HEALTH DATA ANALYSIS

A COURSE PROJECT REPORT

By

RATAN PRIYA SINGH (RA2111027010065)

SNEHAL SUKUNDARI (RA2111027010049)

Under the guidance of

Mrs. D. Hemavathi

Assistant professor

Department of Data Science and Business Systems

In partial fulfilment for the Course

of

18CSE396T – DATA SCIENCE

in

Department of Data Science and Business Systems

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

Kattankulathur, Chenpalpattu District

NOVEMBER 2023

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

BONAFIDE CERTIFICATE

Certified that this mini project report "HEALTH DATA ANALYSIS" is the bonafide work of RATAN PRIYA SINGH (RA2111027010065) and SNEHAL SUKUNDARI(RA2111027010049) who carried out the project work under my supervision.

Mrs. D. Hemavathi

Assistant professor
Department of Data Science and Business Systems
SRM institute of science and technology

ACKNOWLEDGEMENT

We express our heartfelt thanks to our honorable **Vice Chancellor Dr. C. MUTHAMIZHCHELVAN**, for being the beacon in all our endeavors.

We would like to express my warmth of gratitude to our Registrar **Dr. S. Ponnusamy**, for his encouragement.

We express our profound gratitude to our Dean of College of Engineering and Technology, **Dr. T. V.Gopal,** for bringing out novelty in all executions.

We would like to express my heartfelt thanks to the Chairperson, School of Computing **Dr. Revathi Venkataraman**, for imparting confidence to complete my course project

We wish to express my sincere thanks to Course Audit Professor **Dr. Annapurani Panaiyappan**, Professor and Head, Department of Networking and Communications and Course Coordinators for their constant encouragement and support.

We are highly thankful to our course project faculty **Mrs. D.Hemavathi** ,Assistant Professor , Department of DSBS for her assistance, timely suggestion and guidance throughout the duration of this course project.

We extend my gratitude to our HoD, Professor **Dr. M. Lakshmi ,** Department of DSBS and my Departmental colleagues for their Support.

Finally, we thank our parents and friends near and dear ones who directly and in directly contributed to the successful completion of our project. Above all, I thank the almighty for showering his blessings on me to complete my Course project

TABLE OF CONTENTS

ABSTRACT

Study investigates the effects of regular exercise on cardiovascular health in a cohort of 500 participants over a 12-month period. Participants were divided into two groups: an exercise intervention group and a control group. The exercise group engaged in supervised aerobic and strength training sessions three times a week, while the control group maintained their usual activity levels. Cardiovascular health markers including blood pressure, cholesterol levels, and heart rate variability were monitored at baseline, 6 months, and 12 months. Results indicate a significant improvement in cardiovascular health parameters among the exercise group compared to the control group. This study provides valuable insights into the potential benefits of exercise interventions for cardiovascular health improvement.

PROJECT STATEMENT

Health Data Analysis provides data to local health authorities, healthcare providers, and policymakers to guide the allocation of resources. To predict how much a given customer's cost for the medical bill for the year based on some generic features, in this project we'll use R to create regression model.

OBJECTIVE

Big part of life is insurance and insurance companies charge individuals premiums based off some basic information of who they are, where they live and what they live and what they do. Insurers use predictive modeling to estimate future healthcare costs, which can inform pricing and risk management strategies. Analyzing health data can enhance the accuracy of these models. They have this problem that if they charge too much they're going to lose customers but if they charge enough they're going to go out of the business and be bankrupt.

DATA SET

age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.92
18	male	33.77	1	no	southeast	1725.552
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.47
32	male	28.88	0	no	northwest	3866.855
31	female	25.74	0	no	southeast	3756.622
46	female	33.44	1	no	southeast	8240.59
37	female	27.74	3	no	northwest	7281.506
37	male	29.83	2	no	northeast	6406.411
60	female	25.84	0	no	northwest	28923.14
25	male	26.22	0	no	northeast	2721.321
62	female	26.29	0	yes	southeast	27808.73
23	male	34.4	0	no	southwest	1826.843
56	female	39.82	0	no	southeast	11090.72
27	male	42.13	0	yes	southeast	39611.76
19	male	24.6	1	no	southwest	1837.237
52	female	30.78	1	no	northeast	10797.34
23	male	23.845	0	no	northeast	2395.172
56	male	40.3	0	no	southwest	10602.39
30	male	35.3	0	yes	southwest	36837.47
60	female	36.005	0	no	northeast	13228.85
30	female	32.4	1	no	southwest	4149.736
18	male	34.1	0	no	southeast	1137.011
34	female	31.92	1	yes	northeast	37701.88
37	male	28.025	2	no	northwest	6203.902
59	female	27.72	3	no	southeast	14001.13
63	female	23.085	0	no	northeast	14451.84
55	female	32.775	2	no	northwest	12268.63
23	male	17.385	1	no	northwest	2775.192

insurance

ALGORITHM

Input: Algorithms take some input or data as their initial information, which they process to produce an output.

