Impacto da Pandemia COVID-19 na Preferência *Stay at home*

Filipa Pereira[1*], Luís Pinto[1*], Luísa Carneiro[1*] and Rita Peixoto[1*]

[1*]Departamento de Informática, Universidade do Minho, Braga, Portugal.

*Corresponding author(s). E-mail(s): pg46978@alunos.uminho.pt; pg47428@alunos.uminho.pt; pg46983@alunos.uminho.pt; pg46988@alunos.uminho.pt;

Abstract

Desde 2020 que a pandemia COVID-19 tem tido uma grande influência em todos os fatores da sociedade. Um dos fatores mais influenciados pelo COVID-19, foi a preferência das pessoas ficarem em casa (*stay at home*) relativamente ao período pré-pandémico, devido à implementação restrições que condicionaram a mobilidade e o ajuntamento de pessoas. Com este documento pretende-se sensibilizar o leitor acerca das diferentes ferramentas de *big data* existentes e de como podemos aplicá-las no caso prático. Deste modo, iremos começar com uma breve introdução apresentando em maior detalhe o *use case*. De seguida, será apresentado o estado da arte em *big data*, bem como dado a conhecer o sistema desenvolvido e a arquitetura adotada. Posteriormente, na secção de resultados, existirá uma explicação dos resultados obtidos, a qual será acompanhada por análise crítica dos mesmos na secção de discussão. Finalmente, chegamos a conclusão do artigo onde serão apresentadas as considerações finais do mesmo.

Keywords: *big data*, COVID-19, Python, Apache Spark, PowerBi, MongoDB *Batch Pipeline*, *Dashboards*, *ETL*, *Data Analytics*, *Big Data Tools*

1 Introdução

Nestes últimos anos, a COVID-19 influenciou vários aspetos da sociedade. O crescente número de casos e de mortes diários, levou a que as pessoas não se sentissem à vontade em sair de casa sem máscara ou a estarem em espaços com grandes ajuntamentos, correndo o risco de ficarem infetadas (principalmente quando a vacinação não era uma opção). Por outro lado, devido ás novas restrições em vigor, viam-se obrigadas a ficarem confinadas ou a terem os seus eventos públicos cancelados. Desta forma, o clima de insegurança e instabilidade sentido, levou a que parte da população se sentisse mais confortável no seu ambiente domiciliar, manifestando, assim, a sua preferência em ficar em casa mesmo quando as restrições foram aligeiradas.

Por conseguinte, o presente artigo foca-se no estudo do impacto da COVID-19 na vontade de as pessoas em ficarem em casa, procurando encontrar métricas que expliquem e avaliem a relação entre ambos os temas, através do uso de diversas ferramentas de *big data*. Desta forma, pretendemos avaliar como é que o número de contaminados e mortes por COVID-19, assim como o processo da vacinação influenciaram a perspetiva das pessoas em ficar em casa. Além disso, foi também avaliada de que forma as restrições que limitam os ajuntamentos de pessoas e restringem as saídas de casa (restrições *stay at home*) condicionaram a preferência da população em ficar na sua residência.

2 Estado da Arte

Nesta secção culmina a pesquisa das principais tecnologias associadas à área de *Big Data*. Cada uma das ferramentas possui, inerentemente, as suas características tornando-se assim necessário efetuar uma análise prévia das tecnologias que servirão de suporte para a arquitetura a ser implementada. A decisão da arquitetura a adotar teve em conta os artigos individuais desenvolvidos por cada um dos elementos do grupo, tendo sido comparadas as várias abordagens e chegando-se a um entendimento das ferramentas a adotar de modo a fornecer uma abordagem capaz de suportar o problema em mãos. Uma vez escolhidas as tecnologias, o processo de criação de um *pipeline* surgiu naturalmente, este será apresentado em maior detalhe na secção seguinte.

O presente documento também utilizou os temas sobre o Covid-19 e a influência que a pandemia teve em vários fatores da sociedade apresentados nos artigos individuais. Desta forma, o artigo expande o estudo prévio tanto do tema como das ferramentas de *big data*, materializando o mesmo numa solução de arquitetura para o *use case* definido.

3 Arquitetura

Para responder ao use case apresentado foi implementada uma arquitetura que é descrita segundo 4 fases. Na primeira fase vamos **unir os vários *datasets*** utilizados e na segunda fase vamos **armazenar o *dataset* obtido** numa base de dados. Na terceira e quarta fase vamos fazer o **tratamento dos dados** e a **visualização** destes, respetivamente. Na figura abaixo temos a estrutura do pipeline que foi implementado.

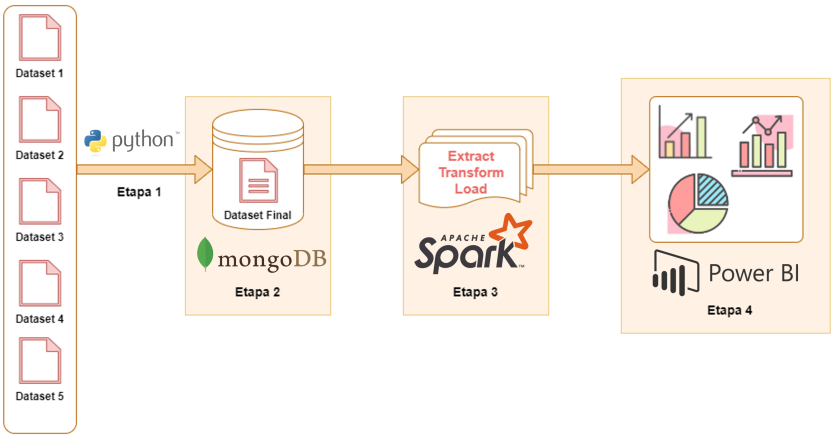

Fig. 1 *Etapas e ferramentas do Pipeline desenvolvido*

3.1 Análise dos dados

Para responder corretamente ao caso de uso apresentado foram utilizados cerca de 5 *datasets*, sendo que 1 deles foi fornecidos pelos docentes e os restantes foram encontrados nos *websites* listados na secção referências. É de notar que todos os *datasets* estão em formato CSV exceto o *dataset* da evolução da vacinação que se encontra em formato JSON e que apesar das observações serem realizadas em períodos de tempo distintos para cada *dataset*, nem sempre os países têm dados associados a todos os dias dentro desse período.

3.1.1 *Dataset* 1 - *WHO Covid*

Dataset fornecido pelos docentes que representa os dados globais associados à pandemia COVID-19, desde 3 de janeiro de 2020 até 1 de março de 2022, em cada um dos 236 países. Estes dados fazem de base para os restantes *datasets*. Para além dos atributos que representam o país, o código do país e a data, este *dataset* tem os seguintes atributos:

- ***New_cases***: número de novos casos de COVID-19, por dia.

- ***Cumalative_cases***: número casos de COVID-19 acumulados ao longo dos dias.
- ***New_deaths***: número de novas mortes por COVID-19, por dia.
- ***Cumalative_deaths***: número de mortes por COVID-19 acumulados ao longo dos dias.

