Case study: Healthcare Costs Analyzing and Predicting, USA.

By Auguste Shikongo

Introduction

The escalating healthcare costs in the US necessitate a thorough understanding of medical expenses. Our project, fueled by Kaggle dataset encompassing vital variables: Age, Sex, BMI, Children, Smoker, Region, and Charges, aims to analyze and predict individual medical costs. By unravelling the complexities of healthcare expenditures, we strive to develop a predictive model to estimate individual medical costs based on personal and health-related factors, contributing valuable insights into the factors influencing medical costs.

Motivation

The rising healthcare costs in the USA necessitate a deeper understanding to address this concern. This project aims to provide insights and predictive capabilities to better comprehend healthcare expenses.

Goal

The goal is to analyze and predict individual medial costs (Charges) based on personal and health-related factors, ultimately developing a predictive model for future cost estimation.

Objectives

- Develop a predictive model for estimating individual medical costs.
- Provide insights into factors significantly impacting medical costs.
- Create a user-friendly interface for data exploration and costs predictions.
- Ensure the highest standards of data privacy and security.

Scope

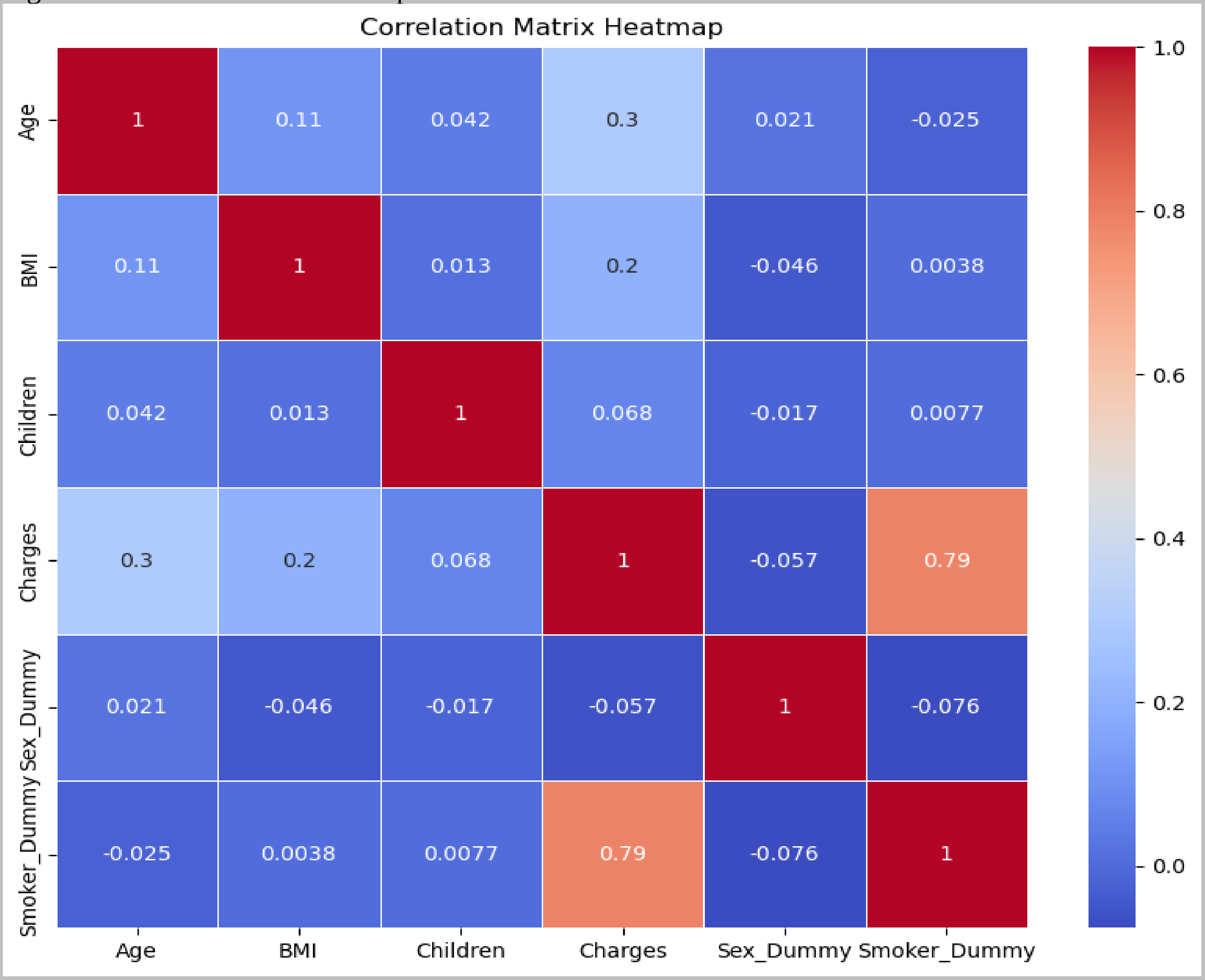
The scope includes data analysis, visualization, and the development of a predictive model. Emphasis is also placed on data privacy and security.

