Project

2022-11-30

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
#Descriptive Analytics Part for IST 687 Final Project - sabdelra

#Library to call at the begining of the code
library (tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.3.6      ✓ purrr   0.3.4
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

```
library(RCurl)
```

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

```
library(jsonlite)
```

```
##
## Attaching package: 'jsonlite'
##
## The following object is masked from 'package:purrr':
##
##     flatten
```

```
library(imputeTS)
```

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

```
library(ggplot2)
library(ggmap)
```

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.
## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```
library (arules)
```

```
## Warning: package 'arules' was built under R version 4.2.2
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
##
## Attaching package: 'arules'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
data <- data.frame(read_csv('HMO_data.csv'))
```

```
## Rows: 7582 Columns: 14
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (8): smoker, location, location_type, education_level, yearly_physical, ...
## dbl (6): X, age, bmi, children, hypertension, cost
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

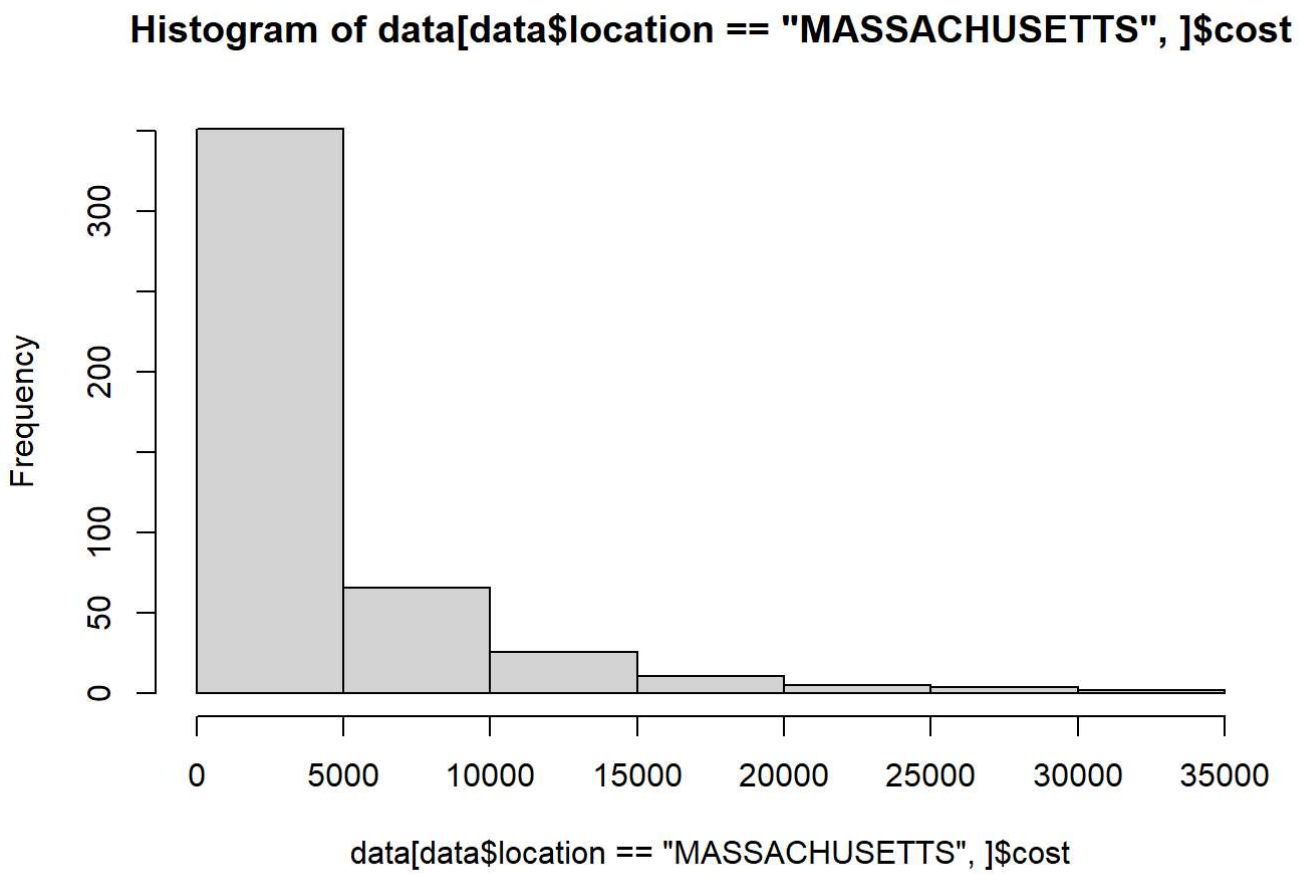
```
summary(data)
```

```
##        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
##
```

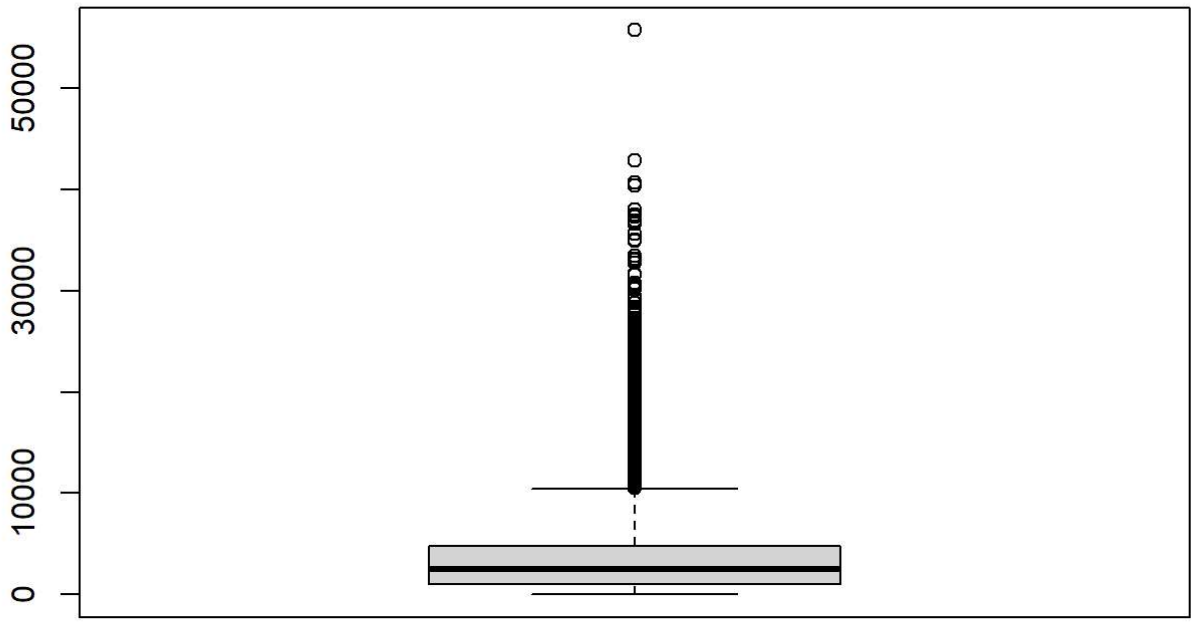
```
table(data$location)
```

```
##
##   CONNECTICUT      MARYLAND MASSACHUSETTS    NEW JERSEY      NEW YORK
##           611           747           465           498           547
##  PENNSYLVANIA  RHODE ISLAND
##          4010           704
```

```
hist(data[data$location == 'MASSACHUSETTS', ]$cost)
```




```
boxplot(data$cost)
```



```
#Below code is trying to estimate the threshold of being expensive --> 5000 was chosen based on professor suggestion
```

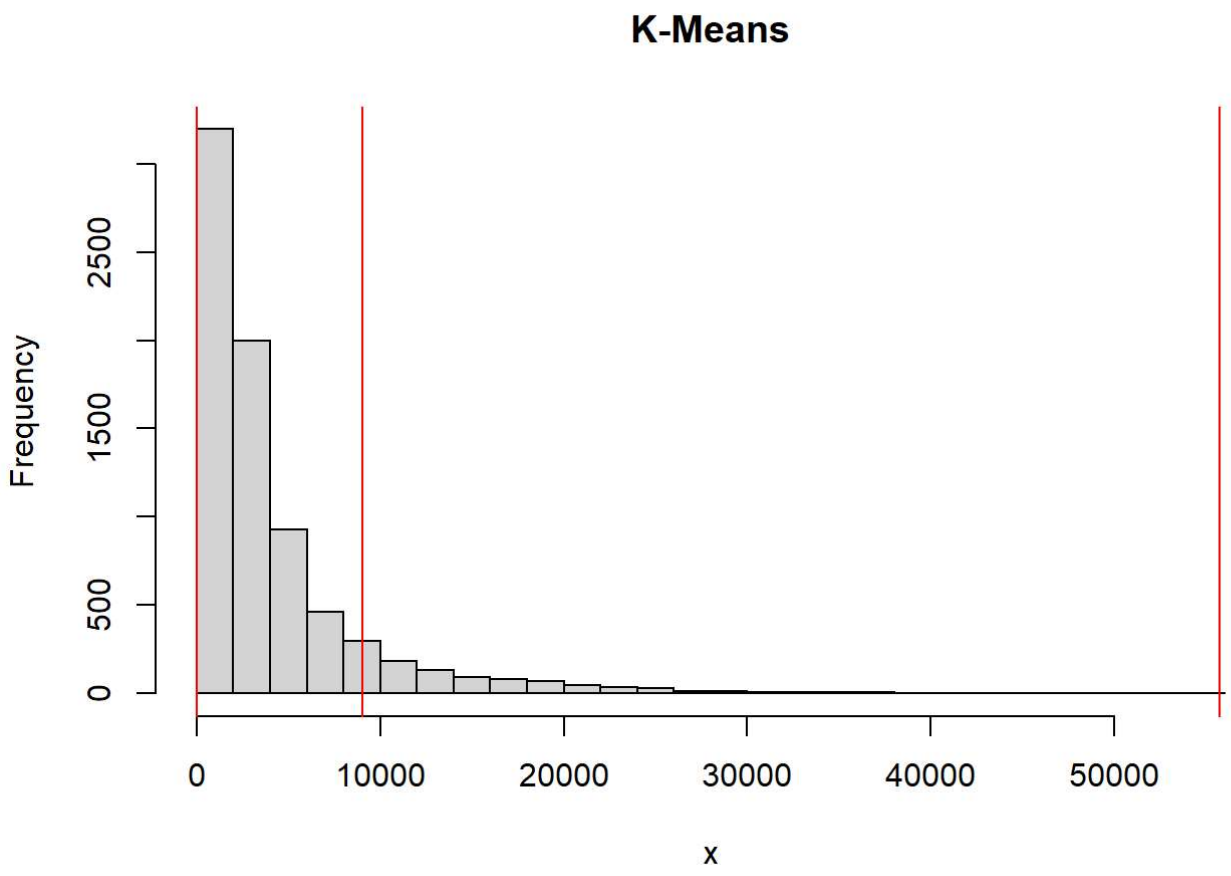
```
### k-means clustering
x= data$cost
table(discretize(x, "cluster", categories=2))
```

```
## Warning in discretize(x, "cluster", categories = 2): Parameter categories is
## deprecated. Use breaks instead! Also, the default method is now frequency!
```

```
##
##        [2,9.05e+03) [9.05e+03,5.57e+04]
##                6747                 835
```

```
hist(x, breaks=20, main="K-Means")
abline(v=discretize(x, method="cluster", categories=2, onlycuts=TRUE),
col="red")
```

```
## Warning in discretize(x, method = "cluster", categories = 2, onlycuts = TRUE):
## Parameter categories is deprecated. Use breaks instead! Also, the default method
## is now frequency!
```




```
discretize(x, method="cluster", categories=2, onlycuts=TRUE)
```

```
## Warning in discretize(x, method = "cluster", categories = 2, onlycuts = TRUE):
## Parameter categories is deprecated. Use breaks instead! Also, the default method
## is now frequency!
```

```
## [1]     2.000  9048.269 55715.000
```

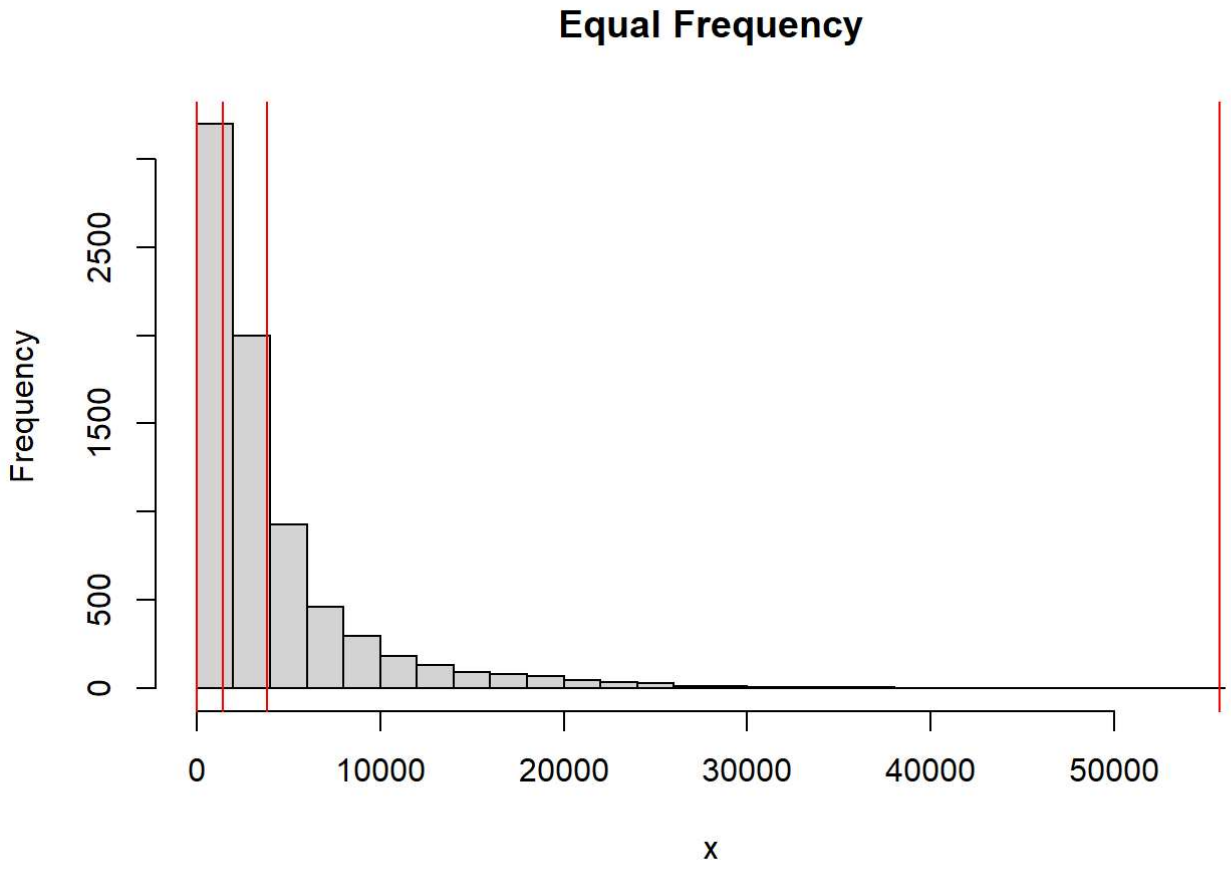
```
### equal frequency
table(discretize(x, "frequency", categories=3))
```

```
## Warning in discretize(x, "frequency", categories = 3): Parameter categories is
## deprecated. Use breaks instead! Also, the default method is now frequency!
```

```
##
##        [2,1.45e+03) [1.45e+03,3.82e+03) [3.82e+03,5.57e+04]
##                2527                2527                2528
```

```
hist(x, breaks=20, main="Equal Frequency")
abline(v=discretize(x, method="frequency", categories=3 , onlycuts=TRUE),
col="red")
```

```
## Warning in discretize(x, method = "frequency", categories = 3, onlycuts = TRUE):
## Parameter categories is deprecated. Use breaks instead! Also, the default method
## is now frequency!
```



```
discretize(x, method="frequency", categories=3 , onlycuts=TRUE)
```

```
## Warning in discretize(x, method = "frequency", categories = 3, onlycuts = TRUE):
## Parameter categories is deprecated. Use breaks instead! Also, the default method
## is now frequency!
```

```
## [1]     2  1449  3819 55715
```

