Practice#1

Natalia

10 May 2019

```
set.seed(42)

load("C://Users//Natalia//Desktop//ITMO//R//Practice#1//CHIS2009_reduced_2.Rdata")

head(adult)
```

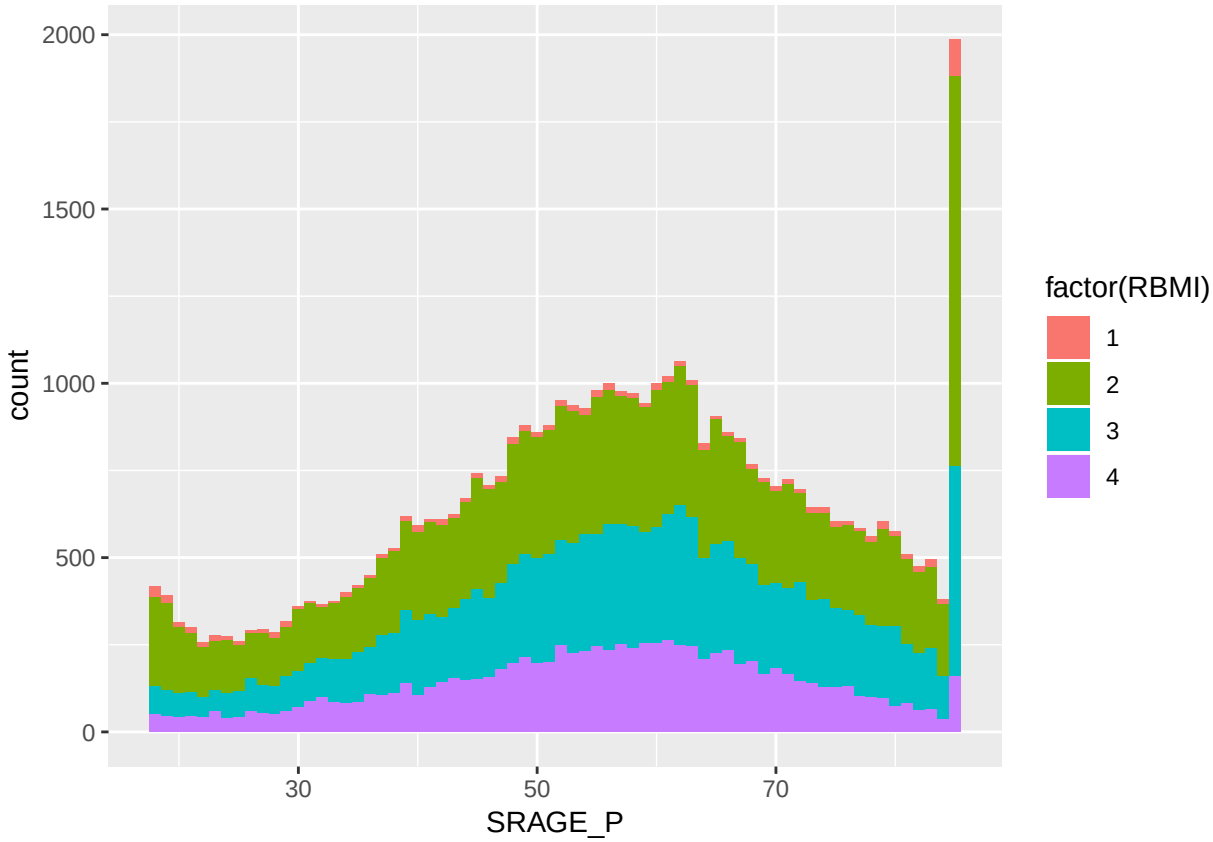
```
##   RBMI BMI_P RACEHPR2 SRSEX SRAGE_P MARIT2 AB1 ASTCUR AB51 POVLL
## 1    3 28.89        6     1      32      1   1      2   -1     4
## 2    3 26.15        6     2      80      3   1      2   -1     4
## 3    3 25.06        6     1      71      1   2      1   -1     4
## 4    2 24.99        6     1      39      4   1      2   -1     4
## 5    3 25.09        6     1      75      1   2      2   -1     4
## 6    4 32.21        6     2      53      1   3      1   -1     4
```

```
summary(adult)
```

```
##       RBMI           BMI_P          RACEHPR2         SRSEX
##  Min.   :1.000   Min.   :12.65   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:22.77   1st Qu.:5.000   1st Qu.:1.000
##  Median :3.000   Median :25.72   Median :6.000   Median :2.000
##  Mean   :2.748   Mean   :26.64   Mean   :5.088   Mean   :1.591
##  3rd Qu.:3.000   3rd Qu.:29.32   3rd Qu.:6.000   3rd Qu.:2.000
##  Max.   :4.000   Max.   :93.72   Max.   :6.000   Max.   :2.000
##     SRAGE_P          MARIT2           AB1             ASTCUR
##  Min.   :18.00   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:44.00   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :57.00   Median :1.000   Median :2.000   Median :2.000
##  Mean   :56.14   Mean   :2.043   Mean   :2.525   Mean   :1.915
##  3rd Qu.:69.00   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:2.000
##  Max.   :85.00   Max.   :4.000   Max.   :5.000   Max.   :2.000
##       AB51             POVLL
##  Min.   :-1.0000   Min.   :1.000
##  1st Qu.:-1.0000   1st Qu.:2.000
##  Median :-1.0000   Median :4.000
##  Mean   :-0.7108   Mean   :3.196
##  3rd Qu.:-1.0000   3rd Qu.:4.000
##  Max.   : 3.0000   Max.   :4.000
```

```
library(ggplot2)

p = ggplot(adult, aes(x = SRAGE_P, fill = factor(RBMI)))
p + geom_histogram(binwidth = 1)
```




```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
SRAGE_P= as.numeric(adult$SRAGE_P)
df = filter(adult, SRAGE_P < 85)
summary(df)
```

```
##       RBMI           BMI_P          RACEHPR2         SRSEX
##  Min.   :1.000   Min.   :12.65   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:22.85   1st Qu.:5.000   1st Qu.:1.000
##  Median :3.000   Median :25.75   Median :6.000   Median :2.000
##  Mean   :2.764   Mean   :26.74   Mean   :5.056   Mean   :1.588
##  3rd Qu.:3.000   3rd Qu.:29.45   3rd Qu.:6.000   3rd Qu.:2.000
##  Max.   :4.000   Max.   :93.72   Max.   :6.000   Max.   :2.000
##     SRAGE_P          MARIT2          AB1             ASTCUR
##  Min.   :18.00   Min.   :1.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:44.00   1st Qu.:1.00   1st Qu.:2.000   1st Qu.:2.000
##  Median :56.00   Median :1.00   Median :2.000   Median :2.000
```