Methods Employed

- **Data Source:** Kaggle dataset
- **Data Cleaning**: Wrangling and cleaning, including handling missing values and extreme values
- **Key Questions:**
    - Factors impacting medical costs.
    - Regional differences in charges.
    - Distribution of medical costs.
    - Gender, BMI, and smoking distributions.
    - Age distribution and more.
- **Visualizations and Techniques**
- **Exploratory Variable Data Analysis:**

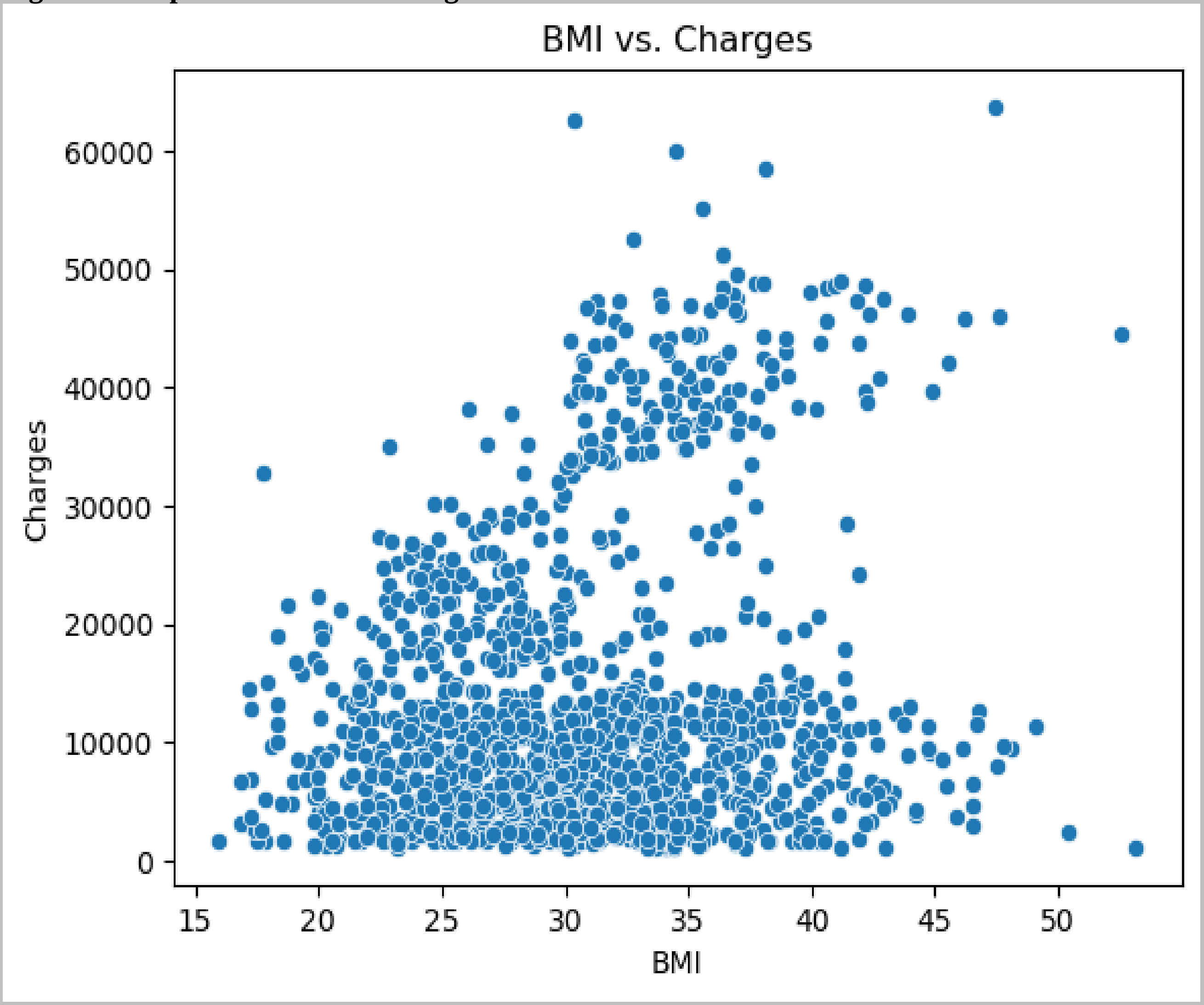
Exploratory Analysis

Fig 1. Correlation Matrix Heatmap



Correlation Matrix Heatmap revealed positive correlations of age, BMI, children, and smoking with healthcare Charges.

Fig 2. Scatterplot for BMI vs. Charges



The fig 2 illustrates a strong relationship between BMI and Charges, indicating an increase in healthcare costs with high BMI.

