Relatório Final do Projeto de Análise de Dados

Título do Projeto: Análise de Dados do Mercado de Veículos Seminovos a partir de Web Scraping

Autores: Hélio Ricardo / Carlos Nascimento

Data: 27 de julho de 2025

Resumo

Este relatório documenta o ciclo completo de uma análise de dados, desde a coleta de informações em fontes abertas na internet (web scraping) até a realização de testes de hipóteses para extrair conclusões estatisticamente válidas. Foram coletados dados de anúncios de veículos do portal iCarros, que passaram por um rigoroso processo de pré-processamento. Esta etapa incluiu não apenas a limpeza, formatação e tratamento de erros, mas também a aplicação de técnicas de Engenharia de Atributos, como Discretização e Normalização, para enriquecer o dataset e prepará-lo para análises avançadas. Na sequência, foi realizada uma análise exploratória para entender o perfil dos dados, seguida por uma análise inferencial para comparar preços de veículos entre diferentes estados. O projeto demonstra com sucesso a aplicação de um fluxo de trabalho robusto de ciência de dados para transformar dados brutos em insights acionáveis.

1. Introdução

O mercado de veículos seminovos e usados é um ambiente dinâmico e de grande volume de dados. A análise desses dados pode revelar tendências de mercado, diferenças de preços regionais e insights valiosos tanto para consumidores quanto para vendedores.

O objetivo deste projeto é aplicar as principais etapas de um fluxo de trabalho de ciência de dados para analisar uma amostra deste mercado. O projeto seguirá quatro passos fundamentais:

1. **Coleta de Dados:** Extração de dados reais e atualizados diretamente de um portal de classificados online.
2. **Pré-processamento de Dados:** Limpeza, formatação, tratamento de erros e dados ausentes para garantir a qualidade e a confiabilidade da análise.
3. **Análise Exploratória de Dados (AED):** Geração de estatísticas descritivas e visualizações para compreender as características principais dos dados.
4. **Análise Inferencial:** Realização de um teste de hipótese para validar uma questão de negócio com rigor estatístico.

2. Metodologia e Execução

2.1. Etapa 1: Coleta de Dados (Web Scraping)

A primeira fase do projeto consistiu na coleta dos dados. Foi desenvolvido um programa em Python utilizando as bibliotecas Requests para requisições HTTP rápidas e

BeautifulSoup para a análise inicial do HTML. Para extrair dados dinâmicos (como o preço, que é carregado por JavaScript), foi utilizada a biblioteca Selenium com o undetected-chromedriver, garantindo a capacidade de navegar nas páginas como um usuário real.

O código foi programado para navegar por 50 páginas de anúncios, extraindo informações como marca, modelo, ano, quilometragem, preço e localização de cada veículo. Para garantir a robustez contra interrupções, cada registro foi salvo em um arquivo CSV (**salva-icarros.csv**) imediatamente após sua coleta.

O resultado foi um dataset bruto com 1000 registros, contendo dados "sujos" e não estruturados, como pode ser visto em uma amostra inicial:

estado	cidade	Preço	km
PR	Curitiba	ERRO	300 Km
SP	Ibitinga	ERRO	220.000 Km
SP	Campinas	R$ 60.090,00	84.287 Km

2.2. Etapa 2: Pré-processamento dos Dados

Com os dados brutos em mãos, a etapa de pré-processamento foi dividida em duas fases:

2.2.1. Limpeza Inicial e Conversão de Tipos

O dataset inicial apresentava diversos desafios:

- As colunas preço e km estavam em formato de texto (object), continham caracteres (R$, Km, pontos) e valores de erro ("ERRO"), impedindo cálculos matemáticos.
- Problemas de codificação de caracteres (UnicodeDecodeError) foram encontrados, exigindo a leitura do arquivo com a codificação correta (latin-1).

Um programa foi executado para realizar a limpeza. Com as seguintes transformações:

- Converteu os valores "ERRO" para NaN (Not a Number), o marcador padrão para dados ausentes.
- Removeu todos os caracteres não numéricos das colunas preço e km.
- Converteu as colunas limpas para o tipo numérico (float64).

