IN HOUSE PROJECT

Report On

Data visualization and prediction of Cancer using ggplot2 and Linear Model Building

Project work submitted to

PARVATHANENI BRAHMAYYA SIDDHARTHA COLLEGE OF ARTS&SCIENCE (AUTONOMOUS)

for the partial fulfillment of the requirements for the award of Degree of Bachelor Science

Done By
KOLLI.SOWMYA SRI
(223411P)

Under the guidance of

Ms. K. DIVYA ,M.Sc
DEPARTMENT OF STATISTICS

MAY- 2024

PARVATHANENI BRAHMAYYA SIDDHARTHA COLLEGE OF ARTS&SCIENCE (AUTONOMOUS)

Vijayawada – 520010

DEPARTMENT OF STATISTICS

CERTIFICATE

This is certified that this is a bonafide record of **INHOUSE PROJECT** work done in DEPARTMENT OF STATISTICS entitled with **Data visualization and prediction of Cancer using ggplot2 and Linear Model Building** done by **KOLLI.SOWMYA SRI** for the partial fulfillment of the requirement for the award of the Bachelor of Science and Arts degree as part of the curriculum during the academic year 2023 - 2024.

Date:

Lecturer in charge **Head of the Department**

Examiner

DECLARATION

I hereby declare that the In house project report entitled **Data visualization and prediction of Cancer using ggplot2 and Linear Model Building** submitted by us in the partial fulfillment of the requirement for the award of degree in Bachelor of Science is the record of work originally carried out by from April - may 2024 under able guidance of **Ms.K.DIVYA** Department of Statistics, P. B. Siddhartha College of Arts & Science, Vijayawada.

KOLLI.SOWMYA SRI(223411P)

2ND B.Sc. MSCS-A

INDEX

CHAPTER-1: INTRODUCTION TO STATISTICS, R PROGRAMMING LANGUAGE AND DATA DESCRIPTION

1.1 Introduction to Statistics

Statistics is the science concerned with developing and studying methods for collecting, analyzing, interpreting and presenting empirical data. The science of statistics is essentially a branch of Applied Mathematics, and may be regarded as mathematics applied to observational data.

Types of Statistics Basically,

There are two types of statistics.

Descriptive Statistics

Inferential Statistics

Descriptive Statistics

The data is summarized and explained in descriptive statistics. The summarization is done from a population sample utilizing several factors such as mean and standard deviation. Descriptive statistics is a way of organizing, representing, and explaining a set of data using charts, graphs, and summary measures. Histograms, pie charts, bars, and scatter plots are common ways to summarize data and present it in tables or graphs. Descriptive statistics are just that: descriptive. They don't need to be normalized beyond the data they collect.

Inferential Statistics

We attempt to interpret the meaning of descriptive statistics using inferential statistics. We utilize inferential statistics to convey the meaning of the collected data after it has been collected, evaluated, and summarized. The probability principle is used in inferential statistics to determine if patterns found in a study sample may be extrapolated to the wider population from which the sample was drawn. Inferential statistics are used to test hypotheses and study correlations between variables, and they can also be used to predict population sizes. Inferential statistics are used to derive conclusions and inferences from samples,

i.e., to create accurate generalizations.

1.2 Applications of Statistics

Here we are going to explore the Application of Statistics and its usage.
Actuarial science is the discipline that applies mathematical and statistical methods to assess risk in the insurance and finance industries.

Biostatistics is a branch of biology that studies biological phenomena and observations by means of statistical analysis, and includes medical statistics.

Business analytics is a rapidly developing business process that applies statistical methods to data sets (often very large) to develop new insights and understanding of business performance & opportunities.

Chemometrics is the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods.

Demography is the statistical study of all populations. It can be a very general science that can be applied to any kind of dynamic population, that is, one that changes over time or space.

Environmental statistics is the application of statistical methods to environmental science. Weather, climate, air and water quality are included, as are studies of plant and animal populations.

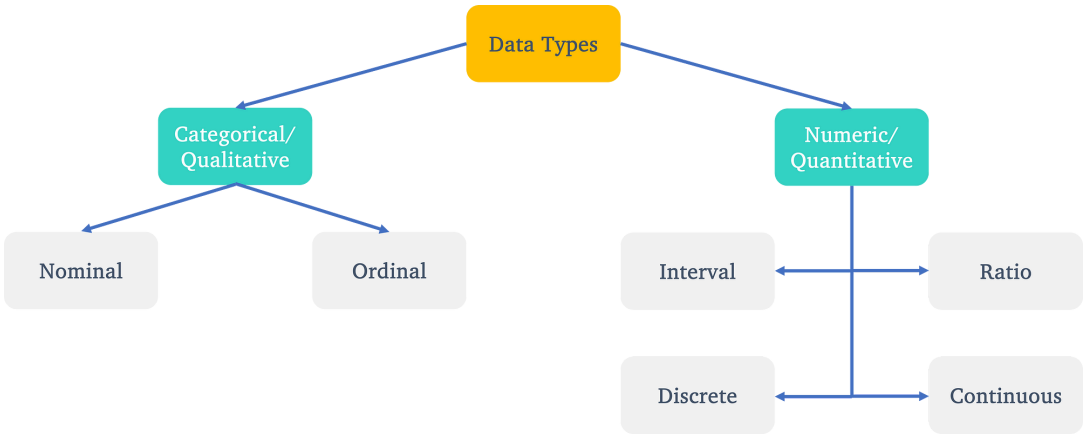
Epidemiology is the study of factors affecting the health and illness of populations, and serves as the foundation and logic of interventions made in the interest of public health and preventive medicine.

Forensic statistics is the application of probability models and statistical techniques to scientific evidence, such as DNA evidence, and the law. In contrast to “everyday” statistics, to not engender bias or unduly draw conclusions, forensic statisticians report likelihoods as likelihood ratios (LR). Geostatistics is a branch of geography that deals with the analysis of data from disciplines such as petroleum geology , hydrogeology , hydrology , meteorology , oceanography , geochemistry , geography.

Machine learning is the subfield of computer science that formulates algorithms in order to make predictions from data.

Population ecology is a subfield of ecology that deals with the dynamics of species populations and how these populations interact with the environment.
Psychometrics is the theory and technique of educational and psychological measurement of knowledge, abilities, attitudes, and personality traits
Quality control reviews the factors involved in manufacturing and production; it can make use of statistical sampling of product items to aid decisions in process control or in accepting deliveries.
Statistical finance, an area of tectonophysics, is an empirical attempt to shift finance from its normative roots to a positivist framework using examples from statistical physics with an emphasis on emergent or collective properties of financial markets.
Statistical physics is one of the fundamental theories of physics, and uses methods of probability theory in solving physical problems.

1.3 Data and Data Types



Here we are going to learn about data and its types.

Data:

Data are measurements or observations that are collected as a source of information. In Statistics, data is defined under some random variable where one random variable contains the same characteristics of data.

