BLUE EDTECH

ANA CRISTINA CHAVES
ANTONIO DUARTE MARCOS JUNIOR

BOOTCAMP - PESQUISA DE IMPACTO DO CORONAVÍRUS 2020

SPRINT 5

BRASIL

FEVEREIRO 2023

1. Introdução

Neste relatório são apresentados os principais resultados obtidos durante o projeto proposto pelo bootcamp.

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

2. Contextualização

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

A COVID-19 é uma doença infecciosa causada pelo coronavírus SARS-CoV-2 e tem como principais sintomas febre, cansaço e tosse seca entre outros sintomas e é altamente contagiosa.

É uma doença que atingiu o patamar mundial, foram meses até surgir a primeira vacina, o planeta teve muitos impactos, as pessoas preocupadas com seu futuro, houveram lockdown, E vamos apresentar as análises feitas desse período e como os dados foram importantes para o controle da doença.

3. Arquitetura

3.1. Fonte de dados

3.1.1. Jhons Hopkins

A Jhons Hopkins é uma universidade de Baltimore, Maryland, Estados Unidos. Foi fundada em 1875 seguindo um modelo de ensino que dá ênfase à pesquisa científica. É reconhecida como umas das instituições acadêmicas mais importantes do mundo e uma das mais prestigiadas nos Estados Unidos. Ao longo da pandemia de COVID-19 se tornou uma das referências no assunto. Ela coleta dados de casos de COVID-19 de quase todos os países e os disponibiliza gratuitamente em seu repositório público no GitHub. O link para acesso é o seguinte: https://github.com/CSSEGISandData/COVID-19

3.1.2. API Twitter

O Twitter é uma das maiores redes sociais da atualidade. É uma rede social e serviço de microblog, que permite aos usuários enviar e receber atualizações pessoais de outros contatos. Foi criado e lançado em 2006 nos Estados Unidos. Na era do Big Data o Twitter conta com um dos maiores acervos de dados públicos. Através de sua API é possível consultar o histórico de tweets publicados em todo o planeta sobre os mais diversos assuntos. Com o caso da pandemia de COVID-19 não seria diferente. Os tweets publicados oferecem um panorama geral da opinião das pessoas sobre a pandemia, medidas públicas adotadas e suas expectativas em relação a mesma.

3.1.3. Woldometer

O Worodometer é um site que disponibiliza dados estatísticos, em tempo real, de diversos temas, entre eles a COVID-19. É operado pela empresa de dados Dadas. Foi criado em 2003 por Andrey Alimetoy e em 2011 foi eleito como um dos melhores sites de referência gratuira pela American Library Association. Atualmente está disponível em 31 idiomas.

3.1.4. Ember

Ember se declara como uma think tank de energia independente que usa insights orientados por dados para mudar o mundo do carvão para a eletricidade limpa. Reúnem, selecionam e analisam dados sobre o setor de energia global e seu impacto no clima. Reúnem uma equipe de especialistas de todo o mundo que entendem a rede elétrica. Em seu site disponibiliza dados de demanda de energia de países de todo o mundo.

3.2. Data Lake

3.2.1. 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.

3.2.1.1. covid_data

Anteriormente este diretório continha apenas os dados fornecidos no dataset original do Kaggle. Porém, foi realizada a coleta de dados a partir do repositório oficial da Universidade de Johns Hopkins. Esta é a mesma utilizada como fonte de dados do dataset original e considerada uma das referências sobre o tema COVID.

Dentro deste diretório são criados diretórios que armazenam os dados para cada dia de coleta. Os dados da Johns Hopkins são atualizados diariamente e disponibilizados no formato csv.

São fornecidos dados, consolidados a partir das fontes oficiais de diversos países, sobre os totais, até a data de coleta, de: casos, mortes, recuperados.

3.2.1.2. 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).

Os dados da EMBER são atualizados mensalmente.

3.2.1.3. twitter

Conjunto de tweets publicados ao longo da pandemia que contém o termo COVID em seu conteúdo. Como o termo COVID foi criado pela OMS 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.

O subdiretório query_covid refere-se a query utilizada para realizar a pesquisa de tweets. No caso foi pesquisado o termo “COVID”.

Para cada país é criado um subdiretório a saber: AR (Argentina), CH (Chile), EC (Equador), ES (Espanha), MX (México).

Os tweets, por país, são armazenados nos respectivos diretórios com a informação da data em que foram coletados. Em cada coleta são coletados tweets das 24 horas que antecedem.

3.2.2. 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:

3.2.2.1. covid_data

Contém os dados tratados de COVID. Em todos os subdiretórios os dados são salvos no formato parquet.

No subdiretório time_series são salvas as séries temporais de casos de COVID, para os países em estudo, desde o dia 01 de janeiro de 2020.

O subdiretório time_series_with_calc_fields contém as séries temporais de casos acrescidas dos seguintes campos calculados: novos casos, novas mortes, novos recuperados, casos ativos.

No subdiretório forecast são salvas as previsões de novos casos para cada dia.

3.2.2.2. twitter

Contém os dados processados dos tweets coletados. Os dados são organizados por query “query_covid” e por país, utilizando a nomenclatura apresentada anteriormente. Neste diretórios os arquivos são salvos no formato json.

No subdiretório tweet_processed são salvos os dados tratados dos tweets. A API do Twitter retorna campos que não são úteis para a proposta do trabalho, tais

campos são removidos e salvos apenas aqueles de interesse. Neste trabalho os campos salvos são: data do tweet, localização e texto do tweet.

No subdiretório tweet_sentiments são salvos os resultados da análise de sentimentos dos tweets.

O arquivo ember_electricity_monthly.csv contém dados mensais do setor de energia elétrica dos países selecionados.

3.2.3. Diretório Gold

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

3.3. Data Pipeline

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.

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 duas DAG's estão operacionais. A DAG necessária para extração e transformação dos tweets e a para extração, transformação e previsão de casos de COVID

A DAG Twitter realiza a extração dos tweets com o termo covid no texto para cada um dos cinco países de estudo. São coletados 100 tweets diários, são coletados tweets nas 24 horas que antecedem o momento de execução da DAG. Os tweets são salvos na pasta bronze no diretório específico. Em seguida, realiza a transformação do tweets para extração das informações: data de postagem, local de postagem e texto do tweet. O resultado desta tarefa é salvo na pasta silver. Finalizada a tarefa anterior é executada a tarefa de tradução de avaliação de sentimentos dos tweets cujo resultado é salvo em diretório específico.

Na DAG Covid são executadas as tarefas de extração, transformação e carregamento dos dados de Covid. A primeira tarefa faz o download dos dados mais recentes sobre os cacos de Covid do repositório da Johns Hopkins. A segunda tarefa realizada é a construção das séries temporais de cada país. O repositório da Johns Hopkins oferece um arquivo para cada dia, nesta tarefa os dados são organizados em um único arquivo no formato parquet. A próxima tarefa executa o cálculo dos seguintes campos: novos casos, novas mortes, novos recuperados, casos ativos. Após esta tarefa é executado o modelo de previsão para gerar a previsão de novos casos para os próximos 7 dias. A última tarefa realiza a carga dos dados no diretório gold.

4. Modelos de Machine learning

4.1. Análise de sentimentos

4.1.1. VADER

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]

4.2. Séries temporais

4.2.1. 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.

