# Explore the dataset with summary and str

> 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

> str(adult)

'data.frame': 44346 obs. of 10 variables:

$ RBMI : num 3 3 3 2 3 4 3 2 3 3 ...

$ BMI\_P : num 28.9 26.1 25.1 25 25.1 ...

$ RACEHPR2: num 6 6 6 6 6 6 6 6 6 6 ...

$ SRSEX : num 1 2 1 1 1 2 1 2 1 2 ...

$ SRAGE\_P : num 32 80 71 39 75 53 42 33 67 52 ...

$ MARIT2 : num 1 3 1 4 1 1 1 1 3 3 ...

$ AB1 : num 1 1 2 1 2 3 2 2 1 5 ...

$ ASTCUR : num 2 2 1 2 2 1 2 2 2 2 ...

$ AB51 : num -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...

$ POVLL : num 4 4 4 4 4 4 4 3 4 4 ...

>

> # Age histogram

>

>

>

> # BMI value histogram

>

>

>

> # Age colored by BMI, binwidth = 1

> # Explore the dataset with summary and str

> 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

> str(adult)

'data.frame': 44346 obs. of 10 variables:

$ RBMI : num 3 3 3 2 3 4 3 2 3 3 ...

$ BMI\_P : num 28.9 26.1 25.1 25 25.1 ...

$ RACEHPR2: num 6 6 6 6 6 6 6 6 6 6 ...

$ SRSEX : num 1 2 1 1 1 2 1 2 1 2 ...

$ SRAGE\_P : num 32 80 71 39 75 53 42 33 67 52 ...

$ MARIT2 : num 1 3 1 4 1 1 1 1 3 3 ...

$ AB1 : num 1 1 2 1 2 3 2 2 1 5 ...

$ ASTCUR : num 2 2 1 2 2 1 2 2 2 2 ...

$ AB51 : num -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...

$ POVLL : num 4 4 4 4 4 4 4 3 4 4 ...

>

> # Age histogram

>

> ggplot(adult, aes(x = SRAGE\_P)) +

geom\_histogram()

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

>

> # BMI value histogram

>

> ggplot(adult, aes(x = BMI\_P)) +

geom\_histogram()

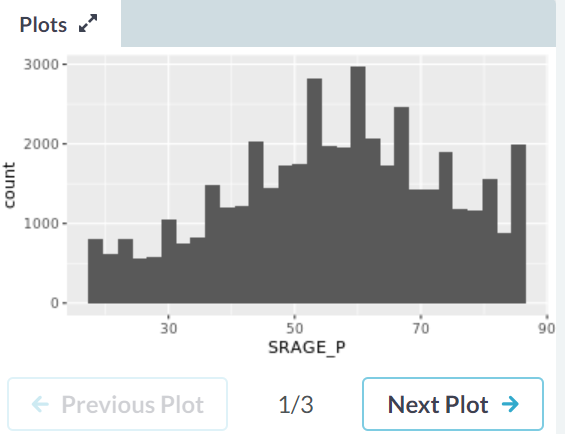
`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

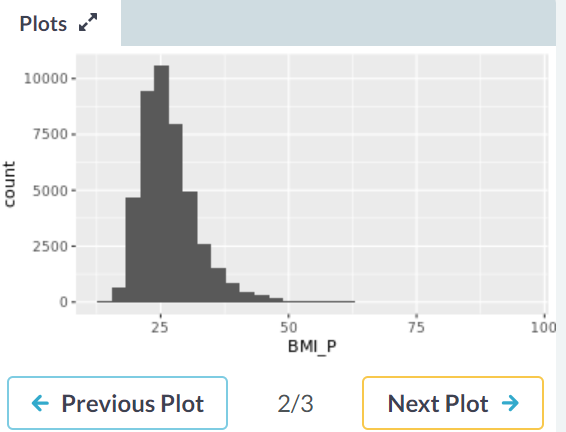
>

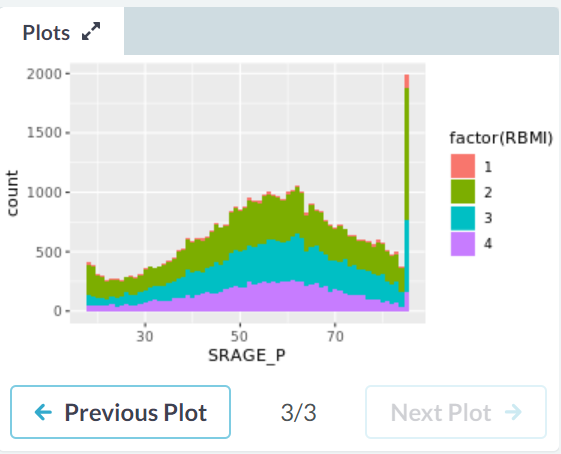
> # Age colored by BMI, binwidth = 1

> ggplot(adult, aes(x = SRAGE\_P, fill= factor(RBMI))) +

geom\_histogram(binwidth = 1)

>





# Remove individual aboves 84

> adult <- adult[adult$SRAGE\_P <= 84, ]

>

> # Remove individuals with a BMI below 16 and above or equal to 52

> adult <- adult[adult$BMI\_P >= 16 & adult$BMI\_P < 52, ]

>

> # Relabel the race variable:

> adult$RACEHPR2 <- factor(adult$RACEHPR2, labels = c("Latino", "Asian", "African American", "White"))

>

> str(adult)

'data.frame': 42167 obs. of 10 variables:

$ RBMI : num 3 3 3 2 3 4 3 2 3 3 ...

