Case – Data Science Digio

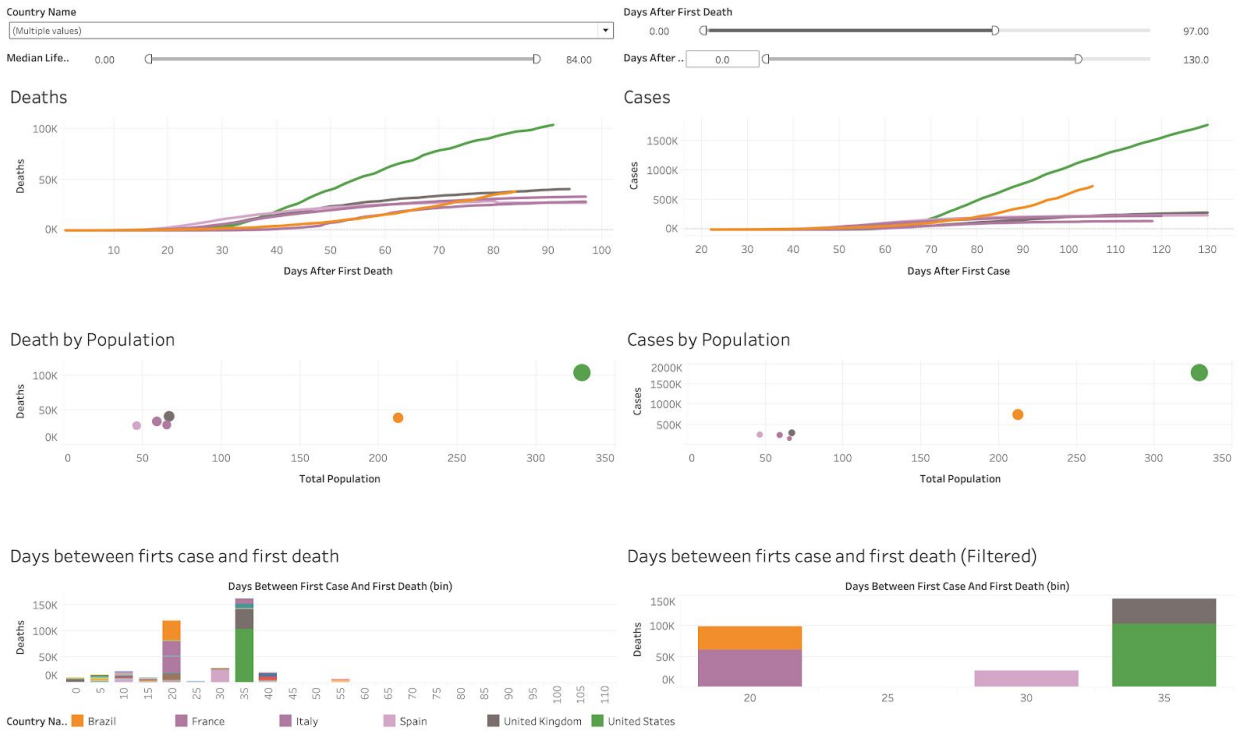
Questões

1. ***Levante informações e gere visualizações que julgue relevantes sobre a Pandemia de COVID-19 no Brasil.***

As visualizações foram desenvolvidas em Tableau e estão anexadas ao e-mail, com filtros para facilitar a manipulação da informação. Também pode ser acessada através desse link:

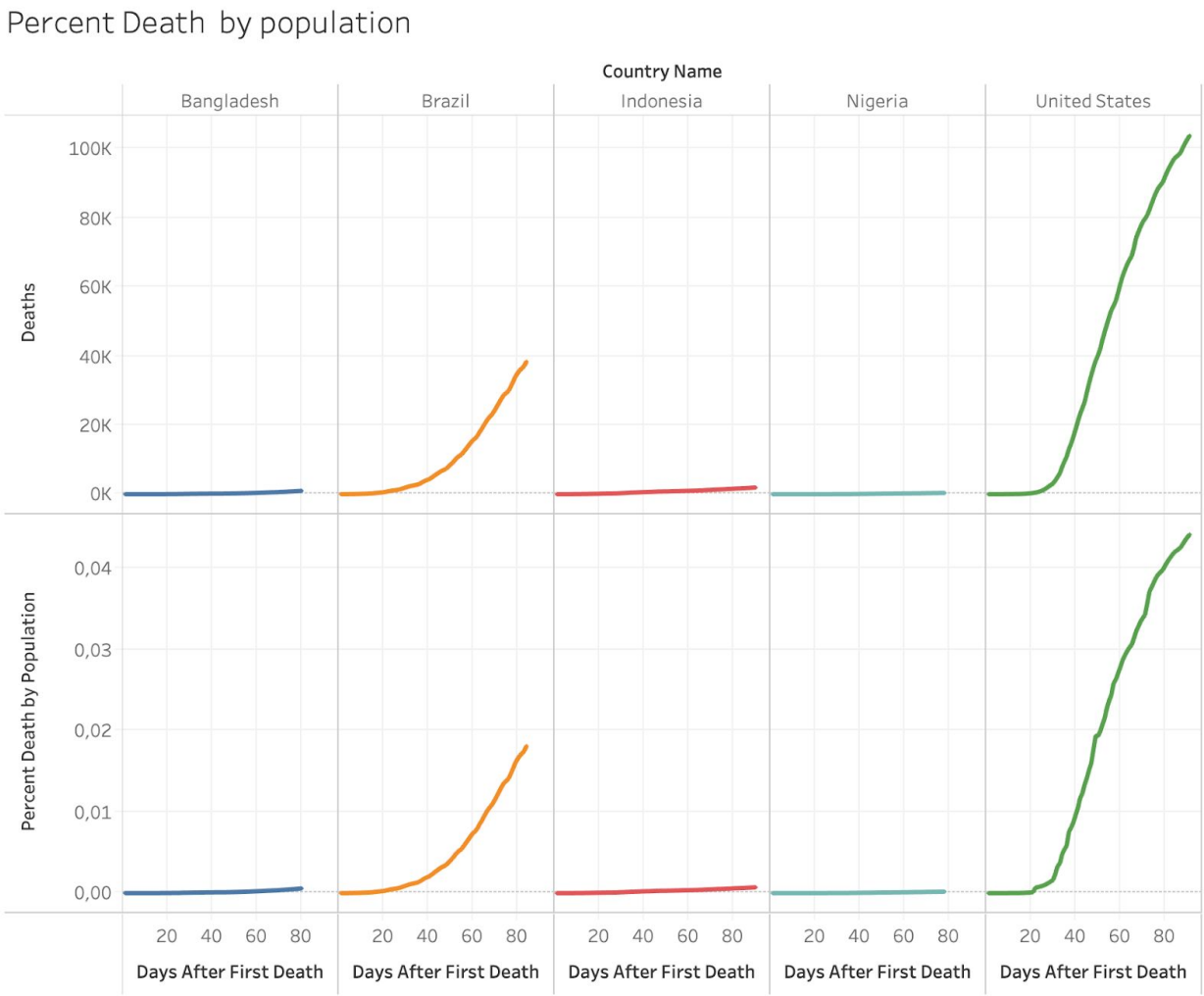
https://github.com/thaina-batista/covid-19-analysis/tree/master/visualization

Caso a visualização via tableau não seja possível, estou anexando um print do painel principal comparando o Brasil com os demais países com maior número de mortos:



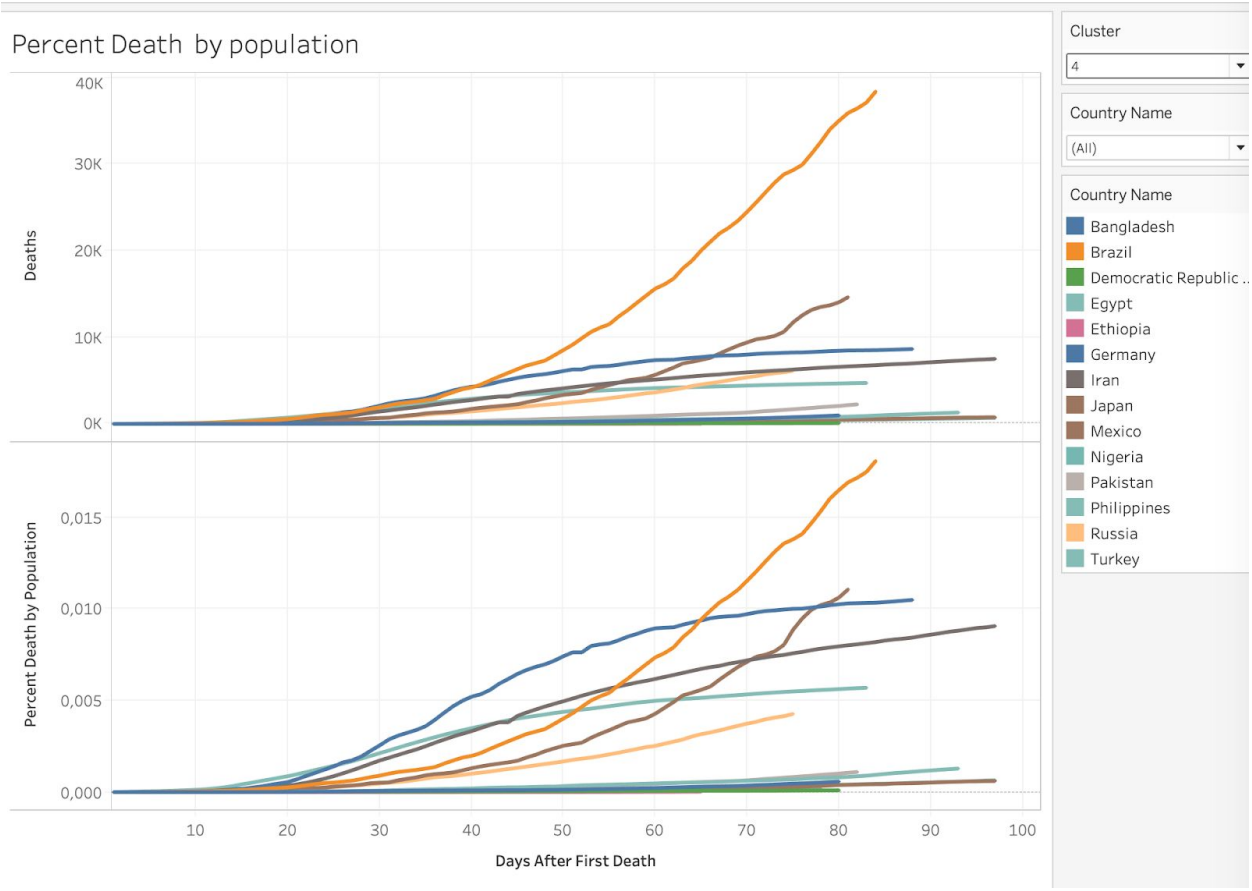
2. ***Qual ou quais seriam indicadores adequados ao fazer a comparação entre diferentes países? Por quê?***

Para realizar comparação entre diferentes países, o ideal é trabalhar com percentual de mortos pela população do país, dado que o número de mortos pode ser maior em países mais populosos, como abaixo:



Outra análise interessante é comparar países geograficamente similares, para realizar essa análise realizei o enriquecimento dos dados sobre a doença com dados de demografia dos países (https://data.world/hdx/749ed4a9-6a89-4a3f-a4c8-b5359966a6e9), após isso processei estes dados usando o algoritmo de agrupamento Kmeans (para encontrar similaridades entre os países e agrupa-los), caso queira pode acessar a implementação desse algoritmo em: https://github.com/thaina-batista/covid-19-analysis/blob/master/Clusters.ipynb

