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

Nome do autor(a)
PEDRO ALEXANDRE BERGO GONÇALVES PINTO

ANÁLISE DE DESPESAS DE COMBUSTÍVEIS DOS MUNICÍPIOS DO ESTADO DE MINAS GERAIS

Belo Horizonte
2020

Nome do(a) Autor(a)
PEDRO ALEXANDRE BERGO GONÇALVES PINTO

ANÁLISE DE DESPESAS DE COMBUSTÍVEIS DE VEÍCULOS DOS MUNICÍPIOS DO ESTADO DE MINAS GERAIS

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

Belo Horizonte
ano

SUMÁRIO

1. Introdução

1.1. Contextualização

Todo o orçamento de receitas e despesas de todos os municípios do Estado de Minas Gerais são enviados para o Tribunal de Contas do Estado de Minas Gerais (TCE-MG) através do sistema SICOM (Sistema Informatizado de Contas do Município), que permite a apuração de resultados e cruzamento de informações.

Posteriormente, o TCE-MG disponibiliza esses dados em plataforma única para serem utilizados pelos cidadãos para controle e estudos, seguindo a Lei de Acesso à Informação, Lei nº 12.527/2011, de 18/11/2011 e conforme Decreto 8.777/2016 de 11/05/2016.

As análises e cruzamentos dos dados realizados pelo TCE-MG seguem as normas descritas na sua Jurisdição e Competência, e tem como objetivo identificar se a execução das despesas dos municípios está conforme as regras e regulamentações estaduais e municipais, suportando a aprovação e/ou reprovação das contas municipais, através de pareceres técnicos e notas informativas.

A plataforma em que os dados são disponibilizados possui várias formas de acesso, como gráficos estáticos e dinâmicos, bases detalhadas em arquivos nos formatos texto e csv, além de relatórios dos processos. Em especial, a seção Fiscalizando com o TCE permite um visão ampla dos principais indicadores de gastos dos municípios e ligação direta com os pareceres emitidos para as contas enviadas.

1.2. O problema proposto

O volume de dados das prestações de contas dos municípios é muito grande, e apesar de ser obtido e também tratado através de softwares, os prazos legais para análises e emissão de pareceres, bem como recursos e ajustes vão de 365 dias a vários anos.

Outro fator importante se refere ao controle dos entes públicos, papel que é de direito e dever do cidadão, que é o principal beneficiário dos gastos e mesmo que decisões tomadas no âmbito dos gastos esteja de acordo com a legislação, sua aplicação pode ter sido definida sem um critério que atenda a maioria ou aos mais necessitados.

Nesse tocante, um cidadão de uma pequena cidade, deve e pode analisar os dados da administração municipal e pela sua proximidade à execução, consegue identificar mais claramente se a aplicação dos recursos está correta ou condizente com os resultados obtidos na sua localidade.

A proposta desse estudo é realizar um recorte dos gastos e fazer uma análise mais detalhada de algumas despesas. Não há intenção em analisar as despesas em geral, nem mesmo se está dentro dos limites de leis específicas, mas sim se há ou houve desvios, muito comuns em alguns tipos de gastos, seja por mera ineficiência da sua aplicação, por incapacidade na gestão dos bens e serviços, mas também por intenção de aumentar ou registrar despesas de forma a não ser percebida pelo mecanismos de controle.

A pretensão desse estudo é analisar os gastos com veículos, mais especificamente, suas despesas de combustível e manutenção, ficando de fora aquisição e outros usos, como serviços de transporte. Ainda assim, entendemos que se trata de um grande volume de dados, de uma disparidade grande, tanto de tipos de veículos, quanto de aplicações possíveis dos gastos. Sendo assim, as abordagens de Data Science para a obtenção, armazenamento, tratamento e análise dessas informações será de grande importância para atingir resultados tangíveis.

Entendemos que como resultados, pode-se descobrir:

- Quais são os maiores gastos por tipo de combustível e por veículo, e sua frequência de utilização
- Quais são os municípios que mais consomem combustível ou que mais realizam abastecimentos.
- Se há variações nos gastos com combustível que indiquem excesso de consumo.

Os dados utilizados nessa análise são todos de origem de órgãos, que regidos sob a Lei de Acesso à Informação, podem ser utilizados para os objetivos mencionados acima.

Esse estudo analisa todo o período disponibilizado pelo site de dados abertos do TCE-MG, que abrange 2014, 2015, 2016, 2017, 2018 e 2019.

2. Diagrama de Atualização

A Figura 1 - Fluxo de Atualização demonstra de onde foram obtidos os dados e como foram aplicados os conceitos de Data Science. Os dados foram baixados manualmente dos sites do IBGE, ANP e TCE-MG, sendo que este último contém o maior volume de informações. Em seguida, esses dados foram ingeridos para uma base Hive no Cloudera Hadoop, usando a ferramenta Knime. Na terceira etapa, alguns tratamentos foram realizados, além de filtragem nas informações através de instruções HQL – Hive Query Language, gerando tabelas Impala, que contém um conjunto de informações a serem exploradas na etapa seguinte.

A quarta etapa permitiu realizar as análises e agrupamentos de dados, além de categorizar os abastecimentos em Outlier, Suspeito e Normal, tudo feito através do R.

A última etapa foi feita através do Qlik Sense, apresentando os resultados obtidos durante todas as análises.

Figura 1 - Fluxo de Atualização

Legenda:

ID	Nome	Descrição	Ferramenta
1	Data Sources	Sites públicos com informações a serem coletadas e analisadas.	
2	Download	Download manual dos arquivos em formato Zip	
3	ELT – Extract Load to Transform	Processo construído para atualizar os dados dentro da infraestrutura para análise.	KNIME 4.0.2 Beeline
4	Infraestrutura	Estrutura para armazenagem e tratamento dos dados e aplicação de algoritmos de análise	Cloudera Quikstart Virtualbox Image 5.13.0
5	Time Series Analysis	Algoritmo para análise e determinação se os Gastos podem ser considerados Anomalias ou apenas Outliers	R 3.6.
7	Advanced Analytics	Aplicação das bibliotecas de análises avançadas para respostas ao self service on-the-fly	Python 3.6 R 3.6
6	Data Viz e Self Service BI	Ambiente de visualização e análises ad-hoc, além de integração com bibliotecas de Analytics	Qlik Sense Desktop November 2019
7	Dashboards	Dashboards de Análises e interação humana	Qlik Sense Desktop November 2019

3. Coleta de Dados

Foram obtidos 4 conjuntos de dados:

1. Dados das Despesas com Frotas:

   Com origem no portal de dados abertos do TCE-MG, se referem aos dados das despesas a serem analisadas, bem como para quem foram realizadas.

   Metadados:

   - Link para Download
   - Formato do Arquivo: CSV / JSON
   - Qtde de Arquivos: 75.064

   Lay-out:

Nome do Arquivo: ORGAO					
Campos que determinam a chave do registro: *seq_orgao*					
seq	**Nome do Campo**	**Tamanho máximo**	**Formato**	**Obrigatório**	**Conteúdo**
1.	***seq_orgao***	Sempre 5	Number	Sim	Chave identificadora do órgão.
2.	num_anoexercicio	Sempre 4	Number	Sim	Representa o exercício de referencia de envio de dados para os órgãos municipais cadastrados
3.	cod_orgao	Sempre 2	Varchar	Sim	Código do órgão – conforme cadastrado no Portal SICOM.
4.	nom_orgao	100	Varchar	Sim	Descrição do órgão conforme cadastro por parte do município junto ao Sistema de Gestão de Identidade do TCEMG
5.	cod_municipio	Sempre 5	Number	Sim	Conforme código cadastrado para o município tabela do IBGE
6.	nom_municipio	100	Varchar	Sim	Conforme descrição dos municípios de Minas Gerais

7.	nom_uf	Sempre 12	Varchar	Sim	Corresponde ao nome do estado onde se encontram os municípios Valor válido: Minas Gerais
8.	sgl_uf	Sempre 2	Varchar	Sim	Corresponde a sigla do estado onde se encontram os municípios Valor válido: MG
9.	num_versaoarq	3	Varchar	Sim	Corresponde quanto a versão do arquivo
10.	dsc_regiaoplanejamento	50	Varchar	Sim	Corresponde à região de planejamento demarcada dentro do estado onde o município se encontra

