Project

2022-12-07

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
#install.packages("ggpubr")
library("ggpubr")
#install.packages("gridExtra")
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
#install.packages("imputeTS")
library(imputeTS)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
#install.packages('e1071')
#install.packages("rpart.plot")
library(e1071)
library(rpart)
library(rpart.plot)
#install.packages("rio")
#install.packages("caret")
#install.packages("kernlab")
#install.packages("rlang")
library(rio)
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(kernlab)
```

```
##
## Attaching package: 'kernlab'
##
## The following object is masked from 'package:purrr':
##
##     cross
##
## The following object is masked from 'package:ggplot2':
##
##     alpha
```

```
library(rlang)
```

```
##
## Attaching package: 'rlang'
##
## The following objects are masked from 'package:purrr':
##
##     %@%, as_function, flatten, flatten_chr, flatten_dbl, flatten_int,
##     flatten_lgl, flatten_raw, invoke, splice
```

```
df <- read.csv("https://intro-datascience.s3.us-east-2.amazonaws.com/HMO_data.csv")
```

```
head(df)
```

```
##   X age    bmi children smoker      location location_type education_level
## 1 1  18 27.900        0    yes   CONNECTICUT         Urban        Bachelor
## 2 2  19 33.770        1     no  RHODE ISLAND         Urban        Bachelor
## 3 3  27 33.000        3     no MASSACHUSETTS         Urban          Master
## 4 4  34 22.705        0     no  PENNSYLVANIA       Country          Master
## 5 5  32 28.880        0     no  PENNSYLVANIA       Country             PhD
## 6 7  47 33.440        1     no  PENNSYLVANIA         Urban        Bachelor
##   yearly_physical   exercise married hypertension gender cost
## 1              No     Active Married            0 female 1746
## 2              No Not-Active Married            0   male  602
## 3              No     Active Married            0   male  576
## 4              No Not-Active Married            1   male 5562
## 5              No Not-Active Married            0   male  836
## 6              No Not-Active Married            0 female 3842
```

```r
dim(df)
```

```
## [1] 7582   14
```

```r
# We have 7582 rows and 14 columns
```

```r
summary(df)
```

```
##        X                  age             bmi           children    
##  Min.   :        1   Min.   :18.00   Min.   :15.96   Min.   :0.000  
##  1st Qu.:     5635   1st Qu.:26.00   1st Qu.:26.60   1st Qu.:0.000  
##  Median :    24916   Median :39.00   Median :30.50   Median :1.000  
##  Mean   :   712602   Mean   :38.89   Mean   :30.80   Mean   :1.109  
##  3rd Qu.:   118486   3rd Qu.:51.00   3rd Qu.:34.77   3rd Qu.:2.000  
##  Max.   :131101111   Max.   :66.00   Max.   :53.13   Max.   :5.000  
##                                      NA's   :78                     
##     smoker            location         location_type      education_level   
##  Length:7582        Length:7582        Length:7582        Length:7582       
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##  yearly_physical      exercise           married           hypertension   
##  Length:7582        Length:7582        Length:7582        Min.   :0.0000  
##  Class :character   Class :character   Class :character   1st Qu.:0.0000  
##  Mode  :character   Mode  :character   Mode  :character   Median :0.0000  
##                                                           Mean   :0.2005  
##                                                           3rd Qu.:0.0000  
##                                                           Max.   :1.0000  
##                                                           NA's   :80      
##     gender               cost      
##  Length:7582        Min.   :    2  
##  Class :character   1st Qu.:  970  
##  Mode  :character   Median : 2500  
##                     Mean   : 4043  
##                     3rd Qu.: 4775  
##                     Max.   :55715  
## 
```

```r
str(df)
```

```
## 'data.frame':    7582 obs. of  14 variables:
##  $ X              : int  1 2 3 4 5 7 9 10 11 12 ...
##  $ age            : int  18 19 27 34 32 47 36 59 24 61 ...
##  $ bmi            : num  27.9 33.8 33 22.7 28.9 ...
##  $ children       : int  0 1 3 0 0 1 2 0 0 0 ...
##  $ smoker         : chr  "yes" "no" "no" "no" ...
##  $ location       : chr  "CONNECTICUT" "RHODE ISLAND" "MASSACHUSETTS" "PENNSYLVANIA" ...
##  $ location_type  : chr  "Urban" "Urban" "Urban" "Country" ...
```

```
##  $ education_level: chr  "Bachelor" "Bachelor" "Master" "Master" ...
##  $ yearly_physical: chr  "No" "No" "No" "No" ...
##  $ exercise       : chr  "Active" "Not-Active" "Active" "Not-Active" ...
##  $ married        : chr  "Married" "Married" "Married" "Married" ...
##  $ hypertension   : int  0 0 0 1 0 0 0 1 0 0 ...
##  $ gender         : chr  "female" "male" "male" "male" ...
##  $ cost           : int  1746 602 576 5562 836 3842 1304 9724 201 4492 ...
```

```
# Checking for null values
colSums(sapply(df,is.na))
```

```
##               X             age             bmi        children          smoker
##               0               0              78               0               0
##        location   location_type education_level yearly_physical        exercise
##               0               0               0               0               0
##         married    hypertension          gender            cost
##               0              80               0               0
```

```
# We have 78 null values in BMI and 80 in hypertension.
```

```
#See all rows with atleat one null value
#Also we do not have NA values in same row for both hypertension and bmi, it's either null for bmi or hy
df %>% filter(if_any(everything(),is.na))
```

```
##          X age    bmi children smoker      location location_type
## 1       23  19     NA        0     no  PENNSYLVANIA         Urban
## 2       39  35     NA        1    yes  RHODE ISLAND         Urban
## 3      123  19     NA        0     no      MARYLAND         Urban
## 4      156  42 39.520        0     no MASSACHUSETTS         Urban
## 5      221  34 33.700        1     no  PENNSYLVANIA       Country
## 6      312  19     NA        0     no  PENNSYLVANIA         Urban
## 7      387  59     NA        0     no  PENNSYLVANIA       Country
## 8      425  48 30.200        2     no      NEW YORK         Urban
## 9      440  26     NA        0     no      MARYLAND       Country
## 10     585  20 20.700        0     no      NEW YORK       Country
## 11     598  35 33.250        1     no   CONNECTICUT       Country
## 12     682  19     NA        0     no  PENNSYLVANIA         Urban
## 13     724  19     NA        0     no      MARYLAND         Urban
## 14     747  34 27.000        2     no      NEW YORK       Country
## 15     771  60 36.100        3     no      MARYLAND       Country
## 16     892  36 29.040        4     no    NEW JERSEY         Urban
## 17    1015  37     NA        0     no  PENNSYLVANIA         Urban
## 18    1092  57     NA        0     no  RHODE ISLAND       Country
## 19    1205  18 27.280        3    yes    NEW JERSEY       Country
## 20    1218  29 37.290        2     no  PENNSYLVANIA       Country
## 21    1240  25 42.130        1     no      NEW YORK         Urban
## 22    1271  26 33.915        1     no MASSACHUSETTS         Urban
## 23    1282  47 27.645        2    yes      NEW YORK         Urban
## 24    1314  20 34.700        2    yes  PENNSYLVANIA       Country
## 25    8311  62     NA        0     no  PENNSYLVANIA       Country
## 26   11111  54     NA        1     no   CONNECTICUT         Urban
## 27    8691  60 23.655        0     no   CONNECTICUT         Urban
```

```
## 28    13021  62     NA         3    yes    NEW JERSEY       Country
## 29    11031  28 38.940         1     no  PENNSYLVANIA       Country
## 30     2281  57     NA         0     no  RHODE ISLAND         Urban
## 31     4681  57 33.820         2     no  PENNSYLVANIA         Urban
## 32     7751  42     NA         2     no   CONNECTICUT       Country
## 33     5571  46 33.440         1     no      NEW YORK         Urban
## 34     5010  35     NA         1    yes  PENNSYLVANIA         Urban
## 35    12551  33 27.720         0     no   CONNECTICUT         Urban
## 36     8181  24 37.100         3     no    NEW JERSEY         Urban
## 37     3541  34     NA         0     no  PENNSYLVANIA       Country
## 38     9331  47 25.800         5     no   CONNECTICUT       Country
## 39     8231  19     NA         0     no  PENNSYLVANIA       Country
## 40     9642  45     NA         3     no  PENNSYLVANIA         Urban
## 41    11462  51 32.775         3     no  PENNSYLVANIA         Urban
## 42   110411  59     NA         0     no  RHODE ISLAND       Country
## 43     4792  23 36.850         0     no  PENNSYLVANIA       Country
## 44     6262  28     NA         0     no  PENNSYLVANIA         Urban
## 45     5031  51 23.210         1    yes      MARYLAND         Urban
## 46     8462  60 32.450         0    yes   CONNECTICUT         Urban
## 47     8100  36     NA         3     no  PENNSYLVANIA         Urban
## 48    12712  25     NA         1     no  RHODE ISLAND         Urban
## 49    57211  18 37.290         1     no  PENNSYLVANIA         Urban
## 50    10622  58     NA         1     no  PENNSYLVANIA       Country
## 51    95011  25 29.700         3    yes      MARYLAND         Urban
## 52    22621  55     NA         3     no  PENNSYLVANIA         Urban
## 53   935111  34     NA         2     no  PENNSYLVANIA         Urban
## 54    11181  24 33.330         2    yes  PENNSYLVANIA         Urban
## 55    12262  34 39.820         1     no  PENNSYLVANIA         Urban
## 56    11342  53     NA         0     no  PENNSYLVANIA         Urban
## 57    45312  24 23.400         0     no  RHODE ISLAND         Urban
## 58    39411  49 31.350         1     no MASSACHUSETTS         Urban
## 59     9373  44 29.735         2     no   CONNECTICUT         Urban
## 60     2122  40     NA         4     no  PENNSYLVANIA       Country
## 61     7401  29     NA         2    yes  PENNSYLVANIA         Urban
## 62    11202  28 19.950         3     no  PENNSYLVANIA         Urban
## 63    13362  18     NA         0     no      NEW YORK         Urban
## 64     5851  19 20.700         0     no  RHODE ISLAND         Urban
## 65    92511  44     NA         0     no      MARYLAND         Urban
## 66    98111  53     NA         1     no   CONNECTICUT         Urban
## 67   115611  37 22.135         3     no  PENNSYLVANIA         Urban
## 68   125611  42 37.900         0     no  PENNSYLVANIA       Country
## 69     6472  39 26.220         1     no MASSACHUSETTS       Country
## 70     8863  33 28.930         1    yes  PENNSYLVANIA         Urban
## 71    84421  57 29.810         0    yes  PENNSYLVANIA         Urban
## 72    91121  22 39.490         0     no  RHODE ISLAND       Country
## 73     6515  49 42.680         2     no  RHODE ISLAND         Urban
## 74    87621  22 28.120         0     no  PENNSYLVANIA         Urban
## 75     4113  25     NA         0     no    NEW JERSEY       Country
## 76     9092  63     NA         3     no  PENNSYLVANIA         Urban
## 77    72911  18 40.280         0     no  RHODE ISLAND         Urban
## 78   103121  46     NA         1    yes    NEW JERSEY         Urban
## 79     1763  63     NA         0    yes  PENNSYLVANIA         Urban
## 80     6013  18 39.160         0     no MASSACHUSETTS       Country
## 81    42721  37 27.265         1     no  PENNSYLVANIA         Urban
```

