DSC465

Data Visualization

Project Final Report

Health Analyst

A. Introduction

Our project centers on a synthetic healthcare dataset obtained from Kaggle, crafted to emulate real-world medical records. Utilizing Python's Faker Library, the dataset replicates authentic healthcare data. In this report, we analyzed various healthcare records to derive insights into healthcare trends. The dataset, sourced from Kaggle, comprises medical records for 10,000 patients. (https://www.kaggle.com/datasets/prasad22/healthcare-dataset)

1. Data Description

Our Dataset contains medical records for 10000 patients. The variables we analyzed in our healthcare dataset are as follows:

- Name: This represents the name of the patient associated with the healthcare record.
- Date of Admission: The date on which the patient was admitted to the healthcare facility.
- Medical Condition: This column specifies the primary medical condition or diagnosis associated with the patient, such as "Diabetes," "Hypertension," "Asthma," and more.
- Insurance Provider: This column indicates the patient's insurance provider, which can be one of several options, including "Aetna," "Blue Cross," "Cigna," "UnitedHealthcare," and "Medicare."
- Billing Amount: The amount of money billed for the patient's healthcare services during their admission. This is expressed as a floating-point number.
- Medication: Identifies a medication prescribed or administered to the patient during their admission. Examples include "Aspirin," "Ibuprofen," "Penicillin," "Paracetamol," and "Lipitor."
- Test Results: Describes the results of a medical test conducted during the patient's admission. Possible values include "Normal," "Abnormal," or "Inconclusive," indicating the outcome of the test.

2. Purpose

The purpose of our project is to explore and discover patterns from the following graphs:

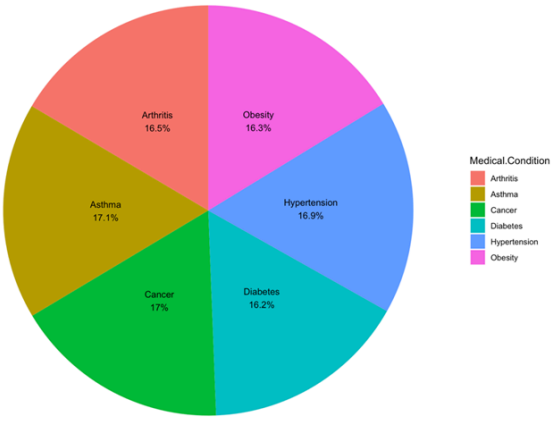
1. Chord diagram
2. Mosaic plot
3. Calendar Heatmap
4. HeatMap
5. Scatter plot
6. 2D Density Plot

We have selected two paths for exploration. Since our data is artificially generated, we initially analyzed the distribution of patients across medical record variables and their distribution over time periods and also their relationship across the billing amount.

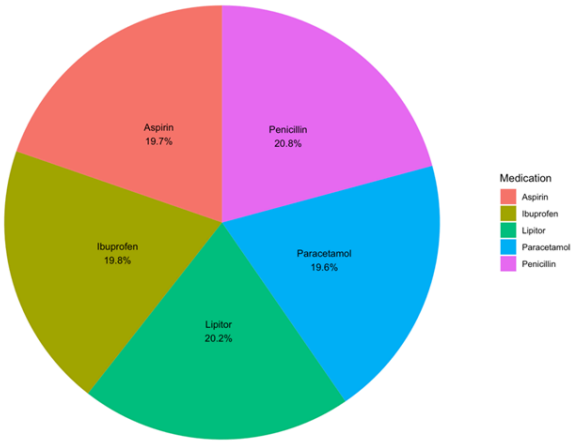
B. Exploratory Analysis

As part of our initial exploratory analysis, we conducted a series of pie charts and bar charts to gain insights into the distribution of our data across categorical variables. These visualizations served as a foundational step in understanding the structure of our dataset.

