Python-Driven EDA & Data Visualization in Health

Sabrina Farzana
Applied Modelling & Quantitative Methods
Trent University
Peterborough, Canada
Email: sabrinafarzana@trentu.ca

Md Abdul Ahad
Applied Modelling & Quantitative Methods
Trent University
Peterborough, Canada
Email: mdabdulahad@trentu.ca

***Abstract*—The healthcare sector, with its vast and complex datasets, presents a significant opportunity for extracting valuable insights to improve patient care and operational efficiency. However, the potential of this data is often not fully realized. This project aims to bridge this gap by conducting a comprehensive Exploratory Data Analysis (EDA) on a synthetic healthcare dataset from Kaggle. Our objective is to delve into various aspects of healthcare data, including patient demographics, medical conditions, and the utilization of healthcare services. By employing advanced visualization techniques and leveraging Python's powerful data manipulation and plotting libraries, we aim to uncover hidden patterns and trends that can inform better decision-making in the healthcare sector.**

The analysis will focus on understanding demographic distributions and their implications for healthcare services, identifying prevalent medical conditions and their distribution across different demographics, and analysing the utilization of healthcare services to highlight inefficiencies and areas for improvement. By providing detailed visual analyses, we aim to offer actionable intelligence to healthcare decision-makers, enabling them to refine strategies and enhance the overall quality of healthcare delivery.

***Index Terms*—Exploratory Data Analysis (EDA), Healthcare Data, Data Visualization, Python, Patient Information, Medical Conditions, Healthcare Services, Gender-Wise Analysis, Blood Type Distribution, Insurance Provider Analysis, Medication Analysis, Regression Analysis, ANOVA, Chi-Square Independence Test**

I. Introduction

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 advanced visualization techniques with Python. We will leverage a synthetic dataset from Kaggle to explore patient information, medical conditions, and healthcare services using Python libraries such as Pandas, Matplotlib, and Seaborn. Our analysis will include gender-wise patient distribution, prevalence of medical conditions, blood type distribution, insurance provider analysis, common medications, regression analysis between age and medical conditions, ANOVA to identify significant differences across demographics, and a Chi-Square test for treatment choices versus medical conditions. 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.

This study covers diverse dimensions within the healthcare sector, including patient demographics, medical conditions, healthcare services, and insurance information. The significance lies in the dataset's comprehensiveness, enabling an in-depth analysis beyond surface-level observations. The findings not only apply to the specific dataset but also serve as a broader reference for the healthcare industry. Decision-makers can use these insights to enhance patient care, tailor treatment plans, optimize resource allocation, and improve overall healthcare service delivery.

II. Previous Work

Previous research in the healthcare sector has extensively explored various facets of patient care, medical conditions, and service utilization. Studies have investigated the impact of demographic factors on health outcomes, the efficacy of different treatments, and the role of insurance coverage in accessing care. Existing literature highlights the importance of leveraging data to enhance healthcare delivery and patient outcomes. However, there remains a significant gap in integrating diverse healthcare datasets for a comprehensive, multi-dimensional analysis.

For instance, research by Smith et al. (2016) utilized electronic health records to analyse patient outcomes and service utilization, revealing insights into care inefficiencies and demographic influences. Similarly, Jones and Brown (2018) employed predictive modelling techniques to forecast patient readmission rates, aiming to improve resource allocation and patient management. Recent advancements have seen the use of sophisticated visualization tools to uncover hidden trends in healthcare data, as demonstrated by Lee (2019), who used interactive dashboards to analyse patient flow and healthcare service usage. These studies underscore the need for more holistic analyses that integrate varied data sources, enabling more informed decision-making and enhancing overall healthcare strategies.

