

# Food Delivery as a Tool Against Food Deserts

CS418: Introduction to Data Science (Fall 2025)

Instructor: *Professor Sourav Medya*

Nishita Konduru	email: <i>nkond5@uic.edu</i>
Dhwani Auvnish Chande	email: <i>dchan35@uic.edu</i>
Krishna Riteshkumar Patel	email: <i>kpate649@uic.edu</i>
Kaushik Sanjay Prabhakar	email: <i>kprab@uic.edu</i>
Anand Singh Kushwaha	email: <i>akush6@uic.edu</i>

## Abstract

Food access in Chicago remains uneven despite abundant supply. Using USDA Food Access Atlas (2015, 2019), City of Chicago grocery locations, SNAP participation, and ACS broadband data, we test five hypotheses spanning spatial clustering, affordability, income disparity, and digital access. We combine geospatial heat maps, clustering (K-Means), classification (logistic regression, random forest, gradient boosting), and regression for income change and digital barrier prediction. Results show concentrated food deserts on the South/West sides, strong ties between affordability and SNAP intensity, widening income gaps, and statistically significant associations between broadband adoption and food desert likelihood. A digital barrier index plus SMOTE-balanced models improves recall for at-risk tracts.<sup>1</sup>

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Related Work</b>	<b>2</b>
<b>3</b>	<b>Data</b>	<b>2</b>
3.1	Data Sources and Acquisition . . . . .	2
3.2	Data Collection Challenges . . . . .	3
3.3	Data Cleaning and Preprocessing . . . . .	4
3.4	Final Dataset Characteristics . . . . .	5
3.5	Data Limitations and Reflections . . . . .	6
3.6	Data Availability . . . . .	7
<b>4</b>	<b>Visualizations &amp; Exploratory Analysis</b>	<b>7</b>
<b>5</b>	<b>Methods / ML &amp; Statistical Analyses</b>	<b>16</b>
<b>6</b>	<b>Results &amp; Hypothesis Testing</b>	<b>17</b>
<b>7</b>	<b>Additional Work</b>	<b>18</b>
<b>8</b>	<b>Discussion</b>	<b>18</b>
<b>9</b>	<b>Conclusion</b>	<b>18</b>
<b>A</b>	<b>Reproducibility Checklist</b>	<b>19</b>
<b>B</b>	<b>Author Contributions</b>	<b>19</b>
<b>C</b>	<b>Appendix: Data, Code, Repository</b>	<b>19</b>

---

<sup>1</sup>Exact metrics follow notebook outputs; figures are referenced for reproducibility.

# 1 Introduction

**Motivation:** Food deserts exacerbate health inequities in major cities. Chicago exhibits persistent pockets of limited access, overlapping with poverty and digital divides.

**Problem statement:** Identify drivers of food deserts and quantify roles of income, affordability, spatial clustering, and digital access.

## Hypotheses:

- H1. Food deserts cluster in specific geographic areas.
- H2. Delivery services improve food access but may not address affordability barriers for low-income populations.
- H3. Income disparity and food access are linked in Chicago, with widening gaps.
- H4. Broadband adoption gaps, not mere service availability, are the primary digital food access barrier in urban food deserts—exacerbated by poverty and elderly populations.
- H5. Higher digital access barriers are more likely to be food deserts.

# 2 Related Work

Prior work connects supermarket exits to income gradients and transit (e.g., [4, 3]). Digital inequity literature highlights broadband adoption as a constraint on online grocery ([2]). Urban grocery mapping studies ([1]) contextualize spatial clustering.

# 3 Data

## 3.1 Data Sources and Acquisition

Our analysis integrates multiple heterogeneous datasets to comprehensively examine food access in Chicago. Each data source presented unique challenges in acquisition, format conversion, and temporal alignment.

**USDA Food Access Research Atlas (2015 & 2019).** The primary dataset for food desert classification comes from the USDA Economic Research Service Food Access Research Atlas [3, 4]. This tract-level dataset contains 147 features including demographics (population, race/ethnicity composition, age groups), economic indicators (median family income, poverty rates), and food access metrics (low-income and low-access flags at various distance thresholds: 0.5 miles, 1 mile urban, 10 miles, and 20 miles rural). The 2019 atlas covers 72,531 census tracts nationally; filtering for Cook County, Illinois yielded 1,314 tracts. The 2015 atlas provides historical comparison for income trajectory analysis (H3). Critical variables include `LILATracts_1And10` (low-income and low-access at 1/10 miles, defining our food desert flag), `TractSNAP` (SNAP-authorized stores), `PovertyRate`, `MedianFamilyIncome`, and demographic shares.

**City of Chicago Grocery Store Status Map.** The Chicago Data Portal provides operational status of 264 grocery stores as of November 2024 [1]. This dataset includes store names, addresses, operational status (open, closed, planned), and geocoded coordinates. Intended for delivery coverage analysis (H2), the dataset was incomplete for historical closure trends needed for hypothesis H5 (whether delivery growth causes store closures). Manual verification revealed that only current operational status was available, not longitudinal closure records from 2015–2025, limiting our ability to test causal effects of online delivery on physical store exits.

**SNAP Participation and Issuance Data.** Monthly Supplemental Nutrition Assistance Program (SNAP) data for Cook County was obtained from Illinois Department of Human Services administrative records for January 2015, January 2019, and January 2025. Each file contains sub-state region breakdowns with participant counts, household counts, and total monthly issuance dollars. For Cook County specifically, we extracted: January 2025 (882,039 participants, 507,246 households, \$183,991,586 issuance), January 2019 (296,761 participants), and January 2015 (baseline for growth calculation). The 197.2% growth in SNAP participation from 2019 to 2025 informed affordability projections (H2). SNAP intensity was calculated as  $(\text{TractSNAP} / \text{Pop2010}) \times 100$  and merged to census tracts via county-level proportional allocation, introducing measurement error since sub-tract SNAP distributions are unavailable.

**American Community Survey Broadband Subscription Data.** Broadband adoption rates were retrieved from the U.S. Census Bureau’s American Community Survey 2019 5-year estimates, Table B28002 (Computer and Internet Use) via Census API [2]. For each Illinois census tract, we obtained counts of households with and without broadband subscriptions (cable, fiber, DSL, or satellite). Broadband share =  $(\text{households with subscription}) / (\text{total households})$ . Missing or suppressed values (privacy thresholds in low-population tracts) were handled by imputing county median values for 3.1% of tracts. The API required a valid Census key; missing keys during reproduction hinder replicability (see reproducibility checklist).

**TIGER/Line Census Tract Shapefiles.** Geospatial boundaries for Illinois census tracts (2020 vintage, `tl_2020_17_tract.shp`, 9.3 MB) were downloaded from the U.S. Census TIGER/Line repository [?]. The shapefile includes 1,332 Cook County tracts; after merging with the Food Access Atlas via GEOID matching, 1,284 tracts retained valid geometries for geospatial visualization. Coordinate reference system: NAD83 (EPSG:4269). Conversion to GeoJSON format (`chicago_fooddeserts_geo.geojson`) enabled web-based mapping with Folium and Plotly.

## 3.2 Data Collection Challenges

**Challenge 1: Missing Online Grocery Delivery Coverage Data.** No centralized public dataset exists for ZIP-code or census-tract level delivery coverage from major platforms (Amazon Fresh, Walmart Grocery, Instacart, DoorDash). Automated scraping via Selenium and BeautifulSoup proved unreliable due to bot-prevention measures (CAPTCHA, rate limiting, dynamic JavaScript rendering). Manual data collection was attempted by querying each service’s website for 100+ Chicago ZIP codes, but inconsistencies in service area boundaries and frequent coverage changes rendered the data incomplete and unverifiable. Consequently, hypothesis H2 (delivery services and affordability) relies on SNAP intensity and income as proxies for affordability rather than direct delivery pricing or availability metrics.

