Impacto do covid na economia e estadia residencial (Data Analysis)

Carlos Ferreira PG47087[1,1*], Joel Martins PG47347[1,1*], Júlio Alves pg47390[1,1*] and Nuno Silva pg42645[1,1*]

[1*]Departmento de Informática, Universidade Uminho, Braga, Portugal.

*Corresponding author(s). E-mail(s): pg47087@alunos.uminho.pt; pg47347@alunos.uminho.pt; pg47390@alunos.uminho.pt; pg42645@alunos.uminho.pt;

Abstract

Este artigo cientifico tem como objetivo descrever o procedimento efetuado para a concretização da planificação realizada no trabalho antecedente, referente ao processo de extração de conhecimento a partir de "datasets" selecionados. Serão abordados todos os passos planeados anteriormente, começando pela arquitetura, onde será feita a descrição dos "datasets" encontrados, a descrição do processo de "ETL" assim como do processo de armazenamento de dados e terminar com a análise dos dados. Seguidamente, nos resultados, irão ser abordados aspetos mais específicos como o código do desenvolvido, como as "queries" efetuadas para a construção dos gráficos. Por último será feita uma explicação dos gráficos efetuados e que informações são possíveis extrair, assim como a discussão das dificuldades analíticas encontradas.

Keywords: covid19, Big Data, Dataset, WHO, kaggle, JupiterLab, Spark, Pipe-Line, Economia, PowerBI, Big-Data

1 Introdução

A finalidade deste artigo consiste no detalhadamento e documentação da concretização de todo o planeamento efetuado no trabalho anterior, portanto, a forma como tratar um conjunto de datasets selecionados de forma a conseguir um produto final capaz de ser facilmente usado para outras finalidades.

Contextualizando o assunto do trabalho, em meados de 2019 a humanidade teve que enfrentar uma nova pandemia devido ao aparecimento do vírus altamente contagioso intitulado covid-19. As consequências gerais de tal acontecimento foram sentidas por toda a população, sendo o interesse em conhecer e perceber algumas destas consequências o principal motivo de todo este trabalho, sendo elas, perceber como é que a progressão da pandemia afetou os diversos fatores/métricas da economia dos países, tendo em consideração também a preferência das pessoas em ficar em casa com o aumento do número de casos e mortes.

2 Arquitetura

Neste capitulo será apresentado todo o procedimento efetuado, desde a apresentação dos datasets, processos de ETL, processos de Data Storage e processos de análise.

2.1 Descrição dos DataSets

A equipa docente forneceu-nos a indicação a um dataset inicial em formato csv, originado no site oficial da organização mundial de saúde, relativo à evolução do número de casos de covid em diversos países ao redor do mundo, num período de aproximadamente dois anos, tendo inicio em janeiro de 2020. Este dataset tem como atributos a seguinte lista:

Table 1 Atributos do dataset covid mundial

Nome	Tipo	Descrição
Date Reported	Date	Data ao qual o registo foi realizado.
Country Code	String	Código ISO Alpha-2 associado ao país do registo.
Country	String	País, território, área onde o registo foi realizado.
WHO region	String	Estado membro da organização mundial de saúde a qual o país faz parte. No total existem: Africa (AFRO), Americas (AMRO), Ásia Oriental (SEARO), Europa (EURO), Mediterrânico Oriental (EMRO), Pacífico Ocidental (WPRO).
New Cases	Integer	Novos casos confirmados, calculado pela subtração dos casos cumulativos atuais pelas casos cumulativos do último registo do país.
Comulative Cases	Integer	Número de casos confirmados reportados para a organização mundial de saúde até a data no determinado país.
New Deaths	Integer	Novas mortes confirmadas, calculada pela subtração das mortes acumuladas atuais pelas mortes acumuladas do último registo do país.
Comulative Deaths	Integer	Número de mortes confirmadas reportadas para a organização mundial de saúde até a data no determinado país.

Após toda a intensiva pesquisa efetuada no planeamento, foi possível reunir mais dois datasets complementares, tendo em atenção que ambos devem ter registos associados a uma data e um país coincidentes com o dataset fornecido, ambos os datasets foram obtidos a partir da plataforma kaggle pelos seguintes links:

- https://www.kaggle.com/datasets/shashwatwork/impact-of-covid19-pandemic-on-the-global-economy?select=transformed_data.csv
- https://www.kaggle.com/datasets/aestheteaman01/people-staying-in-home-during-covid19?select=residential-stay-in-covid19.csv

Um deles, chamado transformed data, contém informações relativas à economia dos países com o surgimento da pandemia, portanto, para aproximadamente o mesmo intervalo de tempo do dataset inicialmente fornecido. Este dataset é constituído pelos seguintes atributos:

Table 2 Atributos do dataset economia mundial

Nome	Tipo	Descrição
Code	String	Código associado ao país onde foi realizado o registo
Country	String	País onde foi realizado o registo.
Date	Date	Data ao qual o registo foi realizado
HDI	Float	Medida comparativa usada para classificar os países pelo seu grau de desenvolvimento humano entre 0 e 1.
TC	Float	A soma de todos os custos na produção de um certo output.
TD	Float	Depósitos temporais realizadores por um país em todos os bancos,
STI	Float	Medida comparativa usada para classificar os países pela sua ciência, tecnologia, inovação e cultura no âmbito de desenvolvimento de um ambiente sustentável.
POP	Float	Área no qual os profissionais de marketing planeiam atividades promocionais em torno dos produtos.
GDPCAP	Float	Produção económica de uma nação por pessoa.

O último dataset é o residential-stay-in-covid19, que apesar de poucos atributos, contém informação relativa a variação da vontade da população em ficar em casa durante o período de pandemia em relação ao inicio desta, ou seja, um valor de 10 significa que há uma vontade 10 vezes maior de permanecer em casa em relação ao inicio da pandemia. É constituído pelos seguintes atributos:

Table 3 Atributos do dataset estadia residencial

Nome	Tipo	Descrição
Entity	String	País onde foi realizado o registo.
Day	Date	Data associada ao registo.
Increase In Residential Stay	Float	Aumento percentual da quantidade de pessoas que preferem ficar em casa relativo a situação anterior a pandemia.