4.2.2. SARIMA

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

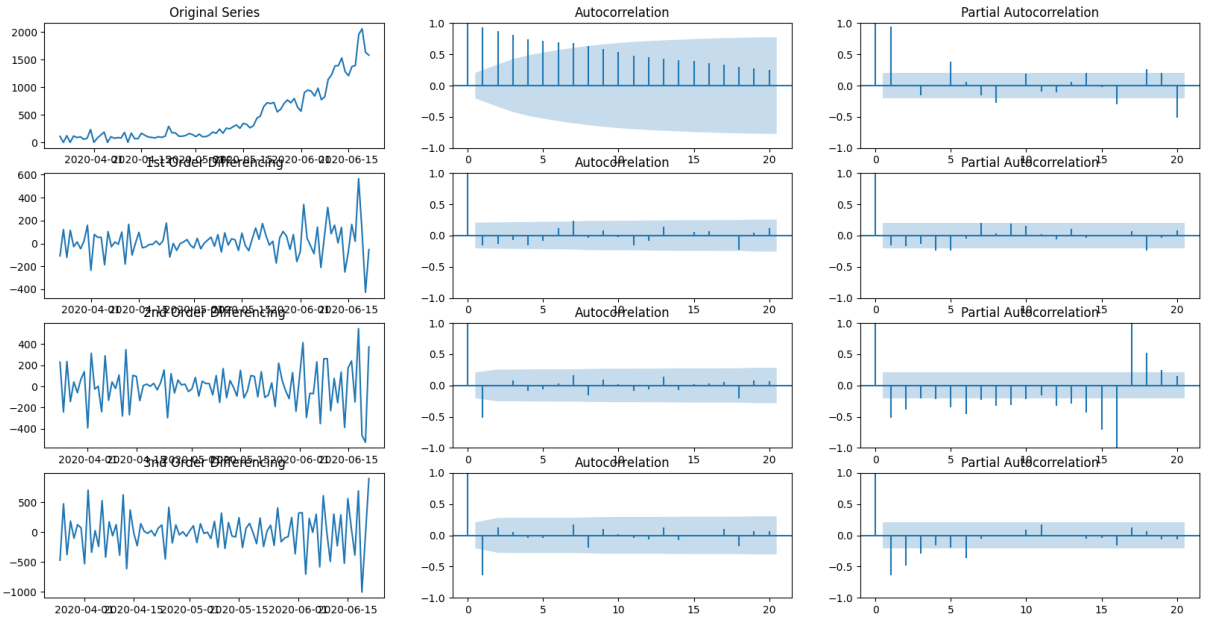
ARIMA(p, d, q)(P, D, Q)m

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

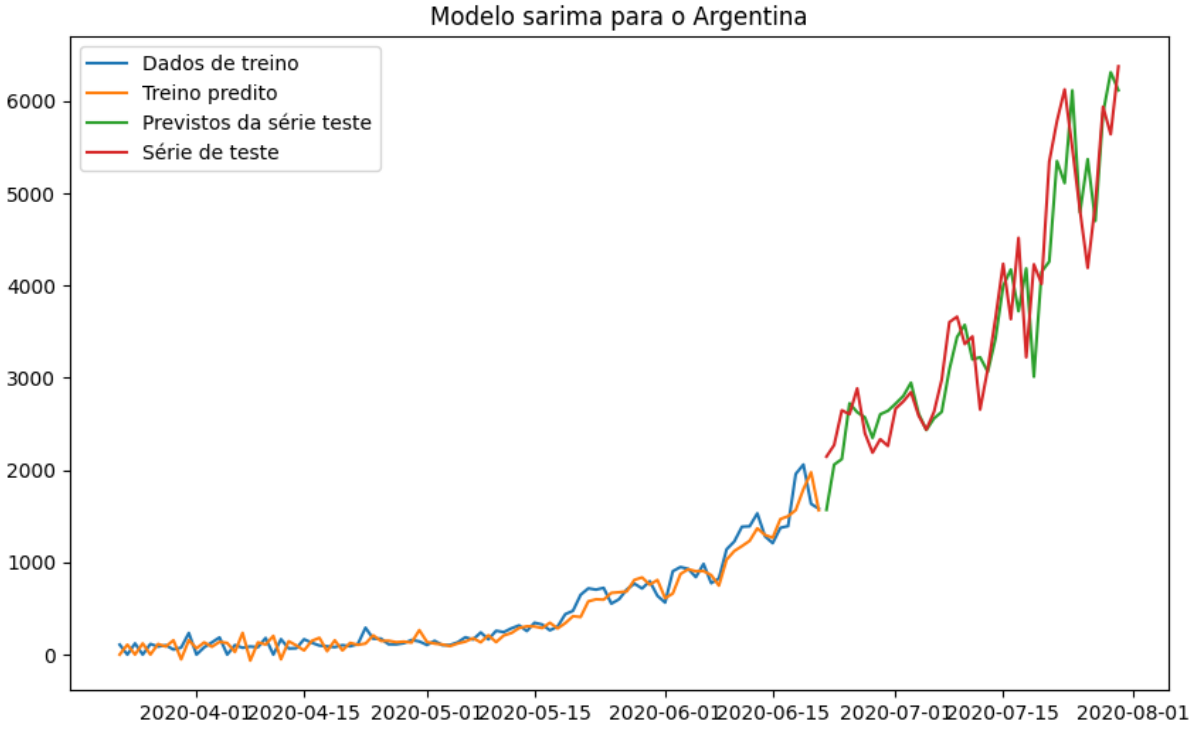
4.2.3. Resultados

Argentina

A figura a seguir mostra a série temporal de novos casos de COVID-19 para a Argentina (Original Series), autocorrelação e autocorrelação parcial (primeira linha da figura). A segunda linha tem a mesma ordem que a anterior mas com a diferenciação de primeira ordem. Na terceira linha são resultados para diferenciação de segunda ordem. E na quarta linha são resultados para diferenciação de terceira ordem.



Diferentes diferenciações foram aplicadas para se saber quantas seriam necessárias para saber quantas seriam necessárias para tornar a série estacionária. A autocorrelação fornece o valor do termo de médias móveis e o auto correlação parcial a do termo auto-regressivo do modelo. Não há sazonalidade expressiva na série e assim foi utilizado o modelo de ordem ARIMA(0, 1 2). Foram obtidos os seguintes valores: AIC 1045 e BIC 1057. A figura a seguir mostra o comparativo entre valores observados e previstos para o horizonte de 1 dia a frente.



A tabela a seguir apresenta o resumo das métricas de avaliação utilizadas para previsões em diferentes horizontes. Na coluna horizonte os valores com “h”

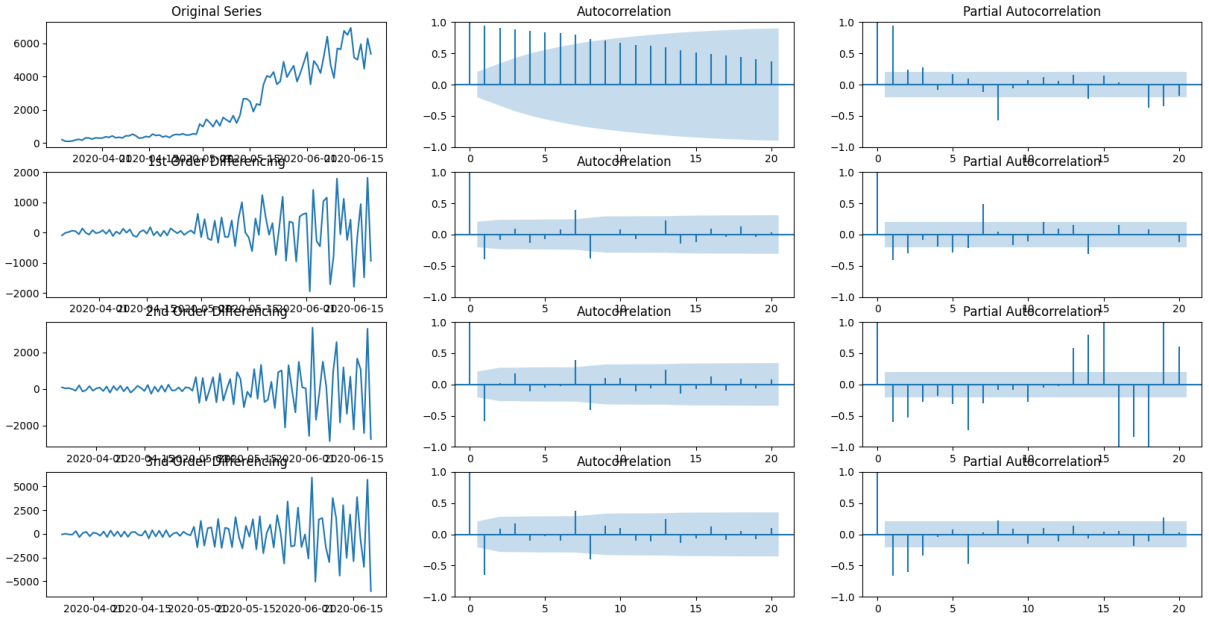
representam previsões n dias a frentes, por exemplo, h7 significa previsões para 7 dias após a data de origem. A linha “7 dias” tem o resultado da média móvel de sete dias.

Avaliação de desempenho para diferentes horizontes de previsão.

horizonte	r²	RMSE	MAE	MAPE	Correl
h1	0.799512	456.366641	356.123002	0.107029	0.901084
h2	0.792222	486.452365	401.421676	0.118690	0.905214
h3	0.699399	589.574870	469.699691	0.133367	0.850269
h4	0.671748	614.113379	510.370183	0.141608	0.833234
h5	0.653159	638.717413	521.053744	0.139860	0.821075
h6	0.728380	599.593259	477.834708	0.127318	0.869602
h7	0.689371	654.914946	543.771461	0.140742	0.840881
7 dias	0.810744	425.178066	338.956960	0.096718	0.913443

Para os demais foram utilizados os mesmos procedimentos adotados para a Argentina assim serão apresentados apenas os resultados.

