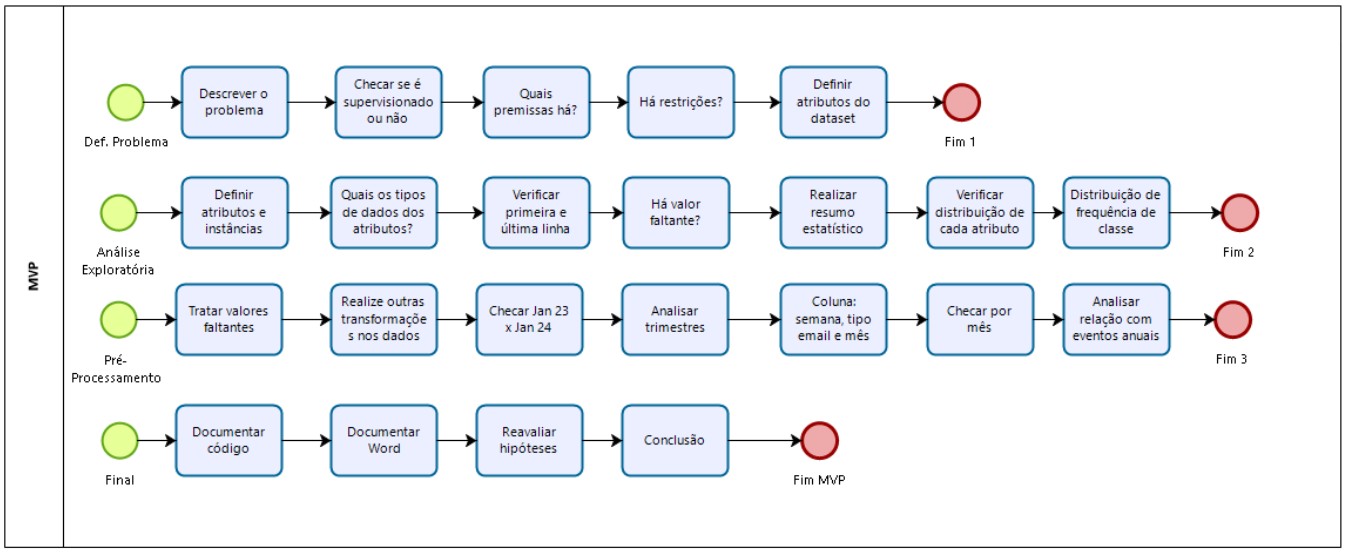
PLANEJAMENTO MVP – E-Commerce

ETAPAS: Definir o problema, hipóteses e descrever a base, realizar a análise exploratória dos dados, realizar o pré-processamento e por fim analisar tudo e concluir.

As etapas do meu MVP podem ser analisadas no fluxo a seguir:



1. Definição do Problema

Encaro como meu problema principal a identificação acurada dos impactos da sazonalidade nas vendas de livro de uma livraria qualquer. Para analisar esses possíveis impactos, irei tratar um dataset com análises exploratórias e pré-processamentos.

A base contém dados de venda de compradores de uma livraria. Irei enunciar o objetivo e hipóteses a seguir.

OBJETIVO: Realizar uma análise em cima do dataset de venda de livros para entender melhor o relacionamento das vendas com suas épocas do ano, assim como o domínio de cada e-mail.

HIPÓTESES:

- Há maior venda de livros no último trimestre do ano provavelmente devido ao Natal;
- Há maior venda de livros no primeiro trimestre do ano devido à retomada das aulas;
- Há maior venda de livros em janeiro de 2023 do que em 2024 devido à proximidade com a pandemia;
- Há maior venda de livros aos sábados;
- Segundo o Veja, Nordeste é a região que mais lê, portanto o estado com maior registro de compra é de lá;

COLETA DE DADOS: Conjuntos de dados do Kaggle, como "Initial_Data".

Atributos e suas instâncias: **CustomerID** (identificador único para cada cliente, numerado de 1 a 2000), **PurchaseDate** (data em que foi feita cada compra de cada cliente, formato date), **Name** (Nome do comprador), **Email** (Email do comprador, formato nome@example.com), **PhoneNumber** (Numero de contato do comprador, formato (551)626-0650x4340), **Address** (Endereço do cliente), **PurchaseAmount** (Total da Venda, em decimal), **BookTitle** (Nome de livro comprado).

2. Análise Exploratória

Na análise exploratória chequei os atributos e suas instâncias, que são diversas e variadas. Os tipos de dados são definidos como: **CustomerID** (int), **PurchaseDate** (date), **Name** (string), **Email** (string), **PhoneNumber** (string), **Address** (string), **PurchaseAmount** (decimal), **BookTitle** (string).

Isso corresponde às colunas já existentes. Mais para frente haverá novas com diferentes datatypes. Olhando a primeira linha e última do dataset, é possível verificar que há valores nulos em algumas colunas, assim como o nome das colunas é duplicado como uma primeira linha, sem necessidade. É um ponto a ser tratado.

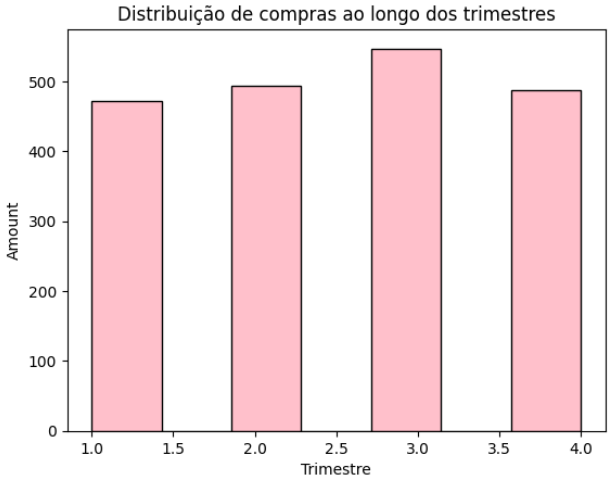
Analisando a coluna email, vemos que o domínio do email é listado, para todos, como @example, o que invalida um de meus objetivos principais que era analisar o domínio de email que mais realiza compra de livros. Isso impossibilitaria possíveis tomadas de decisão que envolvam realizar publicidade com a empresa de domínio do email.

3. Pré-Processamento

Para a etapa de pré-processamento decidi tratar os valores nulos, transformando-os e não os eliminando. Foi necessário, também, outras transformações em datatypes e criação de novas colunas (Dia, Mês, Ano, Dia_da_Semana, Trimestre, DDD) para validação de hipóteses e análise aprofundada do dataset.

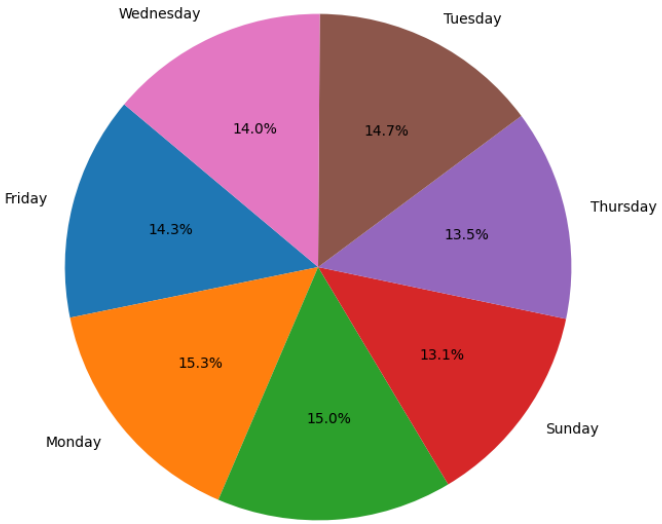
Analisando a coluna de Ano, vemos que o único ano existente é 2023, ou seja, a hipótese “Há maior venda de livros em janeiro de 2023 do que em 2024 devido à proximidade com a pandemia” será descartada visto que não há dados que nos possibilitem comprová-la ou refutá-la.

