Relatório de Mineração de Dados: Análise da Frota de Veículos por Município e Tipo no Rio Grande do Sul

1. Introdução

Este relatório detalha um projeto de mineração de dados focado na análise da frota de veículos no estado do Rio Grande do Sul, segmentada por município e tipo de veículo. O objetivo principal é aplicar técnicas de agrupamento (clustering) para identificar padrões e características distintas entre os municípios gaúchos com base na composição de suas frotas veiculares. A análise visa fornecer insights sobre a heterogeneidade da frota estadual e as possíveis implicações para planejamento urbano, infraestrutura de transporte e políticas públicas regionais.

2. Escolha e Origem do Conjunto de Dados

O conjunto de dados utilizado para esta análise é a "Frota de Veículos em Circulação" do Rio Grande do Sul, obtido através do portal de Dados Abertos do Governo do Estado do Rio Grande do Sul. A fonte específica é:

- **Origem:** Governo do Estado do Rio Grande do Sul - Dados Abertos
- **URL:** https://dados.rs.gov.br/dataset/frota-veiculos-em-circulacao

Este dataset foi escolhido por sua granularidade, abrangência e por ser específico para o estado do Rio Grande do Sul, contendo informações detalhadas sobre cada veículo emplacado, como marca, fabricação, categoria, espécie, tipo, combustível, município de emplacamento, entre outros. Com mais de 6 milhões de registros, ele oferece uma base robusta para uma análise aprofundada de mineração de dados.

3. Pré-processamento e Análise Exploratória dos Dados

O conjunto de dados original, em formato `.csv` e com um volume considerável (mais de 6 milhões de linhas), exigiu um pré-processamento cuidadoso para garantir sua adequação à análise e otimizar o uso de recursos. As principais etapas foram:

1. **Carregamento em Chunks:** Devido ao grande volume de dados, o arquivo foi lido em partes (chunks de 100.000 linhas) utilizando a biblioteca `pandas` do Python. Isso permitiu processar o dataset sem esgotar a memória do sistema.

2. **Identificação de Delimitadores e Codificação:** A inspeção inicial revelou que o arquivo utilizava ponto e vírgula (`;`) como delimitador de colunas e vírgula (`,`) como separador decimal, além da codificação `latin1`. Essas configurações foram especificadas durante a leitura para garantir a correta interpretação dos dados.
3. **Seleção de Colunas Relevantes:** Para a análise de agrupamento da frota por município e tipo, foram selecionadas as seguintes colunas:
    - `especie`: Categoria principal do veículo (e.g., Carga, Passageiro, Misto).
    - `munic_emplamento`: Município onde o veículo está emplacado.
    - `marca`: Marca do veículo.
    - `fabricacao`: Ano de fabricação do veículo.
    - `categoria`: Categoria do veículo (e.g., Particular, Aluguel).
    - `tipo`: Tipo específico do veículo (e.g., Automóvel, Caminhão, Motocicleta).
4. **Agregação dos Dados:** Para transformar os dados de nível de veículo individual para um nível de município-tipo de veículo, foi realizada uma agregação. Os dados foram agrupados por `munic_emplamento` e `especie`, e a contagem de veículos para cada combinação foi calculada. Isso resultou em um dataset onde cada linha representa um município e as colunas representam a contagem total de veículos e a contagem para cada `especie` de veículo naquele município.
5. **Tratamento de Valores Ausentes e Preenchimento:** Após a agregação, tipos de veículos que não existiam em determinados municípios resultaram em valores `NaN`. Esses valores foram preenchidos com zero (0) para garantir a consistência dos dados numéricos para o agrupamento.

O dataset resultante do pré-processamento, `frota_veiculos_rs_agregado.csv`, contém 497 registros (municípios únicos) e 9 colunas, sendo 7 delas numéricas, representando a contagem de diferentes espécies de veículos. A estrutura do dataset após a limpeza e agregação é a seguinte:

MUNICIPIO	TOTAL	Carga	Misto	Passageiro	TraÃ § Ã£o	Especial	CompetiÃ § Ã£o	ColeÃ § Ã£o
ACEGUA	3816	729	83	2658	104	242	0	0
AGUA SANTA	2593	585	101	1750	39	118	0	0
AGUDO	9951	1476	276	7720	118	360	1	0
AJURICABA	5624	1193	255	3798	83	294	0	1
ALECRIM	3766	389	108	3136	5	128	0	0

Esta etapa de pré-processamento foi fundamental para transformar um dataset massivo e detalhado em um formato adequado para a aplicação de técnicas de mineração de dados, permitindo a análise da composição da frota em um nível municipal.

