**Exploring Synthetic Healthcare Data: Report**

***by*  
  
Group 5**Dinesh Kumar Reddy

Hayden Sanders

Hemanth Kumar Katrakuti

Sayan Banerjee

Thameem Ashiq Shaik

Vishal Reddy Chewalla  
  
**Professor Andrew Duncklee  
  
Data Mining with Splunk**

**Application Name: HealthCare\_APP\_Group\_Assignment\_Group 5**

**The Dataset**

Data-set: Synthetic healthcare dataset  ([🩺Healthcare Dataset 🧪 (kaggle.com)](https://www.kaggle.com/datasets/prasad22/healthcare-dataset) )   
Year range: 2018-2013

Row count: 10001

Column count: 15  
  
**Objective:** The primary aim of this synthetic healthcare dataset is to provide an extensive and morally sound educational resource for those who are enthusiastic about data science, machine learning, and data analysis, particularly in relation to the healthcare sector. Inspired by the acknowledged need for diverse and useful healthcare data for research and teaching, this dataset attempts to offer a secure and private substitute for real-world healthcare data, which is sometimes limited by privacy and sensitive laws. The dataset has been carefully created to imitate the normal structure and properties found in real healthcare records by utilising Python's Faker package.  **Introduction:**

The importance of Splunk analysis in the rapidly changing field of healthcare analytics cannot be emphasised. Professionals may get deep insights into the complexities of healthcare data by utilising Splunk's capability, which enables them to identify significant patterns, trends, and correlations. Real-time analysis and decision-making are necessary in the dynamic field of healthcare, and Splunk is an essential tool for doing this. Healthcare professionals may improve patient outcomes, increase operational efficiency, and find crucial insights that lead to game-changing developments in the industry by using Splunk analysis. The synthetic healthcare dataset that is provided here gains even more value because it offers the perfect setting for using Splunk's capabilities in healthcare data exploration, in addition to facilitating conventional analytica lmethodologies.

**Educational Importance, and the Role of Splunk Synthetic Dataset:**

This project addresses the critical need for diverse and practical healthcare data in educational and research contexts. The synthetic dataset not only serves as a platform for practicing data manipulation and predictive modeling but also ensures a secure environment by eliminating concerns related to sensitive healthcare information. It fosters innovation and knowledge sharing, enabling the development and testing of predictive models and encouraging insightful data visualizations. In the evolving landscape of healthcare analytics, the importance of Splunk analysis cannot be overstated. Splunk serves as a pivotal tool for deriving comprehensive insights into healthcare data, enhancing operational efficiency, and uncovering critical patterns and correlations in real-time. The synthetic healthcare dataset becomes an invaluable asset by providing an ideal environment for harnessing the capabilities of Splunk in healthcare data exploration.

**Data-set Information:**  
  
 This comprehensive healthcare dataset encompasses a myriad of columns, each providing specific details crucial for holistic patient analysis and healthcare modeling. Ranging from fundamental patient information such as name, age, and gender to critical medical indicators like blood type and primary medical condition, the dataset offers a rich repository for in-depth data analysis. With additional insights into the date of admission, attending doctor, healthcare facility details, insurance provider, billing amounts, room assignments, admission types, discharge dates, prescribed medications, and medical test results, this dataset becomes a versatile resource for various healthcare-related tasks. The inclusion of diverse parameters, expressed in both categorical and numerical formats, empowers practitioners to explore a multitude of scenarios, making it an ideal dataset for modeling, predictive analytics, and gaining nuanced insights into healthcare trends and patient outcomes.  
  
  
**Data-set Preview:**  
  


**Dashboard Overview: Optional+++**

1. **Average Billing Amount by Insurance Provider and Medical Condition:**

***Requires User Input:*** This scenario allows users to input specific parameters and observe the average billing amount based on insurance providers and medical conditions.

1. **Patient Count by Medical Condition and Age Group:**

***Requires User Input:*** Users can input criteria to explore the distribution of patient counts across different medical conditions and age groups.

1. **Patient Count by Age:**

This scenario provides a straightforward analysis of patient counts based on age groups, offering insights into the age distribution within the dataset.

