

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

FELIPE ALBINO RODRIGUES

**ANÁLISE DA SÉRIE TEMPORAL NAS DESPESAS REALIZADAS PELOS
MINISTÉRIOS DA ECONOMIA, SAÚDE E EDUCAÇÃO**

Belo Horizonte
2021

FELIPE ALBINO RODRIGUES

**ANÁLISE DA SÉRIE TEMPORAL NAS DESPESAS REALIZADAS PELOS
MINISTÉRIOS DA ECONOMIA, SAÚDE E EDUCAÇÃO**

Trabalho de Conclusão de Curso apresentado
ao Curso de Especialização em Ciência de
Dados e Big Data como requisito parcial à
obtenção do título de especialista.

Belo Horizonte

2021

SUMÁRIO

1. Introdução.....	5
1.1. Contextualização.....	5
1.2. O problema proposto.....	6
2. Coleta de Dados.....	7
3. Processamento/Tratamento de Dados.....	9
3.1. Base de Dados Despesas Públicas do Governo Federal.....	9
3.2. Base de Dados IPCA (Índice de Preços no Consumidor).....	10
4. Análise e Exploração dos Dados.....	11
4.1. Importação da Base de Dados das Despesas Públicas do Governo Federal.....	12
4.2. Plotagem dos Ministérios escolhidos para o estudo.....	13
4.3. Transformando Dados Não Estacionários em Estacionários.....	18
4.3.1. Teste de Dickey-Fuller.....	18
4.3.2. Teste KPSS.....	19
4.3.3. Primeiro método de transformação de dados não estacionários para estacionários: Ajuste de Inflação.....	21
4.3.4. Segundo método de transformação de dados não estacionários para estacionários: Diferenciação.....	24
5. Criação de Modelos de Machine Learning.....	28
5.1. AUTO-ARIMA.....	29
5.1.1 AUTO-ARIMA para o Ministério da Economia com valores ajustados pela inflação.....	29
5.1.2. AUTO-ARIMA para o Ministério da Saúde com valores ajustados por diferenciação.....	31
5.1.3. AUTO-ARIMA para o Ministério da Educação com valores ajustados por diferenciação.....	32
5.2. Suavização Exponencial Holt-Winters.....	33
5.2.1 Holt-Winters para a base de despesas do Ministério da Economia com valores ajustados pela inflação.....	33

5.2.2. Holt-Winters para a base de despesas do Ministério da Economia com valores originais.....	34
5.2.3. Holt-Winters para a base de despesas do Ministério da Saúde com valores originais.....	35
5.2.4. Holt-Winters para a base de despesas do Ministério da Educação com valores originais.....	36
5.3. Facebook Prophet.....	37
5.3.1. Facebook Prophet para a base de despesas do Ministério da Economia com os valores originais.....	37
5.3.2. Facebook Prophet para a base de despesas do Ministério da Saúde com valores originais.....	38
5.3.3. Facebook Prophet para a base de despesas do Ministério da Educação com valores originais.....	39
6. Apresentação dos Resultados.....	40
6.1. AUTO-ARIMA – Previsão 12 meses futuros.....	40
6.2. Suavização Exponencial Holt-Winters – Previsão 12 meses futuros.....	42
6.3. Facebook Prophet – Previsão 12 meses futuros.....	43
6.3.1. Facebook Prophet – Ministério da Economia com os valores originais (12 meses futuros).....	43
6.3.2. Facebook Prophet – Ministério da Saúde com os valores originais (12 meses futuros).....	44
6.3.3. Facebook Prophet – Ministério da Educação com os valores originais (12 meses futuros).....	45
6.4. Conclusões.....	46
7. Links.....	47
REFERÊNCIAS.....	48

1. Introdução

1.1. Contextualização

Em um cenário de pandemia em decorrência do vírus COVID-19, popularmente conhecido como “Coronavírus”, líderes mundiais buscam, cada vez mais, recursos financeiros que auxiliem a sobrevivência de seus povos, fomentem seus mercados e projetem um futuro próspero a médio e longo prazo. Para isso ser possível, cada país é obrigado a controlar seus gastos e a programar a melhor forma de receber recursos, criando um ambiente saudável entre despesas e receitas.

No Brasil, o governo atual (eleito em 2019) distribuiu suas áreas de atuação em 21 Ministérios, a saber:

- Casa Civil;
- Justiça e Segurança Pública;
- Defesa;
- Relações Exteriores;
- Economia;
- Infraestrutura;
- Agricultura, Pecuária e Abastecimento;
- Educação;
- Cidadania;
- Saúde;
- Minas e Energia;
- Comunicações;
- Ciência, Tecnologia e Inovações;
- Meio Ambiente;
- Turismo;
- Desenvolvimento Regional;
- Controladoria-Geral da União;
- Mulher, da Família e dos Direitos Humanos;
- Secretaria-Geral da Presidência;
- Secretaria de Governo;
- Gabinete de Segurança Institucional;

Cada Ministério é responsável por uma área específica dentro da política de investimentos da União, cabendo mais recursos para gastos a alguns em detrimento de outros ministérios. Isso faz parte do direcionamento que cada governo dá para o desenvolvimento de sua nação.

Este Trabalho de Conclusão de Curso procura abordar e estudar a parte das Despesas e, para isso, tomou por base o orçamento de Despesas Públicas do Brasil/União, detalhando, especificamente, os Ministérios da Economia, Saúde e Educação. Esses últimos foram os escolhidos para uma melhor abordagem e estudo tendo em vista sua representatividade no orçamento de despesas da União e na sua capacidade de articular políticas públicas que atinjam diretamente a população.

1.2. O problema proposto

O problema analisado neste trabalho visa prever 12 meses de gastos/despesas/investimentos dos Ministérios da Economia, Saúde e Educação, para além da base histórica fornecida. Ou seja, tratamos de Análise Preditiva de Séries Temporais.

Diante do cenário de pandemia do Coronavírus, será possível analisar o tamanho do impacto orçamentário nas despesas destes Ministérios, tanto quanto observar quais áreas sofreram mais investimentos nos últimos anos e quais áreas foram as mais demandadas no cenário atual, pela gravidade do vírus.

Tomamos como base de dados:

- Detalhamento de Despesas do Governo Federal: dados das despesas realizadas pela União com todos os seus Ministérios;
- IPCA/Inflação: dado histórico do IPCA (Índice de preços no consumidor) que servirá de base para atualização histórica dos valores das Despesas ao tempo atual.

O objetivo é analisar a política de gastos/investimentos do Brasil no histórico de 2014 até 2020, e tentar prever a continuidade dos gastos por mais 12 meses, a partir de janeiro de 2021, analisando o comportamento das séries temporais ao

longo do tempo, a forma comparativa entre os Ministérios da Economia, Saúde e Educação, utilizando, para isso, três bibliotecas de modelos preditivos de análise:

- AUTO-ARIMA;
- Suavização Exponencial Holt-Winters; e
- Facebook Prophet.

Utilizaremos a linguagem Python pela plataforma Jupyter Notebook nas análises.

2. Coleta de Dados

Os dados de Detalhamento das Despesas Públicas do Governo Federal foram extraídos diretamente do site do Portal da Transparência¹. Faz-se necessárias as seguintes observações sobre os dados:

- Os dados foram fornecidos em planilhas no formato “XLSX”;
- Os dados são fornecidos por período em Ano/Mês e com limitações de até 1000 linhas por relatório;
- Por este motivo, extraímos 27 planilhas diferentes para obter todos os dados necessários, tendo em vista a limitação expressa do site.
- O leiaute da planilha apresenta as seguintes informações:

Mês Ano	Órgão Superior	Órgão/Entidade Vinculada	Valor Pago
01/2014	53000 - Ministério do Desenvolvimento Regional	53901 - Fundo Constitucional de Financiamento do Norte	99878580.37
01/2014	53000 - Ministério do Desenvolvimento Regional	74918 - Fundo de Desenvolvimento do Nordeste	0.00
01/2014	25000 - Ministério da Economia	25904 - Fundo de Estabilidade do Seguro Rural	0.00
01/2014	53000 - Ministério do Desenvolvimento Regional	53903 - Fundo Constitucional de Financiamento do Nordeste	333773710.86
01/2014	25000 - Ministério da Economia	28501 - Fundo Nacional de Desestatização	0.00
01/2014	25000 - Ministério da Economia	71902 - Fundo Soberano do Brasil	0.00
01/2014	25000 - Ministério da Economia	28500 - Fundo de Garantia para a Promoção da Competitividade	54409.74
01/2014	39000 - Ministério da Infraestrutura	62901 - Fundo Nacional de Aviação Civil	0.00
01/2014	25000 - Ministério da Economia	25914 - Fundo de Garantia à Exportação	0.00
01/2014	53000 - Ministério do Desenvolvimento Regional	53902 - Fundo Constitucional de Financiamento do Centro-Oeste	0.00
01/2014	39000 - Ministério da Infraestrutura	56901 - Fundo Nacional de Segurança e Educação de Trânsito	5465.50
01/2014	20113 - Ministério do Planejamento, Desenvolvimento e Gestão	20113 - Ministério do Planejamento, Desenvolvimento e Gestão - Unidades com vínculo direto	6115671.40
01/2014	54000 - Ministério do Turismo	34902 - Fundo Nacional de Cultura	1286.80
01/2014	44000 - Ministério do Meio Ambiente	44204 - Fundo Nacional do Meio Ambiente	1091.30
01/2014	57000 - Ministério das Mulheres, Igualdade Racial, da Juventude e dos Direitos Humanos	64902 - Fundo Nacional do Idoso	0.00
01/2014	57000 - Ministério das Mulheres, Igualdade Racial, da Juventude e dos Direitos Humanos	64901 - Fundo Nacional para a Criança e o Adolescente	0.00
01/2014	39000 - Ministério da Infraestrutura	62000 - Secretaria de Aviação Civil	1151215.55
01/2014	49000 - Ministério do Desenvolvimento Agrário	49000 - Ministério do Desenvolvimento Agrário - Unidades com vínculo direto	0.00
01/2014	54000 - Ministério do Turismo	20403 - Fundação Casa de Rui Barbosa	2388866.75
01/2014	20000 - Presidência da República	67000 - Secretaria de Políticas de Promoção da Igualdade Racial	56609.70
01/2014	30000 - Ministério da Justiça e Segurança Pública	30211 - Conselho Administrativo de Defesa Econômica	761879.09
01/2014	25000 - Ministério da Economia	25916 - Fundo Contingente da Extinta RFFSA-FCMP	0.00
01/2014	24000 - Ministério da Ciência, Tecnologia, Inovações e Comunicações	20402 - Agência Espacial Brasileira	683429.42
01/2014	26000 - Ministério da Educação	26449 - Universidade Federal do Cariri	0.00
01/2014	54000 - Ministério do Turismo	20203 - Agência Nacional do Cinema	5037698.07
01/2014	54000 - Ministério do Turismo	20404 - Fundação Biblioteca Nacional	5268461.20

¹ <http://www.portaldatransparencia.gov.br/despesas>

- A coluna Órgão Superior representa os Ministérios, e cada Ministério possui diversas Entidades Vinculadas;
- Como afirmado anteriormente, foram 27 planilhas que integram um único arquivo, sendo o link disponibilizado ao final deste Trabalho.

Os dados do IPCA (Índice de Preços no consumidor) foram extraídos do site do IBGE², entidade responsável pelo seu cálculo e fornecimento do mesmo ao público. Faz-se necessárias as seguintes observações sobre os dados:

- Os dados foram fornecidos em planilha no formato “XLS”;
- A série histórica contempla dados de 1994 a 2021, portanto, bastou apenas uma planilha para todas as informações necessárias neste Trabalho;
- O leiaute da planilha contém as seguintes informações:

SÉRIE HISTÓRICA DO IPCA							
(continua)							
ANO	MÊS	NÚMERO ÍNDICE (DEZ 93 = 100)	VARIACÃO (%)				
			NO MÊS	3 MESES	6 MESES	NO ANO	12 MESES
1994	JAN	141.31	41.31	162.13	533.33	41.31	2,693.84
	FEV	198.22	40.27	171.24	568.17	98.22	3,035.71
	MAR	282.96	42.75	182.96	602.93	182.96	3,417.39
	ABR	403.73	42.68	185.71	648.92	303.73	3,828.49
	MAI	581.49	44.03	193.36	695.71	481.49	4,331.19
	JUN	857.29	47.43	202.97	757.29	757.29	4,922.60
	JUL	915.93	6.84	126.87	548.17	815.93	4,005.08
	AGO	932.97	1.86	60.44	370.67	832.97	3,044.89
	SET	947.24	1.53	10.49	234.76	847.24	2,253.15
	OUT	972.06	2.62	6.13	140.77	872.06	1,703.17
	NOV	999.37	2.81	7.12	71.86	899.37	1,267.54
	DEZ	1016.46	1.71	7.31	18.57	916.46	916.46
1995	JAN	1033.74	1.70	6.35	12.86	1.70	631.54
	FEV	1044.28	1.02	4.49	11.93	2.74	426.83
	MAR	1060.47	1.55	4.33	11.95	4.33	274.78
	ABR	1086.24	2.43	5.08	11.75	6.87	169.05
	MAI	1115.24	2.67	6.80	11.59	9.72	91.79
	JUN	1140.44	2.26	7.54	12.20	12.20	33.03
	JUL	1167.35	2.36	7.47	12.92	14.84	27.45
	AGO	1178.91	0.99	5.71	12.89	15.98	26.36
	SET	1190.58	0.99	4.40	12.27	17.13	25.69
	OUT	1207.37	1.41	3.43	11.15	18.78	24.21
	NOV	1225.12	1.47	3.92	9.85	20.53	22.59
	DEZ	1244.23	1.56	4.51	9.10	22.41	22.41

Apenas com esses dados não é possível ajustar/atualizar os valores das despesas com a data proposta pelo projeto, contudo são suficientes para o cálculo que será detalhado adiante.

² <https://www.ibge.gov.br/estatisticas/economicas/precos-e-custos/9256-indice-nacional-de-precos-ao-consumidor-amplo.html?=&t=o-que-e>

3. Processamento/Tratamento de Dados

O tratamento inicial dos dados se deu diretamente nas planilhas, por meio do aplicativo Microsoft Excel. Posteriormente, fizemos pequenos ajustes via Python para encaixar a necessidade da proposta deste Trabalho com o resultado final obtido. Foi preciso, entre outras coisas, agrupar diversas planilhas e transformar alguns dados extraídos para uma base que agregasse todas as informações necessárias para a correta leitura do projeto.

3.1 – Base de Dados Despesas Públicas do Governo Federal

Para transformar em uma única base de dados, agrupamos 27 planilhas distintas em apenas um documento de formato “XLSX”, com o total de **20.552 registros** em **84 meses**. Além disso, realizamos as seguintes etapas:

1. Alterações realizadas no aplicativo Microsoft Excel:

- a) Identificamos 11 linhas em que a coluna “órgão” (Ministério) constava “Sem Informação”, seguido da coluna “Entidade Vinculada” com a informação “-3”. Para essas linhas, as demais colunas vinham com valores informados “zero”. Dessa forma, não sendo possível utilizar essas linhas para a nossa análise, retiramos as 11 da planilha final;
- b) A coluna “Valor Pago” apresentava valores com sinal negativo incorretamente. Tratamos esse atributo para a correta leitura do mesmo posteriormente;
- c) Reformatamos o campo “Valor Pago” para tipo float;
- d) Reformatamos o campo “Mês/Ano” para tipo date.

