**AMOD-5430: Data Visualization**

PROJECT PROPOSAL ON

Python Driven EDA & Data Visualization in Retail

JUNE 17, 2023

# SABRINA FARZANA

STUDENT ID: 0792500

**ABDUL KAMAL AZAD**

STUDENT ID: 0774556

## Abstract

The healthcare sector generates vast amounts of data that are often underutilized. This project seeks to delve into healthcare data to extract meaningful insights through meticulous Exploratory Data Analysis (EDA) and the employment of advanced visualization techniques with Python. The exploration will revolve around dissecting patient information, medical conditions, and healthcare services, utilizing a synthetic dataset sourced from Kaggle. We will leverage Python libraries such as Pandas for data wrangling, Matplotlib and Seaborn for visualization. Our anticipated contribution is a set of detailed visual analyses that illuminate hidden patterns and trends, equipping healthcare decision-makers with actionable intelligence to refine healthcare strategies.

## Problem

The healthcare industry is a complex and dynamic environment where patient needs and medical trends change rapidly. One of the significant challenges within this sector is the effective use of data for informed decision-making. Our project addresses the problem of untapped potential in healthcare data, focusing on uncovering actionable insights that can lead to improved healthcare strategies.

.

## Intended Contribution

## Our contribution will be a detailed exploratory data analysis of healthcare data, providing

## visual insights into patient demographics and medical conditions.

## The EDA phase will involve a series of analyses and data visualization techniques to gain

## a deeper understanding of healthcare performance.

## We expect to conduct the following exploratory data analyses:

## Gender-Wise Analysis of Patient Distribution: Identify the distribution of patients

## across different genders using bar charts and plots.

## Medical Condition Analysis: Understand the prevalence of various medical condition

## among patients.

## Blood Type Distribution: Analyze the distribution of different blood types among

## patients.

## Insurance Provider Analysis: Investigate the distribution of patients across different insurance providers.

## Medication Analysis: Identify the most commonly prescribed medications.

## Regression Analysis between Age and Medical Conditions: Examine the relationship

## between patient age and the occurrence of different medical conditions through regression analysis.

## ANOVA Test: Conduct an analysis of variance (ANOVA) to determine if there were

## significant differences in medical conditions across different age groups, genders, and

## insurance providers.

## Chi-Square Independence Test: Perform a Chi-square test of independence to assess

## whether the choice of treatment was associated with the type of medical condition

## .

## 4.Methodology

## The methodology detailed below will provide a systematic approach to exploring healthcare

## performance using EDA and statistical testing. By implementing these methods, valuable insights and evidence-based conclusions will be drawn from the dataset.

## Data Set: We will use a synthetic healthcare dataset containing patient demographics, medical conditions, and healthcare services. Preprocessing tasks will be performed to ensure dataquality and consistency.

## Programming Language: Python

## Environments/Tools: Jupyter Notebook for interactive development.

## Libraries: Pandas, Numpy, Matplotlib, Seaborn, Scikit-learn, SciPy, Statsmodels

## Computational Resources: Personal computers with adequate processing power and memory.

## EDA Structure: We will perform univariate analysis to understand single-variable distributions, followed by bivariate and multivariate analysis to uncover relationships and patterns.

## 5.Expected Results

We anticipate revealing insightful trends and correlations within the healthcare data, such as:

1. A higher proportion of female patients compared to male patients.
2. The prevalence of specific medical conditions such as asthma.
3. The distribution of different blood types, with a notable presence of AB-.
4. The dominance of specific insurance providers, such as Cigna.
5. The most commonly prescribed medications, such as penicillin.
6. Significant relationships between patient age and medical conditions.
7. Significant differences in medical conditions across age groups, genders, and insurance providers.
8. Associations between treatment choices and medical conditions.

These insights will provide actionable intelligence for healthcare decision-makers, helping to refine healthcare services and improve patient care.

## Team Responsibilities

|  |  |  |
| --- | --- | --- |
| **#** | **Team Member** | **Responsibilities** |
| 1 | Sabrina Farzana  ID: 0792500 | Background research, defining objectives, writing literature review, data preprocessing & cleaning, coding, exploratory data analysis, data visualization and peer review. |
| 2 | Abdul Kamal azad  ID: 0774556 | Writing methodology, coding, exploratory data analysis, visualization, analyzing graphs, interpretating results, writing  conclusion and peer review. |

1. **Key References**

* Abouelmehdi, K., Beni-Hessane, A., & Khaloufi, H. (2018). Big healthcare data: Preserving security and privacy. Journal of Big Data. https://doi.org/10.1186/s40537-017-0110-7
* Agrawal, R. (2022, August 31). Exploratory data analysis using data visualization techniques! Analytics Vidhya. [https://www.analyticsvidhya.com/blog/2021/06/exploratory-](https://www.analyticsvidhya.com/blog/2021/06/exploratory-data-analysis-using-data-visualization-techniques) [data-analysis-using-data-visualization-techniques](https://www.analyticsvidhya.com/blog/2021/06/exploratory-data-analysis-using-data-visualization-techniques).
* Recker, G. (2023, July 21). Predictive analytics in healthcare: An overview. Blog. Categories: Clinical AI, Early Disease Detection, Healthcare AI, Predictive Analytics, Preventive Healthcare.