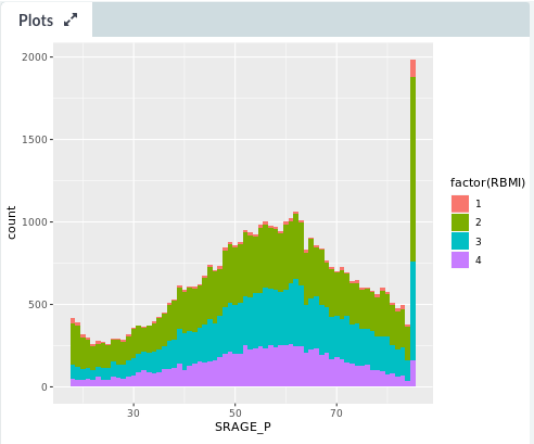
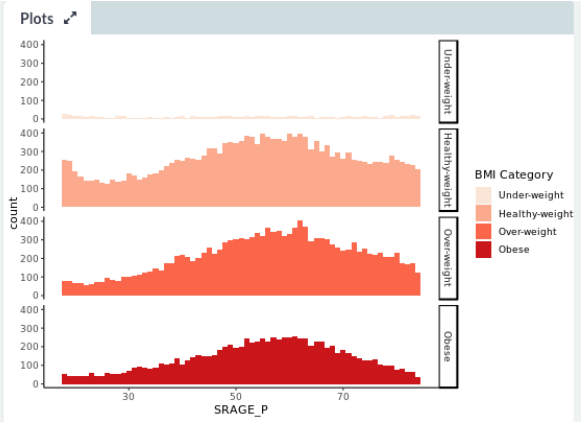
CHIS Exercise: Code

```
# Explore the dataset with summary and str
summary(adult)
str(adult)

# Age histogram
ggplot(adult, aes (x = SRAGE_P)) +
  geom_histogram()

# BMI histogram
ggplot(adult, aes (x = BMI_P)) +
  geom_histogram()

# Age colored by BMI, binwidth = 1
ggplot(adult, aes (x = SRAGE_P, fill= factor(RBMI))) +
  geom_histogram(binwidth = 1)
```



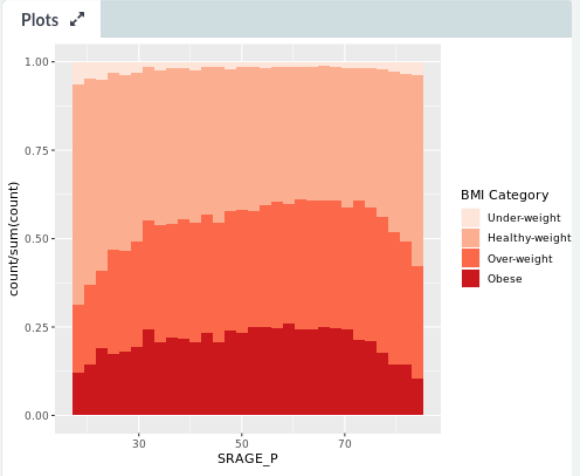

```
# Keep adults younger than or equal to 84
adult <- adult[adult$SRAGE_P <= 84, ]

# Keep adults with BMI at least 16 and less than 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"))
```

```
# Relabel the BMI categories variable
adult$RBMI <- factor(adult$RBMI, labels = c("Under-weight", "Normal-weight", "Over-weight", "Obese"))
```



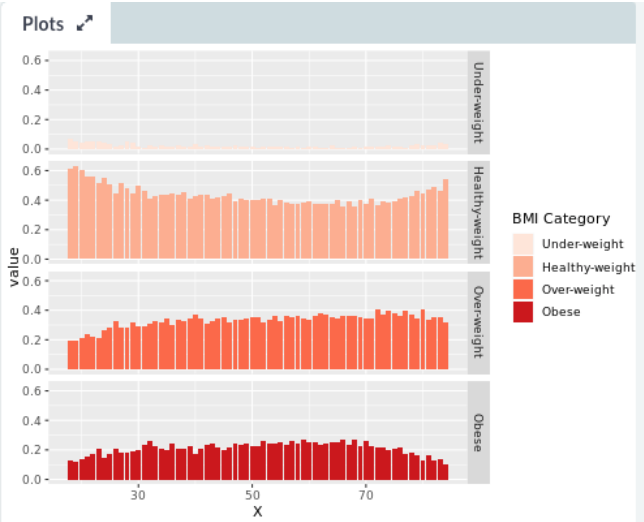

```
# Plot 1 - Count histogram
ggplot(adult, aes (x = SRAGE_P, fill= factor(RBMI))) +
  geom_histogram(binwidth = 1) +
  BMI_fill

# Plot 2 - Density histogram
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
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)) +
geom_histogram(position = "fill") +
  BMI_fill
```



