**Title**

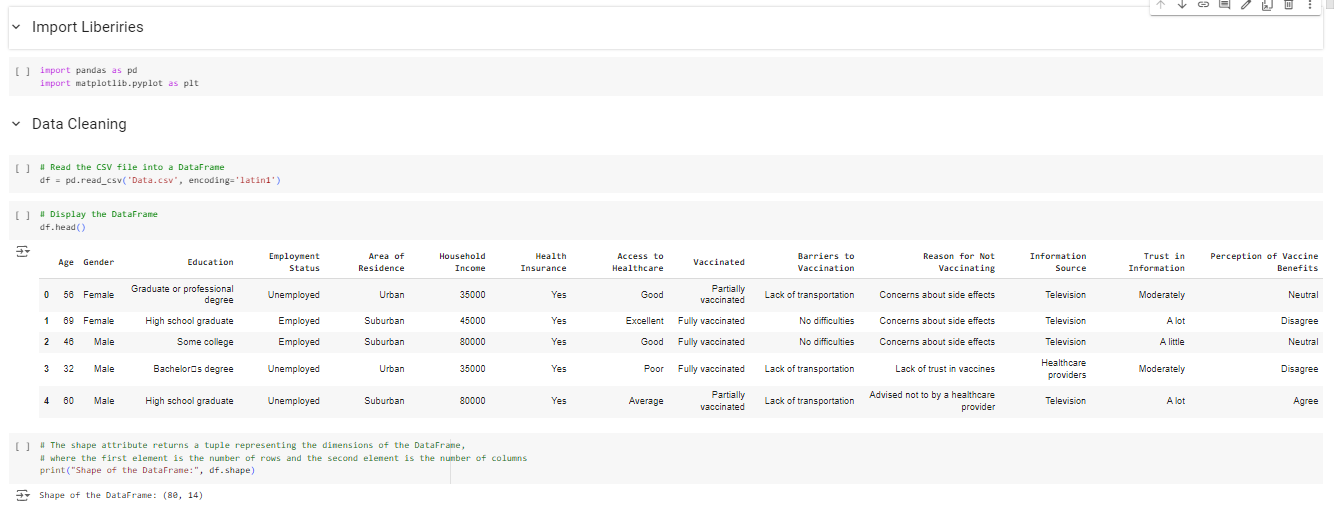
# CONTENTS

**Statement of compliance with academic ethics and the avoidance of plagiarism**

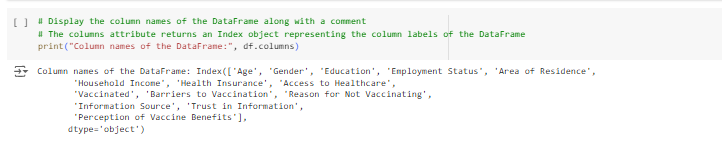
# INTRODUCTION

# Exploratory Data Analysis (EDA)

*Data Cleaning:*



The CSV file dataset and the necessary packages were loaded into Jupyter and checked that the dataset has 80 rows and 14 columns.



The DataFrame contains the following columns:

