# Assignment 1 - Data Analysis and R Programming

Group 5 - Pat, Oat, Shine, Erika, David

2023-02-10

## Assignment 1 - Data Analysis and R Programming

*Author: GBC T405 BUS4066 - Group 5*

1. *Nichapat (Pat) Boonprasertsri, Student ID 101410612*
2. *Chotiros (Oat) Srisiam, Student ID 101411914*
3. *Shine Chen, Student ID 101450231*
4. *Erika Valle, Student ID 101381686*
5. *David Minotas, Student ID 101409821*

***Please import insurance_data.csv before run all code below***

```
insurance_data <- read.csv("https://raw.githubusercontent.com/pat-nb/gbc-t405-bus4066-assignment1-r/mai
```

### Data Preprocessing

#### Print the structure of dataset

```
str(insurance_data)
```

```
## 'data.frame':    1340 obs. of  11 variables:
##  $ index        : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ PatientID    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age          : num  39 24 NA NA NA NA NA 19 20 30 ...
##  $ gender       : chr  "male" "male" "male" "male" ...
##  $ bmi          : num  23.2 30.1 33.3 33.7 34.1 34.4 37.3 41.1 43 53.1 ...
##  $ bloodpressure: int  91 87 82 80 100 96 86 100 86 97 ...
##  $ diabetic     : chr  "Yes" "No" "Yes" "No" ...
##  $ children     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ smoker       : chr  "No" "No" "No" "No" ...
##  $ region       : chr  "southeast" "southeast" "southeast" "northwest" ...
##  $ claim        : num  1122 1132 1136 1136 1137 ...
```

#### List the variables in dataset

```r
names(insurance_data)
```

```
##  [1] "index"         "PatientID"     "age"           "gender"       
##  [5] "bmi"           "bloodpressure" "diabetic"      "children"     
##  [9] "smoker"        "region"        "claim"
```

### Print the top 15 rows of dataset

```r
head(insurance_data, n=15)
```

```
##    index PatientID age gender  bmi bloodpressure diabetic children smoker
## 1      0         1  39   male 23.2            91      Yes        0     No
## 2      1         2  24   male 30.1            87       No        0     No
## 3      2         3  NA   male 33.3            82      Yes        0     No
## 4      3         4  NA   male 33.7            80       No        0     No
## 5      4         5  NA   male 34.1           100       No        0     No
## 6      5         6  NA   male 34.4            96      Yes        0     No
## 7      6         7  NA   male 37.3            86      Yes        0     No
## 8      7         8  19   male 41.1           100       No        0     No
## 9      8         9  20   male 43.0            86       No        0     No
## 10     9        10  30   male 53.1            97       No        0     No
## 11    10        11  36   male 19.8            88      Yes        0     No
## 12    11        12  37   male 20.3            90      Yes        0     No
## 13    12        13  19   male 20.7            81       No        0     No
## 14    13        14  32   male 27.6           100       No        0     No
## 15    14        15  40   male 28.7            81      Yes        0     No
##       region   claim
## 1  southeast 1121.87
## 2  southeast 1131.51
## 3  southeast 1135.94
## 4  northwest 1136.40
## 5  northwest 1137.01
## 6  northwest 1137.47
## 7  northwest 1141.45
## 8  northwest 1146.80
## 9  northwest 1149.40
## 10 northwest 1163.46
## 11 northwest 1241.57
## 12 northwest 1242.26
## 13 northwest 1242.82
## 14           1252.41
## 15           1253.94
```

### Write a user defined function using any of the variables from the data set

```r
cal_yob <- function(age) {
  2021 - age # assume that 2021 is the year that data is created
}
print(head(cal_yob(insurance_data$age), n = 20))
```

```
##  [1] 1982 1997   NA   NA   NA   NA   NA 2002 2001 1991 1985 1984 2002 1989 1981
## [16] 1989 1986 1980 1972 1973
```

