Penda Health

Pius Mbeti

2023-07-23

INTRODUCTION

In this report, I present an analysis of data from Penda Health Clinic. The data includes information on medical visits, revenue, and diagnoses recorded at the clinic. We will use R for data processing, Exploratory Data Analysis (EDA), and analysis, and then create visualizations to gain insights into the clinic's performance.

DATA LOADING AND OVERVIEW

Before the analysis, let's first take a look at the structure of the data.

```
setwd("C:/Users/Pius/OneDrive/Desktop/Penda health Project")



# Set warning behavior to ignore
options(warn = -1)  # -1 suppresses all warnings







# Read the CSV files into data frames
visit_tbl <- read.csv("Visit_Tbl.csv")
invoice_tbl <- read.csv("Invoice_Tbl.csv")
diagnosis_tbl <- read.csv("Diagnosis_Tbl.csv")

# Merging the three tableS
merged_data <- merge(visit_tbl, invoice_tbl, by = "VisitCode", all.x = TRUE)
merged_data <- merge(merged_data, diagnosis_tbl, by = "VisitCode", all.x = TRUE)
head(merged_data)
```

```
##    VisitCode   PatientCode    VisitDateTime MedicalCenter   VisitCategory
## 1 XA-1060253 65503b74-84c1 01/03/2022 00:19      Pipeline In-person Visit
## 2 XA-1060256 091fec77-3906 01/03/2022 01:35      Pipeline In-person Visit
## 3 XA-1060258 4310a085-9c8a 01/03/2022 02:04        Tassia In-person Visit
## 4 XA-1060260 52601bd1-c12c 01/03/2022 02:56      Pipeline In-person Visit
## 5 XA-1060267 36bfabee-26a1 01/03/2022 05:46        Tassia In-person Visit
## 6 XA-1060274 274d97da-04d1 01/03/2022 06:34      Pipeline In-person Visit
##                 Payor NPS.Score Amount                     Diagnosis
## 1                Cash        NA    870                          <NA>
## 2                Cash        NA   2087 tonsillitis, acute bacterial
```

```
## 3                Cash        NA    750                         <NA>
## 4                Cash        NA   2522             acute bronchitis
## 5                Cash        NA     48                         <NA>
## 6 Insurance Company B        NA   4183                         <NA>
```

```
str(merged_data)
```

```
## 'data.frame':    48147 obs. of  9 variables:
##  $ VisitCode    : chr  "XA-1060253" "XA-1060256" "XA-1060258" "XA-1060260" ...
##  $ PatientCode  : chr  "65503b74-84c1" "091fec77-3906" "4310a085-9c8a" "52601bd1-c12c" ...
##  $ VisitDateTime: chr  "01/03/2022 00:19" "01/03/2022 01:35" "01/03/2022 02:04" "01/03/2022 02:56" .
##  $ MedicalCenter: chr  "Pipeline" "Pipeline" "Tassia" "Pipeline" ...
##  $ VisitCategory: chr  "In-person Visit" "In-person Visit" "In-person Visit" "In-person Visit" ...
##  $ Payor        : chr  "Cash" "Cash" "Cash" "Cash" ...
##  $ NPS.Score    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Amount       : int  870 2087 750 2522 48 4183 2250 1840 3230 3390 ...
##  $ Diagnosis    : chr  NA "tonsillitis, acute bacterial" NA "acute bronchitis" ...
```

Exploratory Data Analysis (EDA)
Exploring the data and understanding its characteristics.
Summary statistics for merged_data