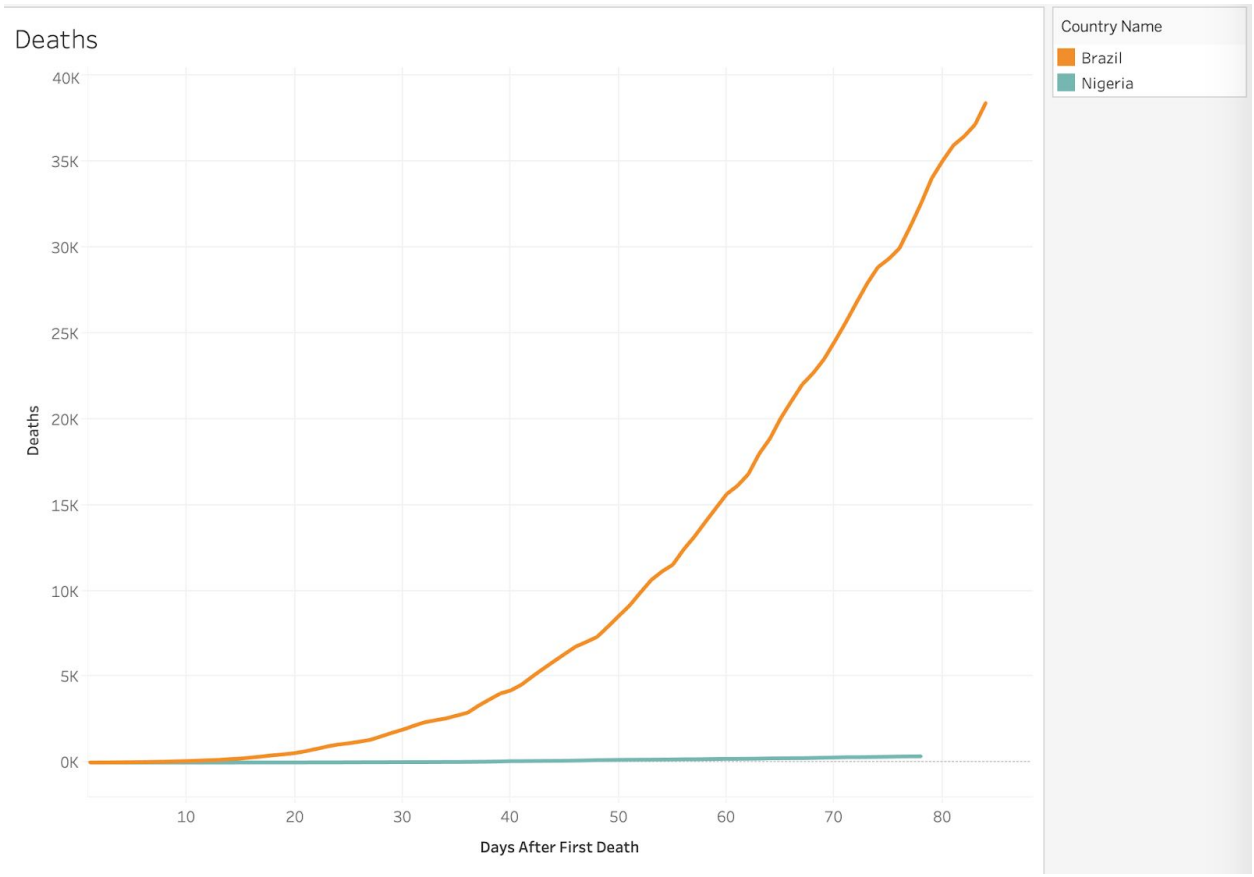
Encontramos então, 5 grupos de países similares, após isso basta comparar os países de determinado grupo. Um exemplo é o grupo ao qual o Brasil faz parte:



Podemos notar que o Brasil é o país com maior número de mortos absoluto e percentual por população dentro do cluster que pertence.

3. ***É possível prever a quantidade de casos que teríamos no Brasil sem as medidas de isolamento social adotadas? Como?***

Sim é possível, podemos utilizar um modelo de impacto casual (https://google.github.io/CausalImpact/CausalImpact.html) para este caso. Também podemos analisar países similares (utilizando kmeans) que tiveram uma postura de isolamento social mais forte, temos como exemplo a Nigéria, que pertence ao mesmo cluster que o Brasil, tem um número de população bem similar ao do Brasil, mas que adotou fortemente o isolamento de sua capital e cidades mais afetadas (https://br.noticias.yahoo.com/covid-19-maior-cidade-africa-lagos-isolamento-social-140715271.html). Podemos notar abaixo a diferença entre o número de mortos entre os dois países:

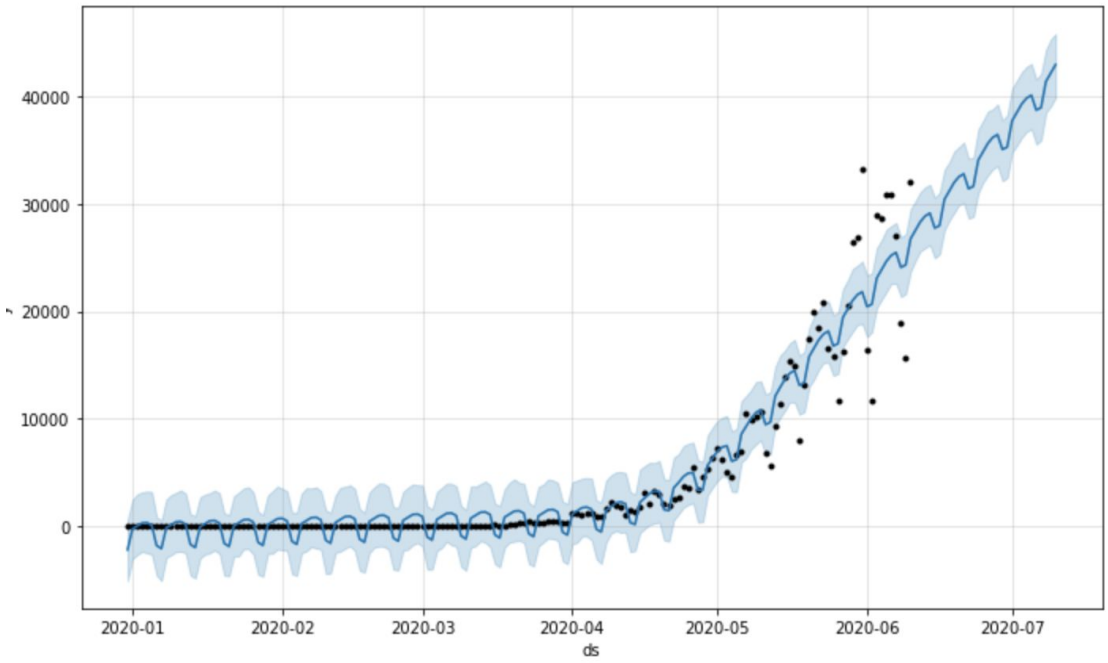


4. ***Baseado no cenário atual que temos no país, faça uma previsão da evolução da pandemia nos próximos 30 dias a partir da data de recebimento deste case. E para os próximos 6 meses? O que você pode dizer sobre a acurácia dessas previsões?***

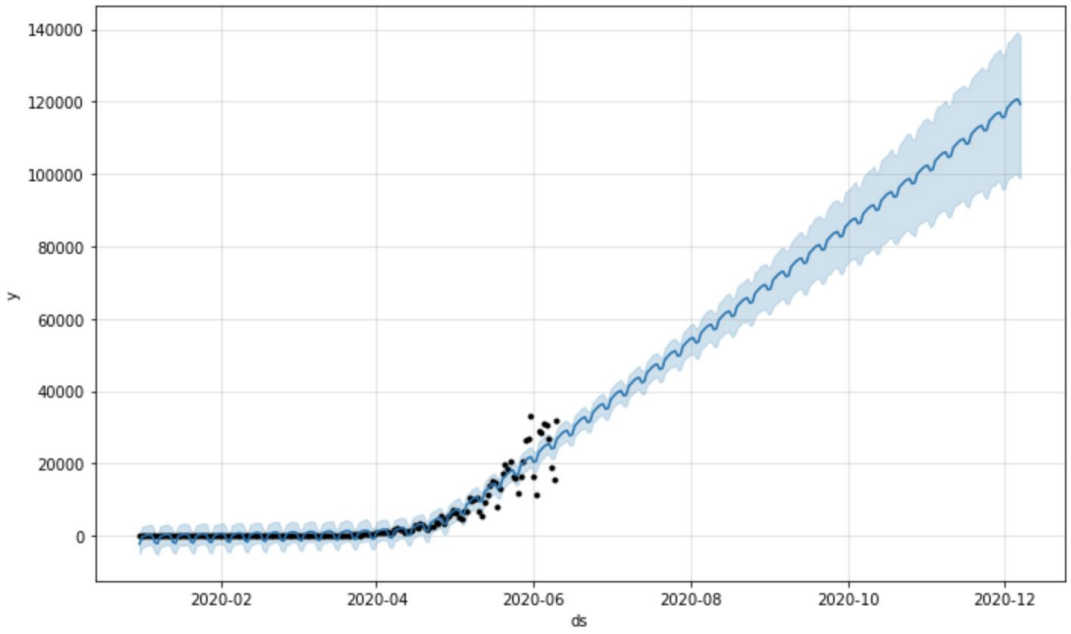
   Para realizar esta predição utilizei duas abordagens: aplicação do algoritmo Gradient Boosting Regression, desta forma pude utilizar outras métricas além do número de casos e mortes, como os dados geográficos dos países a serem analisados. Porém esse modelo não apresentou acurácia relevante, a implementação deste modelo encontra-se em: https://github.com/thaina-batista/covid-19-analysis/blob/master/Gradient%20Boosting.ipynb

