# ASSESSMENT REPORT FOR CERTIFICATE IN INTRODUCTORY DATA ANALYTICS

# GitHub URL (insert URL here)

# 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, subsetted it, 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 timeseries and relational variables. These were two characteristics of datasets that were not really relevant to the VAERs database.

# Dataset (Provide a description of your dataset and source. Also justify why you chose this source)

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, timeseries, database which 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:

### Transforming and Cleaning 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.
* Then 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.
* 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. Following this, the number of columns was reduced from 52 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 canned this dataset “Nounderlying”.
* I then used the Itterows command on that data to print out the age of each patient and their history of underlying illnesses.
* I performed a .mean() command then on the “AGE\_YRS” column of the died dataset to establish that 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)
* Using .isin() I created a separate database for these vaccine reports. The number of rows in this database was established with a df.info() command to be 267,835 entries.
* I repeated this 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. Note there are some ‘unknown’ vaccine manufaturers listed, hence the recombination does not match the original number of rows as I did not subset out the unknowns.
* Similarly I split the dataset by columns and then recombined them afterwards to show this methodology.
* 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 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

### Dataset No. 1: 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:

### Transforming and Cleaning the Data.

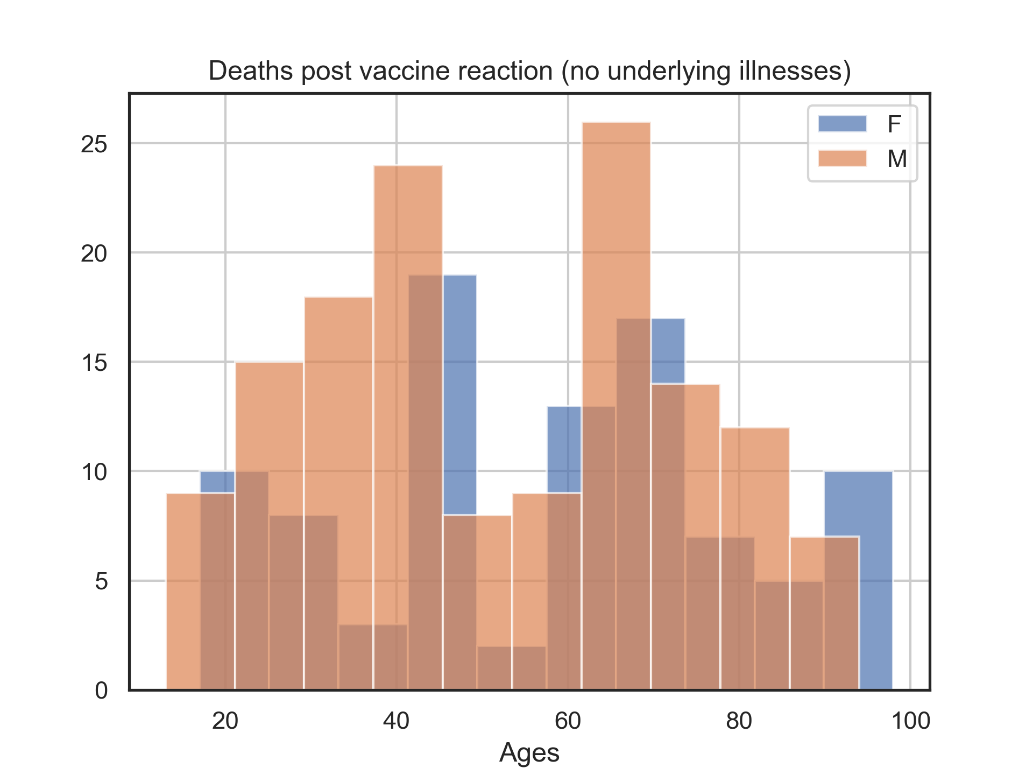
* In advance of importing the data, I created a function to convert the temperature from Celcius to Fahrenheit called convertTemp
* I saved the .CSV file to my Jupyter notebook folder and then imported it using pd.read\_csv command. Note I also imported it directly using the URL and verified that I got the same data, which I then used going fowrads.
* For demonstration, I also imported the file using a JSON request and viewed the contents.
* I then applied my new convertTemp function to the column containing temperature in Celsius, to add a new column with temperature in Fahrenheit

