# Project on Health Data

Hide

```
# Installing Required Packages

install.packages("tidyverse")

library(tidyverse)

library(dplyr)

library(ggplot2)
```

Hide

```
# Reading the data set and naming it as health_df

health_df <- read.csv("Health_Data.csv")
```

Hide

```
# Previewing the data frame using View() function

View(health_df)
```

Hide

```
# Checking the structure of the data frame and the type of variables

str(health_df)
```

```
'data.frame':   100 obs. of  10 variables:

 $ Patient_id          : int  202201001 202201002 202201003 202201004 2022010
05 202201006 202201007 202201008 202201009 202201010 ...

 $ Name                : chr  "Aniket" "Nayan" "Ritoprovo" "Swastika" ...

 $ Age                 : int  25 52 45 24 14 28 58 78 77 47 ...

 $ Gender              : chr  "Male" "Male" "Male" "Female" ...

 $ Disease             : chr  "Diabetics" "Thyroid" "Diabetics" "Thyroid" ...

 $ Duration._of_disease: int  1 2 3 4 5 4 2 3 1 5 ...

 $ Day_of_visit        : int  1 2 4 7 5 7 3 6 1 5 ...

 $ Family_history      : chr  "Yes" "No" "No" "No" ...

 $ Bill_amount         : int  12540 25123 4512 7845 2500 4512 500 879 4562 15
145 ...

 $ Hospital_received   : num  7524 15074 2707 4707 1500 ...
```

Hide

```
# Checking summary statistics of the data frame

summary(health_df)
```

```
   Patient_id             Name                Age             Gender
```

```
 Min.   :202201001   Length:100         Min.   :14.00   Length:100
 1st Qu.:202201026   Class :character   1st Qu.:25.00   Class :character
 Median :202201050   Mode  :character   Median :33.00   Mode  :character
 Mean   :202201050                      Mean   :42.05
 3rd Qu.:202201075                      3rd Qu.:55.25
 Max.   :202201100                      Max.   :91.00
   Disease          Duration._of_disease  Day_of_visit  Family_history
 Length:100         Min.   :0.00         Min.   :1.00   Length:100
 Class :character   1st Qu.:1.00         1st Qu.:3.00   Class :character
 Mode  :character   Median :2.00         Median :4.50   Mode  :character
                    Mean   :2.51         Mean   :4.25
                    3rd Qu.:4.00         3rd Qu.:6.00
                    Max.   :8.00         Max.   :7.00
  Bill_amount    Hospital_received
 Min.   :  100   Min.   :   60.0
 1st Qu.: 1246   1st Qu.:  747.9
 Median : 2658   Median : 1595.1
 Mean   : 4809   Mean   : 2885.6
 3rd Qu.: 7820   3rd Qu.: 4692.1
 Max.   :25123   Max.   :15073.8
```

Hide

```
# Getting column names of the data frame
colnames(health_df)
```

```
 [1] "Patient_id"           "Name"                 "Age"
 [4] "Gender"               "Disease"              "Duration._of_disease"
 [7] "Day_of_visit"         "Family_history"       "Bill_amount"
[10] "Hospital_received"
```

Hide

```
# Obtaining the row and column numbers of the data frame
nrow(health_df)
```

```
[1] 100
```

Hide

```
ncol(health_df)
```

```
[1] 10
```

Hide

```
# Creating one new column called Doc_received by subtracting Hospital bill fr
om Bill received

# Saving it in another data frame

health_01_df <- health_df %>% mutate(Doc_received=Bill_amount-Hospital_receiv
ed)

head(health_01_df)
```

| | Patient_id<br><int> | Name<br><chr> | Age<br><int> | Gender<br><chr> | Disease<br><chr> | Duration._of_disease<br><int> | Day_of_vi<br><ir |
|---|---|---|---|---|---|---|---|
| 1 | 202201001 | Aniket | 25 | Male | Diabetics | 1 | |
| 2 | 202201002 | Nayan | 52 | Male | Thyroid | 2 | |
| 3 | 202201003 | Ritoprovo | 45 | Male | Diabetics | 3 | |
| 4 | 202201004 | Swastika | 24 | Female | Thyroid | 4 | |
| 5 | 202201005 | Aishi | 14 | Female | Thyroid | 5 | |
| 6 | 202201006 | Anindita | 28 | Female | Diabetics | 4 | |

