

# *Análise de fraude em cartão de crédito*

Gutemberg Souza, Juliana Farkuh, Maria Mansour, Thomas Ramos e Vittor Rodrigues



# *Equipe de trabalho*



Gutemberg Souza



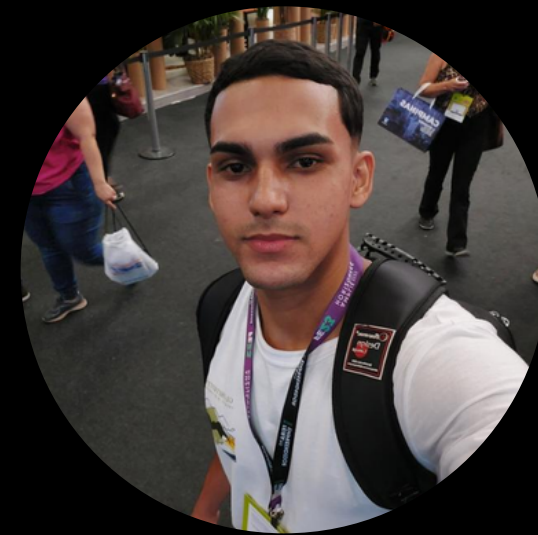
Juliana Farkuh



Maria Mansour



Thomas Ramos



Vittor Rodrigues

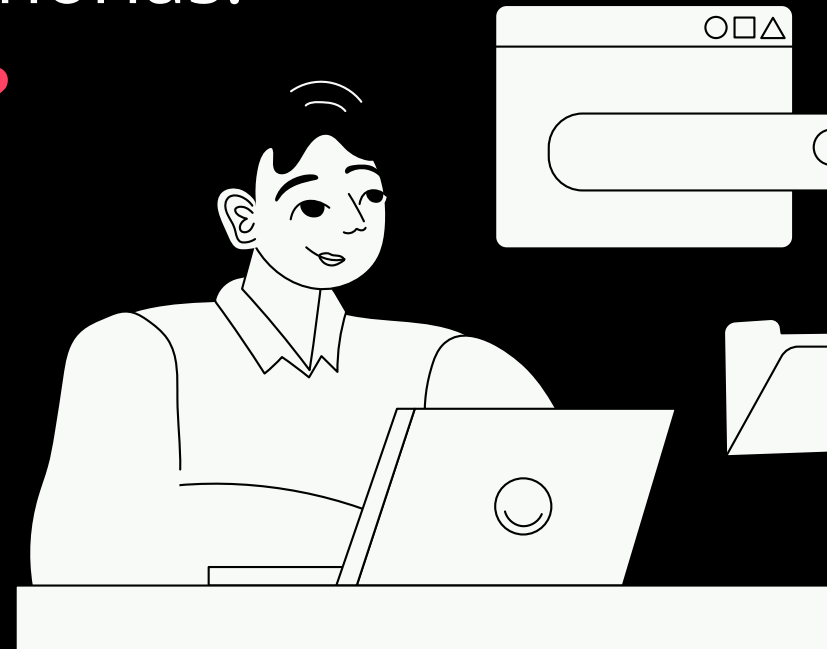
# *Apresentação do Problema:*

Fraudes em setores como **bancos, e-commerces e seguros geram grandes prejuízos** e afetam a confiança dos clientes. O sistema de detecção busca identificar atividades fraudulentas com precisão, minimizando falsos positivos e negativos. O **objetivo é criar um modelo preditivo** que identifique transações suspeitas a partir de características fornecidas, treinando-o com um conjunto de dados e avaliando seu desempenho com outro.



# *Perguntas da pesquisa:*

1. **Quais características das transações estão mais correlacionadas com fraudes?** A pesquisa buscará variáveis como valor, localização e tipo de transação que indicam fraudes.
2. **Qual a precisão do modelo preditivo?** Avaliar a eficácia do modelo com métricas como precisão, sensibilidade, especificidade e falsos positivos.
3. **Como o modelo se comporta em diferentes cenários de teste e pode ser otimizado?** Analisar seu desempenho e buscar melhorias.
4. **Como a distribuição de fraudes varia entre treino e teste?** Verificar discrepâncias entre os dois conjuntos de dados.



