# ASSESSMENT REPORT FOR CERTIFICATE IN INTRODUCTORY DATA ANALYTICS (by Angela morrison)

# GitHub URL

[angmmorrison/UCDPA\_AngelaMorrison: Project for submittal at the end of the UCD Introduction to Data Analytics March 2022. (github.com)](https://github.com/angmmorrison/UCDPA_AngelaMorrison)

# Abstract

For the purpose of my project, I chose two publicly available datasets as follows;

## Dataset No.1: “COVID-19 Vaccine Adverse Reactions (VAERS) Dataset

<https://www.kaggle.com/datasets/landfallmotto/covid19-vaccine-adverse-reactions-vaers-dataset>

This dataset was located on the Kaggle.com website. This data is collected from VAERS website and contains Pfizer/BionTech, Moderna and Johnson & Johnson (Janssen) vaccines. Details in dataset includes Covid-19 symptoms, Age, Gender, Life-threatening illness, Allergies, Vaccination dates, Doses, Hospitalization, Recovery Status, Death, Vaccine Manufacturer, and many more information.

I reduced the size of the dataset down to make it more manageable, cleaned it, analysed it and created some visualizations.

## Dataset No. 2: Cork

[Yesterday's Weather Cork Airport - Yesterday's Weather Cork Airport - data.gov.ie](https://data.gov.ie/dataset/yesterdays-weather-cork-airport/resource/7ee747e2-e10d-4dac-b2b3-6ef647c31516)

This data is available on the Data.Gov.ie website. It shows the weather data from Cork Airport from the previous day.

I used this data to display relational plots between temperature and pressure.

# Introduction

I chose the VAERs dataset (Dataset No. 1) as I have a friend who is concerned about taking the Covid Vaccine due to her fear of developing a heart related illness afterwards. Also, I had noticed myself, that there seems to be more frequent sudden deaths and heart conditions in middle aged men since the onset of the Covid pandemic vaccination program. Therefore, I also wanted to satisfy my own curiosity as well. Since, I work in the pharmaceutical industry I am familiar with FDA (Food and Drugs Administration) terminology also.

I chose the Cork weather (Dataset No.2) as I wanted to find a short dataset with time series and relational variables. These were two dataset characteristics not very relevant to the VAERs (Database No.1).

# Dataset

The VAERs dataset (Dataset No. 1) contains 890,836 rows and 52 columns. This was a large, categorical dataset, which allowed me to perform many data cleaning and transforming activities once I had imported the file into a Pandas Dataframe.

I located this dataset from the Kaggle.com website.

Kaggle is an online community of [data scientists](https://en.wikipedia.org/wiki/Data_science) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning) practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine-learning engineers, and enter competitions to solve data science challenges.

The Cork-yesterday dataset is a smaller, relational, time-series, database that is publicly available on an Irish government website. I chose this data as it has many of the attributes, which were not apparent in the VAERs dataset.

# Implementation Process

## Dataset No. 1:

### Importing the Data.

* In order to get started quickly, I downloaded the dataset from the Kaggle website to my local directory. I then moved it into my Jupityer notebook local folder so that it was immediately accessible for importing.
* Next, I used the Pandas command pd.read\_csv command, assigning the first column (VAERS\_ID) as the index.
* To check the success of this import, I performed a df.head() command. This told me that the data had 51.no. columns, and I could see a sample of the data. I then used df.describe() to assess the number of rows.

