PM 566 Midterm Project

AUTHOR
Ziquan 'Harrison' Liu

Requried packages

```
library(tidyverse)  # dplyr, ggplot2, readr, string tools, plotting
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.1     ✔ stringr   1.5.2
✔ ggplot2   4.0.0     ✔ tibble    3.3.0
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.1.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to
become errors
```

```
library(dplyr)      # explicit for pipes and data wrangling
library(ggplot2)    # ggplot plotting
library(readr)      # fast/read_csv
library(stringr)    # string cleanup / regex
library(lubridate)  # work with dates (for stock update timestamps)
library(knitr)      # nice summary tables via kable()
library(kableExtra) # styling kable() for report-quality tables
```

```
Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

    group_rows
```

```
library(maps)       # state map polygons for U.S. choropleth
```

```
Attaching package: 'maps'

The following object is masked from 'package:purrr':

    map
```

```
library(countrycode) # optional helper if we ever need state mapping
library(forcats)    # reorder factors for nicer plots (top providers)
```

```
library(patchwork)    # to combine multiple ggplots into one figure grid
library(ggcorrplot)   # nice correlation heatmaps; install.packages("ggcorrplot") if miss
```

1. Introduction

Description of Dataset: The dataset I utilzied was a national registry of influenza (flu) vaccination provider locations published by the U.S. Centers for Disease Control and Prevention (CDC) through the Vaccines.gov platform. The dataset contained point-level information on health care sites (e.g., pharmacies, clinics, health departments) that report offering influenza vaccination to the public. The dataset was designed to support public-facing vaccine finder tools and programmatic monitoring of geographic access to flu vaccination. There are 28 colomns and 202652 observations included in this dataset. In other words, 202,652 locations across 53 state or territorial codes. Temporal stamps reflect inventory/operational updates from 2023-08-17 through 2024-08-01.

Primary Question: How are flu vaccination provider locations distributed across the United States, and what patterns emerge when examining provider characteristics such as minimum eligible age?

---

2. Data preparing

```
# 2.1Read the dataset
fluvacall = read.csv("/Users/ldh/Library/Mobile Documents/com~apple~CloudDocs/USC PhD/STU
```

```
# 2.2 Wrangle & standardize variables being analyzed
## provider_chain: a simplified provider “brand” from loc_name
## vaccine_type: the general product category (“Flu Shot”, “Flu Shot (65+)”, etc.) from s
## age_group: extract age eligibility like “65+”
## last_update_date: parsed quantity_last_updated
## days_since_update: how stale the stock info is (use Sys.Date() for “today”)
## state / lat / lon kept for mapping
## in_stock_flag and supply_level for availability
# ——— Data Prep: create cleaned analytic dataset
```

```
# --- Data Prep: create cleaned analytic dataset
fluvac <- fluvacall %>%                                                  # start with ra
  mutate(
    provider_chain = str_squish(loc_name),                                 # collapse ex
    vaccine_type   = str_squish(searchable_name),                           # human-read
    age_group      = if_else(str_detect(searchable_name, "65\\+"),
                             "65_plus",           # pull explicit age info if mentioned
                       if_else(str_detect(searchable_name, "Nasal Spray"),  # nasal spray
                             "younger_adult_child",
                             "all_ages_general")),                          # fallback i
    last_update_date = suppressWarnings(lubridate::ymd(quantity_last_updated)),  # parse
    days_since_update = as.numeric(Sys.Date() - last_update_date),          # how many da
    state = loc_admin_state,                                                # 2-letter st
    lat = latitude,                                                         # latitude co
    lon = longitude,                                                        # longitude c
    in_stock_flag = in_stock,                                               # TRUE/FALSE
    supply_level_clean = supply_level                                      # numeric / or
  ) %>%
  filter(!is.na(lat), !is.na(lon), !is.na(state))                            # drop rows
## Reasons to conduct above:
### selecting the variables that most relavant to the primary question before doing analy
### extracting “age eligibility” signals directly from searchable_name because my researc
### by creating days_since_update leting us discuss supply recency at each site, tying di
### keeping lat, lon, and state, which will be used to check geographic distribution like
```

3. Exploratory Analysis

```
# 3.1 Dataset size and basic structure (Objective: How big is fluvac and what variables d
## basic structure, size, and variable types
fluvac_overview <- tibble(                                                  # create a t
  n_rows         = nrow(fluvac),                                          # total number
  n_cols         = ncol(fluvac),                                          # total number
  last_update_min = min(fluvac$last_update_date, na.rm = TRUE),           # earliest stoc
  last_update_max = max(fluvac$last_update_date, na.rm = TRUE)            # most recent s
)
fluvac_overview
```

```
# A tibble: 1 × 4
  n_rows n_cols last_update_min last_update_max
   <int>  <int> <date>          <date>
1 202640     38 2023-08-17      2024-08-01
```

```
#
var_names   <- names(fluvac)    # list of variable names
var_names
```

```
 [1] "provider_location_guid" "loc_store_no"           "loc_phone"
 [4] "loc_name"               "loc_admin_street1"      "loc_admin_street2"
```

```
 [7] "loc_admin_city"         "loc_admin_state"        "loc_admin_zip"
[10] "sunday_hours"           "monday_hours"           "tuesday_hours"
[13] "wednesday_hours"        "thursday_hours"         "friday_hours"
[16] "saturday_hours"         "web_address"            "pre_screen"
[19] "insurance_accepted"     "walkins_accepted"       "provider_notes"
[22] "searchable_name"        "in_stock"               "supply_level"
[25] "quantity_last_updated"  "latitude"               "longitude"
[28] "category"               "provider_chain"         "vaccine_type"
[31] "age_group"              "last_update_date"       "days_since_update"
[34] "state"                  "lat"                    "lon"
[37] "in_stock_flag"          "supply_level_clean"
```

```
# 3.2 Checking categories that being analyzed later
# EDA: key categorical summaries (provider brand, product type, age group)
provider_counts <- fluvac %>%
  count(provider_chain, sort = TRUE) %>%          # how many rows per provider brand/cha
  mutate(pct = 100 * n / sum(n))                  # share of total locations
#
vaccine_type_counts <- fluvac %>%
  count(vaccine_type, sort = TRUE) %>%            # how many rows per vaccine type label
  mutate(pct = 100 * n / sum(n))                  # share of total
#
age_group_counts <- fluvac %>%
  count(age_group, sort = TRUE) %>%               # how many rows per inferred eligibili
  mutate(pct = 100 * n / sum(n))                  # % of all sites
vaccine_type_counts
```

```
                            vaccine_type     n      pct
1                              Flu Shot 75693 37.35343
2 Flu Shot (65+, high-dose or adjuvanted) 62786 30.98401
3                     Flu Shot (Egg free) 56017 27.64360
4                         Flu Nasal Spray  8144  4.01895
```

```
age_group_counts
```

```
           age_group      n      pct
1    all_ages_general 131710 64.99704
2            65_plus  62786 30.98401
3 younger_adult_child   8144  4.01895
```

```
### the output below was used for providing characteristics such as type [≈provider_chain
```

