PPOL 563 Data Visualization
Data collection & cleaning report

Chau Nguyen

October 3, 2021

Source: For this report, I am working with the Liquor Licenses dataset from DC Open Data.

License: The data is under Creative Commons CC BY 4.0, meaning I am free to share, copy and redistribute if in any medium or format, as well as adapt, transform and build upon it for any purposes as long as I give appropriate credit, provide a link to the license, and indicate if changes were made[1].

Collection method: I downloaded the data. The dataset was created by the DC Geographic Information System (DC GIS) for the DC Office of the Chief Technology Officer (OCTO). The Alcoholic Beverage Regulation Administration (ABRA) provided DC GIS staff with a database identifying locations and attributes of liquor licensees in DC. The DC GIS staff geo-processed the data.

Biases and sampling: I do not expect biases or sampling issues with this dataset, because it's a simply a descriptive dataset listing liquor licensees in Washington DC. The raw data contains 2,130 observations and 32 rows.

Data Cleaning Methods & Identified Issues: I first used the missingno package to identify rows with missing numbers in my dataset. Figure 1 shows the features included in the dataset, as well as the missingness in each column. At first glance, I looks like the columns STORAGEFACILITY, DISTILLERY_PUB and SPORTS_WAGGERING contain all missing values.

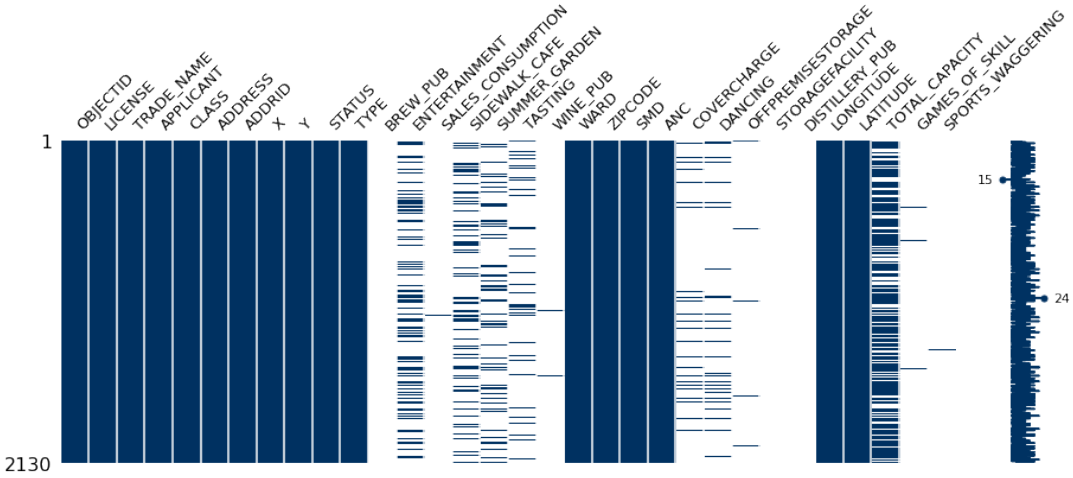

Figure 1: Missingno matrix for DC Liquor Licensees dataset

[1] Source: https://creativecommons.org/licenses/by/4.0/

Figures 1 also tells me that many of the features indicating what type of establishments the liquor lincensee is such as BREW_PUB and SIDEWALK_CAFE should be re-coded as binary or boolean (the current values are "nan" and "CHECKED").