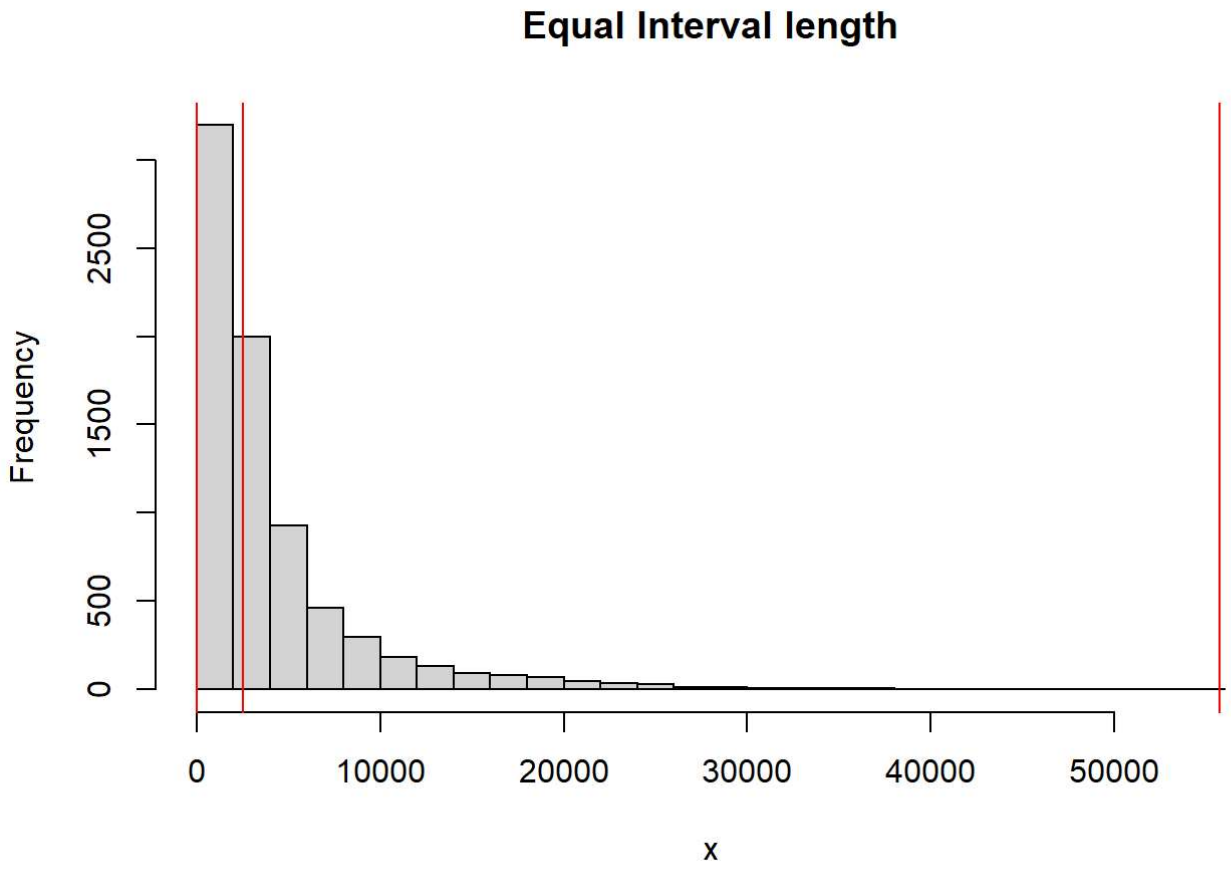
```
table(discretize(x, categories=2))
```

```
## Warning in discretize(x, categories = 2): Parameter categories is deprecated.
## Use breaks instead! Also, the default method is now frequency!
```

```
##
##        [2,2.5e+03) [2.5e+03,5.57e+04]
##               3789               3793
```

```
hist(x, breaks=20, main="Equal Interval length")
abline(v=discretize(x, categories=2, onlycuts=TRUE),
col="red")
```

```
## Warning in discretize(x, categories = 2, onlycuts = TRUE): Parameter categories
## is deprecated. Use breaks instead! Also, the default method is now frequency!
```



```
discretize(x, method="frequency", categories=2, onlycuts=TRUE)
```

```
## Warning in discretize(x, method = "frequency", categories = 2, onlycuts = TRUE):
## Parameter categories is deprecated. Use breaks instead! Also, the default method
## is now frequency!
```

```
## [1]     2  2500 55715
```

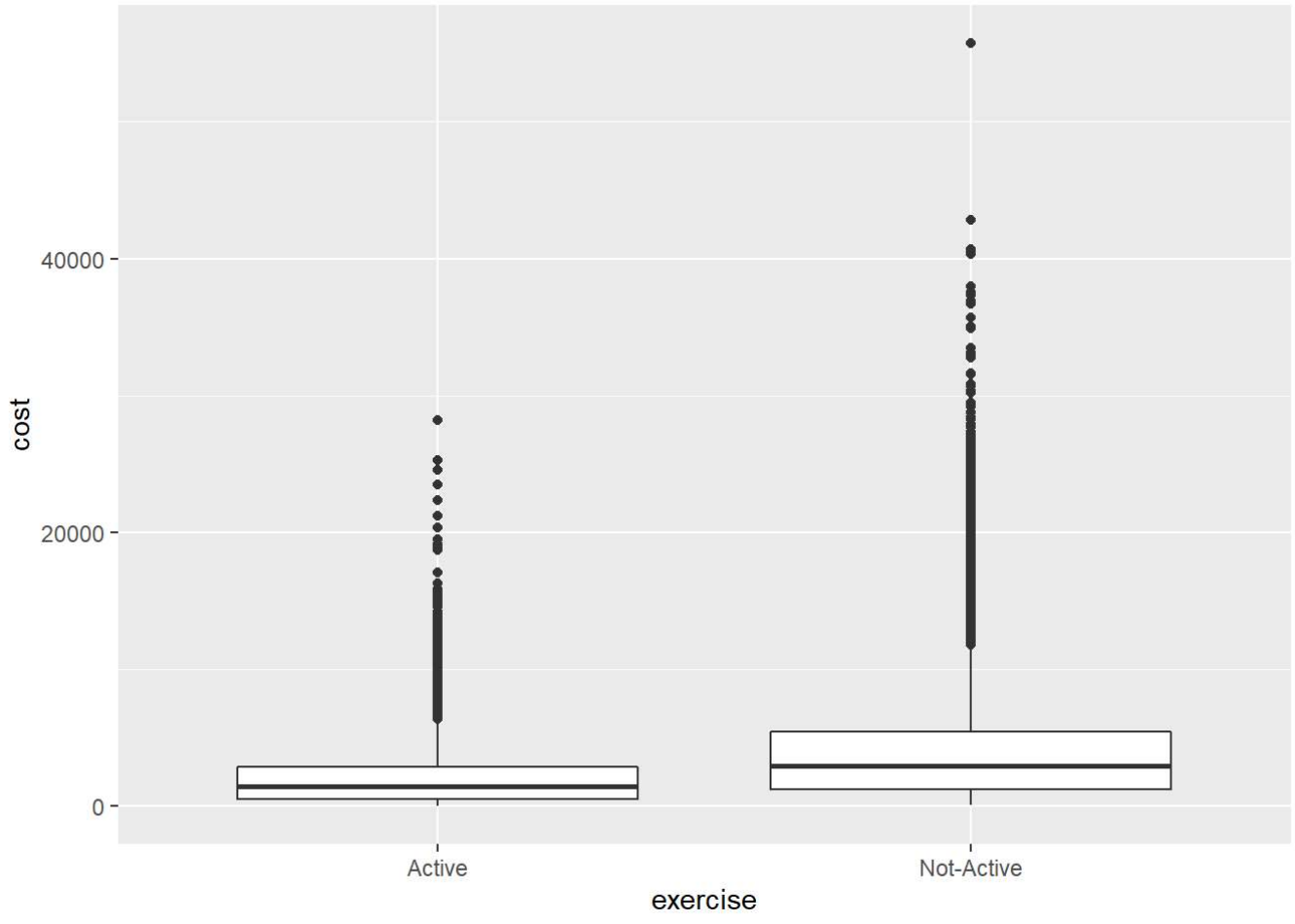
```
#Create the expensive table
#data <- transform(
#  data, expensive= ifelse(cost > 3819, 1,0))


data <- transform(
  data, expensive= ifelse(cost > 5000, 1,0))
```

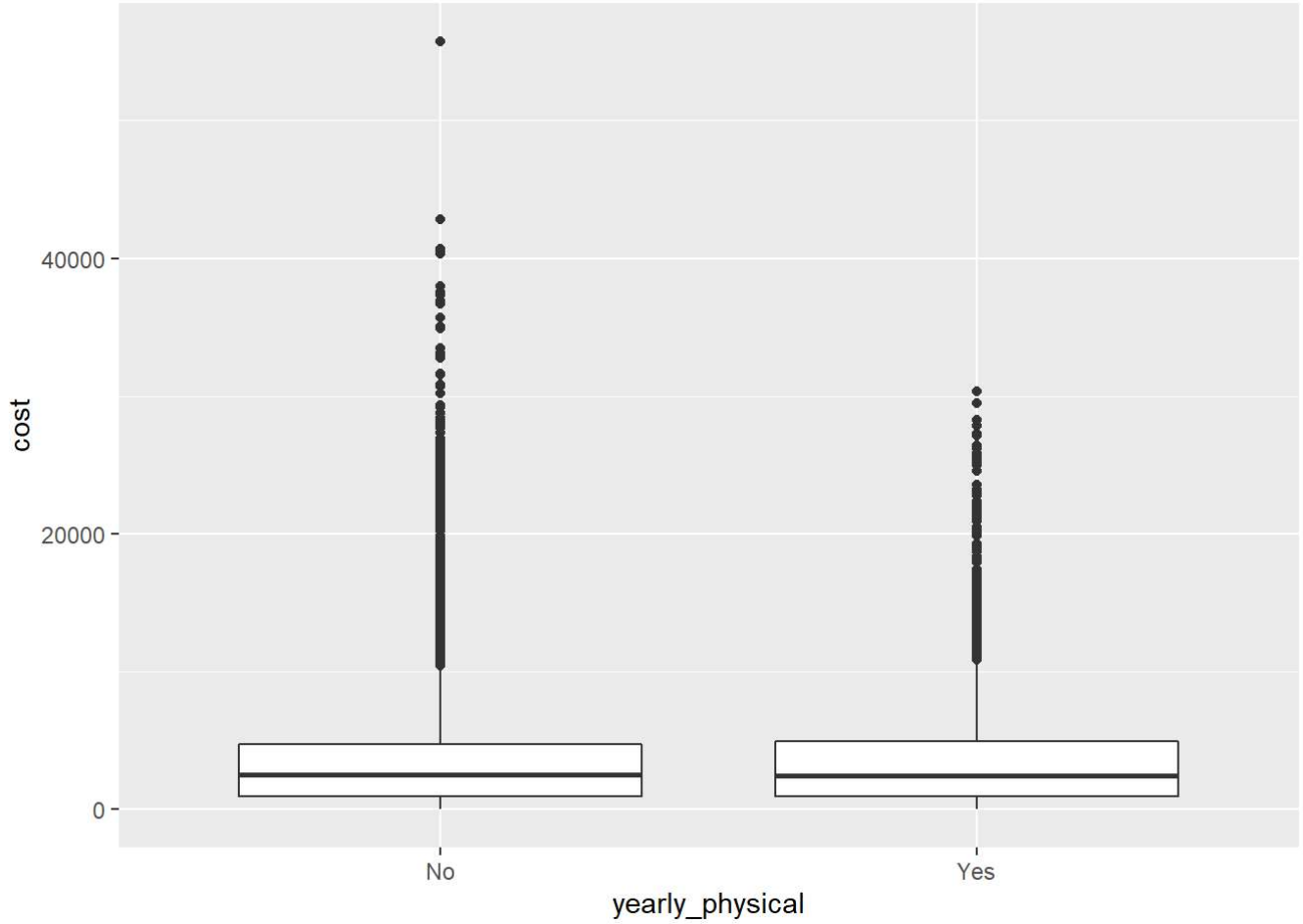
```
table(data$expensive)
```

```
##
##    0    1
## 5780 1802
```

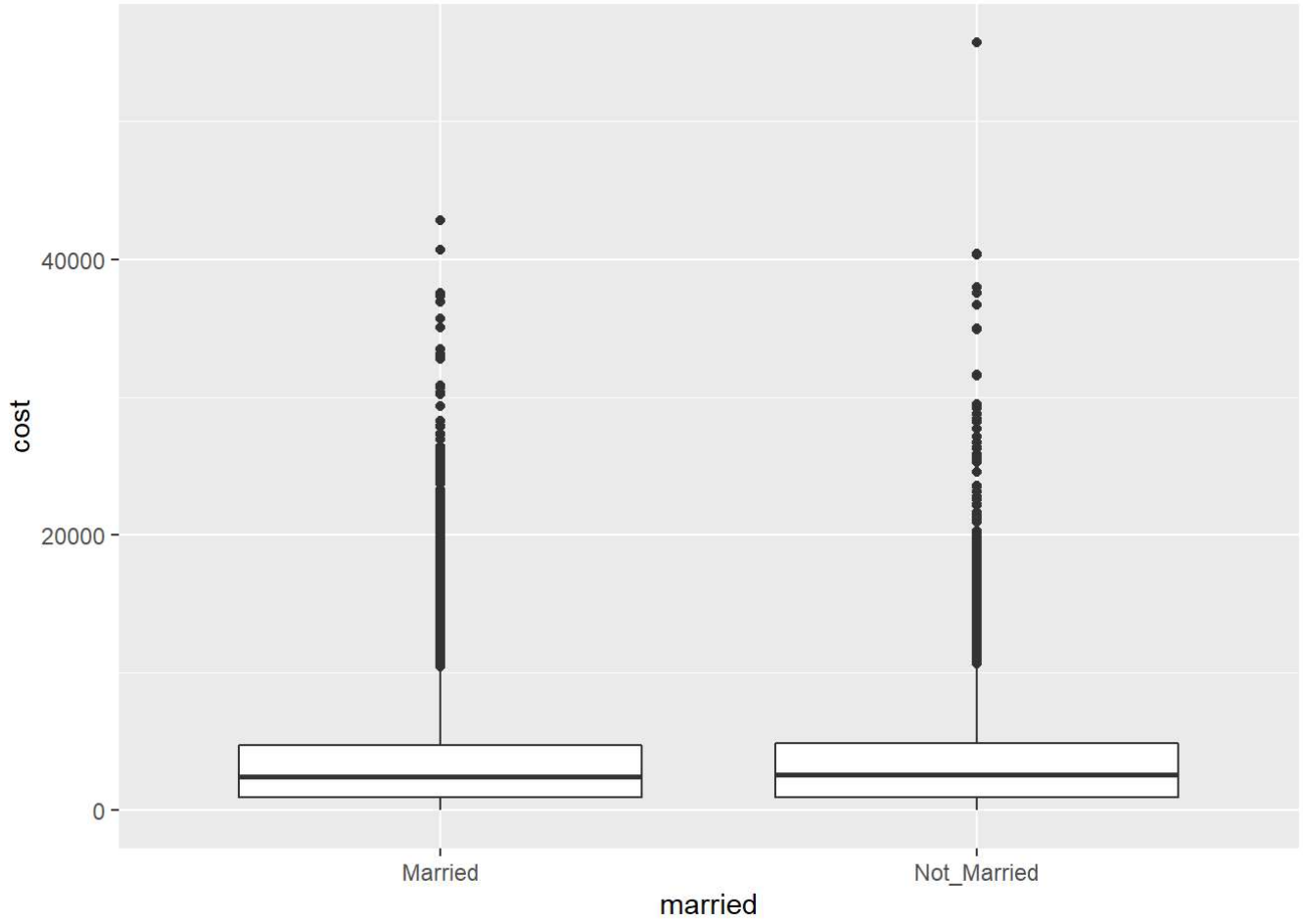
```
ggplot(data, aes(x=exercise , y=cost, fill= expensive)) +
    geom_boxplot()
```



