Relatório inicial de coleta de dados da 4º iteração com a Capital São Luís

INTRODUÇÃO

Trata o presente relatório do resumo da tarefa de “Coleta inicial dos dados” da Metodologia CRISP-DM da Previsão das Receitas Públicas. Todos os dados obtidos atendem aos ditames da Lei Nº 12.527, de 18 de novembro de 2011, Lei de acesso à informação no setor público brasileiro, já que elas estão disponíveis nos sites de Transparência de Entidades do setor público (BRASIL, 2011).

## JUSTIFICATIVA

Todo o trabalho de coleta dos dados foi elaborado na ferramenta *Power BI Desktop*, alguns dados, como o exemplo da “População” que foi necessário um único relatório[[1]](#footnote-1) para fazer extração, transformação e carregamento dos dados, as demais foram no relatório geral, e, na prática, observou-se que quando os relatórios são divididos e unificados em outro relatório seu tempo de processamento reduz, lembrando que a versão da ferramenta é gratuita, por isso algumas limitações. Quanto ao artefato desta atividade, ele fica registrado na ferramenta, podendo incluir metadados, e ficando os títulos no painel para visualização, assim o relatório inicial de coleta de dados, apenas da Capital São Luís, está resumido.

Figura - Tarefas e artefatos da Fase de Entendimento dos dados do CRISP-DM

Interface gráfica do usuário, Texto

Descrição gerada automaticamente com confiança média

Fonte: Adaptado de (CHAPMAN, CLINTON, *et al.*, 2000, p. 20) (tradução nossa).

# PROPOSTA DE TRABALHO

Mediante os objetivos outrora traçados, a proposta deste trabalho de regressão, que visa a comparação entre modelos estatísticos e de técnicas de AM como uma alternativa de apoiar a implantação das regras da SOF para a LRF na previsão das receitas públicas, utilizar-se-á o CRISP-DM para buscar as melhores previsões e variáveis explicativas com base nas séries históricas das oito capitais. Portanto, o capítulo teve com escopo demonstrar como as atividades ocorrerão, desde a compreensão do negócio e dos dados, detalhando a escolha das variáveis exógenas e construção dos *datasets*, a fase de preparação dos dados, com os procedimentos adotados para ajustar as bases para a modelagem avaliação e implementação foram resumidas nas seis primeiras subseções. E a sétima foram discutidos os trabalhos correlatos.

É importante registrar que toda a proposta foi elabora baseando-se na primeira iteração, utilizando os dados da Capital do Maranhão, São Luís, pretendendo aplicar o mesmo processo com foco na análise das questões propostas nos objetivos para a previsão das receitas públicas das 8 capitais que aceitaram participar deste experimento.

## Segunda Fase de Entendimento dos dados

Na primeira iteração realizada, com os dados do Município de São Luís, foi obtida uma aplicação prática das quatro tarefas desta fase, tal fato levou-se a concordar com Plotnikova, Dumas e Milani quando confirmaram a interdependências desta etapa com a anterior e sua forte iteração., os autores ainda comentam que chegam a substituir dados com o intuito de aprimorá-los ou incluir novos. ( (PLOTNIKOVA, DUMAS e MILANI, 2021, p. 6)). Assim, os exemplos dos relatórios e tarefas desta fase estão relacionados aos dados de São Luís, as demais capitais serão incluídas conforme Cronograma do capítulo 4 nos mesmos moldes dos elencados a seguir.

Neste sentindo, a “Coleta dos dados” de São Luís iniciou-se com um pedido formal ao site da Ouvidoria, mas como não foi respondida[[2]](#footnote-2) a solicitação dos dados, todos foram adquiridos nos Portais da Transparência, evitando quaisquer questões de sigilo fiscal, o que poderia ser um risco se fossem tratadas as receitas próprias ( (BRASIL, 2011)). Sendo importante enfatizar que os dados obtidos não ferem nenhum dos ditames da Lei Nº 12.527, de 18 de novembro de 2011, a qual versa sobre o acesso à informação no setor público brasileiro, já que elas estão disponíveis nos sites de Transparência de Entidades do setor público, confirmaram ( (BRASIL, 2011)). Os requisitos para coleta dos dados foram observados na fase de entendimento dos dados, descritas e especificadas as relações de cada Receita Pública com as suas possíveis variáveis preditoras (explicativas ou dependente) na Fundamentação Teórica. Vale comentar que as Transferências da União possuem um padrão divulgação, em detalhamento de repasses no Portal da Transparência do Governo Federal, já as Transferências dos Estados, provavelmente poderão apresentar mais buscas ou alterações desta proposta ao longo da execução por falta de padronização entre os Governos Estaduais.

Para as tarefas de “Descrever dados” e “Explorar dados” ( (CHAPMAN, CLINTON, *et al.*, 2000, p. 19)), foram incluídos os “APÊNDICE D” e “E”. O primeiro, trata do Relatório de descrição de dados, que em uma tabela, elencou todas as 85 variáveis testadas na 1ª iteração com os dados de São Luís, foram identificadas por tabela de origem na tarefa anterior e sua relação com cada receita, uma breve descrição, a periodicidade de publicação e frequência das observações e a identificação da fonte de origem, sendo a variável “Data” separada como índice e a “Realizado” da tabela “*f\_Programaca\_Financeira*” a variável alvo. O segundo, versa sobre o Relatório de Exploração de Dados, este foi elaborado no *Jupyter Notebook*, utilizando o pacote *pandas profiling[[3]](#footnote-3)* na linguagem *Python*.

A atividade de “Verificar a qualidade dos dados” que remete a preparação de um Relatório de qualidade de dados ficou mais ágil com a utilização do pacote do *pandas profiling,* uma vez que foi possível identificar a divergência nos dados das RECEITAS, pois esta variável deveria ter a mesma quantidade de observações para todas, no caso 132 conforme distribuição da Tabela 2, com exceção da CFEM AFETADOS que iniciou a distribuição no exercício de 2019 e as receitas das Emendas Constitucionais (EC) do FPM que são repasses anuais.

Tabela - Distribuição quantitativa e percentual das 928 observações por receita



Fonte: Adaptado do APÊNDICE E.

A questão da identificação, principalmente, no repasse da cota-parte do IPI-Exportação, levou a notificação formal do Agente Financeiro e do Governo do Estado do Maranhão a prestarem esclarecimento, os quais culminaram no depósito dos valores descobertos na 1ª iteração apontados como ausentes dos meses de janeiro e maio, em junho de 2022 (ver APÊNDICE F). Ainda foram encontrados outros valores que não estão bem compreensíveis, quando são comparados os montantes que o Governo Federal divulga com o que o Governo Municipal de São Luís registra em seu Balancete e as publicações do Agente Financeiro. Ademais, os valores atípicos, como exemplo da CFEM do Produtor que foi explicado por uma multa aplicada a uma empresa e foi estornado do montante como um *outlier*. Destarte, pretende-se repassar a cada uma das Capitais o mesmo relatório de descoberta da qualidade dos seus dados, caso tenha ocorrência semelhante.

