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 3](#_Toc70966980)

[Research Questions 3](#_Toc70966981)

[Datasets 4](#_Toc70966982)

[Covid Data 4](#_Toc70966983)

[Demographic Data 7](#_Toc70966984)

[Data Mining Techniques Used 10](#_Toc70966985)

[(1) Decision Tree 10](#_Toc70966986)

[(2) K-mean Clustering 10](#_Toc70966987)

[(3) Logistic Regression 10](#_Toc70966988)

[Logistic regression to identify influencing factors for deaths 10](#_Toc70966989)

[Data cleansing 11](#_Toc70966990)

[Data transformations 12](#_Toc70966991)

[Death rate Insights 14](#_Toc70966992)

[Logistic regression 15](#_Toc70966993)

[Decision Tree Solution Explanation 17](#_Toc70966994)

[Data Transformation 17](#_Toc70966995)

[Age -> Age\_Group 17](#_Toc70966996)

[Risk\_Level 18](#_Toc70966997)

[Cluster Analysis Using Weka 23](#_Toc70966998)

[Data cleansing 23](#_Toc70966999)

[Data Constraints and Assumptions 23](#_Toc70967000)

[Weka Insights 23](#_Toc70967001)

[Cluster Analysis 26](#_Toc70967002)

[Management Implementation 26](#_Toc70967003)

[Summary 26](#_Toc70967004)

[Computation or Database Attachment 26](#_Toc70967005)

[References 27](#_Toc70967006)

# 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.

1. 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.

1. 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.

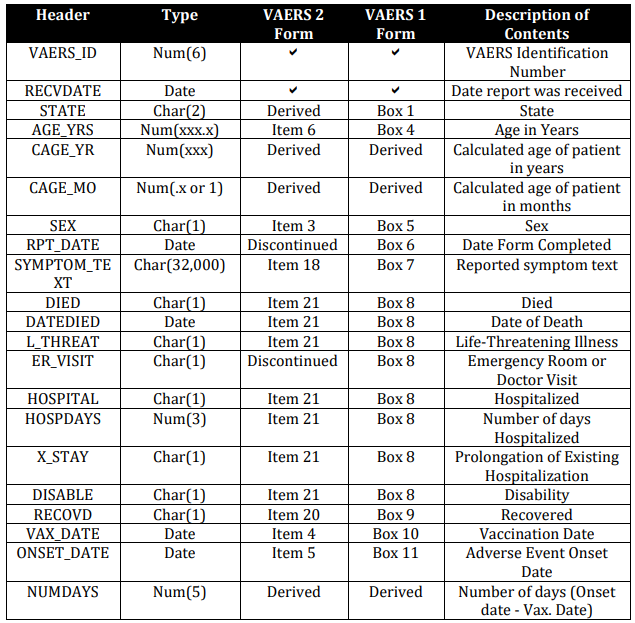


Table 1. Data description (VAERS, November 2020).

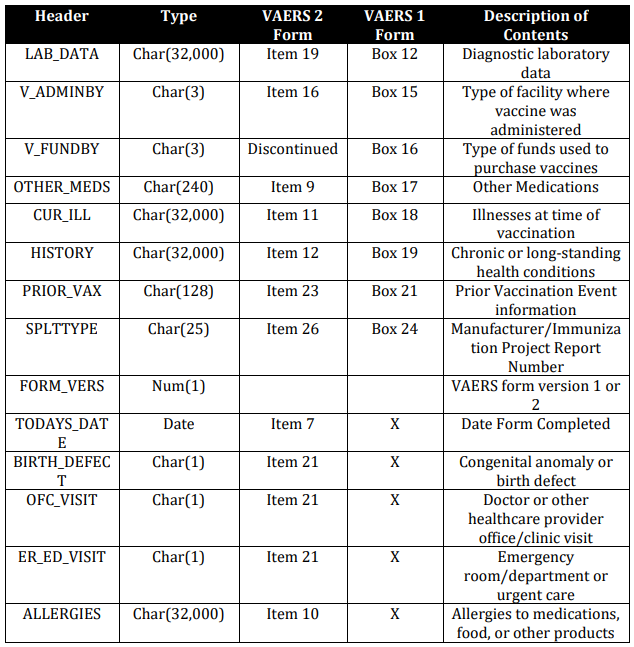


Table 2. Data description (VAERS, November 2020).

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

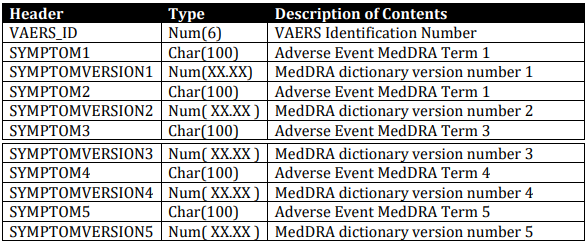


Table 3. Data description (VAERS, November 2020).

(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.

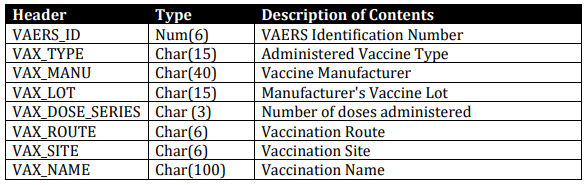


Table 4. Data description (VAERS, November 2020).

