# Statistical Analysis of Food Environment Atlas

Group 5: Carolina Garza, Tola Ouk, Michael Saenz, Ray Thomas

Introduction:

For our project, we have chosen to use the "Food Environment Atlas" provided by the U.S. Department of Agriculture (https://www.ers.usda.gov/data-products/food-environment-atlas/data-access-and-documentation-downloads/#Current%20Version). This dataset covers a wide range of data related to food such as agriculture, government assistance, grocery stores, and demographics in the US. This is the most current release, and was published in August 2020. This data has 3,143 observations, with each observation representing a county in the US. Although this data has over 280 variables overall, we have decided to focus on 13 variables for this exploration. The variables are as follows:

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

Although the available data are not all from the same year, we assume that the values do not vary significantly within the timeframe.

The purpose of our exploration is to answer four questions:

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

Question 2: Is there a correlation between health and poverty rate?

Question 3: What features can determine diabetes rate the most?

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

Methodology:

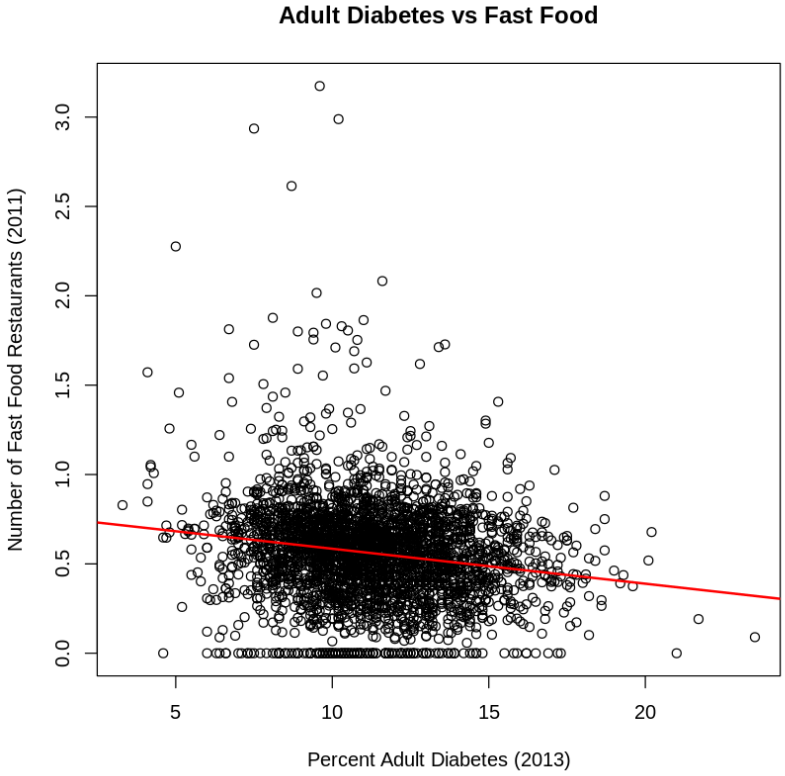
Before analyzing any of our proposed questions, we will do a preliminary analysis and get rid of any outliers in our data. For Question 1 and Question 2 we will plot the variables and use simple correlation, specifically the Pearson Correlation test.

For Questions 3 and 4, we will be using Multiple linear regression, Random Forest, and the associated Variable Importance Plots to answer our questions. For the Multiple Linear Regression analysis, we will use Stepwise Regression to choose the best subset of variables for the model. And we will utilize the R-Square values to determine the efficacy of the models in answering these prediction questions.

Data Analysis:

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

After performing the necessary data cleaning steps, we conduct a preliminary analysis. From this, we identified outliers, and removed them from the data. Then we plot the data to get a visual representation of the problem.



