ANÁLISE E CLASSIFICAÇÃO DE DADOS DE CARROS USADOS

UTILIZANDO MACHINE LEARNING PARA PREVER PREÇOS

PEDRO HENRIQUE
LUCAS LEITE
NICOLAS GOMES

Sumário

- Introdução
- Visão geral do dataset de carros usados;
- Distribuição de dados;
- Como as variáveis se relacionam;
- Boxplot das variáveis;
- Após a limpeza de dados;
- Comparação com dados filtrados;
- Como prever os preços dos carros;
- Resultados;
- Optuna;
- Análise final.

Introdução

- As informações foram extraídas do site cars.com, um dos principais portais de compra e venda de automóveis nos Estados Unidos.
- Os dados foram reunidos em abril de 2023, refletindo o mercado de carros usados naquele período.
- Com 762.091 registros, o dataset oferece uma amostra significativa para análises estatísticas e desenvolvimento de modelos de aprendizado de máquina.

Visão geral do dataset

O dataset contém informações de carros usados, como preço, quilometragem, ano, etc.

0	manufacturer
1	model
2	year
3	mileage
4	engine
5	transmission
6	drivetrain
7	fuel_type
8	mpg
9	exterior_color
10	interior_color

11	accidents_or_damage
12	one_owner
13	personal_use_only
14	seller_name
15	seller_rating
16	driver_rating
17	driver_reviews_num
18	price_drop
19	price

Distribuição dos dados

Valores mínimos, médios e máximos das colunas principais;

Colunas:

- preço (price)
- queda de preços (price_drop)

Sem filtragem de dados:

	Min	Mean	Max
price	1.0	36485.494589	1.000000e+09
price_drop	100.0	996.812298	9.000000e+04

Distribuição dos dados

Após a filtragem de dados:

```python
# Filtrando os dados para remover valores de
cars = cars[cars['price'] < 1000000]
print(cars['price'].describe())
```

```
count    730578.000000
mean      32328.831565
std       21882.092293
min           1.000000
25%       19639.000000
50%       27981.000000
75%       39052.250000
max      899975.000000
Name: price, dtype: float64
```

```python
# Filtrando os dados para remover valores de price_drop
cars = cars[cars['price_drop'] < 60000]

# Verificando os valores estatísticos após o filtro
print(cars['price_drop'].describe())
```

```
count    730601.000000
mean        996.282803
std         952.329375
min         100.000000
25%         539.000000
50%         996.812298
75%         996.812298
max       56000.000000
Name: price_drop, dtype: float64
```

Distribuição dos dados

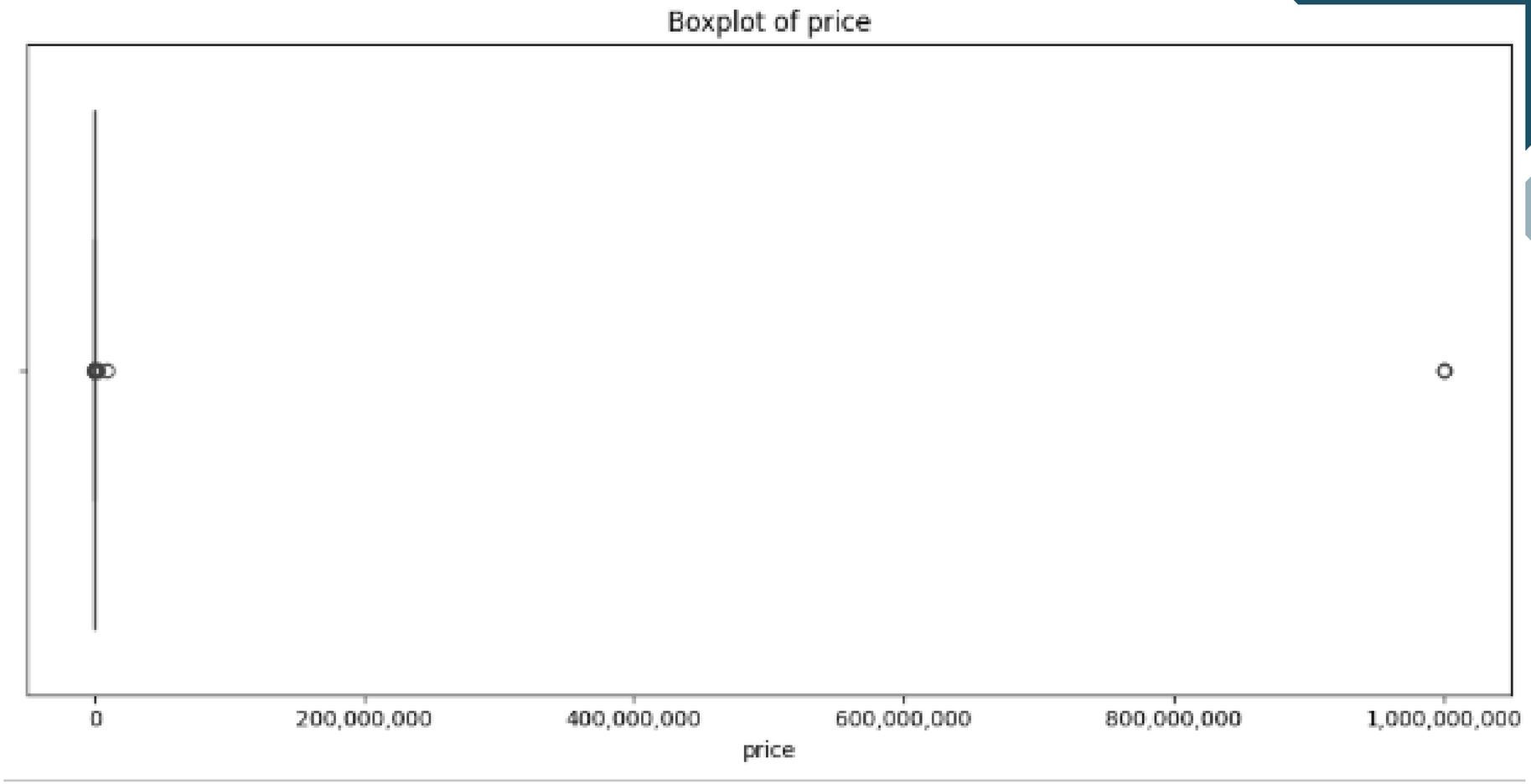
- Distribuição de variáveis numéricas.

