BLUEEDTECH

ANA CRISTINA CHAVES
ANTONIO DUARTE MARCOS JUNIOR
THIAGO CHARLES

BOOTCAMP - PESQUISA DE IMPACTO DO CORONAVÍRUS 2020

SPRINT 2

BRASIL

DEZEMBRO 2022

1. INTRODUÇÃO

Neste relatório são apresentados os principais resultados obtidos durante a segunda sprint do bootcamp. Para esta sprint os objetivos exigidos foram:

- Iniciar a estruturação dos dados em relação a estruturação e armazenamento;
- Desenvolvimento do modelo planejado;
- Avanço de dados complementares ao dataset inicial;
- Apresentação em desenvolvimento;
- Dashboard em desenvolvimento.

Desta forma este relatório foi dividido em quatro seções que mostram os resultados de cada um dos itens exigidos.

2. Estruturação dos dados

Nesta seção são apresentados os resultados relativos à primeira tarefa: Iniciar a estruturação dos dados em relação a estruturação e armazenamento.

Foi iniciada a estrutura inicial do data lake seguindo a arquitetura medallion. Na arquitetura medallion os dados são divididos em três diretórios principais: bronze, silver e gold. A seguir são descritos os conteúdos de cada um dos diretórios.

2.1. Diretório bronze

No diretório bronze são armazenados os dados na forma mais pura, da forma como foram obtidos originalmente. Dentro deste diretório foram criados os seguintes diretórios.

- ember
    - Contém os dados sobre os mercados de energia dos países do projeto obtidos do repositório da EMBER (empresa especializada no uso de dados no mercado de energia a nível mundial).
- kaggle
    - Dados sobre os casos de COVID que foram disponibilizados no Kaggle. Os dados trazem informações sobre os primeiros meses da pandemia de COVID ao redor do mundo. Exemplos de variáveis disponíveis são: número de casos diários, número de mortes diárias, número de recuperados e número de casos ativos.
- twitter
    - Conjunto de tweets publicados ao longo da pandemia que contém o termo COVID em seu conteúdo. Como o termo COVID foi criado pela ONS somente a partir de 10 de fevereiro de 2020 esta é a data inicial dos tweets obtidos. Os tweets foram coletados para cada um dos países do projeto.
- wikipedia

- Dados complementares obtidos da Wikipedia. Os dados obtidos foram: PIB de cada país em 2020 e população de cada país em 2020.

2.2. Diretório Silver

Este diretório contém dados com tratamento prévio. Foram filtrados os dados do diretório bronze referentes apenas aos países de interesse do projeto. Os dados também foram estruturados e organizados para facilitar consultas futuras. Os principais diretórios são:

- twitter
  - Contém os dados processados dos tweets coletados. Dentro deste diretório existem outros dois sub-diretórios:
    - covid
      - Tweets estruturados e organizados entre países
    - sentimental_analysis
      - Resultados da análise de sentimentos dos tweets
- forecasts
  - Contém os resultados da previsões realizadas pelos modelos
- kaggle
  - Dados sobre o cenário inicial da COVID obtidos a partir da filtragem dos dados disponibilizados no Kaggle.

Além dos diretórios referidos há arquivos no diretório sobre o mercado de energia, PIB e população de cada um dos países de interesse do projeto.

2.3. Diretório Gold

Neste diretório são disponibilizados os dados finais que serão utilizados na ferramenta de BI.

2.4. Operacionalização dos dados

Para automatizar o processo de coleta e armazenamento dos dados foi iniciado o projeto de uma data pipeline utilizando o Airflow. O Ariflow é uma ferramenta que permite o agendamento de tarefas (chamadas de DAG's) a serem executadas em intervalos de tempo pré determinados de forma automática.

A data pipeline está disponibilizada no repositório do Github no diretório datapipeline. No diretório estão os scripts criados para automatizar a coleta e armazenamento dos dados. No momento desta sprint apenas das DAG's necessárias para coleta dos tweets está disponível. Ao longo das próximas sprints outras DAG's serão implementadas.

3. Modelos desenvolvidos

Este projeto exige o desenvolvimento de modelos com objetivos distintos. A princípio foi identificada a necessidade de se desenvolver para realizar duas tarefas distintas: análise se sentimentos e previsão do número de casos de COVID em cada país. Portanto, nesta seção são apresentados os principais resultados dos modelos desenvolvidos para cada uma das tarefas mencionadas.

3.1. Análise de sentimentos

A análise de sentimentos dos tweets foi utilizado o modelo VADER que está disponível na biblioteca NLTK. Antes de se utilizar o VADER da NLTK foi necessário realizar a tradução dos tweets para o inglês. Isto foi necessário pois o VADER é um modelo treinado para textos em inglês.

Para realizar a tradução dos textos, inicialmente, foi utilizado o modelo Marian da biblioteca Spark-NLP, descrito no relatório da sprint anterior. Porém, tal modelo exigia demasiado poder computacional pois, o mesmo foi desenvolvido para utilizar processamento em paralelo em clusters avançados. Para solucionar o problema de tradução dos textos foi utilizada a API do Google Tradutor.

A consulta à API do Google Tradutor foi feita por intermédio da biblioteca Translators. Esta biblioteca disponibiliza métodos para realizar a consulta a diferentes serviços de tradução de texto online como: Google Tradutor e Bing Tradutor. O único inconveniente é que ela demanda conexão com a internet e o tempo de resposta com o servidor da Google depende da velocidade de conexão, o que torna o processo lento. Porém, esta alternativa ainda teve melhor desempenho de velocidade do que a alternativa utilizando o Spark-NLP.

Com os tweets pôde-se então realizar a análise de sentimentos utilizando o VADER do NLTK. 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. Por este motivo é o ideal para uso neste projeto.

3.1.1. Resultados

Figura 1 - Análise de sentimentos Argentina

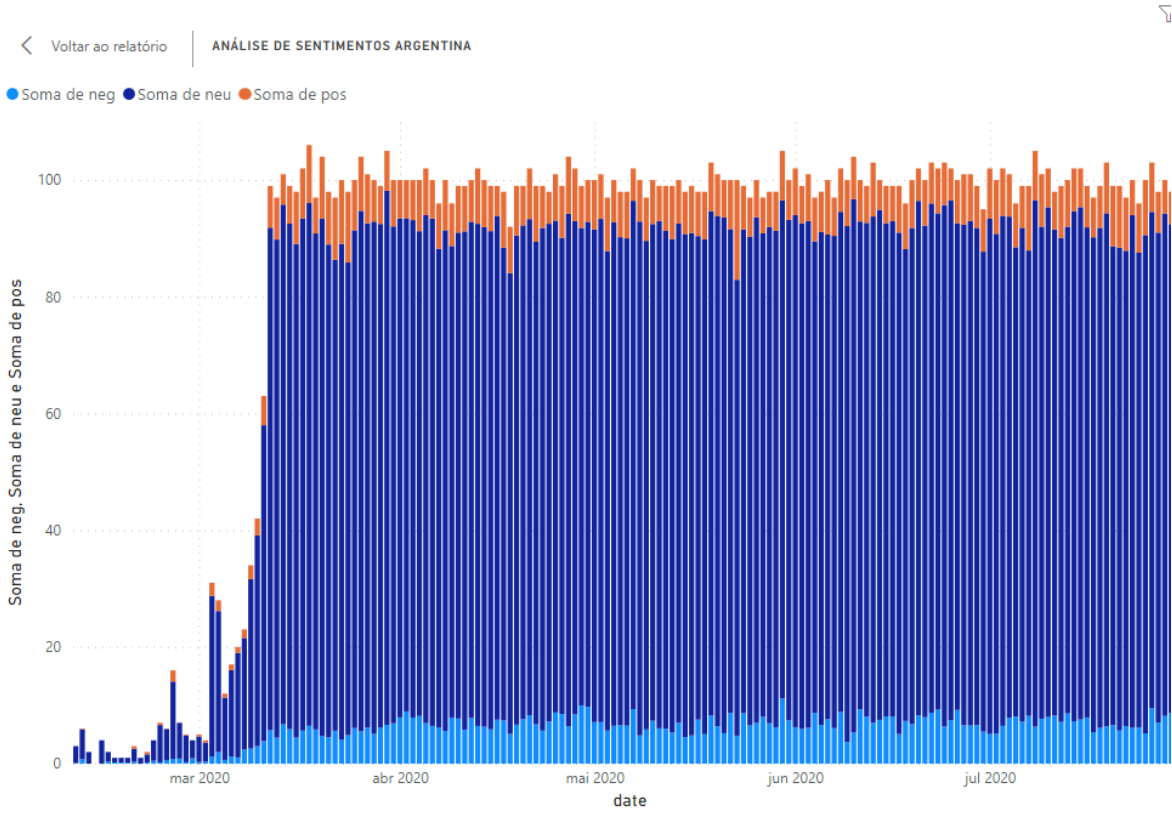


Figura 2 - Análise de sentimentos Chile

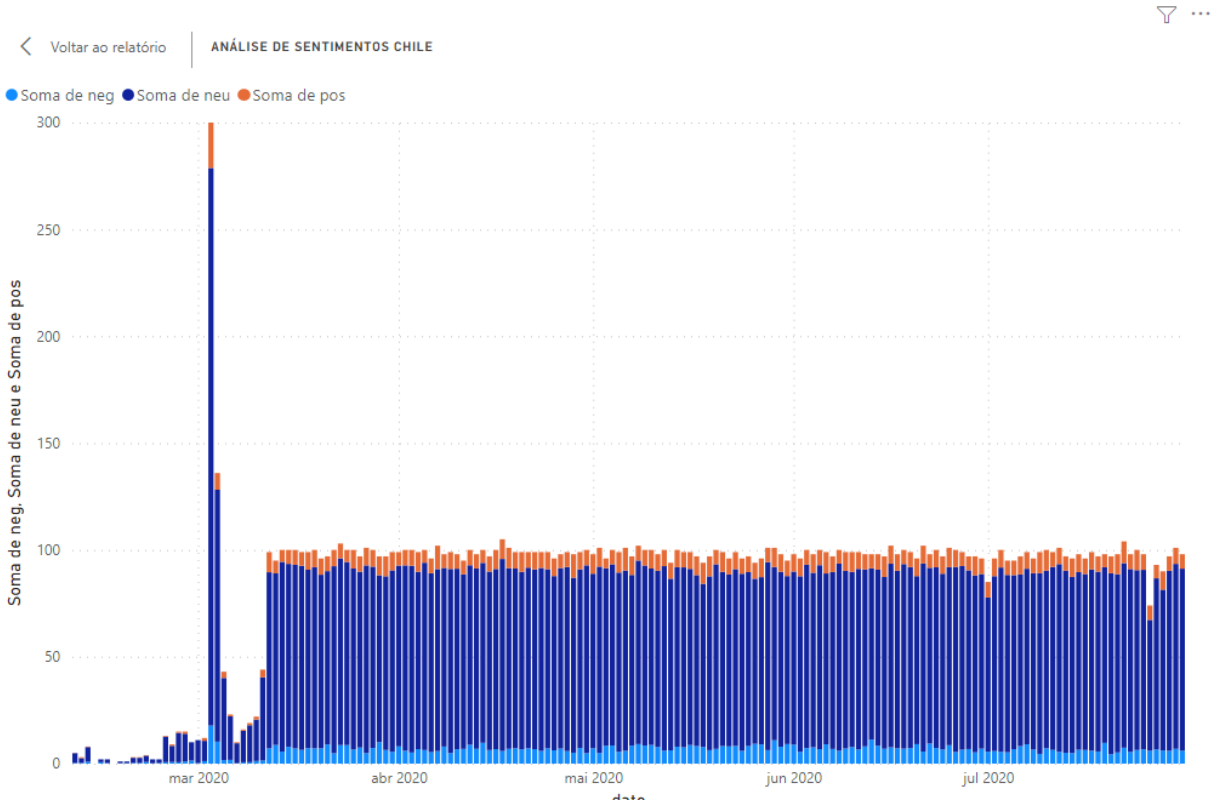


Figura 3 - Análise de sentimentos Equador

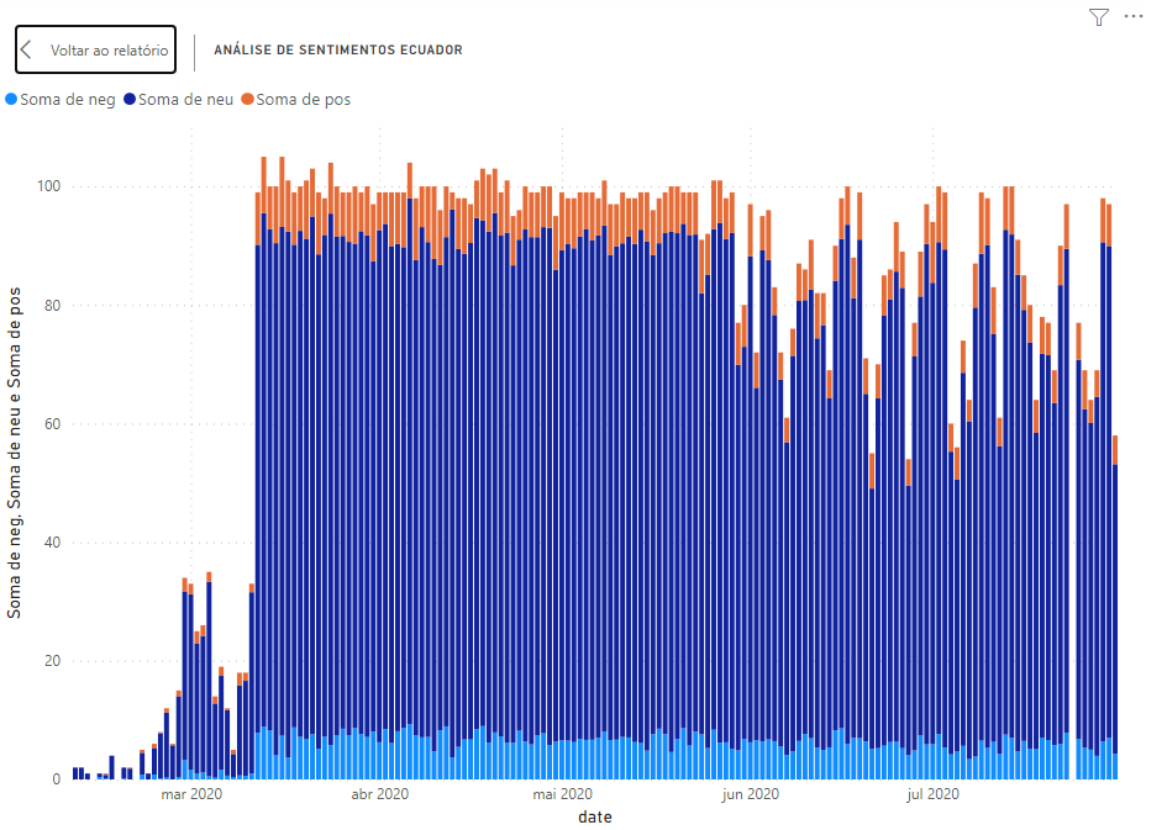


Figura 4 - Análise de sentimentos Espanha

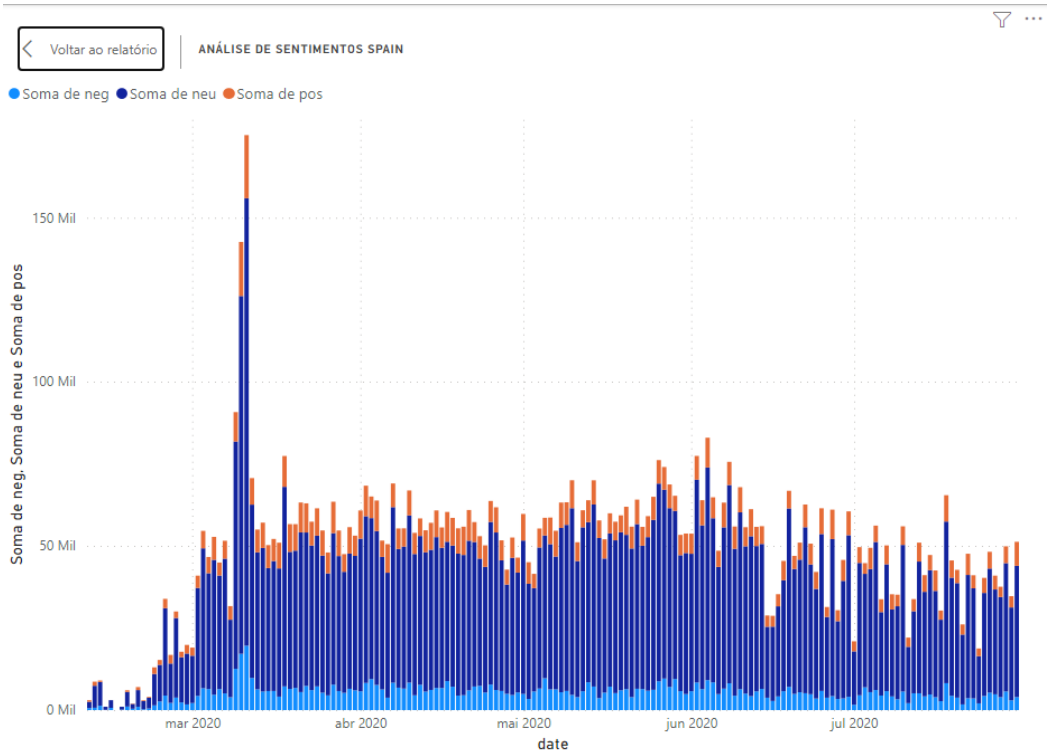
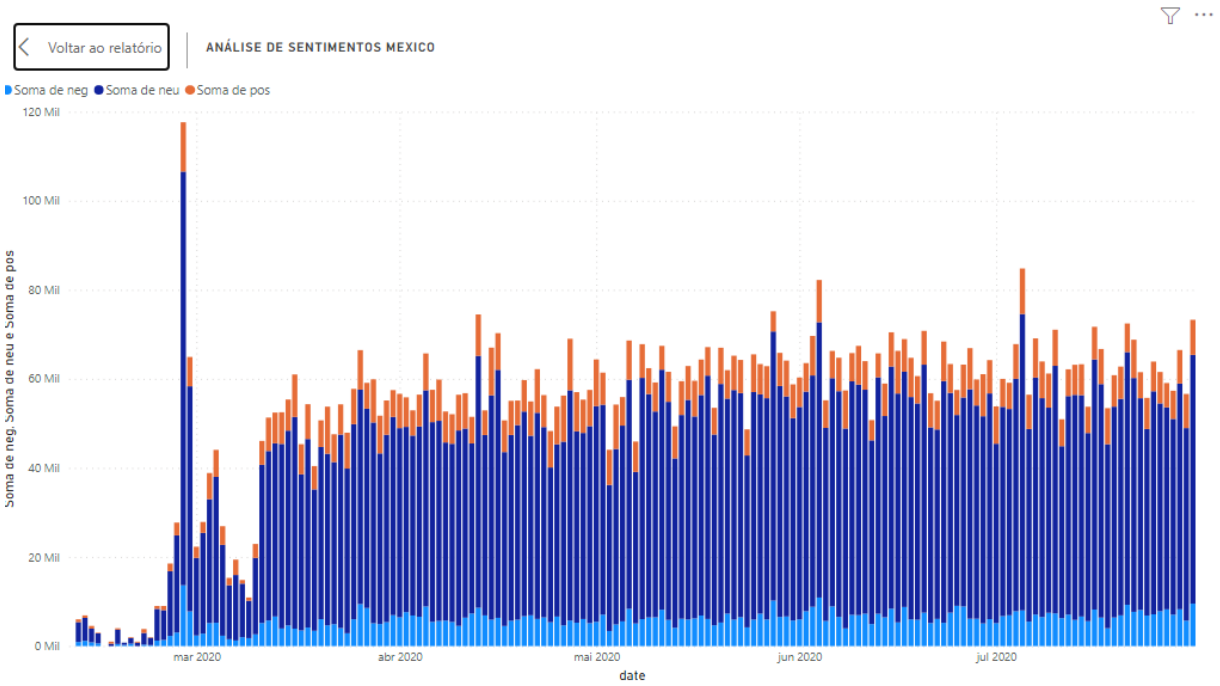