## 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.

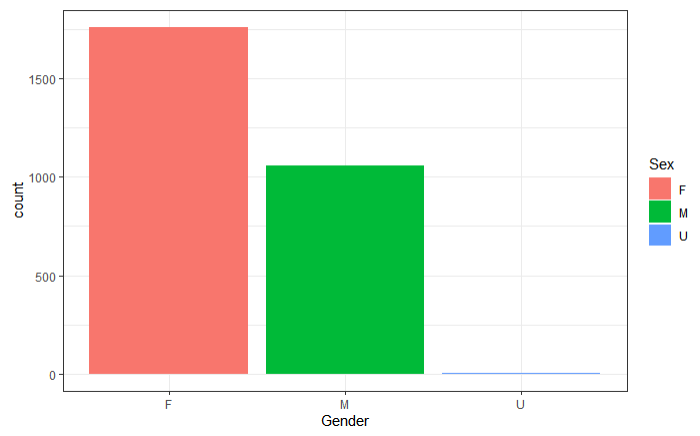


Figure 1. Gender graph

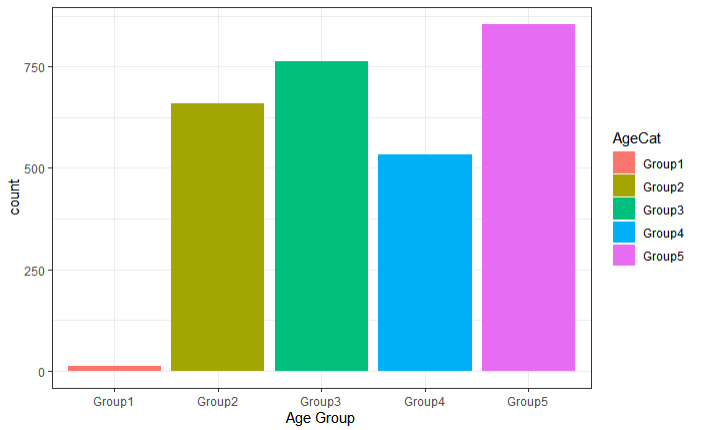


Figure 2. Age group plot

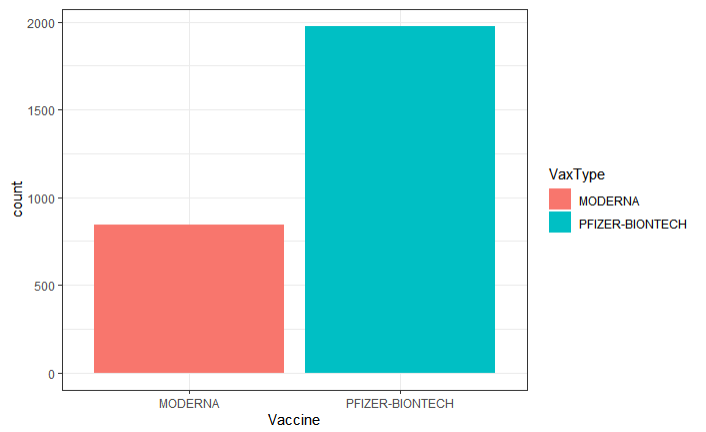


Figure 3. Vaccine type plot indicates the vaccine type received by individuals.

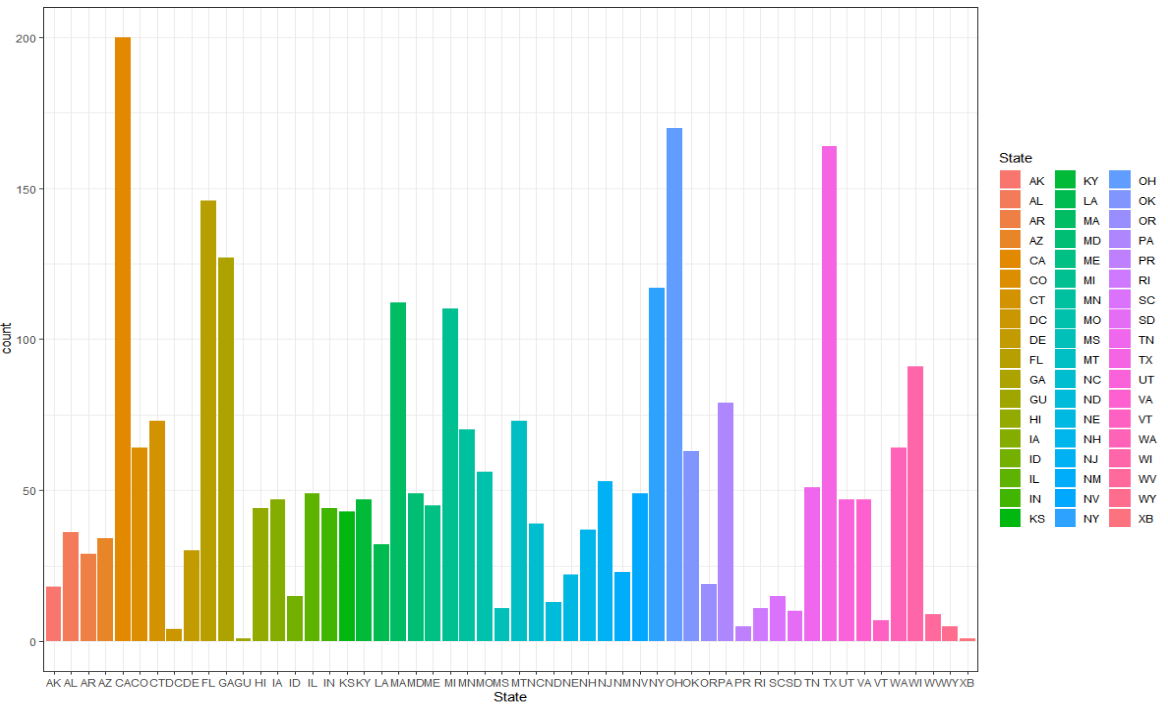


Figure 4. State plot indicates the home state of the vaccinee.

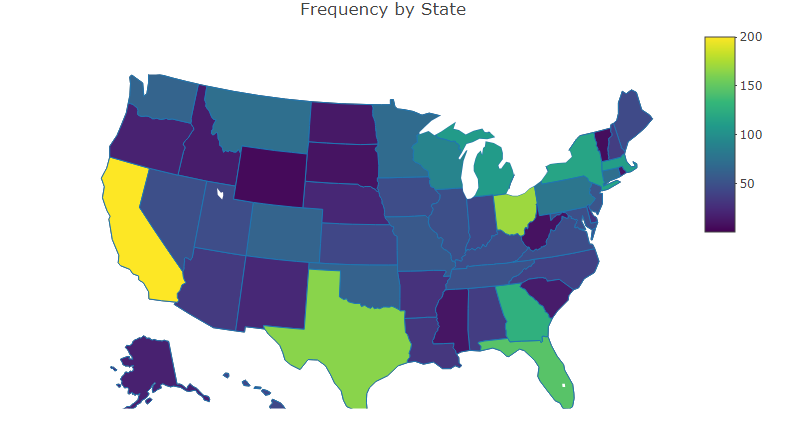


Figure 5. US map plot at state level indicates the top states relate to this study are California, Florida, New York, and Texas.

# 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

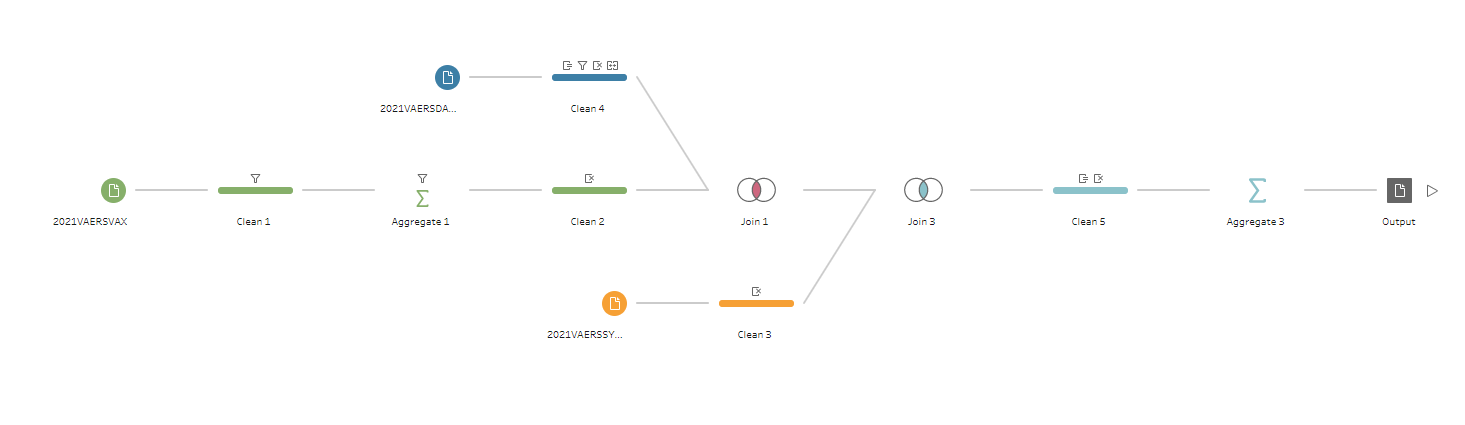
In data mining, Regression is one of the basic tools for analysis, used in classification applications through logistic regression and discriminant analysis, as well as the prediction of continuous data through ordinary least squares (OLS) and other forms (Shi & Olson, 2005). Regression analysis is a 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.