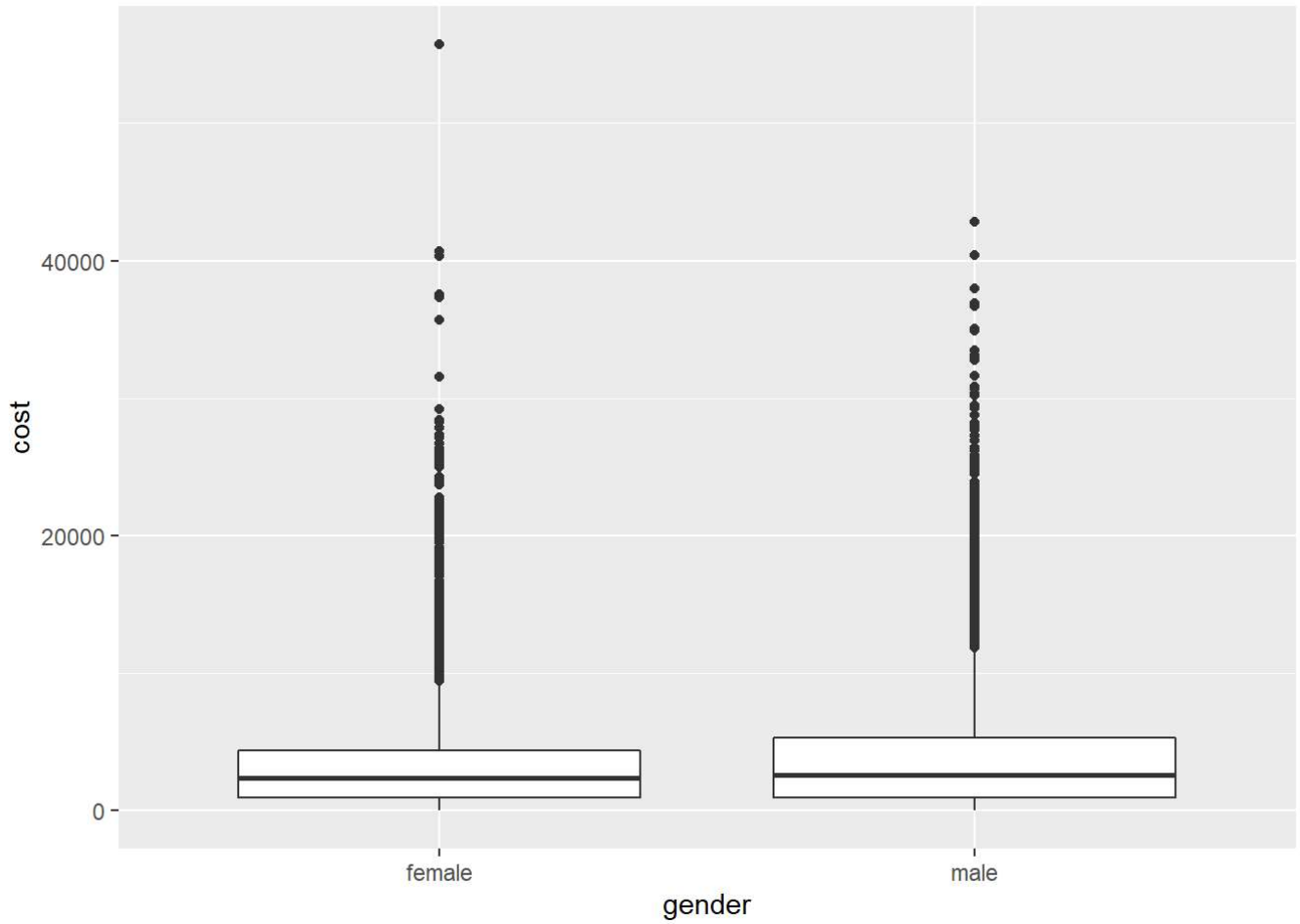
```
ggplot(data, aes(x=yearly_physical , y=cost, fill= expensive)) +
    geom_boxplot()
```



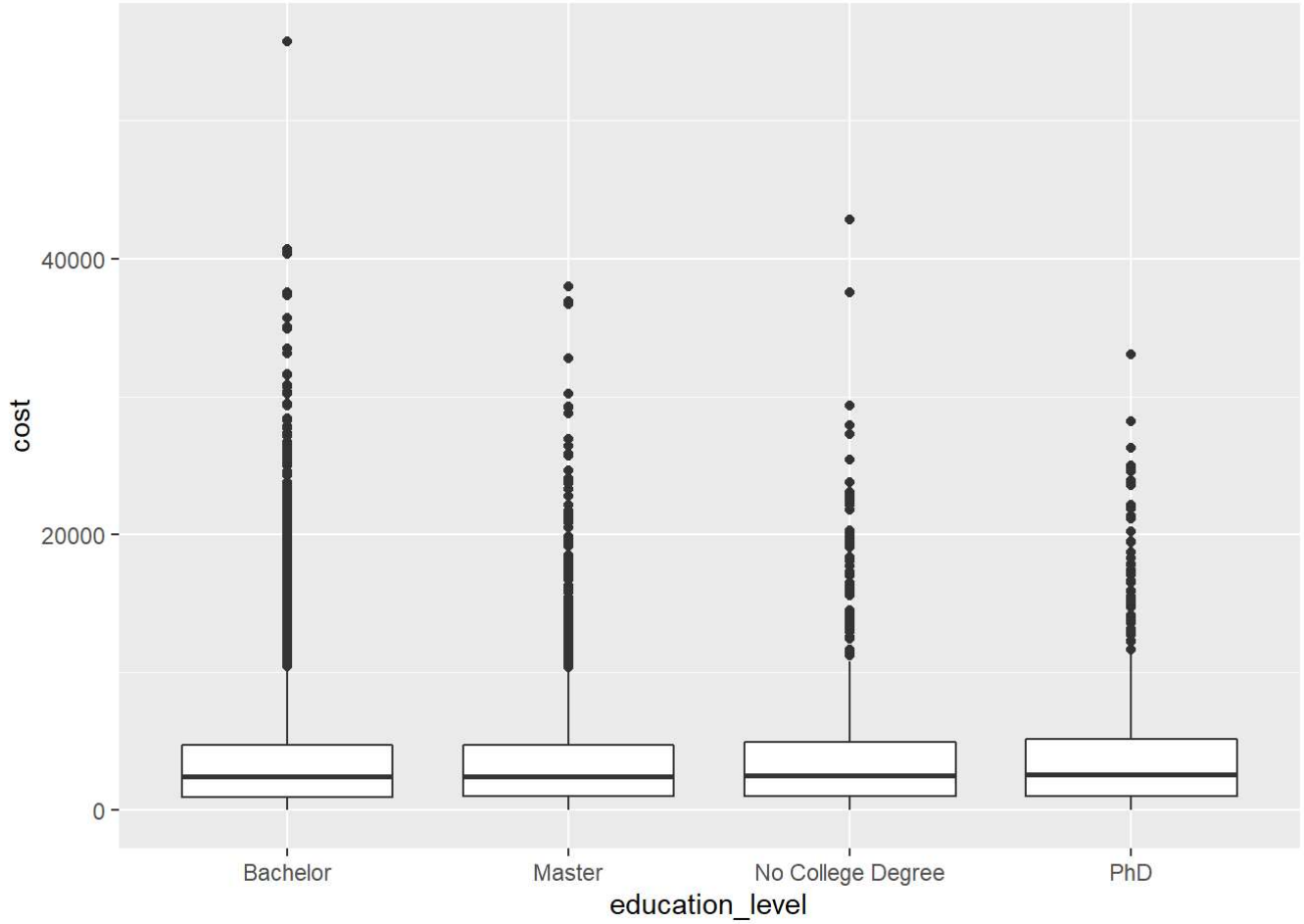

```
ggplot(data, aes(x=married , y=cost, fill= expensive)) +
    geom_boxplot()
```



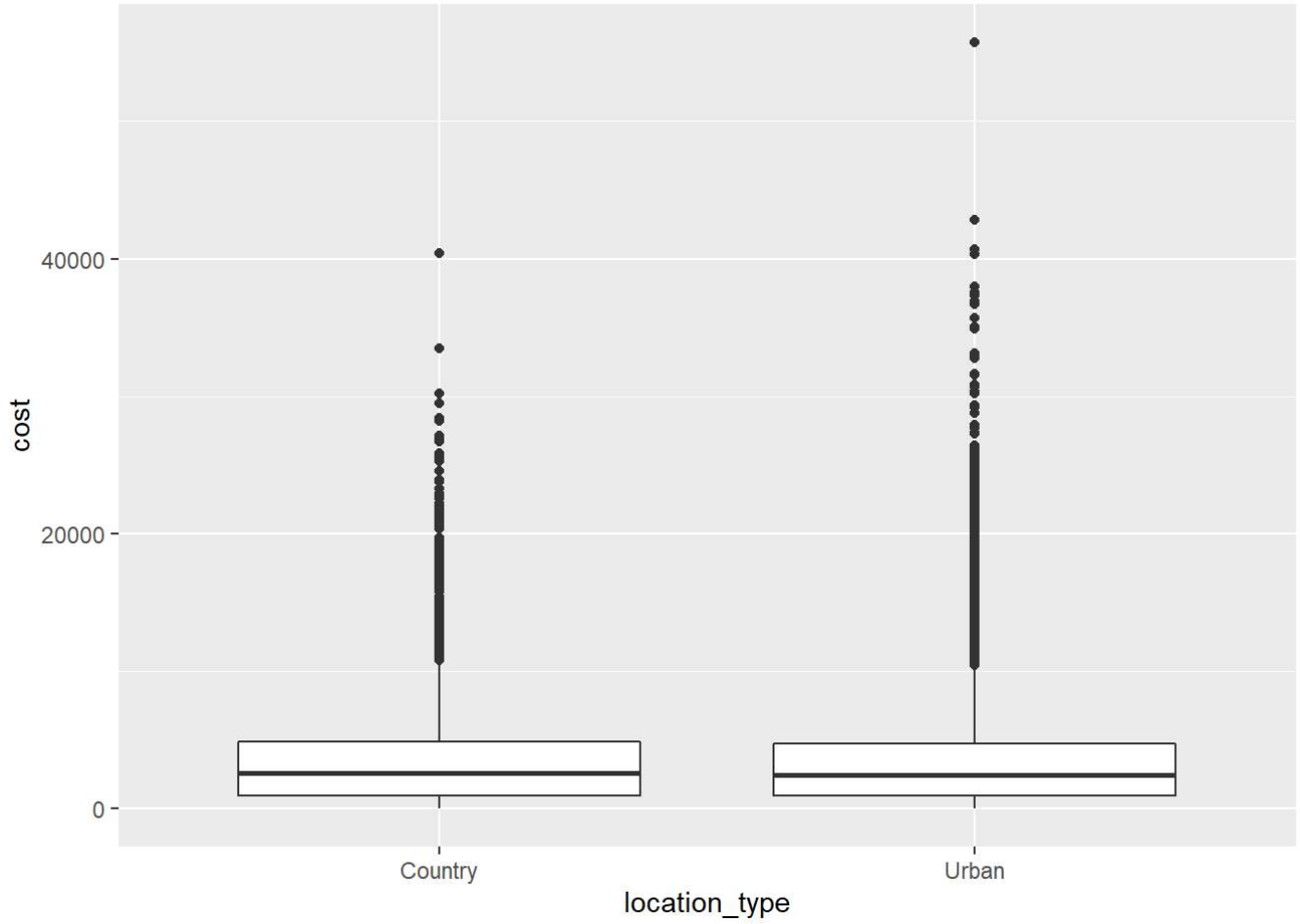

```
ggplot(data, aes(x=gender , y=cost, fill= expensive)) +
    geom_boxplot()
```



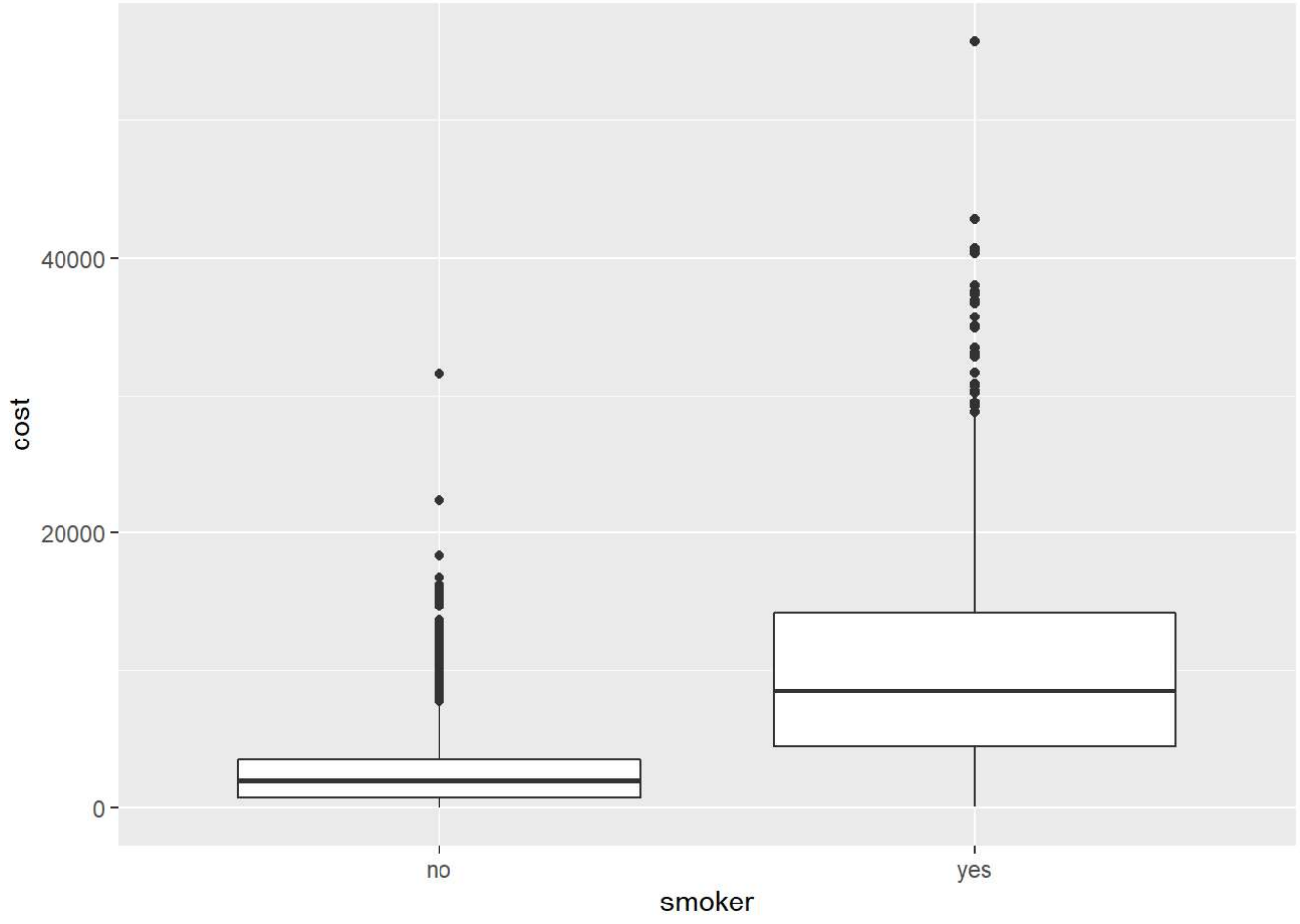
```
ggplot(data, aes(x= education_level , y=cost, fill= expensive)) +
    geom_boxplot()
```



