**PROBLEM\_2(SOLUTION):**

**A)**

***v= matrix(c(-4,1,7),nrow = 3,ncol = 1)***

***> v***

[,1]

[1,] -4

[2,] 1

[3,] 7

**w= matrix(c(0,4,20),nrow = 3,ncol = 1)**

**> w**

[,1]

[1,] 0

[2,] 4

[3,] 20

***scalar=t(v)%\*%w***

***> scalar***

[,1]

[1,] 144

**B.)**

***w*= *matrix(c(0,4,20),nrow = 3,ncol = 1)***

> w

[,1]

[1,] 0

[2,] 4

[3,] 20

***ans\_b=-3\*w***

***> ans\_b***

[,1]

[1,] 0

[2,] -12

[3,] -60

**C*.***

***M= matrix(c(8,3,0,21,42,3,0,34,11),nrow = 3,ncol = 3,byrow = T)***

***> M***

[,1] [,2] [,3]

[1,] 8 3 0

[2,] 21 42 3

[3,] 0 34 11

***v= matrix(c(-4,1,7),nrow = 3,ncol = 1)***

***> v***

[,1]

[1,] -4

[2,] 1

[3,] 7

***ans\_c=M%\*%v***

***> ans\_****c*

[,1]

[1,] -29

[2,] -21

[3,] 111

**D).**

***N= matrix(c(-6,-3,0,0,2,7,5,1,-8),nrow = 3,ncol = 3,byrow = T)***

***> N***

[,1] [,2] [,3]

[1,] -6 -3 0

[2,] 0 2 7

[3,] 5 1 -8

***M= matrix(c(8,3,0,21,42,3,0,34,11),nrow = 3,ncol = 3,byrow = T)***

***> M***

[,1] [,2] [,3]

[1,] 8 3 0

[2,] 21 42 3

[3,] 0 34 11

***ans\_d=M+N***

***> ans\_d***

[,1] [,2] [,3]

[1,] 2 0 0

[2,] 21 44 10

[3,] 5 35 3

**E)**

**ans\_e=M-N**

**> ans\_e**

[,1] [,2] [,3]

[1,] 14 6 0

[2,] 21 40 -4

[3,] -5 33 19

**F)**

***Z= matrix(c(1,-1,1,6,1,-2,1,0),nrow = 4,ncol = 2,byrow = T)***

***> Z***

[,1] [,2]

[1,] 1 -1

[2,] 1 6

[3,] 1 -2

[4,] 1 0

***> ans\_f=t(Z)%\*%Z***

***> ans\_f***

[,1] [,2]

[1,] 4 3

[2,] 3 41

**G)**

***ans\_g=solve(ans\_f)***

***> ans\_g***

[,1] [,2]

[1,] 0.26451613 -0.01935484

[2,] -0.01935484 0.02580645

**H)**

***Y= matrix(c(0,9,0,1),nrow = 4,ncol = 1,byrow = T)***

***> Y***

[,1]

[1,] 0

[2,] 9

[3,] 0

[4,] 1

***ans\_h=t(Z)%\*%Y***

***> ans\_h***

[,1]

[1,] 10

[2,] 54

**I)**

***B= ans\_g%\*%ans\_h***

***> B***

[,1]

[1,] 1.6

[2,] 1.2

**J)**

***ans\_j=det(ans\_f)***

***> ans\_j***

[1] 155

**PROBLEM\_3(SOLUTION):**

I read this article about a new technology that can help to predict the quality of wastewater. The idea is to use the data collected from sensors and other tools to make predictions on how much water will be used for drinking, irrigation and sanitation purposes. It is also possible to use these sensors and other tools to measure the quality in the wastewater treatment plants. This is an important step in improving our understanding of what is happening in the wastewater treatment plants. We have already started to understand how well they are working on this problem. It also says that about to detect some problems with the current system of wastewater treatment. This lack of real- time process variable information limits the effective operation of effluent water quality prediction .Furthermore, the online monitoring instrument needs high economic costs and is difficult to be conducted in the WWTP. Therefore, water quality prediction model is essential to support water quality parameters”.

"To overcome this problem, ridge regression echo state network (RESN) is proposed, in which the ridge regression algorithm is used to calculate the output weights instead of linear regression. Simulation results show that the RESN has better performance than some other existing methods, thus can deal with the ill-posed problem”. (pg-1) The data obtained from the simulations are compared with those obtained by a conventional method. In addition, the accuracy of the models is also tested and compared with the results obtained from traditional methods. In order to achieve the best possible result, they have chosen a simple model for our analysis .