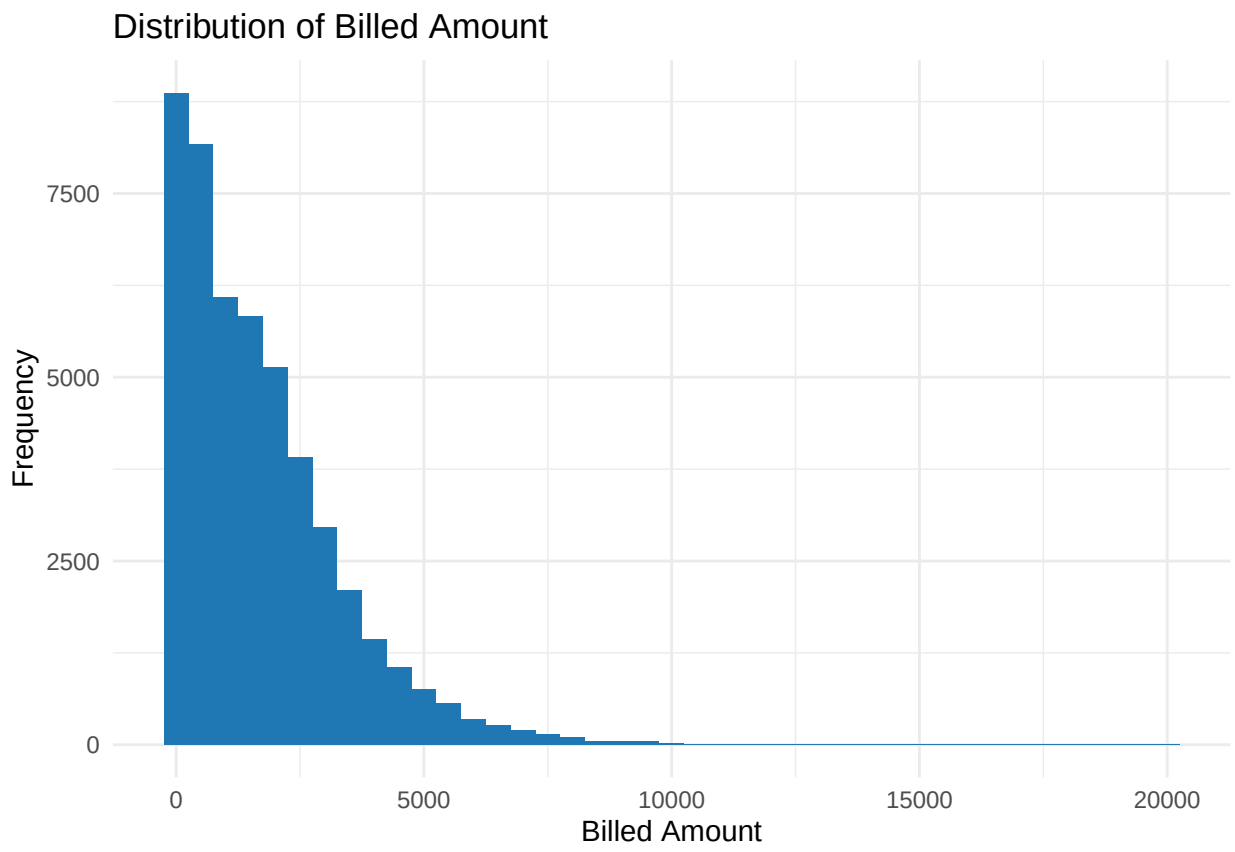
```
summary(merged_data)
```

```
##   VisitCode         PatientCode        VisitDateTime      MedicalCenter
##  Length:48147       Length:48147       Length:48147       Length:48147
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  VisitCategory         Payor             NPS.Score         Amount
##  Length:48147       Length:48147       Min.   : 0.00   Min.   :    0
##  Class :character   Class :character   1st Qu.: 9.00   1st Qu.:  375
##  Mode  :character   Mode  :character   Median :10.00   Median : 1340
##                                        Mean   : 8.84   Mean   : 1715
##                                        3rd Qu.:10.00   3rd Qu.: 2500
##                                        Max.   :11.00   Max.   :20059
##                                        NA's   :46125
##   Diagnosis
##  Length:48147
##  Class :character
##  Mode  :character
##
##
##
##
```

