Nome: Denise do Rocio Maciel
Projeto: Visualização de dados com python

VISUALIZAÇÃO DE DADOS

Essa seção é dividida em três tópicos. No primeiro tópico é apresentado o resultado final do painel projetado, Figura 1, assim como a justificativa para as escolhas de design. No tópico 2 é apresentada codificação referente à análise e pré-processamento dos dados. Finalizando, tem-se o tópico 3, que apresenta a codificação, em python, utilizada para a plotagem de cada representação.

IMPORTANTE: Os dados utilizados para gerar a visualização são os constantes na tabela t_mov.csv. Composta por: id (inteiro), title (string), genre (string), director (string), actors (string), year (int), runtime (int), rating (float), votes (int), revenueMillions (float), metascore (float).

A. Resultado final do painel projetado e justificativa para escolhas de design

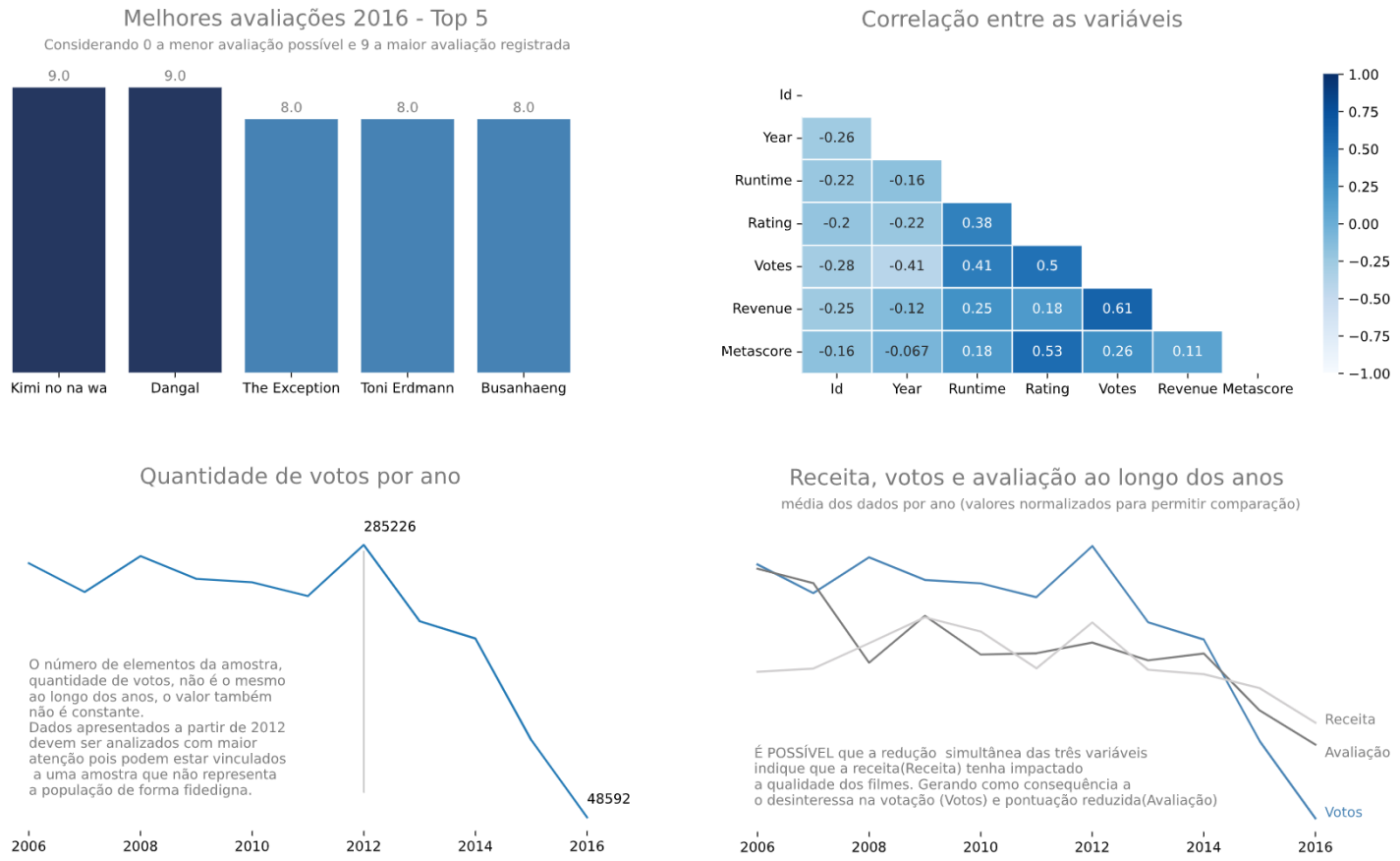


Figura 1 - Resultado final da atividade

Optou-se por utilizar como base as heurísticas de usabilidade de Nielsen [4], leis da Gestalt [3] , normas da engenharia semiótica [3] e orientações provenientes da literatura sobre visualização de dados [1]. Adicionalmente, foram consideradas recomendações de estatística descritiva para visualização de dados [5,6,7]. Como exemplos da aplicação dessas normas, pode-se citar:

