**Healthcare Data – Disease Risk Prediction & Reporting**

## **Project Members & Roles:-**

* **Data Engineer – Devashish Kumar**
* **EDA Specialist – Hrithik Kumar**
* **ML Modeler – Satish**
* **SQL Analyst – Divya Kaushik**
* **BI Developer – Adithya**
* **Report – Divya Kaushik**

## ****Executive Summary****

This project simulates a real-world healthcare analytics scenario. MedXCare Analytics aims to predict patients at high risk of chronic diseases like diabetes and heart disease.

Using EHR data, machine learning models, SQL analysis, and Power BI dashboards, we identified high-risk patients and provided actionable insights for preventive care.

Outcomes include:

1. Identification of high-risk patients using Logistic Regression and Decision Tree models.
2. Segmentation of patients by age, gender, region, and lifestyle factors.
3. Interactive dashboard for hospital decision-makers.
4. Recommendations for targeted health interventions.

## 3. Problem Statement

The healthcare sector generates massive amounts of data through Electronic Health Records (EHRs), but this data often remains underutilized for proactive care. Chronic diseases like diabetes, heart disease, and obesity-related illnesses are increasing globally. Early detection and preventive intervention are key to improving health outcomes.

In this project, we aim to predict patients at high risk of developing chronic diseases using machine learning techniques. The predictions will help hospitals intervene early, and a Power BI dashboard will provide a visual overview of demographic patterns and high-risk segments.

## 4. Tools & Technologies Used

* **Python (pandas, scikit-learn, seaborn):** Data cleaning, model building (Logistic Regression, Decision Tree).
* **SQL :** Patient segmentation, clustering queries.
* **Excel:** Early data profiling, univariate & bivariate analysis.
* **Power BI:** Dashboard creation and visualization.
* **Jupyter / Google Colab:** Coding environment for Python and SQL integration.

## ****5. Execution Summary****

### ****Step 1: Data Cleaning & Feature Engineering****

* Handled missing values using median/mode imputation.
* Normalized numerical features like BMI, blood pressure, and cholesterol.
* Created new features:

1. **BMI Category:** Underweight, Normal, Overweight, Obese
2. **Lifestyle Index:** Combined score from exercise frequency, diet, and smoking status

* Documented assumptions for dataset usage.

### ****Step 2: Exploratory Data Analysis (EDA)****

* Conducted univariate and bivariate analysis in Python and Excel.
* Visualizations included:
* Age vs Disease occurrence
* Gender-wise disease distribution
* Correlation heatmap of lab results
* Observed trends: Older age, higher BMI, and poor lifestyle habits correlate with increased disease risk.

### ****Step 3: Machine Learning Modeling****

### Split dataset: **70% training**, **30% testing**.

* Models used: **Logistic Regression** and **Decision Tree**.
* Hyper-parameter tuning via cross-validation.
* Evaluation metrics:

1. **Accuracy** – Overall correct predictions
2. **Precision & Recall** – Important for minimizing false negatives in healthcare

* **AUC-ROC** – Model’s ability to distinguish high-risk vs low-risk patients
* Feature importance highlighted age, BMI, cholesterol, and exercise frequency as key predictors.

### ****Step 4: SQL Analysis in Google Colab****

* Connecting Colab to MySQL/PostgreSQL
* Queries to find high-risk patients (Age > 60, High BP + Cholesterol)
* Patient segmentation by risk and age groups
* Exported CSV for Power BI
* **Include query & sample outputs**

### ****Step 5: Power BI Dashboard****

Designed visuals:

1. **Age-Risk Map:** Highlights high-risk age groups
2. **Disease Heatmap:** Geographic risk distribution
3. **Risk by Gender & Region:** Interactive filters
4. Dashboard includes filters for gender, lab results, and lifestyle index.
5. Dashboard walkthrough prepared for hospital management.

### ****Step 6: Reporting & Strategic Recommendations****

### Insights from EDA, ML models, SQL queries, and dashboards compiled.

* Recommendations:
* Targeted screening for high-risk age groups
* Lifestyle modification campaigns
* Proactive monitoring of high-risk patients
* Ethical handling of patient data ensured, maintaining privacy and compliance.

## ****Team Contributions****

* **Data Engineer:** Responsible for data cleaning, handling missing values, and feature engineering.
* **EDA Specialist:** Conducted univariate and bivariate analysis, created plots, and identified key patterns.
* **ML Modeler:** Developed Logistic Regression and Decision Tree models, tuned hyperparameters, and evaluated performance.
* **SQL Analyst:** Performed patient segmentation, created risk-based queries, and exported data to CSV for dashboard use.
* **BI Developer:** Designed and implemented the Power BI dashboard, including interactive filters and risk visualizations.
* **Report & Strategy Lead:** Compiled the report, documented findings, and prepared strategic recommendations.

## ****Key Findings****

* Age and BMI are the most significant predictors of disease risk.
* High cholesterol and poor lifestyle habits increase chronic disease risk.
* Certain regions show higher prevalence, guiding targeted interventions.
* Predictive models effectively identify patients for early intervention.

## ****Challenges & Learnings****

* **Challenges:**

1. Incomplete or inconsistent EHR data
2. Class imbalance in disease occurrences
3. Integrating SQL results into BI dashboards

* **Learnings:**

1. Importance of clean, structured data for predictive modeling
2. End-to-end workflow from data cleaning to insights
3. Cross-disciplinary teamwork enhances project effectiveness

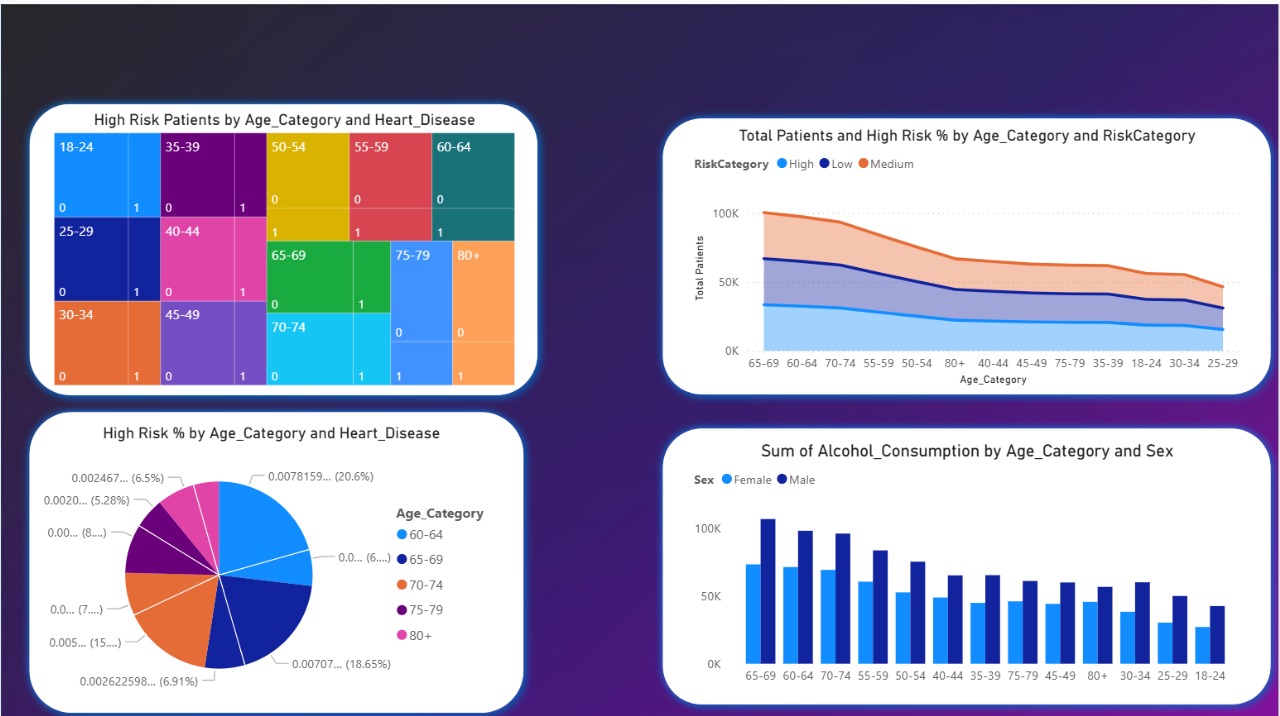
## ****Recommendations****

* Deploy predictive models for real-time risk flagging.
* Implement awareness campaigns for high-risk populations.
* Regularly update EHR data to maintain accuracy.
* Use Power BI dashboards for weekly hospital management reviews.
* Ensure patient data privacy and comply with healthcare regulations.