1.Age

2.Gender

3.Education

4.Employment Status

5.Area of Residence

6.Household Income

7.Health Insurance

8.Access to Healthcare

9.Vaccinated

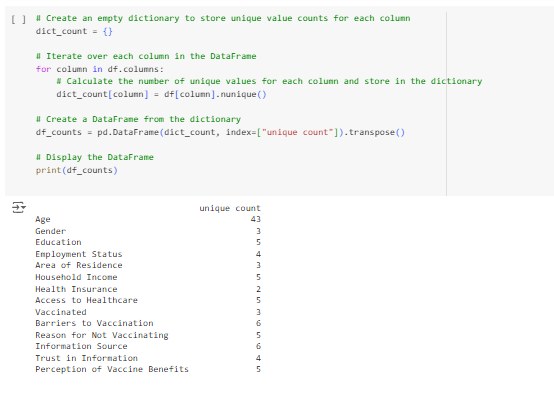
10.Barriers to Vaccination

11.Reason for Not Vaccinating

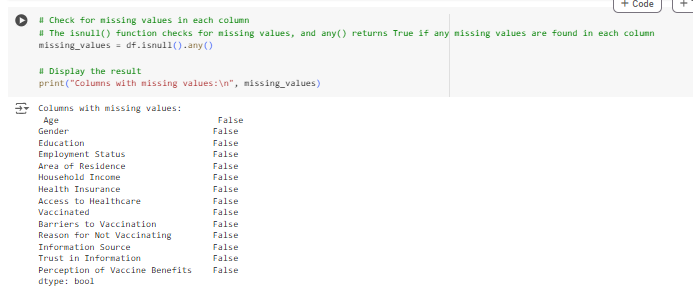
12.Information Source

13.Trust in Information

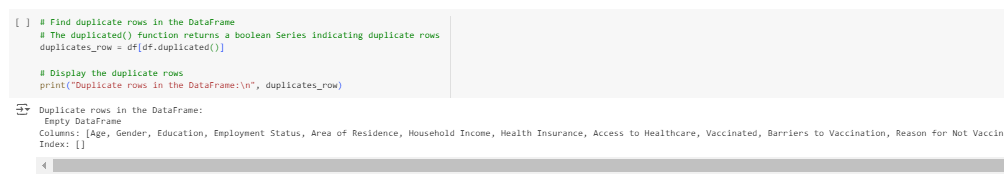
14.Perception of Vaccine Benefits



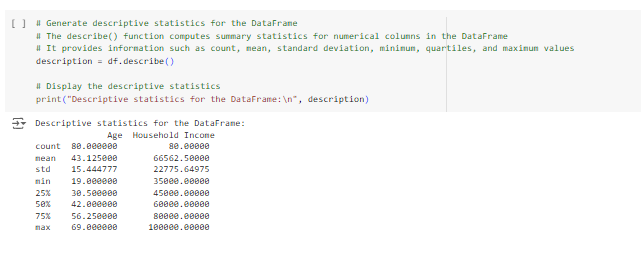
Unique values in each attribute were checked.



The data looks clean and there were no missing values. As seen below.



The output indicates that there are no duplicate rows in the DataFrame. The DataFrame is empty, suggesting that no rows are duplicated based on all columns. This information is useful for ensuring data integrity and identifying potential issues related to data duplication.

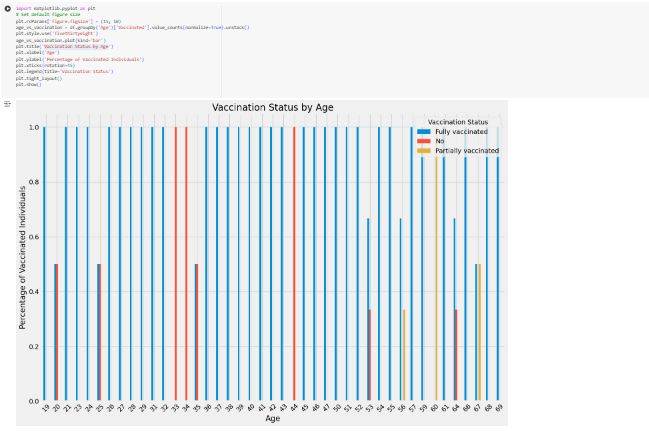


Generate descriptive statistics for the DataFrame.