Nome do Arquivo: Empenhos					
Campos que determinam a chave do registro: *seq_orgao*					
seq	**Nome do Campo**	**Tamanho máximo**	**Formato**	**Obrigatório**	**Conteúdo**
1.	***seq_empenho***	25	Number	Sim	Chave identificadora do empenho.
2.	cod_municipio	Sempre 5	Number	Sim	Conforme código cadastrado para o município tabela do IBGE
3.	***seq_orgao***	Sempre 5	Number	Sim	Chave identificadora do órgão.
4.	cod_unidade	Sempre 5	Varchar	Sim	Código da unidade orçamentária.
5.	cod_subunidade	Sempre 3	Varchar	Sim	Código da subunidade orçamentária.
6.	***seq_licitacao***	25	Number	Sim	Chave identificadora da licitação
7.	***seq_dispensa***	25	Number	Sim	Chave identificadora da dispensa
8.	***seq_convenio***	25	Number	Sim	Chave identificadora do convênio
9.	***seq_contrato***	25	Number	Sim	Chave identificadora do contrato
10.	***seq_termo_aditivo***	25	Number	Sim	Chave identificadora do termo aditivo do contrato
11.	num_anoexercicio	Sempre 4	Number	Sim	Representa o exercício de referencia de envio de dados.
12.	num_mesexercicio	2	Number	Sim	Representa o mês de referencia de envio de dados.
13.	dsc_funcao	25	Varchar	Sim	Código e descrição da função, conforme Portaria n. 42, de 14/04/1999, expedida pelo Ministério do Orçamento e Gestão.
14.	dsc_subfuncao	50	Varchar	Sim	Código e descrição da Subfunção, conforme Portaria n. 42, de 14/04/1999, expedida pelo Ministério do Orçamento e Gestão.
15.	dsc_naturezadespesa	50	Varchar	Sim	Natureza da Despesa, conforme Discriminação das Naturezas de Despesa padronizada pelo TCEMG, disponível no Portal SICOM.
16.	dsc_programa	200	Varchar	Sim	Código do programa.
17.	dsc_acao	200	Varchar	Sim	Código e descrição que identifica a Ação.
18.	dsc_subacao	200	Varchar	Sim	Código e descrição que identifica a SubAção.
19.	dsc_fonterecurso	50	Varchar	Sim	Código e descrição da fonte de recursos, conforme Classificação por

					Fonte e Destinação de Recursos estabelecida pelo TCEMG.
20.	num_empenho	22	Number	Sim	Número do empenho.
21.	dat_empenho	Sempre 8	Varchar	Sim	Data do empenho.
22.	vlr_empenhado	14	Varchar	Sim	Valor empenhado na fonte de recurso.
23.	vlr_reforco	14	Varchar	Sim	Valor do reforço do empenho.
24.	vlr_anulempenho	14	Varchar	Sim	Valor anulação do empenho
25.	vlr_liquidacao	14	Varchar	Sim	Valor liquidação do empenho
26.	vlr_anulliquidacao	14	Varchar	Sim	Valor anulação da liquidação do empenho
27.	vlr_pagamento	14	Varchar	Sim	Valor pagamento do empenho
28.	vlr_anulpagamento	14	Varchar	Sim	Valor anulação do pagamento do empenho
29.	vlr_outrasbaixas	14	Varchar	Sim	Valor de outras baixas do empenho
30.	vlr_anuloutrasbaixas	14	Varchar	Sim	Valor anulação de doutras baixas do empenho
31.	vlr_rspprocessado	14	Varchar	Sim	Valor de restos a pagar processado do empenho
32.	vlr_rspnaoprocessado	14	Varchar	Sim	Valor de restos a pagar não processado do empenho

Nome do Arquivo: Credor empenho					
Campos que determinam a chave do registro: *seq_orgao*					
seq	**Nome do Campo**	**Tamanho máximo**	**Formato**	**Obrigatório**	**Conteúdo**
1.	***seq_credor_empenho***	25	Number	Sim	Chave identificadora do empenho.
2.	***seq_empenho***	25	Number	Sim	Chave identificadora do empenho.
3.	cod_municipio	Sempre 5	Number	Sim	Conforme código cadastrado para o município tabela do IBGE
4.	***seq_orgao***	Sempre 5	Number	Sim	Chave identificadora do órgão.
5.	num_anoexercicio	Sempre 4	Number	Sim	Representa o exercício de referencia de envio de dados.
6.	num_mesexercicio	2	Number	Sim	Representa o mês de referencia de envio de dados.
7.	num_docto	14	Varchar	Sim	Número do documento do credor

Frota

Nome do Arquivo: Frota - Dados referentes à frota de veículos do Município, contendo código, marca, modelo, ano de fabricação, placa, chassi e renavam do veículo, bem como informações sobre o tipo do deslocamento e peças e serviços.					
Campos que determinam a chave do registro: *seq_orgao*					
seq	**Nome do Campo**	**Tamanho máximo**	**Formato**	**Obrigatório**	**Conteúdo**
1.	***seq_orgao***	Sempre 5	Number	Sim	Chave identificadora do órgão.
2.	cod_unidade	Sempre 5	Varchar	Sim	Código da unidade orçamentária.

3.	cod_subunidade	Sempre 3	Varchar	Sim	Código da subunidade orçamentária.
4.	num_empenho	22	Number	Sim	Número do empenho
5.	dat_empenho	Sempre 8	Varchar	Sim	Data do empenho.
6.	num_anoexercicio	Sempre 4	Number	Sim	Representa o exercício de referencia de envio de dados.
7.	num_mesexercicio	2	Number	Sim	Representa o mês de referencia de envio de dados.
8.	cod_municipio	Sempre 5	Number	Sim	Conforme código cadastrado para o município tabela do IBGE
9.	cod_veiculo	10	Varchar	Sim	Número único, a ser criado pelo município
10.	dsc_veiculo	100	Varchar	Sim	Descrição detalhada do veículo ou equipamento.
11.	dsc_marca	50	Varchar	Sim	Marca do Veículo.
12.	dsc_modelo	50	Varchar	Sim	Modelo do Veículo.
13.	num_anofabricacao	Sempre 4	Number	Sim	Ano do Veículo.
14.	dsc_num_placa	8	Varchar	Não	Placa do Veículo.
15.	dsc_numchassi	30	Varchar	Não	Número do Chassis.
16.	num_renavam	14	Number	Não	Número do RENAVAM
17.	dsc_numserie	20	Varchar	Não	Número de Série. Informado quando não houver número do chassis.
18.	dsc_situacaoveiculo	50	Varchar	Não	Situação do veículo: 01 – Compõe o patrimônio do município (veículo próprio); 02 – Terceirizado ou contratado; 03 – Cedido, empréstimo de outro ente, convênio, acordo ou ajuste.
19.	num_docproprietario	14	Varchar	Não	Número do documento. Este campo torna-se obrigatório quando o campo Situação for definido como 02 – Terceirizado ou contratado ou 03 – Cedido, empréstimo de outro ente, convênio, acordo ou ajuste.
20.	dsc_tipdeslocamento	15	Varchar	Sim	Tipo do deslocamento
21.	dsc_tipgasto	20	Varchar	Sim	Tipo do gasto
22.	dsc_origemgasto				
23.	dsc_pecasservicos	50	Varchar	Não	Descrição da peça ou serviço
24.	qtd_inicial	6	Number	Sim	Hodômetro / Horímetro inicial do veículo no período do mês de referência
25.	qtd_utilizada	14	Number	Sim	Quantidade utilizada no período.
26.	qtd_final	6	Number	Sim	Hodômetro / Horímetro final do veículo no período do mês de referência
27.	vlr_gasto	14	Number	Sim	Valor gasto com peças e serviços no período

2. Dados de Veículos:

Com origem no site do INMETRO, agrupa dados sobre o consumo médio de veículos, os dados do Programa Nacional de Etiquetagem apresentam o consumo médio dos veículos por combustível.

Metadados:

- Link para Cálculo
- Formato do Arquivo: PDF
- Qtde de Arquivos: 6

Lay-out:

- O lay out dos dados é em PDF

3. Dados de Combustível:

   Com origem na base de dados da ANP – Agência Nacional de Petróleo, os dados da pesquisa levantamento de preços e de margens de comercialização de combustíveis permitem identificar se os custos estão muito além do praticado.

