AtlantBH

Business Data Quality Report

- Analysis of POI dataset –

Author: Adnan Ovčina

Date, 10 August 2022

Introduction

Company ZeroPoint has put on the market dataset with locations of business in San Francisco. The supplier would like to sell POIs data at the highest market price. To make an informed decision about the potential procurement of the available dataset, I analyzed the data set to check data quality and consistency. The total number of data available in the dataset is 52 315. Data that are considered critical data (i.e., data of the utmost business importance) are stored in columns business name, business longitude, and business latitude. Data analysis was done in R.

Data analysis

Missing data

In the first step, I made an analysis of missing values. In total there are 1 887 missing values in the column business name. Furthermore, missing data were entered inconsistently, i.e., in some columns data were entered as "hidden", while in other cases word hidden was entered with a capital letter. Also, the term Unavailable was used to denote missing values.

To enable consistent tracking of missing values, all values in column business name that were entered as either "hidden", "Hidden", or "Unavailable", and all missing values (NAs) were transformed to "NA" (string) values. This step was taken to enable consistency in filtering, grouping, and summarizing data as well as checking the number of missing data. Analysis revealed that for 1 709 data business name values and 22 674 long. and lat. values are missing in the data set.

Data accuracy and completeness

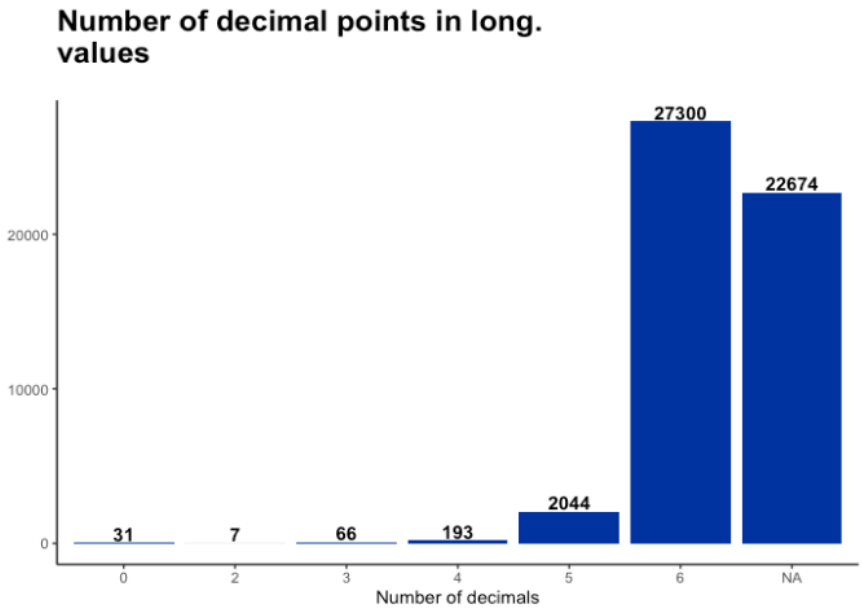
To check the correctness of data available in the dataset, firstly I checked if longitude and latitude values are all between valid ranges (long. values +-90 degrees and latitude values +-180 degrees). Analysis revealed that there are 4 data that fall outside valid ranges.

The precision of long. and lat. values are impacted by the number of decimal points. The relation between the number of decimal points and the precision of long. and lat. values is presented in the table below[1].

[1] Table was taken from https://en.wikipedia.org/wiki/Decimal_degrees

Decimal places	Decimal degrees	DMS	Object that can be unambiguously identified	N/S or E/W at equator	E/W at 23N/S	E/W at 45N/S	E/W at 67N/S
0	1.0	1° 00′ 0″	country or large region	111 km	102 km	78.7 km	43.5 km
1	0.1	0° 06′ 0″	large city or district	11.1 km	10.2 km	7.87 km	4.35 km
2	0.01	0° 00′ 36″	town or village	1.11 km	1.02 km	0.787 km	0.435 km
3	0.001	0° 00′ 3.6″	neighborhood, street	111 m	102 m	78.7 m	43.5 m
4	0.0001	0° 00′ 0.36″	individual street, large buildings	11.1 m	10.2 m	7.87 m	4.35 m
5	0.00001	0° 00′ 0.036″	individual trees, houses	1.11 m	1.02 m	0.787 m	0.435 m
6	0.000001	0° 00′ 0.0036″	individual humans	111 mm	102 mm	78.7 mm	43.5 mm
7	0.0000001	0° 00′ 0.00036″	practical limit of commercial surveying	11.1 mm	10.2 mm	7.87 mm	4.35 mm
8	0.00000001	0° 00′ 0.000036″	specialized surveying (e.g. tectonic plate mapping)	1.11 mm	1.02 mm	0.787 mm	0.435 mm

