FACULDADE SENAI “GASPAR RICARDO JÚNIOR”

CFP 402

PROJETO FINAL – CIÊNCIA DE DADOS

ANDRÉ LUCAS FERREIRA

LEONARDO RODRIGUES VIEIRA

NICKOLAS RAPHAEL MACHADO

ODIRLEI LIMA

SOROCABA

2025

FACULDADE SENAI “GASPAR RICARDO JÚNIOR”

CFP 402

PROJETO FINAL – CIÊNCIA DE DADOS

ANDRÉ LUCAS FERREIRA

LEONARDO RODRIGUES VIEIRA

NICKOLAS RAPHAEL MACHADO

ODIRLEI LIMA

Nº2, N°26, N°33, N°34, Tecnólogo em Análise e Desenvolvimento de Sistemas, 1º semestre.

Trabalho solicitado no componente curricular de banco de dados sob orientação do Prof.º André Souza.

Sorocaba, 16 de maio de 2025.

SUMÁRIO

1 INTRODUÇÃO

1.1 Apresentação do tema escolhido

O setor hoteleiro ocupa uma posição de destaque na economia global, sendo essencial para o desenvolvimento do turismo e para a movimentação econômica de diversas regiões. A crescente competitividade no mercado da hospitalidade exige que empresas adotem estratégias baseadas em dados para otimizar seus serviços, compreender o perfil dos hóspedes e maximizar seus lucros. Neste contexto, a análise estatística aplicada à hotelaria surge como uma ferramenta indispensável, permitindo extrair informações de suma importância a partir dos dados operacionais e financeiros gerados diariamente pelos estabelecimentos.

1.2 Objetivo da análise

O presente trabalho tem por objetivo realizar uma análise estatística exploratória sobre dados simulados do setor hoteleiro, a fim de compreender o comportamento dos clientes, avaliar o desempenho financeiro e identificar padrões que possam subsidiar a tomada de decisão gerencial. Pretende-se aplicar conceitos estatísticos fundamentais, além de técnicas de visualização de dados, para oferecer esclarecimentos práticos e teóricos relevantes.

1.3 Justificativa da escolha dos dados

A escolha dos dados relativos ao setor hoteleiro justifica-se pela representatividade e aplicabilidade prática que este segmento oferece. A análise permite abordar variáveis de interesse como tempo de estadia, faixa etária dos hóspedes, tipo de quarto, avaliações dos serviços, entre outros. Tais informações são cruciais tanto para gestores do setor quanto para pesquisadores que desejam compreender melhor os fenômenos associados à hospitalidade e ao comportamento do consumidor, de forma a tornar possível tomadas de decisões fundamentadas em análises estratégicas.

2 REFERENCIAL TEÓRICO

2.1 Conceitos estatísticos e computacionais utilizados

A análise de dados estatísticos fundamenta-se em conceitos essenciais como escalas de medição, medidas de tendência central (média, mediana e moda), medidas de dispersão (desvio padrão, variância, amplitude e intervalo interquartil), além de análise de correlação, regressão linear simples e séries temporais.

- **Tendência central:** busca identificar o ponto de equilíbrio dos dados;
- **Escalas de medição:** são formas de categorizar e quantificar dados;
- **Dispersão:** mede a variabilidade dos dados em relação à média;
- **Correlação de Pearson:** avalia a intensidade e a direção da relação linear entre duas variáveis quantitativas;
- **Regressão linear:** permite modelar e prever valores de uma variável dependente com base em uma variável independente;
- **Séries temporais:** analisam a evolução dos dados ao longo do tempo, permitindo identificar tendências sazonais e padrões.

2.2 Bibliotecas Python aplicadas

Para a condução da análise, utilizaram-se diversas bibliotecas da linguagem Python, entre elas, estão:

- **Pandas:** responsável pela manipulação e análise de dados estruturados em formato de DataFrame;

- **NumPy:** fornece suporte a operações matemáticas e arrays multidimensionais;

- **Matplotlib e Seaborn:** voltadas à criação de gráficos estáticos e interativos, possibilitando uma visualização clara e eficiente dos dados;

- **Statsmodels:** utilizada para modelagem estatística, especialmente regressão linear e outros testes estatísticos;

- **SciPy:** oferece ferramentas para cálculos estatísticos, incluindo medidas de tendência, dispersão e testes de hipóteses.

3 METODOLOGIA

3.1 Descrição da base de dados

A base de dados foi gerada de maneira simulada, contendo 400 registros de reservas em um hotel. As variáveis contemplam: gênero do hóspede, faixa etária, número de dias de estadia, preço da diária, total pago, data de check-in, avaliação dos serviços prestados e tipo de quarto reservado.