### Dataset No. 2: 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.

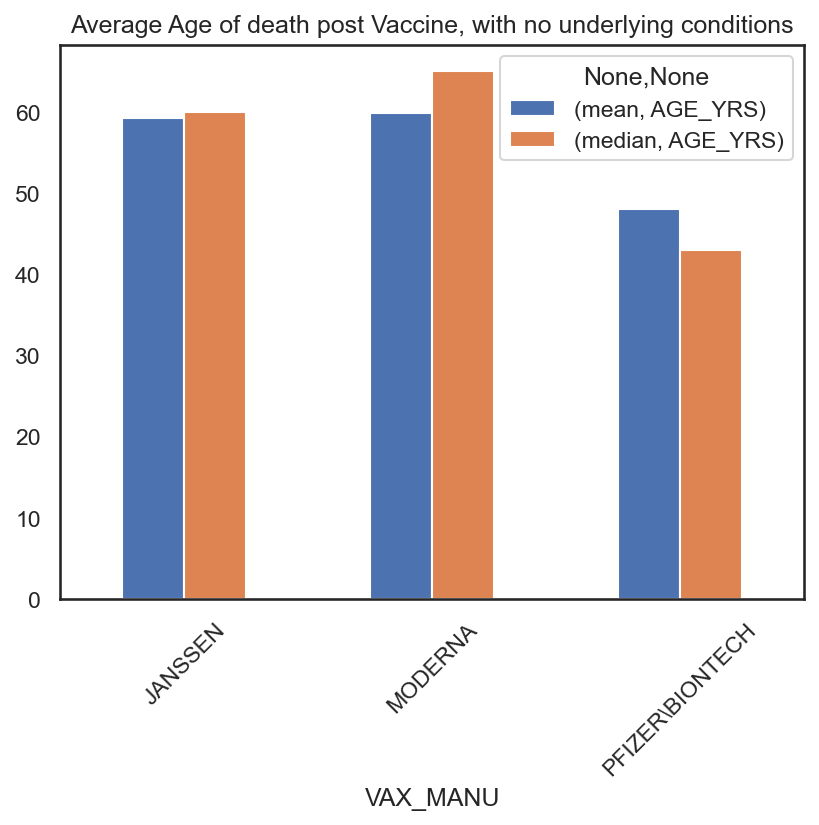
# Results (Include the charts and describe them)

## Chart 1.

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

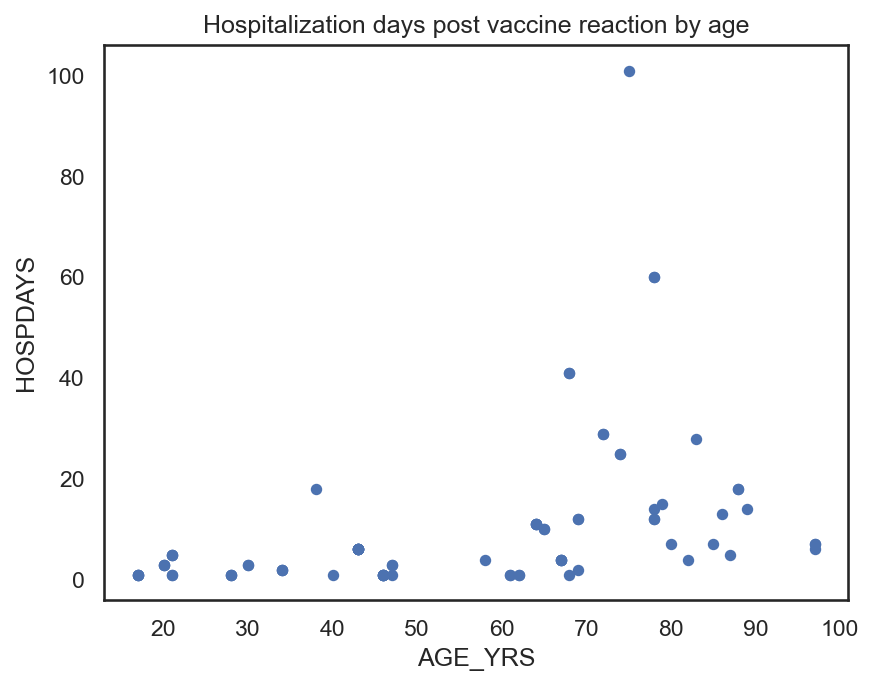
## Chart 2.