Neste sentido, entende-se que há poucas publicações voltadas para o estudo de métodos de ML em previsão de receitas públicas, e menos ainda com a proposta de seleção de variáveis para cada receita, bem como a relevância de manter a sazonalidade da série temporal. Portanto a proposta de aplicação de técnicas de ML para previsão das receitas públicas, em séries temporais com utilização de RNN e LSTM é uma indicação dos pesquisadores Carmo, Komati e Boldt, (2019).

**FERRAMENTA**

Para a coleta e unificação do *Dataset* foi utilizado o s*oftware Power BI Desktop,* versão: 2.96.901.0 64-bit, ferramenta gratuita da *Microsoft.*

**CONFIGURAÇÔES INCIAIS**

Para acessar a API dos dados abertos do Portal da Transparência do Governo Federal será necessário fazer um cadastro no sistema do Gov.br, autorizando uma chave de acesso (token), a qual é fornecida pelo e-mail cadastrado na conta, melhores esclarecimento no site[[4]](#footnote-4).

Para as configurações iniciais foram criadas as variáveis globais que serão utilizadas em várias tabelas, por isso foram concentradas no grupo de configurações, sendo elas elencadas no Quadro 1.

Quadro - Lista de variáveis globais

|  |  |  |  |
| --- | --- | --- | --- |
| Denominação | Tipo | Descrição da variável global | Tabelas dependentes |
| *Caminho* | *Text* | Determina o caminho de um arquivo físico que por algum motivo não está sendo adquirido pela internet. |  |
| *MenorData* | *Date* | Pode ser incluída manualmente ou por código, o qual verifica nas receitas das capitais a maior data entre as menores datas dos dados disponíveis. | d\_Calendario, f\_Dolar\_Comercial, f\_SistNacPesqCustIndConstCivil |
| *DataAtual* | *Text* | Calcula a data atual do sistema operacional convertida para texto, formato “MM-dd-yyyy”. | *MaiorData*, f\_Dolar\_Comercial, f\_SistNacPesqCustIndConstCivil |
| *MaiorData* | *Date* | Também pode ser determina a maior data, configurada para o dia 15 de dezembro do ano da variável *DataAtual*. Podendo ser alterada manualmente. | d\_Calendario, Mese\_MenorAno |
| *VarPerCalendario* | *Text* | Em formato de texto cria uma lista para as consultas das API do IBGE por Mês e Ano, utilizando a lista da coluna “[AnoMes]” da tabela oculta “VarPerCalendario”. Fica oculta para o Painel. | f\_SistNacPesqCustIndConstCivil |
| *AnoMesAtual* | *Text* | Pela data atual do sistema extrai ano e mês anterior ao atual para utilizar nas buscas nas APIs. | *VarPerCalendario*, f\_SistNacPesqCustIndConstCivil |

Uma configuração muito importante a ser feita é a criação de um ambiente virtual para o *Python*, pois os *web scraping* e *crawling* foram executados com scripts desta linguagem, uma vez que na linguagem “M” do Power Query estava consumindo muita memória, tornando o processo mais demorado. Neste sentido deve criar um ambiente virtual no prompt do Anaconda, denominado de *“envpwbicap”*, além de configurar o Power BI Desktop para utilizar este ambiente[[5]](#footnote-5)*,* conforme Figura 2. Devem ser instaladas neste ambiente os pacotes:

* A do*“Jupyter”* foi a IDE definida para utilização neste trabalho;
* A do *“pandas”,* já instala junto a do *“numpy”* e a do “*matplotlib”,*
* *selenium,*
* HTMLParser,
* webdriver-manager,
* beautifulsoup4,
* lxml,
* html5lib,
* bs4,
* collections-extended,
* os-sys.
* pip install zipfile36
* pip install openpyxl
* pip install beautifulsoup4
* pip install fsspec

Pandas: a principal e mais completa biblioteca para análise de dados. Fornece ferramentas e estruturas de dados de alta performance.

Numpy: é usada principalmente para computação científica. Fornece um grande conjunto de funções e operações que ajudam a executar facilmente cálculos numéricos.

Zipfile: fornece ferramentas para criar, ler, escrever, adicionar, e listar um arquivo ZIP.

Requests: é uma biblioteca HTTP para Python simples

urllib3 - esta biblioteca possui recursos que nos permite conectar a um website e recuperar de lá o conteúdo de uma página.

beautifulsoap - esta biblioteca nos facilita a manipulação de HTML da página onde desejamos baixar os arquivos.

wget - uma das várias bibliotecas Python que possuem recursos para realizar download de dados de um endereço na internet.

zipfile - uma biblioteca para trabalhar com arquivos compactados.

Figura - Criação, ativação e instalação de bibliotecas no ambiente virtual *“envpwbicap”*

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamenteInterface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

A questão das credenciais

Site do IPEA precisam ser anônimas,

2.1 Coletar dados iniciais

A Coleta inicial dos dados, na primeira iteração contou com uma topologia estrela em um único arquivo do Power BI, mas com a inclusão das informações das 8 Capitais foi necessário reformular a arquitetura, pois a consulta das API e aos sites, bem como o download e o processamento exigiram bem mais que o modelo anterior.

A topologia passou a ser *snowflake* (floco de neve), pois reúnem-se as tabelas entorno da data do calendário e dos Estados e/ou suas Capitais.

Adquira dentro do projeto os dados (ou acesso aos dados) listados nos recursos do projeto. Essa coleta inicial inclui o carregamento de dados, se necessário, para a compreensão dos dados. Por exemplo, se você pretende usar uma ferramenta específica para compreensão de dados, é lógico carregar seus dados nessa ferramenta.

Saída do relatório de coleta de dados inicial

Liste todos os vários dados usados ​​no projeto, juntamente com qualquer requisitos de seleção para dados mais detalhados. A coleta de dados O relatório também deve definir se alguns atributos são relativamente mais importantes que outros.

Lembre-se de que qualquer avaliação da qualidade dos dados deve ser feita não apenas das fontes de dados individuais, mas também de quaisquer dados provenientes da fusão de fontes de dados. Os dados mesclados podem apresentar problemas que não existem nas fontes de dados individuais devido a inconsistências entre as fontes.

Atividades Planejamento de requisitos de dados

n Planeje quais informações são necessárias (por exemplo, apenas determinados atributos, informação adicional).

n Verifique se todas as informações necessárias (para resolver os objetivos de mineração de dados) estão realmente disponíveis.

Critério de seleção

n Especifique os critérios de seleção (por exemplo, quais atributos são necessários para os objetivos de mineração de dados especificados? Quais atributos foram identificados como irrelevantes? Com ​​quantos atributos podemos lidar

as técnicas escolhidas?).

n Selecione as tabelas/arquivos de interesse.

n Selecione os dados em uma tabela/arquivo.

n Pense em quanto tempo de histórico deve-se usar, mesmo se disponível (por exemplo, mesmo que os dados de 18 meses estejam disponíveis, talvez apenas 12 meses sejam necessários para o exercício).