Variáveis Numéricas:

- year
- mileage
- seller_rating
- driver_rating
- driver_reviews_num
- price_drop
- price
- accidents_or_damage
- one_owner
- personal_use_only
- mpg

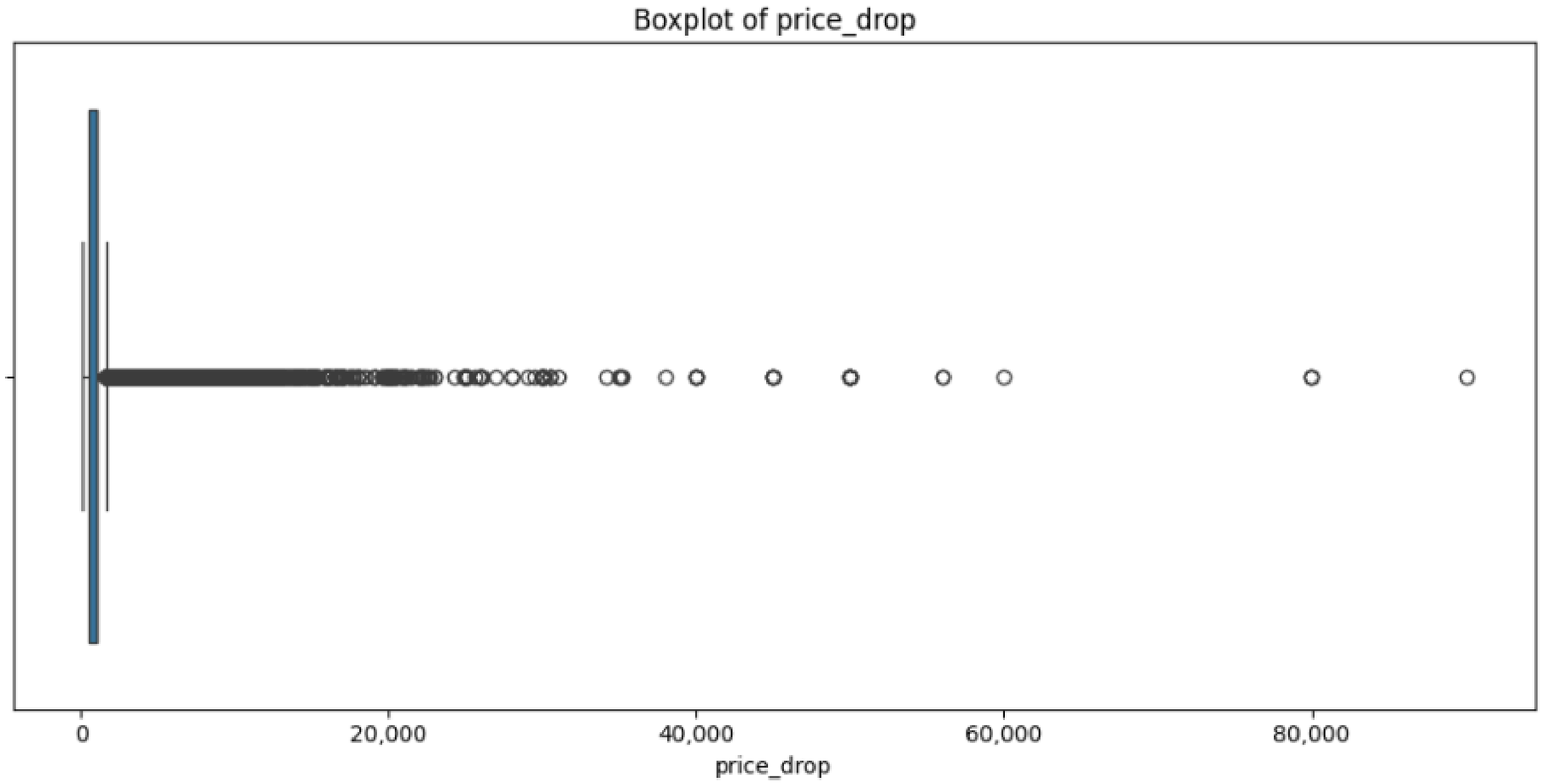
Como as Variáveis se relacionam?

