Fast Food Density Versus Median Income in Northwest Indiana Cities: A Bayesian Analysis

ISyE 6420: Bayesian Statistics

Adam Cuculich

December 2, 2023

Introduction:

This report presents an in-depth exploration of the potential relationship between the concentration of fast food restaurants, relative to dine-in, full-service establishments, and median household income in cities across Northwest Indiana. As someone who grew up in this region, I have some familiarity with the area, which inspired me to focus my research within this familiar context. Driven by a blend of curiosity and a desire for empirical understanding, this study aims to investigate whether a statistically significant relationship exists between these two variables, which, on the surface, might seem disconnected. A significant portion of this project was dedicated to the meticulous process of data acquisition and refinement. Recognizing the inherent limitations and uncertainties in the data, which was compiled from scratch, a Bayesian analytical approach was chosen for its robustness in handling uncertainty with limited data sets. Later in this report, I will discuss strategies for enhancing the dataset to improve the reliability and depth of future studies. In the interest of open-source analysis, all relevant code files are available on GiHub. The goal is to provide a foundation that not only reveals insights about the current data but also paves the way for more comprehensive research in this area.

Methodology

The methodology of this investigation is structured methodically to examine the relationship between fast food restaurant density and median household income in Northwest Indiana. The study focuses on cities and towns within Lake and Porter Counties with populations exceeding 500. The investigative process involves several key steps:

Identification of Cities The first step is to compile a comprehensive list of cities in Northwest Indiana, defined by their geographical location within Lake or Porter County and having a population above the specified threshold.

Geographical Centering For each city, the central point is determined based on its latitudinal and longitudinal coordinates. This geographical center serves as the focal point for subsequent data collection.

Restaurant Data Collection Utilizing the city center coordinates, the next step involves gathering data on restaurants within a 5-kilometer radius of each center. This data collection provides a basis upon which restaurants can be later categorized.

Restaurant Categorization Restaurants are classified into two distinct categories: fast food and full-service dine-in establishments. The classification criteria are based on common understanding and operational definitions for this study. Fast food restaurants are identified as those generally recognized as such or primarily offering takeaway services. In contrast, the remaining restaurants are categorized as full-service, where customers dine in and receive table service.

Income Data Collection Median household income data for each city was gathered as a critical socio-economic variable. This data, combined with the restaurant information, allows for the analysis of potential correlations between fast food restaurant density and economic indicators in the studied cities.

Analysis Preparation Following data collection, the next phase focused on preparing for the analysis. This involved calculating the proportion of fast food establishments relative to the total number of restaurants in each city, thus quantifying the density of fast food options. The study then set out to explore the potential correlation between two primary variables: the density of fast food restaurants and the median household income.

Bayesian Linear Regression Model To analyze the data, a Bayesian linear regression model is employed using probabilistic programming. This approach allows for a nuanced analysis, taking into account the uncertainties and variabilities inherent in the data. The model aims to determine whether or not there is a significant correlation between the density of fast food restaurants and the median household income in the studied areas.

Throughout this process, the methodology was applied with a commitment to thoroughness within the time and resource constraints of the study. While acknowledging the challenges in accurately classifying restaurant types and

handling geographical and demographic data, strong efforts were made to ensure the reliability of the analysis.

Data Acquisition

Identification of Cities The compilation of the list of cities in Lake and Porter Counties was sourced from geographic.org. The website provides an interface where specifying a county name yields a comprehensive list of cities within that county. The list of cities was then refined to only include cities or towns with a population greater than 500. Population numbers were obtained form `api.census.gov`, using FIPS codes for each city. The code that implements this work can be found in the attached `get_population_data.py` file.

Geographical Centering To determine the geographical centers of these cities, the Python library Geopy was employed. By inputting a city name and using the `indiana_geocoder` user agent, Geopy returns the latitude and longitude coordinates for each city. The implementation details and code are available in the attached `get_coordinates.py` file.

Restaurant Data Collection Utilizing the GPS coordinates obtained from the Geopy library, the Google Places API was instrumental in gathering lists of restaurants within a 5-kilometer radius of each city's coordinates. The data for each city was recorded in separate CSV files. All retrieved data is stored in the `restaurants directory`, and the code for this task is outlined in the `get_restaurant_data.py` file.

Restaurant Categorization Categorizing the restaurants posed a significant challenge. Initially, the `meal_takeaway` tag from the Google Places API's returned objects was considered as a categorization criterion. However, this method was found to be unreliable, as many obvious fast food establishments lacked this tag. Consequently, the lists of restaurants were provided to ChatGPT4 for classification into two categories: fast-food, take-out restaurants and full-service, dine-in establishments. While this method may not conform to a strict mathematical categorization technique, it produced reasonable results for the purposes of this analysis. The resulting list of fast food establishments used for the study can be found in the `fastfood.py`

file, stored as a variable named, `FAST_FOOD_LIST`. Further discussion on improving this methodology is included later in the report.