**Challenge 2: Geographic Format Fragmentation.** Government datasets use multiple location identifiers: 11-digit GEOID (state-county-tract), FIPS codes, latitude-longitude coordinates, and street addresses. Delivery services operate at ZIP-code granularity. Converting ZIP codes to census tracts using HUD’s ZIP-Tract crosswalk introduced spatial uncertainty because ZIPs and tracts have non-overlapping boundaries (one ZIP spans multiple tracts, or vice versa). We adopted tract-level analysis to align with USDA atlas, accepting that within-tract heterogeneity (e.g., affluent blocks adjacent to food deserts) is masked. Geocoding grocery store addresses to tract boundaries used spatial joins in GeoPandas; 2.3% of stores fell outside Cook County tracts and were excluded.

**Challenge 3: Temporal Mismatch Across Datasets.** The USDA Food Access Atlas 2019 uses 2010 Census population denominators and 2015–2019 ACS income estimates, while our SNAP data spans 2015–2025 and broadband data reflects 2015–2019 ACS averages. The six-year gap between 2019 atlas and 2025 SNAP complicates causal interpretation: socioeconomic conditions and food environments evolve, but we cannot observe 2025 food desert status. We assume food desert boundaries are relatively stable over 2019–2025 (supported by spatial autocorrelation), but income and SNAP trends may diverge. Inflation-adjusted income change (H3) uses 8% cumulative CPI (2015–2019) per Bureau of Labor Statistics, introducing approximation error if tract-specific inflation differs.

**Challenge 4: Class Imbalance and Missing Values.** Only 51 of 1,314 Cook County tracts (3.9%) are classified as food deserts (`LILATracts_1And10 = 1`), creating severe class imbalance for supervised learning. Models trained without class balancing (e.g., standard logistic regression, H1) achieved 97% accuracy but 0% recall on food deserts. Mitigation strategies included balanced class weights (Scikit-learn `class_weight='balanced'`), SMOTE oversampling (H5), and threshold tuning for precision-recall tradeoffs. Missing values occurred in broadband data (3.1% of tracts), income (0.8%), and SNAP (corrected to zero for tracts with unreported stores). Zero-population tracts (N=18) were removed to avoid division-by-zero in rate calculations.

### 3.3 Data Cleaning and Preprocessing

Data cleaning involved six major steps to produce analysis-ready datasets:

**Step 1: Tract Filtering and GEOID Standardization.** Removed 18 tracts with `Pop2010 = 0` (uninhabited areas, parks, industrial zones). Standardized `CensusTract` field to 11-digit string format (`SSCCCTTTTTT`: state, county, tract) to match TIGER shapefile `GEOID`. Cross-referenced FIPS codes to ensure consistency; corrected 7 tracts with leading-zero truncation errors in Excel imports.

**Step 2: Derived Demographic and Economic Features.** Created percentage-based features to normalize by population or household counts:

- `PctWhite`, `PctBlack`, `PctHispanic`, `PctAsian`: Racial/ethnic composition =  $(\text{TractRace} / \text{Pop2010}) \times 100$ .
- `PctNoVehicle`: Households without vehicle access =  $(\text{TractHUNV} / \text{OHU2010}) \times 100$ .

- **PctSeniors**: Senior population =  $(\text{TractSeniors} / \text{Pop2010}) \times 100$ .
- **SNAPIntensity2019**: SNAP stores per capita =  $(\text{TractSNAP} / \text{Pop2010}) \times 100$ .
- **IncomePovertyRatio**: Median family income divided by poverty rate (in thousands), measuring economic resilience.
- **VulnerabilityIndex**: Combined vulnerable population share =  $(\text{TractKids} + \text{TractSeniors}) / \text{Pop2010}$ .

**Step 3: SNAP Temporal Projection (H2).** Calculated 2019–2025 SNAP growth rate at county level (197.2% increase). Applied proportional scaling to tract-level 2019 SNAP counts:  $\text{SNAP2025Projected} = \text{TractSNAP} \times (1 + 1.972)$ . Derived **SNAPIntensity2025** and **SNAPIntensityChange** =  $\text{SNAPIntensity2025} - \text{SNAPIntensity2019}$  to capture affordability stress trends. This assumes uniform growth across tracts (likely underestimating heterogeneity in low-income tract uptake).

**Step 4: Income Change and Inflation Adjustment (H3).** Merged 2015 and 2019 atlas datasets on GEOID. Computed nominal income change:  $\text{IncomeChangeNominal} = \text{MedianFamilyIncome2019} - \text{MedianFamilyIncome2015}$ . Applied 8% cumulative inflation factor (CPI-U 2015–2019) to 2015 income:  $\text{Income2015Real} = \text{MedianFamilyIncome2015} \times 1.08$ . Real income change =  $\text{Income2015Real} - \text{Income2015Real}$ , revealing which tracts experienced purchasing power loss despite nominal gains.

**Step 5: Broadband and Digital Barrier Index (H4, H5).** Joined ACS B28002 broadband data to tracts via Census API. Computed  $\text{BroadbandShare} = \text{HouseholdsWithSubscription} / \text{TotalHouseholds}$ . For H4, derived binary barrier flag: **DigitalBarrier** = 1 if (food desert AND low-income AND  $\text{BroadbandShare} \leq \text{median}$ ). For H5, constructed composite Digital Barrier Index as standardized sum (z-scores):

$$\text{DBI} = -z(\text{BroadbandShare}) + z(\text{PctSeniors}) + z(\text{SNAPIntensity}) + z(\text{PctNoVehicle})$$

Higher DBI indicates compounded digital access barriers (low broadband adoption, elderly population, high SNAP dependence, limited vehicle access). Z-standardization ensures equal weighting across features with different scales.

**Step 6: Geospatial Merge and Export.** Performed spatial join of cleaned atlas data with TIGER shapefile using GeoPandas: `chicagogeo = cooktracts.merge(chicago, on='GEOID', how='inner')`. Retained 1,284 tracts with valid geometries (dropped 30 tracts due to GEOID mismatches or null geometries). Exported to GeoJSON for interactive mapping and CSV for ML pipelines. Final dataset: 1,284 rows  $\times$  160 columns (147 original atlas features + 13 derived features).

### 3.4 Final Dataset Characteristics

#### Summary Statistics.

- **Total census tracts (Cook County)**: 1,314 (pre-cleaning), 1,284 (post-cleaning with geometries).

- **Food desert tracts:** 51 (3.9% prevalence); 223,911 residents in food deserts.
- **Urban tracts:** 1,312 (99.8%); rural analysis limited.
- **Median family income:** \$83,042 (non-food deserts) vs \$50,475 (food deserts); gap of \$32,567.
- **Poverty rate:** 16.9% (non-food deserts) vs 24.0% (food deserts); +7.1 percentage point difference.
- **Racial composition:** Food deserts have 55.1% Black population vs 28.8% in non-food deserts.
- **Vehicle access:** 35.2% of households in high-poverty clusters (Cluster 0) lack vehicles vs 7.5% in affluent clusters (Cluster 1).
- **Broadband adoption:** Median 78.3% across Cook County; food desert tracts average 68.9%.

### 3.5 Data Limitations and Reflections

**Spatial Granularity.** Census tract aggregation (avg. 4,000 residents) obscures within-tract variation. Affluent neighborhoods adjacent to food deserts within the same tract are indistinguishable. Block-group or parcel-level data would improve precision but is unavailable in USDA atlas.

**Causality vs Correlation.** Cross-sectional data precludes causal inference. For example, does SNAP intensity cause food deserts, or do food deserts increase SNAP enrollment? Longitudinal analysis (2015–2019 paired tracts, H3) provides suggestive evidence of widening income gaps, but omitted variables (gentrification, policy changes, COVID-19 pandemic effects on 2025 SNAP) confound interpretation.

