DATA VISUALIZATION WITH GGPLOT2

California Health Information Survey Descriptive Statistics

Data

- Largest state health survey in US
- Wide variety of variables
- Personal health and economic measurements
- Original dataset

```
> dim(adult)
[1] 47614   536
```

- Reduced dataset

```
> dim(adult)
[1] 44346    10
```

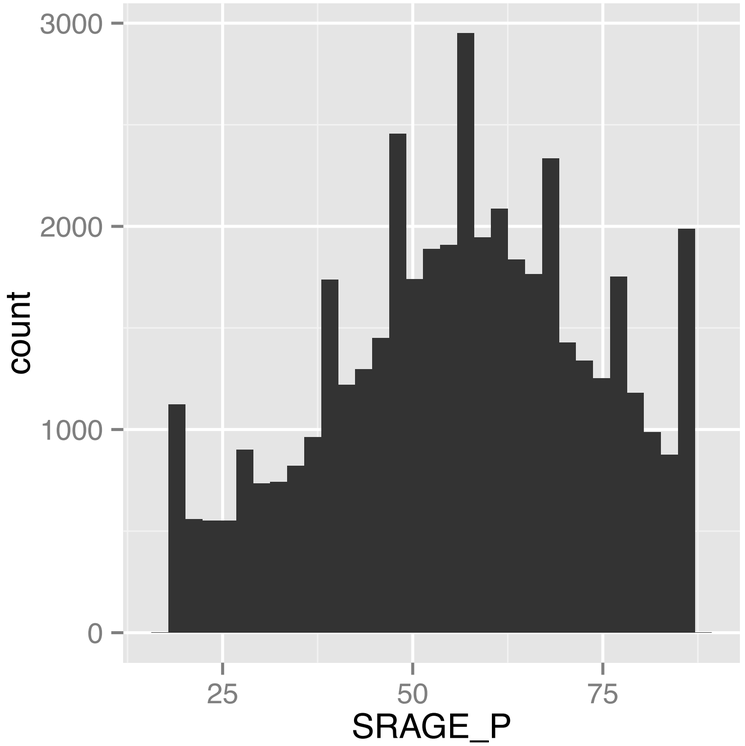
Variables

Variable	Description
RBMI	BMI Category Description
BMI_P	BMI value
RACEHPR2	Race
SRSEX	Sex
SRAGE_P	Age
MARIT2	Marital Status
AB1	General Health Condition
ASTCUR	Current Asthma Status
AB51	Type I or Type II Diabetes
POVLL	Poverty level

http://healthpolicy.ucla.edu/

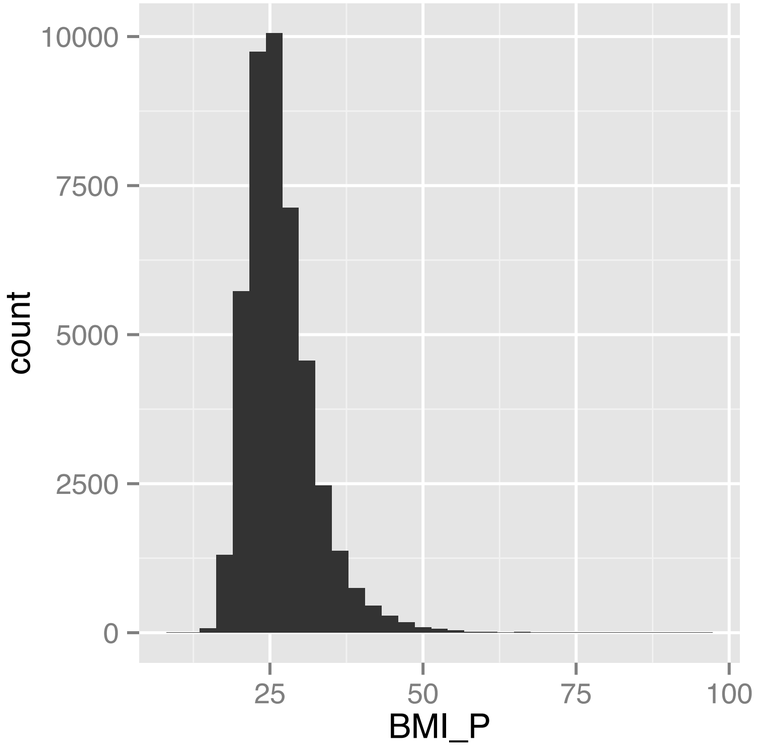
Age

```
> ggplot(adult, aes(x = SRAGE_P)) +
    geom_histogram()
> diff(range(adult$SRAGE_P)) / 30
[1] 2.233333
```



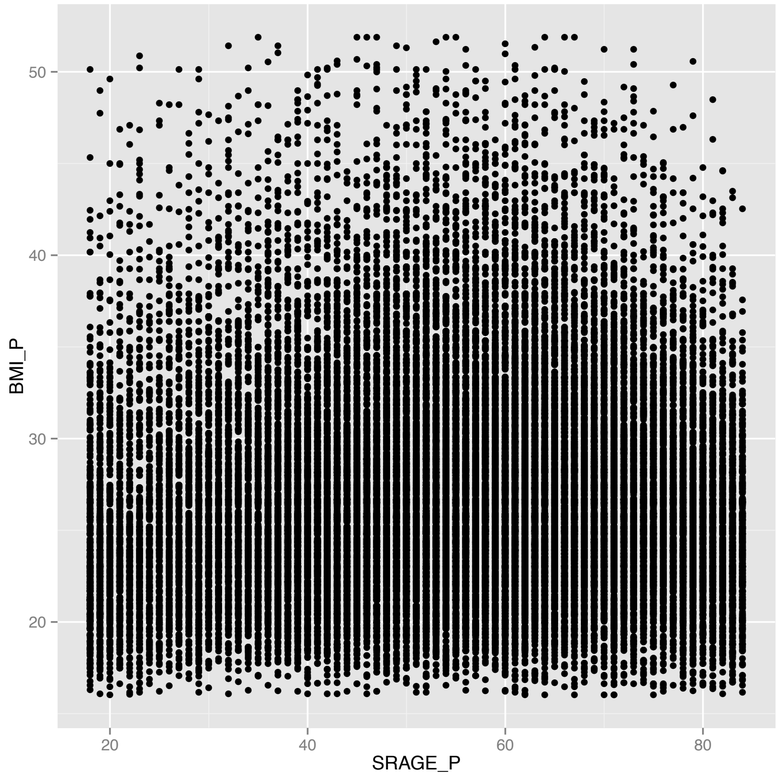
BMI

```
> ggplot(adult, aes(x = BMI_P)) +
    geom_histogram()
```



