CCT College Dublin

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Strategic Thinking |
| **Assessment Title:** | Report 2 |
| **Lecturer Name:** | James Garza |
| **Student Full Name:** |  |
| **Student Number:** |  |
|  |  |
|  |  |
|  |  |
|  |  |
| **Assessment Due Date:** |  |
| **Date of Submission:** |  |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

## Introduction Error! Bookmark not defined.

Business description 3

Hypothesis Error! Bookmark not defined.

[General goal 3](#_TOC_250019)

[Success criteria/indicators 4](#_TOC_250018)

[Technologies used 4](#_TOC_250017)

[Models and machine learning algorithms 4](#_TOC_250016)

[Libraries 4](#_TOC_250015)

Accomplishment 4

Data 4

[source 4](#_TOC_250014)

[Attributes 4](#_TOC_250013)

[Dimensions 5](#_TOC_250012)

[Descriptive statistics and Data visualization 5](#_TOC_250011)

[Correlation: Pair plot 11](#_TOC_250010)

[Data preparation and pre-processing 12](#_TOC_250009)

[Flowchart of Data Preparation and Modelling 13](#_TOC_250008)

[Feature engineering 15](#_TOC_250007)

[Models 15](#_TOC_250006)

[Challenges encountered 15](#_TOC_250005)

[Inclusion of strategies to overcome them 15](#_TOC_250004)

[Results and analysis 16](#_TOC_250003)

[Step 1 - Defining the X and y 16](#_TOC_250002)

[Conclusion 177](#_TOC_250001)

[References 188](#_TOC_250000)

## Introduction

We were tasked with finding a topic that interested us as a team, from there we were to create a business model and come up with a hypothesis or with questions, to which we would need to find the answers. After a couple meetings we realised that Covid was an interesting subject to us all. We also knew that there was a lot of data taken regarding the pandemic so surely there would be some good datasets with which we could do a study and make predictions. We came up with the idea that there might be some work we could do predicting people who have bad reactions to a vaccine. With a little work we were able to find the VAERS dataset from Kaggle.

Vaccine Adverse Event Reporting System was created by the Food and Drug Administration (FDA) and the US Centres for Disease Control and Prevention (CDC) to receive reports of potentially vaccine-related adverse events (Gee et al., 2021). Vaccines protect many people from dangerous illnesses, but can also cause side effects, a small percentage of which may be serious (Classen, 2021). VAERS is used to continually monitor reports to determine whether any vaccine or vaccine lot has a higher than expected rate of adverse events (Moro, Haber and McNeil, 2019).

The Vaers dataset is a large dataset with 890,836 rows, 52 Columns made up of 12 numerical and 40 categorical features. With such a large dataset it seemed like there was a lot of information we could get from the dataset. We continued with the project using the CRISP DM methodology. As you will see some things changed from the first stages of our Business understanding. This is not unexpected as part of the project process is to re-evaluate and re-access the situation and if necessary come up with new ideas and ways of achieving our goals.

# Business understanding

We are trying to resolve if we can predict the number of hospital days (if any) a person may require if they have had the Covid 19 vaccine, report an adverse effect and require hospitalisation. This is important as the lockdowns occurred because of the impact on the hospital system in each country. This will make this a regression problem.

# Data Understanding of Raw Data

This section is about understand our dataset. We will read in the dataset and then filter to California which is the state we will be analysing. We will then reset the index and drop the newly created index feature and use .head() to look at our Data Frame which will have 86,381 observations and 52 features.

## Business description Hypothesis

We are trying to predict the adverse effects of the covid vaccines and predict the number of hospital days a person may require if they contract Covid 19 and require hospitalisation.

## General goal

The main goal of this analysis is to check the adverse effects of different Covid19 vaccinations using the Vears database. We try to resolve the regression, classification and predictions problems in this paper. This may make predictions like what kind adverse effect people may have using different vaccines. The aim is to see which vaccine have shown higher incidents of adverse effects.

## Success criteria/indicators

We have found the correlation between vaccine type and symptoms in people having adverse effects after the vaccination. We have used two different machine learning models and algorithms to find the best predictions and results with higher accuracy. First, we have used linear regression, but it was showing lower accuracy. So, we have also used random forest algorithm in order to predict the outcome. We have achieved 99.8 percent of accuracy through random forest classification.

## Technologies used

## Models and machine learning algorithms

We have used two different supervised machine learning models, Random Forest and Linear regression, that may often use in regression and classification problems.

## Libraries

Different libraries have been used to perform different tasks and modeling of algorithms. These may include: Pandas, Numpy, Seaborn, Matplotlib, scipy, missingno, etc.

## Accomplishment Data

The data for Covid19 Vaccination risk in US have been used for the analysis, having 52 variables (or columns) and 890,836 observations (or rows). There were 12 numerical variables and 40 were categorical variables. Some of the included variables may have symptoms, vaccination name, vaccination doses, age, sex, mortality rate, hospitalized days, previous medical history, and allergies, etc. For symptoms, we have five variables of symptoms such as symptom 1, symptom 2, and symptom 3, symptom 4, and symptom 5 as well.

## Source

The data has been taken from an online source that is Kaggle. Kaggle link needs to be provided and referenced!

## Attributes

Attributes are the variables in the machine learning model that may be used as a predictor (Khanal et al., 2018). In this paper, the main attributes include symptoms of patients, vaccination name, and days spent in hospital after contacting covid19.