Explore the distribution of billed amount

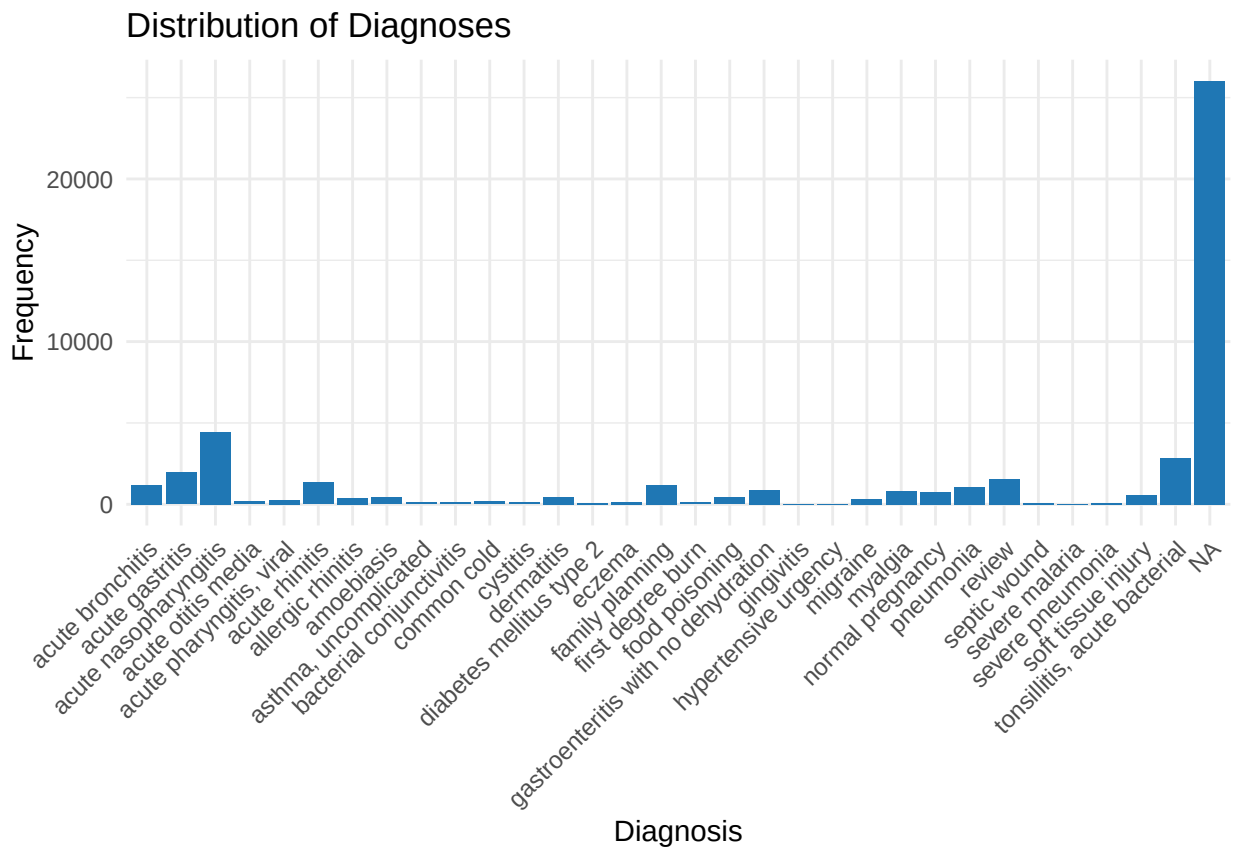
```
library(ggplot2)

ggplot(merged_data, aes(x = Amount)) +
  geom_histogram(binwidth = 500, fill = "#1f77b4") +
  labs(title = "Distribution of Billed Amount",
       x = "Billed Amount",
       y = "Frequency") +
  theme_minimal()
```



Explore the distribution of diagnoses

```
ggplot(merged_data, aes(x = Diagnosis)) +
  geom_bar(fill = "#1f77b4") +
  labs(title = "Distribution of Diagnoses",
       x = "Diagnosis",
       y = "Frequency") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))
```



