Análise Exploratória do Projeto High School and Beyond (HS&B)

1. Introdução

O presente relatório apresenta uma análise exploratória do dataset **hsb2f.csv**, composto por dados coletados por meio de uma pesquisa de base com alunos do último e segundo ano do ensino médio nos Estados Unidos. A primeira survey nacional foi aplicada em 1980. De acordo com o Centro Nacional para Estudos da Educação NCES(s/d), o modelo amostral inicial previa 1.100 escolas com 36 alunos do último ano e 36 do segundo ano por escola. Os instrumentos de pesquisa incluíram: questionário do segundo ano, questionário do último ano, páginas de identificação do aluno, séries de testes cognitivos para cada coorte, questionário escolar, lista de verificação de comentários do professor e questionário dos pais. A pesquisa foi realizada em sequência de 1983 a 1986 e, após um intervalo, em 1992, 1993 e 2015. De acordo com Tatsuoka (1988), até 1986, foram coletados dados de 58.270 estudantes do ensino médio (28.240 veteranos e 30.030 alunos do segundo ano) em 1.015 escolas secundárias. O High School and Beyond Project (em português Projeto Ensino Médio e Além) foi um **estudo longitudinal** dos estudantes do ensino médio e também após o témino de sua formação realizado pelo National Center for Education Statistics (United States Department of Education, 2006). Um *estudo longitudinal* é um tipo de estudo observacional que coleta dados de forma consistente de um mesmo grupo de pessoas ao longo de um período de tempo prolongado. O objetivo é monitorar mudanças de opinião e experiências, identificar problemas, descobrir estratégias para melhorar uma área específica e medir o impacto dessas estratégias (QESTIONPRO, 2024). O dataframe utilizado neste trabalho, denominado **hsb2f.csv** (OPENINTRO, s/d) é uma amostra contendo 200 observações do estudo original, aleatoriamente selecionadas, de características desconhecidas, dos alunos do último ano do ensino médio, originalmente selecionadas das 600 observações utilizadas em Tatsuoka (1988).

2. Carregando as bibiotecas

In [21]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

3. Importando os dados

In [22]:
```
data = pd.read_csv("../dados/hsb2f.csv", sep=";")
df = data.loc[:, data.columns != 'id']
```

3.1 Visualizando o dataframe

In [23]:
```
sns.set(style="whitegrid")
```

In [24]:
```
df.head(10) # visualizando as 10 primeiras linhas do dataframe
```

Out[24]:

	genero	raca	clasocial	tipescola	programa	ler	escrever	matematica	cienc
0	masculino	branca	baixa	pública	básico	57	52	41	
1	feminino	branca	média	pública	técnico	68	59	53	
2	masculino	branca	alta	pública	básico	44	33	54	
3	masculino	branca	alta	pública	técnico	63	44	47	
4	masculino	branca	média	pública	acadêmico	47	52	57	
5	masculino	branca	média	pública	acadêmico	44	52	51	
6	masculino	afro-americana	média	pública	básico	50	59	42	
7	masculino	hispânica	média	pública	acadêmico	34	46	45	
8	masculino	branca	média	pública	básico	63	57	54	
9	masculino	afro-americana	média	pública	acadêmico	57	55	52	