2. Alterações realizadas na plataforma Jupyter Notebook:

- a) Agrupamos a coluna “Entidade Vinculada” para cada “Órgão” respectivo, acumulando seus valores por Ministério, conforme demonstramos:

```
[ ] df_result = df_desp.groupby(['periodo', 'orgao'])['vl_pago'].sum()

[ ] df_result
```

periodo	orgao	
2014-01-01	20000 - Presidência da República	6.519071e+07
	20113 - Ministério do Planejamento, Desenvolvimento e Gestão	6.115671e+06
	22000 - Ministério da Agricultura, Pecuária e Abastecimento	6.272410e+08
	24000 - Ministério da Ciência, Tecnologia, Inovações e Comunicações	7.881719e+08
	25000 - Ministério da Economia	2.307079e+11
	...	
2020-12-01	53000 - Ministério do Desenvolvimento Regional	4.035312e+09
	54000 - Ministério do Turismo	1.805339e+09
	55000 - Ministério da Cidadania	3.453721e+10
	63000 - Advocacia-Geral da União	3.854883e+08
	81000 - Ministério da Mulher, Família e Direitos Humanos	2.592352e+07

Name: vl_pago, Length: 1927, dtype: float64

b) Transformamos as linhas “Órgão” em colunas para melhor tratamento:

```
[ ] df_result3 = df_result2.pivot_table('vl_pago', 'periodo', 'orgao')

[ ] df_result3.head()
```

orgao	20000 - Presidência da República	20113 - Ministério do Planejamento, Desenvolvimento e Gestão	22000 - Ministério da Agricultura, Pecuária e Abastecimento	24000 - Ministério da Ciência, Tecnologia, Inovações e Comunicações	25000 - Ministério da Economia	26000 - Ministério da Educação	30000 - Ministério da Justiça e Segurança Pública	32000 - Ministério de Minas e Energia	33000 - Ministério da Previdência Social
periodo									
2014-01-01	6.519071e+07	6115671.40	6.272410e+08	7.881719e+08	2.307079e+11	4.828033e+09	6.059899e+08	1.434163e+09	19074822.83
2014-02-01	8.479716e+07	7413.41	7.469514e+08	6.870223e+08	8.926030e+10	7.076257e+09	6.565594e+08	3.498208e+09	26530306.94
2014-03-01	8.807581e+07	539261.48	7.540673e+08	6.506453e+08	1.554504e+11	6.493763e+09	6.851543e+08	3.334442e+09	24732270.44
2014-04-01	1.130643e+08	449034.76	7.810936e+08	7.230355e+08	1.923474e+11	9.155573e+09	7.050006e+08	1.946108e+09	26548176.70
2014-05-01	1.147648e+08	131330.31	8.134798e+08	9.396719e+08	6.895201e+10	9.063304e+09	7.442369e+08	3.573487e+09	25381394.24

3.2 – Base de Dados IPCA (Índice de Preços no Consumidor)

Para a base de dados do IPCA, foi preciso adequar por completo o leiaute fornecido pelo site do IBGE à necessidade do trabalho proposto neste projeto.

Com base nesse arquivo, restringimos a série histórica para o período de 2002 até 2020 e mantivemos apenas a coluna da variação(%) “NO MÊS” para fins de cálculo do acumulado. Além disso, alteramos as seguintes colunas/informações à planilha:

- a) Transformamos a coluna “Mês” para numeral;
- b) Adicionamos a coluna “Data” agregando as colunas anteriores “ANO” e “Mês”;
- c) Alteramos o nome da coluna “NO MÊS” para “IPCA”;
- d) Adicionamos coluna representando a quantidade de dias de determinado período em relação à data de atualização/correção da inflação (neste caso do projeto, dezembro de 2020);
- e) Adicionamos a coluna “Acumulado”, trazendo o cálculo de “IPCA” pela quantidade de dias acumulado em relação a dezembro de 2020.

A planilha ficou com o seguinte leiaute:

ANO	Mês	Data	IPCA			Acumulado	Acum
2002	11	11/1/2002	3.02%	103.02%	218	286.74%	2.87
2002	12	12/1/2002	2.10%	105.18%	217	282.92%	2.83
2003	1	1/1/2003	2.25%	107.55%	216	280.43%	2.80
2003	2	2/1/2003	1.57%	109.24%	215	278.04%	2.78
2003	3	3/1/2003	1.23%	110.58%	214	276.27%	2.76
2003	4	4/1/2003	0.97%	111.65%	213	275.61%	2.76
2003	5	5/1/2003	0.61%	112.34%	212	274.62%	2.75
2003	6	6/1/2003	-0.15%	112.17%	211	273.91%	2.74
2003	7	7/1/2003	0.20%	112.39%	210	274.95%	2.75
2003	8	8/1/2003	0.34%	112.77%	209	275.80%	2.76
2003	9	9/1/2003	0.78%	113.65%	208	275.61%	2.76
2003	10	10/1/2003	0.29%	113.98%	207	274.92%	2.75
2003	11	11/1/2003	0.34%	114.37%	206	274.35%	2.74
2003	12	12/1/2003	0.52%	114.97%	205	271.23%	2.71
2004	1	1/1/2004	0.76%	115.84%	204	269.85%	2.70
2004	2	2/1/2004	0.61%	116.55%	203	269.58%	2.70
2004	3	3/1/2004	0.47%	117.09%	202	269.69%	2.70
2004	4	4/1/2004	0.37%	117.53%	201	269.40%	2.69

4. Análise e Exploração dos Dados

Utilizaremos para este tópico a plataforma Jupyter Notebook e a linguagem de programação Python, apresentando telas do código para facilitar a compreensão. Ao final deste Trabalho, será disponibilizado link do notebook completo para consulta.

4.1 Importação da Base de Dados das Despesas Públicas do Governo Federal:

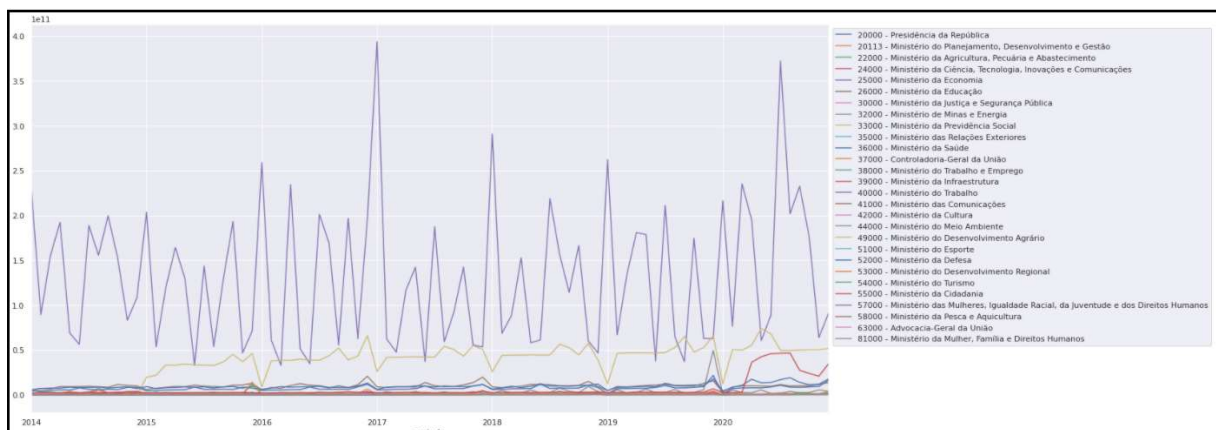
Como já explicado no tópico **3.1.2** deste projeto, após a importação da planilha tratada, no formato Excel (.xlsx), e dos tratamentos daquele tópico, foi gerado o DataFrame seguinte:

```
[19] df_result3.info()

<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 84 entries, 2014-01-01 to 2020-12-01
Data columns (total 28 columns):
 #   Column                                                                 Non-Null Count  Dtype
---  -
 0   20000 - Presidência da República                                     84 non-null    float64
 1   20113 - Ministério do Planejamento, Desenvolvimento e Gestão       45 non-null    float64
 2   22000 - Ministério da Agricultura, Pecuária e Abastecimento        84 non-null    float64
 3   24000 - Ministério da Ciência, Tecnologia, Inovações e Comunicações 84 non-null    float64
 4   25000 - Ministério da Economia                                       84 non-null    float64
 5   26000 - Ministério da Educação                                       84 non-null    float64
 6   30000 - Ministério da Justiça e Segurança Pública                 84 non-null    float64
 7   32000 - Ministério de Minas e Energia                              84 non-null    float64
 8   33000 - Ministério da Previdência Social                           84 non-null    float64
 9   35000 - Ministério das Relações Exteriores                         84 non-null    float64
10  36000 - Ministério da Saúde                                          84 non-null    float64
11  37000 - Controladoria-Geral da União                                84 non-null    float64
12  38000 - Ministério do Trabalho e Emprego                           36 non-null    float64
13  39000 - Ministério da Infraestrutura                                84 non-null    float64
14  40000 - Ministério do Trabalho                                       16 non-null    float64
15  41000 - Ministério das Comunicações                                48 non-null    float64
16  42000 - Ministério da Cultura                                       36 non-null    float64
17  44000 - Ministério do Meio Ambiente                                84 non-null    float64
18  49000 - Ministério do Desenvolvimento Agrário                     48 non-null    float64
19  51000 - Ministério do Esporte                                       41 non-null    float64
20  52000 - Ministério da Defesa                                        84 non-null    float64
21  53000 - Ministério do Desenvolvimento Regional                    84 non-null    float64
22  54000 - Ministério do Turismo                                       84 non-null    float64
23  55000 - Ministério da Cidadania                                    84 non-null    float64
24  57000 - Ministério das Mulheres, Igualdade Racial, da Juventude e dos Direitos Humanos 35 non-null    float64
25  58000 - Ministério da Pesca e Aquicultura                          26 non-null    float64
26  63000 - Advocacia-Geral da União                                    84 non-null    float64
27  81000 - Ministério da Mulher, Família e Direitos Humanos          84 non-null    float64

dtypes: float64(28)
memory usage: 19.0 KB
```

Notamos a presença de 84 entradas, ou seja, 84 meses e 28 colunas (as colunas representando os Ministérios e correlatos a Ministério). Os dados plotados apresentam-se como segue:



É latente observar que o Ministério da Economia é o órgão que mais "gasta" dentro do orçamento da União. A partir disso, vamos listar os Ministérios que mais geram despesas para a União por ranking, facilitando algumas observações futuras:

	TOTAL
25000 - Ministério da Economia	1.092967e+13
33000 - Ministério da Previdência Social	3.149113e+12
26000 - Ministério da Educação	8.315516e+11
36000 - Ministério da Saúde	8.034681e+11
52000 - Ministério da Defesa	6.128835e+11
55000 - Ministério da Cidadania	5.250087e+11

Para fins de estudo nesse Trabalho de Conclusão, optamos por analisar três Ministérios de suma importância para o Brasil e que, pelo ranking demonstrado, estão no top 5 de despesas para a União:

Ministério da Economia

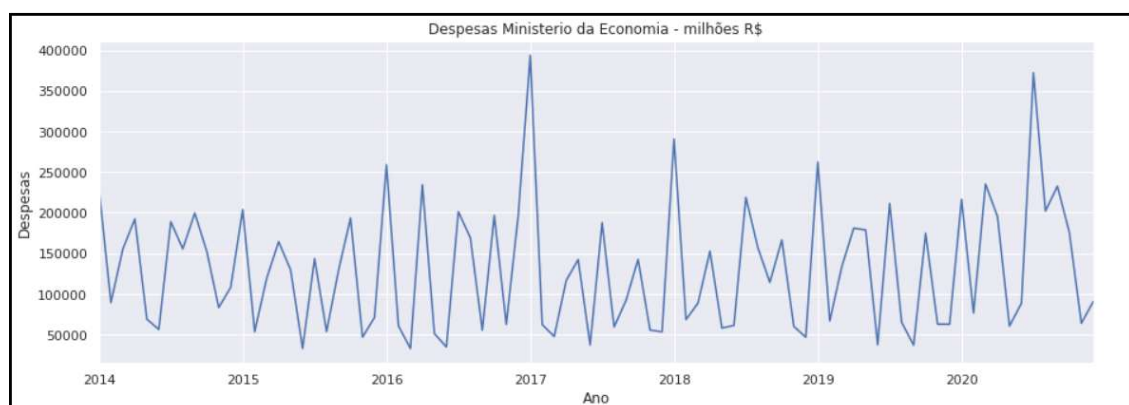
Ministério da Saúde

Ministério da Educação

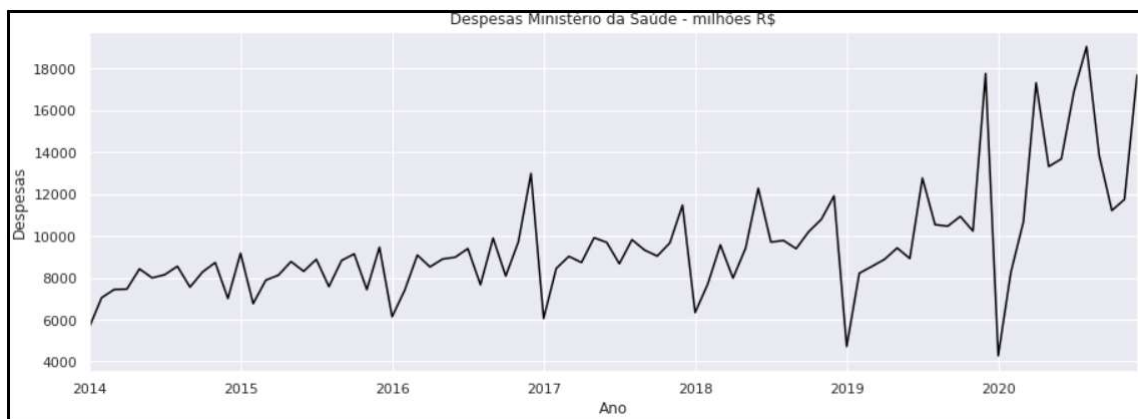
4.2 Plotagem dos Ministérios escolhidos para o estudo

Para facilitar nossa análise quanto a tendência e sazonalidade, demonstramos em gráficos separados os Ministérios trabalhados neste projeto:

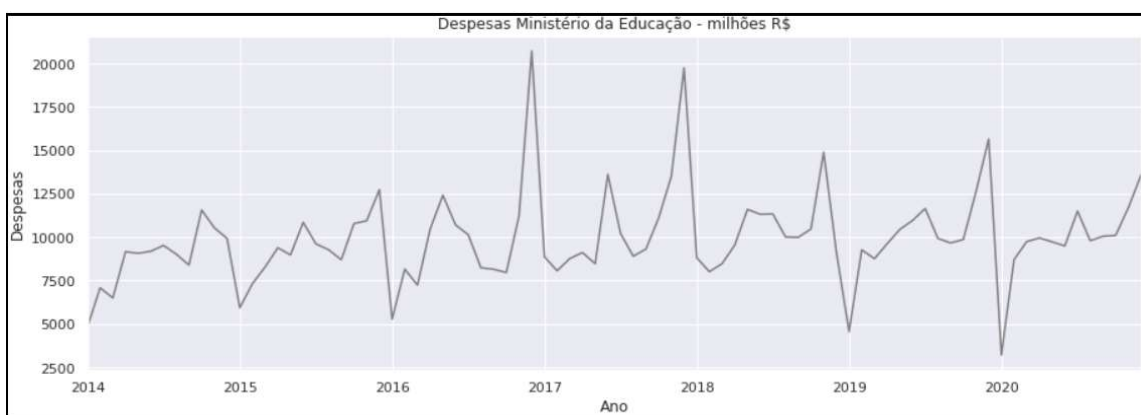
a) Ministério da Economia



b) Ministério da Saúde



c) Ministério da Educação



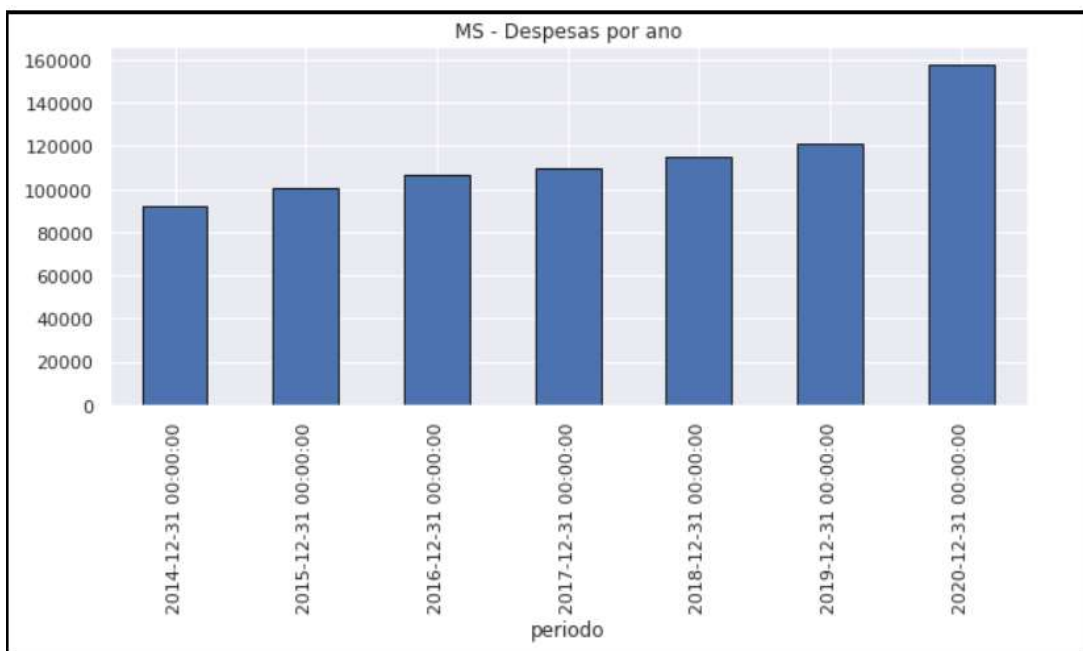
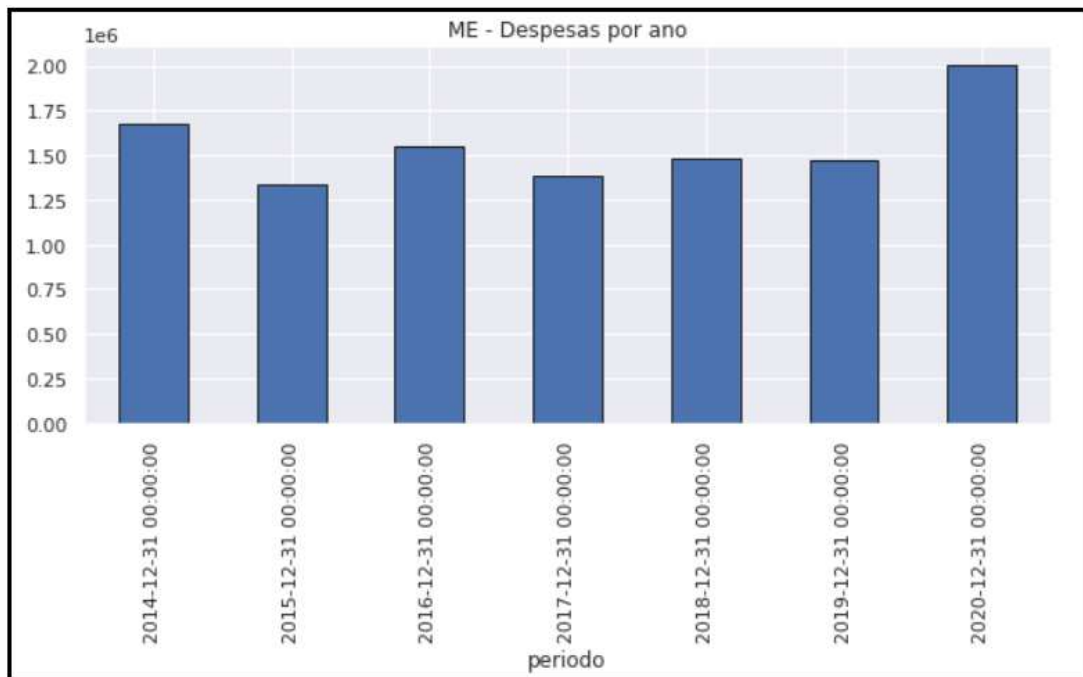
Foi possível observar uma ligeira tendência de aumento/crescimento nas despesas desses Ministérios. Além disso, há sazonalidade nos gastos. Essas informações ficarão mais claras adiante.

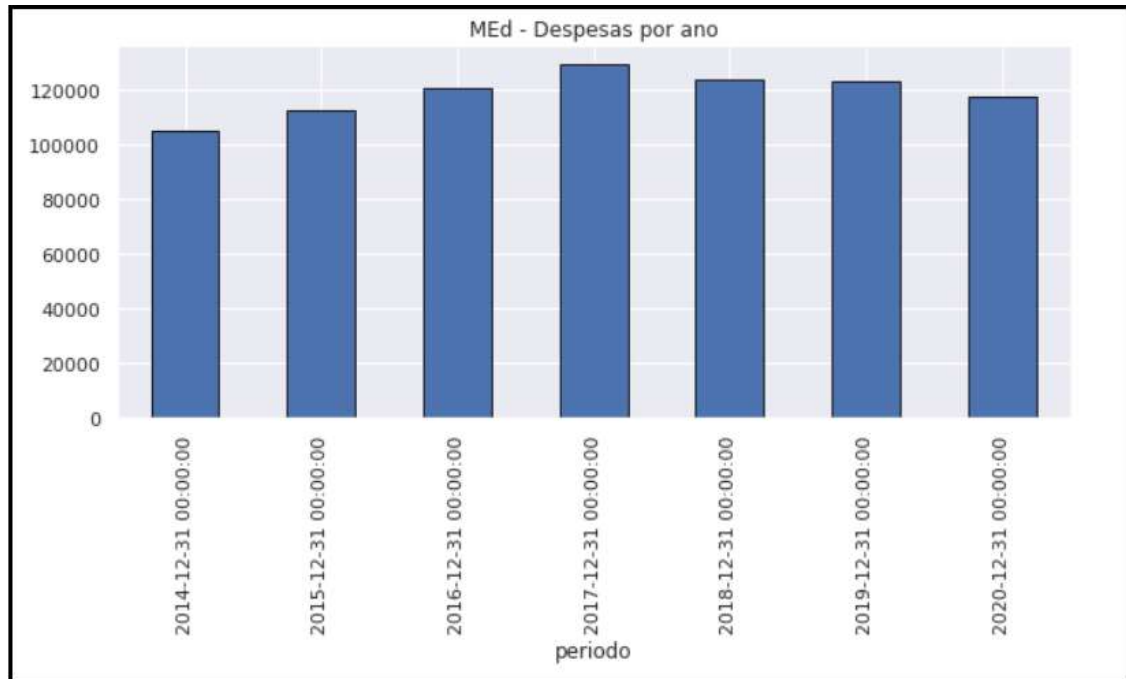
O gráfico das despesas do Ministério da Economia (**ME**) é bem diferente dos outros listados. Por ser o Ministério que gerencia e dita as regras de gastos da União, talvez esse seja um dos motivos de suas despesas ser tão peculiar e de grande volume. Os Ministérios da Saúde(**MS**) e Educação(**MEd**) possuem tendência e sazonalidade semelhantes.

Interessante notar nesse gráfico o aumento de despesas para o ano de 2020 no Ministério da Saúde, com redução da mesma no Ministério da Educação, em decorrência da pandemia do Coronavírus.

Para facilitar a compreensão dos dados, toda base foi dividida por 1.000.000 (um milhão), para que a partir deste momento os dados sejam apresentados em valores por milhão.

Ao analisar as Despesas Públicas por Ministério ao ano/anual, fica ainda mais fácil identificar a tendência de crescimento dos gastos:

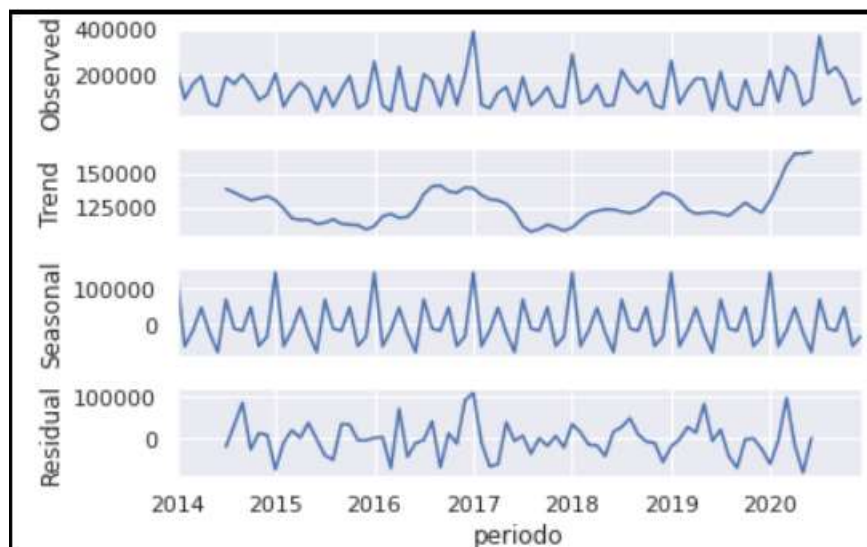




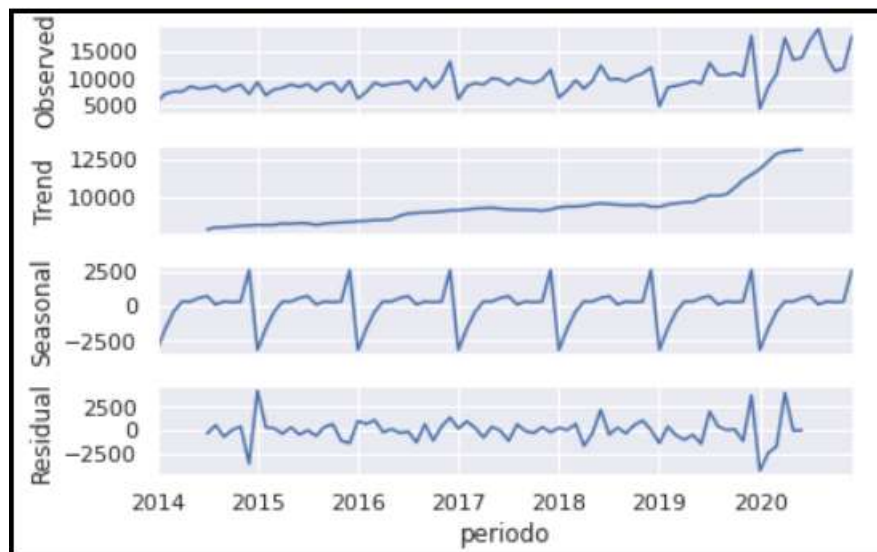
Foi possível observar de forma clara o aumento de despesas para o ano de 2020 nos Ministérios da Economia e Saúde, e a consequente redução de gastos no Ministério da Educação.

Não obstante os gráficos disponibilizados, fez-se necessária a utilização do **STATSMODEL** para obter tendência. Assim, realizamos a **decomposição temporal** como forma de isolar os componentes ETS (Erro, Tendência e Sazonalidade).

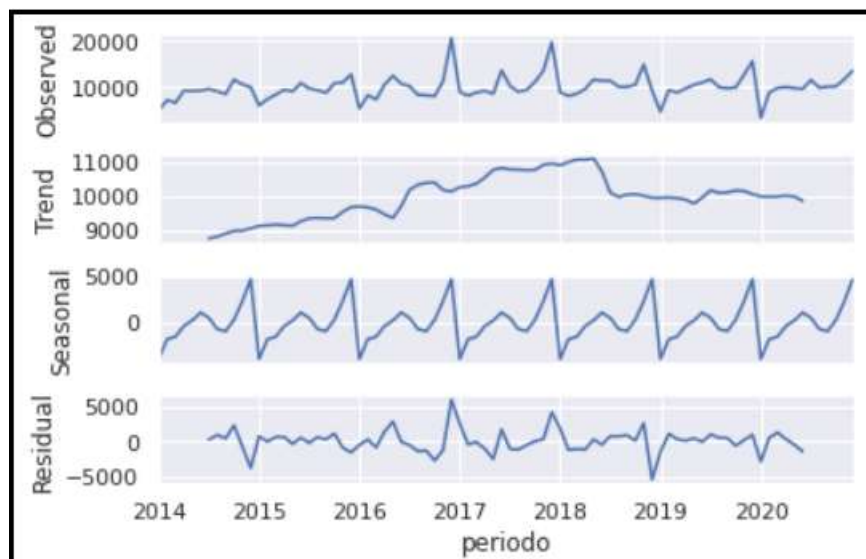
- Ministério da Economia



- Ministério da Saúde



- Ministério da Educação



Isoladas as variáveis Tendência e Sazonalidade, chegamos a conclusão de que, de fato, os três Ministérios estudados possuem aumento de gastos e sazonalidade bem estabelecida.

A partir daqui, é de suma importância para determinar os modelos preditivos que serão utilizados neste estudo sabermos se os dados apresentados são estacionários ou não estacionários.

4.3 Transformando Dados Não Estacionários em Estacionários

Uma série estacionária tem média constante durante o tempo, não existindo tendência de alta ou de baixa. A razão disso é que tendo uma média constante com variações ao redor desta média fica muito mais fácil de extrapolar ao futuro.

Uma série temporal estacionária é aquela cujas propriedades estatísticas como a média, a variância e a autocorrelação são constantes ao longo do tempo. Assim, uma série não estacionária é uma cujas propriedades estatísticas mudam com o tempo.

Para identificar se determinada base de dados apresenta característica estacionária ou não estacionária, podemos, entre outras ferramentas, utilizar dois testes: Teste de Dickey-Fuller³ e teste KPSS⁴.