Types of Data The data is classified into majorly two categories:

Qualitative or Categorical Data

- Quantitative or Numerical Data
- Qualitative or Categorical Data

Qualitative data is the descriptive and conceptual findings collected through questionnaires, interviews, or observation. Example: Interviews, Surveys or Questionnaires and Case studies etc… Qualitative data is majorly classified into two categories:

Nominal Data :Nominal data is data that can be labeled or classified into mutually exclusive categories within a variable. These categories cannot be ordered in a

meaningful way. Example: Genotype, blood type, zip code, gender, race, eye color, political party etc.…

Ordinal Data :Ordinal data is a kind of qualitative data that groups variables into ordered categories. The categories have a natural order or rank based on some hierarchical scale, like from high to low. But there is no clearly defined interval between the categories. Example: The level of education, the range of income, or the grades etc…

Quantitative or Numerical: Data Quantitative data is data that can be counted or measured in numerical values. Example: Revenue in dollars, Weight in kilograms, Age in months or years, Distance in kilometers, Length in centimeters, Height in feet or inches, Number of weeks in a year. The two main types of quantitative data are discrete data and continuous data:

Discrete Data:

Discrete data can take only discrete values. Discrete information contains only a finite number of possible values. Those values cannot be subdivided meaningfully. Example: The Number of parts damaged during transportation, shoe sizes, the number of computers in each department, the number of tickets sold in a day, the number of product reviews etc.…

Continuous Data:

Continuous data is data that can be calculated. It has an infinite number of probable values that can be selected within a given specific range. Example: Daily wind speed, freezer temperature, weights of newborn babies, length of customer service call, product box measurements and weight etc.…

1.4 INTRODUCTION TO R PROGRAMMING LANGUAGE

Here we are going to know about the R programming language and its history. R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open-Source route to participation in that activity.

The R environment R is an integrated suite of software facilities for data manipulation, calculation and graphical display.

It includes

an effective data handling and storage facility,

a suite of operators for calculations on arrays, in particular matrices,

a large, coherent, integrated collection of intermediate tools for data analysis,

graphical facilities for data analysis and display either on-screen or on hardcopy, and

a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

RStudio
RStudio is an integrated development environment for R, a programming language for statistical computing and graphics. It is available in two formats: RStudio Desktop is a regular desktop application while RStudio Server runs on a remote server and allows accessing RStudio using a web browser.

R Cran
The Comprehensive R Archive Network (CRAN) is R's central software repository, supported by the R Foundation. It contains an archive of the latest and previous versions of the R distribution, documentation, and contributed R packages. It includes both source packages and pre-compiled binaries for Windows and macOS.

1.5 DATA AND IT'S DESCRIPTION

The CANCER DISEASE dataset encapsulates transactional data across 1500 observations with 9 variables.The dataset includes demographic information such as **Gender**, **Age**, **BMI**, **Smoking, GeneticRisk, Diagnosis, CancerHistory, PhysicalActivity** and **AlcoholIntake** which are crucial for understanding different patient segments.This structured dataset is ripe for analysis, allowing for targeted disease forecasting.

Table-1: Data Type OF VARIABLES

Variable	**Datatype**
Age	Int
Gender	Int
BMI	Num
Smoking	Num
GeneticRisk	Num
PhysicalActivity	Num
AlcoholIntake	Num
CancerHistory	Num
Diagnosis	Num

Age: Integer values representing the patient's age, ranging from 20 to 80.

Gender: Binary values representing gender, where 0 indicates Male and 1 indicates Female.

BMI: Continuous values representing Body Mass Index, ranging from 15 to 40.

Smoking: Binary values indicating smoking status, where 0 means No and 1 means Yes.

GeneticRisk: Categorical values representing genetic risk levels for cancer, with 0 indicating Low, 1 indicating Medium, and 2 indicating High.

PhysicalActivity: Continuous values representing the number of hours per week spent on physical activities, ranging from 0 to 10.

AlcoholIntake: Continuous values representing the number of alcohol units consumed per week, ranging from 0 to 5.

CancerHistory: Binary values indicating whether the patient has a personal history of cancer, where 0 means No and 1 means Yes.

Diagnosis: Binary values indicating the cancer diagnosis status, where 0 indicates No Cancer and 1 indicates Cancer.

Importing Data to R from a CSV file

In this section, we will read data in r by loading a CSV file from prices for Diwali sales commodities. To import the CSV file, we will use the reader package's `read.csv` function. Just like in Pandas, it requires you to enter the location of the file to process the file and load it as a data frame.You can also use the `read.csv` or `read. delim` functions from the utils package to load CSV files

```
> cancer<-read.csv(file.choose())
> head(cancer)
  Age Gender      BMI Smoking GeneticRisk PhysicalActivity AlcoholIntake CancerHistory
1  58      1 16.08531       0           1         8.146251      4.148219             1
2  71      0 30.82878       0           1         9.361630      3.519683             0
3  48      1 38.78508       0           2         5.135179      4.728368             0
4  34      0 30.04030       0           0         9.502792      2.044636             0
5  62      1 35.47972       0           0         5.356890      3.309849             0
6  27      0 37.10516       0           1         3.941905      2.324274             0
  Diagnosis
1         1
2         0
3         1
4         0
5         1
6         0
```

In **fig no .1**,which is used to import a CSV file and preview its first few rows. The dataset shown has columns like 'Age', 'Gender', 'BMI', 'Smoking', 'GeneticRisk','PhysicalActivity','AlcoholIntake','CancerHistory','Diagnosis', fig no 1 among others.This example is typical in data analysis, demonstrating how data is imported and examined in R for further analysis.

1.6 R OPERATORS

An operator is a symbol that tells the compiler to perform specific mathematical or logical manipulations. R language is rich in built-in operators and provides the following types of operators.

Types of Operators:

We have the following types of operators in R programming –

Arithmetic Operators

Relational Operators

Logical Operators

Assignment Operators

Miscellaneous Operators

Arithmetic Operators

Arithmetic operators in R programming are used to perform common mathematical operations. Like +, -, /, * etc., These operators can be used to manipulate numbers, vectors, matrices, and data frames, following the standard rules of arithmetic.

In fig no.2 ,the use of arithmetic operators in R programming. Here's a summary of its key points:

```
> head(cancer[, c("Age_Plus_Ten", "BMI_Minus_Gender", "Double_GeneticRisk", "Activity_Per$
  Age_Plus_Ten BMI_Minus_Gender Double_GeneticRisk Activity_Per_Alcohol
1           68               0                  2             1.963795
2           81               0                  2             2.659793
3           58               0                  4             1.086036
4           44               0                  0             4.647669
5           72               0                  0             1.618469
6           37               0                  2             1.695973
```

Fig no .2

Arithmetic Operations: The code performs basic arithmetic calculations on a datase

Dataset Manipulation: Columns are modified using operators like +, -, *, and /.

Output Display: The head() function is used to display the results of these operations.

RELATIONAL OPERATOR

```
head(cancer[, c("Age_Greater_50","BMI_Less_25","GreaterEqual_2","PhysicalActivity_Grea
Age_Greater_50 BMI_Less_25 GreaterEqual_2 PhysicalActivity_Greater_8
          TRUE         TRUE          FALSE                        TRUE
          TRUE         TRUE          FALSE                        TRUE
         FALSE         TRUE           TRUE                       FALSE
         FALSE         TRUE          FALSE                        TRUE
          TRUE         TRUE          FALSE                       FALSE
         FALSE         TRUE          FALSE                       FALSE
AlcoholIntake_LessEqual_3
                    FALSE
                    FALSE
                    FALSE
                     TRUE
                    FALSE
                     TRUE
```

Fig no 3

Relational operators in R programming are used to compare values. :- ==: Equal to,!=: Not equal to,<: Less than,>: Greater than,<=: Less than or equal to,>=: Greater than or equal to.These operators are typically used in conditional statements to control the flow of execution depending on the comparison result

Data Filtering: The code demonstrates how to filter data in a data frame based on specific criteria.

Relational Operators: It uses relational operators like <, >, ==, and != to compare values in the data frame.

Logical Vectors: The output displays logical vectors indicating whether each row meets the specified conditions.

