PROJETO – 1ª FASE

CADEIRA: TÓPICOS EM ENGENHARIA DE DADOS

DISCENTE: Rodrigo Morita a22308365

DOCENTE: Bruno Cipriano

UNIVERSIDADE LUSÓFONA

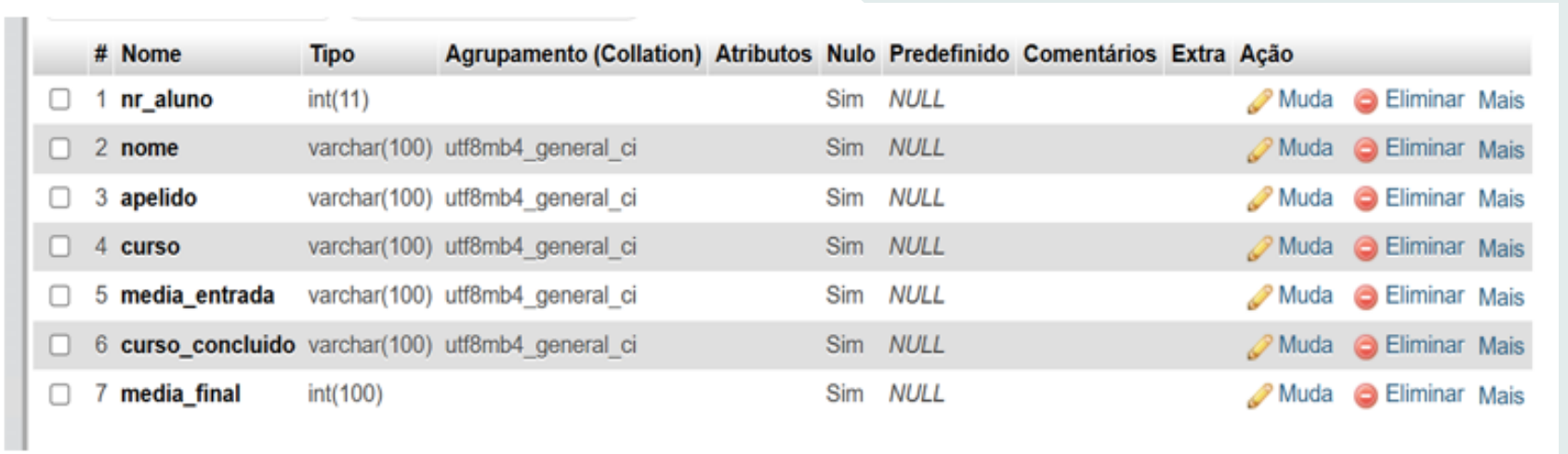
Objetivo da 1ªfase

Encontrar uma relação entre a média de entrada na licenciatura e a média de conclusão da licenciatura.

Para tal foram utilizadas ferramentas como o PHPMYADMIN, JUPYTER NOTEBOOK, GIT BASH e GITHUB

1. Análise do ficheiro “dados_alunos.sql” de forma a perceber qual a estrutura necessária para a Base de Dados.

Criei essa base de dados através do Phpmyadmin com as condições necessárias para que não houvesse erros quando importasse o ficheiro SQL.



#	Nome	Tipo	Agrupamento (Collation)	Atributos	Nulo	Predefinido	Comentários	Extra	Ação
1	nr_aluno	int(11)			Sim	NULL			Muda Eliminar Mais
2	nome	varchar(100)	utf8mb4_general_ci		Sim	NULL			Muda Eliminar Mais
3	apelido	varchar(100)	utf8mb4_general_ci		Sim	NULL			Muda Eliminar Mais
4	curso	varchar(100)	utf8mb4_general_ci		Sim	NULL			Muda Eliminar Mais
5	media_entrada	varchar(100)	utf8mb4_general_ci		Sim	NULL			Muda Eliminar Mais
6	curso_concluido	varchar(100)	utf8mb4_general_ci		Sim	NULL			Muda Eliminar Mais
7	media_final	int(100)			Sim	NULL			Muda Eliminar Mais

