UNIVERSIDADE FEDERAL DE PERNAMBUCO - UFPE
CAMPUS CARUARU
CURSO DE BACHARELADO EM ENGENHARIA DE PRODUÇÃO

Projeto – Análise de Séries Temporais
Um estudo sobre o roubo à propriedade no estado de Pernambuco no contexto da Pandemia e Quarentena de 2020.

Rodrigo Duarte Xavier da Costa

RECIFE – PERNAMBUCO
Período extracurricular
2020

1. Introdução

No ano de 2020 um surto causado por um vírus do grupo dos coronavírus, o Sars-CoV-2 gerou diversos impactos na sociedade, uma das áreas mais afetada foi a economia, pois como medida de saúde foi necessário fechar estabelecimentos de diversos setores e a mudança para o sistema de trabalho home-office. Tais mudanças influenciaram drasticamente na dinâmica da sociedade, com menos pessoas se deslocando diariamente e com diversos comércios fechados é importante analisar como a criminalidade reagiu a essas mudanças.

Dado esse contexto de pandemia, a análise será voltada ao período mais rígido da quarentena que começou no dia 16/04/2020 [1] até o dia 30/05/2020 [2] para o estado de Pernambuco, sendo mais específico o munícipio do Recife. A criminalidade estudada será os Crimes contra o patrimônio (CVP), cuja definição dada pelo Governo de Pernambuco é "*Entende-se por CVP todos os crimes classificados como roubo, extorsão mediante sequestro e roubo com restrição da liberdade da vítima, exceto o roubo seguido de morte (...) Nesse sentido, o roubo é o ato de subtrair coisa alheia móvel, para si ou para outro, mediante grave ameaça ou violência à pessoa (ou não), ou depois de havê-la, por qualquer meio, reduzido à impossibilidade de resistência*." [3]

Assim esse trabalho tem como objetivo analisar as principais características acerca dos CVP para o período mais rígido da quarentena em comparação com o mesmo período do ano de 2019, implementar modelos de análises de séries temporais afim de observar características implícitas da série temporal, realizar previsões e comparar com os dados reais, avaliar a qualidade dos modelos e por fim fornecer essas informações no auxílio da tomada de decisão acerca de políticas públicas para combater os CVP.

2. Revisão da Literatura

Na literatura são encontrados alguns artigos que tratam de assuntos e maneiras similares, porém em períodos diferentes dos aqui tratados. O projeto de TCC "Criminalidade e desempenho econômico: Uma análise em séries temporais para a Região Metropolitana do Recife" [4] analisa o comportamento temporal das ocorrências de crimes em Recife com o fator de renda e ocupação no período de 2007 a 2015. A metodologia utiliza ferramentas de séries temporais, dentre elas, Análise de correlação, teste de estacionariedade Dickey-Fuller, modelo VAR (vetor autorregressivo), decomposição da variância etc. O estudo em questão teve como conclusão de que existem variáveis de crime que são fortemente correlacionadas com variáveis econômicas e existem respostas significativas e persistentes para choques (mudanças drásticas) nas variáveis econômicas através das séries de crimes.

Outro artigo que é similar é o "Alcohol and violent behavior among football spectators: An empirical assessment of Brazilian's criminalization" [5] que análise um comportamento violento dos espectadores de jogos de futebol em Pernambuco em relação ao seu consumo de álcool, tal estudo foi possível pois foi decretado uma lei que impedia a venda de álcool em estádios de futebol. A metodologia consiste em realizar um teste não-paramétrico e a modelagem de média móvel autorregressiva (ARMAX) para analisar os dados antes da lei até o período vigente da sua efetivação para os 3 maiores times de futebol de Pernambuco. O estudo teve como conclusão de que a venda e consumo de álcool exerce pouca influência no comportamento violento dos

torcedores, tendo como fatores mais importantes o ambiente (tamanho do público, competitividade, jogos importantes etc.)

3. Metodologia

A metodologia do artigo consiste em obter estatísticas descritivas dos dados, comparar as variâncias e médias do período para os anos de 2020 e 2019 utilizando o Teste Levene [6] e o Teste t [7] respectivamente a fim de entender se houve uma diminuição ou aumento no número de crimes. Avaliar visualmente características como tendência, sazonalidade e ciclo, analisar os gráficos do ACF e PACF para tentar identificar a melhor ordem para o modelo ARMA, Testar a estacionariedade e invertibilidade da série temporal e assim aplicar a melhor ordem ao modelo para realizar previsões e poder comparar a qualidade do modelo com o de suavização exponencial Holt-Winters através de métricas de avaliação, que será a AIC e MSE. Os algoritmos e códigos utilizados foram implementados na linguagem Python, no ambiente de programação Google Colab.

