DSC540 – Data Preparation – Final Project

By Meenakshi Shankara

Bellevue University

March 5, 2022

Vegan & Vegetarian Restaurants in the US

The objective of this project is to analyze the rise of vegan and vegetarian restaurants across the US amidst fast food and other restaurants. For this analysis, I will be working with 3 datasets. The datasets provide the information related to restaurants and fast foods eateries which include the location, name, cuisine, categories, data added to the system etc. which will help us analyze across various factors.

1. The first data source is a **Flat File Source – Fast food across US downloaded from data.world.**

This CSV provided the information of about 10,000 fast food restaurants Fast-Food restaurants across US, while providing various kinds of information including address, dates, categories.

https://data.world/datafiniti/fast-food-restaurants-across-america

Column descriptions – Important and relevant columns in the dataset

Sno	**Column**	**Type**	**Description**
1	Address	String	Address of the restaurant
2	City	String	City
3	Country	String	country = US
4	Keys	String	search key words
5	location	Geopoint	location point
6	name	String	name of the restaurant
7	postal code	String	zip code
8	Province	String	State
9	websites	url	website of the restaurant if available

For the purposes of this analysis, I have made the below transformations –

a. Renamed headers
b. Dropped columns that are not useful for the analysis
c. Dropped duplicates
d. Formatted Dates to a more readable format
e. Checked for Nulls and missing values
f. Sorted the final data set on Date Added and Date updated fields.

2. The second data source is **Website data – data scraped from HappyCow.net**

Data from Happy Cow provides a detailed information of restaurants from all over the world that are Vegan or Vegetarian or Veg-Friendly depending on our search criteria.

For this analysis, I have selected a consolidated count of such restaurants in each state of the United States.

https://www.happycow.net/

Data to be scraped – Search condition on How many Vegan, Vegetarian and Veg-friendly restaurants are available in each state of US -

Sno	Column	Type	Description
1	State	String	States in US
2	Count	Number	Number of restaurants
3	Type	String	Vegan/Vegetarian/Veg-Friendly

The below transformations were conducted for the data to be usable in the next steps.

a. Cleaning the Number column by removing the brackets
b. Removing junk records by comparing with a valid state code list
c. Sorting the dataset on state names
d. Merging the State Codes from the state list to our data set
e. Formatting the number column to int from object data type

3. The third data source **API Data from Kaggle – list of Vegan and Veg restaurants across the US.**

Using the Kaggle api packages available in Python, I was able to download the data from the datafiniti data source to obtain the list of al vegan and vegetarian restaurants across the country. The columns are very similar to the fast-food data set. Combining both will provide a set of all the restaurants in the US with most categories.

https://www.kaggle.com/datafiniti/vegetarian-vegan-restaurants

Column descriptions – Important and relevant columns in the dataset

Sno	Column	Type	Description
1	id	Unique ID	id
2	address	String	address of the vegan/veg restaurant
3	categories	String	search words
4	city	String	city
5	country	String	country = us
6	cuisines	String	type of cuisines
7	name	String	name of restaurant
8	postal code	String	zip code
9	province	String	state
10	website	url	website

The below transformations were conducted on the data set –

1. Rename headers
2. Drop columns not required for analysis
3. Drop columns with all Null/missing values
4. Drop duplicates in the data

5. Converts dates to readable format
6. Sort the data on date added field

The relationship between each of the data sources will be formed based on the location/region and name of restaurants, if and where possible.

For the Analysis the datasets from each of the above data sources were uploaded to a SQLiteDB. I have also created a table that has the consolidated data from all the sources.

The below Visualizations were created for analysis –

1. Bar graph on the Consolidated data– to visualize which states have the most Veg-friendly restaurants/eating joints
2. Pie chart – to Visualize the distribution of fast-food restaurants state-wise.
3. Line plot – to visualize the year when the list of restaurants was added to the system
4. Horizontal Bar plot – to visualize different cuisines that are veg-friendly
5. Scatter plots – to visualize the difference between data from different data sources.

There are a few ethical implications of cleaning the data sources.

1. When scraping data from HappyCow.net, I encountered the captcha issue. This could be a security measure the website has taken to discourage bot scraping. But I was able to work around this problem by calling the selenium web drivers and forcibly opening the website to scrape the html data. Although this is not a data cleansing ethical issue, this could be a potential issue when the data contains personal information that could be scraped and potentially be utilized incorrectly and unethically.
2. There were some junk state data in the second data source from Happycow.net. They were removed as they did not match with any state names. But the question regarding what those data points represent remains unanswered. Unless we understand how the data was collected, there is no way to predict how the removal of some records affect the results.
3. The data sources that were considered for this project were only a portion of the huge dataset that is not available freely. This could potentially mean that the analysis/visualizations were conducted on only a random sample of the datasets. I had removed some of the columns that had all null values. Since this was only a sample of the entire data set, I could've possibly deleted columns that probably had informative data points that were not available for us.

References –

1. https://data.world/datafiniti/fast-food-restaurants-across-america
2. https://www.happycow.net/
3. https://www.kaggle.com/datafiniti/vegetarian-vegan-restaurants