Group 2 Covid-19 Data Mining & Analysis Project

Spring 2021 Semester

Created by:
Krishnaveni Balusupati
Yaow Hui Chong
Candace Meneely
Kavya Raviprakash

Table of Contents

Introduction

COVID-19 has been a huge topic in 2020 and in 2021. As of March 25, 2021, the global impact has around 126 million COVID-19 cases, while the United States has around 30 million cases. As a comparison, the total amount of COVID-19 vaccine doses distributed across the U.S. is around 173 million doses, while there are around 133 million people who received the first dose. (News Break, March 25, 2021). Currently, three main COVID-19 vaccines brands authorized by FBA for emergency uses are Pfizer-BioNTech, Moderna, and Johnson & Johnson. Based on evidence from clinical trials, the Pfizer-BioNTech vaccine was 95% effective at preventing laboratory-confirmed COVID-19 illness in people without evidence of previous infection. (CDC, Match 4, 2021). On the other hand, Moderna vaccine was 94.1% effective at preventing laboratory-confirmed COVID-19 illness in people who received two doses who had no evidence of being previously infected. (CDC, Match 4, 2021). Also, the J&J/Janssen vaccine was 66.3% effective in clinical trials at preventing laboratory-confirmed COVID-19 illness in people who had no evidence of prior infection 2 weeks after receiving the vaccine; people had the most protection 2 weeks after getting vaccinated. (CDC, Match 4, 2021). Our group is interested to study the results of COVID-19 vaccines of different brands to know which vaccine could possibly perform better to different age groups, so that we could recommend certain vaccine to each age group.

Research Questions

In our research, we are interested to analyze the adverse effect of COVID-19 vaccinations on individuals, on different age groups, and overall which vaccine works better. Our consideration for a better performance of a vaccine is with the less adverse effects.

(1) Which age group performs the best after vaccination?

We will examine the number of adverse effects on certain age groups after vaccination to study the effect of vaccines on age groups.

(2) Which vaccine brand overall performs the best?

We will examine the number of adverse effects after using a certain brand of vaccine to determine which vaccine brand is better.

(3) In each age group, which vaccine brand performs the best?

We will examine the number of adverse effects of certain vaccine brands on each age group to understand which vaccine brand does better in each age group.

These research questions above are how our data mining group project started out. They seemed relatively straight forward and as a group we felt undaunted by the task. As weeks went by and we all dug deep into our data, we had to go back to Dr. Shi for guidance. When we asked our questions and explained our challenges to Professor, he chuckled and verbally expressed what we'd been struggling with: our questions although they seemed simple, were too complex. We needed to simplify our questions. We took Dr. Shi's advice. With our original questions in mind, we combined them and came up with a new, simplified, exploratory based question:

What insights can be found regarding deaths and the Covid-19 vaccines manufactured by Pfizer and Moderna?
More specifically:

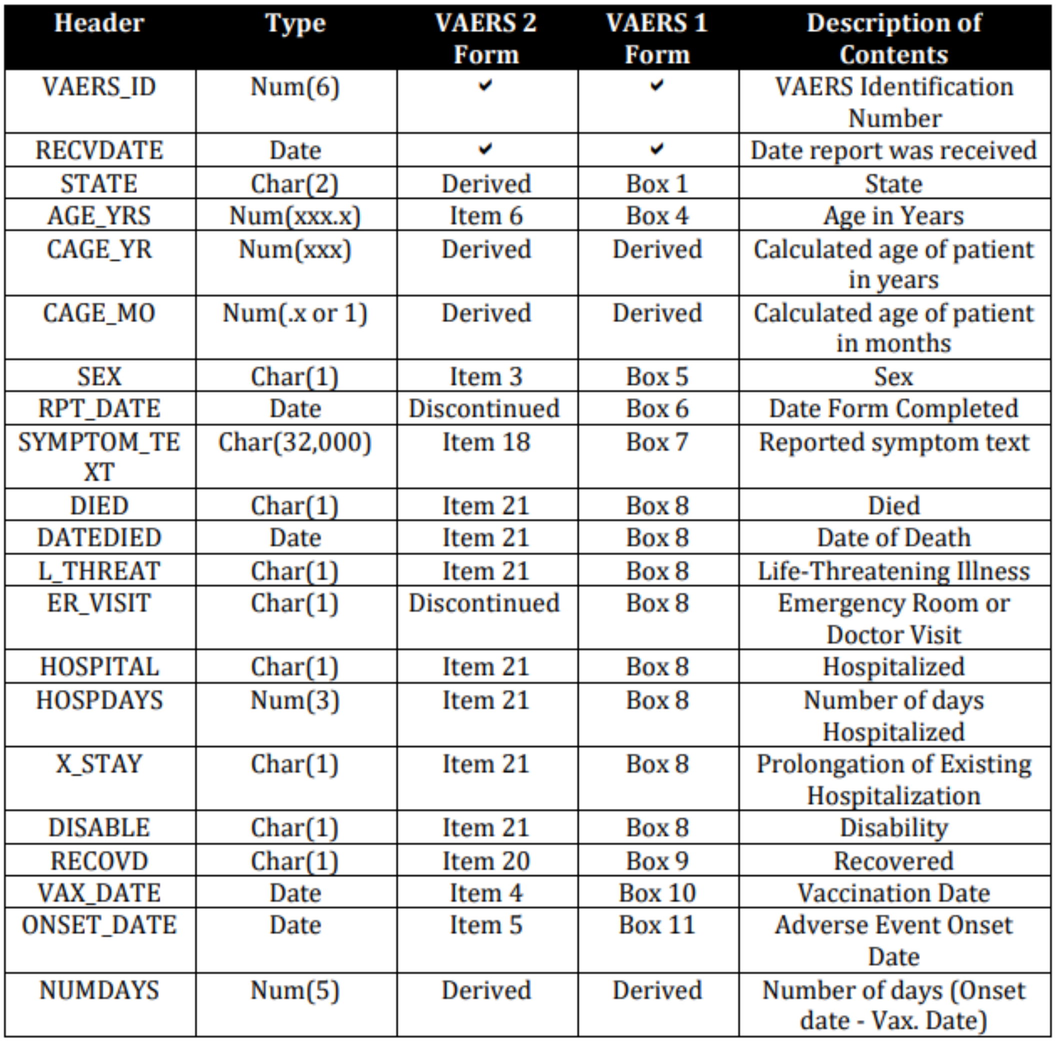
(1) What are the factors that significantly effecting the death rate in COVID vaccines?
(2) How do we predict the adverse effect risk for future vaccine candidates and provide different treatment methods using different risk groups 'Low', 'Medium', and 'High'?
(3) What is the correlation between allergies and deaths, in regard to different vaccines?

Datasets

Covid Data

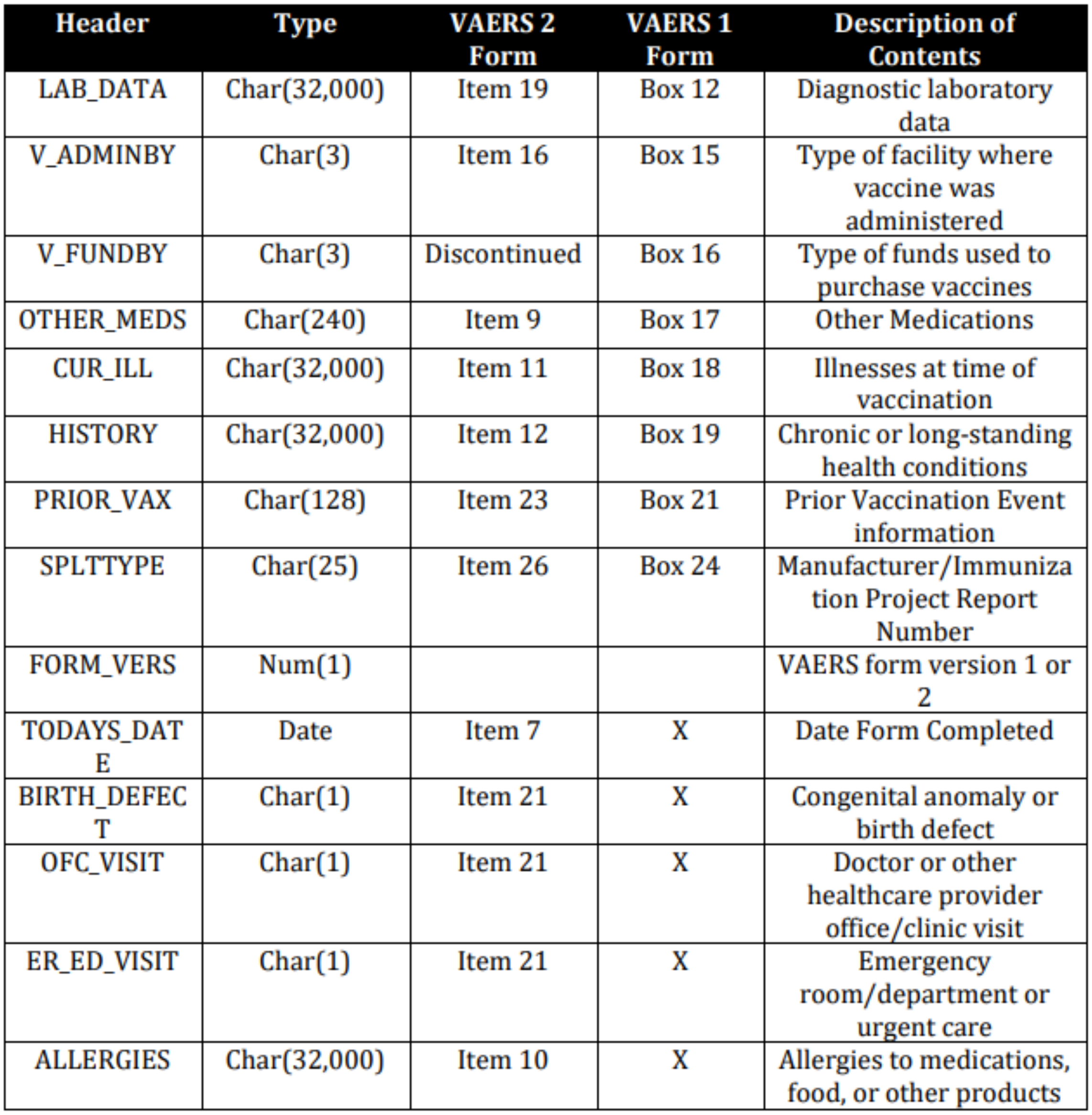
We retrieved the datasets from Kaggle.com. The datasets were reported by the Vaccine Adverse Event Reporting System (VAERS), which was created by the Food and Drug Administration (FDA) and Centers for Disease Control and Prevention (CSC) to receive reports about adverse events that may be associated with vaccines. (VAERS, November 2020). The vaccine adverse conditions were reported in three different datasets, which are:

(1) **2021VAERSDATA.csv**: This dataset includes the demographic background of individuals (e.g. age, sex and state), and some general information of individuals after the vaccination event (e.g. died, other medications, illnesses at time of vaccination). Some of the variables in the dataset and its description are listed in the table below.



Header	Type	VAERS 2 Form	VAERS 1 Form	Description of Contents
VAERS_ID	Num(6)	✓	✓	VAERS Identification Number
RECVDATE	Date	✓	✓	Date report was received
STATE	Char(2)	Derived	Box 1	State
AGE_YRS	Num(xxx.x)	Item 6	Box 4	Age in Years
CAGE_YR	Num(xxx)	Derived	Derived	Calculated age of patient in years
CAGE_MO	Num(.x or 1)	Derived	Derived	Calculated age of patient in months
SEX	Char(1)	Item 3	Box 5	Sex
RPT_DATE	Date	Discontinued	Box 6	Date Form Completed
SYMPTOM_TEXT	Char(32,000)	Item 18	Box 7	Reported symptom text
DIED	Char(1)	Item 21	Box 8	Died
DATEDIED	Date	Item 21	Box 8	Date of Death
L_THREAT	Char(1)	Item 21	Box 8	Life-Threatening Illness
ER_VISIT	Char(1)	Discontinued	Box 8	Emergency Room or Doctor Visit
HOSPITAL	Char(1)	Item 21	Box 8	Hospitalized
HOSPDAYS	Num(3)	Item 21	Box 8	Number of days Hospitalized
X_STAY	Char(1)	Item 21	Box 8	Prolongation of Existing Hospitalization
DISABLE	Char(1)	Item 21	Box 8	Disability
RECOVD	Char(1)	Item 20	Box 9	Recovered
VAX_DATE	Date	Item 4	Box 10	Vaccination Date
ONSET_DATE	Date	Item 5	Box 11	Adverse Event Onset Date
NUMDAYS	Num(5)	Derived	Derived	Number of days (Onset date - Vax. Date)

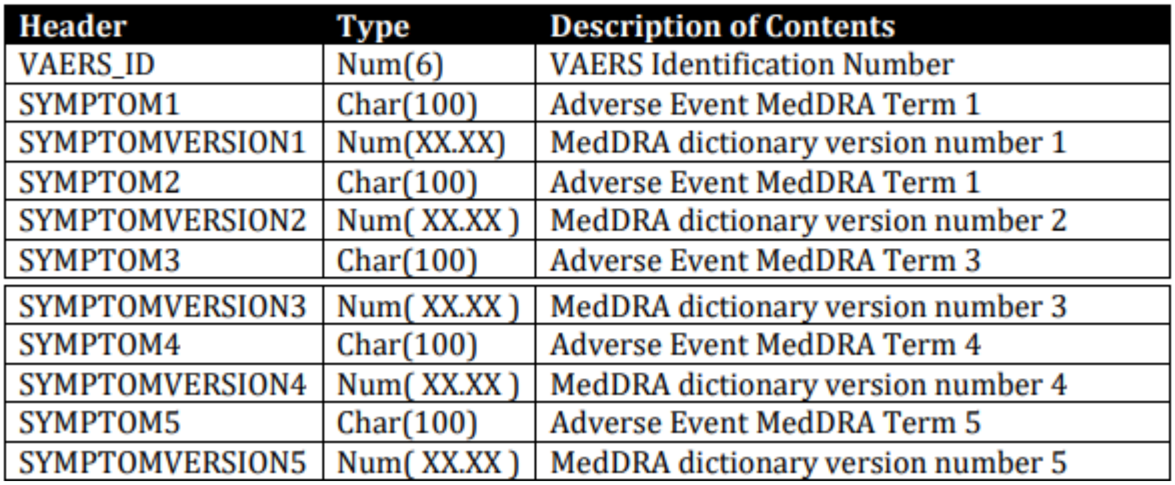
Table 1. Data description (VAERS, November 2020).



Header	Type	VAERS 2 Form	VAERS 1 Form	Description of Contents
LAB_DATA	Char(32,000)	Item 19	Box 12	Diagnostic laboratory data
V_ADMINBY	Char(3)	Item 16	Box 15	Type of facility where vaccine was administered
V_FUNDBY	Char(3)	Discontinued	Box 16	Type of funds used to purchase vaccines
OTHER_MEDS	Char(240)	Item 9	Box 17	Other Medications
CUR_ILL	Char(32,000)	Item 11	Box 18	Illnesses at time of vaccination
HISTORY	Char(32,000)	Item 12	Box 19	Chronic or long-standing health conditions
PRIOR_VAX	Char(128)	Item 23	Box 21	Prior Vaccination Event information
SPLTTYPE	Char(25)	Item 26	Box 24	Manufacturer/Immunization Project Report Number
FORM_VERS	Num(1)			VAERS form version 1 or 2
TODAYS_DATE	Date	Item 7	X	Date Form Completed
BIRTH_DEFECT	Char(1)	Item 21	X	Congenital anomaly or birth defect
OFC_VISIT	Char(1)	Item 21	X	Doctor or other healthcare provider office/clinic visit
ER_ED_VISIT	Char(1)	Item 21	X	Emergency room/department or urgent care
ALLERGIES	Char(32,000)	Item 10	X	Allergies to medications, food, or other products

Table 2. Data description (VAERS, November 2020).

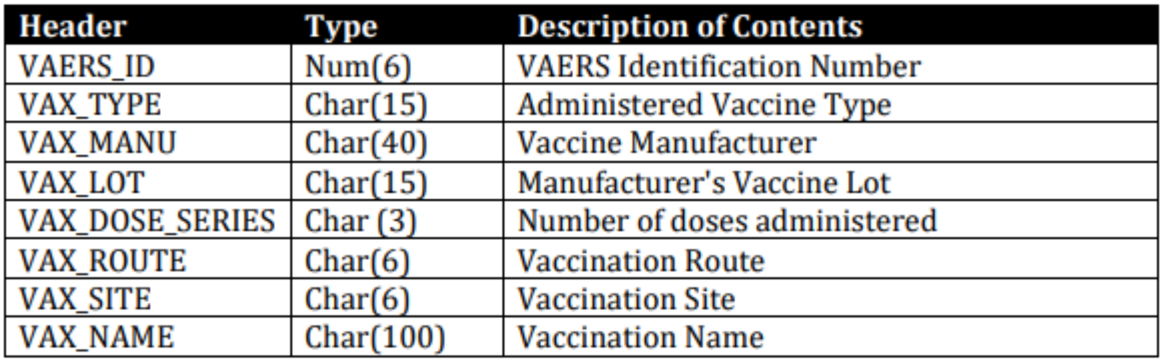
(2) **2021VAERSSYMTOMS.csv**: This dataset includes the reported adverse event of individuals after vaccinated.



Header	Type	Description of Contents
VAERS_ID	Num(6)	VAERS Identification Number
SYMPTOM1	Char(100)	Adverse Event MedDRA Term 1
SYMPTOMVERSION1	Num(XX.XX)	MedDRA dictionary version number 1
SYMPTOM2	Char(100)	Adverse Event MedDRA Term 1
SYMPTOMVERSION2	Num(XX.XX)	MedDRA dictionary version number 2
SYMPTOM3	Char(100)	Adverse Event MedDRA Term 3
SYMPTOMVERSION3	Num(XX.XX)	MedDRA dictionary version number 3
SYMPTOM4	Char(100)	Adverse Event MedDRA Term 4
SYMPTOMVERSION4	Num(XX.XX)	MedDRA dictionary version number 4
SYMPTOM5	Char(100)	Adverse Event MedDRA Term 5
SYMPTOMVERSION5	Num(XX.XX)	MedDRA dictionary version number 5

Table 3. Data description (VAERS, November 2020).

(3) **2021VAERSVAX.csv**: This dataset includes the vaccine information of individuals such as the vaccine name, manufacturer, lot number, route, site, and number of previous doses administered.



Header	Type	Description of Contents
VAERS_ID	Num(6)	VAERS Identification Number
VAX_TYPE	Char(15)	Administered Vaccine Type
VAX_MANU	Char(40)	Vaccine Manufacturer
VAX_LOT	Char(15)	Manufacturer's Vaccine Lot
VAX_DOSE_SERIES	Char (3)	Number of doses administered
VAX_ROUTE	Char(6)	Vaccination Route
VAX_SITE	Char(6)	Vaccination Site
VAX_NAME	Char(100)	Vaccination Name

Table 4. Data description (VAERS, November 2020).

Demographic Data

We have explored some demographics information about the individuals who are involved in the adverse effect datasets. These graphs of gender, state, age group, and the vaccine types are shown in graphs.

