Analyzing Mortality Rates From Nutritional and Metabolic Diseases by U.S. County vs. Determinants of Food Access

Group 6: Adam Cross, Mark Lee, Shreya Musini, Ellie Wang

Introduction

In our project, we decided to study the mortality rates due to nutritional and metabolic diseases and explore some factors of food access that could contribute to these rates. This topic is essential to study because it examines the complex relationship between food accessibility and public health outcomes, more specifically nutritional and metabolic diseases like diabetes, cardiovascular disease, and hypertension, which are some of the leading causes of deaths in the United States. We are particularly interested in this topic because the prevalence of these diseases are closely related to dietary habits, which are significantly influenced by the availability and affordability of healthy food options.

Food deserts are defined as areas with limited access to affordable and nutritious food and disproportionately affect low-income, rural, and minority communities, according to the Food Empowerment Project (https://foodispower.org/access-health/food-deserts/). Within these areas, residents of these communities typically struggle with an overabundance of fast-food outlets and convenience stores, both of which specialize mostly in processed, calorie-dense, and nutrient-poor food options. There are also a lack of options that are culturally appropriate for residents or do not consider dietary restrictions like lactose intolerance, gluten allergies, etc. The environmental factors combined with unhealthy eating patterns might further increase the risk of chronic diseases associated with poor nutrition.

Analyzing mortality rates from nutritional and metabolic diseases alongside food accessability factors gives us valuable insight on the health disparities and the social determinants that could possibly contribute to them. This is significant for developing a multidisciplinary approach that includes public health, urban planning, and social policy to address food insecurity and inaccessibility.

Prior Analysis on Similar Data

While we could not find other open-source analysis on our particular datasets that we used in this project, this topic of analyzing mortality rates due to nutritional and metabolic diseases and the relation with food access determinants has been studied in the past.

For instance, a study from the Journal of the National Cancer Institute analyzed the association between food deserts, food swamps, and obesity-related cancer mortality across the US. They discovered that counties with higher food desert scores faced elevated risks of suffering from obesity-related cancer. The food desert score was calculated through the consideration of factors like distance to the closest supermarket, availability of healthy food, and income. (**Source**: https://www.researchgate.net/publication/370525715_Association_of_Food_Deserts_and_Food_Swamps_With_Obesity-Related_Cancer_Mortality_in_the_US)

Another study from the American Cancer Society found a strong association between both income levels and access to healthy foods with life expectancy. Their research found that living in a food deserts is associated with a shorter life expectancy. While not particularly focusing on nutritional and metabolic diseases, this study portrays the negative health effects that food deserts and other food access indicators have on our health and wellbeing. (**Source**: https://pressroom.cancer.org/releases?item=1237)

Our study in particular focuses on the mortality rate of all types of nutritional and metabolic diseases and aims to explore the effects of some determinants of food access that are found in our datasets.

Data Cleaning

Food Access Dataset

Below is our Food Access dataset, collected from the United States Department of Agriculture's Food Access Research Atlas of 2019 that provides information of food access indicators country-wide. The data is recorded per Census Tract for a county on food access through the dataset. The dataset is large, with 72,531 instances and 147 different features with counties having multiple entries. Because each entry is identifiable by the census tract and not by county, we performed some data cleaning to aggregate the data to make the resulting data frame have each row represent one county's data. This also helped merge our data with the Cause of Death dataset, which identifies each entry by county.

Out[86]:

	CensusTract	State	County	Urban	Pop2010	OHU2010	GroupQuartersFlag	NUMGQTRS	PCTGQTRS	LILATracts_1And10	...
0	1001020100	Alabama	Autauga County	1	1912	693	0	0.0	0.000000	0	...
1	1001020200	Alabama	Autauga County	1	2170	743	0	181.0	8.341014	1	...
2	1001020300	Alabama	Autauga County	1	3373	1256	0	0.0	0.000000	0	...
3	1001020400	Alabama	Autauga County	1	4386	1722	0	0.0	0.000000	0	...
4	1001020500	Alabama	Autauga County	1	10766	4082	0	181.0	1.681219	0	...

5 rows × 147 columns

To change the food access dataset to be 1 entry per county, we combined all the tracts of a county by summing up the features. As well, some variables within the dataset represented a percentage of the tract that was calculated by dividing the county's total population feature for that census tract. Because of this, we removed all the columns that were shares in our dataset to have meaningful aggregates.

Out[89]:

	CensusTract	State	County	Urban	Pop2010	OHU2010	GroupQuartersFlag	NUMGQTRS	PCTGQTRS	LILATracts_1And10	...
0	1001020100	Alabama	Autauga County	1	1912	693	0	0.0	0.000000	0	...
1	1001020200	Alabama	Autauga County	1	2170	743	0	181.0	8.341014	1	...
2	1001020300	Alabama	Autauga County	1	3373	1256	0	0.0	0.000000	0	...
3	1001020400	Alabama	Autauga County	1	4386	1722	0	0.0	0.000000	0	...
4	1001020500	Alabama	Autauga County	1	10766	4082	0	181.0	1.681219	0	...

5 rows × 95 columns

The variables we wanted to aggregate per county were chosen below by including all numerical features and dropping identifiers such as 'CensusTract' and binary variables like 'GroupQuartersFlag' and 'HUNVFlag', which represents group quarters that had a share of >= 67 and vehicle access where >= 100 households did not have access to a vehicle and were 1/2 a mile from the supermarkets. Combined_variables shows the variables of focus to combine when doing our aggregation.

To make our new dataframe, we grouped the instances by State and County to account for the same county existing in multiple states and took the sum of all the values per each state and county, with the new name of total_{variable_name} for each variable_name in our defined combined_variables list of important columns to aggregate. 'MedianFamilyIncome' was the only column that we decided to take the average of outside of the combined_variables list.

After making the dataframe with our new summed values, we created a list of the columns of interest from combined_variables to note down all the variables from the new dataframe, df_food_new, for all the new shares we will calculate from the totaled columns.

Then iteratively, we looped over all the variables that were greated from the summed aggregation and made shares of those variables by dividing the totals by the total population that was summed up for each county. Each of these new share columns were named as {variable}_share, with variable being the columns created from summing up the values per county per state.

Out[93]:

	State	County	avg_median_family_income	total_Urban	total_Pop2010	total_OHU2010	total_NUMGQTRS	total_PCTGQTRS	to
0	Alabama	Autauga County	69337.500000	7	54571	20221	455.0	12.853691	
1	Alabama	Baldwin County	72665.741935	14	182265	73180	2307.0	41.062742	
2	Alabama	Barbour County	44792.444444	2	27457	9820	3193.0	79.735470	
3	Alabama	Bibb County	60645.500000	0	22915	7953	2224.0	26.251382	
4	Alabama	Blount County	60437.666667	1	57322	21578	489.0	7.248861	

5 rows × 183 columns

Before we merged our dataframe with the cause of death dataframe, we created 'State_Abbr' from the 'State' values in our dataframe to effectively merge by both the state and county.

Below, df_food_new is our processed, cleaned version of the food access dataset.

Out[95]:

	County	avg_median_family_income	total_Urban	total_Pop2010	total_OHU2010	total_NUMGQTRS	total_PCTGQTRS	total_LILATr
0	Autauga County	69337.500000	7	54571	20221	455.0	12.853691	
1	Baldwin County	72665.741935	14	182265	73180	2307.0	41.062742	
2	Barbour County	44792.444444	2	27457	9820	3193.0	79.735470	
3	Bibb County	60645.500000	0	22915	7953	2224.0	26.251382	
4	Blount County	60437.666667	1	57322	21578	489.0	7.248861	

5 rows × 183 columns

Death by County Dataset

This dataset is the underlying cause of death dataset, collected from 2018-2023 from the Centers for Disease Control and Prevention (CDC) WONDER database. The dataset has an instance for each county, identified by the County column by county name, state abbreviation and their County Code. The dataset has the columns County for the county, the County Code that identifies each county, the number of deaths recorded due to a nutritional and metabolic diseases in the Deaths column, the Population of the county, and the Crude Rate per 100,000 which is the percentage of deaths per 100,000 of the population.

Out[96]:

	County	County Code	Deaths	Population	Crude Rate per 100,000
0	Autauga County, AL	1001.0	35	55869	62.6
1	Baldwin County, AL	1003.0	108	223234	48.4
2	Barbour County, AL	1005.0	11	24686	Unreliable
3	Blount County, AL	1009.0	18	57826	Unreliable
4	Butler County, AL	1013.0	24	19448	123.4

For data cleaning for this dataset, we created separate columns for the State and County of each county entry for easier merging with the Food Access dataset.

As well, we investigated the missingness in our data, which resulted in one county entry with null values. Since it was only one county that had null values, we removed that one county before further data cleaning and merging.

We filtered the dataset to only include County_State, State, County, and Deaths to merge with our Food Access dataset because the mortality rate per 100,000 will be recaculated using the total population aggregated by county from the food access dataset to have one column representing the mortality rate per county.

Out[99]:

	County_State	State	County	Deaths
0	Autauga County, AL	AL	Autauga County	35
1	Baldwin County, AL	AL	Baldwin County	108
2	Barbour County, AL	AL	Barbour County	11
3	Blount County, AL	AL	Blount County	18
4	Butler County, AL	AL	Butler County	24

We looked at the data types for each column we selected, which is shown below. County_State, State, and County were all objects and Deaths was an integer value. We remade the data types for County, State, and County_State as categories for further analysis.

Out[100…

```
County_State    object
State           object
County          object
Deaths           int64
dtype: object
```

Merging Datasets

After cleaning both the datasets, we merged the two datasets on their State and County values through an inner merge to keep counties that show up in both datasets.