BMI & Age

```
> # Cleaning code left out
> ggplot(adult, aes(x = SRAGE_P, y = BMI_P)) +
    geom_point()
```



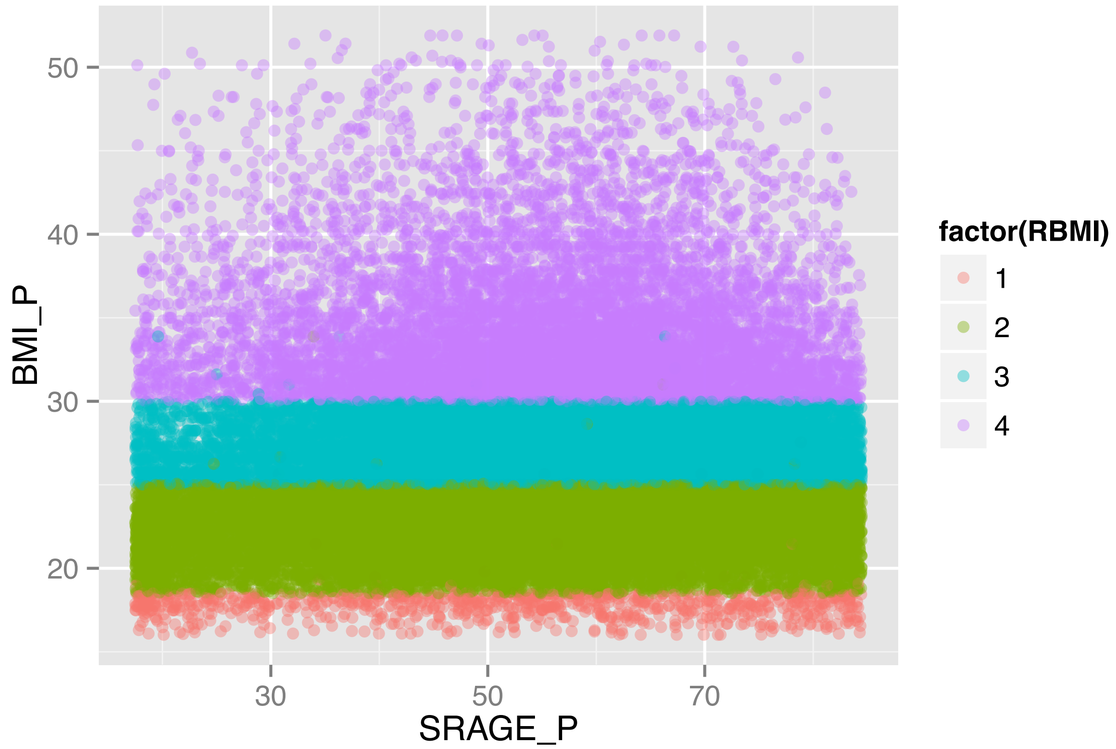
BMI is ordinal

http://healthpolicy.ucla.edu/

Range	Classification
0 - 18.49	Underweight
18.5 - 24.99	Healthy-weight
25.0 - 29.99	Over-weight
30.0+	Obese

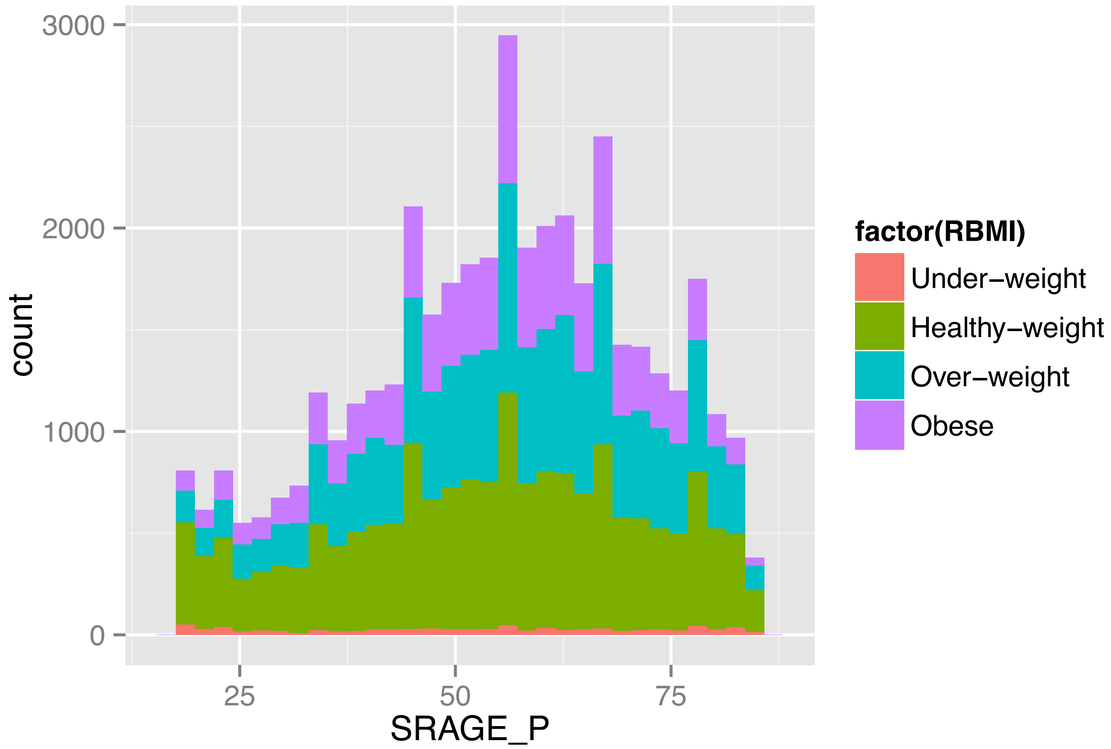
BMI & Age

```
> ggplot(adult, aes(x = SRAGE_P, y = BMI_P, col = factor(RBMI))) +
    geom_point(alpha = 0.4, position = position_jitter(width = 0.5))
```



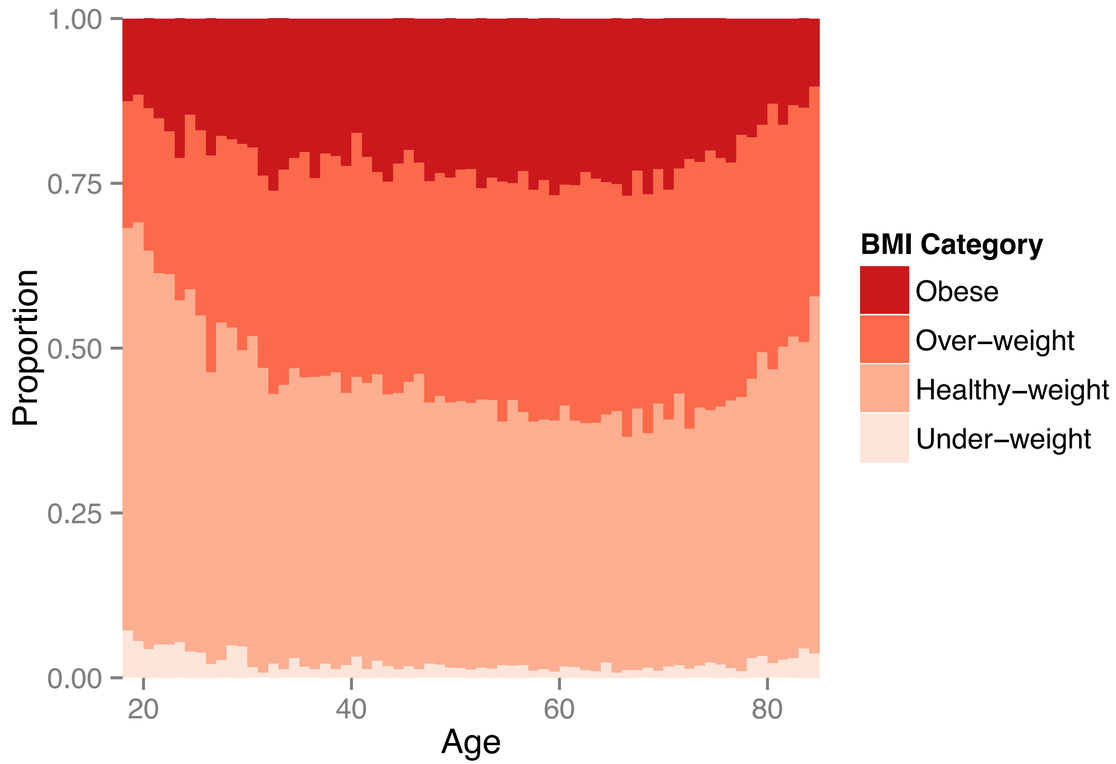
Histogram

```
> ggplot(adult, aes (x = SRAGE_P, fill= factor(RBMI))) +
    geom_histogram()
```



Histogram

```
> ggplot(adult, aes (x = SRAGE_P, fill= factor(RBMI))) +
    geom_histogram(aes(y = ..count../sum(..count..)),
                   binwidth = 1, position = "fill") + ... # left out
```



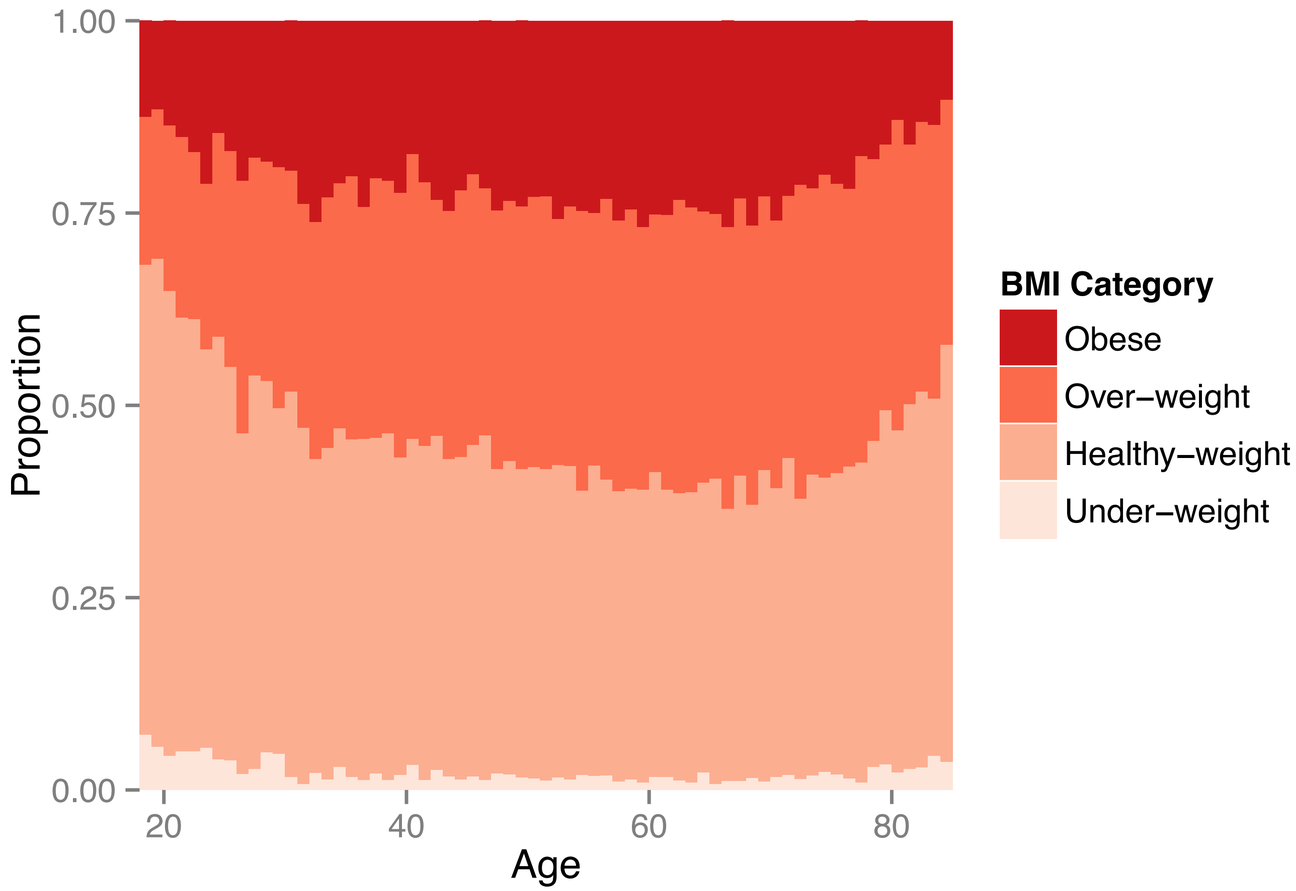
DATA VISUALIZATION WITH GGPLOT2

Let’s practice!