"For the samples' data, the first 200 samples are used for training, 50 samples in training set are

discarded to washout initial transient, and the last 133 values are used to test the network performance. Before the simulation, the inputs and the target outputs are normalized into [-1, 1-]. simulations, the outputs are converted”(pg-6). The output is then analyzed using the following algorithm: This algorithm uses a combination of statistical techniques and statistical models to estimate the overall quality of wastewater treatment. Ridge regression is applied to determine the degree of quality improvement in wastewater treatment. The results show that the quality of wastewater treated by basins has been improved significantly over time.

To conclude The proposed RESN can solve the ill-posed problem by adding a penalty term to the loss function. The simulation results show that the proposed RESN model can get good performance and has a smaller testing error than other models.

**Work cites:**

Zhao, J., Zhao, C., Zhang, F., Wu, G., & Wang, H. (2018). Water Quality Prediction in the Waste Water Treatment Process Based on Ridge Regression Echo State Network. Retrieved April 10, 2021, from <https://iopscience.iop.org/article/10.1088/1757-899X/435/1/012025/meta>

**PROBLEM\_4(SOLUTION)**

This work is intended to provide a framework for ethical decision making in big data. The framework provides guidance on how to conduct research in big data, including the use of statistical methods and Research developed by a working group convened by the

Science," Health and Policy-relevant Ethics in Singapore (SHAPES) Initiative.it had a six domains where big data is currently employed: openness in big data and data repositories, precision medicine and big data, real-world data to generate evidence about healthcare interventions, AI-assisted decision-making in healthcare, public-private partnerships in healthcare and research, and cross-sectoral big data."(abstract) it had ethical framework for big data in health and policy has been developed with the goal of helping practitioners and researchers understand how big data can be used to improve patient outcomes, reduce costs and improve quality of care, and improve patient safety. This framework was developed by the national Institute of health's national center for health statistics and is designed to help practitioners identify the best practices in big data that can improve patient care. The framework is based on a set of principles that are applicable to all healthcare organizations. These principles include: the use of big data to inform decision making , the use of big data to support research , the importance of using large amounts of information in decision making and etc.

In this paper, we will examine how these three domains are integrated into the ethical framework for data collection and analysis of big data. it will also look at how they can be applied to other domains such as clinical trials, patient safety and quality assurance. it discuss how these domains can be used to improve the quality of life. "it did not intend to settle considerable debate in the literature concerning the precise meaning of concepts like justice or privacy, but rather sought definitions that were concise, comprehensible to non-expert stakeholders, and reflective of the core ethical concerns raised by big data in health and research"(applying the values). The paper concludes with a discussion of the implications for health data management and the role of big data in health care. In this paper, it explore the ethical issues surrounding big data and its use in health care.

**Work cites:**

Xafis, V., Schaefer, G. O., Labude, M. K., Brassington, I., Ballantyne, A., Lim, H. Y., . . . Tai, E. S. (2019). An ethics framework for big data in health and research. *Asian Bioethics Review,* *11*(3), 227-254. doi:10.1007/s41649-019-00099-x

**PROBLEM\_5(SOLUTION):**

**A)**

> #look a size of dataset by row and col

> dim(training\_values)

[1] 1338 10

* As we see that in given dataset we have 1338 row which is total number of observations in file that we looking for and we have 10 col which is data variables in our datasets.

> str(training\_values)

'data.frame': 1338 obs. of 10 variables:

$ age : int 19 18 28 33 32 31 46 37 37 60 ...

$ sex : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...

$ gender\_num: int 0 1 1 1 1 0 0 0 1 0 ...

$ bmi : num 27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...

$ children : int 0 1 3 0 0 0 1 3 2 0 ...

$ smoker : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...

$ smoker\_num: int 1 0 0 0 0 0 0 0 0 0 ...

$ region : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...

$ region\_num: int 4 3 3 2 2 3 3 2 1 2 ...

$ expenses : num 16885 1726 4449 21984 3867 ...

* We see that there are 10 variables but 4 numerical variables (age, bmi, children and expenses) and 3 nominal variables (sex, smoker and region) so variables which are nominal are create in dummy variables .

# Check for Missing Values

> sum(is.na(training\_values))

[1] 0

* We see that there is no missing values in our datasets

> #Create new subsets of data

> library(psych)

> training\_dataset2 <- training\_values[,c(1,3:5,7,9,10)]

> str(training\_dataset2)

'data.frame': 1338 obs. of 7 variables:

$ age : int 19 18 28 33 32 31 46 37 37 60 ...