   Metadados:

   - Link para Download
   - Formato do Arquivo: XLSX
   - Qtde de Arquivos: 1 (hum)

   Lay-out:

Seq	Nome da coluna/campo	Descrição	Tipo
1	MÊS	Mês e Ano da Pesquisa	Data
2	PRODUTO	Combustível pesquisado	Texto
3	REGIÃO	Região do pais	Texto
4	ESTADO	Estado do país	Texto
5	MUNICÍPIO	Município da pesqisa	Texto
6	NÚMERO DE POSTOS PESQUISADOS	Qtde de postos em foi realizada a pesquisa	Inteiro
7	UNIDADE DE MEDIDA	Unidade de medida dos valores apresentados	Texto
8	PREÇO MÉDIO REVENDA	Preço médio dos postos de combustíveis pesquisados para venda ao consumidor final	Decimal
9	DESVIO PADRÃO REVENDA	Desvio padrão do preço	Decimal
10	PREÇO MÍNIMO REVENDA	Preço mínimo dos postos de combustíveis pesquisados para venda ao consumidor final	Decimal
11	PREÇO MÁXIMO REVENDA	Preço máximo dos postos de combustíveis pesquisados para venda ao consumidor final	Decimal
12	MARGEM MÉDIA REVENDA	MARGEM MÉDIA REVENDA	Decimal
13	COEF DE VARIAÇÃO REVENDA	COEF DE VARIAÇÃO REVENDA	Decimal
14	PREÇO MÉDIO DISTRIBUIÇÃO	PREÇO MÉDIO DISTRIBUIÇÃO	Decimal
15	DESVIO PADRÃO DISTRIBUIÇÃO	DESVIO PADRÃO DISTRIBUIÇÃO	Decimal
16	PREÇO MÍNIMO DISTRIBUIÇÃO	PREÇO MÍNIMO DISTRIBUIÇÃO	Decimal

17	PREÇO MÁXIMO DISTRIBUIÇÃO	PREÇO MÁXIMO DISTRIBUIÇÃO	Decimal
18	COEF DE VARIAÇÃO DISTRIBUIÇÃO	COEF DE VARIAÇÃO DISTRIBUIÇÃO	Decimal

4. Dados dos Municípios:

   Com origem na base de dados do IBGE – Instituto Brasileiro de Geografia Estatística, dados sobre os municípios, como localização, tamanho da população e outros dados demográficos que permitam contextualizar o município para melhorar as análises.

   Metadados:

   - Link para Download
   - Formato do Arquivo: XLSX / ZIP
   - Qtde de Arquivos: 1 (hum)

   Lay-out:

Sq	Nome da coluna/campo	Descrição	Tipo
1	Ano	Ano dos registros	Inteiro
2	Código da Grande Região	Código da Região do país	Texto
3	Nome da Grande Região	Nome da Região do país	Texto
4	Código da Unidade da Federação	Código do Estado	Texto
5	Sigla da Unidade da Federação	Sigla do estado	Texto
6	Nome da Unidade da Federação	Nome do Estado	Texto
7	Código do Município	Código do Múnicipio seguindo a formação do IBGE	Texto
8	Nome do Município	Nome do Município	Texto
9	Região Metropolitana	Região Metropolitana a que pertence, se pertencer a alguma	Texto
10	Código da Mesorregião	Código da Mesorregião Geográfica	Texto
11	Nome da Mesorregião	Nome da Mesorregião Geográfica	Texto
12	Código da Microrregião	Código da Microrregião Geográfica	Texto
13	Nome da Microrregião	Nome da Microrregião Geográfica	Texto
14	Código da Região Rural	Código da Região Rural	Texto
15	Nome da Região Rural	Nome da Região Rural	Texto
16	Tipo da Região Rural	Tipo da Região Rural	Texto
17	Código da Região Geográfica Imediata	Código da Região Geográfica Imediata	Texto
18	Nome da Região Geográfica Imediata	Nome da Região Geográfica Imediata	Texto
19	Município da Região Geográfica Imediata	Município da Região Geográfica Imediata	Texto
20	Código da Região Geográfica Intermediária	Código da Região Geográfica Intermediária	Texto
21	Nome da Região Geográfica Intermediária	Nome da Região Geográfica Intermediária	Texto
22	Município da Região Geográfica Intermediária	Município da Região Geográfica Intermediária	Texto

23	Amazônia Legal	Se pertence a Amazonia Legal	Texto
24	Semiárido	Se pertence ao Semiárido Brasileiro	Texto
25	Código Concentração Urbana	Código Concentração Urbana	Texto
26	Nome Concentração Urbana	Nome Concentração Urbana	Texto
27	Tipo Concentração Urbana	Tipo Concentração Urbana	Texto
28	Código Arranjo Populacional	Código Arranjo Populacional	Texto
29	Nome Arranjo Populacional	Nome Arranjo Populacional	Texto
30	Tipologia Rural-Urbana	Tipologia Rural-Urbana	Texto
31	Hierarquia Urbana	Hierarquia Urbana	Texto
32	Hierarquia Urbana (principais categorias)	Hierarquia Urbana (principais categorias)	Texto
33	Cidade-Região de São Paulo	Se pertence a região metropolitana de são paulo	Texto
34	Valor adicionado bruto da Agropecuária, a preços correntes (R$ 1.000)	Valor adicionado bruto da Agropecuária, a preços correntes (R$ 1.000)	Decimal
35	Valor adicionado bruto da Indústria, a preços correntes (R$ 1.000)	Valor adicionado bruto da Indústria, a preços correntes (R$ 1.000)	Decimal
36	Valor adicionado bruto dos Serviços, a preços correntes - exclusive Administração, defesa, educação e saúde públicas e seguridade social (R$ 1.000)	Valor adicionado bruto dos Serviços, a preços correntes - exclusive Administração, defesa, educação e saúde públicas e seguridade social (R$ 1.000)	Decimal
37	Valor adicionado bruto da Administração, defesa, educação e saúde públicas e seguridade social (R$ 1.000)	Valor adicionado bruto da Administração, defesa, educação e saúde públicas e seguridade social (R$ 1.000)	Decimal
38	Valor adicionado bruto total, a preços correntes (R$ 1.000)	Valor adicionado bruto total, a preços correntes (R$ 1.000)	Decimal
39	Impostos, líquidos de subsídios, sobre produtos, a preços correntes (R$ 1.000)	Impostos, líquidos de subsídios, sobre produtos, a preços correntes (R$ 1.000)	Decimal
40	Produto Interno Bruto, a preços correntes (R$ 1.000)	Produto Interno Bruto, a preços correntes (R$ 1.000)	Decimal
41	População (Nº de habitantes)	População (Nº de habitantes)	Decimal
42	Produto Interno Bruto per capita (R$ 1,00)	Produto Interno Bruto per capita (R$ 1,00)	Decimal
43	Atividade com maior valor adicionado bruto	Qual a principal atividade do município	Texto
44	Atividade com segundo maior valor adicionado bruto	Qual a segunda atividade do município	Texto
45	Atividade com terceiro maior valor adicionado bruto	Qual a terceira maior atividade do muninicípio	Texto

4. Processamento/Tratamento de Dados

A obtenção dos dados ocorreu de forma manual, isso porque as fontes não oferecem uma ferramenta para conexão ou download automático dos arquivos, como

API´s ou links abertos. Foi necessário acessar cada um dos sites, escolher o conjunto de dados desejado e fazer o download. Como resultado, tivemos:

1. Despesas com Frotas: 14 arquivos compactados em formato ZIP
2. Dados de Veículos: 6 arquivos em formato PDF
3. Dados de Combustíveis: 1 arquivo em formato XLSX
4. Dados dos Municípios: 3 arquivos compactados em formato ZIP

Um outro problema encontrado é com relação ao metadados, que não condiz com todo o conteúdo baixado, encontramos diferenças entre os arquivos baixados e o lay-out apresentado pela documentação no site.

Para usar esses arquivos foi necessário criar estratégias de tratamento para cada conjunto. Foram criados diversos scripts para carga dos dados:

Sq	Nome do arquivo	Descrição	Ferramenta
1	001_ELT_DESPESAS_COM_FROTAS v04	Carga dos dados de frotas para dentro do Cloudera Hive	Knime
2	001_ELT_DADOS_MUNICIPIOS	Carga dos dados dos municípios para dentro do Cloudera Hive	Knime
3	001_ELT_DADOS_COMBUSTIVEIS	Carga dos dados dos preços de combustíveis para dentro do Cloudera Hive	Knime
4	apaga_dw.sh	Shell Script para executar as instruções de limpeza das tabelas Impala via Beeline.	Bash
5	apaga_dw.hql	Instruções de limpeza das tabelas Hive	Bash
6	copia_hql.sh	Shell Script para copiar arquivos do Windows para Linux Cloudera.	Bash
7	cria_dw.sh	Shell Script para executar as instruções de criação das tabelas Impala via Beeline	Bash
8	cria_dim_orgao.hql cria_dim_origemgasto.hql cria_dim_pecasservicos.hql cria_dim_pib.hql cria_dim_produto.hql cria_dim_tipdeslocamento.hql cria_dim_tipogasto.hql cria_dim_veiculos.hql cria_fato_empenho.hql cria_fato_frota.hql	Instruções de criação das tabelas Hive	HQL – Hive Query Language.
9	apaga_dw.hql	Script para apagar tabelas do Hive.	HQL – Hive Query Language.
10	ins_dw.sh	Shell Script para executar as instruções de criação de registros das tabelas Hive via Beeline	Bash
11	cria_dim_orgao.hql cria_dim_origemgasto.hql cria_dim_pecasservicos.hql cria_dim_pib.hql cria_dim_produto.hql cria_dim_tipdeslocamento.hql cria_dim_tipogasto.hql cria_dim_veiculos.hql cria_fato_empenho.hql cria_fato_frota.hql ins_dim_orgao.hql ins_dim_origemgasto.hql ins_dim_pecasservicos.hql	Instruções de inserção de registros das tabelas Hive	HQL – Hive Query Language

	ins_dim_pib.hql ins_dim_produto.hql ins_dim_tipdeslocamento.hql ins_dim_tipogasto.hql ins_dim_veiculos.hql ins_fato_empenho.hql ins_fato_empenho_v02.hql ins_fato_frota.hql		
12	copia_iql.sh	Shell Script para copiar arquivos do Windows para Linux Cloudera.	Bash
13	apaga_fatos_impala.iql cria_fato_classe_gastos.iql cria_fato_municipio.iql cria_fato_mun_per.iql cria_fato_mun_per_gasto.iql cria_fato_veiculos.iql gera_fato_classe_gastos.iql gera_fato_municipio.iql gera_fato_mun_per.iql gera_fato_mun_per_gasto.iql gera_fato_veiculos.iql	Instruções de inserção de registros das tabelas Impala executadas via impala-shell	Impala Query Language

O conjunto de dados principal, Despesas com Frotas, continha cerca de 75 mil arquivos em formato CSV e JSON, separados por ano e município. Usando a ferramenta Knime, foi realizada a atualização de uma base ODS – Operational Data Storage, contendo uma cópia das tabelas originais, sendo disponibilizadas no Hive para acesso e tratamento.

Nessa ingestão de dados, foram desprezados todos os arquivos em formato JSON, pois contém o mesmo conteúdo dos arquivos CSV e também as entidades RESTOS A PAGAR de EMPENHOS e de CREDORES, que apresentaram erros, arquivos nulos ou vazios, mas que não são importantes para as análises, e com isso, os arquivos depois de filtrados, resultaram em mais de 37 mil arquivos no formato CSV, totalizando 27.4Gb de dados, conforme demonstrado na Figura 2 - Arquivos Fonte Despesas de Frotas.

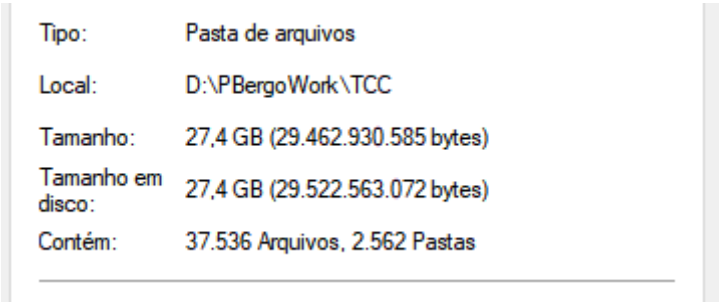

Figura 2 - Arquivos Fonte Despesas de Frotas

Os dados de consumo de combustível por veículo tiveram que ser descartados devido a impossibilidade de leitura dos arquivos disponibilizados no site, que estavam em formato PDF, dificultando em muito seu tratamento.

Numa segunda etapa, foi realizada a carga para dentro do Cloudera Impala, com dados mais tratados para uma lista de 10 tabelas, e usando os seguintes tratamentos aos seus conteúdos:

Sq	Tabela Impala	Tabela origem	Campos	Tratamentos
1	dim_orgao	ods_orgao		
2	dim_origem-gasto	ods_frota	dsc_origemgasto	São gerados registros onde dsc_tipgasto = ('1 - ÁLCOOL (LITRO)','2 - GASOLINA (LITRO)','3 - GÁS NATURAL (M³)','4 - DIESEL (LITRO)');
3	dim_pecasser-vicos	ods_frota	dsc_pecasservicos	São gerados registros onde dsc_tipgasto = ('1 - ÁLCOOL (LITRO)','2 - GASOLINA (LITRO)','3 - GÁS NATURAL (M³)','4 - DIESEL (LITRO)');
4	dim_pib	ods_munpib	c_digo_do_munic_pio AS keymunicipio, ano AS num_anoexercicio, sigla_da_unidade_da_federa__o AS sgl_uf, c_digo_do_munic_pio AS cod_municipio, nome_do_munic_pio AS nom_municipio, c_digo_da_microrregi_o AS cod_microrregiao, nome_da_microrregi_o AS nom_microrregiao, c_digo_da_mesorregi_o AS cod_mesorregiao, nome_da_mesorregi_o AS nom_mesorregiao, produto_interno_bruto__a_pre_os_correntes__r__1_000_ AS vlr_pib, popula__o__n__de_habitantes_ AS nr_habitantes, produto_interno_bruto_per_capita__r__1_00_ AS vlr_pib_percapita	Os campos foram renomeados para comparação com outras tabelas e foram gerados registros somente sigla_da_unidade_da_federa__o = 'MG'
5	dim_produto	ods_precos	produto	Foi incluído um registro manualmente contendo denominação 'OUTROS'.
6	dim_tipdesloca-mento	ods_frota	dsc_tipdeslocamento	São gerados registros onde dsc_tipgasto = ('1 - ÁLCOOL (LITRO)','2 - GASOLINA (LITRO)','3 - GÁS NATURAL (M³)','4 - DIESEL (LITRO)');
7	dim_tipogasto	ods_frota	dsc_tipgasto	São gerados registros onde dsc_tipgasto = ('1 - ÁLCOOL (LITRO)','2 - GASOLINA (LITRO)','3 - GÁS NATURAL (M³)','4 - DIESEL (LITRO)');
8	dim_veiculo	ods_frota	CONCAT_WS("/",cod_veiculo,dsc_num_placa) AS keyveiculo, cod_veiculo, dsc_veiculo, dsc_marca, dsc_modelo, num_anofabricacao, dsc_num_placa, dsc_numchassi, num_renavam, dsc_numserie, dsc_situacaoveiculo, num_docproprietario	São gerados registros onde dsc_tipgasto = ('1 - ÁLCOOL (LITRO)','2 - GASOLINA (LITRO)','3 - GÁS NATURAL (M³)','4 - DIESEL (LITRO)'); Keyveiculo é uma chave a partir de diversos campos.
9	fato_empenho	ods_empenho	CONCAT_WS("/", emp.seq_orgao, emp.cod_unidade, emp.cod_subunidade, emp.num_empenho, emp.dat_empenho, emp.num_anoexercicio, emp.cod_municipio) AS keyempenho, emp.seq_empenho, emp.cod_municipio, emp.seq_orgao, emp.cod_unidade, emp.cod_subunidade,	Keyempenho é uma chave a partir de diversos campos.

			emp.seq_licitacao, emp.seq_dispensa, emp.seq_convenio, emp.seq_contrato, emp.seq_termo_aditivo, emp.num_anoexercicio, emp.num_mesexercicio, emp.dsc_funcao, emp.dsc_subfuncao, emp.dsc_naturezadespesa, emp.dsc_programa, emp.dsc_acao, emp.dsc_subacao, emp.dsc_fonterecurso, emp.num_empenho, emp.dat_empenho, emp.vlr_empenhado, emp.vlr_reforco, emp.vlr_anulempenho, emp.vlr_liquidacao, emp.vlr_anulliquidacao, emp.vlr_pagamento, emp.vlr_anulpagamento, emp.vlr_outrasbaixas, emp.vlr_anuloutrasbaixas, emp.vlr_rsppprocessado, emp.vlr_rspnaoprocessado	
10	fato_frota	ods_frota	row_number() over() AS pk_frota, CONCAT_WS("/",seq_orgao,cod_unidade,cod_subunidade,num_empenho,dat_empenho,num_anoexercicio,cod_municipio) AS keyempenho, CONCAT_WS("/","01",num_mesexercicio,num_anoexercicio) AS keydata, row_number() over() AS idfrota, seq_orgao, cod_unidade, cod_subunidade, num_empenho, dat_empenho, num_anoexercicio, num_mesexercicio, cod_municipio AS KeyMunicipio, cod_municipio, CONCAT_WS("/",cod_veiculo,dsc_num_placa) AS keyveiculo, (dsc_tipdeslocamento) AS keytipodeslocamento, (dsc_tipgasto) AS keytipogasto, (dsc_origemgasto) AS keyorigemgasto, (dsc_pecasservicos) AS keypecassericos, qtd_inicial, if(qtd_utilizada<1, qtd_final - qtd_inicial, qtd_utilizada) AS qtd_utilizada, qtd_final, vlr_gasto, vlr_gasto / if(if(qtd_utilizada<1, if((qtd_final - qtd_inicial)=0,1, (qtd_inicial - qtd_final)), qtd_utilizada)<1,1,qtd_utilizada) AS vlr_gasto_unit	São gerados registros onde dsc_tipgasto = ('1 - ÁLCOOL (LITRO)','2 - GASOLINA (LITRO)','3 - GÁS NATURAL (M³)','4 - DIESEL (LITRO)'); pk_frota e idfrota é uma chave sequencial; key_empenho é uma chave para relacionamento com a tabela fato-empenho; keyveiculo, keytipodeslocamento, keytipogasto, keyorigemgasto, keypecas serviços são chaves para relacionamento com as outras tabelas, mas mantém a mesma condição original. Vlr_gasto_unit contém os dados unitários obtidos a partir da subtração da km inicial e final, que podem ser nulas ou conter dados incorretos, nesse caso, o resultado é o mesmo do gasto total.

4. Análise e Exploração dos Dados

O foco inicial das análises foi identificar nos registros da tabela de gastos com frotas (fato_frotas) a distribuição do Valor por Km (vlr_gasto_unit) categorizada por Tipo de Combustível (keytipogasto). Há 4 tipos de combustíveis:

- 1 - ÁLCOOL (LITRO)
- 2 - GASOLINA (LITRO)

- 3 - GÁS NATURAL (M³)
- 4 - DIESEL (LITRO)

A amplitude dos gastos unitários é extremamente alta, conforme é possível analisar na Figura 3 - Estatísticas da tabela de Gastos Unitários, onde também é possível notar que a média dos registros é bem superior à mediana, o que demonstra uma distribuição desigual das informações.

```
> summary(df_frota_gastos_unit)
    dt_gasto                      keytipogasto     vlr_gasto_unit     num_anoexercicio
 Min.   :2014-01-01   1 - ÁLCOOL (LITRO)  : 238398   Min.   :       0   Min.   :2014
 1st Qu.:2015-08-01   2 - GASOLINA (LITRO):4543988   1st Qu.:       3   1st Qu.:2015
 Median :2017-02-01   3 - GÁS NATURAL (M³):   3945   Median :       4   Median :2017
 Mean   :2016-12-14   4 - DIESEL (LITRO)  :1957829   Mean   :      20   Mean   :2017
 3rd Qu.:2018-05-01                                  3rd Qu.:       4   3rd Qu.:2018
 Max.   :2019-08-01                                  Max.   :61848792   Max.   :2019

   pk_frota           keyveiculo
 Length:6744160     PUB4566:   1453
 Class :character   OPQ9615:   1361
 Mode  :character   PVK6489:   1336
                    PUK6388:   1327
                    OPM2158:   1294
                    PVA5141:   1237
                    (Other):6736152
>
```

Figura 3 - Estatísticas da tabela de Gastos Unitários

Para entender um pouco mais essa distribuição, foi feita análise visual dos gastos, através de histogramas por Valor por Km (gasto unitários) e por combustível, o que confirmou o Positive Skewness, conforme a Figura 4 - Histogramas de Valor por Km.

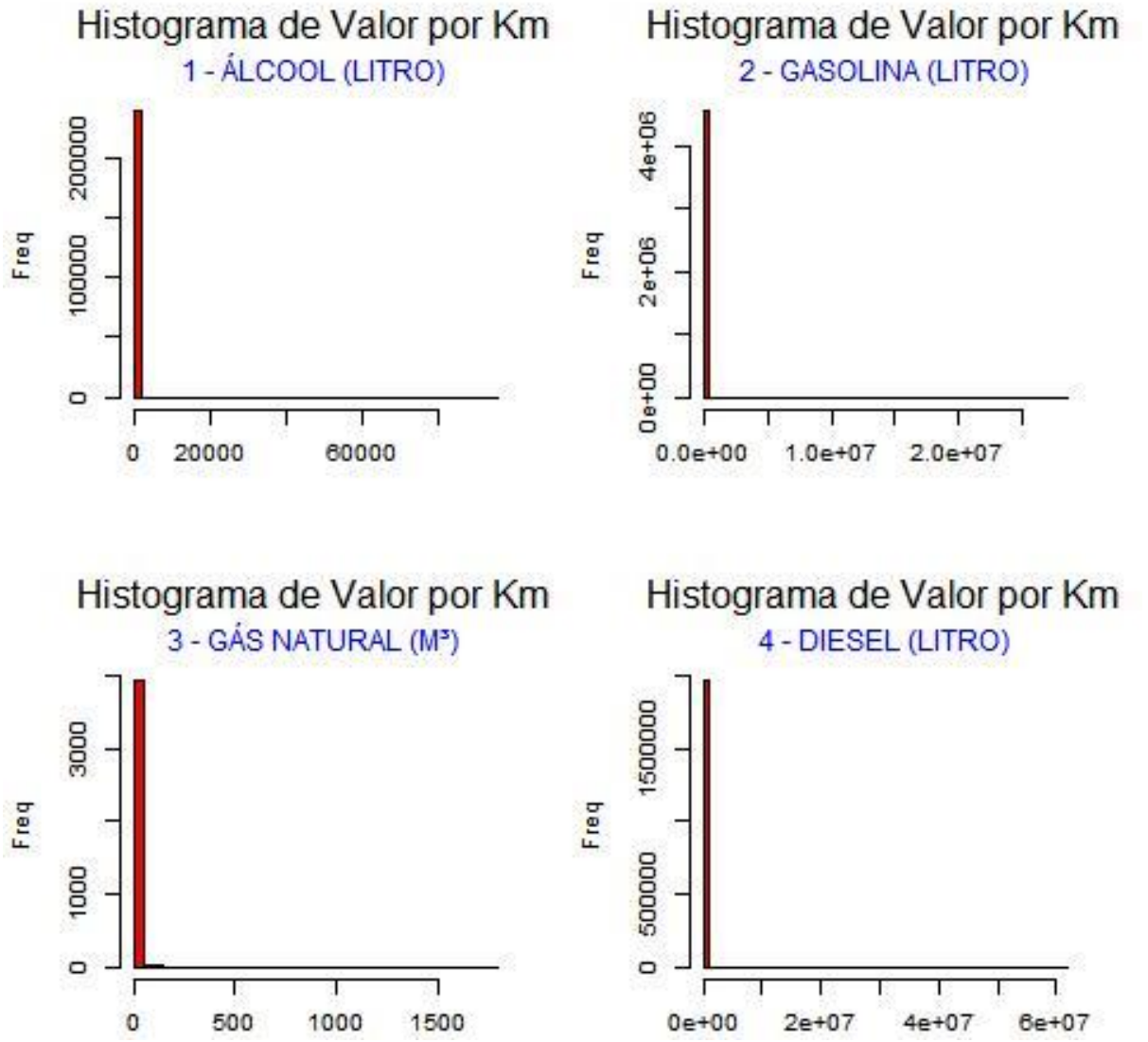

Figura 4 - Histogramas de Valor por Km

Ao analisar a variável de quilometragem rodada pelos veículos (qtd_utilizada), o resultado foi idêntico, com o Positive Skewness com um conjunto de valores extremante amplo, conforme apresentado na Figura 5 - Histograma de Km Rodado por Combustível.