Income Data Collection The acquisition of median household income data commenced with obtaining FIPS codes corresponding to each city's GPS coordinates, facilitated by `api.census.gov`, as detailed in the `get_fips_codes.py` file. Subsequently, these FIPS codes were used to retrieve the median household income data for each city, with the implementation detailed in the `get_income_data.py` file.

Analysis Preparation Following the comprehensive data collection process, the fast food density for each city was computed using the script provided in the `get_fastfood_percentage.py` file. The culmination of the collected data, which forms the basis for our subsequent Bayesian analysis, is succinctly summarized in the table below:

	City	Coordinates	population	fips_code	county	fast_food_percentage	median_income
0	Cedar Lake	(41.3647578, -87.4411473)	13725	1811062	Lake County	0.3667	73611
1	Crown Point	(41.4169806, -87.3653136)	33518	1816138	Lake County	0.3500	87500
2	Dyer	(41.4942021, -87.5217068)	16422	1819270	Lake County	0.3667	90848
3	East Chicago	(41.6397857, -87.4548466)	26502	1819486	Lake County	0.3333	37807
4	Gary	(41.6020962, -87.3370646)	69739	1827000	Lake County	0.5179	34085
5	Griffith	(41.534507, -87.4255305)	16213	1830042	Lake County	0.5500	64265
6	Hammond	(41.5833618, -87.5000081)	77491	1831000	Lake County	0.6167	48107
7	Highland	(41.5536529, -87.4520484)	23695	1833466	Lake County	0.6000	71246
8	Hobart	(41.5322592, -87.2550353)	29516	1834114	Lake County	0.4500	63168
9	Lake Station	(41.5750369, -87.2389246)	13292	1841535	Lake County	0.4833	50852
10	Lowell	(41.2914244, -87.4205903)	10569	1845144	Lake County	0.4333	71463
11	Merrillville	(41.4828144, -87.3328139)	36196	1848528	Lake County	0.4167	61230
12	Munster	(41.5644798, -87.5125412)	23717	1851912	Lake County	0.6167	96938
13	Saint John	(41.45, -87.47)	19805	1866852	Lake County	0.4167	115230
14	Schererville	(41.4789246, -87.4547605)	29381	1868220	Lake County	0.3833	77530
15	Whiting	(41.6797578, -87.4944873)	4571	1884122	Lake County	0.4000	50590
16	Beverly Shores	(41.6925381, -86.9775319)	588	1805158	Porter County	0.0000	104063
17	Burns Harbor	(41.6258708, -87.1333676)	2374	1809370	Porter County	0.4333	87396
18	Chesterton	(41.6105938, -87.0641992)	14045	1812412	Porter County	0.3500	86098
19	Hebron	(41.3186482, -87.2003091)	3397	1832818	Porter County	0.5000	64276
20	Kouts	(41.3167058, -87.0258594)	2311	1840518	Porter County	0.5000	65987
21	Ogden Dunes	(41.6228148, -87.1917022)	1198	1856088	Porter County	0.3214	134250
22	Portage	(41.5758708, -87.1761455)	37540	1861092	Porter County	0.5167	63550
23	Porter	(41.6164, -87.0748)	5175	1861164	Porter County	0.3500	79247
24	South Haven	(41.5429357, -87.1395051)	4717	1871288	Porter County	0.4746	59271
25	Valparaiso	(41.4730948, -87.0611412)	33820	1878326	Porter County	0.3667	56465

Figure 1: Data obtained for Bayesian analysis

Bayesian Linear Regression Model The relationship between fast food density and median income was analyzed using a Bayesian Linear Regression model in PyMC3. Key components of the model included standardized data input, specified priors (Normal for intercept and beta, Gamma for precision), a linear function relating median income to fast food percentage, and a likelihood function. The model was calibrated with 3000 posterior samples to ensure robustness. This approach allowed for an in-depth, probabilistic understanding of the relationship between the two variables. The detailed implementation is available in the bayesian_anaylsis_ffp.ipynb file. The code below illustrates the core components of the model, were `ffp` and `mi` stand for "fast food percentage" and "median income", respectively:

```
with pm.Model() as income_model:
    # Data
    x_data = pm.Data("x_data", ffp_standardized)
    y_data = pm.Data("y_data", mi_standardized)

    # Priors
    intercept = pm.Normal('Intercept', mu=0, sigma=10)
    beta = pm.Normal('Beta', mu=0, sigma=10)
    tau = pm.Gamma("tau", alpha=0.001, beta=0.001)

    # Linear model
    mu = intercept + beta * x_data

    # Likelihood
    likelihood = pm.Normal('likelihood', mu=mu, tau=tau, observed=y_data)

    # Posterior sampling
    trace = pm.sample(3000, target_accept=0.95)
```