3.2 Visualizando informações sobre o dataframe

In [25]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   genero      200 non-null    object
 1   raca        200 non-null    object
 2   clasocial   200 non-null    object
 3   tipescola   200 non-null    object
 4   programa    200 non-null    object
 5   ler         200 non-null    int64 
 6   escrever    200 non-null    int64 
 7   matematica  200 non-null    int64 
 8   ciencias    200 non-null    int64 
 9   estsociais  200 non-null    int64 
dtypes: int64(5), object(5)
memory usage: 15.8+ KB
```

4. Análise Descritiva dos Dados

Análise descritiva é uma técnica de análise de dados que visa resumir, organizar e compreender dados históricos para identificar padrões e relacionamentos. É um dos quatro tipos principais de análise de dados, juntamente com a *análise diagnóstica*, *análise preditiva* e a *análise prescritiva* (MÉTRICAS BOSS, 2023). A análise descritiva é usada para:

- Descrever um evento, fenômeno ou resultado;
- Compreender o que aconteceu no passado;
- Acompanhar tendências;
- Identificar medidas de tendência central, dispersão e distribuição dos dados;
- Visualizar padrões; e
- Identificar outliers. A análise descritiva é essencial para explorar e compreender os dados antes de prosseguir para análises mais avançadas (SIRIUS, 2022). Apesar de ser uma estratégia simples, realizada no início do trabalho com os dados, a análise descritiva pode ter diferentes tipos, e essa classificação depende da quantidade de elementos que serão interpretados. Os três tipos de classificação são:
- Univariada: análise de dados que trabalha com apenas uma variável de forma isolada, sem se relacionar com as outras do dataset analisado. Apresenta apenas uma característica;
- Bivariada: análise feita utilizando-se de duas variáveis. O objetivo é investigar a forma que uma variável se comporta em contato com outra e medir a relação que existe entre as duas;