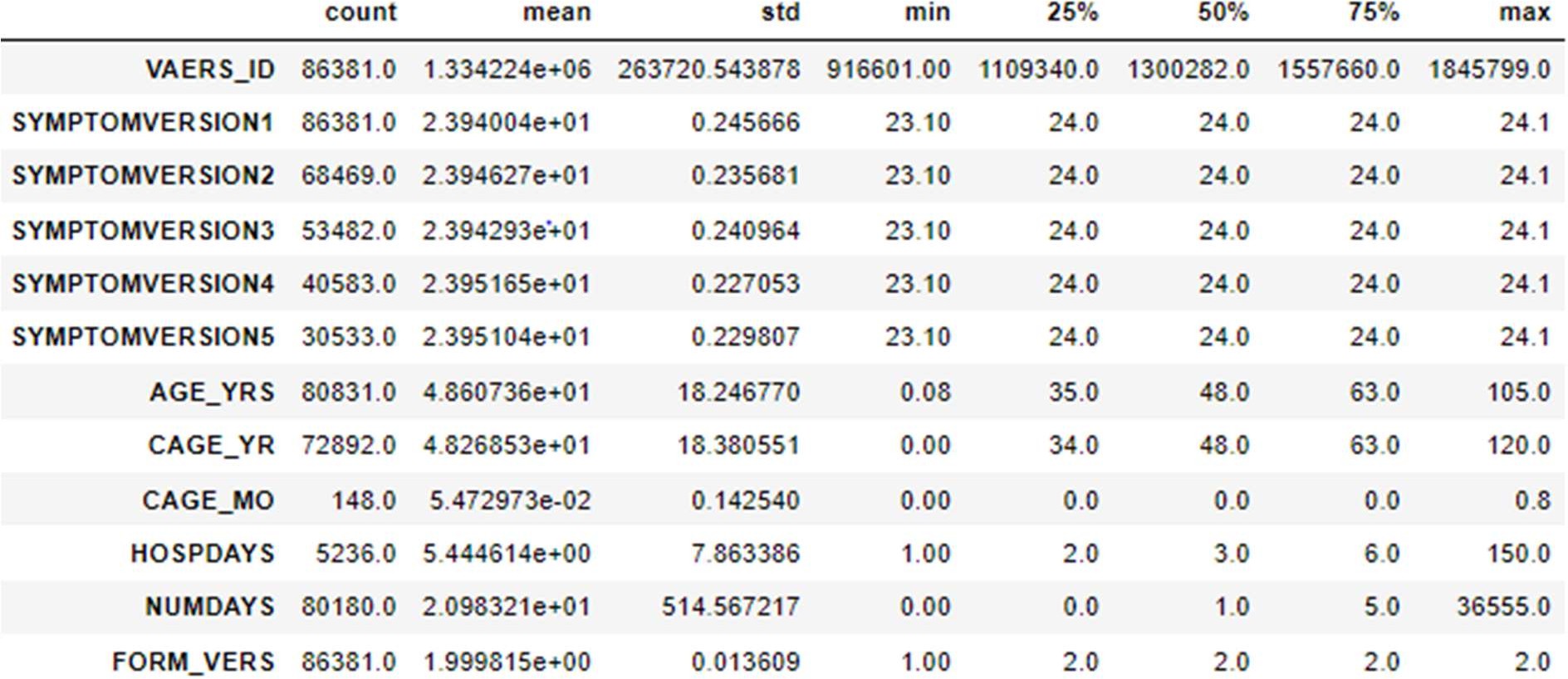
## Dimensions

It includes 52 columns of variable and 86,381 rows as observations.

## Descriptive statistics and Data visualization

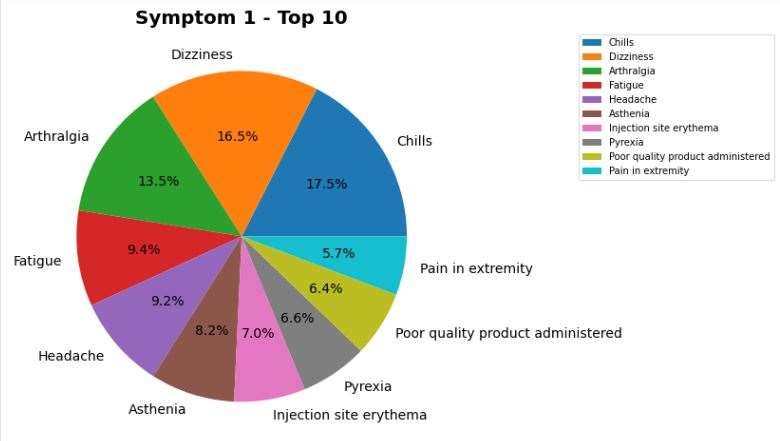
In the descriptive statistics, we have gone through the overview of our dataset using head or simple description codes. The following results showed the statistics of numerical features.

Table1: Results of the statistics of numerical features



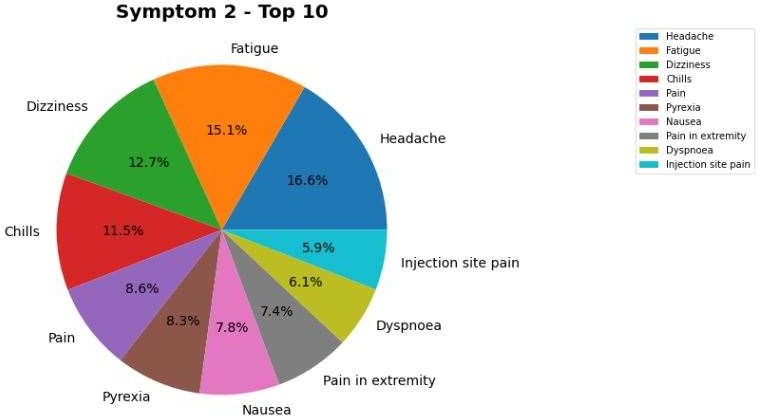
We have done exploratory data analysis in this step to explore the different features of variables in our dataset. There was a data of symptoms in patients showed in five different variables.