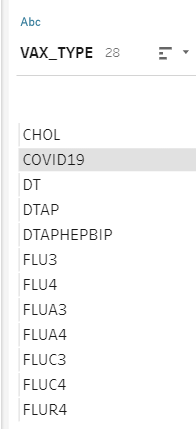
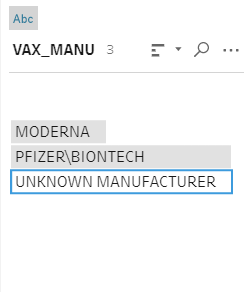
## Data cleansing

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.

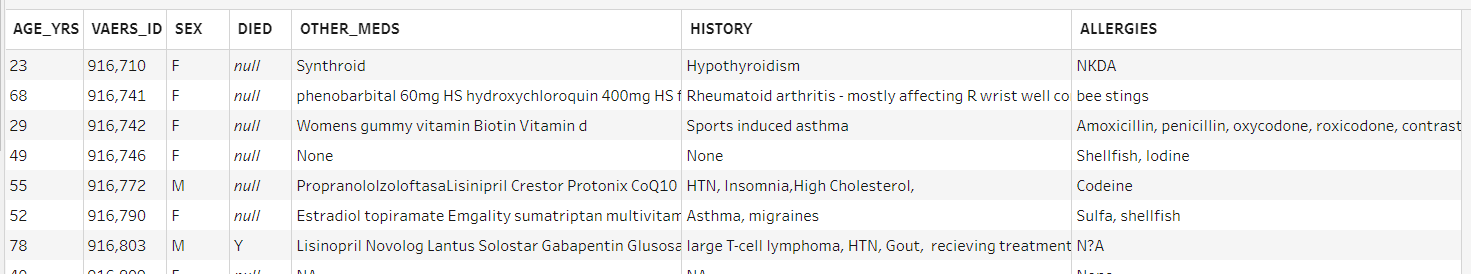
Below shows pictorial view of the cleansing process taken from Tableau prep builder.



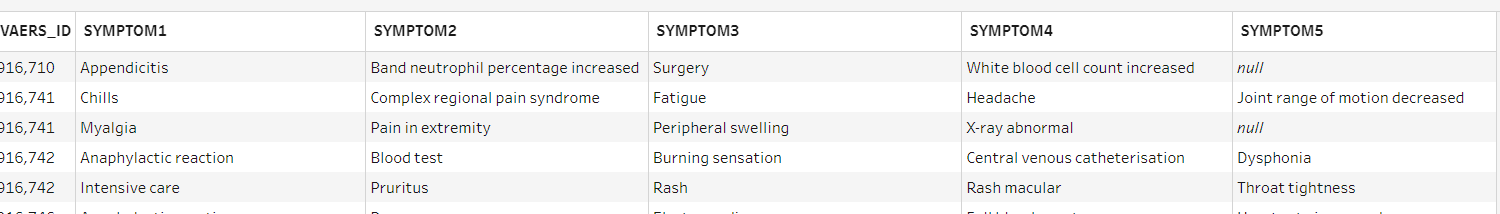
Coming to Initial cleansing process, the dataset 2021VAERSVAX got different vaccines in which we are interested in COVID19. In Covid19, filtered only Pfizer and Moderna. Removed the duplicate rows.

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.



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



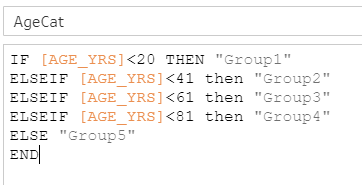
After the initial data cleaning in each dataset, joined 2021VAERSVAX, 2021VAERSDATA, and 2021VAERSSYMPTOMS datasets.

## Data transformations

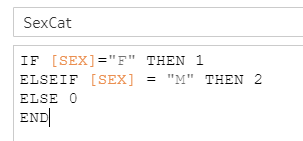
Data transformation is the process of changing the format, structure, or values of data to better organize, improve quality and to facilitate the compatibility of data.

Created categorical data from the existing information for all the columns.

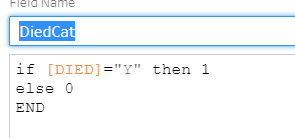
We divided the age into five groups as below.

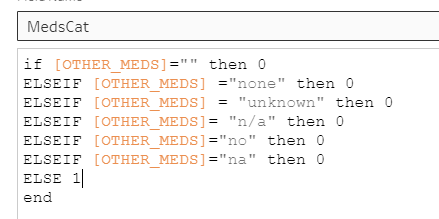


The sex column is divided into three categories as below

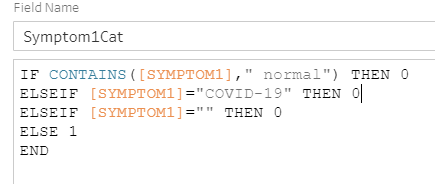


The DIED column is converted into binary column as died 1, otherwise a 0.



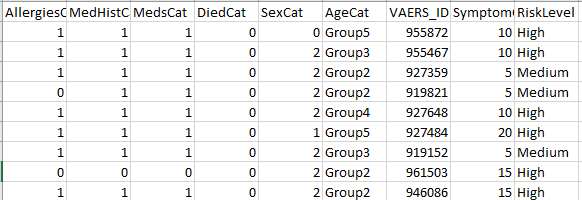
The current medications a patient taking OTHER\_MEDS column is categorized into two with a medication taken as 1, otherwise 0. Same applied for ALLERGIES and HISTORY columns.

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



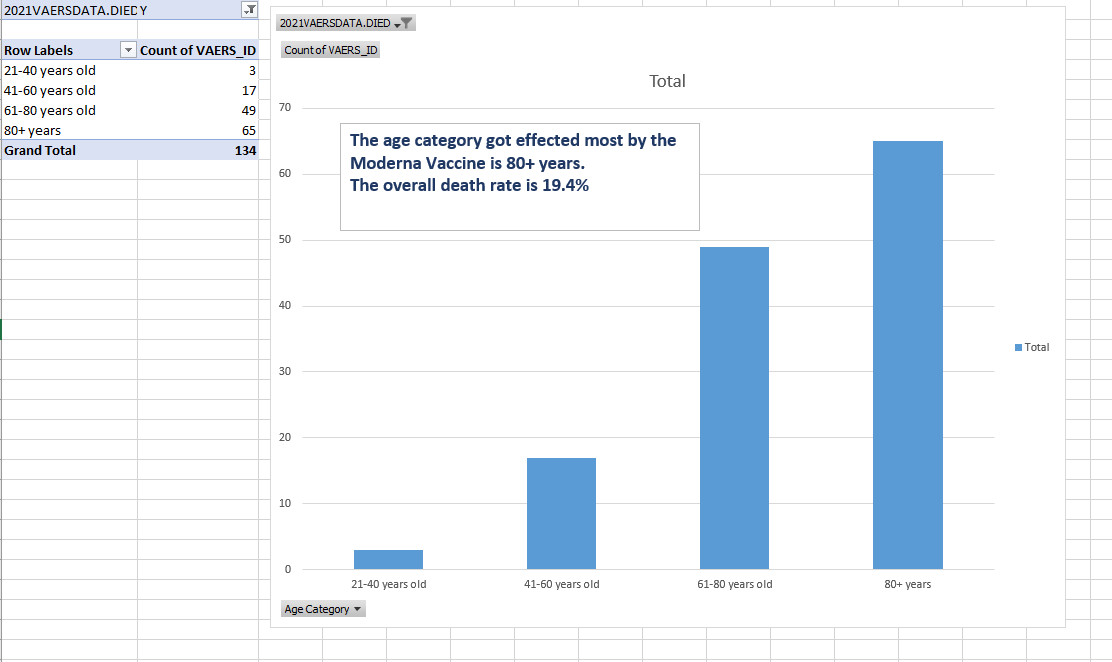
Followed same transformation method for all the other symptoms, Symptom2 to Symptom5. In order to calculate the risk levels, we consolidated the number of symptoms. The data transformation for RiskLevel is explained in decision tree section.

