Healthcare Analysis

Oppy

2023-11-18

Background This report provides a comprehensive overview of key metrics that are vital for hospital management and policy-making. It helps in understanding the demographic profile of patients, the financial aspects of healthcare delivery, and the efficiency of hospital services, all of which are critical for informed decision-making by the hospital board.

Introduction Healthcare is an important part of human existence and would usually come at a cost directly or indirectly. Directly in terms of paying out of pocket or indirectly through a founded program. Also, health challenges vary from person to person due to varying factors. Some health challenges are generally attributed to certain age groups while same may arise due to other socioeconomic factors or lifestyle.

Objective

1. To explore trends in the given healthcare dataset

2. Use visualization techniques to understand and communicate patterns in the dataset

Data Source The data used in this analysis is a public dataset from Kaggle kaggle.. A copy of the dataset can be obtained here: Healthcare Dataset

Scope

*** Explore the data**

*** Descriptive statistics**

*** Use visualization tools to present findings**

*** Document each step in a communicable manner.**

Process Flow

A. Install r

B. Install Rstudio

C. Set work directory

D. Install and load tidyverse - to input dataset into r & data manipulation.

E. Telescope your dataset - to gain an initial overview of the data

F. Identify data types and convert as necessary

```
options(repos = c(CRAN = "https://cran.rstudio.com"))
install.packages("tidyverse")
```

___install package - tidyverse - used for data manipulation_

```
## Installing package into 'C:/Users/fadar/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)

## package 'tidyverse' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\fadar\AppData\Local\Temp\RtmpeWVgjU\downloaded_packages
```

```
library(tidyverse)
```

load tidayverse into the R session

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

___ knit to pdf_document _____

tinytex::install_tinytex()

update.packages(ask = FALSE, checkBuilt = TRUE)

tinytex::tlmgr_update()

```r
healthcare <- read.csv("healthcare_dataset.csv")
View(healthcare)
```

