Universidade de Brasília
Programa de Pós-graduação em Computação Aplicada – PPCA
Análise Estatística de Dados e Informações (AEDI)

Relatório - ANOVA

Aluno: João Robson Santos Martins

Professor:
João Gabriel de Moraes Souza

8 de dezembro de 2024

1 Introdução

Este relatório apresenta uma análise estatística de um *dataset* de imóveis da cidade de Ames, no Iowa, EUA. O dado foi estruturado por [1] e sua documentação pode ser vista aqui.

A ideia é comparar o preço de venda médio desses imóveis em relação a três características presentes no dado e indicar se tais características determinam (com significância estatística) alguma tendência no preço, partindo da análise de variância.

O *notebook* com a análise completa pode ser encontrado aqui.

2 ANOVA

Para determinar se variáveis discretas ou categóricas influenciam significativamente o preço de venda, é possível utilizar a **Análise de Variância (ANOVA)** [2]. Essa técnica compara as médias de dois ou mais grupos para identificar se as diferenças entre elas são estatisticamente significativas.

2.1 Fatores

Na ANOVA, **fatores** são as variáveis independentes categóricas ou discretas que queremos testar para avaliar se elas afetam a variável dependente (neste caso, o preço de venda). Cada fator pode ter dois ou mais **níveis** ou valores (categorias).

No *dataset* da cidade de Ames, por exemplo, podemos ter como fator o bairro (**Neighbourhood**), com os níveis sendo os bairros em si (Bloomington Heights, Bluestem, Briardale, etc.).

2.2 Tipos de ANOVA

One-Way ANOVA (ANOVA de um fator): Utilizada quando há apenas um fator (variável independente) com dois ou mais níveis.
Exemplo: Comparar o preço de venda entre diferentes tipos de bairros (residencial, comercial, industrial).

Two-Way ANOVA (ANOVA de dois fatores): Utilizada para avaliar dois fatores simultaneamente e investigar interações entre eles.
Exemplo: Comparar o preço de venda considerando o tipo de bairro e a qualidade da construção.

Factorial ANOVA (ANOVA de n fatores): Extensão do Two-Way ANOVA, incluindo n fatores, cada um com dois ou mais níveis.
Exemplo: Comparar o preço de venda considerando o tipo de bairro, a qualidade da construção e o tipo de garagem.

Repeated Measures ANOVA: Usada para comparar valores para três ou mais grupos relacionados, em que os mesmos indivíduos são medidos várias vezes sob diferentes condições ou pontos no tempo [7]. Esse método é útil na análise de dados em que as observações não são independentes, pois vêm dos mesmos registros repetidamente.

2.3 Pressupostos da ANOVA

Os pressuspostos que precisam ser satisfeitos para a ANOVA são [3]:

- **Independência das observações:** A observação de uma amostra não deve influenciar os valores de outra amostra, ou seja, o valor de uma observação não deve fornecer informações sobre outra.

- **Homogeneidade de variâncias:** As variâncias de cada grupo devem ser aproximadamente iguais. Isso significa que a dispersão dos valores em torno da média de cada grupo deve ser semelhante entre os diferentes grupos.

$$H_0 = \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 \tag{1}$$

- **Normalidade dos dados:** Os dados de cada grupo devem seguir uma distribuição normal.

2.4 Fórmulas

Para utilizar a ANOVA, deve-se calcular uma estatística F, que é dada pela razão entre a variância intergrupo e a variância intragrupo [3]. Essas duas medidas podem ser obtidas por meio do cálculo da razão entre o **quadrado médio entre os grupos** (QME) e o **quadrado médio dentro dos grupos** (QMD):

$$F = \frac{QME}{QMD} \tag{2}$$

onde:

$$QME = \frac{SQE}{k-1} = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2}{k-1} \tag{3}$$

$$QMD = \frac{SQD}{N-k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2}{N-k} \tag{4}$$

- X_{ij}: Observação j no grupo i.
- $\bar{X}_i$: Média do grupo i.
- $\bar{X}$: Média geral (média de todas as observações).
- n_i: Número de observações no grupo i.
- N: Número total de observações.
- k: Número de grupos.

