Data Analysis Task Report

Elmedina Ćiber

17.01.2024

Summary

Dataset Overview:

Source: San Francisco Restaurant Inspection Scores
Sample Size: 52,315 rows

Objective

The primary objective of this analysis is to examine the San Francisco Restaurant Inspection Scores dataset and identify any elements of unstandardized or suspicious data, as well as analyze any underlying patterns that may exist.

Based on the analysis, the goal is to divide the dataset into three distinct quality-based subsets: high, medium, and low. Such division is essential for ensuring cost-effectiveness in the acquisition of the dataset records. It will enable informed decision-making regarding investment in high-quality records, while also providing leverage in negotiating prices for medium and lower quality records, thereby maintaining the integrity and value of the data presented on www.navigator.ba.

Key Findings

Business Information

1. Majority of entries contain full business details, including names and addresses.
2. Geographic coordinates are available for approximately 56% of the records.
3. Some records are missing crucial information such as business names.

Unstandardized Data and Suspicious Patterns

1. There are instances of unstandardized data, particularly in business names, addresses, city, inspection dates, and state names.

Data Classification

1. Dataset was classified into three quality-based subsets:
    a. Low Quality: Records missing crucial information like business name, precise address or coordinates
    b. Medium Quality: Records with some issues such as unstandardized phone numbers, addresses, zip codes or coordinates not matching addresses.
    c. High Quality: Records with no issues, where all data is standardized and complete.
2. Final classification resulted in 230 businesses in the low-quality subset, 5,173 in the medium-quality subset, and 670 in the high-quality subset.

Analysis

Preliminary Data Assessment

The dataset was observed using Google Sheets, and loaded into a Python project with Pandas library and QGIS for further data analysis. Initial inspection involved understanding various columns and data types.

It was essential to identify which columns are most relevant to our use case, which was needed for better understanding which ‘issues’ within the dataset are of concern.

The dataset consists of **52,315** entities, each representing an individual inspection for a business. A single business can have multiple inspection records, thus can be found multiple times in the dataset.

Data Quality Assessment

Initial task involved field validation, where completeness, frequency, duplicate records and dummy values were checked.

Completeness

The percentage of empty values was calculated for each column.