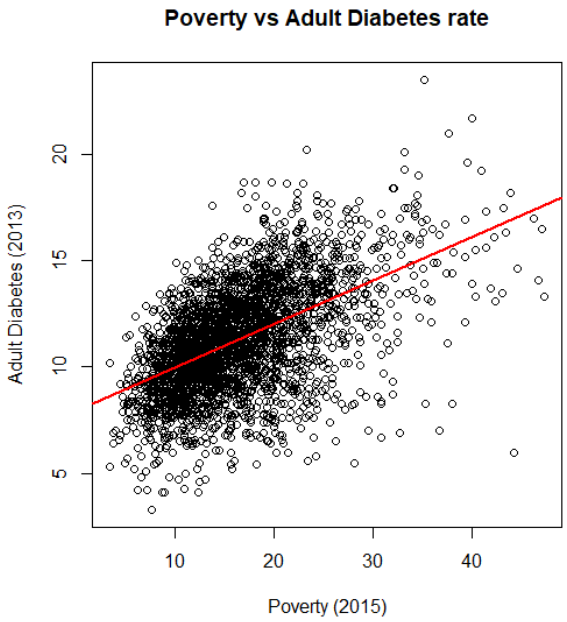

Now we have a gut check as we look into the correlation numbers. We create a linear model using both the rate of diabetes and number of fast food restaurants per 1000 residents. A supplemental correlation test is used for additional context for the data.

```
> food.lm <-lm(Food$FFRPTH11~Food$PCT_DIABETES_ADULTS13)
> summary(food.lm)

Call:
lm(formula = Food$FFRPTH11 ~ Food$PCT_DIABETES_ADULTS13)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6902 -0.1597  0.0059  0.1426  5.1480

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 0.780204   0.024561  31.766   <2e-16 ***
Food$PCT_DIABETES_ADULTS13 -0.019568   0.002134  -9.168   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2973 on 3140 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.02607,   Adjusted R-squared:  0.02576
F-statistic: 84.05 on 1 and 3140 DF,  p-value: < 2.2e-16

> cor.test(Food$FFRPTH11,Food$PCT_DIABETES_ADULTS13)

        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
```

The correlation is strictly negative, as we can verify from the graph. The linear model reveals a negative slope of -0.019. According to the Pearson correlation test, we also notice a very low correlation coefficient of -0.16. These numbers show a very weak, negative correlation. The conclusion that the correlation is not positive reframes the results in a clearer, more concise way.

2. **Is there a correlation between health and poverty rate?**

In order to answer this question, we decided to focus on the poverty rate and the adult diabetes rate. From the scatterplot alone, we notice that there might be a slight correlation between these two variables. We decided to fit a linear regression model to these variables, and obtained a slope of 0.20.