2.5 Variáveis Escolhidas

No caso da análise do preço de venda, a ANOVA será aplicada às seguintes variáveis:

- Número de vagas na garagem (**Garage Cars**);
- Ano de construção (**Year Built**);
- Bairro do imóvel (**Neighborhood**).

As duas primeiras variáveis foram selecionadas devido à alta correlação com o preço de venda. Já o bairro foi escolhido porque, em geral, a localização do imóvel tem grande impacto no preço, mesmo quando dois imóveis possuem características semelhantes.

3 Análise exploratória

Em suma, análise exploratória dos dados destacou os seguintes padrões presentes no dado:

- Muitas *features* possuem uma alta contagem de registros com valor 0 ("Pool Area"e "Mas Vnr Area", por exemplo). Isso possivelmente se deve ao fato de imóveis em que esse critério não se aplica. Ou seja, imóveis sem piscina tem área da piscina sendo 0, por exemplo.
- Os preços de venda possuem uma distribuição assimétrica à direita, contendo muitos *outliers* à direita (valores mais caros), conforme a figura 1.

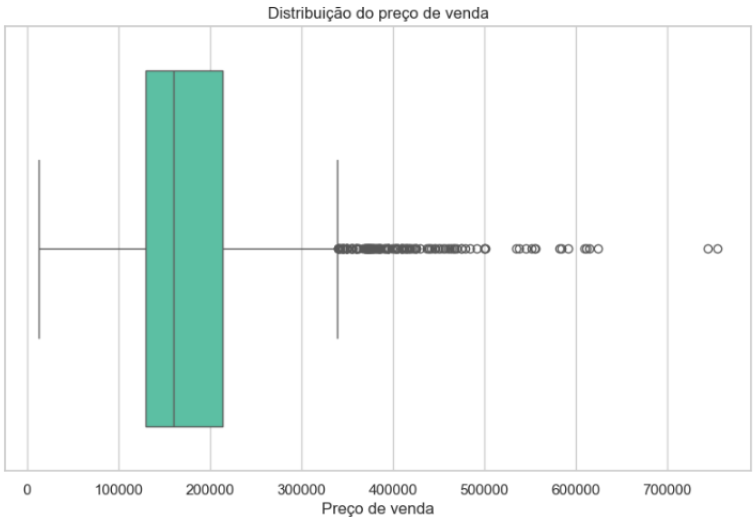

Figura 1: Boxplot dos preços de venda

3.1 Correlação entre *features*

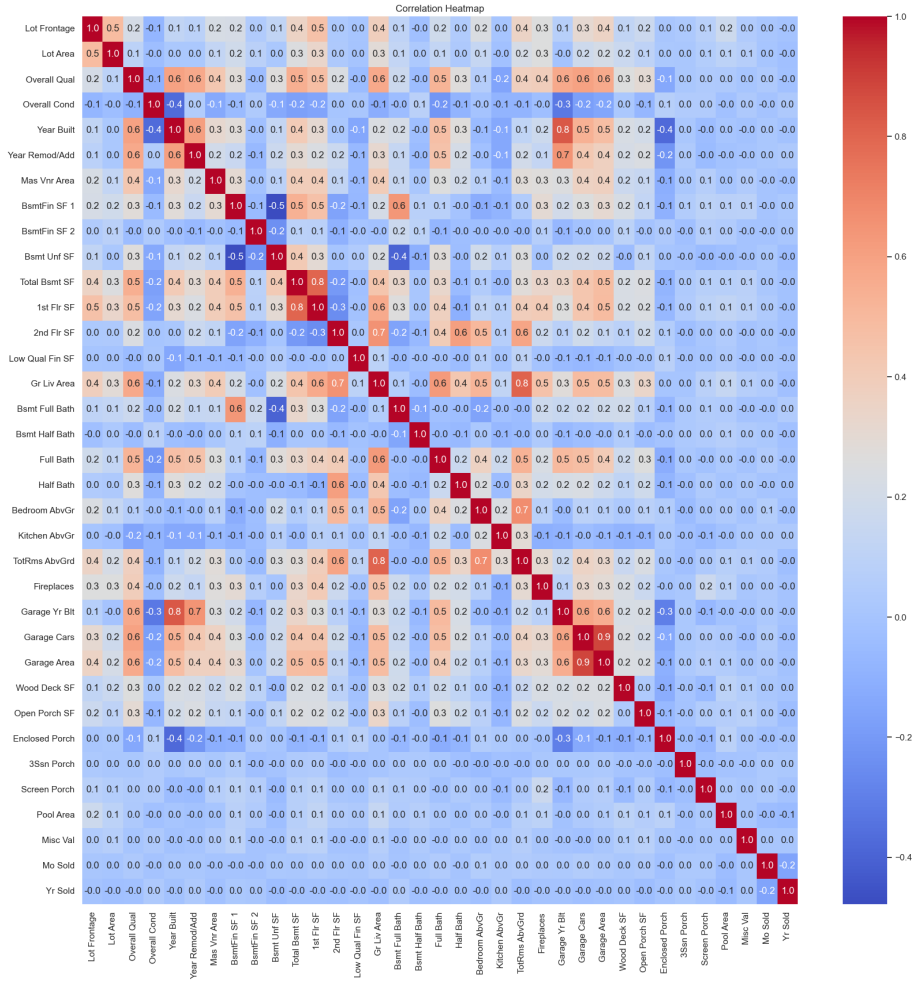

Figura 2: Correlação das *features* com preço de venda