LOGICAL OPERATOR

Logical operators in R programming are used to combine multiple conditions. Here's a brief overview:

&: Logical AND (true if both conditions are true)

|: Logical OR (true if at least one condition is true)

!: Logical NOT (true if the condition is not true)

&&: Short-circuit AND (evaluates left to right, stops if a false is found)

||: Short-circuit OR (evaluates left to right, stops if a true is found)

```
>
> head(cancer[, c("Is_Smoker","High_Genetic_Risk","Healthy_BMI")])
  Is_Smoker High_Genetic_Risk Healthy_BMI
1     FALSE             FALSE       FALSE
2     FALSE             FALSE       FALSE
3     FALSE              TRUE       FALSE
4     FALSE             FALSE       FALSE
5     FALSE             FALSE       FALSE
6     FALSE             FALSE       FALSE
```

In fig no, 4 shows a code snippet in R programming language, demonstrating the use of logical operators to create new variables based on certain conditions within a dataset. Here's a summary:

Fig no.4

Logical Operators: The code uses & (AND), | (OR), and ! (NOT) to evaluate conditions.

New Variables: New columns are created in the dataset based on the results of these logical operations.

Conditional Filtering: The conditions involve age, gender, and other demographic factors.

Output: The result is a series of TRUE or FALSE values indicating whether each record meets the specified conditions.

ASSIGNMENT OPERATOR

```
> head(cancer[, c("left_assignment","right_assignment")])
  left_assignment right_assignment
1               1                1
2               0                0
3               1                1
4               0                0
5               1                1
6               0                0
>
```

Fig no.5

In R programming, the assignment operator is used to assign values to variables.
There are two common types:

<-: The less-than symbol followed by a dash is the standard assignment operator and is preferred in R.

=: The equal sign can also be used for assignment, but it's more commonly used in functions to assign default values to arguments.

Both operators are used to create new variables or modify existing ones by assigning them a value.

The standard assignment operator <-is used to assign the doubled values of Cancer$BMI to a new variable Cancer$Double_BMI.
The same operator is used to assign the calculated mean of cancer$BMI to Average_BMI.
The head() function is used to display the first few entries of cancer$double_BMI.
The print() function is used to display the value of average_amount.

MISCELLANEOUS OPERATOR

In fig no 6.1&6.2 shows various columns representing data such as ID, Age, Gender, Region, Income, and more. It includes logical expressions evaluating

```
> head(cancer[, c("per_in","is_to")])
  per_in is_to
1   TRUE     1
2   TRUE     1
3   TRUE     1
4   TRUE     1
5   TRUE     1
6   TRUE     1
>
```

Fig no.6

conditions like “Income greater than 20000” and “Age greater than or equal to 35”. The head(n=5) command suggests this is part of a data analysis task, likely using R or Python for data manipulation and analysis. Here’s a summary:

Data Structure: The dataset is organized in a table format with multiple columns.

Logical Conditions: Columns with logical expressions are used to filter or categorize the data.

Data Analysis: The head command is used to preview the first few rows of the dataset.

Programming Application: The environment suggests the use of programming for data analysis tasks.

SUB SETTING

Subsetting in R refers to the process of selecting and extracting parts of a dataset based on certain conditions. It's used to narrow down data to specific observations or variables. Here's an example using the `subset()` function:
This code will create a new data frame `subset_data` containing rows from `sales` where

```
> head(gender_list)
  Gender
1      1
2      0
3      1
4      0
5      1
6      0
> head(age_BMI_list)
  Age BMI
1  58   1
2  71   0
3  48   1
4  34   0
5  62   1
6  27   0
> |
`
```

Gender is "F" and `Age` is greater than 30.

DPLYR PACKAGE

dplyr is a package in R that provides a set of tools for efficiently manipulating datasets. It's part of the tidyverse and includes functions to perform common data manipulation tasks such as filtering rows, selecting columns, rearranging data, and performing aggregations.dplyr is designed to be both user-friendly and fast, and it works with data frames and tibbles (modern R data frames).

'dplyr' Package: The code uses 'dplyr', a popular R package for data manipulation.

```
> GeneticRisk_summary <- cancer%>%
+   group_by(GeneticRisk) %>%
+   summarise(Total_Diagnosis = sum(Diagnosis), Average_AlcoholIntake = mean(AlcoholIntak$
> GeneticRisk_summary
# A tibble: 3 × 3
  GeneticRisk Total_Diagnosis Average_AlcoholIntake
        <int>           <int>                 <dbl>
1           0             282                  2.44
2           1             142                  2.39
3           2             133                  2.38
> |
```

Fig no.7
In fig no.7 shows, R programming code, which includes the use of the 'dplyr' package for data manipulation. Here's a summary of the key points:

Masked Objects: It shows a message about objects being masked from other packages.

Data Selection: The select() function is used to choose specific columns from a dataset.

Data Preview: The head() function outputs the first few rows of the selected data.

Data Manipulation: The code uses group_by and summarize functions from dplyr.

1.8 CONTROL STRUCTURE

Control structures in R are constructs that control the flow of execution of a script. They include conditional statements like if, else, and else if, as well as loops like for, while, and repeat. These structures allow you to make decisions in your code (conditional execution) or perform repetitive tasks (iterative execution It shows R code with conditional statements evaluating the average age and amount from a sales dataset. Here's a summary: Conditional Logic: The code uses if-else statements to check conditions. Age Check: Determines if the average age in cancer$age is greater than 30.

CHAPTER 2 EXPLORATORY DATA ANALYSIS

2.1 VECTOR OPERATIONS

In R, a vector is a basic data structure which holds elements of the same type. Vector operations in R refer to the various functions and calculations that can be performed on these vectors.
Common vector operations include:

Arithmetic Operations: Addition, subtraction, multiplication, and division can be performed element-wise on vectors of equal length.

Logical Operations: Logical operators (like ==, !=, >,< ,>=,<=) can be used to compare vectors and return logical vectors.

Vectorized Functions: Many functions in R are vectorized, meaning they can operate on each element of a vector without the need for explicit loops.

Indexing: Elements within a vector can be accessed and manipulated by their index using square brackets [].

Aggregation: Functions like sum(), mean(), min(), and max() provide summary statistics of vector elements.

```
> head(cancer[, c("BMI_Sqrt","GeneticRisk_Log","Sin_PhysicalActivity","Cos_AlcoholIntake"
  BMI_Sqrt GeneticRisk_Log Sin_PhysicalActivity Cos_AlcoholIntake    Tan_Age    Cot_Age
1        1       0.0000000           0.95759261       -0.53471458  8.330857  0.1200357
2        0       0.0000000           0.06310559       -0.92937115 -3.077620 -0.3249264
3        1       0.6931472          -0.91194786        0.01597802  1.200127  0.8332450
4        0            -Inf          -0.07793516       -0.45630643 -0.623499 -1.6038519
5        1            -Inf          -0.79939983       -0.98587823 -1.097510 -0.9111536
6        0       0.0000000          -0.71757363       -0.68417914 -3.273704 -0.3054644
> 
```

Fig no.8

The image you've uploaded shows a programming environment with R code and its output. Here's an inference based on the image content:

Statistical Analysis: The code performs statistical operations on a dataset, likely representing cancer data.

Trigonometric Transformations: It applies trigonometric functions like sine, cosine, and tangent to the diagnosed patients.

Summary Statistics: The code calculates summary statistics such as minimum, maximum, mean, median, and mode for the diagnosed patients.