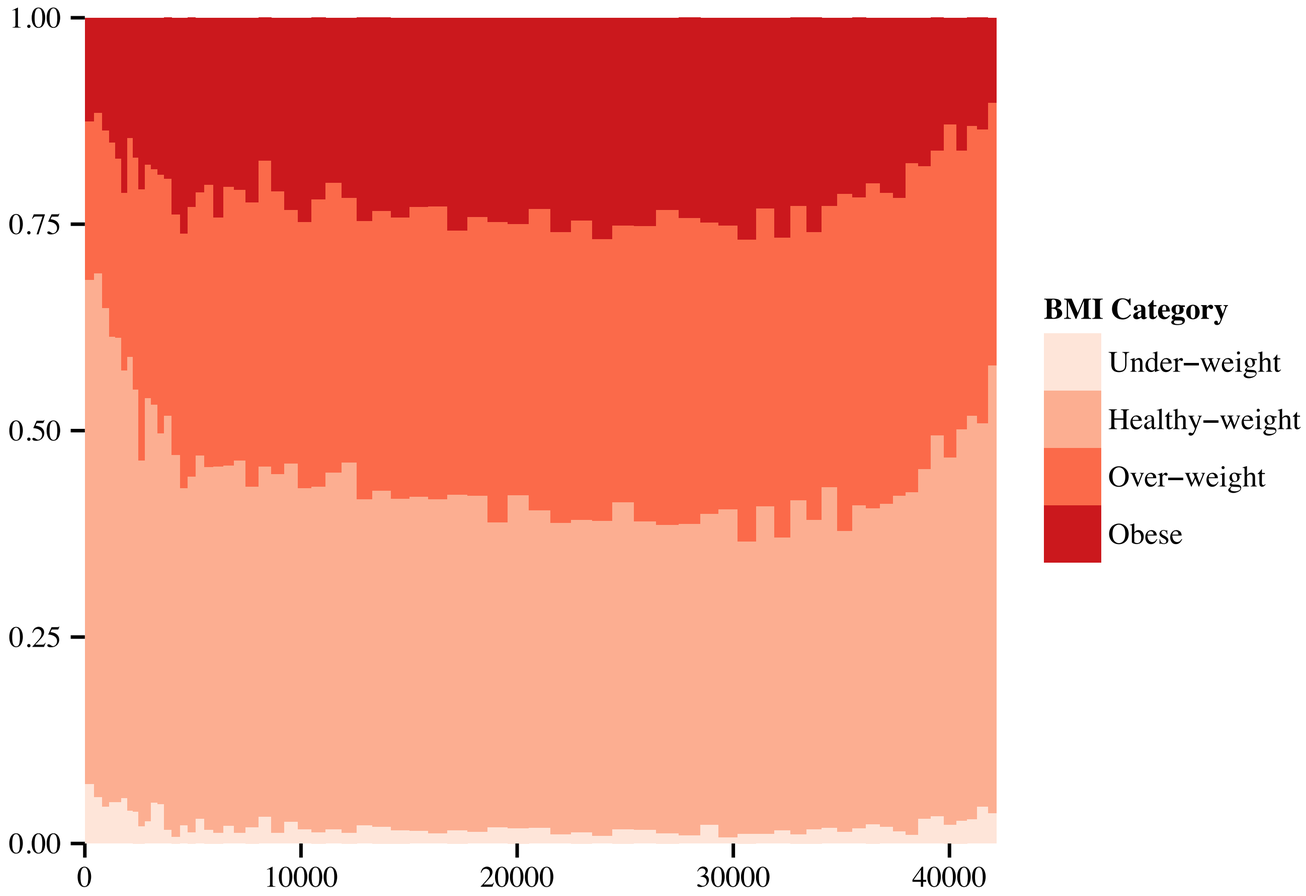
DATA VISUALIZATION WITH GGPLOT2

California Health Information Survey Mosaic Plots

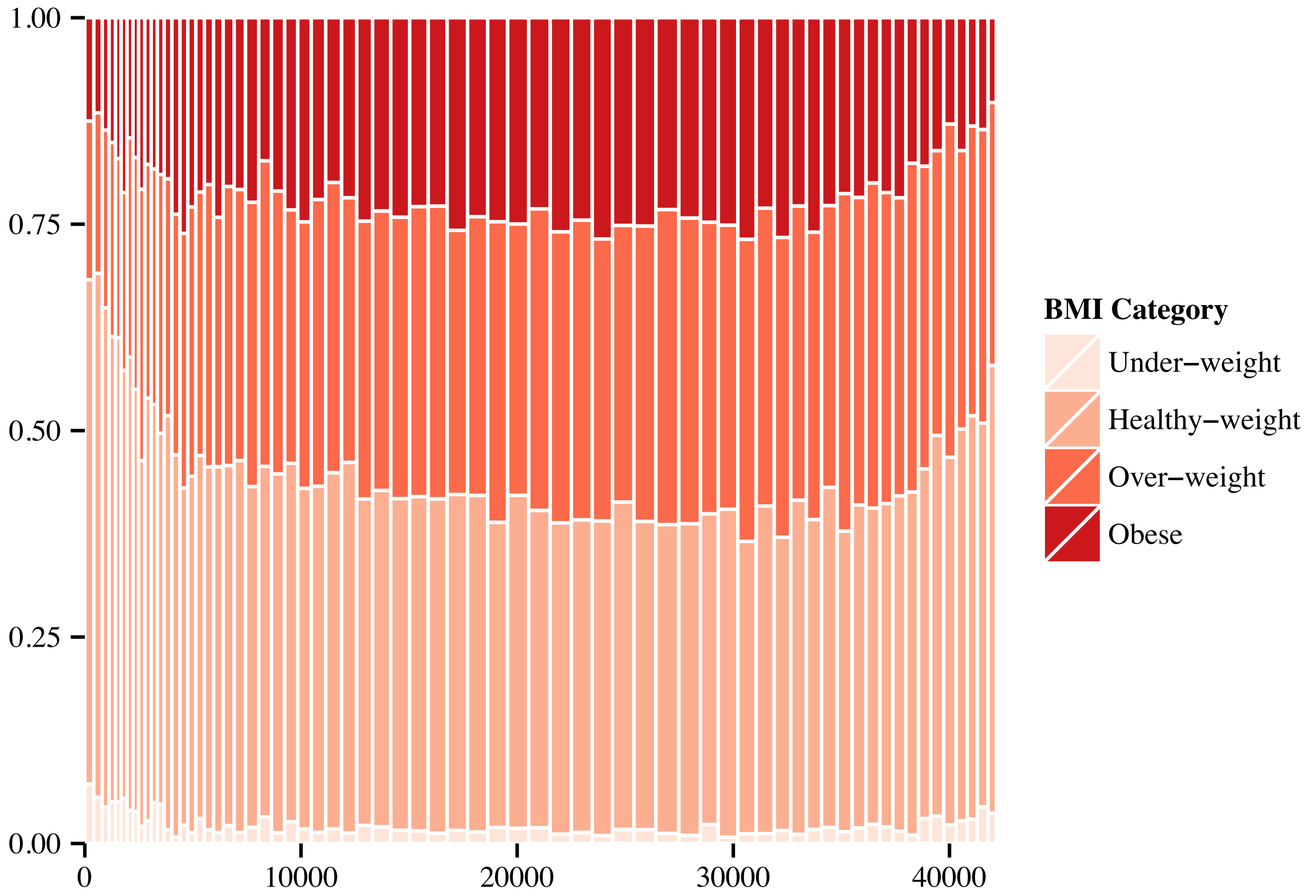
Histogram


1.00
0.75
0.50
0.25
0.00
Proportion
20
40
60
80
Age
BMI Category
Obese
Over−weight
Healthy−weight
Under−weight

Mosaic Plot


1.00
0.75
0.50
0.25
0.00
0
10000
20000
30000
40000
BMI Category
Under−weight
Healthy−weight
Over−weight
Obese

Mosaic Plot


1.00
0.75
0.50
0.25
0.00
0
10000
20000
30000
40000
BMI Category
Under−weight
Healthy−weight
Over−weight
Obese

Chi-Squared Test

observations

```
> cont_table
       party
gender Democrat Independent Republican
     F      762         327        468
     M      484         239        477
```