importing your dataset

```r
summary(healthcare)
```

to have a quick view of the dataset

```
##      Name                Age            Gender            Blood.Type
##  Length:10000       Min.   :18.00   Length:10000       Length:10000
##  Class :character   1st Qu.:35.00   Class :character   Class :character
##  Mode  :character   Median :52.00   Mode  :character   Mode  :character
##                     Mean   :51.45
##                     3rd Qu.:68.00
##                     Max.   :85.00
##  Medical.Condition  Date.of.Admission     Doctor            Hospital
##  Length:10000       Length:10000       Length:10000       Length:10000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Insurance.Provider Billing.Amount   Room.Number    Admission.Type
##  Length:10000       Min.   : 1000   Min.   :101.0   Length:10000
##  Class :character   1st Qu.:13507   1st Qu.:199.0   Class :character
##  Mode  :character   Median :25258   Median :299.0   Mode  :character
##                     Mean   :25517   Mean   :300.1
##                     3rd Qu.:37734   3rd Qu.:400.0
##                     Max.   :49996   Max.   :500.0
##  Discharge.Date      Medication        Test.Results
##  Length:10000       Length:10000       Length:10000
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```r
dim(healthcare) # return number of rows vs column
```

```
## [1] 10000    15
```

```r
str(healthcare)
```

```
## 'data.frame':    10000 obs. of  15 variables:
##  $ Name              : chr  "Tiffany Ramirez" "Ruben Burns" "Chad Byrd" "Antonio Frederick" ...
##  $ Age               : int  81 35 61 49 51 41 82 55 33 39 ...
```

```
##  $ Gender            : chr  "Female" "Male" "Male" "Male" ...
##  $ Blood.Type        : chr  "O-" "O+" "B-" "B-" ...
##  $ Medical.Condition : chr  "Diabetes" "Asthma" "Obesity" "Asthma" ...
##  $ Date.of.Admission : chr  "17/11/2022" "1/06/2023" "9/01/2019" "2/05/2020" ...
##  $ Doctor            : chr  "Patrick Parker" "Diane Jackson" "Paul Baker" "Brian Chandler" ...
##  $ Hospital          : chr  "Wallace-Hamilton" "Burke, Griffin and Cooper" "Walton LLC" "Garcia Ltd"
##  $ Insurance.Provider: chr  "Medicare" "UnitedHealthcare" "Medicare" "Medicare" ...
##  $ Billing.Amount    : num  37491 47304 36875 23303 18086 ...
##  $ Room.Number       : int  146 404 292 480 477 180 161 384 215 310 ...
##  $ Admission.Type    : chr  "Elective" "Emergency" "Emergency" "Urgent" ...
##  $ Discharge.Date    : chr  "1/12/2022" "15/06/2023" "8/02/2019" "3/05/2020" ...
##  $ Medication        : chr  "Aspirin" "Lipitor" "Lipitor" "Penicillin" ...
##  $ Test.Results      : chr  "Inconclusive" "Normal" "Normal" "Abnormal" ...
```

```
head(healthcare)
```

```
##                  Name Age Gender Blood.Type Medical.Condition Date.of.Admission
## 1     Tiffany Ramirez  81 Female         O-          Diabetes        17/11/2022
## 2         Ruben Burns  35   Male         O+            Asthma         1/06/2023
## 3           Chad Byrd  61   Male         B-           Obesity         9/01/2019
## 4   Antonio Frederick  49   Male         B-            Asthma         2/05/2020
## 5 Mrs. Brandy Flowers  51   Male         O-         Arthritis         9/07/2021
## 6      Patrick Parker  41   Male        AB+         Arthritis        20/08/2020
##           Doctor                  Hospital Insurance.Provider Billing.Amount
## 1 Patrick Parker          Wallace-Hamilton           Medicare       37490.98
## 2  Diane Jackson Burke, Griffin and Cooper   UnitedHealthcare       47304.06
## 3     Paul Baker                Walton LLC           Medicare       36874.90
## 4 Brian Chandler                Garcia Ltd           Medicare       23303.32
## 5 Dustin Griffin   Jones, Brown and Murray   UnitedHealthcare       18086.34
## 6    Robin Green                  Boyd PLC              Aetna       22522.36
##   Room.Number Admission.Type Discharge.Date  Medication Test.Results
## 1         146       Elective      1/12/2022     Aspirin Inconclusive
## 2         404      Emergency     15/06/2023     Lipitor       Normal
## 3         292      Emergency      8/02/2019     Lipitor       Normal
## 4         480         Urgent      3/05/2020  Penicillin     Abnormal
## 5         477         Urgent      2/08/2021 Paracetamol       Normal
## 6         180         Urgent     23/08/2020     Aspirin     Abnormal
```

```
tail(healthcare)
```

```
##                      Name Age Gender Blood.Type Medical.Condition
## 9995          Jorge Obrien  69   Male         A+          Diabetes
## 9996            James Hood  83   Male         A+           Obesity
## 9997       Stephanie Evans  47 Female        AB+         Arthritis
## 9998  Christopher Martinez  54   Male         B-         Arthritis
## 9999           Amanda Duke  84   Male         A+         Arthritis
## 10000            Eric King  20   Male         B-         Arthritis
##       Date.of.Admission            Doctor                  Hospital
## 9995         25/12/2021      Frank Miller                 Scott LLC
## 9996         29/07/2022      Samuel Moody  Wood, Martin and Simmons
## 9997          6/01/2022 Christopher Yates              Nash-Krueger
## 9998          1/07/2022  Robert Nicholson           Larson and Sons
## 9999          6/02/2020       Jamie Lewis              Wilson-Lyons
```

```
## 10000        22/03/2023       Tasha Avila Torres, Young and Stewart
##       Insurance.Provider Billing.Amount Room.Number Admission.Type
## 9995    UnitedHealthcare      16793.598         341       Elective
## 9996    UnitedHealthcare      39606.840         110       Elective
## 9997          Blue Cross       5995.717         244      Emergency
## 9998          Blue Cross      49559.203         312       Elective
## 9999    UnitedHealthcare      25236.345         420         Urgent
## 10000             Aetna      37223.966         290      Emergency
##       Discharge.Date Medication Test.Results
## 9995       6/01/2022 Penicillin Inconclusive
## 9996       2/08/2022  Ibuprofen     Abnormal
## 9997      29/01/2022  Ibuprofen       Normal
## 9998      15/07/2022  Ibuprofen       Normal
## 9999      26/02/2020 Penicillin       Normal
## 10000     15/04/2023 Penicillin     Abnormal
```

```
class(healthcare)
```

```
## [1] "data.frame"
```

```
glimpse(healthcare)
```

```
## Rows: 10,000
## Columns: 15
## $ Name               <chr> "Tiffany Ramirez", "Ruben Burns", "Chad Byrd", "Ant~
## $ Age                <int> 81, 35, 61, 49, 51, 41, 82, 55, 33, 39, 45, 23, 85,~
## $ Gender             <chr> "Female", "Male", "Male", "Male", "Male", "Male", "~
## $ Blood.Type         <chr> "O-", "O+", "B-", "B-", "O-", "AB+", "AB+", "O-", "~
## $ Medical.Condition  <chr> "Diabetes", "Asthma", "Obesity", "Asthma", "Arthrit~
## $ Date.of.Admission  <chr> "17/11/2022", "1/06/2023", "9/01/2019", "2/05/2020"~
## $ Doctor             <chr> "Patrick Parker", "Diane Jackson", "Paul Baker", "B~
## $ Hospital           <chr> "Wallace-Hamilton", "Burke, Griffin and Cooper", "W~
## $ Insurance.Provider <chr> "Medicare", "UnitedHealthcare", "Medicare", "Medica~
## $ Billing.Amount     <dbl> 37490.983, 47304.065, 36874.897, 23303.322, 18086.3~
## $ Room.Number        <int> 146, 404, 292, 480, 477, 180, 161, 384, 215, 310, 3~
## $ Admission.Type     <chr> "Elective", "Emergency", "Emergency", "Urgent", "Ur~
## $ Discharge.Date     <chr> "1/12/2022", "15/06/2023", "8/02/2019", "3/05/2020"~
## $ Medication         <chr> "Aspirin", "Lipitor", "Lipitor", "Penicillin", "Par~
## $ Test.Results       <chr> "Inconclusive", "Normal", "Normal", "Abnormal", "No~
```

```
names(healthcare)
```

```
##  [1] "Name"               "Age"                "Gender"
##  [4] "Blood.Type"         "Medical.Condition"  "Date.of.Admission"
##  [7] "Doctor"             "Hospital"           "Insurance.Provider"
## [10] "Billing.Amount"     "Room.Number"        "Admission.Type"
## [13] "Discharge.Date"     "Medication"         "Test.Results"
```