```
## 82     6184 49     NA          2   yes PENNSYLVANIA          Urban
## 83   598111 32 33.250          1    no     NEW YORK          Urban
## 84  5721111 18     NA          1    no PENNSYLVANIA          Urban
## 85    66411 18     NA          0    no PENNSYLVANIA          Urban
## 86     7811 31 24.400          3   yes RHODE ISLAND          Urban
## 87     4383 36     NA          3    no PENNSYLVANIA          Urban
## 88    52111 23 33.630          2    no PENNSYLVANIA          Urban
## 89     8851 25     NA          4    no PENNSYLVANIA          Urban
## 90   104621 43     NA          2   yes PENNSYLVANIA          Urban
## 91    12281 42 37.180          2    no PENNSYLVANIA        Country
## 92    11691 32 35.200          2    no RHODE ISLAND          Urban
## 93   112511 24 42.750          1   yes MASSACHUSETTS         Urban
## 94    83103 22 37.620          1   yes     NEW YORK          Urban
## 95  1161112 42 34.580          1    no PENNSYLVANIA        Country
## 96    53821 47     NA          2    no     NEW YORK        Country
## 97   626111 29     NA          0    no PENNSYLVANIA        Country
## 98   122021 37 30.210          3    no PENNSYLVANIA          Urban
## 99  5661111 19 30.495          0    no PENNSYLVANIA          Urban
## 100    8082 19     NA          0    no     MARYLAND          Urban
## 101  101612 59 25.460          0    no  CONNECTICUT          Urban
## 102  778112 44     NA          0    no PENNSYLVANIA          Urban
## 103  113313 58     NA          0    no     MARYLAND        Country
## 104    6333 29     NA          0    no PENNSYLVANIA          Urban
## 105    4715 27 32.670          0    no PENNSYLVANIA        Country
## 106 5771111 22 26.840          0    no     MARYLAND          Urban
## 107   11394 33 30.250          0    no MASSACHUSETTS       Country
## 108   49911 43 23.980          2    no PENNSYLVANIA          Urban
## 109    1814 23     NA          0    no PENNSYLVANIA          Urban
## 110  122522 42     NA          1    no     MARYLAND        Country
## 111  773111 45     NA          0    no  CONNECTICUT        Country
## 112  473111 19 29.800          0    no PENNSYLVANIA          Urban
## 113  100812 45     NA          3   yes     MARYLAND        Country
## 114   33103 19     NA          5    no PENNSYLVANIA          Urban
## 115   34631 32     NA          3    no  CONNECTICUT        Country
## 116   39111 49 35.625          4    no PENNSYLVANIA        Country
## 117  331211 19     NA          5    no MASSACHUSETTS         Urban
## 118  800111 32     NA          0   yes PENNSYLVANIA          Urban
## 119  986112 45 25.800          1    no PENNSYLVANIA          Urban
## 120   87112 49 36.200          0    no  CONNECTICUT          Urban
## 121   31122 22 35.600          0   yes  CONNECTICUT          Urban
## 122  194112 55     NA          1    no MASSACHUSETTS       Country
## 123  880311 37     NA          2    no     NEW YORK          Urban
## 124   42613 45 24.310          5    no PENNSYLVANIA        Country
## 125   12612 33     NA          0    no  CONNECTICUT          Urban
## 126  808111 18     NA          0    no PENNSYLVANIA          Urban
## 127  484111 53 39.500          1    no PENNSYLVANIA          Urban
## 128   49013 54 31.160          1    no PENNSYLVANIA        Country
## 129    2202 25     NA          0    no PENNSYLVANIA        Country
## 130  363221 20     NA          0   yes     MARYLAND          Urban
## 131  115132 18 30.305          0    no PENNSYLVANIA        Country
## 132 3481111 46 33.345          1    no PENNSYLVANIA          Urban
## 133   21711 53     NA          0    no   NEW JERSEY          Urban
## 134   43141 19     NA          0    no     MARYLAND          Urban
## 135   32831 44     NA          2   yes PENNSYLVANIA          Urban
```

```
## 136   27541  23 27.550         0     no MASSACHUSETTS         Urban
## 137   79022  62 29.920         0     no      NEW YORK         Urban
## 138   96432  45     NA         3     no  RHODE ISLAND         Urban
## 139   99612  40 23.275         3     no MASSACHUSETTS         Urban
## 140   54213  21     NA         2     no  PENNSYLVANIA         Urban
## 141   12895  19 39.400         2    yes  PENNSYLVANIA       Country
## 142    6844  53 24.320         0     no  PENNSYLVANIA         Urban
## 143 1441111  28     NA         2     no   CONNECTICUT         Urban
## 144  100313  25     NA         0     no  PENNSYLVANIA       Country
## 145 1286211  48     NA         0     no  PENNSYLVANIA         Urban
## 146    3321  52 27.360         0    yes   CONNECTICUT         Urban
## 147    3834  54     NA         0     no  PENNSYLVANIA         Urban
## 148    1843  44     NA         0     no  PENNSYLVANIA         Urban
## 149  181211  58 28.595         0     no      NEW YORK         Urban
## 150  804121  18     NA         0    yes  RHODE ISLAND       Country
## 151   12533  20 27.300         0    yes  PENNSYLVANIA       Country
## 152   11952  31     NA         0     no  PENNSYLVANIA         Urban
## 153    9122  19 31.730         0    yes      MARYLAND       Country
## 154  984112  29     NA         1     no  PENNSYLVANIA         Urban
## 155   19511  18     NA         0     no  PENNSYLVANIA         Urban
## 156  520111  32     NA         0     no  PENNSYLVANIA         Urban
## 157 3072111  27     NA         2     no  PENNSYLVANIA         Urban
## 158  811112  45 30.800         3     no  PENNSYLVANIA         Urban
##       education_level yearly_physical   exercise     married hypertension
## 1   No College Degree              No     Active Not_Married            0
## 2            Bachelor             Yes Not-Active     Married            0
## 3                 PhD             Yes Not-Active Not_Married            0
## 4            Bachelor             Yes Not-Active     Married           NA
## 5              Master             Yes Not-Active Not_Married           NA
## 6              Master              No Not-Active     Married            0
## 7              Master              No     Active Not_Married            0
## 8            Bachelor              No Not-Active     Married           NA
## 9            Bachelor              No     Active     Married            0
## 10           Bachelor              No Not-Active     Married           NA
## 11           Bachelor             Yes     Active Not_Married           NA
## 12             Master              No Not-Active     Married            0
## 13             Master              No Not-Active     Married            1
## 14             Master              No Not-Active     Married           NA
## 15           Bachelor              No Not-Active     Married           NA
## 16  No College Degree              No Not-Active     Married           NA
## 17           Bachelor              No Not-Active     Married            0
## 18           Bachelor             Yes Not-Active     Married            0
## 19           Bachelor             Yes Not-Active Not_Married           NA
## 20  No College Degree              No Not-Active Not_Married           NA
## 21  No College Degree              No Not-Active     Married           NA
## 22           Bachelor              No Not-Active Not_Married           NA
## 23           Bachelor              No     Active     Married           NA
## 24           Bachelor              No     Active     Married           NA
## 25           Bachelor              No Not-Active     Married            0
## 26  No College Degree              No Not-Active Not_Married            0
## 27           Bachelor              No     Active     Married           NA
## 28             Master             Yes     Active     Married            0
## 29                PhD             Yes     Active     Married           NA
## 30           Bachelor              No Not-Active     Married            0
```

```
## 31          Bachelor             Yes Not-Active Not_Married           NA
## 32          Bachelor             Yes Not-Active Not_Married            1
## 33          Bachelor             Yes Not-Active     Married           NA
## 34            Master              No Not-Active Not_Married            0
## 35            Master              No Not-Active     Married           NA
## 36          Bachelor              No Not-Active     Married           NA
## 37               PhD              No     Active     Married            0
## 38            Master              No Not-Active     Married           NA
## 39          Bachelor              No Not-Active Not_Married            0
## 40          Bachelor              No Not-Active     Married            0
## 41          Bachelor             Yes Not-Active     Married           NA
## 42          Bachelor              No     Active     Married            0
## 43               PhD             Yes Not-Active     Married           NA
## 44          Bachelor             Yes Not-Active     Married            1
## 45          Bachelor              No Not-Active     Married           NA
## 46          Bachelor              No Not-Active     Married           NA
## 47          Bachelor              No     Active     Married            0
## 48          Bachelor              No Not-Active     Married            0
## 49          Bachelor              No Not-Active     Married           NA
## 50          Bachelor             Yes Not-Active     Married            0
## 51          Bachelor              No Not-Active     Married           NA
## 52          Bachelor              No Not-Active     Married            0
## 53 No College Degree              No     Active Not_Married            0
## 54          Bachelor              No Not-Active     Married           NA
## 55          Bachelor              No Not-Active     Married           NA
## 56          Bachelor              No Not-Active Not_Married            0
## 57          Bachelor             Yes Not-Active Not_Married           NA
## 58          Bachelor              No     Active     Married           NA
## 59 No College Degree              No     Active Not_Married           NA
## 60          Bachelor              No Not-Active     Married            0
## 61          Bachelor              No     Active Not_Married            1
## 62          Bachelor              No Not-Active     Married           NA
## 63               PhD              No Not-Active     Married            0
## 64            Master              No Not-Active     Married           NA
## 65          Bachelor              No Not-Active     Married            0
## 66          Bachelor              No     Active     Married            0
## 67          Bachelor              No Not-Active     Married           NA
## 68            Master              No     Active     Married           NA
## 69          Bachelor              No Not-Active     Married           NA
## 70          Bachelor              No     Active     Married           NA
## 71            Master              No Not-Active Not_Married           NA
## 72               PhD             Yes Not-Active Not_Married           NA
## 73            Master             Yes Not-Active     Married           NA
## 74          Bachelor              No Not-Active     Married           NA
## 75          Bachelor             Yes     Active Not_Married            0
## 76            Master              No     Active Not_Married            0
## 77          Bachelor              No Not-Active Not_Married           NA
## 78          Bachelor             Yes Not-Active     Married            0
## 79               PhD              No     Active     Married            0
## 80               PhD              No Not-Active     Married           NA
## 81               PhD             Yes Not-Active Not_Married           NA
## 82          Bachelor             Yes Not-Active     Married            0
## 83          Bachelor              No Not-Active     Married           NA
## 84 No College Degree              No     Active     Married            0
```

```
## 85           Bachelor              No     Active     Married            0
## 86  No College Degree              No Not-Active Not_Married           NA
## 87           Bachelor              No Not-Active Not_Married            1
## 88           Bachelor             Yes Not-Active     Married           NA
## 89           Bachelor              No Not-Active Not_Married            1
## 90           Bachelor              No Not-Active     Married            1
## 91           Bachelor              No Not-Active Not_Married           NA
## 92             Master             Yes Not-Active     Married           NA
## 93           Bachelor              No     Active     Married           NA
## 94           Bachelor              No Not-Active     Married           NA
## 95           Bachelor              No Not-Active     Married           NA
## 96                PhD             Yes Not-Active Not_Married            0
## 97           Bachelor             Yes Not-Active     Married            0
## 98             Master             Yes     Active Not_Married           NA
## 99           Bachelor              No     Active     Married           NA
## 100          Bachelor             Yes Not-Active Not_Married            0
## 101          Bachelor              No Not-Active Not_Married           NA
## 102          Bachelor              No Not-Active     Married            0
## 103               PhD              No Not-Active Not_Married            1
## 104               PhD              No     Active     Married            1
## 105            Master              No Not-Active     Married           NA
## 106          Bachelor              No Not-Active Not_Married           NA
## 107 No College Degree              No     Active     Married           NA
## 108            Master              No Not-Active     Married           NA
## 109            Master              No Not-Active     Married            0
## 110            Master              No     Active Not_Married            0
## 111          Bachelor              No     Active Not_Married            0
## 112            Master             Yes Not-Active     Married           NA
## 113          Bachelor              No Not-Active Not_Married            0
## 114               PhD             Yes Not-Active     Married            0
## 115            Master             Yes Not-Active     Married            0
## 116          Bachelor              No Not-Active     Married           NA
## 117          Bachelor             Yes Not-Active     Married            0
## 118            Master              No Not-Active     Married            0
## 119          Bachelor              No Not-Active     Married           NA
## 120            Master              No Not-Active Not_Married           NA
## 121          Bachelor              No Not-Active     Married           NA
## 122          Bachelor              No Not-Active Not_Married            0
## 123          Bachelor             Yes Not-Active     Married            0
## 124          Bachelor             Yes Not-Active     Married           NA
## 125 No College Degree              No Not-Active     Married            0
## 126            Master              No     Active     Married            0
## 127 No College Degree              No Not-Active Not_Married           NA
## 128          Bachelor             Yes Not-Active     Married           NA
## 129          Bachelor              No Not-Active     Married            0
## 130               PhD             Yes     Active     Married            0
## 131          Bachelor              No Not-Active Not_Married           NA
## 132          Bachelor             Yes     Active     Married           NA
## 133          Bachelor              No Not-Active Not_Married            0
## 134            Master              No Not-Active     Married            0
## 135 No College Degree              No Not-Active     Married            1
## 136          Bachelor              No Not-Active     Married           NA
## 137          Bachelor              No Not-Active     Married           NA
## 138            Master              No     Active     Married            0
```