- Boxplot da coluna price



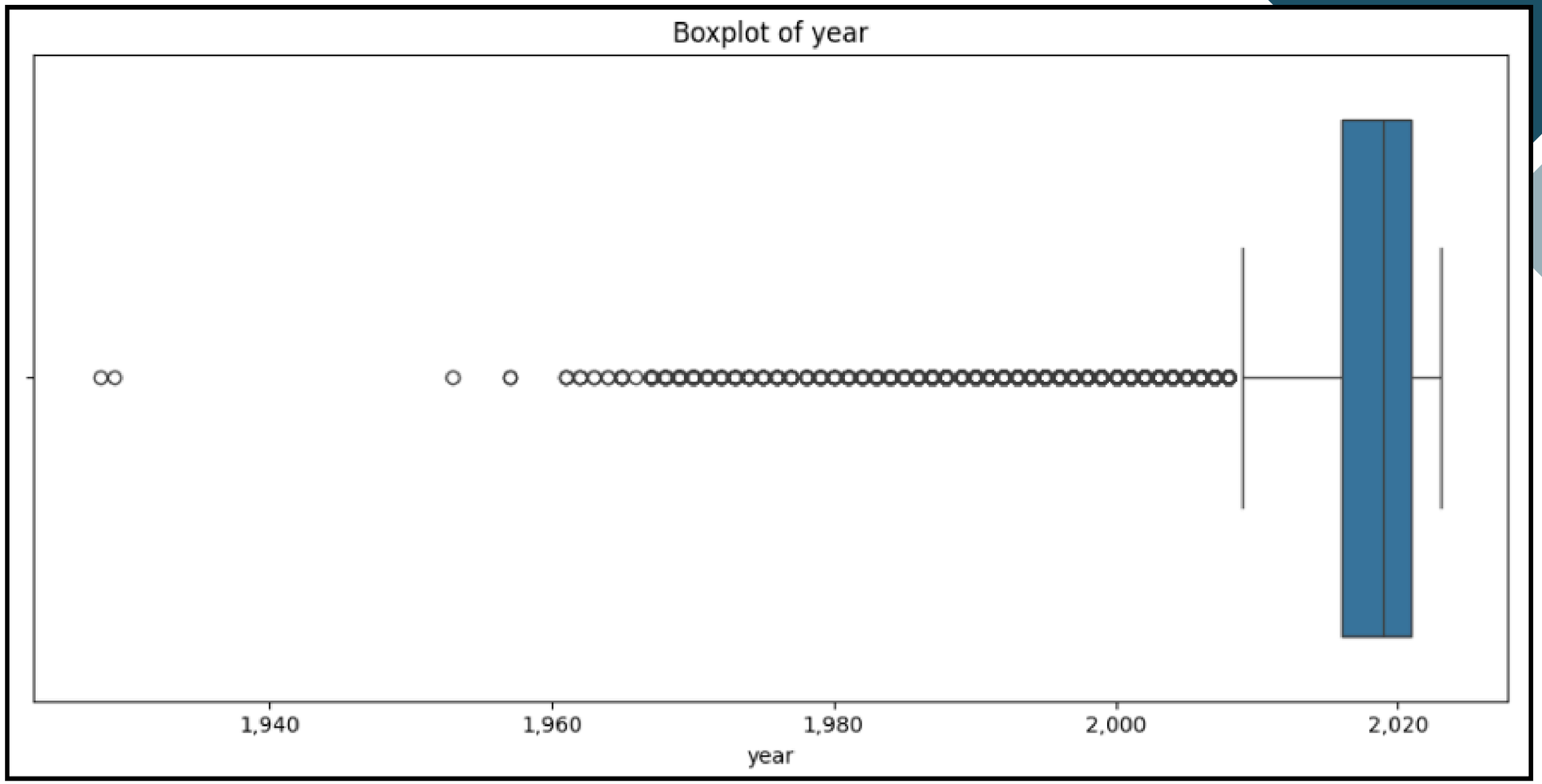
Como as Variáveis se relacionam?

- Boxplot da coluna price_drop



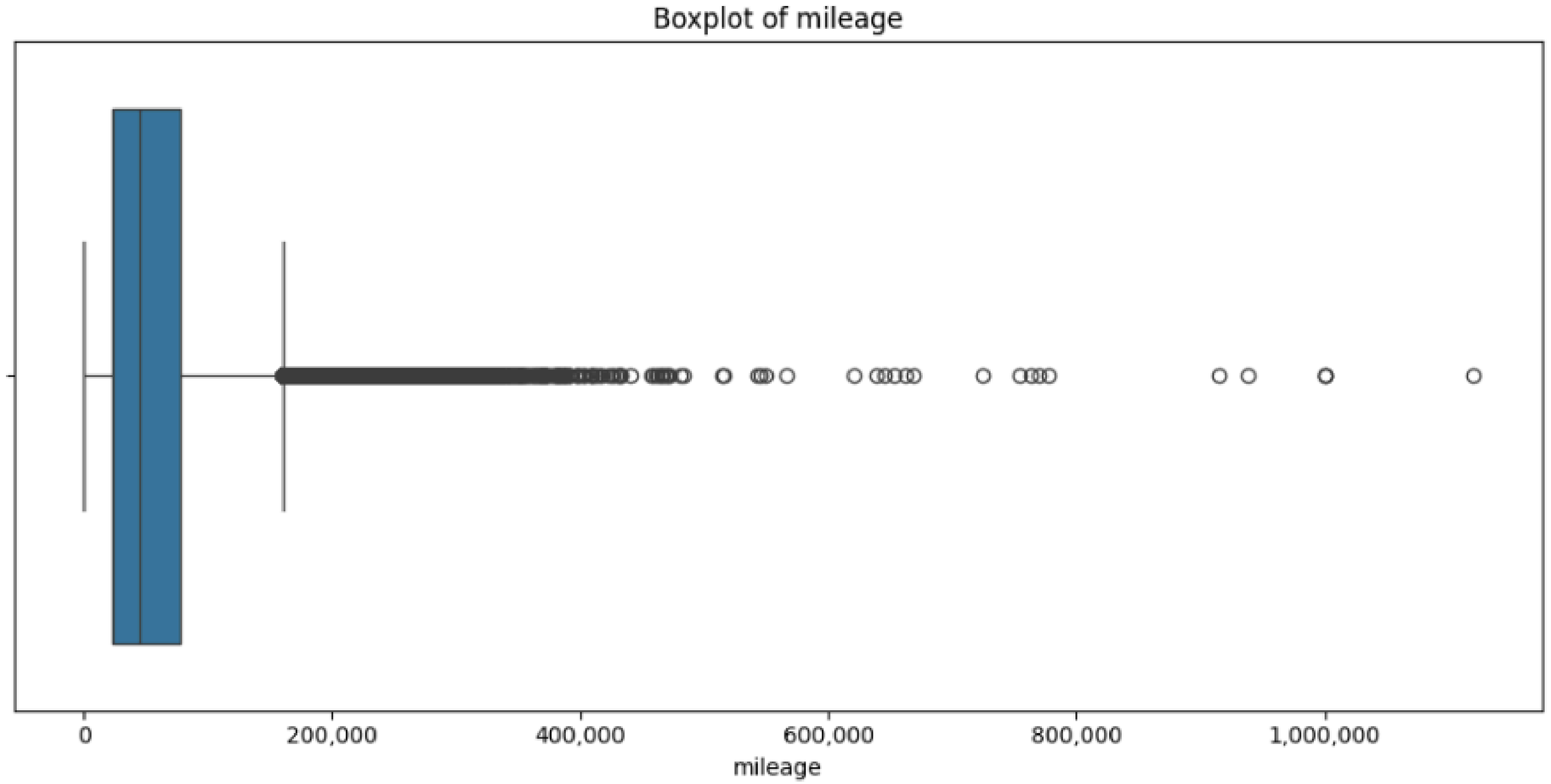
Como as Variáveis se relacionam?