3.1.2 *Dataset* 2 - *Vaccination*[1]:

Este *dataset* representa a evolução do processo de vacinação desde 22 de janeiro de 2021 até 27 de março de 2022 em 235 países. De forma a representar esse processo de vacinação por país foram utilizadas algumas métricas e atributos, contudo é de notar que nem todas as observações possuem valores para todos estes parâmetros. Neste *dataset* existem cerca de 14 atributos representativos do processo de vacinação, no entanto a evolução da vacinação pode ser analisada através das seguintes métricas, sendo que as restantes servem para completar cada observação.

- ***country***: país.
- ***iso_code***: código do país.
- ***date***: dia da observação.
- ***total_vaccinations***: número total de doses administradas. Para vacinas que requerem doses múltiplas, cada dose individual é contada.
- ***daily_vaccinations***: novas doses administradas por dia.
- ***people_vaccinated***: número total de pessoas que receberam pelo menos uma dose de vacina.
- ***people_fully_vaccinated***: número total de pessoas que receberam todas as doses prescritas pelo protocolo de vacinação inicial, isto é, 2 ou 1 dose dependendo da vacina administrada.
- ***total_boosters***: número total de doses de reforço administradas.
- ***daily_people_vaccinado***: número diário de pessoas que receberam a primeira dose da vacina.

3.1.3 *Dataset* 3 - *Restrictions on gatherings*[2]:

Neste *dataset* encontra-se associada a cada dia (1 de janeiro de 2020 a 21 de março de 2022) e a cada um dos 186 países, as medidas aplicadas aos ajuntamentos de pessoas. Desta forma, o *dataset* é constituído por 4 colunas onde a primeira representa o país, a segunda representa o código do país, a terceira o dia e a quarta as restrições aplicadas (***restriction_gatherings***). Estas restrições são caracterizadas por valores de 0 a 4, sendo que:

- **0** - inexistência de medidas;
- **1** - limita os ajuntamentos a um número maior que 1000 pessoas;
- **2** - restringe os ajuntamentos entre 100 a 1000 pessoas;
- **3** - permite ajuntamentos de 10 a 100 pessoas;
- **4** - os ajuntamentos só podem ter menos de 10 pessoas.

3.1.4 *Dataset* 4 - *Stay at home*[3]:

Neste *dataset* encontra-se associada a cada dia (1 de janeiro de 2020 a 21 de março de 2022) e a cada um dos 186 países, as medidas *stay at home* (recomendação de ficar em casa). Desta forma, o *dataset* é constituído por 4 colunas onde a primeira representa o país, a segunda representa o código do país, a terceira o dia e a quarta as restrições aplicadas (***stay_home_requirements***). Estas restrições são caracterizadas por valores de 0 a 3, sendo que:

- **0** - inexistência de medidas;
- **1** - recomendação de não sair de casa;
- **2** - obrigatoriedade de não sair de casa exceto para exercício diário, compras de mercearia ou saídas essenciais;
- **3** - obrigatoriedade de não sair de casa com exceções mínimas (apenas 1 membro da família pode sair, entre outros)

3.1.5 *Dataset* 5 - *Residential Mobility*[4]:

Neste *dataset* encontra-se associada a cada dia (17 de fevereiro de 2020 a 31 de janeiro de 2022) e a cada um dos 129 países, uma métrica que avalia a vontade de as pessoas ficarem em casa (***Increase in Residential Stay***). Esta métrica avalia o aumento da percentagem de pessoas que preferem ficar em casa durante a pandemia COVID-19, em comparação com antes da pandemia. Por exemplo, no primeiro registo do *dataset* podemos ver que para o dia 17 de Fevereiro de 2020 no Afeganistão houve um aumento de 1,33% de pessoas que preferem ficar em casa, quando comparando com a percentagem de pessoas antes da pandemia.

3.2 Etapa 1 - *Merge* do *Dataset*

O pipeline começa por utilizar a biblioteca Pandas[5] da linguagem Python para unir os 5 *datasets* que foram escolhidos. Esta biblioteca foi escolhida pois é direcionada para a representação de tabelas com vários tipos de dados através do uso de *dataframes* permitindo ao utilizador maior facilidade na manipulação destes objetos.

Como forma de fazer a união dos *datasets* foi utilizado o código de cada país e a data associada ao registo para assim unir os novos atributos (colunas) de cada *dataset*.

Para o *dataset* 1 apresentado acima, começou-se por transformar o código do país (*iso_code*) de 3 caracteres para o código de 2 caracteres. Esta transformação deveu-se ao facto dos restantes *datasets* associarem a cada país um código de 2 caracteres em vez de 3 e assim conseguirmos unir os 5 *datasets* por esse código. Para isso vamos utilizar a biblioteca **PyCountry** do Python que nos permite obter a informação variada sobre diversos países. Com esta biblioteca conseguimos obter os vários nomes para cada país assim como o respetivo código de 2 e 3 caracteres (*alpha_2* e *alpha_3*), respetivamente. Desta

forma, associamos o código presente no *dataset* 1 ao código de 3 caracteres respetivo, substituindo a coluna *Code* por esse código.

Por outro lado, no *dataset* 5, verificou-se que a coluna *Day* possuía as suas datas em formato Dia-Mês-Ano, o que não é concidente com os restantes *datasets*, que apresentam as suas datas em formato Ano-Mês-Dia. Deste modo, foi necessário alterar o formato das datas no *dataset* 5, de modo a que as datas em todos os *datasets* fossem compatíveis.

Além disto, foi verificou-se também que existiam registos de dados sem código ISO associado, pelo que, foi necessário retificar este aspeto para que o processo de *merge* pelo código fosse possível. Assim sendo, para cada *dataset* obteve-se a lista dos países que não possuíam código. No *dataset 1* os países sem código são o Kosovo, Namibia, Bonaire Saba,Sint Eustatius e o registo Other, sendo que destes, na biblioteca *Pycountry*, apenas é conhecido o código ISO da Namibia. Por esta razão, os restantes registos serão tratados na etapa 3. No *dataset* 2, 3 e 5 não existem países sem código. Já no *dataset* 4, surgiram como países sem código, o Laos e o Cote d'Ivoire, pelo que, neste caso, a resolução do problema passou apenas em alterar o seu nome para que estivessem de acordo com a biblioteca *Pycountry*. Assim o país Laos passou a Lao People's Democratic Republic e o Cote d'Ivoire passou a Côte d'Ivoire.

Por fim, efetuo-se, iterativamente, o *merge* dos diversos *datasets* sendo que, previamente,foi necessário, renomear as colunas usadas nesta junção, para que tivessem o mesmo nome nos 5 *datasets*.