### Use data manipulation techniques and filter rows based on any logical criteria that exist in dataset

```
# Attach tidyverse packages to use data manipulation, reading, transforming and visualizing datasets
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   1.0.1
## v tibble  3.1.7      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# Select first 10 rows
insurance_data %>%
  select(age, bmi, diabetic, claim) %>%
  filter(age <= 30, diabetic == "Yes") %>%
  slice_head(n = 10)
```

```
##    age  bmi diabetic   claim
## 1   18 35.5      Yes 1532.47
## 2   30 17.5      Yes 1621.34
## 3   29 20.4      Yes 1625.43
## 4   21 22.6      Yes 1628.47
## 5   29 26.8      Yes 1665.00
## 6   30 21.5      Yes 1702.46
## 7   30 23.8      Yes 1705.62
## 8   21 26.1      Yes 1708.93
## 9   21 23.3      Yes 1711.03
## 10  23 28.5      Yes 1712.23
```

### Identify the dependent & independent variables and use reshaping techniques and create a new data frame by joining those variables from dataset

```
# Create a new data frame with smoker and diabetic column
reshap_col_smoker_diab <- cbind(insurance_data$smoker, insurance_data$diabetic)
print(head(reshap_col_smoker_diab, n = 10))
```

```
##       [,1] [,2]
##  [1,] "No" "Yes"
##  [2,] "No" "No"
##  [3,] "No" "Yes"
##  [4,] "No" "No"
##  [5,] "No" "No"
##  [6,] "No" "Yes"
##  [7,] "No" "Yes"
##  [8,] "No" "No"
##  [9,] "No" "No"
## [10,] "No" "No"
```

```
# Find and store patients whose age is under 20
df_age_under_20 <- insurance_data %>%
                select(PatientID, age, bmi, diabetic, claim) %>%
                filter(age < 20) %>%
                arrange(age, by_group = TRUE)

# Find and store patients whose age is from 20 to 30
df_age_20_30 <- insurance_data %>%
              select(PatientID, age, bmi, diabetic, claim) %>%
              filter(age >= 20, age < 30) %>%
              arrange(age, by_group = TRUE)

# Create a new data frame with patients whose age under 30 by merging 2 prepared data frame
df_age_under_30 <- rbind(df_age_under_20, df_age_20_30)
print(head(df_age_under_30, n = 10))
```

```
##    PatientID age  bmi diabetic   claim
## 1         23  18 35.5      Yes 1532.47
## 2         42  18 27.8       No 1635.73
## 3        153  18 27.6      Yes 2523.17
## 4        245  18 25.5      Yes 3645.09
## 5        260  18 30.9       No 3877.30
## 6        327  18 30.8      Yes 4646.76
## 7        463  18 29.8       No 6406.41
## 8        518  18 36.0       No 7160.33
## 9        565  18 26.6       No 7742.11
## 10       581  18 32.0      Yes 8116.27
```

**Remove missing values in dataset.**

```
insurance_data %>%
  select(2:5, 11) %>%
  filter(!is.na(age)) %>%
  slice_head(n = 10)
```

```
##    PatientID age gender  bmi   claim
## 1          1  39   male 23.2 1121.87
## 2          2  24   male 30.1 1131.51
## 3          8  19   male 41.1 1146.80
## 4          9  20   male 43.0 1149.40
## 5         10  30   male 53.1 1163.46
## 6         11  36   male 19.8 1241.57
## 7         12  37   male 20.3 1242.26
## 8         13  19   male 20.7 1242.82
## 9         14  32   male 27.6 1252.41
## 10        15  40   male 28.7 1253.94
```

