Assignment No. 1

Aim: Getting data to work with: Download the sample dataset locally for any application (Kaggle)

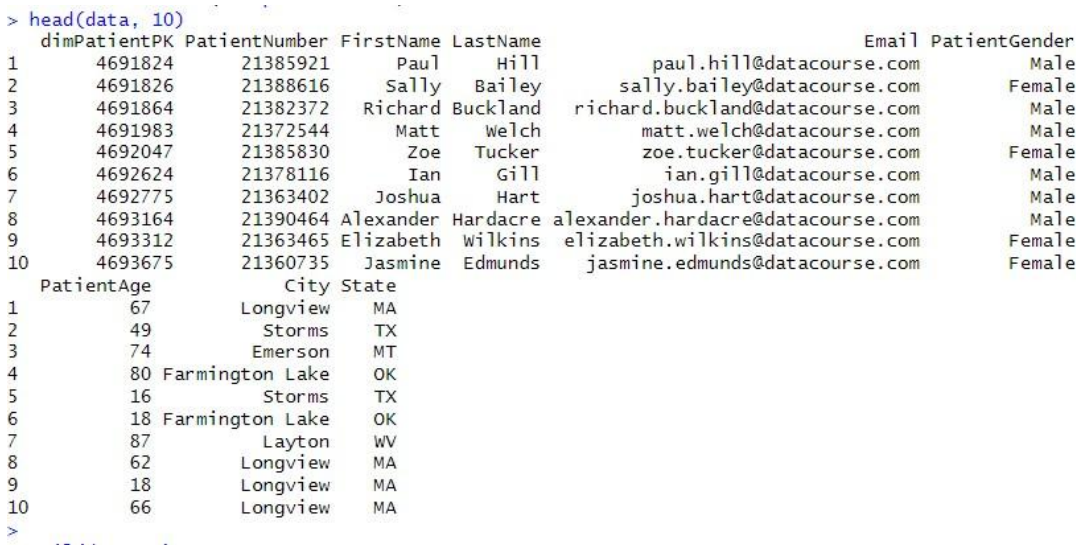
- Setting up the working directory.
- Unpacking the data. Decompress the file locally.
- Looking at the data. Display the top (10) and bottom (10) of the file.
- Measuring the length of the data set. Count the number of lines in the file.
- Encode the categorical data
- Plot a graph and give your insights for the application selected cases.

Theory:

In this practical, we explore a healthcare dataset in R, demonstrating essential data handling and visualization techniques. First, the working directory is set using setwd() and verified with getwd(). The dataset is then decompressed using unzip() and loaded into R with read.csv(). Initial data inspection is performed using head() and tail() to view the top and bottom 10 rows, helping understand the dataset's structure. The dataset size is determined with nrow(), and categorical variables such as PatientGender, City, and State are encoded as factors using as.factor() to enable statistical and visual analysis. Summary statistics are generated with summary() to identify key data characteristics like mean, median, and frequency of categorical variables. Finally, a bar plot is created using ggplot2 to visualize the distribution of patients by gender, providing insights into gender representation in the dataset. This process illustrates the importance of data exploration and visualization in identifying patterns and guiding further analysis.