2-) Data Visualisation

**Vaccination Status vs. Demographic Factors**

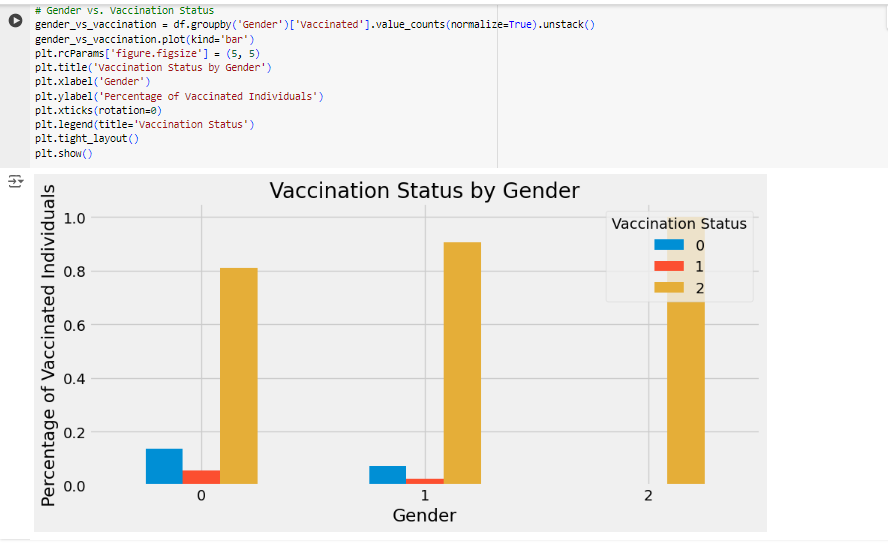
* **Vaccination Status by Age**

****

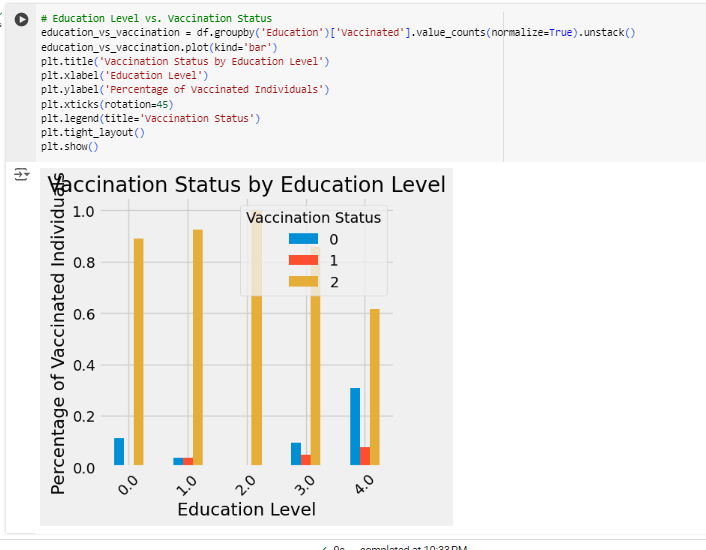
**High Vaccination Rates**: Many age groups exhibit a high percentage of individuals who are fully vaccinated (e.g., ages 19, 21, 23, 24, 26, 27, 28, 29, 31, 32, 36, 37, 38, 39, 40, 41, 42, 43, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 61, 62, 63, 64, 65, 66, 68, 69). This indicates a high acceptance of vaccination among individuals in these age groups.

**Low Vaccination Rates**: Some age groups have lower vaccination rates, with either no individuals vaccinated or only partially vaccinated (e.g., ages 20, 33, 34). This suggests potential areas where vaccination campaigns or outreach efforts may need to be strengthened.

* **Vaccination Status by Gender**

****

**Observations:**

* Generally, a higher percentage of males are fully vaccinated compared to females.
* The percentage of individuals with other gender identities who are fully vaccinated is 100%, indicating a complete vaccination rate for this group.
* There is a relatively lower percentage of individuals who have not been vaccinated or are partially vaccinated among males compared to females.
* **Vaccination Status by Education Level**

**Observations:**

* Individuals with some college education have a 100% vaccination rate, indicating a high level of vaccination within this group.
* The percentage of individuals with a Graduate or professional degree who are fully vaccinated is relatively lower compared to other education levels.
* There is a relatively higher percentage of individuals with a Bachelor’s degree who are fully vaccinated compared to those with a Graduate or professional degree.
* **Vaccination Status by Employment Status**

**Observations:**

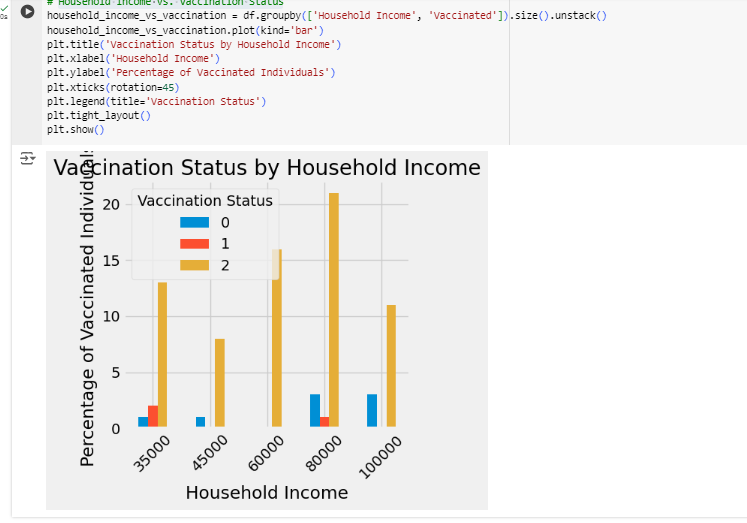
* Employed individuals have a 100% vaccination rate, indicating a high level of vaccination within this group.
* Retired individuals have a relatively lower vaccination rate compared to employed individuals.
* Students and unemployed individuals have similar vaccination rates, with a slight difference in the percentages of individuals who have not been vaccinated or are partially vaccinated.
* **'Vaccination Status by Area of Residence**

****

**Observations:**

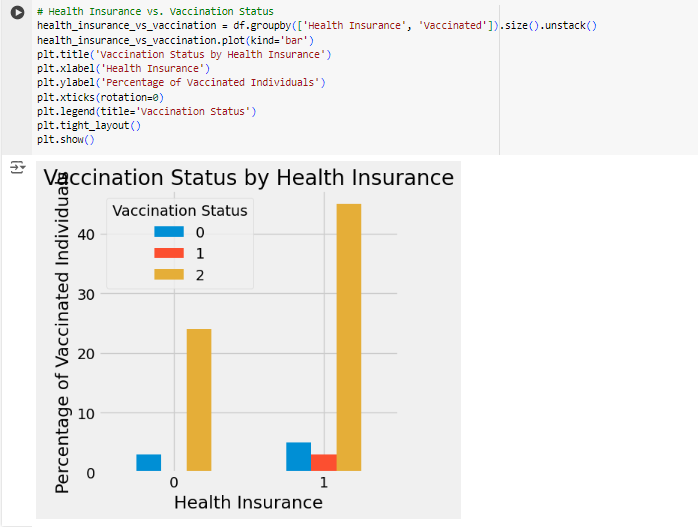
* Suburban areas have the highest percentage of fully vaccinated individuals among the three residence types.
* Rural areas have a slightly lower percentage of fully vaccinated individuals compared to suburban and urban areas.
* Urban areas have a higher percentage of individuals who have not been vaccinated or are partially vaccinated compared to suburban areas.