Na análise de correlação entre *features*, conforme a figura 3, os seguintes pontos se destacam:

- A qualidade geral ("Overall Qual") está fortemente correlacionada com o ano de construção e a capacidade da garagem.
- A área da piscina, que é igual a zero para quase todos os imóveis, e o mês da venda não apresentam correlação significativa com praticamente todas as outras variáveis.
- O tamanho da área de convivência apresenta uma correlação bastante alta (¿ 0.8) com o número de cômodos.
- A área do primeiro piso tem uma correlação elevada (0.8) com a área do porão.
- Existe uma forte correlação negativa entre "Bsmt Unf SF"(espaço do porão inacabado) e "BsmtFin SF 1"(espaço do porão acabado), indicando (de maneira bastante intuitiva) que, à medida que a área do porão acabado aumenta, a área do porão inacabado diminui.

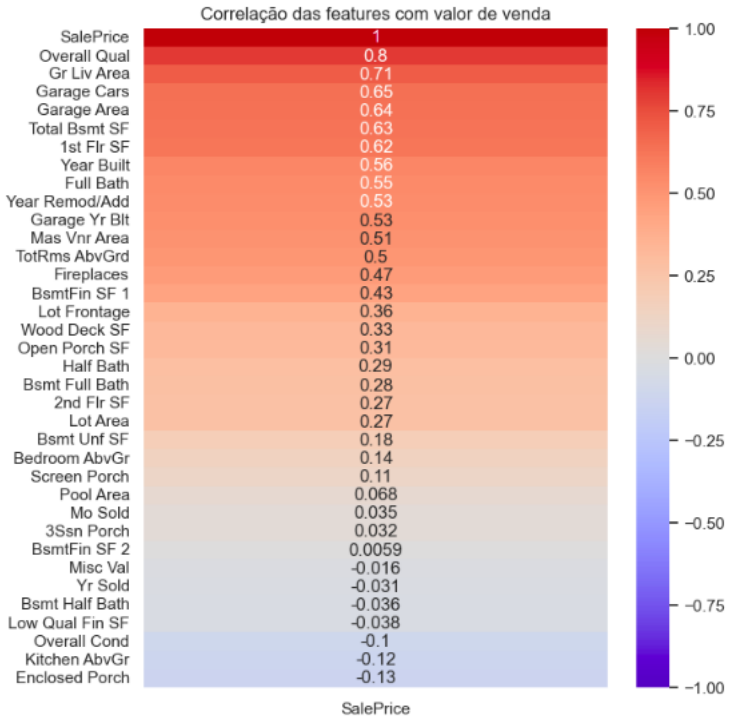

Figura 3: Correlação das *features* com preço de venda

3.2 Correlação entre *features* e preço de venda

Como mostrado na figura 3, das variáveis com maior correlação com o preço de venda, a qualidade geral do imóvel (que indica a qualidade do material e acabamento usado e é representado por uma nota de 1 a 10) é a que apresenta a correlação mais alta com o valor de venda.

