Heart Risk Project; Data Analysis

By: Mahsa Nafei, Jesús Hernández, Fernando Lopez, and Alexandru Arnautu

Photos provided by Pexels

Heart Risk Data Analysis

Overview

Our repository combines two data sources and API Coordinates to predict heart attack risk using machine learning and visualization analysis. It features a Flask-based web application for risk prediction, a data processing script using Postgresql in Amazon RDS, and employs Spark for data cleaning and preparation. We employed data Plotly, Pandas and Folium for visualization and analysis.

Objective

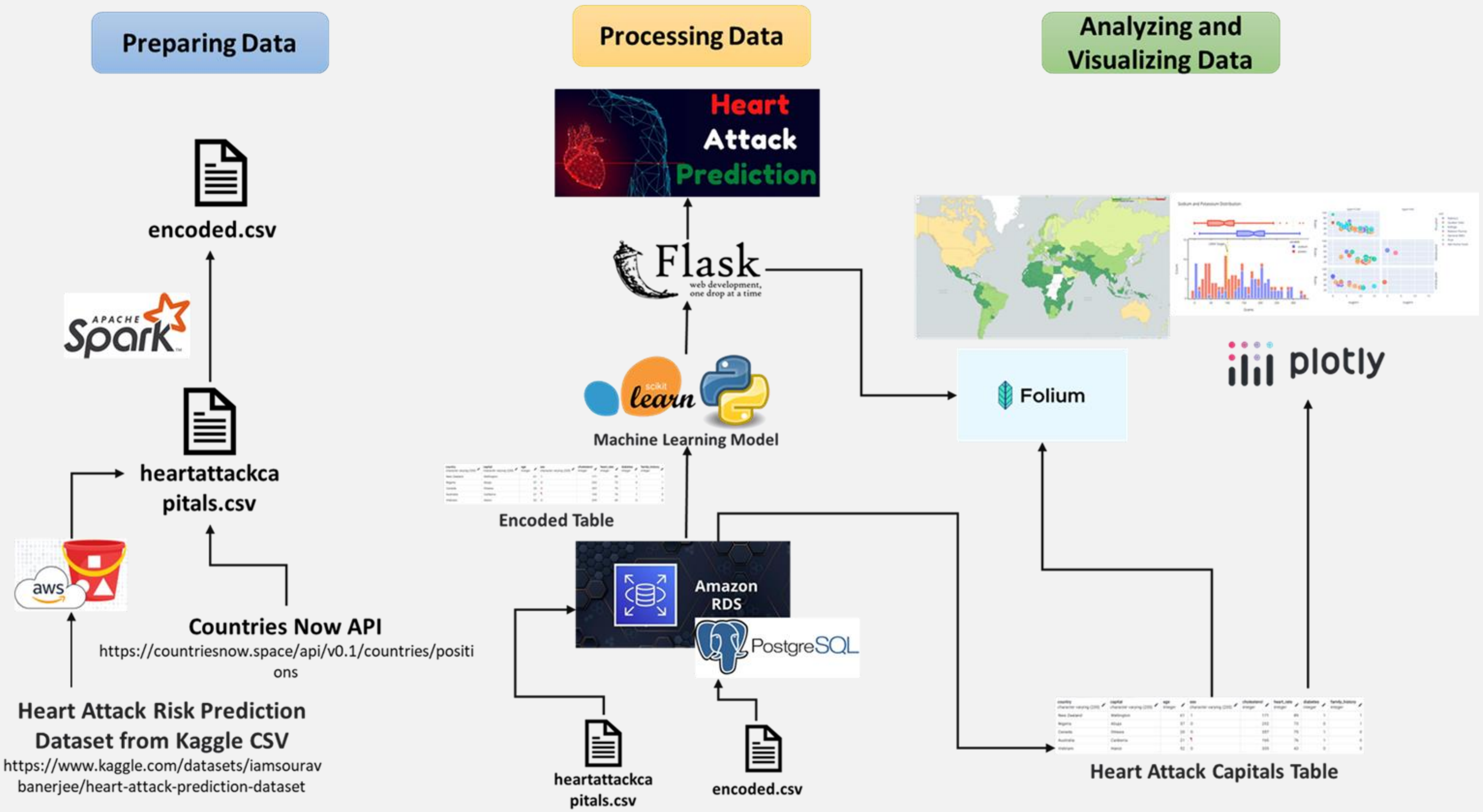
This project focuses on using a dataset to create a predictive model for assessing an individual's risk of having a heart attack. It aims to understand how specific lifestyle factors influence the probability of heart risk

- Age
- Blood Pressure
- Cholesterol Levels
- Diabetic Status
- Hours of Sleep
- Medication Use
- Obesity
- Physical Activity
- Smoking
- Triglycerides

Our project seeks to examine the data outlined above to help insurance companies provide more detail specific health programs for patients affected by heart risk, by looking at related factors. A critical aspect of the project is ensuring fairness and reducing bias in predictions across different demographic groups, and instead focused on creating a global pattern rather than continent/country specific ones. Our goals throughout the making of this analysis, was to focus on improving risk assessment accuracy through supervised machine learning to support informed decision-making in heart attack risk evaluation.