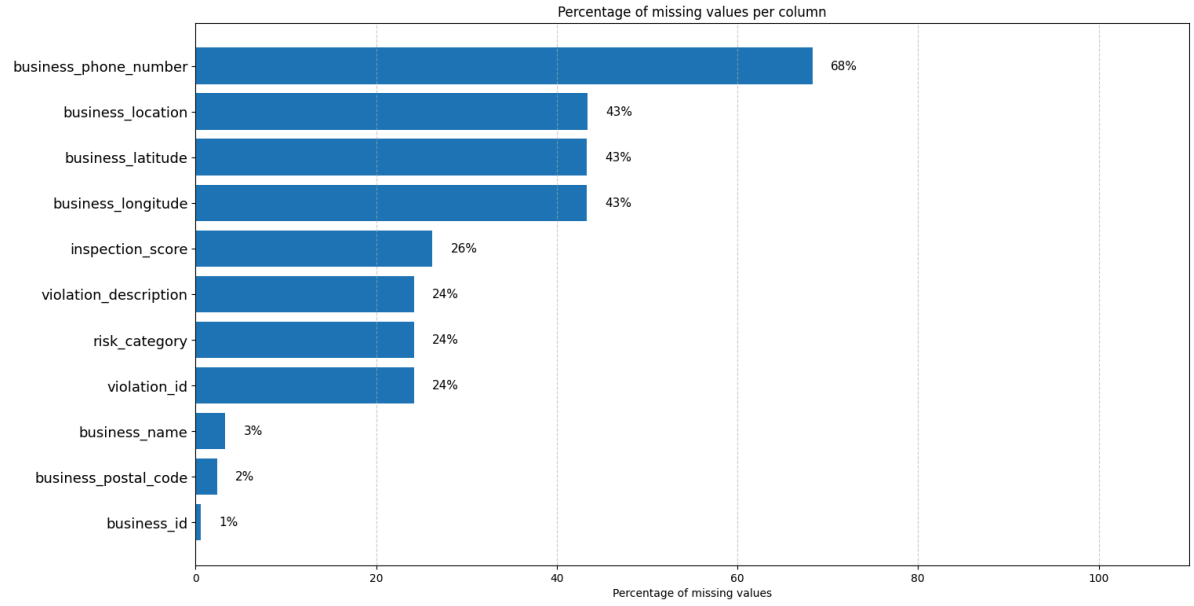

Figure 1

The columns omitted from the graph had 0% missing data. These columns include: **business_address**, **business_state**, **inspection_type**, **inspection_date**, **inspection_id**, and **business_city**.

However, the values in columns **business_name**, **business_latitude**, and **business_longitude**, which are essential for our purposes, are not always present. This information will be beneficial in the next tasks.

Frequency

Checking frequency of values helped to see the range of values within the columns, including the geographical spread of businesses, and it also helped to uncover minor issues in the data, which will be useful in the later tasks.

Frequency check wasn't performed for all columns since some of them were irrelevant such as the ID columns. Frequency analysis was performed for following six columns:

1. Business_city
2. Business_state
3. Inspection_score
4. Inspection_type
5. Violation_description
6. Risk_category

Business City

This column contained two values: San Francisco and SF. This indicated the unstandardized format of the city, and will be important for later.

Business State

There were three unique values in column business_state.

State value	Frequency
CA	99.57%

IL	0.25%
California	0.19%

Figure 2

This was suspicious and needed further investigation since the only city in the **business_city** column was San Francisco, and there is no San Francisco in Illinois.

Inspection Type

There were 15 types of inspections, where the top 5 were as follows:

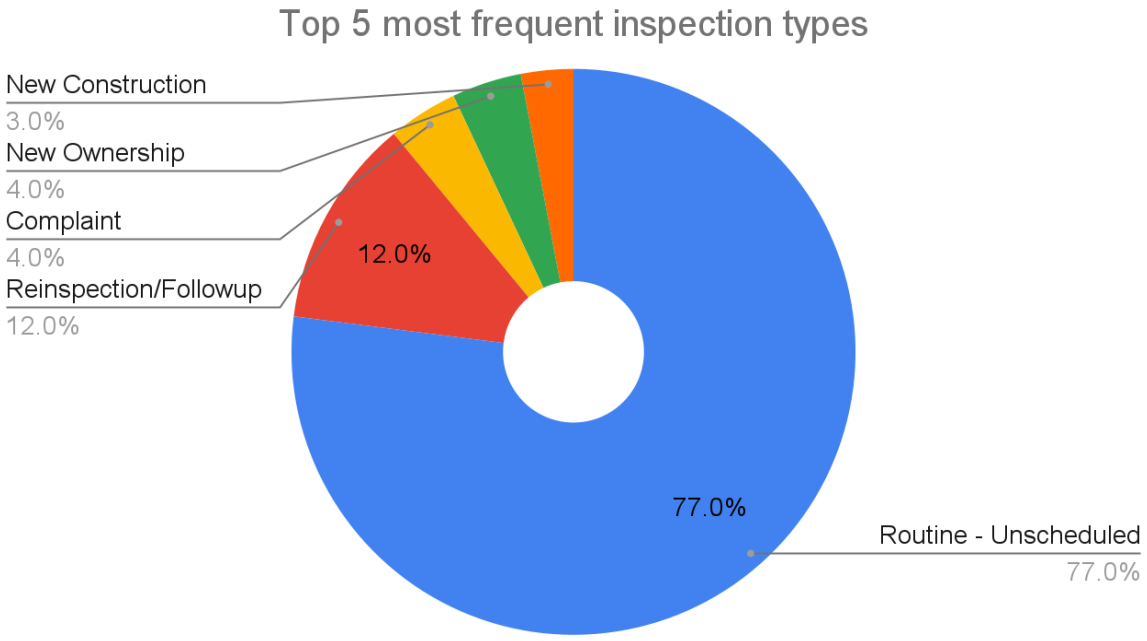

Figure 3

Violation Description

There were 67 different violation descriptions registered throughout the inspections.

