PROJETO DE CIÊNCIA DE DADOS: Linkedin Data Engineer Job Postings

Nome: *Gustavo Navarro Felix*, Email: *gustavo.sempre.estude@gmail.com*

Introdução

Este projeto é uma iniciativa pessoal com o objetivo de praticar e aprofundar meus conhecimentos em ciência de dados.

O dataset trata de posts de vagas de emprego para Data Engineer dentro do Linkedin em 4 países: Estados Unidos, Canadá, Reino Unido e Austrália.

Link para o dataset:
https://www.kaggle.com/datasets/asaniczka/linkedin-data-engineer-job-postings

As tecnologias que utilizei para desenvolver este projeto incluem:

Python 3 - Welcome to Python.org
Anaconda - Anaconda | The World's Most Popular Data Science Platform
Jupyter Notebook - Project Jupyter | Home
Power Bi - Power BI - Visualização de Dados | Microsoft Power Platform

O código e arquivos relacionados ao projeto estão disponíveis no GitHub: GustavoNav/ProjetosDeDados (github.com)

Objetivo

O objetivo deste projeto é preparar os dados para então extrair valor e significado do conjunto de dados.

Análise exploratória

Para realizar a análise exploratória, os arquivos foram carregados no Power BI e no jupyter notebook. A únicas coisas a constatar a respeito dessa análise, é que a coluna "job_skills" precisa receber um tratamento para ser utilizada, além disso, existe um total de 6% de valores nulos em algumas colunas. O tratamento desses 2 problemas é encontrado na próxima sessão.

Limpeza e Transformação

O problema da coluna “job_skills” foi solucionado dentro do jupyter notebook, eu criei um script em python para criar um novo data frame, o qual recebe cada uma das skills separadamente, e conta o número total de repetições, nomeada “Jobs”

O problema dos valores nulos é algo mais complicado de lidar. Após algum tempo cheguei a conclusão que é melhor manter esses dados, eles são equivalentes a 6% do dataset, entretanto os valores nulos só ocorrem em 2 colunas. Acredito que isso não vá influenciar negativamente as Análises.

O Power BI tentou criar uma relação entre as duas tabelas criadas, entretanto eu removi essa relação para não provocar problemas durante as análises.

