Detecção de Fraudes em Transações Financeiras com Big Data

Diego Hiroshi Goto Yanaze - 626457

Guilherme Lorenzetti Bonini - 617547

Diogo Tachibana de Oliveira - 626872

Luciano Moreira Barbosa Junior - 625418

Luís Fernando - 629881

O PROBLEMA DE NEGÓCIO

- Volume massivo de dados.
- Necessidade de identificar comportamentos atípicos.
- Ausência de dados previamente rotulados.

OBJETIVOS DO PROJETO

- Tratamento e limpeza de dados brutos.
- Criação de novas variáveis (Feature Engineering).
- Aplicação de algoritmos de Machine Learning (Isolation Forest e K-Means).
- Criação de um Score de Risco.

O DATASET E LIMPEZA (ETL)

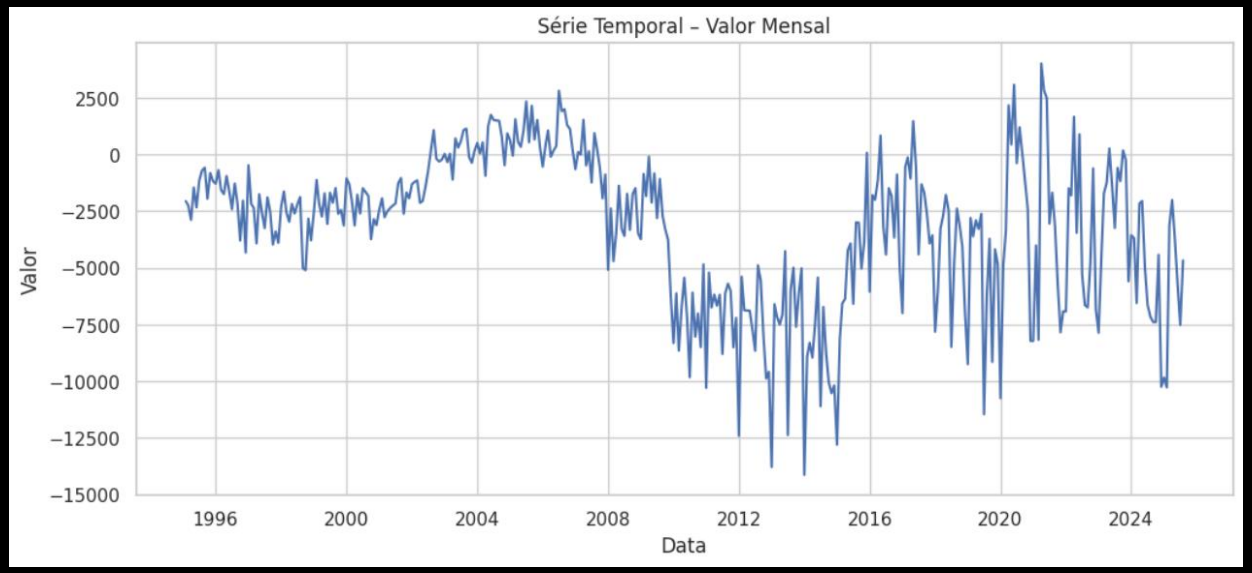
```
# ============================================================
# 2. Carregar SEU dataset
# ============================================================
df = pd.read_csv("/content/Transações correntes - mensal - saldo.22701 1.csv", sep=';')

# ============================================================
# 3. Preparação e criação de atributos
# ============================================================

# Limpar e converter a coluna 'valor'
# Remover as aspas duplas, espaços e substituir vírgulas por pontos antes de converter para numérico
df["valor"] = df["valor"].astype(str).str.replace('"', '').str.replace(' ', '').str.replace(',', '.', regex=False).astype(float)

# Converter datas
df["data"] = pd.to_datetime(df["data"], format="%d/%m/%Y")
```

