Projeto Integrador em Banco de Dados e Big Data II

Monniky S R Pereira

IESB - Banco de Dados e Armazenamento Big Data

Sumário

1. Introdução

Os projetos em Big Data vão além da simples análise de dados e apresentação de gráficos. A coleta, armazenamento e processamento de dados são etapas cruciais para a obtenção de informações relevantes. Este projeto tem como objetivo analisar os gastos públicos a partir de dados da Câmara dos Deputados no ano de 2022, utilizando algoritmos de machine learning para identificar padrões e insights significativos. O projeto busca não apenas relatar os dados, mas também contribuir para a discussão sobre a eficácia e a moralidade das despesas públicas.

2. Motivação

A análise de dados públicos é fundamental para auxiliar na fiscalização e na transparência dos gastos governamentais. Neste projeto, busca-se entender os fatores que influenciam os gastos dos deputados, contribuindo para a discussão sobre a eficácia e a moralidade das despesas públicas. A transparência nos gastos públicos é vital para a confiança da população nas instituições governamentais, e a análise de dados pode influenciar decisões políticas e a alocação de recursos.

3. Script e Banco de Dados

3.1 Estrutura do Banco de Dados

O banco de dados foi criado utilizando PostgreSQL, contendo duas tabelas principais: **deputados** e **despesas**. A tabela "deputados" armazena informações sobre os parlamentares, enquanto a tabela "despesas" registra as despesas realizadas por eles. A estrutura das tabelas é apresentada a seguir:

Tabela: deputados

- **id** (SERIAL): Identificador único do deputado.
- **uri** (VARCHAR): URI do deputado.
- **nome** (VARCHAR): Nome do deputado.
- **siglaPartido** (VARCHAR): Sigla do partido do deputado.
- **uriPartido** (VARCHAR): URI do partido do deputado.
- **siglaUf** (VARCHAR): Sigla da unidade federativa.
- **idLegislatura** (INTEGER): Identificador da legislatura.
- **urlFoto** (TEXT): URL da foto do deputado.
- **email** (VARCHAR): Email do deputado.
- **gabinete** (VARCHAR): Gabinete do deputado.

Tabela: despesas

- **id** (SERIAL): Identificador único da despesa.
- **deputado_id** (INTEGER, FOREIGN KEY): Chave estrangeira que referencia o id na tabela deputados.
- **txNomeParlamentar** (VARCHAR): Nome parlamentar.
- **cpf** (VARCHAR): CPF do deputado.
- **ideCadastro** (INTEGER): Identificação do cadastro.
- **nuCarteiraParlamentar** (VARCHAR): Número da carteira parlamentar.
- **nuLegislatura** (INTEGER): Número da legislatura.
- **sgUF** (VARCHAR): Sigla da unidade federativa.
- **sgPartido** (VARCHAR): Sigla do partido.
- **codLegislatura** (INTEGER): Código da legislatura.
- **numSubCota** (INTEGER): Número da subcota.
- **txtDescricao** (TEXT): Descrição da despesa.
- **numEspecificacaoSubCota** (INTEGER): Número da especificação da subcota.
- **txtDescricaoEspecificacao** (TEXT): Descrição da especificação.
- **txtFornecedor** (TEXT): Fornecedor.
- **txtCNPJCPF** (VARCHAR): CNPJ/CPF do fornecedor.
- **txtNumero** (VARCHAR): Número do documento.
- **indTipoDocumento** (VARCHAR): Tipo do documento.
- **datEmissao** (DATE): Data de emissão do documento.
- **vlrDocumento** (NUMERIC): Valor do documento.
- **vlrGlosa** (NUMERIC): Valor da glosa.
- **vlrLiquido** (NUMERIC): Valor líquido.
- **numMes** (INTEGER): Número do mês.
- **numAno** (INTEGER): Número do ano.
- **numParcela** (INTEGER): Número da parcela.
- **txtPassageiro** (TEXT): Nome do passageiro (se for uma despesa de viagem).
- **txtTrecho** (TEXT): Trecho da viagem (se for uma despesa de viagem).
- **numLote** (INTEGER): Número do lote.
- **numRessarcimento** (INTEGER): Número do ressarcimento.
- **datPagamentoRestituicao** (DATE): Data do pagamento da restituição.
- **vlrRestituicao** (NUMERIC): Valor da restituição.
- **nuDeputadoId** (INTEGER): Identificador do deputado.
- **ideDocumento** (VARCHAR): Identificador do documento.
- **urlDocumento** (TEXT): URL do documento.

3.2 Diagrama de Entidade-Relacionamento (ER)

O diagrama de Entidade-Relacionamento (ER) do banco de dados foi desenvolvido para representar as relações entre as principais tabelas: “**deputados”** e “**despesas”**. Esse diagrama facilita a compreensão da estrutura e das interações entre as entidades, permitindo uma análise eficaz dos dados coletados. Abaixo estão as principais relações definidas:

1. **Relação 1**

   (deputados para despesas)

   - O campo “**id”** na tabela “**deputados”** é a chave primária (PK) que estabelece uma relação de um para muitos com o campo “**deputado_id”** na tabela “**despesas”**. Esta relação indica que cada deputado pode estar associado a várias despesas, mas cada despesa pertence a apenas um deputado.

2. **Relação 1**

   (partido para despesas)

   - O campo “**siglaPartido”** na tabela “**deputados”** é uma chave estrangeira (FK) que relaciona um partido com múltiplos deputados e, consequentemente, com diversas despesas registradas sob essa afiliação partidária. Esse relacionamento permite identificar os gastos agrupados por partido.

3. **Outros Atributos Importantes e Relações Secundárias:**
    - **“txNomeParlamentar”** e “**siglaUF”**: Estes campos auxiliam na segmentação e análise dos dados por parlamentar e unidade federativa, sendo importantes para a geração de relatórios que mostram os principais gastos por UF e os deputados com maiores despesas.

O diagrama ER, com essas relações bem definidas, permite não apenas a visualização da estrutura do banco de dados, mas também a criação de consultas mais eficientes para explorar e identificar padrões nos gastos públicos por deputado, partido e UF.

3.3 Script de Obtenção dos Dados

O script para obter os dados dos deputados foi desenvolvido utilizando a API da Câmara dos Deputados. O código realiza a requisição dos dados e os armazena na tabela correspondente do banco de dados. O script também inclui funções para carregar despesas a partir de um arquivo CSV e para visualizar os dados armazenados.

Web Scraping dos Deputados:

```
def web_scraping_deputados():
    """
    Realiza o web scraping dos dados dos deputados a partir da API da Câmara dos Deputados.

    Returns:
        DataFrame: Um DataFrame contendo os dados dos deputados, ou None se ocorrer um erro.
    """
    try:
        url = "https://dadosabertos.camara.leg.br/api/v2/deputados"
        response = requests.get(url)

        if response.status_code == 200:
            deputados_data = response.json()
            deputados = deputados_data['dados']
            df_deputados = pd.DataFrame(deputados)
            return df_deputados
        else:
            print(f"Erro ao acessar a API: {response.status_code}")
            return None
    except Exception as e:
        print(f"Ocorreu um erro durante o scraping: {e}")
        return None
```

Despesas do CSV:

```python
def carregar_despesas_de_csv(caminho_csv):
    """
    Carrega os dados de despesas a partir de um arquivo CSV.
    Args:
        caminho_csv (str): O caminho do arquivo CSV.
    Returns:
        DataFrame: Um DataFrame contendo os dados das despesas, ou None se ocorrer um erro.
    """
    try:
        # Especifica o delimitador correto e trata as aspas
        df = pd.read_csv(caminho_csv, sep=';', quotechar='"')
        return df
    except Exception as e:
        print(f"Ocorreu um erro ao carregar o CSV: {e}")
        return None
```

4. Relatório Analítico

4.1 Análise dos Gastos Públicos

A análise dos dados permitiu identificar quais deputados gastaram mais no ano de 2022 e quais foram os principais itens de gastos. Também foram levantadas suspeitas sobre gastos excessivos, utilizando gráficos de barras e tabelas para facilitar a visualização. Os dados foram organizados em tabelas que detalham os gastos por categoria e por deputado

4.2 Dashboard no Power BI

Foi criado um dashboard no Power BI contendo diversos gráficos, incluindo:

- ➔ Segmentação de dados por partido
- ➔ Análise de gastos por deputado
- ➔ Distribuição geográfica dos maiores gastos
- ➔ Itens com maiores gastos
- ➔ Análise de suspeitas de gastos

Os gráficos proporcionam uma visão clara e intuitiva sobre os gastos públicos, facilitando a identificação de padrões e anomalias.

4.3 Dashboard de Análise de Gastos Públicos: Visão Geral do Dashboard

Este relatório apresenta uma visão geral do dashboard de análise dos gastos públicos, desenvolvido para fornecer insights sobre a utilização dos recursos pelos deputados. O dashboard está organizado em diferentes seções e gráficos, que permitem uma análise detalhada e segmentada dos gastos, respondendo a todas as questões solicitadas no projeto.

Na página inicial do dashboard, temos as seguintes áreas de visualização:

1.**Partidos**: Segmentação de dados por partido, permitindo identificar os gastos totais de cada legenda e realizar comparações entre os partidos.

2.**Distribuição Geográfica dos Maiores Gastos Públicos**: Mapa que mostra a localização dos maiores gastos públicos por unidade federativa, com o tamanho das bolhas representando o valor gasto. Essa visualização facilita a identificação de regiões com altos custos.

3.**Gastos Totais**: Exibição do valor total dos gastos analisados, oferecendo uma visão ampla da soma dos valores alocados.

4. **Análise de Suspeitas de Gastos**: Gráfico que destaca gastos suspeitos, com um limite de valor estabelecido para facilitar a identificação de possíveis irregularidades.

5. **Parlamentar**: Segmentação de dados por parlamentar, que permite uma análise individual dos gastos de cada deputado.

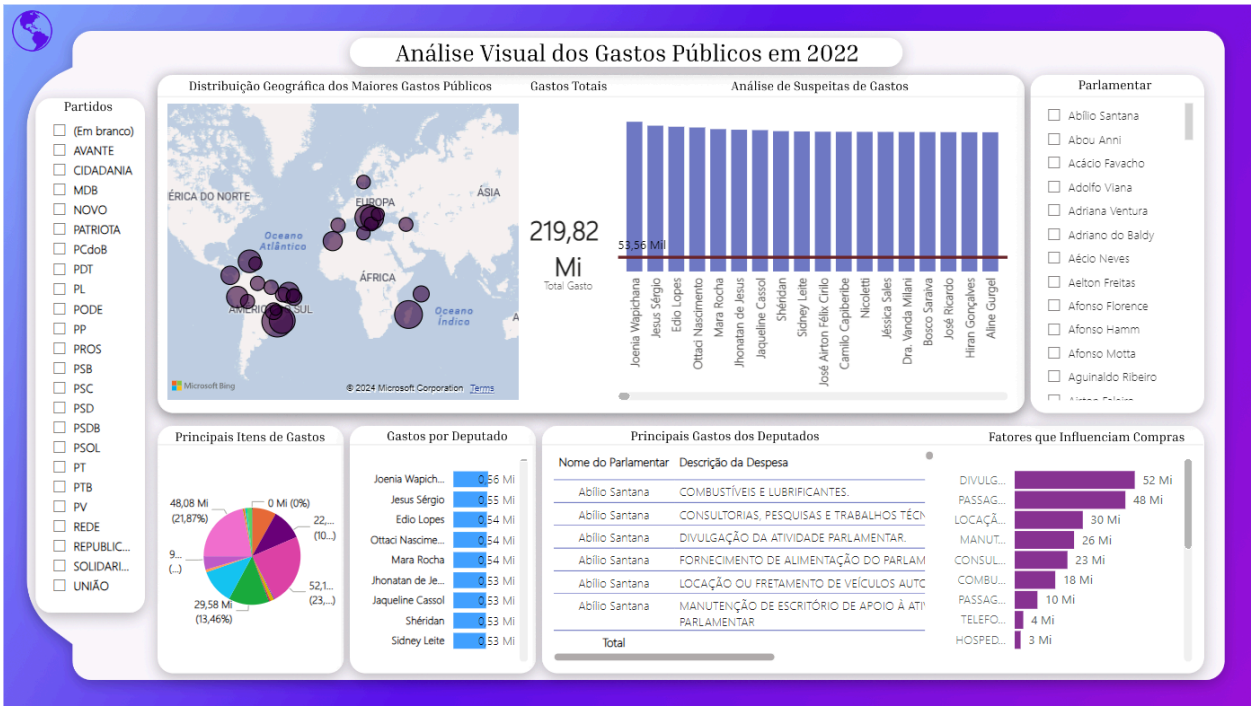
6.**Principais Itens de Gastos**: Gráfico de pizza que representa os itens que receberam os maiores valores de gasto, trazendo insights sobre onde o orçamento foi mais alocado.

7.**Gastos por Deputado**: Gráfico de barras que exibe o valor gasto por cada deputado, oferecendo uma visão comparativa entre os parlamentares.

8.**Principais Gastos dos Deputados**: Tabela com o nome do parlamentar, descrição do gasto, valor e data de emissão, detalhando os maiores gastos de forma acessível.

9.**Fatores que Influenciam Compras**: Gráfico de barras que explora os fatores e tipos de itens que influenciam os gastos dos parlamentares.

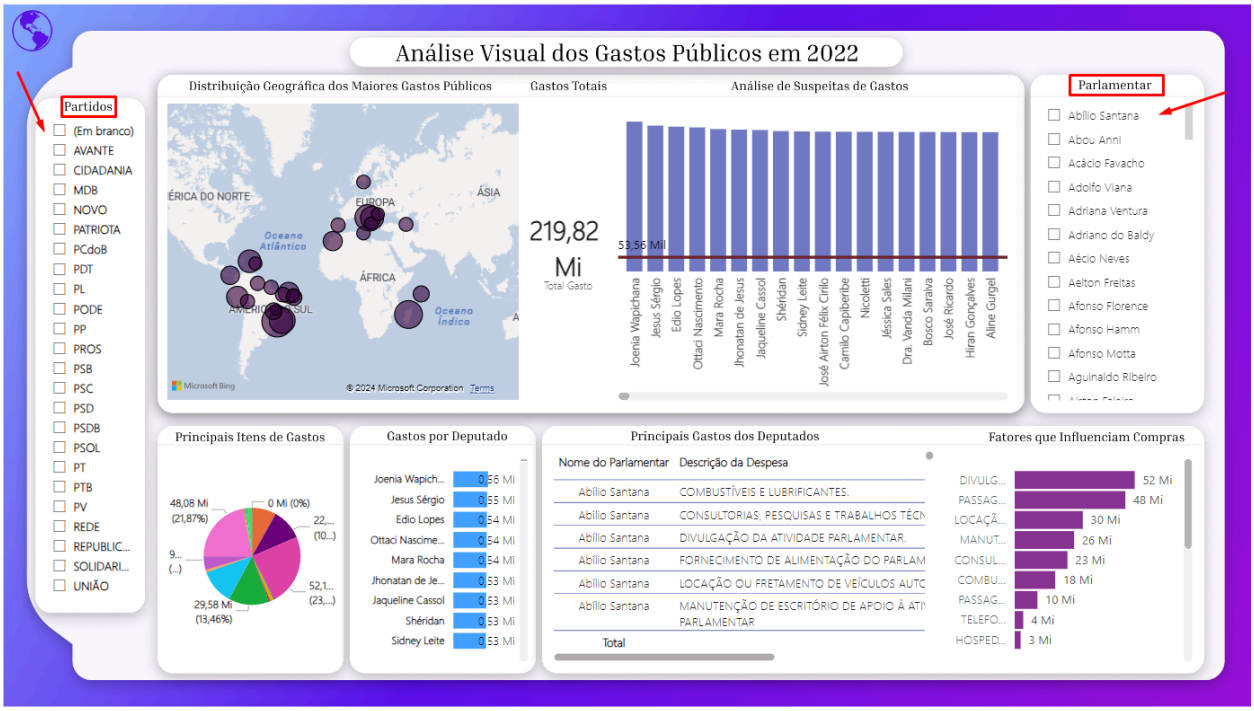
Com essa estrutura, o dashboard possibilita uma análise robusta e multifacetada dos dados, proporcionando informações essenciais para a compreensão do uso dos recursos públicos.



4.4 Dashboard de Filtros Selecionáveis

Este dashboard oferece filtros interativos para a análise personalizada dos dados. Os filtros de **Partidos** e **Parlamentares** permitem que o usuário selecione múltiplas opções simultaneamente, sem limite de quantidade. Essa funcionalidade possibilita uma segmentação detalhada, permitindo visualizar dados específicos de partidos ou parlamentares de interesse.

Após a seleção dos filtros, todos os gráficos e tabelas do dashboard são automaticamente atualizados para refletir as escolhas feitas. Esse recurso dinâmico facilita a análise comparativa e a identificação de padrões ou discrepâncias nos gastos públicos, proporcionando uma visão mais ajustada às necessidades da análise.



4.5 Dashboards com Informações Importantes

Além dos filtros interativos, nosso dashboard apresenta visualizações detalhadas que facilitam a análise e investigação dos dados de gastos públicos. A seguir, uma visão geral dos principais componentes:

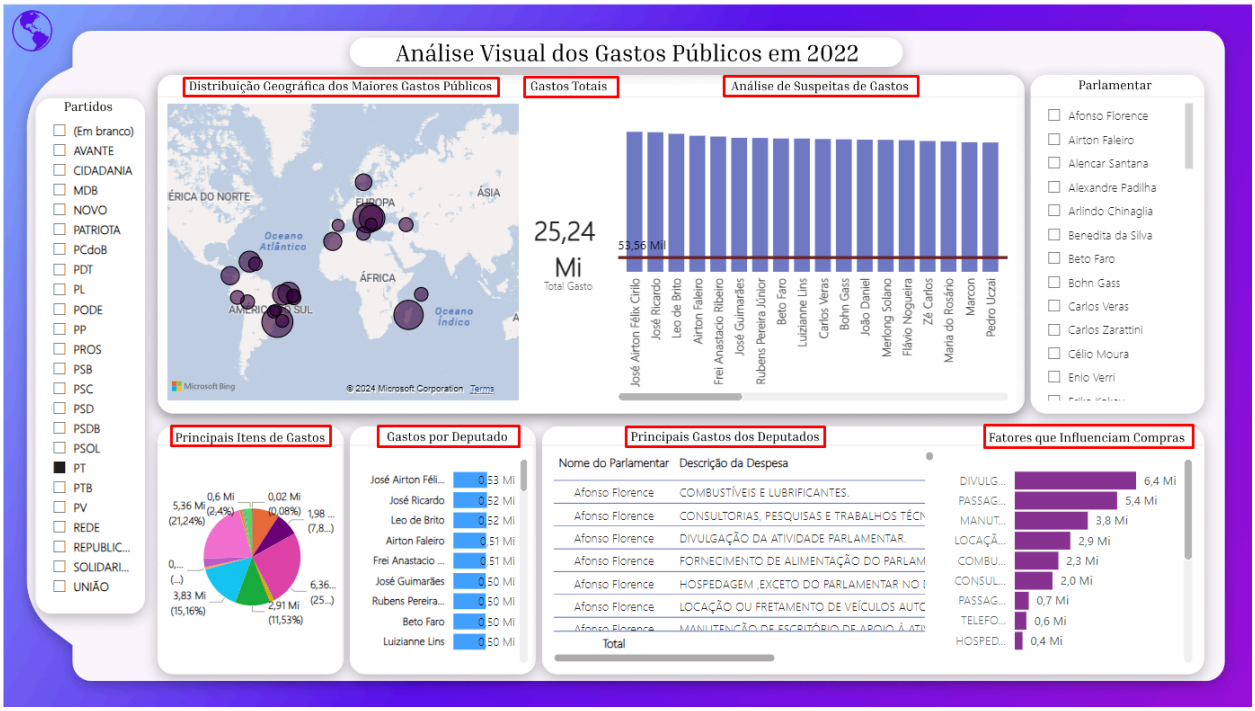
- ➔ **Distribuição Geográfica dos Maiores Gastos Públicos**: Este gráfico utiliza um mapa interativo que permite dar zoom para explorar os locais de gastos em nível regional. Através da distribuição geográfica, o usuário pode observar a localização dos maiores gastos por unidade federativa (**sgUF**), permitindo entender a concentração de despesas públicas em diferentes regiões do país.
- ➔ **Gastos Totais**: Exibe o valor total de gastos públicos, ajustando-se automaticamente conforme os filtros aplicados. Com filtros, apresenta o gasto total específico para as seleções feitas (como partido ou parlamentar), e sem filtros, mostra o total de gastos realizados por todos os deputados ao longo do ano.
- ➔ **Análise de Suspeitas de Gastos**: Esta seção usa um gráfico de barras para identificar valores que ultrapassam um limite preestabelecido. Valores superiores ao limite são destacados com uma linha vermelha, indicando possíveis suspeitas. Esta visualização facilita a identificação de gastos que ultrapassam o padrão esperado, permitindo ao usuário investigar excessos em despesas parlamentares.
- ➔ **Principais Itens de Gastos**: Apresentado em formato de gráfico de pizza, este gráfico destaca os itens com maiores valores de despesas. Além de mostrar a participação dos itens mais onerosos, ele detalha os principais tipos de despesas (representados pela

coluna **txtDescricao**) e os valores associados, permitindo um entendimento das categorias de gastos mais frequentes.

- ➔ **Gastos por Deputado**: Exibido em um gráfico de barras ordenado do maior para o menor gasto, essa visualização apresenta o ranking dos deputados de acordo com o total gasto no ano. Sem filtros, mostra os parlamentares que mais gastaram; com filtros, exibe as despesas de um grupo específico.
- ➔ **Principais Gastos dos Deputados**: Este componente é apresentado em uma tabela, contendo informações detalhadas sobre as despesas dos parlamentares, incluindo nome do deputado (**txNomeParlamentar**), descrição da despesa (**txtDescricao**), valor gasto (**vlrLiquido**), e data de emissão (**datEmissao**). Com essa visão completa, o usuário pode examinar cada gasto individualmente, possibilitando uma análise minuciosa das despesas específicas.
- ➔ **Fatores que Influenciam Compras**: Um gráfico de barras que analisa os fatores que impulsionam os gastos públicos. Essa visualização permite observar as categorias de despesas mais frequentes e suas representações em valores (**vlrLiquido**), possibilitando uma análise das áreas com maior concentração de gastos.

Todos esses gráficos e tabelas são totalmente interativos, e podem ser ajustados conforme o interesse do usuário. Os filtros de **Partidos** e **Parlamentares** afetam simultaneamente todos os dashboards, permitindo a atualização em tempo real e a análise segmentada conforme as necessidades do estudo.

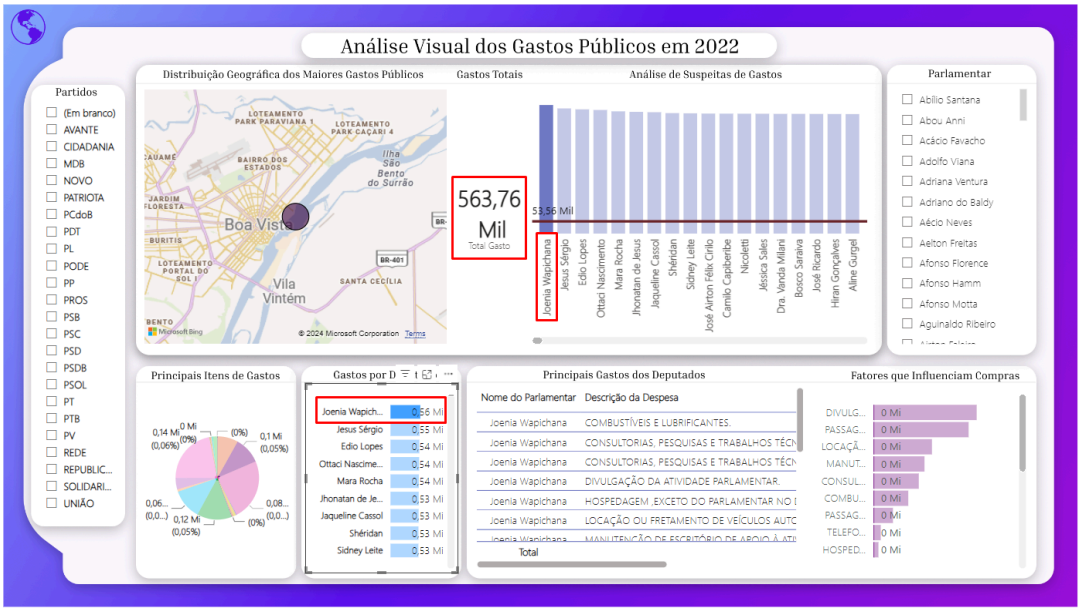
Essa estrutura permite responder às perguntas de análise ao oferecer insights sobre a distribuição de despesas, os maiores gastos por categoria e região, e a identificação de padrões e possíveis anomalias nos gastos públicos.



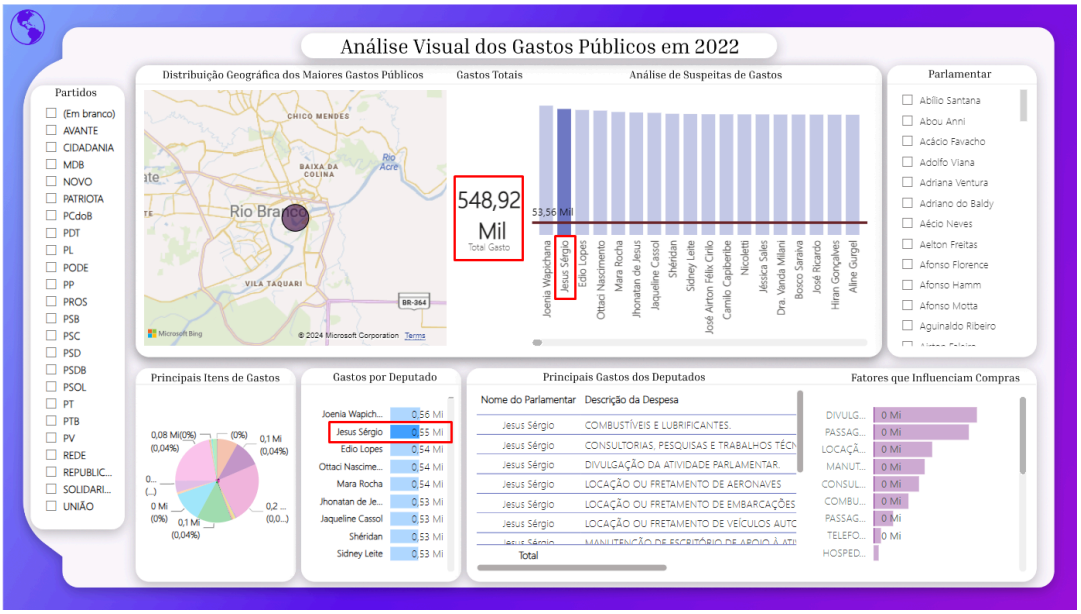
4.6 Relatórios Detalhados

➔ **Deputados com Maiores Gastos no Ano Analisado**: O filtro “Gasto por Deputado”, é possível identificar os parlamentares que mais gastaram ao longo do ano de 2022. A seguir, destacamos o “Top 5” de deputados que registraram os maiores valores em despesas:

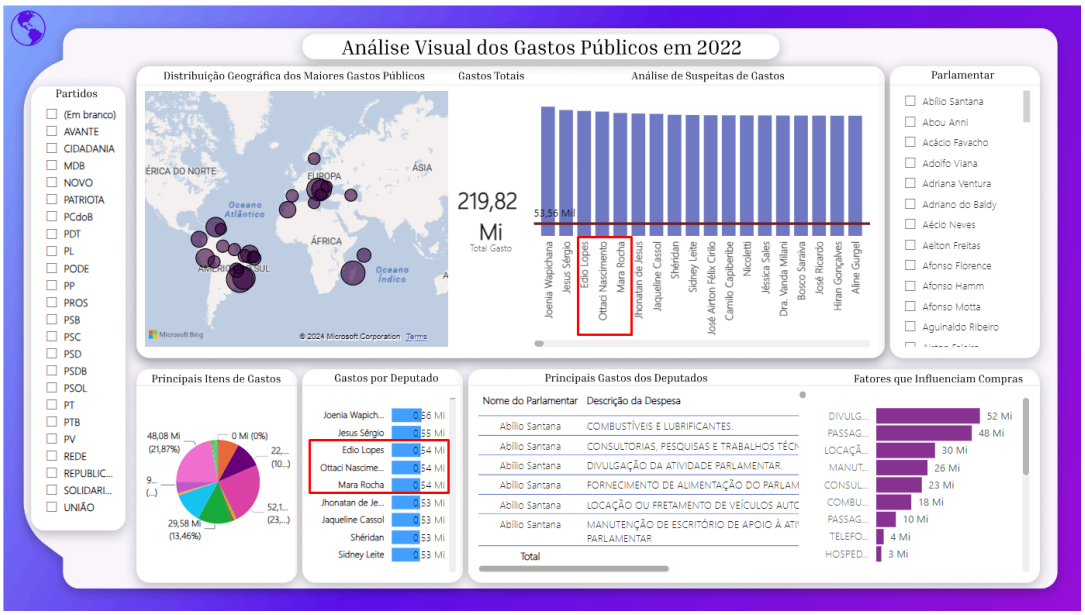
Joenia Wapichana — R$ 56 milhões



➔ **Jesus Sérgio** — R$ 55 milhões



➔ **Edio Lopes**, **Ottaci Nascimento**, **Mara Rocha** — R$ 54 milhões



➔ **Principais Gastos dos Deputados e Fatores Influenciadores das Compras**: Para entender quais foram os maiores gastos dos deputados no ano de 2022, utilizamos os filtros “Principais Gastos dos Deputados”, “Principais Itens de Gastos” e “Fatores que Influenciam Compras” no dashboard. Esses filtros nos permitem visualizar informações detalhadas, como o nome do parlamentar, a descrição das despesas, o valor de cada gasto e o ano correspondente. Dessa forma, conseguimos identificar padrões de despesas e os principais fatores que impactaram as compras realizadas.

Abaixo, destacamos as principais observações sobre os gastos e fatores influenciadores:

1. **Principais Gastos por Categoria:**

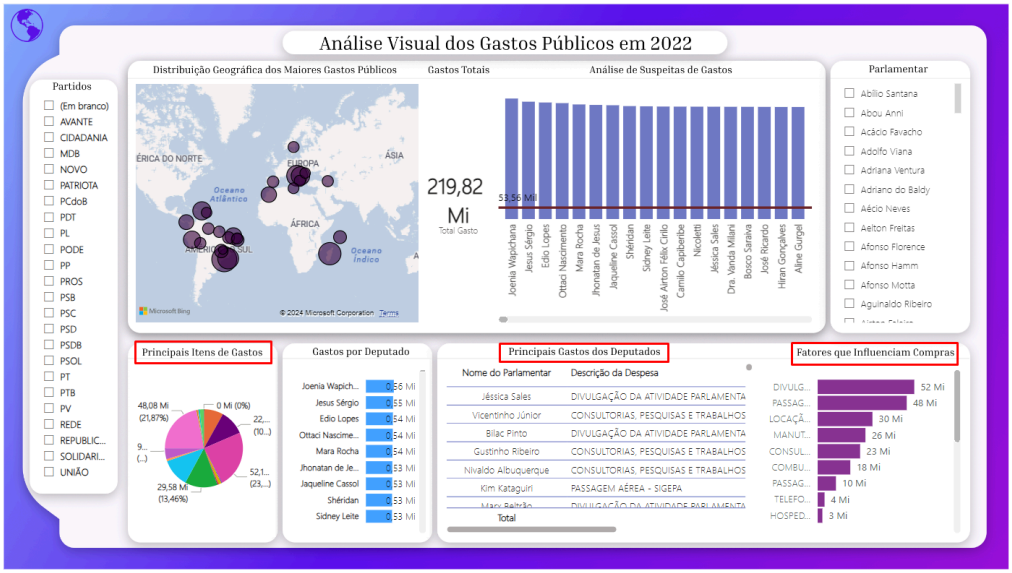
   As duas categorias de maior valor de gasto foram:

   - **Divulgação da Atividade Parlamentar**

- **Consultorias, Pesquisas e Trabalhos Técnicos**

2. Essas categorias representam as despesas com divulgação das atividades do mandato e contratação de serviços técnicos especializados, frequentemente observadas no uso de verbas parlamentares.
3. **Top 5 Fatores que Mais Influenciaram Compras:** Com base no filtro "Fatores que Influenciam Compras", identificamos as cinco categorias com maiores valores totais:
   - **Divulgação da Atividade Parlamentar** – R$ 52 milhões
   - **Passagens Aéreas - SIGEPA** – R$ 48 milhões
   - **Locação ou Fretamento de Veículos Automotores** – R$ 30 milhões
   - **Manutenção de Escritório de Apoio à Atividade Parlamentar** – R$ 26 milhões
   - **Consultorias, Pesquisas e Trabalhos Técnicos** – R$ 23 milhões

Esses fatores revelam as áreas onde os parlamentares tendem a concentrar os maiores gastos, refletindo as atividades necessárias para o exercício do mandato, como deslocamentos, divulgação, e suporte técnico. A análise também permite aplicar filtros específicos para observar os principais gastos individuais e identificar quais deputados mais contribuíram para esses montantes.



- ➔ **Gastos Suspeitos: Análise e Limites Definidos para Identificação**: Para identificar gastos potencialmente suspeitos, desenvolvemos um filtro específico denominado "Análise de Suspeitas de Gastos". Esse filtro aplica limites claros para as despesas, com o objetivo de destacar possíveis desvios e facilitar a análise dos gastos.

Definição dos Limites de Gastos: Para compreender os padrões gerais e identificar exceções, realizamos os seguintes cálculos:

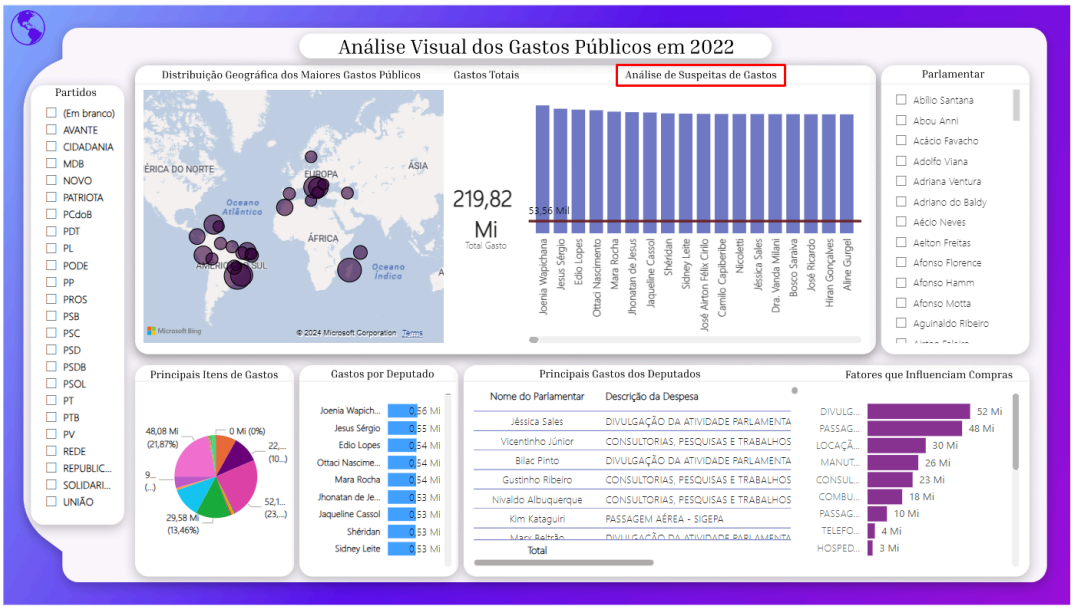
1. **Média Mensal Total de Gastos:**

   Foi calculada uma média mensal total de **R$ 18.318.333,33**, considerando todos os deputados.
2. **Média Mensal por Deputado:**

   Definimos um gasto médio mensal por deputado em **R$ 35.707,24**, que serviu como base para estabelecer limites.
3. **Estabelecimento dos Limites de Alerta:**
   - Limite de 1,5 vezes a média mensal por deputado: **R$ 53.560,86**
   - Limite de 2 vezes a média mensal por deputado: **R$ 71.414,48**

Esses limites ajudam a identificar gastos fora do padrão, oferecendo uma faixa de valores que sinalizam possíveis excessos. No dashboard, os deputados que ultrapassam o limite de 1,5 vezes a média mensal têm seus gastos destacados com uma linha vermelha, facilitando a visualização do valor excedido.

Interpretação dos Resultados e Conclusões: Essa abordagem permite distinguir entre gastos ordinários e aqueles que demandam maior atenção. A linha de limite no gráfico de “Análise de Suspeitas de Gastos” proporciona uma maneira prática de identificar casos de possível excesso, promovendo maior transparência e accountability. A ferramenta oferece um recurso essencial para monitorar e avaliar os gastos públicos com base em uma análise quantitativa sólida, auxiliando na identificação rápida de gastos que necessitam de uma revisão mais detalhada.



5. Machine Learning

5.1 Modelos Utilizados

Utilizaram-se os algoritmos de Regressão Logística e Random Forest para identificar os principais fatores que influenciam os gastos dos deputados. A variável dependente foi definida como "gasto alto", considerando um limite de R$ 53.560,86 anualmente. Os modelos foram escolhidos pela sua capacidade de lidar com dados categóricos e contínuos, permitindo uma análise mais robusta.

Cálculo da Média Mensal por Deputado

- ➔ Média mensal total: R$ 18.318.333,33 (já calculada)
- ➔ Média mensal por deputado: R$ 35.707,24

Cálculo dos Limites por Deputado

- ➔ Limite de 1,5 vezes a média mensal: R$ 53.560,86
- ➔ Limite de 2 vezes a média mensal: R$ 71.414,48

5.2 Resultados Obtidos

Os resultados obtidos a partir da aplicação dos modelos foram apresentados em forma de relatórios de classificação e matrizes de confusão, permitindo avaliar a precisão e eficácia dos modelos. As métricas incluem acurácia, precisão, recall e F1-Score, que fornecem uma visão clara da performance dos algoritmos utilizados.

6. Considerações Finais

O projeto apresenta desafios significativos, como a manipulação e limpeza dos dados, bem como a integração entre as diferentes etapas do processo. As adaptações necessárias foram implementadas para garantir o correto funcionamento do sistema. A experiência adquirida será valiosa para futuras análises de dados e projetos em Big Data, evidenciando a importância da transparência e da fiscalização no uso de recursos públicos..

Referências

BRASIL. Câmara dos Deputados. Dados Abertos. Disponível em: https://dadosabertos.camara.leg.br.

CÂMARA DOS DEPUTADOS. Transparência - Gastos Parlamentares. Disponível em: https://www.camara.leg.br/transparencia/gastos-parlamentares?legislatura=56&ano=2022&mes=&por=deputado&deputado=&uf=&partido=.