PYTHON - NÍVEL INICIANTE

POWERED BY MULHERES EM DADOS

PROJETO SATISFAÇÃO DO CONSUMIDOR

AULA 02 - ANÁLISE EXPLORATÓRIA DE DADOS (EDA)

22 DE JUNHO DE 2022

CRONOGRAMA

Acesse **aqui** o notebook da aula 02

JUNHO-JULHO 2022

DOM	SEG	TER	QUA	QUI	SEX	SAB
12	13	14	15 **Entendimento**	16	17	18
19	20	21	22 **Análise Exploratoria**	23	24	25
26	27	28	29 **Criação de modelo**	30	1	2
3	4	5	6 **Avaliação**	7	8	9

Entendimento, carregamento e pré-processamento dos dados

Entendimento do negócio e configurar o colab para pre-processar os dados

Análise exploratória de dados (EDA)

Analisaremos os dados resultando nos principais insights sobre o negócio

Feature Engineering e criação de modelo de ML

Criaremos o nosso modelo de machine learning a partir das variáveis que mais fazem sentido para o negócio

Avaliação e melhorias no modelo de ML

Entenderemos melhor sobre como avaliar e melhorar a performance do nosso modelo de machine learning

SUMÁRIO

ANÁLISE EXPLORATÓRIA DE DADOS (EDA)

- Importante etapa de análise e investigação dos conjuntos de dados, pois resume suas principais características, muitas vezes usando métodos de visualização de dados.

- Proporciona uma melhor compreensão das variáveis do conjunto de dados e as relações entre eles.

GRÁFICOS - MELHORES PRÁTICAS

- Utilizar gráficos que condizem com quantidade e finalidade daquela observação. Recomendamos checar o tópico 'Para saber mais', que traz indicações de guias de gráficos.
- Pensar nas cores adotadas, cartela que condiza com o que quer passar e na acessibilidade.
- Evitar visualizações de difícil entendimento. Lembre-se do *storytelling* e qual grupo será o grupo receptor.
- Quando possível, utilizar o conhecimento em funções para facilitar na plotagem dos gráficos.

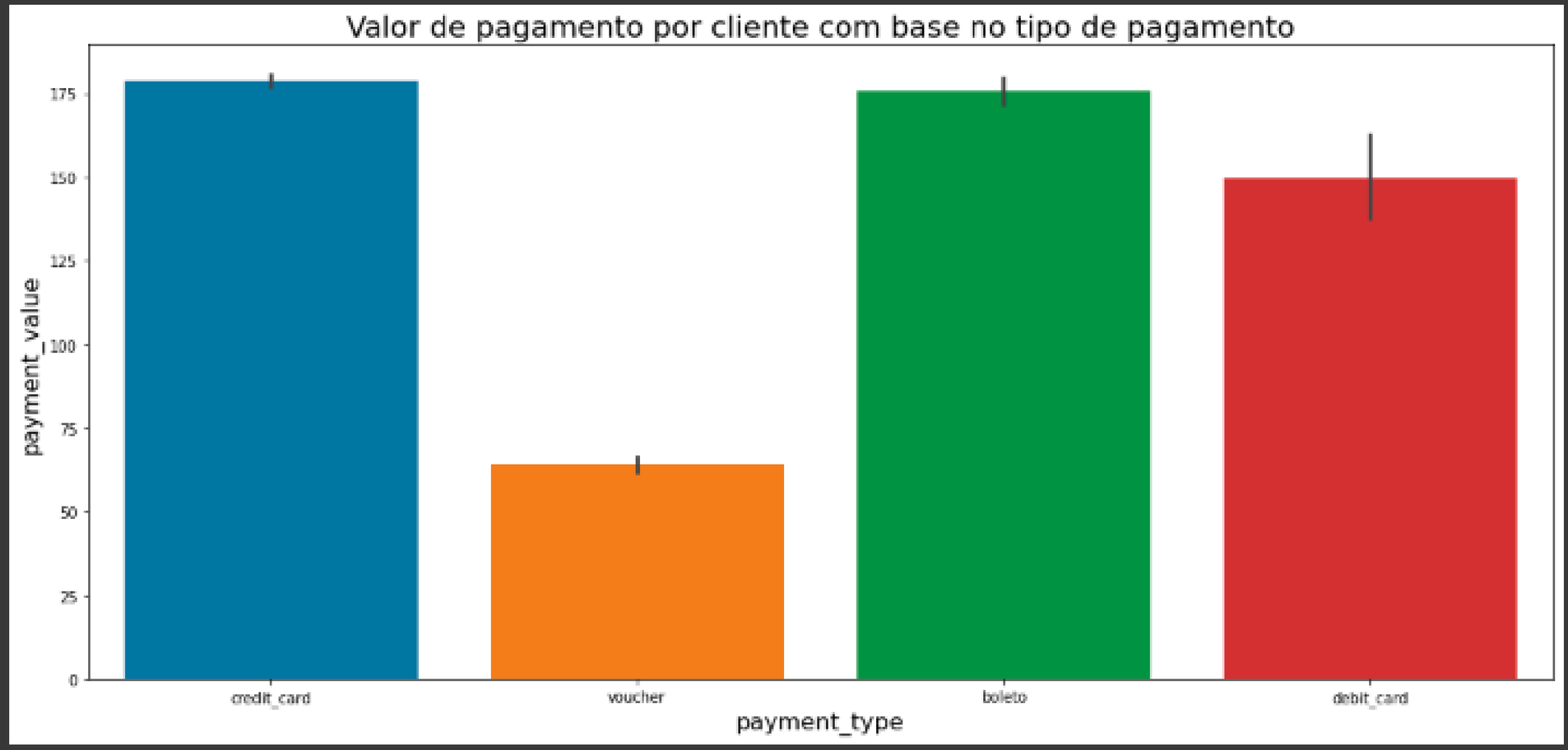
GRÁFICO DE BARRAS (BAR PLOT)

Usado para demonstrar a proporção de vezes que uma variável assume um dado valor.

Um gráfico de barras representa a frequência de uma variável numérica e, na biblioteca Seaborn, pode fornecer também a indicação de incerteza sobre a distribuição de valores por meio de barras de erro (que são os tracinhos em cinza que vemos no exemplo a seguir).

```
[ ]  7  # definição dos nomes dos eixos e titulo
     8  plt.xlabel(x_var, fontsize = font_size)
     9  plt.ylabel(y_var, fontsize = font_size)
    10  plt.title(title, fontsize = title_font_size)
    11
    12  plt.show()
```

```
1 bar_plot_df('payment_type', 'payment_value', 'Valor de pagamento por cliente com base no tipo de pagamento')
```



Exemplo de barplot com indicação de barras de erro

GRÁFICO DE DISPERSÃO (SCATTER PLOT)

- São representações gráficas do relacionamento entre duas variáveis numéricas.
- O scatter plot utiliza pontos de dispersão para demonstrar essa relação, cada ponto representa o valor de uma variável no eixo horizontal e o valor de outra variável no eixo vertical. O que acontece com o exemplo abaixo:

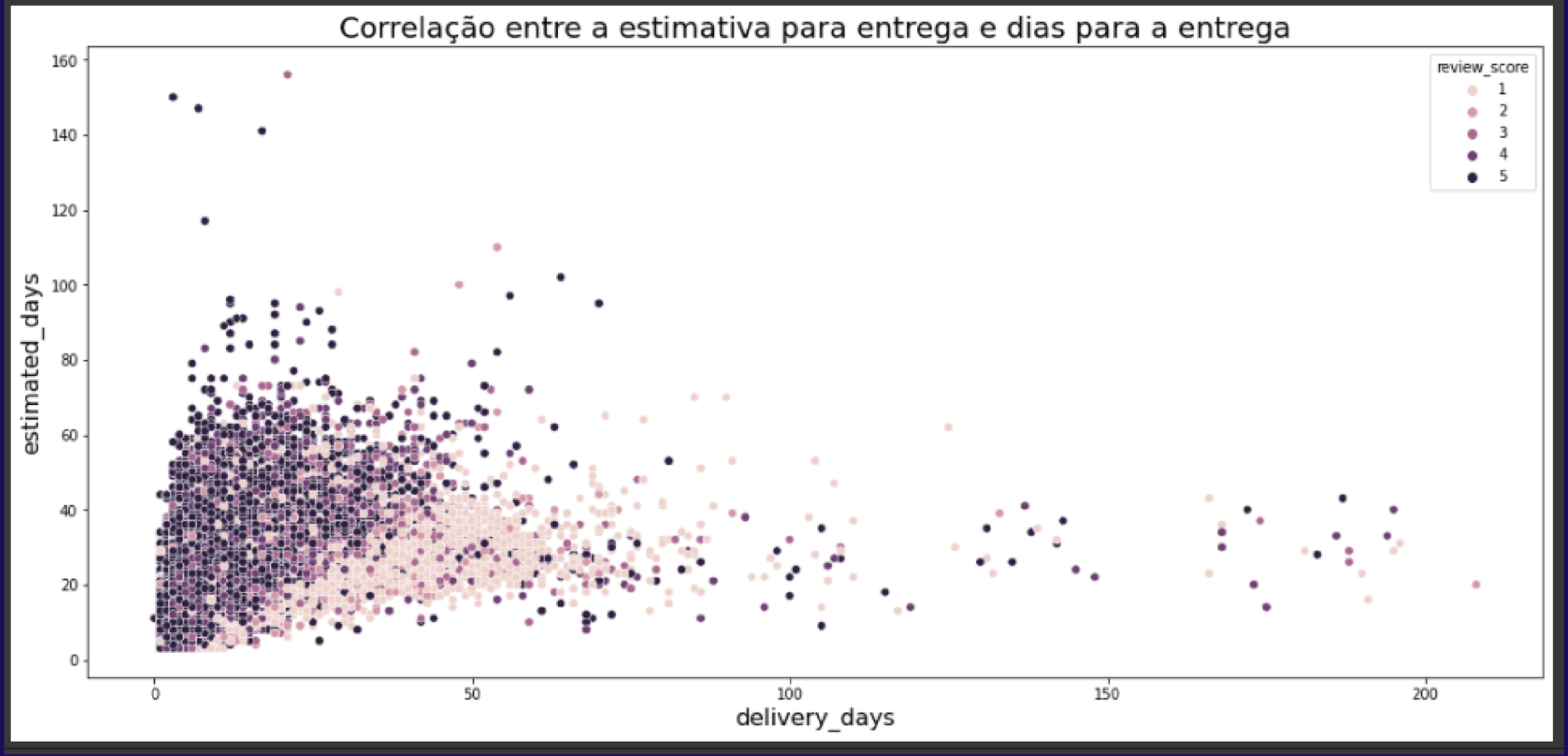
```
def scatter_plot_df(x_var, y_var, title):
  # definição do gráfico de barras
  fig = plt.figure(figsize = fig_size)

  sns.scatterplot(x = x_var, hue='review_score',
                  y = y_var, data=df_ecommerce)

  plt.title(title, fontsize = title_font_size)
  plt.xlabel(x_var, fontsize = font_size)
  plt.ylabel(y_var, fontsize = font_size)
  plt.show()
```

Exemplo de gráfico utilizando scatter plot

```
scatter_plot_df('delivery_days', 'estimated_days', 'Correlação entre a estimativa para entrega e dias para a entrega')
```



Exemplo de gráfico utilizando scatter plot

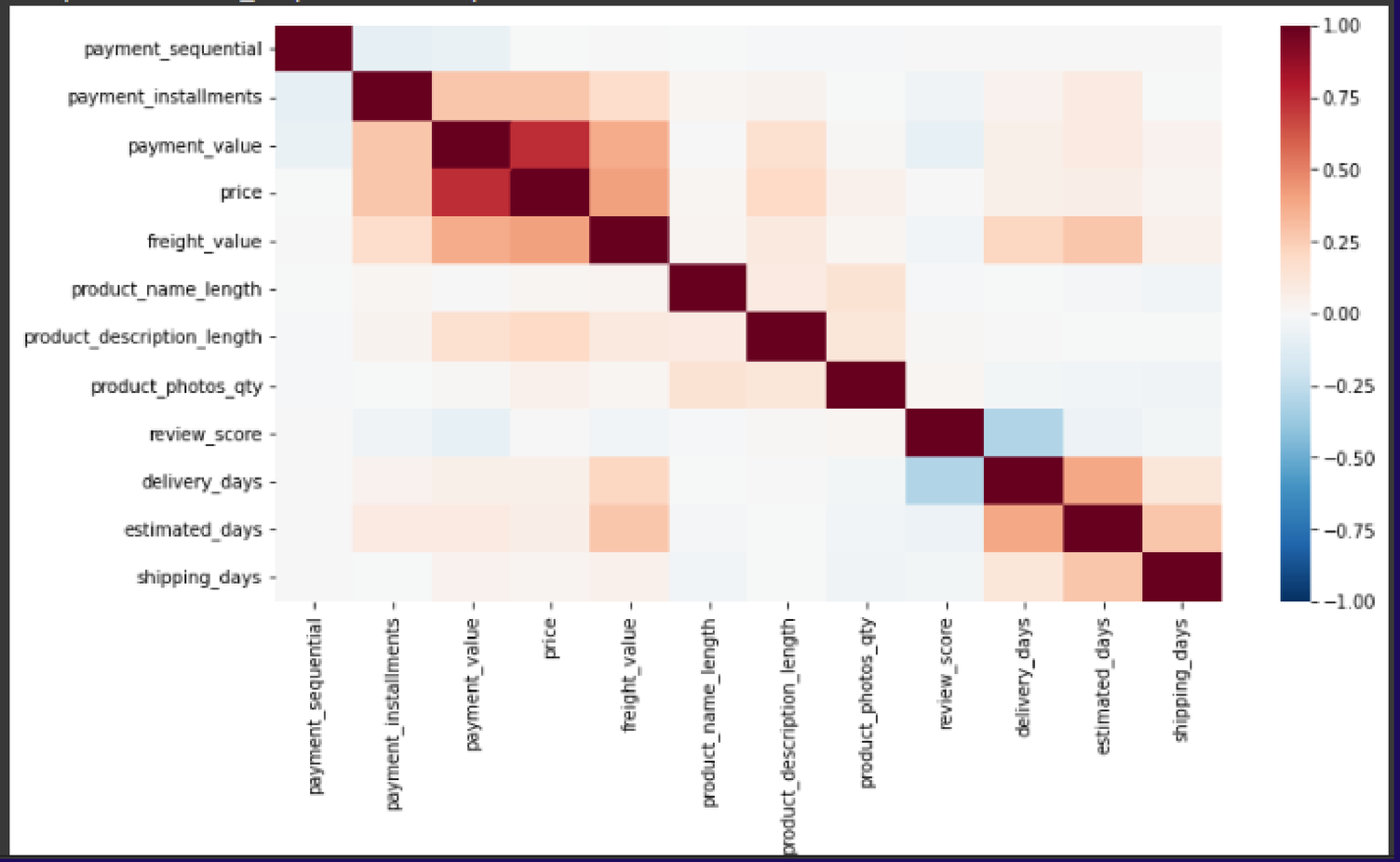
GRÁFICO DE CALOR (HEATMAP)