Cuidado! Esteja ciente de que os dados coletados de diferentes fontes podem dar origem a problemas de qualidade quando mesclados (por exemplo, arquivos de endereço mesclados com a própria base de clientes podem mostrar inconsistências de formato, invalidade de dados, etc.).

Inserção de dados

n Se os dados contiverem entradas de texto livre, precisamos codificá-las

para modelagem ou queremos agrupar entradas específicas?

n Como os atributos ausentes podem ser adquiridos?

n Descreva como extrair os dados.

Boa ideia! Lembre-se de que algum conhecimento sobre os dados pode estar em fontes não eletrônicas (por exemplo, Pessoas, Texto impresso etc.).

Lembre-se que pode ser necessário pré-processar os dados (dados de séries temporais, médias ponderadas, etc.).

Os resultados da fase de compreensão de dados geralmente são documentados em vários relatórios. Idealmente, esses relatórios devem ser escritos durante a execução das respectivas tarefas. Os relatórios descrevem os conjuntos de dados que são explorados durante a compreensão dos dados. Para o relatório final, basta um resumo das partes mais relevantes.

Relatório de coleta de dados inicial

Este relatório descreve como as diferentes fontes de dados identificadas no inventário foram capturadas e extraídas.

Tópicos a serem abordados:

n Histórico dos dados.

n Lista de fontes de dados com ampla área de dados necessários cobertos por cada uma.

n Para cada fonte de dados, método de aquisição ou extração.

n Problemas encontrados na aquisição ou extração de dados.

Divergência entre os dados

Relacionadas as tabelas do IBGE, com dados das UF e Municípios, alguns valores de referência tiveram que ser ajustados, pois os dados da codificação, ou nomenclatura, estavam divergentes em outras tabelas e foram alterados para o padrão dos dados primários do IBGE, conforme listado na Tabela 3.

Quadro - Ajuste de dados de variáveis por divergência entre tabelas

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variável | Tabela Primária | Referência | **Dado Primário** | Tabela dependente | Dado alterado |
| *idUF* | D\_Estado\_API\_IBGE | Sergipe | **28** | f\_Export\_Import\_UF | 31 |
| Bahia | **29** | 32 |
| Minas Gerais | **31** | 33 |
| Espírito Santo | **32** | 34 |
| Rio de Janeiro | **33** | 36 |
| São Paulo | **35** | 41 |
| Paraná | **41** | 42 |
| Santa Catarina | **42** | 44 |
| Rio Grande do Sul | **43** | 45 |
| Mato Grosso do Sul | **50** | 55 |
| Mato Grosso | **51** | 52 |
| Goiás | **52** | 53 |
| Distrito Federal | **53** | 54 |
|  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Transformações de dados:

Quadro - Variáveis que foram transformadas

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variável resultante | Tabela | Descrição da transformação | Variáveis de cálculo | Cálculo utilizado |
| *Balanca\_ComerUF* | f\_Export\_Import\_UF | Valores das exportações e importações muito alto por isso resolveu-se calcular a balança comercial: o saldo da subtração entre as duas variáveis | a = [EXP\_UF]  b = [IMP\_UF] | (a - b) |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Em atendimento aos preceitos da Macroeconomia[[6]](#footnote-6), no que tange a expectativa de depreciação cambial para comparações, as tabelas cujas unidades dos dados são mensuradas em dólar, tiveram suas variáveis multiplicadas pela média da cotação do dólar no mês, transformando-a para real.

**COLETA DOS DADOS**

A primeira coleta foi iniciada em setembro de 2021, com as receitas do APÊNDICE B, que foram codificadas como: 1=CFEM, 2=IPI, 3=ITR, 4=FEP, 5=FPM, 6=ICMS, 7=IPVA e 8=CFEM AFETADOS. Destas, apenas o ITR foi descartado, pela sua menor relevância financeira. Devido às inúmeras fontes de dados e formatos, elas foram transformadas em 27 tabelas, quando do seu carregamento na ferramenta, com suas respectivas variáveis foram resumidas no Quadro 7.

Quadro - Resumo descritivo das tabelas por fonte, formato e variáveis da coleta inicial de dados