Finite Steps: Algorithms consist of a finite number of well-defined steps. Each step must be precise and unambiguous.

Deterministic: Algorithms are deterministic, meaning that for a given input, they will always produce the same output.

Termination: Algorithms are designed to terminate after a finite number of steps, providing a solution or result.

Correctness: An algorithm should produce the correct output for all possible valid inputs. It must solve the problem as intended.

Efficiency: Efficiency is a crucial consideration in algorithm design. An efficient algorithm should complete the task in a reasonable amount of time and with minimal resource usage.

Optimality: In some cases, algorithms aim to find an optimal solution, meaning the best possible outcome given specific constraints.

CODE

```
---
  title: "Medical Charge Predictions"
output: html_document
date: "2022-08-25"
---
{r setup , include=FALSE}
knitr::opts_chunk$set(echo = TRUE)

## Load in the data
{r}
df = read.csv('insurance.csv', header=TRUE)
num_cols <- unlist(lapply(df, is.numeric))
plot(df[,num_cols])
insurance <- read.csv("D:/predict-hospital-bills-main/insurance.csv")
num_cols <- sapply(insurance, is.numeric)
plot(insurance[, num_cols])
insurance <- read.csv("D:/predict-hospital-bills-main/insurance.csv")
num_cols <- sapply(insurance, is.numeric)
print(num_cols)
plot(insurance[, num_cols])
# Read the insurance data into a data frame
insurance <- read.csv("D:/predict-hospital-bills-main/insurance.csv")
# Read the insurance data into a data frame
insurance <- read.csv("D:/predict-hospital-bills-main/insurance.csv")

# Plot all numeric columns in the 'insurance' data frame
plot(insurance)


# Identify numeric columns
num_cols <- sapply(insurance, is.numeric)

# Extract and plot numeric columns
plot(insurance[, num_cols])

round(cor(df[,num_cols]),2)
# Read the insurance data into a data frame
insurance <- read.csv("D:/predict-hospital-bills-main/insurance.csv")

# Extract and plot numeric columns
```