- Boxplot da coluna year



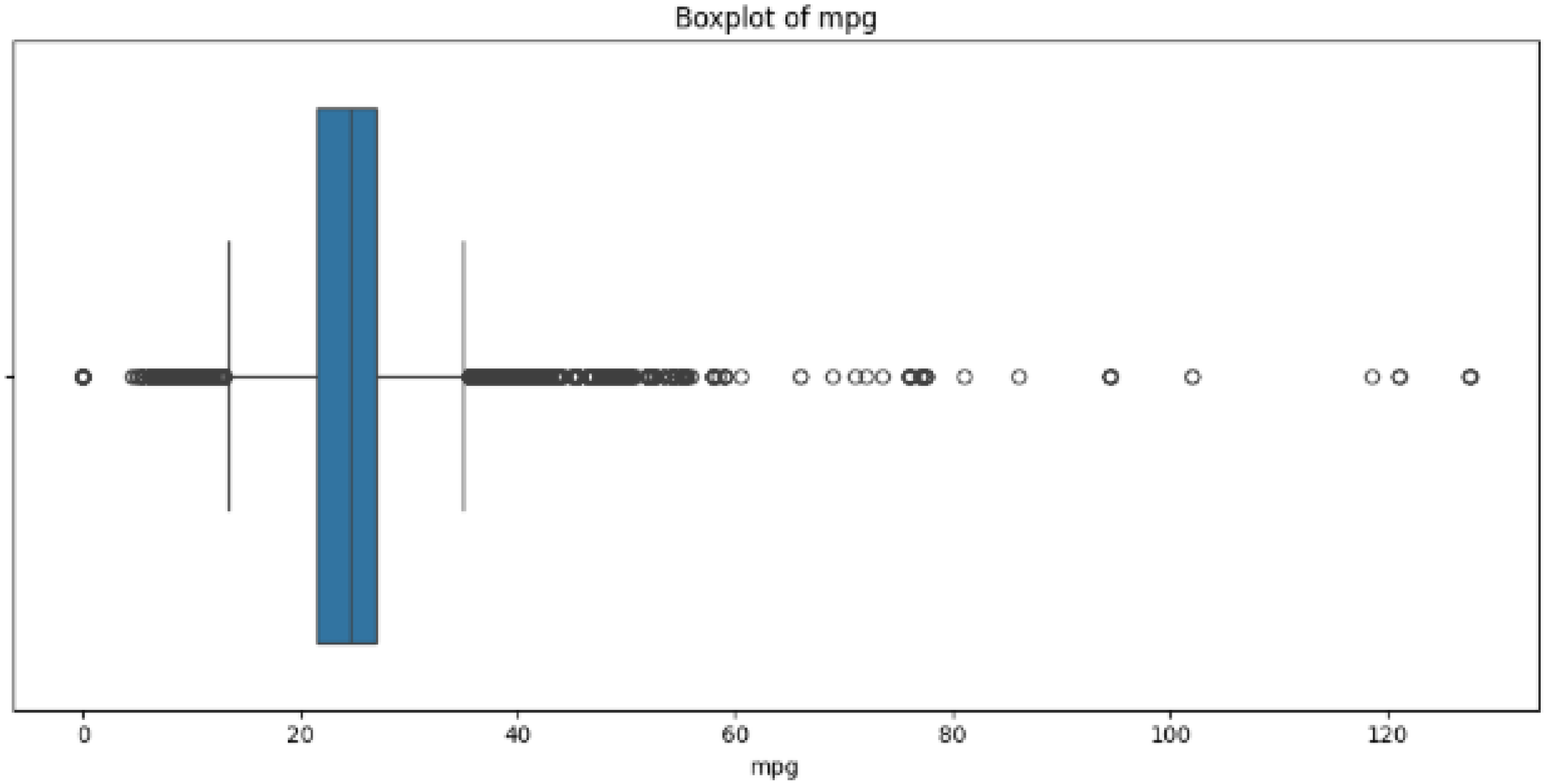
Como as Variáveis se relacionam?

