VAXRISK: ANALYSIS OF VAERS REPORTS FOR VACCINE SAFETY ASSESSMENT

DATA 606 Capstone in Data Science
Under the guidance: Ozgur Ozturk

Chandrikarani Vaidya
Terala Bhuvana Chandrika
Sai Ganesh Donka

TABLE OF CONTENTS

1. Introduction
2. Background
3. Data Description
4. Exploratory Data Analysis and Data engineering
5. Project Scope

INTRODUCTION

In an era of heightened focus on public health and vaccine safety, our project aims to analyze and predict Adverse Events (AEs) associated with vaccines using data from the Vaccine Adverse Event Reporting System (VAERS). Our primary goal is to develop a model that predicts the risk of AEs and create a tool to help users assess vaccine safety.

What is an Adverse Event (AE)?
An Adverse Event is a harmful or unintended outcome that occurs after a patient receives medical care, including vaccination. AEs can range from mild (e.g., soreness at injection site) to severe (e.g., anaphylaxis).

PROJECT RATIONALE

01

Vaccines are crucial for public health, but no vaccine is entirely risk-free

02

Understanding which vaccines cause more severe reactions for certain individuals is vital to minimize risk and improve patient safety. This project is timely, given the increased focus on vaccination safety during the COVID-19 pandemic.

03

This research is particularly relevant given the increased focus on vaccination safety these days and the ongoing need for effective, safe immunization programs.

RESEARCH QUESTION AND HYPOTHESIS

Can we predict whether a vaccine will cause a serious adverse event based on patient demographics, health history, and symptoms?

Certain vaccines have a higher likelihood of causing serious adverse reactions in patients with specific characteristics, and this risk can be predicted using machine learning models.

MOTIVATION BEHIND THE STUDY

Over 1 million COVID-19 vaccine injuries have been reported, highlighting the need to identify factors contributing to adverse outcomes.

Serious adverse events (AEs), though rare, can lead to severe complications include severe allergic reactions, seizures, and life-threatening complications.

A CDC investigation into the Janssen vaccine revealed higher rates of fainting, emphasizing the importance of monitoring vaccine recipients.

OVERVIEW OF SIMILAR APPROACHES

1. State of the Art

Existing Studies: There has been extensive research on vaccine safety, focusing on adverse events reported in the VAERS database. Studies have used machine learning to predict serious vs. non-serious outcomes, but most focus on one specific vaccine or short-term side effects.

VAERS Monitoring Tools: The Food and Drug Administration (FDA) and Centers for Disease Control and Prevention (CDC) already uses data mining techniques to detect safety signals, but these models are general and not tailored to individual patient characteristics.

2. What's Missing?

Personalized Prediction: Most existing approaches do not tailor predictions to specific individuals based on their demographics and medical history. There is also limited focus on predicting future risks for patients considering vaccination.

INTRODUCTION TO DATASET

- **Data source:** VAERS data link
- **Dataset characteristics:**

It consists of 3 CSV files per year (1990-2024):

1.VAERSDATA.CSV: Contains demographic and AE information

2. VAERSVAX.CSV: Vaccine-specific data

3. VAERSSYMPTOMS.CSV: Detailed symptom information

- **Data volume:**

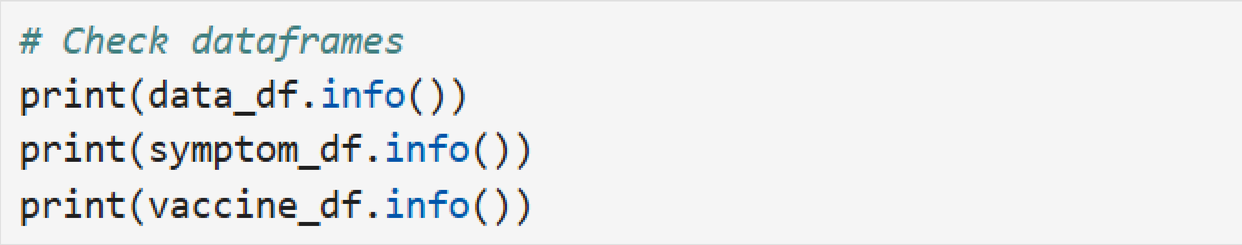
The total size of the zip file is 505.96 MB

We are using data from years 2015 to 2024 which is 1.3 GB which has 47 columns and 2106687 rows

- **Data quality considerations:**

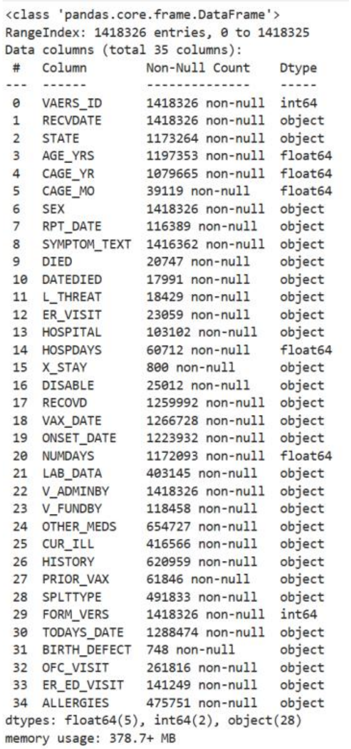
Self-reported data may include biases and inconsistencies, and not all AEs are reported to VAERS

