Flu Vaccination Trends: Costs, Coverage, and Emergency Care Effects

Sheamin Kim, Lomash Sharma, Chris Joon Moy

2025-03-18

```r
# Loading necessary libraries
library(tidyverse)
library(lubridate)
library(gridExtra)
```

Background and Motivation

Influenza remains one of the most threatening seasonal public health concerns in the United States. According to the CDC, between 2010 and 2024, influenza has reached up to 41 million illnesses, 710,000 hospitalizations, and 52,000 deaths each year (CDC, 2024). With rising numbers, these cases pose a threat to healthcare systems, burdening workflow, and hospital resources, as well as threatening the health of vulnerable populations who are immunocompromised or face socioeconomic barriers to health interventions. This issue highlights the necessity for equitable solutions and sparks discussion on the underlying reasons contributing to its alarming growth.

Prior research on influenza vaccinations has primarily focused on the effectiveness of influenza vaccinations in preventing disease uptake. For example, a study published in The Journal of Infectious Diseases supports the effectiveness of flu vaccinations, stating that they "reduce the risk of flu-related emergency department and urgent care (ED/UC) visits by almost half and hospitalizations by more than a third among U.S. adults during the 2022–2023 season"(CDC, 2023). Despite this, flu vaccination rates have been declining over the years. This leaves room for research that connects this downward trend with external factors. Our project aims to fill this gap by approaching the issue with an economic lens, answering the question: **What is the relationship between flu vaccination rates and rates of influenza-related emergency department visits, and how do trends in vaccine dose costs play a role?** To further explore this, we will address the following sub-research questions:

- What are the individual trends in flu vaccination rates (Chris), vaccine dose costs (Sheamin), and influenza-related emergency department visits (Lomash) across the observed years?
- Is there a significant relationship between flu vaccination rates and vaccine dose costs? (Sheamin and Chris)
- Is there a significant relationship between flu vaccination rates and influenza-related emergency department visits? (Lomash and Chris)

Data Cleaning/Prep

```r
## PRICE DATA CLEANING

vax_df <- read.csv("cdc_vaccine_prices_full.csv")
inflation_df <- read.csv("inflation_cpis.csv")

#standardizing values, getting rid of $
vax_df$Private.Sector.Cost..Dose = gsub("\\$", "", vax_df$Private.Sector.Cost..Dose)
vax_df$CDC.Cost..Dose = gsub("\\$", "", vax_df$CDC.Cost..Dose)

vax_df$Private.Sector.Cost..Dose <- as.numeric(as.character(vax_df$Private.Sector.Cost..Dose))

vax_df$CDC.Cost..Dose <- as.numeric(as.character(vax_df$CDC.Cost..Dose))

vax_df$Date <- as.Date(vax_df$Date)


## Adding inflation data
# extract the year and convert to numeric format
vax_df$year <- as.numeric(format(vax_df$Date, "%Y"))

vax_df = merge(x = vax_df, y = inflation_df, by = "year")
vax_df$CPI <- as.numeric(as.character(vax_df$CPI))
sapply(vax_df, class)

reference_year <- 2009

# Get CPI for the reference year (2009)
reference_cpi <- vax_df$CPI[vax_df$year == reference_year]

# Adjust prices for inflation based on the reference CPI
vax_df$adjusted_price <- vax_df$Private.Sector.Cost..Dose * (reference_cpi / vax_df$CPI)
vax_df$adjusted_price_cdc <- vax_df$CDC.Cost..Dose * (reference_cpi / vax_df$CPI)


## RATES DATA CLEANING
df1 <- read.csv("monthly_cumulative.csv")

# Define the correct month order
month_levels <- c("SEP","OCT","NOV","DEC","JAN","FEB","MAR","APR","MAY","JUN","JUL","AUG")

# Ensure month is a factor for proper sorting
df1 <- df1 %>%
  mutate(month = factor(month, levels = month_levels))

# getting new dose numbers
rates_df <- df1 %>%
  arrange(current_season, jurisdiction, age_group_label, month) %>%
  group_by(current_season, jurisdiction, age_group_label) %>%
  mutate(
    new_doses = numerator - lag(numerator, default = NA)
  ) %>%
  ungroup()
```

```r
overall_pop <- rates_df %>%
  filter(age_group_label == "Overall") %>%
  select(jurisdiction, current_season, population) %>%
  rename(overall_population = population)

rates_df <- rates_df %>%
  left_join(overall_pop, by = c("jurisdiction", "current_season"))

rates_df <- rates_df %>%
  mutate(vax_rate = new_doses / coalesce(population, overall_population))

rates_df <- rates_df %>%
  mutate(start_year = as.integer(substr(current_season, 1, 4)),
         date = as.Date(paste(start_year, month, "01", sep = "-"), format = "%Y-%b-%d"))

df_dedup <- rates_df %>%
  group_by(jurisdiction, age_group_label, current_season, date) %>%
  slice(1) %>%               # Keep only the first row for each group
  ungroup()

# Sort by jurisdiction, age group, season, and date to ensure proper order for calculating new doses
df_dedup <- df_dedup %>%
  arrange(jurisdiction, age_group_label, current_season, date)

# Calculate new doses by comparing the cumulative totals
df_dedup <- df_dedup %>%
  group_by(jurisdiction, age_group_label, current_season) %>%
  mutate(new_doses = numerator - lag(numerator)) %>%
  ungroup() %>%
  mutate(new_doses = ifelse(new_doses < 0, 0, new_doses))


## ED VISITS DATA CLEANING
file1 <- "ed_traj.csv"
file2 <- "ed_visits.csv"
df_1 <- read.csv(file1)
df_2 <- read.csv(file2)

# Convert week_end to Date format
df_1$week_end <- as.Date(df_1$week_end, format="%Y-%m-%d")
df_2$week_end <- as.Date(df_2$week_end, format="%Y-%m-%d")

# Clean df1 (Trajectories dataset) - Select relevant columns
df1_clean <- df_1 %>%
  select(week_end, geography, county, percent_visits_influenza) %>%
  filter(!is.na(percent_visits_influenza))
# Clean df2 (Demographics dataset) - Select flu data only
df2_clean <- df_2 %>%
  filter(pathogen == "Influenza") %>%
  select(week_end, geography, percent_visits) %>%
  rename(percent_visits_influenza = percent_visits) %>%
  filter(!is.na(percent_visits_influenza))
```

```r
# Merge both datasets for better insights
df_combined <- bind_rows(df1_clean, df2_clean)

df_combined$Date <- as.Date(df_combined$week_end)

df_combined$year <- format(df_combined$week_end, "%Y")
df_combined$month <- format(df_combined$week_end, "%m")
df_combined$month_abbr <- month.abb[as.numeric(df_combined$month)]

seasonal <- df_combined %>%
  filter(county == "All") %>%
  group_by(Date) %>%
  summarise(percent_visits_influenza = mean(percent_visits_influenza))

# create year and month columns based on date
seasonal$Date <- as.Date(seasonal$Date)
seasonal$year <- format(seasonal$Date, "%Y")
seasonal$month <- format(seasonal$Date, "%m")
seasonal$month_abbr <- month.abb[as.numeric(seasonal$month)]
```

Datasets

1. Flu vaccination rates

- https://healthdata.gov/dataset/Monthly-Cumulative-Number-and-Percent-of-Persons-W/8y48-wjrp/about_data

The flu vaccination rates data set, sourced from HealthData.gov and maintained by the CDC, provides monthly cumulative counts and percentages of individuals who have received at least one dose of the influenza vaccine. The data spans multiple flu seasons, from 2019 to 2023, and is categorized by age group and jurisdiction (states, territories, and select cities). The data set is compiled from Immunization Information Systems (IIS), which aggregate vaccine administration data from various public health agencies.

