Technical Report

Rishi Krishnamurthy, Tina Qin, Kenai Burton-Heckman

11/15/2021

Abstract

The Supplemental Nutrition Assistance Program (SNAP) is a benefits program that provides recipients with funds to redeem at participating food retailers, serving more than 42 million people in 2021. Given that a substantial share of the American population relies on SNAP benefits, it is crucial to understand how spending behavior, operationalized as the average number of SNAP redemptions per authorized SNAP store, relates to county-level factors above and beyond common predictors like income. The present analysis predicts average SNAP redemption from county demographics, participation in other benefit programs, and other socioeconomic conditions, using data from the USDA's Food Environment Atlas. We fit multilinear regression, ridge regression, LASSO, and random forest regression models to data from 2015 to 2017, using county-level average SNAP redemptions per authorized store as the response. We found that a random forest performed best, identifying per-capita benefits, stores per capita, superstores per capita, and number of redemptions in a related benefits program as the best predictors of average redemption. Linear models identified a similar set of predictors as significant, and showed that while spending was generally lower when more stores were nearby, spending was higher when specifically more supercenters were nearby. Our findings reaffirm the importance of having eligible retailers nearby, especially larger stores that may be able to meet more diverse needs. Importantly, the presence of eligible retailers was a better predictor of average redemption than more intuitive predictors, like poverty rate, median income, and race- among SNAP-eligible people, relevant demographic disparities may be overshadowed by a simple lack of access to stores.

Introduction

The Supplemental Nutrition Assistance Program (SNAP), commonly known as "food stamps," is a service of the U.S. Department of Agriculture (USDA). It provides monthly benefits to program participants, which can be redeemed with certain food retailers in order to purchase food. People qualify for SNAP by meeting income, employment, resource, and expenditure-related criteria (USDA, 2020a). Since SNAP applicants' eligibility relies on showing that they cannot buy all of their food with their existing income, demographic trends in the SNAP recipient population reflect existing socioeconomic inequalities in the U.S. SNAP recipients are disproportionately Black or Latinx, more likely than the average U.S. resident to have dependents, and make a maximum of 130% of the federal poverty level by definition (CBPP, 2021; Loveless, 2020; USCB, 2019, 2020). On average, Black, Latinx, and low-income people already experience greater food insecurity and are more likely to live in food deserts (Bower et al., 2014; Rabbitt et al., 2016; Sharma et al., 2015). Differences in the surroundings of marginalized groups may then be reflected in their spending behavior with SNAP benefits. For example, the number of grocery stores in a county, the extent to which county residents use other benefits programs, and county levels of access to cars may all be associated with different average SNAP redemptions. However, the relationship between the various barriers to purchase and average redemption amount is unclear. Having fewer grocery stores in a county may be associated with larger average redemptions, as reaching stores is more difficult and each trip "counts for more," but participation in other benefits programs may be higher in those same counties, reducing average redemptions if enrollees are not dipping into SNAP funds first.

Such questions about the purchasing behavior of SNAP recipients under different environmental and socioeconomic conditions have become increasingly urgent as SNAP enrollment has increased. Enrollment increased from approximately 37 million people in 2020 to 42 million in 2021, reversing a downward trend in SNAP participation. This growth was primarily driven by the economic fallout of the COVID-19 pandemic and the increase in monthly SNAP benefits implemented to compensate for economic hardship (CBPP, 2021; USDA, 2020b). Using recent county-level SNAP data collected by the USDA's Economic Research Service (ERS) from 2015 to 2017, we modeled the number of average SNAP redemptions per authorized SNAP store (measured by county) as a function of socioeconomic predictors such as number of grocery stores, access to grocery stores, and reliance on other benefits programs. This analysis highlights environmental and socioeconomic factors that predict average redemption size above and beyond obvious predictors like income, suggesting that structural differences in U.S. counties explain unique variance in residents' use of government benefits.

Methods

Data were drawn from the USDA Economic Research Service's (ERS) Food Environment Atlas (FEA), which tracks information about the food environment- predictors of dietary choice at a county and state level, such as access to grocery stores and per-capita SNAP benefits. FEA data are open-access and were downloaded from the USDA ERS website. Data were divided into subcategories: Access and Proximity to Grocery Store, Store Availability, Restaurant Availability and Expenditure, Food Assistance, State Food Insecurity, Food Prices and Taxes, Local Foods, Health and Physical Activity, and Socioeconomic Characteristics. In order to provide the model with more training data and consider more local variation in predictors and response, only county-level data within these subcategories were considered- state-level data were omitted. County-level data in each subcategory had 3143 observations, each corresponding to one U.S. county. While the response (the average number of SNAP redemptions per authorized SNAP store) was measured in 2017, relevant predictors were measured in 2017 or previous years, or over a range of previous years, in the case of predictors representing % change over time or multi-year averages. In order to minimize the impact of unmeasured, external factors on the predictors, only predictors measured between 2015 and 2017 were considered. While the inclusion of predictors measured before 2017 introduces unmeasured, external influences on predictor values, the predictive power of the model would suffer more from excluding these variables than from including them with the understanding that different sets of external factors influenced predictors measured at different times.

Data was downloaded as a .xlsx notebook and opened in R using the readxl package. Only counties with values for all relevant predictors were considered. This excluded one empty row, and three counties that were merged into other counties or changed their FIPS number. Additionally, one county was renamed to keep its name consistent across the different sheets in the Environmental Atlas excel file. Sheets were then joined by FIPS number and county name to create one large dataset (the county's state was named differently by sheet and it was therefore excluded as a variable to join by). Of the 281 predictors included in the FEA data, 78 were retained because they were measured in one or multiple years between 2015 and 2017.

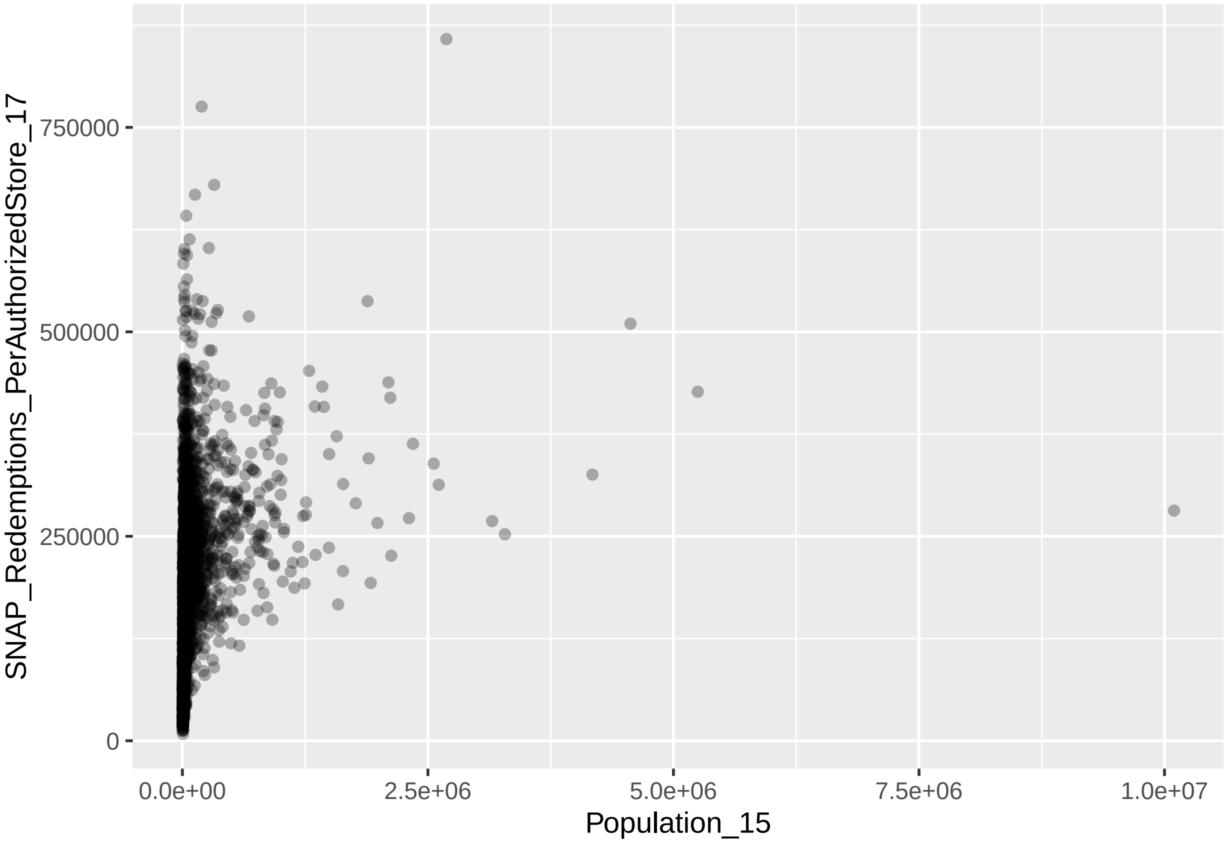
Data

We are using the USDA's Food Environment Atlas, and modeling average SNAP redemption per participating SNAP store in 2017 as a function of several socioeconomic predictors.