Em sequência, tem-se o tamanho da área de convivência, a capacidade e área da garagem e a área do sótão com as maiores correlações.

No caso das variáveis com correlação negativa, tem-se que casas com uma área de varanda cercada maior e com mais cozinhas tem uma tendência (leve) a ter um preço menor.

No caso de varandas cercadas, conforme análise abaixo, se vê uma correlação negativa (-0.4) com a idade do imóvel. Ou seja, quanto mais antiga a casa, maior a área de varanda cercada e vice-versa.

4 Resultados da análise de pressupostos do ANOVA

Os pressupostos necessários para a realização do teste ANOVA foram verificados, com os seguintes resultados:

- **Normalidade**: foi verificada por meio de gráficos Q-Q plots [9] e pelo teste de Shapiro-Wilk [4]. Ambos os métodos indicaram que os dados não seguem uma distribuição normal para todos os níveis dos grupos. Nos Q-Q plots, especialmente nos quantis superiores, as observações se desviaram da linha diagonal, sugerindo a presença de caudas pesadas e valores extremos (outliers). Exemplos de Q-Q plot para os grupos da variável que armazena o número de vagas no imóvel podem ser vistos na figura 4. A distribuição dos valores por grupo também foi gerada, conforme exemplo para número de vagas do gráfico 5.

- **Homocedasticidade**: foi testada utilizando o teste de Levene [5]. Os resultados indicaram que as variâncias não são homogêneas entre os grupos, violando o pressuposto de homocedasticidade.

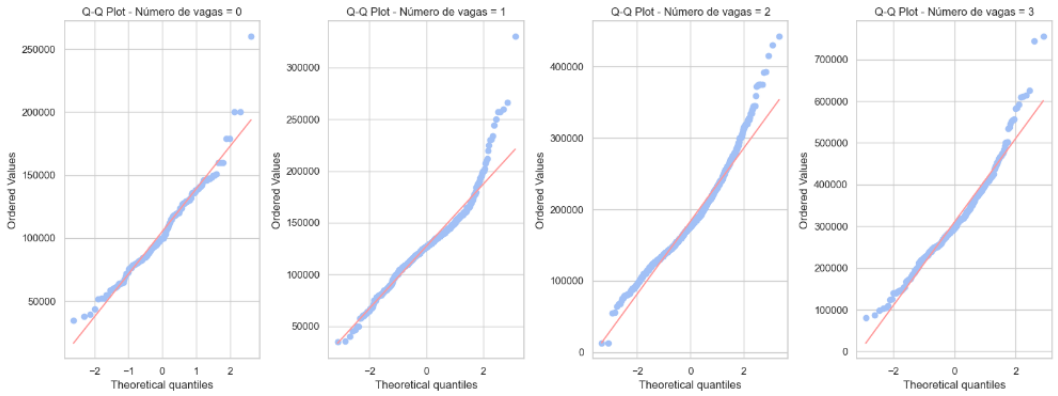

Figura 4: Q-Q plots para número de vagas

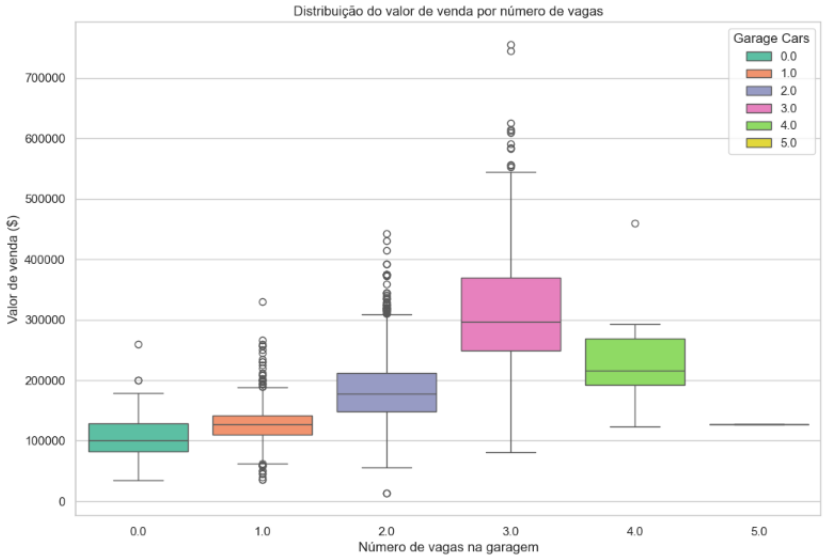

Figura 5: Boxplots para grupos por números de vagas

- **Independência dos Erros**: foi avaliada por meio do gráfico de resíduos (residual plot) e de Q-Q plots dos resíduos. A análise dos gráficos revelou padrões e a presença de outliers, especialmente em valores extremos, sugerindo que os erros não são independentes.

