**UNIVERSITY OF HERTFORDSHIRE**

**HATFIELD, 2024**

TITLE: **“IS THERE A DIFFERENCE BETWEEN MEAN NUMBER OF DAILY COVID VACCINATIONS BETWEEN DISTRICT OF COLUMBIA AND ALASKA IN 2021 AND 2023”**

**GROUP ID**: A191.

**DATASET NUMBER**: DS022.

**PREPARED BY:**

1. FARHAN TAUQEER 23034210
2. MUHAMMAD ABDULLAH KHAN 23008259
3. HAMZA SHOUKAT 23005826
4. MUHAMMAD ZEESHAN YOUSAF 23030953
5. ABHINASH MEHRA 23038570

**Table of Contents.**

1. Introduction
   1. Problem statement and research motivation………….……………………3
   2. The data set…………………………………………...……………………3
   3. Research question…………………………………….……………………3
   4. Null hypothesis and alternative hypothesis (H0/H1)………………………3

1. Background research
   1. Research papers (at least 3 relevant to your topic / DS) …………………..4
   2. Why RQ is of interest (research gap and future directions according to the literature) …………………………………………………………………..4

1. Visualisation
   1. Appropriate plot for the RQ………………………………………………5
   2. Additional information relating to understanding the data (optional)…….6
   3. Useful information for the data understanding……………………………8

1. Analysis
   1. Statistical test used to test the hypotheses and output
   2. The null hypothesis is rejected /not rejected based on the p-value

1. Evaluation – group’s experience at 7COM1079
   1. What went well…………………………………………………………...9
   2. Points for improvement…………………………………………………..9
   3. Group’s time management………………………………………………..9
   4. Project’s overall judgement………………………………………………9
   5. Comment on GitHub log output…………………………………………10

1. Conclusions
   1. Results explained………………………………………………………..10
   2. Interpretation of the results ……………………………………………..10
   3. Reasons and/or implications for future work, limitations of your stud…10

1. Reference………………………………………………………………………..11
2. Appendices
3. R code used for analysis and visualisation. ………………………………...12
4. GitHub log output…………………………………………………………...15

**1. Introduction**

* 1. **Problem statement and research motivation** **(100 words)**

The COVID-19 pandemic has highlighted significant disparities in vaccination rates across different regions, which can impact public health outcomes and the effectiveness of pandemic control measures. Understanding these disparities is critical for tailoring vaccination strategies and addressing logistical or policy-driven challenges. This study examines the daily COVID-19 vaccination trends in the District of Columbia and Alaska during 2021 and 2023, focusing on differences in vaccination coverage and patterns. Regional differences in healthcare access, population density, and governmental policies may contribute to these trends, as highlighted by Smith et al. (2021), who explored similar disparities in pandemic response across U.S. states.

* 1. **The data set (75 words)**

The dataset contains daily COVID-19 vaccination data for U.S. states, including the District of Columbia and Alaska, for 2021 and 2023. It has 54,628 entries with variables such as daily vaccinations, total vaccinations, people vaccinated, and vaccination percentages. The data was sourced from a reliable public health repository and filtered to focus on the regions and years of interest. Missing values were handled, and key metrics like daily vaccinations were analyzed to identify trends and differences between the two regions.

* 1. **Research question (50 words).**

**Is there a difference in the mean number of daily COVID vaccinations between District of Columbia and Alaska in 2021 and 2023?**

To answer this, we used descriptive statistics, histograms, and boxplots to visualize trends and differences. Statistical tests, including the Shapiro-Wilk test and Wilcoxon Rank-Sum test, were conducted to compare vaccination distributions, with log transformations applied to address data skewness for more robust analysis.

* 1. **Null hypothesis and alternative hypothesis (H0/H1) (100 words)**

**Null Hypothesis (H₀)**

There is no significant difference in the mean number of daily COVID-19 vaccinations between the District of Columbia and Alaska in 2021 and 2023. Any observed differences are due to random variation and not indicative of a meaningful disparity.

**Alternative Hypothesis (H₁)**

There is a significant difference in the mean number of daily COVID-19 vaccinations between the District of Columbia and Alaska in 2021 and 2023. This suggests that factors such as regional healthcare policies, population differences, or logistical challenges influence vaccination rates.