2. Analisar os dados, de forma a encontrar dados inválidos e/ou estranhos.

```
SELECT * FROM `aluno` where `media_final` is null;
```

nr_aluno	nome	apelido	curso	media_entrada	curso_concluido	media_final
2003821	Nuno	Laranjeira	Psicologia	16	0	*NULL*
200739	Alice	Pires	Psicologia	17	0	*NULL*
2003485	Alice	Pires	Psicologia	13	0	*NULL*
20211052	Raquel	Castelo	Comunicação	16	0	*NULL*
20211156	Ana	Morais	Comunicação	20	1	*NULL*
20191494	Joana	Castelo	Psicologia	15	0	*NULL*
20211664	Beatriz	Capicua	Informática	16	0	*NULL*
2022848	Carla	Pereira	Psicologia	17	0	*NULL*
2003504	Beatriz	Laranjeira	Comunicação	12	0	*NULL*
2020537	Maria João	Sampaio	Comunicação	12	0	*NULL*
20201546	Anne	Sampaio	Informática	15	1	*NULL*
20031072	Nuno	Pires	Comunicação	14	0	*NULL*
200722	Catarina	da Silva	Psicologia	13	0	*NULL*
20191750	Anne	Cintra	Comunicação	12	0	*NULL*
20221087	Rodrigo	Morais	Informática	16	0	*NULL*
2021847	Patrícia	Pires	Psicologia	10	0	*NULL*
2007103	Catarina	Castelo	Informática	14	0	*NULL*

Quando executo esta query no Phpmyadmin, visualizo que existem 17 valores NULL, por isso mesmo decidi exclui-los para as etapas seguintes de EDA (Exploratory Data Analysis)

2. Continuação

```
SELECT * FROM `aluno` WHERE `curso` is null;
```

nr_aluno	nome	apelido	curso	media_entrada	curso_concluido	media_final
20031371	Leonardo	Cintra	*NULL*	10	1	18
2021840	Sandra	Pires	*NULL*	11	1	19
2003995	Nuno	Silva	*NULL*	16	1	15
2020392	Bruno	Castelo-Branco	*NULL*	19	1	11

Quando executo esta query no Phpmyadmin, visualizo que existem 4 valores NULL, por isso mesmo decidi excluí-los para as etapas seguintes de EDA (Exploratory Data Analysis).

2. Continuação

```
SELECT * FROM `aluno` WHERE `media_entrada` >20 OR `media_entrada`<10;
```

nr_aluno	nome	apelido	curso	media_entrada	curso_concluido	media_final
2003594	Michaelangelo	Laranjeira	Comunicação	0	1	15
2019990	Leonardo	Capicua	Comunicação	22	1	12
2021652	Lucas	Machado	Informática	21	1	17
20201386	Pedro	Pires	Psicologia	22	1	18
20201046	Miguel	Silva	Psicologia	-1	1	17
2019514	Diana	Sampaio	Informática	-2	1	12
2210030	Sininho	(Terra do Nunca)	Cintilar	-20	1	-20

Quando executo esta query no Phpmyadmin, visualizo que existem 3 valores superiores a 20 e 4 valores inferiores a 10, por isso mesmo decidi exclui-los para as etapas seguintes de EDA (Exploratory Data Analysis).
Não faz sentido ter valores superiores a 20 nem inferiores a 10 como média de entrada.

2. Continuação

```
SELECT * FROM `aluno` WHERE `media_final` >20 OR `media_final`<1;
```

nr_aluno	nome	apelido	curso	media_entrada	curso_concluido	media_final
202211	Catarina	Morais	Comunicação	20	1	21
202245	Rui	Pires	Informática	10	1	-2
2007469	Nuno	da Silva	Psicologia	14	1	-1
20071679	Jorge	Sampaio	Comunicação	10	1	-1
20211328	Ana	Sampaio	Informática	12	1	-1
2007975	Rita	Pires	Comunicação	11	1	21
2007691	Miguel	Silva	Informática	14	1	-2
20221456	Lucas	Silva	Informática	18	1	-1
2210030	Sininho	(Terra do Nunca)	Cintilar	-20	1	-20

Quando executo esta query no Phpmyadmin, visualizo que existem 2 valores superiores a 20 e 7 valores inferiores a 10 por isso mesmo decidi exclui-los para as etapas seguintes de EDA (Exploratory Data Analysis).
Não faz sentido ter valores superiores a 20 nem inferiores a 10 como média final de curso.

