BLUEEDTECH

ANA CRISTINA CHAVES
ANTONIO DUARTE MARCOS JUNIOR
THIAGO CHARLES

BOOTCAMP - PESQUISA DE IMPACTO DO CORONAVÍRUS 2020

BRASIL

NOVEMBRO 2022

1. INTRODUÇÃO

Uma pandemia é uma epidemia de doença infecciosa, que se espalha entre a população localizada numa grande região geográfica como, por exemplo, todo o planeta terra.

Em 31 de dezembro de 2019, a Organização Mundial da Saúde (OMS) foi alertada sobre vários casos de pneumonia na cidade de Wuhan, província de Hubei, na República Popular da China. Tratava-se de uma nova cepa (tipo) de coronavírus que não havia sido identificada antes em seres humanos.

Uma semana depois, em 7 de janeiro de 2020, as autoridades chinesas confirmaram que haviam identificado um novo tipo de coronavírus. Os coronavírus estão por toda parte. Eles são a segunda principal causa de resfriado comum (após rinovírus) e, até as últimas décadas, raramente causavam doenças mais graves em humanos do que o resfriado comum.

Com a pandemia, o número de estudos aumentou drasticamente e junto o volume de dados produzidos e distribuídos por todos os países com objetivo de ajudar a combater o vírus também sofreu um aumento considerável.

Para melhor controle da doença, foi gerado muitos dados estatísticos, muita informação, para combater, para prevenção, enfim, os dados foram utilizados da melhor forma em meio ao caos, e agora que a pandemia deu uma trégua, iremos utilizar as informações contidas nesses dados reunidos, referente a cinco países Espanha, equador, Chile, México e Argentina em um intervalo de Janeiro a Julho de 2020.

2. OBJETIVOS

2.1. Objetivo principal

Avaliar a evolução dos casos de COVID nos primeiros meses da pandemia e seus possíveis impactos no humor da população afetada e no mercado de energia elétrica.

2.2. Objetivo específicos

- Analisar os dados fornecidos sobre a pandemia
- Implementar metodologia para coleta de dados de diversas fontes;
- Coletar dados da API do Twitter;
- Analisar o sentimento da população afetada pela pandemia a partir de tweets;
- Criar modelo de previsão de casos de COVID;
- Identificar possíveis padrões no número de casos;
- Avaliar o possível impacto de medidas adotadas para o combate da pandemia no mercado de energia elétrica.
- Construir dashboard interativo para compartilhamento dos resultados.

3. METODOLOGIA

Neste projeto serão avaliados alguns dos possíveis impactos da pandemia em cinco países: Argentina, Chile, Espanha, Equador e México. Para este estudo a principal fonte de dados será o dataset disponibilizado na plataforma Kaggle. Tal dataset contém informação sobre o início da pandemia nos referidos países.

A metodologia de análise e modelagem constará de cinco etapas principais listadas a seguir. As etapas não necessariamente ocorrem sequencialmente, alguma poderão correr em paralelo.

A primeira etapa do trabalho irá consistir na análise exploratória dos dados fornecidos. Utilizando tanto Python quanto PowerBi buscaremos identificar os principais aspectos referente ao conjunto de dados fornecidos. Como trata-se de uma análise de série temporal serão avaliados, inicialmente, os seguintes aspectos: tendência, sazonalidade, distribuição de casos por país, evolução da taxa de contágio, evolução da taxa de mortalidade.

A etapa seguinte será avaliar a possível relação entre os casos e mortalidade dos países, tamanho da população e PIB. Nesta etapa terá o objetivo de responder às seguintes perguntas: Países com PIB maior conseguem combater melhor a pandemia? A pandemia evoluiu mais rápido em países com maior densidade populacional?

Na terceira etapa será avaliado como o humor da população evoluiu ao longo da pandemia. Será que o aumento do número de casos aumentou o negativismo da população. Para esta análise serão utilizados tweets coletados a partir da API do Twitter. A amostra utilizada conta com cerca de 3.000 tweets por mês por país.