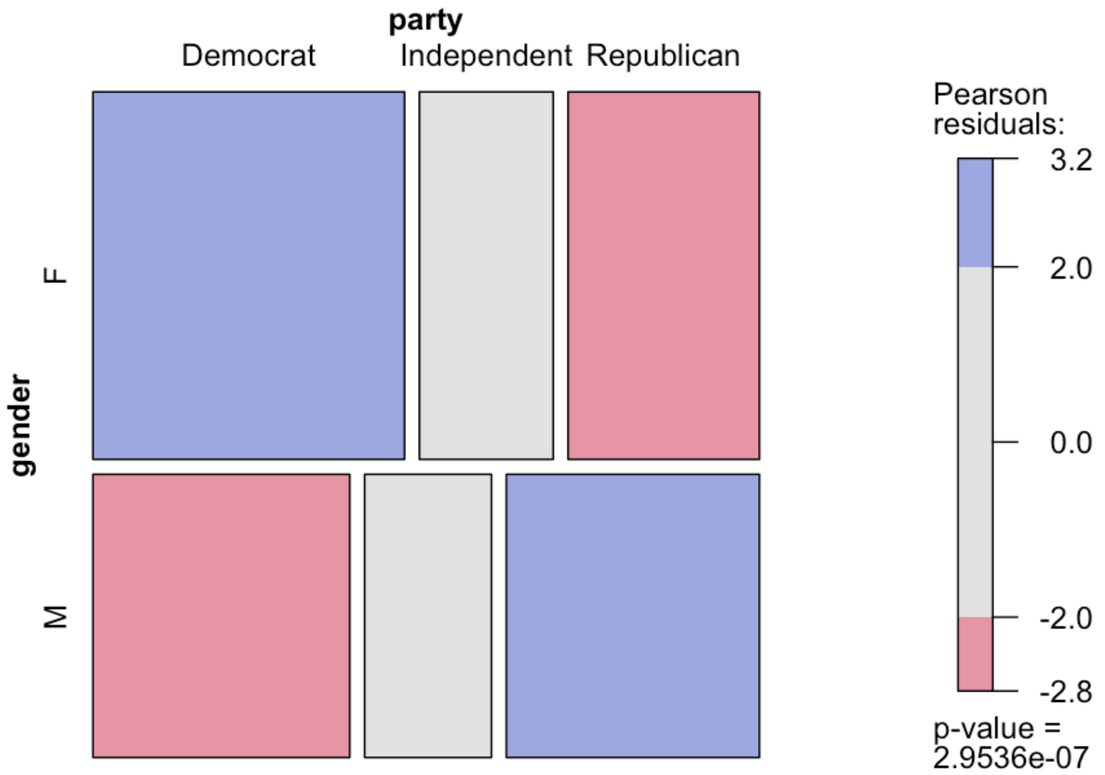
expected

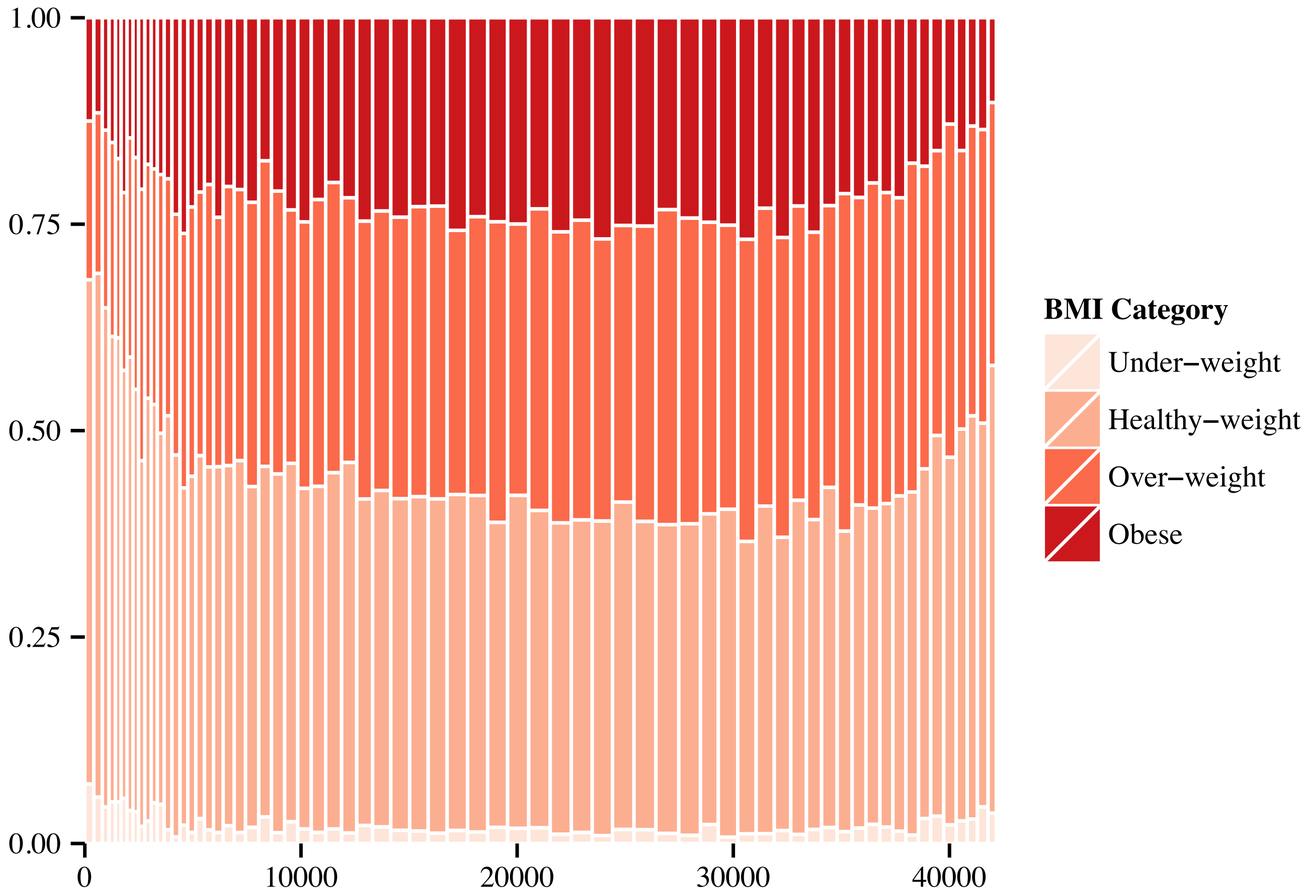
```
> chisq.test(cont_table)$expected
       party
gender Democrat Independent Republican
     F 703.6714    319.6453   533.6834
     M 542.3286    246.3547   411.3166
```