DATA DESCRIPTION

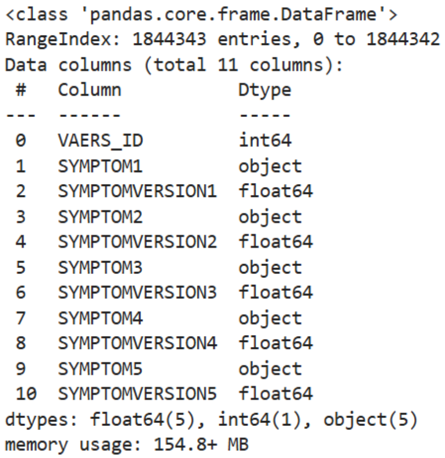


```python
# Check dataframes
print(data_df.info())
print(symptom_df.info())
print(vaccine_df.info())
```

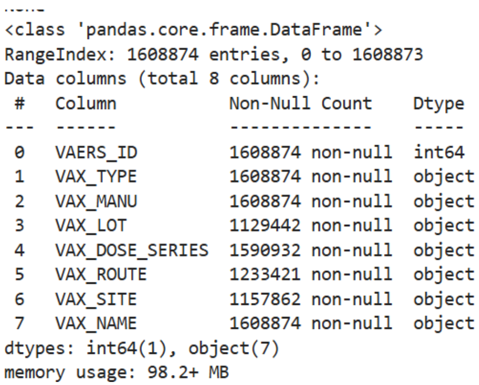
Data_df- dataframe for Vaers Data
Symptom_df- dataframe for vaers Symptoms
Vaccine_df – dataframe for vaers Vaccine



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1418326 entries, 0 to 1418325
Data columns (total 35 columns):
 #   Column         Non-Null Count    Dtype
---  ------         --------------    -----
 0   VAERS_ID       1418326 non-null  int64
 1   RECVDATE       1418326 non-null  object
 2   STATE          1173264 non-null  object
 3   AGE_YRS        1197353 non-null  float64
 4   CAGE_YR        1079665 non-null  float64
 5   CAGE_MO        39119 non-null    float64
 6   SEX            1418326 non-null  object
 7   RPT_DATE       116389 non-null   object
 8   SYMPTOM_TEXT   1416362 non-null  object
 9   DIED           20747 non-null    object
 10  DATEDIED       17991 non-null    object
 11  L_THREAT       18429 non-null    object
 12  ER_VISIT       23059 non-null    object
 13  HOSPITAL       103102 non-null   object
 14  HOSPDAYS       60712 non-null    float64
 15  X_STAY         800 non-null      object
 16  DISABLE        25012 non-null    object
 17  RECOVD         1259992 non-null  object
 18  VAX_DATE       1266728 non-null  object
 19  ONSET_DATE     1223932 non-null  object
 20  NUMDAYS        1172093 non-null  float64
 21  LAB_DATA       403145 non-null   object
 22  V_ADMINBY      1418326 non-null  object
 23  V_FUNDBY       118458 non-null   object
 24  OTHER_MEDS     654727 non-null   object
 25  CUR_ILL        416566 non-null   object
 26  HISTORY        620959 non-null   object
 27  PRIOR_VAX      61846 non-null    object
 28  SPLTTYPE       491833 non-null   object
 29  FORM_VERS      1418326 non-null  int64
 30  TODAYS_DATE    1288474 non-null  object
 31  BIRTH_DEFECT   748 non-null      object
 32  OFC_VISIT      261816 non-null   object
 33  ER_ED_VISIT    141249 non-null   object
 34  ALLERGIES      475751 non-null   object
dtypes: float64(5), int64(2), object(28)
memory usage: 378.7+ MB
```



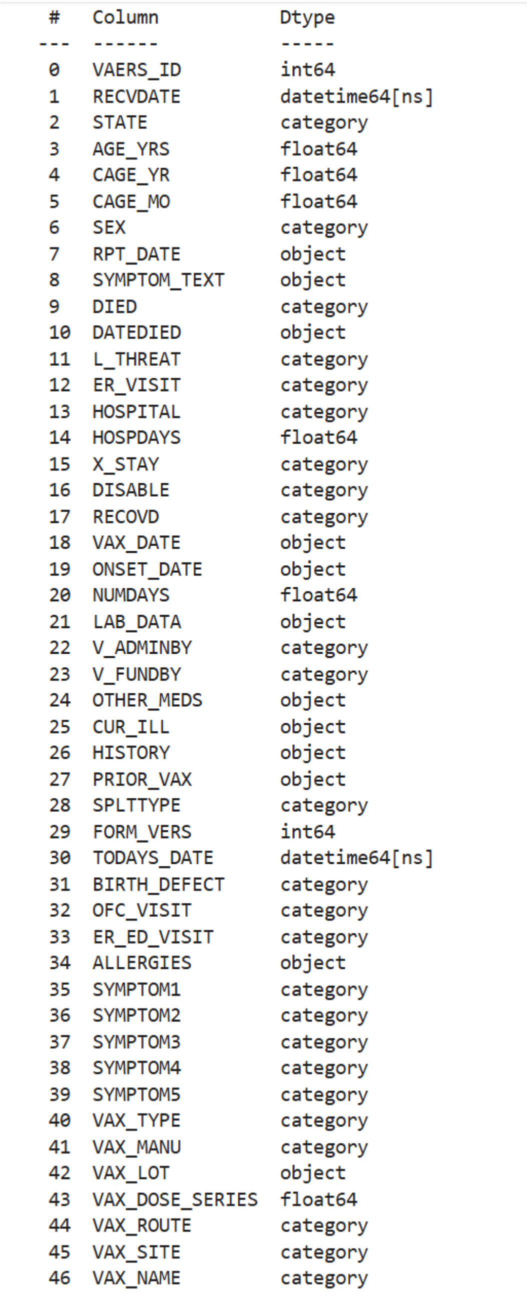
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1844343 entries, 0 to 1844342
Data columns (total 11 columns):
 #   Column           Dtype
---  ------           -----
 0   VAERS_ID         int64
 1   SYMPTOM1         object
 2   SYMPTOMVERSION1  float64
 3   SYMPTOM2         object
 4   SYMPTOMVERSION2  float64
 5   SYMPTOM3         object
 6   SYMPTOMVERSION3  float64
 7   SYMPTOM4         object
 8   SYMPTOMVERSION4  float64
 9   SYMPTOM5         object
 10  SYMPTOMVERSION5  float64
dtypes: float64(5), int64(1), object(5)
memory usage: 154.8+ MB
```



