BDE Assignment

Contents

1 Introduction 1

2 About the datasets 2

3 Data Cleaning and Preparation 2

3.1 Parsing the data 2

3.2 Understanding the datasets 2

3.3 Cleaning the datasets 4

4 Merging the datasets 5

5 Results 7

1 Introduction

This documentation outlines a data engineering project aimed at creating a joined dataset from three distinct sources: Google, Facebook, and Website data.

The report was created by Sándor Kolumbán.

A summary presentation of the results is available here: https://app.pitch.com/app/public/player/41f05329-b0c7-4cb9-98d9-44566e87fdb4

The code belonging to the assignment is available here: https://github.com/Explorer-Consulting/Test-Assignment-Exadel-BDE

2 About the datasets

We have three different files for the datasets:

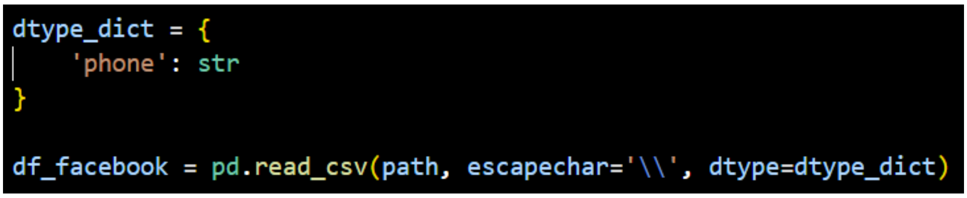
- Facebook_dataset.csv
- Google_dataset.csv
- Website_dataset.csv

These datasets contain some basic information about companies (e.g. Company name, address, categories, phone number, etc.) from different sources.

3 Data Cleaning and Preparation

3.1 Parsing the data

Reading the datasets was not so straightforward. For the Facebook and Google dataset there were columns which data also contained commas escaped by \ characters, so we had to specify escape character parsing the data. These two datasets also contained some strange values in the 'phone' column, so we had to also tell the pandas to parse the 'phone' column as string. For the website dataset, the data was separated by semicolons.

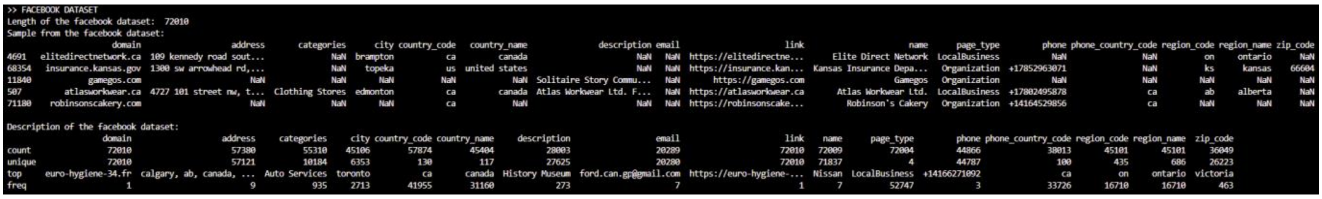


```
dtype_dict = {
    'phone': str
}

df_facebook = pd.read_csv(path, escapechar='\\', dtype=dtype_dict)
```

Figure 1 Phone numbers must be understood as strings and \ character should be understood as escape character.

