Seth Kim

Trong Nguyen

**ETL Project – Yelp Delivery Service**

**Proposed Use of Data**

To select the best candidate businesses for a new food delivery service. The service, unlike existing ones, will ship frozen meals or ingredients along with instructions to all markets within the contiguous United States. To aid the selection process, we will need the following data points:

* Yelp rating of establishment
* Whether or not the establishment caters
* Median income for zip code of establishment
* Number of reviews for the establishment

**EXTRACT**

We used the following two original sources of data:

* Yelp dataset challenge, Business data. Format: JSON
* IRS median income by zip code. Format: CSV

We used the Python Pandas library to read both datasets into DataFrames.

**TRANSFORM**

We transformed our data using the Pandas library. The dataset was cleaned and trimmed to include only the pertinent data. For the Yelp businesses dataset, we selected the following fields:

* business\_id
* name
* address
* state
* zip\_code
* has\_catering
* stars
* review\_count

The original ‘attribute’ column contained dictionaries and was flattened to extract only the ‘caters’ key as a column. The pertinent columns were selected and renamed to match our MySQL database and the income table. Rows (businesses) with ‘None’ or ‘NaN’ values in the ‘has\_catering’ column were removed, leaving only ‘True’ or ‘False’ values. Furthermore, many rows had strange, meaningless alphanumeric values in the ‘zip\_code’ column, and were also removed using a regular expression interpretation.

For the IRS income data, we selected and renamed the following fields:

* Zip Code
* Income Level

**LOAD**

Our final location for the data is a MySQL database. We created our yelp\_delivery\_db with a UTF-8 character set, and created two tables – ‘businesses’ and ‘median\_income’ – with columns that exactly matched those in our DataFrames. Their primary keys were the ‘business\_id’ and ‘zip\_code’ columns, respectively. The tables could presumably be joined in the database on the ‘zip\_code’ columns. This final representation of the data would prove useful for new delivery service, allowing us to analyze various businesses for their attractiveness as potential clients.