## Visuals & Screenshots:-

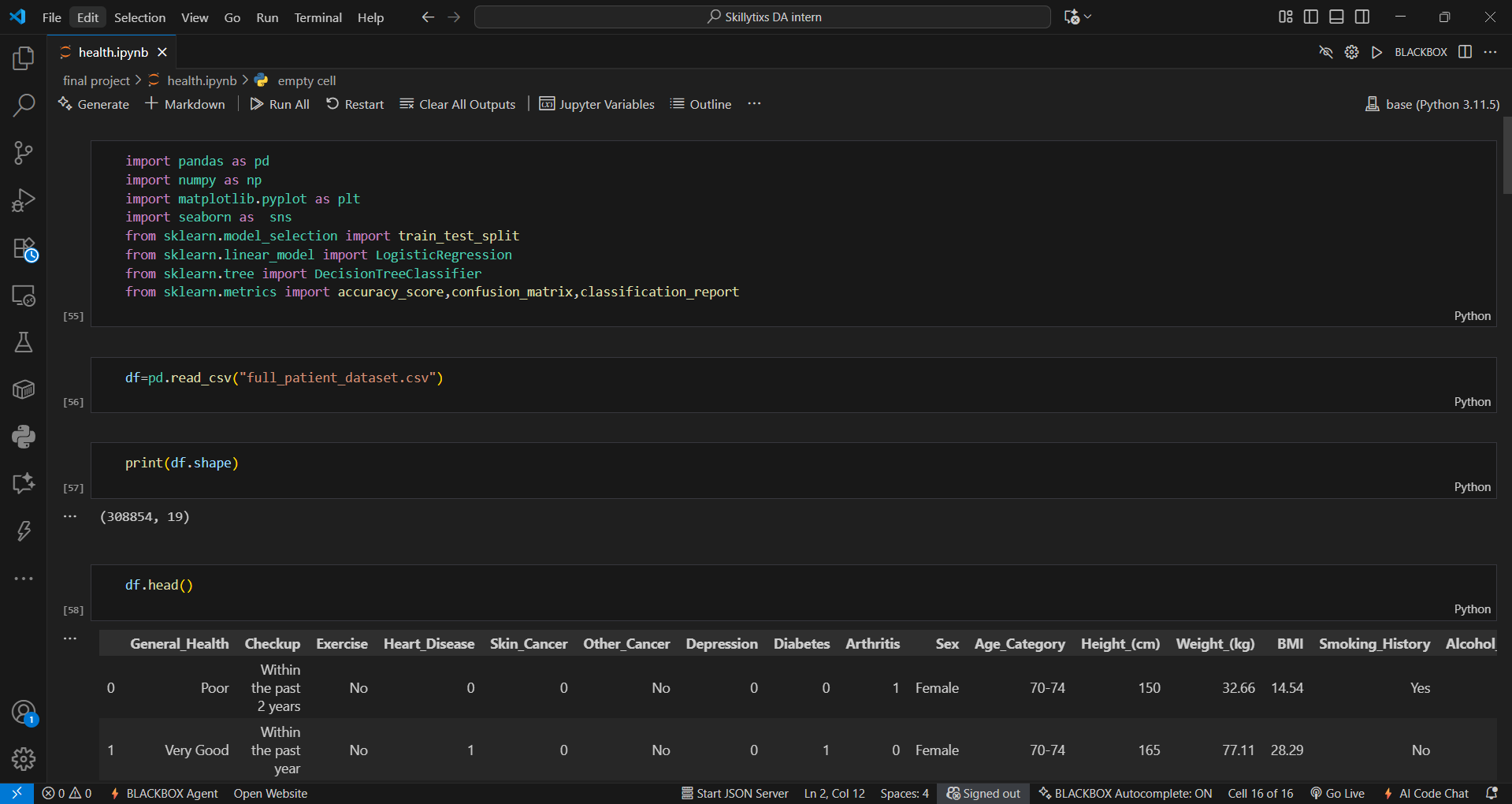
* **Power BI Dashboard**

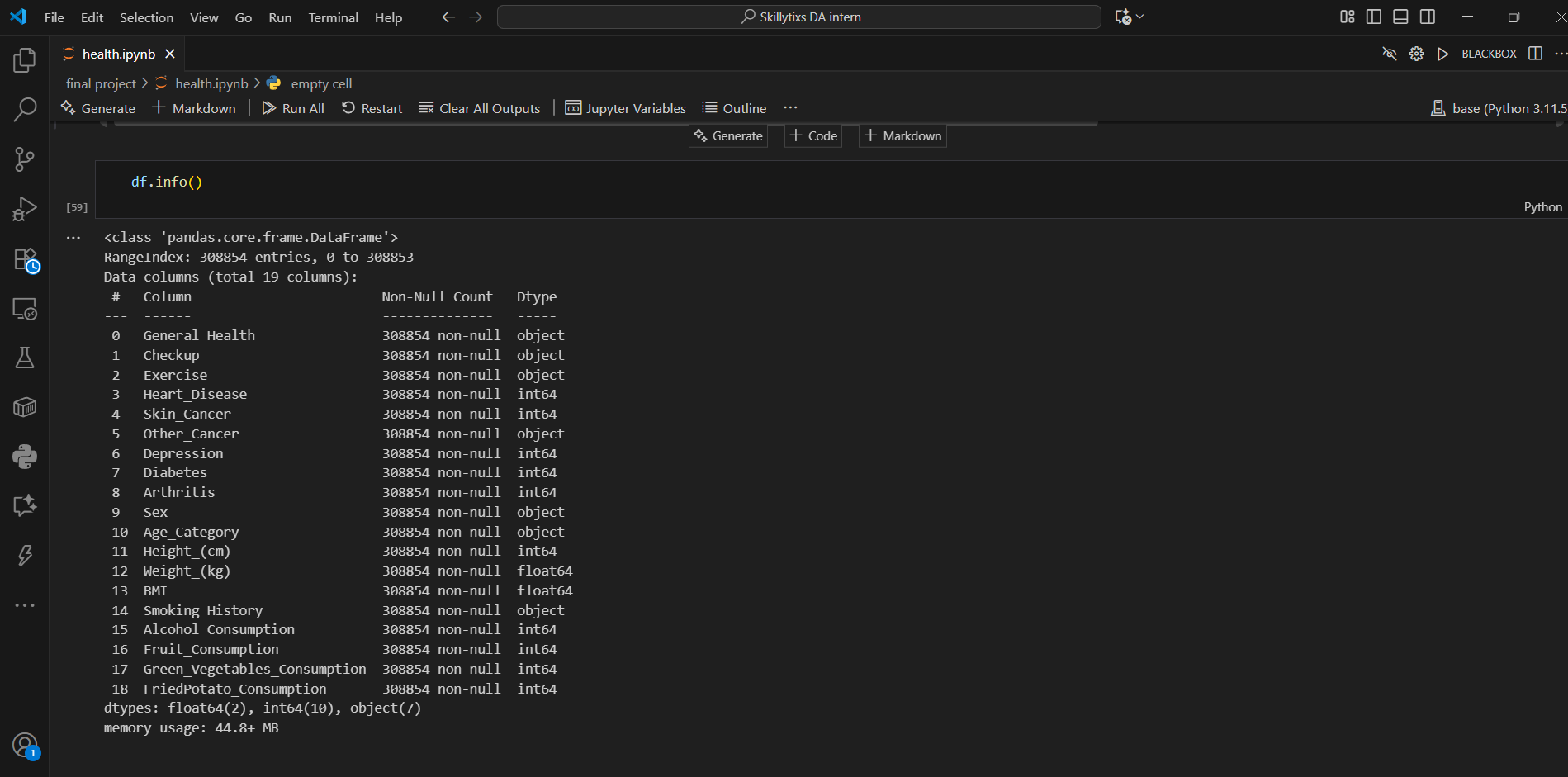


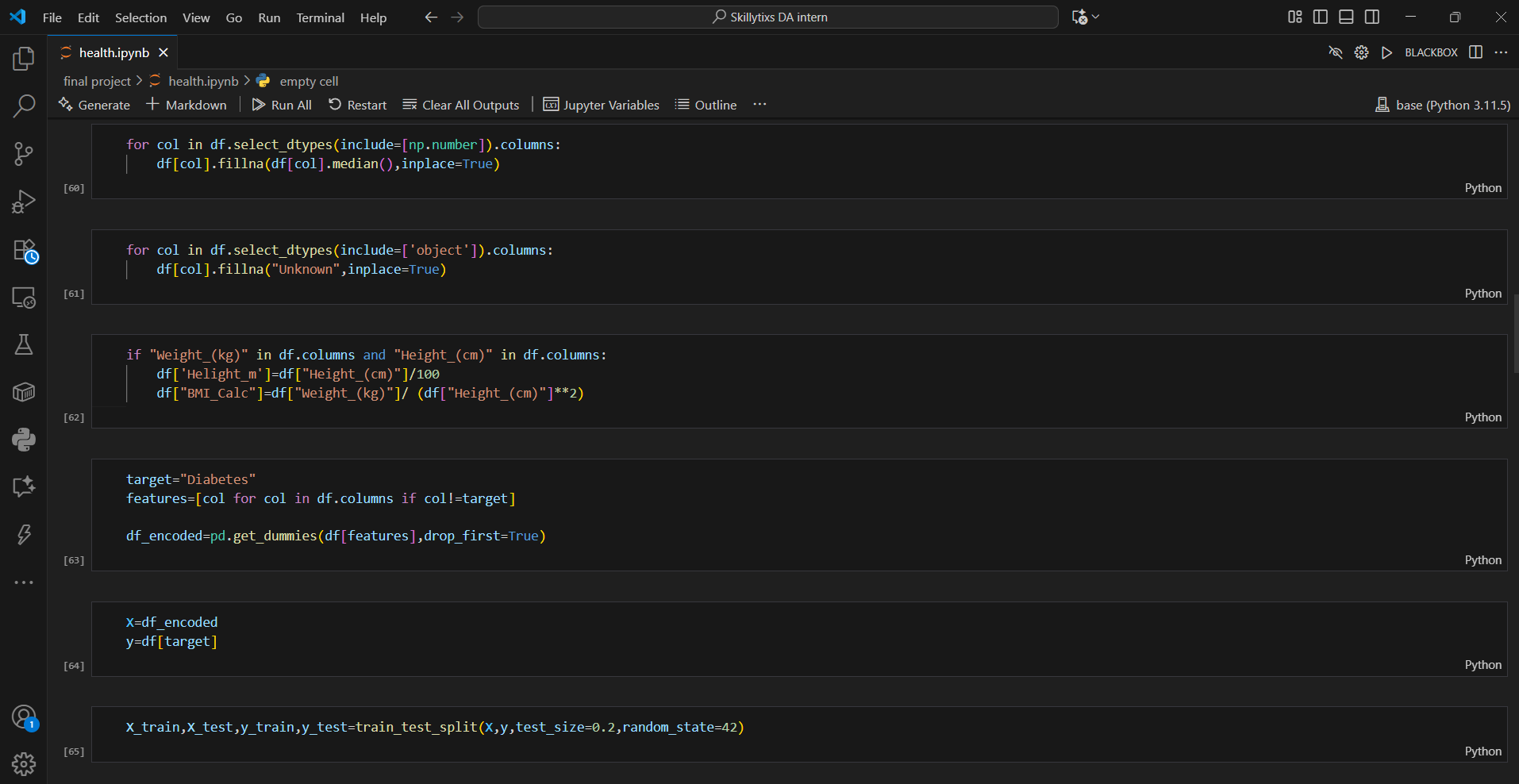


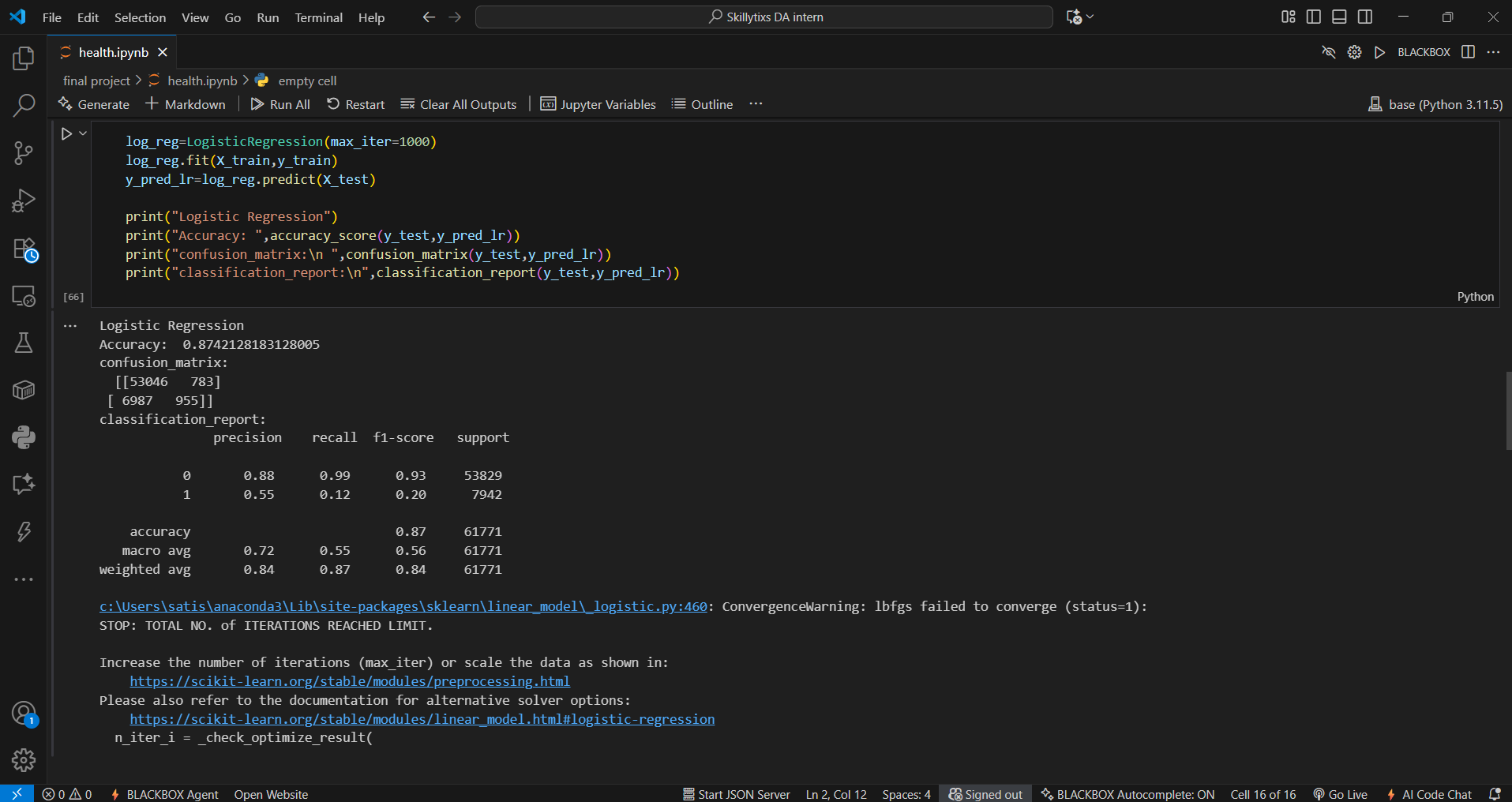
1. Dashboard built on **160K patient records**, avg. risk score **6.31**, with **12K high-risk patients (0.07%)**.
2. Shows **risk categories, disease trends, age–gender breakdown, BMI & lifestyle factors**.
3. Insights: **50% high-risk from one disease**, **females slightly higher risk**, **25–34 & 65+ more vulnerable**.
4. Enables **early risk prediction, preventive care, and better resource allocation**.