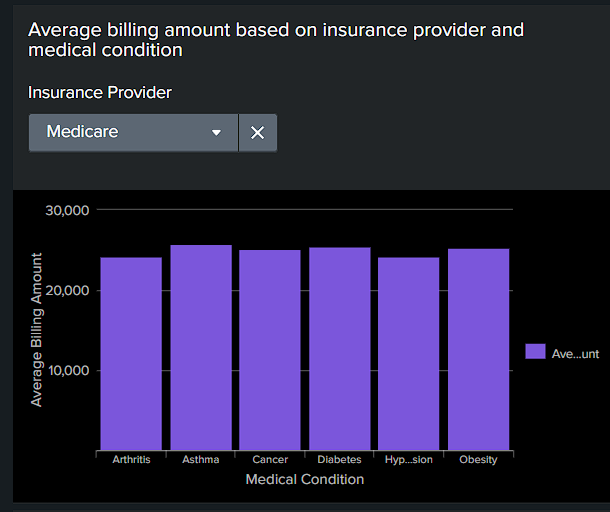
1. **Medical Condition with the Maximum Count for Each Admission Type:**

Here, we identify the medical condition with the highest count for each admission type, shedding light on prevalent conditions for different admission scenarios.

1. **Distribution of Different Test Results ('Inconclusive,' 'Abnormal,' 'Normal'):**

This scenario visualizes the distribution of test results, providing a comprehensive view of the prevalence of 'Inconclusive,' 'Abnormal,' and 'Normal' outcomes in the dataset.  
  
**6. Gender Distribution Pie Chart:**The gender distribution of healthcare datasets is visualised using a Splunk-generated Gender Distribution Pie Chart in the Report Strategy, which helps with resource allocation, strategic planning, and patient care.  
  
7.  
  
8.  
  
  
  
  
  
  
  
  
  
Link: <https://splunk.sps.cs.clarku.edu:8000/en-GB/app/Healthcare/healthcare?form.insuranceprovider=UnitedHealthcare&form.medicalcondition=Hypertension>

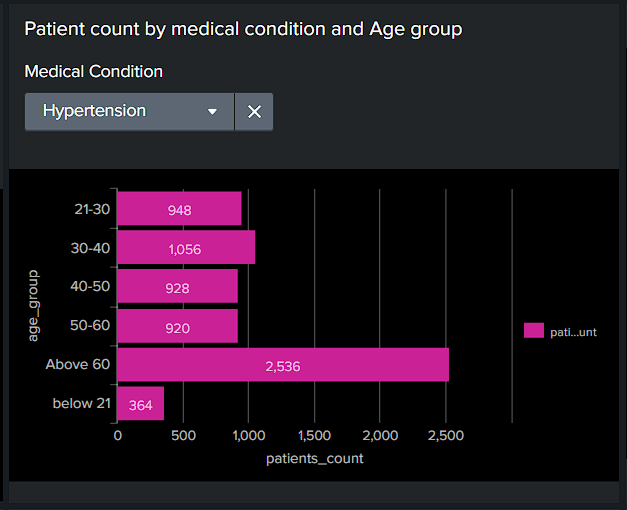
**Report Strategy:**

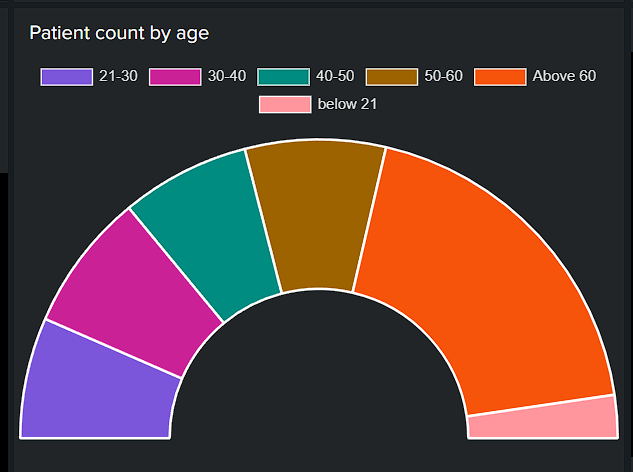
**Scenario 1:** ***Average Billing Amount by Insurance Provider and Medical Condition***  
  
 The goal of this Splunk query is to extract useful information about healthcare billing amounts for Medicare-eligible patients. It determines the average billing amount for every medical condition in the designated dataset, giving Medicare-insured individuals a comprehensive picture of the costs related to various medical disorders. Here user needs to select insurance provider name from the drop down list to fetch the result from dataset.  
  
Query: source="archive.zip:\*" host="upb-spssplunk-01" index="healthcare\_index\_group\_5" "Insurance Provider"=Medicare |stats avg("Billing Amount") as "Average Billing Amount" by "Medical Condition"  
  


In the provided screenshot, the analysis focuses on patients covered by Medicare. Notably, conditions such as Asthma, Diabetes, and Obesity stand out, exhibiting billing amounts exceeding 25,000. These observations shed light on the significant financial aspects associated with these prevalent medical conditions within the Medicare-insured patient population.