Photos provided by Pexels

Data Collection and Preprocessing



Preparing Data

Processing Data

Analyzing and Visualizing Data

encoded.csv

heartattackcapitals.csv

Countries Now API
https://countriesnow.space/api/v0.1/countries/positions

Heart Attack Risk Prediction Dataset from Kaggle CSV
https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset

Heart Attack Prediction

Flask web development, one drop at a time

Machine Learning Model

Encoded Table

Amazon RDS

PostgreSQL

heartattackcapitals.csv

encoded.csv

Folium

plotly

Heart Attack Capitals Table

Preparing the Data

The heart Attack Risk Prediction dataset was sourced from Kaggle in CSV format. Geographical coordinates were obtained via Now API. Data is encoded using Amazon RDS and processed with Apache Spark.

Processing the Data

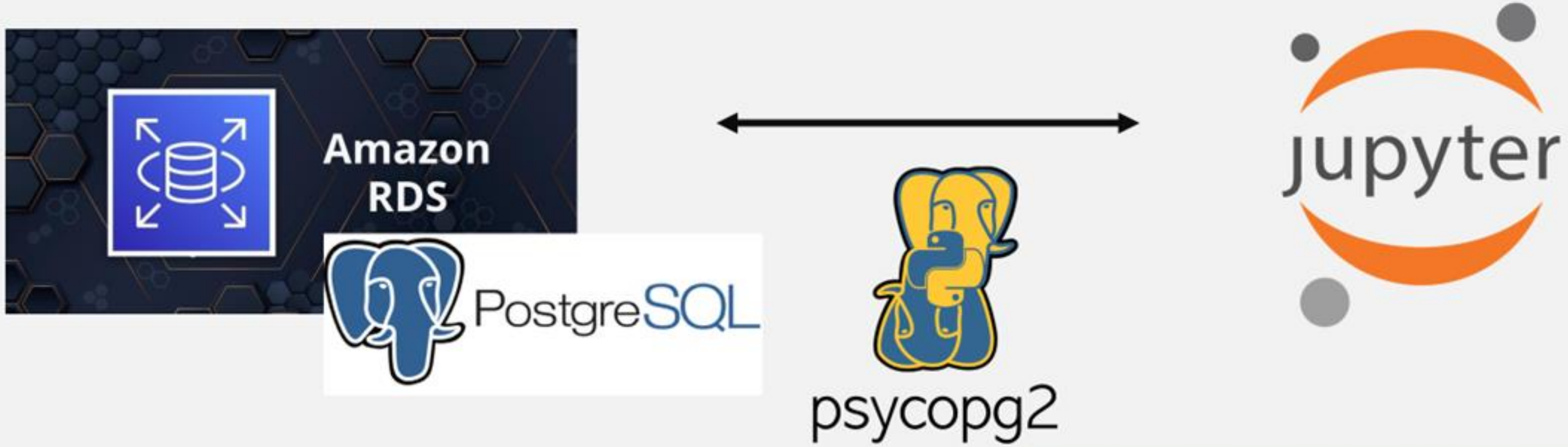
Refined datasets, including Heart Attack Capitals and Encoded CSV, were analyzed using PostgreSQL on Amazon RDS and an optimal machine learning model for Flask's Heart Attack Prediction app..

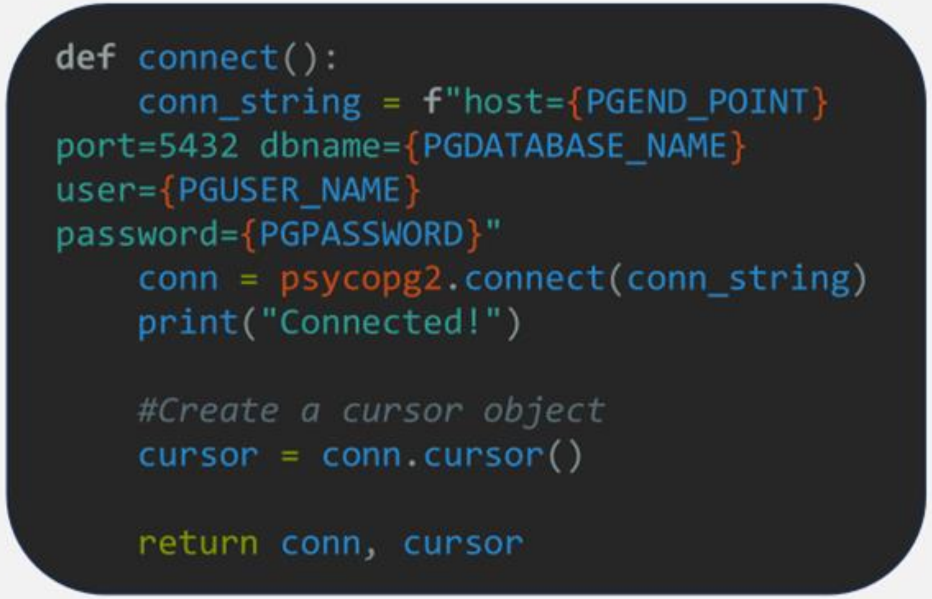
Analyzing and Visualizing the Data

Analytical conclusions are communicated through detailed visualizations crafted with Folium, Pandas, and Plotly.