Figura 5 - Análise de sentimentos México



3.2. Previsão do número de casos

3.2.1. Modelo Auto Regressivo (AR)

3.2.1.1. Metodologia

Em um modelo Auto Regressivo (AR) como o nome sugere as variáveis dependente e independente são as mesmas. Ou seja, em um modelo AR é utilizada a dados da própria variável que se deseja prever como variável preditora. A diferença entre o dado predito e o preditor é um lag (intervalo) de tempo pré-estabelecido. Este tipo de modelo é equacionado da seguinte forma:

$$y(t) = \alpha + p_1 y_{t-1} + p_2 y_{t-2} + \ldots + p_i y_{t-i} + \varepsilon$$

Em que:

$y(t)$ é a variável predita para o tempo t;

α é o termo de interseção;

p_i é o coeficiente de regressão para um determinado lag i;

y_{t-i} é o valor da variável y para o instante de tempo t - i (lag i);

ε é o ruído branco com média igual a zero.

Para se determinar a ordem “p” do modelo AR analisa-se o gráfico de Autocorrelação Parcial (PACF) da série de dados. A PACF pode ser imaginada como a correlação entre a série e sua defasagem, excluindo-se o efeito de defasagens intermediárias. Então a PACF reflete a correlação direta entre um lag e a série. Assim, sabe-se se um lag é necessário ou não no termo AR.

A formulação matemática da PACF é a seguinte:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_i y_{t-1}$$

Em que:

y_t é o valor corrente da série;

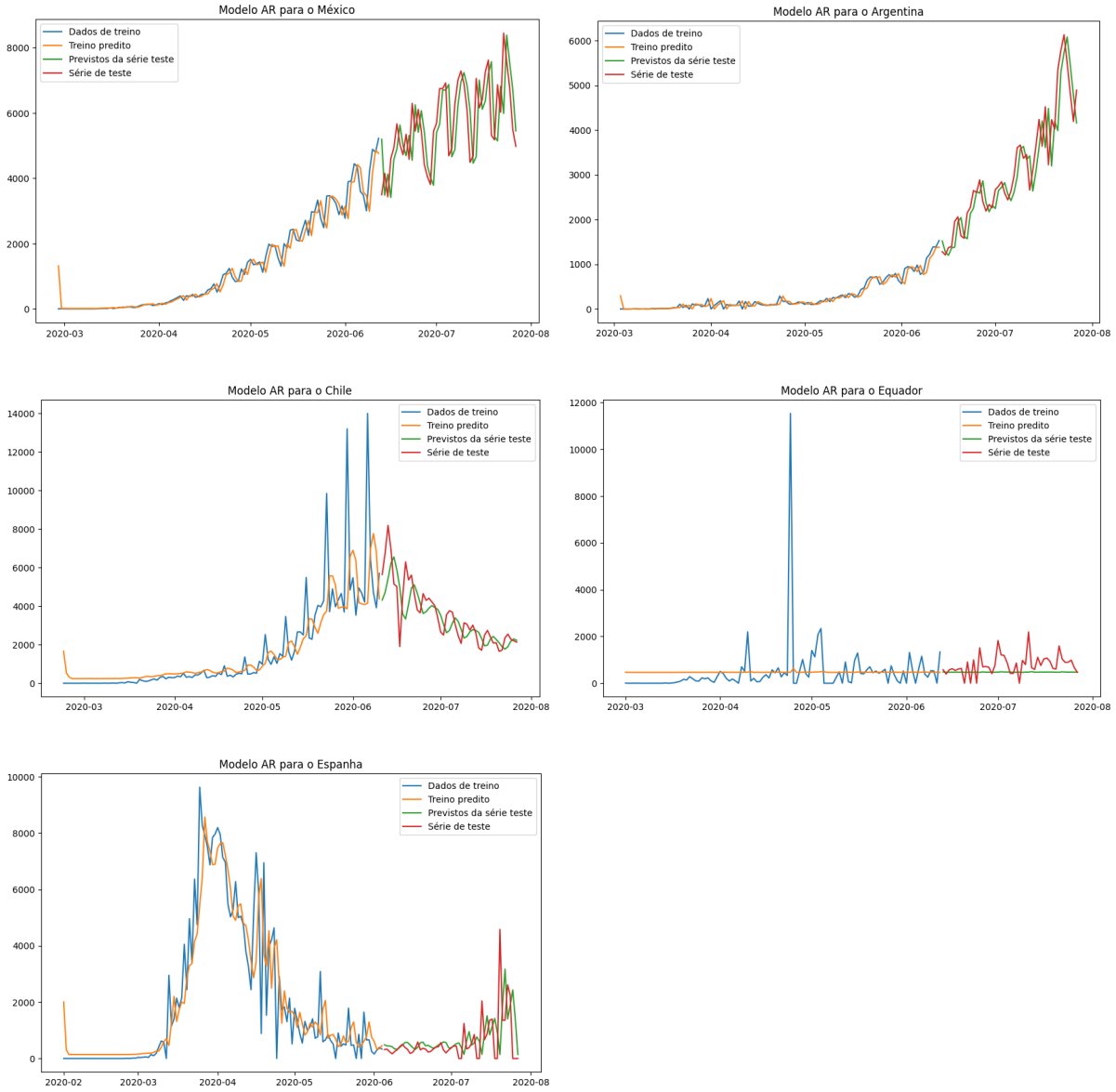
y_{t-i} é o valor de y no lag i

α_i é assumido como sendo o valor da PACF para um lag i.

3.2.1.2. Resultados modelo AR

Na Figura 6 é apresentado o resultado do modelo AR para cada um dos países do projeto. Apenas para a série do Equador foi observado um resultado muito ruim. O motivo do resultado obtido para o Equador é que a série de dados não possui “memória”, não há dependência temporal clara entre os registros.

Figura 6 - Resultado modelo AR para horizonte de 1 dia.



3.2.2. Modelo de Médias Móveis (MA)

3.2.2.1. Metodologia

No modelo de Médias Móveis (MA) é utilizada uma formulação em que os termos de erro, ϵ_t, são capazes de modelar o valor de y em t. Tal modelo têm estacionariedade fraca por ser uma combinação de ruídos brancos.

O modelo de MA é equacionado da seguinte forma:

$$y(t) = \alpha + \epsilon_t + q_1\epsilon_{t-1} + q_2\epsilon_{t-2} + \ldots + q_i\epsilon_{t-i} + \varepsilon$$

Em que:

ϵ_t é um ruído branco com média zero, refere-se aos erros de previsão passados;

q_i é o coeficiente de regressão para um determinado erro e lag i.

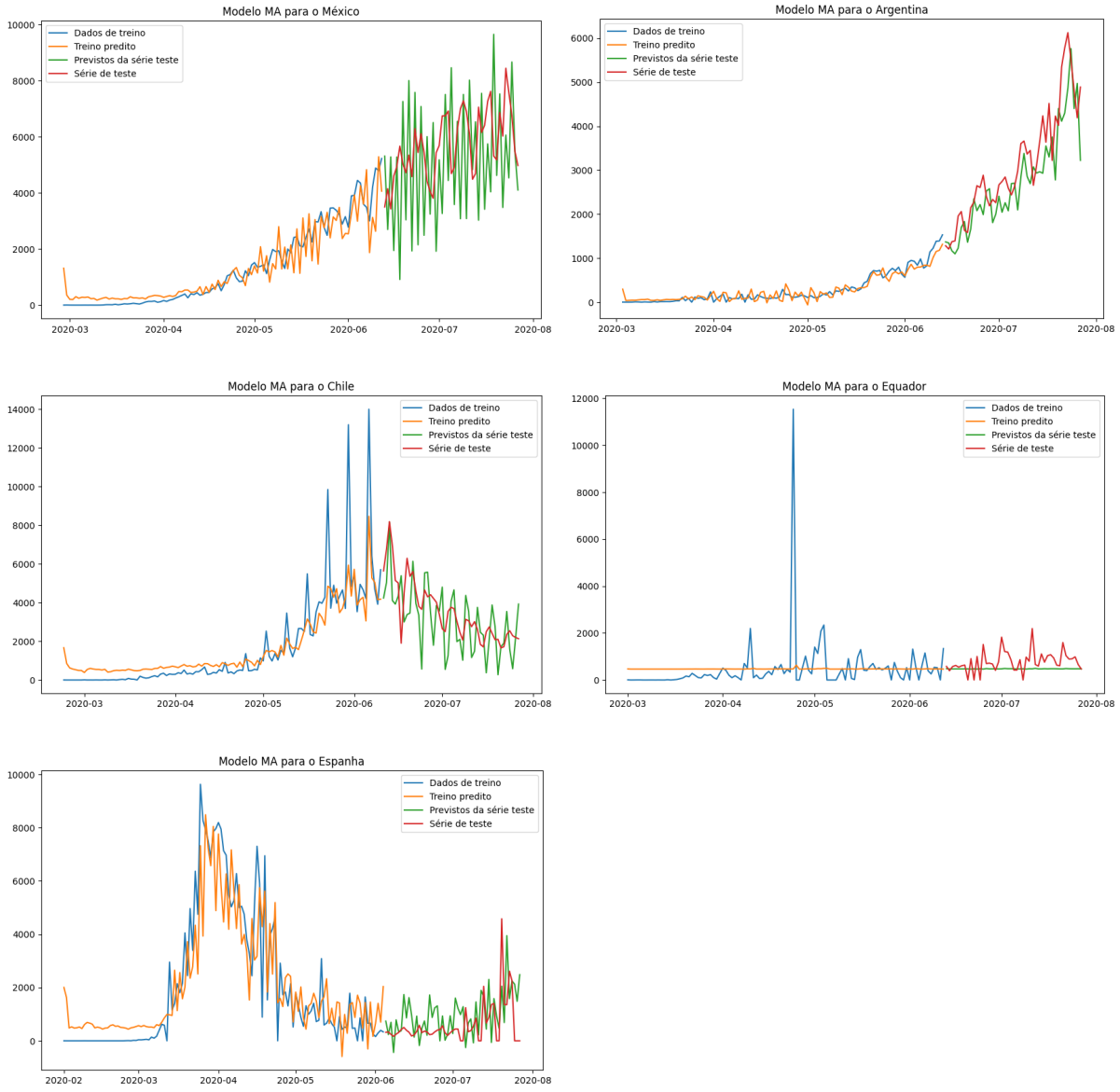
As demais variáveis têm a mesma conceituação do modelo AR.

Para determinar a ordem do termo “q” de um modelo MA é realizada a análise do gráfico da autocorrelação (ACF). A ACF mostra quantos termos MA são necessários para remover qualquer autocorrelação da série. A autocorrelação estima o quanto a série está correlacionada com uma determinada defasagem. Diferentemente da PACF, tanto as influências diretas e indiretas de um lag são contabilizadas na ACF. Sua formulação é a mesma da correlação de Pearson.

3.2.2.2. Resultados modelo MA

Na Figura 7 é apresentado o resultado dos modelos MA para cada um dos países do projeto. No geral os modelos MA tiveram performance inferior à do modelo AR.

Figura 7 - Resultados do modelo MA, previsão de 1 dia.



3.2.3. Modelo ARIMA

3.2.3.1. Metodologia

O modelo ARIMA (Autorregressivo Integrado de Médias Móveis) é a combinação dos modelos AR e MA com a adição de um termo de diferenciação. ARIMA é uma classe de modelos que "explica" uma determinada série temporal com base nos seus próprios valores passados. Os termos de um modelo ARIMA são p, d e q.

O termo "p" refere-se à componente AR do modelo, já mencionada anteriormente. O termo "q" refere-se à componente MA do modelo, também já mencionado. E o termo "d" é a ordem de diferenciação necessária para que a série temporal seja estacionária.

A prática mais comum para tornar uma série estacionária é diferenciá-la. Ou seja, subtrair o valor corrente do valor anterior. Para algumas séries é necessário realizar mais de uma diferenciação. Portanto, o valor de “d” é a quantidade mínima de diferenciações para que a série seja estacionária, caso ela já seja estacionário o valor de “d” é igual a zero.

No modelo ARIMA a série foi diferenciada pelo menos uma vez e sua formulação matemática é:

$$y(t) = \alpha + p_1 y_{t-1} + p_2 y_{t-2} + \ldots + p_i y_{t-i} + \epsilon_t + q_1 \epsilon_{t-1} + q_2 \epsilon_{t-2} + \ldots + q_i \epsilon_{t-i} + \varepsilon$$

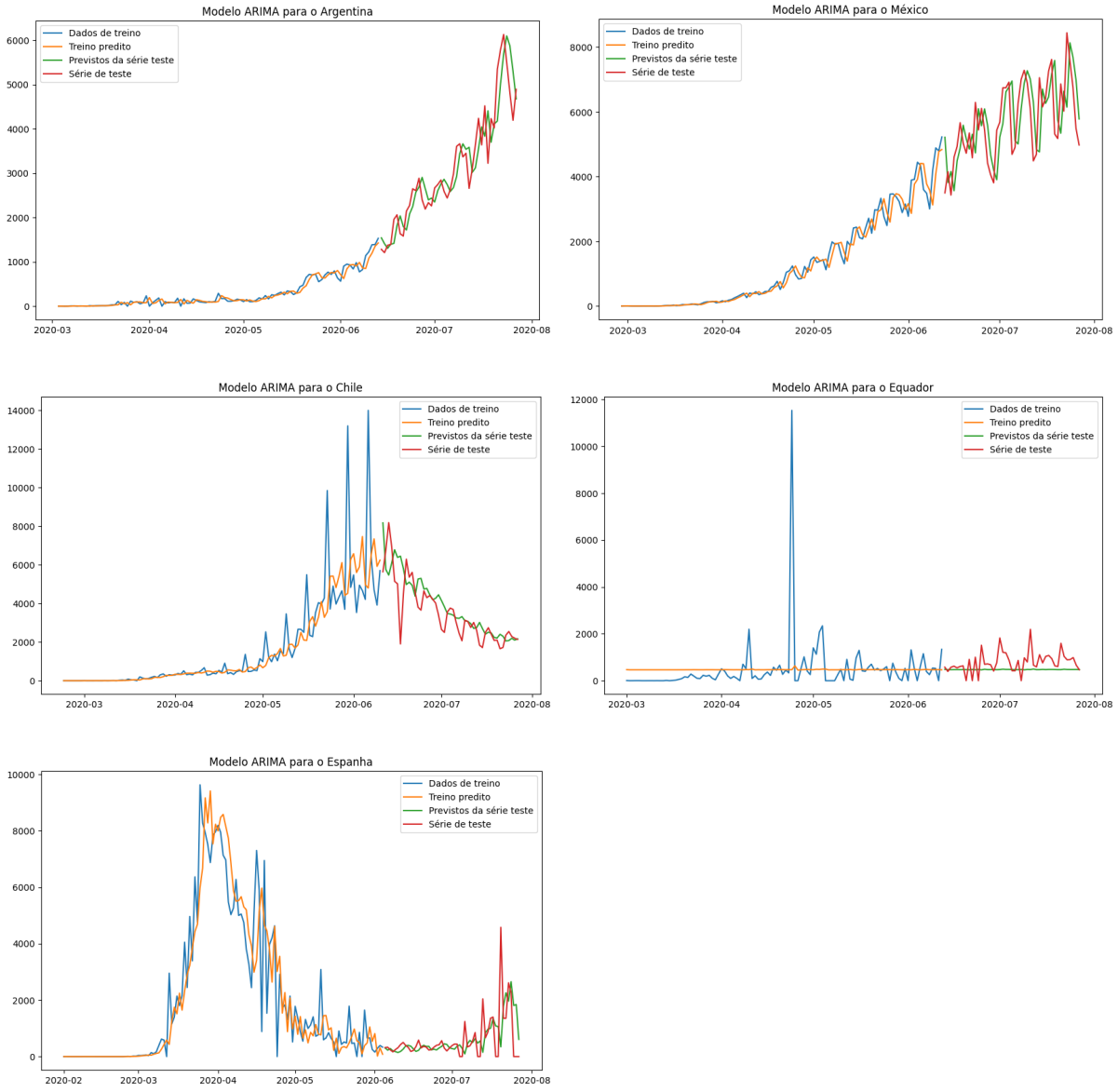
Em termos literais seria:

Valor predito de y no instante t = Uma constante + combinação linear de lags de y + combinação linear de erros de previsão com lags + ruído branco.