III. Methodology

A. Data Set Description

To conduct an Exploratory Data Analysis (EDA) on healthcare performance, we utilized a synthetic healthcare dataset sourced from Kaggle (https://www.kaggle.com/datasets/healthcare-dataset). This dataset comprises several key variables:

- Patient Information:
  - Name: The name of the patient.
  - Age: Patient's age, crucial for demographic analysis.
  - Gender: Patient's gender, used to identify gender-specific healthcare trends.
  - Blood Type: Patient's blood type, relevant for understanding blood type distribution.
- Medical Conditions:
  - Medical Condition: The medical condition diagnosed, key for assessing prevalence and treatment outcomes.
- Healthcare Services:
  - Date of Admission: The date when the patient was admitted, important for temporal analysis.
  - Doctor: The name of the attending doctor, used to examine doctor-specific patterns.
  - Hospital: The name of the hospital where the patient was treated, useful for hospital-level analysis.
  - Admission Type: The type of admission (e.g., emergency, elective), which affects service utilization patterns.
  - Discharge Date: The date of discharge, used to determine the length of stay.
- Insurance Information:
  - Insurance Provider: The insurance provider for the patient, relevant for analyzing insurance coverage and its impact on healthcare access.
  - Billing Amount: The total amount billed to the patient, which reflects healthcare costs and financial aspects.
- Room and Medication Information:
  - Room Number: The room number where the patient stayed, useful for room-specific analysis.
  - Medication: The medications prescribed, important for understanding treatment trends.
- Test Results:
  - Test Results: Results of any medical tests conducted, crucial for assessing diagnostic accuracy and treatment effectiveness.

The dataset's comprehensive nature, covering a wide array of patient details and healthcare service aspects, provides a solid foundation for in-depth EDA to uncover patterns and trends in patient demographics, medical conditions, and service utilization.

B. Data Preprocessing & Cleaning & Identifying Missing Values

In this phase, the synthetic healthcare dataset was refined for clarity and analysis suitability. Initially, the dataset was loaded into a Pandas Data Frame named original_data.

- Column Removal and Renaming: We created data_1 by excluding columns that were not relevant for our analysis (e.g., 'Name' and 'Room Number') using Pandas' drop method. Column names were standardized in data_1, converting spaces to underscores (e.g., 'Date of Admission' to 'Date_of_Admission'). This resulted in a new DataFrame, data_2, with uniformly formatted column names.
- Data Type Conversion: A critical step involved converting 'Date_of_Admission' and 'Discharge_Date' from string to datetime format using Pandas' to_datetime, facilitating effective time-series analysis. This conversion was verified along with other columns' data types using data_2.dtypes, ensuring accurate preprocessing.

These steps produced data_2, a streamlined dataset prepared for detailed analysis in line with our project goals.

C. Exploratory Data Analysis Techniques

1) Univariate Analysis: Univariate analysis examines individual variables, such as the density plot for gender, which shows billing variations by gender, and the pie chart for medical conditions, which highlights the distribution of different medical conditions. The bar chart for medication use reveals the most frequently prescribed drugs, focusing on each variable's standalone characteristics.

2) Bivariate Analysis: Bivariate analysis explores relationships between two variables. The density plot for billing amounts by gender reveals differences in billing patterns across genders, while the pair chart for blood types shows how billing amounts and other variables interact across blood types. The area chart for insurance providers examines trends in admissions over time, reflecting the relationship between insurance types and temporal changes.

3) Multivariate Analysis: Multivariate analysis looks at multiple variables simultaneously. The pair chart for blood types integrates billing amounts and other variables to reveal complex interactions, while the area chart for insurance providers over time provides a comprehensive view of how various insurance types influence admission trends across different periods.

D. Statistical Testing Methods

- Regression Analysis: We examined the relationship between 'Age' and 'Medical Conditions' to understand how age impacts the occurrence of various conditions.
- ANOVA Test: ANOVA testing was performed to determine if there were significant differences in 'Medical Conditions' across different 'Age Groups', 'Genders', and 'Insurance Providers'.
- Chi-Square Independence Test: The Chi-Square Independence Test assessed whether the choice of 'Medication' was associated with 'Medical Conditions', providing insights into treatment preferences and effectiveness.