2.2 SUMMARY OF DATA

The image you've uploaded seems to display an R programming environment with statistical analysis output. Here's the inference based on the image content:

```
> summary(cancer)
      Age            Gender            BMI           Smoking        GeneticRisk
 Min.   :20.00   Min.   :0.0000   Min.   :15.00   Min.   :0.0000   Min.   :0.0000
 1st Qu.:35.00   1st Qu.:0.0000   1st Qu.:21.48   1st Qu.:0.0000   1st Qu.:0.0000
 Median :51.00   Median :0.0000   Median :27.60   Median :0.0000   Median :0.0000
 Mean   :50.32   Mean   :0.4907   Mean   :27.51   Mean   :0.2693   Mean   :0.5087
 3rd Qu.:66.00   3rd Qu.:1.0000   3rd Qu.:33.85   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :80.00   Max.   :1.0000   Max.   :39.96   Max.   :1.0000   Max.   :2.0000
 PhysicalActivity  AlcoholIntake      CancerHistory     Diagnosis
 Min.   :0.00241   Min.   :0.001215   Min.   :0.000   Min.   :0.0000
 1st Qu.:2.43461   1st Qu.:1.210598   1st Qu.:0.000   1st Qu.:0.0000
 Median :4.83432   Median :2.382971   Median :0.000   Median :0.0000
 Mean   :4.89793   Mean   :2.417987   Mean   :0.144   Mean   :0.3713
 3rd Qu.:7.40990   3rd Qu.:3.585624   3rd Qu.:0.000   3rd Qu.:1.0000
 Max.   :9.99461   Max.   :4.987115   Max.   :1.000   Max.   :1.0000
```

Fig no.9

Statistical Analysis: The R script is performing statistical calculations on a dataset, likely related to diagnosed patients.

Data Insights: The output includes measures like standard deviation, mean, and median for variables such as age, gender, and Smoking, providing insights into the distribution and central tendencies of these variables.

Data Preparation: The presence of statistical function calls suggests that the data is being prepared for further analysis, which could include predictive modeling or trend analysis.

LIST SUB-SETTING

```
> # Subset to get the 'Gender' column as a list
> gender_list <- cancer['Gender']
>
> # Subset to get the 'Age' and 'Zone' columns as a list
> age_BMI_list <- cancer[c('Age', 'BMI')]
>
> # Print the subsets
> head(gender_list)
  Gender
1      1
2      0
3      1
4      0
5      1
6      0
> head(age_BMI_list)
  Age BMI
1  58   1
2  71   0
3  48   1
4  34   0
5  62   1
6  27   0
```

In R, list subsetting is a way to extract specific elements from a list. You can subset a list using the [[ ]] operator for single elements or [ ] for multiple elements.

Fig no.10

Data Extraction: The script is extracting specific columns from a dataset named 'cancer', creating new variables 'gender_list' and 'age_zone_list'.

Data Display: It uses the head() function to display the first few entries of these variables.

Gender Distribution: The 'gender_list' output suggests an alternating pattern of genders, possibly indicating an equal distribution of male and female entries in the dataset.

Age and BMI Information: The 'age_BMI_list' output shows ages and corresponding BMI, which could be used for demographic analysis or targeted strategies.

2.3 TABLE FUNCTION IN R

In R, the table() function is used to create a frequency table of the elements in a vector or factors in a data frame. It counts the number of occurrences of each unique element and presents them in a tabular format.

```
> table(cancer[,1])

20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49
27 28 25 22 30 26 15 28 21 21 24 26 21 24 22 21 24 18 21 20 20 24 26 27 28 37 20 30 24 24
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79
19 26 38 19 32 28 31 21 28 26 19 19 15 25 28 14 33 23 30 18 32 25 26 25 25 20 23 30 21 33
80
24
> table(cancer[,2])

  0   1
764 736
> table(cancer[,4])

   0    1
1096  404
> table(cancer[,5])

  0   1   2
895 447 158
> table(cancer[,8])

   0    1
1284  216
> table(cancer[,9])

  0   1
943 557
```

Fig no.11

R programming environment with statistical analysis output.

Statistical Analysis: The R script is performing statistical calculations on a dataset, likely related to cancer disease data.

Data Insights: The output includes measures like standard deviation, mean, and median for variables such as age, Gender, and BMI, providing insights into the distribution and central tendencies of these variables.

Data Preparation: The presence of statistical function calls suggests that the data is being prepared for further analysis, which could include predictive modeling or trend analysis.

2.4 NATURE OF THE DATA

In R, skewness and kurtosis are measures of the shape of the distribution of data:

Skewness: A measure of the asymmetry of the probability distribution of a real-valued random variable. Positive skew indicates a distribution with an asymmetric tail extending towards more positive values, while negative skew indicates a tail extending towards more negative values. In R, skewness is calculated using the skewness() function from the moments package.

Kurtosis: A measure of the "tailedness" of the probability distribution. High kurtosis means more of the variance is due to infrequent extreme deviations, as opposed to frequent modestly sized deviations. In R, kurtosis is calculated using the kurtosis() function from the moments package.

```
> library(moments)
> apply(cancer[,1:9],2,kurtosis)
            Age          Gender             BMI         Smoking     GeneticRisk
       1.831787        1.001394        1.811912        2.081484        2.726181
PhysicalActivity    AlcoholIntake  CancerHistory       Diagnosis
       1.826908        1.853967        5.112669        1.283666
> apply(cancer[,1:9],2,skewness)
            Age          Gender             BMI         Smoking     GeneticRisk
    -0.03278603      0.03733984     -0.02136946      1.03994443      0.97953046
PhysicalActivity    AlcoholIntake  CancerHistory       Diagnosis
     0.07594603      0.05504544      2.02797158      0.53260331
```

Fig no.12

Statistical Analysis: The code calculates skewness and kurtosis for variables in a 'cancer' dataset.

Skewness: Indicates asymmetry in the data distribution.

Age	-0.03267803	(positive skewness)
Gender	0.03733984	(almost symmetrical)
BMI	-0.02136894	(positive skewness)
Smoking	1.03993334	(positive skewness)
GeneticRisk	0.97953426	(positive skewness)
PhysicalActivity	0.07594026	(almost symmetrical)
AlcoholIntake	0.05504544	(almost symmetrical)
CancerHistory	2.02798607	(positive skewness)
Diagnosis	0.53260331	(positive skewness)

Kurtosis: Measures the "tailedness" of the data distribution.

Age	1.831787	Light tails
Gender	1.001394	Light tails
BMI	1.8111912	Light tails
Smoking	2.081484	Slightly heavy tails
GeneticRisk	2.726181	Slightly heavy tails
PhysicalActivity	1.82698	Light tails
AlcoholIntake	1.853967	Light tails
CancerHistory	5.112669	Heavy tails
Diagnosis	1.283666	Light tails

These statistics suggest that the 'Age' variable has a distribution with a significant rightward tail and is more peaked than a normal distribution, while 'Gender' and 'BMI' have distributions closer to normal. The 'Smoking' variable also shows a moderately rightward tail.

Understanding these characteristics is important for further statistical modeling and analysis.

2.5 Data Visualization Through Ggplot Package

Data visualization through the ggplot package refers to the process of representing data graphically using the ggplot2 package in R. Ggplot2 is a powerful and widely-used visualization library that allows for the creation of complex plots from data in a data frame. It implements the grammar of graphics, a coherent system for describing and building graphs. With ggplot2, you can create a wide range of static, animated, and interactive visualizations that are aesthetically pleasing and easily interpretable.