Photos provided by Pexels

AWS -Relational Data Base Services

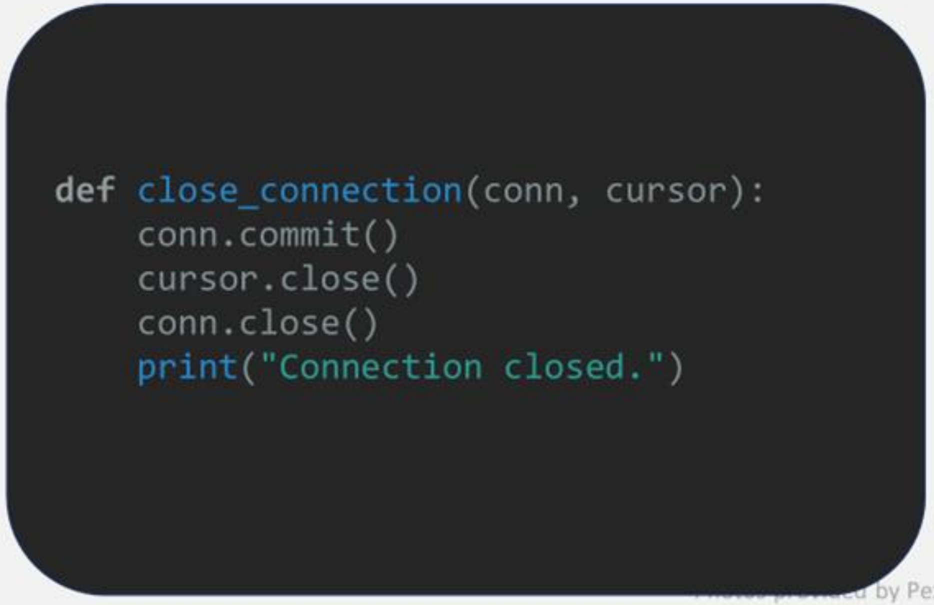




```
def connect():
    conn_string = f"host={PGEND_POINT}
port=5432 dbname={PGDATABASE_NAME}
user={PGUSER_NAME}
password={PGPASSWORD}"
    conn = psycopg2.connect(conn_string)
    print("Connected!")

    #Create a cursor object
    cursor = conn.cursor()

    return conn, cursor
```

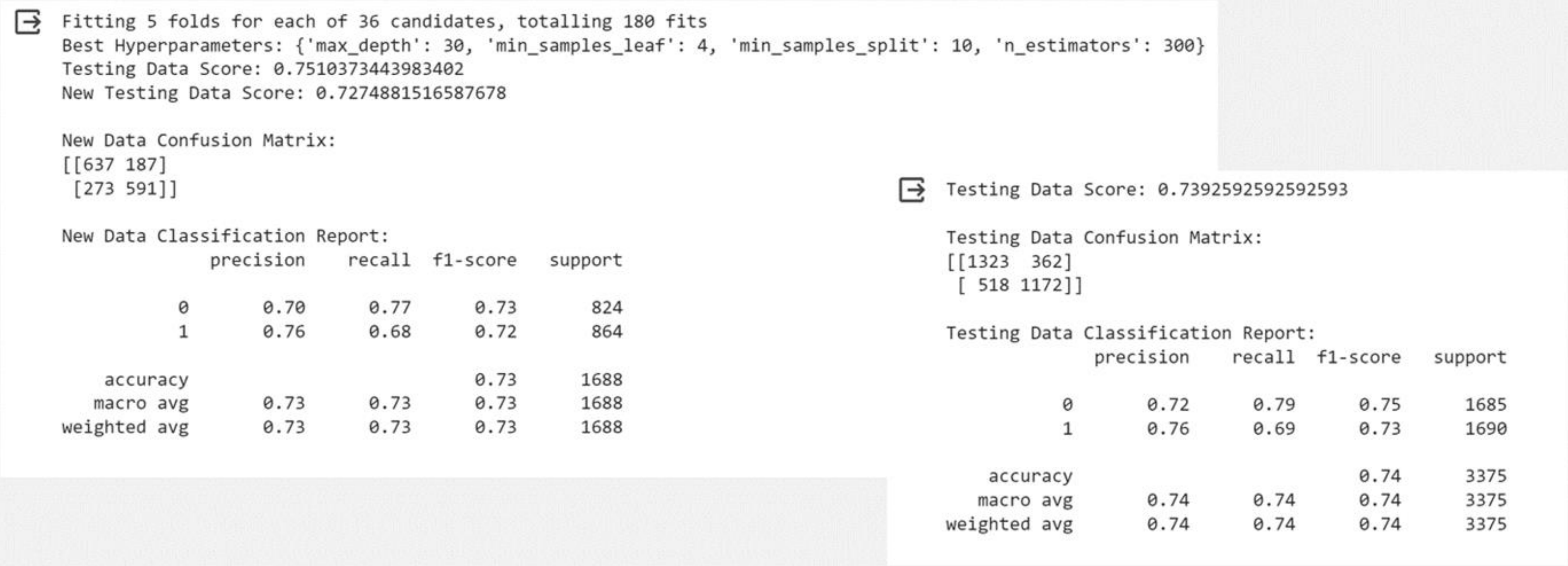


```
def close_connection(conn, cursor):
    conn.commit()
    cursor.close()
    conn.close()
    print("Connection closed.")
```

Photos provided by Pexels

Best Prediction Model

Data Resampling with RandomOverSampler Grid serach and RandomForestClassifier



```
Fitting 5 folds for each of 36 candidates, totalling 180 fits
Best Hyperparameters: {'max_depth': 30, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 300}
Testing Data Score: 0.7510373443983402
New Testing Data Score: 0.7274881516587678

New Data Confusion Matrix:
[[637 187]
 [273 591]]

New Data Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.77      0.73       824
           1       0.76      0.68      0.72       864

    accuracy                           0.73      1688
   macro avg       0.73      0.73      0.73      1688
weighted avg       0.73      0.73      0.73      1688
```

```
Testing Data Score: 0.7392592592592593

Testing Data Confusion Matrix:
[[1323  362]
 [ 518 1172]]

Testing Data Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.79      0.75      1685
           1       0.76      0.69      0.73      1690

    accuracy                           0.74      3375
   macro avg       0.74      0.74      0.74      3375
weighted avg       0.74      0.74      0.74      3375
```