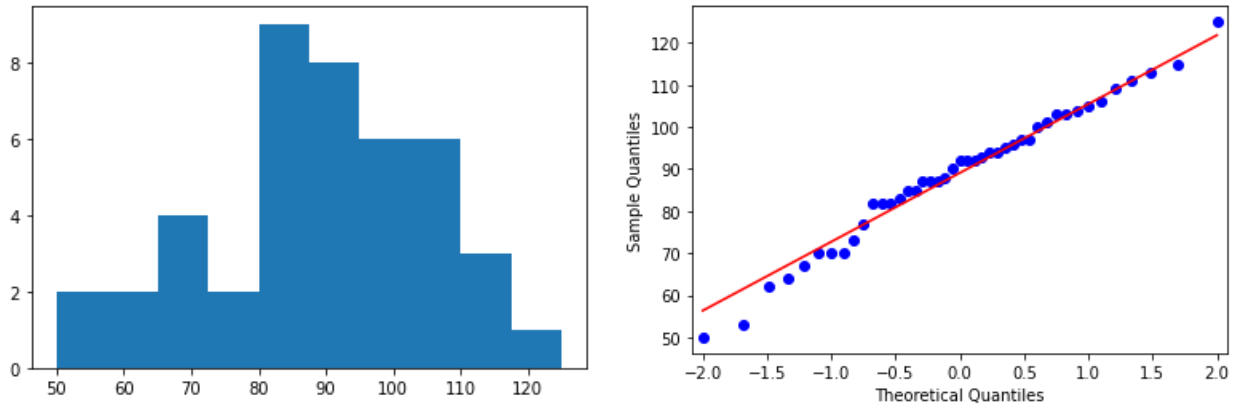
4. Dados

Os dados utilizados foram obtidos no site da Secretaria de Defesa Social do Governo de Pernambuco [x] com o nome “Microdados de CVP”, após o tratamento dos dados para o período de estudo obtemos 45 observações, um número pequeno para uma análise de série temporal, porém para a comparação entre 2019 e 2020 é um número suficiente, deveremos inicialmente levantar estatísticas descritivas dos dois períodos, 2019 e 2020, respectivamente:

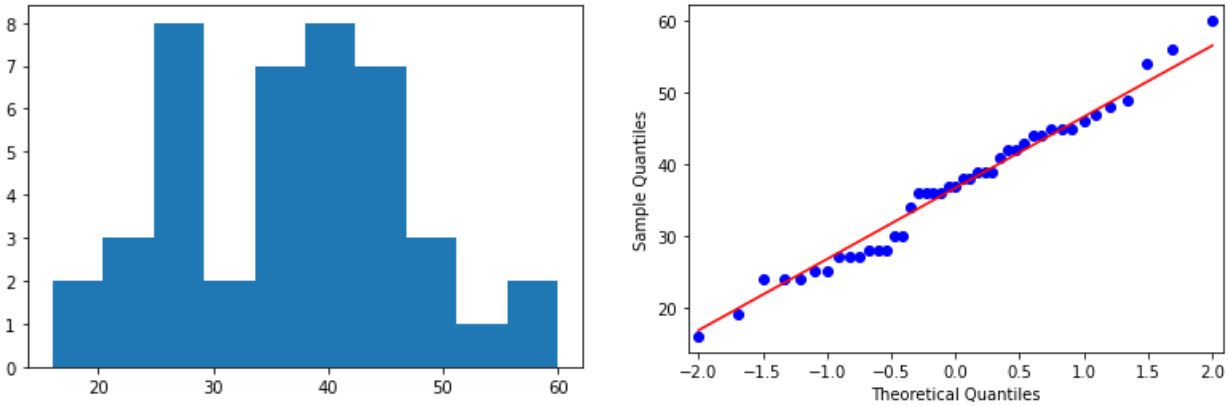
count	45.000000
mean	89.311111
std	16.289738
min	50.000000
25%	82.000000
50%	92.000000
75%	101.000000
max	125.000000

count	45.000000
mean	36.844444
std	10.117392
min	16.000000
25%	28.000000
50%	37.000000
75%	44.000000
max	60.000000

É possível observar que a média e o desvio padrão do período de 2019 é ligeiramente superior ao de 2020, porém os dados que estamos tratando não são todos as ocorrências possíveis no período, tendo em vista que o próprio site onde se obteve os dados informa o seguinte “*sendo considerados apenas os casos registrados nos boletins de ocorrências da Polícia Civil, os quais poderão sofrer modificação posterior em razão das vítimas muitas vezes registrarem suas ocorrências nas delegacias ou delegacia pela internet dias depois do fato*.” [3] Então se faz necessário realizar o teste T para comparar as médias de grupos diferentes. Uma das condições necessárias para realizar esse teste é a que as amostras tenham variâncias diferentes e sigam uma distribuição normal, para isso vamos utilizar as ferramentas teste Levene, histogramas e qqplot.