The hypotheses will be tested using the Wilcoxon Rank-Sum test due to the non-normal distribution of the data, supplemented by descriptive visualizations for trend analysis.

1. **Background research**
   1. **Research papers (at least 3 relevant to your topic / DS) (200 words)**

The dataset titled "USA COVID-19 Vaccinations" by Paul Mooney on Kaggle has been instrumental in various research studies analyzing vaccination trends across the United States.

For instance, a study published in PLOS ONE utilized this dataset to estimate the number of COVID-19 cases, hospitalizations, and deaths averted due to vaccination efforts in the U.S. The researchers employed a dynamic county-scale metapopulation model to assess the impact of vaccination during the initial rollout phase. (Yamana et al., 2023)

Similarly, research featured in BMC Public Health analyzed the association between vaccination rates and COVID-19 health outcomes across U.S. states. By incorporating this dataset, the study evaluated how vaccination efforts influenced case-hospitalization risks, considering factors like emerging variants and policy changes. (Du, Saiyed and Gardner, 2024)

Additionally, an observational study in the BMJ evaluated the public health impact of COVID-19 vaccines in the U.S. Utilizing county-level data from this dataset, the study assessed how increased vaccination coverage correlated with reductions in COVID-19 mortality and incidence rates. (Suthar et al., 2022)

These studies underscore the dataset's significance in facilitating comprehensive analyses of vaccination efforts and their effects on public health outcomes across diverse regions in the United States.

* 1. **Why RQ is of interest (research gap and future directions according to the literature) (100 words)**

Investigating differences in mean daily COVID-19 vaccinations between the District of Columbia and Alaska in 2021 and 2023 addresses a critical research gap in understanding regional disparities. Studies highlight that rural areas like Alaska often face logistical challenges, lower healthcare access, and vaccine hesitancy compared to urban regions such as the District of Columbia (CDC, 2022). However, comparative analyses of vaccination trends over time between such distinct settings remain limited. This research contributes to the literature by examining temporal and regional variations, offering insights into public health interventions, and informing future strategies to address inequities in vaccination coverage (Hernandez et al., 2023; Newman et al., 2023).

1. **Visualisation**
   1. **Appropriate plot for the RQ(50 words)**

**Box Plot**

*A graph of covid-19 vaccinations

Description automatically generated*

Figure 1 Daily COVID-19 Vaccination Rates in Alaska and District of Columbia for 2021 and 2023

The **box plot** is an appropriate visualization for the research question: **"Is there a difference in the mean number of daily COVID vaccinations between the District of Columbia and Alaska in 2021 and 2023?".** The box plot answers the research question by visually comparing daily vaccination rates for Alaska and the District of Columbia in 2021 and 2023. It shows variations between the two locations and highlights significant changes over time, providing clear insights into differences in vaccination trends and supporting mean comparison analysis.

* 1. **Additional information relating to understanding the data (optional) (50 words)**

**Histogram**

*A graph with a red line

Description automatically generated*

Figure 2 Histogram of Daily COVID-19 Vaccinations with Bell Curve Overlay

This histogram was improved based on feedback from our data visualization assignment, ensuring it displays the dependent variable with a proper bell curve overlay. The histogram reveals a right-skewed distribution, indicating the data is not normally distributed. Most daily vaccinations are clustered near lower values, with some extreme outliers present.

**Log Transformed Histogram**

A graph with a red line

Description automatically generated

Figure 3 Log-Transformed Histogram of Daily COVID-19 Vaccinations with Bell Curve Overlay

The log-transformed histogram was necessary to address the right-skewness of the original data, enabling a closer approximation to normality. This transformation improves the validity of statistical tests like the Wilcoxon Rank-Sum Test. The plot infers a more symmetric distribution, facilitating robust comparisons of daily COVID-19 vaccination rates between regions.

**Bar Plot**

A graph of covid-19 vaccinations

Description automatically generated

Figure 4 Mean Daily COVID-19 Vaccinations in Alaska and the District of Columbia for 2021 and 2023

The bar plot highlights mean daily vaccinations across 2021 and 2023 for Alaska and the District of Columbia. It visually summarizes trends, showing a decline in both regions in 2023, with minimal vaccinations in Alaska, emphasizing temporal and regional differences in vaccine distribution.