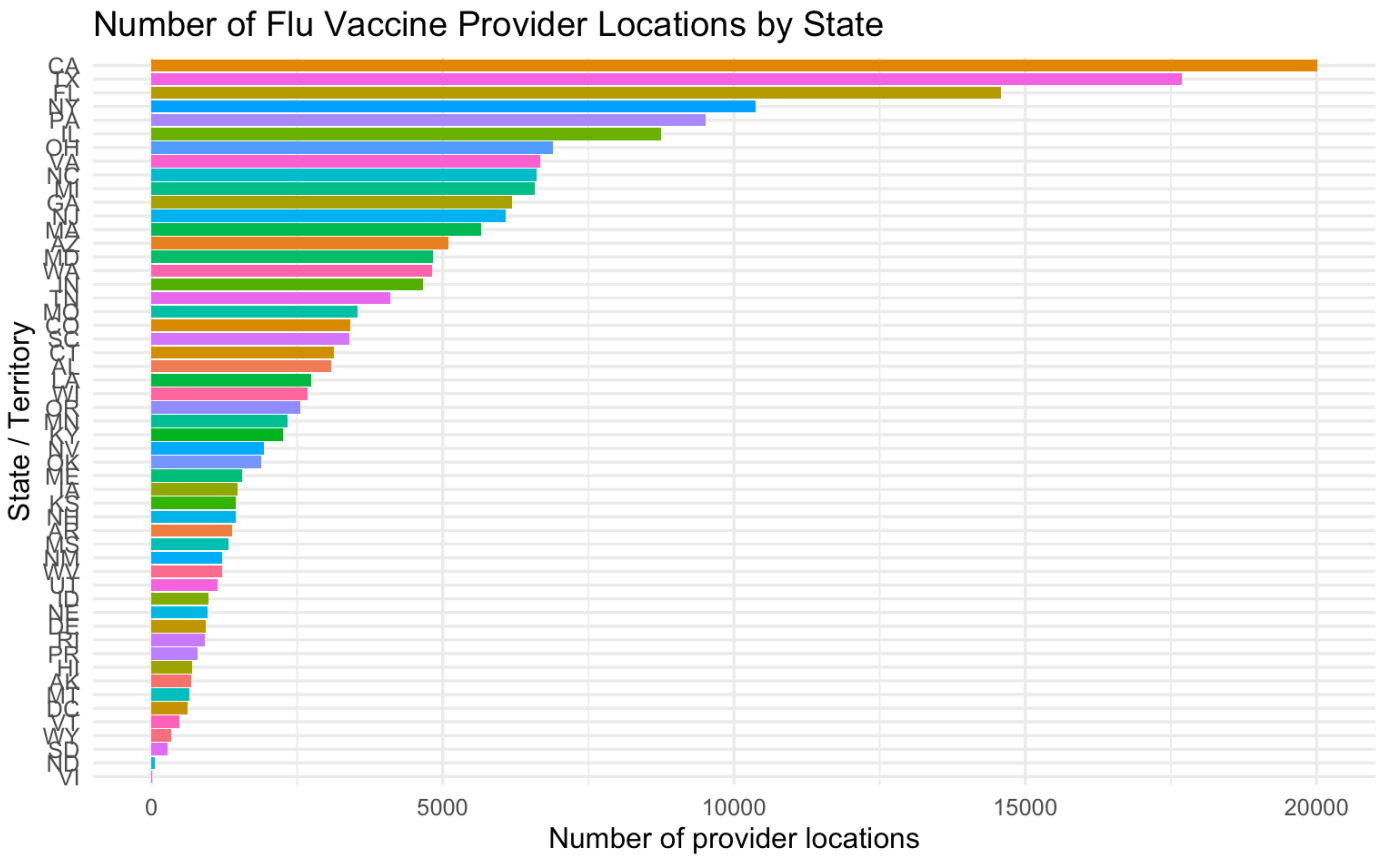
```
# 3.3 numrix summaries with sapply() style (distribution sanity check)
numeric_vars <- fluvac %>%
  select(days_since_update, supply_level_clean, lat, lon)
num_summary <- sapply(numeric_vars, summary)
num_missing <- sapply(numeric_vars, function(x) sum(is.na(x)))
## wrap numeric summary into a nicer long-format table for kable (so no raw console dump)
num_summary_tbl <- as.data.frame(t(num_summary)) %>%
```

```
  mutate(var = rownames(.),
         n_missing = num_missing[ var ]) %>%
  relocate(var)
#
kable(num_summary_tbl,
      caption = "Distribution summaries for numeric variables in `fluvac` (with missing c
  kable_styling(full_width = FALSE, position = "center",
                bootstrap_options = c("striped","hover","condensed","responsive")) %>%
  row_spec(0, bold = TRUE, background = "#756bb1", color = "white") %>%
  row_spec(1:nrow(num_summary_tbl), background = c("#efedf5","white"))
```