5 Teste não-paramétrico

Como os pressupostos da ANOVA não foram atendidos pelos dados, foi aplicado o **teste não-paramétrico de Kruskal-Wallis** para verificar se há diferenças significativas entre os grupos em relação ao preço de venda [6]. Esse teste é uma alternativa robusta à ANOVA quando os pressupostos de normalidade e homocedasticidade são violados. Ele avalia se as amostras provêm da mesma distribuição ou de distribuições diferentes.

5.1 Resultados do Teste de Kruskal-Wallis

- **Número de vagas na garagem:** O teste de Kruskal-Wallis para o número de vagas resultou na rejeição da hipótese nula (H_0). Isso indica que pelo menos um dos grupos possui uma distribuição de preços de venda significativamente diferente.
- **Bairro:** Para a variável **Bairro**, o teste também resultou na rejeição da hipótese nula (H_0), sugerindo que os preços de venda variam significativamente entre diferentes bairros.

- **Ano de construção:** Da mesma forma, o teste aplicado ao **Ano de construção** rejeitou a hipótese nula (H_0). Esse resultado implica que há pelo menos um grupo com ano de construção diferente que apresenta uma distribuição distinta de preços de venda.

6 Teste pós-hoc

O **Teste de Dunn** é um teste pós-hoc não-paramétrico utilizado para realizar comparações múltiplas entre grupos após a aplicação do **Teste de Kruskal-Wallis** [8]. Quando o teste de Kruskal-Wallis rejeita a hipótese nula (H_0), indicando que pelo menos um dos grupos possui uma distribuição diferente, o teste de Dunn pode ser empregado para identificar quais pares de grupos apresentam diferenças estatisticamente significativas.

Este teste é especialmente adequado para situações em que os pressupostos da ANOVA não são atendidos, como quando os dados não seguem uma distribuição normal ou quando as variâncias não são homogêneas. O teste de Dunn considera as classificações (ranks) dos dados, o que o torna robusto a desvios de normalidade e homocedasticidade.

A estatística do teste de Dunn é ajustada pelo número de comparações realizadas para controlar o erro tipo I (falsos positivos) (no caso da presente análise, foi usada a correção de **Bonferroni**). Assim, o teste garante que a probabilidade de se encontrar uma diferença significativa ao acaso seja minimizada, mesmo quando múltiplas comparações estão sendo feitas.

Dessa forma, no contexto desse teste, tem-se as seguintes hipóteses:

- H_0: As distribuições das duas populações são iguais.
- H_1: As distribuições das duas populações são diferentes.

Se existir um valor de p para determinada comparação entre um grupo i e um grupo j menor que um nível de significância (na análise, 0.05), tem-se a rejeição da hipótese nula, ou seja, um indicativo de que os grupos são significativamente diferentes.

6.1 Número de Vagas

O **teste de Dunn** para o número de vagas na garagem indica que quase todos os pares de grupos apresentam distribuições significativamente diferentes, exceto entre os grupos com **2 e 3 vagas** e **2 e 4 vagas**, onde não foi encontrada diferença estatisticamente significativa.

A comparação com o grupo de imóveis com **5 vagas** deve ser interpretada com cautela, pois este grupo contém apenas uma observação, o que torna a comparação pouco confiável.

Em resumo, os resultados sugerem que o número de vagas na garagem influencia significativamente o preço de venda dos imóveis, especialmente quando há uma grande discrepância entre os grupos (por exemplo, imóveis com **0 vagas** comparados a imóveis com **4 vagas**). Diferenças maiores no número de vagas tendem a estar associadas a diferenças maiores nos preços dos imóveis.