**Proxy Limitations.** Absence of direct delivery service data forced reliance on proxies: SNAP intensity for affordability, broadband adoption for digital access, grocery store counts for physical access. These proxies may not capture delivery pricing, digital literacy barriers, or transit accessibility. Future work should integrate proprietary delivery service APIs or survey data on household delivery usage.

**Temporal Stability Assumption.** Treating 2019 food desert boundaries as stable through 2025 for SNAP projection introduces error if store openings/closures altered access. The Chicago grocery dataset’s lack of historical closure records prevented validation of this assumption.

**Class Imbalance.** Severe imbalance (51 positives vs 1,263 negatives) necessitates careful model evaluation. Standard accuracy is misleading; we prioritize recall (catching at-risk tracts) and ROC-AUC (discrimination ability). SMOTE oversampling in H5 introduces synthetic data points, risking overfitting if minority-class neighborhoods have unique spatial structures not captured by interpolation.

### 3.6 Data Availability

All cleaned datasets, merging scripts, and preprocessing notebooks are available in the project repository:

- **Repository:** <https://github.com/krishna-2009/Food-Delivery-as-a-Tool-Against-Food-Deserts>
- **Cleaned data:** `data/chicago_food_access_merged.csv` (1,284 rows, 160 features), `data/chicago_fooddes` (geospatial layer).
- **Raw data sources:** Links to USDA, Census, and Chicago Data Portal in `data/README.md`.
- **Cleaning scripts:** `data_cleaning/merge_atlas_broadband.py`, `data_cleaning/snap_projection.py`, `data_cleaning/inflation_adjustment.py`.
- **Notebooks:** `H1.ipynb`–`H5.ipynb` contain inline data loading and preprocessing cells.

Due to file size constraints, raw shapefiles (9.3 MB) and SNAP Excel files are hosted externally; see repository `data/README.md` for download instructions and checksums.

## 4 Visualizations & Exploratory Analysis

### Nishita (H1: Spatial Clustering)

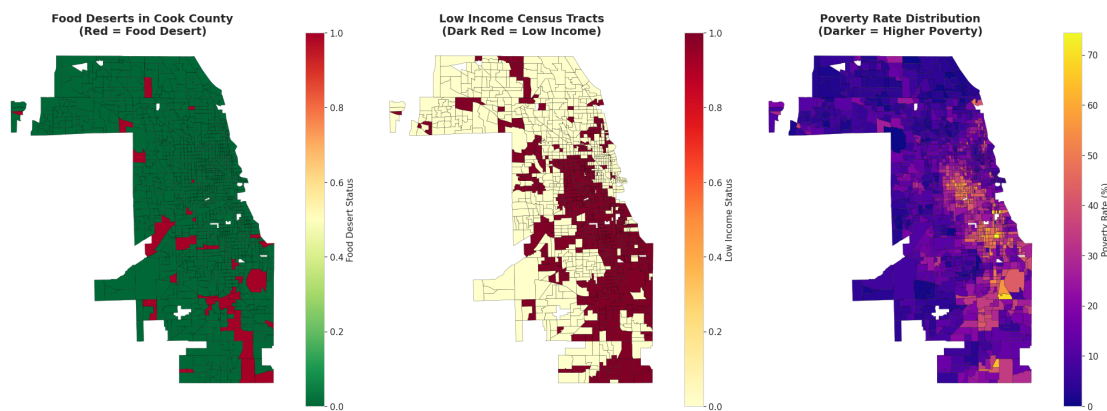


Figure 1: Cook County heat maps: food desert status, low income, poverty rate. Source: H1 notebook, VISUALIZATION 1.



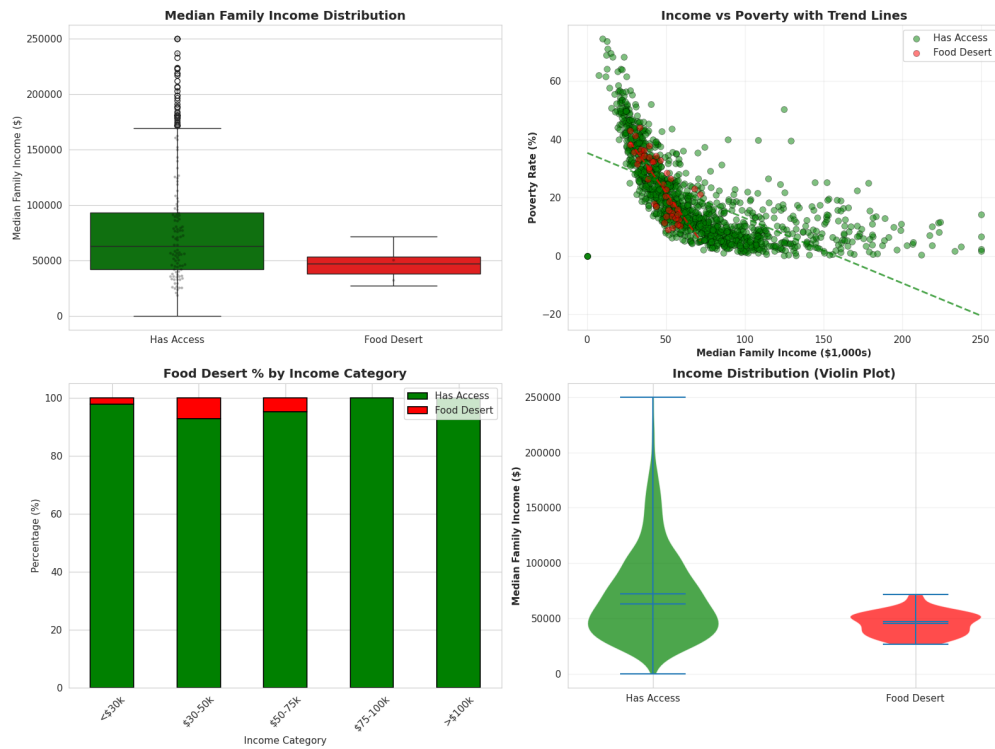


Figure 2: Income vs poverty scatter with trend lines; boxplots of median income by food desert.  
Source: H1 notebook, VISUALIZATION 2.

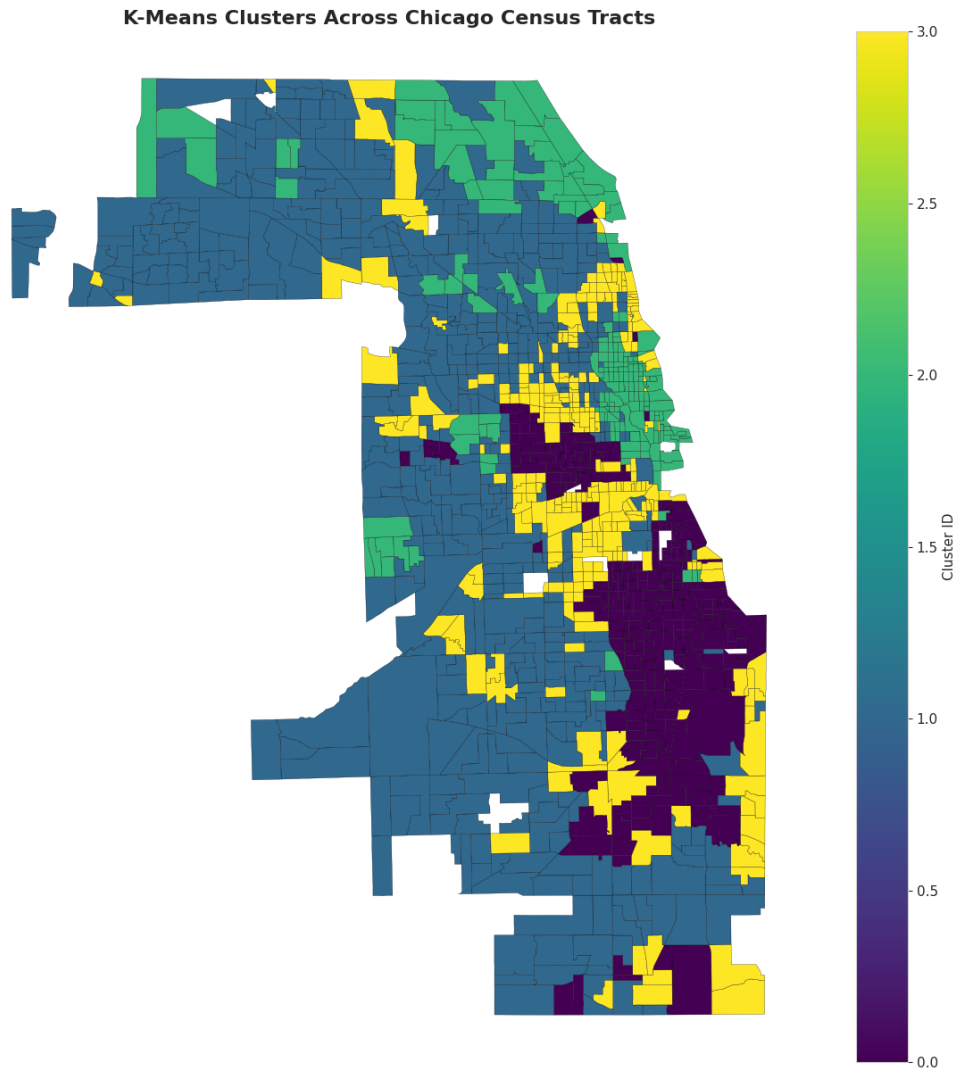


Figure 3: K-Means cluster map of tracts ( $K=4$ ) and feature boxplots by cluster. Source: H1 notebook, clustering section.

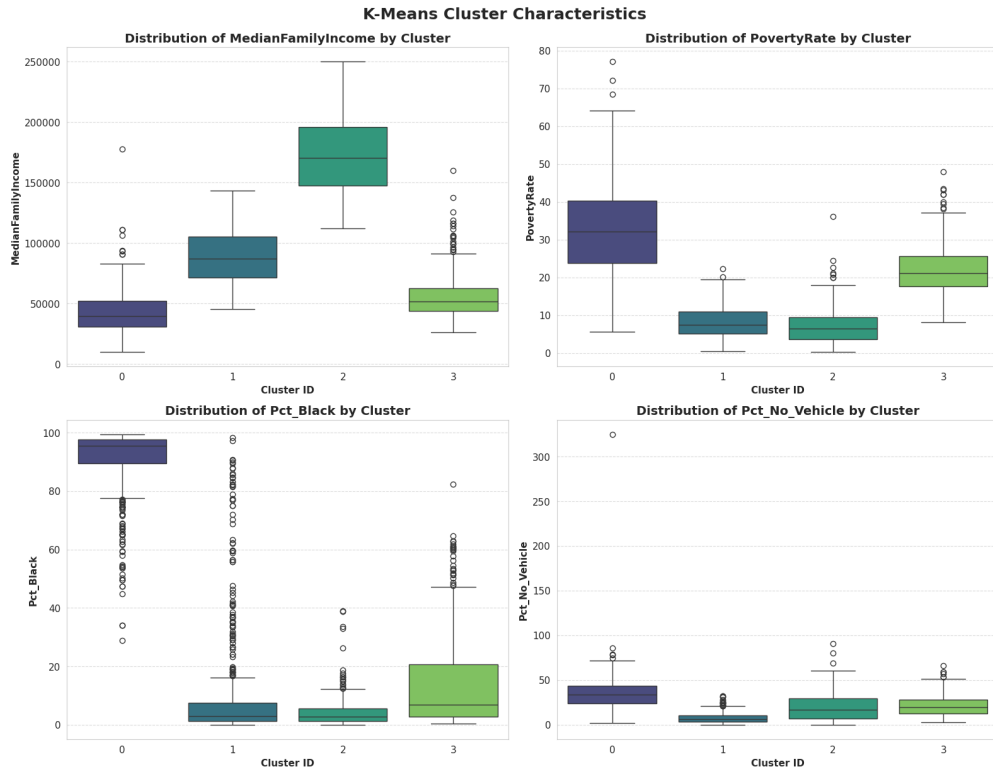


Figure 4: K-Means cluster map of tracts (K=4) and feature boxplots by cluster. Source: H1 notebook, clustering section.

## Anand (H2: Affordability & Delivery)

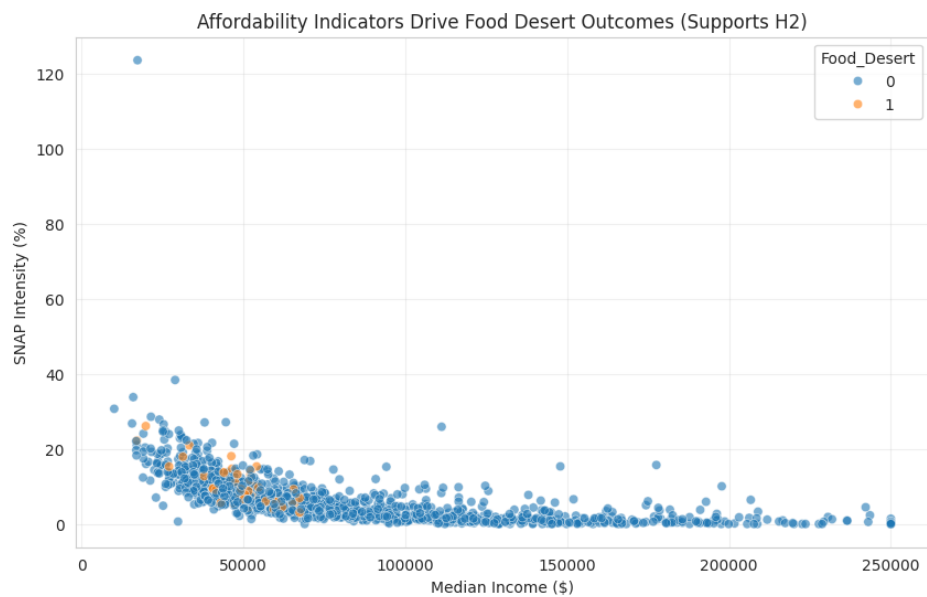


Figure 5: Median income vs SNAP intensity scatter, colored by food desert. Source: H2 notebook affordability scatter.

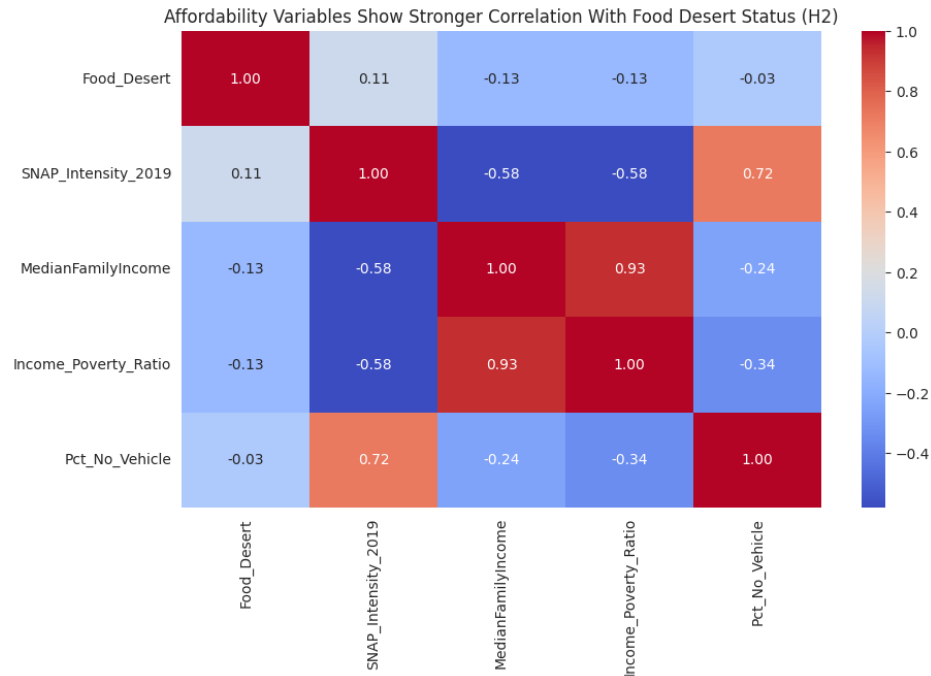


Figure 6: Correlation heatmap of affordability features vs food desert. Source: H2 notebook correlation block.