*Figure 1: Pie chart for 1st symptom after vaccination*



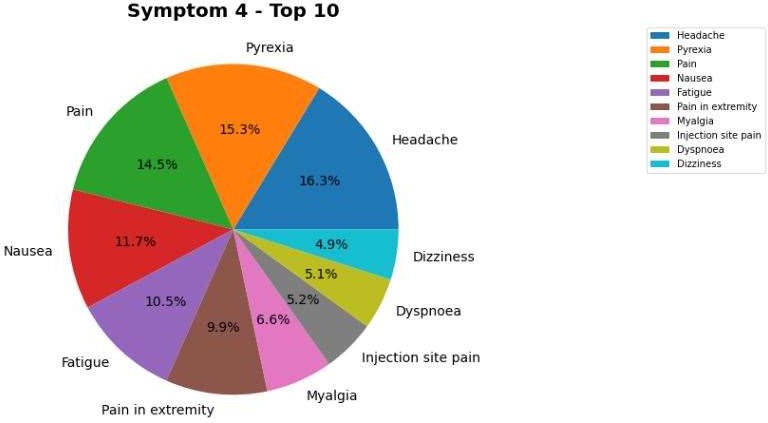
This picture shows the graph for different symptoms in patients after getting the vaccination. Dizziness and chills show the highest frequencies of symptoms.

*Figure 2: Pie chart for 2nd symptom after vaccination*



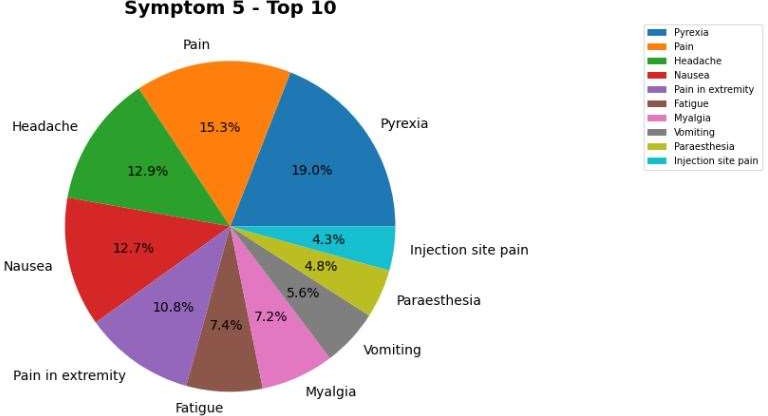
This table also shows the symptoms in patients after the vaccination. This variable showed the high percentages of fatigue, headache, and dizziness as the symptoms of patients.

*Figure 3: Pie chart for 3rd symptom after vaccination.*



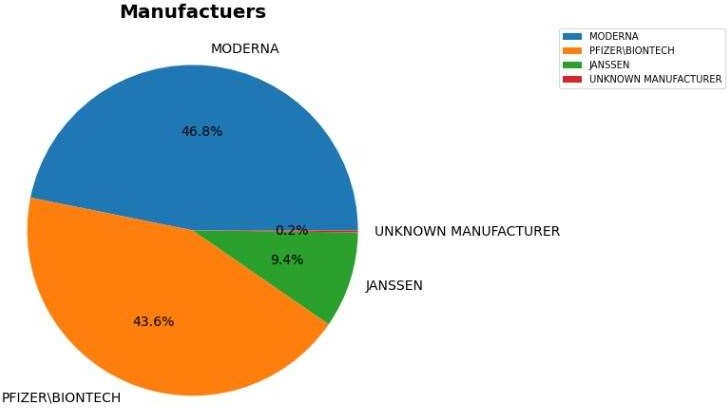
This chart also shows the symptoms in patients after the vaccination. While this also has showed pyrexia and headache as well as the symptoms Covid19 contacted patient in vaccinated people.

*Figure 4. Pie chart for 4th symptom after vaccination*



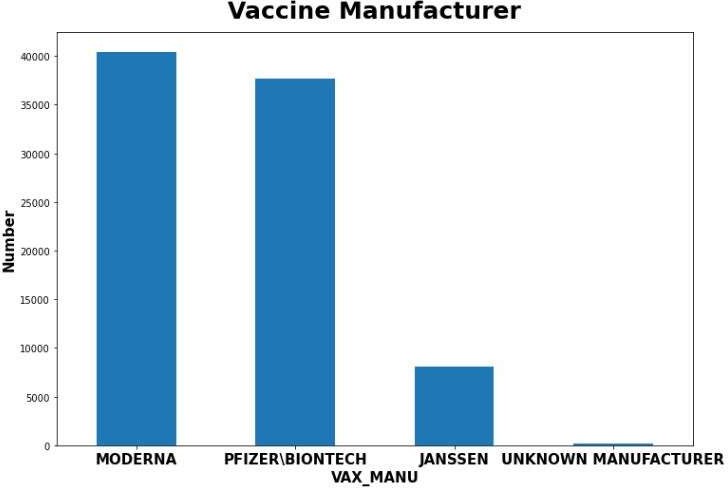
This table also shows the symptoms in patients after the vaccination. These also showed similar results as they have pain, headache, and pyrexia in high percentages.