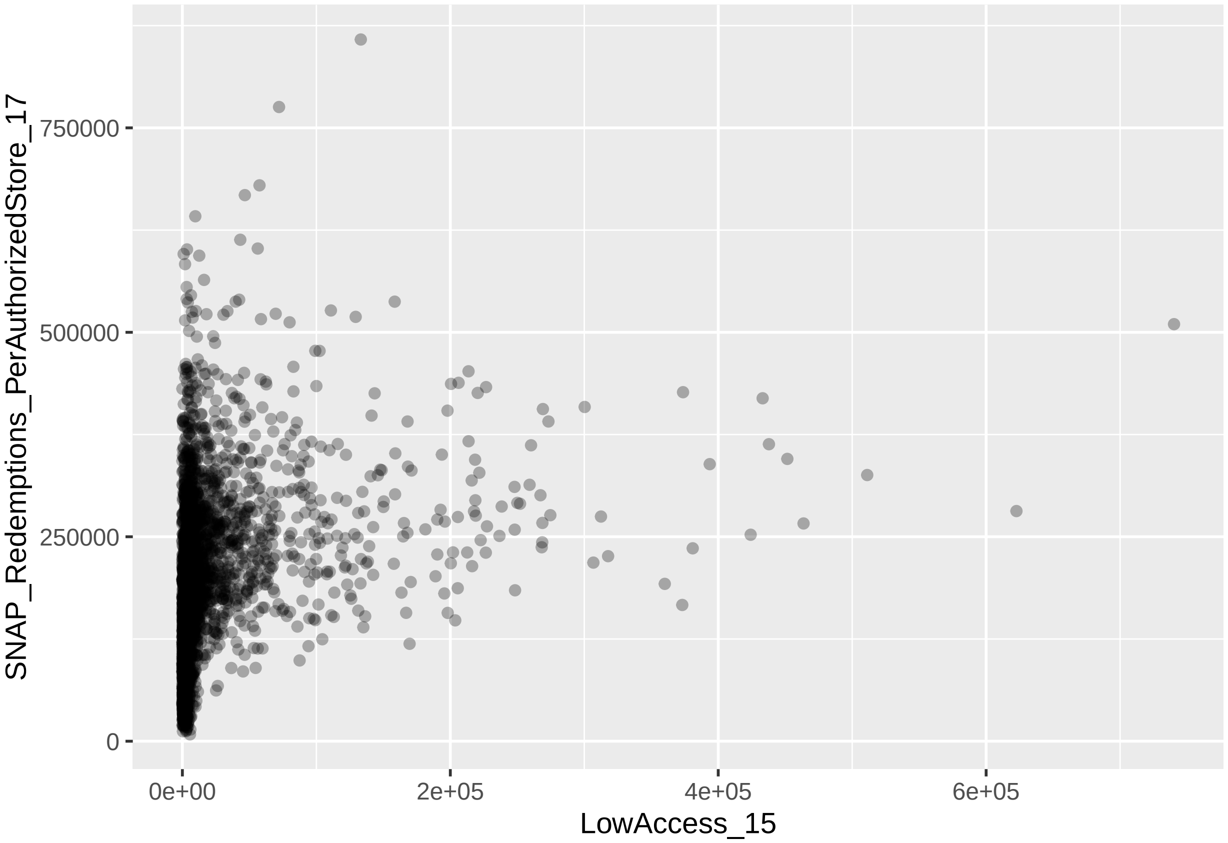
Exploratory Data Analysis

There are 3142 counties included in the data, although some predictors are measured on the state level. There are 78 total variables relevant to the three-year period we are investigating, 2015-2017. Only considering variables measured between 2015 and 2017 drastically reduces the number of available predictors, as most predictors were measured outside of that window. Even within the 2015-2017 period, variables were measured at different times, and were subject to different exogenous factors. Given the staggered measurements of the predictors, predictions made using our models will have to be interpreted cautiously, as environmental and socioeconomic conditions may have changed dramatically over those three years. We fit four different models to the data: multilinear regression, ridge regression, LASSO, and random forest regression. Before models were fit, we conducted exploratory data analysis to assess collinearity between predictors, the normality of the response, and the conformity of the data to model assumptions.

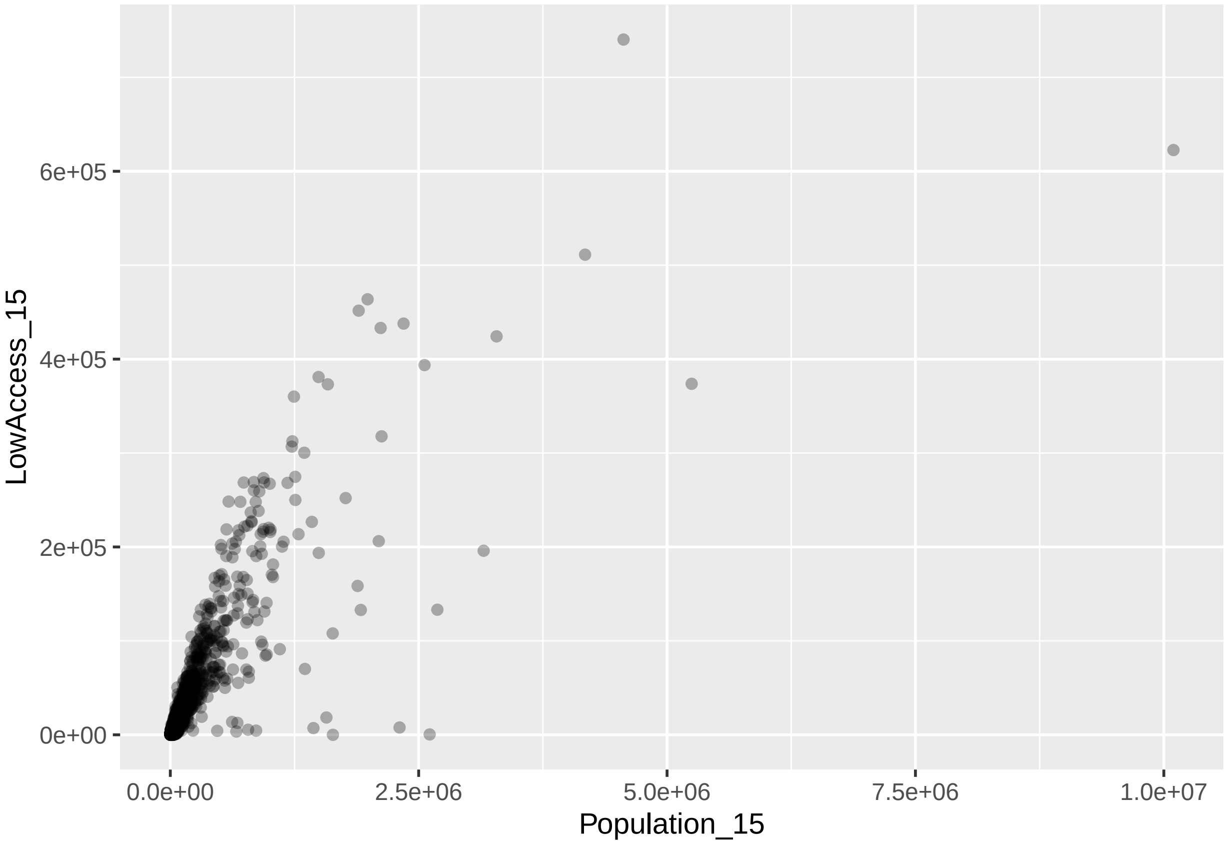
Scatterplots comparing some of the predictors that might influence the response the most:



Response in relation to the population that have low access in 2015:



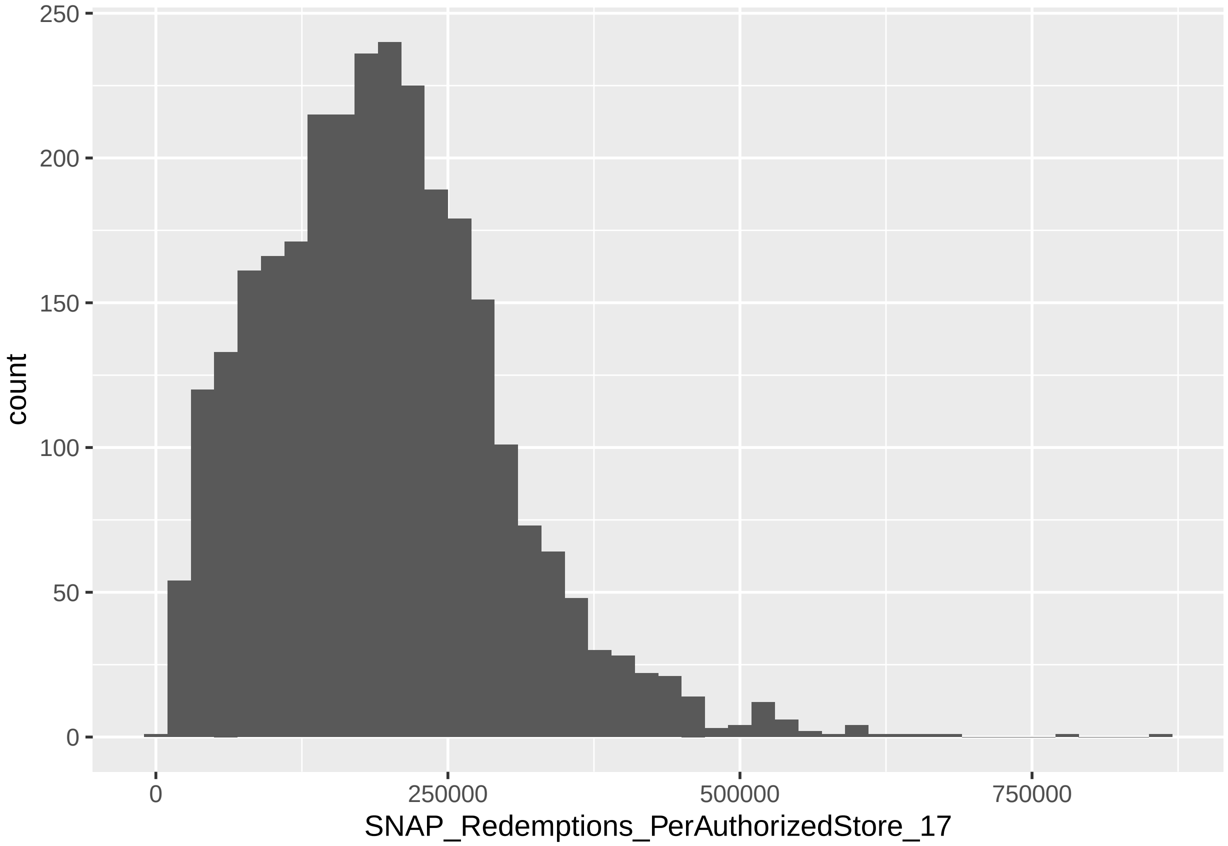
Population of low access in relation to the whole population in 2015:

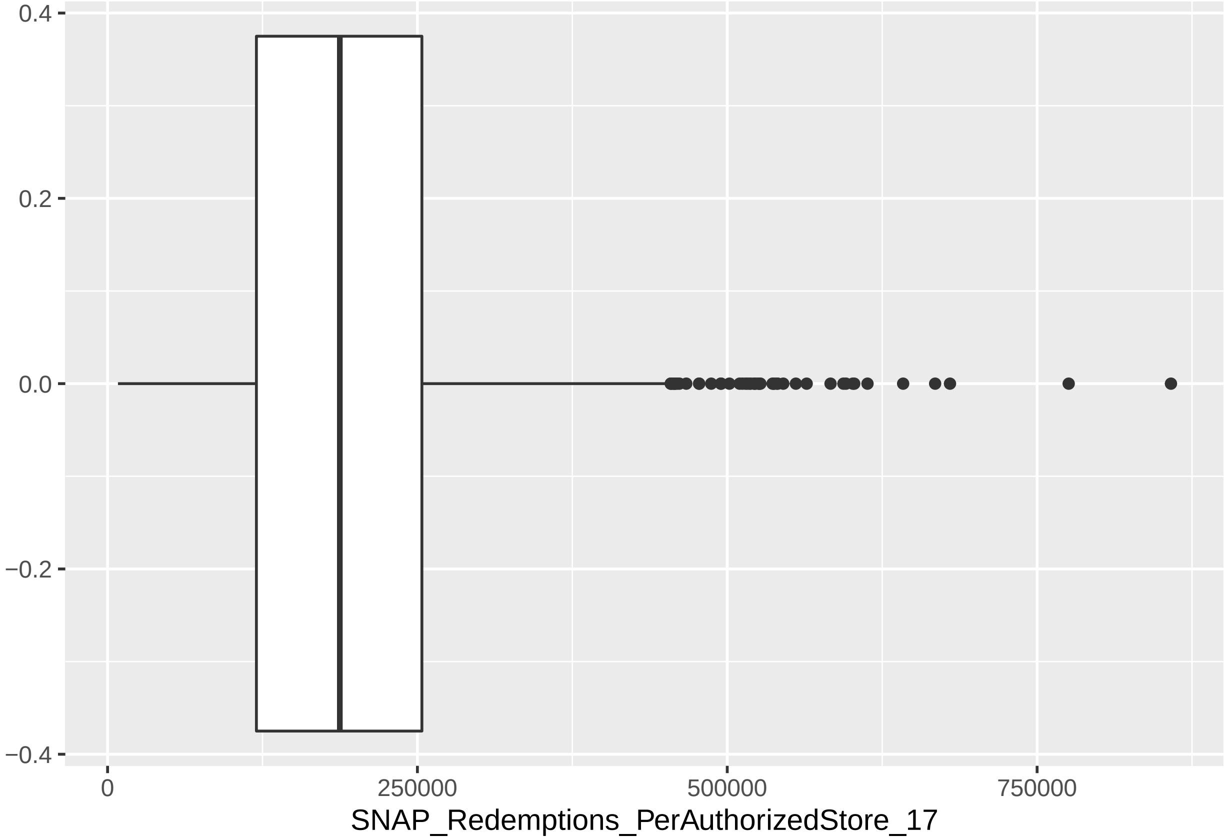


Before inspecting correlations, we removed variables that had an abnormally large number of NAs by inspecting the data summary. A correlation plot revealed several highly correlated variables (|r| > 0.8). Many of these correlations were between count variables, which would be expected to scale with population. Collinearity was mitigated while preprocessing the data by dropping highly correlated variables.

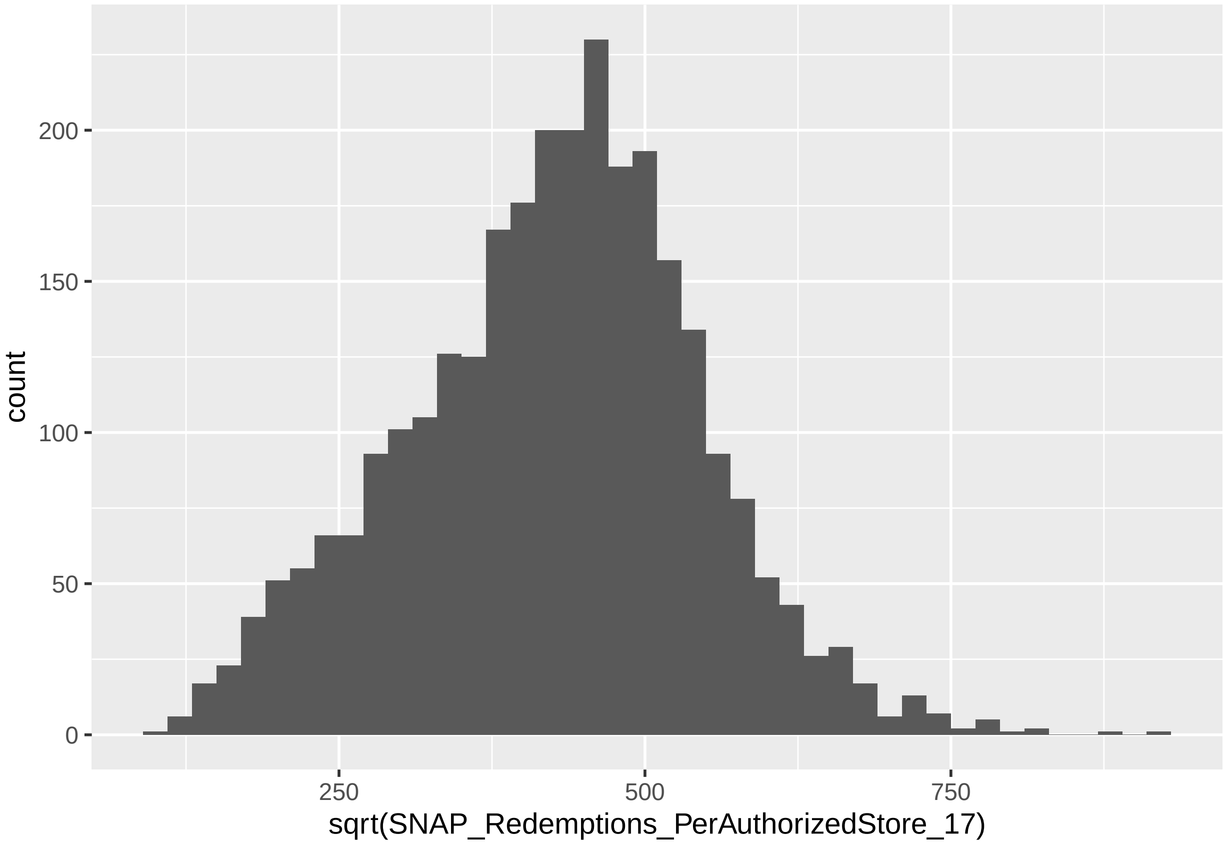
The distribution of the response variable, SNAP redemptions per authorized store, is bell-shaped, with right-skew, a center of 187,397, and an IQR of 74,742. A square root transformation reduces skew but does not actually improve the response variable's normality judging by a Shapiro-Wilk test. It would also decrease interpretability, so the untransformed response is still the most appropriate for the linear model.If our regression methods had failed to produce viable results, we planned to transform the response to a categorical variable with the levels "high redemption" and "low redemption" based on a median split.

Distribution and boxplot of the response:





Distribution of the square root transformation of the response:

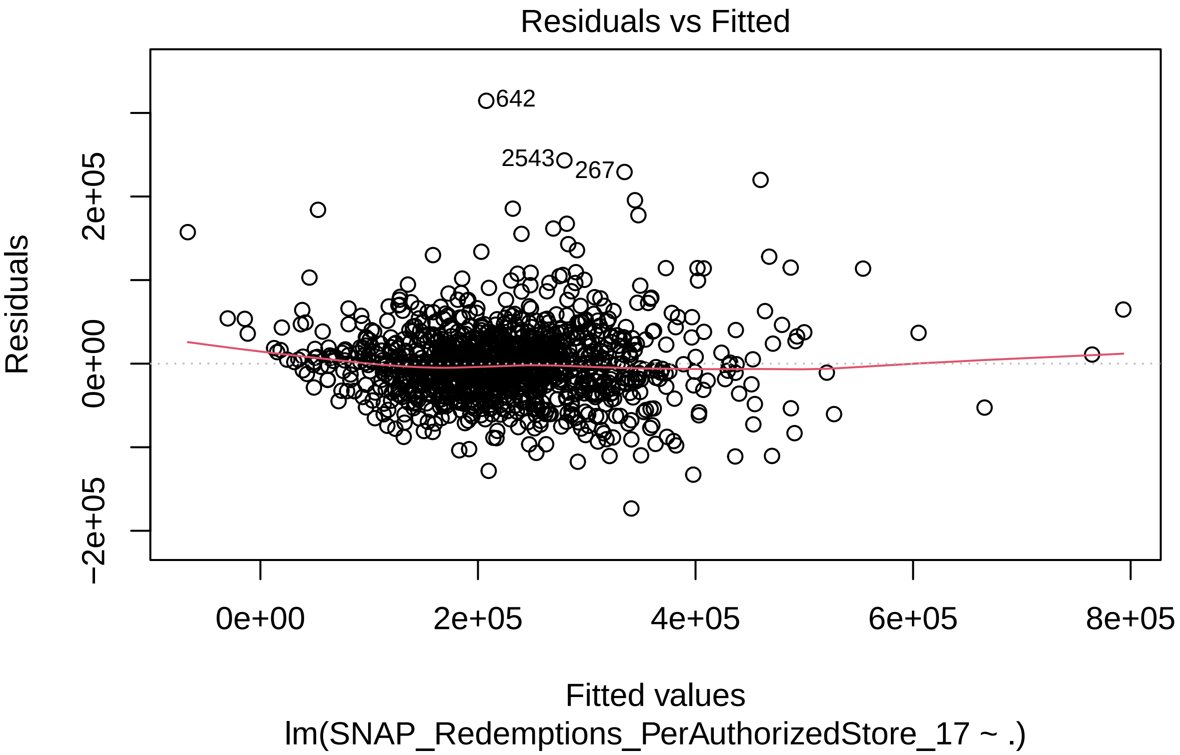


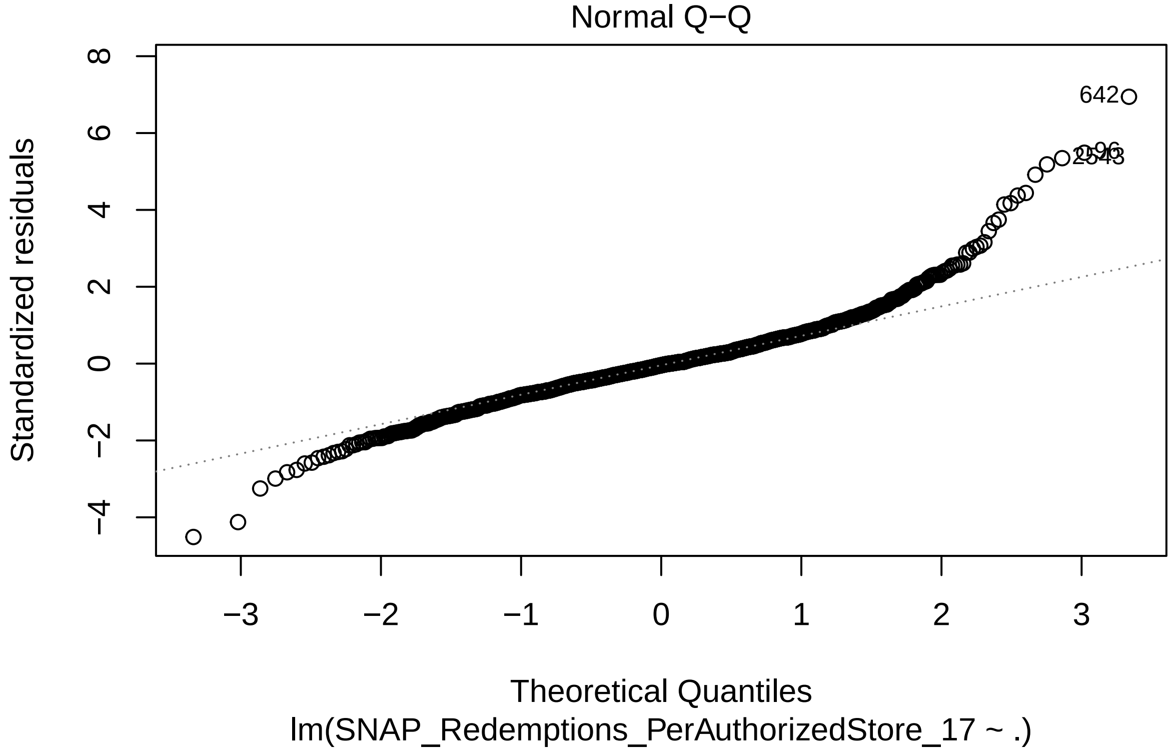
Normality tests:

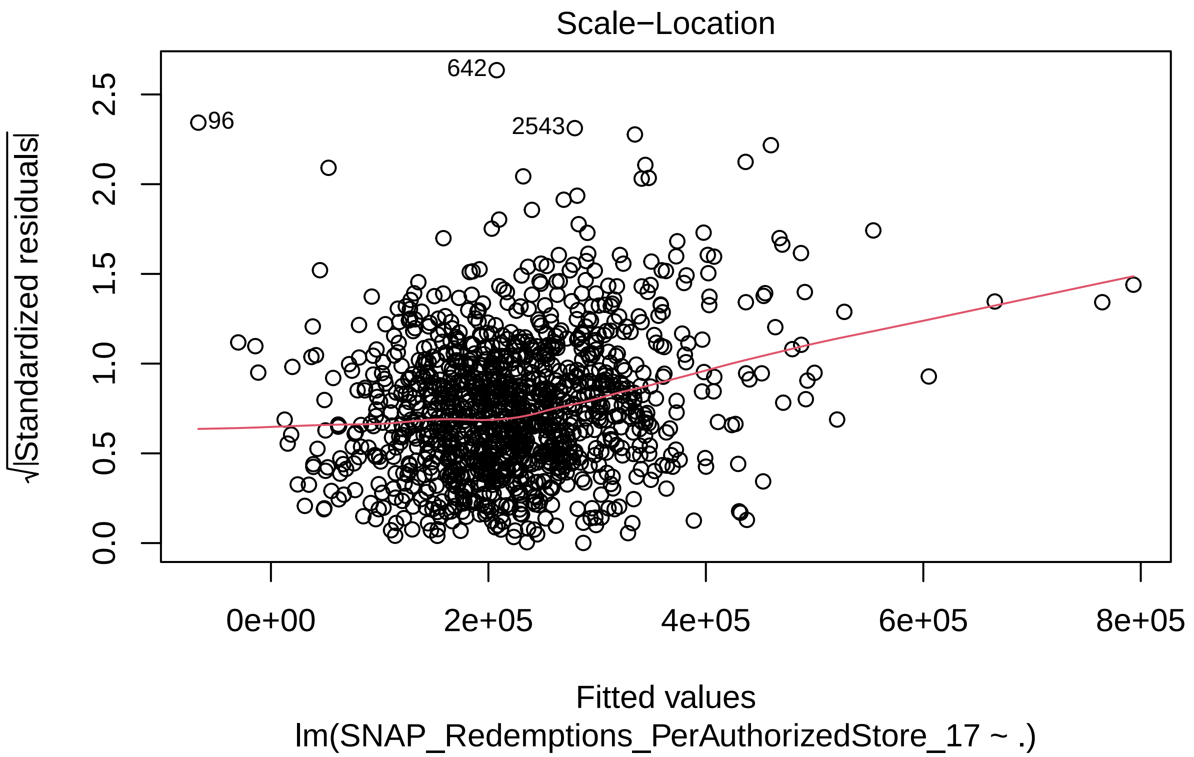
```
##
##  Shapiro-Wilk normality test
##
## data:  atlas$SNAP_Redemptions_PerAuthorizedStore_17
## W = 0.96158, p-value < 2.2e-16
```

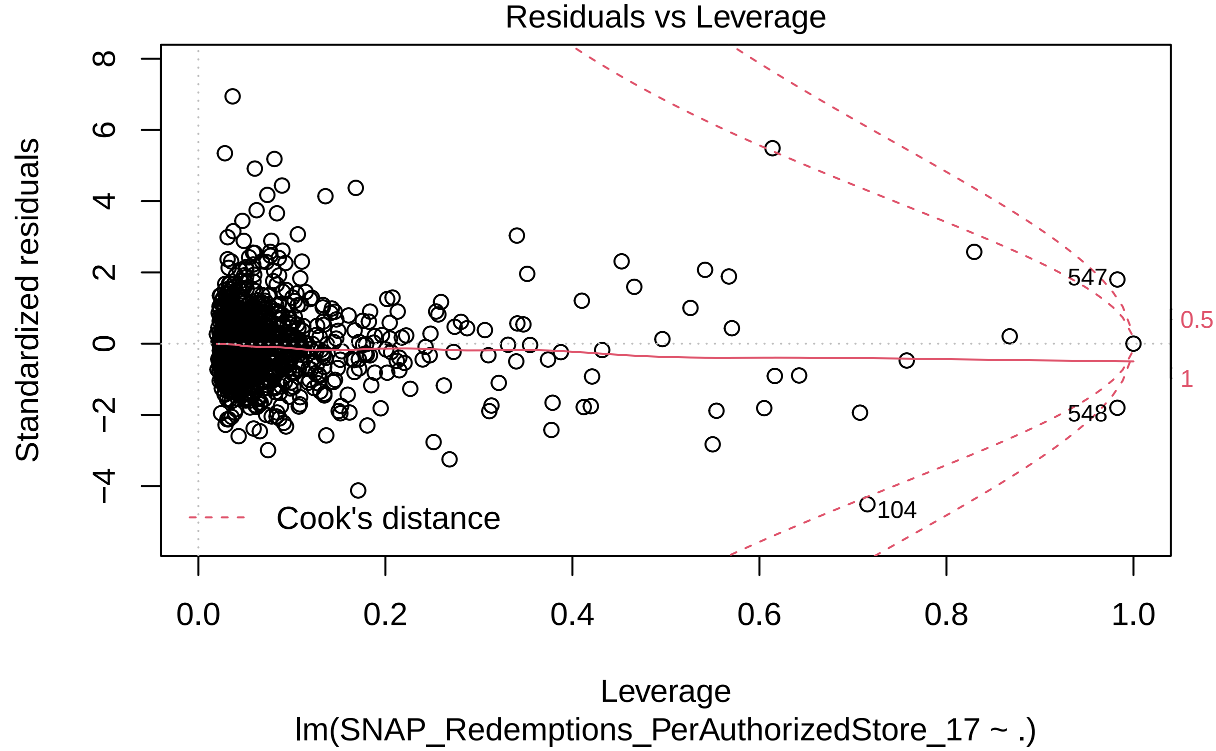
```
##
##  Shapiro-Wilk normality test
##
## data:  sqrt(atlas$SNAP_Redemptions_PerAuthorizedStore_17)
## W = 0.99458, p-value = 7.526e-09
```

Quartile plot for a full linear model:


Residuals vs Fitted
Residuals
−2e+05
0e+00
2e+05
642
2543
267
0e+00
2e+05
4e+05
6e+05
8e+05
Fitted values
lm(SNAP_Redemptions_PerAuthorizedStore_17 ~ .)


