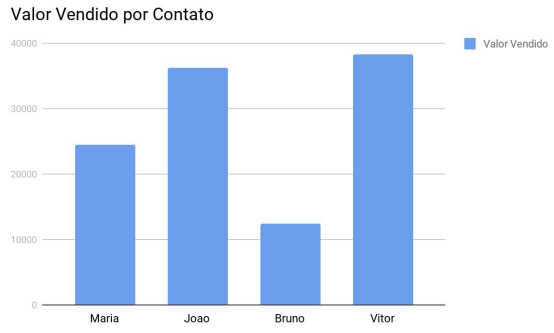
DESAFIO ESTÁGIO DATA ENGINEERING

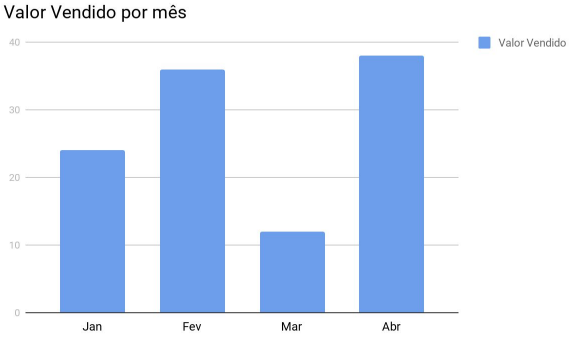
Resumo:

Junto a esse arquivo no pacote zip encontram-se 4 outros arquivos no formato tsv. Esses arquivos representam dados hipotéticos exportados de uma aplicação CRM para empresas B2B. Esses arquivos representam negócios fechados(deals.tsv), pessoas dentro das empresas compradoras (contacts.tsv), empresas compradoras (companies.tsv) e um arquivo mapeando código de setor para nome do setor da empresa (sectors.tsv). O arquivo de negócios possui um campo *contactId* que referencia uma linha no arquivo de contatos e *companyId* que referencia uma linha no arquivo de companhias. No arquivo de empresas existe um *sectorId* que referencia uma linha no arquivo de setores.

A sua tarefa é escrever uma pequena ETL usando esses arquivos de input e gerando dois outputs.

O primeiro deve conseguir servir de base para 2 gráficos: número de vendas por contato e valor total vendido por mês. Segue abaixo imagens exemplificando os gráficos.





O segundo deve ser uma lista dos setores de empresa, ordenado por quanto esse setor representa no total vendido pela empresa no mês. Por exemplo, considerando que a empresa vendeu 10k total, se o 8k foi vendido para empresas do setor 1 e 2k para empresas do setor 2 então a lista resultante seria:

1 Bens De Consumo 0.8
2 Serviços 0.2

Note que o nome do setor deve constar na lista.

O formato do outputs fica a seu critério podendo ser um outro arquivo csv ou uma tabela em um banco de dados relacional por exemplo. O código pode ser escrito em qualquer linguagem.

Observações Importantes:

- O código deve poder ser executado N vezes gerando sempre o mesmo output para o mesmo input(Idempotente).
- Os arquivos que enviados para o desafio são extremamente pequenos, mas imagine que esses arquivos podem ter vários gigabytes em um cenário real, pense em um código que funcionaria nesses casos.
- Podem existir linhas com valores inválidos(com encodings alternativos por exemplo), essas linhas devem ser descartadas e o número de linhas descartadas deve ser informado.
- É mandatório um readme explicando como configurar ambiente e executar o programa. Dockerizar o ambiente para execução será apreciado mas não é obrigatório.
- A entrega pode ser feita em repositório git aberto ou um pacote comprimido(zip, tar..)

Prazo de Entrega:
O prazo de entrega é de 1 semana.

Boa Sorte!