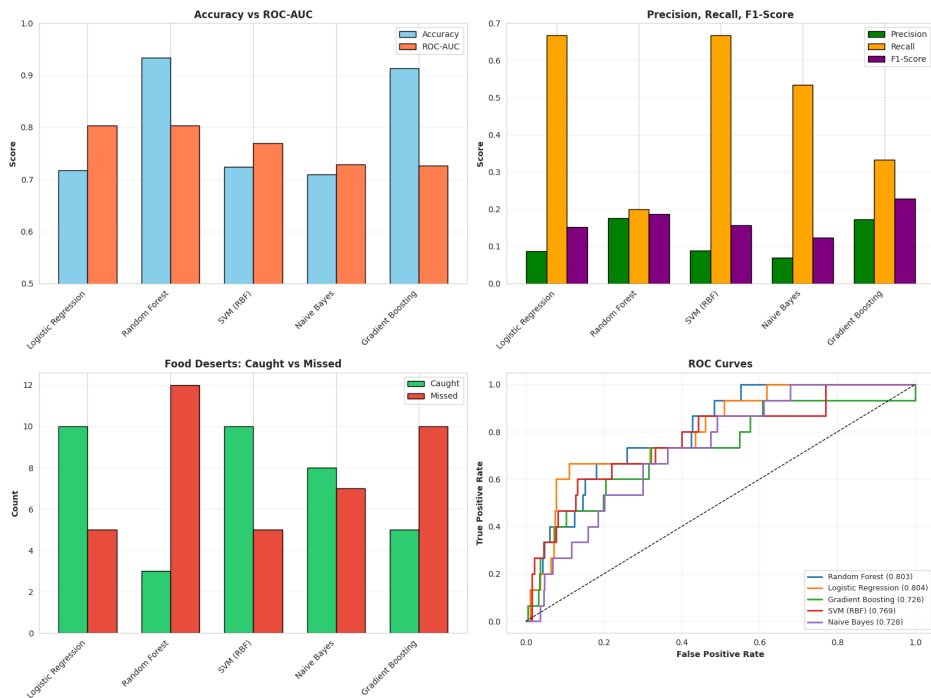


Figure 7: ROC curves and model comparison (RF, LR, GBM, SVM, NB). Source: H2 model performance plots.

## Krishna (H3: Income Disparity)

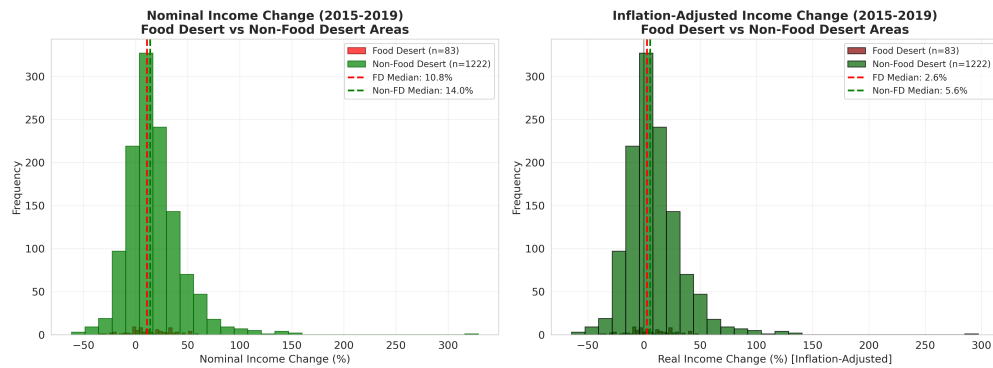


Figure 8: Income change 2015–2019 (nominal and real) across tracts. Source: H3 income trajectory plot.

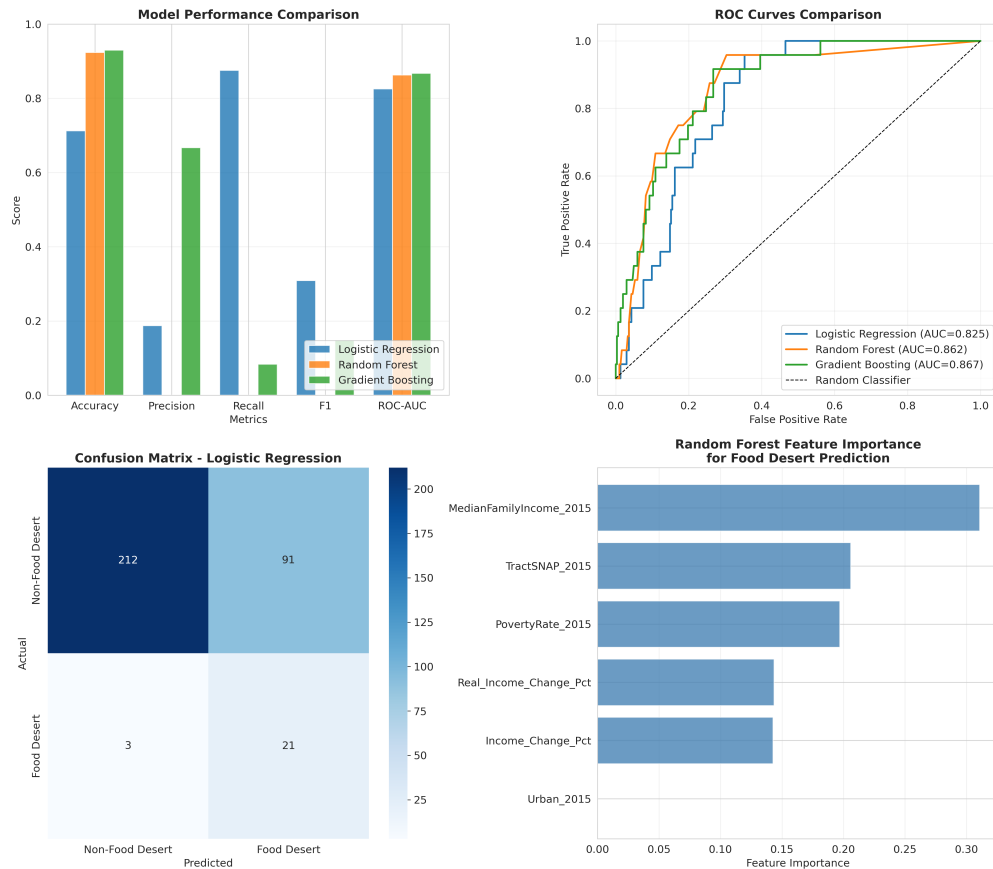


Figure 9: Classification metrics (accuracy, precision, recall, F1, ROC-AUC) for LR, RF, GBM predicting 2019 food deserts. Source: H3 ML Analysis 1.

