CS513: Theory & Practice of Data Cleaning

Individual Submission: Nadia Wood (nadiaw2)

Team115

Git repo:https://github.com/nadiawoodninja/cs513-data-cleaning

1. **Dataset of interest**: farmersmarkets.csv. The source of this data is https://www.ams.usda.gov/local-food-directories/farmersmarkets
   There is an API available to download the farmers market data at this URL:
   `https://www.usdalocalfoodportal.com/api/farmersmarket/`
   According to the website "The Farmers Market Directory lists markets that feature two or more farm vendors selling agricultural products directly to customers at a common, recurrent physical location. Maintained by the Agricultural Marketing Service, the Directory is designed to provide customers with convenient access to information about farmers market listings to include: market locations, directions, operating times, product offerings, accepted forms of payment, and more."

2. **Initial Data Profiling**: I used the Pandas profiling library to develop a profile of the data to understand it better.
   Link to the data profile:
   https://htmlpreview.github.io/?https://github.com/nadiawoodninja/cs513-data-cleaning/blob/main/DataProfiling/farmersDataStats.html

Overview | Alerts 115 | Reproduction

Dataset statistics

Number of variables	59
Number of observations	8665
Missing cells	185185
Missing cells (%)	36.2%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	3.9 MiB
Average record size in memory	472.0 B

Variable types

Numeric	3
Categorical	22
Boolean	34

Names of all the columns in the csv

```
In [8]: my_list = list(df_m)
        print (my_list)

['FMID', 'MarketName', 'Website', 'Facebook', 'Twitter', 'Youtube', 'OtherMedia', 'street', 'city', 'County', 'Stat
e', 'zip', 'Season1Date', 'Season1Time', 'Season2Date', 'Season2Time', 'Season3Date', 'Season3Time', 'Season4Date',
'Season4Time', 'x', 'y', 'Location', 'Credit', 'WIC', 'WICcash', 'SFMNP', 'SNAP', 'Organic', 'Bakedgoods', 'Cheese',
'Crafts', 'Flowers', 'Eggs', 'Seafood', 'Herbs', 'Vegetables', 'Honey', 'Jams', 'Maple', 'Meat', 'Nursery', 'Nuts',
'Plants', 'Poultry', 'Prepared', 'Soap', 'Trees', 'Wine', 'Coffee', 'Beans', 'Fruits', 'Grains', 'Juices', 'Mushroom
s', 'PetFood', 'Tofu', 'WildHarvested', 'updateTime']
```

3. **Use Case U1 (main target):** Given the popularity and usage of credit cards (Apple Pay, Android Pay etc.), an interesting use case to develop would be to identify the markets that accept credit card in a certain geo location. A heatmap could be created to show which geo locations are using credit cards most and which ones are using least. This can be done with some data cleaning efforts given the current dataset.

   U0 use case that requires "zero data cleaning": A possible use case without data cleaning would be to determine the most and least popular products sold by markets by summing the 'Y' for each product's column. This can be pivoted in various other columns such as location of the market.

   U2 is a use case data "never (good) enough": Any use cases surrounding the Season2, Season3 and Season4 columns are never going to be useful due to a lot of missing data. No amount of wrangling, cleaning will be able to give us any insights into the data.