Out[102…

	County_State	State	County	Deaths	avg_median_family_income	total_Urban	total_Pop2010	total_OHU2010	total_NUMGQTRS
0	Autauga County, AL	AL	Autauga County	35	69337.500000	7	54571	20221	455.0
1	Baldwin County, AL	AL	Baldwin County	108	72665.741935	14	182265	73180	2307.0
2	Barbour County, AL	AL	Barbour County	11	44792.444444	2	27457	9820	3193.0
3	Blount County, AL	AL	Blount County	18	60437.666667	1	57322	21578	489.0
4	Butler County, AL	AL	Butler County	24	50170.666667	2	20947	8491	333.0

5 rows × 185 columns

After merging, we calculated the new mortality rate by dividing the Death column by the total population column and multiplying it by 100000 to get a percent value of deaths caused by nutritional and metabolic diseases that is based on the entire county.

And lastly, we set the index to be the County_State to make each row identifiable by the county.

Out[104…

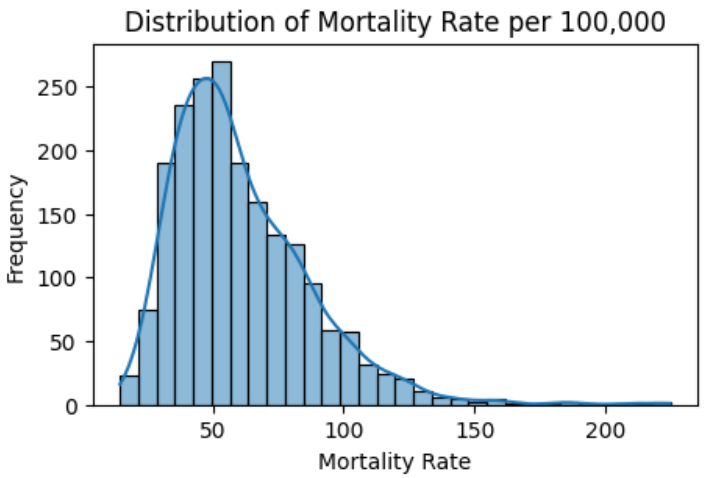
County_State	Deaths	avg_median_family_income	total_Urban	total_Pop2010	total_OHU2010	total_NUMGQTRS	total_PCTGQTRS	t
Autauga County, AL	35	69337.500000	7	54571	20221	455.0	12.853691	
Baldwin County, AL	108	72665.741935	14	182265	73180	2307.0	41.062742	
Barbour County, AL	11	44792.444444	2	27457	9820	3193.0	79.735470	
Blount County, AL	18	60437.666667	1	57322	21578	489.0	7.248861	
Butler County, AL	24	50170.666667	2	20947	8491	333.0	11.934450	

5 rows × 183 columns

EDA

To differentiate between areas with adequate and inadequate access to food, we created a new binary variable, `food_access_indicator`. This variable categorizes areas based on the proportion of the population with access to food within specific distances. Specifically, the value is set to 1 for populations with low food access—defined as areas where at least 50% of the population has limited access to food, within 1/2 mile in urban areas and 10 miles in rural areas. Areas with "better" access to food are assigned a value of 0. This indicator helps to facilitate the analysis of mortality rates in regions based on food access availability.

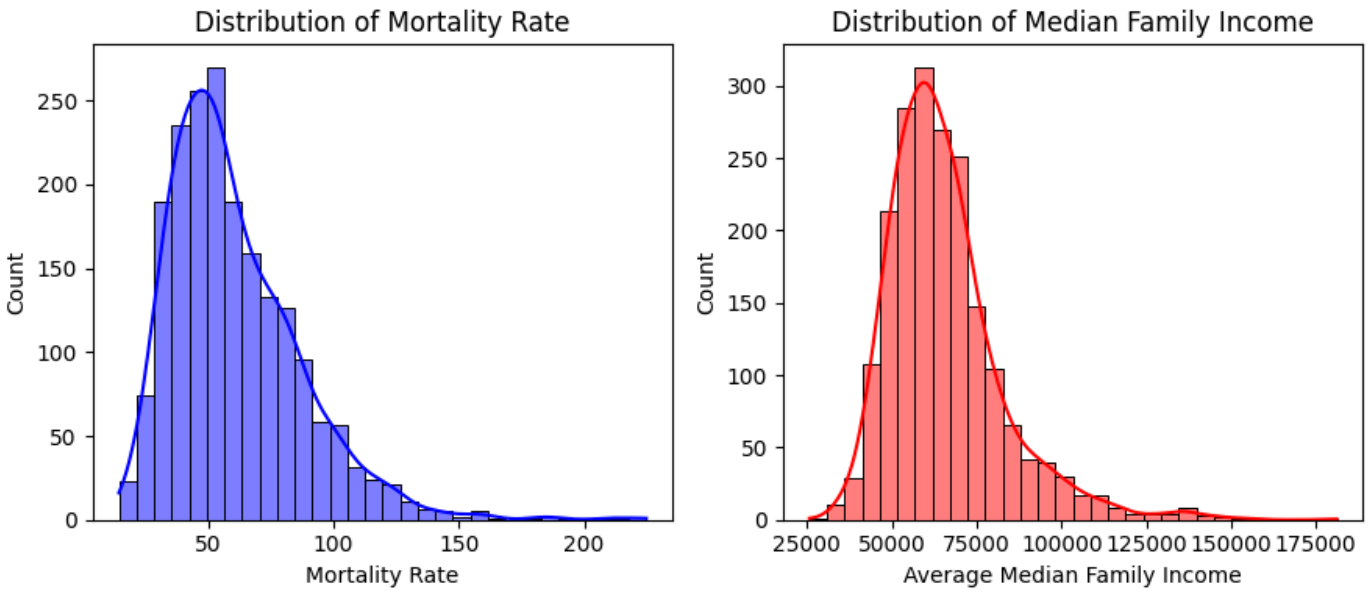
In our EDA, we are aiming to uncover insights on significance of variables and relationships between them. We start off on looking at `mortality_rate` itself. This histogram visualizes the distribution of mortality rates per 100,000 along with a kernel density estimate (KDE). The histogram displays the frequency of different mortality rate values, while the KDE provides a smoothed curve that highlights the overall distribution pattern. This visualization that we created helps in understanding the spread and central tendency of mortality rates across the dataset.



To get a better idea of the spread of the distribution of key variables like `mortality_rate` and `avg_median_family_income`, we created two plots.

1. **Mortality Rate Distribution (Left Plot)**: This plot shows the distribution of mortality rates across the dataset. The histogram, combined with a kernel density estimate (KDE), provides insight into the frequency and spread of mortality rates in the population. The distribution is right-skewed with a peak around 60-70 and a long tail extending beyond 150. This suggests that while most areas have lower mortality rates, a few outliers have significantly higher rates.

2. **Median Family Income Distribution (Right Plot)**: This plot illustrates the distribution of median family incomes. Similar to the mortality rate plot, the histogram and KDE help in understanding the variation in income levels across different counties. The distribution of family income is right-skewed, peaking around $60,000 to 70,000. A long tail extends toward higher income levels, indicating a few regions with much higher median family incomes.

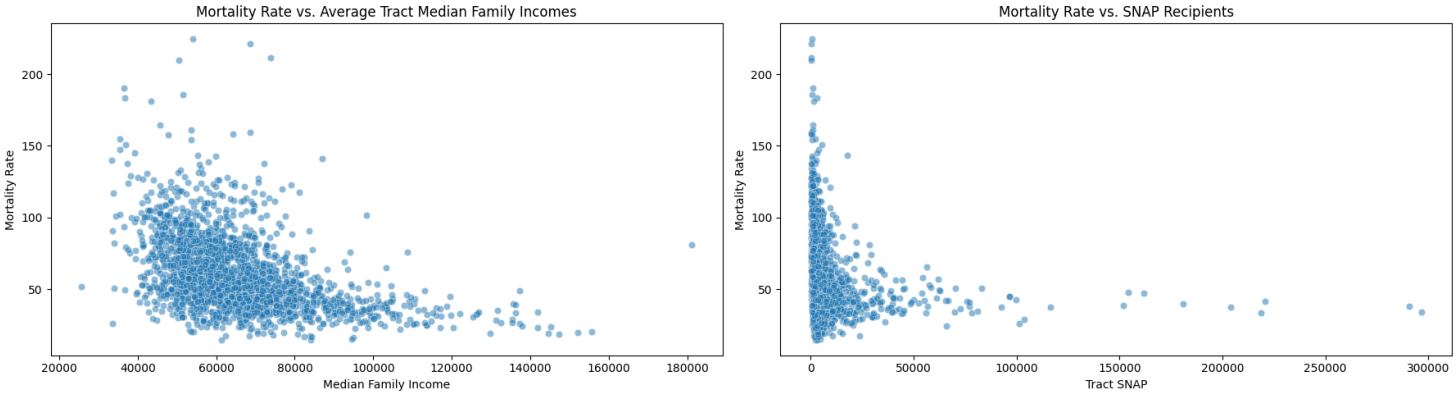


Then, we created two scatter plots to explore potential relationships between the `mortality_rate` and other key variables:

Mortality Rate vs. Average Tract Median Family Income: This plot examines how changes in median family income may be associated with variations in mortality rates. It helps to assess whether areas with higher income have different mortality outcomes. The plot shows a general negative correlation, where higher median family incomes are associated with lower mortality rates. As family income increases, mortality rates tend to cluster toward lower values, suggesting a possible link between socioeconomic status and health outcomes.