3.2 Understanding the datasets

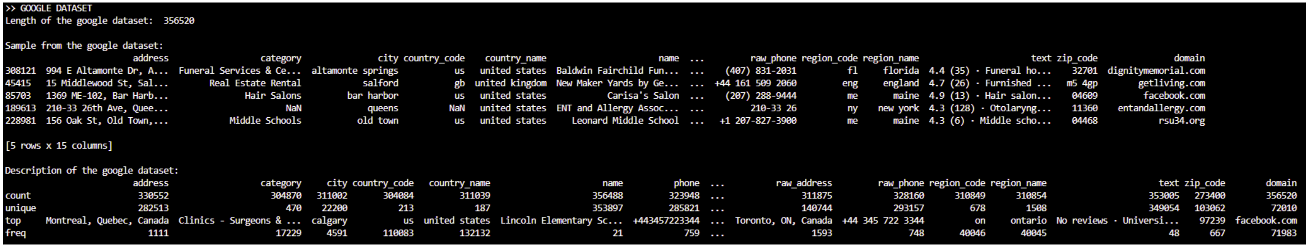
After parsing the datasets, deeper understanding of the data structure was obtained by utilizing the <u>pandas.DataFrame.**describe()**</u> method for each dataset, allowing us to examine the fundamental characteristics of the datasets and better comprehend the nature of the data we are working with.



```
>> FACEBOOK DATASET
Length of the facebook dataset:  72010
Sample from the facebook dataset:
                      domain                  address        categories      city country_code   country_name          description email                     link                  name      page_type         phone phone_country_code region_code region_name zip_code
4691   elitedirectnetwork.ca  109 kennedy road sout...               NaN  brampton           ca         canada                  NaN   NaN https://elitedirectne...  Elite Direct Network  LocalBusiness           NaN                NaN          on     ontario      NaN
68354   insurance.kansas.gov  1300 sw arrowhead rd,...               NaN    topeka           us  united states                  NaN   NaN https://insurance.kan...  Kansas Insurance Depa...   Organization  +17852963071                NaN          ks      kansas    66604
11840            gamegos.com                      NaN               NaN       NaN          NaN            NaN  Solitaire Story Commu...   NaN      https://gamegos.com               Gamegos   Organization           NaN                NaN         NaN         NaN      NaN
507          atlasworkwear.ca  4727 101 street nw, t...  Clothing Stores  edmonton           ca         canada  Atlas Workwear Ltd. F...   NaN https://atlasworkwear.ca   Atlas Workwear Ltd.  LocalBusiness  +17802495878                 ca          ab     alberta      NaN
71180      robinsonscakery.com                      NaN               NaN       NaN           ca            NaN                  NaN   NaN https://robinsonscake...     Robinson's Cakery   Organization  +14164529856                 ca         NaN         NaN      NaN

Description of the facebook dataset:
                    domain                  address     categories     city country_code country_name     description                 email                     link   name      page_type         phone phone_country_code region_code region_name  zip_code
count                72010                    57380          55310    45106        57874        45404           28003                 20289                    72010  72009          72004         44866              38013       45101       45101     36049
unique               72010                    57121          10184     6353          130          117           27625                 20280                    72010  71837              4         44787                100         435         686     26223
top      euro-hygiene-34.fr  calgary, ab, canada, ...  Auto Services  toronto           ca       canada  History Museum  ford.can.gp@gmail.com https://euro-hygiene-...  Nissan  LocalBusiness  +14166271092                 ca          on     ontario  victoria
freq                     1                        9            935     2713        41955        31160             273                     7                        1      7          52747             3              33726       16710       16710       463
```