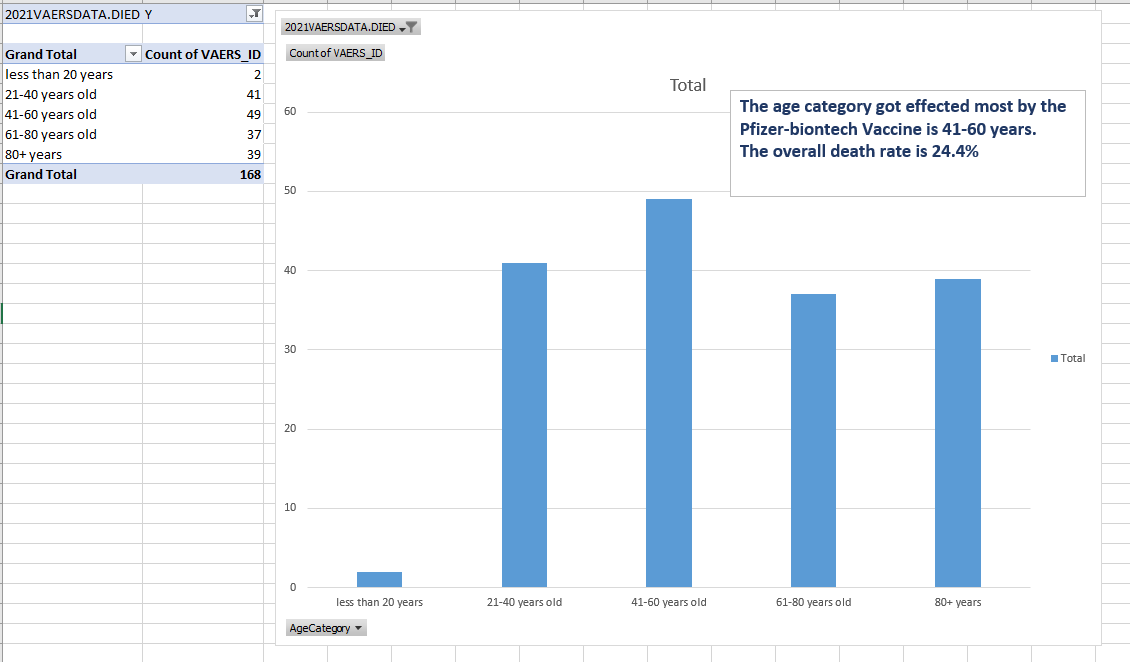
Finally, we have the categorical data which is cleansed and transformed as shown below.



## 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.





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.

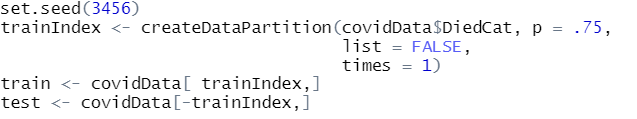
## Logistic regression

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.

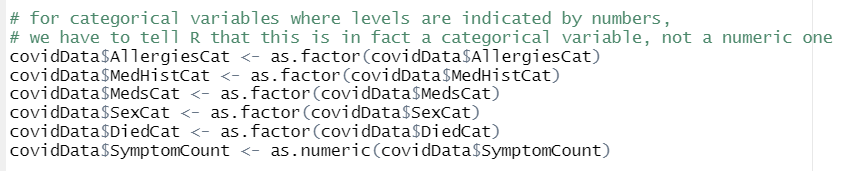
We are using R to run the logistic regression. We used the Generalized Linear Model function glm() to run the logistic regression. The dependent variable is DiedCat. The independent variables are AllergiesCat, MedHistCat, MedsCat, SexCat, SymptomCount.

Steps:

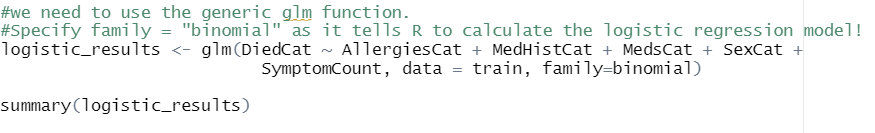
Split the dataset into training set and test set. We are analyzing the regression model with ¾ of data in training set and ¼ of data in test set.



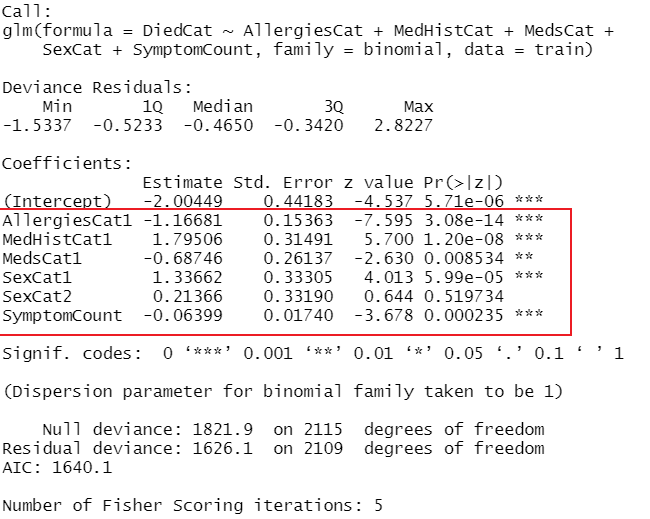
We have to let R know all except SymptomCount are categorical variables.



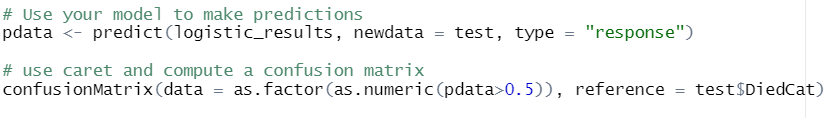
Now we run the logistic regression using glm function. we used the formula notation to specify the variables to the glm function, naming each variable in our formula as below.



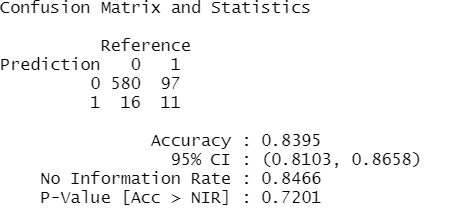
The output shows that all the independent variables show significant effect on death rate in COVID patients.



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



The confusion matrix results shows



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 Risk\_Level, 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.