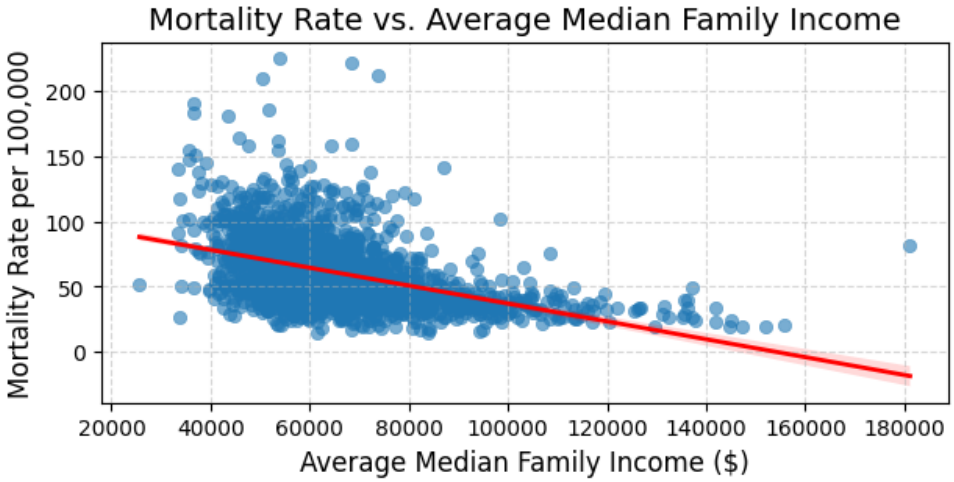
Mortality Rate vs. SNAP Recipients: This plot looks at the relationship between the number of SNAP (Supplemental Nutrition Assistance Program) recipients in a tract and the mortality rate. This plot reveals a strong negative relationship between the number of SNAP recipients and mortality rates, with higher SNAP participation generally associated with higher mortality rates. As the number of SNAP recipients decreases, mortality rates tend to be lower, potentially indicating that areas with more SNAP recipients face higher mortality due to economic hardship or related factors.

Both plots suggest that income and socioeconomic factors may influence mortality rates, with wealthier areas generally experiencing lower mortality.



Next, we created a scatter plot to visualize the relationship between mortality rate and average median family income. The plot includes a regression line (in red) to help highlight the trend between these two variables. This visualization allows us to assess whether areas with higher median family income show different mortality rates.

- There is a clear negative correlation, indicating that as average median family income increases, the mortality rate tends to decrease.
- Most data points are concentrated in the lower income ranges (between $40,000 to 80,000$), and the mortality rate tends to cluster between 50 and 150 per 100,000 people.
- This suggests that higher family incomes are generally associated with better health outcomes and lower mortality rates.

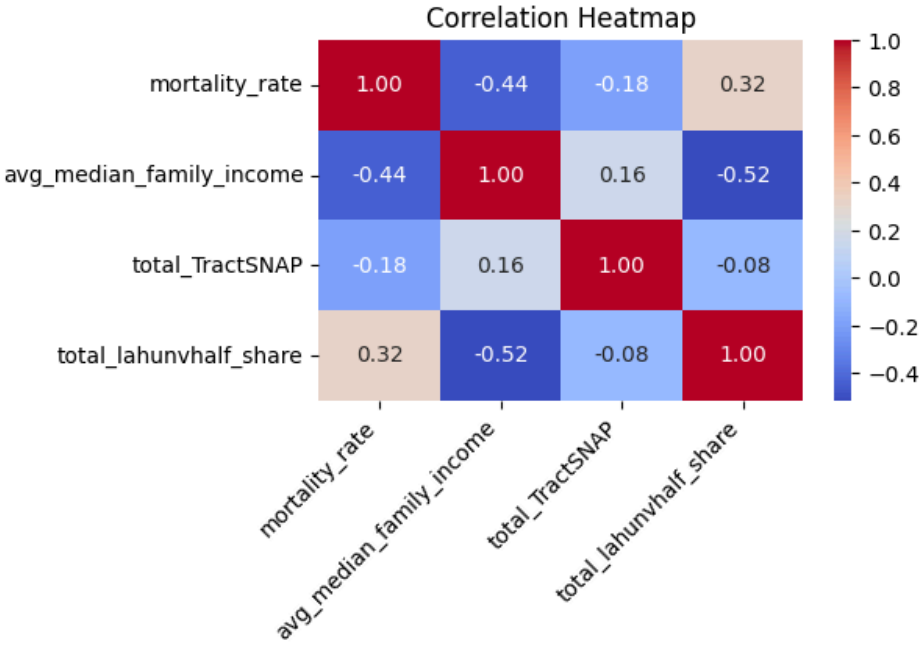


We then created a heatmap that aims to visualize the correlation matrix of a selected set of variables in the dataset. Specifically, it calculates the pairwise correlation coefficients between the mortality rate, average median family income, total SNAP recipients, and share of households with low access to food (total_lahunvhalf_share).

The correlation heatmap reveals key relationships between variables:

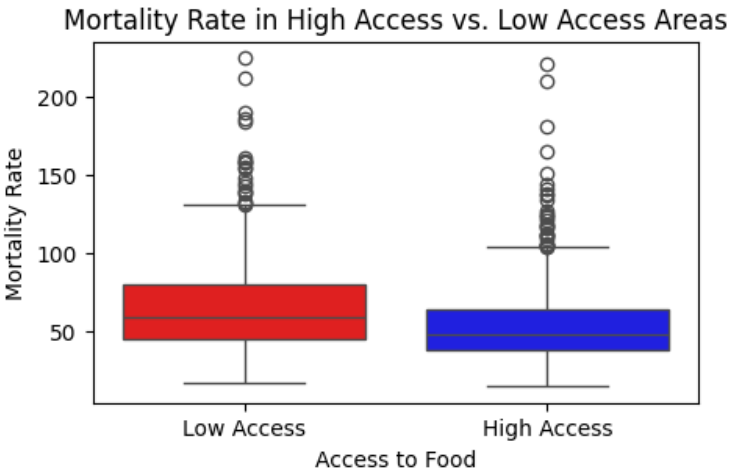
- **Mortality rate** is negatively correlated with **median family income** (-0.44), indicating that areas with higher income tend to have lower mortality rates.
- **Food inaccessibility (total_lahunvhalf_share)** has a moderate positive correlation with **mortality rate** (0.32), suggesting that limited food access may contribute to higher mortality rates.
- **SNAP participation (total_TractSNAP)** has a weak negative correlation with **mortality rate** (-0.18), implying that areas with more SNAP recipients may have slightly lower mortality rates, though the relationship is weak.

- **Income** and **food inaccessibility** are inversely correlated (-0.52), meaning lower-income areas are more likely to experience food inaccessibility.



Finally we generated a boxplot to visualize and compare the mortality rates in areas with high access to food versus areas with low access. Key observations include:

- Low-access areas have a higher median mortality rate than high-access areas.
- The interquartile range (IQR) is slightly wider for low-access areas, indicating more variability in mortality rates.
- Both groups have outliers with extremely high mortality rates, though they appear more frequent in low-access areas.



Then, we grouped the data by the 'food_access_indicator' variable (which indicates high or low food access) and computes summary statistics for the mortality rate within each group to understand the distribution of mortality rates in both high-access and low-access areas and check for any significant differences between them.

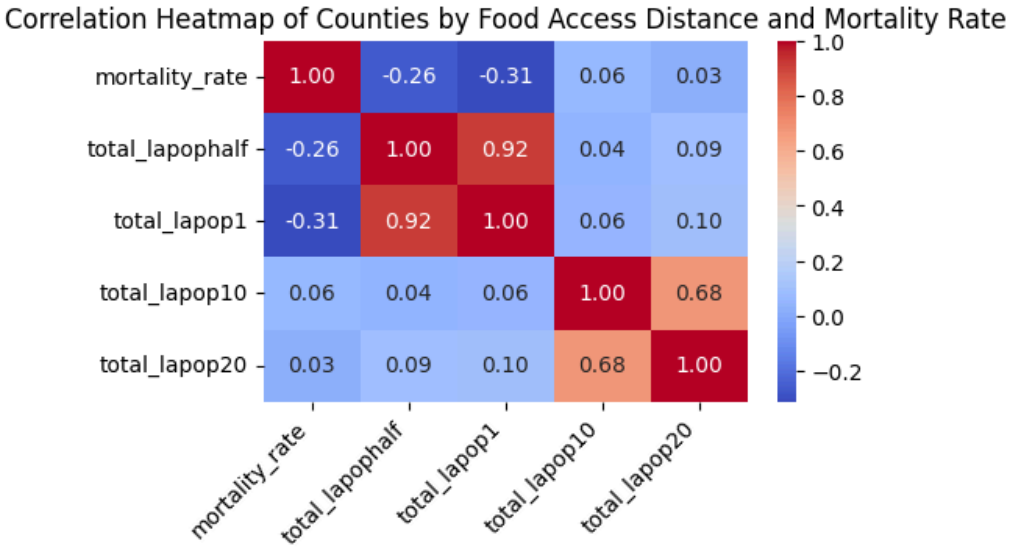
```
                        count       mean        std        min        25%  \
food_access_indicator
0                      1299.0  64.416608  26.382220  16.559755  45.032854
1                       682.0  53.651939  24.817995  14.518327  37.351517

                            50%        75%         max
food_access_indicator
0                     59.253407  79.442278  224.763022
1                     48.200595  63.786636  220.896841
```

The summary statistics show that low-access areas have a higher mean mortality rate (64.42 vs. 53.65 per 100,000) and greater variability than high-access areas. The median mortality rate is also higher in low-access areas (59.25 vs. 48.20). While both groups exhibit wide mortality rate ranges, the slightly lower standard deviation in high-access areas suggests less variability. These differences indicate a potential link between food access and mortality rates, and we explore this idea later.

County Mortality Rate and Supermarket Distances

We next wanted to explore the relationship between a county's mortality rate and its population's distance from a supermarket. The heatmap below shows the correlation between a county's mortality rate and the count of people who live beyond a half mile (total_lapophalf), 1 mile (total_lapop1), 10 miles (total_lapop10), and 20 miles (total_lapop20) from a supermarket.