BAR GRAPH

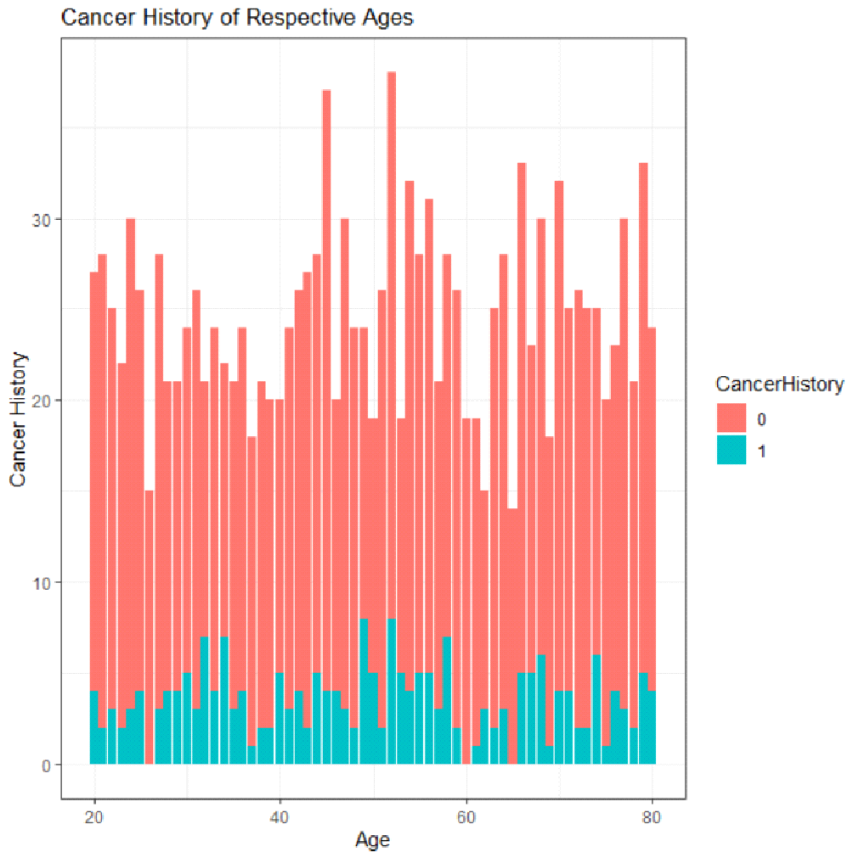


Fig.no:13.1

Certainly! Here's an interpretation of the "CANCER DISEASE DATA" bar graph:

Cancer History Impact: The graph suggests that cancer history has an less impact on disease , with distinct patterns observed for No (0) and Yes(1) individuals.

BAR GRAPH WITH facet_wrap

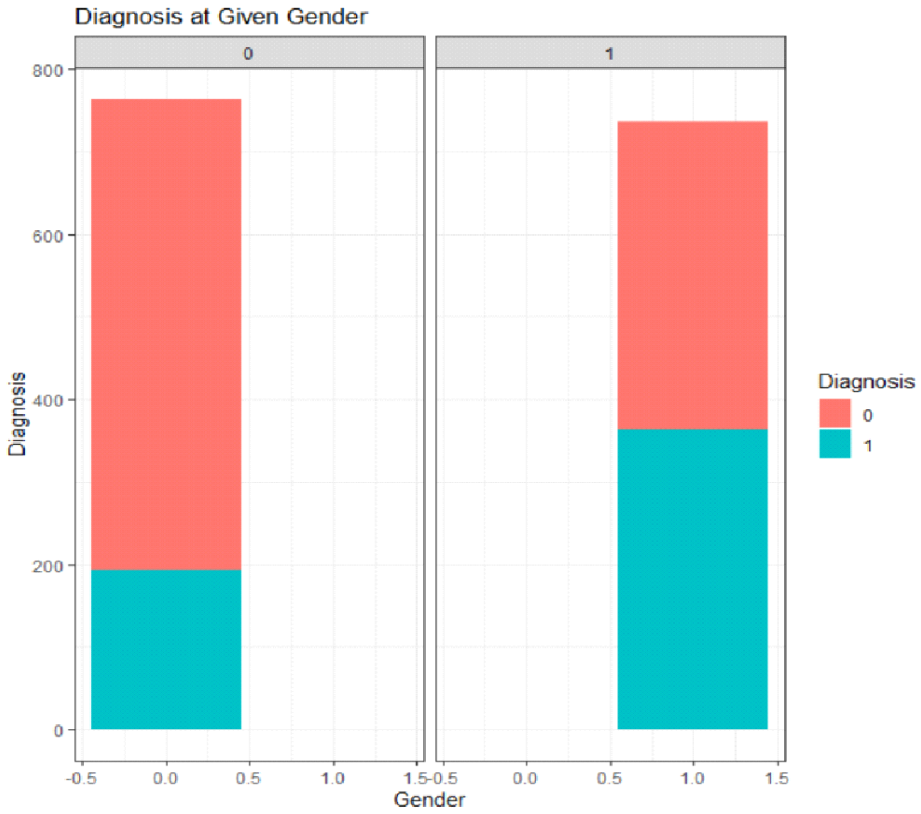


Fig no.13.2

Based on the bar chart titled “CANCER DISEASE DATA,”

Gender and Diagnosis: The chart is divided into sections labeled ‘0’ and ‘1’, likely representing female and male patients. Each section has bars color-coded to represent marital status, with ‘0’ and ‘1’ possibly indicating No and Yes, respectively