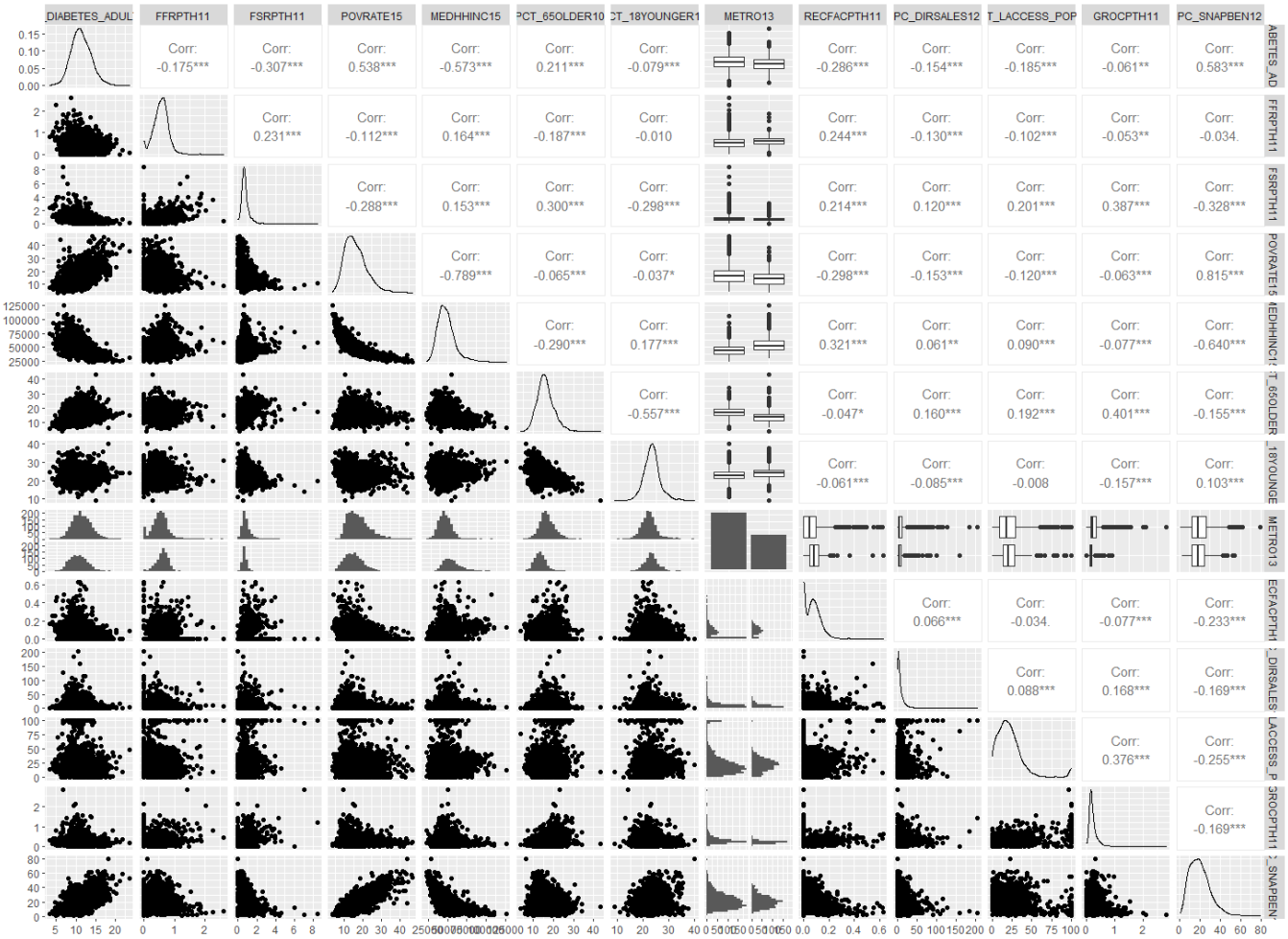



In order to justify the relationship between these two, we ran a Pearson correlation test with an alpha of 0.05. This gave us a correlation coefficient between [0.501, 0.555] with 95% confidence. This means that there is somewhat of a positive correlation between poverty rate and the rate of adult diabetes.

```
        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 rate the most?

Before we fit any model, we would like to see the overall distributions and correlations among the variables. Below is the ggpairs plot of the dataset.




Diabetes rates appear to have correlations with poverty rate, median household income, and SNAP benefit amount (correlation values of 0.538 - 0.583), and have weak correlations with other variables.

Two methods are used to identify which factors can be used to determine the diabetes rate. They are Multiple Linear Regression and Random Forest. Below is the result from the Multiple Linear Regression method.

```
> summary(mlr)

Call:
lm(formula = fea_clean[4:16]$PCT_DIABETES_ADULTS13 ~ ., data = fea_clean[4:16])

