Data Analyst - Case

You are provided with the **data_case.csv** file. This file has registration data from active Brazilian companies of the state of Paraná. This data is freely available from Receita Federal (Brazilian equivalent to United States IRS).
Registration data include columns such:

- Date when the company was created
- Address information (in this case, city, state and zip code)
- Business activity of the company (if the company is a bakery, clothing store, plumbing services etc)

Imagine that you are working on a new product geared towards businesses (a B2B product), and you want to understand the Brazilian companies, especially the ones in the state of Paraná. So you are provided with the CSV file, and are asked to generate insights from this data.

The dataset was **arbitrarily processed**, some columns have weird data formats that need to be processed for full use.

Thinking outside of the box is encouraged, some columns may have information that could be used in different ways, you are free to use different visualizations packages, additional data sources etc.

Task:

- Preprocess the dataset
    - Clean (preprocess all “odd” columns, and pre-process values)
    - Enrich (get data from grouping operations, maybe with extra data)
- Perform EDA (Exploratory Data Analysis) on the dataset
    - Generate visualizations
    - Share your insights and conclusions
- Preferably use Python

Open Questions (optional):

- Any conclusions based on the number of companies created by date, month, year etc?
- Does the number of branches for a CNPJ provided on the dataset is equal to the actual number produced by the dataset?
- What about the business activities (CNAE, MCC)? What are the main types? Can they be aggregated into fewer groups?
- What are the differences between the cities / zip codes?

- Is it possible to catch any spatial relationships? Which visualizations would be best in this case?
- If you were to make any model from the data, which one do you think makes sense?

Deliverables*:

- Code used on processing the files
- Presentation (~20 minutes)

* Files like a jupyter notebook can be used as code & presentation (i.e if you choose such format, you only need to deliver one ipynb file)

Data dictionary:

'document_number':

The full number of CNPJ (Cadastro Nacional da Pessoa Jurídica), i.e an identifier for Brazilian companies.

It is composed by 14 digits, the first 8 identify a company, the next 4 digits define the branch and the final 2 digits are check digits

'cnpj_basico':

The first 8 digits of the full CNPJ

'establishment_type':

The type of establishment

- 'MATRIZ': if Headquarters (or Parent Company)
- 'FILIAL' if Branch Office (or Subsidiary)

'razao_social':

This term is typically translated as "Legal Name" or "Corporate Name." It refers to the official, registered name of the company, often used for legal and administrative purposes.

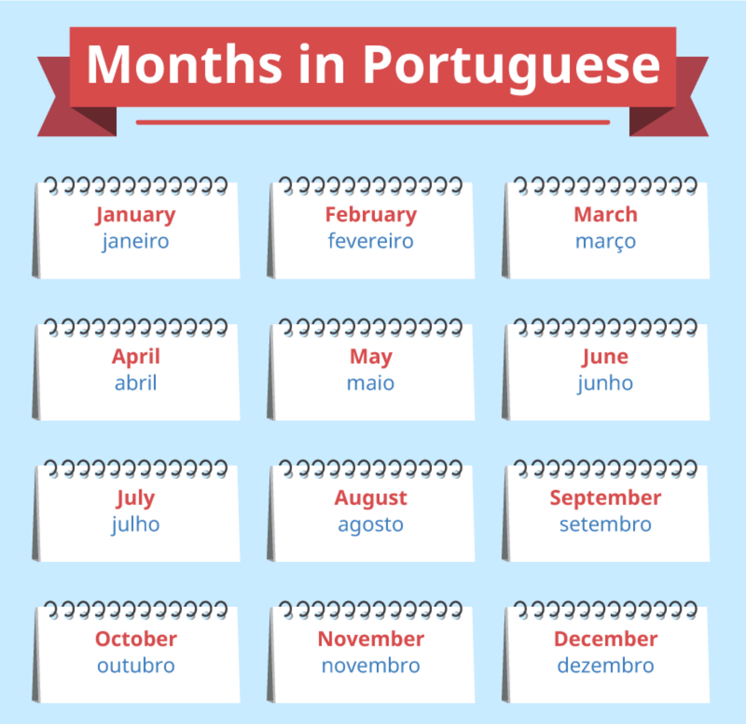
'nome_fantasia':

This term is usually translated as "Trade Name," "Brand Name," or "Doing Business As (DBA) Name." It refers to the name under which a company operates publicly, often used for marketing and branding purposes.

'opening_date':

date when the company was created. The date is written in portuguese in the format of day of month of year. In Portuguese the word “de” has the same meaning as “of”. To get the name of the month you can look into references such as the following picture. Examples include:

- 26 de novembro de 2003 => 26th of November, 2003
- 2 de março de 2020 => March 2nd, 2020
- 21 de junho de 2013 => 21st of June, 2013
- 1 de janeiro de 1990 => 1st of January, 1990
- 15 de setembro de 2005 => 15th of September, 2005
- 7 de agosto de 2018 => 7th of August, 2018



Months in Portuguese

January janeiro	February fevereiro	March março
April abril	May maio	June junho
July julho	August agosto	September setembro
October outubro	November novembro	December dezembro

'cnae':

The CNAE (Classificação Nacional de Atividades Econômicas) code of a company, the IBGE code for the business activities. The code is generated by grouping several hierarchies, based on the digit position, so the first N digits mean an aggregation on the Nth level. There is further documentation provided on the IBGE website.

'cnae_description':

The description of the CNAE - business activity of the company

'mcc':

The Merchant Category Code (MCC) is a four-digit numerical code assigned to businesses by credit card companies and financial institutions to classify the primary type of goods or services they provide. Each MCC corresponds to a specific industry or business type. To get the MCC code, we derived it from the CNAE using a conversion table.

'mcc_description':

The description of the MCC - business activity of the company

OBS: It is recommended the usage of CNAE and its description over MCC **if you know Portuguese**, otherwise feel free to use MCC and its English description. This is because the CNAE is the information originally provided to Receita Federal, while MCC is a translation of the activity to fit into a payments context (and with descriptions in English non-Portuguese speakers can read it to perform the analysis).

'total_branches_and_associates':

JSON-like column with information regarding branches and number of associates:

- total_associates: total number of associates (sócios) of the company
- 'total_branch(es)': total number of branches of the main CNPJ

'city_state':

City and state of the company

'city_code':

IBGE code for the city (it is used in a different array of different data sources, like an identifier for cities, for instance, for the city of Curitiba, its city_code is 4106902, which can be seen in resources such as: https://cidades.ibge.gov.br/brasil/pr/curitiba/panorama, or https://github.com/ipeaGIT/geobr)

'zip_code':

Zip code of the company's address

‘share_capital’:

company's share capital (capital social)

'size_company':

size of the company stated during registration on Receita Federal

'juri_description':

description of legal entity of the company

'juri_description_ENG':

description of legal entity of the company translated to English

'email_provider':

The provider of the email provided by the company legal representative at the time of registry