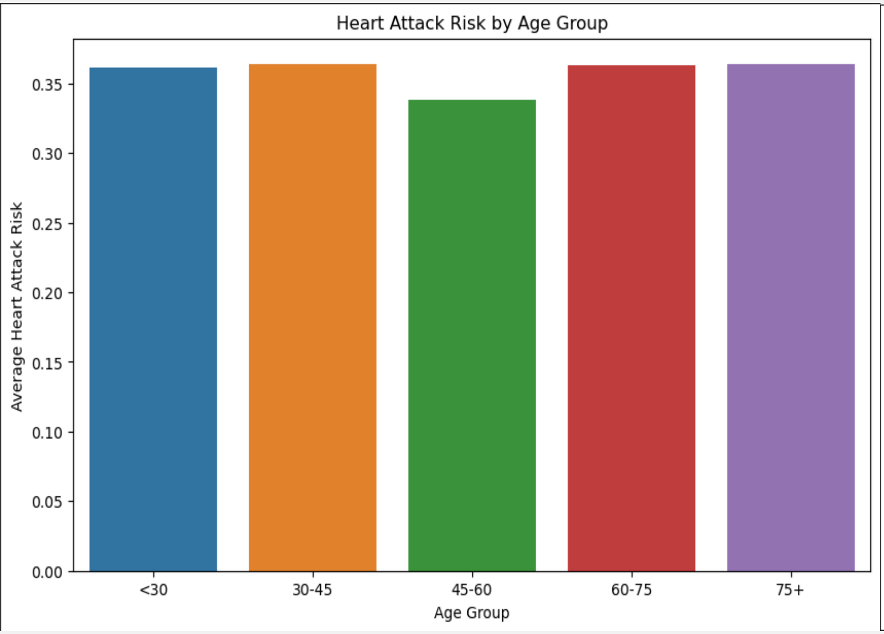
Photos provided by Pexels

Demo
Heart Attack Risk Prediction App

Photos provided by Pexels

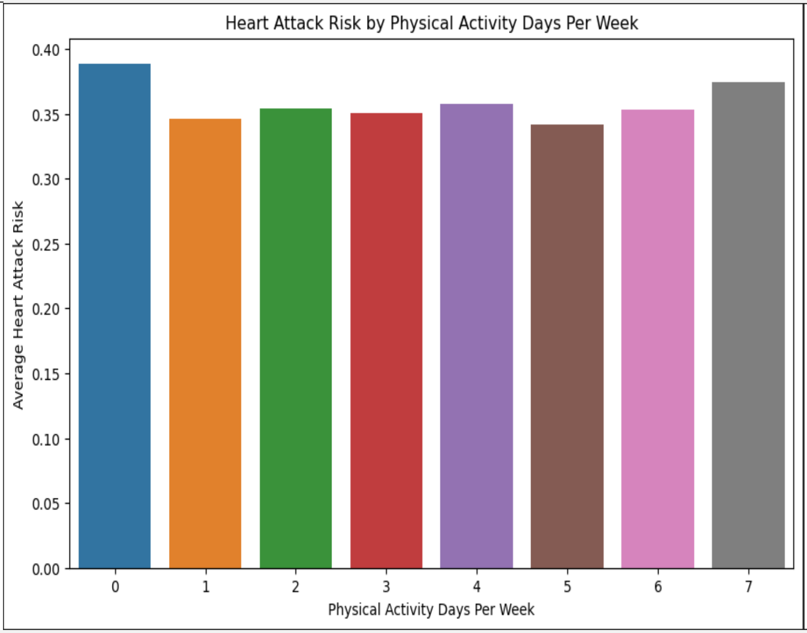
Data Analysis and Visualizations

Heart Attack Risk by Age Group



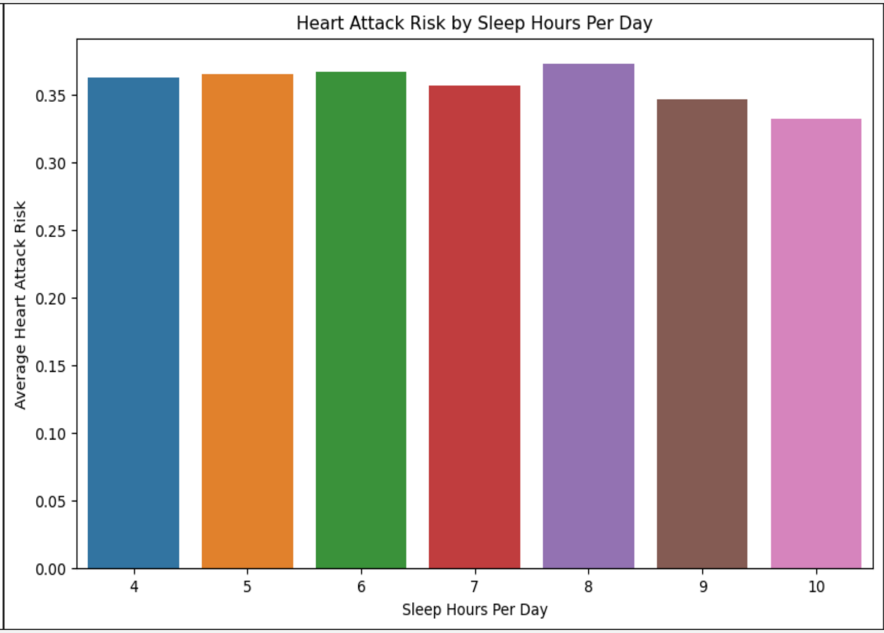
The bar graph provided illustrates a comparative analysis of heart attack risk across different age groups. It highlights that individuals aged 75 and above are at an elevated risk of experiencing a heart attack. Conversely, the graph also indicates that people between the ages of 45 to 60 exhibit the lowest risk of such cardiac events.