### Transforming and Cleaning the Data.

* I used a df.round() and df.dropna() command on the “AGE\_YRS” column as I planned to visualise this data later.
* Since I planned to identify heart related issues, I combined the 5no. individual Symptom columns into one, new column, titled “Allsymptoms” by just adding (appending) the string data together using ‘+’.
* A quick print statement on that new column gave me a look at this new column data. I then used a df.info() statement to confirm the index location of this new column.
* Following this, I used df.dropna() to remove rows that contained all N/As
* After this my dataset was now reduced to 819,268 entries, this was verified by another df.info() command
* Next, I used iloc to slice out the unnecessary columns, reducing the number of columns from 52no. to 21no.
* After this, I used a df.fillna() command on the HOSPDAYS column to put a zero instead of an NA, as I wanted to plot this data later.
* I used df.describe() and len(df.DIED.unique) to establish the content of the ‘DIED’ column. Then I used a df.query() command to filter out records of those who had died into another database called ‘died’. A quick died.head() command confirmed that the data in that column appeared to be all “Y”.
* Following this, I created a list of terms, which I then used to filter out patients who had died but had no previous history of illnesses. I called this dataset “Nounderlying”.
* I then used the Itterows command on that data to print out the age of each patient and the data entry listed under “HISTORY”.
* I performed a .mean() command then on the “AGE\_YRS” column of the ‘died’ dataset to establish the age of those who had died following a report of a vaccine adverse reaction.
* I repeated a mean() command, then, to establish the mean age of those who died, and had no previous illness, using the Nounderlying dataset.
* I then created another list with just the mRNA vaccine manufaturers (Pfizer\BionTech and Moderna), by using .isin(). The number of rows in this database was confirmed with a df.info() command to be 267,835 entries.
* I repeated this isin() command for just the Janssen data, and this resulted in a database with 29,647 entries.
* To display the methodology around recombining data, I used the pd.concat() command to combine them.
* Similarly, I split the dataset by columns and then recombined them afterwards.
* Next, I sorted the Nounderlying dataset by VAX\_DATE
* Using Numpy, I created a pivot table of the mean and median years of deaths post vaccine adverse reaction, where the patient had no underlying illnesses.
* I also created another pivot table that counted the number of records. These pivot tables are exporting to the clipboard when run.
* I also subsetted out the heart conditions post vaccine from the full dataset using the .contains() command, and put them into a new dataset called dfsmall\_heart. In that, I estimated the mean age and created a pivot table by VAX\_MANU, SEX, and AGE\_YRS

### Visualizing the data

* The following commands were used to plot the data;
  + .hist(),
  + .plot(kind = “scatter”), and (kind=”bar”)
  + Using seaborn, I used sns.catplot() to create, point, box and scatter plots.

## Dataset No.2:

### Importing the Data.

* I saved the .CSV file to my Jupyter notebook folder and then imported it using pd.read\_csv command. Note I also imported it directly using the URL and verified that I got the same data, which I then used going forwards.
* For demonstration, I also imported the file using a JSON request and viewed the contents.

### Transforming and Cleaning the Data

* In advance of importing the data, I created a function to convert the temperature from Celcius to Fahrenheit called “convertTemp”
* I then applied my new “convertTemp” function to the column containing temperature in Celsius, to add a new column with temperature in Fahrenheit

### Visualizing the data

* I used a Seaborn plot to visualise the relationship between the temperature and the pressure.
* To export the charts, I used the .savefig() method. The pictures in this chart were then inserted using those files.