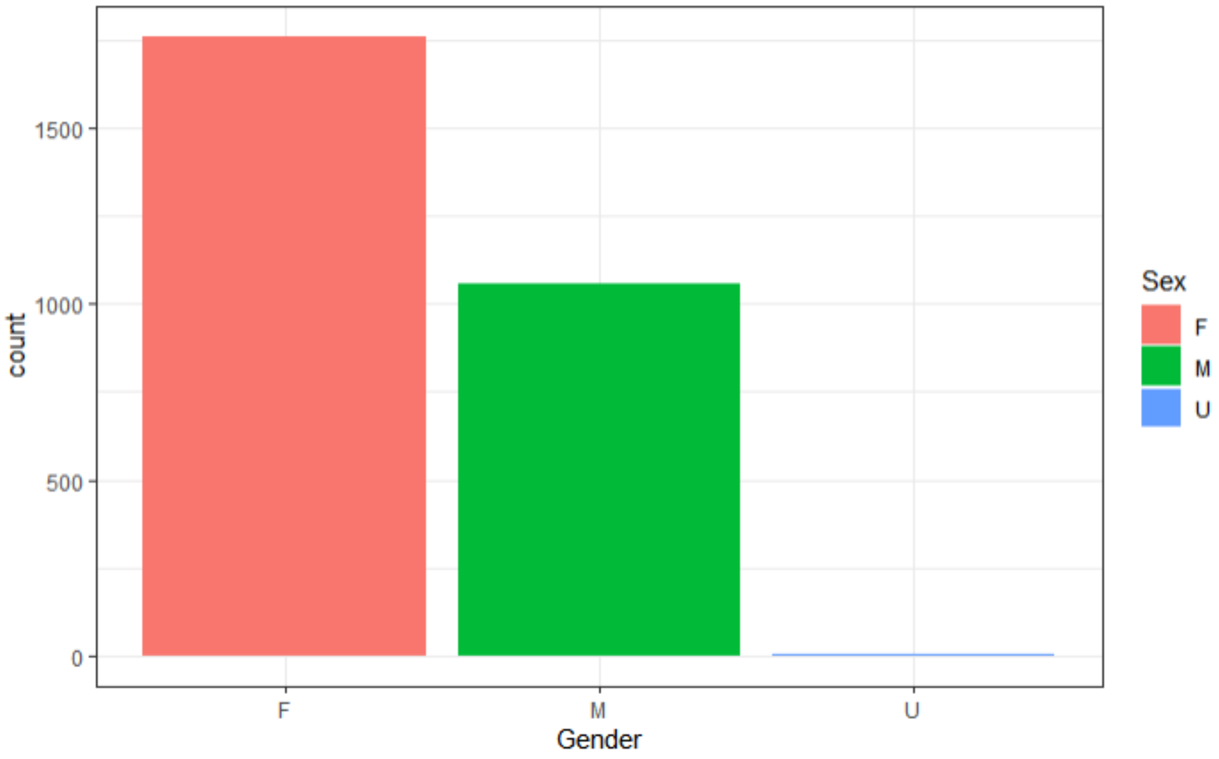


Figure 1. Gender graph

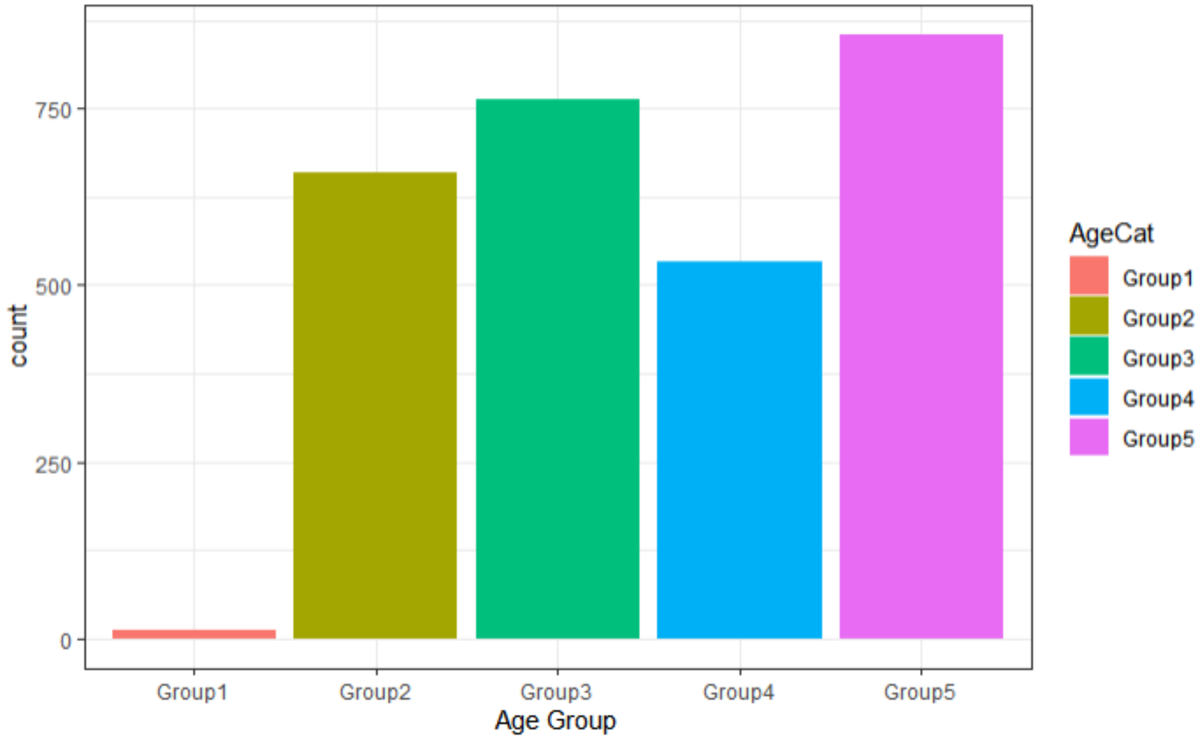


Figure 2. Age group plot

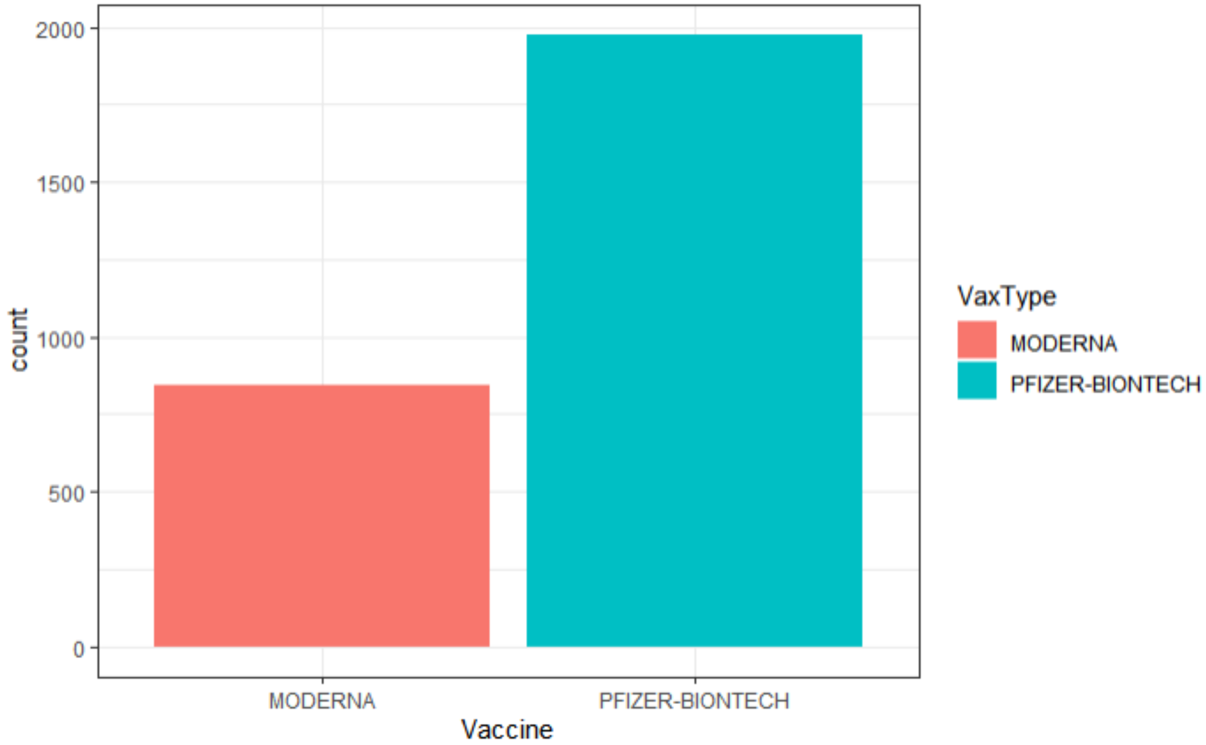


Figure 3. Vaccine type plot indicates the vaccine type received by individuals.

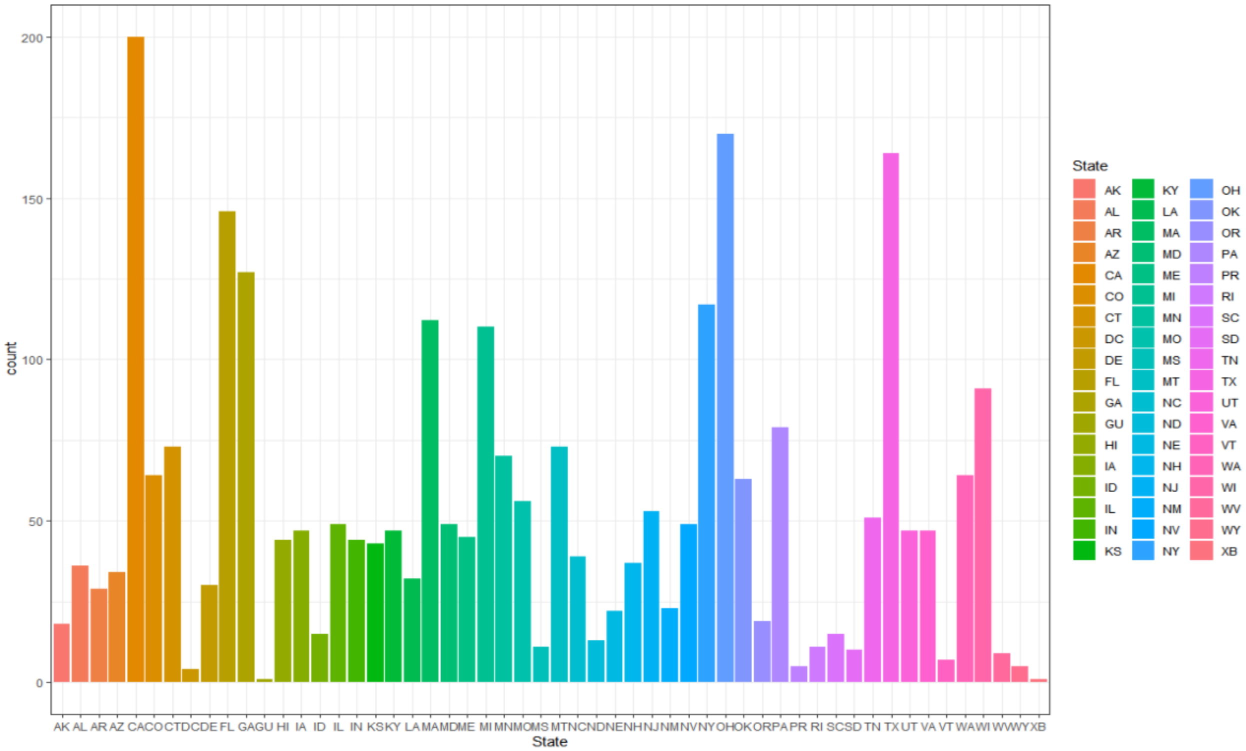


Figure 4. State plot indicates the home state of the vaccinee.

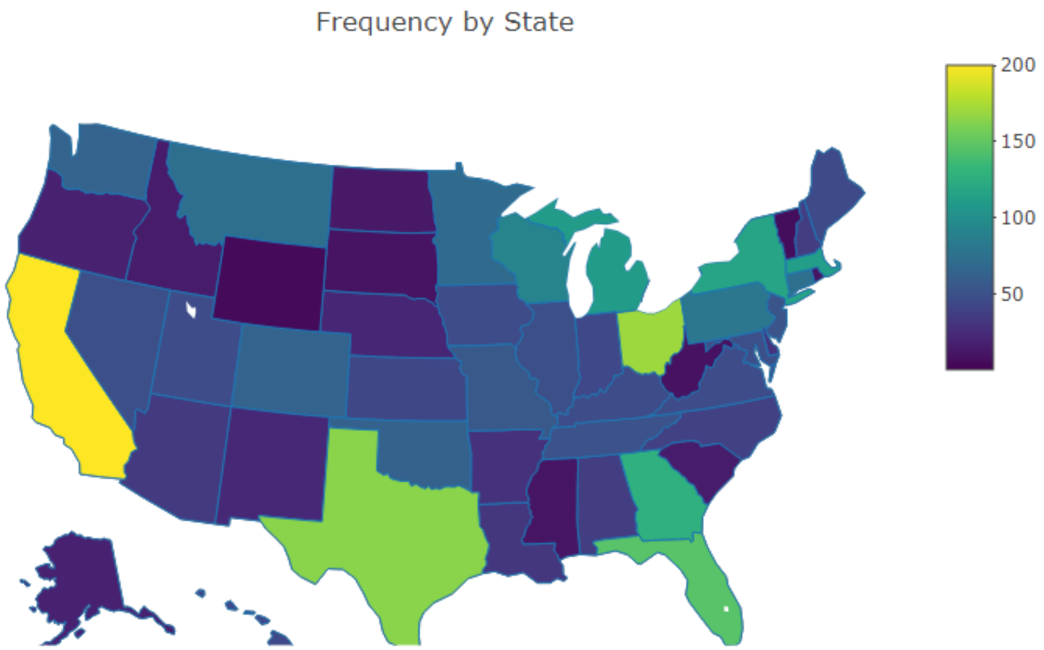


Figure 5. US map plot at state level indicates the top states relate to this study are California, Florida, New York, and Texas.

Data Mining Techniques Used

(1) Decision Tree

Decision tree is a method in supervised machine learning. It is a tree structure of IF-THEN rules. In decision tree, the algorithm automatically determines which variables are most important, based on their ability to sort the data into the correct output category. (Shi & Olson, 2005). Some important elements in decision tree are root, nodes, and branches. Root is the first node where the split in a decision tree begins. Nodes are where the splits in decision tree happen, and the different values of the attribute will split into different branches. The advantages of using decision tree is that it automatically processes data, and it searches through data for patterns and relationships. Thus, it is a pure knowledge discovery algorithm which assumes no prior hypothesis and it disregards human judgment. Decision tree are categorized into classification tree and regression tree. Classification tree makes yes or no decisions, while regression tree predicts values for the target variable.

(2) K-mean Clustering

Clustering analysis is an unsupervised technique, where data is examined without reference to a response variable. (Shi & Olson, 2005). The most general form of clustering analysis allows the algorithm to determine the number of clusters, or the number of clusters may be pre-specified. (Shi & Olson, 2005). In that case, users will define a target number k, which refers to the number of centroids needed in the dataset. A centroid is the imaginary or real location representing the center of the cluster, and every data point is allocated to each of the clusters through reducing the in-cluster sum of squares. (Garbade, September 12, 2018).

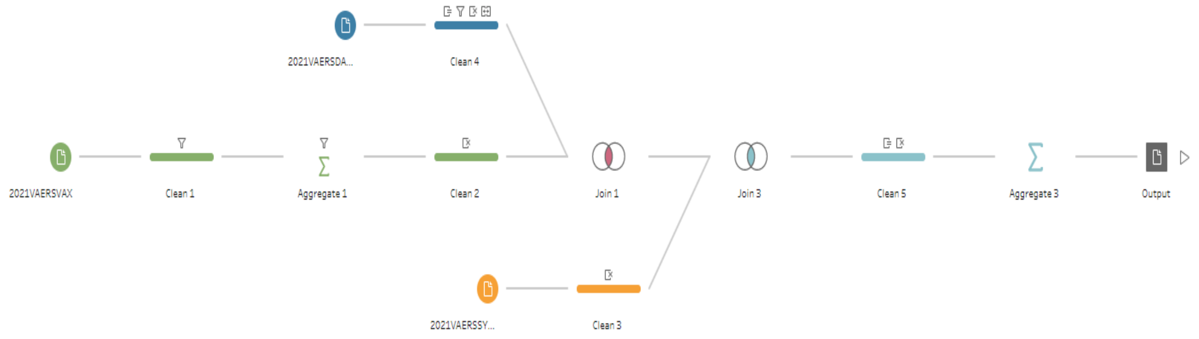
(3) Logistic Regression

In data mining, Regression is one of the basic tools for analysis, used in classification applications through logistic regression and discriminant analysis, as well as the prediction of continuous data through ordinary least squares (OLS) and other forms (Shi & Olson, 2005). Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. An independent variable is an input, assumption, or driver that is changed to assess its impact on a dependent variable. Logistic Regression is used when the dependent variable(target) is categorical.

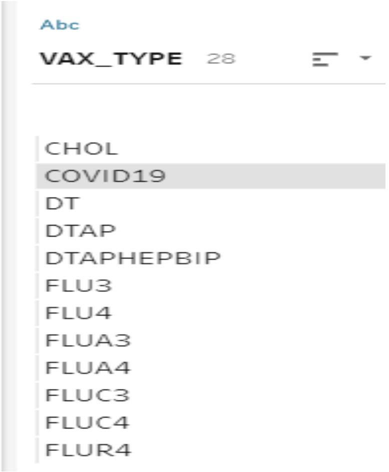
Logistic regression to identify influencing factors for deaths.

Data cleansing

Data cleansing is the process of detecting and correcting corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. To clean the data set 2021VAERSVAX, I used a cleansing tool named Tableau Prep Builder. Tableau prep builder helps in combining, shaping, and cleaning the data for the analysis. Below shows pictorial view of the cleansing process taken from Tableau prep builder.



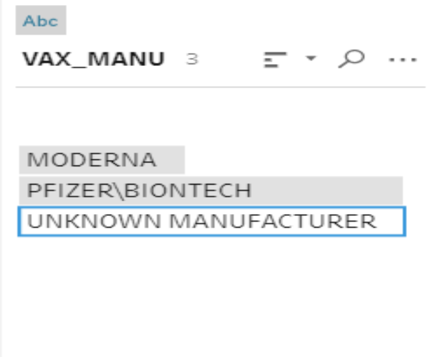
Coming to Initial cleansing process, the dataset 2021VAERSVAX got different vaccines in which we are interested in COVID19. In Covid19, filtered only Pfizer and Moderna. Removed the duplicate rows.



Abc

VAX_TYPE 28

CHOL
COVID19
DT
DTAP
DTAPHEPBIP
FLU3
FLU4
FLUA3
FLUA4
FLUC3
FLUC4
FLUR4

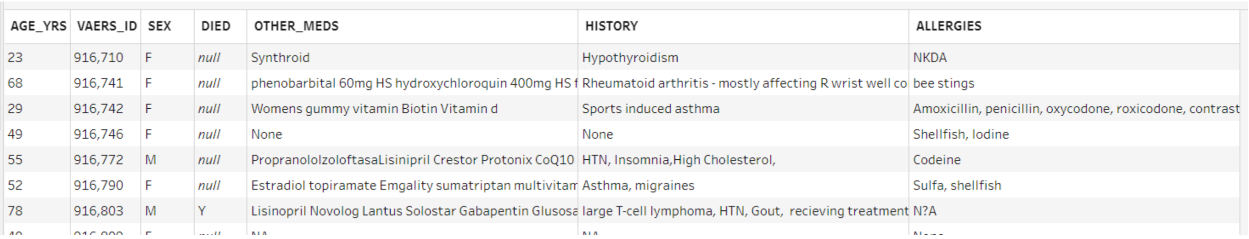


Abc

VAX_MANU 3

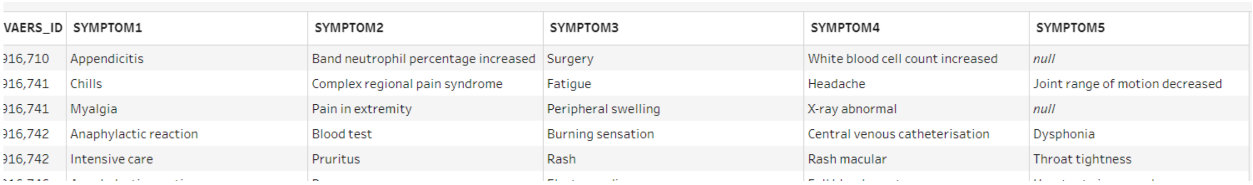
MODERNA
PFIZER\BIONTECH
UNKNOWN MANUFACTURER

For the dataset 2021VAERSDATA: Excluded inappropriate age values. Merged AGE_YRS and CAGE_YR columns. Removed all the other columns and kept following 7 columns for processing.



AGE_YRS	VAERS_ID	SEX	DIED	OTHER_MEDS	HISTORY	ALLERGIES
23	916,710	F	*null*	Synthroid	Hypothyroidism	NKDA
68	916,741	F	*null*	phenobarbital 60mg HS hydroxychloroquin 400mg HS f	Rheumatoid arthritis - mostly affecting R wrist well co	bee stings
29	916,742	F	*null*	Womens gummy vitamin Biotin Vitamin d	Sports induced asthma	Amoxicillin, penicillin, oxycodone, roxicodone, contrast
49	916,746	F	*null*	None	None	Shellfish, Iodine
55	916,772	M	*null*	PropranololzoloftasaLisinipril Crestor Protonix CoQ10	HTN, Insomnia,High Cholesterol,	Codeine
52	916,790	F	*null*	Estradiol topiramate Emgality sumatriptan multivitam	Asthma, migraines	Sulfa, shellfish
78	916,803	M	Y	Lisinopril Novolog Lantus Solostar Gabapentin Glusosa	large T-cell lymphoma, HTN, Gout, recieving treatment	N?A

For the dataset 2021VAERSSYMPTOMS: Removed all the columns with versions. Considered the following 6 columns.

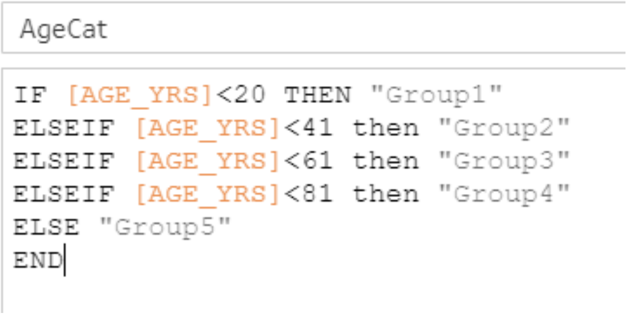


VAERS_ID	SYMPTOM1	SYMPTOM2	SYMPTOM3	SYMPTOM4	SYMPTOM5
916,710	Appendicitis	Band neutrophil percentage increased	Surgery	White blood cell count increased	*null*
916,741	Chills	Complex regional pain syndrome	Fatigue	Headache	Joint range of motion decreased
916,741	Myalgia	Pain in extremity	Peripheral swelling	X-ray abnormal	*null*
916,742	Anaphylactic reaction	Blood test	Burning sensation	Central venous catheterisation	Dysphonia
916,742	Intensive care	Pruritus	Rash	Rash macular	Throat tightness

After the initial data cleaning in each dataset, joined 2021VAERSVAX, 2021VAERSDATA, and 2021VAERSSYMPTOMS datasets.

