**Team Avengers**  
**DATS 6101: Summary Report**  
**Professor Divya Narula**  
**April 30, 2025**

# Food Deserts in the United States

## Introduction

Food deserts are communities where residents have limited access to affordable and nutritious food, particularly fresh fruits and vegetables. These areas are often defined by their lack of supermarkets or grocery stores within convenient traveling distance and typically overlap with low-income or marginalized populations. According to the USDA, food deserts are identified as census tracts meeting both low-income and low-access criteria.

Our project was motivated by the critical public health and social equity implications of food deserts. Millions of Americans—especially those living in rural regions, minority communities, or impoverished neighbourhoods—struggle to access essential nutrition, which may contribute to broader systemic issues including chronic health problems, economic stagnation, and inter-generational poverty. Given the increased awareness of health disparities during the COVID-19 pandemic, we saw an opportunity to use data science tools to provide insights that could support more equitable food systems across the U.S.

This project aimed to use data-driven methods to explore the geographic patterns, contributing factors, and potential policy interventions related to food deserts in the United States. Through exploratory data analysis, statistical modelling, and predictive analytics, our team sought to understand the root causes and potential solutions to this pressing issue.

## SMART questions

Food security is a fundamental determinant of health and well-being, and addressing food deserts requires detailed analysis. We were especially interested in exploring how machine learning models could not only reveal trends but also inform actionable policies. So, we set forth the following SMART questions:

1. Where are food deserts most geographically concentrated across the U.S., and how do these concentrations differ between urban and rural census tracts?
2. What demographics and socioeconomic factors are linked to food deserts?
3. How are food deserts related to health issues like obesity and diabetes?
4. Can we develop a predictive model that accurately identifies high-risk areas likely to be or become food deserts based on social and economic indicators?
5. What national policies could reduce food deserts based on our findings?

## Literature Review

Prior research has shown that food deserts are more prevalent in areas with high poverty, low vehicle ownership, and limited access to transportation. Studies have documented the link between food access and diet-related health outcomes, although the strength of this relationship varies by region and demographic group. Some researchers argue that physical proximity to food is only part of the problem—affordability, cultural preferences, and education also play key roles.

For instance, Ver Ploeg et al. (2009) in a USDA report highlighted that while low-income communities often lack large grocery stores, they may still have small corner stores that offer limited or unhealthy food options. Other studies, such as those by Walker et al. (2010), have shown that the presence of food deserts correlates with increased prevalence of obesity and Type 2 diabetes, but access alone does not guarantee improved health outcomes.

Our project builds on these findings by integrating datasets across health, economic, and demographic dimensions to assess both the presence and the effects of food deserts.

## Description of the Data

We utilized and downloaded 2 public datasets directly from USDA:

1. **Food Access Research Atlas**: Provides tract-level indicators of food access based on income and distance to supermarkets.
2. **Food Environment Atlas**: Provides supplementary food environment factors—such as store/restaurant proximity, food prices, food and nutrition assistance programs, and community characteristics—interact to influence food choices and diet quality.

We chose LILATracts\_1And10 as the proxy variable for Food Desert. This variable sufficiently captures the definition of a food desert as set forth by the USDA.

LILATracts\_1And10 is a **binary indicator variable** (1 = true, 0 = false) that identifies census tracts that meet both of the following criteria:

* **Low-Income Tract (LI):**
  + The tract's poverty rate is ≥ 20%,
  + *or* the tract's median family income is ≤ 80% of the state (or metropolitan area) median.
* **Low-Access Tract (LA):**
  + At least **500 people or 33%** of the population in the tract reside more than:
    - **1 mile (urban)** or
    - **10 miles (rural)**  
      from the nearest supermarket, supercentre, or large grocery store

The Food Access Atlas had 147 features in total including LILATracts\_1And10, and the Food Environment Atlas had a separate table for each cohort like stores, restaurants, access, etc. with a total of 281 features. All tables were joined using CountyFIPS codes to ensure geographic consistency and the combined dataset had 428 features in total. There was a total of around 74k observations in the aggregated data. Since the variable list was large, we needed to implement a robust feature selection process to filter the features for modelling.