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
```

This result suggests that diabetes rate can be determined by the following 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

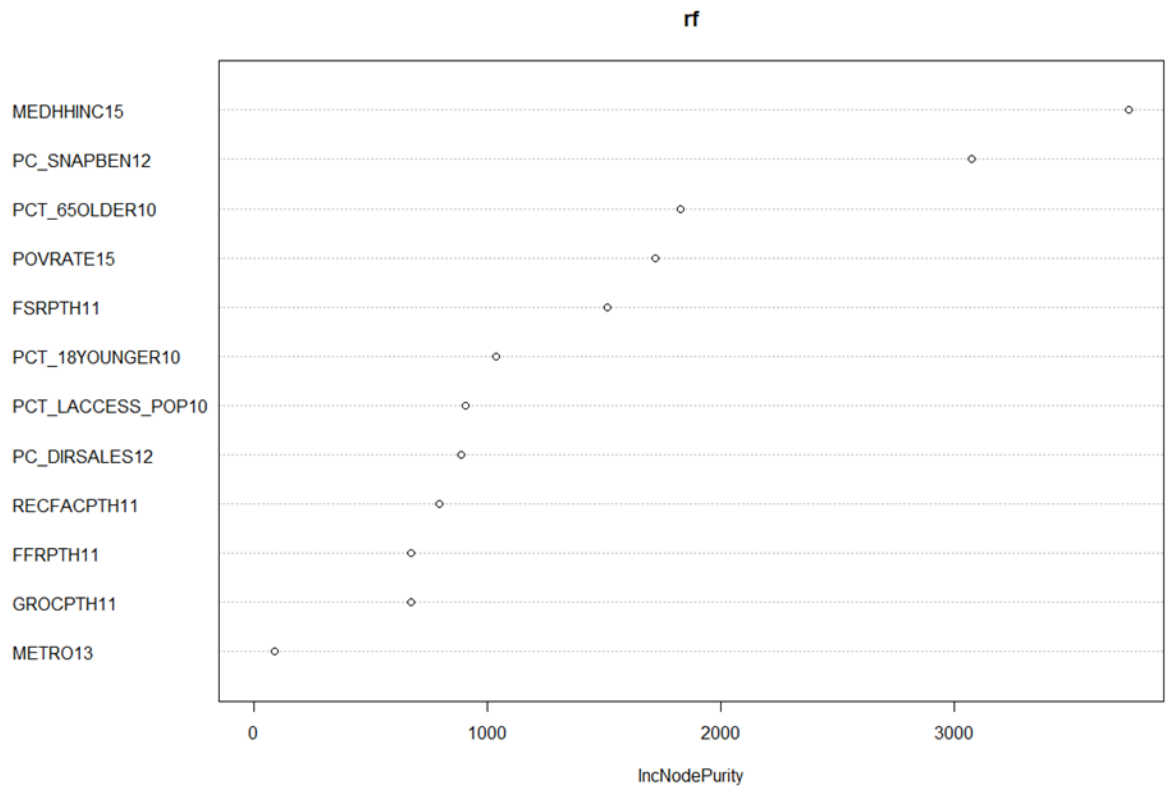
The model is:
8.680 - 0.3702 FFRPTH11 - 0.8479 FSRPTH11 - 3.262 e-05 MEDHHINC15 + 0.2015 PCT_65OLDER10 + 0.2127 METRO13 - 2.619 RECFACPTH11 - 0.01732 PC_DIRSALES12 - 0.01054 PCT_LACCESS_POP10 + 0.1007 PC_SNAPBEN12

An adjusted R-squared of 0.5081 suggests that about half of the variations in the diabetes rate can be explained by the model.

The diabetes rate is positively correlated with percent of population over 65, if the county is in a metropolitan area, and percent of SNAP benefits. The diabetes rate is negatively correlated to the number of fast-food restaurants, number of full-service restaurants, number of recreation and fitness facilities, percent direct farm sales, and percent of population with low access to stores.

Below is the result from the Random Forest method.

