THE IMPACT OF WALKABILITY AND PUBLIC TRANSIT ON HEALTH

An analysis of the health of citizens in areas with high walkability and accessible public transit.

WHY?

Through our team discussions, we came to a realization. We had all experienced how difficult it can be to access necessities when transportation was limited.

Our goal was to explore how limited walkability or restricted access to food could impact a person's health.

Our Hypotheses

NULL HYPOTHESIS

Census Tracts with a **higher walkability score** and i**ncreased access to public transi**t will have a **lower** rate of obesity and a **longer** life expectancy.

ALTERNATIVE HYPOTHESIS

Census Tracts with a **higher walkability score** and i**ncreased access to public transit** will have a **higher** rate of obesity and a **shorter** life expectancy.

Data Sources

Census Data

We used the American Community Survey data collected around the 2010 census to gather the census tract IDs and demographic information for each census tract in the United States.

USDA Food Access Data

This data set from the USDA gives each census tract a score based on how far citizens have to travel to access food. The census tracts are identified as "Low Income, Low Access" areas, otherwise known as Food Deserts.

CDC Life Expectancy Data

This data set provides the average life expectancy of citizens in a given census tract.

Institute for Health Metrics and Evaluation Data

This data set provides health information for citizens of each census tract. We used this to data to gather the rate of obesity in each census tract.

Narrowing the Data

Health Data	City Data	Questions / Solution
There was so much health data to explore. Heart disease, obesity, cancer, mental illness, and more.	The census data set was huge; it included almost every demographic data point you could imagine.	• What data most directly impacts our hypothesis? • What is our capacity as a team?
We determined that we should focus on heart disease, obesity, and life expectancy.	We determined that we should narrow the city data to income and transportation statistics.	• Is our data still too large? • Will a model be able to process this data accurately and clearly?
We determined that the heart disease data was too vast and diverse.	We decided to remove some of the income data that would require a lot of scaling and encoding.	• How do we maintain focus on our hypothesis? • This allowed us to narrow to our final data sets.

We finalized our data sets to measure obesity and life expectancy against transportation and food access data from each census tract.

Choosing a Machine Learning Model

We chose to use the Random Forest Regressor model for this project. This was perfect for our needs because it could predict continuous values without having to use PCA to cut down on the number of features.

This would allow us to use much more of the three features in our model as well as let us calculate the feature importance later.

We needed the feature importance fata in order to determine which aspects of infrastructure affected our targets the most.

Random Forest reduces overfitting and doesn't require feature normalization which helped us feel more confident about the accuracy of our model.

The Modeling Process

Clean the data in Jupyter Notebook.

Load the data into SQL. Join data in SQL.

Connect the model to the SQL database.

Drop the columns describing location.

Drop the target columns (obesity, life expectancy).

