All Food in the Neighborhood

Group Member Names & CNetIDS
Takayuki Nitta (tnitta)
Ryoya Hashimoto (rhashimoto)
Vincent Liu (jliu48)
Liz Colavita (ecolavita)

Project Overview
Our project aims to explore the relationship between food security and crime level throughout 72 Chicago neighborhoods. Using data primarily scraped from the Chicago Data Portal and Chicago Health Atlas, we collected the data using APIs. Chicago Data Portal contains neighborhood-level socioeconomic data, such as income, poverty, and crime rates. Chicago Health Atlas provides us information about several food and health related indicators, such as food access, access to fruits/vegetables, and soda consumption.

There is an interesting relationship between food access, and related variables, and crime rates, and we investigated this relationship from various perspectives. The correlation matrix heatmap in the dashboard most succinctly captured these insights, where we see the most highly correlated variable to crime rate is poverty rate, followed by soda consumption and fruit and vegetable servings - which, notably, is negatively correlated with crime rate. A decision tree classification model using the attributes resulted in a correct prediction of a neighborhood's crime rate about 55% of the time. Regression analysis reinforced findings from the correlation heatmap and concluded that crime rate has the strongest correlation with poverty rate, seconded by soda consumption rate. Through criterias of R^2, AIC, and BIC, the model with only poverty rate and soda consumption succeeded in explaining the crime rates. Lastly, our analysis is further supported by additional visualizations on the Dash interface and the visualization report notebook file.

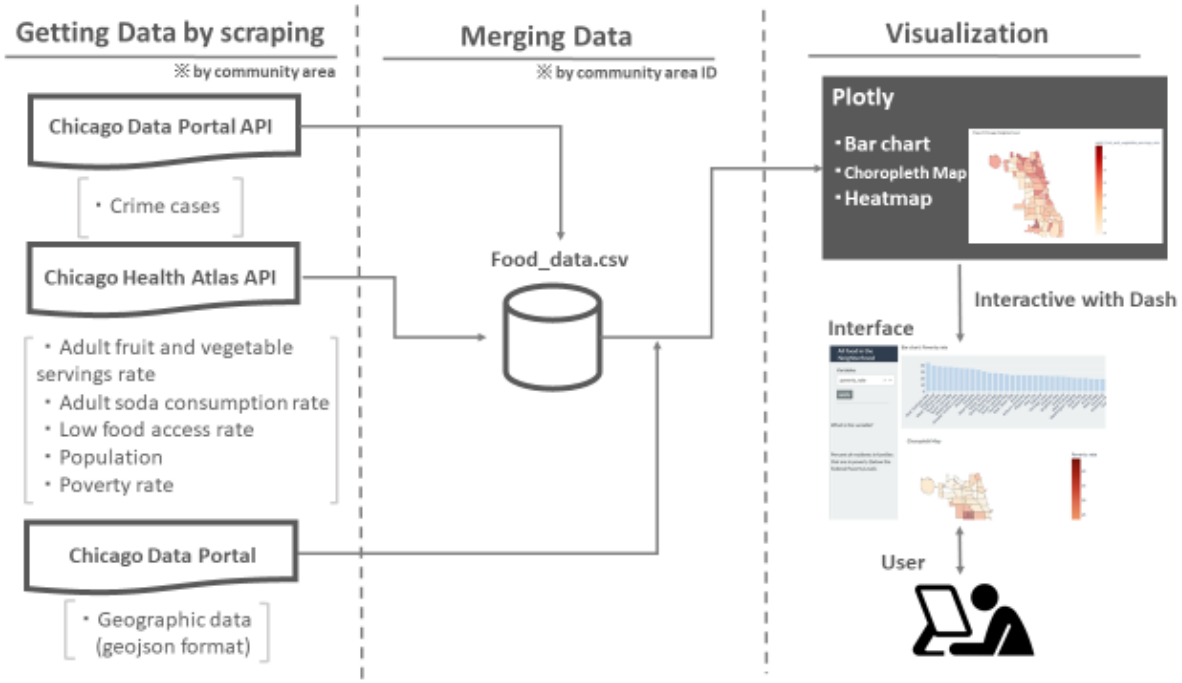
Software Structure
The overall structure of the software can be broken down into 3 main components: data collection, data processing, and data analysis and visualizations. The data collection pieces are primarily included in files "get_data_from_atlas.py", "get_data_from_portal.py", "chicago_community_areas.geojson". In the first two, we accessed the data on Chicago Health Atlas and Chicago Data Portal through web scraping using an API. The last one, "Chicago_community_areas.geojson", can be downloaded from Chicago Health Atlas website. The geojson file is an intermediate step in getting the geographic data into a useful format for building map-related visualizations.