| Nº | Tabela | Descrição | Formato | Local | Variáveis | Fonte de Coleta |
| --- | --- | --- | --- | --- | --- | --- |
| 01 | f\_Programaca\_Financeira | Contém os dados da execução financeira do Município, da arrecadação em sua publicação contábil, publicados no portal da transparência do Governo Municipal, desta tabela retira-se o Valor Executado, a variável alvo. | Arquivo “PDF” e “CSV” | São Luís | Data, DATA, Cod\_Nivel04, DESC\_NV\_04, Classif/Codigo, Espec\_Receita, Fonte, "%", Tipo, RCL, Valor, Ano, Mes, RECEITA, RECEITA\_CODIGO, NOVA\_CLASSIF, RECEITA\_LRF, DUODECIMO, Nivel\_RCL, CodigoRCL, Receita\_RCL\_STN, Rubrica, NrMes, Destinacao, CodNivelRCL, Cód\_Desc | <https://saoluis.giap.com.br/apex/saoluis/wwv_flow.accept> |
| 02 | f\_Dolar\_Comercial | Contém os dados taxa de câmbio é o preço de uma moeda estrangeira medido em unidades ou frações (centavos) da moeda nacional, refletindo o custo de uma moeda em relação à outra. | Planilha eletrônica “XLS” | Aplicado ao Brasil | Data; DolarComercial | <https://www.bcb.gov.br/estabilidadefinanceira/historicocotacoes> |
| 03 | f\_Arrec\_Gov\_Estado | Reúne os dados da arrecadação total do Governo do Estado, neste caso. Este conjunto de dados foi carregado, limpo e transformado em um Relatório de BI exclusivo. | Site, páginas “HTML” | Estado MA | DATA, ICMS, ST\_EXTERNO, ST\_INTERNO, ICMS\_SIMP\_NACIO, ICMS\_DIV\_ATIV\_TRIB, IPVA, ICMS\_TOTAL | <http://sistemas.sefaz.ma.gov.br/arrecadacaoonline/arrecadacaoperiodo.html> |
| 04 | f\_Arrec\_Gov\_MA\_Municipios | Agrupamento das pesquisas feitas no site do Governo do Estado, Portal da Transparência para a arrecadação de tributos por Município. Este conjunto de dados foi carregado, limpo e transformado em um Relatório de BI exclusivo. | Arquivo “CSV” | São Luís | Data", Municipios, ICMS\_DvAtTrb\_Muns, IPVA\_Muns, TOTAL\_ArrecMun, ICMS\_Muns | <http://sistemas.sefaz.ma.gov.br/arrecadacaoonline/arrecadacaomunicipio.html> |
| 05 | f\_IndGeralPrecoMercado | Recolhe o Índice Geral de Preços - Mercado. Indicador mensal do nível de atividade econômica do país, englobando seus principais setores. | Site, páginas “HTML” | Aplicado ao Brasil | DATA, IGPM\_ACUMULADO, IGPM | <https://portal.fgv.br/noticias/igp-m-resultados-2021> |
| 06 | f\_Importacao\_Exportacao\_Brasil | Tabelas com as séries históricas dos Totais mensais dos valores das exportações e importações com os valores em dólar. | Planilha eletrônica “XLS” | Brasil | DATA, Exportacao, Importacao, Saldo, Corrente, Var\_Export\_MesAnoAnt, Var\_Import\_MesAnoAnt, Var\_Saldo\_MesAnoAnt, Var\_Corrente\_MesAnoAnt | <https://balanca.economia.gov.br/balanca/publicacoes_dados_consolidados/pg.html#s%C3%A9ries-hist%C3%B3ricas-detalhadas> ou <http://comexstat.mdic.gov.br/pt/home> |
| 07 | f\_IndNacPrecoCosumidor | Tabela do IBGE do Índice Nacional de Preços ao Consumidor nº 1736- INPC - Série histórica com número-índice, variação mensal e variações acumuladas em 3 meses, em 6 meses, no ano e em 12 meses (a partir de abril/1979). | JSON da API do IBGE | Brasil | DATA, INPC, INPC\_Nr\_Indice, INPC\_AC3M, INPC\_AC6M, INPC\_AC12M, INPC\_AC\_ANO | <https://servicodados.ibge.gov.br/api/docs/agregados?versao=3#api-bq> |
| 08 | f\_IndPrecoProdutor | Tabela do IBGE do Índice de Preços ao Produtor nº 6903- Índice de Preços ao Produtor, por tipo de índice, indústria geral, indústrias extrativas e indústrias de transformação e atividades (dezembro de 2018 = 100) | JSON da API do IBGE | Brasil | DATA, IPP, IPP\_Nr\_Indice, IPP\_M\_M-12 | <https://servicodados.ibge.gov.br/api/docs/agregados?versao=3#api-bq> |
| 09 | f\_PIB\_Vlr\_Corrente\_Brasil\_Milhoes | Unificação dos dados da Tabela 5938 - Produto interno bruto a preços correntes, impostos, líquidos de subsídios, sobre produtos a preços correntes e valor adicionado bruto a preços correntes total e por atividade econômica, e respectivas participações - Referência 2010 e as variação anual, trimestral e mensal da Tabela do PIB  Esta tabela foi elabora em um relatório do *Power Bi Desktop* exclusiva para depois ser juntada na geral | Planilha eletrônica “XLS”, Arquivo “CSV” e JSON da API do IBGE | Brasil, Estados e Municípios | DATA, PIB\_MensalBrasil, Ano, PIB\_Muni, RndPrCap\_Mun, VA\_BrtAdm, VA\_BrtAgro, VA\_BrtIndust, VA\_BrtServExclAdm, VA\_BrtTotal, VA\_ImpostosProd\_, PrcInvRndPrCpMn, FtInvRndPrCap\_Mun, PIB\_CorrTotal, PercPIB\_Total, PIB\_Municipal, RndPrCap\_Mun, VA\_BrutoAdministracao, VA\_BrutoAgropecuaria, VA\_BrutoIndustria, VA\_BrutoServicosExcluiAdm, VA\_BrutoTotal, VA\_ImpostosProdutos | <https://www.ibge.gov.br/estatisticas/economicas/contas-nacionais/9300-contas-nacionais-trimestrais.html?=&t=series-historicas&utm_source=landing&utm_medium=explica&utm_campaign=pib#evolucao-taxa>  <https://www.ibge.gov.br/estatisticas/economicas/contas-nacionais/9088-produto-interno-bruto-dos-municipios.html?t=resultados&c=2111300> |
| 10 | f\_SELIC | Histórico das taxas de juros fixadas pelo Copom e evolução que são registradas diariamente no Sistema Especial de Liquidação e de Custódia (Selic), consiste na taxa média cobrada em negociações com títulos emitidos pelo Tesouro Nacional, uma taxa de juros de referência para os demais juros da economia. | Site, páginas “HTML” | Brasil | Data, Taxa\_SELIC\_Periodo, | <https://www.bcb.gov.br/controleinflacao/historicotaxasjuros> |
| 11 | d\_Populacao\_Estimativa | Dados do Censo e Estimativa da população do IBGE.  Esta tabela foi elaborada em um relatório do *Power BI Desktop* exclusiva para posterior agregação a base geral | JSON da API do IBGE | Brasil, Estados e Municípios | Data, PopulacaoCapitais, FtPopulacao, POPULACAO, Per\_EntreCapitais, PerPopulacaoUF, DataCIFPM | <https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html>  <https://servicodados.ibge.gov.br/api/docs/agregados?versao=3#api-bq> |
| 12 | d\_IPCA | Agrupa o Índice Nacional de Preços ao Consumidor Amplo 15 (IPCA) da Tabela 3065 - IPCA15 - Série histórica com número-índice, variação mensal e variações acumuladas em 3 meses, em 6 meses, no ano e em 12 meses (a partir de maio/2000) do IBGE | Arquivo “CSV” | Brasil | Data, NR\_MÊS, INDICE\_IPCA, IPCA\_ANO, IPCA\_MES, IPCA\_12\_MESES, IPCA\_SEM, IPCA\_TRIM, Ano, TaxaIPCA | <https://sidra.ibge.gov.br/tabela/3065> |
| 13 | f\_Petroleo\_Gasolina | Série Histórica de Preços de Combustíveis, com base na pesquisa de preços da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis, no formato de dados abertos. | Arquivo “CSV” | Brasil | Data , Mês, Petóleo\_Brent\_Brasil, Gasolina\_Brasil | <https://www.gov.br/anp/pt-br/centrais-de-conteudo/dados-abertos/serie-historica-de-precos-de-combustiveis> |