- Boxplot da coluna mileage



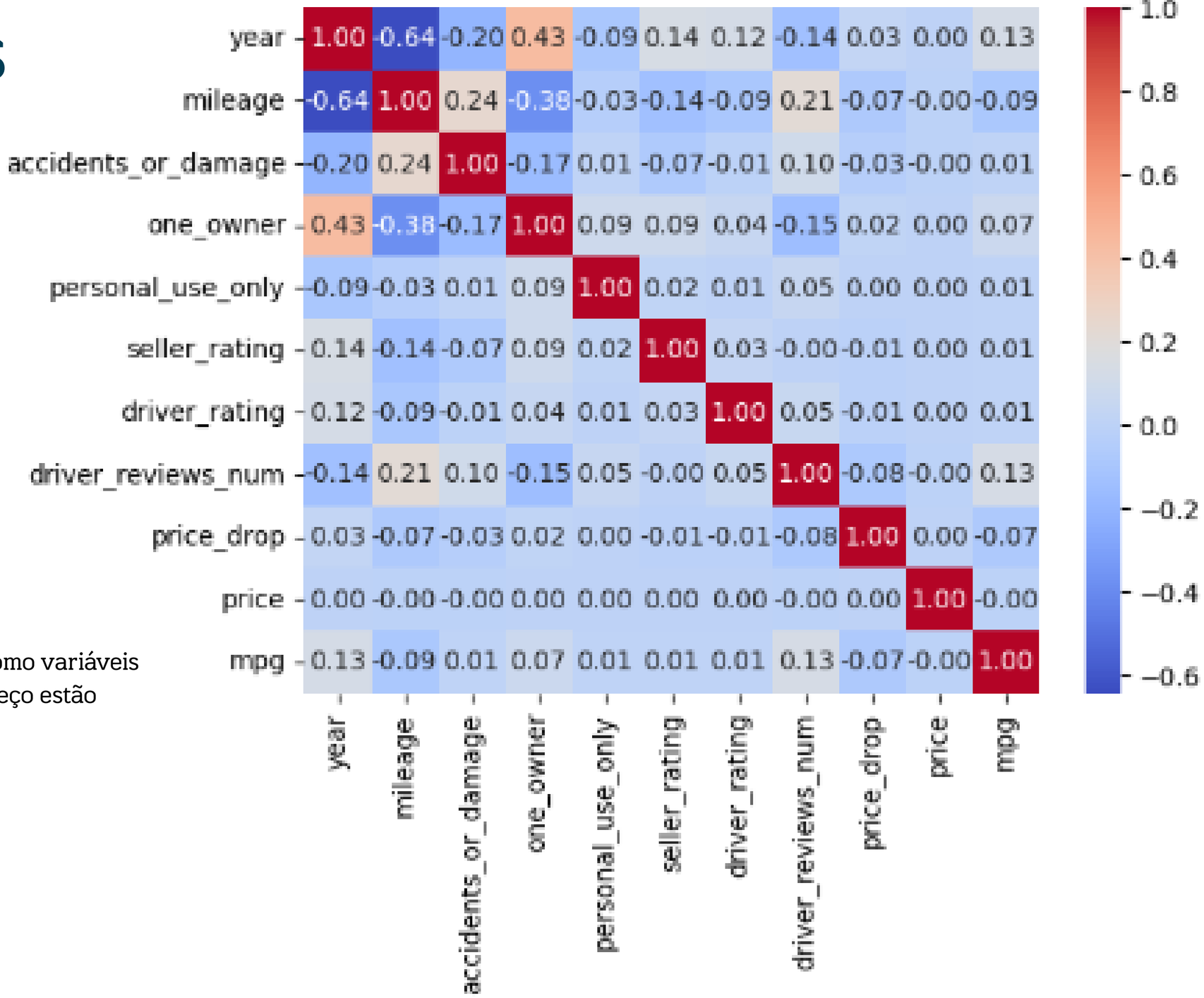
Como as Variáveis se relacionam?

- Boxplot da coluna mpg



Boxplots das variáveis

- Heatmap de correlação



Ao examinar a matriz de correlação, poderá ver como variáveis como o ano de fabricação, quilometragem e preço estão relacionadas.

Após a Limpeza dos Dados

Removemos valores extremos e inconsistentes
Estatísticas atualizadas das colunas price, price_drop, year, mileage

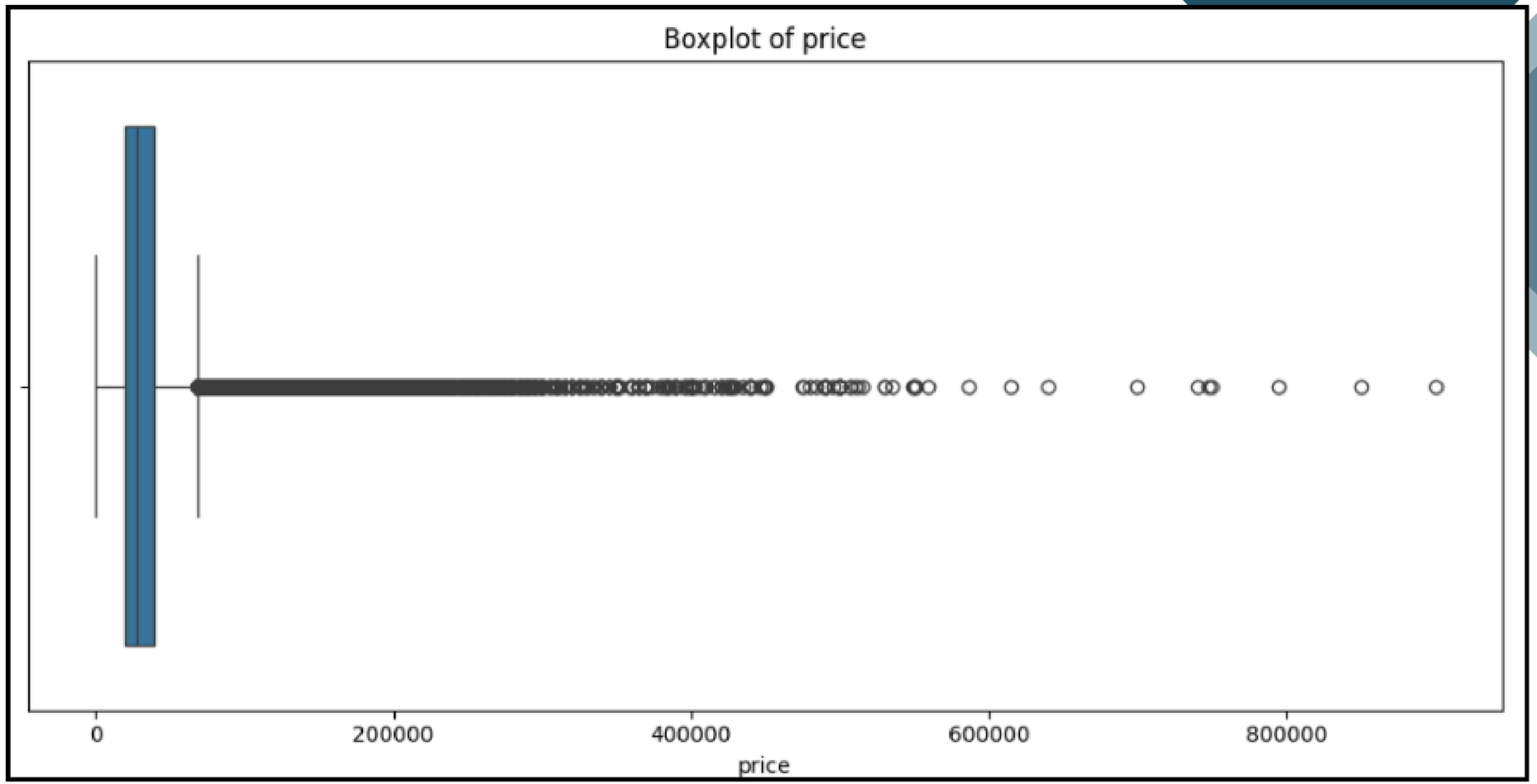
- Distribuição de preços: A distribuição dos preços dos carros pode indicar se há algum valor atípico ou tendências de preços.