2.Continuação

```
SELECT DISTINCT (curso) FROM `aluno` WHERE 1;
```

curso
Psicologia
Informática
Comunicação
NULL
Desconhecido
Medicina Veterinária
Má Vida
Cintilar

Quando executo esta query no Phpmyadmin, visualizo que existem nomes de cursos que não devem entrar para a análise (EDA), ou seja, todos os que forem diferentes de "Psicologia", "informática" e "Comunicação" foram excluídos.

```
# Criar um filtro para selecionar apenas os cursos 'Informática', 'Psicologia' e 'Comunicação'
filtro_cursos = df['curso'].isin(['Informática', 'Psicologia', 'Comunicação'])

# Aplicar o filtro ao DataFrame para manter apenas os valores onde o curso é um dos especificados
df_filtrado = df[filtro_cursos]
```

Este código seleciona apenas os cursos pretendidos.

3. Extração dos dados relevantes da base de dados para um ficheiro CSV

	nr_aluno	nome	apelido	curso	media_entrada	curso_concluido	media_final
0	20031358	Leonardo	Pires	Psicologia	15	1	16.0
1	2022158	Raquel	Laranjeira	Informática	17	1	11.0
2	2003812	Anne	Silva	Informática	16	1	16.0
3	2022901	Patrícia	Castelo-Branco	Informática	11	1	17.0
4	2019303	Avelino	Sampaio	Comunicação	13	1	18.0
...	...	...	...	...	...	...	...
1717	2019869	Alice	Laranjeira	Informática	11	1	17.0
1718	20221093	Patrícia	Pereira	Comunicação	14	1	16.0
1719	20071160	Lucas	Capicua	Informática	12	1	20.0
1720	2020288	Pedro	Capicua	Comunicação	13	1	19.0
1721	2019163820	Diana	da Silva	Comunicação	13	1	18.0

1717 rows × 7 columns

```
dados2.to_csv('fase1_1.csv', index=False)
```

Ficheiro csv criado com a Base de dados sem valores estranhos/inválidos.

4. Criação de (pelo menos) um notebook Jupyter que consiga carregar o(s) ficheiro(s) CSV produzido(s).

In [3]:
```
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import scipy.stats as stats
```

In [5]:
```
df = pd.read_csv('fase1_1.csv')
df
```

Out[5]:

	nr_aluno	nome	apelido	curso	media_entrada	curso_concluido	media_final
0	20031358	Leonardo	Pires	Psicologia	15	1	16.0
1	2022158	Raquel	Laranjeira	Informática	17	1	11.0
2	2003812	Anne	Silva	Informática	16	1	16.0
3	2022901	Patrícia	Castelo-Branco	Informática	11	1	17.0
4	2019303	Avelino	Sampaio	Comunicação	13	1	18.0
...	...	...	...	...	...	...	...
1712	2019869	Alice	Laranjeira	Informática	11	1	17.0
1713	20221093	Patrícia	Pereira	Comunicação	14	1	16.0
1714	20071160	Lucas	Capicua	Informática	12	1	20.0
1715	2020288	Pedro	Capicua	Comunicação	13	1	19.0
1716	2019163820	Diana	da Silva	Comunicação	13	1	18.0

1717 rows × 7 columns

5. EDA

Para responder à pergunta: “Existe alguma relação entre a média de entrada na licenciatura e a média de conclusão da licenciatura?”, fiz uma correlação entre a média de entrada e a média final de cada aluno.