| 14 | f\_ReservasInternacionais | Série histórica da tabela nº 3546 de Reservas internacionais - Total - mensal que, segundo o Banco Central do Brasil, são "Ativos externos do Brasil, geridos pelo BC. São recursos para utilização imediata, se eventualmente for necessário financiar saldos negativos das contas externas e/ou estabilizar o mercado de câmbio" | Arquivo “CSV” | Brasil | Data, ResvInternac | <https://www.bcb.gov.br/estatisticas/grafico/graficoestatistica/reservasInternacionais>  <https://www3.bcb.gov.br/sgspub/localizarseries/localizarSeries.do?method=prepararTelaLocalizarSeries> |
| 15 | f\_Exportacao\_SaoLuis | Base de dados do *Comex Stat* de Exportações de 2. Base de dados detalhada por Município da empresa exportadora/importadora e Posição do Sistema Harmonizado (SH4). | Arquivo “CSV” | Estado MA e São Luís | Data", "TransacoesEspeciais", "Prod\_Minerais", "MetaisCcomuns", "MadCarvVegObrMad", "EXPT\_SLZ\_ProdIndQuim", "Prod\_InduAlim", "Prod\_ReinoVegetal", "PedrasPreciosas", "InstrAparFoto", "MaqAparEletro", "MatTexteis", "MercdProdDiv", "MatTransporte", "PastMadCelulo", "AnimaisVivos", "GordOleosAnimVeget", "PlastBorracha", "ObrPedGesCimCerVid", "CalcChapArtFloresArtif", "PelCourBolsas", "EXPORT\_SLZ\_TOTAL | <http://comexstat.mdic.gov.br/pt/municipio>  <https://www.gov.br/produtividade-e-comercio-exterior/pt-br/assuntos/comercio-exterior/estatisticas/base-de-dados-bruta> |
| 16 | f\_Resultados\_Estado\_Municipios | Diferença entre o que o governo arrecada e gasta, serve para abater a dívida, mas exclui os juros. No caso de ocorrência de superavit, o governo abate sua dívida. Já o resultado nominal é a soma do resultado primário com os juros a pagar. Assim foram unificadas as tabelas do Banco Central do Brasil de: Necessidades de financiamento (14821, 14847 e 14873), Outros fluxos (ajustes, 14899), Dívida líquida (15125, 15538 e 15668), Receita dos estados e municípios (4336 e 4363) e Dívida líquida (5798). Todas terão que ser revisadas pois foram descontinuadas pelo Banco Central. | Arquivo “CSV” | Estado MA e São Luís | DataComp, Trimestre, FinMA\_MnRestPrim, FinMA\_MnRestPrim, FinMA\_Mn\_JrNom, Ajustes\_MA\_Munic, DvLq\_MA\_Munic, DvLq\_Gov\_MA, DvLq\_MA\_Munic, Receita\_Arrec\_ICMS\_MA, Receita\_Transf\_Uniao\_MA, DvLq\_MA\_ExcCapital | <https://www3.bcb.gov.br/sgspub/localizarseries/localizarSeries.do?method=prepararTelaLocalizarSeries> |
| 17 | F\_FROTA\_Municipio | Dados do Departamento Nacional de Trânsito (DENATRAN) do Ministério da Infraestrutura contendo a quantidade de veículos da frota de veículos do Município, advindos da tabela de Frota Nacional, 02 - Frota por Município e Tipo. | Arquivo “CSV” | São Luís | DATA, Veic\_PgtoIPVA | <https://www.gov.br/infraestrutura/pt-br/assuntos/transito/conteudo-Senatran/estatisticas-frota-de-veiculos-senatran> |
| 18 | f\_Arrecadacao\_CFEM | Resgate da divulgação no site da Agência Nacional de Mineração (ANM) da arrecadação dos Estados e Municípios Produtores de bens minerais. | Site, páginas “HTML” | Estado MA e São Luís | DATA, CFEM\_MA, CFEM\_Total, CFEM\_Municipio, DataRepasse | <https://sistemas.anm.gov.br/arrecadacao/extra/relatorios/arrecadacao_cfem.aspx> |
| 19 | f\_SalarioMinimo | Define o limite mínimo para os salários pagos na produção de bens e serviços, assim como benefícios previdenciários. | Site, páginas “HTML” | Brasil | Data, Dispositivo legal, SimboloMoeda, SalarioMinimo, DATAPublicacao, DataVigencia | <http://www.ipeadata.gov.br/exibeserie.aspx?stub=1&serid1739471028=1739471028> e <http://www.yahii.com.br/Salariomi.html> |
| 20 | f\_TaxaDesocupadosBrasil | Total de pessoas sem trabalho, procurando emprego, de 14 anos ou mais de idade. Pesquisa Nacional por Amostra de Domicílios Contínua mensal, Tabela 4092 - Pessoas de 14 anos ou mais de idade, por condição em relação à força de trabalho e condição de ocupação. | Arquivo “CSV” | Brasil | DATA, TxDesocupados | <https://sidra.ibge.gov.br/tabela/4092> |
| 21 | f\_Pesq\_Indus\_Mensal\_Prod\_Física | Tabela da Pesquisa Industrial Mensal - Produção Física, disponível pela Tabela 3653 - Produção Física Industrial, por seções e atividades industriais (série encerrada em janeiro de 2022, vai precisar ser revisada) | Arquivo “CSV” | Brasil | DATA, PIM\_PF | <https://sidra.ibge.gov.br/tabela/3653> |
| 22 | f\_SistNacPesqCustIndConstCivil | De acordo com o IBGE, “o Sistema Nacional de Pesquisa de Custos e Índices da Construção Civil - SINAPI tem por objetivo a produção de séries mensais de custos e índices para o setor habitacional, e de séries mensais de salários medianos de mão de obra e preços medianos de materiais, máquinas e equipamentos e serviços da construção para os setores de saneamento básico, infraestrutura e habitação.” Aqui foi utilizada a Tabela nº 647 - Custo de projeto m², por tipo de projeto e padrão de acabamento | Arquivo “CSV” ou JSON da API do IBGE | Estado do MA | DATA, SINAP\_Alto, SINAP\_Normal, SINAP\_Baixo, SINAP\_Mínimo | <https://sidra.ibge.gov.br/pesquisa/sinapi/tabelas> ou <https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html>  <https://servicodados.ibge.gov.br/api/docs/agregados?versao=3#api-bq> |
| 23 | f\_ResultadoPerc\_Primario\_Nominal | Mesma tabela f\_Resultados\_Estado\_Municipios, mas esta é a nível Nacional os resultados do Governo Federal. | Arquivo “TXT” | Brasil | Data, DataVigencia, PercResSP\_Nom, PercResSP\_Prim | <https://www.bcb.gov.br/estatisticas/grafico/graficoestatistica/resultados> |
| 24 | f\_IBC\_BR\_AtivEcon\_Per | Na descrição do Banco Central esta tabela reúne “O Índice de Atividade Econômica do Banco Central (IBC-Br) reflete as atividades da indústria, dos serviços e da agropecuária.” Foi utilizada a tabela nº 24364 - Índice de Atividade Econômica do Banco Central (IBC-Br) - com ajuste sazonal. | Arquivo “TXT” | Brasil | Data, IBC\_Br, DateTime | <https://www.bcb.gov.br/estatisticas/grafico/graficoestatistica/ibcbr> e [SGS - Sistema Gerenciador de Séries Temporais (bcb.gov.br)](https://www3.bcb.gov.br/sgspub/localizarseries/localizarSeries.do?method=prepararTelaLocalizarSeries) |
| 25 | d\_CoeficientesCIFPM | Tabela com o agrupamento dos coeficientes individuais de distribuição do FPM para os municípios, sendo isolado apenas o de São Luís | Planilha eletrônica “XLS”, Arquivo “PDF” | Municípios de Brasil (São Luís) | Data, Coef\_CIFPM\_SLZ | <https://portal.tcu.gov.br/transferencias-constitucionais-e-legais/coeficientes-fpe-e-fpm/> |
| 26 | d\_Coeficiente\_IPM | Tabela com unificação do coeficiente do Índice de Participação dos Municípios para distribuição no Estado da receita do ICMS, IPI e 50% da CIDE, separado apenas o Município de São Luís | Arquivo “PDF” | Municípios do MA  (São Luís) | Data, Coef\_IPM | <https://sistemas1.sefaz.ma.gov.br/portalsefaz/jsp/pagina/pagina.jsf?codigo=70> |
| 27 | d\_Calendario | Tabela criada por *procedure* em linguagem "M" para criação dos relacionamentos entre as tabelas pela chave primária "Data". Depende de dois parâmetros que recebem 2 variáveis globais: "MenorData" e "MaiorData". | Não se aplica | Não se aplica | Data, Ano, MesNr, Nome\_Mes, Trimestre, Quinzena, Periodo\_Quinzena, BimNr, Bimestre, MES\_ANO | Função que verifica a Maior e a Menor data entre as tabelas elencadas. |

