

DAN RODRIGUES DE MORAIS DE OLIVEIRA

MVP – ENGENHARIA DE DADOS

OBJETIVO

Analisar o Billionaires Statistics Dataset extraído do Kaggle, e avaliar os seguintes questionamentos:

1. Quais ramos de atividades geraram mais bilionários?
2. Qual a idade média? E qual o mais velho e o mais novo?
3. Qual o país de origem da maioria dos bilionários?
4. Quantos são *self-made* e quantos são herdeiros?
5. Eles residem no mesmo país de origem?
6. Qual a distribuição por gênero?

COLETA E MODELAGEM DOS DADOS

Os dados foram extraídos do Kaggle: <https://www.kaggle.com/datasets/nelgiriyeewithana/billionaires-statistics-dataset/>

Com os dados baixados em CSV para a máquina local, foram inseridos manualmente na plataforma Azure.

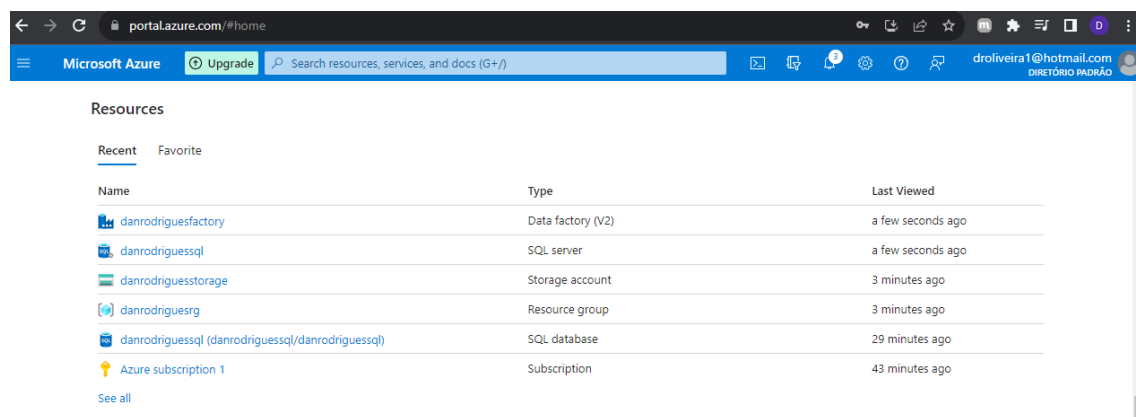
The top screenshot shows the Microsoft Azure portal interface for the 'danrodriguesstorage' storage account. The 'Containers' tab is selected, displaying a list of containers. The table below shows the containers:

Name	Last modified	Anonymous access level	Lease state
slogs	9/30/2023, 5:42:48 PM	Private	Available
danrodriguescontainer	9/30/2023, 5:52:48 PM	Private	Available

The bottom screenshot shows the 'danrodriguescontainer' view. The 'Overview' tab is selected, displaying a table of blobs. The table below shows the blobs:

Name	Modified	Access tier	Archive status	Blob type	Size
Billionaires Statistics Dataset.csv	9/30/2023, 5:52:53 PM	Hot (Inferred)		Block blob	661.

O próximo passo foi criar o SQL databases e Data Factory para manipulação dos dados.