## Data preprocessing and feature selection

The following steps were employed in sequence to clean the data:

1. **Dropping columns based on fill-rate:** Only variables with a fill rate of >=90% were retained.
2. **Data imputation:** Of the remaining columns, the missing values of numerical variables were imputed with the median of its respective column.
3. **Data type conversion:** The binary columns, including the target variable, were converted to factor type, which would be useful when used with the glm model.
4. **Data standardization:** The numeric columns were standardized using z-score transformation which is helpful for faster convergence during modelling.
5. **Random sampling:** Since there were 74k observations, we needed to subset the dataset to use only a 30% random sample (~22k rows) for model training purposes.

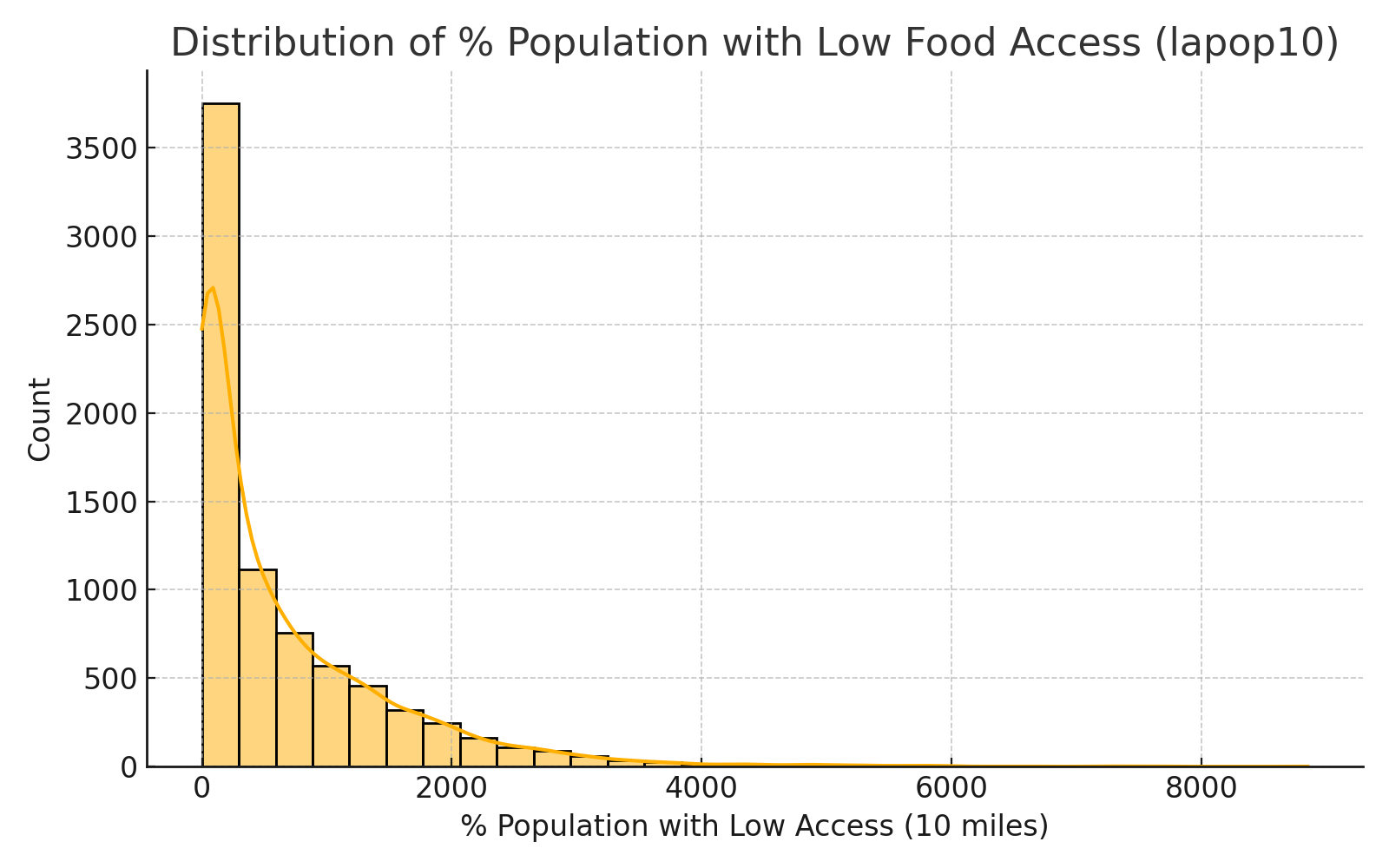
Once the data was adequately cleaned and sub-sampled, we performed the following steps to filter the features:

1. **Iterative Univariate feature selection:** We built a logistic regression model using one variable from the feature list as the independent (X) and FoodDesert as the dependent variable (y). Once all the models were built, only the features wherein the model p-value was statistically significant (<0.05) were selected.
2. **Penalized Lasso regression:** The second step was to perform a lasso penalization on the logit model using all the variables from the previous step. Lasso tends to make the coefficients of the weak variables to zero thereby eliminating them.
3. **Iterative feature elimination with vif:** The features output from the lasso model were then used to build logit model and the corresponding vif scores of the variables were calculated. The feature with the maximum vif score was removed and this modelling process was repeated until the max vif score was <5.
4. **Iterative feature elimination by p-value:** The final step involved filtering the features by the p-value of the coefficients. Since there were about 100 variables at this stage, we put a stringent threshold for the p-value (<0.01) to make sure only the strongest factors affecting FoodDesert were retained.

At the end of the feature selection process, there were a total of 27 variables which were strongly linked to FoodDesert.

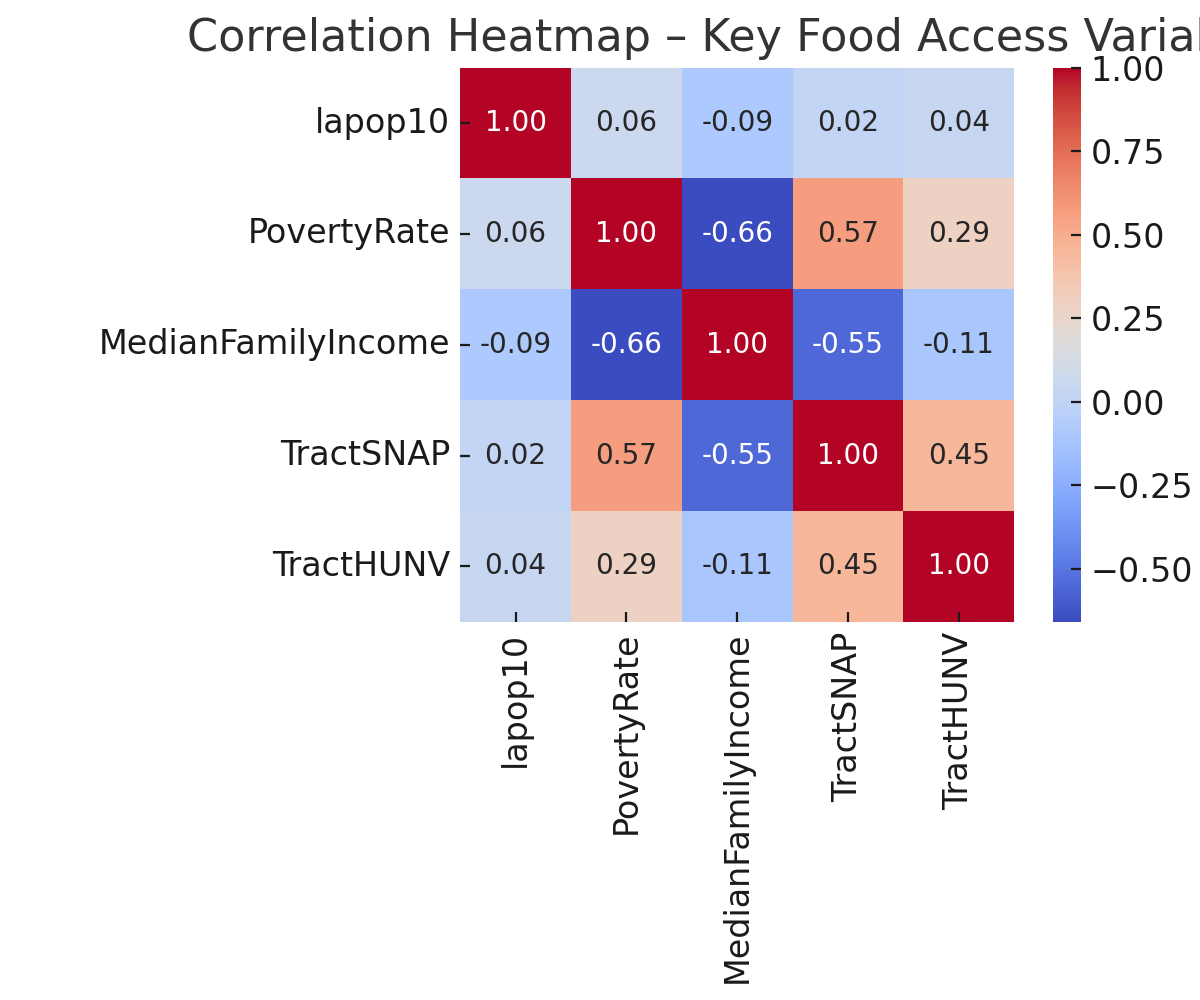
## Exploratory Data Analysis (EDA)

We began our EDA by focusing on the key variables associated with food access and socioeconomic hardship, rather than external health data or spatial mapping.

****

**Distribution Analysis:**

The primary variable of interest, lapop10 — the percentage of the population in a tract with low access to food within 10 miles — showed a positively skewed distribution. This indicates that while most tracts have relatively low percentages of food inaccessibility, a minority of tracts experience significantly higher rates, marking them as potential food deserts.

****

**Correlation Analysis:**

We explored the relationships between lapop10 and other core variables including:

* PovertyRate
* MedianFamilyIncome
* TractSNAP
* TractHUNV

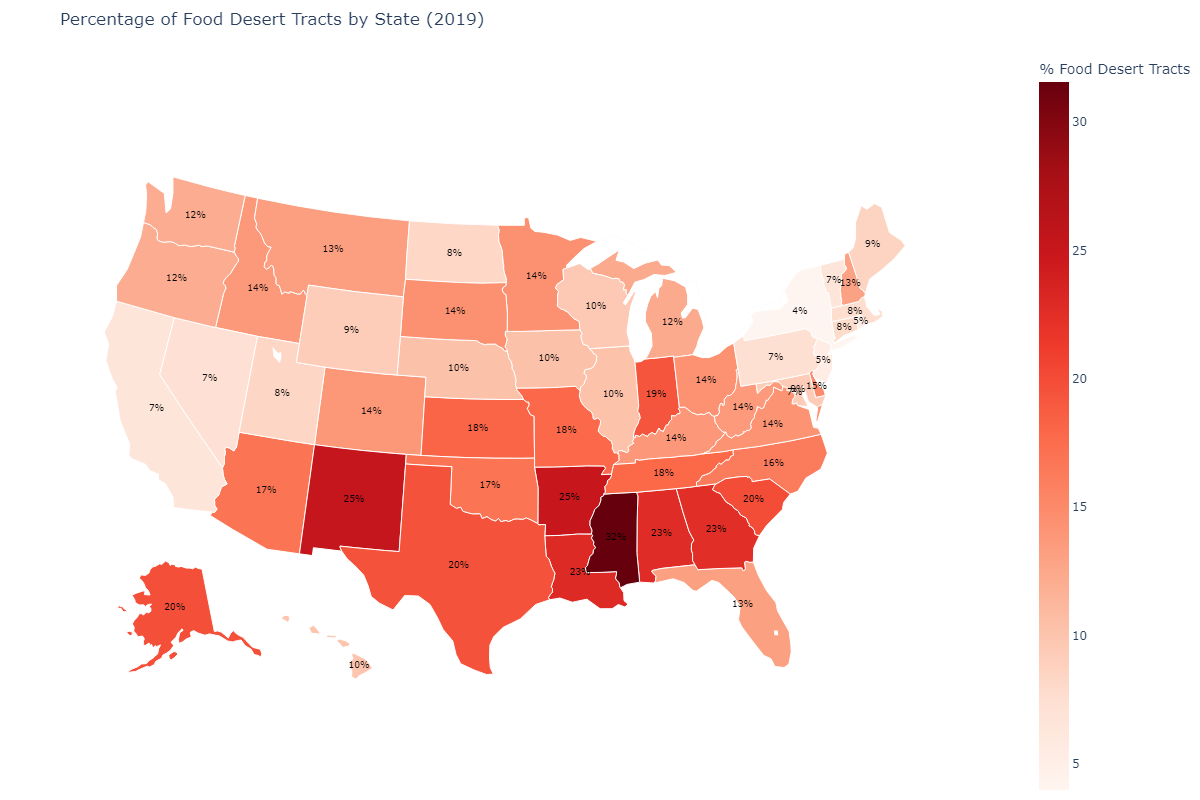
A correlation heatmap revealed strong positive associations between food inaccessibility and poverty and SNAP participation. Conversely, MedianFamilyIncome showed a negative correlation with lapop10, reinforcing the expected link between low income and poor food access. Lack of vehicle ownership (TractHUNV) was also positively correlated, highlighting transportation as a key access barrier.

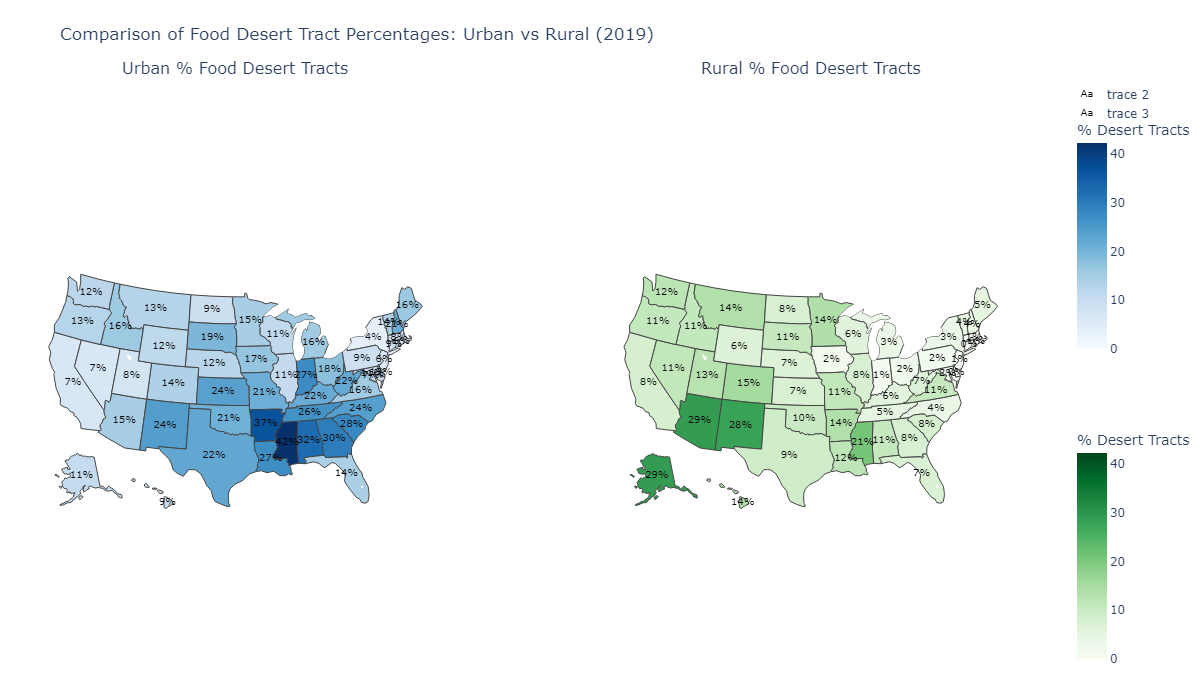
## SMART Questions & Insights

**Q1: Where are food deserts most geographically concentrated across the U.S.?**

**A graph of blue bars

AI-generated content may be incorrect.**





We used tract-level data to identify hotspots of food insecurity. We found the following observations:

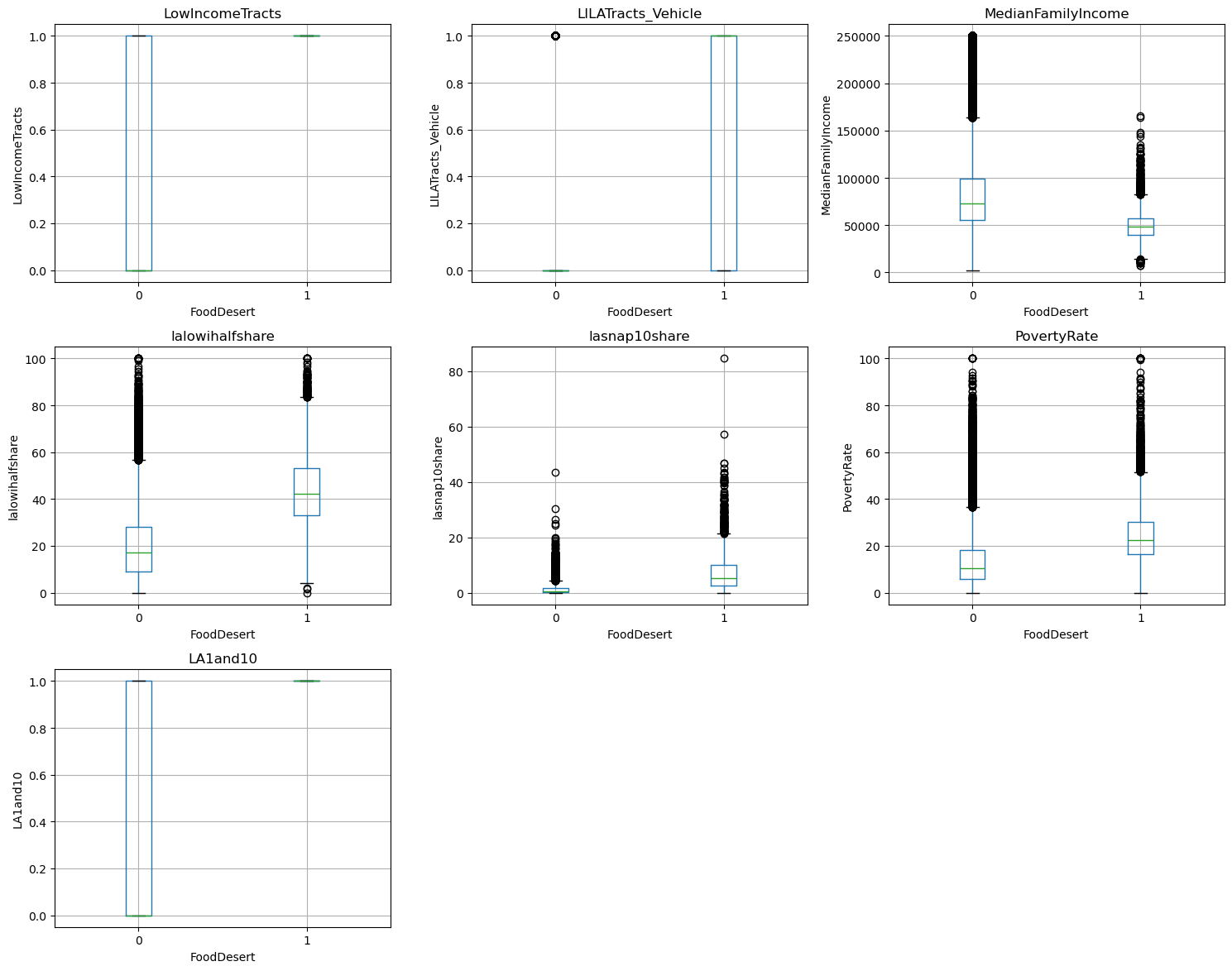
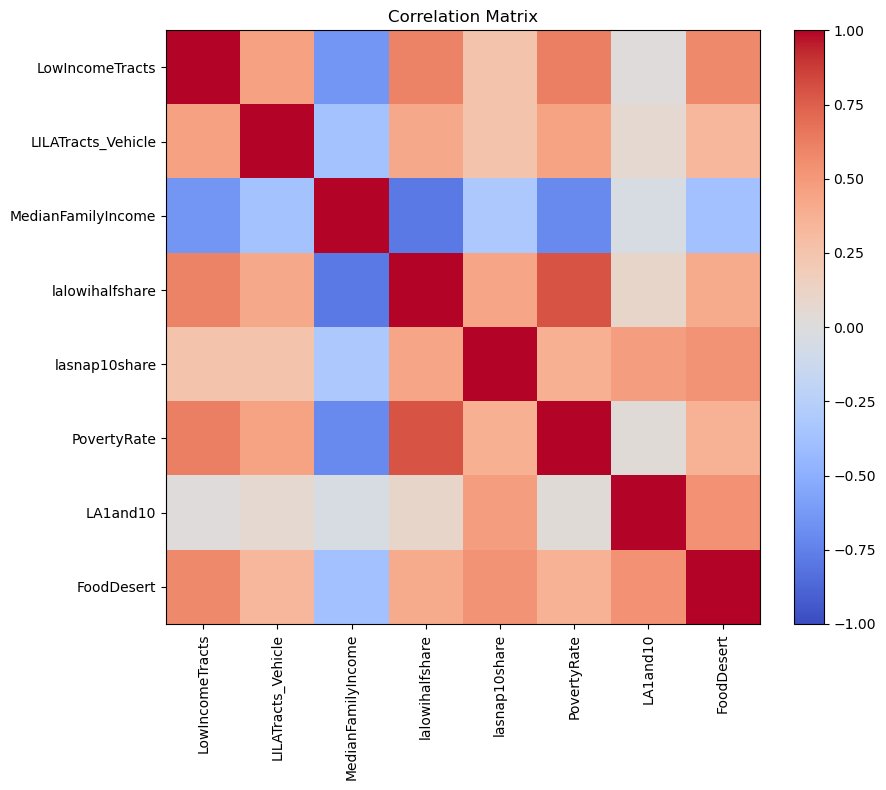
1. By far Texas has the most number of food deserts, but Mississippi has the largest percentage of food desert tracts.
2. States in the southeast have highest concentrations of urban food deserts (Mississippi, Arkansas, and Alabama).
3. States in the southwest region have the highest number of rural food deserts (Arizona and New Mexico). Alaska also has a high percentage of rural food deserts.

Overall, food deserts were concentrated mostly in the southern region of the US. This spatial distribution is likely influenced by factors such as:

* Urban sprawl and zoning in cities
* Lack of commercial incentives in rural markets
* Poor infrastructure and limited transportation

**Q2: What demographics and socioeconomic factors are linked to food deserts?**

Once the final list of features from the feature selection process were obtained, we calculated the spearman’s coefficient of correlation value for all the variables. Then the variables with >0.3 absolute value were visualized with a correlation heatmap as shown below along with the boxplots for each variable.

****

We found that the top key variables associated with food deserts, in terms of the correlation matrix, include the following components:

* **Low-income status**
* **Limited vehicle access**
* **High SNAP dependency**

These indicators suggest that food insecurity stems from structural inequality, not just poor planning. Households may technically live within a mile of a grocery store, but without a car or reliable public transport, access remains a barrier.

**Q3: How are food deserts related to health issues like obesity and diabetes?**

**A graph with red and black lines

AI-generated content may be incorrect.A graph of food desert

AI-generated content may be incorrect.**

**A blue and white squares with text

AI-generated content may be incorrect.**

To find out if health issues were linked to food deserts, we merged tract-level health variables, i.e. indicators for obesity and diabetes rates, with food desert indicators and calculated the correlation matrix along with the boxplots. Our analysis found:

* A weak correlation between food desert status and obesity/diabetes rates
* A strong correlation (r ≈ 0.5) between obesity and diabetes themselves

This suggests that food deserts may contribute to poor health, but they are not the sole driver. Other influences—such as healthcare access, exercise opportunities, diet patterns, and education—must also be considered.

**Q4: Can we develop a predictive model that identifies high-risk areas to be food deserts?**

From the final set of 27 features, we constructed two predictive models to determine the effects of each variable on food deserts:

1. **Logistic Regression** – Provided interpretable coefficients for each variable.
2. **Random Forest** – Offered higher performance since non-linearity relationships were modelled.

**Logistic Regression summary:**

A screenshot of a computer program

AI-generated content may be incorrect.

The final model was constructed using the same set of features from the feature selection process. It is interesting to note that along with the usual key factors like poverty rate, vehicle access and count of SNAP users, supplemental variables like age and race groups, restaurant and farmer’s market access, and food insecurity indicators also came up as strong factors linked with food deserts.

**A graph of a curve

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.**

The model performance was adequate in terms of measuring the effect of variables, but not accurate enough to use for predicting tracts that would likely become food deserts. This is mainly due to the imbalanced training set, wherein most of the observations were 0 (Not Food Desert). This impacted on the F1 score which was a low 24%. It is suggested that further modelling exercises employ class balancing methods along with hyperparameter tuning to improve performance.

**A graph of a forest plot

AI-generated content may be incorrect.Random Forest summary:**

**A screenshot of a computer

AI-generated content may be incorrect.**

The same set of features used for the logit model were used to train the RF model. This model performed slightly better since it was able to capture the non-linear relationships between the variables more accurately. But it was not feasible to be used for any predictions since the F1 score was around 26%. In the future, it is recommended to filter the feature list using RF since this retains the non-linear relationships instead of using the feature list output from logit model. Additionally, implementing class balancing techniques and switching to a more powerful tree-based model like XGB with hyper-parameter tuning will increase the predictiveness of the model.

A graph with a bar graph

AI-generated content may be incorrect.

The feature importance chart from the RF model is consistent with that of the coefficient values from the logit model. Poverty rate shows up as the top variable, followed by the different tract level variables defining SNAP count users, race and age groups. The variables that stand out are PCTGQTRS and NUMGQTRS which indicate the tract population that reside in institutional or communal living arrangements.

**Q5: What national policies could reduce food deserts?**

The main drivers of food deserts were found to be low-income, limited vehicle availability and SNAP user share. Our analysis points to multiple intervention policies aimed at each of these drivers that can be introduced at a federal level as well as at the state level to reduce food deserts:

* **Low-income Tracts:** 
  + Expand the Earned Income Tax Credit: Increase the credit for very low-income households living in federally designated food-desert tracts.
  + Targeted Workforce Development Grants: Embed food-desert zone bonuses into employee training programs, so employers in these areas receive wage subsidies when they hire and upskill local residents.
* **Limited vehicle availability:**
  + Incentivize Mobile Market Operators: Offer incentives to vendors to serve limited vehicle neighbourhoods.
  + Zero-Fare Public transport routes: Partner with state transportation services to waive off fares on designated public-transit routes connecting grocery retailers with food deserts.
* **SNAP enhancements:**
  + Nationwide SNAP Online Ordering & Delivery: Mandate all major grocery chains to accept online SNAP purchases and subsidize delivery fees for EBT users in food desert counties.
  + Mobile EBT Kiosk Grants: Deploy portable EBT terminals throughout rural food deserts to broaden point-of-sale options.

These policies should be multi-tiered, combining federal, state, and local efforts and tailored to regional needs.

## Conclusion

Our analysis confirms that food deserts are complex, multifaceted problems rooted in geography, infrastructure, and inequality. While physical distance to food sources matters, our findings emphasize that income, transportation, and economic vulnerability are stronger predictors of food access.

We found that food deserts are primarily concentrated in the southern region of the United States. Health impacts are not directly caused by food deserts alone, but by other supplementary factors intersecting socio-economic conditions. Thus, any solution must address the full ecosystem—access, affordability, education, and opportunity.

Our predictive models offer a foundation for targeted policy responses and future research. By identifying high-risk areas, we can improve resource allocation and advance equitable access to nutrition for all Americans.

## Limitations and Future Scope

Some limitations of our study include:

* Health-impact analysis was narrow since the study was restricted to only 2 variables – diabetes and obesity. However, there are many more health indicators that can be used along with supplementary environmental factors. All these indicators can be utilized to holistically provide a clear view on the relationship between health and food access.
* Not all relevant variables were sourced—future research should integrate factors like crime rates, school quality, and walkability from other government annotated datasets.
* Our models assume static conditions; longitudinal or time-series data could better capture the dynamics of changing neighbourhoods.

Future directions include:

* Incorporating clustering analysis to identify segments more prone to food deserts and time-series analysis to historically map the point of conversion from a non-food desert tract to a food desert tract.
* Implementing more robust feature-engineering and modelling processes. This includes creating complex variables as well as using more sophisticated models with hyper-parameter tuning.
* Collaborating with urban planners, public health departments, and non-profits for insights and real-world applications.

## References

**1. Ver Ploeg et al. (2009).**  
*Access to affordable and nutritious food: Measuring and understanding food deserts and their consequences: Report to Congress.*  
U.S. Department of Agriculture, Economic Research Service.  
[*https://www.ers.usda.gov/publications/pub-details/?pubid=42729*](https://www.ers.usda.gov/publications/pub-details/?pubid=42729)

**2. Walker, R. E., Keane, C. R., & Burke, J. G. (2010).**  
Disparities and access to healthy food in the United States: A review of food deserts literature. *Health & Place, 16*(5), 876–884.  
[*https://doi.org/10.1016/j.healthplace.2010.04.013*](https://doi.org/10.1016/j.healthplace.2010.04.013)

**3. Food Access Research Atlas (Food Deserts data)**

U.S. Department of Agriculture, Economic Research Service. (2020). *Food Access Research Atlas data (2019)* [Data set].  
[*https://www.ers.usda.gov/data-products/food-access-research-atlas/*](https://www.ers.usda.gov/data-products/food-access-research-atlas/)

**4. Food Environment Atlas (Socioeconomic, Health, Access, Store, Restaurant, Tax, and Local data)**

U.S. Department of Agriculture, Economic Research Service. (2020). *Food Environment Atlas* [Data set].  
[*https://www.ers.usda.gov/data-products/food-environment-atlas/*](https://www.ers.usda.gov/data-products/food-environment-atlas/)