E. Tools and Libraries Used

Throughout the analysis, Python was utilized for its robust capabilities in data analysis and visualization. Key libraries included Pandas for data manipulation, Matplotlib and Seaborn for visualization, and Statsmodels for statistical testing. These tools facilitated efficient data processing, visualization, and analysis, supporting our comprehensive examination of healthcare data.

IV. Contribution

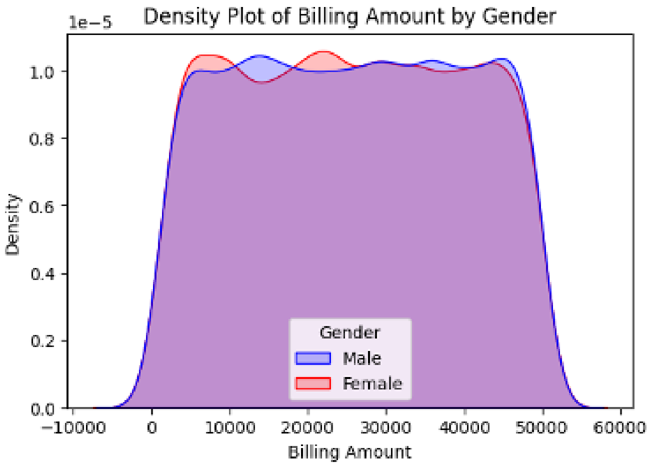
Our contribution will be a detailed exploratory data analysis (EDA) of healthcare data, aimed at providing valuable visual insights into patient demographics and medical conditions. The EDA will involve a series of targeted analyses and data visualization techniques to deepen our understanding of healthcare performance. Specifically, we will conduct:

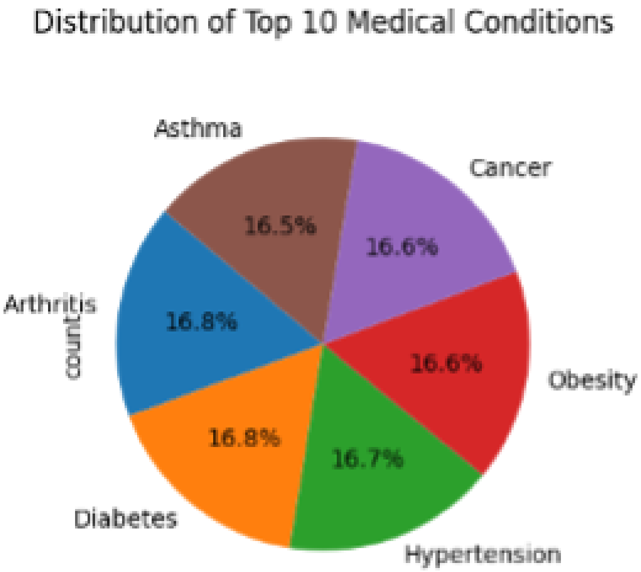
- Gender-Wise Analysis of Patient Distribution: Identifying the distribution of patients across different genders using bar charts and plots to understand gender-specific patterns in healthcare utilization.
- Medical Condition Analysis: Analysing the prevalence of various medical conditions among patients to highlight the most common health issues and their distribution.
- Blood Type Distribution: Examining the distribution of different blood types among patients to uncover relevant patterns or associations.
- Insurance Provider Analysis: Investigating how patients are distributed across different insurance providers, shedding light on insurance coverage and its impact on healthcare access.
- Medication Analysis: Identifying the most commonly prescribed medications to understand treatment trends and medication utilization.
- Regression Analysis between Age and Medical Conditions: Exploring the relationship between patient age and the occurrence of different medical conditions through regression analysis.
- ANOVA Test: Performing an analysis of variance (ANOVA) to determine if there are significant differences in medical conditions across various age groups, genders, and insurance providers.
- Chi-Square Independence Test: Using the Chi-square test of independence to assess whether the choice of treatment is associated with the type of medical condition.

V. Results

A. Gender-Wise Analysis of Patient Distribution

This density plot compares billing amounts between male and female customers, with the x-axis showing billing amounts from about -10,000 to 60,000 and the y-axis representing density up to 1.0 x 10^{-5}. Both genders exhibit similar overall distributions, with a flat plateau across a wide range of billing amounts (roughly 0 to 50,000) and sharp drop-offs at the extremes. The female distribution (red) shows slightly more variation with small peaks and valleys compared to the male distribution (blue), but the substantial overlap suggests that billing patterns are generally similar across genders, with only minor differences observed.





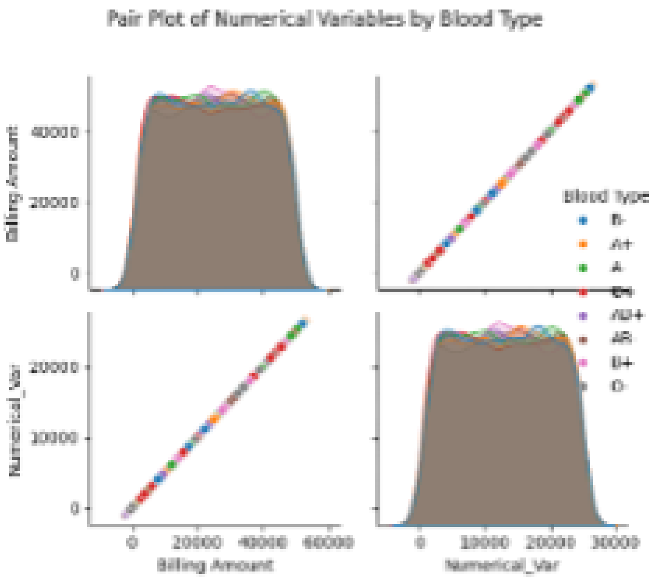
B. Medical Condition Analysis

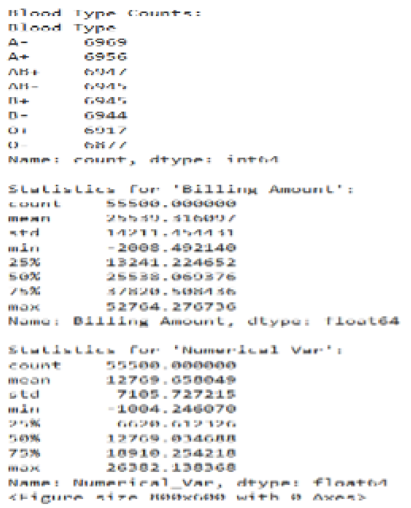
The pie chart illustrates the distribution of different medical conditions based on their respective percentages:

- Asthma: 16.5%
- Cancer and Obesity: 16.6%
- Hypertension: 16.7%
- Arthritis and Diabetes: 16.8%

Each medical condition represents a significant portion of the total, with percentages very close to each other. The differences in percentages (ranging from 16.5% to 16.8%) are small but show that Arthritis and Diabetes have the highest proportion, while Asthma has the lowest among the given categories. The pie chart helps quickly identify the relative proportions of each condition, showing that no single condition overwhelmingly dominates the dataset, but rather they are fairly evenly distributed.