- Multivariada: análise realizada simultâneamente entre diversos elementos do dataset, relacionando-os entre si permitindo obter inferências mais elaboradas.

4.1. Análise Descritiva Univariada

Análise descritiva univariada é utilizada na **Análise Exploratória de Dados** (AED) para sumarizar ou descrever a distribuição de uma única variável de um conjunto de dados. O processo de análise univariada consiste basicamente em, para cada uma das variáveis individualmente:

- classificar a variável quanto a seu tipo: qualitativa ou quantitativa; e
- elaborar gráficos e/ou medidas que resumam a variável analisada.

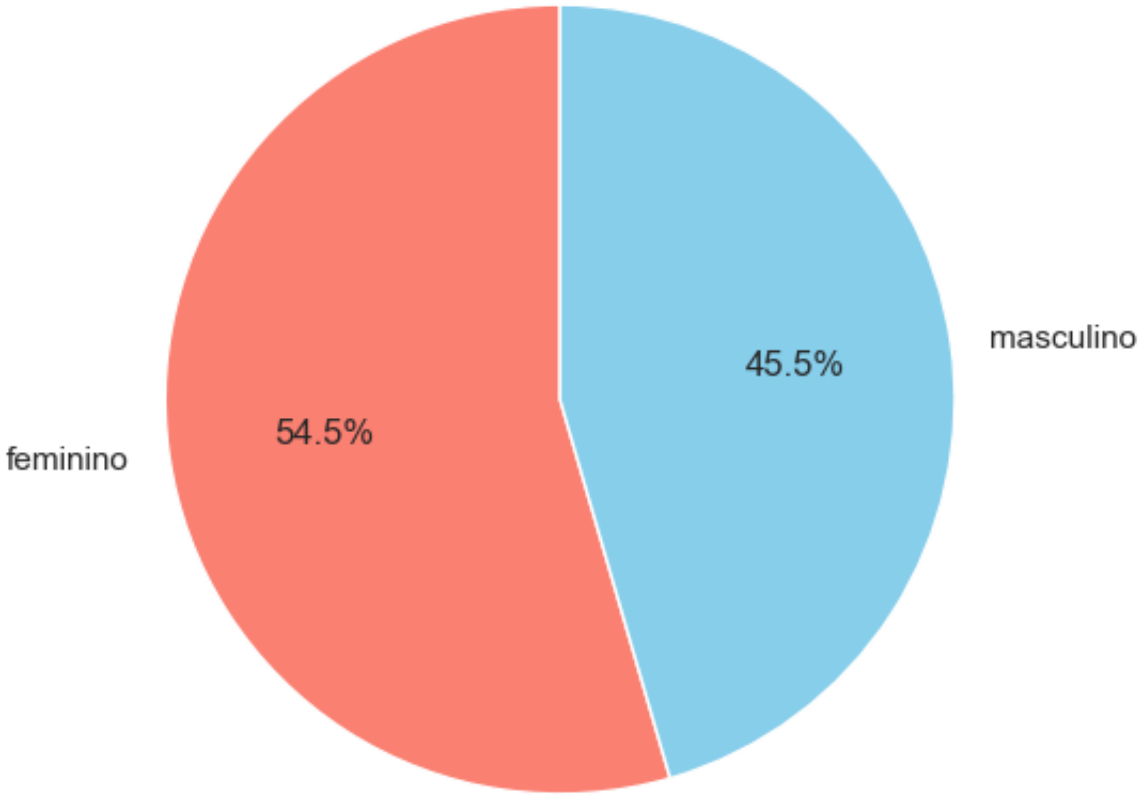
Os gráficos mostrados a seguir nessa análise buscam mostrar, atravéz da visualização dos dados no formato *pizza*. . Os gráficos de pizza normalmente são usados quando se necessita visualizar uma porcentagem ou partes de um todo. A visualização do gráfico no formato pizza pode sermais eficaz para exibir dados rápidos. Como visto nos parágrafos anteriores, a *análise descritiva univariada* permite entender como uma variável se comporta com relação a um todo e demonstrar a contribuição de cada elemento analisado para o todo de forma ágil e fornecendo informações importantes para a tomada de decisões.

4.1.1. Análise descritiva sobre gênero

Para verificar o perfil da amostra pesquisada, foram analisados os dados sobre gênero no dataset. O gráfico da **Figura 1** mostra a distribuição de gênero observado na amostra pesquisada, realizada de forma aleatoriamente na High School and Beyond (2024), representado por um gráfico de pizza.

In [26]:
```
plt.figure(figsize=(6, 6))
df['genero'].value_counts().plot.pie(autopct='%1.1f%%', startangle=90, co
plt.title('Figura 1. Distribuição de gênero')
plt.ylabel('')
plt.show()
```

Figura 1. Distribuição de gênero



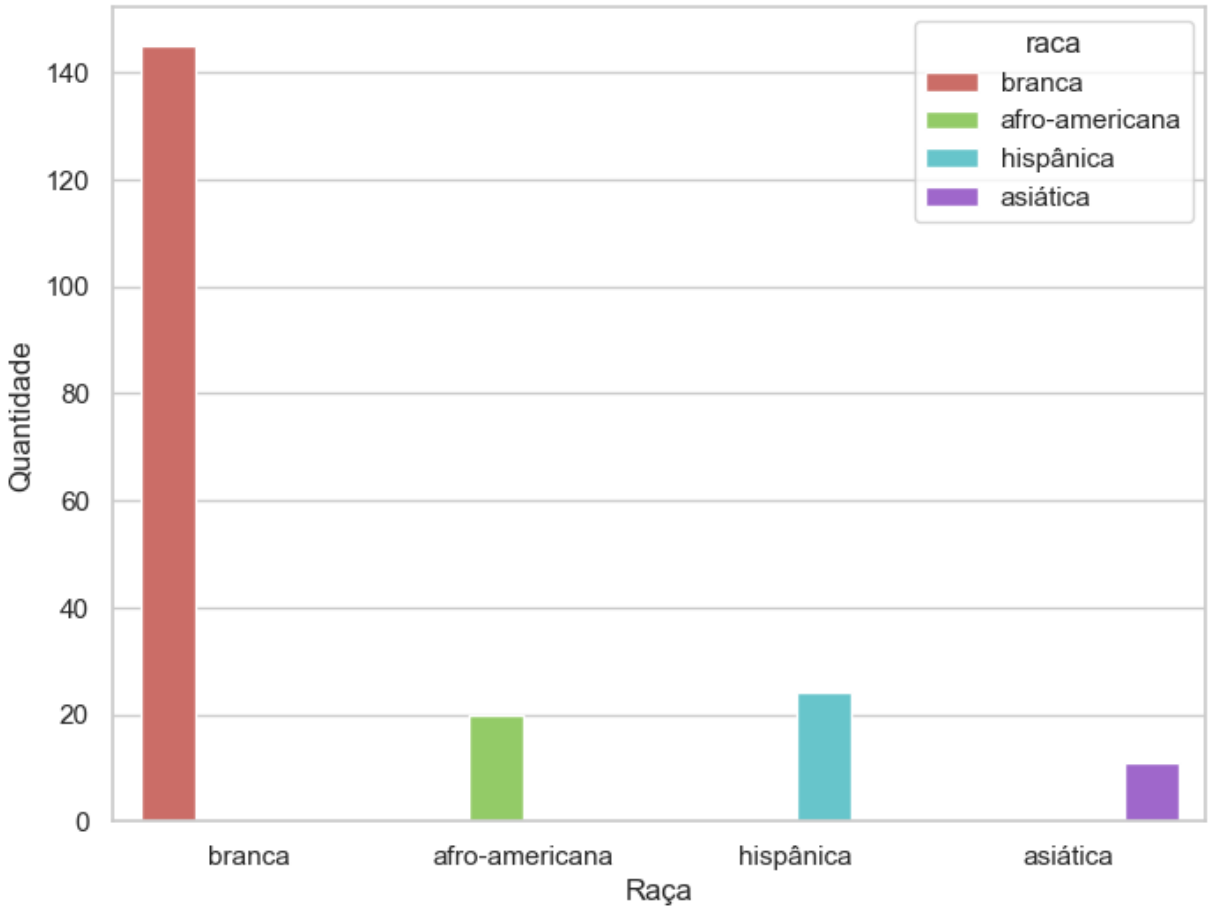
A porcentagem de indivíduos do gênero feminino é de 54,5%, enquanto a porcentagem de indivíduos do gênero masculino é de 45,5%. Isso indica que há uma leve predominância de mulheres em relação aos homens no grupo analisado, com uma diferença de 9% entre os dois gêneros.

4.1.2. Análise descritiva sobre raça

O gráfico de barras na **Figura 2** mostra a distribuição da variável rotulada "raça", entre as diversidades étnicas raciais encontrada nos dados observados de forma aleatoriamente na High School and Beyond (2024). Os gráficos de barras foram considerados adequados para representar esta variável. O gráfico de barras é formado pelas categorias no eixo X, e pela frequência no eixo Y. A frequência utilizada pode ser tanto a absoluta quanto a relativa, conforme é o caso da variável raça.

In [27]:
```
plt.figure(figsize=(8, 6))
sns.countplot(data=data, x='raca', palette='hls', hue='raca')
plt.title('Figura 2. Distribuição da raça')
plt.xlabel('Raça')
plt.ylabel('Quantidade')
plt.show()
```

Figura 2. Distribuição da raça

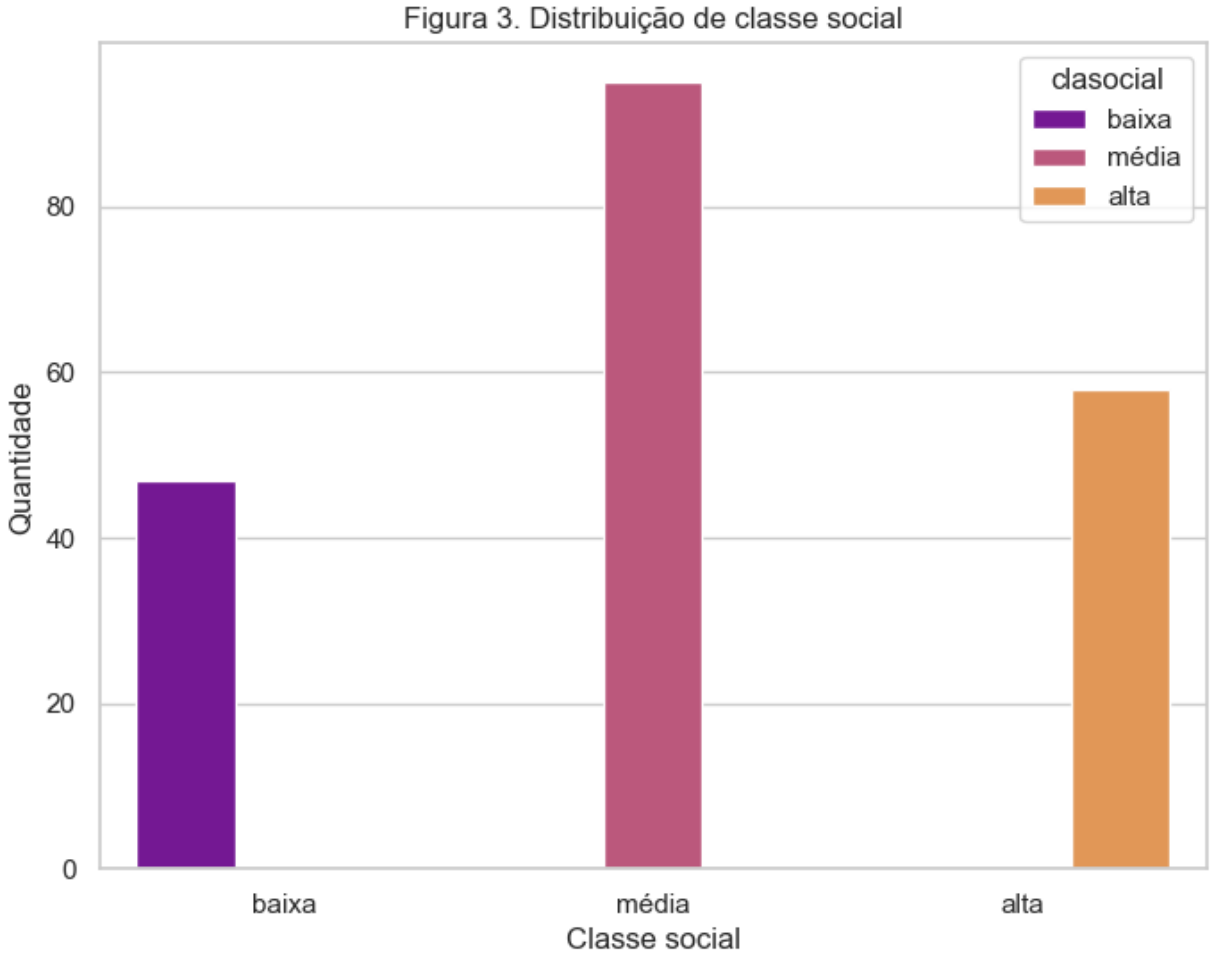


As categorias apresentadas são "branca", "afro-americana", "hispânica" e "asiática".
Observações com relação ao gráfico representado pela **Figura 2**:
1.**Categoria Dominante** : A categoria "branca" tem a maior quantidade, com um valor acima de 140. Isso sugere que os indivíduos nesta categoria constituem a maioria absoluta do conjunto de dados.
2.**Categorias Minoritárias**: As outras categorias raciais ("afro-americana", "hispânica" e "asiática") possuem contagens significativamente mais baixas. Os grupos "afro-americana" e "hispânica" possuem quantidades quase semelhantes, enquanto a categoria "asiática" tem a contagem mais baixa.
3.**Disparidade**: Há uma clara disparidade na representação entre a categoria "branca" e as outras categorias, indicando um desequilíbrio na composição racial deste conjunto de dados.
Essa distribuição pode ser relevante para análises demográficas, e a representação significativa de um grupo pode impactar interpretações em estudos onde a diversidade racial é um fator.