Use sklearn train _ test _ split to determine targets and features.
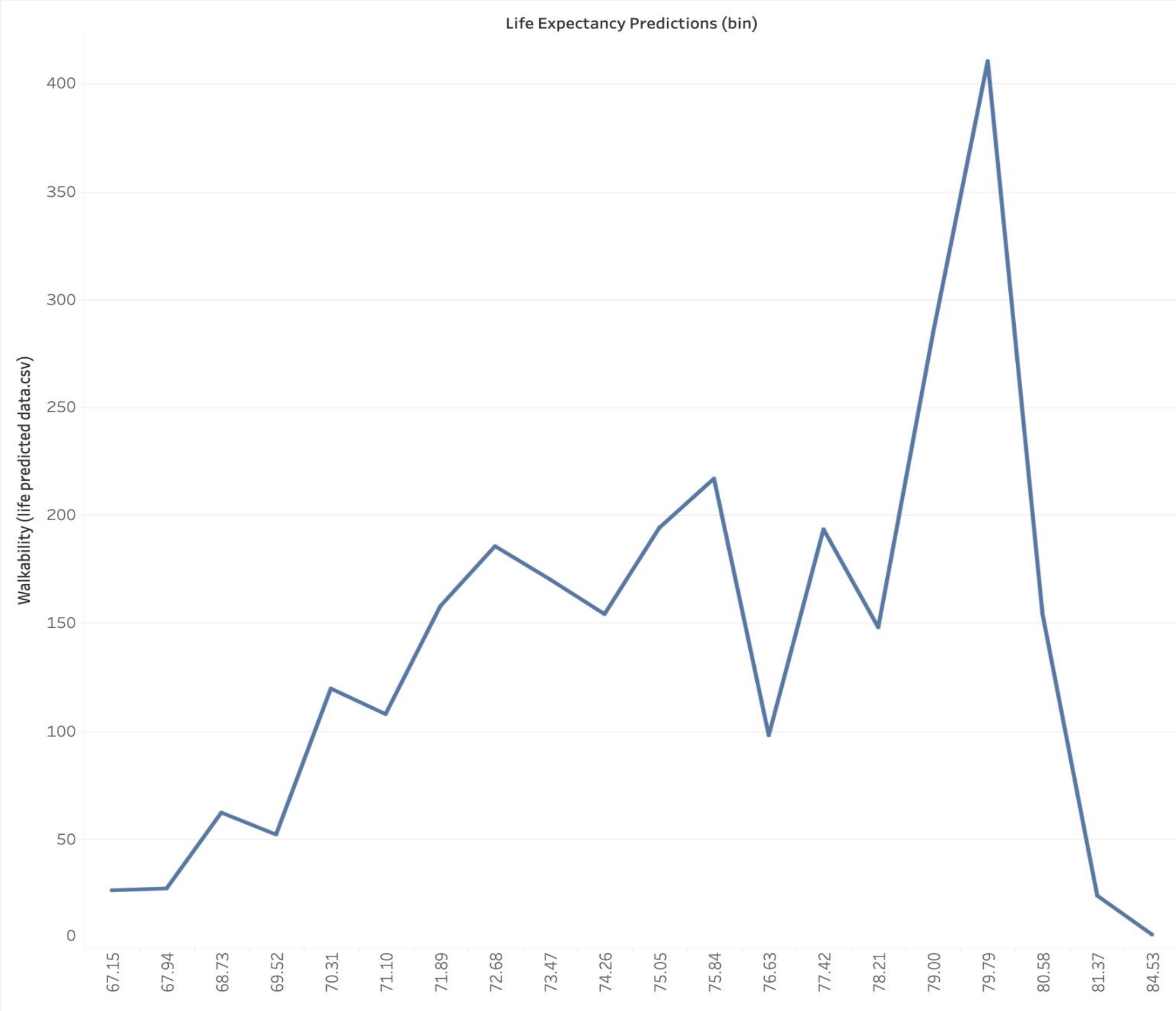
Train the model and run a preliminary test.

Drop the % Low Income column to balance results.

Run final Random Forest Regressor models.

Feature Importance
Life Expectancy

This visualization illustrates describes the most important city features in determining life expectancy of an area's citizens.

Life Expectancy Predictions (bin)