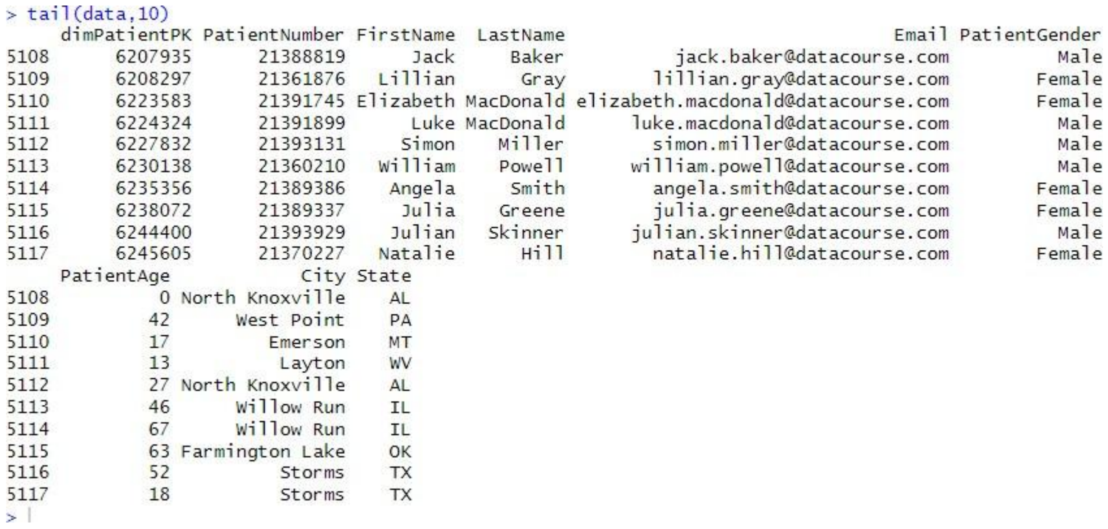
Code:

```
setwd("D:\\MIT ADT\\LY - Sem 1\\BDA Lab\\Amreen Mam\\Assign 1")

getwd()

unzip("archive.zip", exdir = "D:\\MIT ADT\\LY - Sem 1\\BDA Lab\\Amreen Mam\\Assign 1")

data <- read.csv("Dimpatient.csv")

head(data, 10)
```



```
> head(data, 10)
   dimPatientPK PatientNumber FirstName LastName                          Email PatientGender
1       4691824      21385921      Paul     Hill         paul.hill@datacourse.com          Male
2       4691826      21388616     Sally   Bailey       sally.bailey@datacourse.com        Female
3       4691864      21382372   Richard Buckland   richard.buckland@datacourse.com          Male
4       4691983      21372544      Matt    Welch         matt.welch@datacourse.com          Male
5       4692047      21385830       Zoe   Tucker         zoe.tucker@datacourse.com        Female
6       4692624      21378116       Ian     Gill           ian.gill@datacourse.com          Male
7       4692775      21363402    Joshua     Hart        joshua.hart@datacourse.com          Male
8       4693164      21390464 Alexander Hardacre alexander.hardacre@datacourse.com          Male
9       4693312      21363465 Elizabeth  Wilkins   elizabeth.wilkins@datacourse.com        Female
10      4693675      21360735   Jasmine  Edmunds     jasmine.edmunds@datacourse.com        Female
   PatientAge            City State
1          67        Longview    MA
2          49          Storms    TX
3          74         Emerson    MT
4          80 Farmington Lake    OK
5          16          Storms    TX
6          18 Farmington Lake    OK
7          87          Layton    WV
8          62        Longview    MA
9          18        Longview    MA
10         66        Longview    MA
>
```

tail(data,10)



```
> tail(data,10)
     dimPatientPK PatientNumber FirstName  LastName                            Email PatientGender
5108      6207935      21388819      Jack     Baker           jack.baker@datacourse.com          Male
5109      6208297      21361876   Lillian      Gray         lillian.gray@datacourse.com        Female
5110      6223583      21391745 Elizabeth MacDonald elizabeth.macdonald@datacourse.com        Female
5111      6224324      21391899      Luke MacDonald      luke.macdonald@datacourse.com          Male
5112      6227832      21393131     Simon    Miller        simon.miller@datacourse.com          Male
5113      6230138      21360210   William    Powell      william.powell@datacourse.com          Male
5114      6235356      21389386    Angela     Smith         angela.smith@datacourse.com        Female
5115      6238072      21389337     Julia    Greene         julia.greene@datacourse.com        Female
5116      6244400      21393929    Julian   Skinner       julian.skinner@datacourse.com          Male
5117      6245605      21370227   Natalie      Hill         natalie.hill@datacourse.com        Female
     PatientAge            City State
5108          0 North Knoxville    AL
5109         42      West Point    PA
5110         17         Emerson    MT
5111         13          Layton    WV
5112         27 North Knoxville    AL
5113         46      Willow Run    IL
5114         67      Willow Run    IL
5115         63 Farmington Lake    OK
5116         52          Storms    TX
5117         18          Storms    TX
> |
```

num_lines <- nrow(data)

cat("Number of lines in the dataset:", num_lines, "\n")

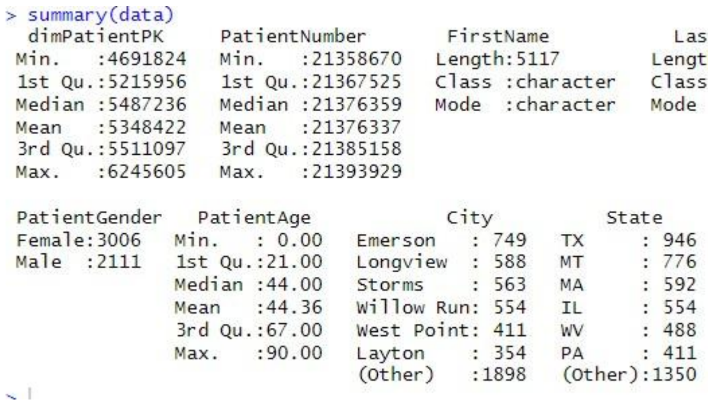
Number of lines in the dataset: 5117

data$PatientGender <- as.factor(data$PatientGender)

data$City <- as.factor(data$City)