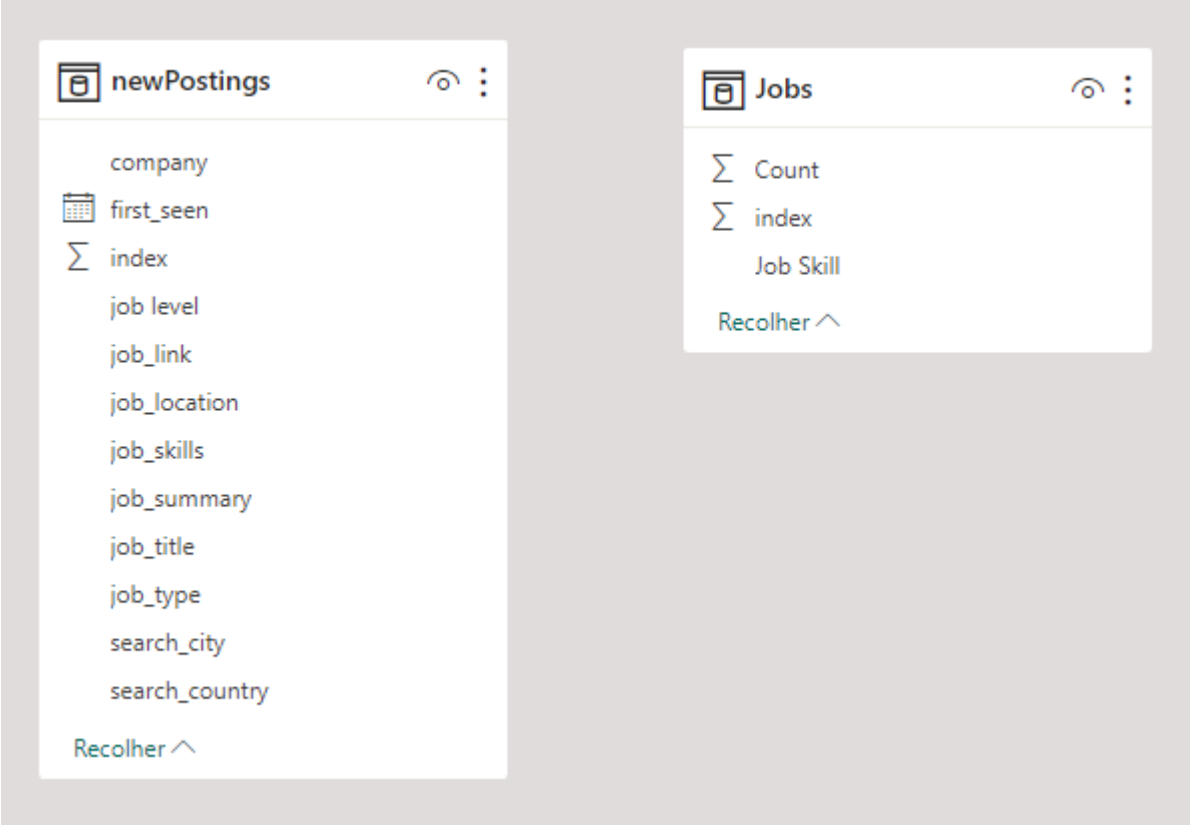

Figura 1: Exibição de modelo

Eu tentei fazer tratamento do título com Processamento de Linguagem Natural utilizando o pacote nltk e re, entretanto o resultado não ficou compatível para uso no Power BI, então eu o mantive no Jupyter para acrescentar à análise.

Análise e Insights

Todas as imagens abaixo foram criadas no Power BI e portanto são interativas caso baixe o dashboard disponível no github.

Eu criei ao todo 3 dashboards, segue o índice abaixo:

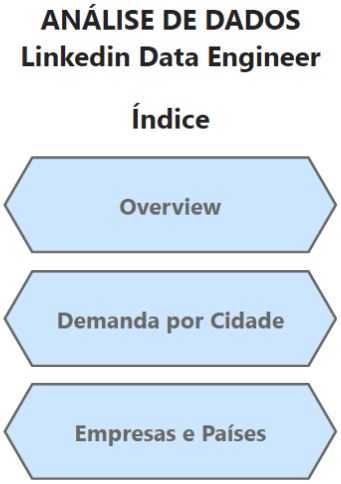

Figura 2: Índice

A imagem abaixo mostra uma visão geral do conjunto de dados.

Figura 3: Overview

O próximo dashboard utiliza apenas 1 recurso, mas não é viável de utilizar com vários outros recursos, então preferi deixar uma página inteiramente para ele.