the dataset can be grouped into three types of data i. categorical

ii. date

iii. numeric

Categorical data an initial review of the data suggest the following are categorical data to be converted to factors Gender, Blood.Type, Admission.Type, Test results

```
healthcare$Gender <- as.factor(healthcare$Gender)
healthcare$Blood.Type <- as.factor(healthcare$Gender)
healthcare$Admission.Type <- as.factor(healthcare$Admission.Type)
healthcare$Test.Results <- as.factor(healthcare$Test.Results)
```

date date convert columns with date to date format

```
healthcare$Date.of.Admission <- as.Date(healthcare$Date.of.Admission, format="%d/%m/%Y")
healthcare$Discharge.Date <- as.Date(healthcare$Discharge.Date, format="%d/%m/%Y")
```

numeric data

```
summary(healthcare$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   35.00   52.00   51.45   68.00   85.00
```

Findings Basic descriptive statistics for the age variable. The minimum age is 18 years. 1st Qu. (First Quartile): 25% of the patients are aged 35 or younger.Median: the 50th percentile. Half of the patients are younger than 52 years, and half are older. Mean: The mean age in your dataset is approximately 51.45 years.3rd Qu. 75% of the patients are aged 68 or younger. Max The maximum age in your dataset is 85 years. These statistics give a general idea of the age distribution of the patients. The median and mean being close to each other suggests that the age distribution is fairly symmetrical. The range from the minimum to the maximum age (18 to 85 years) indicates the breadth of ages covered.

to check the skewness install the e1071 package

```
install.packages("e1071")
```

```
## Installing package into 'C:/Users/fadar/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

```
## package 'e1071' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\fadar\AppData\Local\Temp\RtmpeWVgjU\downloaded_packages
```

```
library(e1071)
```

to load the e1071 package

```r
age.skewness <- skewness(healthcare$Age)
print(age.skewness)
```

to check the skewness of age, Billing.Amount

```
## [1] -0.01214421
```

Findings

The skewness value for the Age variable is -0.01214421, which is very close to 0. This small negative value indicates a very slight skew to the left. However, given how close the value is to zero, it suggests that the age distribution of the patients is almost symmetrical. In practical terms, this means that the ages of the patients are fairly evenly distributed around the median, with no significant skew towards younger or older ages

```r
BillingAmount.skewness <- skewness(healthcare$Billing.Amount)
print(BillingAmount.skewness)
```

```
## [1] 0.01271741
```

Findings

BillingAmount, data does not have a pronounced long tail on either the right or left side and is approximately symmetric. This symmetry means that there aren't extreme values (outliers) that are significantly distorting the distribution in one direction.

check if there is a relationship between the age and admission period

check if there is a relationship between billing amount and admission period

```r
Doctors <- unique(healthcare$Doctor)
View(Doctors)
```

```r
Medications <- unique(healthcare$Medication)
View(Medications)
```

```r
Test.Results <- unique(healthcare$Test.Results)
View(Test.Results)
```

Unique Counts