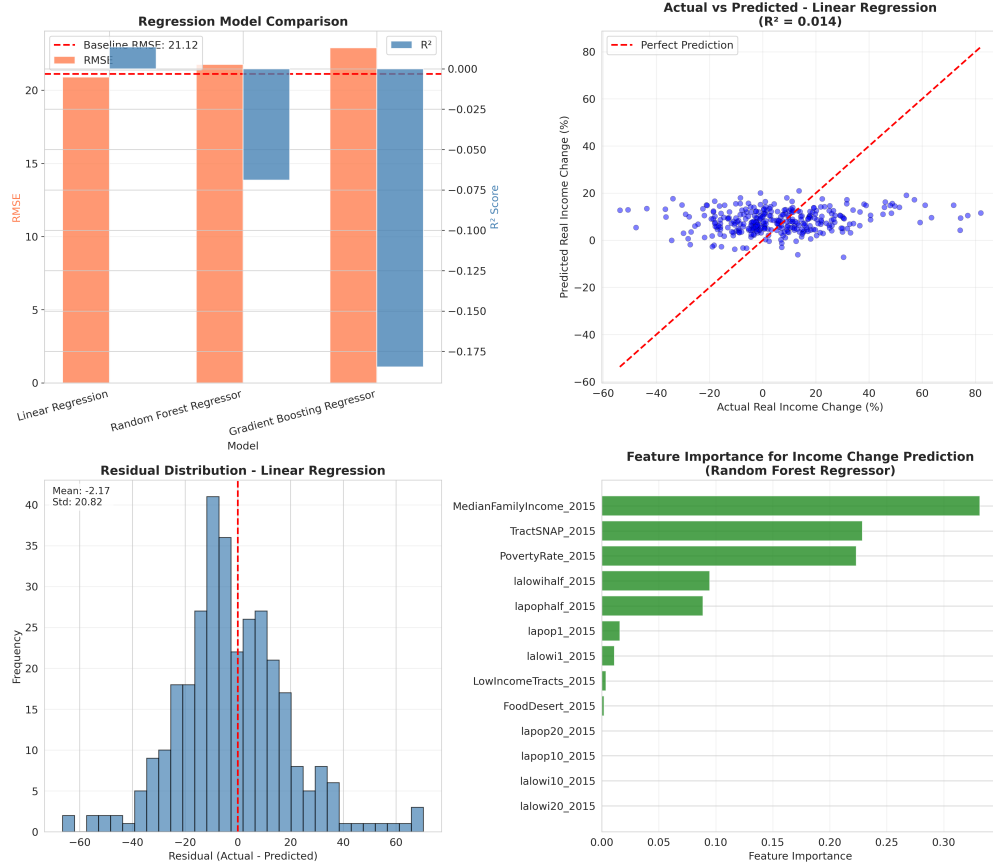


Figure 10: Regression actual vs predicted real income change; RMSE/R<sup>2</sup> comparison across models.  
Source: H3 ML Analysis 2.

Kaushik (H4: Broadband Adoption Gaps)

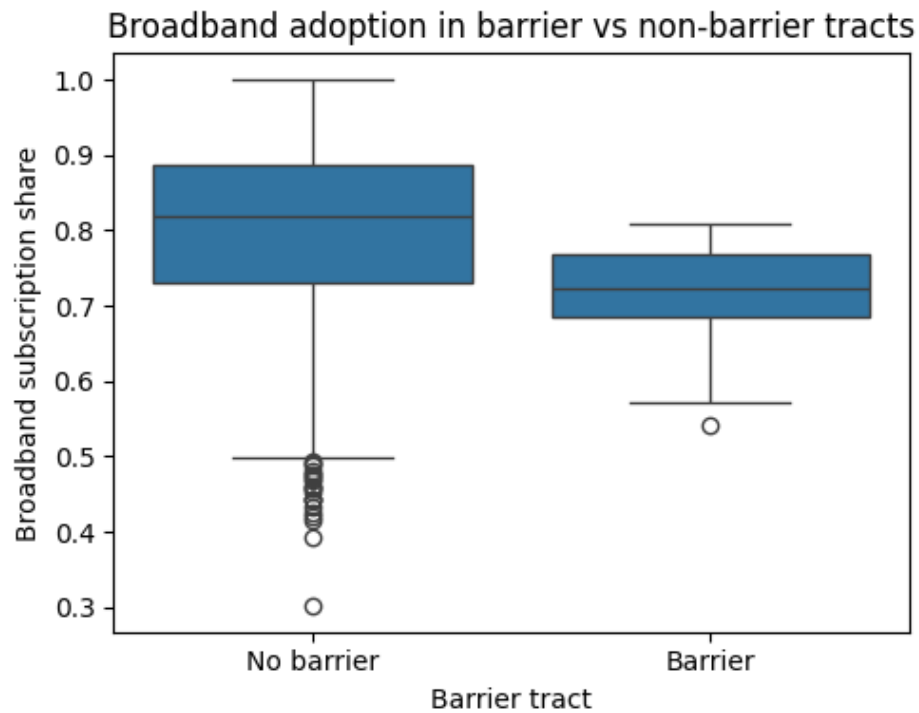


Figure 11: Broadband adoption by tract merged with food desert status. Source: H4 broadband merge section.

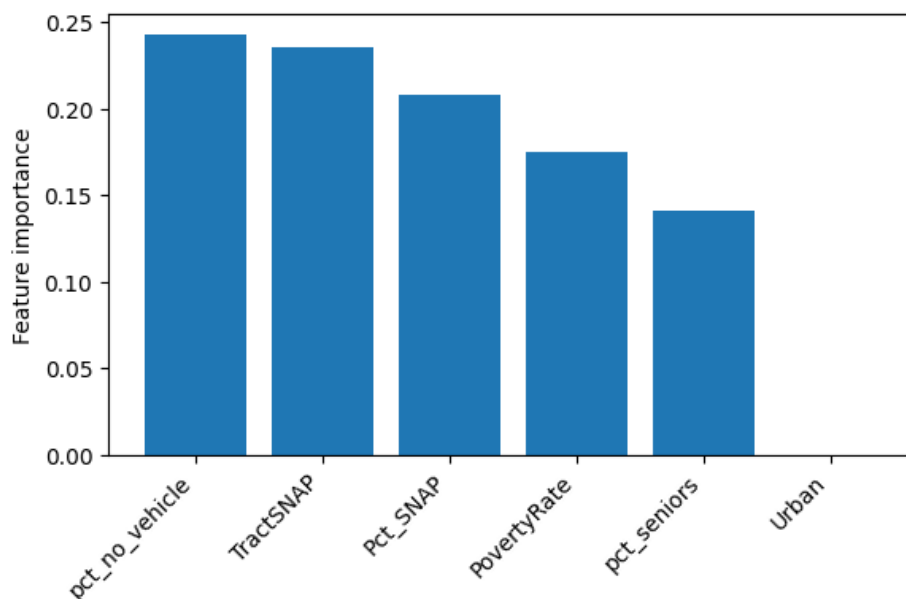


Figure 12: Logistic regression odds ratios: seniors, poverty, no vehicle predicting digital barrier. Source: H4 logit output.

