Business Data Quality

ABH Data Analyst Tasks

Name: Ivana Kolorici-Livnjak

Date: 07.08.2022

To Buy or not to Buy?

As an owner of www.navigator.ba I can relate to what author Cassie Graham said: "There isn't any one big test or way to validate ourselves in the world. There's just a long, quiet process of finding our place in it". This is in a way what we do. We know that one test isn't enough, and even though the process of validating and understanding data isn't easy, we know that doing that can result in you finding your special place.

That being said, my wish is to expand our services outside Bosnia and Herzegovina and I came across a POI (point of interest) dataset that I'm interested in purchasing. Spending money for clean data is nothing in comparison to paying the price of working and fixing bad decisions made due to bad data. I am willing to pay the price of 20$ per record for the highest quality data, but for the lower quality one, not so much.

Detailed Steps of the Analysis:

In order to assess the quality of the given dataset, I will evaluate it against 6 dimensions:

1. Completeness, meaning that the main features don't have missing values
2. Uniqueness, meaning there are no unnecessary duplicates
3. Timeliness, meaning the data is up to date
4. Consistency, meaning the data is presented in the same form
5. Validity, meaning that the data is of the right type, format and range
6. Accuracy, meaning correct/accurate

Data analysis will be performed using SQL (BigQuery), R (Kaggle), Excel, and Python (Jupyter Notebook).

Importing data:

As I felt like combining different tools, the source data which is in a CSV format was loaded into a BigQuery (SQL) table, Kaggle(R), and Jupyter Notebook (Python).

*The code, scripts, SQL queries, and Excel sheets are all provided in a separate file.

General information about the dataset

business_id	business_name	business_address	business_city	business_state	business_postal_code	business_latitude	business_longitude	business_location
<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<chr>
19	Nrgize Lifestyle Cafe	1200 Van Ness Ave, 3rd Floor	SF	CA	94109	37.78685	-122.4215	(37.786848, -122.421547)
19	Nrgize Lifestyle Cafe	1200 Van Ness Ave, 3rd Floor	SF	CA	94109	37.78685	-122.4215	(37.786848, -122.421547)
19	Nrgize Lifestyle Cafe	1200 Van Ness Ave, 3rd Floor	SF	CA	94109	37.78685	-122.4215	(37.786848, -122.421547)
19	Nrgize Lifestyle Cafe	1200 Van Ness Ave, 3rd Floor	SF	CA	94109	37.78685	-122.4215	(37.786848, -122.421547)
19	Nrgize Lifestyle Cafe	1200 Van Ness Ave, 3rd Floor	SF	CA	94109	37.78685	-122.4215	(37.786848, -122.421547)
19	Nrgize Lifestyle Cafe	1200 Van Ness Ave, 3rd Floor	SF	CA	94109	37.78685	-122.4215	(37.786848, -122.421547)

business_phone_number	inspection_id	inspection_date	inspection_score	inspection_type	violation_id	violation_description	risk_category
<chr>	<chr>	<chr>	<dbl>	<chr>	<chr>	<chr>	<chr>
NA	19_20171211	12/11/2017 12:00:00 AM	94	Routine - Unscheduled	19_20171211_103144	Unapproved or unmaintained equipment or utensils	Low Risk
NA	19_20171211	12/11/2017 12:00:00 AM	94	Routine - Unscheduled	19_20171211_103116	Inadequate food safety knowledge or lack of certified food safety manager	Moderate Risk
NA	19_20160513	05/13/2016 12:00:00 AM	94	Routine - Unscheduled	19_20160513_103144	Unapproved or unmaintained equipment or utensils	Low Risk
NA	19_20160513	05/13/2016 12:00:00 AM	94	Routine - Unscheduled	19_20160513_103157	Food safety certificate or food handler card not available	Low Risk
NA	19_20160513	05/13/2016 12:00:00 AM	94	Routine - Unscheduled	19_20160513_103154	Unclean or degraded floors walls or ceilings	Low Risk
NA	19_20180607	06/07/2018 12:00:00 AM	96	Routine - Unscheduled	19_20180607_103116	Inadequate food safety knowledge or lack of certified food safety manager	Moderate Risk

Table #1: Quick visual of what the columns contain

Understanding the data:

business_id - A unique id for each business.

business_name - Name of the business (hotel, restaurant…) of interest.

business_address - Indicates the address of the business.

business_city- City where the business is located.

business_state - State where the business is located.

business_postal_code - An expected form of 5 numbers.

business_latitude - The measurement of distance north or south of the Equator.

business_longitude- The measurement east or west of the prime meridian.

business_location - A combination of the latitude and longitude and the format of (latitude, longitude).

business_phone_number - Phone number.

inspection_id - id is in the format of businessid_date(yyyymmdd).

inspection_date - date of the inspection conducted. Format: mm/dd/yyyy, where mm is the month, dd is the day and yyyy is the year.

inspection_score - integer with the highest possible value of 100 (best)

inspection_type - Explains what was inspected.

violation_id - inspectionid_violationdescription(Code).

violation_description - explanation of the inspected violation of the protocol .

risk_category - low, moderate, and high risk.