*Figure 5: Share of vaccines produced by different manufacturers.*



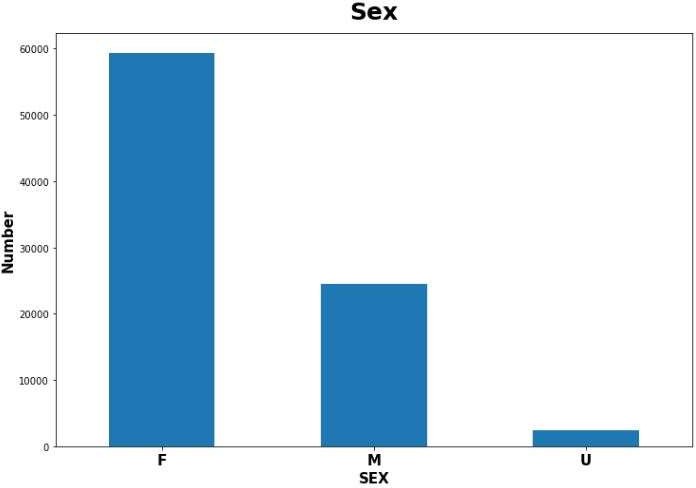
This pie chart shows the share of vaccines produced by different manufacturers. We can see that more than forty percent shares of manufacturing, was produced by Moderna and Pfizer as 46.8 % and 43.6% respectively.

*Graph 1: The numerical numbers of vaccines produced by different manufacturers.*



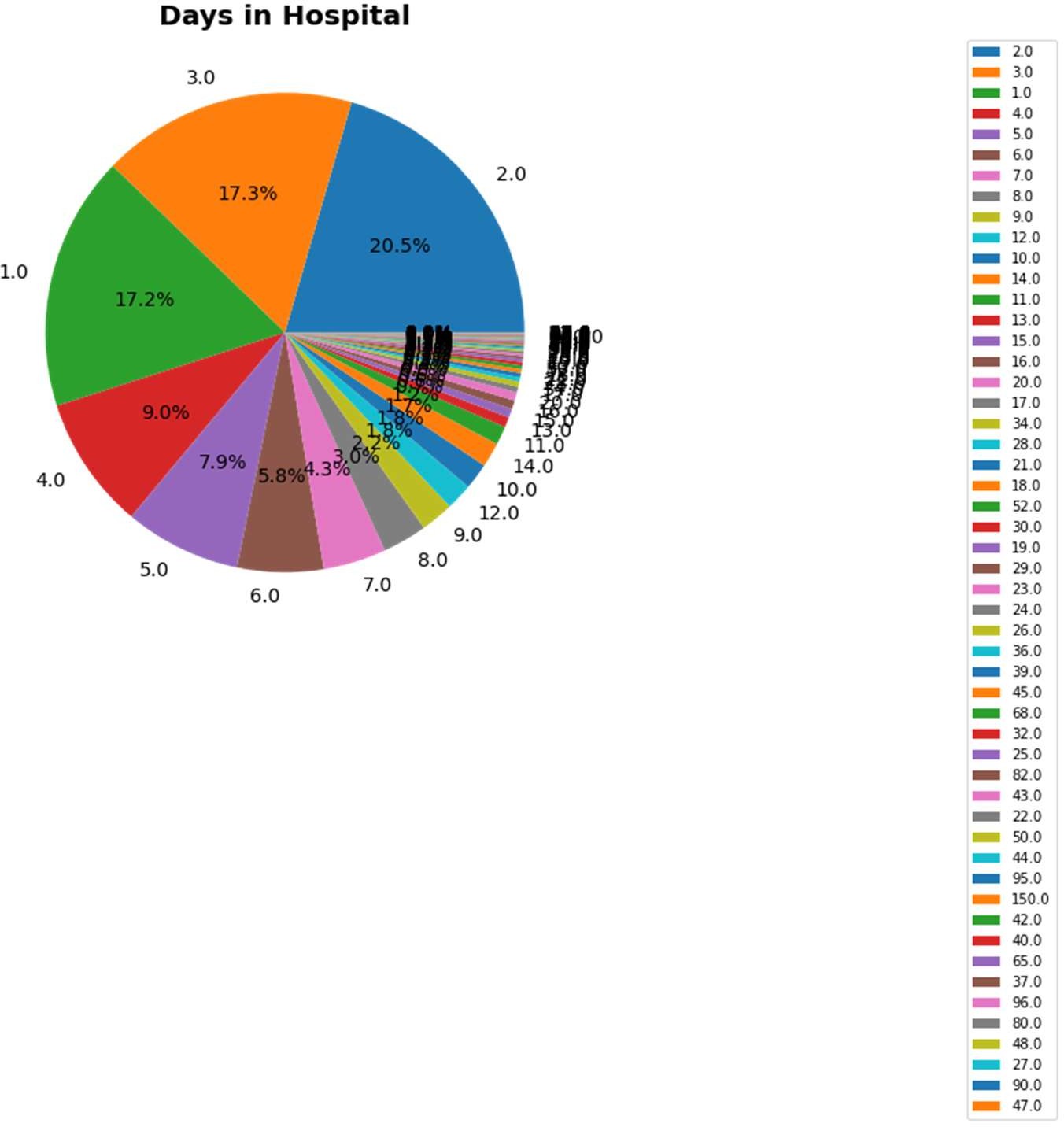
This table also shows the table showing the numerical numbers of vaccines produced by different manufacturers.

