



CLÁUDIO HENRIQUE DE ARAÚJO COUTINHO
VICTOR MIRANDA DE SOUZA

PROJETO BIGDATA - GASTOS PÚBLICOS NA CÂMARA DOS DEPUTADOS

RIO DE JANEIRO

2020



CLÁUDIO HENRIQUE DE ARAÚJO COUTINHO
VICTOR MIRANDA DE SOUZA

PROJETO BIGDATA - GASTOS PÚBLICOS NA CÂMARA DOS DEPUTADOS

Trabalho de conclusão de curso
apresentado à Faculdade Senac como
requisito para a obtenção de aprovação
no curso de Pós-graduação lato sensu
Especialização em Big Data.

Docente orientador: Clayton Escouper
das Chagas

RIO DE JANEIRO

2020

C871 Coutinho, Cláudio Henrique de Araújo.

Projeto Big Data: gastos públicos na câmara dos deputados / Cláudio Henrique de Araújo Coutinho; Victor Miranda de Souza – Rio de Janeiro, 2020.

37 f.; il. ; 30 cm.

Orientador: Clayton Escouper das Chagas.

Trabalho de conclusão de curso (Pós-graduação lato sensu em Big Data) – Faculdade de Tecnologia Senac Rio, 2020.

Inclui bibliografia.

1. Big data. 2. Internet das coisas. 3. Recursos eletrônicos de informação. 4. Inovações tecnológicas. I. Coutinho, Cláudio Henrique de Araújo. II. Souza, Victor Miranda de. III. Título.

CDD 005.75



Faculdade de Tecnologia Senac

Curso: Pós-Graduação lato sensu Especialização em Big Data

Ano: 2020

Nome dos alunos: Cláudio Henrique de Araújo Coutinho

Victor Miranda de Souza

Título: Gastos Públicos na Câmara dos Deputados

Nome do docente orientador: Clayton Escouper das Chagas

Conceito: “O”

Recomendações:

Sem pendências ou recomendações, trabalho de conclusão de curso aprovado. Liberado para os trâmites finais do curso e solicitação do certificado, após verificação de outros requisitos (aprovação nas outras disciplinas/módulos do curso, documentação e financeiro).

Rio de Janeiro, 10 de OUTUBRO de 2020.

CLAYTON ESCOUPER DAS CHAGAS - Orientador

AGRADECIMENTOS

Primeiramente, agradecemos a Deus por ter dado a oportunidade de chegarmos até esta etapa em nossas vidas, e que, em todos os momentos, deu-nos forças para seguir em frente.

Agradecimento aos colegas da Pós-Graduação pelo conhecimento compartilhado.

Ao nosso amigo Vanderson Dutra por todo apoio desde o início da Pós até as últimas fases do nosso Projeto.

Agradecemos as nossas famílias e amigos.

Agradecemos a todos os nossos professores do SENAC RJ.

RESUMO

O projeto consiste em realizar uma análise dos gastos com dinheiro público e apresentar dados coletados relacionados às Despesas cobertas pela Cota para Exercício da Atividade Parlamentar de cada deputado de 2012 até o ano de 2020. Através dessas análises elaborar a construção de métricas para que se possa compreender e entender pontualmente as particularidades dos gastos da Câmara dos Deputados.

Propomos esse projeto com o intuito de medir e entender tais gastos de forma a demonstrar através de gráficos em um dashboard com o máximo de detalhes possíveis.

Ao final deste projeto, evidenciaremos quais despesas estão sendo gastas de forma desproporcional ao que deveria e propor melhorias que ajudem a manutenção dos gastos do dinheiro público.

Palavra-chave: Despesas, cota, atividade parlamentar, métricas, câmara dos deputados, análises, gráficos, dashboard, gastos, dinheiro.

ABSTRACT

The project consists of carrying out an analysis of public money expenditures and presenting data collected related to the Expenses covered by the Quota for the Exercise of Parliamentary Activity of each deputy from 2012 to 2020. Through these analyzes, elaborate the construction of metrics so that to understand and understand the specificities of the expenses of the Chamber of Deputies on time.

We propose this project in order to measure and understand such expenses from forming to demonstrating through graphics on a dashboard with as much detail as possible

At the end of this project, we will highlight which expenses are being spent disproportionately to what they should and propose improvements to help maintain the spending of public money.

Keywords: Expenses, quota, parliamentary activity, metrics, chamber of deputies, analysis, graphs, dashboard, expenses, money.

LISTA DE FIGURAS

| | |
|---|----|
| Figura 1 – Diagrama de arquitetura do projeto. | 15 |
| Figura 2 – Projeto no GITHUB. | 16 |
| Figura 3 – Obtenção dos dados em Python. | 17 |
| Figura 4 – Lista dos bancos de dados. | 19 |
| Figura 5 – Configuração do banco de dados no Google Cloud Platform. | 20 |
| Figura 6 – Limpeza dos dados. | 21 |
| Figura 7 – Tratamento dos dados. | 21 |
| Figura 8 – Instalação do mysql. | 22 |
| Figura 9 – Conexão com o banco de dados mysql. | 22 |
| Figura 10 – Lista de Conexões no Google Cloud Platform. | 23 |
| Figura 11 – Tela de conexão com o Banco no Power Bi. | 24 |
| Figura 12 – Modelo de relacionamento no Power Bi. | 26 |
| Figura 13 – Tabela período criada no Power Bi. | 26 |
| Figura 14 – Dashboard despesa de deputados. | 27 |
| Figura 15 – Dashboard despesa partidos. | 28 |
| Figura 16 – Dashboard análise despesa por quantidade de deputados por partidos. | 29 |
| Figura 17 – Dashboard despesa fornecedor. | 30 |
| Figura 18 – Dashboard proporção gênero por deputados | 31 |

SUMÁRIO

| | |
|---|-----------|
| 1. INTRODUÇÃO | 10 |
| 2. CONTEXTUALIZAÇÃO | 11 |
| 2.1. GASTOS PÚBLICOS EM TEMPO DE PANDEMIA | 11 |
| 3. REFERENCIAL TEÓRICO | 12 |
| 3.1 DADOS ABERTO DA CÂMARA DOS DEPUTADOS | 12 |
| 3.2. PYTHON COMO LINGUAGEM DE PROGRAMAÇÃO | 12 |
| 3.3 BANCO DE DADOS | 13 |
| 3.4 BIG DATA | 13 |
| 3.5 POWER BI | 14 |
| 4. ARQUITETURA E INFRAESTRUTURA | 15 |
| 4.1. DIAGRAMA DE ARQUITETURA | 15 |
| 4.2. GITHUB DO PROJETO | 16 |
| 5. ANÁLISE EXPLORATÓRIA | 17 |
| 5.1. DIAGRAMA DE ARQUITETURA | |
| 5.2. DATASETS | 18 |
| 5.3. ARMAZENAMENTO DOS DADOS | |
| 5.4. LIMPEZA DOS DADOS | 20 |
| 5.5. TRATAMENTO DOS DADOS | 21 |
| 5.6. CONEXÃO COM BANCO DE DADOS | 22 |
| 5.7. DEFINIÇÃO DE MÉTRICAS | 23 |
| 5.8. CONEXÃO COM O POWER BI | 24 |
| 6. VISUALIZAÇÃO DOS DADOS E ANÁLISE DOS RESULTADOS | 25 |
| 6.1. ANÁLISE DOS DATASETS | |
| 6.2. MODELO DE DADOS | 26 |
| 6.3. VISUALIZAÇÃO DOS DADOS | 27 |
| 6.3.1. DESPESA DE DEPUTADOS | 27 |
| 6.3.2. DESPESA PARTIDOS | 28 |
| 6.3.3. ANÁLISE DESPESA POR QUANTIDADE DE DEPUTADOS POR PARTIDOS | 29 |
| 6.3.4. DESPESA FORNECEDOR | 30 |
| 6.3.5. PROPORÇÃO GÊNERO POR DEPUTADOS | 31 |
| 6.4. MÉTRICAS | 32 |
| 6.5. ANÁLISE DOS RESULTADOS | 34 |
| 7. CONCLUSÃO | 35 |
| 8. REFERÊNCIAS | 36 |

1. INTRODUÇÃO

A câmara dos deputados foi criada pela primeira Constituição brasileira (1824-1891) juntamente com Senado, em 25 de março 1824. Na época era composta por 102 integrantes eleitos em eleições indiretas.

A constituição de 1988(a sétima do Brasil), estabeleceu o cenário que atualmente está em vigor: o quantitativo de 513 deputados eleitos pelo sistema do voto proporcional, onde cada deputado exerce seu cargo por quatro anos, sendo no mínimo 8 e no máximo 70 por unidade federativa, em totais que variam conforme a população do país.

A Cota para o Exercício da Atividade Parlamentar (CEAP) é uma cota única mensal destinada a custear as despesas do mandato, como passagens aéreas e conta de celular, vinculados ao exercício da atividade parlamentar. O valor máximo da cota depende do Estado que o deputado representa.

Cada deputado tem um montante de R\$ 111.657,59 por mês para pagamento de salários de um total de até 25 secretários parlamentares contratados diretamente pelos deputados, que trabalham para o mandato em Brasília ou nos estados.

O projeto consiste em realizar uma análise dos gastos com dinheiro público e apresentar dados coletados relacionados às Despesas cobertas pela Cota para Exercício da Atividade Parlamentar de cada deputado de 2012 até o ano de 2020. Através dessas análises elaborar a construção de métricas para que se possa compreender e entender pontualmente as particularidades dos gastos da Câmara dos Deputados.

Propomos esse projeto com o intuito de medir e entender tais gastos de forma a demonstrar através de gráficos em um dashboard com o máximo de detalhes possíveis.

A base utilizada neste projeto, assim como o script utilizado para a geração do Dashboard das análises estão disponíveis em:

https://github.com/projetopossenac/claudio_vanderson_victor

2. CONTEXTUALIZAÇÃO

2.1. GASTOS PÚBLICOS EM TEMPOS DE PANDEMIA

Segundo Cláudio Humberto, do Diário do Poder (2020), mostra que “O pagador de impostos não teve alívio do setor público na pandemia do coronavírus no Brasil. Nem com propaganda própria. Deputados federais e senadores torraram, nos últimos quatro meses, mais de R\$11,3 milhões na “divulgação da atividade parlamentar”, segundo a ONG Operação Política Supervisionada. Essa conta faz parte do cotão parlamentar, que ressarce parlamentares em cerca de R\$ 45 mil por mês por qualquer tipo de despesa, de tapiocas a “consultorias”. “

Conforme o site INFOSAJ (2020), aborda o gasto desenfreado com combustíveis pelos deputados em tempos de pandemia. De acordo com a reportagem, todos os 513 deputados teriam gasto no período de março até setembro verba suficiente para um carro dar 268 voltas pelo planeta terra. “Um veículo que roda uma média de 10 quilômetros por litro, com gasolina a R\$ 4, se abastecido com R\$ 372.150, conseguiria dar 23 voltas na Terra. Esse foi o valor gasto com combustíveis pelos deputados federais baianos entre março e setembro, período em que a Câmara estava com atividades presenciais suspensas e o país em isolamento decretado em razão da pandemia do novo coronavírus.”.

De acordo com o RADARAMAZONICO (2020), a câmara dos deputados pagou cerca de R\$ 11,8 milhões em passagens, estampa a matéria de forma a criticar e fugar o leitor. Num período crítico com diversas campanhas e diversas medidas preventivas, observamos estes tipos de despesas com passagens aéreas.

3. REFERENCIAL TEÓRICO

3.1. DADOS ABERTO DA CÂMARA DOS DEPUTADOS

É um site que prove um serviço de dados abertos que permite qualquer cidadão obter informações anonimamente, ou seja, sem precisar se identificar através da internet e realizar consultas capazes de fiscalizar, monitorar, conhecer e discutir os gastos, ações e as decisões de cada entidade.

Os dados podem ser obtidos de diversas formas diferentes cada uma em um formato de arquivo. Através de sua API (API RESTful) pode-se obter os dados puros nos formatos JSON e XML. Outra forma de se obter os dados é o download de arquivos nas versões em XML, JSON, CSV, XLSX e ODS.

Atualmente a API encontra-se na versão: 0.4.23 – 18/09/2020, com atualizações constantes, não tendo uma versão completa, estando sujeita a mudanças rotineiras em suas estruturas.

A API, possui algumas limitações, por padrão todos os seus serviços de listagem retornam 15 itens, sendo o limite 100 itens por requisição. Esse detalhe, dificultou na consulta por essa estrutura, já que seriam muitos anos de busca e teríamos pouco tempo para elaborar algum script ou método mais sofisticado de sondar os dados.

3.2. PYTHON COMO LINGUAGEM DE PROGRAMAÇÃO

Python é uma linguagem de programação criada por Guido van Rossum em 1991. A linguagem tinha como objetivo a produtividade e legibilidade. O Python foi desenvolvido para ser produzido códigos limpos e de fácil usabilidade.

A linguagem de programação escolhida para o desenvolvimento do projeto é o Python inicialmente em sua versão 3.7 depois atualizada para versão 3.8 ao longo do projeto.

Foram utilizadas as bibliotecas: pandas, datetime, sqlalchemy, pymysql para obtenção, tratamento e limpeza dos dados.

- python datetime: biblioteca que fornece classes para manipulação de datas e horas.

- python pandas: biblioteca utilizada para análise e manipulação dos dados em diferentes tipos de arquivos.
- python sqlalchemy: biblioteca que permite a comunicação, manipulação, conexão e análise de banco de dados em python.
- pymysql: biblioteca que permite a comunicação, manipulação, conexão e análise de banco de dados MySQL.

3.3. BANCO DE DADOS

Segundo Christopher J. Date (Introdução a Sistemas de Dados, Editora Bookman, 2004), “Um banco de dados é uma coleção de dados persistentes, usados pelos sistemas de aplicação de uma determinada empresa.”.

O banco de dados que será utilizado no projeto é o MySQL em sua versão 5.7. dentro da infraestrutura do console do Google Platform.

3.4. BIGDATA

“Big Data não é uma única tecnologia, mas uma combinação de tecnologias novas e antigas que ajudam as empresas a conseguirem ideias viáveis (Alan Nugent, 2013).”.

“O Big Data é ferramenta tecnológica que possibilita às empresas criação de experimentos controlados para testar hipóteses que guiarão a tomada de decisão em, por exemplo, novos investimentos ou mudanças operacionais, possibilitando centenas ou milhares de experimentações. É possível distinguir entre simples correlação de eventos daqueles que realmente possuem uma ligação de causa e efeito (BROWN, 2011).”.

“Tão importante quanto gerar informação é a capacidade de processamento de dados volumosos em alta velocidade. Isso se comprova pelo fato de que, nas últimas décadas, presenciamos o desenvolvimento de supercomputadores que atendam essa necessidade: quanto mais a tecnologia foi penetrando no meio social, mais informações as pessoas foram gerando e consumindo (Volpato, Rufino e Dias, 2014).”.

“As novas tecnologias e ideias, originadas da grande quantidade de dados e análises de Big Data promovem inovações em tecnologias, produtos, na gestão e na estratégia das organizações (Zhang, Chen, & Li, 2013).”.

O conceito de Big Data pode também ser definido pelos 5 Vs: Volume, Variedade, Velocidade, Veracidade e Valor. Além do grande volume de dados vindo de diversos locais dos mais variados tipos, um grande fator que determina é a velocidade que se obtém os dados. Neste caso, é necessário saber a veracidade desses dados e ser capaz de analisar se tais valores, agregam valor ou não para a empresa.

3.5. POWER BI

Conforme demonstrado na documentação oficial da Microsoft(docs.microsoft.com), “O Power BI é uma coleção de serviços de software, aplicativos e conectores que trabalham juntos para transformar suas fontes de dados não relacionadas em informações coerentes, visualmente envolventes e interativas. Os dados podem estar em uma planilha do Excel ou em uma coleção de data warehouses híbridos locais ou baseados na nuvem. Com o Power BI, você pode se conectar facilmente a fontes de dados, visualizar e descobrir conteúdo importante e compartilhá-lo com todas as pessoas que quiser.”.

Ou seja, é uma poderosa ferramenta de análise de dados para tomada de decisão através de relatórios gerenciais nos mais diversos dispositivos eletrônicos, obtidos através das mais variadas fontes de dados.

4. ARQUITETURA E INFRAESTRUTURA

4.1. DIAGRAMA DE ARQUITETURA

O diagrama de arquitetura do projeto foi dividido em processos, são eles: coleta dos dados através do site da Câmara dos Deputados, linguagem de programação Python, tratamento dos dados, armazenamento em banco de dados e visualização de dados através de Dashboards, conforme pode ser observado na Figura 1, abaixo.

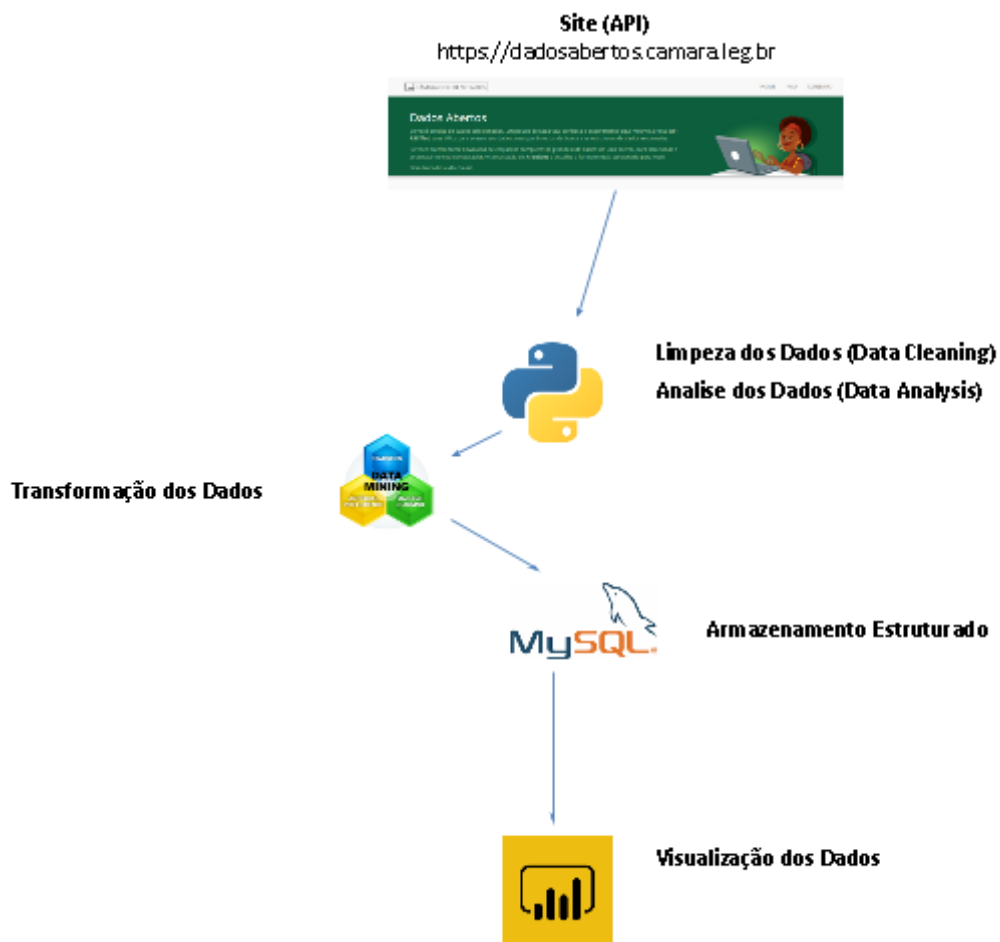


Figura 1

A escolha pela linguagem Python para utilização no projeto foi pela facilidade e usabilidade, além da grande comunidade e fóruns de ajuda.

O banco de dados MySQL era de conhecimento de boa parte dos integrantes do grupo, o que facilitou no momento da análise, manipulação e conexão da linguagem com o banco de dados.

O Power BI foi escolhido como ferramenta de visualização de dados, pelo fato de ser altamente conceituado e fácil de manipulação, estruturação e demonstração de resultado e utilização de métricas.

4.2. GITHUB DO PROJETO

Todo o projeto está armazenado na plataforma do github podendo ser consultado por qualquer usuário oriundo da internet. Disponibilizado através da url: https://github.com/projetopossenac/claudio_vanderson_victor e demonstrado na Figura 2, abaixo.

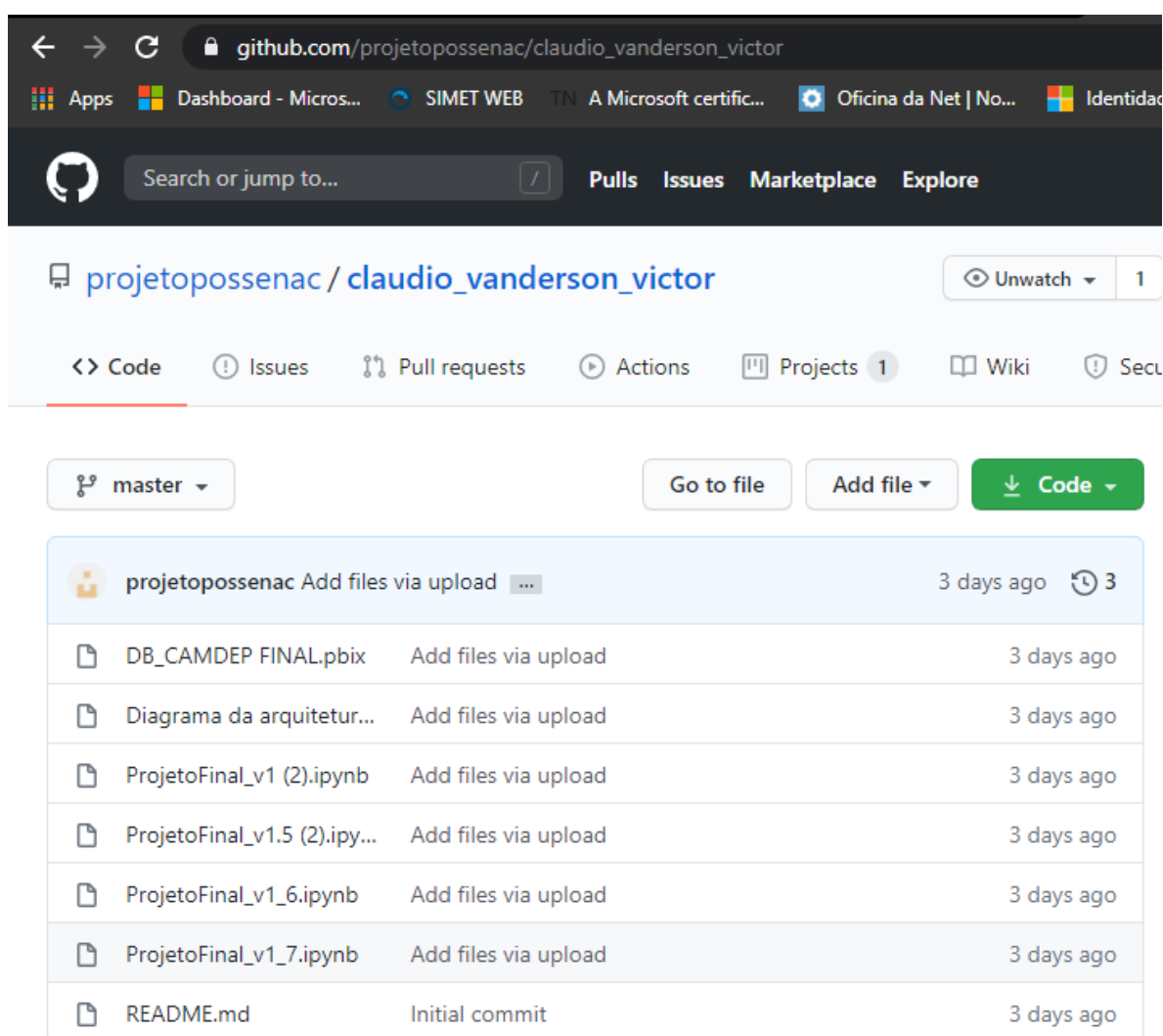


Figura 2

5. ANÁLISE EXPLORATÓRIA

5.1. OBTENÇÃO DOS DADOS

Para coleta e extração dos dados a primeira opção foi através da API RESTful por meio da linguagem de programação Python (versão 3.7) via o ambiente de notebook Jupyter Google Colaboratory onde não seria necessário configuração do ambiente, sendo totalmente executado via nuvem.

O acesso a API foi feito através da biblioteca do Python chamada *requests* onde era passada via URL o local dos dados e armazenados em um Data-Frame. Porém, a grande dificuldade foi para lidar com os constantes problemas que ocorriam nas consultas das diversas páginas onde tinham um limite de 100 itens.

Como estamos falando de um projeto de grandes proporções de dados, com mais de 5 milhões de linhas, o tempo para aperfeiçoar a técnica de extração foi se mostrando cada vez mais escasso e novas formas deveriam que ser desenvolvidas em tempo hábil.

Com isso a estratégia mudou, foi decidido que todos os dados deveriam ser obtidos por meio de arquivos no formato XLXS.

```
[ ] #Define qual é o ano atual para que o WHILE execute do ano 2012 até o ano atual
now = datetime.now()
ano = now.year
# Definição do dataset dfdespesa
df_despesa = pd.DataFrame()
#Estrutura de Repetição para buscar o DataSet ANO-DESPESA de varios anos.
while ano >= 2012:
    url = 'http://www.camara.leg.br/cotas/Ano-' + str(ano) + '.xlsx'
    df = pd.read_excel (url)
    ano = ano - 1
    df_despesa = pd.concat([df,df_despesa])
```

Figura 3

Conforme a imagem 3 acima, foram criadas 2 estruturas para lidar com os dados do tipo data (datetime.now e now.year). Criado um dataframe para armazenar os dados extraídos diretamente da url da Câmara dos Deputados no formato XLSX. Onde os valores dos dados percorriam na estrutura de repetição while do ano de 2012 até o ano presente 2020.

5.2. DATASETS

Para a construção do projeto, foram escolhidos os datasets Deputados e Despesa. Verificamos que esses seriam os principais para serem analisados, limpos e tratados pelas diversas informações que esses possuem.

O dataset de Deputados, contem as informações dos parlamentares necessárias para interligar com o dataset de Despesas.

O dataset de Despesas, verificamos que era a principal fonte de obtenção das informações que necessitaríamos demonstrar em nosso projeto.

Com esses datasets seria possível verificar o gasto dos deputados, partidos e construir métricas para realizar as apurações e medições necessárias para elaborar dashboards capazes de identificar as despesas e informações que fossem pertinentes para realizar as análises dos dados.

5.3. ARMAZENAMENTO DOS DADOS

O armazenamento dos dados em um primeiro momento foi pensando a ser utilizado através do MongoDB Atlas, um serviço de banco de dados em nuvem de documentos flexível e escalável, desenvolvido para ser totalmente gerenciado na AWS, Azure ou GCP.

Entretanto, seu limite para uso gratuito é de menos de 500 megabytes o que tornaria todo o projeto inviável.

Como alternativa, foi utilizado o Google Cloud Platform, uma plataforma na nuvem da Google capaz de executar diversos serviços em Cloud Computer (Computação em nuvem) com alta performance, segurança e confiabilidade.


Através dessa forma foi possível resolver o problema do armazenamento, sem custo e com espaço disponibilizado de até 20GB pela plataforma da Google.

Foi criado um script em Python para importação dos datasets para dentro do banco de dados MySQL na versão 5.7.


Lista dos bancos de dados

Bancos de dados

Todas as instâncias > projetocamara

 **projetocamara**

MySQL 5.7

 **CRIAR BANCO DE DADOS**

| Nome ↑ | Compilação | Conjunto de caracteres | Tipo | |
|--------------------|-----------------|------------------------|---------|---|
| db_camdep | utf8_general_ci | utf8 | Usuário | ⋮ |
| information_schema | utf8_general_ci | utf8 | Sistema | ⋮ |
| mysql | utf8_general_ci | utf8 | Sistema | ⋮ |
| performance_schema | utf8_general_ci | utf8 | Sistema | ⋮ |
| sys | utf8_general_ci | utf8 | Usuário | ⋮ |

Figura 4

Configuração do banco de dados no Google Cloud Platform

Versão do banco de dados
MySQL 5.7

Opções de configuração

☒ **Conectividade**
IP público ativado

☒ **Tipo de máquina e armazenamento**

Tipo de máquina ?

Para melhor desempenho, escolha um tipo de máquina com memória suficiente para manter sua maior tabela



db-n1-standard-1

vCPUs

Memória

1

3,75 GB

Alterar

Taxa de transferência da rede (MB/s) ?

250 de 2.000



Tipo de armazenamento

SSD

Capacidade de armazenamento ?

20 – 30720 GB. Maior capacidade melhora o desempenho, até os limites definidos pelo tipo de máquina. A capacidade não pode ser reduzida posteriormente.

20 GB

☐ Ativar aumento automático de armazenamento

Se ativado, sempre que você estiver próximo da capacidade, o armazenamento será aumentado de maneira incremental (e permanente). [Saiba mais](#)

Figura 5

5.4. LIMPEZA DOS DADOS

Limpeza de dados é um dos processos mais importante quando queremos gerar insights para tomar alguma decisão importante. Basicamente é a preparação dos dados para torna a análise mais clara e objetiva, removendo todo o conteúdo desnecessário.

Nesta etapa ficou decidido que todo o processo seria realizado através da linguagem de programação Python. Os dados extraídos através do site da Câmara dos Deputados vieram de forma bruta com informações incompletas e algumas desnecessárias que iriam atrapalhar, confundir e tornar a análise complexa.

Por isso, foi decidido remover diversas colunas e linhas que prejudicasse o entendimento dos dados.

```
##LIMPANDO O DATASET ANO-DESPESA

df_despesa = df_despesa.drop(columns=['cpf'])
df_despesa = df_despesa.drop(columns=['nuLegislatura'])
df_despesa = df_despesa.drop(columns=['nuCarteiraParlamentar'])
df_despesa = df_despesa.drop(columns=['codLegislatura'])
df_despesa = df_despesa.drop(columns=['urlDocumento'])
df_despesa = df_despesa.drop(columns=['ideDocumento'])
df_despesa = df_despesa.drop(columns=['numLote'])
df_despesa = df_despesa.drop(columns=['txtTrecho'])
df_despesa = df_despesa.drop(columns=['numParcela'])
df_despesa = df_despesa.drop(columns=['vlrGlosa'])
df_despesa = df_despesa.drop(columns=['vlrDocumento'])
df_despesa = df_despesa.drop(columns=['txtNumero'])
df_despesa = df_despesa.drop(columns=['indTipoDocumento'])
df_despesa = df_despesa.drop(columns=['numRessarcimento'])
df_despesa = df_despesa.drop(columns=['txtPassageiro'])
df_despesa = df_despesa.drop(columns=['vlrRestituicao'])
df_despesa = df_despesa.drop(columns=['txtDescricaoEspecificacao'])
df_despesa = df_despesa.drop(columns=['numSubCota'])
df_despesa = df_despesa.drop(columns=['numEspecificacaoSubCota'])
```

Figura 6

5.5. TRATAMENTO DOS DADOS

Após realizar a limpeza dos dados e entender quais informações não são mais necessárias. Foi identificado que haviam muitas datas com formato errado e que precisariam de correção, para que pudesse fazer em um processo mais a frente a conexão de uma tabela somente com datas que ligasse e identificasse os períodos de despesas de cada deputado, partido ou fornecedor.

O tratamento inicial foi realizado pela linguagem de programação Python, sendo melhor refinada no processo de criação dos dashboards pelo Power BI.

```

Limpeza Deputados

#Substituição da coluna URI
dfdeputados['ideCadastro'] = dfdeputados['uri'].replace('https://dadosabertos.camara.leg.br/api/v2/deputados/', '',
                                                       regex=True)

# Tratamento de Dataset substituindo nomes
dfdeputados['DataNascimento'] = dfdeputados['dataNascimento'].str.split('T').str[0]

#Tratando novamente a coluna URI
dfdeputados['uri'] = dfdeputados['uriTratada'].replace('https://dadosabertos.camara.leg.br/api/v2/deputados/',
                                                       '', regex=True)
#Organizando as colunas
dfdeputados = dfdeputados[['uri', 'idLegislaturaInicial', 'idLegislaturaFinal', 'nomeCivil', 'cpf', 'siglaSexo',
                           'DataNascimento', 'dataFalecimento', 'ufNascimento', 'municipioNascimento']]
#Tratando id Deputados
dfdeputados = dfdeputados.rename(columns={'uri': 'nuDeputadoId'})

[ ]

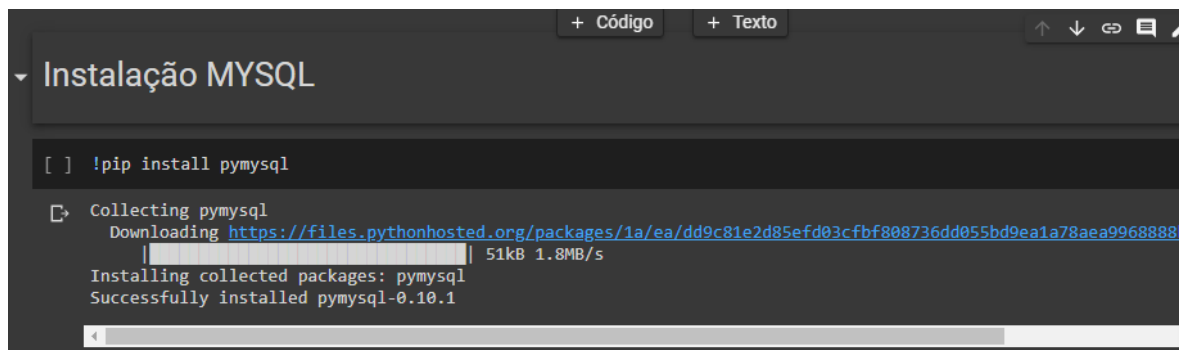
# Limpando as colunas
dfdeputados = dfdeputados.drop(columns=['uri'])
dfdeputados = dfdeputados.drop(columns=['cpf'])
dfdeputados = dfdeputados.drop(columns=['urlRedeSocial'])
dfdeputados = dfdeputados.drop(columns=['urlWebsite'])
dfdeputados = dfdeputados.drop(columns=['idLegislaturaInicial'])
dfdeputados = dfdeputados.drop(columns=['idLegislaturaFinal'])
dfdeputados = dfdeputados.drop(columns=['dataNascimento'])
dfdeputados = dfdeputados.drop(columns=['dataFalecimento'])
dfdeputados = dfdeputados.drop(columns=['ufNascimento'])

```

Figura 7

A imagem 7, apresenta como foi o tratamento do dataset deputados. Nele foram removidas colunas e tratado uma coluna do tipo data.

5.6. CONEXÃO COM BANCO DE DADOS



```

+ Código + Texto
Instalação MYSQL

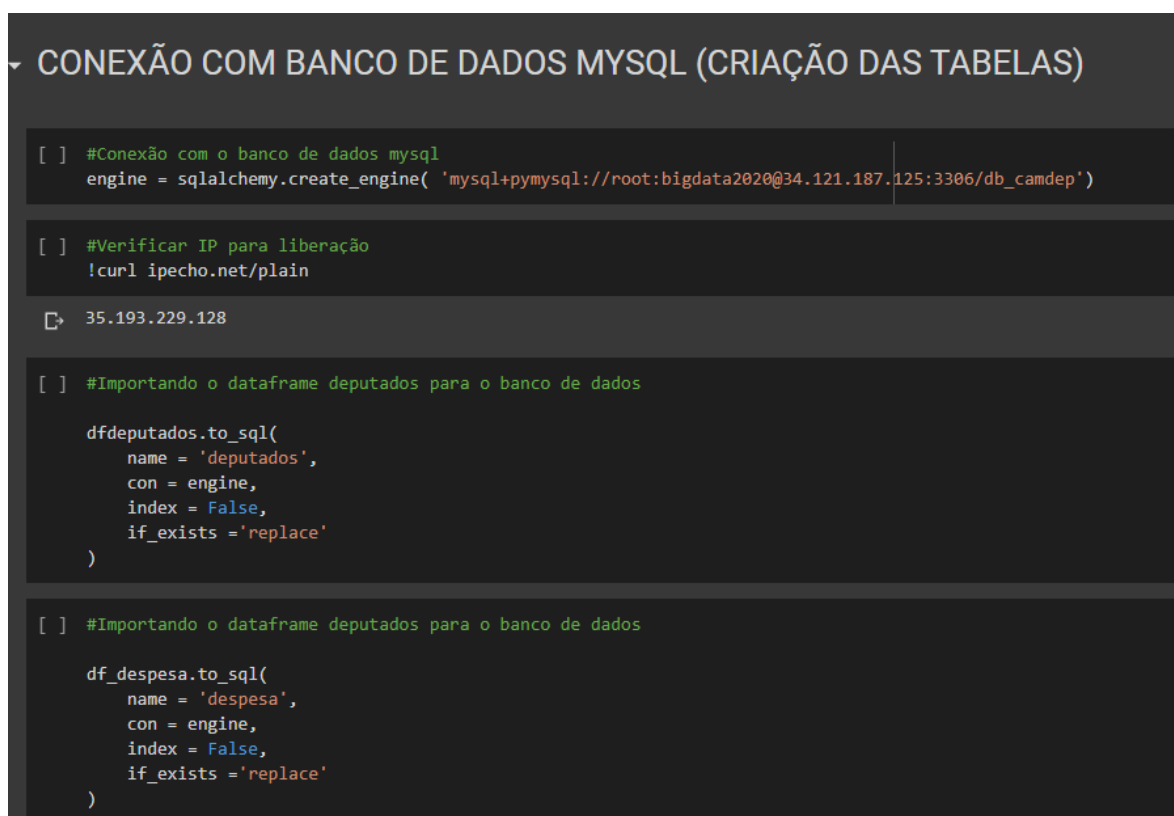
[ ] !pip install pymysql

Collecting pymysql
  Downloading https://files.pythonhosted.org/packages/1a/ea/dd9c81e2d85efd03cfbf808736dd055bd9ea1a78aea996888b...
    | 51kB 1.8MB/s
Installing collected packages: pymysql
Successfully installed pymysql-0.10.1

```

Figura 8

Para realizar a conexão com o banco de dados foi necessária a instalação da biblioteca “pymysql” dentro do Google Colaboratory.



```

CONEXÃO COM BANCO DE DADOS MYSQL (CRIAÇÃO DAS TABELAS)

[ ] #Conexão com o banco de dados mysql
engine = sqlalchemy.create_engine('mysql+pymysql://root:bigdata2020@34.121.187.125:3306/db_camdep')

[ ] #Verificar IP para liberação
!curl ipecho.net/plain

35.193.229.128

[ ] #Importando o dataframe deputados para o banco de dados

dfdeputados.to_sql(
    name = 'deputados',
    con = engine,
    index = False,
    if_exists = 'replace'
)

[ ] #Importando o dataframe deputados para o banco de dados

df_despesa.to_sql(
    name = 'despesa',
    con = engine,
    index = False,
    if_exists = 'replace'
)

```

Figura 9

Na Figura 9 acima através da biblioteca sqlalchemy foi criado uma variável que chamamos de “engine” que fará conexão diretamente com o banco de dados.

Após a conexão foi necessário descobrir qual o IP(Endereço de Rede) do computador virtual que foi criado no Google Colaboratory, para então libera-lo

dentro da plataforma do Google (Google Cloud Platform). Conforme Figura 10, abaixo.

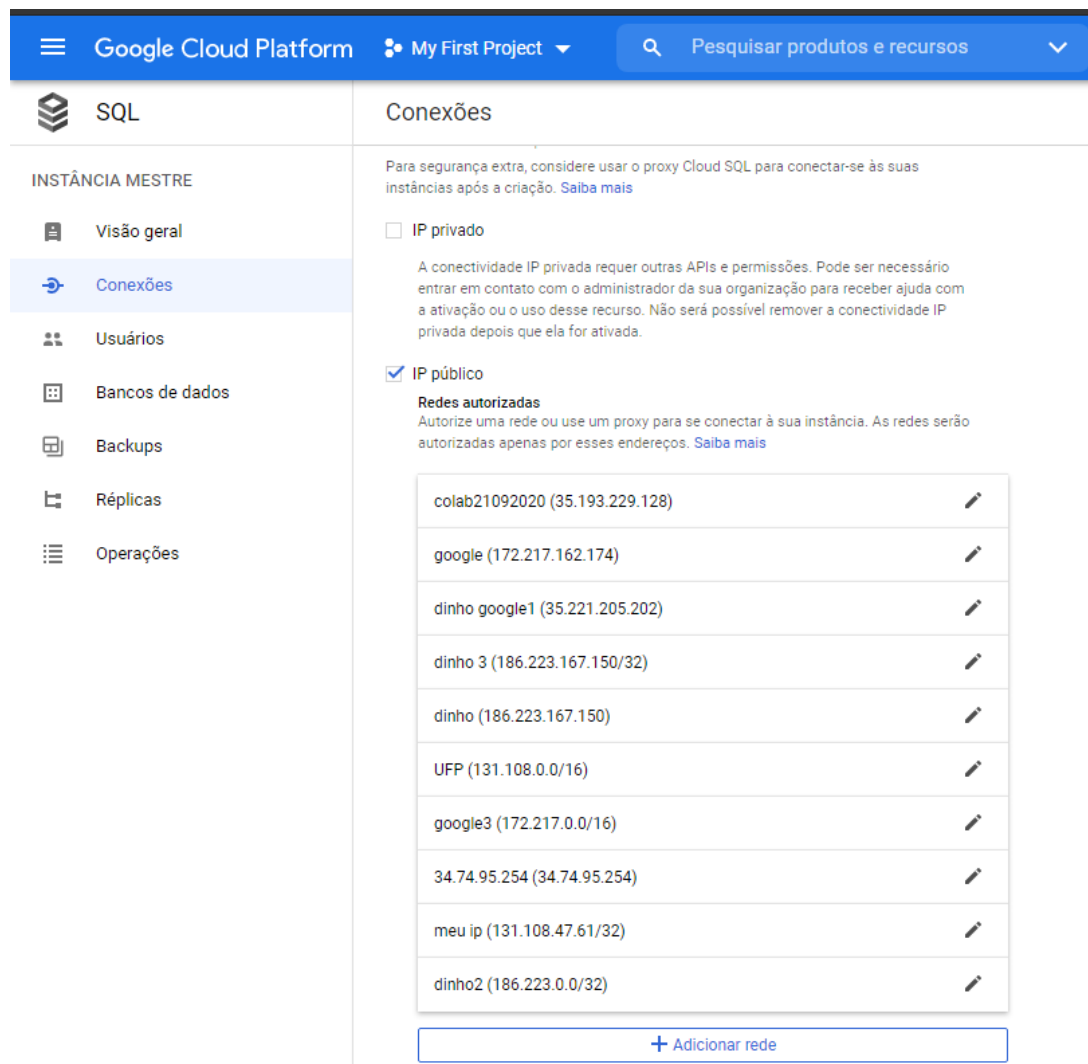


Figura 10

5.7. DEFINIÇÃO DE MÉTRICAS

O processo de criação das métricas foi o mais demorado, visto que, a grande quantidade de informações dificultou em parametrizar e estrutura o que viria a ser destacado.

Inicialmente foi estipulado o total de 9 datasets para serem limpos, analisados e formalização das métricas. Porém, com os problemas enfrentados foi decidido manter somente o de despesa e deputados.

Com a definição dos datasets que comporiam a análise, foi decidido que o métricas seriam criadas dentro do Power BI.

5.8. CONEXÃO COM O POWER BI

Configurações da fonte de dados

Gerenciar configurações para fontes de dados que você conectou usando o Power BI Desktop.

☒ Fontes de dados no arquivo atual ☐ Permissões globais

Pesquisar configurações da fonte de dados ⌵

| |
|--|
|  34.121.187.125;db_camdep |
|--|

Alterar Fonte... Editar Permissões... Limpar Permissões ▾

Fechar

Figura 11

A conexão com o Power BI foi criada a partir das informações configuradas no banco de dados dentro do Google Cloud Platform.

6. VISUALIZAÇÃO DOS DADOS E ANÁLISE DOS RESULTADOS

6.1. ANÁLISE DOS DATASETS

Depois de todas as explorações, tratamento, limpeza dos dados e importação dos dados, foi necessário um ambiente para criação das métricas e montagem dos dashboards para realizar as análises e realizar as conclusões.

Dataset Despesa

| Coluna | Tipo de Dados |
|-------------------|---------------|
| txNomeParlamentar | string |
| ideCadastro | float |
| sgUF | string |
| sgPartido | string |
| TxtDescricao | string |
| txtFornecedor | string |
| txtCNPJCPF | string |
| datEmissão | string |
| vlrLiquido | float |
| numMes | int |
| numAno | int |
| nuDeputadold | int |

Dataset Deputado

| Coluna | Tipo de Dados |
|---------------------|---------------|
| nome | string |
| nomeCivil | string |
| siglaSexo | string |
| municipioNascimento | string |
| ideCadastro | string |
| DataNascimento | string |

Os datasets acima foram importados para dentro do Power BI.

6.2. MODELO DE DADOS

Foi criado um modelo de dados conforme demonstrado na Figura 12, abaixo:

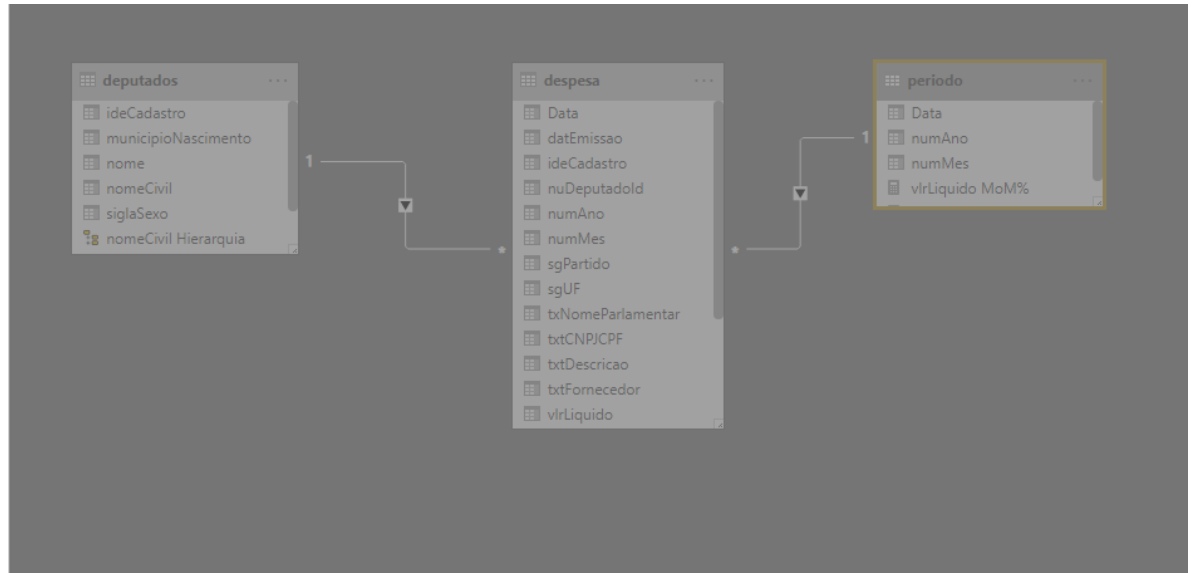


Figura 12

Houve a necessidade de criar uma tabela chamada “período” para interligar e ajudar na análise temporal dos dados. Demonstrado na Figura 13, abaixo.

| numMes | numAno | Data |
|--------|--------|--------------------------------------|
| 7 | 2012 | domingo, 1 de julho de 2012 |
| 6 | 2012 | sexta-feira, 1 de junho de 2012 |
| 8 | 2012 | quarta-feira, 1 de agosto de 2012 |
| 12 | 2012 | sábado, 1 de dezembro de 2012 |
| 5 | 2012 | terça-feira, 1 de maio de 2012 |
| 10 | 2012 | segunda-feira, 1 de outubro de 2012 |
| 11 | 2012 | quinta-feira, 1 de novembro de 2012 |
| 9 | 2012 | sábado, 1 de setembro de 2012 |
| 3 | 2012 | quinta-feira, 1 de março de 2012 |
| 4 | 2012 | domingo, 1 de abril de 2012 |
| 1 | 2012 | domingo, 1 de janeiro de 2012 |
| 2 | 2012 | quarta-feira, 1 de fevereiro de 2012 |
| 2 | 2013 | sexta-feira, 1 de fevereiro de 2013 |
| 3 | 2013 | sexta-feira, 1 de março de 2013 |

Figura 13

6.3. VISUALIZAÇÃO DOS DADOS

Foram elaborados 5 Dashboards para compor as análises de dados:

6.3.1. DESPESA DE DEPUTADOS

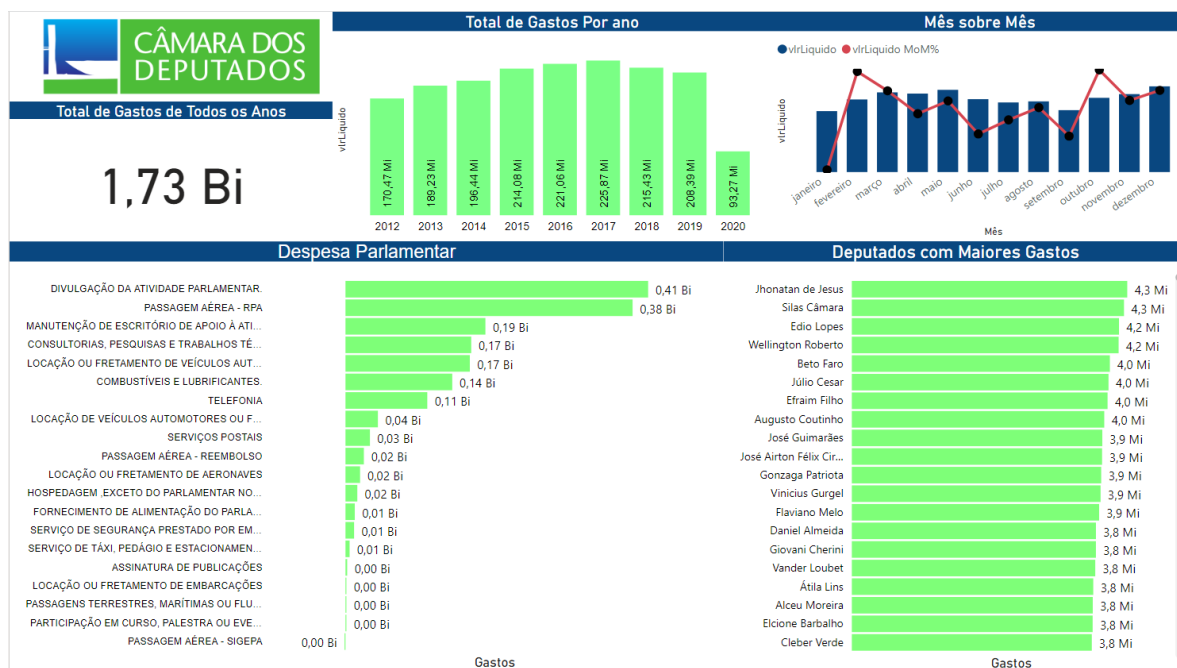


Figura 14

O Dashboard da Figura 14, aborda os gastos dos deputados dos anos de 2012 até 2020.

Analisando os dados apresentando podemos identificar que os maiores gastos são com divulgação da atividade parlamentar e com passagens aéreas.

É possível identificar também os Deputados que mais gastaram durante todos os períodos analisados.

O total de gastos chega a quase o valor de 2 Bilhões de reais.

6.3.2. DESPESA PARTIDOS

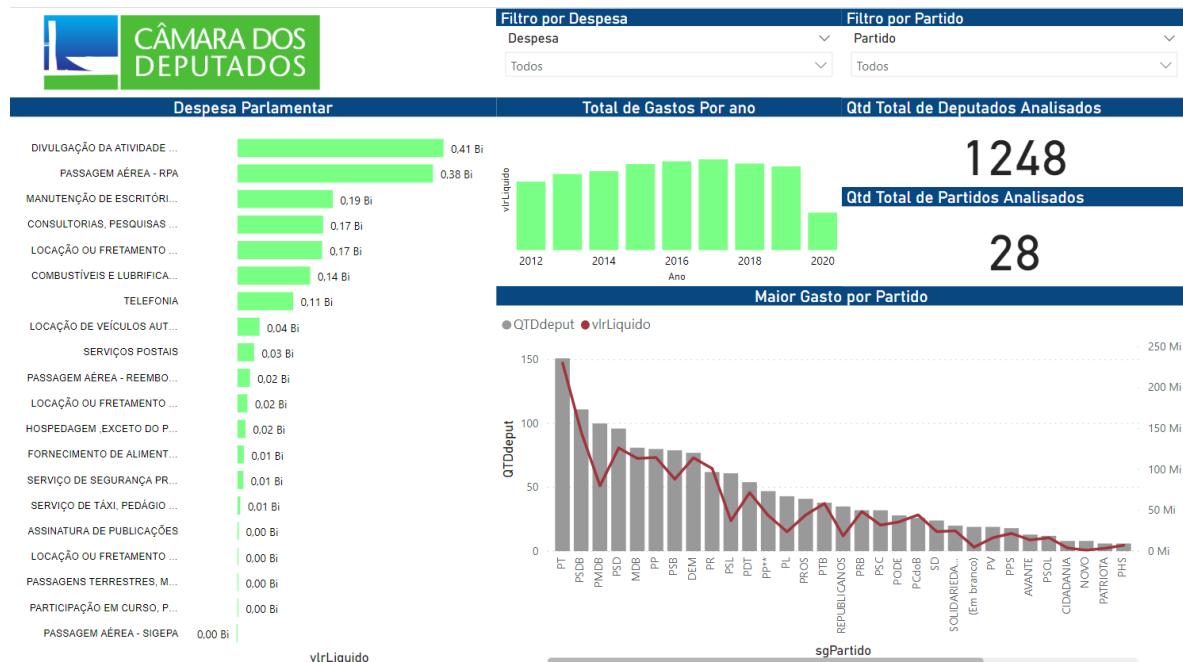
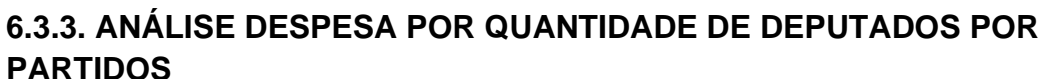


Figura 15

Através do Dashboard da Figura 15, podemos analisar os gastos de todos os partidos. O PT foi o que mais gastou durante todos os períodos.

Foram criados dois filtros um por Despesa e o outro por Partido, com o intuito de realizar um detalhamento maior dos gastos públicos.

Ao todo foram analisados 28 partidos e 1248 deputados.



6.3.4. DESPESA FORNECEDOR

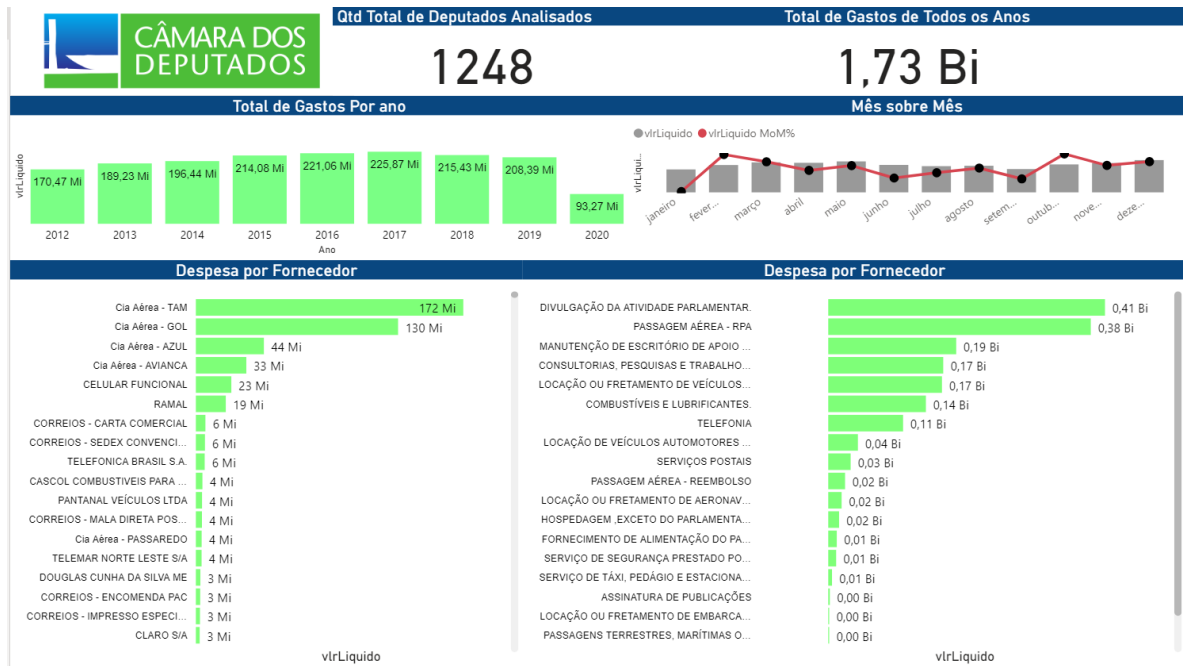


Figura 17

No Dashboard Despesa Fornecedor, demonstra o gasto obtido ao longo dos anos, o tipo de gasto e qual fornecedor foi beneficiado.

6.3.5. PROPORÇÃO GÊNERO POR DEPUTADOS

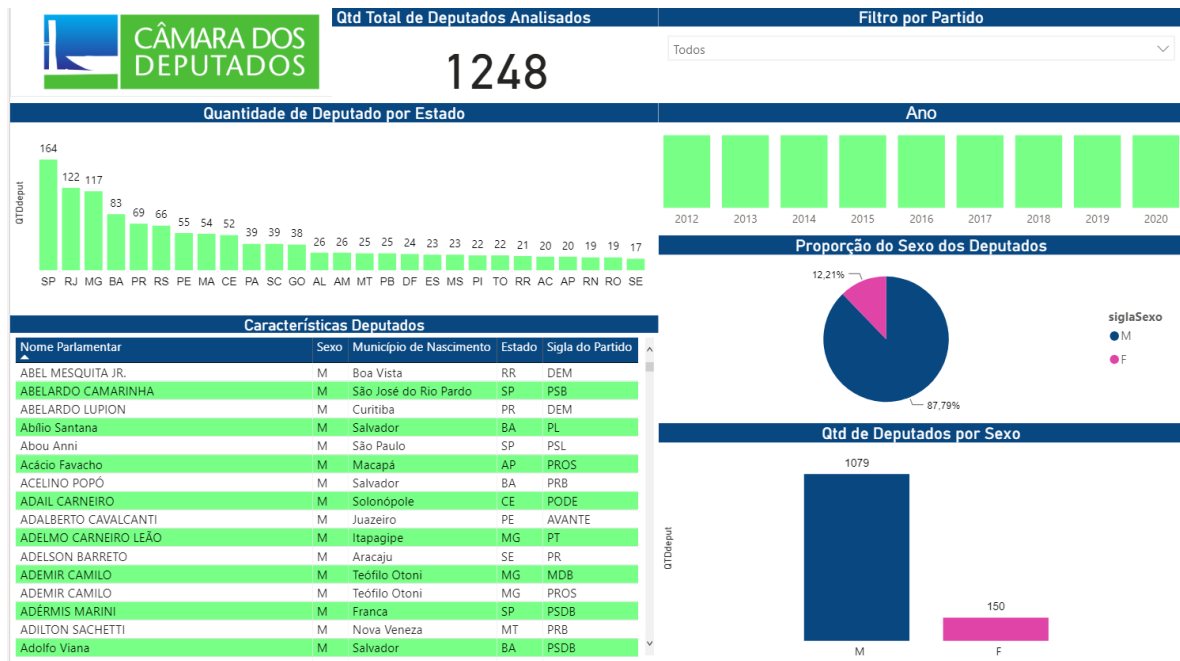


Figura 18

Neste Dashboard demonstra a proporção dos gêneros masculino e feminino. Nota-se que existem mais homens ao longo de todos os anos do que deputadas mulheres.

Outro ponto é a quantidade de gêneros em cada Estado, onde os homens são predominantes.

Pelo Filtro por Partido é possível analisar o quantitativo proporcional dos gêneros masculino e feminino.

6.4. MÉTRICAS

Foram criadas diversas métricas no Power Bi para compor a análise de todos os dados. São elas:

```
Filtro Despesa Ano Deputado =
CALCULATE(DISTINCTCOUNT(deputados[ideCadastro]);FILTER(periodo;periodo[numAno]);FILTER(despesa;despesa[ideCadastro]))
```

Filtro indica a despesa do deputado no ano.

```
Filtro Qtd deputado Valor = CALCULATE(SUM(despesa[vlrLiquido]);FILTER(periodo;
periodo[numAno]))
Gasto por Combustivel =
CALCULATE(SUM(despesa[vlrLiquido]);FILTER(despesa;despesa[txtDescricao] = "COMBUSTÍVEIS E
LUBRIFICANTES."))
```

Filtro indica o gasto de combustíveis e lubrificantes por ano do deputado.

```
QTDdeput = DISTINCTCOUNT(despesa[txNomeParlamentar])
```

Filtro indica a quantidade de deputados.

```
QTDpart = CALCULATE(DISTINCTCOUNT(despesa[sgUF]))
```

Filtro indica a quantidade de partidos.

```
Total Ano partido Deputado = DIVIDE(CALCULATE(SUM(despesa[vlrLiquido]);FILTER(periodo;
periodo[numAno]));CALCULATE(DISTINCTCOUNT(deputados[ideCadastro]);FILTER(periodo;periodo[
numAno]);FILTER(despesa;despesa[ideCadastro]));BLANK())
```

Filtro indica o valor da despesa com um outro filtro por período do ano do pela quantidade de deputados.

```
vlrLiquido MoM% =
IF(
    ISFILTERED('periodo'[Data]);
    ERROR("Medidas rápidas de inteligência de tempo somente podem ser agrupadas ou
filtradas pela hierarquia de data fornecida pelo Power BI ou pela coluna de data
primária.");
    VAR __PREV_MONTH =
        CALCULATE(
            SUM('despesa'[vlrLiquido]);
            DATEADD('periodo'[Data].[Date]; -1; MONTH)
        )
    RETURN
        DIVIDE(SUM('despesa'[vlrLiquido]) - __PREV_MONTH; __PREV_MONTH)
)
```

Filtro indica a quantidade de valor mês a mês da despesa.


```

vlrLiquido YoY% =
IF(
    ISFILTERED('periodo'[Data]);
    ERROR("Medidas rápidas de inteligência de tempo somente podem ser agrupadas ou
filtradas pela hierarquia de data fornecida pelo Power BI ou pela coluna de data
primária.");
    VAR __PREV_YEAR =
        CALCULATE(
            SUM('despesa'[vlrLiquido]);
            DATEADD('periodo'[Data].[Date]; -1; YEAR)
        )
    RETURN
        DIVIDE(SUM('despesa'[vlrLiquido]) - __PREV_YEAR; __PREV_YEAR)
)

```

Filtro indica a quantidade de valor ano a ano da despesa.

Além das criadas pelo Power Bi, foram utilizadas outras para análise:

- Total de Gasto de todos os anos;
- Gasto por despesa parlamentar;
- Ranking dos deputados com maiores gastos;
- Gastos por ano;
- Gastos mês a mês;
- Total de Deputados;
- Total de Partidos;
- Maior Gasto por partido;
- Total de Gasto por partido ao longo dos anos;
- Análise dos Deputados com maiores gastos proporcional a quantidade de membros no Partido em comparação com outros Partidos;
- Despesas por pelo tipo de gasto;
- Despesas por Fornecedor;
- Proporção de deputados por gênero;
- Proporção de deputados por estados em relação ao seu gênero;
- Proporção de deputados por gênero por partido;

Entre outras análises que são possíveis de se realizar ao cruzar os dados, existem inúmeras possibilidades de se conseguir insights e gerar muitas outras informações.

6.5. ANÁLISE DOS RESULTADOS

Os resultados apresentados nos dashboards concluem que os maiores gastos são por campanhas de divulgação de atividade parlamentar e passagens aéreas.

A quantidade de dinheiro público usado para diversos fins chega na casa dos milhões. Foi observado que alguns deputados e partidos estão sempre em destaque nos diversos filtros aplicados durante a análise.

Ainda existe uma diferença desigual a proporção de deputados e deputadas que são eleitos pelo povo que em sua grande maioria ainda é composta por homens.

Nas análises feitas pode perceber que alguns partidos que possuem uma quantidade menor de membros gastam quase na mesma proporção de outros que possuem quase que dez vezes o número de ocupantes.

7. CONCLUSÃO

O presente projeto se propôs a analisar os gastos da Câmara dos Deputados com base em suas despesas entre os anos de 2012 até 2020.

Durante o período analisado, foi possível entender melhor as despesas de cada partido e deputado ao longo dos seus mandatos. Muitos dos gastos vistos durante as análises remetiam a passagens aéreas ou divulgações da atividade do parlamentar. Contudo, uma das despesas que chamaram atenção foi o serviço postal, ainda sendo usado por grande parte dos deputados. Onde hoje existem grandes facilidades por meio do uso de tecnologia sendo estranho que este serviço ainda seja usado amplamente. Os gastos com combustíveis, locação de fretamento de veículos, manutenção de escritório e telefonia ainda são também os grandes vilões da máquina de gastos públicos.

Nota-se que houve um grande crescimento dos gastos entre o ano de 2012 e 2017, vindo a começar a despencar entre 2018 e 2019 depois que uma nova gestão de governo começou. Não se pode concluir que foi devido a mudança de gestão, ou implementação de fiscalizações mais rígidas. Porém, essa observação torna-se interessante se formos analisar por esse ponto de vista.

A partir dos Dashboards foi possível realizar várias possibilidades de análises que evidenciaram o uso indevido do dinheiro público. Outro ponto que chamou a atenção é quantidade mínima de integrantes do sexo feminino, cujo, poderiam ter uma maior participação no pleito, promovendo melhorias que se fazem necessárias.

Dada à importância do assunto, torna-se necessário uma maior fiscalização das despesas que sejam capazes de averiguar irregularidades e tomar medidas cabíveis. Podendo economizar não só o tempo como recursos que são necessários para outros fins de suma importância para população.

8. REFERÊNCIAS

Supremo derruba regra que mudou número de deputados de 13 estados:

<http://g1.globo.com/politica/eleicoes/2014/noticia/2014/06/supremo-derruba-regra-que-alterou-bancadas-de-13-estados.html>, acessado em 23/09/20, p. 10.

A HISTÓRIA DA CÂMARA DOS DEPUTADOS: <https://www2.camara.leg.br/a-camara/conheca/historia/oimperio.html>

acessado em 23/09/20, p. 10.

LEGISLAÇÃO: <https://www2.camara.leg.br/legin/fed/consti/1824-1899/constituicao-35041-25-marco-1824-532540-norma-pl.html> acessado em 23/09/20, p. 10.

Constituição brasileira de 1824:

https://pt.wikipedia.org/wiki/Constitui%C3%A7%C3%A3o_brasileira_de_1824 acessado em 23/09/20, p. 10.

Sobre o congresso nacional:

<https://www.congressonacional.leg.br/institucional/sobre-o-congresso-nacional> acessado em 23/09/20, p. 10.

Câmara dos deputados do Brasil

https://pt.wikipedia.org/wiki/C%C3%A2mara_dos_Deputados_do_Brasil acessado em 23/09/20, p. 10.

História e arquivo - CONHEÇA A HISTÓRIA DA CÂMARA:

<https://www.camara.leg.br/historia-e-arquivo/> acessado em 23/09/20, p. 10.

ACESSO À INFORMAÇÃO: Cota para o exercício da atividade parlamentar:

https://www2.camara.leg.br/transparencia/acesso-a-informacao/copy_of_perguntas-frequentes/cota-para-o-exercicio-da-atividade-parlamentar acessado em 23/09/20, p. 10.

Deputados e senadores torram R\$11 milhões em propaganda pessoal:

<https://diariodopoder.com.br/politica/deputados-e-senadores-torram-r11-milhoes-em-propaganda-pessoal> acessado em 23/09/20, p. 11.

Deputados federais baianos gastam na pandemia combustível suficiente para 23 voltas na terra, diz pesquisa:

<https://infosaj.com.br/deputados-baianos-gastam-na-pandemia-combustivel-suficiente-para-23-voltas-na-terra/> acessado em 23/09/20, p. 11.

Durante a pandemia, Câmara dos Deputados pagou R\$ 11,8 milhões em passagens:

<https://radaramazonico.com.br/durante-a-pandemia-camara-dos-deputados-pagou-r-118-milhoes-em-passagens/> acessado em 23/09/20, p. 11.

Dados Abertos

<https://dadosabertos.camara.leg.br/swagger/api.html#staticfile> acessado em 23/09/20, p. 12.

Documentação Python:

<https://docs.python.org/pt-br/> acessado em 23/09/20, p. 12.

SQLAlchemy 1.3 Documentation

<https://docs.sqlalchemy.org/en/13/> acessado em 23/09/20, p. 12.

O que é big data

<https://canaltech.com.br/big-data/o-que-e-big-data/> acessado em 23/09/20, p. 13.

INOVAÇÃO NOS NEGÓCIOS POR MEIO DA ANÁLISE DE BIG DATA

<https://www.redalyc.org/jatsRepo/5536/553658821001/html/index.html> acessado em 23/09/20, p. 13.

Inovação na inteligência analítica por meio do Big Data: características de diferenciação da abordagem tradicional

<http://www.pos.cps.sp.gov.br/files/artigo/file/488/839f2e27fa0fa7f5776622a62a48a776.pdf> acessado em 23/09/20, p. 13.

BIG DATA <https://webgui.com.br/big-data/> acessado em 23/09/20, p. 13.

Power BI Overview

<https://docs.microsoft.com/pt-br/power-bi/fundamentals/power-bi-overview> acessado em 23/09/20, p. 14.

BIG DATA PARA LEIGOS

https://books.google.com.br/books/about/Big_Data_Para_Leigos.html?id=j8hYCwAAQBAJ&printsec=frontcover&source=kp_read_button&redir_esc=y#v=onepage&q&f=false, acessado em 23/09/20, p13.

Big Data O Futuro dos Dados e Aplicações

https://books.google.com.br/books/about/Big_Data_O_Futuro_dos_Dados_e_Aplica%C3%A7%C3%B5es.html?id=2LdiDwAAQBAJ&printsec=frontcover&source=kp_read_button&redir_esc=y#v=onepage&q&f=false, acessado em 23/09/20, p 13