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1608874 entries, 0 to 1608873
Data columns (total 8 columns):
 #   Column           Non-Null Count    Dtype
---  ------           --------------    -----
 0   VAERS_ID         1608874 non-null  int64
 1   VAX_TYPE         1608874 non-null  object
 2   VAX_MANU         1608874 non-null  object
 3   VAX_LOT          1129442 non-null  object
 4   VAX_DOSE_SERIES  1590932 non-null  object
 5   VAX_ROUTE        1233421 non-null  object
 6   VAX_SITE         1157862 non-null  object
 7   VAX_NAME         1608874 non-null  object
dtypes: int64(1), object(7)
memory usage: 98.2+ MB
```

MERGING DATAFRAMES

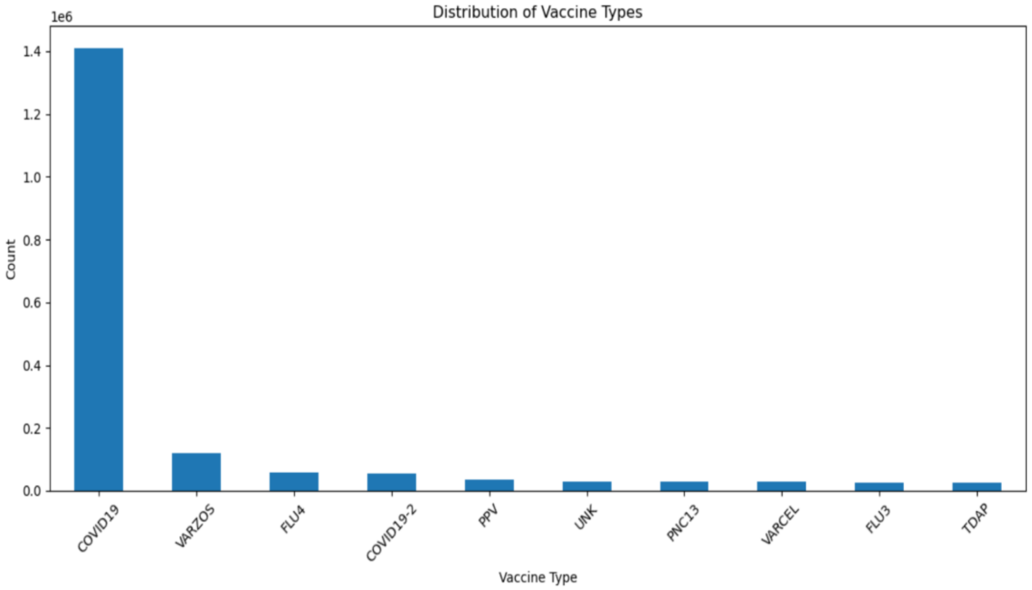
After merging all the 3 CSV we have:
2106687 entries, 0 to 2106686
With 47 columns

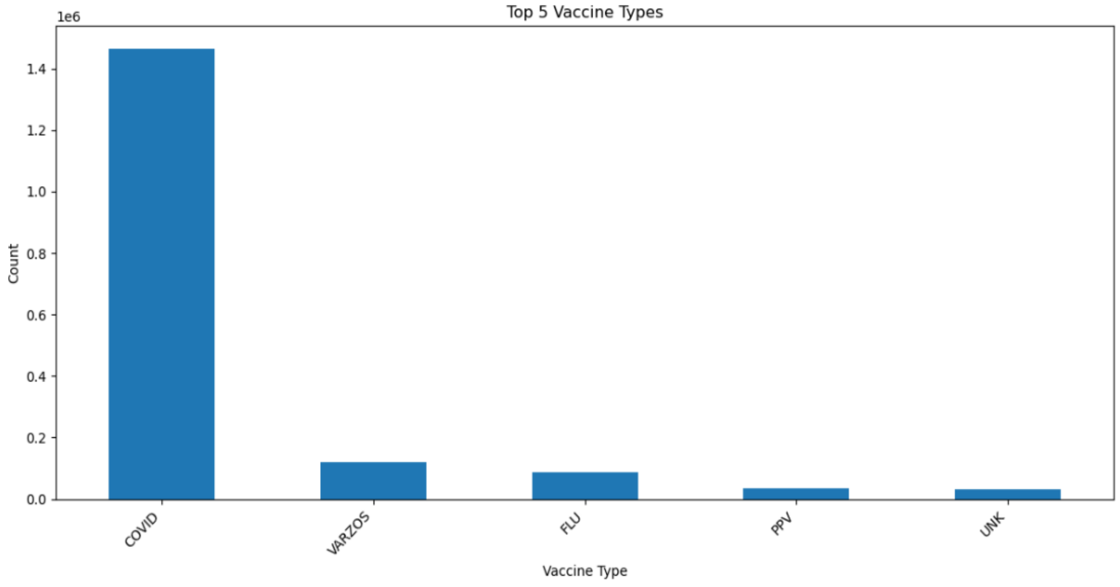