4. Aplicação da Técnica de Mineração de Dados: Agrupamento K-Means

Com o dataset `frota_veiculos_rs_agregado.csv` preparado, a próxima etapa foi a aplicação de uma técnica de mineração de dados para identificar padrões nos dados da frota. O agrupamento K-Means foi escolhido por sua eficácia em particionar dados em grupos (clusters) com base em suas similaridades, o que é ideal para identificar perfis de municípios com frotas veiculares semelhantes.

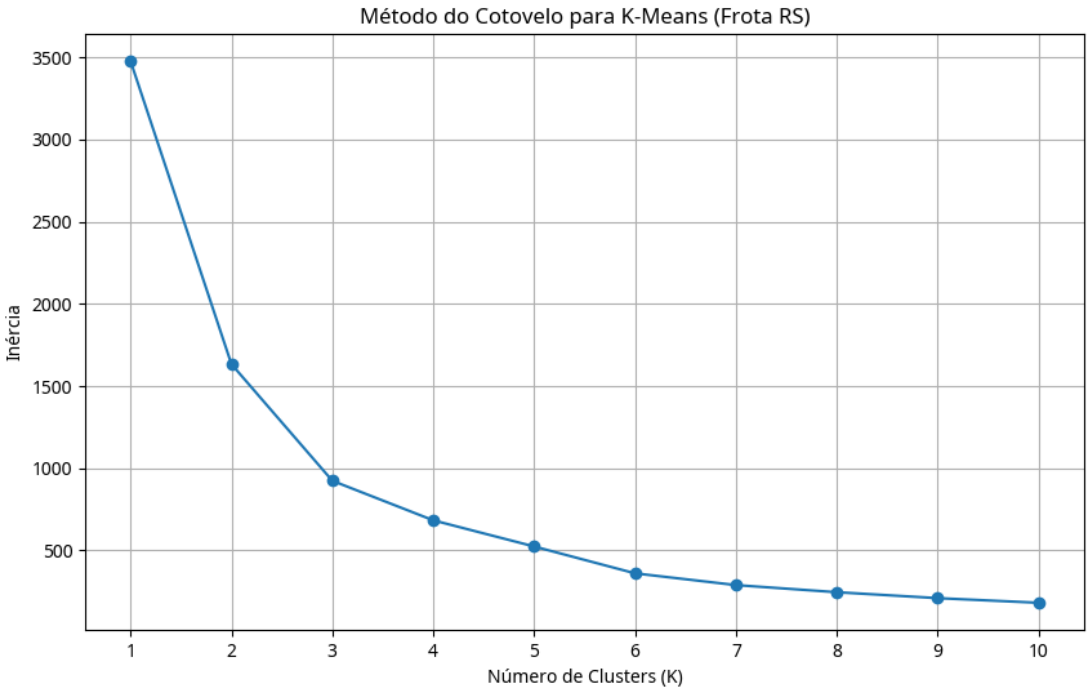
4.1. Normalização dos Dados

Antes de aplicar o K-Means, as colunas numéricas que representam a contagem de tipos de veículos foram normalizadas usando o `StandardScaler`. A normalização é crucial para o K-Means, pois ele calcula distâncias euclidianas entre os pontos de dados. Se as variáveis tiverem escalas muito diferentes (por exemplo, a contagem de automóveis é muito maior que a de triciclos), as variáveis com maior magnitude podem dominar o cálculo da distância, levando a clusters enviesados. A normalização garante que todas as variáveis contribuam igualmente para a formação dos clusters.

4.2. Determinação do Número Ótimo de Clusters (Método do Cotovelo)

Um dos desafios do K-Means é a escolha do número ideal de clusters (`K`). Para isso, foi utilizado o método do cotovelo (Elbow Method). Este método envolve o cálculo da soma dos quadrados das distâncias das amostras para o centro de cluster mais próximo (inércia) para diferentes valores de `K`. O gráfico resultante mostra a inércia em função de `K`. O "cotovelo" no gráfico, onde a taxa de diminuição da inércia se torna menos acentuada, indica o número ótimo de clusters.

O gráfico do método do cotovelo gerado para este dataset (`elbow_method_rs.png`) é apresentado abaixo:



Analisando o gráfico, observa-se uma queda significativa na inércia de K=1 para K=2, e uma diminuição mais gradual a partir de K=3 ou K=4. Embora não haja um "cotovelo" extremamente pronunciado, a região em torno de K=4 ou K=5 parece ser um ponto de inflexão razoável. Para esta análise, optou-se por `K=4` clusters, buscando um equilíbrio entre a redução da inércia e a interpretabilidade dos clusters.