$ gender\_num: int 0 1 1 1 1 0 0 0 1 0 ...

$ bmi : num 27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...

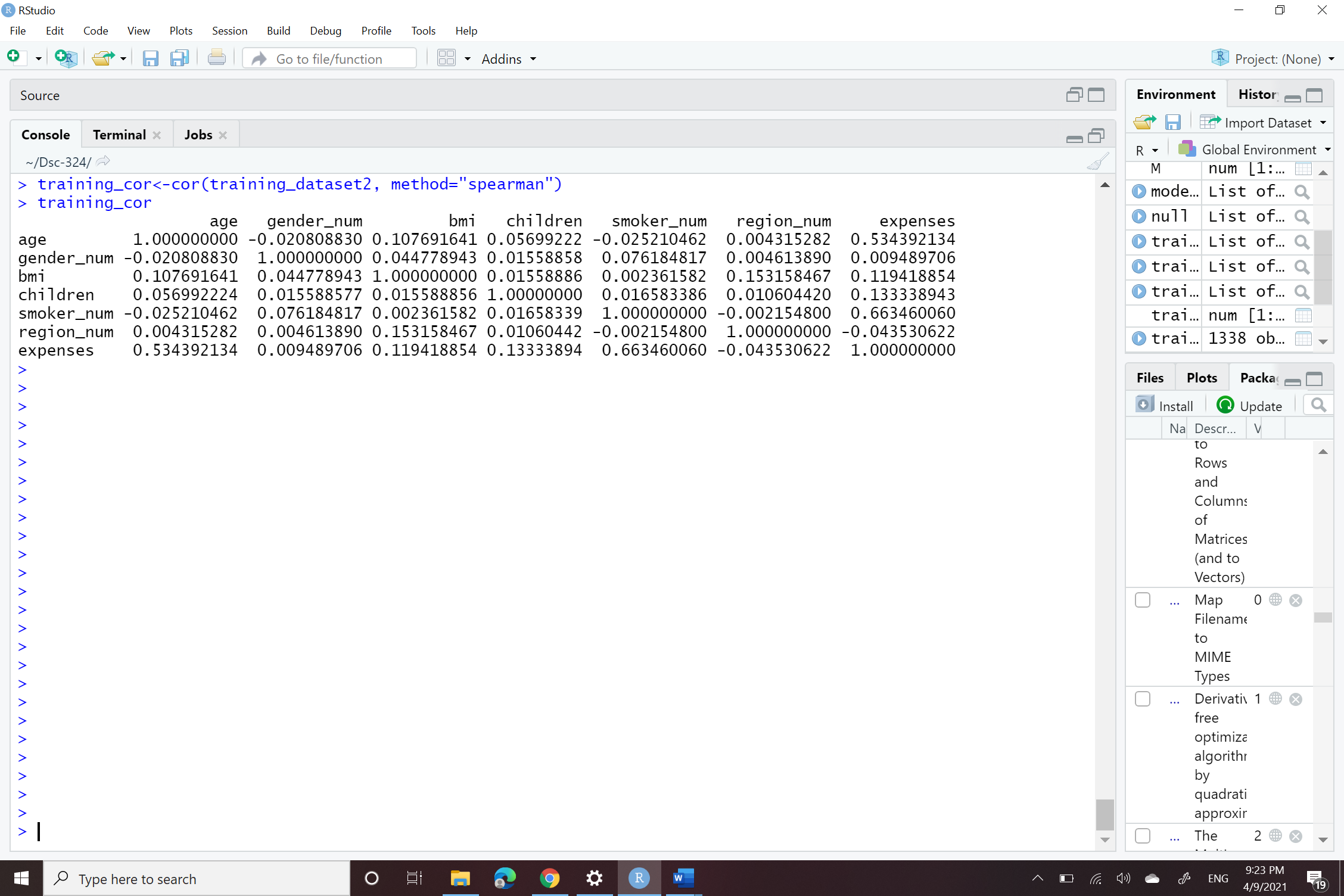
$ children : int 0 1 3 0 0 0 1 3 2 0 ...

$ smoker\_num: int 1 0 0 0 0 0 0 0 0 0 ...

$ region\_num: int 4 3 3 2 2 3 3 2 1 2 ...

$ expenses : num 16885 1726 4449 21984 3867 ...

* So we can see that our clean datasets that has 7 variables which 4 numerical variables (age, bmi, children and expenses) and 3 nominal variables (sex, smoker and region) so variables which are nominal are create in dummy variables .

Chart, scatter chart

Description automatically generated

* As we see that Correlation values conclusion with corr(expenses and smoker\_num)=0.6634, highest , the correlation values of expenses with smoker\_num is indicating that there is positive correlation then after corr(expenses and age)= 0.534. and Correlation values conclusion with corr(expenses and region\_num)= - 0.0435, very smallest the correlation values of expenses with region\_num.

> VIF(model1)

age gender\_num bmi children smoker\_num region\_num

1.015411 1.008888 1.040583 1.002481 1.006468 1.025925

* As see output of our VIF . there is a no problem of multicollinearity in our datasets because every value is less than 10 by compute the VIF it show that it have not collinearity. If the variables having multicollinearity higher than 10 that variables will not be included in the regression model.

**B)**

1. **)**

> summary(train\_Forward)

Call:

lm(formula = expenses ~ smoker\_num + age + bmi + children + region\_num,

data = training\_dataset2)

Residuals:

Min 1Q Median 3Q Max

-11402 -2808 -989 1392 29683

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -11516.14 973.93 -11.824 < 2e-16 \*\*\*

smoker\_num 23807.10 410.53 57.991 < 2e-16 \*\*\*

age 257.39 11.88 21.669 < 2e-16 \*\*\*

bmi 332.12 27.68 11.999 < 2e-16 \*\*\*

children 478.64 137.58 3.479 0.000519 \*\*\*

region\_num -353.29 151.87 -2.326 0.020152 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6058 on 1332 degrees of freedom

Multiple R-squared: 0.7507, Adjusted R-squared: 0.7498

F-statistic: 802.3 on 5 and 1332 DF, p-value: < 2.2e-16

> summary(train\_Backward)

Call:

lm(formula = expenses ~ age + bmi + children + smoker\_num + region\_num,

data = training\_dataset2)

Residuals:

Min 1Q Median 3Q Max

-11402 -2808 -989 1392 29683

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -11516.14 973.93 -11.824 < 2e-16 \*\*\*

age 257.39 11.88 21.669 < 2e-16 \*\*\*

bmi 332.12 27.68 11.999 < 2e-16 \*\*\*

children 478.64 137.58 3.479 0.000519 \*\*\*

smoker\_num 23807.10 410.53 57.991 < 2e-16 \*\*\*

region\_num -353.29 151.87 -2.326 0.020152 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6058 on 1332 degrees of freedom

Multiple R-squared: 0.7507, Adjusted R-squared: 0.7498

F-statistic: 802.3 on 5 and 1332 DF, p-value: < 2.2e-16

> summary(train\_Step)

Call:

lm(formula = expenses ~ smoker\_num + age + bmi + children + region\_num,

data = training\_dataset2)

Residuals:

Min 1Q Median 3Q Max

-11402 -2808 -989 1392 29683

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -11516.14 973.93 -11.824 < 2e-16 \*\*\*

smoker\_num 23807.10 410.53 57.991 < 2e-16 \*\*\*

age 257.39 11.88 21.669 < 2e-16 \*\*\*

bmi 332.12 27.68 11.999 < 2e-16 \*\*\*

children 478.64 137.58 3.479 0.000519 \*\*\*

region\_num -353.29 151.87 -2.326 0.020152 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6058 on 1332 degrees of freedom

Multiple R-squared: 0.7507, Adjusted R-squared: 0.7498

F-statistic: 802.3 on 5 and 1332 DF, p-value: < 2.2e-16

* Above three automatic regression selection techniques are applied BACKWARD, FORWARD and STEPWISE. All the three techniques give the same result.
* I will go with the forward selection technique because selecting one over other doesn’t make any difference.
* All the given variables are used in the model also all then passes the t-test.
* The adjusted R squared of the final model is 0.7498 means that 74.98% variability of the dependent variable is been explained by the independent variables.

1. **)**

* Boxplot of the dependent variable Expenses

**Chart, box and whisker chart

Description automatically generated**

* Gg-boxplot of the dependent variable Expenses

**Chart

Description automatically generated**

**> summary(training\_dataset2$expenses)**

**Min. 1st Qu. Median Mean 3rd Qu. Max.**

**1122 4740 9382 13270 16640 63770**

* As we see that in boxplot of depended variables expenses We can say that expenses is skewed to the right because mean > Median and we see it also has outliner.

**Code in R:**

**##---------------------------------------------------------------**

**#------------- Solution- 2-------------------------------------**

**#-------------A-----------------------------**

**v= matrix(c(-4,1,7),nrow = 3,ncol = 1)**

**v**

**w= matrix(c(0,4,20),nrow = 3,ncol = 1)**

**w**

**scalar=t(v)%\*%w**

**scalar**

**#---------------B-------------------------**

**w= matrix(c(0,4,20),nrow = 3,ncol = 1)**

**w**

**ans\_b=-3\*w**

**ans\_b**

**#-------------C----------------------------**

**M= matrix(c(8,3,0,21,42,3,0,34,11),nrow = 3,ncol = 3,byrow = T)**

**M**

**ans\_c=M%\*%v**

**ans\_c**

**#----------------D------------------------**

**N= matrix(c(-6,-3,0,0,2,7,5,1,-8),nrow = 3,ncol = 3,byrow = T)**

**N**

**ans\_d=M+N**

**ans\_d**

**#---------------E-----------------------**

**ans\_e=M-N**

**ans\_e**

**#-------------F----------------------------**

**Z= matrix(c(1,-1,1,6,1,-2,1,0),nrow = 4,ncol = 2,byrow = T)**

**Z**

**ans\_f=t(Z)%\*%Z**

**ans\_f**

**#--------------G------------------------**

**ans\_g=solve(ans\_f)**

**ans\_g**

**#------------------H-----------------------**

**Y= matrix(c(0,9,0,1),nrow = 4,ncol = 1,byrow = T)**

**Y**

**ans\_h=t(Z)%\*%Y**

**ans\_h**

**#----------------------I-----------------**

**B= ans\_g%\*%ans\_h**

**B**

**#------------------J-------------------**

**ans\_j=det(ans\_f)**

**ans\_j**

**##---------------------------------------------------------------**

**#------------- Solution- 5-------------------------------------**

**library(Hmisc) #Describe Function**

**library(psych) #Multiple Functions for Statistics and Multivariate Analysis**

**library(GGally) #ggpairs Function**

**library(ggplot2) #ggplot2 Functions**

**library(vioplot) #Violin Plot Function**

**library(corrplot) #Plot Correlations**

**library(DescTools) #VIF Function**

**library(leaps) #Best Set Linear Regression Functions**

**#set a dataset form file**

**setwd("~/Dsc-324")**

**training\_values <- read.csv(file="insurance\_dataset.csv", header=TRUE, sep=",")**

**#look a size of dataset by row and col**

**dim(training\_values)**

**# look a data is a num or category**

**str(training\_values)**

**# Show names of the variables in datasets**

**names(training\_values)**

**# looking a datasets table /Show for first 6 rows of data**

**head(training\_values)**

**# Check for Missing Values**

**sum(is.na(training\_values))**

**#Create new subsets of data**

**library(psych)**

**training\_dataset2 <- training\_values[,c(1,3:5,7,9,10)]**

**str(training\_dataset2)**

**head(training\_dataset2)**

**library(corrplot) #Plot Correlations**

**library(ggplot2)**

**#library(dplyr)**

**#Check for Multicollinearity with Correlations**

**training\_cor<-cor(training\_dataset2, method="spearman")**

**training\_cor**

**corrplot(training\_cor, method = "circle")**

**model1 <- lm( expenses~ ., data=training\_dataset2)**

**model1**

**library(DescTools)**

**VIF(model1)**

**#Creating Automatic Models**

**library(leaps)**

**null = lm( expenses~ 1, data=training\_dataset2)**

**null**

**full = lm(expenses ~ ., data=training\_dataset2)**

**full**

**#Forward Regression**

**train\_Forward = step(null, scope = list(lower=null, upper=full), direction="forward")**

**summary(train\_Forward)**

**#Backward Regression**

**train\_Backward = step(full, direction="backward")**

**summary(train\_Backward)**

**#Stepwise Regression**

**train\_Step = step(null, scope = list(upper=full), direction="both")**

**summary(train\_Step)**

**#Boxplots**

**boxplot(training\_dataset2$expenses, col = "blue", main="expenses",**

**ylab="expenses" )**

**summary(training\_dataset2$expenses)**

**#GG Boxplot**

**ggboxplot<-ggplot(training\_dataset2, aes(y=expenses)) +**

**geom\_boxplot(col="blue") +**

**labs(**

**title="expenses",**

**y="expenses(100K)")**

**ggboxplot**