A quarta etapa consiste em avaliar como a pandemia afetou o mercado de energia elétrica dos países. Para tanto serão coletados dados de demanda de eletricidade nos países listados para os meses em que há dados sobre a pandemia. O mercado de energia elétrica é um dos primeiros a ser afetados quando há eventos que modificam o comportamento da sociedade como: guerras, catástrofes naturais, crises econômicas e também pandemias. Assim, espera-se que as ações tomadas no combate à pandemia, em especial o lockdown, refletem em mudanças significativas no padrão de consumo dos países.

A quinta etapa é a construção de um modelo de machine learning para a previsão dos números futuros de casos de COVID nos países em estudo. A ideia inicial é utilizar o modelo ARIMA que é amplamente utilizado em séries temporais. Outras variações poderão ser testadas como: SARIMA e ARIMAX. Mas tais variações dependem de outros dados. O modelo SARIMA se aplica a séries temporais com sinal de sazonalidade significativo. Então para utilizá-lo será preciso que os testes de sazonalidade na etapa um sejam positivos. o ARIMAX depende de dados externos a série temporal. Portanto, para utilizar o modelo ARIMAX ainda seria necessário coletar outras séries de dados que possam ter relação com o número de casos de COVID. Outros modelos também poderão ser utilizados.

Ao final de todas as etapas será montado um dashboard no PowerBi para a visualização das principais métricas e insights encontrados durante o projeto. Também será arquitetada uma estrutura de DAGs no Airflow para que o processo de coleta, transformação e carregamento dos dados no datalake seja automatizado. Assim, com todo o processo automatizado a empresa poderá, em tempo real, ver a evolução da pandemia e tirar insights e decisões sobre ações futuras baseadas nos dados coletados continuamente.

O modelo de previsão que será construído será avaliado a partir de métricas clássicas para modelos de regressão como: R^2, RMSE e MAE. Também será avaliado o horizonte máximo de previsão; qual seria o número máximo de dias a frente que o modelo pode prever com o mínimo de incerteza possível?

Além das etapas planejadas há a expectativa de se utilizar: Cadeias ocultas de Markov e LSTM. As cadeias ocultas de Markov foram utilizadas com o principal objetivo de identificar padrões ocultos na série de dados. A LSTM trata-se de uma rede neural bastante utilizada para problemas com séries temporais. Ela poderá ser testada como um modelo para previsão dos casos futuros de COVID. Porém, estas duas metodologias referidas estão listadas apenas como potenciais, seu uso estará condicionado à disponibilidade de tempo hábil para que sejam exploradas adequadamente.

4. DESCRIÇÃO DOS MODELOS UTILIZADOS

Nesta seção são listadas as principais características dos diferentes modelos de machine learning que serão utilizados neste projeto.

4.1. Tradução dos textos

Optou-se por realizar a tradução dos textos do espanhol para o inglês, haja visto que a maioria dos modelos de NLP foram treinados com textos em inglês. Para realizar tal processamento foram utilizadas ferramentas disponíveis na biblioteca Spark-NLP.

Segundo o site oficial do Spark-NLP esta ferramenta é o estado da arte em processamento de texto. Esta biblioteca é desenvolvida pelo Jhon Snow Labs e conta com os modelos mais recentes para processamento de texto. O Spark NLP da John Snow Labs é uma biblioteca de processamento de texto de código aberto para Python, Java e Scala. Ele fornece versões de nível de produção, escaláveis e treináveis das pesquisas mais recentes em processamento de linguagem natural.

O modelo de tradução utilizado pelo Spark-NLP é o Marian. Marian é uma estrutura Neural Machine Translation escrita em C++. Está sendo desenvolvido principalmente pela equipe do Microsoft Translator. Muitos colaboradores acadêmicos (principalmente a Universidade de Edimburgo e, no passado, a Universidade Adam Mickiewicz em Poznań) e comerciais ajudam no seu desenvolvimento.

4.2. Análise de sentimentos

4.2.1. Spark NLP

O Spark NLP também oferece um modelo para análise de sentimentos. Porém, por limitações computacionais tal modelo não pode ser utilizado neste projeto. Mas a seguir segue uma breve descrição do mesmo.

O Sark NLP conta com uma pipeline pré-treinada para classificação de sentimentos utilizando tweets. O que é o ideal para este projeto. Para esta tarefa ele utiliza o Universal Sentence Encoder.

O Universal Sentence Encoder codifica texto em vetores de alta dimensão que podem ser usados para classificação de texto, similaridade semântica, agrupamento e outras tarefas de linguagem natural.

