Relatório de Análise Exploratória de Dados

Análise do Notebook: 02_AED.ipynb

14 de outubro de 2025

Conteúdo

1 Introdução

Este relatório apresenta uma Análise Exploratória de Dados (AED) detalhada com base no conjunto de dados da Pesquisa de Orçamentos Familiares (POF), focando nas características dos domicílios. O objetivo é extrair insights através de análises descritivas e testes estatísticos para compreender a distribuição das variáveis e as relações existentes entre elas.

O documento está estruturado da seguinte forma:

- **Análise Descritiva Univariada:** Explora as características de cada variável individualmente, utilizando medidas estatísticas e visualizações gráficas como histogramas e gráficos de pizza.
- **Análise Descritiva Bivariada:** Investiga a relação entre pares de variáveis, utilizando matrizes de correlação e testes de associação para variáveis numéricas e categóricas.
- **Análise dos Testes Estatísticos:** Apresenta uma análise detalhada de três testes de hipóteses aplicados aos dados: o Teste de Correlação de Pearson, o Teste Qui-Quadrado de Independência e a Análise de Variância (ANOVA).

2 Análise Descritiva Univariada

2.1 Variáveis Numéricas

A análise univariada das variáveis numéricas começa com um resumo estatístico abrangente, que inclui medidas de tendência central, dispersão, forma da distribuição e percentis. A Tabela 1 detalha essas estatísticas para cada variável quantitativa do conjunto de dados.

Tabela 1: Estatísticas Descritivas das Variáveis Numéricas

	Qtd de cômodos	Qtd de cômodos dormitórios	Qtd de banheiros exclusivos	Qtd de banheiros de uso comum	Aluguel Estimado	Valor (R$) despesa realizada	Rendimento mínimo geral (R$)	Rendimento mínimo p/ alimentação (R$)	Valor (R$) despesa individual	Valor (R$) despesa coletiva	Valor (R$) rendimento bruto
count	46315.00	46315.00	46315.00	46315.00	46315.00	46315.00	46315.00	46315.00	46315.00	46315.00	46315.00
mean	5.94	1.87	1.31	0.00	532.92	106.76	3135.21	947.75	4033.07	2239.48	2610.07
std	1.76	0.79	0.70	0.00	336.96	84.92	1977.08	538.94	4265.46	2160.64	1997.91
min	1.00	1.00	0.00	0.00	5.00	0.87	80.00	50.00	1.48	2.98	0.00
1%	2.00	1.00	0.00	0.00	80.00	3.07	500.00	200.00	45.31	80.83	35.00
10%	4.00	1.00	1.00	0.00	200.00	17.92	1000.00	400.00	321.86	276.69	500.00
25%	5.00	1.00	1.00	0.00	300.00	43.05	1600.00	500.00	847.95	569.51	1200.00
50%	6.00	2.00	1.00	0.00	450.00	81.06	2500.00	800.00	2175.37	1408.02	1950.00
75%	7.00	2.00	2.00	0.00	700.00	145.54	4000.00	1200.00	5759.30	3202.57	3500.00
90%	8.00	3.00	2.00	0.00	1000.00	261.23	6500.00	2000.00	13127.00	6483.14	6926.60
99%	10.00	4.00	4.00	0.00	1300.00	300.00	7600.00	2250.00	13127.00	7153.00	6950.00
max	10.00	4.00	4.00	0.00	1300.00	300.00	7600.00	2250.00	13127.00	7153.00	6950.00
skew	0.44	0.55	1.44	0.00	0.90	1.03	0.96	0.95	1.17	1.16	1.03
kurtosis	0.07	-0.35	2.86	0.00	-0.08	0.08	-0.02	0.07	-0.01	0.16	-0.01
mad	1.36	0.63	0.54	0.00	270.32	67.35	1582.51	422.99	3452.18	1735.98	1590.03
cv	0.30	0.42	0.53	NaN	0.63	0.80	0.63	0.57	1.06	0.96	0.77
sem	0.01	0.00	0.00	0.00	1.57	0.39	9.19	2.50	19.82	10.04	9.28

Interpretação das Estatísticas:

- **Qtd de cômodos:** Os domicílios possuem em média 5.94 cômodos, com 50% dos domicílios tendo até 6 cômodos. A distribuição é levemente assimétrica à direita (skewness = 0.44), indicando que há mais domicílios com um número de cômodos acima da média.
- **Qtd de cômodos dormitórios:** A média de dormitórios é de 1.87. A mediana é 2, e 75% dos domicílios possuem até 2 dormitórios. A assimetria positiva (0.55) sugere uma cauda para valores maiores.