Fonte: Elaborado pela Autora (2022).

**ETAPAS DE EXTRAÇÃO, CARREGAMENTO E TRANSFORMAÇÃO DOS DADOS**

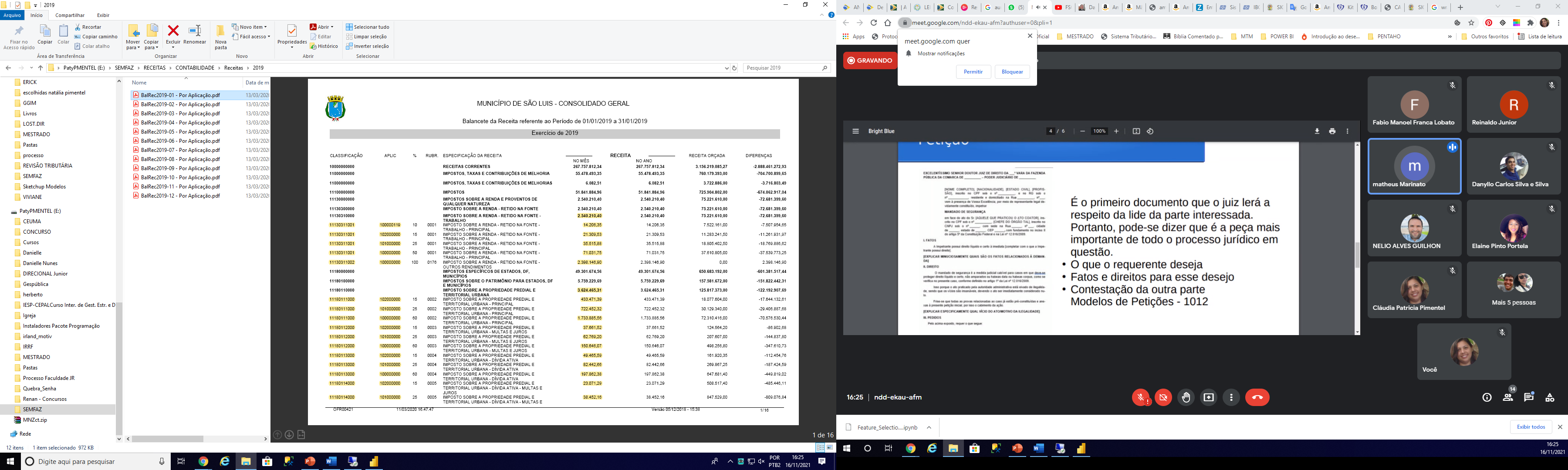
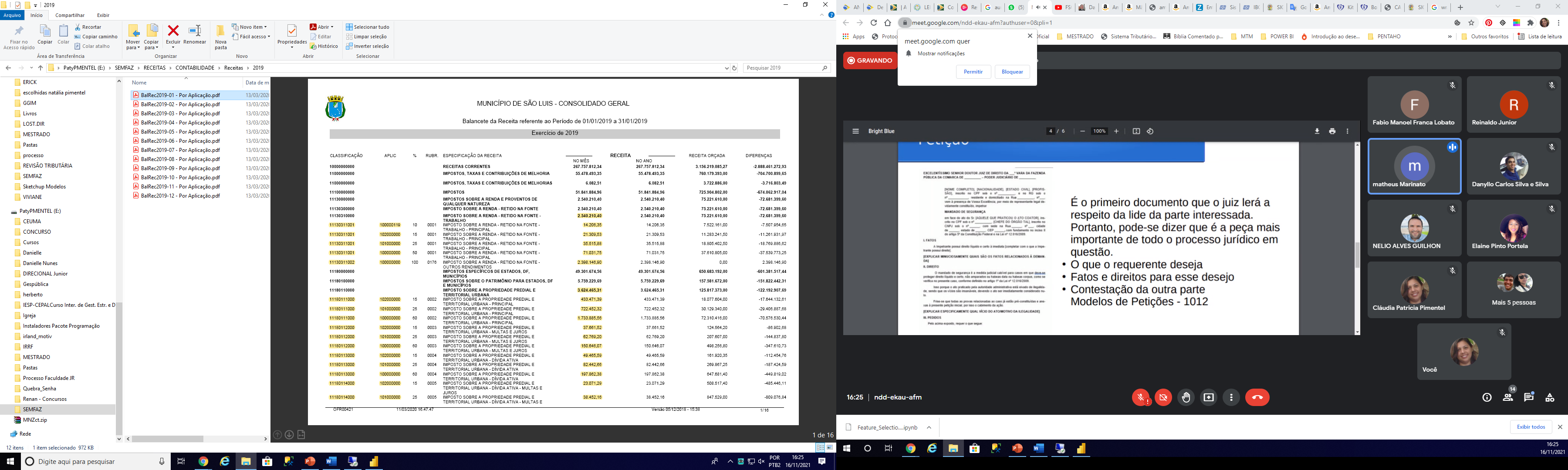
**Tabela 01 - *f\_Programaca\_Financeira***

**Descrição:** Contém os dados da execução financeira do Município, da arrecadação em sua publicação contábil, publicados no portal da transparência do Governo Municipal.

**Fonte de coleta:** <https://saoluis.giap.com.br/apex/saoluis/wwv_flow.accept>.

**Formato da fonte:** Parte dos arquivos foram reutilizados do Trabalho de conclusão de curso, os quais estavam em formato “PDF” e os demais foram baixados no portal do Governo Municipal de São Luís em formato “CSV”. A Figura 6 apresenta um exemplo do arquivo em “PDF” do Balancete da Receita do Governo Municipal de São Luís.

