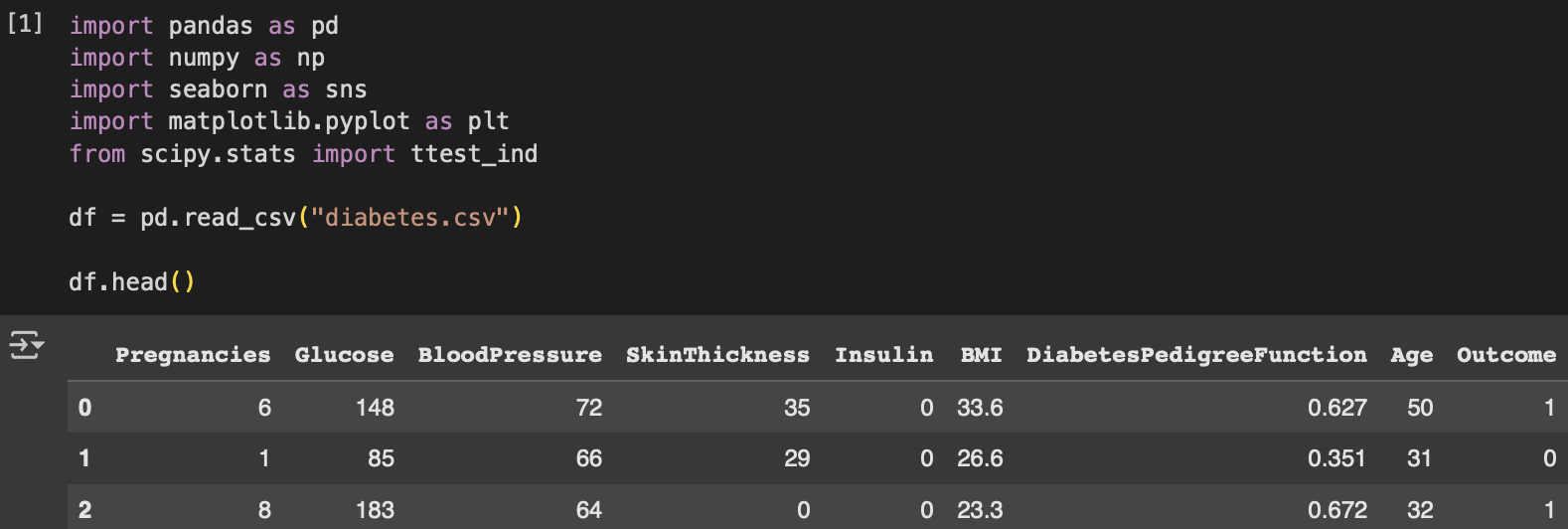
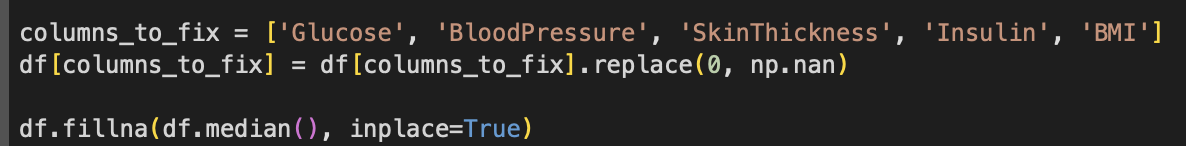
**BIS Assignment - Report**