Data transformations

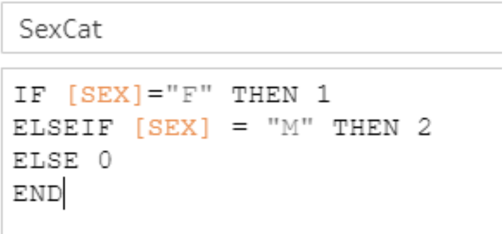
Data transformation is the process of changing the format, structure, or values of data to better organize, improve quality and to facilitate the compatibility of data.
Created categorical data from the existing information for all the columns.
We divided the age into five groups as below.



AgeCat

```
IF [AGE_YRS]<20 THEN "Group1"
ELSEIF [AGE_YRS]<41 then "Group2"
ELSEIF [AGE_YRS]<61 then "Group3"
ELSEIF [AGE_YRS]<81 then "Group4"
ELSE "Group5"
END
```

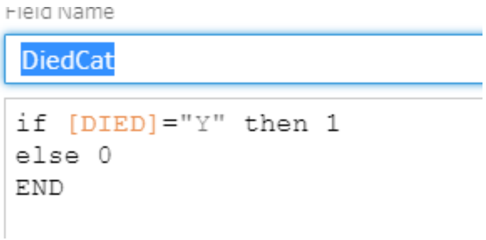
The sex column is divided into three categories as below



SexCat

```
IF [SEX]="F" THEN 1
ELSEIF [SEX] = "M" THEN 2
ELSE 0
END
```

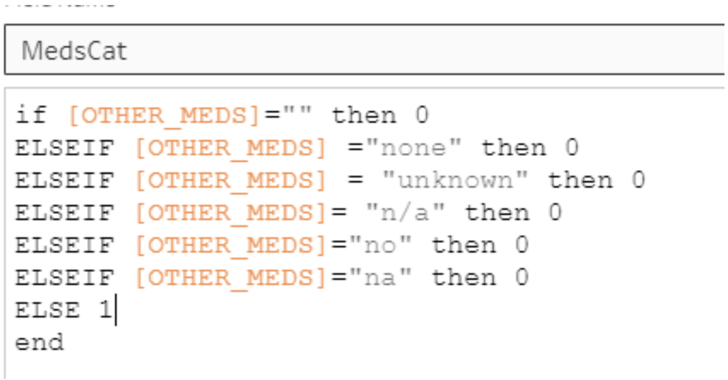
The DIED column is converted into binary column as died 1, otherwise a 0.



DiedCat

```
if [DIED]="Y" then 1
else 0
END
```

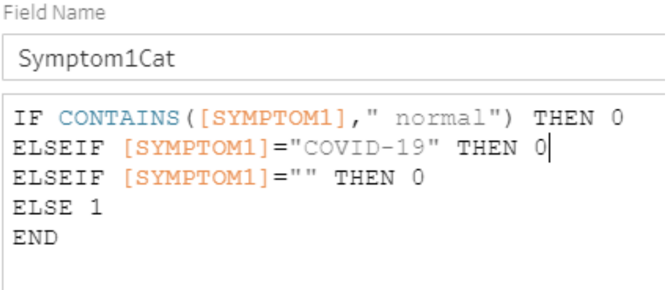
The current medications a patient taking OTHER_MEDS column is categorized into two with a medication taken as 1, otherwise 0. Same applied for ALLERGIES and HISTORY columns.



MedsCat

```
if [OTHER_MEDS]="" then 0
ELSEIF [OTHER_MEDS] ="none" then 0
ELSEIF [OTHER_MEDS] = "unknown" then 0
ELSEIF [OTHER_MEDS]= "n/a" then 0
ELSEIF [OTHER_MEDS]="no" then 0
ELSEIF [OTHER_MEDS]="na" then 0
ELSE 1
end
```

A SYMPTOM1 column is categorized as 1 for any symptoms shown otherwise a 0.



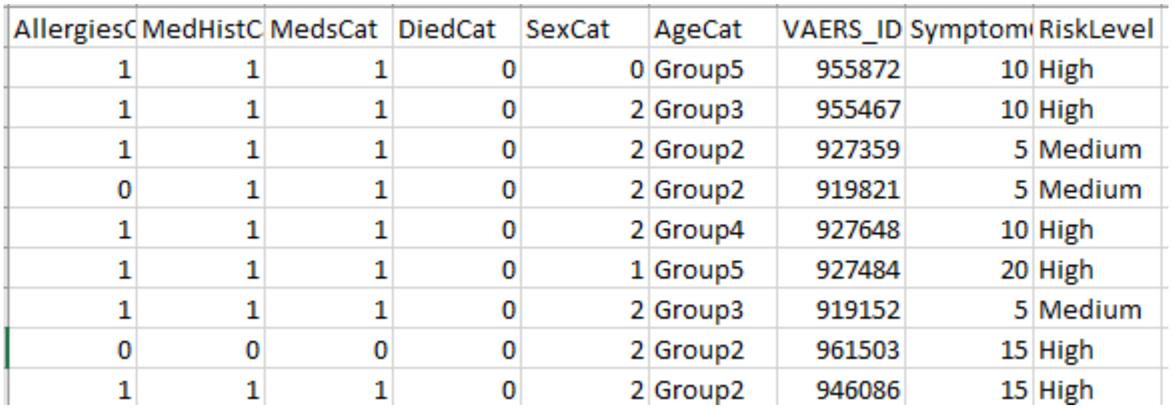
Field Name

Symptom1Cat

```
IF CONTAINS([SYMPTOM1]," normal") THEN 0
ELSEIF [SYMPTOM1]="COVID-19" THEN 0
ELSEIF [SYMPTOM1]="" THEN 0
ELSE 1
END
```

Followed same transformation method for all the other symptoms, Symptom2 to Symptom5. In order to calculate the risk levels, we consolidated the number of symptoms. The data transformation for RiskLevel is explained in decision tree section.

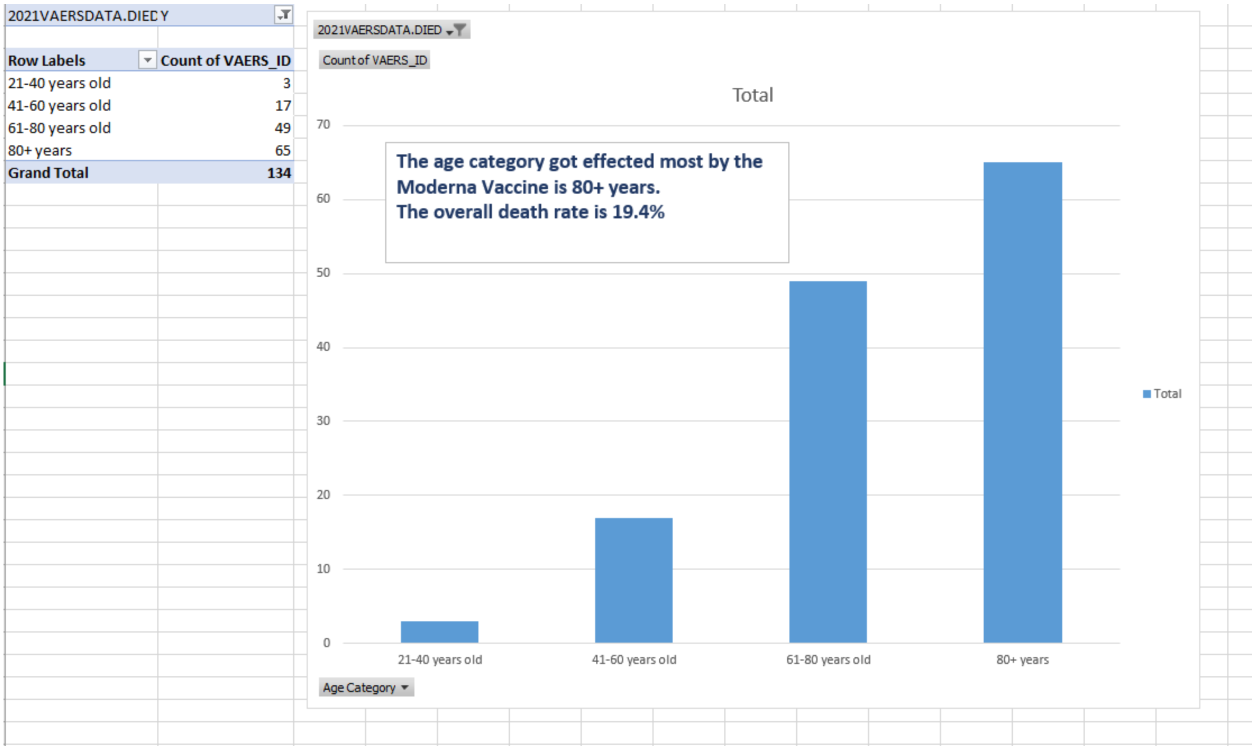
Finally, we have the categorical data which is cleansed and transformed as shown below.



AllergiesC	MedHistC	MedsCat	DiedCat	SexCat	AgeCat	VAERS_ID	Symptom	RiskLevel
1	1	1	0	0	Group5	955872	10	High
1	1	1	0	2	Group3	955467	10	High
1	1	1	0	2	Group2	927359	5	Medium
0	1	1	0	2	Group2	919821	5	Medium
1	1	1	0	2	Group4	927648	10	High
1	1	1	0	1	Group5	927484	20	High
1	1	1	0	2	Group3	919152	5	Medium
0	0	0	0	2	Group2	961503	15	High
1	1	1	0	2	Group2	946086	15	High

Death rate Insights

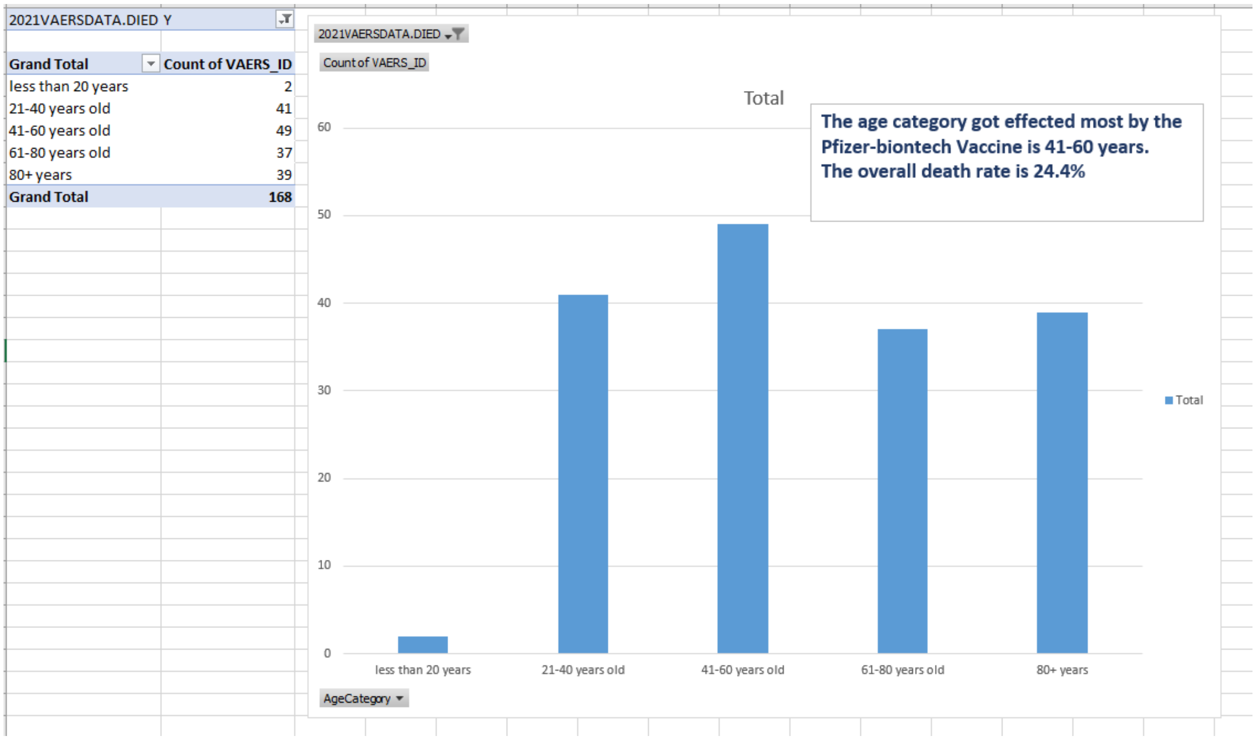
The known fact is that COVID virus is affecting the elderly the most. we got interested in finding how the death rate is spread across age categories and considering death count does Moderna or Pfizer works better. Here we got some results from Modern and Pizer death count.



2021VAERSDATA.DIEC Y

Row Labels	Count of VAERS_ID
21-40 years old	3
41-60 years old	17
61-80 years old	49
80+ years	65
Grand Total	134

The age category got effected most by the Moderna Vaccine is 80+ years.
The overall death rate is 19.4%



2021VAERSDATA.DIED Y

Grand Total	Count of VAERS_ID
less than 20 years	2
21-40 years old	41
41-60 years old	49
61-80 years old	37
80+ years	39
Grand Total	168

The age category got effected most by the Pfizer-biontech Vaccine is 41-60 years.
The overall death rate is 24.4%

Vaccine comparison based on death count:

	Moderna	Pfizer
Sample size	688	688

Successful cases	554	520
Number of deaths	134	168
Success rate	80.5%	75.5%

As the table above shows the Success rate with no deaths is more for Moderna vaccine.

Logistic regression

As we have already the Age is a significant factor for COVID vaccine death rate, Using the logistic regression, we are planning to find out the influencing factors for deaths of the people's, who took COVID19 vaccine. We had seen the Age was an influencing factor of COVID deaths. We are now interested in finding does allergies, past medical history, current medication, sex or number of total symptoms after vaccination has any significance on death.

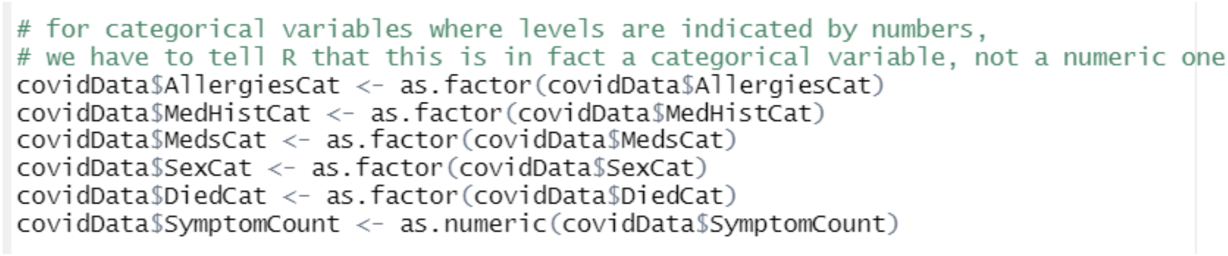
We are using R to run the logistic regression. We used the Generalized Linear Model function glm() to run the logistic regression. The dependent variable is DiedCat. The independent variables are AllergiesCat, MedHistCat, MedsCat, SexCat, SymptomCount.

Steps:

Split the dataset into training set and test set. We are analyzing the regression model with ¾ of data in training set and ¼ of data in test set.

```
set.seed(3456)
trainIndex <- createDataPartition(covidData$DiedCat, p = .75,
                                  list = FALSE,
                                  times = 1)
train <- covidData[ trainIndex,]
test <- covidData[-trainIndex,]
```

We have to let R know all except SymptomCount are categorical variables.

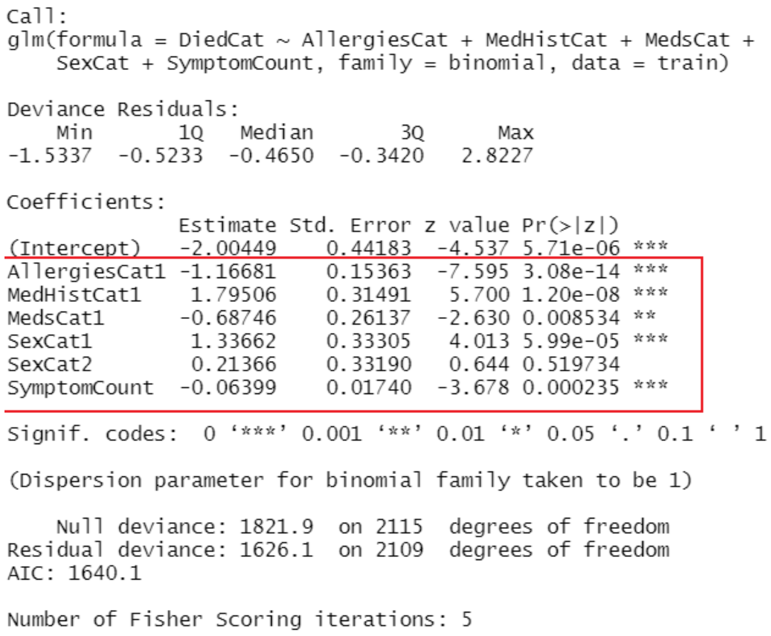


```
# for categorical variables where levels are indicated by numbers,
# we have to tell R that this is in fact a categorical variable, not a numeric one
covidData$AllergiesCat <- as.factor(covidData$AllergiesCat)
covidData$MedHistCat <- as.factor(covidData$MedHistCat)
covidData$MedsCat <- as.factor(covidData$MedsCat)
covidData$SexCat <- as.factor(covidData$SexCat)
covidData$DiedCat <- as.factor(covidData$DiedCat)
covidData$SymptomCount <- as.numeric(covidData$SymptomCount)
```

Now we run the logistic regression using glm function. we used the formula notation to specify the variables to the glm function, naming each variable in our formula as below.

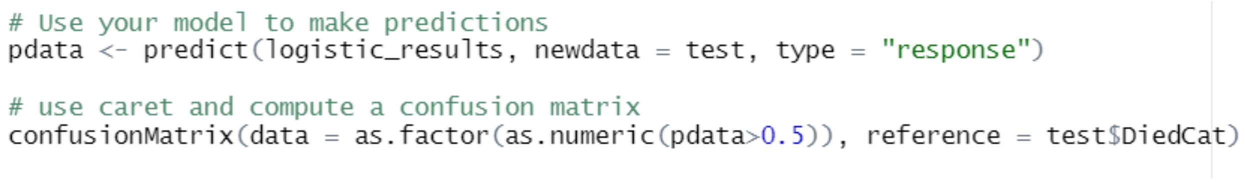
```
#we need to use the generic glm function.
#Specify family = "binomial" as it tells R to calculate the logistic regression model!
logistic_results <- glm(DiedCat ~ AllergiesCat + MedHistCat + MedsCat + SexCat +
                         SymptomCount, data = train, family=binomial)

summary(logistic_results)
```

The output shows that all the independent variables show significant effect on death rate in COVID patients.