```r
table(healthcare$Test.Results)
```

Frequency Analysis

```
##
##     Abnormal Inconclusive       Normal
##         3456         3277         3267
```

Findings_

The table indicates a relatively even distribution among the three categories of test results. This imply a diverse patient population with varying health conditions.

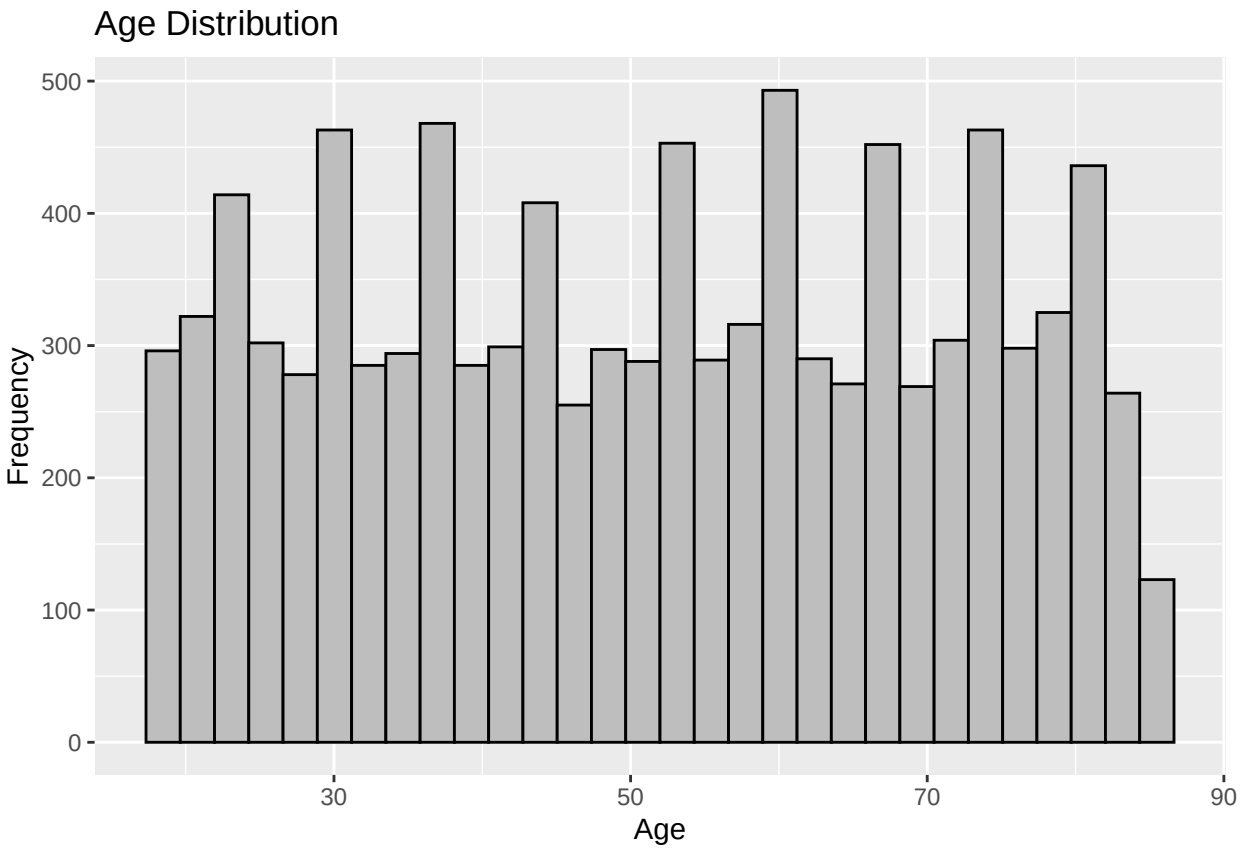
visualization to load the ggplot2

```r
library(ggplot2)
```

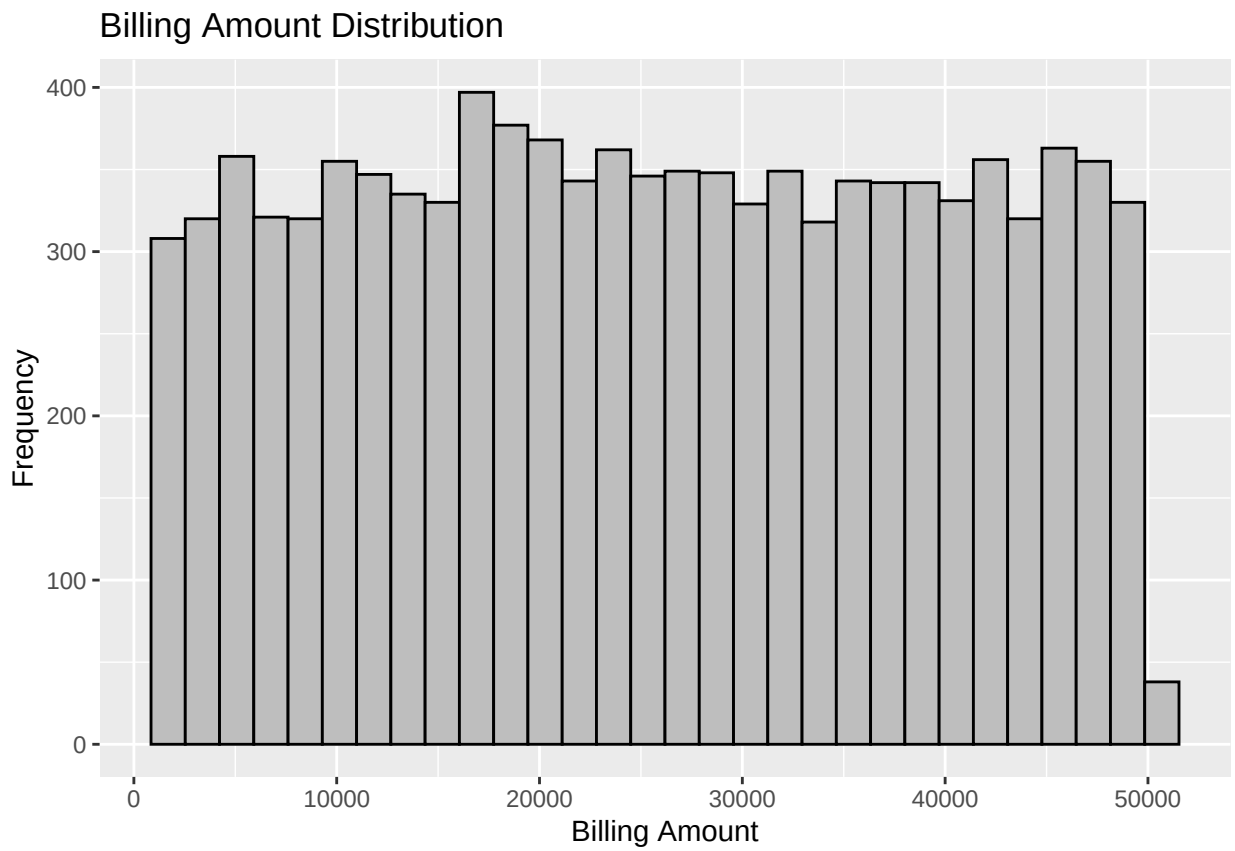
Histograms 1. Age

2. Billing.Amount