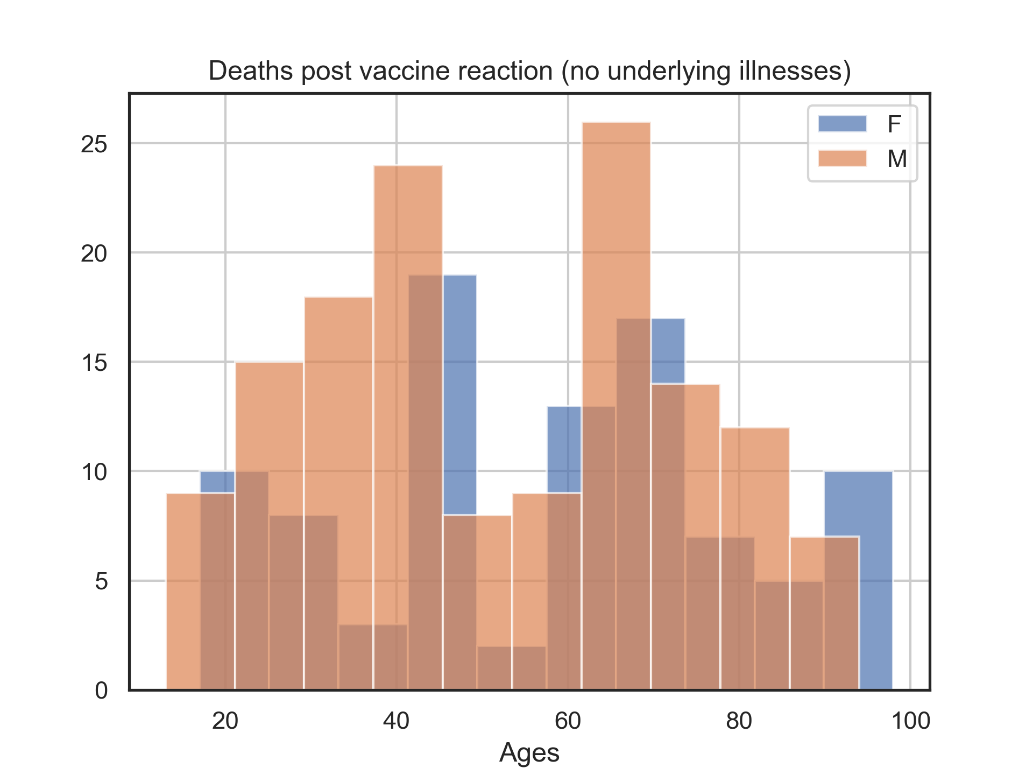
# Results

## Chart 1

This chart shows the relationship between age and the number of days in hospital post vaccine adverse reactions where patients had no underlying illness at the time of vaccination.

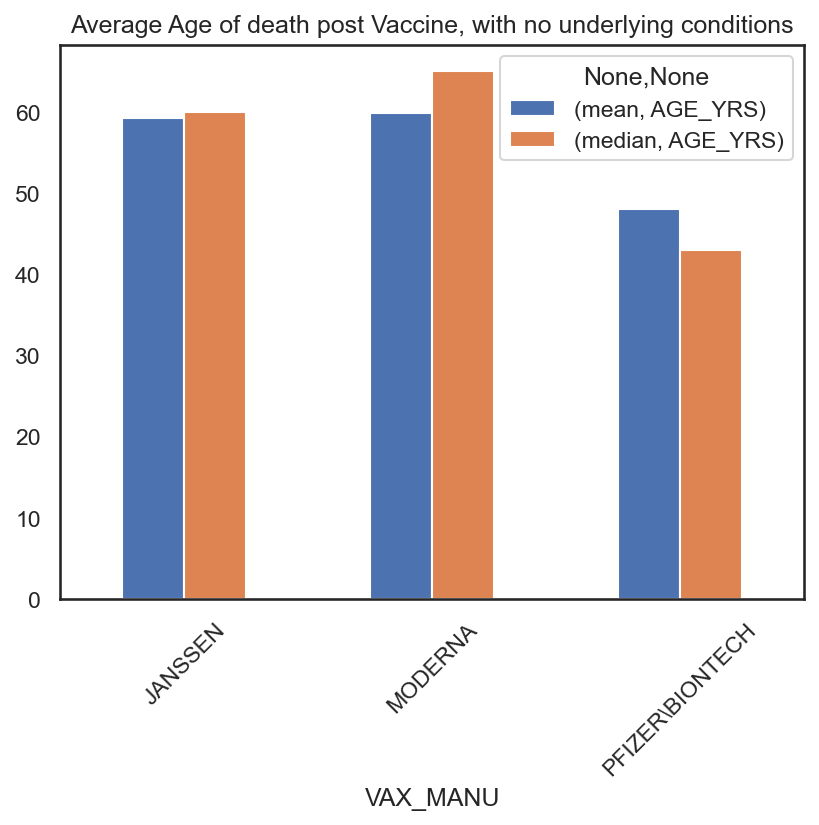
## 

## Chart 2.

This plot shows the distribution of deaths following a record of a vaccine adverse reaction, plotted by Male and Female. 