Distribution summaries for numeric variables in `fluvac` (with missing counts)

	var	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
days_since_update	days_since_update	451.00000	451.00000	454.00000	498.636701	571.00000	801.00000
supply_level_clean	supply_level_clean	-1.00000	0.00000	0.00000	0.024393	0.00000	4.00000
lat	lat	17.96489	33.69613	38.60332	37.397446	41.31921	71.29726
lon	lon	-176.65900	-97.56295	-85.65537	-90.829963	-78.69288	-64.88893

```
# 3.4 Visualize basic distributions:
## distribution of sites per state (bar plot)
state_counts <- fluvac %>%
  count(state, name = "num_sites") %>%                          # count locations by 2-l
  arrange(desc(num_sites))                                      # rank states by number
#
ggplot(state_counts, aes(x = reorder(state, num_sites), y = num_sites, fill = state)) +
  geom_col(show.legend = FALSE) +                               # bar height = number of
  coord_flip() +                                                # flip for readability
  labs(
    title = "Number of Flu Vaccine Provider Locations by State", # figure title
    x = "State / Territory",                                    # x-axis label
    y = "Number of provider locations"                          # y-axis label
  ) +
  theme_minimal(base_size = 12)                                 # clean theme
```




```
### The output below showed states with the highest counts of listed flu vaccine provider
```

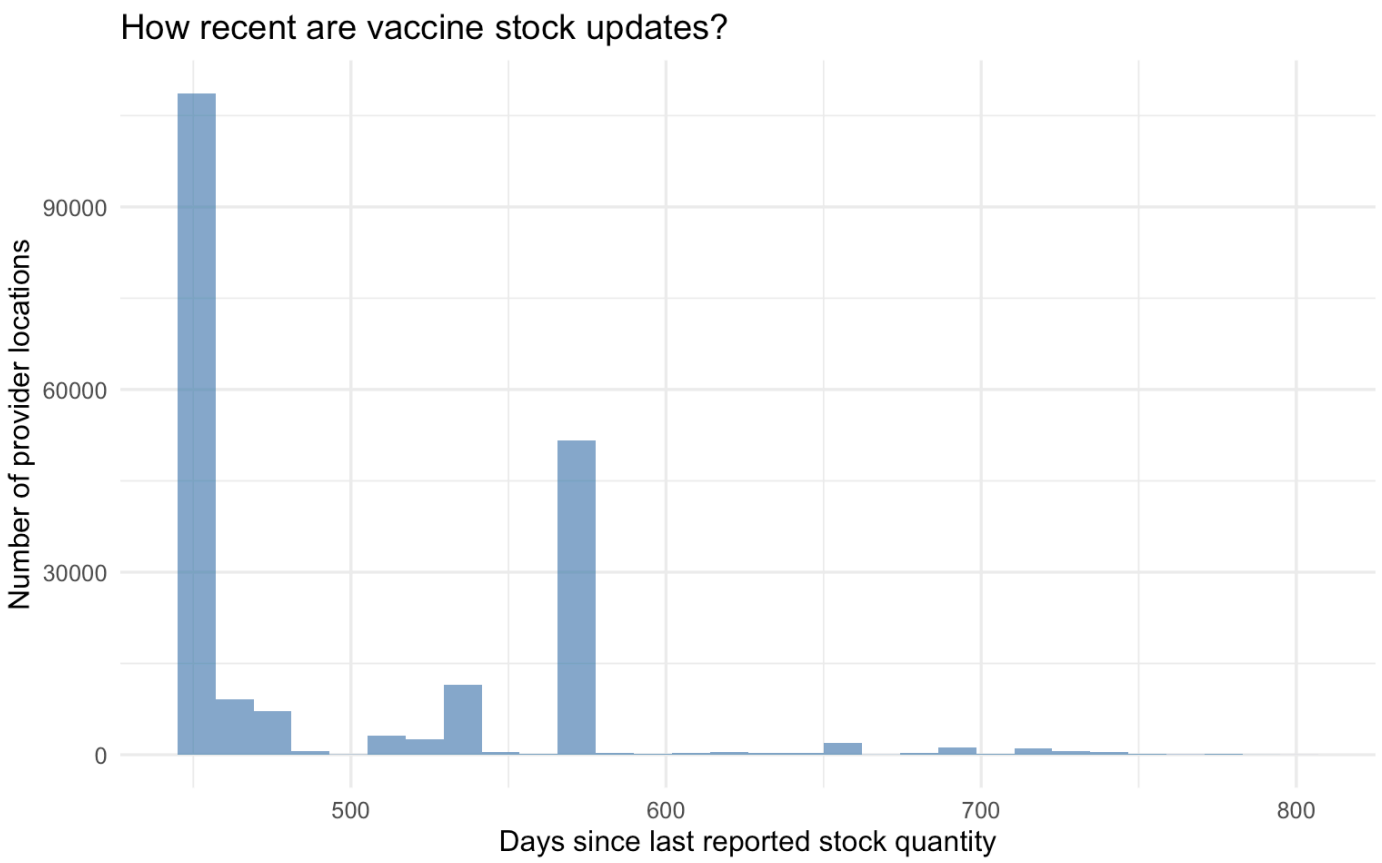
```
## checking in-stock status and stock recency
stock_summary <- fluvac %>%
  mutate(
    stock_status = if_else(isTRUE(in_stock_flag), "In stock", "Not in stock / Unknown")
  ) %>%
  group_by(stock_status) %>%
  summarise(
    n_sites = n(),                                          # number of sites in
    median_days_since_update = median(days_since_update, na.rm=TRUE),  # recency of last
    .groups = "drop"
  )
#
kable(stock_summary,
      caption = "Stock status & reporting recency across flu vaccine provider locations")
  kable_styling(full_width = FALSE, position = "center",
                bootstrap_options = c("striped","hover","condensed","responsive")) %>%
  row_spec(0, bold = TRUE, background = "#756bb1", color = "white") %>%        # styled h
  row_spec(1:nrow(stock_summary), background = c("#efedf5","white"))           # alternat
```