Risk Category

This column had 3 values, not including empty values:

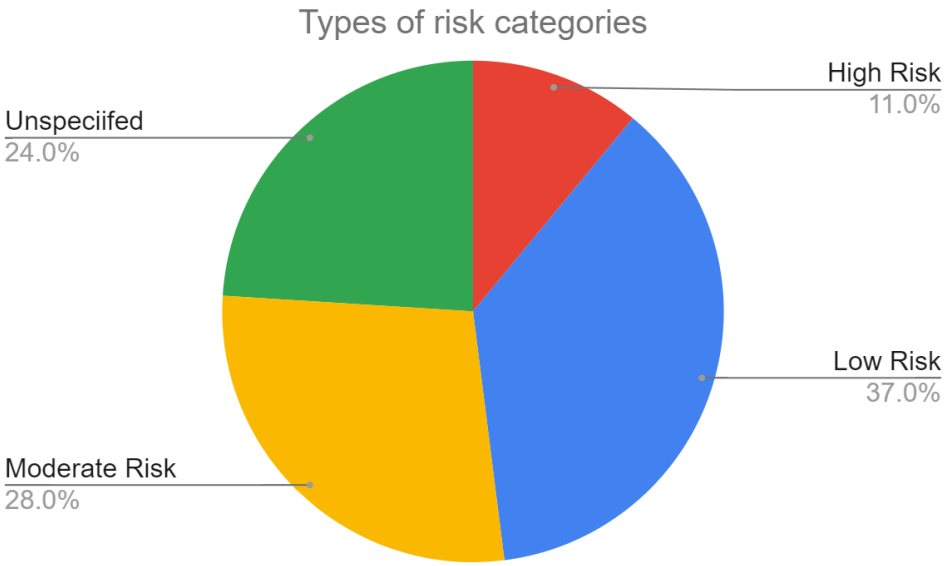

Figure 4

Very small number of businesses were in the high risk category, however there was a significant percentage of businesses that didn't have specified risk categories.

Beside this, the percentage of empty values in the **inspection_score** column that is visible in Figure 1 is 26%. For that reason, risk and inspection related columns were not taken into account in the later classification into quality subsets.

Furthermore, it was interesting to see what type of violations qualify the business as a high risk category. Figure 4 shows the top eight violations most commonly associated with high risk categories.