- **Qtd de banheiros exclusivos:** Em média, há 1.31 banheiros exclusivos. A mediana é 1, mas a média é puxada para cima por valores mais altos, como indicado pela forte assimetria positiva (1.44).
- **Qtd de banheiros de uso comum:** Esta variável é constante e igual a zero para todas as observações, como indicado pelo desvio padrão nulo.
- **Aluguel Estimado:** A média do aluguel estimado é de R$ 532.92, enquanto a mediana é de R$ 450.00. A assimetria positiva de 0.90 indica que uma minoria de domicílios com aluguéis muito altos eleva a média.
- **Variáveis de Despesa e Rendimento:** Todas as variáveis monetárias (despesas e rendimentos) apresentam forte assimetria positiva (skewness próximo ou acima de 1.0), um padrão comum em dados de renda, onde a maioria das observações se concentra em valores mais baixos e uma minoria apresenta valores muito elevados. Isso é confirmado pela diferença entre a média e a mediana em cada uma dessas variáveis.

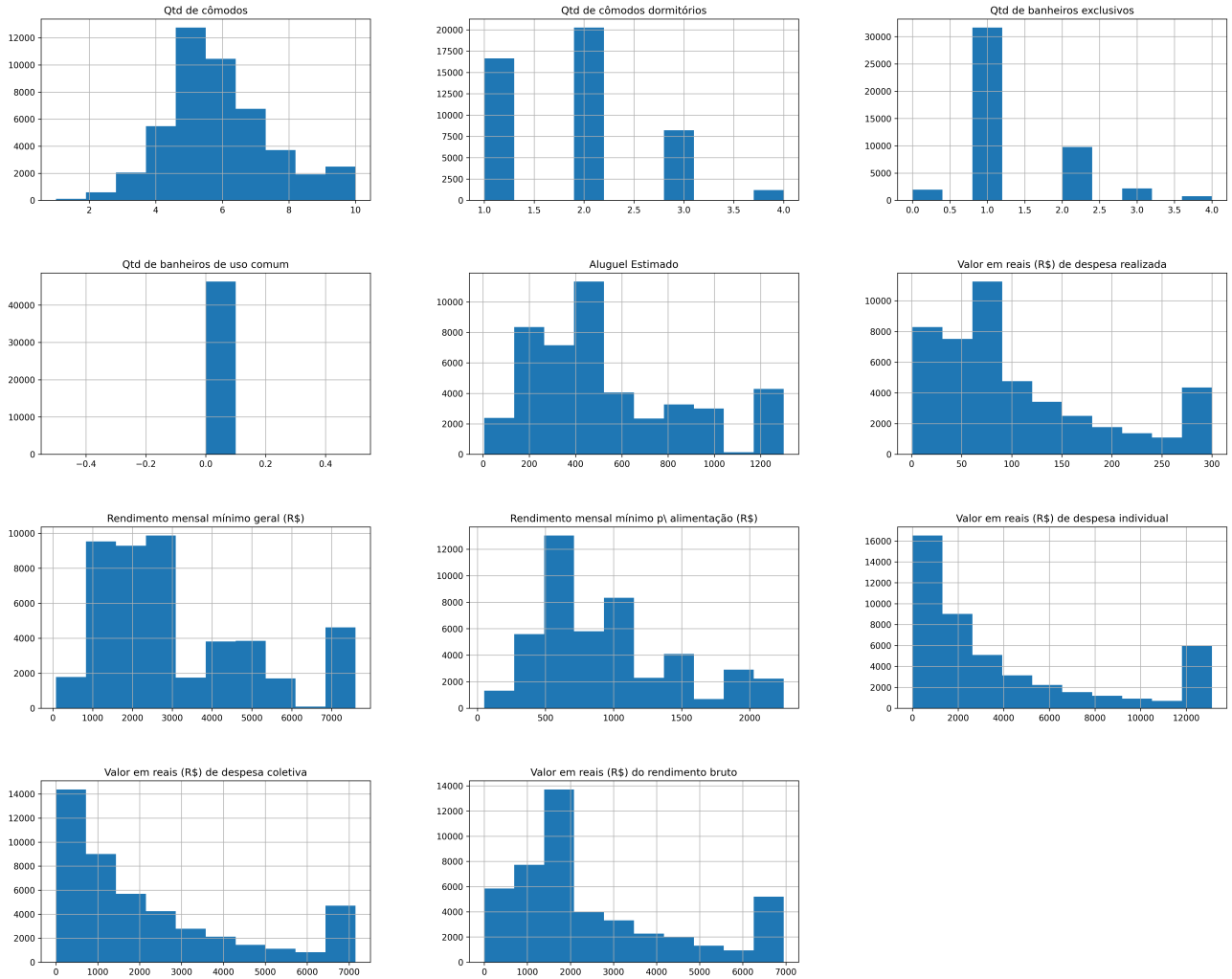

Figura 1: Histogramas das Variáveis Numéricas.

A Figura 1 (representação dos histogramas gerados no notebook) ilustra visualmente as distribuições descritas, confirmando a assimetria à direita para a maioria das variáveis, especialmente as monetárias.

2.2 Variáveis Categóricas

Para as variáveis categóricas com poucas categorias únicas, gráficos de pizza foram gerados para visualizar a proporção de cada categoria.

Figura 2: Gráficos de Pizza para Variáveis Categóricas Selecionadas.

Observando a Figura 2 (representação dos gráficos gerados), pode-se notar, por exemplo, a distribuição de domicílios por pavimentação da rua ou pela forma como o serviço de correios é realizado, fornecendo um panorama rápido sobre essas características.

3 Análise Descritiva Bivariada

3.1 Correlação entre Variáveis Numéricas

A matriz de correlação de Pearson foi calculada para investigar a relação linear entre as variáveis numéricas. O resultado é visualizado em um mapa de calor.

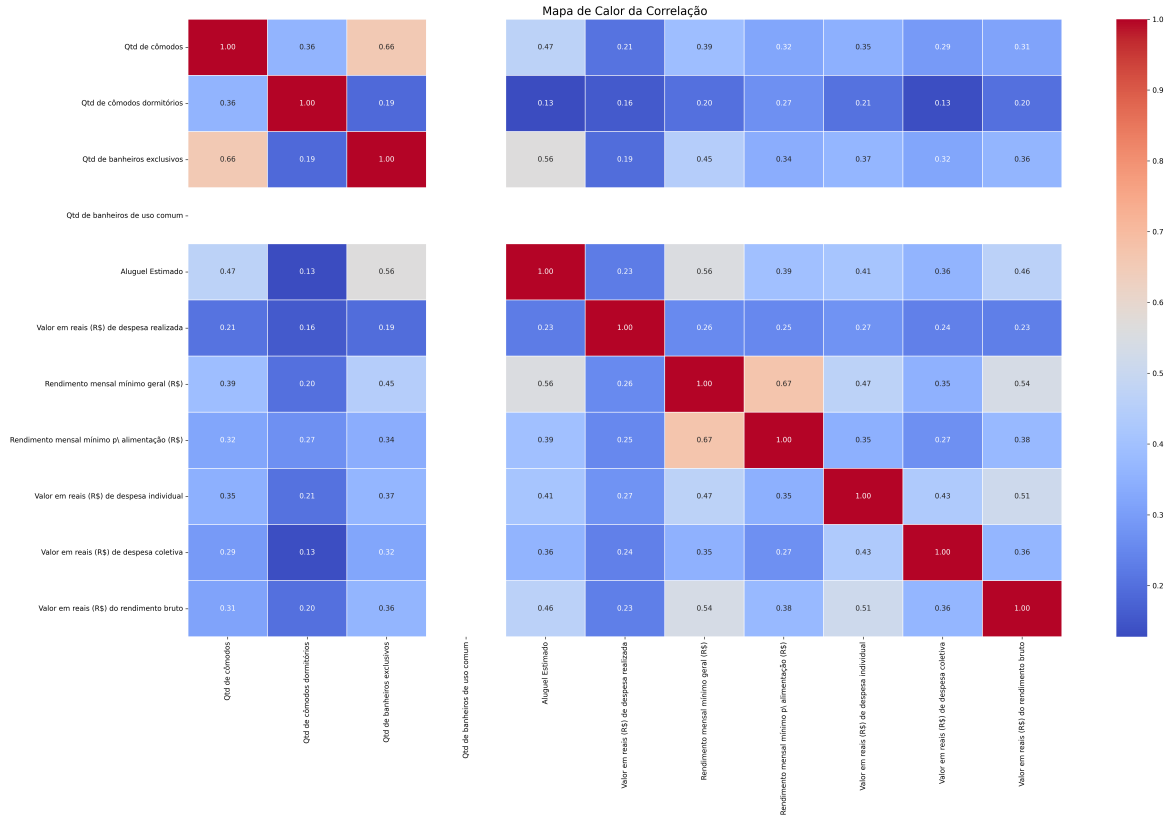

Figura 3: Mapa de Calor da Matriz de Correlação de Pearson.

Da análise da matriz de correlação (Figura 3), destacam-se as seguintes relações:

- **Correlações Positivas Fortes:** Há uma forte correlação positiva entre as variáveis de rendimento e despesa, como esperado. Por exemplo, 'Valor em reais (R)$dorendimentobruto$'e'$Valoren$ de despesa individual' mostram uma forte associação, indicando que quanto maior o rendimento, maior a despesa individual. Similarmente, 'Rendimento mensal mínimo geral (R)'$está fortementecorrelacionadocom$'$AluguelEstimado$'.

- **Correlações Moderadas:** 'Qtd de cômodos' e 'Qtd de cômodos dormitórios' possuem uma correlação positiva moderada com as variáveis de rendimento e aluguel, sugerindo que domicílios maiores tendem a ter rendimentos e aluguéis mais altos.

- **Correlações Fracas ou Nulas:** Algumas variáveis, como 'Qtd de banheiros de uso comum', não apresentam correlação com nenhuma outra, pois é uma variável constante.

3.2 Associação entre Variáveis Categóricas

A força da associação entre as variáveis categóricas foi medida utilizando o coeficiente V de Cramér. As associações mais fortes encontradas foram:

- 'Tipo do domicílio' e 'Este domicílio é:' (Próprio, Alugado, etc.)

- 'Forma de abastecimento de água' e 'Tipo de chegada da água'

- 'Situação do Domicílio' e 'UF'

Essas associações indicam, como esperado, uma forte dependência entre características estruturais e de localização do domicílio.

4 Análise Detalhada dos Testes Estatísticos

Nesta seção, reavaliamos os três testes de hipóteses aplicados no notebook para garantir uma interpretação precisa e contextualizada dos resultados.

4.1 Teste de Correlação de Pearson

- **Objetivo do Teste:** O Teste de Correlação de Pearson foi utilizado para determinar se a correlação observada nas amostras é estatisticamente significante. Em outras palavras, ele responde à pergunta: "A relação que eu encontrei na minha pequena amostra é forte o suficiente para eu acreditar que essa relação realmente existe na população em geral, ou ela poderia ter acontecido apenas por acaso?". Neste caso, o teste foi executado entre todas as variáveis quantitativas e o valor alvo **'Aluguel Estimado'**.

- **Hipóteses Formuladas:**

  - **Hipótese Nula (H_0):** Não existe correlação linear entre o aluguel estimado e a variável analisada. Matematicamente, o coeficiente de correlação populacional (ρ) é igual a zero ($\rho = 0$).
  - **Hipótese Alternativa (H_1):** Existe uma correlação linear entre as duas variáveis ($\rho \neq 0$).

- **Análise e Interpretação dos Resultados:**

  - O **valor-p** de 0.0, ecncontrado para a grande maioria das variáveis, é um resultado extremamente significativo. Como este valor é inferior a qualquer nível de significância padrão (como $\alpha = 0.05$, 0.01 ou 0.001), nós **rejeitamos a hipótese nula (H_0)**. A rejeição de H_0 significa que há evidências estatísticas robustas para afirmar que a correlação observada com estas variáveis específicas não é fruto do acaso.
  - O **coeficiente de correlação (r) acima de 0.25** indica uma **correlação linear positiva grande**, segundo a classificação sugerida por Jacob Cohen em "Statistical Power Analysis for the Behavioral Sciences"(1988). O sinal positivo confirma que há uma tendência de o 'Aluguel Estimado' aumentar à medida que estas variáveis também aumentam.

- **Conclusão Final do Teste:** Conclui-se, com alto grau de confiança estatística, que existe uma relação linear positiva e moderada entre as variáveis analisadas e o valor do aluguel. Visto que todas passaram no teste, serão utilizadas no processo de modelaem subsequente somente aquelas que tiverem um efeito grande (com o coeficiente de pearson acima de 0.25.