4.3.1 Teste de Dickey-Fuller

O teste foi criado por Dickey e Fuller em 1979, daí vem o nome, e tem como base verificar se um modelo autorregressivo tem ou não raiz unitária.

$$y_t = \rho y_{t-1} + u_t$$

Usamos o valor de P base 5%, ou seja, caso o valor P esteja abaixo desse percentual, significa que a série é estatisticamente estacionária.

Assim, executamos o teste para os Ministérios estudados e os resultados são os seguintes:

a) Ministério da Economia

E nao-estacionaria	
Teste Estatístico Dickey Fuller	-2.6843
Valor-P	0.0768
Lags Usados	12.0000
Número de observações usadas	71.0000
Valores Críticos (1%)	-3.5260
Valores Críticos (5%)	-2.9032
Valores Críticos (10%)	-2.5890
dtype: float64	

³ https://en.wikipedia.org/wiki/Dickey%E2%80%93Fuller_test

⁴ https://en.wikipedia.org/wiki/kpss_test

b) Ministério da Saúde

```

É não-estacionária
Teste Estatístico Dickey Fuller      2.3060
Valor-P                             0.9990
Lags Usados                         11.0000
Número de observações usadas       72.0000
Valores Críticos (1%)              -3.5246
Valores Críticos (5%)              -2.9026
Valores Críticos (10%)             -2.5887
dtype: float64

```

c) Ministério da Educação

```

É não-estacionária
Teste Estatístico Dickey Fuller     -1.7018
Valor-P                             0.4302
Lags Usados                         11.0000
Número de observações usadas       72.0000
Valores Críticos (1%)              -3.5246
Valores Críticos (5%)              -2.9026
Valores Críticos (10%)             -2.5887
dtype: float64

```

O teste de Dickey-Fuller nos informa que os três Ministérios em estudo possuem característica de dados **não estacionários**. Para confirmar o teste, aplicaremos o segundo a seguir.

4.3.2 Teste KPSS

O teste foi criado por Denis Kwiatkowski, Peter C. B. Phillips, Peter Schmidt e Yongcheol Shin, denominado teste KPSS devido a seus nomes. Tem por finalidade determinar a estacionariedade em uma série temporal.

O teste KPSS já pressupõe que a série é estacionária e só não será se o valor P for inferior a 5%, ou se o teste estatístico for menor que algum valor crítico escolhido.

Da mesma forma, executamos o teste para os Ministérios estudados e os resultados são os seguintes:

a) Ministério da Economia

Teste Statístico KPSS	0.1495
Valor-P	0.1000
Lags Usados	12.0000
Valores Críticos (10%)	0.3470
Valores Críticos (5%)	0.4630
Valores Críticos (2.5%)	0.5740
Valores Críticos (1%)	0.7390
dtype: float64	
dtype: float64	

b) Ministério da Saúde

Teste Statístico KPSS	0.6311
Valor-P	0.0198
Lags Usados	12.0000
Valores Críticos (10%)	0.3470
Valores Críticos (5%)	0.4630
Valores Críticos (2.5%)	0.5740
Valores Críticos (1%)	0.7390
dtype: float64	

c) Ministério da Educação

Teste Statístico KPSS	0.4899
Valor-P	0.0439
Lags Usados	12.0000
Valores Críticos (10%)	0.3470
Valores Críticos (5%)	0.4630
Valores Críticos (2.5%)	0.5740
Valores Críticos (1%)	0.7390
dtype: float64	

Dadas as análises dos testes apresentados, ficou evidente que os três Ministérios possuem dados não estacionários. A partir desse momento, tratamos de transformar esses dados para **estacionários**, assim poderemos dispor de três modelos preditivos, um tratando especificamente dados estacionários e outros dois tratando dados não estacionários. Ou seja, usaremos nesse Trabalho de Conclusão de Curso modelos preditivos que utilizam a biblioteca AUTO-ARIMA, Suavização Exponencial Holt-Winters e Facebook Prophet.

4.3.3 Primeiro método de transformação de dados não estacionários para estacionários: Ajuste de Inflação

Valores expressos em qualquer tipo de moeda fiduciária, como no caso do Brasil, o Real (R\$), sofrem desvalorização e são relativos ao tempo que foram utilizados. Em janeiro de 2014, por exemplo, a despesa do Ministério da Economia estava algo em torno de R\$ 230.000.000,00 (duzentos e trinta milhões de reais). Se a despesa deste Ministério fosse gasta/investida nos dias de hoje, certamente muitos programas sociais ou políticas públicas do Governo Federal seriam afetados, pois o valor está defasado em relação aos dias atuais.

Para tentar tornar a série estacionária, colocaremos a série toda com base nos valores atuais pelo IPCA, acumulando do final do Trabalho proposto (dez/20) até o início do estudo. O link para o arquivo trabalhado nesta seção estará ao final deste projeto.

Assim, após a leitura do arquivo na plataforma Jupyter Notebook, temos o seguinte leiaute:

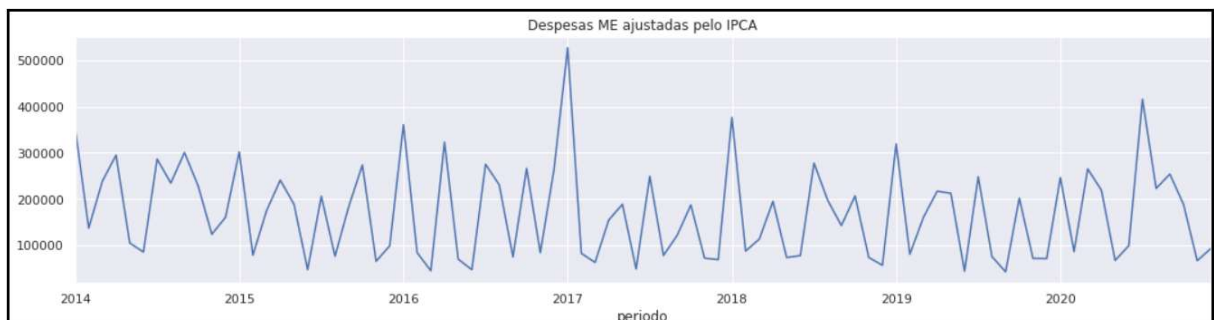
	ANO	Mês	IPCA	Unnamed: 4	Unnamed: 5	Acumulado	Acum
Data							
2002-11-01	2002	11	0.0302	1.030200	218	2.867418	2.867418
2002-12-01	2002	12	0.0210	1.051834	217	2.829223	2.829223
2003-01-01	2003	1	0.0225	1.075500	216	2.804265	2.804265
2003-02-01	2003	2	0.0157	1.092386	215	2.780354	2.780354
2003-03-01	2003	3	0.0123	1.105822	214	2.762673	2.762673

A partir de então, passamos a mesclar o campo referente à despesa de cada Ministério estudado, de acordo com o período, com o campo “Acumulado” do Data-Frame IPCA. Para isso utilizamos o comando merge do Python, conforme a seguir:

```
#Mesclando
index = bkp_gastos.index
bkp_gastos = bkp_gastos.merge(infl.loc[:,['ANO','Mês','Acumulado']], how='left', on=['ANO','Mês'])
bkp_gastos['Despesa ME Ajustada'] = bkp_gastos['25000 - Ministério da Economia'] * bkp_gastos['Acumulado']
bkp_gastos['Despesa MS Ajustada'] = bkp_gastos['36000 - Ministério da Saúde'] * bkp_gastos['Acumulado']
bkp_gastos['Despesa MEd Ajustada'] = bkp_gastos['26000 - Ministério da Educação'] * bkp_gastos['Acumulado']
bkp_gastos.set_index(index, inplace=True)
```

Como resultado, temos três novas colunas na entidade que representam os valores/gastos dos Ministérios corrigidos pela inflação. Passamos, então, a realizar o teste de Dickey-Fuller para cada valor ajustado, procurando observar se a correção pela inflação foi suficiente para transformar os dados em estacionários.

Infelizmente, para esse primeiro método, apenas o Ministério da Economia obteve êxito na transformação dos dados para estacionário, conforme podemos observar:



Teste Estatístico Dickey Fuller	-3.1597
Valor-P	0.0224
Lags Usados	12.0000
Número de observações usadas	71.0000
Valores Críticos (1%)	-3.5260
Valores Críticos (5%)	-2.9032
Valores Críticos (10%)	-2.5890
dtype:	float64

O teste estatístico de Dickey-Fuller apontou valor P abaixo do considerado não estacionário, ou seja, abaixo de 5%, o que confirma sua transformação em dados estacionários. É possível reparar, inclusive, uma maior linearidade no gráfico plotado.

Para os outros dois Ministérios não foi possível a transformação. Apesar de os gráficos apresentarem uma melhor linearidade, o teste de Dickey-Fuller considerou o valor de P acima da faixa dos 5% para cada um deles:

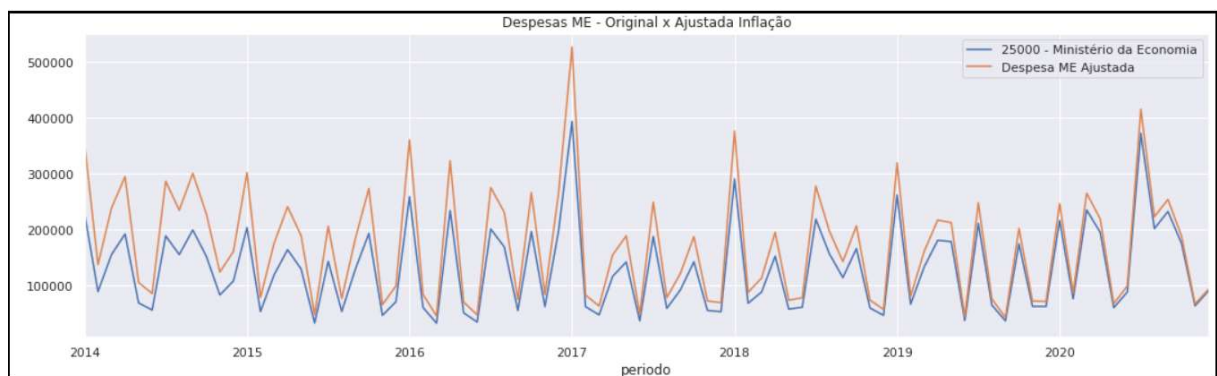
Teste Estatístico Dickey Fuller	0.4192
Valor-P	0.9822
Lags Usados	11.0000
Número de observações usadas	72.0000
Valores Críticos (1%)	-3.5246
Valores Críticos (5%)	-2.9026
Valores Críticos (10%)	-2.5887
dtype: float64	

Ministério da Saúde

Teste Estatístico Dickey Fuller	-0.0392
Valor-P	0.9552
Lags Usados	11.0000
Número de observações usadas	72.0000
Valores Críticos (1%)	-3.5246
Valores Críticos (5%)	-2.9026
Valores Críticos (10%)	-2.5887
dtype: float64	

Ministério da Educação

Como curiosidade, o gráfico a seguir representa as Despesas do Ministério da Economia (ME) com os valores originais comparados às Despesas do ME com os valores ajustados pela inflação:



Dessa forma, não sendo possível transformar os três Ministérios em dados estacionários pelo método da correção pela inflação, utilizaremos um segundo método muito prático: a **diferenciação**.

4.3.4 Segundo método de transformação de dados não estacionários para estacionários: Diferenciação

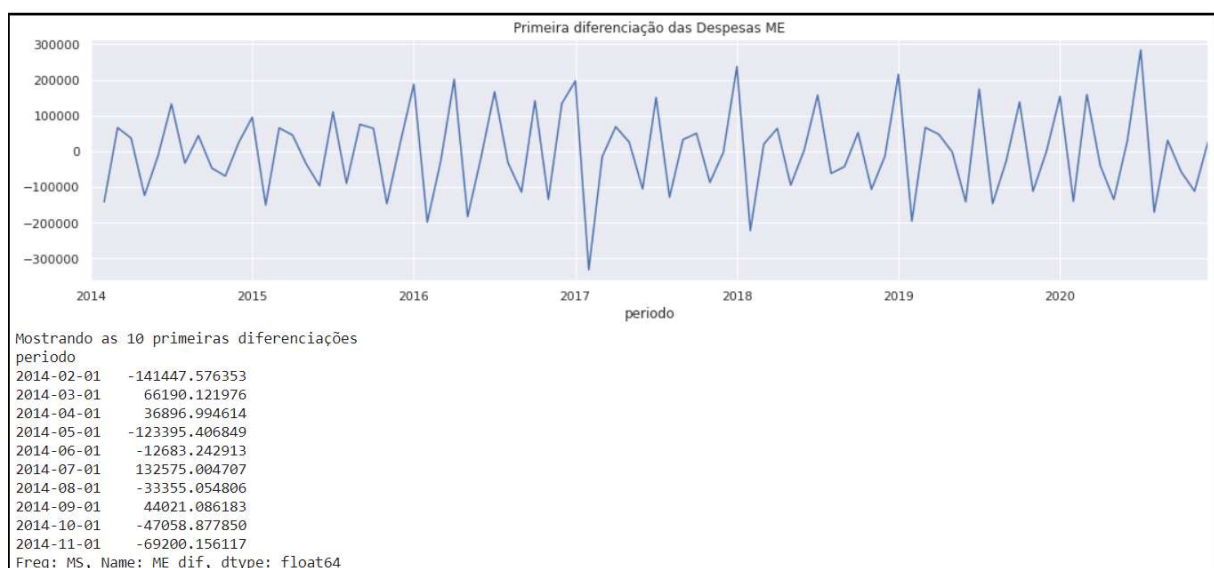
A diferenciação é utilizada para remover os sinais de tendências e reduzir a variância. É simplesmente a diferença do valor do período T com o valor do período anterior T-1.

Normalmente só é necessário uma diferenciação para transformar uma série em estacionária, mas caso seja necessário, pode-se aplicar uma segunda diferenciação (neste caso, a diferenciação será sobre a primeira diferenciação, sendo raro haver casos com mais do que 2 diferenciações).

