1. Até que ponto os problemas e cálculos exigidos se assemelham a situações reais do ambiente corporativo? Por exemplo, em uma empresa, como uma comparação de comportamento de consumo pré e pós pandemia seria exigido e com que recursos e informações vocês normalmente teriam?

Se assemelham bastante!! Os desafios foram feitos a partir de situacoes/ analises reais que como analistas ou cientistas de dados lidamos no dia a dia.

Por exemplo no caso 1: As equipes de marketing realmente estao tendo que se adaptar no pós pandemia, e geralmente o setor responsável por analisar diferencas de comportamento do cliente é o dados.

No caso 2: Em toda empresa de logistica, existe uma área de planejamento de demanda ou estoque. Profissionais de dados que trabalham nessas áreas lidam com essas informacoes do challenge todos os dias em uma escala muito maior (Para diferentes produtos, diferentes niveis de servico, centros de distribuicao e principalmente diferentes distribuicoes de dados, como poisson).

2. No primeiro desafio fiz um groupby pela coluna pre-pandemia e apliquei um describe nesses dados, gostaria de confirmar por favor se esse procedimento está correto.

Nesse caso, o procedimento correto seria na verdade utilizar a coluna com o indicativo se o dado é pre-pandemia ou pos como filtro. Pois ao agrupar voce perderia informacoes, iria ter somas, medias, etc. Mas nao teria mais o dado cru para trabalhar.

3. No primeiro desafio seria uma boa prática construir tabelas de frequência para as variáveis qualitativas? Seria uma excelente pratica!

4. Qual biblioteca mais usada para os gráficos (matplotlib ou seaborn)? As duas sao muito utilizadas. Eu particularmente prefiro seaborn, mas para alguns graficos específicos só temos no matplotlib, por exemplo o gráfico de pizza.

5. É possível fazer um crosstab com os dados de produtos e região BR antes e depois da pandemia?
Eu particularmente prefiro pivot_table a crosstab, entao nao uso muito no meu dia a dia. Mas acho q seria possivel. Ou voce poderia dividir em duas tabelas.

6. Como eu adiciono Filtro no crosstab? Por exemplo:

```
#Agrupando os produtos em pré e pós pandemia
labels = {0: 'Pós Pandemia', 1: 'Pré Pandemia'}

produtos_pandemia = pd.crosstab(
   dataset['produto'],
   dataset['pre-pandemia'],
   rownames = ['Produtos'],
   colnames = ['Fase']
)
```

Se eu quisesse selecionar esses dados apenas para uma regiao, como eu faria?

Essa pergunta é bem interessante! vc poderia utilizar o filtro no próprio data frame. por exemplo, ao inves de chamar df['produto'], vc chamaria df.loc[df['pre-pandemia']==0]['produto']. Esse conceito em python é chamado de mascara ou mask em ingles.

7. Tive dificuldades em realizar os cálculos sobre tamanho da amostra e também tive dificuldades em definir se uma base está na forma normal.

Entendo a dificuldade, realmente existem diferentes formulas de tamanho da amostra. Na solucao do challenge, fiz a proposicao de duas delas, mas a mais simples seria a fórmula de Slovin

Sobre como definir se uma base / dado está na forma normal ou nao. Isso pode ser feito de diferentes formas e qualquer uma delas está correta vou listar algumas delas aqui.

1. Análise visual dos dados: Plotar o histograma e verificar se ele tem forma de sino, é simétrico ao redor da média,
2. Analise dos gráficos QQ: os graficos QQ, comparam os quartis da distribuicao que voce está analisando em questao com os quartis da distribuicao normal, quanto mais proximos eles forem, mais o seu dado se assemelha a uma normal.
3. Outra possibilidade seria fazer um teste de hipóteses, como o KS. Nesse caso o teste vai analisar o quao distante a curva de densidade cumulativa do seu dado é da da normal.