Histograma e qqplot para o período de 2019

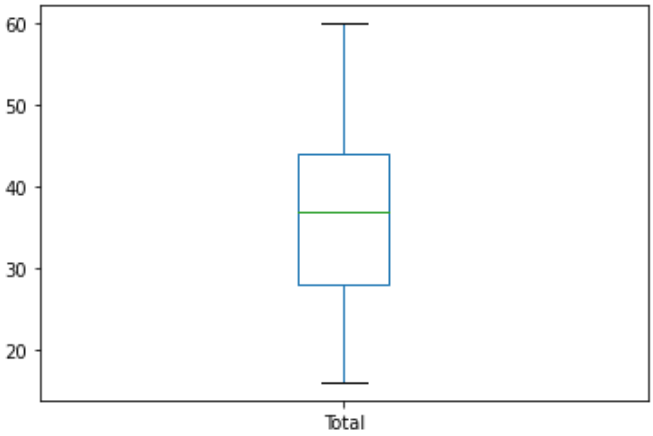


Histograma e qqplot para o período de 2020

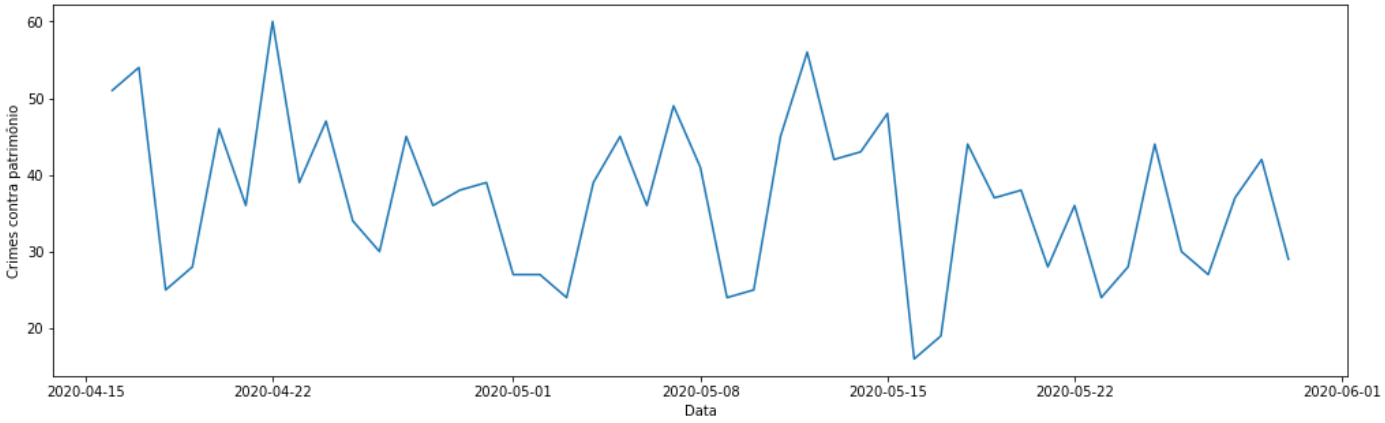
É possível observar que pelos histogramas e o qqplot que as duas amostras seguem uma distribuição normal, o teste Levene teve o resultado $p - value = 0.0083269$, indicando que podemos rejeitar a hipótese nula de que as variâncias são iguais. O teste T teve um $p - value = 1.969199e - 27$, assim podemos afirmar com 95% de certeza que as médias diferem no mesmo período e há fortes indícios que o período de 2020 teve uma média inferior ao de 2019, não se pode afirmar que a quarentena é uma causa mas é um importante fator pela diminuição.

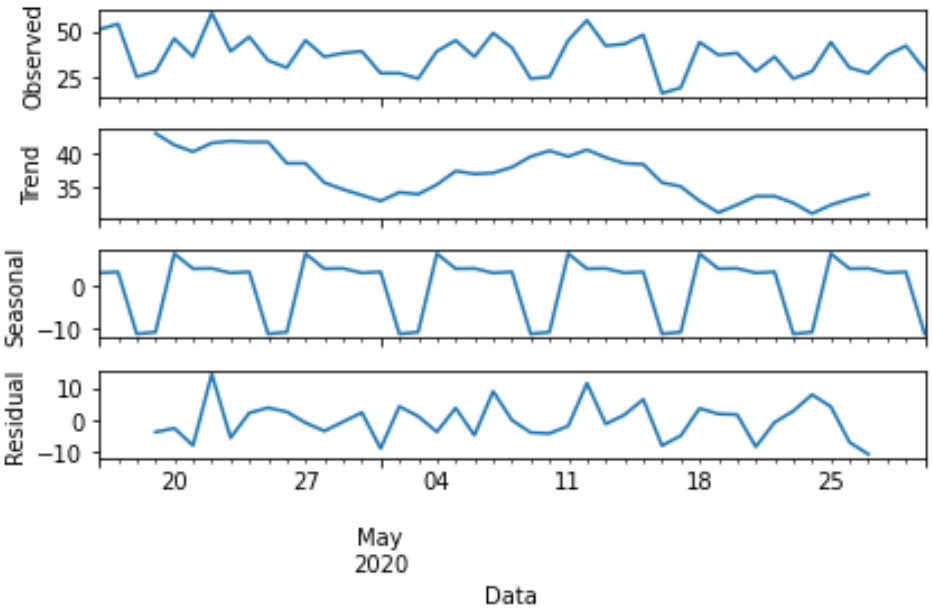
5. Avaliação

Agora iremos utilizar de ferramentas de série temporal para analisar os dados, iremos focar apenas os dados do período de 2020, foi gerado um boxplot para identificar a ocorrência de outliers e como é observado abaixo os dados não contém outliers.

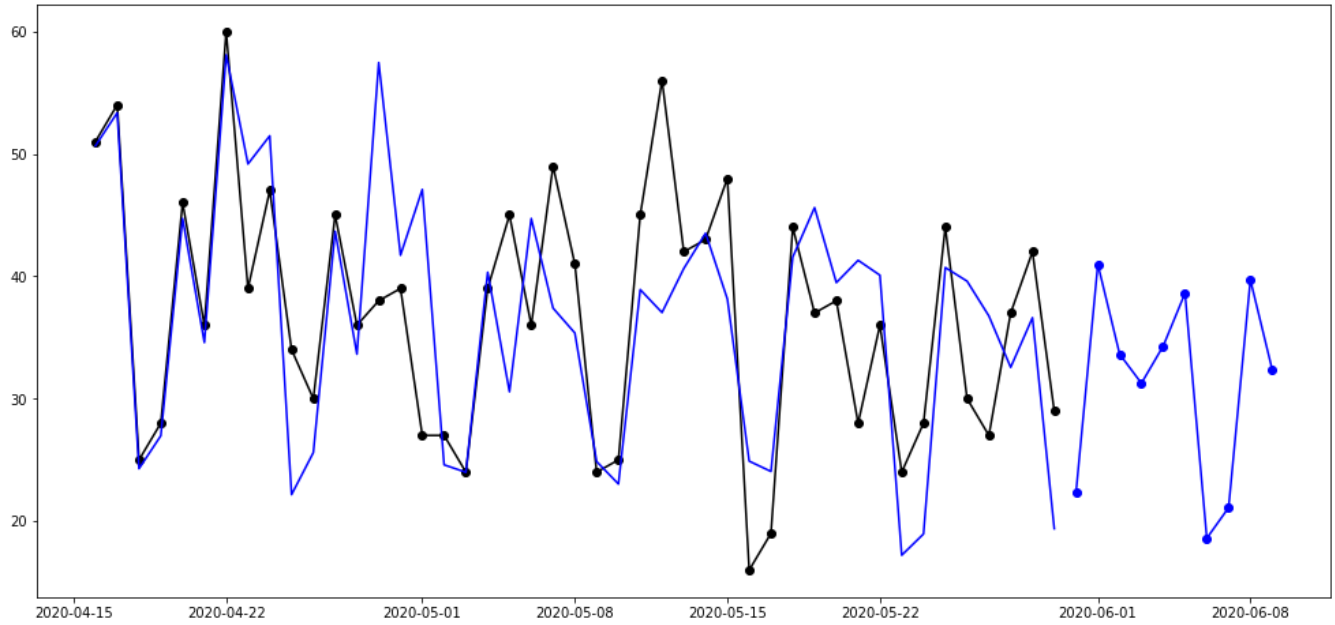


Gerando um gráfico da série temporal podemos ver que a série não aparenta ter uma tendência, sazonalidade, nem ciclos evidentes, é possível analisar isso mais precisamente ao plotar o gráfico da série temporal separando a tendência, sazonalidade e os erros. Os erros por sua vez estão concentrados em torno de uma média 0, indicando que é possível que a série não seja tão errática. Tais características levam a crer que nossa série é estacionária e invertível, futuramente será testado essas hipóteses.



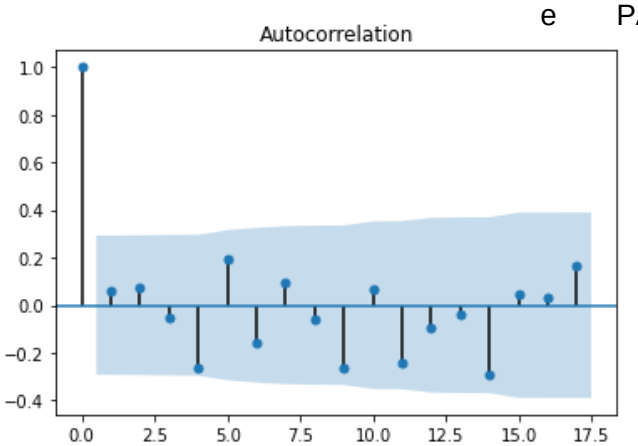


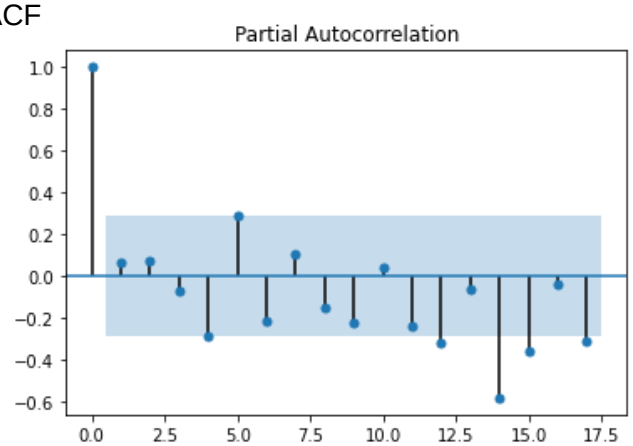
O primeiro modelo utilizado para tentar modelar a nossa série temporal é o de suavização exponencial Holt-Winters, esse modelo serve para prever dados de uma série que apresenta ambos tendências e sazonalidade, como foi visto, essas características não foram identificadas nas nossas séries, porém esse modelo foi oque teve melhor desempenho dentre outros modelo de suavização exponencial (suavização exponencial simples, suavização exponencial de Holt). Os parâmetros otimizados foram $\alpha = 0.0522360$, $\beta = 0.0522360$ e $\Upsilon = 0.4701271$, o gráfico dos valores dos modelos com os dados e a previsão para os próximos 10 dias são apresentados abaixo.



2020-05-31 22.374817	2020-06-03 31.258856	2020-06-06 18.602042	2020-06-09 32.378132
2020-06-01 40.972324	2020-06-04 34.302276	2020-06-07 21.098045	
2020-06-02 33.639037	2020-06-05 38.613848	2020-06-08 39.703510	

O RMSE e AIC deste modelo foi 9.1432 e 209.394, o valor do RMSE ser baixo faz sentido pois os parâmetros são otimizados utilizando o MSE como métrica para a decisão. A análise de resíduos é necessária para identificar se capturamos toda a estrutura do processo, uma forma de avaliar a qualidade do modelo, para isso é necessário verificar se os resíduos estão limpos, ou seja, testar se os resíduos do modelo são não auto correlacionados e se os resíduos seguem uma distribuição normal. Para o primeiro quesito vamos utilizar o teste Ljung-Box e é necessário analisar o ACF e PACF

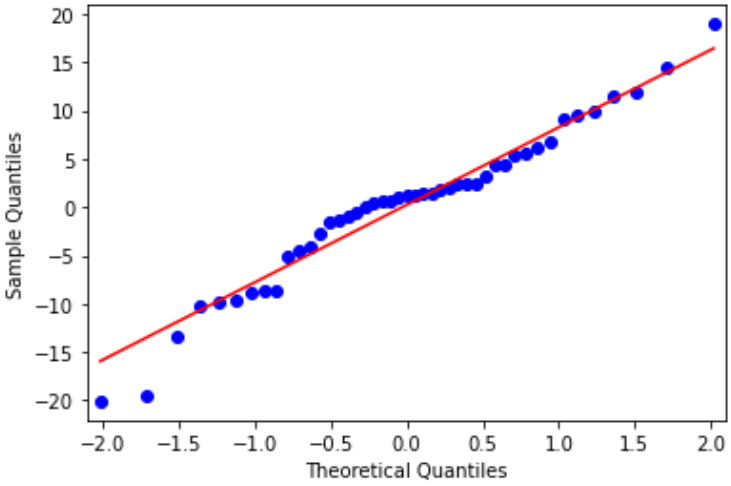




dos resíduos do modelo.

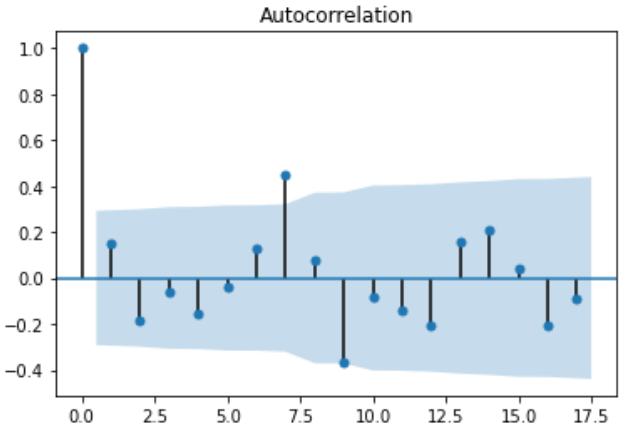
É possível retirar algumas interpretações sobre os gráficos acima, o PACF indica que os resíduos são não correlacionados até o 12º lag, tanto o ACF quanto o PACF têm nenhum padrão ou lag um número de lag significante, caracterizando um Ruído Branco, esse é um importante fator para a qualidade do modelo. O teste Ljung-Box tem como hipótese nula que os dados não são auto correlacionados, como podemos ver, todos os lags até o 20 a hipótese nula é rejeitada, então os resíduos não são auto correlacionados.

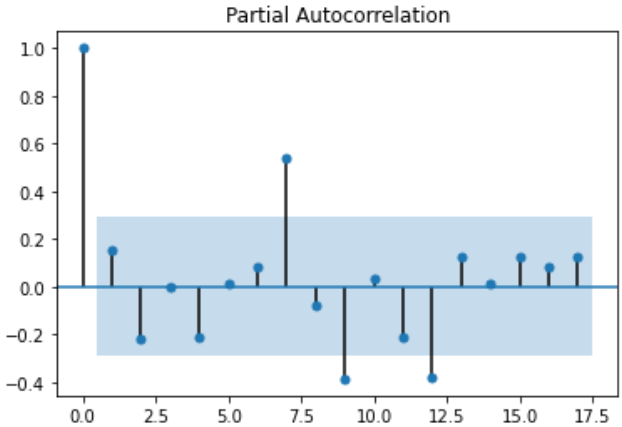
lag	p-value	lb_stat	rejectH0
1	0.66	0.19	False
2	0.80	0.45	False
3	0.89	0.61	False
4	0.39	4.10	False
5	0.29	6.12	False
6	0.28	7.48	False
7	0.33	7.99	False
8	0.42	8.17	False
9	0.20	12.19	False
10	0.25	12.48	False
11	0.14	16.09	False
12	0.16	16.70	False
13	0.21	16.81	False
14	0.07	22.64	False
15	0.09	22.79	False
16	0.12	22.86	False
17	0.09	24.99	False
18	0.09	26.29	False
19	0.11	26.69	False
20	0.07	30.21	False



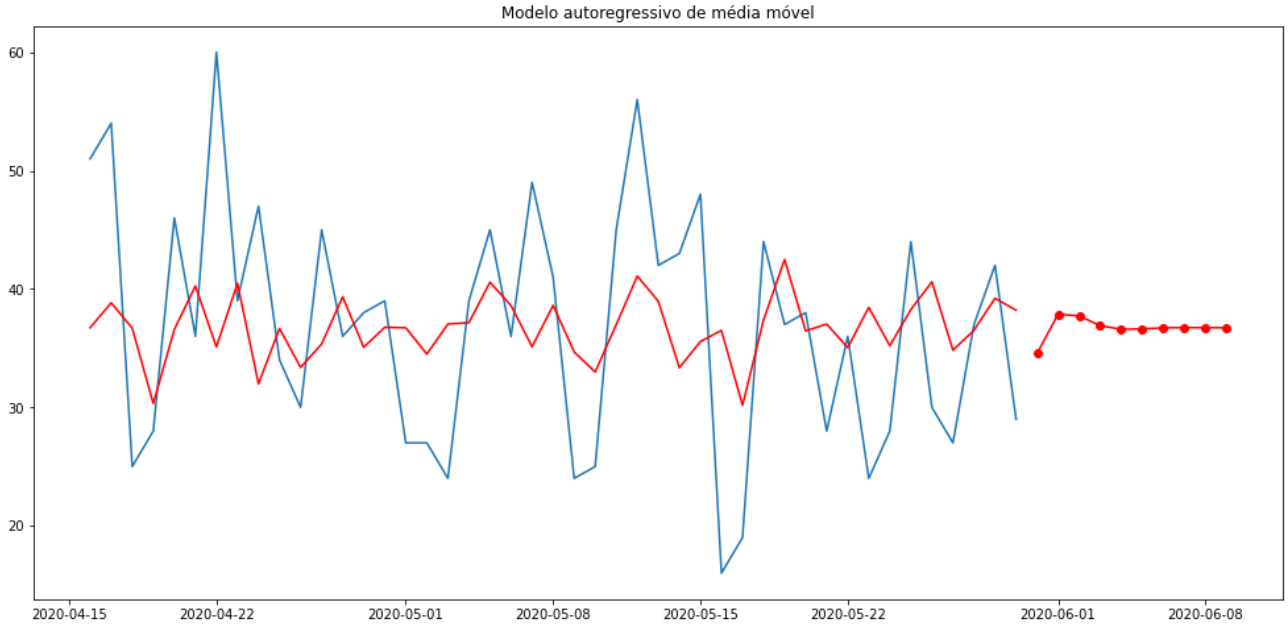
Para a segunda parte da análise de resíduos vamos testar a normalidade do mesmo através do teste de normalidade Shapiro-Wilk, que a hipótese nula é que a série é normalmente distribuída, $p - value = 0.2880$, então não podemos rejeitar que a hipótese é nula. Vamos utilizar o qqplot para suportar a nossa hipótese. Dado essas análises, podemos afirmar que os resíduos estão limpos e que o nosso modelo de Holt-Winters teve um bom desempenho de interpretar a série temporal.

O segundo modelo utilizado para será o Modelo autorregressivo com média móvel, ARMA, inicialmente é importante identificar a ordem AR(p) e MA(q) através dos gráficos de PACF e ACF respectivamente. Esses gráficos também nos indicarão se nossa série é um Passeio aleatório ou Ruído Branco. É observado então um ARMA(3, 1) e que nossa série não é nem Passei aleatório e nem Ruído Branco.



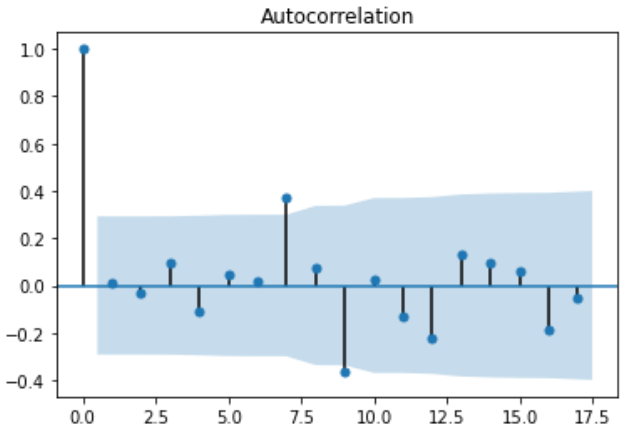


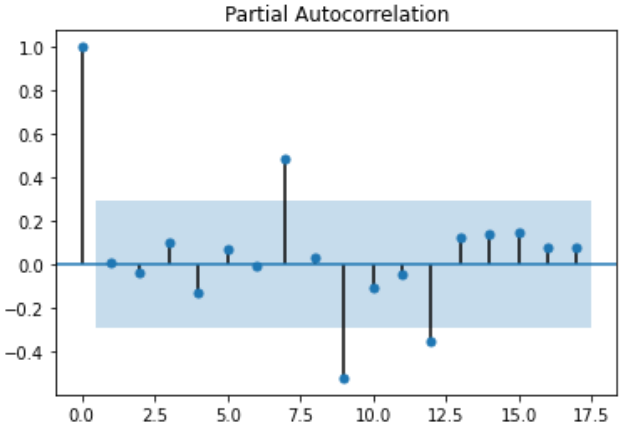
Para os parâmetros de ordem $p = 3$ e $q = 1$ temos que nossa série estacionária é estacionaria e invertível, essas características são relevantes pois indica que a maneira como a série se modifica ao longo do tempo não muda, tendo um padrão na sua variação e condiz com as primeiras análises da série de que ela não tem tendência e nem sazonalidade evidentes, pois esses são pressupostos da estacionaridade, já a invertibilidade indica que as nossas previsões se aproximam da realidade. Para esse modelo obtivemos o AIC de 343.686 e RMSE de 6.7602. É possível observar que a predição começa a se estabilizar ao longo do tempo, isso pode ter sido causado pelo parâmetro q =1, dando mais importante para valores mais próximos do final da série do que no início.



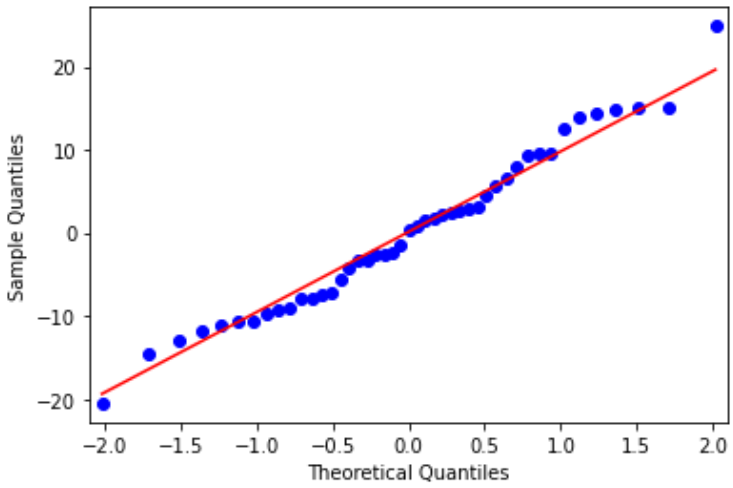
2020-05-31 34.568018	2020-06-03 36.903634	2020-06-06 36.713698	2020-06-09 36.723138
2020-06-01 37.871975	2020-06-04 36.588909	2020-06-07 36.738680	
2020-06-02 37.722647	2020-06-05 36.630732	2020-06-08 36.731384	

Agora vamos analisar os resíduos para verificar se podemos levar em conta esse modelo na nossa análise. Do gráfico do ACF e PACF podemos retirar que os resíduos não são ruídos brancos. O teste Ljung-Box trata informações mais precisas.



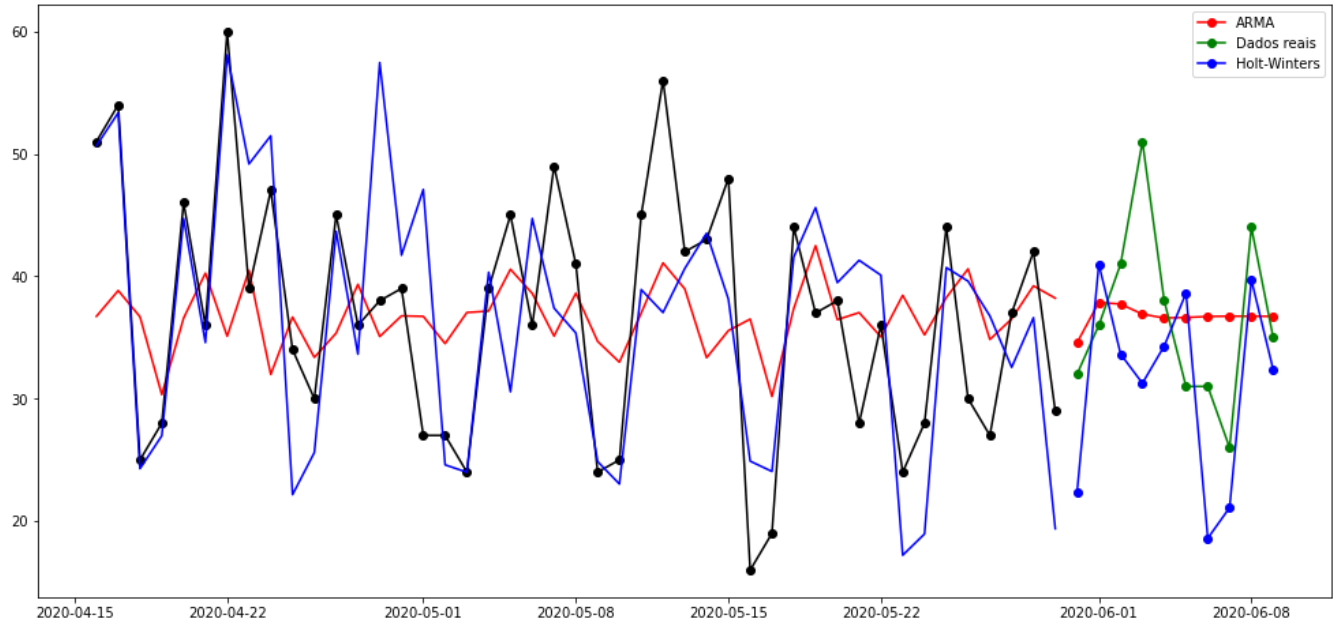


lag	p-value	lb_stat	rejectH0
1	0.94	0.01	False
2	0.97	0.06	False
3	0.92	0.49	False
4	0.89	1.12	False
5	0.94	1.24	False
6	0.97	1.26	False
7	0.25	8.98	False
8	0.32	9.27	False
9	0.05	17.00	True
10	0.07	17.04	False
11	0.08	18.07	False
12	0.05	21.32	True
13	0.05	22.47	True
14	0.06	23.12	False
15	0.08	23.37	False
16	0.05	25.95	False
17	0.07	26.15	False
18	0.08	26.99	False
19	0.08	28.19	False
20	0.06	30.62	False



Como podemos ver entre os lags 9 e 13 os resíduos rejeitam a hipótese que os resíduos são auto correlacionados. Como é um número pequeno e só ocorre no lag 9 podemos considerar que nossos resíduos não são auto correlacionados. No teste de normalidade temos um $p - value = 0.6242$ e o gráfico de qqplot acima, o que indica que nossos resíduos seguem uma distribuição normal, podemos então considerar esse modelo para análise.

Plotando os modelos e as previsões em um mesmo gráfico podemos ter uma visão mais ampla dos modelos em relação aos dados reais, logo abaixo se encontra dados descritivos sobre as previsões do modelo de holt-winters, ARMA e os dados reais após 10 dias.



count	10.000000
mean	31.294289
std	8.015891
min	18.602042
25%	24.595827
50%	33.008584
75%	37.535955
max	40.972324

count	10.000000
mean	36.719281
std	0.883351
min	34.568018
25%	36.651474
50%	36.727261
75%	36.862395
max	37.871975

count	10.000000
mean	36.500000
std	7.321961
min	26.000000
25%	31.250000
50%	35.500000
75%	40.250000
max	51.000000

6. Conclusão

Podemos concluir então com a análise feita nesse trabalho que houve sim uma diminuição nos CVP no período de uma quarentena mais rígida no Munícipio do Recife em relação ao ano anterior. Os modelos escolhidos para modelar a nossa série temporal tiveram uma RMSE próxima, porém um AIC distante, portanto é observado que o modelo que melhor representou a série temporal foi o de Holt-Winters.

Utilizando dos dados descritivos é possível ver que a média dos 10 dias após a saída da quarentena rígida está próxima do período em que ela estava mais rígida, oque pode indicar que logo após a saída da quarentena rígida não teve um significativo aumento dos CVP, ambos os modelos demostraram previsões com média inferior do que os dados reais (após 10 dias) e visualmente também é visto que o ARMA não conseguiu prever bem e o de Holt-Winters tem um bom desempenho nas últimas 3 observações, porém o holt-Winters forneceu valores abaixo do que o real, tanto nos dados de treinamento, quanto na previsão.

Os modelos desenvolvidos aqui podem ter um desempenho inferior a outra técnicas de modelagem de série temporal, então a forma como modelamos a série temporal pode também ser modificada com outras análises que demonstrem características desconhecidas da nossa série. Para trabalhos futuros, se faz pendente uma análise de toda a série histórica desde 2014 à 2020 e após o período mais rígido da quarentena, apenas com o isolamento social, para pode analisar se o período de quarentena realmente teve um efeito redutivo nos crimes violentos ao patrimônio.

7. Referências

[1] Pernambuco adota medidas mais rígidas para combate ao coronavírus, Disponível em: <Pernambuco adota medidas mais rígidas para combate ao coronavírus - Folha PE>. Acesso em: 27/11/2020.

[2] DECRETO Nº 49.055, DE 31 DE MAIO DE 2020, Disponível em: <Alepe Legis - Portal da Legislação Estadual de Pernambuco>. Acesso em: 27/11/2020.

[3] Crimes contra o Patrimônio, Disponível em: <https://www.sds.pe.gov.br/estatisticas/40-estatisticas/177-ccp> .Acesso em: 27/11/2020.

[4] Junior, I. Criminalidade e desempenho econômico: Uma análise em séries temporais para a Região Metropolitana do Recife, UFRPE, 2018.

[5] Nepomuceno, T. Moura, J. Silva, L. Costa, A. Alcohol and violent behavior among football spectators: An empirical assessment of Brazilian's criminalization. International Journal of Law, Crime and Justice, Vol. 51, 2017, 34 - 44.

[6] Gastwirth, J. Gel, Y. Miao, W. The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice. Statistical Science 2009, Vol. 24, No. 3, 343 – 360.

[7] Kim, T. T test as a parametric statistic. Korean Journal of Anesthesiology, Vol. 68, No. 6, 2015, 540 – 546.

Advanced Time Series Analysis with ARMA and ARIMA, Disponível em: <https://towardsdatascience.com/advanced-time-series-analysis-with-arma-and-arima-a7d9b589ed6d>. Acesso em: 27/11/2020.

White Noise Time Series with Python, Disponível em: <White Noise Time Series with Python (machinelearningmastery.com)>. Acesso em: 27/11/2020.

A Gentle Introduction to the Random Walk for Times Series Forecasting with Python, Disponível em: <A Gentle Introduction to the Random Walk for Times Series Forecasting with Python (machinelearningmastery.com)>. Acesso em: 27/11/2020.

How to Model Residual Errors to Correct Time Series Forecasts with Python, Disponível em: <How to Model Residual Errors to Correct Time Series Forecasts with Python (machinelearningmastery.com)> Acesso em: 27/11/2020.

Apêndice A:

Os algoritmos e códigos foram feitos na linguagem Python, tendo como suporte os pacotes Pandas, Numpy, Statsmodels, Scipy, Math, Matplotlib, dentre outros. A rotina desenvolvida para os códigos teve como apoio aas atividades desenvolvidas na cadeira, que auxiliaram a ter um panorama geral do problema e como poder resolver. O tratamento dos dados se deu em filtrar o município de Recife, as datas necessárias e a coluna Total, que é o cumulativo dos CVP naquela data. O link para o código se encontra logo abaixo:

https://colab.research.google.com/drive/1lFc-TYtrMD9d-5akJ_NvZ-STbovlZzMr?usp=sharing

Apêndice B:

2020-04-16	51	2020-05-05	45	2020-05-24	28
2020-04-17	54	2020-05-06	36	2020-05-25	44
2020-04-18	25	2020-05-07	49	2020-05-26	30
2020-04-19	28	2020-05-08	41	2020-05-27	27
2020-04-20	46	2020-05-09	24	2020-05-28	37
2020-04-21	36	2020-05-10	25	2020-05-29	42
2020-04-22	60	2020-05-11	45	2020-05-30	29
2020-04-23	39	2020-05-12	56		
2020-04-24	47	2020-05-13	42		
2020-04-25	34	2020-05-14	43		
2020-04-26	30	2020-05-15	48		
2020-04-27	45	2020-05-16	16		
2020-04-28	36	2020-05-17	19		
2020-04-29	38	2020-05-18	44		
2020-04-30	39	2020-05-19	37		
2020-05-01	27	2020-05-20	38		
2020-05-02	27	2020-05-21	28		
2020-05-03	24	2020-05-22	36		
2020-05-04	39	2020-05-23	24		