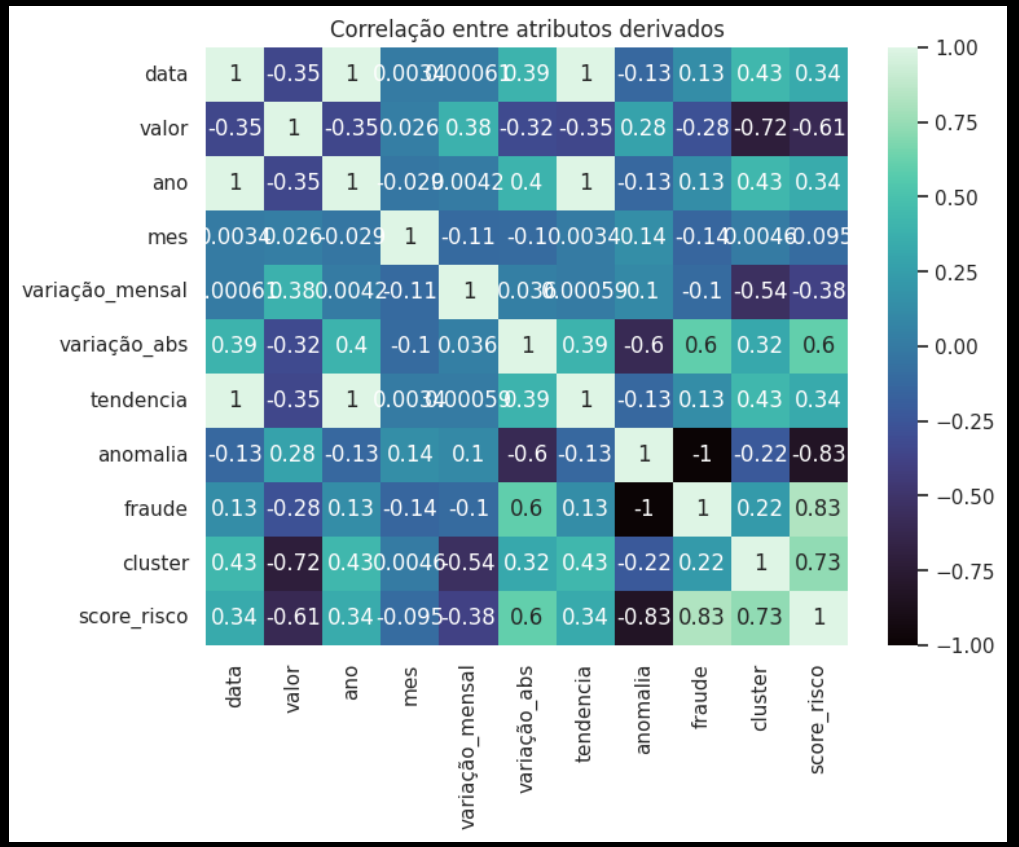
ENGENHARIA DE ATRIBUTOS (FEATURE ENGINEERING)

novas colunas criadas:
ano, **mes**,
variação_mensal,
variação_abs,
tendencia

```python
# Criar atributos derivados
df["ano"] = df["data"].dt.year
df["mes"] = df["data"].dt.month
df["variação_mensal"] = df["valor"].diff()
df["variação_abs"] = df["valor"].diff().abs()
df["tendencia"] = np.arange(len(df))
```