Data Field Validation

	Null	Invalid	Inconsistent	Unique
business_id		Those that have letters.		Uniqueness not relevant as we can have different violation id which defines each row as unique and therefore, we don' view them as unnecessary duplicates.
business_name		Those that have #, "floor",@,_,/ ,),(, symbols in their name which we don't expect. Next to that, inputs as NA,0, none...were inspected.	I managed to come across some cases where capital letters were used, " ' " was missing	
business_address				
business_city		" business_city " was the only invalid variable.	There was a mix of "San Francisco" and "SF"	1 San Francisco is in California but there was also Illinois in the dataset. By checking the location, I came to the conclusion that all recurrences of "IL" data is invalid.
business_state		"IL" results were interpreted as invalid and " business_state".	There was a mix of "California" and "CA" inputs.	2
business_postal_code		I have searched for postal codes specific for SF and selected those as valid.		
business_latitude		Those that don't match the coordinates of the SF location.		
business_longitude		Those that don't match the coordinates of the SF location or don't have a "-" or ",".	Those that don't have the same number of figures or are missing a bracket or a space.	
business_location				
business_phone_number		Numbers not starting with any of the San Francisco Local Area Codes, are too long or too short (together with No, na, null, NULL, unde, websites...)	None of the date satisfied one of the two standard formats which is having 10 numbers or 11 with a + in front.	
inspection_id		Fined, available, Available and similar values were detected.		
inspection_date			Data wasn't in a YYYY starting format.	
inspection_score		All values were in a 0-100 range with one invalid result "inspection_score"		
inspection_type				
violation_id				
violation_description				
risk_category				

