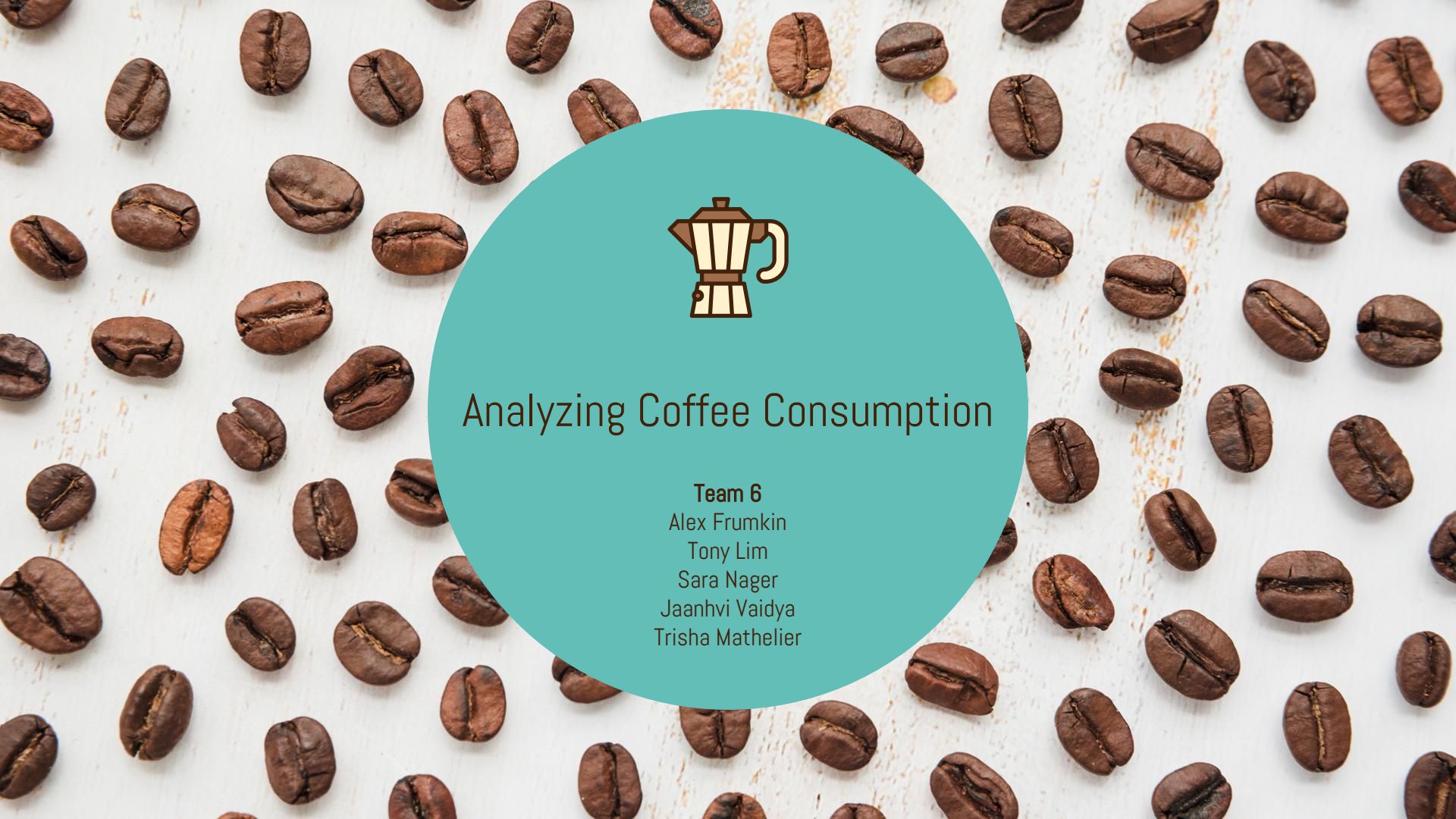


# Analyzing Coffee Consumption

**Team 6**
Alex Frumkin
Tony Lim
Sara Nager
Jaanhvi Vaidya
Trisha Mathelier

# Which variables can predict coffee consumption?

Our team performed a multiple linear regression on data from the National Health and Nutrition Examination Survey (NHANES) 2015-2016.