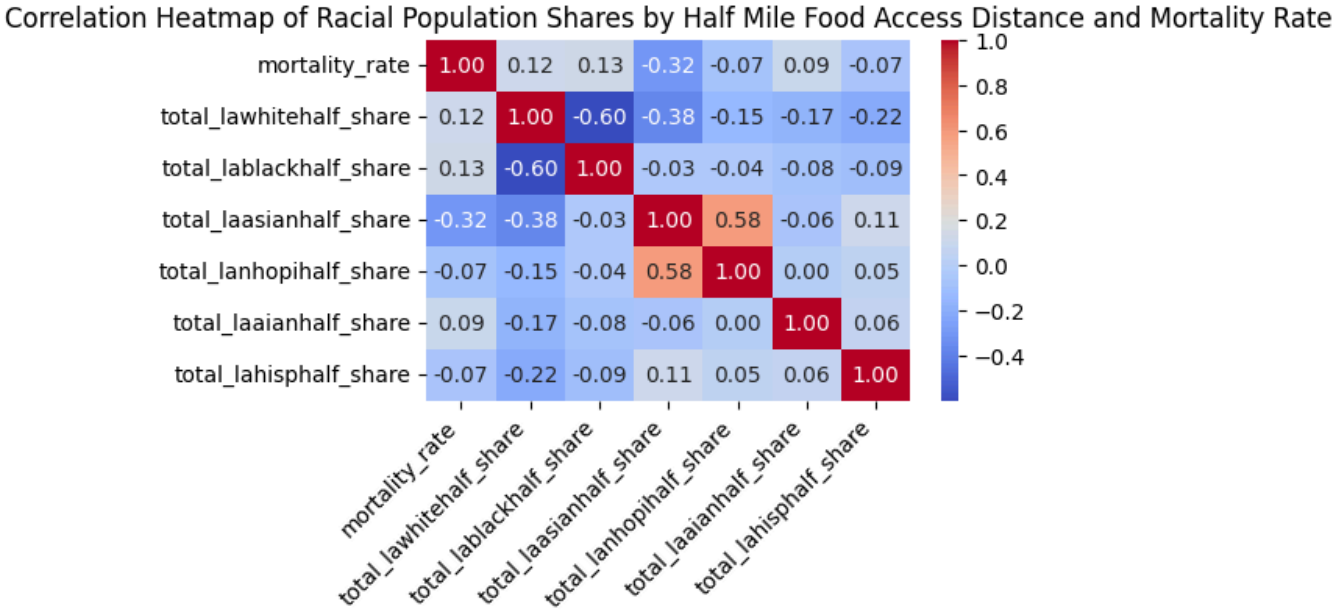
We can see that a distance beyond a half to whole mile seems to have the greatest correlation with a county's mortality rate. This result informs our further exploration of shorter minimum distances from supermarkets.



After identifing that a population with a shorter minimum distance from a supermarket has having a greater correlation with a county's mortality rate, we wanted to explore the correlation between subgroups within the half mile distance category.

The heatmap below shows the correlation between mortaility rate and the percent of its population beyond a half mile from a supermarkt broken down by race. The race categories are as follows:

- White population beyond a half mile (total_lawhitehalf_share)
- Black or African American population beyond a half mile (total_lablackhalf_share)
- Asian population beyond a half mile (total_laasianhalf_share)
- Native Hawaiian or Pacific Islander population beyond a half mile (total_lanhopihalf_share),
- American Indian or Alaskan Native population beyond a half mile (total_laaianhalf_share),
- Other or mixed race population beyond a half mile (total_laomultirhalf_share)
- Hispanic population beyond a half mile (total_lahisphalf_share).



Correlation Significance

After examining the correlation between a county population's distance from a supermarket and mortality rate we wanted to ensure that these measured relationships were statistically significant. The P-value below shows the statistical significance of the percenatage of people in a county who live beyond a half mile from a supermarket on mortality.

```
Pearson Correlation: 0.2585942453002774
P-value: 1.2454974446155588e-31
The correlation is statistically significant.
```

The correlation between the average medain family income across cencus tracts in a county and mortality rate is also obersved below. The moderate correclation between the two can be seen as statistically significant as well.

```
Pearson Correlation: -0.4435579811025938
P-value: 3.2113295857416646e-96
The correlation is statistically significant.
```

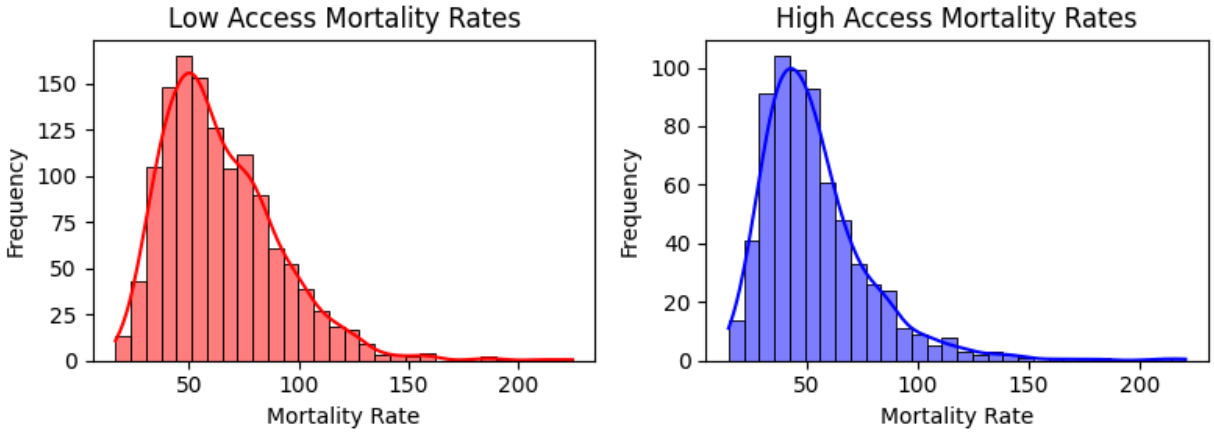
The observed significance of both these variables on mortality rate reinforces their further analysis in this report.

Hypothesis Testing

To begin our hypothesis testing, we thought of using a t-test to determine whether there is a significant difference between the means of two groups, `mortality_rate` and `food_access_indicator`. Before proceeding with a t-test, we had to check the three key assumptions of a t-test:

1. **Normality** – the data in each group should be approximately normally distributed
2. **Equal Variance** – the variability (spread) of the two groups should be similar
3. **Independence** – the observations in one group should not influence observations in the other group

Therefore, we plotted histograms to visualize the distribution mortality rates in low and high food access areas.



From these histograms, we observe that both distributions appear to be right-skewed, indicating potential deviation from normality. This suggests the need for further assessment of normality before conducting a parametric test like the t-test. To further check for normality, we then proceeded to create QQ plots. We know that if the data follows a normal distribution, the points should align closely with the red diagonal line.

QQ plots of low vs high food access areas

In our plots, we observe significant deviations from normality, particularly at the tails, which suggests the presence of skewness and outliers. Given the results from both the histograms and Q-Q plots, it is evident that the normality assumption for a t-test is likely violated. Therefore, we proceeded with an alternative path.

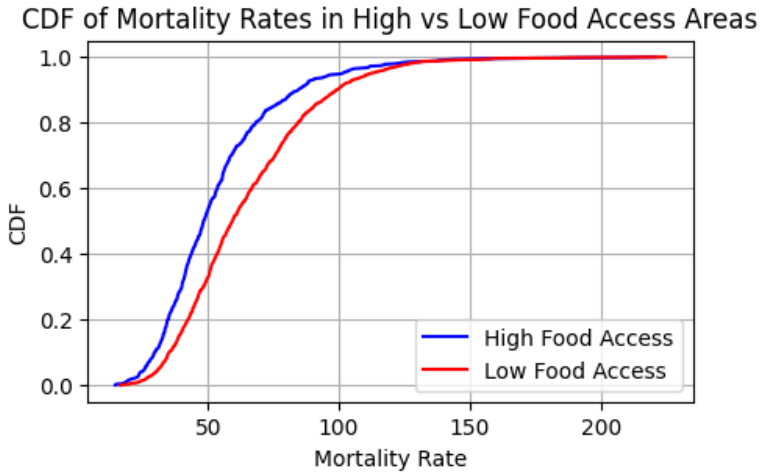
Kolmogorov-Smirnov (KS) test

We decided to use the Kolmogorov-Smirnov (KS) test. This test is a non-parametric statistical test that compares the cumulative distribution functions (CDFs) of two independent samples. Unlike the t-test, it does not assume normality or equal variance, making it more robust for skewed or non-normally distributed data, which was what we had.

```
KS Test statistic: 0.21065493646138808, p-value: 7.224501751217153e-18
```

With our KS test statistic of 0.211 and a p-value of basically 0, since this p-value is much smaller than 0.05, we reject the null hypothesis, meaning there is a statistically significant difference in the distributions of mortality rates between low and high food access areas. We can assume that mortality rates are distributed differently in counties with low vs. high food access, supporting the idea that food access may be associated with mortality outcomes.

To further investigate the difference in mortality rate distributions between areas with high and low food access, we plot their Cumulative Distribution Functions (CDFs).



Comparing the two lines, the blue line (high food access) lies above the red line (low food access), suggesting that mortality rates tend to be lower in areas with high food access. The red line shifts to the right, indicating that mortality rates tend to be higher in these areas.

The plot suggests that areas with low food access experience higher mortality rates, as their CDF accumulates more slowly than areas with high food access. We can further say that mortality rates are significantly different between high and low food access areas, with higher mortality rates in low food access regions.

Permutation Test

Then, we decided to perform a permutation test as it is also a non-parametric method that does not rely on assumptions about normality, which were violated in our dataset. The permutation test directly evaluates the difference in mortality rates by randomly shuffling group labels and comparing observed differences to a null distribution, providing a more robust assessment of statistical significance.

We first began to prepare the data by treating a single DataFrame that combines mortality rates from both high and low food access areas, along with labels indicating which group each data point belongs to.