Data Processing and Wrangling
Before creating visualizations, I need to process and clean the data to ensure it's in the right format.

```
# Convert "VisitDateTime" to a proper date-time format
merged_data$VisitDateTime <- as.POSIXct(merged_data$VisitDateTime, format = "%m/%d/%Y %H:%M")

# Convert "NPS.Score" to numeric and handle missing values
merged_data$NPS.Score <- as.numeric(merged_data$NPS.Score)

# Convert "Amount" to numeric
merged_data$Amount <- as.numeric(merged_data$Amount)

# Handle missing values in the "Diagnosis" column
merged_data$Diagnosis[is.na(merged_data$Diagnosis)] <- "Unknown"

head(merged_data)
```

```
##    VisitCode   PatientCode       VisitDateTime MedicalCenter   VisitCategory
## 1 XA-1060253 65503b74-84c1 2022-01-03 00:19:00      Pipeline In-person Visit
## 2 XA-1060256 091fec77-3906 2022-01-03 01:35:00      Pipeline In-person Visit
## 3 XA-1060258 4310a085-9c8a 2022-01-03 02:04:00        Tassia In-person Visit
## 4 XA-1060260 52601bd1-c12c 2022-01-03 02:56:00      Pipeline In-person Visit
## 5 XA-1060267 36bfabee-26a1 2022-01-03 05:46:00        Tassia In-person Visit
## 6 XA-1060274 274d97da-04d1 2022-01-03 06:34:00      Pipeline In-person Visit
##                Payor NPS.Score Amount                   Diagnosis
## 1               Cash        NA    870                     Unknown
```

```
## 2                Cash        NA   2087 tonsillitis, acute bacterial
## 3                Cash        NA    750                       Unknown
## 4                Cash        NA   2522             acute bronchitis
## 5                Cash        NA     48                       Unknown
## 6 Insurance Company B        NA   4183                       Unknown
```

```
str(merged_data)
```

```
## 'data.frame':    48147 obs. of  9 variables:
##  $ VisitCode    : chr  "XA-1060253" "XA-1060256" "XA-1060258" "XA-1060260" ...
##  $ PatientCode  : chr  "65503b74-84c1" "091fec77-3906" "4310a085-9c8a" "52601bd1-c12c" ...
##  $ VisitDateTime: POSIXct, format: "2022-01-03 00:19:00" "2022-01-03 01:35:00" ...
##  $ MedicalCenter: chr  "Pipeline" "Pipeline" "Tassia" "Pipeline" ...
##  $ VisitCategory: chr  "In-person Visit" "In-person Visit" "In-person Visit" "In-person Visit" ...
##  $ Payor        : chr  "Cash" "Cash" "Cash" "Cash" ...
##  $ NPS.Score    : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Amount       : num  870 2087 750 2522 48 ...
##  $ Diagnosis    : chr  "Unknown" "tonsillitis, acute bacterial" "Unknown" "acute bronchitis" ...
```

DATA VISUALIZATION \ *Key Performance Indicators*
Calculating and displaying some key performance indicators for the clinic.

```
# Load required libraries

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Calculating KPIs
total_visits <- sum(merged_data$VisitCategory == "In-person Visit" | merged_data$VisitCategory == "Tele
total_revenue <- sum(merged_data$Amount)
total_patients <- n_distinct(merged_data$PatientCode)

# Display KPIs
cat("Total Visits: ", total_visits, "\n")
```

```
## Total Visits:  48147
```

```
cat("Total Revenue: Sh.", total_revenue, "\n")
```

```
## Total Revenue: Sh. 82593465
```