**Gender**, **age**, **race**, **education**, and **household size** are variables that can be used to predict coffee consumption

# Raw Data Overview

## DEMO_I.XPT

Demographic dataset that includes

- Individual
- Family, and
- Household-level information

## DR1IFF_I.XPT

Dietary intake information used to estimate

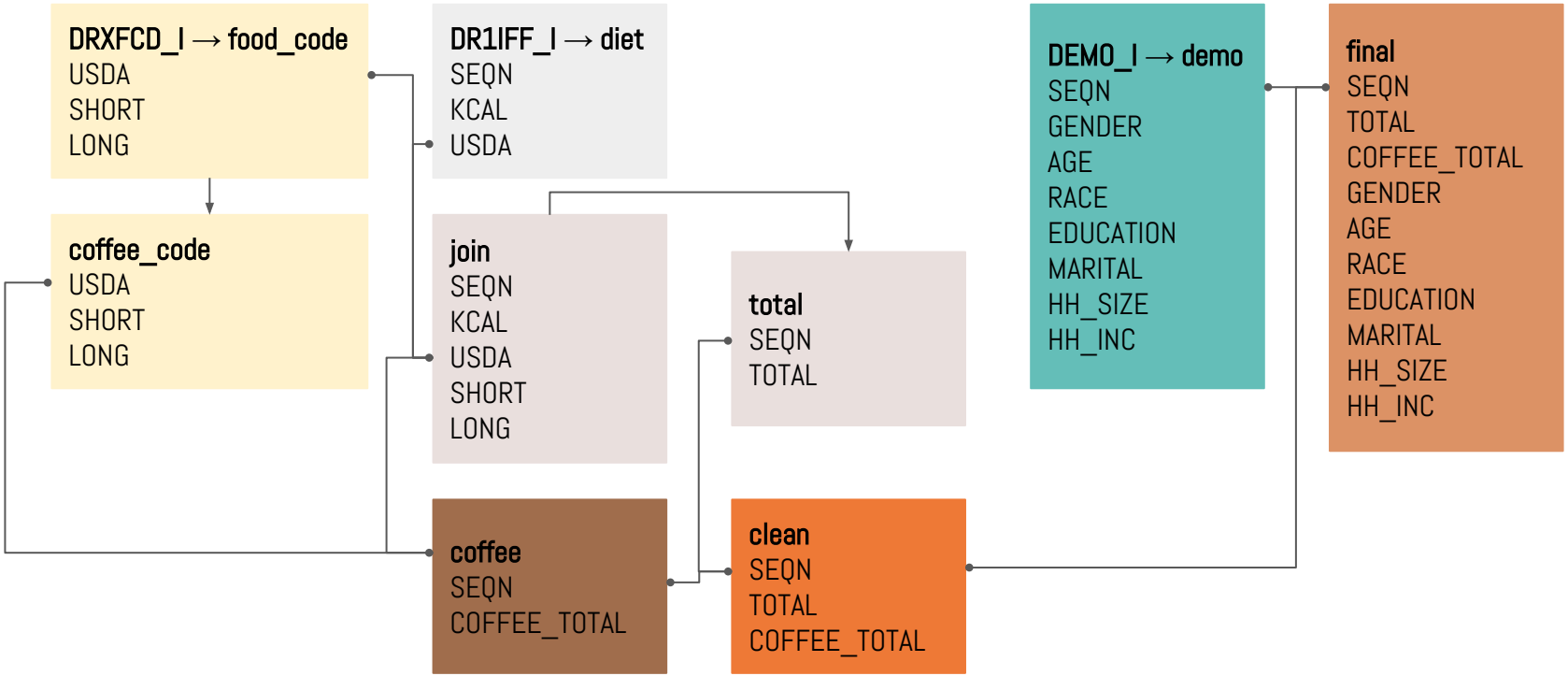
- Types and amounts of foods and beverages consumed
- Intakes of energy, nutrients, and other food components from those foods and beverages

## DRXFCD_I.XPT

Supporting file to add food code descriptions to DR1IFF_I and includes

- USDA food codes
- Abbreviated descriptions (up to 60 characters)
- Complete descriptions (up to 200 characters)