Figure 2: A Bayesian linear regression model

Results

The Bayesian analysis conducted on the data resulted in a posterior distribution for the model parameters, which provides insights into the relationship between fast food restaurant density and median household income. The summary statistics of the posterior distributions for each parameter are presented in the following table:

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
Intercept	-0.003	0.196	-0.377	0.361	0.002	0.002	10212.0	7819.0	1.0
Beta	-0.394	0.197	-0.767	-0.031	0.002	0.001	10047.0	7469.0	1.0
tau	1.101	0.318	0.540	1.689	0.003	0.002	9450.0	8269.0	1.0

Figure 3: Posterior Summary Statistics

The HDI for the beta parameter, ranging from -0.767 to -0.031, lies entirely below zero, suggesting with 94% confidence that an increase in fast-food restaurant density correlates with a decrease in median household income. This negative association, while statistically significant, comes with a caveat — the broad range of the HDI indicates considerable uncertainty about the magnitude of this effect. The scatter of data points around the following regression line further suggests that the predictive power of fast-food density on median income, while present, may not be strong.

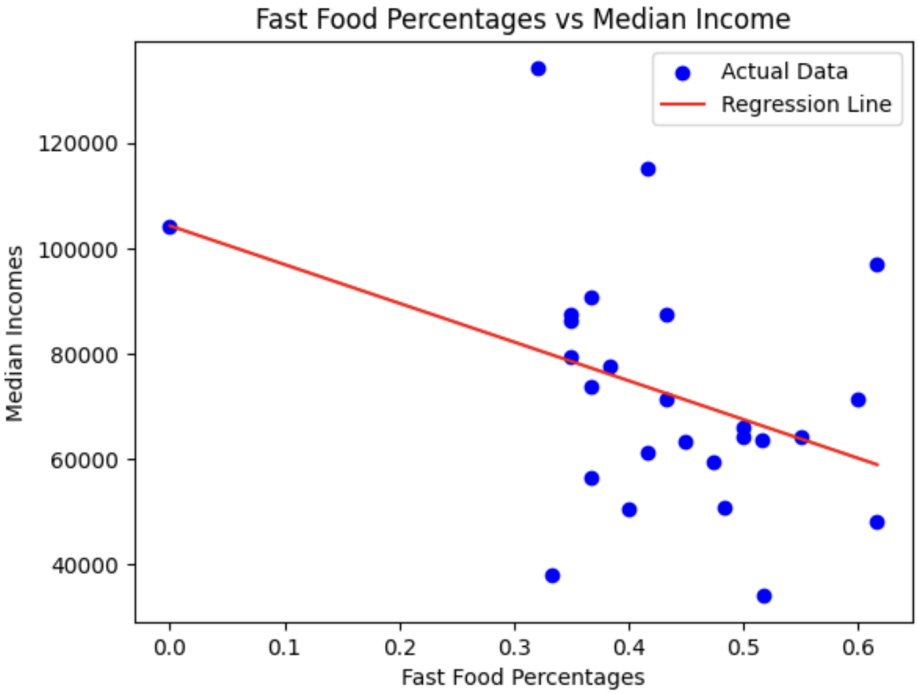

Figure 4: Fast food density compared to median household income

Suggested Improvements

Restaurant Data The current analysis is limited by a Google Places API's restriction, which yields a maximum of 60 restaurants for a given area. This constraint almost certainly results in an incomplete dataset, potentially affecting the robustness of the study. To mitigate this limitation, incorporating additional data sources, such as the Yelp API, could enhance the comprehensiveness of restaurant discovery and provide a more accurate representation of the fast-food landscape.

Restaurant Categorization The method of categorizing restaurants employed in this study relies on ChatGPT4's classification, which, while efficient, lacks systematic rigor. Future research could benefit from a more structured approach, possibly by leveraging multiple tags from both Google and Yelp data sources to classify restaurants more precisely. Furthermore, incorporating metrics such as the average duration of restaurant visits, if

accessible, could serve as a more objective measure to differentiate between fast-food and full-service, sit-down restaurants. Such data could refine the categorization process, leading to potentially more insightful results.

Conclusion:

Through Bayesian analysis, this study has identified a negative correlation between fast-food restaurant density and median household income within Northwest Indiana. Given the data collection limitations and the significant variance observed in the data points, the results suggest a cautious interpretation. The extensive HDIs underscore the necessity of further, more rigorous research before making definitive assertions about the relationship identified. Recommendations for future research include expanding the dataset beyond the Google Places API's restrictions and enhancing restaurant categorization methods. The study provides a preliminary understanding of the economic factors associated with fast-food restaurant distribution and emphasizes the need for a more detailed investigation to ascertain these findings with greater confidence. The analysis, while not conclusive on its own, sets the stage for subsequent inquiries into these complex socio-economic interactions.