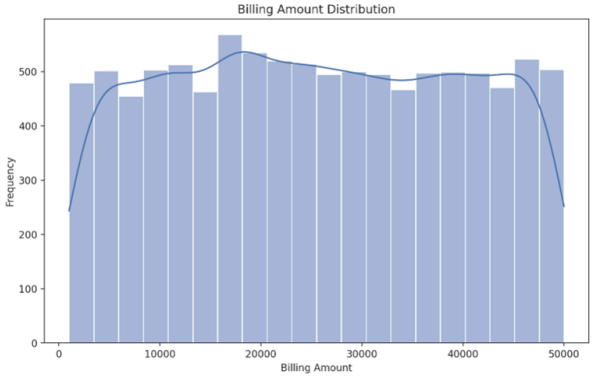
(a)



(b)



(c)



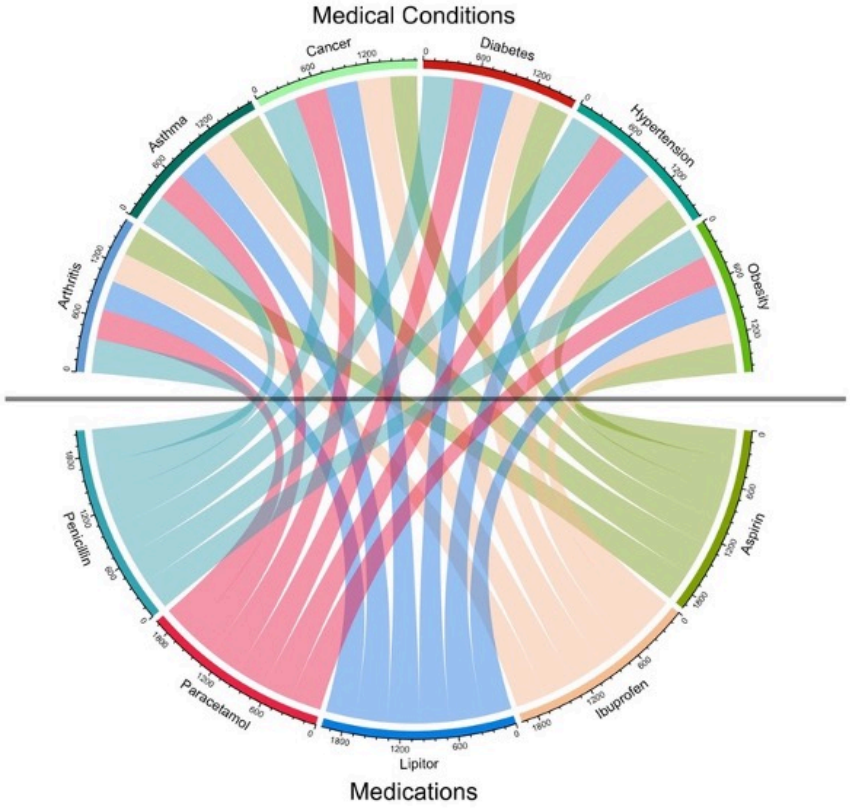
From the pie charts depicting medical conditions (a) and medications (b), we observed a uniform distribution across all categories. This consistency extended to other categorical variables we explored, indicating similar frequency occurrences and distributions. Consequently, we recognized the potential to showcase this uniform distribution as a central theme in our analysis.

However, one variable stood out from the others in terms of distribution: the billing amount. The bar chart (c) illustrating the distribution of billing amounts revealed notable discrepancies in frequency. This observation prompted us to add another direction to explore insights related to financial trends within the dataset.

C. Visualizations

1. Chord Diagram:

One of our pivotal visualizations created was a chord diagram using the circlize package in R. This visualization depicted the relationship between medical conditions and medications based on their frequency of occurrence in the dataset.