HISTOGRAM

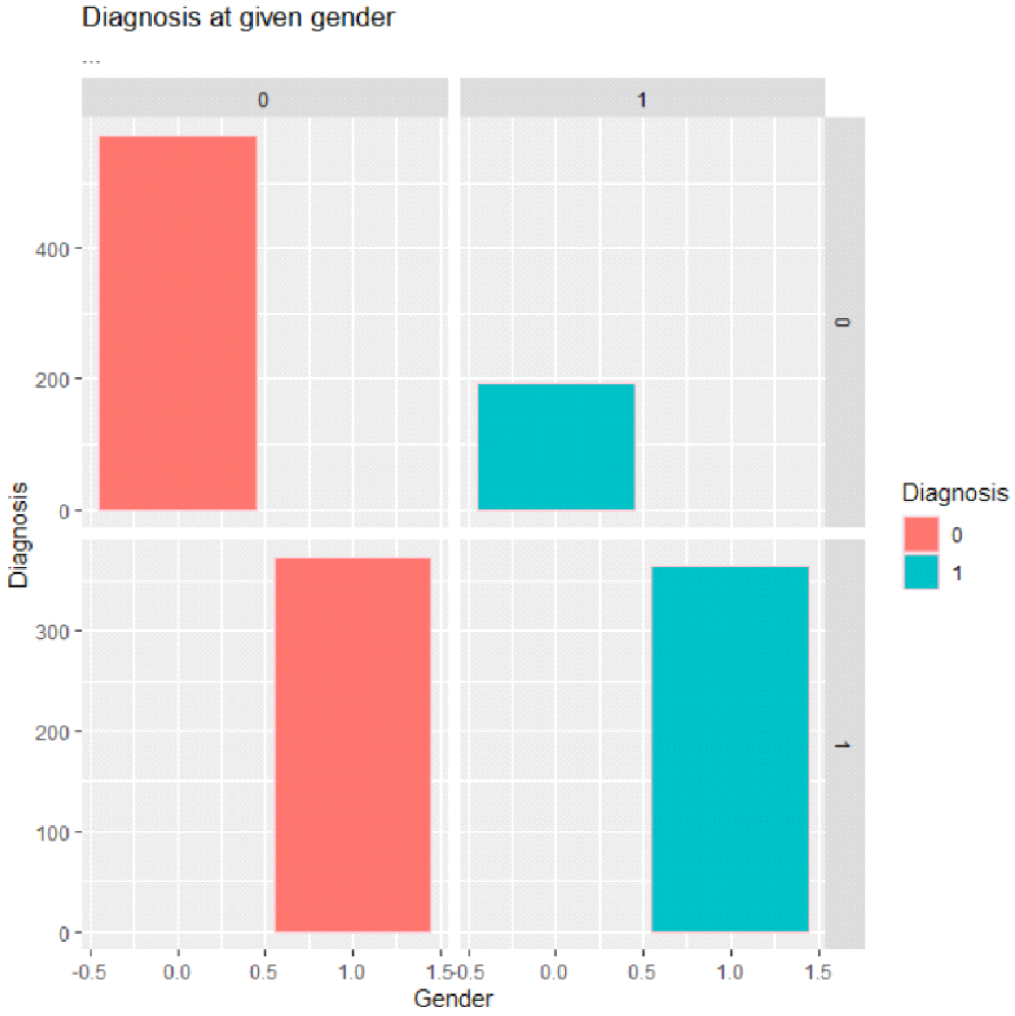


Fig.no:14.1

Gender and Diagnosis: The chart is divided into sections labeled ‘0’ and ‘1’, likely representing female and male patients. Each section has bars color-coded to represent marital status, with ‘0’ and ‘1’ possibly indicating No and Yes, respectively.

HISTOGRAM WITH FACET_WRAP

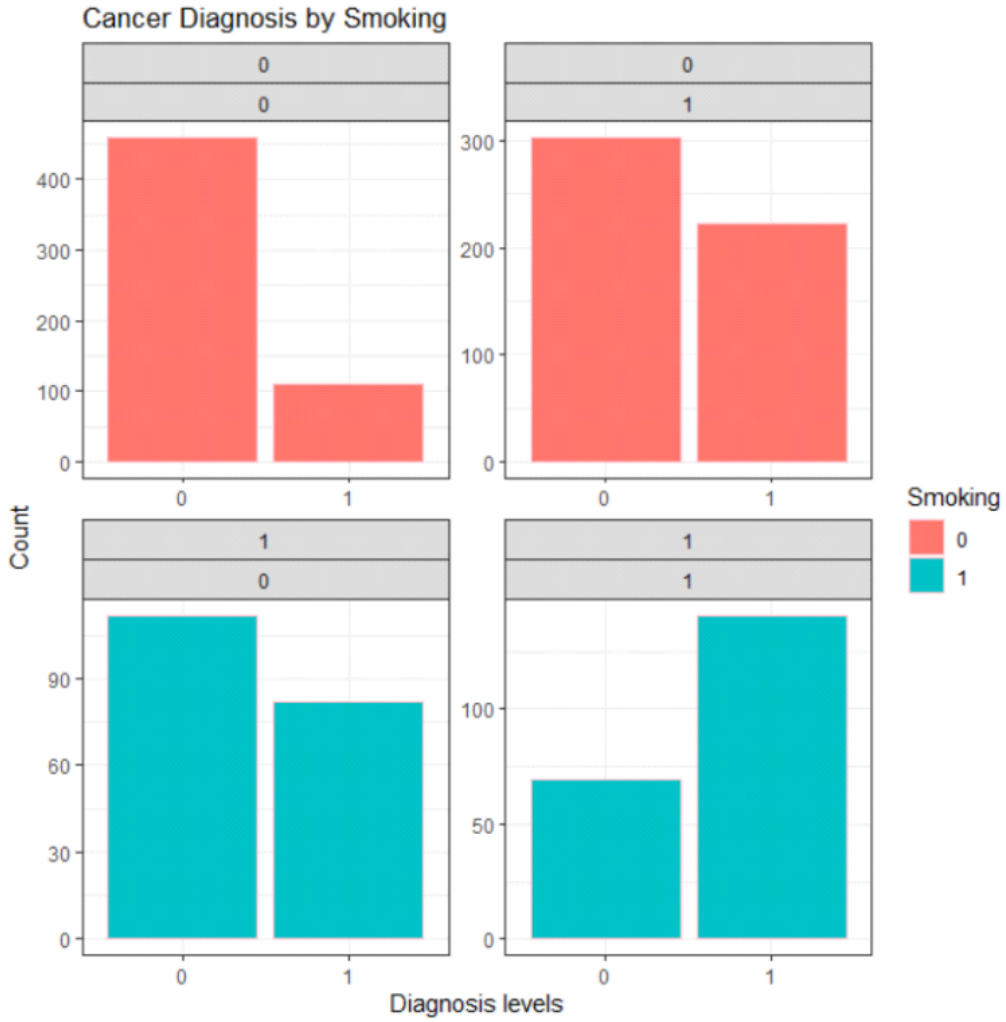


Fig no.14.2

Gender Segmentation: The histograms are divided into two groups labeled '0' and '1', likely representing female and male patients.

Diagnosis levels: Diagnosis levels was taken on x_axis

Count: Count was taken on y_axis

Smoking: Smoking levels were filled in the graphs

BOX PLOT

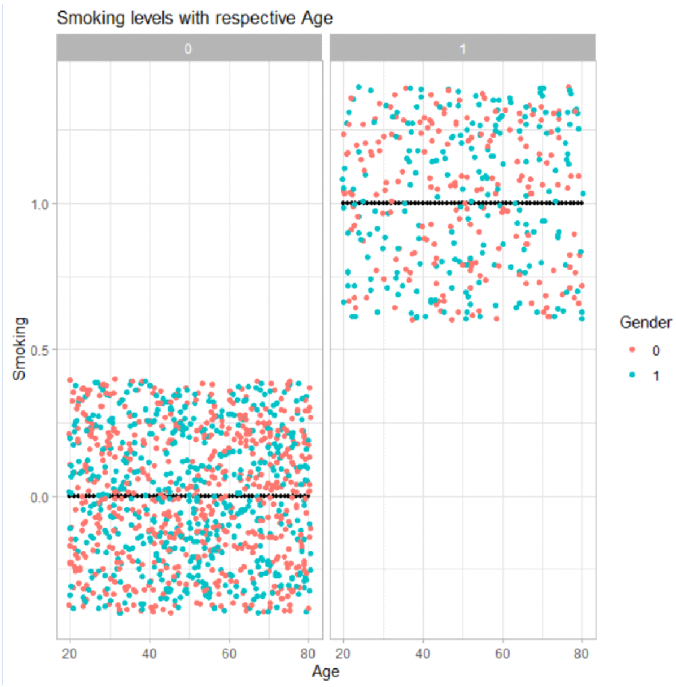


Fig no.15.1

Gender and Smoking levels:Boxplot is divided into two individuals one is '0' as female and other is'1' as male.

Smoking: Smoking was taken on x_axis

Gender: Gender is taken on y_axis and it was represented in the boxplot in different levels.