Chile



Termo autoregressivo: 1

Número de diferenciações: 1

Termo de médias móveis: 1

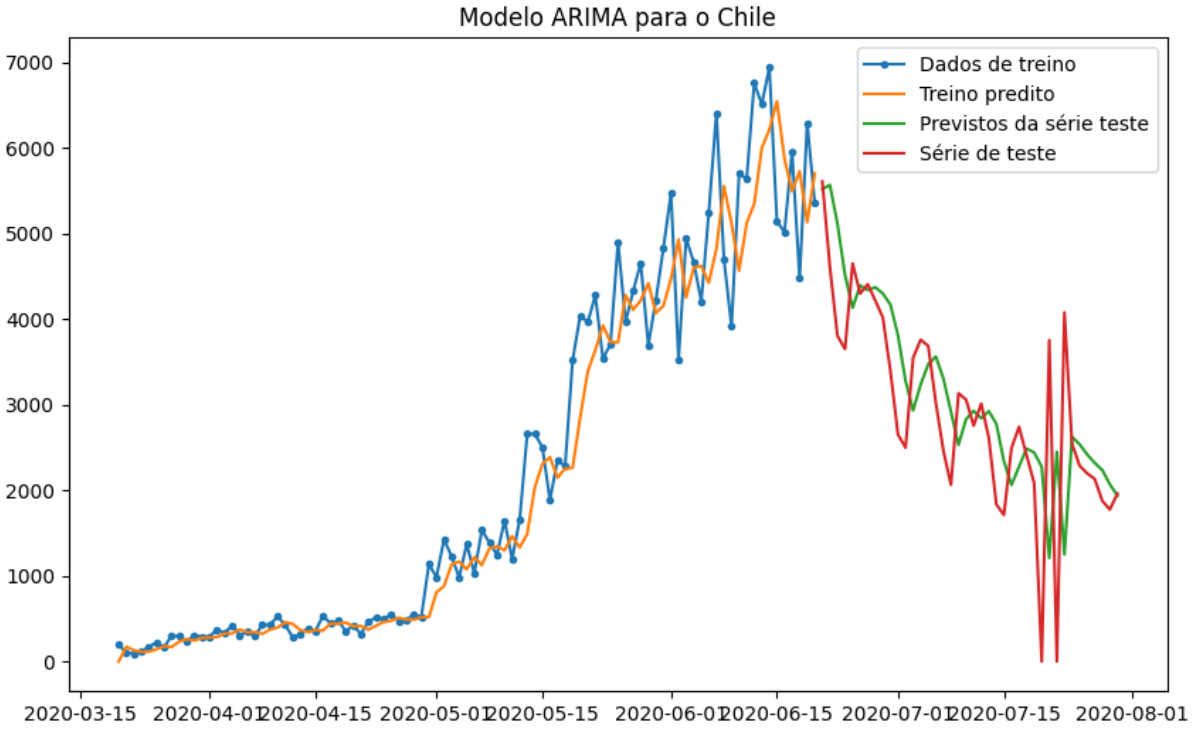
Sazonalidade: Não

Modelo: ARIMA(1, 1, 1)

AIC: 1433

BIC: 1441

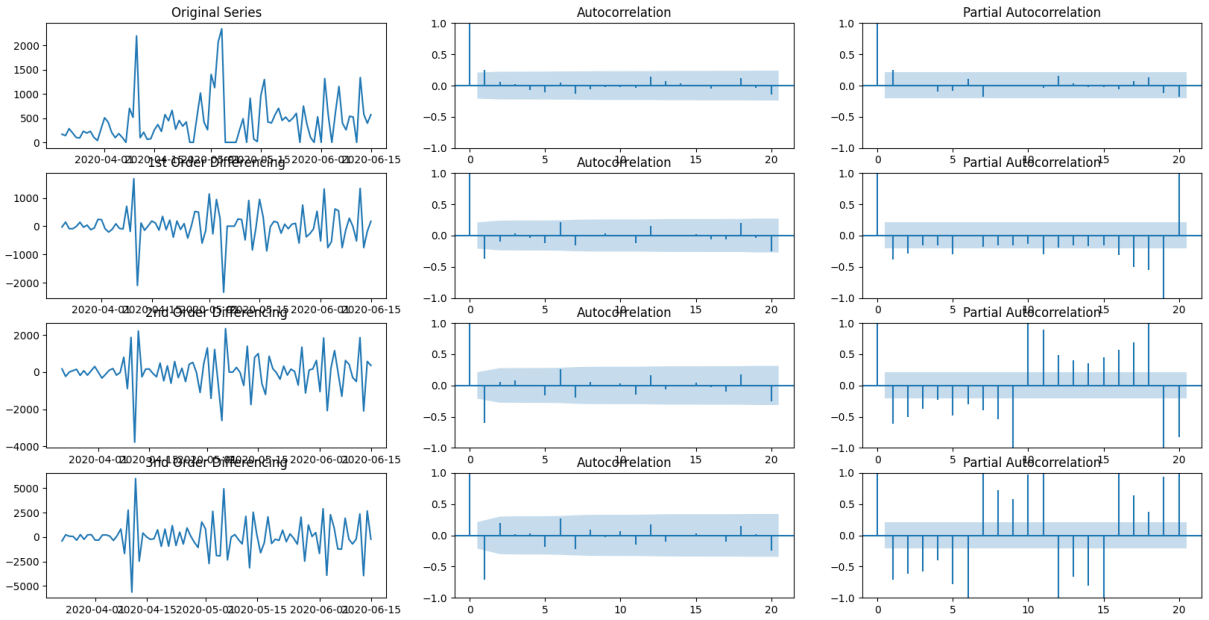
Comparativo entre observado e previsto para horizonte de 1 dia.



Avaliação de desempenho para diferentes horizontes de previsão.

horizonte	r²	RMSE	MAE	MAPE	Correl
h1	0.393765	714.819626	547.706769	0.176050	0.767251
h2	-0.047928	843.087266	659.370091	0.214619	0.686668
h3	-0.251248	888.161394	670.550947	0.218914	0.646482
h4	-0.131423	847.634956	655.008021	0.214642	0.687337
h5	0.000304	803.073816	616.049173	0.201572	0.730443
h6	-0.228330	851.503321	643.754110	0.211542	0.695302
h7	-0.509156	921.615709	716.157272	0.243672	0.634512
7 dias	-0.148137	638.189604	502.698770	0.155634	0.936243

Equador



Termo autoregressivo: 1

Número de diferenciações: 1

Termo de médias móveis: 1

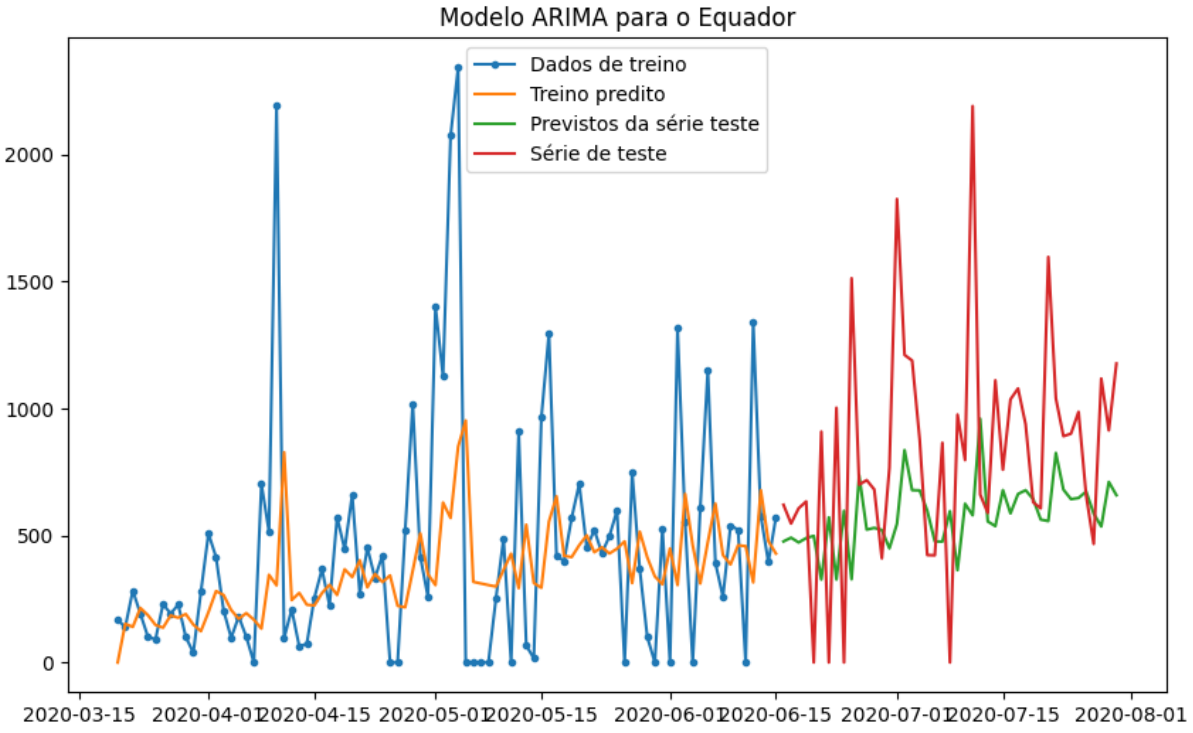
Sazonalidade: Não

Modelo: ARIMA(1, 1, 1)