Fig 3. Scatterplot for Age vs Charges

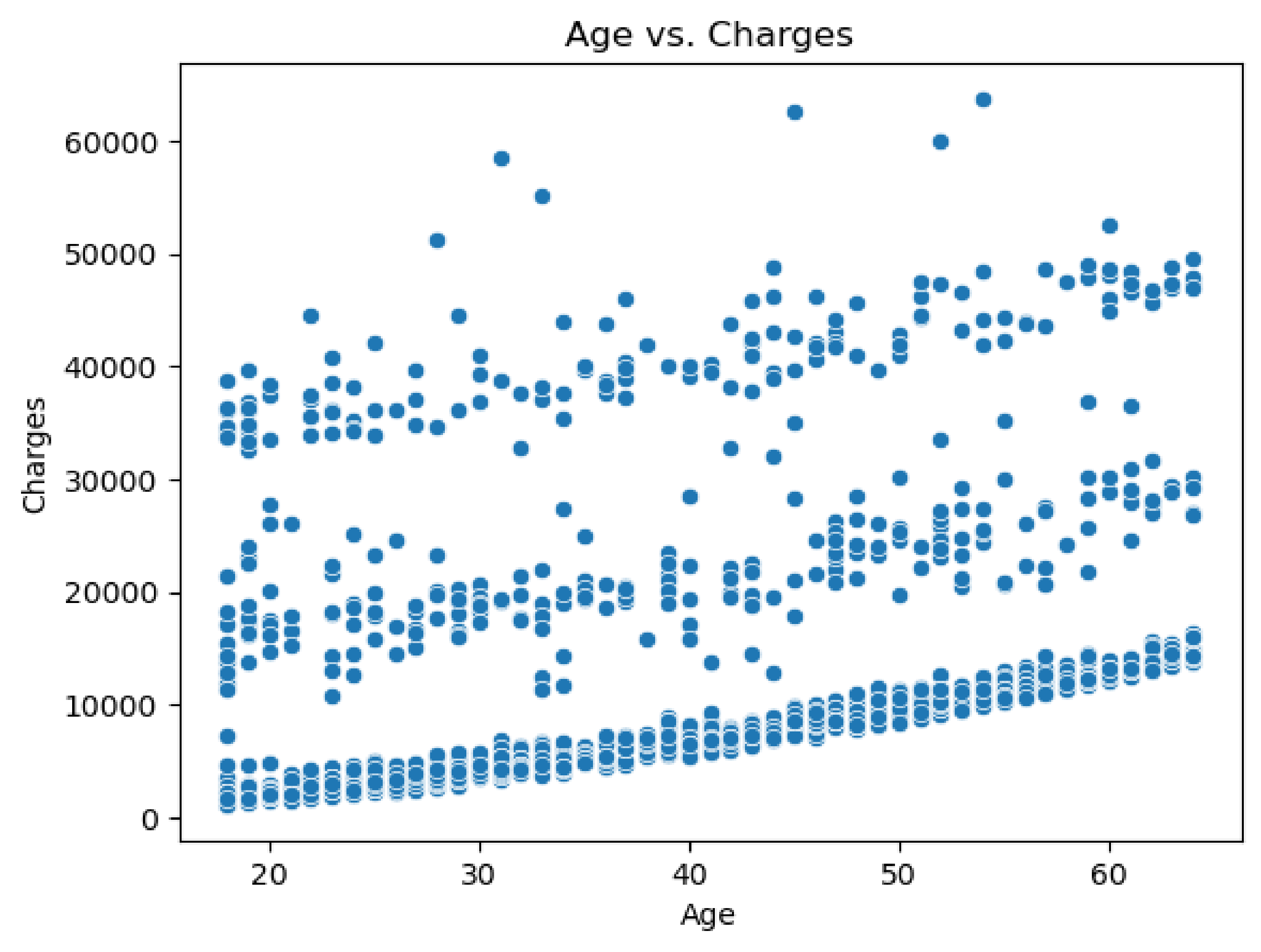


Fig 3 illustrated a strong relationship between Charges and Age, indicating that as individuals become older, their healthcare costs increase.