Figura - Visualização do Balancete da Receita



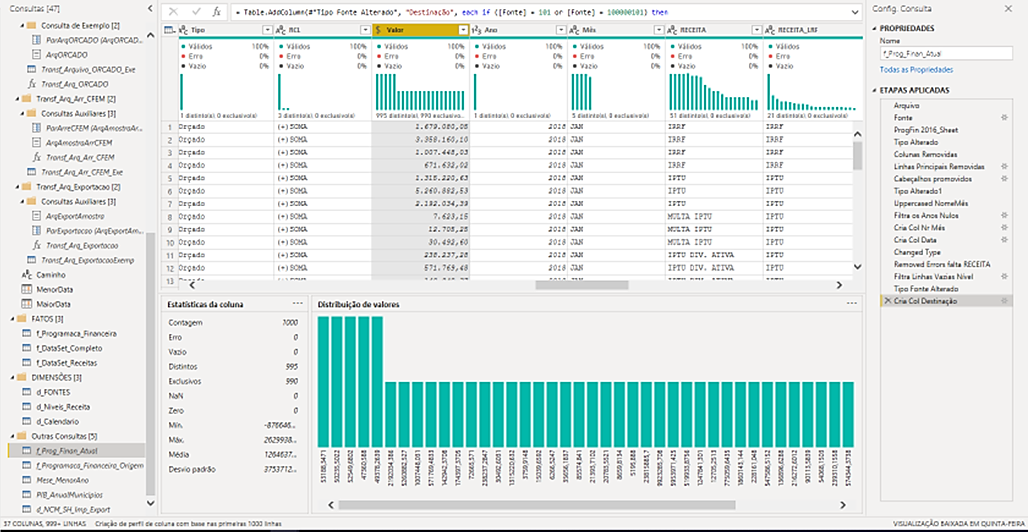
Fonte: Balancete da Receita (SÃO LUÍS-MA, 2021b).

**Periodicidade da frequência dos dados:** Anual e bimestral.

**Periodicidade de Publicação:** ao final de cada bimestre.

**Limpeza e Transformação:** Os procedimentos efetuados para ajustar os dados carregados foram efetivados no Relatório Principal do Power BI Desktop, no arquivo denominado V6\_BI\_RECEITAS\_TRANSF\_DADOS\_CAPITAIS, em que serão unificadas todas as informações das Capitais. Por intermédio da Figura 7 é possível ler o roteiro das “ETAPAS APLICADAS”, as estatísticas descritivas de cada variável da tabela, quantidade de erros, valores nulos e ausentes.

Figura - Limpeza e Transformação dos dados da tabela “*f\_Programaca\_Financeira”* do Município de São Luís



Fonte: Imagem do componente “*Power Query Editor*” da ferramenta *Power BI Desktop*.

**Tabela 01 -** *f\_Dolar\_Comercial*

**Descrição:** De acordo com o Instituto de Pesquisa Econômica Aplicada (IPEA), *“a taxa de câmbio é o preço de uma moeda estrangeira medido em unidades ou frações (centavos) da moeda nacional, refletindo o custo de uma moeda em relação à outra. [...] Também conhecida como PTAX, esta cotação corresponde à média aritmética das taxas de compra das consultas realizadas diariamente”* (BRASIL; IPEA, 2022)*.*

Ajuste dos dados para mensal Média entre o maior e o menor valor cotado do dólar comercia no mês. Cotação do Dólar dos EUA. Cotações de fechamento de uma moeda em um período.

**Fonte de coleta:** <https://www.bcb.gov.br/estabilidadefinanceira/historicocotacoes>.

**Formato da fonte:** página na web e planilhas eletrônicas tipo “XLS”.

**Periodicidade da frequência dos dados:** Diário.

**Periodicidade de Publicação:** ao final de cada dia.

**Limpeza e Transformação:** Os procedimentos efetuados para ajustar os dados carregados foram efetivados no Relatório Principal do Power BI Desktop, no arquivo denominado V6\_BI\_RECEITAS\_TRANSF\_DADOS\_CAPITAIS, em que serão unificadas todas as informações das Capitais. Por intermédio da Figura 8 é possível ler o roteiro das “ETAPAS APLICADAS”, as estatísticas descritivas de cada variável da tabela, quantidade de erros, valores nulos e ausentes.

Figura - Limpeza e Transformação das variáveis da tabela “*f\_Dolar\_Comercial*”

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

Fonte: Imagem do componente “*Power Query Editor*” da ferramenta *Power BI Desktop*.

Concluída a coleta dos dados, todas as tabelas foram relacionadas por intermédio do campo “Data” da tabela “*d\_Calendario*” e suas respectivas colunas de datas na cardinalidade de 1 para muitos, utilizando a topologia estrela, conforme ilustração da

Figura - Unificação das tabelas para criação do *Dataset*

Fonte: Imagem do componente “Modelagem” da ferramenta *Power BI Desktop*.

Referências

ACKERMANN, A. E. F.; SELLITTO, M. A. Métodos de previsão de demanda: uma revisão de literatura. **Innovar**, 32, Julo-Setembro 2022. 83-99.

BRASIL. Constituição (1988). **Constituição da República Federativa do Brasil de 1988**, outubro 1988. Disponivel em: <http://www.planalto.gov.br/ccivil\_03/Constituicao/Constituicao.htm>. Acesso em: 16 set. 2019.

BRASIL. Lei Complementar nº 101. **Estabelece normas de finanças públicas voltadas para a responsabilidade na gestão fiscal e dá outras providências**, Brasília, 4 Maio 2000.

BRASIL. Lei nº 12.527. **Regula o acesso a informações previsto no inciso XXXIII do art. 5º , no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga aLei nº 11.111, de 5 de maio de 2005, e dispositiv**, Brasília/DF, 18 Novembro 2011.

BRASIL, CONGRESSO NACIONAL. **Nota Técnica Conjunta nº 03: Subsídios à apreciação do Projeto de Lei de Diretrizes Orçamentárias para 2022**. Brasília. 2021.

BRASIL, CONTROLADORIA GERAL DA UNIÃO. Orçamento público. **Portal da Transparência**, s.d. Disponivel em: <https://www.portaltransparencia.gov.br/entenda-a-gestao-publica/orcamento-publico>. Acesso em: 26 Novembro 2021.

BRASIL, TRIBUNAL DE CONTAS DA UNIÃO. Política pública em dez passos, 2021. Disponivel em: <https://portal.tcu.gov.br/data/files/1E/D0/D4/DF/12F99710D5C6CE87F18818A8/Politica%20Publica%20em%20Dez%20Passos\_web.pdf>. Acesso em: 26 Novembro 2021.

BRASIL; IPEA. Taxa de câmbio comercial para compra: real (R$) / dólar americano (US$) - média. **Instituto de Pesquisa Econômica Aplicada**, 2022. Disponivel em: <http://www.ipeadata.gov.br/ExibeSerie.aspx?serid=38590&module=M>. Acesso em: 20 Julho 2022.

BRASIL; SOF. CLASSIFICAÇÃO DA RECEITA. In: BRASIL; FEDERAL, S. D. O. **Manual Técnico de Orçamento**. [Online]: [s.n.], 2021a. Cap. 3, p. n.p. Disponivel em: <https://www1.siop.planejamento.gov.br/mto/doku.php/mto2021:cap3#previsao>. Acesso em: 20 Novembro 2021.