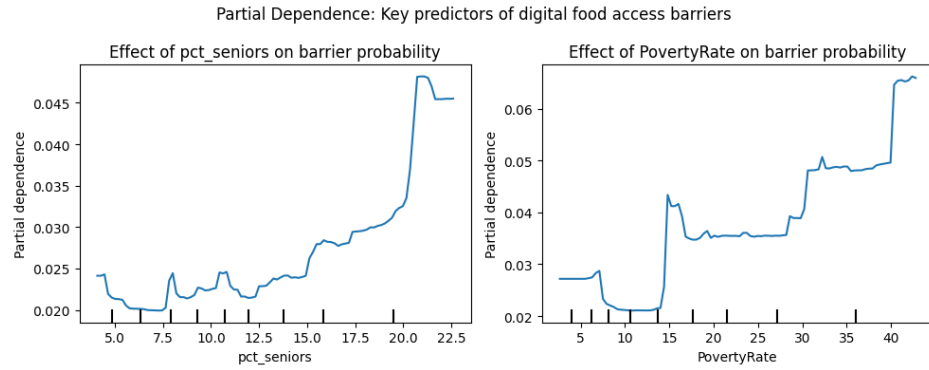


Figure 13: Interaction plot: senior share vs predicted barrier probability at low/high urbanicity. Source: H4 interaction curves.

### Dhwani (H5: Digital Barrier Index)

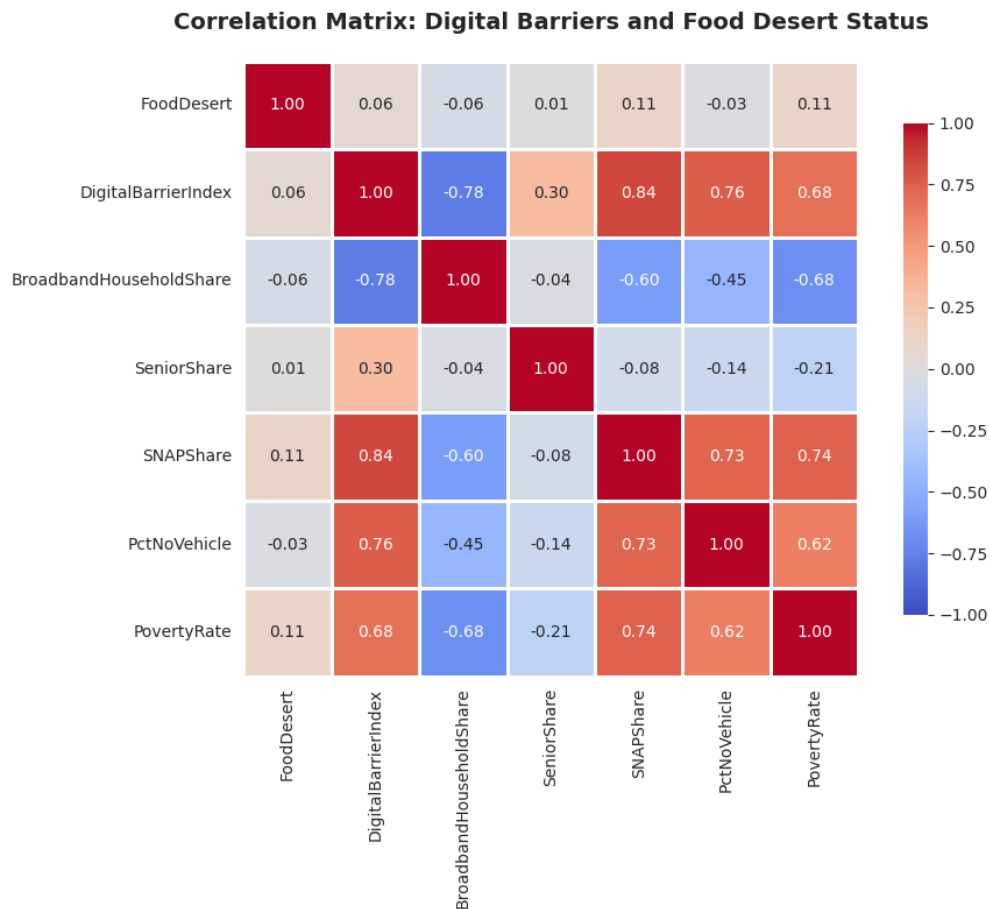


Figure 14: Correlation matrix: Digital Barrier Index, broadband, SNAP, seniors, no-vehicle vs food desert. Source: H5 EDA.



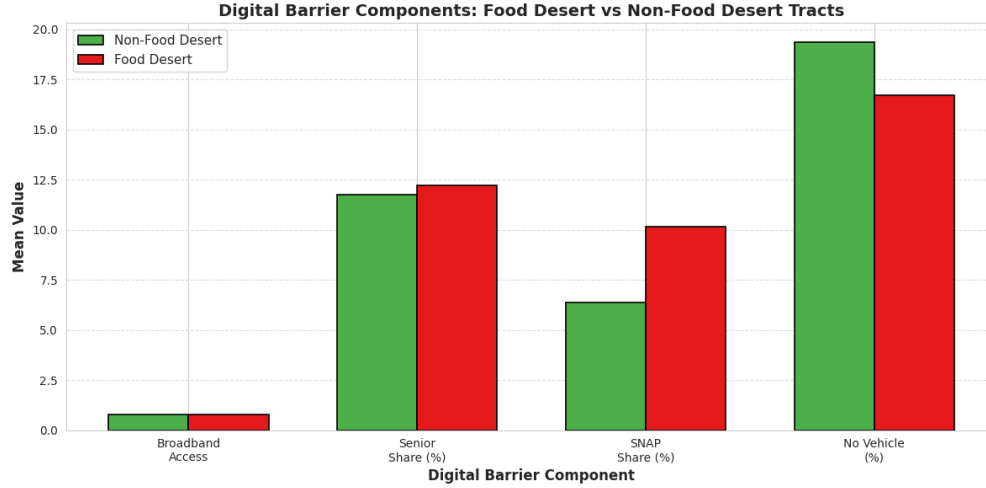


Figure 15: Component means by food desert status (broadband inverted, seniors, SNAP, no-vehicle). Source: H5 component plot.

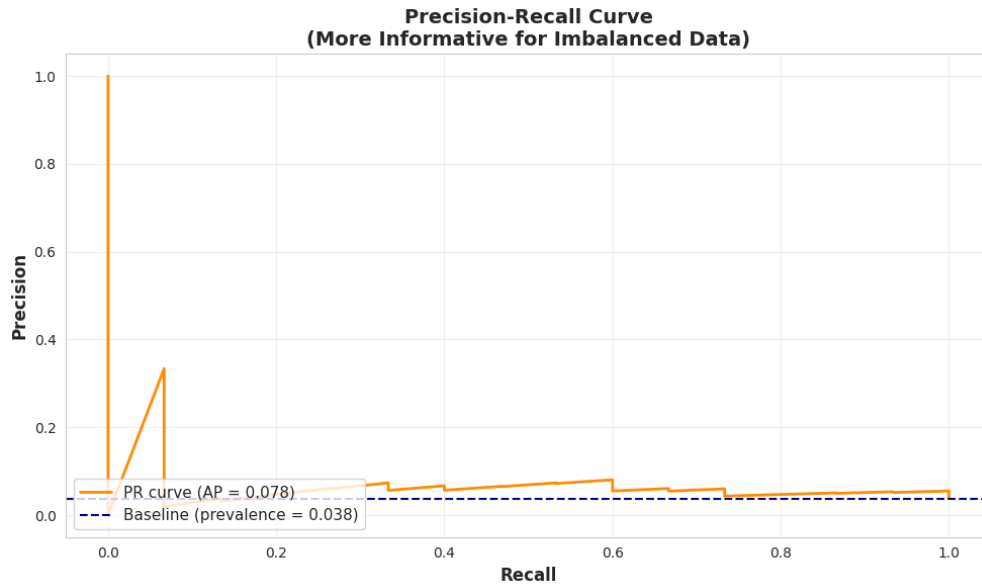


Figure 16: Precision-recall and threshold sweep for SMOTE-balanced logistic/RF predicting food deserts. Source: H5 threshold and PR plots.