- Gráfido de calor é definido como uma representação gráfica dos dados usando cores para visualizar o valor da matriz.

- Para representar valores mais comuns ou atividades mais altas, são usadas cores mais profundas e, para representar valores menos comuns , tons mais claros são selecionados.

```
plt.figure(figsize=(12,6))
sns.heatmap(df_ecommerce.corr(), cmap='RdBu_r', vmin=-1, vmax=1)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f61b679e050>
```



Exemplo de gráfico utilizando heatmap

CORRELAÇÃO

- Avalia a relação linear entre duas variáveis, o quanto uma variável tende a influenciar outra. Aplicação prática: qual é o valor esperado de y para um determinado valor de x?
- Na aula 02, usamos o **coeficiente de correlação de Pearson (p com intervalo de -1 a +1):**

- **correlação positiva** (valores próximos a +1) mostra que quando uma variável x aumenta seu valor, a outra (y) também aumenta;
- **correlação negativa** mostra que, enquanto uma variável x aumenta seu valor, a outra y sempre diminui (valores próximos a -1);
- quando uma variável não possui correlação uma com a outra, o valor é 0;
  - **Importante**: correlação não implica em causalidade (que uma variável causou a outra)

REFERÊNCIAS

(CLIQUE NOS LINKS)

Análise de Dados @ FLAI

Correlação e causalidade @ Canal Professora Fernanda Maciel

Estatística Descritiva Bivariada 2 - Relacionamento entre Duas Quantitativas @ Canal R, Estatística e Aprendizado de Máquina

GRUS, Joel. Visualizando Dados. In.: **Data Science do Zero: noções fundamentais com Python**. 2ª edição. Rio de Janeiro: Alta Books, 2021.

PARA SABER MAIS:

(CLIQUE NOS LINKS)

Choose an effective visual with the SWD Chart Guide @ Blog do Storytelling with Data

GRUS, JOEL. Estatística. In.: **Data Science do Zero: noções fundamentais com Python**. 2ª edição. Rio de Janeiro: Alta Books, 2021.

KNAFLIC, Cole Nussbaumer. **Storytelling com dados: Um guia sobre visualização de dados para profissionais de negócios**. 2ª edição. Rio de Janeiro: Alta Books, 2019.

SALSBURG, David. As Distribuições Assimétricas. In: **Uma Senhora Toma Chá... como a estatística revolucionou a ciência no século XX**. Rio de Janeiro: Zahar, 2009.

Trilha de Estudos para Analista de Dados @ GitHub das Mulheres em Dados

Trilha de Estudos para Cientista de Dados @ GitHub das Mulheres em Dados

OBRIGADA PELA PARTICIPAÇÃO!

POWERED BY MULHERES EM DADOS

- Próxima aula em 29/06, às 19h, no Discord
- Dúvidas e sugestões no canal #python

Equipe Python:
Andressa Apio, Crislane Maria, Érika Santos e Joice Oliveira