**Scenario 2:** ***Patient Count by Medical Condition and Age Group***  
  
 This Splunk query is designed to analyze and categorize patients with selected medical condition based on their age groups. The eval command is used to create a new field called age\_group, which categorizes patients into specific age ranges. The stats command is then employed to calculate the count of patients in each age group.

The breakdown of age groups, such as "below 21," "21-30," "30-40," and so on, provides a detailed distribution of patients with different medical condition across different age categories. User needs to select specific medical condition to fetch results from dataset. This analysis can offer valuable insights into the age demographics of individuals affected by certain medical condition within the specified healthcare dataset.  
  
Query: source="archive.zip:\*" host="upb-spssplunk-01" index="healthcare\_index\_group\_5" "Medical Condition"="Hypertension"| eval age\_group=case( Age >= 18 AND Age < 21, "below 21", Age >= 21 AND Age < 30, "21-30", Age >= 30 AND Age < 40, "30-40", Age >= 40 AND Age < 50, "40-50", Age >= 50 AND Age < 60, "50-60", Age >= 60 AND Age <=85, "Above 60", true(), "Other" ) | stats count as patients\_count by age\_group

  
  
 In the presented data, patients with the selected medical condition 'Hypertension' show that individuals aged above 60 suffer the most, with their count nearly doubling that of the second most affected age group, 30-40. This insight sheds light on the prevalence of hypertension within the specified age demographics.  
  
**Scenario 3:** ***Patient Count by Age***  
  
 The purpose of this Splunk query is to use age groups to classify and examine patient data from the given healthcare dataset. Patients are categorised into various age groups using the eval command, which also creates a new variable called age\_group. To categorise patients into age groups according to their ages, the case function is utilised.

The stats command determines the number of patients in each age group after classifying patients based on their age. This study offers a condensed perspective of the patient distribution within the designated healthcare dataset across different age groups. The number of patients in each age group is included in the output that is produced, providing information on the demographic makeup of the patient population.  
  
Query: source="archive.zip:\*" host="upb-spssplunk-01" index="healthcare\_index\_group\_5" | eval age\_group=case( Age >= 18 AND Age < 21, "below 21", Age >= 21 AND Age < 30, "21-30", Age >= 30 AND Age < 40, "30-40", Age >= 40 AND Age < 50, "40-50", Age >= 50 AND Age < 60, "50-60", Age >= 60 AND Age <=85, "Above 60", true(), "Other" ) | stats count as patients\_count by age\_group  
  


In the visualized data, patients aged over 60 exhibit the highest count, exceeding 15,000. This insight suggests a higher demand for precautionary measures and healthcare support to ensure the well-being of this demographic  
  
  
  
**Scenario 4: *Medical Condition with the Maximum Count for Each Admission Type***

This Splunk query focuses on analyzing healthcare data by counting occurrences of different medical conditions based on admission types. The stats command is utilized to calculate the count of occurrences for each unique combination of "Medical Condition" and "Admission Type."The eventstats command is then employed to find the maximum count for each "Admission Type."   
  
 Subsequently, the query filters the results to include only the rows where the count is equal to the maximum count for each "Admission Type." The final output is presented in a tabular format, displaying the "Admission Type," "Medical Condition," and the corresponding count, sorted in descending order based on the count.This query is designed to identify and highlight the medical conditions that have the highest count for each admission type, offering valuable insights into prevalent conditions associated with different admission scenarios in the healthcare dataset.  
  
Query: source="archive.zip:\*" host="upb-spssplunk-01" index="healthcare\_index\_group\_5"

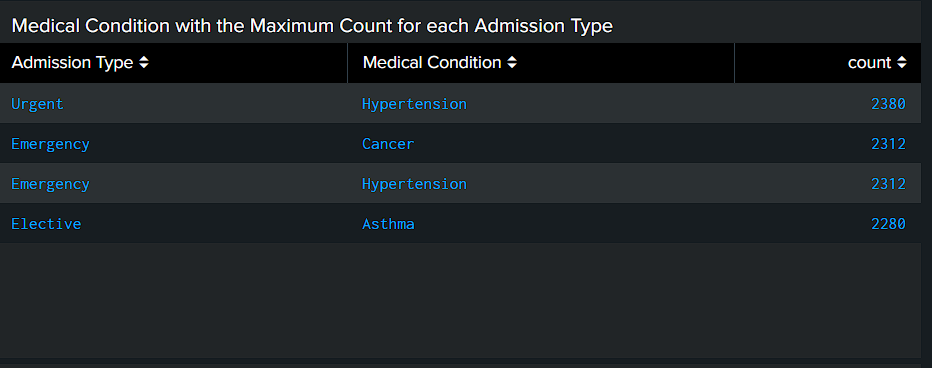
| stats count by "Medical Condition", "Admission Type"

| eventstats max(count) as max\_count by "Admission Type"

| where count = max\_count

| table "Admission Type", "Medical Condition", count

| sort - count

  
  