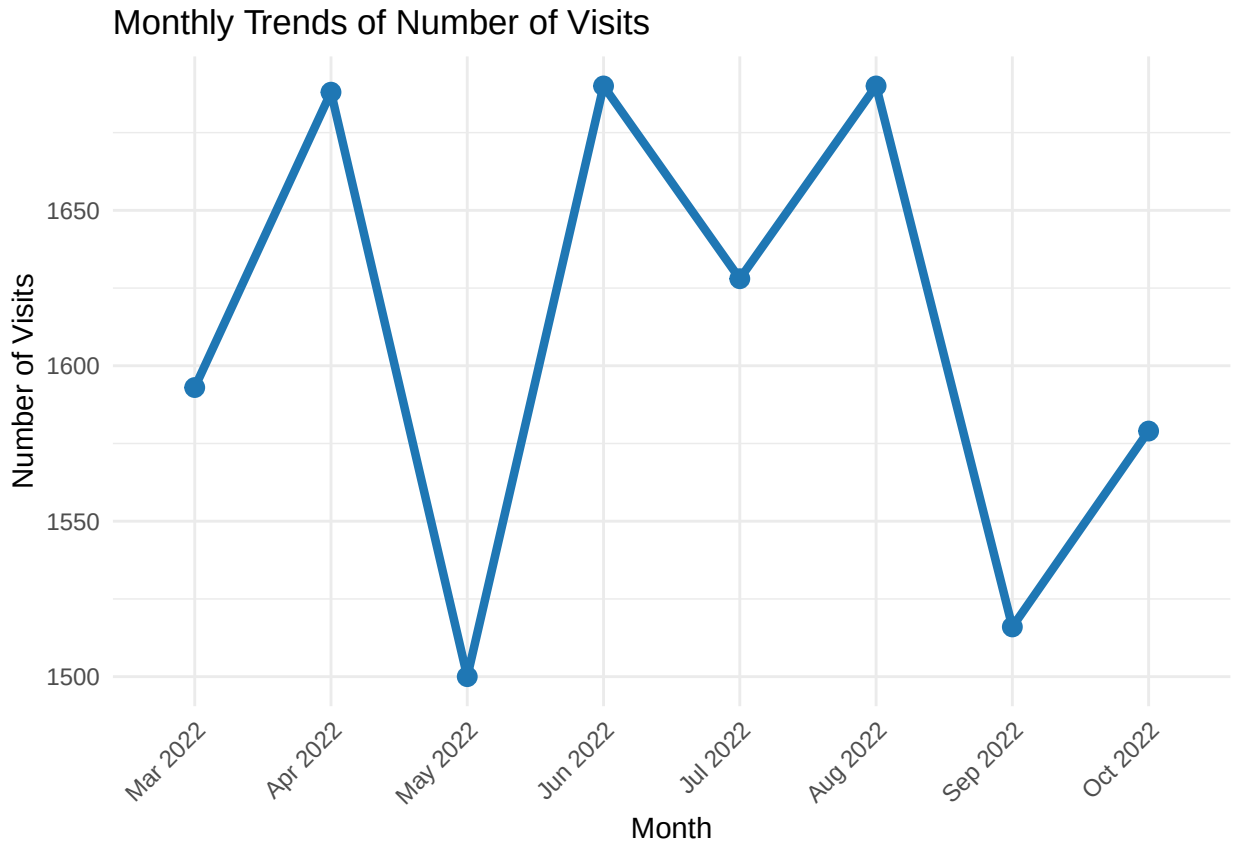
```r
cat("Total Patients: ", total_patients, "\n")
```

```
## Total Patients:  28416
```