```r
ggplot(healthcare, aes(x = Age)) +
  geom_histogram(bins = 30, fill = "grey", color = "black") +
  ggtitle("Age Distribution") +
  xlab("Age") +
  ylab("Frequency")
```




```r
ggplot(healthcare, aes(x = Billing.Amount)) +
  geom_histogram(bins = 30, fill = "grey", color = "black") +
  ggtitle("Billing Amount Distribution") +
  xlab("Billing Amount") +
  ylab("Frequency")
```

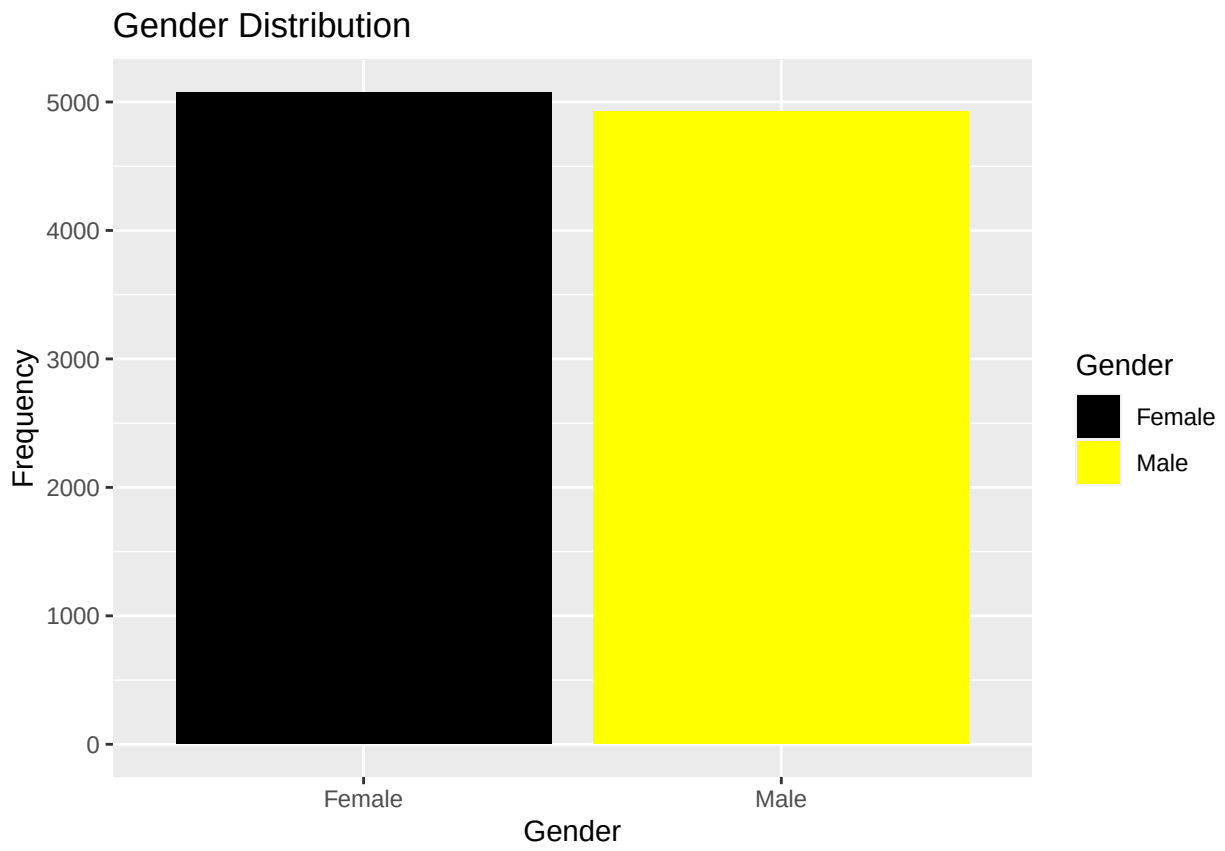


Bar chart 1. Gender

2. Blood.Type

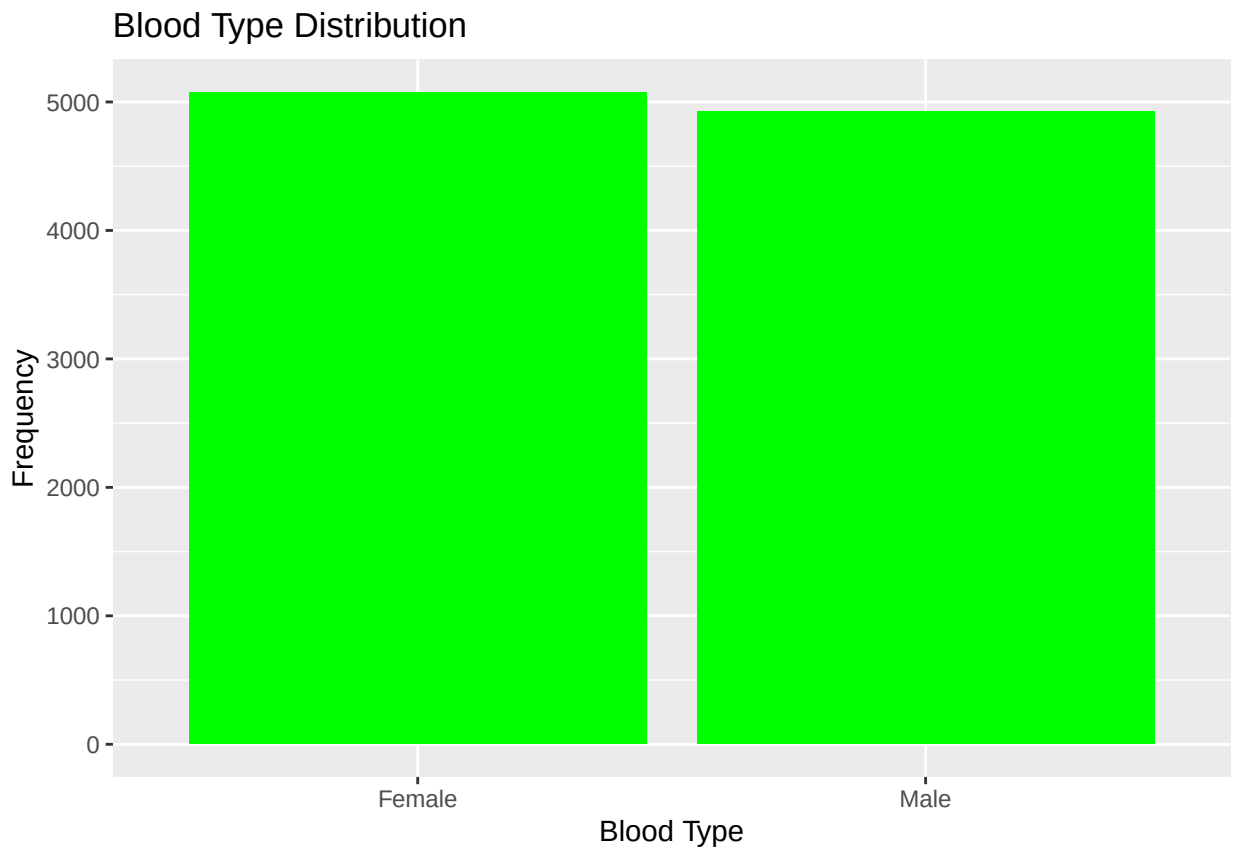
3. Medical.Condition