6 rows | 1-9 of 11 columns

Hide

```
NA
```

Hide

```
# Creating a feedback column using if-else condition

# Storing it in another variable

health_02_df <- health_01_df %>% mutate(Feedback=ifelse(Bill_amount > mean(Bi
ll_amount),'Bad','Good'))

head(health_02_df)
```

| | Patient_id<br><int> | Name<br><chr> | Age<br><int> | Gender<br><chr> | Disease<br><chr> | Duration._of_disease<br><int> | Day_of_visit<br><int> | Family_history<br><chr> |
|---|---|---|---|---|---|---|---|---|
| 1 | 202201001 | Aniket | 25 | Male | Diabetics | 1 | 1 | Yes |
| 2 | 202201002 | Nayan | 52 | Male | Thyroid | 2 | 2 | No |
| 3 | 202201003 | Ritoprovo | 45 | Male | Diabetics | 3 | 4 | No |
| 4 | 202201004 | Swastika | 24 | Female | Thyroid | 4 | 7 | No |
| 5 | 202201005 | Aishi | 14 | Female | Thyroid | 5 | 5 | Yes |
| 6 | 202201006 | Anindita | 28 | Female | Diabetics | 4 | 7 | Yes |

6 rows | 1-9 of 12 columns

Hide

```
# Getting the sum of total Good and Bad feedback

health_02_df %>% group_by(Feedback) %>% count()
```

| **Feedback**<br>&lt;chr&gt; |
|---|
| Bad |
| Good |

2 rows

Hide

```
# Selecting Age,Disease and Duration of disease column from the data frame

health_02_df %>% select(Age,Disease,Duration._of_disease)
```

| **Age**<br>&lt;int&gt; | **Disease**<br>&lt;chr&gt; | **Duration._of_disease**<br>&lt;int&gt; |
|---:|---|---:|
| 25 | Diabetics | 1 |
| 52 | Thyroid | 2 |
| 45 | Diabetics | 3 |
| 24 | Thyroid | 4 |
| 14 | Thyroid | 5 |
| 28 | Diabetics | 4 |
| 58 | Thyroid | 2 |
| 78 | Diabetics | 3 |
| 77 | Thyroid | 1 |
| 47 | Diabetics | 5 |

Next
123456
...
10
Previous
1-10 of 100 rows

Hide

```
# Renaming a column into a new column

# Saving and storing the data in a new data frame

health_final_df <- health_02_df %>% rename(Duration_of_disease = Duration._of_disease)

head(health_final_df)
```

| | Patient_id | Name | Age | Gender | Disease | Duration_of_disease | Day_of_visit | Family_history |
|---|---|---|---|---|---|---|---|---|
| | <int> | <chr> | <int> | <chr> | <chr> | <int> | <int> | <chr> |
| 1 | 202201001 | Aniket | 25 | Male | Diabetics | 1 | 1 | Yes |
| 2 | 202201002 | Nayan | 52 | Male | Thyroid | 2 | 2 | No |
| 3 | 202201003 | Ritoprovo | 45 | Male | Diabetics | 3 | 4 | No |
| 4 | 202201004 | Swastika | 24 | Female | Thyroid | 4 | 7 | No |
| 5 | 202201005 | Aishi | 14 | Female | Thyroid | 5 | 5 | Yes |
| 6 | 202201006 | Anindita | 28 | Female | Diabetics | 4 | 7 | Yes |

6 rows | 1-9 of 12 columns

Hide

```
# Finnaly getting the colnames of the data frame

colnames(health_final_df)
```

```
 [1] "Patient_id"         "Name"               "Age"
 [4] "Gender"             "Disease"            "Duration_of_disease"
 [7] "Day_of_visit"       "Family_history"     "Bill_amount"
[10] "Hospital_received"  "Doc_received"       "Feedback"
```