# *Objetivo da pesquisa:*

**Desenvolver e avaliar modelos preditivos para identificar transações fraudulentas de forma eficiente. Especificamente, o projeto visa:**

1. Identificar as características mais relevantes para diferenciar transações legítimas e fraudulentas.
2. Comparar diferentes técnicas de aprendizado de máquina para detecção de fraudes.
3. Criar estratégias para lidar com o desbalanceamento de classes no dataset.
4. Minimizar falsos positivos e negativos, otimizando a eficiência do modelo. O objetivo é criar um sistema de detecção aplicável em cenários reais, prevenindo perdas financeiras e protegendo transações.

# *Dataset e Data Acquisition*

O dataset foi retirado do site kaggle e pode ser encontrado nesse link:

<https://www.kaggle.com/datasets/dermisfit/fraud-transactions-dataset>

Os dados utilizados referem-se a transações de cartões de crédito, contendo informações sobre o titular do cartão, o comerciante, a transação em si, e uma classificação indicando se a transação foi fraudulenta.

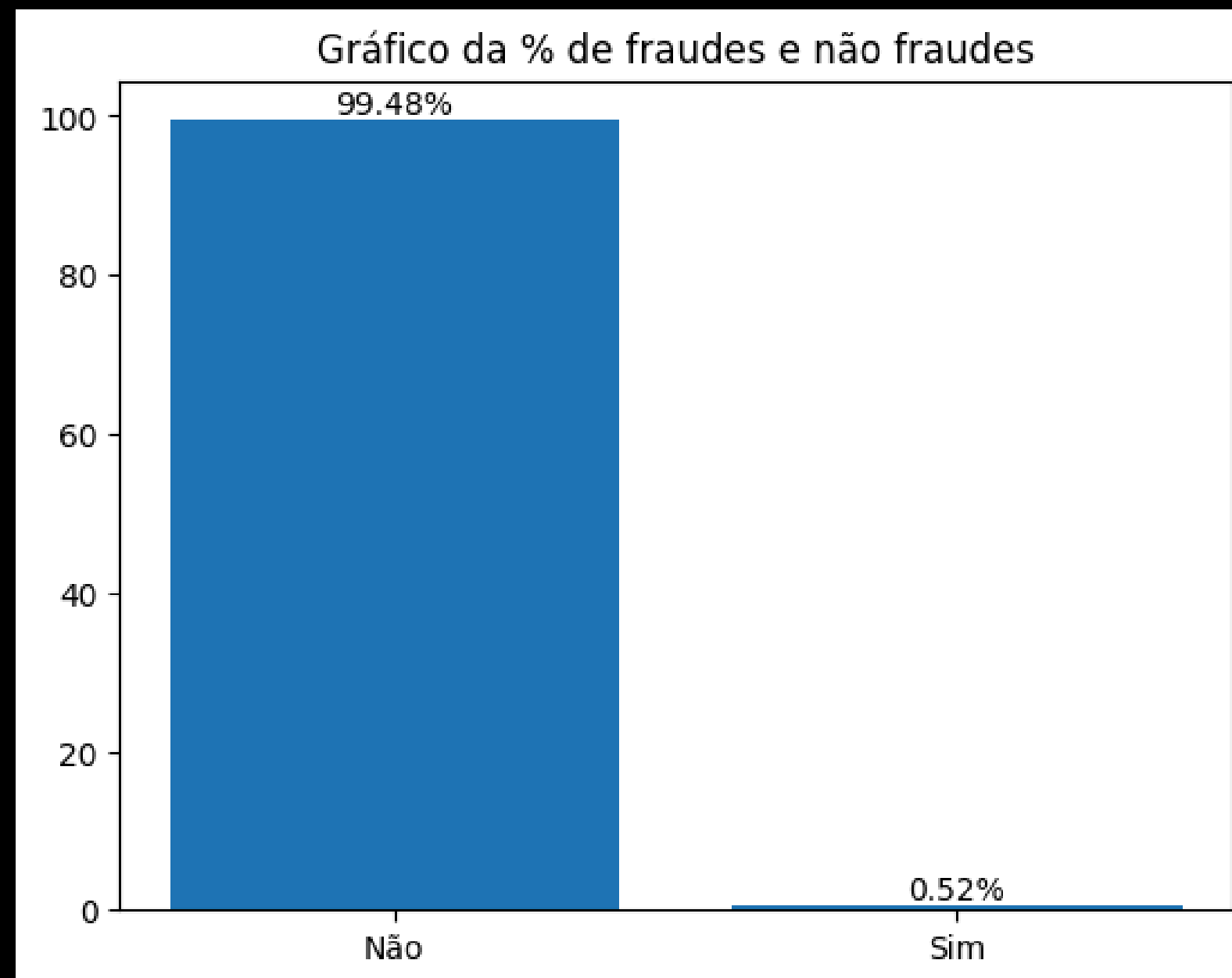


# *Uso do modelo*

O conjunto de dados tem como objetivo identificar fraudes, utilizando a variável “is\_fraud” como target. Após a limpeza dos dados, a biblioteca PyCaret foi aplicada para comparar modelos de aprendizado de máquina, destacando Random Forest e XGBoost. O Random Forest se destacou por sua capacidade de lidar com dados imperfeitos e identificar características importantes. O modelo foi ajustado através de hiperparâmetros e implementado, provando ser o mais adequado para a base, devido à sua robustez e habilidade de corrigir amostras com dados ausentes.

