



CS513: Theory & Practice of Data Cleaning  
Individual Submission: Nadia Wood (nadiaw2)  
Team115

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

Tableau Dashboard: <https://public.tableau.com/app/profile/nadia.wood/viz/cs513/Dashboard1>

## 1. Dataset of interest:

- **Initial dataset:** 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.”
- **Output dataset:**
  - farmersmarkets\_output.csv



- farmersMarket\_location.csv
- farmeresmarkets\_payments.csv
- farmersmarkets\_products.csv

## 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 profiling report:**

<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)

['FMIID', 'MarketName', 'Website', 'Facebook', 'Twitter', 'Youtube', 'OtherMedia', 'street', 'city', 'County', 'State', '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', 'Mushrooms', 'PetFood', 'Tofu', 'WildHarvested', 'updateTime']
```

## 3. Project Use Cases

### 3.1. 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.



### 3.2. 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.

### 3.3. 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. Initial Assessment of the dataset

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
```

### 4.1. 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

## 4.2. Data quality problems

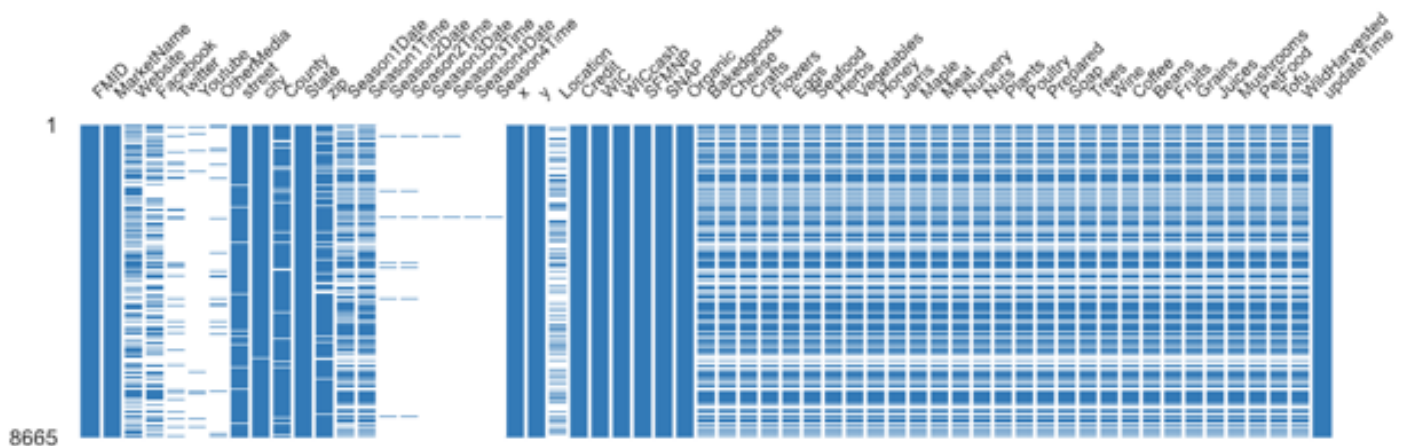
**U1:** Given the popularity and usage of credit cards (Apple Pay, Android Pay etc.), an interesting use case to develop would be to identify the markets that accept **credit card** in a certain **geo location**.

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 quickly visually pick out patterns in data completion.

Matrix

Heatmap

Dendrogram



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.

<b>street</b> Categorical  HIGH CARDINALITY MISSING UNIFORM	<b>Distinct</b>	8188	Main Street	30
	<b>Distinct (%)</b>	97.7%	Courthouse Square	20
	<b>Missing</b>	285	City Park	13
	<b>Missing (%)</b>	3.3%	Main St.	12
	<b>Memory size</b>	67.8 KiB	Main St	6
			Other values (8183)	8299

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

<b>city</b> Categorical  HIGH CARDINALITY	<b>Distinct</b>	5015	Chicago	51
	<b>Distinct (%)</b>	58.1%	Washington	45
	<b>Missing</b>	40	New York	45
	<b>Missing (%)</b>	0.5%	Philadelphia	43
	<b>Memory size</b>	67.8 KiB	Brooklyn	42
			Other values (5010)	8399

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

<b>County</b> Categorical  HIGH CARDINALITY MISSING	<b>Distinct</b>	1487	Los Angeles	122
	<b>Distinct (%)</b>	18.3%	Cook	95
	<b>Missing</b>	538	Washington	94
	<b>Missing (%)</b>	6.2%	Jefferson	93
	<b>Memory size</b>	67.8 KiB	Middlesex	77
			Other values (1482)	7646

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

<b>zip</b> Categorical  HIGH CARDINALITY MISSING UNIFORM	<b>Distinct</b>	6281	60602	13
	<b>Distinct (%)</b>	81.3%	79902	8
	<b>Missing</b>	944	95814	7
	<b>Missing (%)</b>	10.9%	53703	6
	<b>Memory size</b>	67.8 KiB	20032	6
			Other values (6276)	7681



<b>Credit</b> Boolean <b>HIGH CORRELATION</b>	<b>Distinct</b>	2	True	4872
	<b>Distinct (%)</b>	< 0.1%	False	3793
	<b>Missing</b>	0		
	<b>Missing (%)</b>	0.0%		
	<b>Memory size</b>	8.6 KiB		

## 5. Data Cleaning methods and process

- 5.1. Zip Code: Records were filtered where zip was missing and geo attributes (Longitude & Latitude) available. TomTom's Reverse Geocode API was used to derive zip code. (See farmersmarket\_impute\_zip.ipynb). After this step, I had only 0.3% missing zip code vs. 10.9%. The resulting dataset only had 3 records where it did not have any geographics attributes such as street, city, state or longitude, latitude.