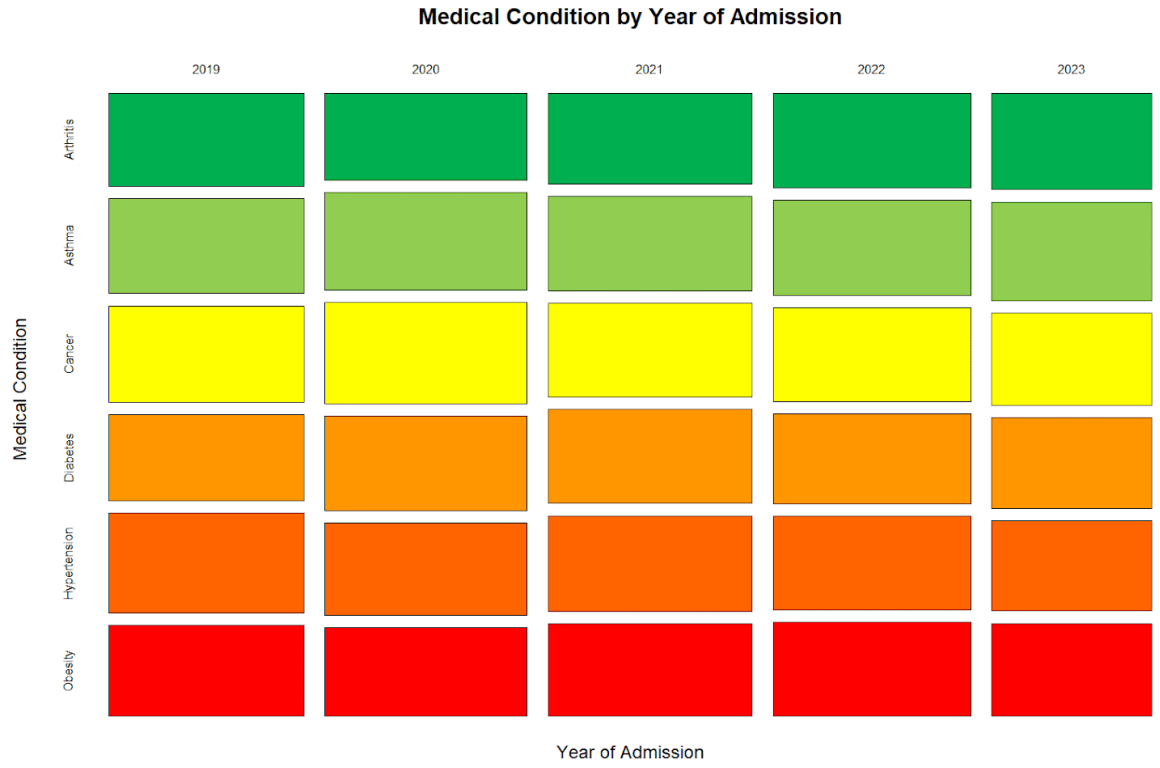
In the upper portion of the diagram, Medical Conditions such as Arthritis, Asthma, Cancer, Diabetes, Hypertension, and Obesity are represented. Medications like Penicillin, Paracetamol, Lipitor, Ibuprofen, and Aspirin are displayed in the lower portion.

Each ribbon connecting a Medication to a Medical Condition indicates the frequency of that specific combination. The ribbon color was chosen to represent the medication used, with distinct colors assigned to each medication category to show their difference. For example, Penicillin is represented by a sky-blue color, Paracetamol by red, Lipitor by blue, Ibuprofen by skin color, and Aspirin by green. The ribbon's width or arc length represents the occurrence frequency, with wider ribbons indicating higher frequencies.

While our dataset may exhibit a uniform distribution, no discernible trend is evident with all medications connected to each medical condition. However, this chord diagram serves as a valuable tool for uncovering trends in medication usage patterns across different medical conditions. Analyzing such trends can offer insights into treatment preferences and associations within the dataset, aiding healthcare providers in making informed decisions and optimizing patient care.

2. Mosaic Plot:

From the earlier analysis, we observed a uniform distribution of medical conditions and medications. To delve deeper, we investigated the distribution of patients admitted each year based on their medical condition. We utilized a mosaic plot, mapping medical conditions on the y-axis and the year of admission on the x-axis.