Monthly Trends of Number of Visits

```r
# Data preparation for monthly trends
merged_data$VisitDateTime <- as.Date(merged_data$VisitDateTime, format = "%m/%d/%Y %H:%M")
monthly_visits <- merged_data %>%
  filter(format(VisitDateTime, "%b") %in% c("Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct")) %>%
  group_by(Month = factor(format(VisitDateTime, "%b %Y"), levels = c("Mar 2022", "Apr 2022", "May 2022"
  summarise(Visits = n())

# Visualization: Line Graph
ggplot(monthly_visits, aes(x = Month, y = Visits, group = 1)) +
  geom_line(color = "#1f77b4", size = 1.5) +
  geom_point(color = "#1f77b4", size = 3) +
  labs(title = "Monthly Trends of Number of Visits",
       x = "Month",
       y = "Number of Visits") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



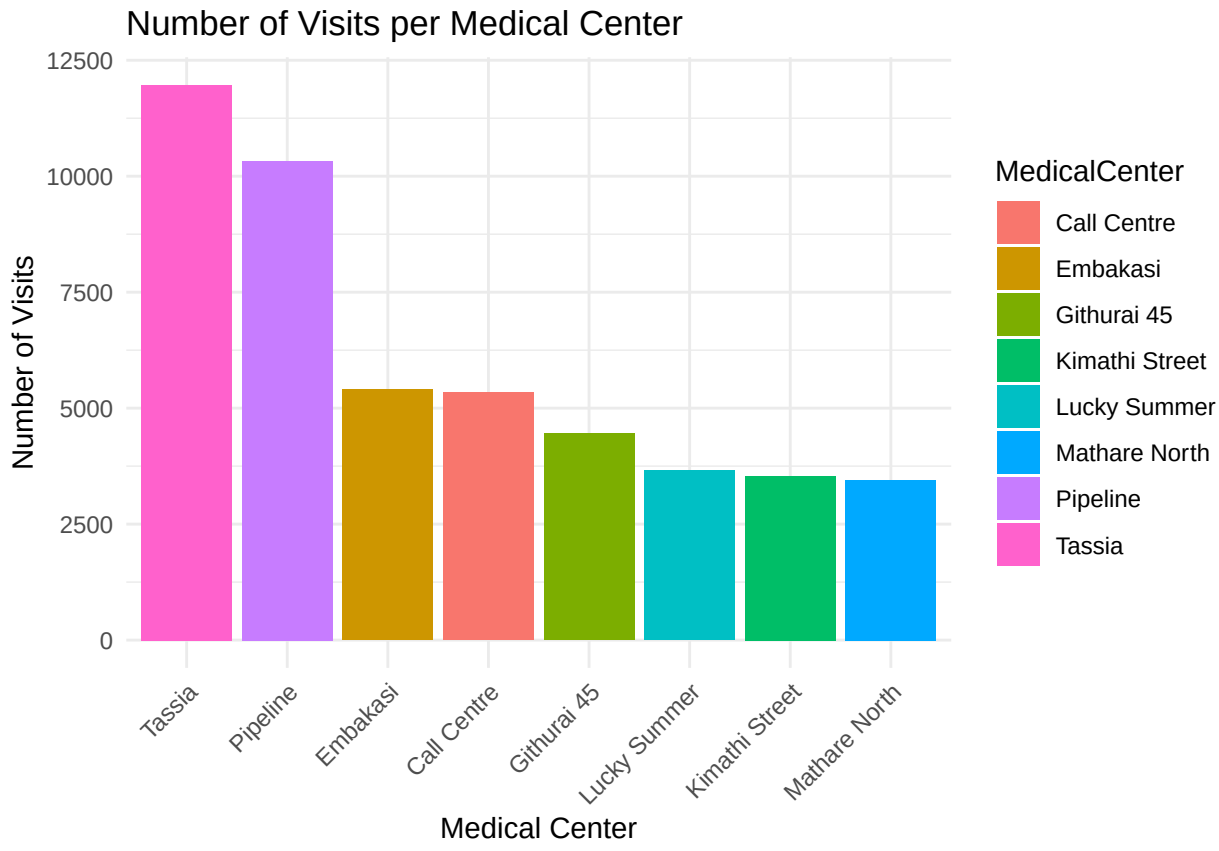
The chart revealed that the month of June recorded the highest number of visits, while the month of October had the lowest. This insight allows the clinic to better prepare for peak months and efficiently

manage patient flow.

Number of Visits per Medical Center