Out[124…

	mortality_rate	is_high_food_access
0	64.136629	1.0
1	30.361299	1.0
2	38.583082	1.0
3	53.624437	1.0
4	46.927252	1.0

Our permutation test works to compare the mean mortality rates between two groups: high food access and low food access. It first calculates the observed difference in the means of the mortality rates between these two groups. Then, it randomly permutes the

combined data of both groups multiple times (10,000 permutations in this case) and recalculates the mean difference for each permutation. The test then calculates the p-value by determining the proportion of permutations where the absolute difference in means is greater than or equal to the observed difference. A p-value of 0.0 indicates that the observed difference in means is highly unlikely to have occurred by chance.

Conclusion: The observed difference in mortality rates between high and low food access groups is statistically significant, with a p-value of 0.0, suggesting that the observed difference is highly unlikely to have occurred randomly. This implies that food access may have a significant impact on mortality rates.

```
Observed difference in means: -10.764668907420216
Permutation test p-value: 0.0
```

Regression Modeling

We want to take a deeper look into the key risk factors that contribute to the highest county-wide mortality rates in our data. To accomplish this, we can contruct a regression model with an optimal set of variables to predict the mortality rate. This model can potentially give us insight on what populations suffer the most from food deserts and how to focus public health efforts to reduce food insecurity among US counties.

Interested Variables

There are over 180 variables associated with each county in our cleaned, merged dataset, so we need to cut down and focus just on the variables we are most interested in. We identified the following 10 variables to include in our model, based on their description and correlation matrix defined above in the exploratory data analysis section. Intuitively, these variables can shed valuable insight on the nutritional and metabolic mortality rate in these US counties.

The variables are described as follows:

1. 'avg_median_family_income' [continuous quantitative]: The arithmetic average of the median income of each tract within a county.
2. 'total_lapophalf_share' [continuous quantitative]: Share of county population that are beyond 1/2 mile from supermarket
3. 'total_lalowihalf_share' [continuous quantitative]: Share of county population that are low income individuals beyond 1/2 mile from supermarket
4. 'total_lawhitehalf_share' [continuous quantitative]: Share of county population that are white beyond 1/2 mile from supermarket
5. 'total_lablackhalf_share' [continuous quantitative]: Share of county population that are Black or African American beyond 1/2 mile from supermarket
6. 'total_laasianhalf_share' [continuous quantitative]: Share of county population that are Asian beyond 1/2 mile from supermarket
7. 'total_laaianhalf_share' [continuous quantitative]: Share of county population that are American Indian or Alaska Native beyond 1/2 mile from supermarket
8. 'total_lahisphalf_share' [continuous quantitative]: Share of county population that are of Hispanic or Latino ethnicity beyond 1/2 mile from supermarket
9. 'total_lahunvhalf_share' [continuous quantitative]: Share of county housing units that are without vehicle and beyond 1/2 mile from supermarket
10. 'total_Pop2010' [discrete quantitative]: Population count from 2010 census for each county

The response variable 'mortality_rate' is a continuous quantitative data type, which is the crude mortality rate due to nutritional and metabolic diseases per 100,000 in a given US county.

It is important for our features to have limited (and ideally none) missing values. As shown below, each of our 10 interested variables have no missing values (except for 'avg_median_family_income' with a singular missing value). This is a result of the usage of groupby, which summed together relevant numerical data for each tract of a county to give a county total. Values from tracts that were missing were simply not included in this total.

Consequently, we do not need to apply any imputation techniques to handle missing values.

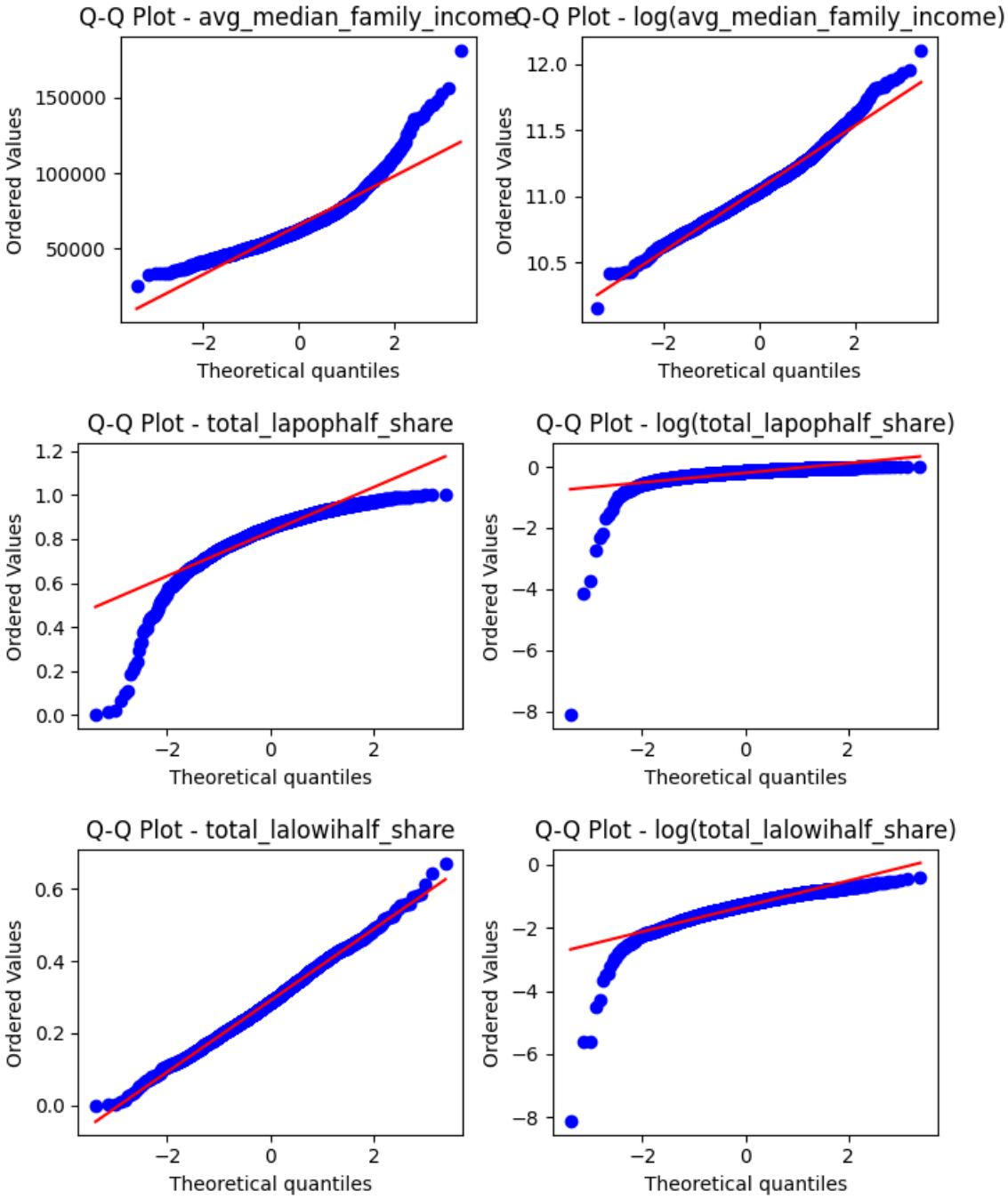
```
Missing values for feature avg_median_family_income: 1 (0.0% missing)
Missing values for feature total_lapophalf_share: 0 (0.0% missing)
Missing values for feature total_lalowihalf_share: 0 (0.0% missing)
Missing values for feature total_lawhitehalf_share: 0 (0.0% missing)
Missing values for feature total_lablackhalf_share: 0 (0.0% missing)
Missing values for feature total_laasianhalf_share: 0 (0.0% missing)
Missing values for feature total_laaianhalf_share: 0 (0.0% missing)
Missing values for feature total_lahisphalf_share: 0 (0.0% missing)
Missing values for feature total_lahunvhalf_share: 0 (0.0% missing)
Missing values for feature total_Pop2010: 0 (0.0% missing)
Missing values for feature mortality_rate: 0 (0.0% missing)
```

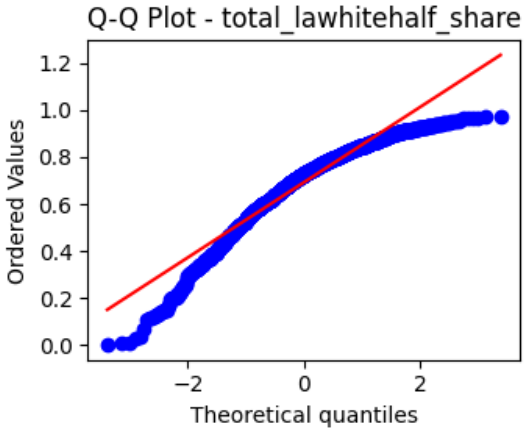
We drop the one county that had missing data for its family income.

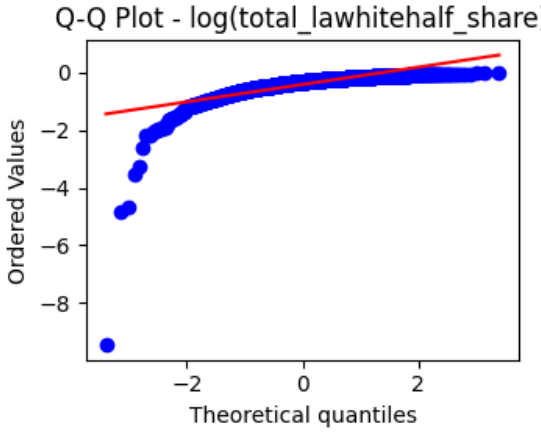
Transforming Data

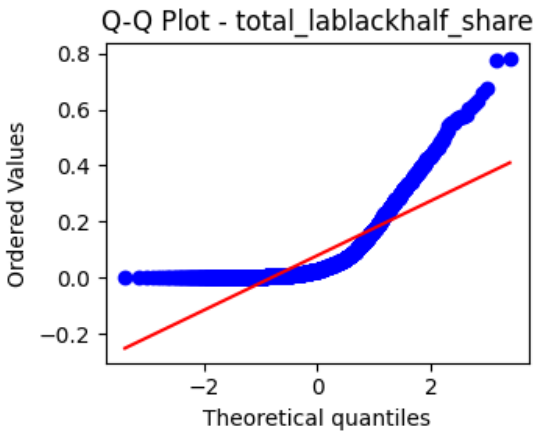
To better capture the relationship between mortality rate (dependent variable) and our key independent variables, we applied a log transformation to our data. This transformation helps linearize relationships, making them more suitable for OLS regression. Additionally, it mitigates the skewness often observed in variables such as income and population.

