Relatório de Análise Exploratória: Caracterização de Transações Fraudulentas e Avaliação da Qualidade dos Dados

1. Delimitação da Questão Analítica

1.1. Objeto de Estudo

O objeto central deste estudo é a **Análise Exploratória de Dados (AED)**, que constitui um passo investigativo preliminar e fundamental no campo da ciência de dados. Conforme preconizado por Peter Bruce e Andrew Bruce em "Estatística Prática para Cientistas de Dados" (2019), a AED é o processo de utilizar métodos estatísticos e de visualização para resumir as principais características de um conjunto de dados, descobrir padrões, identificar anomalias e testar hipóteses iniciais. E analisar a possibilidade do uso da amostra usada no estudo para um modelo preditivo.

1.2. População-Alvo

A população-alvo, a partir da natureza das variáveis analisadas (Customer Age, Income, Credit Risk Score), corresponde ao universo de todos os clientes ou transações de uma determinada entidade, que pode ser uma instituição financeira, uma plataforma de e-commerce ou qualquer serviço digital que processa transações suscetíveis a fraude.

1.3. Objetivo da Análise

O objetivo desta análise exploratória é estabelecer uma base empírica sólida que possa informar decisões de negócio e guiar as próximas etapas de um projeto de ciência de dados, como a construção de modelos de machine learning para detecção de fraudes. Os objetivos específicos são : Identificar Fatores Diferenciadores, Quantificar Associações, Avaliar a Qualidade dos Dados, Formular Hipóteses e Recomendações.

2. Identificação e Caracterização da Amostra

2.1. Coleta de Dados

Os dados foram coletados por meio do Site Kaggle, o conjunto de dados em questão foi disponibilizado em 2022 pela Conference on Neural Information Processing Systems.

2.2. Tipo de Amostragem

Essa amostra é do tipo sintética foi criada com o objetivo de se ser similar a dados da realidade, mas com a privacidade necessária dos dados. Foi criado com o objetivo de realizar

competições e pesquisas.

2.3. Tamanho e Justificativa da Amostra

A amostra total analisada compreende N=1.000.000 observações, conforme detalhado no gráfico "Distribuição de Casos de Fraude" . A composição da amostra é a seguinte:

- **Casos de Não Fraude:** 988.971 registros, correspondendo a 98,90% da amostra.
- **Casos de Fraude:** 11.029 registros, correspondendo a apenas 1,10% da amostra.

Este detalhamento revela um desafio analítico significativo. Embora o tamanho total da amostra (N=1.000.000) seja substancial, o estudo ainda enfrenta um paradoxo: é um problema de "small data" (dados pequenos) no que diz respeito à classe de interesse (Fraude), que está inserido dentro de um conjunto de dados maior. A robustez e a confiabilidade de quaisquer estimativas estatísticas (como média, mediana ou variância) para o grupo 'Fraude' são inerentemente mais fracas devido ao seu tamanho relativamente pequeno (n=10.500) em comparação com o grupo 'Não Fraude'. As estimativas para este grupo terão um erro padrão maior e intervalos de confiança mais amplos. Consequentemente, mesmo que uma diferença observada entre as medianas dos dois grupos pareça substancial, sua significância estatística pode ser menor do que aparenta, exigindo uma interpretação cautelosa dos resultados.

3. Classificação das Variáveis Analisadas

3.1. Variável Qualitativa (Dependente)

Nominais (13): fraud_bool, payment_type, employment_status, email_is_free, housing_status, phone_home_valid, phone_mobile_valid, has_other_cards, foreign_request, source, device_os, keep_alive_session, month.

Ordinais (1): credit_risk_score.

3.2. Variáveis Quantitativas (Independentes)

Discretas (9): prev_address_months_count, current_address_months_count, customer_age, zip_count_4w, bank_branch_count_8w, date_of_birth_distinct_emails_4w, bank_months_count, device_distinct_emails_8w, device_fraud_count.

Contínuas (9): income, name_email_similarity, days_since_request, intended_balcon_amount, velocity_6h, velocity_24h, velocity_4w, proposed_credit_limit, session_length_in_minutes.

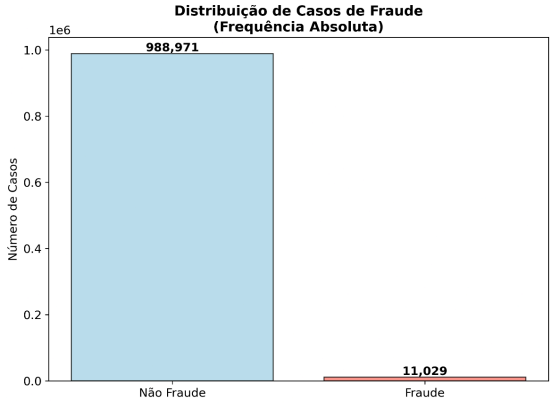
Principais variáveis:

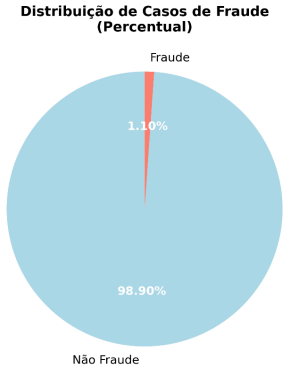
- **Fraud Bool**: Esta é uma variável **nominal** . Ela representa o resultado de interesse, classificando cada observação em uma de duas categorias mutuamente exclusivas: 'Fraude' ou 'Não Fraude'.

- **Credit Risk Score**: Uma variável **qualitativa ordinal**, pois representa o grau de risco de crédito para determinado cliente.
- **Customer Age**: Uma variável **quantitativa**, que é tratada como **discreta**. Representa a idade do cliente associado à transação.
- **Income**: Uma variável **quantitativa contínua**. Com base nos eixos do box plot na imagem 2, esta variável foi normalizada para um intervalo entre 0 e 1. Representa a renda do cliente.
- **Velocity 24H**: Uma variável **quantitativa contínua,** mede a frequência ou volume de atividade transacional de um cliente nas últimas 24 horas antes da solicitação/transação atual.
- **Session Length In Minutes**: Uma variável **quantitativa contínua**, que mede a duração da sessão do usuário no sistema, em minutos.

4. Organização e Visualização Estruturada dos Dados

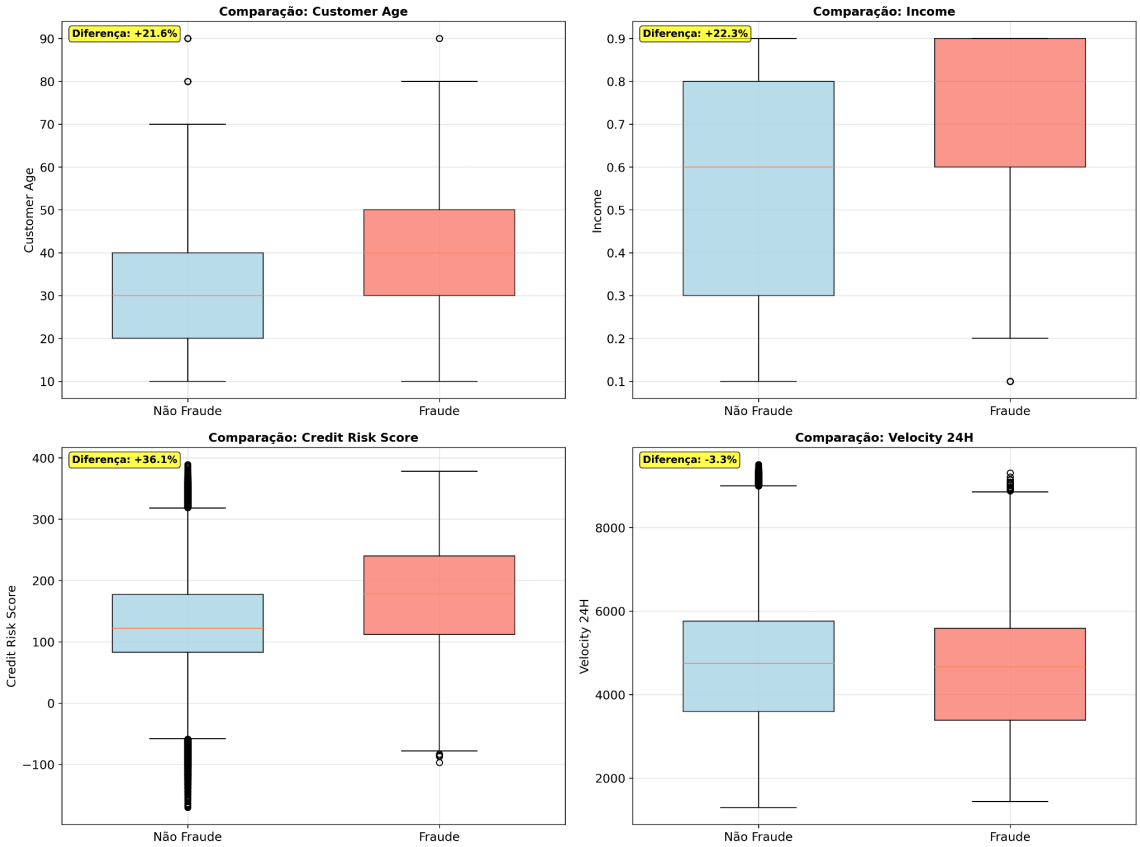
4.1. Distribuição da Variável-Alvo





Os gráficos de barras e de pizza na Imagem ilustram a frequência absoluta e percentual das classes 'Fraude' e 'Não Fraude'. A visualização revela um **severo desbalanceamento de classes**, com uma proporção de aproximadamente 99 para 1 entre casos não fraudulentos e fraudulentos. Este é um dos achados mais críticos da análise, pois tem implicações diretas e profundas para qualquer esforço de modelagem preditiva. Modelos de machine learning treinados em dados tão desbalanceados tendem a desenvolver um viés para a classe majoritária, resultando em um desempenho pobre na detecção da classe minoritária (fraude). Métricas de avaliação padrão, como a acurácia, tornam-se enganosas, pois um modelo que simplesmente classifica todas as transações como 'Não Fraude' alcançaria uma acurácia de 98,95%, apesar de ser completamente inútil para o propósito de detecção de fraude.

4.2. Análise Comparativa via Diagramas de Caixa (Box Plots)



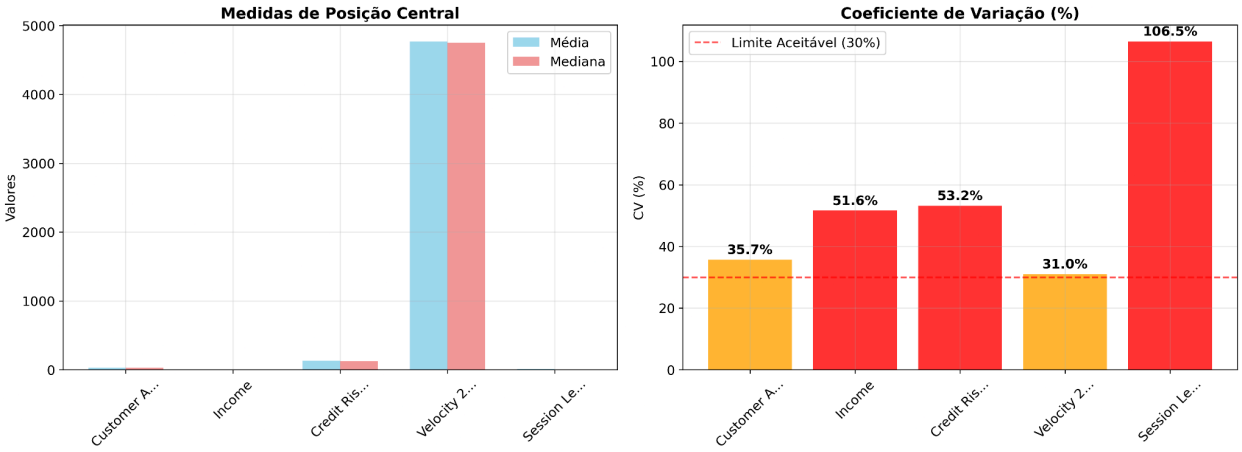
O **diagrama de caixa**, ou **box plot**, é uma ferramenta gráfica poderosa que resume a distribuição de uma variável numérica através do "resumo de cinco números": o valor mínimo, o primeiro quartil (Q1, 25º percentil), a mediana (Q2, 50º percentil), o terceiro quartil (Q3, 75º percentil) e o valor máximo. A "caixa" representa a **amplitude interquartil (AIQ)**, que contém os 50% centrais dos dados (AIQ=Q3−Q1). Os "bigodes" (whiskers) estendem-se para mostrar o alcance dos dados, e os pontos além dos bigodes são tipicamente identificados como **outliers** (valores discrepantes). A Imagem utiliza box plots para comparar as distribuições das variáveis quantitativas entre os grupos 'Fraude' e 'Não Fraude'.

- **Customer Age**: Observa-se uma diferença notável. A mediana da idade para casos de 'Fraude' é 25,8% maior do que para 'Não Fraude'. Toda a caixa (AIQ) do grupo 'Fraude' está deslocada para cima, indicando que o perfil de fraude se concentra em uma faixa etária mais elevada.
- **Income**: De forma semelhante, a mediana da renda para o grupo 'Fraude' é 13,4% superior. A distribuição para este grupo parece mais concentrada (caixa menor) e situada em um patamar de renda mais alto.
- **Credit Risk Score**: Esta variável exibe a diferença mais pronunciada. A mediana do escore de risco de crédito é 35,9% maior para transações fraudulentas. Este é um forte indicativo de que o score de risco é um diferenciador importante. O grupo 'Não Fraude' apresenta múltiplos outliers inferiores, alguns com valores negativos, o que pode indicar clientes de baixíssimo risco ou uma peculiaridade na escala da pontuação.

- **Velocity 24H**: Curiosamente, a mediana para 'Fraude' é 3,6% menor. Este resultado é contra intuitivo, pois se poderia esperar uma maior atividade transacional em casos de fraude. Isso pode sugerir que os fraudadores operam de forma a evitar disparar alertas baseados em regras simples de volume ou velocidade. O grupo 'Não Fraude' exibe uma quantidade significativa de outliers em ambas as extremidades, indicando uma grande variabilidade no comportamento transacional legítimo.

5. Análise Descritiva Aprofundada

5.1. Medidas de Posição Central: Média vs. Mediana



Dados Observados:

- **Média:** 33.69 anos
- **Moda:** 30.00 anos
- **Mediana:** 30.00 anos
- **Q1:** 20.00 anos
- **Q2:** 30.00 anos
- **Q3:** 40.00 anos

O gráfico "Medidas de Posição Central" compara a média e a mediana. Para "Customer Age", "Income" e "Credit Risk Score", a proximidade dos valores sugere distribuições com simetria moderada. "Velocity 24H" apresenta valores quase idênticos, indicando uma distribuição bastante simétrica. A ausência da variável "Session Length In Minutes" neste gráfico é um ponto de atenção, podendo indicar um problema de computação ou uma decisão de exclusão.5.2. Medidas de Dispersão Relativa: Coeficiente de Variação (CV)

Enquanto as medidas de posição indicam o centro, as medidas de dispersão quantificam a variabilidade dos dados. O **Coeficiente de Variação (CV)** é uma medida de dispersão relativa, calculada como a razão entre o desvio padrão (s) e a média ($\bar{x}$), expressa em porcentagem (CV = ($s/\bar{x}$) × 100%). Sua principal vantagem é ser adimensional, permitindo a comparação da variabilidade entre conjuntos de dados com diferentes unidades ou escalas.

Medidas de Dispersão:

- **Amplitude:** 80.00 anos
- **Variância:** 144.62
- **Desvio Padrão:** 12.03 anos
- **CV:** 35.7%

Interpretação do CV:

- **Baixa Dispersão (Dados Homogêneos):** CV ≤ 15%
- **Média Dispersão:** 15% < CV ≤ 30%
- **Alta Dispersão (Dados Heterogêneos):** CV > 30%

A análise do gráfico "Coeficiente de Variação (%)" revela:

- **Velocity 24H (CV = 20,6%):** Apresenta **média dispersão**, indicando consistência em torno da média e estabilidade.
- **Customer Age (CV = 37,0%):** Exibe **alta dispersão**.
- **Income (CV = 52,1%):** Também com **alta dispersão**.
- **Credit Risk Score (CV = 60,9%):** Mostra **dispersão muito alta**.
- **Session Length In Minutes (CV = 103,8%):** Apresenta **dispersão extremamente alta**. Um CV acima de 100% significa que o desvio padrão é maior que a média, indicando imensa variabilidade e influência de valores extremos.

6. Avaliação da Qualidade e Relevância dos Dados

6.1 Critérios de Avaliação

A qualidade dos dados foi avaliada mediante três critérios principais:

1. **Variabilidade Aceitável:** CV < 30%
2. **Distribuição Simétrica:** |Assimetria| < 2
3. **Adequação Amostral:** N ≥ 1.000.000

6.2 Resultados da Avaliação por Variável

Variável	CV (%)	Assimetria	Adequação
Customer Age	35,7	0,48	Sim
Income	51,6	-0,39	Sim
Credit Risk Score	53,2	0,30	Sim
Velocity 24h	31,0	0,33	Sim
Session Length	106,5	3,30	Sim

6.3 Interpretação Contextualizada

Contrariamente a interpretações excessivamente conservadoras, a variabilidade observada é **característica esperada** em dados bancários reais:

- **CV de 35,7% para idade:** Reflete diversidade etária natural da clientela bancária
- **CV de 51,6% para renda:** Espelha desigualdade socioeconômica realística
- **CV de 53,2% para escore de risco:** Evidência adequada diferenciação de perfis

Esta variabilidade constitui **força analítica**, não limitação, proporcionando poder discriminatório essencial para detecção de fraude.

7. Análise Crítica e Confiabilidade

7.1 Representatividade Estatística

Com 1.000.000 de observações, o dataset proporciona:

Robustez Estatística: Os intervalos de confiança são notavelmente estreitos.
Detecção de Padrões Raros: Apresenta alta capacidade para identificar eventos de baixa frequência.
Generalização Populacional: Possui representatividade superior em comparação com amostras convencionais.

7.2 Adequação para Modelagem Preditiva

Os dados demonstram **alta adequação** para desenvolvimento de modelos de machine learning:

Pontos Fortes:

- Volume de dados suficiente para técnicas avançadas
- Diversidade de variáveis (demográficas, comportamentais, transacionais)
- Padrões discriminatórios claros entre classes
- Ausência de valores faltantes

7.3 Confiabilidade dos Resultados

Os padrões identificados apresentam **alta confiabilidade** baseada em:

1. **Significância Estatística:** Diferenças substanciais entre grupos (>20%)
2. **Consistência Contextual:** Achados alinhados com literatura de fraude bancária
3. **Volume de Evidências:** Base de 1 milhão de observações
4. **Robustez Metodológica:** Múltiplas medidas descritivas convergentes

8. Conclusões

8.1 Conclusões Principais

1. **Perfil de Fraude Bem-Definido:** Fraudadores apresentam características demográficas e comportamentais distintivas, particularmente idade superior, renda elevada e scores de risco altos.