# Methodology Summary



# Final Dataset

| SEQN <dbl> | TOTAL <dbl> | COFFEE_TOTAL <dbl> | GENDER <fctr> | AGE <fctr> | RACE <fctr> | EDUCATION <fctr> | MARITAL <fctr> | HH_SIZE <dbl> | HH_INC <fctr> |
|---|---|---|---|---|---|---|---|---|---|
| 83732 | 1781 | 723.92 | Male | 55-65 | White | College graduate or above | Married | 2 | 45-75k |
| 83733 | 2964 | 480.00 | Male | 45-55 | White | High school graduate/GED | Divorced | 1 | Below 20k |
| 83734 | 2482 | 776.75 | Male | Above 75 | White | High school graduate/GED | Married | 2 | 20-45k |
| 83735 | 1340 | 480.00 | Female | 55-65 | White | College graduate or above | Living with partner | 1 | 45-75k |
| 83736 | 604 | 0.00 | Female | 35-45 | Black | Some college/AA | Divorced | 5 | 20-45k |
| 83737 | 1304 | 0.00 | Female | 65-75 | Hispanic | Some high school | Separated | 5 | Above 75k |
| 83738 | 1239 | 0.00 | Female | Below 15 | Hispanic | Less than high school | *NA* | 5 | 20-45k |
| 83739 | 1242 | 0.00 | Male | Below 15 | White | *NA* | *NA* | 5 | Above 75k |
| 83740 | 1151 | 0.00 | Male | Below 15 | Hispanic | *NA* | *NA* | 7 | Refused |
| 83741 | 2338 | 0.00 | Male | 15-25 | Black | Some college/AA | Never married | 3 | 20-45k |