Normal Q−Q
Standardized residuals
−4
−2
0
2
4
6
8
642
296
2543
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(SNAP_Redemptions_PerAuthorizedStore_17 ~ .)





For multilinear regression, ridge regression, and LASSO, nonlinearity in the data is quite possible given the complex system being modeled, so we may assume the models will have high bias. However, correlation of the error terms should be less of an issue because observations are related spatially and not temporally. We also see some heteroscedasticity present in the data from the scale-location plot in addition to the expected non-normality in the response. Given the large number of observations and probable nonlinearity, we will also have to account for outliers and high-leverage points in these models. Given that all the variables are somewhat interrelated, we might conjecture that ridge regression will outperform LASSO, since most variables should be similarly important. We would also want to research what the exact qualifications for SNAP are, as there might be variables in the dataset that measure something similar, which would then dominate the models without providing meaningful information.

Of the 78 variables measured between 2015 and 2017, some can be dropped because they measure the same quantities as counts and percentages. Wherever possible, we retain variables measured as percentages, which are not impacted by county population in the way that count variables would be. Accordingly, our ratio of n to p-with the exception of variables measured on the state level- is acceptably large. This is important for tree-based models, especially so if we choose to construct a random forest regressor. However, if we are unable to eliminate many variables or have a particularly high concentration of state-level variables, linear models may be preferable because of their lack of flexibility- the random forest might overfit with more predictors.

Results

`tidyverse` was used to preprocess the data. First, the variables with the most missing observations were removed. After removing the number of WIC redemptions per capita in 2016 and the percent of active high schoolers in 2017, all observations with missing values were removed. 2380 of 3142 original observations were retained this way, approximately 76% of the data. Next, ID variables were removed. A categorical variable denoting state would have had too many levels for linear regression, so geographical information was recoded by census region instead of by state and county. A standard 0.75/0.25 train/test split was created split using `rsample`. `tidymodels` was used to construct the models. First, a recipe was created for preprocessing of the training and test sets (the aforementioned data cleaning was done outside of a `recipes` framework due to some issues with `step_select` and `step_mutate`). Dummy variables were created, predictors with zero variance were removed, highly correlated predictors were removed, and all predictors were normalized. The four models were then fitted. A full linear model was run as a baseline. Feature selection was not performed because additional regularization models were also fitted. Regularization was performed via ridge regression and LASSO, which were tuned over the shrinkage penalty λ. The final ridge regressor had $\lambda = 0$, and the final LASSO model had $\lambda = 769$. Finally, random forest regression was run and tuned over the size of the random subset of predictors at each split and over the minimum number of observations required to split a node. Tuning was performed using 10-fold cross-validation, and using 100 trees per model, for efficiency. The final random forest regressor had an optimal `mtry=35` and `min_n=2`, and was calculated with 1000 trees for precision.

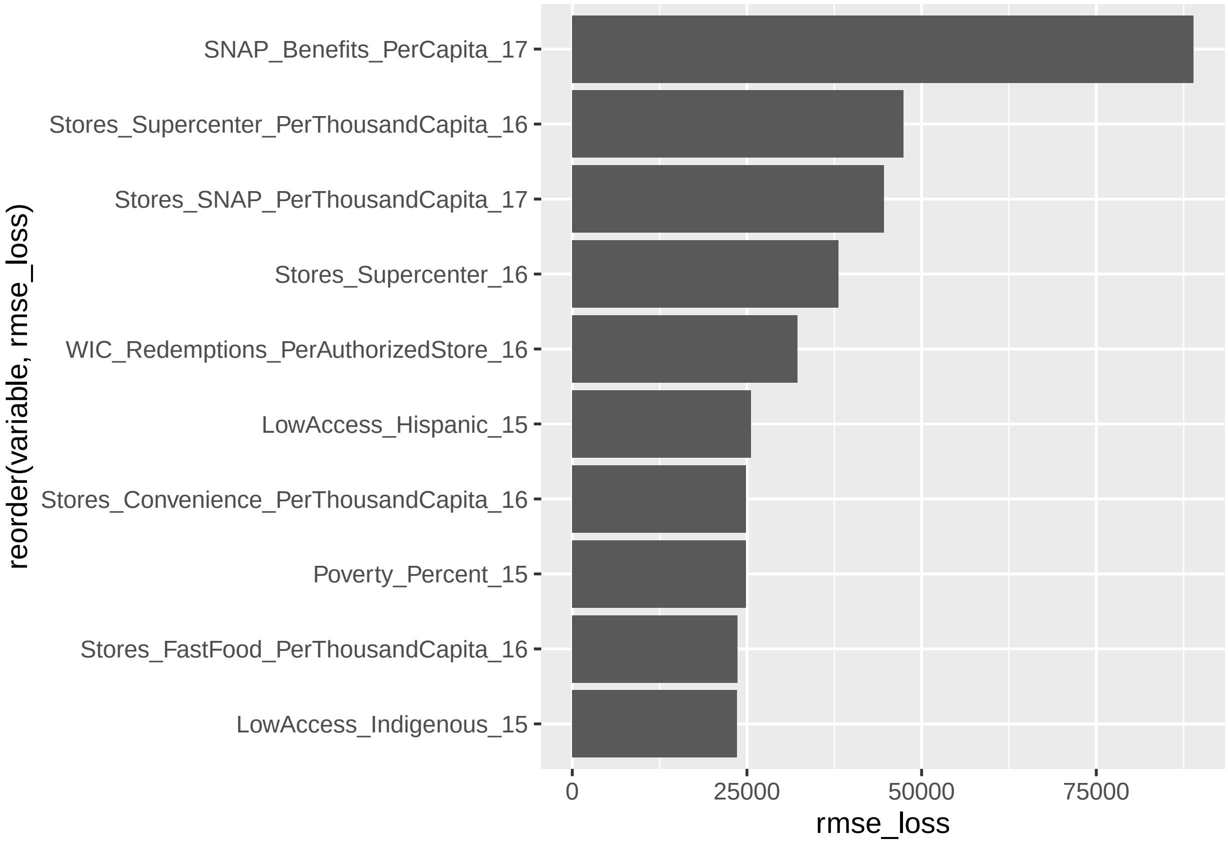
```
##   Metric Full Linear Regression Ridge Regression        LASSO
## 1   RMSE           4.938595e+04     4.959929e+04 4.933487e+04
## 2    RSQ           7.453464e-01     7.465611e-01 7.471837e-01
## 3    MAE           3.704440e+04     3.762208e+04 3.683807e+04
## 4     SE           2.498477e+03     1.923979e+03 1.536688e+03
##   Random Forest Regression
## 1             4.548128e+04
## 2             7.860423e-01
## 3             3.420767e+04
## 4             1.814135e+03
```

The full linear model, ridge regression, and LASSO had very similar RMSE, but the test error for the full linear model was estimated to be very unstable, since it had the highest standard error. LASSO had the lowest standard error of any model, but its RMSE was not competitive with the random forest. The clear winner is the random forest model, with the lowest RMSE and the second lowest standard error. Furthermore, random forests does not violate any model assumptions because the sample is representative.

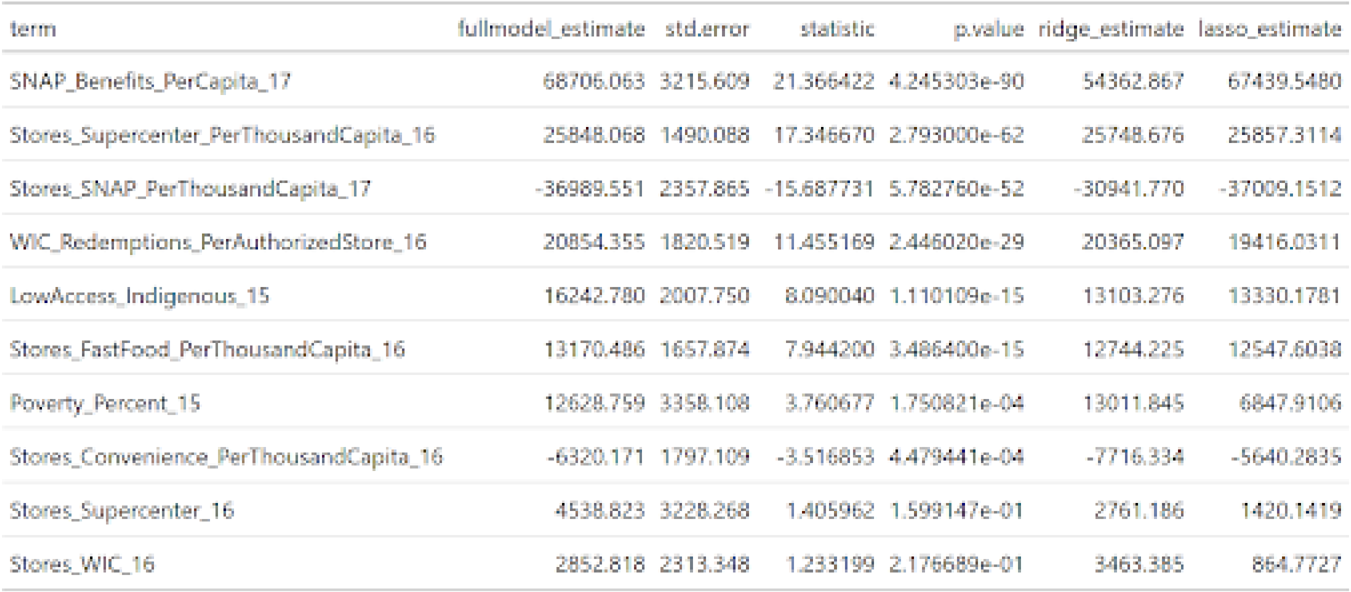
We can look at some sample predictions:

```
## # A tibble: 10 x 2
##      .pred SNAP_Redemptions_PerAuthorizedStore_17
##      <dbl>                                  <dbl>
##  1 220923.                                116757.
##  2 181330.                                        157910.
##  3 137800.                                        150181.
##  4 341310.                                        363339.
##  5 227444.                                        284061.
##  6  72382.                                         84816.
##  7 177034.                                        150364.
##  8 236475.                                        299664.
##  9 172761.                                        251679.
## 10 216497.                                        252481.
```