Stock status & reporting recency across flu vaccine provider locations

stock_status	n_sites	median_days_since_update

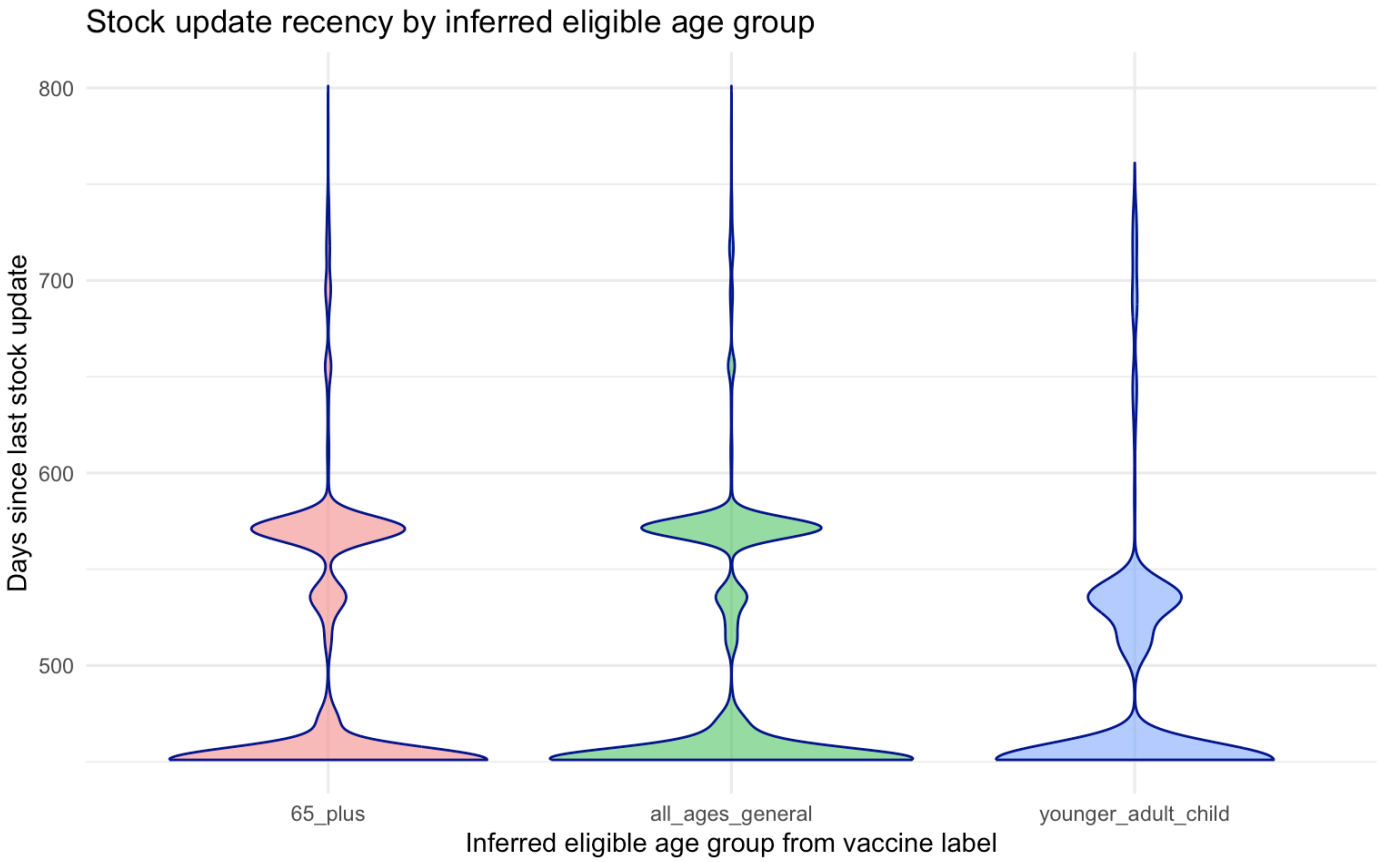
stock_status	n_sites	median_days_since_update
Not in stock / Unknown	202640	454

```
# distribution of days since last stock update
ggplot(fluvac, aes(x = days_since_update)) +
  geom_histogram(bins = 30, fill = "steelblue", alpha = 0.7) +   # histogram of reporting
  labs(
    title = "How recent are vaccine stock updates?",
    x = "Days since last reported stock quantity",
    y = "Number of provider locations"
  ) +
  theme_minimal(base_size = 12)
```