O modelo é treinado e otimizado para texto com comprimento maior que a palavra, como sentenças, frases ou parágrafos curtos. Ele é treinado em uma variedade de fontes de dados e uma variedade de tarefas com o objetivo de acomodar dinamicamente uma ampla variedade de tarefas de compreensão de linguagem natural. A entrada é um texto em inglês de tamanho variável e a saída é um vetor de 512 dimensões.

4.2.2. NLTK

A biblioteca NLTK é dedicada para métodos de processamento de linguagem natural. Entre as funções oferecidas está o Sentiment Intensity Analyser. Um analisador de sentimentos de textos que além de classificar um texto como sendo de sentimentos positivo, negativo ou neutro também informa o quanto o texto se enquadra em cada categoria de sentimentos. Como motor de classificação o NLTK utiliza o VADER.

O VADER (Valence Aware Dictionary and Sentiment Reasoner) é uma ferramenta de análise de sentimentos baseada em léxico e regras que está especificamente sintonizada com os sentimentos expressos nas mídias sociais. O VADER usa uma combinação de um léxico de sentimento e uma lista de recursos lexicais (por exemplo, palavras) que geralmente são rotulados de acordo com sua orientação semântica como positiva ou negativa. O VADER não apenas informa

sobre a pontuação de positividade e negatividade, mas também nos informa sobre o quão positivo ou negativo é um sentimento. [SENTIMENTAL ANALYSIS USING VADER. interpretation and classification of… | by Aditya Beri | Towards Data Science]

4.3. Previsão de séries temporais

Os modelos iniciais que são planejados de serem utilizados são o ARIMA e sua variação para sazonal o SARIMA.

4.3.1. ARIMA

O modelo Autoregressivo Integrado de Média Móvel (ARIMA) é uma combinação dos modelos Auto Regressivo (AR) e o de Médias Móveis (MA) foi popularizado no trabalho de referência de Box e Jenkins (1970). Além de considerar os padrões AR e MA leva em conta a diferenciação, uma forma de remover tendências e tornar a série temporal estacionária.

Matematicamente o modelo pode ser descrito como:

$$y_t = c + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \theta_1 + \theta_1 e_{t-1} = + \ldots + \theta_q e_{t-q} + e_t$$

Onde y_t é a série diferenciada. A equação acima descreve o modelo **ARIMA(p, d, q)**, onde:

- **p** é a ordem do modelo autoregressivo;
- **d** é o grau de diferenciação;
- **q** é a ordem do modelo de média móvel.

4.3.2. SARIMA

Modelos ARIMA são capazes também de modelar séries que apresentam um componente sazonal, sendo descrito como:

ARIMA(p, d, q)(P, D, Q)m

Onde o primeiro parênteses se refere à parte não-sazonal do modelo e o segundo à parte sazonal. corresponde ao número de períodos sazonais.

5. ANÁLISE EXPLORATÓRIA

5.1. Descrição do conjunto de dados

Conjunto de dados COVID-19 *, Kaggle* Este conjunto possui dados do número de casos confirmados, mortes e recuperados por dia em vários países entre Janeiro de 2020 a Julho de 2020.

- country_wise_latest.csv
- Covid_19_clean_complete.csv
- day_wise.csv
- Full_grouped.csv
- Usa_county_wise.csv
- worldometer_data.csv

API do Twitter A rede social Twitter.

WebScraping Utilização da técnica de garantia de dados de sites, como Wikipédia.

5.2. Dicionário de Dados

worldometer_data.csv	
Country/Region	País/Região
Continent	Continente
Population	População
TotalCases	Total de casos
NewCases	Novos casos
TotalDeaths	Total de mortes
NewDeaths	Novas mortes
TotalRecovered	Total de recuperados
NewRecovered	Novos casos recuperados
Active Cases	Casos ativos
Serius, Critical	Casos sérios/ críticos
Tot Cases/1m pop	Total de casos por 1m habitantes
Death/ 1m pop	Morte por 1m de habitantes
TotalTests	Total de testes
Tests/1M pop	Testes por 1m de habitantes
Who Region	Região da OMS