```
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
```



The result is slightly better than MLR method. 58.43% of variance can be explained by this model. The top 5 most important variables are

**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

Among these 5 variables, poverty rate is the only variable that does not show up in the multiple linear regression. However, other variables might be correlated to poverty level already. For example, the coefficient for the number of full-service restaurants is more than double of that of the number of fast-food restaurants (-0.85 to -0.37). This could be correlated to poverty level as the higher number of full-service restaurants suggests there are a higher number of people that can afford to eat out at full-service restaurants.

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

Once again, both Multiple Linear Regression (MLR) and Random Forest (RF) models were utilized to answer the question of whether SNAP participation can be predicted for participants. Using stepwise regression, we determined the optimal subset of factors for the MLR model. Below is the output from RStudio for the resultant 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
```

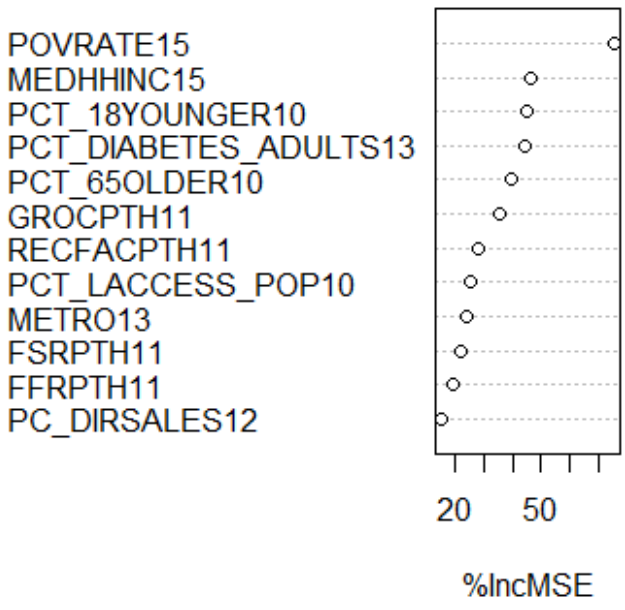
As seen in the output, 10 of the original factors were used to build the model. The Adjusted R-squared of 0.7476 suggests that about 74.76 % of the variation in SNAP participation can be explained by the model.

Subsequently, we developed the Random Forest model after performing a preliminary 80-20 split of the data to training and test subsets. Then we built the model with 1000 trees and 4 variables used at each split. Below are the outputs from the random forest modeling.

```
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
```

The output shows that the model is able to explain 78.18% variation in the training data and 75.38% variation in the test data.

The variable importance plots for the random forest model is presented below.




| | %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 |

We can see from this output that the MSE decreases by 74.82% if Poverty Rate is excluded from the model. This tells us that Poverty Rate is a very strong indicator of whether a person participates in SNAP.

Furthermore, we explored the possibility of adding State and County variables to the models to see their effects on the two models. As before, we proceeded with stepwise regression to whittle away at the extraneous factors. Below is resultant model output.