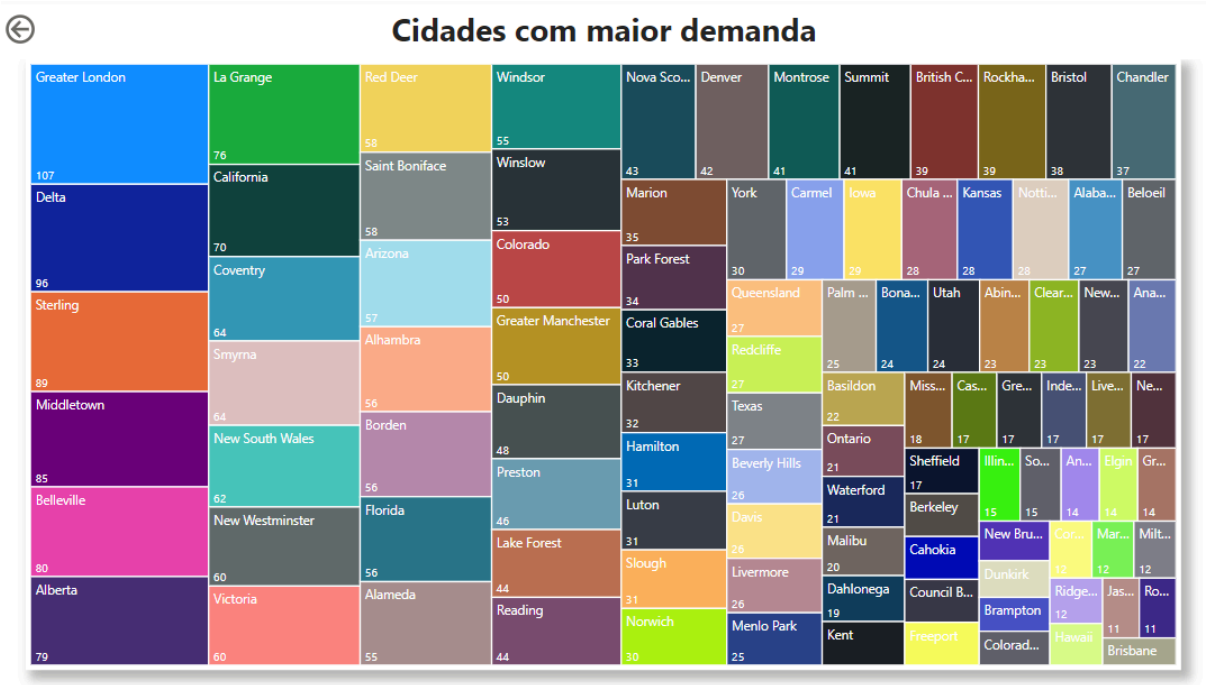

Figura 4: Cidades com maior demanda

O último dashboard junta informações referentes das empresas e dos países, eu escolhi manter os 2 juntos pois eles não sobrecarregam o usuário de informação, diferente da imagem anterior.

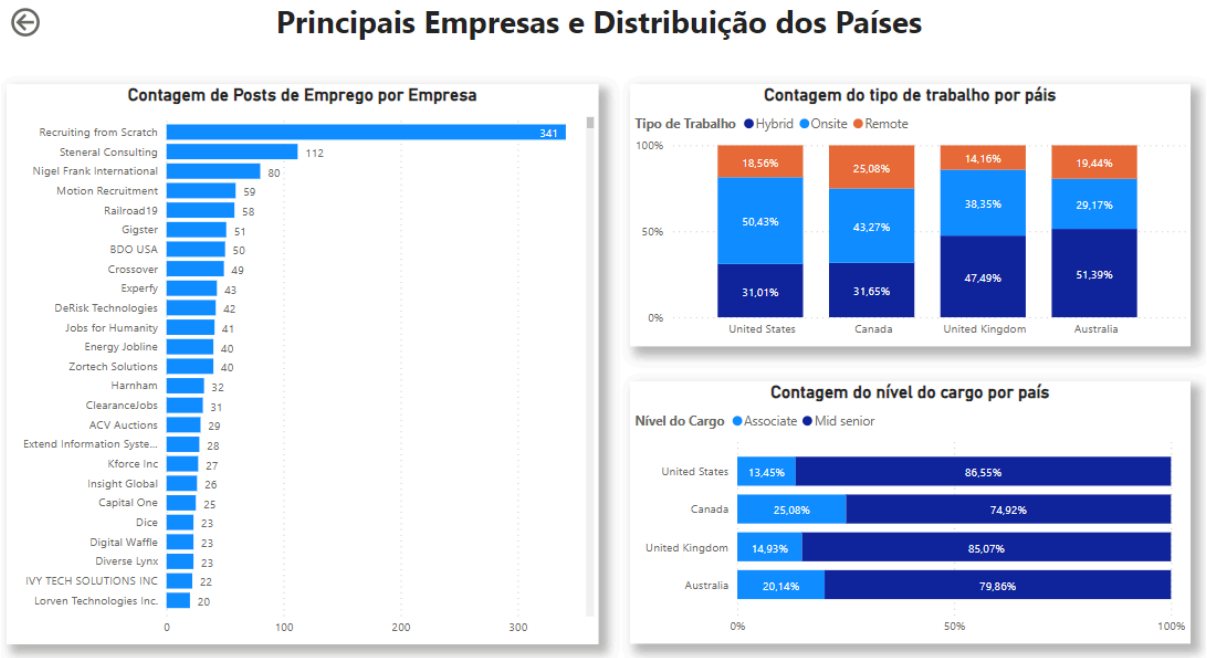

Figura 5: Principais empresas e distribuição dos países

Considerações

Inicialmente quando eu comecei a análise exploratória, eu imaginei que o foco desse projeto estaria principalmente na criação de gráficos, pois os dados pareciam já bem preparados. Todavia eu me enganei, trabalhar com linguagem natural é realmente complicado, entretanto consegui achar maneiras de lidar com esse problema para extrair o máximo de insights do dataset.

Eu tive a oportunidade de estudar mais a respeito de NPL e consegui treinar o uso de Regex, foi cogitado o uso de machine learning para fazer agrupação de textos, entretanto dentro do tempo não foi viável, porém consegui encontrar um rumo para tentar aplicar essa técnica em futuros projetos.