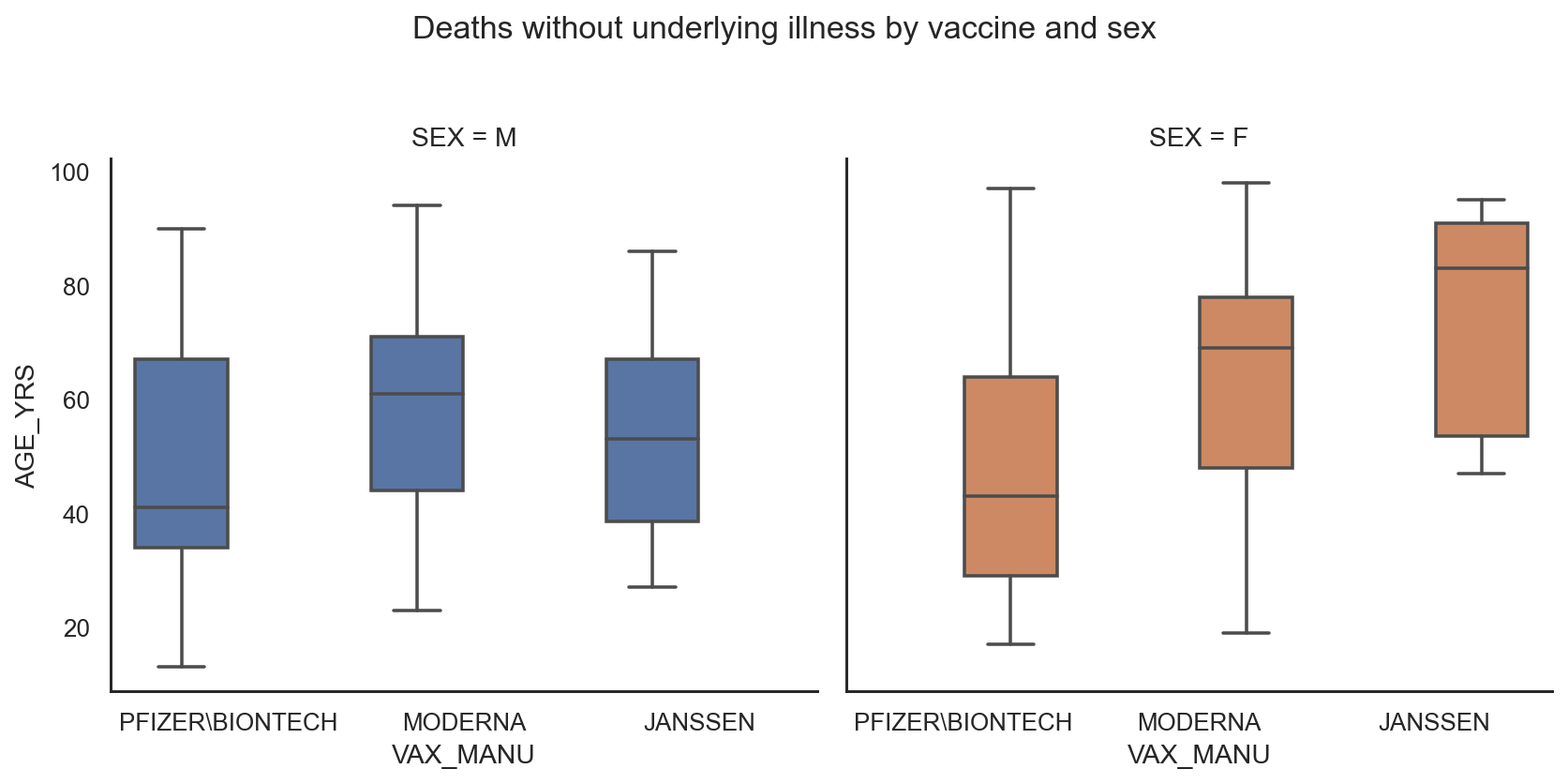
## Chart 3.