Heart Attack Risk by Physical Activity Days Per Week



The bar chart indicates that individuals inactive throughout the week face the highest heart attack risk, while those active daily are also at a notable risk. People with five active days per week show the least risk for heart attacks.

Heart Attack Risk by Hours of Sleep per day

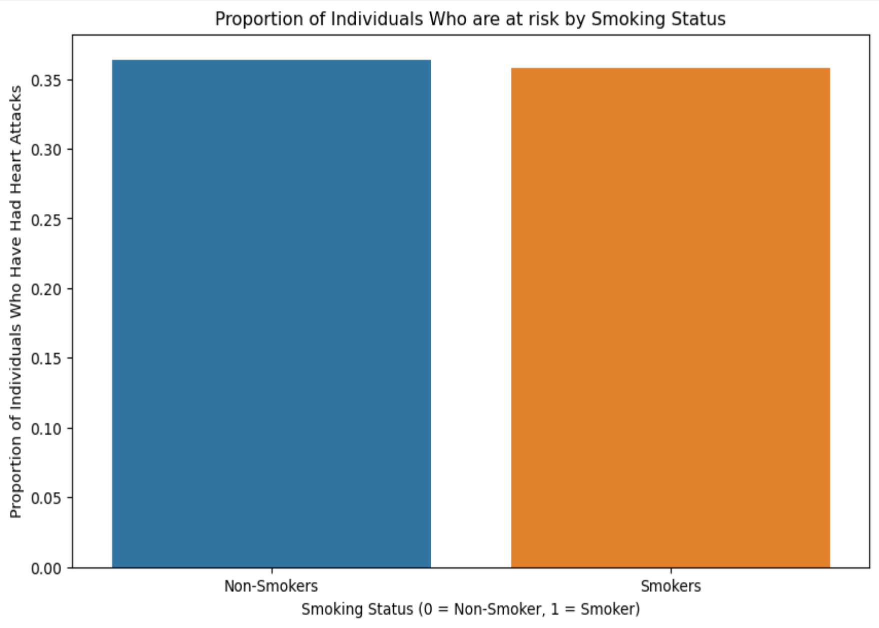


The bar graph suggests that a daily rest of 10 hours minimizes the risk of heart attacks. It also notes that an 8-hour sleep duration appears to have the highest associated risk, potentially attributable to it being the most common sleep duration.