4. **Describe the dataset D**:
   There are 8665 entries in the dataset and 59 columns. Further details can be found in the data profiling report generated by pandas_profiling report.

```
In [24]: df_m.info(verbose=True)

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8665 entries, 0 to 8664
Data columns (total 59 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   FMID           8665 non-null   int64
 1   MarketName     8665 non-null   object
 2   Website        5207 non-null   object
 3   Facebook       3796 non-null   object
 4   Twitter        997 non-null    object
 5   Youtube        161 non-null    object
 6   OtherMedia     638 non-null    object
 7   street         8380 non-null   object
 8   city           8625 non-null   object
 9   County         8127 non-null   object
 10  State          8665 non-null   object
 11  zip            7721 non-null   object
 12  Season1Date    5386 non-null   object
 13  Season1Time    5525 non-null   object
 14  Season2Date    429 non-null    object
 15  Season2Time    414 non-null    object
 16  Season3Date    79 non-null     object
 17  Season3Time    75 non-null     object
 18  Season4Date    7 non-null      object
 19  Season4Time    7 non-null      object
 20  x              8636 non-null   float64
 21  y              8636 non-null   float64
 22  Location       2936 non-null   object
 23  Credit         8665 non-null   object
 24  WIC            8665 non-null   object
 25  WICcash        8665 non-null   object
 26  SFMNP          8665 non-null   object
 27  SNAP           8665 non-null   object
 28  Organic        8665 non-null   object
 29  Bakedgoods     5642 non-null   object
 30  Cheese         5642 non-null   object
 31  Crafts         5642 non-null   object
 32  Flowers        5642 non-null   object
 33  Eggs           5642 non-null   object
 34  Seafood        5642 non-null   object
 35  Herbs          5642 non-null   object
 36  Vegetables     5642 non-null   object
 37  Honey          5642 non-null   object
 38  Jams           5642 non-null   object
 39  Maple          5642 non-null   object
 40  Meat           5642 non-null   object
 41  Nursery        5642 non-null   object
 42  Nuts           5642 non-null   object
 43  Plants         5642 non-null   object
 44  Poultry        5642 non-null   object
 45  Prepared       5642 non-null   object
 46  Soap           5642 non-null   object
 47  Trees          5642 non-null   object
 48  Wine           5642 non-null   object
 49  Coffee         5642 non-null   object
 50  Beans          5642 non-null   object
 51  Fruits         5642 non-null   object
 52  Grains         5642 non-null   object
 53  Juices         5642 non-null   object
 54  Mushrooms      5642 non-null   object
 55  PetFood        5642 non-null   object
 56  Tofu           5642 non-null   object
 57  WildHarvested  5642 non-null   object
 58  updateTime     8665 non-null   object
```

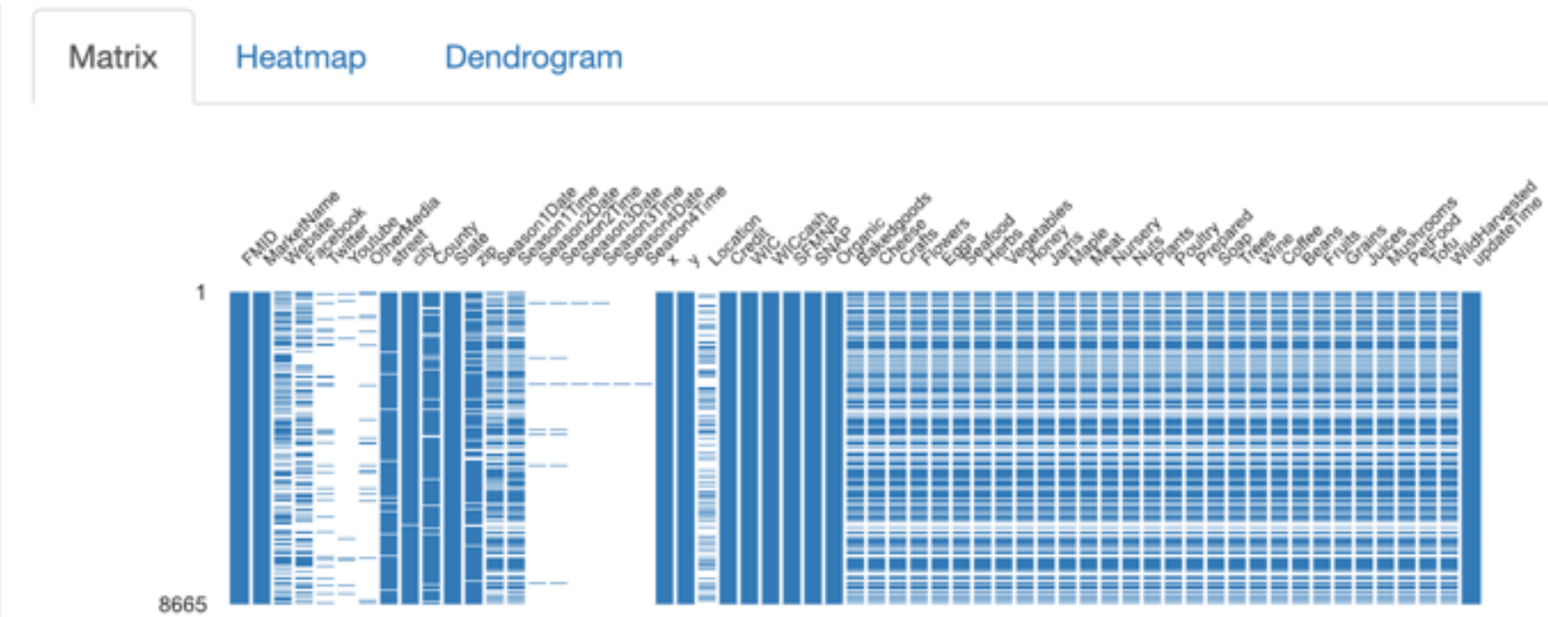
Columns and their description:

- FMID – a 7 digit integer unique identifier for each farmers' market
- MarketName - a string containing the name of the farmers' market
- Website, Facebook, Twitter, Youtube, OtherMedia - a string containing URL or other information that identifies the social media information
- street, city, County, State, zip - strings containing data corresponding to the column name that identifies the location of the farmers' market
- Season1Date, Season1Time, Season2Date, Season2Time, Season3Date, Season3Time, Season4Date, Season4Time - date fields representing the start date and end date for the given farmers' market or the times in which the farmers' markets are opened
- x, y - latitude and longitude coordinates
- location - a string describing the location of the farmers' market
- Credit, WIC, WICcash, SFMNP, SNAP - Y/N (boolean) character to indicate whether or not a given payment method is accepted
- Organic, Bakedgoods, Cheese...PetFood, Tofu, WildHarvested (30 columns) - Y/N (boolean) column to indicate whether or not a given product is offered

5. List obvious data quality problems (i.e., which are easy to spot during Phase-I).

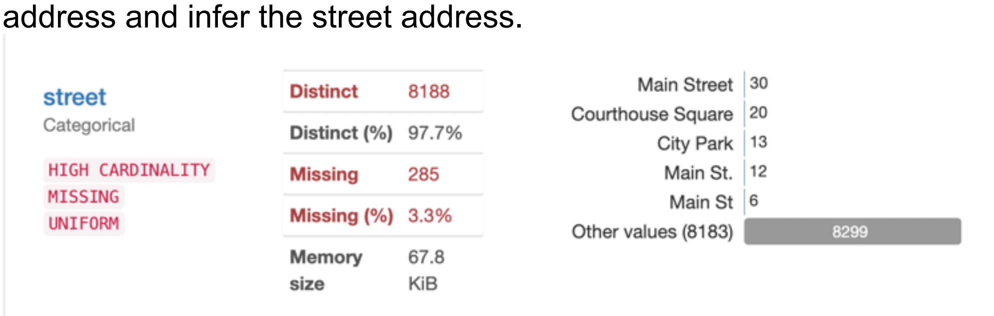
***U1**: Given the popularity and usage of credit cards (Apple Pay, Android Pay etc.), an interesting use case to develop would be to identify the markets that accept **credit card** in a certain **geo location**.*

Let’s observe the missing values by looking at the nullity matrix of the data generated by pandas_profiling. The nullity matrix is a data-dense display which allows one to quicky visually pick out patterns in data completion.

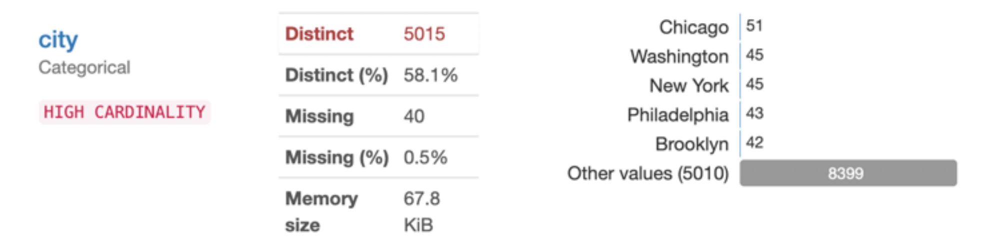


Doing a quick visual analysis, we can see that there are some missing data fields in the **street**, **city**, **county** and **zip** columns. Credit column does not have any missing data and seems to have either true or false values for each entry. Let’s deep dive into these columns.