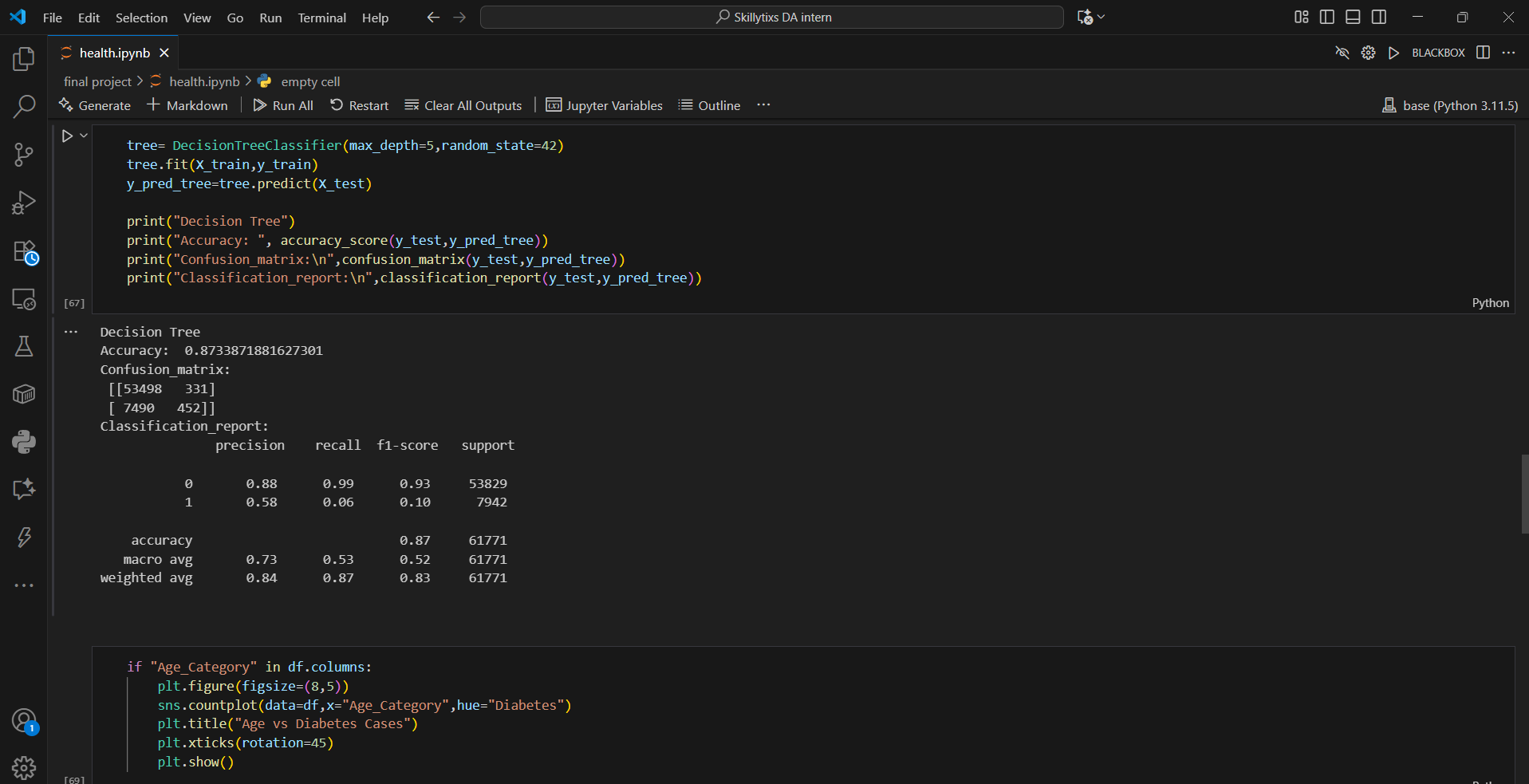
* **Python Code Supporting the Dashboard**