2.1.1 Relacionamento com os 5 Vs do Big Data

- **Volume**: A sua junção não cria uma quantidade de dados capaz de ocupar gigabytes, no entanto, consideramos ser uma quantidade adequada.
- **Variedade**: Os datasets não tem muita variação nas fontes dos dados, uma vez que, ambos foram retirados de uma só fonte, o Kaggle, cujos links constam nas referências deste relatório.
- **Velocidade**: Uma vez que não temos uma quantidade de dados exorbitante, o seu processamento e tratamento tenderá a ser rápido.
- **Veracidade**: Uma vez que os dados foram extraídos de fontes fidedignas, tudo indica que sejam verídicos. Uma vez que não foi pedida a realização do processo de extração dos dados para a criação dos datasets.

- **Valor**: O valor gerado por estes dados pode ser incalculável para a sociedade, uma vez que perceber determinados aspetos, fará com que numa nova pandemia estejamos aptos para a combater economicamente, de forma rápida e precisa.

2.2 Processo de ETL

Todo o processamento e tratamento efetuado aos datasets, tal como previsto, foi feito recorrendo ao uso da ferramenta **JupiterLab** juntamente da linguagem **Python** e das suas bibliotecas **Pandas** e **Numpy**.

2.2.1 Junção dos DataSets

Neste capitulo serão abordadas as etapas feitas para a junção dos diferentes datasets num único, sendo apresentados todos os métodos aplicados para a resolução de cada um dos problemas.

Numa primeira fase importamos os datasets guardados em formato csv no sistema de arquivos local para o **JupiterLab**. De seguida, e antes de avançar para a fase de junção concretamente, achamos importante confirmar se tudo foi importado corretamente, portanto, efetuar uma breve visualização dos cabeçalhos de cada um dos datasets, de modo a confirmar se todos os atributos estava presentes e se os registos não desformataram por algum motivo.

Feito isso, prosseguimos para a fase de junção, onde imediatamente nos deparados com o possível surgimento de dois tipos de problemas diferentes: O primeiro problema e o mais óbvio, é o facto de algum dos datasets possuir menor quantidade de datas e ou países que, como estes são atributos a ser usados para fazer a junção, irão surgir problemas de compatibilidade. O segundo problema é o facto de atributos com o mesmo significado, terem formatos de valoração diferentes,por exemplo, países que são representados com nomes diferentes.

Para perceber a gravidade de cada um dos problemas, verificamos a quantidade de valores únicos, tanto para os atributos que representam os países como para os que representam as datas em cada um dos datasets importados.

Efetuado este procedimento, foi nos possível concluir que a solução para o primeiro problema passa por, uma vez que que o dataset relativo aos casos de covid-19 está muito completo face aos restantes dois, aplicar o método de junção "left" dos restantes datasets ao dataset mais completo, que por outras palavras, significa completar o dataset relativo aos casos de covid-19 com os dados contidos nos restantes datasets, garantindo assim que mantemos a maior quantidade de dados possíveis. Ao não perder informação relativa ao dataset fornecido permito-nos por exemplo calcular com confiança o numero de casos novos por ano mundialmente, na qual numa junção "inner" entre os datasets já não seria possível.

Relativamente ao segundo problema a solução para o caso das datas foi simplesmente garantir que todas as datas estavam no mesmo formato string sendo este Ano-Mês-Dia relativo ao dataset covid, transformando se necessário

usando a biblioteca pandas o atributo para o tipo data e de seguida converter para um string num formato novo correto.

Já para o caso dos países,este envolve uma abordagem muito mais complexa, pelo qual das 2 opções que equipa pensou, uma delas envolvia a adição e junção de outro dataset capaz de relacionar o nome dos países com os seus respetivos códigos, a outra envolvia o cálculos dos valores que ambos os datasets não têm em comum, ou seja, aqueles onde não é capaz de realizar uma junção, e procurar manualmente dessa lista designações diferentes para o mesmo país, alterando num dos datasets o nome dos seus países para equivaler ao outro.

A adotada foi a segunda abordagem, ao qual, mesmo tendo as suas desvantagens permitiu não perder informação relativa a países considerados importantes para o caso de estudo.

Durante a resolução das seguintes etapas foi detetado mais um problema que exigiu alterações no processo de ETL, foi detetado um enorme aumento de GDP mundial num curto período de tempo. Este aumento originou-se pelo facto de que dependendo da data existem mais ou menos países com o seu GDP registado nessa data, por exemplo, num determinado período existem uma dezena de países que não tem o seu GDP registado no dataset, ficando este a null, fazendo com que a soma do GDP de todos os países varia-se muito dependendo da data. Para combater este problema a equipa originou uma nova coluna **GDPCAP_T**, não para substituir a coluna original de GDP mas para fornecer uma alternativa na construção de gráficos que envolvem o cálculos de GDP mundialmente. Nessa nova coluna houve um tratamento de valores nulos no intervalo de datas do dataset de economia, o tratamento foi feito via interpolação pelo facto dos registos de um mesmo país se encontrarem agrupados mas envolve certos erros, sendo estes erros a justificação pela qual não se eliminou a coluna original.

Por fim, para além de serem eliminadas certas colunas não necessárias para o caso de estudo como os códigos e regiões de um país também foram eliminadas colunas com valor repetido de país e data provenientes da junção dos 3 datasets, adicionalmente de modo a facilitar a escrita das colunas no futuro redefinimos o nome de alguns atributos de forma a tornámos menos verbosos.

2.3 Data Storage

Completada a fase anterior, o próximo passo de acordo com o planeado, é guardar os dados, ou seja o dataset resultante da fusão no sistema de arquivos. Para tal, recorrendo à biblioteca do **Fastparquet** do **Python**, bastou um simples comando para assim gerar um ficheiro em formato parquet com todo o conteúdo do dataset resultante.

