Exploring Chicago Neighborhood Venues and Predicting Economic Hardship

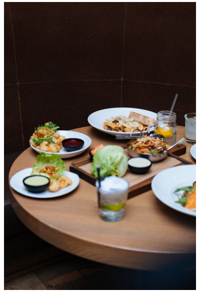
By: Eduardo Murillo

What is Economic Hardship?

- To study the economic difficulties that different communities across Chicago face, the University of Illinois at Chicago (UIC) developed the "**hardship index**".
- The economic hardship score is an average of six standardized variables: *Unemployment*, *Education*, *per capita income level*, *poverty*, *crowded housing*, *dependency*
- It is a score from **0 - 100**, 100 being the most economic hardship.

The Study - Predicting Hardship using Venue Types

- We will explore how the types of **venues** in a given neighborhood may affect the **hardship score**.
- The goal is for Chicago city officials and developers to be able to determine or predict how different venues may affect a neighborhoods' economy.
- Using this information, developers may also more effectively decide which venues should be constructed or demolished to benefit the community.



www.shutterstock.com · 439390204

Data Acquisition

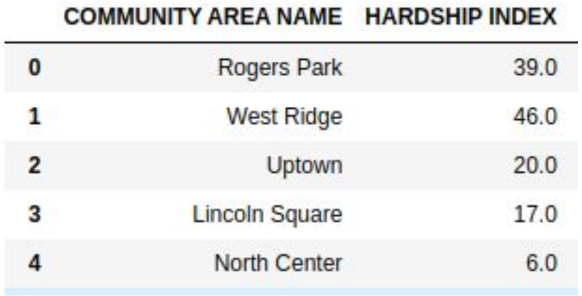
- Hardship index scores for all 77 neighborhoods in Chicago, from the City of Chicago's database: https://data.cityofchicago.org/Health-Human-Services/Hardship-Index/5kdt-irec

- Geographical coordinate data for the neighborhoods in order to determine the nearby venues: *Geocoder*

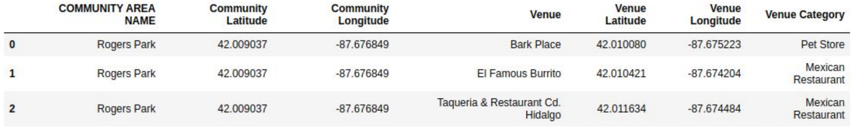
- A dataset of the nearby venues for each neighborhood: *Foursquare API*.
  - Will be used to try and predict a given neighborhoods' hardship level

Data Samples and Data Cleaning

- Removed any non-neighborhoods ("CHICAGO")
- Corrected some community area names
- 0 nearby venues found for **Riverdale**
- Economic Data - Only *Hardship Index* studied



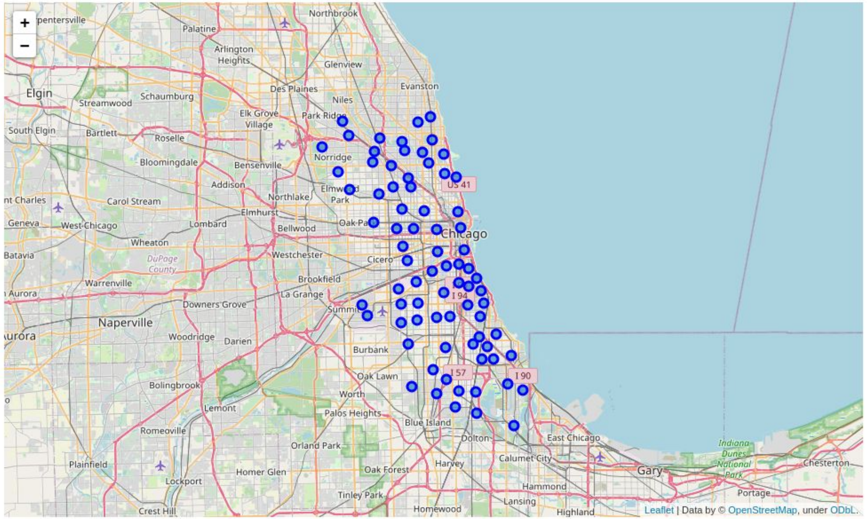
	COMMUNITY AREA NAME	HARDSHIP INDEX
0	Rogers Park	39.0
1	West Ridge	46.0
2	Uptown	20.0
3	Lincoln Square	17.0
4	North Center	6.0