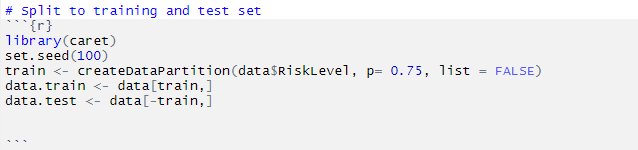
|  |  |
| --- | --- |
| Age | Age Group |
| 0-20 | Group 1 |
| 21-40 | Group 2 |
| 41-60 | Group 3 |
| 61-80 | Group 4 |
| >80 | Group 5 |

### Risk\_Level

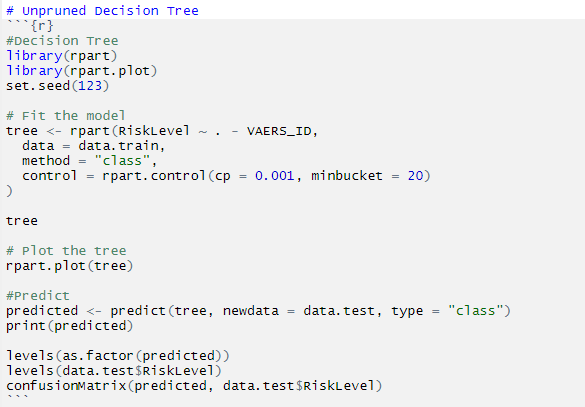
We created a new variable ‘Risk level’ based on the death condition and the total number of symptoms of a particular patient. The new variable ‘Risk level’ contains three levels: ‘Low’, ‘Medium’, and ‘High’. It describes the level of risk of each patient is facing when he/she takes the vaccine. The reason 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:

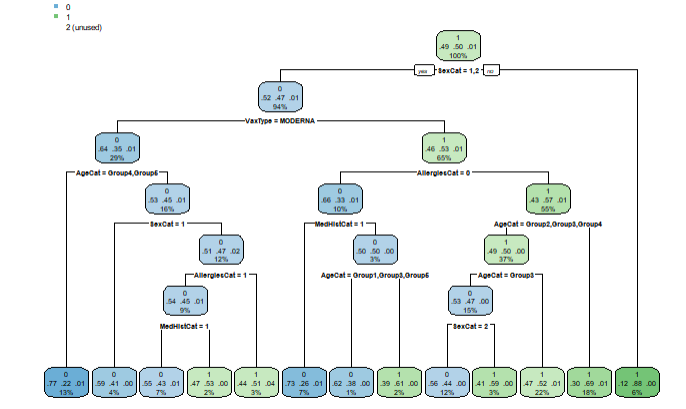
|  |  |  |
| --- | --- | --- |
| Risk Level | Algorithm | Description |
| High | Died = ‘Yes’ or Number of symptoms > 6 | User in this group might have life threatening symptoms. He/she is expected to experience more than 6 symptoms after taking vaccine. Thus, they should be closely monitored. |
| Medium | 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.

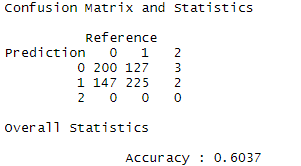


1. After that, we build a classification tree using Allergies, Other\_Meds, History, Sex, Age\_Group, Vax\_Type as input variables, and Risk Level as output. The unpruned decision tree code and output are shown below:

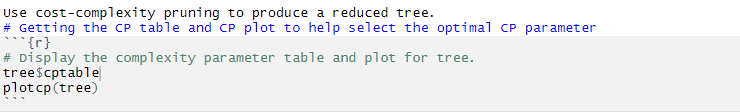


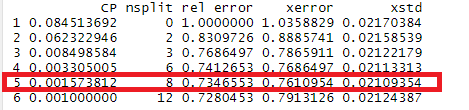


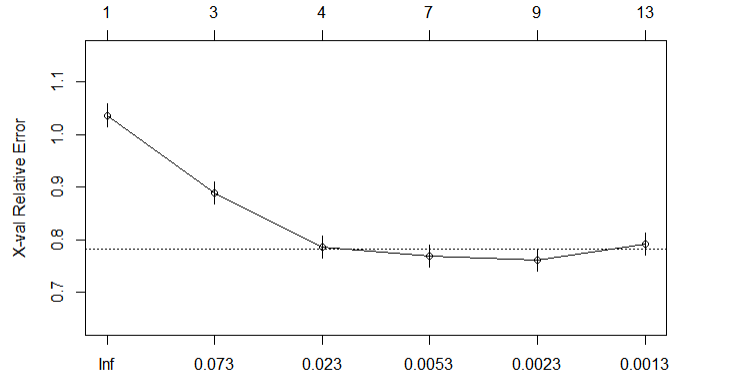
1. The confusion matrix and statistics for unpruned tree is shown below:



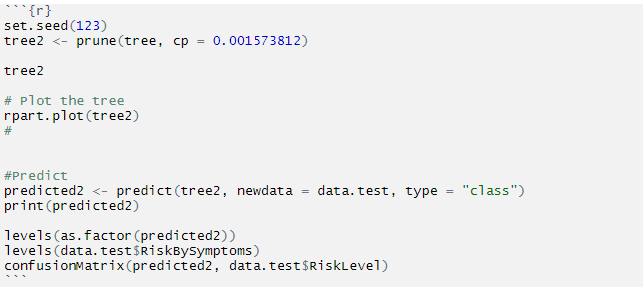
1. Next, we use cost-complexity pruning to see if we can build a simpler tree. The code and output are shown below. We will use the cp value which contributes to least ‘xerror’ to build a pruned tree.

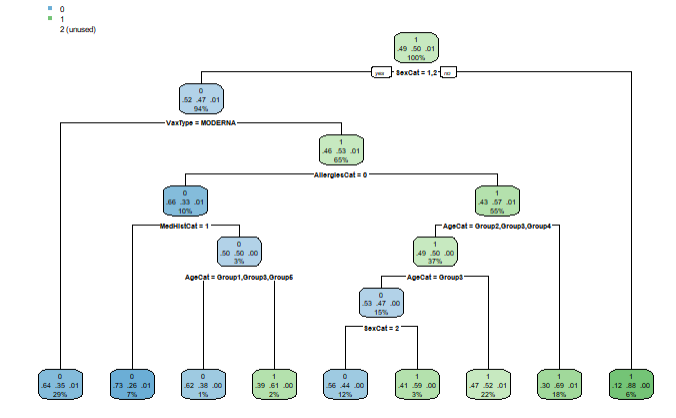




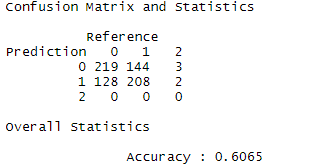


1. We built a pruned tree using cp value 0.001573812. The code and output are shown below:





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



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.

Graphical user interface, text, application

Description automatically generated

1. 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:

Graphical user interface, text, application, email

Description automatically generated

Diagram

Description automatically generated

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

Text

Description automatically generated with low confidence

**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.

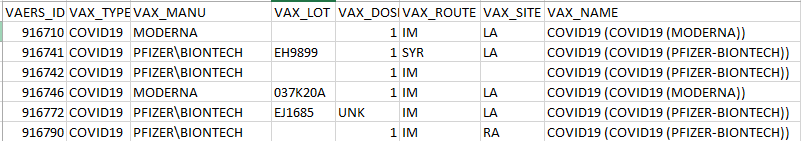
# Cluster Analysis Using Weka

## Data cleansing

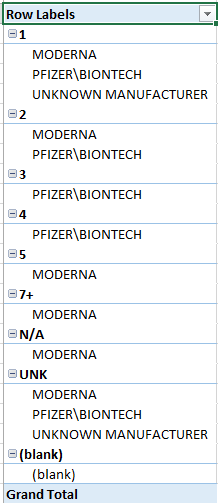
As mentioned in the section above on datasets, the data set 2021VAERSVAX listed all different types of vaccines. As there were 28 different types of vaccines across 3014 rows we had to cleanse the data to only retain the COVID 19 vaccines.

|  |  |
| --- | --- |
| **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** |

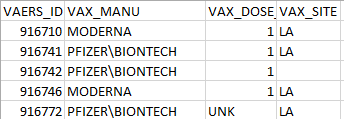
The raw data also consisted of 8 different columns.



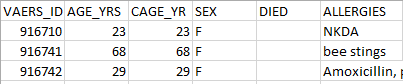
To narrow the scope of analysis as well as to aid the Weka data mining tool, we chose to reduce the number of columns to only keep ones essential to our research as well as those that sparked curiosity within us. One such column was vaccine doses. Going into this research we assumed that people only 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.



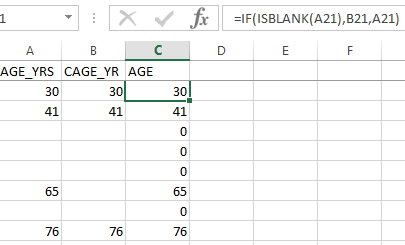
Ultimately, only half of the original columns were kept. The VAERS\_ID was kept so that we could link a vaccine to a patient, VAX\_MANU so that we knew which manufacturer made the vaccine received by the patient, VAX\_DOSE so that we could know (if indicated) how many doses a patient received and VAX\_SITE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



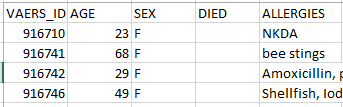
The next data set that we were interested in is the one that included patient data. This file was quite wide with 35 columns. As our team was interested in data mining information related to death, age and vaccine manufacturer, we again trimmed the number of columns to only those we felt essential.



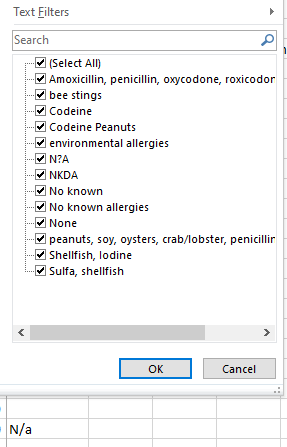
We kept 6 columns. As much work is often needed to cleanse the data more worked needed to be done. Specifically, the age columns were not consistent in that sometimes AGE\_YRS had a value while CAGE\_YR was blank. To solve this inconsistency we merged the columns into 1 single column keeping the value 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.



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

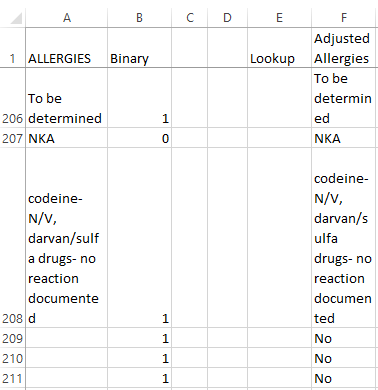


As the data was free form text, we felt it would be best to classify it into a binary format. 0 for no allergies or unknown and 1 for allergies. This was not so simple due to the nature of the data.

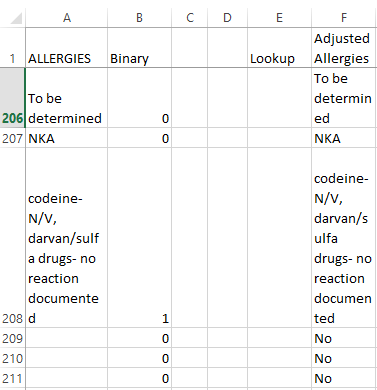


We had to take a step back from the binary approach when it was discovered that there was an issue with blanks. What we had to do was first replace the blanks with “No” which was a value in our lookup list. Then we ran our binary function again on this new list which no longer contained missing values.

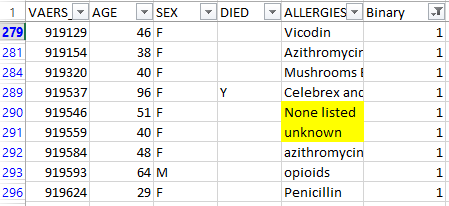
Below shows an example of the issues with the data that I had to solve for:



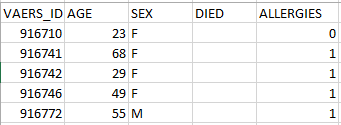
The Binary column for rows 209 to 211 shows a classic Type 2 error where the data is misclassified. The Adjusted Allergies columns shows our formula correcting this. However note that “To be determined” was not accounted for so it is being classified as an allergy. For this analysis exercise the assumption 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.



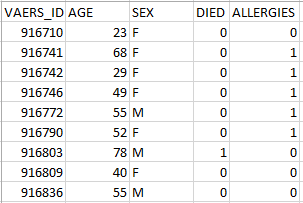
But even this is not enough. After combining the binary classification back to the main data set we had to manually test it as free form text allows for many combinations. We found that we missed some combinations.



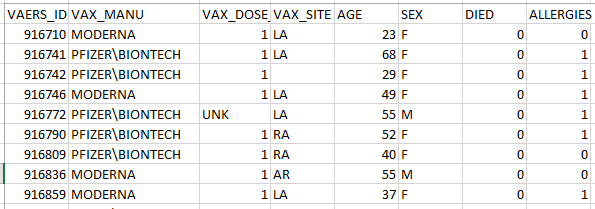
So now, just like we learned in our cluster analysis lecture, we must reiterate the process in order to get the data as clean as possible with the smallest margin for error. Our list of negative responses that started at 14 grew to almost 160. After many iterations we were feeling good about the allergies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



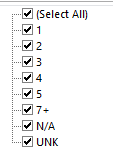
Now we needed to move on to the DIED column. Our data mining tool does not do well with missing data. The good news is that cleaning the DIED column would be very easy. 1 will indicate that the patient died and 0 will indicate that they did not die.



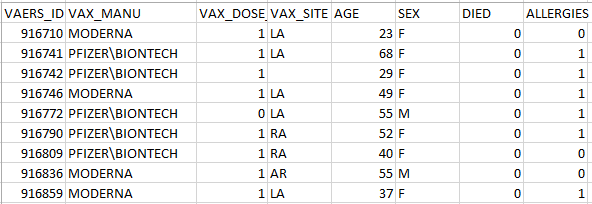
We’ve now merged our 2 datasets into one dataset. The data is looking almost ready to feed into Weka, our data mining tool however as we look at the merged outcome, still we see areas where there are missing values that we need to make some decisions on.



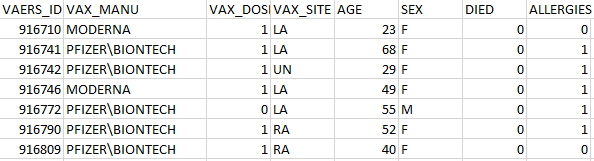
The VAX\_DOSE data types are not uniform and there are missing values in the VAX\_SITE column. We know that Weka is sensitive to these data anomalies. Starting with the list of unique values for VAX\_DOSE we can see 3 data values that will need to change.



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.



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.



## Data Constraints and Assumptions

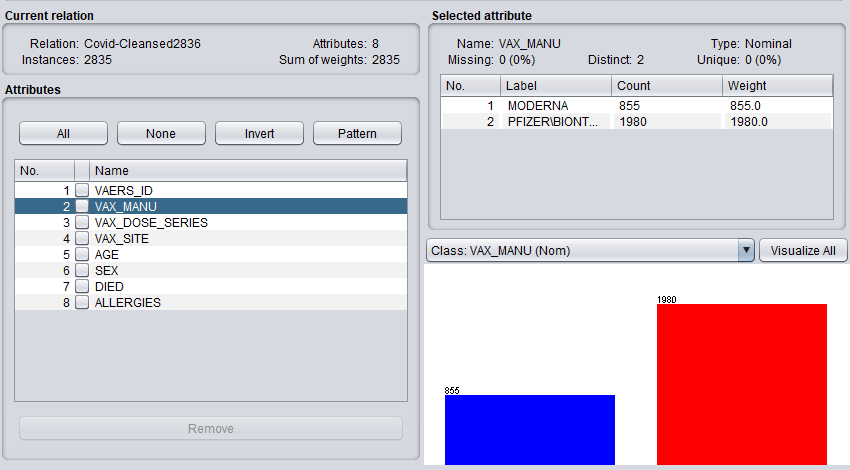
In summary, here are the assumptions and constraints of the data:

* 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

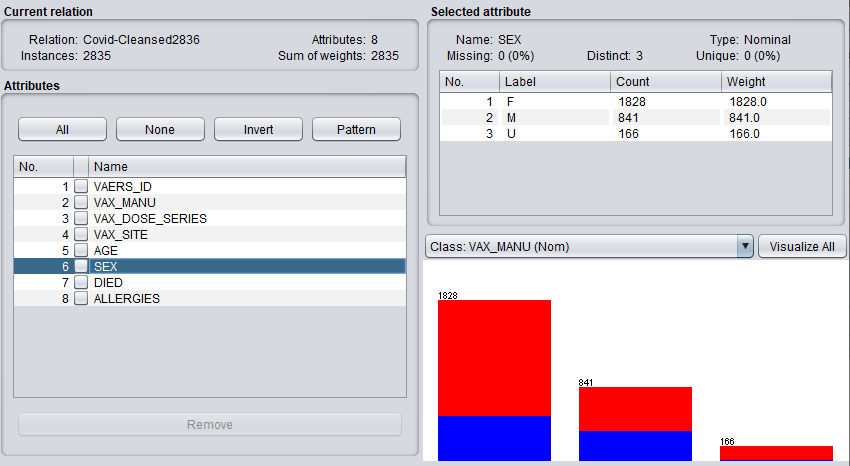
Now, let’s begin to see some insights using Weka!

## Weka Insights

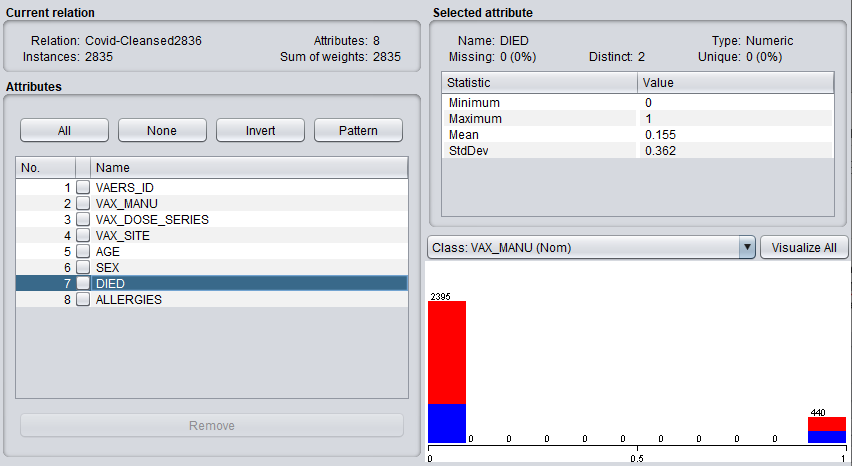
Observations of 2835 instances are confirmed. We are able to easily see that Pfizer has been more widely administered than Moderna. With 1980 doses of Pfizer administered it is coming in at more than double the Moderna vaccines administered at the time our data was captured.



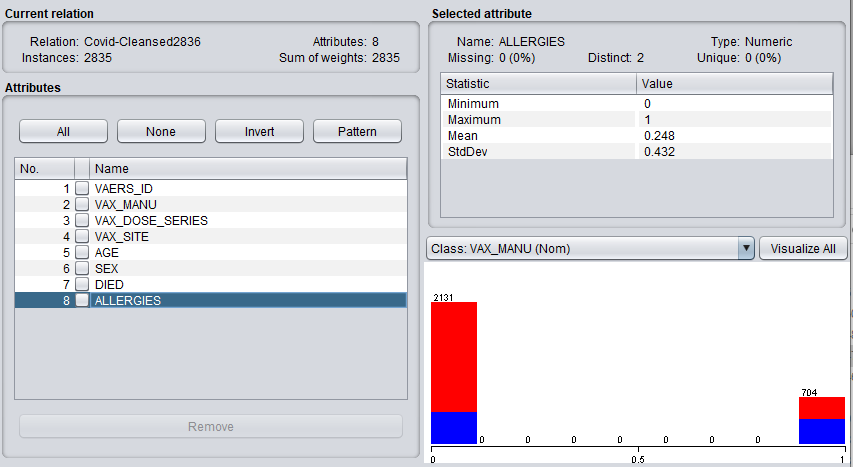
Based on gender, Pfizer has been significantly more prevalent with people classified as Female. Males are almost a 50/50 split between Pfizer and Moderna and people classified as non-binary show significantly more vaccinations with Moderna than Pfizer.



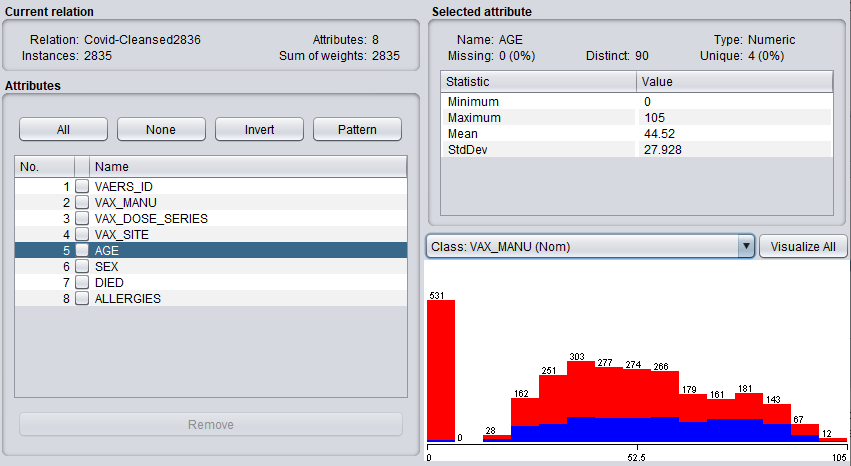
In terms of deaths observed, Pfizer has only slightly more reported deaths than Moderna which is interesting given that over 50% more vaccines were administered by that manufacturer. The bar on the left of the graph shows patients that are still alive. The data shows in favor of survival even with the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



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.

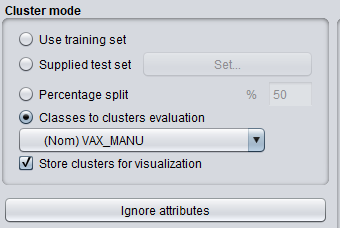


Age shows another interesting story. Given the large number of Pfizer patients with unreported ages, one might deduce that perhaps that information was not deemed important at the start of the vaccines being rolled out. However over time as people were wanting to do more analysis, it was determined that 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.

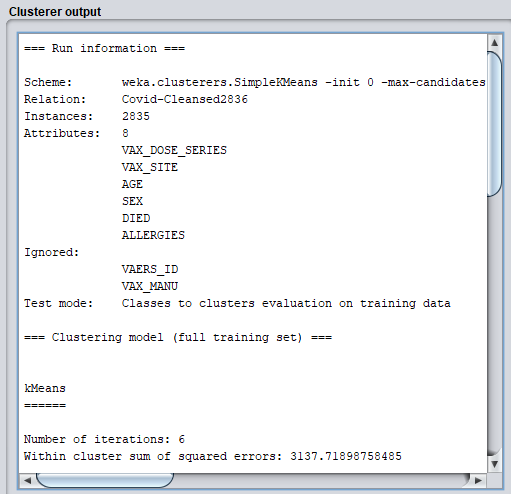


## Cluster Analysis

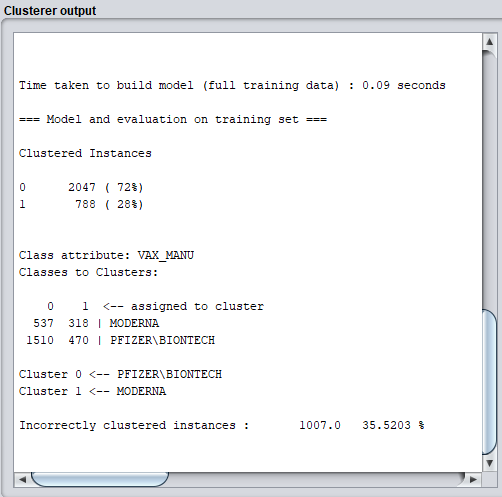
We ran a cluster analysis using a simple K means method with Weka and chose 2 clusters evaluated by vaccine manufacturer. We ignored the VAERSID as there was no need to individually identify patients.



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

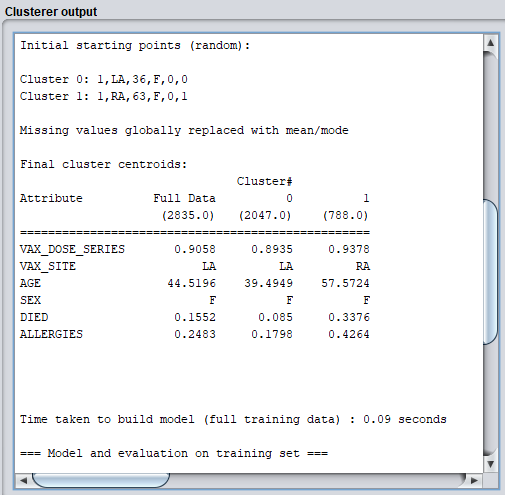


The model calculates 35% cluster errors based on 6 iterations through the training data. It has assigned Cluster 0 to Pfizer and Cluster 1 to Moderna.

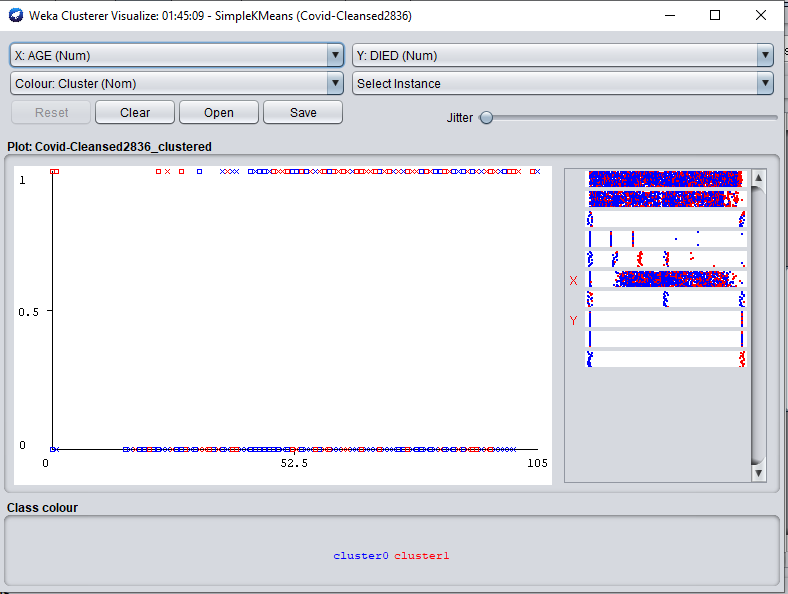


Below are the final cluster centroids. Due to our diligence cleansing our data, there were no missing values so nothing had to be replaced with the mean/mode. Weka is remarkable fast when it comes to running data mining algorithms. We can see that running 6 iterations across 6 attributes only took Weka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

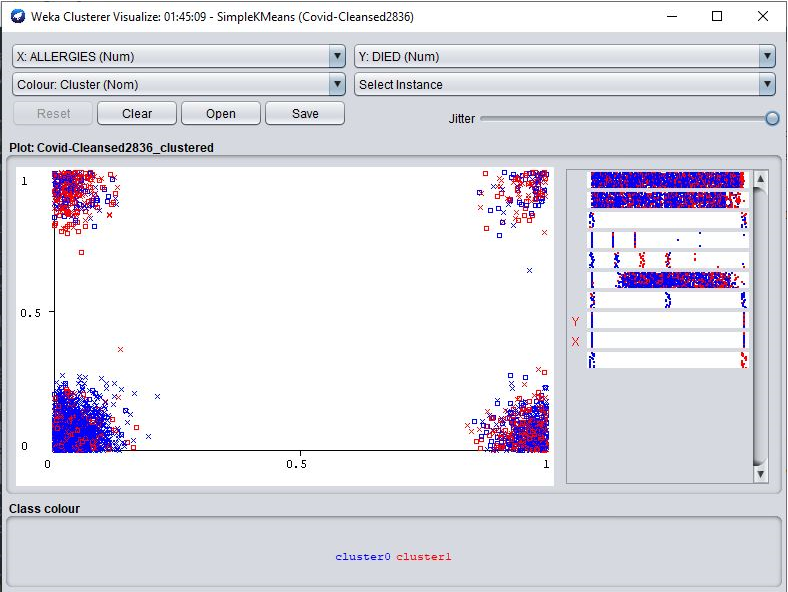
Based on the final cluster centroids the model predicts that more Moderna patients will choose to receive their vaccine in their right arm while Pfizer patients will tend to favor their left arm.



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.



# Management Implementation

# Summary

# Computation or Database Attachment

# 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>