Não foi necessária a utilização de ferramentas com armazenamentos distribuídos como o **Apache Hadoop** para guardar o dataset resultante da fase anterior uma vez que tal como o previsto inicialmente, não existe uma quantidade de dados na ordem dos gigabytes que assim o justifique, no

entanto reconhece-se que a adição desta etapa não envolve grande esforço ou modificações por parte da equipa sendo somente necessário a instalação da ferramenta na maquina, e a modificações de dois comandos relativos ao salvamento e carregamento do dataset.

2.4 Analytical Data Store

Feita a fase anterior, segue-se a preparação e o processamento dos dados num formato mais legível e especifico, iremos então recorrer à ferramenta **Pyspark** continuando no ambiente **JupiterLab**.

Inicialmente foi preciso estabelecer uma conexão ao **Spark**. Feito isso bastou ler o parquet gerado na fase anterior para ficar com todo o dataset pronto a ser usado pelo **Pyspark**.

Numa primeira fase ao observar quais os tipos de dados inferidos pelo **Pyspark** com a leitura do parquet reparamos que o atributo correspondente à data constava no tipo string, pelo que, para que nos processos seguintes possam ser usados métodos de manipulação de datas mais facilmente, convertemos imediatamente para um tipo corresponde à data.

De seguida procedemos à realização das queries , cujos respetivos resultados servirão de input para a realização dos gráficos na fase seguinte. De cada query vai resultar um ficheiro csv que na fase seguinte será importado para o **PowerBI** para assim criar o gráfico correspondente.

É importante realçar a importância do **Pyspark** nesta fase, uma vez que é uma ferramenta facilmente utilizável, com grande adaptabilidade e potencial.

2.5 Analytics and Reporting

Por fim, foi usada a ferramenta analítica **PowerBI** para visualizar os dados através de uma interface iterativa. Os métodos usados para a visualização dos dados foram principalmente gráficos.

Nesta fase surgiram problemas de incompatibilidade no formato de certos valores, como por exemplo, valores do tipo float com o separado "," ao invés de ".", no entanto, facilmente contornamos o problema.

3 Resultados

Neste capitulo serão abordados os resultados obtidos e código desenvolvido após realizados todos os processos descritos anteriormente, mais propriamente, conteúdo especifico e técnico.

3.1 Extract Transform Load

O seguinte código e imagem demonstra o processo realizado para reconhecer e procurar estratégias relativas ao problema de cada dataset conter diferentes variedades de países e datas, resolvendo-se, como dito anteriormente, á utilização de um "Left Join" devido a maior variedade de valores no dataset relativo aos casos de covid-19 como se pode observar pela imagem.

```
print("paises covid:",covidCasos['Country'].nunique())
print("paises economia:",economia['COUNTRY'].nunique())
print("paises permanencia:",permanencia['Entity'].nunique())

print("datas covid:",covidCasos['Date_reported'].nunique())
print("datas economia:",economia['DATE'].nunique())
print("datas permanencia:",permanencia['Day'].nunique())
```

```
paises covid: 237
paises economia: 210
paises permanencia: 129
datas covid: 790
datas economia: 294
datas permanencia: 714
```

O seguinte código demonstra o processo realizado para combater a situação da presença de diferentes formatos na data.

```
#Transformar a string data no tipo datatime
permanencia['Day'] = pd.to_datetime(permanencia.Day,format='%d-%m-%Y')
#Transformar datatime na string com o formato certo
permanencia['Day'] = permanencia['Day'].dt.strftime('%Y-%m-%d')
```

A implementação da estratégia mencionada anteriormente que visa a combater as diferentes denominações do mesmo país em cada dataset, originou o seguinte código:

```
#(A \/ B), a soma de todos os paises em ambos os datasets
union = pd.Series(np.union1d(covidCasos['Country'],
    economia['COUNTRY']))

# (A /\ B), os paises em comum de ambos os datasets
intersect = pd.Series(np.intersect1d(covidCasos['Country'],
    economia['COUNTRY']))

# (A \/ B) - (A /\ B) , os pases nao em comum
notcommonseries = union[~union.isin(intersect)]

print(notcommonseries.tolist())
```

```
# Substituio  no dataset covid dos valores "The United Kingdom" para
    "United Kingdom"
covidCasos.loc[covidCasos['Country'] == 'The United Kingdom',
    ['Country']] = 'United Kingdom'
```

Relativo ao último problema mencionado, o seguinte código demonstra então a sua resolução, pelo qual consegue obter um novo dataset resultante do merge do dataset de economia e covid mas com uma nova coluna contendo o tratamento de valores nulos do atributo GDPCAP. É importante mencionar que o tratamento dos valores nulos é realizado no dataset resultante de fazer a junção dos dois datasets, e não somente ao dataset original da economia.

```
#Restringir as datas as dos valores registados no dataset de economia
novoEconomia=casosEconomia.where((casosEconomia['Date_reported'] >=
    '2020-01-03') & (casosEconomia['Date_reported'] <= '2020-10-19'))

#Criacao de nova coluna com interpolao para tratar valores nulos
novoEconomia['GDPCAP_T']=novoEconomia['GDPCAP']
novoEconomia['GDPCAP_T'].interpolate(method = 'linear',inplace=True)

#Filtrar colunas somente associadas ao dataset economia
novoEconomia = novoEconomia.filter(items=['Date_reported',
    'Country','GDPCAP','GDPCAP_T','POP','HDI','TC','TD','STI'])
novoEconomia['Date_reported'] =
    novoEconomia['Date_reported'].dt.strftime('%Y-%m-%d')

#Fazer novamente merge
casosEconomia=pd.merge(covidCasos,novoEconomia,how='left',
left_on=['Date_reported','Country'],right_on=['Date_reported','Country'])
```

Como se pode observar pela seguinte figura referente a quantidade de valores nulos em cada atributo do dataset final, a realização da técnica "Left Join", levou a presença de uma grande quantidade de valores em falta, que exigira uma grande atenção por parte da equipa na fase respetiva a realização das queries.