```
##  Mean   :54.79   Mean   :2.02   Mean   :2.511   Mean   :1.914
##  3rd Qu.:67.00   3rd Qu.:3.00   3rd Qu.:3.000   3rd Qu.:2.000
##  Max.   :84.00   Max.   :4.00   Max.   :5.000   Max.   :2.000
##       AB51             POVLL
##  Min.   :-1.0000   Min.   :1.000
##  1st Qu.:-1.0000   1st Qu.:2.000
##  Median :-1.0000   Median :4.000
##  Mean   :-0.7147   Mean   :3.203
##  3rd Qu.:-1.0000   3rd Qu.:4.000
##  Max.   : 3.0000   Max.   :4.000
```

```
df %>% filter(BMI_P >= 16) %>% filter(BMI_P < 52) %>% head()
```

```
##   RBMI BMI_P RACEHPR2 SRSEX SRAGE_P MARIT2 AB1 ASTCUR AB51 POVLL
## 1    3 28.89        6     1      32      1   1      2   -1     4
## 2    3 26.15        6     2      80      3   1      2   -1     4
## 3    3 25.06        6     1      71      1   2      1   -1     4
## 4    2 24.99        6     1      39      4   1      2   -1     4
## 5    3 25.09        6     1      75      1   2      2   -1     4
## 6    4 32.21        6     2      53      1   3      1   -1     4
```

```
library(dplyr)
df$RACEHPR2 = factor(df$RACEHPR2, labels = c("Latino", "Asian", "African American", "White"))
levels(df$RACEHPR2)
```

```
## [1] "Latino"           "Asian"            "African American"
## [4] "White"
```

```