So although the RMSE of the random forest is high, it is more effective than the linear models in predicting the response relative to its large scale (from 8375 to 858018). In this context, the R^2 value of 0.78 might give a better intuitive feel for this being a fairly good fit. Feature importance for the random forest is plotted below, and ranks the variables by how much their removal would increase model error:



The most important predictors were `SNAP_Benefits_PerCapita_17`, `Stores_Supercenter_PerThousandCapita_16`, `Stores_SNAP_PerThousandCapita_17`, `Stores_Supercenter_16`, and `WIC_Redemptions_PerAuthorizedStore_16`. The relative importance of these predictors is somewhat reflected in the linear models, as 7 of the 10 most important predictors in the random forest were significant ($p < .05$) in the full linear model, and LASSO retained 9 of those 10 predictors.



term	fullmodel_estimate	std.error	statistic	p.value	ridge_estimate	lasso_estimate
SNAP_Benefits_PerCapita_17	68706.063	3215.609	21.366422	4.245303e-90	54362.867	67439.5480
Stores_Supercenter_PerThousandCapita_16	25848.068	1490.088	17.346670	2.793000e-62	25748.676	25857.3114
Stores_SNAP_PerThousandCapita_17	-36989.551	2357.865	-15.687731	5.782760e-52	-30941.770	-37009.1512
WIC_Redemptions_PerAuthorizedStore_16	20854.355	1820.519	11.455169	2.446020e-29	20365.097	19416.0311
LowAccess_Indigenous_15	16242.780	2007.750	8.090040	1.110109e-15	13103.276	13330.1781
Stores_FastFood_PerThousandCapita_16	13170.486	1657.874	7.944200	3.486400e-15	12744.225	12547.6038
Poverty_Percent_15	12628.759	3358.108	3.760677	1.750821e-04	13011.845	6847.9106
Stores_Convenience_PerThousandCapita_16	-6320.171	1797.109	-3.516853	4.479441e-04	-7716.334	-5640.2835
Stores_Supercenter_16	4538.823	3228.268	1.405962	1.599147e-01	2761.186	1420.1419
Stores_WIC_16	2852.818	2313.348	1.233199	2.176689e-01	3463.385	864.7727

Furthermore, since these predictors have been normalized, we may interpret the coefficients relative to each other. All of the models show that `SNAP_Benefits_PerCapita_17` is the most important predictor, as it has the highest magnitude and the smallest p-value. Comparing coefficients across models also confirms that the regularization methods tended to have smaller coefficients than the full linear model, which is consistent with ridge and LASSO's functionality as shrinkage methods related to simple linear regression. While we cannot hypothesize the direction that a random forest splits along a given feature, the linear models converged on a similar set of important coefficients. Based on the linear models, counties with a higher per-capita SNAP benefits generally have higher average SNAP redemptions per authorized store. Relative changes in per-capita SNAP benefits will tend to have an outsized influence on predicted response values relative to the other predictors (and the remaining coefficients could be interpreted similarly).

Discussion

The present analysis modeled average SNAP redemption as a function of socioeconomic and demographic predictors at the county level to test whether those predictors explained variance in average redemption above and beyond traditional explanations like income and race. Multilinear regression, ridge regression, LASSO, and random forests were used to evaluate these relationships. Random forests made the most accurate predictions, and its results are discussed in the context of the directional predictions generated by the linear models.

A random forest model identified per-capita SNAP benefits, supercenter stores per 1000 capita, SNAP stores per 1000 capita, WIC redemptions per store, percentage of Hispanic residents with low access to stores, and poverty rate as the most important predictors. Taken together, these predictors indicate that the number of SNAP-eligible stores best explains average SNAP redemption, above and beyond the demographic variables included in the FEA data. In the linear models, the number of SNAP stores is negatively associated with average redemption, suggesting that spending is generally lower with more stores available to county residents. However, the linear relationship between the number of supercenters per 1000 capita and average redemption is positive, suggesting that SNAP spending behavior may differ depending on the types of stores at which participants redeem benefits. That is, people may be spending more at supercenters in a reversal of the overall relationship between number of stores and average redemption. These relationships reflect previous work in the food access literature, which argues that supercenters often provide a wider range of products to food-insecure residents than are available at local convenience and grocery stores (Neff et al., 2009). However, supercenters are less likely to be located in or near economically and racially marginalized communities, suggesting that these relationships may be moderated by demographic variables beyond the scope of the present analysis.

While the FEA is a rich source of data consolidated from multiple government agencies, the manner in which it measured race kept us from testing for clearer direct and moderated relationships between race and the response. Of the important predictors identified by both the linear and nonlinear models, only one racial predictor was identified. The linear models identified the percentage of Indigenous residents with low access to stores, while the random forest identified the percentage of Hispanic residents with low access to stores. Given the overall pattern of findings, these data would suggest that racial disparities in store access are poor predictors of average redemption amount. However, past work has demonstrated striking racial disparities in food security and access, prompting alternative explanations of race's minimal role in our model. In each county, the FEA records the percentage of each racial minority group that is low-income and low-income-and-low-store-access. While some of these variables reached significance or made substantial contributions to the models, they omit the overall percentage of residents who identify as a specific minority group. We expect that a more comprehensive race variable that examines full minority populations rather than subsets would have yielded different results and been more useful in testing moderation for other relationships.

Average redemption amount may also not have been an effective response variable when exploring racial disparities. SNAP enrollees in different racial groups may have similar redemption amounts given similar benefits received, but simply struggle more to access stores. Past work suggests that it may be especially difficult for members of marginalized groups to access stores with higher-quality offerings like supermarkets and supercenters (Neff et al., 2009). Consequently, the main effects observed in our linear models may be moderated by race, underscoring the importance of integrating these data with more comprehensive race data for future work.

Poverty rate, however, was a clearly measured variable with an interpretable relationship to average redemption. Given that effect direction is unavailable from random forests, the output of other models suggests that poverty rate is positively related to average redemption. If average redemption is higher in counties with more residents under the poverty line, those residents may be using more of their benefits regardless of the stores available to them, motivating a test of whether the relationship between number of stores and average redemption is weaker in poorer counties. Such a finding would help define the limits of an approach centered on access to stores, and license a focus on helping people acquire the resources necessary to purchase food.

Participation in WIC also predicted average redemption, and highlights a productive direction for future research. If there is a relationship between use of SNAP and use of WIC (and potentially other benefits programs), it may be productive to construct profiles from benefits usage data to get a broad sense of how people enrolled in different programs allocate those funds.