ANÁLISE EXPLORATÓRIA INICIAL



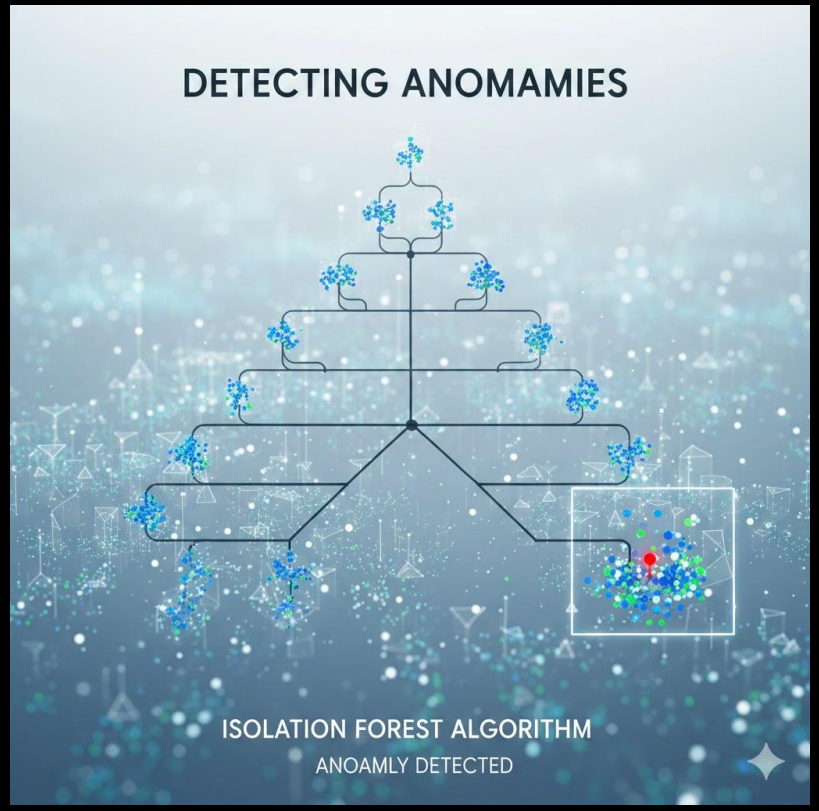


ESTRATÉGIA DE MACHINE LEARNING

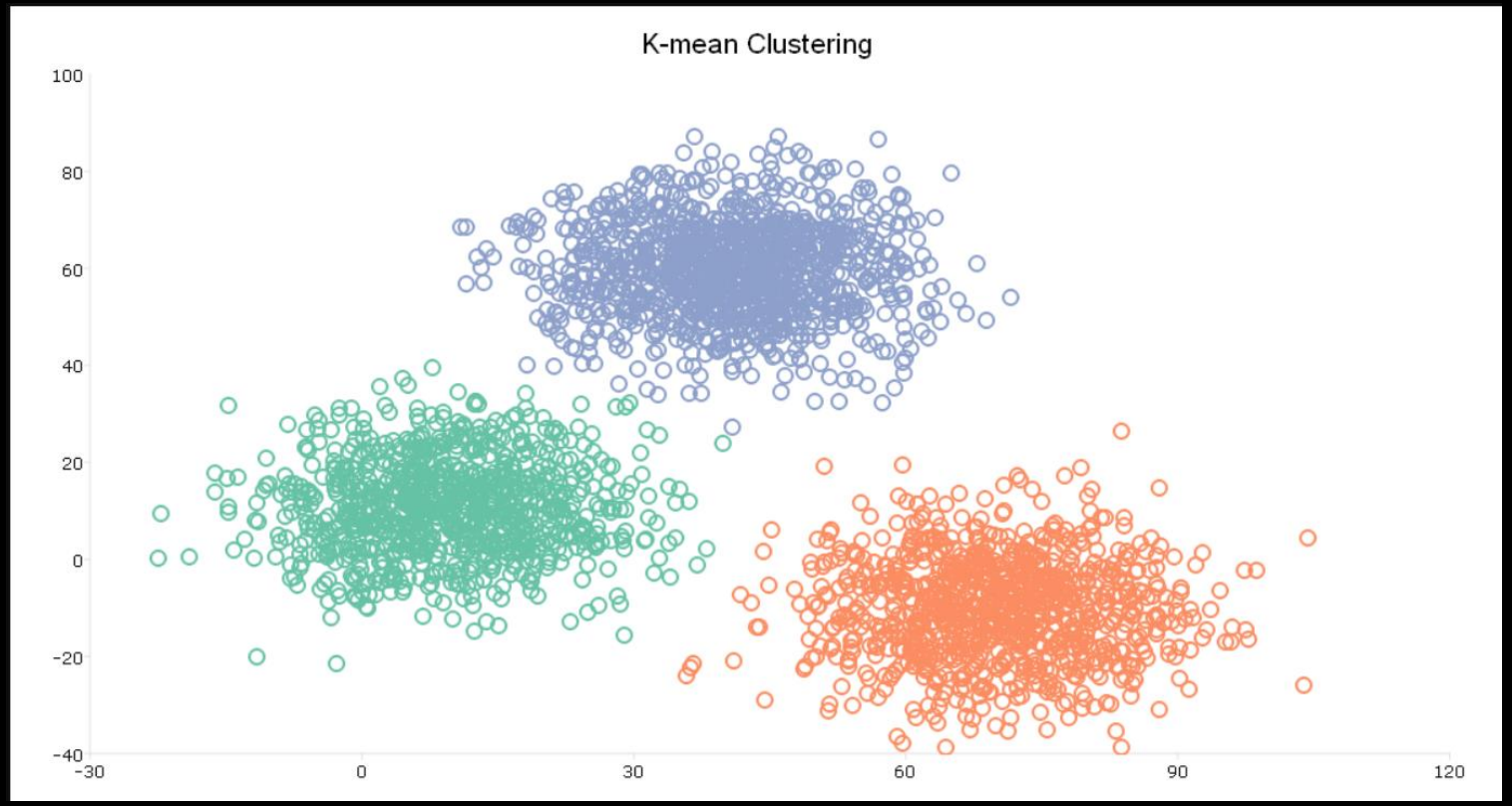
Dados -> Padronização (StandardScaler)

