Data Analyst - Case

You are provided with the **data_case.csv** file. This file has registration data from Brazilian companies of the state of Paraná. This data is freely available from Receita Federal[1]. 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 enrich in different ways 👀)
- 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)? What are the main types? Can they be aggregated into fewer groups?

---

[1] https://www.gov.br/receitafederal/pt-br/assuntos/orientacao-tributaria/cadastros/consultas/dados-publicos-cnpj

- 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 head office

'FILIAL' if branch office

'razao_social':

company name

'nome_fantasia':

trade name of a company

'opening_date':

date when the company was created

'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

'total_branches_and_socios':
    JSON-like column with information regarding branches and number of associates
        'tot_branches': total number of branches of the main CNPJ
        'tot_socios': total number of associates of the company

'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)

'zip_code':
    Zip code of the company's address

'capital_social':
    company's share capital

'size_company':
    size of the company

'juri_description':
    description of legal entity of the company

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