[Skills Bootcamp for Data Analytics Portfolio]

[Ameena Hamid]

[Early Stage Diabetes Risk Prediction Dataset]

One of the datasets I have decided to analyse for my portfolio has a theme of health, as this relates to my background in science. The dataset I have selected consists of data for the key signs and symptom of people who are either at risk of developing diabetes and are showing early signs/symptoms of the disease have tested positive.

Can we predict an onset of Diabetes from the symptoms an individual is showing?

Early Stage Diabetes Risk Prediction Dataset - What does the data look like?

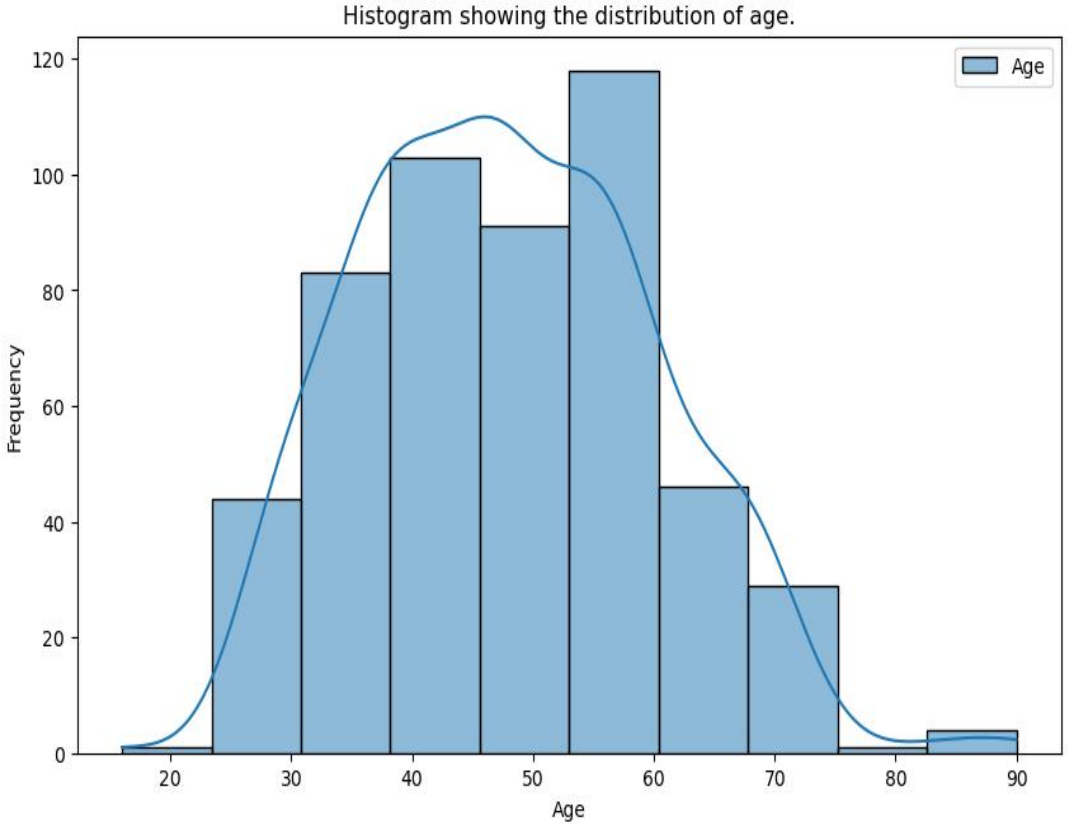
- Data for 520 individuals of a varied age range
- 14 different symptoms of a possible diabetes patient
- Data shows which symptom is present/not present for each individual
- Class confirms if an individual is positive or negative for Diabetes

Early Stage Diabetes Risk Prediction Dataset - Exploratory data analysis

- The link below shows the Jupyter Notebook with the exploratory data analysis and code