- aplicamos o **`StandardScaler`** para colocar todos os dados na mesma escala numérica, evitando que valores absolutos muito grandes distorcessem os cálculos de distância dos algoritmos.

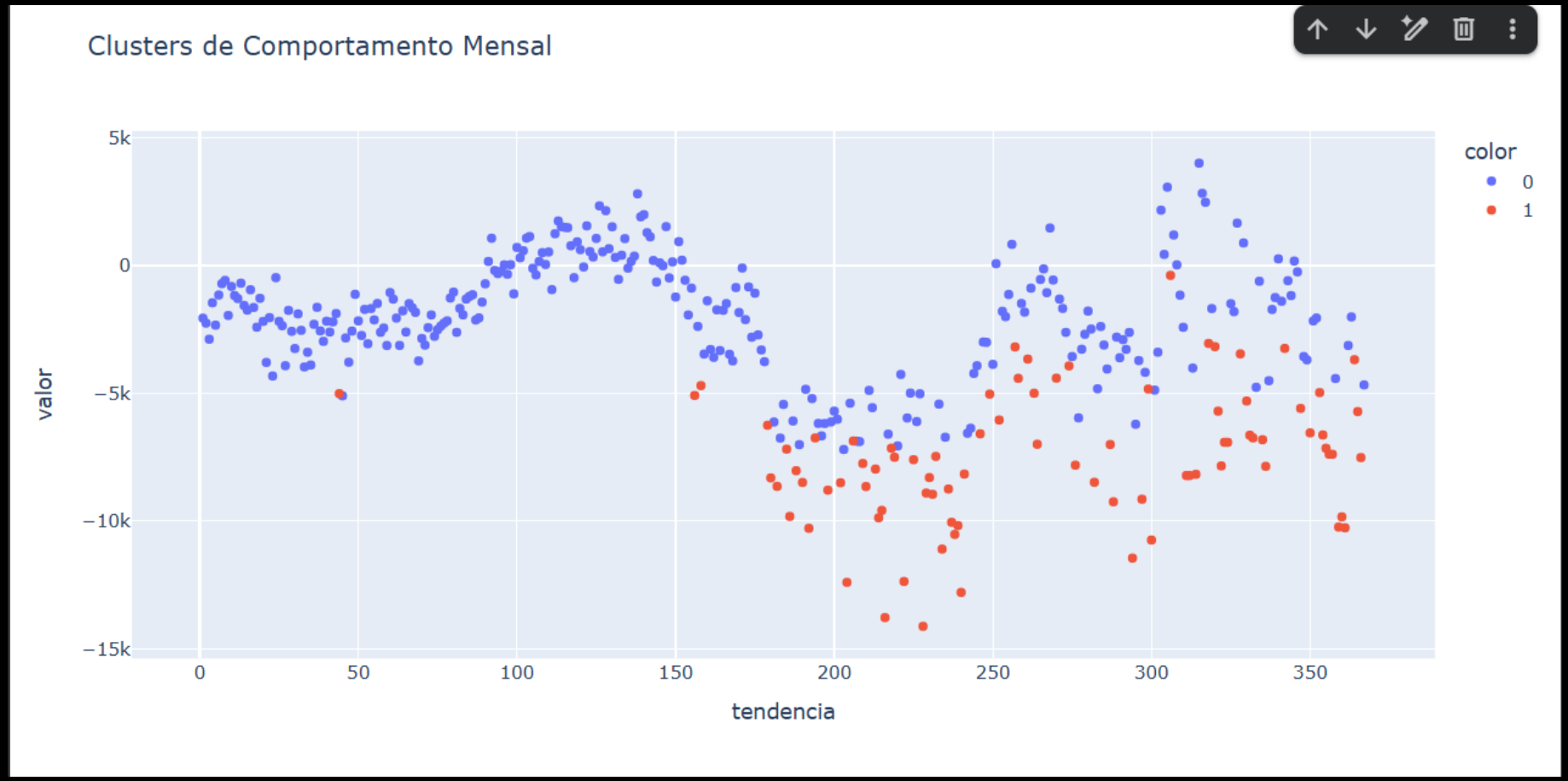
O ALGORITMO ISOLATION FOREST



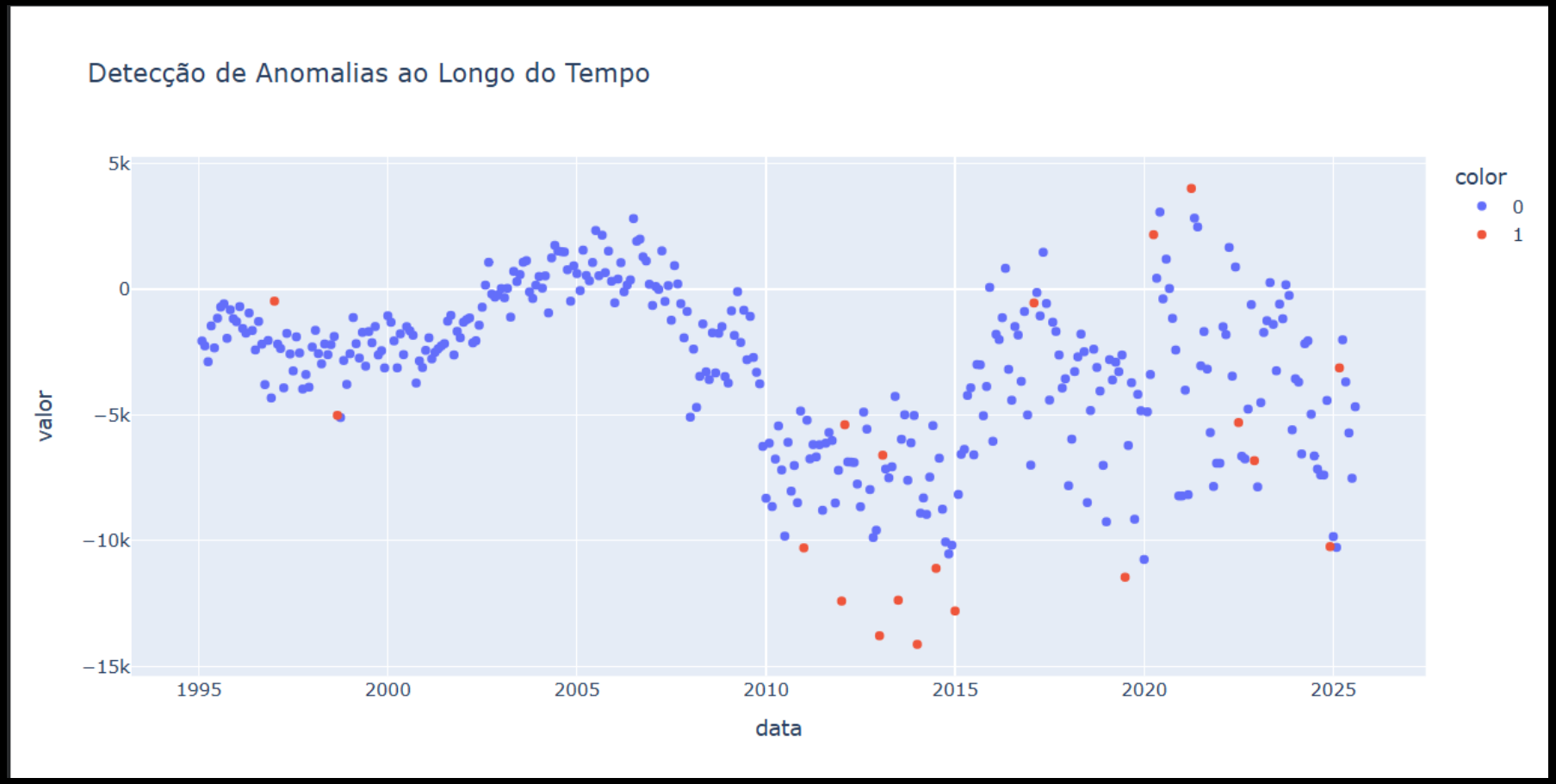
O ALGORITMO K-MEANS



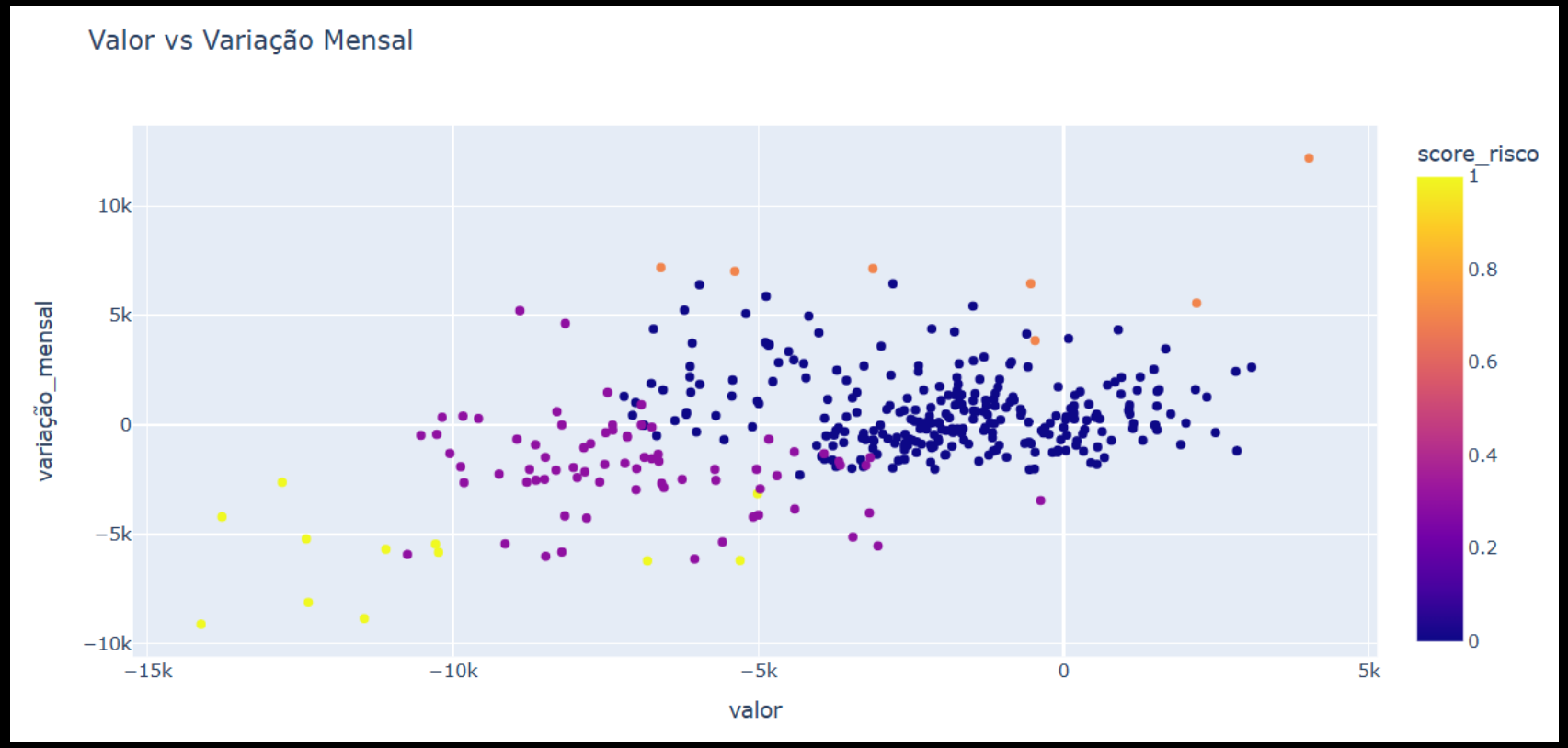