Figure 2 Results of describe() for the Facebook data

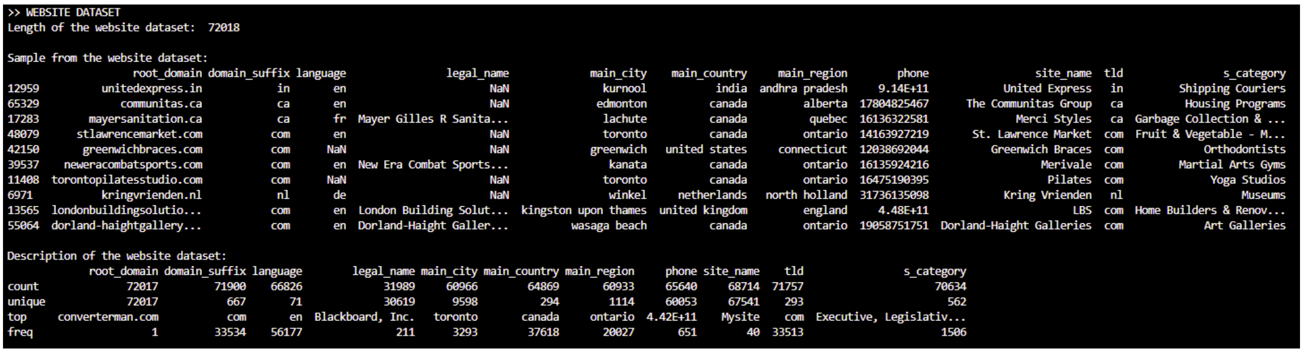
In the Facebook dataset, we observed a total of 72,010 records (see Figure 2). It's noteworthy that the 'domain' column was unique across all records, with each row featuring a distinct domain entry. This unique characteristic sets the 'domain' column apart from the others in the dataset and made it a good candidate for joining.



```
>> GOOGLE DATASET
Length of the google dataset:  356520

Sample from the google dataset:
                          address               category               city country_code   country_name                      name  ...        raw_phone region_code region_name                       text zip_code               domain
308121  994 E Altamonte Dr, A...  Funeral Services & Ce...  altamonte springs           us  united states  Baldwin Fairchild Fun...  ...   (407) 831-2031          fl     florida  4.4 (35) · Funeral ho...    32701  dignitymemorial.com
45415   15 Middlewood St, Sal...      Real Estate Rental            salford           gb  united kingdom  New Maker Yards by Ge...  ...  +44 161 509 2060         eng     england  4.7 (26) · Furnished ...   m5 4gp          getliving.com
85703   1369 ME-102, Bar Harb...             Hair Salons         bar harbor           us  united states          Carisa's Salon  ...   (207) 288-9444          me       maine  4.9 (13) · Hair salon...    04609         facebook.com
189613  210-33 26th Ave, Quee...                     NaN             queens          NaN  united states  ENT and Allergy Assoc...  ...        210-33 26          ny    new york  4.3 (128) · Otolaryng...    11360      entandallergy.com
228981  156 Oak St, Old Town,...          Middle Schools           old town           us  united states    Leonard Middle School  ...  +1 207-827-3900          me       maine  4.3 (6) · Middle scho...    04468             rsu34.org

[5 rows x 15 columns]

Description of the google dataset:
                          address               category     city country_code   country_name                      name          phone  ...           raw_address         raw_phone region_code region_name                          text zip_code        domain
count                      330552                 304870   311002       304084         311039                    356488         323948  ...                311875            328160      310849      310854                        353005   273400        356520
unique                     282513                    470    22200          213            187                    353897         285821  ...                140744            293157         678        1508                        349054   103062         72010
top     Montreal, Quebec, Canada  Clinics - Surgeons & ...  calgary           us  united states  Lincoln Elementary Sc...  +443457223344  ...  Toronto, ON, Canada  +44 345 722 3344          on     ontario  No reviews · Universi...    97239  facebook.com
freq                         1111                  17229     4591       110083         132132                        21            759  ...                  1593               748       40046       40045                            48      667         71983
```

Figure 3 Results of describe() for the Google data

Within the Google dataset (see Figure 3), we encountered a total of 356,520 records. None of the columns exhibited the characteristics of being both unique and set in every row of the dataset, setting it apart from the Facebook dataset. However, an observation was made regarding the 'domain' column, where the unique count equaled 72,010, precisely matching the number of records in the Facebook dataset. This observation suggests a potential correspondence between the domains in the two datasets, indicating that multiple rows within the Google dataset may share the same domain.