3.3 Etapa 2 - Armazenamento do *Dataset*

Esta etapa do *pipeline* teve como objetivo armazenar o *dataset* obtido na etapa anterior na base dados MongoDB. Decidiu-se armazenar o ficheiro original com valores em falta, isto é, valores não tratados, pois caso haja necessidade de obter o *dataset* original sem dados sintéticos gerados, conseguimos obtê-lo com esta fase. Além disso, com o armazenamento deste ficheiro em bruto conseguimos identificar facilmente os dados sintéticos que serão gerados em fases futuras.

Para este armazenamento, foi utilizado a biblioteca PyMongo[6] do Python para assim conseguirmos integrar o código desenvolvido na etapa 1 nesta fase do armazenamento, mantendo desta forma o pipeline contínuo e com as suas fases integradas umas nas outras.

Esta biblioteca foi utilizada pois permite a comunicação entre o MongoDB e a linguagem Python, sendo que o MongoDB é uma ferramenta familiar que armazena documentos do tipo JSON de forma robusta e flexível. Além disso, esta base de dados documental tem uma linguagem estilo *query* que permite retirar informação do documento original de forma eficiente e rápida.

Como o MongoDB utiliza o ficheiros em JSON para armazenar a informação, vamos começar por passar o *dataframe* do Pandas que contém a união dos *datasets* para um dicionário através do método *to_dict* da biblioteca Pandas. De seguida vamos utilizar o método *MongoClient* da biblioteca Pymongo para criar uma conexão com a coleção e base dados onde vamos

armazenar o ficheiro JSON. Utilizando o método ***insert_one*** desta biblioteca conseguimos armazenar o documento na coleção pretendida.

Para obtermos os dados armazenados vamos conectar, novamente, com a base de dados e coleção através do método *MongoClient* e vamos utilizar uma *query* que nos vai devolver toda informação presente nessa coleção. Essa informação será armazenada num ficheiro CSV, que posteriormente será utilizado para o tratamento.

3.4 Etapa 3 - Processo ETL

Após o armazenamento dos dados no MongoDB, é necessário efetuar o tratamento do *dataset* obtido na etapa 1, sendo que no fim, os dados resultantes (e já tratados) serão carregados para o Power BI. É de realçar que as etapas do *pipeline* apresentado, bem como a sua ordem, tornam isto um processo **ETL**, visto que, em primeiro lugar obtêm-se os dados previamente armazenados na base de dados MongoDB (*Extract*), de seguida estes são tratados a (*Transform*) e por fim, é feito o seu armazenamento num ficheiro excel *xlsx* (*Load*), para posteriormente passar para a plataforma de visualização dos dados.

Para realizar o tratamento desenvolvido nesta secção foi utilizado a biblioteca PySpark[7] da linguagem Python. Esta biblioteca, oferece aos utilizadores uma API do Apache Spark que aplica os métodos e funcionalidades desta ferramenta em ambiente Python, mantendo desta forma consistência com as bibliotecas de Python que também serão utilizadas. Esta ferramenta é mais eficaz na implementação de métodos iterativos e complexos.

Antes de começar o tratamento propriamente dito, foi necessário iniciar uma *Spark Session* visto que o *Pyspark* foi a ferramenta usada nesta etapa do *pipeline*. Posto isto, e estando a *session* criada, basta converter o *dataset* num objeto *Pyspark dataframe*, sobre o qual se vão efetuar as alterações necessárias.

3.4.1 Renomeação de países com caracteres não reconhecidos

Em primeiro lugar, e de forma a facilitar a posterior visualização de dados, foram alterados os nomes dos países que possuíam caracteres não reconhecidos. Nos dados deste caso de estudo, apenas o país Costa de Marfim (em português) apresentava este problema, sendo que se alterou o seu nome de "Côte d'Ivoire" para "Cote de Ivoire".

3.4.2 Remoção de colunas redundantes

Por outro lado, e após uma análise ponderada, concluiu-se que no *dataset* existiam algumas colunas redundantes e que apresentavam pouca relevância para o caso de estudo em questão, pelo que estas poderiam ser removidas. As colunas retiradas foram: *daily_vaccinations_raw*, *total_boosters_per_hundred*, *daily_vaccinations_per_million* e *people_vaccinated_per_hundred*.

3.4.3 Tratamento de valores nulos

No caso dos tratamento dos valores nulos do *dataset*, foi necessário analisar diferentes casos e tratá-los de diferentes formas, tal como explícito nas diferentes secções abaixo.

Registos anteriores ao início da vacinação

O primeiro tratamento de valores nulos teve como alvo os registos anteriores ao início da vacinação, para cada país. Visto que o *dataset* final é uma junção de vários *datasets* distintos e o intervalo de datas em cada um varia, no caso dos dados referentes à vacinação apenas existem registos após finais de 2020 e inícios de 2021 (dependendo do país em questão). Desta forma, com a junção de todos os dados, os registos anteriores às datas previamente mencionadas, apresentam valores nulos que deveriam ser substituídos por zero. Assim, foi necessário efetuar um tratamento que, para todas as colunas referentes à vacinação e para cada país, verifica qual a data do primeiro registo ocorrido e, para as datas anteriores, são colocados os valores a zero.

Desta forma, poderemos ter uma visão mais realista dos dados na parte da visualização e avalia-los de forma mais precisa.

Remoção dos países com muitos valores em falta

Após colocar a 0 todos os campos da vacinação antes do início desta, vamos determinar quais os países que não tem dados suficientes para conseguir fazer a uma boa geração de valores sintéticos que serão utilizados para preencher os campos em falta. Esta abordagem tem como objetivo eliminar os países com colunas com muitos valores a *null*, pois nesses casos os dados que se encontram nessas colunas não são suficientes para gerar valores sintéticos realistas e que sigam a mesma distribuição dos que já se encontram nessas colunas.

Para isso, foi analisada a percentagem de valores a *null* que existem no *dataset* e reparou-se que em média cerca de 20% dos campos, em cada país, encontram-se por preencher. Desta forma, conseguimos inferir que para a maioria dos países vamos conseguir gerar valores sintéticos realistas. Contudo, ainda podem existir alguns países com colunas que não tenham dados suficientes para fazer essa geração dos dados. Para lidar com esses casos, vamos percorrer os países e para cada um analisar a percentagem de valores em falta que existe em cada coluna. Caso exista uma coluna com mais de 60% de valores a null, então esse país, juntamente com todos os seus registos, vai ser eliminado do *dataset*. Decidiu-se eliminar os países com mais de 60%, pois assim garante-se que os países que ficam no *dataset* após este tratamento, possuem, para cada coluna, um conjunto de valores significativos, isto é, acima de 50% que permitem fazer uma geração de valores sintéticos de forma realista.

Geração de valores sintéticos

Finalmente, vamos gerar os valores sintéticos para os campos que ainda se encontrem em falta. Esta etapa do tratamento pode ser divida em duas partes.

Numa primeira fase vamos preencher esses campos com a média de valores em cada semana e numa segunda fase vamos preencher os campos em falta que sobram com o resultado da interpolação linear utilizando os valores das semanas anteriores e seguintes.