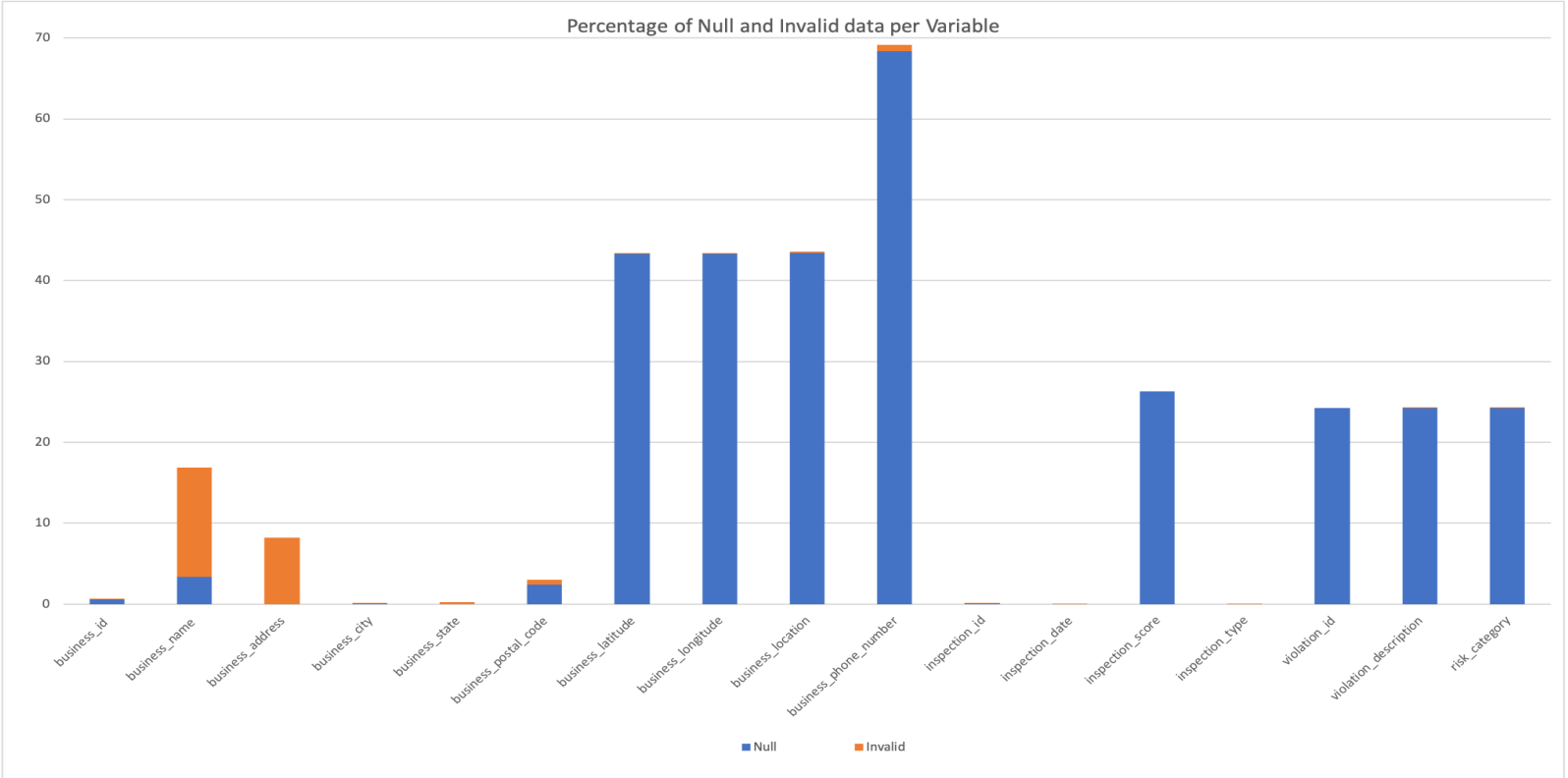
*Table #3: Data Field Validation

Data Quality Summary

	Number	%					
Rows	52315,00	100,00					
Duplicate rows	47,00	0,09					
Incoplete	47059,00	89,95					
	Null number	Null %	Invalid	Invalid %	Inconsistant	Inconsistant %	Unique
business_id	297,00	0,57	8,00	0,02			
business_name	1732,00	3,31	7099,00	13,57			
business_address	0,00	0,00	4267,00	8,16			
business_city	31,00	0,06	1,00	0,00	135,00	0,26	1,00
business_state	0,00	0,00	130,00	0,25	97,00	0,19	2,00
business_postal_code	1246,00	2,38	326,00	0,62			
business_latitude	22674,00	43,34	37,00	0,07	1276,00	2,44	
business_longitude	22674,00	43,34	43,00	0,08	2368,00	4,53	
business_location	22725,00	43,44	79,00	0,15	1,00	0,00	
business_phone_number	35762,00	68,36	416,00	0,80			1738,00
inspection_id	29,00	0,06	39,00	0,07			25107,00
inspection_date	0,00	0,00	1,00	0,00	52315,00	100,00	2015-2018
inspection_score	13725,00	26,24	1,00	0,00	0,00	0,00	45-100
inspection_type	0,00	0,00	1,00	0,00	0,00	0,00	15,00
violation_id	12669,00	24,22	0,00	0,00		0,00	39601,00
violation_description	12669,00	24,22	1,00	0,00		0,00	67,00
risk_category	12669,00	24,22	1,00	0,00	0,00	0,00	3,00

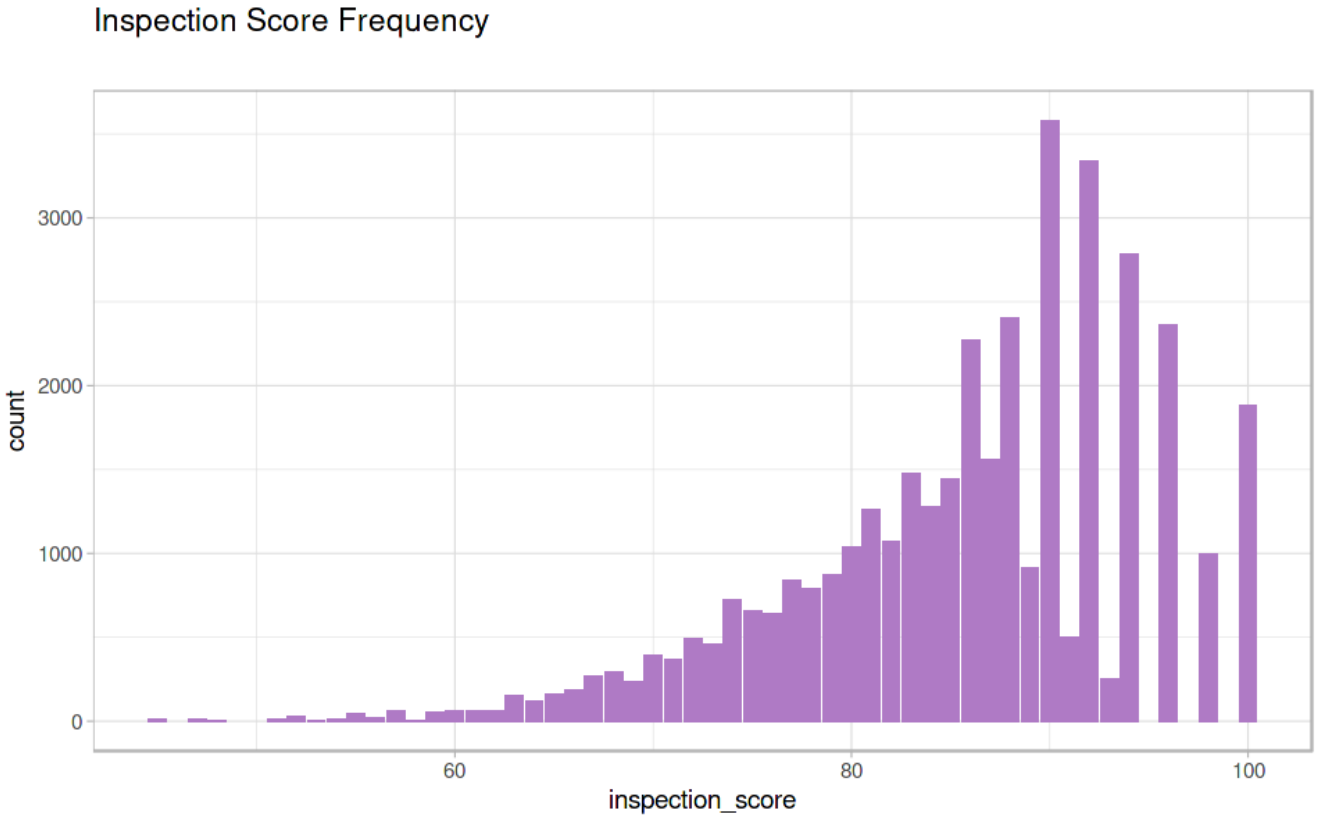
* Tabel #2: Number of null, invalid, Inconsistent, and unique data present for each column