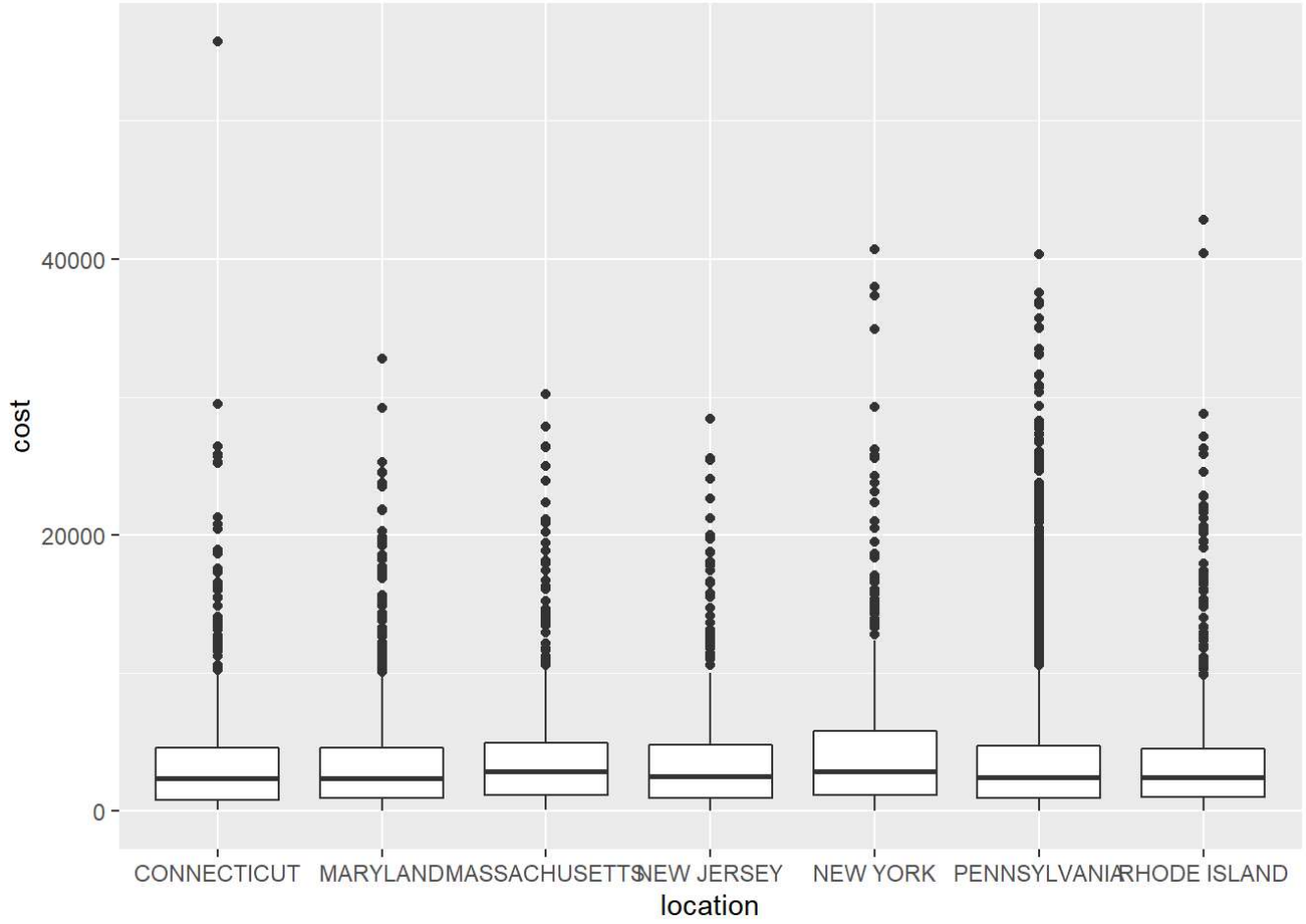

```
ggplot(data, aes(x= location_type , y=cost, fill= expensive)) +
    geom_boxplot()
```




```
ggplot(data, aes(x= smoker , y=cost, fill= expensive)) +
    geom_boxplot()
```



```
ggplot(data, aes(x= location , y=cost, fill= expensive)) +
    geom_boxplot()
```




```
numerical_data <- data[,c('age','bmi', 'cost')]
summary(numerical_data)
```

```
##       age             bmi             cost
##  Min.   :18.00   Min.   :15.96   Min.   :    2
##  1st Qu.:26.00   1st Qu.:26.60   1st Qu.:  970
##  Median :39.00   Median :30.50   Median : 2500
##  Mean   :38.89   Mean   :30.80   Mean   : 4043
##  3rd Qu.:51.00   3rd Qu.:34.77   3rd Qu.: 4775
##  Max.   :66.00   Max.   :53.13   Max.   :55715
##                  NA's   :78
```

```
res <- cor(numerical_data)
round(res, 2)
```

```
##       age bmi cost
## age  1.00  NA 0.32
## bmi    NA   1   NA
## cost 0.32  NA 1.00
```

```
 data[is.na(data$bmi) , ]
```

```
##            X age bmi children smoker      location location_type
## 20        23  19  NA        0     no  PENNSYLVANIA         Urban
## 32        39  35  NA        1    yes  RHODE ISLAND         Urban
## 93       123  19  NA        0     no      MARYLAND         Urban
## 231      312  19  NA        0     no  PENNSYLVANIA         Urban
## 281      387  59  NA        0     no  PENNSYLVANIA       Country
## 320      440  26  NA        0     no      MARYLAND       Country
## 514      682  19  NA        0     no  PENNSYLVANIA         Urban
## 548      724  19  NA        0     no      MARYLAND         Urban
## 768     1015  37  NA        0     no  PENNSYLVANIA         Urban
## 828     1092  57  NA        0     no  RHODE ISLAND       Country
## 1014    8311  62  NA        0     no  PENNSYLVANIA       Country
## 1023   11111  54  NA        1     no   CONNECTICUT         Urban
## 1057   13021  62  NA        3    yes    NEW JERSEY       Country
## 1099    2281  57  NA        0     no  RHODE ISLAND         Urban
## 1292    7751  42  NA        2     no   CONNECTICUT       Country
## 1340    5010  35  NA        1    yes  PENNSYLVANIA         Urban
## 1417    3541  34  NA        0     no  PENNSYLVANIA       Country
## 1469    8231  19  NA        0     no  PENNSYLVANIA       Country
## 1532    9642  45  NA        3     no  PENNSYLVANIA         Urban
## 1625  110411  59  NA        0     no  RHODE ISLAND       Country
## 1767    6262  28  NA        0     no  PENNSYLVANIA         Urban
## 1976    8100  36  NA        3     no  PENNSYLVANIA         Urban
## 2050   12712  25  NA        1     no  RHODE ISLAND         Urban
## 2106   10622  58  NA        1     no  PENNSYLVANIA       Country
## 2251   22621  55  NA        3     no  PENNSYLVANIA         Urban
## 2431  935111  34  NA        2     no  PENNSYLVANIA         Urban
## 2523   11342  53  NA        0     no  PENNSYLVANIA         Urban
## 2708    2122  40  NA        4     no  PENNSYLVANIA       Country
## 2762    7401  29  NA        2    yes  PENNSYLVANIA         Urban
## 2785   13362  18  NA        0     no      NEW YORK         Urban
## 2850   92511  44  NA        0     no      MARYLAND         Urban
## 2883   98111  53  NA        1     no   CONNECTICUT         Urban
## 3370    4113  25  NA        0     no    NEW JERSEY       Country
## 3431    9092  63  NA        3     no  PENNSYLVANIA         Urban
## 3535  103121  46  NA        1    yes    NEW JERSEY         Urban
## 3559    1763  63  NA        0    yes  PENNSYLVANIA         Urban
## 3648    6184  49  NA        2    yes  PENNSYLVANIA         Urban
## 3921 5721111  18  NA        1     no  PENNSYLVANIA         Urban
## 3954   66411  18  NA        0     no  PENNSYLVANIA         Urban
## 4004    4383  36  NA        3     no  PENNSYLVANIA         Urban
## 4086    8851  25  NA        4     no  PENNSYLVANIA         Urban
## 4153  104621  43  NA        2    yes  PENNSYLVANIA         Urban
## 4464   53821  47  NA        2     no      NEW YORK       Country
## 4467  626111  29  NA        0     no  PENNSYLVANIA       Country
## 4545    8082  19  NA        0     no      MARYLAND         Urban
## 4704  778112  44  NA        0     no  PENNSYLVANIA         Urban
## 4738  113313  58  NA        0     no      MARYLAND       Country
## 4740    6333  29  NA        0     no  PENNSYLVANIA         Urban
## 4816    1814  23  NA        0     no  PENNSYLVANIA         Urban
## 4828  122522  42  NA        1     no      MARYLAND       Country
## 4923  773111  45  NA        0     no   CONNECTICUT       Country
## 5009  100812  45  NA        3    yes      MARYLAND       Country
## 5068   33103  19  NA        5     no  PENNSYLVANIA         Urban
## 5168   34631  32  NA        3     no   CONNECTICUT       Country
```

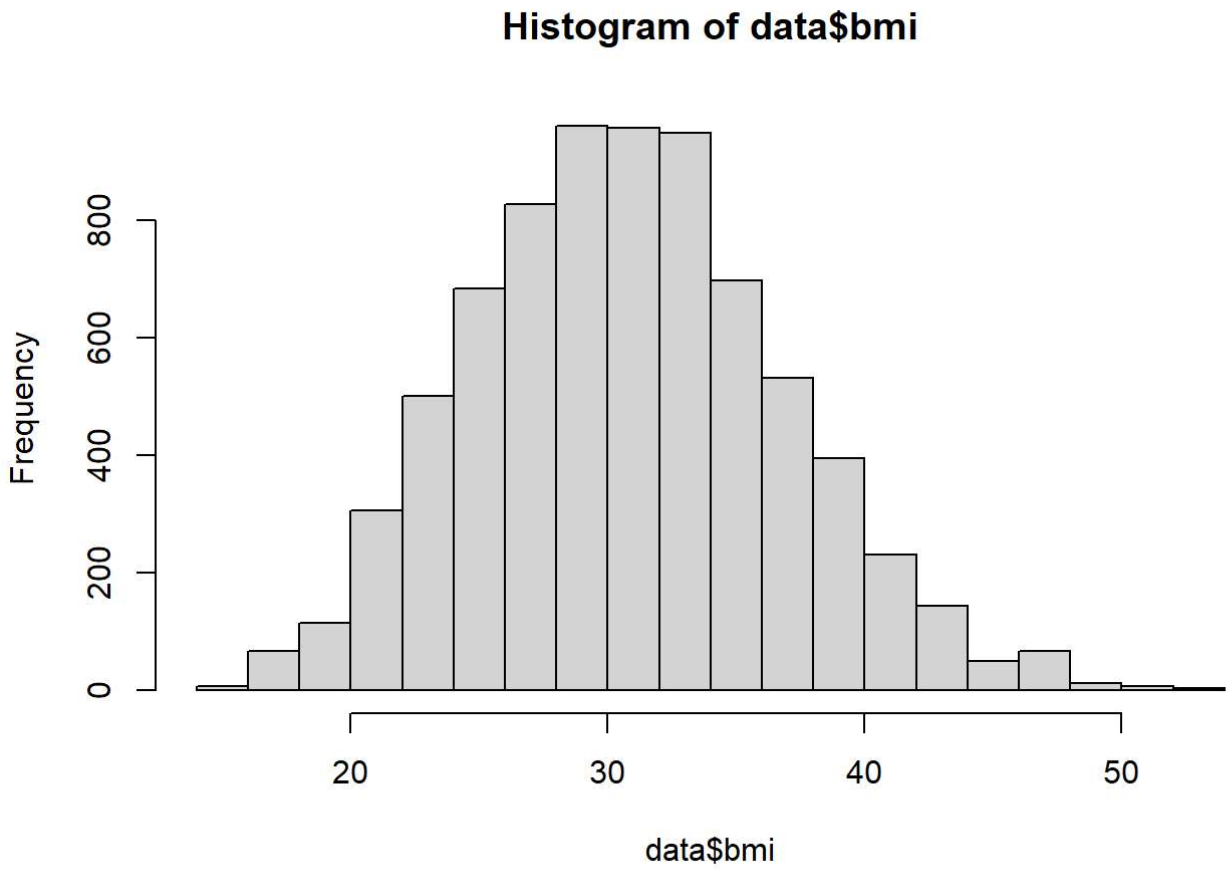
```
## 5353  331211  19  NA        5     no MASSACHUSETTS         Urban
## 5354  800111  32  NA        0    yes  PENNSYLVANIA         Urban
## 5486  194112  55  NA        1     no MASSACHUSETTS       Country
## 5525  880311  37  NA        2     no      NEW YORK         Urban
## 5616   12612  33  NA        0     no   CONNECTICUT         Urban
## 5637  808111  18  NA        0     no  PENNSYLVANIA         Urban
## 5811    2202  25  NA        0     no  PENNSYLVANIA       Country
## 5824  363221  20  NA        0    yes      MARYLAND         Urban
## 6022   21711  53  NA        0     no    NEW JERSEY         Urban
## 6025   43141  19  NA        0     no      MARYLAND         Urban
## 6234   32831  44  NA        2    yes  PENNSYLVANIA         Urban
## 6386   96432  45  NA        3     no  RHODE ISLAND         Urban
## 6452   54213  21  NA        2     no  PENNSYLVANIA         Urban
## 6681 1441111  28  NA        2     no   CONNECTICUT         Urban
## 6696  100313  25  NA        0     no  PENNSYLVANIA       Country
## 6763 1286211  48  NA        0     no  PENNSYLVANIA         Urban
## 6940    3834  54  NA        0     no  PENNSYLVANIA         Urban
## 7016    1843  44  NA        0     no  PENNSYLVANIA         Urban
## 7145  804121  18  NA        0    yes  RHODE ISLAND       Country
## 7175   11952  31  NA        0     no  PENNSYLVANIA         Urban
## 7251  984112  29  NA        1     no  PENNSYLVANIA         Urban
## 7346   19511  18  NA        0     no  PENNSYLVANIA         Urban
## 7372  520111  32  NA        0     no  PENNSYLVANIA         Urban
## 7385 3072111  27  NA        2     no  PENNSYLVANIA         Urban
##        education_level yearly_physical   exercise     married hypertension
## 20   No College Degree              No     Active Not_Married            0
## 32             Bachelor             Yes Not-Active     Married            0
## 93                  PhD             Yes Not-Active Not_Married            0
## 231              Master              No Not-Active     Married            0
## 281              Master              No     Active Not_Married            0
## 320            Bachelor              No     Active     Married            0
## 514              Master              No Not-Active     Married            0
## 548              Master              No Not-Active     Married            1
## 768            Bachelor              No Not-Active     Married            0
## 828            Bachelor             Yes Not-Active     Married            0
## 1014           Bachelor              No Not-Active     Married            0
## 1023 No College Degree               No Not-Active Not_Married            0
## 1057             Master             Yes     Active     Married            0
## 1099           Bachelor              No Not-Active     Married            0
## 1292           Bachelor             Yes Not-Active Not_Married            1
## 1340             Master              No Not-Active Not_Married            0
## 1417                PhD              No     Active     Married            0
## 1469           Bachelor              No Not-Active Not_Married            0
## 1532           Bachelor              No Not-Active     Married            0
## 1625           Bachelor              No     Active     Married            0
## 1767           Bachelor             Yes Not-Active     Married            1
## 1976           Bachelor              No     Active     Married            0
## 2050           Bachelor              No Not-Active     Married            0
## 2106           Bachelor             Yes Not-Active     Married            0
## 2251           Bachelor              No Not-Active     Married            0
## 2431 No College Degree               No     Active Not_Married            0
## 2523           Bachelor              No Not-Active Not_Married            0
## 2708           Bachelor              No Not-Active     Married            0
## 2762           Bachelor              No     Active Not_Married            1
## 2785                PhD              No Not-Active     Married            0
## 2850           Bachelor              No Not-Active     Married            0
```

```
## 2883          Bachelor              No     Active     Married            0
## 3370          Bachelor             Yes     Active Not_Married            0
## 3431            Master              No     Active Not_Married            0
## 3535          Bachelor             Yes Not-Active     Married            0
## 3559               PhD              No     Active     Married            0
## 3648          Bachelor             Yes Not-Active     Married            0
## 3921 No College Degree              No     Active     Married            0
## 3954          Bachelor              No     Active     Married            0
## 4004          Bachelor              No Not-Active Not_Married            1
## 4086          Bachelor              No Not-Active Not_Married            1
## 4153          Bachelor              No Not-Active     Married            1
## 4464               PhD             Yes Not-Active Not_Married            0
## 4467          Bachelor             Yes Not-Active     Married            0
## 4545          Bachelor             Yes Not-Active Not_Married            0
## 4704          Bachelor              No Not-Active     Married            0
## 4738               PhD              No Not-Active Not_Married            1
## 4740               PhD              No     Active     Married            1
## 4816            Master              No Not-Active     Married            0
## 4828            Master              No     Active Not_Married            0
## 4923          Bachelor              No     Active Not_Married            0
## 5009          Bachelor              No Not-Active Not_Married            0
## 5068               PhD             Yes Not-Active     Married            0
## 5168            Master             Yes Not-Active     Married            0
## 5353          Bachelor             Yes Not-Active     Married            0
## 5354            Master              No Not-Active     Married            0
## 5486          Bachelor              No Not-Active Not_Married            0
## 5525          Bachelor             Yes Not-Active     Married            0
## 5616 No College Degree              No Not-Active     Married            0
## 5637            Master              No     Active     Married            0
## 5811          Bachelor              No Not-Active     Married            0
## 5824               PhD             Yes     Active     Married            0
## 6022          Bachelor              No Not-Active Not_Married            0
## 6025            Master              No Not-Active     Married            0
## 6234 No College Degree              No Not-Active     Married            1
## 6386            Master              No     Active     Married            0
## 6452          Bachelor              No     Active     Married            0
## 6681 No College Degree             Yes Not-Active     Married            1
## 6696          Bachelor              No Not-Active Not_Married            0
## 6763            Master              No     Active     Married            1
## 6940 No College Degree             Yes Not-Active Not_Married            0
## 7016          Bachelor             Yes Not-Active     Married            0
## 7145 No College Degree              No Not-Active Not_Married            0
## 7175 No College Degree              No Not-Active     Married            0
## 7251            Master             Yes Not-Active     Married            1
## 7346          Bachelor              No Not-Active     Married            1
## 7372            Master             Yes Not-Active     Married            0
## 7385          Bachelor             Yes Not-Active Not_Married            0
##      gender  cost expensive
## 20     male   146         0
## 32     male 16448         1
## 93   female   605         0
## 231  female   194         0
## 281  female  2389         0
## 320    male   556         0
## 514    male   169         0
## 548    male   322         0
```

```
## 768  female  1855         0
## 828  female  4503         0
## 1014   male  4993         0
## 1023 female  2987         0
## 1057   male 11766         1
## 1099 female 12459         1
## 1292   male  6295         1
## 1340   male 17026         1
## 1417   male  2830         0
## 1469 female   515         0
## 1532   male  2185         0
## 1625   male  1021         0
## 1767 female  1172         0
## 1976 female  1181         0
## 2050   male  1182         0
## 2106   male  3206         0
## 2251   male  2681         0
## 2431   male    64         0
## 2523 female  1948         0
## 2708   male  2520         0
## 2762   male  8801         1
## 2785 female   341         0
## 2850   male  3046         0
## 2883   male  2515         0
## 3370 female   286         0
## 3431   male  5146         1
## 3535 female  8620         1
## 3559 female 14701         1
## 3648   male  5306         1
## 3921 female   450         0
## 3954   male   208         0
## 4004   male  2230         0
## 4086   male   934         0
## 4153 female  7904         1
## 4464 female  3119         0
## 4467 female   404         0
## 4545 female   472         0
## 4704   male  2147         0
## 4738   male 11074         1
## 4740 female   639         0
## 4816   male   380         0
## 4828   male  1062         0
## 4923 female  1077         0
## 5009   male 16879         1
## 5068 female   823         0
## 5168 female  2704         0
## 5353 female  1286         0
## 5354   male  5696         1
## 5486 female  1725         0
## 5525 female  3312         0
## 5616 female   976         0
## 5637 female   637         0
## 5811 female  5018         1
## 5824 female  1071         0
## 6022 female  3820         0
## 6025   male  9257         1
```

```
## 6234   male 19752         1
## 6386   male  2013         0
## 6452 female   437         0
## 6681   male  4504         0
## 6696   male   318         0
## 6763 female   373         0
## 6940   male  6367         1
## 7016 female  2568         0
## 7145 female 12927         1
## 7175 female  1089         0
## 7251 female  3513         0
## 7346   male   274         0
## 7372   male   877         0
## 7385 female  3913         0
```

```
hist(data$bmi)
```




```
data <- data %>% mutate(across(bmi, ~replace_na(., mean(., na.rm=TRUE))))
```

```
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