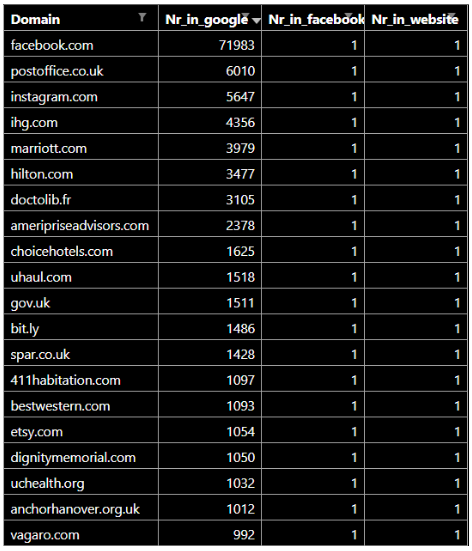


```
>> WEBSITE DATASET
Length of the website dataset:  72018

Sample from the website dataset:
                    root_domain domain_suffix language                legal_name            main_city   main_country     main_region       phone                site_name  tld                s_category
12959          unitedexpress.in            in       en                       NaN              kurnool          india  andhra pradesh    9.14E+11           United Express   in         Shipping Couriers
65329             communitas.ca            ca       en                       NaN             edmonton         canada         alberta 17804825467     The Communitas Group   ca          Housing Programs
17283        mayersanitation.ca            ca       fr  Mayer Gilles R Sanita...              lachute         canada          quebec 16136322581             Merci Styles   ca  Garbage Collection & ...
48079      stlawrencemarket.com           com       en                       NaN              toronto         canada         ontario 14163927219      St. Lawrence Market  com  Fruit & Vegetable - M...
42150       greenwichbraces.com           com      NaN                       NaN            greenwich  united states     connecticut 12038692044         Greenwich Braces  com             Orthodontists
39537    neweracombatsports.com           com       en  New Era Combat Sports...               kanata         canada         ontario 16135924216                 Merivale  com         Martial Arts Gyms
11408  torontopilatesstudio.com           com      NaN                       NaN              toronto         canada         ontario 16475190395                  Pilates  com              Yoga Studios
6971           kringvrienden.nl            nl       de                       NaN               winkel    netherlands   north holland 31736135098           Kring Vrienden   nl                   Museums
13565  londonbuildingsolutio...           com       en  London Building Solut... kingston upon thames united kingdom         england    4.48E+11                      LBS  com  Home Builders & Renov...
55064  dorland-haightgallery...           com       en  Dorland-Haight Galler...         wasaga beach         canada         ontario 19058751751 Dorland-Haight Galleries  com             Art Galleries

Description of the website dataset:
             root_domain domain_suffix language        legal_name main_city main_country main_region     phone site_name   tld                s_category
count              72017         71900    66826             31989     60966        64869       60933     65640     68714 71757                     70634
unique             72017           667       71             30619      9598          294        1114     60053     67541   293                       562
top     converterman.com           com       en  Blackboard, Inc.   toronto       canada     ontario  4.42E+11    Mysite   com  Executive, Legislativ...
freq                   1         33534    56177               211      3293        37618       20027       651        40 33513                      1506
```

Figure 4 Results of describe() for the Website data

We identified 72,018 records within the Website dataset (see Figure 4), which is slightly surpassing the number of records in the Facebook dataset. Similarly to the Facebook dataset, the **domain** column in the Website dataset was unique, and the count of unique domains closely matched those in the other datasets. This consistent pattern of unique domains across the dataset indicates that the **domain** column held distinct values in each dataset and was pivotal in our data analysis.

With the understanding that the **domain** column could serve as the key for joining the dataset, we sought to determine the presence of domains in each of the datasets. Our objective was to assess whether all domains were represented in every dataset and how many times each domain appeared in which dataset.
To achieve this, we counted the occurrence of every domain within each dataset. A portion of the results is given in Figure 5.



Domain	Nr_in_google	Nr_in_facebook	Nr_in_website
facebook.com	71983	1	1
postoffice.co.uk	6010	1	1
instagram.com	5647	1	1
ihg.com	4356	1	1
marriott.com	3979	1	1
hilton.com	3477	1	1
doctolib.fr	3105	1	1
ameripriseadvisors.com	2378	1	1
choicehotels.com	1625	1	1
uhaul.com	1518	1	1
gov.uk	1511	1	1
bit.ly	1486	1	1
spar.co.uk	1428	1	1
411habitation.com	1097	1	1
bestwestern.com	1093	1	1
etsy.com	1054	1	1
dignitymemorial.com	1050	1	1
uchealth.org	1032	1	1
anchorhanover.org.uk	1012	1	1
vagaro.com	992	1	1

Figure 5 Occurrence nr. of domains in various datasets (the list continues).

Our analysis now provides a clear overview of the frequency with which domains are represented in the datasets. It’s clear now that only the Google dataset contains domains appearing multiple times.