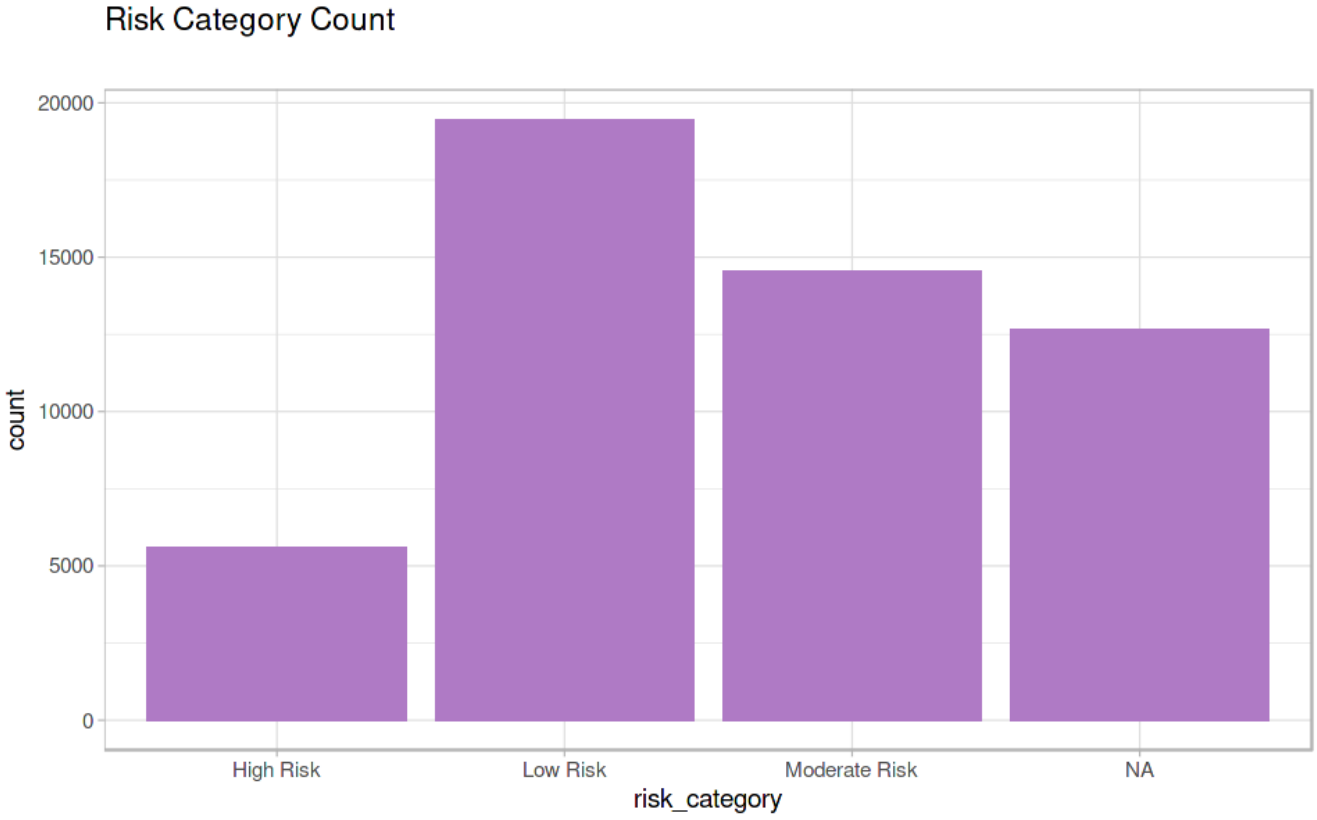
Data Quality Visualization



*Graph #1: Visualization of Table #2 data

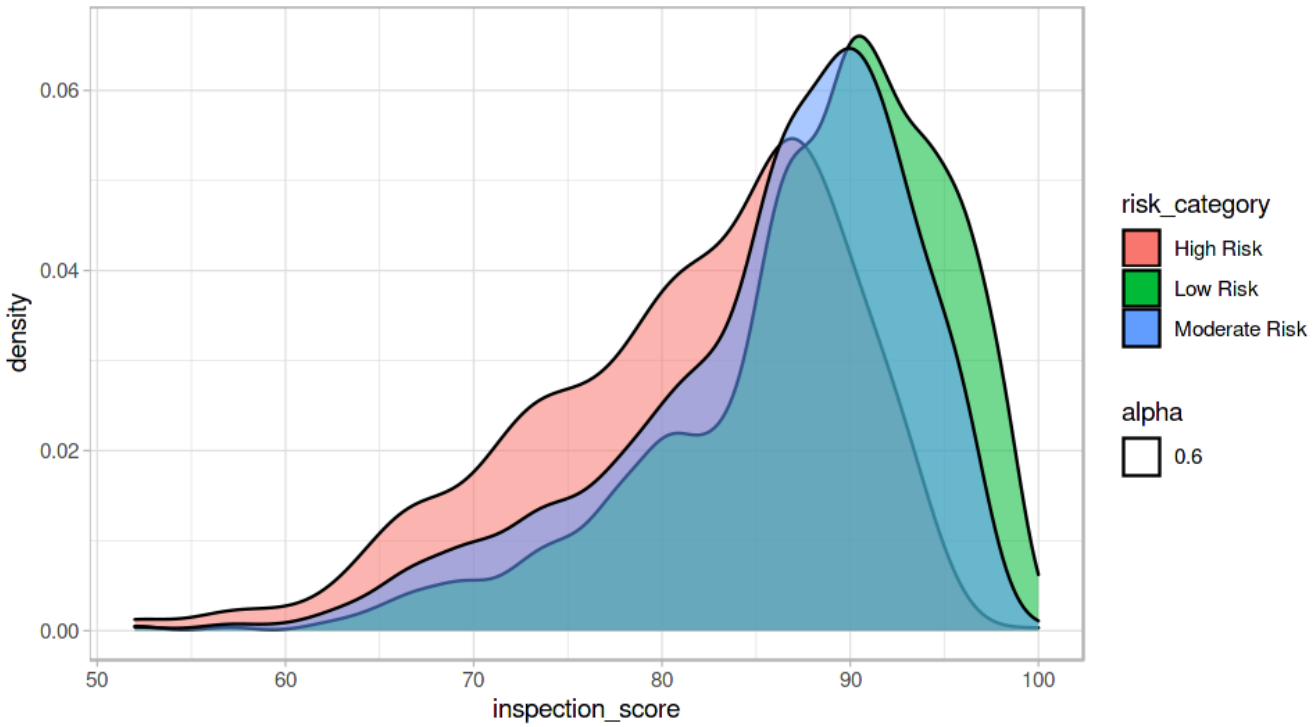
For the inspection_score, I wanted to see if the data shows validity in terms of being correlated to the risk_category as a higher score should indicate lower risk. This was performed in Kaggle (R).