Dito isso, aplicamos a diferenciação para os três Ministérios estudados, ou seja, mesmo o Ministério da Economia apresentando valores estacionários quando ajustado à inflação, optamos por utilizar esse método para os três órgãos para uma melhor comparação. Inclusive, mais adiante utilizaremos os dois métodos de transformação como fontes para os modelos preditivos que serão utilizados neste estudo.

a. Ministério da Economia

O gráfico do Ministério da Economia com os valores ajustados por diferenciação ficou da seguinte forma:



Realizando o teste de Dickey-Fuller para o método da diferenciação, temos o seguinte resultado:

```

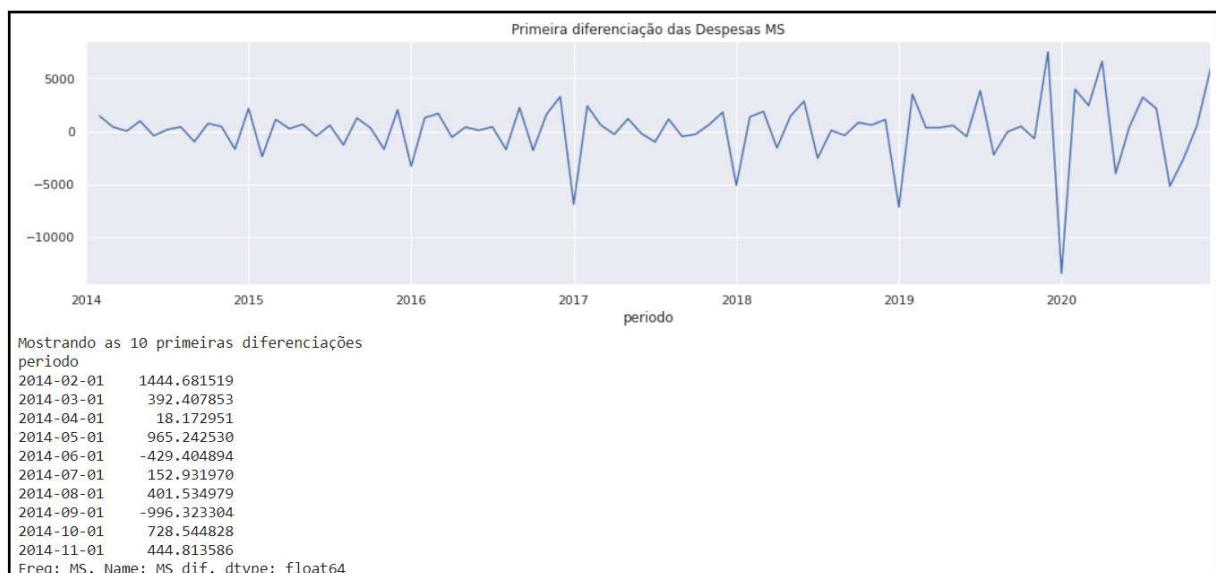
Teste de Dickey-Fuller Aumentado:
ADF teste estatístico    -4.582948
p-value                  0.000139
# lags used              10.000000
# observações           72.000000
valor crítico (1%)      -3.524624
valor crítico (5%)      -2.902607
valor crítico (10%)     -2.588679
Fortes evidências contra a hipótese nula
Rejeita a hipótese nula
É estacionário

```

Assim como o primeiro método de transformação (pela inflação), o segundo método também apresentou resultados com valores de P abaixo dos 5%.

b. Ministério da Saúde

O gráfico do Ministério da Saúde com os valores ajustados por diferenciação ficou da seguinte forma:



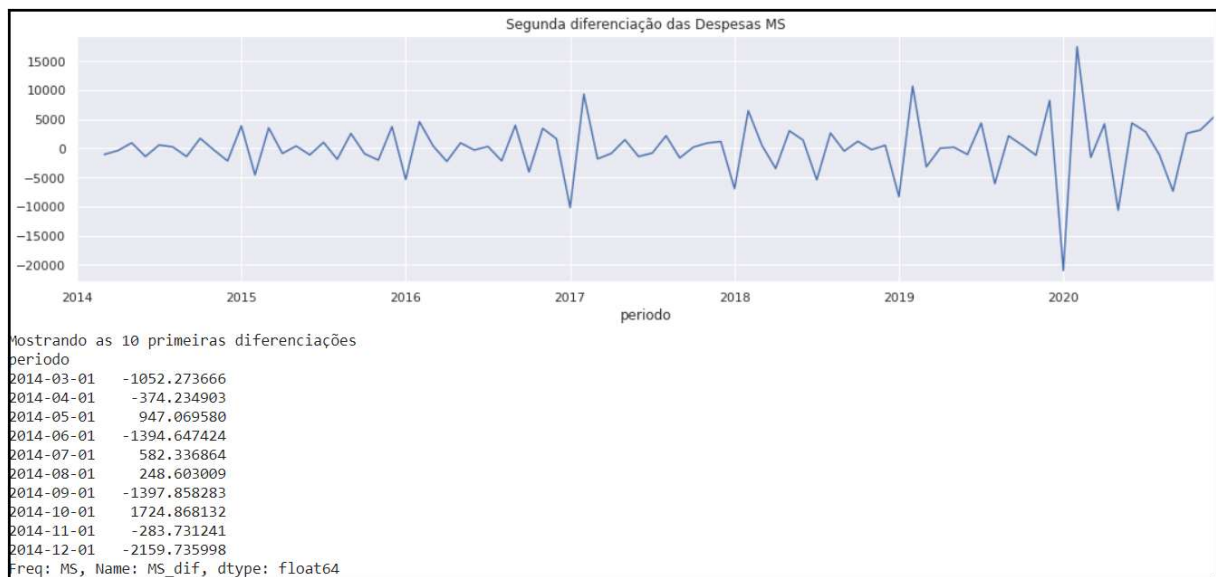
Realizando o teste de Dickey-Fuller para o método da diferenciação, temos o seguinte resultado:

```

Teste de Dickey-Fuller Aumentado:
ADF teste estatístico    -2.405226
p-value                  0.140267
# lags used              12.000000
# observações           70.000000
valor crítico (1%)      -3.527426
valor crítico (5%)      -2.903811
valor crítico (10%)     -2.589320
Fracas evidências contra a hipótese nula
Falha ao rejeitar a hipótese nula
É não-estacionária

```

No caso do Ministério da Saúde, uma única diferenciação não foi suficiente para transformar os dados em estacionários. Assim, faz-se necessário a utilização de uma segunda diferenciação. Após a aplicação do comando, realizamos novo teste Dickey-Fuller que apresentou os seguintes resultados:



```

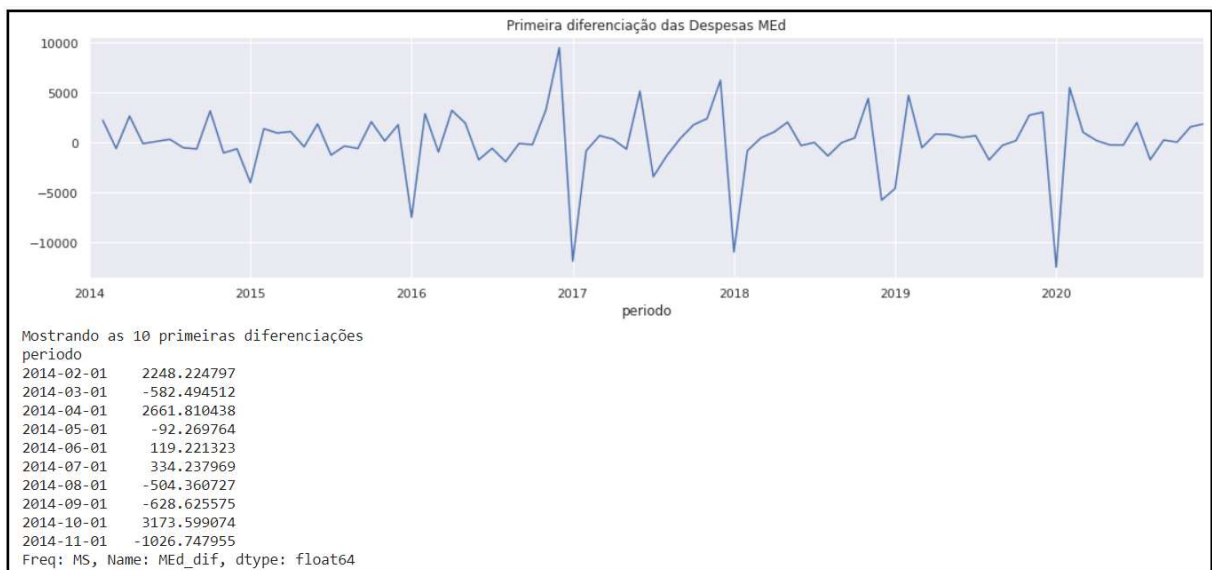
Teste de Dickey-Fuller Aumentado:
ADF teste estatístico    -8.246901e+00
p-value                  5.522832e-13
# lags used              1.200000e+01
# observações           6.900000e+01
valor crítico (1%)      -3.528890e+00
valor crítico (5%)      -2.904440e+00
valor crítico (10%)     -2.589656e+00
Fortes evidências contra a hipótese nula
Rejeita a hipótese nula
É estacionário

```

Agora sim, os dados foram transformados em estacionários a partir da aplicação da segunda diferenciação sobre os valores originais.

c. Ministério da Educação

O gráfico do Ministério da Educação com os valores ajustados por diferenciação ficou da seguinte forma:



Realizando o teste de Dickey-Fuller para o método da diferenciação, temos o seguinte resultado:

```

Teste de Dickey-Fuller Aumentado:
ADF teste estatístico    -8.013862e+00
p-value                  2.165045e-12
# lags used              1.000000e+01
# observações            7.200000e+01
valor crítico (1%)       -3.524624e+00
valor crítico (5%)       -2.902607e+00
valor crítico (10%)      -2.588679e+00
Fortes evidências contra a hipótese nula
Rejeita a hipótese nula
É estacionário
  
```

A partir deste momento, e como já previamente mencionado, temos todas as bases com dados estacionários e não estacionários. Desta forma, seguiremos neste estudo com aplicações de modelos de previsão de Série Temporal utilizando ora dados estacionários (**AUTO-ARIMA**), ora não estacionários (**Holt-Winter e Facebook Prophet**).

5. Criação de Modelos de Machine Learning

Conforme já anunciamos anteriormente, a modelagem dar-se-á utilizando três bibliotecas distintas:

- AUTO-ARIMA;
- Suavização Exponencial de Holt-Winters; e
- Facebook Prophet.

Os três métodos trabalham com séries que apresentam tendência e sazonalidade. Optamos por utilizar a biblioteca AUTO-ARIMA para determinar os melhores parâmetros ao se trabalhar com dados transformados em estacionários, seja pela correção da inflação, seja pela diferenciação. Os modelos de Holt-Winters e Facebook Prophet trabalharão com os dados originais, ou seja, não estacionários.

A biblioteca AUTO-ARIMA tem como ferramenta o pacote pmdarima. Nele, é utilizado para aplicar uma espécie de força bruta em um modelo ARIMA, testando diversos parâmetros a fim de encontrar o melhor cenário. Um modelo ARIMA significa Média Móvel integrada Autorregressiva, ou seja, este tipo de modelo trabalha com a Média Móvel diante de dados históricos para prever o futuro. Diferente de alguns outros modelos, o ARIMA suporta trabalhar tanto com dados sazonais (SARIMA) como dados não sazonais. Tudo isso é especificado através de uma série de parâmetros que basicamente categoriza seus dados em tendência, ruído e sazonalidade.

O método elaborado por Holt (1957) e Winters (1960) tem como pressuposto fundamental lidar com sazonalidades. Ele possui três equações básicas: uma para ajuste de nível, outra para ajustar o crescimento e outra para sazonalidade.

Já o método Facebook Prophet foi criado em 2017 pela Facebook Research como ferramenta *open source*. É um modelo que captura e interpreta a sazonalidade e a tendência em dados de séries temporais, podendo ser, por exemplo, anualmente, semanalmente e diariamente, além de séries personalizadas.

Espera-se, ao final deste estudo, poder analisar qual melhor modelo preditivo se encaixou com os dados tanto na forma estacionária, quanto na forma não estacionária. Para isso, os dados serão divididos em bases de treino e de teste da seguinte forma:

- **Treino:** 01/2014 até 12/2018;
- **Teste:** 01/2019 até 12/2020.

A divisão procura chegar mais próximo possível de 80% de dados para treino, mas de forma que terminasse em dezembro de algum ano.

```
train = bkp_gastos.loc[:'2018-12-01']
test = bkp_gastos.loc['2019-01-01':]
```

5.1 AUTO-ARIMA

Utilizamos a ferramenta `pmdarima.auto_arima` para indicar o melhor modelo para cada base estudada.

```
!pip install pmdarima
from pmdarima import auto_arima
```

5.1.1 AUTO-ARIMA para o Ministério da Economia com valores ajustados pela inflação

Com a biblioteca AUTO-ARIMA, foi possível estabelecer o seguinte modelo indicado para os dados apresentados:

Statespace Model Results						
Dep. Variable:	y	No. Observations: 84				
Model:	SARIMAX(3, 0, 5)x(2, 0, 0, 12)			Log Likelihood	-1056.220	
Date:	Fri, 09 Apr 2021			AIC	2136.440	
Time:	18:58:51			BIC	2165.610	
Sample:	0			HQIC	2148.166	
- 84						
Covariance Type: opg						
	coef	std err	z	P> z	[0.025	0.975]
intercept	2.654e+05	1.25e-05	2.12e+10	0.000	2.65e+05	2.65e+05
ar.L1	-1.0124	0.509	-1.989	0.047	-2.010	-0.015
ar.L2	-1.0102	0.501	-2.015	0.044	-1.993	-0.027
ar.L3	-0.0146	0.504	-0.029	0.977	-1.003	0.974
ma.L1	1.1737	0.557	2.109	0.035	0.083	2.265
ma.L2	1.1366	0.694	1.639	0.101	-0.223	2.496
ma.L3	0.0286	0.724	0.040	0.968	-1.391	1.448
ma.L4	-0.1499	0.305	-0.492	0.623	-0.747	0.447
ma.L5	-0.1484	0.198	-0.748	0.455	-0.537	0.240
ar.S.L12	0.1639	0.178	0.922	0.356	-0.184	0.512
ar.S.L24	0.3318	0.175	1.901	0.057	-0.010	0.674
sigma2	5.44e+09	1.55e-10	3.51e+19	0.000	5.44e+09	5.44e+09
Ljung-Box (Q):	21.86	Jarque-Bera (JB): 7.89				
Prob(Q):	0.99	Prob(JB): 0.02				
Heteroskedasticity (H):	0.95	Skew: 0.71				
Prob(H) (two-sided):	0.89	Kurtosis: 3.48				

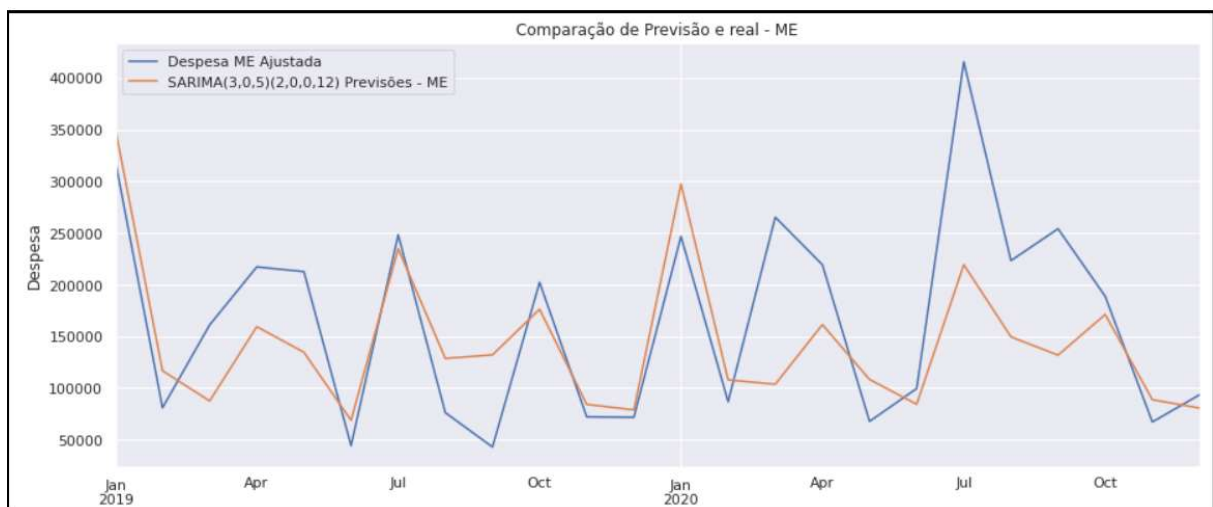
Dessa forma, ajustamos o modelo e previmos a base de teste:

Ministério da Economia com despesas ajustadas pela Inflação

```
[85] model_sarima_ME = SARIMAX(train['Despesa ME Ajustada'].dropna(),order=(3,0,5),seasonal_order=(2,0,0,12))
      results_sarima_ME = model_sarima_ME.fit()
      results_sarima_ME.summary()
```

```
# Obtendo a previsão
inicio = len(train)
fim = len(train)+len(test)-1
predictions_sarima_ME = results_sarima_ME.predict(start=inicio, end=fim, dynamic=False, typ='levels').rename('SARIMA(3,0,5)(2,0,0,12) Previsões - ME')
```

Após determinarmos o melhor modelo para a base, plotamos a previsão comparando com a base de teste:



5.1.2 AUTO-ARIMA para o Ministério da Saúde com valores ajustados por diferenciação

Com a biblioteca AUTO-ARIMA, foi possível estabelecer o seguinte modelo indicado para os dados apresentados:

Statespace Model Results						
Dep. Variable:	y	No. Observations: 82				
Model:	SARIMAX(1, 0, 1)x(1, 0, 0, 12)	Log Likelihood -747.801				
Date:	Fri, 09 Apr 2021	AIC 1505.602				
Time:	18:59:14	BIC 1517.635				
Sample:	0	HQIC 1510.433				
- 82						
Covariance Type: opg						
	coef	std err	z	P> z	[0.025	0.975]
intercept	-0.1219	6.588	-0.019	0.985	-13.034	12.790
ar.L1	-0.4673	0.078	-6.029	0.000	-0.619	-0.315
ma.L1	-0.9787	0.091	-10.802	0.000	-1.156	-0.801
ar.S.L12	0.6792	0.110	6.152	0.000	0.463	0.896
sigma2	4.198e+06	5.83e+05	7.205	0.000	3.06e+06	5.34e+06
Ljung-Box (Q): 70.43 Jarque-Bera (JB): 25.13						
Prob(Q): 0.00 Prob(JB): 0.00						
Heteroskedasticity (H): 5.68 Skew: 0.50						
Prob(H) (two-sided): 0.00 Kurtosis: 5.52						

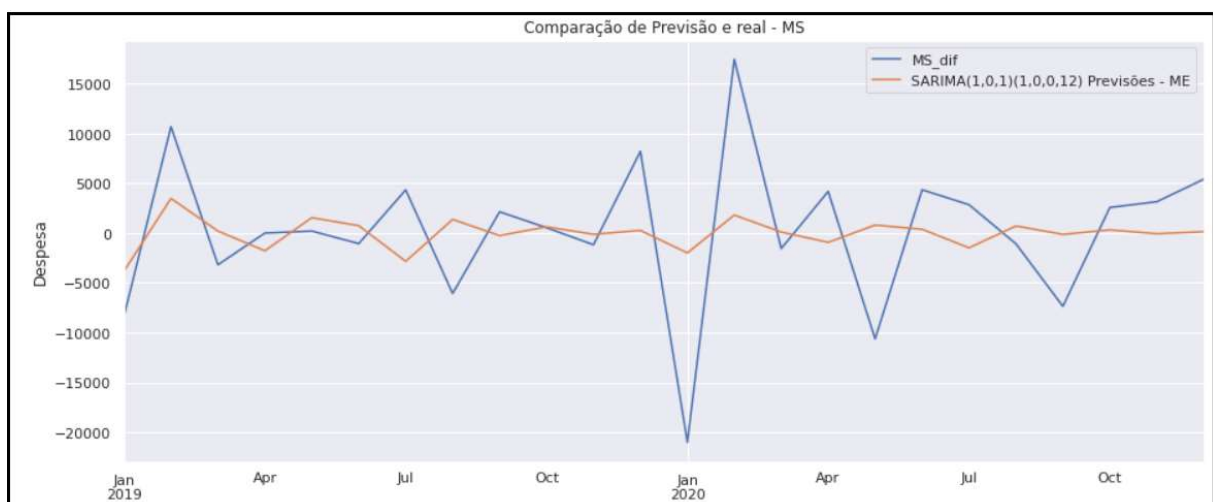
Dessa forma, ajustamos o modelo e previmos a base de teste:

Ministério da Saúde com despesas ajustadas por Diferenciação

```
[89] model_sarima_MS = SARIMAX(train.MS_dif.dropna(),order=(1,0,1),seasonal_order=(1,0,0,12))
      results_sarima_MS = model_sarima_MS.fit(dispatch = -1)
      results_sarima_MS.summary()
```

```
[90] # Obtendo a previsão
      inicio = len(train)-2
      fim = len(train)+len(test)-3
      predictions_sarima_MS = results_sarima_MS.predict(start=inicio, end=fim, dynamic=False, typ='levels').rename('SARIMA(1,0,1)(1,0,0,12) Previsões - ME')
```

Após determinarmos o melhor modelo para a base, plotamos a previsão comparando com a base de teste:



5.1.3 AUTO-ARIMA para o Ministério da Educação com valores ajustados por diferenciação

Com a biblioteca AUTO-ARIMA, foi possível estabelecer o seguinte modelo indicado para os dados apresentados:

Statespace Model Results						
Dep. Variable:	y				No. Observations:	83
Model:	SARIMAX(5, 0, 1)x(1, 0, 1, 12)				Log Likelihood	-754.147
Date:	Fri, 09 Apr 2021				AIC	1528.295
Time:	19:00:49				BIC	1552.483
Sample:	0				HQIC	1538.012
	- 83					
Covariance Type: opg						
	coef	std err	z	P> z	[0.025	0.975]
intercept	2.0270	16.322	0.124	0.901	-29.963	34.017
ar.L1	0.0358	0.256	0.140	0.889	-0.465	0.537
ar.L2	-0.2453	0.334	-0.735	0.463	-0.900	0.409
ar.L3	-0.1214	0.425	-0.286	0.775	-0.954	0.712
ar.L4	-0.1746	0.352	-0.496	0.620	-0.864	0.515
ar.L5	-0.0830	0.260	-0.320	0.749	-0.592	0.426
ma.L1	-0.8649	0.221	-3.916	0.000	-1.298	-0.432
ar.S.L12	0.9022	0.182	4.969	0.000	0.546	1.258
ma.S.L12	-0.6161	0.333	-1.847	0.065	-1.270	0.038
sigma2	6.266e+06	1.53e-05	4.09e+11	0.000	6.27e+06	6.27e+06
Ljung-Box (Q):	17.77	Jarque-Bera (JB):	138.74			
Prob(Q):	1.00	Prob(JB):	0.00			
Heteroskedasticity (H):	2.32	Skew:	0.20			
Prob(H) (two-sided):	0.03	Kurtosis:	9.32			

Dessa forma, ajustamos o modelo e previmos a base de teste:

Ministério da Educação com despesas ajustadas por Diferenciação

```
[93] model_sarima_MEd = SARIMAX(train.MEd_dif.dropna(), order=(5,0,1), seasonal_order=(1,0,1,12))
      results_sarima_MEd = model_sarima_MEd.fit(dispatch = -1)
      results_sarima_MEd.summary()
```

```
[94] # Obtendo a previsão
      inicio = len(train)-1
      fim = len(train)+len(test)-2
      predictions_sarima_MEd = results_sarima_MEd.predict(start=inicio, end=fim, dynamic=False, typ='levels').rename('SARIMA(5,0,1)(1,0,1,12) Previsões - MEd')
```

Após determinarmos o melhor modelo para a base, plotamos a previsão comparando com a base de teste:



Os modelos apresentaram resultados satisfatórios na comparação com as bases de testes. É sempre bom lembrar que o ano de 2020 foi marcado pela pandemia do Coronavírus. Este é o motivo de termos, em alguns gráficos, outliers que deixam a previsão um pouco aquém do realizado na base de testes.

5.2 Suavização Exponencial Holt-Winters

A chamada à biblioteca deste modelo é dada pelo seguinte comando:

```
from statsmodels.tsa.holtwinters import ExponentialSmoothing
```

5.2.1 Holt-Winters para a base de despesas do Ministério da Economia com valores ajustados pela inflação

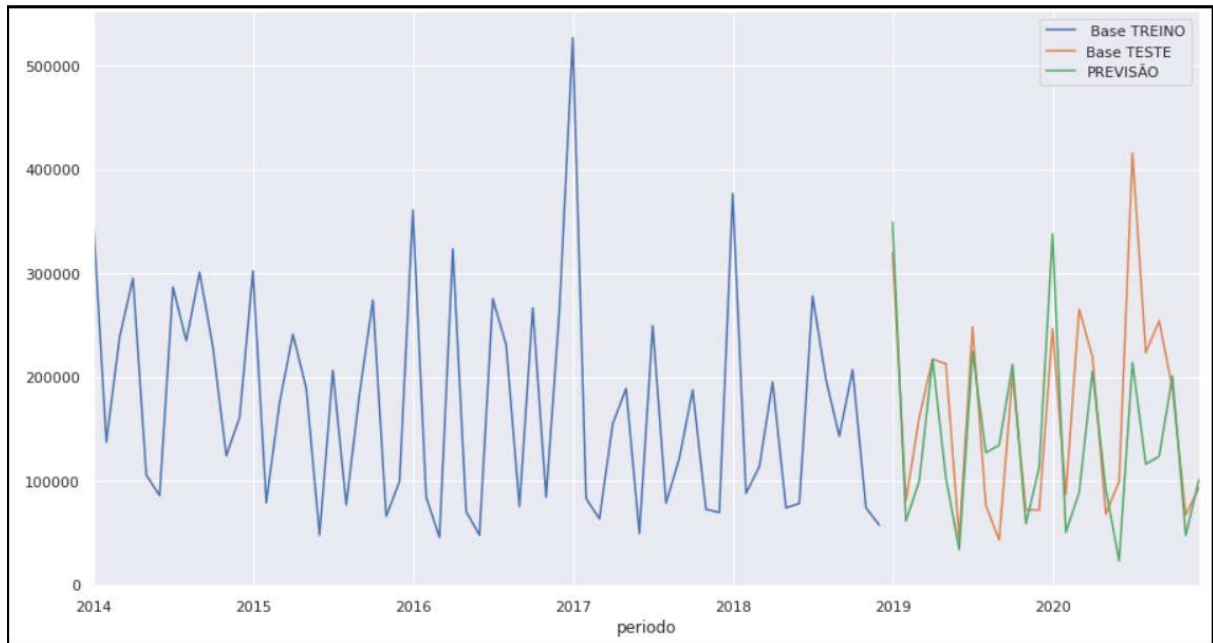
A intenção ao utilizar dados estacionários para o modelo preditivo de Suavização Exponencial de Holt-Winters é facilitar a comparação sobre qual base teve melhor desempenho. Por isso, trabalharemos neste tópico o modelo sobre dados estacionários e mais adiante sobre dados não estacionários. A base do Ministério da Economia possui sazonalidade bem definida, por isso achamos conveniente utilizá-la como parte do estudo.

Assim, iniciamos a modelagem dos dados com o seguinte código:

```
fitted_model_ME = ExponentialSmoothing(train['Despesa ME Ajustada'],trend='add',seasonal='add',seasonal_periods=12).fit()
```

```
test_predictions_ME = fitted_model_ME.forecast(24).rename('Previsão - Holt-Winters - Despesa ME Ajustada')
test_predictions_ME
```

Comparando a base de teste e a previsão feita a partir da base de treino, o resultado é a figura a seguir:



No próximo tópico trataremos do mesmo modelo preditivo para os dados não estacionários do Ministério da Economia. Com isso, poderemos comparar em qual tipo de dados o modelo resultou melhor desempenho.

5.2.2 Holt-Winters para a base de despesas do Ministério da Economia com valores originais

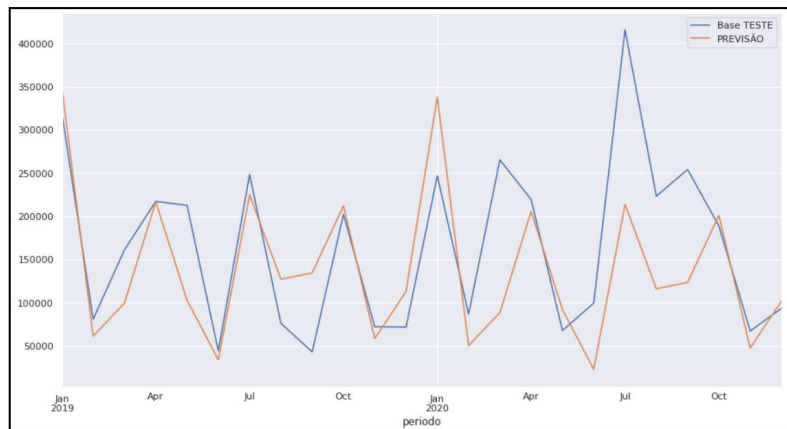
Assim como anteriormente, aplicamos o seguinte comando para obter a previsão na base de teste:

```
fitted_model_ME_orig = ExponentialSmoothing(train['25000 - Ministério da Economia'],trend='add',seasonal='add',seasonal_periods=12).fit()
test_predictions_ME_orig = fitted_model_ME_orig.forecast(24).rename('Previsão - Holt-Winters - ME origem')
test['25000 - Ministério da Economia'].plot(legend=True,label='Base TESTE',figsize=(15,8))
test_predictions_ME_orig.plot(legend=True,label='PREVISÃO ME_orig',xlim=['2019-01-01','2020-12-01']);
```

Plotando o resultado, temos o seguinte:



Acima, vemos a plotagem dos dados sobre a base não estacionária. Abaixo, a plotagem modelada no tópico 5.2.1 (dados estacionários):



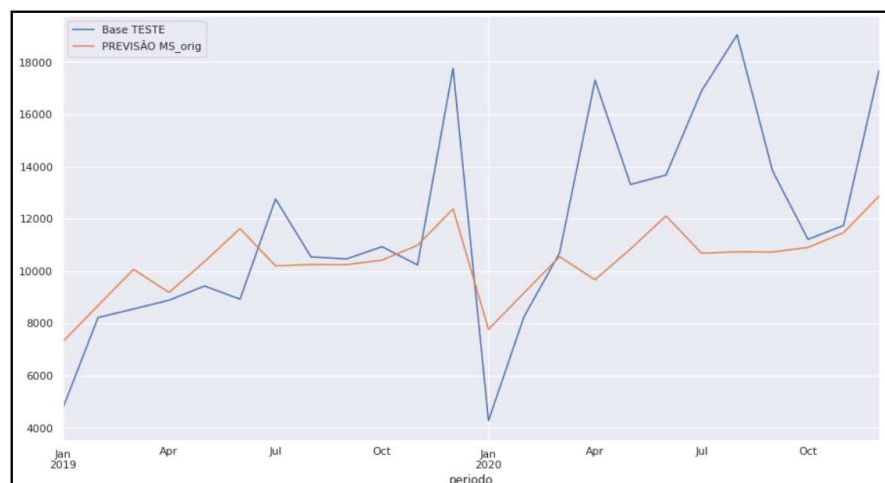
Ou seja, o modelo se apresentou de forma muito próxima com os dois tipos de dados, possuindo bom rendimento para o nosso estudo adiante. Dessa forma, analisaremos os outros dois Ministérios restantes apenas com dados não estacionários.

