Part 1

The data chosen to be cleaned and analyzed was farmer’s market data across the entire United States including territories of Puerto Rico and the U.S Virgin Islands. The unedited data originally has 8812 rows with 59 columns all in one csv file. The data contains much of the information you would except to see for farmers market which includes the market name, social media information, address information, types of items sold at the market, time and date of operating hours, and time the information was last updated. A high overview of the data shows much of the information is fairly structured and relatively easy to read. There are however many distinctive problems with the data that would be a problem for further analysis.

All of the Market names are shown to be present in the data but are not always written in the same manner. Some punctuation and capitalization for similar names is different which can be difficult for more precise analysis. Social media information seems to be only present if the market itself has any. A lot of this information is null since its assuming many don’t have a digital presence. Address information is also not all updated exactly, some markets have coordinates but do not have city, county, zip code, or state information. This is later discussed on how exactly the data was added in the cleaning process. Information about times and dates are very difficult to read with most of the times for different days of the week all put in the same column. Having the data spread out for each day of the week and time would have been a lot easier to read and better for analysis to handle. Coordinate data using longitude and latitude all seem to be populated which is important to use to be able to locate the markets on a map even without exact address information. Payment information as well as food sold is shown with either Y as yes N as no or blank for null values. Some columns have dashes as null values which should be ignored or changed to null. Finally all update times seem to be accurate as well. However any column that has a date is not in a consistent date format. Much of the cleaning process will involve changing the dates to a suitable date time format for easier analysis.

The data use cases identified was to be able to easily find a farmers market close to oneself by address and be able to look up times for that markets open hours, while also being able to identify the items that could be found at the specific market. Other higher level cases would show what days of the week are most open, number of markets per state, and the different items sold at each market showing which ones are most present from greatest to least. One of the main goals was to completely fill out all the address information for each market that had missing data for city, county, state and zip code information. This is essential to identifying where the actual market is located. Other information like having different times for each day of the week for each season of the farmers market is also essential to be able to show. Having all the date time formats in the same manner and formatting all the strings in a consistent matter will also be done.

The dataset in itself has enough clean data that already allows for some analysis. The coordinate data could be put into a tableau map that would show all the places where a market exists in a geographical format, since almost every row has coordinate values. The items found in each market is also in the necessary format and can be analyzed further right away. Social media is something that is very difficult since many of these markets need to be online anyway to have something. If they only advertise locally then there is no need for this data. For data analysis it is not very useful. Also there is a column that describes the location venue of the actual market. This column is missing lots of values and someone would physically need to describe every farmers market in the dataset to fill in these values. Something that would take lots of time and effort to do.

Part 3

The data is loaded into an SQL database file (.db) using the python sqlite3 library. The file that is loaded is the cleaned csv file after it has been cleaned in Open Refine. Once the database file is measured appropriate queries can be run using a GUI like tool called sqlite studio. The integrity constraints checked for the database is to first check that each market has its own unique id (FMID) this ensure that there are no duplicates. Once that is done then the database is checked for markets with the same name, street, city, and state. This shows that a market was input into the database twice. The query will update the database based off of the newest update time and delete the duplicate values. The final integrity constraints look at season dates and times. Because there are 4 different seasons to input if a season column has the same dates and times as another season column then the larger season number would be left blank given this is duplicate data. This is done for all seasons and all data.

Part 5

Further work was done to show the final results of the clean data set using tableau as a way to visualize the data for the user.