Para fazer este tratamento vamos começar por associar a cada registo qual o ano e a semana do ano em que este foi realizado. Esta abordagem foi utilizada para assim conseguimos agrupar os registos por país, ano e semana do ano, sendo este agrupamento utilizado para gerar os dados em falta de forma realista e consistente.

Numa primeira fase de geração os valores sintéticos para os campos em falta, vamos, para cada coluna do *dataset* que possua valores numéricos, determinar a média para cada semana do ano e país, sendo que esta média será utilizada para preencher os campos a nulo existentes nessa coluna. Para as colunas com valores decimais utilizamos a média gerada, contudo para as colunas com valores inteiros vamos utilizar a média arredondada ás unidades. Desta forma, vamos gerar para cada campo a nulo dentro de cada semana do ano e para cada país gerar valores sintéticos consistentes e próximos da realidade. Para isso utilizamos o método to *toPandas* para passar o *dataframe* do *Spark* para o *dataframe* do *Pandas*. Nesse *dataframe* utilizamos o método *transform* para obter a média dos valores agrupados.

No entanto, quando não existem valores para nenhum dia de uma semana num determinado país então não será possível determinar a média para essa semana. Nesses casos avançamos para uma segunda fase na geração dos valores sintéticos que tem como objetivo calcular valores para as semanas a nulo através da interpolação linear utilizando os valores das semanas anteriores e/ou seguintes desse país. Por exemplo, quando numa semana os valores da vacinação estão a aumentar então na semana seguinte, se esta não tiver nenhum valor e utilizando a interpolação linear, os valores da vacinação para essa semana também estão a aumentar segundo o mesmo padrão. Assim conseguimos obter valores realistas que se encontram dentro da mesma distribuição dos valores das semana seguinte e/ou anterior. Para isso utilizamos o método to *toPandas* para passar o *dataframe* do *Spark* para o *dataframe* do *Pandas*. Nesse *dataframe* utilizamos o método *interpolate* com o método linear para assim obter a distribuição para os valores em falta.

3.5 Etapa 4 - Visualização dos dados

Nesta última etapa do *pipeline* vamos implementar diversos gráficos e *dasboards* de forma a analisar a influência que a pandemia COVID-19 teve na preferência de ficar em casa.

Para realizar esta visualização dos dados vamos utilizar o PowerBi[8] que permite a criação de gráficos que se enquadram no problema de forma a conseguimos obter uma resposta para o caso de uso. Esta ferramenta foi escolhida pois é intuitiva e fácil de utilizar e como os dados que serão utilizados para visualização não são obtidos em tempo real, não há a necessidade de utilizar ferramentas mais complexas.

Após armazenar o ficheiro EXCEL processado e tratado, este será importado para o PowerBi onde vamos conseguir associar as variáveis e atributos do ficheiro de forma a criar um gráfico representativo do problema.

Iniciamos a análise com a criação de uma *dashboard* que permite analisar temporalmente e para cada país a influencia que o número de mortes e o número de contaminados por COVID-19 tiveram nas restrições que limitam o ajuntamento e obrigatoriedade *stay at home*. A esta *dashboard* também foi acrescentada uma análise que avalia a influência que o processo de vacinação e os valores da pandemia tiveram não só nas restrições aplicada como também na preferência de as pessoas ficarem em casa (*residential stay*). Assim, como esta *dashboard* podemos ver como é que a pandemia COVID-19 e o processo de vacinação condicionaram as restrições aplicadas e a vontade das pessoas em *stay at home*.

Finalmente, foram criadas outras 2 *dashboards* que analisam para cada país e região WHO (Euro, AMRO, AFRO, etc) a influência que a pandemia COVID-19 e vacinação tiveram tanto nas restrições aplicada como na evolução da preferência de ficar em casa. Desta forma, conseguimos analisar os países com valores mais peculiares e excecionais comparando-os entre si para assim chegar a conclusões mais gerais e abrangentes.

4 Resultados

Nesta secção irão ser apresentadas as *dashboards* criadas no *PowerBI* para a visualização de dados para responder ao caso de estudo em questão.

Dashboard geral por país

É de notar que a primeira *dashboard* apresentada foca-se numa análise generalizada para num único país, neste caso será Portugal. Nesta *dashboard* podemos ver a influência que a vacinação e os valores da pandemia tiveram nas restrições dos ajuntamentos e *stay at home* e na vontade das pessoas ficarem em casa.

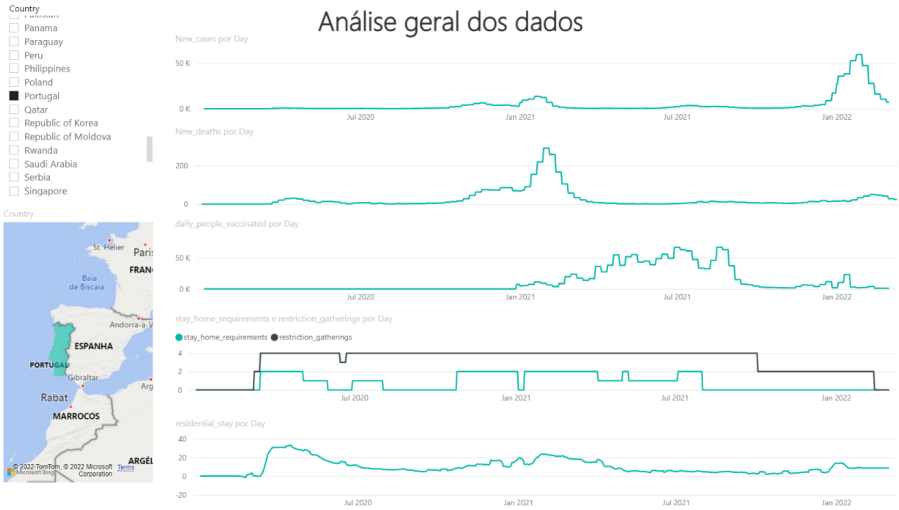

Fig. 2 Visão geral do caso de estudo

Dashboard novos casos e mortes

Além disso, também consideramos relevante criar uma *dashboard* para analisar em maior detalhe para cada zona WHO e para cada país o número de contaminados e mortes. Desta forma, conseguimos ver influência que a pandemia teve pelo mundo dando mais relevância à quantidade de mortes e casos por COVID-19 e não tanto à evolução destes valores ao longo do tempo.