head(df)
```

```
##   RBMI BMI_P RACEHPR2 SRSEX SRAGE_P MARIT2 AB1 ASTCUR AB51 POVLL
## 1    3 28.89    White     1      32      1   1      2   -1     4
## 2    3 26.15    White     2      80      3   1      2   -1     4
## 3    3 25.06    White     1      71      1   2      1   -1     4
## 4    2 24.99    White     1      39      4   1      2   -1     4
## 5    3 25.09    White     1      75      1   2      2   -1     4
## 6    4 32.21    White     2      53      1   3      1   -1     4
```

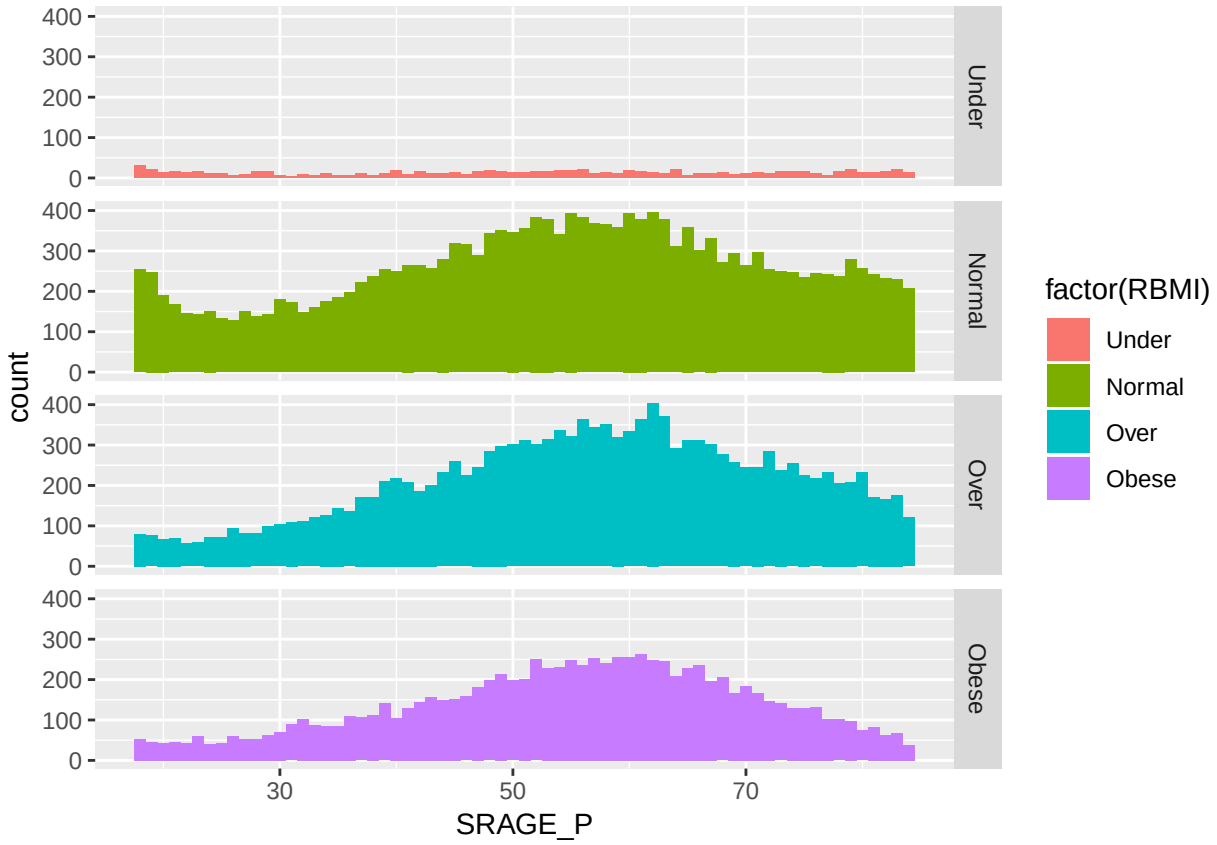
```
library(dplyr)
df$RBMI = factor(df$RBMI, labels = c("Under", "Normal", "Over", "Obese"))
levels(df$RBMI)
```

```
## [1] "Under"  "Normal" "Over"   "Obese"
```

```
head(df)
```

```
##     RBMI BMI_P RACEHPR2 SRSEX SRAGE_P MARIT2 AB1 ASTCUR AB51 POVLL
## 1   Over 28.89    White     1      32      1   1      2   -1     4
## 2   Over 26.15    White     2      80      3   1      2   -1     4
## 3   Over 25.06    White     1      71      1   2      1   -1     4
## 4 Normal 24.99    White     1      39      4   1      2   -1     4
## 5   Over 25.09    White     1      75      1   2      2   -1     4
## 6  Obese 32.21    White     2      53      1   3      1   -1     4
```

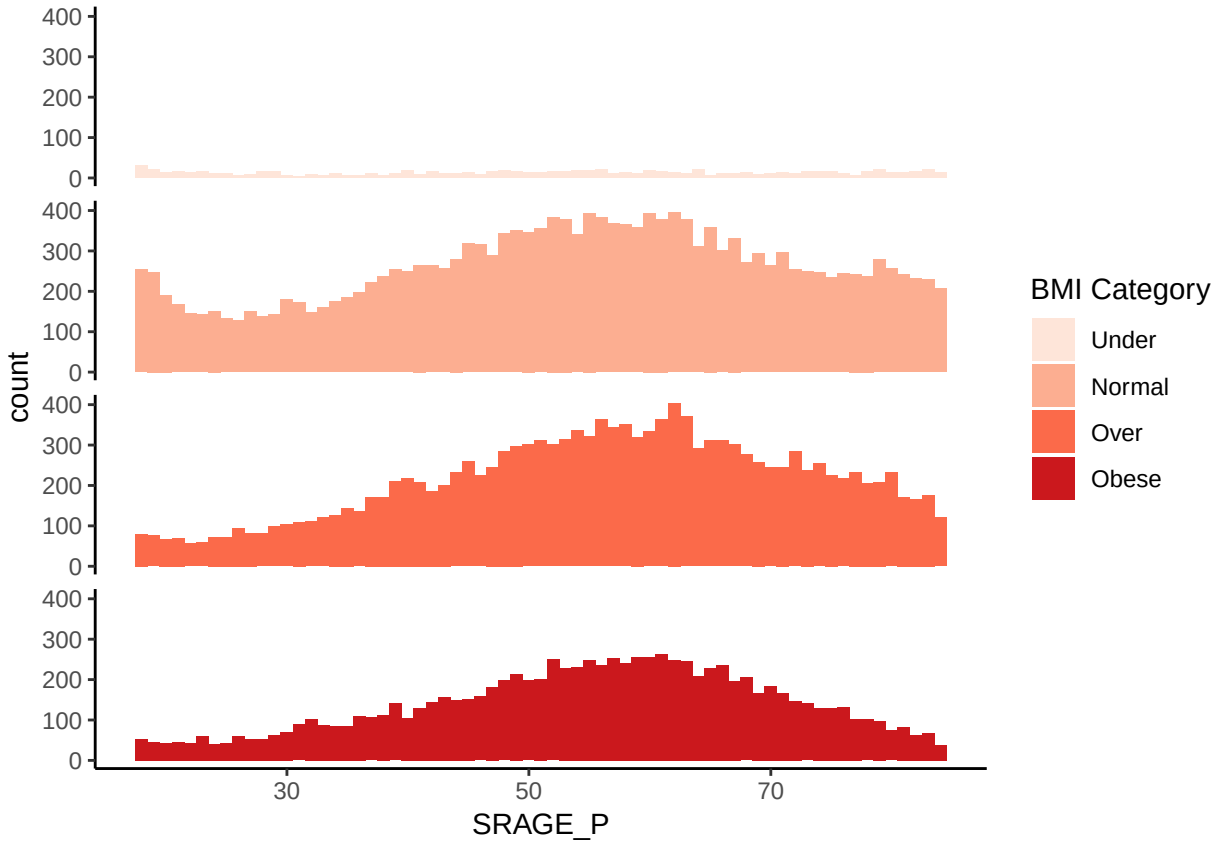
```
ggplot(df, aes(x = SRAGE_P, fill = factor(RBMI))) +
  geom_histogram(binwidth = 1) +
  facet_grid(RBMI ~.)
