Information Visualization
Checkpoint II

Checkpoint II: Data Cleaning & Proces

Group: G13
Date: 2021/10/09

Initial Dataset

As we described on Checkpoint I, we initially had 5 different datasets. The “portugal_ads_proprieties.csv” (dataset 1) table contains information on the properties in the Portuguese Real Estate market. The datasets “portugal_criminality_rates.csv” (dataset 2), “purchase_power.csv” (dataset 3), “population.csv” (dataset 4) and “portugal_area.csv” (dataset 5) contain, respectively, the ‰ of different crimes committed, the purchase power, the population and the area of each city in Portugal.

Dataset samples:

```
from(“portugal_ads_proprieties.csv”)
Locations,Rooms,Price,Area,Bathrooms,Condition,AdsType,ProprietyType
Glória e Vera Cruz, Aveiro,2,850.0,95.0,2,Used,Rent,Apartament
```

Selected/Derived Data

- We decided to produce five different datasets for our visualization, one with all the properties in the Portuguese Real Estate market (“portuguese_real_estate_market.csv”), other with the information about all the Portuguese parishes (“locations_parishes_info.csv”), other with the information about all the Portuguese districts (“locations_districts.csv”), other with the mean price for each type of property (“median_informations_district.csv”) and the other with the mean price for each type of property (“district_median_price_for_each_type_of_house.csv”). The attributes that were selected to appear in our datasets are called “**HM**”, “**Locations**”, “**Rooms**”, “**Price**”, “**PropertyArea**”, “**Area**”, “**Bathrooms**”, “**Condition**”, “**AdsType**”, “**ProprietyType**”, “**CrimeRate**” and “**PurchasePower**”, plus the derived measures that we later created.
- We decided to create several derived measures. “**PopulationDensity**” calculates, for each item, its value in the “HM” attribute (from dataset 4) divided by its value in the “Area” attribute (from dataset 5)., which is good to check if the population density of a certain location influences the price of the properties in it. “**TotalCrimes**” measures the amount of crimes committed calculated by multiplying the “CrimeRate” attribute (from dataset 2) with the “HM” (about the population from dataset 4) divided by 1000, useful to find out how dangerous a city is. We also created a derived measure called “**AverageSellPrice**” and “**AverageRentPrice**”, which do an average of all the values of the “Price” attribute from selling or renting respectively, from items of a certain location, allowing us to get an idea of the property search in the location and check if it has some relation with some attributes of the dataset with informations about each location.

Data Abstraction

- Regarding the dataset types we will use in our visualization, the Real EstateMarket dataset is a table with items and attributes and the Location Informations dataset is spatial as it has information on different locations on Portugal’s surface that will be used for spatial analysis.

- For the Real Estate market dataset, the attributes that describe each property (item) are:
  Nominal attributes: “**Locations**” (physical location of each property), “**Condition**” (if the property is renovated, new or other), “**AdsType**” (if the property is for sale or for renting), “**ProprietyType**” (if the property is an apartment or a house);
  Ordinal, Sequential attributes: “**Rooms**” (amount of bedrooms in the property);
  Ratio, Sequential attributes:“**Price**” (property’s price), “**PropertyArea**” (area of the whole property), “**Bathrooms**” (amount of bathrooms in the property);
- For the Location Informations dataset, the attributes of each location (item) are:
  Nominal attributes: “**Location**” (name that identifies the location);
  Ratio, Sequential attributes: “**CrimesRate**” (rate of crimes per a thousand inhabitants in a location) , “**TotalCrimes**” (total crimes per location in a year), “**PurchasePower**” (number of goods or services that one unit of money can buy in a location), “**PopulationDensity**” (number of people per unit of area in a location), “**AveragePropertyPrice**” (average price of a property that is for sale in a location), “**AverageRentPrice**” (average rent price of a property that is for rent in a location)

Data Processing

We did the Data Processing with the Pandas Python library. In our five original datasets, we only found 10 missing values in the dataset 3, in which we discarded the items with those missing values in order to fix them. We computed the Interquartile Range (IQR) formula to find the outliers, finding 138432 outliers in dataset 1, 46 in dataset 2, 58 in dataset 3, 1006 in dataset 4 and 88 in dataset 5. We decided to maintain every item with a value detected as an outlier in the datasets 2, 3, 4 and 5 since there wasn’t any value that differed a lot from the rest in order to distort the visualization. We merged the different datasets by comparing their “Locations” attribute. We also removed the duplicated locations in our produced dataset with the information of every location, by sorting the dataset by the “PopulationDensity” attribute in descending order and then keeping only an item for each “Location”. We also dropped some irrelevant columns especially in the dataset 2. There were also missing values in the “AveragePropertyPrice” and “AverageRentPrice” attributes in the dataset 2, so we replaced them with the value 0. We did a Group by of the various attributes in the “portuguese_real_estate_market” dataset in order to show the mean “Price” value of the properties with the same values in the other attributes.

Mapping (Data sample/Questions)

To answer to:

- **Question 1:** We look at the Locations dataset and compare all the attributes.
- **Question 2:** We look at the “Price” attribute from the properties dataset with the adsType being “Rent” and to the “AverageRentPrice” and “PurchasePower” in the Real Estate Market dataset. With the location’s city being the same in both datasets.
- **Question 3:** We look at the Real Estate Market dataset and compare all attributes.
- **Question 4:** In the “locations_info” dataset, we look at the values of the “PopulationDensity” and try to see if the values for the “AverageSellPrice” or the “AverageRentPrice” are higher or lower depending on the “PopulationDensity” values.
- **Question 5:** We look at the “locations_info” dataset and check the values of the “CrimeRate” and try to see if the “AverageSellPrice” or the “AverageRentPrice” values depend on the “PopulationDensity” values.