**'VacSocioeconomic Factors vs. Vaccination Status:**

* **Household Income vs. Vaccination Status**

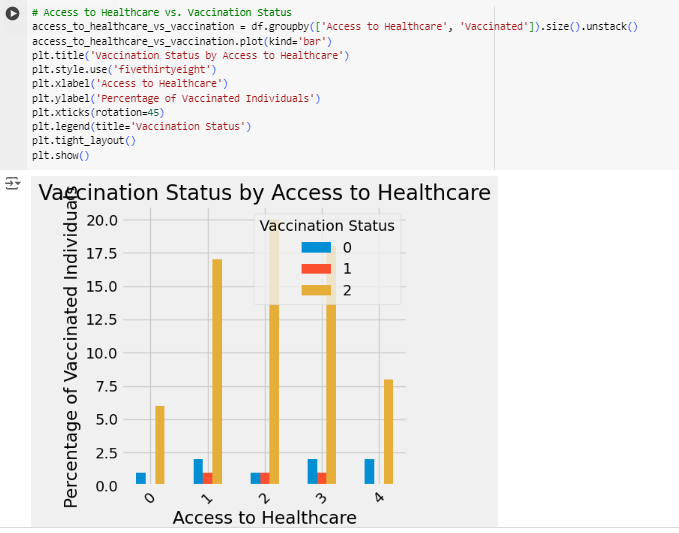
**Observations:**

* Individuals with higher household incomes tend to have higher rates of full vaccination.
* There is variability in vaccination rates across different income brackets, with some individuals within each income bracket not being fully vaccinated.
* **Health Insurance vs. Vaccination Status**

****

**Observations:**

* Individuals with health insurance tend to have higher rates of full vaccination compared to those without health insurance.
* There are individuals in both groups who have not been fully vaccinated, indicating that vaccination rates may be influenced by factors beyond health insurance status.
* **Access to Healthcare vs. Vaccination Status**