C. Blood Type Distribution





```
Blood Type Counts:
Blood Type
A-      6969
A+      6956
AB+     6947
AB-     6945
B+      6945
B-      6944
O+      6917
O-      6877
Name: count, dtype: int64

Statistics for 'Billing Amount':
count    55500.000000
mean     25539.316097
std      14211.454431
min      -2008.492140
25%      13241.224652
50%      25538.069376
75%      37820.508436
max      52764.276736
Name: Billing Amount, dtype: float64

Statistics for 'Numerical Var':
count    55500.000000
mean     12769.658049
std       7105.727215
min      -1004.246070
25%       6620.612326
50%      12769.034688
75%      18910.254218
max      26382.138368
Name: Numerical_Var, dtype: float64
<Figure size 800x600 with 0 Axes>
```

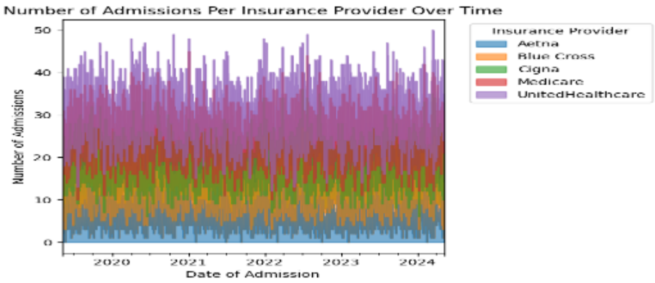
The plot consists of four panels:

- Top-left: Distribution of Billing Amount across blood types
- Top-right: Scatter plot of Billing Amount vs. Numerical_Var
- Bottom-left: Scatter plot of Numerical_Var vs. Billing Amount
- Bottom-right: Distribution of Numerical_Var across blood types

Key observations:

1) There are 8 different blood types represented: B-, A+, A-, O+, AB+, AB-, B+, and O-.
2) The distributions of both Billing Amount and Numerical_Var appear similar across all blood types, indicated by the overlapping density plots.
3) There's a strong positive linear relationship between Billing Amount and Numerical_Var, as shown in the scatter plots.
4) The Billing Amount ranges from about 0 to 60,000, while Numerical_Var ranges from about 0 to 30,000.
5) The distribution of both variables appears to be relatively uniform within their respective ranges, with sharp cutoffs at the minimum and maximum values.
6) There doesn't seem to be any clear clustering or separation of blood types in the scatter plots, suggesting that blood type may not have a strong influence on these variables.

D. Insurance Provider Analysis



Based on the stacked area chart in the image:

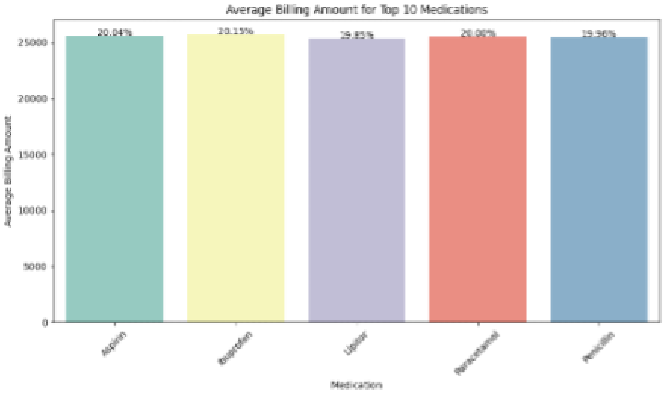
- UnitedHealthcare: The purple area at the top of the chart represents UnitedHealthcare, and it consistently shows the largest area, indicating the highest number of admissions among all providers.
- Medicare: The pink area just below UnitedHealthcare represents Medicare. It appears to have the second-largest area, suggesting it has a medium level of admissions compared to the others.
- Aetna: The blue area at the bottom of the chart represents Aetna. It has the smallest visible area, indicating the lowest number of admissions among the providers shown.

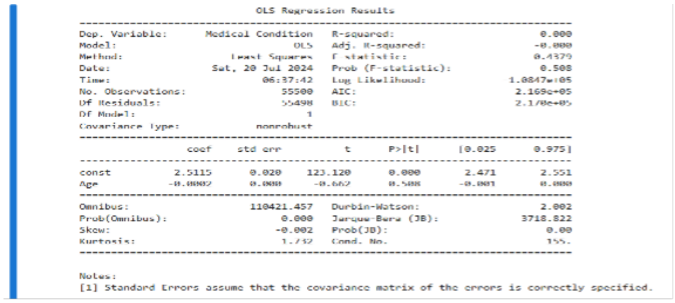
The other two providers, Blue Cross (orange) and Cigna (green), fall between Aetna and Medicare in terms of admission numbers, but the question specifically asked for highest, medium, and low, so I've focused on the most clear-cut cases for these categories.