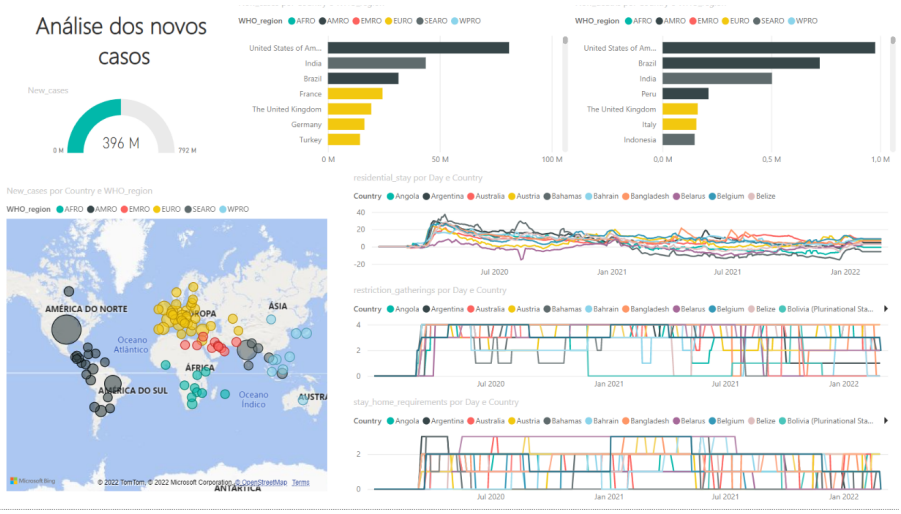

Fig. 3 Análise dos novos casos e mortes pelo mundo

Com esta *dashboard*, podemos analisar mais relevante para os países que tiveram foram mais influenciados pela pandemia. Assim, vamos analisar para os EUA, India e Brasil a influência que os valores do COVID-19 tiveram nas restrições e na vontade de ficar em casa, visto que são os países que tiveram mais casos e mortes por COVID-19. Com estes gráficos, podemos fazer uma análise comparativa entre estes três países e como o elevado número de casos e mortes condicionaram as restrições e a preferência das pessoas de ficar em casa.

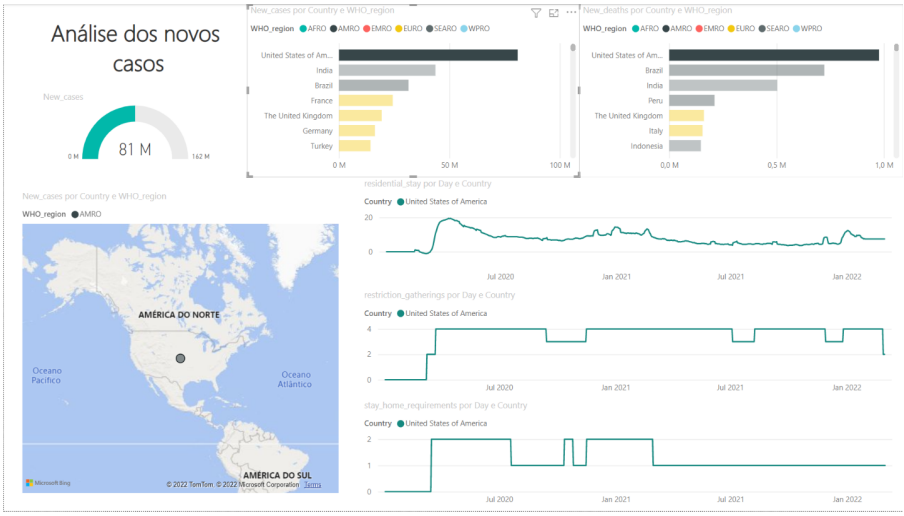

Fig. 4 Influência que a pandemia teve no caso de uso nos EUA

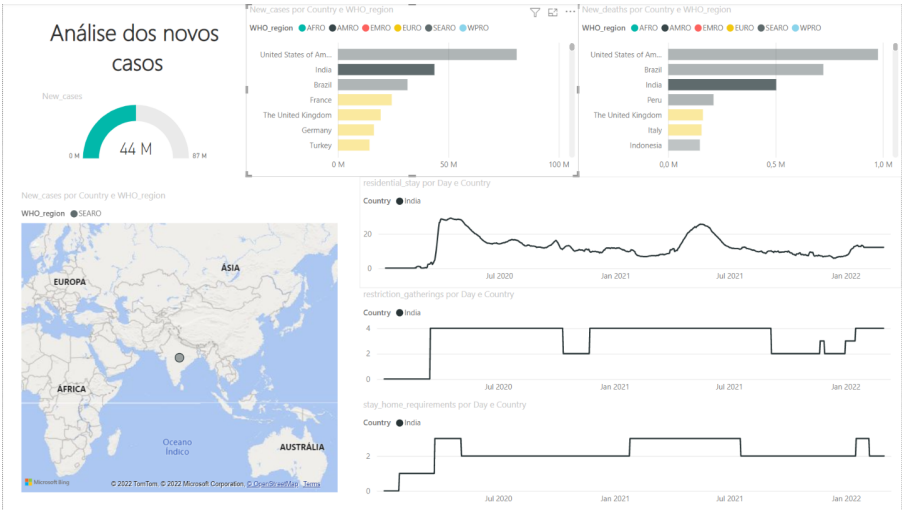

Fig. 5 Influência que a pandemia teve no caso de uso nos India

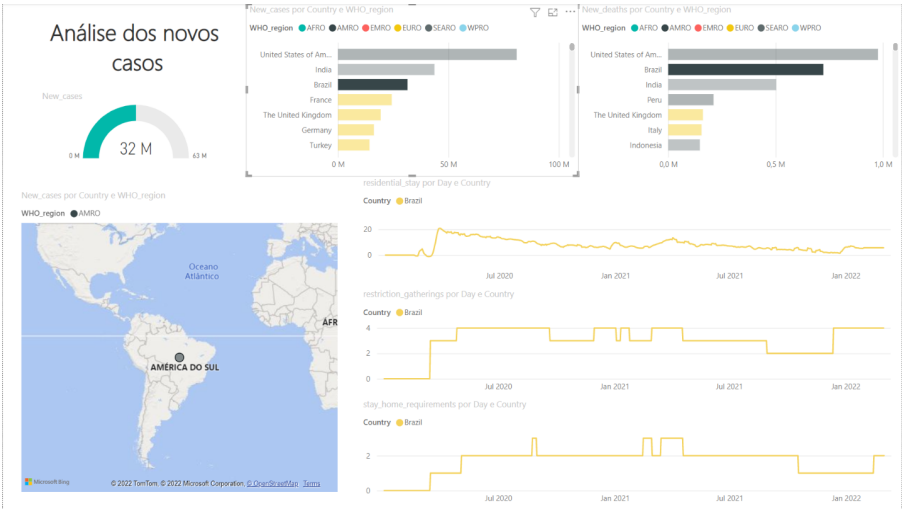

Fig. 6 Influência que a pandemia teve no caso de uso no Brasil

Dashboard Vacinação

Relativamente à vacinação, a *dashboard* criada foi análoga à anterior, na medida em que é possível fazer uma análise tanto por país como por região da WHO. Nesta *dashboard* podemos ver como é que o número de pessoas vacinadas influência no caso de uso apresentado pelo mundo dando mais relevância à quantidade vacinados e não tanto à evolução destes valores ao longo do tempo.