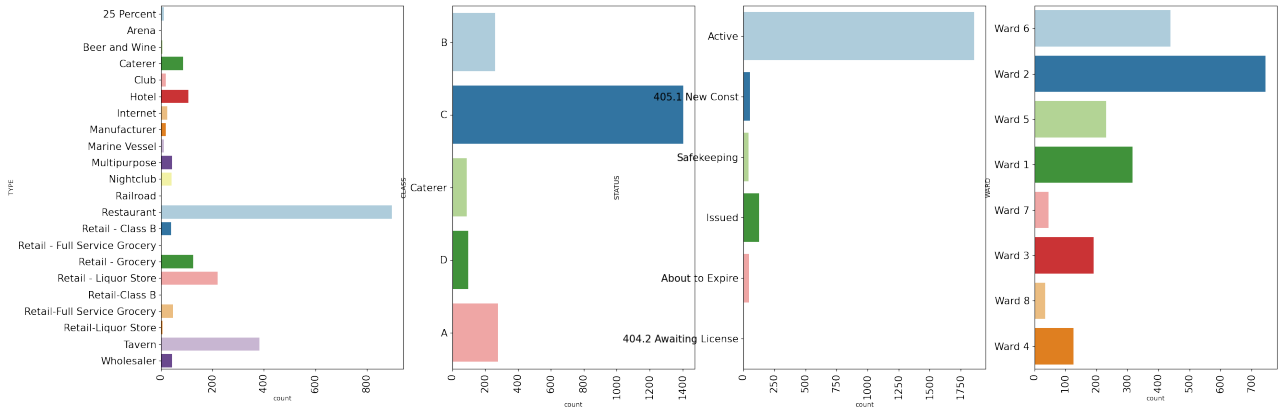

Figure 2: Categorical data

Figure 2 shows the number of observations for 4 categorical variables. From the left-most plot, we can see identify several issues with categories in the TYPE column: for example, "Retail - Full Service Grocery" and "Retail-Full Service Grocery" and "Retail - Full Service Grocery" or "Retail-Class B" and "Retail - Class B" are different categories due to the strings not matching.

A challenge with this dataset is that the metadata from Open DC does not tell me what the categories should be or their definitions, but refer questions to the Alcoholic Beverage Regulation Administration (ABRA) who does not a great job in differentiating the categories either. For now, I decided to keep "Catering" as a license class in addition to A, B, C, D, leave the categories for status as is, and recoded the Wards as integers instead of string. I also found 13 instances where LICENSE was duplicated across columns. After checking the duplicated license numbers, I decided it was ok to only keep the last row in case of duplication[2]. After cleaning, the dataset has 2,123 rows and no missing data for the binary variables.

Figure 1 also shows me that there are many liquor licensees in Washington with missing capacity data. I wanted to check the distribution of the available data without dropping rows with missing values, so I created a histogram for log(TOTAL_CAPACITY + 1) in figure 4. Total capacity for observations where data is available seem to follow a normal distribution.

Data Dictionary of Key Features

For this dataset, the important features can be grouped into several categories:

- **Identifiers:**

  LICENSE: ABRA License

- **Characteristics - categorical:**

  CLASS: License class

  STATUS: License status

  TYPE: Establishment type

[2]There were 6 licenses with 1 duplicate and 1 license with 2 duplicates - Gallaudet College. After checking, 2 of the duplicates for Gallaudet have the wrong street address - 700 Florida Ave NE when it should be 800. I kept the observation with the correct street address.

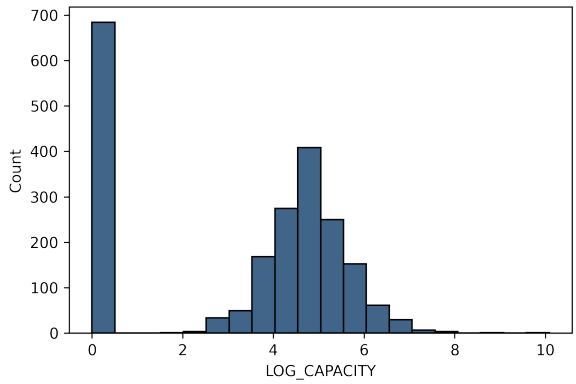

Figure 3: Histogram for log capacity of liquor licensees in DC

- **Characteristics - numerical:**

  TOTAL_CAPACITY: Establishment capacity

- **Location:**

  WARD

  ZIPCODE

  LONGITUDE

  LATITUDE

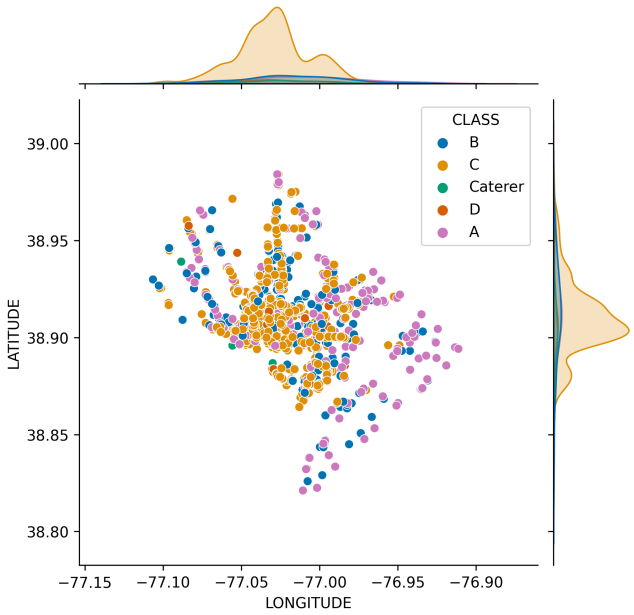

Figure 4: DC Liquor Licensees

This dataset itself can then be merged with different datasets such as Census Tract data for income and unemployment, DC crime maps, shapefiles for Ward and DC neighborhoods. I hope to be able to create different maps of DC neighborhoods & wards in relations to the kind of establishments in each clusters.