Hide

```
# Filtering the latest data frame on Age< 30 and duraion of disease is less t
han 2 years

health_final_df %>% filter(Age < 30 & Duration_of_disease < 2)
```

| Patient_id | Name | Age | Gender | Disease | Duration_of_disease | Day_of_visit | Family_history |
|---|---|---|---|---|---|---|---|
| <int> | <chr> | <int> | <chr> | <chr> | <int> | <int> | <chr> |
| 202201001 | Aniket | 25 | Male | Diabetics | 1 | 1 | Yes |
| 202201035 | Kiyara | 28 | Female | Thyroid | 0 | 5 | Yes |
| 202201037 | Poran | 20 | Male | Diabetics | 1 | 3 | No |
| 202201046 | Iliana | 28 | Female | Thyroid | 0 | 7 | Yes |
| 202201057 | Dipti | 28 | Female | Diabetics | 1 | 5 | No |
| 202201066 | Piyali | 23 | Female | Diabetics | 1 | 4 | No |
| 202201076 | Suvendu | 28 | Male | Diabetics | 0 | 6 | Yes |
| 202201080 | Monalisa | 21 | Female | Diabetics | 0 | 5 | Yes |
| 202201081 | Amit | 20 | Male | Diabetics | 0 | 6 | Yes |
| 202201085 | Hrittika | 24 | Female | Diabetics | 0 | 5 | No |

Next
12
Previous
1-10 of 12 rows | 1-8 of 12 columns

Hide

```
# Filtering the data frame on Diabetics disease and female gender

health_final_df %>% filter(Disease == 'Diabetics' & Gender == 'Female')
```

| Patient_id | Name | Age | Gender | Disease | Duration_of_disease | Day_of_visit | Family_history |
|---|---|---|---|---|---|---|---|
| <int> | <chr> | <int> | <chr> | <chr> | <int> | <int> | <chr> |
| 202201006 | Anindita | 28 | Female | Diabetics | 4 | 7 | Yes |
| 202201010 | Bidisha | 47 | Female | Diabetics | 5 | 5 | No |
| 202201016 | Gargi | 25 | Female | Diabetics | 2 | 5 | Yes |
| 202201018 | Shreya | 49 | Female | Diabetics | 4 | 2 | No |
| 202201022 | Anjali | 20 | Female | Diabetics | 2 | 4 | No |
| 202201029 | Nandita | 29 | Female | Diabetics | 4 | 5 | Yes |
| 202201034 | Alia | 31 | Female | Diabetics | 0 | 3 | No |
| 202201041 | Sinjini | 23 | Female | Diabetics | 4 | 5 | No |
| 202201053 | Trisha | 91 | Female | Diabetics | 4 | 5 | Yes |
| 202201057 | Dipti | 28 | Female | Diabetics | 1 | 5 | No |

Next
12
Previous
1-10 of 19 rows | 1-8 of 12 columns

Hide

```
# Finding some Mathematical values for Bill_amount using group_by() and summa
rize()

health_final_df %>% group_by(Gender) %>% summarize(avg_bill=mean(Bill_amount)
,min_bill=min(Bill_amount),max_bill=max(Bill_amount))
```

| Gender | avg_bill | min_bill | max_bill |
|---|---|---|---|
| <chr> | <dbl> | <int> | <int> |
| Female | 5013.519 | 789 | 23154 |
| Male | 4588.000 | 100 | 25123 |

2 rows

Hide

```
# Finding some Mathematical values for Hospital_received using group_by() and
summarize()
```

```
health_final_df %>% group_by(Gender) %>% summarize(avg_hospital=mean(Hospital
_received),min_hospital=min(Hospital_received),max_hospital=max(Hospital_rece
ived))
```

| Gender | avg_hospital | min_hospital | max_hospital |
|---|---:|---:|---:|
| <chr> | <dbl> | <dbl> | <dbl> |
| Female | 3008.112 | 473.4 | 13892.4 |
| Male | 2752.800 | 60.0 | 15073.8 |