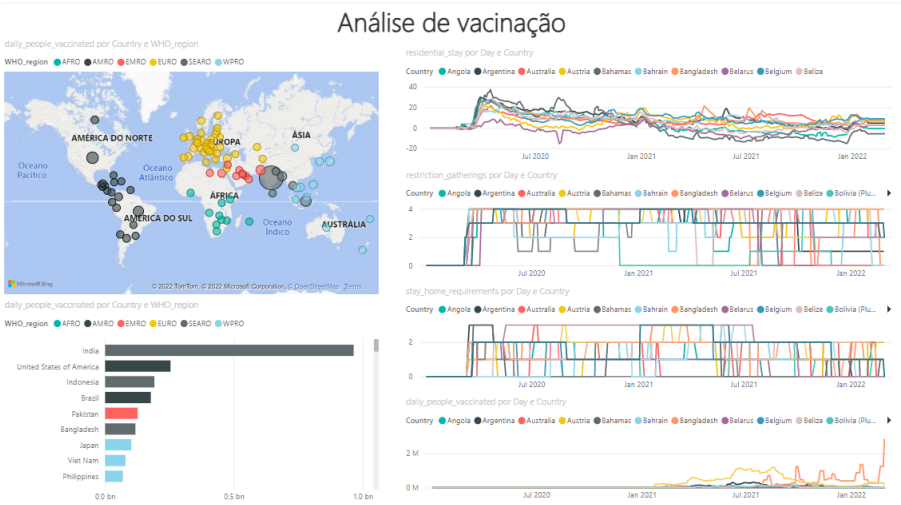

Fig. 7 Análise da vacinação pelo o mundo

Com esta *dashboard*, podemos analisar influencia que a vacinação teve nos países analisados na *dashboard* anterior. Assim, vamos analisar para os EUA, India e Brasil a influência que os valores do vacinação tiveram nas restrições e na vontade de ficar em casa, visto que são os países que tiveram números maiores de pessoas vacinados. Com estes gráficos, podemos fazer uma análise comparativa entre estes três países e como o elevado número de vacinados condicionaram as restrições e a preferência das pessoas de ficar em casa, para além de conseguir comparar e analisar com o número de contaminados e mortes por COVID-19.

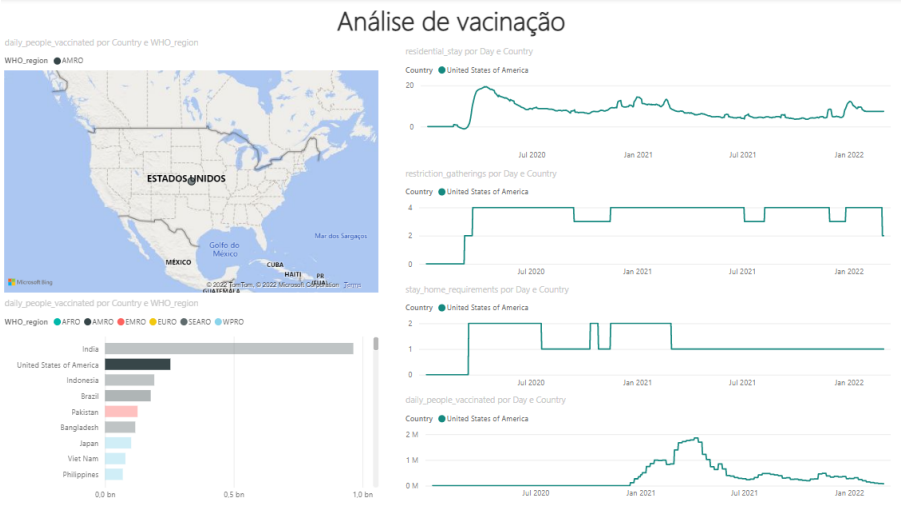

Fig. 8 Influência que a vacinação teve no caso de uso nos EUA

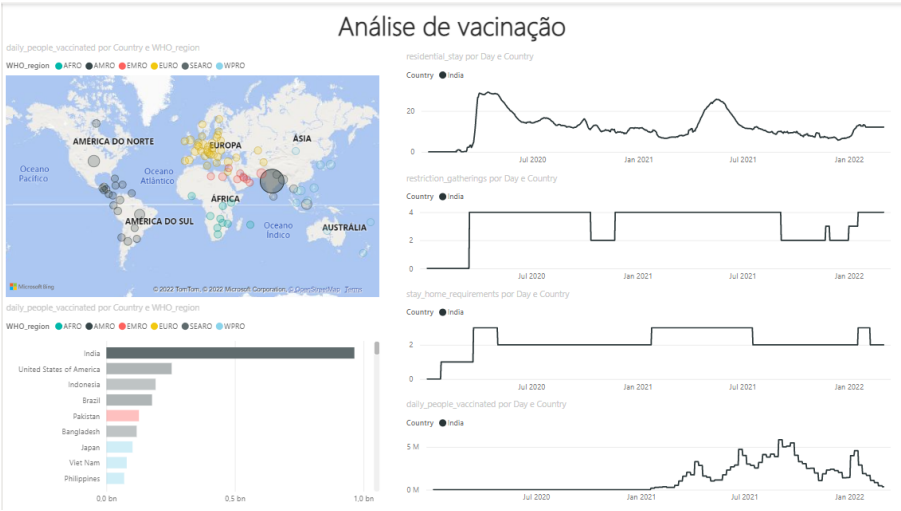

Fig. 9 Influência que a vacinação teve no caso de uso na India

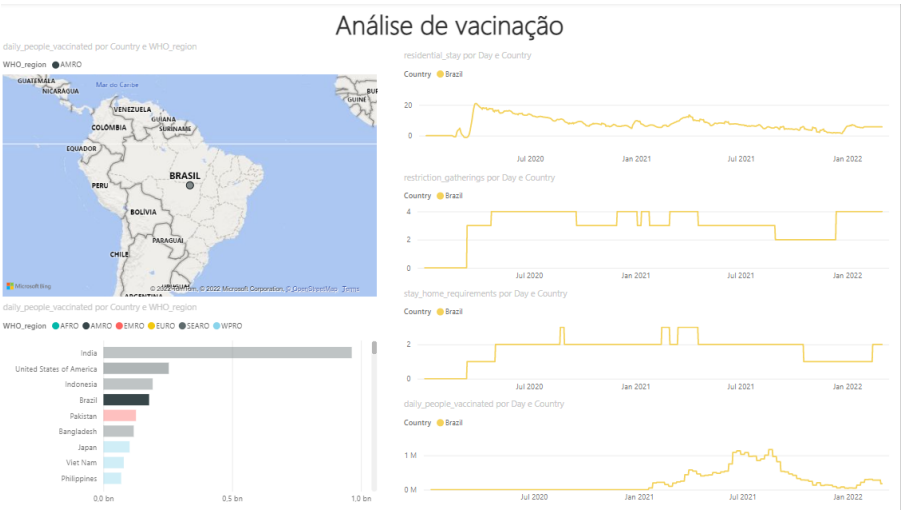

Fig. 10 Influência que a vacinação teve no caso de uso no Brasil

Análise de métricas pertinentes

Finalmente, foi criada uma *dashboard* de forma a conseguir analisar qual será a melhor métrica para analisar os valores tanto da pandemia como da vacinação quando queremos comparar.