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



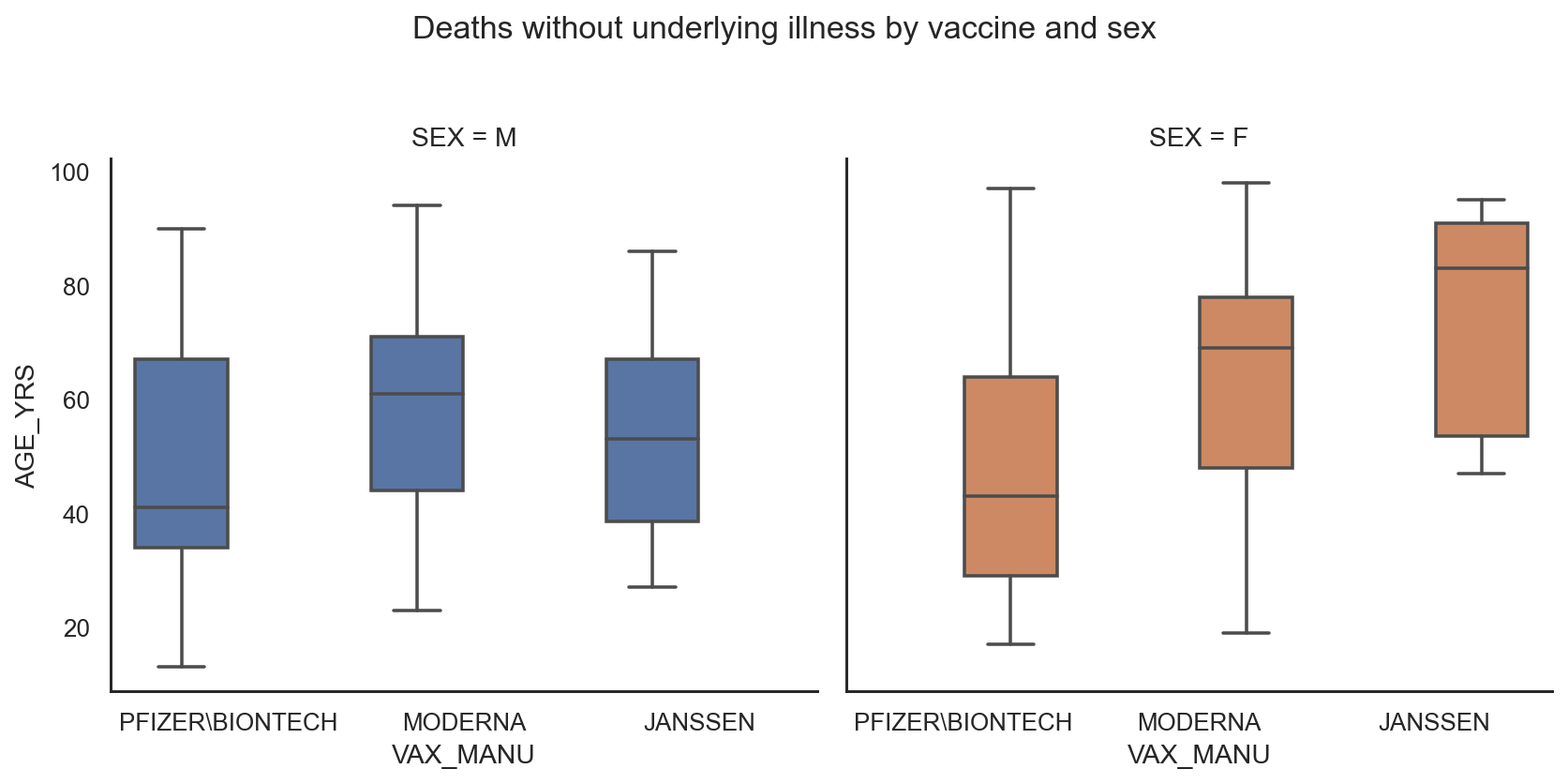
## Chart 3

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.



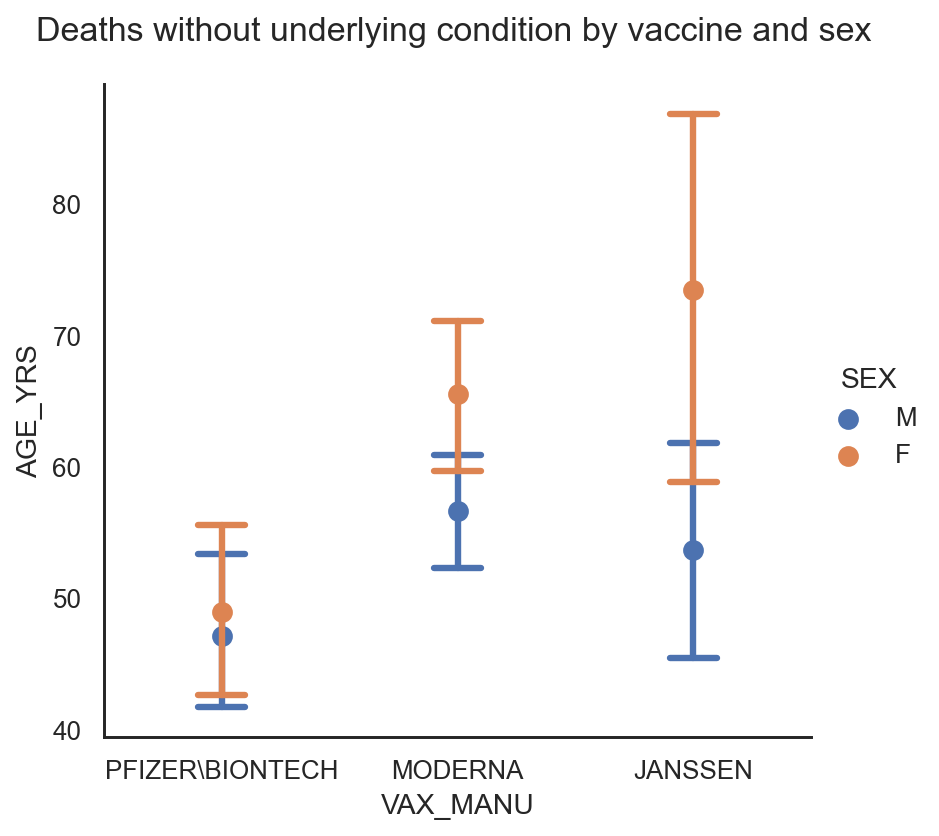
## Chart 4

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



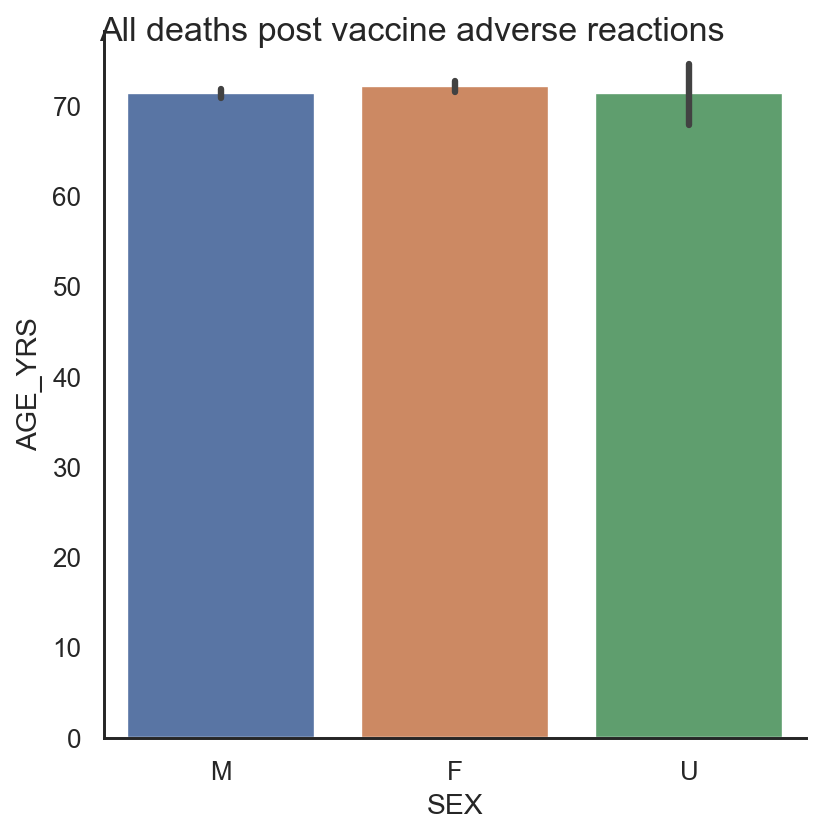
## Chart 5

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

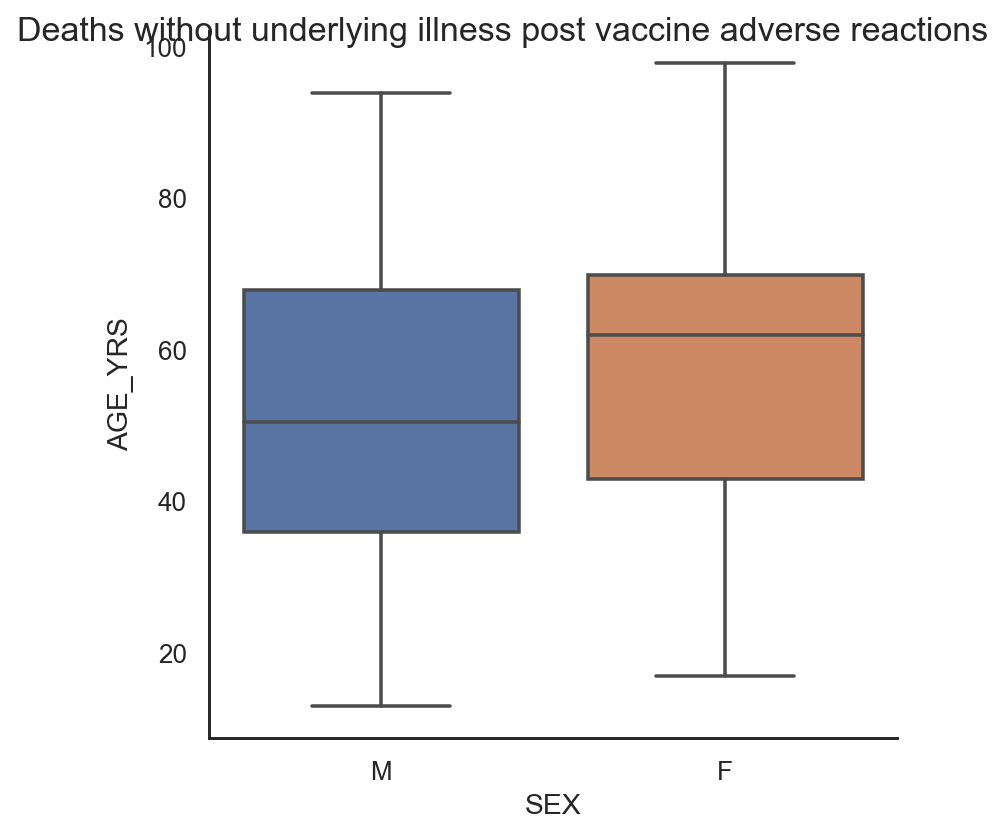


## Chart 6

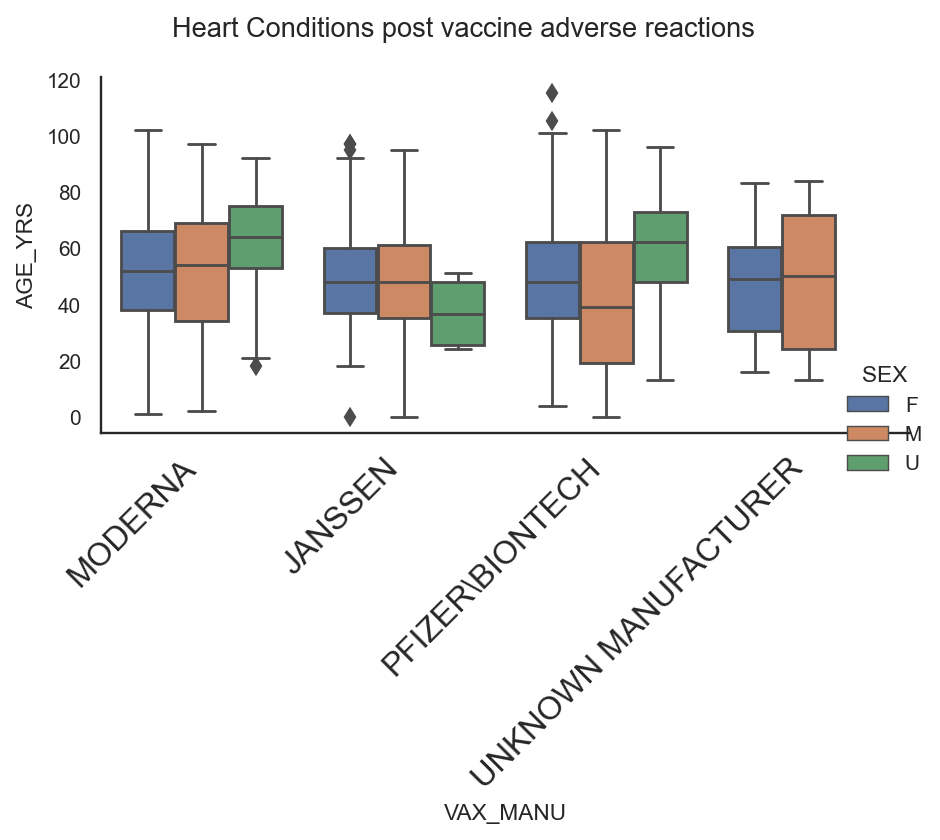
This chart shows all deaths post vaccine adverse reaction, by age and sex.



## Chart 7

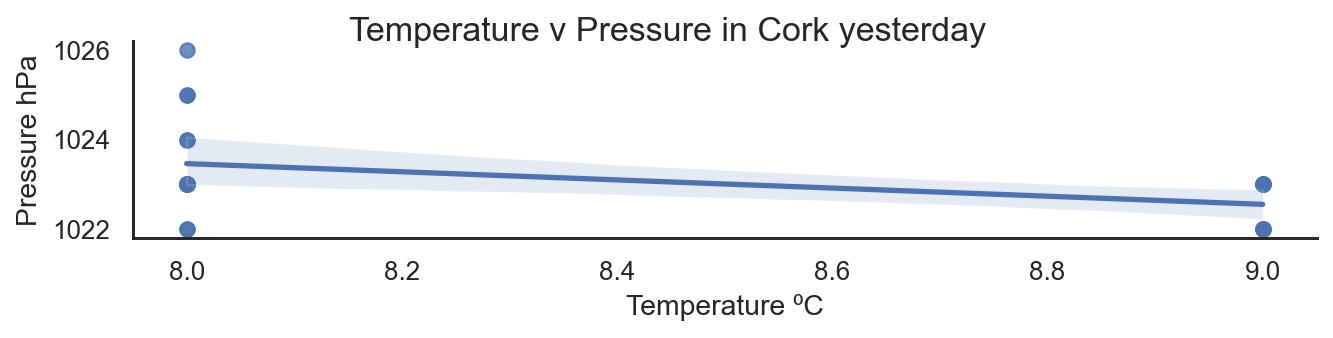


## Chart 8



## 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)

## Insight #1

From section 6.1, it seems that more men died post vaccine adverse reactions than women, and their ages tend to be slightly younger.

## Insight #2

From section 6.2, 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 vaccines.

## Insight #3

From section 6.3, 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 #4

From section 6.4, 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 6.7 indicates that across all vaccines, where the patients had no previous underlying illnesses, the median age of men at death post vaccine was about 10 years younger than the women, aged 50 versus 60. Looking at the vaccines individually (sections 6.5), the Pfizer\Biontech vaccine appears to show the youngest median age of death with the male deaths significantly younger than the female. This subgroup data is also reflected in the age profile for heart conditions post vaccine adverse effect (section 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