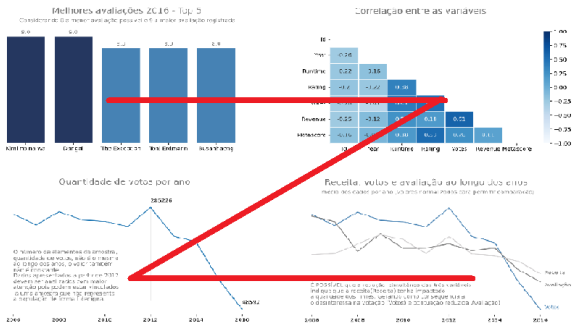
- Foi utilizado design minimalista voltado a apresentar apenas informações elementares;

- Optou-se pelas cores azul, cinza e branco por conta de usuários daltônicos. Caso fosse utilizada a paleta verde, cinza e branco, esse público poderia equivocadamente interpretar o contraste das cores como verde e vermelho (tons que geram problema para esse público). Na cultura brasileira, comumente associa-se o verde e o vermelho com o símbolo semiótico de bom /ruim ou permitido/proibido, respectivamente. Como essa não é a informação que se objetiva transmitir, optou-se pelas cores azul e branco e cinza;

- As fontes sem serifa foram escolhidas porque considerou-se que os dados seriam apresentados em ambiente digital. Caso a análise fosse realizada de forma impressa seria utilizada a versão com serifa, própria para essa finalidade;

- A apresentação dos rótulos de texto foi alinhada à direita em respeito ao fato de que tanto a leitura quanto a escrita, em cadernos, por exemplo, é realizada da margem direita em direção à margem esquerda;

- A apresentação dos gráficos seguiu o modelo de leitura em “ordem Z”, sugerido por [1]. Ou seja, espera-se que a leitura dos dados siga a seguinte ordem:



- A escolha dos dados a serem representados nos gráficos ocorreu porque através da coluna “Votes” nota-se que as amostras são de tamanho diferente. Nesse cenário, algumas representações poderiam induzir o usuário a inferências indevidas provenientes de amostragem não significativa. A seguir são apresentadas informações que justificam a escolha de cada representação::

Gráfico 1 Melhores avaliações 2016 - Top 5	Optou-se por analisar apenas os dados de um ano para evitar comparações com dados de amostras diferentes. Importante: o gráfico de colunas não é o mais adequado para dados de natureza contínua. Ainda assim, foi utilizado porque apesar de a coluna estar formatada para dados contínuos, havia apenas dados discretos. Além disso, o gráfico de colunas é bastante conhecido, o que facilita a interpretação por parte dos usuários.
Gráfico 2 Correlação entre as variáveis	A análise de frequências poderia ser comprometida pelo tamanho da amostra. Logo, optou-se por, através da correlação, verificar o quanto o crescimento da amostra impacta nas demais variáveis. Foi verificada, e posteriormente explorada, a relação entre as colunas Votes, Rating e Revenue.
Gráfico 3 Quantidade de votos por ano	O gráfico foi projetado para visualização temporal do tamanho da amostra. Importante: Nesse gráfico, inicialmente foi projetada a apresentação da linha média. No entanto, o valor do coeficiente de variação ((desvio padrão/ média)*100) para a coluna foi superior a 100. Esse dado indica que comparações com o valor médio não trazem informações significativas devido a alta dispersão dos dados.
Gráfico 4 Receita, votos e avaliação ao longo dos anos	Comparou-se os resultados das correlações mais fortes obtidas através do Gráfico 2. Como as unidades de medida são diferentes entre as variáveis Votes, Rating e Revenue, os dados foram normalizados. A cor azul foi utilizada para os Votos porque é a variável com a qual se objetiva prioritariamente realizar a comparação. As demais variáveis foram coloridas em tons cinza. Importante:a análise foi iniciada investigando o impacto da quantidade de votos. No entanto, através do gráfico 4 pode-se levantar a hipótese de que a redução das variáveis Votes e Rating estão relacionadas a redução da variável Rating.

