Statistical Analysis of Food Environment Atlas

By: Carolina Garza, Tola Ouk, Michael Saenz, and Ray Thomas

Introduction

Data: "Food Environment Atlas" from the U.S. Department of Agriculture (August 2020)
Number of Observations: 3,143
Number of Features: 280

This dataset examines how agriculture, government assistance, grocery store availability, and demographics relate to food availability to households in the US. It is broken down to the State and County level.

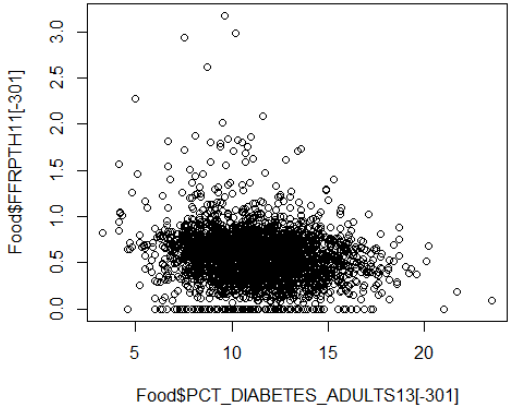
We have chosen a subset of 13 variables to answer 4 analytical questions.

1. **PCT_DIABETES_ADULTS13**: Adult diabetes rate, 2013
2. **FFRPTH11**: Fast-food restaurants/1,000 pop, 2011
3. **FSRPTH11**: Full-service restaurants/1,000 pop, 2011
4. **POVRATE15**: Poverty rate, 2015
5. **MEDHHINC15**: Median household income, 2015
6. **PCT_65OLDER10**: % Population 65 years or older, 2010
7. **PCT_18YOUNGER10**: % Population under age 18, 2010
8. **METRO13**: Metro/nonmetro counties, 2010 (Categorical variable)
9. **RECFACPTH11**: Recreation & fitness facilities/1,000 pop, 2011
10. **PC_DIRSALES12**: Direct farm sales per capita, 2012
11. **PCT_LACCESS_POP10**: Population, low access to store (%), 2010
12. **GROCPTH11**: Grocery stores/1,000 pop, 2011
13. **PC_SNAPBEN12**: SNAP benefits per capita, 2012

1. Is there a correlation between diabetes rate and the number of fast-food restaurants?

- Very weak negative correlation
- [-0.195, -0.127] with 95% confidence

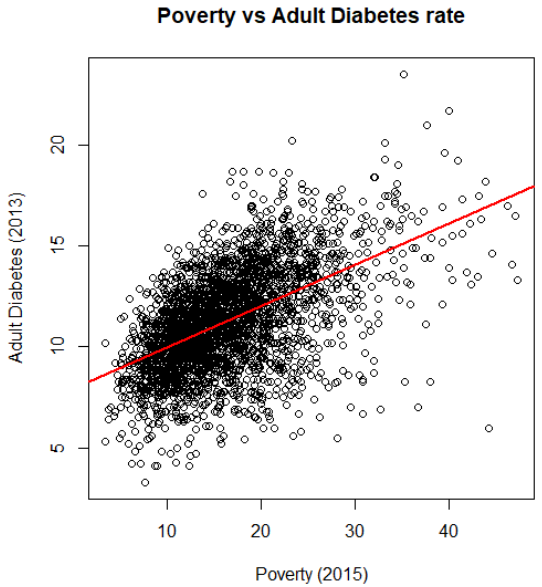
```
        Pearson's product-moment correlation

data:  Food$FFRPTH11 and Food$PCT_DIABETES_ADULTS13
t = -9.1681, df = 3140, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1953308 -0.1272152
sample estimates:
       cor
-0.1614653
```



2. Is there a correlation between health and poverty rates?

- Moderate positive correlation
- [0.505, 0.555] with 95% confidence
- Slope of 0.204

```
        Pearson's product-moment correlation

data:  data$POVRATE15 and data$PCT_DIABETES_ADULTS13
t = 35.062, df = 3137, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5049997 0.5552873
sample estimates:
      cor
0.5306102
```



3. What features can determine diabetes rates the most?

- **Multiple Linear Regression Model**

Significant Variables:

FFRPTH11: Fast-food restaurants/1,000 pop, 2011
FSRPTH11: Full-service restaurants/1,000 pop, 2011
MEDHHINC15: Median household income, 2015
PCT_65OLDER10: % Population 65 years or older, 2010
METRO13: Metro/nonmetro counties, 2010 (Categorical variable)
RECFACPTH11: Recreation & fitness facilities/1,000 pop, 2011
PC_DIRSALES12: Direct farm sales per capita, 2012
PCT_LACCESS_POP10: Population, low access to store (%), 2010
PC_SNAPBEN12: SNAP benefits per capita, 2012

Variation explained: 51%

- **Random Forest Model**

Top 5 Variables:

MEDHHINC15: Median household income, 2015
PC_SNAPBEN12: SNAP benefits per capita, 2012
PCT_65OLDER10: % Population 65 years or older, 2010
POVRATE15: Poverty rate, 2015
FSRPTH11: Full-service restaurants/1,000 pop, 2011

Variation explained: 58%

4. Can we predict the amount of SNAP benefits per capita?

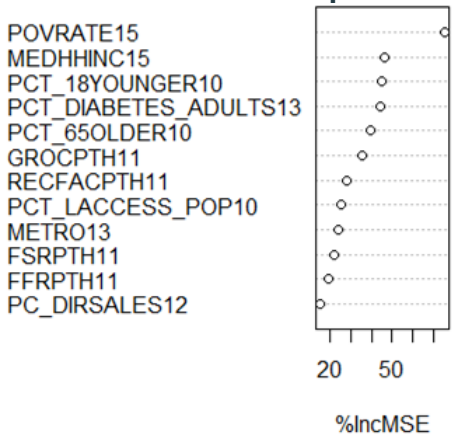
- **Multiple Linear Regression Model**

```
Call:
lm(formula = PC_SNAPBEN12 ~ PCT_DIABETES_ADULTS13 + FFRPTH11 +
    POVRATE15 + MEDHHINC15 + PCT_65OLDER10 + PCT_18YOUNGER10 +
    METRO13 + RECFACPTH11 + PCT_LACCESS_POP10 + GROCPTH11, data =
 df_no_state_county)

Residuals:
    Min      1Q  Median      3Q     Max
-21.389  -3.107  -0.303   2.893  34.250

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.394e+01  2.050e+00  -6.801 1.26e-11 ***
PCT_DIABETES_ADULTS13    8.473e-01  5.120e-02  16.549  < 2e-16 ***
FFRPTH11                 1.965e+00  3.622e-01   5.423 6.34e-08 ***
POVRATE15                1.103e+00  3.007e-02  36.692  < 2e-16 ***
MEDHHINC15              -5.751e-05  1.631e-05  -3.526 0.000429 ***
PCT_65OLDER10           -5.452e-02  3.858e-02  -1.413 0.157658
PCT_18YOUNGER10          4.007e-01  3.752e-02  10.680  < 2e-16 ***
METRO13                  2.252e+00  2.343e-01   9.612  < 2e-16 ***
RECFACPTH11              3.627e+00  1.445e+00   2.510 0.012126 *
PCT_LACCESS_POP10       -5.617e-02  5.464e-03 -10.280  < 2e-16 ***
GROCPTH11               -8.091e-01  5.646e-01  -1.433 0.151975
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.085 on 2867 degrees of freedom
Multiple R-squared:  0.7485,    Adjusted R-squared:  0.7476
F-statistic: 853.3 on 10 and 2867 DF,  p-value: < 2.2e-16
```

- **Random Forest Model**

```
Call:
 randomForest(formula = PC_SNAPBEN12 ~ ., data = train, xtest = X.test,
     ytest = y.test, ntree = 1000, mtry = sqrt(p), importance = TRUE)
               Type of random forest: regression
                     Number of trees: 1000
No. of variables tried at each split: 3

          Mean of squared residuals: 22.5492
                    % Var explained: 78.18
                      Test set MSE: 24.32
                    % Var explained: 75.38
```

- **Variable Importance Plot**



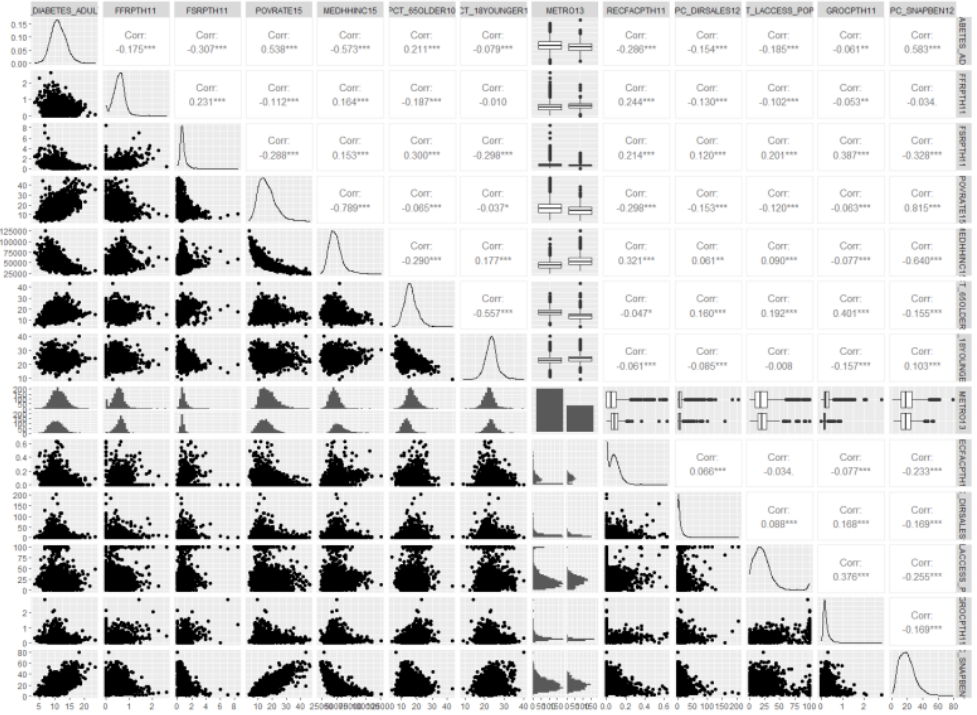
	%IncMSE
POVRATE15	74.82153
MEDHHINC15	46.56223
PCT_18YOUNGER10	44.80507
PCT_DIABETES_ADULTS13	44.33111
PCT_65OLDER10	39.49203
GROCPTH11	35.64165
RECFACPTH11	28.45903
PCT_LACCESS_POP10	25.60713
METRO13	24.11148
FSRPTH11	22.58747
FFRPTH11	19.85260
PC_DIRSALES12	15.82888

