Preliminary EDA and data quality analysis

Preliminary EDA

Preliminary EDA is an initial process in data analysis that aims to explore and understand the data before applying more advanced techniques. During this stage, the raw data is examined to identify patterns, trends, distributions, relationships, and possible outliers or errors. Some of the common techniques used in preliminary EDA are data visualization, descriptive statistical calculations, graphs, histograms, scatterplots, and correlation analysis. The preliminary EDA helps data scientists gain an initial understanding of the data and make informed decisions about subsequent steps in the analysis.

Data quality analysis

Data quality analysis refers to evaluating and ensuring the quality of the data used in a data science project. This involves identifying problems or errors in the data that may affect the accuracy and reliability of the results. Some common aspects of data quality that are analyzed include completeness (completeness of the data), precision (accuracy of the data), consistency (consistency of the data), and validity (conformance to rules or constraints). In addition, other aspects such as temporal consistency, consistency of formats and detection of outliers can be verified. Data quality analysis seeks to correct or eliminate any problems identified before proceeding with the analysis and interpretation of the data.

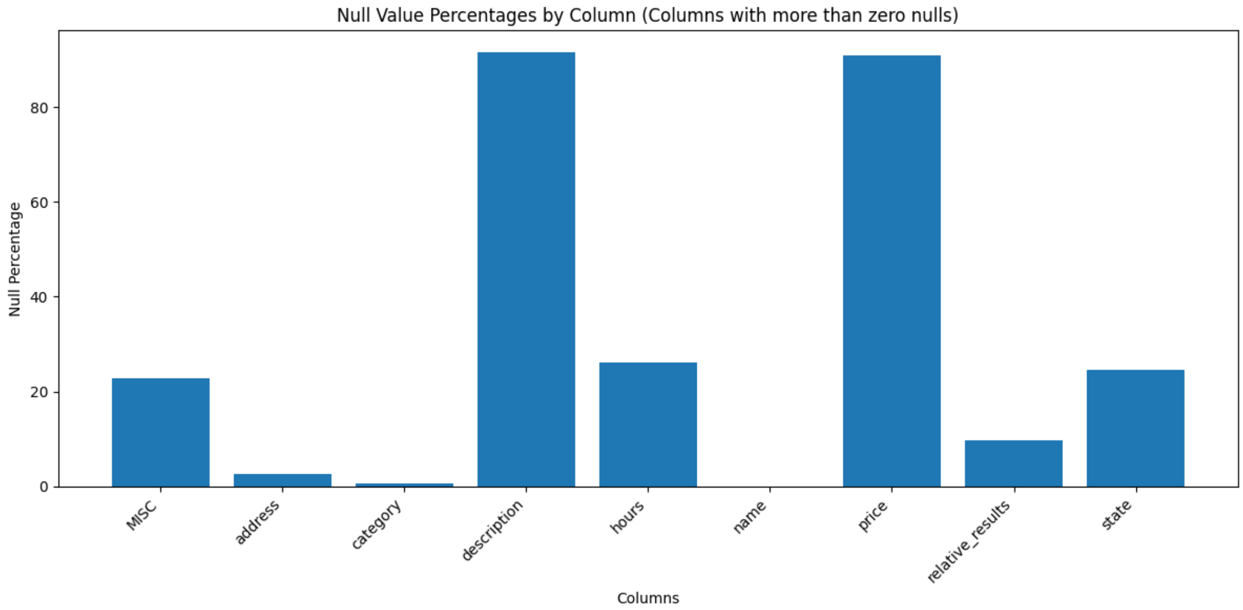
In summary, the preliminary EDA focuses on exploring and understanding the data, while the data quality analysis focuses on ensuring that the data used is reliable and free of errors. Both are important steps in the data analysis process in data science and contribute to more accurate and meaningful results.