$ BMI\_P : num 28.9 26.1 25.1 25 25.1 ...

$ RACEHPR2: Factor w/ 4 levels "Latino","Asian",..: 4 4 4 4 4 4 4 4 4 4 ...

$ SRSEX : num 1 2 1 1 1 2 1 2 1 2 ...

$ SRAGE\_P : num 32 80 71 39 75 53 42 33 67 52 ...

$ MARIT2 : num 1 3 1 4 1 1 1 1 3 3 ...

$ AB1 : num 1 1 2 1 2 3 2 2 1 5 ...

$ ASTCUR : num 2 2 1 2 2 1 2 2 2 2 ...

$ AB51 : num -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...

$ POVLL : num 4 4 4 4 4 4 4 3 4 4 ...

>

> # Relabel the BMI categories variable

> adult$RBMI <- factor(adult$RBMI, labels = c("Under-weight", "Normal-weight", "Over-weight", "Obese"))

> str(adult)

'data.frame': 42167 obs. of 10 variables:

$ RBMI : Factor w/ 4 levels "Under-weight",..: 3 3 3 2 3 4 3 2 3 3 ...

$ BMI\_P : num 28.9 26.1 25.1 25 25.1 ...

$ RACEHPR2: Factor w/ 4 levels "Latino","Asian",..: 4 4 4 4 4 4 4 4 4 4 ...

$ SRSEX : num 1 2 1 1 1 2 1 2 1 2 ...

$ SRAGE\_P : num 32 80 71 39 75 53 42 33 67 52 ...

$ MARIT2 : num 1 3 1 4 1 1 1 1 3 3 ...

$ AB1 : num 1 1 2 1 2 3 2 2 1 5 ...

$ ASTCUR : num 2 2 1 2 2 1 2 2 2 2 ...

$ AB51 : num -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...

$ POVLL : num 4 4 4 4 4 4 4 3 4 4 ...

> # The dataset adult is available

>

> # The color scale used in the plot

> BMI\_fill <- scale\_fill\_brewer("BMI Category", palette = "Reds")

>

> # Theme to fix category display in faceted plot

> fix\_strips <- theme(

strip.text.y = element\_text(angle = 0, hjust = 0, vjust = 0.1, size = 14),

strip.background = element\_blank(),

legend.position = "none")

>

> # Histogram, add BMI\_fill and customizations

> ggplot(adult, aes (x = SRAGE\_P, fill= factor(RBMI))) +

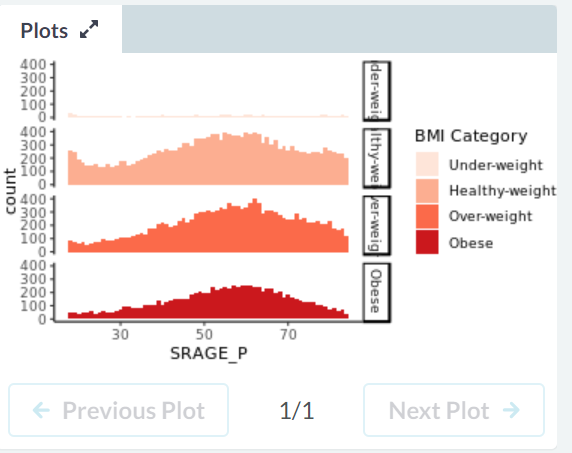
geom\_histogram(binwidth = 1) +

fix\_strips +

BMI\_fill +

facet\_grid(RBMI ~ .) +

theme\_classic()

>

> # Plot 1 - Count histogram

> ggplot(adult, aes (x = SRAGE\_P, fill= factor(RBMI))) +

geom\_histogram(binwidth = 1) +

BMI\_fill

>

> # Plot 2 - Density histogram

> ## This plot looks really strange, because we get the density within each BMI category, not within each age group!

> ggplot(adult, aes(x = SRAGE\_P, fill= factor(RBMI))) +

geom\_histogram(aes(y = ..density..), binwidth = 1) +

BMI\_fill

>

> # Plot 3 - Faceted count histogram

>

> ggplot(adult, aes (x = SRAGE\_P, fill= factor(RBMI))) +

geom\_histogram(binwidth = 1) +

BMI\_fill +

facet\_grid(RBMI ~ .)

>

>

> # Plot 4 - Faceted density histogram

>

> ## Plots 3 and 4 can be useful if we are interested in the frequency distribution within each BMI category.

> ggplot(adult, aes(x = SRAGE\_P, fill= factor(RBMI))) +

geom\_histogram(aes(y = ..density..), binwidth = 1) +

BMI\_fill +

facet\_grid(RBMI ~ .)

>

>

> # Plot 5 - Density histogram with position = "fill"

>

> ggplot(adult, aes(x = SRAGE\_P, fill= factor(RBMI))) +

geom\_histogram(aes(y = ..density..), binwidth = 1, position = "fill") +

BMI\_fill

>

>

> # Plot 6 - The accurate histogram

>

> ggplot(adult, aes(x = SRAGE\_P, fill= factor(RBMI))) +

geom\_histogram(aes(y = ..count../sum(..count..)), binwidth = 1, position = "fill") +

BMI\_fill

