Bryan Winmill
William Cherry

Final Report

We used the Farmers Markets Dataset to do our project. We wanted to be able to take the data that was provided, and clean it up so that it was a more accurate/cleaner dataset that could be used by the “client”. If clients want to be able to know who sells what types of food items, this is a very useful dataset. Another useful aspect of this dataset is the fact that you can see who accepts credit and who doesn’t. This would be helpful information if you were looking for somewhere to purchase stuff. In addition, the dataset provides a website, facebook page, twitter, youtube, and other media for each of the markets to obtain more information. It also includes the location and times that these markets are open depending on the season and what you are looking for.

If you were just looking to find where a market is located, or what items they provide, the data is already clean enough to get what you need. However, if you were hoping to use the x and y coordinates to map out the location, or even wanting to check when it was last updated, you would need to do some cleaning because the info is currently only stored as text and would be much easier to navigate if the x and y values were stored as numbers, and the date was stored as a date object. Also, another thing that would cause a lot of confusion is the fact that there are some markets that have multiple different naming conventions, even though they all refer to the same one. This could cause some confusion and would be much easier to use if it was consistent throughout the data.

Looking at the data, we realized that there were a lot of columns. As we tried to figure out exactly what was present, we tried to break it down as to what was there, and what relationships existed. We soon realized that a market could have multiple locations. As well as a single market having multiple locations, we also noticed that they could provide different products, as well as have different seasons (When they are open and when they aren’t). Once we realized that, it was a lot clearer as to what was included in this dataset. The data provided is a bunch of different companies/stores, each one having its own specific hours of operation for each season, a list of products that are provided at each location, as well as address/locations that these stores exist.

Looking at the data, there were many markets that had multiple spellings for the same place. This was one important part of the dataset that we thought should be taken care of first. So we created text facets and analyzed all the different spellings. We went with a consistent spelling throughout so that none of them would be spelled with all upper case, and also so that each market was only referred to by a single name (No more multiple spellings). After that we then looked at the x and y columns and figured those would be best as numbers rather than text. We then updated the last column so that it was stored as an actual date object instead of just text of a date. We then looked at the remaining text fields such as Website, Facebook, Twitter, Youtube, OtherMedia, street, city, County, and State. We modified these so that they were all consistent with spelling. We then noticed that there were a lot of non URL’s in some of the columns that should be URL’s. We didn’t want to remove the information if it wasn’t a URL because then you lose data. So we decided to use GREL to modify the fields to become URL links. We wanted them to be consistent throughout. We did the same with the other URL columns. Once this was done, we felt that our table was ready to use.

One of the difficulties we ran in to was the fact that in the URL columns, some had just regular text. Most of them were either in there by error (and didn't help), or just contained the location inside of Facebook, or Youtube or whatever column it was under. That is why we created them all to be links, so that everything would go to the correct place. However, if it did not, then the link would basically not work which is better than them having to try and find the page manually. We almost decided to use a Python script for this, but after some research and testing found that it could be done directly in OpenRefine with GREL.

Looking at some of the integrity constraints. We wanted to make sure that each of the FMID's were unique. Two stores shouldn't have the same FMID. We also wanted to ensure that none of the x and y columns were null. If the x was null and had a y value, the y value does nothing for you. So we made sure that all of the values were provided. We also wanted to make sure that when you used the x and y coordinate, that it uniquely identified a specific location. We also made sure that if there were any null values in any of the products provided, that it would result in a FALSE/N value.

While cleaning the data, each category was dependent on its column. As we modified them, we had to make sure everything was consistent within the columns to ensure that no problems were introduced into the data. As we began cleaning it, it was taking the existing information as input, and as we selected what we thought was best/correct, it would output to the decision we made.

Changes to the dataset:
Modified MarketName – 667 cells edited
Modified street – 154 cells edited
Modified city – 2656 cells edited
Modified County – 734 cells edited
Modified Website – 160 cells edited
Modified Facebook – 122 cells edited
Modified Twitter – 24 cells edited
Modified OtherMedia – 47 cells edited
Modified x – 8656 cells edited
Modified y – 8656 cells edited
Modified updateTime – 8146 cells edited
Modified street – 11 cells edited
Modified Facebook – 909 cells edited
Modified Twitter – 346 cells edited

The Farmers Markets dataset is a lot cleaner in the sense that there is no confusion between multiple naming conventions, and they will be able to use the x and y coordinates and date field more effectively now. The clients would be able to understand the changes made and would hopefully be happy that it is cleaner and easier to use now. Some of the problems encountered were how to modify certain columns so that it wasn't losing any data, but was more correct than it was previously. If we had more time, we would do a more in depth modification of the URL columns to try and get them all to be very accurate. This would take more analysis and possible use of another language to get the results we desire. We really learned that data, even if it looks good, can't always be trusted. Sometimes you have to really look closely to make sure it is ready for use, and if it isn't, we learned numerous skills that can help us to make these corrections to make a dataset useful.