6:26 (Top Level) R Script

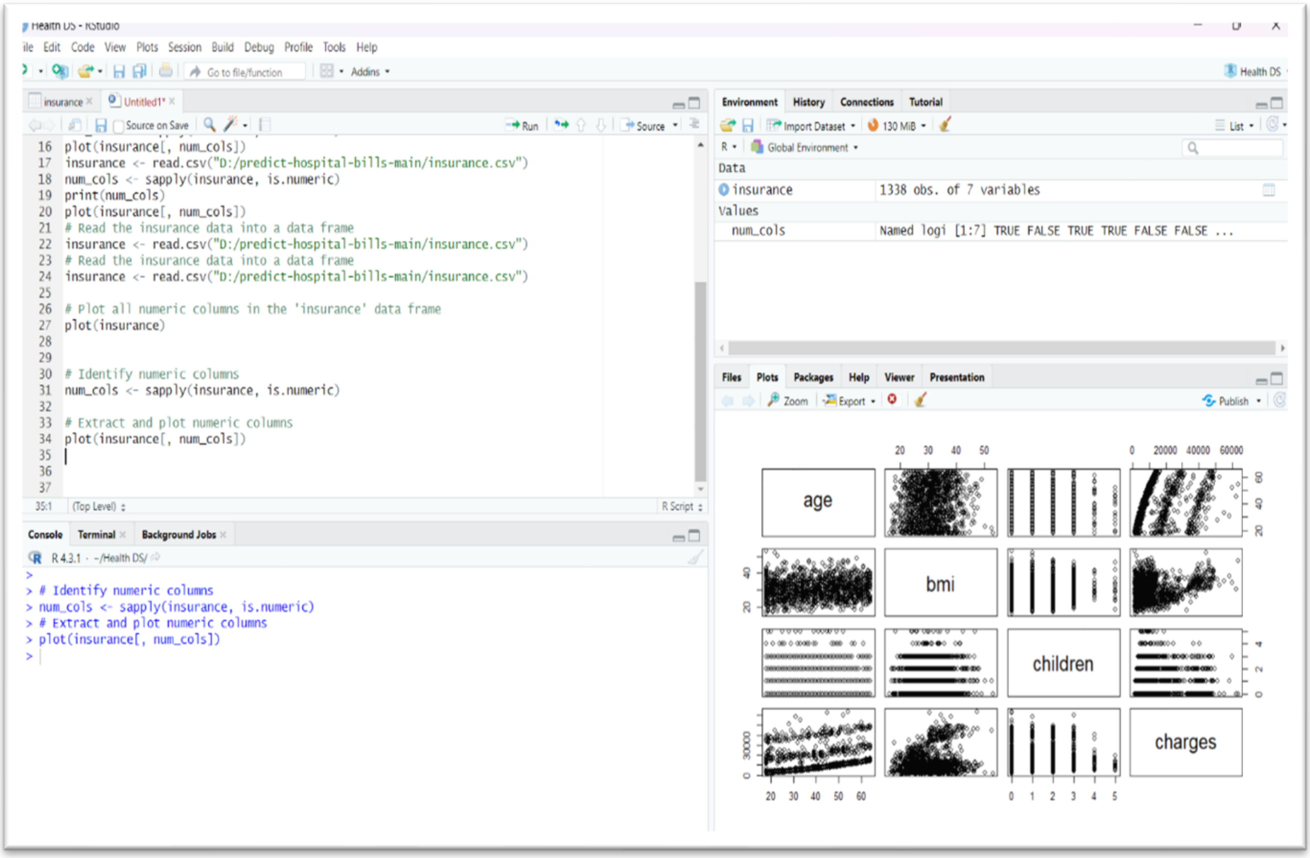
Console

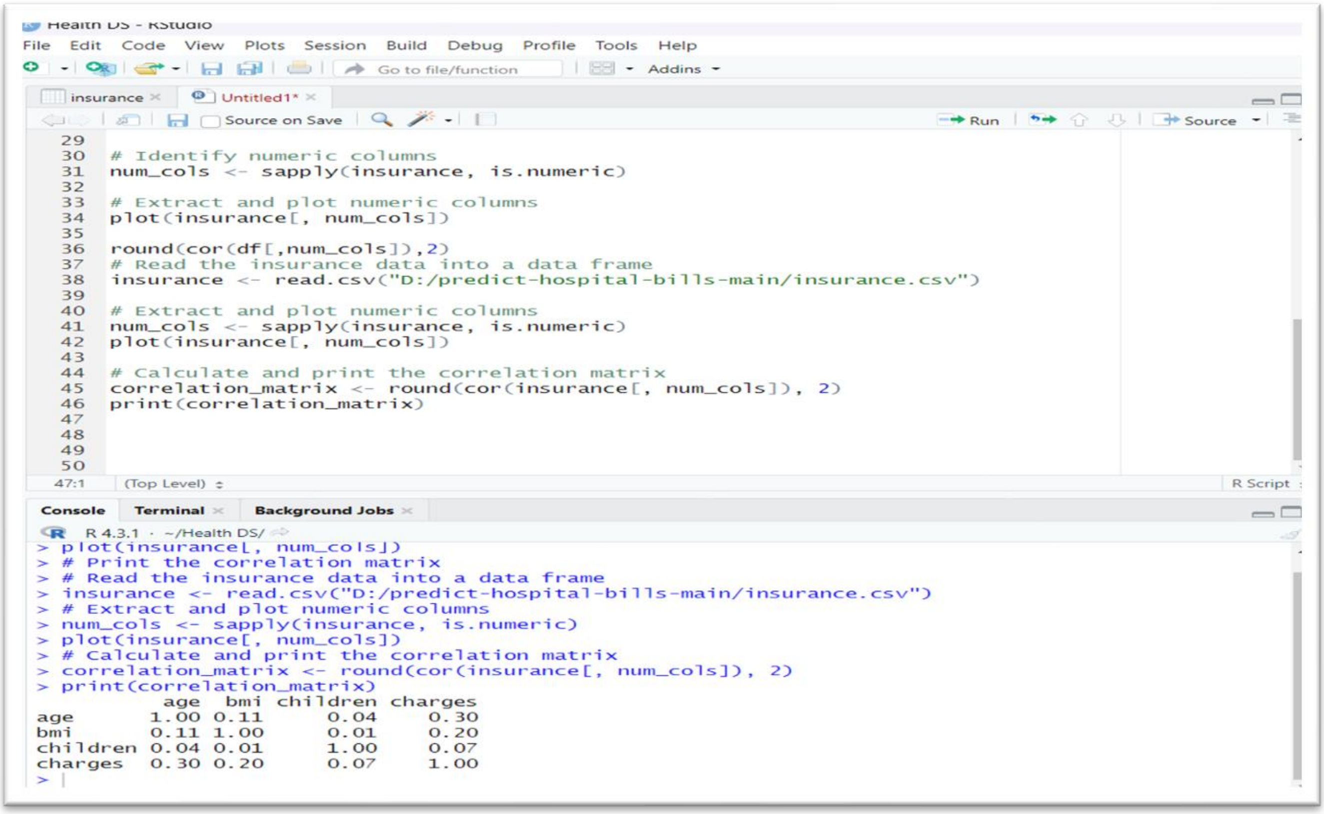
```
# Extract and plot numeric columns
num_cols <- sapply(insurance, is.numeric)
plot(insurance[, num_cols])

# Calculate and print the correlation matrix
correlation_matrix <- round(cor(insurance[, num_cols]), 2)
print(correlation_matrix)

smoker = as.factor(df$smoker)
sex = as.factor(df$sex)
region = as.factor(df$region)

boxplot(df$charges ~ smoker, main ='Smoker')
boxplot(df$charges ~ sex, main ='sex')
boxplot(df$charges ~ region, main ='region')
# Read the insurance data into a data frame
insurance <- read.csv("D:/predict-hospital-bills-main/insurance.csv")

# Convert the 'smoker' and 'sex' columns to factors
insurance$smoker <- as.factor(insurance$smoker)
insurance$sex <- as.factor(insurance$sex)
insurance$region <- as.factor(insurance$region)

# Create boxplots
boxplot(insurance$charges ~ insurance$smoker, main = 'Smoker')
boxplot(insurance$charges ~ insurance$sex, main = 'Sex')

model1 = lm(charges ~. , data =df)
# Read the insurance data into a data frame
insurance <- read.csv("D:/predict-hospital-bills-main/insurance.csv")

# Fit a linear regression model using all variables in the 'insurance' data frame
model1 <- lm(charges ~ ., data = insurance)

summary(model1)
```