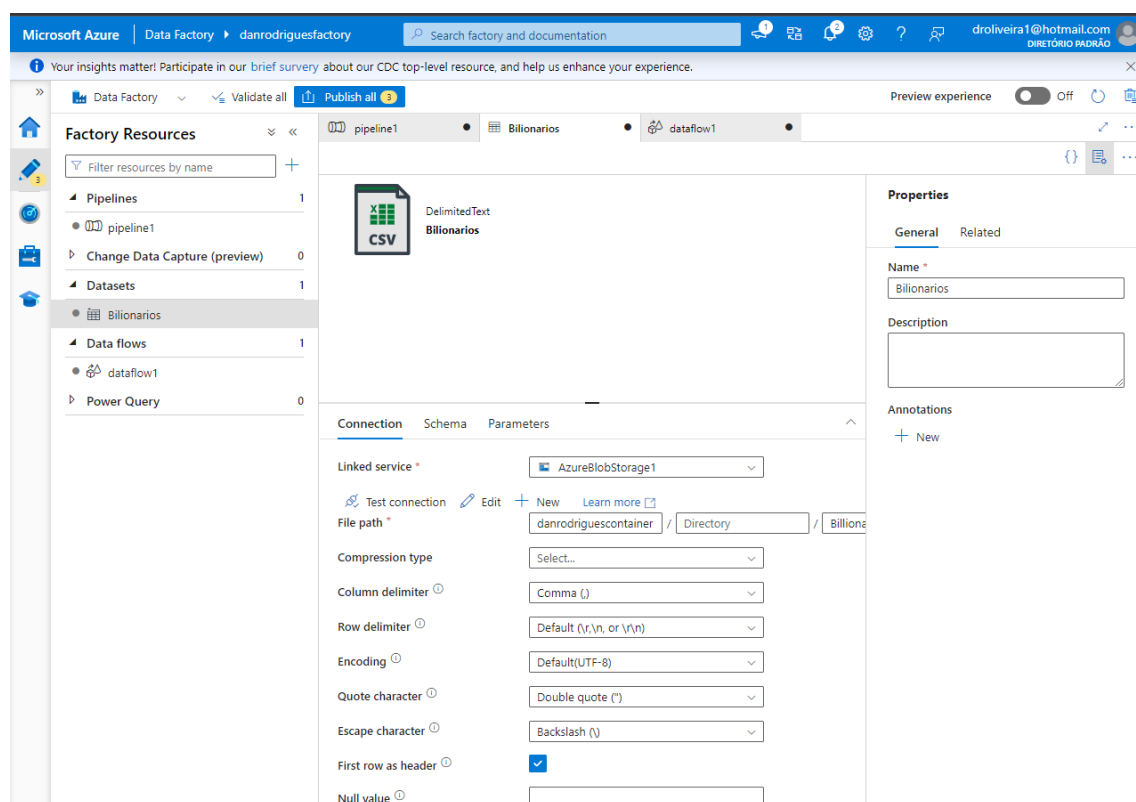


The screenshot shows the Microsoft Azure portal interface. The top navigation bar includes the 'Microsoft Azure' logo, an 'Upgrade' button, a search bar, and a user profile dropdown. The main content area is titled 'Resources' and has two tabs: 'Recent' and 'Favorite'. The 'Recent' tab is selected, showing a table of resources. The table has three columns: 'Name', 'Type', and 'Last Viewed'. The resources listed are:

Name	Type	Last Viewed
danrodriguesfactory	Data factory (V2)	a few seconds ago
danrodriguessql	SQL server	a few seconds ago
danrodriguesstorage	Storage account	3 minutes ago
danrodriguesrg	Resource group	3 minutes ago
danrodriguessql (danrodriguessql/danrodriguessql)	SQL database	29 minutes ago
Azure subscription 1	Subscription	43 minutes ago

Below the table, there is a 'See all' link.

No Data Factory vamos criar o pipeline. Primeiro criamos o Dataset.



The screenshot shows the Microsoft Azure Data Factory interface for configuring a dataset named 'Bilionarios'. The left sidebar shows the 'Factory Resources' tree with 'Bilionarios' selected under 'Datasets'. The main area displays the dataset configuration for 'Bilionarios' (DelimitedText).

Properties

- Name:** Bilionarios
- Description:** (empty text box)
- Annotations:** + New

Connection

- Linked service:** AzureBlobStorage1
- File path:** danrodriguescontainer / Directory / Bilionarios
- Compression type:** Select...
- Column delimiter:** Comma (,)
- Row delimiter:** Default (\r\n, or \n)
- Encoding:** Default(UTF-8)
- Quote character:** Double quote (")
- Escape character:** Backslash (\)
- First row as header:** ☒
- Null value:** (empty text box)

Com o Dataset criado, o próximo passo é criar o DataFlow. Utilizamos a função select para selecionar as colunas que iremos utilizar para análise dos dados. O dataset possuía 35 colunas e selecionamos 6 para análise.

The screenshot shows the Microsoft Azure Data Factory interface. The left sidebar displays 'Factory Resources' including Pipelines, Datasets, and Data flows. The main canvas shows a pipeline named 'dataflow1' with two activities: 'source1' (Import data from Billionarios) and 'select1' (Columns: 6 total). The 'Properties' panel on the right shows the name 'dataflow1' and a description field. The bottom section shows 'Input columns' with 6 mappings and 29 unmapped columns.