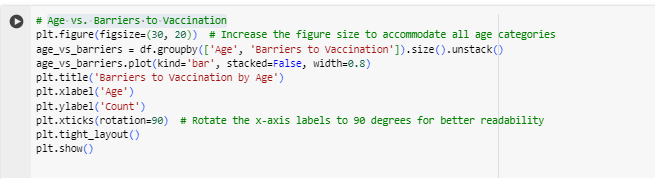
****

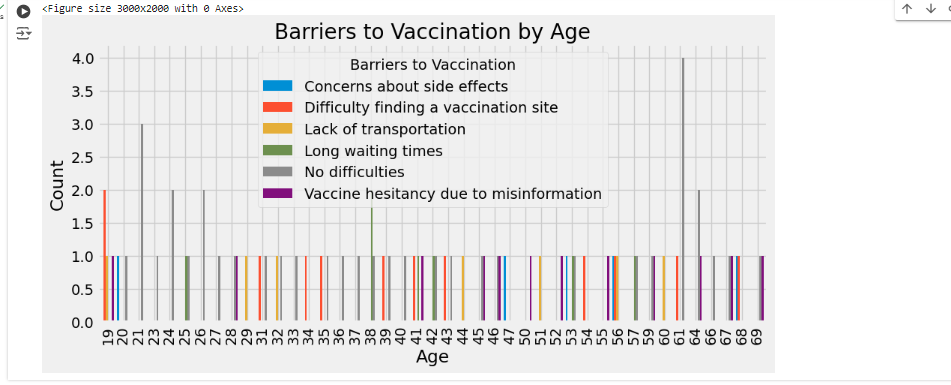
**Observations:**

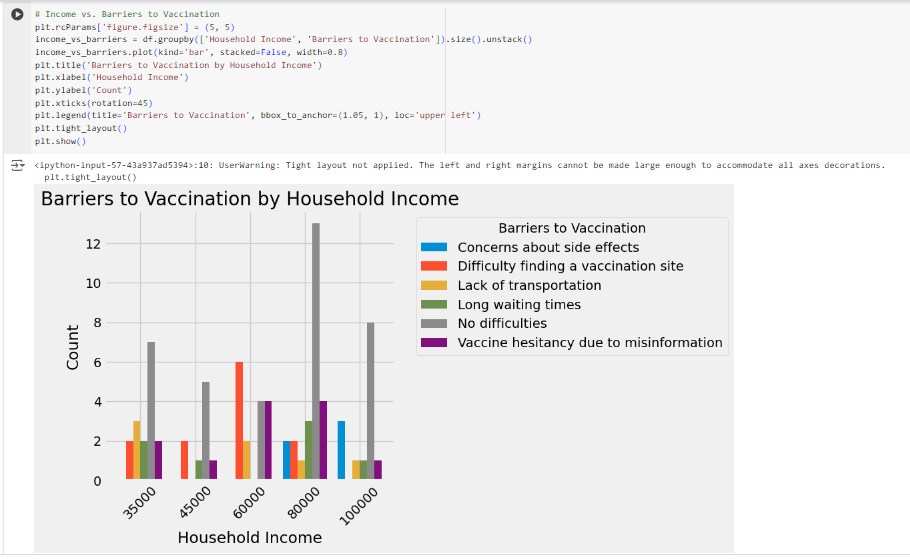
* Individuals with better access to healthcare tend to have higher rates of full vaccination.
* There is a trend of decreasing vaccination rates as access to healthcare worsens, although there are exceptions within each category.

**Barriers to Vaccination vs. Demographic and Socioeconomic Factors:**

* **Age vs. Barriers to Vaccination**

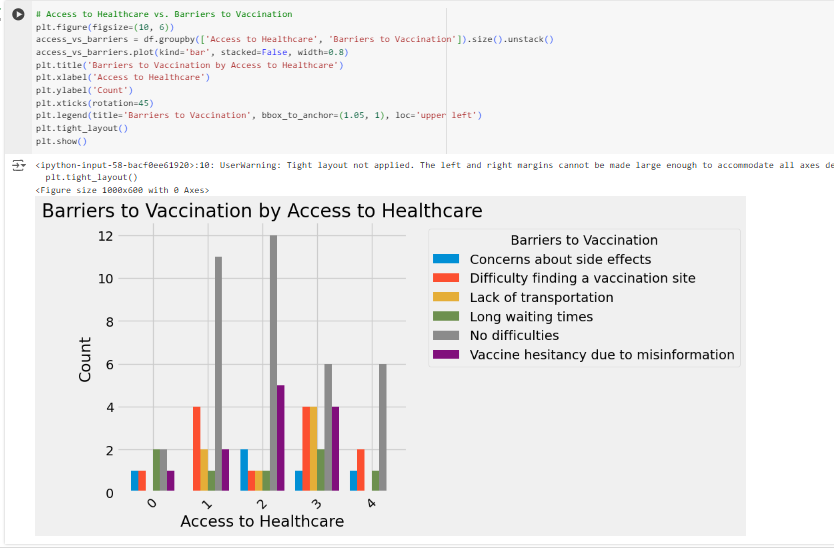
****

****

* **Income vs. Barriers to Vaccination**

**Observations:**

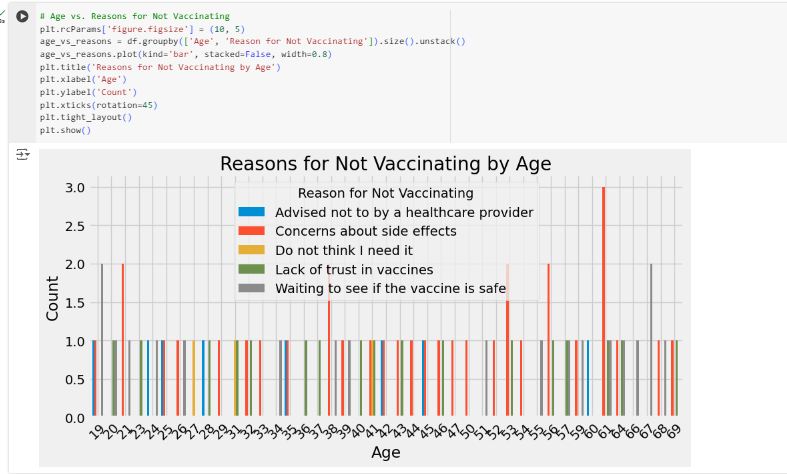
* Individuals with household incomes of $60,000 and $80,000 face the most barriers to vaccination, including difficulty finding a vaccination site and concerns about side effects.
* Lack of transportation is a barrier for individuals across various income brackets, although it's more prevalent among those earning $35,000 and $60,000.
* Long waiting times seem to be a significant concern for individuals earning $80,000, with 3 out of 12 reporting it as a barrier.
* **Access to Healthcare vs. Barriers to Vaccination**