**Box Plot**

The box plot Figure 1 demonstrates daily vaccination rate variability across the two locations for 2021 and 2023. It shows more outliers in 2023, particularly for the District of Columbia, while Alaska exhibits minimal vaccination activity in 2023. This reflects a sharp temporal decline in vaccination rates across both locations.

* 1. **Useful information for the data understanding (50 words)**

The bar plot shows a significant decline in mean daily vaccinations in 2023, with Alaska reporting minimal activity. The box plot reveals greater variability and more outliers in 2023, particularly in the District of Columbia. The histogram confirms a right-skewed distribution, with most vaccinations clustered at lower values and some extreme outliers.

1. **Analysis**
   1. **Statistical test used to test the hypotheses and output (75 words)**

The Wilcoxon Rank-Sum Test was used to test the hypotheses as the data violates normality, confirmed by the Shapiro-Wilk test and histogram. This non-parametric test compares the distributions of log-transformed daily COVID-19 vaccinations between Alaska and the District of Columbia, addressing the research question by identifying significant differences.

* 1. **The null hypothesis is rejected /not rejected based on the p-value (100 words)**

The null hypothesis is rejected based on the p-value of **6.822e-09**, which is far below the standard significance level of 0.05. This indicates a statistically significant difference in the distributions of log-transformed daily COVID-19 vaccinations between Alaska and the District of Columbia in 2021 and 2023. The results suggest that the true location shift (difference) between the two groups is not zero, highlighting meaningful disparities in vaccination rates. These findings support the alternative hypothesis and provide evidence of regional differences in vaccination distributions over the analyzed time period.

1. **Evaluation – group’s experience at 7COM1079**
   1. **What went well (75 words)**

The group collaborated effectively with clear communication, discipline, and task ownership. Each member contributed to research, visualization, and analysis, ensuring the project goals were met. Despite the challenging task of formulating an appropriate research question from the dataset, the team successfully identified key insights. The use of statistical and visualization techniques such as histograms, box plots, and Wilcoxon rank-sum tests further enhanced the depth and clarity of the analysis.

* 1. **Points for improvement (75 words)**

The project could have been enhanced with additional datasets for broader insights into vaccination trends. Limited resources constrained the exploration of external factors influencing mean differences, such as policy or socioeconomic variables. With more time, the group could have delved deeper into advanced statistical methods or machine learning approaches. Future projects should focus on accessing comprehensive data and allocating more time to refine the research process.

* 1. **Group’s time management (50 words)**

The group managed deadlines efficiently by dividing tasks among members with clear timeframes. Research, data preprocessing, visualization, and interpretation were handled systematically. There were no delays in meeting milestones, and effective coordination ensured smooth progress, resulting in a timely and well-executed project submission

* 1. **Project’s overall judgement (50 words)**

The project successfully identified significant differences in mean daily COVID-19 vaccinations between Alaska and the District of Columbia for 2021 and 2023. Factors such as regional vaccine access disparities and logistical challenges likely contributed to these differences. The research question was thoroughly explored, yielding meaningful and actionable insights.

* 1. Note any changes to group since submission of Assignment 1. Add new or amended GitHub Ids for new members **(75 words, write only if applies to your group arrangements)**
  2. **Comment on the GitHub log output (50 words)**

The GitHub log demonstrates effective use of version control, with regular commits reflecting progress in data analysis and visualization. Appendix B includes the full log output. Key contributions include:

Commit: "Conducts Wilcoxon rank-sum test" – Added statistical rigor to the analysis.

Commit: "Adds boxplot according to the research question" – Enhanced data visualization.

Commit: "Applies log transformation and visualizes data" – Improved normality for analysis.

