Group 21 – Project 1 Report

Suhaas Kanneganti, Pranav Arora, Victor Cruz Contreras, Derek Sprincis, Ayush Acharya, & Carlos Revilla

Summary #Suhaas

This study examines the influences of demographic indicators, notably age, gender, and marital status, on flu vaccination rates, by analyzing a dataset from the General Social Survey (GSS). Initially, data from the years 2018, 2021, and 2022 are filtered out of the larger dataset, which includes various variables from different years. The primary focus is on the 'age', 'sex', 'marital', and 'fluvax' columns, which indicate a person's age, gender, marital status, and whether they received the flu vaccine, respectively. After cleaning the dataset of missing values, histograms, and count plots are generated to visually assess the relationship between vaccine uptake and age. Moreover, count plots are also created to depict the relationship of vaccine uptake with gender and marital status. The same visualizations are repeated for the year 2022 alone, allowing for a more specific temporal insight. Using histograms for visualization facilitated an immediate comparison of vaccination statuses among various groups. One standout finding was the pivotal role age played: older individuals, especially those over 60, were significantly more inclined to get vaccinated, suggesting a pronounced awareness of their susceptibility to the flu. In contrast, gender didn't seem to have any discernible effect on vaccination decisions, as both genders mirrored similar patterns. Marital status, on the other hand, unfolded a more intricate picture. Individuals who'd never married potentially exhibited higher vaccination rates, which might be attributed to their higher social exposure, hinting at a relationship between social activity and health precautions. However, the research encountered certain limitations. The most noticeable was the absence of a broader temporal dataset, especially from years preceding the global COVID-19 pandemic. The pandemic undoubtedly influenced public perceptions about vaccinations, rendering the singular-year data from 2022 less comprehensive in gauging long-term trends or attitudes. While the current study carves out an essential understanding, the scope for further exploration remains vast. Investigating other potential determinants, like educational background, economic status, regional influences, or even political inclinations, could be monumental. This becomes especially pertinent in a post-pandemic world where health decisions and public perceptions are in constant change, necessitating robust research endeavors.

Data #Suhaas

Using Python as the tool for this exploration, the dataset was sourced from a CSV file titled gss.csv. The first step in the analysis involved importing essential libraries, namely NumPy and pandas. These libraries offer the capabilities needed for data manipulation and analysis.

The initial observation revealed the dataset to be extensive, with 72,390 rows and 6,645 columns. Given the scope of the study, which intended to focus on the years 2018, 2021, and 2022, a subset of the data was isolated. Surprisingly, the data for 2019 and 2020 was missing. The analysis then honed in on the crucial variables, comprising three subject variables: Age, which is a continuous representation of the respondents' years; Sex, a categorical indication of gender; Marital status, which delves into the personal realm of an individual's marital choices, with categories like 'divorced', 'never married', 'widowed', and 'married'. The target variable, fluvax, was a binary indication of whether an individual has been vaccinated against the flu.

One of the pivotal steps in data analysis is data cleaning. Here, the integrity and quality of the data were paramount. Notably, rows containing missing values in the 'fluvax' and 'age' columns were removed. The rationale behind this was clear: an analysis revolving around vaccination rates would be skewed without these crucial data points. This decision underscores the significance of having complete and accurate data for drawing meaningful conclusions.

Upon further analysis, a significant challenge arose. The initial plan was to conduct an in-depth analysis spanning the three years: 2018, 2021, and 2022. Yet, it soon became apparent that after filtering out incomplete entries, only data from 2022 remained usable. This realization dramatically altered the trajectory of the study. What began as a multi-year analysis was narrowed down to focus solely on the most recent year, 2022.

Such unexpected changes during the data analysis process highlight the unpredictability and challenges inherent in research. While the scope of the project evolved, the data from 2022 still presented valuable insights. This study underscores the importance of a robust data cleaning and preparation process. As seen, the trajectory of a project can be substantially altered based on the quality and completeness of the data at hand. In the realm of health research, where decisions can directly impact public health policies and strategies, the accuracy and depth of data analysis are paramount.