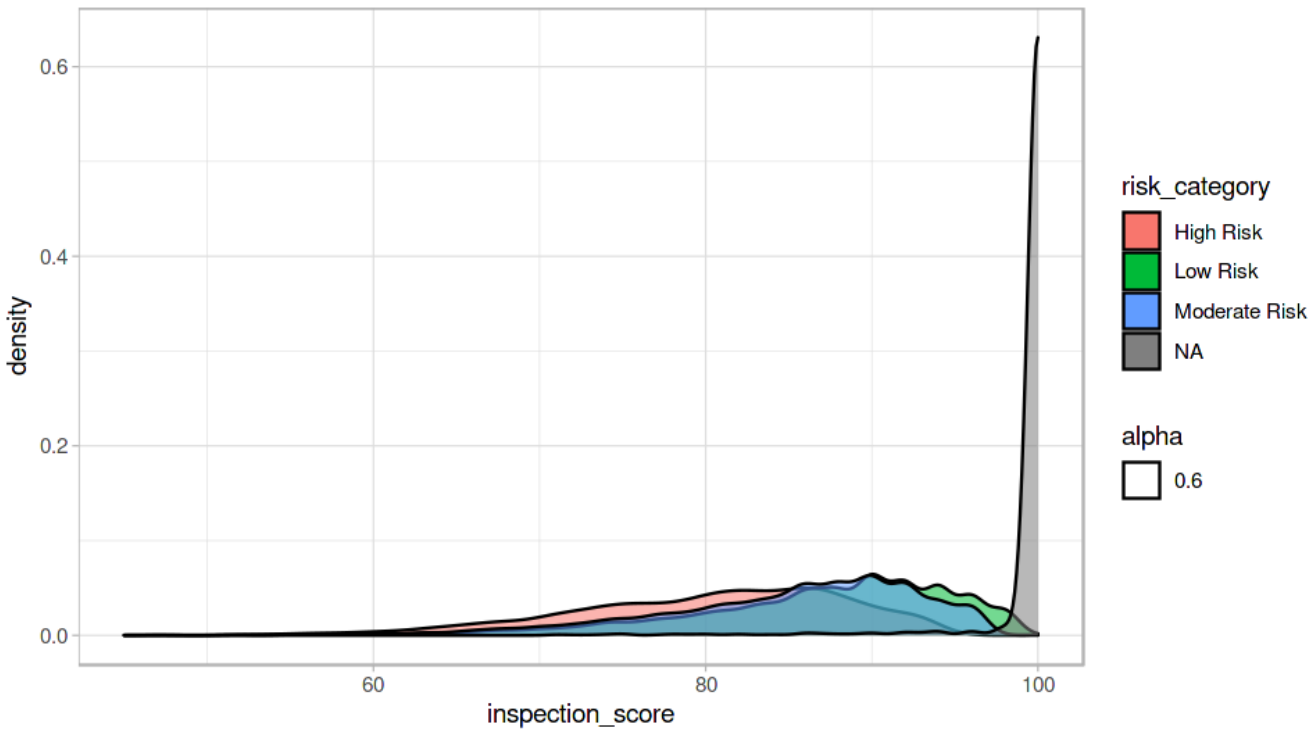
Inspection Score Distribution for Three Risk Categories