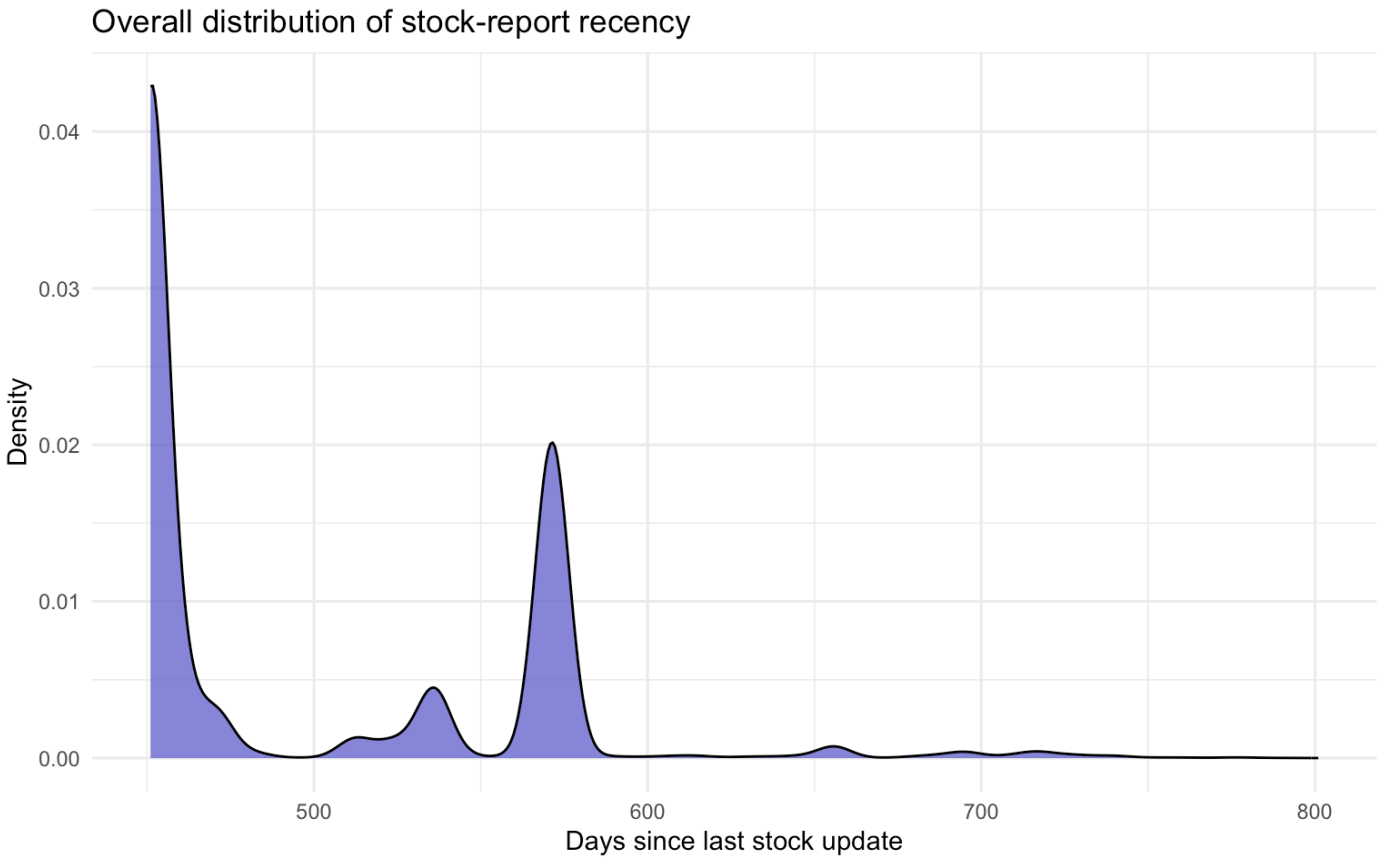



```
## Violin plot of days_since_update by age_group
p_violin_age <- ggplot(
  fluvac,
  aes(x = age_group, y = days_since_update, fill = age_group)
) +
  geom_violin(alpha = 0.5, color = "darkblue") +
  scale_y_continuous(name = "Days since last stock update") +
  labs(
    title = "Stock update recency by inferred eligible age group",
    x = "Inferred eligible age group from vaccine label"
  ) +
```

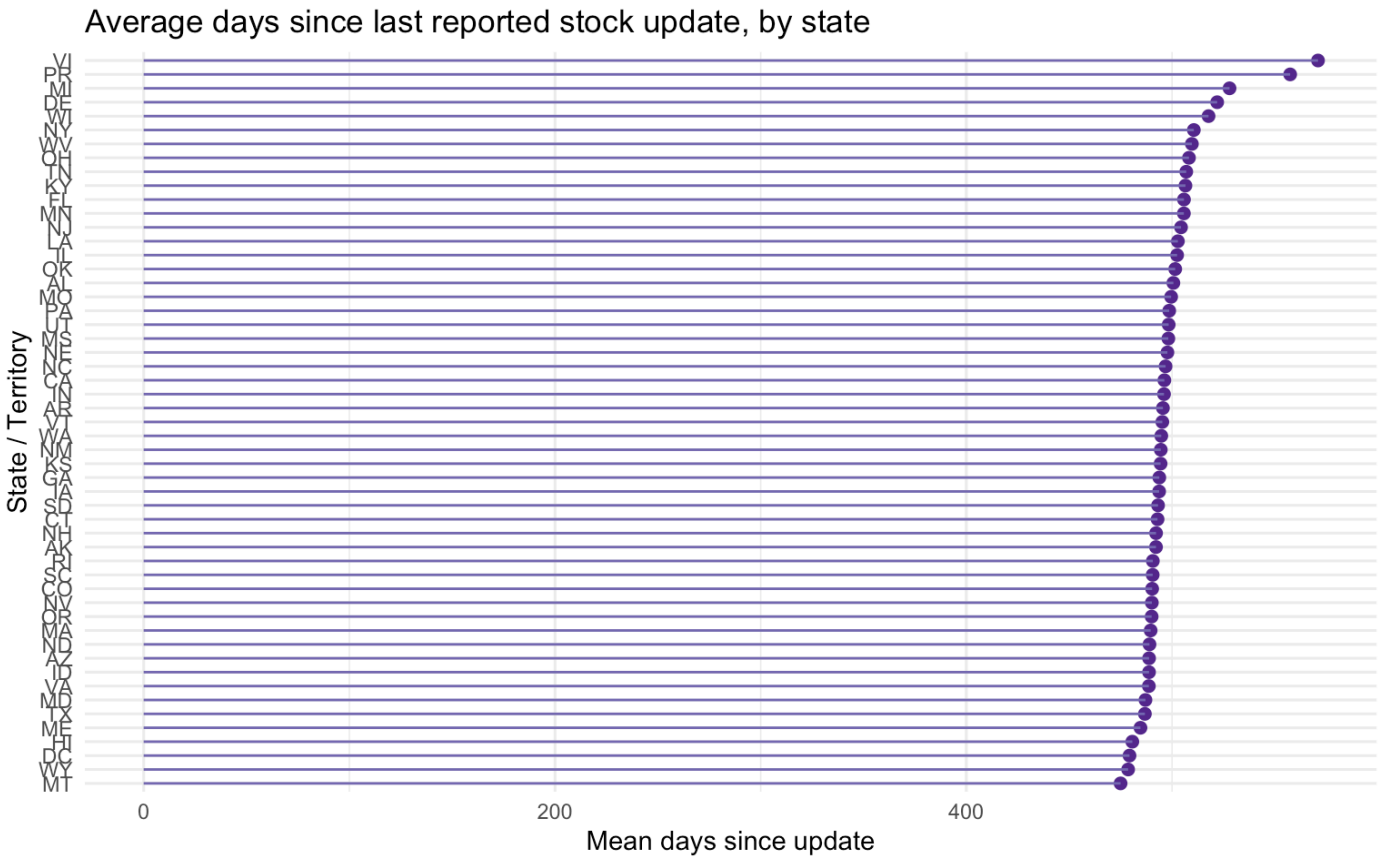
```
  theme_minimal(base_size = 11) +
  theme(legend.position = "none")
p_violin_age
```




```
## Density of days_since_update (freshness overall)
p_density_fresh <- ggplot(
  fluvac,
  aes(x = days_since_update)
) +
  geom_density(fill = "#66c", alpha = 0.8) +                                    #
  labs(
    title = "Overall distribution of stock-report recency",
    x = "Days since last stock update",
    y = "Density"
  ) +
  theme_minimal(base_size = 11)
p_density_fresh
```




```
## Average days_since_update by state (line or lollipop-style)
state_fresh <- fluvac %>%
  group_by(state) %>%
  summarise(mean_days = mean(days_since_update, na.rm = TRUE)) %>%
  arrange(desc(mean_days))
#
p_state_line <- ggplot(
  state_fresh,
  aes(x = reorder(state, mean_days), y = mean_days)
) +
  geom_point(color = "#54278f", size = 2) +
  geom_segment(aes(xend = reorder(state, mean_days),
                   y = 0,
                   yend = mean_days),
               color = "#756bb1") +
  coord_flip() +
  labs(
    title = "Average days since last reported stock update, by state",
    x = "State / Territory",
    y = "Mean days since update"
  ) +
  theme_minimal(base_size = 11)
p_state_line
```

