

# Relatório MVP Engenharia de Dados

## Objetivo

Com os dados sobre as estatísticas dos jogadores no campeonato PGL CS2 Major Copenhagen 2024, fazer as seguintes análises:

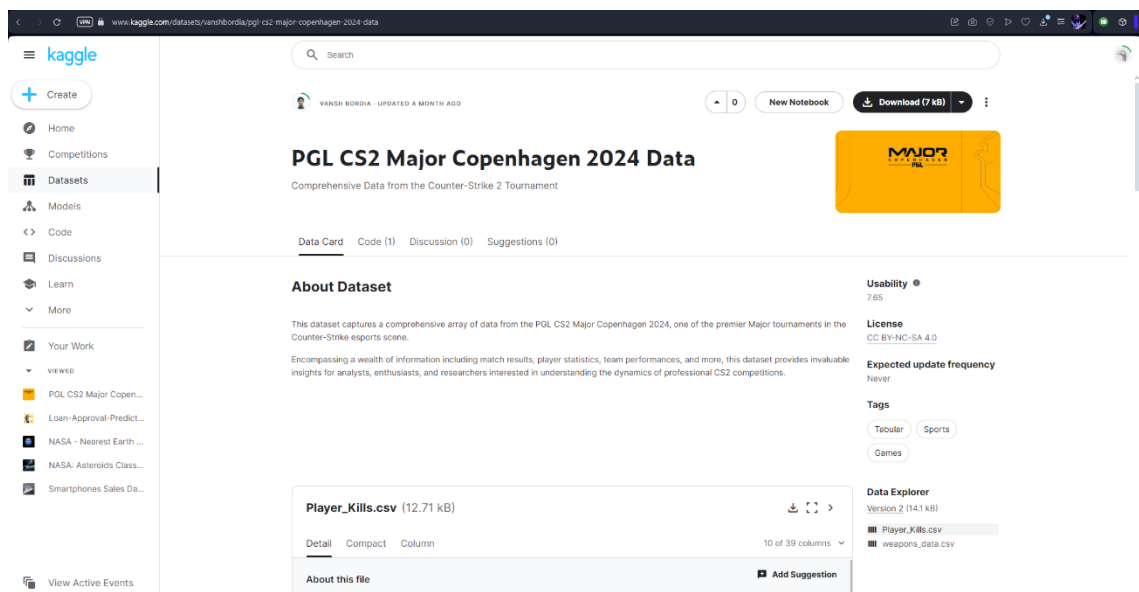
- 1 - Qual o número de kills(inimigos derrubados) do jogador que teve o maior número de MVP(prêmio de melhor jogador do round)?
- 2 - Qual o número de kills por headshot(inimigos derrubados com tiro na cabeça) teve o jogador com o maior número de kills(inimigos derrubados)?
- 3 - Quantos MVP(prêmio de melhor jogador do round) teve o jogador com o maior ADR(média de dano por round)?
- 4 - Qual o número de deaths(mortes) do jogador com o maior número de first kills(derrubou o primeiro inimigo do round)?
- 5 - Quantas assistências teve o jogador com o maior kd(Divisão de kills/deaths)?

## Detalhamento

### 1. Busca pelos dados

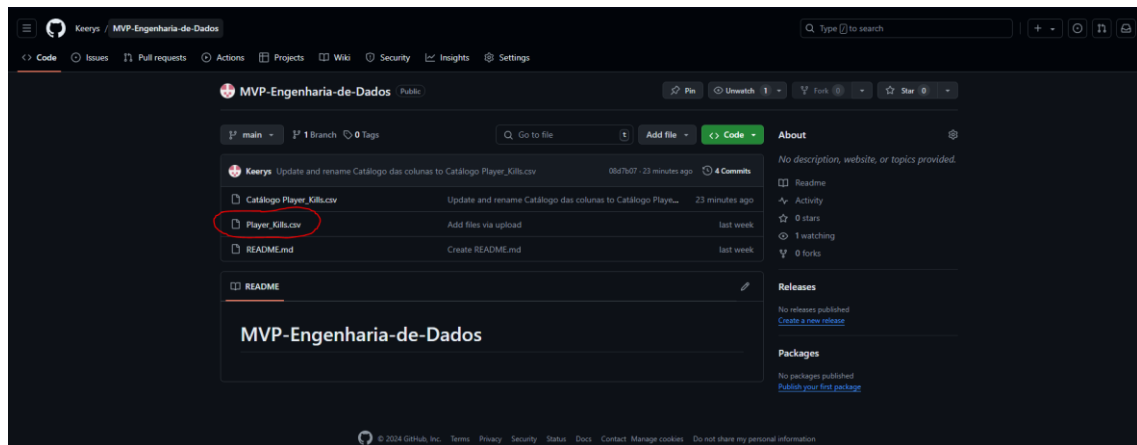
Pesquisa feita no Kaggle, encontrando o seguinte dataset:

<https://www.kaggle.com/datasets/vanshbordia/pgl-cs2-major-copenhagen-2024-data>



## 2. Coleta

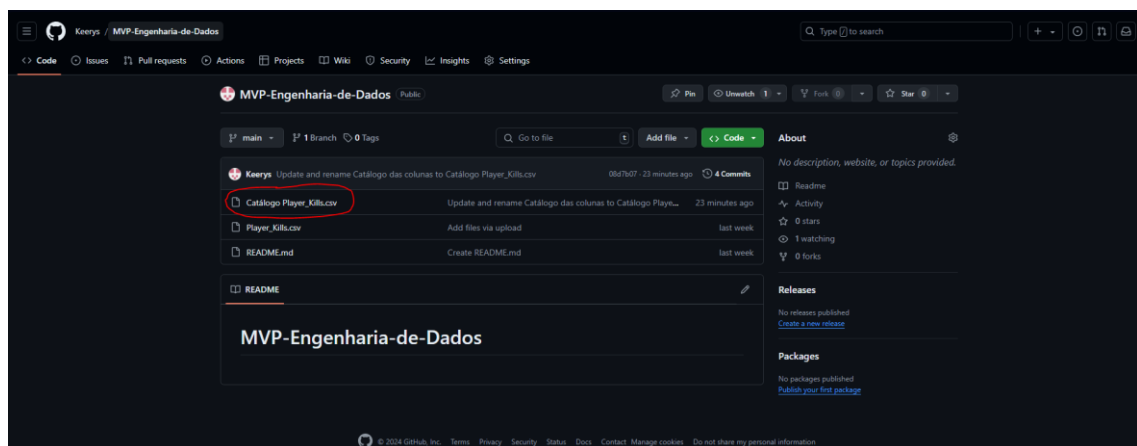
Dados foram baixados manualmente e feito o upload em um repositório público do github



## 3. Modelagem

O modelo de dados escolhido é o Modelo Flat. Todos os dados estão em uma única tabela, não sendo necessário fazer nenhuma junção ou outro tipo de modelagem.

Foi feito um Catálogo dos Dados que está disponibilizado em um repositório público do github com o nome “Catálogo Player\_Kills.csv”.

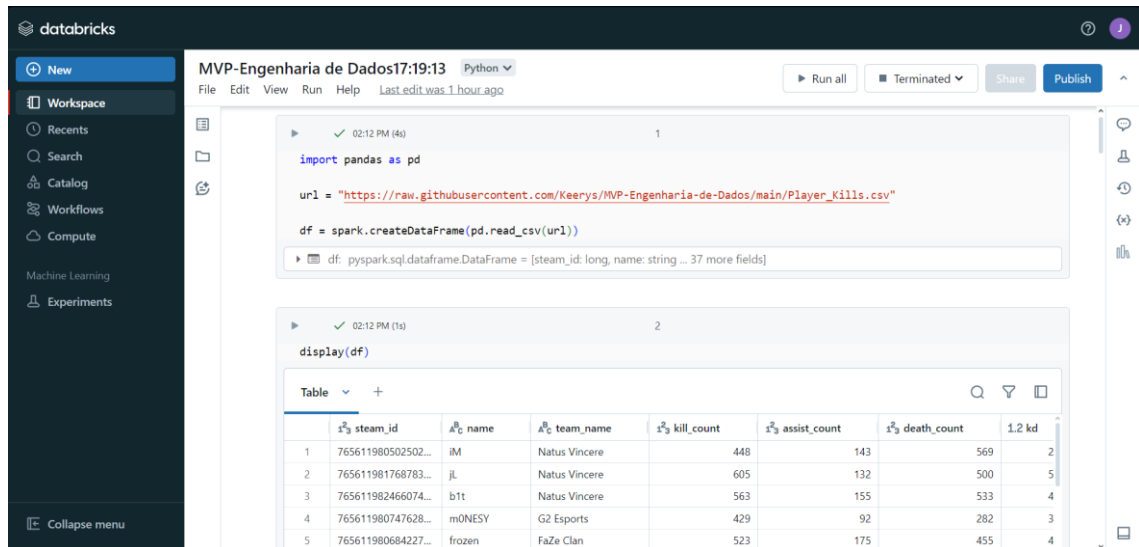


## 4. Carga

O ETL foi feito usando o Databricks Community através da criação de um notebook, com as seguintes etapas:

### Etapa 1 (Extract):

Foi feito o código para extrair os dados do repositório do Github e criado um Dataframe a partir dos dados extraído.



```
import pandas as pd

url = "https://raw.githubusercontent.com/Keerays/MVP-Engenharia-de-Dados/main/Player_Kills.csv"

df = spark.createDataFrame(pd.read_csv(url))
```

df: pyspark.sql.dataframe.DataFrame = [steam\_id: long, name: string ... 37 more fields]

```
display(df)
```

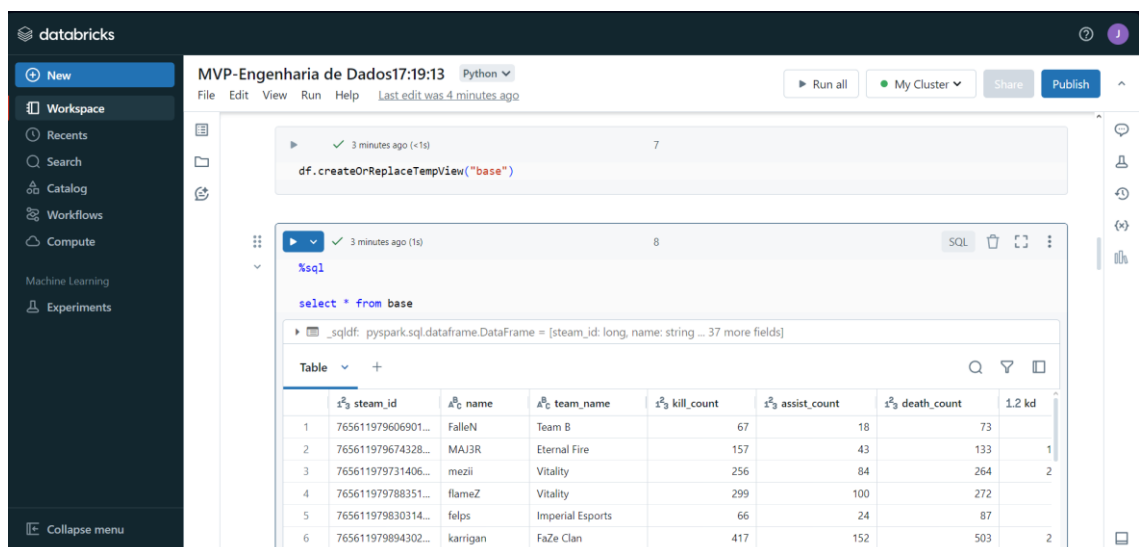
	steam_id	name	team_name	kill_count	assist_count	death_count	kd
1	765611980502502...	iM	Natus Vincere	448	143	569	2
2	765611981768783...	jL	Natus Vincere	605	132	500	5
3	765611982466074...	b1t	Natus Vincere	563	155	533	4
4	765611980747628...	mONESY	G2 Esports	429	92	282	3
5	765611980684227...	frozen	FaZe Clan	523	175	455	4

### Etapa 2 (Transform):

Não foi necessário fazer nenhuma transformação ou modelagem. Como o Modelo é Flat todos os dados já estão em uma única tabela nos formatos desejados para análise.

### Etapa 3 (Load):

Foi criada uma tabela no Databricks Community para carregar os dados no banco de dados.



```
df.createOrReplaceTempView("base")
```

```
%sql

select * from base
```

\_sqldf: pyspark.sql.dataframe.DataFrame = [steam\_id: long, name: string ... 37 more fields]

	steam_id	name	team_name	kill_count	assist_count	death_count	kd
1	765611979606901...	FalleN	Team B	67	18	73	
2	765611979674328...	MAJ3R	Eternal Fire	157	43	133	1
3	765611979731406...	mezli	Vitality	256	84	264	2
4	765611979788351...	flameZ	Vitality	299	100	272	
5	765611979830314...	felps	Imperial Esports	66	24	87	
6	765611979894302...	karrigan	FaZe Clan	417	152	503	2

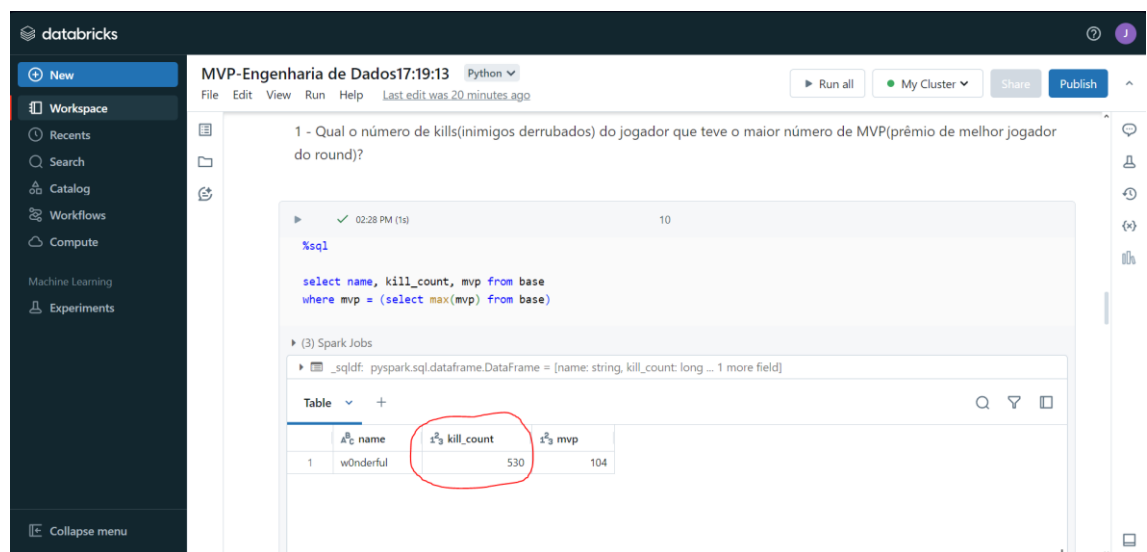
## 5. Análise

### a) Qualidade dos dados

Ao analisar os dados foi encontrado um problema na coluna “kd”, uma vez que esse dado é o resultado da divisão entre as colunas “kill\_count” e “death\_count”. O cálculo feito no dataset baixado está errado, impossibilitando usar os dados da coluna “kd” para obter resultados precisos.

### b) Solução do problema

1 - Qual o número de kills(inimigos derrubados) do jogador que teve o maior número de MVP(prêmio de melhor jogador do round)?



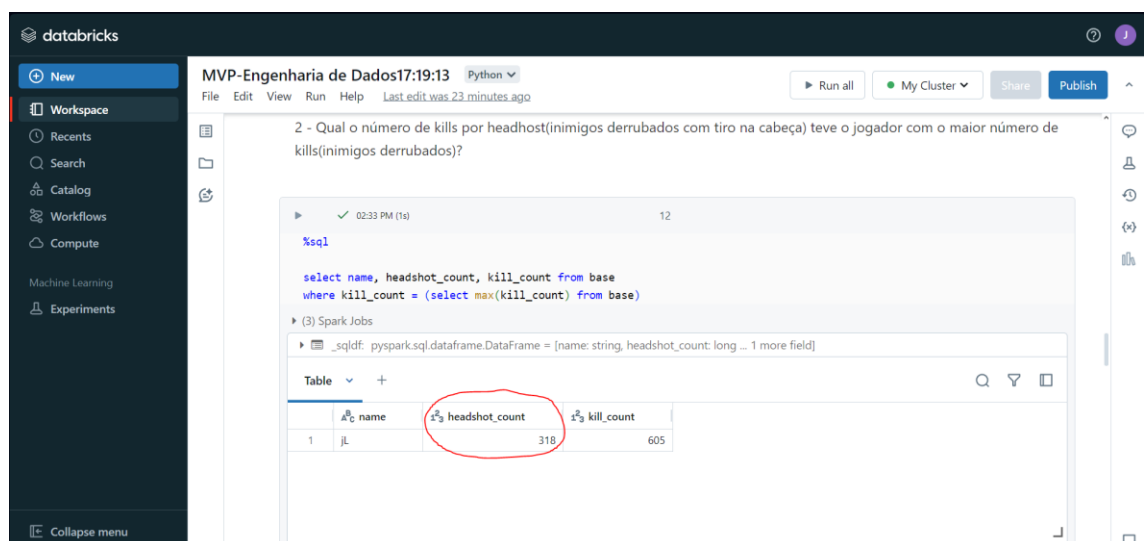
The screenshot shows a Databricks workspace with a notebook titled "MVP-Engenharia de Dados17:19:13". The notebook contains a SQL query that selects the name, kill\_count, and mvp from a table named 'base', where mvp is the maximum mvp value from the same table. The query is executed, and the results are displayed in a table. The table has three columns: name, kill\_count, and mvp. The first row shows a player named 'wonderful' with a kill\_count of 530 and an mvp of 104. The kill\_count value 530 is circled in red.

```
%sql
select name, kill_count, mvp from base
where mvp = (select max(mvp) from base)
```

	name	kill_count	mvp
1	wonderful	530	104

Resposta: 530 kills.

2 - Qual o número de kills por headshot(inimigos derrubados com tiro na cabeça) teve o jogador com o maior número de kills(inimigos derrubados)?



The screenshot shows a Databricks workspace with a notebook titled "MVP-Engenharia de Dados17:19:13". The notebook contains a SQL query that selects the name, headshot\_count, and kill\_count from a table named 'base', where kill\_count is the maximum kill\_count value from the same table. The query is executed, and the results are displayed in a table. The table has three columns: name, headshot\_count, and kill\_count. The first row shows a player named 'jL' with a headshot\_count of 318 and a kill\_count of 605. The headshot\_count value 318 is circled in red.

```
%sql
select name, headshot_count, kill_count from base
where kill_count = (select max(kill_count) from base)
```

	name	headshot_count	kill_count
1	jL	318	605

Resposta: 318 kills por headshot.

3 - Quantos MVP(prêmio de melhor jogador do round) teve o jogador com o maior ADR(média de dano por round)?

The screenshot shows a Databricks workspace with a notebook titled "MVP-Engenharia de Dados17:19:13". The notebook contains a SQL query that filters for the player with the highest ADR and then counts their MVPs. The results table shows one row for player "b1t" with 76 MVPs and an ADR of 2937.905184. The "mvp" column is circled in red.

```
%sql
select name, mvp, adr from base
where adr = (select max(adr) from base)
```

	name	mvp	adr
1	b1t	76	2937.905184

Resposta: 76 MVPs.

4 - Qual o número de deaths(mortes) do jogador com o maior número de first kills(derrubou o primeiro inimigo do round)?

The screenshot shows a Databricks workspace with a notebook titled "MVP-Engenharia de Dados17:19:13". The notebook contains a SQL query that filters for the player with the highest first kill count and then shows their death count. The results table shows one row for player "iM" with 569 deaths and 115 first kills. The "death\_count" column is circled in red.

```
%sql
select name, death_count, first_kill_count from base
where first_kill_count = (select max(first_kill_count) from base)
```

	name	death_count	first_kill_count
1	iM	569	115

Resposta: 569 deaths.

5 - Quantas assistências teve o jogador com o maior kd(Divisão de kills/deaths)?

Resposta: Como foi identificado um erro no cálculo para a obtenção do kd nos dados baixados, não é possível fazer uma análise precisa sobre esta questão levantada.

## **Autoavaliação**

Esse trabalho em um primeiro momento tive muita dificuldade. Como estou mudando de carreira e ainda estou no começo, não possuo muito intimidade com os programas e alguns conceitos.

O objetivo foi fácil de definir, uma vez que consegui usar um banco de dados sobre um assunto que domino. Porém para começar a desenvolver o trabalho tive muitas dificuldades. Se não fosse outros colegas que me auxiliariam pelo Discord com minhas dúvidas eu não teria conseguido realizar o trabalho.

Com as dicas dos colegas, consegui ir desenvolvendo por etapas. Criar tudo a partir de um notebook no Databricks Community facilitou o processo, pois consegui fazer tudo por ele, desde criar o cluster ate as tabelas e o código.

Sobre o resultado do trabalho, somente uma questão proposta pelo objetivo não teve resposta, em função de um dado incorreto que encontrado no banco de dados, o que impossibilita de realizar uma análise precisa. Fora esse problema os outros objetivos foram possíveis ser alcançados sem muitos problemas.

Outro ponto interessante de ressaltar é o fato de que antes de realizar o trabalho, a matéria que estudamos no decorrer do sprint não estava fazendo tanto sentido para mim, contudo ao tentar aplicar e ir fazendo o MVP, foi tudo fazendo sentido e encaixando. Esse projeto me ajudou a entender melhor o conteúdo passado e a desmitificar a matéria. Não foi fácil o processo mas fiquei muito satisfeito de ter conseguido fazer e de fato entender de um modo melhor a matéria.