Fig 4. Pair plot for the entire Dataset

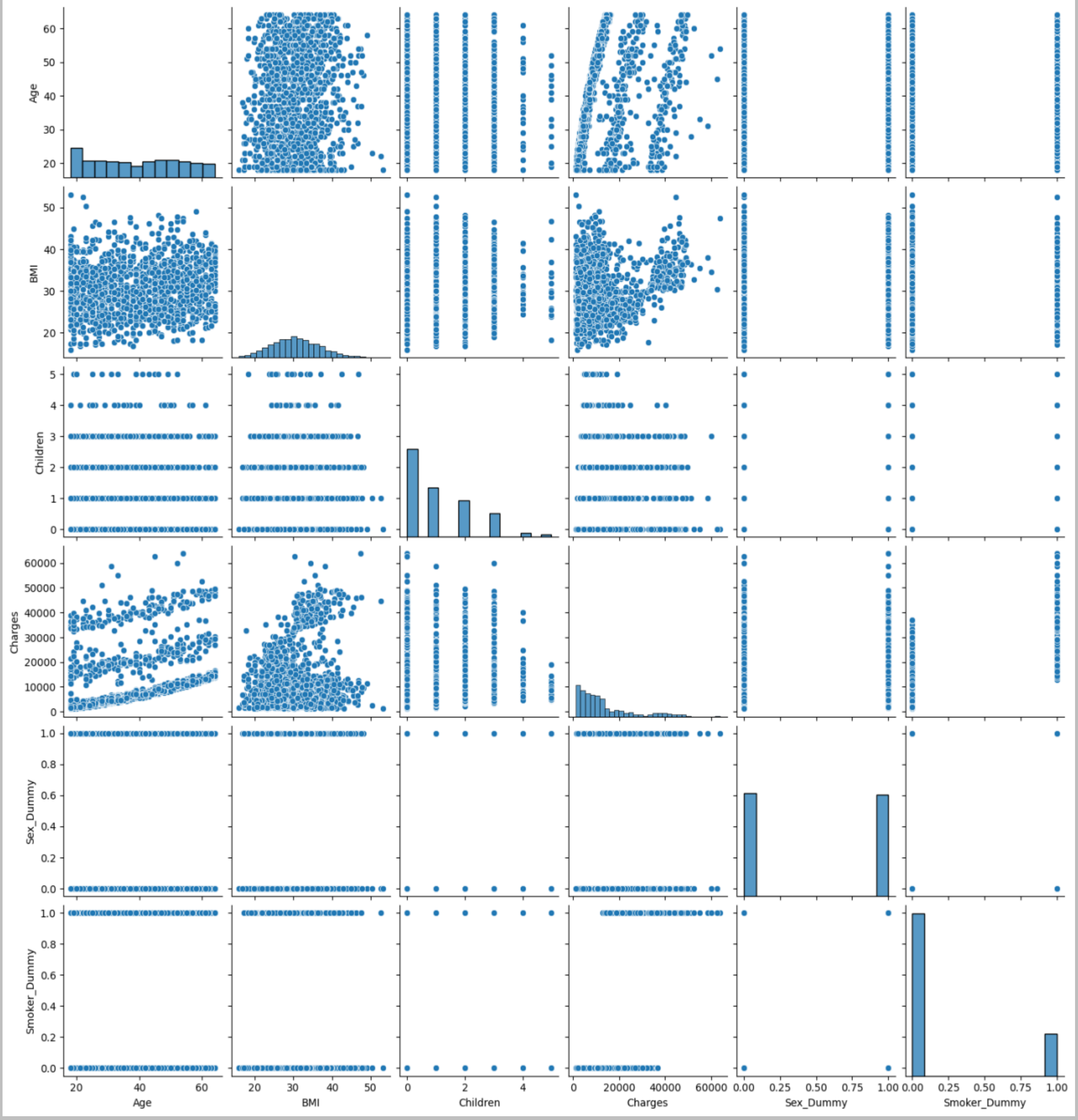


Fig 4 explored relationships between Age, BMI, the number of children, Sex, and smoking status. Highlights a substantial group with high BMI and high Charges, prompting further investigation.