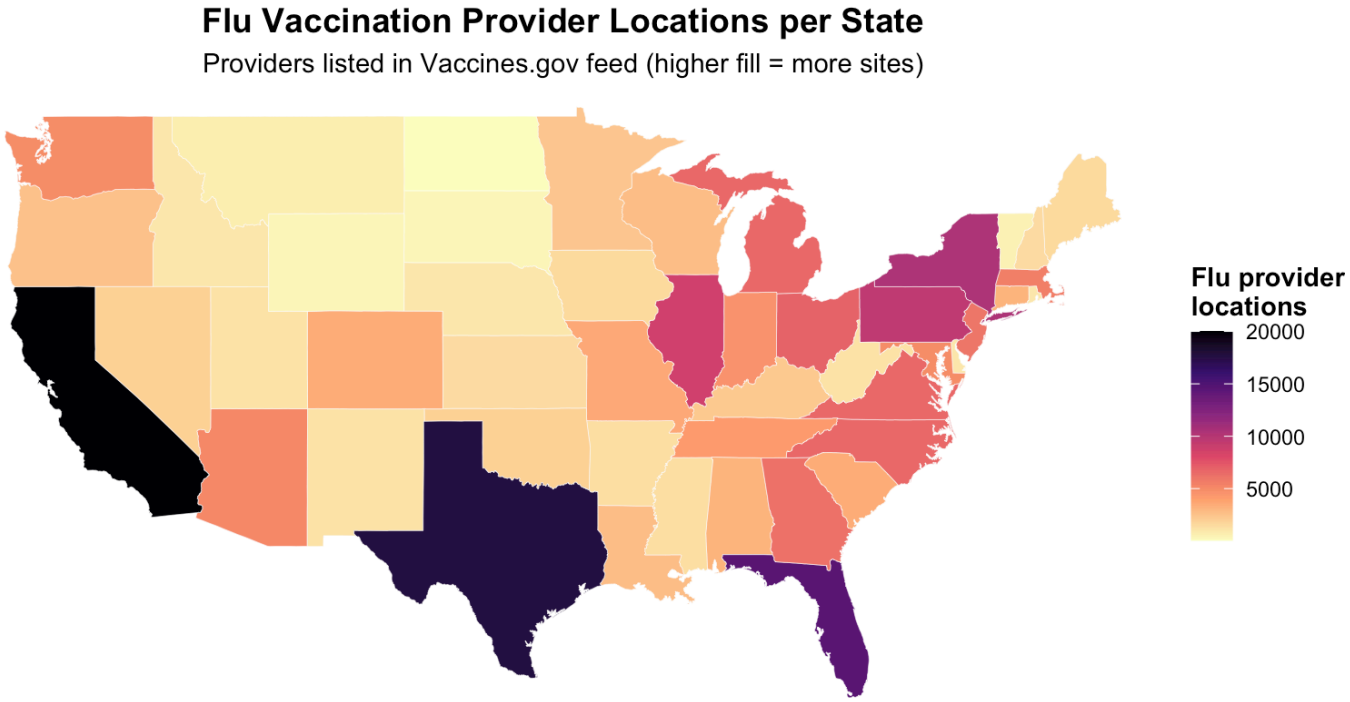


4. Preliminary Results

```
# 4.1 Results: provider locations per state mapped
state_counts_map <- fluvac %>%
  mutate(state_full = tolower(state.name[match(state, state.abb)])) %>%   # convert "CA"
  count(state_full, name = "num_sites")                                   # how many prov
#
usa_map <- map_data("state")                                              # polygon coord
#
chorodata <- usa_map %>%
  left_join(state_counts_map, by = c("region" = "state_full"))            # attach counts
#
ggplot(chorodata, aes(long, lat, group = group, fill = num_sites)) +
  geom_polygon(color = "white", size = 0.1) +                             # draw each sta
  scale_fill_viridis_c(option = "magma", direction = -1, na.value = "grey90",
                       name = "Flu provider\nlocations") +                # color scale s
  coord_quickmap() +                                                      # keeps aspect
  labs(
    title = "Flu Vaccination Provider Locations per State",
    subtitle = "Providers listed in Vaccines.gov feed (higher fill = more sites)",
    x = NULL, y = NULL
  ) +
  theme_void() +                                                          # map-focused t
  theme(
```

```
    plot.title = element_text(face = "bold", hjust = 0.5, size = 14),
    plot.subtitle = element_text(hjust = 0.5, size = 11),
    legend.title = element_text(face = "bold")
  )
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.
```