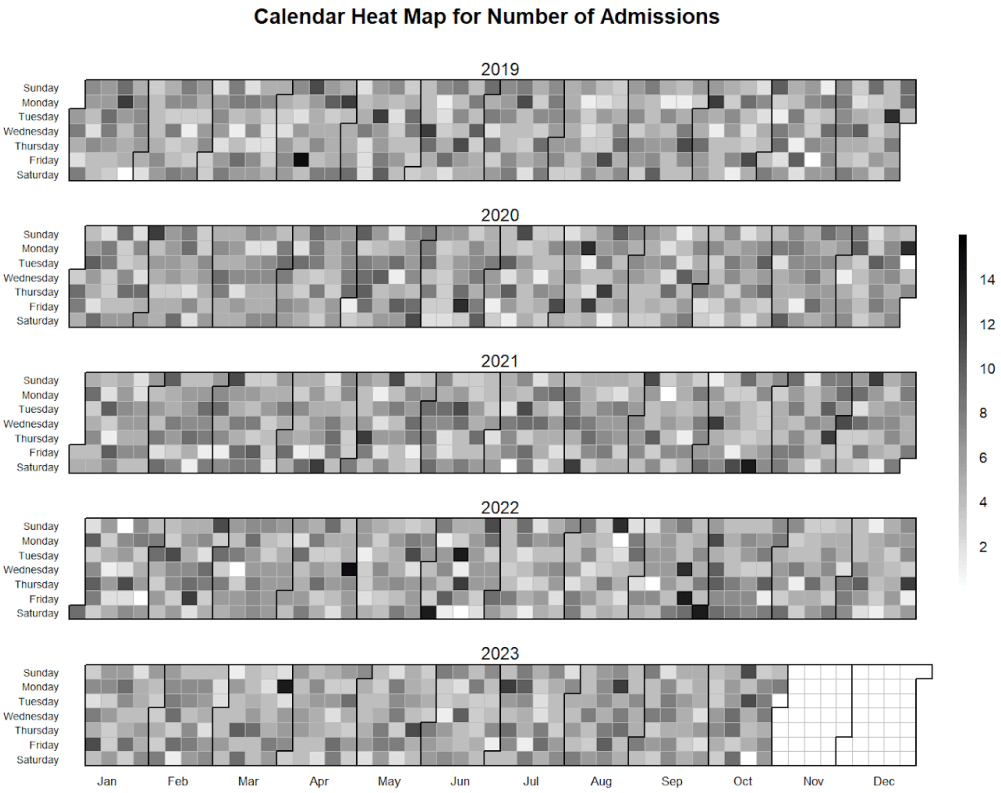


The mosaic plot reveals a relatively even distribution of medical conditions across the years of admission. Initially, the plot lacked clarity, as the medical conditions were not differentiated by color. To enhance comparability, distinct colors were assigned to each medical condition.

While there are slight variations, such as a decrease in Asthma cases from 2021 to 2023, and a slight increase in Diabetes cases from 2019 to 2020, these differences are minimal. Overall, the plot suggests a nearly equal distribution of medical conditions across each year, akin to the distribution of medications across different medical conditions.

3. Calendar Heatmap:

Given the uniform distribution of medical conditions over the years, we sought to explore the distribution of admissions across individual days. We employed a calendar heatmap to visualize the count of patients admitted throughout the period of 2019-2023.



The calendar heatmap displays varying shades of color, with lighter shades indicating fewer admissions and darker shades representing higher admission counts. Each year is delineated into its respective calendar, with the y-axis representing the days of the week and the x-axis representing the months.