```
## 139          Bachelor              No Not-Active Not_Married           NA
## 140          Bachelor              No     Active     Married            0
## 141          Bachelor              No Not-Active     Married           NA
## 142               PhD              No Not-Active     Married           NA
## 143 No College Degree             Yes Not-Active     Married            1
## 144          Bachelor              No Not-Active Not_Married            0
## 145            Master              No     Active     Married            1
## 146          Bachelor              No     Active     Married           NA
## 147 No College Degree             Yes Not-Active Not_Married            0
## 148          Bachelor             Yes Not-Active     Married            0
## 149          Bachelor              No Not-Active Not_Married           NA
## 150 No College Degree              No Not-Active Not_Married            0
## 151          Bachelor              No Not-Active Not_Married           NA
## 152 No College Degree              No Not-Active     Married            0
## 153          Bachelor              No Not-Active     Married           NA
## 154            Master             Yes Not-Active     Married            1
## 155          Bachelor              No Not-Active     Married            1
## 156            Master             Yes Not-Active     Married            0
## 157          Bachelor             Yes Not-Active Not_Married            0
## 158          Bachelor              No Not-Active     Married           NA
##     gender  cost
## 1     male   146
## 2     male 16448
## 3   female   605
## 4     male  4507
## 5   female  1335
## 6   female   194
## 7   female  2389
## 8     male  3182
## 9     male   556
## 10    male   187
## 11  female  1024
## 12    male   169
## 13    male   322
## 14    male  3694
## 15    male 13036
## 16  female  1939
## 17  female  1855
## 18  female  4503
## 19  female  2812
## 20    male  1052
## 21  female  1245
## 22    male   980
## 23  female  6540
## 24  female  2321
## 25    male  4993
## 26  female  2987
## 27    male  1658
## 28    male 11766
## 29    male   306
## 30  female 12459
## 31  female  5281
## 32    male  6295
## 33    male  6066
```

```
## 34    male 17026
## 35  female  1314
## 36    male   961
## 37    male  2830
## 38    male  4187
## 39  female   515
## 40    male  2185
## 41    male  5175
## 42    male  1021
## 43    male   702
## 44  female  1172
## 45    male  5736
## 46  female 11926
## 47  female  1181
## 48    male  1182
## 49  female   433
## 50    male  3206
## 51    male  2089
## 52    male  2681
## 53    male    64
## 54    male 13908
## 55  female  2013
## 56  female  1948
## 57    male   369
## 58    male  3472
## 59    male  8098
## 60    male  2520
## 61    male  8801
## 62  female  1056
## 63  female   341
## 64    male   171
## 65    male  3046
## 66    male  2515
## 67  female  1858
## 68  female  1882
## 69    male  1998
## 70    male  2956
## 71  female  4214
## 72  female   547
## 73  female  3098
## 74  female   804
## 75  female   286
## 76    male  5146
## 77  female   474
## 78  female  8620
## 79  female 14701
## 80  female   565
## 81  female  2170
## 82    male  5306
## 83  female  2268
## 84  female   450
## 85    male   208
## 86    male  4092
## 87    male  2230
```

```
## 88  female   932
## 89    male   934
## 90  female  7904
## 91    male  3198
## 92    male  2318
## 93  female 10612
## 94    male  8250
## 95  female  3718
## 96  female  3119
## 97  female   404
## 98  female  3051
## 99  female   279
## 100 female   472
## 101   male  3369
## 102   male  2147
## 103   male 11074
## 104 female   639
## 105   male   766
## 106   male   123
## 107   male   635
## 108 female  3524
## 109   male   380
## 110   male  1062
## 111 female  1077
## 112 female  1040
## 113   male 16879
## 114 female   823
## 115 female  2704
## 116   male  6097
## 117 female  1286
## 118   male  5696
## 119 female  2288
## 120   male  3652
## 121   male 14022
## 122 female  1725
## 123 female  3312
## 124   male  3173
## 125 female   976
## 126 female   637
## 127 female  7016
## 128   male  4164
## 129 female  5018
## 130 female  1071
## 131 female   956
## 132   male  2509
## 133 female  3820
## 134   male  9257
## 135   male 19752
## 136   male   498
## 137 female  4174
## 138   male  2013
## 139 female  6971
## 140 female   437
## 141   male 10949
```

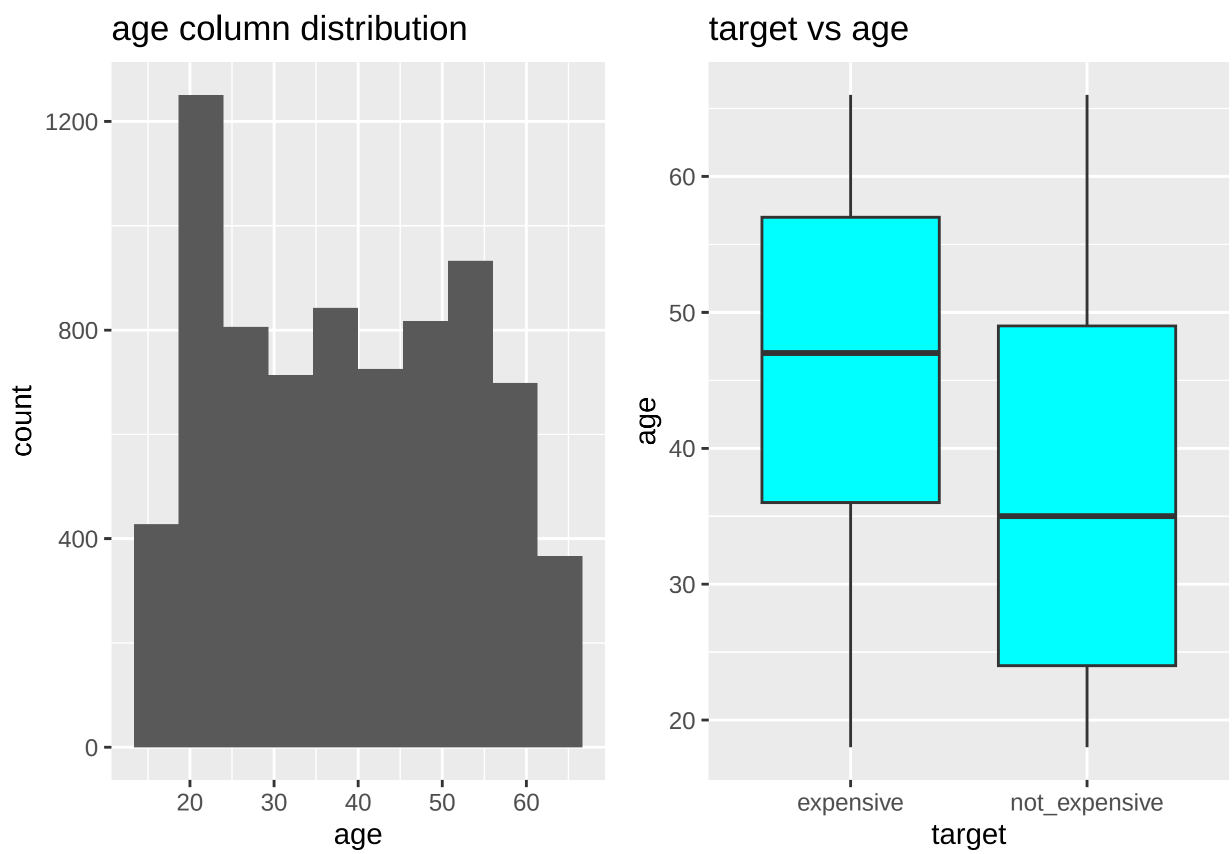
```
## 142   male  2140
## 143   male  4504
## 144   male   318
## 145 female   373
## 146   male  3755
## 147   male  6367
## 148 female  2568
## 149   male  3147
## 150 female 12927
## 151   male  3177
## 152 female  1089
## 153   male  6400
## 154 female  3513
## 155   male   274
## 156   male   877
## 157 female  3913
## 158 female  1836
```

```
# Add new column Expensive or not where cost is more than 4775
df <-df %>% mutate(target=if_else(cost>4775,'expensive','not_expensive'))
```

```
head(df)
```

```
##   X age    bmi children smoker      location location_type education_level
## 1 1  18 27.900        0    yes   CONNECTICUT         Urban        Bachelor
## 2 2  19 33.770        1     no  RHODE ISLAND         Urban        Bachelor
## 3 3  27 33.000        3     no MASSACHUSETTS         Urban          Master
## 4 4  34 22.705        0     no  PENNSYLVANIA       Country          Master
## 5 5  32 28.880        0     no  PENNSYLVANIA       Country             PhD
## 6 7  47 33.440        1     no  PENNSYLVANIA         Urban        Bachelor
##   yearly_physical   exercise married hypertension gender cost        target
## 1              No     Active Married            0 female 1746 not_expensive
## 2              No Not-Active Married            0   male  602 not_expensive
## 3              No     Active Married            0   male  576 not_expensive
## 4              No Not-Active Married            1   male 5562     expensive
## 5              No Not-Active Married            0   male  836 not_expensive
## 6              No Not-Active Married            0 female 3842 not_expensive
```

```
#lets find the distribution of each variable,especially the numeric values
p1<-ggplot(df,aes(x=age))+geom_histogram(bins=10)+ggtitle('age column distribution')


p2<- ggplot(df,aes(x=target,y=age))+geom_boxplot(fill='cyan')+ggtitle('target vs age')
ggarrange(plotlist=list(p1,p2),ncol = 2,nrow=1)
```

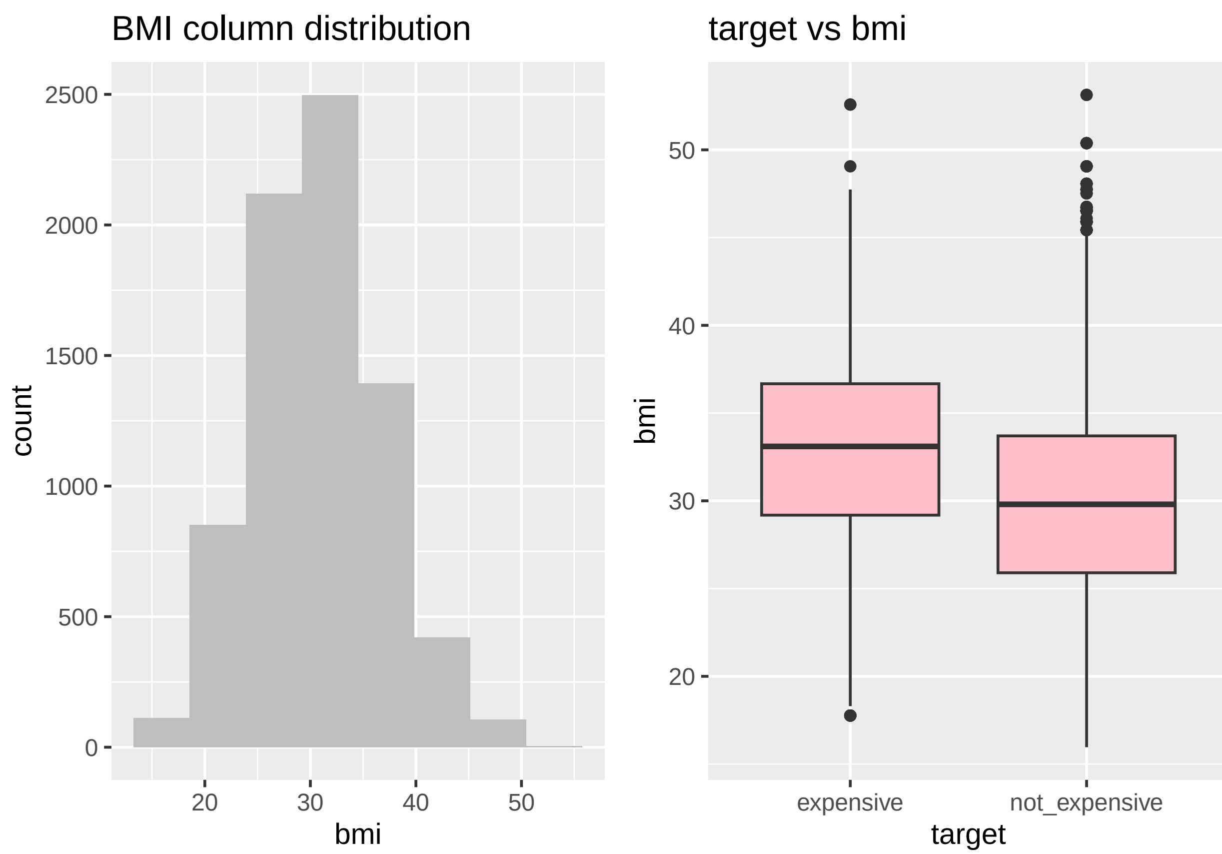



```
# People under 20 are more in count, followed by 20-25 and 45-55, least data is from people of age 65-7
# As seen below, aged people have more medical expense than young ones. Mean age of expensive is around
```