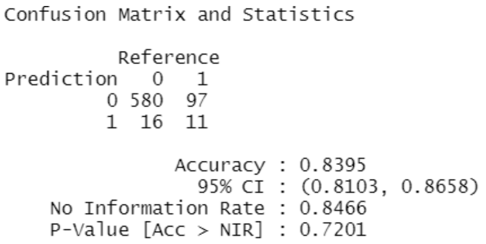
```
Call:
glm(formula = DiedCat ~ AllergiesCat + MedHistCat + MedsCat +
    SexCat + SymptomCount, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5337  -0.5233  -0.4650  -0.3420   2.8227

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.00449    0.44183  -4.537 5.71e-06 ***
AllergiesCat1 -1.16681    0.15363  -7.595 3.08e-14 ***
MedHistCat1    1.79506    0.31491   5.700 1.20e-08 ***
MedsCat1      -0.68746    0.26137  -2.630 0.008534 **
SexCat1        1.33662    0.33305   4.013 5.99e-05 ***
SexCat2        0.21366    0.33190   0.644 0.519734
SymptomCount  -0.06399    0.01740  -3.678 0.000235 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1821.9  on 2115  degrees of freedom
Residual deviance: 1626.1  on 2109  degrees of freedom
AIC: 1640.1

Number of Fisher Scoring iterations: 5
```

Running confusion matrix on the results of logistic regression shows following results.



```
# Use your model to make predictions
pdata <- predict(logistic_results, newdata = test, type = "response")

# use caret and compute a confusion matrix
confusionMatrix(data = as.factor(as.numeric(pdata>0.5)), reference = test$DiedCat)
```

The confusion matrix results shows



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 580  97
         1  16  11

               Accuracy : 0.8395
                 95% CI : (0.8103, 0.8658)
    No Information Rate : 0.8466
    P-Value [Acc > NIR] : 0.7201
```

The results shows that the regression model we built has 84% accuracy in predicting the actual death rate.
Using regression model, we found out that the patient’s medical history, current medications taking, the allergies, sex and number of symptoms count has a significant effect on death rate.

Decision Tree Solution Explanation

Scenario 1: Use a classification tree to predict the risk group of patients belong to.

Data Transformation

The purpose of this analysis is to predict which risk group people are more likely to fall into using the input data when they fill in the survey. In this analysis, we use attributes such as Allergies, Other_Meds, History, Sex, Age_Group, Vax_Type as input variables. The output variable is RiskLevel, which we categorize the people who are taking vaccine into three risk groups 'Low', 'Medium', and 'High'. (refer to the rules below for more details).

Age -> Age_Group

We converted the numeric data to categorical data so that we can apply the variable more effectively. The detail of age group variable is listed in table below.

Age	Age Group
0-20	Group 1
21-40	Group 2
41-60	Group 3
61-80	Group 4
>80	Group 5

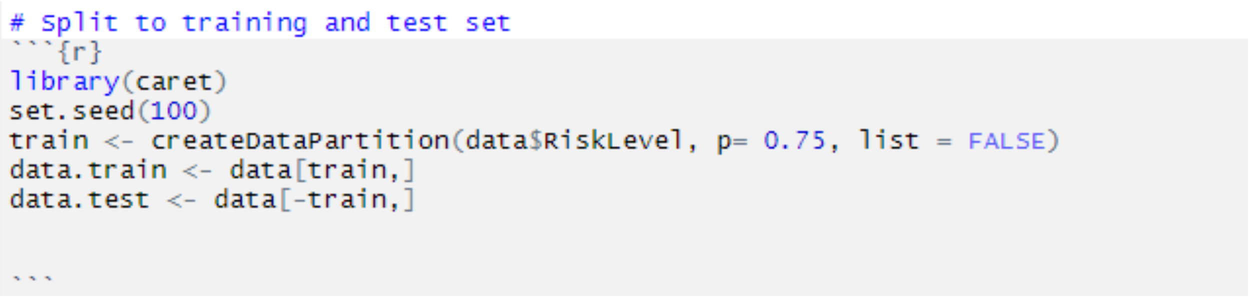
Risk_Level

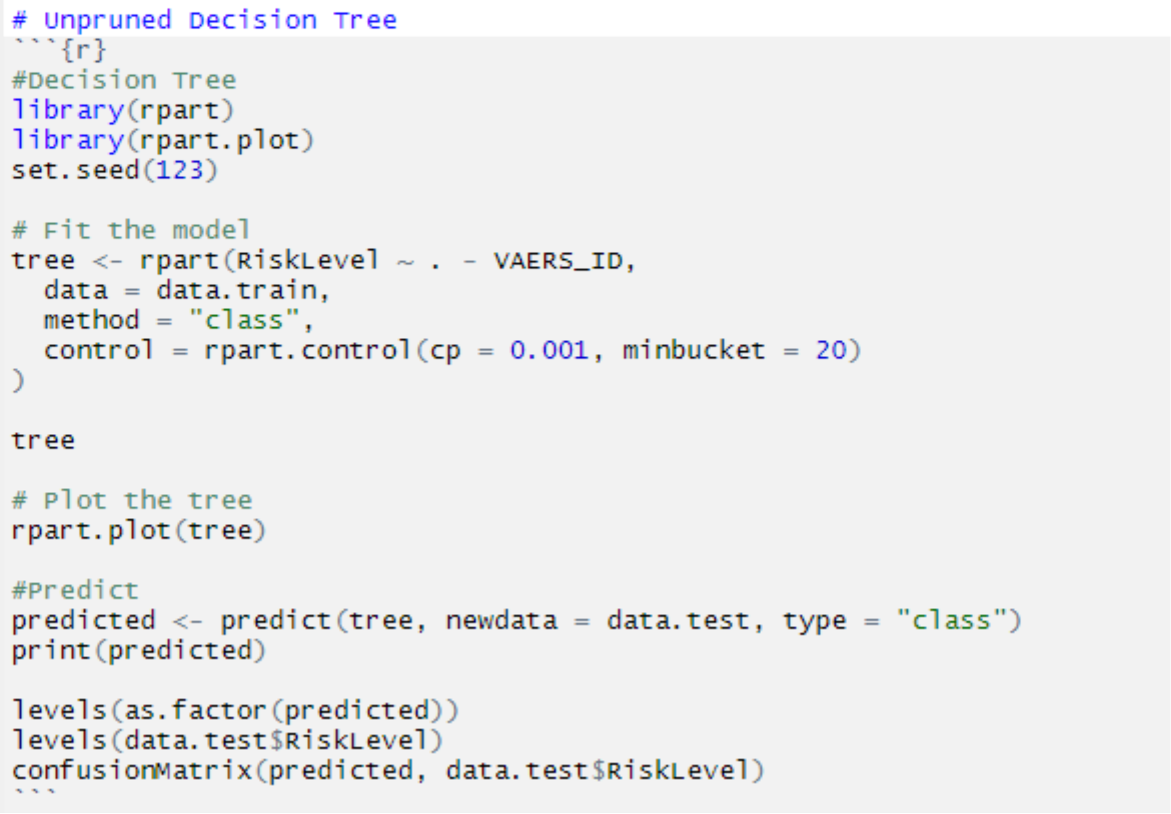
We created a new variable 'Risk level' based on the death condition and the total number of symptoms of a particular patient. The new variable 'Risk level' contains three levels: 'Low', 'Medium', and 'High'. It describes the level of risk of each patient is facing when he/she takes the vaccine. The reason being of the transformation is for the users to have a better understanding of which risk level a patient is belongs to. Also, in future, we could predict which risk level a patient with similar background will most likely fall into using machine learning algorithm. The detail of risk level variable is listed in table below:

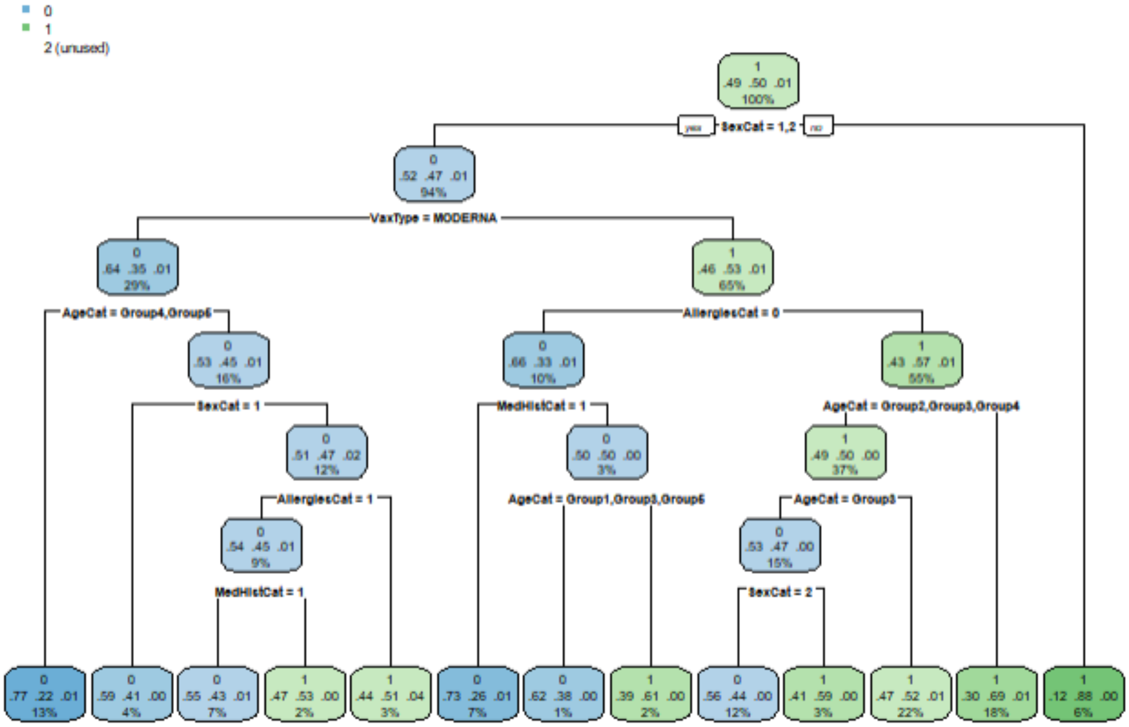
Risk Level	Algorithm	Description
High	Died = 'Yes' or Number of symptoms > 6	User in this group might have life threatening symptoms. He/she is expected to experience more than 6 symptoms after taking vaccine. Thus, they should be closely monitored.
Medium	Died = 'No' and Number of symptoms is in range [4,6]	User in this group is not likely to have life threatening symptoms. He/she is expected to experience more than 4 to 6 symptoms after taking vaccine. Thus, closely monitoring is optional.
Low	Died = 'No' and Number of symptoms is in range [0,3]	User in this group is not likely to have life threatening symptoms. He/she is expected to experience more than 0 to 3 symptoms after

		taking vaccine. Thus, closely monitoring is not needed.

1. We started the analysis by splitting the dataset into training set and test set. In this case, we use 75% of data in training set and 25% of data in test set.



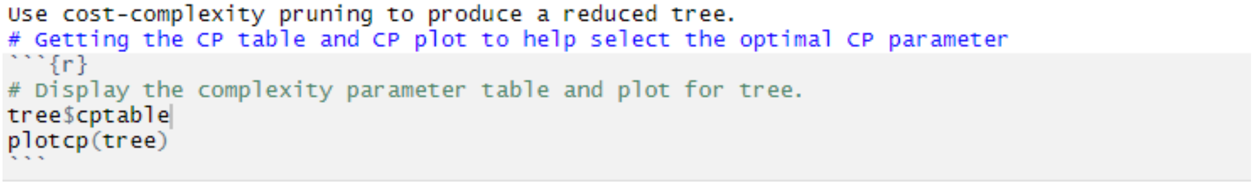
~~~
# Split to training and test set
```{r}
library(caret)
set.seed(100)
train <- createDataPartition(data$RiskLevel, p= 0.75, list = FALSE)
data.train <- data[train,]
data.test <- data[-train,]


~~~
```

2. After that, we build a classification tree using Allergies, Other_Meds, History, Sex, Age_Group, Vax_Type as input variables, and Risk Level as output. The unpruned decision tree code and output are shown below:



```
# Unpruned Decision Tree
```{r}
#Decision Tree
library(rpart)
library(rpart.plot)
set.seed(123)

# Fit the model
tree <- rpart(RiskLevel ~ . - VAERS_ID,
  data = data.train,
  method = "class",
  control = rpart.control(cp = 0.001, minbucket = 20)
)

tree

# Plot the tree
rpart.plot(tree)

#Predict
predicted <- predict(tree, newdata = data.test, type = "class")
print(predicted)

levels(as.factor(predicted))
levels(data.test$RiskLevel)
confusionMatrix(predicted, data.test$RiskLevel)
```
```
```



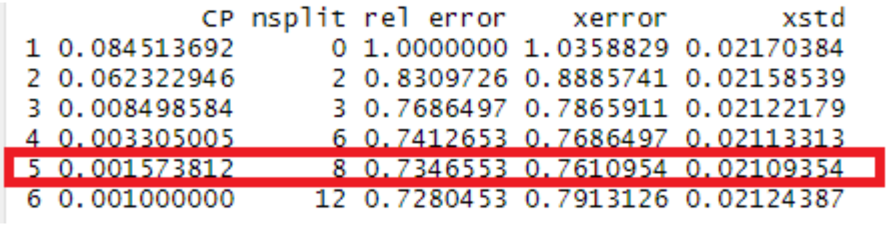
3. The confusion matrix and statistics for unpruned tree is shown below:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1   2
         0 200 127   3
         1 147 225   2
         2   0   0   0

Overall Statistics

               Accuracy : 0.6037
```

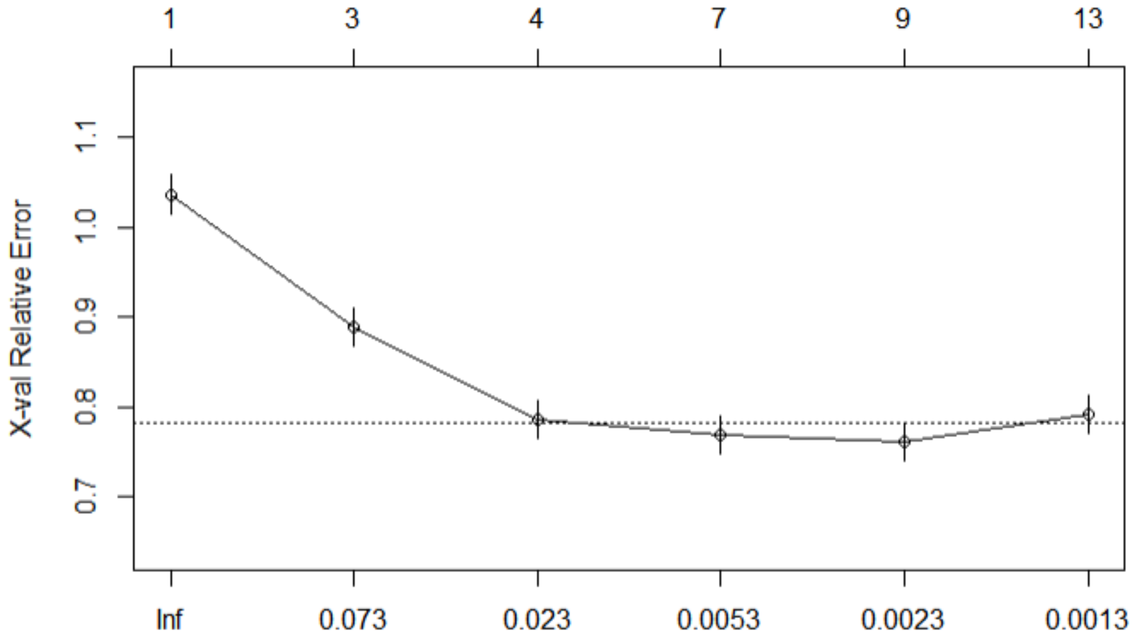
4. Next, we use cost-complexity pruning to see if we can build a simpler tree. The code and output are shown below. We will use the cp value which contributes to least 'xerror' to build a pruned tree.



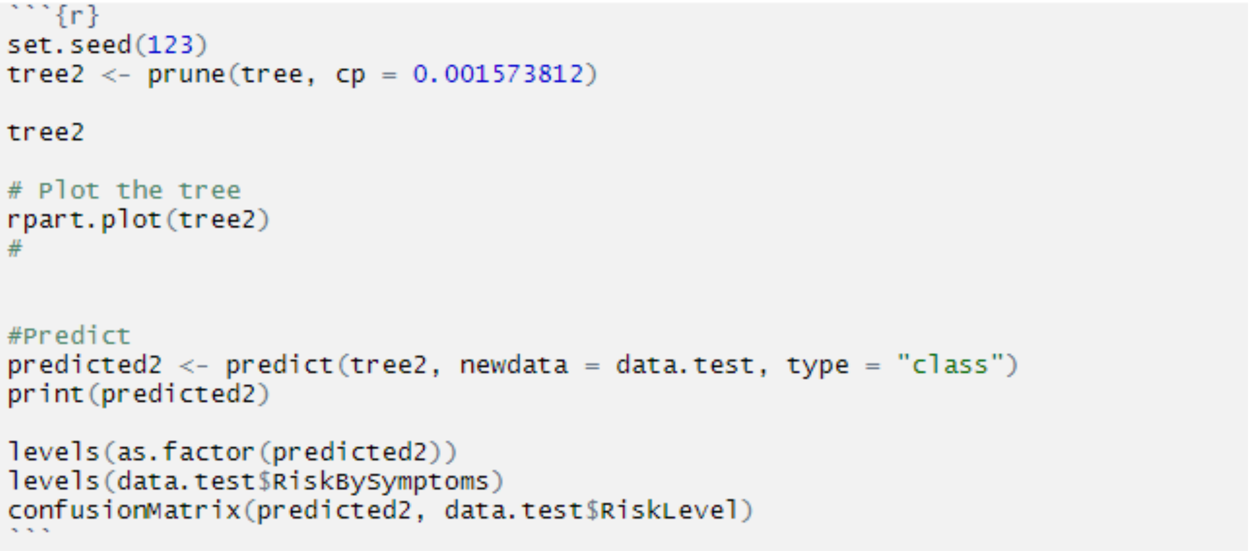
~~~
Use cost-complexity pruning to produce a reduced tree.
# Getting the CP table and CP plot to help select the optimal CP parameter
```{r}
# Display the complexity parameter table and plot for tree.
tree$cptable
plotcp(tree)
```
~~~



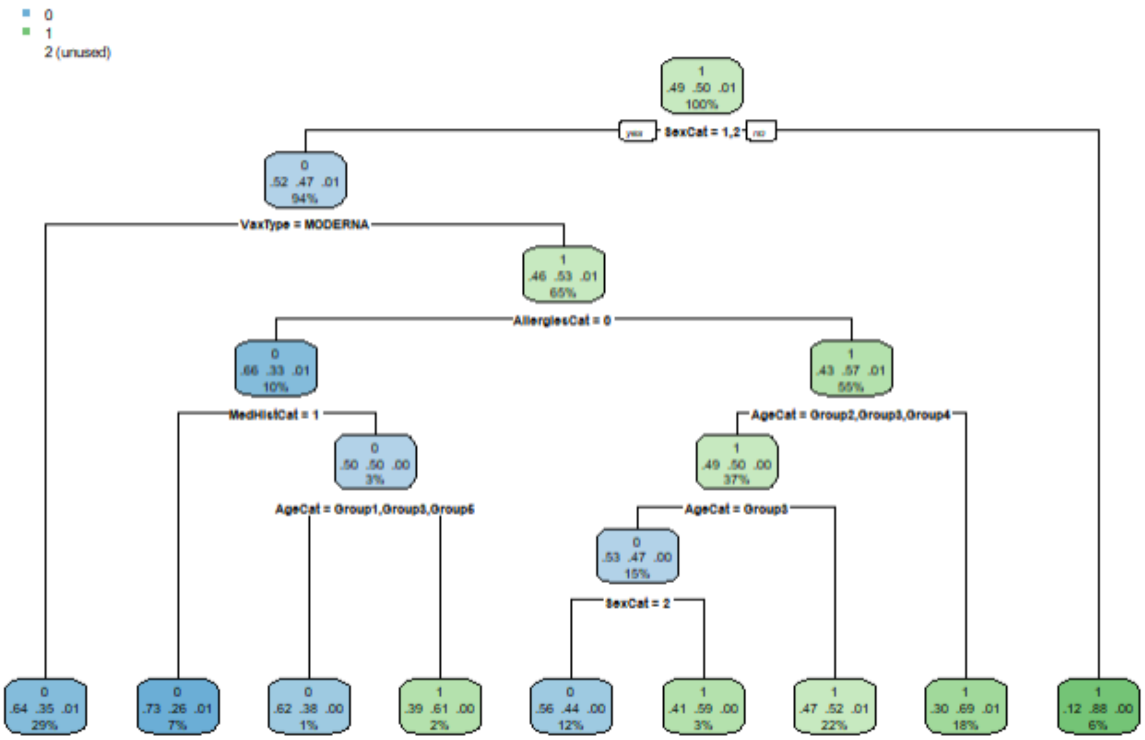
```
            CP nsplit rel error    xerror       xstd
1 0.084513692      0 1.0000000 1.0358829 0.02170384
2 0.062322946      2 0.8309726 0.8885741 0.02158539
3 0.008498584      3 0.7686497 0.7865911 0.02122179
4 0.003305005      6 0.7412653 0.7686497 0.02113313
5 0.001573812      8 0.7346553 0.7610954 0.02109354
6 0.001000000     12 0.7280453 0.7913126 0.02124387
```



5. We built a pruned tree using cp value 0.001573812. The code and output are shown below:



~~~
```{r}
set.seed(123)
tree2 <- prune(tree, cp = 0.001573812)

tree2

# Plot the tree
rpart.plot(tree2)
#


#Predict
predicted2 <- predict(tree2, newdata = data.test, type = "class")
print(predicted2)

levels(as.factor(predicted2))
levels(data.test$RiskBySymptoms)
confusionMatrix(predicted2, data.test$RiskLevel)
```
~~~



6. The confusion matrix for pruned tree is shown below:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1   2
         0 219 144   3
         1 128 208   2
         2   0   0   0

Overall Statistics

               Accuracy : 0.6065
```

We used the pruned tree since it is simpler, and it has a higher accuracy. We concluded the model result in the table below:

Condition	Prediction
Sex = Male or Female, VaxType = Moderna	High

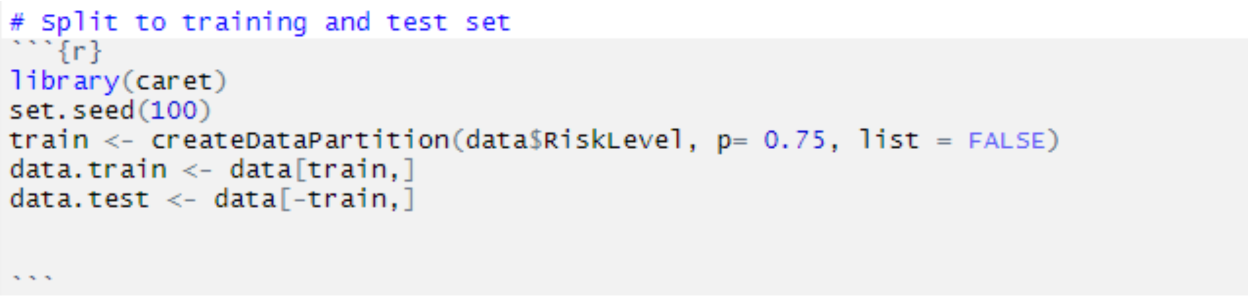
Sex = Male or Female, VaxType = Pfizer/BioNTech, Medical History = Yes, Allergies = No	
Sex = Male or Female, VaxType = Pfizer/BioNTech, Age Group = Group 1, Group 3, Group 4, Allergies = No, Medical History = No	
Sex = Female, VaxType = Pfizer/BioNTech, Allergies = Yes, Age Group = Group 3	
Sex = Male, VaxType = Pfizer/BioNTech, Allergies = No, Medical History = No, Age Group = Group 2, Group 5	Medium
Sex = Female, VaxType = Pfizer/BioNTech, Allergies = No, Medical History = No, Age Group = Group 3	
Sex = Male or Female, VaxType = Pfizer/BioNTech, Allergies = Yes, Age Group = Group 2,Group 4	
Sex = Male or Female, VaxType = Pfizer/BioNTech, Allergies = Yes, Age Group = Group 1, Group 5	
Sex ≠ 'Male' or 'Female'	

By categorizing people into 'High', 'Medium', and 'Low' groups, vaccine staff can manage the members from the high-risk group differently in future. For example, they could request the members from the high-risk group to be hospitalized for one day after taking vaccine, while request the members from the medium-risk group to be monitored on site for 15 minutes.