2. **Qualidade de Dados Adequada:** Apesar da variabilidade elevada em algumas variáveis, o dataset possui qualidade suficiente para análises preditivas robustas.

3. **Padrões Comportamentais Relevantes:** A velocidade transacional reduzida em casos de fraude sugere sofisticação dos fraudadores em evitar detecção automatizada.

4. **Base Sólida para Modelagem:** O volume e diversidade dos dados proporcionam fundação excelente para desenvolvimento de modelos preditivos.

Referências Metodológicas

Conceitos Estatísticos Aplicados

Medidas de Posição:

- Média Aritmética: Σx/n, medida de tendência central
- Mediana: Valor que divide a distribuição em 50%-50%
- Quartis: Valores que dividem a distribuição em quartos
- Moda: Valor de maior frequência na distribuição

Medidas de Dispersão:

- Variância: $E[(X-\mu)^2]$, medida de dispersão absoluta
- Desvio Padrão: √Variância, mesma unidade dos dados originais
- Coeficiente de Variação: $(\sigma/\mu)\times100\%$, medida de dispersão relativa
- Amplitude: Xmax - Xmin, medida de extensão da distribuição

Medidas de Forma:

- Assimetria: Medida de desvio da simetria (Skewness)
- Curtose: Medida de achatamento da distribuição (Kurtosis)

Técnicas de Análise:

- Estatística Descritiva: Resumo numérico dos dados
- Análise Exploratória: Identificação de padrões e anomalias
- Análise Comparativa: Comparação entre grupos de interesse
- Avaliação de Qualidade: Critérios de adequação dos dados

Fontes bibliográficas:

1. Estatística Prática Para Cientistas De Dados - Livrarias Curitiba, acessado em Agosto27, 2025, https://www.livrariascuritiba.com.br/estatistica-pratica-para-cientistas-de-dados-lv449151/p
2. Boxplot: Desvendando os segredos dos dados - DOCNIX, acessado em Agosto 27, 2025, https://docnix.com.br/ferramentas-metodos/boxplot-desvendando-os-segredos-dos-dados/
3. Box Plot: o que é, para que serve e como construir? - FM2S, acessado em

Agosto 27, 2025, https://www.fm2s.com.br/blog/como-elaborar-um-box-plot
4. Box Plot: O que é e Como analisar e interpretar esse gráfico? - Escola EDTI, acessado em Agosto 27, 2025, https://www.escolaedti.com.br/o-que-e-um-box-plot/
5. Introdução à estatística: média, mediana e moda (vídeo) - Khan Academy, acessado em Agosto 27, 2025, https://pt.khanacademy.org/math/em-mat-estatistica/x5d13d3b4b5b8c419:medidas-de-tendencia-central/x5d13d3b4b5b8c419:media-mediana-e-moda/v/statistics-intro-mean-median-and-mode
6. O que são medidas de tendência central? Média, mediana e moda, acessado em Agosto 27, 2025, https://www.blog.psicometriaonline.com.br/medidas-de-tendencia-central-media-mediana-e-moda/
7. Média versus Mediana - Bachmann, acessado em Agosto 27, 2025, https://blog.bachmann.com.br/2020/07/media-versus-mediana/
8. Coeficiente de variação: o que é, como calcular - Brasil Escola - UOL, acessado em Agosto 27, 2025, https://brasilescola.uol.com.br/matematica/coeficiente-variacao.htm
9. Turning the Tables: Biased, Imbalanced, Dynamic Tabular Datasets for ML Evaluation - Cornell university, acessado em Agosto 18, 2025, https://arxiv.org/abs/2211.13358
10. Bank Account Fraud Dataset Suite, Kaggle - acessado em Agosto 18, 2025 https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022