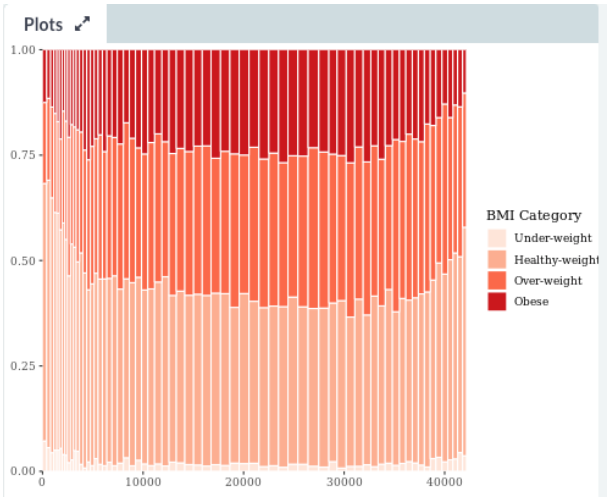

```
# An attempt to facet the accurate frequency histogram from before (failed)
ggplot(adult, aes (x = SRAGE_P, fill= factor(RBMI))) +
  geom_histogram(aes(y = ..count../sum(..count..)), binwidth = 1, position = "fill") +
  BMI_fill +
  facet_grid(RBMI ~ .)

# Create DF with table()
DF <- table(adult$RBMI, adult$SRAGE_P)

# Use apply on DF to get frequency of each group
DF_freq <- apply(DF, 2, function(x) x/sum(x))

# Load reshape2 and use melt on DF to create DF_melted
library(reshape2)
DF_melted <- melt(DF_freq)

# Change names of DF_melted
names(DF_melted) <- c("FILL", "X", "value")

# Add code to make this a faceted plot
ggplot(DF_melted, aes(x = X, y = value, fill = FILL)) +
  geom_col(position = "stack") +
```

```
  BMI_fill +
  facet_grid(FILL ~ .) # Facets
```

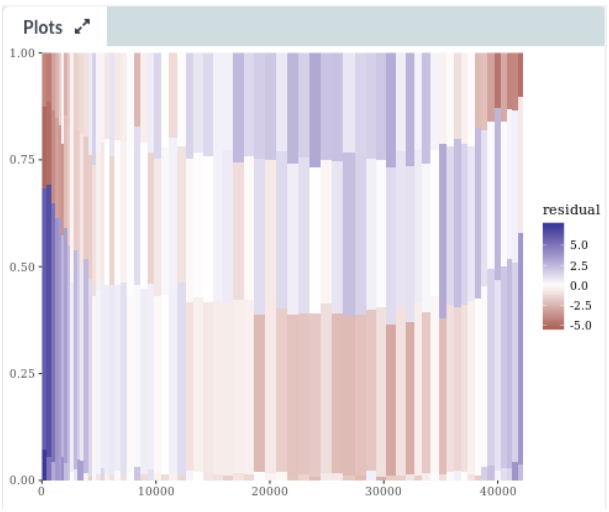



```
# The initial contingency table
DF <- as.data.frame.matrix(table(adult$SRAGE_P, adult$RBMI))

# Create groupSum, xmax and xmin columns
DF$groupSum <- rowSums(DF)
DF$xmax <- cumsum(DF$groupSum)
DF$xmin <- DF$xmax - DF$groupSum

# The groupSum column needs to be removed; don't remove this line
DF$groupSum <- NULL

# Copy row names to variable X
DF$X <- row.names(DF)

# Melt the dataset
library(reshape2)
DF_melted <- melt(DF, id.vars = c("X", "xmin", "xmax"), variable.name = "FILL")

# dplyr call to calculate ymin and ymax - don't change
library(dplyr)
DF_melted <- DF_melted %>%
  group_by(X) %>%
  mutate(ymax = cumsum(value/sum(value)),
```

```
        ymin = ymax - value/sum(value))

# Plot rectangles - don't change
library(ggthemes)
ggplot(DF_melted, aes(ymin = ymin,
                      ymax = ymax,
                      xmin = xmin,
                      xmax = xmax,
                      fill = FILL)) +
 geom_rect(colour = "white") +
  scale_x_continuous(expand = c(0,0)) +
  scale_y_continuous(expand = c(0,0)) +
  BMI_fill +
  theme_tufte()
```