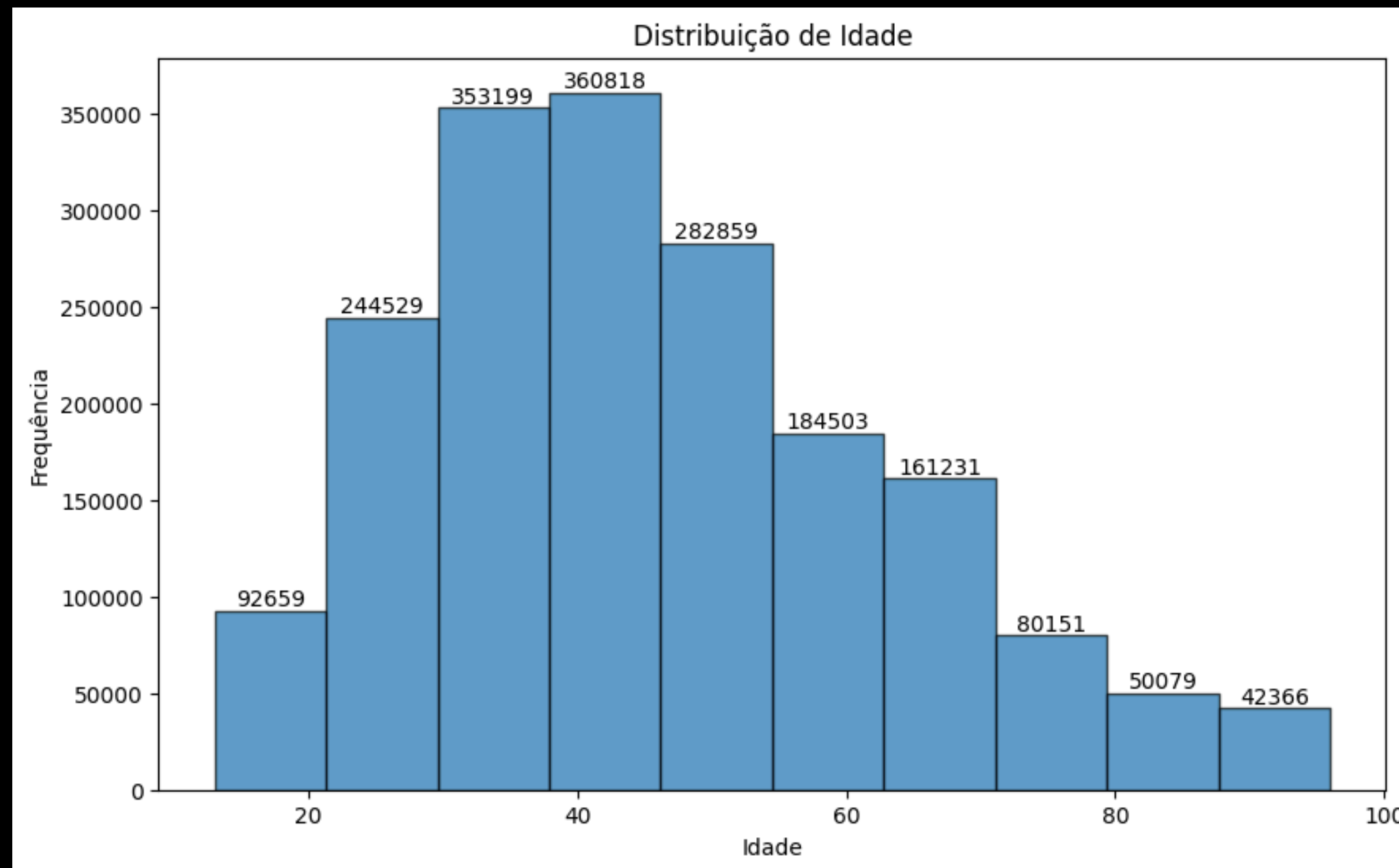
# EDA

Qual porcentagem são fraudes?