```
 #   Column           Dtype
---  ------           -----
 0   VAERS_ID         int64
 1   RECVDATE         datetime64[ns]
 2   STATE            category
 3   AGE_YRS          float64
 4   CAGE_YR          float64
 5   CAGE_MO          float64
 6   SEX              category
 7   RPT_DATE         object
 8   SYMPTOM_TEXT     object
 9   DIED             category
 10  DATEDIED         object
 11  L_THREAT         category
 12  ER_VISIT         category
 13  HOSPITAL         category
 14  HOSPDAYS         float64
 15  X_STAY           category
 16  DISABLE          category
 17  RECOVD           category
 18  VAX_DATE         object
 19  ONSET_DATE       object
 20  NUMDAYS          float64
 21  LAB_DATA         object
 22  V_ADMINBY        category
 23  V_FUNDBY         category
 24  OTHER_MEDS       object
 25  CUR_ILL          object
 26  HISTORY          object
 27  PRIOR_VAX        object
 28  SPLTTYPE         category
 29  FORM_VERS        int64
 30  TODAYS_DATE      datetime64[ns]
 31  BIRTH_DEFECT     category
 32  OFC_VISIT        category
 33  ER_ED_VISIT      category
 34  ALLERGIES        object
 35  SYMPTOM1         category
 36  SYMPTOM2         category
 37  SYMPTOM3         category
 38  SYMPTOM4         category
 39  SYMPTOM5         category
 40  VAX_TYPE         category
 41  VAX_MANU         category
 42  VAX_LOT          object
 43  VAX_DOSE_SERIES  float64
 44  VAX_ROUTE        category
 45  VAX_SITE         category
 46  VAX_NAME         category
```

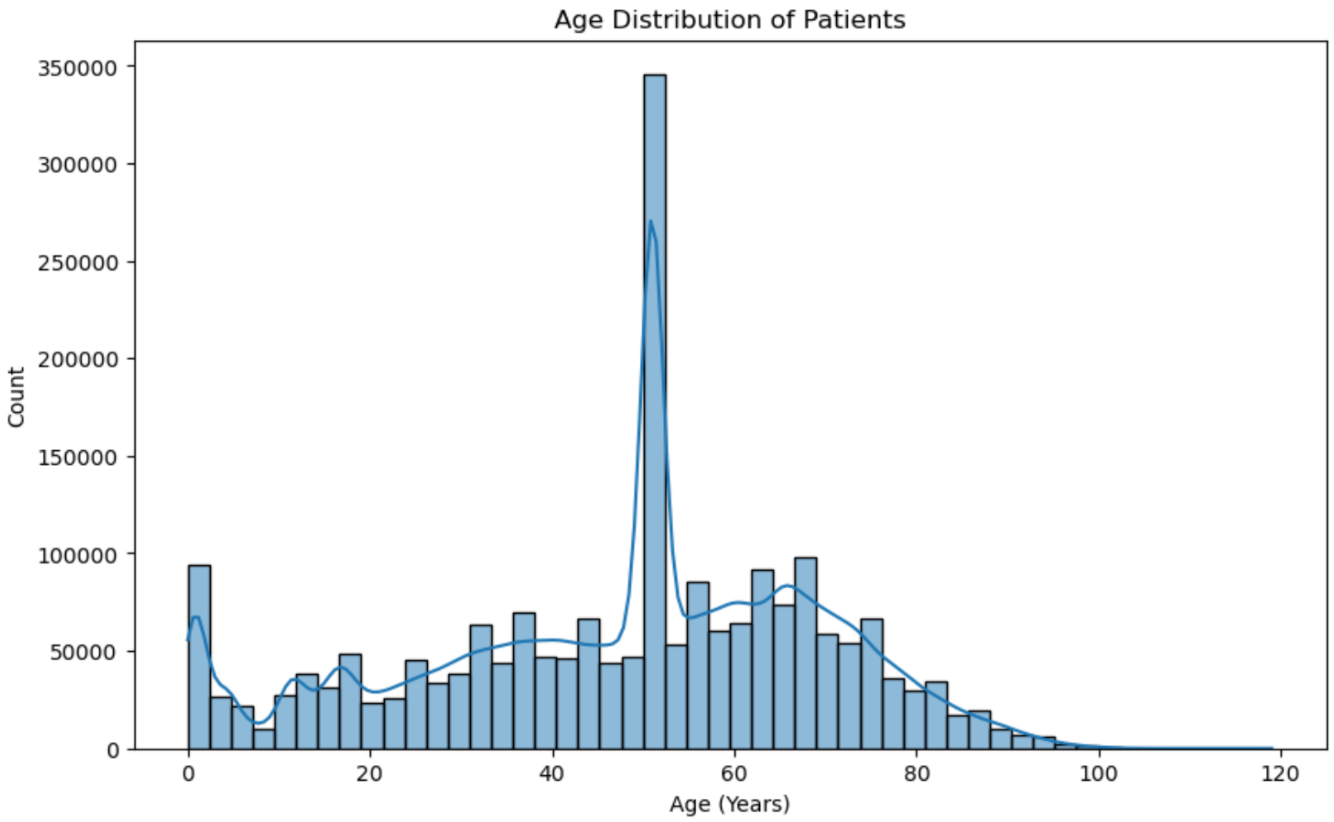
EXPLORATORY DATA ANALYSIS

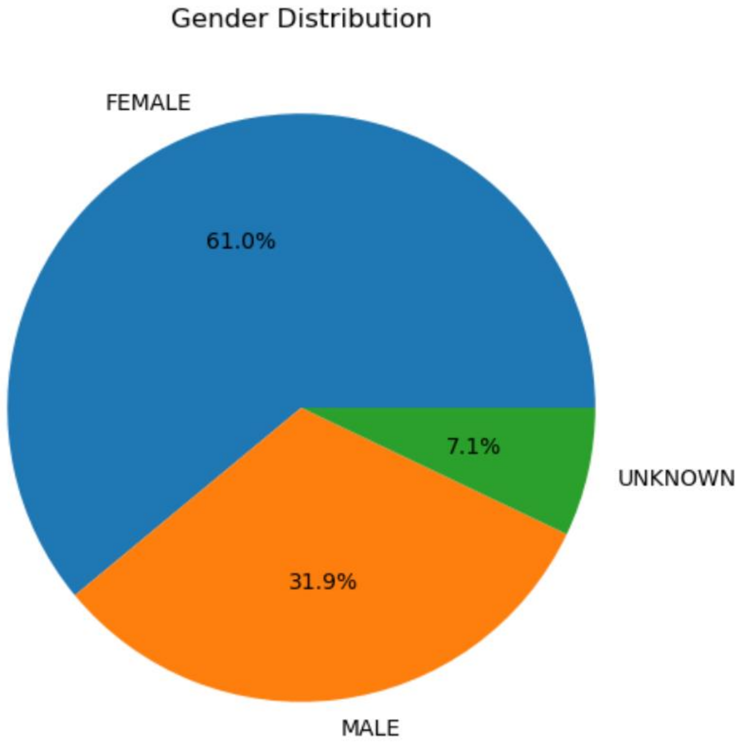
