Unraveling Real Estate Ads

Felipe Lana Machado[1]

[1]Software Engineer – Data Scientist
felipe.krids@gmail.com

1. Introduction

Data Sprints used one of its crawlers to mine data from multiple real estate ad portals and these data were obtained through web requests, with results in *json* format. The data came unstructured and containing several problems, such as duplicate data, incomplete lines and discrepant data. The main nodes of the *json* received for analysis were:

- **_id**, has a unique identifier for each of the lines.
- **_index**, has a unique value on all lines: *"realties"*.
- **_score**, is a crawler metric, but all lines have a *"1"*.
- **_source**, is a complex node that is composed of all the data obtained by the crawler.
- **_type**, represents the category, but all lines contain *"imovel"*.

Therefore, the first step was to remove the nodes that contained the same information on all lines, because they add nothing to the analysis to be performed. Thus, the removed nodes were: (I) *_index*, (II) *_score* and (III) *_type*. Leaving only the *_id* and *_source* nodes in json.

2. Reading and Cleaning Data

The next step was to transform the *json _source* node into a dataframe to be able to handle it more easily. The dataframe was created containing 63 columns and 149968 rows. However, not all columns were complete, with *"NaN"* (Not a number) being recorded in spaces that don't have recorded data.

3. Exploratory Data Analysis

4. Conclusion

Referências