**Index – 20020406 (2020/IS/040)**

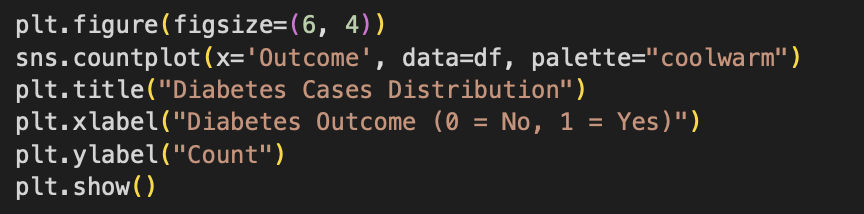
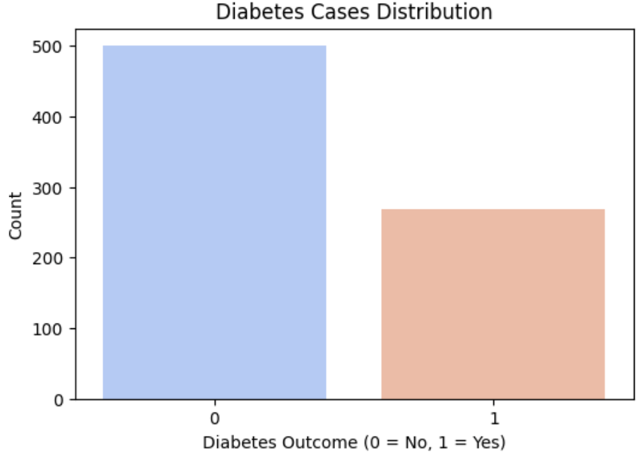
* Business Domain – **Healthcare - Predicting Diabetes Risk**
* Dataset - Pima Indians Diabetes Database
* Source - Kaggle (“uciml/pima-indians-diabetes-database")
* The healthcare industry relies on data analytics to predict diseases and improve preventive healthcare measures.
* Diabetes is a chronic condition that needs early detection to prevent complications.
* This dataset contains 768 patient records with 8 medical features.
* Target Variable - Outcome (1 = Diabetic, 0 = Non-Diabetic)
* Key Features - *Glucose, Blood Pressure, BMI, Age, Insulin, Skin Thickness, Diabetes Pedigree Function, and Pregnancies.*
* Business Question - **"How can we identify high-risk individuals for diabetes based on key health indicators, and what preventive measures can be suggested?"**
* Analytics Process
  + Data Collection – Extract dataset from Kaggle.
  + Data Preprocessing – Handle missing values & clean the data.
  + Exploratory Data Analysis – Identify diabetes prevalence and feature impact.
  + Statistical & Machine Learning Analysis – Perform Linear Regression, Correlation Analysis, and Clustering.
  + Data Visualization – Graphically represent trends using Matplotlib & Seaborn.
* Loading the data



* Preprocessing – handling missing values



* Exploratory Data Analysis
  + Diabetes Outcome Distribution



**Interpretation -**  Approximately **35% of the dataset**

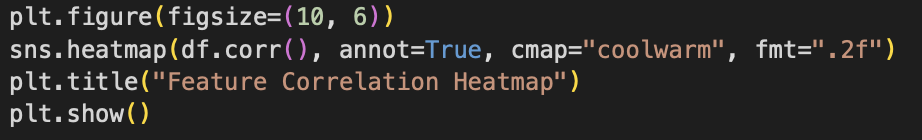
Consists of diabetic patients. The dataset is

slightly **imbalanced**, which is important when

applying machine learning models.

* + A diagram of a heatmap

    AI-generated content may be incorrect.Feature Correlation Analysis (Heatmap)



**Interpretation**

1. Glucose has the highest correlation

(0.49) with diabetes.

1. BMI (0.31) and Age (0.24) also contribute to

diabetes risk.

1. Blood Pressure & Insulin have weak

correlations and may not be strong predictors.