3.2.3.2. Resultados do modelo ARIMA

Na Figura 8 é apresentado o resultado, para a previsão de 1 dia, do modelo ARIMA de cada país. O modelo teve performance ligeiramente melhor que o modelo AR para a maioria dos países. Na Tabela 1 são listadas as principais métricas de comparação e a ordem dos modelos AR, MA e ARIMA.

Figura 8 - Resultados do modelo ARIMA, previsão de 1 dia, para a Argentina.



As métricas de avaliação utilizadas para os modelos ARIMA são:

AIC - O critério de informação de Akaike (AIC) é um estimador de erro de previsão e, portanto, qualidade relativa de modelos estatísticos para um determinado conjunto de dados.Dada uma coleção de modelos para os dados, o AIC estima a qualidade de cada modelo, em relação a cada um dos outros modelos. Assim, o AIC fornece um meio para a seleção do modelo. Quanto menor o valor de AIC melhor a performance do modelo

BIC - Em estatística, o critério de informação bayesiano (BIC) ou critério de informação de Schwarz (também SIC, SBC, SBIC) é um critério para seleção de modelo entre um conjunto finito de modelos; modelos com menor BIC são

geralmente preferidos. É baseado, em parte, na função de verossimilhança e está intimamente relacionado ao critério de informação de Akaike (AIC).

HQIC - Em estatística, o critério de informação Hannan-Quinn (HQC) é um critério para a seleção do modelo. É uma alternativa ao critério de informação de Akaike (AIC) e ao critério de informação bayesiano (BIC).

Tais métricas só podem ser utilizadas para comparar modelos que utilizem a mesma série de dados.

Tabela 1 - Métricas estatísticas de avaliação dos modelos ARIMA

País	Modelo	AIC	BIC	HGIC
México	AR(1)	1521,999	1529,990	1525,238
México	MA(9)	1669,284	1698,581	1681,158
México	ARIMA(2,2,1)	1491,282	1501,860	1495,567
Argentina	AR(1)	1233,110	1241,014	1236,311
Argentina	MA(7)	1297,174	1320,887	1306,779
Argentina	ARIMA(3,2,1)	1200,545	1213,620	1205,838
Chile	AR(3)	1933,746	1947,203	1939,203
Chile	MA(7)	1874,267	1898,490	1884,090
Chile	ARIMA(4,2,1)	1894,252	1910,289	1900,753
Equador	AR(1)	1773,446	1781,379	1776,660
Equador	MA(1)	1773,500	1781,433	1776,714
Equador	ARIMA(1,0,1)	1775,499	1786,076	1779,784
Espanha	AR(1)	2121,474	2132,787	2126,070
Espanha	MA(8)	2152,609	2180,893	2164,099
Espanha	ARIMA(3,2,2)	2086,828	2103,701	2093,682

3.2.4. Modelo Prophet

3.2.4.1. Metodologia

Além do ARIMA, também foi utilizado o Prophet.

Em 2017 o Facebook, necessitando de uma ferramenta de análise de séries temporais, criou o Prophet.

O Prophet é um framewok open source, o seu método de análise pode ser identificado com 4 componentes: sazonalidade, erros, tendência e feriados ou uma personalização de datas.

Algumas das vantagens do Prophet são a sua velocidade, facilidade de uso e sua personalização de “feriados”, no qual utilizamos o Lockdown de cada País.

Para trabalhar com o Prophet, devemos apenas trocar os nomes das colunas data para DS e o target para y, apenas com essas configurações o Prophet já é funcional.

Inicialmente, apenas fizemos o teste com o Prophet no modo default, o que gerou uma previsão com uma variação muito grande.

Para melhorar a previsão do Prophet, utilizamos o cross-validation para encontrar as melhores métricas do modelo para cada País, e aplicação das datas dos lockdown de cada País como holiday, obtendo valores de métricas melhores.

Os parâmetros que foram alterados, foram changepoint_prior_scale e seasonality_prior_scale.

O changepoint_prior_scale é responsável por determinar a flexibilidade da tendência, ele trabalha entre 0,001 - 0,5, se o valor é baixo a tendência é tratada como ruído e quando for alta é identificado tendência e em casos extremos a sazonalidade anual.

O seasonality_prior_scale é responsável por determinar a flexibilidade da sazonalidade, ele trabalha entre 0,01 – 10, um valor grande permite que a sazonalidade se ajuste a grandes flutuações, um valor pequeno diminui a magnitude da sazonalidade.

3.2.4.2. Resultados do Prophet

A figura 9 representa a previsão de novos casos da Espanha, com o período de 14 dias. Os pontos pretos são os novos casos informados no DF, a linha azul é a previsão e a sombra azul é o intervalo de confiança.

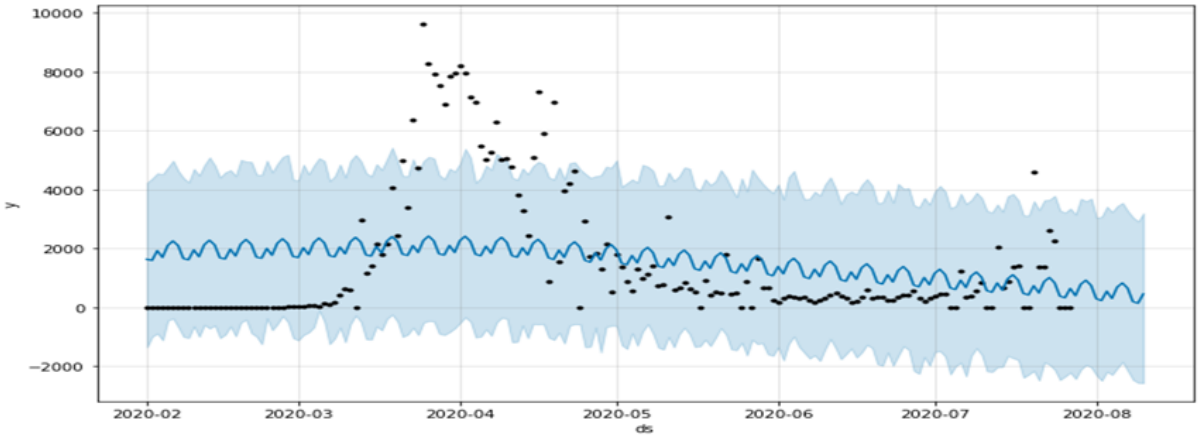


Figura 9 - Previsão da Espanha com o Prophet default.

Agora na figura 10, representa o Prophet com os melhores parâmetros e o lockdown. Podemos ver que a sombra azul é muito pequena comparada a figura anterior.

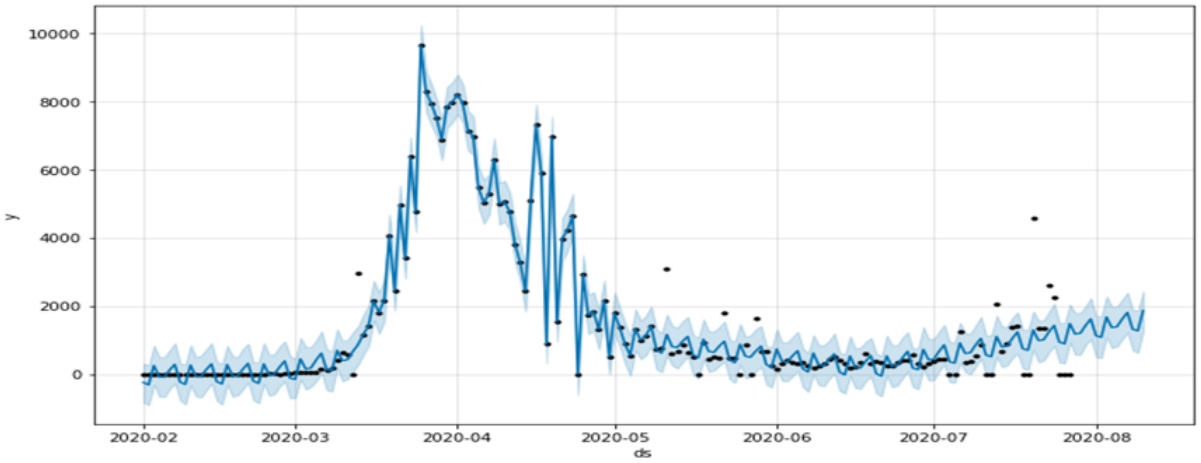


Figura 10 - Previsão da Espanha com o Prophet com os melhores parâmetros.

Esse caso se repetiu nos outros Países, apenas o Chile mostrou-se diferente, enquanto a previsão com os parâmetros no default indicava uma queda, junto com os valores reais, já o Prophet com os melhores parâmetros, mostrou uma previsão que indicava um aumento nos casos.

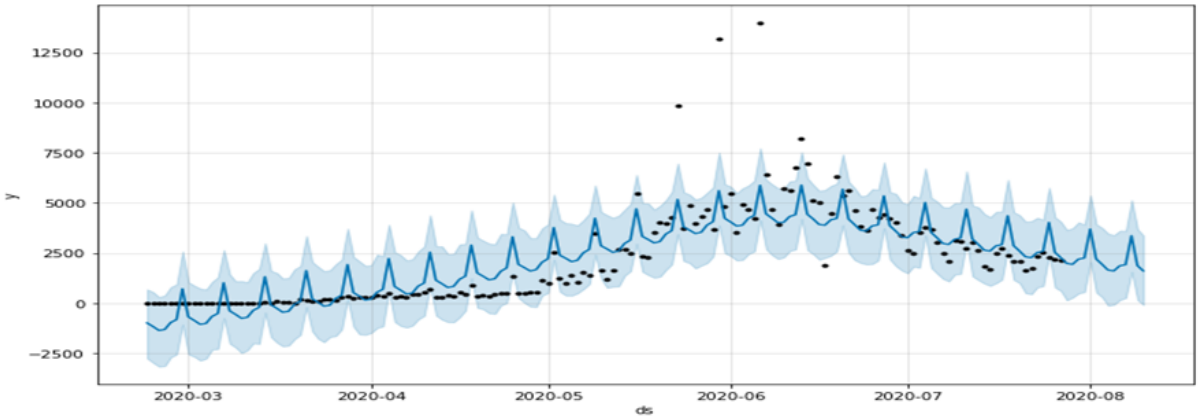


Figura 11 - Previsão do Chile com o Prophet default.

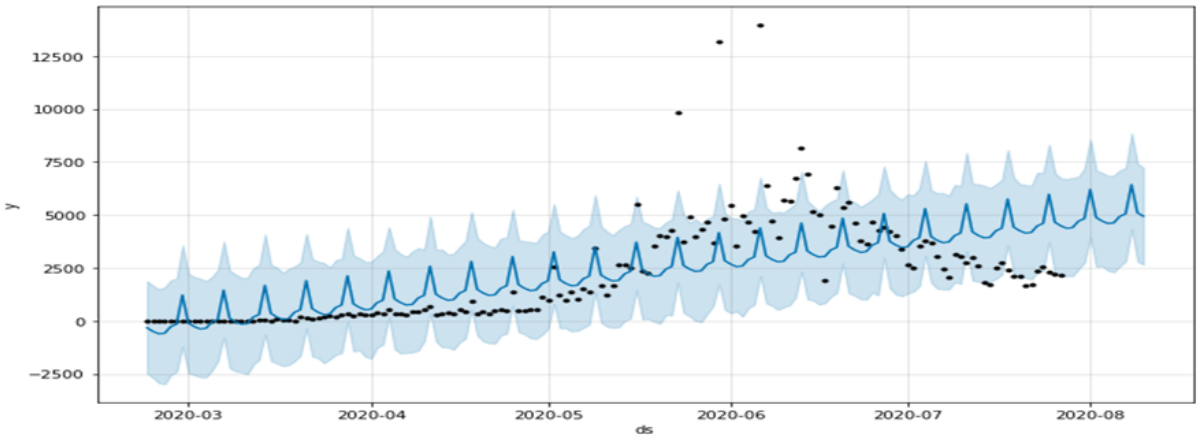


Figura 12 - Previsão do Chile com o Prophet com os melhores parâmetros.

Na Tabela 2 segue as métricas do modelo referente a cada País. A previsão informada é referente a cada dia, com isso optamos apenas pela informação obtida de 7 dias de previsão.

Tabela 2 - Previsão do Chile com o Prophet com os melhores parâmetros.

País	horizonte	mse	rmse	mae	mdape	smape	coverage
Argentina	7 dias	192184.10	438.38	299.38	0.277	0.513	0.030
Espanha	7 dias	1672120334.26	40891.56	19684.17	0.835	1.274	0.276
Chile	7 dias	2936109.02	1713.50	1246.23	0.287	0.500	0.373
Equador	7 dias	312922.05	559.39	389.99	0.659	0.858	0.845
México	7 dias	483169.65	695.10	518.55	0.173	0.274	0.098

4. Dados Complementares

Para complementar o estudo foram coletados dados sobre o mercado de energia dos países em estudo e também informações sobre as datas de lockdown em cada país.

4.1. Mercado de energia

4.2. Datas de lockdown

As datas de ocorrência de lockdown foram consideradas importantes pois trata-se de uma das primeiras medidas adotadas pelos países na tentativa de conter o avanço da COVID. Cada país adotou esta medida em períodos diferentes, alguns mais de uma vez, e com grau de controle sobre o deslocamento da população diferente. A ideia do lockdown é que com restrição do fluxo de pessoa também haja diminuição da dispersão do vírus e consequentemente da taxa de contágio.

A informação das datas de início e fim dos lockdowns foram obtidas do site da Wikipedia pelo seguinte link: (*https://en.wikipedia.org/wiki/COVID-19_lockdowns*). Na

Tabela 3 estão registrados apenas os lockdowns a nível nacional de cada país nos primeiros meses da pandemia.

Tabela 3 - Dados dos lockdowns de cada país.

País	Data de Início	Data de Fim	Total de Dias
Argentina	19/03/2020	10/05/2020	52
Equador	16/03/2020	31/03/2020	15
México	23/03/2020	01/06/2020	70
Espanha	14/03/2020	09/05/2020	56

O Chile não adotou medidas rígidas de isolamento, a nível nacional, e ao longo do período algumas cidades em específico decretaram lockdown.

5. Apresentação e Dashboard

Dashboard iniciado no powerbi, carregado as planilhas:

- country_wise_latest_filtrado;
- full_grouped_filtro(1);
- Wordeometer;

Incluímos a planilha '*1_quarentena*', que consta as informações dos lockdowns que houve nos países;

Incluímos a planilha '*classificação*', que possui informações sobre a classificação do índice de felicidades dos países e '*nota'* é a nota para a classificação.

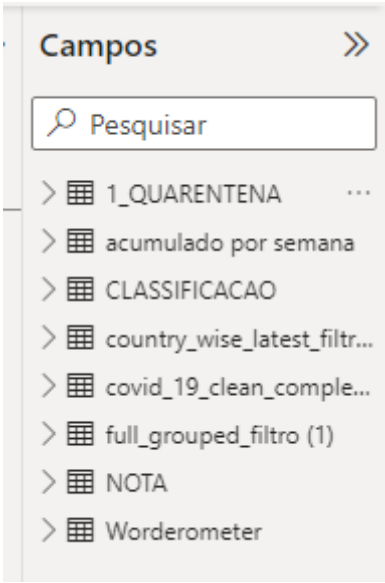
> *O* ***Relatório Mundial da Felicidade*** *(em inglês:* ***World Happiness Report*** *) é uma medição da felicidade publicado pela Rede de Soluções para o Desenvolvimento Sustentável da ONU (SDSN, na sigla em inglês), com base em dados coletados pelo Gallup World Poll.[1] O relatório, cuja primeira edição foi divulgada em 2012 com base em dados de 2011, é feito por especialistas independentes e em 2021-22 foi editado pelo professor Saïd Jan-Emmanuel De Neve da Universidade de Oxford. (Wikipédia)*

Utilizando dessa classificação, queremos observar o índice de felicidade no período pré e pós covid/19, se os países obtiveram alteração na classificação, e como se esse índice reflete na análise de sentimentos que iremos efetuar com dados do *twitter*. Temos que observar que os dados apresentados no ano, são de um período anterior, por exemplo 2019, ele reflete entre os anos 2016 a 2018, neste caso, o índice de 2022 está refletindo os anos de 2019 a 2021(antes e durante a pandemia).

Tabela 4 - Ranking de felicidade dos países do projeto.

País	2019 (2016-2018)	2020 (2017-2019)	2021 (2018-2021)	2022 (2019-2021)
Espanha	30	28	27	29
México	23	24	36	46
Chile	26	39	43	44
Equador	50	58	66	76
Argentina	47	55	57	57