AIC: 1324

BIC: 1331

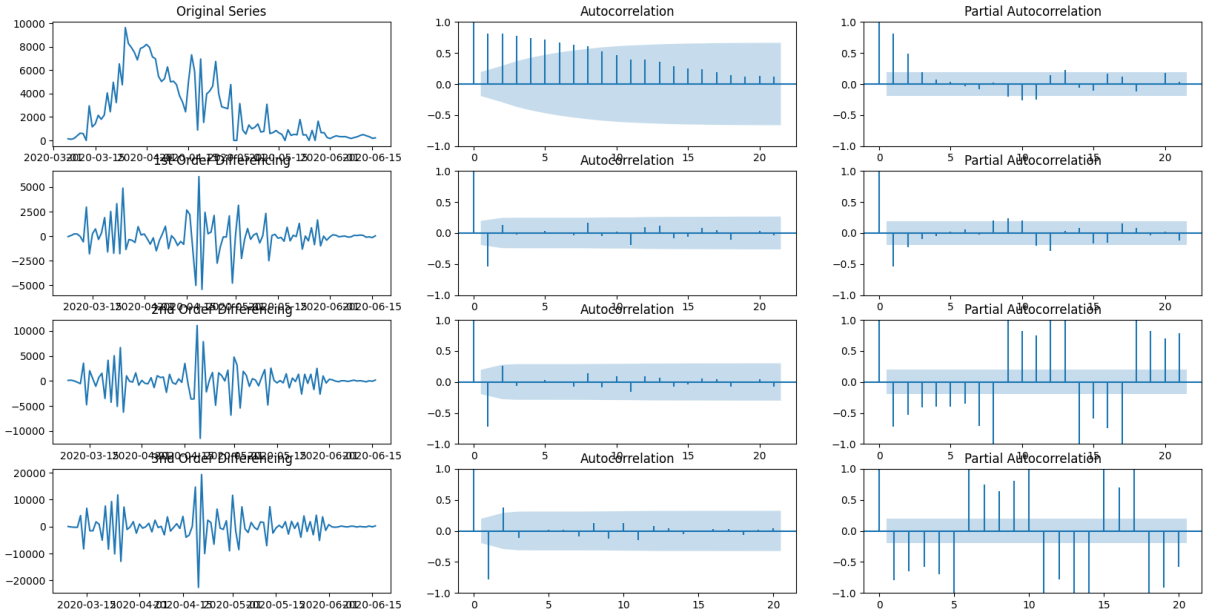
Comparativo entre observado e previsto para horizonte de 1 dia.



Avaliação de desempenho para diferentes horizontes de previsão.

horizonte	r²	RMSE	MAE	MAPE	Correl
h1	-0.166868	409.302422	274.879404	0.308645	0.096763
h2	-0.167460	406.675266	270.738969	0.308904	0.022990
h3	-0.158714	400.720203	259.619594	0.288899	0.039624
h4	-0.189927	405.002749	268.502949	0.302791	-0.002482
h5	-0.161667	404.661111	267.102663	0.301176	0.022515
h6	-0.084417	388.191782	259.085926	0.284043	0.155999
h7	-0.053433	382.603727	258.745471	0.278161	0.198520
7 dias	-2.151132	168.409737	143.255215	0.158145	0.274549

Espanha



Termo autoregressivo: 2

Número de diferenciações: 1

Termo de médias móveis: 8

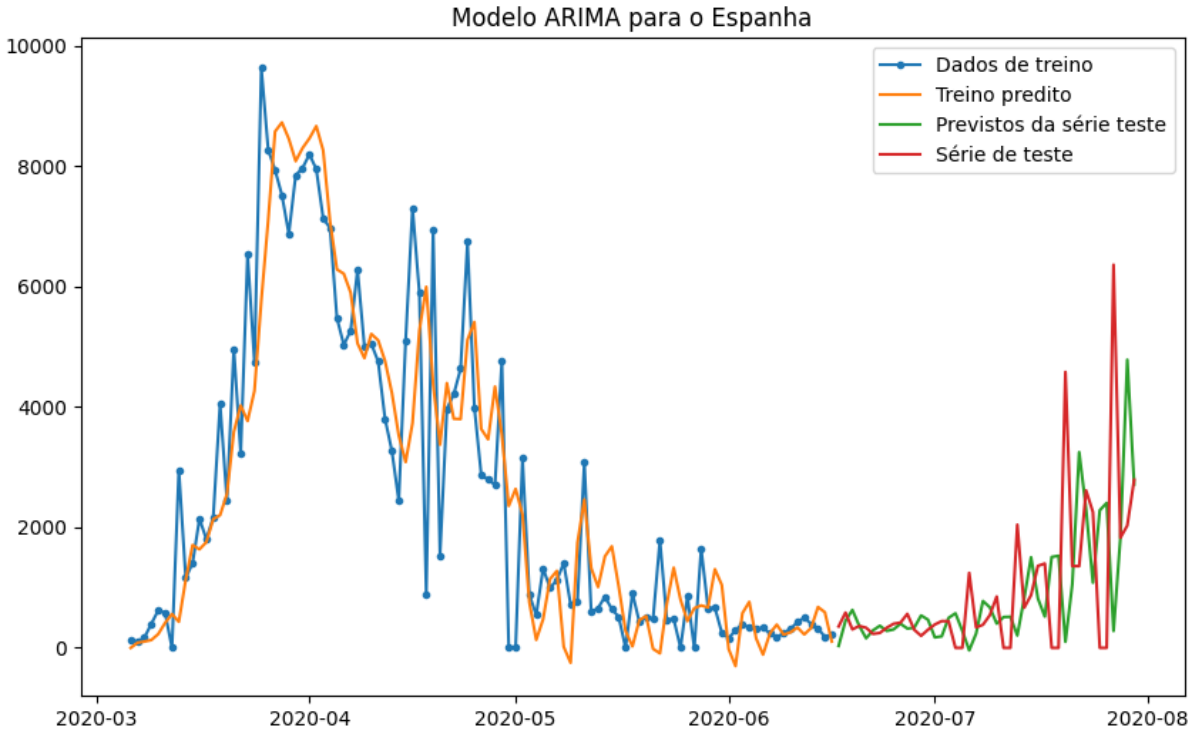
Sazonalidade: Não

Modelo: ARIMA(2, 1, 8)

AIC: 1858

BIC: 1787

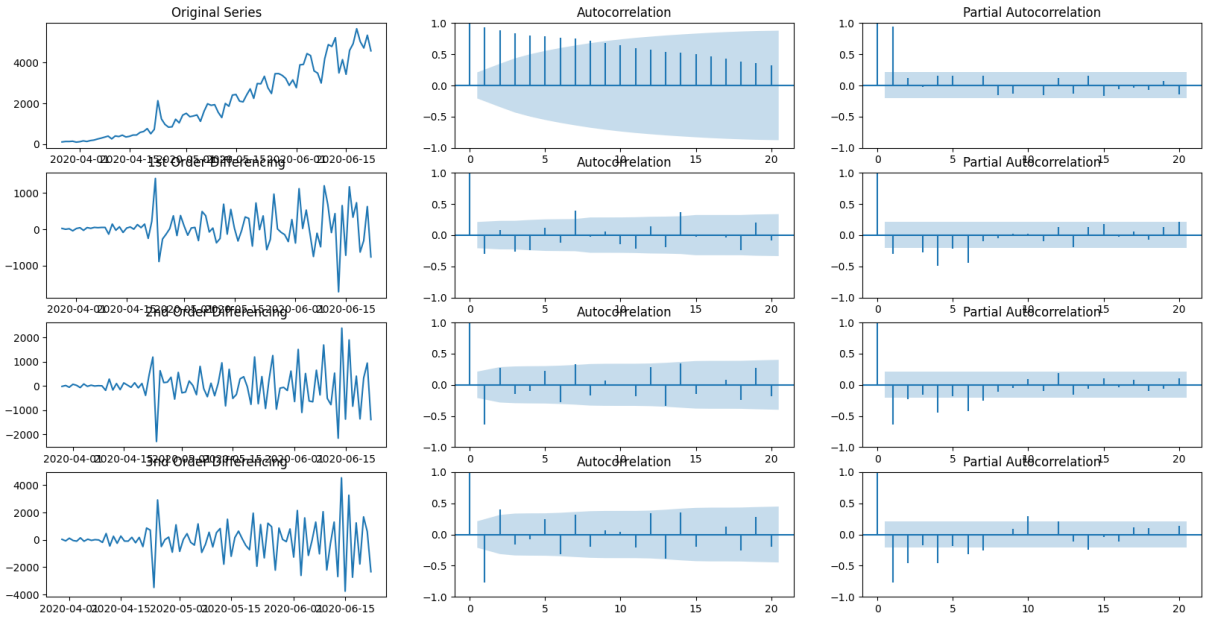
Comparativo entre observado e previsto para horizonte de 1 dia.