* Statistical Analysis - T-Test for Glucose Levels

**A screenshot of a computer program

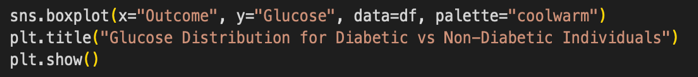
AI-generated content may be incorrect.Interpretation**

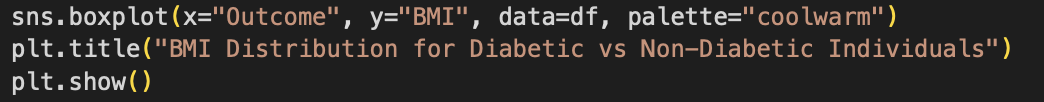
Glucose levels are significantly higher in

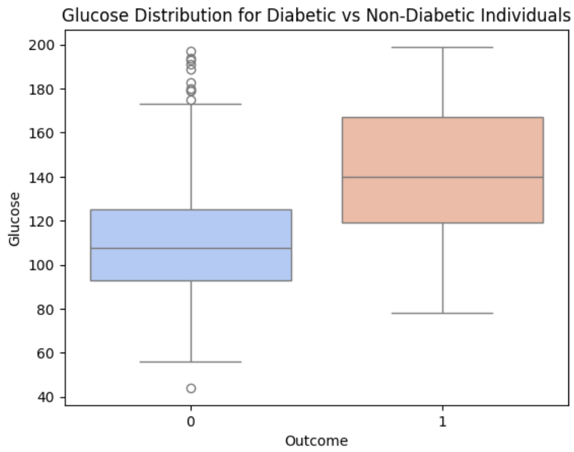
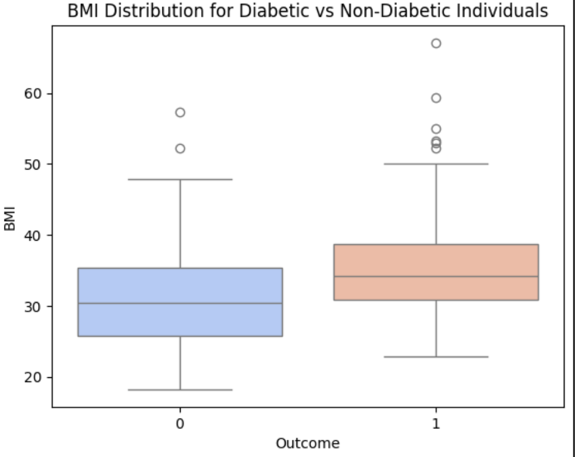
diabetic individuals (p-value < 0.05).

This confirms glucose is a key predictor for

diabetes.