4.3. Aplicação do K-Means e Atribuição de Clusters

Com `K=4` definido, o algoritmo K-Means foi aplicado aos dados normalizados. Cada município foi então atribuído a um dos quatro clusters, com base na similaridade de sua composição de frota com os centroides dos clusters. O DataFrame resultante, `frota_veiculos_rs_com_clusters.csv`, inclui uma nova coluna "Cluster" indicando a qual grupo cada município pertence.

A distribuição dos municípios por cluster é a seguinte:

Cluster	count
1	475
0	20
2	1
3	1

Esta distribuição inicial já sugere que a maioria dos municípios se agrupa em um cluster principal (Cluster 1), enquanto outros clusters representam grupos menores, possivelmente com características de frota mais distintas ou extremas.

4.4. Análise das Características dos Clusters

Para entender o que cada cluster representa, foram analisadas as características médias dos tipos de veículos dentro de cada grupo. Isso é feito examinando os centroides dos clusters (após a transformação inversa da normalização para retornar aos valores originais da frota). Os centroides fornecem uma "média" ou "protótipo" da composição da frota para os municípios em cada cluster.

As características médias dos clusters são:

Cluster	Carga	Misto	Passageiro	Tração	Especial	Competição	Coleção
0	11523	5013	79684	1190	2627	12	63
1	1086	297	6020	84	259	0	2
2	37589	19449	241750	2794	9363	47	370
3	72653	70783	679960	3400	15602	30	628

Esta tabela é fundamental para a interpretação dos clusters. Por exemplo:

- **Cluster 1 (Maioria dos Municípios):** Caracteriza-se por ter uma frota média de passageiros (automóveis, motocicletas, etc.) em torno de 6.000 veículos, com uma proporção menor de veículos de carga, mistos e especiais. Este cluster provavelmente representa a maioria dos municípios de pequeno e médio porte do RS.
- **Cluster 0:** Apresenta uma frota de passageiros significativamente maior (quase 80.000 veículos), além de números mais elevados em todas as outras categorias em comparação com o Cluster 1. Este cluster pode englobar municípios de médio a grande porte.
- **Cluster 2 e 3 (Municípios Singulares):** Estes clusters, com apenas um município cada, representam os extremos da frota. O Cluster 3, com uma frota de passageiros de quase 680.000 veículos, e o Cluster 2, com 240.000, indicam municípios com frotas massivamente maiores que os demais, provavelmente as grandes cidades do estado. A análise dos municípios atribuídos a esses clusters na próxima seção confirmará essa hipótese.

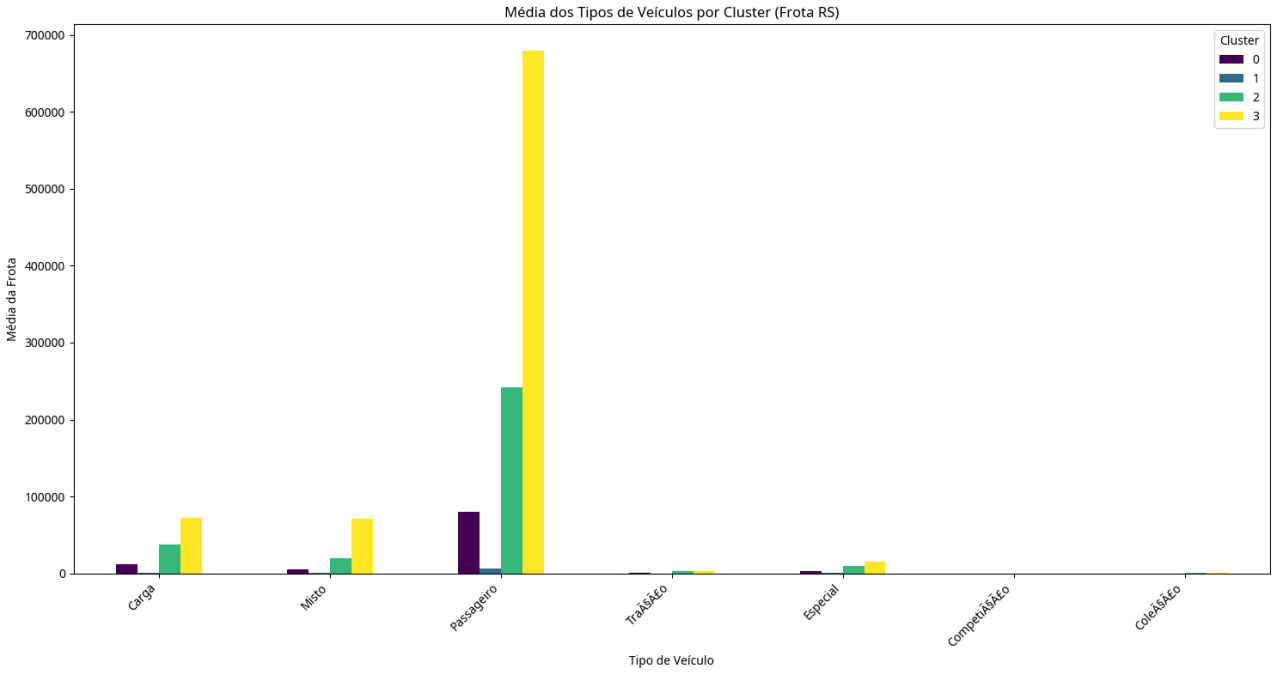
Essa análise dos centroides permite uma compreensão aprofundada dos perfis de frota que o algoritmo K-Means identificou, servindo como base para as visualizações e conclusões.