  
 In the provided statistical table, it's evident that the medical condition 'Hypertension' is associated with the highest count in the 'Urgent' admission type. Aligning with our earlier visualization, where individuals over 60 were observed to be most affected by 'Hypertension,' it indicates that a majority of those requiring 'Urgent' admission were in the above 60 age group. Additionally, interestingly, both 'Cancer' and 'Hypertension' shared an equal count of 2312 for the 'Emergency' admission type.  
 **Scenario 5: *Distribution of Different Test Results ('Inconclusive,' 'Abnormal,' 'Normal')***

This Splunk query is designed to categorize and analyze test results from the specified healthcare dataset. The eval command is utilized to create a new field named result\_category based on the values in the 'Test Results' field. The case function is employed to assign results to specific categories such as "Inconclusive," "Abnormal," "Normal".

Following the categorization, the stats command calculates the count of occurrences for each result category. The query then employs the eval command again to calculate the percentage of each result category relative to the total count. The resulting output provides a clear distribution of test results, with the associated counts and percentages for each category. This analysis aids in understanding the prevalence of different test outcomes within the healthcare dataset.  
  
Query: source="archive.zip:\*" host="upb-spssplunk-01" index="healthcare\_index\_group\_5"

| eval result\_category=case(

'Test Results'="Inconclusive", "Inconclusive",

'Test Results'="Abnormal", "Abnormal",

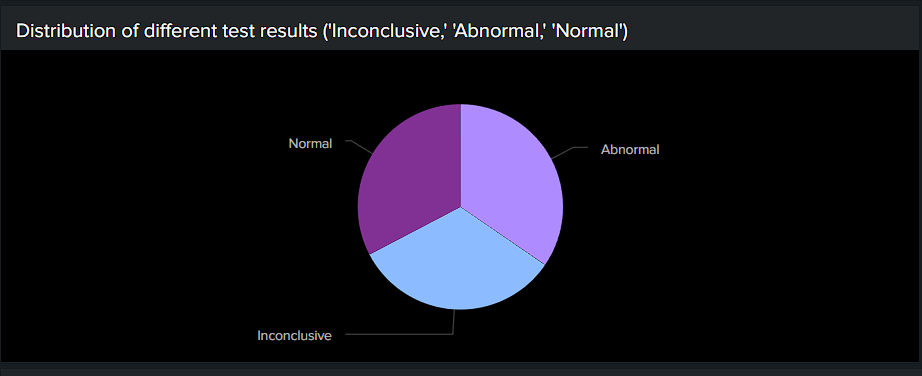
'Test Results'="Normal", "Normal",

true(), "Other"

)

| stats count by result\_category

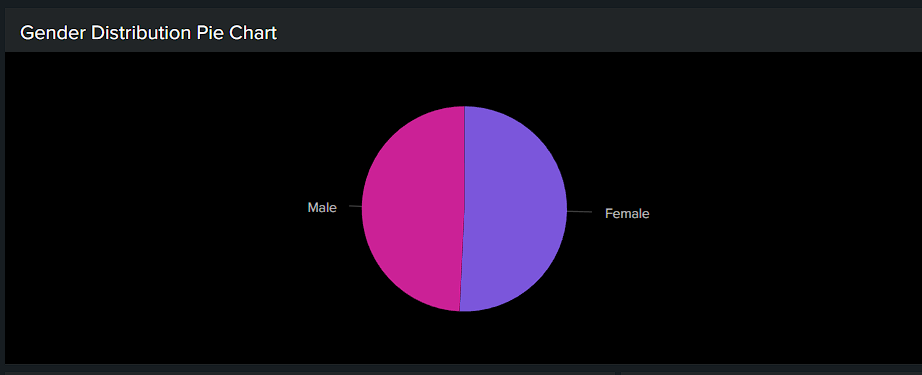
| eval Percentage = count / sum(count) \* 100