VISUALIZANDO OS CLUSTERS



DETECÇÃO DE ANOMALIAS NO TEMPO

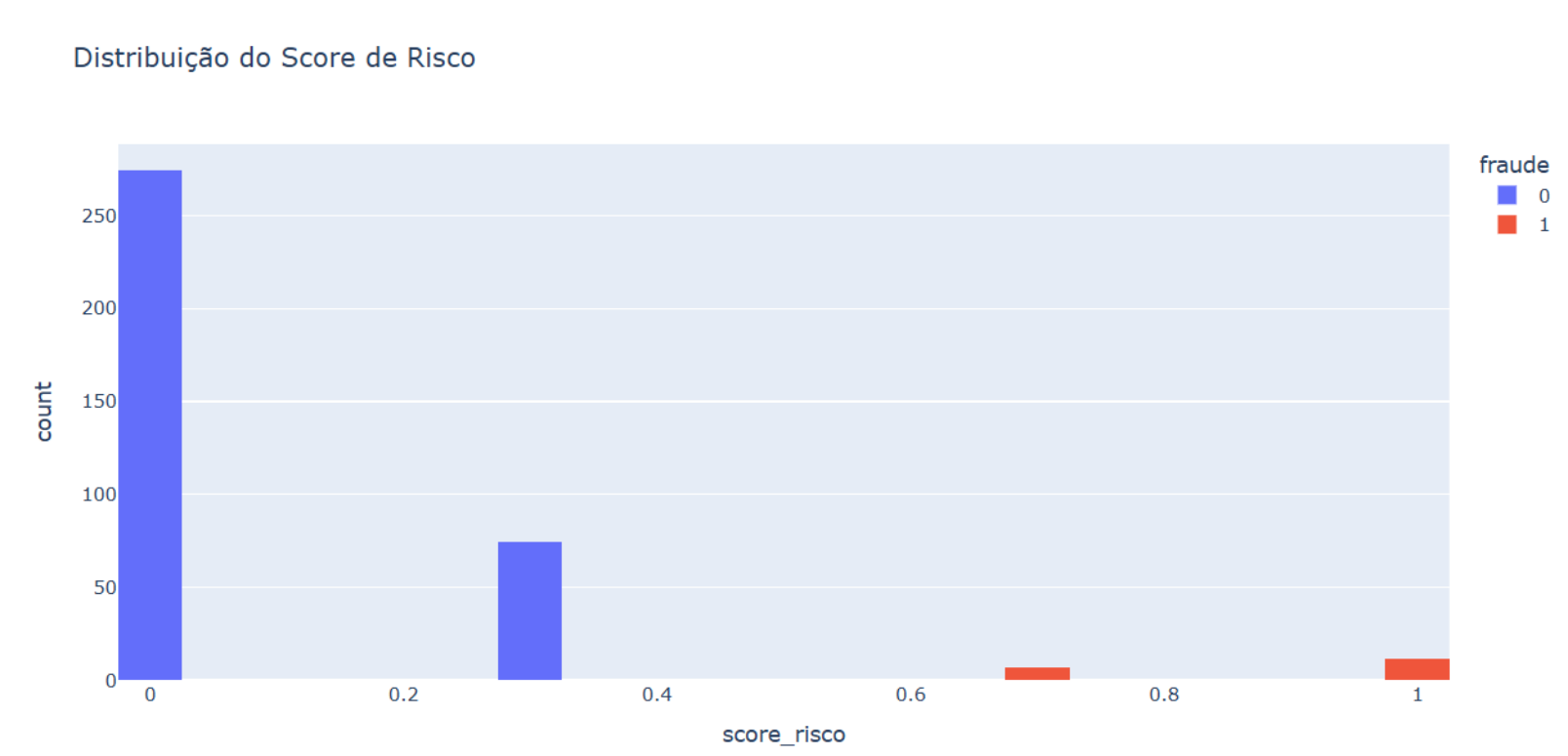


RELAÇÃO VALOR X VARIAÇÃO



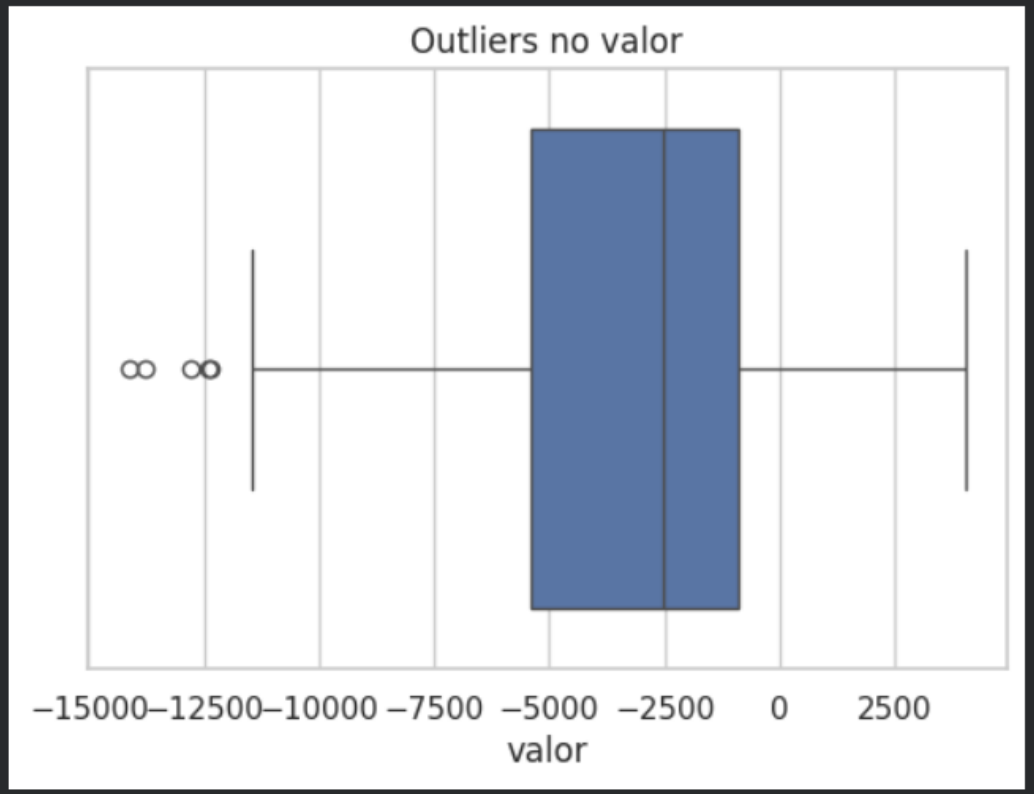
O SCORE DE RISCO COMPOSTO

```
# Score de risco baseado nas duas análises
df["score_risco"] = (df["fraude"] * 0.7) + (df["cluster"] * 0.3)
```



ANÁLISE DE OUTLIERS (BOXPLOT)

```python
# Boxplot de Outliers
plt.figure(figsize=(6,4))
sns.boxplot(x=df["valor"])
plt.title("Outliers no valor")
plt.show()
```



CONCLUSÃO E VALOR AGREGADO

OBRIGADO PELA ATENÇÃO!