Understanding Vaccine type



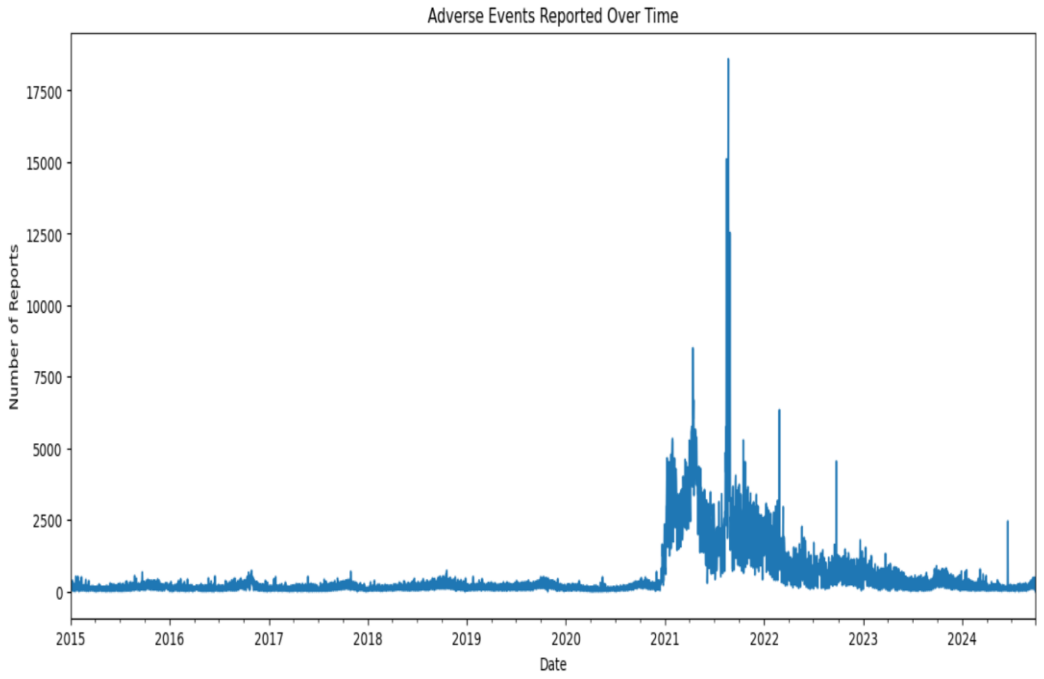


Understanding the demographics: Age and Gender

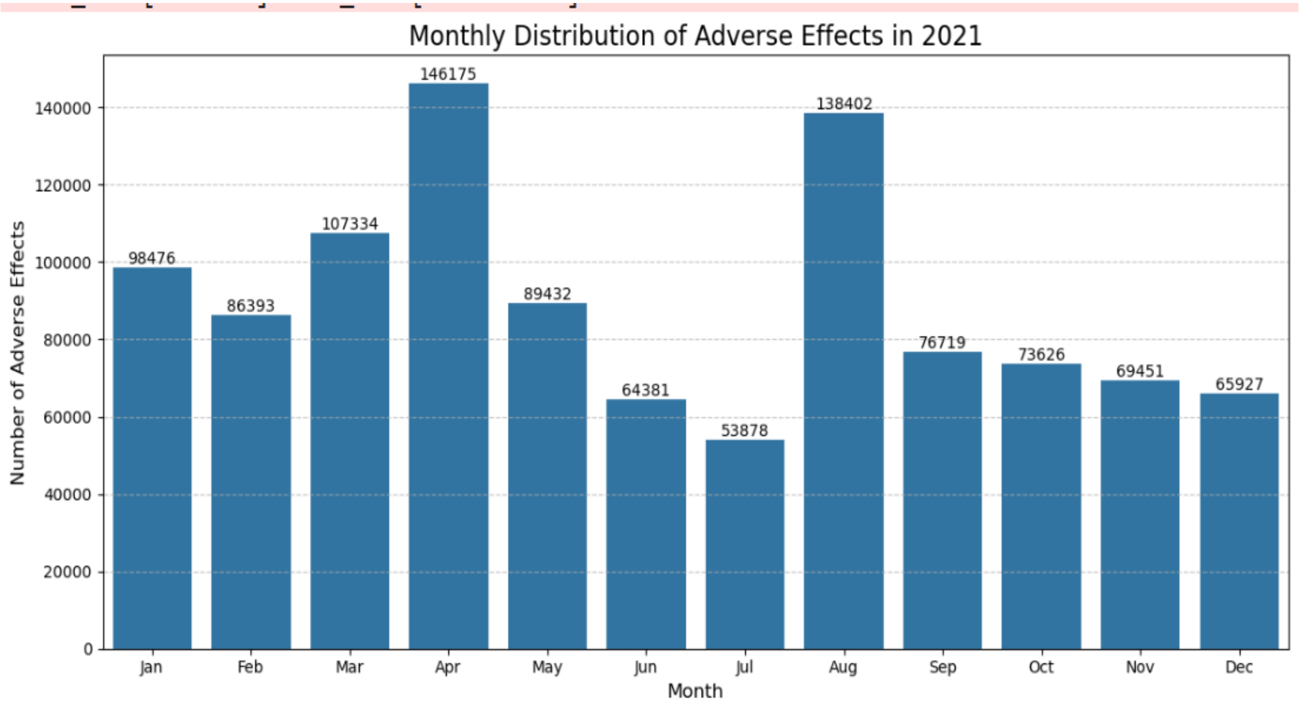




Understanding Adverse events over the time

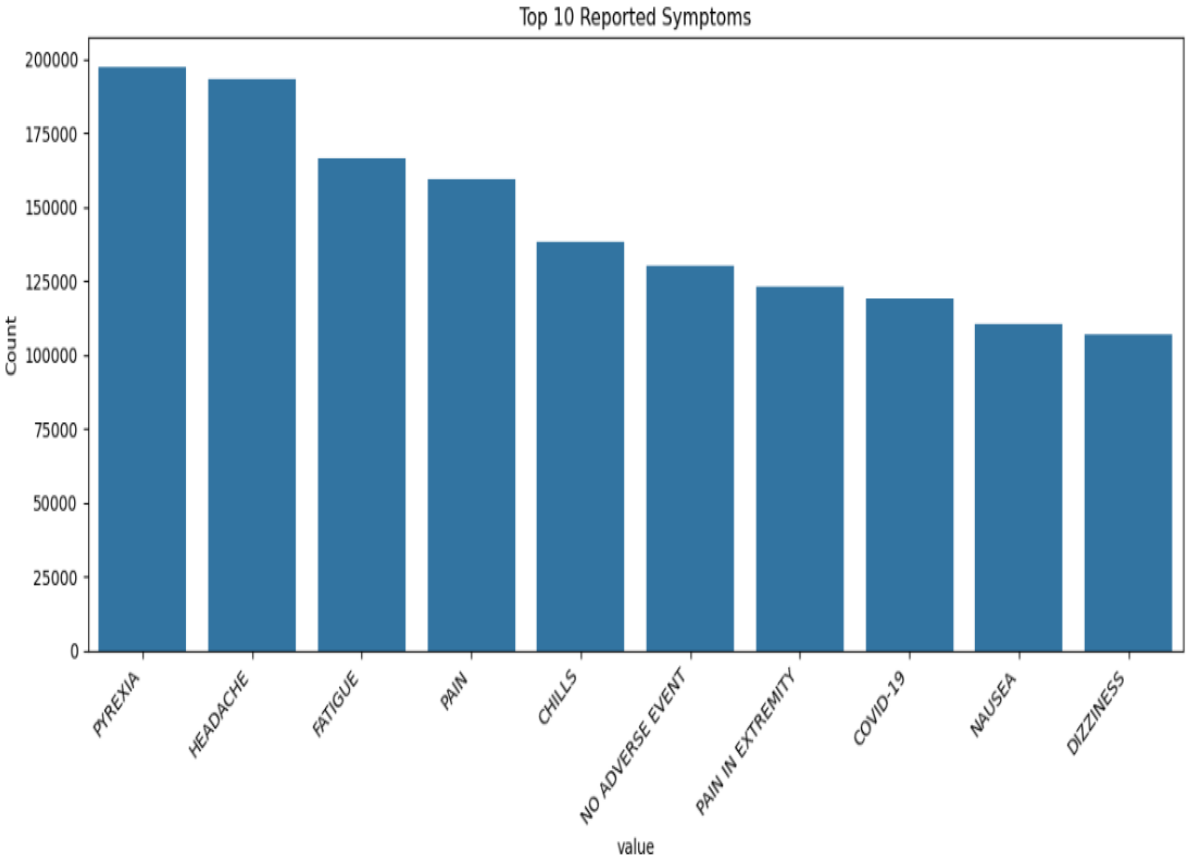


Adverse Events Reported Over Time

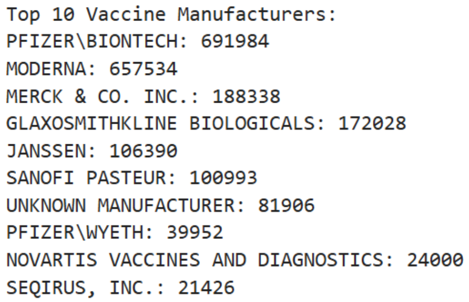


Monthly Distribution of Adverse Effects in 2021

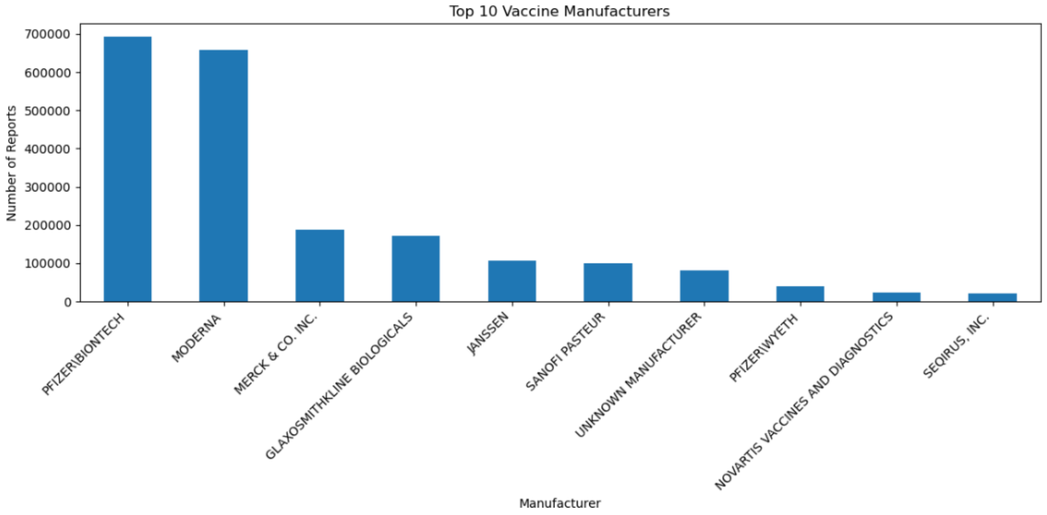
What are the top 10 symptoms?



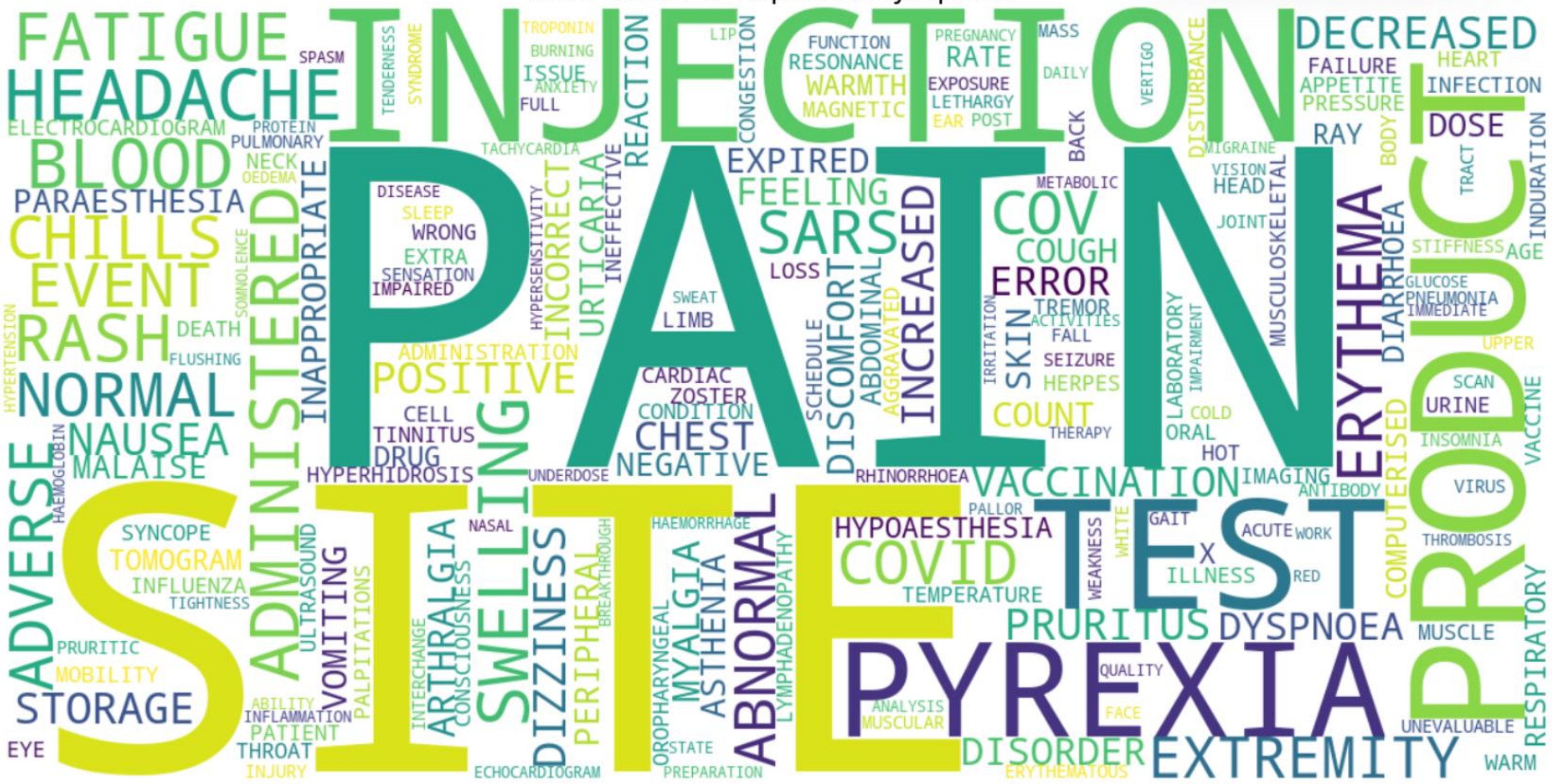
Top 10 Manufacturers