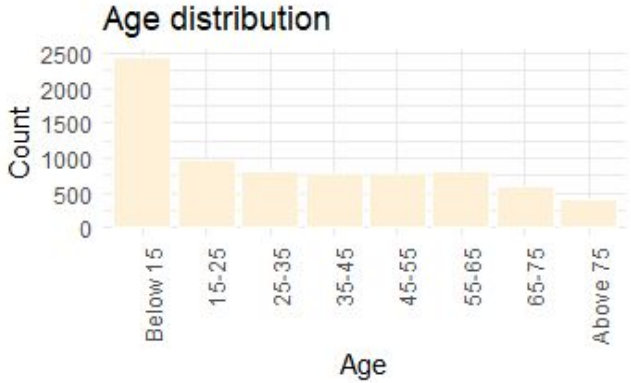
1-10 of 7,923 rows

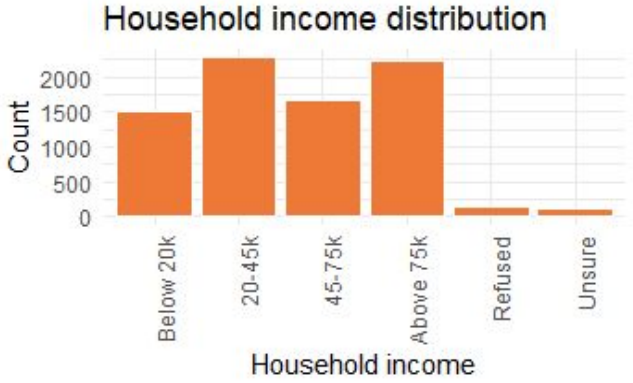
Previous 1 2 3 4 5 6 ... 100 Next

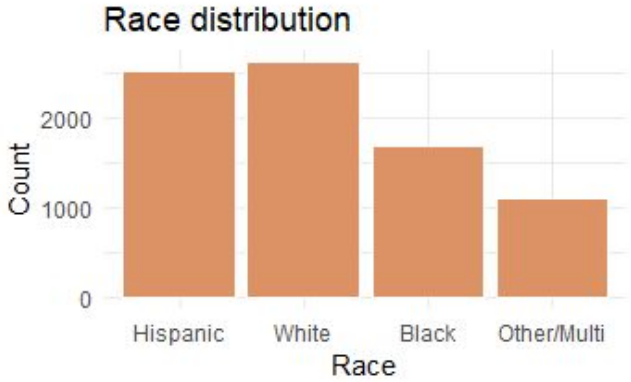
# Variables

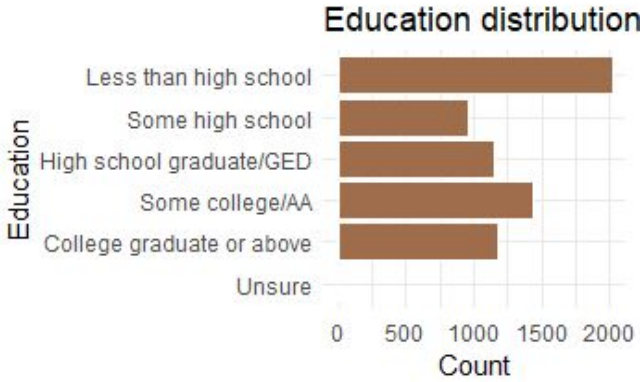
- **SEQN**: Respondent sequence number
- **TOTAL**: Total number of calories consumed that day
- **COFFEE_TOTAL**: Total number of grams consumed from coffee-related products
- **GENDER**: Gender
- **AGE**: Age in years at screening
  - Below 15
  - 15-25
  - 25-35
  - 35-45
  - 45-55
  - 55-65
  - 65-75
  - Above 75
- **RACE**: Race
  - Black
  - Hispanic
  - White
  - Other/Multi
- **EDUCATION**: Education level
  - Less than high school
  - Some high school
  - High school graduate/GED
  - Some college/AA
  - College graduate or above
  - Unsure
- **MARITAL**: Marital status
  - Married
  - Widowed
  - Divorced
  - Separated
  - Never married
  - Living with partner
  - Refused
  - Unsure
- **HH_SIZE**: Total number of people in the household
- **HH_INC**: Total household income
  - Below 20k
  - 20-45k
  - 45-75k
  - Above 75k
  - Refused
  - Unsure

# Data Exploration

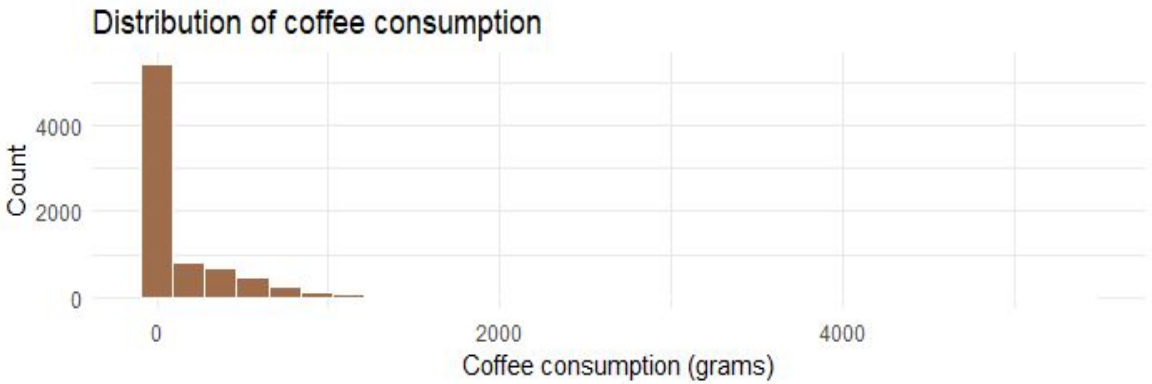


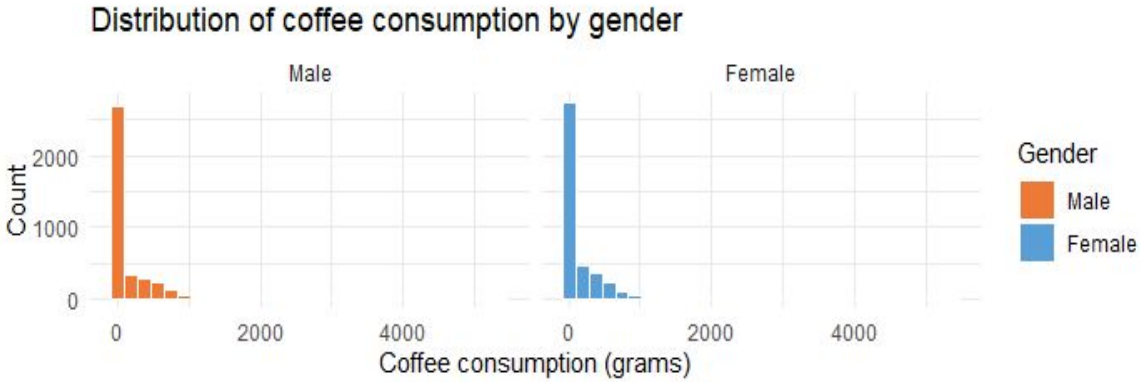




- Many more samples in the **Below 15** age group and **Less than high school** education group
- Fewer individuals identifying as **Black** and **Other/Multi**