### Identify and remove duplicated data in dataset

```
# Identify duplicated data
insurance_data[duplicated(insurance_data)]
```

```
## data frame with 0 columns and 1340 rows
```

```
# Remove duplicated rows in a data frame
insurance_data %>%
  select(2:5, 11) %>%
  distinct() %>%
  slice_head(n = 10)
```

```
##    PatientID age gender  bmi   claim
## 1          1  39   male 23.2 1121.87
## 2          2  24   male 30.1 1131.51
## 3          3  NA   male 33.3 1135.94
## 4          4  NA   male 33.7 1136.40
## 5          5  NA   male 34.1 1137.01
## 6          6  NA   male 34.4 1137.47
## 7          7  NA   male 37.3 1141.45
## 8          8  19   male 41.1 1146.80
## 9          9  20   male 43.0 1149.40
## 10        10  30   male 53.1 1163.46
```

```
# Remove duplicated rows based on age
insurance_data %>%
  select(2:5, 11) %>%
  distinct(age)
```

```
##    age
## 1   39
## 2   24
## 3   NA
## 4   19
## 5   20
## 6   30
## 7   36
## 8   37
## 9   32
## 10  40
## 11  35
## 12  41
## 13  49
## 14  48
## 15  45
## 16  34
## 17  18
## 18  42
## 19  50
## 20  23
## 21  58
## 22  29
## 23  21
## 24  52
## 25  43
## 26  47
## 27  28
## 28  44
## 29  31
## 30  51
## 31  60
## 32  27
## 33  26
## 34  22
## 35  38
## 36  53
## 37  54
## 38  33
## 39  59
## 40  55
## 41  46
## 42  57
## 43  25
## 44  56
```

### Reorder multiple rows in descending order

```
insurance_data %>%
  select(2:5, 11) %>%
  arrange(-age, -claim) %>%
  slice_head(n = 10)
```

```
##    PatientID age gender  bmi    claim
## 1       1302  60 female 35.0 44641.20
## 2       1225  60 female 32.5 36898.73
## 3       1124  60 female 30.6 24059.68
## 4       1105  60 female 28.1 22331.57
## 5       1047  60 female 18.3 19023.26
## 6       1021  60 female 23.7 17626.24
## 7       1009  60 female 27.9 16884.92
## 8        865  60 female 37.5 12265.51
## 9        782  60 female 39.8 11090.72
## 10       773  60 female 41.5 10977.21
```

### Rename some of the column names in dataset

```
insurance_data %>%
  select(2:5, 11) %>%
  rename(patient_id=PatientID) %>%
  slice_head(n = 10)
```

```
##    patient_id age gender  bmi   claim
## 1           1  39   male 23.2 1121.87
## 2           2  24   male 30.1 1131.51
## 3           3  NA   male 33.3 1135.94
## 4           4  NA   male 33.7 1136.40
## 5           5  NA   male 34.1 1137.01
## 6           6  NA   male 34.4 1137.47
## 7           7  NA   male 37.3 1141.45
## 8           8  19   male 41.1 1146.80
## 9           9  20   male 43.0 1149.40
## 10         10  30   male 53.1 1163.46
```

### Add new variables in data frame by using a mathematical function

```
insurance_data %>%
  filter(!is.na(age)) %>%
  mutate(yob = 2021 - age) %>%
  slice_head(n = 10)
```

```
##    index PatientID age gender  bmi bloodpressure diabetic children smoker
## 1      0         1  39   male 23.2            91      Yes        0     No
## 2      1         2  24   male 30.1            87       No        0     No
## 3      7         8  19   male 41.1           100       No        0     No
## 4      8         9  20   male 43.0            86       No        0     No
## 5      9        10  30   male 53.1            97       No        0     No
## 6     10        11  36   male 19.8            88      Yes        0     No
## 7     11        12  37   male 20.3            90      Yes        0     No
## 8     12        13  19   male 20.7            81       No        0     No
## 9     13        14  32   male 27.6           100       No        0     No
## 10    14        15  40   male 28.7            81      Yes        0     No
##       region   claim  yob
## 1  southeast 1121.87 1982
## 2  southeast 1131.51 1997
## 3  northwest 1146.80 2002
## 4  northwest 1149.40 2001
## 5  northwest 1163.46 1991
## 6  northwest 1241.57 1985
## 7  northwest 1242.26 1984
## 8  northwest 1242.82 2002
## 9            1252.41 1989
## 10           1253.94 1981
```