2 rows

Hide

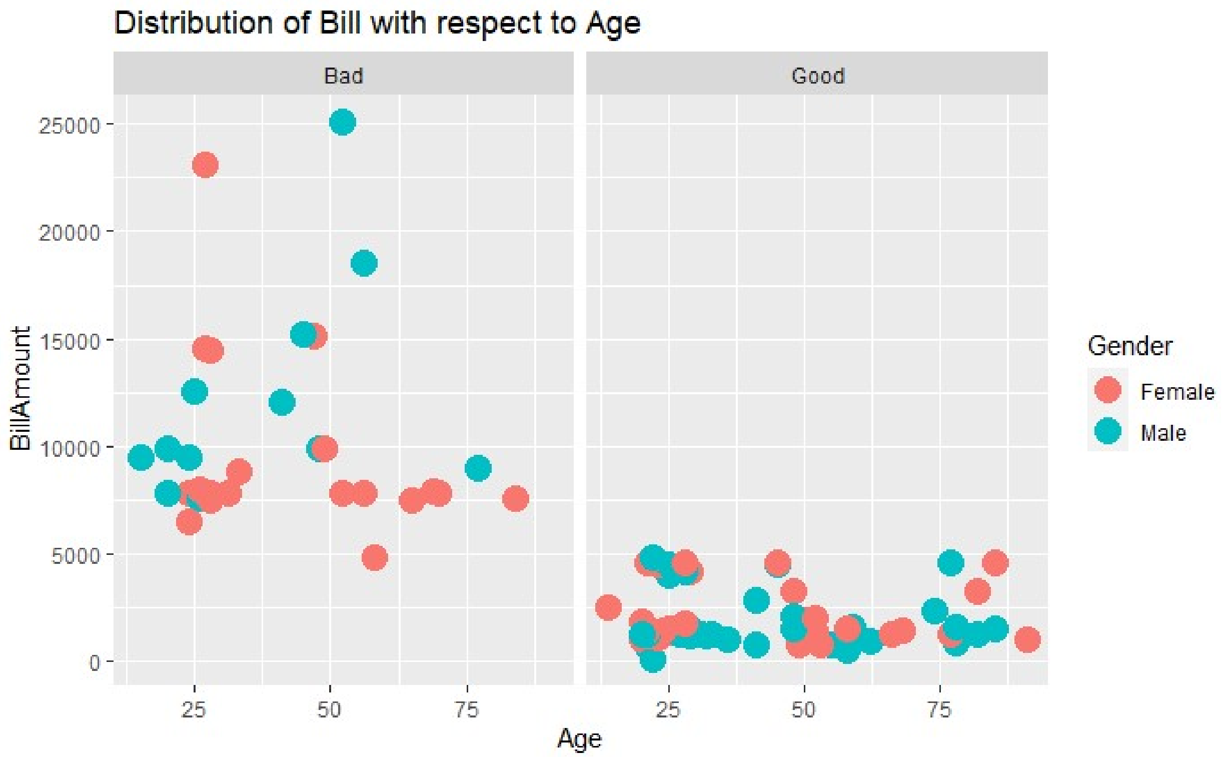
```
# Finding some Mathematical values for Doc_received using group_by() and summ
arize()

health_final_df%>%group_by(Gender)%>%summarize(avg_doc=mean(Doc_received),min
_doc=min(Doc_received),max_doc=max(Doc_received))
```

| Gender | avg_doc | min_doc | max_doc |
|---|---:|---:|---:|
| <chr> | <dbl> | <dbl> | <dbl> |
| Female | 2005.408 | 315.6 | 9261.6 |
| Male | 1835.200 | 40.0 | 10049.2 |

2 rows

Hide

```
# Plotting Age vs Bill_amount for Good and Bad feedback

p1=ggplot(data= health_final_df)+ geom_point(mapping=aes(x= Age,y=Bill_amount
,color= Gender),size=5)+labs(title="Distribution of Bill with respect to Age"
,x="Age",y="BillAmount")+facet_wrap(~Feedback)

p1
```



Hide

```
# Selecting Age and Disease where Age <30 and disease is diabetics

health_sp <- health_final_df %>% filter(Age<30 & Disease =='Diabetics') %>% s
elect(Age,Disease)

View(health_sp)
```