3.2 Etapas de tratamento e limpeza de dados

Foram realizadas etapas essenciais de pré-processamento dos dados, incluindo:

- Conversão de variáveis de data para o formato adequado;
- Verificação e eliminação de eventuais inconsistências ou dados ausentes;
- Transformação de variáveis categóricas em formatos numéricos para possibilitar análises quantitativas;
- Criação de variáveis derivadas, como cálculo do total pago pela estadia (produto entre dias de estadia e preço da diária).

3.3 Ferramentas e processos utilizados

A análise foi desenvolvida integralmente em ambiente Python, com o uso do Jupyter Notebook como interface para desenvolvimento e execução dos scripts. As bibliotecas Pandas, NumPy, Matplotlib, Seaborn, Statsmodels e SciPy foram fundamentais para o desenvolvimento das análises estatísticas, modelagem e geração das visualizações.

4 ANÁLISE DE DADOS

4.1 Aplicação dos tópicos selecionados

Ao total, para uma análise estratégica dos dados, foram aplicadas seis abordagens estatísticas distintas:

1. **Análise Descritiva:** foram calculadas medidas de tendência central e dispersão para variáveis como “Estadia_Dias”, “Preço_Diária” e “Total_Pago”, proporcionando uma visão geral do comportamento dos dados.

2. **Escalas de Medição:** as variáveis foram classificadas segundo suas escalas (nominal, ordinal, intervalar e razão), facilitando a definição dos métodos estatísticos apropriados para cada uma.

3. **Distribuição de Frequências e Visualização:** foram elaborados histogramas e boxplots, especialmente para analisar a variável “Faixa Etária”, permitindo uma interpretação visual dos dados e identificação de possíveis outliers.

4. **Série Temporal:** a análise do “Total_Pago” ao longo do tempo, considerando a data de check-in, revelou tendências sazonais. Aplicou-se uma média móvel de sete dias para suavização das oscilações e melhor interpretação do comportamento temporal.

5. **Correlação:** através do coeficiente de correlação de Pearson, identificou-se uma forte correlação positiva entre os dias de estadia e o total pago, o que era esperado, porém necessário confirmar estatisticamente.

6. **Regressão Linear Simples:** foi ajustado um modelo de regressão linear simples, tendo “Estadia_Dias” como variável independente e “Total_Pago” como variável dependente. O modelo revelou uma relação linear forte, validando a expectativa de que o aumento na quantidade de dias implica proporcionalmente em um maior faturamento.

4.2 Visualizações e interpretações

As visualizações geradas (gráficos de dispersão, linhas, histogramas e boxplots) facilitaram a identificação de padrões relevantes, como a predominância de reservas em quartos padrões, maior concentração de hóspedes na faixa etária de 26 a 35 anos, e picos sazonais de faturamento.

4.3 Discussão dos resultados obtidos

A análise estatística conduzida sobre a base de dados simulada de hotelaria proporcionou uma compreensão aprofundada das principais características dos hóspedes e do desempenho financeiro do estabelecimento. A seguir, apresenta-se a discussão detalhada dos resultados, separando-os conforme cada abordagem estatística empregada.

4.3.1 Tendência central e Dispersão

Os cálculos das medidas de tendência central revelaram que o número médio de dias de estadia foi relativamente baixo, evidenciando um comportamento típico de hóspedes que realizam viagens de curta duração. A média do preço da diária apresentou um valor de aproximadamente R$ 250,00, com um desvio padrão considerável, indicando que os preços praticados variam de forma significativa em função de diferentes fatores como tipo de quarto, sazonalidade e demanda.

A análise do total pago por reserva demonstrou alta variabilidade, como evidenciado pela amplitude e pelo intervalo interquartil, sugerindo que o hotel atende tanto a hóspedes que optam por estadias curtas e econômicas quanto àqueles que investem em hospedagens mais longas e com serviços diferenciados.

IMAGEM 01 - Dispersão

```
Medidas de Dispersão:

--- Estadia_Dias ---
Desvio Padrão: 4.21
Variância: 17.73
Amplitude: 17.04
Q1: 9.03905 | Q3: 15.997150000000001 | IQR: 6.958100000000002

--- Preco_Diaria ---
Desvio Padrão: 50.15
Variância: 2515.07
Amplitude: 288.78
Q1: 212.48000000000002 | Q3: 281.475 | IQR: 68.995

--- Total_Pago ---
Desvio Padrão: 1400.53
Variância: 1961472.95
Amplitude: 7266.52
Q1: 2098.5475 | Q3: 4134.995 | IQR: 2036.4474999999998

--- Avaliacao ---
Desvio Padrão: 1.26
Variância: 1.58
Amplitude: 4.00
Q1: 2.0 | Q3: 4.0 | IQR: 2.0
```

4.3.2 Escalas de medição

Ao classificar as variáveis de acordo com suas respectivas escalas de medição, foi possível compreender melhor as possibilidades analíticas de cada campo de dados. Por exemplo, as variáveis "Gênero" e "Tipo de Quarto", por serem nominais, permitiram apenas análises de frequência e proporção, enquanto "Faixa Etária" e "Avaliação", como variáveis ordinais, permitiram ordenação e categorização por nível de satisfação ou idade. Já as variáveis de razão, como "Estadia_Dias", "Preço_Diária" e "Total_Pago", possibilitaram a realização de operações matemáticas complexas, como cálculo de média, variância e regressão.

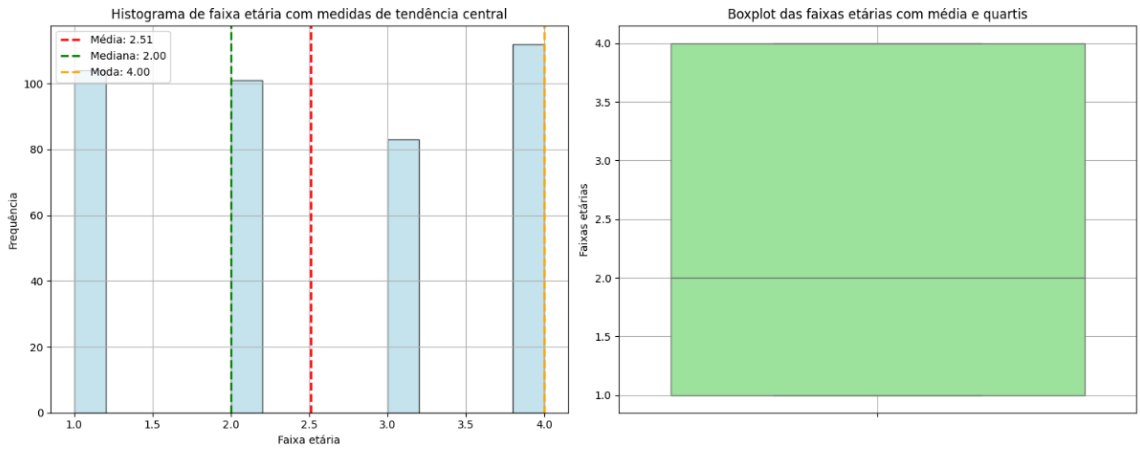
A correta identificação dessas escalas foi fundamental para selecionar os métodos estatísticos apropriados a cada variável, garantindo a validade dos resultados.

4.3.3 Distribuição de frequências e visualização (Faixa etária)

A distribuição das faixas etárias, representada por histogramas e boxplots, indicou que a maioria dos hóspedes se concentra entre 26 e 35 anos, seguidos por aqueles com idade entre 36 e 50 anos. Essa concentração etária pode ser associada ao perfil de viajantes economicamente ativos, geralmente com maior propensão a viagens de lazer e negócios.

Além disso, o boxplot permitiu identificar uma distribuição relativamente assimétrica, com uma leve concentração em faixas etárias mais jovens. A dispersão entre os quartis reforçou a importância de estratégias de marketing segmentadas, focadas no público predominante.