Photos provided by Pexels

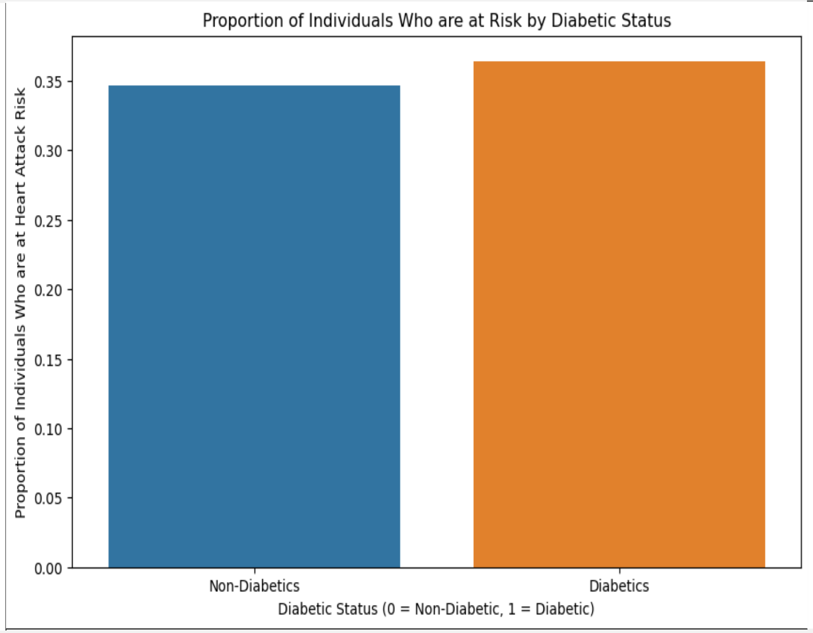
Data Analysis and Visualizations

Proportion of Individuals who are at risk by Smoking Status



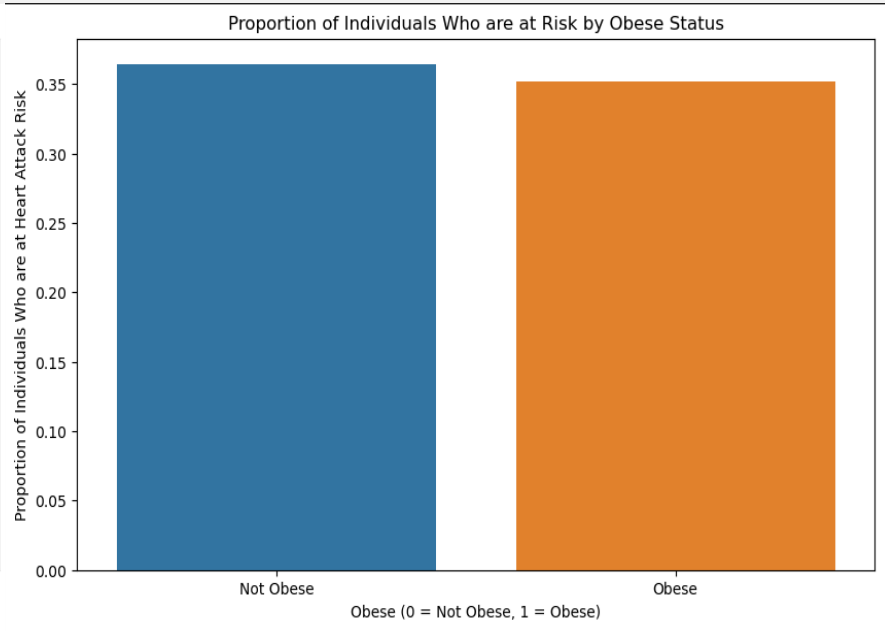
The graph displays a surprising trend: smokers seemingly have a reduced heart attack risk. This could be an instance of the "smoker's paradox," where smokers may exhibit a misleading resilience to heart attacks due to variables not captured in the data.

Proportion of Individuals who are at risk by Diabetic Status



The graph demonstrates that individuals with diabetes have an elevated risk of experiencing heart attacks.

Proportion of Individuals who are at risk by Obesity Status

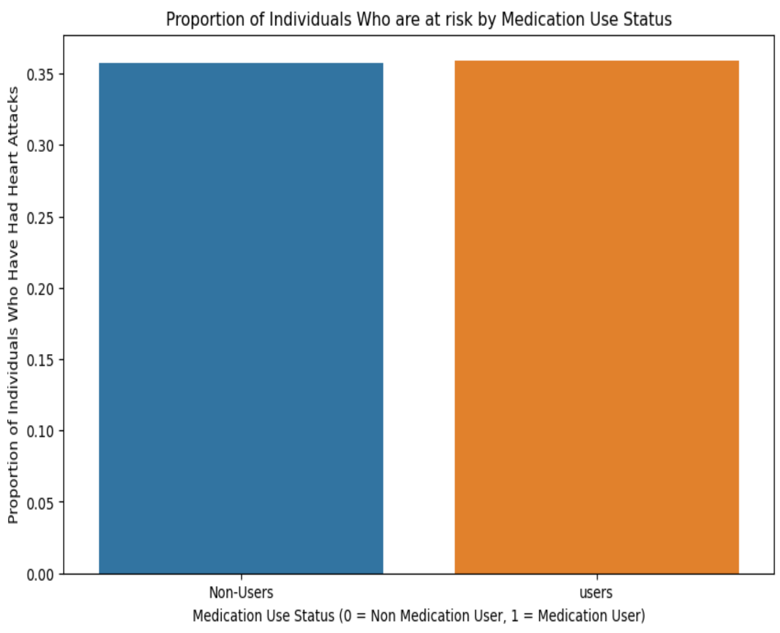


The graph indicates that obese individuals appear to have a lower risk of heart attacks, possibly due to protective factors or more intensive medical treatments often given to overweight patients.

