Desafio Final

A *We Shop* é uma grande rede atacadista, localizada em Santa Catarina, especializada em atender estabelecimentos como panificadores, pousadas, restaurantes, cantinas, bares, supermercados e similares desde 2017. A empresa oferecem uma variedade de mais de 2.700 itens de produtos em seu portfólio. Além disso, estamos sempre atentos às novidades do mercado para oferecer sempre um atendimento diferenciado.

DESAFIO

O ano é 2023 e você foi contratado como Cientista de Dados Júnior pela *We Shop* para compor a squad de Data Insights. Na sua primeira semana de trabalho, seu tech lead solicitou para que você desenvolvesse algumas análises exploratórias com o objetivo de encontrar algumas oportunidades de negócio. A *We Shop* deseja entrar no e-commerce e para isso precisa que suas análises consigam responder as perguntas mais comuns da área (listadas a seguir).

Base de dados

A base de dados disponibilizada corresponde a quatro arquivos em Excel (estão na pasta Desafio):

Dimensões.xlsx: possui os dados que compõem a dimensão do modelo. As tabelas estão separadas em:

- **dCliente:** representa o cadastro dos clientes bem como todos os seus atributos como Categoria, Cidade, UF e país.

- **dProduto:** representa o cadastro de produtos e possui o grupo e a linha que cada produto pertence.
- **dVendedor:** representa o cadastro de vendedores. Cada vendedor possui um Supervisor e um Gerente.

fMetas.xlsx: representa a tabela fato com todas as metas de vendas por cdVendedor e mês/ano. Essa planilha costuma ser preenchida pelos supervisores após reuniões mensais com os gerentes.

Vendas 2017.xlsx, Vendas 2018.xlsx, Vendas 2019.xlsx, Vendas 2020.xlsx, Vendas 2021.xlsx e Vendas 2022.xlsx: representa a tabela fato com todos os registros de vendas ocorridos entre Jan/2017 e Mar/2022. A extração foi feita de forma separada pelo sistema (por ano). Os dados da coluna Peso estão na unidade Kg, mas o cliente deseja obter o total de peso por tonelada.

1. **Análise Exploratória**

- Valor total vendido e Valor total da meta por ano?
- Crie um gráfico de linhas para exibir o comportamento das vendas ao longo do ano de 2021.
- Quantos clientes ativos e inativos temos em nossa empresa?
- Qual a linha de produtos mais vendida?
- Qual foi a quantidade total de produtos vendidos por ano?
- Qual foi o total de peso em toneladas dos produtos vendidos em todo o período?
- Comparando a performance por gerente, qual dos gerentes teve o maior faturamento? Qual o percentual relativo a esse faturamento?
- Qual equipe teve a melhor performance de vendas?

2. **Crie relatórios por Gerente, Supervisor e Vendedor. Faça um script para criar uma pasta com o nome Relatórios, e dentro dessa pasta deve ser criado pastas com nomes conforme o tipo de relatório.**

Machine Learning

Parabéns! Devido ao seu excelente trabalho, seu líder solicitou também que você desenvolvesse um algoritmo de Machine Learning para com base em alguns dados de uma empresa americana de comércio eletrônico. Este estudo será utilizado em algumas reuniões de benchmarkings, e servirá como base para a construção das estratégias de vendas para o e-commerce da We Shop. A empresa analisada vende roupas on-line, mas eles também oferecem sessões de aconselhamento sobre estilo e roupas na loja. Os clientes entram na loja, têm sessões/reuniões com um personal stylist, depois podem ir para casa e encomendar através de um aplicativo móvel ou site as roupas que desejam.

A empresa está tentando decidir se concentrará seus esforços na experiência do aplicativo móvel ou no site.

Eles contam com você para ajudá-los a descobrir! Vamos começar!

Basta seguir os passos abaixo para analisar os dados do cliente (são falsos, não se preocupe, não forneci números de cartão de crédito ou e-mails reais).