## Data Summary

- Heavily skewed to the right
- Median = 0 g
- Mean = ~161 g
  - **Men** = ~172 g
  - **Women** = ~150 g

## Grams to Fl. Oz. Conversion (Starbucks)

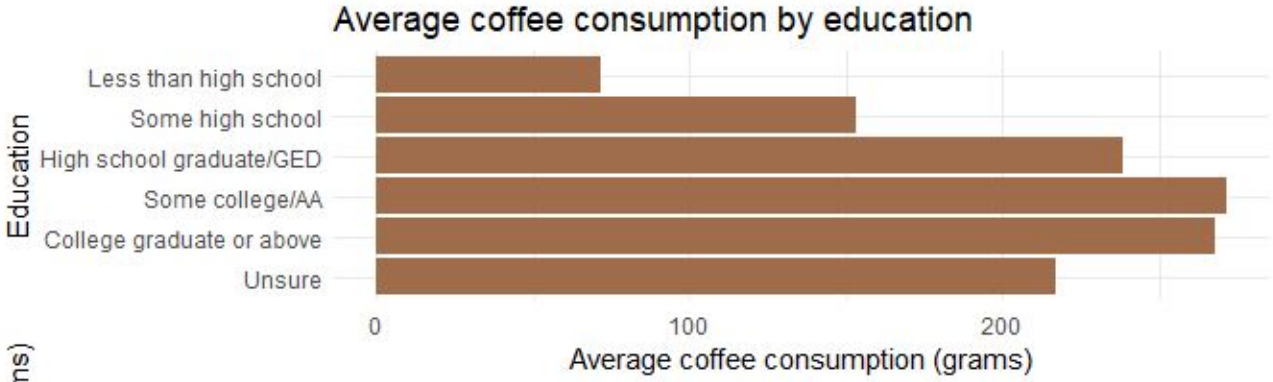
- Short (8 fl. oz. ≈ 237 g)
- Tall (12 fl. oz. ≈ 355 g)
- Grande (16 fl. oz. ≈ 473 g)
- Venti® Hot (20 fl. oz. ≈ 591 g)
- Venti® Cold (24 fl. oz. ≈ 710 g)
- Trenta® Cold (31 fl. oz. ≈ 917 g)

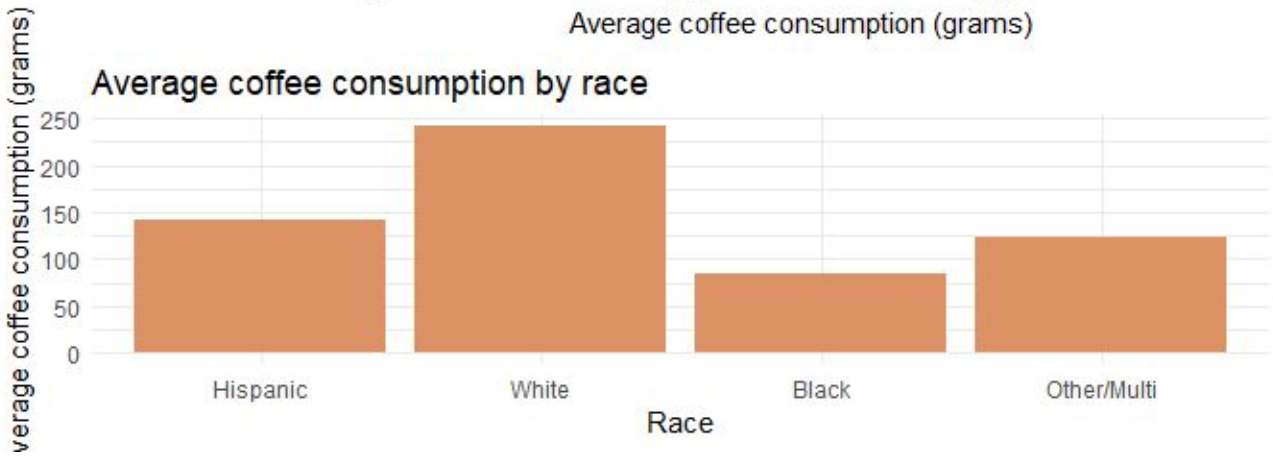
# Coffee by Gender

```
final %>%
  t.test(COFFEE_TOTAL ~ GENDER, data = ., var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  COFFEE_TOTAL by GENDER
## t = 2.7773, df = 7921, p-value = 0.005495
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.255004 36.270375
## sample estimates:
##   mean in group Male mean in group Female
##             171.5573             150.2946
```

