**ETL Project**

Group Project #2

**The Group:**

Zach Johnson

Mary Kilmer

Paul Cousin

Ryan Johnson

Matt Garber

**Extract:** What are the original data sources and how are they formatted?

* Population by zip csv file - <https://blog.splitwise.com/2013/09/18/the-2010-us-census-population-by-zip-code-totally-free/>
* Chicago Sidewalk Cafe Permits csv file - <https://www.kaggle.com/chicago/chicago-sidewalk-cafe-permits>
* Crime by ward json file - <https://data.cityofchicago.org/resource/3uz7-d32j.json>

Final database, tables and collections:  
  
Key Table: Primary Key – Wards  
 Foreign Key – Zip Codes  
  
 Wards was chosen as the primary key in a one to many scenario.  
  
Tables: café\_permit, Pop\_by\_zip, Crime\_ward

**Transform:** What cleaning and transformation was required?

After the three data sources were selected, we used SQLalchemy to import a panda dataframe to a mySQL database. The data files we used were in good clean condition. Most of the cleaning and transforming related to eliminating or deleting columns from the files to be used in the database. At that point we created three tables, one from each data source. In mySQL we created a new table from the crime\_ward table to find the number of robberies in each ward using the count function and named this table robbery\_ward. For joining the pop\_by\_zip and the cafe\_permit data, we inner joined on zip code. We then joined the robbery\_ward data with the pop\_by\_zip data using a inner join on zip code, and a left join outer join on ward. We put this data into a table called cafe\_data. This will all be found in the DB file chicagoData20190221.

The final table will have a list of permits for cafes, robberies, and the population of the ward where the business is established.

We will store the data as a relational MySQL database. This will allow us to combine population data, and crime by ward data from three separate sources, and query for each by ward and zip code.

**Load:** Show the final database and why this was chosen?

We have uploaded the final database, combining the three tables to house the following columns:

Permit number

Legal name

Dba

Address

City

State

Zip

Ward

Population of zip

Number of Robberies

Why was this chosen:  
  
Potential customers will use the database to find sidewalk cafes and where they are located. They can use this database to find what wards have a high number of robberies and also which zip codes have a high population within these wards.