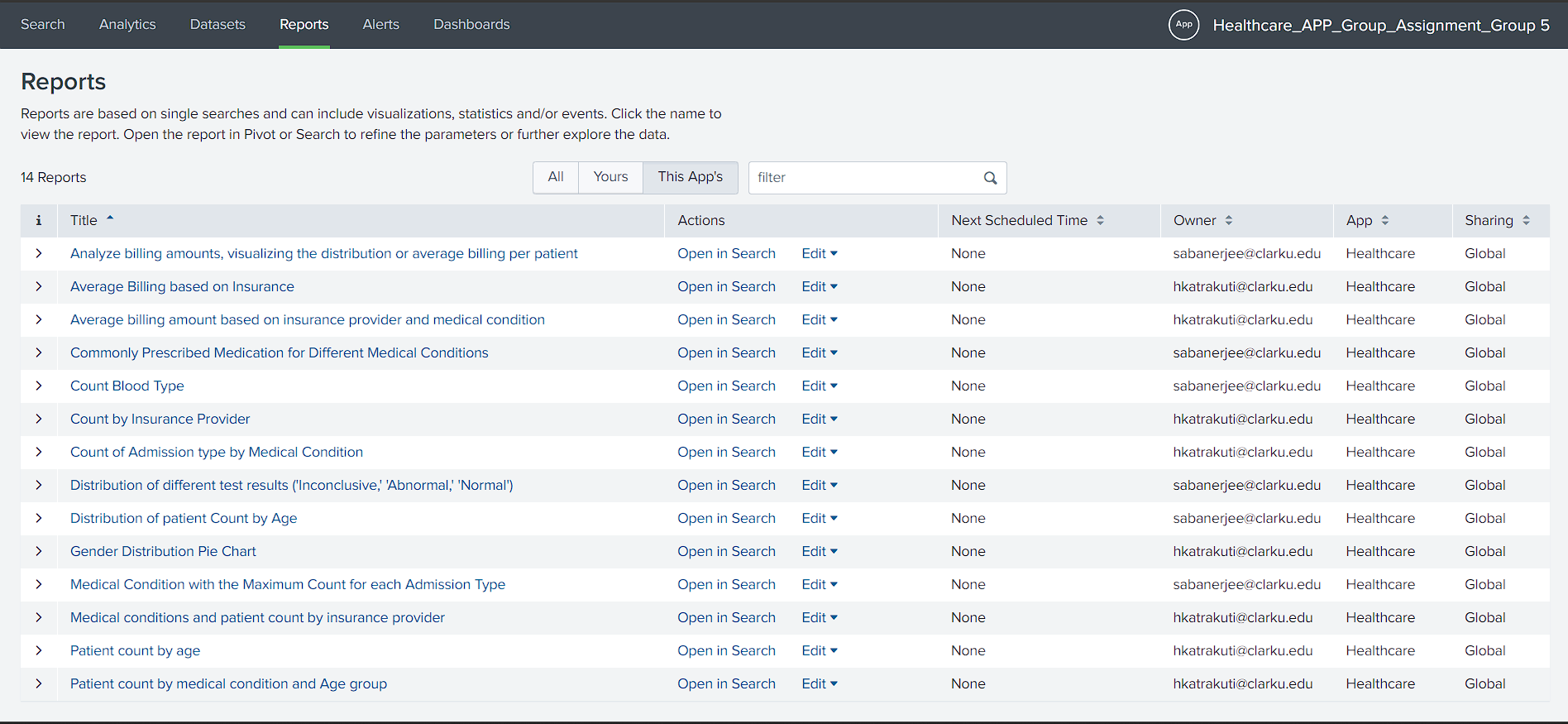
  
  
   
 Examining patient test results reveals that approximately 32% of the total dataset results are inconclusive, raising concerns in healthcare. To prioritize patient well-being, our goal should be to reduce this percentage below 25% next year, with an ultimate target of around 5%. Inconclusive results pose potential risks to patients, necessitating an investigation into the root causes and, if needed, the implementation of efficient testing procedures and machines across hospitals.

**Scenario 6: *Gender Distribution Pie Chart***

The Report Strategy uses a Splunk-generated Gender Distribution Pie Chart to show healthcare dataset gender allocation. It helps identify gender representation trends and imbalances quickly. Healthcare personnel may make patient care, resource allocation, and strategic planning decisions using this visualization. Initial gender-based analysis using it enables more focused healthcare research and precise treatments, resulting in better informed healthcare plans and results.  
  
Query: source="archive.zip:\*" host="upb-spssplunk-01" index="healthcare\_index\_group\_5"

| stats count by Gender

| chart sum(count) as Count by Gender  
  
  
  
  
  
  
**Scenario 7:**

**Report Overview:**  
  
  
  
  
**Report link: <https://splunk.sps.cs.clarku.edu:8000/en-GB/app/Healthcare/reports>**