1. Importe as bibliotecas Pandas, Numpy, Matplotlib, Seaborn e Scikitlearn;

Obtenha os dados

Trabalharemos com o arquivo csv de clientes de comércio eletrônico da empresa. Possui informações do cliente, como e-mail, endereço e avatar colorido. Depois também possui colunas de valores numéricos:

- **Média Duração da sessão:** sessão média de sessões de aconselhamento de estilo na loja.
- **Tempo no aplicativo:** tempo médio gasto no aplicativo em minutos
- **Tempo no site:** tempo médio gasto no site em minutos
- **Duração da associação:** há quantos anos o cliente é membro.

2. Abra o arquivo Ecommerce Customers.csv;
3. Verifique algumas informações dos clientes, utilize os métodos head, tail, info e describe.

Análise Exploratória

Vamos explorar os dados!

No restante do exercício usaremos apenas os dados numéricos do arquivo csv.

4. Use o seaborn para criar do tipo pairplot para comparar as colunas Tempo no site e Valor gasto anualmente. A correlação faz sentido?

```
sns.set_style('whitegrid')
```

```
# More time on site, more money spent.
sns.jointplot(x='Time on Website',y='Yearly Amount Spent',data=customers)
```

5. Faça o mesmo, mas com a coluna Tempo no aplicativo.
6. Use jointplot para criar um gráfico hexadecimal 2D comparando o tempo no aplicativo e a duração da associação.

```
sns.jointplot(x='Time on App',y='Length of Membership',kind='hex',data=customers)
```

7. Vamos explorar esses tipos de relacionamento em todo o conjunto de dados. Use pairplot para recriar o gráfico abaixo. Com base neste gráfico, qual parece ser a variável mais correlacionada com o valor gasto anualmente?

```
sns.pairplot(customers)
```

8. Crie um gráfico linear model plot (usando lmplot do seaborn) do valor gasto anualmente versus duração da associação.

```
sns.lmplot(x='Length of Membership',y='Yearly Amount Spent',data=customers)
```

DADOS DE TREINAMENTO E TESTE

```
y = customers['Yearly Amount Spent']
```

```
X = customers[['Avg. Session Length', 'Time on App','Time on Website', 'Length of Membership']]
```

9. Use model_selection.train_test_split do sklearn para dividir os dados em conjuntos de treinamento e teste. Defina test_size=0,3 e random_state=101.

TREINANDO O MODELO

Agora é hora de treinar nosso modelo em nossos dados de treinamento.

10. Crie uma instância de um modelo LinearRegression() chamado lm.
11. Treinar o modelo com os dados de treinamento.
12. Imprima os coeficientes do modelo.

PREVISÃO DE DADOS DE TESTE

Agora que ajustamos nosso modelo, vamos avaliar seu desempenho prevendo os valores de teste!

13. Use lm.predict() para prever o conjunto de dados X_test.
14. Crie um gráfico de dispersão dos valores reais do teste versus os valores previstos.

AVALIANDO O MODELO

Vamos avaliar o desempenho do nosso modelo calculando a soma residual dos quadrados e a pontuação de variância explicada (R^2).

15. Calcule o erro médio absoluto, o erro quadrático médio e a raiz do erro quadrático médio. Consulte na internet sobre essas fórmulas.

```
from sklearn import metrics

print('MAE:', metrics.mean_absolute_error(y_test, predictions))
print('MSE:', metrics.mean_squared_error(y_test, predictions))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, predictions)))
```

RESÍDUOS

Você deve ter conseguido um modelo muito bom com um bom ajuste. Vamos explorar rapidamente os resíduos para ter certeza de que está tudo bem com nossos dados.

Trace um histograma dos resíduos e certifique-se de que parece distribuído normalmente. Use distplot do seaborn.

```
sns.distplot((y_test-predictions),bins=50);
```