This chart shows the age of deaths post Vaccine, plotted by vaccine



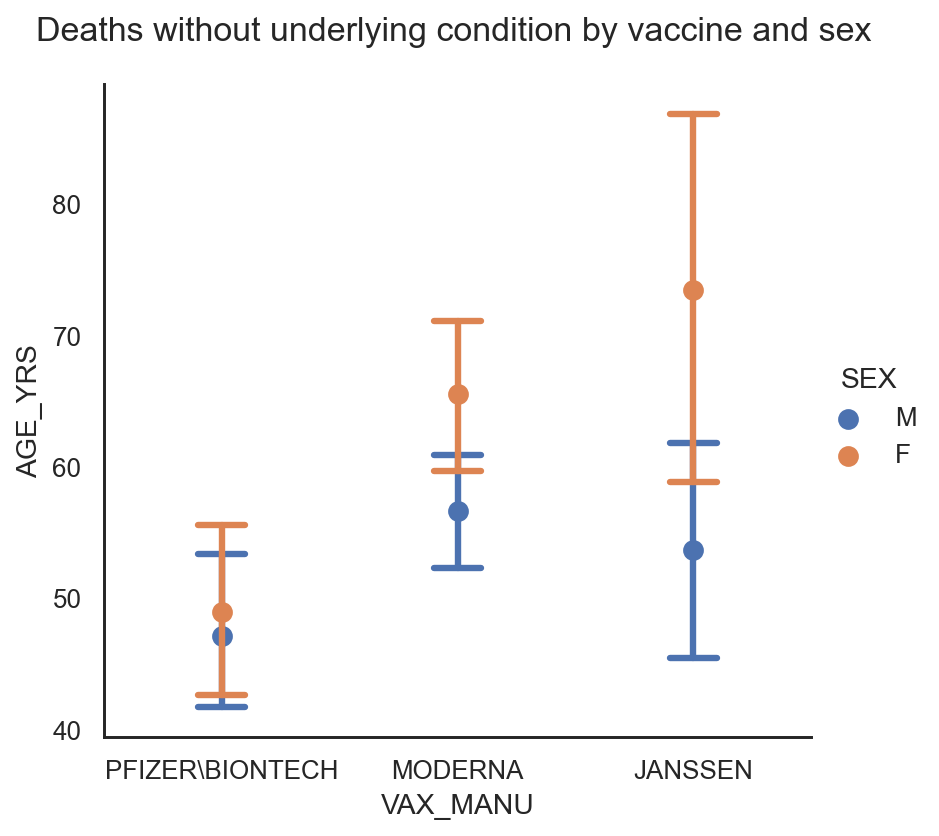
## Chart 4

This chart shows the profile of deaths by vaccine manufacturer, show for male and female patients separately.



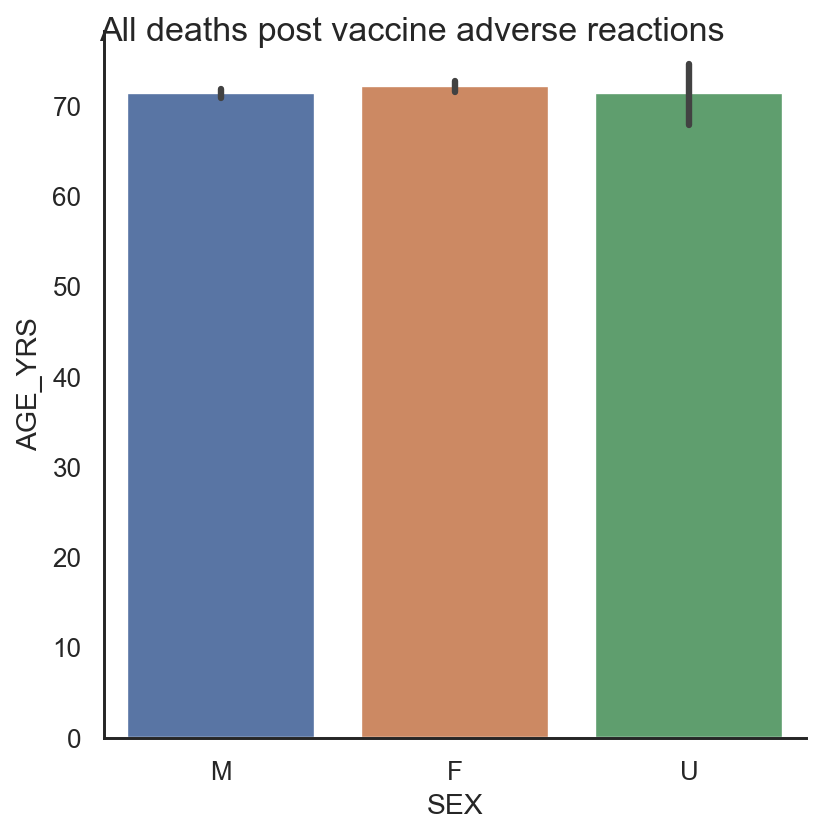
## Chart 5

This is the same data as Chart 4, but shown as a point plot.

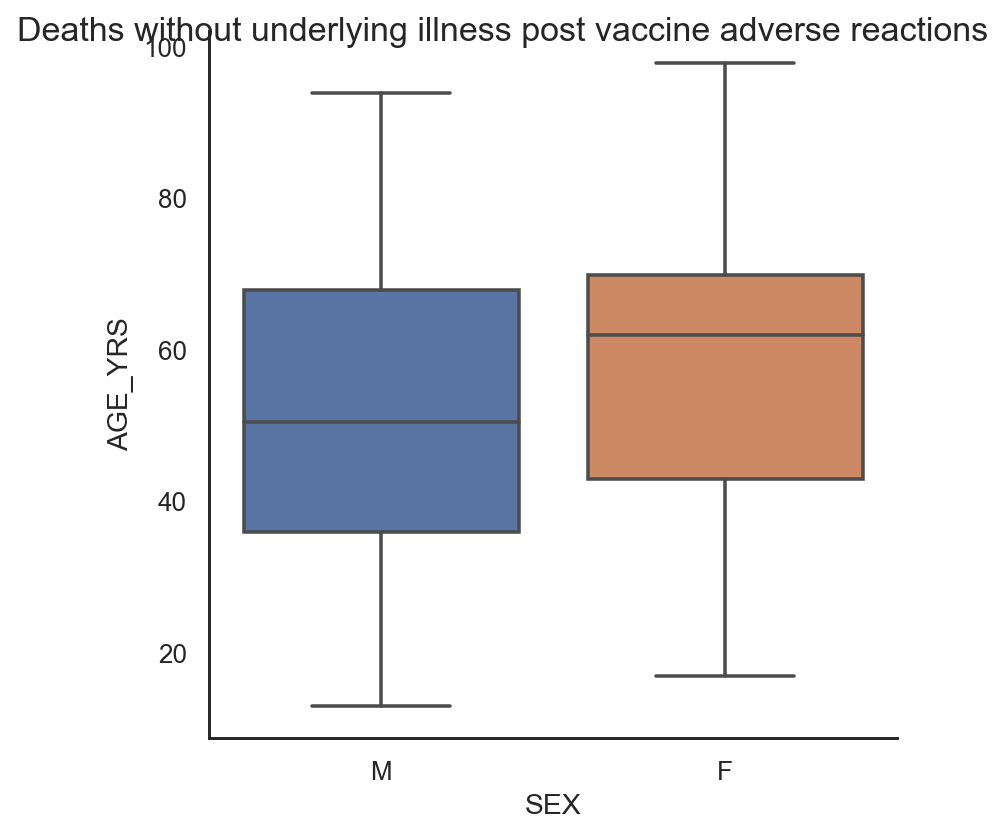


## Chart 6

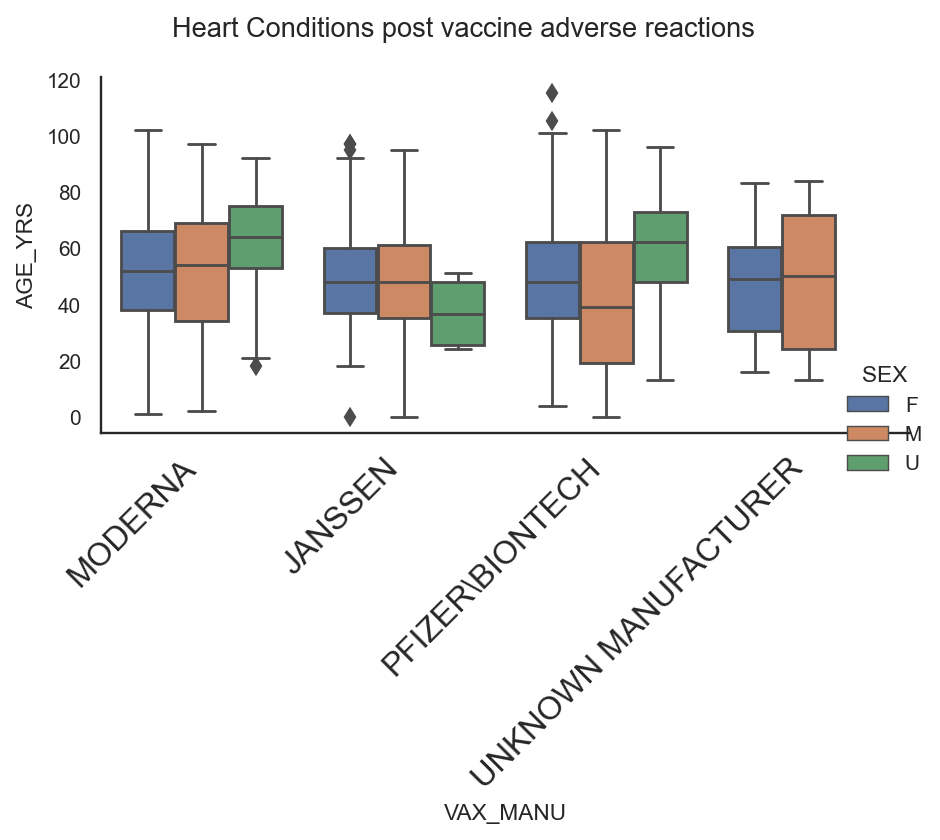
This chart shows all deaths post vaccine adverse reaction, by age and sex.