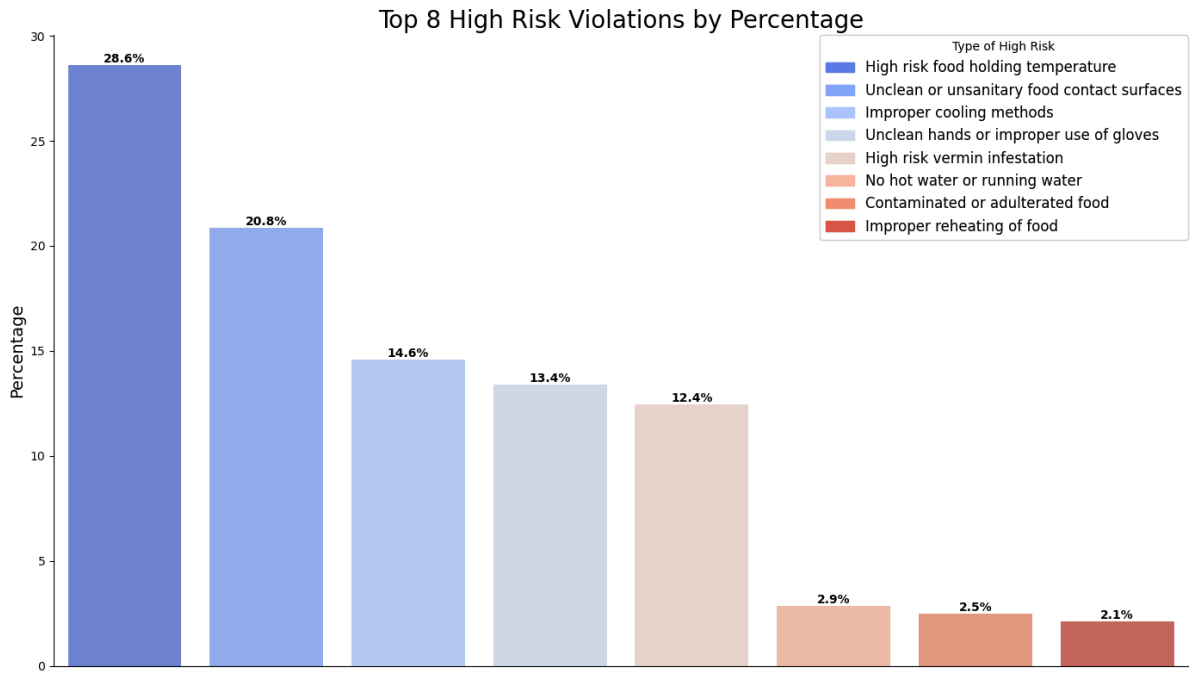

Figure 5

Duplicate records

In the dataset, the number of unique businesses was determined based on a combination of the **business_name** and **business_address** columns. This approach was necessary due to the presence of empty values in the **business_id** column. It was recognized that businesses with identical names could be registered at different addresses.
The analysis revealed that there are 6,073 unique businesses in the dataset, which is only 12% of total records.

Dummy values

Dummy values refer to values in the dataset that, while present, lack meaningful information.

Top 7 values in each column were checked. If there were any values used as dummy values, they would have appeared in the top 7, since dummy values would be used instead of blanks and have a larger count in comparison to some other columns.

Despite this approach, predicting dummy values is hard and requires manual checking. It was only later noticed that there are in fact dummy values in the dataset.

Dummy values were found in three columns:

1. Business_phone_number
2. Business_address
3. Business_name

Business Phone Number

This column included values such as *‘unde’, ‘null’, ‘not’, ‘na’,* and *‘NA’.*

Business Address

This column included values such as *‘Unnamed ST’, ‘Private Locations’, ‘Private & Public’, ‘Off The Grid’,* and *‘OTG’*.

This could present a potential problem for businesses that have these values as addresses, if those businesses do not have correct coordinates.

Business Name

This column had values such as ‘*NA*’, *‘hidden’, ‘Hidden’,* and *‘Unavailable’*.

Suspicious Patterns and Unstandardized Data

Following the initial data analysis, a search was conducted for suspicious patterns or unstandardized data.

Much of this phase relied on extending the analysis conducted earlier. Microsoft Excel was used to filter values in all columns, aiming to identify any irregularities within the dataset and to make an educated decision on how to separate the dataset into low, medium and high quality subsets.

The analysis revealed that several columns contained unstandardized data, as detailed in the following sections. The rest of the columns had standardized data.

Business Name

Some names are all uppercase, and some are written in title case.

Name	Address
Norman's Ice Cream and Freezes	2801 Leavenworth St
CHARLIE'S DELI CAFE	3202 FOLSOM St

Figure 6

Business Address

This column has multiple different types of unstandardized data:

1. Some values are all uppercase, some are title case
2. Some values do not have street number, therefore the address is not precise
3. Some values are PO BOX

business_name	business_address	business_city
BUS STOP PIZZA #1	**Divisadero St**	San Francisco
Garden of Eden	**340 PO BOX**	San Francisco

Figure 7

Business City

This column has values “*San Francisco*” and “*SF*” which means the same but it is unstandardized.