Photos provided by Pexels

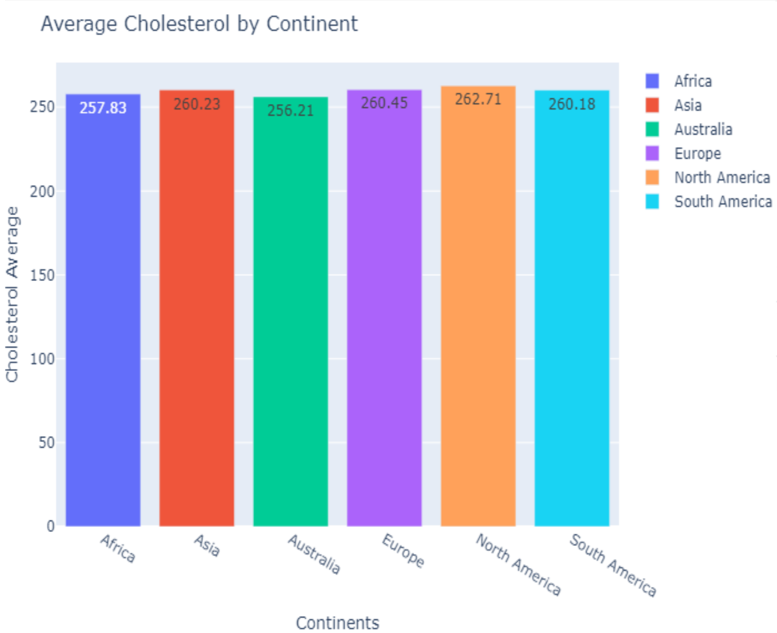
Data Analysis and Visualizations

Proportion of patients at risk who use Medication



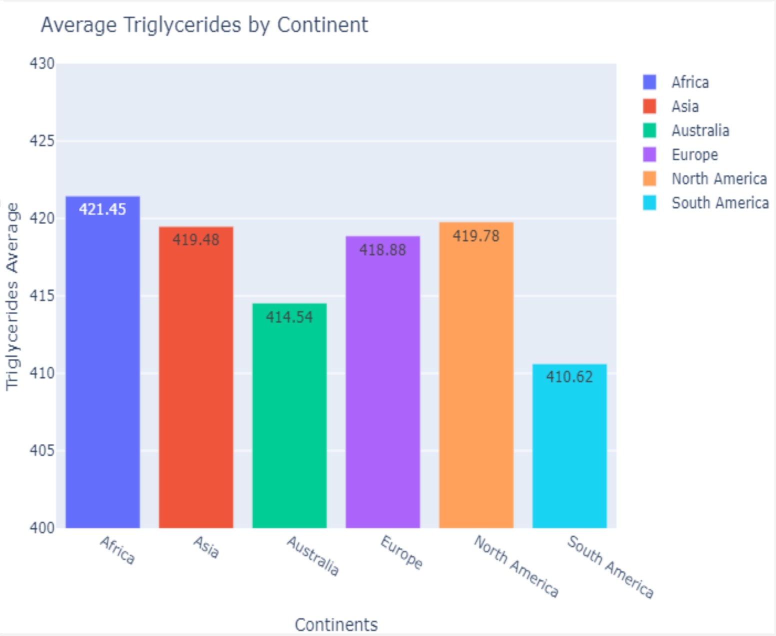
The bar graph vividly conveys that individuals who adhere to a medication regimen exhibit a diminished risk of heart attacks.

Average Cholesterol Levels per Continent



The bar graph elegantly displays the comparative average cholesterol levels across continents, revealing North America as the region with the highest readings, while Australia boasts the lowest cholesterol figures.

Average Triglyceride Levels per Continent

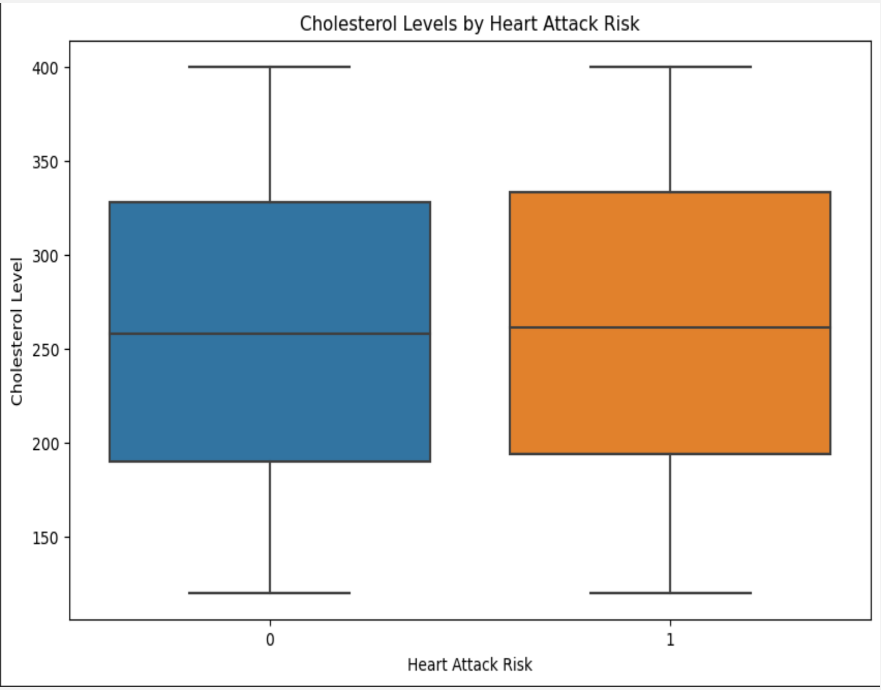


The bar graph elegantly delineates the triglyceride levels by continent, pinpointing South America as the region with the minimal levels, while Africa emerges with the maximum.