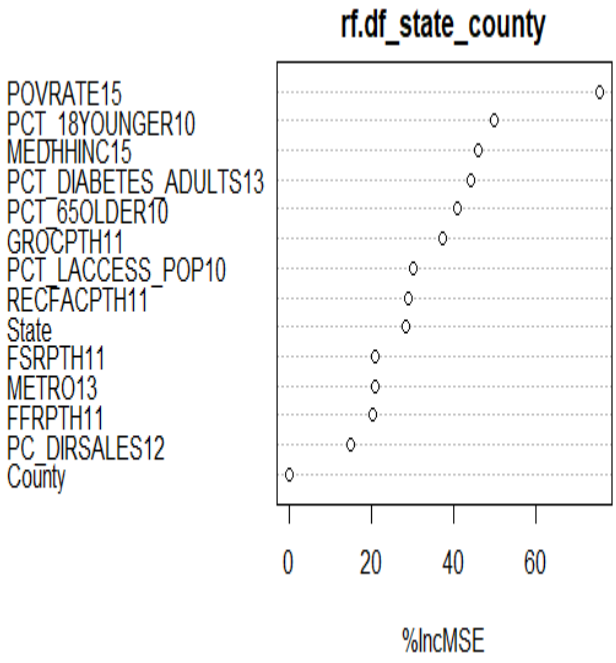
```
Call:
lm(formula = PC_SNAPBEN12 ~ State + PCT_DIABETES_ADULTS13 +
FFRPTH11 +
    POVRATE15 + MEDHHINC15 + PCT_65OLDER10 + PCT_18YOUNGER10 +
    METRO13 + RECFACPTH11 + PC_DIRSALES12 + PCT_LACCESS_POP10 +
    GROCPTH11, data = df_state_county)

Residuals:
     Min       1Q   Median       3Q      Max
-21.8412  -2.2844  -0.0705   2.2948  21.8126

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.329e+01  2.711e+00  -4.904 9.93e-07 ***
StateAL        -1.043e+01  2.161e+00  -4.828 1.45e-06 ***
StateAR        -1.293e+01  2.142e+00  -6.037 1.77e-09 ***
StateAZ        -7.713e+00  2.370e+00  -3.254 0.001151 **
StateCA        -9.174e+00  2.135e+00  -4.297 1.79e-05 ***
StateCO        -5.990e+00  2.127e+00  -2.816 0.004901 **
StateCT        -1.247e+00  2.518e+00  -0.495 0.620584
StateDE        -2.276e+00  3.148e+00  -0.723 0.469684
StateFL        -5.457e+00  2.144e+00  -2.546 0.010957 *
StateGA        -5.584e+00  2.107e+00  -2.650 0.008094 **
StateHI         1.611e+01  3.140e+00   5.129 3.11e-07 ***
StateIA        -7.954e+00  2.108e+00  -3.773 0.000164 ***
StateID        -1.126e+01  2.164e+00  -5.204 2.09e-07 ***
StateIL        -5.682e+00  2.107e+00  -2.696 0.007053 **
```

```
StateIN              -9.730e+00  2.120e+00  -4.589 4.65e-06 ***
StateKS              -1.303e+01  2.113e+00  -6.169 7.88e-10 ***
StateKY              -6.986e+00  2.116e+00  -3.301 0.000976 ***
StateLA              -7.547e+00  2.148e+00  -3.513 0.000449 ***
StateMA              -1.312e+00  2.337e+00  -0.561 0.574578
StateMD              -4.119e+00  2.239e+00  -1.839 0.065978 .
StateME               2.332e+00  2.312e+00   1.009 0.313277
StateMI              -2.776e+00  2.120e+00  -1.309 0.190519
StateMN              -8.629e+00  2.110e+00  -4.089 4.45e-05 ***
StateMO              -9.649e+00  2.110e+00  -4.574 5.00e-06 ***
StateMS              -1.149e+01  2.146e+00  -5.355 9.24e-08 ***
StateMT              -1.026e+01  2.150e+00  -4.774 1.90e-06 ***
StateNC              -8.356e+00  2.119e+00  -3.944 8.22e-05 ***
StateND              -7.450e+00  2.156e+00  -3.456 0.000557 ***
StateNE              -1.156e+01  2.116e+00  -5.464 5.05e-08 ***
StateNH              -4.593e+00  2.437e+00  -1.885 0.059529 .
StateNJ              -5.765e+00  2.270e+00  -2.540 0.011142 *
StateNM              -2.842e+00  2.200e+00  -1.292 0.196580
StateNV              -8.532e+00  2.332e+00  -3.659 0.000258 ***
StateNY              -4.244e+00  2.129e+00  -1.994 0.046276 *
StateOH              -7.625e+00  2.122e+00  -3.593 0.000332 ***
StateOK              -9.838e+00  2.137e+00  -4.604 4.34e-06 ***
StateOR               7.137e-01  2.179e+00   0.328 0.743237
StatePA              -6.630e+00  2.133e+00  -3.108 0.001900 **
StateRI              -2.216e+00  2.762e+00  -0.802 0.422374
StateSC              -5.853e+00  2.184e+00  -2.680 0.007409 **
StateSD              -8.781e+00  2.141e+00  -4.102 4.22e-05 ***
StateTN              -4.323e+00  2.130e+00  -2.030 0.042494 *
StateTX              -1.167e+01  2.087e+00  -5.590 2.48e-08 ***
StateUT              -1.436e+01  2.212e+00  -6.492 9.98e-11 ***
StateVA              -6.326e+00  2.127e+00  -2.974 0.002964 **
StateVT               5.945e+00  2.353e+00   2.527 0.011564 *
StateWA              -1.032e+00  2.167e+00  -0.476 0.633923
StateWI              -5.992e+00  2.120e+00  -2.826 0.004743 **
StateWV              -8.067e+00  2.164e+00  -3.727 0.000197 ***
StateWY              -1.068e+01  2.242e+00  -4.764 1.99e-06 ***
PCT_DIABETES_ADULTS13  8.647e-01  6.106e-02  14.162  < 2e-16 ***
FFRPTH11             2.035e+00  3.054e-01   6.663 3.22e-11 ***
POVRATE15            1.023e+00  2.743e-02  37.313  < 2e-16 ***
MEDHHINC15          -1.151e-04  1.458e-05  -7.896 4.08e-15 ***
PCT_65OLDER10        5.952e-02  3.495e-02   1.703 0.088628 .
PCT_18YOUNGER10      8.056e-01  3.542e-02  22.745  < 2e-16 ***
METRO13              1.574e+00  1.987e-01   7.920 3.38e-15 ***
RECFACPTH11          2.062e+00  1.188e+00   1.736 0.082648 .
PC_DIRSALES12       -4.173e-02  6.665e-03  -6.262 4.39e-10 ***
PCT_LACCESS_POP10   -2.892e-02  4.769e-03  -6.064 1.50e-09 ***
GROCPTH11           -1.658e+00  4.809e-01  -3.449 0.000572 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.098 on 2817 degrees of freedom
Multiple R-squared:  0.8395,       Adjusted R-squared:  0.836
F-statistic: 245.5 on 60 and 2817 DF,  p-value: < 2.2e-16
```