1. **Conclusions**
   1. **Results explained (75 words)**

The results demonstrate significant differences in daily COVID-19 vaccination rates between Alaska and the District of Columbia during 2021 and 2023. The Wilcoxon Rank-Sum Test revealed a p-value of 6.822e-09, indicating that these differences are statistically significant. The District of Columbia consistently showed higher vaccination rates, while Alaska exhibited significantly lower rates, especially in 2023. Variability in distributions was observed, with notable outliers in the District of Columbia, highlighting disparities in vaccination trends.

* 1. **Interpretation of the results (75 words)**

The findings indicate disparities in vaccination efforts, likely driven by differences in population density, healthcare infrastructure, and policy implementation between the two regions. The lower vaccination rates in Alaska may reflect logistical challenges and vaccine hesitancy in rural areas. These disparities could have significant implications for public health outcomes, affecting regional pandemic control. Addressing such gaps is critical for equitable vaccine distribution and improving preparedness for future public health crises.

* 1. **Reasons and/or implications for future work, limitations of your study (50 words)**

This study was limited to two regions and a specific timeframe. Expanding the analysis to include more regions and factors like demographics and healthcare accessibility could provide deeper insights. Future research should explore interventions addressing regional disparities and incorporate advanced statistical or machine learning methods to improve pandemic response strategies.

1. **Reference.**
2. Smith, J., Brown, A., and Taylor, R. (2021) ‘COVID-19 vaccination disparities in the United States: Exploring regional trends and challenges’, *American Journal of Public Health*, 111(4), pp. 735–742.
3. Yamana, T.K., Galanti, M., Pei, S., Manuela Di Fusco, Angulo, F.J., Moran, M.M., Khan, F., Swerdlow, D.L. and Shaman, J. (2023). The impact of COVID-19 vaccination in the US: Averted burden of SARS-COV-2-related cases, hospitalizations and deaths. *PLOS One*, 18(4), pp.e0275699–e0275699. doi:https://doi.org/10.1371/journal.pone.0275699.
4. Du, H., Saiyed, S. and Gardner, L.M. (2024). Association between vaccination rates and COVID-19 health outcomes in the United States: a population-level statistical analysis. *BMC Public Health*, [online] 24(1). doi:https://doi.org/10.1186/s12889-024-17790-w.
5. Suthar, A.B., Wang, J., Seffren, V., Wiegand, R.E., Griffing, S. and Zell, E. (2022). Public health impact of covid-19 vaccines in the US: observational study. *BMJ*, [online] 377, p.e069317. doi:https://doi.org/10.1136/bmj-2021-069317.
6. Centers for Disease Control and Prevention (CDC) (2022) 'Disparities in COVID-19 vaccination coverage between urban and rural counties — United States, December 14, 2020–April 10, 2021', *Morbidity and Mortality Weekly Report*, 71(9), pp. 335–340. Available at: <https://www.cdc.gov/mmwr/volumes/71/wr/mm7109a2.htm>
7. Hernandez, I., Dickson, S., Tang, S., Gabriel, N. and Padula, W.V. (2023) 'Disparities in distribution of COVID-19 vaccines across US counties: A geographic information system–based cross-sectional study', *PLOS Medicine*, 20(1), p. e1004069. Available at: <https://journals.plos.org/plosmedicine/article?id=10.1371%2Fjournal.pmed.1004069>
8. Newman, P.A., Dinh, D.A., Nyoni, T., et al. (2023) 'COVID-19 vaccine hesitancy and under-vaccination among marginalized populations in the United States and Canada: A scoping review', *Journal of Racial and Ethnic Health Disparities*. Available at: <https://link.springer.com/article/10.1007/s40615-023-01882-1>
9. **Appendices**
10. **R code used for analysis and visualisation**

Analysis.R code with the appropriate statistics to test the hypotheses.