```
# Data preparation for visits per medical center
visits_per_center <- merged_data %>%
  group_by(MedicalCenter) %>%
  summarise(Visits = n()) %>%
  arrange(desc(Visits))

# Visualization: Bar Chart
ggplot(visits_per_center, aes(x = reorder(MedicalCenter, -Visits), y = Visits, fill = MedicalCenter)) +
  geom_bar(stat = "identity") +
  labs(title = "Number of Visits per Medical Center",
       x = "Medical Center",
       y = "Number of Visits") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

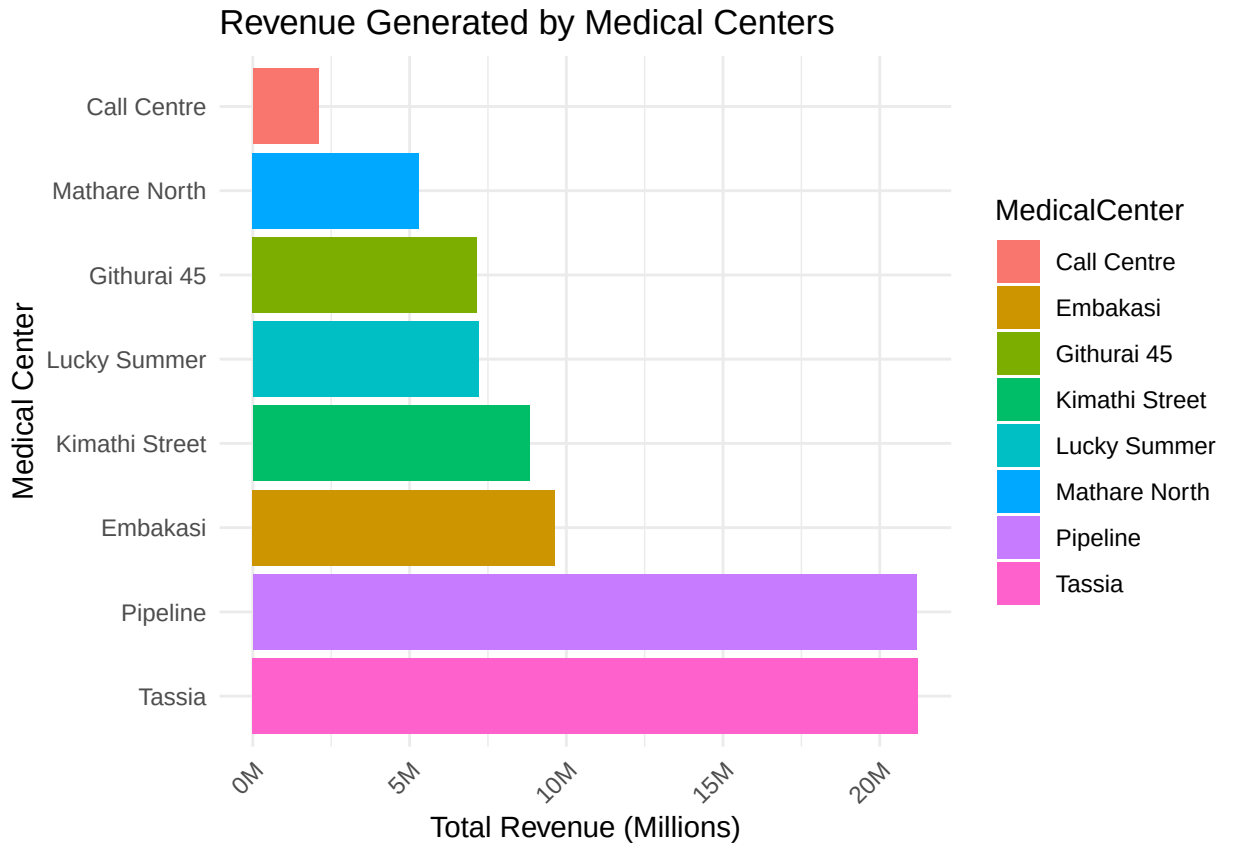


The bar chart comparing the number of visits per medical center highlighted "Tassia" as the most visited center, with "Mathare North" recording the least number of visits. This insight provides actionable information to optimize service distribution and improve underperforming centers.

