**Projeto Final**

**Exposição do problema**

Com o passar dos anos e o avanço da tecnologia muitas ferramentas surgem e/ou são aperfeiçoadas e com isso aparece a necessidade de atualizar-se tanto no que diz respeito ao manuseio de tais ferramentas quanto na sua aplicação no mundo coorporativo e dos negócios.

Um ramo científico que vem crescendo com o tempo é o da Ciência de Dados. Esta, juntamente com o processo de Análise de Dados servem como base de apoio para entendimento do *business* em questão e para futura tomada de decisões que venham a impactar (positiva ou negativamente) o futuro de uma empresa, comércio, etc.

Dentre os muitos exemplos sobre aplicação de Ciência de Dados, podemos citar a segmentação de clientes de uma empresa, assunto este que iremos abordar em detalhes neste trabalho. Neste contexto, afim de aumentar as vendas ou verificar o engajamento do público com relação aos serviços oferecidos, é fundamental entender o ponto de vista dos consumidores. O que eles compram, seu nível de satisfação com o serviço oferecido, o total de receita gerado pelas transações, etc. Ao responder estas e outras perguntas é possível traçar um perfil bem preciso dos clientes e posteriormente tomar decisões para fidelizá-los, como oferecer promoções, enviar propaganda ou anúncios específicos, entre outros.

Neste trabalho iremos realizar a citada segmentação de clientes e para isso usaremos um modelo denominado **RFM** (do inglês *recency*, *frequency*, *monetary*) onde nosso objetivo principal é agrupar os clientes de uma empresa de acordo com as variáveis ditas acima. Além disso, exploraremos as compras para tirar informações sobre os departamentos que mais lograram êxito nas vendas. Por fim, iremos criar um modelo para tentar prever quanto dinheiro será ganho nos próximos 30 dias de acordo com vendas futuras.

**Coleta ou Importação dos dados**

É fato que qualquer empresa, até para próprio controle interno, possui dados históricos acerca de suas operações e estes dados podem (e devem) ser explorados no sentido de trazer *insights* sobre o que realmente está acontecendo e o que se pode fazer a partir daí para seu crescimento.

Para nosso trabalho de segmentação de clientes iremos fazer uso da chamada coleta de dados secundária, que consiste em coletar estes dados já disponibilizado previamente. Neste caso, iremos utilizar uma base de dados pública da Olist, publicada na plataforma de ciência de dados *Kaggle* e que pode ser conferida em: [*https://www.kaggle.com/olistbr/brazilian-ecommerce*](https://www.kaggle.com/olistbr/brazilian-ecommerce).

Esta base de dados consiste em 8 (oito) arquivos no formato *.csv* mas para os propósitos deste trabalho utilizamos apenas 4 (quatro) cujos conteúdos são:

* olist\_customers\_dataset.csv: contém informações dos clientes como seu id único e dados de localização (como cidade, estado e cep);
* olist\_order\_items\_dataset.csv: apresenta informações dos itens que compõem cada ordem de compra do cliente, como id do produto, preço, frete etc;
* olist\_orders\_dataset.csv: possui a ordem de cada compra. Aqui encontramos informações como data da compra, data de aprovação do pedido e data da entrega.
* olist\_products\_dataset.csv: possui descrições do produto como número de fotos no anúncio, peso, dimensões, categoria, entre outros.

A partir destes *datasets* foi possível uni-los em um só para realizar todas a análises necessárias.

Biblioteca(s) utilizada(s) nesta etapa:

* Pandas

**Preparação dos dados**

Como era de se esperar, os dados obtidos via Kaggle não vieram perfeitos e nem seque apropriados para sua pronta utilização. Desta forma foi necessária uma grande limpeza nestes dados que incluiu:

* Eliminação de colunas indesejadas.
* Conversão dos tipos das colunas para os formatos adequados (como *datetime* por exemplo).
* Agregação de colunas para obter uma contagem ou soma de seus valores.
* Remoção de outliers

Uma importante ressalva é a de que, embora os *datasets* apresentassem valores *missing* (faltantes), não foi necessário nenhum tipo de tratamento adicional uma vez que tais problemas foram verificados nas colunas que posteriormente foram excluídas do estudo.

Biblioteca(s) utilizada(s) nesta etapa:

* Pandas
* Timedelta (via datetime)

**Modelo RFM**

Para o processo de segmentação de clientes utilizamos o famoso e consolidado Modelo RFM, que leva em consideração a **recência**, a **frequência** e a **monetaridade** de cada cliente. Temos que as definições destes termos são dadas por:

* Recência: Tempo (em dias) que o cliente não realiza uma compra.
* Frequência: Número de compras realizadas.
* Monetaridade: Quantidade de dinheiro gasto pelo cliente.

Diante deste panorama, identificamos os seguintes pesos de importância para cada uma das variáveis:

* Recência: Quanto menor, melhor (menos dias decorridos desde a última compra).
* Frequência: Quanto maior, melhor (mais vezes o cliente efetuou uma compra).
* Monetaridade: Quanto maior, melhor (mais dinheiro o cliente gastou).

As situações opostas a estas representam um menor peso de importância.