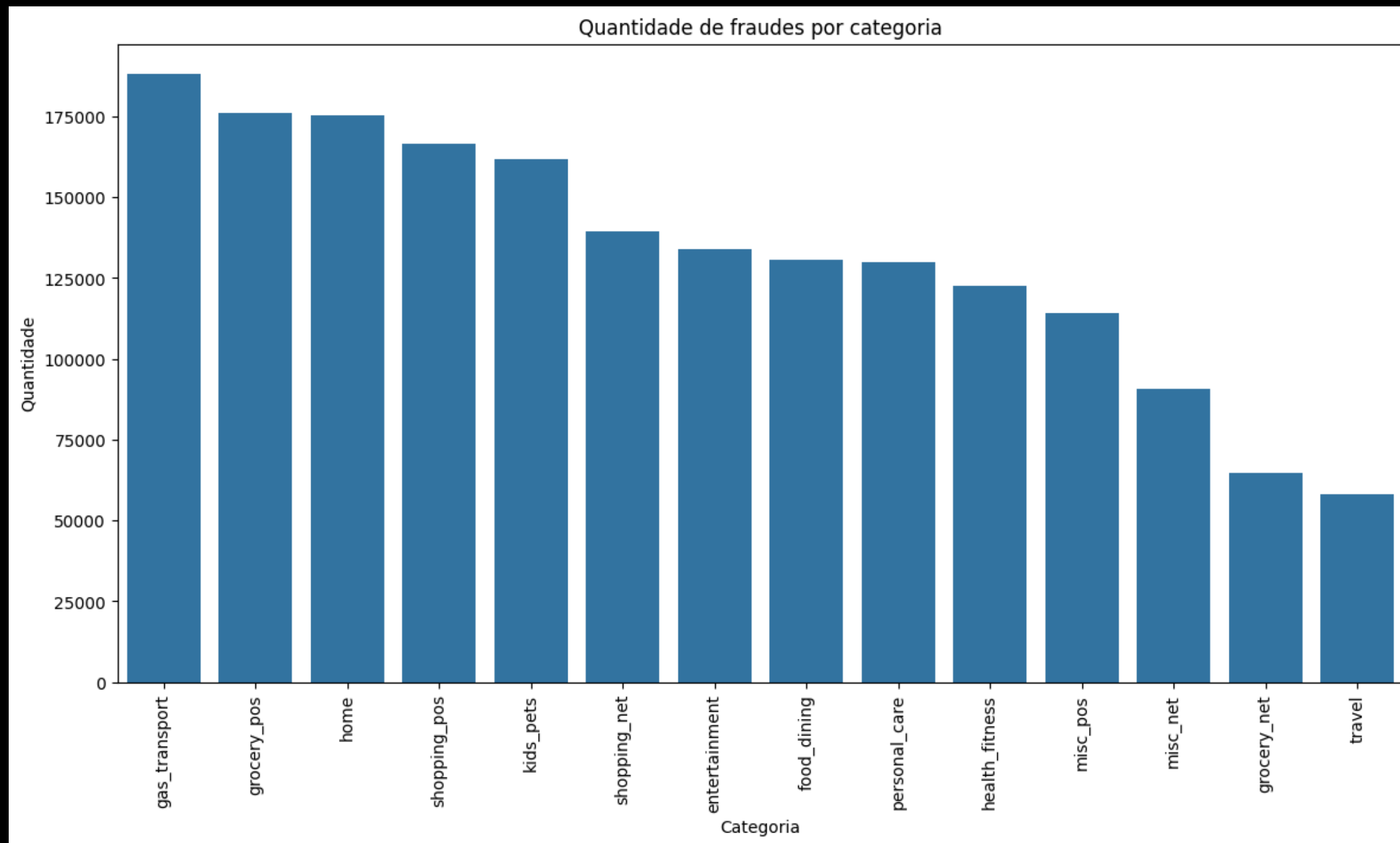




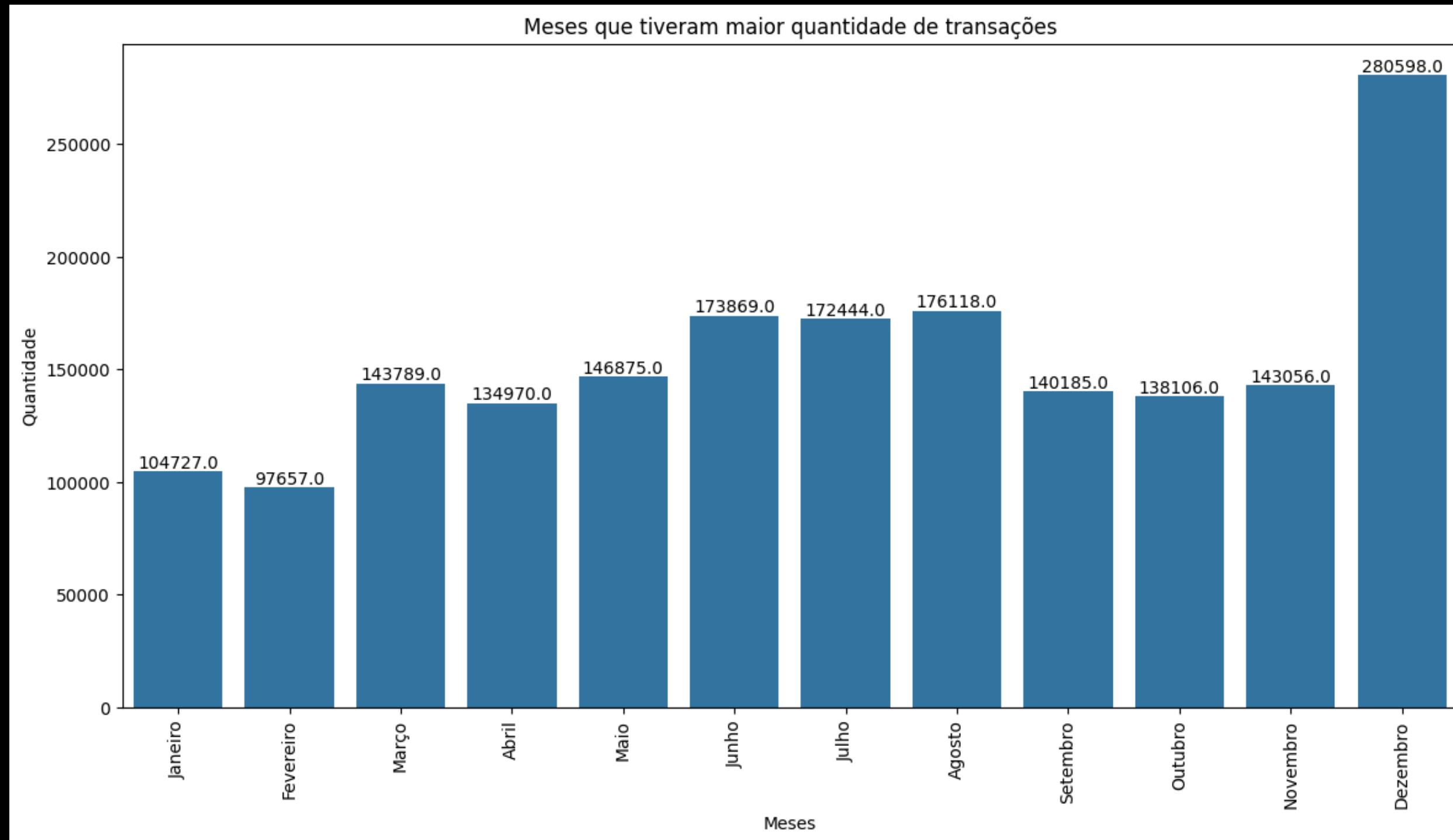
**Qual é a faixa etária com a maior ocorrência de fraude? E com menor ocorrência?**