   A partir disso, decidi então utilizar o modelo de forecast Prophet (https://facebook.github.io/prophet/) e obtive o seguinte resultado:

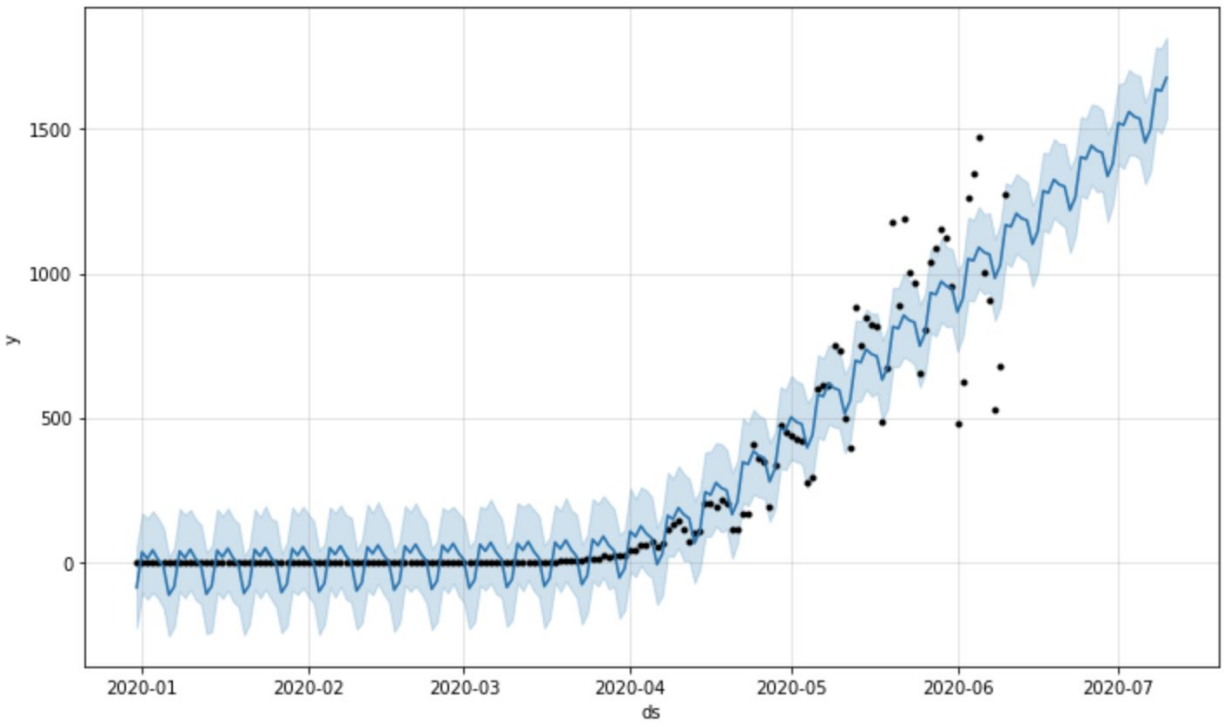
Quantidade de casos confirmados por dia (30 dias):



Quantidade de casos confirmados por dia (180 dias):



Quantidade de mortes confirmadas por dia (30 dias):



Quantidade de mortes confirmadas por dia (180 dias):

Acurácia do modelo: Para todos os modelos realizados o erro padrão médio ficou bem próximo a zero, o que é excelente para modelos de forecast.