Utilizamos a função sink para levar os dados para tratamento em SQL. Em seguida, rodar o pipeline.

The screenshot shows the Microsoft Azure Data Factory interface. The left sidebar displays 'Factory Resources' including Pipelines, Datasets, and Data flows. The main canvas shows a pipeline named 'dataflow1' with three activities: 'source1' (Import data from Billionarios), 'select1' (Columns: 6 total), and 'sink1' (Columns: 6 total). The 'Properties' panel on the right shows the name 'dataflow1' and a description field.

The screenshot shows the Microsoft Azure Data Factory interface. The left sidebar displays 'Factory Resources' including Pipelines, Datasets, and Data flows. The main canvas shows a pipeline named 'dataflow1' with three activities: 'source1' (Import data from Billionarios), 'select1' (Columns: 6 total), and 'sink1' (Columns: 6 total). The 'Properties' panel on the right shows the name 'dataflow1' and a description field. The bottom section shows 'Input columns' with 6 mappings and 29 unmapped columns.

ANÁLISE DOS DADOS

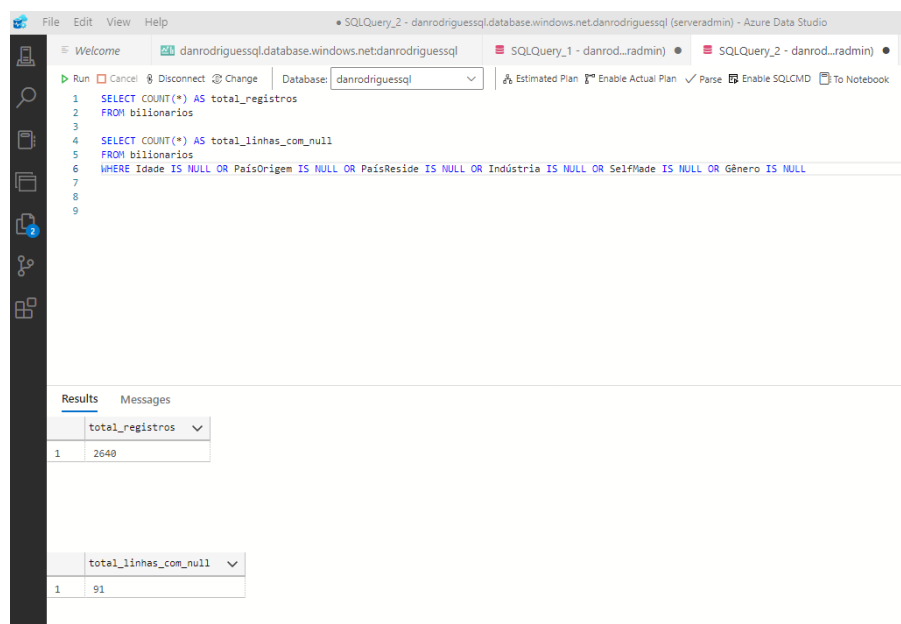
Utilizaremos o Azure Data Studio para fazermos as consultas em SQL e responder as perguntas propostas pelo trabalho.

A base de dados é composta por:

- Idade: idade
- PaísOrigem: país de nascimento
- PaísReside: país de residência
- Indústria: ramo de atuação
- Gênero: gênero (masculino ou feminino)
- Self-Made: fortuna própria (True) ou herdada (False)

Primeiro vamos analisar a qualidade dos dados da base, que possui 2640 registros sendo que 91 são NULL.

Considerando que a quantidade de registros faltantes representa 3% da base e que dado são informações que não seriam tratáveis, devido a sua especificidade, iremos excluir os registros.



The screenshot shows the Azure Data Studio interface with two SQL queries executed against the 'danrodriguesql' database.