```



Add facet_grid

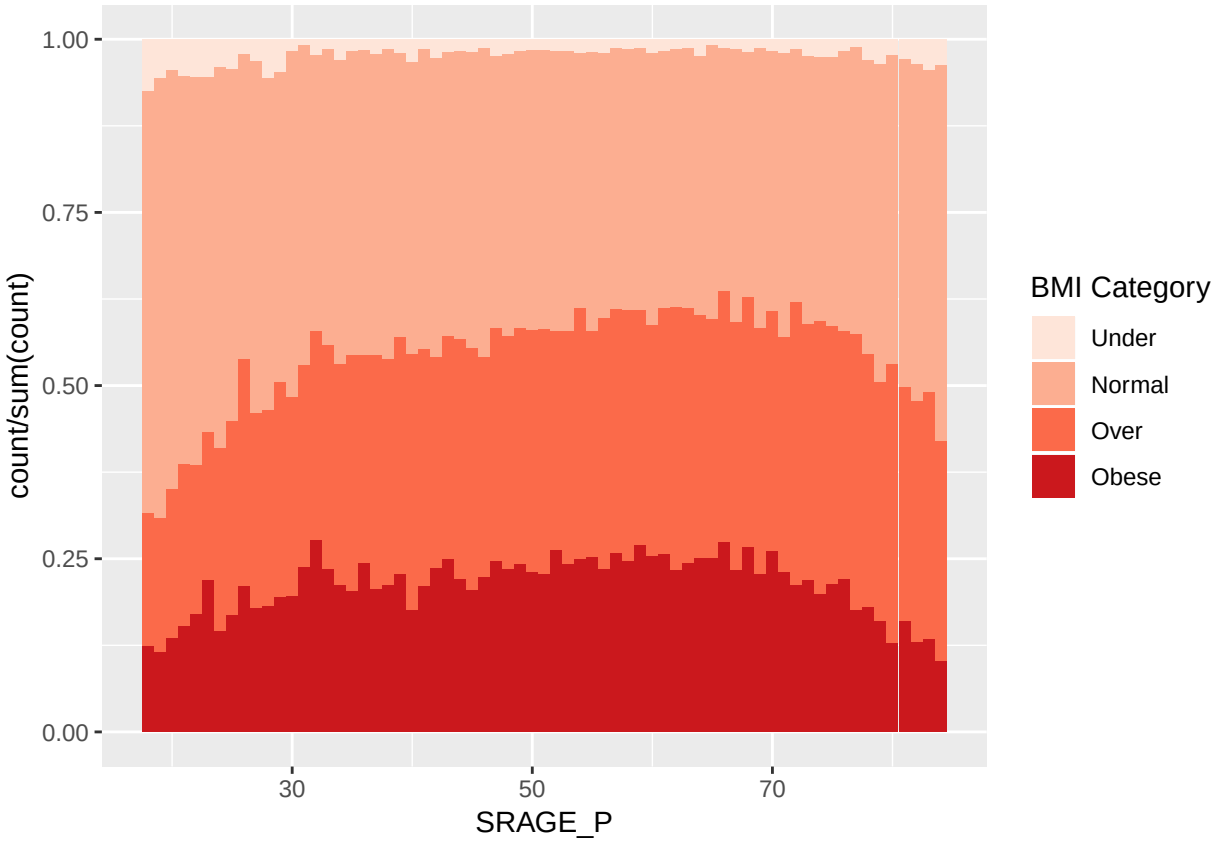
```
BMI_fill = scale_fill_brewer("BMI Category", palette = "Reds")
strip = theme(strip.text.y = element_blank())

ggplot(df, aes(x = SRAGE_P, fill = factor(RBMI))) +
  geom_histogram(binwidth = 1) +
  facet_grid(RBMI ~.) +
  BMI_fill + theme_classic() + strip
```



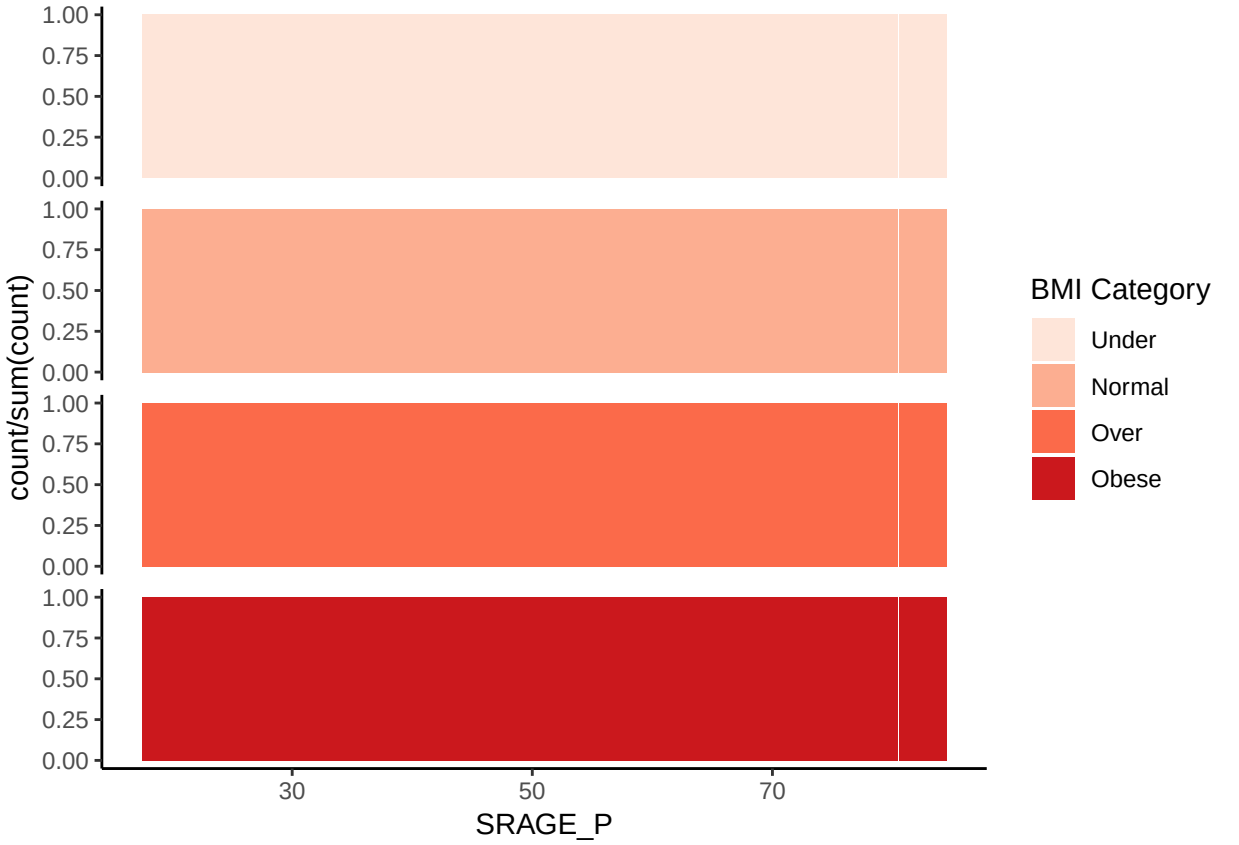
Proportional histogram

```
ggplot(df, aes (x = SRAGE_P, fill= factor(RBMI))) +