- Análise de correlação: Ao examinar a matriz de correlação, você poderá ver como variáveis como o ano de fabricação, quilometragem e preço estão relacionadas.

Comparando com os Dados Filtrados

- Boxplot da coluna price, removeu as linhas onde o valor de price (preço) era maior ou igual a 1.000.000.

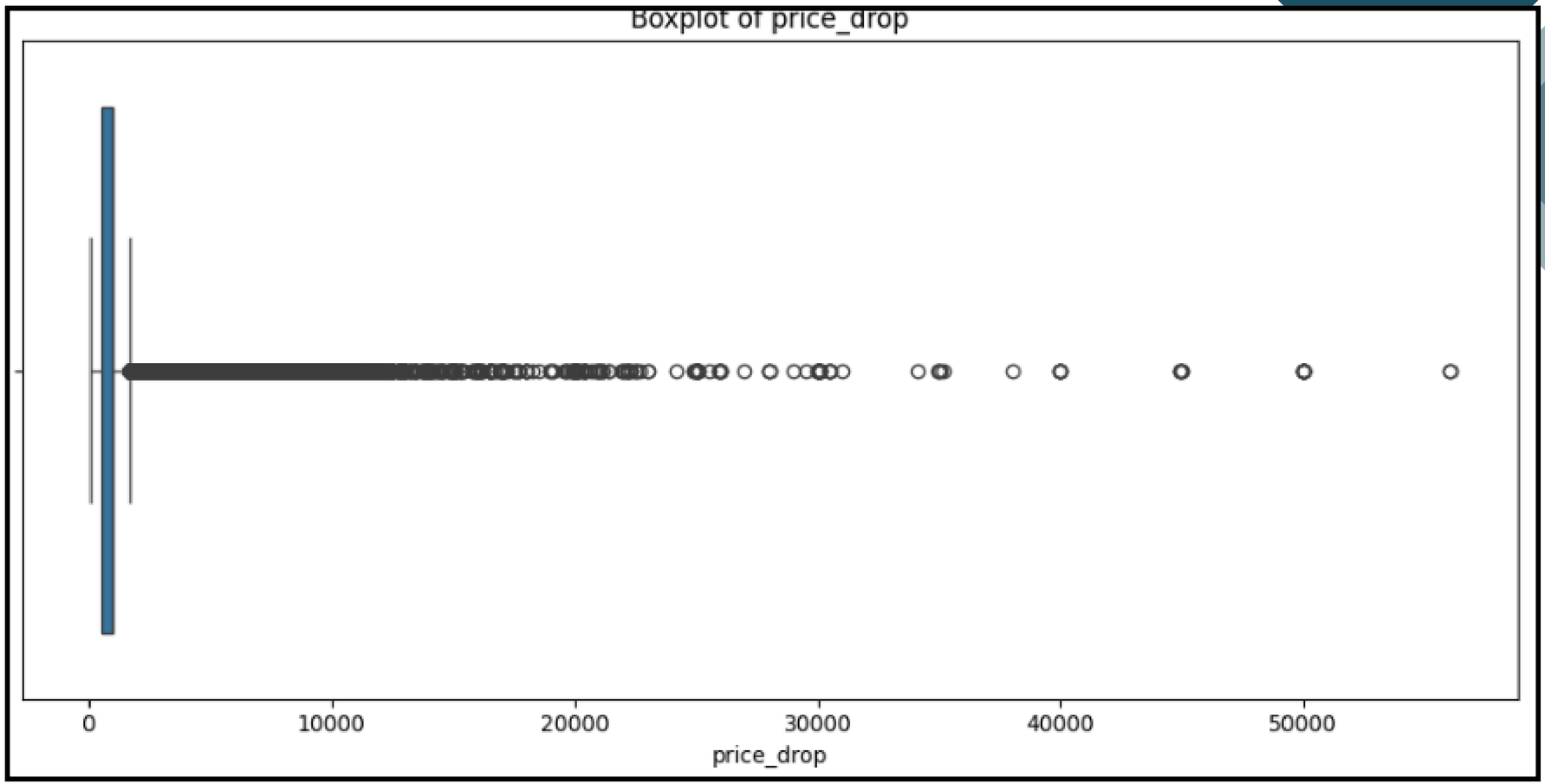
```
count    730578.000000
mean      32328.831565
std       21882.092293
min           1.000000
25%       19639.000000
50%       27981.000000
75%       39052.250000
max      899975.000000
Name: price, dtype: float64
```



Comparando com os Dados Filtrados

- Boxplot da coluna price_drop, removeu as linhas em que o valor de price_drop era maior ou igual a 60.000.

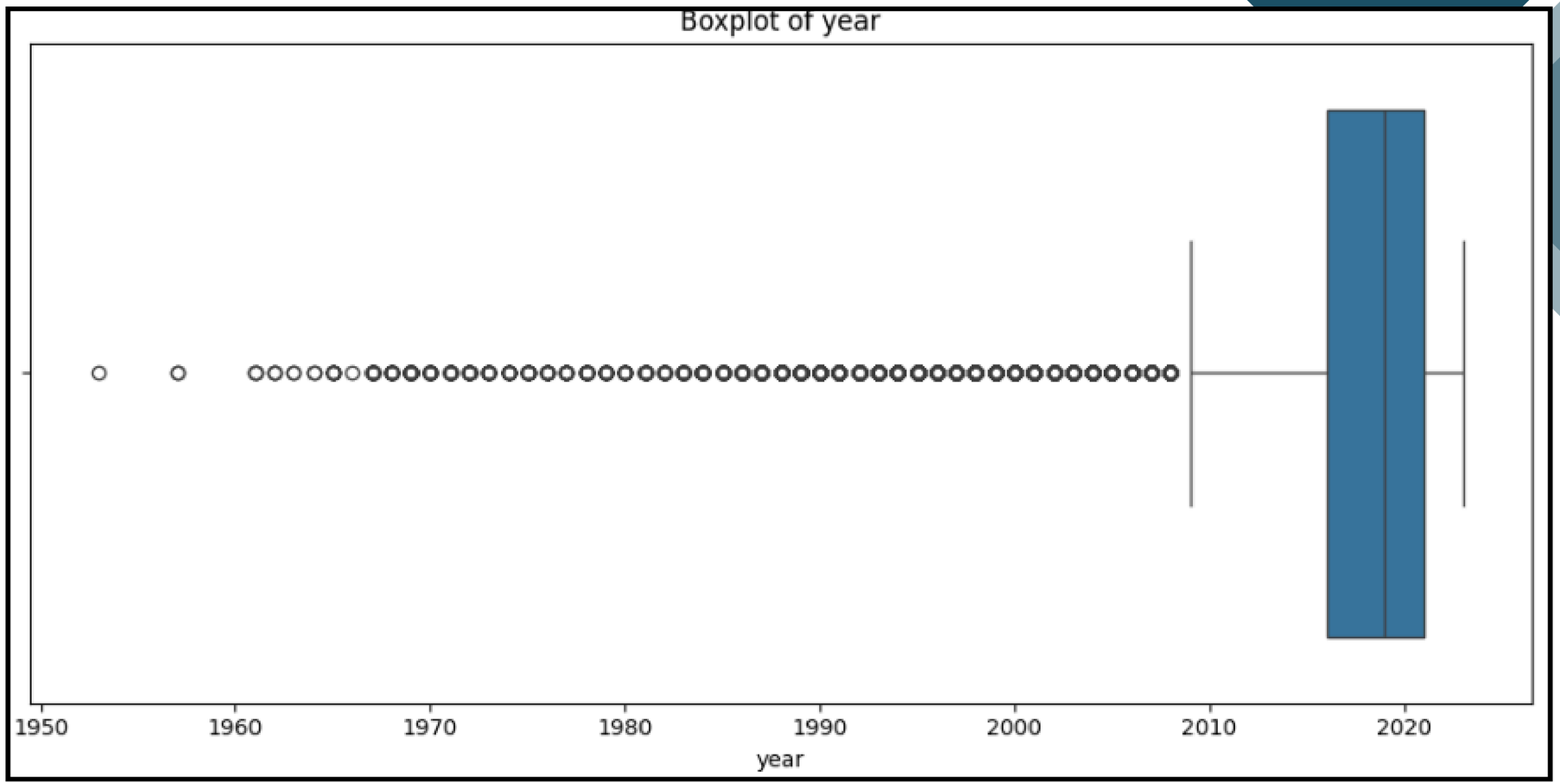
```
count    730601.000000
mean        996.282803
std         952.329375
min         100.000000
25%         539.000000
50%         996.812298
75%         996.812298
max       56000.000000
Name: price_drop, dtype: float64
```



Comparando com os Dados Filtrados

- Boxplot da coluna year, após filtragem manteve apenas os carros com ano de fabricação maior que 1950.

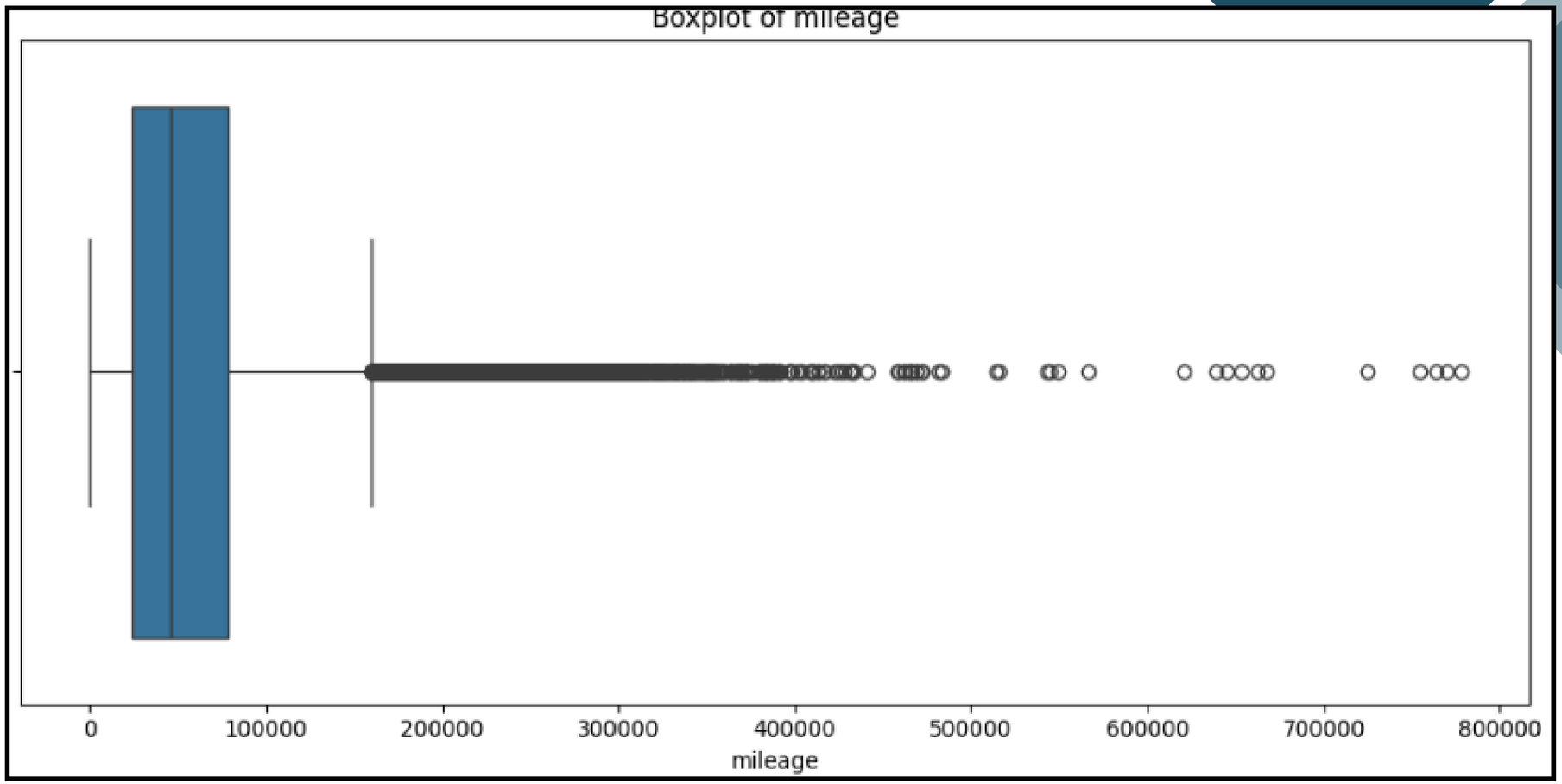
```
count    730606.000000
mean       2017.970682
std           4.233533
min        1953.000000
25%        2016.000000
50%        2019.000000
75%        2021.000000
max        2023.000000
Name: year, dtype: float64
```