5.2.3 Holt-Winters para a base de despesas do Ministério da Saúde com valores originais

Aplicamos o seguinte comando para obter a previsão na base de teste:

```
fitted_model_MS_orig = ExponentialSmoothing(train['36000 - Ministério da Saúde'],trend='add',seasonal='add',seasonal_periods=12).fit()
test_predictions_MS_orig = fitted_model_MS_orig.forecast(24).rename('Previsão - Holt-Winters - MS origem')
test['36000 - Ministério da Saúde'].plot(legend=True,label='Base TESTE',figsize=(15,8))
test_predictions_MS_orig.plot(legend=True,label='PREVISÃO MS_orig',xlim=['2019-01-01','2020-12-01']);
```

Plotando o resultado, temos o seguinte:



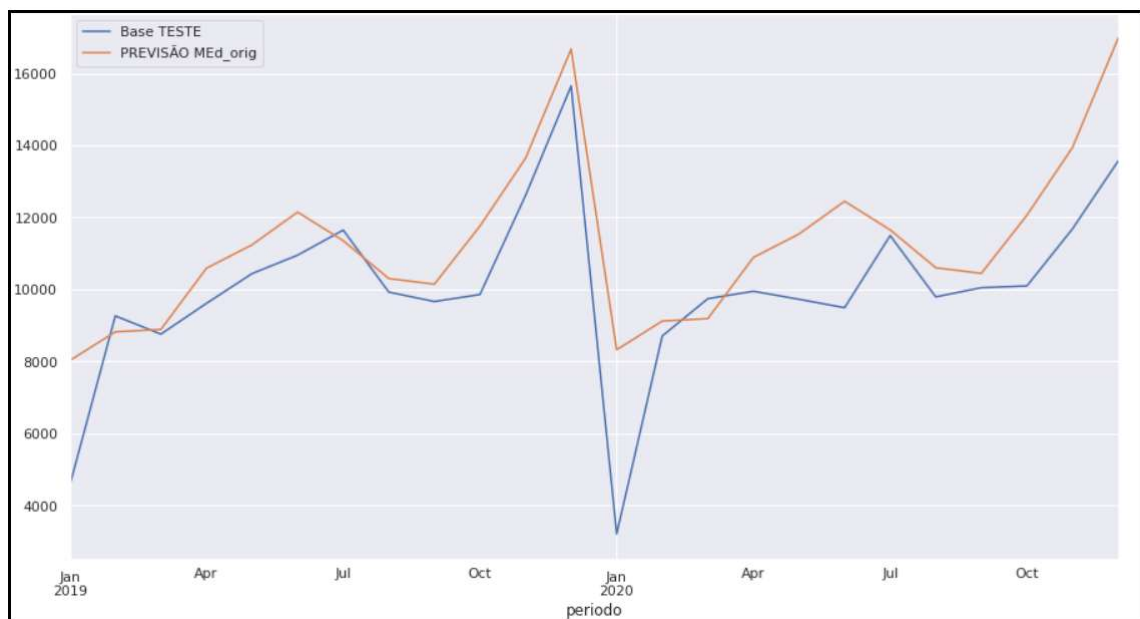
Neste caso há um pequeno descolamento da previsão com a base de teste. Como já citado anteriormente, o ano de 2020 foi marcado pelo surgimento do Coronavírus e, com isso, as políticas públicas de investimento/gasto sofreram mudanças em relação a série histórica. O Ministério da Saúde, por evidência, necessitou dispor de mais investimentos/gastos para suprir a necessidade da população.

5.2.4 Holt-Winters para a base de despesas do Ministério da Educação com valores originais

Aplicamos o seguinte comando para obter a previsão na base de teste:

```
fitted_model_MED_orig = ExponentialSmoothing(train['26000 - Ministério da Educação'],trend='add',seasonal='add',seasonal_periods=12).fit()
test_predictions_MED_orig = fitted_model_MED_orig.forecast(24).rename('Previsão - Holt-Winters - MED origem')
test['26000 - Ministério da Educação'].plot(legend=True,label='Base TESTE',figsize=(15,8))
test_predictions_MED_orig.plot(legend=True,label='PREVISÃO MED_orig',xlim=['2019-01-01','2020-12-01']);
```

Plotando o resultado, temos o seguinte:



A análise deste Ministério segue a mesma do anterior (Saúde). Neste caso, o descolamento é ligeiramente menor. Esperava-se, pela previsão, um maior investimento/gasto do governo, o que não ocorreu, muito provavelmente pela necessidade de repasse de recursos para Ministérios que atendessem a causas sociais/saúde.

5.3 Facebook Prophet

A chamada à biblioteca deste modelo é dada pelo seguinte comando:

```
from fbprophet import Prophet
from statsmodels.tools.eval_measures import rmse
```

5.3.1 Facebook Prophet para a base de despesas do Ministério da Economia com os valores originais

Utilizaremos o Facebook Prophet apenas para a base de dados não estacionário, ou seja, com os valores originais.

Dito isso, ainda é preciso esclarecer que, ao se trabalhar com o modelo preditivo desenvolvido pelo Facebook Research, é preciso reduzir o dataset de treinamento em apenas duas colunas, com o objetivo de atender as especificações da biblioteca. Passamos a informar este comando e o cabeçalho do mesmo apenas para o Ministério da Economia, contudo, é possível identificar os mesmos comandos para os demais Órgãos consultando o notebook de análise (link ao final deste Trabalho).

```
df_prophet_ME = pd.DataFrame(bkp_gastos_prophet, columns=['PERIODO', '25000 - Ministério da Economia'])
```

PERIODO 25000 - Ministério da Economia		
periodo		
2014-01-01	2014-01-01	230707.875070
2014-02-01	2014-02-01	89260.298717
2014-03-01	2014-03-01	155450.420693
2014-04-01	2014-04-01	192347.415307
2014-05-01	2014-05-01	68952.008458

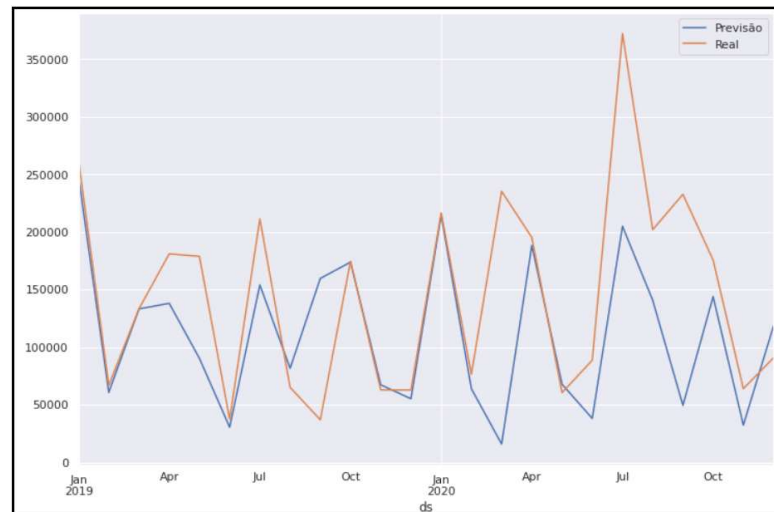
```
df_prophet_ME.columns = ['ds', 'y']
```

A coluna “PERIODO” se transformou em “ds” e a coluna “25000 – Ministério da Economia” em “y”.

O código do modelo ficou da seguinte forma:

```
m = Prophet()
m.fit(train_prophet_ME)
future = m.make_future_dataframe(periods=24,freq='MS')
forecast = m.predict(future)
```

A previsão na base de teste é a que segue:



É possível observar um excelente desempenho do modelo preditivo para os dados não estacionários.

5.3.2 Facebook Prophet para a base de despesas do Ministério da Saúde com valores originais

Após reduzir a base para atender às especificações da biblioteca, passamos ao seguinte comando para obter a previsão na base de teste:

```
m = Prophet()
m.fit(train_prophet_MS)
future = m.make_future_dataframe(periods=24,freq='MS')
forecast = m.predict(future)
fig = m.plot_components(forecast)
```

A previsão na base de teste é a seguinte:



Da mesma forma como nos outros modelos preditivos, o Facebook Prophet sofreu um descolamento na previsão na base de teste para o Ministério da Saúde. Cada vez mais, fica evidente o outlier em decorrência da pandemia do Coronavírus, o que levou este Órgão a dispor de mais recursos durante o ano de 2020.

5.3.3 Facebook Prophet para a base de despesas do Ministério da Educação com valores originais

Após reduzir a base para atender às especificações da biblioteca, passamos ao seguinte comando para obter a previsão na base de teste:

```
m = Prophet()
m.fit(train_prophet_MEd)
future = m.make_future_dataframe(periods=24, freq='MS')
forecast = m.predict(future)
fig = m.plot_components(forecast)
```

A previsão na base de teste é a seguinte:



Assim como analisamos este Ministério com os outros modelos preditivos anteriores, é possível observar pelo Facebook Prophet a mesma conclusão dos demais: há um pequeno descolamento para o período de 2020, com o realizado menor que o previsto pelo modelo.

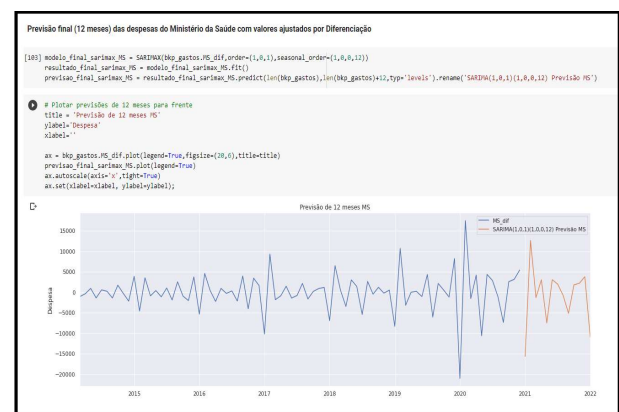
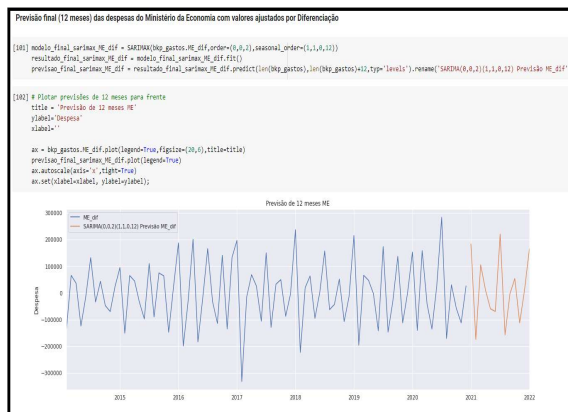
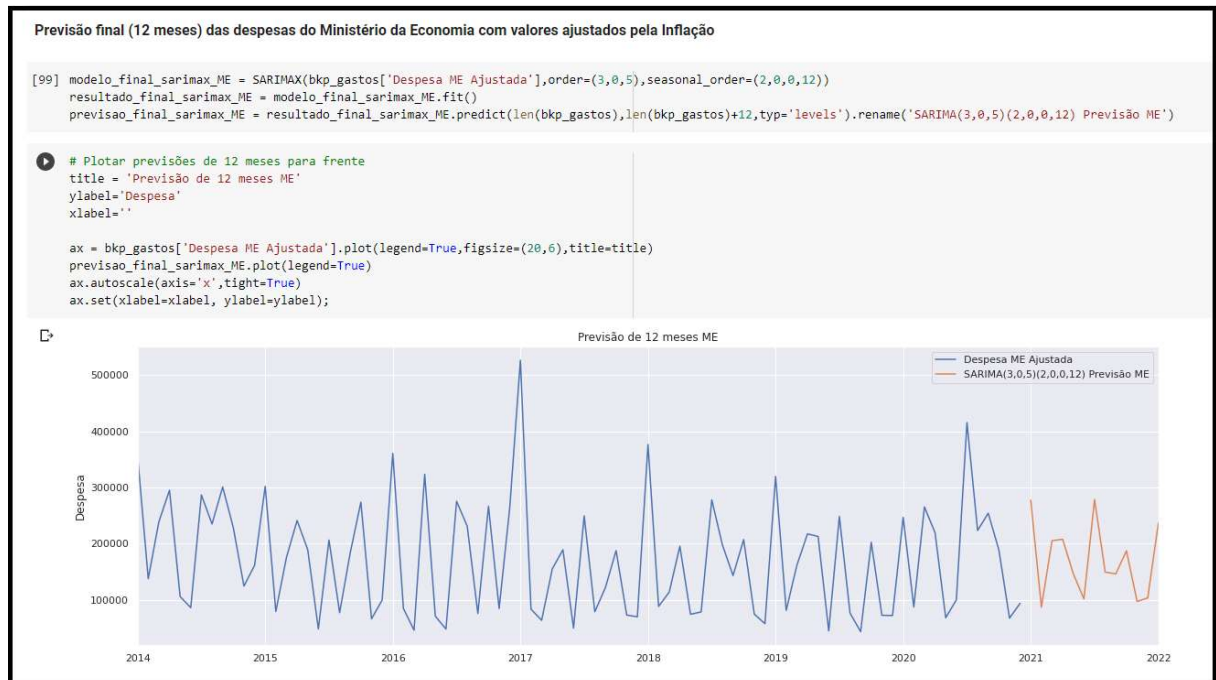
6. Apresentação dos Resultados

A partir desta etapa, apresentaremos os resultados das previsões futuras. Como informado previamente, serão demonstrados comandos e gráficos com a previsão de 12 meses após nossa base de teste, ou seja, janeiro/2021 a dezembro/2021.

Apresentaremos as previsões para os três modelos preditivos propostos (AUTO-ARIMA, Holt-Winters e Facebook Prophet).