Avaliação de desempenho para diferentes horizontes de previsão.

horizonte	r²	RMSE	MAE	MAPE	Correl
h1	0.344496	667.513891	322.702316	0.323941	0.601605
h2	0.439583	638.686622	286.276901	0.283874	0.678521
h3	0.540339	592.130293	237.346355	0.225248	0.752099
h4	0.441137	666.922591	304.560424	0.246998	0.692424
h5	0.375864	983.801917	478.505901	0.283922	0.671392
h6	0.430307	938.592099	488.807380	0.286153	0.725684
h7	0.553905	829.304256	476.155149	0.315068	0.842278
7 dias	0.788244	355.348951	260.008112	0.247846	0.937579

México



Termo autoregressivo: 1

Número de diferenciações: 1

Termo de médias móveis: 1

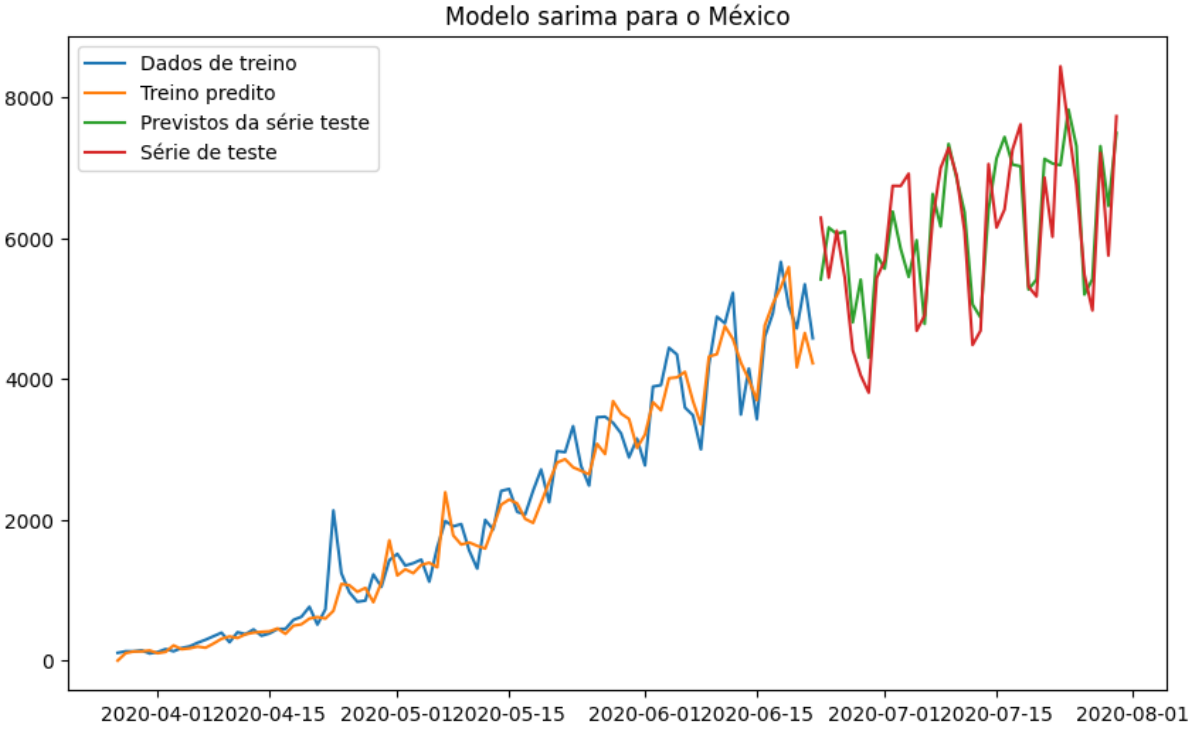
Sazonalidade: Sim

Modelo: SARIMA(1, 1, 1)(1, 1, 0)7

AIC: 1169

BIC: 1173

Comparativo entre observado e previsto para horizonte de 1 dia.



Avaliação de desempenho para diferentes horizontes de previsão.

horizonte	r²	RMSE	MAE	MAPE	Correl
h1	0.399268	858.313312	701.711820	0.119417	0.636729
h2	0.288806	962.593487	771.119251	0.132872	0.544463
h3	0.223205	1007.467552	784.245911	0.135987	0.504155
h4	0.222753	1011.920854	795.858883	0.139460	0.512190
h5	0.229350	1017.223217	820.218845	0.144093	0.525062
h6	0.217175	1006.340851	832.670578	0.143505	0.532247
h7	0.190735	967.296366	794.796413	0.131252	0.532402
7 dias	-0.467170	621.471575	495.548774	0.084379	0.375565

5. Análise dos dados

Após a extração e transformações dos dados, montamos um dashboard, como foi a evolução da pandemia do período de janeiro a julho de 2020, dos países selecionados, Argentina, Equador, Chile, México e Espanha.

5.1. Casos Covid

5.1.1. Totais

Os casos confirmados dos cinco países são 2 milhões e 99 mil óbitos.

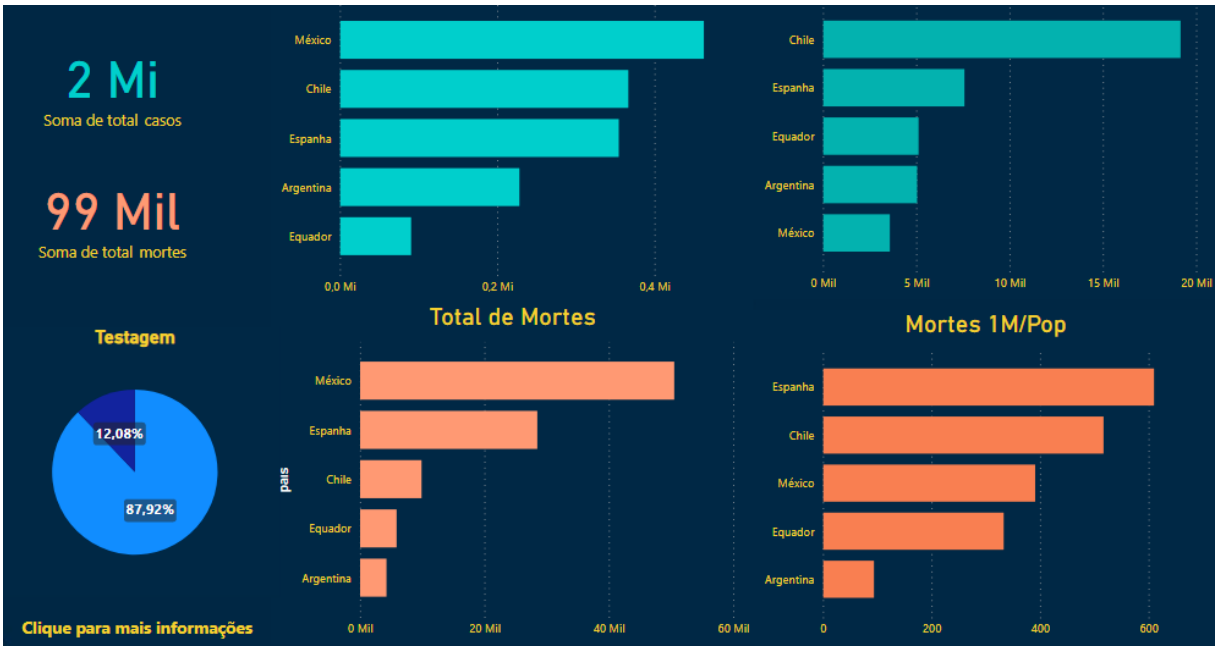
Nos casos confirmados por país temos o México em primeiro seguido do Chile e no total de óbitos o México também está em primeiro, seguido da Espanha.

Quando trata-se de avaliar por 1 milhão de habitantes, a perspectiva muda, cujo Chile fica em primeiro nos casos confirmados, seguido da Espanha. E nos casos de mortes a Espanha está em primeiro, seguido do Chile.

O México é um País populoso o que terá uma maior taxa se avaliando em quantidades, mas quando há um parâmetro o mais afetado é o Chile, por ser um país pequeno.

O percentual de testagem nos mostra o total de testes realizados quantos deram positivo, e esse índice é importante para averiguar se fará necessário medidas restritivas.

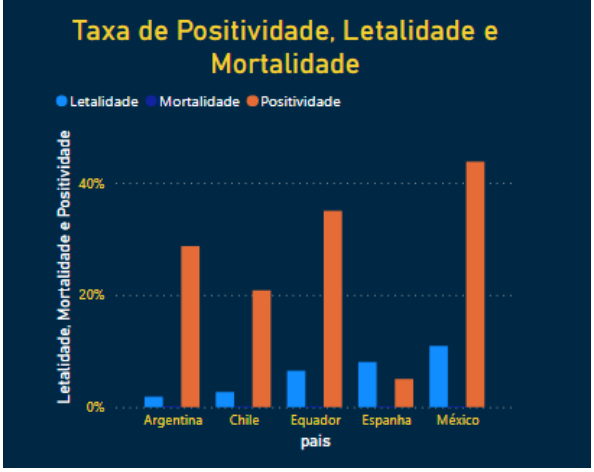
1. Representação gráfica do covid em relação aos países.



5.1.2. Taxas de letalidade, mortalidade e positividade.

A taxa de letalidade se refere a quantidade de casos confirmados quantos foram a óbito, taxa de mortalidade é a relação do tamanho populacional com a quantidade de óbitos e a taxa de positividade do total de testes realizados quantos deram positivo.