Photos provided by Pexels

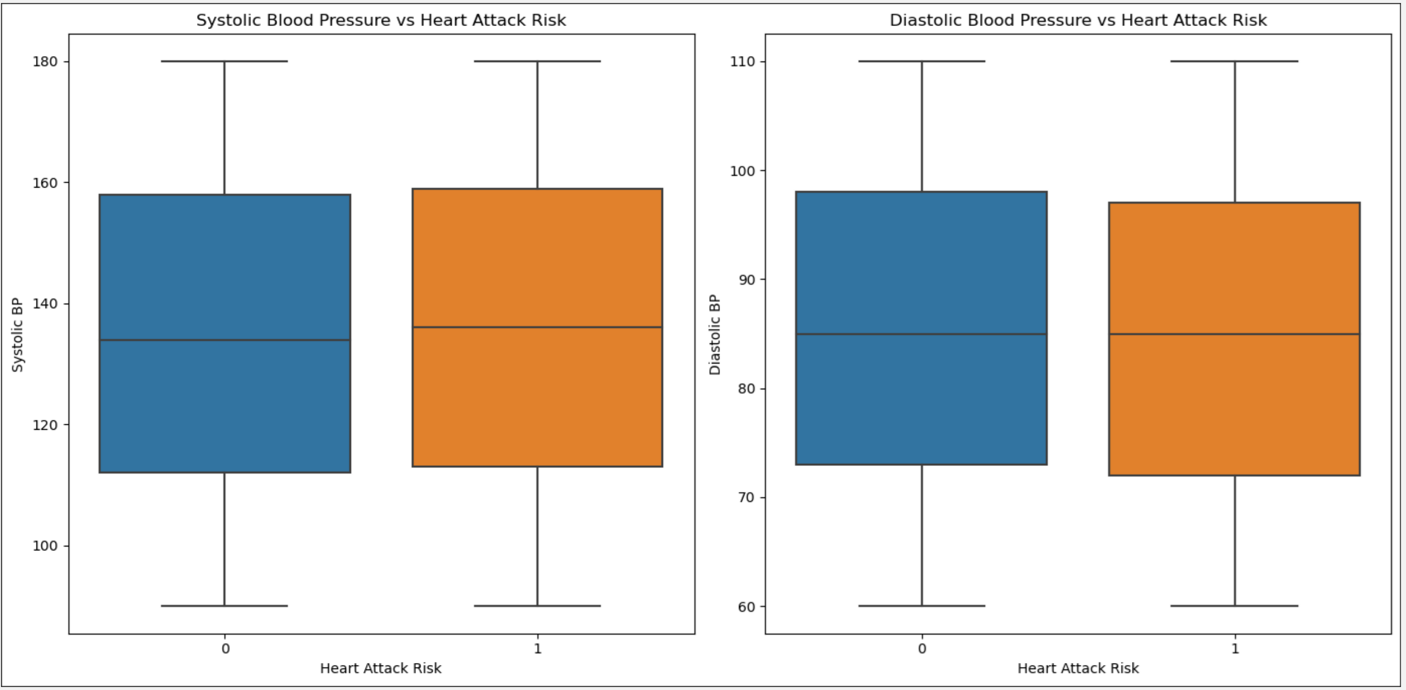
Data Analysis and Visualizations

Cholesterol Levels by Heart Attack Risk



The box plot graph elegantly illustrates that elevated cholesterol levels are associated with an increased risk of cardiac events.

Systolic and Diastolic Blood Pressure vs Heart Attack Risk



These box plots reveal a nuanced cardiovascular insight: individuals with elevated systolic blood pressure face an increased susceptibility to heart attacks, whereas those with higher diastolic blood pressure exhibit a surprisingly reduced risk.

Photos provided by Pexels

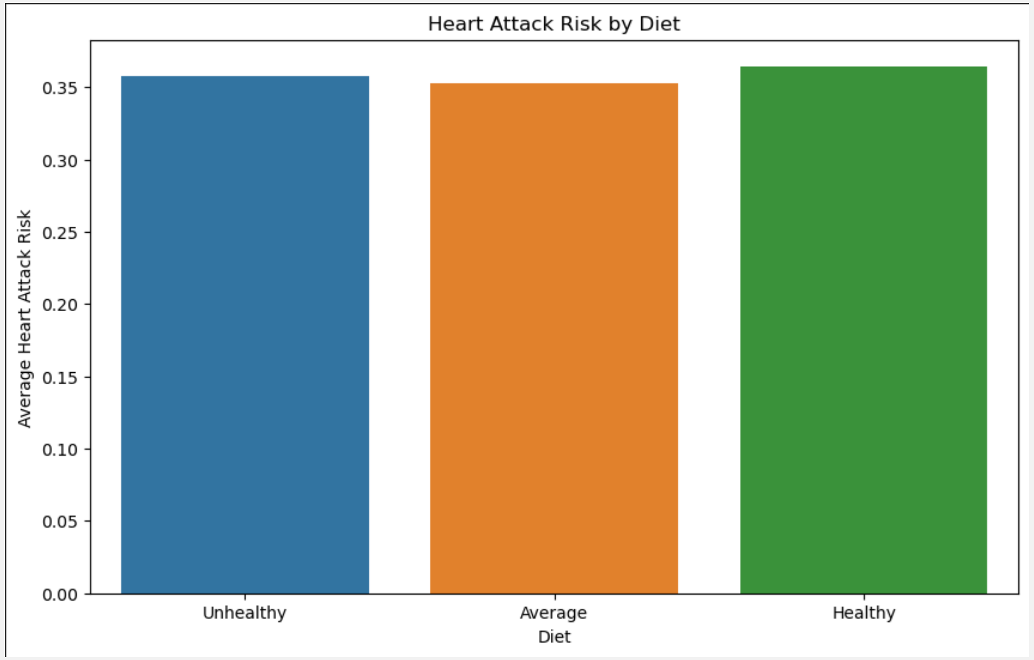
Insights and Findings

Key Insights

While the majority of the data collected aligns with expectations and demonstrates predictable patterns, certain unanticipated results have emerged due to key influence. Given that the data compilation spanned globally and accounted for numerous variable, it is conceivable that the sample size of 8,763 patients may not have been sufficiently large to encapsulate a definitive accuracy in the datas representation.

Trends and Patterns

The analysis reveals that adequate sleep and medication compliance are linked to a lower incidence of heart attacks. Moreover, it's observed that people adhering to exceptionally healthy diets often register a higher risk of heart attacks. This counterintuitive result could be related to a phenomenon known as the "health-conscious worker effect," where individuals who are proactive about their health are more likely to get regular check-ups. Such vigilance may lead to a higher reported incidence of heart issues simply because their conditions are more likely to be diagnosed then those less health-conscious



Photos provided by Pexels

Conclusion

Actionable Steps

To minimize your heart attack risk, ensure ample sleep, adhere to a nutritious diet, maintain a five-day exercise regimen per week, comply with medication schedules, and keep cholesterol levels in check.

Future Research Opportunities

Further examination could be conducted into the median age of individuals adhering to a nutritious diet and an in-depth analysis of familial heart disease history among both smokers and non-smokers, to glean additional insights in the occurrence of the “smoker paradox” phenomenon.

Potential Interventions

To achieve a representative dataset with global scope, and expanded sample size is necessitated.

Photos provided by Pexels