B.Plotagem das representações gráficas

Representação gráfica 1

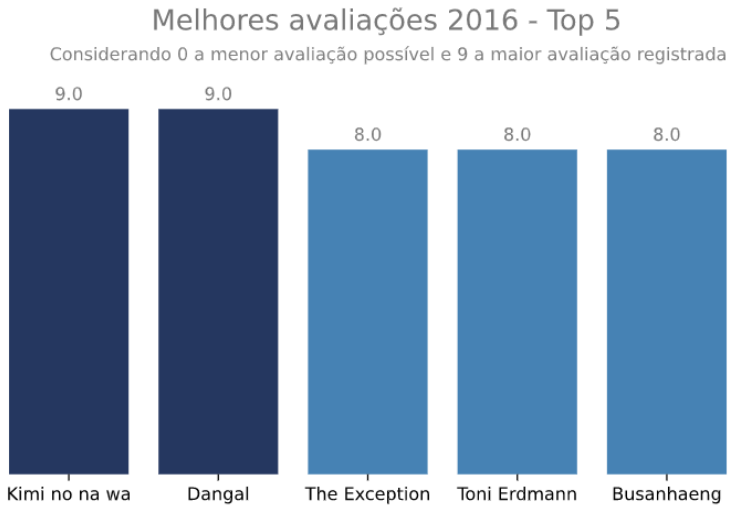


Figura 2 - Representação gráfica 1

```
#gráfico 1
#dados a serem utilizados
df_movie2016=df_imdbMovies.loc[df_imdbMovies['Year'] ==2016,
['Year','Title','Rating']].sort_values('Rating',ascending=False).head(5)
x_dMovie = df_movie2016['Title']
dados = df_movie2016['Rating']

#plotagem
cores = ['#253760','#253760','#4682B4','#4682B4','#4682B4']
fig, ax = plt.subplots(figsize=(8, 4))
fig.text(x = 0.3, y = 1, s= 'Melhores avaliações 2016 - Top 5', fontsize=16, color = '#787878')
fig.text(x = 0.2, y = 0.94, s= 'Considerando 0 a menor avaliação possível e 9 a maior avaliação
registrada', fontsize=10, color = '#808080')

barras = ax.bar(x=x_dMovie, height=dados, color = cores )

#funcao posiciona o label acima da respectiva barra
def set_label_y(bar):
    for b in bar:
        alturaBarra = b.get_height()
        ax.annotate(
                    (alturaBarra),
                    xy=(b.get_x() + b.get_width()/2,  alturaBarra),
                    xytext=(0, +3),
                    textcoords="offset points",
                    ha='center',
                    va='bottom',
                    fontsize=10,
                    color='#787878',
                    )
set_label_y(barras)

#Remocao bordas e eixos
ax.spines['right'].set_visible(False)
ax.spines['left'].set_visible(False)
ax.spines['top'].set_visible(False)
ax.spines['bottom'].set_visible(False)
plt.gca().axes.get_yaxis().set_visible(False)
plt.show()
```

Representação gráfica 2

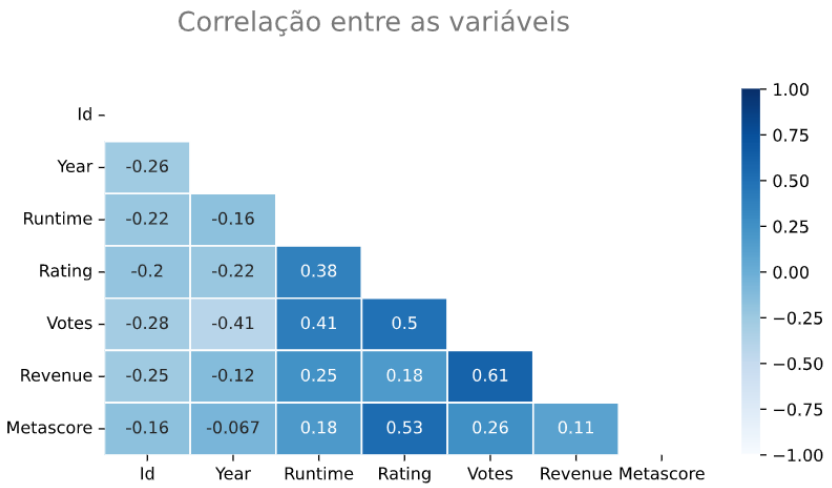


Figura 3 - Representação gráfica 2

```
#grafico 2
#dados a serem utilizados
mask = np.triu(df_imdbMovies.corr())

#plotagem
fig, ax = plt.subplots(figsize=(8, 4))
fig.text(x = 0.2, y = 1, s= 'Correlação entre as variáveis', fontsize=16, color = '#787878')
sns.heatmap(df_imdbMovies.corr(), annot=True, vmin=-1, vmax=+1, cmap='Blues',mask=mask,linewidths=1)

# a máscara foi utilizada como 'triu' (triângulo inferior) em vez de tril (triângulo superior) em
consideração ao fato de a escrita/leitura ocorre da esquerda para a direita.
```