A lot of the data processing involved extracting relevant variables, blending different variations of the same variable from each data source, and reconciling differences in time frames of different variable measurements. The code for this part is also included in the abovementioned

files. For example, Chicago Health Atlas contained a variety of variables to choose from so that we needed to pull out the variables we wanted to use in our analysis. Much of the processing was related to crime rate and population measurements, which involved generating the crime rate from the absolute number of crimes and reconciling that crime and population measurements were from different years/year ranges. Finally, data were broken down differently by geographic measure, such as zip codes or other identifications, such as “community_id” or “geo_id”. We had to keep this in mind when building out the csv files to make sure they all utilized “community_id”/”geo_id” and not zip code, because “community_id”/”geo_id” was the commonality between the two data sources. This work resulted in several csv files, and most notably “food_data.csv”, “poverty_and_crime.csv”, and “total_crime.csv” were used in our analyses.

The csv files were then used for visualizations and analysis. “Manage.py” creates a visualization dashboard using dash, and this dashboard includes a bar chart (interactive), choropleth map (interactive), and a correlation matrix (interactive).

Figure 1. Diagram of how our dashboard is developed



Regression analysis is primarily used for statistical inference, that is, finding out correlations between an outcome variable and predictor variable(s). Regression models explain both the strength (p value) and direction (coefficient) of statistical relationships. In “regression.ipynb”, three regression models check if demographic and food-related variables in “food_data.csv” affected the crime rates in these 72 neighborhoods and select the best model in inferring and predicting the crime rates.

“Decision_tree.py” is another way of relating our explanatory variables (attributes) in food_data.csv – for example low food access, fruit and vegetables serving rate, poverty rate, and others – to crime rate by applying a decision tree classification model to the data. This first required categorizing the attributes and target class into categorical variables into the following five categories: very low, low, medium, high, very high.

Results of both analyses were further examined using visualizations on the Dash interface and “visualization_report.ipynb”. These graphs shed light on which community had the highest level of crime rate, poverty rate, food accessibility, and other indicators as well as longitudinal trend of crime rates of crime rates within the last 5 years. Through the uses of heatmaps, two-year bar charts, and two-way scatter plots, relationships between variables were visualized in different manners and from diverse angles.

Code Responsibilities

Vincent Liu

- Visualization Report - visualization_report.ipynb
- Regression analysis - regression.ipynb
- Project deliverables

Takayuki Nitta

- Dash dashboard - manage.py
- Concatenating data based on community id - get_data_from_atlas.py
- Chicago Data Atlas API - get_data_from_atlas.py
- Building geojson file for mapping neighborhoods - chicago_community_areas.geojson

Ryoya Hashimoto

- Dash dashboard - manage.py
- Concatenating data based on community id - get_data_from_atlas.py
- Chicago Data Atlas API - get_data_from_atlas.py
- Chicago Data Portal API - get_data_from_portal.py

Liz Colavita

- Virtual Environment - install.sh, requirements.txt
- Decision Tree classification model - decision_tree.py
- README.md file
- Project deliverables

Application Interaction

To run our application, a user can run the following from the command line within the app directory:

1. `bash install.sh`
2. `source env/bin/activate`
3. `python3 decision_tree.py & python3 manage.py`
4. Navigate to the address from the previous step, e.g. http://127.0.0.1:8500/, in a web browser.
5. `jupyter notebook regression.ipynb` and follow one of the links starting with http://localhost: (e.g. http://localhost:8888/?token=59fa90841a008fbc90400d4ebdca537ae241d9ced4f6f0cf) or http://127.0.0.1: (e.g. http://127.0.0.1:8888/?token=59fa90841a008fbc90400d4ebdca537ae241d9ced4f6f0cf) in a web browser and select "regression.ipynb" (Note: statsmodels package may need to be installed locally; for some group members this package was causing problems in loading the notebook)
6. (Optional): `jupyter notebook visualization_report.ipynb` and follow one of the links starting with http://localhost: or http://127.0.0.1: in a web browser and select "visualization_report.ipynb" (Note: You may need to run the notebook file line by line in order to see the visualizations because of how Plotly works)

The result from running decision_tree.py is a dictionary where the keys are the following strings: ‘all variables’, 'adult_fruit_and_vegetable_servings_rate', 'adult_soda_consumption_rate', 'low_food_access', 'poverty_rate', and 'population'. The value associated with ‘all variables’ is the rate at which a model using all attribute variables was successful in predicting the crime rate. That is, how well did a model with all attributes in food_data.csv as predictors do at predicting the target class? A model with all variables predicted the crime rate correctly about 55% of the time. The remaining keys are associated with values that show how well a model did at predicting the crime rate excluding that variable from building the model. For example, the “poverty_rate” key is associated with a value of about 41%, meaning a model built without the poverty rate attribute predicted the crime rate correctly 41% of the time. Excluding poverty rate changed the success of the model the most, which matches the correlation matrix heatmap in the interactive dashboard and regression analysis that says that poverty rate is most highly correlated with crime rate.