# Quantidade de fraude por categorias

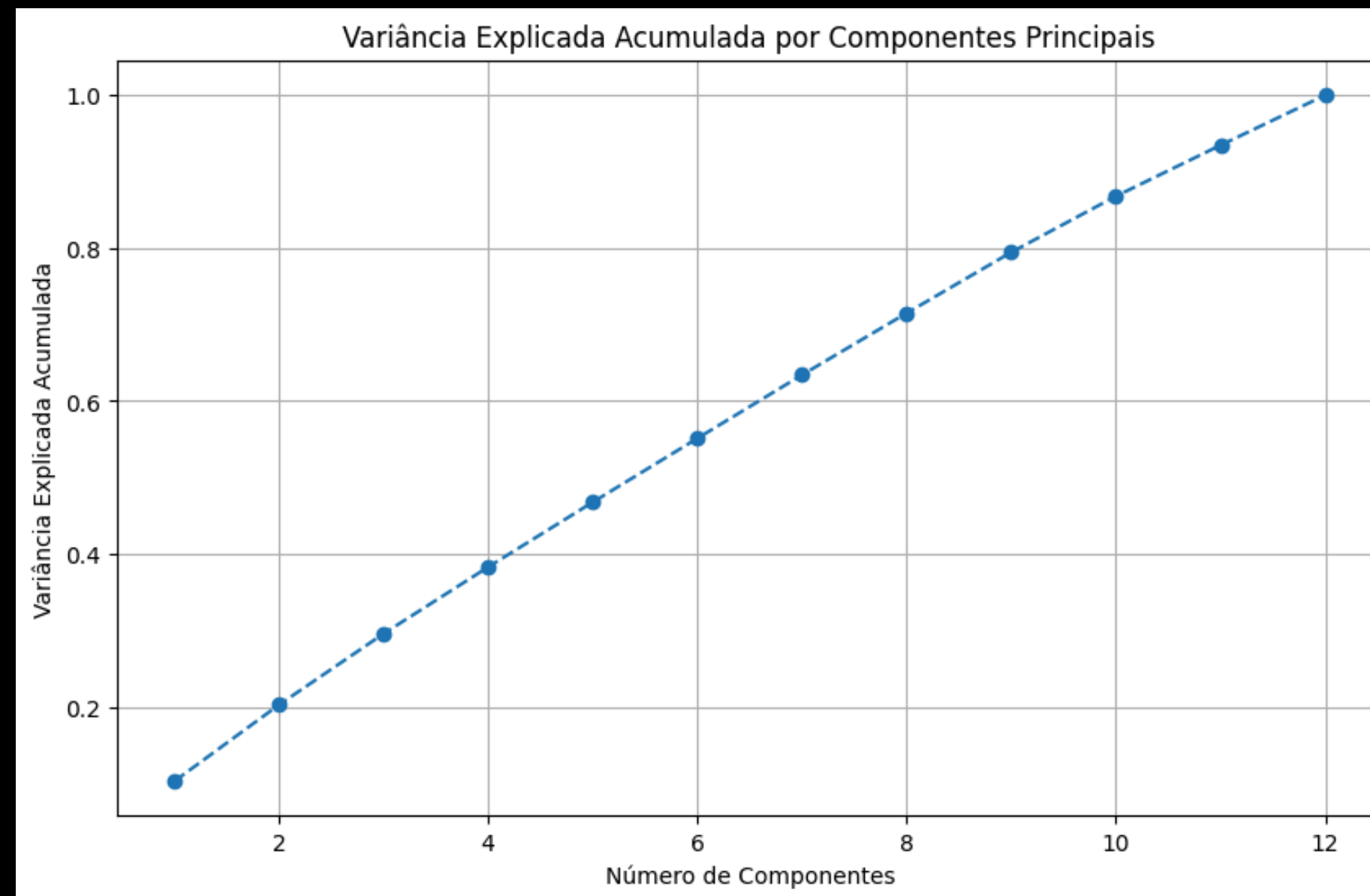


# Qual é o mês que registrou mais transações (não fraudulentas e fraudulentas)?



# PCA

**Número de componentes para explicar 95% da variância: 12**  
**Dimensões dos dados após PCA: (1852394, 12)**



# PCA

