

Propósito

Este desafio tem como objetivo avaliar sua forma de resolver problemas. Ao solucioná-lo, você nos mostrará:

- Sua capacidade de articular questões de negócios em código
- Sua capacidade de extrair dados de uma fonte, processá-los e criar algumas visualizações para esses dados

Sobre o Desafio

O [Stack Overflow](#) é uma plataforma amplamente conhecida na comunidade de tecnologia e permite que usuários façam perguntas e também as respondam. Mas além disso podem, através do registro e da participação ativa, votar em questões e respostas mais ou menos úteis. Provavelmente você já acessou ele para sanar dúvidas de código que tinha.

Todo ano o Stack Overflow faz uma pesquisa com sua comunidade de desenvolvedores sobre vários temas, que vão desde as suas preferências tecnológicas até questões profissionais. E nós do Labs estamos super curiosos para saber o que os desenvolvedores andam falando por aí. Queremos saber quais tecnologias usam, como se comunicam, quanto ganham em média, onde moram e mais algumas coisas.

Seu desafio é nos ajudar a responder essas perguntas usando os resultados da pesquisa aplicada em janeiro de 2018. Dividimos o desafio em duas partes principais:

1. Ler os dados e aplicar algumas limpezas e regras de negócio no mesmo
2. Criar gráficos que nos ajudem a responder de forma visual a nossas questões

Montagem do banco de dados

Nós te daremos um arquivo de texto (formato CSV) contendo uma parte dos resultados da pesquisa realizada pelo Stack Overflow e outro arquivo texto (formato CSV) contendo a descrição das colunas de respostas presentes no primeiro arquivo (ou seja, ele te fala quais perguntas foram feitas e que geraram as respostas).

Você vai usar uma linguagem de programação para ler esse arquivo, processar ele de acordo com as regras de negócio descritas abaixo e depois criar gráficos usando uma biblioteca de sua escolha (vide seção [Stack de Tecnologias](#)).

Fonte de dados

Você encontrará no anexo desse projeto dois arquivos, o primeiro dos quais contém uma amostra de apenas 10 mil linhas de respostas à pesquisa, e o segundo, uma explicação do

significado das colunas de respostas. O primeiro arquivo se chama `base_de_respostas_10k_amostra.csv` e o segundo, `base_de_conhecimento.csv`

Caso queira ver os resultados completos da pesquisa, basta acessar esse [link do Kaggle](#).

Regras de negócio

- As colunas *Hobby* e *OpenSource* devem ter seu conteúdo traduzido para português, ou seja, em todas as linhas que contém o valor “Yes” ele deve ser substituído por “Sim”.
- A coluna *ConvertedSalary* não deve conter valores vazios. Precisamos preencher valores vazios com a média de salário dos demais respondentes que estejam no mesmo país. Caso não haja outras pessoas no mesmo país, o salário deve ser preenchido com zero.
- Salário deve ser sempre calculado em dólar e mensal. Para esse cálculo você usará a coluna *ConvertedSalary*, que contém o salário anual. Crie uma coluna nova chamada *MonthlySalary* que vai conter esse valor.
- Precisamos calcular também o salário do respondente em reais. Considere que 1 dólar equivale a R\$3,81. Crie uma coluna nova chamada *BrazilMonthlySalary* que vai conter esse valor
- Todos os valores “NA” das colunas não-numéricas devem ser alterados por “Unknown”

Criação dos gráficos

Com os dados limpos e padronizados, você poderá construir alguns gráficos e matar nossa curiosidade. A lista abaixo contém tudo que precisamos visualizar:

- ☐ Um histograma dos salários (em reais) dos respondentes brasileiros e de outros dois países de sua escolha. (use a coluna *BrazilMonthlySalary* para isso)
- ☐ Um gráfico de barras contendo a média de salários (em reais) de cada país. (use a coluna *BrazilMonthlySalary* para isso)
- ☐ Um gráfico de barras empilhadas contendo a quantidade de usuários que usam cada linguagem de programação (eixo Y) por sistema operacional (eixo X). (use a coluna *LanguageWorkedWith* para isso)
- ☐ Um gráfico de barras com a distribuição da quantidade de usuários que usam cada tipo de IDE
- ☐ Precisamos saber se a média salarial (*BrazilMonthlySalary*) das pessoas que tiveram educação formal de mais alto grau (Bacharelado - *Bachelor degree*, Mestrado - *Master degree* e Doutorado - *Other doctoral degree*) é maior que a média do todo em pelo menos três países. Você é livre para escolher quais países deseja mostrar e qual o melhor gráfico a ser construído.
- ☐ Precisamos saber se existe algum tipo de relação entre a satisfação no trabalho (*JobSatisfaction*) e o status da busca por trabalho (*JobSearchStatus*). Você é livre para escolher alguma visualização de dados (seja ela tabela ou não) que nos demonstre isso. Aproveite a oportunidade e discorra em algumas frases suas conclusões.

- ❑ Precisamos saber se existe algum tipo de relação entre a satisfação na carreira (*CareerSatisfaction*) e o tempo de permanência em um emprego (*LastNewJob*). Você é livre para escolher alguma visualização de dados (seja ela tabela ou não) que nos demonstre isso. Aproveite a oportunidade e discorra em algumas frases suas conclusões.
- ❑ Queremos saber o quanto percentualmente a média de salário dos Estados Unidos é maior ou menor que a média de salário dos demais países.
- ❑ Qual o percentual de usuários querem usar uma linguagem de programação nova (*LanguageDesireNextYear*) em relação ao que já sabe (*LanguageWorkedWith*). Faça essa mesma análise para uma das seguintes categorias: *Database*, *Platform* ou *Framework*.
- ❑ Que outra ideia ou descoberta você fez ao estudar esses dados? Nos conte, através de uma visualização de dados (seja ela tabela ou gráficos), alguma coisa que não tenhamos pensado nos itens anteriores

Stack de tecnologias

O time do Labs é muito diversificado em termos de tecnologia. Para esse projeto especificamente, selecionamos a seguinte *stack* de tecnologias para você usar:

- **Linguagem de programação:** qualquer linguagem gratuita (sugestões: Python com a biblioteca Pandas, R ou Octave)
- **Biblioteca para criação de gráficos:** sugerimos que use as bibliotecas padrão da linguagem selecionada (ex: Matplotlib caso use o Python ou Ggplot caso use o R)

Sabemos que há muito código disponível na internet e que muitas vezes eles nos ajudam a resolver desafios que enfrentamos ao desenvolver projetos. ***But, be careful!*** Queremos conhecer bastante o código que você é capaz de desenvolver, usando sua capacidade analítica e criatividade. Não esperamos um código específico para resolver esse desafio, mas sim que ele reflita seus conhecimentos.

Fique a vontade caso queira utilizar alguma IDE para otimizar seu tempo de trabalho. É interessante apenas que você deixe isso claro no seu relatório final.

O que esperamos ver ao final?

Nosso time está curioso para ver o seu projeto. Esperamos que seu entregável final contenha os seguintes itens:

1. Um arquivo de introdução nos explicando a visão geral do seu projeto e quais tecnologias utilizou (ex: Pandas 0.22.0)
2. Imagens dos gráficos solicitados
3. Os arquivos com o seu código utilizado para manipular os dados e construir os gráficos

Caso você tenha dificuldade de finalizar o seu projeto, nós encorajamos fortemente que nos envie a sua evolução (código, descrição de como resolveria o problema, imagens, etc.)

BOA SORTE! =)