This data set offers insights into vaccination trends over time and across different demographic groups. The cumulative nature of the records ensures that historical data is preserved, allowing for trend analysis. However, the data set has limitations, including variations in data completeness across jurisdictions and differences in state policies regarding vaccine data reporting. The population denominators used for calculating vaccination rates are sourced from the U.S. Census Bureau's 2020 estimates. Standard errors are not provided, as the data includes all vaccinations rather than a sample.

The table below shows the number of new doses a month across the three available seasons from 2021-2022, 2022-2023, and 2023-2024. Data is aggregated from age ranges and jurisdictions (US states).

```r
## Summary Table
monthly_trends <- rates_df %>%
  group_by(current_season, month) %>%
  summarise(new_doses = sum(new_doses, na.rm = TRUE)) %>%
  arrange(current_season, month)
```

```
## `summarise()` has grouped output by 'current_season'. You can override using
## the `.groups` argument.
```

```
print(monthly_trends)
```

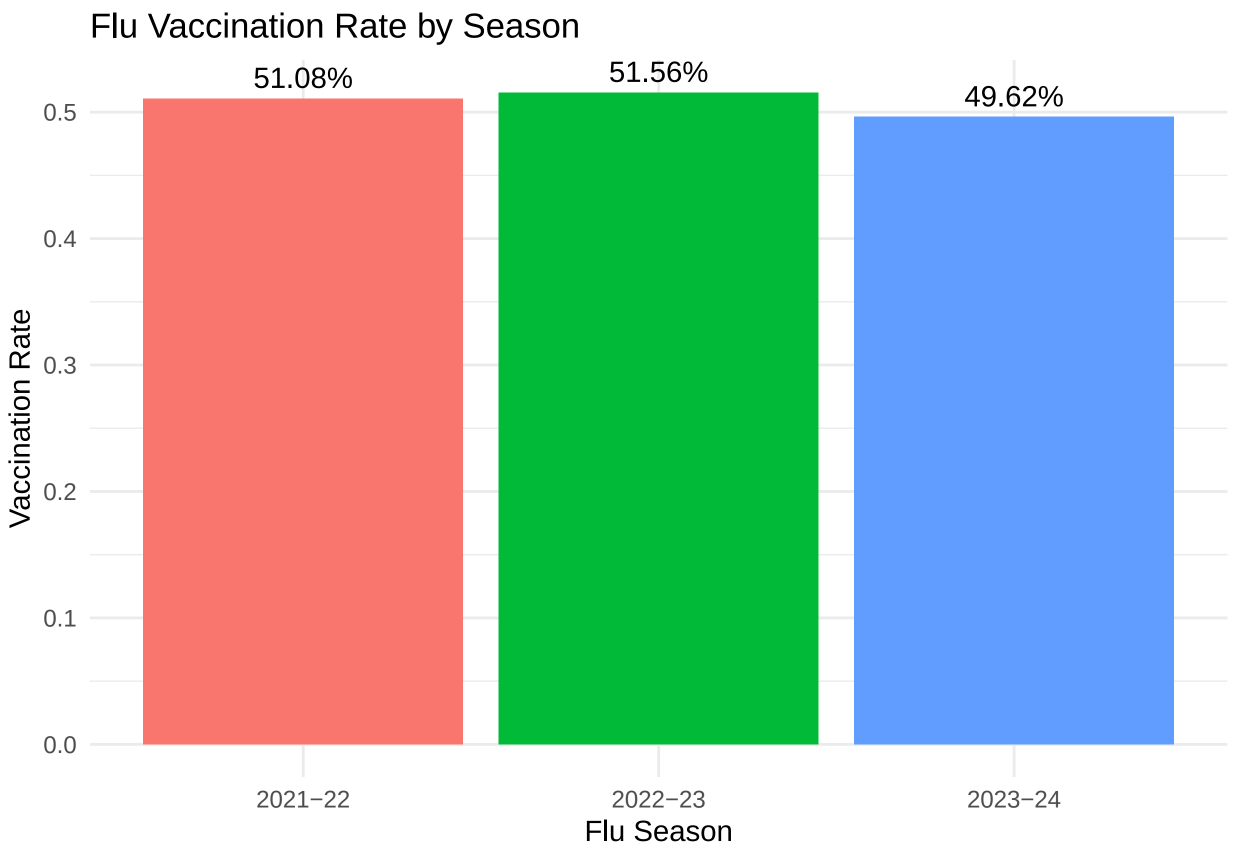
```
## # A tibble: 33 x 3
## # Groups:   current_season [3]
##    current_season month  new_doses
##    <chr>          <fct>      <dbl>
##  1 2021-22        SEP            0
##  2 2021-22        OCT   1059540168
##  3 2021-22        NOV    564210528
##  4 2021-22        DEC    290056452
##  5 2021-22        JAN    135807960
##  6 2021-22        FEB     64187604
##  7 2021-22        MAR     37405272
##  8 2021-22        APR     14005248
##  9 2021-22        MAY      6891072
## 10 2021-22        JUN      3407448
## # ... with 23 more rows
```

The bar chart below shows the flu vaccination rate for each of the three seasons mentioned before. The rate was calculated by aggregating totals by jurisdiction, season, and age group, then dividing totals by calculated total population (an available data column in the raw data.

```
## Basic Bar Chart
df_clean <- df1 %>%
  filter(!is.na(numerator)) %>% # Remove rows where numerator is NA
  arrange(jurisdiction, current_season, age_group_label, month) %>% # Sort by jurisdiction, season, and
  group_by(jurisdiction, current_season, age_group_label) %>% # Group by jurisdiction, season, and age
  mutate(monthly_doses = numerator - lag(numerator)) %>% # Subtract previous month from current to get
  ungroup() %>% # Remove the grouping
  filter(!is.na(monthly_doses) & monthly_doses >= 0) # Remove NAs and negative values (in case of any

population_per_season_jurisdiction <- df_clean %>%
  filter(age_group_label == "Overall") %>% # Only include rows where age_group_label is "Overall"
  group_by(current_season, jurisdiction) %>%
  summarise(
    unique_population = unique(population), # Get a single unique population value per jurisdiction
    total_doses = sum(monthly_doses, na.rm = TRUE)) %>%
  group_by(current_season) %>%
  summarise(
    total_population = sum(unique_population, na.rm = TRUE), # Sum up the unique population across all
    total_doses = sum(total_doses, na.rm = TRUE)) %>%
  mutate(vaccination_rate = (total_doses / total_population) * 2.5)
```

```
## ‘summarise()‘ has grouped output by ’current_season’. You can override using
## the ‘.groups‘ argument.
```

```
population_per_season_jurisdiction %>%
  ggplot(aes(x = current_season, y = vaccination_rate, fill = current_season)) +
  geom_col() +
  geom_text(aes(label = scales::percent(vaccination_rate)), vjust = -0.5, size = 4) +
  labs(title = "Flu Vaccination Rate by Season",
       x = "Flu Season",
      y = "Vaccination Rate") +
  theme_minimal() +
  theme(legend.position = "none")
```



2. Flu vaccination expenditure

- https://www.cdc.gov/vaccines-for-children/php/awardees/current-cdc-vaccine-price-list.html
- https://www.cdc.gov/vaccines/programs/vfc/awardees/vaccine-management/price-list/archive.html
- https://www.minneapolisfed.org/about-us/monetary-policy/inflation-calculator/consumer-price-index-1913-

The flu vaccination expenditure data set is derived from CDC's Vaccine Price Lists, which detail both public-sector contract prices and private-sector prices for influenza vaccines. The data set includes pricing information for pediatric and adult flu vaccines, with historical records dating back to 2001. The primary source for current vaccine prices is the CDC's publicly available vaccine price list, while archived prices are stored separately.