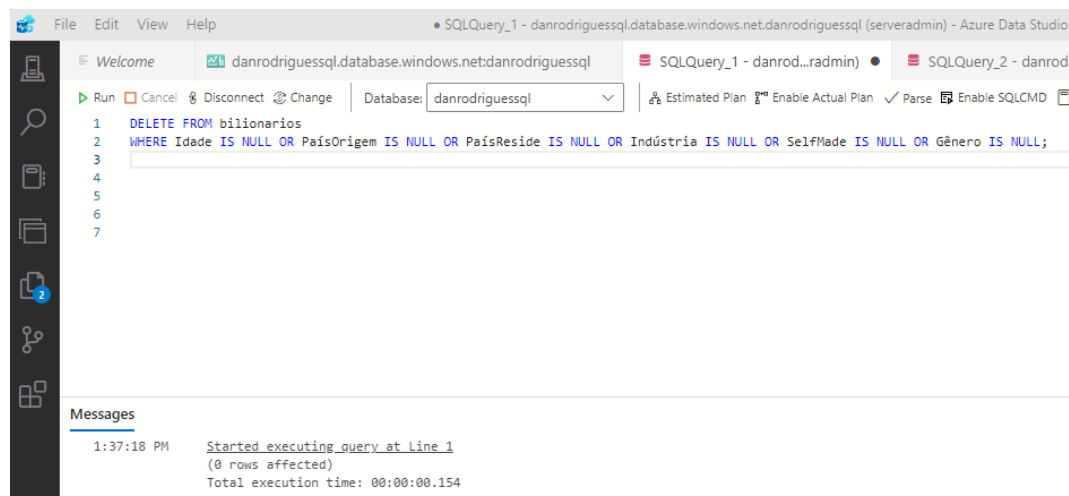
Query 1:

```
1 SELECT COUNT(*) AS total_registros
2 FROM billionaires
3
4 SELECT COUNT(*) AS total_linhas_com_null
5 FROM billionaires
6 WHERE Idade IS NULL OR PaísOrigem IS NULL OR PaísReside IS NULL OR Indústria IS NULL OR SelfMade IS NULL OR Gênero IS NULL
7
8
9
```

Results:

total_registros	
1	2640

total_linhas_com_null	
1	91



The screenshot shows the Azure Data Studio interface with a DELETE query executed against the 'danrodriguesql' database.

Query 1:

```
1 DELETE FROM billionaires
2 WHERE Idade IS NULL OR PaísOrigem IS NULL OR PaísReside IS NULL OR Indústria IS NULL OR SelfMade IS NULL OR Gênero IS NULL;
3
4
5
6
7
```

Messages:

```
1:37:18 PM Started executing query at Line 1
(0 rows affected)
Total execution time: 00:00:00.154
```

Avaliando os ramos de atividade, Finance & Investments foi a que mais gerou bilionários, seguido por Manufacturing, Technology e Fashion & Mall, que representam mais de 48% das fortunas.

SQLQuery_1 - danrodriguessql.database.windows.net:da

File Edit View Help

Welcome danrodriguessql.database.windows.net:danrodriguessql SQLQuery_1 - danr

Run Cancel Disconnect Change Database: danrodriguessql Estimated Plan Er

```

1 SELECT
2     Indústria AS Indústria,
3     COUNT(*) AS Total,
4     (COUNT(*) * 100.0 / (SELECT COUNT(*) FROM bilionarios)) AS Percentual
5 FROM
6     bilionarios
7 GROUP BY
8     Indústria
9 ORDER BY
10    Percentual DESC;
11
12
13
14
15
16

```

Results Messages

	Indústria	Total	Percentual
1	Finance & Investments	361	14.162416633974
2	Manufacturing	311	12.200863083562
3	Technology	307	12.043938799529
4	Fashion & Retail	255	10.003923107100
5	Food & Beverage	200	7.846214201647
6	Healthcare	195	7.650058846606
7	Real Estate	189	7.414672420557
8	Diversified	182	7.140054923499
9	Energy	97	3.805413887799
10	Media & Entertainment	85	3.334641035700
11	Metals & Mining	71	2.785406041584
12	Automotive	70	2.746174970576
13	Service	53	2.079246763436
14	Construction & Engineering	41	1.608473911337
15	Sports	39	1.530011769321
16	Logistics	38	1.490780698313
17	Telecom	30	1.176932130247
18	Gambling & Casinos	25	0.980776775205

Analisando o Gênero entre os bilionários, vemos uma maior presença do gênero masculino, representando 88% da amostra.