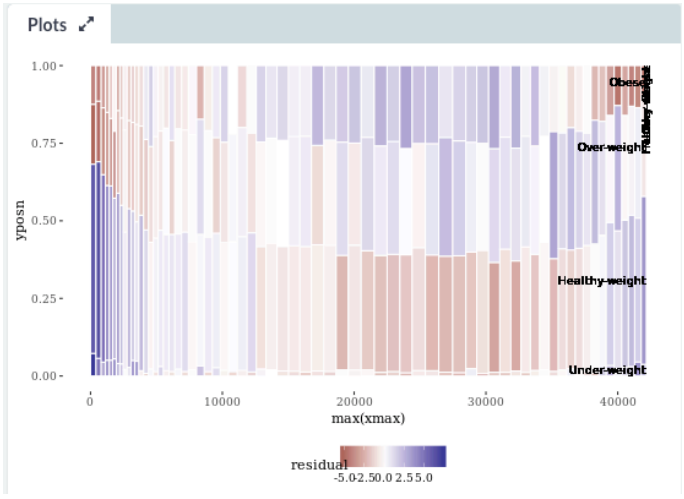

```
# Perform chi.sq test (RBMI and SRAGE_P)
results <- chisq.test(table(adult$RBMI, adult$SRAGE_P))

# Melt results$residuals and store as resid
resid <- melt(results$residuals)

# Change names of resid
names(resid) <- c("FILL", "X", "residual")

# merge the two datasets:
DF_all <- merge(DF_melted, resid)
```




```
# Update plot command
library(ggthemes)
ggplot(DF_all, aes(ymin = ymin,
                   ymax = ymax,
                   xmin = xmin,
                   xmax = xmax,
                   fill = residual)) +
  geom_rect() +
  scale_fill_gradient2() +
  scale_x_continuous(expand = c(0,0)) +
  scale_y_continuous(expand = c(0,0)) +
  theme_tufte()

# Plot so far
p

# Position for labels on y axis (don't change)
index <- DF_all$xmax == max(DF_all$xmax)
DF_all$yposn <- DF_all$ymin[index] + (DF_all$ymax[index] - DF_all$ymin[index])/2

# Plot 1: geom_text for BMI (i.e. the fill axis)
p1 <- p %+% DF_all +
  geom_text(aes(x = max(xmax),
                y = yposn,
                label = FILL),
            size = 3, hjust = 1,
            show.legend  = FALSE)
```

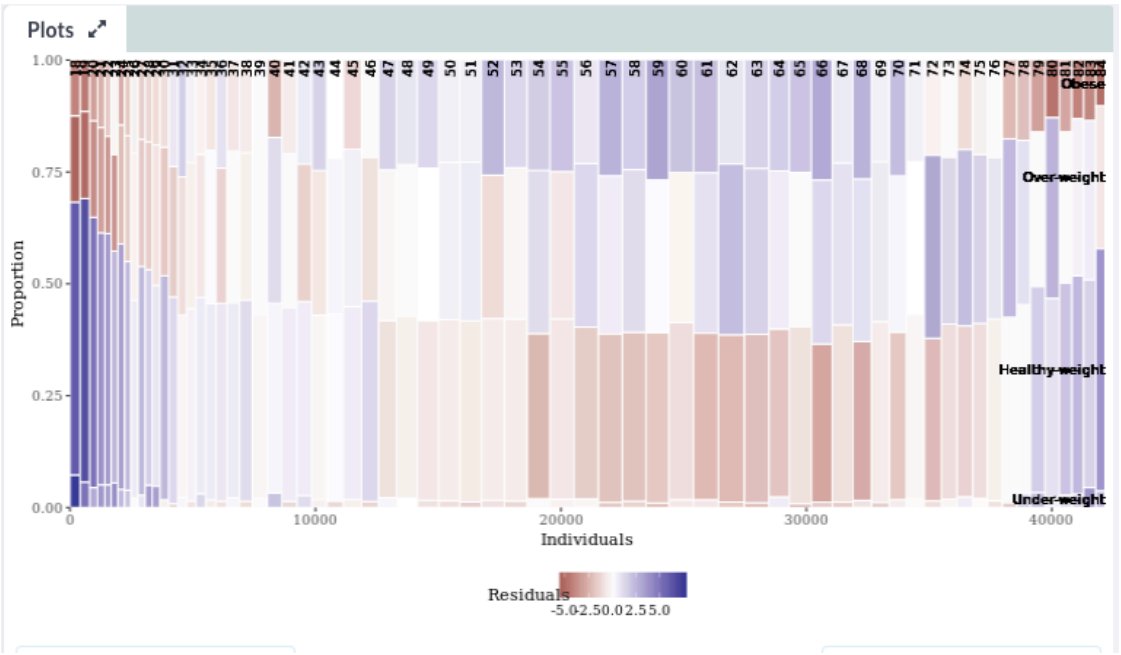
p1

```
# Plot 2: Position for labels on x axis
DF_all$xposn <- DF_all$xmin + (DF_all$xmax - DF_all$xmin)/2

# geom_text for ages (i.e. the x axis)
p1 %+% DF_all +
  geom_text(aes(x = max(xmax), label = FILL),
            y = 1, angle = 90,
            size = 3, hjust = 1,
            show.legend = FALSE)
```