Statistically significant that men consume **more** coffee than women




- Average coffee consumption increases with higher education levels
- Highest coffee consumption for **white** individuals

# Multiple Linear Regression

```
Call:
glm(formula = COFFEE_TOTAL ~ GENDER + AGE + RACE + EDUCATION +
    MARITAL + HH_SIZE + HH_INC, data = final)

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       156.2062    37.5421   4.161 3.23e-05 ***
GENDERFemale                      -41.7051    11.7013  -3.564 0.000369 ***
AGE25-35                           43.1737    24.0667   1.794 0.072891 .
AGE35-45                           96.7948    25.1835   3.844 0.000123 ***
AGE45-55                          181.5582    25.7628   7.047 2.09e-12 ***
AGE55-65                          215.4746    26.2969   8.194 3.24e-16 ***
AGE65-75                          158.9035    28.3880   5.598 2.30e-08 ***
AGEAbove 75                       141.1928    31.8752   4.430 9.66e-06 ***
RACEWhite                         124.9840    15.8132   7.904 3.36e-15 ***
RACEBlack                        -105.9548    17.1879  -6.165 7.67e-10 ***
RACEOther/Multi                   -26.9213    19.7035  -1.366 0.171905
EDUCATIONSome high school            5.8851   24.8704   0.237 0.812953
EDUCATIONHigh school graduate/GED    3.7126   22.5010   0.165 0.868955
EDUCATIONSome college/AA            25.1751   22.3026   1.129 0.259041
EDUCATIONCollege graduate or above  -0.5094   24.2029  -0.021 0.983209
EDUCATIONUnsure                     54.3476  276.9407   0.196 0.844429
MARITALWidowed                      11.7491   25.9438   0.453 0.650666
MARITALDivorced                     10.2728   20.3127   0.506 0.613069
MARITALSeparated                    -3.6636   32.6995  -0.112 0.910798
MARITALNever married               -42.8972   18.4990  -2.319 0.020445 *
MARITALLiving with partner          21.9353   21.2939   1.030 0.303006
MARITALRefused                     433.8209  390.5636   1.111 0.266731
HH_SIZE                             -9.3475    3.9879  -2.344 0.019124 *
HH_INC20-45k                        13.4352   17.4574   0.770 0.441579
HH_INC45-75k                        19.5981   19.4588   1.007 0.313911
HH_INCAbove 75k                     25.6866   19.7809   1.299 0.194162
HH_INCRefused                      -31.0664   45.0891  -0.689 0.490858
HH_INCUnsure                        -5.6526   46.1937  -0.122 0.902614
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Marital status** and **household income** don't appear to be great predictor variables

# Multiple Linear Regression, cont.

```
Call:
glm(formula = COFFEE_TOTAL ~ GENDER + AGE + RACE + EDUCATION +
    HH SIZE, data = final)
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                         51.363     18.037   2.848 0.004417 **
GENDERFemale                       -27.080      8.071  -3.355 0.000798 ***
AGE15-25                            44.973     19.032   2.363 0.018152 *
AGE25-35                           130.290     19.494   6.684 2.52e-11 ***
AGE35-45                           187.405     19.239   9.741  < 2e-16 ***
AGE45-55                           276.838     19.266  14.369  < 2e-16 ***
AGE55-65                           309.678     19.321  16.028  < 2e-16 ***
AGE65-75                           262.423     20.343  12.900  < 2e-16 ***
AGEAbove 75                        254.431     22.056  11.536  < 2e-16 ***
RACEWhite                           87.669     10.673   8.214 2.55e-16 ***
RACEBlack                          -83.519     11.595  -7.203 6.53e-13 ***
RACEOther/Multi                    -24.989     13.342  -1.873 0.061120 .
EDUCATIONSome high school            2.034     16.982   0.120 0.904663
EDUCATIONHigh school graduate/GED   16.308     17.195   0.948 0.342965
EDUCATIONSome college/AA            44.773     16.800   2.665 0.007717 **
EDUCATIONCollege graduate or above  24.023     17.706   1.357 0.174910
EDUCATIONUnsure                     38.420    190.966   0.201 0.840559
HH_SIZE                             -6.754      2.732  -2.472 0.013451 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Women** with negative coefficient
- Coefficient generally increases with age with the peak at age **55-65** and decreases after
- **White** individuals with most positive coefficient; **black** individuals with most negative coefficient
- Increasing coffee consumption as education increases
    - Highest coefficient for **Some college/AA**
- Decreasing coffee consumption with increasing **household size**