Code Appendix

```r
#loading, cleaning, and formatting data into a single data frame

#packages
library(tidymodels)
library(stringr)
library(readxl)
tidymodels_prefer()

#setting wd
#setwd('your working directory')

#loading data
atlas <- "FoodEnvironmentAtlas/FoodEnvironmentAtlas.xls"

#extracting sheets to a list
sheet_names <- excel_sheets(path = atlas)
atlas <- lapply(sheet_names, function(x) read_excel(path = atlas, sheet = x))
```

```r
#picking specific sheets that we want
atlas <- atlas[c(3,5:13)]

#remove " County" from supplemental county data to facilitate joining
atlas[[1]] <- atlas[[1]] %>%
  mutate(County = str_replace_all(County, c(
    " Borough" = "",
    " Census Area" = "",
    " city" = "",
    " County" = "",
    " Municipality" = "",
    " Parish" = ""
  ))) %>%
  mutate(County = str_replace(County, " City and", "")) %>%
  mutate(County = ifelse(FIPS == 22059, "La Salle", County))

#joining sheets
init <- F
temp <- NULL
for (sheet in atlas) {
  if (init == F) {
    temp <- sheet
    init <- T
  } else {
    temp <- full_join(temp, sheet, by = c("FIPS","County"))
  }
}
atlas <- temp

#subsetting the features
atlas <- atlas %>%
  select(FIPS, County, State.x, contains(c("15","16","17"))) %>%
  select(!contains(c("10_15","11_16","12_14","12_15","12_17","14_16","14_17")))

#cleaning outliers
atlas <- atlas %>%
  filter(is.na(FIPS) == F & FIPS != "02270" & FIPS != 46113 & FIPS != 51515)

#02270 changed to 02158
#22059 named wrong
#46113 changed to 46102
#51515 changed into 51019

#renaming variables to be more legible
colnames(atlas) <- c(
  "FIPS",
  "County",
  "State",
  "Population_15",
  "LowAccess_15",
  "LowAccess_Percent_15",
  "LowAccess_LowIncome_15",
  "LowAccess_LowIncome_Percent_15",
  "LowAccess_NoCar_Houses_15",
  "LowAccess_NoCar_Houses_Percent_15",
  "LowAccess_SNAPHouses_15",
  "LowAccess_SNAPHouses_Percent_15",
  "LowAccess_Children_15",
  "LowAccess_Children_Percent_15",
  "LowAccess_Seniors_15",
  "LowAccess_Seniors_Percent_15",
  "LowAccess_White_15",
  "LowAccess_White_Percent_15",
  "LowAccess_Black_15",
  "LowAccess_Black_Percent_15",
  "LowAccess_Hispanic_15",
  "LowAccess_Hispanic_Percent_15",
  "LowAccess_Asian_15",
  "LowAccess_Asian_Percent_15",
  "LowAccess_Indigenous_15",
  "LowAccess_Indigenous_Percent_15",
  "LowAccess_HawaiianPI_15",
  "LowAccess_HawaiianPI_Percent_15",
  "LowAccess_Multiracial_15",
  "LowAccess_Multiracial_Percent_15",
  "FreeLunch_Children_Percent_15",
  "ReducedPriceLunch_Children_Percent_15",
  "FDPIR_Sites_15",
  "Houses_FoodInsecure_AveragePercent_15_17",
  "Houses_VeryLowFoodSecurity_AveragePercent_15_17",
  "FSP_Available_15",
  "HouseholdIncome_Median_15",
  "Poverty_Percent_15",
  "Poverty_Children_Percent_15",
  "Population_16",
  "Stores_Grocery_16",
  "Stores_Grocery_PerThousandCapita_16",
  "Stores_Supercenter_16",
  "Stores_Supercenter_PerThousandCapita_16",
  "Stores_Convenience_16",
  "Stores_Convenience_PerThousandCapita_16",
  "Stores_SpecialtyFood_16",
  "Stores_SpecialtyFood_PerThousandCapita_16",
  "Stores_WIC_16",
  "Stores_WIC_PerThousandCapita_16",
  "Stores_FastFood_16",
  "Stores_FastFood_PerThousandCapita_16",
  "Stores_FullService_16",
  "Stores_FullService_PerThousandCapita_16",
  "SNAP_Eligible_ParticipatingPercent_16",
  "SNAP_OnlineApplication_Available_16",
  "SNAP_CAP_Available_16",
  "SNAP_BroadBasedCategoricalEligibility_Available_16",
  "SNAP_SimplifiedReporting_Available_16",
  "WIC_Redemptions_PerCapita_16",
  "WIC_Redemptions_PerAuthorizedStore_16",
  "WIC_InfantsChildren_ParticipatingPercent",
  "WIC_Women_ParticipatingPercent",
  "RecreationFitnessFacilities_16",
  "RecreationFitnessFacilities_PerThousandCapita_16",
  "Population_17",
  "Stores_SNAP_17",
  "Stores_SNAP_PerThousandCapita_17",
  "SNAP_Redemptions_PerAuthorizedStore_17",
  "SNAP_Participants_Percent_17",
  "SNAP_Benefits_PerCapita_17",
  "NSLP_Children_ParticipatingPercent_17",
  "SBP_Children_ParticipatingPercent_17",
  "SFSP_Children_ParticipatingPercent_17",
  "WIC_Participants_Percent_17",
  "CACFP_Eligible_Percent_17", #unclear whether this is participants or eligible people,
  "Obesity_Adults_Percent_17", #guessing eligible by trends in other variables
  "Active_HighSchoolers_Percent_17"
)
```

```
# scatterplots comparing some of the predictors that might influence the response the most
# response in relation to the estimated population in 2015
atlas %>% ggplot(aes(x = Population_15, y = SNAP_Redemptions_PerAuthorizedStore_17)) +
  geom_point(na.rm = T, alpha = 0.3)
```

```
# response in relation to the population that have low access in 2015
atlas %>% ggplot(aes(x = LowAccess_15, y = SNAP_Redemptions_PerAuthorizedStore_17)) +
  geom_point(na.rm = T, alpha = 0.3)
```

```
# population of low access in relation to the whole population in 2015
atlas %>% ggplot(aes(x = Population_15, y = LowAccess_15)) +
  geom_point(na.rm = T, alpha = 0.3)
```

```
# distribution and boxplot of the response
atlas %>% ggplot(aes(x = SNAP_Redemptions_PerAuthorizedStore_17)) + geom_histogram(binwidth = 20000)
atlas %>% ggplot(aes(x = SNAP_Redemptions_PerAuthorizedStore_17)) + geom_boxplot()
```

```
# distribution of the square root transformation of the response
atlas %>% ggplot(aes(x = sqrt(SNAP_Redemptions_PerAuthorizedStore_17))) + geom_histogram(binwidth = 20)
```

```
#normality tests
shapiro.test(atlas$SNAP_Redemptions_PerAuthorizedStore_17)
shapiro.test(sqrt(atlas$SNAP_Redemptions_PerAuthorizedStore_17))
```

```
#quartile plot
df <- atlas %>%
  select(!c(FIPS,County))

plot(lm(SNAP_Redemptions_PerAuthorizedStore_17 ~ ., data = df))
```

```
set.seed(1)

#initial split and preprocessing
df <- atlas %>%
  mutate(
    Census_Region = case_when( #encoding geographic information
      State %in% c(
        "Connecticut",
        "Maine",
        "Massachusetts",
        "New Hampshire",
        "Rhode Island",
        "Vermont",
        "New Jersey",
        "New York",
        "Pennsylvania"
      ) ~ "Northeast",
      State %in% c(
        "Indiana",
        "Illinois",
        "Michigan",
        "Ohio",
        "Wisconsin",
        "Iowa",
        "Kansas",
        "Minnesota",
        "Missouri",
        "Nebraska",
        "North Dakota",
        "South Dakota"
      ) ~ "Midwest",
      State %in% c(
        "Delaware",
        "Florida",
        "Georgia",
        "Maryland",
        "North Carolina",
        "South Carolina",
        "Virginia",
        "West Virginia",
        "Alabama",
        "Kentucky",
        "Mississippi",
        "Tennessee",
        "Arkansas",
        "Louisiana",
        "Oklahoma",
        "Texas",
        "District of Columbia"
      ) ~ "South",
      State %in% c(
        "Arizona",
        "Colorado",
        "Idaho",
        "New Mexico",
        "Montana",
        "Utah",
        "Nevada",
        "Wyoming",
        "Alaska",
        "California",
        "Hawaii",
        "Oregon",
        "Washington"
      ) ~ "West"
    )
  ) %>%
  select(!c(FIPS, County, State, WIC_Redemptions_PerCapita_16, Active_HighSchoolers_Percent_17)) %>%
  drop_na() %>% #id vars/vars w/ large number of levels/nas
  initial_split() #0.75 proportion, since i don't have a particular justification otherwise...

train <- training(df)
test <- testing(df)
```

```
#workflow
rec <- train %>% #i'm not doing any stepwise regression since i'm also running two regularized models,
  recipe(SNAP_Redemptions_PerAuthorizedStore_17 ~ .) %>% #this is just the baseline
  step_zv(all_predictors()) %>% #the code, counterintuitively, breaks if i place this after step_dummy
  step_dummy(all_nominal_predictors()) %>%
  step_corr(all_predictors()) %>%
  step_normalize(all_predictors()) %>%
  prep()

#base workflow
wf <- workflow() %>%
  add_recipe(rec)
```

```
set.seed (2)

#full linear model
model_lm <- linear_reg() %>%
  set_engine("lm")

#linear workflow
wf_lm <- wf %>%
  add_model(model_lm)

#standard error
se_lm <- fit_resamples(wf_lm, vfold_cv(train)) %>%
  collect_metrics() %>%
  filter(.metric == "rmse") %>%
  select(std_err) %>%
  as.numeric()

#fitting model
fit_lm <- fit(wf_lm, train)

#predictions and true values
res_lm <- predict(fit_lm, test) %>%
  bind_cols(select(test, SNAP_Redemptions_PerAuthorizedStore_17))
```

```
set.seed(3)

#ridge regression
```

```
#tuning model
model_tune <- linear_reg(penalty = tune(), mixture = 0) %>%
  set_engine("glmnet", num.threads = 12)

#tuning workflow
wf_tune <- wf %>%
  add_model(model_tune)

#tuning by lambda
tune <- tune_grid(
  wf_tune,
  resamples = vfold_cv(train),
  grid = grid_regular(penalty(range = c(-10,20)), levels = 150)
)

#best lambda = 0
best_tune <- select_best(tune, "rmse")$penalty

#standard error
se_ridge <- tune %>%
  collect_metrics() %>%
  filter(penalty == best_tune & .metric == "rmse") %>%
  select(std_err) %>%
  as.numeric()

#tuned model
model_ridge <- linear_reg(penalty = best_tune, mixture = 0) %>%
  set_engine("glmnet")

#ridge workflow
wf_ridge <- wf %>%
  add_model(model_ridge)

#fitting model
fit_ridge <- fit(wf_ridge, train)

#predictions and true values
res_ridge <- predict(fit_ridge, test) %>%
  bind_cols(select(test, SNAP_Redemptions_PerAuthorizedStore_17))
```

```
set.seed(4)

#lasso
#tuning model
model_tune <- linear_reg(penalty = tune(), mixture = 1) %>%
  set_engine("glmnet", num.threads = 12)

#tuning workflow
wf_tune <- wf %>%
  add_model(model_tune)

#tuning by lambda
tune <- tune_grid(
  wf_tune,
  resamples = vfold_cv(train),
  grid = grid_regular(penalty(range = c(-10,20)), levels = 150)
)

#best lambda = 769
best_tune <- select_best(tune, "rmse")$penalty

#standard error
se_lasso <- tune %>%
  collect_metrics() %>%
  filter(penalty == best_tune & .metric == "rmse") %>%
  select(std_err) %>%
  as.numeric()

#tuned model
model_lasso <- linear_reg(penalty = best_tune, mixture = 1) %>%
  set_engine("glmnet")

#lasso workflow
wf_lasso <- wf %>%
  add_model(model_lasso)

#fitting model
fit_lasso <- fit(wf_lasso, train)

#predictions and true values
res_lasso <- predict(fit_lasso, test) %>%
  bind_cols(select(test, SNAP_Redemptions_PerAuthorizedStore_17))
```

```
set.seed(5)

#random forest
#tuning model
model_tune <- rand_forest(mode = "regression", mtry = tune(), trees = 100, min_n = tune()) %>%
  set_engine("randomForest", num.threads = 12)

#tuning workflow
wf_tune <- wf %>%
  add_model(model_tune)

#tuning by mtry, min_n
tune <- tune_grid(
  wf_tune,
  resamples = vfold_cv(train),
  grid = grid_regular(parameters(mtry(range = c(39,40)), min_n(range = c(2, 3))), levels = 2)
) #this grid is much smaller to save time, the full range WAS tested

#best mtry = 39, best min_n = 2
best_tune <- select_best(tune, "rmse")$mtry
best_tune_ <- select_best(tune, "rmse")$min_n

#standard error
se_rf <- tune %>%
  collect_metrics() %>%
```

```r
  filter(mtry == best_tune & min_n == best_tune_ & .metric == "rmse") %>%
  select(std_err) %>%
  as.numeric()

#tuned model
model_rf <- rand_forest(mode = "regression", mtry = best_tune, trees = 1000, min_n = best_tune_) %>%
  set_engine("randomForest", num.threads = 12)

#random forest workflow
wf_rf <- wf %>%
  add_model(model_rf)

#fitting model
fit_rf <- fit(wf_rf, train)

#predictions and true values
res_rf <- predict(fit_rf, test) %>%
  bind_cols(select(test, SNAP_Redemptions_PerAuthorizedStore_17))
```

```r
#compiling metrics for each model
metrics <- data.frame(Metric = c("RMSE", "RSQ", "MAE","SE"))

models <- list(res_lm, res_ridge, res_lasso, res_rf)

se <- list(se_lm, se_ridge, se_lasso, se_rf)

rep <- 0

#making a nice table
for (mod in models) {
  rep <- rep + 1
  temp = c(metrics(mod, .pred, SNAP_Redemptions_PerAuthorizedStore_17)$.estimate, se[[rep]])
  metrics <- cbind(metrics,temp)
}

colnames(metrics) <- c(
  "Metric",
  "Full Linear Regression",
  "Ridge Regression",
  "LASSO",
  "Random Forest Regression"
)

metrics
```

```r
set.seed(6)

#sample predictions
sample_n(res_rf, 10)
```

```r
library(DALEXtra)

#feature importance
#making an explainer, whatever that is
```

```r
explainer_rf <-
  explain_tidymodels(
    fit_rf,
    data = train,
    y = train$SNAP_Redemptions_PerAuthorizedStore_17,
    label = "random forest",
    verbose = FALSE
  )

#calculating permutation importance
vip_rf <- model_parts(explainer_rf, loss_function = loss_root_mean_square)

#plotting permutation importance
vip_rf %>%
  group_by(variable) %>%
  summarize(rmse_loss = mean(dropout_loss)) %>%
  dplyr::filter(variable != "_baseline_") %>%
  arrange(desc(rmse_loss)) %>%
  head(10) %>%
  ggplot(aes(rmse_loss,reorder(variable, rmse_loss)))+geom_col()
```

```r
#for exploratoratory data analysis
#missing values
summary(atlas)[7,]

#determining high magnitude correlations
correlations <- atlas %>%
  select(where(is.numeric)) %>%
  select(!c(
    FreeLunch_Children_Percent_15,
    ReducedPriceLunch_Children_Percent_15,
    FSP_Available_15,
    WIC_Redemptions_PerCapita_16,
    SNAP_Redemptions_PerAuthorizedStore_17,
    Stores_WIC_PerThousandCapita_16,
    Active_HighSchoolers_Percent_17,
    WIC_Redemptions_PerAuthorizedStore_16,
    Stores_WIC_16
  )) %>%
  drop_na() %>%
  cor()

for(i in correlations) {
  if(abs(i) >= 0.8 & abs(i) < 1 & is.na(abs(i)) == F) {
    print(i)
    print(which(correlations == i, arr.ind = TRUE))
  }
}
```

```r
#deciding what variables need to be dropped

#counting missing values
df <- atlas %>%
  drop_na(SNAP_Redemptions_PerAuthorizedStore_17)
```

```
nas <- c()
for (col in colnames(df)){
  temp = df %>%
    select(col) %>%
    filter(is.na(.data[[col]]) == T) %>%
    nrow()
  nas <- c(nas,temp)
}
nas <- nas %>%
  bind_cols(colnames(df))

#looks like only Active_HighSchoolers_Percent_17 and WIC_Redemptions_PerCapita_16
#really need to be dropped, everything else has <300 nas
```

```
#creating coefficient table
lr_table <- tidy(fit_lm) %>% tibble()
lasso_table <- tidy(fit_lasso) %>% tibble() %>% dplyr::filter(estimate != 0)
lasso_table <- lasso_table %>% rename(lasso_estimate = estimate) %>% select(-penalty)
lr_lasso_table <- left_join(x = lasso_table, y = lr_table, "term")
ridge_table <- tidy(fit_ridge) %>% tibble() %>% select(-penalty) %>% rename(ridge_estimate = estimate)

linear_models <- left_join(x = lr_lasso_table, y = ridge_table, "term") %>% rename(fullmodel_estimate =
```

```
#creating coefficient table
linear_rf_shared <- linear_models %>%
  dplyr::filter(term %in% c("SNAP_Benefits_PerCapita_17",
                            "Stores_Supercenter_PerThousandCapita_16",
                            "Stores_SNAP_PerThousandCapita_17",
                            "Stores_Supercenter_16",
                            "WIC_Redemptions_PerAuthorizedStore_16",
                            "Poverty_Percent_15",
                            "Stores_Convenience_PerThousandCapita_16",
                            "Stores_FastFood_PerThousandCapita_16",
                            "Stores_WIC_16",
                            "LowAccess_Indigenous_15"))
linear_rf_shared <- linear_rf_shared[,c(1, 3, 4, 5, 6, 7, 2)]
linear_rf_shared <- linear_rf_shared %>% arrange(p.value)
#library(gt)
#gt(linear_rf_shared)
```

```
#variables that lasso dropped
temp <- fit_lasso %>%
  extract_fit_parsnip() %>%
  tidy() %>%
  filter(term != "(Intercept)" & estimate !=  0)

colnames(juice(rec)) %>%
  setdiff(temp$term)
```