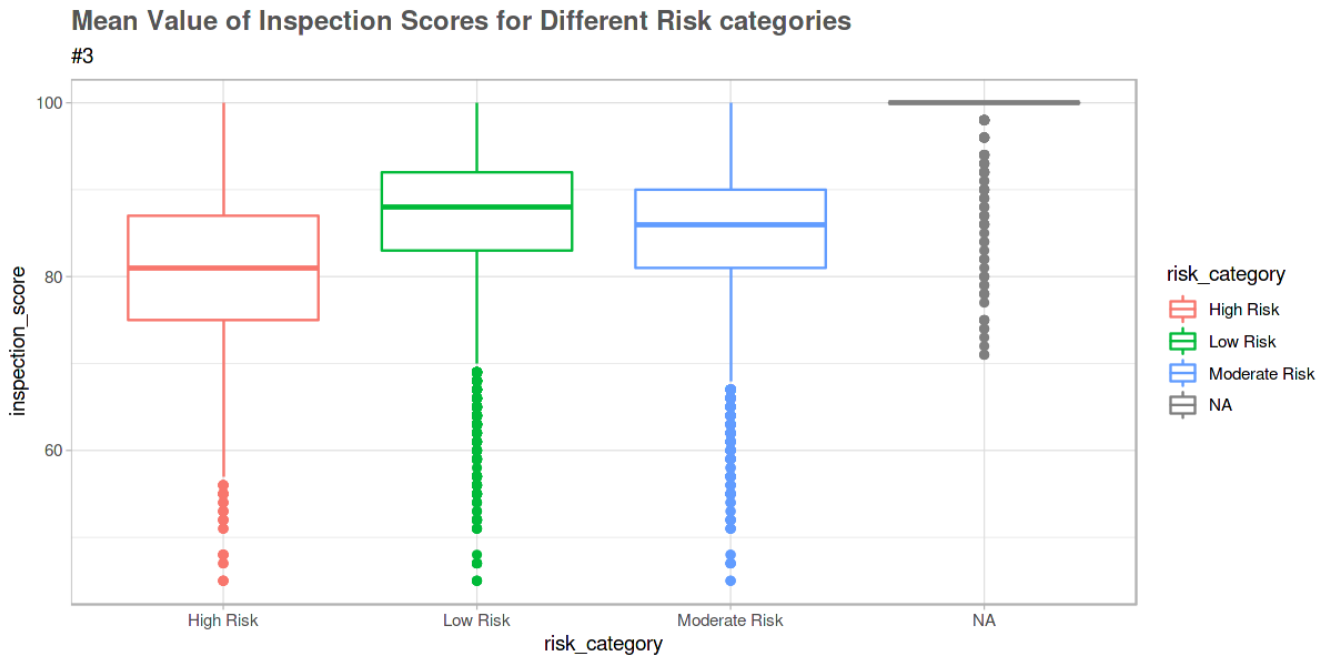
#1



Inspection Score Distribution for Different Risk Categories

#2





- **The low-risk category is the highest part of the inspection score, which makes sense.**
- **There is a significant amount of data that occupies a score region of around 100 whose risk_category isn't known.**

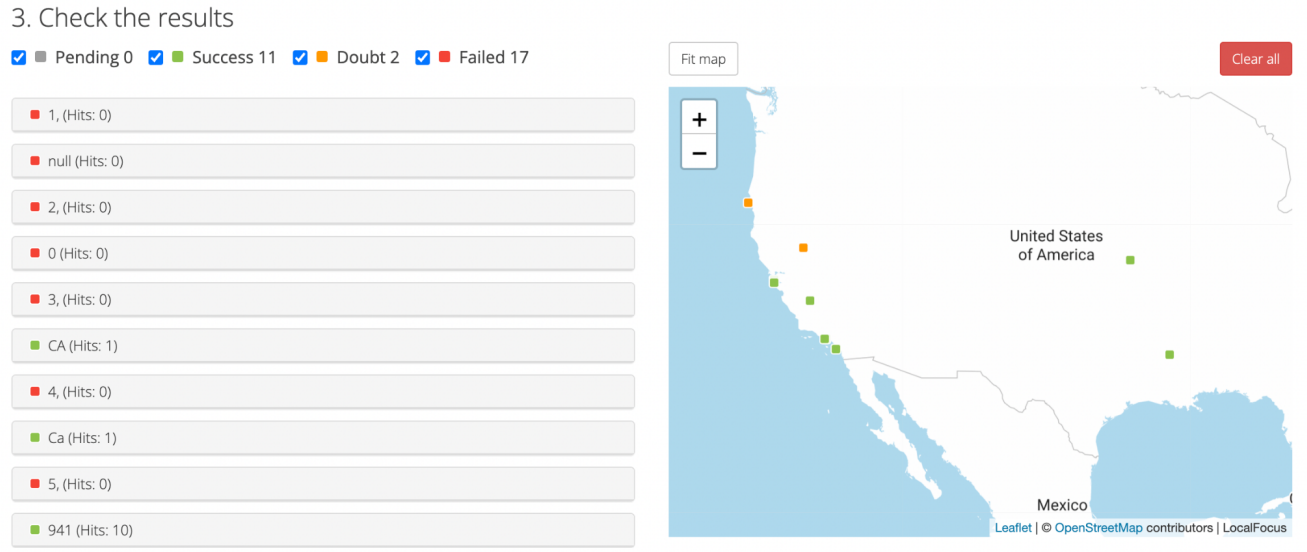
Methods of data verification

External verifications were conducted for:

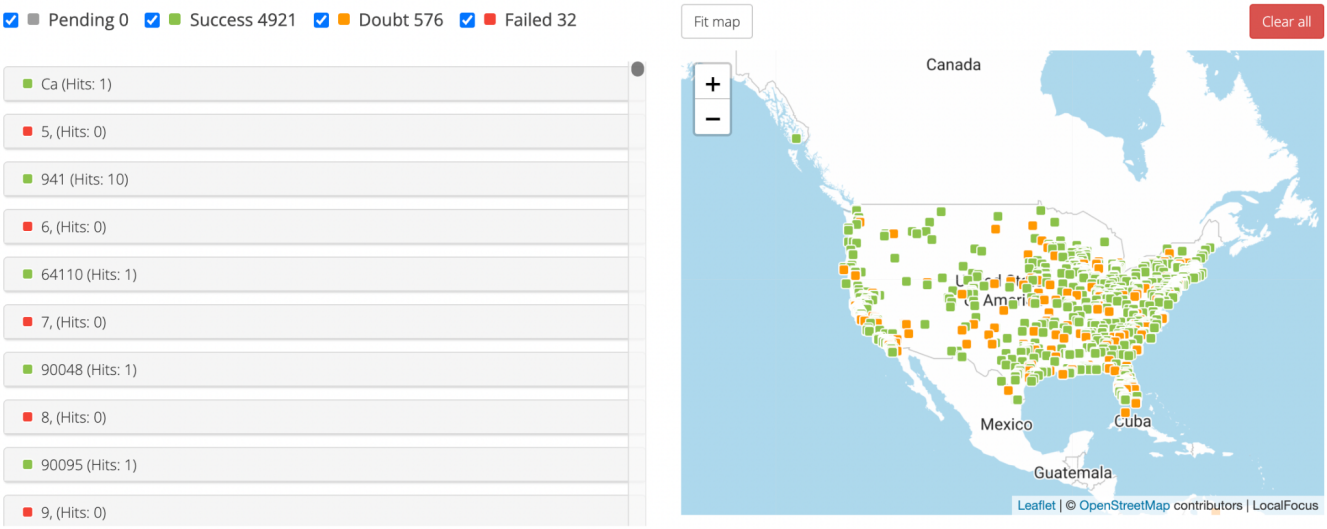
The business_phone_number:
Phone numbers have specific country and area codes that can help in distincting valid from invalid phone numbers. I used this website: https://grasshopper.com/numbers/local-numbers/san-francisco-phone-numbers/ .

The business_postal_code:
1st way:
Checking this website https://www.usmapguide.com/california/san-francisco-zip-code-map/ , allowed me to find ZIP codes specific to my area of interest (San Francisco)

2nd way:
External validation can also be performed by inputting values in https://geocode.localfocus.nl/ and seeing their location on the map to get an idea of how many misplaced zip codes and business addresses are there. I checked the sample of 11(#2) and 4921(#3) distinct data rows for ZIP and address correctness respectively. In addition, the results are provided with an address longitude and latitude value which is great.



*** Picture #2 : Verifying the business location by using distinct ZIP codes**



***Picture #3: Verifying the business location by using distinct business addresses**

By looking at the locations provided on the map, I do see a significant number of points in Illinois and as those results did appear in the data with unmatching coordinates I would have to look at it again later to see what is the main mistake in the row.

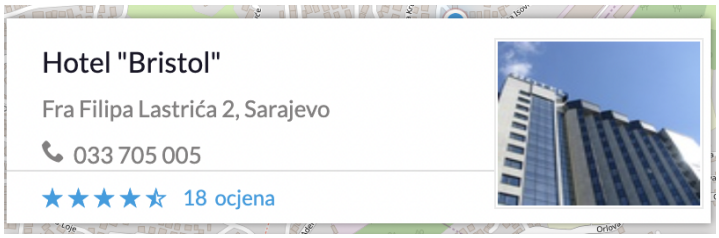
Internal verifications could be conducted for:

- **The variables business_id, inspection_date are combined into the inspection_id.**
- **The variable inspection_id and violation_description in a form of a code, create the violation_id variable.**
- **The business_latitude and the business_longitude combine into the business_location variable.**

Cross matching those values could validate them.

Quality of the data

I defined the data quality based on the existence of data in selected variables. The selection of those variables was based on the main data provided on the www.navigator.ba website.



