**Springboard Foundations of Data Science**

**Capstone Project**

**Diabetes**

By: Swati Jani Joshi

Project Description:

The Dataset is from Kaggle – Prima Indians Diabetes Data set.

The objective of the capstone project is to provide analysis on the data set so as to derive a model that can predict the onset of diabetes.

Diabetes:

This project would be beneficial to scientists/researchers in the health services field as it would give a better insight of who would be pre disposed to diabetes. This is turn will benefit doctors to monitor people who are determined to be predisposed to diabetes. They can conduct frequent tests on such patients to monitor them and provide them with guidance on how to prevent or delay the onset of diabetes.

Diabetes is the seventh leading cause of deaths in the US.

Data:

The Kaggle Data set has nine features in the dataset.

The dataset contains data on women only.

Data Wrangling:

The Dataset had many observations with values of 0 for features such as Blood Pressure, Skin Thickness, Insulin etc. Since it is not medically possible for an alive person to have such values, those records were deleted. Since it is possible for women to have no children, observations where pregnancies were zero, were unchanged in the dataset.

Data Analysis:

A dot graph was plotted to visualize the correlations between the nine features. The corrplot function was used for the dot graph.

Some interesting observations were made based on the corrplot.

Age, Glucose, BMI and Insulin have significant impact on Outcome.

Age had a significant impact on pregnancies, which is puzzling, as generally with older women, number of pregnancies should decrease.

Modelling:

Logistic Regression Model

The glm function was used to determine logistic regression model fitting.

Logistic regression was chosen as the Outcome variable is a 1 or 0 i.e. yes or no which is a binary format.