IMAGEM 02 - Tendência central

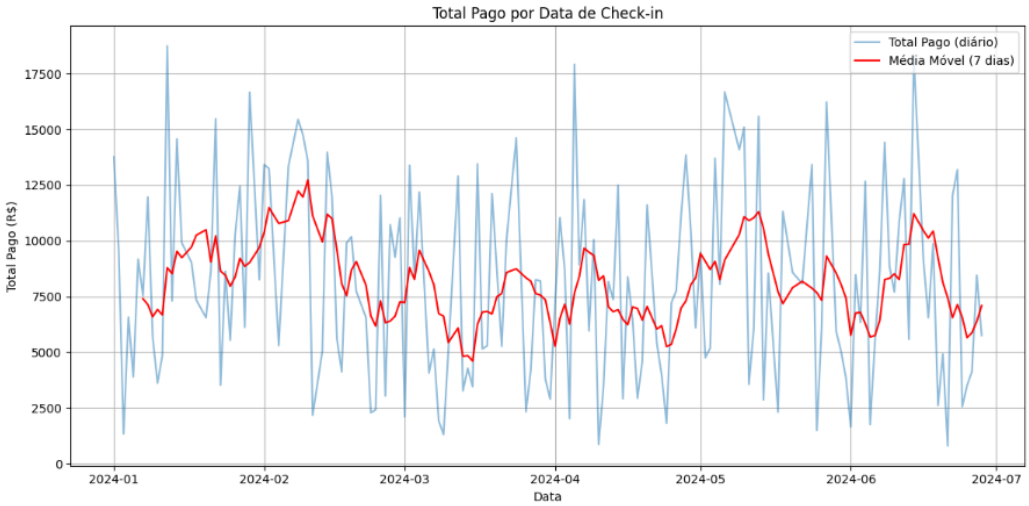


4.3.4 Análise de séries temporais (Total pago por data de check-in)

A análise temporal do “Total Pago”, ao longo de aproximadamente seis meses, revelou flutuações consideráveis no faturamento diário do hotel. A aplicação da média móvel de sete dias foi crucial para suavizar variações pontuais, evidenciando tendências sazonais.

Foi possível observar picos de receita em determinados períodos, sugerindo a influência de eventos sazonais ou promoções pontuais. Mesmo sendo uma base simulada, o comportamento identificado é representativo da realidade de muitos empreendimentos hoteleiros, onde a demanda é afetada por períodos de alta temporada, feriados prolongados e outros fatores externos.

IMAGEM 03 - Séries temporais

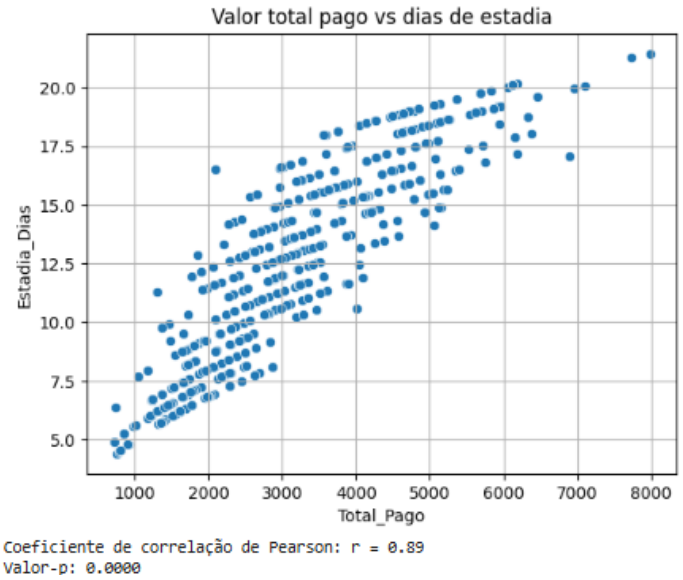


4.3.5 Correlação de Pearson (Estadia vs Total Pago)

O coeficiente de correlação de Pearson entre “Estadia_Dias” e “Total_Pago” foi elevado, demonstrando uma forte correlação positiva. Este resultado confirma uma relação direta entre o tempo de permanência do hóspede e o valor total pago, uma relação esperada no setor, mas que merece atenção no momento de elaborar estratégias de precificação e pacotes promocionais.

É importante destacar que a significância estatística foi confirmada pelo valor-p muito inferior ao nível usual de significância (α = 0,05), reforçando a robustez desta conclusão.