Fig 5. Histogram

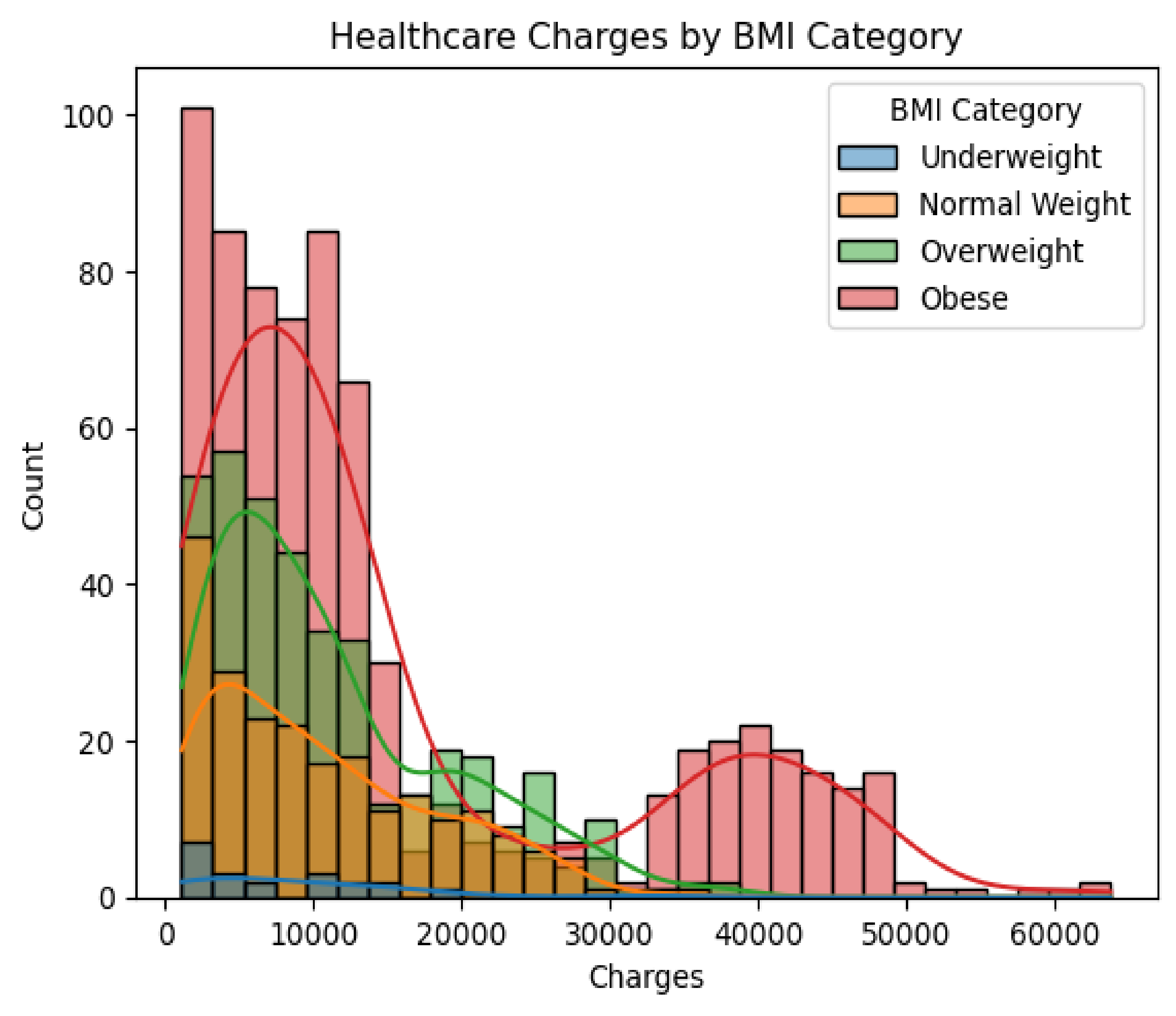


Figure 5 categorizes BMI into Underweight, Normal Weight, Overweight, and obese, revealing varying healthcare charges associated with each category.