Results # Writup by Derek, Carlos, and Victor; coding completed by Ayush & Pranav

The research objective was to compare vaccination rates based on several demographic variables. This objective was achieved through the creation of histograms that compared groups based on vaccination rates. Histograms are an effective way to present data because they provide a clear visual representation of data distribution, facilitate the identification of outliers, and make it easy to compare multiple datasets. They are easily understandable, offer insights into data skewness and kurtosis, and assist in decision-making by summarizing key features of the data. Histograms are valuable for a wide range of applications and help with data preprocessing and analysis. The data that was used for this project was demographic data. Histograms are highly effective for analyzing demographic data due to their ability to visually represent the distribution

of discrete variables as represented by population characteristics. Histograms are especially effective for side-by-side comparison of two categories that are constant between one another with the expectation of a single independent variable. The histogram displayed has no axial distortions so the distance between variables is not skewed for effect.

Basic quantitative features are not addressed directly in our conclusions due to a lack of a compelling need to include them. Showing the mean using the current mode of representing the data would be impractical, and there is no meaningful way to incorporate variance, median, quantiles, and correlation since each variable has a boolean state (vaccinated / not vaccinated). The variables that the data analyzed: gender, marital status, and age by its nature are either boolean or naturally bounded. Due to this, outliers in data did not need to be characterized and addressed.

It was found that older populations tend to be more vaccinated than younger populations, 60+ years are much more likely to be vaccinated vs. not vaccinated, and gender appears to not affect the vaccinated status of the population.