Scenario 2: Using Classification tree predict the death count of Covid 19 vaccinated people among different age category, gender and type of vaccination accounted.

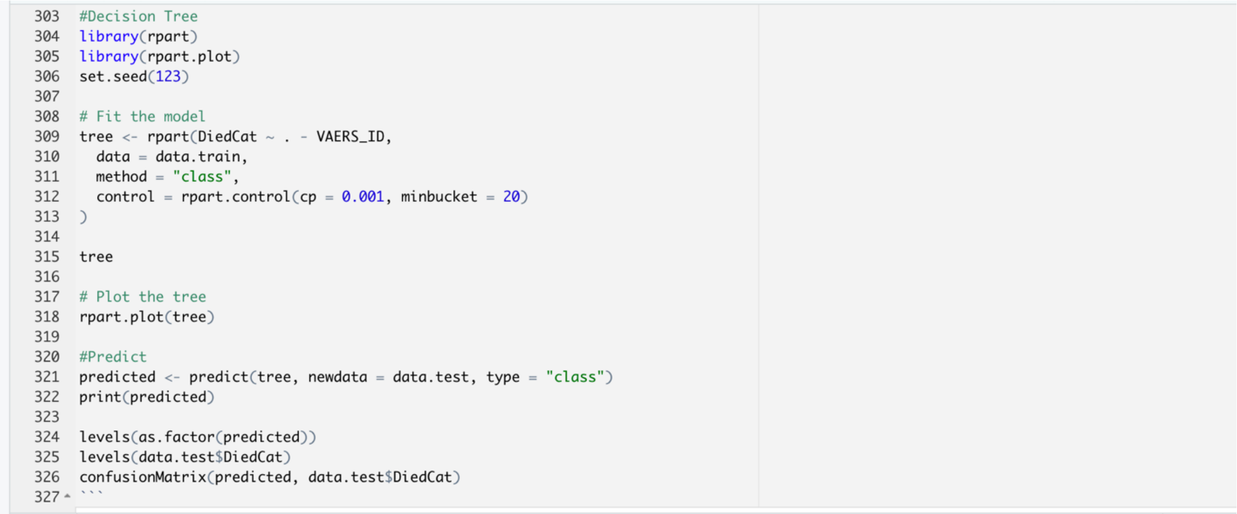
Data Cleansing and Data Transformation are done based on Scenario 1.

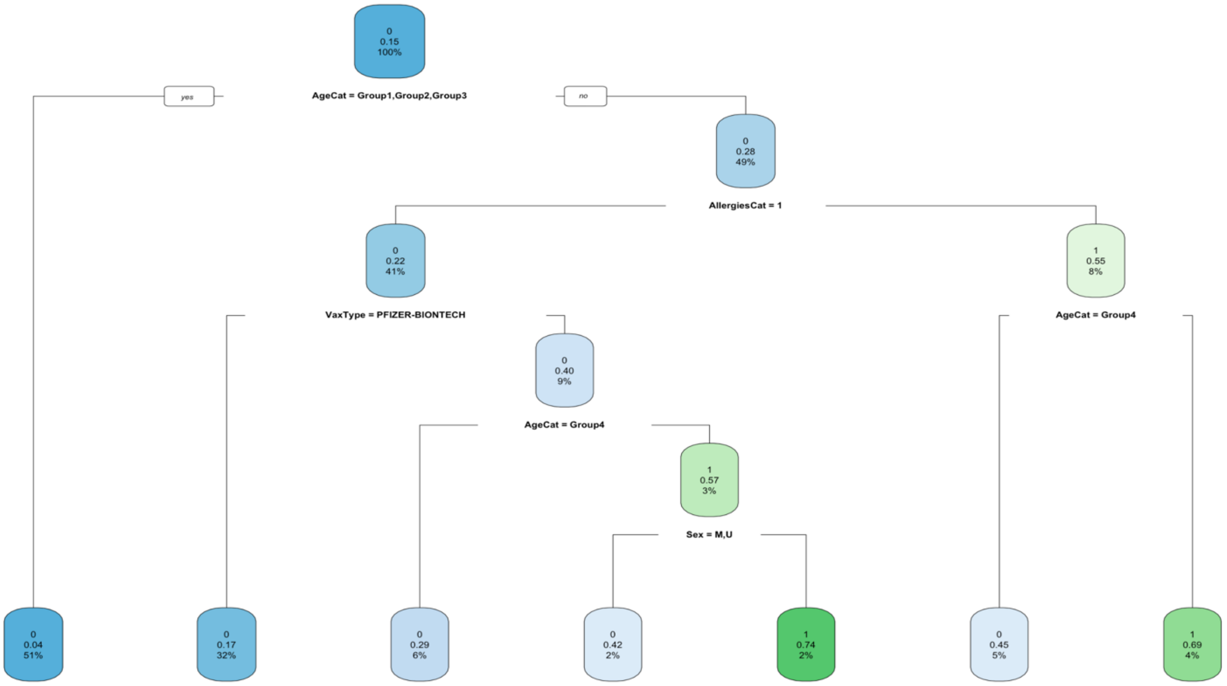
1. We started the analysis by splitting the dataset into training set and test set. In this case, we use 75% of data in training set and 25% of data in test set.



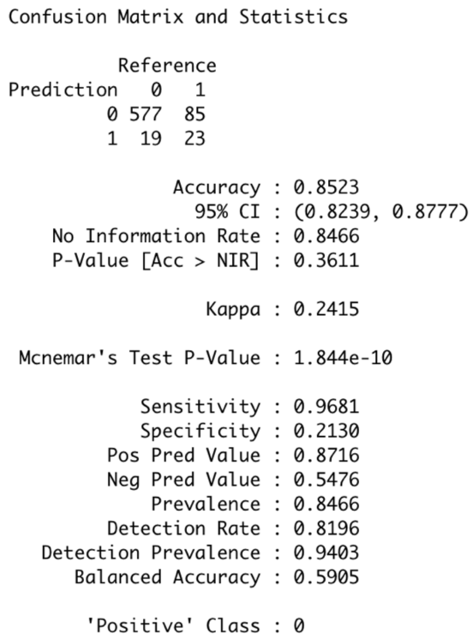
~~~
# split to training and test set
```{r}
library(caret)
set.seed(100)
train <- createDataPartition(data$RiskLevel, p= 0.75, list = FALSE)
data.train <- data[train,]
data.test <- data[-train,]


```
~~~

2. After that, we build a classification tree using Allergies, Other_Meds, History, Sex, Age_Group, Vax_Type as input variables, and DiedCat as output. The unpruned decision tree code and output are shown below:



```
#Decision Tree
library(rpart)
library(rpart.plot)
set.seed(123)

# Fit the model
tree <- rpart(DiedCat ~ . - VAERS_ID,
  data = data.train,
  method = "class",
  control = rpart.control(cp = 0.001, minbucket = 20)
)

tree

# Plot the tree
rpart.plot(tree)

#Predict
predicted <- predict(tree, newdata = data.test, type = "class")
print(predicted)

levels(as.factor(predicted))
levels(data.test$DiedCat)
confusionMatrix(predicted, data.test$DiedCat)
```
```


```

3. The confusion matrix and statistics for unpruned tree is shown below:



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 577  85
         1  19  23

               Accuracy : 0.8523
                 95% CI : (0.8239, 0.8777)
    No Information Rate : 0.8466
    P-Value [Acc > NIR] : 0.3611

                  Kappa : 0.2415

 Mcnemar's Test P-Value : 1.844e-10

            Sensitivity : 0.9681
            Specificity : 0.2130
         Pos Pred Value : 0.8716
         Neg Pred Value : 0.5476
             Prevalence : 0.8466
         Detection Rate : 0.8196
   Detection Prevalence : 0.9403
      Balanced Accuracy : 0.5905

       'Positive' Class : 0
```

Overall, we received an accuracy of 85.23% in predicting the actual death rate among Covid Vaccinated patients.

Conclusion on Scenario 2:

Condition	Prediction
Age category = Group 1, Group2, Group 3	Not Died
Age category = Group4, Group 5, allergies = yes, vaccination = Pfizer Biotech	
Age category = Group4, allergies = yes, vaccination = Moderna	
Age category = Group4, allergies = yes, Vaccination = Not vaccinated	
Age category = Group5, Group 6, allergies = yes, vaccination = Moderna, Sex = Male or Female	
Age category = Group 6, Group 6, vaccination = not vaccinated, allergies = yes, Age category = Group5, Group 6, allergies = yes, vaccination = Moderna, Sex = Male or Female	Died

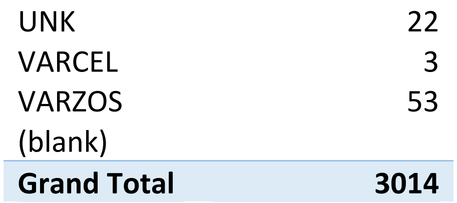
By using classification tree, we can predict and analyze that the major death count was seen in the patients among Group 5 and Group 6 with age around 60-80 years wherein very few of vaccinated patients among this groups are accounted for death as compared to death rate is on patients who haven't been vaccinated.

Cluster Analysis Using Weka

Data cleansing

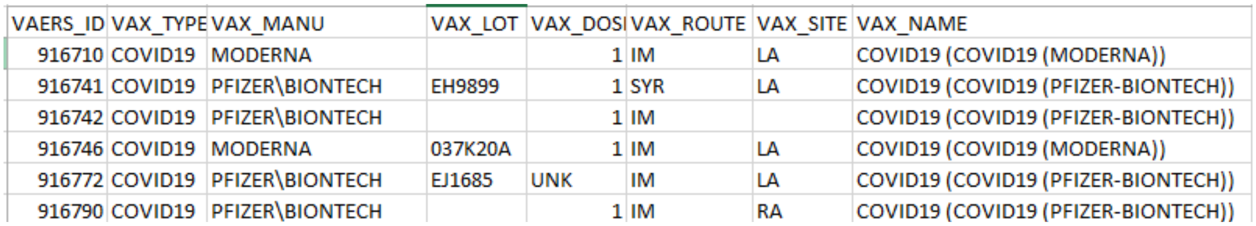
As mentioned in the section above on datasets, the data set 2021VAERSVAX listed all different types of vaccines. As there were 28 different types of vaccines across 3014 rows we had to cleanse the data to only retain the COVID 19 vaccines.

Row Labels	Count of VAX_TYPE
CHOL	1
COVID19	2844
DT	1
DTAP	2
DTAPHEPBIP	1
FLU3	1
FLU4	20
FLUA3	1
FLUA4	2
FLUC3	1
FLUC4	2
FLUR4	4
FLUX	17
HEP	5
HEPA	2
HIBV	1
HPV4	1
HPV9	2
MMR	6
MNQ	2
PNC13	2
PPV	12
RV1	1
TDAP	3
TTOX	2



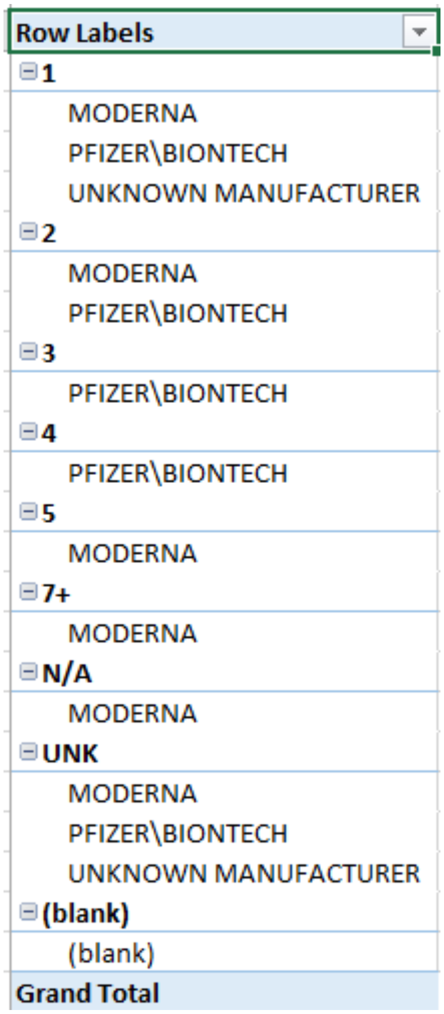
UNK	22
VARCEL	3
VARZOS	53
(blank)	
Grand Total	**3014**

The raw data also consisted of 8 different columns.



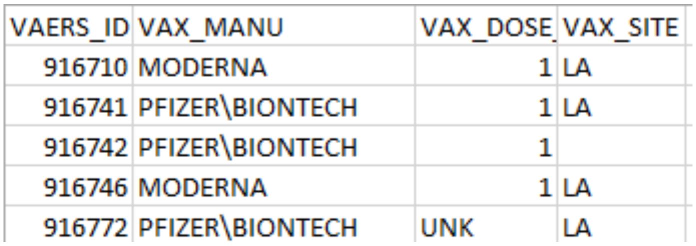
VAERS_ID	VAX_TYPE	VAX_MANU	VAX_LOT	VAX_DOSI	VAX_ROUTE	VAX_SITE	VAX_NAME
916710	COVID19	MODERNA		1	IM	LA	COVID19 (COVID19 (MODERNA))
916741	COVID19	PFIZER\BIONTECH	EH9899	1	SYR	LA	COVID19 (COVID19 (PFIZER-BIONTECH))
916742	COVID19	PFIZER\BIONTECH		1	IM		COVID19 (COVID19 (PFIZER-BIONTECH))
916746	COVID19	MODERNA	037K20A	1	IM	LA	COVID19 (COVID19 (MODERNA))
916772	COVID19	PFIZER\BIONTECH	EJ1685	UNK	IM	LA	COVID19 (COVID19 (PFIZER-BIONTECH))
916790	COVID19	PFIZER\BIONTECH		1	IM	RA	COVID19 (COVID19 (PFIZER-BIONTECH))

To narrow the scope of analysis as well as to aid the Weka data mining tool, we chose to reduce the number of columns to only keep ones essential to our research as well as those that sparked curiosity within us. One such column was vaccine doses. Going into this research we assumed that people only received 1 or 2 doses of a Covid vaccine however the data showed that some people received more. One that especially caught our attention was with Moderna patients having 7+ doses of the vaccine.



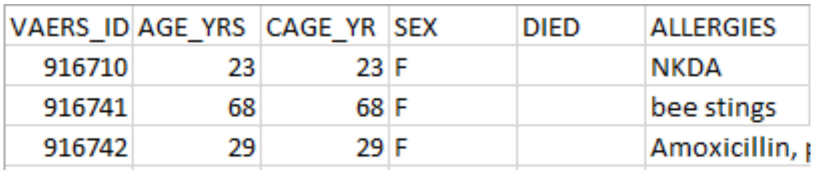
Row Labels
⊟1
MODERNA
PFIZER\BIONTECH
UNKNOWN MANUFACTURER
⊟2
MODERNA
PFIZER\BIONTECH
⊟3
PFIZER\BIONTECH
⊟4
PFIZER\BIONTECH
⊟5
MODERNA
⊟7+
MODERNA
⊟N/A
MODERNA
⊟UNK
MODERNA
PFIZER\BIONTECH
UNKNOWN MANUFACTURER
⊟(blank)
(blank)
Grand Total

Ultimately, only half of the original columns were kept. The VAERS_ID was kept so that we could link a vaccine to a patient, VAX_MANU so that we knew which manufacturer made the vaccine received by the patient, VAX_DOSE so that we could know (if indicated) how many doses a patient received and VAX_SITE so that we could know which arm (or other place) if captured that the dose was administered.



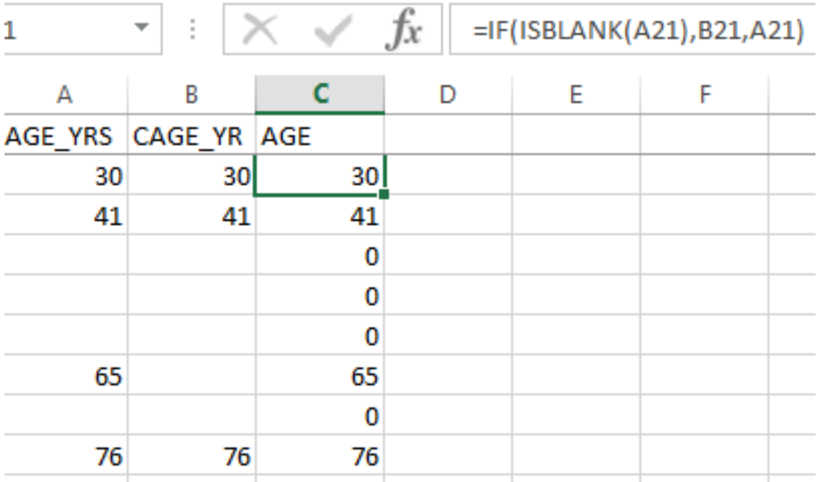
VAERS_ID	VAX_MANU	VAX_DOSE	VAX_SITE
916710	MODERNA	1	LA
916741	PFIZER\BIONTECH	1	LA
916742	PFIZER\BIONTECH	1	
916746	MODERNA	1	LA
916772	PFIZER\BIONTECH	UNK	LA

The next data set that we were interested in is the one that included patient data. This file was quite wide with 35 columns. As our team was interested in data mining information related to death, age and vaccine manufacturer, we again trimmed the number of columns to only those we felt essential.



VAERS_ID	AGE_YRS	CAGE_YR	SEX	DIED	ALLERGIES
916710	23	23	F		NKDA
916741	68	68	F		bee stings
916742	29	29	F		Amoxicillin, p

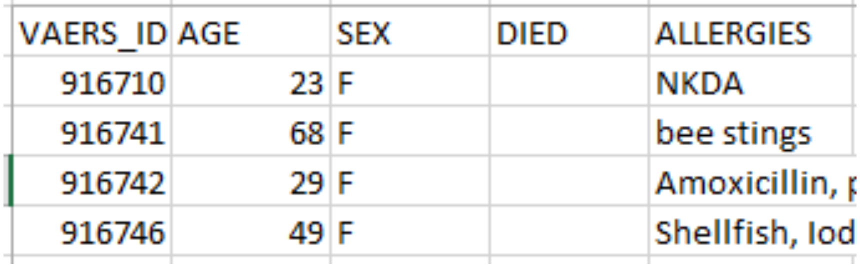
We kept 6 columns. As much work is often needed to cleanse the data more worked needed to be done. Specifically, the age columns were not consistent in that sometimes AGE_YRS had a value while CAGE_YR was blank. To solve this inconsistency we merged the columns into 1 single column keeping the value where the age was not blank. Using Excel we replaced all blank AGE_YRS with CAGE_YR if it had a value. If neither column had a value then a zero was inserted. Inserting zeros where the age was unknown also helped with solve potential future issues with Nulls in data analysis.