Fig 6. USA Map Regions

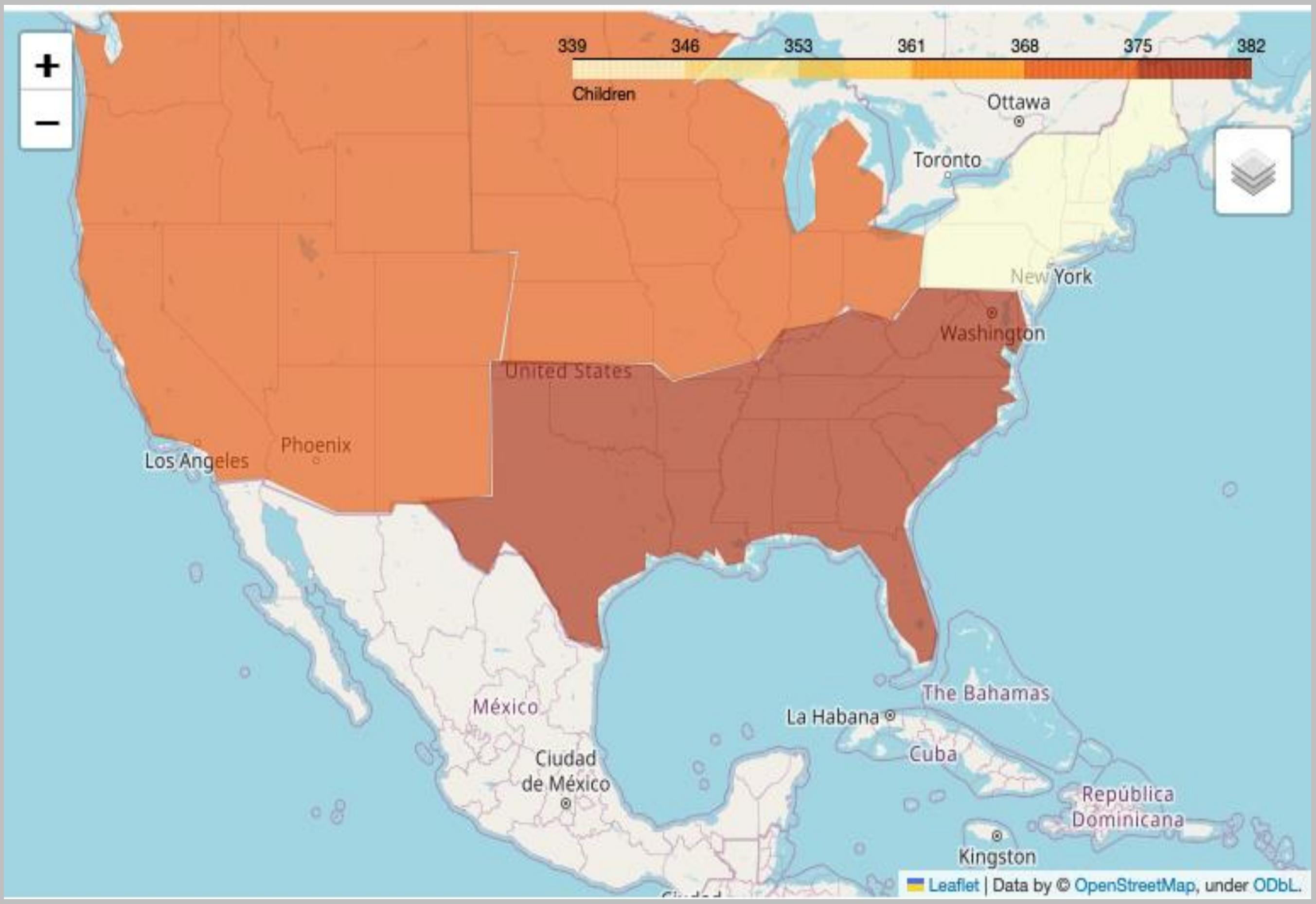
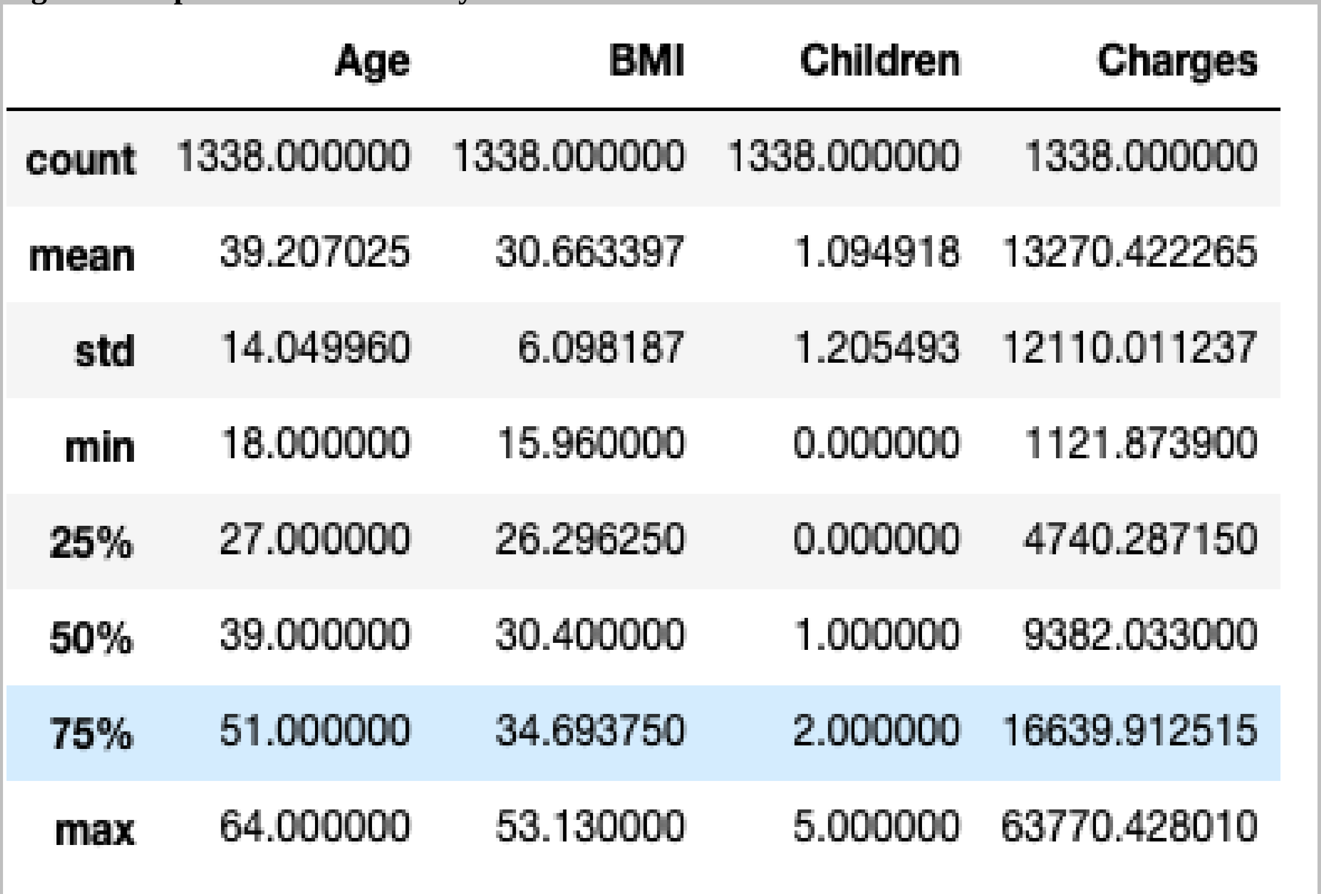


Fig 6 visualizes regional variations in healthcare charges, with the Midwest having the highest average charges. Further analysis is needed to understand underlying factors.

Fig 7. Descriptive Statistical Analysis



	Age	BMI	Children	Charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

Descriptive Statistical Analysis indicates the Age range in the dataset spans from 18 to 64, which is within the expected range.