```
ggplot(healthcare, aes(x = Gender, fill = Gender)) +
  geom_bar() +
  scale_fill_manual(values = c("black", "yellow")) +
  ggtitle("Gender Distribution") +
  xlab("Gender") +
  ylab("Frequency")
```



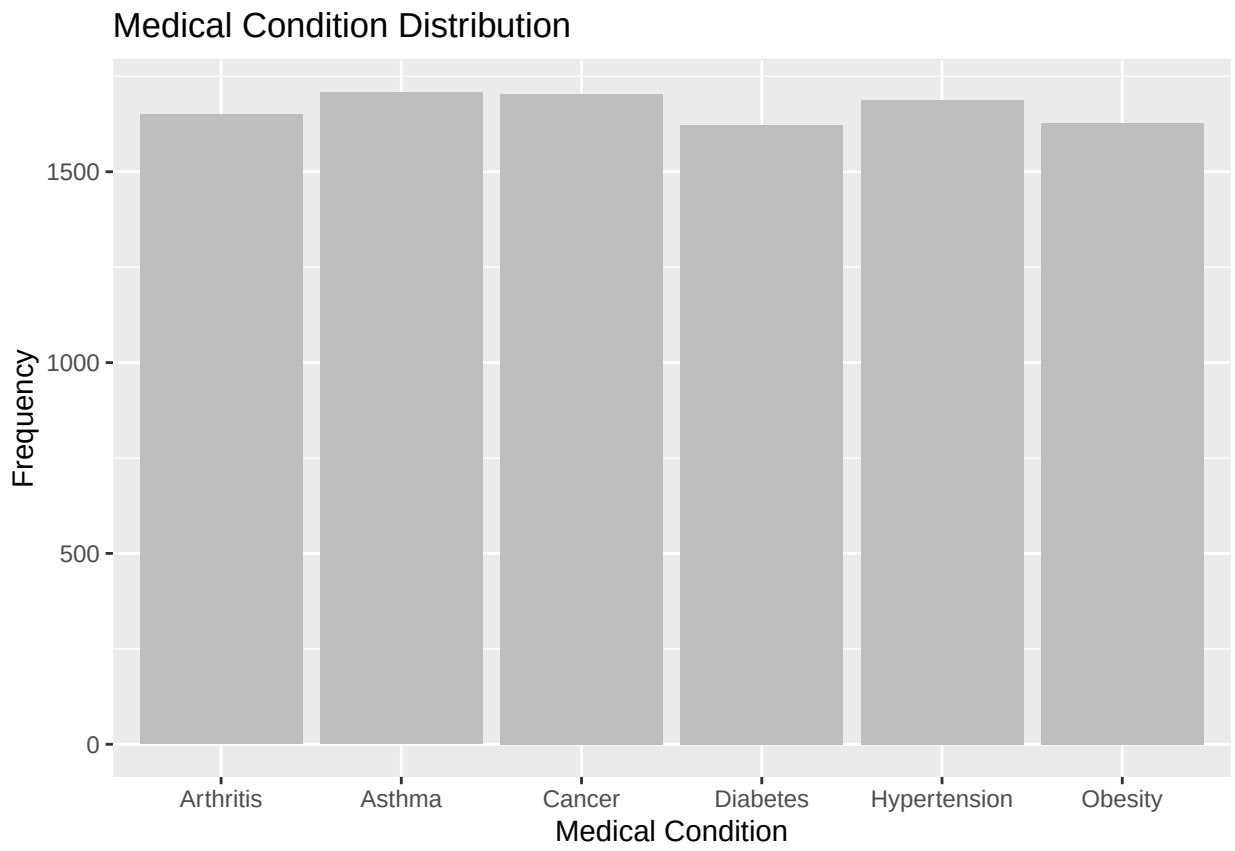
Blood.Type