Business State

This column had values “*CA*”, “*IL*”, and “*California*”.

Business Postal Code

Most of the businesses have 5 digit postal code, however after filtering the others, there were several unstandardized values such as *‘94104-0291’, ‘941102019’, ‘941’, ‘941033148’, ‘CA’* and *’Ca’*.

Inspection Date

This column contains date values in both 12 and 24 hour formats, which is not standardized.

business_name	inspection_date
RHODA GOLDMAN PLAZA	**02/15/2017 12:00:00 AM**
CAFE X + O	**08/02/2016 00:00:00**

Figure 8

Business Latitude and Business Longitude

These two columns were analyzed using QGIS, in order to visually check the exact location where the businesses are located. Figure 9 shows green points representing a business from the dataset.

Figure 9

The points were distributed all over the US as well as some far off the coordinate system, which was very unusual. Since all of the addresses in the dataset were in San Francisco, the San Francisco shapefile was loaded into QGIS and checked which POI coordinates fall within those boundaries.

San Francisco city borders shapefile was not available on the internet, therefore San Francisco counties were used instead. (There was a San Francisco neighborhoods shapefile available, but it dated to 2004, so the county's file was a better fit). After loading both the POIs and San Francisco shapefile, it was clearly visible that there were many businesses outside of San Francisco, as can be seen in the figure 10. The green points are again businesses from the dataset, and the little red area in the left part of the picture is the San Francisco area.

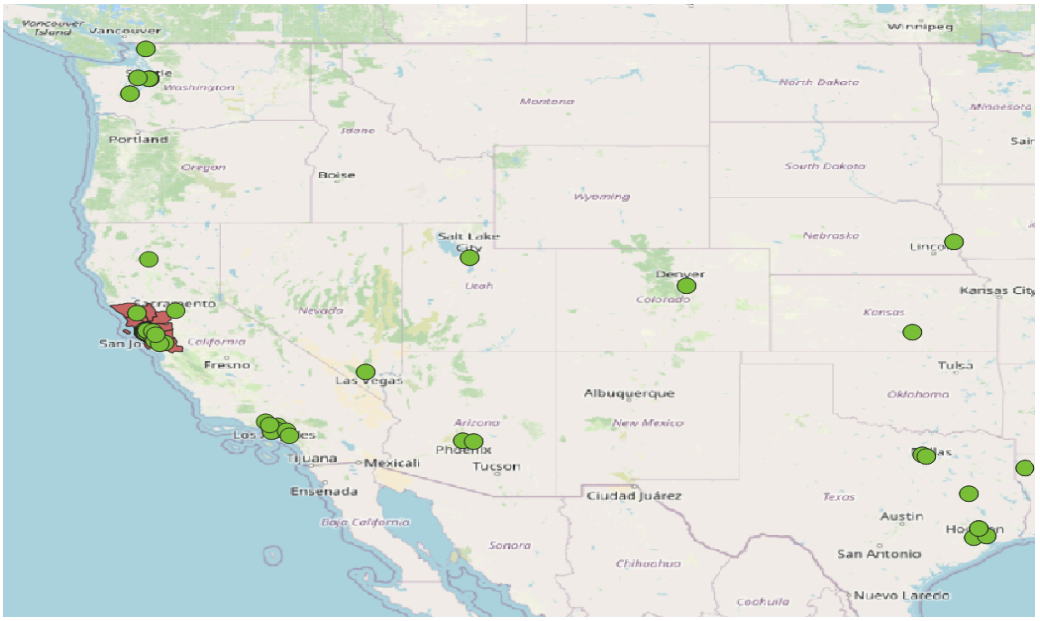

Figure 10

There were many businesses whose addresses were in San Francisco, but their coordinates pointed well outside San Francisco. A sample of these records was cross referenced with Google Maps, and the investigation showed that the businesses were in fact located at the provided address in San Francisco, but their coordinates were wrong.