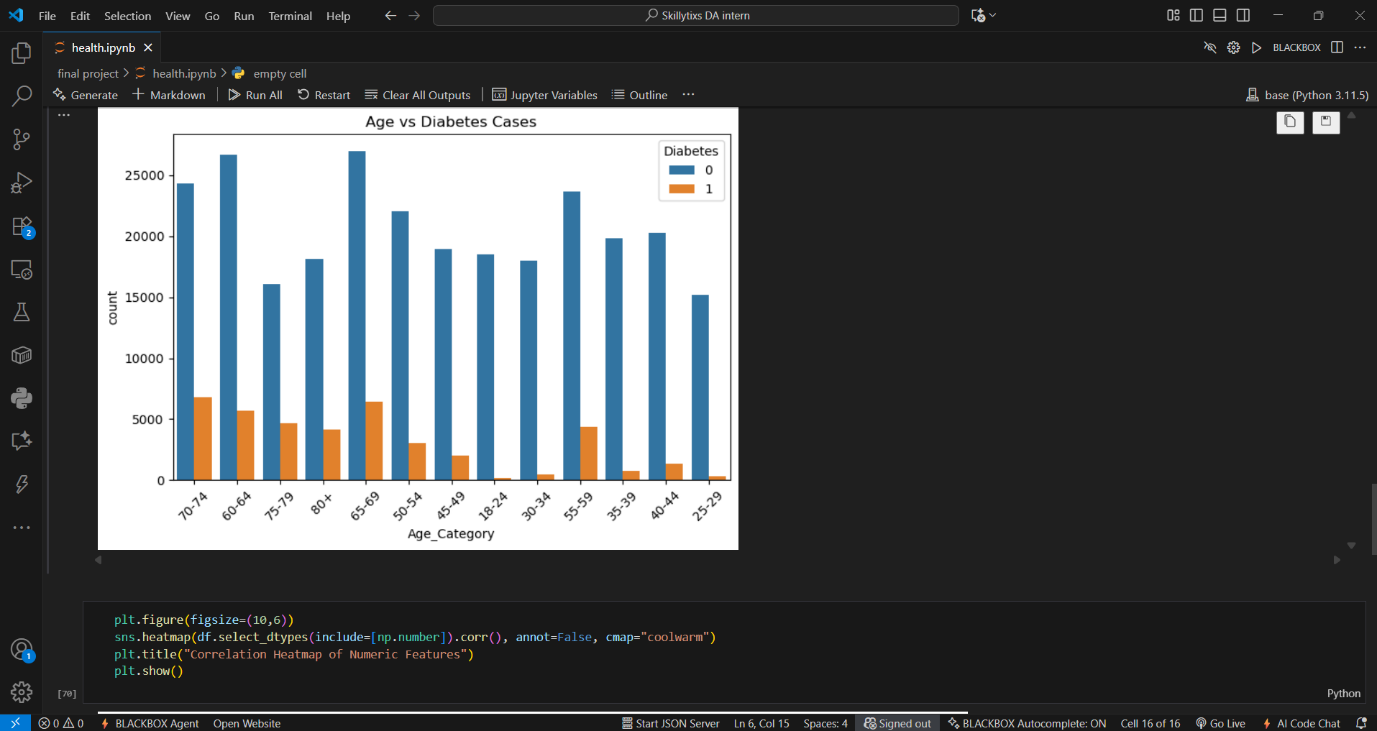


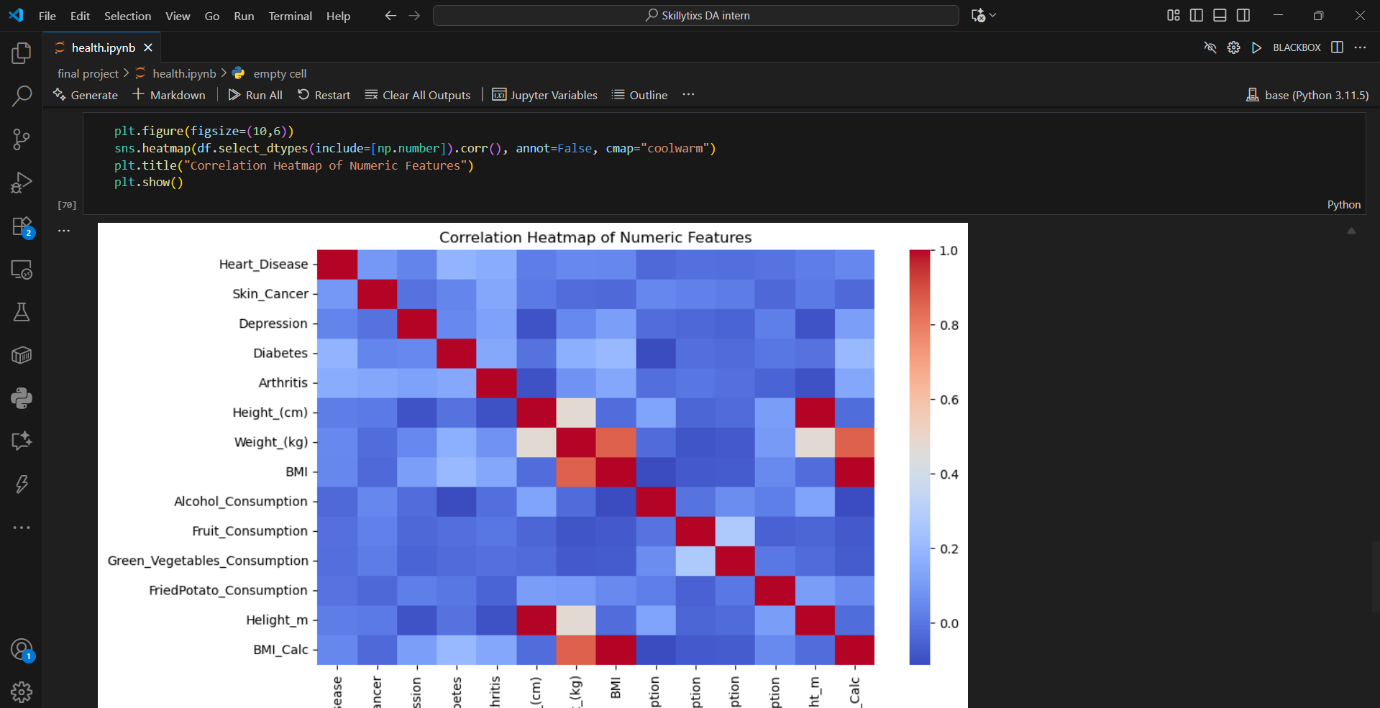






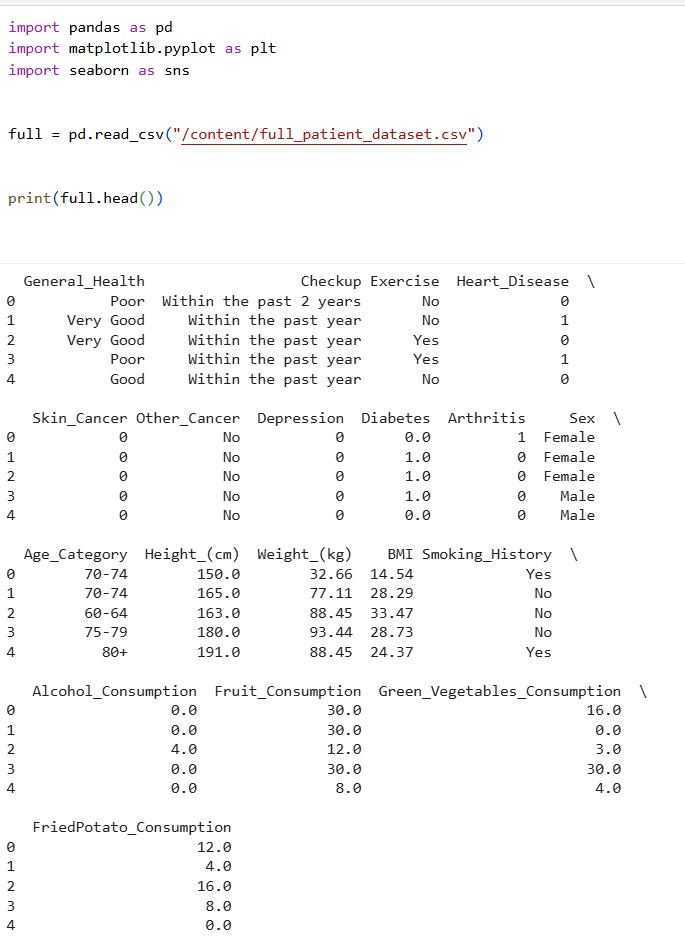




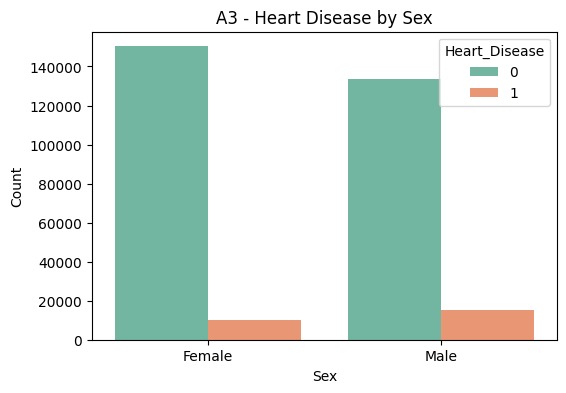
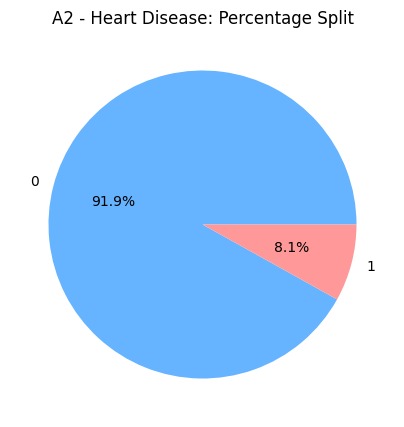
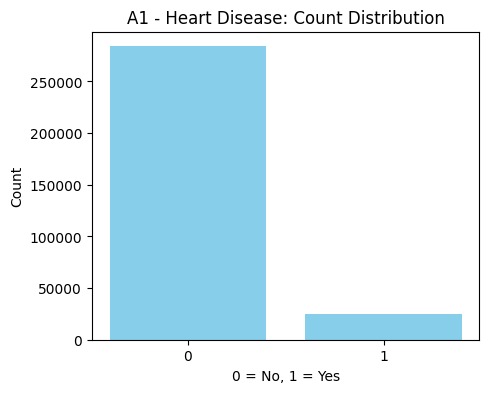
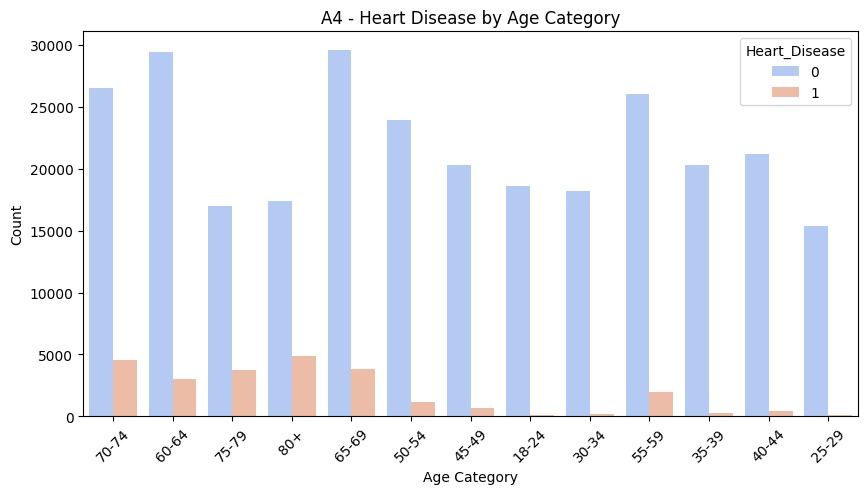
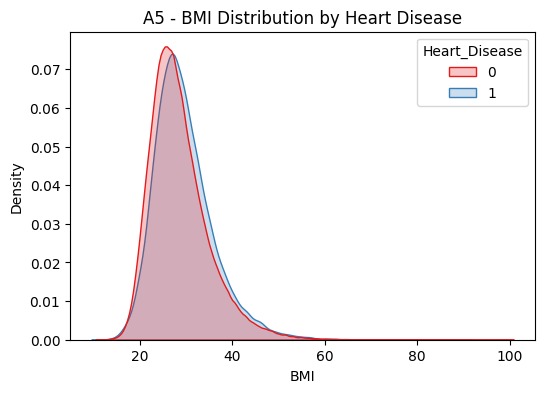


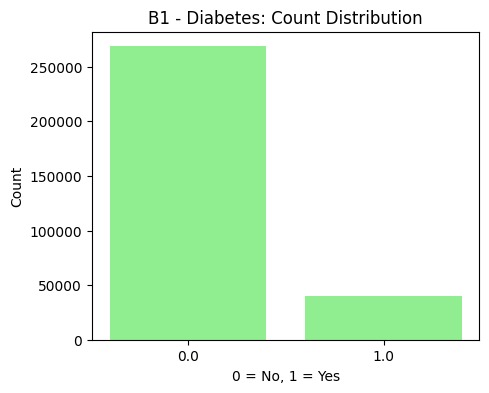
1. **Imported libraries** for data manipulation, visualization, and modeling.
2. **Loaded and inspected** the dataset; checked for missing values and target distribution.
3. Analyzed **feature relationships** using correlations and heatmaps.
4. **Split data** into training and testing sets.
5. Trained **Logistic Regression** and **Decision Tree** models and made predictions.
6. **Evaluated model performance** using accuracy, confusion matrix, and classification report.
7. Identified **important features** from the Decision Tree for insights.