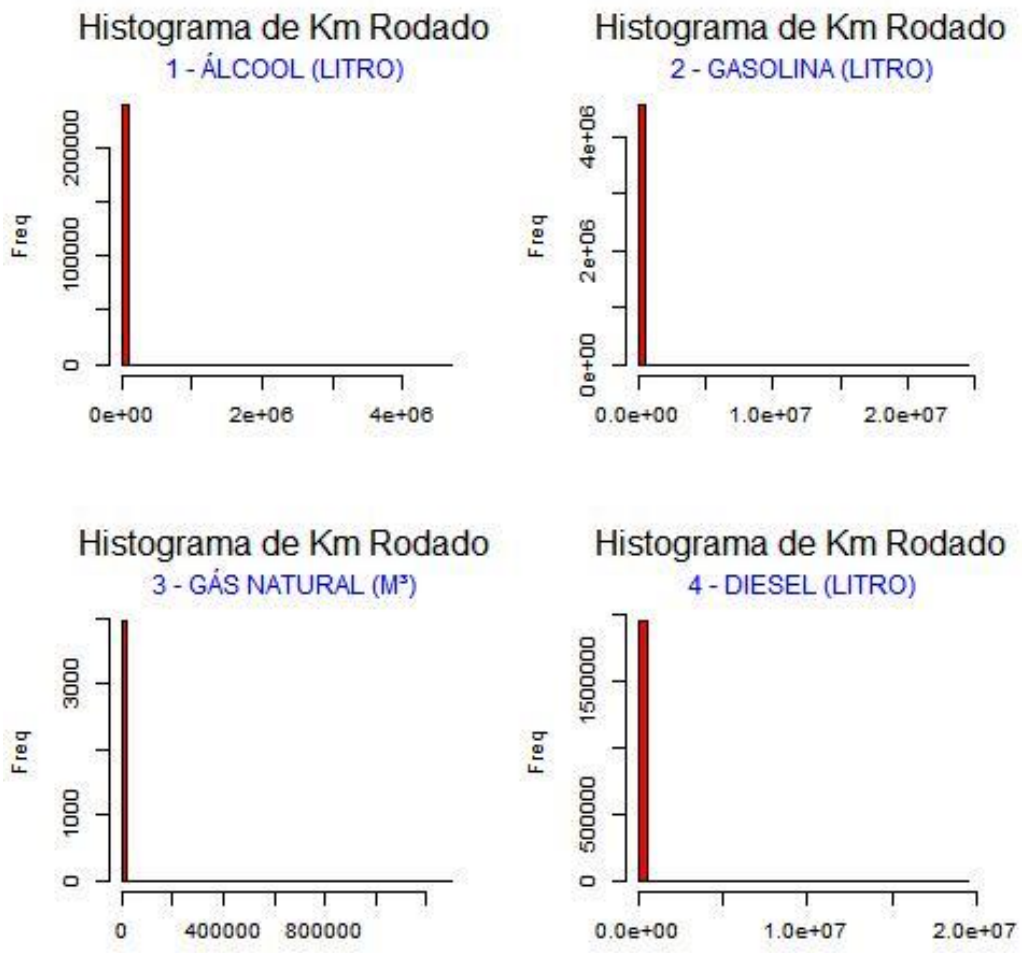

Figura 5 - Histograma de Km Rodado por Combustível

Devido a amplitude de valor por km ser extremamente alta, e que aparentam ruídos no conjunto de dados, foi necessário expurgá-los para uma análise ampla. Para isso, foi aplicada a análise de quartis com BoxPlot anual, retirando-se os outliers, a fim de identificar os limites inferiores e superiores do universo de dados.

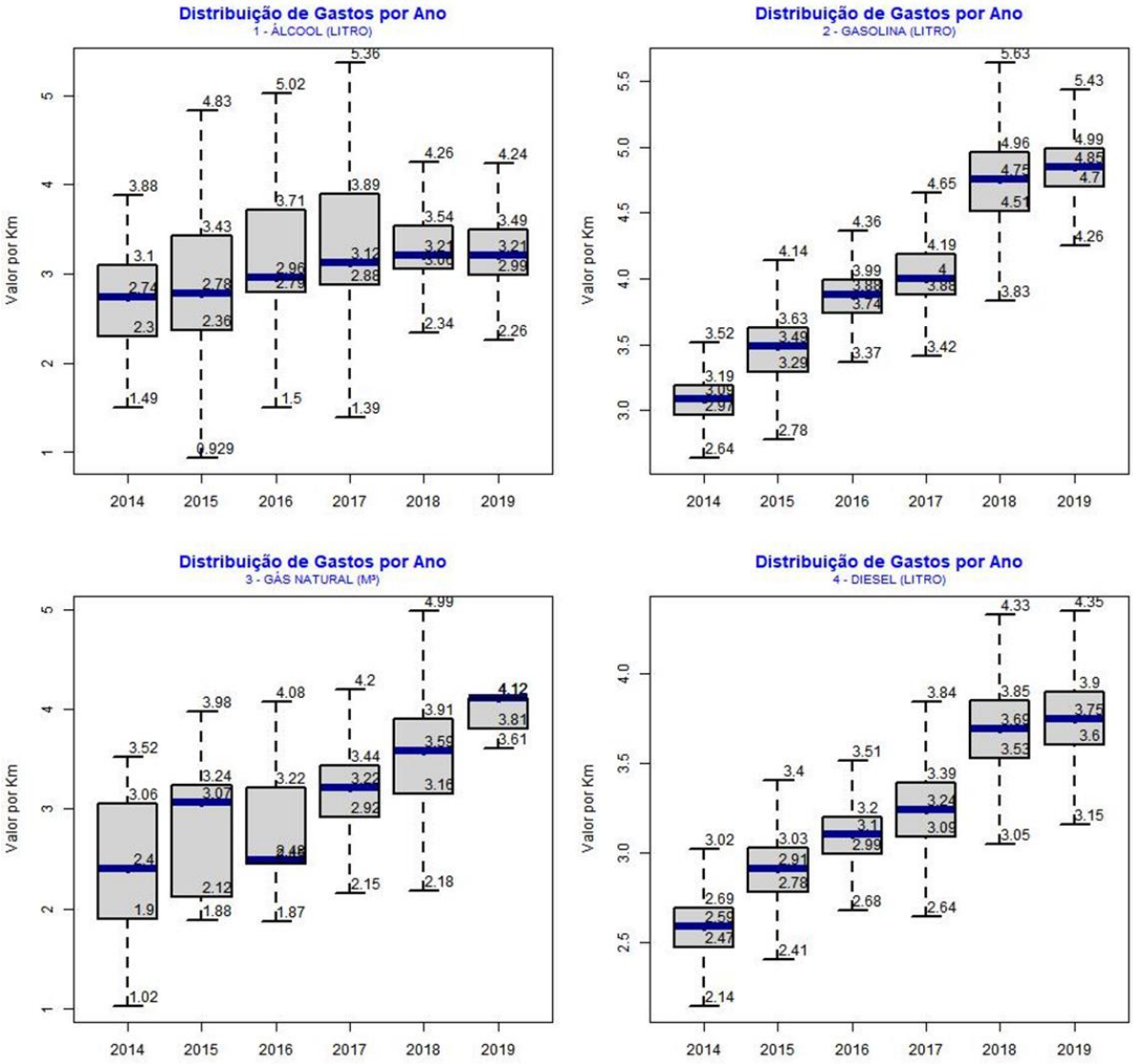

Figura 6 - Distribuição dos Gastos com Boxplot

Com isso, podemos observar as variações no valores por Combustível, representado na Tabela 1 - Limites por Combustível e Ano.

TABELA DE LIMITES POR COMBUSTÍVEL E ANO						
Combustível	Ano	Vl Limite Inferior	Vl 1º Quartil	Vl Inter Quartil	Vl 3º Quartil	Vl Limite Superior
1 - ÁLCOOL (LITRO)	2014	1,49	2,30	2,74	3,10	3,88
1 - ÁLCOOL (LITRO)	2015	0,93	2,36	2,78	3,43	4,83
1 - ÁLCOOL (LITRO)	2016	1,50	2,79	2,96	3,71	5,02
1 - ÁLCOOL (LITRO)	2017	1,39	2,88	3,12	3,89	5,36
1 - ÁLCOOL (LITRO)	2018	2,34	3,06	3,21	3,54	4,26
1 - ÁLCOOL (LITRO)	2019	2,26	2,99	3,21	3,49	4,24
2 - GASOLINA (LITRO)	2014	2,64	2,97	3,09	3,19	3,52
2 - GASOLINA (LITRO)	2015	2,78	3,29	3,49	3,63	4,14
2 - GASOLINA (LITRO)	2016	3,37	3,74	3,88	3,99	4,36
2 - GASOLINA (LITRO)	2017	3,42	3,88	4,00	4,19	4,65

2 - GASOLINA (LITRO)	2018	3,83	4,51	4,75	4,96	5,63
2 - GASOLINA (LITRO)	2019	4,26	4,70	4,85	4,99	5,43
3 - GÁS NATURAL (M^3)	2014	1,02	1,90	2,40	3,06	3,52
3 - GÁS NATURAL (M^3)	2015	1,88	2,12	3,07	3,24	3,98
3 - GÁS NATURAL (M^3)	2016	1,87	2,45	2,48	3,22	4,08
3 - GÁS NATURAL (M^3)	2017	2,15	2,92	3,22	3,44	4,20
3 - GÁS NATURAL (M^3)	2018	2,18	3,16	3,59	3,91	4,99
3 - GÁS NATURAL (M^3)	2019	3,61	3,81	4,12	4,12	4,12
4 - DIESEL (LITRO)	2014	2,14	2,47	2,59	2,69	3,02
4 - DIESEL (LITRO)	2015	2,41	2,78	2,91	3,03	3,40
4 - DIESEL (LITRO)	2016	2,68	2,99	3,10	3,20	3,51
4 - DIESEL (LITRO)	2017	2,64	3,09	3,24	3,39	3,84
4 - DIESEL (LITRO)	2018	3,05	3,53	3,69	3,85	4,33
4 - DIESEL (LITRO)	2019	3,15	3,60	3,75	3,90	4,35

Tabela 1 - Limites por Combustível e Ano

Os dados dessa tabela demonstram que qualquer registro que esteja fora da faixa de valores entre o 1º e 4º Quartil são considerados outliers ou anomalias. Também demonstra que os valores dentro da faixa do 3º e 4º Quartil são considerados valores abusivos ou suspeitos.

Ainda para analisar melhor o impacto do aumento de preços foi necessário analisar a evolução dos gastos por Km ao longo dos períodos da série histórica.

Uma rápida análise da tendência anual dos gastos não apresentou nenhuma avaliação de sazonalidade aparente, mas demonstrou uma evolução contínua do valor dos gastos ao longo dos anos analisados. A figura seguir apresenta a tendência evolutiva do Valor por Km.

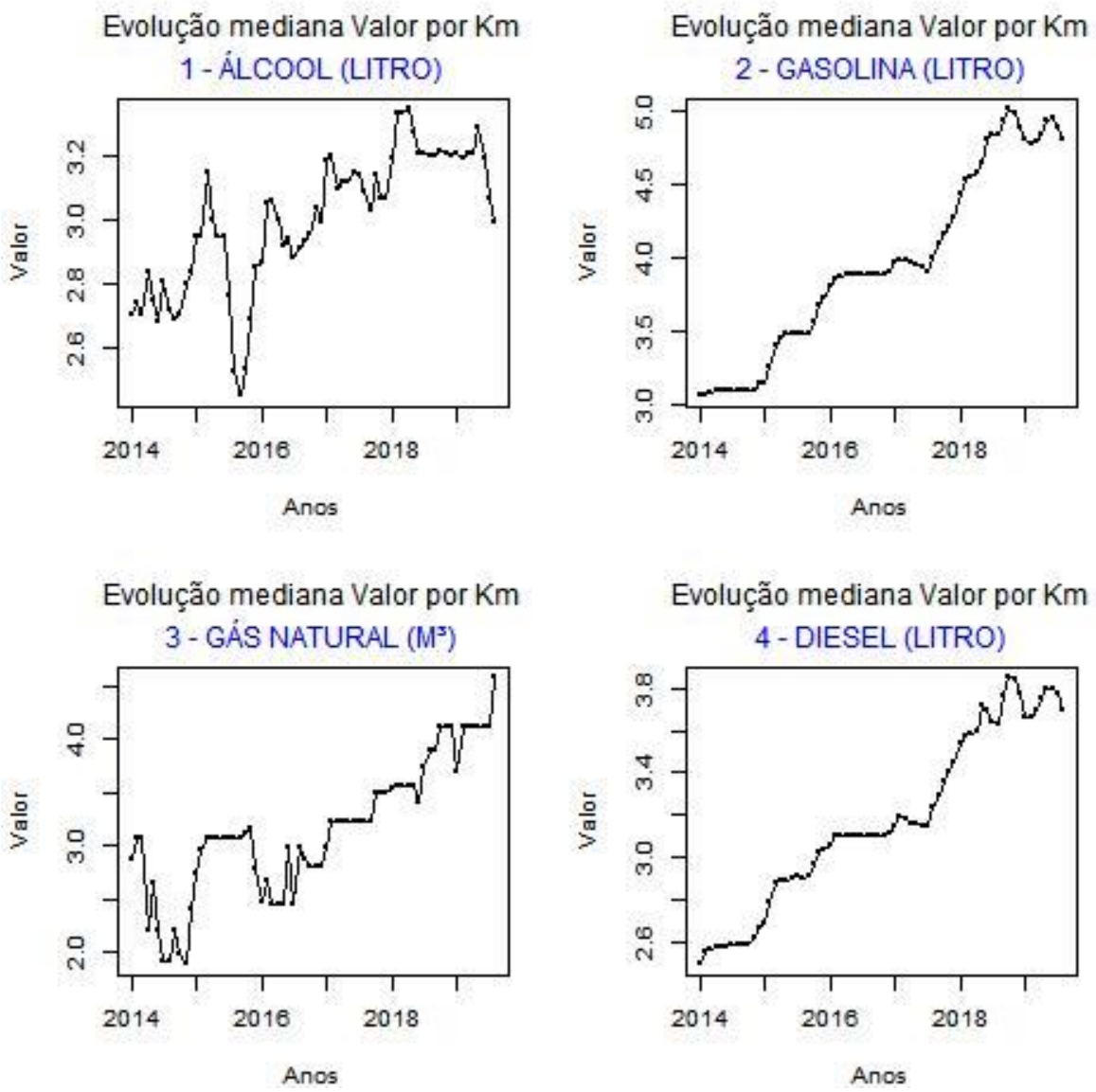

Figura 7 - Análise da Série Histórica dos Gastos por Combustível

Com isso, entende-se que uma avaliação mensal das medianas diante dos limites estabelecidos na Tabela 1 - Limites por Combustível e Ano

pode ser suficiente para avaliar se os gastos unitários com combustíveis são padrões, ou seja, através de um algoritmo construído em R, foi possível criar uma tabela de limites mensais que, uma vez ultrapassados determinam se os valores são OUTLIERS ou necessitam ATENÇÃO, sendo que o primeiro é considerado um erro ou falha, enquanto o segundo apresenta-se com um sobrepreço.

Um programa aplicado diretamente no Impala classificou cada registro de gastos unitário, gerando uma base para análises tanto agrupadas quanto detalhadas. Esse programa resultou em 23,39% de registros requerendo ATENÇÃO (ver Tabela 2 - Resumo da Classificação de Valor por Km dos registros), ou seja, há cerca de 1,6 milhões de

abastecimentos com preços por quilômetro rodado entre o 3º quartil e o limite superior estabelecido na Tabela 1 - Limites por Combustível e Ano.

CLASSE VALOR POR KM	NR. REG	%
NORMAL	4.864.311	72,13%
ATENÇÃO	1.577.505	23,39%
OUTLIER	302.344	4,48%
TOTAL DE REGISTROS	**6.744.160**	**100,00%**

Tabela 2 - Resumo da Classificação de Valor por Km dos registros

Aplicando-se a mesma análise de quartis na variável quilometragem rodada (qtd_utilizada), porém sem aplicar o efeito de anualização, uma vez que essa variável não sofre os efeitos de evolução de preços, resultamos num conjunto de limites de quilometragem que apresentam resultados na Tabela 3 - Limites de Km por Combustível.