At the cost of added computation time, there was an 8.84% increase in accuracy for the Multiple Linear Regression model when having added those factors. This is heavily influenced by a surge of new variables being added by way of several states being introduced into the model.

We pursued the same inquiry regarding the influence of State and County variables on the Random Forest model. The results and the variable importance plots are shown below:

```
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: 4

          Mean of squared residuals: 21.65764
                    % Var explained: 79.04
                       Test set MSE: 24.09
                    % Var explained: 75.61
```




| | %IncMSE |
|---|---|
| POVRATE15 | 75.67059400 |
| PCT_18YOUNGER10 | 50.07907550 |
| MEDHHINC15 | 45.96598022 |
| PCT_DIABETES_ADULTS13 | 44.35394348 |
| PCT_65OLDER10 | 40.97395208 |
| GROCPTH11 | 37.45501841 |
| PCT_LACCESS_POP10 | 30.25464177 |
| RECFACPTH11 | 28.95424935 |
| State | 28.53379309 |
| FSRPTH11 | 21.07490298 |
| METRO13 | 20.94500443 |
| FFRPTH11 | 20.38441558 |
| PC_DIRSALES12 | 14.93552985 |
| County | 0.04684459 |

We find that in the case of the random forest model, there is no significant increase in accuracy when the two new factors are added. The training set explained variance increases minimally by 0.86%, and the test set explained variance increases by 0.23%. We also see from the variable importance plot that State and County factors rank low in importance, decreasing the MSE by only 28.53% and 0.05% respectively.

Conclusion:

We were able to answer the four research questions. There were some surprises such as the weak negative correlation between diabetes rate and number of fast food restaurants per capita. The diabetes rate has some correlations with poverty rate (about 0.53 correlation). This somewhat agrees with our intuition as access to healthy food options and good living conditions can affect health. The diabetes rate can be determined by our model at 58% explained variance rate. Determining the amount of SNAP benefits achieved better results by our model (up to about 75% variance explained rate). For further research, latest data such as post-pandemic one might be interesting to analyze and compare to this study.

Reference:
Economic Research Service (ERS), U.S. Department of Agriculture (USDA). Food Access Research Atlas, https://www.ers.usda.gov/data-products/food-access-research-atlas/