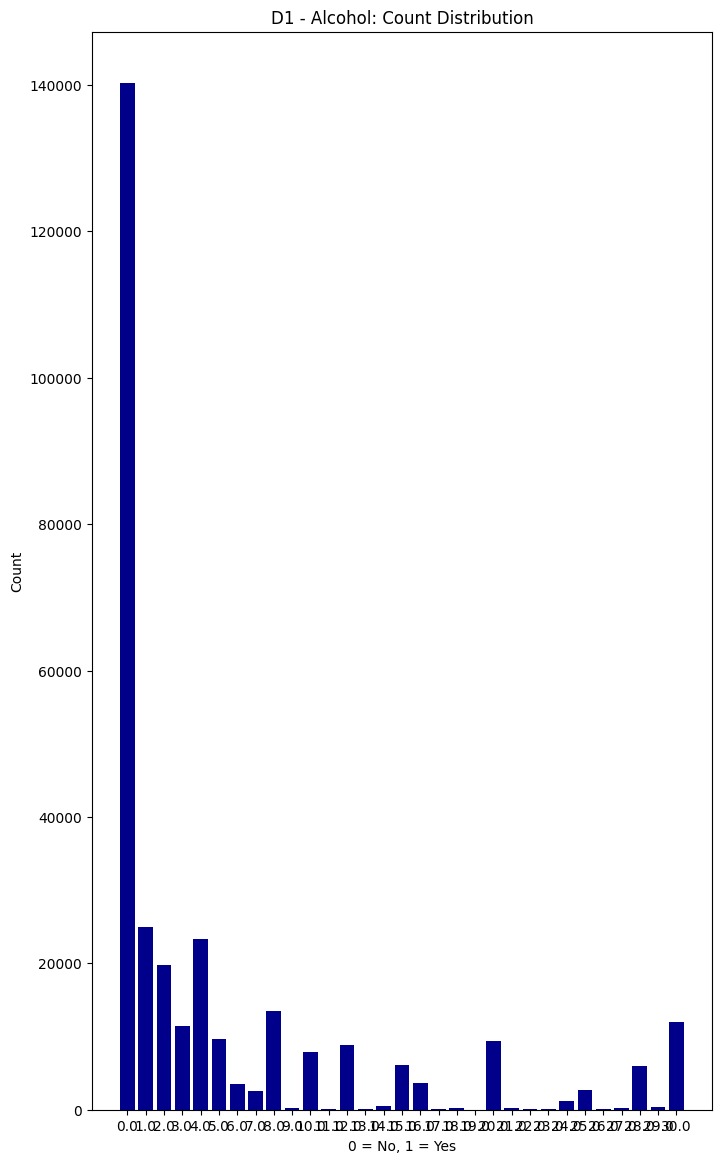
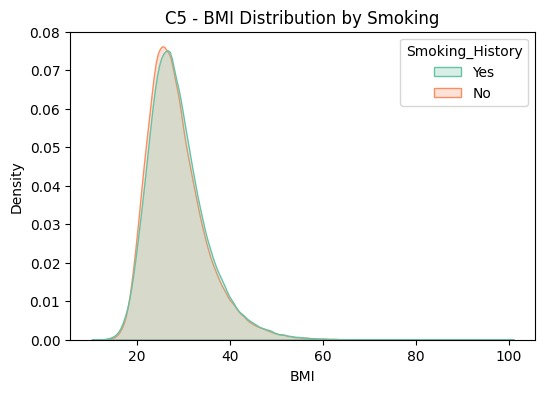
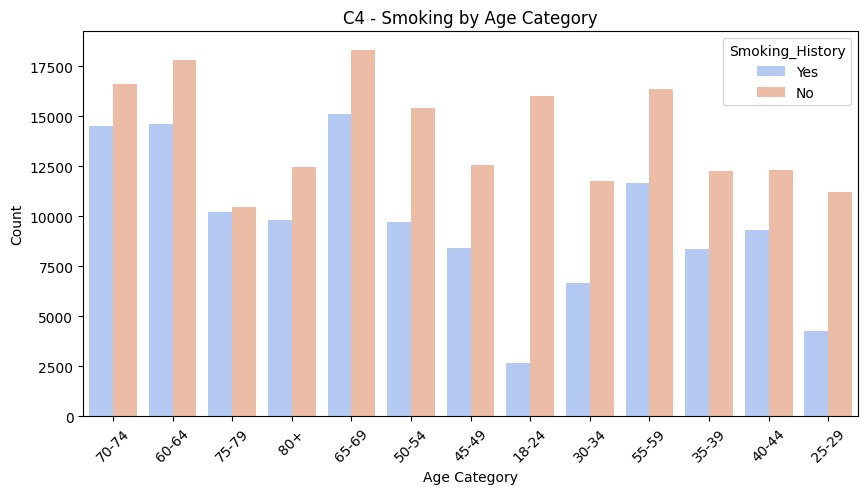
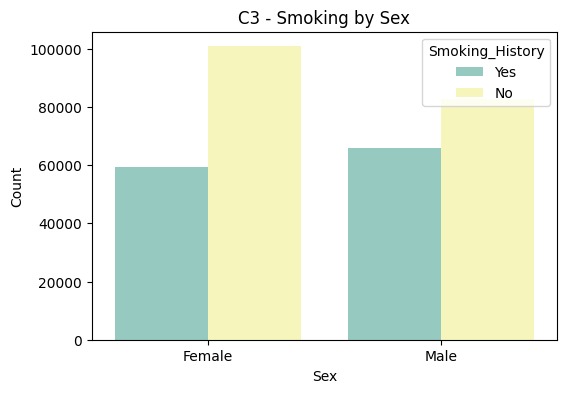
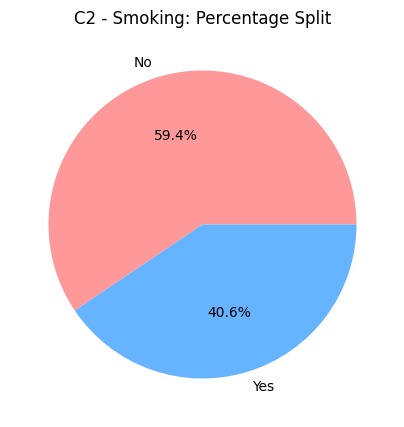
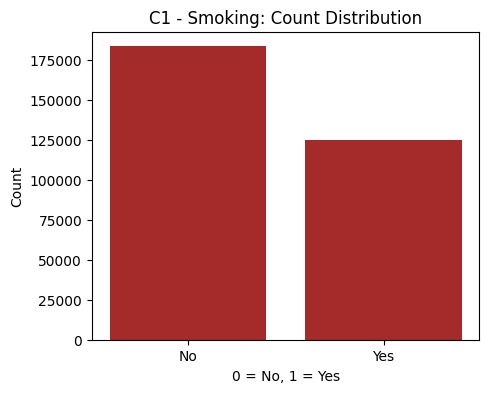
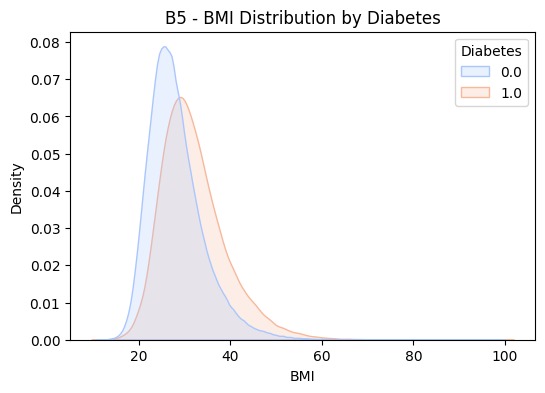
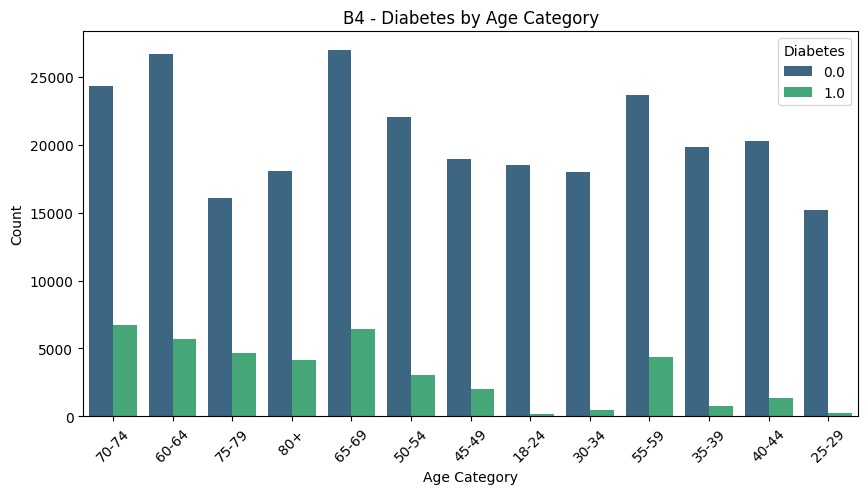
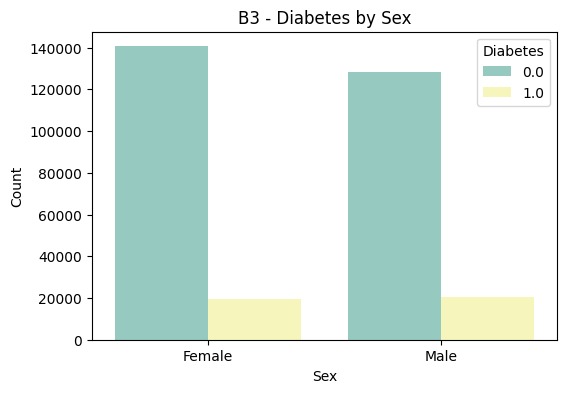
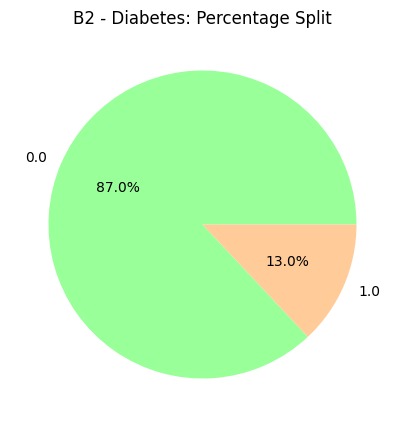
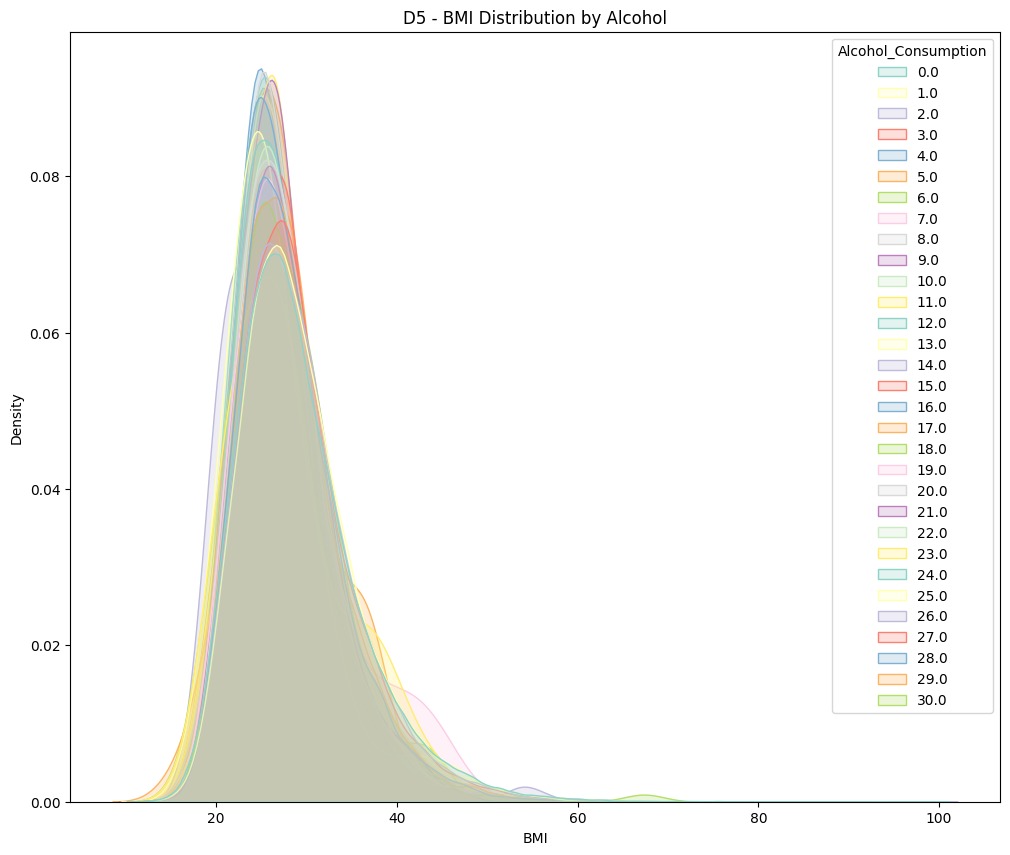
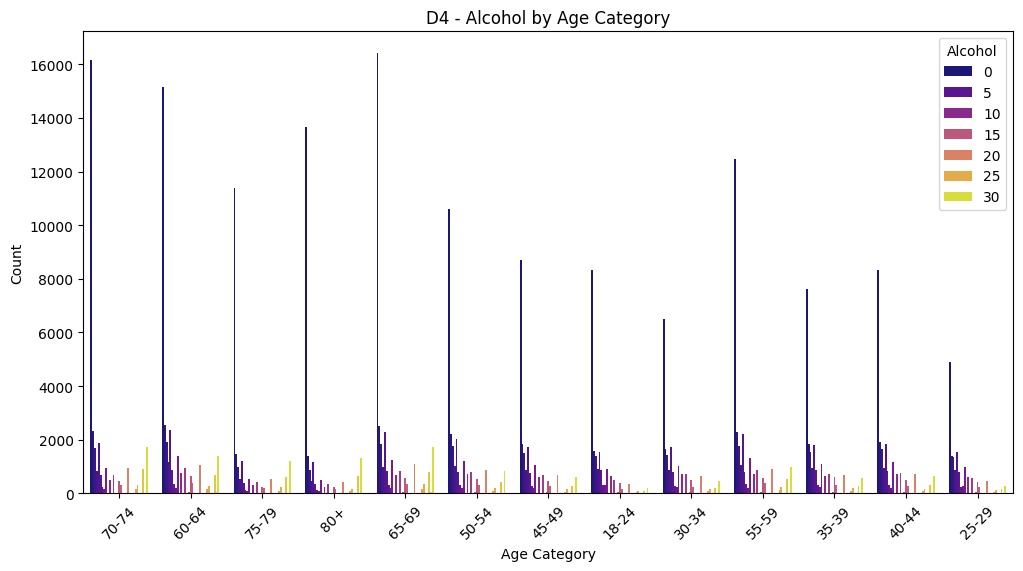
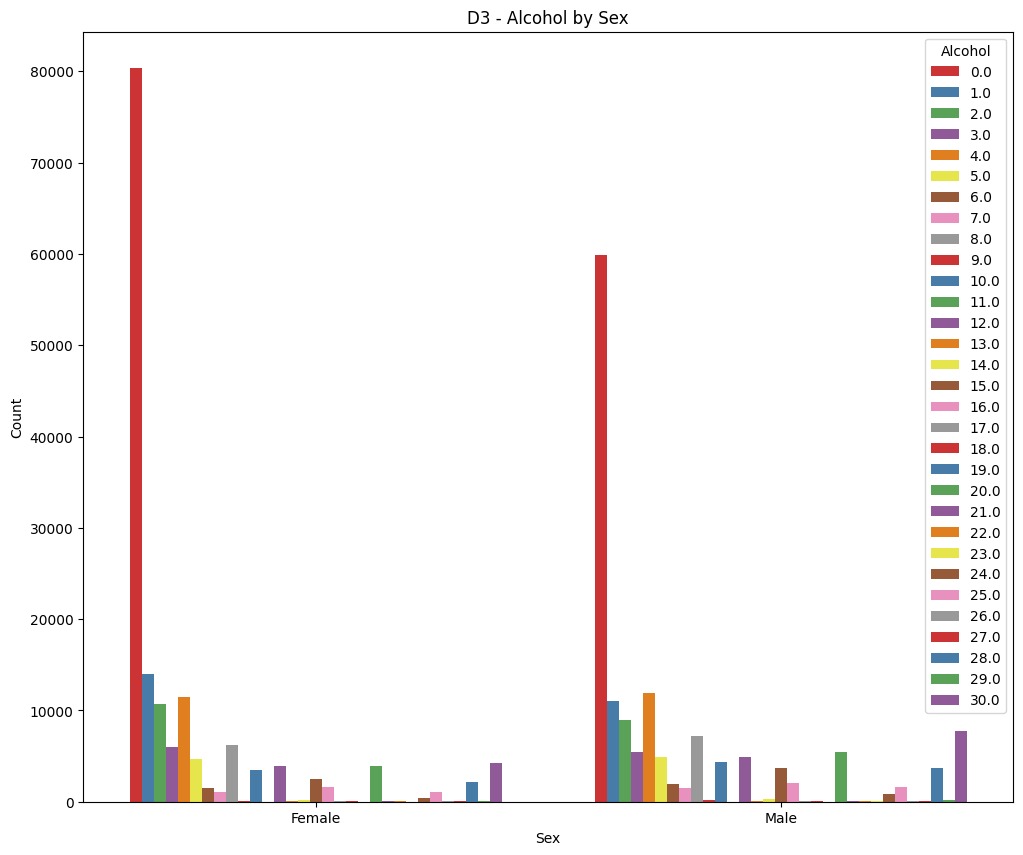
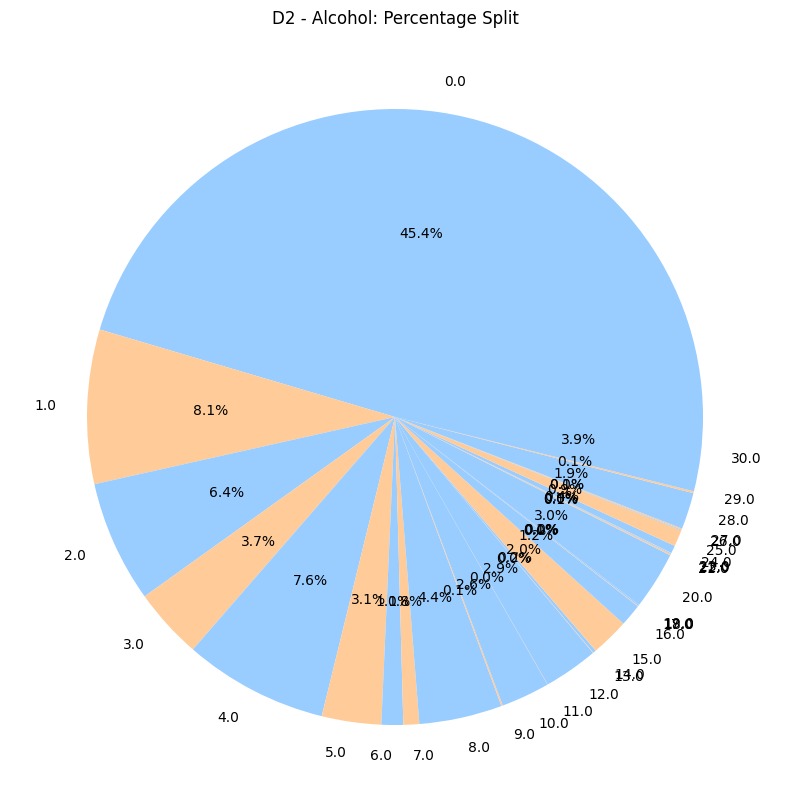
* **Visual EDA: Patient Risk and Health Trends**