```
ggplot(healthcare, aes(x = Blood.Type)) +
  geom_bar(fill = "green") +
  ggtitle("Blood Type Distribution") +
  xlab("Blood Type") +
  ylab("Frequency")
```



Medical.Condition

```
ggplot(healthcare, aes(x = Medical.Condition)) +
  geom_bar(fill = "grey") +
  ggtitle("Medical Condition Distribution") +
  xlab("Medical Condition") +
  ylab("Frequency")
```




```
healthcare$Length.of.Stay <- as.numeric(difftime(healthcare$Discharge.Date, healthcare$Date.of.Admissio
View(healthcare$Length.of.Stay)
head(healthcare$Length.of.Stay)
```

Calculate the length of stay for each patient

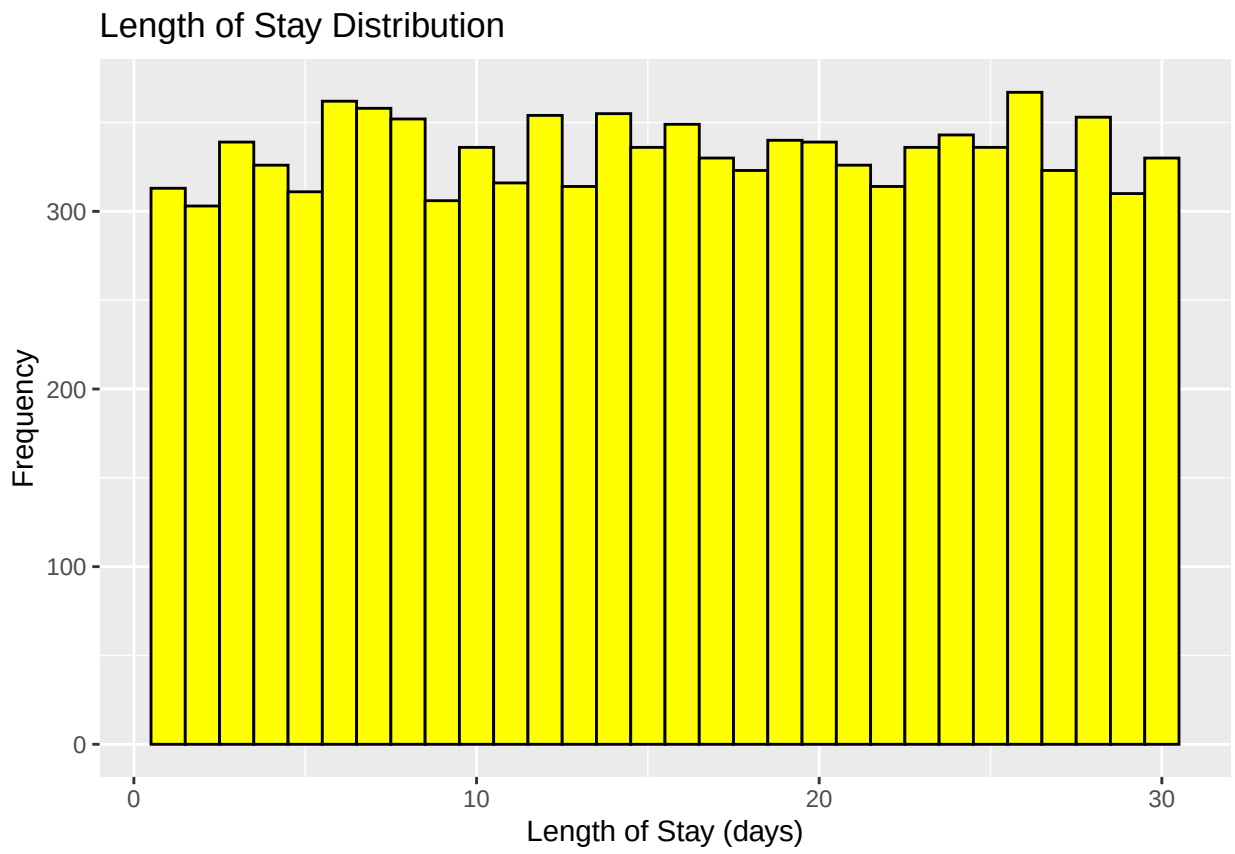
```
## [1] 14 14 30  1 24  3
```