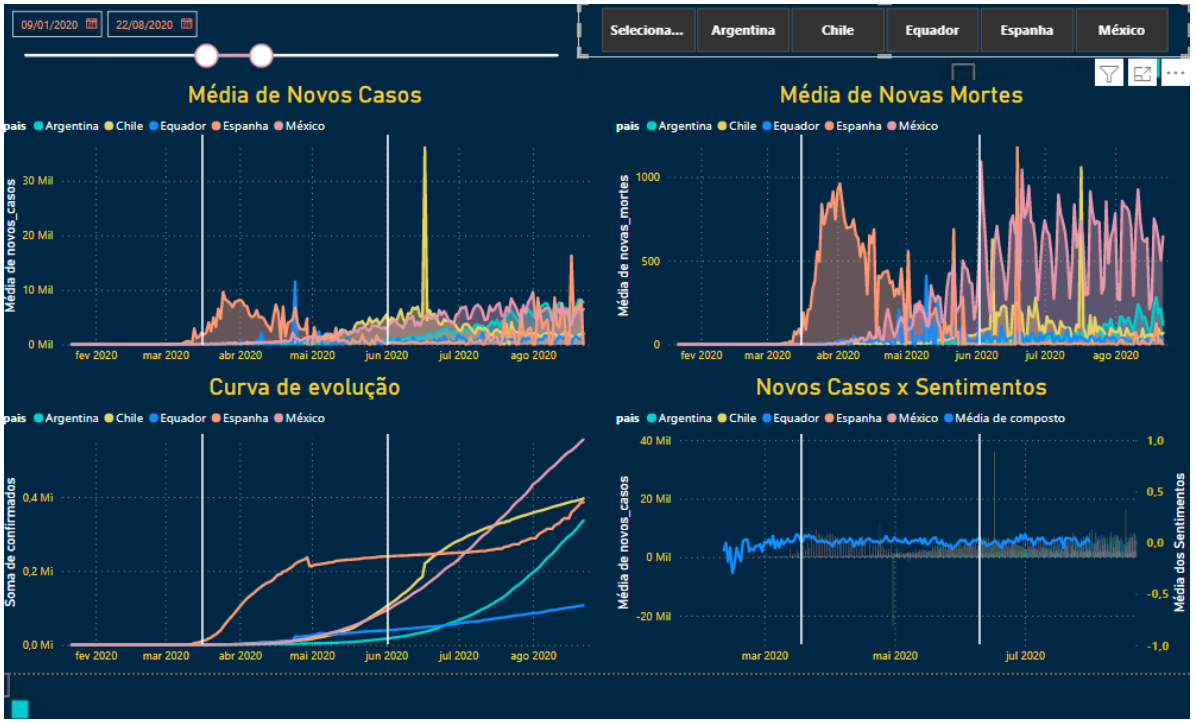
pais	Letalidade	Mortalidade	Positividade
Argentina	1,86%	0,01%	28,72%
Chile	2,70%	0,05%	20,83%
Equador	6,49%	0,03%	35,01%
Espanha	8,04%	0,06%	5,02%
México	10,92%	0,04%	43,78%



Na taxa de letalidade o México está com os maiores índices de mortalidade, a Espanha está mais alta e na taxa de positividade o México está com o índice maior.

5.1.3. Evolução em cada país

Cada país teve uma evolução durante o período observado, mas nos atentamos entre os países da américa latina e a Espanha no continente Europeu, la ja estava ocorrendo a pandemia, um país turístico e com uma população idosa grande e o que obteve os maiores resultados. Os países latinos seguraram um pouco melhor a evolução da doença.

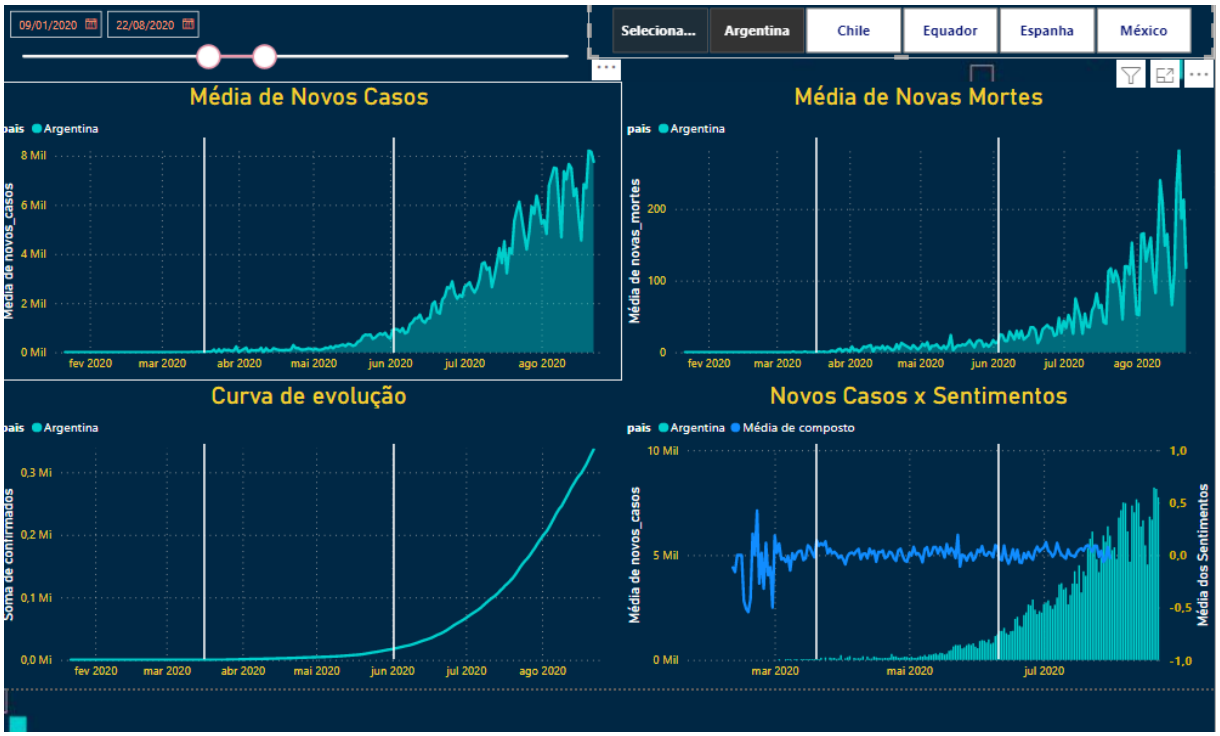


Aqui podemos comprar a Argentina um país latino a Espanha, Casos de covid começaram a aumentar após ao lockdown, não tinham muitos casos, e houve um maior atraso da evolução da doença.

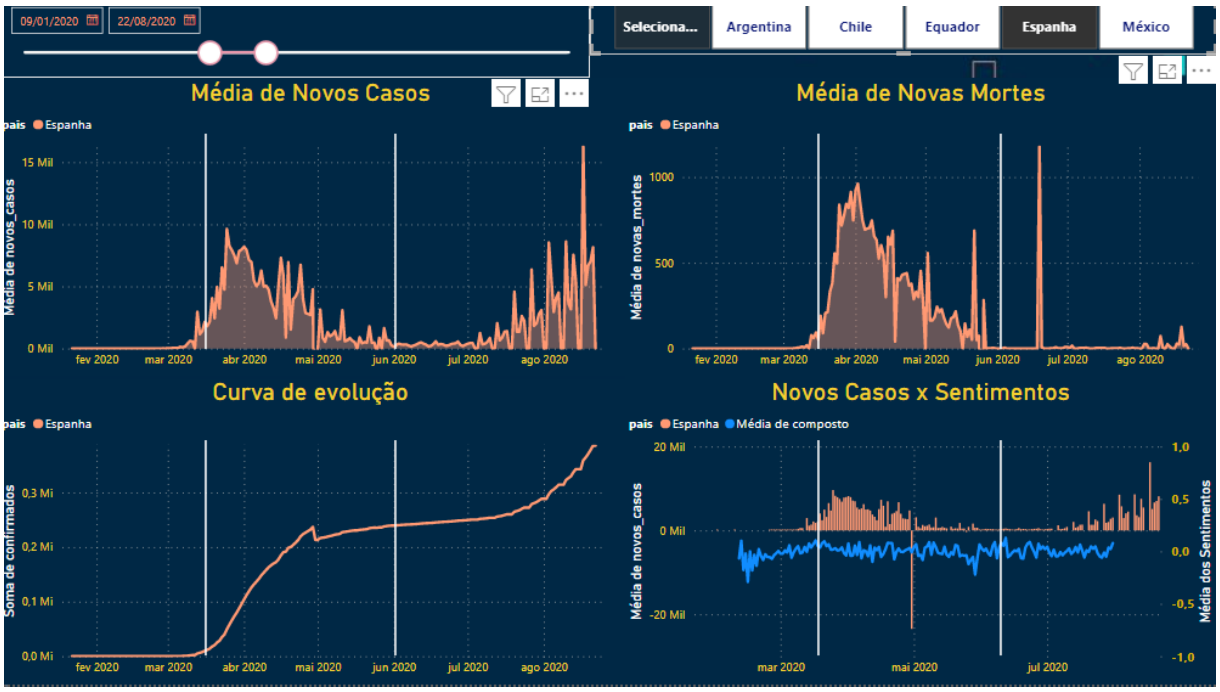
Já a Espanha estava no meio da primeira onda, e pode-se observar que o lockdown foi fundamental para o controle da pandemia.

No gráfico está representando a média móvel dos casos diários, a média móvel dos casos de óbito e a curva de evolução da doença.