*Graph 2: The ratio of patients with different sex included in the dataset*



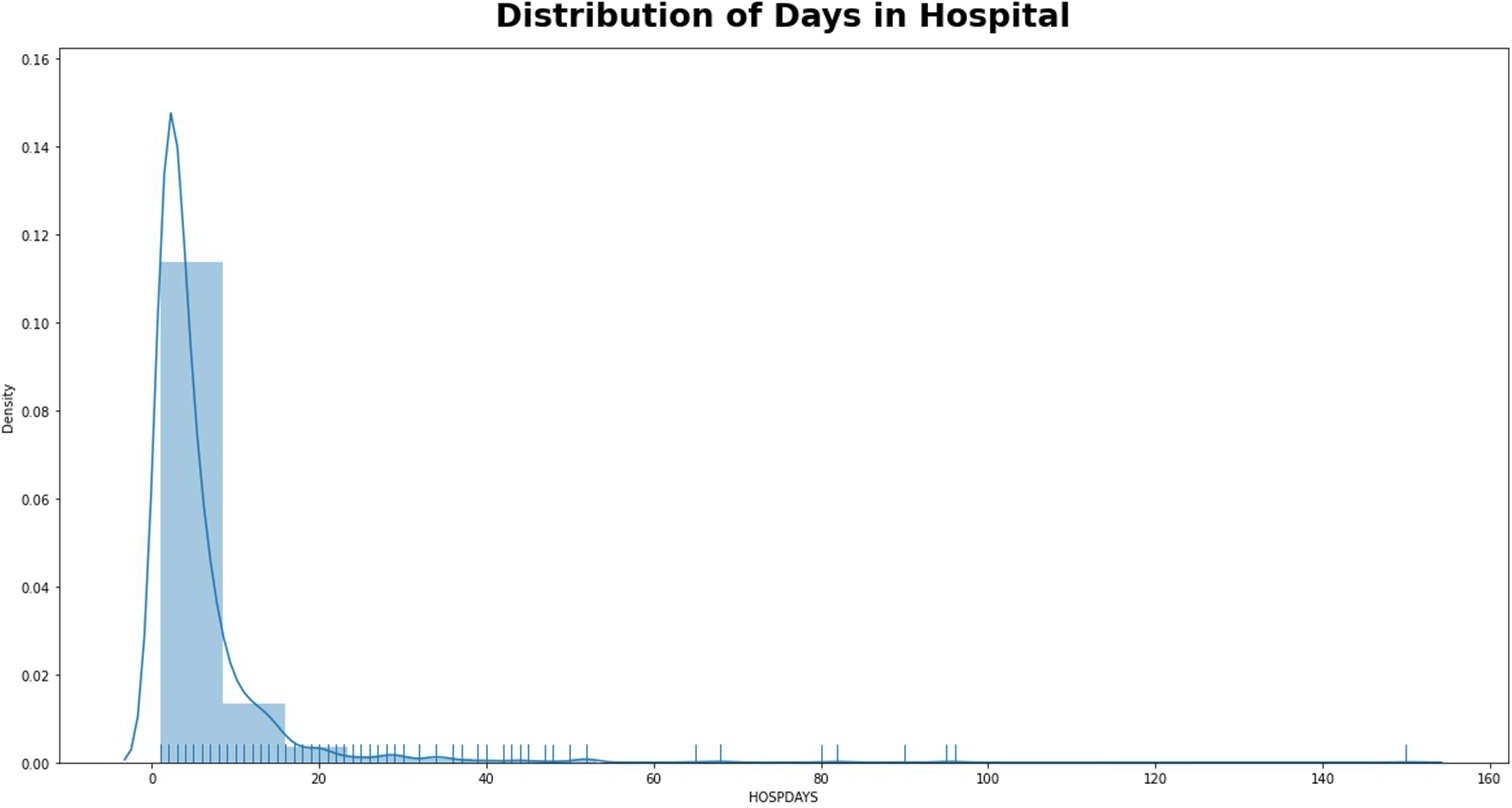
While this table shows the ratio of patients with different sex included in the dataset. This has showed that highest number of females were present in our dataset concluded as females may have contacted more to the covid19 after vaccination.

*Figure 6: The total number of days, the Covid19 patients may have to spend in hospitals*



This pie chart shows the total number of days patients may have to spend in hospitals. A Large share of the pie chart shows that the majority of data represents there will be one, two, or either three days of hospitalization.