****

**Observations:**

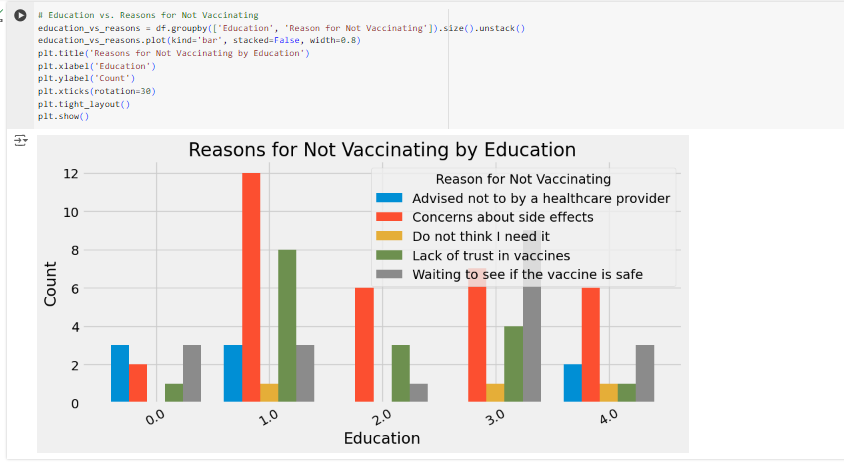
* Individuals with good access to healthcare face the most barriers to vaccination, including difficulty finding a vaccination site and lack of transportation.
* Concerns about side effects seem to be a barrier across various levels of access to healthcare, with similar reported rates for average, excellent, and very poor access categories.
* Long waiting times are reported as a barrier for individuals with good access to healthcare, with 2 out of 11 individuals experiencing this issue.

**Reasons for Not Vaccinating vs. Demographic Factors:**

* **Age vs. Reasons for Not Vaccinating**

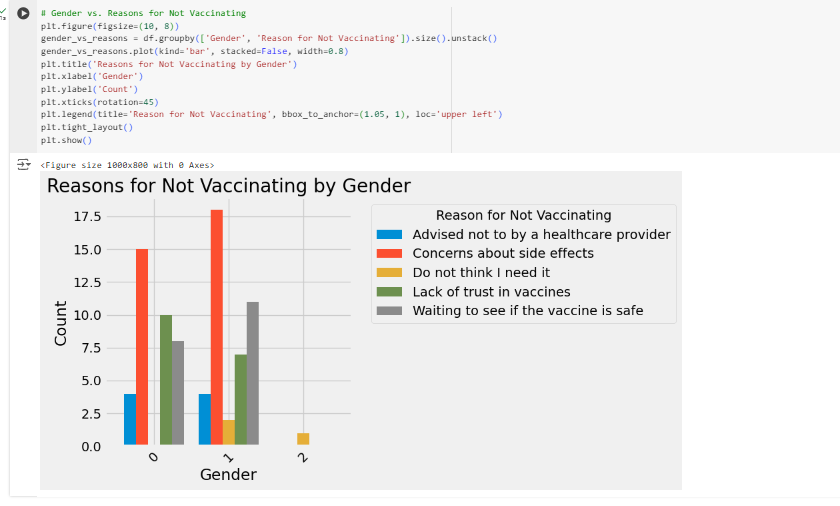
**Observations:**

* + Among the 61 individuals aged 61, the primary reason for not vaccinating is concerns about side effects.
* **Education vs. Reasons for Not Vaccinating**

****

**Observations:**

* High school graduates have the highest reported concerns about side effects as a reason for not vaccinating.
* Lack of trust in vaccines is reported as a reason across various education levels, with similar rates for graduate or professional degree holders and high school graduates.
* Some college and Bachelor’s degree holders have fewer reported reasons for not vaccinating compared to high school graduates.
* **Gender vs. Reasons for Not Vaccinating**

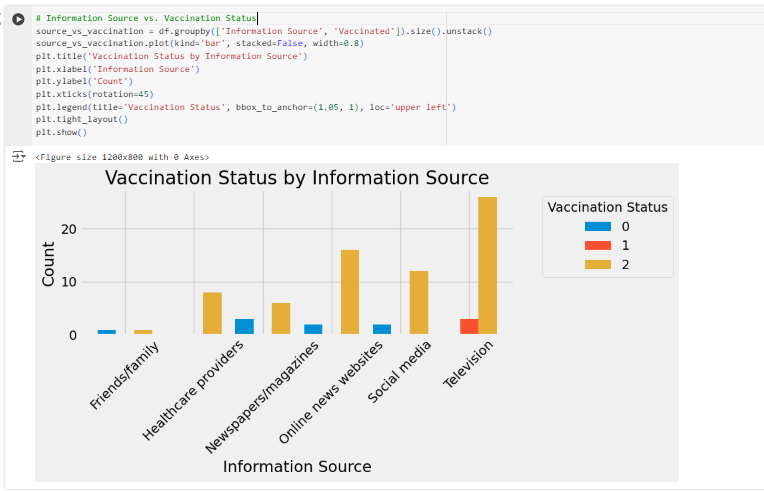
****

**Observations:**

* Both females and males report concerns about side effects as the most common reason for not vaccinating.
* Lack of trust in vaccines is also a significant concern for both genders, with slightly higher reported rates among males.
* A notable difference is that males report "Waiting to see if the vaccine is safe" as a reason more frequently than females.