```
=IF(ISBLANK(A21),B21,A21)
```

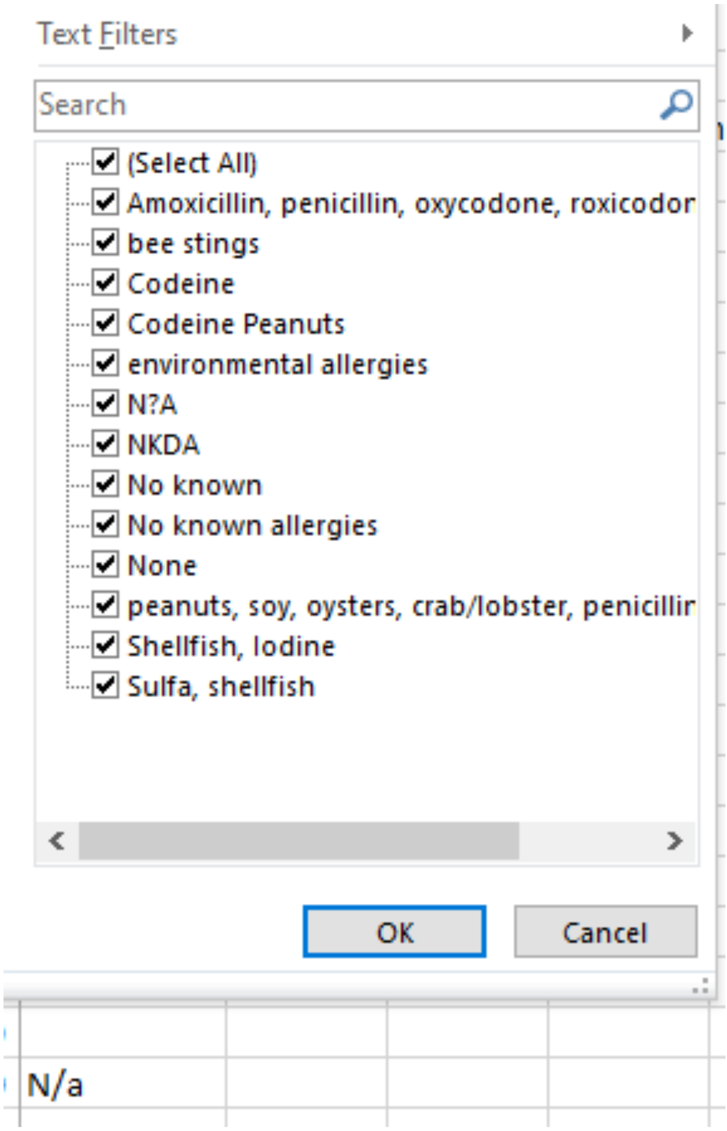
A	B	C	D	E	F
AGE_YRS	CAGE_YR	AGE			
30	30	30			
41	41	41			
		0			
		0			
		0			
65		65			
		0			
76	76	76			

The next data issue that we felt needed to be cleaned up was allergies.



VAERS_ID	AGE	SEX	DIED	ALLERGIES
916710	23	F		NKDA
916741	68	F		bee stings
916742	29	F		Amoxicillin, p
916746	49	F		Shellfish, Iod

As the data was free form text, we felt it would be best to classify it into a binary format. 0 for no allergies or unknown and 1 for allergies. This was not so simple due to the nature of the data.

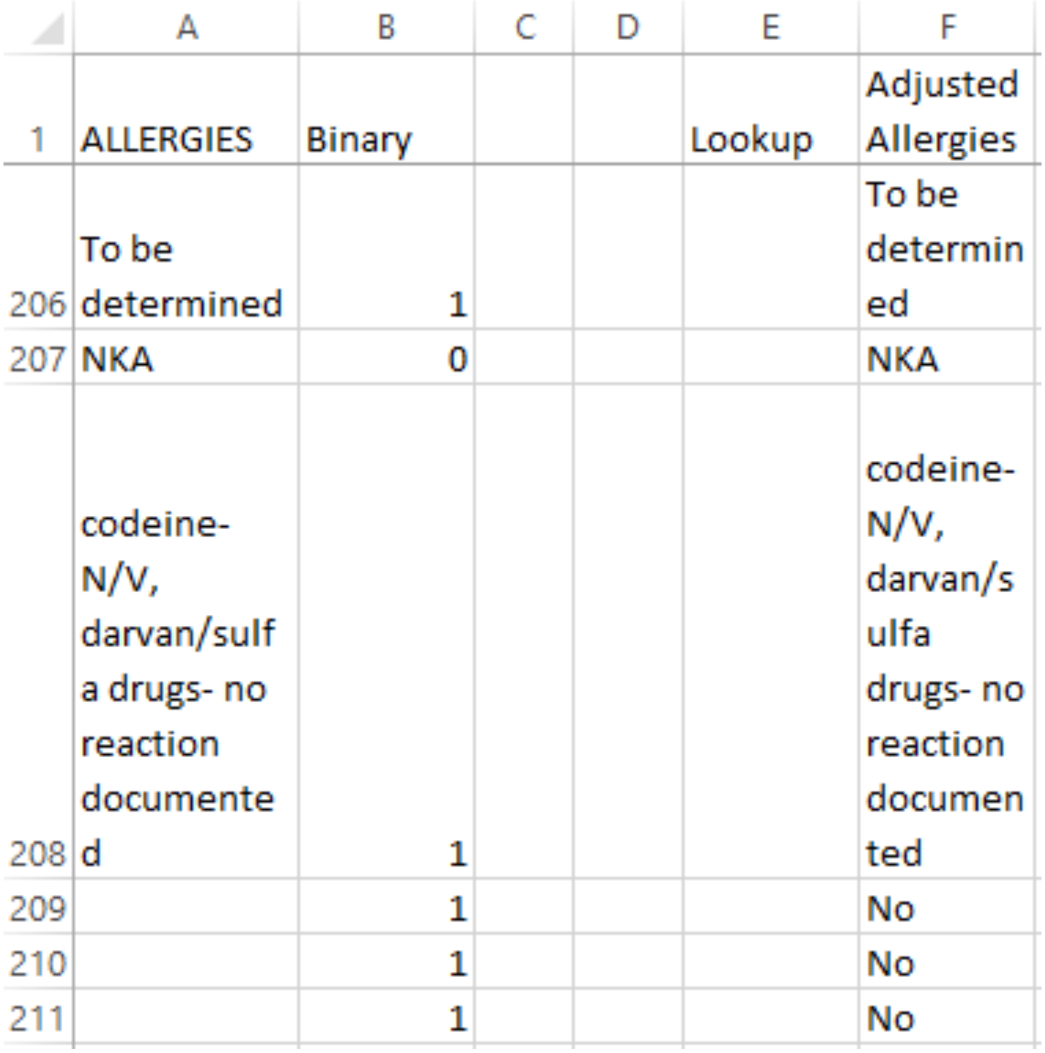


☑ (Select All)
☑ Amoxicillin, penicillin, oxycodone, roxicodor
☑ bee stings
☑ Codeine
☑ Codeine Peanuts
☑ environmental allergies
☑ N?A
☑ NKDA
☑ No known
☑ No known allergies
☑ None
☑ peanuts, soy, oysters, crab/lobster, penicillir
☑ Shellfish, Iodine
☑ Sulfa, shellfish

N/a

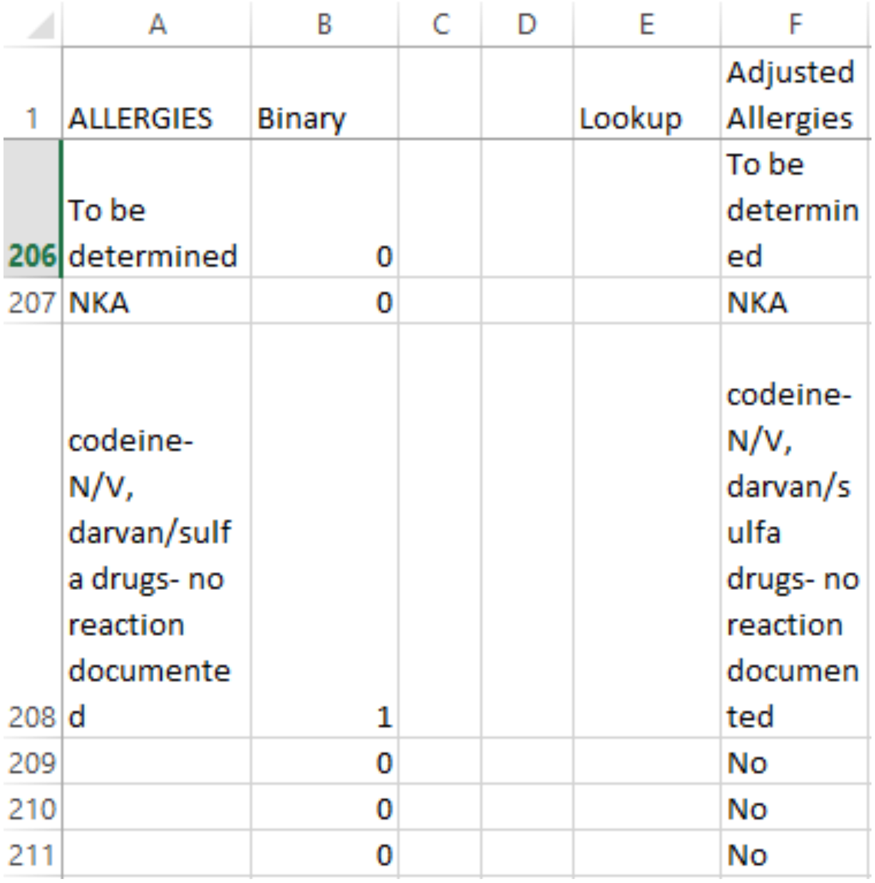
We had to take a step back from the binary approach when it was discovered that there was an issue with blanks. What we had to do was first replace the blanks with “No” which was a value in our lookup list. Then we ran our binary function again on this new list which no longer contained missing values.

Below shows an example of the issues with the data that I had to solve for:



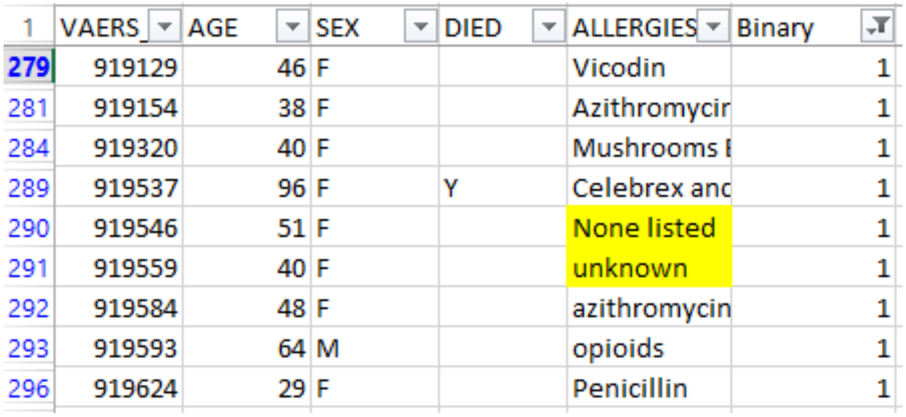
	A	B	C	D	E	F
1	ALLERGIES	Binary			Lookup	Adjusted Allergies
206	To be determined	1				To be determined
207	NKA	0				NKA
208	codeine-N/V, darvan/sulfa drugs- no reaction documented	1				codeine-N/V, darvan/sulfa drugs- no reaction documented
209		1				No
210		1				No
211		1				No

The Binary column for rows 209 to 211 shows a classic Type 2 error where the data is misclassified. The Adjusted Allergies columns shows our formula correcting this. However note that “To be determined” was not accounted for so it is being classified as an allergy. For this analysis exercise the assumption will be that “To be determined” means that no allergy is present. Now the algorithms were rerun against the Adjusted Allergies and the classifications look better.



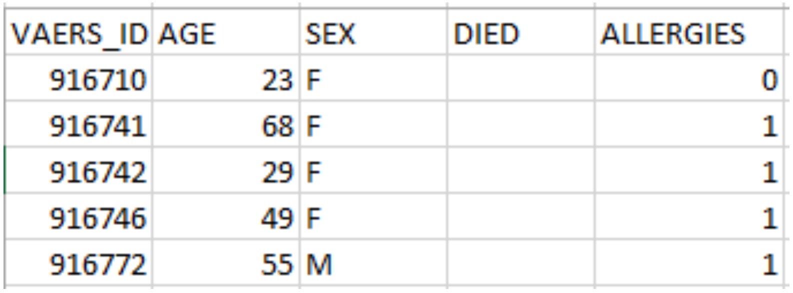
	A	B	C	D	E	F
1	ALLERGIES	Binary			Lookup	Adjusted Allergies
206	To be determined	0				To be determined
207	NKA	0				NKA
208	codeine-N/V, darvan/sulfa drugs- no reaction documented	1				codeine-N/V, darvan/sulfa drugs- no reaction documented
209		0				No
210		0				No
211		0				No

But even this is not enough. After combining the binary classification back to the main data set we had to manually test it as free form text allows for many combinations. We found that we missed some combinations.



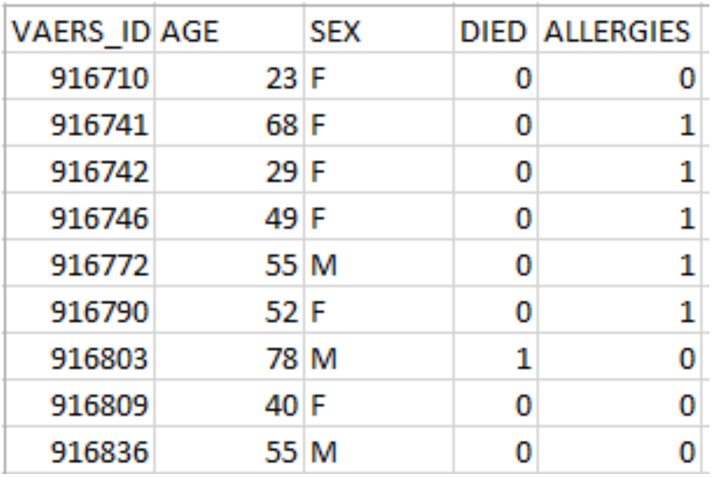
1	VAERS_	AGE	SEX	DIED	ALLERGIES	Binary
279	919129	46	F		Vicodin	1
281	919154	38	F		Azithromycir	1
284	919320	40	F		Mushrooms I	1
289	919537	96	F	Y	Celebrex anc	1
290	919546	51	F		None listed	1
291	919559	40	F		unknown	1
292	919584	48	F		azithromycin	1
293	919593	64	M		opioids	1
296	919624	29	F		Penicillin	1

So now, just like we learned in our cluster analysis lecture, we must reiterate the process in order to get the data as clean as possible with the smallest margin for error. Our list of negative responses that started at 14 grew to almost 160. After many iterations we were feeling good about the allergies column.



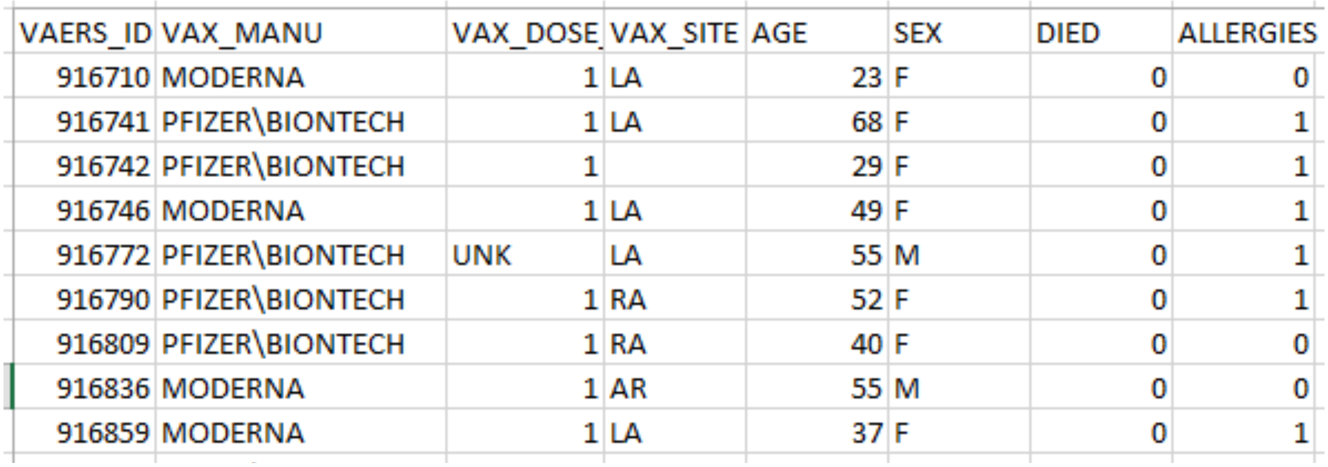
VAERS_ID	AGE	SEX	DIED	ALLERGIES
916710	23	F		0
916741	68	F		1
916742	29	F		1
916746	49	F		1
916772	55	M		1

Now we needed to move on to the DIED column. Our data mining tool does not do well with missing data. The good news is that cleaning the DIED column would be very easy. 1 will indicate that the patient died and 0 will indicate that they did not die.



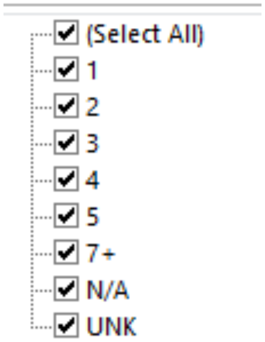
VAERS_ID	AGE	SEX	DIED	ALLERGIES
916710	23	F	0	0
916741	68	F	0	1
916742	29	F	0	1
916746	49	F	0	1
916772	55	M	0	1
916790	52	F	0	1
916803	78	M	1	0
916809	40	F	0	0
916836	55	M	0	0

We've now merged our 2 datasets into one dataset. The data is looking almost ready to feed into Weka, our data mining tool however as we look at the merged outcome, still we see areas where there are missing values that we need to make some decisions on.



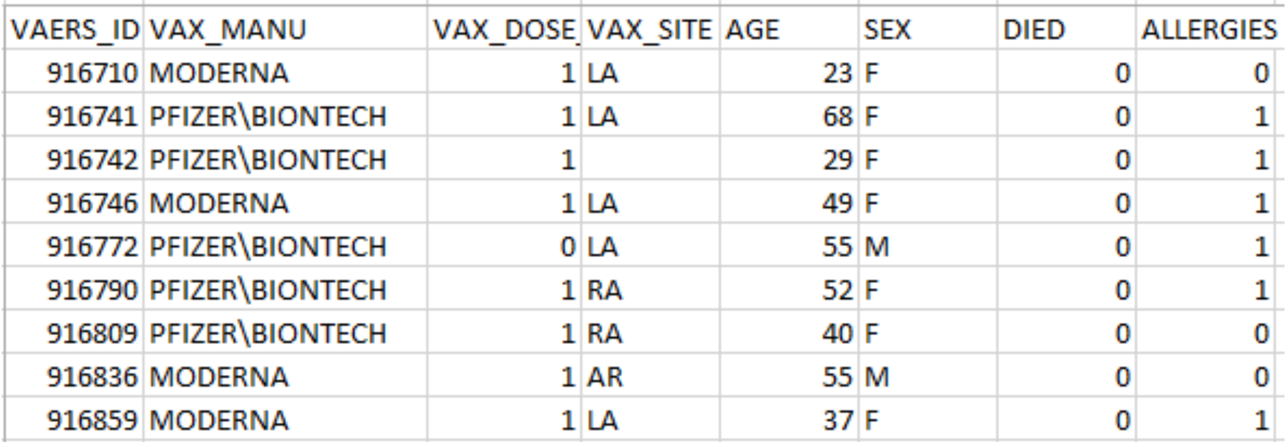
VAERS_ID	VAX_MANU	VAX_DOSE_	VAX_SITE	AGE	SEX	DIED	ALLERGIES
916710	MODERNA	1	LA	23	F	0	0
916741	PFIZER\BIONTECH	1	LA	68	F	0	1
916742	PFIZER\BIONTECH	1		29	F	0	1
916746	MODERNA	1	LA	49	F	0	1
916772	PFIZER\BIONTECH	UNK	LA	55	M	0	1
916790	PFIZER\BIONTECH	1	RA	52	F	0	1
916809	PFIZER\BIONTECH	1	RA	40	F	0	0
916836	MODERNA	1	AR	55	M	0	0
916859	MODERNA	1	LA	37	F	0	1

The VAX_DOSE data types are not uniform and there are missing values in the VAX_SITE column. We know that Weka is sensitive to these data anomalies. Starting with the list of unique values for VAX_DOSE we can see 3 data values that will need to change.



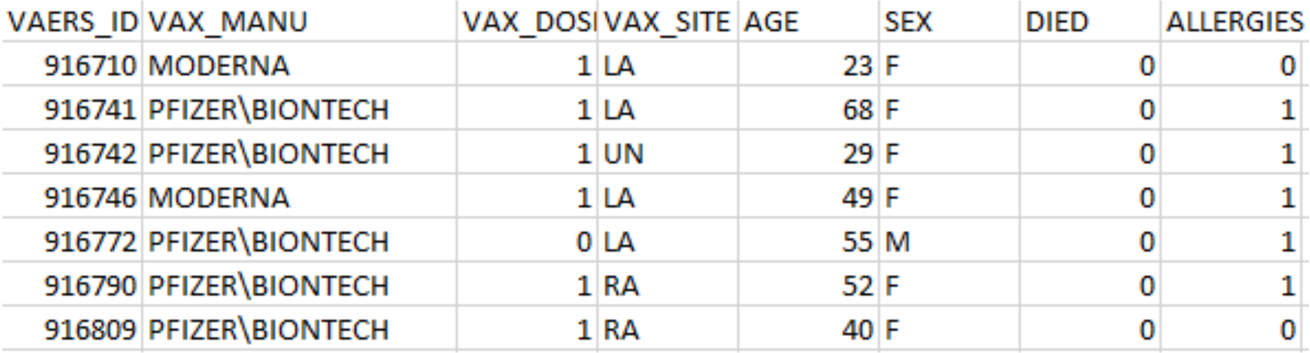
- ☑ (Select All)
- ☑ 1
- ☑ 2
- ☑ 3
- ☑ 4
- ☑ 5
- ☑ 7+
- ☑ N/A
- ☑ UNK

We made the decision that if a patient has had 7 or more doses then we will cap the value at 7. If it is unknown (UNK) or not available (N/A) then we will use the value 0 to represent that.



VAERS_ID	VAX_MANU	VAX_DOSE	VAX_SITE	AGE	SEX	DIED	ALLERGIES
916710	MODERNA	1	LA	23	F	0	0
916741	PFIZER\BIONTECH	1	LA	68	F	0	1
916742	PFIZER\BIONTECH	1		29	F	0	1
916746	MODERNA	1	LA	49	F	0	1
916772	PFIZER\BIONTECH	0	LA	55	M	0	1
916790	PFIZER\BIONTECH	1	RA	52	F	0	1
916809	PFIZER\BIONTECH	1	RA	40	F	0	0
916836	MODERNA	1	AR	55	M	0	0
916859	MODERNA	1	LA	37	F	0	1

Next we move on to cleanse the VAX_SITE. If the value is missing or NA then it will be changed to UN for Unknown. After making this change we finally feel ready to start doing analysis in Weka on this cleansed dataset.



VAERS_ID	VAX_MANU	VAX_DOSI	VAX_SITE	AGE	SEX	DIED	ALLERGIES
916710	MODERNA	1	LA	23	F	0	0
916741	PFIZER\BIONTECH	1	LA	68	F	0	1
916742	PFIZER\BIONTECH	1	UN	29	F	0	1
916746	MODERNA	1	LA	49	F	0	1
916772	PFIZER\BIONTECH	0	LA	55	M	0	1
916790	PFIZER\BIONTECH	1	RA	52	F	0	1
916809	PFIZER\BIONTECH	1	RA	40	F	0	0

Data Constraints and Assumptions

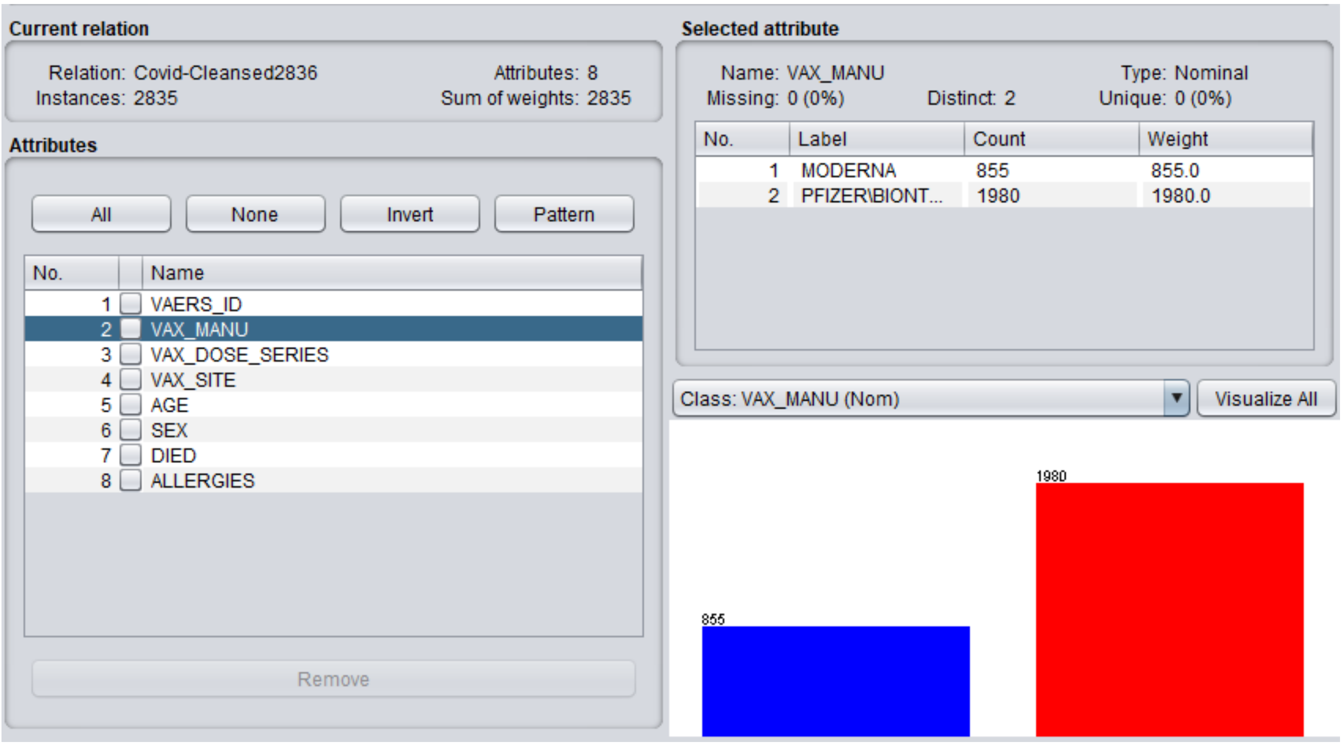
In summary, here are the assumptions and constraints of the data:

- The merged and cleansed dataset contains 2835 rows
- The data was captured on February 18, 2021 and is a point in time snapshot
- Only Covid19 vaccines administered in the USA exist within the dataset
- There are only 2 manufacturers in the data set, Pfizer and Moderna
- If the age of the patient is unknown then it is classified as 0
- Unknown number of vaccines received by a patient is listed as 0
- The maximum number of vaccines received by a patient is capped at 7
- If a patient died it is categorized as 1
- If a patient is still alive then it is categorized as 0
- If a patient had no prior allergies it is categorized as 0
- If a patient had prior allergies then it is categorized as 1

Now, let's begin to see some insights using Weka!

Weka Insights

Observations of 2835 instances are confirmed. We are able to easily see that Pfizer has been more widely administered than Moderna. With 1980 doses of Pfizer administered it is coming in at more than double the Moderna vaccines administered at the time our data was captured.



Current relation

Relation: Covid-Cleansed2836 Attributes: 8
Instances: 2835 Sum of weights: 2835

Attributes

No.	Name
1	□ VAERS_ID
2	□ VAX_MANU
3	□ VAX_DOSE_SERIES
4	□ VAX_SITE
5	□ AGE
6	□ SEX
7	□ DIED
8	□ ALLERGIES

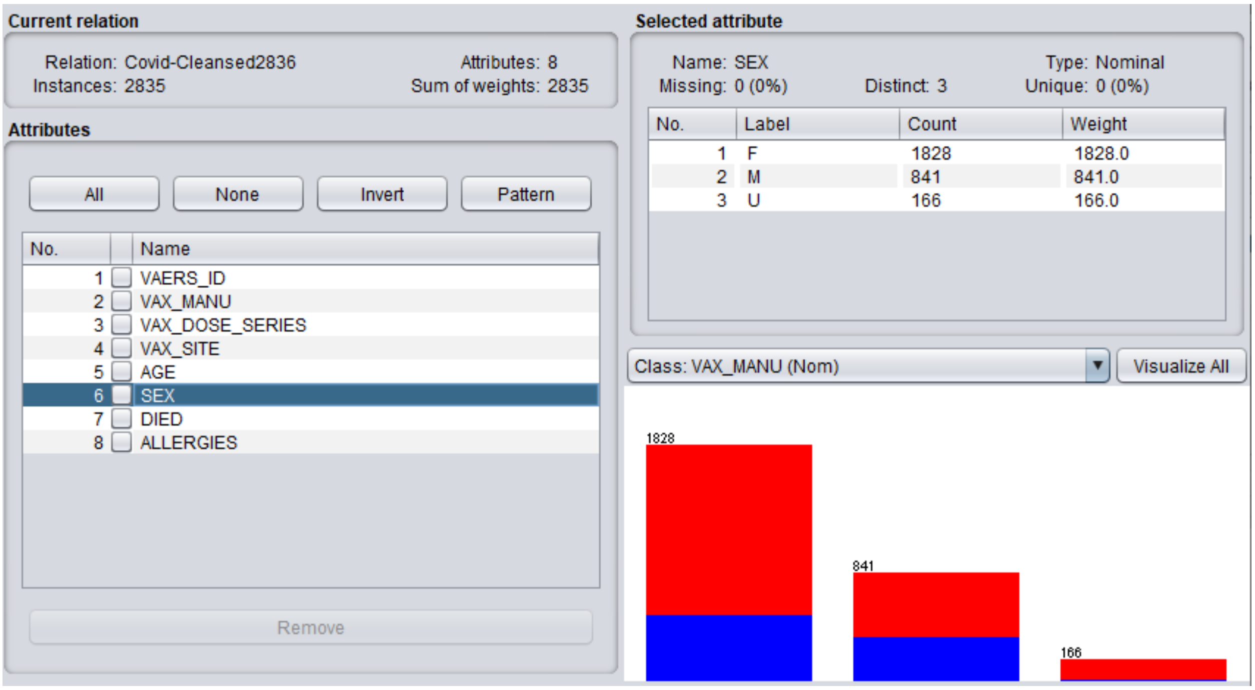
Selected attribute

Name: VAX_MANU Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	MODERNA	855	855.0
2	PFIZER\BIONT...	1980	1980.0

Class: VAX_MANU (Nom)

Based on gender, Pfizer has been significantly more prevalent with people classified as Female. Males are almost a 50/50 split between Pfizer and Moderna and people classified as non-binary show significantly more vaccinations with Moderna than Pfizer.



Current relation

Relation: Covid-Cleansed2836 Attributes: 8
Instances: 2835 Sum of weights: 2835

Attributes

No.	Name
1	VAERS_ID
2	VAX_MANU
3	VAX_DOSE_SERIES
4	VAX_SITE
5	AGE
6	SEX
7	DIED
8	ALLERGIES

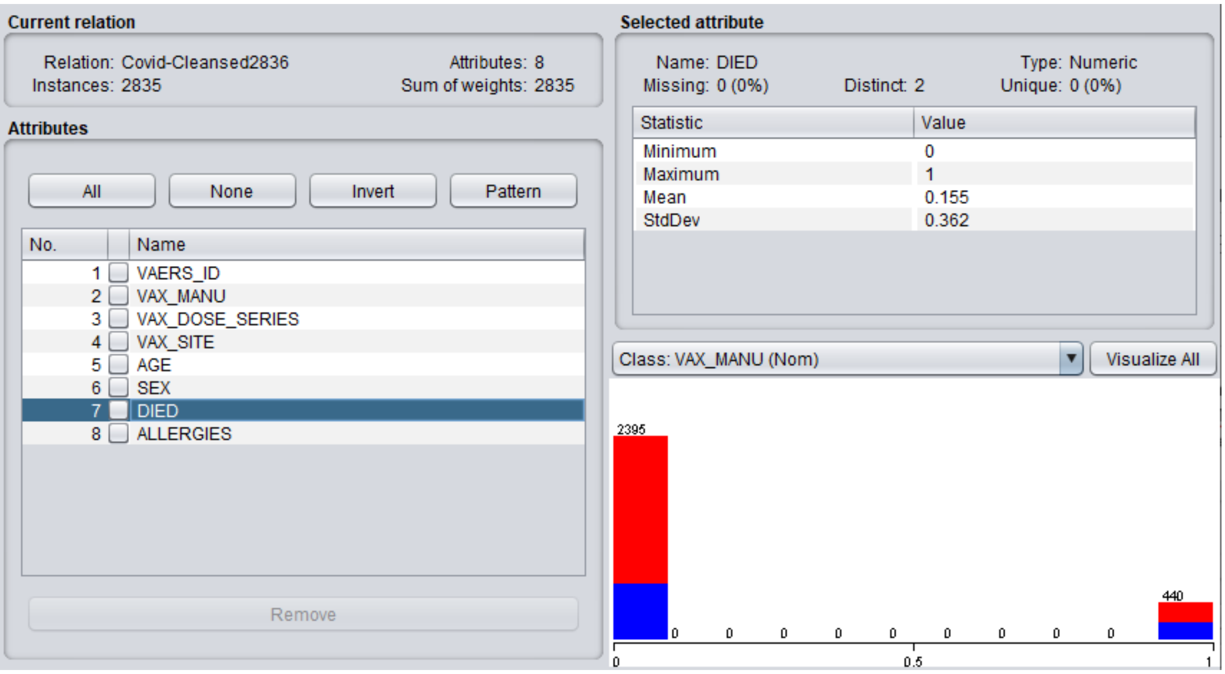
Selected attribute

Name: SEX Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	F	1828	1828.0
2	M	841	841.0
3	U	166	166.0

Class: VAX_MANU (Nom)

In terms of deaths observed, Pfizer has only slightly more reported deaths than Moderna which is interesting given that over 50% more vaccines were administered by that manufacturer. The bar on the left of the graph shows patients that are still alive. The data shows in favor of survival even with the Moderna vaccine.



Current relation

Relation: Covid-Cleansed2836 Attributes: 8
Instances: 2835 Sum of weights: 2835

Attributes

No.	Name
1	VAERS_ID
2	VAX_MANU
3	VAX_DOSE_SERIES
4	VAX_SITE
5	AGE
6	SEX
7	DIED
8	ALLERGIES

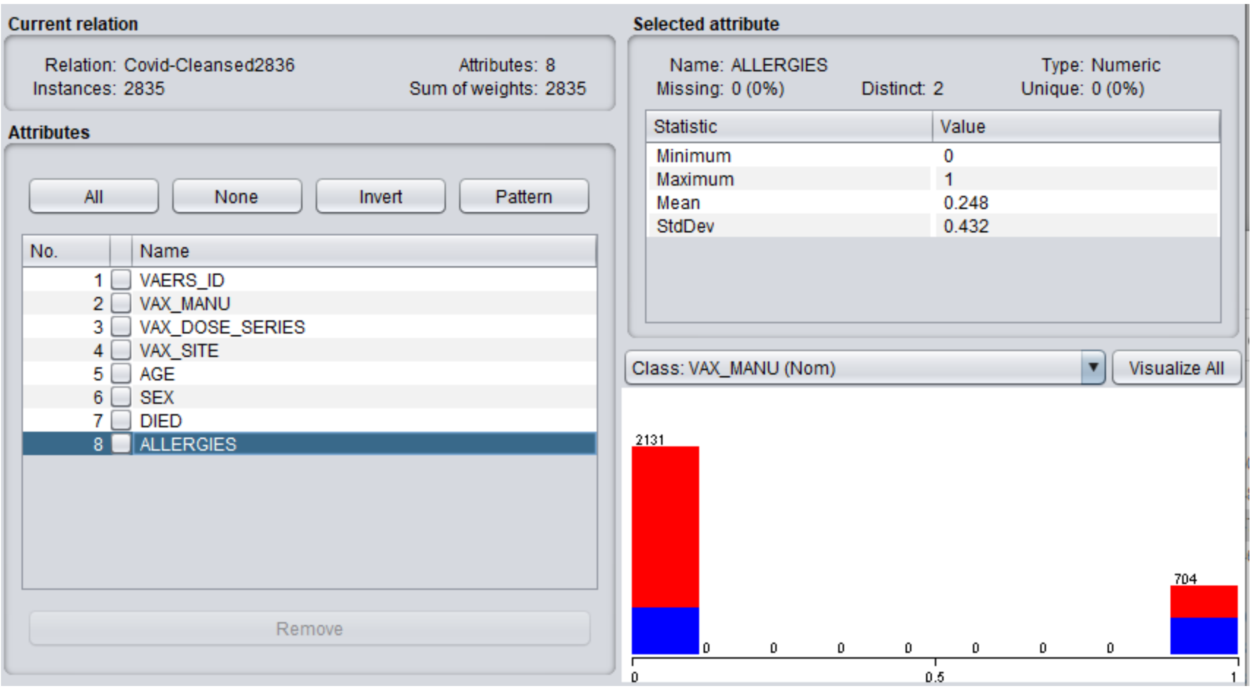
Selected attribute

Name: DIED Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.155
StdDev	0.362

Class: VAX_MANU (Nom)

Allergies shows a particularly interesting story. Most people who received the Pfizer vaccine did not have a pre-existing allergy condition. The same was not true for Moderna. It is almost half and half with only slightly more people who reported as allergy free having received the Moderna vaccine.



Current relation

Relation: Covid-Cleansed2836 Attributes: 8
Instances: 2835 Sum of weights: 2835

Attributes

No.		Name
1	□	VAERS_ID
2	□	VAX_MANU
3	□	VAX_DOSE_SERIES
4	□	VAX_SITE
5	□	AGE
6	□	SEX
7	□	DIED
8	□	ALLERGIES

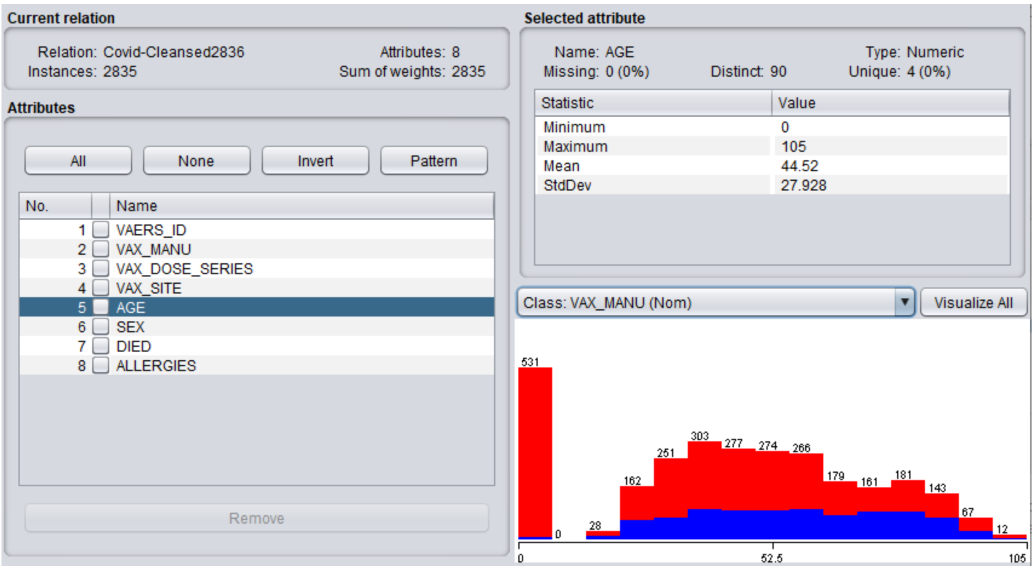
Selected attribute

Name: ALLERGIES Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.248
StdDev	0.432

Class: VAX_MANU (Nom)

Age shows another interesting story. Given the large number of Pfizer patients with unreported ages, one might deduce that perhaps that information was not deemed important at the start of the vaccines being rolled out. However over time as people were wanting to do more analysis, it was determined that age was an important attribute to have. This deduction can be made because Moderna was released after Pfizer. The number of unreported ages for Moderna patients is negligible.



Current relation

Relation: Covid-Cleansed2836 Attributes: 8
Instances: 2835 Sum of weights: 2835

Attributes

No.		Name
1	□	VAERS_ID
2	□	VAX_MANU
3	□	VAX_DOSE_SERIES
4	□	VAX_SITE
5	□	AGE
6	□	SEX
7	□	DIED
8	□	ALLERGIES

Selected attribute

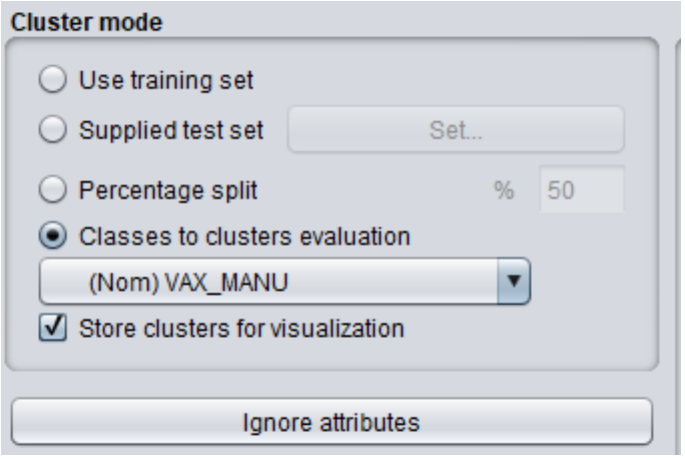
Name: AGE Type: Numeric
Missing: 0 (0%) Distinct: 90 Unique: 4 (0%)

Statistic	Value
Minimum	0
Maximum	105
Mean	44.52
StdDev	27.928

Class: VAX_MANU (Nom)

Cluster Analysis

We ran a cluster analysis using a simple K means method with Weka and chose 2 clusters evaluated by vaccine manufacturer. We ignored the VAERSID as there was no need to individually identify patients.



Cluster mode

○ Use training set

○ Supplied test set

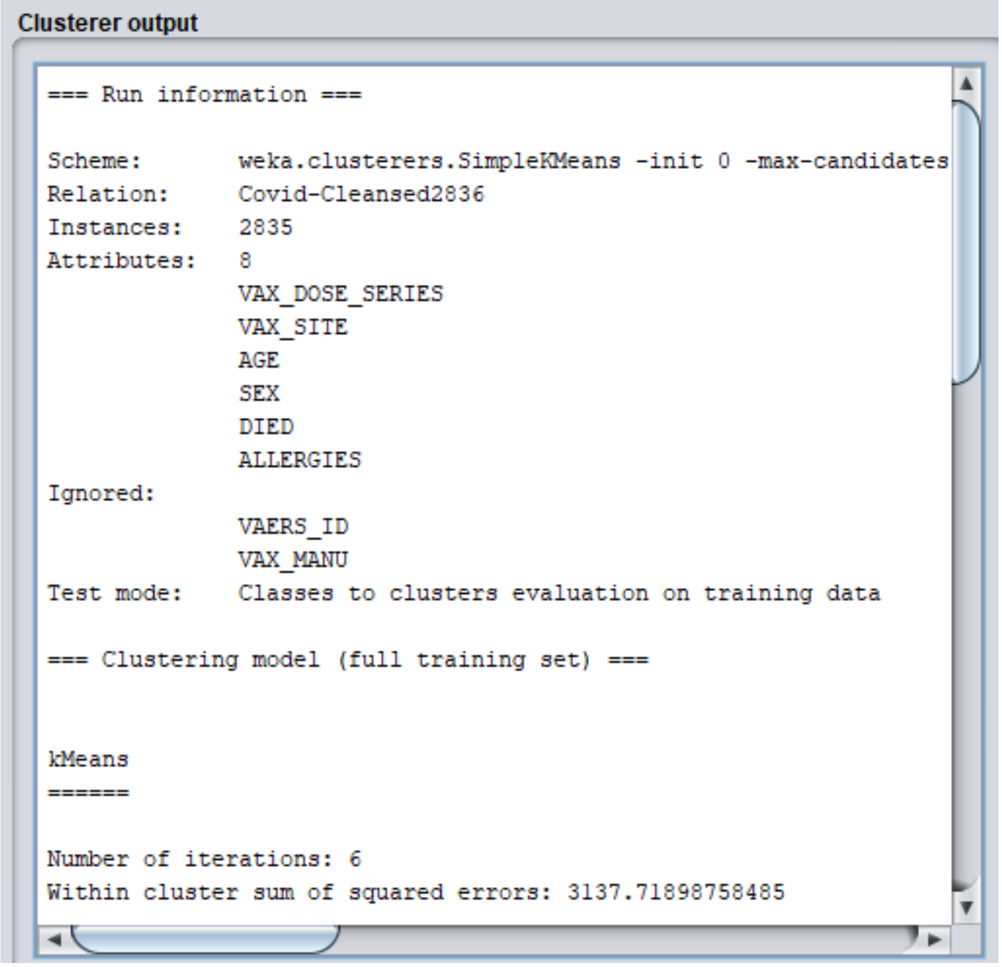
○ Percentage split % 50

● Classes to clusters evaluation

(Nom) VAX_MANU

☑ Store clusters for visualization

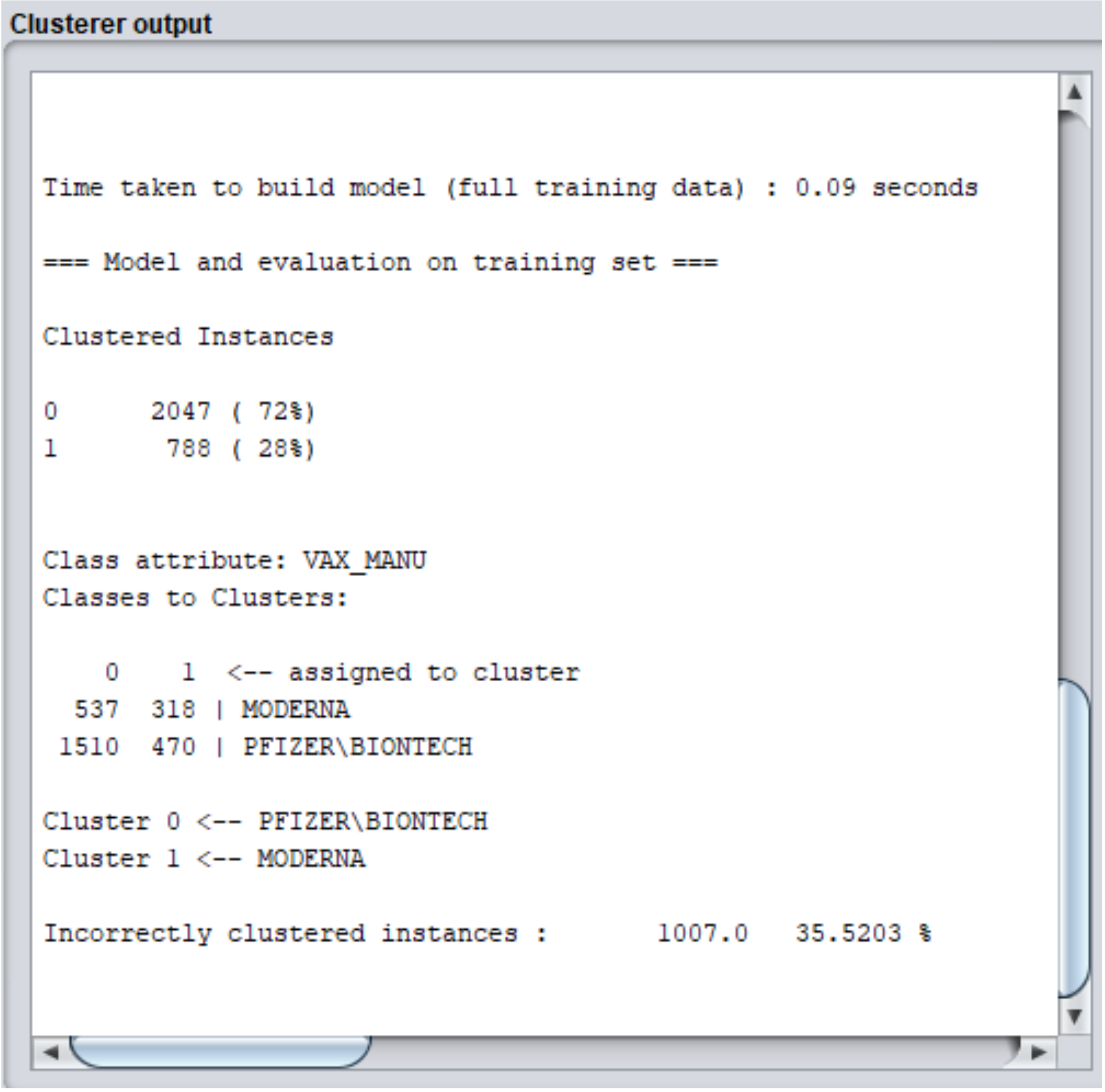
Weka provided interesting run information as seen below. 6 iterations were completed.



Clusterer output

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates
Relation:     Covid-Cleansed2836
Instances:    2835
Attributes:   8
              VAX_DOSE_SERIES
              VAX_SITE
              AGE
              SEX
              DIED
              ALLERGIES
Ignored:
              VAERS_ID
              VAX_MANU
Test mode:    Classes to clusters evaluation on training data

=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 3137.71898758485
```