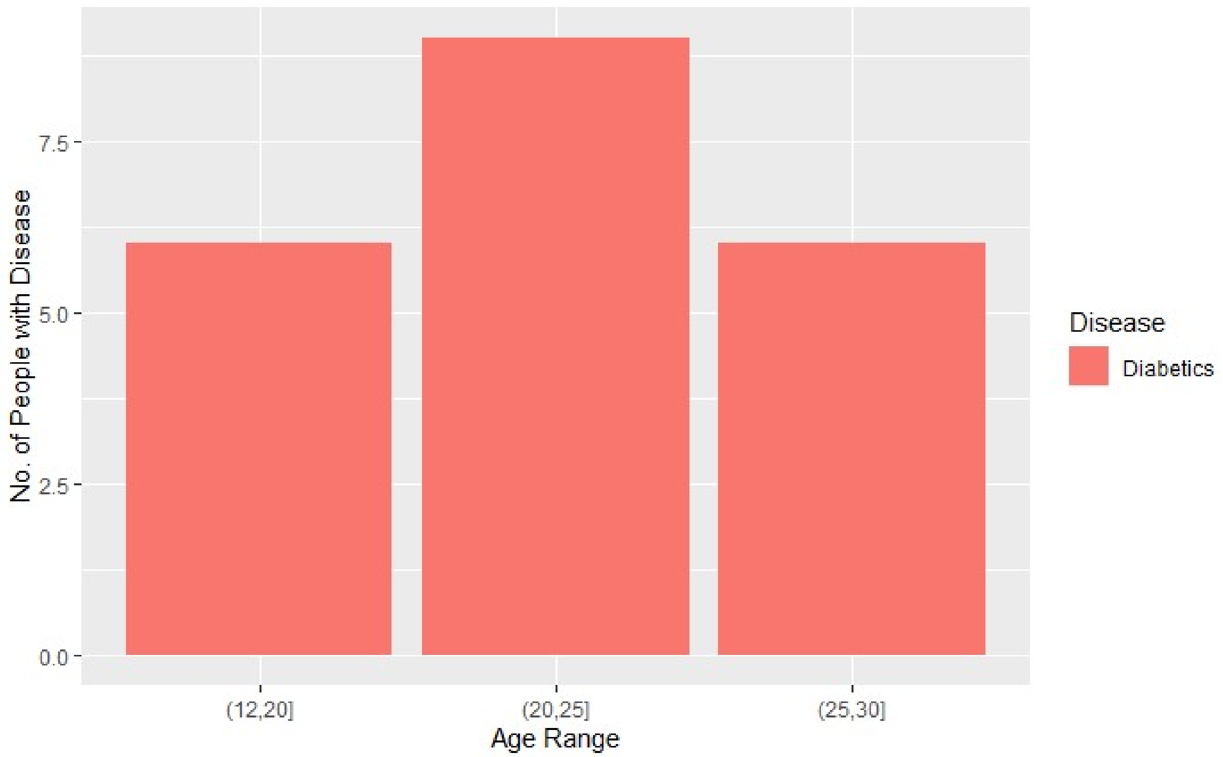
Hide

```
# Constructing an age range and plotting a bar graph

age_range <- cut(health_sp$Age, breaks=c(0, 6, 12, 20, 25, 30))

age_range
```

```
 [1] (20,25] (25,30] (20,25] (20,25] (12,20] (25,30] (25,30] (12,20] (12,20]
(20,25]

[11] (25,30] (12,20] (20,25] (20,25] (25,30] (20,25] (12,20] (20,25] (25,30]
(20,25]

[21] (12,20]

Levels: (0,6] (6,12] (12,20] (20,25] (25,30]
```

Hide

```
p2=ggplot(data=health_sp)+geom_bar(mapping=aes(x=age_range,fill=Disease))+ la
bs(x="Age Range", y="No. of People with Disease")

p2
```



Hide

```
# Counting the number of each type of disease that are in the population
health_final_df %>% group_by(Disease) %>% count()
```

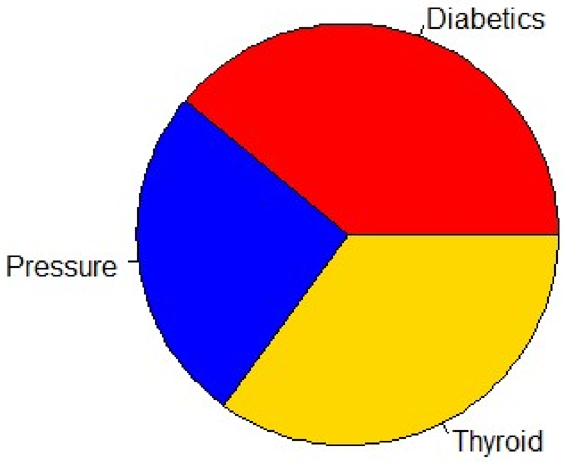
| **Disease** |
|---|
| <chr> |
| Diabetics |
| Pressure |
| Thyroid |

3 rows

Hide

```
# Plotting a pie chart to know the contribution of each disease in the popula
tion
values <- c(39,26,35)
labels <- c('Diabetics','Pressure','Thyroid')
radius <- 1
colors <- c('red','blue','gold')
main <- 'Distribution of Disease in the popiulation'
percentages <- round(values/sum(values)*100,2)
p3=pie(values,labels,radius=radius,main=main,col=colors)
```

**Distribution of Disease in the popiulation**


Diabetics
Pressure
Thyroid