QGIS functions were used in order to preserve POIs with correct coordinates. After detecting points that belong inside the borders of san francisco (points that fall within the red area), a freehanded select tool was used to separate POIs with San Francisco coordinates, then they were saved to another layer, and exported in a separate CSV file.

Business Phone Number

Values in this column usually were in the format 1415, where 1 is the phone code for the US, and 415 is the phone code for San Francisco.

However, there were other values that had + in front, or more or less digits than what would correspond to the San Francisco phone number format. Some of them even had a website url instead of a phone number.

Figure 11 shows various unstandardized phone numbers. The row shows the correct phone number, and the rest of them are wrong or suspicious.

Name	Phone Number
RHODA GOLDMAN PLAZA	**14155345060**
J & M A-1 CAFE RESTAURANT LLC	**unde**
NORDSTROM CAFE BISTRO	**9999999999**
PACIFIC PLAZA CAFE	**+82-64-732 5222**
Peralta	**5555555555**
Beach Street Grill	**51105**
Farm Fresh Underground	**253353597**
Yummy Sticks	**https://www.sspaonline.com**

Figure 11

Data Classification

In this final step of our analysis, previous insights about the dataset were used and an algorithm was developed using Python to create high, medium and low quality subsets, each of which was exported as a CSV for further data analysis and algorithm optimisation.

Preparation

First step was filtering out duplicate records, which has left us with 6,073 unique businesses. Each duplicate record had the same values for columns, therefore there was no need to check each duplicate record if it should be prioritized over the other if it had more data.

Risk categories, inspections and violations were not taken into consideration because they are not a crucial component of the objective. It may be argued that they be included in the decision, since only the highest quality data is wanted, and that means restaurants which are essentially low risk with best inspection score. However, that risk category is temporary and can be improved or worsened in subsequent inspections.

Criteria for Data Classification

Based on the previous analysis and knowing what data is crucial, it was easiest to classify data from low to high quality based on a certain criteria.

Low Quality Criteria

Low quality data is any record which is missing **business_name** value, vague **business_address,** and missing **business_latitue** or **business_longitude** columns**.**

If a business is missing both address and coordinates, it cannot be placed on a map.

A vague address, besides dummy values, is any address which does not have a street number or combination of streets to determine that the business resides at the intersection. Some examples include: *'Divisadero St', 'B Hayes St', 'Union St', 'Vicente St', 'Close to Park', 'Close to Golden Gate bridge', 'Unnamed ST', '377', 'Golden Gate Park', 'Fort Mason', 'Off The Grid-Upper Haight', 'Ocean Ave', 'Public', 'Off The Grid', 'Treasure Island Flea Market', 'Soma Street Food Park', 'Various Farmers Markets', 'Treasure Island', 'Approved Private Locations', 'Private & Public', 'Cortland Ave', 'Approved Locations', 'Approved Private Locations & Special Events', 'TFF Event Operations', 'Private Locations', 'OTG', 'Treasure Fest'*.

These addresses were found through filtering the data using Microsoft Excel.

Wrong coordinates were detected through QGIS, by plotting the data points along with the boundary of San Francisco and observing any data points which are out of bounds. All of those data points had a latitude of 37 and longitude of -122.

Figure 12 shows two examples of low quality data, where the first one is missing a name, and the second one is missing coordinates plus it has vague address.

business_name	business_address	business_latitude	business_longitude
	925 Larkin St		
J-Shack	**Off The Grid**		

Figure 12

Medium Quality Criteria

After filtering low quality data, a subset containing medium and high quality data is left. Next step was to extract medium quality data.