### Create a training set using random number generator engine

```
# Total number of data
count(insurance_data)
```

```
##      n
## 1 1340
```

```r
# Number of selected 5% data
count(insurance_data) * 0.05
```

```
##    n
## 1 67
```

```r
# Random select 5% of records
set.seed(1234)
insurance_data %>%
  sample_frac(0.05, replace = FALSE)
```

```
##    index PatientID age gender  bmi bloodpressure diabetic children smoker
## 1   1307      1308  47   male 31.4           137       No        3    Yes
## 2   1017      1018  32 female 28.3            92       No        0    Yes
## 3   1124      1125  46 female 23.8            86      Yes        3    Yes
## 4   1003      1004  42 female 24.8            83       No        0    Yes
## 5    622       623  36   male 23.6            86       No        2     No
## 6    904       905  25   male 39.9            88       No        0     No
## 7    644       645  43   male 43.9            89      Yes        3     No
## 8    933       934  28 female 31.8           110      Yes        2     No
## 9    399       400  27 female 19.9            81       No        0     No
## 10   899       900  33   male 33.7            98       No        4     No
## 11    97        98  44   male 25.2            91       No        0     No
## 12  1126      1127  51 female 41.9           106       No        0     No
## 13   725       726  36 female 23.2           104      Yes        0     No
## 14   325       326  31 female 30.2            83       No        3     No
## 15  1102      1103  40 female 24.0            87      Yes        1     No
## 16   883       884  57 female 25.7            94       No        2     No
## 17   269       270  49 female 41.1            88      Yes        0     No
## 18   183       184  37   male 27.2            87      Yes        0     No
## 19   573       574  42 female 37.0            80       No        1     No
## 20     3         4  NA   male 33.7            80       No        0     No
## 21   551       552  37 female 25.8            87       No        1     No
## 22  1235      1236  19   male 34.4           106       No        0    Yes
## 23   951       952  50 female 27.8            80      Yes        3     No
## 24  1218      1219  37   male 34.4           126       No        0    Yes
## 25   995       996  55 female 39.1            83      Yes        3     No
## 26   478       479  38   male 41.2            81      Yes        1     No
## 27   633       634  29   male 25.4            87      Yes        0     No
## 28   900       901  40   male 38.4            82       No        0     No
## 29   577       578  25   male 32.3            98       No        1     No
## 30  1131      1132  43 female 27.6           107       No        2    Yes
## 31   130       131  30 female 31.9            89       No        0     No
## 32  1064      1065  28 female 20.0            96      Yes        2    Yes
## 33  1013      1014  45   male 27.4            83       No        1    Yes
## 34   739       740  37   male 29.2           109       No        1     No
## 35   297       298  40 female 29.3            87      Yes        1     No
## 36   257       258  33   male 28.9            87       No        0     No
## 37    78        79  42 female 31.5           100       No        0     No
## 38  1205      1206  20   male 34.9           124      Yes        0    Yes
## 39   304       305  32   male 31.7            83      Yes        2     No
## 40   695       696  26   male 30.2            99       No        1     No
## 41   306       307  44   male 27.0           100      Yes         2     No
## 42   901       902  19   male 21.4            85      Yes         