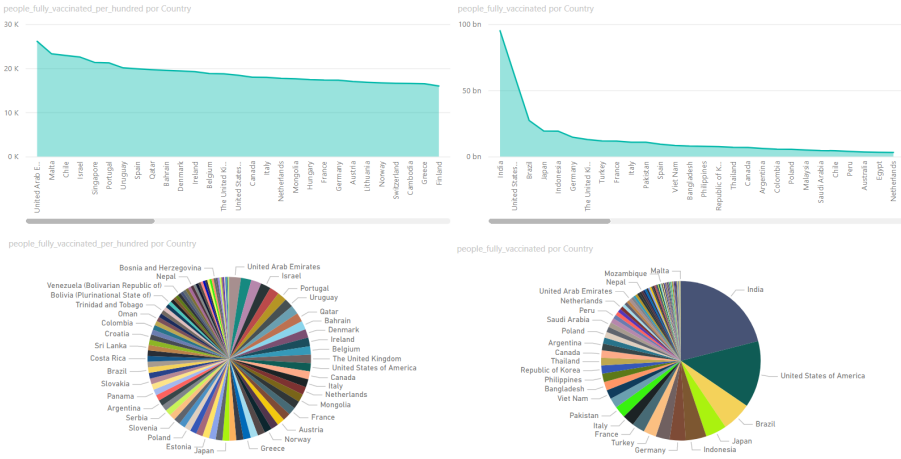

Fig. 11 Comparação da vacinação total com a vacinação por 100 habitantes

5 Discussão

Dashboard geral por país

Com base na figura 2, iremos começar por analisar com detalhe, de que forma os diversos fatores relacionados com a COVID-19 se relacionam entre si e o modo como estes impactaram não só as restrições impostas pelo governo como também a preferência da população em ficar em casa. É de realçar que esta exploração será feita apenas para os dados de **Portugal**, no entanto, a análise seria análoga para qualquer outro país.

Em primeiro lugar, é interessante salientar que relativamente ao número de casos, Portugal verificou dois picos, um em janeiro de 2021 com aproximadamente 13500 casos num dia e outro em janeiro de 2022, bastante mais drástico, com aproximadamente 59000 casos num dia. Por outro lado, observou-se também um pico do número de mortes diário em janeiro de 2021 e um aumento, não muito significativo, em janeiro de 2022. Além disto, tendo em conta o gráfico da vacinação no país, pode-se ver que a vacinação foi gradual, tendo início em dezembro de 2020 e sofrendo uma redução em setembro de 2021. Desta forma, podemos concluir que apesar do segundo pico de casos diários ser bastante significativo, este não provocou muitas mortes visto que grande parte da população já se encontrava vacinada. No entanto, no primeiro pico de casos, apesar de este ser pouco significativo, houve um número de mortes bastante alto, pois o processo de vacinação ainda estava no início.

Por outro lado, e referente ás restrições impostas, é possível observar que a após ter havido um aumento no número de mortes em abril de 2020, houve também um agravamento das restrições aos ajuntamentos e no confinamento. As restrições aos aglomerados de pessoas, mantiveram-se praticamente no seu valor mais alto, 4, até outubro de 2021, quando o processo de vacinação verificou uma desaceleração, ou seja, havia mais pessoas vacinadas. Por outro lado, as restrições do confinamento, foram sofrendo variações ao longo do tempo, sendo que, o valor mais alto atingido foi o 2. Após o mês de abril de 2020 houve uma redução no número de mortes, o que se traduziu num aliviamento do confinamento, sendo que as restrições passaram de 2 para 1 e por fim para 0. No entanto, no início do processo de vacinação, houve também um pico do número de mortes, que se traduziu numa passagem das restrições de confinamento para os níveis 2 e 1. Quando a vacinação foi evoluindo e o número de mortes foi decrescendo, estas restrições foram aliviadas para o valor 0.

Posto isto, e focando na vontade das pessoas em ficar em casa comparativamente ao período pré pandemia, pode-se verificar que, o já referido, aumento do número de mortes e as consequentes restrições aplicadas, traduziram-se num aumento da vontade da população em ficar em casa, sendo que esta preferência foi diminuindo quando o número de mortes reduziu e o confinamento aliviado. Por outro lado, quando se verifica o pico de mortes em janeiro de 2021, a preferência da população em ficar casa sofre também um aumento. No entanto, quando o processo de vacinação vai evoluindo, e a percentagem da população vacinada vai aumentando, a vontade das pessoas em ficar em

ambiente domiciliar, vai diminuindo até valores próximos de zero, ou seja, a vontade das pessoas em ficar em casa assemelha-se à registada antes da pandemia. Por fim, em janeiro de 2022, com o aumento significativo do número de casos, houve também um aumento da vontade em ficar em casa, sendo uma provável causa disto, o facto de as pessoas deixarem de se sentir tão seguras no exterior.

Dashboard novos casos e mortes

Relativamente à **análise dos novos casos e mortes**, figura 3, podemos tirar diversas conclusões relativamente ao número de casos e mortes com a evolução da pandemia, como por exemplo o número total de casos globais de Covid que corresponde a 396 milhões, durante o período de análise. Além disso, numa visão geral, comparando o número de casos e mortes por país, conseguimos observar que se encontram equilibrados na medida em que os países com mais casos apresentam maior número de mortes. Também é possível obter informações relativamente aos países que tiveram mais casos e mais mortes, o que será útil para posteriormente efetuar uma análise mais detalhada destes países.

Desta forma, podemos focar a nossa análise no **top 3 de países com mais casos e mortes**. No caso dos **EUA**, este trata-se do país com um número de casos e mortes mais significativos. Podemos observar na figura 4 que houve 81 milhões de casos durante o período de análise e que houve aproximadamente 1 milhão de mortes, correspondendo a cerca de 1% dos casos de Covid. Além disso, existe uma forte correlação entre as restrições impostas neste país e a vontade das pessoas ficarem em casa. Pela curva acentuada no gráfico de "residential_stay" podemos concluir que no início da pandemia as pessoas preferiam manter-se em casa e, nesta altura, as restrições de ajuntamentos e confinamento também eram altas. Com o passar do tempo, as restrições foram variando o que se refletiu também numa variação da vontade das pessoas ficarem em casa. Finalmente, podemos concluir que este país foi o países com mais casos e mortes por Covid pois, a evolução da preferência de as pessoas ficarem manteve-se com valores baixos e as restrições *stay at home* nível 3 nunca chegaram a ser aplicadas que podem ter promovido mais contactos entre pessoas e feito com que as contaminações por Covid aumentassem mais rapidamente.