covid *19* clean_complete.csv - Dia a dia do país não. de casos (não possui dados em nível de condado/estado/província)	
Province/States	Províncias e estados
Country/Region	País/Região
Lat	Latitude
Long	Longitude
Date	data
Confirmed	Confirmados
Deaths	Mortes
Recovered	Recuperados
Actives	Ativos
Who Region	Região da OMS

full_grouped.csv - Dia a dia do país não. de casos (tem dados em nível de condado/estado/província)	
Date	Data
Country/Region	País/Região
Confirmed	Casos confirmados
Deaths	mortes
Recovered	recuperados
Active	Casos ativos
New Cases	Novos casos em 24h
New deaths	Novas mortes
new recovered	Novos casos recuperados 24h
who region	Região da OMS

country *wise* later.csv - Nº do nível do país mais recente. de casos	
Country/Region	País/Região
Confirmed	Casos confirmados
Deaths	mortes
Recovered	recuperados
Active	Ativos
New cases	Novos casos em 24h
New deaths	Novas mortes
New recovered	Novos casos recuperados 24h
Deaths / 100 Cases	Mortes/ 100 casos
Recovered / 100 Cases	Recuperados/ 100 casos
Deaths / 100 Recovered	Mortes / 100 Casos
Confirmed last week	Confirmados na semana passada]
1 week change	Mudança de 1 semana
1 week % increase	semana % de aumento
WHO Region	Região da OMS

day_wise.csv - Dia sábio não. de casos (não possui dados em nível de país)
Não iremos utilizar por não ser específico por país como foi solicitado
usa *county* wise.csv - Dia a dia no nível do condado. de casos
Informações somente dos Estados Unidos

5.3. ANÁLISE DE DADOS DESCRITIVOS

Análise descritiva dos dados fornecidos nos datasets, fizemos um filtro dos países solicitados, Argentina, Chile, Equador, México e Espanha, reduzindo os datasets originais. conforme segue:

Dataset: worldometer_data.csv

```
#   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Country/Region  5 non-null      object
 1   Continent       5 non-null      object
 2   Population      5 non-null      float64
 3   TotalCases      5 non-null      int64
 4   NewCases        1 non-null      float64
 5   TotalDeaths     5 non-null      float64
```

```
 6   NewDeaths          1 non-null      float64
 7   TotalRecovered     4 non-null      float64
 8   NewRecovered       1 non-null      float64
 9   ActiveCases        4 non-null      float64
 10  Serious,Critical   5 non-null      float64
 11  Tot Cases/1M pop   5 non-null      float64
 12  Deaths/1M pop      5 non-null      float64
 13  TotalTests         5 non-null      float64
 14  Tests/1M pop       5 non-null      float64
 15  WHO Region         5 non-null      object
dtypes: float64(12), int64(1), object(3)
memory usage: 680.0+ bytes
```

Ao analisar os dados obtidos pelo worldometer_data.csv, verificou-se que o dataset contém os dados acumulados por países sobre o covid-19. Nos datasets que contém datas, elas estão dentro de 22/01/2020 à 27/07/2020 .

Outro ponto verificado, foi a ausência de alguns dados, para a obtenção desses dados utilizamos os dados da dataset covid_19_clean_complete.csv.

	Country/Region	Continent	Population	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecovered	NewRecovered	ActiveCases	Serious,Critical	Tot Cases/1M pop	Deaths/1M pop	TotalTests	Tests/1M pop	WHO Region
5	Mexico	North America	129066160.0	462690	6590.0	50517.0	819.0	308848.0	4140.0	103325.0	3987.0	3585.0	391.0	1056915.0	8189.0	Americas
7	Chile	South America	19132514.0	366671	NaN	9889.0	NaN	340168.0	NaN	16614.0	1358.0	19165.0	517.0	1760615.0	92022.0	Americas
9	Spain	Europe	46756648.0	354530	NaN	28500.0	NaN	NaN	NaN	NaN	617.0	7582.0	610.0	7064329.0	151087.0	Europe
17	Argentina	South America	45236884.0	228195	NaN	4251.0	NaN	99852.0	NaN	124092.0	1150.0	5044.0	94.0	794544.0	17564.0	Americas
27	Ecuador	South America	17668824.0	90537	NaN	5877.0	NaN	71318.0	NaN	13342.0	378.0	5124.0	333.0	258582.0	14635.0	Americas

Figura 1 - Análise descritiva do dataset antes do tratamento.

	Column1	Country/Region	Continent	Population	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecovered	NewRecovered	ActiveCases	Serious,Critical	Tot Cases/1M pop	Deaths/1M pop	TotalTests	Tests/1M pop	WHO Region
0	5	Mexico	North America	1290166160	400462	4973	48995	4973	312398	8588	47657	39870	31	4	1569150	121	America
1	7	Chile	South America	191325140	366671	2133	9262	75	321813	1859	18782	13580	1916	48	17606150	92022	America
2	9	Spain	Europe	467566480	354530	908	27855	146	154306	3930	53521	6170	758	59	70643290	151087	Europe
3	17	Argentina	South America	45236884	172306	4890	3179	120	74632	2057	1240920	11500	3808	70	7945440	175640	America
4	27	Ecuador	South America	176688240	81352	658	5523	8	35554	352	133420	3780	460	31	2585820	14635	America

Figura 2 - Análise descritiva do dataset após o tratamento.

Como o worldometer_data.csv apresenta apenas o acumulado, não temos muita noção da evolução dos casos, mas com os dados acumulados, podemos analisar que o país que apresentou a maior quantidade é casos é o México com 400 mil casos confirmados e a quantidade de mortos por 1 milhão de habitante é de 380 pessoas, enquanto o Chile que apresenta uma população aproximadamente 6 vezes menor que do México, apresentou a sua taxa de morte/1M é de mortes 484 pessoas.

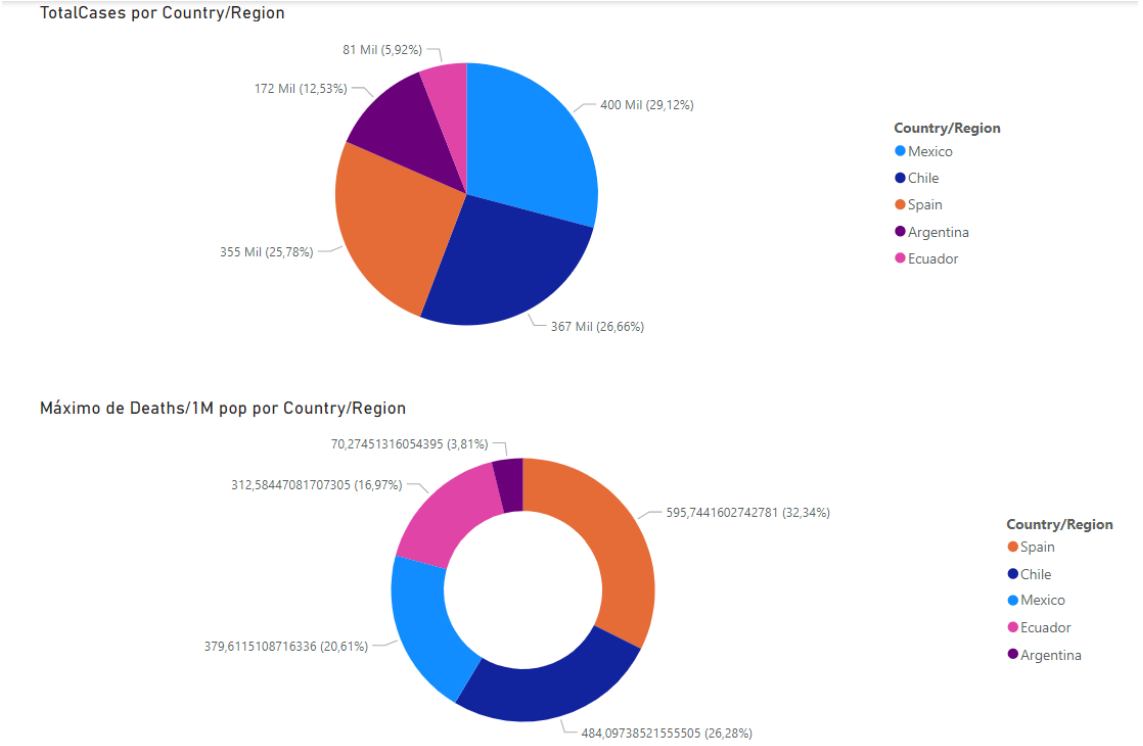

Figura 3 - Total de casos por país e mortes por milhão por país.

Dataset: covid_19_clean_complete.csv

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 940 entries, 6 to 49006
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Province/State  0 non-null      object
 1   Country/Region  940 non-null    object
 2   Lat             940 non-null    float64
 3   Long            940 non-null    float64
 4   Date            940 non-null    object
 5   Confirmed       940 non-null    int64
 6   Deaths          940 non-null    int64
 7   Recovered       940 non-null    int64
 8   Active          940 non-null    int64
 9   WHO Region      940 non-null    object
dtypes: float64(2), int64(4), object(4)
memory usage: 80.8+ KB
```

Possui apenas uma coluna com valores faltantes, que é a identificação do Estado, cada linha contém o dia a dia de cada país.

Ao analisar os dados obtidos do covid_19_clean_complete.csv, verificamos que o primeiro caso identificado foi em 02/02/2020 na Espanha, após o primeiro caso, o COVID-19 apresentou um crescimento dos contaminados mês a mês.

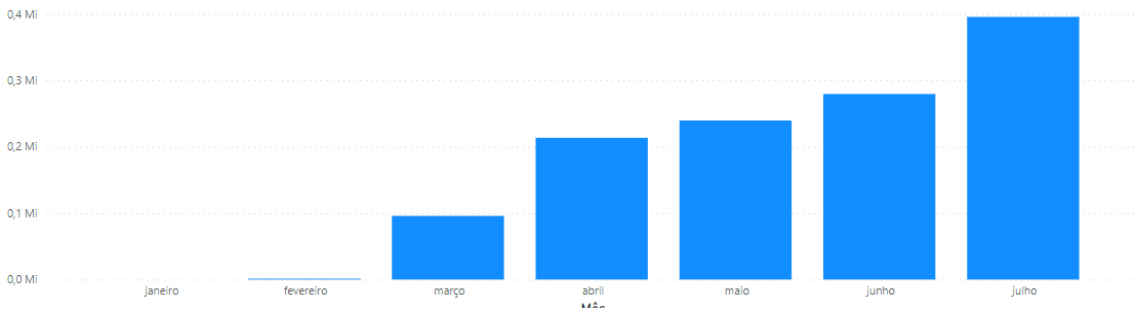

Figura 4 - Taxa de contaminação total por mês.

Junto com o crescimento de contaminados, verificamos que as mortes acompanharam esse crescimento, no gráfico abaixo, as barras representam as mortes e a linha, os casos confirmados.

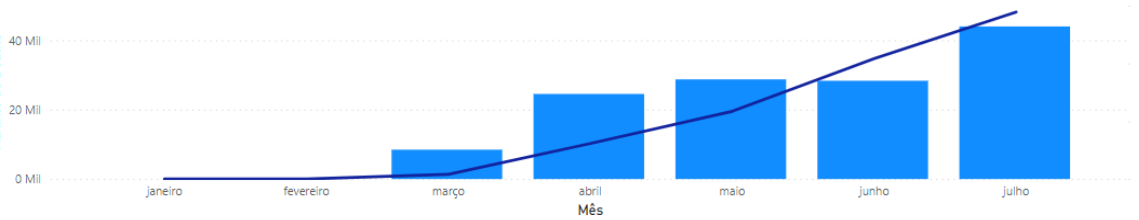

Figura 5 - Total de casos (linha sólida) e total de mortes (colunas) por mês.

Dataset: FullGrouped.csv (pegar as informações filtradas)

```
#   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Date            35156 non-null  object
 1   Country/Region  35156 non-null  object
 2   Confirmed       35156 non-null  int64
 3   Deaths          35156 non-null  int64
 4   Recovered       35156 non-null  int64
 5   Active          35156 non-null  int64
 6   New cases       35156 non-null  int64
 7   New deaths      35156 non-null  int64
 8   New recovered   35156 non-null  int64
 9   WHO Region      35156 non-null  object
dtypes: int64(7), object(3)
```

Não há valores nulos, é um dataset com sequência de dias, para análises temporais. Região, confirmados, mortes, recuperados, ativos, novos casos, novas mortes, novos recapados e região da OMS.

Dataset: country_wise_latest.csv

```
#   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Country/Region        5 non-null      object
 1   Confirmed             5 non-null      int64
 2   Deaths                5 non-null      int64
 3   Recovered             5 non-null      int64
 4   Active                5 non-null      int64
 5   New cases             5 non-null      int64
 6   New deaths            5 non-null      int64
 7   New recovered         5 non-null      int64
 8   Deaths / 100 Cases    5 non-null      float64
 9   Recovered / 100 Cases 5 non-null      float64
 10  Deaths / 100 Recovered  5 non-null    float64
 11  Confirmed last week   5 non-null      int64
 12  1 week change         5 non-null      int64
 13  1 week % increase     5 non-null      float64
 14  WHO Region            5 non-null      object
dtypes: float64(4), int64(9), object(2)
memory usage: 640.0+ bytes
```

Não há valores nulos, cada linha indica um país, com um acumulado do período, mas não está explícito qual é o período utilizado no dataset, vamos averiguar se o período do dataset fullgroupe.csv corresponde as informações acumuladas no country_wise_latest.csv. Há um acumulado de casos confirmados, mortes, recuperados, casos ativos, novos casos, novas mortes, novas recuperações, entretanto não sabemos se os novos casos são por dia, por semana. Óbitos a cada 100 casos, recuperação a cada 100 casos, mortes por 100 casos recuperados, confirmados da semana passada, mudança na semana, aumento na semana, região por OMS.

Utilizando os dados que foram fornecidos no dataset country_wise_latest.csv e obtendo as primeiras informações, ainda de um modo simples e com intuito exploratório.

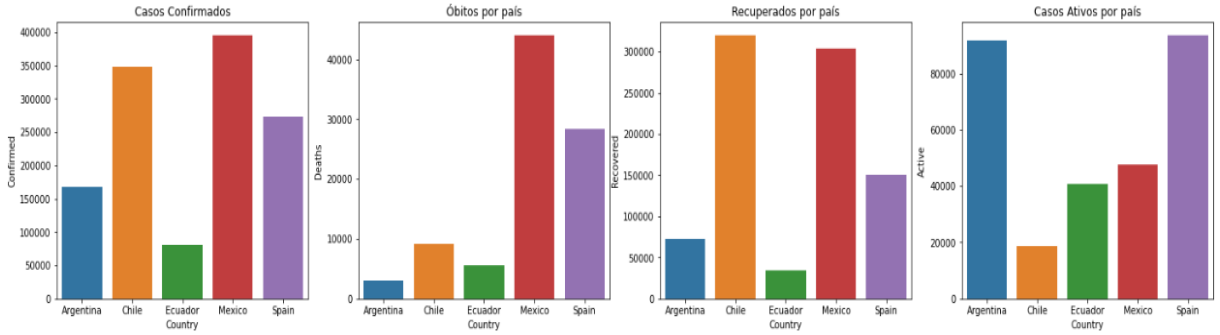

Figura 6 - Panorama geral dos totais por país.

Apresentando as primeiras análises feitas sobre o dataset, nos países selecionados, sem padrões para comparações, à primeira vista podemos observar que México e Espanha são os campeões nos casos de óbito.

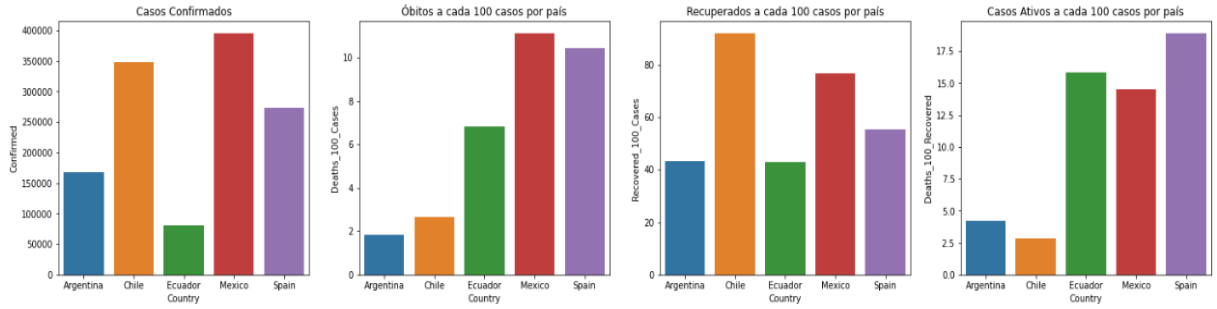

Figura 7 - Resumo das métricas por 100 casos confirmados.

Já com os um índice para cada 100 casos, já temos um linear a ser apresentado para comparações. E podemos observar que o equador aumentou na questão de óbitos.

Observando nesse gráfico, Espanha está na terceira posição de confirmados, mas o segundo em óbitos, primeiro em casos ativos e o terceiro em recuperados, mostrando um desempenho ruim contra a pandemia, visto que é um País com um fluxo grande de turistas, já o Chile é o segundo em casos confirmados, mas o quarto em óbitos e o primeiro em casos recuperados e o menor em casos ativos. Claro, que precisamos buscar as informações por tamanho populacional, que este dataset não fornece.

Tweets

Os tweets coletados da API do Twitter compreendem o período de 11 de fevereiro de 2020 a 31 de dezembro de 2021. Não há registros anteriores a 11 de fevereiro pois, o termo utilizado para realizar a busca dos tweets foi apenas “covid”. O termo COVID foi criado pela OMS apenas em 11 de fevereiro de 2020 e por isso não há tweets, anteriores a esta data, que utilizem o termo de busca utilizado.

Para a análise deste projeto serão, inicialmente, utilizados apenas dos tweets publicados entre fevereiro e julho de 2020. Este é o mesmo período de dados sobre a pandemia disponível no dataset do Kaggle. A busca foi limitada a 100 tweets por dia, o que dá um total de aproximadamente 3.000 tweets por mês e por país.

A análise preliminar mostra que alguns tweets estão em idiomas diferentes do espanhol.

6. DASHBOARD

Após a coleta dos dados que foram armazenados na pasta bronze, a limpeza dos dados que ficam na área Silver e chegamos aos dados Gold serão explorador no PowerBI, informações com base em taxa de incidência por 100mil/habitantes e podemos verificar pelo PIB no País.

Países selecionados:

- Espanha
- Equador
- Chile
- México
- Argentina

Os indicadores de saúde podem ser definidos como uma variável mensurada para monitorar o progresso ou avaliar o que funciona e o que não funciona quanto às ações relacionadas à saúde. Os indicadores possuem uma importância fundamental para a saúde pública e o surgimento de novos problemas. A disponibilidade de indicadores é essencial para se conhecer uma nova doença, sua

dinâmica de transmissão e seus impactos na saúde e na vida das populações. E a partir deles é possível traçar um plano de ação em saúde e adotar estratégias com vistas a controlar o avanço da epidemia de forma eficaz.

- · **Incidência acumulada**: é considerado uma média da intensidade com a qual a doença ocorre na população. Pode ser definida como a relação entre o número de casos novos de uma doença em um determinado intervalo de tempo e local.
- · **Mortalidade:** A taxa de mortalidade é a proporção entre a frequência absoluta de óbitos e o número de indivíduos expostos ao risco de morrer, no mesmo período de referência e no mesmo local. Total de óbitos pelo número de indivíduos expostos ao risco de morrer, os que tem e os que não tem a doença.
- · **Letalidade:** A letalidade expressa a gravidade de uma doença. A taxa de letalidade é a proporção entre a frequência absoluta de óbitos pelo número de indivíduos com determinada doença.
- · **Taxa de positividade:** A proporção de testes positivos dentre os testes realizados, também chamada de taxa de positividade. (coletar a informação).
-
- **Média móvel de casos e de morte (7 ou 15 dias)**: é um indicador de tendência da doença, no período de 7 ou 15 dias.
- **Curva de evolução:** é o crescimento da doença ao longo do tempo**.**
- **Sazonalidade:** como a doença se comporta no decorrer do tempo.
- **Taxa de Contágio:**

REFERÊNCIAS

Barreto, Evandro Fernandes Ciência de Dados aplicada à pandemia do coronavirus no Brasil, uma análise socioeconômica. Universidade Estadual Paulista (Unesp), 2021. Disponível em: <http://hdl.handle.net/11449/213821>.

https://covid.saude.gov.br/