Ao final desta etapa, foi gerado o arquivo **icarros_preprocessado.csv**, com os dados estruturados e prontos para a análise numérica inicial.

2.2.2. Análise e Tratamento de Outliers

Após a limpeza inicial, foi realizada uma análise descritiva que revelou a presença de **outliers** — valores extremos que não condizem com a realidade. A quilometragem

máxima encontrada era de **9.999.999 Km**, um valor claramente irreal que distorcia todas as métricas e gráficos.

Para corrigir isso, foi aplicado um filtro para remover todos os registros com quilometragem acima de um limiar realista (1.000.000 Km). O resultado desta limpeza é visível na comparação dos histogramas de quilometragem "antes" (Figura 1) e "depois" (Figura 2).

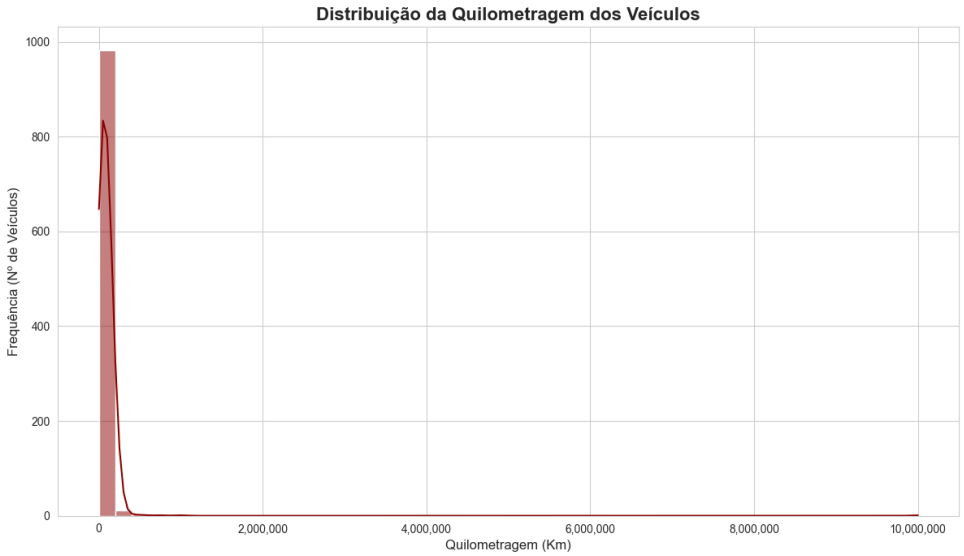

Figura 1: Distribuição da Quilometragem ANTES da limpeza. (Gráfico ilegível devido ao outlier)