# Prediction

```
tony <- data.frame(
  GENDER = "Male",
  AGE = "15-25",
  RACE = "Other/Multi",
  EDUCATION = "College graduate or above",
  HH_SIZE = 1
  )
```

```
predict(reg2, newdata = tony)
```

```
##        1
## 88.61694
```

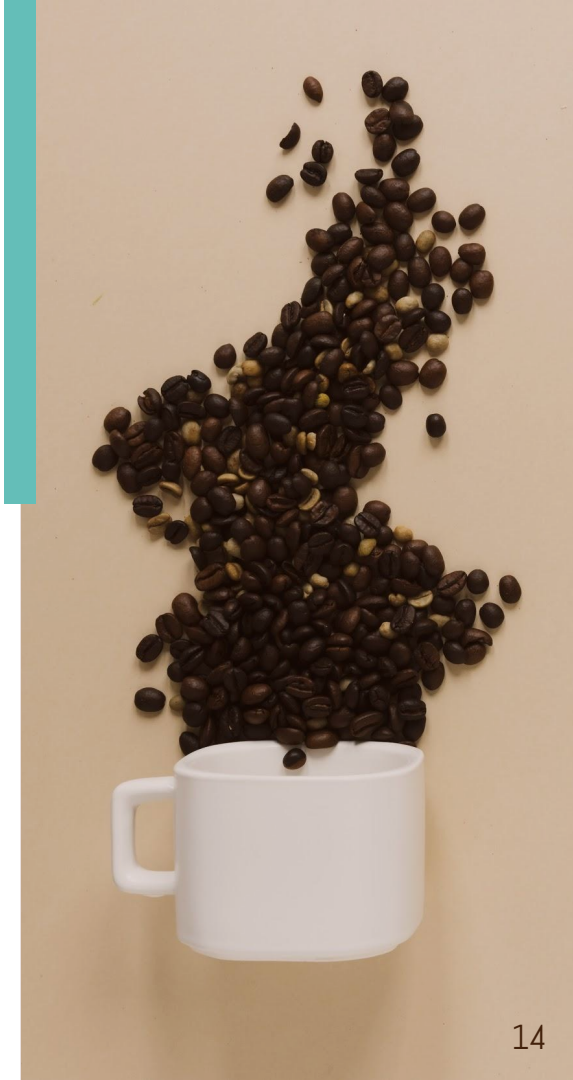
From our regression, it's predicted Tony would consume

**~89 g (~3 fl. oz.)**

of coffee.