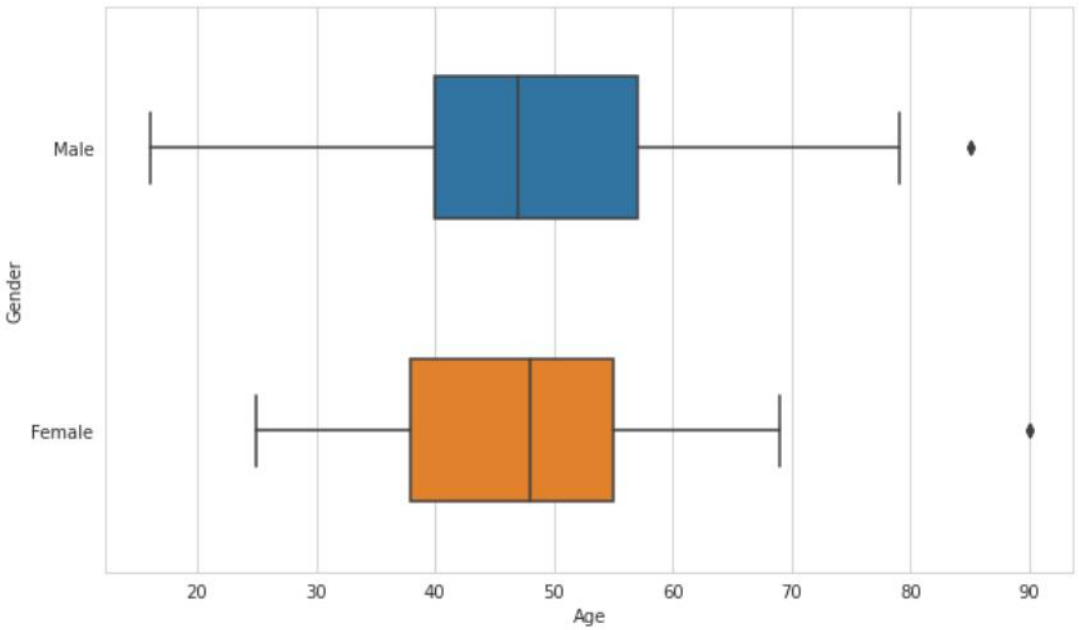
- https://anaconda.cloud/share/notebooks/496e882b-39f1-4e71-98d4-ef32a37aa94a/overview

Early Stage Diabetes Risk Prediction Dataset - Data Visualization



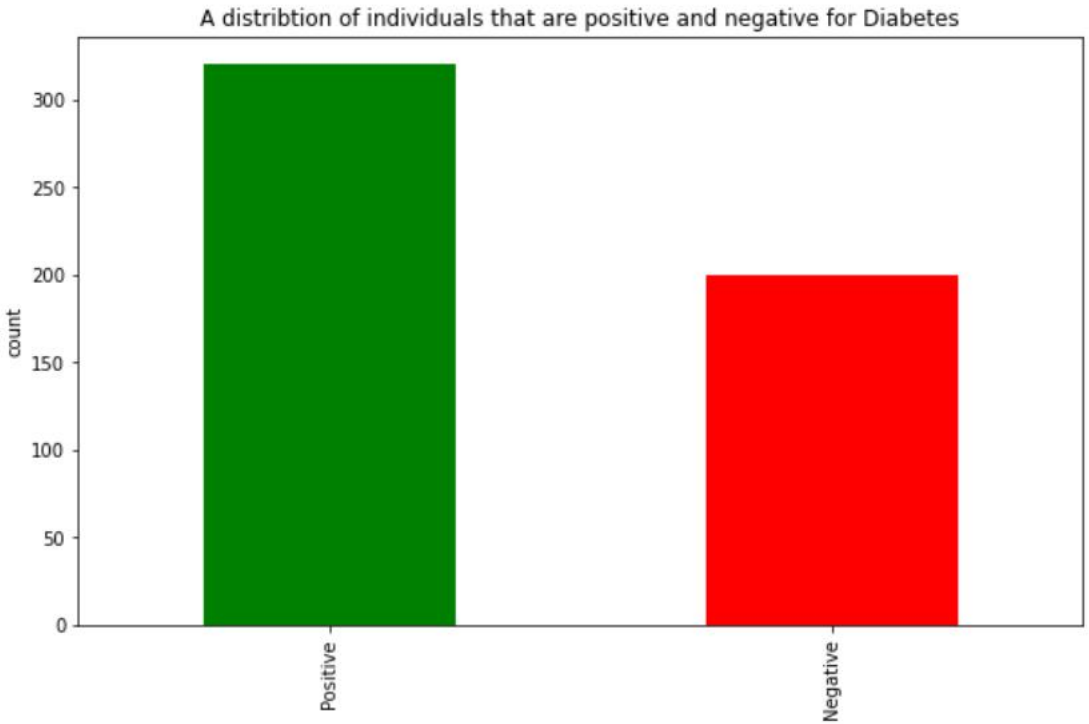
- The Histogram is showing a fairly normal distribution of ages, with a bell-shaped curve
- We can see a cluster of individuals between 40-60
- The mean, median and mode of the data are also similar, signifying a normal distribution (see Jupyter notebook)