## Chart 7

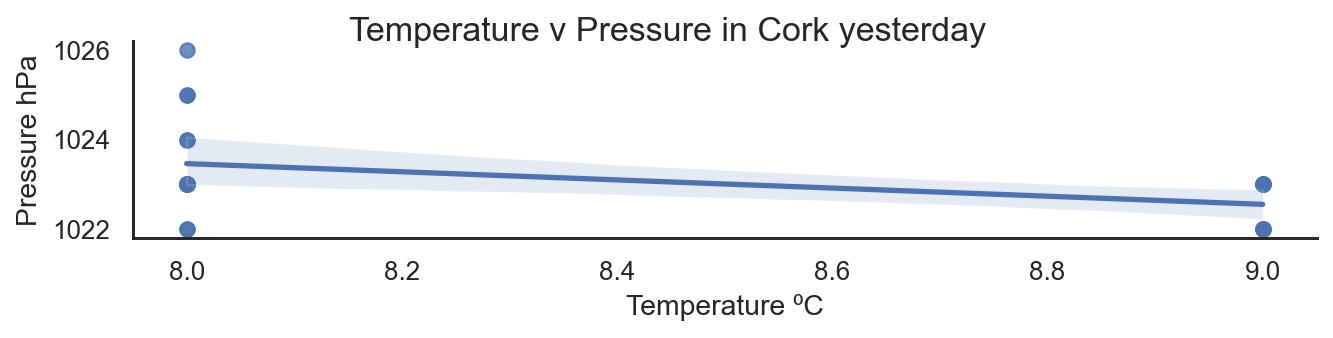


## Chart 8



## Chart 9

This chart shows the relationship between the temperature and pressure in Cork yesterday.



# Insights (Point out at least 5 insights in bullet points)

Note the insights below were based on the data available in the data log of vaccine adverse reactions. This dataset does not show how many vaccines were administered.. Further investigation should be done to statistically compare the vaccine adverse reaction data to the number of vaccine administered in order to determine the risk factors associated with these vaccines.

## Insight #1

From section 6.1, the number of days in hospital post vaccine adverse reactions where patients had no underlying illness at the time of vaccination, tended to be higher, the older the patient was

## Insight #2

From section 6.2, it seems that more men died post vaccine adverse reactions than women, and their ages tend to be slightly younger.

## Insight #3

From section 6.3 and 6.4, is seems that the age profile of deaths post vaccine reactions where there were no underlying illness is younger for the Pfizer\BionTech vaccine. The median age is comparable for the other two vaccines. This trend is also apparent in the heart conditions post vaccine adverse reactions, section 6.8

## Insight #4

From section 6.4 and section 6.5, the chart indicates that the age profile of women that died post adverse reactions to the Janssen vaccine tended to be significantly older than the men. For the other vaccines, there appeared to be no significant difference between the ages of the men and women.

## Insight #5

The chart in section 6.6 indicates that across all vaccines, the median age of death of men, women and unknown sex was comparable. However when the data was subsetted to look at those without previous underlying illnesses, section 6.7, the median age of men at death post vaccine was about 10 years younger than the women, aged 50 versus 60. 6.8).

## Insight #6

Section 6.9 shows us that as the air temperature rises, the air pressure drops.

# References (Include any references if required)

Kaggle.com

Data.Gov.ie

count

Allsymptoms

VAX\_MANU

JANSSEN 32594

MODERNA 139185

PFIZER\BIONTECH 139034

UNKNOWN MANUFACTURER 703