5. Geração de Visualizações e Análise de Resultados

Para complementar a análise quantitativa dos clusters, foram geradas visualizações que permitem uma compreensão mais intuitiva das características de cada grupo e da distribuição dos municípios. As visualizações são cruciais para comunicar os padrões descobertos e validar as interpretações dos centroides dos clusters.

5.1. Distribuição Média dos Tipos de Veículos por Cluster

O gráfico de barras `vehicle_type_distribution_by_cluster_rs.png` ilustra a média da frota para cada tipo de veículo em cada um dos quatro clusters identificados. Esta visualização reforça as observações feitas a partir da tabela de centroides, mas de forma gráfica, facilitando a comparação entre os clusters.



Observações chave a partir do gráfico:

- **Domínio de Veículos de Passageiros:** Em todos os clusters, a categoria "Passageiro" (que inclui automóveis, motocicletas, etc.) é a mais representativa em termos de volume, o que é esperado para a frota geral de veículos.
- **Diferenças de Escala entre Clusters:** O Cluster 3 (representado pela barra amarela) e o Cluster 2 (barra verde-água) mostram volumes de frota significativamente maiores em todas as categorias em comparação com o Cluster 0 (roxo escuro) e, especialmente, o Cluster 1 (azul-petróleo). Isso confirma que os Clusters 2 e 3 representam municípios com frotas muito maiores.
- **Perfil do Cluster 1 (Maioria):** Este cluster, que abrange a grande maioria dos municípios, apresenta volumes de frota mais modestos em todas as categorias, com uma predominância clara de veículos de passageiros, mas com números de carga e mistos em proporções esperadas para municípios de menor porte.
- **Perfil do Cluster 0:** Embora menor em número de municípios que o Cluster 1, o Cluster 0 exibe uma frota média consideravelmente maior, especialmente em veículos de passageiros e carga, indicando municípios de porte intermediário a grande.
- **Clusters 2 e 3 (Grandes Centros):** Estes clusters se destacam pela magnitude de suas frotas. O Cluster 3, em particular, mostra uma frota de passageiros que se aproxima de 700.000 veículos, e volumes substanciais em outras categorias, sugerindo que este cluster representa a maior

metrópole do estado. O Cluster 2, embora menor que o 3, ainda possui uma frota muito superior aos Clusters 0 e 1, indicando outro grande centro urbano.

5.2. Distribuição Geográfica dos Clusters

Para entender a localização dos municípios em cada cluster, foi realizada uma listagem dos top 5 municípios (quando aplicável) para cada grupo. Esta análise ajuda a contextualizar os resultados do agrupamento e a identificar quais tipos de municípios se encaixam em cada perfil de frota.

- **Top 5 Municípios no Cluster 0:**

MUNICIPIO
BENTO GONCALVES
CANOAS
CARAZINHO
ERECHIM
FARROUPILHA

Este cluster inclui municípios de médio a grande porte no interior do estado, que são polos regionais e possuem frotas consideráveis, mas não atingem o volume das maiores cidades.

- **Top 5 Municípios no Cluster 1:**

MUNICIPIO
ACEGUA
AGUA SANTA
AGUDO
AJURICABA
ALECRIM

Este é o cluster mais populoso, contendo a maioria dos municípios do Rio Grande do Sul. São predominantemente cidades de pequeno e médio porte, com frotas veiculares mais homogêneas e menores em volume total.