Number of decimal points in long. and lat. values impact the precision of data and its ability to unambiguously identify points of interest, and according to the table above values with 4 decimal points can unambiguously identify an individual street and large building. Three decimal points are used for the unambiguous identification of neighborhoods and streets. Identification with values that are composed of less than 3 decimal points is used for identification of towns, large cities, and countries which does not hold any relevance for navigator.ba. Data analysis shows 29 537 (56%) long. values have 4 or more than 4 digits and, while 66 (< 1%) have 3 decimal points. Finally, 38 data have less than three decimal points and can be considered as data of low quality. The precision of lat. values are less affected by the number of decimal points, hence data related to the number of decimal points presented below are limited only to long. values.



To analyze the precision of available data I have conducted row wise analysis of long. and lat. values. The analysis included a group summation of each level of decimal points available in the data set. I converted the data type in the business longitude column from character to factor. After the conversion of the data, type was completed I grouped and summarized the data based on each available factor level to get the total number of longs. values per number of decimal points. The table below shows the number of decimal points in all examined data. There are 22 674 (43%) missing long. and lat. values in the dataset out of which 9 rows are also missing data in the column business name.

Number of decimal points	Total
0	31
2	7
3	66
4	193
5	2044
6	27300
NA	22674

Data standardization

I also conducted an analysis of the standards used in the data set as they are critical when it comes to processing and mapping data. For the business city, data analysis shows that there are inconsistencies in entering the name of San Francisco. Data below show that there are 52 148 values "San Francisco", 136 SF, and 31 values are missing.

City	Total
NA	31
San Francisco	52148
SF	136

Analysis of state name data show that there 52 089 values entered as CA and 97 as California. Also, the analysis revealed that there are 129 data entered as ISO country code for Illinois which are incorrect data as San Francisco is in the state of California, USA.

State	Total
CA	52089
California	97
IL	129

Data verification

The sample was selected randomly from rows with complete critical data (i.e. business name, longitude, and latitude). Duplicate values for business names were excluded from the dataset prior to sampling. Data were shuffled to get random values from the dataset. Additionally, from column business location which includes full business location data (i.e. long. and lat. values), brackets were removed to enable me to fetch data from an external API. I planned to take 10% of the total data for the data verification but due to the technical issues with API call, I managed to sample 497 data (9% of total data with unique values)

Sample data were exported in Excel worksheet and then imported into Google sheets. API connector add-in, available in Google sheets was used to make an API request to position stack API. Positionstack has the option of sending batch API call by referencing column with full location data (long. and lat. values separated by comma). API URL to which GET request was sent is <http://api.positionstack.com/v1/forwar> ? access_key = <USER ACCESS KEY> & query=+++Master_sheet!I2:I498+++&limit=1. All data from Excel file were copied to Master_sheet and API results were stored in ValidationSheet. Returned API data shows that out of **497 data sampled from the original dataset, there are 39 (8%) values in business name column with identical long. and lat. values.** Based on the findings from sampled data I engaged in the analysis of duplicate values for a total of 5 423 unique business names. **Data shows that 370 unique business names (7%) share identical long. and lat. values.**

Analysis of sample data was also done in an Excel file (data.xlsx). Data retrieved from API call in Google sheet were copied in sheet validatedData and then using VLOOKUP in Excel file returned addresses were compared to addresses available in Master sheet to check if long. and lat. values are the same in both sheets. **Analysis of data showed that out of 497 data,**

474 (95%) addresses returned from API did not match with addresses available from the original dataset from the supplier. Considering that almost 95 percent of the data did not match I engaged in the manual analysis of the first 15 entries used in the API call to check their accuracy on google maps and the data returned from API were consistent with data returned by Google maps.

Final remarks and recommendations

Analysis of data as presented above identified many errors in terms of the number of missing values, data completeness, and data accuracy. Considering the findings from the analysis of a random sample collected from the original data set conclusion can be drawn that all data available in the data set are of questionable quality and require extensive cleaning before they can be used in navigator.ba app for displaying data to end-user.

The recommendation is not to engage in further negotiation with the supplier. Efforts should be made to identify another source of POIs.