* Data Visualization

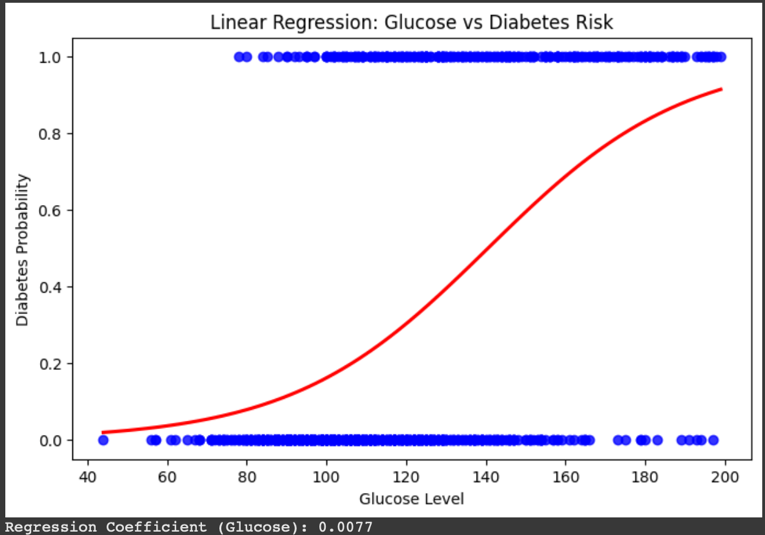


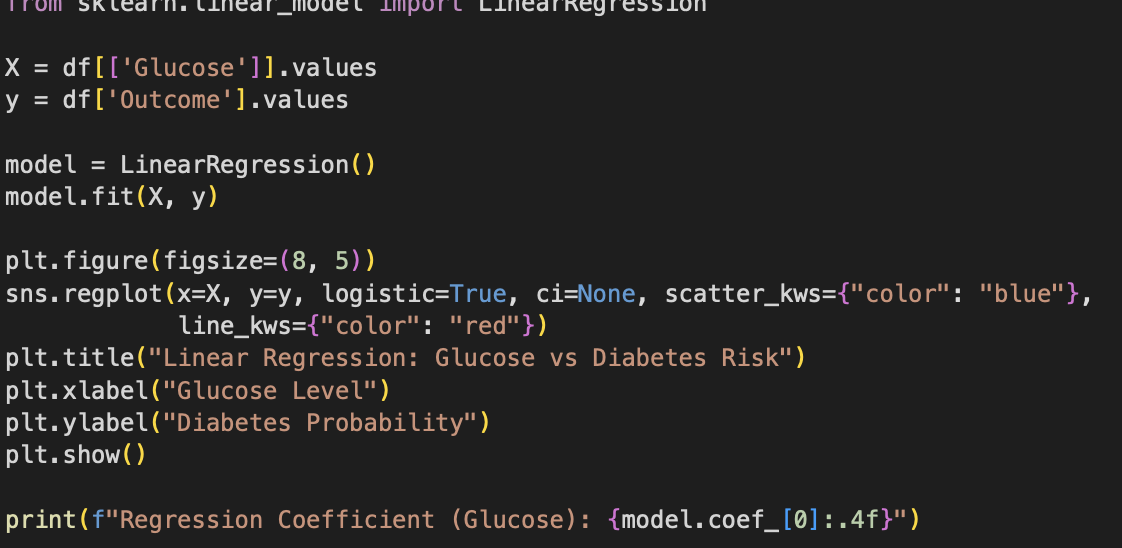


**Interpretation**

1. Diabetic patients have higher BMI & Glucose levels than non-diabetic patients.
2. Preventive strategies should focus on reducing BMI & glucose levels.

* Regression Analysis (Linear Regression) - Does Glucose Predict Diabetes Risk?





**Interpretation**

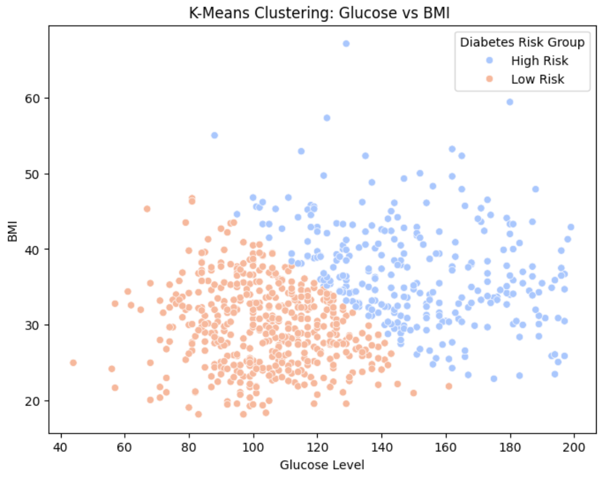
1. Higher glucose levels increase diabetes