4.1.3. Análise descritiva sobre classe social**

O gráfico apresentado na **Figura 3** representa a distribuição da classe social em um determinado grupo de pessoas. Cada barra representa uma classe social (baixa, média e alta), e a altura da barra indica a quantidade de pessoas que se encaixam nessa classificação.

In [28]:
```
plt.figure(figsize=(8, 6))
sns.countplot(data=data, x='clasocial', palette='plasma', hue='clasocial'
plt.title('Figura 3. Distribuição de classe social')
plt.xlabel('Classe social')
plt.ylabel('Quantidade')
plt.show()
```



Pode-se observar na **Figura 3** que:

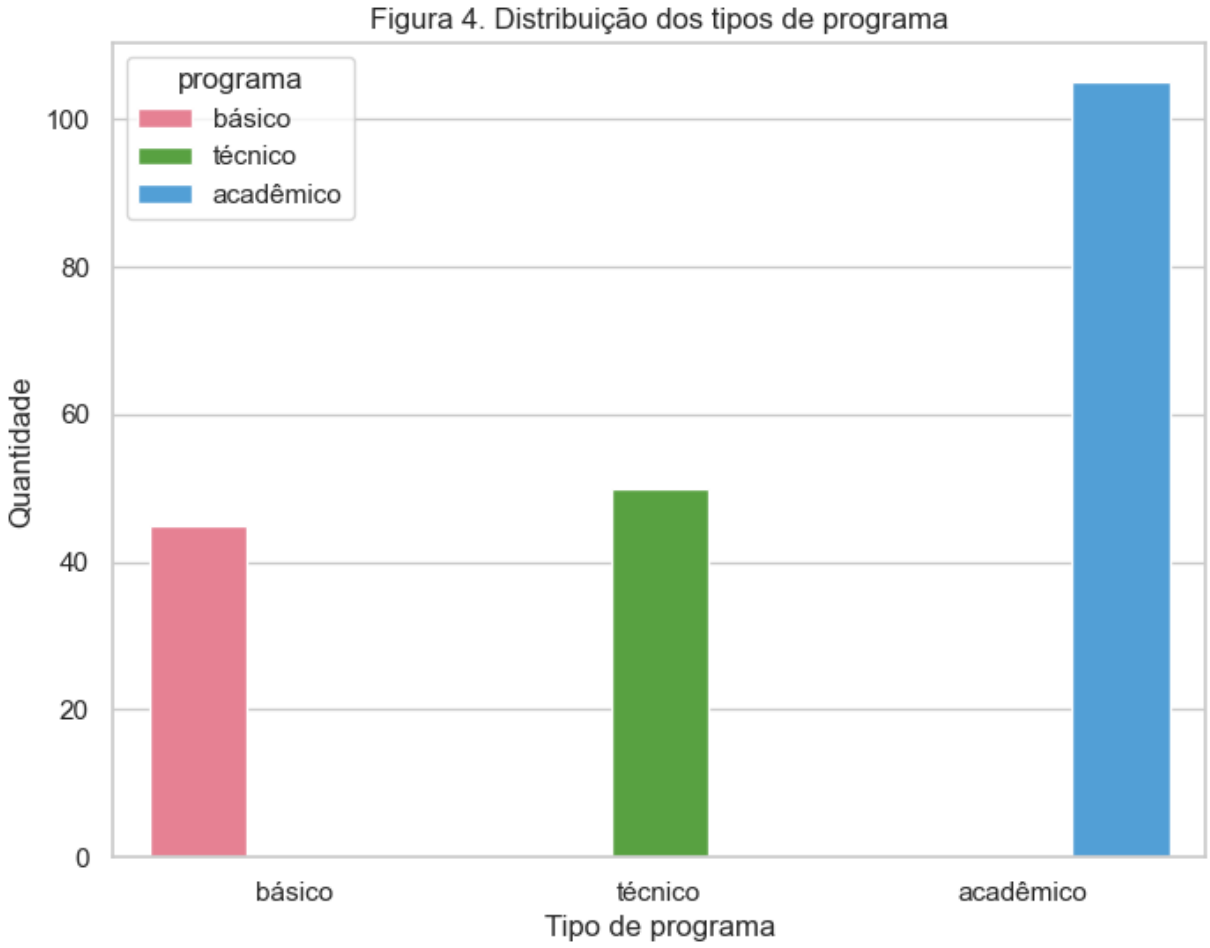
- **Maioria classe média:** A barra correspondente à classe média é a mais alta, indicando que a maior parte das pessoas no grupo analisado se classifica como classe média.
- **Classe alta em segundo lugar:** A classe alta possui uma quantidade intermediária de pessoas, ficando em segundo lugar em relação à classe média.
- **Menor quantidade classe baixa:** A extensão da barra representando a classe baixa é a menor, sugerindo que há uma menor quantidade de pessoas nessa categoria.

  Este gráfico nos permite visualizar a distribuição da classe social de forma clara e concisa. Ele indica que, no grupo analisado, a classe média é predominante, seguida pela classe alta e, por fim, pela classe baixa.

4.1.4. Análise descritiva sobre tipo de currículo**

A **Figura 4** demonstra a quantidade de cada tipo de programa ou currículo (básico, técnico e acadêmico) das escolas da amostra. Cada barra representa um tipo de programa e a altura da barra indica a quantidade total desse tipo.

In [29]:
```
plt.figure(figsize=(8, 6))
sns.countplot(data=data, x='programa', palette='husl', hue='programa')
plt.title('Figura 4. Distribuição dos tipos de programa')
plt.xlabel('Tipo de programa')
plt.ylabel('Quantidade')
plt.show()
```



No gráfico da **Figura 4** pode-se constatar que:

- **Maioria dos programas:** Nota-se que a maior parte dos programas é do tipo acadêmico.
  A barra correspondente a esse tipo é significativamente mais alta que as demais, indicando um número consideravelmente maior de programas nessa categoria.

- **Menor quantidade:** Os programas básicos apresentam a menor quantidade entre os três tipos.
  A barra correspondente é a mais baixa do gráfico.

- **Quantidade intermediária:** Os programas técnicos ocupam uma posição intermediária, com uma quantidade maior que os básicos, mas menor que os acadêmicos.

Sendo assim, com os dados desta amostra, pode-se inferir que:

- Foco em programas acadêmicos: As instituições ou áreas analisadas possuem um forte foco em programas acadêmicos, com uma oferta muito maior desse tipo de programa em comparação aos demais.
- Diversidade de opções: A instituição oferece uma variedade de programas, abrangendo desde os básicos até os acadêmicos.
- Desigualdade na distribuição: A distribuição dos programas não é uniforme, com uma concentração significativa nos programas acadêmicos.

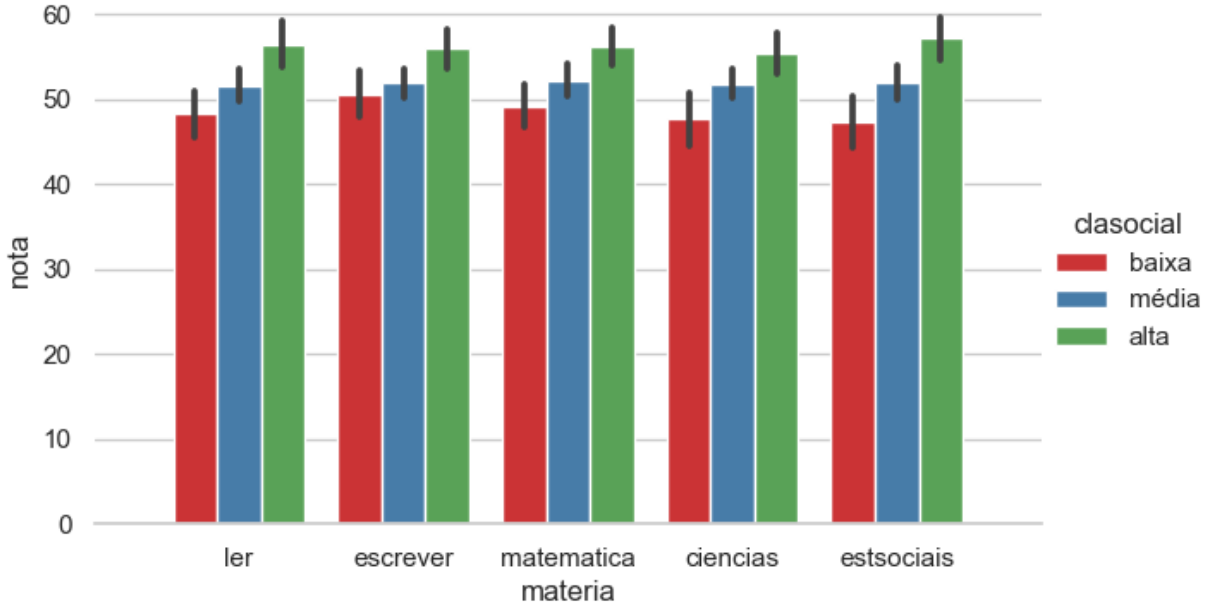
4.1.5. Análise descritiva multivariada sobre classe social e disciplinas cursadas**

A análise multivariada é uma técnica estatística que analisa a relação entre diferentes variáveis de um conjunto de dados. É uma ferramenta que permite a análise simultânea de três ou mais variáveis, e pode ser usada para identificar padrões e tendências que não seriam percebidos em uma análise isolada e/ou bivariada. Com o intuito de verificar o desempenho de alunos em diferentes disciplinas (leitura, escrita, matemática, ciências e estudos sociais), divididos por classe social (baixa, média e alta) o gráfico da **Figura 5** foi elaborado.

In [30]:
```
data_long = data.melt(id_vars=["clasocial"],
                      value_vars=["ler", "escrever", "matematica", "cienc
                      var_name="materia", value_name="nota")

sns.set_style('whitegrid')
g = sns.catplot(x="materia", y="nota", hue="clasocial", data=data_long, k
for ax in g.axes.flat:
    sns.despine(ax=ax, left=True)
    ax.margins(x=0.1)
plt.title('Figura 5. Desempenho em Disciplinas por Classe Social')
plt.subplots_adjust(wspace=0, bottom=0.18, left=0.06)
plt.show()
```

Figura 5. Desempenho em Disciplinas por Classe Social



Cada barra representa a nota média nas disciplianas cursadas dos grupos de alunos agrupados pelas cores representando as classes sociais.
Uma análise mais detalhado da **Figura 5** demonstra que:

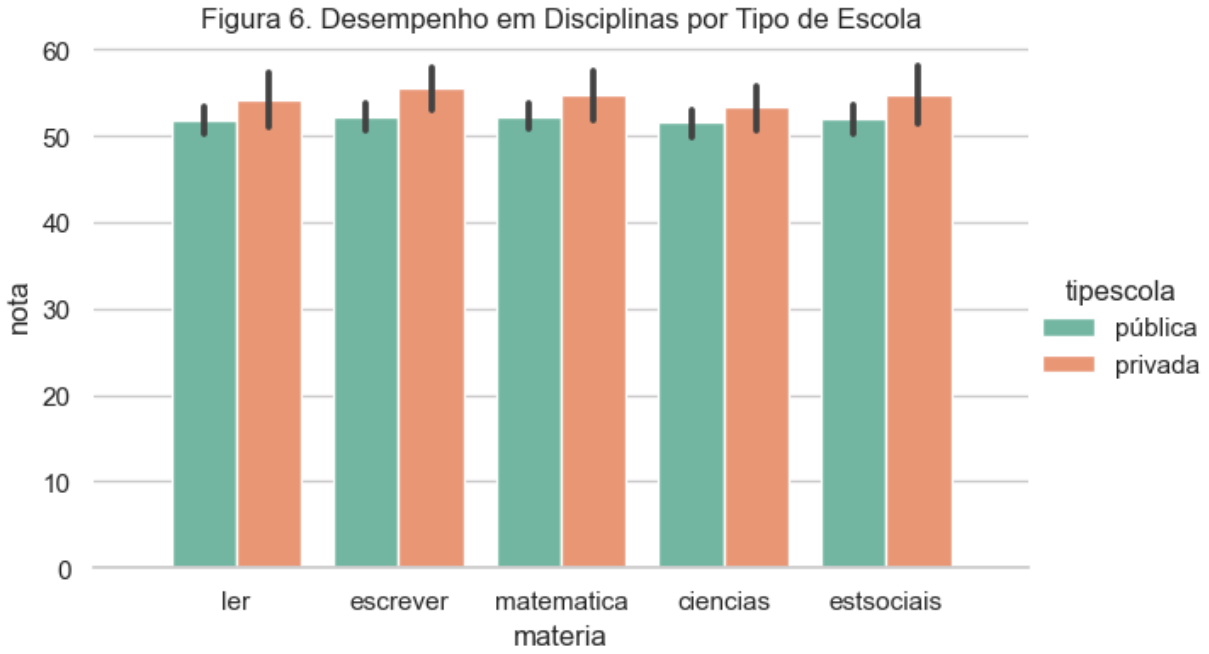
- **Desempenho geral:** De forma geral, o gráfico sugere que o desempenho dos alunos varia de acordo com a matéria e a classe social.
- **Influência da classe social:** As barras verdes (classe alta) tendem a ser mais altas, indicando um desempenho geralmente superior em comparação aos alunos de classe média (azul) e baixa (vermelha).
  Essa diferença é mais evidente em algumas matérias do que em outras.
- **Desempenho por matéria:** O desempenho varia significativamente entre as diferentes matérias.
  Algumas discioplinas, tais como ciências e estudos sociais, apresentam uma diferença mais acentuada entre as classes sociais, enquanto que em outras, como leitura e escrita, essa diferença parece ser menor.
- **A classe social influencia o desempenho escolar:** Alunos de classe social mais alta tendem a apresentar melhores resultados nas diferentes matérias.
- **As diferenças de desempenho variam entre as matérias:** Algumas matérias são mais influenciadas pela classe social do que outras, de acordo com o gráfico.
- **Fatores além da classe social:** É importante ressaltar que a classe social é apenas um dos fatores que podem influenciar o desempenho escolar. Outros fatores, como a qualidade do ensino, recursos disponíveis em casa, expectativas familiares, etc., também podem desempenhar um papel importante.

4.1.6. Análise descritiva multivariada sobre tipo de escola e disciplinas cursadas**

O gráfico na **Figura 6** nos dá uma visão geral do desempenho dos alunos em escolas públicas e privadas.
O gráfico mostra que, em geral, as notas são bastante semelhantes nas duas redes de ensino, com pequenas variações em algumas disciplinas.

In [31]:
```
data_long = data.melt(id_vars=["tipescola"],
                      value_vars=["ler", "escrever", "matematica", "cienc
                      var_name="materia", value_name="nota")

sns.set_style('whitegrid')
g = sns.catplot(x="materia", y="nota", hue="tipescola", data=data_long, k
for ax in g.axes.flat:
    sns.despine(ax=ax, left=True)
    ax.margins(x=0.1)
plt.title('Figura 6. Desempenho em Disciplinas por Tipo de Escola')
plt.subplots_adjust(wspace=0, bottom=0.18, left=0.06)
plt.show()
```



Para entender completamente o que os dados da **Figura 6** significam, é preciso considerar o contexto em que os dados foram coletados e outras informações relevantes. Deve-se observar no entanto que se os dados comparam duas turmas de uma mesma série, as notas podem ser muito parecidas, mas cada aluno tem sua própria história e suas próprias dificuldades.
Com base nas informações fornecidas, pode-se inferir que a pesquisa realizada na High School and Beyond (2024) revela algumas tendências importantes:

1. Distribuição de Gênero: Há uma leve predominância de mulheres (54,5%) em relação aos homens (45,5%) na amostra analisada.
2. Diversidade Étnica: A categoria "branca" é a mais representada, com uma quantidade significativamente maior do que as categorias "afro-americana", "hispânica" e "asiática". Isso indica uma disparidade na composição racial do grupo.
3. Classe Social: A maioria das pessoas pertence à classe média, seguida pela classe alta e, por fim, pela classe baixa.
   Isso mostra uma predominância da classe média no grupo analisado.
4. Tipos de Programas: A maioria dos programas oferecidos é do tipo acadêmico, com uma quantidade menor de programas técnicos e básicos.
   Isso sugere um foco maior em programas acadêmicos nas instituições analisadas.
5. Desempenho Escolar por Classe Social: O desempenho dos alunos varia de acordo com a matéria e a classe social, com alunos de classe alta geralmente apresentando melhores resultados.
   No entanto, essa diferença é mais acentuada em algumas matérias do que em outras.
6. Desempenho em Escolas Públicas e Privadas: As notas dos alunos são bastante semelhantes nas duas redes de ensino, com pequenas variações em algumas disciplinas.
   Isso indica que, apesar das diferenças contextuais, o desempenho acadêmico é comparável entre escolas públicas e privadas.
   Essas conclusões destacam a importância de considerar fatores como gênero, diversidade étnica, classe social e tipo de programa ao analisar dados educacionais, pois eles podem influenciar significativamente os resultados e interpretações dos estudos.

5. Considerações Finais

Referências

ALVES, Ana. Estatística Aplicada: Análise de Dados. Editora Aprender Estatística Fácil, 2022. MÉTRICAS BOSS. Os 4 tipos de análise de dados e como fazê-los. Blog de Web Analytics. 2023. Disponível em: https://metricasboss.com.br/artigos/os-4-tipos-de-analise-de-dados-e-como-faze-los. Acesso em: 26 nov. 2024. OPENINTRO. High School and Beyond survey. s/d. Disponível em: https://www.openintro.org/data/index.php?data=hsb2. Acesso em: 26 nov. 2024. NCES. High School & Beyond, National Center for Educational Studies,US Department of Education. Disponivel em: https://nces.ed.gov/surveys/hsb/surveydesign.asp. Acesso em: 26 nov. 2024. QUESTIONPRO, O que é uma investigação longitudinal? Blog do Software de pesquisa QuestionPro. 2024. Disponível em: https://www.questionpro.com/blog/pt-br/investigacao-longitudinal/#:~:text=O%20que%20%C3%A9%20uma%20investiga%C3%A7%C3%A3o, Acesso em: 24 nov. 2024. SIRIUS. Entenda o que é análise descritiva, quais são os tipos e o passo a passo para fazer uma! Blog Sirius Educação. 07 set. 2022. Disponível em: https://blog.sirius.education/analise-descritiva/#:~:text=A%20an%C3%A1lise%20descritiva%20%C3%A9%20usada,algum%2 Acesso em: 26 nov. 2024. TATSUOKA, Maurice M. Análise multivariada: técnicas para pesquisa educacional e psicológica (2ª ed.) Nova York: Macmillan, Apêndice F, pp: 430-442, 1988. UNITED STATES DEPARTMENT OF EDUCATION. Institute of Education Sciences. National Center for Education Statistics. High School and Beyond, 1980: A Longitudinal Survey of Students in the United States. Inter-university Consortium for Political and Social Research, 2006-01-12. Disponível em: https://doi.org/10.3886/ICPSR07896.v2. Acesso em 26 nov. 2024.

In [ ]: