**Etapa 2 do CRISP-DM - Entendimento dos Dados**

**1. Coleta Inicial dos Dados**

Quais fontes de dados estão disponíveis para o projeto?

**R:** DataSet em formato de um arquivo CSV.

De onde esses dados foram obtidos? Eles são internos, externos, públicos?

**R:** Extraímos essas informações do DataSet presente no Kagle. Esse DataSet fica disponível de forma pública, também, consideramos esses dados confiáveis.

Em que formato os dados foram entregues (CSV, SQL, API, Excel, etc.)?

**R:** Coletamos em formato .CSV

**2. Descrição dos Dados**

Qual é a dimensão do conjunto de dados (número de linhas e colunas)?

**R:** Em nossa base de dados há em torno de 99 mil linhas, já em colunas existem 8

Quais são os tipos de variáveis presentes (numéricas, categóricas, textuais,

temporais, etc.)?

**R:** Em nossa extração de dados temos as seguintes variáveis numéricas, categóricas e temporais.

Há variáveis que claramente se relacionam com o objetivo do projeto?

Quais?

Sim, numérica e temporal!

**3. Exploração Inicial**

Que estatísticas descritivas podem ser calculadas (médias, medianas,

valores máximos/mínimos)?

**R:** Inicialmente, estamos considerando somente a média.

Existem distribuições interessantes ou inesperadas nas variáveis?

**R:** Não identificamos no momento.

Foi identificada alguma correlação inicial entre variáveis?

**R:** Sim, a maior quantidade de produtos vendidos correspondem a produto abaixo de 1kg.

**4. Qualidade dos Dados**

Existem valores ausentes? Se sim, em que proporção e em quais variáveis?

**R:** Sim, não conseguimos definir quais são os produtos, pois não contém os nomes.

Há dados duplicados ou inconsistentes?

**R:** Não identificamos no momento.

Foram identificados outliers? Eles parecem ser erros ou observações

legítimas?

**R:** Na análise atual, não conseguimos identificar

Há variáveis com muitos valores únicos que podem dificultar a análise?

**R:** Não identificamos.

**5. Primeiras Hipóteses**

Com base nessa exploração inicial, que hipóteses podem ser levantadas

sobre os dados?

**R:** Com base na análise inicial, verificamos que os componentes que mais foram emitido vendas, correspondem a produtos com um peso inferior a 1kg e que pertencem a categoria de Som/Eletrônica.

Quais problemas de qualidade podem impactar diretamente os objetivos

definidos na Etapa 1 (Entendimento do Negócio)?

**R:** A não identificação dos produtos, dificulta a clareza dos dados que temos.

Quais dados adicionais poderiam ser úteis, mas não estão disponíveis?

**R:** Nome do produto e sua definição.

**6. Conclusões e Próximos Passos**

Quais etapas de pré-processamento ou limpeza provavelmente serão

necessárias antes de avançar para a etapa de Preparação dos Dados?

**R:** Nosso foco será segmentar em dois gráficos principais as informações inicialmente coletadas, para que assim, consigamos direcionar melhor a análise dos dados. Também, vamos identificar possíveis valores(outliers e etc) que impactam negativamente em nossa coleta/análise.