0     No
## 43  1244      1245  22   male 36.3           123       No         2    Yes
## 44   560       561  50   male 26.1            96      Yes         2     No
## 45   135       136  29   male 23.7            89      Yes         0     No
## 46  1168      1169  36 female 27.6           118       No         1     No
## 47   958       959  28 female 25.1           103       No         0     No
## 48   122       123  52 female 35.6            89       No         0     No
## 49  1257      1258  43   male 36.7           139      Yes         1    Yes
## 50   607       608  50 female 27.8            80       No         2     No
## 51   494       495  32   male 26.0            90       No         0     No
## 52   533       534  32 female 33.1            94      Yes         0     No
## 53   802       803  42   male 49.1           109      Yes         0     No
## 54   207       208  48 female 40.2            82       No         0     No
## 55  1154      1155  31   male 23.8           126      Yes         0    Yes
## 56   853       854  27 female 30.8            97      Yes         3     No
## 57   568       569  58 female 38.3            87      Yes         0     No
## 58   950       951  32 female 27.7            86      Yes         3     No
## 59   247       248  31   male 30.3            92       No         0     No
## 60   664       665  43 female 34.1            81       No         0     No
## 61   594       595  33 female 33.3            93      Yes         0     No
## 62   433       434  36   male 37.1            88       No         1     No
## 63   756       757  26   male 35.6           106      Yes         4     No
## 64   759       760  34   male 29.0           110      Yes         0     No
## 65  1241      1242  32   male 33.4           112      Yes         2    Yes
## 66   275       276  37   male 33.2            90      Yes         2     No
## 67   168       169  26 female 25.7            88       No         1     No
##       region    claim
## 1  northwest 46130.53
## 2  northwest 17468.98
## 3  northeast 24106.91
## 4  southeast 16577.78
## 5  northeast  8603.82
## 6  southeast 12982.87
## 7  southeast  8944.12
## 8  northeast 13607.37
## 9  southeast  5458.05
## 10 southeast 12949.16
## 11 northwest  2045.69
## 12 southeast 24227.34
## 13 northeast 10197.77
## 14 northwest  4618.08
## 15 southeast 22192.44
## 16 southeast 12629.17
## 17 southwest  3989.84
## 18 southwest  2866.09
## 19 northwest  8023.14
## 20 northwest  1136.40
## 21 southwest  7624.63
## 22 southeast 37742.58
## 23 southeast 14001.29
## 24 southwest 36197.70
## 25 southeast 16085.13
## 26 southeast  6610.11
## 27 southwest  8782.47
## 28 northwest 12950.07
## 29 southwest  8062.76
## 30 northwest 24535.70
## 31 northwest  2261.57
## 32 northeast 19798.05
## 33 northeast 17178.68
## 34 southwest 10436.10
## 35 southeast  4350.51
## 36 northwest  3866.86
## 37 southeast  1877.93
## 38 southwest 34828.65
## 39 northwest  4433.39
## 40 southwest  9724.53
## 41 southeast  4435.09
## 42 southwest 12957.12
## 43 southwest 38711.00
## 44 southeast  7729.65
## 45 northwest  2352.97
## 46 northwest 28340.19
## 47 northwest 14254.61
## 48 southeast  2211.13
## 49 northeast 39774.28
## 50 southeast  8515.76
## 51 southeast  6837.37
## 52 southwest  7345.08
## 53 southeast 11381.33
## 54 northwest  3201.25
## 55 southeast 26926.51
## 56 southwest 12105.32
## 57 northeast  7935.29
## 58 southeast 14001.13
## 59 southeast  3704.35
## 60 southeast  9283.56
## 61 southeast  8283.68
## 62 southeast  6079.67
## 63 northeast 10736.87
## 64 northeast 10796.35
## 65 southwest 38415.47
## 66 northwest  4058.71
## 67 northwest  2710.83
```