FileEditViewHelpSQLQuery_1 - danrodriguesql.database.windows.net.danrodriguesql (serveradmin) - Azur

danrodriguesql.database.windows.net:danrodriguesqlSQLQuery_1 - danrod...radminSQLQuer

RunCancelDisconnectChangeDatabase: danrodriguesqlEstimated PlanEnable Actual PlanParseEnabl

```
1 SELECT
2   Gênero AS Gênero,
3   COUNT(*) AS Total,
4   (COUNT(*) * 100.0 / (SELECT COUNT(*) FROM bilionarios)) AS Percentual
5 FROM
6   bilionarios
7 GROUP BY
8   Gênero
9 ORDER BY
10  Percentual DESC;
```

ResultsMessages

	Gênero	Total	Percentual
1	M	2247	88.15221655511
2	F	302	11.847783444488

Ao avaliarmos se os bilionários são *self-made*, vemos que 69% da amostra construiu a sua fortuna. O restante foi herdado.

FileEditViewHelpSQLQuery_1 - danrodriguesql.database.windows.net.danrodriguesql (serveradmin) -

danrodriguesql.database.windows.net:danrodriguesqlSQLQuery_1 - danrod...radminSQLC

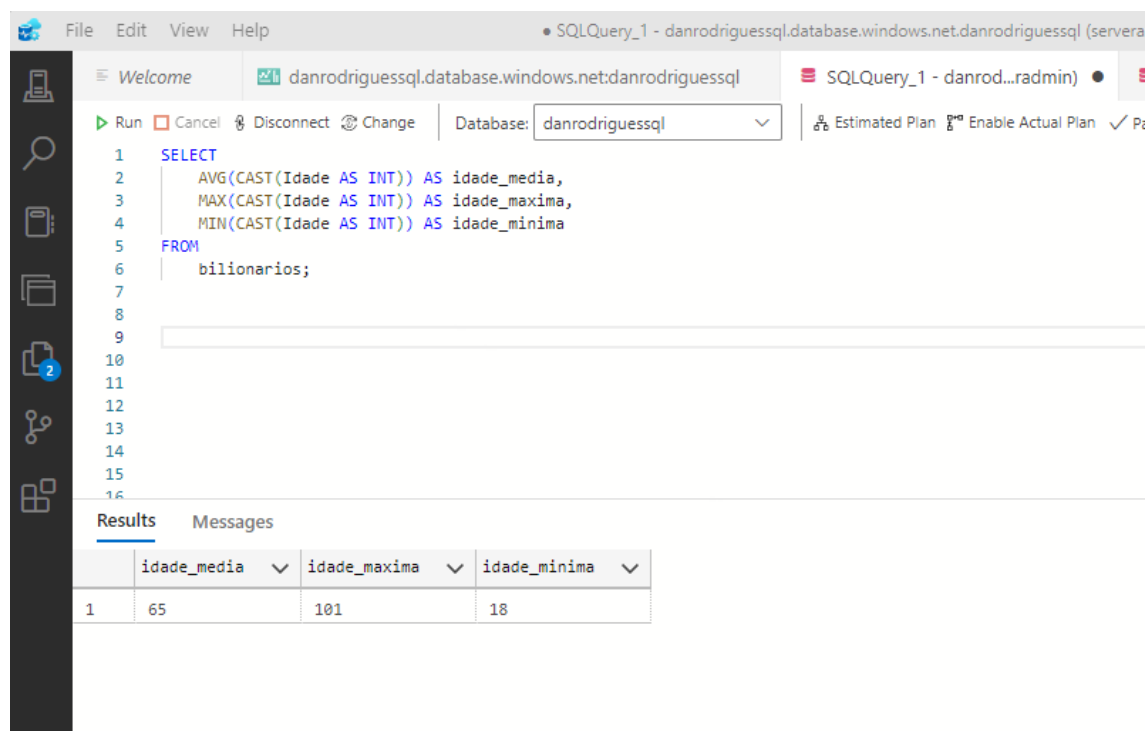
RunCancelDisconnectChangeDatabase: danrodriguesqlEstimated PlanEnable Actual PlanParseE

```
1 SELECT
2   SelfMade AS SelfMade,
3   COUNT(*) AS Total,
4   (COUNT(*) * 100.0 / (SELECT COUNT(*) FROM bilionarios)) AS Percentual
5 FROM
6   bilionarios
7 GROUP BY
8   SelfMade
9 ORDER BY
10  Percentual DESC;
```

ResultsMessages

	SelfMade	Total	Percentual
1	TRUE	1780	69.831306394664
2	FALSE	769	30.168693605335

Ao analisarmos a idade dos bilionários, vemos que a idade média é de 65 anos, sendo que o mais velho tem 101 e o mais novo tem 18 anos.