```
p1<-ggplot(df,aes(x=bmi))+geom_histogram(bins=8,fill='grey')+ggtitle('BMI column distribution')


p2<- ggplot(df,aes(x=target,y=bmi))+geom_boxplot(fill='pink')+ggtitle('target vs bmi')
ggarrange(plotlist=list(p1,p2),ncol = 2,nrow=1)
```

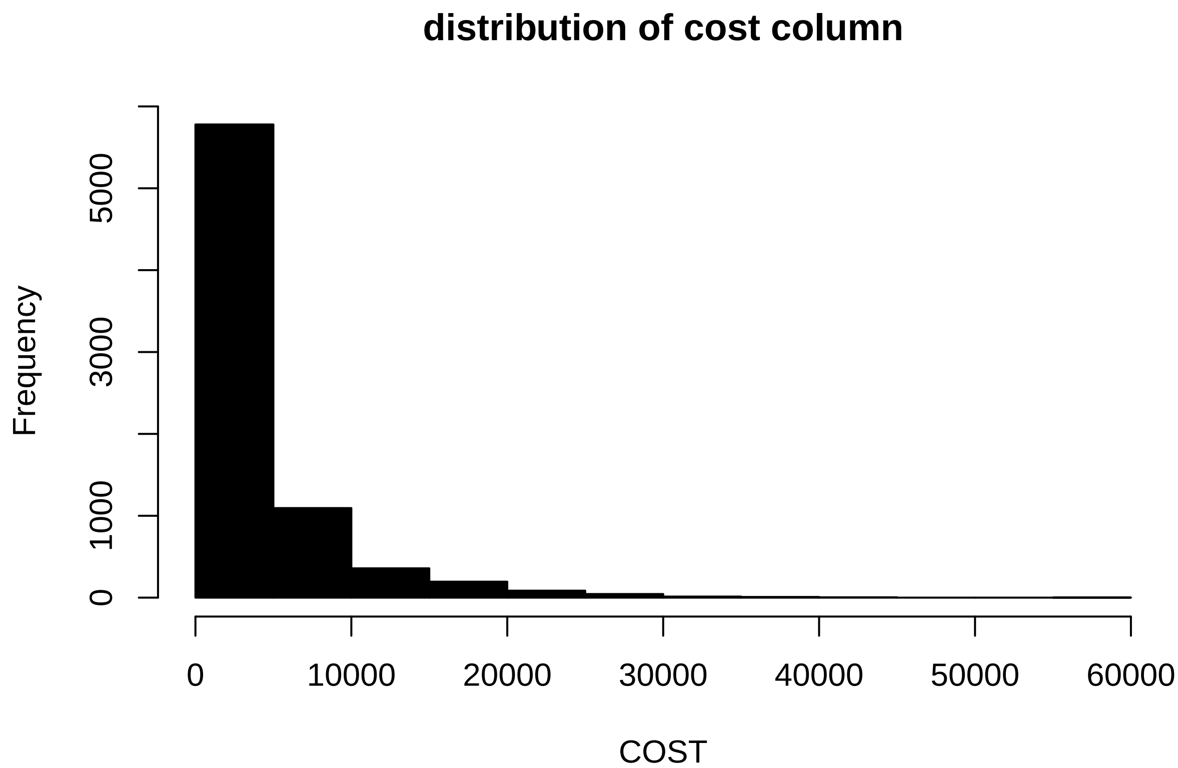
```
## Warning: Removed 78 rows containing non-finite values (‘stat_bin()‘).
```

```
## Warning: Removed 78 rows containing non-finite values (‘stat_boxplot()‘).
```




```
# Its a normal distribiution. Mean BMI is 30, which according to study is obesity zone. Indicating more
# Probably obese people have higher medical expense. mean is 33 , whereas non_exp its is almost 30.
```

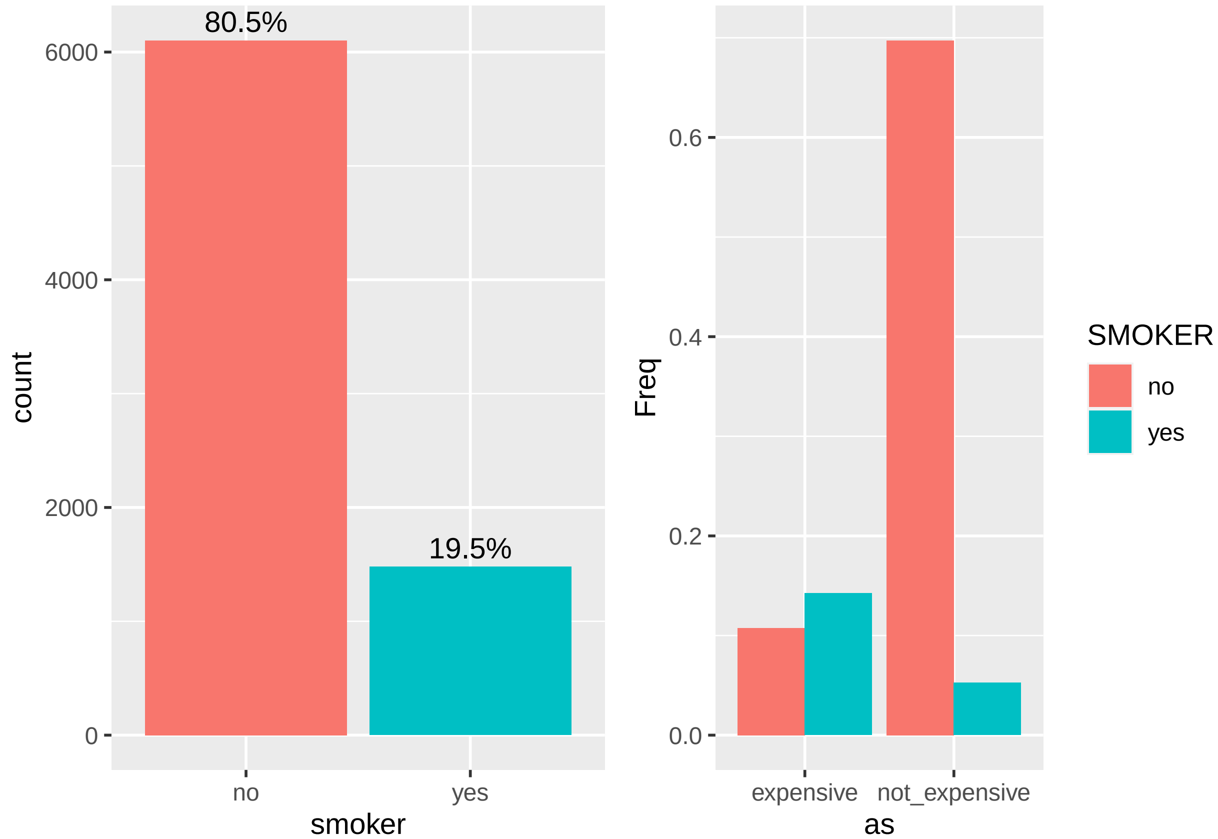
```
hist(df$cost,main='distribution of cost column',col='black',xlab='COST')
```




```
# Target variable is highly skewed. we have mean of 4043 and 3rd quant 4775 and max  55715.
# we need to convert this variable into categorical as expensive or non_expensive.
# going with data we are considering not_expensive till 3rd quartile and above that its expensive.ie(>4
```

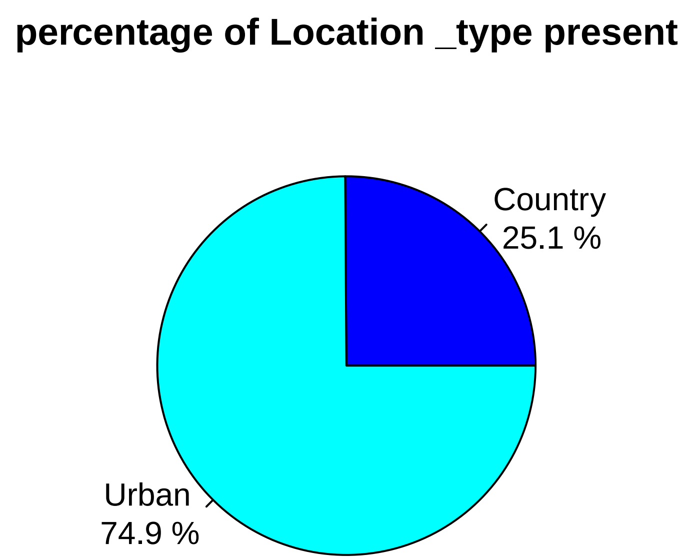
```
p1<-ggplot(df,aes(x=smoker, fill=smoker)) +geom_bar(show.legend = FALSE) +# add percentages on top of b
    geom_text(
        stat='count',
        aes(label=paste0(round(after_stat(prop*100), digits=1), "%"),group=1),
        vjust=-0.4,
        size=4
    )

P <- (prop.table(table(df$smoker,df$target)))
p2 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positi

ggarrange(plotlist=list(p1,p2),ncol=2)
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```

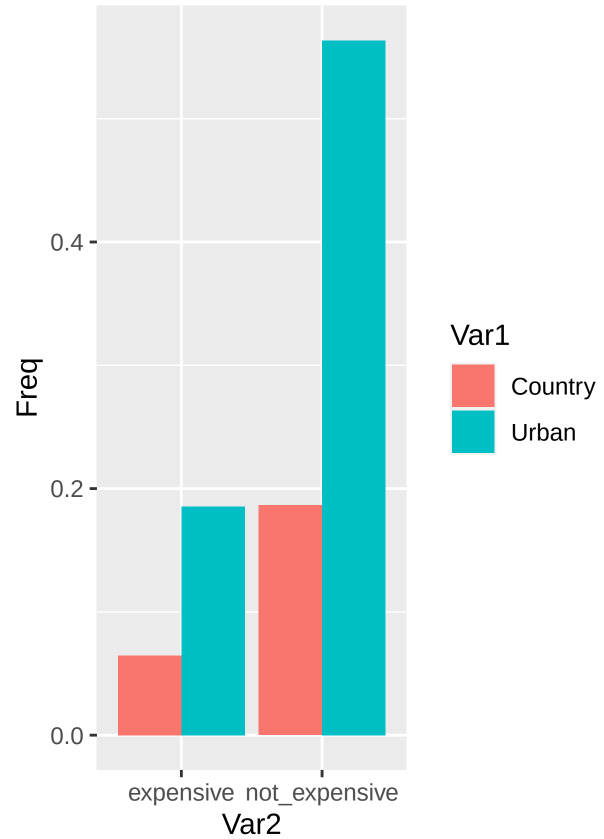



```
#80.5% are non_smokers and 70% of them have medical expense less costly. out of 19.5% smoker, 15% have
#it seems like smoking is related to our target
```

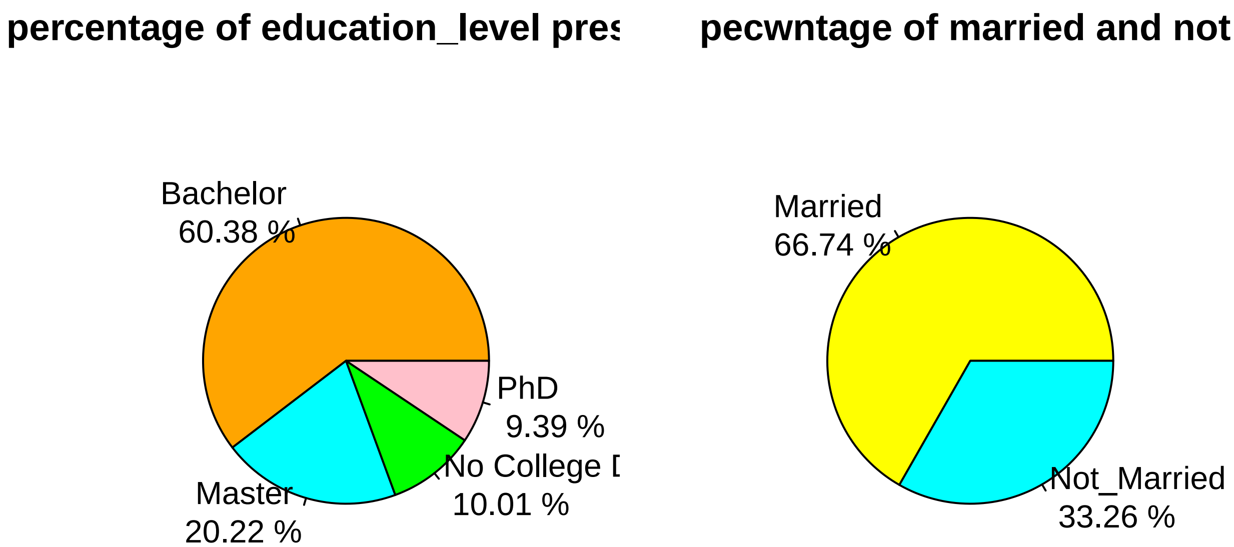
```
mytable = round(prop.table(table(df$location_type))*100,2)
lbls <- paste(names(mytable), "\n", paste(mytable,'%'), sep=" ")
p1<- pie(mytable, labels = lbls,main="percentage of Location _type present",col=c('blue','cyan'))
```