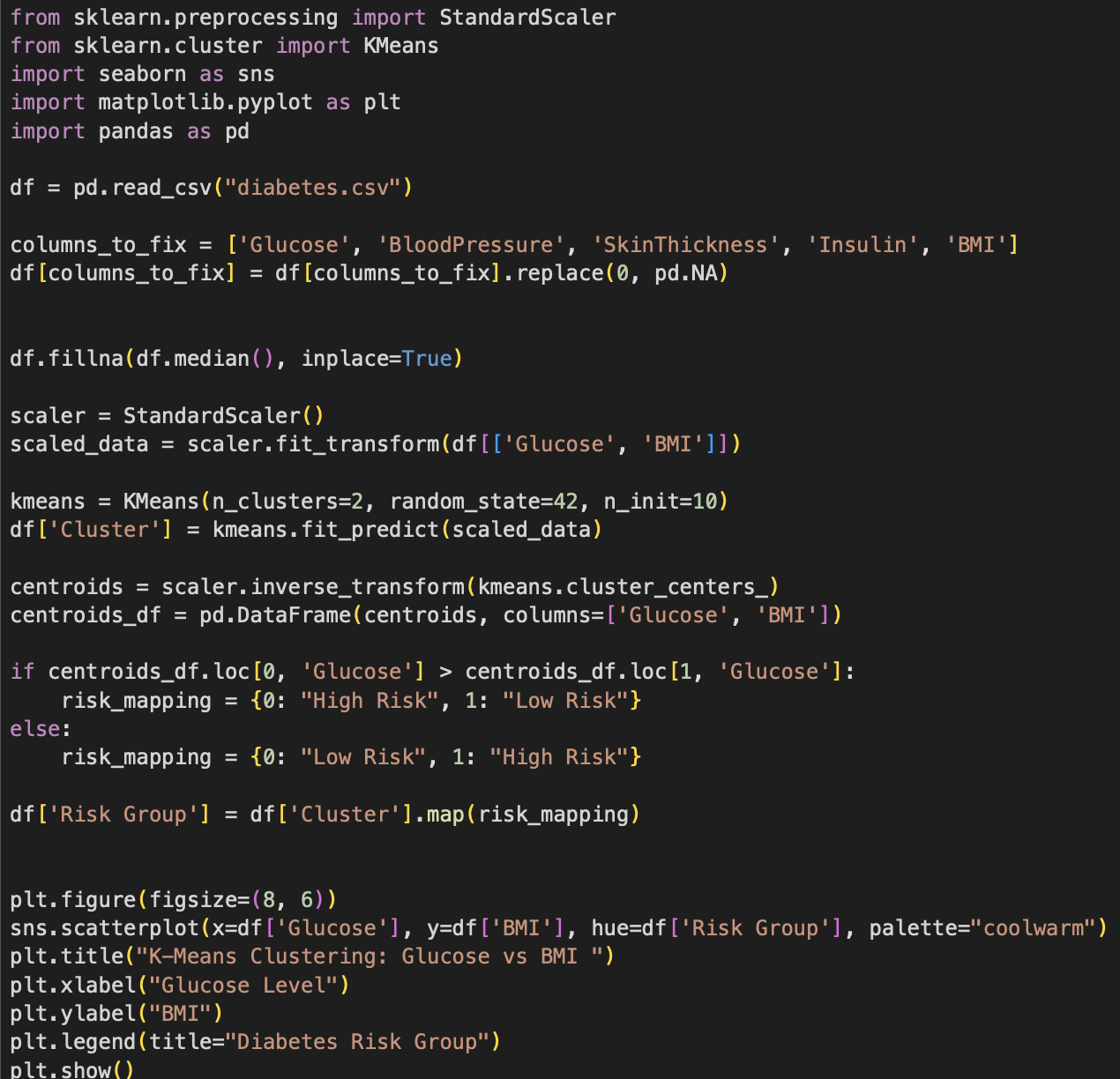
probability.

1. Glucose has a positive regression

coefficient, confirming its impact on diabetes.

* Clustering Analysis (K-Means) - Identifying Diabetes Risk Groups





**Interpretation**

1. K-Means clustering divides patients into  high-risk (red) and low-risk (blue) diabetes groups.
2. Higher glucose & BMI indicate higher diabetes risk.
3. This clustering helps in identifying high-risk individuals for early intervention
4. **Findings & Business Implications**

* Glucose is the strongest predictor of diabetes.
* BMI and Age also significantly impacting diabetes risk.
* Diabetes prevalence is higher in older individuals (>40 years).
* For Healthcare Providers
  + Target high-risk groups (High BMI & Glucose) for early screening.
  + Personalized treatment plans for weight loss & lifestyle changes.
* Insurance Companies can adjust premiums based on diabetes risk factors.
* Government & Health Agencies can spread awareness on diet & exercise.