The multiple occurrences of domains in the Google dataset may have different underlying reasons. So, the next step was to see what is the relationship between entries in the Google dataset sharing the same domain.

Our investigation also extended to examining the differences between the sets of domains in the three datasets, which resulted in the following observations:

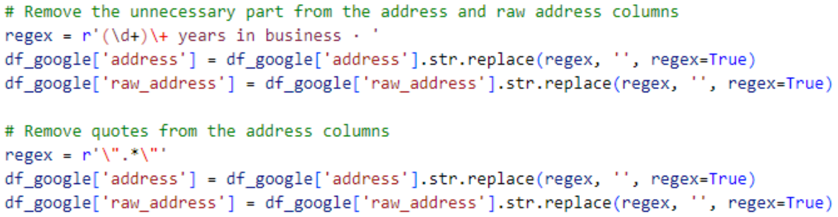
- Unique to Google: An empty set, indicating no domain is unique to the Google dataset.
- Unique to Facebook: An empty set, implying no domain is exclusively found in the Facebook dataset.
- Unique to Website: Several unique entries were identified, and all of them were non-domain entries like 'Fitzwilliam NH 03447’, ‘Avenida Food Hall & Fresh market’, ‘Paint &’, ‘MARTIN-LAFLAMME’ and others.

These distinctive entries are specific to the Website dataset and are not shared with the other datasets.

3.3 Cleaning the datasets

Prior to commencing the dataset integration process, we recognized the need to standardize and prepare the data. To begin this data preparation, we applied the **_(dataset_name)** suffix to every column in each of the datasets. The first step involved working with the Google dataset.

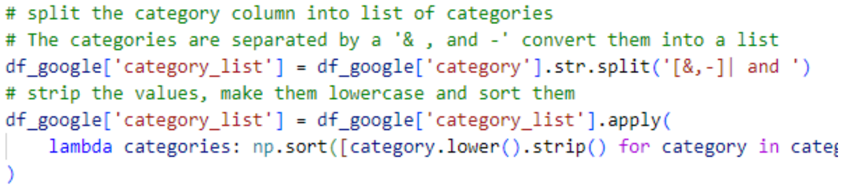
In the Google dataset, we encountered non-conforming values within the **address** and **raw_address** columns. These values included unnecessary information, which we successfully removed using regular expressions (regex), see Figure 6.



```
# Remove the unnecessary part from the address and raw address columns
regex = r'(\d+)\+ years in business · '
df_google['address'] = df_google['address'].str.replace(regex, '', regex=True)
df_google['raw_address'] = df_google['raw_address'].str.replace(regex, '', regex=True)

# Remove quotes from the address columns
regex = r'\".*\"'
df_google['address'] = df_google['address'].str.replace(regex, '', regex=True)
df_google['raw_address'] = df_google['raw_address'].str.replace(regex, '', regex=True)
```

Figure 6 Remove years in business and quotation marks from adress fields in Google data

Next, we turned our attention to the **categories** column, where we observed variations in how categories were presented. In order to standardize this data and improve its consistency, we made the decision to separate the category values and transform them into a list of categories within a dedicated column.



```
# split the category column into list of categories
# The categories are separated by a '& , and -' convert them into a list
df_google['category_list'] = df_google['category'].str.split('[&,-]| and ')
# strip the values, make them lowercase and sort them
df_google['category_list'] = df_google['category_list'].apply(
    lambda categories: np.sort([category.lower().strip() for category in categ
)
```

Figure 7 Splitting category information into a list format

We also performed column renaming to establish a more standardized and consistent naming convention, to ensure that the column names across the datasets followed a uniform pattern, making it easier to work with and integrate the data seamlessly.

Our data cleaning process for the Facebook dataset followed a pattern, like the one for the Google dataset. We maintained consistency in naming conventions by renaming specific columns. The category values were transformed into a list of categories also.

In the Website dataset, we executed transformations akin to those performed in the other datasets. We converted its category column named **s_category** into a list of categories, and column renaming was applied to standardize the naming conventions as well.