<b>zip</b> Categorical <b>HIGH CARDINALITY</b> <b>UNIFORM</b>	<b>Distinct</b>	6863	60602	13
	<b>Distinct (%)</b>	79.4%	13601	10
	<b>Missing</b>	27	03256	10
	<b>Missing (%)</b>	0.3%	79902	8
	<b>Memory size</b>	67.9 KiB	10029	8
			Other values (6858)	8599

Toggle details

- 5.2. I used OpenRefine to clean the data. Please note that the tool works for smaller datasets and not large datasets
- 5.3. Cleaning started with MarketName column by first trimming the leading and trailing whitespace and then collapsing any consecutive whitespaces. Then a text facet was used and clustering to group similar MarketNames together. As seen below, the key collision method and the fingerprint keying function was used.



## Cluster & Edit column "MarketName"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method key collision

Keying Function fingerprint

209 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value	# Choices in Cluster
3	12	<ul style="list-style-type: none"><li>Main Street Farmers Market (10 rows)</li><li>Main Street Farmer's Market (1 rows)</li><li>Main Street Farmers' Market (1 rows)</li></ul>	<input checked="" type="checkbox"/>	Main Street Farmers Market	
3	5	<ul style="list-style-type: none"><li>Rochester Downtown Farmers Market (3 rows)</li><li>Downtown Rochester Farmers Market (1 rows)</li><li>Downtown Rochester Farmers' Market (1 rows)</li></ul>	<input checked="" type="checkbox"/>	Rochester Downtown Farmers	
3	3	<ul style="list-style-type: none"><li>Harrison Farmer's Market (1 rows)</li><li>Harrison Farmers Market (1 rows)</li><li>Harrison Farmers' Market (1 rows)</li></ul>	<input checked="" type="checkbox"/>	Harrison Farmer's Market	
3	4	<ul style="list-style-type: none"><li>Goshen Farmers Market (2 rows)</li><li>Goshen Farmer's Market (1 rows)</li><li>Goshen Farmers' Market (1 rows)</li></ul>	<input checked="" type="checkbox"/>	Goshen Farmers Market	
3	5	<ul style="list-style-type: none"><li>Irvinton Farmers Market (3 rows)</li><li>Irvinton Farmer's Market (1 rows)</li><li>Irvinton Farmers' Market (1 rows)</li></ul>	<input checked="" type="checkbox"/>	Irvinton Farmers Market	
3	4	<ul style="list-style-type: none"><li>Northfield Farmers' Market (2 rows)</li><li>Northfield Farmer's Market (1 rows)</li><li>Northfield Farmers Market (1 rows)</li></ul>	<input checked="" type="checkbox"/>	Northfield Farmers' Market	

Select All Unselect All

Export Clusters

Merge Selected & Re-Cluster

Merge Selected & Close

Close

- 5.4. Next, removed some of the columns that are **irrelevant** to both main use case and other potential use cases. The social media data quality was very poor and so those columns were also removed: Website, Facebook, Twitter, Youtube, OtherMedia. The time and date columns for Season2 onwards were also removed because there was very little data for these columns.



Facet / Filter **Undo / Redo** 13 / 14 **8687 rows**

Show as: **rows** records Show: 5 10 25 50 rows

Filter:

Extract... Apply...

1. Create project

2. Text transform on 392 cells in column MarketName: value.trim()

3. Text transform on 43 cells in column MarketName: value.replace(/s+/, '')

4. Mass edit 653 cells in column MarketName

5. Remove column Website

6. Remove column Facebook

7. Remove column Twitter

8. Remove column Youtube

9. Remove column OtherMedia

10. Remove column Season2Date

11. Remove column Season2Time

12. Remove column Season3Date

13. Remove column Season4Date

14. Remove column Season4Time

State	zip	Season1Date	Season1Time	Season4Time	x	y	Location	Credit	WIC	
Vermont	05828	06/14/2017 to 08/30/2017	Wed: 9:00 AM-1:00 PM;	Facet	140337	44.411036		Y	Y	N
Ohio		06/24/2017 to 09/30/2017	Sat: 9:00 AM-1:00 PM;	Text filter						
South Carolina	29682			Edit cells	7339387	41.3748009		Y	N	N
Missouri	64759	04/02/2014 to 11/30/2014	Wed: 3:00 PM-6:00 PM;Sat: 8:00 AM-1:00 PM;	Edit column					N	N
New York	10029	July to November	Tue:8:00 am - 5:00 pm;Sat:8:00 am - 8:00 pm;	Transpose					N	N
Tennessee	37204	05/05/2015 to 10/27/2015	Tue: 3:30 PM-6:30 PM;	Sort...					N	N
New York	10027	06/10/2014 to 11/25/2014	Tue: 10:00 AM-7:00 PM;	View					N	Y
Delaware	19801	05/16/2014 to 10/17/2014	Fri: 8:00 AM-11:00 AM;	Reconcile	-75.534480	39.742117	On a farm from: a barn, a greenhouse, a tent, a stand, etc	N	N	N
District of Columbia	20009	05/03/2014 to 11/22/2014	Sat: 9:00 AM-1:00 PM;		-77.0320505	38.9169984	Other	Y	Y	Y
District of Columbia	20011	04/09/2016 to 11/19/2016	Sat: 9:00 AM-1:00 PM;		-77.0334486	38.9559783		Y	Y	Y