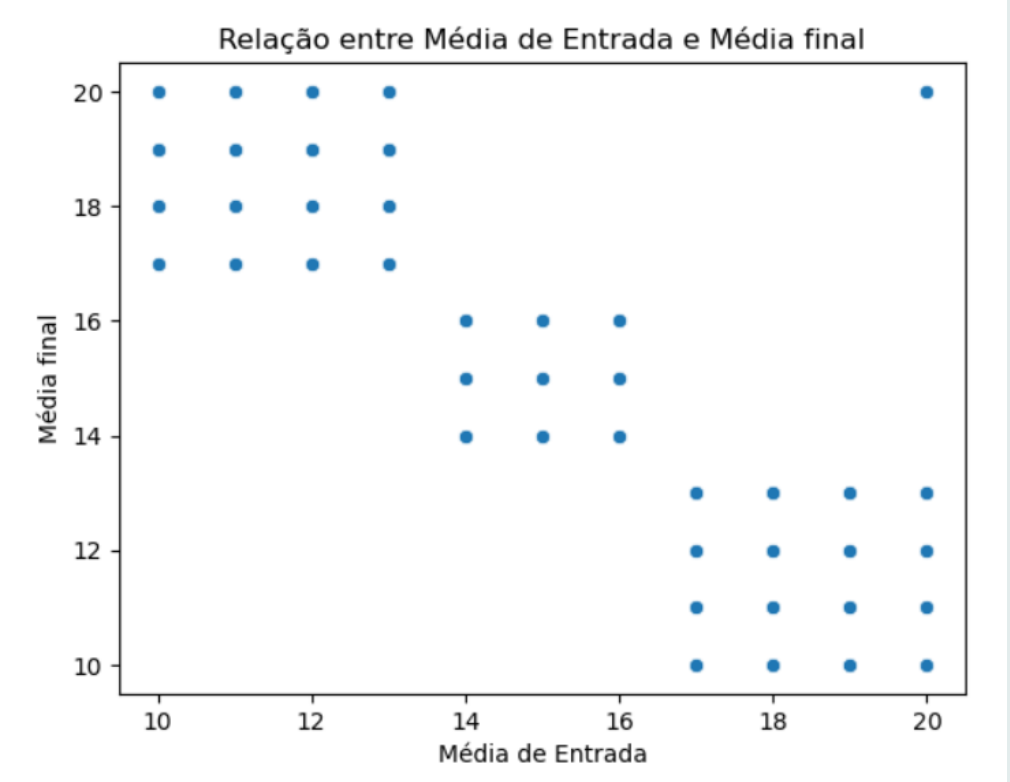
```
# Calcular a correlação
correlation = df['media_entrada'].corr(df['media_final'])
print(f"Correlação entre a média de entrada e a média final {correlation}")

# Criar um gráfico de dispersão
sns.scatterplot(x='media_entrada', y='media_final', data=df)
plt.title('Relação entre Média de Entrada e Média final')
plt.xlabel('Média de Entrada')
plt.ylabel('Média final')
plt.show()
```

```
Correlação entre a média de entrada e a média final -0.8801223419637589
```

Interpretação da correlação

Correlação Forte e Negativa:

- Uma correlação de -0,88 indica uma relação forte e negativa entre as duas variáveis.
- Isso significa que, em geral, alunos com uma média de entrada mais alta tendem a ter uma média de conclusão mais baixa, e vice-versa.

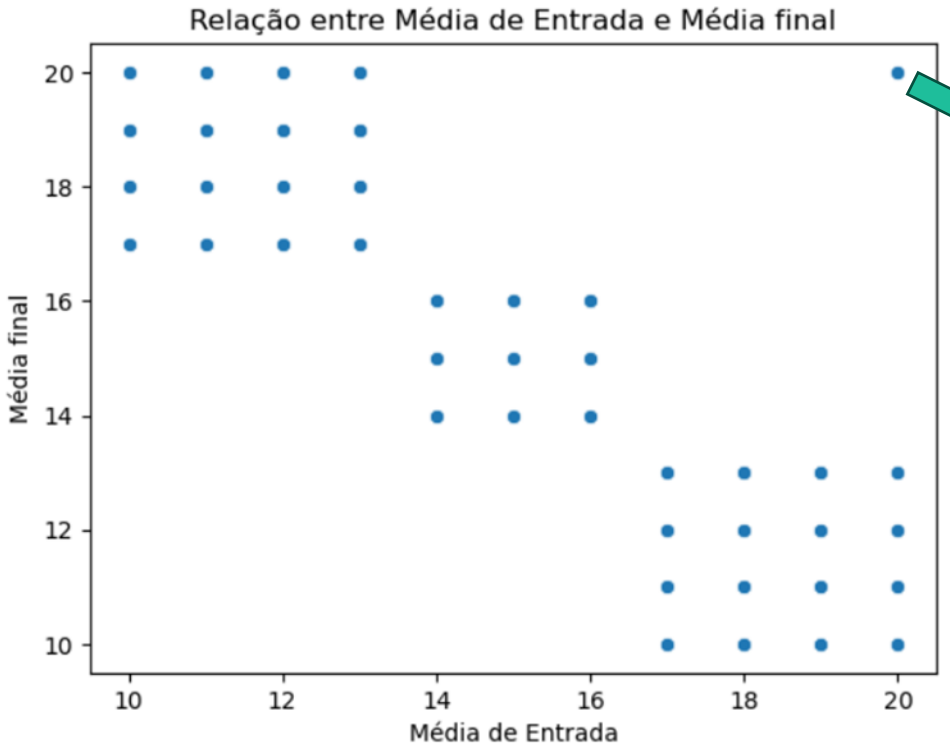


Possíveis Conclusões

Desempenho Inversamente Relacionado:

- Os alunos que entram com melhores notas na licenciatura podem-se desleixar durante o curso que resulta numa média final mais baixa.
- Alternativamente, pode haver fatores externos (por exemplo, pressão ou expectativas) que fazem com que esses alunos não mantenham o seu desempenho inicial.