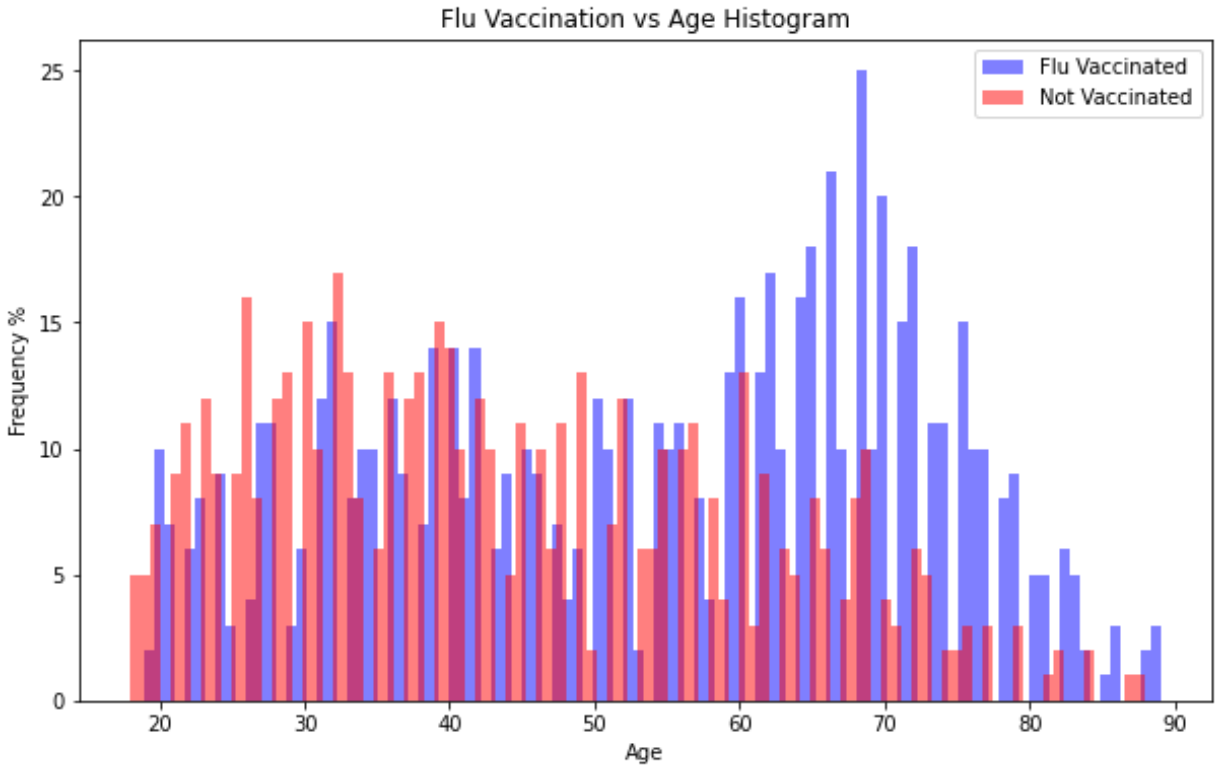

Figure 1

It was found that the population of the not vaccinated is generally a right-skewed normal distribution and the vaccinated appears to be a bimodal distribution but heavily skewed left. This may be the case since older people may prefer to be vaccinated to reduce the risk of suffering major health issues from the flu since they are overall in a weaker state of health.

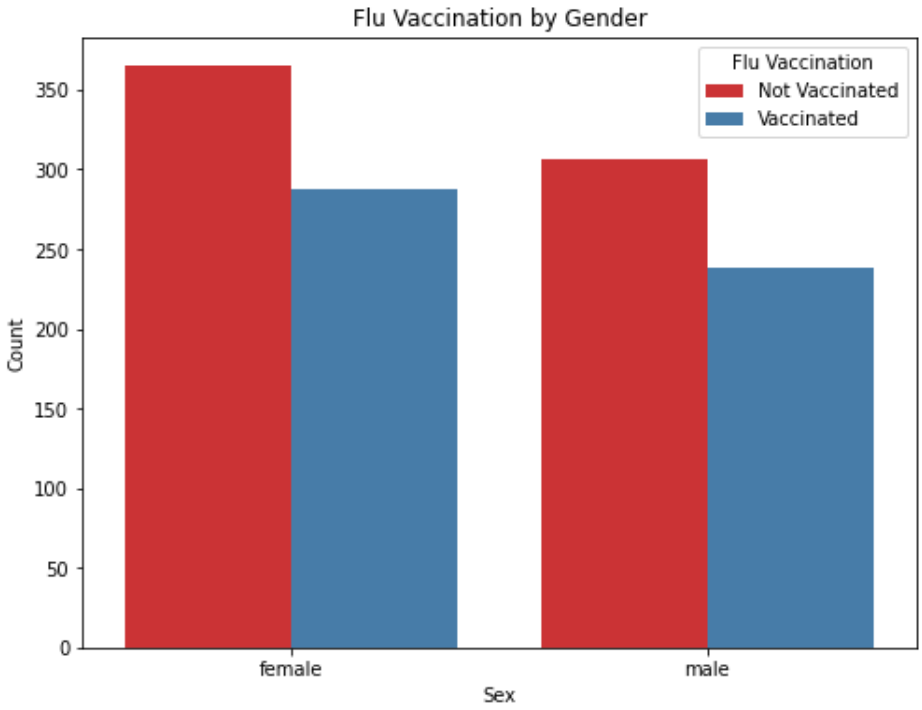

Figure 2

It was found that gender does not affect whether someone is more or less likely to be vaccinated with the difference between not vaccinated and vaccinated females being similar to the difference between not vaccinated and vaccinated males. The histogram also reveals that the majority of the surveyed population is unvaccinated rather than vaccinated.