  geom_histogram(aes(y = ..count../sum(..count..)), binwidth = 1, position = "fill") +
  BMI_fill
```



Combine facet & proportional histogram

```
ggplot(df, aes(x = SRAGE_P, fill = as.factor(RBMI))) +
  geom_histogram(aes(y = ..count../sum(..count..)), binwidth = 1, position = "fill") +
  facet_grid(RBMI ~.) +
  BMI_fill + theme_classic() + strip
```



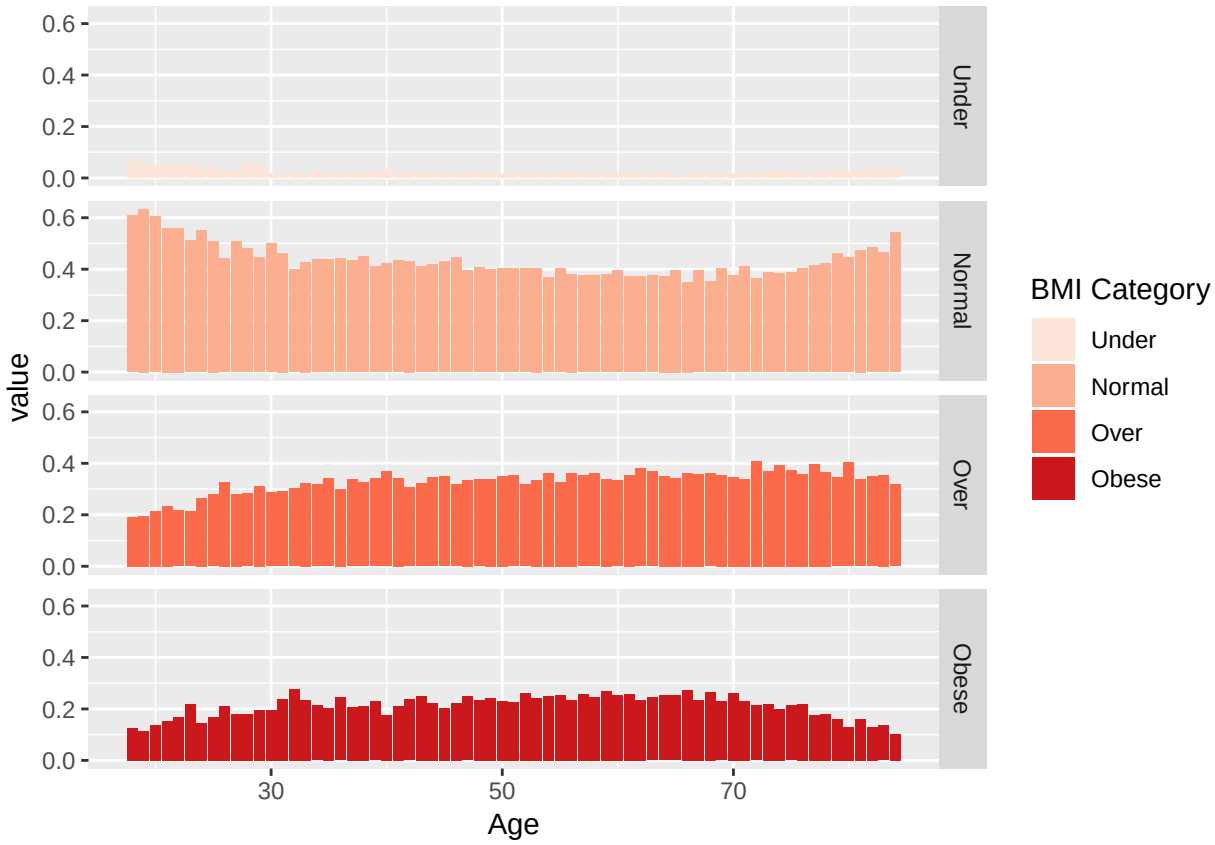
Create frequency table

```
df1 = table(df$RBMI, df$SRAGE_P)
proportion = apply(df1, 2, function(x) x/sum(x))
```

Plot frequency table

```
library(reshape2)
df_prop = melt(proportion)
names(df_prop) = c("FILL", "Age", "value")

ggplot(df_prop, aes(x = Age, y = value, fill = as.factor(FILL))) +
  geom_col(position = "stack") +
  BMI_fill +
  facet_grid(FILL ~ .)
```


value
0.6
0.4
0.2
0.0
Under
Normal
Over
Obese
30
50
70
Age
BMI Category
Under
Normal
Over
Obese