Porem, podemos observar que os atributos referentes ao estado da pandemia se encontram totalmente completos, assim como também se pode ver o resultado da realização de tratamento de valores nulos na coluna **GDPCAP_T** em relação a coluna **GDPCAP**

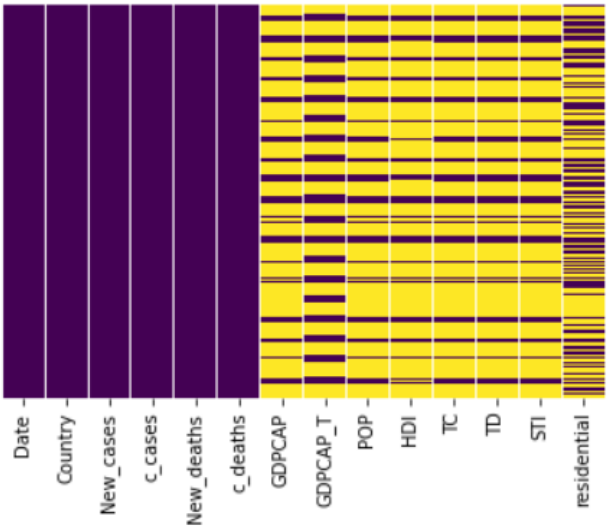

Fig. 1 HeatMap de valores nulos do dataset Final

3.2 Queries desenvolvidas

Nesta secção será apresentada uma descrição, explicação e justificação do objetivo de todas as queries realizadas, assim como, a apresentação do código desenvolvido.

3.2.1 Casos por Ano

Na **primeira** query precisamos de determinar o número de casos de covid por ano no mundo inteiro. Para procedermos na sua construção, primeiro foi necessário acrescentar um atributo chamado "ano" ao dataset, correspondente à data de cada registo, de seguida, foi preciso depois agrupar os registos feitos em cada ano, conseguindo assim obter a soma dos novos casos registados nesse ano, adicionalmente deu-se um novo nome a coluna resultante.

```
gf1 = df_pyspark.withColumn('year',year(df_pyspark.date))
gf1 = gf1.groupBy('year').agg({'New_cases':'sum'})
gf1 = gf1.withColumnRenamed("sum(New_cases)", "cases")
gf1.toPandas().to_csv('Graficos/gf1.csv')
```

3.2.2 Evolução do GDP mundial

Na **segunda** query precisamos de determinar a evolução do GDPCAP mundial, em relação ao primeiro valor detetado. Para tal, primeiramente filtramos os registos onde o GDP não é nulo e, seguidamente, agrupámos os registos que têm a mesma data, somando o GDP respetivo, de forma a conseguirmos o GDP para cada data. Posteriormente ordenou-se os resultados pela data de forma ascendente e obtivemos o valor do GDP da primeira data, removendo este a todos os valores da coluna GDP.

```
gf2=df_pyspark.filter(df_pyspark.GDPCAP_T.isNotNull()).groupBy('date')
.agg({'GDPCAP_T':'sum'}).sort('date', ascending=True)
first = (gf2.head())[1]
gf2 = gf2.select('date',gf2[1] - first)
gf2 = gf2.withColumnRenamed(gf2.columns[1], "GDPCAP_Diff")
```

3.2.3 Países com mais casos acumulados por ano

A **terceira** e **quarta** query são muito semelhantes, pois os respetivos objetivos são, os países com mais casos cumulativos, no final do ano 2020 e no final do ano 2021. Para tal, bastou-nos filtrar os registos pela data correspondente ao final do ano 2020 e 2021 e ordenar pelo número de casos, limitando depois as linhas resultantes.

```
gf3=df_pyspark.where(df_pyspark.date=='2020-12-31')
.select('Country','c_cases').sort('c_cases',ascending=False).limit(10)
```

```
gf4=df_pyspark.where(df_pyspark.date=='2021-12-31')
.select('Country','c_cases').sort('c_cases',ascending=False).limit(10)
```

3.2.4 Distribuição do GDP por países

A **quinta** query consiste em obter o valor do GDP para cada país para uma determinada data aleatória,no entanto, como existem muitos países foi decidido assim como costume na realização deste tipo de grafos, agrupar os resultados finais numa única fatia. Posto isto, primeiramente, filtrámos os registos com a respetiva data igual à pretendida e ordenámos pelo o GDP. De seguida, replica-mos este dataset, limitando o original a somente 15 países, após retirar da cópia esses 15 países, esta foi agrupada e transformada numa única nova entrada que foi adicionada ao original com o nome de "Outros". É importante de salientar que este tipo de operações semi-complexas só conseguem ser feita de forma bastante facilitada devido as capacidades da ferramenta **PySpark**

```
gf5=df_pyspark.where(df_pyspark.date=='2020-09-01')
.select('Country','GDPCAP').sort('GDPCAP',ascending=False)
temp = gf5
gf5 = gf5.limit(15)
temp=temp.subtract(gf5)
temp=temp.agg({'GDPCAP':'sum'}).head()
value=temp[0]
newRow = spark.createDataFrame([('Outros',value)],['Country','GDPCAP'])
gf5=gf5.union(newRow)
```

3.2.5 Evolução da pandemia

A **sexta** query consiste apenas em obter para cada data a soma dos novos casos registados mundialmente, para tal foi simplesmente feito um agrupamento pela data, seguido de uma ordenação por esta mesma.

```
gf6=df_pyspark.groupBy('date').agg({'New_cases':'sum'})
.sort('date',ascending=True)
```

3.2.6 Impacto da pandemia na economia em Portugal e no Mundo

A **sétima** e **oitava** query são muito semelhantes, sendo que a primeira consiste em obter dados relativos concretamente a Portugal e a segunda a nível mundial,e sses dados sendo a evolução do GDP e casos acumulados ao longo do tempo. As técnicas utilizadas não diferenciam das outras queries, sendo necessário a frequente restrição de valores não nulos, agrupamento por data, ordenações ou seleções para restringir as colunas oferecidas como resultado.

```
gf7=df_pyspark.filter(df_pyspark.GDPCAP.isNotNull())
```

```
.where(df_pyspark.Country=='Portugal').select('date','c_cases','GDPCAP')

gf8=df_pyspark.filter(df_pyspark.GDPCAP_T.isNotNull()).groupBy('date')
.agg({'c_cases':'sum','GDPCAP_T':'sum'}).sort('date',ascending=True)
```

3.2.7 Quarentena em Portugal

A **nona** query consiste em obter dados relativos a Portugal, portanto, obter o número de novos casos assim como a medida de residência para cada data. Para tal foi apresentada as colunas data, novos casos e residência no dataset resultante da filtração dos registos de Portugal em que o valor de residência não é nulo.

```
gf9=df_pyspark.filter(df_pyspark.residential.isNotNull())
.where(df_pyspark.Country=='Portugal')
.select('date','New_cases','residential')
```

3.2.8 Evolução da pandemia e economia em 6 países diferentes

A **décima** query consiste em obter o número de casos acumulados assim como o GDPCAP para cada mês em seis países escolhidos, relativamente ao ano 2020. Para isso, primeiramente, filtrámos os registos dos países em estudo, seguidamente, filtrámos os registos ao ano 2020 e criámos um novo atributo que guarda o mês dos registos, por fim, fizemos a agregação dos registos por país e mês, calculando assim o máximo de casos acumulados e o GDP máximo nesse mês, naquele país. No final ordenámos os registos pelo mês.

```
gf10=df_pyspark.filter(df_pyspark.GDPCAP.isNotNull())
gf10=gf10.where(gf10.Country.isin(['United
    States','China','Spain','Ukraine','Qatar','India']))
gf10=gf10.where(year(gf10.date)==2020).withColumn('month',month(gf10.date))
gf10=gf10.groupBy('Country','month').agg({'c_cases':'max','GDPCAP':'max'})
.sort('month',ascending=True)
```

3.2.9 Impacto do índice de desenvolvimento humano na pandemia

A **décima primeira** query consiste em obter para cada país, os casos acumulados e o seu grau de desenvolvimento numa determinada data, tentando então obter uma relação para estes. No entanto como existem bastantes países foi utilizada novamente outra funcionalidade do **PySpark** para retirar uma amostra do dataset original.

```
gf11=df_pyspark.filter(df_pyspark.HDI.isNotNull())
.where(df_pyspark.date=='2020-10-01').select('Country','c_cases','HDI')
gf11 = gf11.sample(False, 0.1, seed=20)
```

3.3 Gráficos desenvolvidos

De seguida irão ser apresentados todos os gráficos desenvolvidos, juntamente com uma breve descrição da apresentação do seu tipo.

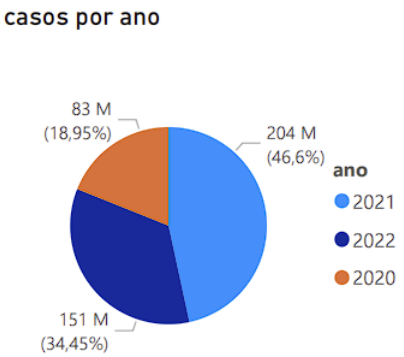

Fig. 2 Gráfico de fatias para análise do número de casos.

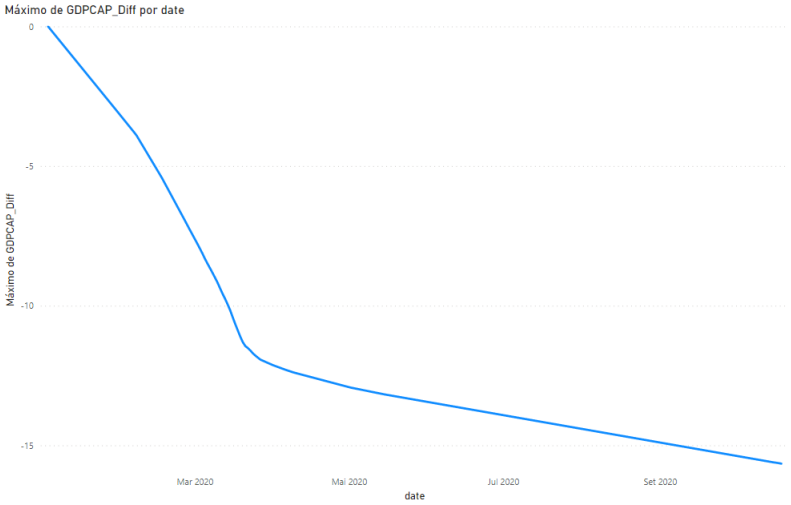

Fig. 3 Gráfico de barras para análise da distribuição de casos pelos top 10 países em 2020.

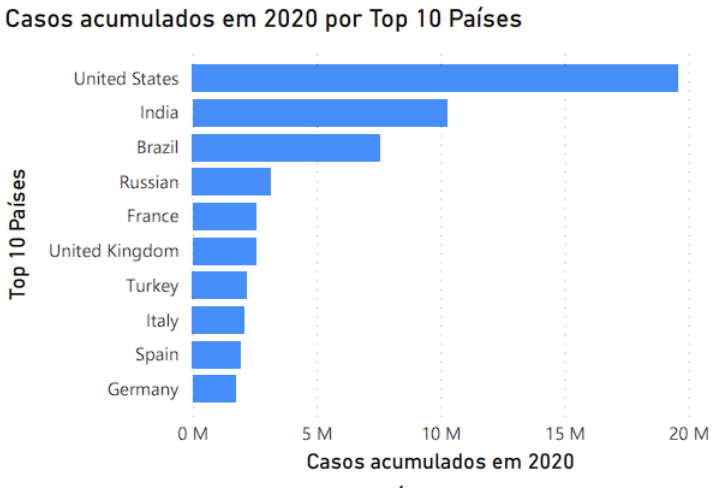

Fig. 4 Gráfico de barras para análise da distribuição de casos pelos top 10 países em 2021.

Fig. 5 Gráfico para análise da variação o GDPCAP.

Country	GDPCAP
Outros	1.378,73
Qatar	11,67
Luxembourg	11,45
Singapore	11,36
Ireland	11,12
United Arab Emirates	11,12
Kuwait	11,09
Norway	11,08
Switzerland	10,96
San Marino	10,95
United States	10,90
Bermuda	10,83
Cayman Islands	10,82
Saudi Arabia	10,80
Netherlands	10,79
Sweden	10,76
Total	**1.544,42**

Fig. 6 Gráfico para análise do GDPCAP em cada país.

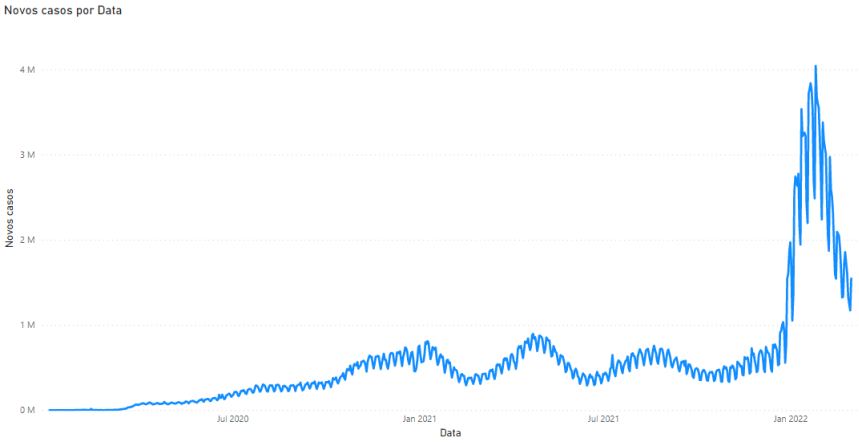

Fig. 7 Gráfico para análise do aumento do número de casos no mundo.

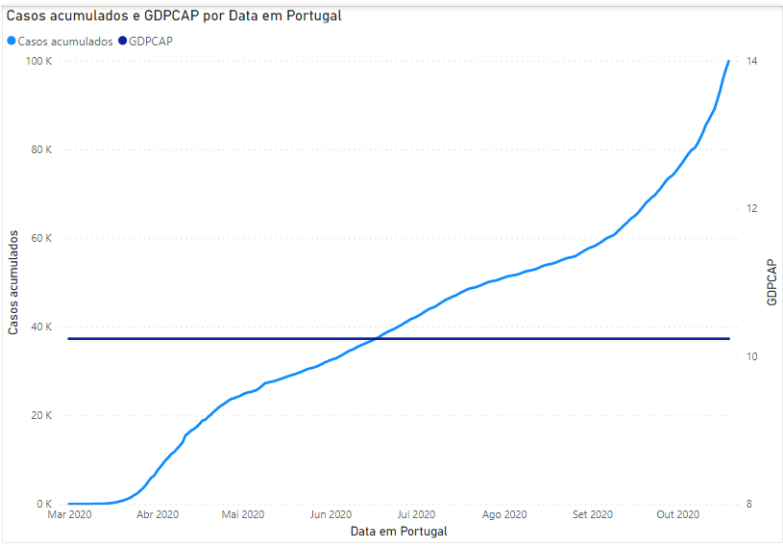

Fig. 8 Gráfico para análise da relação entre o número de casos acumulados e o GDPCAP em Portugal.

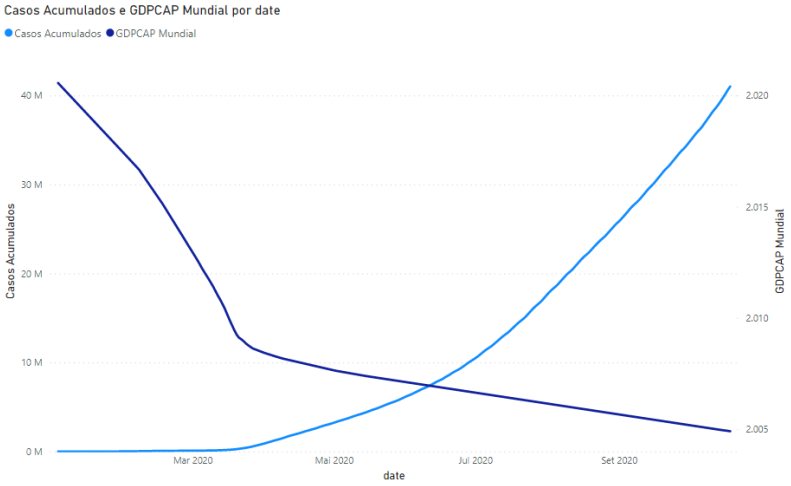

Fig. 9 Gráfico para análise da relação entre o número de casos acumulados e o GDPCAP no mundo.

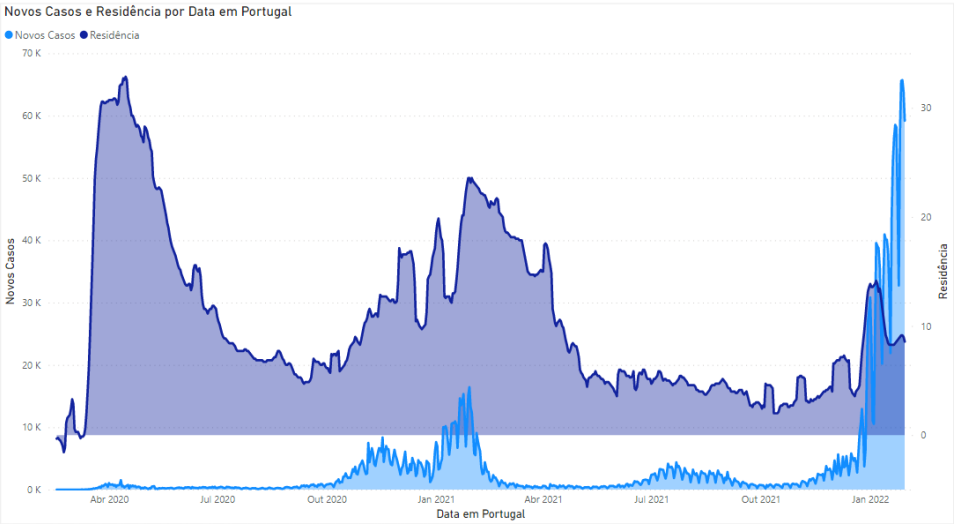

Fig. 10 Gráfico para análise da relação entre o aumento do número de novos casos e o índice de residência (preferência em permanecer em casa).

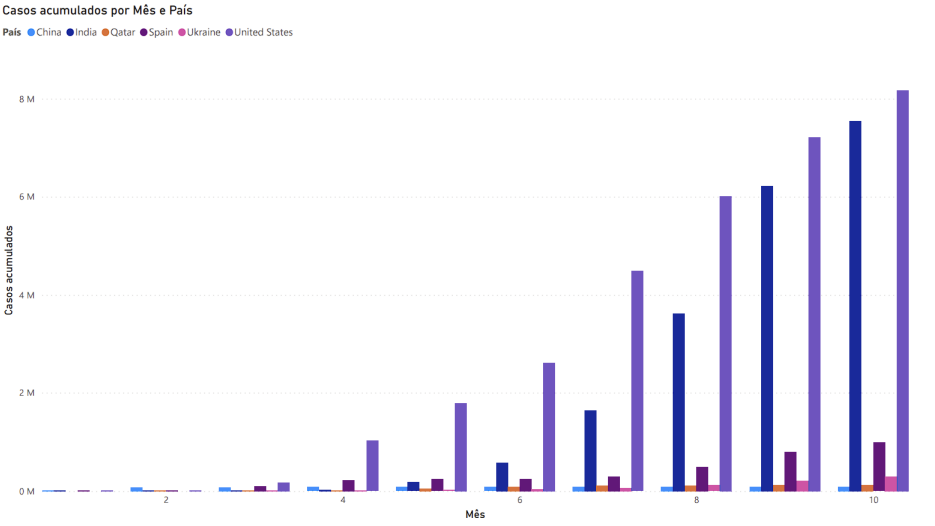

Fig. 11 Gráfico para análise do número de casos acumulados em cada mês, em cada país.

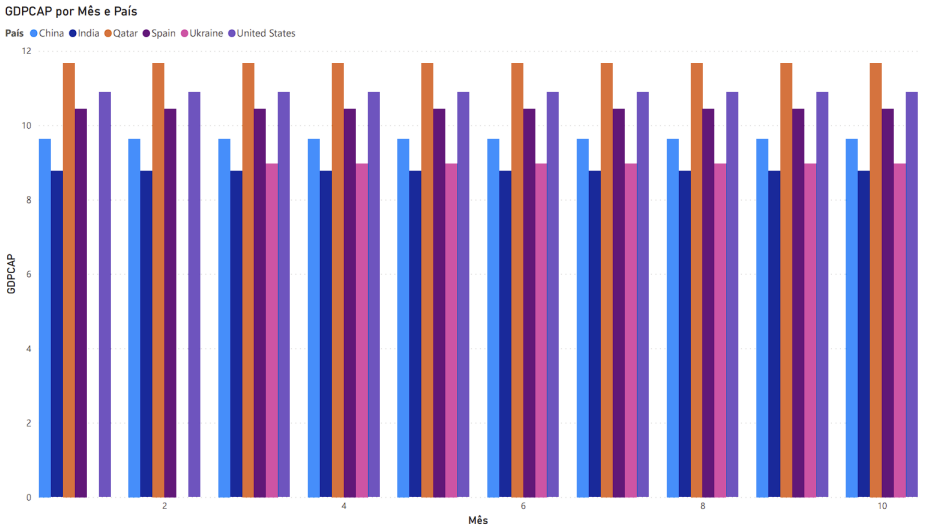

Fig. 12 Gráfico para análise do GDPCAP por mês, em cada país.

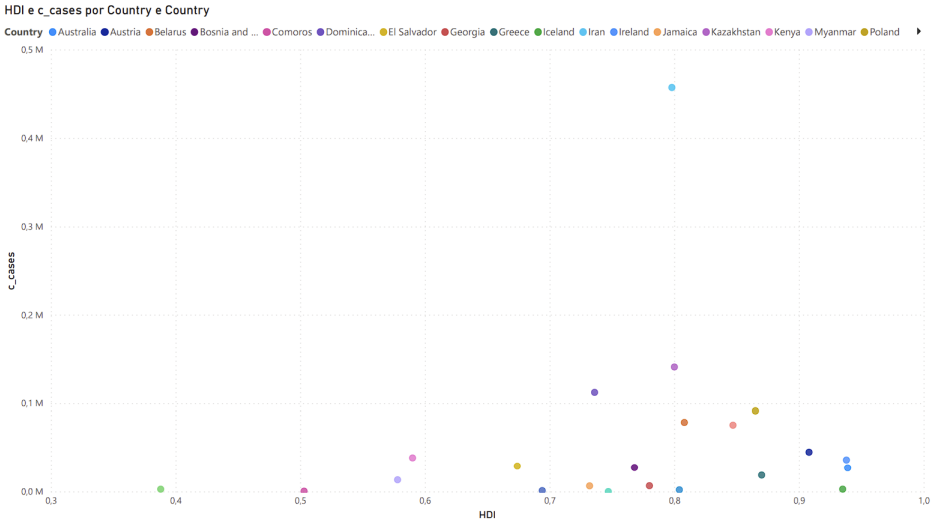

Fig. 13 Gráfico para análise da relação entre o HDI e o número de casos acumulados por país.

4 Discussão

4.1 Gráficos

Nesta fase do relatório iremos falar do conhecimento que a equipa conseguiu retirar dos gráficos realizados e que conclusões se podem tirar destes.

