FINAL SEMESTER PROJECT

TITLE – Medical Insurance Premium Prediction

BSc. Honours In Statistics

Batch : 2019-2022

Current Semester : 6

Registration No. - 544-1111-0469-19

Roll No. - 193544-21-0146

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I am indebted to number of person for helping me in the preparation of this project. Firstly, Dr. Tapan Kumar Poddar, Principal, Vivekananda College, University of Calcutta; without whose help I couldn't have been a part of this prestigious college. I owe a deep debt of gratitude to my supervisor Sri Nilkanta Mukherjee for necessary guidance, for this presentation of this dissertation, valuable comments and suggestions. I am extremely grateful to him for the necessary stimulus, support and valuable time. Special thanks to Prof. Sri Nilkanta Mukherjee, Head of the Department of Statistics, Vivekananda College. I am greatly indebted to Smt. Sutapa Biswas, Riddhi Das Majumdar (Faculty members) often took pains and stood by me in adverse circumstances. Without their encouragement and inspiration it was not possible for me to complete this project. Finally my earnest thanks go to my friends who were always beside me when I needed them without any excuses and made these three years worthwhile. This project is not only a mere project. It is the memories spend with the whole department which has created a mutual understanding among us. There are many emotions related to this piece of work, especially respect and duty towards teachers and vice versa; educational attachment with my friends; social attachment with my college.

Department of statistics

Vivekananda College, C.U.

ABSTRACT

❖ Introduction:

Health insurance or medical insurance (also known as medical aid in South Africa) is a type of insurance that covers the whole or a part of the risk of a person incurring medical expenses. As with other types of insurance is risk among many individuals. By estimating the overall risk of health risk and health system expenses over the risk pool, an insurer can develop a routine finance structure, such as a monthly premium or payroll tax, to provide the money to pay for the health care benefits specified in the insurance agreement. The benefit is administered by a central organization, such as a government agency, private business, or not for profit entity. According to the Health Insurance Association of America, health insurance is defined as "coverage that provides for the payments of benefits as a result of sickness or injury. It includes insurance for losses from accident, medical expense, disability, or accidental death and dismemberment"

A health insurance policy is:

1. A contract between an insurance provider (e.g. an insurance company or a government) and an individual or his/her sponsor (that is an employer or a community organization). The contract can be renewable (annually, monthly) or lifelong in the case of private insurance. It can also be mandatory for all citizens in the case of national plans. The type and amount of health care costs that will be covered by the health insurance provider are specified in writing, in a member contract or "Evidence of Coverage" booklet for private insurance, or in a national [health policy] for public insurance.

2. (US specific) In the U.S., there are two types of health insurance - tax payer-funded and private-funded. An example of a private-funded insurance plan is an employer sponsored selffunded ERISA plan. The company generally advertises that they have one of the big insurance companies. However, in an ERISA case, that insurance company "doesn't engage in the act of insurance", they just administer it. Therefore, ERISA plans are not subject to state laws. ERISA plans are governed by federal law under the jurisdiction of the US Department of Labor (USDOL). The specific benefits or coverage details are found in the Summary Plan Description (SPD). An appeal must go through the insurance company, then to the Employer's Plan Fiduciary. If still required, the Fiduciary's decision can be brought to the USDOL to review for ERISA compliance, and then file a lawsuit in federal court.

❖ Data source:

The link of the data set is given here
https://www.kaggle.com/tejashvi14/medical-insurance-premium-prediction

❖ Software used: RStudio version 1.4.110

❖ Objective:

The major focus of this study is to estimate the premium price due to various factors in the population. The premium price is dependent on various factors like age, diabetes, blood pressure problems, any chronic diseases, any transplants, height, weight, etc. In this study, we aimed to predict changes in the premium price for the clients based on certain features such as clients' health insurance costs and to identify factors contributing substantially to this prediction. Linear Regression was applied to as Age, diabetes, blood pressure problems, any chronic diseases, any transplants, height, weight, etc.

❖ Data Handling:

A Medical Insurance Company Has Released Data For Almost 1000 Customers. Create A Model That Predicts The Yearly Medical Cover Cost. The Data Is Voluntarily Given By Customers.

The dataset contains health related parameters of the customers. We will use them to build a model and also perform EDA on the same.

The Premium Price Is In INR(₹) Currency And Showcases Prices For A Whole Year.

❖ Methodology:

I. Definition
II. Exploring the dataset
III. Data Visualization (**various kinds of plots**)
IV. Analysis of data
V. Hypothesis testing
VI. Calculating Multiple Correlation Coefficient (**linear and polynomial**)
VII. Prediction using Multiple Regression Model (**linear and polynomial**)
VIII. Conclusion

❖ Variables:

I. Age - Age Of Customer
II. Diabetes - Whether The Person Has Abnormal BloodSugar Levels
III. BloodPressureProblems - Whether The Person Has Abnormal Blood Pressure Levels
IV. AnyTransplants - Any Major Organ Transplants
V. AnyChronicDiseases - Whether Customer Suffers From Chronic Ailments Like Asthama, Etc.
VI. Height - Height Of Customer
VII. Weight - Weight Of Customer
VIII. KnownAllergies - Whether The Customer Has Any Known Allergies
IX. HistoryOfCancerInFamily - Whether Any Blood Relative Of The Customer Has Had Any Form Of Cancer
X. NumberOfMajorSurgeries - The Number Of Major Surgeries That The Person Has Had
XI. PremiumPrice - Yearly Premium Price

ANALYSIS OF DATA

❖ INSTALLING NEEDED PACKAGES FOR ANALYSIS:

CODES AND OUTPUTS-

```
install.packages('tidyverse') #for data manipulation
install.packages('dummies') #for creating dummy variables
install.packages('ggpubr') #would be used to arrange plots on a single page with ggarange
install.packages('caTools') #For randomly splitting data into test and train
install.packages('rpart') #for creating a decision tree model
install.packages('rpart.plot') #for plotting the decision tree
install.packages('ggplot2') #for boxplot and necessary diagrams


library(tidyverse)
library(dummies)
library(ggpubr)
library(caTools)
library(rpart)
library(rpart.plot)
library(ggplot2)
```

❖ GETTING OUR DATA AND HAVING A GLIMPSE OF IT:

CODES AND OUTPUTS-

```
> #Getting our data
>Data=read.csv("C:/Users/RANIT/OneDrive/Desktop/RANITPROJECT/Medicalpremium.csv",header=
T)
#having a glimpse of the data
>glimpse(Data)
```

```
Rows: 986
Columns: 11
$ Age                  <int> 45, 60, 36, 52, 38, 30, 33, 23, 48, 38, 60, 66~
$ Diabetes             <int> 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0~
$ BloodPressureProblems   <int> 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0~
$ AnyTransplants       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0~
$ AnyChronicDiseases     <int> 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ Height               <int> 155, 180, 158, 183, 166, 160, 150, 181, 169, 1~
$ Weight               <int> 57, 73, 59, 93, 88, 69, 54, 79, 74, 93, 74, 67~
$ KnownAllergies       <int> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1~
$ HistoryOfCancerInFamily <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ NumberOfMajorSurgeries  <int> 0, 0, 1, 2, 1, 1, 0, 0, 0, 0, 2, 0, 1, 0, 1, 1~
$ PremiumPrice         <int> 25000, 29000, 23000, 28000, 23000, 23000, 2100~
"There are 986 observations in the dataset. This tallies with the metadata of the dataset"

> View(Data)
> summary(Data)
     Age         Diabetes      BloodPressureProblems AnyTransplants
 Min.   :18.00   Min.   :0.0000  Min.   :0.0000       Min.   :0.00000
 1st Qu.:30.00   1st Qu.:0.0000  1st Qu.:0.0000       1st Qu.:0.00000
 Median :42.00   Median :0.0000   Median :0.0000       Median :0.00000
 Mean   :41.75   Mean   :0.4199  Mean   :0.4686       Mean   :0.05578
 3rd Qu.:53.00   3rd Qu.:1.0000   3rd Qu.:1.0000       3rd Qu.:0.00000
 Max.   :66.00   Max.   :1.0000  Max.   :1.0000       Max.   :1.00000
 AnyChronicDiseases     Height         Weight      KnownAllergies
 Min.   :0.0000     Min.   :145.0   Min.   : 51.00   Min.   :0.000
 1st Qu.:0.0000     1st Qu.:161.0   1st Qu.: 67.00   1st Qu.:0.000
 Median :0.0000     Median :168.0   Median : 75.00   Median :0.000
 Mean   :0.1805     Mean   :168.2   Mean   : 76.95   Mean   :0.215
 3rd Qu.:0.0000     3rd Qu.:176.0   3rd Qu.: 87.00   3rd Qu.:0.000
```

Max. :1.0000 Max. :188.0 Max. :132.00 Max. :1.000

HistoryOfCancerInFamily NumberOfMajorSurgeries PremiumPrice

Min. :0.0000 Min. :0.0000 Min. :15000

1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:21000

Median :0.0000 Median :1.0000 Median :23000

Mean :0.1176 Mean :0.6673 Mean :24337

3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:28000

Max. :1.0000 Max. :3.0000 Max. :40000

• Observations from Summary:

CODES AND OUTPUTS-

➢ Resting BP and Cholesterol has zero as minimum which is unusual.

➢ There may be outliers/Massings in Cholesterol and Resting BP being presented as zero.

➢ The interquartile range of Max HR suggests that there may be outliers from Min to 1st quartile.

> cor(Data)

Age Diabetes BloodPressureProblems

Age 1.000000000 0.21090797 0.24488846

Diabetes 0.210907965 1.00000000 0.12772652

BloodPressureProblems 0.244888460 0.12772652 1.00000000

AnyTransplants -0.008549118 -0.03665192 -0.02453793

AnyChronicDiseases 0.051071698 -0.08942838 0.04542434

Height 0.039879419 -0.00378252 -0.03792591

Weight -0.018590495 -0.02456310 -0.06101603

KnownAllergies -0.024415906 -0.08010240 -0.01154995

HistoryOfCancerInFamily -0.027623152 -0.05552683 0.04823876

NumberOfMajorSurgeries 0.429181489 0.12272247 0.25156767

PremiumPrice 0.697539966 0.07620924 0.16709675

AnyTransplants AnyChronicDiseases Height

Age -0.008549118 0.051071698 0.03987942

Diabetes -0.036651925 -0.089428377 -0.00378252

BloodPressureProblems -0.024537927 0.045424341 -0.03792591

AnyTransplants 1.000000000 0.035284503 -0.03154273

AnyChronicDiseases 0.035284503 1.000000000 0.04741889

Height -0.031542729 0.047418892 1.00000000

Weight 0.002086965 -0.033317642 0.06694566

KnownAllergies 0.001876436 -0.027417905 -0.01019988

HistoryOfCancerInFamily -0.020170803 0.008665551 0.01054894

NumberOfMajorSurgeries -0.004153805 0.014835207 0.03728857

PremiumPrice 0.289055937 0.208609860 0.02690951

Weight KnownAllergies HistoryOfCancerInFamily

Age -0.018590495 -0.024415906 -0.027623152

Diabetes -0.024563103 -0.080102401 -0.055526833

BloodPressureProblems -0.061016030 -0.011549955 0.048238755

AnyTransplants 0.002086965 0.001876436 -0.020170803

AnyChronicDiseases -0.033317642 -0.027417905 0.008665551

Height 0.066945661 -0.010199878 0.010548935

Weight 1.000000000 0.037491804 0.003480510

KnownAllergies 0.037491804 1.000000000 0.115382761

HistoryOfCancerInFamily 0.003480510 0.115382761 1.000000000

NumberOfMajorSurgeries -0.006108025 0.103923397 0.212657455

PremiumPrice 0.141507405 0.012102791 0.083139417

NumberOfMajorSurgeries PremiumPrice

Age 0.429181489 0.69753997

Diabetes 0.122722474 0.07620924

BloodPressureProblems 0.251567674 0.16709675

AnyTransplants -0.004153805 0.28905594

AnyChronicDiseases 0.014835207 0.20860986

Height 0.037288570 0.02690951

Weight -0.006108025 0.14150741

KnownAllergies 0.103923397 0.01210279

HistoryOfCancerInFamily 0.212657455 0.08313942

NumberOfMajorSurgeries 1.000000000 0.26424953

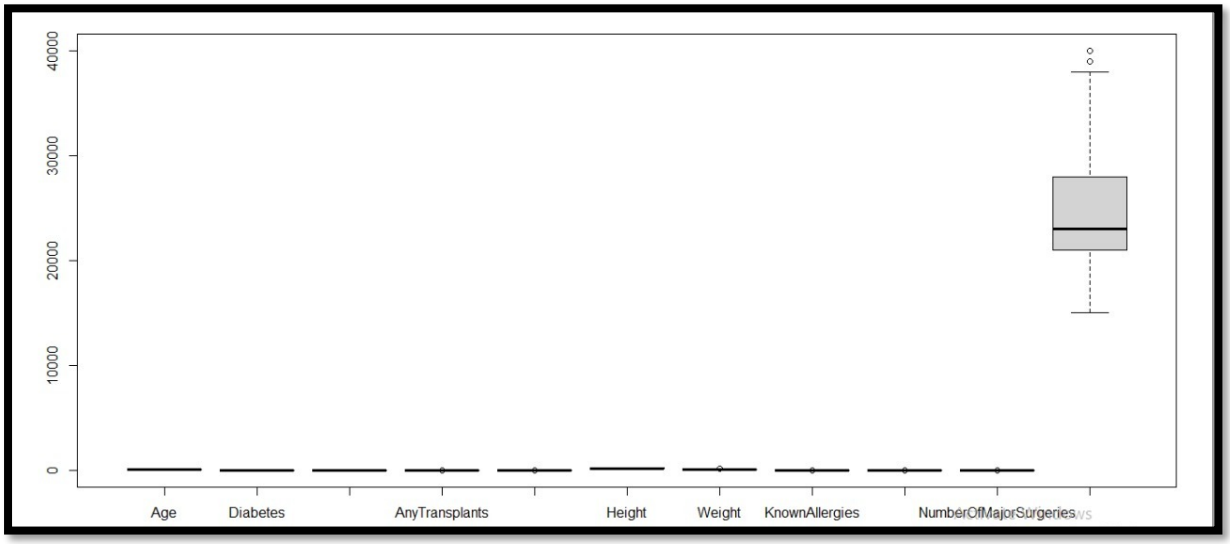
PremiumPrice 0.264249529 1.00000000

• CHECKING FOR OUTLIERS:

CODES AND OUTPUTS-

> #Boxplot

> boxplot(Data)



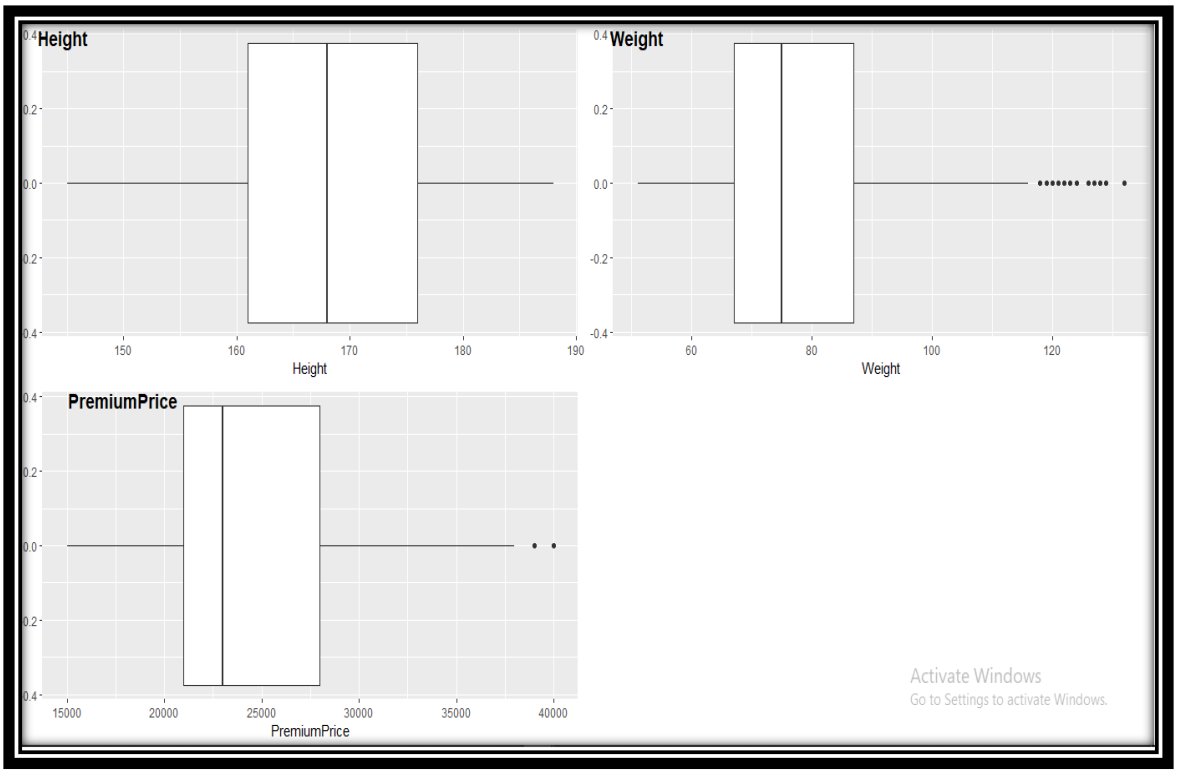
(Figure – 0.1.a)

> heightout<- ggplot(data = Data, aes(x = Height))+geom_boxplot()

> weightout <- ggplot(data = Data, aes(x = Weight))+geom_boxplot()

> premiumpriceout <- ggplot(data = Data, aes(x = PremiumPrice))+geom_boxplot()

>ggarrange(heightout, weightout,labels = , premiumpriceout c("Height", "Weight","PremiumPrice"))



(Figure – 0.1.b)

```
> boxplot(Data$Weight)
> boxplot(Data$Weight, plot= FALSE)$out
 [1] 118 121 119 129 127 132 120 128 120 123 126 121 118 128 124 122
> boxplot(Data$PremiumPrice)
> boxplot(Data$PremiumPrice, plot= FALSE)$out
[1]          39000          40000          39000          39000          39000          39000
```

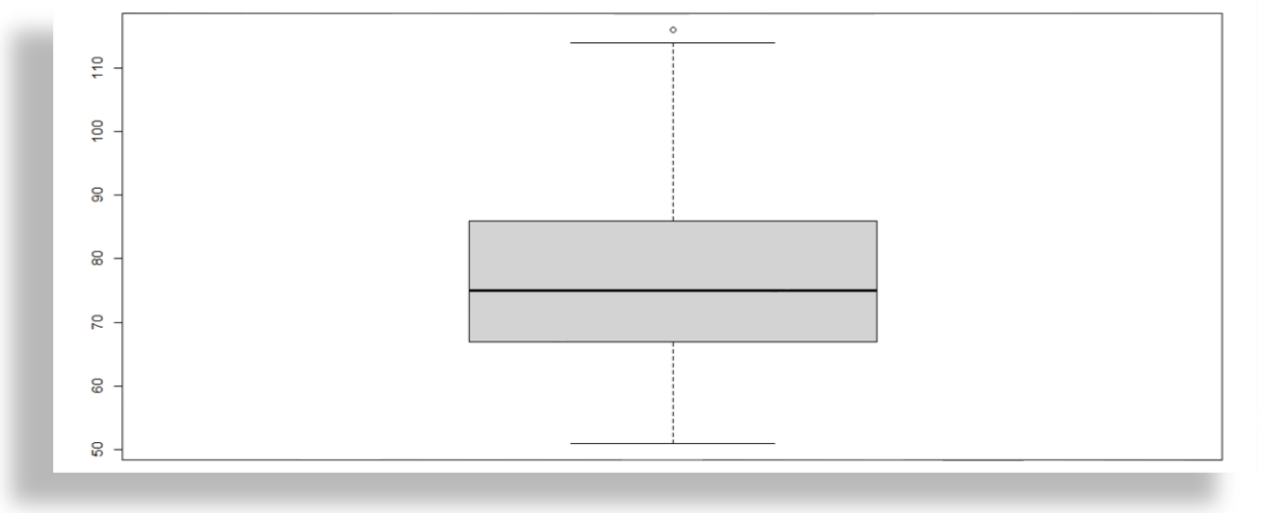
❖ DATA CLEANING:

➢ *Removing outliers:*

CODES AND OUTPUTS-

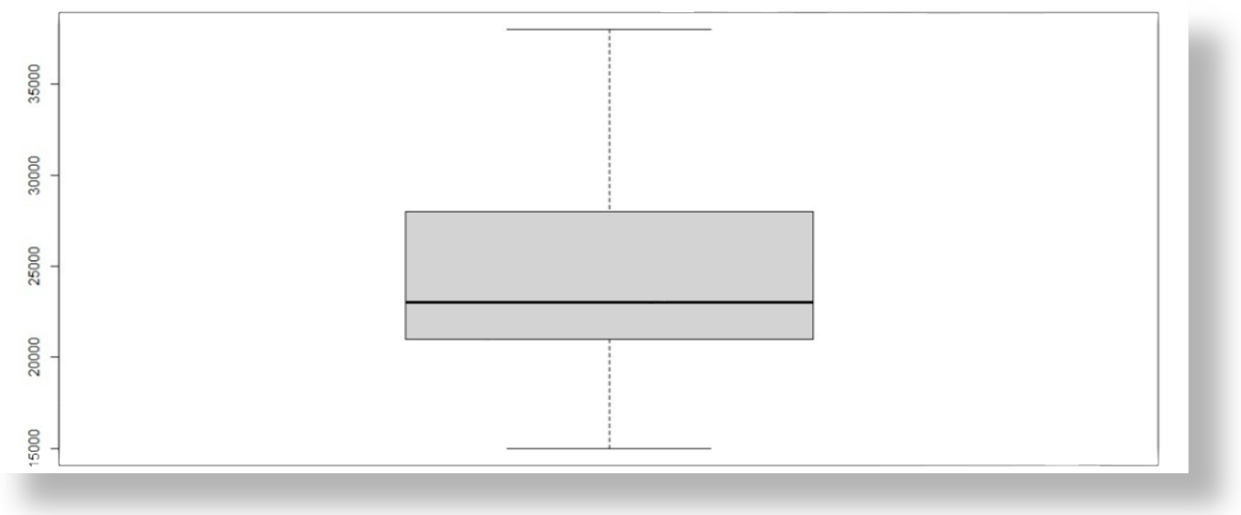
```
> #remoivng the outlier
> outlier <- boxplot(Data$Weight,plot= FALSE)$out
> outlier
 [1] 118 121 119 129 127 132 120 128 120 123 126 121 118 128 124 122
```

```
> newData <- Data[-which(Data$Weight %in% outlier),]
> summary(newData$Weight)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  51.00   67.00   75.00   76.18   86.00  116.00
> boxplot(newData$Weight)
```



(Figure – 0.1.c)

```
> outlier <- boxplot(Data$PremiumPrice,plot= FALSE)$out
> newData3 <- Data[-which(Data$PremiumPrice %in% outlier),]
> summary(newData3$PremiumPrice)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15000   21000   23000   24246   28000   38000
> boxplot(newData3$PremiumPrice)
```



(Figure – 0.1.d)

➢ Checking Missing Values:

CODES AND OUTPUTS-

Checking missing Values

table(is.na(Data))

FALSE

10846

➢ Checking if there are any zero values in our data:

CODES AND OUTPUTS-

Our business knowledge of the dataset suggests that zero values in age , height , weight and premium price data are illogical , so we need to identify the zero values in our data if exist.

> is.null(Data)

[1] FALSE

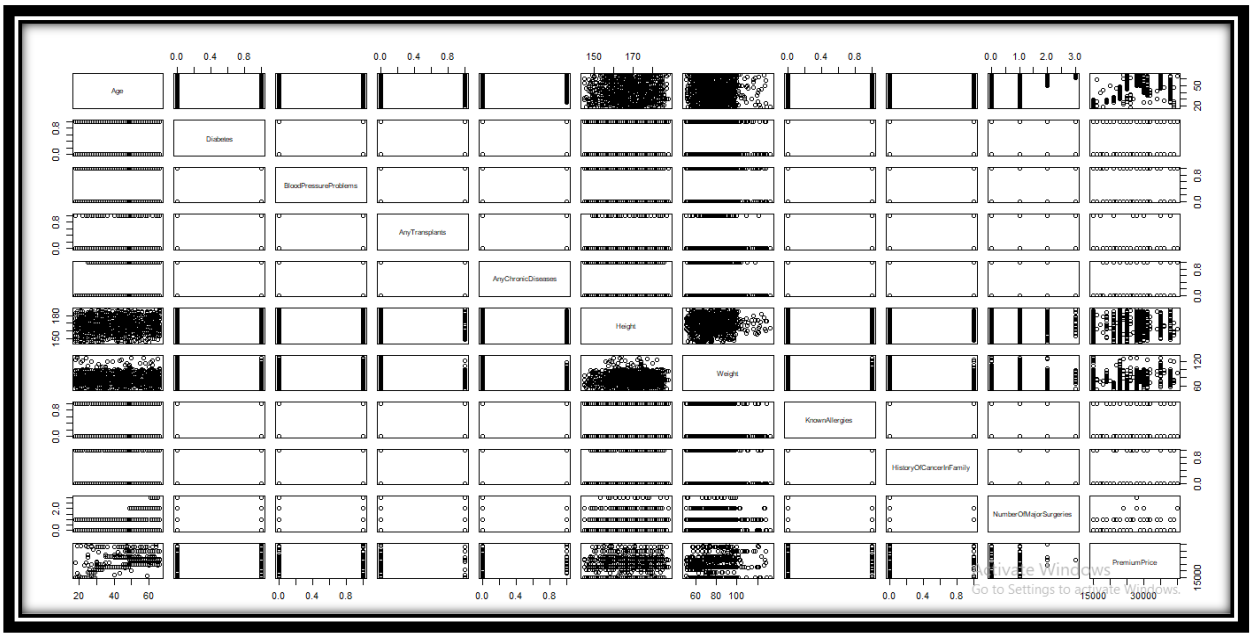
❖ VISUALIZATIONS AND DATA EXPLORATIONS:

This section contains visualizations from the cleaned dataset.

➢ Structure of the data:

CODES AND OUTPUTS-

> plot(Data)



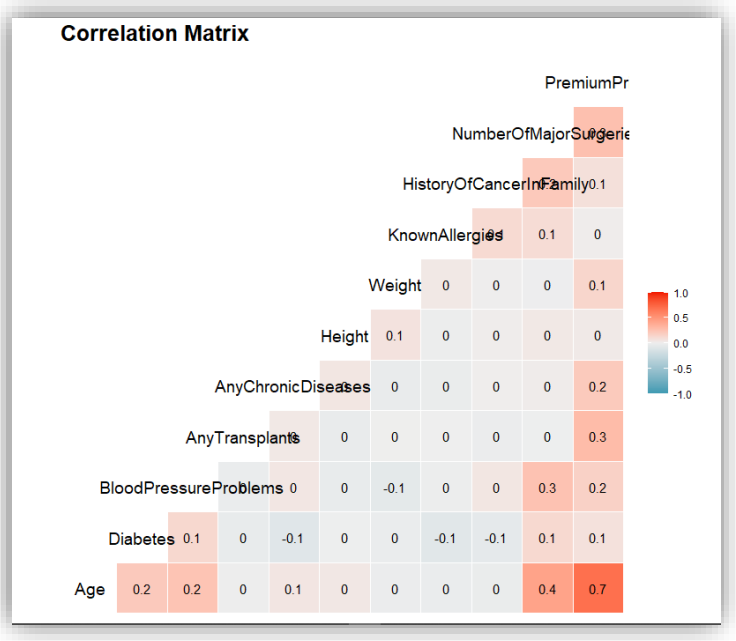
(Figure – 0.2)

➢*Correlation matrix:*

CODES AND OUTPUTS-

> #Creating the correlation matrix

> ggcorr(Data, label = T, color = "black", size = 5)+labs(title = "Correlation Matrix")+theme(plot.title = element_text(family = "Roboto Condensed", size = 19, face = "bold",vjust = 1),plot.subtitle = element_text(family = "Roboto Condensed", size = 16,vjust = 0))



(Figure – 0.3)

It can be easily seen here that factor affecting Premium Price are Age, Number of Major Surgery, Any Transplant, Any Chronic Disease, Blood Pressure Problems, History of Cancer in Family, Diabeties, Weight. While factors that hardly matters are height and Known Allergies.

But it should be noted that correlation between factor might not neccessary mean a causation.

FINDINGS-

It can be easily seen here that factor affecting Premium Price are Age, Number of Major Surgery, Any Transplant, Any Chronic Disease, Blood Pressure Problems, History of Cancer in Family, Diabeties, Weight. While factors that hardly matters are height and Known Allergies. But it should be noted that correlation between factor might not neccessary mean a causation.

From the correlation matrix we can conclude that all the independent variables have a **positive(+ve)** relationship / correlation with the response variable i.e. “Age” , “Diabetes” , “BloodPressureProblems” , “AnyTransplants”, “AnyChronicDiseases” , “Height” , “Weight” , “KnownAllergies” , “HistoryOfCancerInFamily” & “ NumberOfMajorSurgeries” have a positive correlation with the dependent variable “PremiumPrice”. Here “Age” has the **highest correlation** coefficient with “PremiumPrice” which is **equal to o.7** and , “Height” & “KnownAllergies” have **the lowest correlation coefficient** with “PremiumPrice” which are **equal to 0.**

■Applying factors to different columns

CODES AND OUTPUTS-

```
> Data$Diabetes <- as.factor(Data$Diabetes)
> Data$BloodPressureProblems <- as.factor(Data$BloodPressureProblems)
> Data$AnyTransplants <- as.factor(Data$AnyTransplants)
> Data$AnyChronicDiseases <- as.factor(Data$AnyChronicDiseases)
> Data$KnownAllergies <- as.factor(Data$KnownAllergies)
> Data$HistoryOfCancerInFamily <- as.factor(Data$HistoryOfCancerInFamily)
> Data$NumberOfMajorSurgeries <- as.factor(Data$NumberOfMajorSurgeries)

> str(Data)
'data.frame':   986 obs. of  11 variables:
 $ Age                 : int  45 60 36 52 38 30 33 23 48 38 ...
 $ Diabetes            : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 1 ...
 $ BloodPressureProblems  : Factor w/ 2 levels "0","1": 1 1 2 2 1 1 1 1 1 1 ...
 $ AnyTransplants      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ AnyChronicDiseases  : Factor w/ 2 levels "0","1": 1 1 1 2 2 1 1 1 1 1 ...
 $ Height              : int  155 180 158 183 166 160 150 181 169 182 ...
 $ Weight              : int  57 73 59 93 88 69 54 79 74 93 ...
 $ KnownAllergies      : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 2 2 1 ...
 $ HistoryOfCancerInFamily: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ NumberOfMajorSurgeries : Factor w/ 4 levels "0","1","2","3": 1 1 2 3 2 2 1 1 1 1 ...
 $ PremiumPrice        : int  25000 29000 23000 28000 23000 23000 21000 15000 23000 23000 ...
```

➢ ***Calculatting BMI:***

CODES AND OUTPUTS-

```
> #Calculatting BMI
> Data$bmi <- 10000*(Data$Weight/(Data$Height)^2)
```

➢ ***Assigning different categories to different BMI ranges:***

CODES AND OUTPUTS-

```
> Data <- data %>%mutate( bmiCategory = case_when(
                        bmi<18.49999 ~ "under weight",
                         bmi>18.5 & bmi<24.99999 ~ "normal weight",
                         bmi>25 & bmi<29.99999 ~ "over weight",
                         bmi>30 ~ "obesity"))
```

Note that 0 means the absence of an attribute while 1 shows it's presence

• Creating histogram for distribution of Age

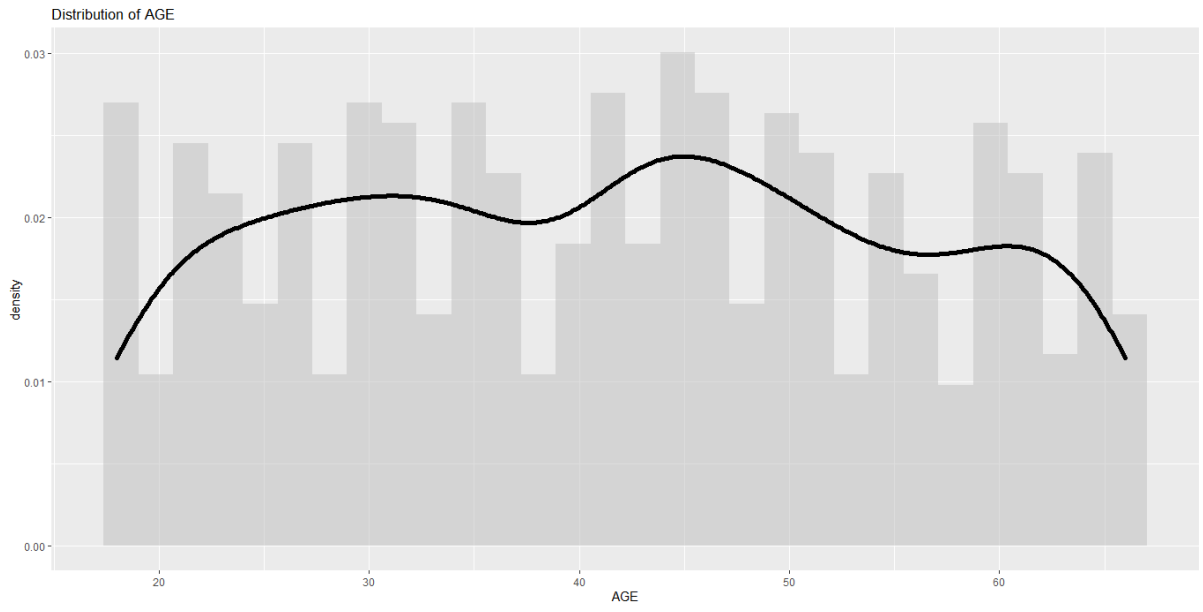
CODES AND OUTPUTS-

 # ASSIGNING THE 'AGE' VALUES TO A VARIABLE

```
AGE = Data$Age
```

 # CREATING HISTOGRAM FOR DISTRIBUTION OF AGE

```
ggplot(data.frame(Data$Age), aes(x=AGE)) + geom_histogram(aes(y=..density..),
fill="grey",alpha=0.5)+ geom_density(alpha=.2,col="black",size=2) + labs(title="Distribution of AGE")
```



(Figure – 1.a)

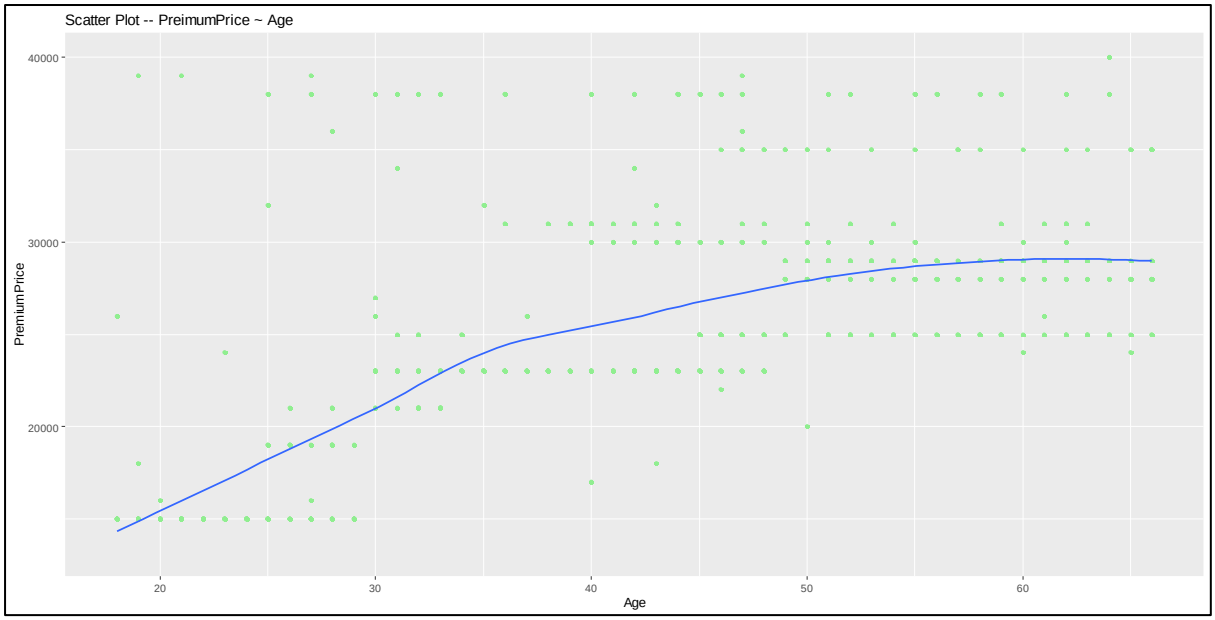
FINDINGS-

From the Histogram for 'Age',
(1) We can see that the density of the age represents that maximum people have age around 20 – 60 and very few people have age around 10, similarly very few people have age around 70.
(2) We can see that the distribution of age with respect to premium price is a comb distribution, the bars are alternately tall and short & it is multimodal.
(3) The distribution of age with respect to premium price is left skewed i.e. mean < median.

SCATTER PLOT BETWEEN AGE AND PREMIUMPRICE

CODES AND OUTPUTS-

```
> ggplot(train,aes(x=Age,y=PremiumPrice))+geom_point(col="light
green")+geom_smooth(method="auto", se=TRUE, fullrange=FALSE, level=0.95)+labs(title="Scatter
Plot -- PreimumPrice ~ Age")
```



(Figure – 1.b)

FINDINGS-

Scatterplots display the direction, strength, and linearity of the relationship between two variables.
(1) From the above diagram we can see that the relationship between age and premium price is positive i.e. they are positively correlated.

(2) Stronger relationships produce a tighter clustering of data points. From this diagram as the data points don't cluster that much tightly so they provide a moderately strong relationship.

CREATING DENSITY PLOT FOR DISTRIBUTION OF DIFFERENT AGES

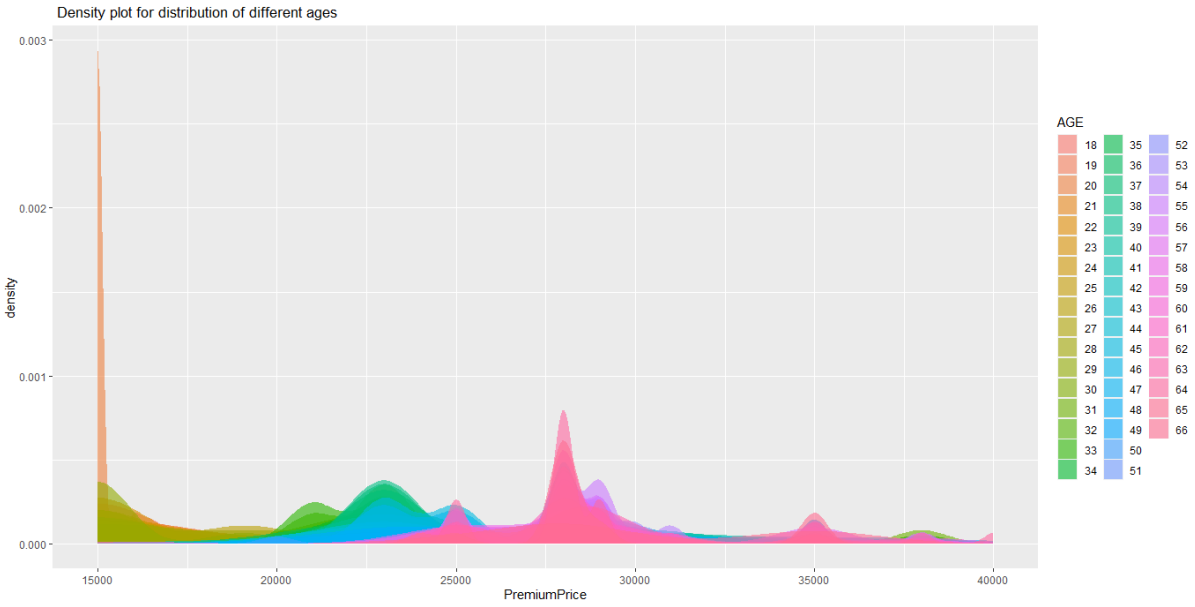
CODES AND OUTPUTS-

Applying factors to 'Age' column

```
AGE <- as.factor(Data$Age)
```

#Density plot

```
ggplot(Updated_Data,aes(PremiumPrice))+geom_density(aes(fill=AGE),color=NA,
alpha=0.6)+labs(title = " Density plot for distribution of different ages ")
```



(Figure – 1.c)

• Diabetics People Premium Analysis

➢ *Pie chart for Diabetes:*

CREATING FREQUENCY TABLE WITH RESPECT TO THE 'DIABETES' COLUMN

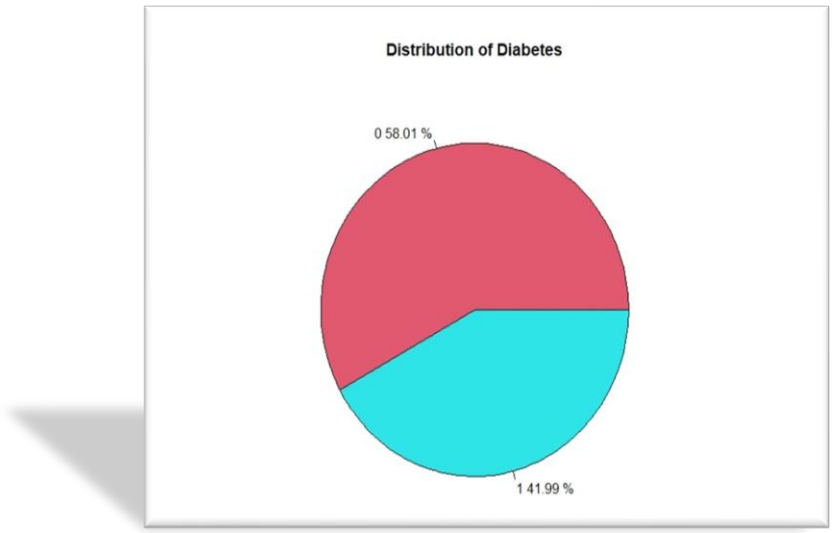
CODES AND OUTPUTS-

```
> Diabetes_freq <- table(Data$Diabetes)

> # Finding percentages of 'Yes-No'

> percentageYN <- round(Diabetes_freq/986*100,digits=2)
```

DRAWING A PIE CHART FOR "DIABETES"

CODES AND OUTPUTS-

>pie(Diabetes_freq,labels = paste(names(Diabetes_freq),percentageYN,"%",sep = " "),main = "Distribution of Diabetes",border="black",col = c(2,5))



(Figure – 2.a)

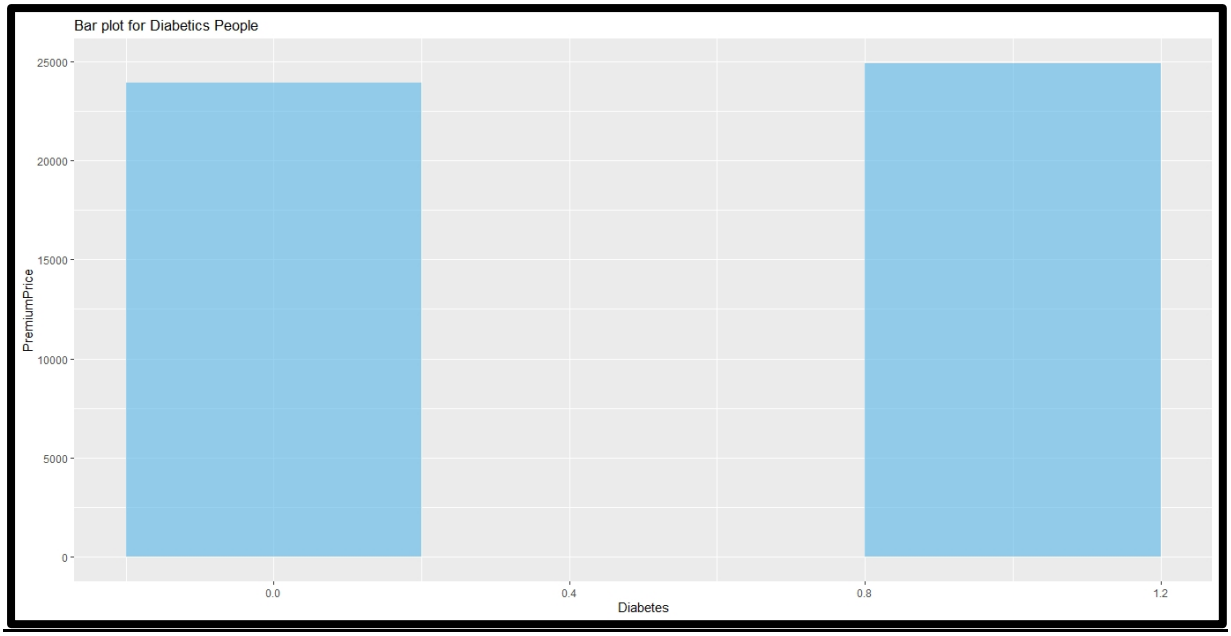
FINDINGS-

From the pie-chart for “Diabetes” we can see that 41.99% of the total population are diabetic and 58.01% of the total population are non-diabetic.

#AVERAGE DIFFERENCE IN PREMIUM PRICES FOR DIABETIC AND NON-DIABETIC PEOPLE

CODES AND OUTPUTS-

```
> Data %>%
  select(Diabetes,PremiumPrice) %>%
  group_by(Diabetes) %>%
  summarise( PremiumPrice = mean(PremiumPrice)) %>%
  ggplot(.,aes(Diabetes,PremiumPrice))+
  geom_bar(stat = "identity",width = 0.4, fill = "#56B4E9", alpha = 0.6)+
  labs(title = "Bar plot for Diabetics People")
```



(Figure – 2.b)

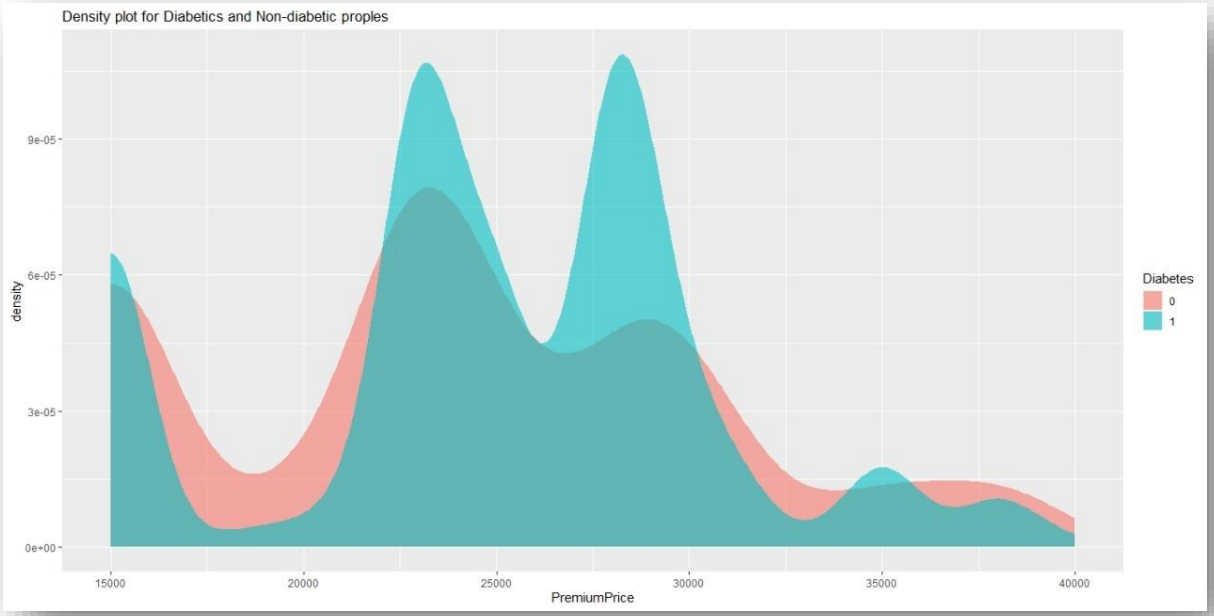
FINDINGS-

The above bar graph assess the differences between bars to evaluate how the metric changes between categorical values. Also it identifies the groups that have the highest and lowest values. From the above figure we can see that the people having diabetes gets more medical premium than the people who don't have diabetes

#DISTRIBUTION OF PREMIUM PRICES FOR DIABETIC AND NON-DIABETIC PEOPLES

CODES AND OUTPUTS-

```
> ggplot(Data, aes(PremiumPrice))+
 geom_density(aes(fill = Diabetes), color = NA, alpha = 0.6)+
 labs(title = "Density plot for Diabetics and Non-diabetic proples")
```



(Figure – 2.c)

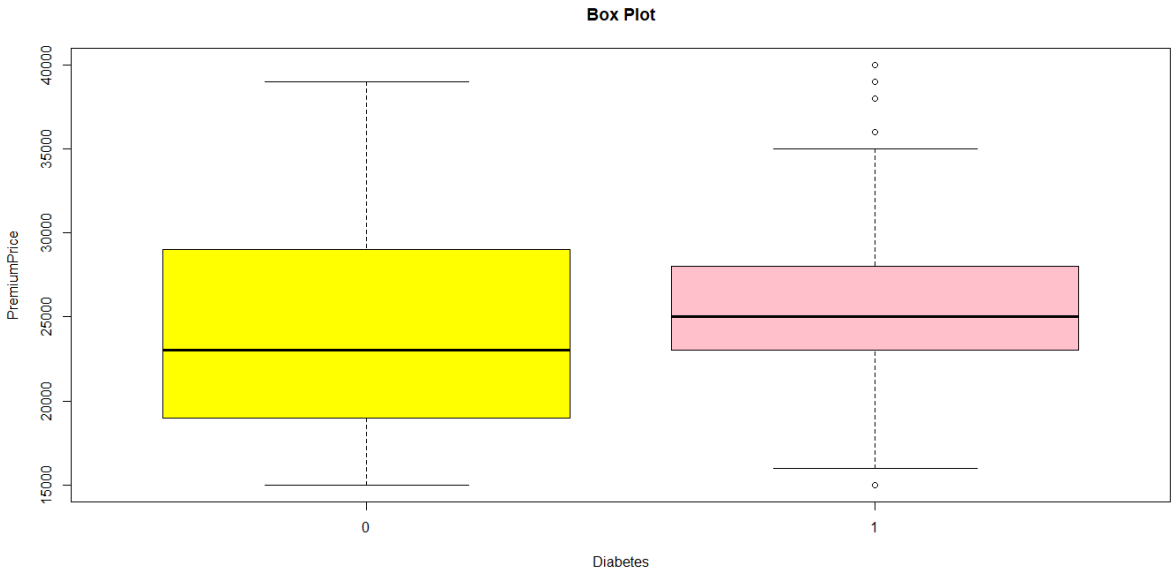
FINDINGS-

From the above diagram the density plot for diabetic and non-diabetic people are both left skewed i.e. mean Is less than the median (as the mean of the premium price is 24337). As the number of peaks for diabetic and non-diabetic people are more than two so they have a multimodal distribution.

CREATING BOXPLOTS FOR PREMIUMPRICE ~ DIABETES

CODES AND OUTPUTS-

```
boxplot(Data$ PremiumPrice ~Data$Diabetes, xlab = " Diabetes ",ylab = " PremiumPrice ", main =
"Box Plot",col=c("yellow","pink"))
```



(Figure – 2.d)

FINDINGS-

From the above diagram -

1. The box plot for the people with diabetes is comparatively short**.** This suggests that overall values have a high level of agreement with each other.

2. The box plot for the people without diabetes is comparatively tall. This suggests the people hold quite different values of premium price.

3. The box plot for the people with diabetes has five outliers.

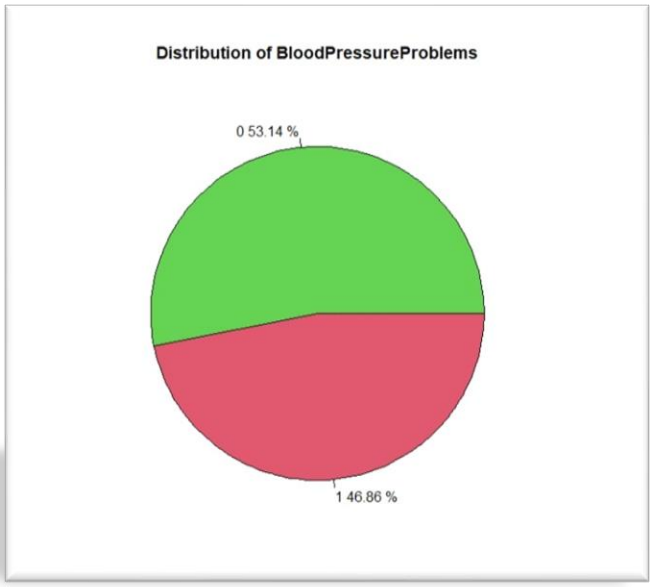
• Blood Pressure Patients Premium Analysis

➢*Pie chart for BloodPressureProblems:*

CODES AND OUTPUTS-

```
> # Creating frequency table with respect to the 'BloodPressureProblems' column
> BloodPressureProblems_freq <- table(Data$BloodPressureProblems)
> # Finding percentages of 'Yes-No'
> percentageYN <- round(BloodPressureProblems_freq/986*100,digits=2)
> # Drawing a Pie Chart for "BloodPressureProblems"
```

>pie(BloodPressureProblems_freq,labels=paste(names(BloodPressureProblems_freq),percentageYN, "%",sep = " "),main = "Distribution of BloodPressureProblems",border="black",col = c(3,2))



(Figure – 3.a)

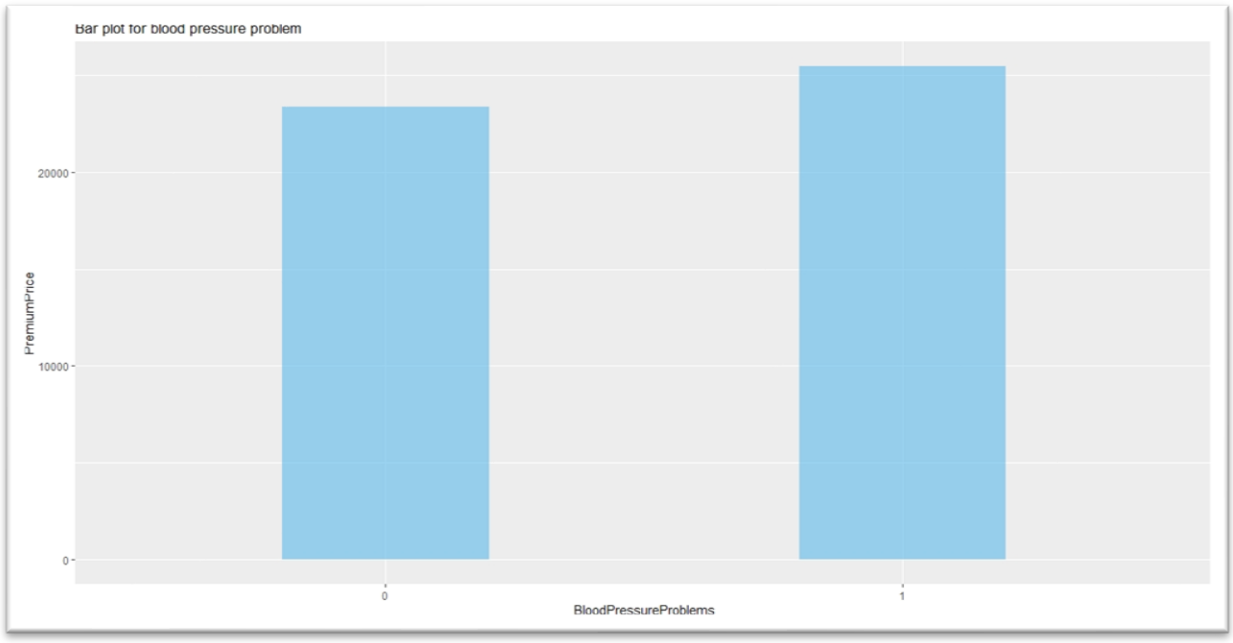
FINDINGS-

From the pie-chart for "BloodPressureProblems" we can observe that 53.14% of the total population have no blood pressure problem and rest 46.86% people of the total population have blood pressure problems.

#AVERAGE DIFFERENCE IN PREMIUM PRICES FOR BLOOD PRESSURE PATIENTS AND NON-BLOOD PRESSURE PATIENTS

CODES AND OUTPUTS-

```
> Data %>%
  select(BloodPressureProblems,PremiumPrice) %>%
  group_by(BloodPressureProblems) %>%
  summarise( PremiumPrice = mean(PremiumPrice)) %>%
  ggplot(.,aes(BloodPressureProblems,PremiumPrice))+
  geom_bar(stat = "identity",width = 0.4, fill = "#56B4E9", alpha = 0.6)+
  labs(title = "Bar plot for blood pressure problem")
```



(Figure – 3.b)

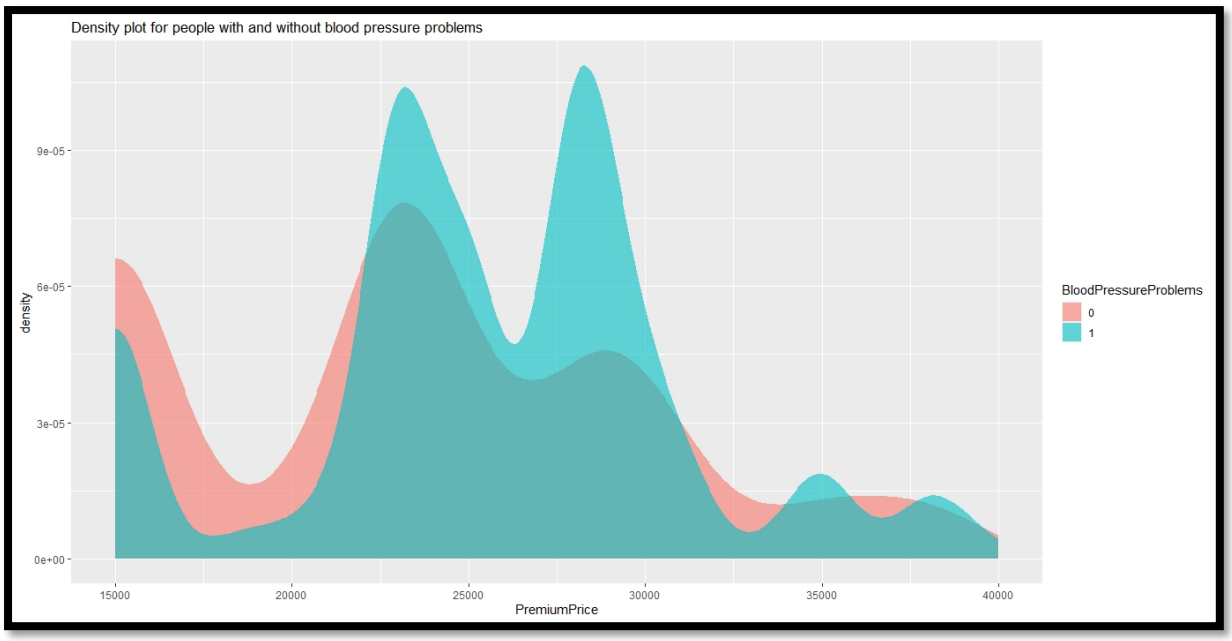
FINDINGS-

The above bar graph assess the differences between bars to evaluate how the metric changes between categorical values. Also it identifies the groups that have the highest and lowest values. From the above figure we can see that the people having blood pressure problems gets more medical premium than the people who don't have blood pressure prolems.

#DISTRIBUTION OF PREMIUM PRICES FOR BLOOD PRESSURE PATIENTS AND NON-BLOOD PRESSURE PATIENTS

CODES AND OUTPUTS-

```
ggplot(Data, aes(PremiumPrice))+
  geom_density(aes(fill = BloodPressureProblems), color = NA, alpha = 0.6)+
  labs(title = "Density plot for people with and without blood pressure problems")
```



(Figure – 3.c)

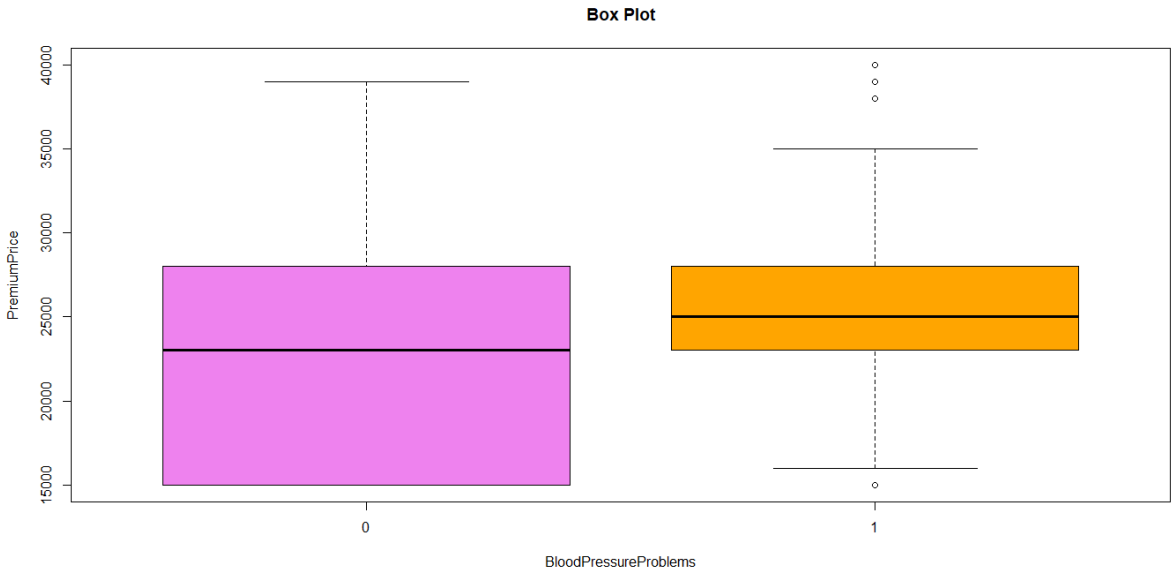
FINDINGS-

From the above diagram the density plot for the people with and without any BP problems are left skewed i.e. mean is less than the median (as the mean of the premium price is 24337). As the number of peaks for for the people with and without any BP problems are more than or equals to two so they are multimodal distributions .

CREATING BOXPLOTS FOR PREMIUMPRICE ~ BLOODPRESSUREPROBLEMS

CODES AND OUTPUTS-

boxplot(Data$ PremiumPrice ~Data$BloodPressureProblems, xlab = " BloodPressureProblems ",ylab = " PremiumPrice ", main = "Box Plot",col=c("violet","orange"))



(Figure – 3.d)

FINDINGS-

From the above diagram -

1. The box plot for the people with blood pressure problems is comparatively short**.** This suggests that overall values have a high level of agreement with each other.

2. The box plot for the people without blood pressure problems is comparatively tall. This suggests the people hold quite different values of premium price.

3. The box plot for the people with blood pressure problems has four outliers.

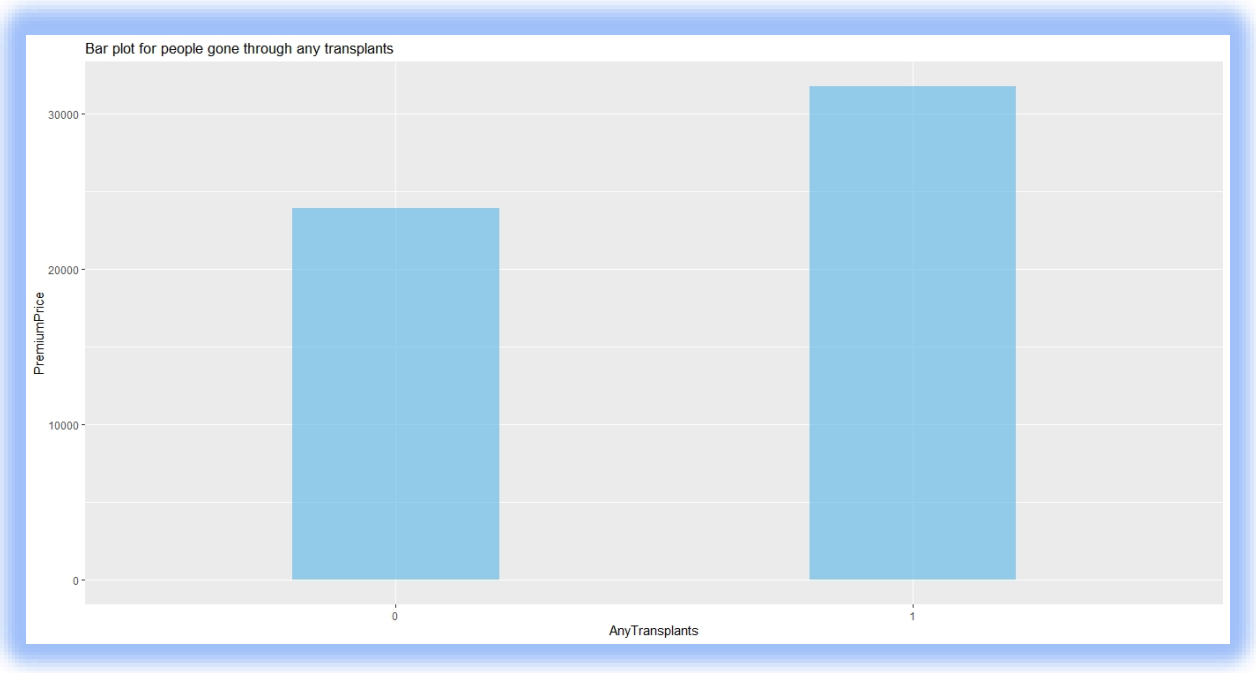
• People Gone Through Any Transplants Premium Analysis

#AVERAGE DIFFERENCE IN PREMIUM PRICES FOR PEOPLE GONE THROUGH ANY TRANSPLANTS VS THOSE WHO HAVEN'T GONE THROUGH ANY TRANSPLANTS

CODES AND OUTPUTS-

```
Data %>%
  select(AnyTransplants,PremiumPrice) %>%
  group_by(AnyTransplants) %>%
```

```
 summarise( PremiumPrice = mean(PremiumPrice)) %>%
 ggplot(.,aes(AnyTransplants,PremiumPrice))+
 geom_bar(stat = "identity",width = 0.4, fill = "#56B4E9", alpha = 0.6)+
 labs(title = "Bar plot for people gone through any transplants")
```



(Figure – 4.a)

FINDINGS-

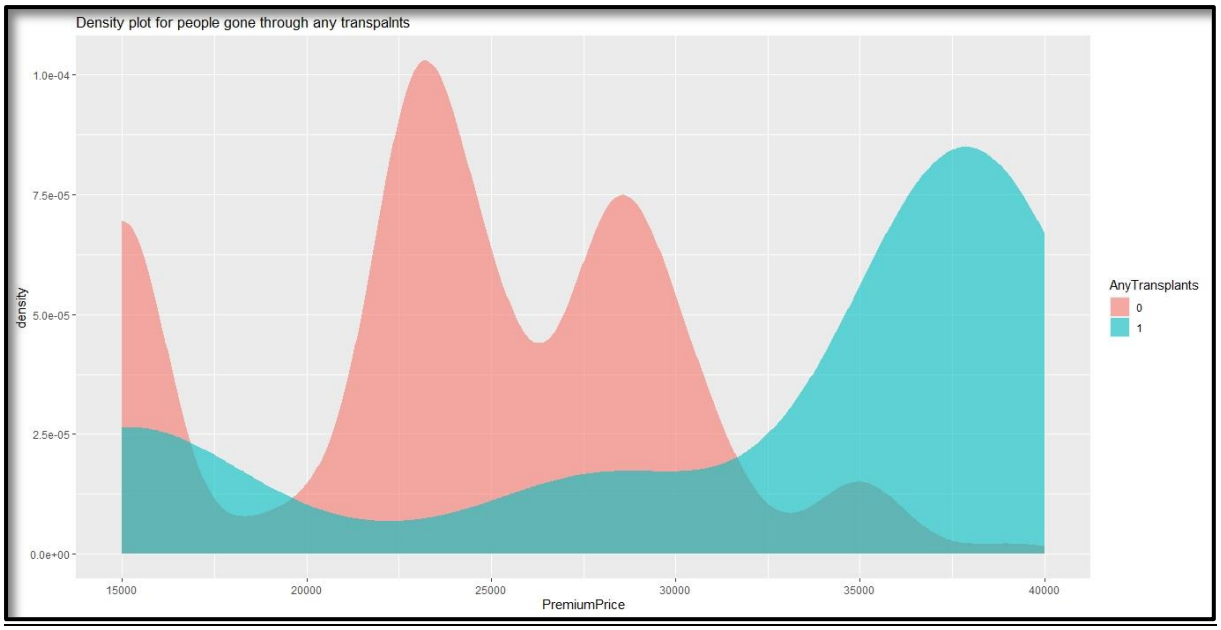
The above bar graph assess the differences between bars to evaluate how the metric changes between categorical values. Also it identifies the groups that have the highest and lowest values. From the above figure we can see that the people who have gone through any transplants gets more medical premium than the people who have no transplants.

#DISTRIBUTION OF PREMIUM PRICES FOR PEOPLE GONE THROUGH ANY TRANSPLANTS VS THOSE WHO HAVEN'T GONE THROUGH ANY TRANSPLANTS

CODES AND OUTPUTS-

```
ggplot(Data, aes(PremiumPrice))+
 geom_density(aes(fill = AnyTransplants), color = NA, alpha = 0.6)+
```

labs(title = "Density plot for people gone through any transpalnts")



(Figure – 4.b)

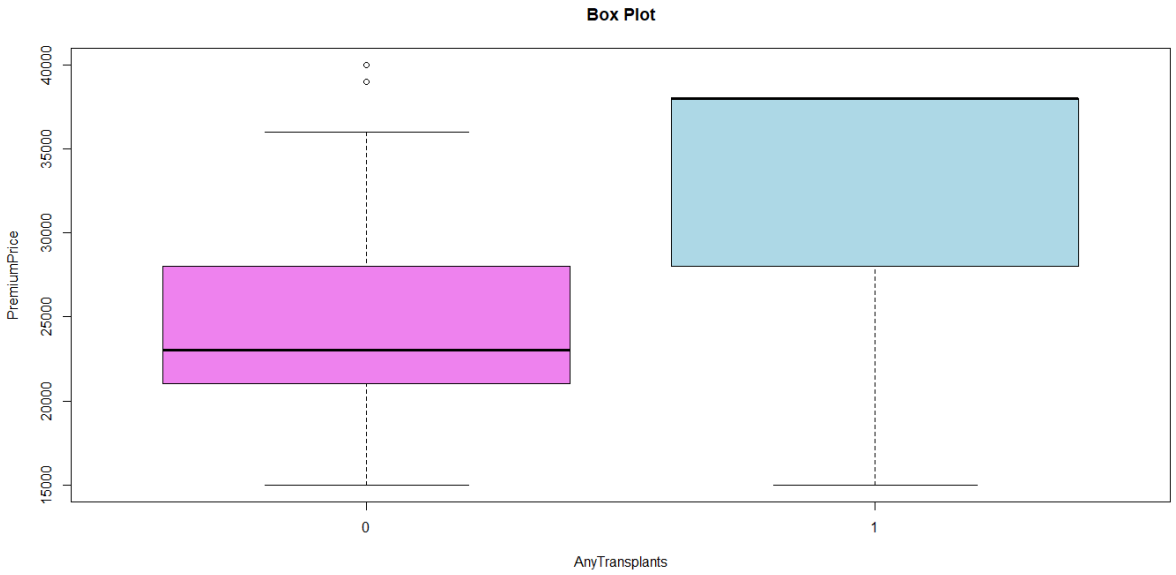
FINDINGS-

From the above diagram the density plot for who have not gone through any transplants is left skewed i.e. mean is less than the median. The density plot for the people who have gone through any transplants is right skewed i.e. mean is greater than median (as the mean of the premium price is 24337). As the number of peaks for people who have & have not gone through any transplants are more than or equals to two so they have a multimodal distribution.

CREATING BOXPLOTS FOR PREMIUMPRICE ~ ANYTRANSPLANTS

CODES AND OUTPUTS-

boxplot(Data$ PremiumPrice ~Data$AnyTransplants, xlab = " AnyTransplants ",ylab = " PremiumPrice ", main = "Box Plot",col=c("violet","light blue"))



(Figure – 4.c)

FINDINGS-

From the above diagram -

1. The box plot for the people without any transplants is comparatively short**.** This suggests that overall values have a high level of agreement with each other.

2. The box plot for the people with any transplants is comparatively tall. This suggests the values hold quite different values of premium price.

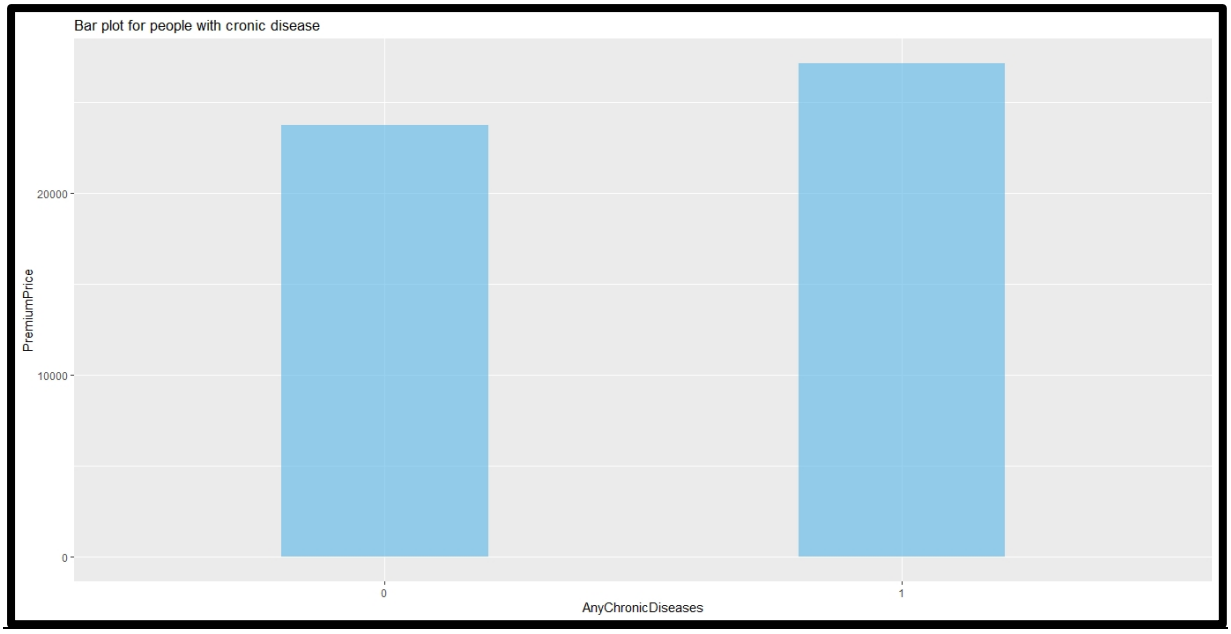
3. The box plot for the people without any transplants has two outliers.

• People With Chronic Disease Premium Analysis

#AVERAGE DIFFERENCE IN PREMIUM PRICES FOR PEOPLE WITH CHRONIC DISEASE AND PEOPLE WITH NO CHRONIC DISEASE

CODES AND OUTPUTS-

```
Data%>%select(AnyChronicDiseases,PremiumPrice)%>%group_by(AnyChronicDiseases)%>%summarise(PremiumPrice=mean(PremiumPrice))%>%ggplot(.,aes(AnyChronicDiseases,PremiumPrice))+geom_bar(stat = "identity",width = 0.4, fill = "#56B4E9", alpha = 0.6)+labs(title = "Bar plot for people with cronic disease")
```



(Figure – 5.a)

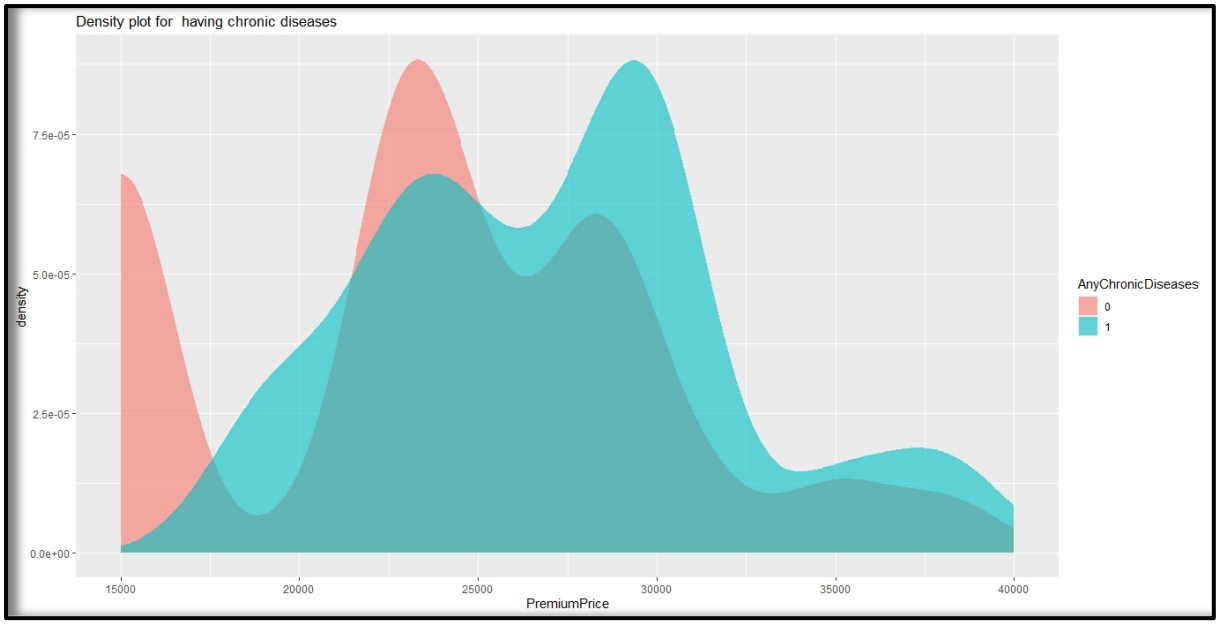
FINDINGS-

The above bar graph assess the differences between bars to evaluate how the metric changes between categorical values. Also it identifies the groups that have the highest and lowest values. From the above figure we can see that the people having any chronic diseases gets more medical premium than the people who don't have any chronic diseases.

#DISTRIBUTION OF PREMIUM PRICES FOR PEOPLE WITH CHRONIC DISEASE AND PEOPLE WITH NO CHRONIC DISEASE

CODES AND OUTPUTS-

```
ggplot(Data, aes(PremiumPrice))+geom_density(aes(fill = AnyChronicDiseases), color = NA, alpha = 0.6)+labs(title = "Density plot for  having chronic diseases")
```



(Figure – 5.b)

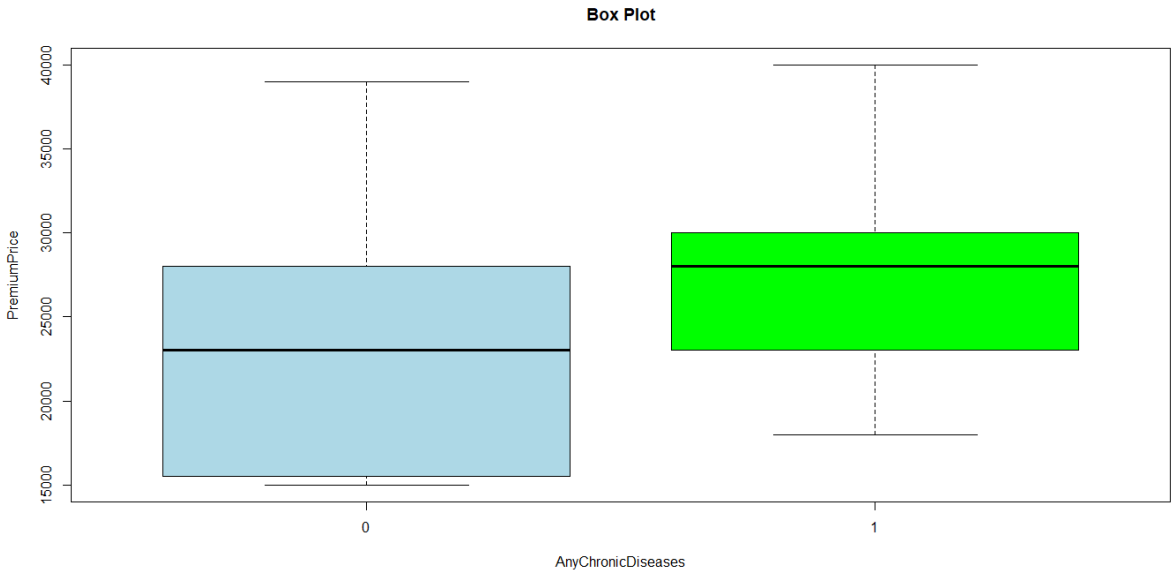
FINDINGS-

From the above diagram the density plot for the people who have no chronic diseases is left skewed i.e. mean is less than the median. The density plot for the people who have chronic diseases is right skewed i.e. mean is greater than median (as the mean of the premium price is 24337). As the number of peaks for the people who have & have no chronic diseases are more than or equals to two so they have a multimodal distribution .

CREATING BOXPLOTS FOR PREMIUMPRICE ~ ANYCHRONICDISEASES

CODES AND OUTPUTS-

```
boxplot(Data$ PremiumPrice ~Data$ AnyChronicDiseases, xlab = " AnyChronicDiseases ",ylab = "
PremiumPrice ", main = "Box Plot",col=c("light blue","green"))
```



(Figure – 5.c)

FINDINGS-

From the above diagram -

1. The box plot for the people with any chronic diseases is comparatively short**.** This suggests that overall values have a high level of agreement with each other.

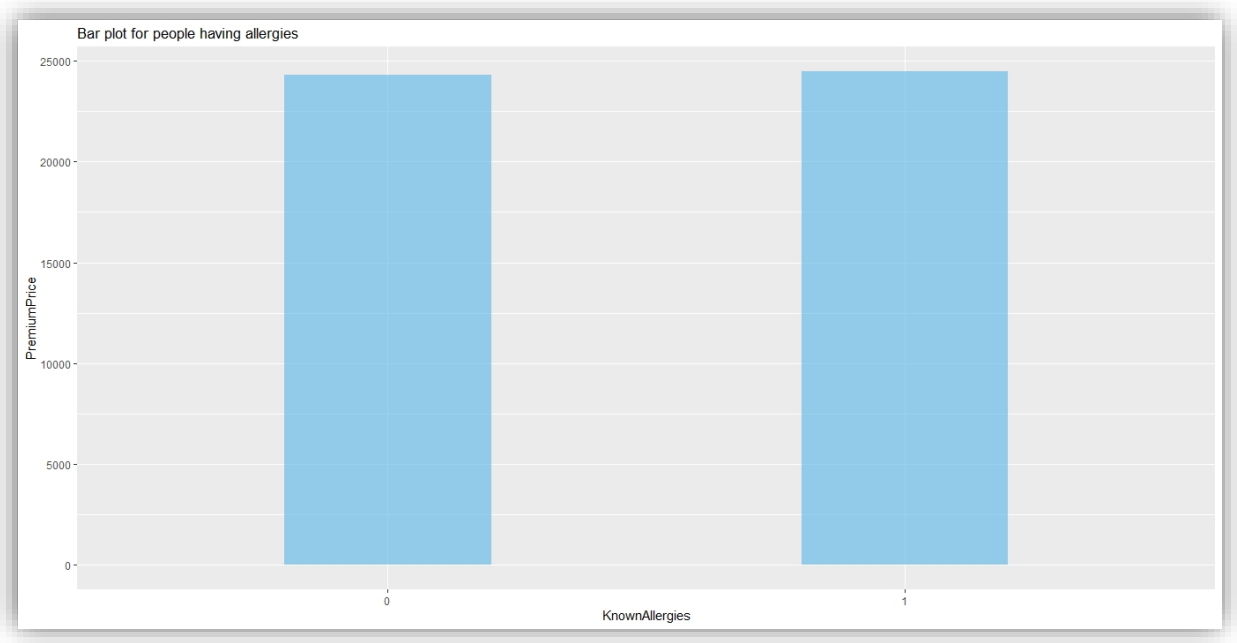
2. The box plot for the people without chronic diseases is comparatively tall. This suggests the values hold quite different values premium price.

• Allergy Patients Premium Analysis

#AVERAGE DIFFERENCE IN PREMIUM PRICES FOR ALLERGY PATIENTS AND NO ALLERGY PATIENTS

CODES AND OUTPUTS-

```
Data %>% select(KnownAllergies,PremiumPrice) %>%group_by(KnownAllergies) %>% summarise(
PremiumPrice = mean(PremiumPrice)) %>%  ggplot(.,aes(KnownAllergies,PremiumPrice))+
geom_bar(stat = "identity",width = 0.4, fill = "#56B4E9", alpha = 0.6)+labs(title = "Bar plot for people having allergies")
```



(Figure – 6.a)

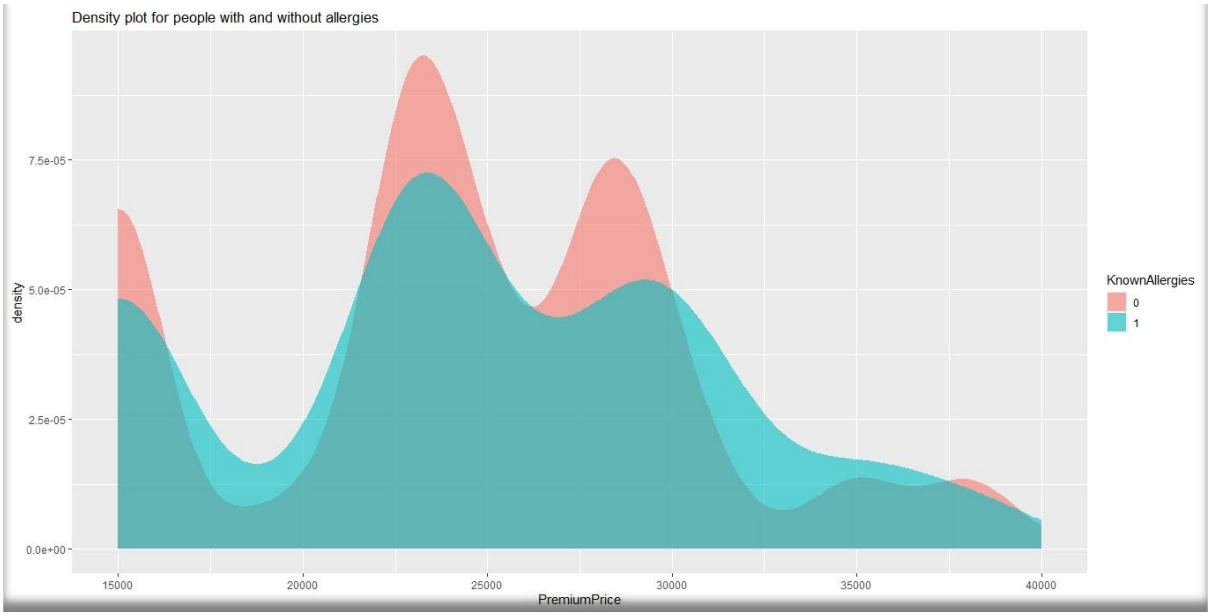
FINDINGS-

The above bar graph assess the differences between bars to evaluate how the metric changes between categorical values. Also it identifies the groups that have the highest and lowest values. From the above figure we can see that the people having allergies gets more medical premium than the people who don't have allergies.

#DISTRIBUTION OF PREMIUM PRICES FOR ALLERGY PATIENTS AND NO ALLERGY PATIENTS

CODES AND OUTPUTS-

```
ggplot(Data, aes(PremiumPrice))+
  geom_density(aes(fill = KnownAllergies), color = NA, alpha = 0.6)+
  labs(title = "Density plot for people with and without allergies")
```



(Figure – 6.b)

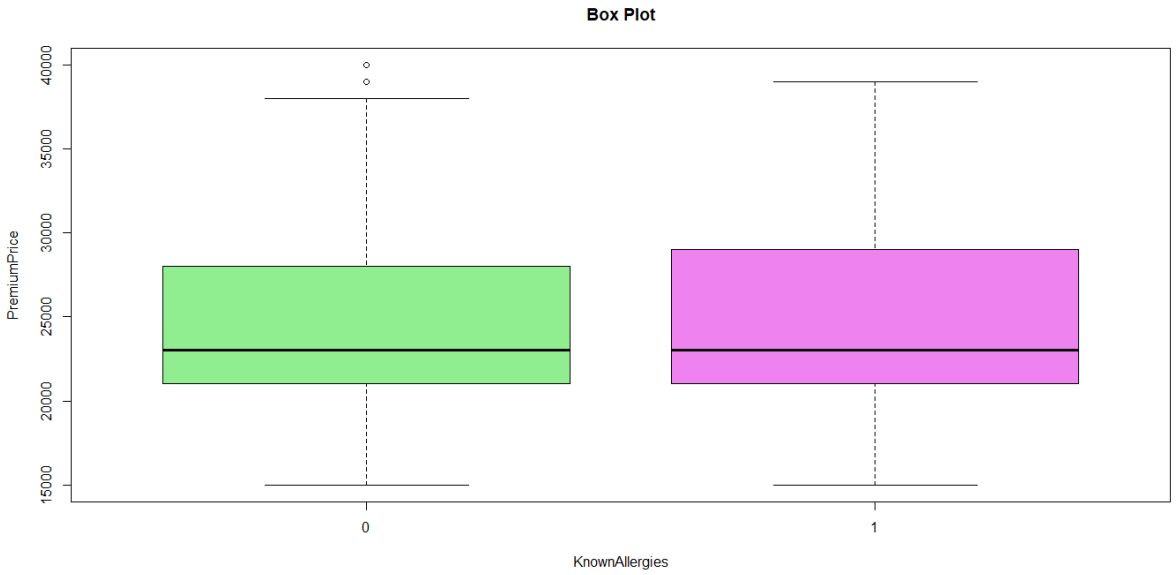
FINDINGS-

From the above diagram the density plot for the people with and without any allergies are left skewed i.e. mean is less than the median (as the mean of the premium price is 24337). As the number of peaks for with and without any allergies are more than or equals to two so they have a multimodal distribution.

CREATING BOXPLOTS FOR PREMIUMPRICE ~ KNOWNALLERGIES

CODES AND OUTPUTS-

boxplot(Data$ PremiumPrice ~Data$KnownAllergies, xlab = " KnownAllergies ",ylab = " PremiumPrice ", main = "Box Plot",col=c("light green","violet"))



(Figure – 6.c)

FINDINGS-

From the above diagram -

1. The boxplot for the people without any known allergies is comparatively short**.** This suggests that overall values have a high level of agreement with each other.

2. The boxplot for the people with any known allergies is comparatively tall. This suggests the values hold quite different values premium price.

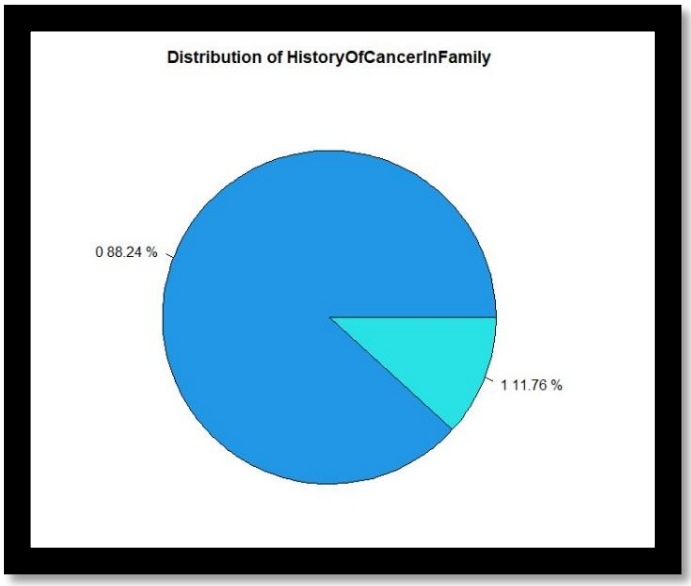
3. The boxplot for the people without any known allergies has two outliers.

• Patients with History of Cancer in Family Premium Analysis

➢*Piechart for HistoryOfCancerInFamily:*

CODES AND OUTPUTS-

```
>#Creatingfrequency table with respect to the 'HistoryOfCancerInFamily' column

> HistoryOfCancerInFamily_freq <- table(Data$HistoryOfCancerInFamily)

> # Finding percentages of 'Yes-No'

> percentageYN <- round(HistoryOfCancerInFamily_freq/986*100,digits=2)

> # Drawing a Pie Chart for "HistoryOfCancerInFamily"

>pie(HistoryOfCancerInFamily_freq,labels=paste(names(HistoryOfCancerInFamily_freq),percentageYN,"%",sep = " "),main = "Distribution of HistoryOfCancerInFamily",border="black",col =c(4,5))
```



(Figure – 7.a)

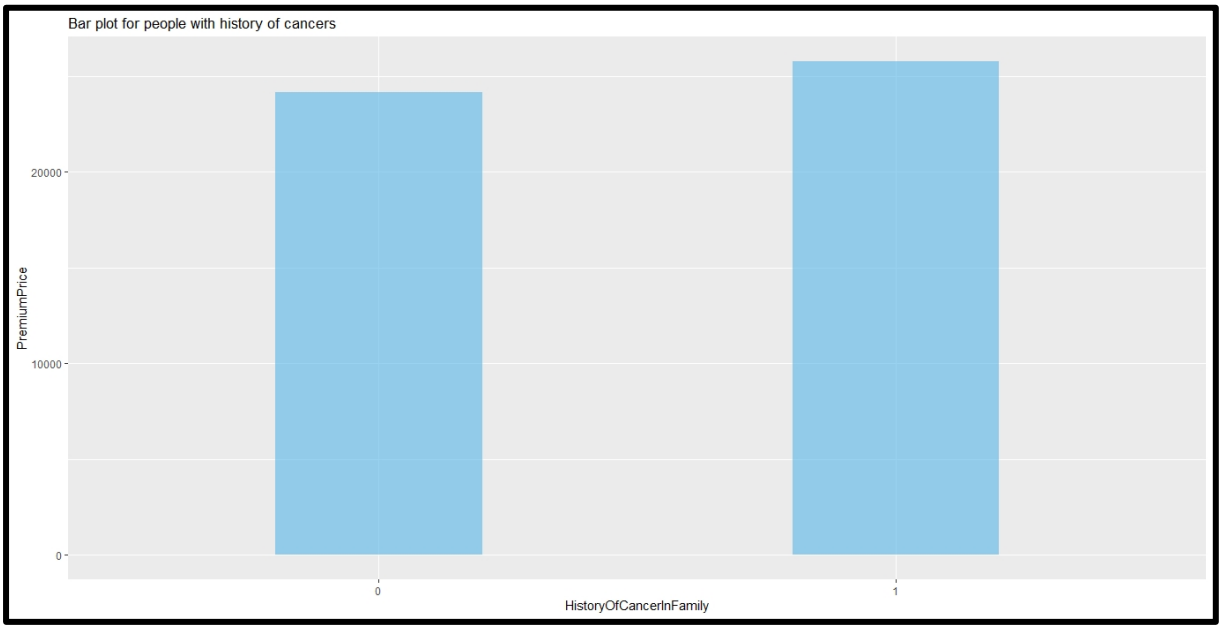
FINDINGS-

From the pie-chart for “HistoryOfCancerInFamily” we can see that 88.24% of the total population have no history of cancer in family and the rest 11.76% people of the total population have a history of cancer in family.

#AVERAGE DIFFERENCE IN PREMIUM PRICES FOR PATIENTS WITH HISTORY OF CANCER AND PATIENTS WITHOUT HISTORY OF CANCER

CODES AND OUTPUTS-

```
Data %>%
 select(HistoryOfCancerInFamily,PremiumPrice) %>%
 group_by(HistoryOfCancerInFamily) %>%
 summarise( PremiumPrice = mean(PremiumPrice)) %>%
 ggplot(.,aes(HistoryOfCancerInFamily,PremiumPrice))+
 geom_bar(stat = "identity",width = 0.4, fill = "#56B4E9", alpha = 0.6)+
 labs(title = "Bar plot for people with history of cancers")
```



(Figure – 7.b)

FINDINGS-

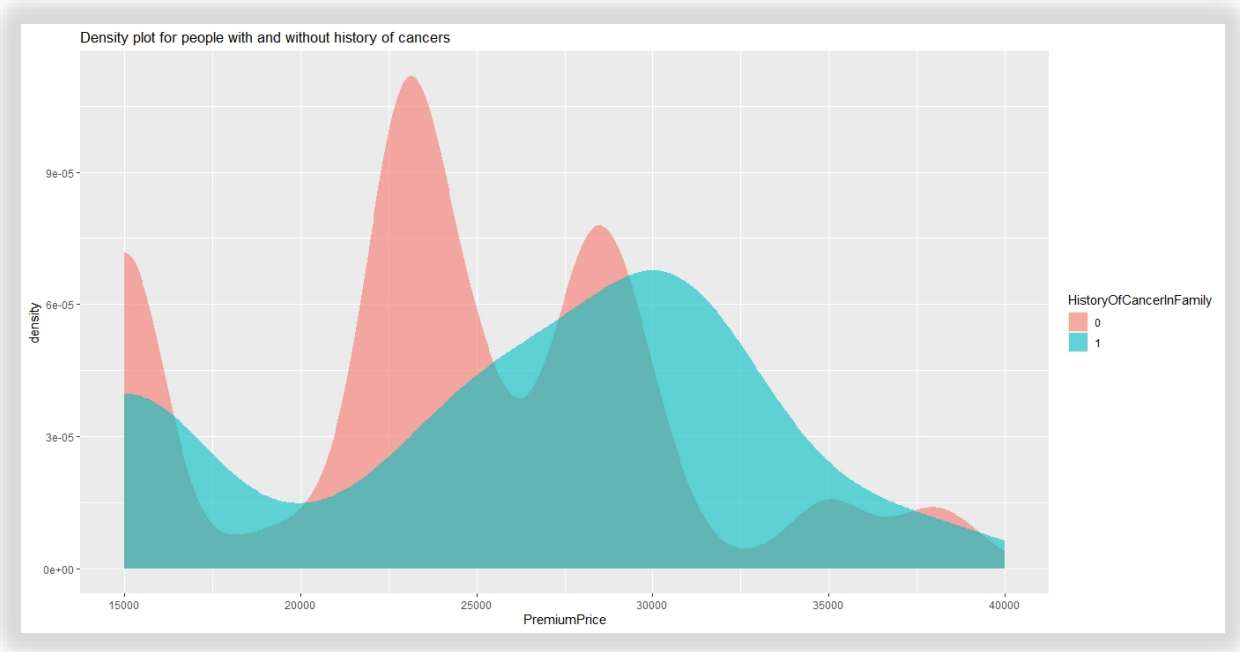
The above bar graph assess the differences between bars to evaluate how the metric changes between categorical values. Also it identifies the groups that have the highest and lowest values. From

the above figure we can see that the people having history of cancer in family gets more medical premium than the people who don't have history of cancer in family.

#DISTRIBUTION OF PREMIUM PRICES FOR PATIENTS WITH HISTORY OF CANCER AND PATIENTS WITHOUT HISTORY OF CANCER

CODES AND OUTPUTS-

```
ggplot(Data, aes(PremiumPrice))+
  geom_density(aes(fill = HistoryOfCancerInFamily), color = NA, alpha = 0.6)+
  labs(title = "Density plot for people with and without history of cancers")
```



(Figure – 7.c)

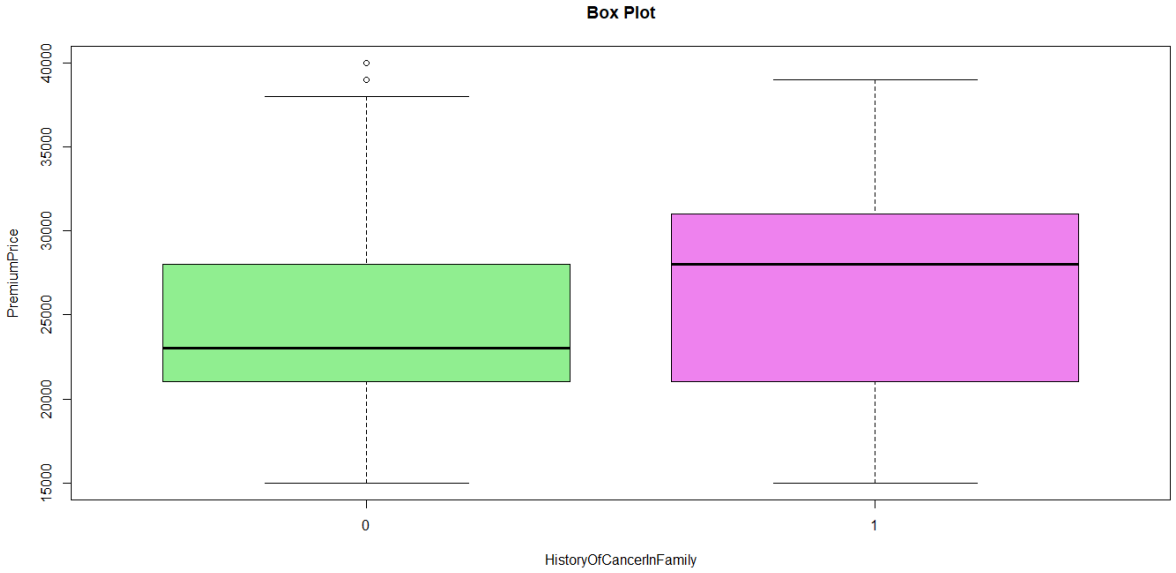
FINDINGS-

From the above diagram the density plot for the people without a history of cancer is left skewed i.e. mean is less than the median. The density plot for the people with a history of cancer is right skewed i.e. mean is greater than the median (as the mean of the premium price is 24337). As the number of peaks for the people with & without a history of cancer are more than or equals to two so they have multimodal distribution .

CREATING BOXPLOTS FOR PREMIUMPRICE ~ HISTORYOFCANCERINFAMILY

CODES AND OUTPUTS-

>boxplot(Data$ PremiumPrice ~Data$ HistoryOfCancerInFamily, xlab = " HistoryOfCancerInFamily ",ylab = " PremiumPrice ", main = "Box Plot",col=c("light green","violet"))



(Figure – 7.d)

FINDINGS-

From the above diagram -

1. The boxplot for the people without a history of cancer is comparatively short**.** This suggests that overall values have a high level of agreement with each other.

2. The boxplot for the people with a history of cancer is comparatively tall. This suggests the values hold quite different values premium price.

3. The boxplot for the people without a history of cancer has two outliers.

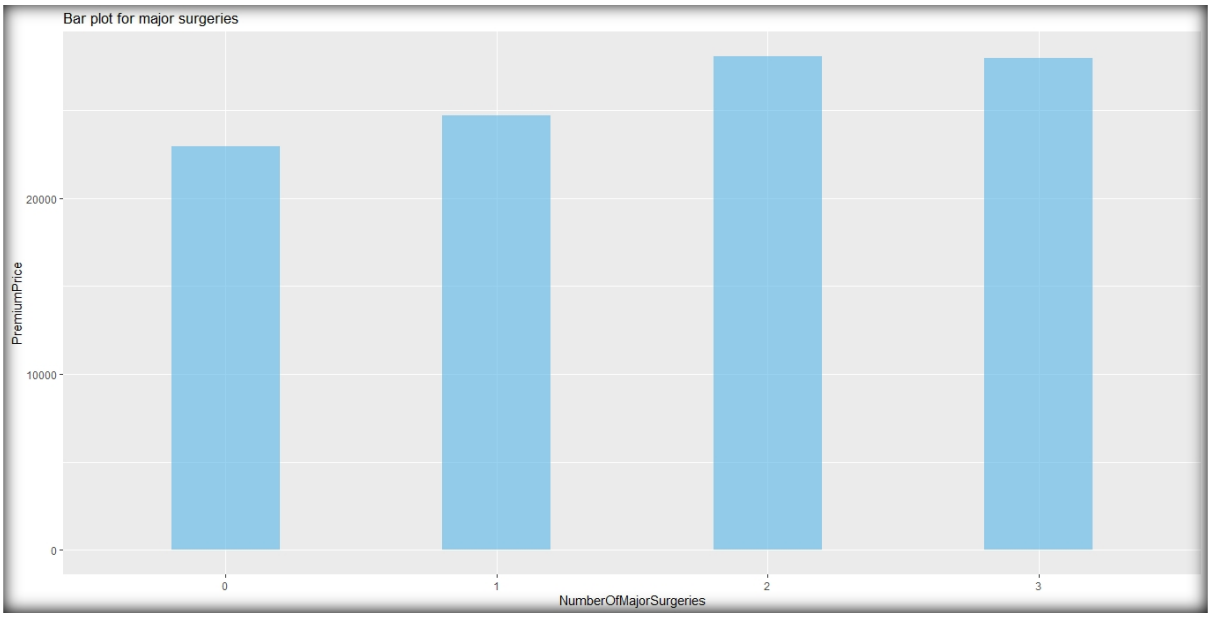
• People Gone Through Major Surgeries Premium Analysis

#AVERAGE DIFFERENCE IN PREMIUM PRICES FOR PEOPLE GONE THROUGH MAJOR SURGERIES

CODES AND OUTPUTS-

```
Data %>%
  select(NumberOfMajorSurgeries,PremiumPrice) %>%
  group_by(NumberOfMajorSurgeries) %>%
  summarise( PremiumPrice = mean(PremiumPrice)) %>%
  ggplot(.,aes(NumberOfMajorSurgeries,PremiumPrice))+
```

```
geom_bar(stat = "identity",width = 0.4, fill = "#56B4E9", alpha = 0.6)+
labs(title = "Bar plot for major surgeries")
```



(Figure – 8.a)

FINDINGS-

The above bar graph assess the differences between bars to evaluate how the metric changes between categorical values. Also it identifies the groups that have the highest and lowest values. From the above figure we can see that the people having two or three major surgeries gets more medical premium than the people who don't have or have 1 major surgery.

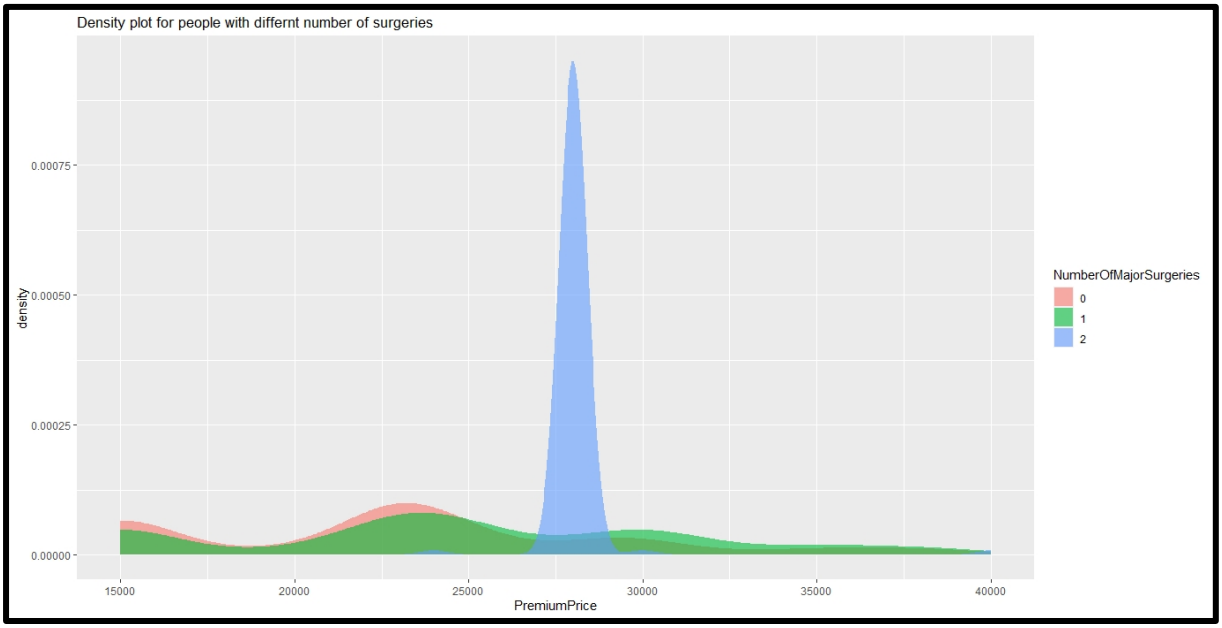
#DISTRIBUTION OF PREMIUM PRICES FOR PEOPLE GONE THROUGH MAJOR SURGERIES

CODES AND OUTPUTS-

Here people with 4 are charged constant 28000 hence neglected that uniform distribution.

```
ggplot(Data %>%

select(NumberOfMajorSurgeries,PremiumPrice) %>%

    filter(!NumberOfMajorSurgeries == 3),

    aes(PremiumPrice))+

 geom_density(aes(fill = NumberOfMajorSurgeries), color = NA, alpha = 0.6)+
```

labs(title = "Density plot for people with differnt number of surgeries")



(Figure – 8.b)

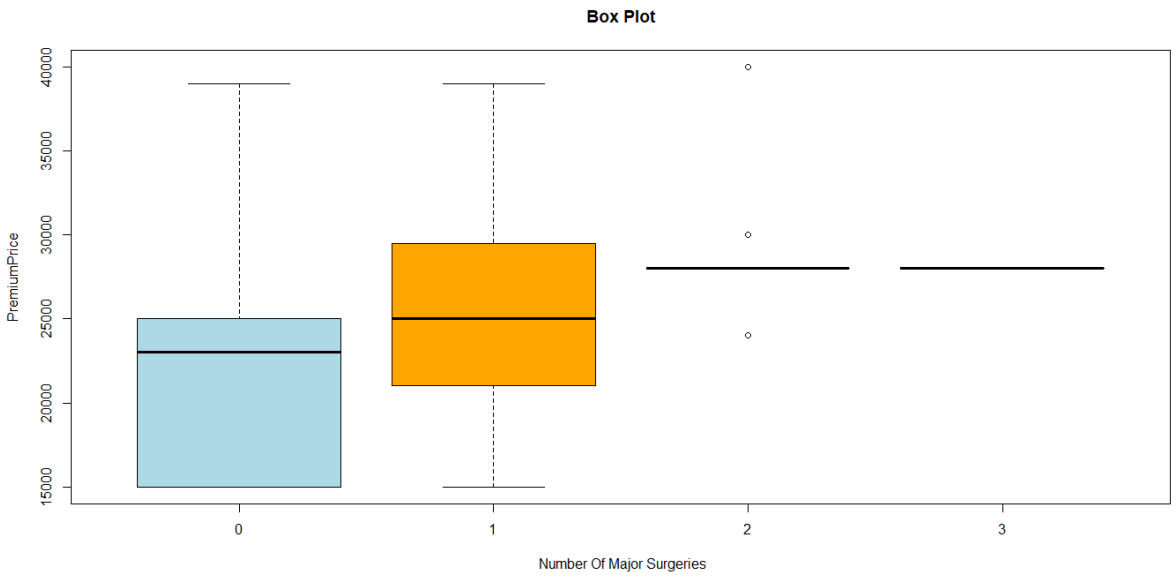
FINDINGS-

From the above diagram the density plot for “0” & “1” major surgeries are left skewed i.e. mean is less than the median. The density plot for “2” major surgeries is right skewed i.e. mean is greater than median (as the mean of the premium price is 24337). As the number of peaks for “0” and “1” surgeries are more than or equals to two so they are multimodal distributions & the density plot for “2” has one peak so this is a unimodal distribution.

 # CREATING BOXPLOTS FOR CHARGES ~ NUMBER OF NUMBER OF MAJOR SURGERIES

CODES AND OUTPUTS-

>boxplot(Data$PremiumPrice~Data$NumberOfMajorSurgeries, xlab = " Number Of Major Surgeries", ylab = "PremiumPrice", main = "Box Plot",col=c("light blue","orange","yellow","light green"))



(Figure – 8.c)

FINDINGS-

From the above diagram -

1. The boxplots for the people with 2 & 3 major surgeries are comparatively short**.** This suggests that overall values have a high level of agreement with each other.

2. The boxplot for the people with no & 1 major surgeries is comparatively tall. This suggests the values hold quite different values premium price.

3. The boxplot for the people with 2 major surgeries has three outliers.

• Distribution of BMI

 # ASSIGNING THE 'BMI VALUES TO A VARIABLE

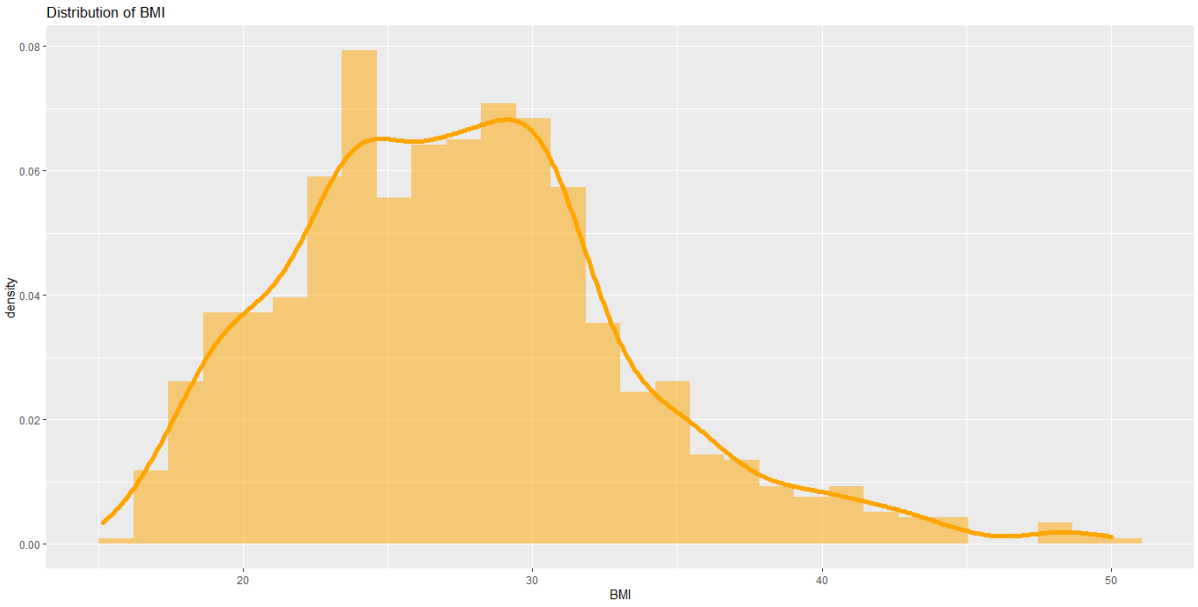
CODES AND OUTPUTS

```
BMI = Data$bmi
```

 # CREATING HISTOGRAM FOR DISTRIBUTION OF BMI

CODES AND OUTPUTS

```
ggplot(data.frame(Data$bmi), aes(x=BMI)) +
geom_histogram(aes(y=..density..),fill="orange",alpha=0.5) +
geom_density(alpha=.2,col="orange",size=2) + labs(title="Distribution of BMI")
```



(Figure – 9.a)

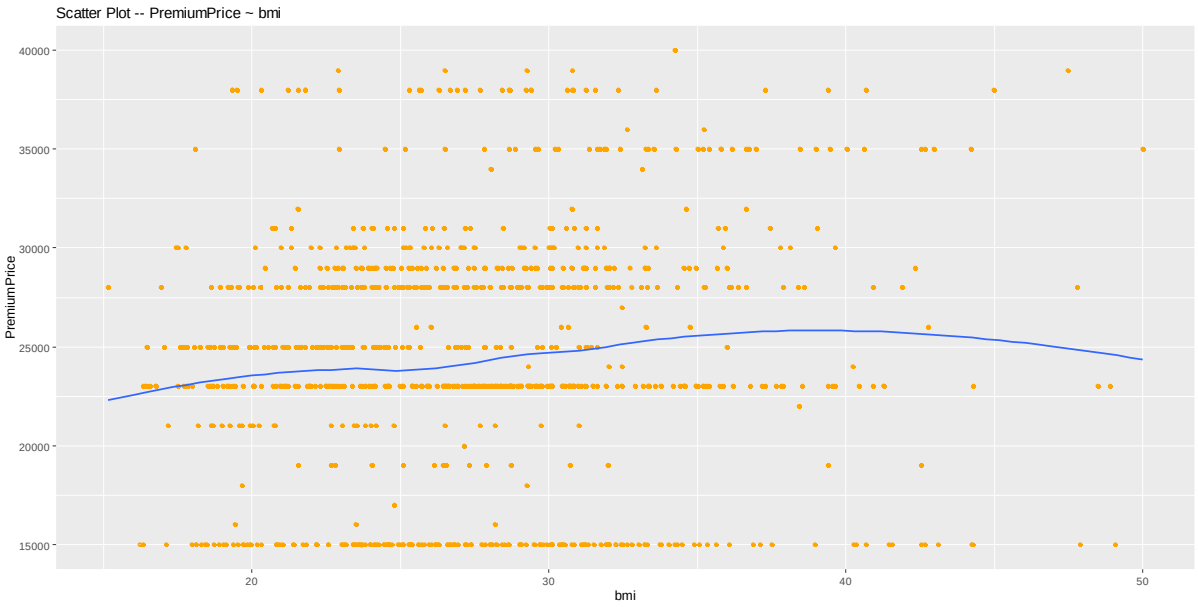
FINDINGS-

From the Histogram for 'BMI',
(1) We can see that the histogram of the BMI represents that maximum people have BMI around 20 – 40 and very few people have BMI around 10, similarly very few people have BMI around 50.
(2) We can see that the distribution of BMI is an undefined Distribution & it is bimodal.
(3) The distribution of BMI is right skewed i.e. mean > median.

 # CREATING SCATTER PLOTS TO SHOW THE VARIATION

CODES AND OUTPUTS

```
>ggplot(Data,aes(x=bmi,y=PremiumPrice))+geom_point(col="orange")+geom_smooth(method="auto", se=TRUE, fullrange=FALSE, level=0.95)+labs(title="Scatter Plot -- PremiumPrice ~ bmi")
```



(Figure – 9.b)

FINDINGS-

Scatterplots display the direction, strength, and linearity of the relationship between two variables. (1) From the above diagram we can see that the relationship between BMI and premium price is positive i.e. they are positively correlated.

(2) Stronger relationships produce a tighter clustering of data points. From this diagram as the data points don't cluster that much tightly so they provide a moderately strong relationship.

• Distribution of PremiumPrice

 # ASSIGNING THE 'PREMIUMPRICE' VALUES TO A VARIABLE

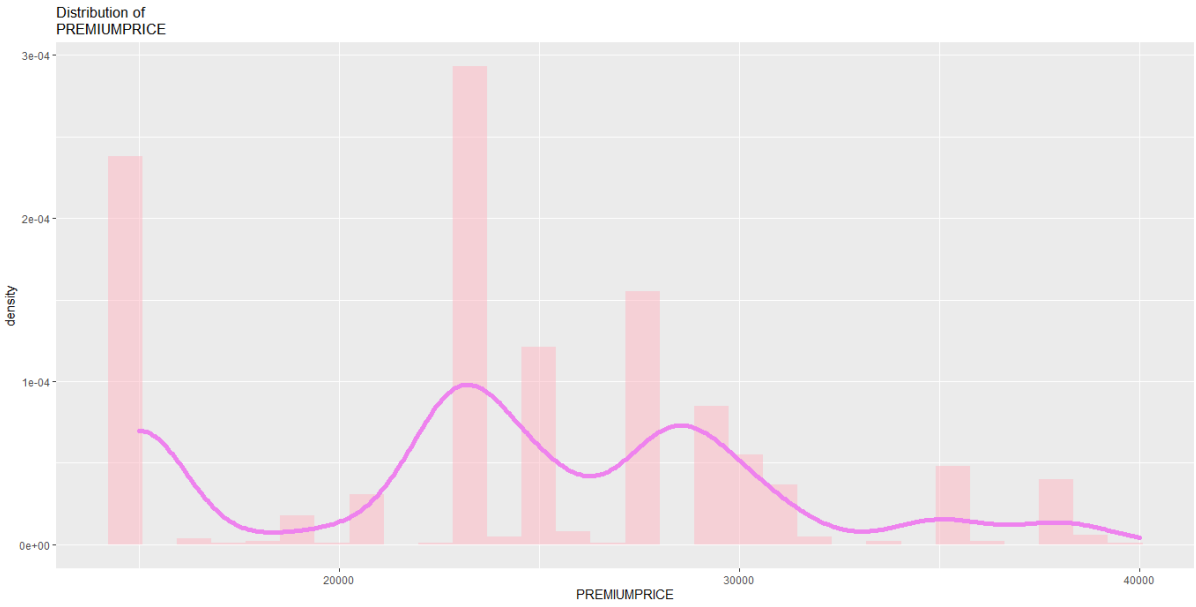
CODES AND OUTPUTS

```
PREMIUMPRICE = Data$PremiumPrice
```

 # CREATING HISTOGRAM FOR DISTRIBUTION OF PREMIUMPRICE

CODES AND OUTPUTS

```
ggplot(data.frame(Data$PremiumPrice), aes(x=PREMIUMPRICE)) +
geom_histogram(aes(y=..density..),fill="light
pink",alpha=0.5)geom_density(alpha=.2,col="violet",size=2) + labs(title="Distribution of
PREMIUMPRICE")
```



(Figure – 10.a)

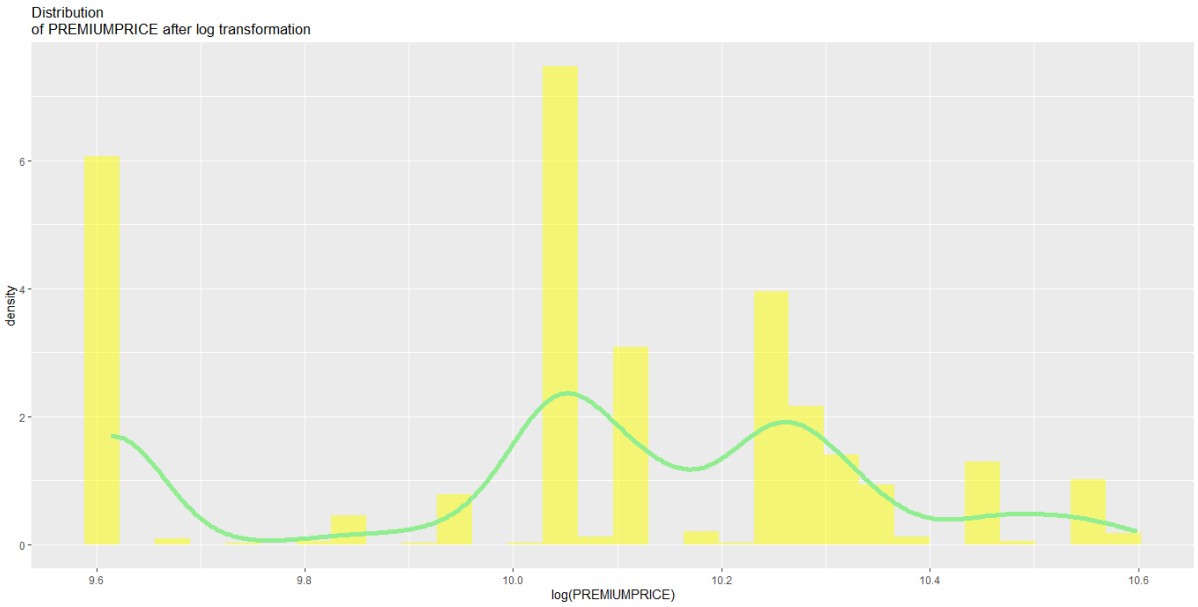
FINDINGS-

From the Histogram for 'PremiumPrice',
(1) We can see that the histogram of the PremiumPrice represents that maximum people get PremiumPrice around 20k – 30k and few people get PremiumPrice around 10k , similarly very few people get PremiumPrice around 40k.
(2) We can see that the distribution of PremiumPrice is an undefined distribution & it is multimodal.
(3) The distribution of PremiumPrice is right skewed i.e. mean > median.

 # CREATING HISTOGRAM FOR DISTRIBUTION OF PREMIUMPRICE

CODES AND OUTPUTS-

```
ggplot(data.frame(Data$PremiumPrice), aes(x=log(PREMIUMPRICE)))+
geom_histogram(aes(y=..density..),fill="yellow",alpha=0.5) + geom_density(alpha=.01,col="light
green",size=2) + labs(title="Distribution of PREMIUMPRICE after log transformation")
```



(Figure – 10.b)

FINDINGS-

From the Histogram for 'PremiumPrice',
(1) We can see that the histogram of the PremiumPrice represents that maximum people get PremiumPrice around 20k – 30k and few people have PremiumPrice around 10k , similarly very few people have PremiumPrice around 40k.
(2) We can see that the distribution of PremiumPrice is an undefined distribution & it is multimodal.
(3) The distribution of PremiumPrice is right skewed i.e. mean > median.

We have used the Bubble Chart to represent the relation of Charges with Bmi, Age, and Diabetes–

 # PLOTTING THE RELATION OF PREMIUMPRICE WITH BMI, AGE, & DIABETES

CODES AND OUTPUTS

```
ggplot(Data,aes(x= PremiumPrice,y=bmi,color=Diabetes,size=Age)) + geom_point(alpha = .6) +
labs(title = "Relation of PremiumPrice with Bmi, Age, and Diabetes")
```

(Figure – 10.c)

FINDINGS-

Scatterplots display the direction, strength, and linearity of the relationship between two variables. (1) From the above diagram we can see that the relationship between BMI , age , diabetes and premium price is positive i.e. they are positively correlated.

(2) Stronger relationships produce a tighter clustering of data points. From this diagram as the data points don't cluster that much tightly so they provide a moderately strong relationship.

❖ HYPOTHESIS TESTING:

Normality Test:-

CODES AND OUTPUTS-

```
> data<-rnorm(100)

> shapiro.test(data)

        Shapiro-Wilk normality test

data:  data

W = 0.98936, p-value = 0.6132
```

TESTING-

Here we want to test whether our data is normally distributed or not.

Our null hypothesis is H_0: Our data is normally distributed vs. H_1: Our data is not normally distributed.

To test the above we take a random sample of 100 data. And we apply the "shapiro.test" code to get the p-value.

Now as the p-value(0.6132) is greater than 0.05 so at 5% level of significance we accept the null hypothesis and conclude that our data is normally distributed.

t-Test:-

1. Test to check the significance of avg. premium price for diabetic and non-diabetic patients:

CODES AND OUTPUTS-

```
> t.test(PremiumPrice~Diabetes,mu=0,conf=0.95,var.eq=F,Paired=F)

        Welch Two Sample t-test

data:  PremiumPrice by Diabetes

t = -2.4489, df = 949.09, p-value = 0.01451

alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0

95 percent confidence interval:
```

-1737.0795 -191.5546

sample estimates:

mean in group 0 mean in group 1

23931.82 24896.14

TESTING-

Here we want to test if there is any significant difference between the average premium price for diabetic and non-diabetic patients is present or not.

Our Null hypothesis (H_0) is , there is no significant difference & our alternative hypothesis (H_1) is , there exists true difference between average premium prices for diabetic and non-diabetic people.

i.e. $H_0 : \mu A - \mu B = 0$ VS. $H_1 : \mu A - \mu B \neq 0$.

Where μA= average premium price for diabetic people .

μB= average premium price for non-diabetic people.

As the p-value (=0.01451) < 0.05 so we reject H_0 at 5% level of significance , so we reject the null hypothesis and conclude that the alternative hypothesis is true.

So , there is a significant difference between means of premium prices for diabetic and non-diabetic people . The predicted values of μA & μB are 24896.14 and 23931.82 respectively.

2.Test to check the significance of avg. premium price for blood pressure and non-blood pressure patients:

CODES AND OUTPUTS-

```
> t.test(PremiumPrice~BloodPressureProblems,mu=0,conf=0.95,var.eq=F,Paired=F)

                    Welch Two Sample t-test

data:  PremiumPrice by BloodPressureProblems

t = -5.3703, df = 982.76, p-value = 9.813e-08

alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0

95 percent confidence interval:

-2855.331 -1327.033

sample estimates:

mean in group 0  mean in group 1

23356.87         25448.05
```

TESTING-

Here we want to test if there is any significant difference between the average premium price for the people with blood pressure problems and without pressure problems is present or not.

Our Null hypothesis (H_0) is , there is no significant difference & our alternative hypothesis (H_1) is , there exists true difference between average premium prices for the people with blood pressure problems and without pressure problems.

i.e. $H_0 : \mu A - \mu B = 0$ VS. $H_1 : \mu A - \mu B \neq 0$.

Where μA= average premium price for the people with blood pressure problems.

μB= average premium price for the people without blood pressure problms.

As the p-value (=9.813e-08) < 0.05 so we reject H_0 at 5% level of significance , so we reject the null hypothesis and conclude that the alternative hypothesis is true.

So , there is a significant difference between means of premium prices for the people with blood pressure problems and without pressure problems. The predicted values of μA & μB are 25448.05 and 23356.87 respectively.

3.Test to check the significance of avg. premium price for people gone through any transplants and who haven't gone through any transplants:

CODES AND OUTPUTS-

```
> t.test(PremiumPrice~AnyTransplants,mu=0,conf=0.95,var.eq=F,Paired=F)

                Welch Two Sample t-test

data:  PremiumPrice by AnyTransplants

t = -6.2599, df = 56.504, p-value = 5.545e-08

alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0

95 percent confidence interval:

 -10382.274  -5349.081

sample estimates:

mean in group 0 mean in group 1

      23897.96        31763.64
```

TESTING-

Here we want to test if there is any significant difference between the average premium price for the people with any transplants and without any transplants is present or not.

Our Null hypothesis (H_0) is , there is no significant difference & our alternative hypothesis (H_1) is , there exists true difference between average premium prices for the people with any transplants and without any transplants.

i.e. $H_0 : \mu A - \mu B = 0$ VS. $H_1 : \mu A - \mu B \neq 0$.

Where μA= average premium price for the people with any transplants.

μB= average premium price for the people without any transplants..

As the p-value (=5.545e-08) < 0.05 so we reject H_0 at 5% level of significance , so we reject the null hypothesis and conclude that the alternative hypothesis is true.

So , there is a significant difference between means of premium prices for the people with any transplants and without any transplants . The predicted values of μA & μB are 31763.64 and 23897.96 respectively.

4.Test to check the significance of avg. premium price for people with any chronic disease and without any chronic disease:

CODES AND OUTPUTS-

```
> t.test(PremiumPrice~AnyChronicDiseases,mu=0,conf=0.95,var.eq=F,Paired=F)

        Welch Two Sample t-test

data:  PremiumPrice by AnyChronicDiseases

t = -7.7077, df = 311.78, p-value = 1.728e-13

alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0

95 percent confidence interval:

 -4251.769 -2522.455

sample estimates:

mean in group 0 mean in group 1

      23725.25        27112.36
```

TESTING-

Here we want to test if there is any significant difference between the average premium price for the people with and without any chronic diseases is present or not.

Our Null hypothesis (H_0) is , there is no significant difference & our alternative hypothesis (H_1) is , there exists true difference between average premium prices for the people with and without any chronic diseases.

i.e. $H_0 : \mu A - \mu B = 0$ VS. $H_1 : \mu A - \mu B \neq 0$.

Where μA= average premium price for the people with any chronic diseases.

μB= average premium price for the people without any chronic diseases.

As the p-value (=1.728e-13) < 0.05 so we reject H_0 at 5% level of significance , so we reject the null hypothesis and conclude that the alternative hypothesis is true.

So , there is a significant difference between means of premium prices for the people with and without any chronic diseases. The predicted values of μ A & μ B are 27112.36 and 23725.25 respectively.

5.Test to check the significance of avg. premium price for people having known allergies and not having known allergies:

CODES AND OUTPUTS-

```
> t.test(PremiumPrice~KnownAllergies,mu=0,conf=0.95,var.eq=F,Paired=F)

        Welch Two Sample t-test

data:  PremiumPrice by KnownAllergies

t = -0.36669, df = 320.56, p-value = 0.7141

alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0

95 percent confidence interval:

 -1171.0486   803.0997

sample estimates:

mean in group 0 mean in group 1

      24297.16        24481.13
```

TESTING-

Here we want to test if there is any significant difference between the average premium price for the people with allergies and without allergies is present or not.

Our Null hypothesis (H_0) is , there is no significant difference & our alternative hypothesis (H_1) is , there exists true difference between average premium prices for the people with allergies and without allergies.

i.e. $H_0 : \mu A - \mu B = 0$ VS. $H_1 : \mu A - \mu B \neq 0$.

Where μ A= average premium price for the people with allergies.

μ B= average premium price for the people without allergies.

As the p-value (=0.7141) > 0.05 so we reject H_0 at 5% level of significance , so we accept the null hypothesis and conclude that the alternative hypothesis is false.

So , there is no significant difference between means of premium prices for the people with allergies and without allergies . The predicted values of μ A & μ B are 24481.13 and 24297.16 respectively.

6.Test to check the significance of avg. premium price for people who have history of cancer in family and who have no history of cancer in family :

CODES AND OUTPUTS-

```
> t.test(PremiumPrice~HistoryOfCancerInFamily,mu=0,conf=0.95,var.eq=F,Paired=F)

        Welch Two Sample t-test

data:  PremiumPrice by HistoryOfCancerInFamily

t = -2.3568, df = 139.3, p-value = 0.01983

alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0

95 percent confidence interval:

 -2963.4186  -259.5699

sample estimates:

mean in group 0 mean in group 1

      24147.13        25758.62
```

TESTING-

Here we want to test if there is any significant difference between the average premium price for the people with and without a history of cancer is present or not.

Our Null hypothesis (H_0) is , there is no significant difference & our alternative hypothesis (H_1) is , there exists true difference between average premium prices for the people with and without a history of cancer.

i.e. $H_0 : \mu A - \mu B = 0$ VS. $H_1 : \mu A - \mu B \neq 0$.

Where μA= average premium price for the people with a history of cancer.

μB= average premium price for the people without a history of cancer.

As the p-value (=0.01983) < 0.05 so we reject H_0 at 5% level of significance , so we reject the null hypothesis and conclude that the alternative hypothesis is true.

So , there is a significant difference between means of premium prices for diabetic and non-diabetic people . The predicted values of μA & μB are 25758.62 and 24147.13 respectively.

•ANOVA Test of avg. premium price for people gone through no, 1, 2, 3 major surgeries:-

CODES AND OUTPUTS-

```
> aov(Data$PremiumPrice~factor(Data$NumberOfMajorSurgeries))
Call:
  aov(formula = Data$PremiumPrice ~ factor(Data$NumberOfMajorSurgeries))

Terms:
                factor(Data$NumberOfMajorSurgeries)   Residuals
Sum of Squares                           2843295212 35610915742
Deg. of Freedom                                   3        982

Residual standard error: 6021.932
```

```
> a<-aov(Data$PremiumPrice~factor(Data$NumberOfMajorSurgeries))
```

```
> summary(a)
                                             Df    Sum Sq       Mean Sq   F value   Pr(>F)
factor(Data$NumberOfMajorSurgeries)  3    2.843e+09    947765071   26.14    2.87e-16
Residuals                                    982    3.561e+10    36263662

factor(Data$NumberOfMajorSurgeries) ***
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

TESTING-

Here we want to test whether there is a significant difference between the average premium prices for the people who have gone through no, 1, 2, 3 major surgeries.

So our test can be looked upon as a one factor ANOVA test.

The null and alternative hypotheses of one-way ANOVA can be expressed as:

H0: $\mu1 = \mu2 = \mu3 = \mu4$ ("all 4 population means are equal")

H1: At least one μi different ("at least one of the 4 population means is not equal to the others")

where

μi is the population mean of the ith group (i = 1, 2, 3, 4)

The One-Way ANOVA is considered an omnibus (Latin for “all”) test because the F test indicates whether the model is significant overall—i.e., whether or not there are any significant differences in the means between any of the groups. (Stated another way, this says that at least one of the means is different from the others.) However, it does not indicate which mean is different. Determining which specific pairs of means are significantly different requires either contrasts or post hoc (Latin for “after this”) tests.

The test statistic for a One-Way ANOVA is denoted as F. For an independent variable with k groups, the F statistic evaluates whether the group means are significantly different. Because the computation

of the F statistic is slightly more involved than computing the paired or independent samples t test statistics, it's extremely common for all of the F statistic components to be depicted in a table like the following:

ANOVA TABLE:

SOURCE OF VARIATION	SUM OF SQUARES	D.F.	MEAN SQUARE	F-STATISTIC
TREATMENT	SSR=2843295212	$df_r = 3$	MSR=947765071	MSR/MSE=26.14
ERROR	SSE=35610915742	$df_e = 982$	MSE=36263662	--
TOTAL	TSS=38454210954	$df_t = 985$	--	--

SSR = the regression sum of squares, SSE = the error sum of squares, SST = the total sum of squares (SST = SSR + SSE), dfr = the model degrees of freedom (equal to dfr = k - 1) ,dfe = the error degrees of freedom (equal to dfe = n - k) ,k = the total number of groups (levels of the independent variable)

n = the total number of valid observations, dfT = the total degrees of freedom (equal to dfT = dfr + dfe = n - 1), MSR = SSR/dfr = the regression mean square, MSE = SSE/dfe = the mean square error

Then the F statistic itself is computed as

F=MSR/MSE

As the p-value (=2.87e-16) is less than 0.05 , so at 5% level of significance we conclude that the avg. premium prices are significant and we reject the null hypothesis and accept the alternative hypothesis.

❖ PREDICTION MODEL:

➢ *Splitting Data into two Subset for Training and Testing use:*

CODES AND OUTPUTS-

```
> sample <- sample.split(Data$PremiumPrice, SplitRatio = 0.75)

> train <- subset(Data, sample == TRUE)

> test <- subset(Data, sample == FALSE)

> dim(train)

[1] 743  11

> dim(test)

[1] 243  11
```

FINDINGS-

Here we have split our data randomly into two portions namely train and test. The train portion contains 75% & the test portion contains 25% of the overall data.

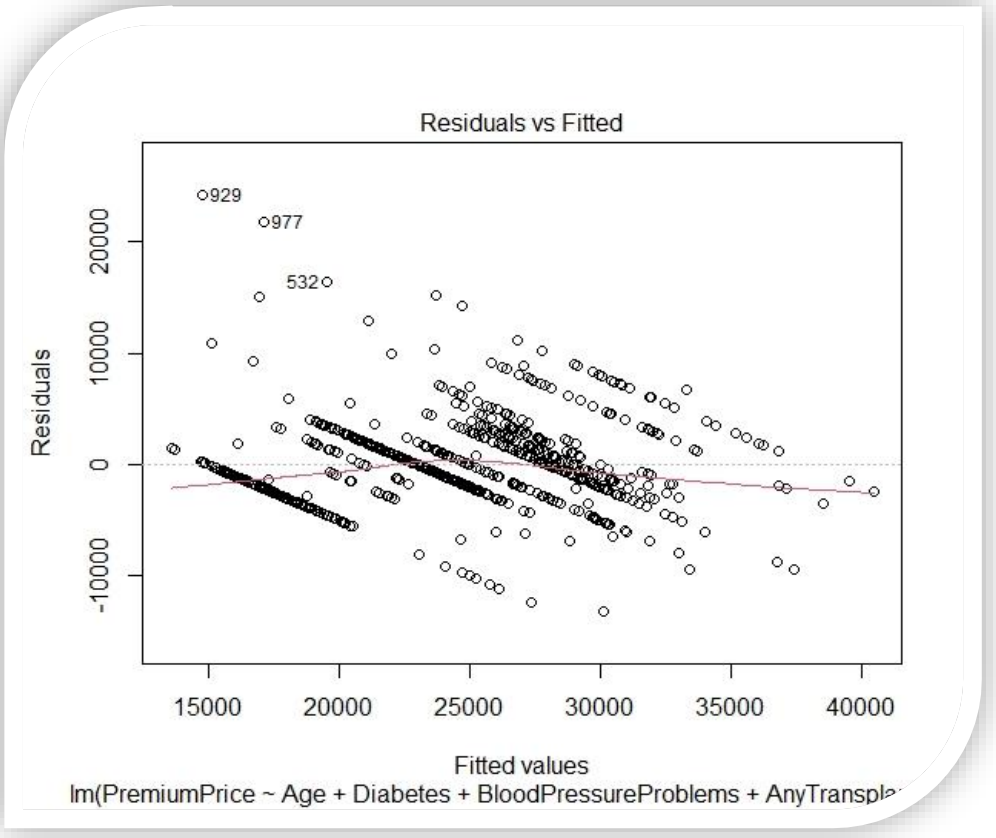
➢*Fitting Multiple Linear Regression Model:*

A. WITH ALL VARIABLES-

CODES AND OUTPUTS-

```
>multiple.regression<-
lm(PremiumPrice~Age+Diabetes+BloodPressureProblems+AnyTransplants+AnyChronicDiseases+Hei
ght+Weight+KnownAllergies+HistoryOfCancerInFamily+NumberOfMajorSurgeries,data = train)


> summary(multiple.regression)


Call:
lm(formula = PremiumPrice ~ Age + Diabetes + BloodPressureProblems +
    AnyTransplants + AnyChronicDiseases + Height + Weight + KnownAllergies +
    HistoryOfCancerInFamily + NumberOfMajorSurgeries, data = train)
Residuals:
    Min      1Q  Median      3Q     Max
-12794.0  -2211.5   -377.6   1904.8  24235.0
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           3571.639   2474.905   1.443  0.14941
Age                    327.495     11.539  28.382  < 2e-16 ***
Diabetes              -319.528    296.433  -1.078  0.28143
BloodPressureProblems   16.645    303.487   0.055  0.95628
AnyTransplants        7270.831    593.086  12.259  < 2e-16 ***
AnyChronicDiseases    2714.237    368.236   7.371 4.59e-13 ***
Height                   1.950     14.068   0.139  0.88982
Weight                  78.679      9.819   8.013 4.42e-15 ***
KnownAllergies         485.166    342.922   1.415  0.15755
```

```
HistoryOfCancerInFamily 1798.643    453.609   3.965 8.05e-05 ***
NumberOfMajorSurgeries  -633.876    222.271  -2.852  0.00447 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 3842 on 732 degrees of freedom
Multiple R-squared:  0.6305,    Adjusted R-squared:  0.6254
F-statistic: 124.9 on 10 and 732 DF,  p-value: < 2.2e-16

> sigma(multiple.regression)/mean(train$PremiumPrice)
[1] 0.1577148

> plot(multiple.regression)
```



(Figure – 12.a)

FINDINGS-

From the above fitting-

1. Model accuracy assessment:

As we have seen in simple linear regression, the overall quality of the model can be assessed by examining the R-squared (R2) and Residual Standard Error (RSE).

R-squared:
In multiple linear regression, the R2 represents the correlation coefficient between the observed values of the outcome variable (y) and the fitted (i.e., predicted) values of y. For this reason, the value of R will always be positive and will range from zero to one.

R2 represents the proportion of variance, in the outcome variable y, that may be predicted by knowing the value of the x variables. An R2 value close to 1 indicates that the model explains a large portion of the variance in the outcome variable.

A problem with the R2, is that, it will always increase when more variables are added to the model, even if those variables are only weakly associated with the response . A solution is to adjust the R2 by taking into account the number of predictor variables.
The adjustment in the "Adjusted R Square" value in the summary output is a correction for the number of x variables included in the prediction model.

So, in this model the value of R2 is equal to 0.6305 that means our model has a 63.05% accuracy. The value of the adjusted R2 is equal to 0.6254 which is close to our R2 value. This indicates that there might be some unimportant variables in our model.

2. Residual Standard Error (RSE):

The RSE estimate gives a measure of error of prediction. The lower the RSE, the more accurate the model (on the data in hand).

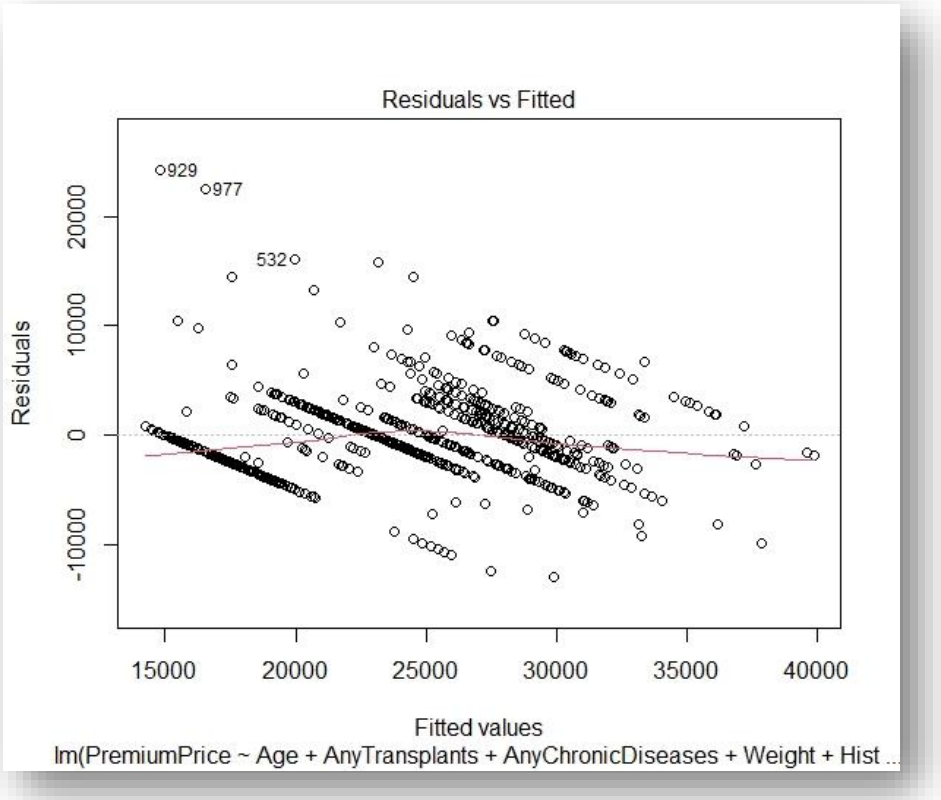
In this model the value of RSE is equal to 3842 corresponding to 15.77% error rate.

B. ELIMINATING VARIABLES WITH p-VALUE >0.05-

CODES AND OUTPUTS-

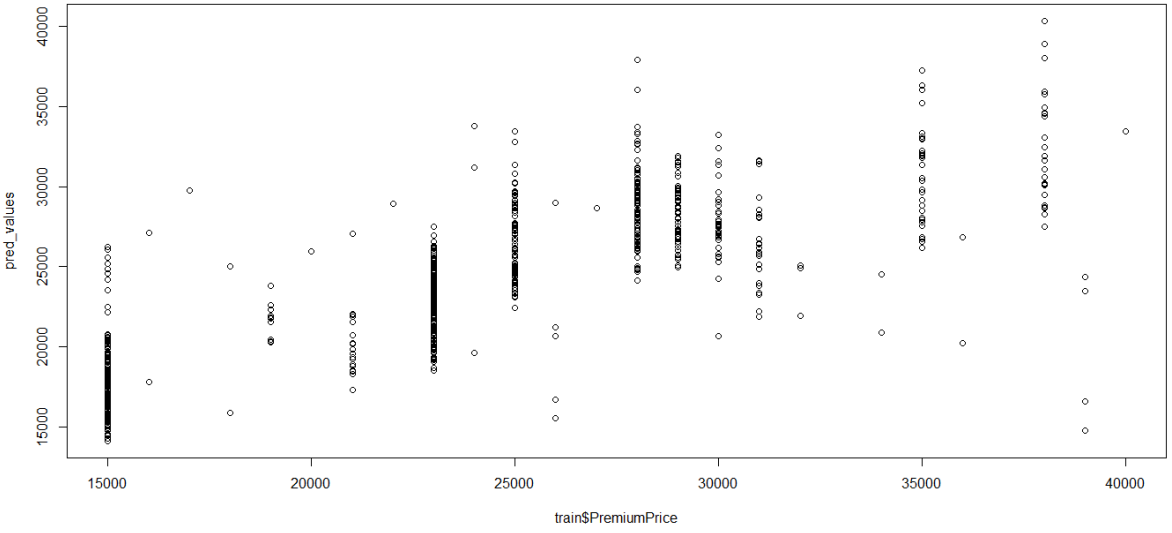
```
>multiple.regression1<-
lm(PremiumPrice~Age+AnyTransplants+AnyChronicDiseases+Weight+HistoryOfCancerInFamily+Nu
mberOfMajorSurgeries,data = train)

> summary(multiple.regression1)

Call:

lm(formula = PremiumPrice ~ Age + AnyTransplants + AnyChronicDiseases +

   Weight + HistoryOfCancerInFamily + NumberOfMajorSurgeries,

   data = train)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-12758.3  -2254.8   -347.7   1877.9  24224.0
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          3894.70     896.93   4.342 1.61e-05 ***
Age                   324.84      11.20  29.000  < 2e-16 ***
AnyTransplants       7305.01     591.73  12.345  < 2e-16 ***
AnyChronicDiseases   2761.72     365.25   7.561 1.19e-13 ***
Weight                 79.50       9.76   8.145 1.62e-15 ***
HistoryOfCancerInFamily  1889.16     450.55   4.193 3.09e-05 ***
NumberOfMajorSurgeries   -617.06     217.17  -2.841  0.00462 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 3840 on 736 degrees of freedom
Multiple R-squared:  0.6287,    Adjusted R-squared:  0.6257
F-statistic: 207.7 on 6 and 736 DF,  p-value: < 2.2e-16

> sigma(multiple.regression1)/mean(train$PremiumPrice)
[1] 0.1575975

> plot(multiple.regression1)
```



(Figure – 12.b)

```
> predict(multiple.regression1,train)
> pred_values=predict(multiple.regression1,data=train)
> train$pred_premium=pred_values
> plot(train$PremiumPrice,pred_values)
```



(Figure – 12.c)

```
> write_xlsx(train,path="F:\\PROJECT2022\\RRR\\TEST-ds.xlsx")
```

FINDINGS-

From the above fitting-

1. Model accuracy assessment:

As we have seen in simple linear regression, the overall quality of the model can be assessed by examining the R-squared (R2) and Residual Standard Error (RSE).

R-squared:
In multiple linear regression, the R2 represents the correlation coefficient between the observed values of the outcome variable (y) and the fitted (i.e., predicted) values of y. For this reason, the value of R will always be positive and will range from zero to one.

R2 represents the proportion of variance, in the outcome variable y, that may be predicted by knowing the value of the x variables. An R2 value close to 1 indicates that the model explains a large portion of the variance in the outcome variable.

A problem with the R2, is that, it will always increase when more variables are added to the model, even if those variables are only weakly associated with the response . A solution is to adjust the R2 by taking into account the number of predictor variables.
The adjustment in the "Adjusted R Square" value in the summary output is a correction for the number of x variables included in the prediction model.

So, in this model the value of R2 is equal to 0.6287 that means our model has a 62.87% accuracy. The value of the adjusted R2 is equal to 0.6257 which is very close to our R2 value. This indicates that our model is free of any unimportant variables.

2. Residual Standard Error (RSE):

The RSE estimate gives a measure of error of prediction. The lower the RSE, the more accurate the model (on the data in hand).

In this model the value of RSE is equal to 3840 corresponding to 15.76% error rate.

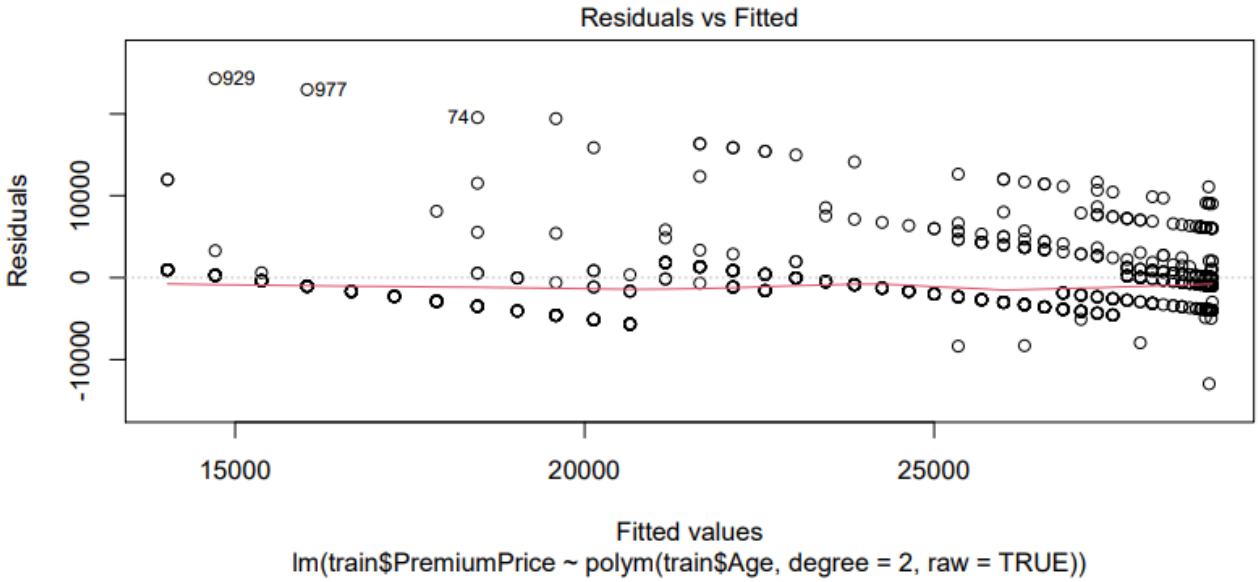
➢*Fitting Simple Polynomial Regression Model:*

1. BETWEEN PREMIUM PRICE & AGE-

CODES AND OUTPUTS-

```
>  PR1.1=lm(train$PremiumPrice ~ polym(train$Age,degree=2, raw=TRUE),data=train)

> summary(PR1.1)

Call:

lm(formula = train$PremiumPrice ~ polym(train$Age, degree = 2,
```

```
   raw = TRUE), data = train)
Residuals:
    Min      1Q  Median      3Q     Max
-12934.0  -2880.1   -826.2    878.9  24286.5
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -938.9668  1395.3441  -0.673    0.501
polym(train$Age, degree = 2, raw = TRUE)1  974.6207    71.2628  13.676   <2e-16
polym(train$Age, degree = 2, raw = TRUE)2   -7.9373     0.8409  -9.439   <2e-16

(Intercept)
polym(train$Age, degree = 2, raw = TRUE)1 ***
polym(train$Age, degree = 2, raw = TRUE)2 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 4264 on 740 degrees of freedom
Multiple R-squared:  0.5399,    Adjusted R-squared:  0.5386
F-statistic: 434.1 on 2 and 740 DF,  p-value: < 2.2e-16
>plot(PR1.1)
```



(Figure – 13.a)

FINDINGS-

(1) In this model the value of R2 is equal to 0.5399. The value of the adjusted R2 is equal to 0.5386 which is not that close to our R2 value.

(2) The p-value is less than 2.2e-16 which is less than 0.05 , so we will take the variable Age in our multiple polynomial model.

2.BETWEEN PREMIUM PRICE & DIABETES-

CODES AND OUTPUTS-

> PR1.2=lm(train$PremiumPrice ~ polym(train$Diabetes,degree=2, raw=TRUE),data=train)

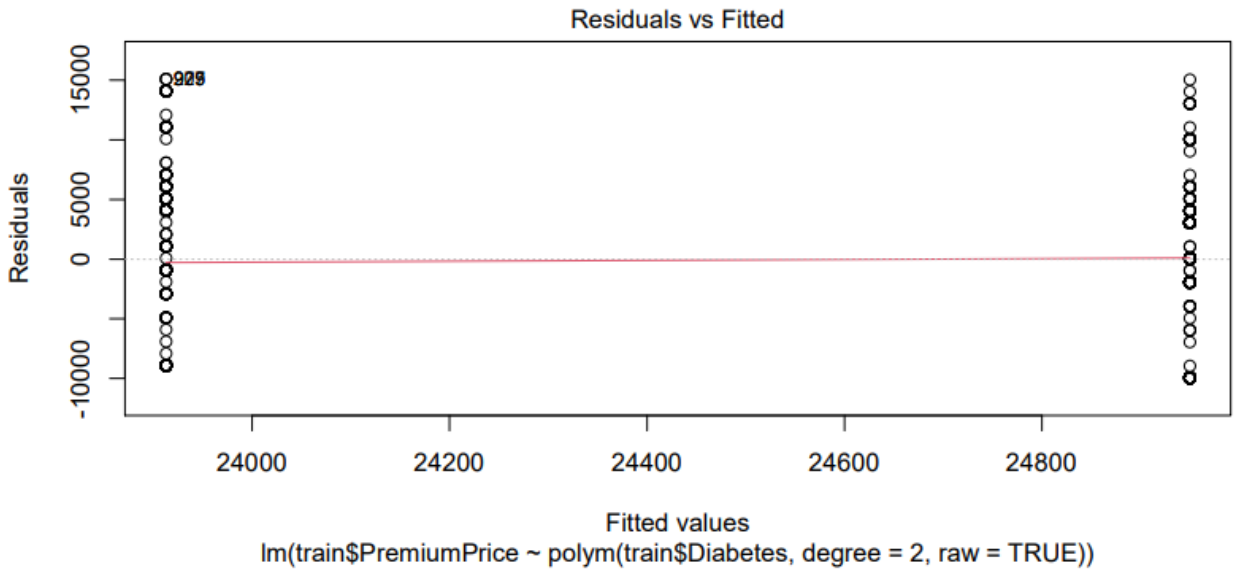
> summary(PR1.2)

Call:lm(formula = train$PremiumPrice ~ polym(train$Diabetes, degree = 2, raw = TRUE), data = train)

Residuals:

Min 1Q Median 3Q Max

```
       -9949.7 -2912.9  -912.9  4087.1 15087.1

Coefficients: (1 not defined because of singularities)
                                    Estimate Std. Error t value
(Intercept)                          23912.9      303.7  78.744
polym(train$Diabetes, degree = 2, raw = TRUE)1   1036.7      464.2   2.233
polym(train$Diabetes, degree = 2, raw = TRUE)2       NA         NA      NA
                                    Pr(>|t|)
(Intercept)                          <2e-16 ***
polym(train$Diabetes, degree = 2, raw = TRUE)1   0.0258 *
polym(train$Diabetes, degree = 2, raw = TRUE)2       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 6261 on 741 degrees of freedom
Multiple R-squared:  0.006687,  Adjusted R-squared:  0.005346
F-statistic: 4.988 on 1 and 741 DF,  p-value: 0.02582
>plot(PR1.2)
```



(Figure – 13.b)

FINDINGS-

(1) In this model the value of R2 is equal to 0.006687. The value of the adjusted R2 is equal to 0.005346 which is not that close to our R2 value.

(2)The p-value is equals to 0.02582 which is less than 0.05 , so we will take the variable Diabetes in our multiple polynomial model.

3.BETWEEN PREMIUM PRICE &BLOOD PRESSURE PROBLEMS-

CODES AND OUTPUTS-

```
>PR1.3=lm(train$PremiumPrice~polym(train$BloodPressureProblems,degree=2,raw=TRUE),data=tra
in)

> summary(PR1.3)


Call:

lm(formula = train$PremiumPrice ~ polym(train$BloodPressureProblems,

        degree = 2, raw = TRUE), data = train)


Residuals:

      Min       1Q   Median       3Q      Max

-10531.6  -2531.6   -321.5   3468.4  15678.5


    Coefficients: (1 not defined because of singularities)

                                                  Estimate Std. Error

    (Intercept)                                    23321.5      311.1

    polym(train$BloodPressureProblems, degree = 2, raw = TRUE)1   2210.1      454.6

    polym(train$BloodPressureProblems, degree = 2, raw = TRUE)2       NA         NA

                                                  t value Pr(>|t|)

    (Intercept)                                    74.956  < 2e-16

    polym(train$BloodPressureProblems, degree = 2, raw = TRUE)1   4.861 1.42e-06

    polym(train$BloodPressureProblems, degree = 2, raw = TRUE)2      NA       NA


    (Intercept)                                   ***
```

polym(train$BloodPressureProblems, degree = 2, raw = TRUE)1 ***

polym(train$BloodPressureProblems, degree = 2, raw = TRUE)2

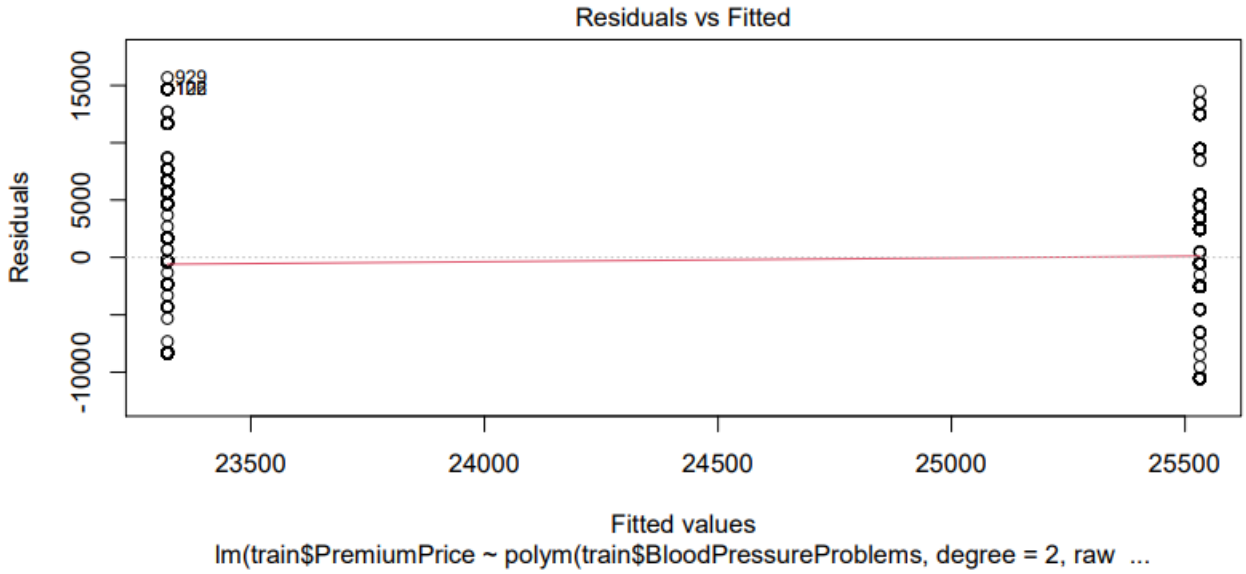
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6184 on 741 degrees of freedom

Multiple R-squared: 0.03091, Adjusted R-squared: 0.0296

F-statistic: 23.63 on 1 and 741 DF, p-value: 1.424e-06

>plot(PR1.3)



(Figure – 13.c)

FINDINGS-

(1) In this model the value of R2 is equal to 0.03091. The value of the adjusted R2 is equal to 0.0296 which is not that close to our R2 value.

(2) The p-value is equals to 1.424e-06 which is less than 0.05 , so we will take the variable BloodPressureproblems in our multiple polynomial model.

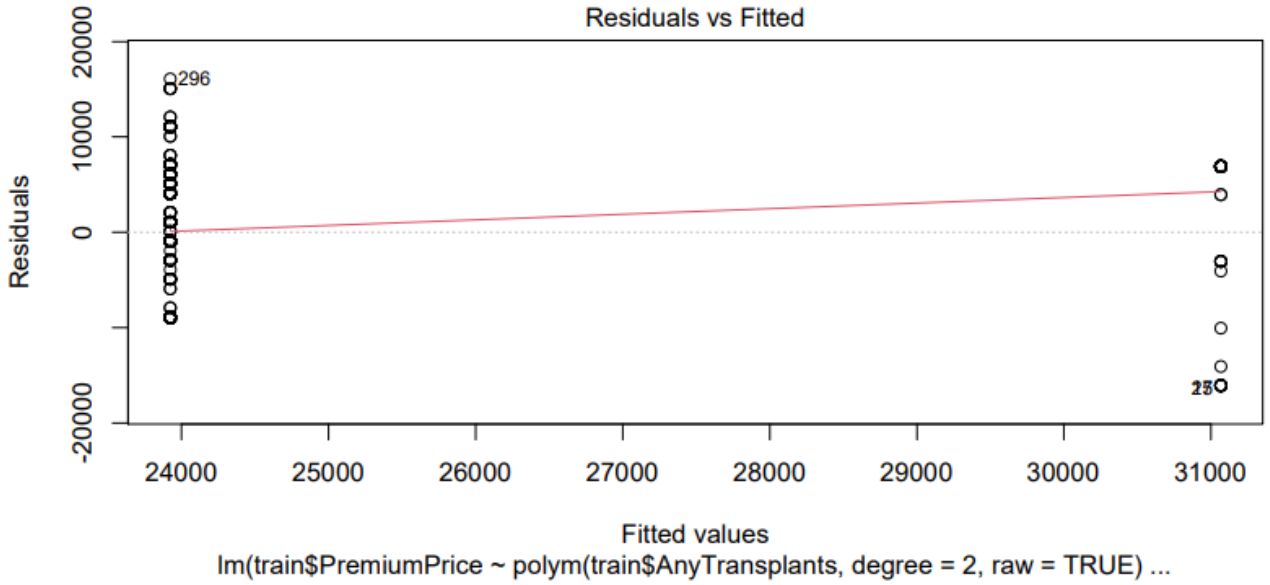
4.BETWEEN PREMIUM PRICE & ANY TRANSPLANTS-

CODES AND OUTPUTS-

> PR1.4=lm(train$PremiumPrice ~ polym(train$AnyTransplants,degree=2, raw=TRUE),data=train)

```
> summary(PR1.4)

Call:
lm(formula = train$PremiumPrice ~ polym(train$AnyTransplants,
    degree = 2, raw = TRUE), data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-16066.7  -2924.1   -924.1   4075.9  16075.9

Coefficients: (1 not defined because of singularities)
                                        Estimate Std. Error
(Intercept)                              23924.1      228.8
polym(train$AnyTransplants, degree = 2, raw = TRUE)1   7142.6      929.8
polym(train$AnyTransplants, degree = 2, raw = TRUE)2       NA         NA
                                        t value Pr(>|t|)
(Intercept)                             104.552  < 2e-16 ***
polym(train$AnyTransplants, degree = 2, raw = TRUE)1   7.682 4.97e-14 ***
polym(train$AnyTransplants, degree = 2, raw = TRUE)2      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6045 on 741 degrees of freedom
Multiple R-squared:  0.07376,   Adjusted R-squared:  0.07251
F-statistic: 59.01 on 1 and 741 DF,  p-value: 4.972e-14
```

>plot(PR1.4)



(Figure – 13.d)

FINDINGS-

(1) In this model the value of R2 is equal to 0.07376. The value of the adjusted R2 is equal to 0.07251 which is not that close to our R2 value.

(2) The p-value is equals to 4.972e-14 which is less than 0.05 , so we will take the variable AnyTransplants in our multiple polynomial model.

5.BETWEEN PREMIUM PRICE & ANY CHRONIC DISEASES-

CODES AND OUTPUTS-

>PR1.5=lm(train$PremiumPrice~polym(train$AnyChronicDiseases,degree=2, raw=TRUE),data=train)

> summary(PR1.5)

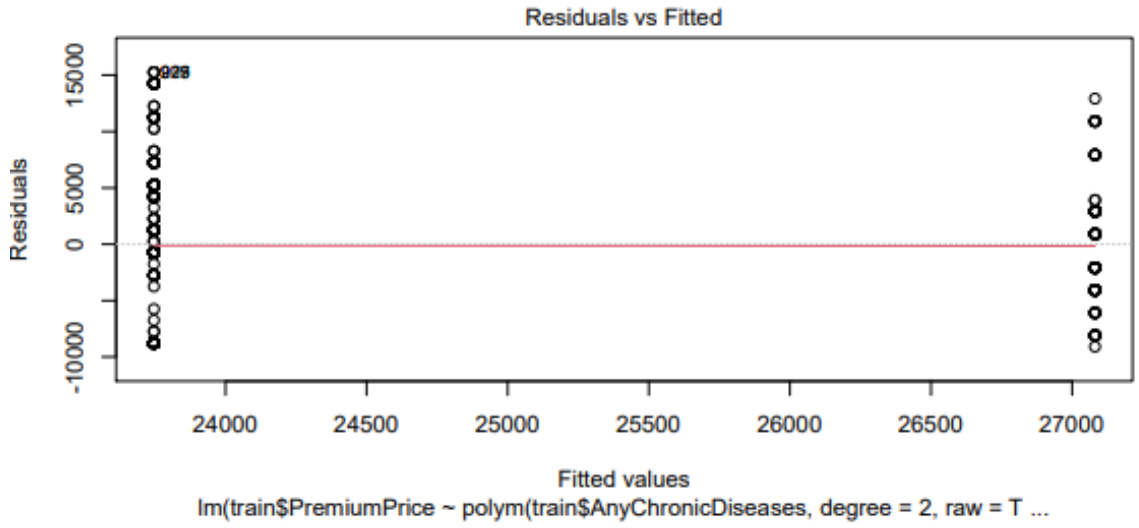
Call:

lm(formula = train$PremiumPrice ~ polym(train$AnyChronicDiseases,

degree = 2, raw = TRUE), data = train)

Residuals:

```
   Min     1Q  Median     3Q    Max
-9080.9 -4080.9  -746.3  4253.7 15253.7

Coefficients: (1 not defined because of singularities)
                                          Estimate Std. Error
(Intercept)                                23746.3      249.5
polym(train$AnyChronicDiseases, degree = 2, raw = TRUE)1   3334.6      583.2
polym(train$AnyChronicDiseases, degree = 2, raw = TRUE)2       NA         NA
                                          t value Pr(>|t|)
(Intercept)                                95.169  < 2e-16 ***
polym(train$AnyChronicDiseases, degree = 2, raw = TRUE)1   5.718 1.57e-08 ***
polym(train$AnyChronicDiseases, degree = 2, raw = TRUE)2      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6147 on 741 degrees of freedom
Multiple R-squared:  0.04225,   Adjusted R-squared:  0.04096
F-statistic: 32.69 on 1 and 741 DF,  p-value: 1.566e-08
> plot(PR1.5)
```



(Figure – 13.e)

FINDINGS-

(1) In this model the value of R2 is equal to 0.04225. The value of the adjusted R2 is equal to 0.04096 which is not that close to our R2 value.

(2) The p-value is equals to 1.566e-08 which is less than 0.05 , so we will take the variable AnyChronicDiseases in our multiple polynomial model.

6.BETWEEN PREMIUM PRICE & HEIGHT-

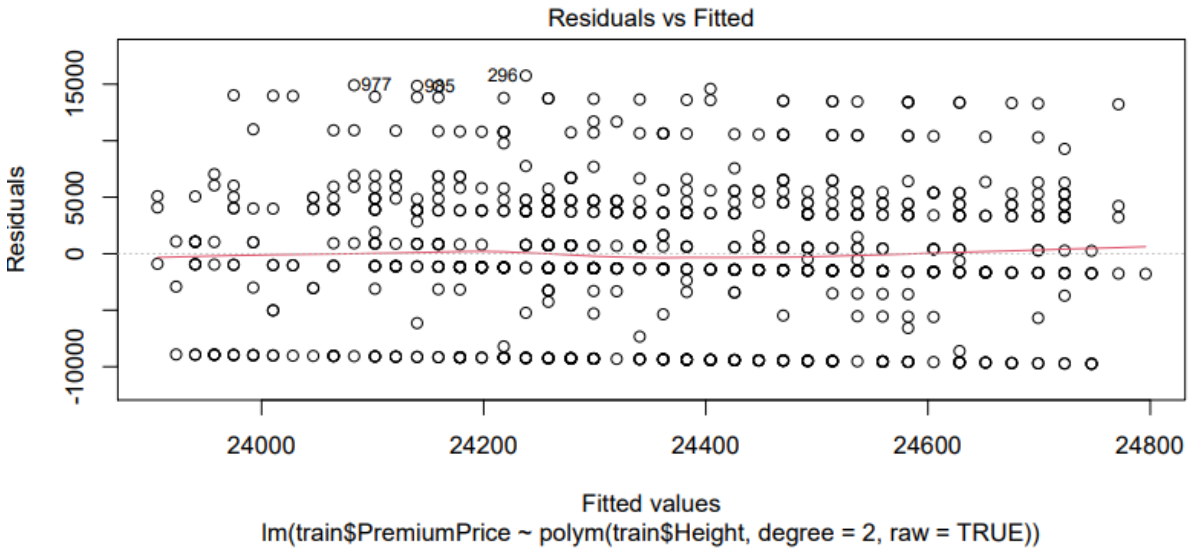
CODES AND OUTPUTS-

```
> PR1.6=lm(train$PremiumPrice ~ polym(train$Height,degree=2, raw=TRUE),data=train)

> summary(PR1.6)


Call:

lm(formula = train$PremiumPrice ~ polym(train$Height, degree = 2,

    raw = TRUE), data = train)


Residuals:

   Min     1Q Median     3Q    Max

 -9747  -3046   -975   3972  15762


Coefficients:

                                                Estimate Std. Error t value

(Intercept)                                     2.337e+04  5.714e+04   0.409

polym(train$Height, degree = 2, raw = TRUE)1 -9.463e+00  6.843e+02  -0.014

polym(train$Height, degree = 2, raw = TRUE)2  9.054e-02  2.044e+00   0.044

                                               Pr(>|t|)

(Intercept)                                       0.683

polym(train$Height, degree = 2, raw = TRUE)1      0.989

polym(train$Height, degree = 2, raw = TRUE)2      0.965


Residual standard error: 6282 on 740 degrees of freedom
```

Multiple R-squared: 0.001126, Adjusted R-squared: -0.001574

F-statistic: 0.4171 on 2 and 740 DF, p-value: 0.6591

> plot(PR1.6)



(Figure – 13.f)

FINDINGS-

(1) In this model the value of R2 is equal to 0.07193. The value of the adjusted R2 is equal to 0.06942 which is not that close to our R2 value.

(2) The p-value is equals to 0.6591 which is greater than 0.05 , so we will not take the variable Height in our multiple polynomial model to get a better model with lesser difference between multiple R2 & adjusted R2.

7.BETWEEN PREMIUM PRICE AND WEIGHT-

CODES AND OUTPUTS-

> PR1.7=lm(train$PremiumPrice ~ polym(train$Weight,degree=2, raw=TRUE),data=train)

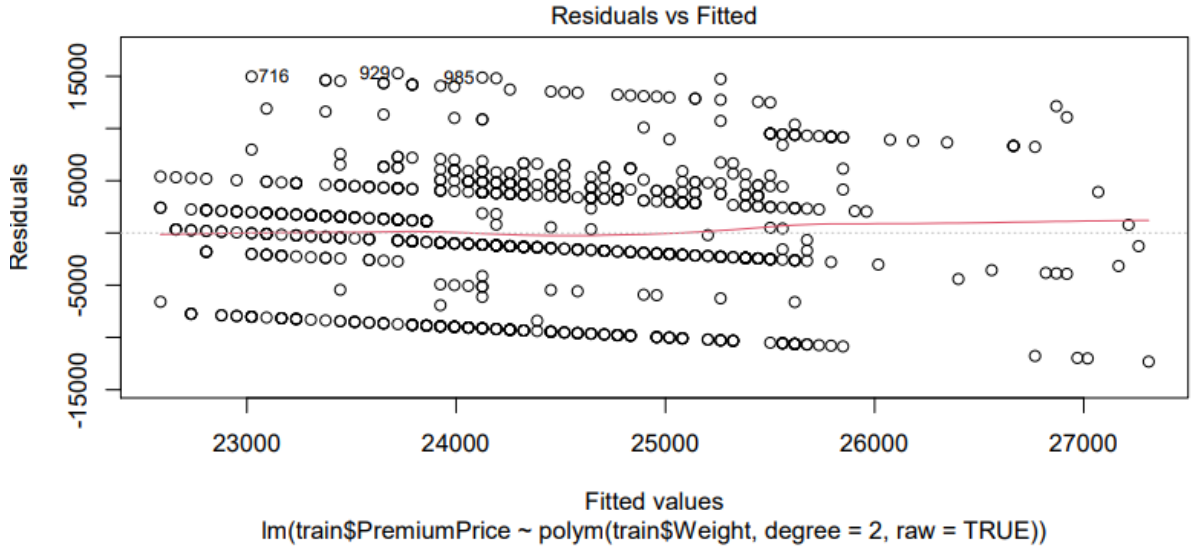
> summary(PR1.7)

Call:

lm(formula = train$PremiumPrice ~ polym(train$Weight, degree = 2,

raw = TRUE), data = train)

```
Residuals:
    Min      1Q  Median      3Q     Max
-12310.4  -2572.4    -22.3   4278.4  15278.4

Coefficients:
                                    Estimate Std. Error t value
(Intercept)                         18396.7049  5019.1951   3.665
polym(train$Weight, degree = 2, raw = TRUE)1    90.6915   124.3180   0.730
polym(train$Weight, degree = 2, raw = TRUE)2    -0.1674     0.7545  -0.222
                                   Pr(>|t|)
(Intercept)                         0.000265 ***
polym(train$Weight, degree = 2, raw = TRUE)1 0.465919
polym(train$Weight, degree = 2, raw = TRUE)2 0.824480
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6218 on 740 degrees of freedom
Multiple R-squared:  0.0214,    Adjusted R-squared:  0.01875
F-statistic:  8.09 on 2 and 740 DF,  p-value: 0.0003346
> plot(PR1.7)
```



(Figure – 13.g)

FINDINGS-

(1) In this model the value of R2 is equal to 0.0214. The value of the adjusted R2 is equal to 0.01875 which is not that close to our R2 value.

(2) The p-value is equals to 0.0003346 which is less than 0.05 , so we will take the variable Weight in our multiple polynomial model.

8.BETWEEN PREMIUM PRICE & KNOWN ALLERGIES-

CODES AND OUTPUTS-

```
> PR1.8=lm(train$PremiumPrice ~ polym(train$KnownAllergies,degree=2, raw=TRUE),data=train)
> summary(PR1.8)


Call:
lm(formula = train$PremiumPrice ~ polym(train$KnownAllergies,
    degree = 2, raw = TRUE), data = train)


Residuals:
   Min     1Q Median     3Q    Max
-9808  -3226  -1226   3774  15774
```

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	24225.7	261.5	92.629	<2e-16	***
polym(train$KnownAllergies, degree = 2, raw = TRUE)1	582.7	551.7	1.056	0.291	
polym(train$KnownAllergies, degree = 2, raw = TRUE)2	NA	NA	NA	NA	

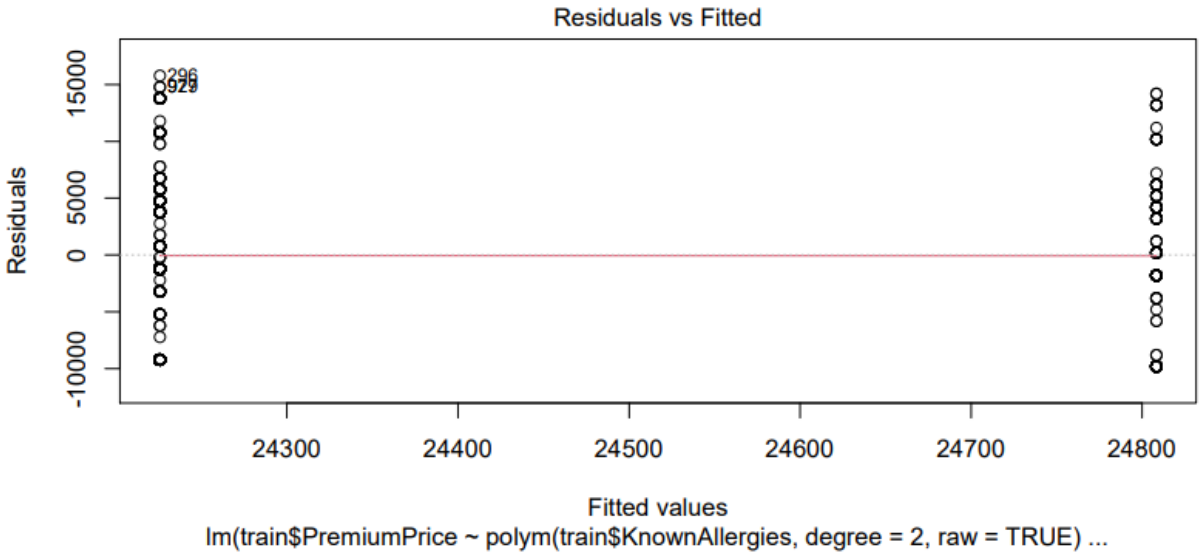
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6277 on 741 degrees of freedom

Multiple R-squared: 0.001503, Adjusted R-squared: 0.0001559

F-statistic: 1.116 on 1 and 741 DF, p-value: 0.2912

> plot(PR1.8)



(Figure – 13.h)

FINDINGS-

(1) In this model the value of R2 is equal to 0.001503. The value of the adjusted R2 is equal to 0.0001559 which is not that close to our R2 value.

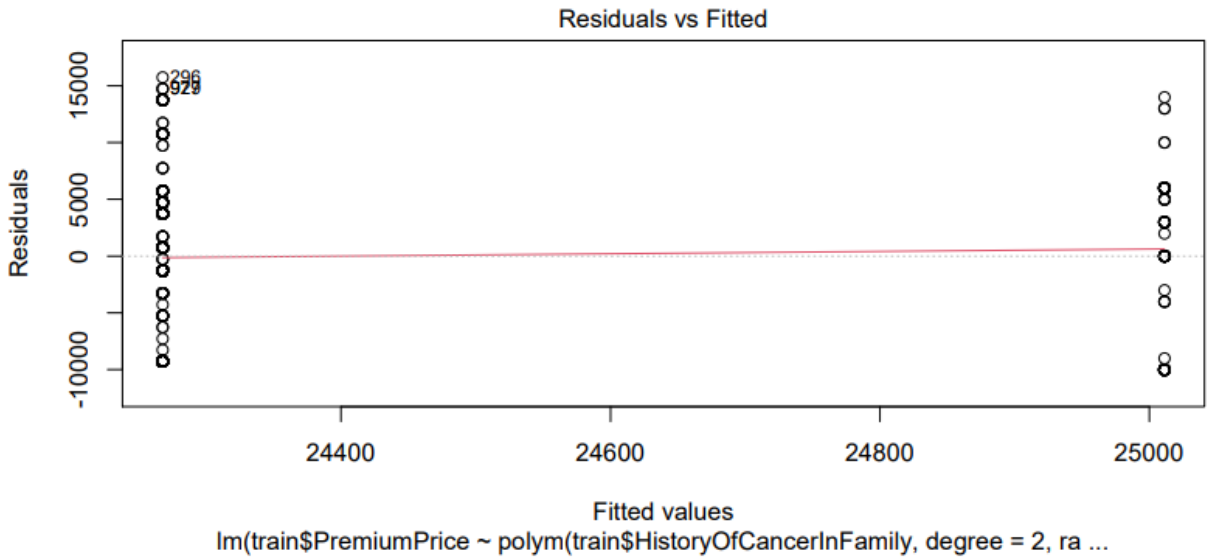
(2) The p-value is equals to 0.2912 which is greater than 0.05 , so we will not take the variable KnownAllergies in our multiple polynomial model to get a better model with lesser difference between multiple R2 & adjusted R2.

9.BETWEEN PREMIUM PRICE & HISTORY OF CANCER IN FAMILY-

CODES AND OUTPUTS-

```
>PR1.9=lm(train$PremiumPrice~polym(train$HistoryOfCancerInFamily,degree=2,raw=TRUE),dat
a=train)

> summary(PR1.9)


Call:

lm(formula = train$PremiumPrice ~ polym(train$HistoryOfCancerInFamily,

   degree = 2, raw = TRUE), data = train)


Residuals:

  Min    1Q Median    3Q   Max

-10011  -3268  -1268   3732  15732


Coefficients: (1 not defined because of singularities)

                                                    Estimate

(Intercept)                                          24267.6

polym(train$HistoryOfCancerInFamily, degree = 2, raw = TRUE)1    743.7

polym(train$HistoryOfCancerInFamily, degree = 2, raw = TRUE)2      NA

                                                    Std. Error

(Intercept)                                            245.4

polym(train$HistoryOfCancerInFamily, degree = 2, raw = TRUE)1     709.2

polym(train$HistoryOfCancerInFamily, degree = 2, raw = TRUE)2       NA

                                                    t value Pr(>|t|)

(Intercept)                                          98.871   <2e-16

polym(train$HistoryOfCancerInFamily, degree = 2, raw = TRUE)1   1.049    0.295
```

```
polym(train$HistoryOfCancerInFamily, degree = 2, raw = TRUE)2      NA       NA
(Intercept)                                                       ***
polym(train$HistoryOfCancerInFamily, degree = 2, raw = TRUE)1
polym(train$HistoryOfCancerInFamily, degree = 2, raw = TRUE)2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 6277 on 741 degrees of freedom
Multiple R-squared:  0.001482,  Adjusted R-squared:  0.0001342
F-statistic:   1.1 on 1 and 741 DF,  p-value: 0.2947
> plot(PR1.9)
```



(Figure – 13.i)

FINDINGS-

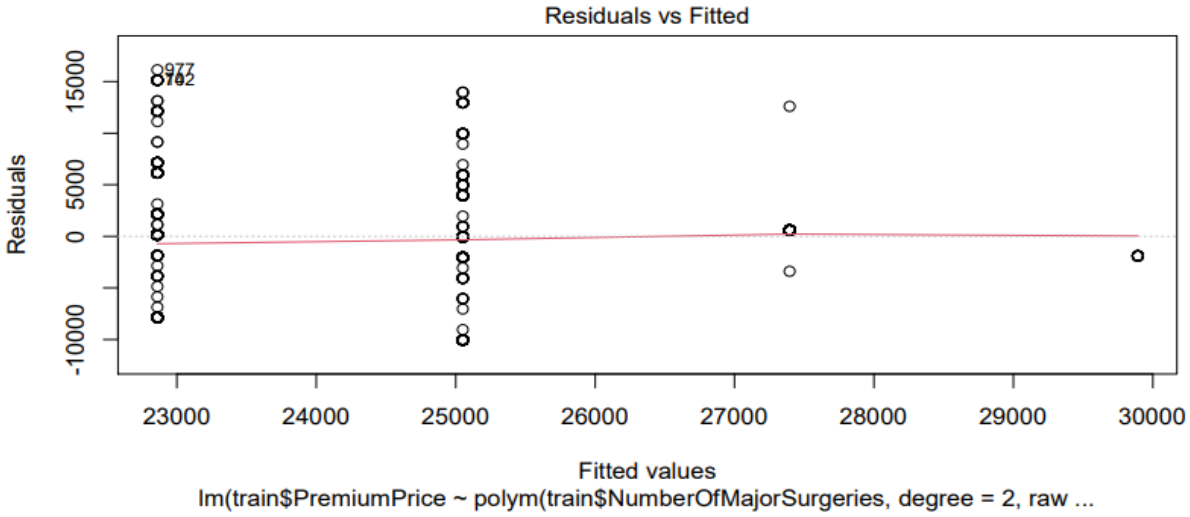
(1) In this model the value of R2 is equal to 0.001482. The value of the adjusted R2 is equal to 0.0001342 which is not that close to our R2 value.

(2) The p-value is equals to 0.2947 which is greater than 0.05 , so we will not take the variable HistoryOfCancerInFamily in our multiple polynomial model to get a better model with lesser difference between multiple R2 & adjusted R2.

10.BETWEEN PREMIUM PRICE & NUMBER OF MAJOR SURGERIES-

CODES AND OUTPUTS-

```
>PR1.10=lm(train$PremiumPrice~polym(train$NumberOfMajorSurgeries,degree=2,raw=TRUE),d
ata=train)

> summary(PR1.10)

Call:

lm(formula = train$PremiumPrice ~ polym(train$NumberOfMajorSurgeries,

    degree = 2, raw = TRUE), data = train)

Residuals:

     Min       1Q   Median       3Q      Max

-10050.6  -2050.6    139.9   2139.9  16139.9

Coefficients:

                                                    Estimate

(Intercept)                                          22860.05

polym(train$NumberOfMajorSurgeries, degree = 2, raw = TRUE)1  2113.65

polym(train$NumberOfMajorSurgeries, degree = 2, raw = TRUE)2    76.87

                                                    Std. Error t value

(Intercept)                                           318.94  71.676

polym(train$NumberOfMajorSurgeries, degree = 2, raw = TRUE)1    749.10   2.822

polym(train$NumberOfMajorSurgeries, degree = 2, raw = TRUE)2    350.71   0.219

                                                    Pr(>|t|)

(Intercept)                                          < 2e-16 ***

polym(train$NumberOfMajorSurgeries, degree = 2, raw = TRUE)1  0.00491 **

polym(train$NumberOfMajorSurgeries, degree = 2, raw = TRUE)2  0.82656

---

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

Residual standard error: 6056 on 740 degrees of freedom

Multiple R-squared:  0.07193,   Adjusted R-squared:  0.06942

F-statistic: 28.67 on 2 and 740 DF,  p-value: 1.013e-12

> plot(PR1.10)
```



(Figure – 13.j)

FINDINGS-

(1) In this model the value of R2 is equal to 0.07193. The value of the adjusted R2 is equal to 0.06942 which is not that close to our R2 value.

(2) The p-value is equals to 1.013e-12 which is less than 0.05 , so we will take the variable NumberOfMajorSurgeries in our multiple polynomial model.

➢*Fitting Multiple Polynomial Regression Model:-*

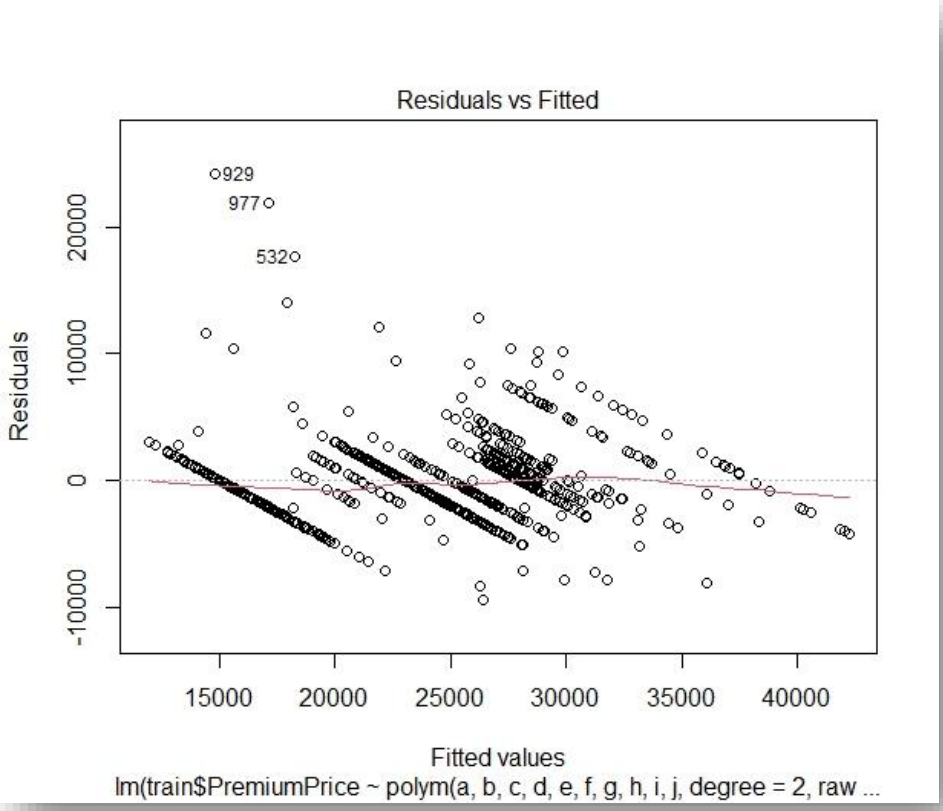
- ***POLYNOMIAL OF DEGREE 2***

A. WITH ALL VARIABLES-

CODES AND OUTPUTS-

```
>PR1=lm(PremiumPrice~polym(Age,Diabetes,BloodPressureProblems,AnyTransplants,AnyChronicDiseases,Height,Weight,KnownAllergies,HistoryOfCancerInFamily,NumberOfMajorSurgeries,degree=2,raw=TRUE),data=train)

> summary(PR1)

Call:

lm(formula = PremiumPrice ~ polym(Age, Diabetes, BloodPressureProblems,

    AnyTransplants, AnyChronicDiseases, Height, Weight, KnownAllergies,
```

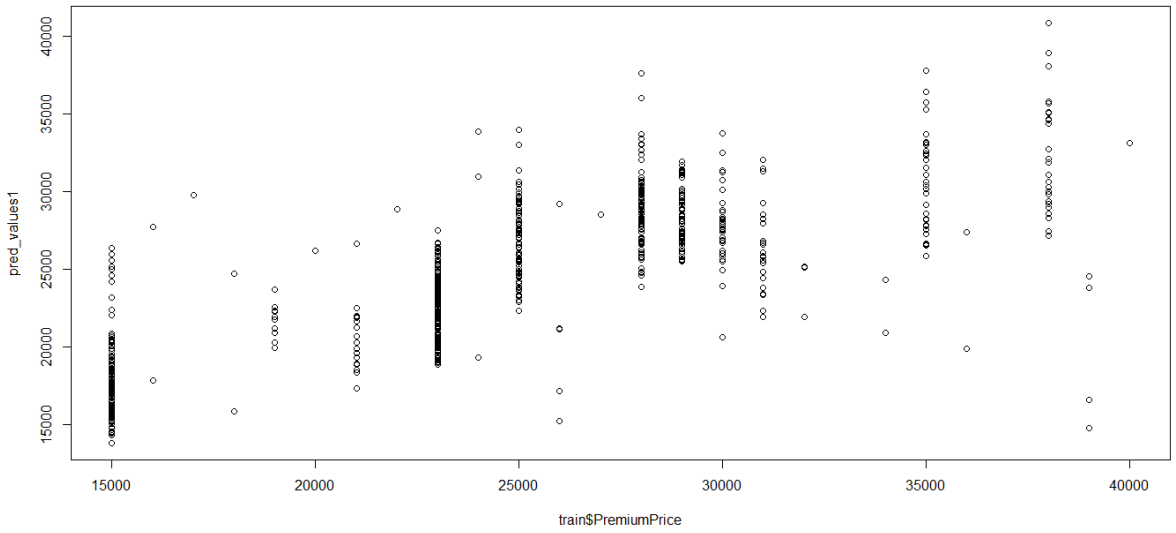
```
   HistoryOfCancerInFamily, NumberOfMajorSurgeries, degree = 2,
   raw = TRUE), data = train)
Residuals:
    Min      1Q  Median      3Q     Max
-12507.6  -1810.2   -186.3   1289.3  24674.6
Residual standard error: 3389 on 683 degrees of freedom
Multiple R-squared:  0.7316,    Adjusted R-squared:  0.7085
F-statistic: 31.56 on 59 and 683 DF,  p-value: < 2.2e-16
> sigma(PR1)/mean(train$PremiumPrice)
[1] 0.1391576
>plot(PR1)
```



(Figure – 14.a)

```
> predict(PR1,train)

> pred_values=predict(PR1,data=train)
```

```
> train$pred_premium=pred_values1

> plot(train$PremiumPrice,pred_values1)
```



(Figure – 14.b)

```
> write_xlsx(train,path="F:\\PROJECT2022\\RRR\\TEST-ds.xlsx")
```

FINDINGS-

From the above fitting-

1. Model accuracy assessment:

As we have seen in simple polynomial regression, the overall quality of the model can be assessed by examining the R-squared (R2) and Residual Standard Error (RSE).

R-squared:
In multiple polynomial regression, the R2 represents the correlation coefficient between the observed values of the outcome variable (y) and the fitted (i.e., predicted) values of y. For this reason, the value of R will always be positive and will range from zero to one.

R2 represents the proportion of variance, in the outcome variable y, that may be predicted by knowing the value of the x variables. An R2 value close to 1 indicates that the model explains a large portion of the variance in the outcome variable.

A problem with the R2, is that, it will always increase when more variables are added to the model, even if those variables are only weakly associated with the response . A solution is to adjust the R2 by taking into account the number of predictor variables.
The adjustment in the “Adjusted R Square” value in the summary output is a correction for the number of x variables included in the prediction model.

So, in this model the value of R2 is equal to 0.7316 that means our model has a 73.16% accuracy. The value of the adjusted R2 is equal to 0.7085 which is not that close to our R2 value. This indicates that there might be some unimportant variables in our model.

2. Residual Standard Error (RSE):

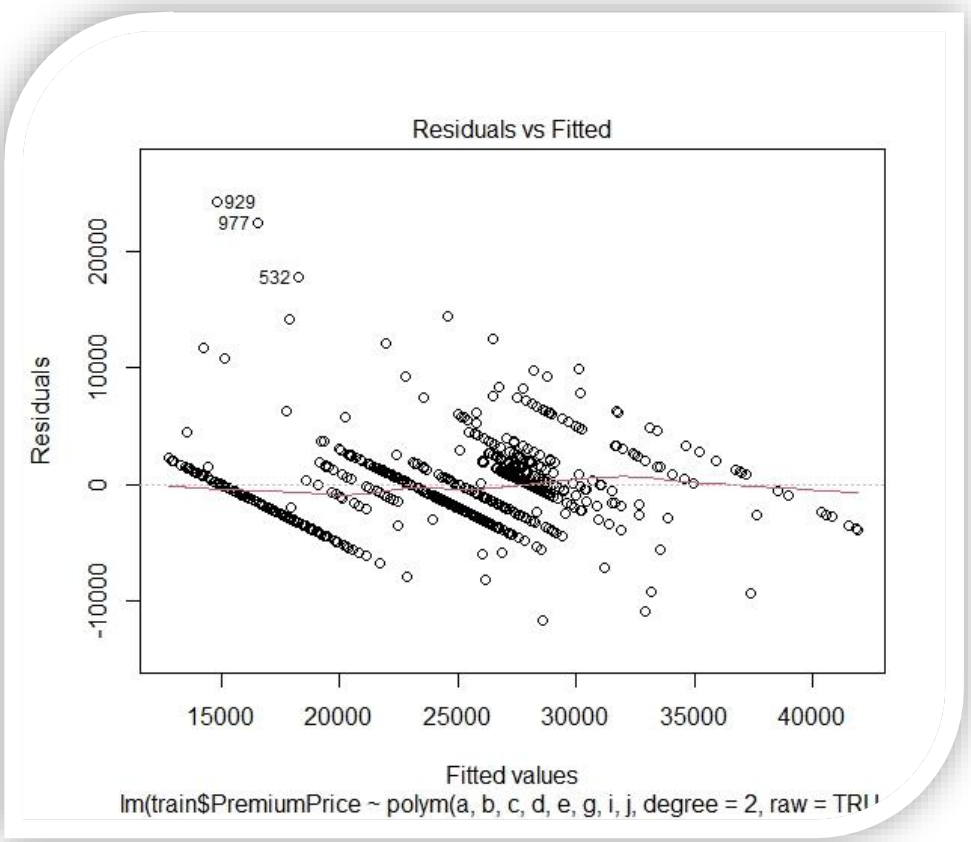
The RSE estimate gives a measure of error of prediction. The lower the RSE, the more accurate the model (on the data in hand).

In this model the value of RSE is equal to 3389 corresponding to 13.91576% error rate.

B.ELIMINATING VARIABLES WITH p-VALUE > 0.05-

CODES AND OUTPUTS-

```
>PR2=lm(PremiumPrice~polym(Age,Diabetes,BloodPressureProblems,AnyTransplants,AnyChronicDiseases,Weight,NumberOfMajorSurgeries,degree=2,raw=TRUE),data=train)

> summary(PR2)

Call:

lm(formula = PremiumPrice ~ polym(Age, Diabetes, BloodPressureProblems,

    AnyTransplants, AnyChronicDiseases, Weight, NumberOfMajorSurgeries,

    degree = 2, raw = TRUE), data = train)


Residuals:

     Min       1Q   Median       3Q      Max

-13287.1  -1985.3   -207.4   1493.9  24364.8

Residual standard error: 3487 on 711 degrees of freedom

Multiple R-squared:  0.7043,    Adjusted R-squared:  0.6914

F-statistic: 54.62 on 31 and 711 DF,  p-value: < 2.2e-16


> sigma(PR2)/mean(train$PremiumPrice)

[1] 0.1431769

>plot(PR2)
```



(Figure – 14.c)

FINDINGS-

From the above fitting-

1. Model accuracy assessment:

As we have seen in simple polynomial regression, the overall quality of the model can be assessed by examining the R-squared (R2) and Residual Standard Error (RSE).

R-squared:
In multiple polynomial regression, the R2 represents the correlation coefficient between the observed values of the outcome variable (y) and the fitted (i.e., predicted) values of y. For this reason, the value of R will always be positive and will range from zero to one.

R2 represents the proportion of variance, in the outcome variable y, that may be predicted by knowing the value of the x variables. An R2 value close to 1 indicates that the model explains a large portion of the variance in the outcome variable.

A problem with the R2, is that, it will always increase when more variables are added to the model, even if those variables are only weakly associated with the response . A solution is to adjust the R2 by taking into account the number of predictor variables.

The adjustment in the "Adjusted R Square" value in the summary output is a correction for the number of x variables included in the prediction model.

So, in this model the value of R2 is equal to 0.7043 that means our model has a 70.43% accuracy. The value of the adjusted R2 is equal to 0.6914 which is very close to our R2 value. This indicates that our model is free of any unimportant variables.

2. Residual Standard Error (RSE):

The RSE estimate gives a measure of error of prediction. The lower the RSE, the more accurate the model (on the data in hand).

In this model the value of RSE is equal to 3487 corresponding to 14.31769% error rate.

- ***POLYNOMIAL OF DEGREE 4-***

A. WITH ALL VARIABLES-

CODES AND OUTPUTS-

>PR3=lm(PremiumPrice~polym(Age,Diabetes,BloodPressureProblems,AnyTransplants,AnyChronicDiseases,Height,Weight,KnownAllergies,HistoryOfCancerInFamily,NumberOfMajorSurgeries,degree=4,raw=TRUE),data=train)

> summary(PR3)

Call:lm(formula=train$PremiumPrice~polym(Age,Diabetes,BloodPressureProblems,AnyTransplants,AnyChronicDiseases,Height,Weight,KnownAllergies,HistoryOfCancerInFamily,NumberOfMajorSurgeries,degree=4,raw=TRUE),data=train)

Residual standard error: 2931 on 299 degrees of freedom

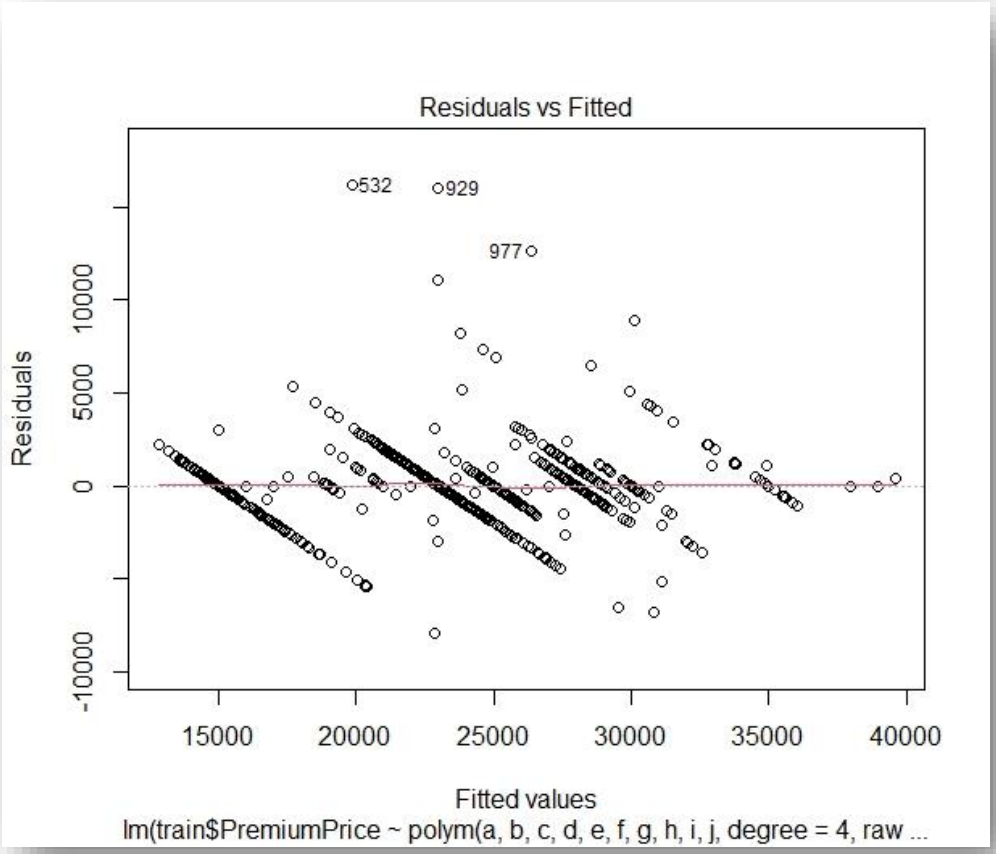
Multiple R-squared: 0.9121, Adjusted R-squared: 0.782

F-statistic: 7.007 on 443 and 299 DF, p-value: < 2.2e-16

> sigma(PR3)/mean(train$PremiumPrice)

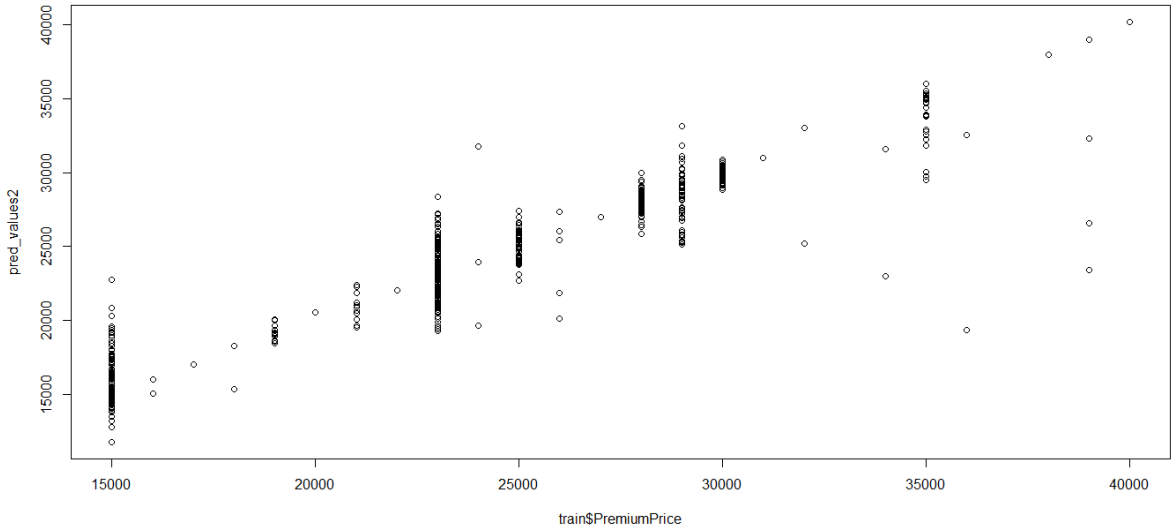
[1] 0.1218055

plot(PR3)



(Figure – 14.d)

```
> predict(PR3,train)
> pred_values2=predict(PR3,data=train)
> train$pred_premium=pred_values2
> plot(train$PremiumPrice,pred_values2)
```



(Figure – 14.e)

> write_xlsx(train,path="F:\\PROJECT2022\\RRR\\TEST-ds2.xlsx")

FINDINGS-

From the above fitting-1. Model accuracy assessment:

As we have seen in simple polynomial regression, the overall quality of the model can be assessed by examining the R-squared (R2) and Residual Standard Error (RSE).

R-squared:
In multiple polynomial regression, the R2 represents the correlation coefficient between the observed values of the outcome variable (y) and the fitted (i.e., predicted) values of y. For this reason, the value of R will always be positive and will range from zero to one.

R2 represents the proportion of variance, in the outcome variable y, that may be predicted by knowing the value of the x variables. An R2 value close to 1 indicates that the model explains a large portion of the variance in the outcome variable.

A problem with the R2, is that, it will always increase when more variables are added to the model, even if those variables are only weakly associated with the response . A solution is to adjust the R2 by taking into account the number of predictor variables.
The adjustment in the "Adjusted R Square" value in the summary output is a correction for the number of x variables included in the prediction model.

So, in this model the value of R2 is equal to 0.9121 that means our model has a 91.21% accuracy. The value of the adjusted R2 is equal to 0.782 which is not that close to our R2 value. This indicates that our model might have some unimportant variables.

2. Residual Standard Error (RSE):
The RSE estimate gives a measure of error of prediction. The lower the RSE, the more accurate the model (on the data in hand).

In this model the value of RSE is equal to 2931 corresponding to 12.18055% error rate.

B. ELIMINATING VARIABLES WITH p-VALUE > 0.05-

CODES AND OUTPUTS-

>PR4=lm(PremiumPrice~polym(Age,Diabetes,BloodPressureProblems,AnyTransplants,AnyChronicDiseases,Weight,NumberOfMajorSurgeries,degree=4,raw=TRUE),data=train)

> summary(PR4)

Call:lm(formula=train$PremiumPrice~polym(Age,Diabetes,BloodPressureProblems,AnyTransplants,AnyChronicDiseases,Weight,NumberOfMajorSurgeries,degree=4,raw=TRUE),data=train)

Residuals: Min 1Q Median 3Q Max

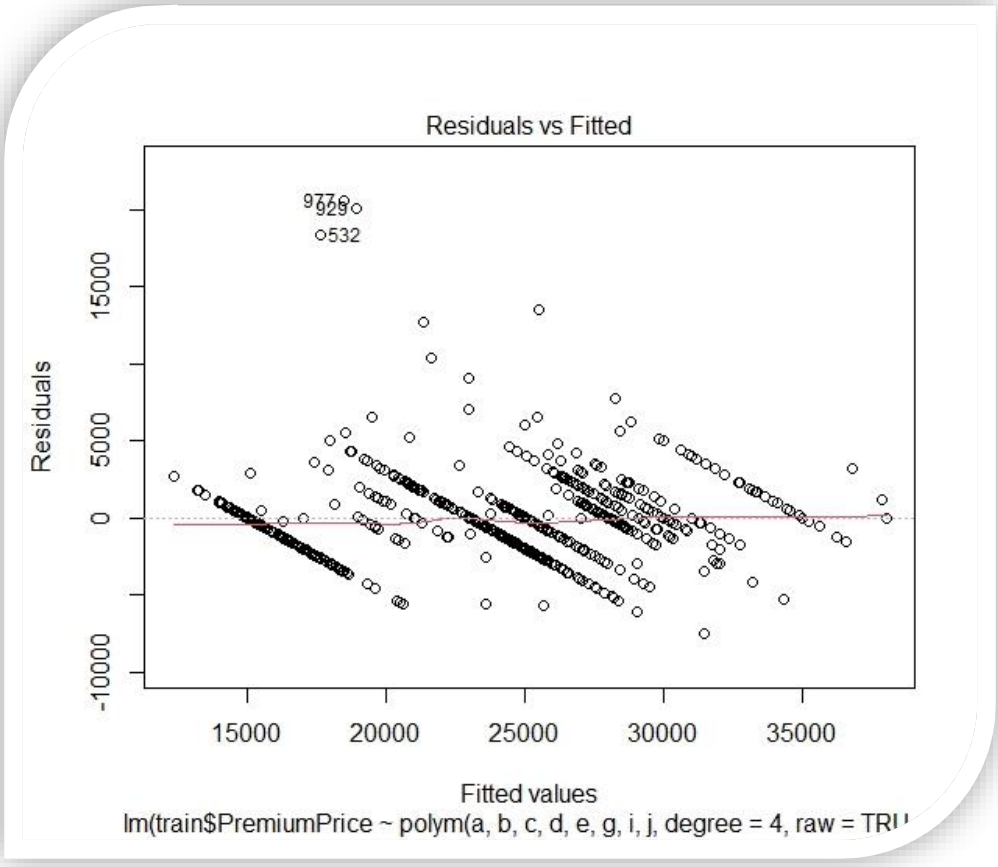
-7444.7 -1194.2 -43.0 773.3 20559.7

Residual standard error: 3086 on 573 degrees of freedom

Multiple R-squared: 0.8134, Adjusted R-squared: 0.7584

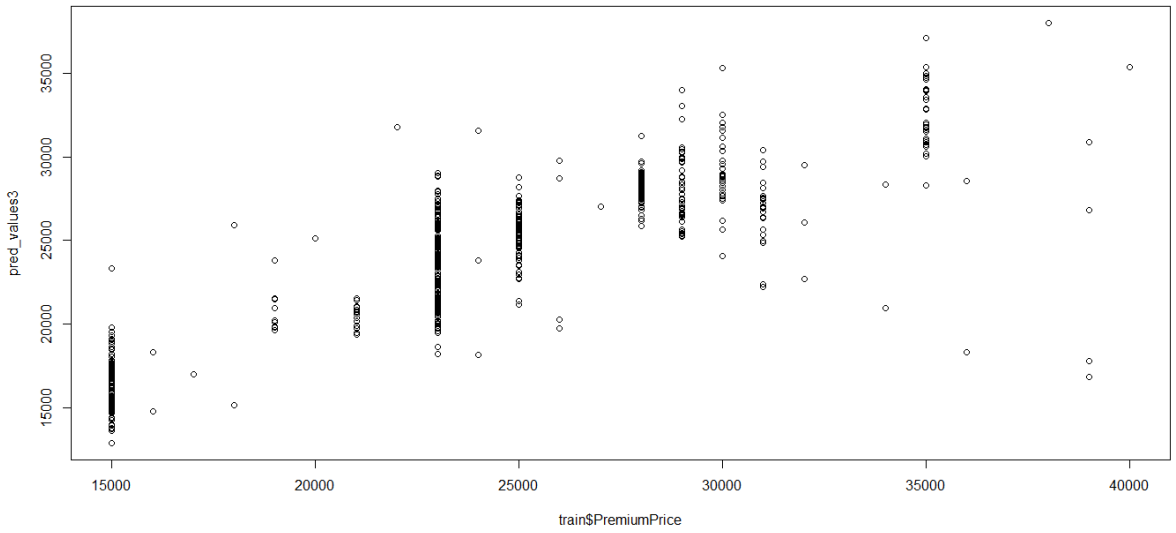
F-statistic: 14.78 on 169 and 573 DF, p-value: < 2.2e-16

plot(PR4)



(Figure – 14.f)

```
> sigma(PR4)/mean(train$PremiumPrice)

[1] 0.1267638

> predict(PR4,train)

     4      5      6      7      8     10     13     14

28296.38 27512.66 19574.13 20341.97 14871.71 23642.37 18180.25 28316.43

    15     16     17     18     19     20     21     22

15000.00 23577.72 28676.38 22472.04 25094.12 17017.17 26182.66 15136.33 …..etc.

> pred_values3=predict(PR4,data=train)

> train$pred_premium=pred_values3

> plot(train$PremiumPrice,pred_values3)
```



(Figure – 14.g)

```
> write_xlsx(train,path="F:\\PROJECT2022\\RRR\\TEST-ds3.xlsx")
```

FINDINGS-

From the above fitting-1.

Model accuracy assessment:

As we have seen in simple polynomial regression, the overall quality of the model can be assessed by examining the R-squared (R2) and Residual Standard Error (RSE).

R-squared:
In multiple polynomial regression, the R2 represents the correlation coefficient between the observed values of the outcome variable (y) and the fitted (i.e., predicted) values of y. For this reason, the value of R will always be positive and will range from zero to one.

R2 represents the proportion of variance, in the outcome variable y, that may be predicted by knowing the value of the x variables. An R2 value close to 1 indicates that the model explains a large portion of the variance in the outcome variable.

A problem with the R2, is that, it will always increase when more variables are added to the model, even if those variables are only weakly associated with the response . A solution is to adjust the R2 by taking into account the number of predictor variables.
The adjustment in the "Adjusted R Square" value in the summary output is a correction for the number of x variables included in the prediction model.

So, in this model the value of R2 is equal to 0.8134 that means our model has a 81.34% accuracy. The value of the adjusted R2 is equal to 0.7584 which is close to our R2 value. This indicates that our model may or may not have any unimportant variable.

2. Residual Standard Error (RSE):The RSE estimate gives a measure of error of prediction. The lower the RSE, the more accurate the model (on the data in hand).**In this model the value of RSE is equal to 3086 corresponding to 12.67638% error rate.**

❖ PREDICTED VALUES:

(1) Predicted values of PremiumPrice From multiple linear regression model eliminating the variables with p-value > 0.05 (page no.- 58)

1	Age	Diabetes	ressurePr	yTransplai	hronicDise	Height	Weight	ownAllerg	OfCancerIr	OfMajorS	remiumPric	ed_premium
2	52	1	1	0	1	183	93	0	0	2	28000	29707.49
3	38	0	0	0	1	166	88	0	0	1	23000	25379.25
4	30	0	0	0	0	160	69	1	0	1	23000	18508.31
5	33	0	0	0	0	150	54	0	0	0	21000	18907.43
6	23	0	0	0	0	181	79	1	0	0	15000	17646.45
7	38	0	0	0	0	182	93	0	0	0	23000	23632.08
8	24	0	0	0	0	178	57	1	0	1	15000	15605.28
9	46	0	1	0	0	184	97	0	0	0	35000	26548.81
10	18	0	0	1	0	150	76	0	0	1	15000	22471.69
11	38	0	0	0	0	160	68	1	0	1	23000	21027.56
12	42	0	0	0	1	149	67	0	0	0	30000	25626.22
13	38	1	0	0	0	154	82	0	0	0	23000	22757.6
14	57	1	0	0	0	156	61	0	0	0	25000	27260.16
15	21	0	1	0	0	186	97	0	0	0	15000	18427.73
16	49	1	0	0	0	160	97	0	0	2	28000	26289.23
17	20	1	0	0	0	181	81	0	0	0	15000	16830.92
18	35	0	0	0	0	163	92	0	0	1	32000	21960.99
19	35	0	1	0	0	175	83	0	0	1	23000	21245.51
20	31	0	0	0	0	172	57	0	0	0	21000	18496.24
21	22	0	0	1	0	151	97	0	0	0	15000	26057.58
22	30	0	0	0	1	162	73	1	0	0	23000	22205.09
23	33	1	1	0	1	153	58	0	0	0	21000	21987.14
24	22	0	0	0	0	168	96	1	0	1	15000	18056.02
25	28	0	0	0	0	158	68	0	0	0	15000	18396.19
26	44	1	0	0	0	157	55	0	0	0	23000	22560.21
27	58	0	1	0	1	147	61	0	0	1	25000	29729.66
28	43	0	0	0	1	173	81	0	1	1	30000	28336.14
29	24	1	1	1	0	168	91	1	0	0	15000	26230.28
30	20	0	1	0	0	163	68	0	0	0	15000	15797.44

(2) Predicted values of PremiumPrice From multiple polynomial regression (degree 2) model eliminating the variables with p-value > 0.05 (page no.- 58)

1	Age	Diabetes	ressurePre	yTransplar	hronicDise	Height	Weight	ownAllerg	OfCancerIn	OfMajorS	remiumPri	ed_premium
2	52	1	1	0	1	183	93	0	0	2	28000	29418.9
3	38	0	0	0	1	166	88	0	0	1	23000	25344.2
4	30	0	0	0	0	160	69	1	0	1	23000	18988.6
5	33	0	0	0	0	150	54	0	0	0	21000	18920.1
6	23	0	0	0	0	181	79	1	0	0	15000	18157.7
7	38	0	0	0	0	182	93	0	0	0	23000	23688.4
8	24	0	0	0	0	178	57	1	0	1	15000	16114.5
9	46	0	1	0	0	184	97	0	0	0	35000	26643.7
10	18	0	0	1	0	150	76	0	0	1	15000	22375.6
11	38	0	0	0	0	160	68	1	0	1	23000	21529.9
12	42	0	0	0	1	149	67	0	0	0	30000	25602.7
13	38	1	0	0	0	154	82	0	0	0	23000	22448.8
14	57	1	0	0	0	156	61	0	0	0	25000	27022.9
15	21	0	1	0	0	186	97	0	0	0	15000	18460.2
16	49	1	0	0	0	160	97	0	0	2	28000	25975.4
17	20	1	0	0	0	181	81	0	0	0	15000	16527.9
18	35	0	0	0	0	163	92	0	0	1	32000	21956.3
19	35	0	1	0	0	175	83	0	0	1	23000	21288.3
20	31	0	0	0	0	172	57	0	0	0	21000	18544
21	22	0	0	1	0	151	97	0	0	0	15000	25973.6
22	30	0	0	0	1	162	73	1	0	0	23000	22655.3
23	33	1	1	0	1	153	58	0	0	0	21000	21652
24	22	0	0	0	0	168	96	1	0	1	15000	18508.5
25	28	0	0	0	0	158	68	0	0	0	15000	18399.7
26	44	1	0	0	0	157	55	0	0	0	23000	22295.3
27	58	0	1	0	1	147	61	0	0	1	25000	29749.4
28	43	0	0	0	1	173	81	0	1	1	30000	28243.2
29	24	1	1	1	0	168	91	1	0	0	15000	26372
30	20	0	1	0	0	163	68	0	0	0	15000	15806.1

(3) Predicted values of PremiumPrice From multiple polynomial regression (degree 4) model with all variables (page no.- 58)

1	Age	Diabetes	ressurePre	yTransplar	hronicDise	Height	Weight	ownAllerg	OfCancerIn	OfMajorS	remiumPri	ed_premium
2	52	1	1	0	1	183	93	0	0	2	28000	28209.1
3	38	0	0	0	1	166	88	0	0	1	23000	22845.9
4	30	0	0	0	0	160	69	1	0	1	23000	21946.2
5	33	0	0	0	0	150	54	0	0	0	21000	21022.9
6	23	0	0	0	0	181	79	1	0	0	15000	14939.9
7	38	0	0	0	0	182	93	0	0	0	23000	23949
8	24	0	0	0	0	178	57	1	0	1	15000	15883.3
9	46	0	1	0	0	184	97	0	0	0	35000	32540.3
10	18	0	0	1	0	150	76	0	0	1	15000	15000
11	38	0	0	0	0	160	68	1	0	1	23000	24172.6
12	42	0	0	0	1	149	67	0	0	0	30000	30735.5
13	38	1	0	0	0	154	82	0	0	0	23000	22562.7
14	57	1	0	0	0	156	61	0	0	0	25000	26627.7
15	21	0	1	0	0	186	97	0	0	0	15000	15129
16	49	1	0	0	0	160	97	0	0	2	28000	27480.7
17	20	1	0	0	0	181	81	0	0	0	15000	16039.6
18	35	0	0	0	0	163	92	0	0	1	32000	25204.7
19	35	0	1	0	0	175	83	0	0	1	23000	21352.7
20	31	0	0	0	0	172	57	0	0	0	21000	19648.6
21	22	0	0	1	0	151	97	0	0	0	15000	15000
22	30	0	0	0	1	162	73	1	0	0	23000	23000
23	33	1	1	0	1	153	58	0	0	0	21000	19545.7
24	22	0	0	0	0	168	96	1	0	1	15000	15530.7
25	28	0	0	0	0	158	68	0	0	0	15000	17210.4
26	44	1	0	0	0	157	55	0	0	0	23000	23369
27	58	0	1	0	1	147	61	0	0	1	25000	24915.5
28	43	0	0	0	1	173	81	0	1	1	30000	30000
29	24	1	1	1	0	168	91	1	0	0	15000	15000
30	20	0	1	0	0	163	68	0	0	0	15000	20815.2

(4) Predicted values of PremiumPrice From multiple polynomial regression (degree 4) model eliminating the variables with p-value > 0.05 (page no.- 58)

1	Age	Diabetes	ressurePro	yTranspla	hronicDise	Height	Weight	ownAllerg	OfCancerIn	OfMajorS	remiumPri	ed_premium
2	52	1	1	0	1	183	93	0	0	2	28000	28296.4
3	38	0	0	0	1	166	88	0	0	1	23000	27512.7
4	30	0	0	0	0	160	69	1	0	1	23000	19574.1
5	33	0	0	0	0	150	54	0	0	0	21000	20342
6	23	0	0	0	0	181	79	1	0	0	15000	14871.7
7	38	0	0	0	0	182	93	0	0	0	23000	23642.4
8	24	0	0	0	0	178	57	1	0	1	15000	18180.3
9	46	0	1	0	0	184	97	0	0	0	35000	28316.4
10	18	0	0	1	0	150	76	0	0	1	15000	15000
11	38	0	0	0	0	160	68	1	0	1	23000	23577.7
12	42	0	0	0	1	149	67	0	0	0	30000	28676.4
13	38	1	0	0	0	154	82	0	0	0	23000	22472
14	57	1	0	0	0	156	61	0	0	0	25000	25094.1
15	21	0	1	0	0	186	97	0	0	0	15000	17017.2
16	49	1	0	0	0	160	97	0	0	2	28000	26182.7
17	20	1	0	0	0	181	81	0	0	0	15000	15136.3
18	35	0	0	0	0	163	92	0	0	1	32000	22671.8
19	35	0	1	0	0	175	83	0	0	1	23000	23505.2
20	31	0	0	0	0	172	57	0	0	0	21000	19357
21	22	0	0	1	0	151	97	0	0	0	15000	15000
22	30	0	0	0	1	162	73	1	0	0	23000	20803.4
23	33	1	1	0	1	153	58	0	0	0	21000	19881.9
24	22	0	0	0	0	168	96	1	0	1	15000	15451.3
25	28	0	0	0	0	158	68	0	0	0	15000	17285.9
26	44	1	0	0	0	157	55	0	0	0	23000	25157.7
27	58	0	1	0	1	147	61	0	0	1	25000	23965.8
28	43	0	0	0	1	173	81	0	1	1	30000	28919.1
29	24	1	1	1	0	168	91	1	0	0	15000	15000
30	20	0	1	0	0	163	68	0	0	0	15000	17828.3

CONCLUSION

[I] On Visualization And Data Exploration-

(1) CORRELATION MATRIX: From the correlation matrix **(Fig.-0.3)** we can conclude that all the independent variables has a positive(+ve) relationship / correlation with the response variable i.e. "Age" , "Diabetes" , "BloodPressureProblems" , "AnyTransplants", "AnyChronicDiseases" , "Height" , "Weight" , "KnownAllergies" , "HistoryOfCancerInFamily" & " NumberOfMajorSurgeries" have a positive correlation with the dependent variable "PremiumPrice". Here "Age" has the highest correlation coefficient with "PremiumPrice" which is equal to o.7 and , "Height" & "KnownAllergies" have the lowest correlation coefficient with "PremiumPrice" which are equal to 0.

- **Distribution of Age:**

(1.a) From the diagram **(fig.-1.a)** , we can see that the distribution of age with respect to premium price is a comb distribution, the bars are alternately tall and short & it is multimodal. The distribution of age with respect to premium price is left skewed i.e. mean < median.

(1.b) From the diagram **(fig.-1.b)** , we can see that the relationship between age and premium price is positive i.e. they are positively correlated.
As the data points don't cluster that much tightly so they provide a moderately strong relationship.

- **Diabetics People Premium Analysis:**

(2.a) From the pie-chart for "Diabetes" we can see that 41.99% of the total population are diabetic and 58.01% of the total population are non-diabetic.

(2.b) From the figure(2.b) we can see that the people having diabetes gets more medical premium than the people who don't have diabetes.

(2.c) From the density plot for "Diabetes" we can conclude that people having diabetes get more medical premium than the non-diabetic people as the density of diabetic people has a higher peak than non-diabetic people at 20k to 30k premium price.

(2.d) The box plot for the people with diabetes is comparatively short. This suggests that overall values have a high level of agreement with each other. The box plot for the people without diabetes is comparatively tall. This suggests the people hold quite different values of premium price.

- **Blood Pressure Patients Premium Analysis:**

(3.a) From the pie-chart for "BloodPressureProblems" we can observe that 53.14% of the total population have no blood pressure problem and rest 46.86% people of the total population have blood pressure problems.

(3.b) From the above figure we can see that the people having blood pressure problems gets more medical premium than the people who don't have blood pressure prolems.

(3.c) From the density plot for people with and without blood pressure problems we can easily conclude that the people who have blood pressure problems get more medical premium than people having no blood pressure problems as the density of the people having blood pressure problems is more at 20k to 30k premium price than the people who don't have any blood pressure problems.

(3.d) The box plot for the people with blood pressure problems is comparatively short. This suggests that overall values have a high level of agreement with each other. The box plot for the people without blood pressure problems is comparatively tall. This suggests the people hold quite different values of premium price.

- **People Gone Through Any Transplants Premium Analysis**

(4.a) From the above figure we can see that the people who have gone through any transplants gets more medical premium than the people who have no transplants.

(4.b) From the density plot for the people gone through any transplants and the people having no transplants we can conclude that people who have transplants get more medical premium than the people who have no transplants as the density for people who gone through any transplants is higher at 20k to 30k premium price than the people who don't have any transplants.

(4.c) The box plot for the people without any transplants is comparatively short. This suggests that overall values have a high level of agreement with each other. The box plot for the people with any transplants is comparatively tall. This suggests the values hold quite different values of premium price.

- **People With Chronic Disease Premium Analysis**

(5.a) From the above figure we can see that the people having any chronic diseases gets more medical premium than the people who don't have any chronic diseases.

(5.b) From the density plot for people having chronic disease we can see that the density for the people having chronic disease is much higher than for than for the people who have no chronic disease at 25k-30k premium price i.e. people having chronic disease get more medical premium.

(5.c) The box plot for the people with any chronic diseases is comparatively short. This suggests that overall values have a high level of agreement with each other. The box plot for the people without chronic diseases is comparatively tall. This suggests the values hold quite different values premium price.

- **Allergy Patients Premium Analysis**

(6.a) From the above figure we can see that the people having allergies gets more medical premium than the people who don't have allergies.

(6.b) From the density plot for people with & without allergies we can conclude that people who don't have allergies get more premium than people who have allergies as the density at premium price 20k-30k is higher for people who have no allergies than who have allergies. But the density is higher at premium price 30k-40k for the people who have allergies than the people who don't have allergies concluding that people who have allergies get more premium than people who don't have allergies at 30k-40k premium price.

(6.c) The boxplot for the people without any known allergies is comparatively short. This suggests that overall values have a high level of agreement with each other. The boxplot for the people with any known allergies is comparatively tall. This suggests the values hold quite different values premium price.

• **Patients with History of Cancer in Family Premium Analysis**

(7.a) From the pie-chart for "HistoryOfCancerInFamily" we can see that 88.24% of the total population have no history of cancer in family and the rest 11.76% people of the total population have a history of cancer in family.

(7.b) From the above figure we can see that the people having history of cancer in family gets more medical premium than the people who don't have history of cancer in family.

(7.c) From the Fig(7.c) we can conclude that people who have history of cancer gets more premium than the people who don't have history of cancer as the density is more spread for people who have history of cancer.

(7.d) The boxplot for the people without a history of cancer is comparatively short. This suggests that overall values have a high level of agreement with each other. The boxplot for the people with a history of cancer is comparatively tall. This suggests the values hold quite different values premium price.

• **People Gone Through Major Surgeries Premium Analysis**

(8.a) From the above figure we can see that the people having two or three major surgeries gets more medical premium than the people who don't have or have 1 major surgery.

(8.b) From Fig(8.b) we can conclude that people who have 2 major surgeries have more density at premium price 25k-30k.

(8.c) The boxplots for the people with 2 & 3 major surgeries are comparatively short. This suggests that overall values have a high level of agreement with each other. The boxplot for the people with no & 1 major surgeries is comparatively tall. This suggests the values hold quite different values premium price.

• **Distribution of BMI:**

(9.a) We can see that the distribution of BMI is an undefined Distribution & it is bimodal. The distribution of BMI is right skewed i.e. mean > median.

(9.b) From the above diagram we can see that the relationship between BMI and premium price is positive i.e. they are positively correlated.

• **Distribution of PremiumPrice:**

(10.a) We can see that the distribution of PremiumPrice is an undefined distribution & it is multimodal. The distribution of PremiumPrice is right skewed i.e. mean > median.

(10.b) We can see that the distribution of PremiumPrice is an undefined distribution & it is multimodal. The distribution of PremiumPrice is right skewed i.e. mean > median.

(10.c) From the above diagram we can see that the relationship between BMI , age , diabetes and premium price is positive i.e. they are positively correlated.

[2] On Hypothesis Testing-

•Normality Test:-

As the p-value(0.6132) is greater than 0.05 so at 5% level of significance we accept the null hypothesis and conclude that our data is normally distributed.

•t-Test:-

(1) As the p-value (=0.01451) < 0.05 so we reject H0 at 5% level of significance , so we reject the null hypothesis and conclude that the alternative hypothesis is true. So , there is a significant difference between means of premium prices for diabetic and non-diabetic people . The predicted values of μ A & μ B are 24896.14 and 23931.82 respectively.

(2) As the p-value (=9.813e-08) < 0.05 so we reject H0 at 5% level of significance , so we reject the null hypothesis and conclude that the alternative hypothesis is true. So , there is a significant difference between means of premium prices for the people with blood pressure problems and without pressure problems. The predicted values of μ A & μ B are 25448.05 and 23356.87 respectively.

(3) As the p-value (=5.545e-08) < 0.05 so we reject H0 at 5% level of significance , so we reject the null hypothesis and conclude that the alternative hypothesis is true. So , there is a significant difference between means of premium prices for the people with any transplants and without any transplants . The predicted values of μ A & μ B are 31763.64 and 23897.96 respectively.

(4) As the p-value (=1.728e-13) < 0.05 so we reject H0 at 5% level of significance , so we reject the null hypothesis and conclude that the alternative hypothesis is true. So , there is a significant difference between means of premium prices for the people with and without any chronic diseases. The predicted values of μ A & μ B are 27112.36 and 23725.25 respectively.

(5) As the p-value (=0.7141) > 0.05 so we reject H0 at 5% level of significance , so we accept the null hypothesis and conclude that the alternative hypothesis is false. So , there is no significant difference between means of premium prices for the people with allergies and without allergies . The predicted values of μ A & μ B are 24481.13 and 24297.16 respectively.

(6) As the p-value (=0.01983) < 0.05 so we reject H0 at 5% level of significance , so we reject the null hypothesis and conclude that the alternative hypothesis is true. So , there is a significant difference between means of premium prices for diabetic and non-diabetic people . The predicted values of μ A & μ B are 25758.62 and 24147.13 respectively.

•ANOVA Test:-

As the p-value (=2.87e-16) is less than 0.05 , so at 5% level of significance we conclude that the avg. premium prices are significant and we reject the null hypothesis and accept the alternative hypothesis.

[2] On Prediction Model-

MULTIPLE LINEAR REGRESSION:

A. WITH ALL VARIABLES-

In this model the value of R2 is equal to 0.6305 that means our model has a 63.05% accuracy. The value of the adjusted R2 is equal to 0.6254 which is close to our R2 value. This indicates that there might be some unimportant variables in our model.

B. ELIMINATING VARIABLES WITH p-VALUE >0.05-

In this model the value of R2 is equal to 0.6287 that means our model has a 62.87% accuracy. The value of the adjusted R2 is equal to 0.6257 which is very close to our R2 value. This indicates that our model is free of any unimportant variables.

• **As the overall conclusion we can conclude that, the difference between multiple R2 & adjusted R2 for the multiple linear regression eliminating variables with p-value>0.05 (B) is LESS than the difference between multiple R2 & adjusted R2 for the multiple linear regression with all variables (A) , so the multiple linear regression in (B) is BETTER than the multiple linear regression in (A) , though the accuracy of multiple linear regression in (A) (which is equal to 63.05%) is greater than the accuracy of multiple linear regression in (B) (which is equal to 62.87%) .**

SIMPLE POLYNOMIAL REGRESSION:

DEGREE-2

1. The p-value is less than 2.2e-16 which is less than 0.05 , so we will take the variable Age in our multiple polynomial model.

2. The p-value is equals to 0.02582 which is less than 0.05 , so we will take the variable Diabetes in our multiple polynomial model.

3. The p-value is equals to 1.424e-06 which is less than 0.05 , so we will take the variable BloodPressureproblems in our multiple polynomial model.

4. The p-value is equals to 4.972e-14 which is less than 0.05 , so we will take the variable AnyTransplants in our multiple polynomial model.

5. The p-value is equals to 1.566e-08 which is less than 0.05 , so we will take the variable AnyChronicDiseases in our multiple polynomial model.

6. The p-value is equals to 0.6591 which is greater than 0.05 , so we will not take the variable Height in our multiple polynomial model to get a better model with lesser difference between multiple R2 & adjusted R2.

7. The p-value is equals to 0.0003346 which is less than 0.05 , so we will take the variable Weight in our multiple polynomial model.

8. The p-value is equals to 0.2912 which is greater than 0.05 , so we will not take the variable KnownAllergies in our multiple polynomial model to get a better model with lesser difference between multiple R2 & adjusted R2.

9. The p-value is equals to 0.2947 which is greater than 0.05 , so we will not take the variable HistoryOfCancerInFamily in our multiple polynomial model to get a better model with lesser difference between multiple R2 & adjusted R2.

10. The p-value is equals to 1.013e-12 which is less than 0.05 , so we will take the variable NumberOfMajorSurgeries in our multiple polynomial model.

• **As the overall conclusion we can conclude that , the variables Height, KnownAllergies, HistoryOfCancerInFamily have p-values greater than 0.05 so in the construction of a multiple polynomial regression model we will exclude these variables to get a lower difference between the multiple R2 & the adjusted R2.**

MULTIPLE POLYNOMIAL REGRESSION:

DEGREE-2

A. WITH ALL VARIABLES-

In this model the value of R2 is equal to 0.7316 that means our model has a 73.16% accuracy. The value of the adjusted R2 is equal to 0.7085 which is not that close to our R2 value. This indicates that there might be some unimportant variables in our model.

B. ELIMINATING VARIABLES WITH p-VALUE >0.05-

In this model the value of R2 is equal to 0.7043 that means our model has a 70.43% accuracy. The value of the adjusted R2 is equal to 0.6914 which is very close to our R2 value. This indicates that our model is free of any unimportant variables.

• **As the overall conclusion we can conclude that, the difference between multiple R2 & adjusted R2 for the multiple polynomial regression eliminating variables with p-value>0.05 (B) is LESS than the difference between multiple R2 & adjusted R2 for the multiple polynomial regression with all variables (A) , so the multiple polynomial regression in (B) is BETTER than the multiple polynomial regression in (A) , though the accuracy of multiple polynomial regression in (A) (which is equal to 73.16%) is greater than the accuracy of multiple polynomial regression in (B) (which is equal to 70.43%) .**

DEGREE-4

A. WITH ALL VARIABLES-

In this model the value of R2 is equal to 0.9121 that means our model has a 91.21% accuracy. The value of the adjusted R2 is equal to 0.782 which is not that close to our R2 value. This indicates that our model might have some unimportant variables.

B. ELIMINATING VARIABLES WITH p-VALUE >0.05-

In this model the value of R2 is equal to 0.8134 that means our model has a 81.34% accuracy. The value of the adjusted R2 is equal to 0.7584 which is close to our R2 value. This indicates that our model may or may not have any unimportant variable.

- **As the overall conclusion we can conclude that, the difference between multiple R2 & adjusted R2 for the multiple polynomial regression eliminating variables with p-value>0.05 (B) is LESS than the difference between multiple R2 & adjusted R2 for the multiple polynomial regression with all variables (A) , so the multiple polynomial regression in (B) is BETTER than the multiple polynomial regression in (A) , though the accuracy of multiple polynomial regression in (A) (which is equal to 91.21%) is greater than the accuracy of multiple polynomial regression in (B) (which is equal to 81.34%) .**

FURTHER STUDIES

In our project ……. Are done. Due to Covid protocols, certain aspects of the project are still not exploited to the fullest.

In the pandemic era, it was not possible for us to physically go and collect data, thus we took data available on different websites. We assumed them to be true and reliable. Moreover, the size of the sample data was not sufficiently large, so we were bound to take few assumptions which ultimately resulted in few approximate inferences.

However, in spite of these difficulties tried my best to include as many topics as possible to be analysed in the report. The few notable topics which can be analysed based on this data and can be added to the current report are:

Non parametric Tests: When the assumptions of normality are not met, and the sample means are not normally distributed parametric tests can lead to erroneous results. Non-parametric tests (distribution-free test) are used in such situation as they do not require the normality assumption.

Principal Component Analysis and factor Analysis: It is mainly done to decrease the number of variables and take into account only the effective ones which contributes for maximum of the variation.

Chi square analysis: It is used for categorical data to test independence.

Time series: Here our object is to predict future trends from the data available.

Furthermore, Simple Linear models are rarely found in real life problems. Thus, multiple linear regression is used, even then the variable is not perfectly expressed. In such cases, exponential or logistic models can be used.] to increase the accuracy of results.

REFERENCE

I. Introductory Statistics with R: Peter Daalgard
II. Fundamentals of statistics, Vol- I & Vol- II, Gun, Gupta, Dasgupta
III. GUPTA KAPOOR
IV. github.com
V. www.kaggle.com
VI. www.wellbeingatschool.org.nz
VII. Linear Statistical Inference & It's Applicatins : C.R. Rao , etc.

DATA USED IN THE PROJECT

Age	Diabetes	BloodPr	AnyTran	AnyChro	Height	Weight	KnownA	HistoryC	Number(	PremiumPrice
45	0	0	0	0	155	57	0	0	0	25000
60	1	0	0	0	180	73	0	0	0	29000
36	1	1	0	0	158	59	0	0	1	23000
52	1	1	0	1	183	93	0	0	2	28000
38	0	0	0	1	166	88	0	0	1	23000
30	0	0	0	0	160	69	1	0	1	23000
33	0	0	0	0	150	54	0	0	0	21000
23	0	0	0	0	181	79	1	0	0	15000
48	1	0	0	0	169	74	1	0	0	23000
38	0	0	0	0	182	93	0	0	0	23000
60	0	1	0	0	175	74	0	0	2	28000
66	1	0	0	0	186	67	0	0	0	25000
24	0	0	0	0	178	57	1	0	1	15000
46	0	1	0	0	184	97	0	0	0	35000
18	0	0	1	0	150	76	0	0	1	15000
38	0	0	0	0	160	68	1	0	1	23000
42	0	0	0	1	149	67	0	0	0	30000
38	1	0	0	0	154	82	0	0	0	23000
57	1	0	0	0	156	61	0	0	0	25000
21	0	1	0	0	186	97	0	0	0	15000
49	1	0	0	0	160	97	0	0	2	28000
20	1	0	0	0	181	81	0	0	0	15000
35	0	0	0	0	163	92	0	0	1	32000
35	0	1	0	0	175	83	0	0	1	23000
53	0	1	0	0	151	97	0	1	1	35000
31	0	0	0	0	172	57	0	0	0	21000
22	0	0	1	0	151	97	0	0	0	15000
60	0	1	0	0	151	88	0	0	2	28000
30	0	0	0	1	162	73	1	0	0	23000
33	1	1	0	1	153	58	0	0	0	21000
22	0	0	0	0	168	96	1	0	1	15000
26	0	1	0	1	152	91	0	0	0	19000
28	0	0	0	0	158	68	0	0	0	15000
26	0	0	0	0	154	88	0	0	0	15000
64	1	0	0	0	172	85	0	0	3	28000
50	0	0	0	0	161	79	0	0	2	28000
44	1	0	0	0	157	55	0	0	0	23000
58	0	1	0	1	147	61	0	0	1	25000
43	0	0	0	1	173	81	0	1	1	30000
24	1	1	1	0	168	91	1	0	0	15000
20	0	1	0	0	163	68	0	0	0	15000
66	1	1	0	0	179	96	0	0	2	28000
25	0	1	0	0	184	55	0	1	1	15000
52	0	0	0	0	181	82	1	0	1	29000
26	1	0	0	0	165	57	0	0	0	15000
44	0	1	0	0	178	78	1	0	1	23000
25	1	1	0	0	179	69	0	0	0	32000
54	1	0	0	1	174	96	0	0	0	35000
64	0	1	0	0	156	56	0	0	1	25000
20	0	0	0	0	160	83	0	0	0	15000
30	0	1	0	0	162	83	0	0	0	23000
63	0	1	0	0	161	77	0	0	2	28000
61	1	1	0	1	185	58	0	0	2	28000
43	0	0	0	0	172	91	0	0	0	32000

55	43	0	0	0	0	172	91	0	0	0	32000
56	58	1	0	0	1	165	59	0	0	0	25000
57	41	1	1	0	0	160	63	0	0	0	23000
58	60	0	1	0	0	177	75	1	0	0	29000
59	43	1	0	0	0	164	59	0	0	2	28000
60	23	0	1	0	0	156	79	0	0	0	24000
61	46	1	1	0	0	181	72	0	0	0	23000
62	36	0	1	0	0	154	70	0	0	0	23000
63	55	1	1	0	0	148	61	0	0	0	25000
64	23	0	1	0	0	169	80	0	0	0	15000
65	31	0	0	0	1	163	78	0	0	0	23000
66	64	0	1	0	0	170	63	0	0	2	28000
67	26	1	1	0	0	183	79	0	0	1	15000
68	49	1	1	0	0	154	97	0	0	2	28000
69	19	0	0	0	0	148	60	0	1	1	15000
70	40	0	0	0	0	180	53	0	0	0	23000
71	32	0	0	0	0	177	57	0	0	0	21000
72	27	0	1	0	0	147	53	0	0	1	15000
73	55	0	1	0	1	157	94	0	0	1	30000
74	63	1	0	0	0	150	68	0	0	0	25000
75	25	1	1	1	1	179	68	0	0	0	38000
76	53	1	0	0	0	154	54	0	0	2	28000
77	66	1	1	0	0	161	75	0	0	2	28000
78	23	0	1	0	0	162	84	0	0	1	15000
79	36	1	1	0	0	149	74	1	0	0	23000
80	49	1	1	0	0	158	54	1	0	2	28000
81	30	0	0	1	1	166	87	0	0	0	38000
82	44	0	0	0	0	185	64	0	0	0	23000
83	63	1	0	0	0	181	54	0	0	0	25000
84	27	1	1	0	0	149	59	0	0	0	15000
85	46	1	0	0	0	187	75	0	0	1	23000
86	46	1	0	0	0	152	56	0	0	0	25000
87	31	0	0	0	1	150	81	1	1	1	25000
88	46	0	1	1	0	152	94	0	1	1	38000
89	20	1	1	0	0	167	91	0	1	1	15000
90	35	1	1	0	0	157	67	1	1	1	31000
91	33	0	0	0	1	167	58	0	0	0	21000
92	46	0	0	0	0	178	61	1	1	1	25000
93	63	1	1	0	0	175	95	0	0	3	28000
94	52	0	0	0	0	159	76	1	1	1	31000
95	54	1	1	0	1	185	66	0	0	2	28000
96	53	1	1	0	0	159	77	0	0	2	28000
97	21	0	0	0	0	165	95	1	0	1	15000
98	34	0	1	0	0	162	74	1	0	1	23000
99	31	0	0	0	1	186	80	0	0	0	23000
100	52	1	0	0	0	161	63	0	0	0	25000
101	42	0	0	0	0	148	97	0	0	0	23000
102	24	0	0	0	0	170	96	0	1	1	15000
103	45	1	0	1	0	152	91	0	0	0	38000
104	59	1	1	0	0	160	92	0	1	1	31000
105	21	0	1	0	0	185	56	1	0	0	15000
106	23	0	0	0	0	160	96	0	0	0	15000
107	33	1	0	0	0	168	68	0	0	0	23000
108	60	0	0	0	0	155	58	0	0	0	25000
109	31	0	0	0	0	188	96	1	0	1	23000

106	23	0	0	0	0	160	96	0	0	0	15000
107	33	1	0	0	0	168	68	0	0	0	23000
108	60	0	0	0	0	155	58	0	0	0	25000
109	31	0	0	0	0	188	96	1	0	1	23000
110	28	0	0	0	1	164	75	1	0	1	19000
111	64	1	0	0	0	187	53	0	0	3	28000
112	55	0	1	0	0	159	91	1	0	1	23000
113	35	1	1	0	0	186	58	0	0	0	23000
114	29	0	0	0	0	154	74	1	1	1	15000
115	35	0	0	0	0	165	62	0	0	0	23000
116	47	1	1	0	0	147	76	1	0	1	23000
117	25	0	0	0	1	167	70	0	0	1	19000
118	55	1	0	0	0	182	86	0	0	2	28000
119	59	0	1	0	0	163	62	1	0	1	25000
120	66	1	0	0	0	169	55	0	0	0	25000
121	32	0	1	0	0	178	59	1	1	1	21000
122	54	0	1	0	1	164	64	0	0	2	28000
123	63	0	1	0	0	157	72	0	0	1	29000
124	38	0	0	0	0	156	72	0	0	0	23000
125	23	0	0	0	0	152	90	0	0	0	15000
126	33	0	0	0	0	149	84	0	0	0	23000
127	31	1	0	1	0	187	95	1	0	0	38000
128	43	0	0	0	1	174	76	0	0	0	30000
129	39	0	1	0	0	151	56	1	1	1	31000
130	59	1	0	0	0	154	74	1	0	1	29000
131	19	1	0	0	0	185	62	0	0	1	15000
132	28	0	1	0	0	171	53	0	1	1	15000
133	59	1	1	0	1	167	83	0	0	2	28000
134	38	1	1	0	0	163	95	0	0	0	23000
135	23	1	0	0	0	150	96	0	0	0	15000
136	34	0	0	0	0	152	91	1	0	1	23000
137	27	1	0	0	0	186	81	0	0	0	15000
138	37	0	1	0	0	182	86	0	0	0	23000
139	66	1	1	0	0	188	75	0	0	2	28000
140	51	1	0	0	0	183	86	0	0	0	29000
141	48	0	1	1	1	176	63	0	0	0	38000
142	40	0	0	0	0	156	95	1	1	1	31000
143	54	1	0	0	0	151	59	0	0	2	28000
144	33	1	1	0	0	149	66	0	0	1	21000
145	33	0	1	0	0	159	67	0	0	1	21000
146	59	1	1	0	0	185	81	0	0	2	28000
147	52	1	0	0	0	187	75	0	0	0	29000
148	35	0	0	0	0	180	65	0	0	0	23000
149	59	0	1	1	0	149	68	0	0	0	38000
150	53	1	1	0	1	154	79	0	0	2	28000
151	50	1	1	0	0	147	53	0	0	0	25000
152	56	0	1	0	0	167	72	0	0	2	28000
153	47	0	0	0	0	169	116	0	0	1	35000
154	57	1	1	0	0	166	70	1	0	1	23000
155	42	0	0	0	0	171	90	0	0	0	23000
156	21	0	0	0	0	157	116	1	0	1	15000
157	62	1	0	0	1	167	110	0	0	0	35000
158	62	1	1	0	0	164	69	0	0	0	25000
159	20	0	0	0	0	174	76	0	0	1	15000
160	43	0	0	0	0	158	121	0	0	0	23000

160	43	0	0	0	0	158	121	0	0	0	23000
161	46	1	0	0	0	175	69	0	0	0	25000
162	58	1	1	0	0	180	116	0	0	1	35000
163	19	0	0	0	0	165	78	1	1	1	15000
164	34	1	1	0	0	182	98	0	0	0	23000
165	53	1	0	0	0	168	73	0	0	1	29000
166	29	0	0	0	0	180	76	0	0	0	15000
167	32	0	0	0	0	167	104	0	0	0	23000
168	18	0	0	0	0	164	77	0	0	0	15000
169	28	1	0	0	0	181	89	0	0	0	15000
170	44	0	0	0	0	175	73	0	0	0	23000
171	42	1	0	1	0	170	80	0	0	1	38000
172	48	0	1	0	1	156	92	0	0	0	30000
173	30	1	1	0	0	161	80	0	0	0	23000
174	31	1	1	0	0	171	79	0	1	1	25000
175	47	1	0	1	0	168	83	0	0	1	38000
176	24	1	0	0	0	178	73	0	0	1	15000
177	55	0	0	0	0	162	116	0	0	1	35000
178	42	0	0	0	0	176	71	1	0	1	23000
179	35	0	1	0	0	174	106	1	0	1	23000
180	33	1	1	0	0	164	74	0	0	0	23000
181	66	1	0	0	0	164	78	1	0	2	28000
182	38	0	0	0	1	165	68	1	0	1	23000
183	37	1	1	0	1	180	99	0	0	1	23000
184	30	0	0	0	1	160	86	1	0	1	23000
185	36	0	0	0	0	156	119	0	0	0	23000
186	30	1	1	0	0	161	107	0	0	0	23000
187	46	1	0	0	1	174	91	0	0	0	30000
188	19	0	0	0	0	173	129	1	0	0	15000
189	25	1	1	1	0	157	109	0	0	0	15000
190	42	0	0	0	0	167	93	0	0	1	23000
191	35	1	0	0	0	164	100	0	0	0	23000
192	30	0	1	0	0	174	91	0	0	0	23000
193	48	0	0	0	0	176	109	0	0	0	35000
194	50	1	0	0	0	163	127	0	0	2	28000
195	33	0	0	0	0	161	107	1	0	1	23000
196	43	1	0	0	0	165	84	0	1	1	31000
197	19	0	0	0	0	164	132	0	0	0	15000
198	39	0	1	0	0	174	120	0	0	0	23000
199	45	1	1	0	0	179	90	0	0	0	23000
200	48	1	0	0	0	182	100	0	0	1	35000
201	20	1	1	0	0	172	86	0	0	0	15000
202	40	1	1	0	0	170	91	1	0	1	23000
203	32	1	1	0	1	176	82	0	0	0	23000
204	30	0	0	0	0	177	96	1	0	1	26000
205	24	0	1	0	0	173	128	0	0	1	26000
206	27	0	1	0	0	159	120	1	1	1	39000
207	60	0	0	0	0	172	107	1	0	0	35000
208	31	1	1	0	0	174	114	0	0	1	23000
209	18	0	0	0	0	172	123	0	1	1	15000
210	32	0	1	0	0	176	93	0	0	0	23000
211	43	1	0	0	0	160	110	1	1	1	35000
212	41	0	0	0	1	178	95	0	0	0	30000
213	29	0	0	0	1	174	80	1	0	1	19000
214	60	1	0	0	0	175	98	0	0	0	24000

	A	B	C	D	E	F	G	H	I	J	K
214	60	1	0	0	0	175	98	0	0	0	24000
215	46	0	0	0	0	170	111	0	1	1	22000
216	52	0	1	0	0	167	88	0	0	2	28000
217	22	0	0	0	0	168	91	0	0	0	15000
218	66	1	1	0	0	175	103	0	0	2	28000
219	65	0	1	0	0	177	126	0	0	2	24000
220	62	0	1	1	0	164	121	1	0	1	38000
221	35	0	0	0	1	173	104	0	1	1	26000
222	50	1	1	0	1	163	102	0	1	2	28000
223	61	0	1	0	1	174	118	0	0	1	35000
224	49	0	1	0	0	162	105	0	0	1	35000
225	25	1	1	0	0	176	92	0	0	0	15000
226	37	0	0	0	0	153	93	1	0	1	23000
227	57	1	0	0	0	160	128	0	0	0	35000
228	60	0	1	0	0	177	96	1	0	2	28000
229	44	0	0	0	0	182	124	0	1	1	31000
230	22	1	1	0	0	166	122	0	0	0	15000
231	56	1	1	0	1	177	87	1	0	2	28000
232	63	1	1	0	0	167	58	1	1	1	25000
233	36	1	0	0	0	180	74	0	0	0	23000
234	53	0	0	0	0	168	73	0	0	2	28000
235	29	1	1	0	1	172	59	1	0	1	21000
236	32	0	0	0	0	160	60	1	0	1	21000
237	18	0	0	0	0	160	71	1	0	1	15000
238	35	0	1	0	0	161	93	0	0	0	23000
239	34	0	0	0	0	158	78	0	1	1	25000
240	62	0	1	0	1	175	89	0	0	2	28000
241	54	1	0	0	1	161	93	0	0	0	30000
242	49	0	1	0	0	177	95	1	0	1	35000
243	21	1	0	0	0	174	84	1	1	1	15000
244	20	0	0	0	0	173	58	0	0	0	15000
245	49	0	1	0	0	179	88	0	0	2	28000
246	47	1	1	0	0	161	86	1	0	0	23000
247	55	1	0	0	0	169	58	0	0	2	28000
248	29	0	0	0	0	171	71	0	0	1	15000
249	52	0	1	0	0	181	73	0	0	0	29000
250	47	0	1	0	1	166	64	0	0	1	25000
251	19	0	0	0	0	169	87	0	0	0	15000
252	25	0	1	0	0	170	90	0	0	0	15000
253	65	1	0	0	1	168	63	1	0	0	25000
254	32	0	0	0	0	164	93	1	0	0	23000
255	53	1	0	0	0	173	97	0	0	0	35000
256	60	1	0	0	0	159	89	1	0	2	28000
257	37	0	0	0	1	173	77	1	0	0	23000
258	20	0	0	0	0	166	75	0	0	0	15000
259	22	0	1	0	0	164	56	0	0	0	15000
260	57	0	0	0	0	158	96	0	0	0	35000
261	33	0	0	0	0	174	86	0	0	1	23000
262	48	1	0	0	0	173	85	0	0	0	23000
263	25	0	0	0	0	182	61	0	0	0	15000
264	54	1	1	0	0	160	62	0	0	1	25000
265	36	1	0	0	0	180	95	0	0	0	23000
266	24	1	0	0	0	178	78	0	0	0	15000
267	24	1	0	0	0	159	84	0	0	1	15000
268	55	0	1	0	0	182	66	0	0	1	25000

	A	B	C	D	E	F	G	H	I	J	K
268	55	0	1	0	0	182	66	0	0	1	25000
269	30	1	1	0	1	182	91	0	0	1	23000
270	40	0	1	0	1	163	83	0	0	0	30000
271	49	0	0	0	0	169	70	0	0	1	29000
272	46	0	1	0	1	166	58	0	0	1	25000
273	62	1	1	0	0	181	82	0	0	2	28000
274	44	1	1	0	1	177	63	0	1	1	30000
275	36	1	0	0	0	163	69	1	0	1	23000
276	64	1	0	0	0	163	88	0	0	3	28000
277	48	1	1	0	0	171	84	1	0	0	23000
278	20	0	0	0	0	181	60	1	0	0	15000
279	28	0	0	0	1	179	63	0	1	1	21000
280	46	0	0	0	0	168	78	1	0	1	23000
281	52	1	0	0	0	157	79	0	0	0	29000
282	18	1	1	0	0	169	73	0	0	0	26000
283	35	1	1	0	0	182	70	0	0	1	23000
284	64	1	0	0	0	175	81	0	0	3	28000
285	66	0	1	0	0	159	73	0	0	1	35000
286	50	0	1	0	1	165	81	0	1	2	28000
287	48	0	1	0	0	172	87	1	0	1	23000
288	66	1	1	0	0	157	62	1	0	1	35000
289	52	0	0	0	0	180	57	0	0	1	25000
290	48	1	1	0	0	168	64	0	0	1	25000
291	64	0	1	1	0	176	71	0	1	1	38000
292	49	0	0	0	0	162	73	1	0	1	29000
293	29	0	0	0	0	156	76	0	0	0	15000
294	43	1	0	0	0	173	70	0	0	1	23000
295	45	0	1	0	0	174	88	0	0	0	23000
296	40	0	0	1	0	164	87	0	0	0	38000
297	64	1	1	0	1	163	91	0	0	2	40000
298	40	0	1	0	0	158	71	0	1	1	31000
299	32	0	0	0	0	163	84	0	1	1	25000
300	52	1	1	0	0	158	88	0	0	2	28000
301	35	0	1	0	0	168	87	1	0	0	23000
302	26	1	0	0	0	181	74	0	0	0	15000
303	25	0	1	0	0	179	71	0	1	1	15000
304	25	0	1	0	0	176	75	1	0	0	15000
305	40	0	1	0	0	157	88	1	1	1	31000
306	26	0	0	0	0	157	66	0	0	0	15000
307	43	0	0	0	0	169	81	0	0	0	23000
308	56	0	0	0	0	169	95	1	0	0	35000
309	38	1	1	0	1	180	71	0	0	0	23000
310	65	1	1	0	0	168	81	0	0	2	28000
311	64	1	0	0	0	177	94	0	0	3	28000
312	23	0	0	0	0	166	70	0	0	0	15000
313	25	1	0	0	0	181	96	0	0	0	24000
314	36	0	1	0	1	166	86	1	0	1	23000
315	49	0	0	0	0	162	95	0	0	0	23000
316	59	1	1	1	0	158	78	0	0	1	38000
317	32	0	0	0	1	182	70	0	0	0	23000
318	52	1	1	0	0	168	68	0	0	1	25000
319	53	0	0	0	0	156	78	0	0	0	23000
320	55	1	0	1	0	159	85	0	0	0	38000
321	42	1	1	0	0	182	63	0	0	1	23000
322	28	0	0	0	1	179	77	0	0	0	19000

	A	B	C	D	E	F	G	H	I	J	K
877	48	1	0	0	0	150	67	0	1	1	25000
878	51	1	1	0	0	164	66	0	0	1	25000
879	23	0	0	0	0	148	69	0	0	0	15000
880	53	0	1	0	1	158	68	0	0	2	28000
881	31	1	1	0	0	147	68	0	0	0	23000
882	44	0	0	0	1	148	66	1	0	1	30000
883	51	0	1	0	0	170	67	0	1	1	25000
884	37	0	1	0	1	165	67	0	0	0	23000
885	22	0	1	0	0	171	72	0	1	1	15000
886	46	0	0	0	1	154	75	1	0	1	30000
887	22	1	0	0	0	157	72	0	1	1	15000
888	52	0	0	0	0	170	66	0	0	0	25000
889	32	1	1	0	0	161	68	0	0	0	23000
890	45	1	1	0	0	164	68	0	0	1	25000
891	43	0	1	0	0	151	70	1	0	1	23000
892	58	1	0	0	0	147	75	0	0	0	29000
893	42	1	1	0	0	147	67	0	0	0	23000
894	62	0	1	0	0	158	67	1	0	1	25000
895	45	0	0	0	0	164	72	0	0	0	23000
896	47	0	1	0	0	175	69	1	0	1	25000
897	27	0	0	0	0	170	71	0	0	0	15000
898	18	0	0	0	0	153	75	0	0	0	15000
899	55	0	1	0	0	160	71	0	0	2	28000
900	38	1	0	0	0	171	70	0	0	0	23000
901	37	0	0	0	1	157	71	1	0	1	23000
902	51	0	1	0	0	178	69	0	1	1	25000
903	19	1	0	0	0	150	72	0	0	0	15000
904	34	1	1	0	1	157	73	0	0	0	23000
905	57	1	1	0	1	166	73	1	0	2	28000
906	34	0	0	0	0	149	75	0	0	0	23000
907	39	1	1	0	1	175	74	0	0	1	23000
908	55	1	0	1	0	170	70	0	0	2	28000
909	30	0	1	0	0	159	69	0	1	1	25000
910	62	0	0	0	1	162	71	0	0	0	30000
911	25	0	1	0	0	177	74	0	0	1	15000
912	19	0	0	0	0	154	71	1	0	0	15000
913	51	0	1	0	0	153	72	0	0	1	29000
914	30	1	1	0	0	177	71	1	0	1	23000
915	40	0	0	0	0	151	70	0	0	0	23000
916	21	0	0	0	0	149	69	0	0	0	15000
917	54	0	1	1	1	156	73	0	0	2	28000
918	63	1	1	0	0	163	66	0	0	1	25000
919	66	1	1	0	0	178	66	0	0	2	28000
920	43	1	0	0	0	173	75	0	0	1	23000
921	47	0	1	0	0	172	71	1	0	1	23000
922	29	0	0	0	0	170	67	1	0	0	15000
923	53	1	0	0	0	145	75	0	0	0	29000
924	33	0	0	1	0	162	70	0	0	0	38000
925	21	0	0	0	0	148	73	0	0	0	15000
926	62	1	1	0	0	159	68	0	0	0	25000
927	41	1	0	0	0	148	67	0	1	1	31000
928	24	0	1	0	0	159	67	0	0	0	39000
929	31	0	0	1	0	154	73	0	1	1	38000
930	19	0	0	0	0	171	67	0	0	1	39000
931	22	0	1	0	0	175	72	0	0	0	15000

	A	B	C	D	E	F	G	H	I	J	K
823	65	1	1	0	0	169	80	0	0	2	28000
824	18	0	1	0	0	163	82	0	0	0	15000
825	62	1	1	0	0	150	64	0	0	2	28000
826	36	1	1	0	0	157	95	0	0	0	23000
827	32	0	1	0	0	177	98	0	1	1	25000
828	52	0	1	0	1	148	70	0	1	2	28000
829	27	0	0	0	0	175	86	0	0	1	15000
830	65	1	1	0	0	158	83	0	0	0	29000
831	31	0	0	0	0	150	91	1	0	1	23000
832	52	1	1	0	0	176	82	0	1	1	31000
833	26	0	0	0	1	147	67	0	0	0	21000
834	40	0	0	0	0	170	83	0	0	0	23000
835	25	1	0	0	0	151	92	0	1	1	15000
836	21	0	1	0	0	156	99	0	0	0	15000
837	21	0	0	0	0	176	88	0	0	0	15000
838	29	1	0	0	0	145	62	0	0	0	15000
839	50	0	1	0	0	164	74	0	0	2	28000
840	44	0	0	0	0	174	57	0	0	0	23000
841	56	1	1	0	1	159	59	0	0	2	28000
842	48	0	0	0	0	150	89	0	0	0	23000
843	56	1	0	1	0	149	65	0	0	2	28000
844	50	0	0	0	1	174	72	0	0	0	30000
845	60	1	1	0	1	145	63	0	0	2	28000
846	18	0	0	0	0	154	60	0	0	0	15000
847	41	0	0	0	0	166	58	0	0	0	23000
848	35	0	1	0	0	160	75	0	0	0	23000
849	24	1	0	0	0	174	100	0	0	0	15000
850	22	0	0	0	0	169	86	1	1	1	15000
851	20	0	0	0	0	169	66	0	0	0	15000
852	48	1	0	0	0	153	74	1	1	1	31000
853	18	0	0	0	0	170	72	0	0	1	15000
854	61	0	1	0	1	163	66	0	0	1	25000
855	37	1	1	0	0	150	69	0	0	0	23000
856	54	0	1	0	1	156	67	0	0	2	28000
857	29	1	0	0	0	156	74	0	0	0	15000
858	21	0	0	0	0	170	75	0	0	0	15000
859	55	0	0	0	0	153	73	0	0	0	29000
860	65	0	1	0	1	166	73	0	0	1	35000
861	28	1	0	0	0	160	69	0	0	0	15000
862	61	1	0	0	0	161	66	0	0	3	28000
863	44	0	0	0	0	162	75	0	0	0	23000
864	53	0	1	0	1	172	75	0	0	1	30000
865	25	0	0	1	0	175	74	0	0	0	15000
866	33	1	1	0	0	169	73	1	0	1	23000
867	66	0	1	1	0	154	66	1	0	1	35000
868	28	1	1	0	0	147	69	0	0	1	15000
869	62	1	1	0	0	166	67	0	0	1	25000
870	25	0	0	0	0	164	66	0	0	0	15000
871	18	0	1	0	0	164	68	0	0	1	15000
872	39	1	0	0	0	168	68	1	1	1	31000
873	28	1	0	0	0	175	68	0	0	1	15000
874	42	0	0	0	0	155	69	0	0	1	23000
875	35	1	0	0	0	173	69	0	0	0	23000
876	51	1	1	0	0	152	72	0	0	1	29000
877	48	1	0	0	0	150	67	0	1	1	25000

	A	B	C	D	E	F	G	H	I	J	K	
934	42	0	0	1	0	156	75	1	0	1	38000	
935	54	0	1	0	0	161	69	0	1	1	25000	
936	54	1	1	0	0	160	70	0	1	1	31000	
937	38	0	1	0	1	170	71	0	1	1	31000	
938	19	1	0	0	0	165	69	1	0	1	15000	
939	48	1	1	0	1	168	68	0	0	0	25000	
940	31	0	0	1	1	174	66	0	0	1	38000	
941	43	0	0	0	0	159	72	0	0	1	23000	
942	40	1	1	0	1	155	66	0	0	1	30000	
943	35	1	0	0	0	176	68	0	0	0	23000	
944	25	1	1	0	1	151	70	0	0	0	19000	
945	53	0	0	0	0	173	66	0	1	1	25000	
946	43	0	0	0	0	166	73	0	0	1	23000	
947	55	0	0	0	0	157	73	0	0	0	23000	
948	45	0	1	0	0	165	73	1	0	0	23000	
949	23	0	0	0	0	154	72	0	0	0	15000	
950	60	1	1	0	0	155	69	0	0	0	25000	
951	30	1	1	0	0	152	71	0	0	1	23000	
952	52	1	1	0	0	161	67	0	0	2	28000	
953	25	0	0	1	0	161	69	1	0	1	15000	
954	44	1	0	1	0	174	66	0	0	1	38000	
955	62	1	1	0	0	157	66	0	0	3	28000	
956	63	1	1	0	0	158	73	0	0	3	28000	
957	21	1	0	0	0	147	66	0	0	0	15000	
958	46	1	1	0	0	146	67	0	0	0	25000	
959	27	1	1	0	0	155	75	0	0	1	15000	
960	18	0	0	0	0	162	70	0	0	0	15000	
961	66	0	1	0	0	153	75	0	0	2	28000	
962	23	0	0	0	0	175	73	0	0	0	15000	
963	59	1	1	0	0	154	66	0	1	1	25000	
964	32	1	1	0	0	166	70	1	0	0	23000	
965	43	1	0	0	0	147	67	0	0	1	25000	
966	27	1	0	0	0	162	69	0	0	0	15000	
967	35	1	1	0	0	165	67	0	0	1	23000	
968	66	0	1	0	1	176	71	1	0	1	35000	
969	42	1	1	0	0	152	67	1	0	1	23000	
970	18	0	0	0	0	160	73	0	0	0	15000	
971	45	0	1	0	0	168	66	0	0	0	25000	
972	46	1	1	0	0	168	75	0	0	1	23000	
973	26	0	0	0	0	178	66	0	0	0	15000	
974	31	0	1	0	0	152	75	1	0	1	23000	
975	28	0	0	0	0	167	66	0	0	0	15000	
976	47	1	0	0	0	170	66	0	0	0	25000	
977	44	0	1	0	0	161	75	0	0	0	23000	
978	21	0	1	0	0	155	74	0	0	0	39000	
979	45	0	1	0	1	157	67	0	0	1	25000	
980	40	0	1	1	0	168	70	0	0	0	17000	
981	24	0	0	0	0	161	71	0	0	0	15000	
982	40	0	1	1	0	171	74	0	0	0	38000	
983	18	0	0	0	0	169	67	0	0	0	15000	
984	64	1	1	0	0	153	70	0	0	3	28000	
985	56	0	1	0	0	155	71	0	0	1	29000	
986	47	1	1	0	0	158	73	1	0	1	39000	
987	21	0	0	0	0	158	75	1	0	1	15000	