Early Stage Diabetes Risk Prediction Dataset - Data Visualization



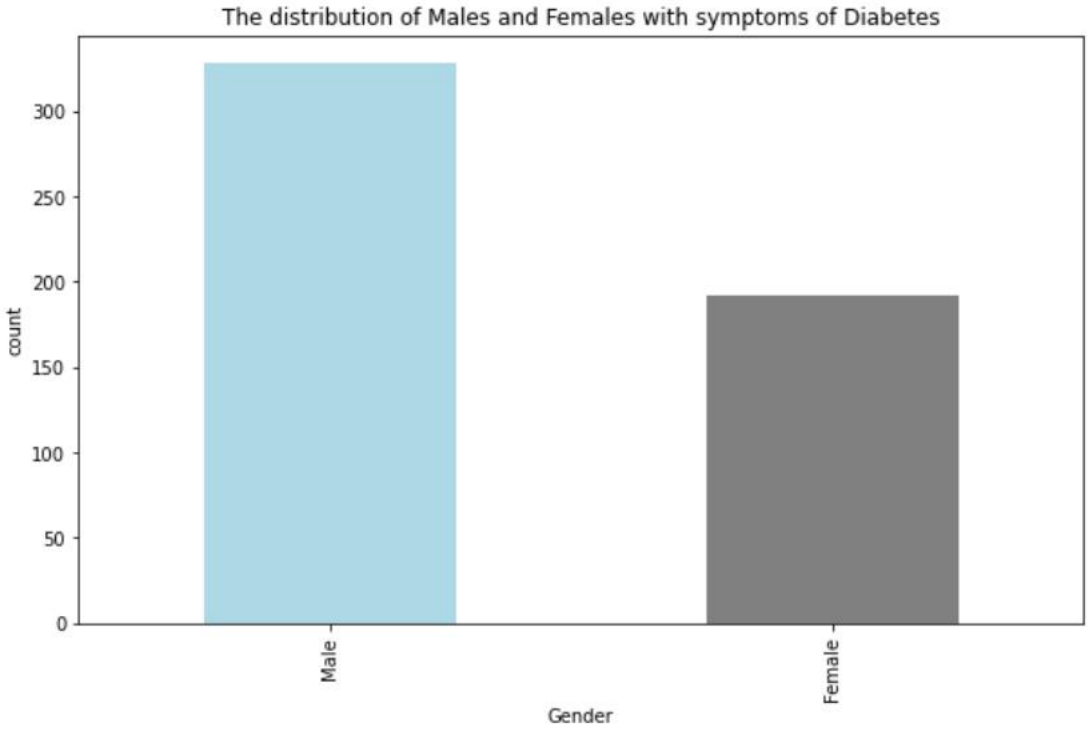
- The Box plot is showing the distribution of ages for each Gender in more detail
- The visualisation shows that the age range for Male individuals is right skewed and for Female individuals is left skewed.
- The middle 50% of the data for Males lies between 40 - 57 and for Females lies between 37 - 55
- Are females more likely to be diagnosed with diabetes at a younger age than males?

Early Stage Diabetes Risk Prediction Dataset - Data Visualization



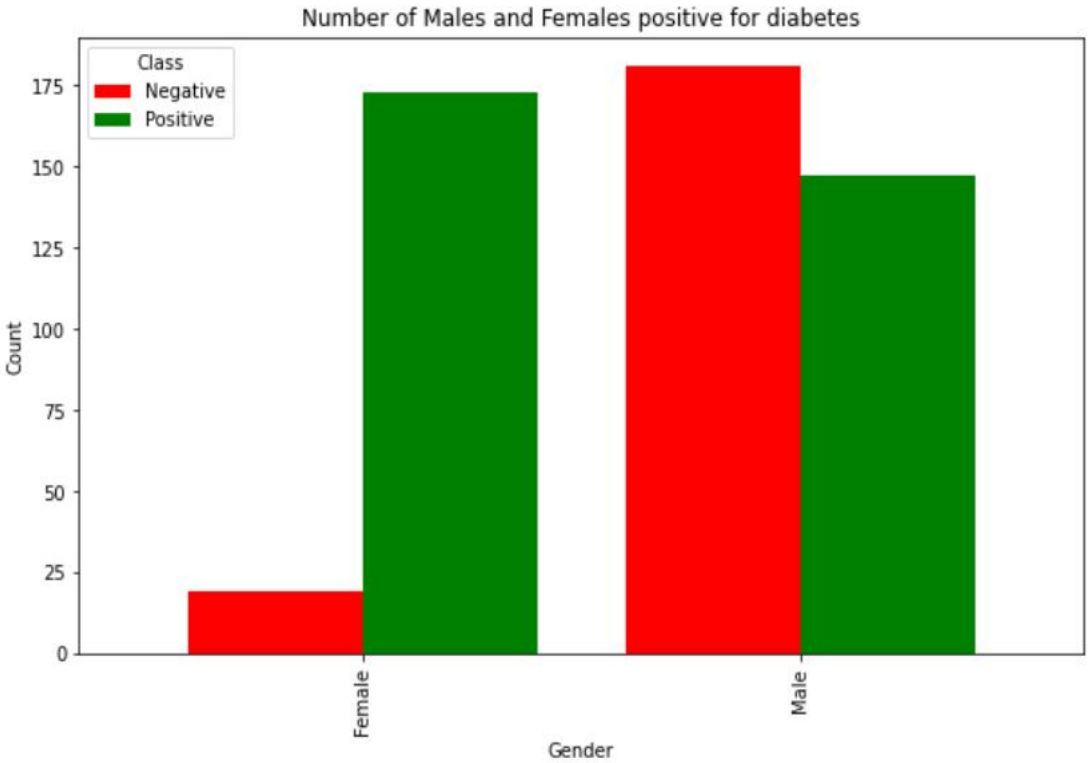
- The number of people positive for Diabetes are significantly higher than those that were negative but may still have had some of the symptoms

Early Stage Diabetes Risk Prediction Dataset - Data Visualization



- From this visualization we can see that the number of males with symptoms of diabetes is almsot 50% more than females with the symptoms
- Can we do further research to see why this could be a possibility?

Early Stage Diabetes Risk Prediction Dataset - Data Visualization



- Further analysis shows that the number of Females positive for Diabetes is slightly higher than the number of Males that are positive.
- The number of females that are negative for Diabetes but may be displaying symptoms are significantly lower than the number of males that are positive
- However, the number of females overall with symptoms are also considerably lower than males

Early Stage Diabetes Risk Prediction Dataset - Data Visualization

- Further analysis shows that the number of Females positive for Diabetes is slightly higher than the number of Males that are positive.
- The number of females that are negative for Diabetes but may be displaying symptoms are significantly lower than the number of males that are positive
- However, the number of females overall with symptoms are also considerably lower than males

Early Stage Diabetes Risk Prediction Dataset - Summary

- The typical age range for individuals with symptoms of diabetes or on the onset of diabetes is between 30-60
- Males are more prone to diabetes than females
- Females may be diagnosed with diabetes at a younger age than Males
- Do we need look at data that has an equal number of females to males ratio for a more accurate analysis?