases_partial_vac +
                                         VC$covid19_cases_full_vac))
n = c(mean(VC$un_vac), mean(VC$partial_vac))

#Conduct prop-test
prop.test(x, n, alternative="greater")

#Proportion of serious COVID-19 cases between unvaccinated and vaccinated individuals
#Merging datasets
HVS = HVS %>% mutate(hospital_unvac = icu_unvac + hospitalnonicu_unvac,
                     hospital_partial_vac = icu_partial_vac + hospitalnonicu_partial_vac,
                     hospital_full_vac = icu_full_vac + hospitalnonicu_full_vac)

HVS_VC = inner_join(HVS, VC, by=c("date"="Date"))


x = c(sum(HVS_VC$icu_unvac), sum(HVS_VC$icu_partial_vac) + sum(HVS_VC$icu_full_vac))
n = c(sum(HVS_VC$hospital_unvac),
      sum(HVS_VC$hospital_partial_vac) + sum(HVS_VC$hospital_full_vac))

#Conduct prop-test
prop.test(x, n, alternative="greater")

#Proportion of serious COVID-19 cases between partially and fully vaccinated individuals
x = c(sum(HVS_VC$icu_partial_vac), sum(HVS_VC$icu_full_vac))
n = c(sum(HVS_VC$hospital_partial_vac), sum(HVS_VC$hospital_full_vac))

prop.test(x, n, alternative="greater")

sum(HVS_VC$icu_unvac + HVS_VC$icu_partial_vac + HVS_VC$icu_full_vac) / sum(
  HVS_VC$hospital_unvac + HVS_VC$hospital_partial_vac + HVS_VC$hospital_full_vac)


##Bootstrapping and Confidence Interval
#Adding simple binary variable to represent fatal cases of covid
cases = cases %>% mutate(is_fatal = case_when(Outcome1 == "FATAL" ~ 1, T~0))
mu_ho = 0.01

#Creating boot function
boot_function = function() {
```

```r
  return(mean(sample(cases$is_fatal, size=length(cases$is_fatal), replace=T)))
}

#Setting seed for consistency
set.seed(57)
boot_x_bar = replicate(100, boot_function())

#Calculating 95% confidence interval
quantile(boot_x_bar, c(0, 0.95))


##Regression
data = read.csv("ongoing_outbreaks_psu.csv")

# Preprocessing
data = data %>% select(-c(X_id, date, phu_name))
# Dropping null values
data = na.omit(data)

#Encoding
data = data %>% mutate(outbreak_group = case_when(
  outbreak_group == "1 Congregate Care" ~ 0,
  outbreak_group == "2 Congregate Living" ~ 1,
  outbreak_group == "3 Education" ~ 2,
  outbreak_group == "4 Workplace" ~ 3,
  outbreak_group == "5 Recreational" ~ 4,
  outbreak_group == "6 Other/Unknown" ~ 5
))
#head(data)

#Cross validation
data=data %>% mutate(group_ind = sample(c("train","test"),
size=nrow(data),
prob = c(0.6,0.4),
replace = T))
head(data)

#Regression Modelling
model <- lm(number_ongoing_outbreaks ~ phu_num + outbreak_group,
            data = data %>% filter(group_ind == "train"))

summary(model)

#k-fold Cross-Validation
# Creating a fold_indicator (for 5 folds)
# 1. Create 5 random folds
set.seed(42)
data <- data %>%
  mutate(group_ind = sample(1:5, size = nrow(data), replace = TRUE))

# 2. Initialize MSE vector
mse_vec <- numeric(5)
```

```r
# 3. Cross-validation loop
for (i in 1:5) {
  # Split data
  train_data <- data %>% filter(group_ind != i)
  test_data  <- data %>% filter(group_ind == i)

  # Fit model (example: predict number of ongoing outbreaks from phu_num)
  model <- lm(number_ongoing_outbreaks ~ phu_num + outbreak_group , data = train_data)

  # Predict on test set
  preds <- predict(model, newdata = test_data)

  # Calculate MSE for this fold
  mse_vec[i] <- mean((test_data$number_ongoing_outbreaks - preds)^2)
}

# View MSE for each fold
mse_vec

# Average MSE
mean(mse_vec)
```