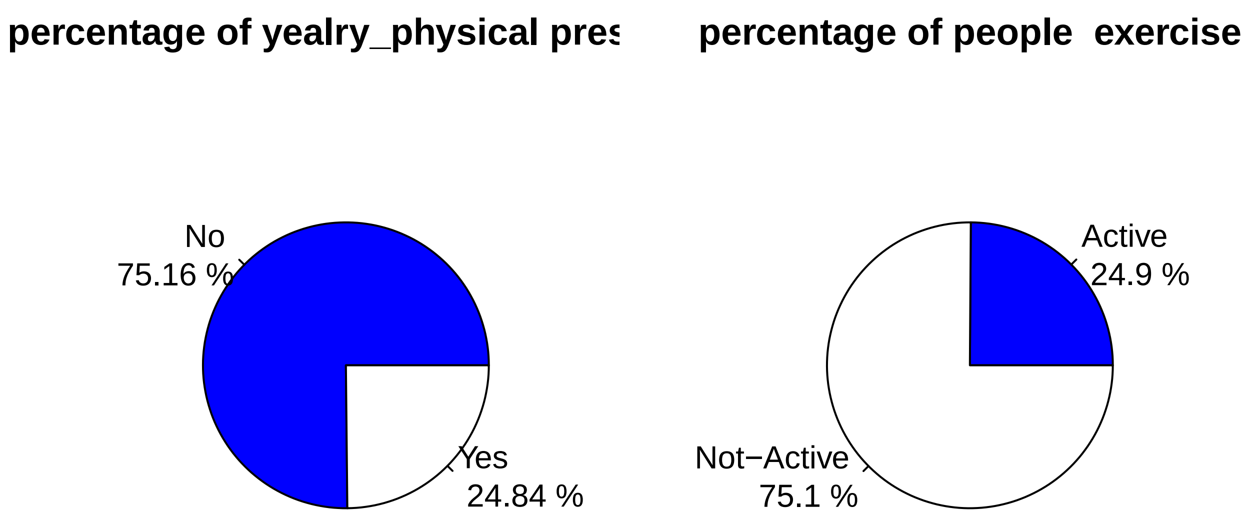

```
P <- as.data.frame(prop.table(table(df$location_type,df$target)))

p2 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positi

cowplot::plot_grid(p1,p2)
```



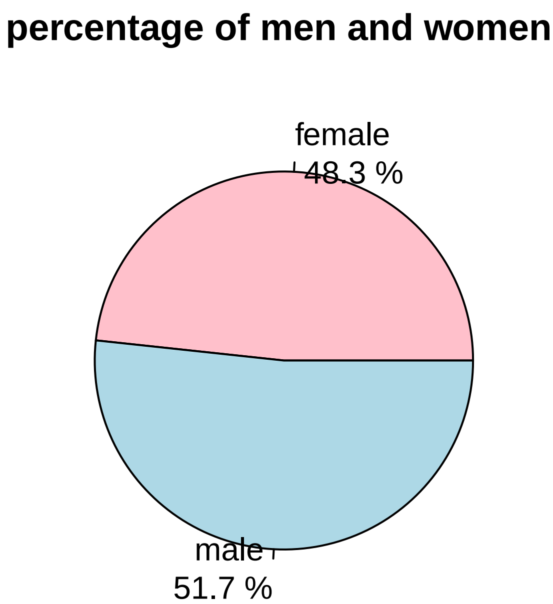

```
par(mfrow=c(1,2))
mytable = round(prop.table(table(df$education_level))*100,2)
lbls <- paste(names(mytable), "\n", paste(mytable,'%'), sep=" ")
pie(mytable, labels = lbls,main="percentage of education_level present",col=c('orange','cyan','green','

mytable = round(prop.table(table(df$married))*100,2)
lbls <- paste(names(mytable), "\n", paste(mytable,'%'), sep=" ")
pie(mytable, labels = lbls,main="pecwntage of married and not ",col=c('yellow','cyan'))
```



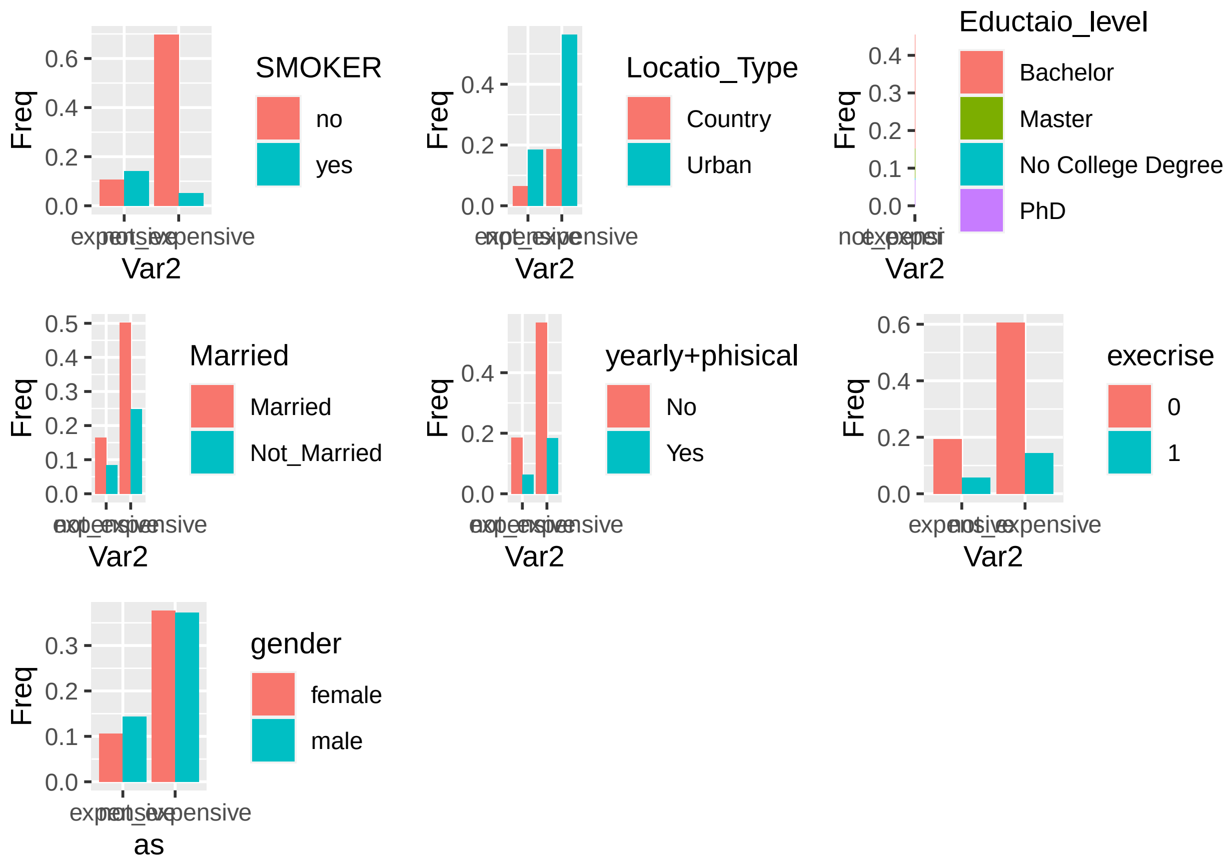

```
figsize <- options(repr.plot.width=10, repr.plot.height=10)
par(mfrow=c(1,2))
mytable = round(prop.table(table(df$yearly_physical))*100,2)
lbls <- paste(names(mytable), "\n", paste(mytable,'%'), sep=" ")
pie(mytable, labels = lbls,
   main="percentage of yealry_physical present",col=c('blue','white'))

mytable = round(prop.table(table(df$exercise))*100,2)
lbls <- paste(names(mytable), "\n", paste(mytable,'%'), sep=" ")
pie(mytable, labels = lbls,
   main="percentage of people  exercise",col=c('blue','white'))
```




```r
mytable = round(prop.table(table(df$gender))*100,2)
lbls <- paste(names(mytable), "\n", paste(mytable,'%'), sep=" ")
pie(mytable, labels = lbls,
   main="percentage of men and women ",col=c('pink','light blue'))
```



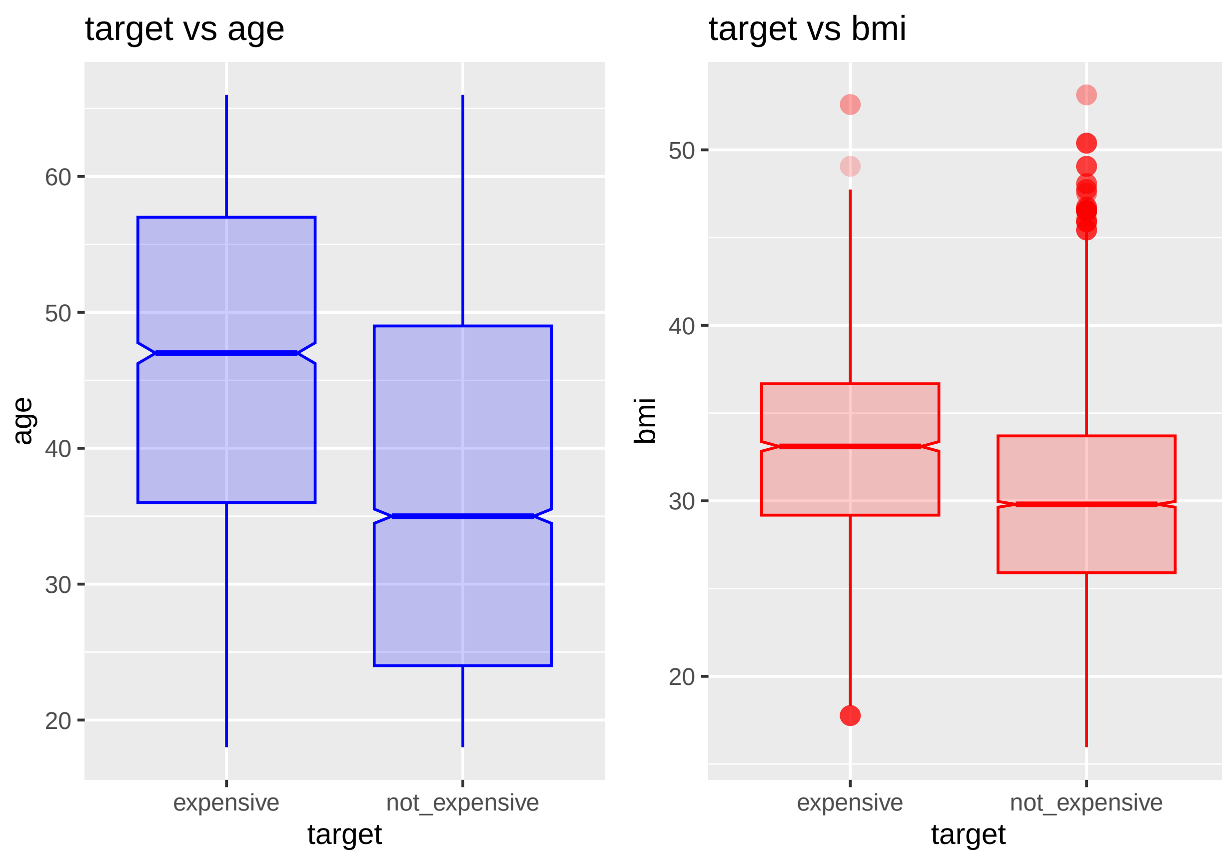

```
P <- (prop.table(table(df$smoker,df$target)))
p1 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positio

P <- (prop.table(table(df$location_type,df$target)))
p2 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positio

P <- (prop.table(table(df$education_level,df$target)))
p3 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positio

P <- (prop.table(table(df$married,df$target)))
p4 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positio

P <- (prop.table(table(df$yearly_physical,df$target)))
p5 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positio

P <- (prop.table(table(df$exercise,df$target)))
p6 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positio

P <- (prop.table(table(df$hypertension,df$target)))
p6 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positio

P <- (prop.table(table(df$gender,df$target)))
p7 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positio

figsize <- options(repr.plot.width=20, repr.plot.height=16)
cowplot::plot_grid(p1,p2,p3,p4,p5,p6,p7,ncol = 3,nrow = 3)
```



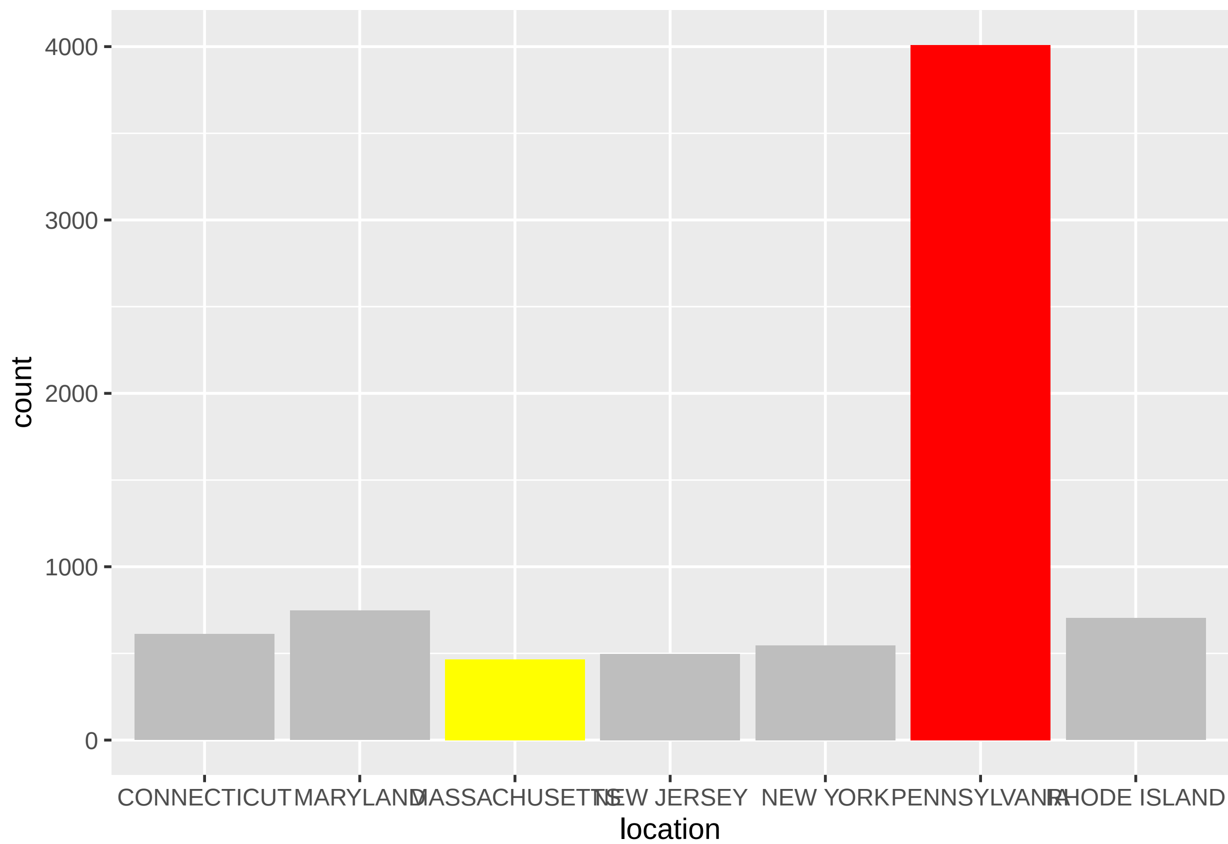

```
#out of 75% urban data, 58% have less medical expenses. remainng 25 % country people, 18% have expensiv
```

```
figsize <- options(repr.plot.width=10, repr.plot.height=10)
p1 <- ggplot(df, aes(x = target,y=age))+geom_boxplot(color="blue",fill="blue",alpha=0.2,notch=TRUE,notc
p2 <- ggplot(df, aes(x = target,y=bmi))+geom_boxplot(color="red",fill="red",alpha=0.2,notch=TRUE,notchw
cowplot::plot_grid(p1,p2,ncol = 2,nrow = 1)
```

```
## Warning: Removed 78 rows containing non-finite values (`stat_boxplot()`).
```




```r
#what are the top 10 locations particiapting in data
ggplot(df, aes(x = location)) +geom_bar(fill=c('grey','grey','yellow','grey','grey','red','grey'))
```

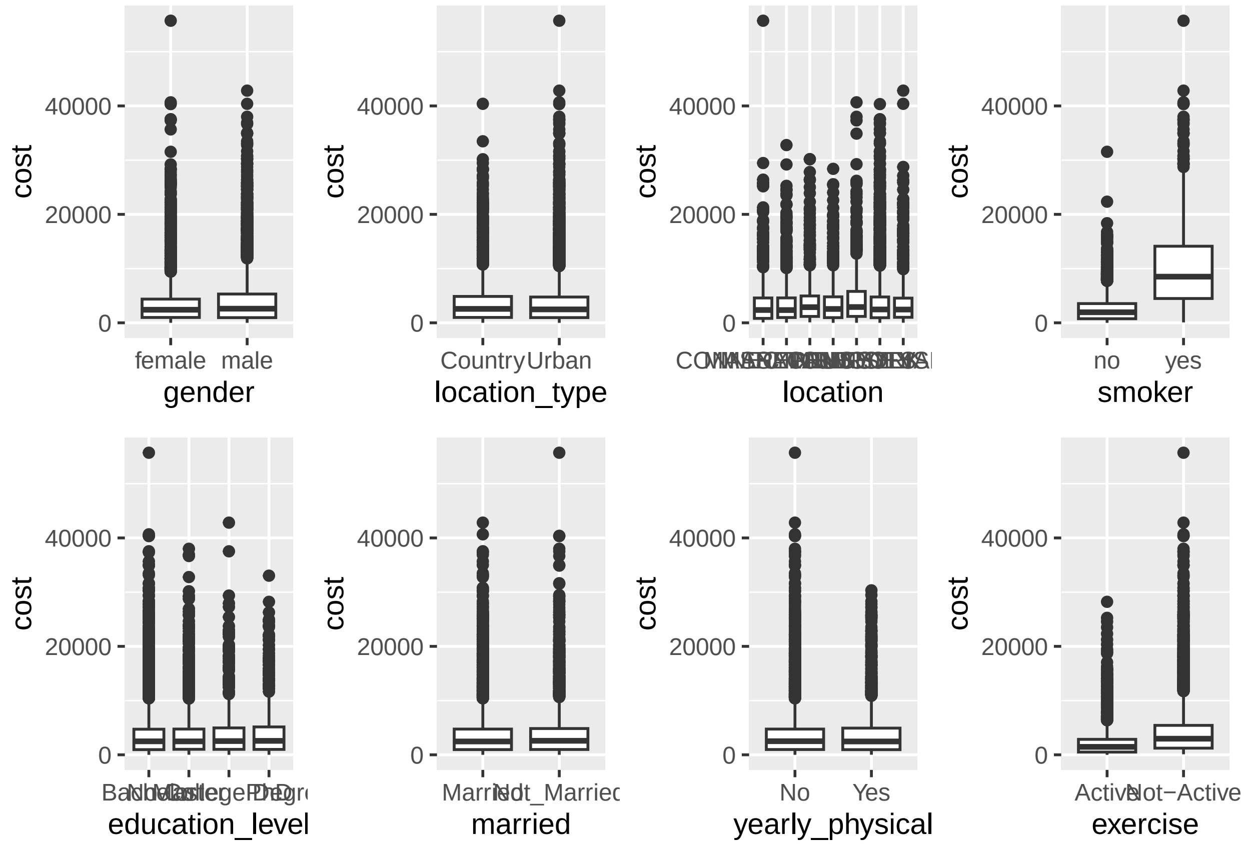



```
#Most of data is from Pennsilvania and least data from Massachusets
```

```
#for visulaizatiuon purpose, we take cost as target
names(df)
```

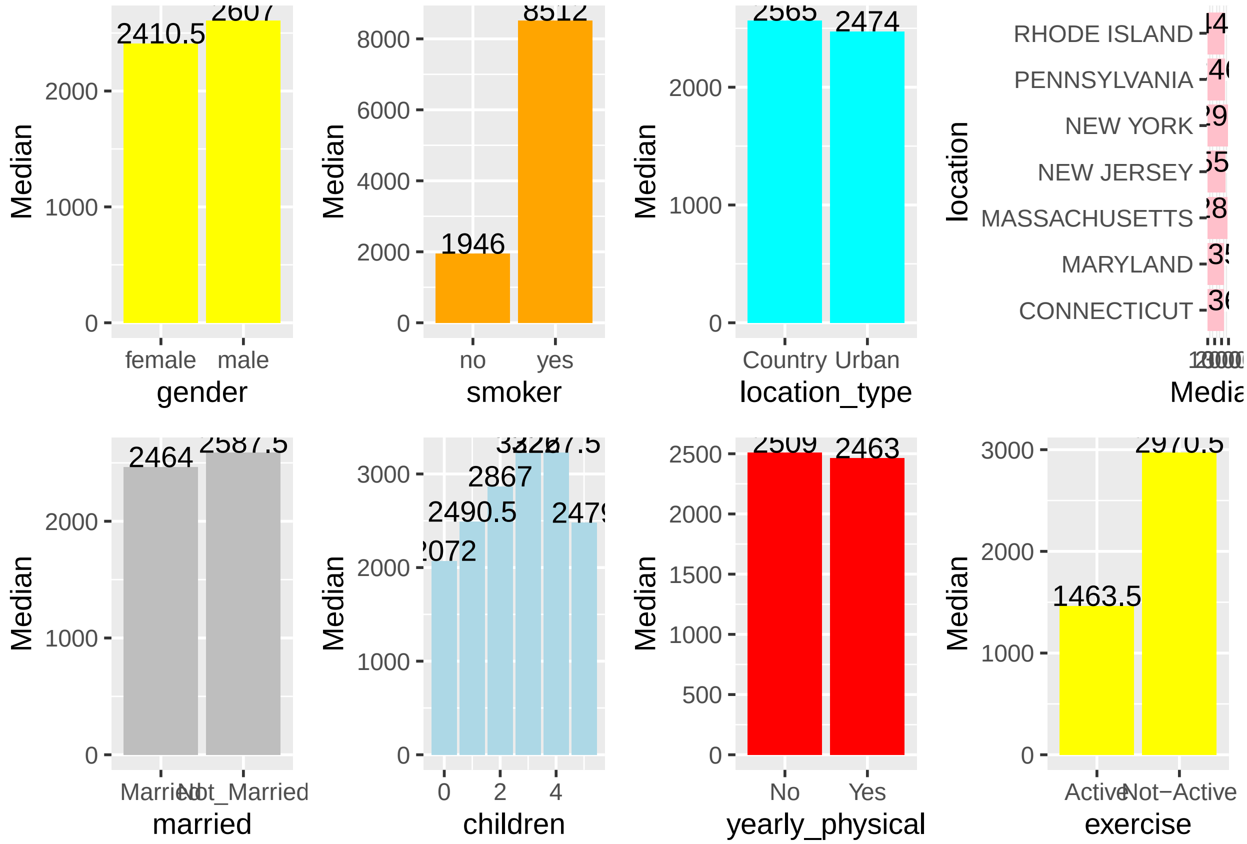
```
##  [1] "X"               "age"             "bmi"             "children"
##  [5] "smoker"          "location"        "location_type"   "education_level"
##  [9] "yearly_physical" "exercise"        "married"         "hypertension"
## [13] "gender"          "cost"            "target"
```

```
p1 <- ggplot(df, aes(x =gender , y=cost)) + geom_boxplot()
p2 <- ggplot(df, aes(x =location_type , y=cost)) + geom_boxplot()
p3 <- ggplot(df, aes(x =location , y=cost)) + geom_boxplot()
p4 <- ggplot(df, aes(x =smoker , y=cost)) + geom_boxplot()
p5 <- ggplot(df, aes(x =education_level , y=cost)) + geom_boxplot()
p6 <- ggplot(df, aes(x =married , y=cost)) + geom_boxplot()
p7 <- ggplot(df, aes(x =yearly_physical , y=cost)) + geom_boxplot()
p8 <- ggplot(df, aes(x =exercise , y=cost)) + geom_boxplot()


figsize <- options(repr.plot.width=20, repr.plot.height=16)
cowplot::plot_grid(p1,p2,p3,p4,p5,p6,p7,p8,ncol = 4,nrow = 2)
```




```
p<-df%>%group_by(gender)%>%summarise(Median =median(cost))
p1<-ggplot(as.data.frame(p),aes(x=gender,y=Median))+geom_bar(stat="identity", position = "dodge",fill='y

p <-df%>%group_by(smoker)%>%summarise(Median =median(cost))
p2 <-ggplot(as.data.frame(p),aes(x=smoker,y=Median))+geom_bar(stat="identity", position = "dodge",fill=

p <-df%>%group_by(location_type)%>%summarise(Median =median(cost))
p3 <-ggplot(as.data.frame(p),aes(x=location_type,y=Median))+geom_bar(stat="identity", position = "dodge


p <-df%>%group_by(location)%>%summarise(Median =median(cost))
p4 <-ggplot(as.data.frame(p),aes(x=location,y=Median))+geom_bar(stat="identity", position = "dodge",fil

p <-df%>%group_by(married)%>%summarise(Median =median(cost))
p5 <-ggplot(as.data.frame(p),aes(x=married,y=Median))+geom_bar(stat="identity", position = "dodge",fill=


p <-df%>%group_by(children)%>%summarise(Median =median(cost))
p6 <-ggplot(as.data.frame(p),aes(x=children,y=Median))+geom_bar(stat="identity", position = "dodge",fil


p <-df%>%group_by(yearly_physical)%>%summarise(Median =median(cost))
p7 <-ggplot(as.data.frame(p),aes(x=yearly_physical,y=Median))+geom_bar(stat="identity", position = "dodg

p <-df%>%group_by(exercise)%>%summarise(Median =median(cost))
p8 <-ggplot(as.data.frame(p),aes(x=exercise,y=Median))+geom_bar(stat="identity", position = "dodge",fil
```

```r
figsize <- options(repr.plot.width=20, repr.plot.height=16)
cowplot::plot_grid(p1,p2,p3,p4,p5,p6,p7,p8,nrow=2,ncol=4)
```