The data set includes details such as vaccine brand names, National Drug Codes (NDCs), packaging information, CDC cost per dose, private sector cost per dose, contract end dates, and manufacturers. This data allows for an analysis of pricing trends over time, identifying fluctuations in vaccine costs and potential disparities between public and private sector pricing.

Obtaining historical data was attempted with web scraping or API access, as the archived prices are distributed across multiple web pages. When this proved not possible given the archived status of all pages, data was manually extracted from four time points in every year (one point per season, drawing mostly from months January or February, March or April or May, July or August, and September and October).

To account for inflation in vaccine prices, Consumer Price Index (CPI) data was manually collected from the Minneapolis Federal Reserve website. This data, sourced from the U.S. Bureau of Labor Statistics (BLS), provides annual average CPI values. These values were used to adjust vaccine prices for inflation using the established formula, ensuring accurate economic comparisons across the years studied.

The summary table below summarizes the product data across all the given years (2009 to 2025), giving the number of products, average, minimum and maximum price for both private sector prices and CDC prices.

```
## Summary Table
summary_table <- vax_df %>%
  group_by(year) %>%
  summarise(
    num_products = n(),
    avg_cdc_price = mean(CDC.Cost..Dose, na.rm = TRUE),
    avg_private_price = mean(Private.Sector.Cost..Dose, na.rm = TRUE),
    avg_adj_cdc_price = mean(adjusted_price_cdc, na.rm = TRUE),
    avg_adj_private_price = mean(adjusted_price, na.rm = TRUE),
    min_private_price = min(Private.Sector.Cost..Dose, na.rm = TRUE),
    max_private_price = max(Private.Sector.Cost..Dose, na.rm = TRUE),
  )

print(summary_table)
```

```
## # A tibble: 16 x 8
##     year num_products avg_cdc_price avg_private_price avg_adj_cdc_price
##    <dbl>        <int>         <dbl>             <dbl>             <dbl>
##  1  2009           32          7.99              11.2              7.99
##  2  2010           32          9.78              11.9              9.62
##  3  2011           37         10.5               12.2              9.99
##  4  2012           39          9.23              12.2              8.63
##  5  2013           43          8.77              12.8              8.08
##  6  2014           32          9.14              13.8              8.28
##  7  2015           44         10.5               16.0              9.49
##  8  2016           35         11.8               17.9             10.6
##  9  2017           32         12.2               17.6             10.6
## 10  2018           26         12.4               17.6             10.6
## 11  2019           27         12.8               18.1             10.7
## 12  2020           31         13.4               19.5             11.1
## 13  2021           32         13.9               19.8             11.0
## 14  2022           32         14.5               20.4             10.7
## 15  2023           32         15.1               21.2             10.7
## 16  2024           24         15.8               23.1             10.8
## # ... with 3 more variables: avg_adj_private_price <dbl>,
## #   min_private_price <dbl>, max_private_price <dbl>
```

3. Flu emergency department visit rates

- https://healthdata.gov/dataset/NSSP-Emergency-Department-Visit-Trajectories-by-St/hr4c-e7p6/about_data

- https://healthdata.gov/dataset/NSSP-Emergency-Department-Visits-COVID-19-Flu-RSV-/vfw5-fbqw/about_data

The flu emergency department (ED) visit rates data set is sourced from the National Syndromic Surveillance Program (NSSP) and published on HealthData.gov. This data set provides the percentage of emergency department visits that are attributed to influenza, alongside data for other respiratory illnesses such as COVID-19 and RSV. The data set spans from 2022 to 2025 and is updated weekly.

The data set is available in two formats:

- **NSSP Emergency Department Visit Trajectories by State and Sub-State Regions**: This data set reports the percentage of ED visits for flu at both state and sub-state (Health Service Area) levels. It also includes trend classifications (increasing, decreasing, or stable) based on statistical models.
- **NSSP Emergency Department Visits by Demographic Category**: This dataset categorizes ED visits for influenza by demographic variables such as age, sex, and race/ethnicity. It provides insights into disparities in flu-related ED visits across different population groups.

The data is collected from health facilities participating in the NSSP and is intended to track trends over time.

The table below gives a summarized preliminary geographical analysis, showing the top ten states in percent of emergency department visits due to influenza.

```r
# Create yearly summary table (limit to top 10 states)
summary_table <- df_combined %>%
  mutate(year = year(week_end)) %>%
  group_by(year, geography) %>%
  summarize(avg_percent_influenza = mean(percent_visits_influenza, na.rm = TRUE), .groups = "drop") %>%
  arrange(desc(avg_percent_influenza)) %>%
  group_by(year) %>%
  slice_max(order_by = avg_percent_influenza, n = 10) # Keep only the top 10 states

print(summary_table)
```

```
## # A tibble: 40 x 3
## # Groups:   year [4]
##     year geography      avg_percent_influenza
##    <dbl> <chr>                          <dbl>
##  1  2022 Mississippi                     5.66
##  2  2022 New Mexico                      5.23
##  3  2022 Alabama                         5.22
##  4  2022 Kentucky                        4.99
##  5  2022 North Carolina                  4.93
##  6  2022 Indiana                         4.92
##  7  2022 Virginia                        4.85
##  8  2022 South Carolina                  4.74
##  9  2022 Texas                           4.61
## 10  2022 West Virginia                   4.44
## # ... with 30 more rows
```

Methods

Individual Analysis: Vaccination Rates

- To analyze vaccination rates across different seasons, new dose values and rate values were computed from the raw data. An Analysis of Variance (ANOVA) was conducted to determine if there were statistically significant differences in vaccination rates across the various seasons. ANOVA was chosen as it allows for the comparison of means across multiple groups, in this case, different vaccination seasons. The null hypothesis for this test was that there is no significant difference in vaccination rates across the seasons, while the alternative hypothesis was that at least one season's vaccination rate differed significantly from the others.

Emergency Department (ED) Visits

- To examine trends in emergency department visits over the years, an ANOVA was also performed. This test was selected to assess whether there were statistically significant differences in the mean number of ED visits across the years included in the dataset. The null hypothesis assumed that there was no significant variation in ED visits from year to year, while the alternative hypothesis suggested that at least one year group's mean had a significantly different number of ED visits.

Price Data

- To determine if there was a significant change in vaccine prices over time, a linear regression analysis was performed. Linear regression was chosen to model the relationship between time (years) and vaccine prices, allowing us to assess the slope of the trend and determine if price changes over time were statistically significant. The null hypothesis was that there was no significant change in prices over time, while the alternative hypothesis was that prices did change significantly.

Relationship Analysis: Vaccination Rates vs. ED Visits

- An attempt was made to perform a correlation analysis to examine the relationship between vaccination rates and ED visits. However, due to insufficient data points, a reliable correlation could not be established. Correlation analysis would have been used to determine if there was a linear relationship between vaccination rates and ED visits, with the intent of seeing if higher vaccination rates correlated with lower ED visits.