Argentina:



Espanha:

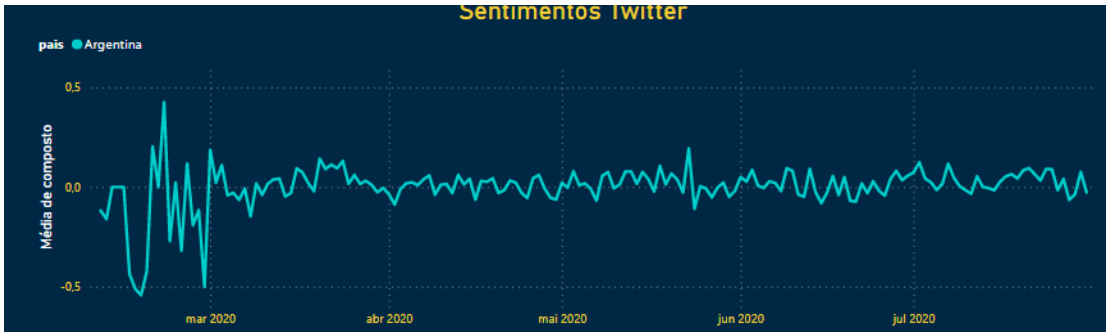


5.2. Sentimentos

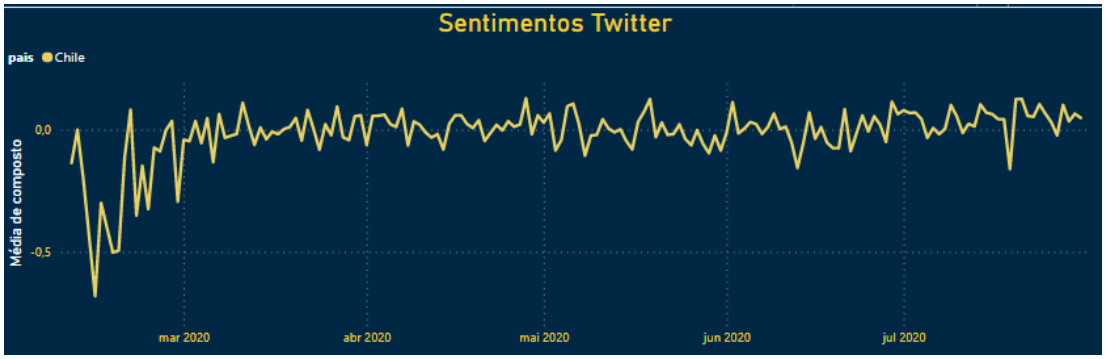
Na extração dos twitter dos países, conseguimos uma representação gráfica deste período, desde fevereiro de 2020 quando a OMS decretou que estávamos em uma pandemia mundial até julho de 2020.

5.2.1. Sentimentos por país

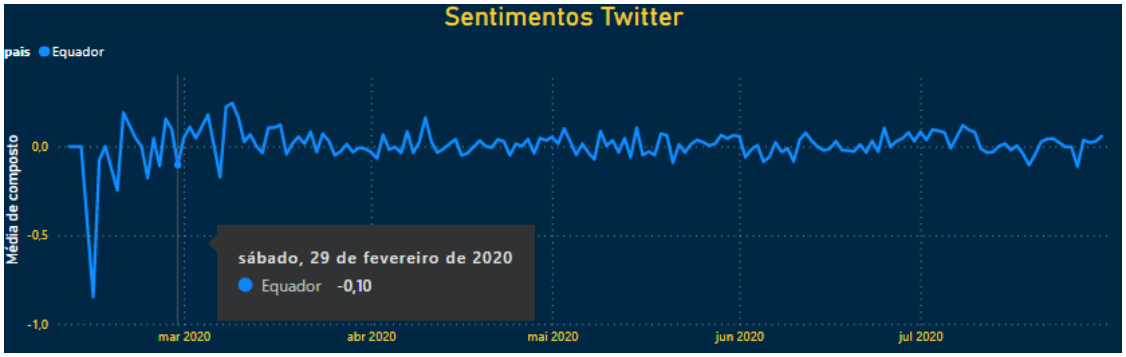
Argentina:



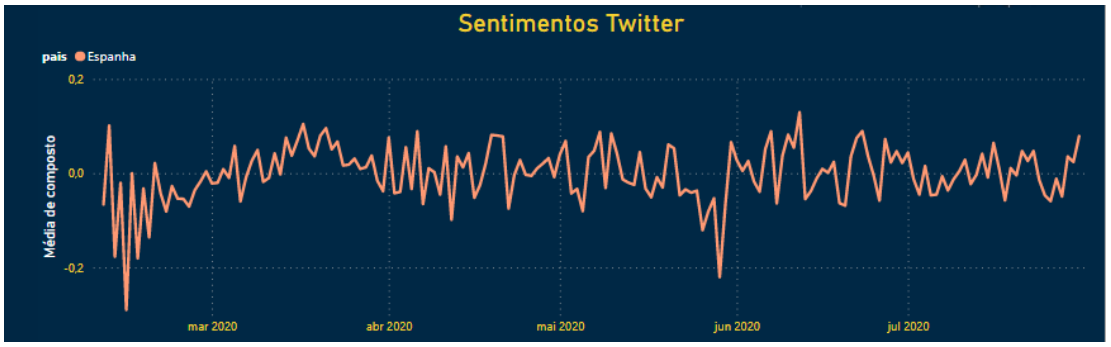
Chile:



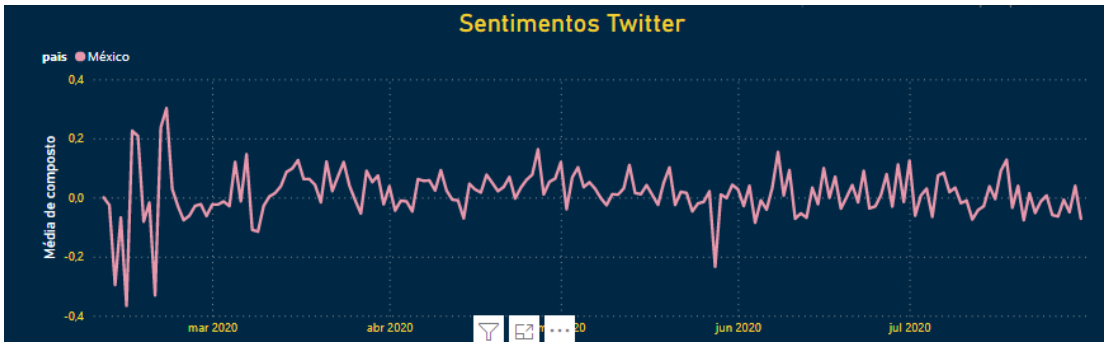
Equador:



Espanha:



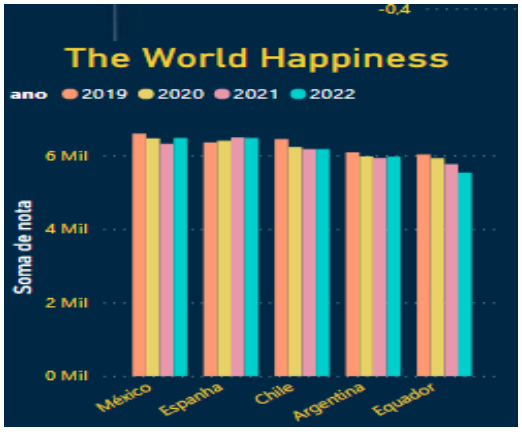
México:



Podemos observar que no começo da do decreto da pandemia, houve uma confusão de sentimentos em todos os países, estabilizando no decorrer do período.

5.2.2. Rank de felicidade

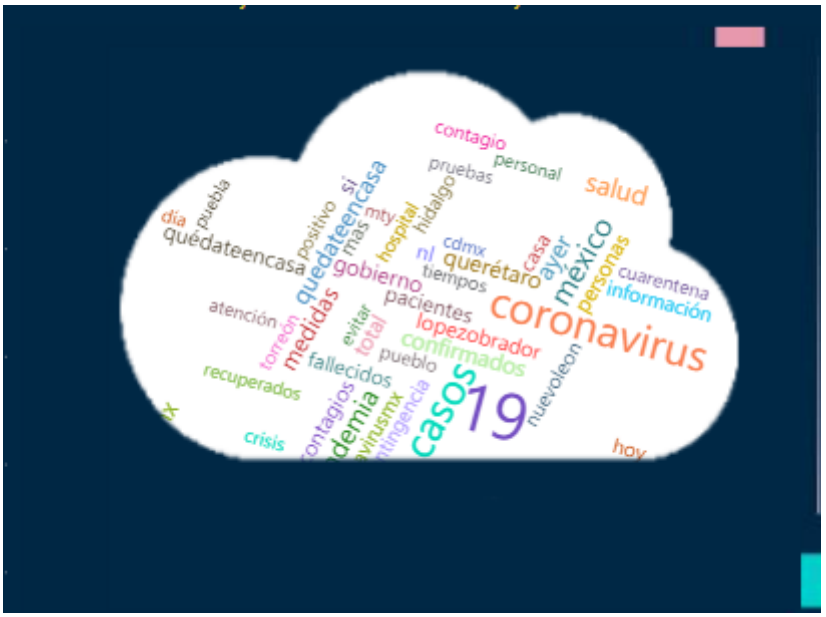
Rank da felicidade é um estudo com notas e classificação dos países e a felicidade, obtivemos o período antes, durante e pós pandemia, para verificar se houve uma queda ou aumento da classificação desses países.