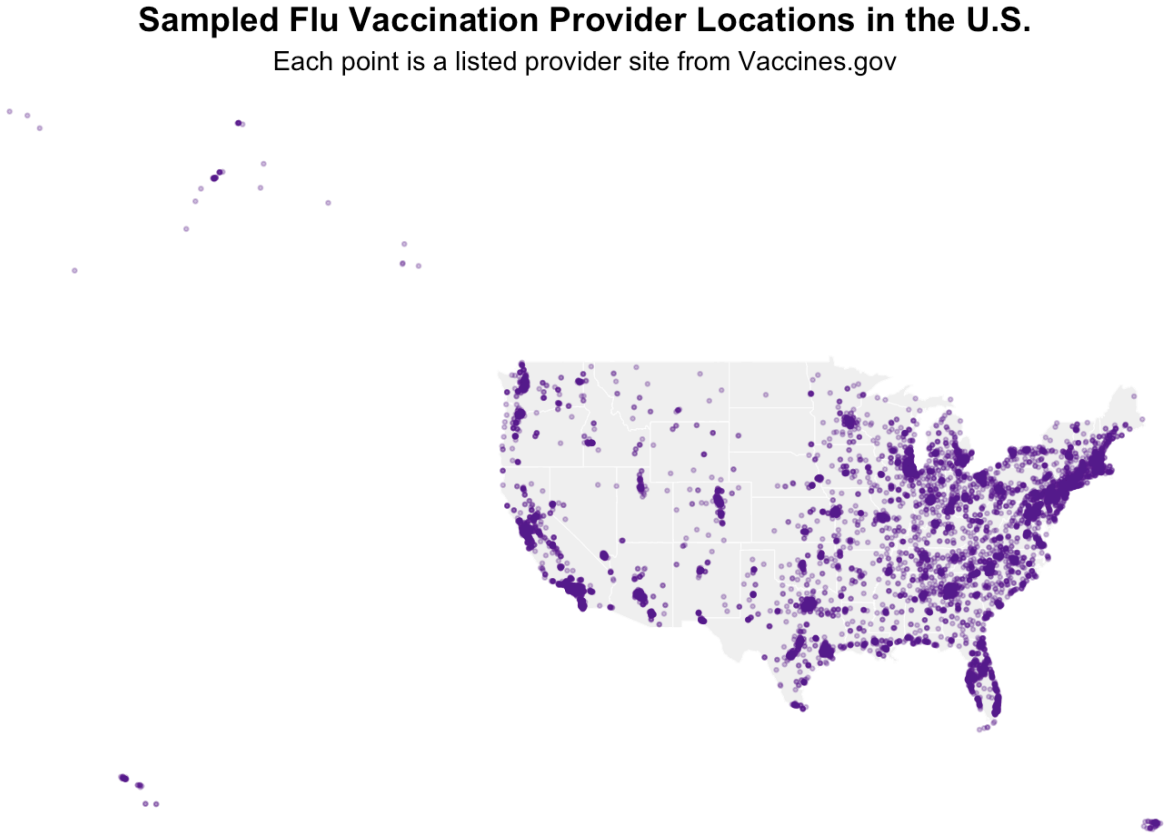


Brief Reports: Provider locations are dense in populous states (e.g., CA, TX, FL, and NY) and more sparse in smaller-population states and certain rural regions. This mirrors what we saw in environmental monitoring site maps, where spatial coverage clusters around populated corridors.The reason of such differences might be states with large urban populations show more listed vaccination sites.

```
# 4.2 Point map of all provider locations
## quick point map of provider locations (sampled)
set.seed(12)                                                    # reproducible
fluvac_sample <- fluvac %>% sample_n(min(10000, n()))           # take up to

usa_states <- map_data("state")                                 # U.S. basemap
```

```
ggplot() +
  geom_polygon(data = usa_states,
               aes(x = long, y = lat, group = group),
               fill = "grey95", color = "white", linewidth = 0.2) +        # light gray U
  geom_point(data = fluvac_sample,
             aes(x = lon, y = lat),
             alpha = 0.3, size = 0.5, color = "purple4") +                 # purple-ish d
  coord_quickmap() +
  labs(
    title = "Sampled Flu Vaccination Provider Locations in the U.S.",
    subtitle = "Each point is a listed provider site from Vaccines.gov",
    x = NULL, y = NULL
  ) +
  theme_void() +
  theme(
    plot.title = element_text(face = "bold", size = 14, hjust = 0.5),
    plot.subtitle = element_text(size = 11, hjust = 0.5)
  )
```



Brief Reports: The result above revealed the spatial distribution of influenza vaccine provider sites across the United States. A point map representation of sampling provider sites (up to 10,000 locations) illustrated that flu vaccine providers are extensively

distributed throughout the country, with much larger densities in highly populated areas, especially along the East Coast, California, Texas, and some sections of the Midwest. In contrast, rural and less densely inhabited regions, such as the Mountain West and some sections of the Great Plains, had a relatively lower number of provider locations. This geographical distribution pattern corresponded with anticipated population and healthcare infrastructure density, indicating increased provider supply in metropolitan areas where demand is greatest. infrastructure patterns.

```
# 4.3 Results: Minimum eligible age: looking at what fraction of locations explicitly adv
age_mix <- fluvac %>%
  group_by(age_group) %>%                                    # "65_plus", "younge
  summarise(
    n_sites = n(),                                           # number of sites of
    pct_sites = 100 * n_sites / nrow(fluvac),                # share of all sites
    .groups = "drop"
  )
#
kable(age_mix,
      caption = "Flu Vaccine Eligibility Focus (based on product labeling at site)") %>%
  kable_styling(full_width = FALSE, position = "center",
                bootstrap_options = c("striped","hover","condensed","responsive")) %>%
  row_spec(0, bold = TRUE, background = "#756bb1", color = "white") %>%
  row_spec(1:nrow(age_mix), background = c("#efedf5","white"))
```

Flu Vaccine Eligibility Focus (based on product labeling at site)

age_group	n_sites	pct_sites
65_plus	62786	30.98401
all_ages_general	131710	64.99704
younger_adult_child	8144	4.01895