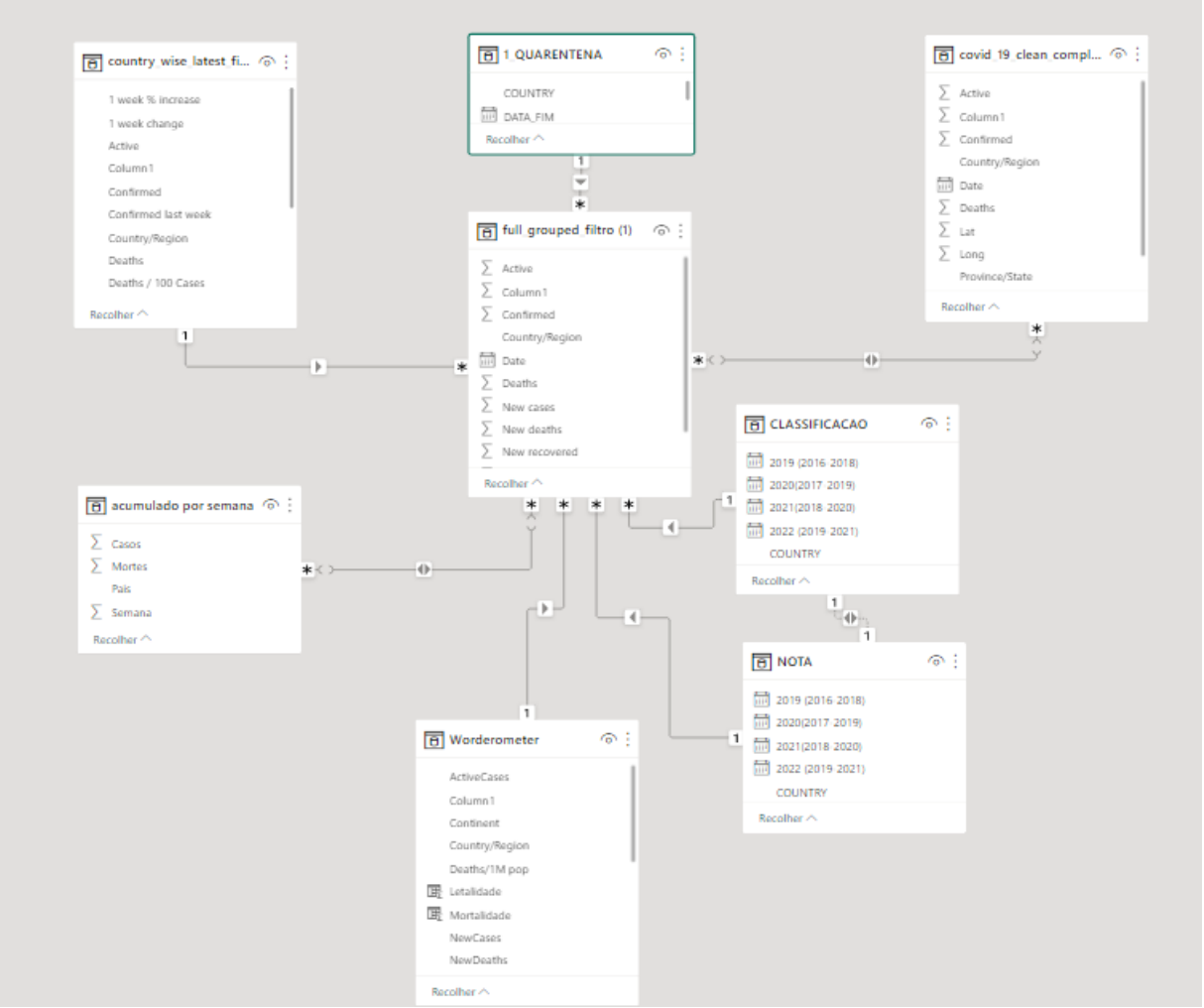
Calculamos o acumulado por semana, que é a planilha do '*full_grouped_filtro'*, separados por semana, para verificarmos a média de casos e morte por semana;



Na planilha ‘Worderometer’. utilizamos o DAX, para calcularmos a:

- Letalidade: Divisão do total de mortes, pelo total de casos.
    - `Letalidade = divide(Worderometer[TotalDeaths], Worderometer[TotalCases])`
- Mortalidade: Divisão do total de mortes, pela população.
    - `Mortalidade= DIVIDE(Worderometer[TotalDeaths],Worderometer[Population]`
- Taxa de positividade: divisão do total de casos positivos, pelo total de testes.
    - `Taxa Pisitividade = divide(Worderometer[TotalCases],Worderometer[TotalTests])`

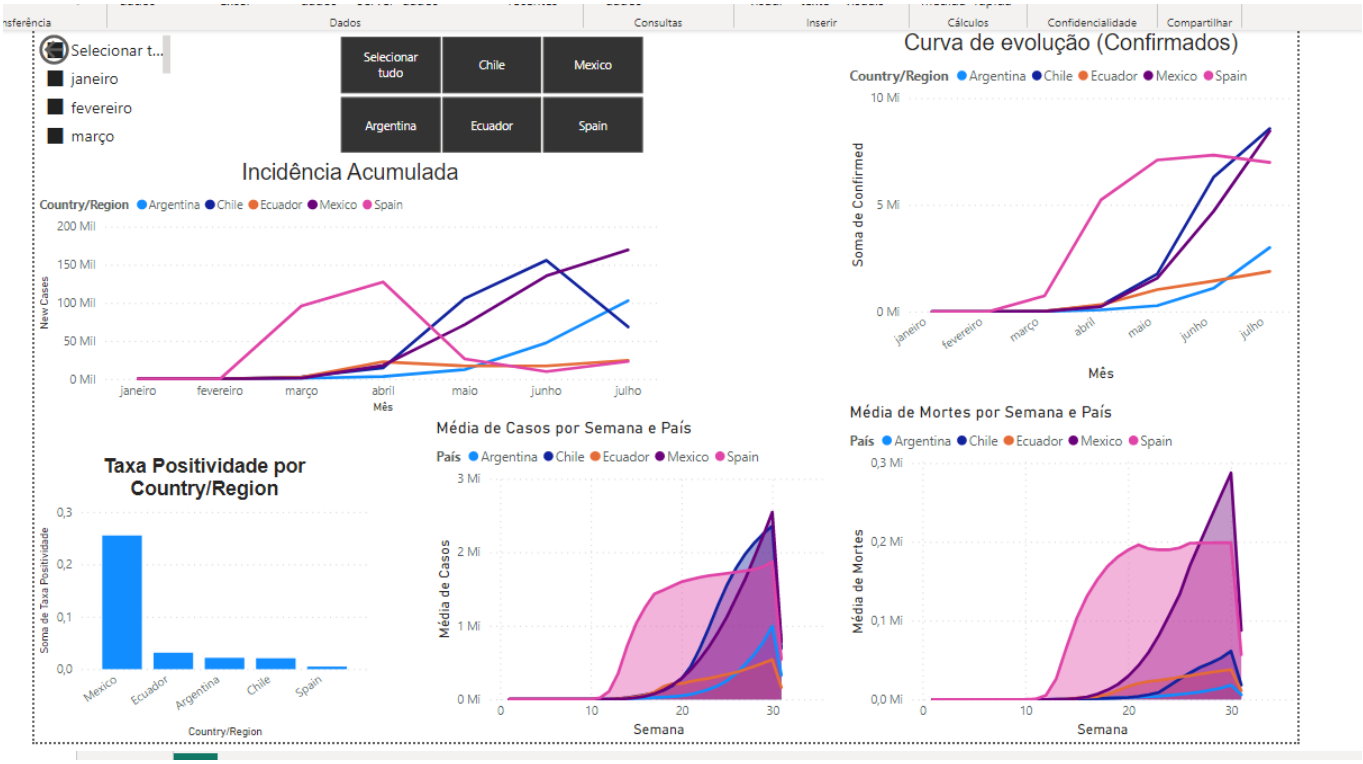
No relacionamento de planilhas, montamos da seguinte maneira:



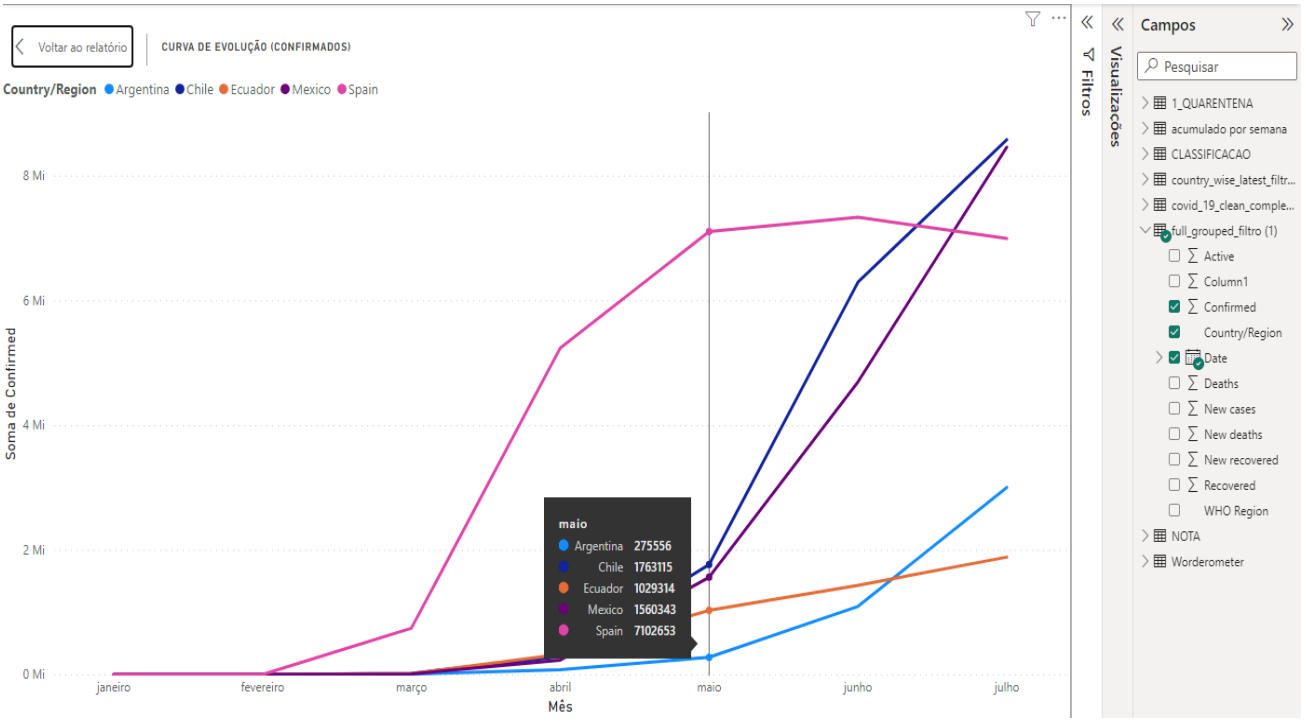
As tabelas foram relacionadas pelo País (Country/Region), formando um modelo star schema.

A princípio, montamos os gráficos conforme as tabelas incluídas referente ao kaggle, e conforme as taxas que iremos apresentar:

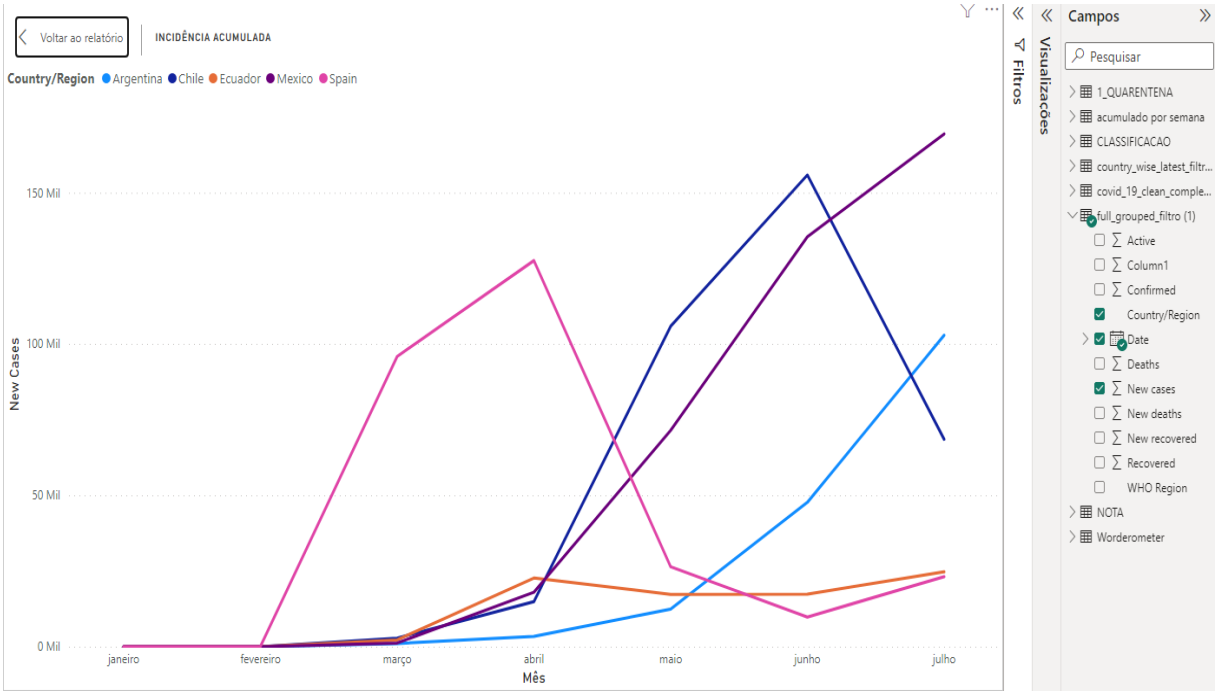
- Gráfico curva de evolução: apresenta os casos confirmados por dia e país.
- Gráfico: Incidência acumulada: mostra os novos casos, por dia e país.
- Gráfico: Taxa de positividade: São o total de testes realizados, pela quantidade de casos confirmados.
- Gráfico média de casos e morte por semana: é a média dos casos acumulados na semana, entretanto, teremos que recalcular esses índices, visto que pegamos as informações acumuladas, e não os casos confirmados no dia.



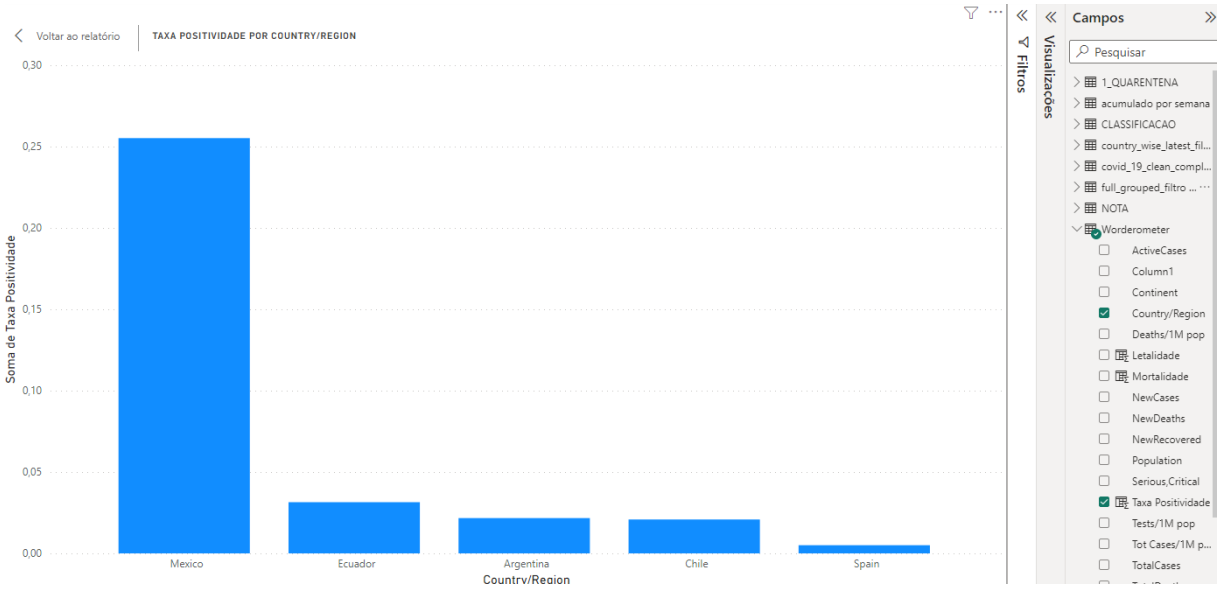
- Gráfico: Curva de evolução:



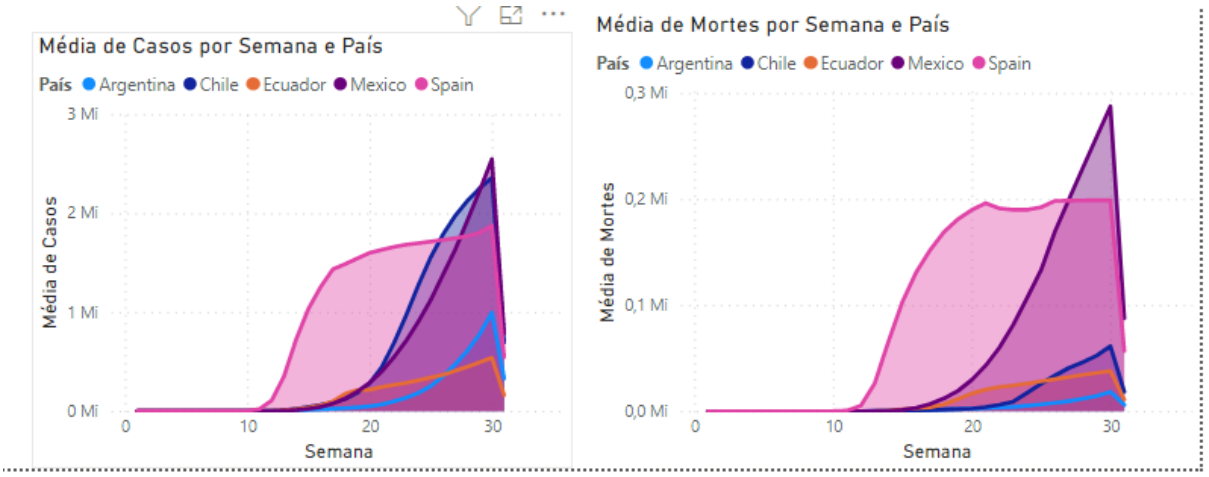
- Gráfico: Incidência acumulada:



- Gráfico: Taxa de positividade



- Gráfico média de casos e morte por semana (Gráfico será recalculado e alterado)



Filtros: O dashboard possui filtros por mês e país para facilitar a visualização específica.

REFERÊNCIAS

Wikipedia. **World Happiness Report**. Disponível em: <(https://en.wikipedia.org/wiki/World_Happiness_Report)>. Acessado 05/12/2022.

Wikipedia. **COVID-19 lockdowns**. Disponível em: <(*https://en.wikipedia.org/wiki/COVID-19_lockdowns*)>. Acessado 05/12/2022.