*Graph 3: Distribution of days in hospital*

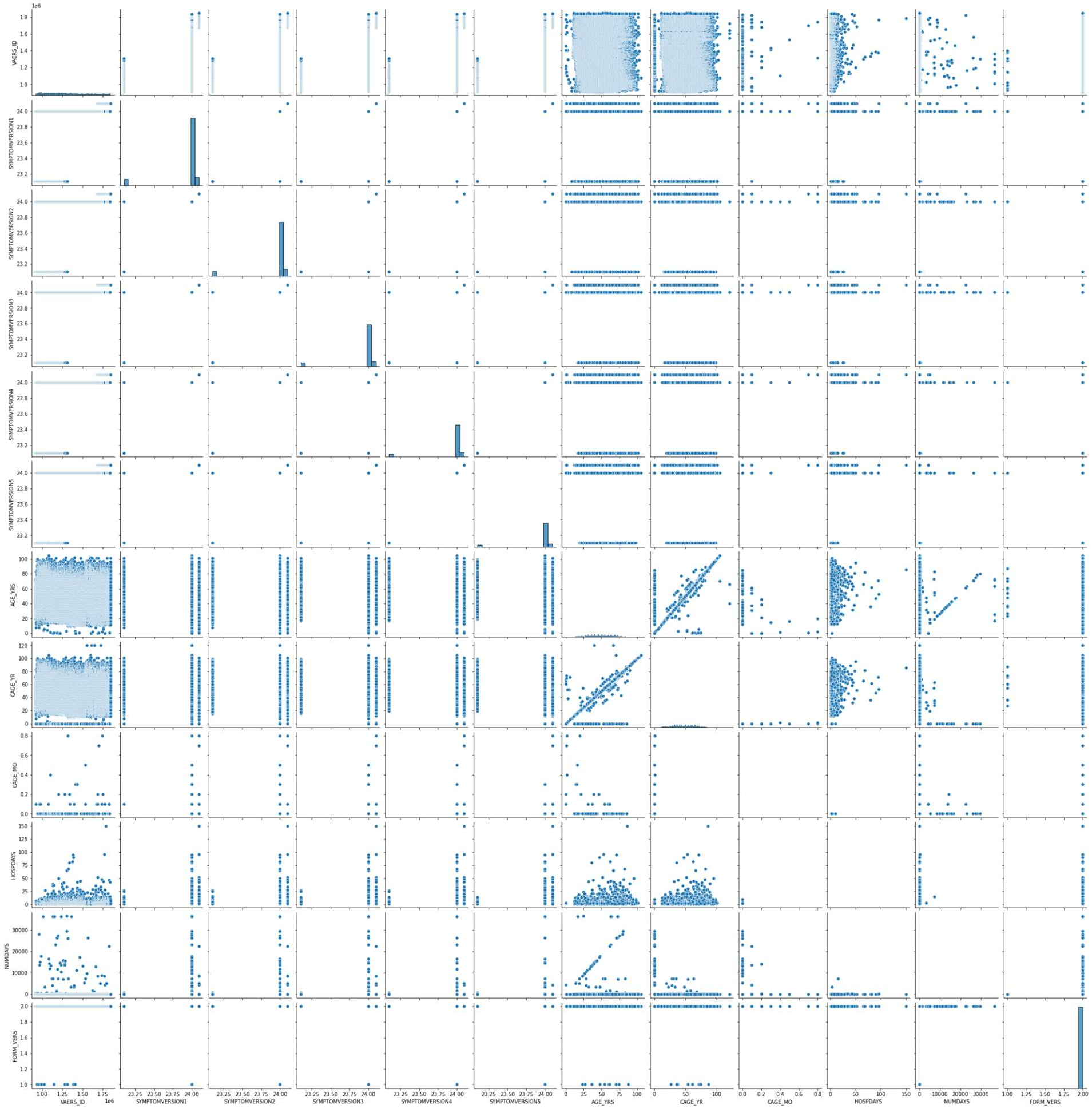


While this chart shows the similar results (patients may have to spend between five to ten days in the hospitals), we have shown the results from pair plot and heat maps to see the correlation between the featured variables.

## Correlation: Pair plot

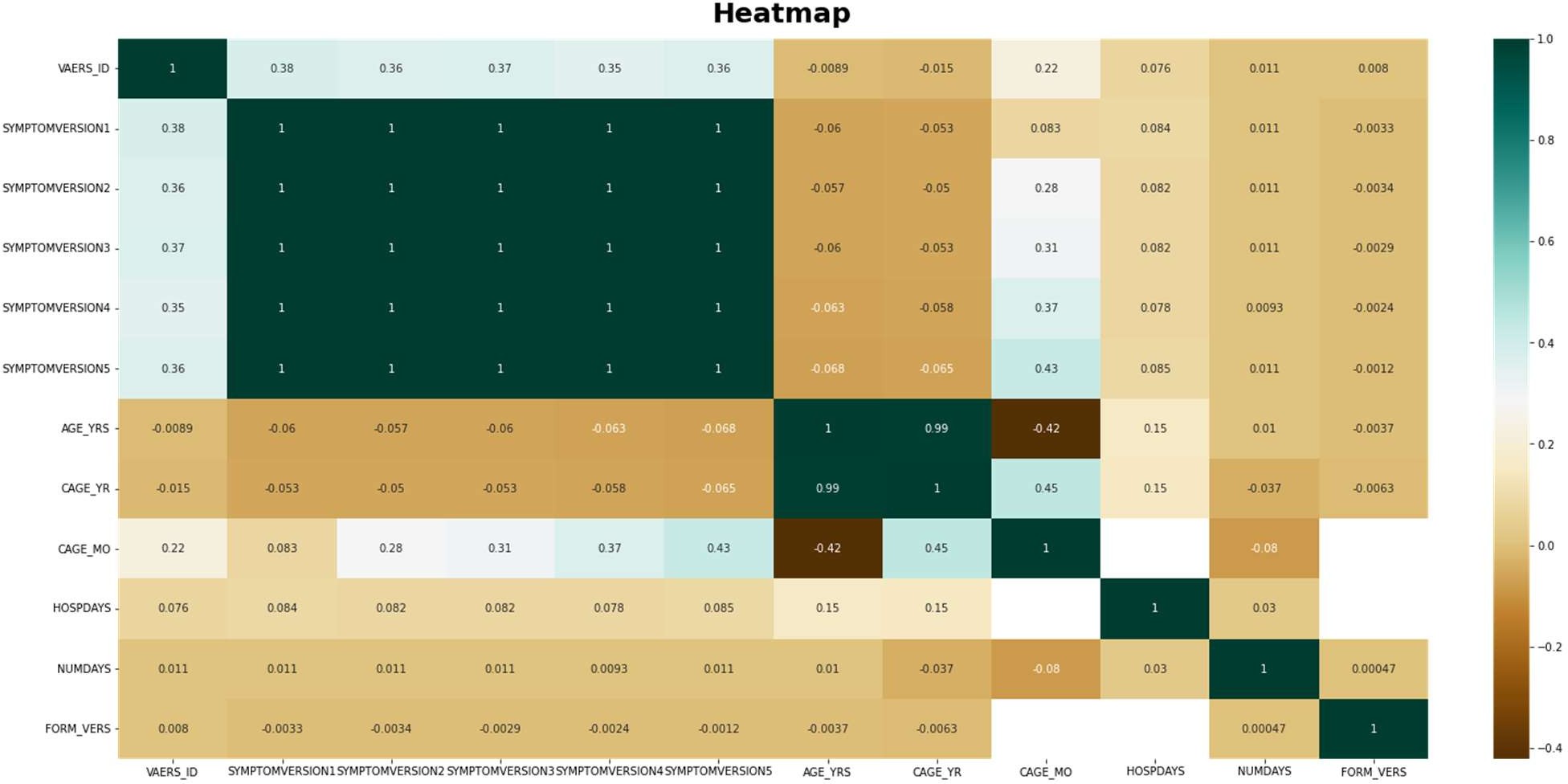
Pair plot is used to find the correlations between different variables to see the any specific relationship among them.

*Figure 7: Pair plot for the variables*



**Heat map** is the visualization with aims to give the better understanding of the dataset using different colour coding in high volume of locations with most data matters.

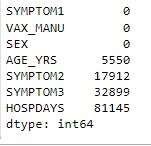
*Figure 8: Heat map showing different colour coding in high volume of locations*



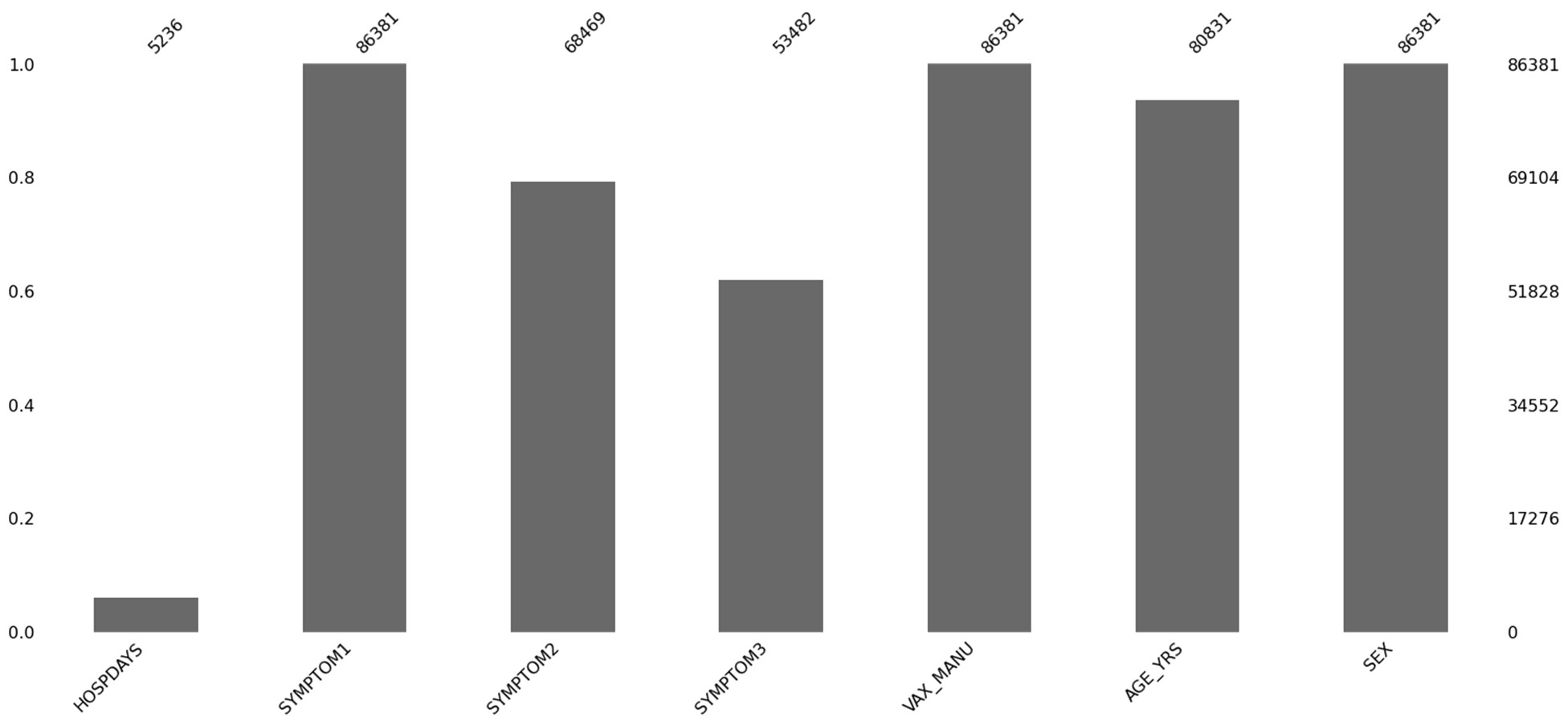
## Data preparation and pre processing

We have used different steps in data preparation such as cleaning the data, resizing of data, and feature engineering. We wanted to predict how many days a person will spend in hospital (if any days) after they advise they have had an adverse symptom to the Covid19 vaccine. From looking at the data dictionary, we dropped the features we believed were not necessary to complete our regression prediction problem i.e. we initially concentrated on the features we believed important for completing the regression problem. It is easier to choose the features we require than drop the features we don't require. This reduces the number of features from 52 columns to only 6 columns. Now, we have looked for missing values in our dataset and remove those values. There were tens of thousands of missing values in at least 4 of our selected features as shown in the below table and graph.

*Figure 9. Table of missing values*



*Graph 5: The data variables*



## Flowchart of Data Preparation and Modelling

Dataset check- understanding the raw data

Data Dictionary- check for quantitative and qualitative data

Analyze Features relevance/significance in regards to the project objectives

Decide the US state that we are analyzing



Descriptive Statistics trough visualization Top 10 Symptoms

Count of how many died Top 10 allergies

Top covid vaccine manufacturers Distributions of Sex and age

Correlation analysis

Feature Selection- Initial removal/dropping of the unwanted columns to avoid noise and reduce the complexity of the EDA process

Text processing and using the top symptoms as recorded in the 3 symptoms features, reducing the dimension of the dataset

Feature Engineering:

* Change PFIZER\BIONTECH to PFIZER
* Remove any rows where the SEX is recorded as 'U' and the VAX\_MANU is is unknown
* Change the na values in HOSPDAYS to 0
* Change the values to the SEX feature to 1 and 0
* Change the AGE\_YRS and HOSPDAYS from floats to ints.
* Change all strings to lower caps.
* Drop any remaining null values.