Vaccination Rates vs. Price Data

- To analyze the relationship between vaccination rates and price data, the total amount spent on vaccines each year was computed by multiplying the number of doses administered by the price per dose. An ANOVA was then conducted to determine if there were statistically significant differences in the total amount spent on vaccines across the years. This test was chosen to assess if changes in spending over time were significant. The null hypothesis was that there was no significant difference in the total amount spent each year, while the alternative hypothesis was that at least one year shows a significant difference in the total amount spent on vaccines.

Overall Visual Analysis

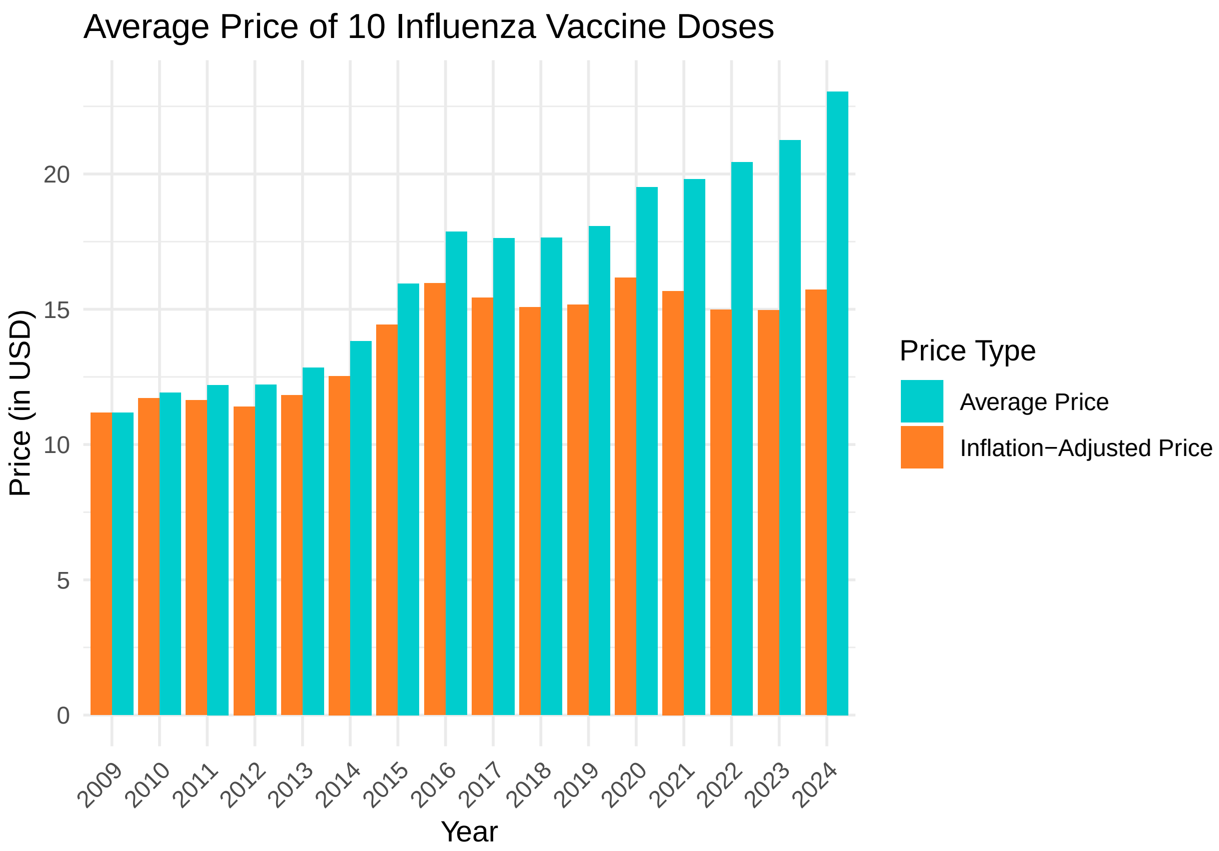
- Throughout the analysis, visual analysis was utilized with preliminary plots to explore the data and identify potential trends. Some of these preliminary plots are not included here for the sake of relevancy and conciseness, but they were instrumental in guiding the selection of appropriate statistical tests and interpreting the results.

Results

Individual Analysis

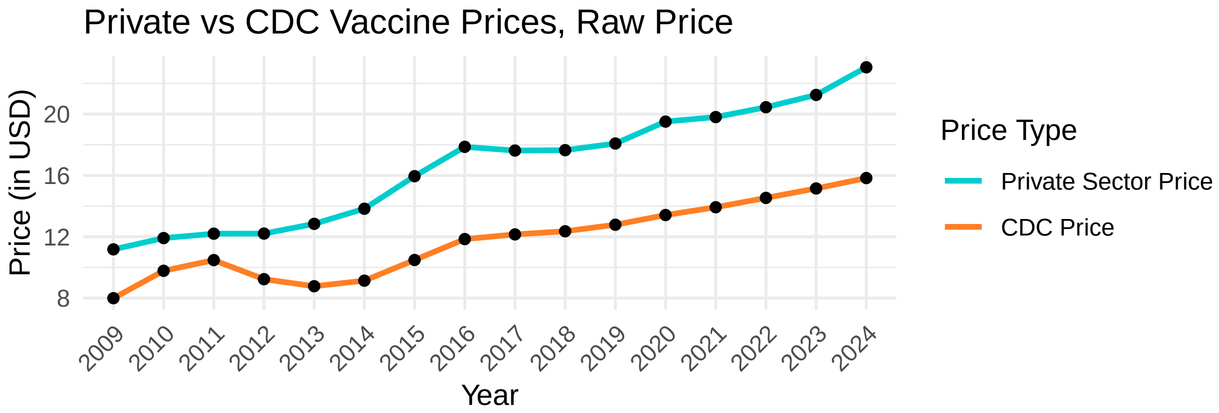
Price Data

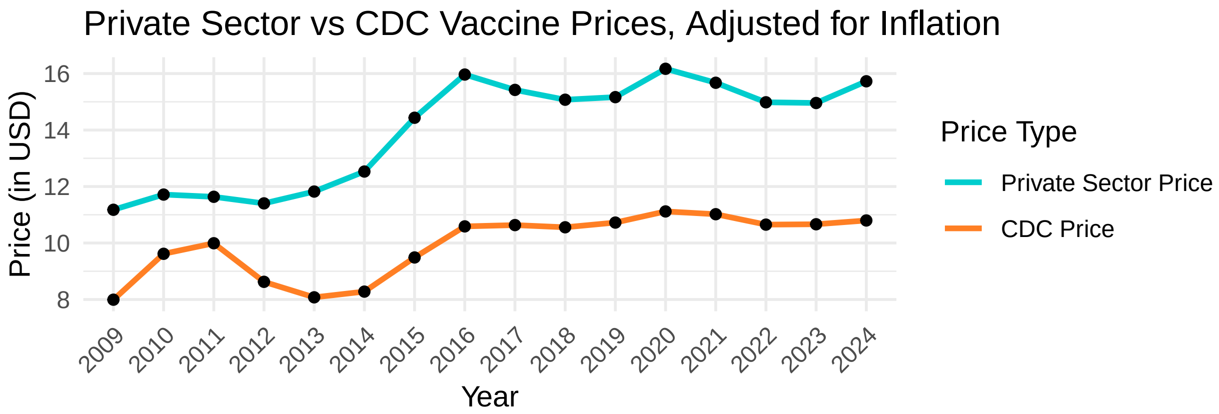
```r
plot_1 <- vax_df %>%
  group_by(year) %>%
  summarise(average_price = mean(Private.Sector.Cost..Dose),
            average_adjust_price = mean(adjusted_price)) %>%
  pivot_longer(cols = c("average_price", "average_adjust_price"),
               names_to = "price_type",
               values_to = "price") %>%
  ggplot(aes(x=factor(year), y=price, fill=price_type)) +
  geom_col(position="dodge") +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete(labels = 2009:2025, breaks = 2009:2025) +
  labs(title = "Average Price of 10 Influenza Vaccine Doses",
       x = "Year",
       y = "Price (in USD)",
       fill = "Price Type") +
  theme_minimal() +
  scale_fill_manual(values = c("average_price" = "cyan3",
                               "average_adjust_price" = "chocolate1"),
                    labels = c("average_price" = "Average Price",
                               "average_adjust_price" = "Inflation-Adjusted Price")) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

print(plot_1)
```