3. What features can determine diabetes rates the most?

Top correlations with Diabetes rates:

- poverty rate
- median household income
- snap benefit amount

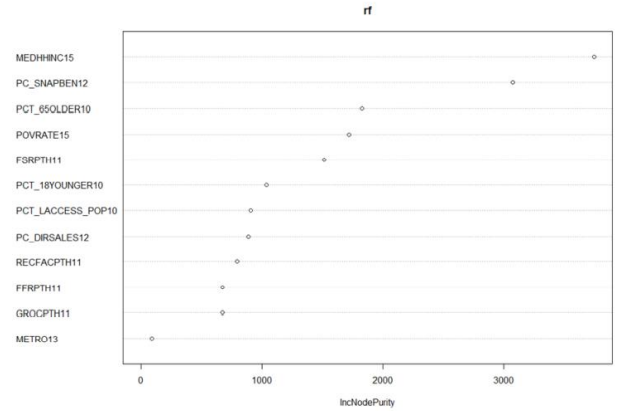
(correlation values of 0.54 - 0.58)



3. What features can determine diabetes rates the most?

- **Random Forest Model**

```
Call:
 randomForest(formula = df$PCT_DIABETES_ADULTS13 ~ ., data = df)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 2.51639
                    % Var explained: 58.43
```



Top Variables:

MEDHHINC15: Median household income, 2015
PC_SNAPBEN12: SNAP benefits per capita, 2012
PCT_65OLDER10: % Population 65 years or older, 2010
POVRATE15: Poverty rate, 2015
FSRPTH11: Full-service restaurants/1,000 pop, 2011

3. What features can determine diabetes rates the most?

- **Multiple Linear Regression Model**

```
Residuals:
    Min      1Q  Median      3Q     Max
-7.1528 -1.1293 -0.0189  1.0616  8.0220

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        8.680e+00  6.979e-01  12.437  < 2e-16 ***
FFRPTH11          -3.702e-01  1.385e-01  -2.674  0.00755 **
FSRPTH11          -8.479e-01  8.033e-02 -10.555  < 2e-16 ***
POVRATE15          1.897e-03  1.246e-02   0.152  0.87896
MEDHHINC15        -3.262e-05  5.521e-06  -5.909 3.86e-09 ***
PCT_65OLDER10      2.015e-01  1.261e-02  15.979  < 2e-16 ***
PCT_18YOUNGER10    1.419e-02  1.323e-02   1.073  0.28353
METRO131           2.127e-01  8.147e-02   2.611  0.00908 **
RECFACPTH11       -2.619e+00  4.943e-01  -5.299 1.25e-07 ***
PC_DIRSALES12     -1.732e-02  2.597e-03  -6.671 3.03e-11 ***
PCT_LACCESS_POP10 -1.054e-02  1.893e-03  -5.568 2.82e-08 ***
GROCPTH11         -2.193e-01  2.006e-01  -1.093  0.27441
PC_SNAPBEN12       1.007e-01  6.060e-03  16.620  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.726 on 2864 degrees of freedom
Multiple R-squared:  0.5102,    Adjusted R-squared:  0.5081
F-statistic: 248.6 on 12 and 2864 DF,  p-value: < 2.2e-16
```

Significant Variables:

FFRPTH11: Fast-food restaurants/1,000 pop, 2011
FSRPTH11: Full-service restaurants/1,000 pop, 2011
MEDHHINC15: Median household income, 2015
PCT_65OLDER10: % Population 65 years or older, 2010
METRO13: Metro/nonmetro counties, 2010 (Categorical variable)
RECFACPTH11: Recreation & fitness facilities/1,000 pop, 2011
PC_DIRSALES12: Direct farm sales per capita, 2012
PCT_LACCESS_POP10: Population, low access to store (%), 2010
PC_SNAPBEN12: SNAP benefits per capita, 2012