Furthermore, in the Website dataset it was required to determine the most accurate company name, by evaluating and choosing between two of its columns: **legal_name** and **site_name.** Following consideration and analysis of the data, we made the decision to prioritize **site_name** whenever it was available for determining the company name**.** In most cases it seemed more accurate than the value in the **legal_name** column, however there may still be room for improvement.

4 Joining the datasets

Upon gathering the insights into the data, we quickly determined that joining the datasets based on the **domain** columns was the most logical and effective approach. However, we also recognized the unique challenge posed by the Google dataset, which contained multiple rows for each domain. To successfully address this, we would need to find out a suitable strategy to handle these multiple entries withing the Google dataset while integrating it with the others.

Our approach was to merge the datasets based on the **domain** column, creating rows that included the columns of each dataset for every domain entry. Subsequently, we undertook the task of selecting the most accurate values for the final columns that held significance for us, such as **Company Name, Address, Category,** and more.

In most cases, the values found in the Google dataset proved to be the most suitable and reliable choice for the final columns, so we chose them when they were available. However, in cases where the Google dataset values were not present, we selected the Facebook dataset as our secondary choice, favoring it over the Website dataset.

Of all the columns, two specific columns that required distinct treatment were the **category_list** and **company_name** columns.

4.1 Category lists

For the **category_list** column, our approach involved consolidating the content of category lists from all the dataset columns. This method was effective in most cases, allowing us to merge category lists seamlessly.
However, it's important to note that in some instances, we encountered values that did not align with the other category list values (e.g. a business doing car repairs and car part sales was also

marked as yoga studio). As a continuation of the data cleaning process, currently popular LLMs could be asked to determine of there is an outlier in a given category list.

4.2 Company name

Resolving the **company_name** column presented a more complex challenge, primarily due to the data in the Google dataset. In certain rows, the company names in the Google dataset included human names, which could be interpreted as employees or individuals. However, in other cases, there were sub-company names associated with a single domain. Based on this insight we classified domains in the Google dataset in three categories:

- A domain representing a single company, with proper data.
- A domain appearing multiple times, and the multiple occurrences are due to employees or individuals of the same company being recorded multiple times.
- A domain that can be thought of as an aggregator.

The case of aggregator domains (a large number of entries in the Google dataset sharing the same domain) can occur when smaller businesses don't own their own domain and their most visible online presence is under an aggregator domain (like facebook.com).

To achieve this classification, we needed to determine when to rely exclusively on the data from the Google dataset (in case of aggregator domains, i.e. facebook, there is no point in joining data belonging to facebook.com to every small business only visible there) and when to choose the most appropriate value from the available values in all datasets. For this, we looked at the domain count table (Figure 5) and decided that when a domain is encountered more than a specific number (in our case 500), we prioritize using the **company_name_google** value whenever it's available. This approach addresses scenarios where multiple companies are associated with a single website, such as in the case of **facebook.com**.

We also had to address the second situation, where the values in the company_name column from the Google dataset appeared to be employee names. To tackle this issue, we implemented a relatively straightforward solution involving the use of a regular expression (regex) to identify titles resembling human names, like M.D., Ph.D., and others. While this method may not be entirely foolproof, it effectively resolves the issue in the majority of cases. There are packages that allow recognition of human names as well, utilizing such a package might improve the detection of such cases.

All other cases were thought of as regular companies, represented by proper entries in the Google dataset, having matching data in other datasets. In order to get an improved version of the Company name, we utilized the **Levenshtein** distance metric. The choice for the company name in the joined dataset was the most central one according to this metric. If not all three datasets contained a company name, then the central one is not uniquely defined, in such cases we prioritized Google over Facebook, and the data from Website was used as last resort.

Following the completion of the joining process, we shifted our focus to address the potential issue of duplicated values, which had not been previously handled. To resolve this, we made the decision to eliminate duplicate entries based on the **address** and **phone** columns.

5 Results

In the final stages of this process, we generated two datasets. The first dataset containes all the merged columns from the three original datasets, along with our chosen values. The second dataset contains only the selected values. Both datasets were organized and sorted based on the **domain** column.