* Picture #1: Screenshot from the site www.navigator.ba website

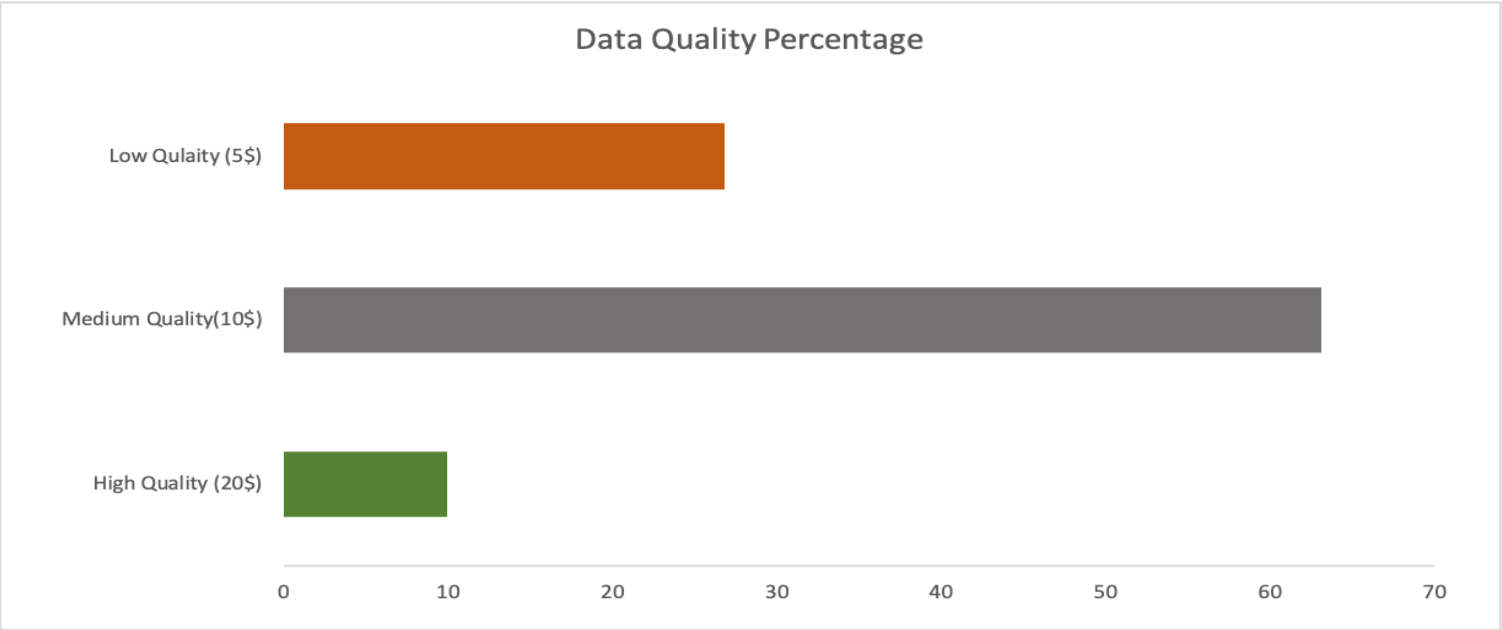
Therefore the variables playing the biggest role in determining the data quality are business_name, business_location, business_phone_number, inspection_score(for the 5-star rating), and violation_id (for validation). Variables are chosen in a way that they all have the same weight in terms of their contribution to the score. Other variables can be derived from the mentioned ones.

Quality Summary

Data Qulaity	Data Count	Quality Data %	Business Name	Business Location	Business Phone	Inspection Score	Violation Id	Price($)
High Quality (20$)	5213,00	9,96	Not Null	Not Null	Not Null	Not Null	Not Null	104260,00
Medium Quality(10$)	33008,00	63,09						330080,00
Low Qulaity (5$)	14046,00	26,85		Null	null			70230,00
							Tota price:	504570,00

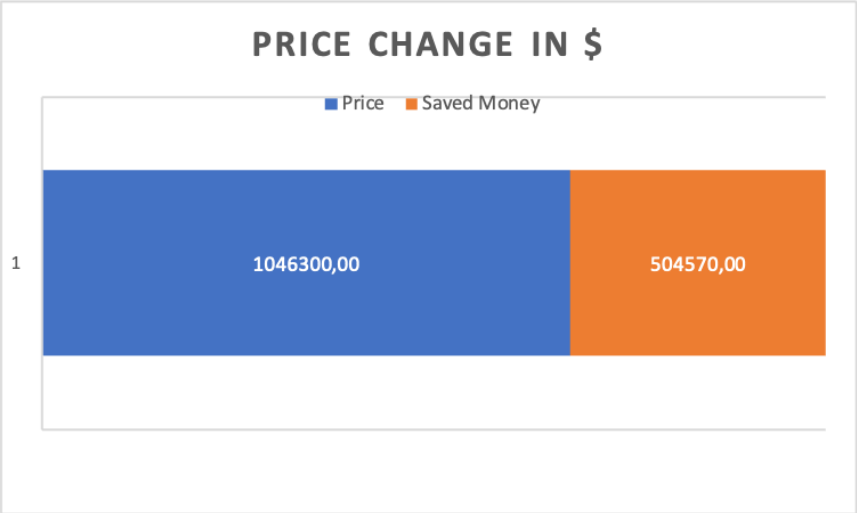
Total Number of Data	Price($)
52315,00	1046300,00
Duplicates,invalid rows	
48	

*** Table #3: Cost Summary**



***Graph #2: Count of the data by Quality (Table #3 Visualization)**

We see that only 10% is high-quality data, whereas 27% is low-quality data. Medium-quality data can still be useful but requires additional time and money to clean, therefore it isn't worth 20$ per record (row) but somewhere around 10$. I estimated the low-quality data to go for 5$.



Graph #3: Amount of money that could be saved (orange) by paying different prices for different quality data.

- **I will improve my price estimates by looking into the prices for the POI data in 2022.**

Conclusion

I started this data investigation with a wish to buy new data and expand the world seen through my website (www.navigator.ba). Going through the data and looking at 5 out of 6 data quality assessments, I gained some deeper insight into what this data can offer.

1. Most of the data (63%) are medium quality as they have some null values in columns specified as important for the quality assessment. 10% of the data is high-quality and 20% is characterized as low-quality.

2. Different data validation methods showed that there are mismatching pieces of information in rows that have to be examined further.

3. The phone number which was taken as a variable of importance to distinct data by their quality had almost 70% of null values.

4. By recognizing the quality of data to purchase, we could save up to 48% of our money.

5. The one data quality assessment wasn't mentioned and that is timeliness. As the inspection_date column suggests, the inspection was conducted in a period from 2015 to 2018 meaning all the other data could be 4 years out of date, making this data not suitable for accomplishing my business goal.