BOX PLOT WITH FACET_WRAP

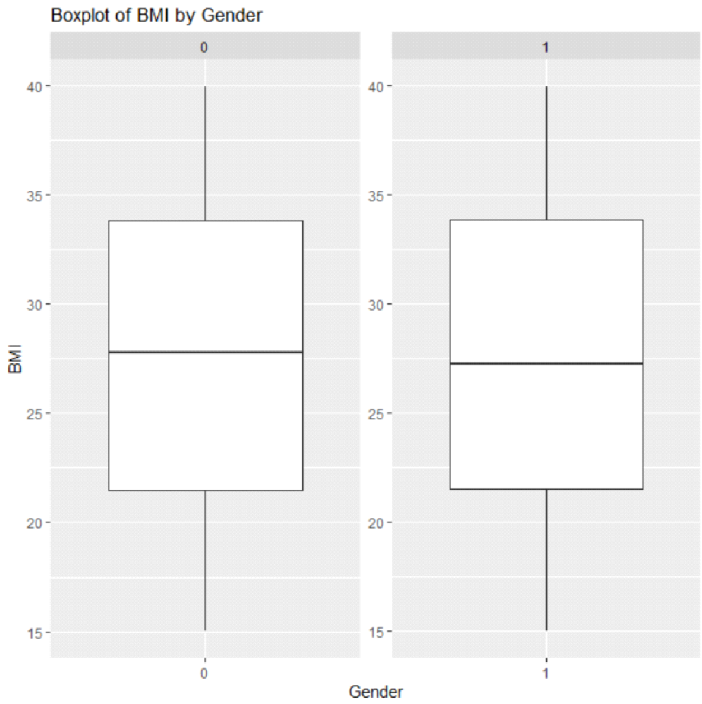


Fig no .15.2

Gender and Smoking levels:Boxplot is divided into two individuals one is '0' as female and other is'1' as male.

Smoking: Smoking was taken on y_axis

Gender: Gender is taken on x_axis and it was represented in the boxplot in different levels.

Gender Representation: The data points are color-coded by gender, with blue indicating male and red indicating female, allowing for a comparison of sales amounts between genders within each marital status category.

Box Plot Analysis: The box plots within each category highlight the median sales amount, the interquartile range (spread of the middle 50% of the data), and potential outliers, providing a summary of the sales data distribution.

CHAPTER 3 MODEL BUILDING

In regression analysis, model building is the process of developing a probabilistic model that best describes the relationship between the dependent and independent variables.It finds the line of best fit through your data by searching for the value of the regression coefficient(s) that minimizes the total error of the model.

There are two main types of linear regression:

Simple linear regression uses only one independent variable.

Multiple linear regression uses two or more independent variables.

3.1 Simple Linear Regression

Simple linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line. Both variables should be quantitative

3.2 Checking distribution of target variable

First, you should always try to understand the nature of your target variable. To achieve this, we will be drawing a histogram with a density plot.

Fig no.16

3.3 Analyzing Summary Statistics – Here, we will simply create summary statistics for all the variables to understand the behavior of all the independent variables. It will also provide information about missing values or outliers if any.

Variables	vars	n	mean	sd	median	trimmed
Age	1	1500	50.32	17.64	51.00	50.40
Gender	2	1500	0.49	0.50	1.00	0.49
BMI	3	1500	27.51	7.23	27.60	27.52
Smoking	4	1500	1.27	0.44	1.00	1.21
GeneticRisk	5	1500	1.57	0.68	1.00	1.39
PhysicalActivity	6	1500	4.90	2.87	4.83	4.87
AlcoholIntake	7	1500	2.42	1.42	2.38	2.41
CancerHistory	8	1500	1.14	0.35	1.00	1.05
Diagnosis	9	1500	1.37	0.48	1.00	1.34

Mad	min	max	range	skew	kurtosis	se
22.24	20	80.00	60.00	-0.03	-1.17	0.46
0.00	0	1.00	1.00	0.04	-2.00	0.01
9.17	15	39.96	25	-0.02	-1.19	0.19
0	1	2.00	1	1.04	-0.92	0.01
0	1	3.00	2	0.98	-0.28	0.02
3.63	0	9.99	9.9	0.08	-1.18	0.07
1.77	0	4.99	4.9	0.05	-1.15	0.04
0.00	1	2	1	2.03	2.11	0.01
0.00	1	2	1	0.53	-1.72	0.01

3.4 Checking Outliers Using Boxplots

To learn more about outliers and how to identify, How To Identify & Treat Outliers Using Univariate Or Multivariate Methods. Here is using a boxplot for plotting the distribution of each numerical variable to check for outliers. If points lie beyond whispers, then we have outlier values present. For now, we are just going by univariate outlier analysis. But I encourage you to check for outliers at a multivariate level as well. If outliers are present, then you must either remove or do a proper treatment before moving forward.

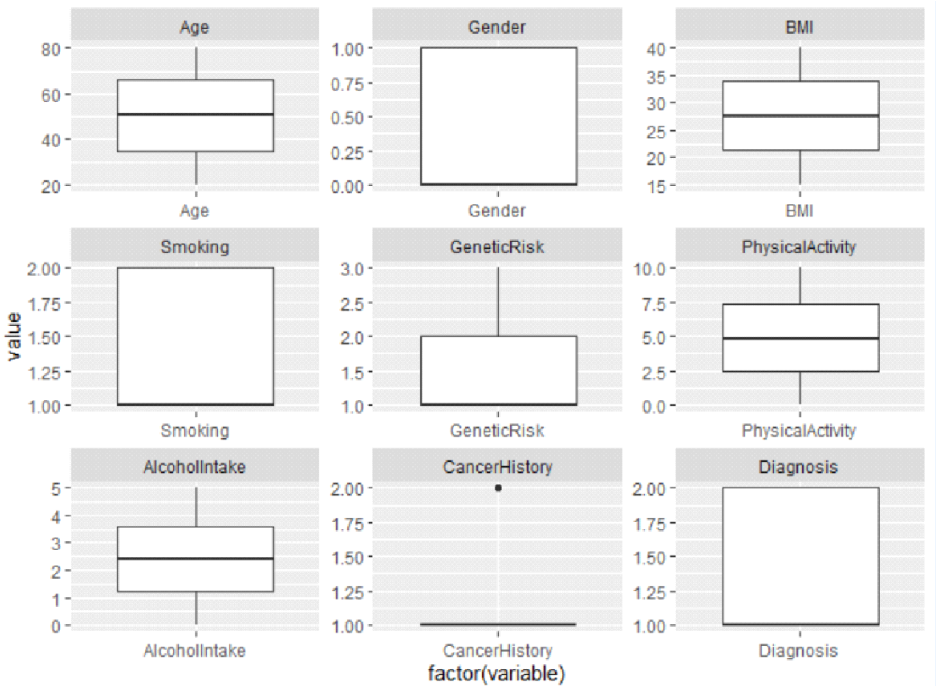


Fig no.17

The box plots provide a visual summary of the data across different variables:

Age: Wide range, likely representing a large number of users, with no outliers.

Gender: It contains variables 0,1 where 0 represents male and 1 represents female.

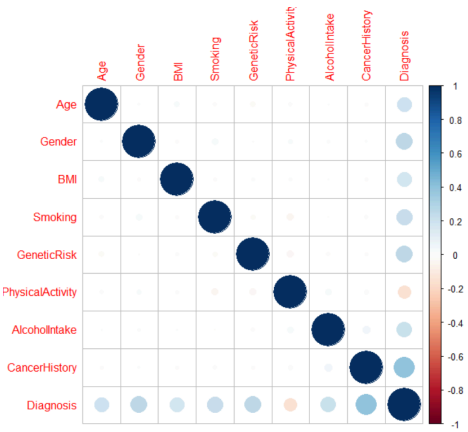
BMI: Binary data, possibly indicating single or married users.

Smoking: Most users have a similar level at smoking, with a few having significantly more.