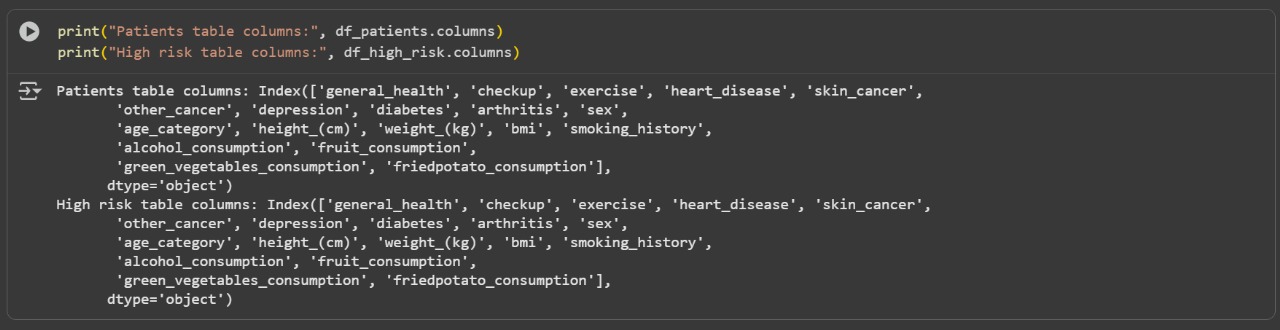
  

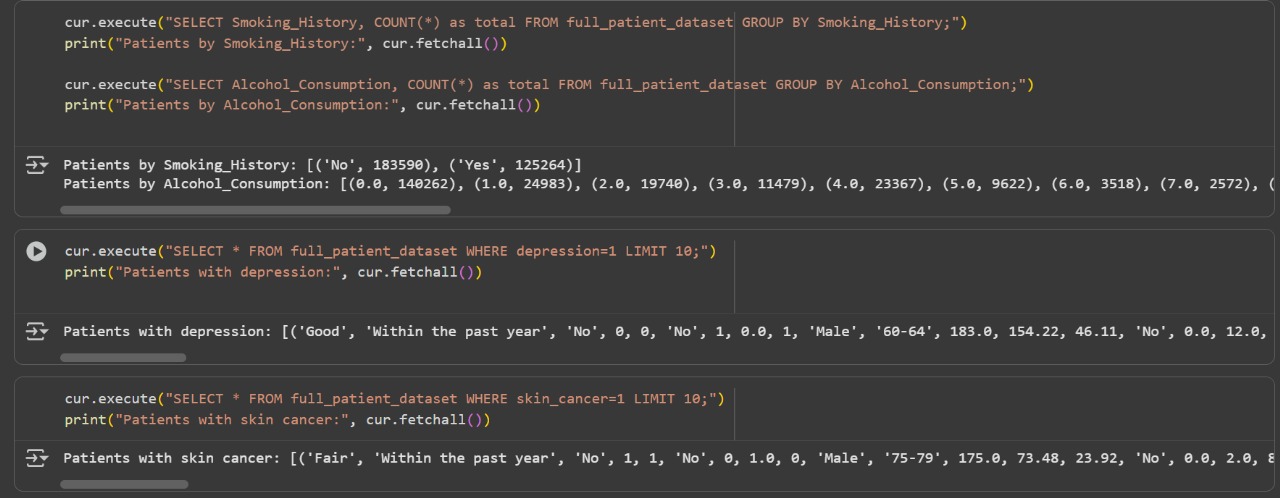


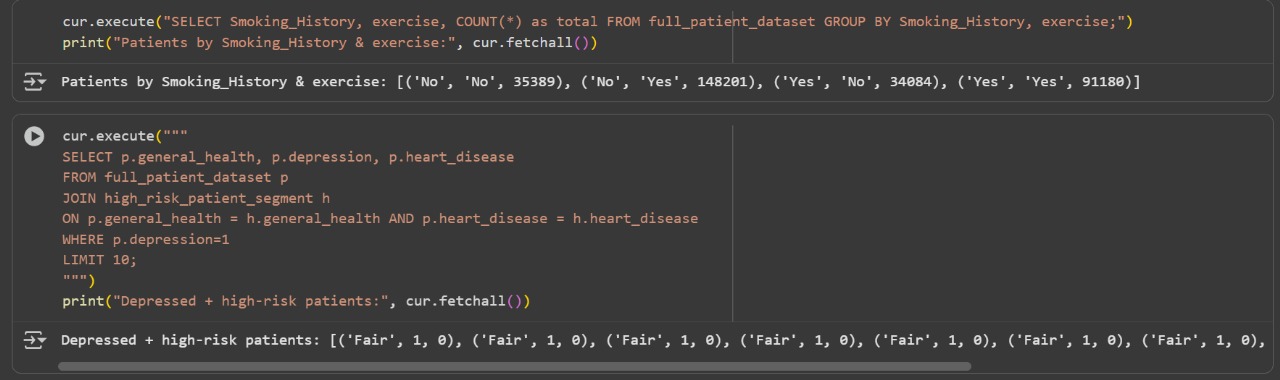
****

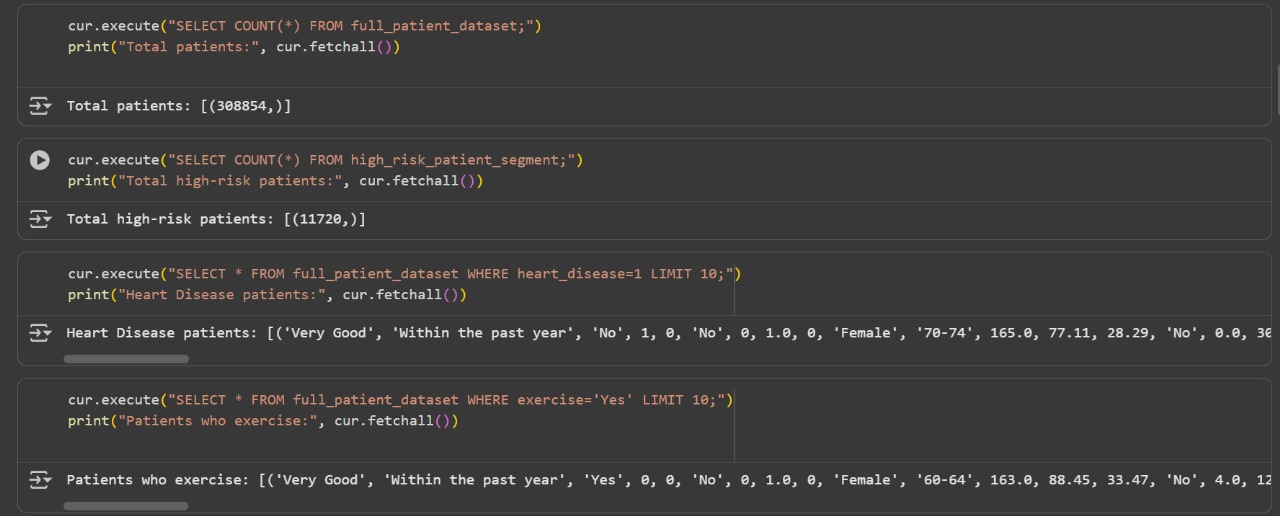
* **SQL-Driven Analysis of Healthcare Dataset**

****

****

****

****

****

1. Established the **patient\_data table** structure and imported the dataset.
2. Identified **high-risk patients** based on **age**, **BMI**, and **heart disease**.
3. Conducted **patient counts** by **age group** and assessed **risk prevalence**.
4. Performed **gender-based analysis** of major **health conditions**.
5. Computed average **BMI**, **alcohol consumption**, and **smoking statistics** by gender.
6. Classified patients into **High Risk** and **Low/Medium Risk** categories.
7. Detected **missing lifestyle information** to ensure **data completeness**.
8. Extracted **high-risk patient data** for **Power BI visualization** and reporting.