Feature
Importance
Obesity

This visualization illustrates describes the most important city features in determining the rate of obesity in a particular area.
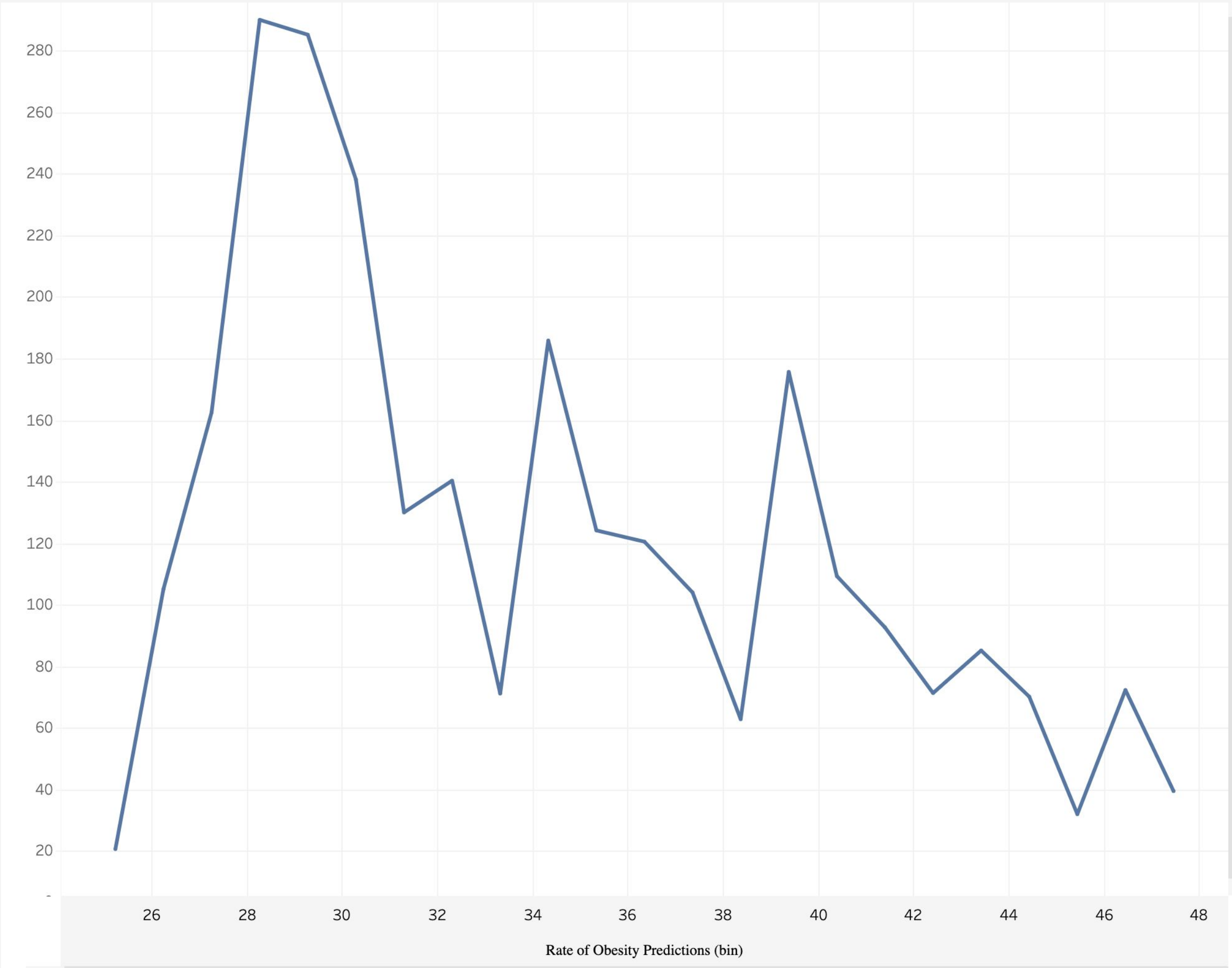
Tableau Dashboard

Obesity
Life Expectancy
Median Income by County
Poverty Rate by County
Population by County

CONCLUSION

Our Null Hypotheses was: Census Tracts with a **higher walkability score** and i**ncreased access to public transi**t will have a **lower** rate of obesity and a **longer** life expectancy.

In the end we did see a correlation between higher walkability scores and increased access to public transit to lower rate of obesity and a longe life expectancy.

Challenges

Github
Github collaboration can be complicated. Files need to be clean and organized on the local computer.

Data Cleaning
Data Cleaning may involve a lot of extra research. Even though it seems solitary, it may require teamwork!

Database Administration
Using SQL meant that our database administrator needed to run everything on the local computer.

Thank you for your interest in our project. Please check out our github for more detailed information about our process.

Commit to the Git

Rowena Quinn

Chris Swan

Kyle Kramer

Savannah McIntyre

Charles Bolton

Citations

Each resource is linked. Please feel free to click the links and explore our resources.

Data Sources

INFRASTRUCTURE DATA

FOOD ACCESS DATA

LIFE EXPECTANCY DATA

OBESITY DATA

CENSUS TRACT DATA

Coding Help and Research

Python: How To Convert SQL To DataFrame In Pandas

Approving a pull request with required reviews