Healthcare Data Analysis

Executive Summary:

This report outlines the comprehensive analysis and preprocessing of a medical appointments dataset using Python and libraries such as Pandas, Matplotlib, and Seaborn. The dataset contains a wealth of information about medical appointments, patient demographics, and appointment outcomes. The report covers initial setup, data overview, data transformation, cleaning, exploration, and findings. The insights gained from this analysis will serve as a foundation for further analysis or modelling tasks.

Introduction:

The purpose of this project was to analyze and preprocess a medical appointments dataset to uncover patterns, trends, and factors influencing appointment no-shows. The dataset contains diverse attributes such as patient details, appointment schedules, and outcomes, making it valuable for understanding patient behaviour and improving appointment attendance rates.

Methodology:

Initial Setup:

The project began by importing essential Python libraries, including Pandas, NumPy, Matplotlib, and Seaborn. The dataset, named 'Data.csv,' was loaded using Pandas for further analysis.

Data Overview:

The dataset comprises 110,527 records and 14 columns. These columns encompass information about patient ID, appointment ID, gender, scheduled day, appointment day, age, neighborhood, scholarship, hypertension, diabetes, alcoholism, handicap, SMS received, and no-show status.

Data Information:

The info() method was employed to obtain insights into data types, memory usage, non-null counts, and data types for each column.

Data Transformation:

The 'ScheduledDay' and 'AppointmentDay' columns were converted into datetime objects using the pd.to_datetime function.
New columns 'sch_weekday' and 'app_weekday' were introduced to store the weekdays for scheduled and appointment days.
Column names were refined to address spelling errors using the rename method.

Data Cleaning:

Columns that had minimal impact on analysis were eliminated using the drop method. The 'No-show' column was transformed into binary values (0 for 'No' and 1 for 'Yes') to facilitate further analysis.

Data Exploration:
The distribution of each predictor variable in relation to the 'NoShow' column was visualized through bar plots. The percentage of 'No-show' appointments was calculated to gauge the overall no-show rate.

Missing Data Analysis:
A point plot was generated to visualize the percentage of missing values in the dataset. Fortunately, no missing values were identified.

Age Grouping:
The 'Age' column was grouped into age ranges using the pd.cut function, creating a new 'Age_group' column. The original 'Age' column was dropped.

Bivariate Analysis:
Univariate plots were constructed for various variables concerning the 'NoShow' column, considering factors such as gender, age group, hypertension, etc.

Findings:

1. Female patients had a higher number of appointments compared to male patients.
2. Age groups 0 and 1 displayed an 80% show rate, while other age groups exhibited a similar show rate.
3. Each neighbourhood had an approximate 80% show rate.
4. Patients without scholarships had an 80% show rate, whereas approximately 75% of patients with scholarships attended.
5. Patients without hypertension had a 78% show rate, compared to 85% among patients with hypertension.
6. Patients without diabetes had an 80% show rate, whereas around 83% of patients with diabetes attended.
7. Patients who didn't receive SMS had an 84% show rate, while those who received SMS had a 72% show rate.
8. Appointments were not scheduled on Sundays, and appointments on Saturdays were relatively infrequent compared to other weekdays.

Conclusion:

The analysis and preprocessing of the medical appointments dataset provided valuable insights into various factors influencing appointment no-shows. Through rigorous data transformation, cleaning, exploration, and visualization, the dataset has been refined for further analysis or modelling endeavours. These insights contribute to a deeper understanding of patient behaviour and can guide strategies to enhance appointment attendance rates.