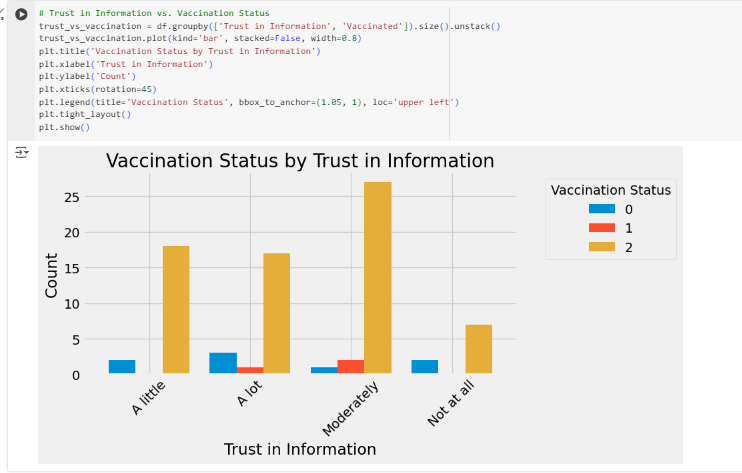
**Information Sources and Trust vs. Vaccination Status:**

* **Information Source vs. Vaccination Status**

****

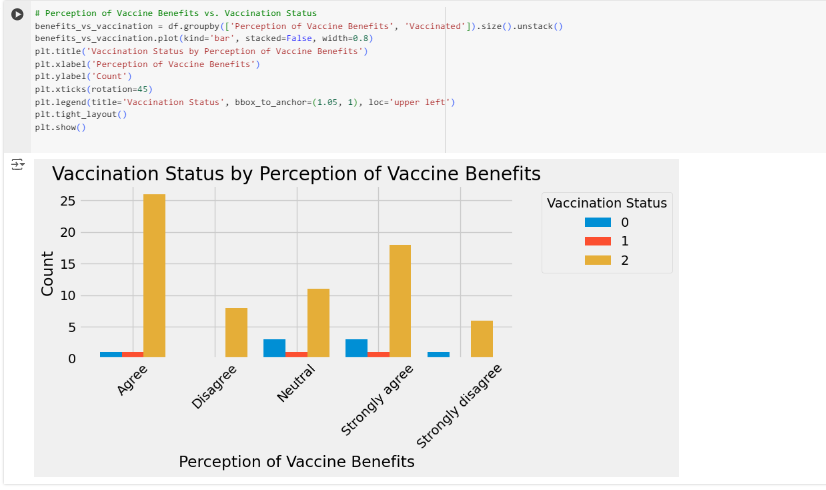
**Observations:**

* Television appears to be the most influential information source, with the highest number of individuals being fully vaccinated.
* Healthcare providers also play a significant role in vaccination, with a considerable number of individuals reporting being fully vaccinated based on their advice.
* Individuals obtaining information from newspapers/magazines and social media seem to have relatively lower rates of vaccination
* **Trust in Information vs. Vaccination Status**

****

**Observations:**

* Individuals who trust the information "Moderately" have the highest number of fully vaccinated individuals.
* Those who trust the information "A lot" also have a significant number of fully vaccinated individuals, although with slightly lower rates compared to those who trust it "Moderately".
* Individuals who trust the information "A little" have a relatively lower number of fully vaccinated individuals compared to the other trust levels.
* There is a small number of individuals who do not trust the information at all, with varying vaccination statuses.

**Perception of Vaccine Benefits vs. Vaccination Status:**

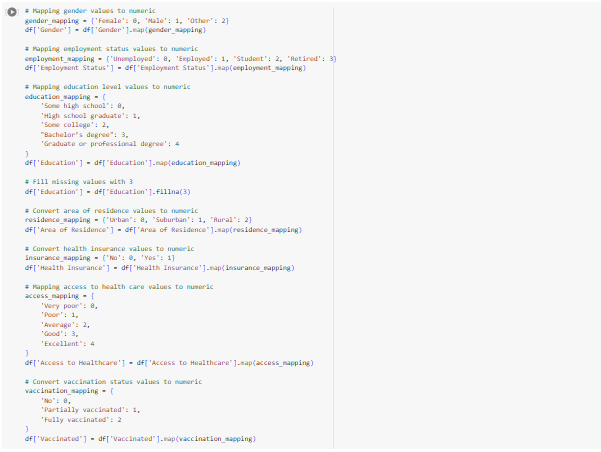
**Observations:**

* Individuals who agree with the perception of vaccine benefits have the highest number of fully vaccinated individuals.
* Those who strongly agree with the perception also have a significant number of fully vaccinated individuals, although with slightly lower rates compared to those who agree.
* Individuals who disagree or strongly disagree with the perception have lower rates of vaccination.