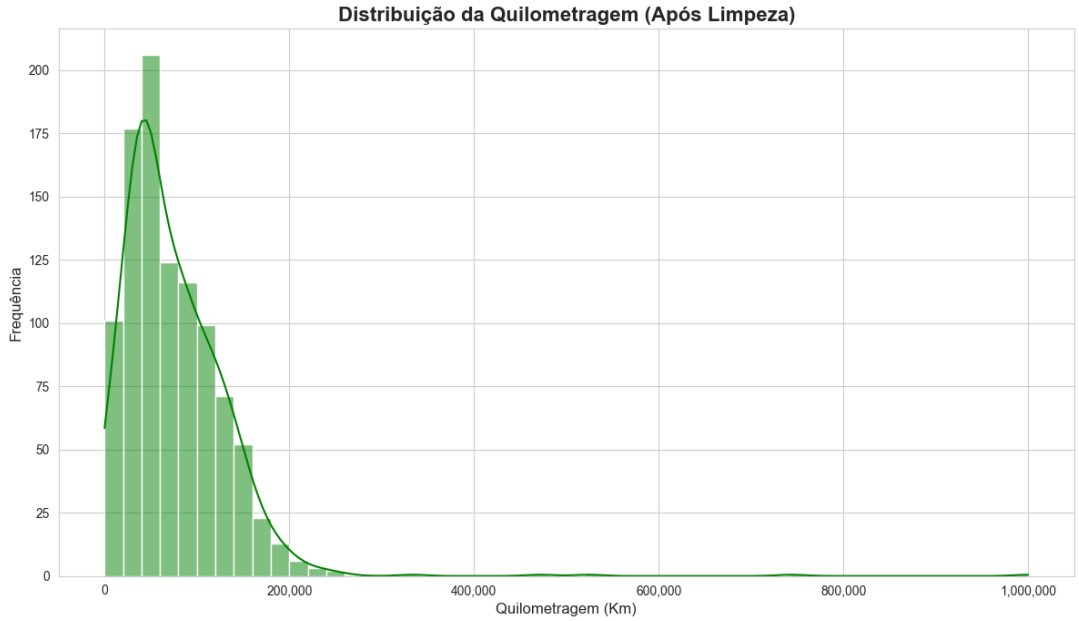

Figura 2: Distribuição da Quilometragem APÓS a limpeza. (Gráfico claro e informativo)

Ao final desta etapa, o dataset totalmente limpo foi salvo como **icarros_final_limpo.csv**, servindo como base para todas as análises subsequentes.

2.2.3. Engenharia de Atributos (Discretização e Normalização)

Embora não fossem necessárias para o Teste t subsequente, as técnicas de Discretização e Normalização foram aplicadas para demonstrar uma etapa mais avançada de pré-processamento, útil para futuras análises de Machine Learning.

- **Discretização:** A variável contínua preço foi transformada em uma variável categórica (faixa_preço), com as seguintes faixas: "Popular" (até R$ 70k), "Intermediário" (de R$ 70k a R$ 120k) e "Premium" (acima de R$ 120k). A distribuição dos veículos nessas faixas pode ser vista na Figura 3.

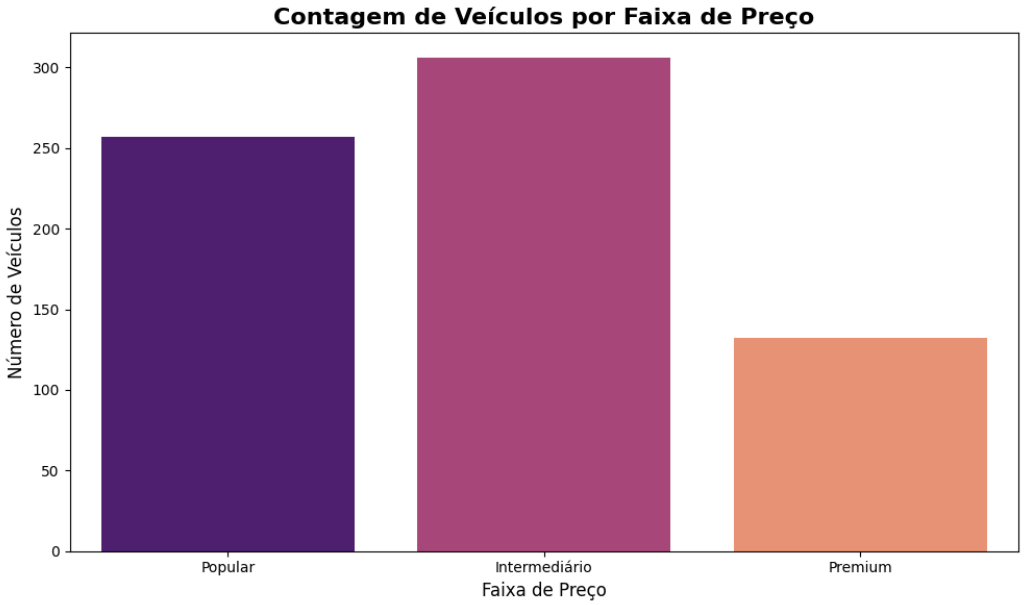

Figura 3: Contagem de Veículos por Faixa de Preço Discretizada.

- **Normalização:** As variáveis preço e km foram normalizadas pela técnica Min-Max, que ajusta seus valores para uma escala comum de 0 a 1. Isso cria as colunas preço_norm e km_norm, preparando os dados para algoritmos sensíveis à escala.