6. Verificação de Outliers



nr_aluno	nome	apelido	curso	media_entrada	curso_concluido	media_final
2210000	Thomas Anderson	(Neo)	Informática	20	1	20
2290110	Anakin Skywalker	(Lord Vader)	Psicologia	20	1	20
2210100	Luke	Skywalker	Comunicação	20	1	20

Estes são chamados de outliers por se diferenciarem dos restantes.

7. Verificação do número de alunos por cada nota da média final

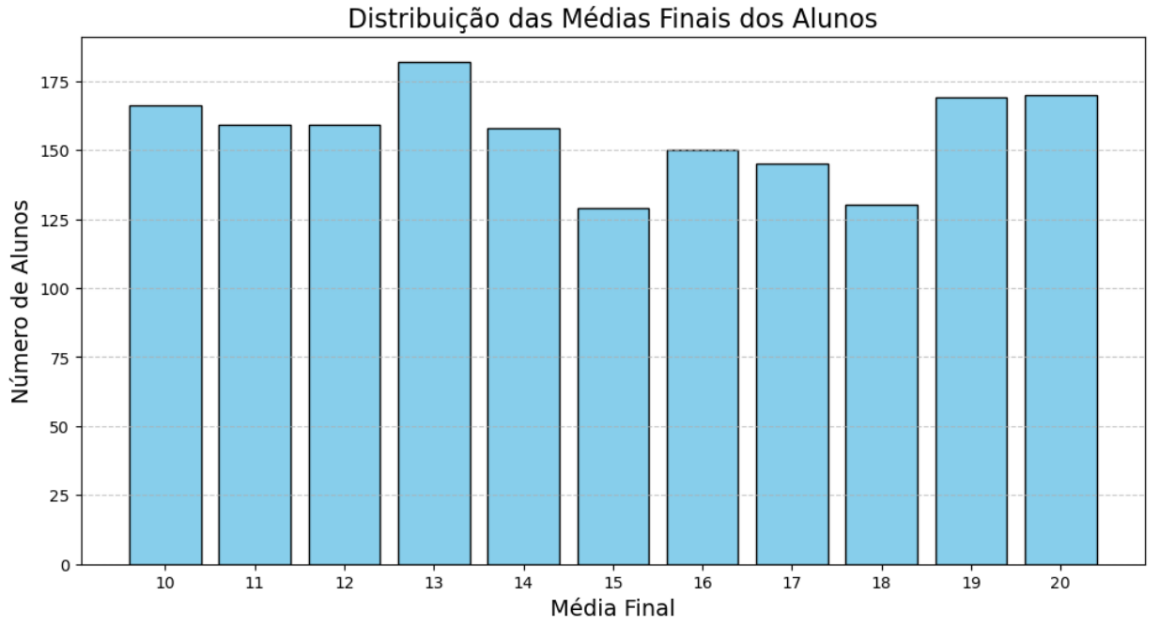
```
: # Contagem do número de alunos para cada nota na coluna 'media_final'
contagem_notas = df['media_final'].value_counts().sort_index()

# Exibir o resultado
print(contagem_notas)
```

```
media_final
10.0    166
11.0    159
12.0    159
13.0    182
14.0    158
15.0    129
16.0    150
17.0    145
18.0    130
19.0    169
20.0    170
Name: count, dtype: int64
```

Com este código consegui descobrir a quantidade de alunos para cada nota de 10 a 20 valores.

7. Verificação do número de alunos por cada nota da média final

```
# Crie o histograma
plt.figure(figsize=(12, 6))
plt.bar(contagem_notas.index, contagem_notas.values, color='skyblue', edgecolor='black')
plt.xlabel('Média Final', fontsize=14)
plt.ylabel('Número de Alunos', fontsize=14)
plt.title('Distribuição das Médias Finais dos Alunos', fontsize=16)
plt.xticks(range(10, 21))  # Defina os rótulos do eixo x de 0 a 20
plt.grid(axis='y', linestyle='--', alpha=0.7)

# Exibe o histograma
plt.show()
```



Com este histograma consegue-se visualizar o numero de alunos que teve a mesma nota e a distribuição das médias finais dos alunos.