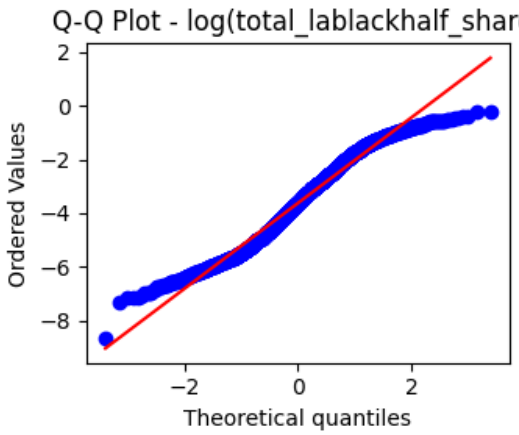
The QQ-plots below illustrate how the log transformation affects the distribution of these variables. Notably, many variables become more normally distributed, which improves model performance and interpretation.

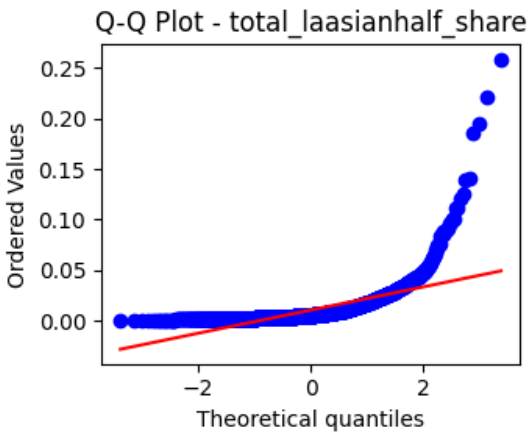


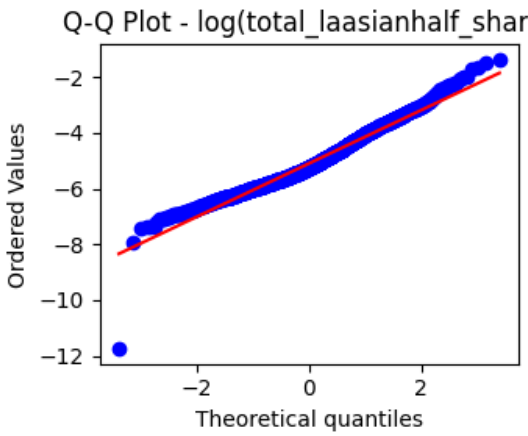

Q-Q Plot - total_lawhitehalf_share
Ordered Values
Theoretical quantiles

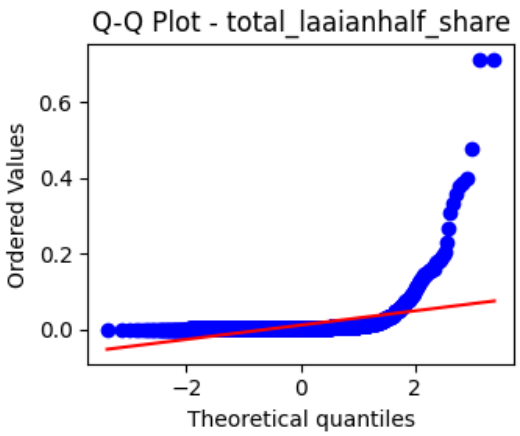

Q-Q Plot - log(total_lawhitehalf_share)
Ordered Values
Theoretical quantiles

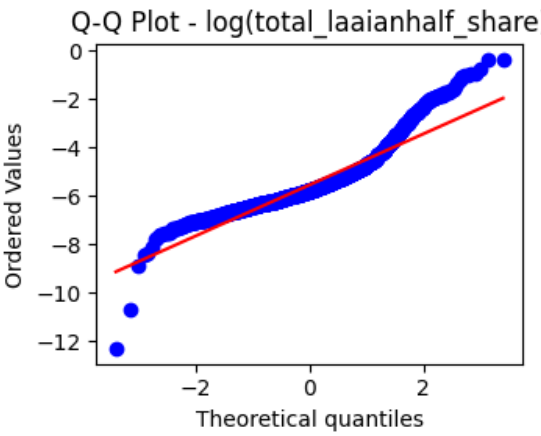

Q-Q Plot - total_lablackhalf_share
Ordered Values
Theoretical quantiles

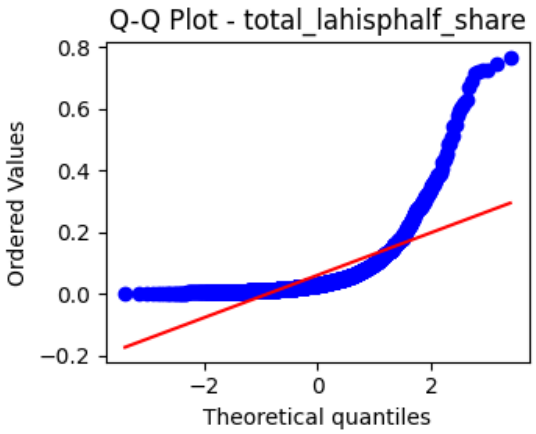

Q-Q Plot - log(total_lablackhalf_share)
Ordered Values
Theoretical quantiles

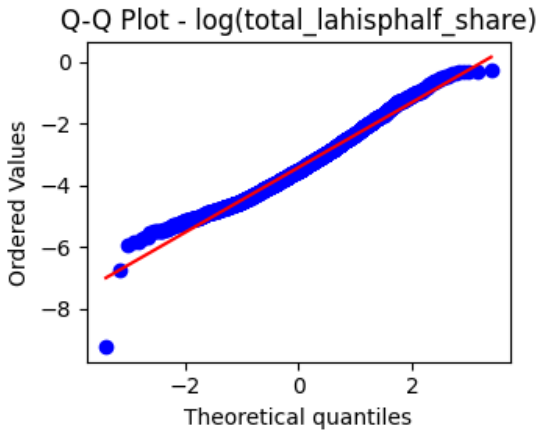

Q-Q Plot - total_laasianhalf_share
Ordered Values
Theoretical quantiles

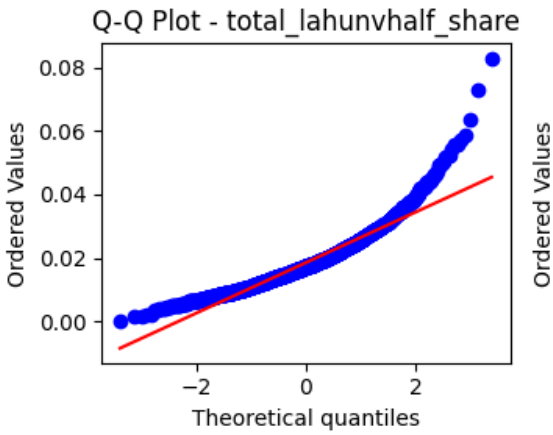

Q-Q Plot - log(total_laasianhalf_share)
Ordered Values
Theoretical quantiles

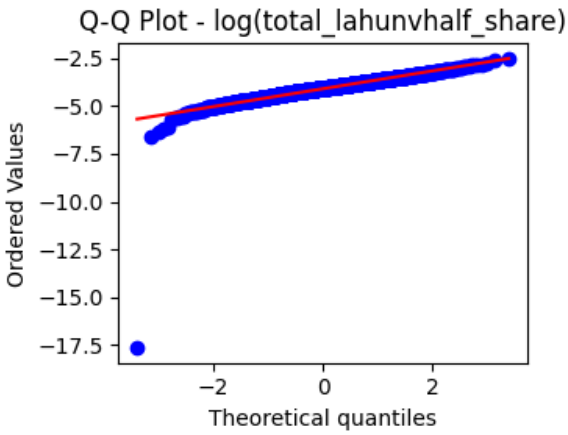

Q-Q Plot - total_laaianhalf_share
Ordered Values
Theoretical quantiles

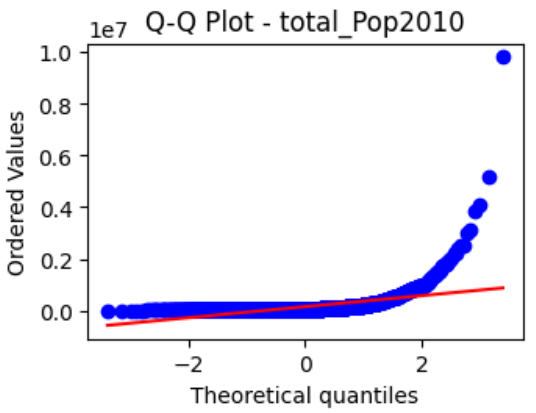

Q-Q Plot - log(total_laaianhalf_share)
Ordered Values
Theoretical quantiles

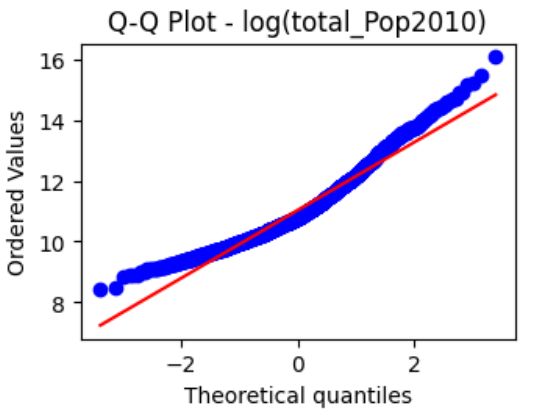


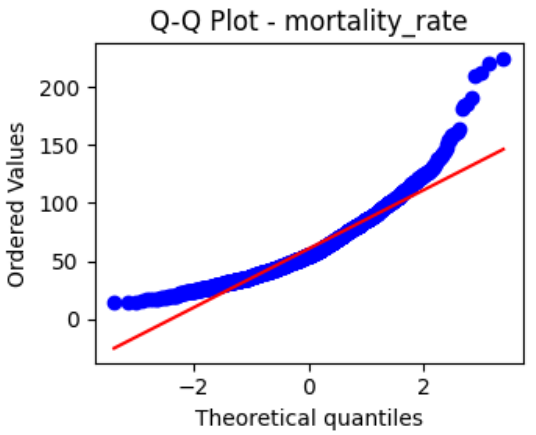


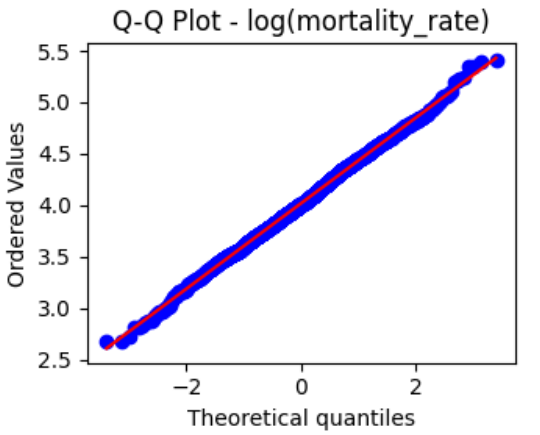












The new new log adjust values of our data can be seen below. All variables moving moving forward will be refering to thier log transformed selves.

Out[131…

County_State	avg_median_family_income	total_lapophalf_share	total_lalowihalf_share	total_lawhitehalf_share	total_lablackhalf_shar
Autauga County, AL	11.146741	-0.097563	-1.257507	-0.343675	-1.80363
Baldwin County, AL	11.193625	-0.095768	-1.331863	-0.245874	-2.50035
Barbour County, AL	10.709795	-0.144512	-0.835999	-0.842027	-0.94476
Blount County, AL	11.009368	-0.042590	-1.062628	-0.117505	-4.38344
Butler County, AL	10.823186	-0.147994	-0.884602	-0.724489	-1.02484

Split Data

To evaluate our model's performance, we implemented a train-test split, dividing the dataset into an 80% training set and a 20% test set. This approach allows us to assess how well the model generalizes to unseen data while reducing the risk of overfitting. Our goal is to ensure the model produces results that are robust and generalizable, rather than overly tailored to the training data.

Baseline Model

We begin by creating a model using all of our features and running an OLS regression on the training set. This serves as a baseline model from which we can refine our final version. The model summary reveals that 6 out of 10 key independent variables are statistically significant at the 5% level, providing insights into which predictors have a meaningful impact on mortality rate.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         mortality_rate   R-squared:                       0.379
Model:                            OLS   Adj. R-squared:                  0.375
Method:                 Least Squares   F-statistic:                     95.83
Date:                Sun, 16 Mar 2025   Prob (F-statistic):          1.05e-154
Time:                        14:11:16   Log-Likelihood:                -476.04
No. Observations:                1584   AIC:                             974.1
Df Residuals:                    1573   BIC:                             1033.
Df Model:                          10
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                       12.4063      1.119     11.091      0.000      10.212      14.600
avg_median_family_income        -0.6568      0.103     -6.364      0.000      -0.859      -0.454
total_lapophalf_share            0.2701      0.113      2.398      0.017       0.049       0.491
total_lalowihalf_share          -0.2327      0.075     -3.122      0.002      -0.379      -0.087
total_lawhitehalf_share         -0.0981      0.055     -1.791      0.074      -0.206       0.009
total_lablackhalf_share         -0.0035      0.008     -0.448      0.654      -0.019       0.012
total_laasianhalf_share         -0.0599      0.015     -3.998      0.000      -0.089      -0.031
total_laaianhalf_share           0.0088      0.009      0.973      0.330      -0.009       0.026
total_lahisphalf_share           0.0079      0.010      0.780      0.436      -0.012       0.028
total_lahunvhalf_share           0.1077      0.027      3.950      0.000       0.054       0.161
total_Pop2010                   -0.1104      0.012     -9.533      0.000      -0.133      -0.088
==============================================================================
Omnibus:                       29.484   Durbin-Watson:                   1.984
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               32.344
Skew:                          -0.293   Prob(JB):                     9.48e-08