O resultado deste enriquecimento foi salvo em um novo arquivo, **icarros_com_features.csv**, para não interferir na análise principal.

3. Análise Exploratória de Dados (AED)

Com um dataset limpo e confiável, a análise exploratória teve como objetivo descrever o perfil da amostra de veículos.

3.1. Estatísticas Descritivas

As principais métricas para preço e quilometragem foram:

Métrica	Preço (R$)	Quilometragem (Km)
Média	95.512	75.610
Mediana (50%)	79.990	63.731

Métrica	Preço (R$)	Quilometragem (Km)
Mínimo	14.499	11
Máximo	579.900	999.999

3.2. Distribuição de Preços

O histograma de preços (Figura 3) mostra a distribuição dos valores dos veículos na amostra.

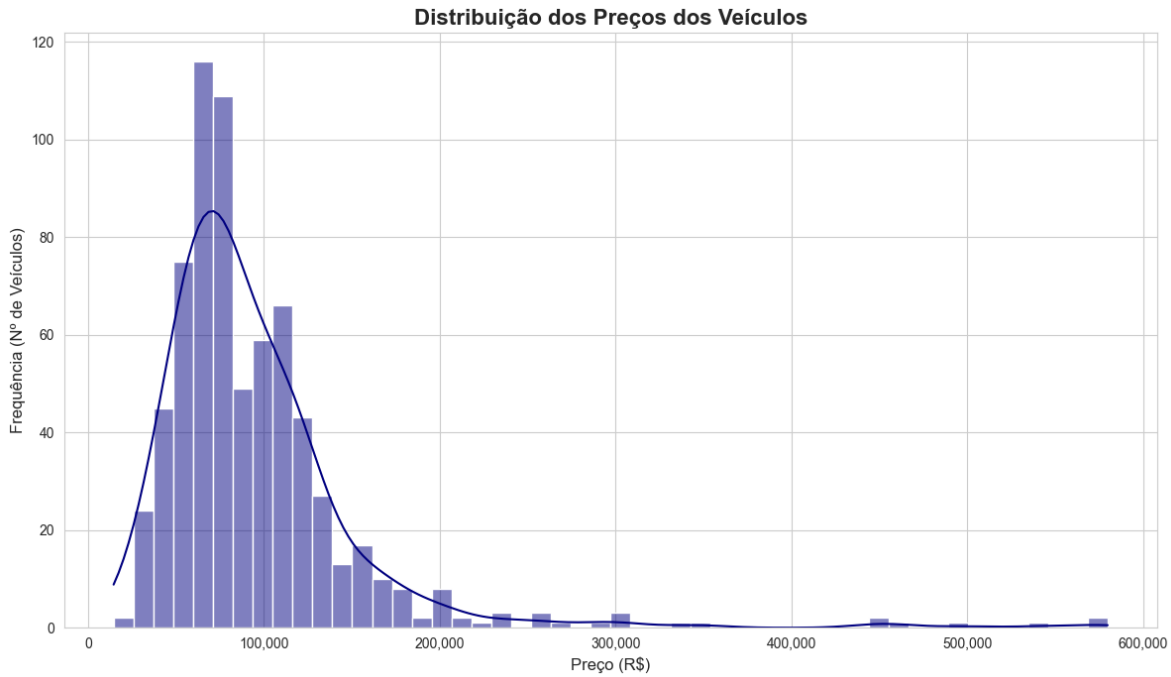

Figura 3: Distribuição dos Preços dos Veículos.

A análise do gráfico indica uma concentração maior de veículos na faixa de R$ 50.000 a R$ 100.000, com a distribuição se estendendo para valores mais altos de forma menos frequente.

4. Análise Inferencial (Teste de Hipóteses)

Etapa final do projeto foi utilizar a estatística para responder a uma pergunta de negócio: **"Existe uma diferença estatisticamente significativa no preço médio dos carros anunciados em São Paulo (SP) em comparação com os do Paraná (PR)?"**

- **Hipótese Nula (H_0):** Não há diferença entre o preço médio dos carros em SP e no PR.
- **Hipótese Alternativa (H_1):** Existe uma diferença entre o preço médio dos carros em SP e no PR.