The screenshot shows the SQL Server Enterprise Manager interface. The query editor displays the following SQL query:

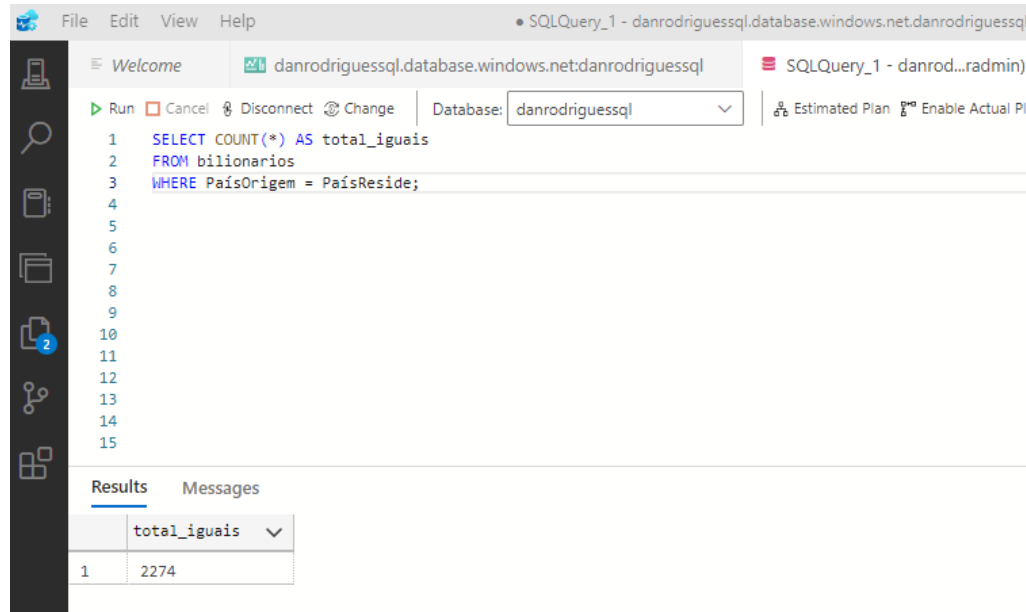
```
1 SELECT
2     AVG(CAST(Idade AS INT)) AS idade_media,
3     MAX(CAST(Idade AS INT)) AS idade_maxima,
4     MIN(CAST(Idade AS INT)) AS idade_minima
5 FROM
6     bilionarios;
```

The Results pane shows the following data:

	idade_media	idade_maxima	idade_minima
1	65	101	18

Analisando o País de origem, temos uma concentração de mais de 61% nos 5 primeiros países, Estados Unidos, China, Índia, Rússia e Alemanha, respectivamente.

Ainda por essa ótica, vemos que 2274 bilionários residem no mesmo país de origem, o que representa mais de 89% da base de 2549 registros, após exclusão dos NULL.



The screenshot shows the SQL Server Enterprise Manager interface. The query editor displays the following SQL query:

```
1 SELECT COUNT(*) AS total_iguais
2 FROM bilionarios
3 WHERE PaísOrigem = PaísReside;
```

The Results pane shows the following data:

	total_iguais
1	2274

CONCLUSÃO

O objetivo proposto pelo trabalho foi atingido. As perguntas foram analisadas e respondidas considerando a base utilizada após tratamento no Azure.

A base apresentava boa qualidade dos dados. Havia alguns registros como NULL (3% da amostra), e dado o típico de dado presente nos registros (idade, país de nascimento e residência, ramo de negócio, gênero e fortuna *self-made*), optou-se por excluir os registros com NULL.

A fim de se aprofundar a análise apresentada e enriquecer os resultados, alguns caminhos poderiam ser tomados:

- Há alguma correlação do país de origem dos bilionários com o nível de desenvolvimento desse país? Comparar com dados como: IDH, nível de educação da população, taxa de crescimento do país.
- Percebe-se uma presença maior do gênero masculino. Uma linha seria incluir a análise de etnias e indicadores de educação e sociais desses países.
- As maiores fortunas vieram do mercado financeiro, porém, fazer um comparativo de taxa de retorno e volatilidade entre ativos financeiros e físicos/materiais, e tentar identificar qual investimento teve a melhor relação risco e retorno.
- Comparar o comportamento dos diferentes ramos de atuação com os ciclos de mercado e como eles se comportarem em momentos de crise e euforia.