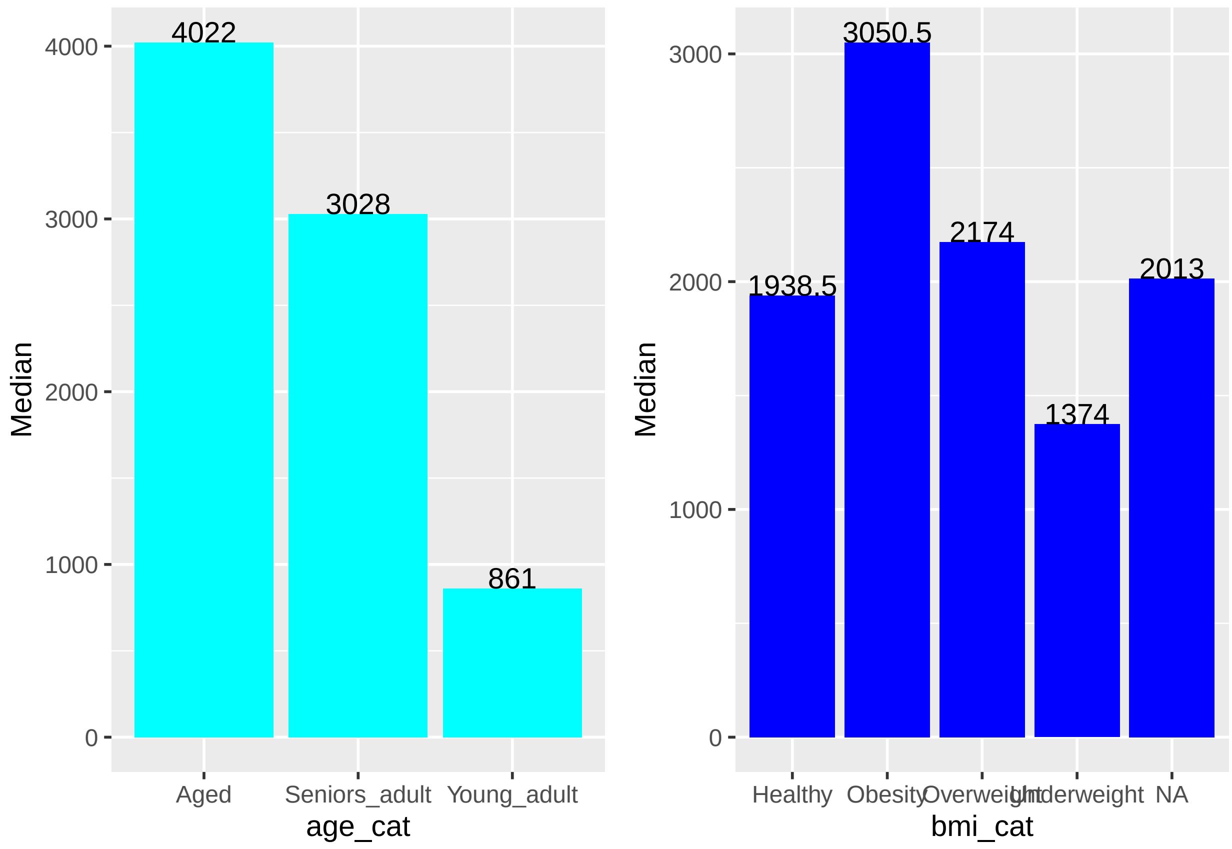



```r
#why did we choose Median??
#Since we are considering Cost column, its highly right skewed. so Mean gets easily affected by Skewene

#Seems Men Median expense is 200$ greater than Female
#Smokers Median expenses is  almost 4 times  that of Non_Smokers
#Country and Urban have almost similar medical expense.
#From the count graph above, we knew Penn has more count and Massachusets has less. But, Median Medical
#Seems Massachusetts is costlier to leave(or is it because of aged and more smokers lest check that aft

#Parents with 3 or 4 have median cost of 3226-27$
#Exercise may stongly relate to cost. Non_Active have to pay almost 3000$ that is 1500$ more than those

#we can make the same plots of age and bmi by subcategorizing them .for ex: age into --young, elderly a
```

```r
df = df%>%mutate(age_cat=case_when(
                                   age>=18 & age<35 ~'Young_adult',
                                   age>=35 & age<50 ~'Seniors_adult',
                                   age>=50  ~'Aged',
))
```

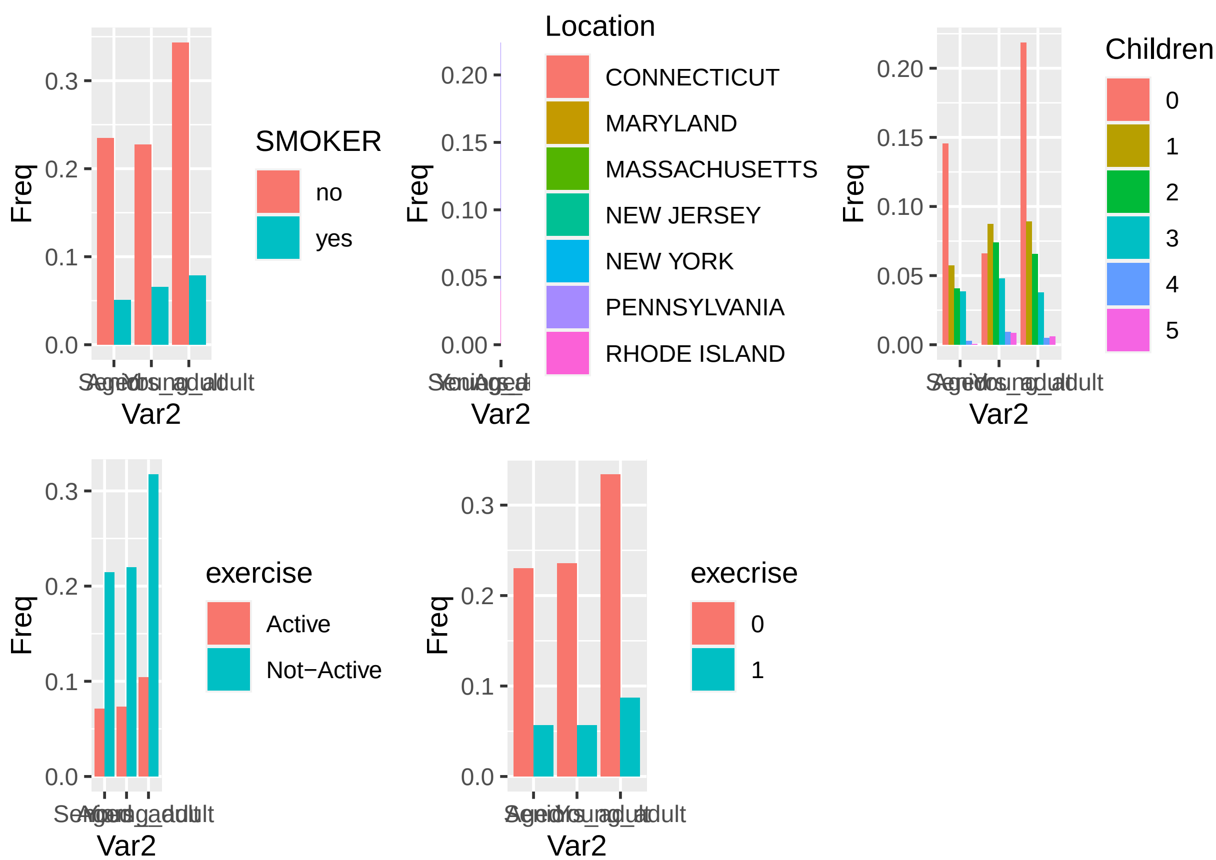
```
#https://www.cdc.gov/obesity/basics/adult-defining.html
#If your BMI is less than 18.5, it falls within the underweight range.
#If your BMI is 18.5 to <25, it falls within the healthy weight range.
#If your BMI is 25.0 to <30, it falls within the overweight range.
#If your BMI is 30.0 or higher, it falls within the obesity range.
df = df%>%mutate(bmi_cat=case_when(
                                bmi<18.5 ~'Underweight',
                                bmi>=18.5 & bmi<25 ~'Healthy',
                                bmi>=25 & bmi<30 ~'Overweight',
                                bmi>30 ~'Obesity'
))
```

```
p <-df%>%group_by(age_cat)%>%summarise(Median =median(cost))
p1 <-ggplot(as.data.frame(p),aes(x=age_cat,y=Median))+geom_bar(stat="identity", position = "dodge",fill=

p <-df%>%group_by(bmi_cat)%>%summarise(Median =median(cost))
p2 <-ggplot(as.data.frame(p),aes(x=bmi_cat,y=Median))+geom_bar(stat="identity", position = "dodge",fill=

figsize <- options(repr.plot.width=8, repr.plot.height=8)
cowplot::plot_grid(p1,p2,nrow=1,ncol=2)
```