data$State <- as.factor(data$State)

summary(data)



```
> summary(data)
  dimPatientPK      PatientNumber        FirstName           LastName            Email          
 Min.   :4691824   Min.   :21358670   Length:5117        Length:5117        Length:5117       
 1st Qu.:5215956   1st Qu.:21367525   Class :character   Class :character   Class :character  
 Median :5487236   Median :21376359   Mode  :character   Mode  :character   Mode  :character  
 Mean   :5348422   Mean   :21376337                                                           
 3rd Qu.:5511097   3rd Qu.:21385158                                                           
 Max.   :6245605   Max.   :21393929                                                           
                                                                                              
 PatientGender   PatientAge             City           State     
 Female:3006   Min.   : 0.00   Emerson   : 749   TX     : 946  
 Male  :2111   1st Qu.:21.00   Longview  : 588   MT     : 776  
               Median :44.00   Storms    : 563   MA     : 592  
               Mean   :44.36   Willow Run: 554   IL     : 554  
               3rd Qu.:67.00   West Point: 411   WV     : 488  
               Max.   :90.00   Layton    : 354   PA     : 411  
                               (Other)   :1898   (Other):1350  
```

library(ggplot2)

# Bar plot for PatientGender

gender_plot <- ggplot(data, aes(x = PatientGender)) +

  geom_bar(fill = "lightgreen", color = "black") +

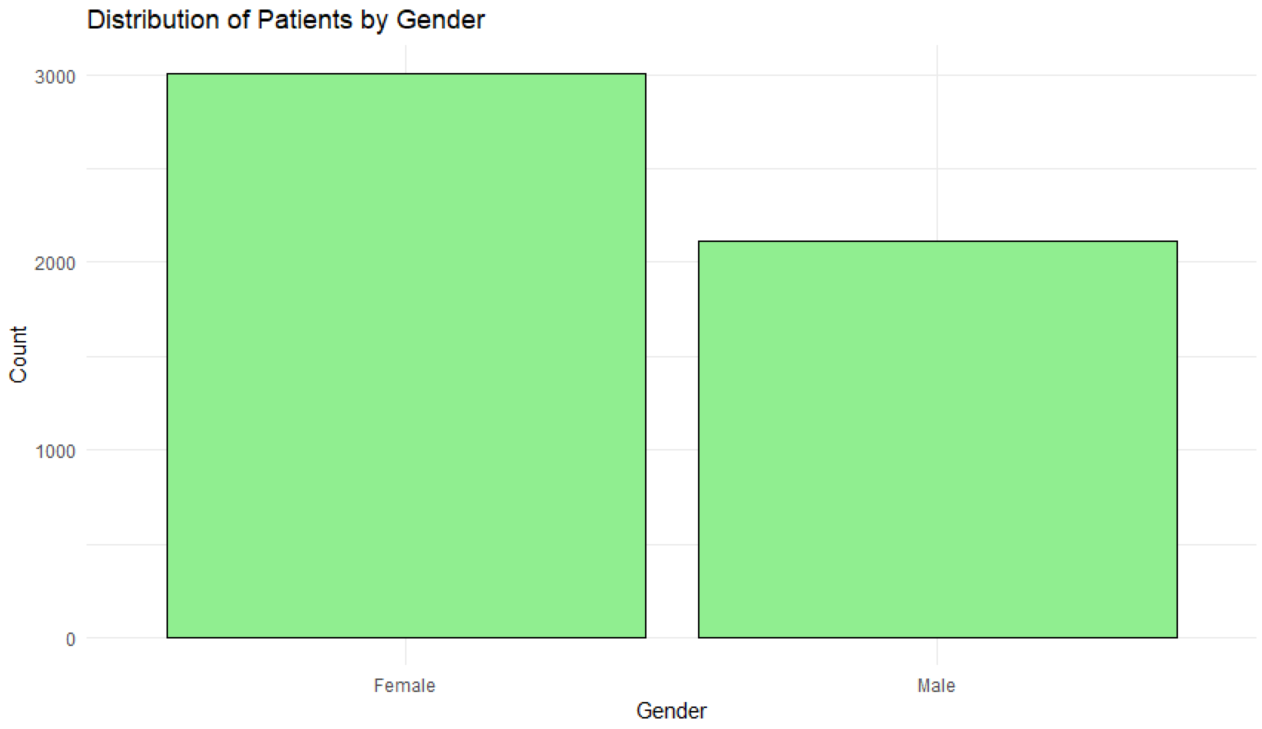
  labs(title = "Distribution of Patients by Gender", x = "Gender", y = "Count") +

  theme_minimal()

# Display the plot

print(gender_plot)

Output:


Distribution of Patients by Gender
Count
Female
Male
Gender