Brief Reports: A substantial fraction of listed locations explicitly offer senior-focused formulations (‘65+’, high-dose or adjuvanted vaccines) (about 31%). This result may highlight targeted access for older adults, a high-risk group for severe influenza outcomes. Additionally, most locations advertise a general ‘Flu Shot’, assumed

to cover standard quadrivalent inactivated influenza vaccine for the general population (6 months+).

```
# 4.4. Within-State Market Share of Flu Vaccination Providers in the United States
dominance_share <- fluvac %>%
  group_by(state) %>%
  mutate(state_total_sites = n()) %>%                                    # total sites
  ungroup() %>%
  group_by(state, provider_chain) %>%
  summarise(
    n_sites = n(),
    state_total_sites = first(state_total_sites),
    share_pct = 100 * n_sites / state_total_sites,                        # % share of
    .groups = "drop"
  ) %>%
  arrange(desc(share_pct)) %>%
  group_by(state) %>%
  slice_max(order_by = share_pct, n = 1, with_ties = FALSE) %>%           # which chain
  ungroup()
#
kable(dominance_share,
      caption = "Within-state market share: which chain accounts for the largest share of
  kable_styling(full_width = FALSE, position = "center",
                bootstrap_options = c("striped","hover","condensed","responsive")) %>%
  row_spec(0, bold = TRUE, background = "#756bb1", color = "white") %>%
  row_spec(1:nrow(dominance_share), background = c("#efedf5","white"))      #
```

Within-state market share: which chain accounts for the largest share of listed flu vaccine locations in each state?

state	provider_chain	n_sites	state_total_sites	share_pct
AK	Walmart Inc #Oct-74	12	676	1.7751479
AL	Walmart Inc #Oct-48	24	3081	0.7789679
AR	Walmart Inc #31-Oct	18	1381	1.3034033
AZ	Walmart Inc #3-Oct	18	5092	0.3534957
CA	Walmart Inc #1-Oct	36	20012	0.1798921
CO	Walmart Inc #Oct-82	18	3413	0.5273953
CT	Stop & Shop #2605	12	3133	0.3830195
DC	Giant Food #2376	12	627	1.9138756
DE	Giant Food #2351	12	932	1.2875536
FL	Walmart Inc #17-Oct	30	14582	0.2057331
GA	Walmart Inc #9-Oct	24	6188	0.3878474
HI	SAFEWAY PHARMACY #1087	11	705	1.5602837
IA	OSCO DRUG #1118	11	1474	0.7462687

state	provider_chain	n_sites	state_total_sites	share_pct
ID	Walmart Inc #Oct-94	18	982	1.8329939
IL	Walmart Inc #Oct-99	18	8752	0.2056673
IN	Walmart Inc #31-Oct	12	4666	0.2571796
KS	Walmart Inc #Oct-55	18	1454	1.2379642
KY	Walmart Inc #18-Oct	12	2264	0.5300353
LA	Walmart Inc #22-Oct	18	2740	0.6569343
MA	Hannaford #8003	12	5659	0.2120516
MD	Giant Food #100	12	4831	0.2483958
ME	Hannaford #8107	12	1551	0.7736944
MI	Walmart Inc #Oct-44	18	6578	0.2736394
MN	Walmart Inc #Oct-52	12	2337	0.5134788
MO	Walmart Inc #27-Oct	18	3539	0.5086183
MS	Walmart Inc #19-Oct	12	1328	0.9036145
MT	Walmart Inc #Oct-47	12	648	1.8518519
NC	Walmart Inc #4-Oct	30	6608	0.4539952
ND	CVS Pharmacy, Inc. #8611	7	59	11.8644068
NE	SAFEWAY PHARMACY #2555	11	966	1.1387164
NH	Walmart Inc #30-Oct	18	1447	1.2439530
NJ	Walmart Inc #12-Oct	18	6077	0.2961988
NM	Walmart Inc #1-Oct	12	1216	0.9868421
NV	Walmart Inc #6-Oct	12	1924	0.6237006
NY	Walmart Inc #Oct-97	24	10367	0.2315038
OH	Walmart Inc #9-Oct	36	6896	0.5220418
OK	Walmart Inc #Oct-95	18	1890	0.9523810
OR	Walmart Inc #20-Oct	12	2562	0.4683841
PA	Walmart Inc #Oct-45	30	9508	0.3155238
PR	Walmart Inc #2-Oct	12	787	1.5247776
RI	Stop & Shop #2701	12	912	1.3157895
SC	Walmart Inc #Oct-87	36	3397	1.0597586
SD	SAFEWAY PHARMACY #1554	11	271	4.0590406
TN	Walmart Inc #Oct-35	18	4105	0.4384896
TX	Walmart Inc #Oct-63	42	17683	0.2375163
UT	Walmart Inc #Oct-68	24	1132	2.1201413
VA	Walmart Inc #5-Oct	24	6671	0.3597662

state	provider_chain	n_sites	state_total_sites	share_pct
VI	Walgreens Co. #13846	5	5	100.0000000
VT	Hannaford #8121	12	483	2.4844720
WA	Walmart Inc #Oct-37	18	4820	0.3734440
WI	Walmart Inc #Oct-68	18	2682	0.6711409
WV	Food Lion #2194	12	1215	0.9876543
WY	OSCO PHARMACY #2061	11	332	3.3132530

Brief Reports:The analysis of intra-state market share revealed that major retail pharmacy chains, notably Walmart, CVS, Walgreens, and Safeway, often represented the predominant proportion of designated influenza vaccine provider locations in several states. Walgreens constituted 100% of the reported locations in the U.S. Virgin Islands.

---

5. Conclusion:According to the U.S. Centers for Disease Control and Prevention (CDC) via the Vaccines.gov platform, flu vaccine provider sites were extensively scattered across the United States, exhibiting distinct geographic and provider-related trends. The maps showed concentrated clusters in densely populated regions, like the East Coast corridor, California, Texas, and Florida, but rural areas in the Mountain West and Upper Plains have less coverage, indicating differences in access related to population density. Market share analysis indicates that major retail chains like Walmart, CVS, and Walgreens frequently constitute the largest share of locations within states. Concerning provider characteristics, the majority of sites cater to the general public, around one-third specifically promote vaccinations for older persons (65+), and a very minor percentage concentrate on pediatric or younger adult demographics. The findings indicated that

general adult access is widely available, older adults are adequately served in numerous locations, yet child-focused vaccination options are scarce, underscoring the necessity for targeted strategies to enhance equity in flu vaccination access, particularly in rural regions and for pediatric populations.