:

	ds	yhat	yhat_lower	yhat_upper	y	cutoff
0	2020-01-08	0.0	-1.651646e-09	1.592634e-09	0	2020-01-07 23:59:59.999999970
1	2020-01-09	0.0	-1.470258e-09	1.628497e-09	0	2020-01-08 23:59:59.999999970
2	2020-01-10	0.0	-1.614915e-09	1.485383e-09	0	2020-01-09 23:59:59.999999970
3	2020-01-11	0.0	-1.462034e-09	1.371185e-09	0	2020-01-10 23:59:59.999999970
4	2020-01-12	0.0	-1.465059e-09	1.428801e-09	0	2020-01-11 23:59:59.999999970

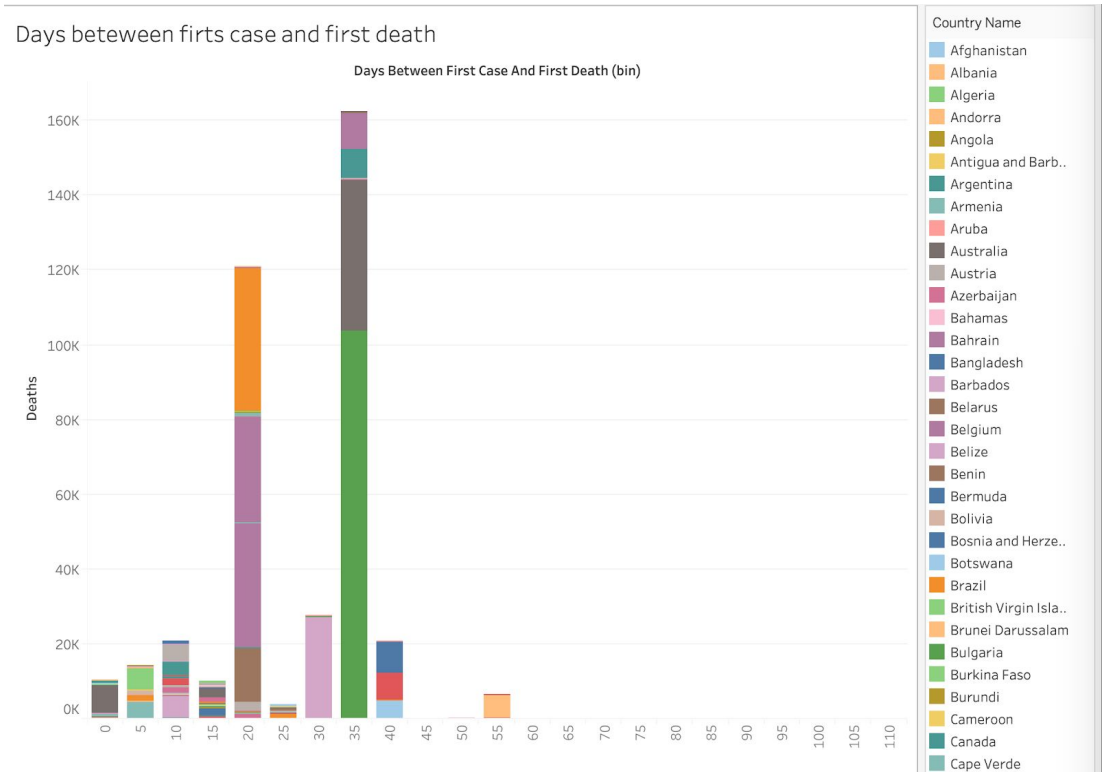
:
```
df_p = performance_metrics(df_cv)
```
```
INFO:fbprophet:Skipping MAPE because y close to 0
```

Porém sabemos que quanto maior o número de períodos a serem preditos menor a acurácia do modelo. Outro ponto relevante é que ainda não atingimos o pico da pandemia, desta forma, ainda não tivemos redução no número de casos e mortes, logo o modelo não apresenta sazonalidade anual, temos apenas a tendência de crescimento.

5. ***Se se sentir a vontade, insira análises adicionais, visualizações ou comentários que julgar relevantes para a sua pesquisa.***

Um dado interessante que pode ser notado durante essa pesquisa é a hipótese de que os países com maior número de mortos não tenham realizado grandes politicas de isolamento social nos primeiros momentos pois levou mais tempo entre o primeiro caso confirmado e a primeira morte.

Neste gráfico podemos notar a quantidade de mortes pela diferença entre datas da primeira morte e do primeiro caso confirmado:



Se isolarmos o gráfico apenas para os 5 países com o maior número de mortos podemos notar que estão concentrados entre 20 e 35 dias de diferença entre o primeiro caso e a primeira morte: