Data Prep & Descriptive Statistics

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse
1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.4
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
## Warning: package 'purrr' was built under R version 4.0.4
```

```
## Warning: package 'dplyr' was built under R version 4.0.4
```

```
## Warning: package 'forcats' was built under R version 4.0.5
```

```
## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(RCurl)
```

```
##
## Attaching package: 'RCurl'
```

```
## The following object is masked from 'package:tidyr':
##
##     complete
```

#Importing the data

```
train <- read.csv("https://raw.githubusercontent.com/crisajose/CIND-820-Big-Data-Analytics-Project/main/train.csv", header=T)
```

number of observations and number of variables in dataset

```
dim(train)
```

```
## [1] 59381   128
```

displaying structure and a subset of values of dataset

```
glimpse(train)
```

```
## Rows: 59,381
## Columns: 128
## $ Id                  <int> 2, 5, 6, 7, 8, 10, 11, 14, 15, 16, 17, 18, 19,
...
## $ Product_Info_1      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1,
1,...
## $ Product_Info_2      <chr> "D3", "A1", "E1", "D4", "D2", "D2", "A8",
"D2",...
## $ Product_Info_3      <int> 10, 26, 26, 10, 26, 26, 10, 26, 26, 21, 26,
26,...
## $ Product_Info_4      <dbl> 0.07692308, 0.07692308, 0.07692308,
0.48717949,...
## $ Product_Info_5      <int> 2, 2, 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2,...
## $ Product_Info_6      <int> 1, 3, 3, 3, 3, 1, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3,...
## $ Product_Info_7      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1,...
## $ Ins_Age             <dbl> 0.64179104, 0.05970149, 0.02985075,
0.16417910,...
## $ Ht                  <dbl> 0.5818182, 0.6000000, 0.7454545, 0.6727273,
0.6...
## $ Wt                  <dbl> 0.1485356, 0.1317992, 0.2887029, 0.2050209,
0.2...
## $ BMI                 <dbl> 0.3230080, 0.2722877, 0.4287804, 0.3524377,
0.4...
## $ Employment_Info_1   <dbl> 0.0280, 0.0000, 0.0300, 0.0420, 0.0270,
0.3250,...
## $ Employment_Info_2   <int> 12, 1, 9, 9, 9, 15, 1, 12, 9, 1, 9, 3, 9, 9,
3,...
## $ Employment_Info_3   <int> 1, 3, 1, 1, 1, 1, 3, 1, 1, 3, 1, 1, 1, 1, 1,
1,...
## $ Employment_Info_4   <dbl> 0, 0, 0, 0, 0, 0, NA, 0, 0, 0, NA, NA, 0, NA,
0...
## $ Employment_Info_5   <int> 3, 2, 2, 3, 2, 2, 3, 2, 2, 3, 2, 2, 2, 2, 2,
2,...
## $ Employment_Info_6   <dbl> NA, 0.0018, 0.0300, 0.2000, 0.0500, 1.0000,
0.8...
## $ InsuredInfo_1       <int> 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 2,
1,...
## $ InsuredInfo_2       <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2,...
## $ InsuredInfo_3       <int> 6, 6, 8, 8, 6, 8, 3, 6, 3, 3, 4, 3, 8, 3, 3,
3,...
## $ InsuredInfo_4       <int> 3, 3, 3, 3, 3, 3, 3, 3, 2, 3, 2, 3, 3, 3, 3,
3,...
## $ InsuredInfo_5       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1,
```

```
1,...
## $ InsuredInfo_6       <int> 2, 2, 1, 2, 2, 1, 2, 1, 1, 2, 1, 1, 2, 1, 2,
2,...
## $ InsuredInfo_7       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1,...
## $ Insurance_History_1 <int> 1, 2, 2, 2, 2, 2, 1, 1, 1, 2, 1, 2, 1, 2, 2,
1,...
## $ Insurance_History_2 <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1,...
## $ Insurance_History_3 <int> 3, 3, 1, 1, 1, 3, 3, 3, 3, 3, 3, 1, 3, 1, 3,
3,...
## $ Insurance_History_4 <int> 1, 1, 3, 3, 3, 2, 2, 1, 2, 1, 1, 3, 1, 3, 1,
2,...
## $ Insurance_History_5 <dbl> 0.000666667, 0.000133333, NA, NA, NA,
0.0050000...
## $ Insurance_History_7 <int> 1, 1, 3, 3, 3, 1, 1, 1, 1, 1, 1, 3, 2, 3, 1,
1,...
## $ Insurance_History_8 <int> 1, 3, 2, 2, 2, 3, 1, 1, 1, 3, 1, 2, 1, 2, 3,
1,...
## $ Insurance_History_9 <int> 2, 2, 3, 3, 3, 2, 2, 2, 2, 2, 2, 3, 2, 3, 2,
2,...
## $ Family_Hist_1       <int> 2, 2, 3, 3, 2, 2, 3, 2, 3, 3, 3, 2, 3, 3, 3,
3,...
## $ Family_Hist_2       <dbl> NA, 0.1884058, 0.3043478, 0.4202899,
0.4637681,...
## $ Family_Hist_3       <dbl> 0.5980392, NA, NA, NA, NA, 0.2941176, NA,
0.490...
## $ Family_Hist_4       <dbl> NA, 0.08450704, 0.22535211, 0.35211268,
0.40845...
## $ Family_Hist_5       <dbl> 0.5267857, NA, NA, NA, NA, NA, NA, 0.6339286,
N...
## $ Medical_History_1   <int> 4, 5, 10, 0, NA, 6, 5, 6, 4, NA, 1, 4, 5, NA,
1...
## $ Medical_History_2   <int> 112, 412, 3, 350, 162, 491, 600, 145, 16, 162,
...
## $ Medical_History_3   <int> 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 2, 2, 3, 2, 2,
2,...
## $ Medical_History_4   <int> 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 2, 1, 2,
1,...
## $ Medical_History_5   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1,...
## $ Medical_History_6   <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 1, 3, 3, 3, 3,
3,...
## $ Medical_History_7   <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1,
2,...
## $ Medical_History_8   <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
3,...
## $ Medical_History_9   <int> 1, 1, 2, 2, 2, 2, 1, 1, 1, 2, 1, 2, 2, 2, 2,
1,...
## $ Medical_History_10  <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
```

```
NA,...
## $ Medical_History_11  <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3,...
## $ Medical_History_12  <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2,...
## $ Medical_History_13  <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3,...
## $ Medical_History_14  <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3,...
## $ Medical_History_15  <int> 240, 0, NA, NA, NA, NA, NA, NA, NA, NA, NA, 8,
...
## $ Medical_History_16  <int> 3, 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3, 1, 1, 1,
3,...
## $ Medical_History_17  <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3,...
## $ Medical_History_18  <int> 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1,...
## $ Medical_History_19  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2,
1,...
## $ Medical_History_20  <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2,...
## $ Medical_History_21  <int> 1, 1, 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1,...
## $ Medical_History_22  <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2,...
## $ Medical_History_23  <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 1, 1, 3, 1, 3, 3,
1,...
## $ Medical_History_24  <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA,...
## $ Medical_History_25  <int> 1, 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1,...
## $ Medical_History_26  <int> 3, 3, 2, 3, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3,...
## $ Medical_History_27  <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3,...
## $ Medical_History_28  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1,
1,...
## $ Medical_History_29  <int> 3, 3, 3, 3, 3, 3, 1, 3, 1, 3, 1, 3, 3, 3, 3,
1,...
## $ Medical_History_30  <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2,...
## $ Medical_History_31  <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3,...
## $ Medical_History_32  <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA,...
## $ Medical_History_33  <int> 1, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 1, 3, 1, 3,
3,...
## $ Medical_History_34  <int> 3, 1, 3, 3, 3, 1, 3, 3, 3, 3, 1, 3, 3, 3, 3,
3,...
## $ Medical_History_35  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
```

```
1,...
## $ Medical_History_36  <int> 2, 2, 3, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2,...
## $ Medical_History_37  <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2,...
## $ Medical_History_38  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1,...
## $ Medical_History_39  <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 1, 3, 3,
3,...
## $ Medical_History_40  <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
3,...
## $ Medical_History_41  <int> 3, 1, 1, 1, 1, 3, 3, 1, 3, 1, 1, 3, 3, 1, 1,
3,...
## $ Medical_Keyword_1   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_2   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_3   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_4   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_5   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_6   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_7   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_8   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_9   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_10  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1,...
## $ Medical_Keyword_11  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_12  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_13  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_14  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_15  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0,
1,...
## $ Medical_Keyword_16  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_17  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_18  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_19  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
```

```
0,...
## $ Medical_Keyword_20  <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_21  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_22  <int> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_23  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0,
0,...
## $ Medical_Keyword_24  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_25  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,
1,...
## $ Medical_Keyword_26  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_27  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_28  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_29  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_30  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
0,...
## $ Medical_Keyword_31  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_32  <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_33  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_34  <int> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_35  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_36  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_37  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_38  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_39  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_40  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_41  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_42  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_43  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_44  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
```

```
0,...
## $ Medical_Keyword_45  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_46  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_47  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
0,...
## $ Medical_Keyword_48  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
0,...
## $ Response            <int> 8, 4, 8, 8, 8, 8, 8, 1, 8, 1, 6, 2, 7, 3, 8,
5,...
```

#summary of predictive variable(Response), BMI, Weight & Height.

```
summary(train$Response)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   6.000   5.637   8.000   8.000

summary(train$BMI)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3855  0.4513  0.4695  0.5329  1.0000

summary(train$Wt)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2259  0.2887  0.2926  0.3452  1.0000

summary(train$Ht)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.6545  0.7091  0.7073  0.7636  1.0000
```

#Checking missing values present in dataset

```
colSums(is.na(train))

##                   Id      Product_Info_1      Product_Info_2
Product_Info_3
##                    0                   0                   0
0
##      Product_Info_4      Product_Info_5      Product_Info_6
Product_Info_7
##                    0                   0                   0
0
##             Ins_Age                  Ht                  Wt
BMI
##                    0                   0                   0
0
##   Employment_Info_1   Employment_Info_2   Employment_Info_3
Employment_Info_4
```

```
##                  19                    0                    0
6779
##   Employment_Info_5   Employment_Info_6       InsuredInfo_1
InsuredInfo_2
##                   0               10854                    0
0
##       InsuredInfo_3       InsuredInfo_4       InsuredInfo_5
InsuredInfo_6
##                   0                    0                    0
0
##       InsuredInfo_7 Insurance_History_1 Insurance_History_2
Insurance_History_3
##                   0                    0                    0
0
## Insurance_History_4 Insurance_History_5 Insurance_History_7
Insurance_History_8
##                   0               25396                    0
0
## Insurance_History_9       Family_Hist_1       Family_Hist_2
Family_Hist_3
##                   0                    0                28656
34241
##       Family_Hist_4       Family_Hist_5   Medical_History_1
Medical_History_2
##               19184               41811                 8889
0
##   Medical_History_3   Medical_History_4   Medical_History_5
Medical_History_6
##                   0                    0                    0
0
##   Medical_History_7   Medical_History_8   Medical_History_9
Medical_History_10
##                   0                    0                    0
58824
##  Medical_History_11  Medical_History_12  Medical_History_13
Medical_History_14
##                   0                    0                    0
0
##  Medical_History_15  Medical_History_16  Medical_History_17
Medical_History_18
##               44596                    0                    0
0
##  Medical_History_19  Medical_History_20  Medical_History_21
Medical_History_22
##                   0                    0                    0
0
##  Medical_History_23  Medical_History_24  Medical_History_25
Medical_History_26
##                   0               55580                    0
0
```

```
##  Medical_History_27  Medical_History_28  Medical_History_29
Medical_History_30
##                   0                   0                   0
0
##  Medical_History_31  Medical_History_32  Medical_History_33
Medical_History_34
##                   0               58274                   0
0
##  Medical_History_35  Medical_History_36  Medical_History_37
Medical_History_38
##                   0                   0                   0
0
##  Medical_History_39  Medical_History_40  Medical_History_41
Medical_Keyword_1
##                   0                   0                   0
0
##   Medical_Keyword_2   Medical_Keyword_3   Medical_Keyword_4
Medical_Keyword_5
##                   0                   0                   0
0
##   Medical_Keyword_6   Medical_Keyword_7   Medical_Keyword_8
Medical_Keyword_9
##                   0                   0                   0
0
##  Medical_Keyword_10  Medical_Keyword_11  Medical_Keyword_12
Medical_Keyword_13
##                   0                   0                   0
0
##  Medical_Keyword_14  Medical_Keyword_15  Medical_Keyword_16
Medical_Keyword_17
##                   0                   0                   0
0
##  Medical_Keyword_18  Medical_Keyword_19  Medical_Keyword_20
Medical_Keyword_21
##                   0                   0                   0
0
##  Medical_Keyword_22  Medical_Keyword_23  Medical_Keyword_24
Medical_Keyword_25
##                   0                   0                   0
0
##  Medical_Keyword_26  Medical_Keyword_27  Medical_Keyword_28
Medical_Keyword_29
##                   0                   0                   0
0
##  Medical_Keyword_30  Medical_Keyword_31  Medical_Keyword_32
Medical_Keyword_33
##                   0                   0                   0
0
##  Medical_Keyword_34  Medical_Keyword_35  Medical_Keyword_36
Medical_Keyword_37
```

```
##                   0                   0                   0
0
##  Medical_Keyword_38  Medical_Keyword_39  Medical_Keyword_40
Medical_Keyword_41
##                   0                   0                   0
0
##  Medical_Keyword_42  Medical_Keyword_43  Medical_Keyword_44
Medical_Keyword_45
##                   0                   0                   0
0
##  Medical_Keyword_46  Medical_Keyword_47  Medical_Keyword_48
Response
##                   0                   0                   0
0
```

#Find columns with missing values

```
Missingcol<-colnames(train)[colSums(is.na(train)) > 0]
Missingcol
```

```
##  [1] "Employment_Info_1"   "Employment_Info_4"   "Employment_Info_6"
##  [4] "Insurance_History_5" "Family_Hist_2"       "Family_Hist_3"
##  [7] "Family_Hist_4"       "Family_Hist_5"       "Medical_History_1"
## [10] "Medical_History_10"  "Medical_History_15"  "Medical_History_24"
## [13] "Medical_History_32"
```

#11 out of 13 columns having NA have very high missing values

Checking Correlation between BMI, Wt and Ht

```
cor(train$BMI,train$Wt)
```

```
## [1] 0.8540833
```

```
cor(train$BMI,train$Ht)
```

```
## [1] 0.1231248
```

```
cor(train$Ht,train$Wt)
```

```
## [1] 0.6104248
```

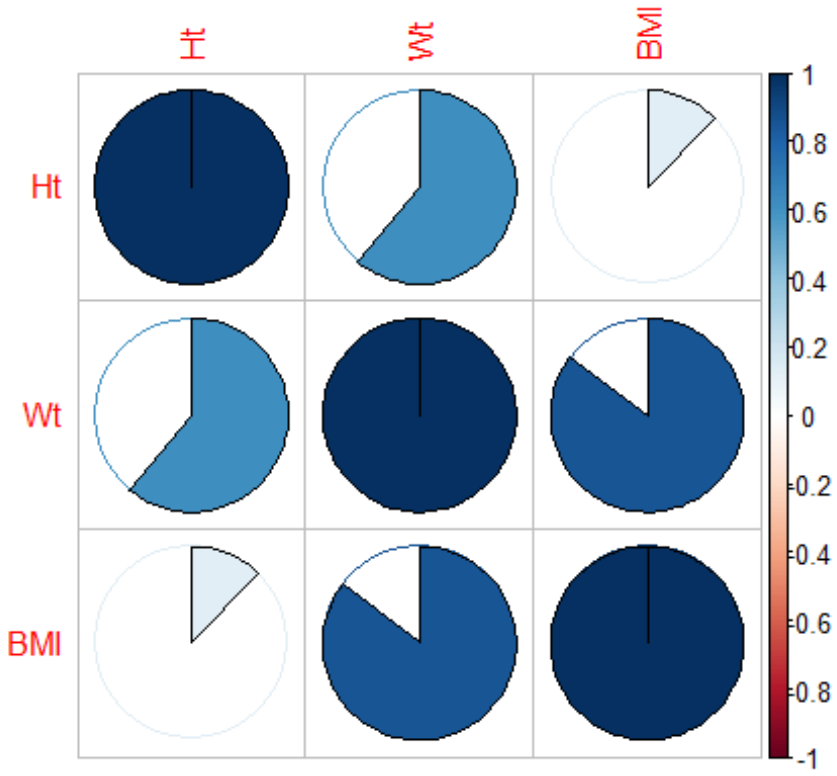
We find that BMI is highly correlated with Wt.

```
library(corrplot)
```

```
## corrplot 0.90 loaded
```

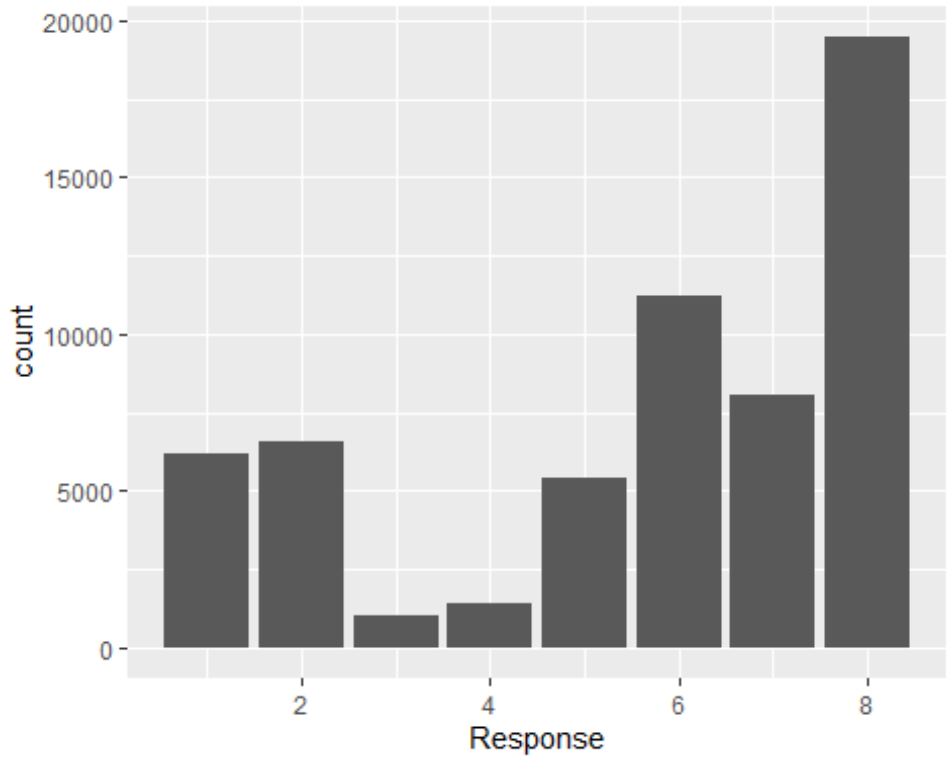
```
newclean<- train[,c(10,11,12)]
clean_corr <-  as.matrix(cor(newclean))
corrplot(clean_corr, method="number")
```


```
corrplot(clean_corr, method="pie")
```



#Examine variable Response

```r
library(ggplot2)
responseg <- ggplot(data = train, aes(Response)) + geom_bar()
responseg
```

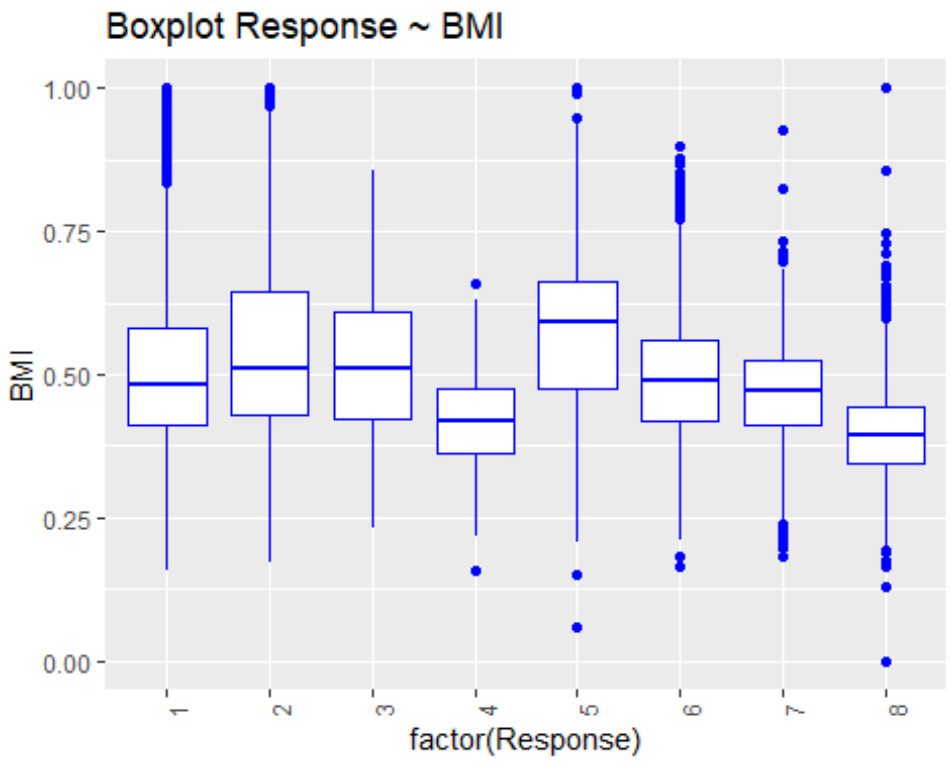



```r
#Response values for 3 and 4 are very rare. Response 8 is by far the most common value
```

#BMI seem to have an impact on the response variable based on the graph

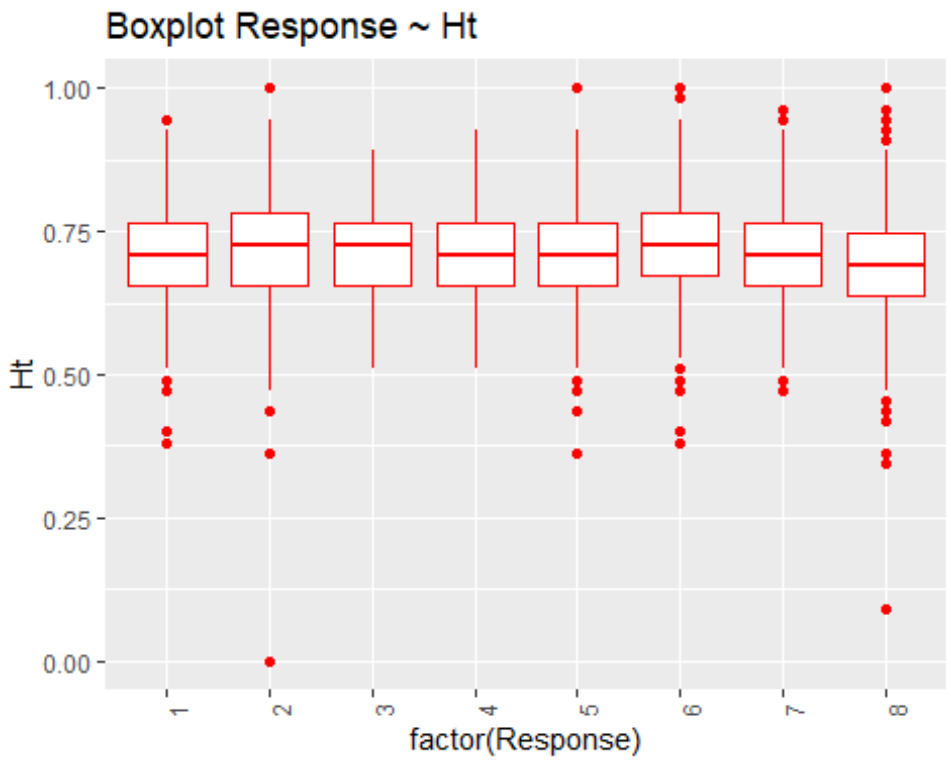
#Response value 8 has lower value of BMI compared to response value 1.Therefore we see that BMI has a strong relationship with Response variable

```r
# Boxplot Response ~ BMI
ggplot(train, aes(x=factor(Response), y=BMI)) + ggtitle("Boxplot Response ~ BMI") +
  geom_boxplot(colour="blue")+
theme(axis.text.x=element_text(angle=90,hjust=1))
```



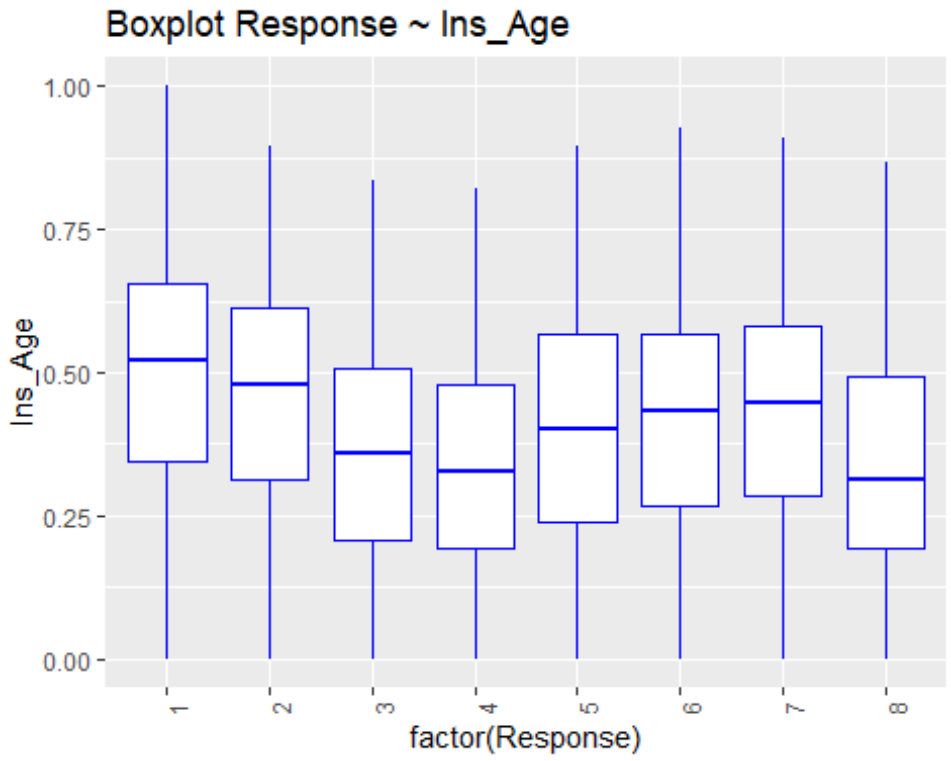
#we see that Ht does not have a strong relationship with Response variable and does not affect it as much

```
# Boxplot Response ~ Ht
ggplot(train, aes(x=factor(Response), y=Ht)) + ggtitle("Boxplot Response ~ Ht") +
  geom_boxplot(colour="red")+
theme(axis.text.x=element_text(angle=90,hjust=1))
```



#we see that Age has a relationship with Response variable

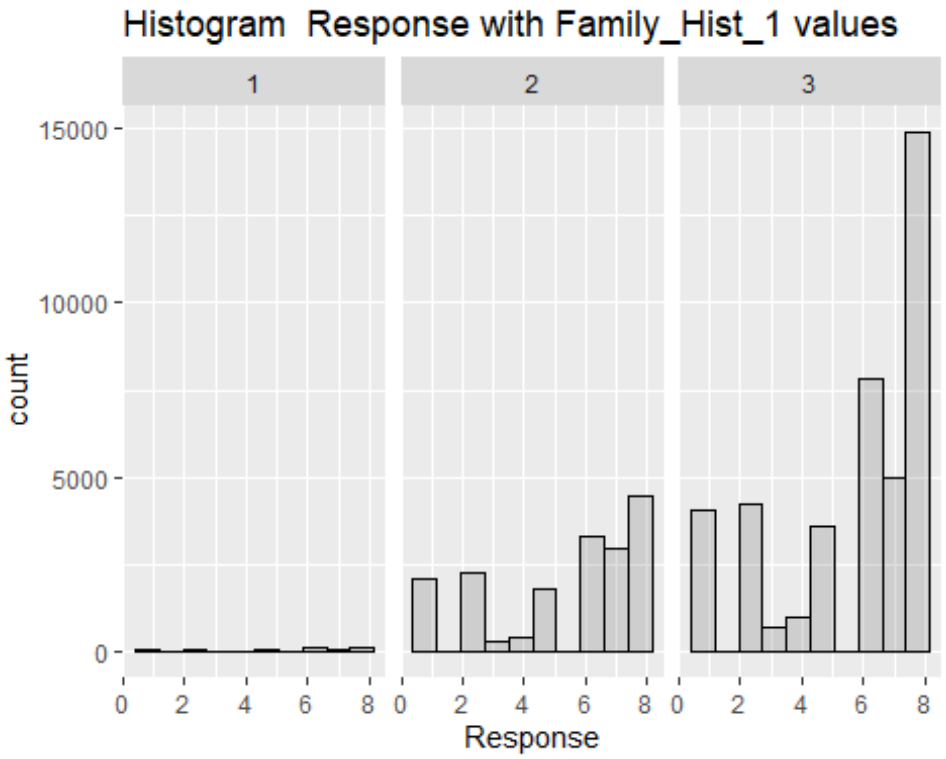
```
ggplot(train, aes(x=factor(Response), y=Ins_Age)) + ggtitle("Boxplot Response
~ Ins_Age") +
  geom_boxplot(colour="blue")+
theme(axis.text.x=element_text(angle=90,hjust=1))
```



#We can see that distribution of response variable is dependent on Family_Hist_1

```
# histogram of Response on Family_Hist_1

  ggplot(train, aes(x=Response)) + ggtitle("Histogram  Response with
Family_Hist_1 values") +
    geom_histogram(position="identity", colour="black", alpha=0.2, bins =
10)+
    facet_grid(. ~ Family_Hist_1)
```


Histogram Response with Family_Hist_1 values
1
2
3
15000
10000
5000
0
count
0
2
4
6
8
Response