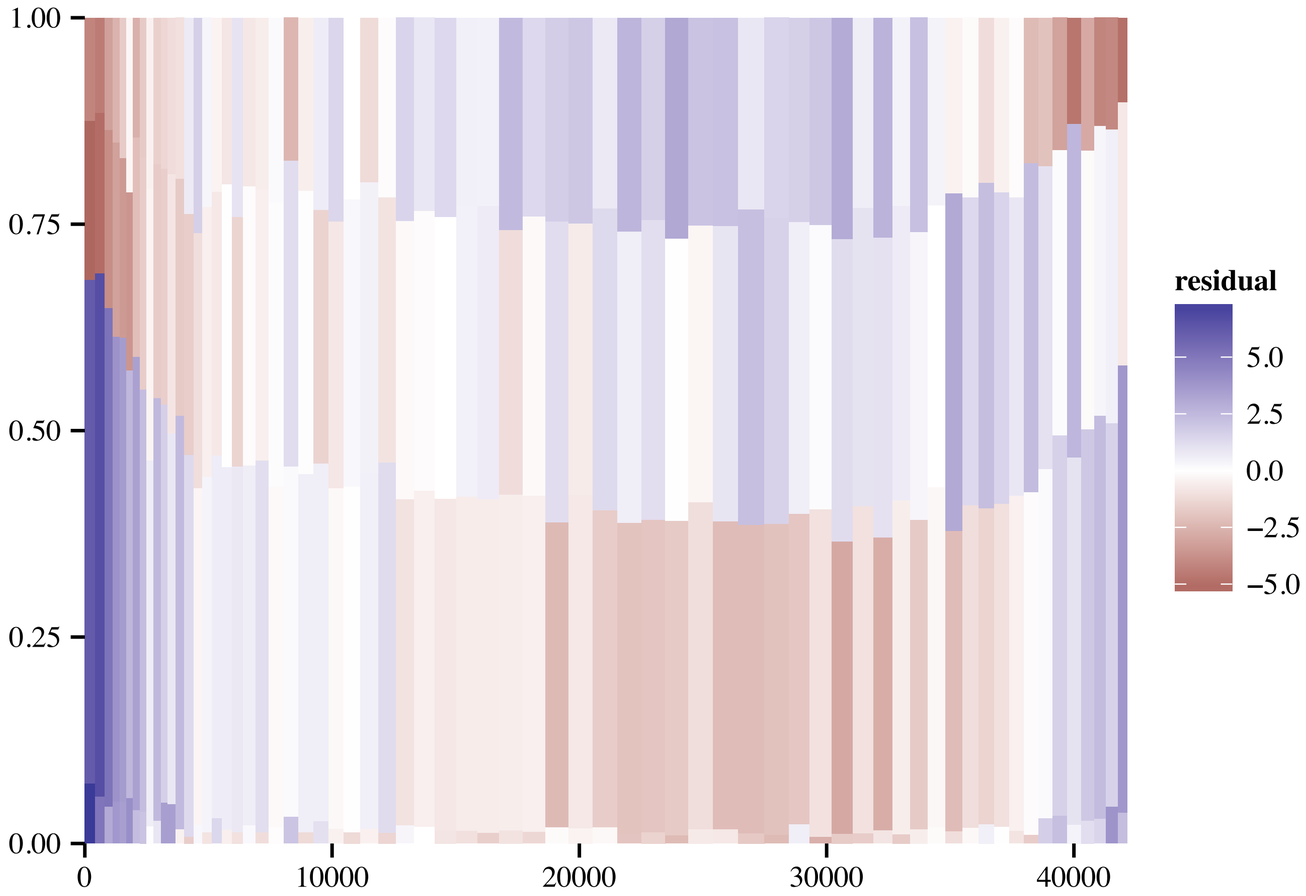
residuals

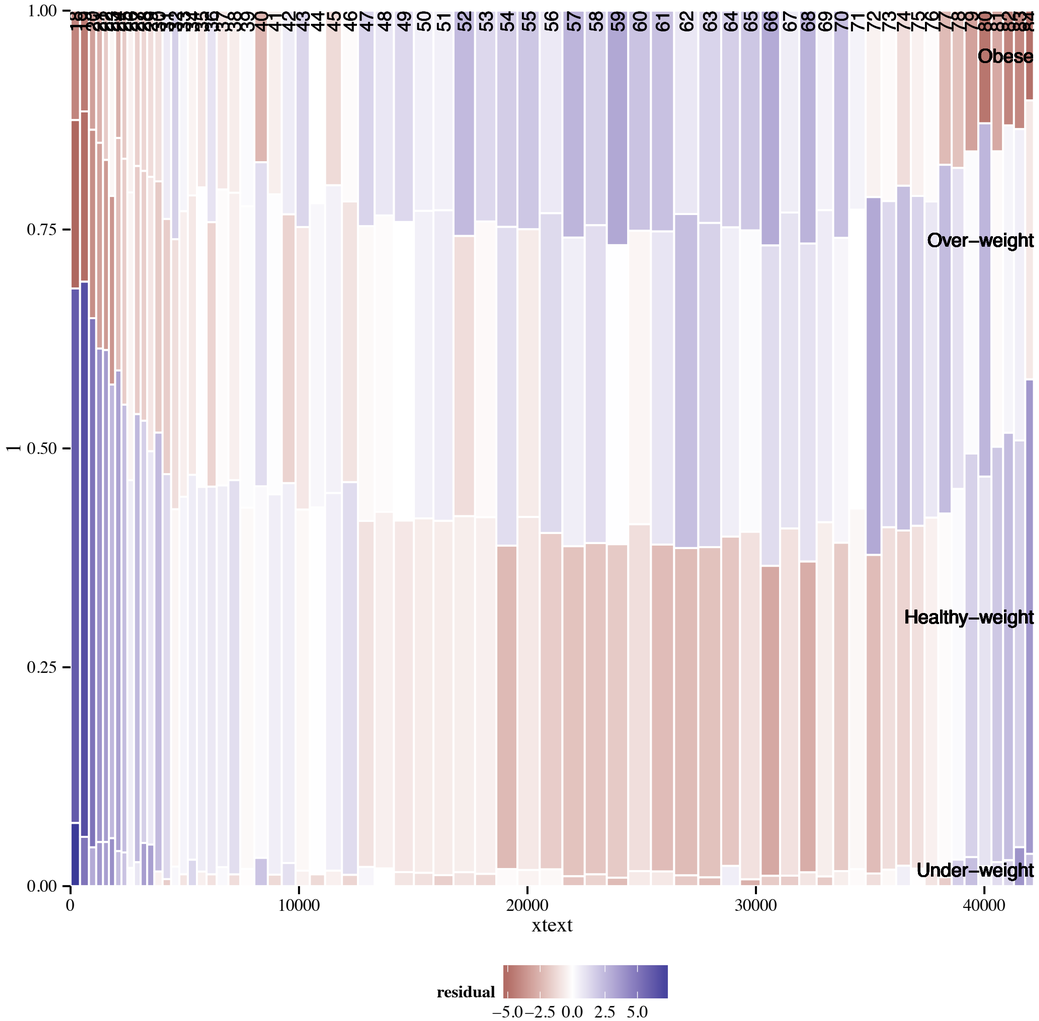
```
> chisq.test(cont_table)$residuals
       party
gender    Democrat Independent Republican
     F  2.1988558   0.4113702 -2.8432397
     M -2.5046695  -0.4685829  3.2386734
```

```
> library(vcd)
> mosaic(M, shade = TRUE, color = TRUE)
```