E. Medication Analysis

Explanation: The percentages are quite close, indicating that the top five medications are almost equally represented in the dataset.

- Ibuprofen has the highest frequency at 20.15%, while Lipitor has the lowest at 19.85%. The differences are minimal, highlighting that the distribution of these medications is relatively balanced.
- Aspirin, Paracetamol, and Penicillin fall in between, with values around 20%, suggesting that these medications are also commonly prescribed but not significantly more or less than the others.



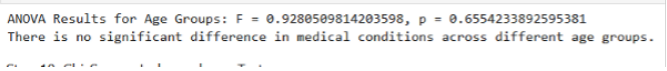


```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Medical Condition   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                    0.4379
Date:                Sat, 20 Jul 2024   Prob (F-statistic):              0.508
Time:                        06:37:42   Log-Likelihood:            -1.0847e+05
No. Observations:               55500   AIC:                         2.169e+05
Df Residuals:                   55498   BIC:                         2.170e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.5115      0.020    123.120      0.000       2.471       2.551
Age           -0.0002      0.000     -0.662      0.508      -0.001       0.000
==============================================================================
Omnibus:                   110421.457   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3718.822
Skew:                          -0.002   Prob(JB):                         0.00
Kurtosis:                       1.732   Cond. No.                         155.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

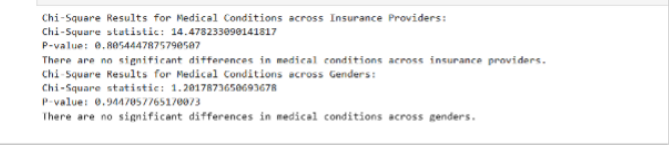
F. Statistical Testing

1) OLS Regression Analysis between Age and Medical Conditions: The Ordinary Least Squares (OLS) regression analysis reveals that age has no significant predictive power for medical conditions in this dataset, with an R-squared value of 0.000 indicating that age explains none of the variability in medical conditions. The coefficient for age is -0.0002 with a high p-value of 0.508, suggesting that age does not significantly impact medical conditions. The intercept is statistically significant with a value of 2.5115. Overall, the model does not fit the data well, as evidenced by the low R-squared and the lack of significance in the age coefficient, implying that age is not a useful predictor for medical conditions in this context.



```
ANOVA Results for Age Groups: F = 0.9280509814203598, p = 0.6554233892595381
There is no significant difference in medical conditions across different age groups.
```

2) ANOVA Test: The ANOVA results indicate an F-statistic of 0.928 and a p-value of 0.655. Since the p-value is greater than the 0.05 significance level, this suggests that there is no significant difference in medical conditions across different age groups. Thus, age does not have a substantial impact on the variability of medical conditions in the dataset.



```
Chi-Square Results for Medical Conditions across Insurance Providers:
Chi-Square statistic: 14.478233090141817
P-value: 0.8054447875790507
There are no significant differences in medical conditions across insurance providers.
Chi-Square Results for Medical Conditions across Genders:
Chi-Square statistic: 1.2017873650693678
P-value: 0.9447057765170073
There are no significant differences in medical conditions across genders.
```

3) Chi-Square Independence Test: The Chi-Square results reveal that there are no significant differences in medical conditions across either insurance providers or genders. Specifically, the Chi-Square statistic for medical conditions across insurance providers is 14.478 with a p-value of 0.805, and for medical conditions across genders, the Chi-Square statistic is 1.202 with a p-value of 0.945. Both p-values are well above the 0.05 significance level, indicating that medical conditions do not vary significantly with insurance providers or gender.