8. Para o desafio 2, quando eu analisei o comportamento dos dados de lead-time, pelo histograma os dados aparentam ter um comportamento normal, porém pelo Qqplot e teste de Shapiro eles não aparentam ter um comportamento normal, porém este comportamento pode ocorrer porque os dados de lead-time fazem parte do conjunto de número naturais e a distribuição normal existe para os números reais. Fiquei com dúvida se para este caso podemos considerar os dados de lead-time como normais ou não?
É normal sim!! é que nesse caso o lead time é uma variável discreta. Ele é o numero de dias que o produto demora para chegar na loja. Entao para essa situacao, é mais recomendável a analise visual do que o teste.

9. Quais as melhores maneiras de filtrar os dados de um enunciado em busca de palavras-chave pra elaboração de dados estatísticos pertinentes?
Nos exemplos dados na Alura, o desvio padrão, a variância, o erro, e outros componentes estatísticos estavam explícitos. Senti dificuldade de identificar esses componentes nos enunciados pra criar resultados relevantes (ou ter certeza que criei os dados certos).

A ideia do challenge é voce ter a oportunidade de fazer um desafio, uma resolucao de case, como os que aparecem em entrevista, e muitas vezes, infelizmente isso vem pouco explicito. Recomendo nesse caso especifico:

1. Focar em entender o problema de negócio que vem sendo perguntado. O que o problema quer resolver?
2. Quais sao as variaveis que resolvem esse problema? E assim decidir o que utilizar.

10. fiquei com uma dúvida no segundo exercício na letra b, seria o ideal mostrar a diferença no perfil dos consumidores por meio de gráficos ou valores.
A ideia seria mais analisar os dados que voce ja plotou anteriormente. Entao nesse caso aqui, poderia comentar sobre os graficos / valores encontrados.

11. Seria possível responder o item 2 do desafio 1 sem utilizar o teste de hipótese? Quais seriam as outras opções para provar a diferença estatisticamente?
Dica: Sempre que te perguntarem se é estatisticamente relevante é via teste de hipóteses. Por isso, a solucao seria explicar que seria necessário utilizar um teste de hipóteses. Mas a pergunta era opcional

12.Tive muita dificuldade em calcular o estoque de segurança. A minha principal dúvida foi se no cálculo eu deveria considerar a média e desvio padrão da demanda e lead-time na base mensal ou diária, os dados estão na semanal. Achei o meu resultado muito elevado. Poderia solucionar o exercício na aula ao vivo?

Vamos resolver juntas na aula :) . Mas, voce pode escolher o que preferir. Geralmente, se o dado está semanal, voce poderia dividir o lead time por 7 para ficar na mesma escala, e devolver os resultados semanais. Ou vc poderia agregar de outra forma também todas as formas estariam corretas.

13. Quando é viável trocar um tipo de dado inteiro para booleano . No caso do campo pré-pandemia por exemplo . Teria alguma vantagem ? Tanto faz nesse caso, o campo, de qualquer maneira somente seria utilizado como filtro. Entao nao precisaria mudá-lo.

14. Qual o tipo de dado para id_cliente? O id cliente era somente um indicativo que a base esta na dimensao cliente, mas no enunciado nao foi pedido para analisálo.

15. Entendo a variável renda como quantitativa continua, porém no arquivo o dado é do tipo inteiro. Isso influencia nessa análise? Nao influencia.

16. Qual a melhor maneira de analisar a idade ? Visto que em uma base de dados a distribuição costuma ser bem grande. (Dividi em classes será que esta correto? Quando fui fazer o gráfico não consegui usar com as classes que fiz) Voce poderia categoriza-la e mostrar as frequencias, contagens das categorias ou trabalhar com ela como inteiro também. Nesse caso sem problemas, mas por exemplo, se voce fosse utilizar essa variável em algum modelo, aí provavelmente precisaria sim categoriza-la.

17. Gostaria de entender melhor sobre as provas estatísticas Acredito que será o proximo tópico da alura, por isso a pergunta era opcional/extra ;)