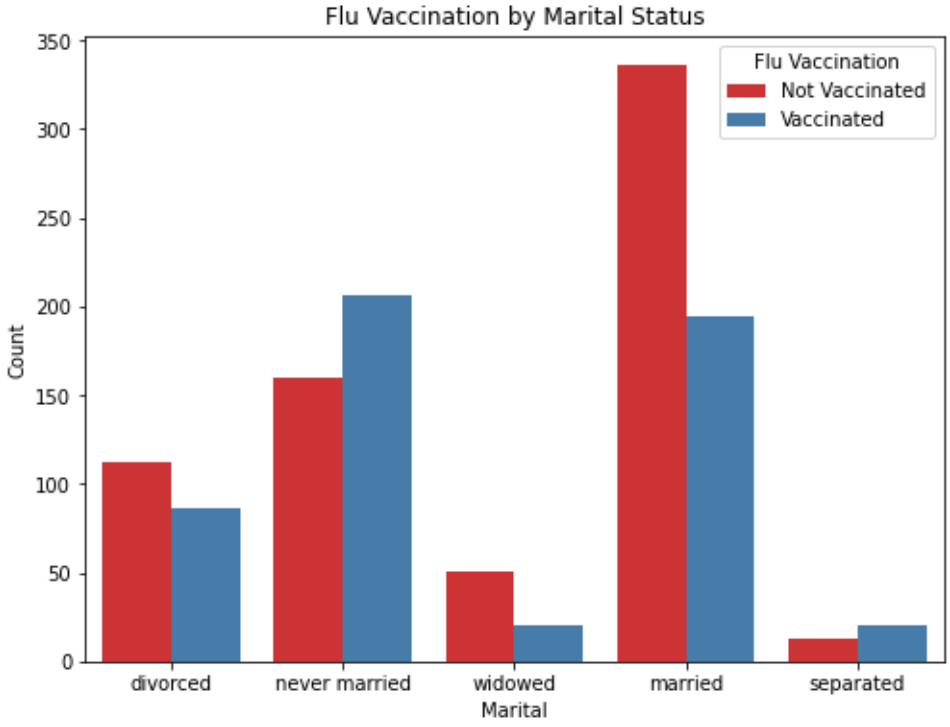

Figure 3

Vaccination rates based on marital status were correlated. A possible explanation for why ‘never married’ has a high vaccinated status is that people are more likely to be in more social situations trying to find a partner when they are never married and more likely to place a higher benefit of the flu shot when they are around a greater number of people. On the other hand, we overall notice a correlation between individuals who have been married before (‘divorced”, “married”, and “widowed marital”) and not being vaccinated.

Results were interpreted by looking at vaccinated / not vaccinated status, between different groups within that variable. To determine the statistical significance of differences in vaccination rates between demographic groups (age, marital status, gender), we employed statistical tests, such as chi-squared tests, and analyzed histograms. For instance, our analysis revealed that categories within the martial column are statistically correlated with the vaccination status as depicted by the p-value of 0.000000004, which is less than 0.05 (our chosen significance level). This indicates a statistically significant correlation between marital status and vaccination rates. Conversely, we found no significant correlation between gender and vaccination rates, as the p-value for gender was .94. Additionally, we observed that the older population tended to have higher vaccination rates.

The data we analyzed was solely from the year 2022. We would have preferred to analyze data from multiple years, specifically before the worldwide COVID-19 pandemic that occurred in 2020. An analysis would have been conducted both before and after the pandemic to observe if there were any significant changes in vaccination rates among different groups based on marital status, gender, and age. Although the dataset included data for 2019, we were unable to utilize it due to the presence of missing data, which hindered our ability to gain a comprehensive understanding across multiple years. COVID-19 may have had significant implications on how people perceive disease. Some individuals took it seriously, while others compared it to the flu. Additionally, there may have been a lack of trust in receiving any type of vaccination, as some major companies faced problems such as side effects or insufficient testing. To address the missing data from other years, we can search for additional datasets from different sources that meet our requirements to provide a more comprehensive picture of what factors lead to vaccination.

Conclusion: #Completed by Carlos

The research aimed to understand the vaccination rates among various demographic groups. The primary focus was on the effects of age, gender, and marital status on the likelihood of an individual receiving the flu vaccine. The implementation of the data collected into a histogram for visualizations provides a clear understanding of the distribution of the vaccination across different ages, genders, and marital statuses. The findings indicate that there is a higher inclination towards vaccination among the older population, specifically those aged 60 and

above. This trend underscores the perceived vulnerability of older individuals to the flu. On the other hand, the younger population exhibited a decreased tendency to get vaccinated, indicating potential areas of health promotion and awareness towards this demographic. In the case of gender, interestingly, this variable did not significantly influence the decision to get vaccinated, both male and female groups displayed similar trends of a higher proportion of individuals unvaccinated against individuals who were vaccinated. Regarding marital status, we discovered that never-married individuals tend to be more vaccinated, likely due to the social situations that they live within.