- **Top 5 Municípios no Cluster 2:**

MUNICIPIO
CAXIAS DO SUL

O Cluster 2 é composto por um único município, Caxias do Sul. Isso indica que Caxias do Sul possui um perfil de frota que é significativamente diferente dos municípios do Cluster 0 e 1, mas ainda distinto da maior metrópole do estado. Sua frota é grande e diversificada, refletindo sua importância econômica e populacional.

- **Top 5 Municípios no Cluster 3:**

MUNICIPIO
PORTO ALEGRE

O Cluster 3 também é composto por um único município, Porto Alegre. Este resultado é esperado, pois a capital do estado possui a maior frota veicular, com volumes muito superiores a qualquer outro município, justificando sua formação em um cluster isolado. A composição de sua frota reflete a densidade populacional e a atividade econômica de uma grande metrópole.

As visualizações e a análise da distribuição geográfica dos clusters fornecem uma validação empírica dos resultados do K-Means, confirmando que o algoritmo conseguiu agrupar os municípios em perfis de frota coerentes e interpretáveis, que refletem a realidade demográfica e econômica do Rio Grande do Sul.

6. Conclusão

Este projeto de mineração de dados demonstrou a eficácia do agrupamento K-Means na identificação de padrões e perfis distintos na frota de veículos do Rio Grande do Sul. Através da análise de um dataset robusto e detalhado, foi possível categorizar os municípios gaúchos em quatro clusters principais, cada um com características de frota bem definidas:

- **Cluster 1 (Municípios de Pequeno e Médio Porte):** Representa a vasta maioria dos municípios, com frotas de volume moderado e predominância de veículos de passageiros, refletindo a realidade de cidades com menor densidade populacional e atividade econômica.
- **Cluster 0 (Municípios de Porte Intermediário a Grande):** Agrupa cidades com frotas significativamente maiores que o Cluster 1, indicando polos regionais com maior movimentação e diversidade de veículos, incluindo uma proporção maior de veículos de carga e mistos.
- **Cluster 2 (Caxias do Sul):** Um cluster singular que destaca Caxias do Sul como um centro urbano com uma frota veicular de grande volume e diversidade, mas ainda distinta da capital.
- **Cluster 3 (Porto Alegre):** Outro cluster singular, representando Porto Alegre, a capital do estado, que possui a maior e mais complexa frota veicular, com volumes muito superiores aos demais municípios.

Os resultados obtidos fornecem insights valiosos para diversas áreas. Para o planejamento urbano e de transporte, a identificação desses perfis de frota pode auxiliar na alocação de recursos, no desenvolvimento de infraestrutura viária e na implementação de políticas de mobilidade urbana mais eficazes e direcionadas às necessidades específicas de cada grupo de municípios. Por exemplo, municípios do Cluster 1 podem focar em melhorias de transporte público local e infraestrutura para

veículos leves, enquanto os Clusters 0, 2 e 3 demandam soluções mais complexas para o tráfego intenso, logística de carga e transporte de massa.

Para o setor automotivo e de serviços, a compreensão desses clusters pode guiar estratégias de mercado, como a distribuição de concessionárias, oficinas e serviços de manutenção, adaptando a oferta de produtos e serviços ao perfil da frota local. Além disso, para a segurança pública e fiscalização de trânsito, a análise da composição da frota pode informar a distribuição de recursos e o foco de operações em áreas com maior concentração de determinados tipos de veículos.

Em termos de limitações, a análise se baseou em dados de 2019. A frota de veículos é dinâmica, e futuras análises poderiam se beneficiar de dados mais recentes para capturar mudanças no comportamento de compra e uso de veículos. Além disso, a inclusão de outras variáveis, como dados demográficos, econômicos e de infraestrutura viária dos municípios, poderia enriquecer ainda mais o agrupamento e fornecer insights mais profundos sobre os fatores que influenciam a composição da frota.

Este trabalho demonstra o potencial da mineração de dados para transformar grandes volumes de informações em conhecimento acionável, contribuindo para uma compreensão mais aprofundada da realidade da frota veicular no Rio Grande do Sul e subsidiando decisões estratégicas em múltiplos setores.

7. Referências

- **Dataset Frota de Veículos em Circulação - Rio Grande do Sul:** https://dados.rs.gov.br/dataset/frota-veiculos-em-circulacao
- **Documentação Pandas:** https://pandas.pydata.org/docs/
- **Documentação Scikit-learn (K-Means):** https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- **Documentação Matplotlib:** https://matplotlib.org/stable/contents.html
- **Documentação Seaborn:** https://seaborn.pydata.org/