Medium subset data had some issues, some of which include:

1. Coordinates outside San Francisco or belonging to Illinois state
    a. Manual check was performed using Google maps to verify whether business is located on the address, which proved that the coordinates were incorrect
2. Missing or unstandardized phone number
    a. Phone number is usually optional data
3. Unstandardized address
    a. If the coordinates are available, geocoding can be used to retrieve the address, and vice-versa
4. Unstandardized zip code

Figure 13 shows two records from the medium subset. First record is missing a phone number. The second record has a vague address, however it has coordinates so its address can be retrieved using external resources.

business_name	business_address	business_state	business_latitude	business_longitude	business_phone_number
KISS SEAFOOD	1708 LAGUNA St	**IL**	37.786838	-122.43	
HAI SUN RESTAURANT	**Close to Park**	CA	**37.800133**	**-122.43**	

Figure 13

High Quality Criteria

The remaining high quality data must have no issues. All the data must be standardized and there must be no empty fields or invalid values for important columns.

Figure 14 shows two records from the high quality subset. They both have names, precise address, coordinates inside San Francisco, as well as phone number.

business_name	business_address	business_state	business_latitude	business_longitude	business_phone_number
DEJA VU PIZZA & PASTA	16th St	CA	37.76	-122.42	14155251600
CAFE X + O	1799 Church St	CA	37.74	-122.43	14155823535

Figure 14

Classification Summary

There were 230 businesses in the low quality subset, 5173 in medium subset, and 670 in high quality subset.

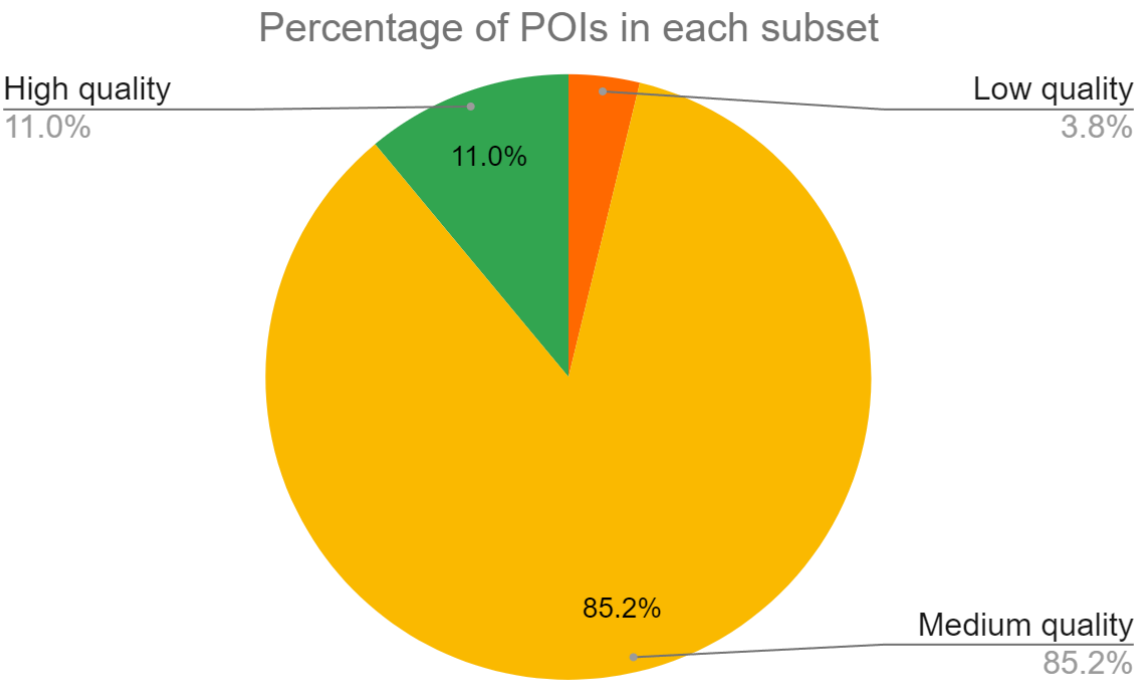

Figure 15