	COMMUNITY AREA NAME	Community Latitude	Community Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rogers Park	42.009037	-87.676849	Bark Place	42.010080	-87.675223	Pet Store
1	Rogers Park	42.009037	-87.676849	El Famous Burrito	42.010421	-87.674204	Mexican Restaurant
2	Rogers Park	42.009037	-87.676849	Taqueria & Restaurant Cd. Hidalgo	42.011634	-87.674484	Mexican Restaurant

Geolocation Data - Mapping

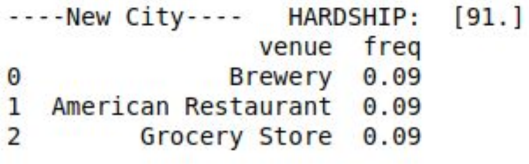
Using **folium** together with our latitude and longitude coordinates, we may create a **Map of Chicago Neighborhoods:**



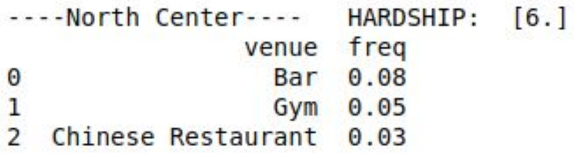
Model Development

- Data Preprocessing:
  - **One Hot Encoding** to convert categorical dependant variables to binary integer values
  - 278 dependant variables (venue categories)
  - Then convert to continuous: Group rows by neighborhood and take the mean of the frequency of occurrence of each category

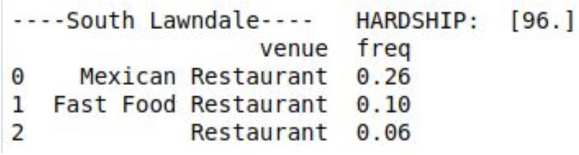
- **Top Venues** per Community:



```
----New City----    HARDSHIP:  [91.]
                 venue  freq
0              Brewery  0.09
1  American Restaurant  0.09
2        Grocery Store  0.09
```



```
----North Center----    HARDSHIP:  [6.]
                venue  freq
0                 Bar  0.08
1                 Gym  0.05
2  Chinese Restaurant  0.03
```



```
----South Lawndale----    HARDSHIP:  [96.]
                 venue  freq
0   Mexican Restaurant  0.26
1  Fast Food Restaurant  0.10
2           Restaurant  0.06
```

Model Development - Multi Linear Regression

- The Model: **Multi Linear Regression Model** (MLR)
  - **Dependant variables:** Venue Categories   Target/**Dependant Variable:** Hardship Index
  - Theory - an aggregate of a community's different features would most affect its wellbeing. That is, one ATM or even all the ATMs combined would have a very significant impact on a neighborhood's hardship index.
  - A linear relationship is usually a good place to start exploration.
  - Will use **linear_model** from **scikit-learn** package
- Model Training -
  - Trained via data sample obtained using **train_test_split** module from scikit-learn

Data
Training
Evaluation

ML Model Training Workflow

Model Prediction and Evaluation

- Predictions for hardship index made for both test dataset and whole dataset
- Evaluated results using the model's **score()** function and **residual sum of squares**

- **TEST Data Evaluation:**
  - Residual sum of squares: 1395.55
  - Variance score: -0.67
  - **Not very good**, and the variance is negative, which is not expected.
- **WHOLE Dataset Evaluation:**
  - Residual sum of squares: 362.48
  - Variance score: 0.55
  - Score is a bit better, and a more typical positive variance score is observed. However it is still not very accurate, test data is best suited to evaluate accuracy.