Research Questions

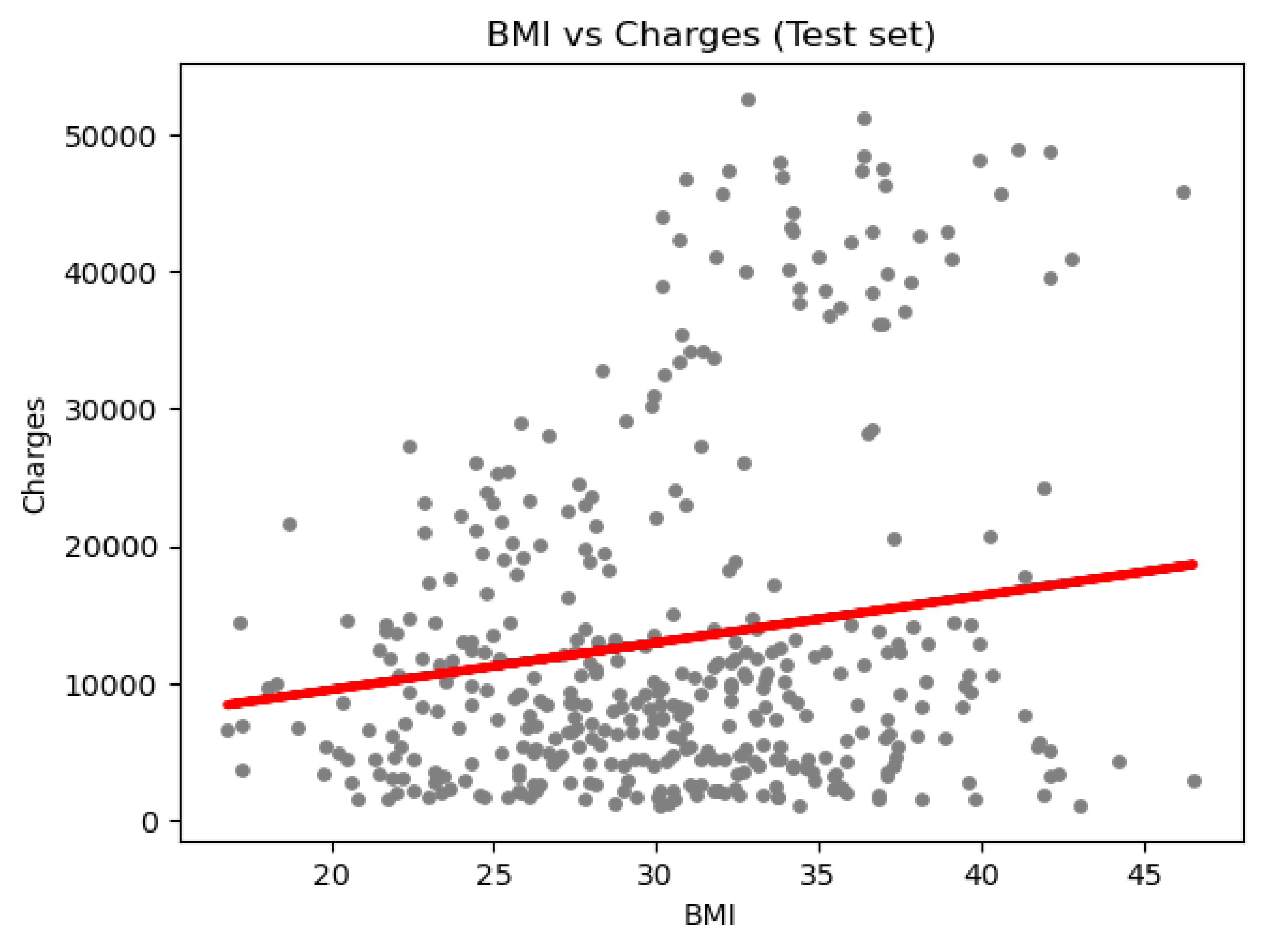
Explored factors impacting medical costs, regional differences, charge distribution, gender distribution, BMI distribution, number of children covered, smoking proportions, and age distribution.

Hypothesis Test

Tested the hypothesis that individuals with a BMI of 30 and above incur higher healthcare charges.