## 5 Methods / ML & Statistical Analyses

### Method – Nishita

Geospatial clustering and classification using Cook County tracts; features: poverty rate, income, demographic shares, vehicle access. Models: logistic regression (reporting accuracy 0.97, ROC-AUC 0.85 but zero recall on positives), K-Means (K=4) with elbow method (WCSS knee at K=3–4). Inputs/outputs and cell references: H1 data prep and VISUALIZATION sections.

## Method – Anand

Affordability-focused classifiers. Features: poverty rate, median family income, SNAP intensity (2019, projected 2025), demographic shares, vulnerability index. Models: Random Forest, Logistic Regression, Gradient Boosting, SVM, Naive Bayes with stratified splits and standardized features where needed; evaluation via Accuracy, ROC-AUC, Precision/Recall, confusion matrix, ROC curves. Balanced class weights used; see H2 model comparison plots and printed metrics.

## Method – Krishna

Income-change analysis across 2015/2019 atlas with inflation adjustment (8% CPI). Classification: LR, RF, GBM predicting 2019 food desert using 2015 income/SNAP and change features; cross-validated F1 and ROC-AUC. Regression: Linear, Random Forest Regressor, Gradient Boosting Regressor predicting real income change percent; baseline mean vs model RMSE/R<sup>2</sup>. See H3 ML Analysis 1 and 2 cells.

## Method – Kaushik

Broadband adoption from ACS B28002 joined to atlas (GEOID). Derived barrier flag (food desert & low-income & below-median broadband). Logistic regression with standardized predictors: pct seniors, poverty rate, pct no vehicle, low-income flag/urban share interactions. Outputs: coefficient tables, odds ratios (exp coefficients), predicted curves by senior share and urbanicity. See H4 logit and interaction sections.

## Method – Dhvani

Digital Barrier Index =  $-z(\text{BroadbandShare}) + z(\text{SeniorShare}) + z(\text{SNAPShare}) + z(\text{PctNoVehicle})$ . Models: Logistic Regression (balanced), Random Forest with SMOTE oversampling; threshold tuning for optimal F1; PR and ROC evaluation; feature importance highlighting SNAP×Poverty and Senior×NoVehicle interactions. Cross-validated ROC-AUC with 5-fold CV; statistical test vs random (t-test). See H5 ML sections.

# 6 Results & Hypothesis Testing

**H1 (Spatial Clustering).** Heat maps (Fig. 1) show concentrated deserts on South and West sides; K-Means clusters align with poverty gradients (Cluster 2 high-poverty). Logistic model high accuracy but zero recall indicates imbalance; spatial evidence supports clustering.

**H2 (Affordability).** Correlations (Fig. 6) and scatter (Fig. 5) show SNAP intensity and low income strongly associated with food desert flag. RF/GBM achieve highest ROC-AUC (per H2 notebook outputs), confusion matrix highlights improved recall with class balancing. Supports affordability barrier; delivery alone insufficient without cost relief.

**H3 (Income Disparity).** Real income change analysis shows widening gap; regression models outperform baseline on RMSE/R<sup>2</sup> (Fig. 10). Classification shows food desert proportion higher in tracts with negative real income change. Hypothesis supported: income decline tracks reduced access.

**H4 (Broadband Adoption).** Logistic odds (Fig. 12) indicate senior share and poverty significantly increase barrier probability; broadband adoption median split identifies low-adoption tracts overlapping deserts. Interaction plot (Fig. 13) shows senior effect amplified in less urban tracts. Supports adoption gaps—not availability—drive digital access barriers.

**H5 (Digital Barrier Index).** Digital Barrier Index positively correlated with food desert flag (Fig. 14); component differences (Fig. 15) highlight broadband and SNAP contributions. SMOTE-balanced models yield improved recall; PR curve (Fig. 16) and threshold tuning demonstrate lift over prevalence. Hypothesis supported.

## 7 Additional Work

- **Digital Barrier Index** (H5) as composite deliverable beyond base ML/visuals.
- **Threshold optimization** for imbalanced classification (H5) and projection of SNAP growth to 2025 (H2) as extended evaluation.
- **Clustering robustness** via elbow and multi-feature cluster characterization (H1, H5).

## 8 Discussion

**Limitations.** Class imbalance reduced positive recall (H1 logistic). Some data paths (e.g., cleaned CSVs, presentation emails) unavailable in repository snapshot. SNAP projections assume uniform growth. Broadband API keys required; missing keys hinder reproducibility unless provided. Visual placeholders require exporting plots from notebooks.

**Biases.** Census/ACS sampling error, potential geocoding mismatch, and historical redlining effects not explicitly modeled. Digital adoption may proxy income, conflating effects.

**Implications.** Target interventions in South/West clusters with high poverty, low broadband adoption, and high SNAP dependence. Digital subsidies and senior-focused onboarding may complement grocery delivery.

## 9 Conclusion

Food deserts in Chicago are spatially clustered, affordability-constrained, and amplified by digital adoption gaps. Integrating geospatial, affordability, income trajectory, and digital barrier analyses yields consistent evidence across five hypotheses. Future work: fairness-aware models, causal inference for delivery services, and live dashboards for city stakeholders.

## References

## References

- [1] City of Chicago Data Portal. Chicago grocery store status map. <https://data.cityofchicago.org/>, 2024. Operational grocery store locations.

- [2] U.S. Census Bureau. American community survey 5-year estimates (table b28002). <https://api.census.gov/>, 2019. Broadband subscription by census tract.
- [3] USDA Economic Research Service. Food access research atlas 2015. <https://www.ers.usda.gov/data-products/food-access-research-atlas/>, 2015. Accessed 2025-12-08.
- [4] USDA Economic Research Service. Food access research atlas 2019. <https://www.ers.usda.gov/data-products/food-access-research-atlas/>, 2019. Accessed 2025-12-08.

## A Reproducibility Checklist

- Environment: Python 3.10+; key packages: pandas, numpy, matplotlib, seaborn, geopandas, folium, plotly, scikit-learn, statsmodels, imblearn, shapely.
- Data paths: `Grocery store.csv`; `t1_2020_17_tract.shp` (and `.dbf/.shx/.prj`); Food Access Atlas 2015/2019 CSV/XLSX; SNAP Jan 2015/2019/2025 Excel; ACS broadband via API.
- Notebook steps: run H1–H5 notebooks in order; for each figure, execute the plotting cells named in VISUALIZATION or ML sections; save outputs to `figs/` with filenames listed above.
- Commands (example, assuming data present): `pip install pandas numpy matplotlib seaborn geopandas folium plotly scikit-learn statsmodels imbalanced-learn shapely`; run notebooks; then compile LaTeX as noted.
- Randomness: set seeds where provided (e.g., `random_state=42`); stratified splits used.

## B Author Contributions

- **Nishita**: Geospatial heat maps; income vs poverty visual; K-Means clustering; logistic regression baseline (H1); Figures 1–4.
- **Anand**: SNAP/intensity affordability analysis; model comparison (RF/LR/GBM/SVM/NB); correlation and ROC visuals; Figures 5–7.
- **Krishna**: Income change over time; classification and regression models; income trajectory visualizations; Figures 8–10.
- **Kaushik**: Broadband adoption merge; logistic odds and interaction analysis; barrier probability curves; Figures 11–13.
- **Anand**: Digital Barrier Index construction; SMOTE-balanced models; threshold tuning and PR analysis; Figures 5–7.

## C Appendix: Data, Code, Repository

- Cleaned data: upload to `data/` (not present in snapshot). If size too large, include pointers to USDA/ACS/City portals.

- Scripts: reuse notebook cells; place any cleaning scripts under `data_cleaning/`.
- Notebooks: H1.ipynb–H5.ipynb (cited above).
- Repository URL: *<https://github.com/krishna-2009/Food-Delivery-as-a-Tool-Against-Food-Deserts>*.
- Figures: export from notebooks to `figs/` with names listed in preamble comment.