****

*AtlantBH Interview*

Data Analytics Tasks

*Candidate: Kanita Krdžalić-Korić*

*SARAJEVO, APRIL 2019.*

# Task 1

## Summary

The aim of the first task was to perform analysis of dataset and categorize data into high, medium and low-quality data. It is in the interest of the owner of [www.navigator.ba](http://www.navigator.ba) website to identify all high quality data because it should be paid 20$ per record and to detect all low quality data since the owner would not like to buy bad records for the same price. In order to do so, analysis and categorization of inspection dataset was performed.

Since the purpose of buying this data was to expand the location coverage for [www.navigator.ba](http://www.navigator.ba), some attributes from inspection dataset were not useful for the categorization. An owner would be interested only in data that could be useful on map so fields like inspection\_score or violation\_description are not important at this point. That is the reason why significant and richness attributes were detected and analyzed.

At the beginning of analysis, it was concluded there are businesses inspected multiple times which means there are duplicated data with respect to significant and richness fields.

IMPORTANT NOTE:

The owner would not like to pay multiple times for the record that seems the same (inspection details are different but address, name, location etc. are the same) because it represents the same POI for the website. That is why only distinct records with respect to significant and richness fields are taken into consideration for further analysis, meaning all inspection details were neglected at this point.

The criteria for data categorization were standardization, validity and completeness.

The results of the first task can be represented in chart in Figure 1.

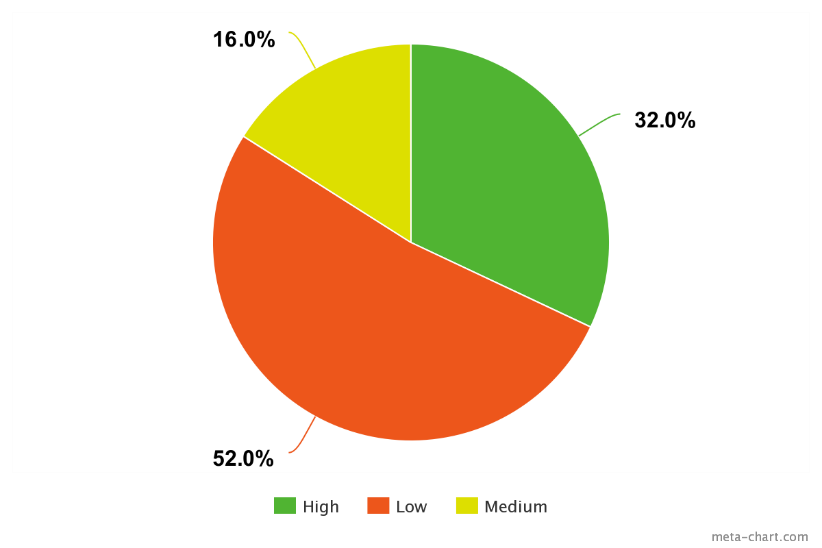


Figure 1. Results of data categorization

## Analysis Explanation

The analysis was consisted of field validation, completeness and standardization checks and cross-validation. At the beginning, the dataset was imported into PostgreSQL database. Only significant and richness fields are chosen for the analysis in this task. Those fields are following:

* business\_id
* name
* Address
* latitude
* longitude
* city
* postal\_code
* state
* phone\_number

Even though there were data that have the same values for those fields but different for other fields, they were considered duplicates because other details about inspection were not important for location coverage. From the initial dataset, only distinct records were extracted and it resulted in creation of new table that has 6080 records which is around 11% of initial dataset.

The criteria for categorization were:

* Low quality data: all records that have missing value in one of the significant attributes and all records that are invalid
* Medium quality data: all records that have one of the significant values unstandardized and all records that do not include phone number as richness field
* High quality data: all other data that is not in low or medium bucket

Analysis of significant fields

Name

Completeness for this field is 100%. Since it is hard to decide if some name is invalid (one can give whatever as name), only standardization is checked.

All names with only uppercase letters are detected. Some names included character “@” which might be an explanation that certain restaurant belongs to some larger building. It was standardization issue because we are interested only in POI itself.

There are examples of unstandardized data:

* Name is only number: “903” (There were less than 1% such data)
* All uppercase letters: "CHARLIE'S DELI CAFE", "ART'S CAFE" (10%)
* Names that included “@”: "Bon Appetit @ Twitter", "Bon Appetit Management Company @ Yahoo SF" (1%)
* Same names written in different way: “MCDONALD'S #7413", "McDonald's" (<1%)

Address

The completeness for the address field was 100% as well. There were some invalid values and number of unstandardized addresses. Invalid records were those that included dummy value or had only number, without street.

Examples of invalid data are:

* Addresses with dummy values: : “OTD”, “OFF THE GRID”, “Approved Locations” (1%)
* Only number included: “377” (<1%)

Unstandardized data are records that have all uppercase and those that do not have house number in address.

The examples of such data are:

* All uppercase: "15 MARINA BL", "255 MENDELL STREET" (2%)
* Address does not include house number: "Foot of Baker Street St", "Golden Gate Park" (2%)

Latitude and longitude

Completeness of those fields is 49%. There were many missing values for those fields, yet they are very important for an owner of navigator.ba, probably the most important because of visualization and geometry.

There was geometry field created out of those two fields so that visualization is easier.

The range of latitude (-90:90) and longitude (-180:180) was checked and all records were within the proper range. There were few records that have location (0,0) which is by cross-validation proved wrong since it is not location of San Francisco.