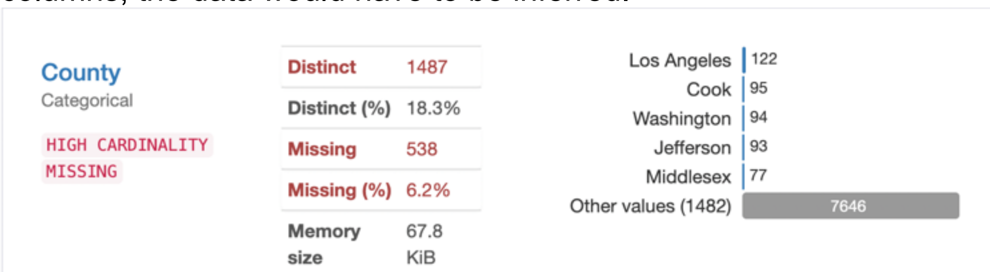
Street Column: It has 285 missing values. To generate a geo heatmap it would this column is not crucial to have. We can also look at other markets which may have similar address and infer the street address.



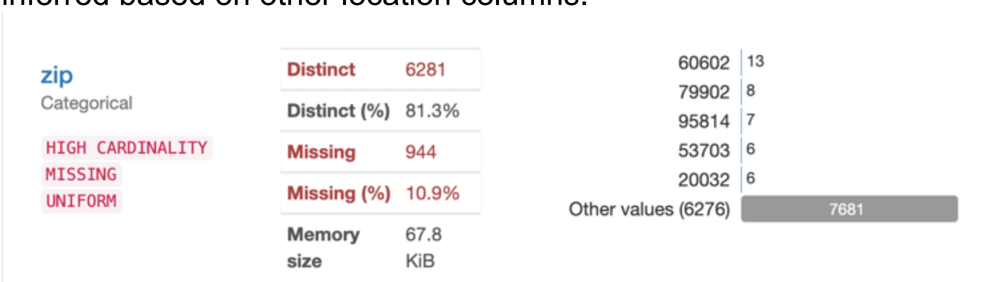
City: This column has 40 missing values. We would have to look at other location related columns to determine the city.

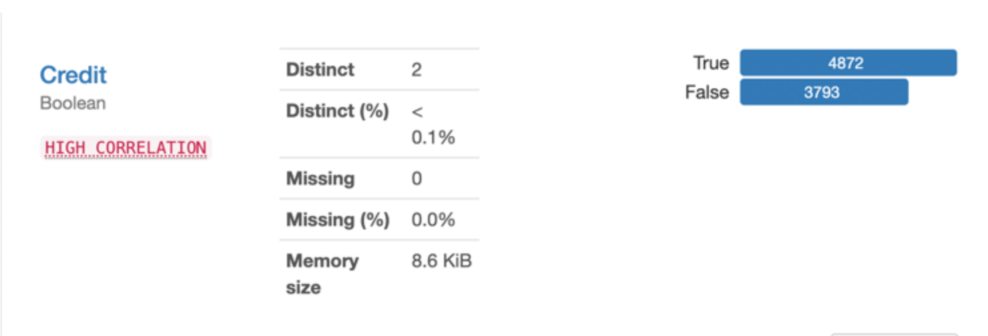


County: There is 6.2% missing data in this column. Again, looking at other location columns, the data would have to be inferred.



Zip: There is 10.9% of the data missing in this column. Again, the data would have be inferred based on other location columns.





6. Devise an initial plan that outlines how you intend to clean the dataset in Phase-II. A typical plan for the overall project will include the following steps:
   - **S1 & S2**: See the description of the dataset for U1 above and the profiling of the data. I used pandas_profiling for understanding the data and assessing what needs to be done for the use case I was targeting and that

if it is possible to answer the question given the dataset. OpenRefine and Python will be used to further clean the data such as inconsistent city names, any bad state names.

- **S3**: The tools I am targeting for this use case is OpenRefine, Pythong and Tableau. Tableau public is available for public to use and generate visualizations and analyze data. The tool has a very good toolset for geography based dataset. There are full datasets for locations that can be merged with the existing dataset to populate missing datapoints.
- **S4**: The nullability matrix should not show any missing data points for the new and improved dataset
- **S5**: Running the profiling against both datasets will give us the amount of changes made between the two datasets.