TABELA DE LIMITES DE KM POR COMBUSTÍVEL					
Combustível	Limite Inferior	1º Quartil	Mediana	3º Quartil	Limite Superior
1 - ÁLCOOL (LITRO)	0,00	27,68	38,60	57,79	102,95
2 - GASOLINA (LITRO)	0,00	24,71	35,01	47,94	82,79
3 - GÁS NATURAL (M^3)	1,00	46,55	113,58	202,75	436,85
4 - DIESEL (LITRO)	0,00	54,89	97,00	168,00	337,67

Tabela 3 - Limites de Km por Combustível

Um programa aplicado diretamente no Impala classificou cada registro de km rodada, gerando uma base para análises tanto agrupadas quanto detalhadas. Esse programa resultou em 23,39% de registros requerendo ATENÇÃO (ver Tabela 4 - Resumo da Classificação por Km Rodada), ou seja, há cerca de 1,6 milhões de abastecimentos com preços por quilômetro rodado entre o 3º quartil e o limite superior estabelecido na Tabela 3 - Limites de Km por Combustível.

CLASSE KM RODADA	NR. REG	%
NORMAL	877.847	13,0%
ATENÇÃO	5.058.384	75,0%
OUTLIER	807.929	12,0%
TOTAL DE REGISTROS	**6.744.160**	**100,00%**

Tabela 4 - Resumo da Classificação por Km Rodada

6. Apresentação dos Resultados

Com base na classificação realizada, foi criado um Ranking com os Municípios que mais gastam em combustíveis, se considerar somente os municípios que contém registros do tipo NORMAL, o município de NOVA SERRANA tem o maior gasto, totalizando 6 milhões de reais entre 2014 e 2019 para abastecer um total de 323 veículos.

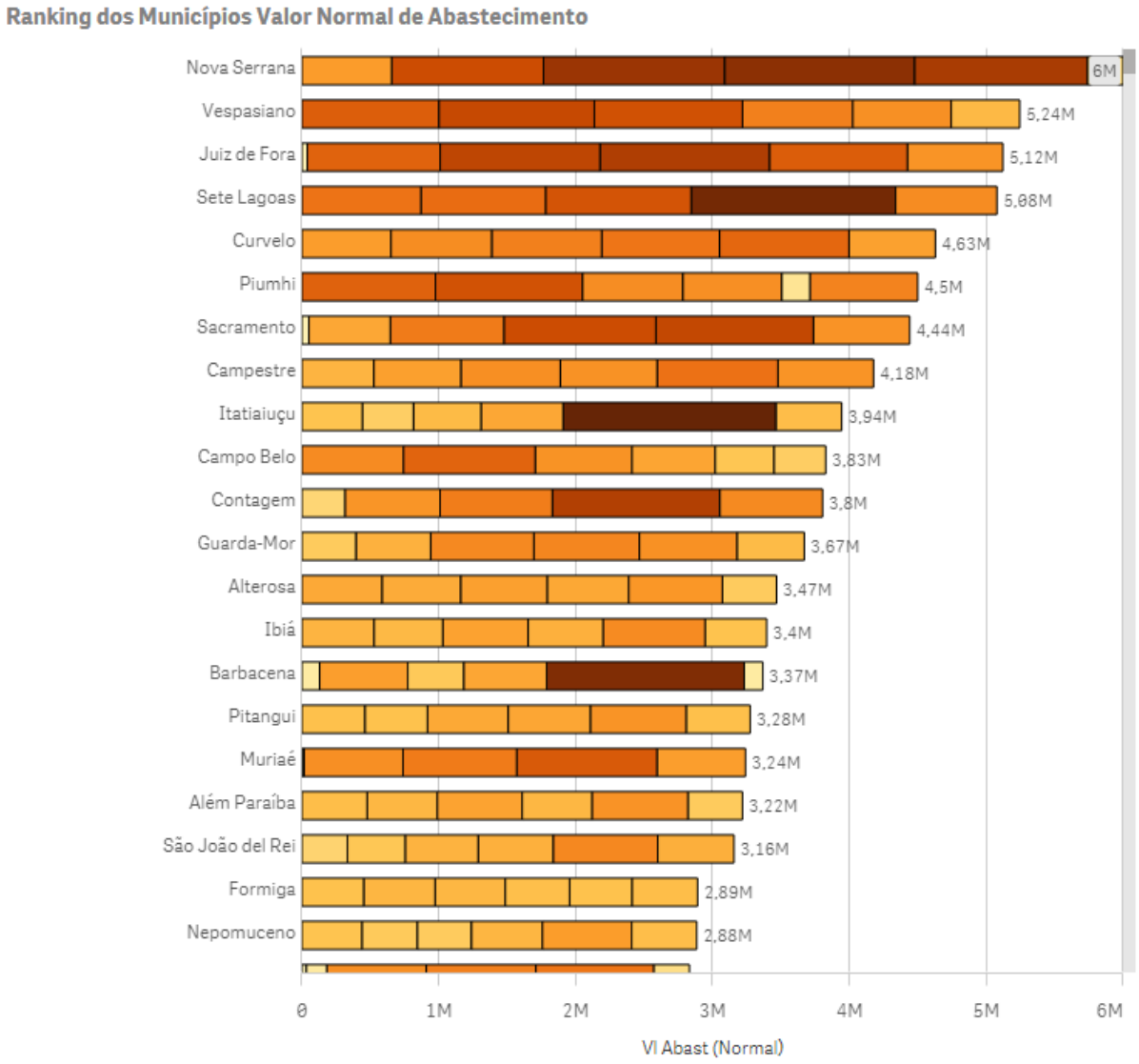

Figura 8 - Ranking dos Gastos dos Munícipios

Através da análise do Raio-X do Município, conseguimos entender qual o comportamento das informações, onde é possível analisar que diversos abastecimentos contêm valores acima da mediana do período, totalizando um valor superior a 450 mil reais acima do esperado, constantes na

Além da análise do município de Nova Serrana, podemos aplicar o método para descobrir quais municípios contém valores discrepantes que podem ser inqueridos e analisado com base nas tabelas de limites determinada pela análise de quartis.

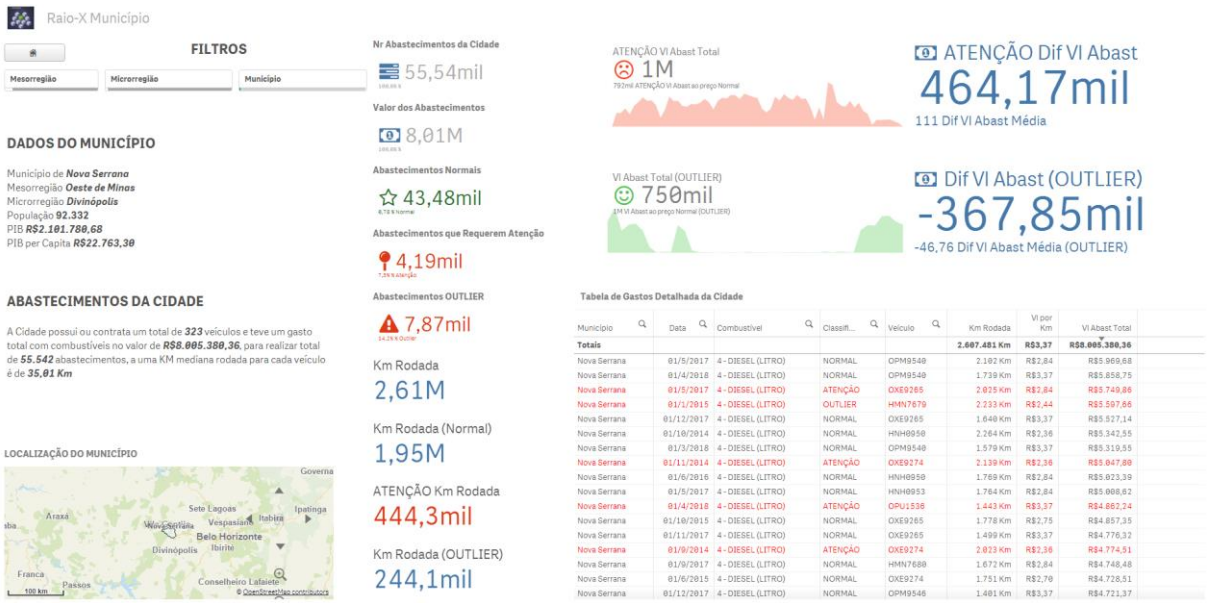

Figura 9 - Raio-X do Município

7. Links

Todos os dados do projeto inclusive o vídeo de apresentação encontra-se no link a seguir:

https://github.com/pbergo/TCC_Entrega