6.2 Ano de Construção

Para os anos de construção, dos **6.903 pares possíveis** de anos ($\frac{118\times117}{2}$), o teste de Dunn encontrou apenas **896 pares** (aproximadamente **12%**) com diferenças estatisticamente significativas entre os grupos.

Além disso, percebe-se que, em média, a diferença de idade entre os pares significativos é de aproximadamente **53 anos**.

Dessa forma, conclui-se que, em geral, o ano de construção não é um fator fortemente determinante para o preço de venda. Contudo, quando há diferença significativa, a discrepância na idade dos imóveis tende a ser grande (em média, **53 anos**).

6.3 Bairro

Para os bairros, **50% dos pares** apresentaram diferenças significativas entre as distribuições de preços de venda. Isso indica que o **bairro** é um fator bastante determinante para o valor de venda dos imóveis.

Uma possível extensão dessa análise seria investigar a **distância geográfica entre os bairros**, testando a hipótese de que, para os pares de bairros com diferença não significativa, a distância média entre eles seja menor.

7 Conclusão

Com base nos resultados apresentados, a análise revelou que, devido à violação dos pressupostos da ANOVA, o **teste não-paramétrico de Kruskal-Wallis** foi uma alternativa adequada para identificar diferenças significativas entre os grupos em relação ao preço de venda dos imóveis. Os resultados desse teste indicaram que as variáveis **número de vagas na garagem**, **bairro** e **ano de construção** influenciam de maneira significativa o preço de venda.

A aplicação do **teste de Dunn** como método pós-hoc permitiu identificar quais pares de grupos apresentavam diferenças estatisticamente significativas. Em resumo:

- O **número de vagas na garagem** demonstrou uma influência significativa no preço de venda, especialmente em casos de grandes discrepâncias entre os grupos. Comparações envolvendo grupos com **2 e 3 vagas** ou **2 e 4 vagas** não apresentaram diferenças significativas.
- Para o **ano de construção**, apenas cerca de **12%** dos pares de anos apresentaram diferenças significativas, sugerindo que, embora o ano de construção afete o preço de venda, essa influência ocorre principalmente quando há grandes diferenças de idade entre os imóveis (em média, **53 anos**).
- O **bairro** mostrou-se um fator importante para determinar o preço de venda, com **50% dos pares** de bairros apresentando diferenças significativas. Isso destaca a importância da localização do imóvel no mercado imobiliário.

Em conclusão, tanto o número de vagas na garagem quanto o bairro são fatores robustos na determinação do preço de venda dos imóveis, enquanto o ano de construção exerce influência em casos de grandes diferenças de idade.

Referências

[1] Iowa Ames. Ames, iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3):161–169, 2011.

[2] Vinicius CAMPOS. Análise de variância (anova). `https://www2.ufpel.edu.br/biotecnologia/gbiotec/site/content/paginadoprofessor/uploadsprofessor/9d165cbedea8a431bfc9c480fd259f27.pdf`. Accessed: 2024-12-08.

[3] DataCamp. Anova test: An in-depth guide with examples. `https://www.datacamp.com/tutorial/anova-test`. Accessed: 2024-12-08.

[4] Universidade Estadual de Londrina. Teste de shapiro-wilk. `https://www.uel.br/projetos/experimental/pages/arquivos/Shapiro.html`. Accessed: 2024-12-08.

[5] William F. Guthrie. Nist/sematech e-handbook of statistical methods (nist handbook 151), 2020.

[6] Samantha Lomuscio. Getting Started with the Kruskal-Wallis Test — UVA Library — library.virginia.edu. https://library.virginia.edu/data/articles/getting-started-with-the-kruskal-wallis-test. [Accessed 08-12-2024].

[7] University of Texas at Austin. Repeated measures anova. https://sites.utexas.edu/sos/guided/inferential/numeric/onecat/more-than-2/more-than-two-groups/repeated-measures-anova/. Accessed: 2024-12-08.

[8] Statology. Dunn's test for multiple comparisons. https://www.statology.org/dunns-test/. Accessed: 2024-12-08.

[9] Statology. How to use q-q plots to check normality. https://www.statology.org/q-q-plot-normality/. Accessed: 2024-12-08.