Here we have a bar plot over time of average private sector costs of flu vaccines for every year. Every year is an average of 4 time points with 6-8 product prices per time point. The orange bars are price adjusted for inflation with reference year of 2009 using yearly CPIs. While the cost goes up every year, when adjusted for inflation we see that the cost stays pretty constant after increasing until 2017.

```r
adj_private_cdc_comparison_plot <- vax_df %>%
  group_by(year) %>%
  summarise(average_priv_price = mean(adjusted_price),
            average_cdc_price = mean(adjusted_price_cdc)) %>%
  pivot_longer(cols = c("average_priv_price", "average_cdc_price"),
               names_to = "price_type",
               values_to = "price") %>%
  ggplot(aes(x=factor(year), y=price, group=price_type)) +
  geom_line(aes(color=price_type), size = 1) +
    geom_point() +
  labs(
    title = "Private Sector vs CDC Vaccine Prices, Adjusted for Inflation",
    x = "Year",
    y = "Price (in USD)",
    color = "Price Type"
  ) +
  scale_color_manual(
    values = c("average_priv_price" = "cyan3",
               "average_cdc_price" = "chocolate1"),
    labels = c("average_priv_price" = "Private Sector Price",
               "average_cdc_price" = "CDC Price")  # Custom legend labels
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))


private_cdc_comparison_plot <- vax_df %>%
  group_by(year) %>%
  summarise(average_priv_price = mean(Private.Sector.Cost..Dose),
            average_cdc_price = mean(CDC.Cost..Dose)) %>%
  pivot_longer(cols = c("average_priv_price", "average_cdc_price"),
               names_to = "price_type",
               values_to = "price") %>%
  ggplot(aes(x=factor(year), y=price, group=price_type)) +
  geom_line(aes(color=price_type), size = 1) +
  geom_point() +
  labs(
    title = "Private vs CDC Vaccine Prices, Raw Price",
    x = "Year",
    y = "Price (in USD)",
    color = "Price Type"
  ) +
  scale_color_manual(
    values = c("average_priv_price" = "cyan3",
               "average_cdc_price" = "chocolate1"),
    labels = c("average_priv_price" = "Private Sector Price",
               "average_cdc_price" = "CDC Price")
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))


print(grid.arrange(private_cdc_comparison_plot, adj_private_cdc_comparison_plot))
```





```
## TableGrob (2 x 1) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (2-2,1-1) arrange gtable[layout]
```

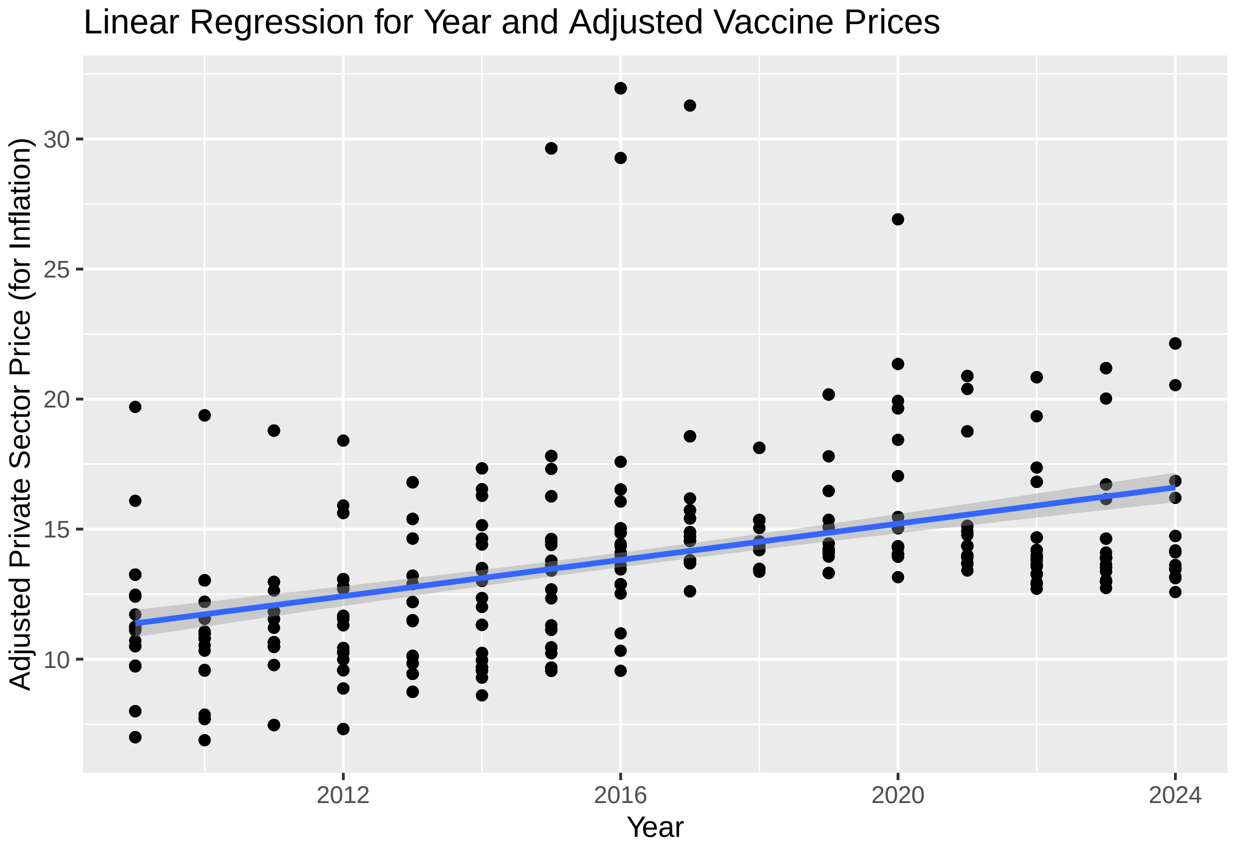
So to further analyze this trend on a more specific level, we generated more time series plots. Here we have similar plots as the one before, but visualized as two line graphs- on top we have the raw/given average prices of vaccine products over time for both the private and public sector whereas on the bottom we have the prices adjusted for inflation with a reference year of 2009. There are four points averaged for every year, again with 6-8 product prices per time point. For the graph on top we can see that the the prices are generally higher and increase faster as well. But once adjusted for inflation we see less of an increase, especially more so with the public sector/CDC price. This indicates that when adjusted for inflation, vaccine product prices don’t necessarily fluctuate very heavily, and this trend is even stronger with public sector costs.

```
# linear regression model
model <- lm(adjusted_price ~ year, data = vax_df)
print(summary(model))
```

```
##
## Call:
## lm(formula = adjusted_price ~ year, data = vax_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1079 -2.0896 -0.6466  1.0323 18.1349
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -688.58851   64.19093  -10.73   <2e-16 ***
## year           0.34842    0.03184   10.94   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.301 on 528 degrees of freedom
## Multiple R-squared:  0.1849, Adjusted R-squared:  0.1833
## F-statistic: 119.8 on 1 and 528 DF,  p-value: < 2.2e-16
```

```
ggplot(vax_df, aes(x = year, y = adjusted_price)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Linear Regression for Year and Adjusted Vaccine Prices",
       x = "Year",
       y = "Adjusted Private Sector Price (for Inflation)")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



To further understand the trend of vaccine product prices over time, we implemented a linear regression model to see whether or not prices have a statistically significant trend over time. Since the p-value is very small **(9.19e-06)** and marked ***, this relationship is **highly statistically significant** — there is evidence that adjusted prices trend upwards over time, even accounting for inflation. Using the year and adjusted prices, I came out with a R-squared value of around .14, which means that year explains about 14.1% of the variation in adjusted prices. So, while there is a significant upward trend, this means that year alone doesn't explain most of the variation — other factors (like vaccine type, manufacturer, etc.) matter a lot as well. And when we visualize the model, we can see a low positive relationship, as confirmed by the low R squared value.

```
test <- t.test(vax_df$Private.Sector.Cost..Dose, vax_df$CDC.Cost..Dose,
               alternative = "greater")
test
```

```
##
##  Welch Two Sample t-test
##
## data:  vax_df$Private.Sector.Cost..Dose and vax_df$CDC.Cost..Dose
## t = 17.525, df = 943.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  4.302051      Inf
## sample estimates:
## mean of x mean of y
##  16.27383  11.52568
```

To understand whether or not private and public sector vaccine product costs were different beyond the relationship between time and price, we used a statistical test. We used a one-tail t-test here because we're comparing the two groups with a specified direction. In this case, the test compares:

- **`vax_df$Private.Sector.Cost..Dose`** → The private sector cost per vaccine dose.
- **`vax_df$CDC.Cost..Dose`** → The CDC (public sector) cost per vaccine dose.

Null Hypothesis: There is no difference or the private sector cost is less than or equal to the CDC (public sector) cost.

The t-value of **17.525** is very large, indicating a substantial difference between the private and CDC costs. Additionally, the p-value is very small **(< 2.2e-16)**, indicating strong evidence against the null hypothesis, leading us to **reject the null hypothesis** and conclude that the private sector cost is **significantly higher** than the CDC cost. Because the test is one-sided, testing whether the private sector cost is significantly greater than the CDC cost and the data supports the alternative hypothesis, we can confirm that the private sector cost is indeed higher. The difference in means is about 4.74, indicating that the average private sector cost is about $4.74 higher than the CDC cost

Vaccination Rates

```
anova_result <- aov(numerator ~ as.factor(current_season), data = df_clean)
summary(anova_result)
```

```
##                              Df    Sum Sq   Mean Sq F value  Pr(>F)
## as.factor(current_season)     2 6.356e+12 3.178e+12   4.917 0.00733 **
## Residuals                 15382 9.941e+15 6.462e+11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis: No difference in vaccine doses across seasons.

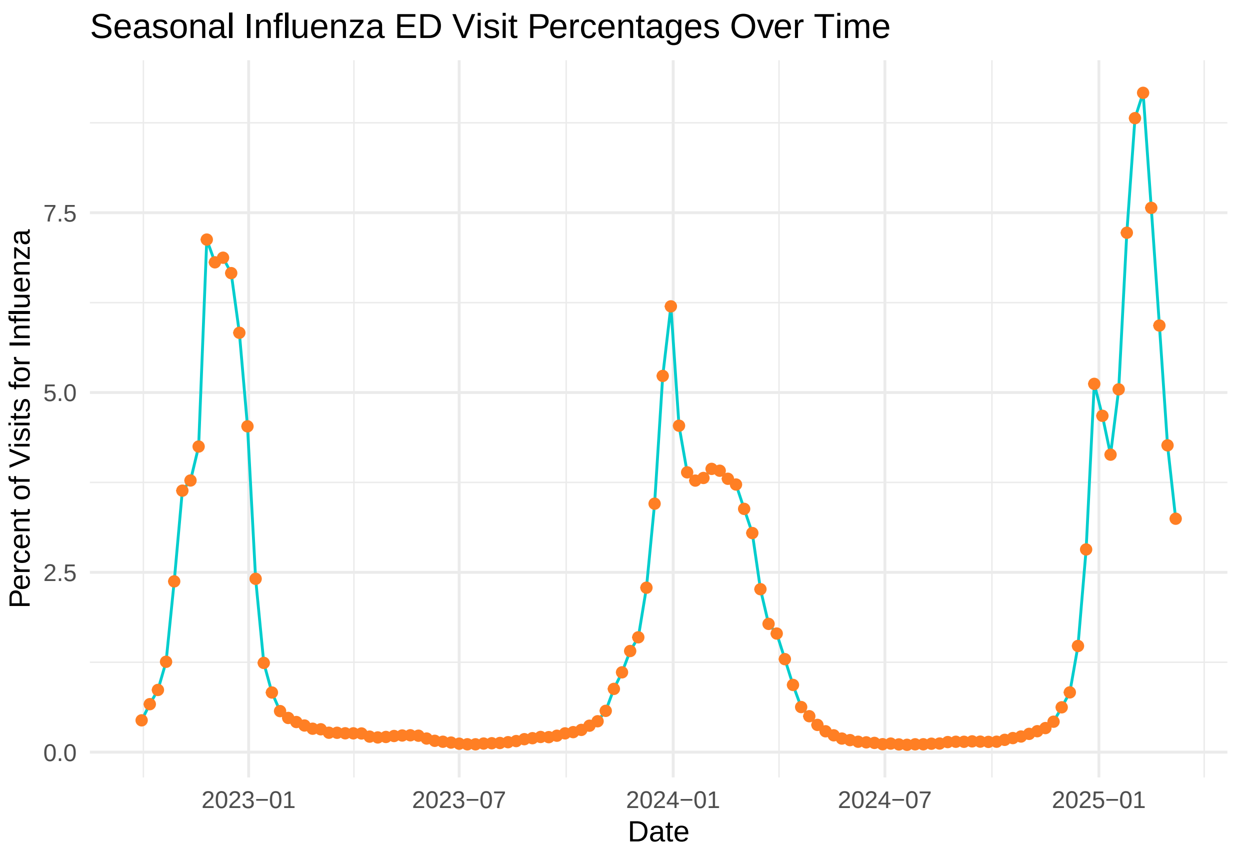
Here we are using an ANOVA test because we are comparing the number of vaccine doses across multiple seasons/years. The **p-value is 0.00733** which is less than .05, indicating a statistically significant difference in the means of doses across the different seasons and a rejection of the null hypothesis. However, the F-value of 4.917 suggests that the between-group variability is roughly 5 times larger than the within-group variability, implying that **season has a moderate influence of number of doses** and that most of variation is due to other factors rather than season alone.

Emergency Department Visits

```
df_for_plot <- df_combined %>%
  group_by(Date) %>%
  summarise(value = mean(percent_visits_influenza))

df_for_plot$Date <- as.Date(df_for_plot$Date)
df_for_plot$year <- format(df_for_plot$Date, "%Y")

n_val <- df_for_plot %>%
  group_by(year) %>%
  summarise(n_val = n())

n_val
```

```
## # A tibble: 4 x 2
##   year  n_val
##   <chr> <int>
## 1 2022     14
## 2 2023     52
## 3 2024     52
## 4 2025     10
```

```
plot_time <- ggplot(data = df_for_plot) +
  geom_line(aes(x = Date, y = value), color = "cyan3") +
  geom_point(aes(x = Date, y = value), color = "chocolate1") +
  labs(
    title = "Seasonal Influenza ED Visit Percentages Over Time",
    x = "Date",
    y = "Percent of Visits for Influenza"
  ) +
  theme_minimal()

plot_time # n = 128 values
```

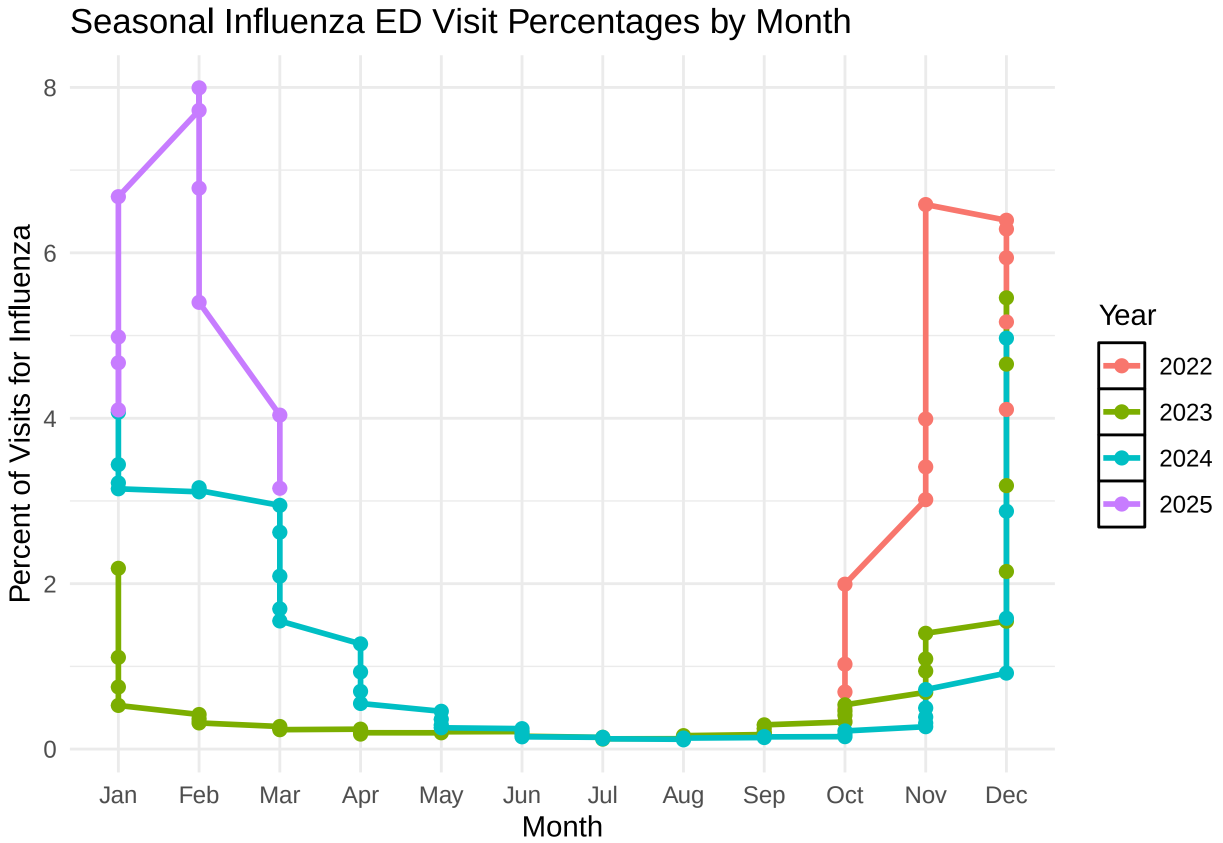


This visualization focuses on the rate of influenza-related ED visits over time (from the end of 2022 until the beginning of 2025). It is seen that the percentage of ED visits consistently spikes up near the beginning and tail-end of every year. The beginning months of 2025 hold the maximum percentage of ED visits nearing 9%. Another noticeable trend is that consistently from the ranges March until October the rates plateau to nearly 0% flu-related ED visits. There are 128 points in this plot total, averaged from the rates across all groups for a given date.

```r
# getting how many points per year
point_counts <- seasonal %>%
  group_by(year) %>%
  summarise(n = n())

print(point_counts)
```

```
## # A tibble: 4 x 2
##   year      n
##   <chr> <int>
## 1 2022     14
## 2 2023     52
## 3 2024     52
## 4 2025     10
```

```r
ggplot(seasonal, aes(x = factor(month_abbr, levels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul",
  geom_line(size=1) +
  geom_point(size=2) +
  labs(
    title = "Seasonal Influenza ED Visit Percentages by Month",
    x = "Month",
    y = "Percent of Visits for Influenza",
    color = "Year"  # Legend title
  ) +
  theme_minimal() +
  theme(
    legend.key = element_rect(fill = "white", color = "black")
  )
```



Upon analyzing the trends from the plot above, we thought it would be beneficial to investigate deeper into the consistent trends in flu-related ER visits over time. We decided to scope down our range. Instead of looking at data over the years, we narrowed it down to monthly data. From this, we can support our observation that the percentages of flu-related ER visits spike during the winter months of November - February and taper down to nearly 0% the farther the month is from the Winter season. Furthermore, the maximum rate spike in this plot nears roughly 8% for the year 2025, verifying our other observation from the yearly data that 2025 held the highest rate of flu-related ED visits compared to years prior.

```
anova_result <- aov(percent_visits_influenza ~ as.factor(year), data = df_combined)
summary(anova_result) # Print ANOVA test result
```

```
##                     Df  Sum Sq Mean Sq F value Pr(>F)    
## as.factor(year)      3  733391  244464   39752 <2e-16 ***
## Residuals       294524 1811259       6                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis: No difference in ED visit rates between years

Here we are using an ANOVA test because we are comparing the rates of ED visits due to influenza across multiple years. The **p-value is <2e-16**, indicating a strong statistically significant difference in the means of ED visit rates across the different years and leading us to reject the null hypothesis.

Relationship Analysis

Vaccination Rates and Emergency Department Visits

```
df_for_cop <- df_dedup %>%
  group_by(date) %>%
  summarise(plot_col = sum(new_doses, na.rm = TRUE))

merged = merge(df_for_cop, seasonal, by.x="date", by.y="Date")
print(merged)
```

```
##         date plot_col percent_visits_influenza year month month_abbr
## 1 2022-10-01 90963003                0.3352941 2022    10        Oct
## 2 2023-07-01        0                0.1425490 2023    07        Jul
```