Kurtosis:                       3.382   Cond. No.                     2.54e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.54e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model Refinement

Next, we implement backward selection to refine our model by minimizing the Bayesian Information Criterion (BIC). This process systematically removes less significant variables, leading to a more parsimonious and optimized model. Based on our selection, the optimal independent variables are:

- avg_median_family_income
- total_Pop2010
- total_laasianhalf_share
- total_lahunvhalf_share

This refined model improves interpretability while maintaining strong predictive performance.

```
Criterion: 1025.9498611911288
Criterion: 1019.1317910285813
Criterion: 1013.7411974354857
Criterion: 1011.6342021043536
Criterion: 1006.0008457120337
Criterion: 1005.5930035342363
```

Out[136…
```
{'avg_median_family_income',
 'total_Pop2010',
 'total_laasianhalf_share',
 'total_lahunvhalf_share'}
```

Final Model

We then fit an OLS regression model using the optimal independent variables identified through backward selection. The model summary provides insights into its performance, helping us assess improvements in predictive power and interpretability compared to the initial model.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         mortality_rate   R-squared:                       0.372
Model:                            OLS   Adj. R-squared:                  0.370
Method:                 Least Squares   F-statistic:                     233.8
Date:                Sun, 16 Mar 2025   Prob (F-statistic):          9.10e-158
Time:                        14:11:22   Log-Likelihood:                -484.38
No. Observations:                1584   AIC:                             978.8
Df Residuals:                    1579   BIC:                             1006.
Df Model:                           4
Covariance Type:            nonrobust
================================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                    9.8858      0.557     17.751      0.000       8.793      10.978
avg_median_family_income    -0.4277      0.051     -8.410      0.000      -0.527      -0.328
total_lahunvhalf_share       0.0746      0.022      3.321      0.001       0.031       0.119
total_Pop2010               -0.1025      0.011     -9.758      0.000      -0.123      -0.082
total_laasianhalf_share     -0.0586      0.014     -4.304      0.000      -0.085      -0.032
==============================================================================
Omnibus:                       30.409   Durbin-Watson:                   1.973
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               34.214
Skew:                          -0.289   Prob(JB):                     3.72e-08
Kurtosis:                       3.430   Cond. No.                     1.15e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.15e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Interpretation of the coefficients

This refined final model highlights four key factors influencing mortality rate in a log-log regression framework:

- Income (avg_median_family_income) – A 1% increase in average tract median family income is associated with a 0.43% decrease in a county's mortality rate, indicating a strong negative relationship.
- Population size (total_Pop2010) – A 1% increase in population size corresponds to a 0.10% decrease in a county's mortality rate, suggesting possible benefits from urban infrastructure or healthcare access.
- Low vehicle access (total_lahunvhalf_share) – A 1% increase in the share of individuals with limited supermarket access leads to a 0.07% increase in a county's mortality rate, highlighting the importance of food accessibility in public health.
- Asian low access share (total_laasianhalf_share) – A 1% increase in the share of Asian individuals in low-access areas is associated with a 0.06% decrease in mortality rate. This suggests that other factors, such as socioeconomic status or community resources, may mitigate the negative effects of low food access for this demographic.

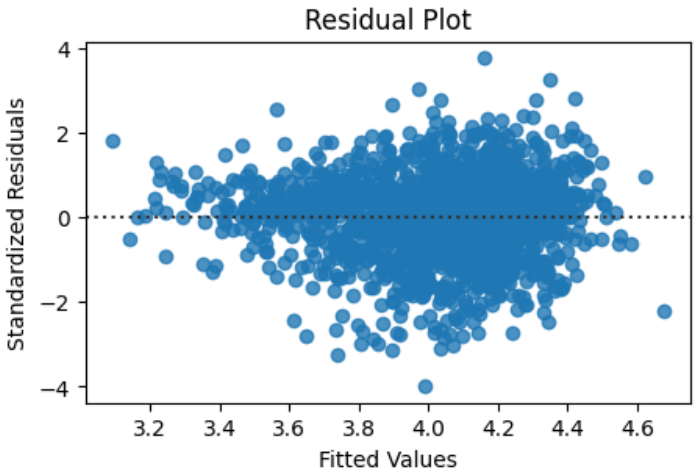
Since the model uses log-transformed variables, the coefficients can be interpreted as elasticities—showing percentage changes in mortality rate in response to percentage changes in the predictors.

Evaluate Model

The final model's training and test RMSE values are very similar, indicating that the model generalizes well to unseen data and is not overfitting. This suggests that the model captures meaningful patterns without being overly complex or tailored to the training data.

```
Training RMSE: 0.32852418304741166
Test RMSE: 0.3364711117603792
```

A residual plot of the final model's standardized residuals shows no systematic patterns. This suggests that the model captures the main linear relationships well and does not exhibit clear signs of nonlinearity. This reduces concerns about omitted variable bias.



Based on the Variance Inflation Factor (VIF) calculations, the level of multicollinearity in the model is within acceptable limits for our analysis. None of the independent variables have a VIF value exceeding 5, indicating that multicollinearity is not a significant concern.

```
VIF: avg_median_family_income:  2.176
VIF: total_lahunvhalf_share:  1.504
VIF: total_Pop2010:  2.108
VIF: total_laasianhalf_share:  2.451
```

Model Assumptions

Now that we have an optimal model, it is essential that we check the 4 model assumptions. The 4 assumptions that we will test are as follows:

1. Normality assumption: The residuals should be normally distributed.
2. Linearity assumption: The relationship between the independent variables and the response is linear.
3. Identical assumption: The residuals have constant variance.
4. Independence assumption: The residuals are independent from each other.

If any of these assumptions are violated, this could potentially mean that our model is biased/inaccurate, coefficients to be uninterpretable, and overall leading to poor performance. We performed the following visual verifications to ensure that all of the assumptions are satisfied.

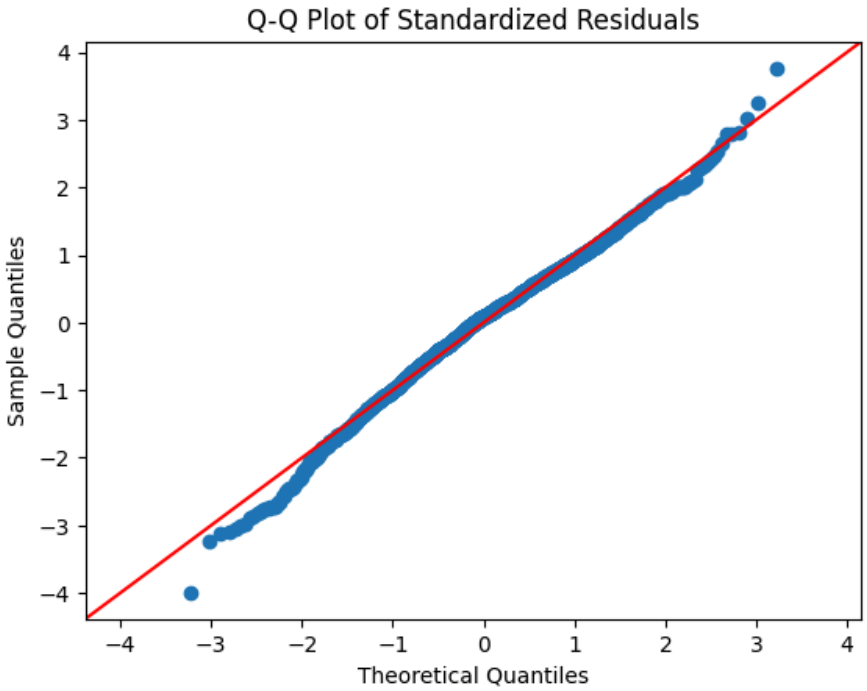
Normality Assumption

For the normality assumption, we can display a Q-Q plot and analyze how well it fits the reference line. We perform the test here using the standardized residuals.

As shown in the Q-Q plot, the residuals follow the 45 degree reference line very closely. Because the plot essentially compares the theoretical values (expected values) and the observed samples values, the plot clearly shows that the normality assumption is satisfied.

So, the **normality assumption is satisfied**.

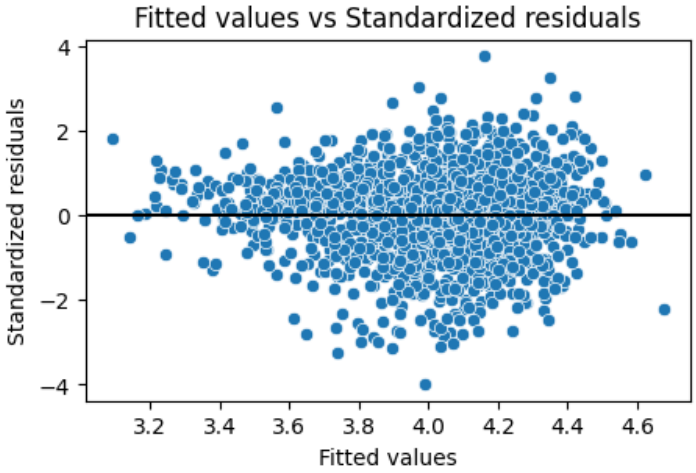
```
<Figure size 100x100 with 0 Axes>
```



Linearity and Identical Assumptions

For both the linearity assumption and the identical assumption, we can use the plot of predicted values vs. the standardized residuals. As shown, the data points are randomly scattered about the x-axis, with no distinctive pattern or slope.

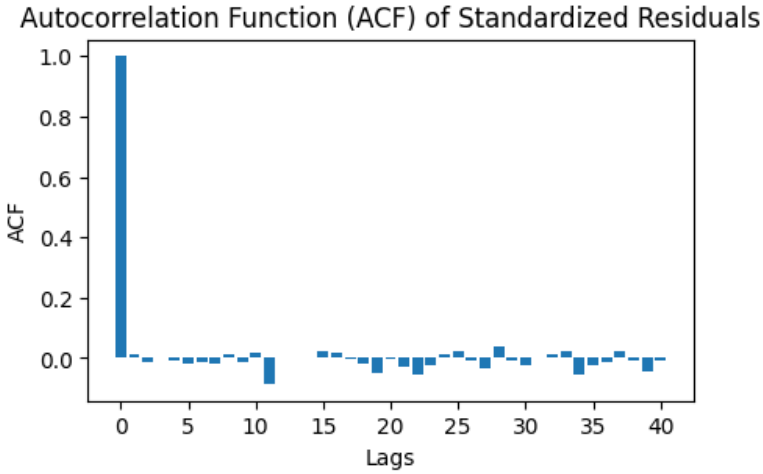
This proves that the **linearity and identical assumptions are satisfied**.



Independence Assumption

Finally, we move to the independence assumption which can be proved using the autocorrelation function. As shown in the autocorrelation function graph of the standardized residuals, there does not seem to be a noticeable pattern of the function from one lag to the next.

So, we can claim that the **independence assumption is satisfied**.



Potential Limitations and Shortcomings of Analysis

One limitation from our hypothesis test is because the mortality rate data did not pass the assumption of normality, we could not use parametric tests to model mortality rate using a known distribution such as the normal distribution. This limited the analysis we could've performed for the mortality rates of low access versus high access counties such as predicting confidence levels for the mortality rate in each category of the county. With our Kolmogorov-Smirnov test, the result only limited our analysis on determining if the distribution of mortality rate of low access versus high access are part of the same distribution. Because of this, nothing more can be said about the distributions in terms of how different the distributions are because we are not aware of the underlying distribution the mortality rate follows.

One of the main limitations of our analysis was the availability of the data that we were looking for. While many of the features in our food access dataset were interesting, many of them were very similiar and in some cases, highly correlated with each other, limiting our use of them to avoid multicollinearity. Even so, with our optimal regression model, there may be some correlation between variables, which may have distorted some relationships.

There are certainly other variables that affect mortality rate that we were interested in analyzing, like a binary variable on whether the county was rural or not, educational level, food affordability, and much more. These features could have given us a better picture of the current landscape of food accessibility within the US, and further insight into how we can address the prevalence of metabolic and nutritional diseases.

As with many datasets, the limited timeframe of the data may limit some of the external validity that our results can show. The data from our datasets were collected relatively recently, with the food access data from 2019 and the mortality data from the range of 2018-2023. However, more recent data can always help the validity of our results. Particularly with public health studies, data from before and after the COVID-19 pandemic can be drastically different so having up-to-date data would help with providing accurate and relevant results.

Conclusion

From our hypothesis test, we ran a Kolmogorov-Smirnov test on the mortality rates of low access counties vs high access counties. The data failed the assumptions for parametric tests, which prompted our use of non-parametric tests such as the Kolmogorov-Smirnov test and the permutation test. The test resulted in a p-value of 7.224501751217153e-18 < 0.05, rejeceting the null hypothesis in support for the alternative: the distribution of mortality rate of low access counties is different from the distribution of mortality rate of high access counties. The empirical CDFs for both the mortality rates in low access versus high access counties showed different distributions, which aligned with the result of our KS test. We also ran a permutation test to measure the difference in means of the mortality rate between low access and high access counties using random shuffling to sample mortality rates. With a p-value of 0, we reject the null hypothesis that the difference between the means is 0 in favor for the alternative. Both of the non-parametric hypothesis tests indicated that mortality rate is different in low access counties from high access counties.

From our regression model, we can observe that there are a few paramount variables that explain the mortality rate due to metabolic and nutritional diseases in a given county. From our model we can see that the feature with the largest effect size (or coefficient) is 'avg_median_family_income'. As one might be able to expect, as the median income of a county goes up, it has a negative affect on the response variable 'mortality_rate', bringing it down.

As mentioned in our introduction, nutritional and metabolic diseases are one of the most prevalent, yet preventable diseases that affect the communities around us. These diseases are often contributed through factors like diet and socioeconomic conditions. As shown from our regression model, counties with higher income tend to lower its respective mortality rates to these preventable diseases. Counties with

higher median income are typically associated with better-funded infrastructure that can support the wellbeing and livelihood of its residents, like access to supermarkets with fresh, healthy food.

Another significant feature was 'total_lahunvhalf_share', which had a positive effect on 'mortality_rate'. This feature explained the total share of individuals in the county that did not have a vehicle and were more than a half-mile away from a supermarket. Consequently, these residents may result to relying on closer, more accessible food sources like convenience stores and fast food outlets to satisfy their diet.

What does this mean from a public health perspective? It implies that there needs to be an increase of funding into the infrastructure of lower-income communities with limited access to fresh, healthy food. From building a comprehensive public transportation system to increasing educational awareness on the benefits of healthy food, there are many policy changes that can be made to address the high mortality rate due to these nutritional and metabolic diseases. With enough advocacy, we can hope to see a future where healthy food is easily accessible to all and a declining mortality rate due to these preventable diseases.