IMAGEM 04 - Correlação de Pearson

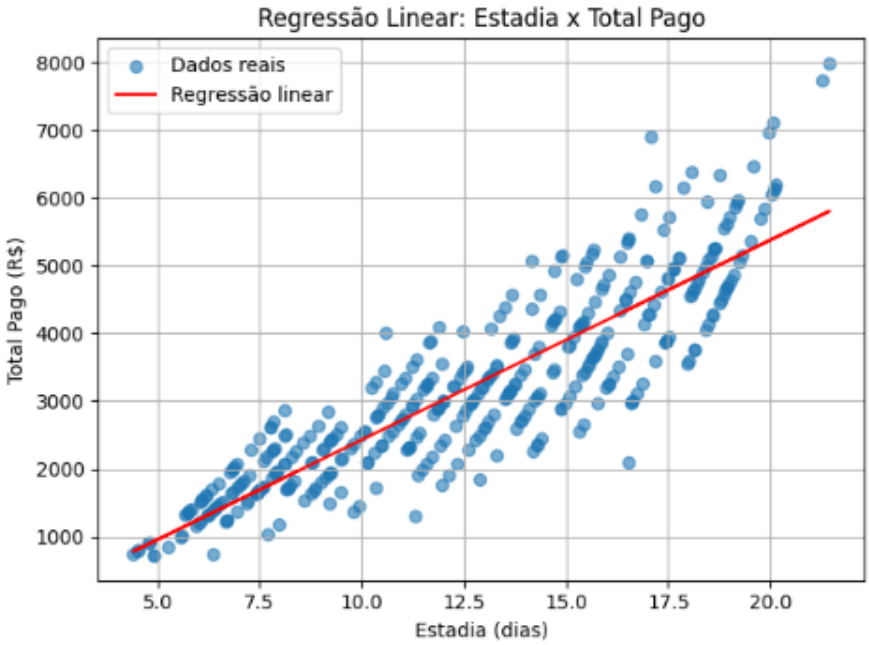


4.3.6 Regressão linear simples (Previsão do total pago com base na estadia)

A aplicação do modelo de regressão linear simples proporcionou um ajuste estatisticamente significativo entre o número de dias de estadia e o total pago. A reta de regressão apresentou inclinação positiva, evidenciando que, a cada dia adicional de hospedagem, há um incremento previsível no faturamento gerado pelo hóspede.

O gráfico de dispersão com a linha de regressão evidenciou que, apesar de alguns pontos de dispersão fora da tendência central (outliers), o modelo explica de forma satisfatória a variabilidade dos dados. Este tipo de análise pode servir de base para modelagens preditivas futuras, visando à previsão de receita com base nas reservas.

IMAGEM 05 - Regressão linear



5 CONCLUSÃO

5.1 Resumo dos principais achados

A análise estatística da base de dados simulada de hotelaria permitiu compreender padrões relevantes sobre o perfil dos hóspedes e o desempenho financeiro. Identificou-se que a maioria das estadias é de curta a média duração, refletindo o perfil de viajantes a negócios ou turismo de fim de semana, o que sugere a necessidade de estratégias focadas nesse público.

Observou-se também uma variação significativa nos preços das diárias, indicando influência de fatores como tipo de quarto e sazonalidade. As faixas etárias predominantes, de 26 a 50 anos, destacam um público economicamente ativo, orientando ações de marketing e personalização dos serviços.

A análise temporal revelou oscilações no faturamento, típicas de sazonalidade, enquanto a forte correlação entre dias de estadia e total pago confirmou o impacto direto do tempo de hospedagem na receita. A regressão linear reforçou essa relação, oferecendo um modelo simples e eficaz para previsão de faturamento.

De forma geral, os achados reforçam a importância da análise de dados para a tomada de decisão no setor hoteleiro, permitindo otimizar preços, ofertas e a gestão de demanda.

5.2 Limitações encontradas

Por se tratar de uma base de dados simulada, algumas limitações estão presentes, como a ausência de variáveis externas (feriados, eventos, sazonalidade real do mercado) que poderiam impactar diretamente nas reservas e no faturamento. Além disso, não foram considerados cancelamentos, upgrades, ou promoções específicas.

5.3 Sugestões para análises futuras

Para estudos futuros, sugere-se a inclusão de dados reais do setor, análise preditiva com modelos de machine learning, segmentação de clientes por meio de algoritmos de clusterização e análise de sentimento aplicada às avaliações dos hóspedes. Estas abordagens poderiam fornecer insights ainda mais robustos e aplicáveis ao contexto empresarial da hotelaria.

6 REFERÊNCIAS

- **Livro:** BARBETTA, Pedro Alberto; REIS, Marcelo Menezes; BORNIA, Antonio Cezar. *Estatística: para cursos de engenharia e informática*. 2. ed. São Paulo: Atlas, 2008.
- **Livro:** BONAFINI, Fernanda César (org.). *Estatística*. 2. ed. São Paulo: Pearson, 2019.
- **Artigo complementar:** "Tipos de variáveis e escalas de medição em estatística" – disponível em sites acadêmicos ou blogs de Data Science.
- **Site:** SOUZA, André. Material - Ciência de dados. Github. https://github.com/profAndreSouza/Material/tree/main/Ciência%20de%20Dados. Acesso em 10/06/2025.
- Triola, M. F. (2011). *Introdução à Estatística*. Pearson.
- **Site:** FERREIRA, André Lucas; RODRIGUES, Leonardo; MACHADO, Nickolas Raphael; LIMA, Odirlei. Projeto final - Ciência de dados. Github. https://github.com/AndreLucas0/projeto-final-cd. Acesso em 18/06/2025.