References

Bower, K. M., Thorpe Jr, R. J., Rohde, C., & Gaskin, D. J. (2014). The intersection of neighborhood racial segregation, poverty, and urbanicity and its impact on food store availability in the United States [online]. Preventive medicine, 58, 33-39. Available at https://www.sciencedirect.com/science/article/pii/S0091743513003988

Center on Budget and Policy Priorities (2021). States Are Using Much-Needed Temporary Flexibility in SNAP to Respond to COVID-19 Challenges [online]. Available at (https://www.cbpp.org/research/food-assistance/states-are-using-much-needed-temporary-flexibility-in-snap-to-respond-to).

Hall, L. (2021). A Closer Look at Who Benefits from Snap: State-by-State Fact Sheets [online]. Available at https://www.cbpp.org/research/food-assistance/a-closer-look-at-who-benefits-from-snap-state-by-state-fact-sheets.

Loveless, T.A. (2020). Supplemental Nutrition Assistance Program (SNAP) Receipt for Households: 2018 [online]. Available at https://www.census.gov/content/dam/Census/library/publications/2020/demo/acsbr 20-01.pdf.

Matthew P. Rabbitt, Michael D. Smith, and Alisha Coleman-Jensen (2016). Food Security Among Hispanic Adults in the United States, 2011-2014 [online]. Available at https://www.ers.usda.gov/publications/pub-details/?pubid=44083.

Neff, R. A., Palmer, A. M., McKenzie, S. E., & Lawrence, R. S. (2009). Food systems and public health disparities. Journal of Hunger & Environmental Nutrition [online], 4(3-4), 282-314.

Sharma, S. V., Hernandez, D. C., Hoelscher, D. M., & Yaroch, A. L. (2015). Multidisciplinary approaches to address food insecurity and nutrition among youth and their families. Journal of Applied Research on Children: Informing Policy for Children at Risk [online], 6(2), 1. Available at https://digitalcommons.library.tmc.edu/childrenatrisk/vol6/iss2/1/

U.S. Census Bureau (2020). Census Bureau Releases New Estimates on America's Families and Living Arrangements [online]. Available at https://www.census.gov/newsroom/press-releases/2020/estimates-families-living-arrangements.html.

U.S. Census Bureau (2021). U.S. Census Bureau Quickfacts: United States. Available at https://www.census .gov/quickfacts/US.

USDA ERS (2020). About the Atlas [online]. Available at https://www.ers.usda.gov/data-products/food-environment-atlas/about-the-atlas/.

USDA FNS (2021) Characteristics of Supplemental Nutrition Assistance Program Households: Fiscal Year 2019 [online]. Available at https://www.fns.usda.gov/snap/characteristics-snap-households-fy-2019

USDA FNS (2021). SNAP Data Tables [online]. Available at https://www.fns.usda.gov/pd/supplemental-nutrition-assistance-program-snap.