**3. Data Preparativo**

Após uma análise cuidadosa dos dados coletados, preparamos esta seção detalhada para tratar e organizar os dados de forma que fiquem prontos para a modelagem e as análises futuras. Este processo envolve a seleção, limpeza, construção, integração e formatação dos dados com análises avançadas e insights preliminares, explorando tanto os padrões lineares quanto os não lineares.

**3.1 Selecionar Dados**

**Critérios de Seleção e Exclusão:** Foram priorizados os dados mais relevantes para os objetivos do projeto, considerando atributos críticos para os processos de otimização, identificação de defeitos e análise de desempenho operacional.

**Dados Selecionados:**

* **Produção:**
  + Máquina.
  + Data.
  + Quantidade Produzida.
  + Tempo de Inatividade.
  + Custo de Produção.
* **Devoluções:**
  + Produto.
  + Motivo da Devolução.
  + Quantidade Devolvida.
  + Data.
* **Qualidade:**
  + Produto.
  + Tipo de Defeito.
  + Cor.
  + Taxa de Defeito.

**Dados Excluídos:**

* **Logs de manutenção**: Baixa correlação com as devoluções e defeitos.
* **Descrição textual de máquinas**: Informação redundante.
* **Dados geográficos detalhados**: Não são críticos para os objetivos do projeto.

**Análise Estatística Inicial:**

* **Produção:** A variância nas quantidades produzidas revelou clusters claros de máquinas com alta e baixa eficiência. Esses clusters servirão para categorizar os grupos posteriormente.
* **Devoluções:** Uma análise descritiva inicial indicou que 15% das devoluções correspondem a problemas de cor, representando o maior motivo de devolução.
* **Qualidade:** Observou-se que 40% dos defeitos estão concentrados em dois produtos específicos, sugerindo uma análise mais aprofundada.

**3.2 Limpando os Dados**

**Tratamento de Valores Faltantes:**

* Para **Custo de Produção**, foi utilizada a imputação baseada em modelos regressivos, onde o custo foi previsto com base no tipo de máquina, quantidade produzida e tempo de inatividade.
* **Tempo de Inatividade:** Valores ausentes foram imputados utilizando KNN (K-Nearest Neighbors), considerando máquinas de configurações similares.
* **Taxa de Defeito:** Preenchida com a média ponderada por tipo de produto, priorizando os registros mais recentes para refletir mudanças operacionais.

**Detecção e Correção de Outliers:**

* **Custo de Produção:** Boxplot revelou outliers extremos em 3% dos registros. Para validar, aplicamos o método IQR combinado com inspeção manual dos registros.
* **Taxa de Devolução:** Valores acima de 3 desvios-padrão foram investigados e correlacionados com falhas específicas no processo produtivo.

**Inconsistências Temporais:**

* Identificou-se registros com **data de devolução anterior à data de produção**, que foram corrigidos ou excluídos.
* Consolidou-se o formato de data para "YYYY-MM-DD" para todas as tabelas.

**Explorações Visuais:**

* Histogramas confirmaram uma distribuição assimétrica do Custo de Produção, sugerindo a necessidade de normalização.
* Gráficos de dispersão indicaram uma relação inversa entre tempo de inatividade e quantidade produzida.

**Insights:**

* Máquinas com maior tempo de inatividade estão fortemente associadas a produtos com maiores taxas de defeito, destacando a necessidade de priorizar manutenção preventiva.

**3.3 Construindo os Dados**

**Novos Atributos Derivados:**

1. **Eficiência de Produção:**

**Eficiência = Quantidade\_Produzida / Capacidade\_Maxima\_da\_Maquina**

* + Insights: A eficiência média por máquina revelou um desvio padrão elevado, indicando inconsistências operacionais.

1. **Impacto Financeiro das Devoluções:**

**Impacto== (Custo\_de\_Producao) \* (Quantidade\_Devolvida)**

* + Insights: Produtos de alto custo, como os da linha premium, têm impacto financeiro desproporcional nas devoluções.

1. **Sazonalidade na Taxa de Defeito:** Criou-se uma variável categórica que identifica "Alta" ou "Baixa" sazonalidade com base em padrões mensais.

**Registros Gerados:**

* **Clusters Temporais:** Agrupados por trimestres para capturar tendências de eficiência e devoluções.
* **Pontos Críticos:** Registro de eventos extremos, como semanas com mais de 10% de devoluções.

**Insights Avançados:**

* Produtos na cor "Vermelho" apresentaram os menores índices de eficiência, enquanto "Preto" se destacou pela menor taxa de devoluções.

**3.4 Integrando os Dados**

**Operações de Integração Avançadas:**

* Integração de **tabelas multi-fonte**: Criou-se uma tabela unificada que combina dados de produção, devoluções e qualidade com mais de 20 atributos derivados.
* **Agregação Avançada:** Métricas como desvio padrão de defeitos por tipo de máquina e correlação cruzada entre cor e tipo de defeito foram adicionadas.

**Técnicas Aplicadas:**

* **Mesclagem:** Unificação por **Produto** e **Data**, criando registros consolidados que permitem análises multidimensionais.
* **Agregação Temporal:** Criou-se um resumo semanal e trimestral para cada máquina.

**Insights Avançados:**

* Máquinas com maior eficiência também apresentaram as menores taxas de devolução, reforçando a hipótese de correlação entre manutenção e qualidade.

**3.5 Formatando os Dados**

**Transformações Realizadas:**

* Variáveis categóricas foram convertidas em **one-hot encoding** para análises em algoritmos baseados em regressão.
* Dados numéricos normalizados usando **Min-Max Scaling** para padronizar as variáveis.
* Reorganização do dataset consolidado com as seguintes colunas:
  + Primeiro campo: Identificador único de máquina.
  + Últimos campos: Variáveis alvo (Taxa de Defeitos e Eficiência de Produção).

**Preparação para Ferramentas:**

* Exportação do dataset para **CSV**, **Parquet** e **SQL** para garantir compatibilidade com diferentes ferramentas analíticas.
* Validação final para verificar se os formatos atendem aos requisitos de ferramentas como Python, Power BI e Excel.

**Insights:** A normalização revelou padrões sutis em variações de custos que antes estavam mascarados por escalas heterogêneas. Esses padrões podem indicar melhorias específicas no processo.