Em relação ao segundo país alvo de análise, a **Índia**, podemos observar que o seu número de casos no período de análise atingiu 44 milhões e o número de mortes ronda os 500 mil (1% dos casos). Pela figura 5 observamos que este país teve muita variação nas restrições ao longo do tempo, sendo que no inicio da pandemia as restrições aumentaram de forma proporcional à vontade das pessoas ficarem em casa. Além disso, pode-se observar que quando a pandemia começou as restrições de ajuntamentos aumentaram significativamente. No entanto, as restrições de confinamento foram variando e aumentando gradualmente, e apenas atingiram o seu pico num curto período de tempo em 2020, o que pode ser reflexo do elevado número de casos e mortes. Tal como pode

ser inferido por esta analise, a India tentou aplicar restrições à medida que os casos e mortes fossem aumentando. Esta abordagem podem ter tido impacto positivo no número de casos e mortes contudo, como a índia é um país com grande população esta acaba sempre por ficar no top de países que mais sofreram com a pandemia principalmente devido ao número de casos e mortes por Covid.

Por fim, iremos analisar o caso do **Brasil**, que atingiu 32 milhões de casos no período de análise e cerca de 700 mil mortes, o que corresponde a cerca de 2% dos casos de Covid. Podemos observar na figura 6 que as restrições implementadas aumentaram de forma brusca no inicio da pandemia assim como a vontade das pessoas ficar em casa. Com o passar do tempo as restrições de confinamento foram variando e apenas em curtos períodos de tempo atingiram um pico, o que pode ter potenciado o número de casos e mortes tão elevado.

Dashboard Vacinação

Nesta *dashboard* podemos observar no cômputo geral como variaram as diferentes restrições e a vacinação, assim como a sua influência no caso de estudo (na vontade das pessoas ficarem em casa). Pela figura 7 podemos observar no mapa os diferentes países do caso de estudo em que as bolhas representam o número de pessoas vacinadas. Este mapa também oferece uma visão agrupada dos países pela sua região da WHO para uma análise por região.

Os países alvo de estudo nesta secção serão os países referidos anteriormente, de forma a completar a análise efetuada. Começando pelos **EUA**, visível na figura 8, podemos observar que os EUA não estão no top dos países com maior vacinação, com cerca de 76% de pessoas vacinadas durante o período de análise. Apesar de ser um número significativo, houve muitas pessoas a vacinar-se ao longo do ano 2021, pelo que aquele valor apenas foi atingido em 2022. Isto pode ser um indicador do grande número de casos e mortes verificado acima. Além disso, podemos visualizar pelo gráfico que a vacinação começou em Janeiro de 2021 e que isso já se refletiu no alivio das restrições de confinamento. No entanto, as restrições de ajuntamentos foram variando, refletindo períodos de mais restrições e de menos restrições.

No caso da **Índia**, visível na figura 9, o número de vacinados também não é muito significativo tendo em conta o número de casos de Covid, sendo cerca de 69%. Para este país a vacinação também começou em Janeiro de 2021, mas a evolução do número de pessoas vacinadas foi gradual e mais lenta, em comparação com os EUA. Esta demora e a percentagem de vacinados em 1 ano pode ser um indicador do grande número de casos e mortes verificados neste país. Além disso, em comparação com as restrições aplicadas no país, podemos observar que quando foi atingido um grande pico de vacinação houve também um alivio nas restrições de confinamento e ajuntamentos, assim como uma diminuição na vontade das pessoas ficarem em casa.

Por fim, analisando o caso do **Brasil** (figura 10) podemos observar que a vacinação começou também em Janeiro de 2021, e atingiu 82% de vacinados durante o período de análise. Neste país verifica-se um crescimento

gradual do número de pessoas vacinadas que se reflete também no alivio das restrições de ajuntamentos e confinamento. A diminuição das restrições após o pico de vacinação também levou a que a vontade das pessoas ficarem em casa diminuísse. Desta forma, existe uma forte correlação e influência mútua entre as métricas em estudo.

Análise de métricas pertinentes

Tendo em conta aos resultados apresentados no tópico anterior, nomeadamente a figura 11 é também possível concluir que uma análise baseada apenas na coluna *People_fully_vaccinated* poderá não ser muito informativa, uma vez que estes valores irão depender da população de cada país. A título de exemplo, no caso da Índia, um dos países mais populosos do mundo, o número de pessoas totalmente vacinadas é bastante alto, especialmente quando comparado com os restantes países. No entanto, quando é analisado número o total de pessoas vacinadas por cada 100 habitantes (*People_fully_vaccinated_per_hundred*), verifica-se que estes valores são bastante mais equilibrados. Desta forma, esta coluna transmite informação mais relevante no que diz respeito ao estado da vacinação (em cada país), permitindo também uma comparação mais realista entre os diversos países.

6 Conclusão

Dado por concluído este trabalho prático, consideramos que se tratou de uma boa oportunidade para colocar em prática os conceitos relativos à manipulação e tratamento de *Big Data*. Para aplicar esses conceitos começámos por juntar os 5 *datasets* necessários para a obtenção da informação relacionada com o caso de uso apresentado. De seguida decidimos armazenar o *dataset* obtido numa base dados que posteriormente será tratado de forma a remover os valores a nulo. Finalmente, efetuou-se a visualização dos dados de forma a obter respostas para o nosso caso de estudo.

O *pipeline* criado assim como as ferramentas utilizadas e a metodologia desenvolvidos em cada etapa foram fundamentais para a obtenção de resultados coerentes, completos e realistas. Em suma, considera-se que a arquitetura apresentada cumpre com todos os requisitos do problema.

References

[1] in Data, O.W.: Vaccinations - Covid-19 Data (Github). Github - https://github.com/owid/covid-19-data/blob/master/public/data/vaccinations/vaccinations.json (2022)

[2] in Data, O.W.: Restrictions on public gatherings in the COVID-19 pandemic. Link - https://ourworldindata.org/grapher/public-gathering-rules-covid (2022)

[3] in Data, O.W.: Stay-at-home requirements during the COVID-19 pandemic. Link - https://ourworldindata.org/grapher/stay-at-home-covid (2022)

[4] Kaggle: Worldwide Residential Mobility in COVID-19. Link - https://www.kaggle.com/datasets/aestheteaman01/people-staying-in-home-during-covid19 (2022)

[5] Pandas: User Guide - Merge, join, concatenate and compare. Link - https://pandas.pydata.org/docs/user_guide/merging.html (2022)

[6] PyMongo: PyMongo 4.1.1 documentation. Link - https://pymongo.readthedocs.io/en/stable/tutorial.html

[7] PySpark: PySpark Documentation. Link - https://spark.apache.org/docs/latest/api/python/index.html

[8] Microsoft: Tutorial: Introdução ao serviço de criação no Power BI. Link - https://docs.microsoft.com/pt-pt/power-bi/fundamentals/service-get-started