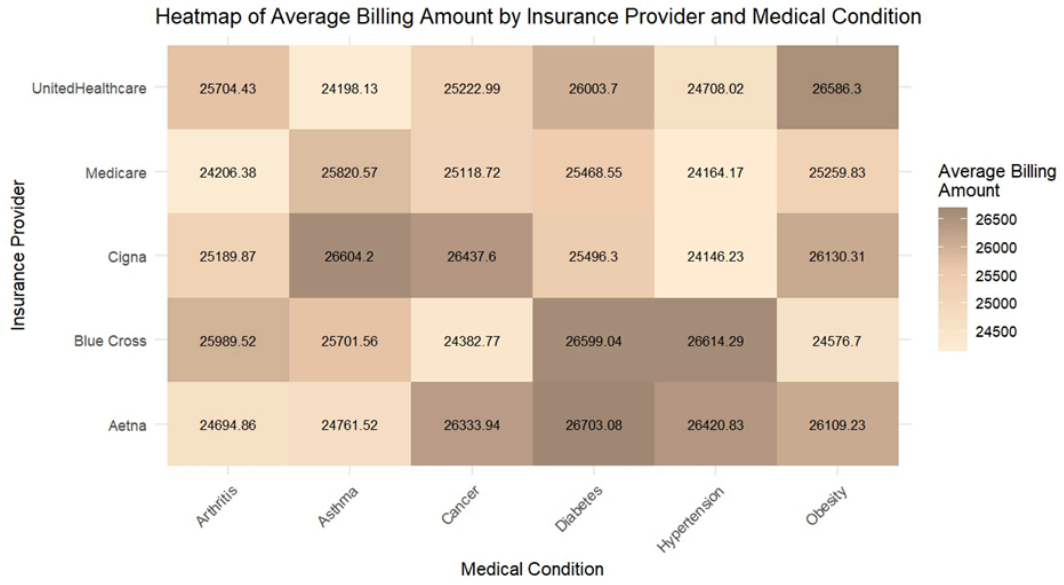
Initially, a diverging color scheme was utilized, but it proved challenging to identify days with the highest and lowest admission counts. Consequently, a sequential color scale from white to black was adopted, facilitating clearer trend identification.

The heatmap reveals a non-uniform distribution of admissions across the years, with no discernible seasonal or weekly trends apparent at first glance. However, upon closer examination, subtle patterns emerge. Darker shades tend to cluster in December, January, and February, suggesting increased admissions during colder months. Additionally, Saturdays consistently exhibit darker shades, indicating higher admission counts compared to other days, potentially reflecting a tendency for individuals to seek hospital admission on weekends.

While these differences are subtle due to the synthetic nature of our dataset, they offer insights into admission patterns that may inform further analysis.

4. Heatmap

Based on our previous visualizations, which indicated uniform distributions, we aim to investigate whether this uniformity persists when plotting variables such as insurance providers and medical conditions against billing amounts.



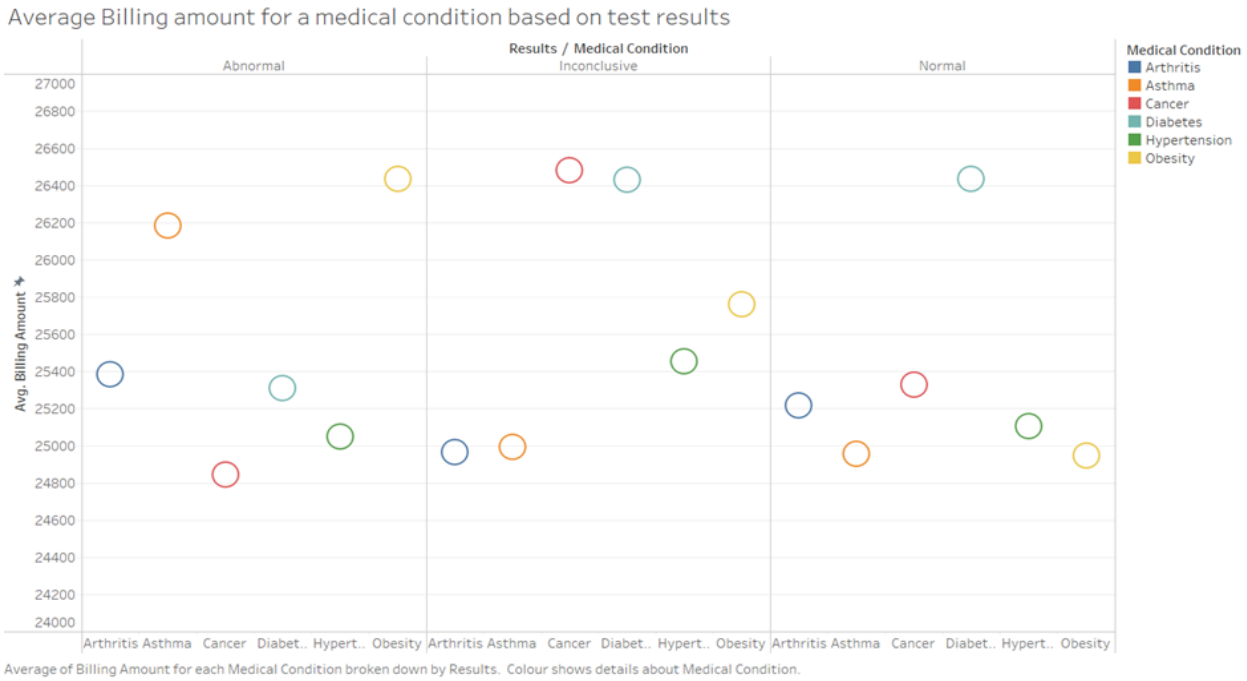
We generated a heatmap with medical conditions on the x-axis and insurance providers on the y-axis, depicting the average billing amount using a color legend. This heat map illustrates the average billing amount associated with each medical condition, categorized by insurance provider.

Initially, we created bar charts for each medication plotted against insurance providers. However, we later opted for a heatmap visualization. Initially, the color scale ranged from 0 to the highest billing amount, resulting in similar shades of color. To enhance clarity, we adjusted the color scale to span from the smallest to the highest variable available, ensuring a sequential representation.

Our analysis revealed interesting patterns. For instance, for cancer patients, UnitedHealthcare insurance tended to yield lower billing amounts, while Cigna was less favorable. Similarly, for hypertension, avoiding Blue Cross and opting for Cigna appeared beneficial. However, on closer examination, we observed minimal variations in average billing amounts overall, suggesting a relatively uniform distribution across insurance providers.

5. Scatterplot

The HeatMap presented above offers a comprehensive understanding of billing amounts across various medical conditions provided by each insurance provider. The below scatter plot helps us further understand the billing amounts of various medical conditions by illustrating a relationship between test results and average billing amounts across various medical conditions. We can further look at the visualization to better understand the trends and potential insights crucial for decision-making.



This visualization is based on test results and the average billing amount for various medical conditions. Because of its ability to visualize two variables and highlight connections or patterns

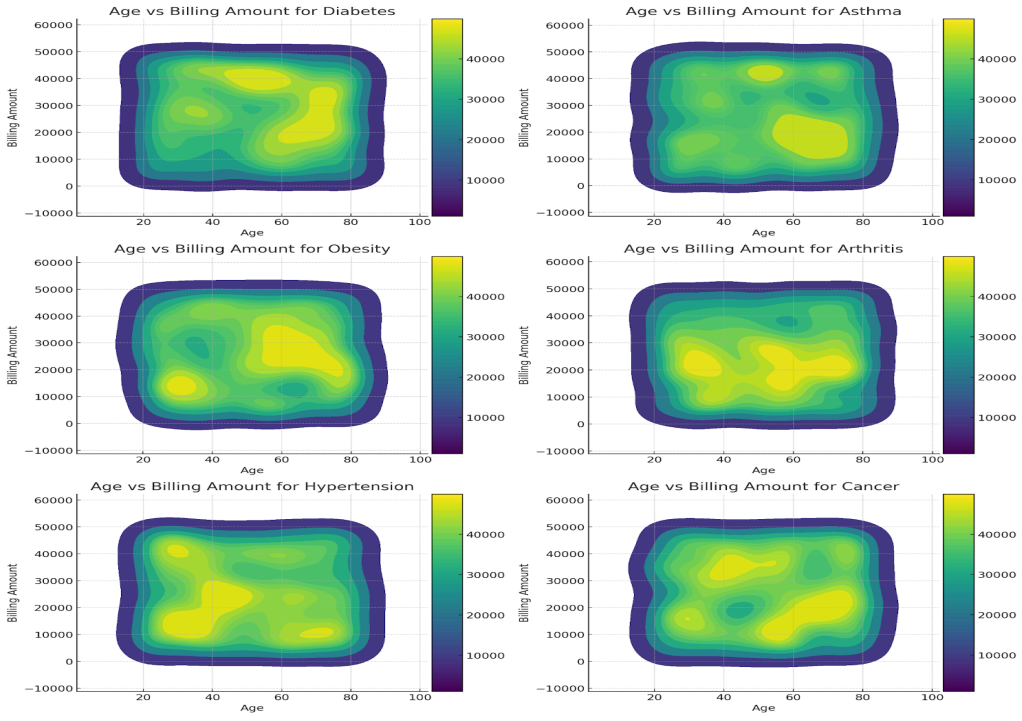
between them, the scatter plot was selected. It's appropriate since it provides easily interpretable information on trends and possible insights.

The above scatter plot was achieved after going through a sequence of drafts. At first, a bar graph was plotted, but it wasn't insightful. So the data was then visualized as a scatter plot and even the scatter plot wasn't conveying any trend or pattern. So, at last, the y-axis was changed, and after the change, we can see a pattern and interpret the values.

From the visualization, we can say that the average billing amount varies from $24,000 to $27,000, and for certain illnesses, such as arthritis(blue), remains relatively constant regardless of the test outcome. This could imply that the expense of treating arthritis is significant even in cases where the tests come out normal. The amounts that are billed for various conditions differ. For instance, treating hypertension seems to have a lower average cost than treating cancer. There is a noticeable increase in the average billing amount from normal to abnormal test findings for certain illnesses, such as obesity (yellow). This may suggest that when test findings are abnormal, more rigorous treatment is required, which would increase expenditures.

6. 2D density Plot:

As we are more focused on cost-effectiveness, let's proceed to dive deeper into the concepts. The density plots illustrate the distribution of billing amounts across different age groups for six medical conditions: Diabetes, Asthma, Obesity, Arthritis, Hypertension, and Cancer.



First, we focused on getting insights about the billing amounts and age limits, so we created a bar graph of different ages patients with medical conditions, and then we created for billing amounts, we noticed that there might be a correlation among the diseases and the billing amounts and we can also get some insights for the specific age limit groups with same medical conditions.

There we generated the 2d Density plots where we have age distributions from ages till 90 years starting from 15 on the X-axis, and then the charges and the billing amounts are on the Y-axis where the density is based on the number of patients for specific age and their billing charges.

We have equally distributed data, which we can clearly see from the density ranges. The Outlines, which are in blue color, show fewer data points where they go yellowish as the density of data points increases. Finally, we can conclude from these plots that healthcare providers and policymakers with valuable information on which age groups may require more financial resources and medical attention for specific conditions. Additionally, these insights could inform targeted prevention and management strategies for these conditions across different age demographics.

D. Analysis and Discussion

We discovered the following from our graphs:

1. The chord diagram revealed the frequency of occurrence for different combinations of medical conditions and medications, providing insights into treatment preferences and associations within the dataset. We showed this through the widths of the ribbons connecting medications to medical conditions, with wider ribbons indicating higher frequencies of that combination. While our dataset may exhibit a uniform distribution, no discernible trend exists with all medications connected to each medical condition. However, this chord diagram can serve as a valuable tool for uncovering trends in medication usage patterns across different medical conditions.
2. The mosaic plot showed a relatively even distribution of medical conditions across the years of admission, with only minor variations observed. We visualized this by mapping medical conditions on the y-axis and years of admission on the x-axis, with the area of each rectangle representing the frequency of that condition in that year.
3. The calendar heatmap uncovered subtle patterns in patient admissions, such as higher admission counts during colder months and on Saturdays. We depicted this using a sequential color scale from white to black, with darker shades representing higher admission counts, allowing patterns to emerge across days, weeks, and months.
4. The heatmap comparing medical conditions and insurance providers suggested a relatively uniform distribution of average billing amounts across insurance providers. We showed this by plotting medical conditions on the x-axis and insurance providers on the

y-axis and using a color legend to represent the average billing amount for each combination.

5. The scatterplot between test results and average billing amounts revealed patterns like higher billing for abnormal test results and varying average costs across different conditions. We demonstrated this by plotting test results on the x-axis and average billing amounts on the y-axis, allowing patterns to be observed based on the distribution of data points.
6. The 2D density plots highlighted the distribution of billing amounts across different age groups for specific medical conditions, informing targeted prevention and management strategies. We visualized this by creating 2D density plots with age on the x-axis and billing amounts on the y-axis, with color gradients representing the density of data points for each combination of age and billing amount.

1. Code

```
# Load required libraries
library(circlize)
library(dplyr)
library(lubridate)
library(ggplot2)
library(tidyr)

# Read dataset
healthcare <- read.csv("healthcare_dataset.csv")

# Calculate the frequency of each combination of medication and medical condition for Chord Diagram
Medication_freq <- healthcare %>%
  group_by(Medication, Medical.Condition) %>%
  summarise(Frequency = n()) %>%
  ungroup()

# Creates the chord diagram for medications by medical conditions
chordDiagram(Medication_freq, transparency = 0.5)


# Convert Date of Admission to Date format for Mosaic Plot and Calendar Heatmap
healthcare$Date.of.Admission <- as.Date(healthcare$Date.of.Admission)

# Create Mosaic Table for Medical Condition by Year of Admission
mosaic_table <- table(year(healthcare$Date.of.Admission), healthcare$Medical.Condition)

# Define Color Palette for Mosaic Plot
color_palette <- c(
  "Obesity" = "#00B050",
```

```
  "Hypertension" = "#92D050",
  "Diabetes" = "#FFFF00",
  "Cancer" = "#FF9900",
  "Asthma" = "#FF6600",
  "Arthritis" = "#FF0000"
)

# Plot Mosaic Plot
mosaicplot(mosaic_table,
           main = "Medical Condition by Year of Admission",
           color = color_palette,
           legend = TRUE,
           xlab = "Year of Admission",
           ylab = "Medical Condition")

# Generate Calendar Heatmap for Admission Counts
admission_counts <- table(healthcare$Date.of.Admission)
admission_counts_df <- data.frame(Date = as.Date(names(admission_counts)), Count =
as.numeric(admission_counts))
source("calendarHeat.R")
calendarHeat(admission_counts_df$Date, admission_counts_df$Count, varname = "Number of Admissions")

# Create 2D Density Plot for Age vs. Billing Amount by Medical Condition
healthcare$Age <- as.numeric(as.character(healthcare$Age))
healthcare$Billing.Amount <- as.numeric(as.character(healthcare$Billing.Amount))

p <- ggplot(healthcare, aes(x = Age, y = Billing.Amount)) +
  stat_density_2d(aes(fill = ..level..), geom = "polygon") +
  scale_fill_viridis_c() +
  facet_wrap(~ Medical.Condition, scales = "free") +
  theme_minimal() +
  labs(title = "Age vs Billing Amount by Medical Condition",
       x = "Age", y = "Billing Amount")
print(p)

# Create Heatmap of Average Billing Amount by Insurance Provider and Medical Condition
heatmap_data <- healthcare %>%
  group_by(Insurance.Provider, Medical.Condition) %>%
  summarise(AverageBillingAmount = mean(Billing.Amount, na.rm = TRUE)) %>%
  ungroup()
heatmap_data_wide <- heatmap_data %>%
  pivot_wider(names_from = Medical.Condition, values_from = AverageBillingAmount)

median_value <- median(heatmap_data$AverageBillingAmount, na.rm = TRUE)

ggplot(heatmap_data, aes(x= Medical.Condition, y= Insurance.Provider, fill=AverageBillingAmount)) +
  geom_tile() +
  geom_text(aes(label=round(AverageBillingAmount, 2)), color="black", size=3) +
```

```
  scale_fill_gradient2(low = "papayawhip", high = "peachpuff4", mid = "peachpuff2", midpoint = median_value,
space = "Lab",
                       name="Average Billing\nAmount") +
  labs(title = "Heatmap of Average Billing Amount by Insurance Provider and Medical Condition",
       x = "Medical Condition", y = "Insurance Provider") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

2. References:

- https://www.kaggle.com/datasets/prasad22/healthcare-dataset/data
- https://gist.github.com/aaronwolen/4d93d7e676ed01351afb
- https://r-graph-gallery.com/123-circular-plot-circlize-package-2.html