Podemos observar que houve uma pequena redução do período da pandemia e pós pandêmico.

5.2.3. Termos principais nos tweets

Elaboramos uma nuvem de palavras mais utilizadas no twitter dos países.



5.3. Demanda de Energia elétrica

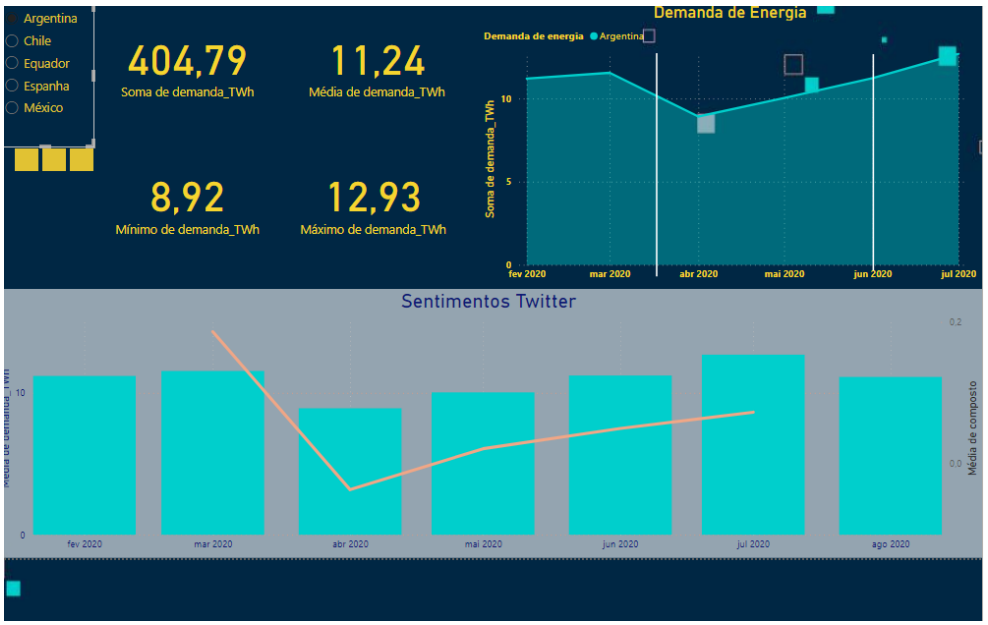
O intuito de comparar a série de demanda de energia em conjunto com os sentimentos é ver, ainda que indiretamente, o quanto medidas de lockdown impactaram no ânimo da população.

Durante os lockdown as principais atividades econômicas dos países pararam. Muitas pessoas ficaram em casa, muitas com medo de saber se teriam seus trabalhos de volta. E sem trabalhar também ficaram sem dinheiro para pagar despesas básicas. É fácil perceber o período de lockdown pois há redução da demanda de energia. Pois, apesar do maior número de consumidores estar no segmento residencial a maior demanda de energia elétrica de um país provém da atividade industrial e comercial.

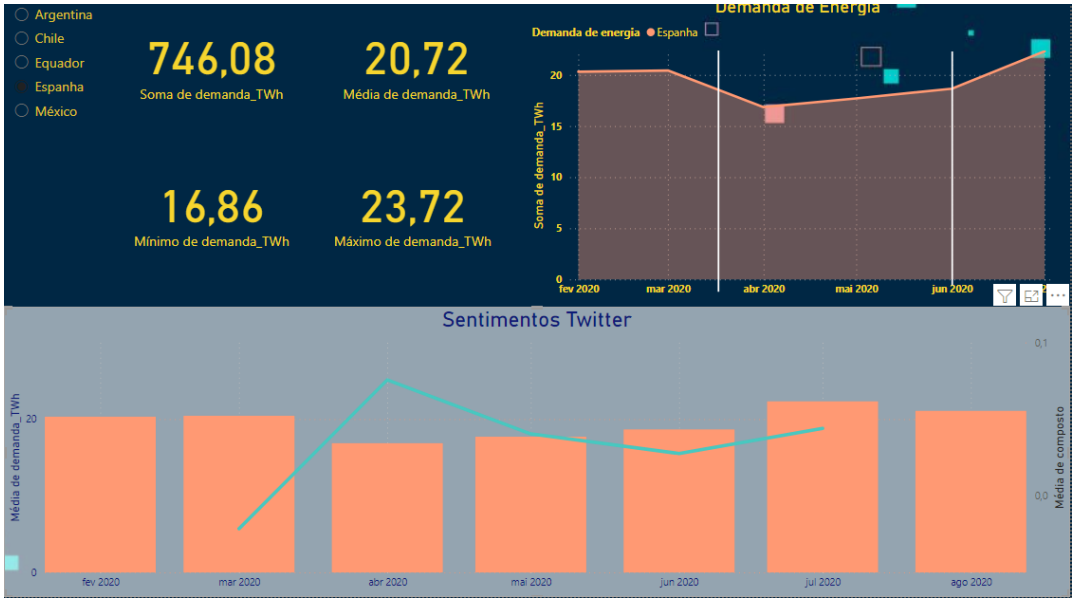
E como tais atividades estavam paradas houve significativa redução da demanda. Observa-se então que durante os períodos de lockdown, principalmente nos países que tiveram lockdowns mais intensos há também a redução do

sentimento das pessoas. Reflexo da angústia em saber o que irá acontecer no futuro.

Abaixo o exemplo da Argentina, que houve uma queda na demanda de energia, e no mesmo período houve um momento negativo nos sentimentos das pessoas.

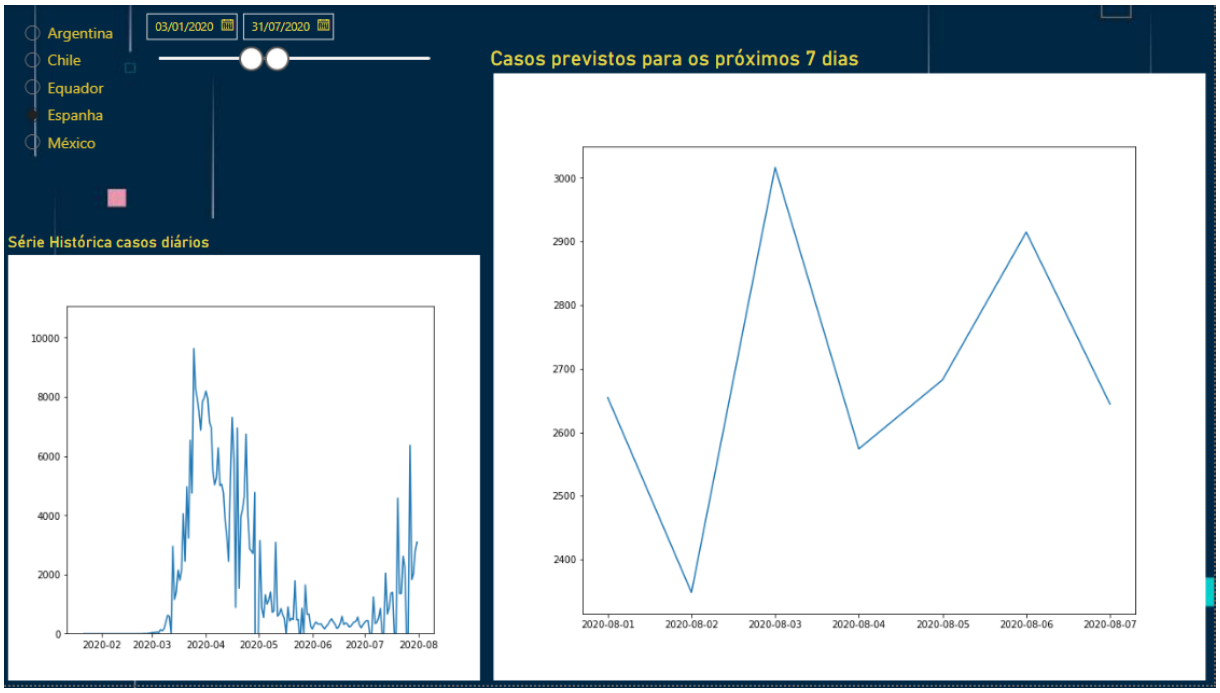


Outro exemplo a Espanha, houve uma queda na demanda energética, porém comparado aos sentimentos expostos no twitter que foram positivo, pois nesse lockdown em específico da Espanha houve redução dos casos confirmados e de óbitos, o que deixou as pessoas com esperança:



5.4. Modelo preditivo

No modelo utilizado, apresentamos a série histórica dos casos diários, e no lado o resultado do modelo previsto para os sete próximos dias. Abaixo a representação da Espanha:



6. Referências

https://pt.wikipedia.org/wiki/Universidade_Johns_Hopkins

https://pt.wikipedia.org/wiki/Twitter

https://pt.wikipedia.org/wiki/Worldometer

https://ember-climate.org/about/