The next result from running the command above is an interactive dashboard interface that opens in a browser. The dashboard includes three visualizations: a bar chart (interactive), choropleth map (interactive), and correlation matrix heatmap. The bar chart shows the values of a particular variable (y-axis) for each Chicago neighborhood (x-axis), and the choropleth map uses color intensity and a geographic representation of Chicago to show the same information. The user can choose which variable - adult fruit and vegetable servings rate, adult soda consumption rate, low food access rate, poverty rate, crime rate, and population - they would like to see represented in the bar chart and the choropleth map using the drop-down menu on the left sidebar of the dashboard. Also on the left sidebar, below the drop-down menu, is a full description of the variable selected by the user. The correlation matrix heatmap is a static visual

but is a clear concise way of presenting the correlations between each of the variables with one another. The most highly correlated to crime rate is poverty rate with a positive correlation of 0.79. This is followed by adult soda consumption and adult fruit and vegetables servings rate with correlations of 0.45 and -0.3 respectively. Interestingly, low food access has the lowest correlation (in magnitude) with crime rate, perhaps because we are including other related covariates. Though low food access has the lowest correlation it is still valuable to isolate the effect of low food access on crime rates.

With the command on step 5, a jupyter notebook file named regression.ipynb will be opened in the browser. In the jupyter notebook file, three models were constructed based on "food_data.csv" with the goal of finding the best model that balanced interpretability and performance. In the first model, all demographic and food related variables were used as predictors of crime rate. This model indicates that poverty rate was the only statistically significant effect on crime rates. This is consistent with what we concluded from the decision tree algorithm and the correlation matrix heatmap. The second model removed population, poverty rate, and other socioeconomic indicators as explanatory variables in order to find out the power of food security. The model showed that the adult soda consumption rate was the only significant predictor, though now as significant as poverty rate in the previous model. In the last model, only poverty rate and soda consumption were used as regressors, and this model rejected soda consumption as a significant indicator. Three models were compared using four criterias: $R^2, R^2_{adj}, AIC, and\ BIC$. The rule is that for the first two, the larger the value the better the model (as they indicate the amount of outcome explained by predictors), whereas, for the last two, the lower the better. The last model with only poverty rate and soda consumption had the largest $R^2\ and\ R^2_{adj}$ and the lowest AIC and BIC scores. Therefore, poverty rate and sofa consumption should be treated collectively as predictors of crime rates.

Lastly, "visualization_report.ipynb" offers additional interactive visualizations using Plotly. These visualizations were made utilizing three data sources: "food_data.csv", "poverty_and_crime.csv", "total_crime.csv". This selection was intentional in order to present information from data to the fullest extent. Three types of graphs were made and with disparate purposes. The first section of the report contained two one-way bar charts and two two-way bar charts that ranked these 72 communities by the number of assault and homicide, crime rates by poverty rates (poverty rate is indicated by the thickness and lightness of the color), level of low food access, and food accessibility by adult vegetable serving rate (indicated by the color). These graphs strengthened our analytical findings that poverty rate is the single most important indicator of crimes but also enabled readers to know which neighborhood had the highest crime and food inaccessibility levels. The second portion of the report presented the trend of crime level changes in the top6 hyper-criminal activity neighborhoods of Chicago. The line chart marked 2015 and 2018 as two breaking points of the trend and left spaces for researchers to dig into "why". It found that Austin, Near North Side, and Loop had the largest fluctuations across these 5 years, whereas the remaining three were more stable in terms of criminal activities. The last chunk of the report shed light on crime rate, poverty rate, educational level, and food accessibility are related through scatter plots. The graphs reinforced findings from the decision tree algorithm and regression analysis.

Project's Intended and Actual Accomplishments

The project is a successful exploratory look at the relationship between food access and crime rates at the neighborhood level in Chicago. The data collection and data processing components show success in using an API to gather web data and transform it into more useful data forms for analysis, combining variables from both Chicago Data Portal and Chicago Health Atlas. The subsequent analysis and visualizations are a good first-stage exploratory analysis and would show their values in providing a foundation for further policy analysis and implications.

Given more time, we would have liked to incorporate more data sources, for example grocery store location data and other relevant measurements surrounding food insecurity. More data likely would have allowed us to do additional interesting data analysis. Also, the intent was to have all the analysis appear on the dashboard, but integrating multiple visualizations into one dashboard ended up being a more complicated task than we envisioned. However, given our limited prior experience with dash and gathering data from the web, we are overall satisfied with how the project turned out.