BRASIL; SOF. ESTIMATIVAS DE RECEITAS ORÇAMENTÁRIAS. In: BRASIL; FEDERAL, S. D. O. **Manual Técnico de Oraçmento**. [Online]: [s.n.], 2021b. Cap. 7, p. n.p. Disponivel em: <https://www1.siop.planejamento.gov.br/mto/doku.php/mto2021:cap7>. Acesso em: 26 Novembro 2021.

CHAPMAN, P. et al. **CRISP-DM 1.0:** Step-by-step data mining guide. [S.l.]: [s.n.], 2000. Disponivel em: <https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72>. Acesso em: 10 Julho 2022.

GÉRON, A. **Mãos à Obra Aprendizado de Máquina com Scikit-Learn & TensorFlow:** Conceitos, Ferramentas e Técnicas Para a Construção de Sistemas Inteligentes. Tradução de Rafael Contatori. 2. ed. Rio de Janeiro: Alta Books, 2021. 576 p.

HART, M. et al. Conceitos básicos para consumidores do serviço do Power BI. **Documentação do Power BI**, 14 Julho 2022. Disponivel em: <https://docs.microsoft.com/pt-br/power-bi/consumer/end-user-basic-concepts>. Acesso em: 20 Julho 2022.

MARTÍNEZ-PLUMED, F. et al. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. **IEEE Transactions on Knowledge and Data Engineering**, v. 33, p. 3048-3061, 01 Agosto 2021. ISSN 8. Disponivel em: <https://ieeexplore.ieee.org/document/8943998>. Acesso em: 08 Junho 2022.

NAIBERT, N. T. **Orçamento público: metodologia alternativa para previsão de receitas municipais**. Dissertação (Mestrado em Administração Pública) - FGV. Rio de Janeiro, p. 66. 2021.

PARMEZAN, A. R. S.; SOUZA, V. M. A.; BATISTA, G. E. A. P. A. Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. **Information Sciences**, 484, Maio 2019. 302-337. Disponivel em: <https://www.sciencedirect.com/science/article/pii/S0020025519300945>. Acesso em: 22 Junho 2022.

PLOTNIKOVA, V.; DUMAS, M.; MILANI, F. P. Adapting the CRISP-DM Data Mining Process: A Case Study in the Financial Services Domain. **Research Challenges in Information Science. RCIS 2021. Lecture Notes in Business Information Processing**, Limassol, Cyprus, 415, 08 Maio 2021. 55-71. Disponivel em: <https://www.semanticscholar.org/paper/Adapting-the-CRISP-DM-Data-Mining-Process%3A-A-Case-Plotnikova-Dumas/a4e44d5ce058e6b815631947db23044a22f8b114>. Acesso em: 10 Julho 2022.

SAMPAIO, I. G. et al. Avaliação de modelos de predição e previsão construídos por algoritmos de aprendizado de máquina em problemas de Cidades Inteligentes. In: DAVID, J. M. N., et al. **Tópicos em sistemas de informação [recurso eletrônico]:** minicursos do XIV Simpósio Brasileiro de Sistemas de Informação. 15. ed. Aracajú/SE: Sociedade Brasileira de Computação, 2019. Cap. 4, p. 81-113.

SÃO LUÍS-MA. Decreto Municipal n° 56.835. **Estabelece Normas para a Programação e Execução Orçamentária e Financeira dos Orçamentos Fiscal e da Seguridade Social do exercício 2021, e dá outras providências**, São Luis, 4 Janeiro 2021a. Disponivel em: <http://sistemas.semad.saoluis.ma.gov.br/easysearch/cachedownloader?collection=default&docId=ec69d2c164d0ccc69a4342ab3d463c9824f5ff02&fieldName=Visualizar&extension=bht.zip#q=execu%C3%A7%C3%A3o%20execu%C3%A7%C3%A3o%20financeira%202021>. Acesso em: 22 Outubro 2021.

SÃO LUÍS-MA. Receitas Total e detalhada. **Portal da Transparência**, 2021b. Disponivel em: <https://saoluis.giap.com.br/apex/saoluis/wwv\_flow.accept>. Acesso em: 15 Setembro 2021.

SCHRÖER, C.; KRUSE, F.; GÓMEZ, J. M. A Systematic Literature Review on Applying CRISP-DM Process Model. **Procedia Computer Science**, 181, 2021. 526-534. Disponivel em: <https://www.sciencedirect.com/science/article/pii/S1877050921002416>. Acesso em: 8 Junho 2022.

SEKAR, M. Introduction to AI, Data Science, and Machine Learning. In: SEKAR, M. **Machine Learning for Auditors:** Automating Fraud Investigations Through Artificial Intelligence. Calgary, AB, Canada: Apress, 2022. Cap. 4-6, p. 43-72.

WIRTH, R.; HIPP, J. Crisp-dm: towards a standard process modell for data mining, 2000. 11. Disponivel em: <https://www.semanticscholar.org/paper/Crisp-dm%3A-towards-a-standard -process-modell-for-Wirth-Hipp/48b9293cfd4297f855867ca278f7069abc6a9c24>. Acesso em: 25 Outubro 2021.

1. O termo “relatório do *Power BI* é composto por uma ou mais páginas de visualizações, grafos e texto. Todas as visualizações em um relatório vêm de um único conjunto de dados[...]” ( (HART, RATHORE, *et al.*, 2022)). [↑](#footnote-ref-1)
2. Foi cadastrada no dia 22 de novembro de 2021, como o e-mail institucional de aluno da UEMA, sob protocolo nº 11111000064202111, o acesso às séries históricas dos dados das receitas municipais no site e-OUV, no Sistema de Ouvidorias do Poder Executivo Municipal de São Luís, disponível no endereço: <http://eouv.saoluis.ma.gov.br:82/publico/Manifestacao/RegistrarManifestacaoBs.aspx>. [↑](#footnote-ref-2)
3. Maiores detalhamento disponíveis na documentação do pacote em <https://pandas-profiling.ydata.ai/docs/master/index.html>. [↑](#footnote-ref-3)
4. Portal da Transparência. API de dados – Cadastro. Disponível em: <https://www.portaldatransparencia.gov.br/api-de-dados/cadastrar-email>. Acessado em 11 outubro de 2022. Um exemplo, na linguagem *Python* pode ser encontrado no *Jupyter Notebook* disponível em: <https://colab.research.google.com/drive/1lRbpw9NfxiTrGTEcSuh4QNzX0hUjJV8c?usp=sharing#scrollTo=WII6A77Z2QMH>. [↑](#footnote-ref-4)
5. Detalhes para instalação, criação e execução de ambientes virtuais no Power Bi Desktop consultar o site da Microsoft, disponível em: <https://learn.microsoft.com/pt-br/power-bi/connect-data/desktop-python-scripts>. Acessado em 08 de outubro de 2022. [↑](#footnote-ref-5)
6. Curso de Introdução à Análise Macroeconômica, Coursera, 3 semana, Aula Determinantes da taxa de câmbio no curto prazo, em 10:51 a 13:55 minutos, [↑](#footnote-ref-6)