# Conclusion

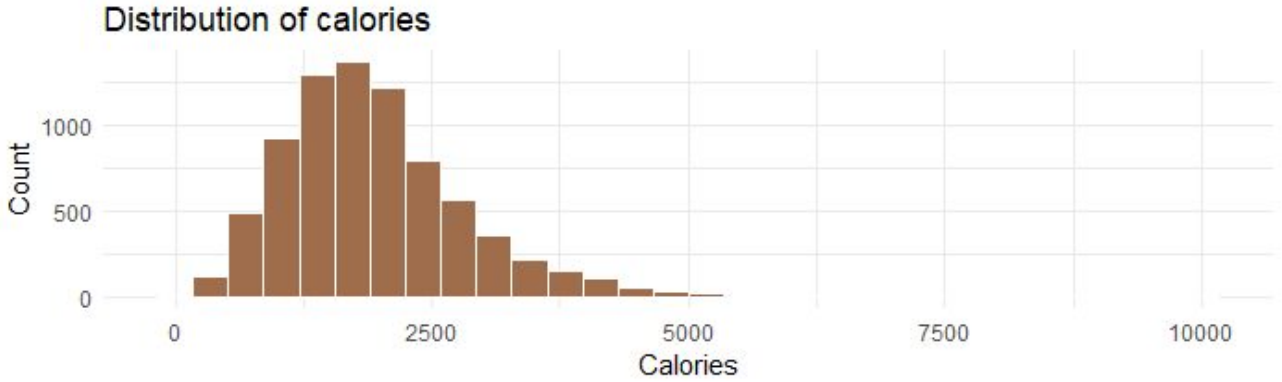
- On average, men consume **more** coffee compared to women
- Gender, age, race, education, and household size are variables that can be used to predict coffee consumption
    - **Middle-aged** or those with **some college/AA** education background predicted to consume more coffee compared to counterparts
    - Groups such as **women** and **black** individuals predicted to consume less coffee
    - Increasing **household size** also predicts lower coffee consumption

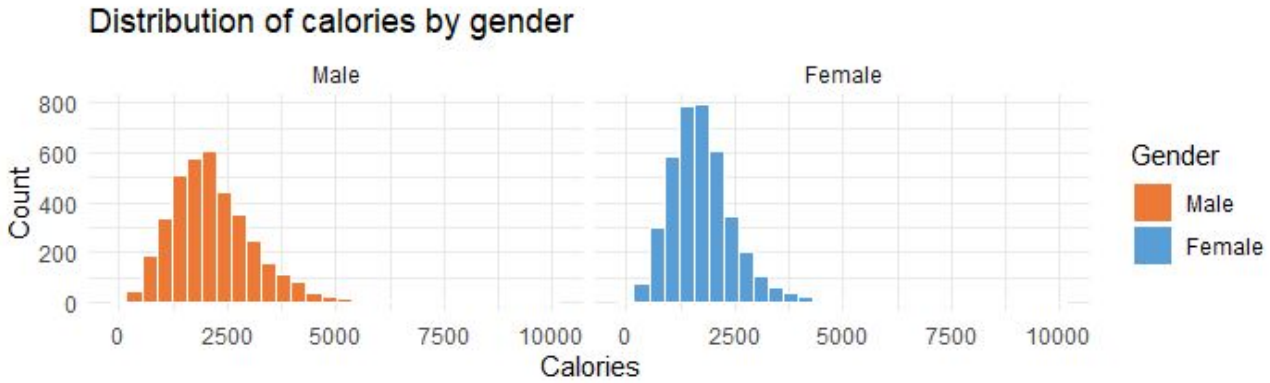


# Thank you!

Any questions?

# Appendix




- Relatively normal
- Slightly skewed to the right
- Mean = 1,953 calories
  - **Men** = ~2,175 calories
  - **Women** = ~1,741 calories

# Methodology - DEMO_I.XPT

- Filtered out samples younger than 20 years old
- Selected and renamed only the desired demographic variables
- Factorized certain variables
    - AGE
    - GENDER
    - EDUCATION
    - RACE
    - MARITAL
    - HH_INC

# Methodology - DRXFCD_I.XPT

Subsetted with numeric food codes related to coffee consumption (n = 135)

- Coffee creamer
- Various coffee types
  - Bottled
  - Brewed
  - Cafe mocha
  - Cappuccino
  - Cuban
  - Espresso
  - Frozen
  - Iced
  - Instant
  - Latte
  - Macchiato
  - Mocha
  - Etc.

Omitted unrelated coffee products

- Chocolate-covered espresso coffee beans
- Coffee cake
- Coffee-flavored cordial or liqueur
- Irish coffee
- Spanish coffee bread

# Methodology - DR1IFF_I.XPT

- Created **total** (a dataset with participant sequence number and the total number of calories consumed that day)

- Inner joined **diet** with **food_code** to create **join**
- Inner joined **join** with **coffee_code** to create **coffee** ( a new dataset with participant sequence number the total calories from coffee consumption)

- Left joined **total** and **coffee** together to create **clean** (new dataset with participant sequence number, total calories, and total calories from coffee consumption)
- Replaced NAs in COFFEE_TOTAL with 0 in **clean**

- Inner joined **clean** with **demo** to create **final**