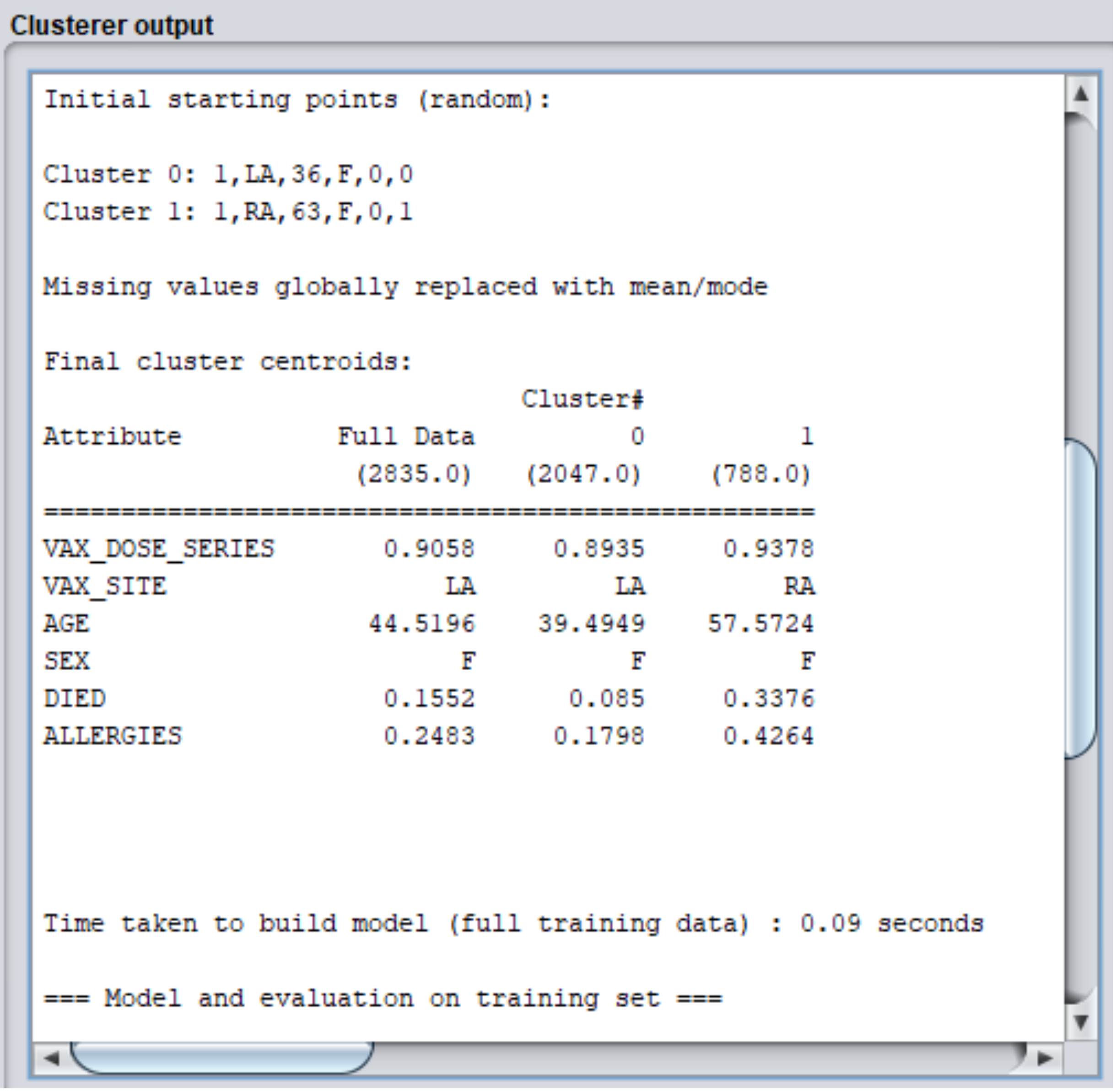
The model calculates 35% cluster errors based on 6 iterations through the training data. It has assigned Cluster 0 to Pfizer and Cluster 1 to Moderna.



```
Time taken to build model (full training data) : 0.09 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      2047 ( 72%)
1       788 ( 28%)


Class attribute: VAX_MANU
Classes to Clusters:

    0    1  <-- assigned to cluster
  537  318 | MODERNA
 1510  470 | PFIZER\BIONTECH

Cluster 0 <-- PFIZER\BIONTECH
Cluster 1 <-- MODERNA

Incorrectly clustered instances :       1007.0   35.5203 %
```