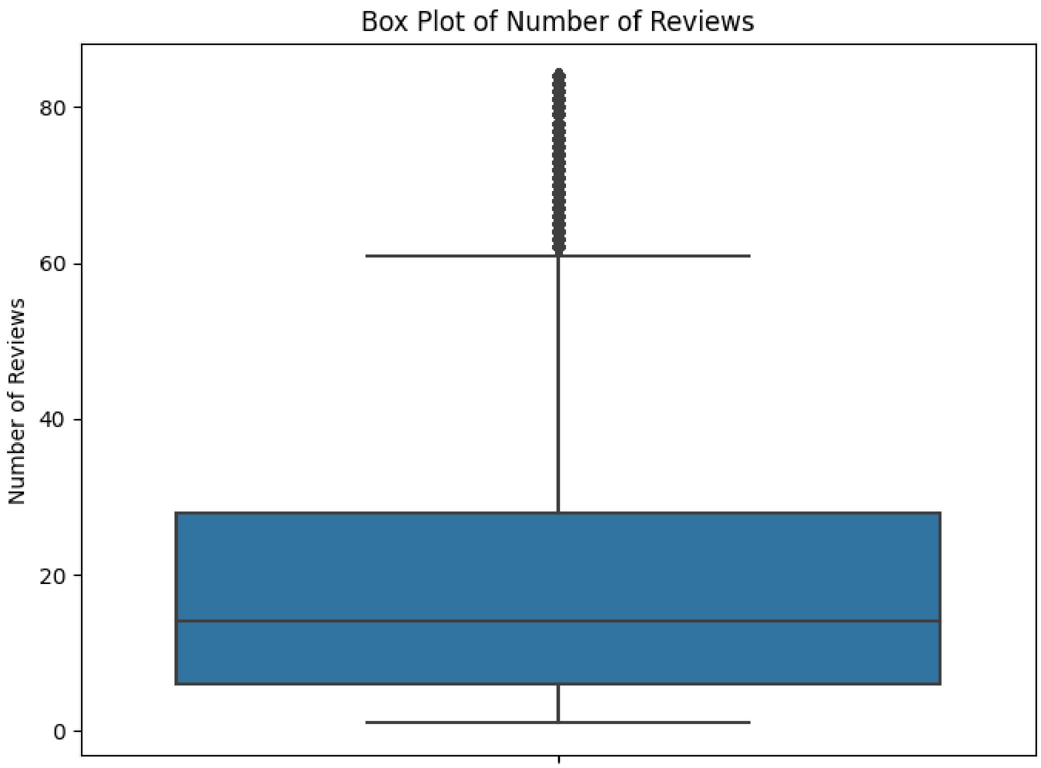
Google Maps dataset

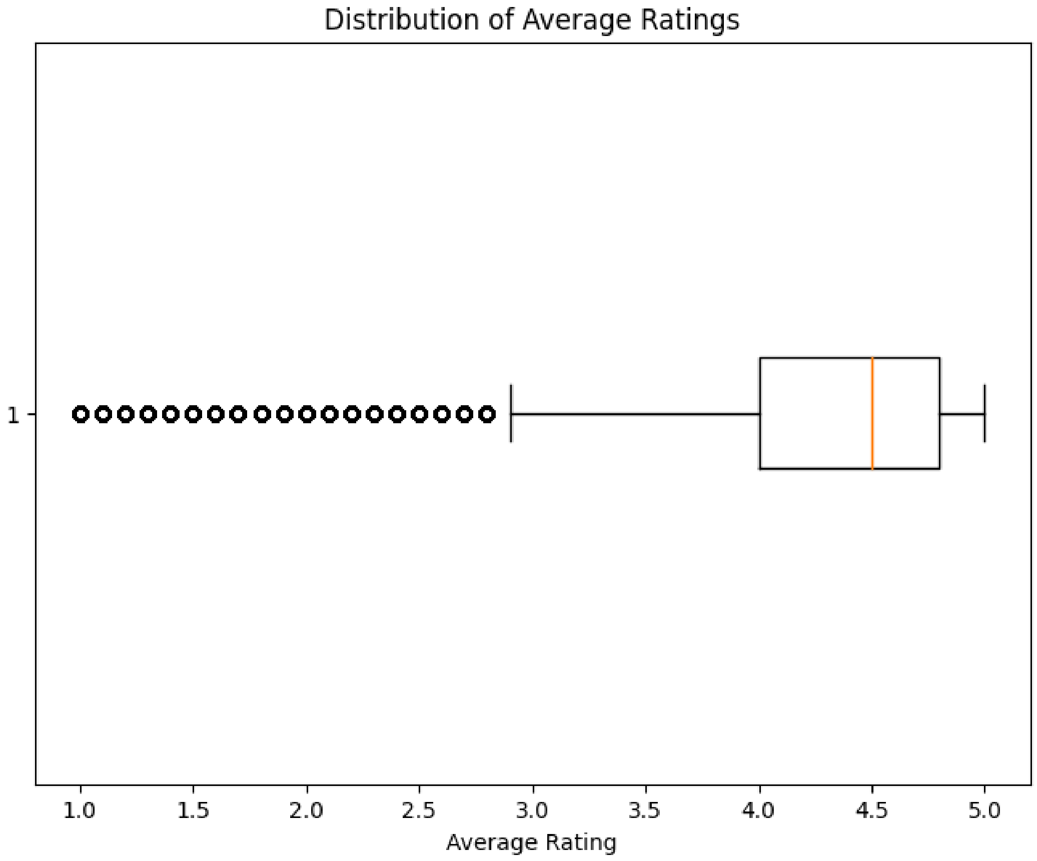
metadata-sitios

Observations:

- There are 3.025.011 rows and 15 columns.
- Number of nulls:
  - "MISC": 690.834
  - "address": 80.511
  - "category": 17.419
  - "description": 2.770.722
  - "hours": 787.405
  - "name": 37
  - "price": 2.749.808
  - "relative_results": 295.058
  - "state": 746.455
- Seems to be 26.573 duplicates



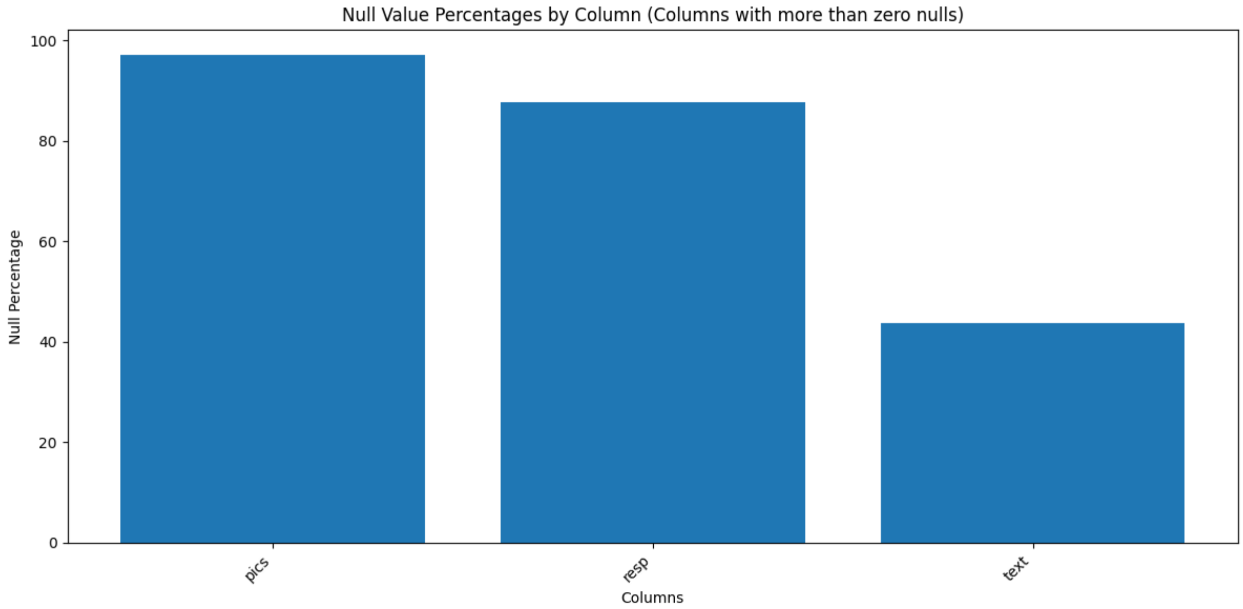

Box Plot of Number of Reviews
Number of Reviews
0
20
40
60
80


Distribution of Average Ratings
1
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Average Rating

reviews-estados

Observations:

- There are 89.946.359 rows and 9 columns.
- There are nulls in columns:
    - "pics" (87.450.680)
    - "resp" (78.917.249)
    - "text" (39.307.744)
- Seems to be 1.386.843 duplicates
- The “time” column is in Unix Time Stamp so we should transform it into date type for analysis.