```
Top 10 Vaccine Manufacturers:
PFIZER\BIONTECH: 691984
MODERNA: 657534
MERCK & CO. INC.: 188338
GLAXOSMITHKLINE BIOLOGICALS: 172028
JANSSEN: 106390
SANOFI PASTEUR: 100993
UNKNOWN MANUFACTURER: 81906
PFIZER\WYETH: 39952
NOVARTIS VACCINES AND DIAGNOSTICS: 24000
SEQIRUS, INC.: 21426
```



Word Cloud of Reported Symptoms



DATA ENGINEERING

Data Preprocessing Steps

1. Merging the three CSV files (VAERSDATA, VAERSVAX, VAERSSYMPTOMS) for each year using glob module
2. Handling missing values and data inconsistencies.
3. Encoding categorical variables to binary parameters.
4. Preprocessed data by removing outliers and dropped unnecessary columns.
5. Defined the datatypes for columns and standardized the text.

PROJECT SCOPE

Vaccines Considered:

- COVID-19
- Varicella-Zoster (VARZOS)
- Pneumococcal vaccine polyvalent (PPV)
- Influenza (FLU) vaccines

Predicting the risk of serious AEs based on patient demographics, symptoms, and health history. Developing insights to identify which vaccines are safer for different patient profiles.

VACCINE CONSIDERED

- COVID vaccine safeguards from the respiratory illness in humans caused by a coronavirus, capable of producing severe symptoms.
- VARICELLA-ZOSTER VACCINE (VARZOS): Vaccine that reduces the incidence of herpes zoster (shingles), a disease caused by reactivation of the varicella-zoster virus (VZV), which is also responsible for chickenpox.
- FLU (Influenza) vaccine protects against FLU and from its potential serious complications.
- PPV vaccine protects against infections like pneumonia and meningitis caused by Streptococcus pneumoniae bacteria, especially in older adults and those with weakened immune systems.

OBSERVATION AND FUTURE MODELING

- Data has more Sparsity
- Create a target variable - Serious and Non-serious cases
- Perform One Hot Encoding on the features required for modeling
- Building Machine learning models to predict the risk of serious adverse events
- Evaluating the model performance
- Deploying a predictive tool

REFERENCES

1. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/safety/adverse-events.html

2. https://pubmed.ncbi.nlm.nih.gov/15071280/

3. https://www.cdc.gov

4. https://stackoverflow.com/questions/45787782/combine-multiple-columns-in-pandas-excludingnans.

5. https://stackoverflow.com/questions/17679089/pandas-dataframe-groupby-two-columns-and-get-counts

THANK YOU