```
data_mode <- getmode
data <- data %>%
  mutate(hypertension = if_else(is.na(hypertension),
                         getmode(hypertension),
                         hypertension))
```

```
data[is.na(data$bmi) , ]
```

```
##  [1] X               age             bmi             children
##  [5] smoker          location        location_type   education_level
##  [9] yearly_physical exercise        married         hypertension
## [13] gender          cost            expensive
## <0 rows> (or 0-length row.names)
```

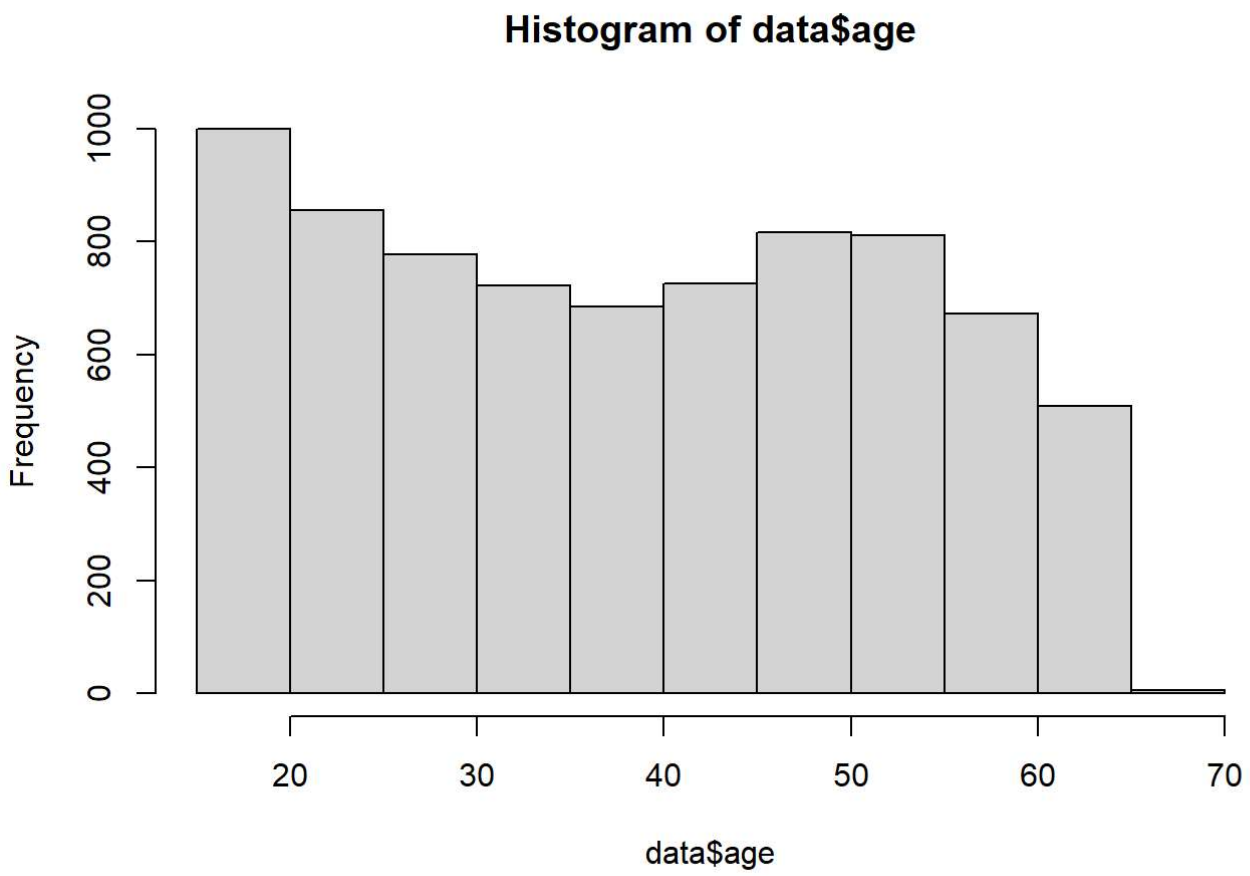
```
numerical_data <- data[,c('age','bmi', 'cost')]
summary(numerical_data)
```

```
##       age             bmi             cost
##  Min.   :18.00   Min.   :15.96   Min.   :    2
##  1st Qu.:26.00   1st Qu.:26.60   1st Qu.:  970
##  Median :39.00   Median :30.50   Median : 2500
##  Mean   :38.89   Mean   :30.80   Mean   : 4043
##  3rd Qu.:51.00   3rd Qu.:34.60   3rd Qu.: 4775
##  Max.   :66.00   Max.   :53.13   Max.   :55715
```

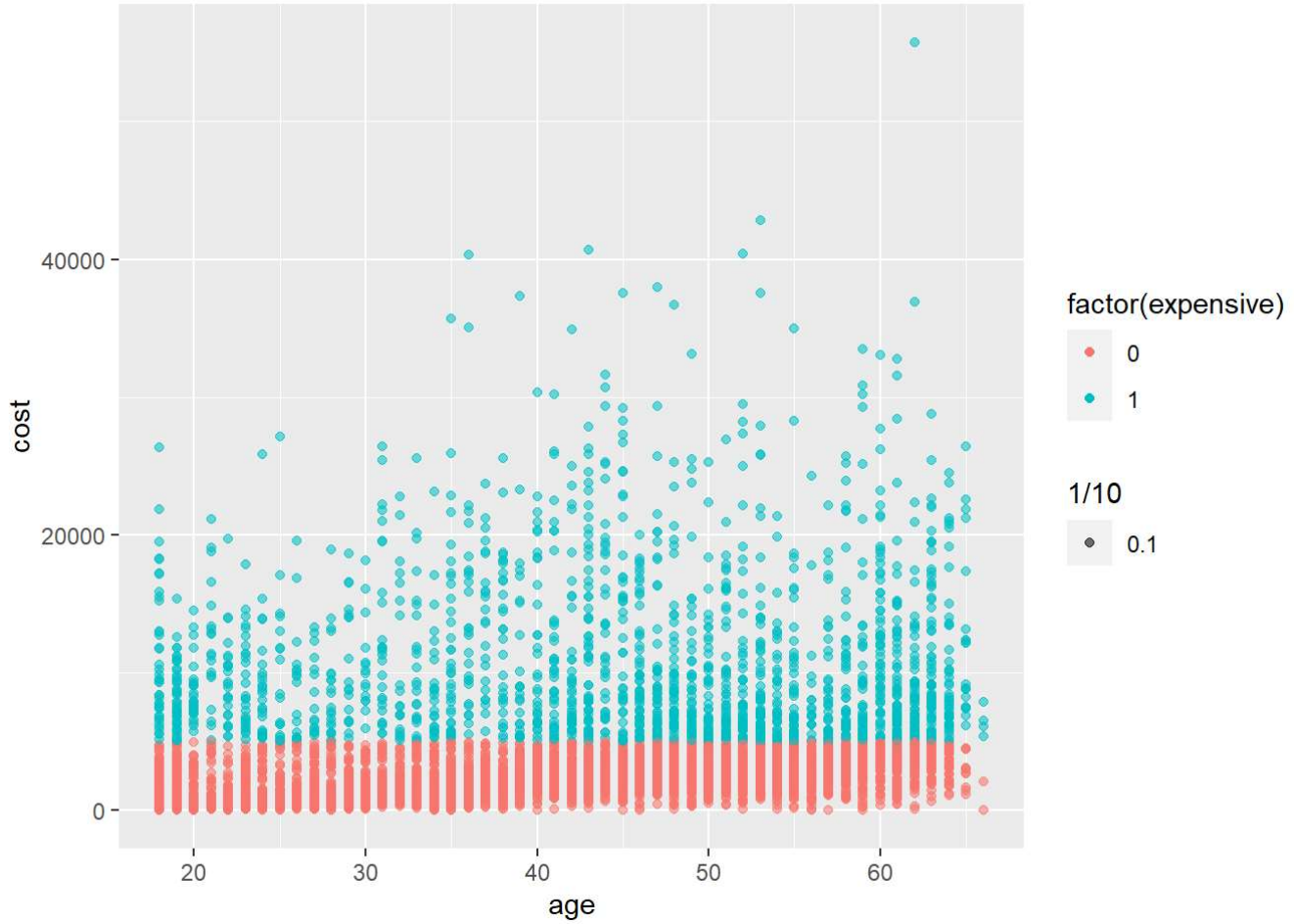
```
res <- cor(numerical_data)
round(res, 2)
```

```
##       age  bmi cost
## age  1.00 0.09 0.32
## bmi  0.09 1.00 0.25
## cost 0.32 0.25 1.00
```

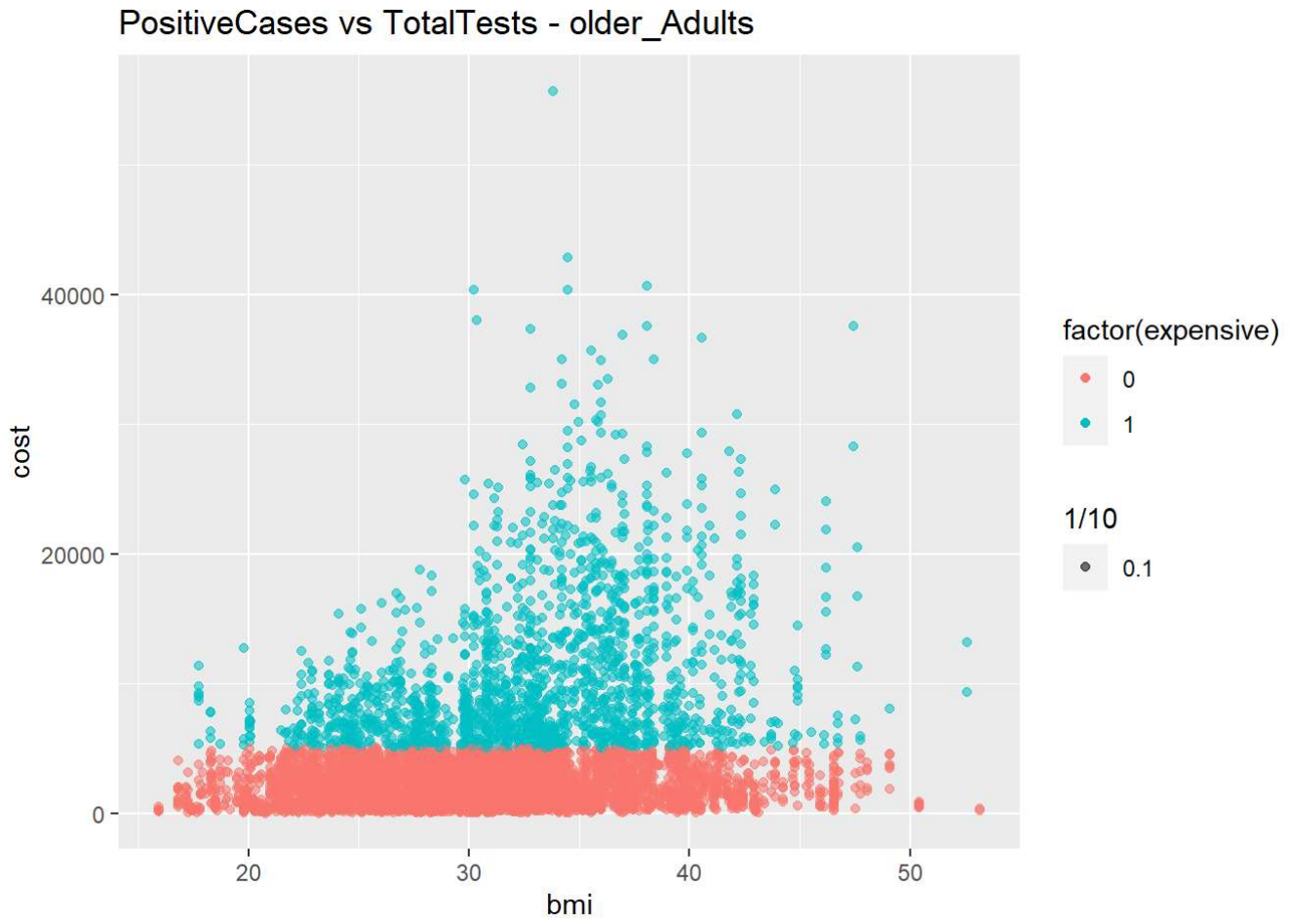
```
hist(data$age)
```




```
#Scattor plot usinf GGPLOT2
my_plot <- ggplot(data)
my_plot <- my_plot + geom_point(aes(x=age, y= cost, color = factor(expensive),alpha = 1/10 ))
my_plot <- my_plot
my_plot
```

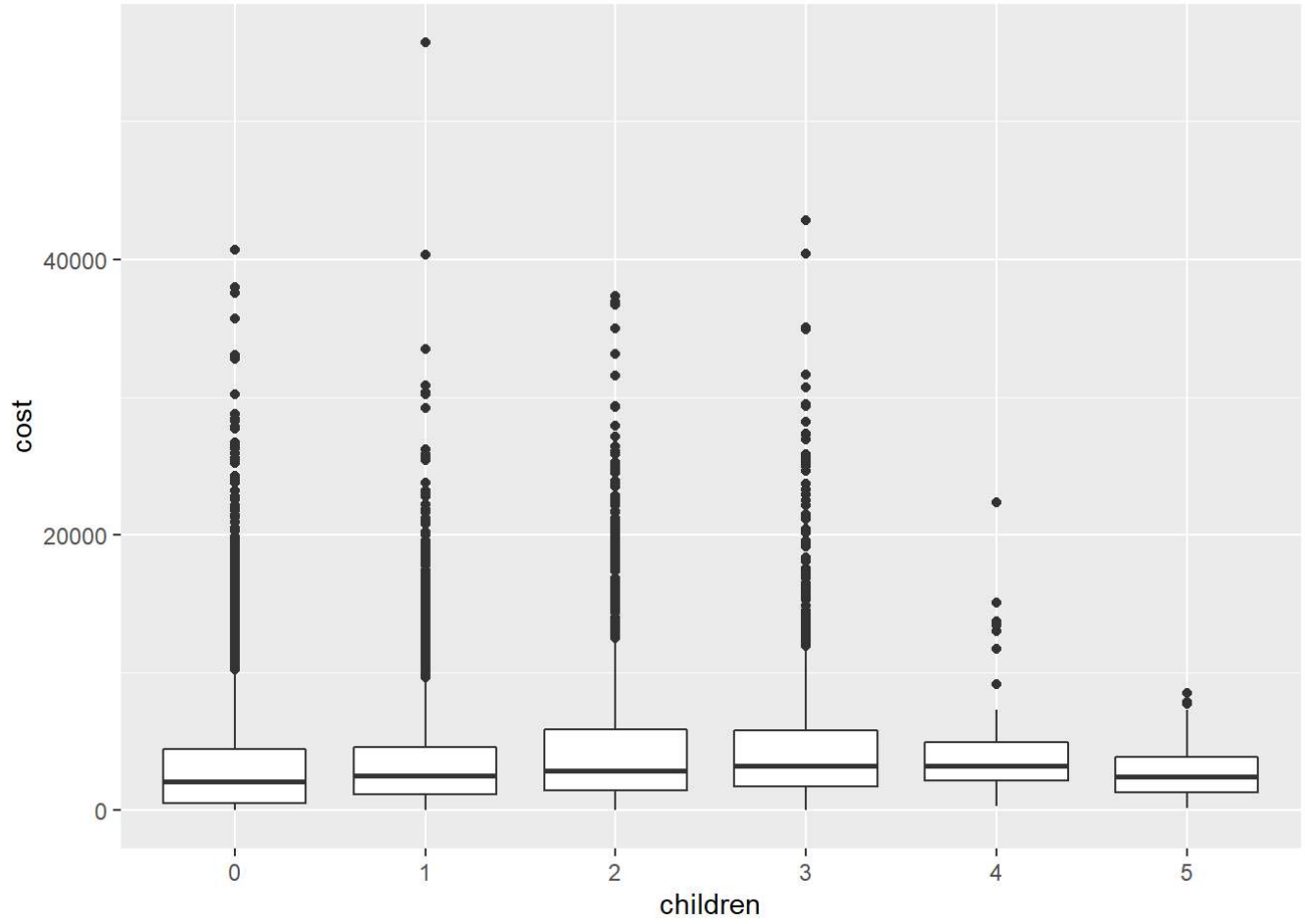



```
#Scattor plot usinf GGPLOT2
my_plot <- ggplot(data)
my_plot <- my_plot + geom_point(aes(x=bmi, y= cost, color = factor(expensive),alpha = 1/10 ))
my_plot <- my_plot + ggtitle("PositiveCases vs TotalTests - older_Adults")
my_plot
```

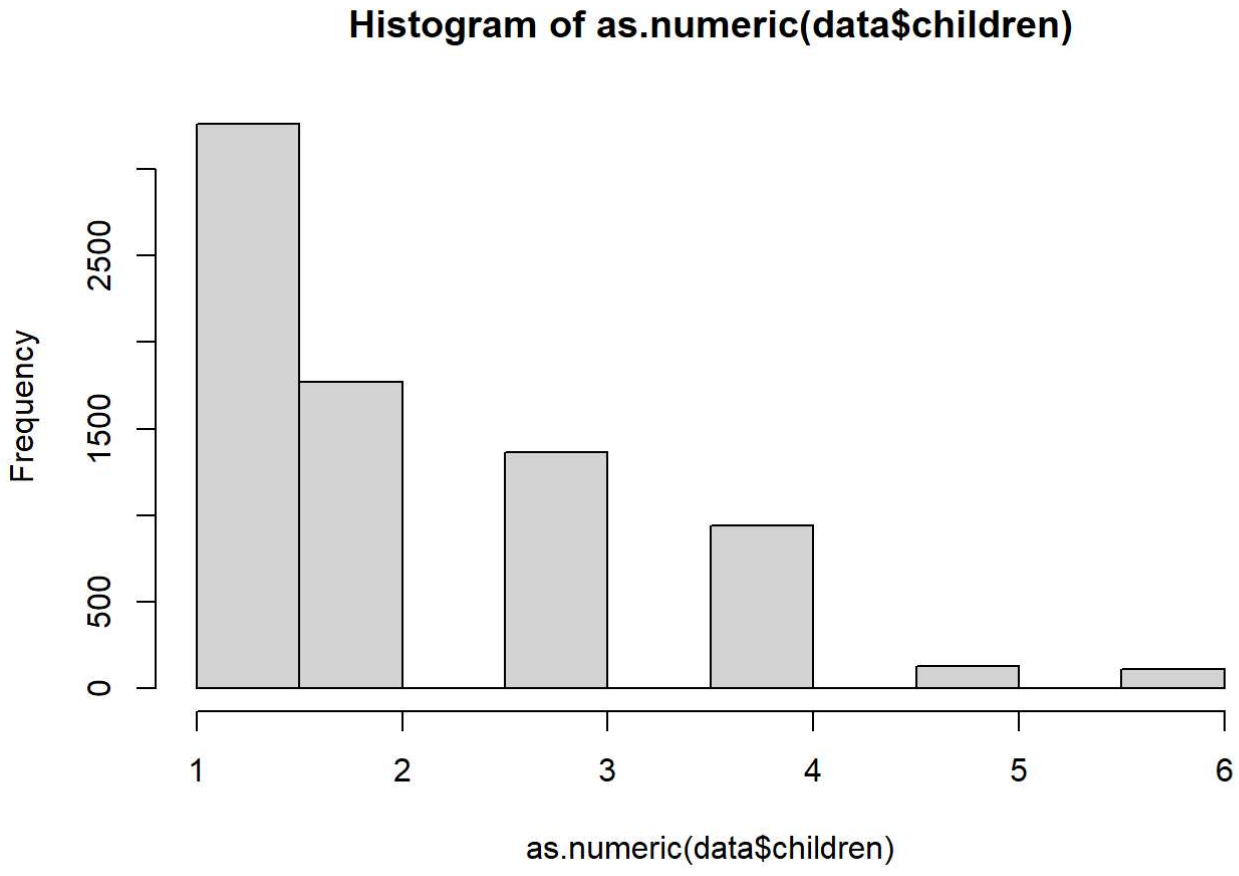



```
data$children <- as.factor(data$children)
data$hypertension <- as.factor(data$hypertension)
```

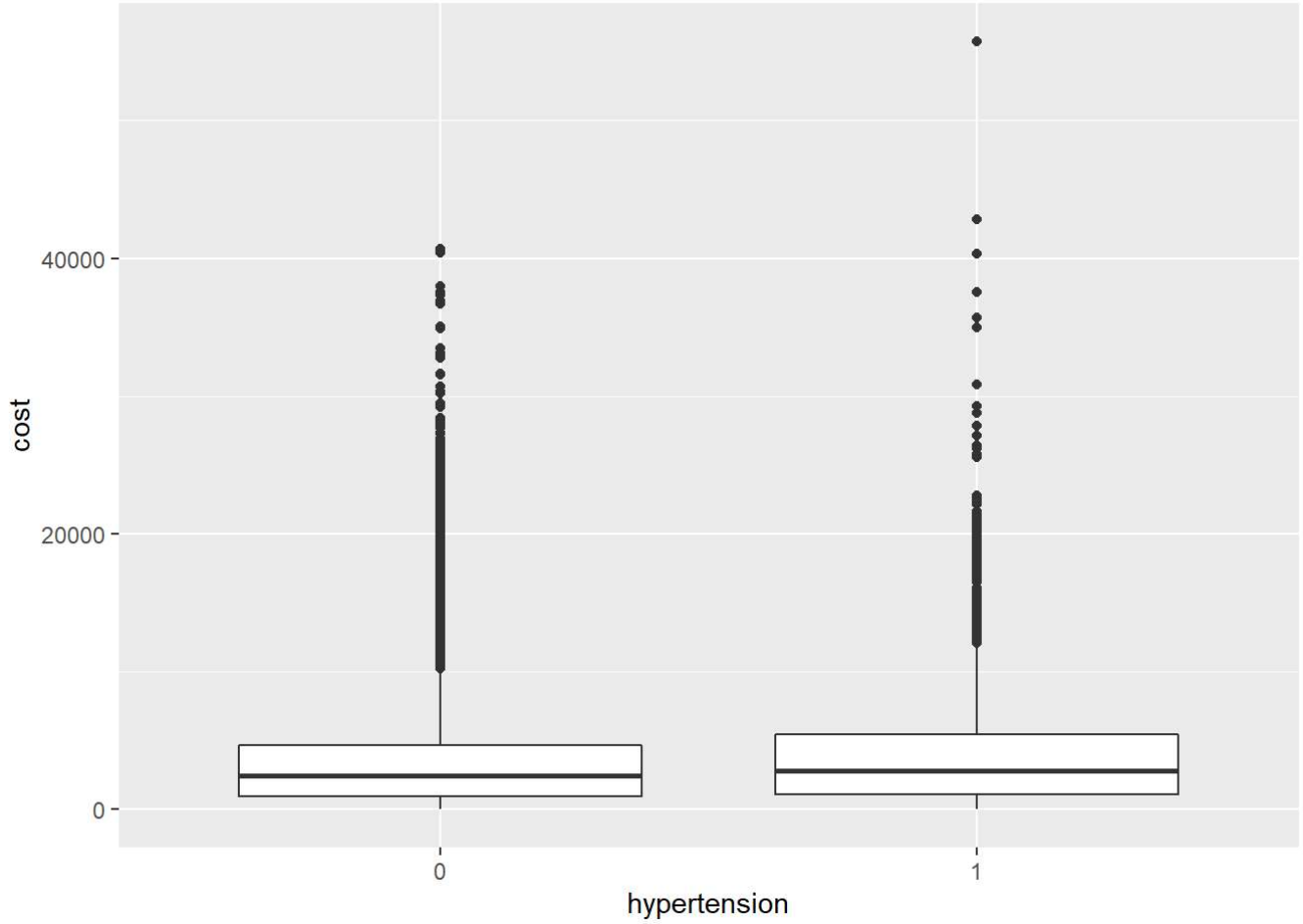
```
ggplot(data, aes(x= children , y=cost, fill= expensive)) +
    geom_boxplot()
```




```
hist(as.numeric(data$children))
```



```
ggplot(data, aes(x= hypertension , y=cost, fill= expensive)) +
    geom_boxplot()
```

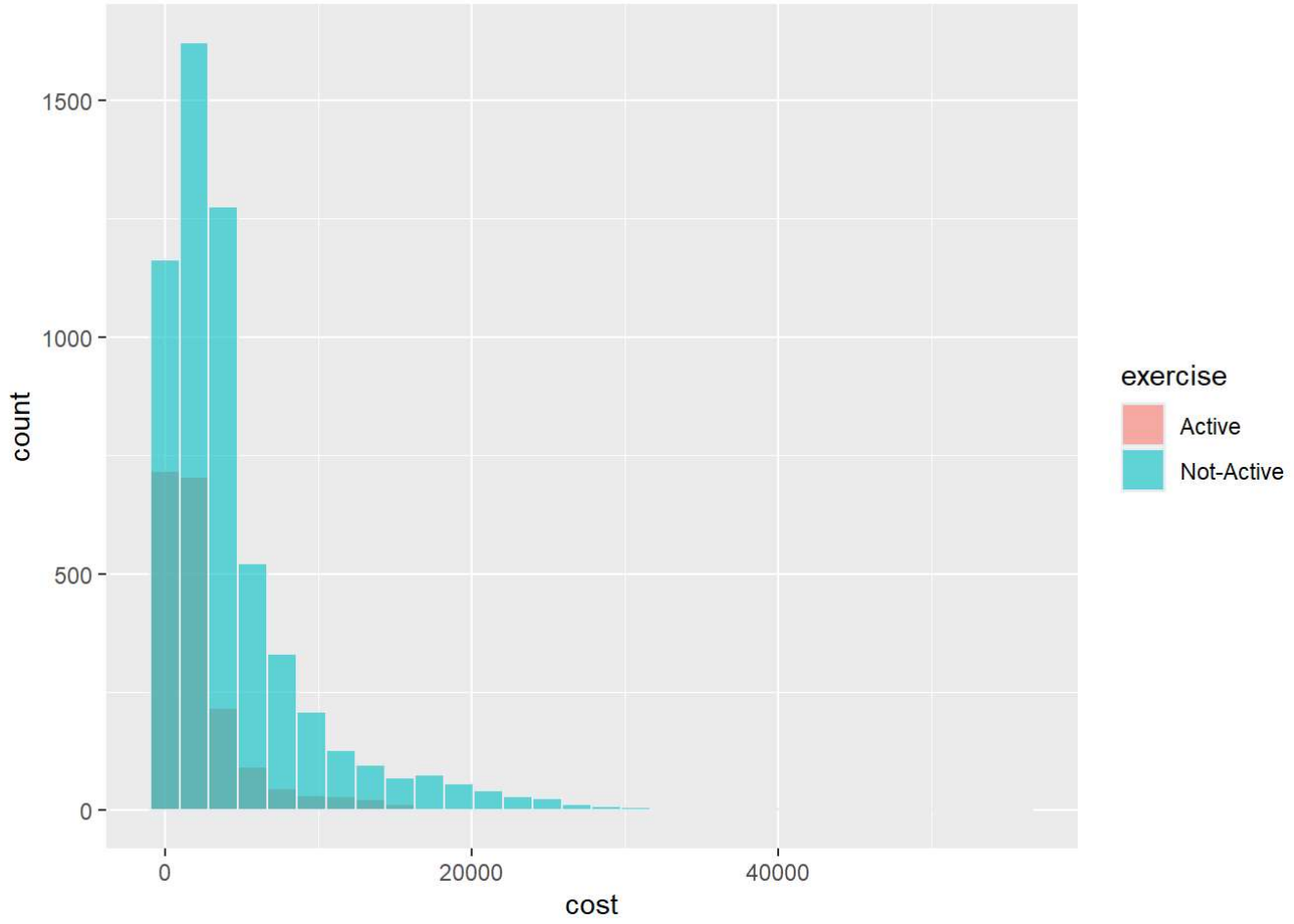



```
table(data$hypertension)
```

```
##
##    0    1
## 6078 1504
```

```
ggplot(data , aes(x=cost, fill=exercise)) +
  geom_histogram( color='#e9ecef', alpha=0.6, position='identity')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
table(data$exercise)
```

```
##
##     Active Not-Active
##       1888       5694
```

```
data %>% group_by(location) %>% summarise(median = median(cost))
```

```
## # A tibble: 7 × 2
##   location      median
##   <chr>          <dbl>
## 1 CONNECTICUT    2362
## 2 MARYLAND       2352
## 3 MASSACHUSETTS  2887
## 4 NEW JERSEY     2552.
## 5 NEW YORK       2910
## 6 PENNSYLVANIA   2462
## 7 RHODE ISLAND   2448.
```

```
data %>% group_by(gender) %>% summarise(median = median(cost))
```

```
## # A tibble: 2 × 2
##   gender median
##   <chr>   <dbl>
## 1 female  2410.
## 2 male    2607
```

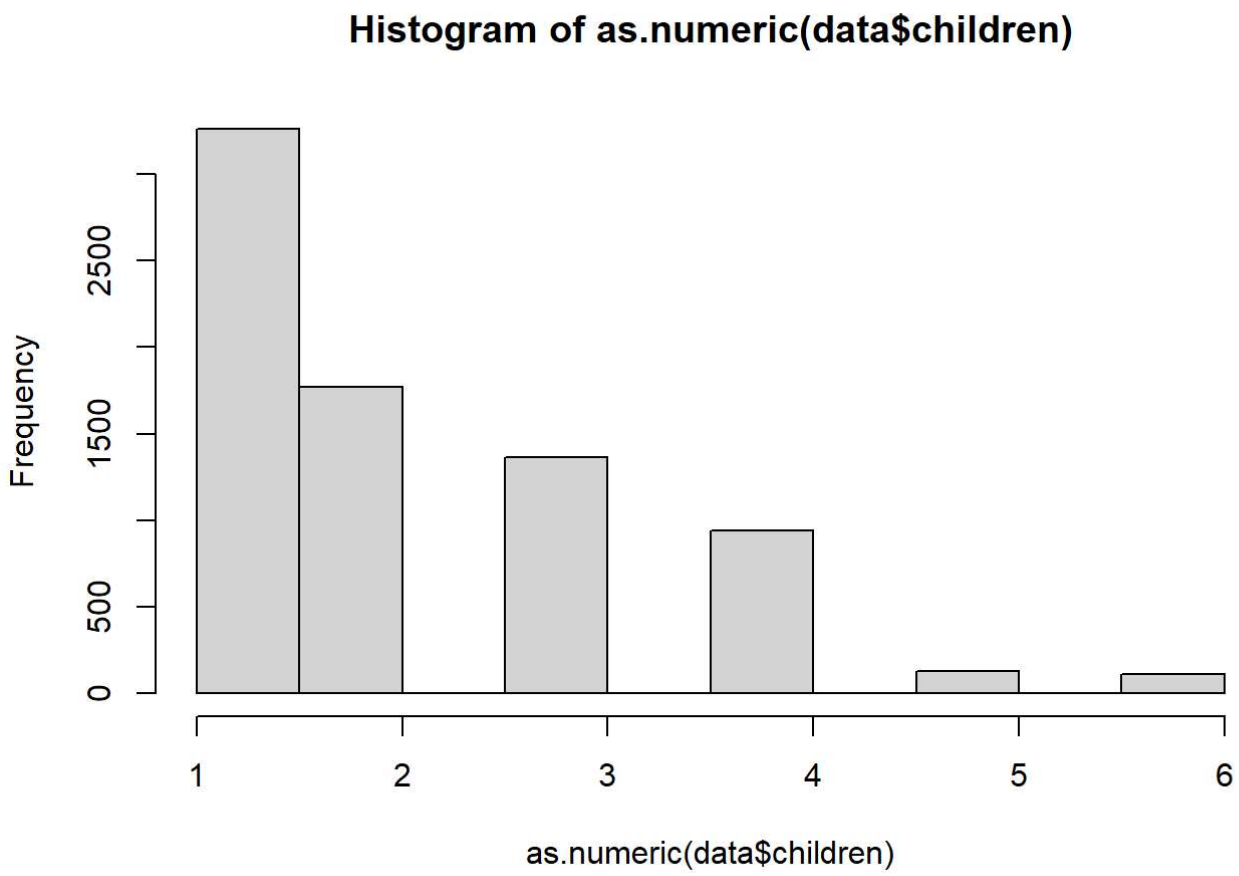
```
data %>% group_by(hypertension) %>% summarise(median = median(cost))
```

```
## # A tibble: 2 × 2
##   hypertension median
##   <fct>         <dbl>
## 1 0             2449
## 2 1             2778.
```

```
data %>% group_by(children) %>% summarise(median = median(cost))
```

```
## # A tibble: 6 × 2
##   children median
##   <fct>     <dbl>
## 1 0         2072
## 2 1         2490.
## 3 2         2867
## 4 3         3226
## 5 4         3228.
## 6 5         2479
```

```
hist(as.numeric(data$children))
```




```
location_data <- data %>% group_by(location) %>% summarise(sum = sum(cost), median = median(c
ost))
```

```
location_data
```

```
## # A tibble: 7 × 3
##   location           sum median
##   <chr>            <dbl>  <dbl>
## 1 CONNECTICUT    2350834  2362
## 2 MARYLAND       2826778  2352
## 3 MASSACHUSETTS  1984406  2887
## 4 NEW JERSEY     1957421  2552.
## 5 NEW YORK       2549844  2910
## 6 PENNSYLVANIA  16132692  2462
## 7 RHODE ISLAND   2851757  2448.
```

```
table(data$location)
```

```
##
##   CONNECTICUT      MARYLAND MASSACHUSETTS    NEW JERSEY      NEW YORK
##           611           747           465           498           547
##  PENNSYLVANIA  RHODE ISLAND
##          4010           704
```

```
us <- map_data("state")
```

```
location_data$location <- tolower(location_data$location)
```

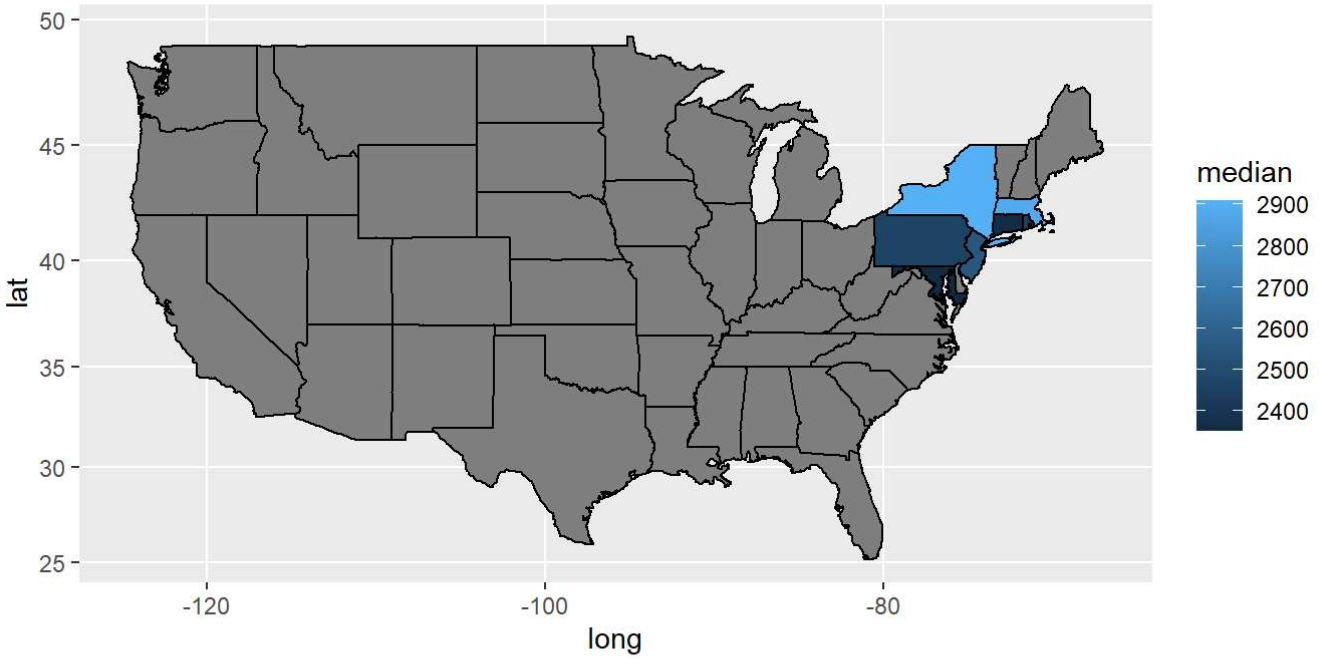
```
dfStatesWithGeom <- merge(us, location_data, all.x=TRUE, by.x="region", by.y = "location")
```

```
d <- merge(us, location_data, by.x="region", by.y = "location")
```

```
dfStatesWithGeom <- dfStatesWithGeom %>% arrange(order)
```

```
myMap <- ggplot(dfStatesWithGeom) + geom_polygon( color = "black", aes(x=long,y=lat, group =
group, fill =median)) +coord_map()
```

```
myMap
```




```
bb <- c(left = min(d$lon),
bottom = min(d$lat),
right = max(d$lon),
top = max(d$lat))
```

```
stateCenter <- data.frame(state= tolower(state.name),
                          x= state.center$x,
                          y =state.center$y
                          )
```

```
#dfStatesWithGeom
```

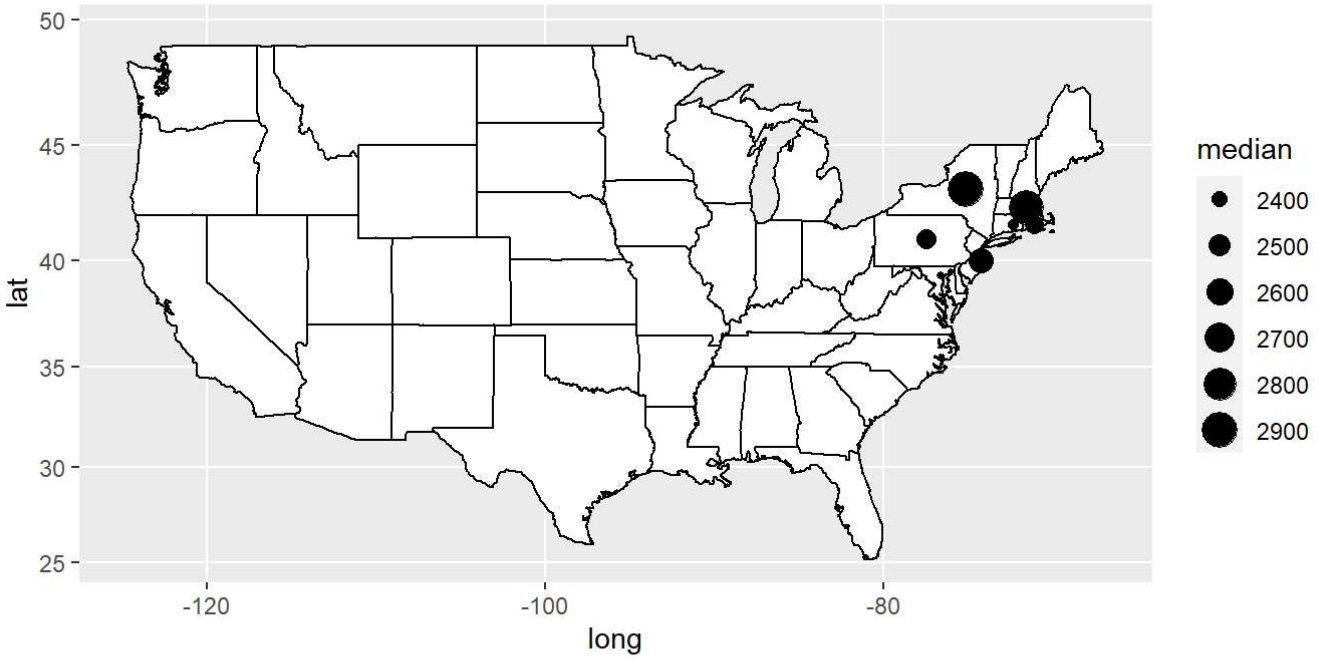
```
dfStatesWithCenter <- merge(dfStatesWithGeom,stateCenter, by.x="region", by.y = "state")
```

```
dfStatesWithCenter <- dfStatesWithCenter %>% arrange(order)
```

```
myMap2 <- ggplot(dfStatesWithCenter)+geom_polygon(fill="white", color = "black", aes(x= long, y = lat, group = group)) + geom_point(aes(x=x, y=y,size = median)) +coord_map()
```