Para a hipótese “Há maior venda de livros no último trimestre do ano provavelmente devido ao Natal” vemos que o trimestre com maior número de vendas não é o último (próximo ao Natal), inclusive este valor reduz para o último trimestre.



Não há relação também com o primeiro trimestre do ano ter um aumento devido à retomada das aulas, pois não há uma grande quantidade de vendas.

Analisei, também, os dias da semana com maior venda para validar a hipótese “Há maior venda de livros aos sábados”. Vemos que sábado está no top 2 com 15% do total de vendas.



Já para analisar a hipótese “Segundo o Veja, Nordeste é a região que mais lê, portanto o estado com maior registro de compra é de lá” vimos que o DDD com mais venda é um criado, 581. Este não se relaciona com nenhuma região do Brasil.

Para maior entendimento, analisar o *notebook* com os códigos, anotações e gráficos.

4. Conclusão

Com este trabalho, conclui-se que, a base é extremamente fictícia com um viés para estudos de análise exploratória e pré-processamento. Ela cumpre seu papel, auxiliando a aluna em seu desenvolvimento na sprint, porém não permitiu validar muitas hipóteses, visto que os dados, além de bastante fictícios, são muito balanceados.

No que diz respeito às hipóteses pode-se concluir o seguinte:

- Há maior venda de livros no último trimestre do ano provavelmente devido ao Natal: **REFUTADA. Há maior venda no terceiro trimestre, sendo que elas caem para o último, próximo ao Natal.**
- Há maior venda de livros no primeiro trimestre do ano devido à retomada das aulas: **REFUTADA. Há maior venda no terceiro trimestre. Pode-se tentar refletir sobre o fim das férias escolares de meio do ano e retomada das aulas, porém como o mesmo não ocorre para fevereiro e janeiro (primeiro trimestre), essa hipótese não faz muito sentido. É mais intuitivo pensar que há um super balanceamento de dados fictícios da base do que uma explicação racional para tais resultados.**
- Há maior venda de livros em janeiro de 2023 do que em 2024 devido à proximidade com a pandemia: **INVALIDADA. Impossível realizar a análise tendo somente dados para o ano de 2023.**
- Há maior venda de livros aos sábados: **PARCIALMENTE VALIDADA. Sábado é o segundo dia com mais venda durante o ano de 2023, sendo o primeiro uma segunda feira. Assim, pode-se concluir que é de fato um dia próspero para venda de livros e possíveis campanhas e divulgação.**
- Segundo o Veja, Nordeste é a região que mais lê, portanto o estado com maior registro de compra é de lá: **INVALIDADA. Não é possível realizar tal análise visto que os valores para DDD são fictícios. O resultado do DDD com maior número de vendas não se relaciona com nenhuma região do Brasil.**

Para fins de aprendizado, posso afirmar que o trabalho em questão me auxiliou muito no desenvolvimento e entendimento do processo de pré-processamento e análise exploratória, assim como no desenvolvimento de minhas skills de programação e análise de dados. Embora não tenha sido criado um dash atrelado à base de dados, o que envolveria modelagem e outros processos, consegui realizar diversas análises por meio das visualizações da biblioteca importada.

Minha maior dificuldade rondou o entendimento da diferença das etapas de pré-processamento e análise exploratória. Compreender o que deveria ser realizado primeiro. Dificuldade, esta, combatida com maestria por meio do auxílio acadêmico da faculdade, permitindo que eu concluísse o projeto final. Para trabalhos futuros no meu ramo profissional planejo certamente colocar em prática meus aprendizados nessa sprint.