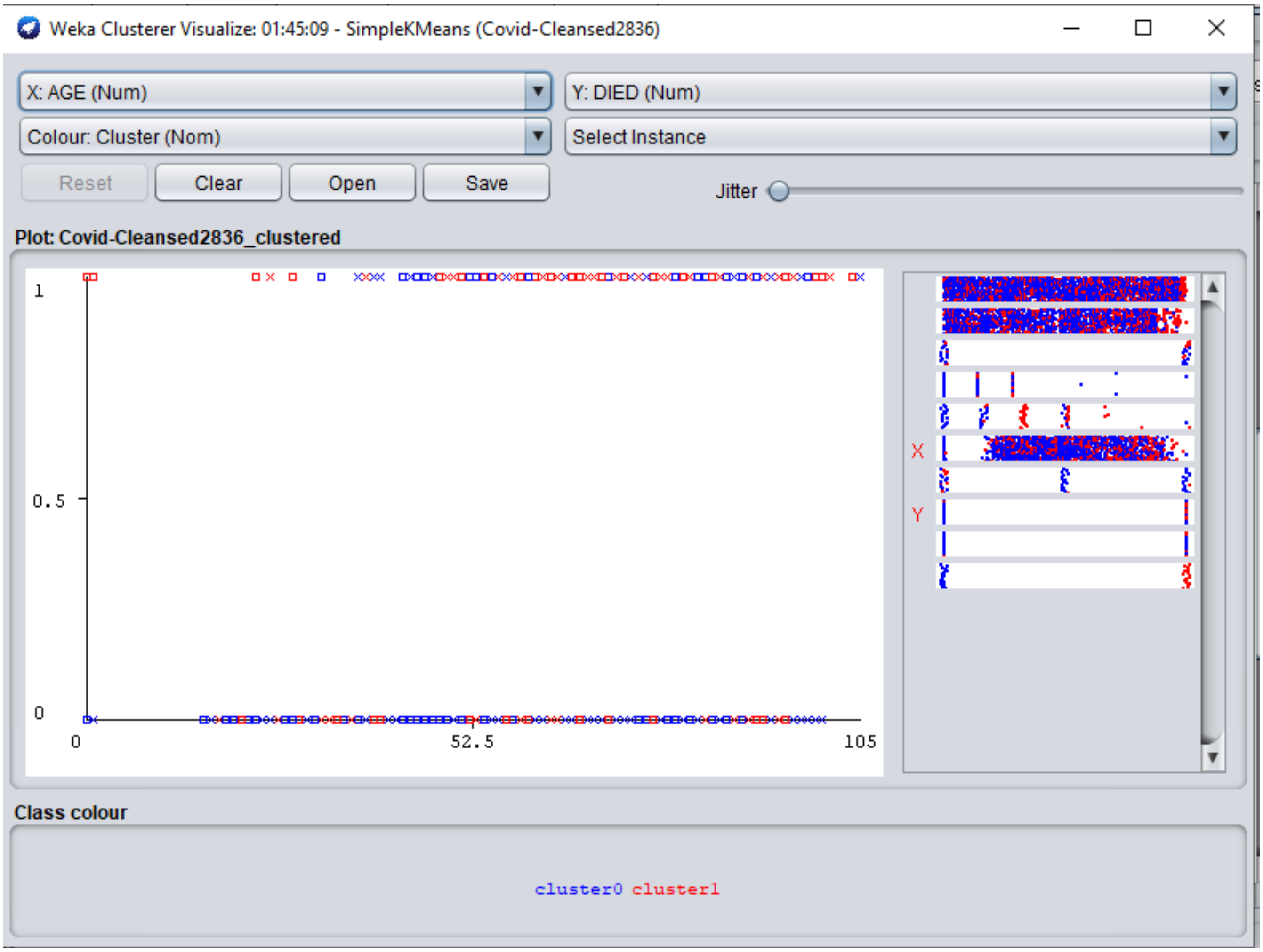
Below are the final cluster centroids. Due to our diligence cleansing our data, there were no missing values so nothing had to be replaced with the mean/mode. Weka is remarkable fast when it comes to running data mining algorithms. We can see that running 6 iterations across 6 attributes only took Weka 0.09 seconds!

Based on the final cluster centroids the model predicts that more Moderna patients will choose to receive their vaccine in their right arm while Pfizer patients will tend to favor their left arm.

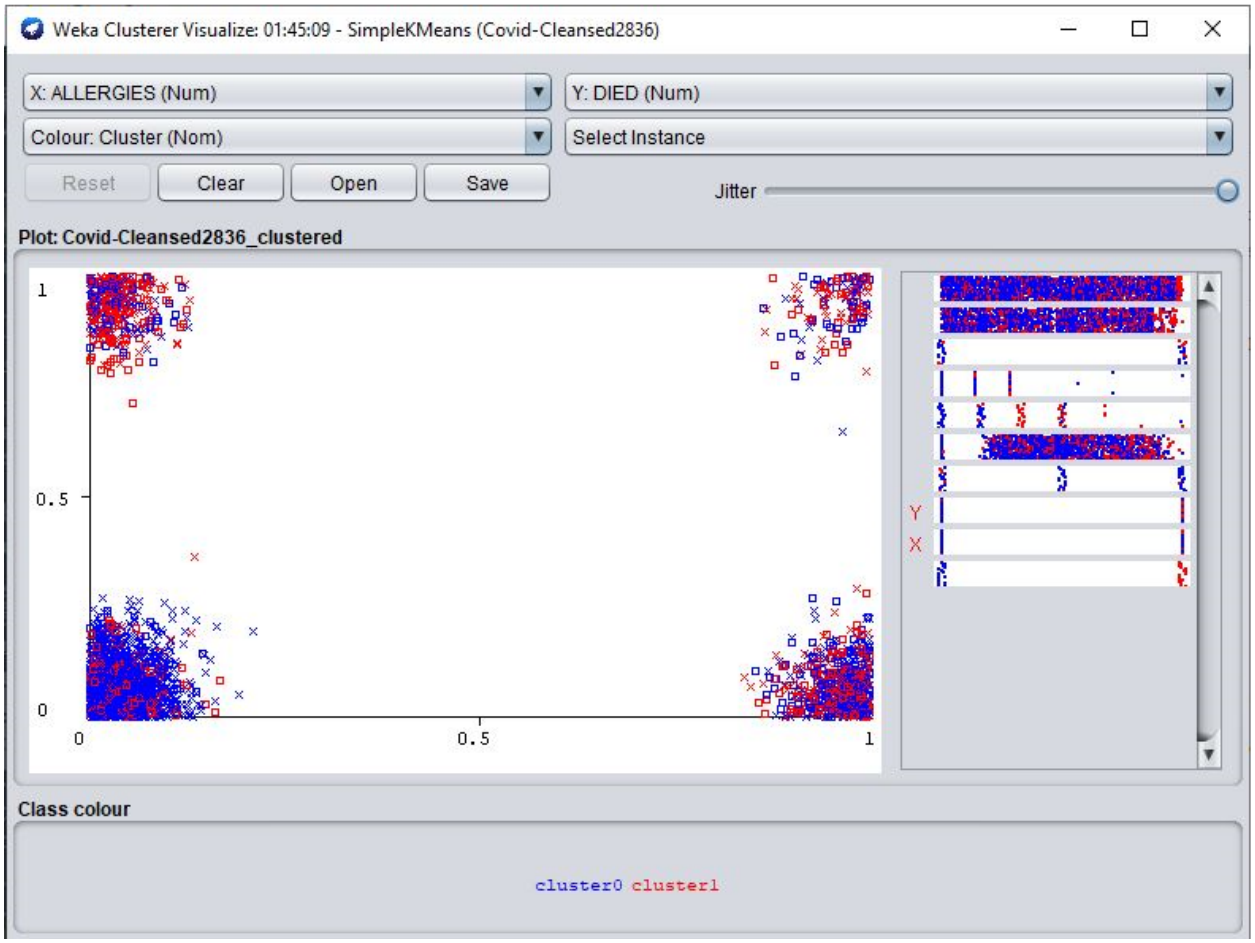


```
Initial starting points (random):

Cluster 0: 1,LA,36,F,0,0
Cluster 1: 1,RA,63,F,0,1

Missing values globally replaced with mean/mode

Final cluster centroids:
                               Cluster#
Attribute           Full Data         0          1
                     (2835.0)  (2047.0)    (788.0)
==================================================
VAX_DOSE_SERIES        0.9058    0.8935     0.9378
VAX_SITE                   LA        LA         RA
AGE                   44.5196   39.4949    57.5724
SEX                         F         F          F
DIED                   0.1552     0.085     0.3376
ALLERGIES              0.2483    0.1798     0.4264




Time taken to build model (full training data) : 0.09 seconds

=== Model and evaluation on training set ===
```

Looking at the cluster visualization of deaths by age where Died is on the Y axis and Age is on the X axis, and 1 is that the patient died, we can see the Moderna deaths jump out in red for those patients age 52.5 and older. It is clear to see that this model predicts that there will be few to no deaths for patients under the age of 52.5 receiving the Pfizer vaccine.



Finally, we were curious about what else Weka could show us through cluster analysis visualization. We ran another visualization to see what correlation the model predicted between deaths and allergies by vaccine. In order to produce a more meaningful illustration we had to increase the jitter to maximum. The visualization that emerged was indeed interesting. Allergies is on the X axis where 0 is had no allergies and 1 is has allergies. Died is on the Y axis where 1 is died and 0 is not died. The model predicts that a relatively even distribution of people with allergies will die no matter which vaccine they choose. Although this sounds grim, the good news is that the model also predicts that regardless of which vaccine they receive, significantly more people with allergies will live than will die.



Weka Clusterer Visualize: 01:45:09 - SimpleKMeans (Covid-Cleansed2836)

X: ALLERGIES (Num)

Y: DIED (Num)

Colour: Cluster (Nom)

Select Instance

Plot: Covid-Cleansed2836_clustered

Class colour

cluster0 cluster1

Summary

The purpose of our project is to provide some insights in the COVID vaccines manufactured by Pfizer and Moderna. We did some analysis from different perspective, such as finding the factors that significantly effecting the death rate in COVID vaccines, finding the correlation between allergies and deaths in regard to different vaccines, and we also split the vaccine candidates into different risk groups using measurement levels ‘Low’, ‘Medium’, and ‘High’. We are trying to predict which risk groups the future vaccinating candidates will fall into so we can provide different treatments and care to them. This will help to save some time and resources and we always have budget constraints in real life. Also, by studying the key factors effecting the death rate in COVID vaccines, we can take cautious to prevent tragedy events happen.

In our project, we spent much time on the data cleansing and data transformation works. The data cleansing is an important process in data analysis to ensure there is no missing or inaccurate records in the dataset. Thus, we are taking extra cares on the data cleansing and data transformation to make sure the processes are correct. We used two data cleansing tools for our project: Tableau Prep Builder and Weka. It makes sense to have different data cleansing steps when we are looking for different insights using one dataset.

In the data mining part, we used three data mining techniques such as Decision Tree, K-mean Clustering, and Logistic Regression in our analysis. Each data mining technique is used to look for different insight, but overall, we are all doing classification analysis. In real life, there is not any best supervised machine learning model when doing analysis. So, the evaluation of the performance of each model is needed in data analysis. In our project, confusion matrixes are provided for each of the models, so we will know the performance of each model when doing data analysis.

The comparison of the accuracy of confusion matrix of each model is shown in table below:

Model	Accuracy
Logistic Regression	0.8395
Decision Tree (RiskLevel)	0.6037
Decision Tree (DeathCount)	0.8523
K-Mean Cluster	0.6448

We observed that the Decision Tree (RiskLevel) model has the lowest accuracy. The reason being is that we split the risk level into three different levels: ‘Low’, ‘Medium’, and ‘High’. Thus, we have a higher chance of misplacing the classification in the model. We can see the higher accuracy rate when using Decision Tree model in a two-level output (0, 1) for Death. Overall, we can conclude that no single model is best in data mining, we will need to evaluate the performance of each model to determine which model is suitable to be used in certain condition.

References

1. News Break (March 25, 2021). Coronavirus Real Time Updates. https://www.newsbreak.com/topics/coronavirus
2. Centers for Disease Control and Prevention (CDC, Match 4, 2021). Pfizer-BioNTech. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines/Pfizer-BioNTech.html
3. Centers for Disease Control and Prevention (CDC, Match 4, 2021). Moderna. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines/Moderna.html
4. Centers for Disease Control and Prevention (CDC, Match 4, 2021). Johnson & Johnson's Janssen. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines/janssen.html

5. Garbade M.J. (September 12, 2018). Understanding K-means clustering in machine learning. Towards data science. https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
6. Shi Y. & Olson D.L. (2005). *Introduction to business data mining*, McGraw-Hill Companies
7. VAERS (November 2020). VAERS data use guide. https://vaers.hhs.gov/docs/VAERSDataUseGuide_November2020.pdf