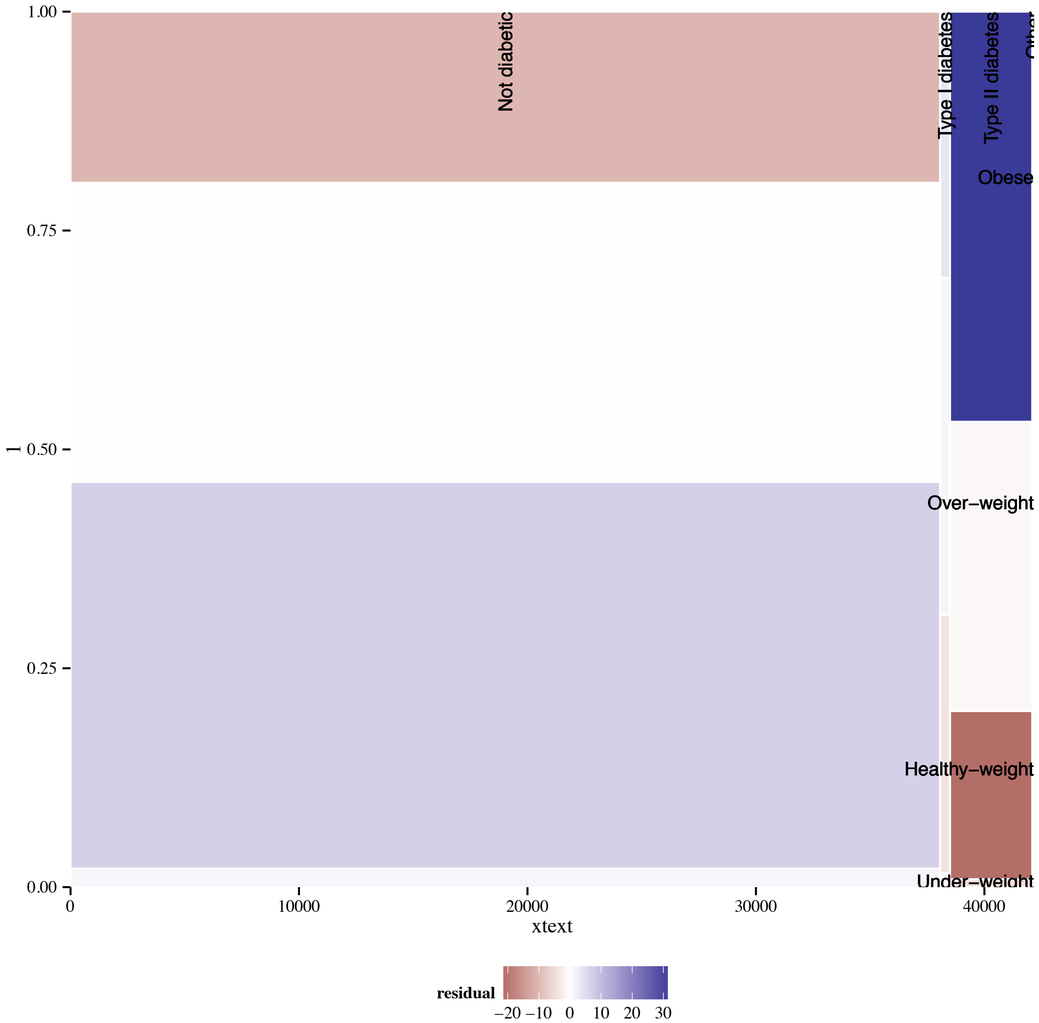



1.00
0.75
0.50
0.25
0.00
0
10000
20000
30000
40000
BMI Category
Under−weight
Healthy−weight
Over−weight
Obese

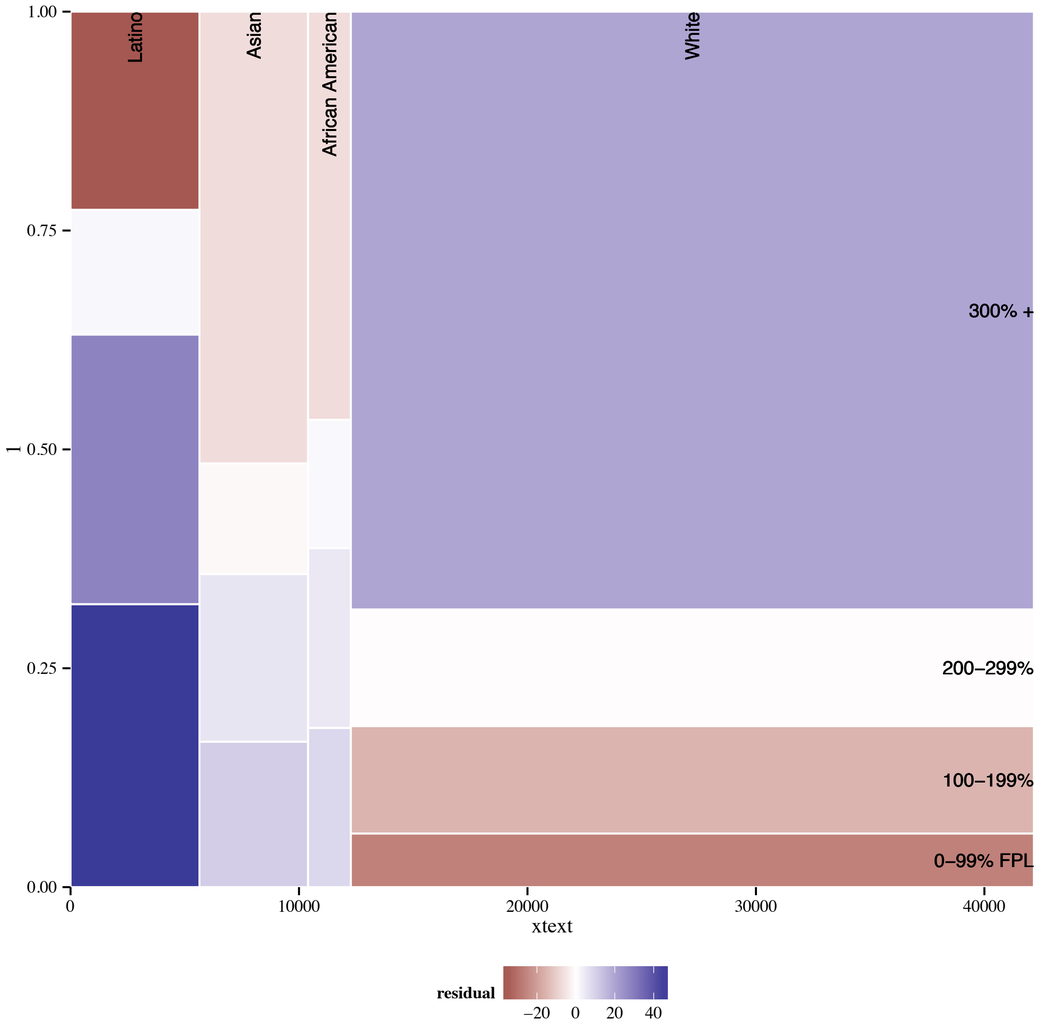

1.00
0.75
0.50
0.25
0.00
0
10000
20000
30000
40000
residual
5.0
2.5
0.0
−2.5
−5.0


1.00
0.75
0.50
0.25
0.00
1
0
10000
20000
30000
40000
xtext
Obese
Over−weight
Healthy−weight
Under−weight
residual
−5.0 −2.5 0.0 2.5 5.0

```
> mosaicGG(adult, "AB51", "RBMI")
```



```
> mosaicGG(adult, "RACEHPR2", "POVLL")
```



Closing Notes

- Concepts and Practice
- Essential skill
- Consider purpose & audience
- Course 3:
  - Advanced statistical plots
  - ggplot2 internals
  - accessory packages

DATA VISUALIZATION WITH GGPLOT2

Let’s practice!