Representação gráfica 3

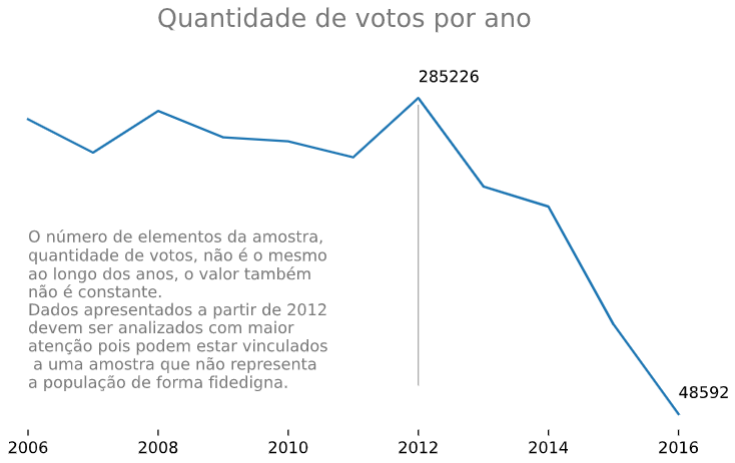


Figura 4 - Representação gráfica 3

```
#gráfico 3
#dados a serem utilizados
x_dYear = [2006,2007,2008,2009,2010,2011,2012,2013,2014,2015,2016]
y_dVotes=df_imdbMovies.groupby(['Year'])['Votes'].mean()

#plotagem
fig, ax = plt.subplots(figsize=(8, 4))
fig.text(x = 0.3, y = 1, s= 'Quantidade de votos por ano', fontsize=16, color = '#787878')
plt.plot(x_dYear,y_dVotes)


t = ("O número de elementos da amostra,\n"
     "quantidade de votos, não é o mesmo\n"
     "ao longo dos anos, o valor também\n"
     "não é constante.\n"
     "Dados apresentados a partir de 2012\n"
     "devem ser analizados com maior\n"
     "atenção pois podem estar vinculados\n "
     "a uma amostra que não representa\n"
     "a população de forma fidedigna.")
plt.text(2006,70000,t,wrap=True, color="#787878")
plt.vlines(2012, 280000, 70000, color ="#787878", linewidth=0.5)

for x,y in zip(x_dYear,y_dVotes):
    if (x== 2012 or x== 2016):
       label = "{:.0f}".format(y)

       plt.annotate(label, # this is the text
                 (x,y), # these are the coordinates to position the label
                 textcoords="offset points", # how to position the text
                 xytext=(0,10), # distance from text to points (x,y)
                 ha='left') # horizontal alignment can be left, right or center


#Remocao bordas e eixos
ax.spines['right'].set_visible(False)
ax.spines['left'].set_visible(False)
ax.spines['top'].set_visible(False)
ax.spines['bottom'].set_visible(False)
plt.gca().axes.get_yaxis().set_visible(False)
plt.show()
```

Representação gráfica 4

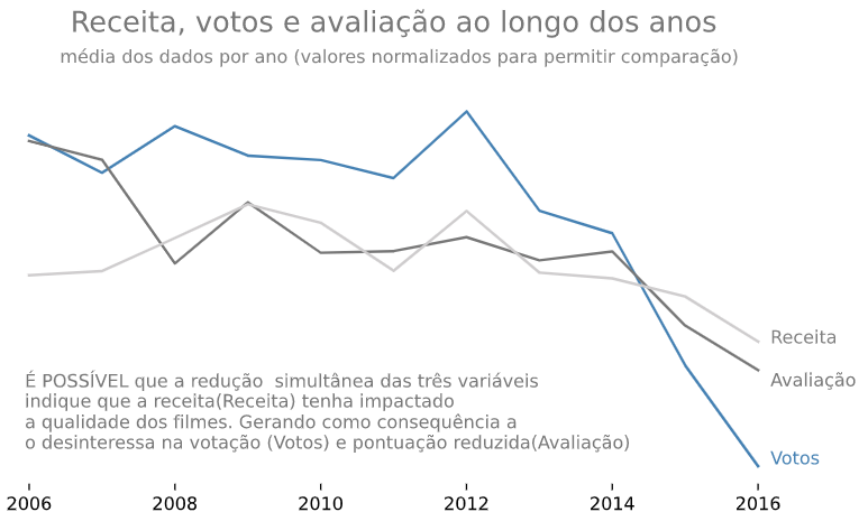


Figura 5 - resultado da representação gráfica 4

```
#gráfico 4
#dados a serem utilizados
# colunas 'Votes', 'Rating' e 'Revenue' estão em diferentes unidades de mensuração. Para que as mesmas
sejam comparadas em um mesmo gráfico será realizada normalização das mesmas
scaler = StandardScaler()
df_imdbMovies['VotesNorm'] = scaler.fit_transform(np.asarray(df_imdbMovies.loc[:, 'Votes']).reshape(-1,1))
df_imdbMovies['RatingNorm'] = scaler.fit_transform(np.asarray(df_imdbMovies.loc[:,
'Rating']).reshape(-1,1))
df_imdbMovies['RevenueNorm'] = scaler.fit_transform(np.asarray(df_imdbMovies.loc[:,
'Revenue']).reshape(-1,1))

x_dYear = [2006,2007,2008,2009,2010,2011,2012,2013,2014,2015,2016]
y_dVotes=df_imdbMovies.groupby(['Year'])['VotesNorm'].mean()
y_dRating=df_imdbMovies.groupby(['Year'])['RatingNorm'].mean()
y_dRevenue=df_imdbMovies.groupby(['Year'])['RevenueNorm'].mean()
# apesar de o teorema do limite central afirmar que a média das amostras tende a acertar
#o estimador da população, a média não foi informada no gráfico porque o coeficiente de
#variação retornou valor elevado (inclusive acima de 100).
#Segundo a literatura estatística, isso implica que há significativa variação entre os dados,
#logo, não é adequado utilizar a média como parâmetro para a tomada de decisão

#plotagem
fig, ax = plt.subplots(figsize=(8, 4))
fig.text(x = 0.2, y = 1, s= 'Receita, votos e avaliação ao longo dos anos', fontsize=16, color =
'#787878')
fig.text(x = 0.19, y = 0.94, s= 'média dos dados por ano (valores normalizados para permitir comparação)',
fontsize=10, color = '#808080')

plt.plot(x_dYear,y_dVotes, color='#4682B4')
plt.text(0.97, 0.05, 'Votos', transform=ax.transAxes, color = '#4682B4')
plt.plot(x_dYear,y_dRating, color ="#787878")
plt.text(0.97, 0.25, 'Avaliação', transform=ax.transAxes, color = "#787878")
plt.plot(x_dYear,y_dRevenue, color = "#CFCDCE")
plt.text(0.97, 0.36, 'Receita', transform=ax.transAxes, color = "#808080")

t = ("É POSSÍVEL que a redução  simultânea das três variáveis\n"
     "indique que a receita(Receita) tenha impactado \n"
     "a qualidade dos filmes. Gerando como consequência a  \n"
     "o desinteressa na votação (Votos) e pontuação reduzida(Avaliação)")
plt.text(0.04, 0.09, t, transform=ax.transAxes, color="#787878")

#Remocao bordas e eixos
ax.spines['right'].set_visible(False)
ax.spines['left'].set_visible(False)
ax.spines['top'].set_visible(False)
ax.spines['bottom'].set_visible(False)
plt.gca().axes.get_yaxis().set_visible(False)
plt.show()
```

REFERÊNCIAS

[1] KNAFLIC, Cole Nussbaumer. **Storytelling com Dados: Um guia sobre visualização de dados para profissionais de negócios**. Alta Books, 2019.

[2] KNAFLIC, Cole Nussbaumer. **Storytelling with data: lets practice!**. Wiley, 2019.

[3] BARBOSA, Simone; SILVA, Bruno. **Interação humano-computador**. Elsevier Bras, 2019.

[4] NORMAN, Nielsen. 10 Usability Heuristics for User Interface Design. Disponível em:<https://www.nngroup.com/articles/ten-usability-heuristics/>. Acesso em:<dezembro de 2021>

[5] LARSON, Ron. Estatística Aplicada. Editora Pearson, 2015.

[6] HUFF, Darrell. Como mentir com estatística. Editora Intrinseca, 2016.

[7] USP e-Aulas - Heitor Marques Honório. Disponível em: <https://eaulas.usp.br/portal/VMSResources/search.action?professor=eitor+Marques+Hon%C3%B3rio#:~:text=Atualmente%20%C3%A9%20respons%C3%A1vel%20pela%20Disciplina,HRAC%2FCentrinho)%20da%20USP.>. Acesso em:<dezembro de 2021>