Foi realizado um **Teste t de Amostras Independentes** utilizando a biblioteca scipy. Os resultados foram:

- **Preço Médio em SP:** R$ 99.261,44
- **Preço Médio no PR:** R$ 80.838,47
- **P-valor:** 0.0161

Como o p-valor (0.0161) é menor que o nível de significância padrão de 0.05, a Hipótese Nula foi rejeitada. Isso significa que a diferença de quase R$ 20.000 entre as médias não é fruto do acaso.

O boxplot abaixo (Figura 4) ilustra essa diferença de forma clara, mostrando que a faixa de preços em São Paulo é consistentemente mais alta que no Paraná.

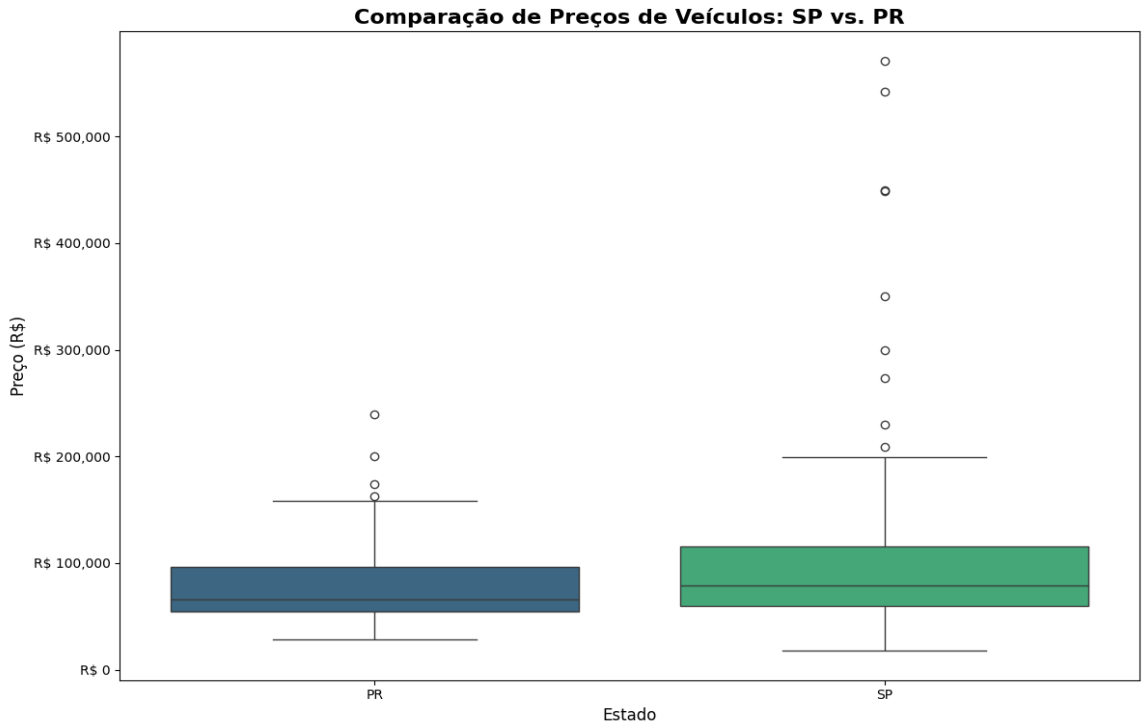

Figura 4: Comparação da Distribuição de Preços entre SP e PR.

5. Conclusão

Este projeto demonstrou com sucesso um ciclo completo de análise de dados. A partir de dados brutos e não estruturados coletados na web, foi possível, através de um processo sistemático de limpeza e análise, chegar a uma conclusão estatisticamente robusta: **existe uma diferença significativa e real no preço médio de veículos seminovos entre os estados de São Paulo e Paraná**, com base na amostra estudada.

O trabalho reforça a importância do pré-processamento como etapa fundamental para garantir a validade de qualquer análise e evidencia como as ferramentas de ciência de dados podem ser aplicadas para extrair insights valiosos de fontes de dados públicas.