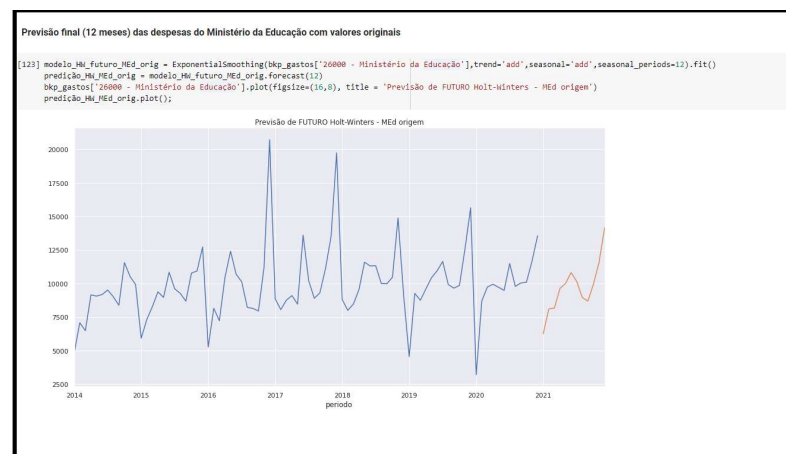
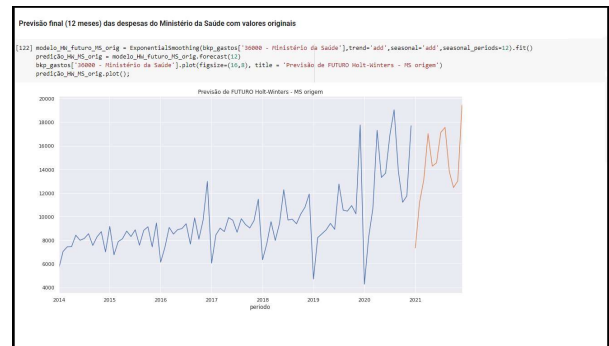
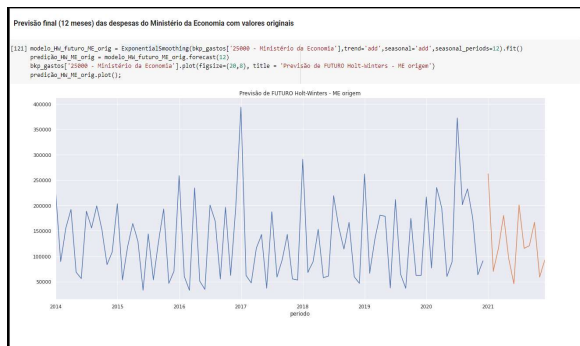
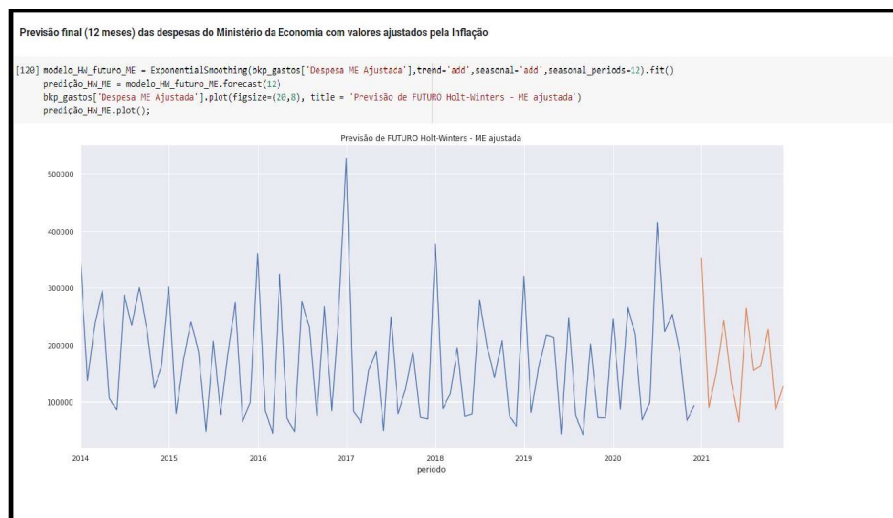
6.1 AUTO-ARIMA – Previsão 12 meses futuros

Nesta seção optamos por apresentar o comando aplicado e o gráfico plotado do Ministério da Economia com os valores ajustados pela inflação, contudo, para facilitar a visualização dos demais Ministérios estudados, serão dispostos todos os resultados finais nesta mesma seção.



6.2 Suavização Exponencial Holt-Winters – Previsão 12 meses futuros

Da mesma forma como fizemos no tópico anterior, passaremos a dispor do comando e gráfico do Ministério da Economia com os valores ajustados pela inflação, seguidos do mesmo órgão com os valores originais e demais Ministérios (Saúde e Educação, sucessivamente):



6.3 Facebook Prophet – Previsão 12 meses futuros

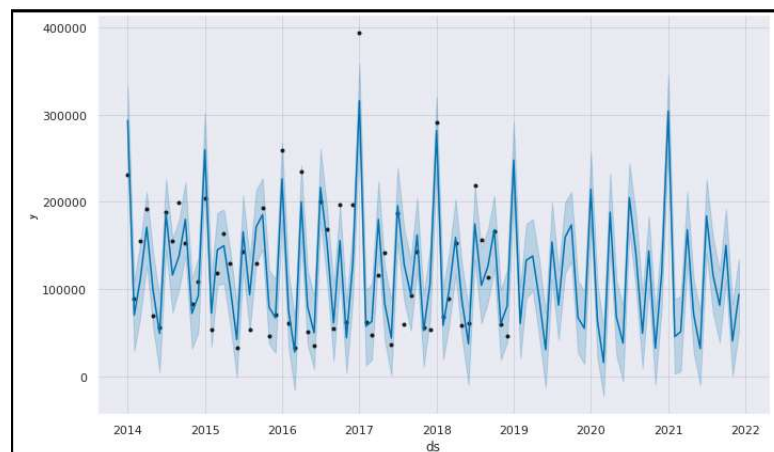
Para este tópico, passamos a demonstrar de forma mais separada os resultados para facilitar a visualização.

6.3.1 Facebook Prophet – Ministério da Economia com os valores originais (12 meses futuros)

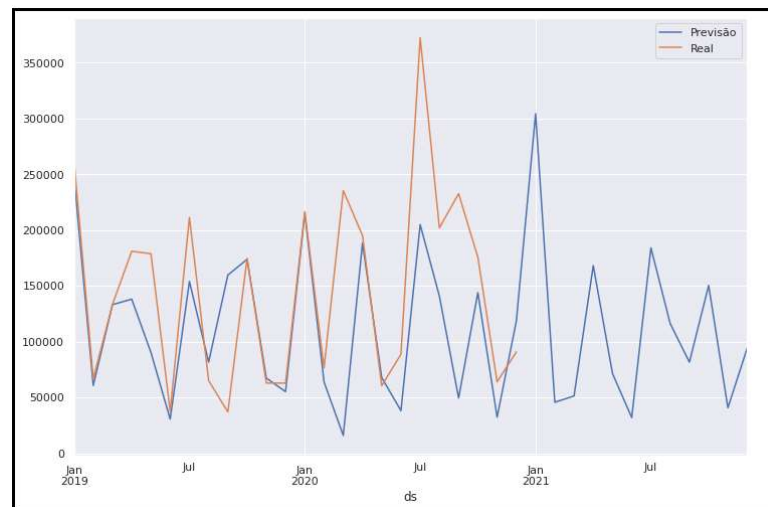
Para modelar o futuro, consideramos o seguinte comando:

```
m = Prophet()
m.fit(train_prophet_ME)
future = m.make_future_dataframe(periods=36, freq='MS')
forecast = m.predict(future)
```

O gráfico do comando forecast para este Ministério segue adiante:



Por fim, a previsão de futuro é demonstrada na figura abaixo:

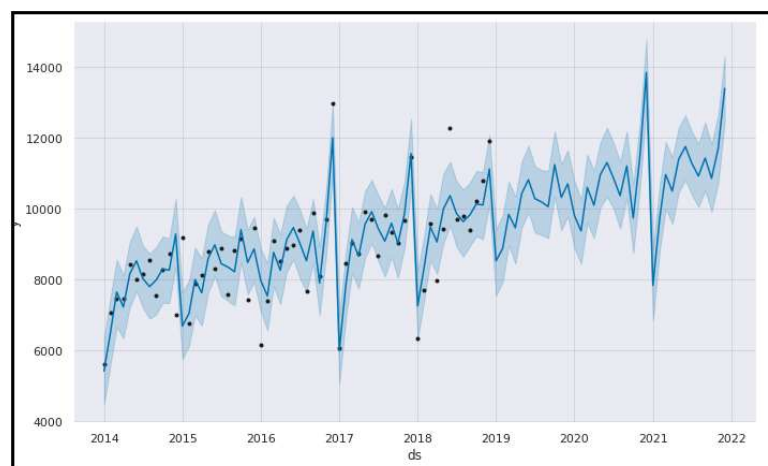


6.3.2 Facebook Prophet – Ministério da Saúde com os valores originais (12 meses futuros)

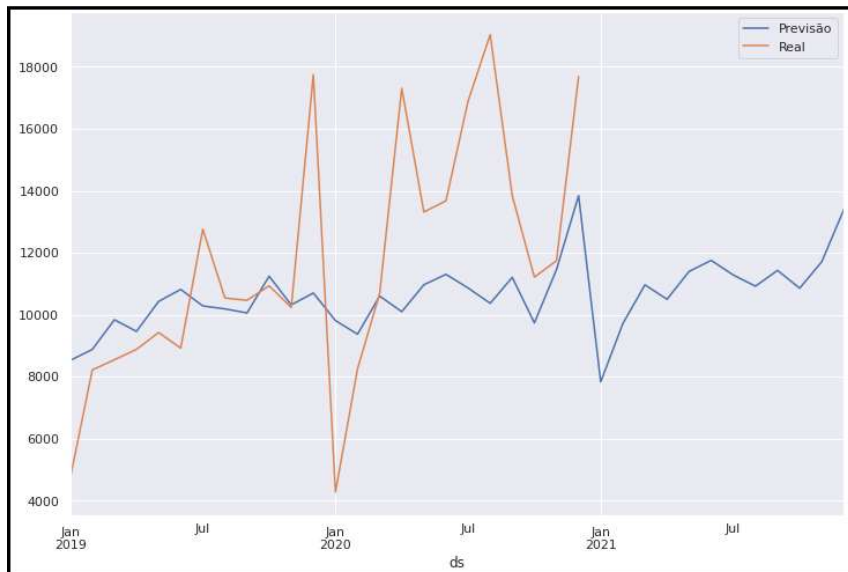
Para modelar o futuro, consideramos o seguinte comando:

```
m = Prophet()
m.fit(train_prophet_MS)
future = m.make_future_dataframe(periods=36, freq='MS')
forecast = m.predict(future)
m.plot(forecast);
```

O gráfico do comando forecast para este Ministério segue adiante:



Por fim, a previsão de futuro é demonstrada na figura abaixo:

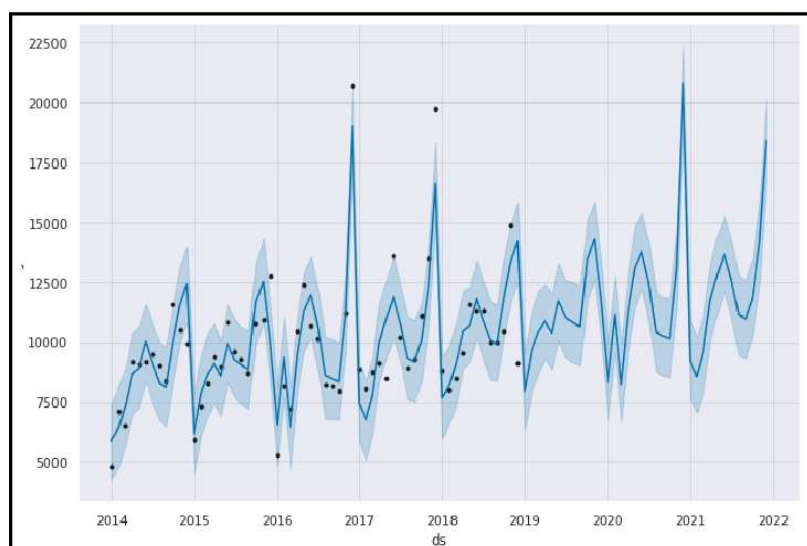


6.3.3 Facebook Prophet – Ministério da Educação com os valores originais (12 meses futuros)

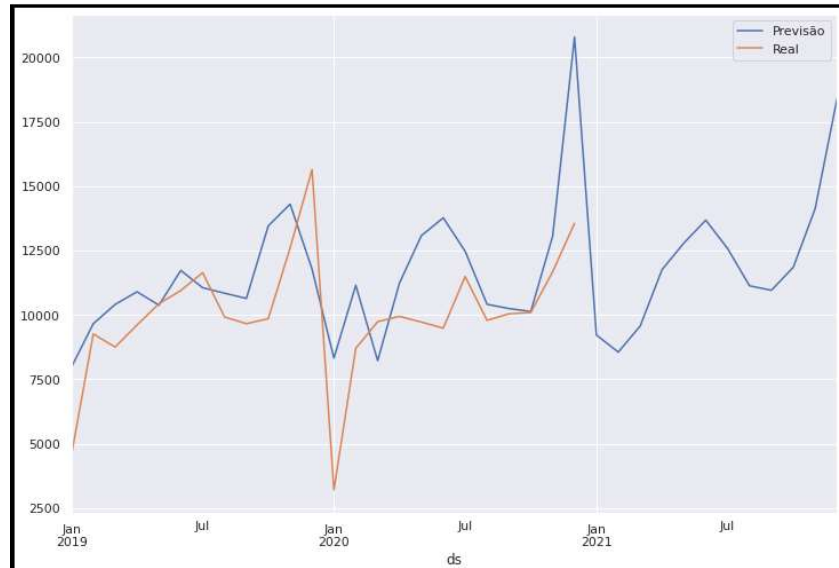
Para modelar o futuro, consideramos o seguinte comando:

```
m = Prophet()
m.fit(train_prophet_MEd)
future = m.make_future_dataframe(periods=36, freq='MS')
forecast = m.predict(future)
m.plot(forecast);
```

O gráfico do comando forecast para este Ministério segue adiante:



Por fim, a previsão de futuro é demonstrada na figura abaixo:



6.4 Conclusões

Interessante notar como todos os modelos estudados se comportaram de maneira adequada na previsão da base de testes. Por consequência, as previsões futuras também apresentam boa correlação.

A pandemia do Coronavírus mudou significativamente a forma como os gastos de alguns Ministérios são reproduzidos. Isso faz bastante sentido, já que o Governo Federal dispõe de seus “braços” ministeriais para aplicar políticas públicas e direcionar o investimento/gasto nas atividades emergenciais do país. Foi possível verificar que o Ministério da Saúde sofreu um aumento significativo dos gastos/despesas em 2020, em contraponto ao Ministério da Educação, que teve seus investimentos reduzidos no mesmo período. Ademais, todos os modelos preditivos se adaptaram bem a esses dois órgãos.

No Ministério da Economia é que foi possível verificar a melhor compatibilidade entre previsão e real. Como já falado, por ser o órgão superior responsável pela aplicação e direcionamento dos recursos investidos/gastos, talvez este tenha sido o motivo de que seus gráficos apresentassem maior linearidade. Dessa forma, mesmo

com o aumento expressivo nos gastos para o período de 2020, todos os modelos preditivos conseguiram reagir de maneira satisfatória ao proposto pelo projeto.

7. Links

Link para o vídeo: <https://youtu.be/gmArl-uAQWQ>

Link para o repositório: https://github.com/felipealbinorodrigues/TCC_PUC

Neste repositório temos os seguintes arquivos:

- Bases de dados originais e tratadas (Despesas Públicas e IPCA);
- Script do Jupyter Notebook;
- PDF do Jupyter Notebook;
- Vídeo de apresentação compactado; e
- Trabalho de Conclusão do Curso em formato PDF.

REFERÊNCIAS

FACEBOOK Prophet. Disponível em: <<https://opensource.facebook.com/>>. Acesso em: 10 abr. 2021

PORTILLA, **Jose Marcial**. *Using Python and Auto ARIMA to Forecast Seasonal Time Series*. Disponível em: <<https://medium.com/@josemarcialportilla/usingpython-and-auto-arima-to-forecast-seasonal-time-series-90877adff03c>>. Acesso em:

05 mar. 2021

Brownlee, J. (2019). *Deep Learning for Time Series Forecasting*. Machine Learning Mastery.