- **Casos por ano**
  Do primeiro gráfico pode-se deduzir que já foram registados mais de 400 milhões de casos e estes vem aumentando drasticamente por ano sendo que nem metade do ano 2022 representa quase 50% desses casos.

- **Evolução do GDP mundial**
  Como previsto, houve um impacto negativo na economia mundial num dos anos mais impactados pela pandemia 2020,ou seja, uma descida no GDP mundial, no entanto, é importante salientar que os valores representados são a soma dos GDP de todos os países e não a média, pois individualmente, cada país não desceu mais do que umas meras décimas.

- **Países com mais casos acumulados**
  Apesar de haverem pequenas alterações nos países com mais casos registados ao longo do ano de 2021, alguns se destacam como USA, Índia e Brasil, principalmente, devido a sua grande população,no entanto, convém mencionar que ter mais registos não significa que é o país com mais infetados, visto que, é provável haverem muitos países com casos não registados pela organização mundial de saúde.

- **Distribuição do GDP por países**
  Pouco se pode retirar sobre o caso de estudo a partir deste gráfico, sendo mais realizado para testar as técnicas que o PySpark nos permite efetuar, no entanto, serve para observar alguns dos países com maior GDP por população.

- **Evolução da pandemia**
  Este gráfico permito-nos observar globalmente a evolução da pandemia, pelo que se pode deduzir quais as épocas com mais novos casos registados. Dessas épocas destaca-se o mês de Janeiro de 2022 coincidente ao levantamento das restrições da quarentena mesmo já com a presença da vacina.

- **Impacto da pandemia em Portugal e no mundo**
  A medida que o numero de casos acumulados aumentou ,Portugal manteve o seu GDP constante o que foi bom para todos os portugueses que como podemos ver pelo outro gráfico, globalmente não foi o caso. Adicionalmente, se pode observar que Portugal teve uma evolução ou propagação de casos acumulados de covid-19 um pouco semelhante a registada pelo mundo como um todo.

- **Quarentena em Portugal**

  A equipa acha que este gráfico é um dos mais importantes realizados ao qual se notou uma grande correlação entre o numero de casos e a preferência da população ficar nas suas residências demonstrando um pouco a veracidade de ambos os datasets. Deste gráfico pode-se retirar bastante informação. No inicio da pandemia apesar dos casos serem poucos, as pessoas preferiam ficar em casa, provavelmente pelo impacto de uma nova pandemia e não tendo ideia do intervalo que esta iria afetar as suas vidas diárias. Durante alguns meses aos quais a vontade de novamente sair aumentou constantemente finalmente foram retiradas algumas medidas de quarentena aumentando imediatamente o numero de casos no intervalo de Outubro á Fevereiro e consequentemente o retorno da preferência da população permanecer nas suas residências. A partir de Abrir os casos diminuíram bastante, provavelmente o efeito das vacinas e as pessoas preferiram novamente não estar em casa. No inicio de 2022 os casos aumentaram bruscamente principalmente por causa do final das restrições da quarentena.

- **Evolução da pandemia em 6 países**

  Dos 6 países escolhidos não se notou uma variação nos GDP de cada, no entanto notou-se uma clara distinção na evolução de casos acumulados. A China, por exemplo, um país com uma enorme população,apesar de ter começado com vários casos registados não aumentou muito no decorrer do ano 2020. Pode haver várias razões para tal acontecer, poderá ser pela eficiência das medidas de quarentena do país ou simplesmente o numero de casos não foi divulgado, ou as pessoas não mostraram necessidade de os registar, isto leva-nos a verificar que a data usada é sempre restrita e influenciada por inúmeros fatores políticos, sócio-económicos, culturais ,entre outros, que o analista não tem controlo sobre, e sempre deve-se observa-los com olho critico.

- **Impacto do ínice de desenvolvimento humano**

  Neste gráfico o objetivo seria tentar observar a correlação entre o grau de desenvolvimento de um país e os seus casos acumulados, em principio, lógica indica que os países menos desenvolvidos terão mais casos pois não tem a facilitação de ficar em casa, no entanto, esta lógica não leva em atenção o numero de população por pais, ou o facto de países mais desenvolvidos terem maior capacidade de registar, demonstrando assim resultados contrários aos esperados.

4.2 Dificuldades de análise

Algumas das grandes dificuldades sentidas devem-se principalmente ao desconhecimento do domínio do problema. Este desconhecimento, juntamente com a da falta de experiência em análise de dados geram diversos tipos de problemas tanto nas fases inicias, onde se tenta perceber o significado de cada atributo, assim como perceber as suas relações, mas também nas fases finais, onde normalmente se decide que dados recolher do dataset para construir os respetivos gráficos ou como fazer a análise dos resultados obtidos.

Nas fases intermédias, relacionadas a parte de programação, uma vez que apenas é necessário instalar e utilizar as ferramentas necessárias ou tratar e corrigir algum determinado atributo ou conjunto de dados, estas não geram para a equipa dificuldades comparáveis ás apresentadas anteriormente.

5 Conclusão

A realização deste trabalho permitiu expandir o nosso conhecimento sobre as áreas relativas a "Big Data", com especial ênfase à utilidade das ferramentas já existentes na ajuda do tratamento e extração de conhecimento de "Big Data".

Em conclusão, pensamos ter conseguido cumprir todo o planeamento anteriormente efetuado para tratamento do conjunto de datasets encontrados de forma eficaz, assim como extrair a toda a informação útil pretendida dos mesmos. Tudo isso recorrendo a ferramentas investigadas selecionadas, que demonstraram um enorme potencial.

Em relação ao problema em questão, pensamos ter observado relações óbvias sobre os temas da pandemia, economia e quarentena, e achamos ter retirado com sucesso informação pertinente demonstrando esta mesma via uma apresentação útil, eficaz e valiosa.

Em projetos futuros estaremos consideravelmente à vontade para trabalhar novamente com as ferramentas utilizadas, o que nos proporcionará maior produtividade devido à experiência ganha com este trabalho.