```
myMap2
```

```
## Warning: Removed 13646 rows containing missing values (geom_point).
```




```
map <- get_stamenmap(bbox = bb, zoom = 6)
```

```
## Source : http://tile.stamen.com/terrain/6/17/23.png
```

```
## Source : http://tile.stamen.com/terrain/6/18/23.png
```

```
## Source : http://tile.stamen.com/terrain/6/19/23.png
```

```
## Source : http://tile.stamen.com/terrain/6/17/24.png
```

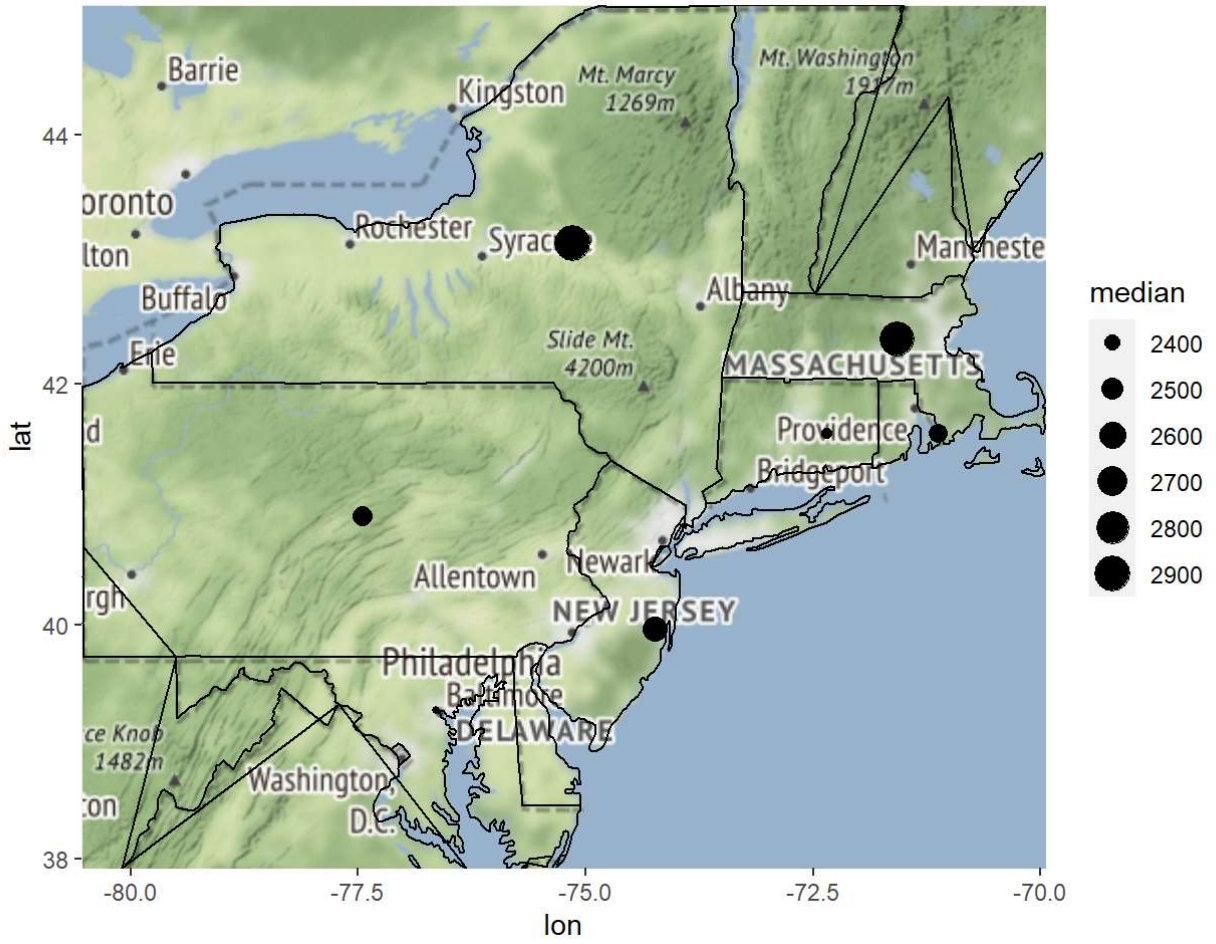
```
## Source : http://tile.stamen.com/terrain/6/18/24.png
```

```
## Source : http://tile.stamen.com/terrain/6/19/24.png
```

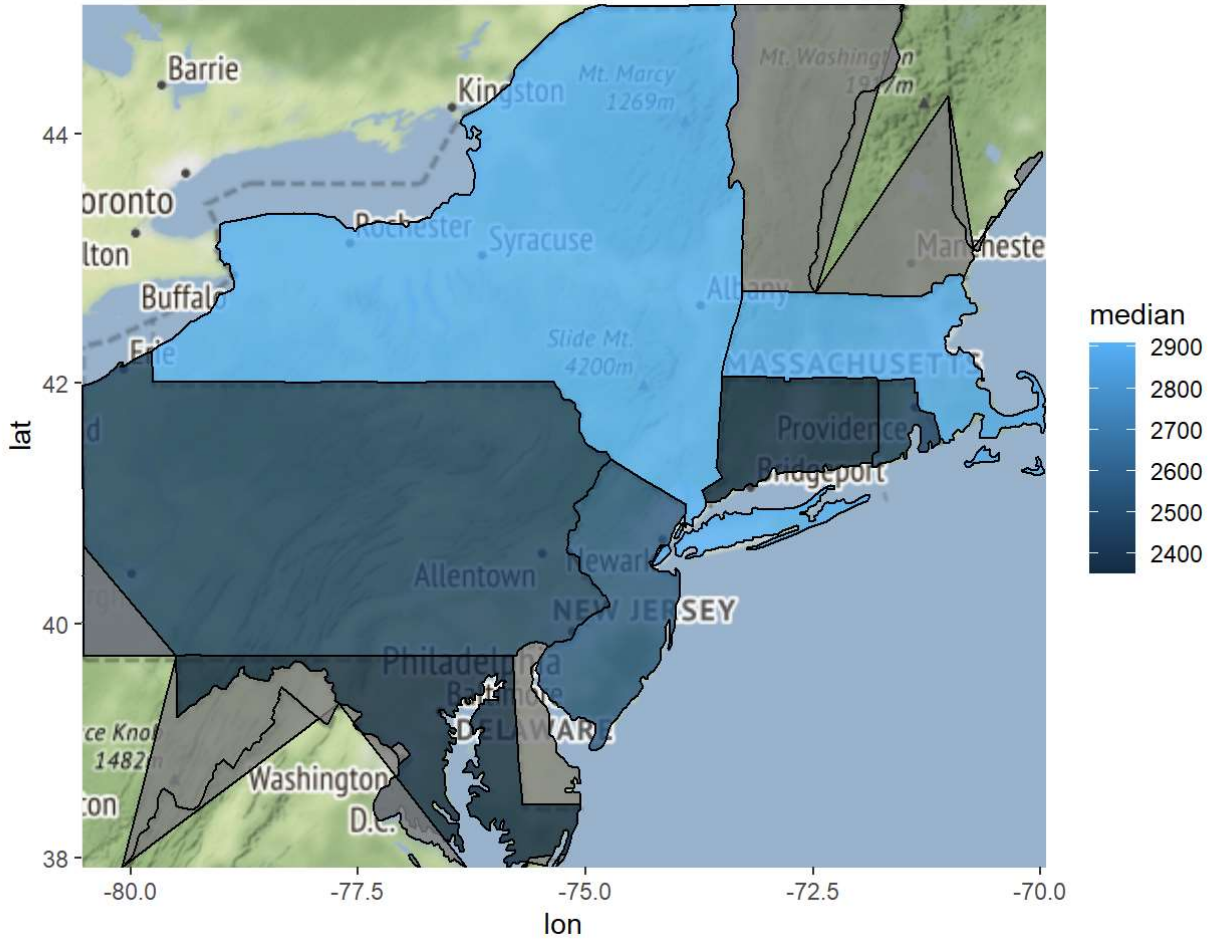
```
library(ggmap)
```

```
ggmap(map) +geom_polygon(data = dfStatesWithCenter, fill="NA", color = "black", aes(x= long,y = lat, group = group)) + geom_point(data = dfStatesWithCenter, aes(x=x, y=y,size = median))
```

```
## Warning: Removed 13646 rows containing missing values (geom_point).
```



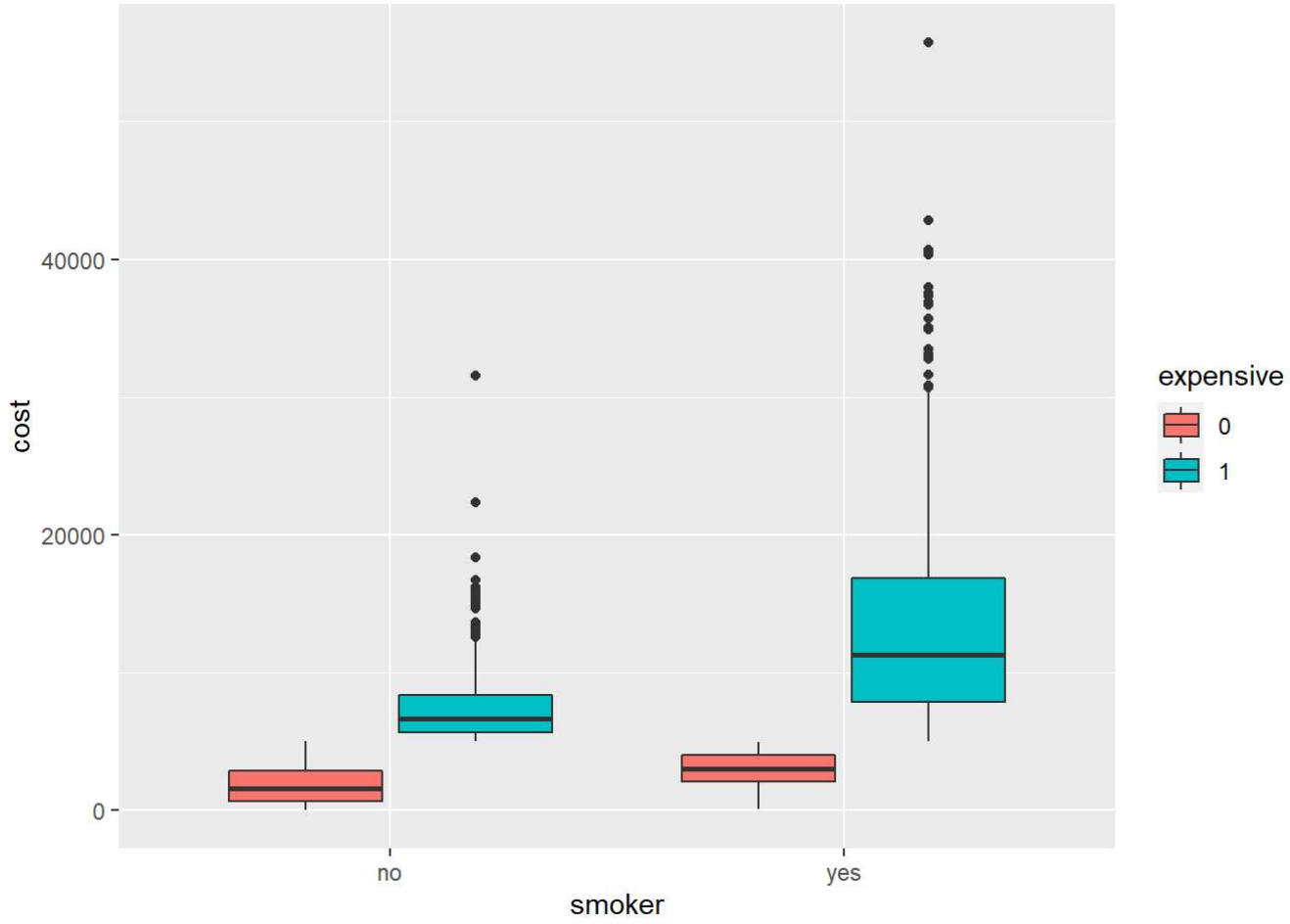

```
ggmap(map) + geom_polygon(data= dfStatesWithGeom, alpha = 0.8,aes(x = long, y= lat, group =group,fill = median), color ="black")
```



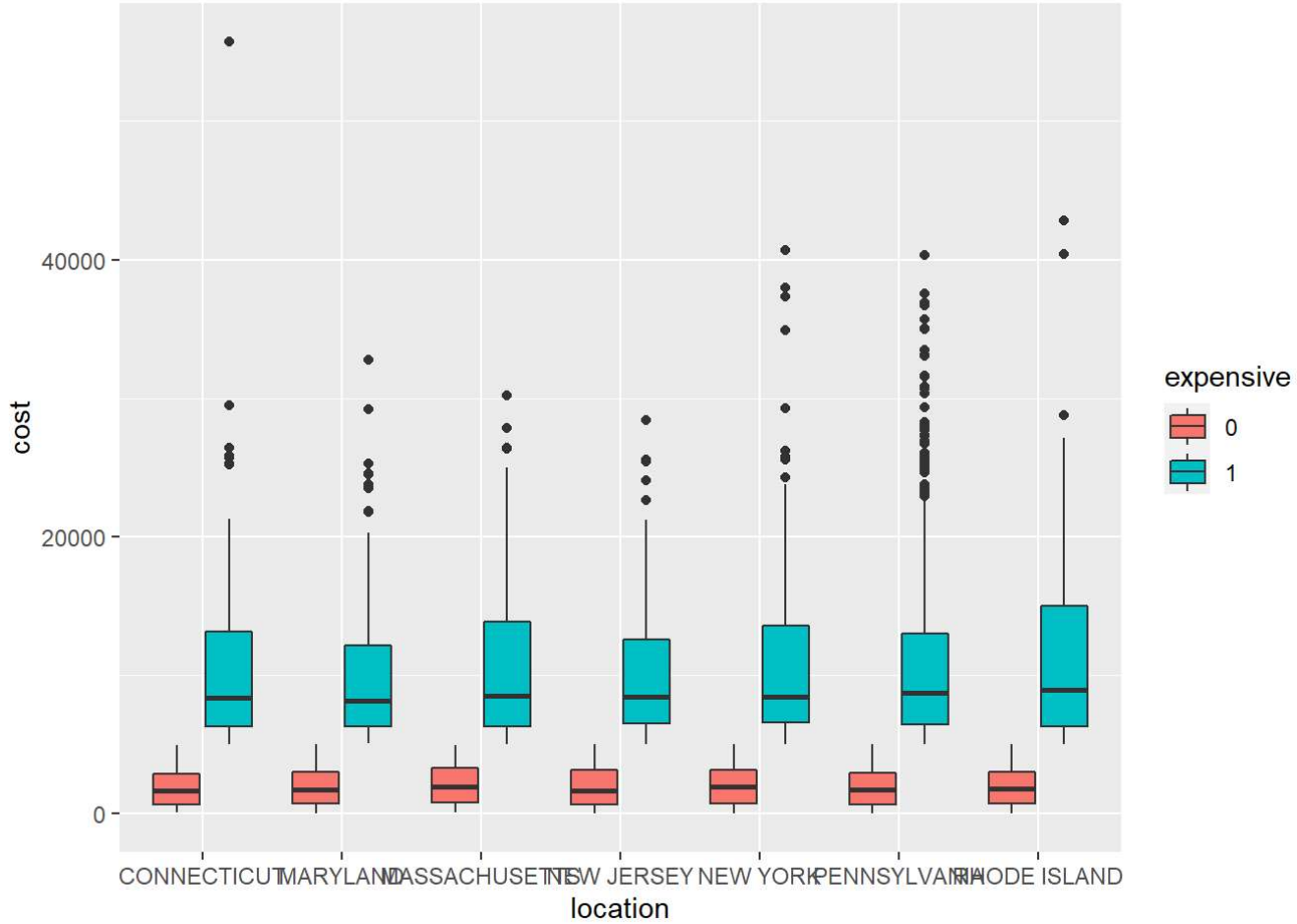

```
HMO_data <- data[, c('children', 'smoker', 'location', 'exercise', 'hypertension', 'gender',
'expensive', 'cost')]
```

```
HMO_data$expensive <- as.factor(HMO_data$expensive)
```

```
ggplot(HMO_data, aes(x= smoker , y=cost, fill= expensive)) +
    geom_boxplot()
```




```
ggplot(HMO_data, aes(x= location , y=cost, fill= expensive)) +
    geom_boxplot()
```



```
ggplot(data, aes(x= as.factor(expensive) , y= age, fill= expensive)) +
    geom_boxplot()
```