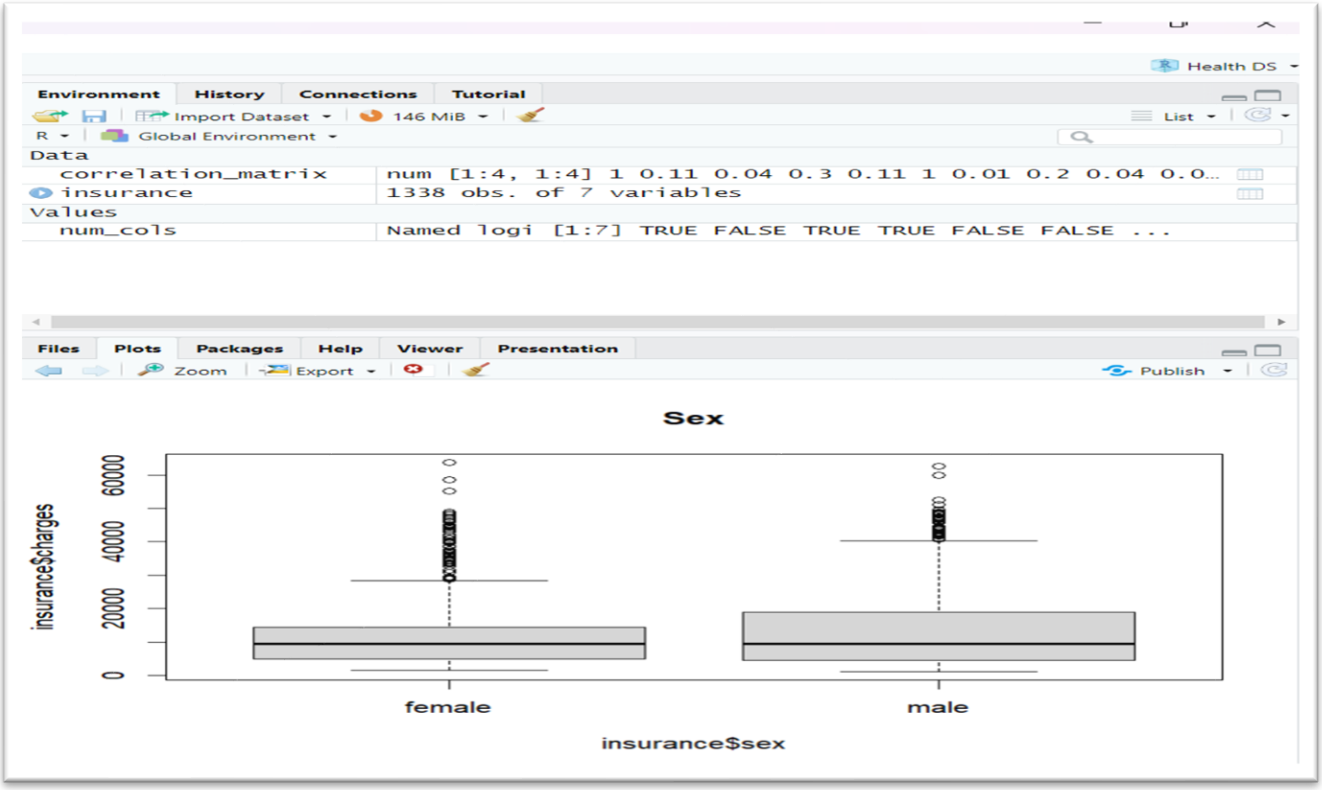
6:26 (Top Level) R Script

Console

OUTPUT


Health DS - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
insurance
Untitled1*
Source on Save
Run
Source
plot(insurance[, num_cols])
insurance <- read.csv("D:/predict-hospital-bills-main/insurance.csv")
num_cols <- sapply(insurance, is.numeric)
print(num_cols)
plot(insurance[, num_cols])
# Read the insurance data into a data frame
insurance <- read.csv("D:/predict-hospital-bills-main/insurance.csv")
# Read the insurance data into a data frame
insurance <- read.csv("D:/predict-hospital-bills-main/insurance.csv")
# Plot all numeric columns in the 'insurance' data frame
plot(insurance)
# Identify numeric columns
num_cols <- sapply(insurance, is.numeric)
# Extract and plot numeric columns
plot(insurance[, num_cols])
Console Terminal Background Jobs
R 4.3.1 · ~/Health DS/
> # Identify numeric columns
> num_cols <- sapply(insurance, is.numeric)
> # Extract and plot numeric columns
> plot(insurance[, num_cols])
Environment History Connections Tutorial
Import Dataset
130 MiB
Global Environment
Data
insurance 1338 obs. of 7 variables
Values
num_cols Named logi [1:7] TRUE FALSE TRUE TRUE FALSE FALSE ...
Files Plots Packages Help Viewer Presentation
Zoom
Export
Publish
age
bmi
children
charges


Health DS - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
insurance
Untitled1*
Source on Save
Run
Source
# Identify numeric columns
num_cols <- sapply(insurance, is.numeric)
# Extract and plot numeric columns
plot(insurance[, num_cols])
round(cor(df[,num_cols]),2)
# Read the insurance data into a data frame
insurance <- read.csv("D:/predict-hospital-bills-main/insurance.csv")
# Extract and plot numeric columns
num_cols <- sapply(insurance, is.numeric)
plot(insurance[, num_cols])
# Calculate and print the correlation matrix
correlation_matrix <- round(cor(insurance[, num_cols]), 2)
print(correlation_matrix)
(Top Level)
R Script
Console Terminal Background Jobs
R 4.3.1 · ~/Health DS/
> plot(insurance[, num_cols])
> # Print the correlation matrix
> # Read the insurance data into a data frame
> insurance <- read.csv("D:/predict-hospital-bills-main/insurance.csv")
> # Extract and plot numeric columns
> num_cols <- sapply(insurance, is.numeric)
> plot(insurance[, num_cols])
> # Calculate and print the correlation matrix
> correlation_matrix <- round(cor(insurance[, num_cols]), 2)
> print(correlation_matrix)
age bmi children charges
age 1.00 0.11 0.04 0.30
bmi 0.11 1.00 0.01 0.20
children 0.04 0.01 1.00 0.07
charges 0.30 0.20 0.07 1.00


Health DS
Environment
History
Connections
Tutorial
Import Dataset
146 MiB
List
R
Global Environment
Data
correlation_matrix
num [1:4, 1:4] 1 0.11 0.04 0.3 0.11 1 0.01 0.2 0.04 0.0...
insurance
1338 obs. of 7 variables
Values
num_cols
Named logi [1:7] TRUE FALSE TRUE TRUE FALSE FALSE ...
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Sex
insurance$charges
0
20000
40000
60000
female
male
insurance$sex

CONCLUSION

In conclusion, using R Studio for health data analysis provides a powerful and versatile tool for extracting valuable insights from complex healthcare datasets. The integration of R's vast libraries, along with the ability to generate publication-quality graphs and reports, makes it an essential resource in the pursuit of better health outcomes and medical research.