Encoding the Symptoms and VAX\_Manu feature

Modelling using Random Forest Algorithm Defining the X and Y features

Splitting the dataset into the test and training set Creating and fitting the regressor to the training set

Calculating the accuracy of the training and the test set.

## Feature engineering

We have feature engineering analysis to improve the predictions of the results by extracting the information from features of raw text. We may do the following steps in order to extract the information. We have changed “PFIZER\BIONTECH” to “PFIZER”. Removed any rows where the SEX is recorded as 'U' and the VAX\_MANU is unknown. Changed the na values in HOSPDAYS to 0. Changed the values to the SEX feature to 1 and 0. Changed the AGE\_YRS and HOSPDAYS from floats to integers. Changed all strings to lower caps. Dropped any remaining null values.

## Models

We have used two different machine learning algorithms including linear regression and random forest algorithm. Both the models have focused on to resolve the regression and classification prediction of featured variable. We have done different steps to make the data useful. Such as removing unnecessary variables and making the dummy variables of symptoms using age of the patients. After encoding there were 4239 rows and 44 feature variables in this dataset.

## Challenges encountered

Different challenges have been encountered in working with this dataset and applying models on it. These challenges may include the big size of dataset, categorizing the descriptive text of symptoms. There were a lot of missing values, cleaning and pre-processing of the data, sparse data after encoding, finding the perfect model that may give greater accuracy.

## Inclusion of strategies to overcome them

All of the 52 features (variables) in the dataset were not be useful for analysis. So, in order to reduce the size of large dataset, we have removed unnecessary features by selecting the features of our choice and removing all-other features. After the removal process, we remained with only 6 features to work on. There was a lot of missing values in our selected features. With the library of missing numbers, we have removed the missing values.

## Results and analysis

We don't need to scale our data for the ensemble machine learning model Random Forest. We have set the number of estimators at 10 and used ' entropy ' as the criterion parameter. The overall accuracy score of the model is 99.8% when predicting the number of days trough classification (if any) a person would require if reporting an adverse effect to the vaccine if we are provided with the SYMPTOM1, SYMPTOM2, SYMPTOM3, VAX\_MANU, AGE\_YRS and SEX of the patient.

## Step 1 - Defining the X and y

In this step we may be dividing the testing and training dataset. Separating out the dependent or target variable

(y) and the independent variables (X) and convert them into arrays using values as ML algorithms require arrays (Menduni et al., 2022). I'm looking to predict the number of days a person will stay in hospital after reporting an adverse effect of the Covid19 vaccine so the 'y' variable will be HOSPDAYS. The 'X' features are the features that will feed into our target feature.

Firstly, we have imported Random Forest Classification from sklearn.ensemble. Then we have created the classifier variable and use n estimators to set the number of parameters i.e., it's the number of trees we want to add. It is recommended to work with ten decision trees and ‘entropy’ as the criterion, so we apply these as our parameter for our model. We then fit the classifier onto our X\_train and y\_train variables.

The results in this model have showed a very high accuracy of 99.8 percent. In the evaluation of the model, we have used Crisp DM model also as shown in figure. So, we must re-evaluate to see what else can be done to improve and make our prediction accurate classifying/ predicting the model.

*Figure 10: Crisp DM model*



## Conclusion

In this paper we have focused to analyse the Vaers dataset which explains different adverse symptoms of patients after their vaccination been done. We have tried to see and predict the results of how many days people may have to be hospitalized with adverse symptoms after the vaccination process. So the number of days were analysed using repressors that almost 10 to 20 days patients may have to hospitalize in this situation. Using the Random Forrest classifier we have been able to predict with 99% accuracy the amount of days someone with certain symptoms may have to stay in hospital. However our results may not be as accurate as they first appear. The data is severely imbalanced which may be causing the classifier to get such good results. We are now entering the evaluation stage of the CRISP DM model which means we may have to re-visit some of the earlier steps to obtain confidence in our predictions. As a team we feel we are on the right track and we have made great progress up to this point with a little more effort we are well on our way to achieving our goals.

## References

Classen, J.B., 2021. Review of COVID-19 vaccines and the risk of chronic adverse events including neurological degeneration. *Journal of Medical-Clinical Research and Reviews*, *5*(4), pp.1-7.

Gee, J., Marquez, P., Su, J., Calvert, G., Liu, R., Myers, T., Nair, N., Martin, S., Clark, T., Markowitz, L., Lindsey, N., Zhang, B., Licata, C., Jazwa, A., Sotir, M. and Shimabukuro, T., 2021. First Month of COVID-19 Vaccine Safety Monitoring — United States, December 14, 2020–January 13,

2021. *MMWR. Morbidity and Mortality Weekly Report*, 70(8), pp.283-288.

Khanal, S., Fulton, J., Klopfenstein, A., Douridas, N. and Shearer, S., 2018. Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. *Computers and electronics in agriculture*, *153*, pp.213-225.

Menduni, G., Zifarelli, A., Sampaolo, A., Patimisco, P., Giglio, M., Amoroso, N., Wu, H., Dong, L., Bellotti, R. and Spagnolo, V., 2022. High-concentration methane and ethane QEPAS detection employing partial least squares regression to filter out energy relaxation dependence on gas matrix composition. *Photoacoustics*, *26*, p.100349.

Moro, P.L., Haber, P. and McNeil, M.M., 2019. Challenges in evaluating post-licensure vaccine safety: observations from the Centers for Disease Control and Prevention. *Expert Review of*

*Vaccines*, *18*(10), pp.1091-1101.