CS513- Data Cleaning Project Phase- I

Prepared By : Zoheb Satta(satta2@illinois.edu)/Bhushan Bathani(bbath2@illinois.edu)

We are going to execute the below plan for our data cleaning project.

1. **Identify a dataset :** We will be using farmers market dataset as provided.

2. **Develop a target main use Case**

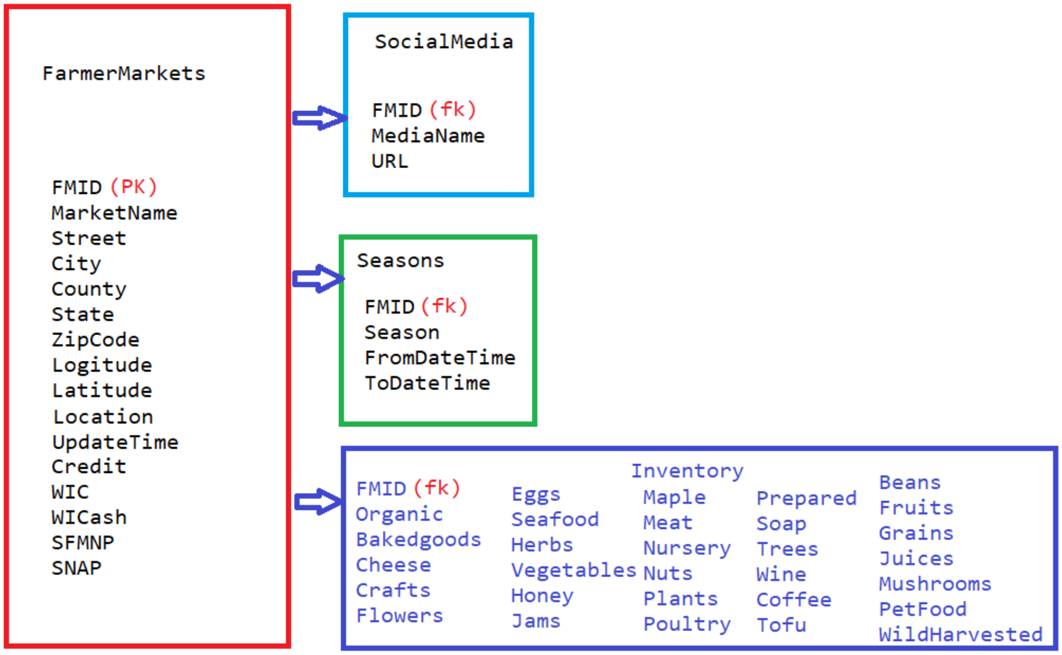
Use Case	**FMID**	**Data Quality issues**
U0 - Data cleaning is not required	1012158	
U1 - Data cleaning is required and possible	1000986	Street address is in caps.
	1001683	Season1 date is not in ISO Date format
	1009994 and 1010376	Market Name is same but in different State
	1009028	Market Name is in caps
	1009876 and 1001094	Zipcode is 4 digit
	1012802 and 1012121	State code is populated in zipcode column
	1007585	Market Name has grammar error - Farmers's
	1001490 and 1001491	Duplicate entry for the same market. Market Name has been adjusted in one of the entry
	1009990 and 1009991	Seems same market but street address is little different and some fields are not matching
U2- Data cleaning is not possible	2000035 and 2000007	Most of the fields is empty and update time is in year

Because the entire data with all fields is not possible to fit in, We have put only FMID here. Also above are only a few examples for given data quality issues. For the same category there are many others.

3. Describe the dataset

Dataset consists of basically 4 different informations about farmers market

1) Market Name, Address, location and payment type - FarmerMarkets
2) Social Media website - SocialMedia
3) Different Season and their open timings - Seasons
4) What each market sells - Inventory



4. Data Quality Problem

As described in #2 , we see below data quality problems at this point and we expect to find it more as we look deeper into it.

- Many of the fields do not have a consistent case across all data.
- Season1 date is not in ISO Date format in many places and only months are written.

- Zip Code is only 4 digit
- State code is populated in zip code column
- Market Name has grammar error - Farmers's
- Duplicate entry for the same market. Some of the fields has different names.

5. **Plan**

We are going to execute the below plan to fix each and every field.

1) Correct market name, street name, city, Country, state, location to title case.
2) Trim white spaces of all fields
3) Replace zip code to empty cell where it is not a number.
4) Make the zipcode to 5 digits where it is not by prefixing 0 and transform it to number.
5) Convert season date, Season Time, update Time to ISO Format
6) Convert season date which is in month to ISO Format date by taking first and last day as default date
7) Identify and remove duplicates or very closely matching records
8) Apply clustering and facets to convert data to uniformly.
9) Sort by state and market name to have a better view.
10) Remove the records about markets which do not have any information which are not possible to clean.
11) Apply and check integrity constraints on state and market name which should be unique.
12) Load the data in SQLLite and DB and apply integrity constraints. Also split data in 4 different data tables as above.
13) Prepare and fire different sql query to analyze data which should provide user some useful information e.g. please give me any market open currently in my state where vegetables are available and cash payment is acceptable.

Team Member Distribution : We are thinking of distributing the work as below.

1) Initial cleaning using openrefine - Bhushan
2) Remove data which can't be cleaned - Bhushan
3) Write down a datalog script to fix any integrity constraint - Zoheb
4) Create 4 different tables in relational database - Zoheb
5) Write down a script to parse and load data into different tables - Zoheb and Bhushan
6) Validate and create different use case queries - Zoheb and Bhushan

Tools : We are planning to use OpenRefine, Data Log, R/Python, Regex And Tablea but that may vary as we go through the different use cases and scenarios.