5.5. Next. Switching the focus to the location columns - street, city, County, State, and zip. For street, GREL expressions were used to remove any special characters and substitute the ampersand with 'AND':

```
value.replace(/[!@#~!~?~:~;~, "']/, ' ').replace(/-[\ ]\(\)/, ' ').replace("&", 'AND')
```

### Custom text transform on column street

Expression

Language **General Refine Expression Language (GREL)**

```
value.replace(/[!@#~!~?~:~;~, "']/, ' ').replace(/-[\ ]\(\)/, ' ').replace("&", 'AND')
```

No syntax error.

**Preview** History Starred Help

row	value	value.replace(/[!@#~!~?~:~;~, "']/, ...
1.	null	Error: replace expects 3 strings, or 1 string, 1 regex, and 1 string
2.	6975 Ridge Road	6975 Ridge Road
3.	106 S. Main Street	106 S Main Street
4.	10th Street and Poplar	10th Street and Poplar
5.	112th Madison Avenue	112th Madison Avenue
6.	3000 Granny White Pike	3000 Granny White Pike
7.	400 West 40th Street and Adams Street	400 West 40th Street and Adams Street

On error

- ☒ keep original
- ☐ set to blank
- ☐ store error

☐ Re-transform up to 10 times until no change**OK** **Cancel**





5.6. Then, the leading and trailing whitespace were trimmed and collapsed any consecutive whitespaces and converted to uppercase.

OpenRefine FarmersMarket Permalink Or

Facet / Filter Undo / Redo 17 / 17 8687 rows

Extract... Apply... Show as: rows records Show: 5 10 25 50 rows « first < prev

Filter:

- Create project
- Text transform on 392 cells in column MarketName: value.trim()
- Text transform on 43 cells in column MarketName: value.replace(/s+/, '')
- Mass edit 653 cells in column MarketName
- Remove column Website
- Remove column Facebook
- Remove column Twitter
- Remove column Youtube
- Remove column OtherMedia
- Remove column Season2Date
- Remove column Season2Time
- Remove column Season3Date
- Remove column Season3Time
- Remove column Season4Date

F MID	MarketName	street	city	County	State	zip	Season1Date	Season1Time	x
1018261	Caledonia Farmers Market Association - Danville	Facet		Caledonia	Vermont	05828	06/14/2017 to 08/30/2017	Wed: 9:00 AM-1:00 PM;	-72.140337
1018318	Stearns Homestead Farmers' Market	Text filter					06/24/2017 to 08/30/2017	Sat: 9:00 AM-1:00 PM;	-81.7339387
1009364	106 S. Main Street Farmers Market	Edit cells							-2.8187
1010691	10th Steet Community Farmers Market	Edit column							-4.2746191
1002454	112st Madison Avenue	Transpose							-3.9493
1011100	12 South Farmers Market	Sort...							-6.790709
1009845	125th Street Fresh Connect Farmers' Market	View							-3.9482477
1005586	12th & Brandywine Urban Farm Market	Reconcile							-75.534460
1008071	14&U Farmers'	Cluster and edit...							-77.0320505

Common transforms

- Trim leading and trailing whitespace
- Collapse consecutive whitespace
- Unescape HTML entities
- To titlecase
- To uppercase
- To lowercase
- To number
- To date
- To text
- To null
- To empty string

5.7. Next step was to merge any logical clusters together using the key collision method and fingerprint keying function followed by ngram-fingerprint keying.

#### Cluster & Edit column "street"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method key collision Keying Function fingerprint 8 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	3	<ul style="list-style-type: none"><li>5TH AND MAIN STREET (1 rows)</li><li>5TH STREET AND MAIN STREET (1 rows)</li><li>MAIN STREET AND 5TH STREET (1 rows)</li></ul>	<input checked="" type="checkbox"/>	5TH AND MAIN STREET
3	3	<ul style="list-style-type: none"><li>DOWNTOWN MAIN STREET (1 rows)</li><li>MAIN STREET DOWNTOWN (1 rows)</li><li>MAIN STREET- DOWNTOWN (1 rows)</li></ul>	<input checked="" type="checkbox"/>	DOWNTOWN MAIN STREET
2	2	<ul style="list-style-type: none"><li>555 14TH STREET WEST (1 rows)</li><li>555 WEST 14TH STREET (1 rows)</li></ul>	<input checked="" type="checkbox"/>	555 14TH STREET WEST
2	2	<ul style="list-style-type: none"><li>1ST NORTH ST (1 rows)</li><li>NORTH 1ST ST (1 rows)</li></ul>	<input checked="" type="checkbox"/>	1ST NORTH ST
2	2	<ul style="list-style-type: none"><li>124TH STREET AND 5TH AVENUE/ MARCUS GARVEY PARK (1 rows)</li><li>MARCUS GARVEY PARK 124TH STREET AND 5TH AVENUE (1 rows)</li></ul>	<input checked="" type="checkbox"/>	124TH STREET AND 5TH AVE
2	2	<ul style="list-style-type: none"><li>3RD AND MAIN ST (1 rows)</li><li>MAIN ST AND 3RD (1 rows)</li></ul>	<input checked="" type="checkbox"/>	3RD AND MAIN ST

# Choices in Cluster

# Rows in Cluster

Average Length of Choices

Length Variance of Choices

Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close



## Cluster & Edit column "street"

Use

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

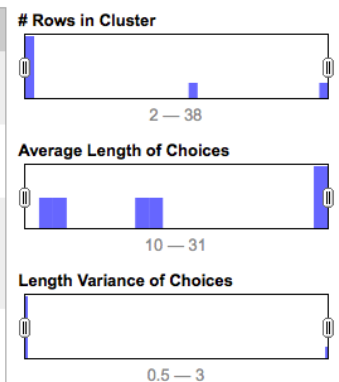
Method key collision

Keying Function ngram-fingerprint

Ngram Size 2

6 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	38	<ul style="list-style-type: none"><li>MAIN STREET (37 rows)</li><li>MAIN STREEET (1 rows)</li></ul>	<input checked="" type="checkbox"/>	MAIN STREET
2	2	<ul style="list-style-type: none"><li>12TH STREET (1 rows)</li><li>12THSTREET (1 rows)</li></ul>	<input checked="" type="checkbox"/>	12TH STREET
2	2	<ul style="list-style-type: none"><li>EAST SIDE OF COURT HOUSE SQUARE (1 rows)</li><li>EAST SIDE OF COURTHOUSE SQUARE (1 rows)</li></ul>	<input checked="" type="checkbox"/>	EAST SIDE OF COURT HOU
2	2	<ul style="list-style-type: none"><li>WEST SIDE OF COUNTY COURT HOUSE (1 rows)</li><li>WEST SIDE OF COUNTY COURTHOUSE (1 rows)</li></ul>	<input checked="" type="checkbox"/>	WEST SIDE OF COUNTY COL
<a href="#">Browse this cluster</a>				
2	2	<ul style="list-style-type: none"><li>COMMUNITY CENTER (1 rows)</li><li>UNITY COMMUNITY CENTER (1 rows)</li></ul>	<input type="checkbox"/>	COMMUNITY CENTER
2	22	<ul style="list-style-type: none"><li>COURTHOUSE SQUARE (20 rows)</li><li>COURT HOUSE SQUARE (2 rows)</li></ul>	<input checked="" type="checkbox"/>	COURTHOUSE SQUARE



Select All Unselect All

Export Clusters

Merge Selected & Re-Cluster

Merge Selected & Close Close

5.8. Same steps were used (remove special characters, trim and collapse whitespace, convert to uppercase, clustering) for the city, County, and State columns. After this process, the address information is much cleaner and more consistent.

OpenRefine FarmersMarket [Permalink](#) Open... Export Help

Facet / Filter Undo / Redo 36 / 37 Extract... Apply...

8687 rows Show as: rows records Show: 5 10 25 50 rows Extensions: Wikidata

Filter:

23. Text transform on 2 cells in column city: value.replace(/s+/,'')

24. Text transform on 8547 cells in column city: value.toUpperCase()

25. Mass edit 5 cells in column city

26. Text transform on 24 cells in column city: value.replace(" ", "")

27. Mass edit 63 cells in column city

28. Text transform on 126 cells in column County: grel.value.replace(/%@#?;:~"/, ""); replace(/-|\_|\\|/,"").replace("&","AND")

29. Text transform on 0 cells in column County: value.trim()

30. Text transform on 0 cells in column County: value.replace(/s+/,'')

31. Text transform on 8140 cells in column County: value.toUpperCase()

32. Mass edit 17 cells in column County

33. Text transform on 0 cells in column State: grel.value.replace(/%@#?;:~"/, ""); replace(/-|\_|\\|/,"").replace("&","AND")

34. Text transform on 0 cells in column State: value.trim()

35. Text transform on 0 cells in column State: value.replace(/s+/,'')

36. Text transform on 8687 cells in column State: value.toUpperCase()

FMID	MarketName	street	city	County	State	zip	Season1Date	Season1Time	x	y
1018261	Caledonia Farmers Market Association - Danville	6975 RIDGE ROAD	DANVILLE	CALEDONIA	VERMONT	05828	06/14/2017 to 08/30/2017	Wed: 9:00 AM-1:00 PM;	-72.140337	44.411036
1018318	Stearns Homestead Farmers' Market	106 S. Main Street	PARMA	CUYAHOGA	OHIO		06/24/2017 to 09/30/2017	Sat: 9:00 AM-1:00 PM;	-81.7339387	41.3748009
1009364	106 S. Main Street Farmers Market	106 S. Main Street	SIX MILE		SOUTH CAROLINA	29682			-82.8187	34.8042
1010691	10th Street Community Farmers Market	10TH STREET AND POPLAR	LAMAR	BARTON	MISSOURI	64759	04/02/2014 to 11/30/2014	Wed: 3:00 PM-6:00 PM; Sat: 8:00 AM-1:00 PM;	-94.2746191	37.4956280
1002454	112st Madison Avenue	112TH MADISON AVENUE	NEW YORK	NEW YORK	NEW YORK	10029	July to November	Tue: 8:00 am - 5:00 pm; Sat: 8:00 am - 8:00 pm;	-73.9493	40.7939
1011100	12 South Farmers Market	3000 GRANNY WHITE PIKE	NASHVILLE	DAVIDSON	TENNESSEE	37204	05/05/2015 to 10/27/2015	Tue: 3:30 PM-6:30 PM;	-86.790709	36.118370
1009845	125th Street Fresh Connect Farmers' Market	163 WEST 125TH STREET AND ADAM CLAYTON POWELL JR BLVD	NEW YORK	NEW YORK	NEW YORK	10027	06/10/2014 to 11/25/2014	Tue: 10:00 AM-7:00 PM;	-73.9482477	40.8089533
1005586	12th & Brandywine Urban Farm Market	12TH AND BRANDYWINE STREETS	WILMINGTON	NEW CASTLE	DELAWARE	19801	05/16/2014 to 10/17/2014	Fri: 8:00 AM-11:00 AM;	-75.534460	39.742117
1008071	14&U Farmers' Market	1400 U STREET NW	WASHINGTON	DISTRICT OF COLUMBIA	DISTRICT OF COLUMBIA	20009	05/03/2014 to 11/22/2014	Sat: 9:00 AM-1:00 PM;	-77.0320505	38.9169994
1012710	14th & Kennedy Street Farmers Market	5500 COLORADO AVENUE NW	WASHINGTON	DISTRICT OF COLUMBIA	DISTRICT OF COLUMBIA	20011	04/09/2016 to 11/19/2016	Sat: 9:00 AM-1:00 PM;	-77.0334486	38.9559783
1016782	175th Street Greenmarket	175TH STREET BETWEEN WADSWORTH AND BROADWAY	NEW YORK	NEW YORK	NEW YORK	10033	06/29/2017 to 11/22/2017	Tue: 8:00 AM-6:00 PM;	-73.938049	40.846354

5.9. Next step was to switch focus to Season1Date and Season1Time. These columns were removed as they are not relevant to the use case in focus.

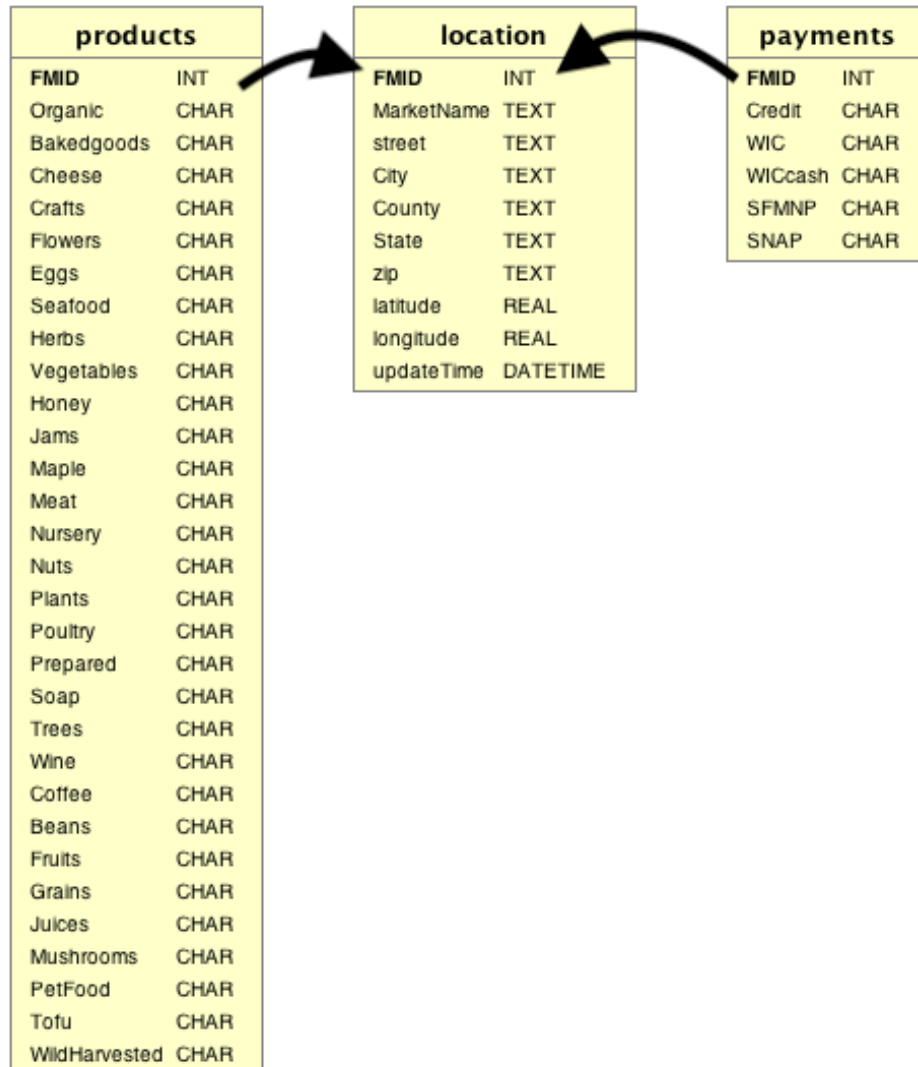


- 5.10. The x and y columns were renamed to latitude and longitude respectively, and then convert to numeric. The Location column was removed as it is not helpful for our purposes and is generally blank.
- 5.11. The values in the updateTime column were converted to ISO format using the GREL expression: `value.toDate('d/M/y H:m:s')` after trimming and collapsing whitespace.

## 6. Final Dataset

### 6.1. Relational Database Schema

The following Entity Relationship shows the schema developed for the final dataset. The cleaned dataset was broken into three separate tables (found in 2CleanedData): location, payments, and products, with the FMID as the primary key for all of them. This ER diagram was generated using DBVisualizer after loading the separate tables into sqllite.



Then, a few integrity constraints were developed using in sql-lite/sql-lite.ipynb notebook.

- Ensure that FMID is an appropriate primary non-null and unique key
- Ensure that data for my use case is non-null (specifically latitude, longitude, state)
- Ensure latitude is between [0,90] and longitude is between [-180, 180]
- Ensure FMID has unique address (street, City, County, State, zip) if it exists

## 6.2. Modeling the Workflow

I modeled the workflow using or2yw tool, and the instructions found at

<https://pypi.org/project/or2ywtool/>, I was able to create a workflow model of our data cleaning

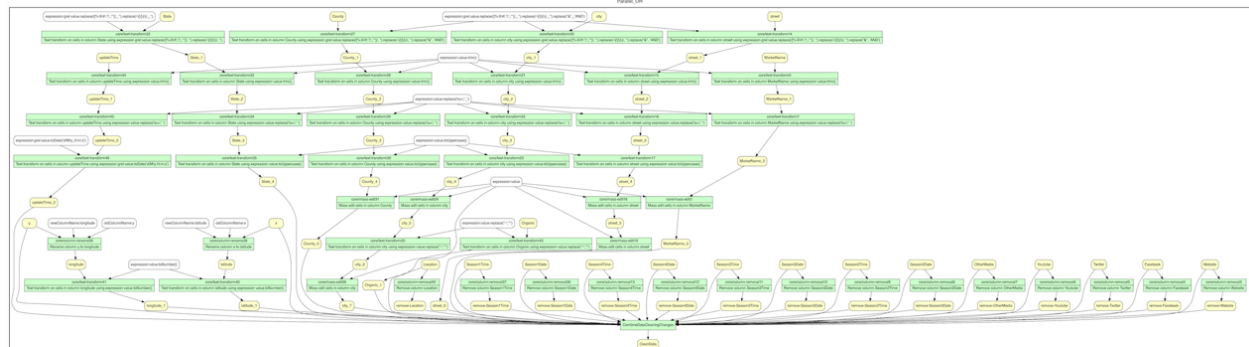


process very easily. I used the operations history json file from the OpenRefine Data Cleaning steps. Here the commands that I used:

1. pip install of the or2yw tool:  
`>pip install or2ywtool`
  2. Then, we generate both a serial and parallel yw file  
`>or2yw -i farmersmarkets_OperationHistory.json -o farmersmarkets_serial.yw`  
`>or2yw -i farmersmarkets_OperationHistory.json -o farmersmarkets_parallel.yw -t parallel`
  3. Install graphviz:  
`>brew install graphviz`
- and then generate the model using the following commands:
- ```
>or2yw -i farmersmarkets_OperationHistory.json -o farmersmarkets_parallel.png -ot png -t parallel
>or2yw -i farmersmarkets_OperationHistory.json -o farmersmarkets_linear.png -ot png
```

```
unset it.
Make sure your username (nadiawood) matches the one in your $HOME path.
See the "macOS Troubleshooting" section in the docs for more information.

The default interactive shell is now zsh.
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
(base) Nadias-M1-MBP:OpenRefine nadiawood$ pwd
/Users/nadiawood/Documents/GitHub/cs513-data-cleaning/2CleanedData/OpenRefine
(base) Nadias-M1-MBP:OpenRefine nadiawood$ ls
1_MarketPlace.png
2_Remove.png
3a_street_transform.png
3b_street_upper_case.png
3c_street_clustering_fingerprint.png
3d_street_clustering_ngram-fingerprint.png
3e_location_finished.png
6a_replace_Organic.png
6b_updateTime.png
farmersmarkets_OperationHistory.json
(base) Nadias-M1-MBP:OpenRefine nadiawood$ or2yw -i farmersmarkets_OperationHistory.json -o farmersmarkets_parallel.png -ot png -t parallel
File farmersmarkets_parallel.png generated.
(base) Nadias-M1-MBP:OpenRefine nadiawood$
```



### 6.3. Overview of changes

|                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|--------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| FMID                                                                                                   |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| MarketName                                                                                             | 392 leading/trailing whitespaces trimmed; 43 whitespace collapsed; 653 clustered                                                                                                                                                                                                                                                                                                                                                                                                      |
| Website, Facebook, Twitter, Youtube, OtherMedia                                                        | columns removed                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| street, city, County, State, zip                                                                       | street: 3175 cells had special characters removed/replaced; 305 had whitespaces trimmed, and 108 had whitespace collapsed, and 8301 were converted to uppercase, and 84 total were clustered<br>city: 917 had whitespaces trimmed and 2 had whitespace collapsed; 68 total clustered cells clustered, 24 had "-" replaced,<br>County: 126 cells had punctuation/special characters removed or replaced; 8140 were converted to uppercase<br>State: all converted to uppercase<br>zip: |
| Season1Date, Season1Time, Season2Date, Season2Time, Season3Date, Season3Time, Season4Date, Season4Time | Removed                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| x, y                                                                                                   | columns renamed to latitude and longitude, and 8658 cells converted to                                                                                                                                                                                                                                                                                                                                                                                                                |

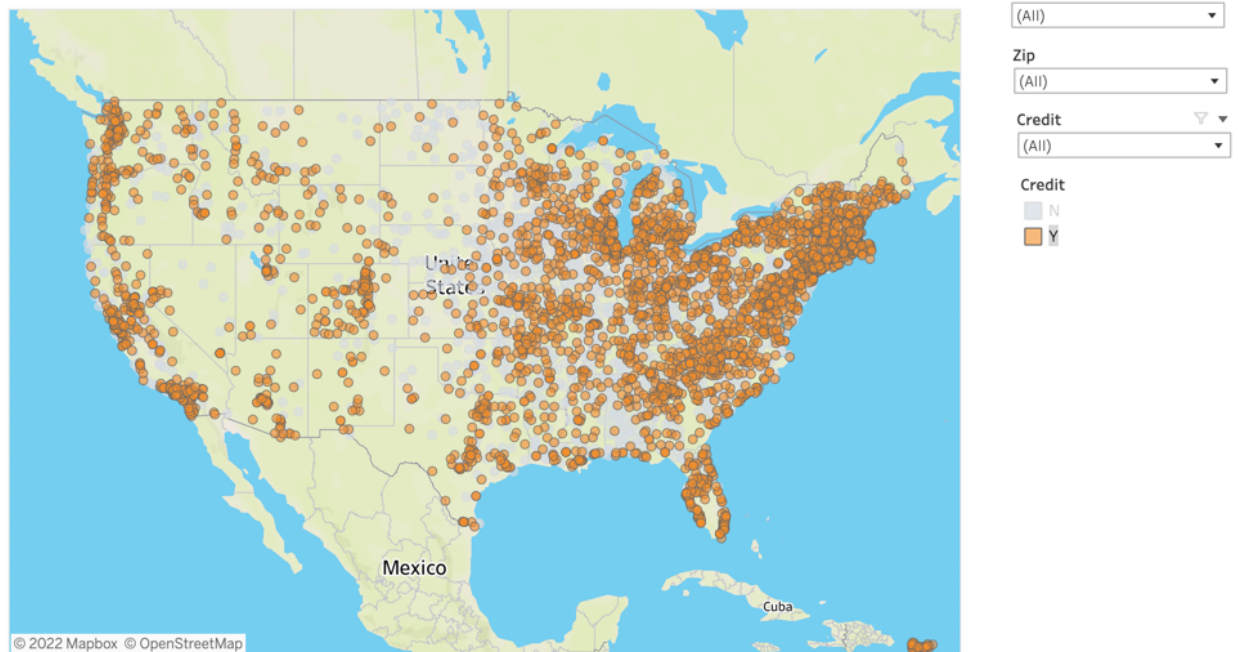


|                                                                         |                                                                    |
|-------------------------------------------------------------------------|--------------------------------------------------------------------|
|                                                                         | Numeric                                                            |
| location                                                                | removed                                                            |
| Credit, WIC, WICcash, SFMNP, SNAP                                       | No change                                                          |
| Organic, Bakedgoods, Cheese...PetFood, Tofu, WildHarvested (30 columns) | 5043 cells had "-" replaced in Organic column                      |
| updateTime                                                              | 219 cells had whitespace collapsed; 8384 cells changed to ISO date |

## 7. Conclusion

Now that we have a clean dataset, we can dive into our focus use case. I used tableau to create a dashboard (<https://public.tableau.com/app/profile/nadia.wood/viz/cs513/Dashboard1>) with drop downs (State, Zip Code and Credit card yes or no) . This will allow us to explore the data. From visualization, we can see that East coast, West coast and some parts of Midwest are dense with credit card acceptance.

Farmers Market

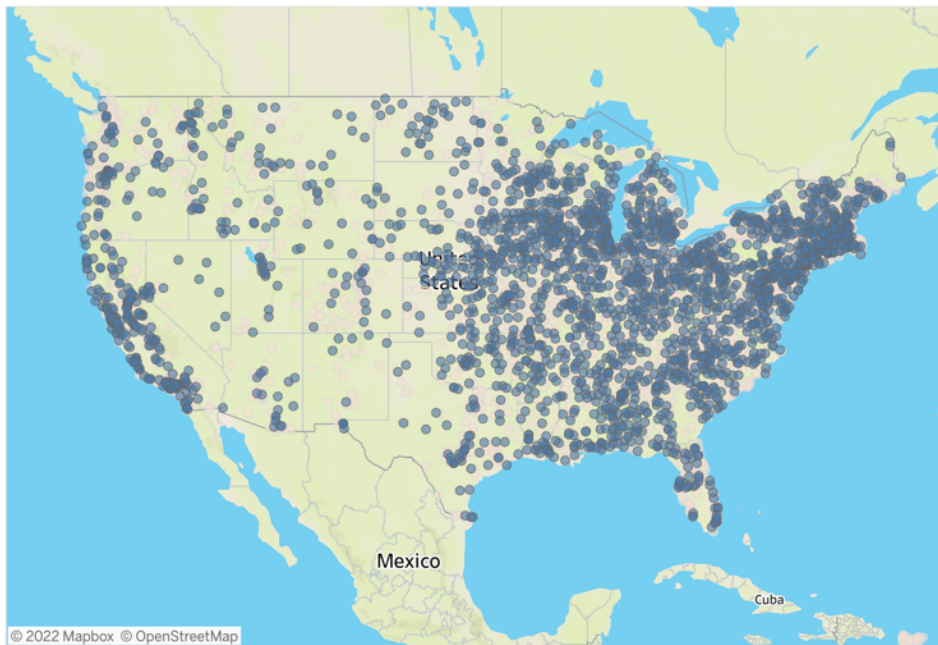


Here is another visualization of the zip codes that do not accept credit card. Comparing the two visualizations, it is hard to determine the difference between the two cases. So let's break it down by state.



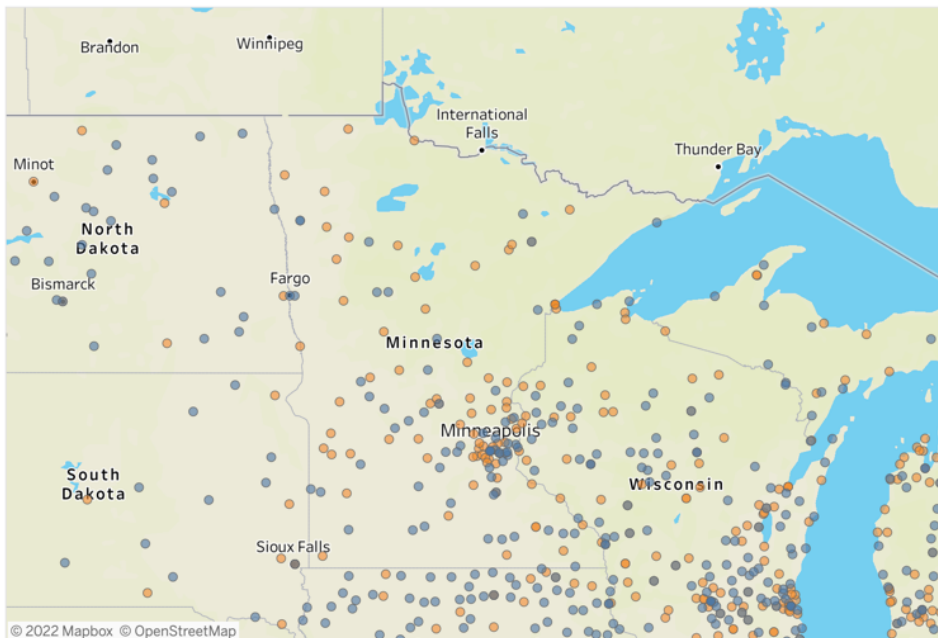


## Farmers Market



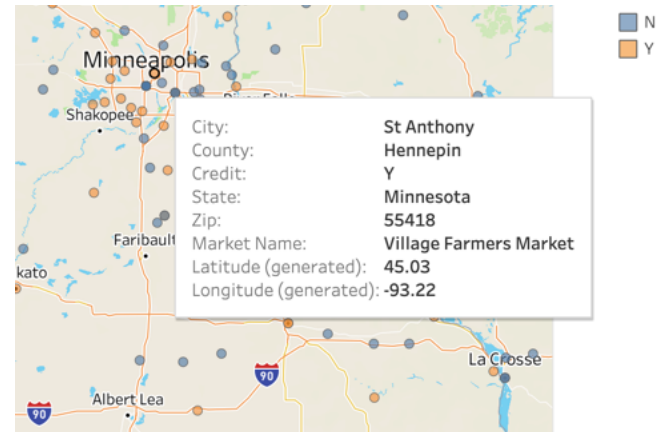
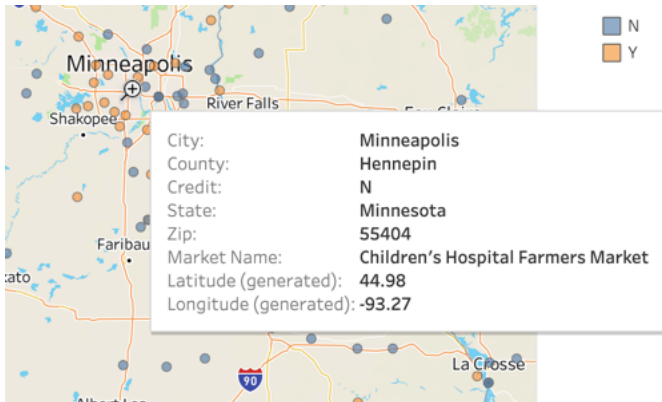
Here we observe just the state of Minnesota. The places closer to bigger cities like Minneapolis and Rochester accept credits cards vs. some of the smaller cities do not.

## Farmers Market



Within Minnesota, even within Hennepin County, there are markets which do not accept credit card. See below.





## 8. Future Considerations

The data can allow to slice and dice information based on products, credit cards, city, zip code etc. The data can also be enriched by population data. There are many different permutations of the data that can be done to understand the data. Another thing to note that Tableau itself, gives the ability to clean and manipulate the data as well. The downside of cleaning the data in tableau is that there is no tracking history of changes unlike OpenRefine and Yesworkflow. These tools are great to be able to reproduce the steps taken to clean and manipulate the data. Datasets can also be used to develop predictive models based on population growth and credit card acceptance rate.