```
summary(healthcare$Length.of.Stay)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    8.00   16.00   15.56   23.00   30.00
```

Histogram for 'Length of Stay'

```
ggplot(healthcare, aes(x = Length.of.Stay)) +
  geom_histogram(bins = 30, fill = "yellow", color = "black") +
  ggtitle("Length of Stay Distribution") +
  xlab("Length of Stay (days)") +
  ylab("Frequency")
```



Correlation Analysis - Age

- Billing Amount

- Length of Stay

```
cor(healthcare[c("Age", "Billing.Amount", "Length.of.Stay")], use = "complete.obs")
```

```
##                         Age Billing.Amount Length.of.Stay
## Age             1.000000000   -0.009483329    0.009111433
## Billing.Amount -0.009483329    1.000000000   -0.013506706
## Length.of.Stay  0.009111433   -0.013506706    1.000000000
```

Findings

Age and Billing.Amount: The correlation coefficient is approximately -0.0095, indicating a very weak negative correlation between age and billing amount. This suggests that there's almost no linear relationship between these two variables. Age and Length.of.Stay: The correlation coefficient is approximately 0.0091, suggesting a very weak positive correlation between age and length of stay. Again, this implies that there's almost no linear relationship. Billing.Amount and Length.of.Stay: The correlation coefficient is approximately -0.0135, indicating a very weak negative correlation between billing amount and length of stay. This suggests there's almost no linear relationship between these variables.

In summary, the output suggests that there are no strong linear relationships between age, billing amount, and length of stay among the patients. All these correlations are very close to zero, indicating that changes in one of these variables do not reliably predict changes in the others.

Grouped Analyses Mean Age by Gender

```
aggregate(Age ~ Gender, data = healthcare, mean)
```

```
##   Gender      Age
## 1 Female 51.60847
## 2   Male 51.29117
```

Findings

The average age of both male and female patients are the same. All patient's regardless of their gender fall within the same age group.

Mean Billing Amount by Medical Condition

```
aggregate(Billing.Amount ~ Medical.Condition, data = healthcare, mean)
```

```
##   Medical.Condition Billing.Amount
## 1         Arthritis       25187.63
## 2            Asthma       25416.87
## 3            Cancer       25539.10
## 4          Diabetes       26060.12
## 5      Hypertension       25198.03
## 6           Obesity       25720.84
```

Finding

For instance, the data suggests that on average, diabetes has the highest associated billing amount, while arthritis and hypertension have relatively lower average billing amounts.

Average Length of Stay by Admission Type

```
aggregate(Length.of.Stay ~ Admission.Type, data = healthcare, mean)
```

```
##   Admission.Type Length.of.Stay
## 1       Elective       15.60117
## 2      Emergency       15.60974
## 3         Urgent       15.47656
```

Finding

Elective: The average length of stay is approximately 15.60 days. Emergency: The average length of stay is approximately 15.61 days. Urgent: The average length of stay is approximately 15.48 days. These results suggest that the length of hospital stay is quite similar across the three types of admission, with very slight variations. In particular, there's only a marginal difference in the average length of stay between elective and emergency admissions.

Summary

Overview of Patient Demographics and Admission Patterns The analysis of the data set reveals insightful trends and characteristics about patient population and hospital admissions. Patient age range spans from 18 to 85 years, with a median and mean age close to 52 years, indicating a well-distributed age profile across young, middle-aged, and elderly patients. This diversity in age groups underscores the need for a wide range of medical services and specialized care to cater to the varied health needs of different age groups. Additionally, admissions data show a near-even split between genders, suggesting that healthcare services are accessed equally by both male and female patients. In terms of medical conditions, the average billing amounts vary slightly across different diagnoses such as Arthritis, Asthma, Cancer, Diabetes, Hypertension, and Obesity, but not significantly, pointing towards a consistent billing policy.

Analysis of Billing and Length of Stay The financial aspect of healthcare services was examined through the billing amounts and length of stay associated with various medical conditions and admission types. The skewness of the billing amount distribution is remarkably low (0.0127), indicating a symmetrical distribution with no significant outliers, which suggests a consistent and equitable billing practice. This is crucial for maintaining trust and transparency with patients and their families. Furthermore, the average length of stay for patients, regardless of the admission type (Elective, Emergency, or Urgent), hovers around 15.5 days, suggesting effective and efficient patient care and hospital resource utilization. This consistency in length of stay, irrespective of the urgency of admission, demonstrates hospital's commitment to providing timely and quality care to all patients.