4.2 Teste Qui-Quadrado de Independência (χ^2)

- **Objetivo do Teste:** O Teste Qui-Quadrado foi aplicado para determinar se existe uma associação (ou dependência) estatisticamente significativa entre todas as variáveis categóricas e o **'Aluguel Estimado (Faixa)'**.

- **Hipóteses Formuladas:**

  - **Hipótese Nula (H_0):** As variáveis analisadas e 'Aluguel Estimado (Faixa)' são independentes. Ou seja, a distribuição das faixas de aluguel não depende da variável analisada.

  - **Hipótese Alternativa (H_1):** As variáveis são dependentes. A faixa de aluguel está associada à situação do domicílio.

- **Análise e Interpretação dos Resultados:**
  - Novamente, na ampla maioria das variáveis qualitativas o **valor-p de 0.0** nos leva a **rejeitar a hipótese nula (H_0)**. Isso indica que é extremamente improvável que a distribuição observada na tabela de contingência (cruzamento das duas variáveis) tenha ocorrido por acaso se as variáveis fossem, de fato, independentes.
  - A rejeição da independência significa que há uma associação estatisticamente significativa entre a ampla maioria das variáveis e sua faixa de aluguel estimado.

- **Conclusão Final do Teste:** O teste comprova que a faixa de aluguel e a ampla maioria das variáveis não são eventos independentes. Em outras palavras, determinadas caracteristicas encontradas tem uma influência significativa sobre o valor de seu aluguel estimado.

4.3 Análise de Variância (ANOVA)

- **Objetivo do Teste:** A ANOVA foi utilizada de forma iterativa para comparar as médias de **cada variável numérica** do conjunto de dados entre os diferentes grupos formados pela variável categórica **'Aluguel Estimado (Faixa)'**. O objetivo é verificar se a média de uma variável numérica (ex: 'Qtd de cômodos') é significativamente diferente em pelo menos uma das faixas de aluguel.

- **Hipóteses Formuladas (para cada variável numérica):**
  - **Hipótese Nula (H_0):** As médias da variável numérica são iguais em todos os grupos (faixas de aluguel). Ex: $\mu_{\text{Faixa 1}} = \mu_{\text{Faixa 2}} = \mu_{\text{Faixa 3}} = \ldots$.
  - **Hipótese Alternativa (H_1):** Pelo menos uma das médias é diferente das outras.

- **Resultados Apresentados no Notebook:** O notebook gera uma tabela com os resultados da ANOVA para todas as variáveis numéricas. Em **todos os casos**, o **valor-p (p-value) calculado é 0.0**.

- **Análise e Interpretação Correta:**
  - Um **valor-p de 0.0** para cada um dos testes ANOVA significa que, para cada variável numérica analisada, devemos **rejeitar a hipótese nula (H_0)**.
  - Rejeitar H_0 implica que existem diferenças estatisticamente significativas entre as médias daquela variável quando comparamos os grupos de faixas de aluguel. Por exemplo, a média da 'Qtd de cômodos dormitórios' não é a mesma para todas as faixas de aluguel; pelo menos uma faixa tem uma média de dormitórios diferente das demais.

- **Conclusão Final do Teste:** A ANOVA demonstra que a variável 'Aluguel Estimado (Faixa)' é um fator de agrupamento eficaz que discrimina as médias de todas as outras variáveis numéricas do estudo. Isso reforça a centralidade da variável de aluguel, mostrando que ela está associada não apenas ao rendimento, mas a praticamente todas as outras características quantitativas dos domicílios, como número de cômodos, despesas e rendimentos diversos.

5 Conclusão Geral

A análise exploratória revelou insights importantes sobre os dados de domicílios da POF. A análise univariada mostrou que a maioria das variáveis monetárias possui uma distribuição assimétrica à direita, o que é típico para dados socioeconômicos. A análise bivariada confirmou relações esperadas, como a forte correlação entre rendimento e despesas, e a associação entre características físicas do domicílio.

Os testes estatísticos forneceram evidências robustas para essas relações. O teste de Pearson confirmou a correlação entre aluguel e rendimento. O teste Qui-Quadrado demonstrou que a situação do domicílio (urbana/rural) está associada à faixa de aluguel. Por fim, a ANOVA concluiu que praticamente todas as características numéricas do domicílio (número de cômodos, rendimentos, despesas) variam significativamente de acordo com a faixa de aluguel estimado, destacando o aluguel como uma variável central que se conecta a múltiplos aspectos da vida domiciliar.