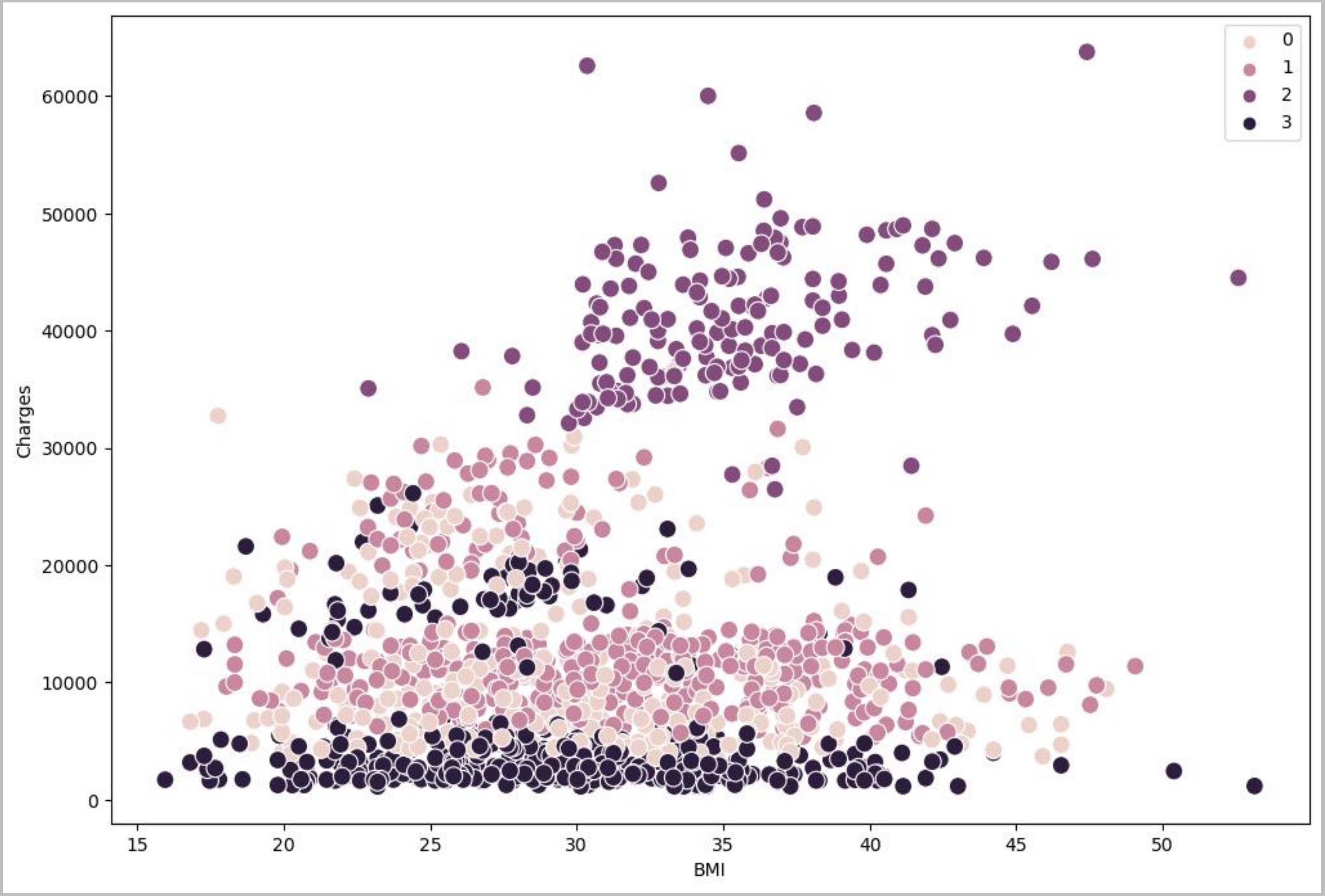
Exploring and Gaining Insights into the Complex Relationship between BMI and Healthcare Charges through Regression and Clustering Models.

Fig 8. Regression model



In this regression model, the red regression line suggests that, on average, higher BMI corresponds to higher charges, as evidence by the upward trend. However, it’s essential to note that the red regression line doesn’t perfectly encapsulate all the data points. Notably, in the BMI range of 40 to 45, there are numerous data points indicating low charges, while a low BMI of 25 has data points indicating high charges, this finding contradicts my initial hypothesis.

Fig 9. Clustering Models



In this cluster examining the color-coded data points, where pink (code 0), red (code 1), and blank (code 3) clusters are closely grouped, charges consistently range from 0 to 30000 across all BMI values. In contrast, the purple (code 2) clusters tightly group around a BMI of 30 and above, exhibiting notable high charges ranging from 30000 to 50000. This suggests that individuals with a BMI of 30 and above generally incur higher charges compared to those with other BMI values. The clustering patterns appear coherent, capturing distinct relationships between BMI and medical charges. The separation into low, high BMI clusters aligns with the expected behavior in a k-means clustering scenario. These insights offer valuable perspectives on the intricate interplay between BMI and healthcare charges, providing a nuanced understanding of the data.

Skills and Tools Used

- Visualizations
- Python-Jupyter Notebook
- Tableau
- Excel
- Github
- Machine learning (Algorithm: Linear Regression)
- Unsupervised Machine learning (Algorithim: K-mean clustering)
- Descriptive Statistics
- Hypothesis Testing.

Challenges Faced

Navigating through diverse data sources, cleaning, and preprocessing data, and selecting appropriate machine learning techniques were initial challenges. Additionally, interpreting the complex relationships between variables required thorough exploration and analysis.

Decision-Making

After initial exploratory analysis, a focus on specific variables, especially the relationship between Charges and BMI, guided further analysis. The use of Tableau provided a dynamic and interactive platform for in -depth exploration.

Conclusion

The project successfully addressed the rising healthcare costs by providing valuable insights and a predictive model. Challenges were overcome through a systematic approach, utilizing a variety of tools influencing medical costs.

Links

- Tableau analyzing BMI and Charges
- Github python code project

Source

Kaggle website
https://www.kaggle.com/datasets/mirichoi0218/insurance