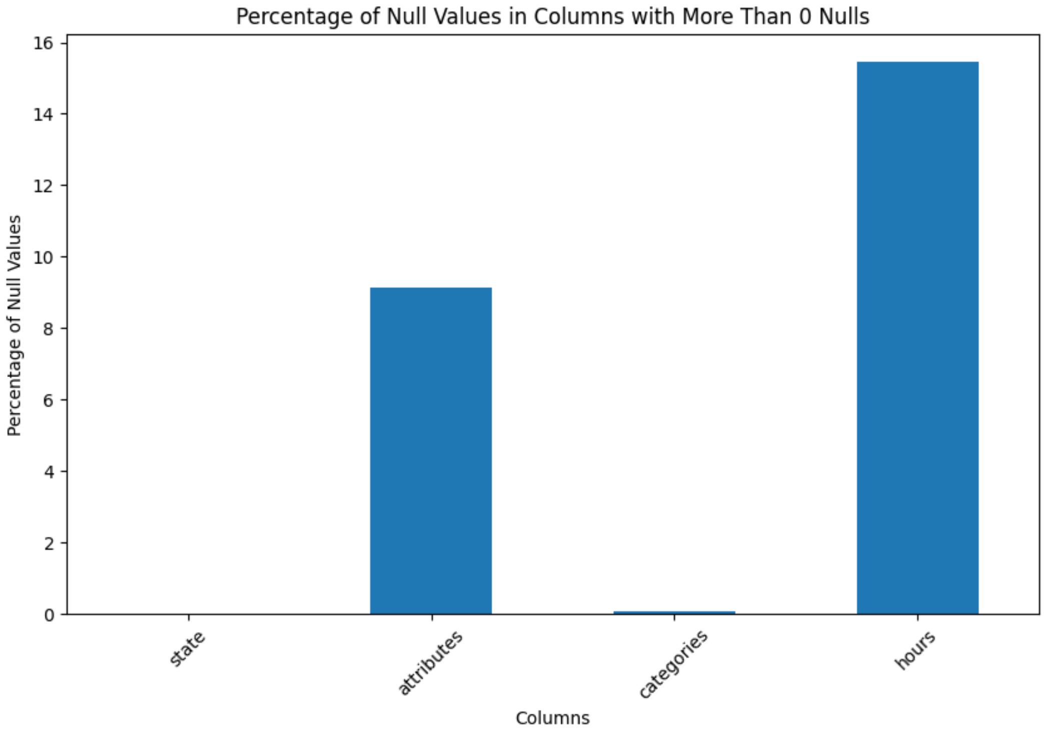
Yelp! dataset

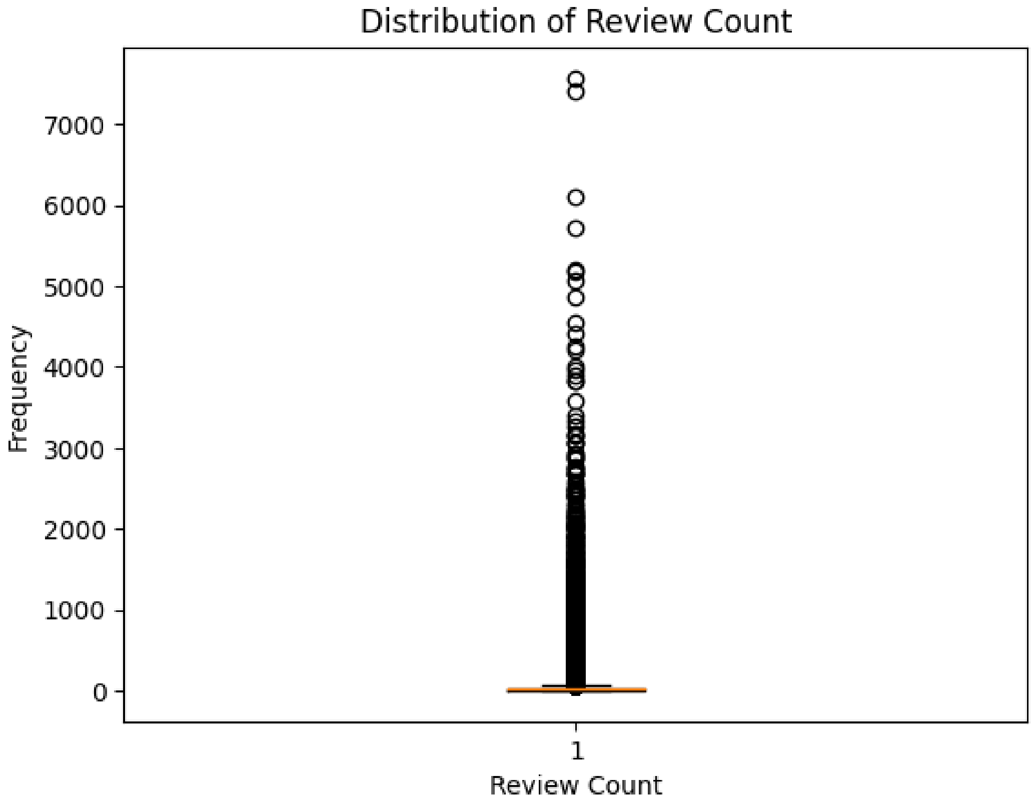
business.pkl

Observations:

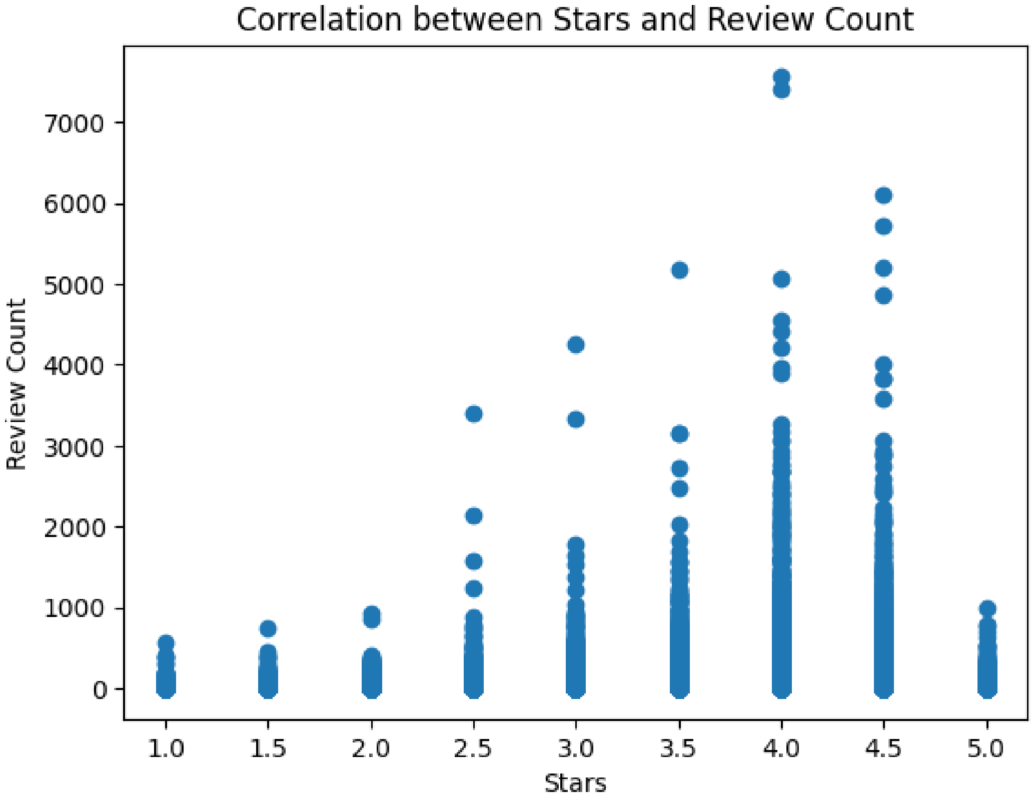
- There are 150.346 rows and 28 (14) columns.

- Columns are repeated twice!
- Aparently there is no duplicate
- Everything is object type. 'stars', 'latitude' and 'longitude' columns should be float type, 'review_count' and 'is_open' columns integer type
- There are null values in columns: state(3), attributes(13744), categories(103), hours(23223)
- There are a few places with relatively much more reviews (possible outliers).
- Seems like there are many reviews between 2.5 and 4.5 stars and less between 1 and 2, and specifically 5





Distribution of Review Count
Frequency
7000
6000
5000
4000
3000
2000
1000
0
1
Review Count



Correlation between Stars and Review Count
Review Count
7000
6000
5000
4000
3000
2000
1000
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Stars

review.json

Observations:

- There are 6.990.280 rows and 9 columns.
- Seems to be no null or duplicated value
- date column is string type

user.parquet

Observations:

- There are 2.105.597 rows and 22 columns.
- Seems to be no null
- Seems to be 117.700 duplicated values
- I do not know if there are outliers