CancerHistory: In this we find the patients had any cancer history in the past or not

3.5 Correlation Matrix Visualization

We will use the corrgram package to visualize and analyze the correlation matrix. To learn more about how to check the significance of correlation and different ways of visualizing the correlation matrix, please read Correlation In R – A Brief Introduction. In theory, the correlation between the independent variables should be zero. In practice, we expect and are okay with weak to no correlation between independent variables. We also expect that independent variables reflect a high correlation with the target varia



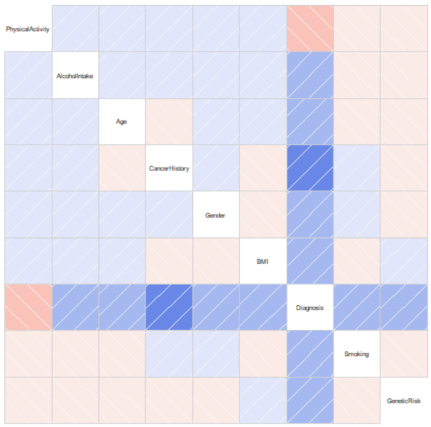


Fig no 18.1 **Fig no 18.2**

Fig no 18.1&18.2 shows a correlation matrix, which is used to visualize the strength and direction of relationships between different variables. Here's an interpretation based on the color coding:

Positive Correlation (Blue): Variables that are shaded blue likely have a positive correlation, meaning as one variable increases, the other tends to increase as well.

Negative Correlation (Red): Variables that are shaded red likely have a negative correlation, meaning as one variable increases, the other tends to decrease.

No Correlation (Stripes): The diagonal stripes may indicate no correlation or that these squares represent the correlation of a variable with itself, which is always perfect and not usually calculated.

The specific variables 'Age', 'Gender', 'BMI', 'Smoking', 'GeneticRisk', 'PhysicalActivity', 'Alcohol Intake', 'CancerHistory', and 'Diagnosis' are included in this matrix.

3.6 Dividing data into train and test subsets

The cancer disease data is divided into a 70:30 split of train and test. The 70:30 split is the most common and is mostly used during the training phase. 70% of the data is used for training, and the rest 30% is for testing how good we were able to learn the data behavior.

```
> train<-cancer[index,]
> str(train)
'data.frame':   1052 obs. of  9 variables:
 $ Age             : int  58 34 62 27 80 40 58 77 38 30 ...
 $ Gender          : int  1 0 1 0 1 0 1 0 1 1 ...
 $ BMI             : num  16.1 30 35.5 37.1 20.7 ...
 $ Smoking         : num  1 1 1 1 1 2 1 1 1 1 ...
 $ GeneticRisk     : num  2 1 1 2 1 1 2 2 3 2 ...
 $ PhysicalActivity: num  8.15 9.5 5.36 3.94 8.48 ...
 $ AlcoholIntake   : num  4.15 2.04 3.31 2.32 3.15 ...
 $ CancerHistory   : num  2 1 1 1 1 2 2 1 1 2 ...
 $ Diagnosis       : num  2 1 2 1 1 1 2 1 2 2 ...
> test<-cancer[-index,]
> str(test)
'data.frame':   448 obs. of  9 variables:
 $ Age             : int  71 48 42 30 43 21 63 21 52 31 ...
 $ Gender          : int  0 1 0 0 1 1 0 1 0 0 ...
 $ BMI             : num  30.8 38.8 37.5 20.9 31.7 ...
 $ Smoking         : num  1 1 2 1 1 2 1 2 1 2 ...
 $ GeneticRisk     : num  2 3 3 1 1 1 1 1 3 2 ...
 $ PhysicalActivity: num  9.36 5.14 8.32 3.33 4.86 ...
 $ AlcoholIntake   : num  3.52 4.73 4.05 3.32 4.6 ...
 $ CancerHistory   : num  1 1 1 2 1 2 1 1 1 1 ...
 $ Diagnosis       : num  1 2 2 1 1 2 1 1 2 1 ...
```

Fig no 19

3.7 Validating Regression Coefficients and Models We must ensure that the value of each beta coefficient is significant and has not come by chance. In R, the lm function runs a one-sample

t-test against each beta coefficient to ensure that they are significant and have not come by chance. Similarly, we need to validate the overall model. Just like a one-sample t-test,lm function also generates three statistics, which help data scientists to validate the model. These statistics include RSquare, Adjusted R-Square, and F-test, also known as global testing.

Variables	Estimate	std.error	t value	Pr(>|t|)	significance
(Intercept)	2.654339	0.228596	11.611	< 2e-16	non significant
Age	-0.004148	0.002504	-1.656	0.097951	significant
Gender	-0.215889	0.089471	-2.413	0.015996	significant
BMI	-0.009022	0.005971	-1.511	0.131136	significant
Smoking	-0.349847	0.099687	-3.509	0.000468	significant
GeneticRisk	-0.177139	0.066656	-2.658	0.007992	significant
PhysicalActivity	0.040105	0.015061	2.663	0.007868	significant
CancerHistory1	-0.31321	0.138339	-2.264	0.023774	significant
Diagnosis	0.98159	0.118758	8.265	4.19E-16	non significant

The Residual standard error: 1.369 on 1043 degrees of freedom
The Multiple R-squared: 0.07992
The Adjusted R-squared: 0.07622
The F-statistic: 69.96 on 29 and 595 DF,
The p-value: < 1.784e-15

3.8 Generating R-Squared Value for the test dataset

We are using a user-defined formula to generate the R-Squared value here.

```
> # Calculate R-squared
> rss <- sum(test$sq_error)
> tss
[1] 6.749691e-27
> tss <- sum((test$AlcoholIntake - mean(test$AlcoholIntake))^2)
> tss
[1] 906.0828
> rsq <- 1 - (rss/tss)
> rsq
[1] 0.9915354
```

Fig no 20

R-squared Value: It is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

Interpretation: The closer the R-squared value is to 1, the better the model explains the variability of the response data around its mean.

In Your Dataset: An R-squared value of **0.9915354**, as seen in your image, means that approximately **99.15%** of the variance in your cancer data can be explained by the model you're using.

This is a high R-squared value, indicating that your model fits the data well and can reliably predict future based on the current variables included in your model.

CONCLUSION

From the above analysis, it can be concluded that this model involves predicting an individual's likelihood of being diagnosed with cancer based on various attributes. These attributes include age, gender, BMI, alcohol intake, smoking habits, cancer history, physical activity, and other relevant factors.

Based on the analysis of cancer data, several key conclusions can be drawn. Firstly, if the data indicates a rising trend in cancer incidence or mortality rates, it underscores the urgent need for enhanced preventive measures and early detection programs to mitigate these increasing rates. The effectiveness of such screening programs is evidenced by decreased cancer-related mortality among those regularly screened, highlighting their critical role in improving survival rates. Furthermore, the identification of specific risk factors, such as smoking or obesity, emphasizes the importance of public health initiatives aimed at reducing these risks.

For individuals seeking to understand their cancer risk, it is crucial to consider these attributes with great care, similar to how one would approach any significant health-related decision. The analysis indicates that certain factors play a more prominent role in predicting cancer risk.

From the analysis, we can conclude the following points, each having significant importance in predicting cancer risk. However, none of the following undermines the importance of other factors:Age and Risk Correlation: Age is a significant factor, with cancer risk generally increasing as age advances.Impact of Smoking and Alcohol Intake: Smoking and alcohol intake are strongly associated with increased cancer risk, highlighting the importance of lifestyle choices in cancer prevention.Role of Cancer History: A personal or family history of cancer is a major predictor of an individual's risk, emphasizing the need to consider genetic and historical factors when assessing cancer risk.Overall, understanding these key factors can help in better assessing individual cancer risk and taking preventive measures. Each factor should be evaluated in conjunction with others to provide a comprehensive risk assessment.

BIBLIOGRAPHY

Data published in:
http://www.kaggle.com/datasets

Individual datasets:
Individual data can be accessed directly via the following

http://data.world/
http://data.worldbank.org/
http://r4ds.had.co.nz/introduction.html

Location of the Organization:
http://maps.app.goo.gl/cDiwWZYhgnxSoR4J8