There are two main potential points of contention we aim to address within our research, one being the lack of quantitative metrics for analysis, and the second being the choice of histograms as the primary tool of histograms for data visualization. While one could critique the research for not including more quantitative aspects like median variance and correlation, it is important to note that the main variable under study is vaccination status, which is inherently binary (whether the individual is vaccinated or not). Because of the nature of the research question at hand, the quantitative metrics become less relevant in the justification of its analysis.

When we considered the use of histograms as a primary tool for data visualization, we used the tool to have a straightforward comparison between our discrete variables. Their simplicity ensured that the data's nuances were easily accessible without the distraction of more complex visualization techniques. Despite this, our statistical findings further refine the research's conclusions. For instance, the p-value for the correlation between Gender and Fluvax being 0.9264 firmly validates the observation that gender does not significantly influence vaccination choice. In contrast, the statistically significant p-value of 0.000000004 indicates a correlation between Marital Status and Fluvax and hints at an unexplored area of investigation. This underscores the potential influence of marital status on one's vaccination decision.

To reiterate, the question at hand was to research the binary nature of the result (vaccinated or not vaccinated) due to various other factors involving age, gender, and marital status. To maintain a clear and easily understandable report, it was important to provide clear and simple metrics to properly convey the results of the study. While more quantitative and descriptive metrics could have provided more evidence to support findings, they were not necessary in the context of the report at hand.

In reflecting upon the final report of our study, we noticed potential areas where additional research could have been provided. Some areas that could have given more information on the decision of whether or not to be vaccinated could have included, but are not limited to:

1. *Education Level:* It could be useful to study the correlation between an individual's level of Education and their vaccination status. This could provide insight into the information

and awareness of vaccination at different education levels. Important questions could be asked such as: “Does an individual understand why they need to be vaccinated?” “Are individuals who are not vaccinated skeptical of the vaccine because they are unaware of the contents of the vaccine?” And if so, “is there a distrust in the institution administering the vaccination for this reason or other reasons? Why?”

2. *Economic Status:* “Does one's economic status affect whether they choose to get vaccinated?” Exploring this question could provide insights into whether cost or accessibility barriers exist for certain individuals
3. *Region/ Location:* More broadly, if we can determine that region/ location correlates with the choice to get vaccinated, we could potentially point to more pressing issues at hand within a region. For instance, we could examine if the region has a greater or lower awareness campaign for the flu vaccination, or if there is more accessibility to health care in a region.
4. *Political Affiliation:* “Does one’s political views impact their decision on whether they choose to get vaccinated?” Some parties have different agendas and views on vaccination, leading to potential variances in vaccination statuses.
5. *COVID-19:* “Did COVID-19 impact vaccination rates?” With mask-wearing and isolation reducing the risk of getting sick, did people still choose to get vaccinated for the flu? Or did some individuals opt for COVID-19 vaccination while neglecting the flu vaccine? The extensive media coverage of COVID-19 overshadowed the flu, and there was significant skepticism surrounding the COVID-19 vaccine, which could have impacted the trust in other vaccines.
6. *Additional Data:* It would be beneficial to combine data from multiple datasets to conduct a more comprehensive analysis across different years and observe if these trends remained consistent over time.

Regarding the report's analysis, it could have offered a more comprehensive exploration of the primary factors influencing the vaccination decision. Such additional insights could have greatly contributed to addressing the questions raised earlier. This could have been collected through survey/ qualitative interviews which would have been crucial in observing these trends. While this research has shed valuable light on the decision-making of an individual getting vaccinated, there is always more to be done, as the realm of public health is ever-evolving and new information regarding the issue at hand appears daily (which is only accelerated with the rise of information and misinformation through social media). Nonetheless, this project serves as a foundational step in paving the way for more in-depth investigations in the future/