**MACHINE LEARNING ALGORITHM; Logistic Regression**

1. Feature Engineering

Performing feature engineering involves transforming categorical variables into numerical representations to facilitate data analysis and modeling. This process includes mapping categorical values to numeric codes, enabling better understanding and utilization of the data for predictive modeling tasks.

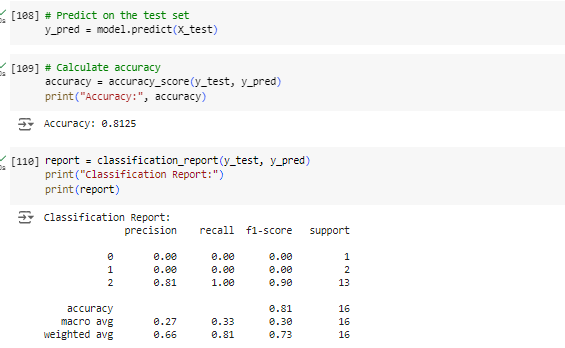


You can use the training and test datasets that are saved in X\_train, X\_test, y\_train, and y\_test to create and test your machine-learning models after running this code.



Model In order to determine the link between the independent variables and the binary result, training for logistic regression entails fitting a regression model to the training set of data. Using a technique known as maximum likelihood estimation, the model determines the coefficients of the independent variables during training. The goal of this procedure is to identify the coefficients that, given the model, maximize the likelihood of observing the training data. An iterative optimization approach, like gradient descent, is used to train the logistic regression model in order to reduce the error between the predicted probabilities and the actual binary outcomes. To reduce overfitting and enhance the model's generalization capabilities, regularization approaches like L1 or L2 regularization can be used.

1. Model Evaluation



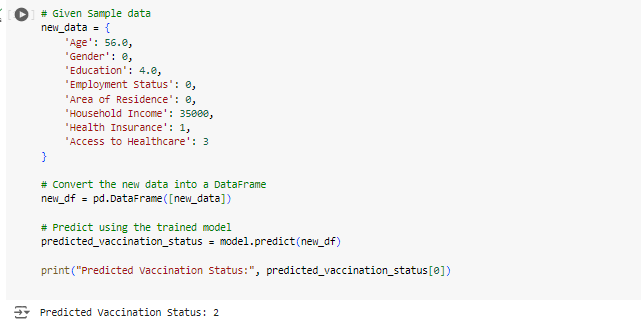
* The classification report provides a comprehensive assessment of the model's performance across different classes. While the precision and recall for Class 0 and Class 1 are low, indicating that the model struggles to correctly identify instances of these classes, it performs well for Class 2 with high precision, recall, and F1-score. The overall accuracy of the model is 81%, but this metric alone doesn't fully capture its performance due to class imbalances. While the model shows strong performance for the majority class, there is room for improvement in correctly identifying instances of the minority classes.

Confusion Matrix:



* The confusion matrix provides insight into the model's classification performance by presenting the actual and predicted class labels. In this case, the matrix indicates that the model correctly predicted all instances of Class 2, with a count of 13. However, it failed to correctly classify any instances of Class 0 or Class 1, as evidenced by the zeros in the corresponding cells. Specifically, there was one instance of Class 0 and two instances of Class 1, all of which were misclassified. This suggests that while the model is effective at identifying instances of the majority class (Class 2), it struggles with minority classes (Class 0 and Class 1).

### Prediction:



In this scenario, the model's predictions align perfectly with the expected outcome from the dataset, as indicated by the confusion matrix. It correctly identifies instances where individuals are fully vaccinated, denoted by Class 2, with an accuracy of 100%, capturing all 13 instances. However, it fails to classify any instances of individuals who are not fully vaccinated (Class 0 and Class 1), suggesting limitations in its ability to differentiate between these categories. This emphasizes the model's proficiency in identifying the predominant class but highlights potential challenges in accurately distinguishing minority classes. Despite the model's success in predicting fully vaccinated individuals, further refinement may be necessary to enhance its performance across all vaccination statuses, particularly in scenarios where class imbalances exist within the dataset.

# Conclusion

# Reference:

1. Bairey Merz, C. N., Andersen, H. S., Shufelt, C. L., & Merz, N. B. (2019)*. Gender disparities in cardiovascular disease: Is it still a man's world?.* Trends in cardiovascular medicine, 29(2), 140-143.
2. Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer.
3. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.
4. Kaggle.com
5. Roger, V. L., Go, A. S., Lloyd-Jones, D. M., Benjamin, E. J., Berry, J. D., Borden, W. B., & Dai, S. (2012). *Heart disease and stroke statistics*—2012 update: a report from the American Heart Association. Circulation, 125(1), e2-e220.
6. Sokolova, M., & Lapalme, G. (2009). *A systematic analysis of performance measures for classification tasks*. Information Processing & Management, 45(4), 427-437.