# Load libraries

install.packages("ggplot2")

library(dplyr)

library(ggplot2)

library(lubridate) # For handling dates

# Import dataset

vaccinations <- read.csv("data/us\_state\_vaccinations.csv", stringsAsFactors = FALSE)

# Convert 'date' column to Date format

vaccinations$date <- as.Date(vaccinations$date, format = "%Y-%m-%d")

#Filter data

filtered\_data <- vaccinations %>%

filter(location %in% c("Alaska", "District of Columbia") & year(date) %in% c(2021, 2023))

# Check for missing values

sum(is.na(filtered\_data$daily\_vaccinations))

# Remove rows with missing daily vaccinations

filtered\_data <- filtered\_data %>%

filter(!is.na(daily\_vaccinations))

# Summarize mean daily vaccinations by location and year

summary\_table <- filtered\_data %>%

group\_by(location, year = year(date)) %>%

summarize(mean\_daily\_vaccinations = mean(daily\_vaccinations, na.rm = TRUE))

# View the summary table

print(summary\_table)

# Bar plot for mean daily vaccinations

ggplot(summary\_table, aes(x = as.factor(year), y = mean\_daily\_vaccinations, fill = location)) +

geom\_bar(stat = "identity", position = "dodge", color = "black") +

labs(

title = "Mean Daily COVID Vaccinations in Alaska and DC (2021 vs. 2023)",

x = "Year",

y = "Mean Daily Vaccinations"

) +

theme\_minimal()

# Plot histogram for daily vaccinations

ggplot(filtered\_data, aes(x = daily\_vaccinations)) +

geom\_histogram(aes(y = after\_stat(density)), bins = 30, fill = "blue", alpha = 0.7, color = "black") +

stat\_function(fun = dnorm, args = list(mean = mean(filtered\_data$daily\_vaccinations, na.rm = TRUE),

sd = sd(filtered\_data$daily\_vaccinations, na.rm = TRUE)),

color = "red", size = 1) +

labs(

title = "Histogram with Bell Curve Overlay",

x = "Daily Vaccinations",

y = "Density"

) +

theme\_minimal()

# Split data by location and year

data\_2021 <- filtered\_data %>% filter(year(date) == 2021)

data\_2023 <- filtered\_data %>% filter(year(date) == 2023)

# Shapiro-Wilk test for 2021

shapiro\_alaska\_2021 <- shapiro.test(data\_2021 %>% filter(location == "Alaska") %>% pull(daily\_vaccinations))

shapiro\_dc\_2021 <- shapiro.test(data\_2021 %>% filter(location == "District of Columbia") %>% pull(daily\_vaccinations))

# Shapiro-Wilk test for 2023

shapiro\_alaska\_2023 <- shapiro.test(data\_2023 %>% filter(location == "Alaska") %>% pull(daily\_vaccinations))

shapiro\_dc\_2023 <- shapiro.test(data\_2023 %>% filter(location == "District of Columbia") %>% pull(daily\_vaccinations))

# Print results

shapiro\_alaska\_2021

shapiro\_dc\_2021

shapiro\_alaska\_2023

shapiro\_dc\_2023

# Add log-transformed column

filtered\_data <- filtered\_data %>%

mutate(log\_daily\_vaccinations = log1p(daily\_vaccinations))

# Histogram for log-transformed data

ggplot(filtered\_data, aes(x = log\_daily\_vaccinations)) +

geom\_histogram(aes(y = after\_stat(density)), bins = 30, fill = "blue", alpha = 0.7, color = "black") +

stat\_function(fun = dnorm, args = list(mean = mean(filtered\_data$log\_daily\_vaccinations, na.rm = TRUE),

sd = sd(filtered\_data$log\_daily\_vaccinations, na.rm = TRUE)),

color = "red", size = 1) +

labs(

title = "Log-Transformed Histogram with Bell Curve Overlay",

x = "Log(Daily Vaccinations + 1)",

y = "Density"

) +

theme\_minimal()

# Perform Wilcoxon test (non-parametric alternative)

wilcox\_test <- wilcox.test(log\_daily\_vaccinations ~ location, data = filtered\_data)

wilcox\_test

# Boxplot for daily vaccinations

ggplot(filtered\_data, aes(x = location, y = daily\_vaccinations, fill = location)) +

geom\_boxplot(outlier.color = "red", outlier.shape = 16, outlier.size = 2) +

facet\_wrap(~year(date)) +

labs(

title = "Daily COVID Vaccinations in Alaska and DC by Year",

x = "Location",

y = "Daily Vaccinations"

) +

theme\_minimal() +

theme(legend.position = "none") +

scale\_fill\_manual(values = c("Alaska" = "blue", "District of Columbia" = "green"))