### Print the summary statistics of dataset

```
insurance_data %>%
  group_by(age) %>%
  summarise(mean(claim)) %>%
  slice_head(n = 10)
```

```
## # A tibble: 10 x 2
##      age ‘mean(claim)‘
##    <dbl>         <dbl>
##  1    18        10724.
##  2    19        13785.
##  3    20        16147.
##  4    21        11827.
##  5    22        21664.
##  6    23        12455.
##  7    24         8709.
##  8    25        13165.
##  9    26        13743.
## 10    27        15201.
```

**Use any of the numerical variables from the dataset and perform the following statistical functions: Mean, Median, Mode, Range**

```
# Find mean of claim
cat("Mean claim = ", as.character(mean(insurance_data$claim)))
```

```
## Mean claim =  13252.745641791
```

```
# Find median of claim
cat("Median claim = ", median(insurance_data$claim))
```

```
## Median claim =  9369.615
```

```
# Find mode of age
cal_mode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
cat("Mode age = ", cal_mode(insurance_data$age))
```

```
## Mode age =  43
```

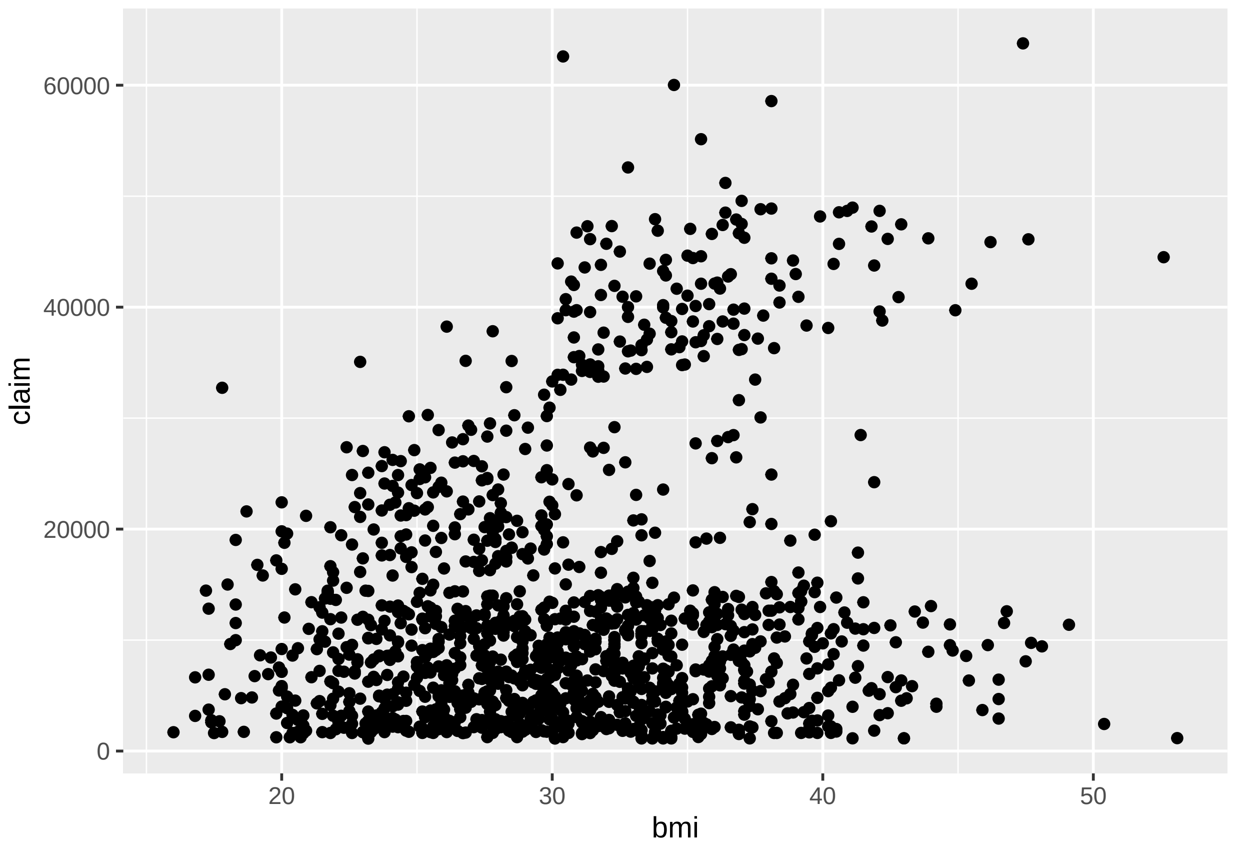
```
# Find range of claim
cat("Range claim = ",(range(insurance_data$claim)))
```

```
## Range claim =  1121.87 63770.43
```