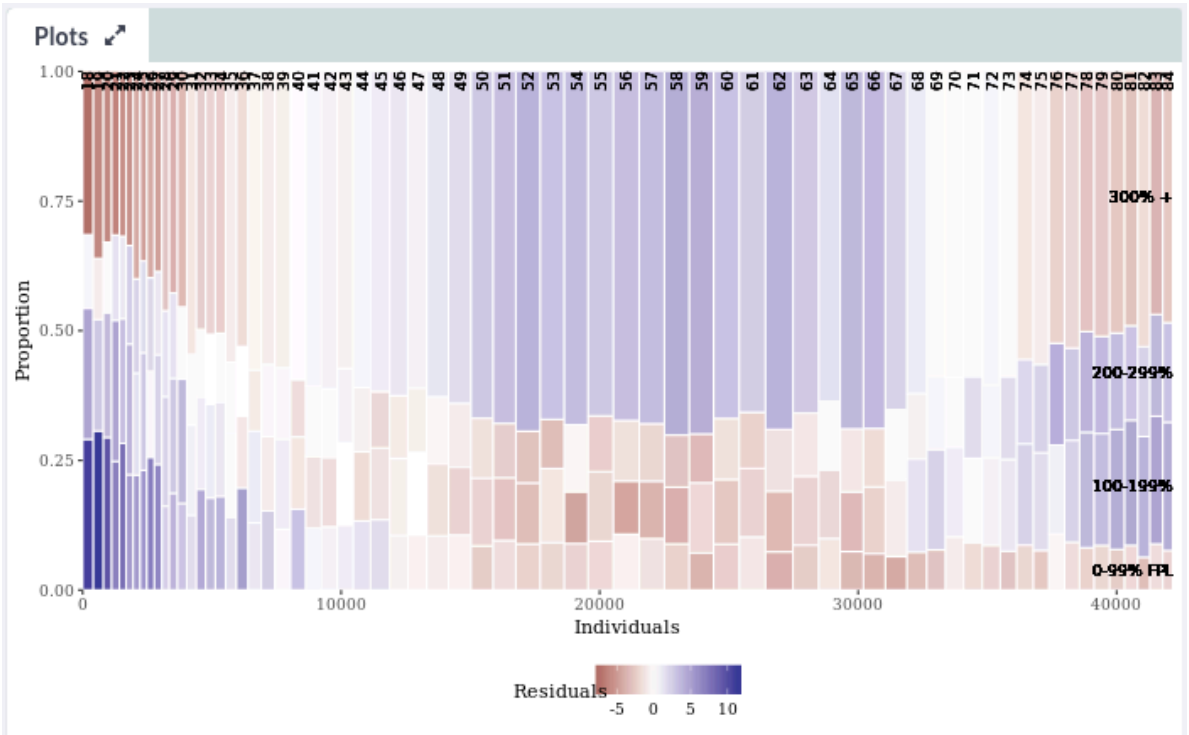



```
# Load all packages
library(ggplot2)
library(reshape2)
library(dplyr)
library(ggthemes)

# Script generalized into a function
mosaicGG
```

# BMI described by age (as previously seen)
mosaicGG(adult, X = "SRAGE_P", FILL = "RBMI")



# Poverty described by age
mosaicGG(adult, X = "SRAGE_P", FILL = "POVLL")



CHIS Exercise: Explanation

According to the final mosaic plot where BMI is described by age, it seems that individuals 47- to 76-years-old who were under-weight or healthy-weight are marked with red tiles that indicate significant negative residuals, where the frequency is less than expected. Those in the same age category who were over-weight or obese are marked with blue tiles that indicate significant positive residuals, where the frequency is greater than expected. The intensity of the color represents the magnitude of the residual.

According to the final mosaic plot where poverty is described by age, it seems that the majority of individuals 46- to 84-years-old who were 0% to 299% Federal Poverty Level are marked with red tiles that indicate significant negative residuals, where the frequency is less than expected. Those in the same age category who were 300%+ Federal Poverty Level are marked with blue tiles that indicate significant positive residuals, where the frequency is greater than expected. The intensity of the color represents the magnitude of the residual.