Revenue Generated by Medical Centers

```r
# Data preparation for revenue per medical center
revenue_per_center <- merged_data %>%
  group_by(MedicalCenter) %>%
  summarise(TotalRevenue = sum(Amount)) %>%
  arrange(desc(TotalRevenue))

# Visualization: Stacked Bar Chart with revenue in millions
ggplot(revenue_per_center, aes(x = reorder(MedicalCenter, -TotalRevenue), y = TotalRevenue, fill = Medi
  geom_bar(stat = "identity") +
  labs(title = "Revenue Generated by Medical Centers",
       x = "Medical Center",
       y = "Total Revenue (Millions)") +
  scale_y_continuous(labels = scales::label_number(scale = 1e-6, suffix = "M")) +  # Format y-axis labe
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_flip()
```



The chart showcasing revenue generated by medical centers identified "Tassia" as the highest revenue-generating center, while "Call Centre" contributed the least. This insight guides revenue enhancement strategies, focusing on high-revenue centers to further boost overall revenue.

Diagnoses and Total Patients Recorded

```r
# Data preparation for diagnoses and total patients recorded
diagnoses_table <- merged_data %>%
  group_by(Diagnosis) %>%
```

```
  summarise(TotalPatients = n()) %>%
  arrange(desc(TotalPatients))

# Display the matrix table
pander::pander(diagnoses_table, caption = "Diagnoses and Total Patients Recorded")
```

Table 1: Diagnoses and Total Patients Recorded The table shows the total patients recorded for each diagnosis. "Acute Nasopharyngitis" emerged as the most common diagnosis with 4441 patients, allowing the clinic to prioritize and allocate resources for managing prevalent health conditions.

Diagnosis	TotalPatients
Unknown	26028
acute nasopharyngitis	4441
tonsillitis, acute bacterial	2864
acute gastritis	1951
review	1568
acute rhinitis	1345
acute bronchitis	1192
family planning	1189
pneumonia	1061
gastroenteritis with no dehydration	891
myalgia	821
normal pregnancy	775
soft tissue injury	567
dermatitis	469
amoebiasis	413
food poisoning	413
allergic rhinitis	373
migraine	308
acute pharyngitis, viral	268
acute otitis media	166
common cold	166
eczema	124
first degree burn	110
asthma, uncomplicated	106
bacterial conjunctivitis	103
cystitis	101
diabetes mellitus type 2	89
septic wound	79
severe pneumonia	65
severe malaria	38
hypertensive urgency	32
gingivitis	31

CONCLUSION

In conclusion, this data analysis report provides valuable insights into the performance of Penda Health Clinic. I have performed Exploratory Data Analysis (EDA) to understand the data distribution and then analyzed various key performance indicators, monthly visit trends, number of visits per medical center, revenue generated by medical centers, and diagnoses with the highest number of patients recorded.

These insights can help the clinic in making data-driven decisions for improved performance and patient care.

RECOMMENDATIONS

1. Resource Optimization: Allocate resources based on monthly visit trends to efficiently handle higher patient volumes during peak months like June. This will improve patient experience and reduce wait times.

2. Focus on High-Revenue Centers: Implement targeted marketing and service enhancement strategies for medical centers like "Tassia," which generate higher revenue. This will further maximize revenue generation for the clinic.

3. Enhanced Diagnoses-Specific Care: Given the prevalence of "Acute Nasopharyngitis," develop specialized care plans and resources to manage and treat this condition effectively. Proactive measures can improve patient outcomes and satisfaction.

4. Data-Driven Decision Making: Continue utilizing data analytics and interactive dashboards to drive informed decision-making in areas such as resource allocation, revenue enhancement, and patient care.