**Farmers Market Dataset**

**1. Introduction and Overview**

The Farmers Market dataset is an extensive dataset containing valuable data regarding the location, season, days and time of operation, social media handles, products on sale, mode of accepted payment to name a few. The main aim of this project is to help users find the famers market in their locality which are either nut free, vegan, organic, gluten free or markets selling prepared food. This would save the shoppers a lot of time as they can choose a market easily depending on the choice of produce/items they are particularly looking for.

The shoppers with specific food restrictions can now easily enjoy a great variety of fresh, local produce in seasons when its flavor and nutrition are at its peak. The prices of organic fruits and veggies are usually a lot cheaper and would help the buyer.

**2. Initial assessment of the dataset**

This dataset like many other datasets requires a fair bit of cleaning to be used for further analysis.

The address of each farmers market comprises of a street, city, county, state and zip. There are many records that are missing either a part or a few parts which makes the address incomplete.

The operating seasons are divided into Season1Date, Season2Date, Season3Date and Season4Date. These columns have a from and to date which needs to be extracted and formatted into separate columns making it easy for further analysis.

The season time column per season has cell values with day and time. This column needs to be split into operating days columns with operating times as cell values for further analysis.

**3. Data Cleaning methods and process**

OpenRefine has been extensively used for this project for the data cleaning phase. All the white spaces have been collapsed and trimmed. Incorrect data types have been inspected and converted into their correct data type.

**Common Transformations:**

The following columns have been transformed to number.

1. Zip
2. x co-ordinates
3. Y co-ordinates

The following columns have been transformed to date format

1. Season1Date
2. Season2Date
3. Season3Date
4. Season4Date
5. UpdateTime

**Facets and Clustering:**

Text facets were created for the following columns

1. MarketName
2. City
3. County
4. State

Clustering for performed on the following columns

1. MarketName

2. City

3. County

**Column splitting**

Major part of the data cleaning for this project for the data to be fit for further analysis involves column splitting. The available seasons and time are extremely vital data so splitting and formatting was required. The following columns were split into multiple columns Season1Date

1. Season1Time
2. Season2Date
3. Season2Time
4. Season3Date
5. Season3Time
6. Season4Date
7. Season4Time

| SeasonDate | |
| --- | --- |
| Existing Columns | New Columns |
| Season1Date | Season1StartDate |
|  | Season1EndDate |
| Season2Date | Season2StartDate |
|  | Season2EndDate |
| Season3Date | Season3StartDate |
|  | Season3EndDate |
| Season4Date | Season4StartDate |
|  | Season4EndDate |

| SeasonTime | | | | |
| --- | --- | --- | --- | --- |
| Existing Columns | Intermediate Columns | Intermediate Columns | New Columns | New Columns cell value |
| Season1Time | Days | Time | 1Sunday | Time |
|  |  |  | 1Monday | Time |
|  |  |  | 1Tuesday | Time |
|  |  |  | 1Wednesday | Time |
|  |  |  | 1Thursday | Time |
|  |  |  | 1Friday | Time |
|  |  |  | 1Saturday | Time |
| Season2Time | Days | Time | 2Sunday | Time |
|  |  |  | 2Monday | Time |
|  |  |  | 2Tuesday | Time |
|  |  |  | 2Wednesday | Time |
|  |  |  | 2Thursday | Time |
|  |  |  | 2Friday | Time |
|  |  |  | 2Saturday | Time |
| Season3Time | Days | Time | 3Sunday | Time |
|  |  |  | 3Monday | Time |
|  |  |  | 3Tuesday | Time |
|  |  |  | 3Wednesday | Time |
|  |  |  | 3Thursday | Time |
|  |  |  | 3Friday | Time |
|  |  |  | 3Saturday | Time |
| Season4Time | Days | TIme | 4Sunday | Time |
|  |  |  | 4Monday | Time |
|  |  |  | 4Tuesday | Time |
|  |  |  | 4Wednesday | Time |
|  |  |  | 4Thursday | Time |
|  |  |  | 4Friday | Time |
|  |  |  | 4Saturday | Time |

**4. Data cleaning Results**

1. Relational Database schema

b. Workflow Model

c. Showcase overall changes

**5. Conclusion and Future Work**

As part of the future scope of work we plan to validate the zip codes info using a Python script with some help from the available APIs. There are some parts of the address that can be identified and rendered based on data from the other columns. We plan to populate the missing address parts in all possible cells that can be correctly rendered. For example if all fields like street, city, county and state are available, empty zip code value can be replaced with the correct zip code. The only column as part of the complete address that cannot be generated is “street”.

We have begun work in this direction and also attached a snapshot of the populated set of rows.