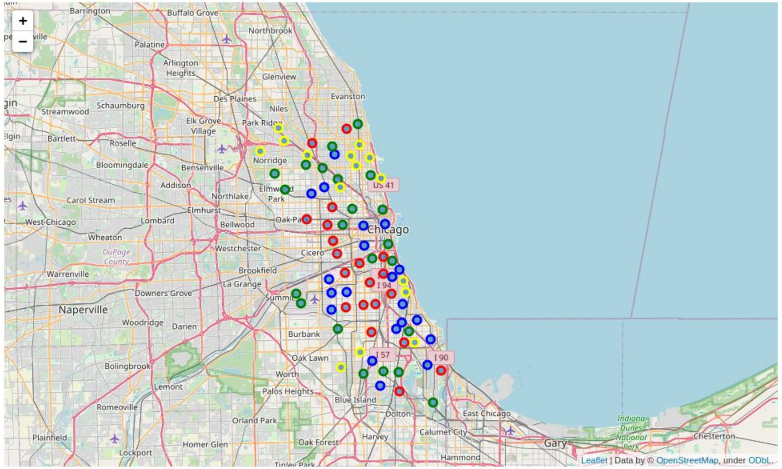
Key:

- Yellow: 0-25
- Green: 25-50
- Blue: 50-75
- Red: 75-100

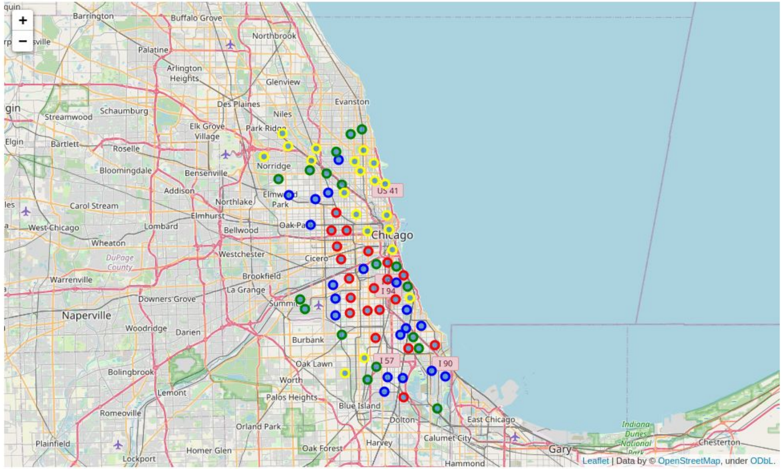
Visualization of Model Predictions

- **Map Predicted v.s. Actual Hardship Index Scores:**

Predicted



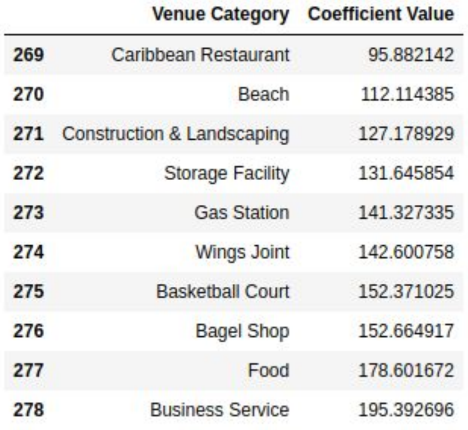
Actual



Note that we are using value ranges for the marker colors and using predictions for the whole dataset

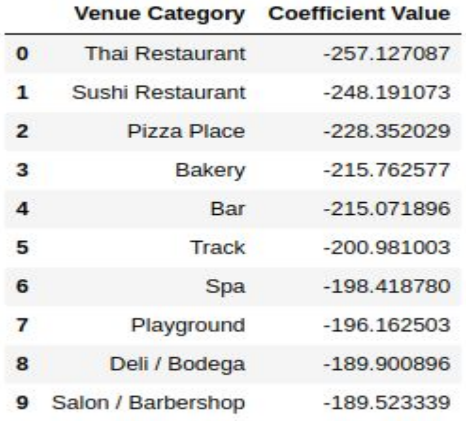
Further Data Analysis

Model Coefficient Analysis - To determine which venue types may most affect hardship, the model coefficients were ordered by magnitude:



	Venue Category	Coefficient Value
269	Caribbean Restaurant	95.882142
270	Beach	112.114385
271	Construction & Landscaping	127.178929
272	Storage Facility	131.645854
273	Gas Station	141.327335
274	Wings Joint	142.600758
275	Basketball Court	152.371025
276	Bagel Shop	152.664917
277	Food	178.601672
278	Business Service	195.392696

Large + Value -> Increases Hardship

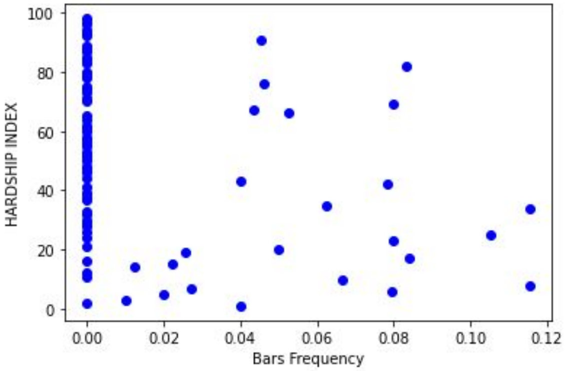


	Venue Category	Coefficient Value
0	Thai Restaurant	-257.127087
1	Sushi Restaurant	-248.191073
2	Pizza Place	-228.352029
3	Bakery	-215.762577
4	Bar	-215.071896
5	Track	-200.981003
6	Spa	-198.418780
7	Playground	-196.162503
8	Deli / Bodega	-189.900896
9	Salon / Barbershop	-189.523339

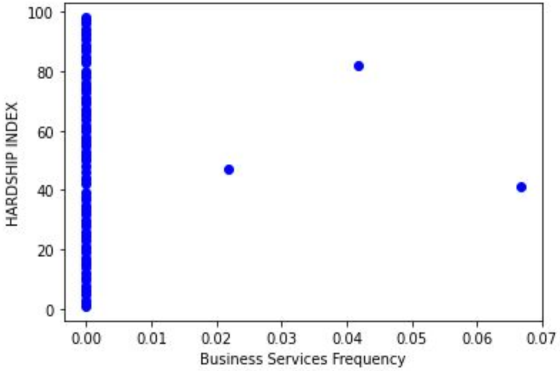
Large - Value -> Decreases Hardship

Further Data Analysis

Using our coefficient analysis, can we find other relationships between ‘significant’ dependant variables and our target variable?



Appears to be a **slight clustering** of points up to 0.04 frequency, however, there is **not a clear relationship**. Most points lie around 0



Here, there is definitely no clear relationship, and again, most of the x values are 0.

Conclusion and Further Directions

- The goal of this study was to explore how the types of venues in a given Chicago neighborhood may affect its hardship score.
- This would help city officials and land developers determine the types of venues which may be beneficial or detrimental to a community, and how large that effect may be.
- We began our study of this question by collecting data on the hardship scores of the 77 communities in Chicago. Then we also collected geolocation and venue data on each of the categories.
- After initial exploration of our dataset, we then began our model development. We chose to primarily explore a multi linear regression model because of our assumptions on the dynamics of a community.
- We also explored some of the venue categories individually to check for further predictions.

Conclusion and Further Directions (Continued)

- There was definitely a **wide variety of venue** categories across the Chicago neighborhoods. A lot of them were food related.
- Interestingly enough, the community with the highest hardship index had 0 nearby venues found. This observation would be interesting to explore further using a larger dataset or different variables.
- From our MLR model itself, what we determined is that our **data likely does not have a multi linear relationship**, thus our model would not be the best to use to predict hardship scores or make recommendations on the type of new venue a community should have.
- Perhaps one insight Chicago officials and developers may gain from this study is that there is **a lot more factors that go into a community's economic wellbeing**. Just changing the types of venues in a community may not have a significant impact on its economy. However one must keep in mind the scope of this study.
- Only the relationship between venue categories and hardship was studied and only on Chicago neighborhoods using Foursquare. It would be useful to conduct **future studies** where **larger datasets** were used or **different variables** were examined.