## Visualization

### Plot a scatter plot for any 2 variables in dataset

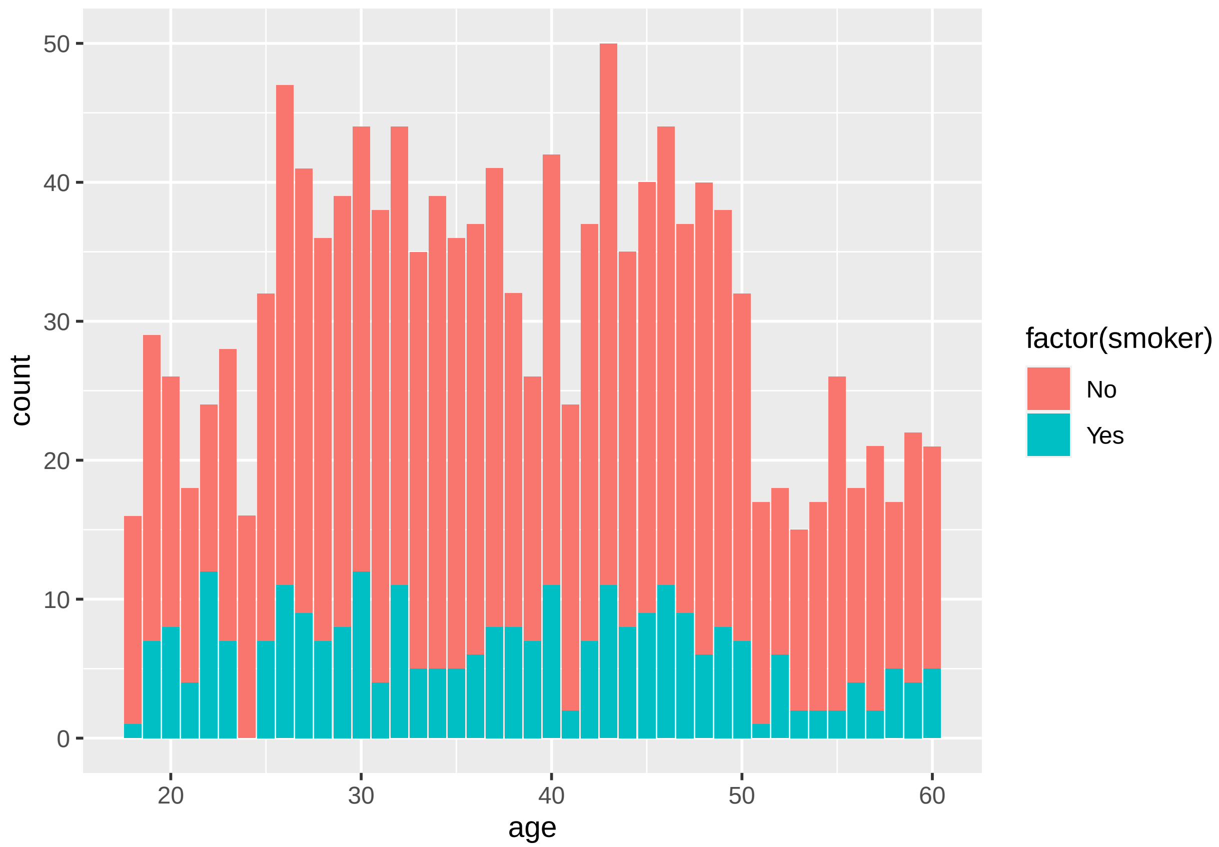
```
ggplot(data = insurance_data,
       mapping = aes(x=bmi,
                     y=claim)) +
  geom_point()
```



**Plot a bar plot for any 2 variables in dataset**

```
ggplot(data = insurance_data,
       aes(x = age, fill = factor(smoker))) +
  geom_bar()
```

```
## Warning: Removed 5 rows containing non-finite values (stat_count).
```



## Correlation

**Find the correlation between any 2 variables by applying least square linear regression model**

```
corr_bmi_claim = cor(insurance_data$claim, insurance_data$bmi, method="pearson")
print(corr_bmi_claim)
```

```
## [1] 0.1974013
```