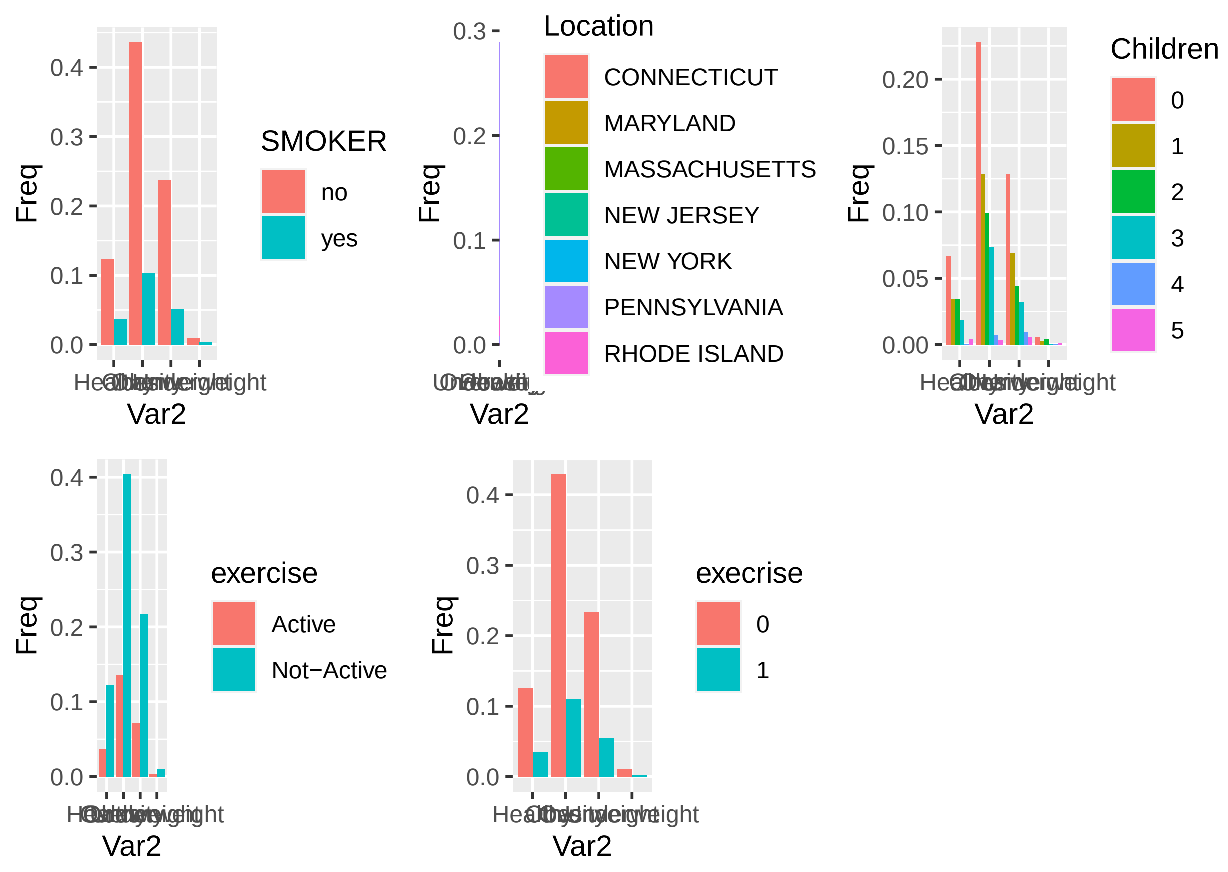

```
#As seen , aged factor and BMI plays a role. Aged person pays higher median cost where as young ones pa
#3050.5$ is median health cost for Obese people followed by overweight and healthy.  Interestingly, Und
```

```r
P <- (prop.table(table(df$smoker,df$age_cat)))
p1 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positio

P <- (prop.table(table(df$location,df$age_cat)))
p2 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positio


P <- (prop.table(table(df$children,df$age_cat)))
p3 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positio

P <- (prop.table(table(df$exercise,df$age_cat)))
p4 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positio

P <- (prop.table(table(df$hypertension,df$age_cat)))
p5 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positio


figsize <- options(repr.plot.width=20, repr.plot.height=16)
cowplot::plot_grid(p1,p2,p3,p4,p5,ncol = 3,nrow = 2)
```




```r
#As we saw from previous plots where exercise,location,smoker and children had affect on Median Cost, S
#Guess was more aged people are smoking , More aged might be living in NY and MASSACHUSETTS,May be they
#but since data has more rows about young and middle age we cant be sure of the above guess/hypothesis
```

```
P <- (prop.table(table(df$smoker,df$bmi_cat)))
p1 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positi

P <- (prop.table(table(df$location,df$bmi_cat)))
p2 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positi


P <- (prop.table(table(df$children,df$bmi_cat)))
p3 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positi

P <- (prop.table(table(df$exercise,df$bmi_cat)))
p4 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positi

P <- (prop.table(table(df$hypertension,df$bmi_cat)))
p5 <- ggplot(as.data.frame(P), aes(x = Var2, y = Freq, fill = Var1)) + geom_bar(stat="identity", positi


figsize <- options(repr.plot.width=20, repr.plot.height=16)
cowplot::plot_grid(p1,p2,p3,p4,p5,ncol = 3,nrow = 2)
```




```
#obese people have more smokers compared to other categories
#Same with Location, Obese people are more in NY and Massachusetts
#they stand 1st with 3 children
#they dont excersie and are not active.
```

```r
#BMI surely playing a direct role in cost involved.
```

```r
# Hypertension and BMI have null values. We can use impute to try adding values
# but that returns values such as 0.5 in BMI which is not possible as that column
# denotes that the person either has hypertension or not.
# We are therefore removing those rows from the dataset.

df <- na.omit(df)
colSums(sapply(df,is.na))
```

```
##               X             age             bmi        children          smoker
##               0               0               0               0               0
##        location   location_type education_level yearly_physical        exercise
##               0               0               0               0               0
##         married    hypertension          gender            cost          target
##               0               0               0               0               0
##         age_cat         bmi_cat
##               0               0
```

```r
# This results in us removing a total of 158 rows.
```

```r
# Convert all char columns to factor
df <- df %>% mutate_if(is.character, as.factor)
# LM does not accept factor so converted it to numeric
df <- df %>% mutate_if(is.factor, as.numeric)
str(df)
```

```
## 'data.frame':    7413 obs. of  17 variables:
##  $ X              : int  1 2 3 4 5 7 9 10 11 12 ...
##  $ age            : int  18 19 27 34 32 47 36 59 24 61 ...
##  $ bmi            : num  27.9 33.8 33 22.7 28.9 ...
##  $ children       : int  0 1 3 0 0 1 2 0 0 0 ...
##  $ smoker         : num  2 1 1 1 1 1 1 1 1 2 ...
##  $ location       : num  1 7 3 6 6 6 6 6 6 1 ...
##  $ location_type  : num  2 2 2 1 1 2 2 1 2 2 ...
##  $ education_level: num  1 1 2 2 4 1 1 1 1 3 ...
##  $ yearly_physical: num  1 1 1 1 1 1 1 1 1 1 ...
##  $ exercise       : num  1 2 1 2 2 2 1 2 1 1 ...
##  $ married        : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ hypertension   : int  0 0 0 1 0 0 0 1 0 0 ...
##  $ gender         : num  1 2 2 2 2 1 2 1 2 1 ...
##  $ cost           : int  1746 602 576 5562 836 3842 1304 9724 201 4492 ...
##  $ target         : num  2 2 2 1 2 2 2 1 2 2 ...
##  $ age_cat        : num  3 3 3 3 3 2 2 1 3 1 ...
##  $ bmi_cat        : num  3 2 2 1 3 2 3 3 3 3 ...
##  - attr(*, "na.action")= 'omit' Named int [1:169] 20 32 93 118 167 231 281 309 320 339 ...
##   ..- attr(*, "names")= chr [1:169] "20" "32" "93" "118" ...
```

```r
lm_out <- lm(data = df, target ~ age+bmi+children+smoker+location_type+education_level+yearly_physical+
                                        married+hypertension+gender)
summary(lm_out)
```

```
##
## Call:
## lm(formula = target ~ age + bmi + children + smoker + location_type +
##     education_level + yearly_physical + exercise + married +
##     hypertension + gender, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.14887 -0.12756  0.05885  0.20577  0.94604
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.4957034  0.0387688  90.168  < 2e-16 ***
## age             -0.0073700  0.0002710 -27.201  < 2e-16 ***
## bmi             -0.0126240  0.0006395 -19.739  < 2e-16 ***
## children        -0.0116733  0.0031446  -3.712 0.000207 ***
## smoker          -0.5944971  0.0096637 -61.518  < 2e-16 ***
## location_type    0.0101584  0.0088178   1.152 0.249348
## education_level  0.0002466  0.0038564   0.064 0.949020
## yearly_physical -0.0228914  0.0088317  -2.592 0.009562 **
## exercise        -0.1704060  0.0088221 -19.316  < 2e-16 ***
## married         -0.0084090  0.0080957  -1.039 0.298982
## hypertension    -0.0332006  0.0095234  -3.486 0.000493 ***
## gender          -0.0132373  0.0076799  -1.724 0.084818 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3283 on 7401 degrees of freedom
## Multiple R-squared:  0.4264, Adjusted R-squared:  0.4255
## F-statistic: 500.1 on 11 and 7401 DF,  p-value: < 2.2e-16
```

```r
# Choosing only relevant columns
df <- data.frame(age = df$age, bmi = df$bmi, children=df$children, smoker=df$smoker, exercise = df$exer

trainList <- createDataPartition(y=df$target,p=.75,list=F)
# Putting 75% data in training and 25% in testing
training <- df[trainList,]
testing <- df[-trainList,]

str(df)
```

```
## 'data.frame':    7413 obs. of  8 variables:
##  $ age            : int  18 19 27 34 32 47 36 59 24 61 ...
##  $ bmi            : num  27.9 33.8 33 22.7 28.9 ...
##  $ children       : int  0 1 3 0 0 1 2 0 0 0 ...
##  $ smoker         : num  2 1 1 1 1 1 1 1 1 2 ...
##  $ exercise       : num  1 2 1 2 2 2 1 2 1 1 ...
##  $ hypertension   : int  0 0 0 1 0 0 0 1 0 0 ...
##  $ yearly_physical: num  1 1 1 1 1 1 1 1 1 1 ...
##  $ target         : Factor w/ 2 levels "1","2": 2 2 2 1 2 2 2 1 2 2 ...
```

```r
dim(training)
```

```
## [1] 5561    8
```

```r
dim(testing)
```

```
## [1] 1852    8
```

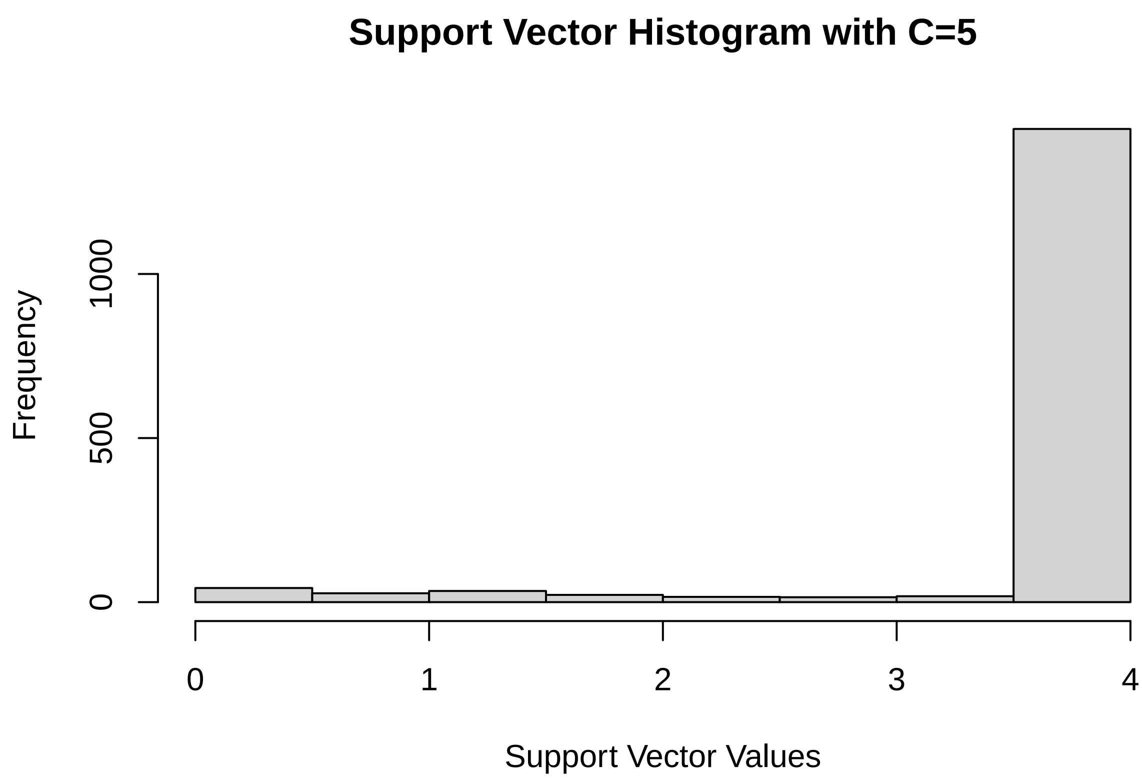
```r
head(df) #2 is not expensive and 1 is expensive
```

```
##   age    bmi children smoker exercise hypertension yearly_physical target
## 1  18 27.900        0      2        1            0               1      2
## 2  19 33.770        1      1        2            0               1      2
## 3  27 33.000        3      1        1            0               1      2
## 4  34 22.705        0      1        2            1               1      1
## 5  32 28.880        0      1        2            0               1      2
## 6  47 33.440        1      1        2            0               1      2
```

```r
csvm <- ksvm(target~age+bmi+children+smoker+exercise+hypertension+yearly_physical, data=training, type =
csvm
```

```
## Support Vector Machine object of class "ksvm"
##
## SV type: C-svc  (classification)
##  parameter : cost C = 4
##
## Gaussian Radial Basis kernel function.
##  Hyperparameter : sigma =  0.112320479111284
##
## Number of Support Vectors : 1617
##
## Objective Function Value : -5785.193
## Training error : 0.115986
## Cross validation error : 0.12354
```

```r
hist(alpha(csvm)[[1]], main="Support Vector Histogram with C=5", xlab="Support Vector Values")
```




```r
svmPred <-predict(csvm,testing)
confusionMatrix(svmPred,testing$target)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    2
##          1  275   46
##          2  188 1343
##
##                Accuracy : 0.8737
##                  95% CI : (0.8577, 0.8885)
##     No Information Rate : 0.75
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6247
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.5940
##             Specificity : 0.9669
##          Pos Pred Value : 0.8567
##          Neg Pred Value : 0.8772
##              Prevalence : 0.2500
##          Detection Rate : 0.1485
```

```
##    Detection Prevalence : 0.1733
##       Balanced Accuracy : 0.7804
##
##        'Positive' Class : 1
##
```