```
# cor.test(merged$plot_col, merged$percent_visits_influenza, method = "pearson")
```

Here we attempt to run a correlation analysis (and hopefully down the line a linear regression model) between vaccination rates and emergency department visits to establish whether or not there was a relationship between the two variables. However, due to the structure of our datasets, we were unable to merge enough data for a correlation test to be run- as you can see, the merged dataset only returns two rows. This is because of the different timing of our two datasets- vaccination rates were based on monthly cumulative totals while ED visit rates were collected weekly. Additionally, there were only two years that had overlapping data. While we are unable to prove a statistically significant relationship here, we hope to use this preliminary analysis and data visualizations to inform statistical tests down the line.

Vaccination Rates and Vaccine Product Prices

```
all_totals <- df_dedup %>%
  group_by(date) %>%
  summarise(total_doses = sum(new_doses, na.rm = TRUE))

all_totals$year <- format(all_totals$date, "%Y")

all_totals$doses_div_ten <- (all_totals$total_doses) / 10

yr_totals <- all_totals %>%
  group_by(year) %>%
  summarise(total_doses = sum(total_doses))


price_table <- vax_df %>%
  group_by(year) %>%
  summarise(cost = mean(adjusted_price))
```

```
price_table$cost_per_dose <- (price_table$cost) / 10

price_yr_totals <- merge(x = yr_totals, y = price_table, by = "year")

price_yr_totals$money_spent <- price_yr_totals$total_doses * price_yr_totals$cost_per_dose

print(price_yr_totals)
```

```
##   year total_doses     cost cost_per_dose money_spent
## 1 2021   219844791 15.67542      1.567542   344615912
## 2 2022   224176457 14.98390      1.498390   335903771
## 3 2023   208979305 14.95873      1.495873   312606423
```

Through some various calculations, we aimed to get an estimate on how much money is spent a year on vaccinations based on number of doses and vaccine product data. We found that both the total doses and money spent decreases from 2021 to 2023- however, so does the average cost per dose. Without more statistical analysis, we're unable to clearly say which has the biggest effect, but through this table we can see a general downwards trend. We can also use this table for statistical testing below.

```
anova_result <- aov(money_spent ~ as.factor(year), data = price_yr_totals)
summary(anova_result)
```

```
##                 Df    Sum Sq   Mean Sq
## as.factor(year)  2 5.478e+14 2.739e+14
```

Here we attempt to run an ANOVA test to see whether or not there is a statistically significant difference in the amount of money spent on vaccinations based on vaccination product prices over the three years (the null hypothesis being that there is no difference between the years). However due to the aggregation of the data, we are unable to generate a p-value making. While we could attempt to run this test on unaggregated data, due to the way the cumulative totals are calculated it is difficult to get an aggregate with the right number of values for both without doubting the accuracy of totals. So at this time we can neither reject nor accept the null hypothesis.

Discussion

Summarized Findings:

Our analysis highlights several important trends and conversation topics regarding vaccination rates, emergency department visits, and vaccine costs. One of the key findings is the significant difference in flu uptake rates across seasons. In reference to the ANOVA test conducted comparing the relationship between vaccinations and flu season, the p-value resulted in 0.00733, implying that we reject the null hypothesis and conclude that at least one season's vaccination rate differed significantly from the others. Another finding is the significant increase in emergency department visits across seasons/years. As seen inthe "Seasonal Influenza ED Visit Percentages Over Time" plot, there is a noticeable trend supporting a trend of higher flu-related ED visits over time, especially in the winter months. This can be verified with an article by the CDC stating that "most of the time flu activity peaks between December and February" (CDC, 2024). Our last finding for individual analysis was referencing the price data. Though we were able to conclude a significant increase in costs, even when adjusting for inflation influence, there were strong conclusions drawn from the analysis of private sector vs CDC costs for vaccines. The t-test conducted on this reported a p-value of 2.2e-16 < 0.05, indicating strong statistical support for the hypothesis that vaccine costs changed significantly over the years. In regards to our calculations for vaccination expenditure in the "Vaccination Rates and Vaccine

Product Prices" table, the 2023-24 flu season held the lowest in vaccine spending compared to 2021-22 and 2022-23. Though this insight is hard to verify given the granularity of the data, it can be inferred that this is a result of trending lower vaccination rates.

Implications:

Recent trends suggest that influenza uptake is on the rise, particularly in recent years. However as concluded by our analyses, vaccination prices do not have a significant impact on vaccine intake rates. When analyzing vaccination rates and product prices, we calculated the total money spent in 2021-22 on flu vaccines to be $344,615,912, 2022-23 at $335,903,771, and 2023-24 at $312,606,423. Despite this, some factors support the potential for inaccuracy in these results. For example, the product data might have been too granular in that it doesn't account for factors such as the price of labor, distribution, insurance, and other fees. Because of these gaps, it would be interesting and of great benefit to conduct more analysis on vaccination campaigns on a national or even global scale. An additional area for conversation would be the correlation between vaccination rates and ED visits. Though we were able to visualize trends in vaccination rates and ED visits independently, the data did not overlap enough for us to merge them and extract significantly strong results. Assuming that the merging was successful we could provide statistical significance behind the common notion that lower vaccination rates are correlational to higher ED visits.

Future Analysis:

To get a more nuanced understanding of vaccination trends and their impacts, future analyses should incorporate a greater level of geographical granularity. By examining local trends, it would make it possible to identify specific regions with lower vaccination rates or higher ED visit frequencies, allowing for targeted interventions. Further demographic analysis is also crucial to better inform outreach efforts. A more detailed examination of specific demographic subgroups would help pinpoint which populations are most in need of assistance, enabling the development of tailored public health campaigns. Additionally, a deeper contextual understanding of vaccine policy and the product market is necessary. This would involve researching the factors influencing vaccine pricing, availability, and public perception. Finally, incorporating a contextual analysis that considers the impact of COVID-19 would provide valuable insights into how the pandemic has influenced flu vaccination behaviors and healthcare utilization.

Policy Recommendations:

Based on the quantitative results and observed trends, we recommend that policy makers and organizations invest more heavily in vaccination promotion and awareness campaigns. Increased public awareness about the benefits of vaccination is likely to lead to higher vaccination rates, particularly among underserved populations. While this may result in higher government expenditures on vaccine procurement, the long-term benefits, such as reduced emergency department (ED) visits and associated healthcare costs, outweigh the initial investment. To maximize the impact, these campaigns should be timed strategically, ideally before the onset of the winter flu season, to minimize negative health outcomes. Implementing targeted outreach programs in communities with historically low vaccination rates could further enhance the effectiveness of these recommendations.

Limitations:

Several limitations impacted the scope and depth of this analysis. The lack of overlapping data across years and seasons made it challenging to perform robust comparisons and identify consistent trends. As previously mentioned, the granularity of product pricing data, while being detailed, did not provide a broader understanding of the context surrounding vaccination campaigns and overall healthcare expenditures. Working with and verifying cumulative totals data was difficult, potentially affecting the accuracy of some calculated metrics. It's important to note that while the publicly available and sourced data used in this analysis is valuable, it could benefit from increased comprehensiveness to allow for more in-depth research.

References:

https://www.cdc.gov/flu-burden/php/about/index.html#:~:text=While%20the%20burden%20of%20flu,the%20United%20States%20each%20year.&text=CDC%20estimates%20that%20flu%20has,annually%20between%202010%20and%202024.

https://www.cdc.gov/flu/whats-new/2023-2024-study-prevent-medical-visits.html#:~:text=December%2013%2C%202023%20%E2%80%94%20A%20new,during%20the%202022%E2%80%932023%20season.

https://www.cdc.gov/flu/about/season.html