Comparando com os Dados Filtrados

- Boxplot da coluna mileage após filtragem, que agora só contém os carros com quilometragem abaixo de 800.000.

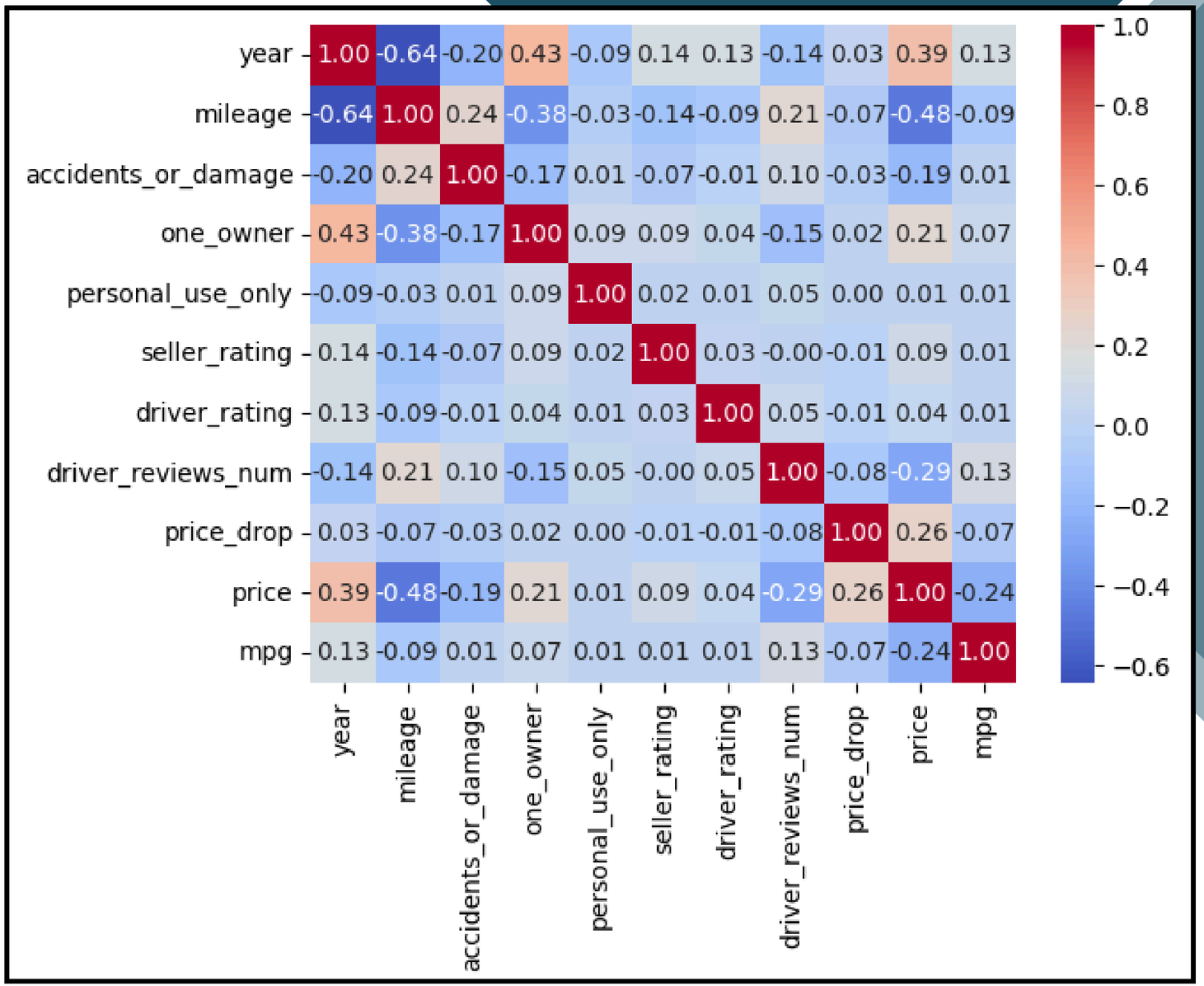
```
count    730572.000000
mean      55747.607651
std       43072.382193
min           0.000000
25%       23542.750000
50%       45655.000000
75%       78136.000000
max      777698.000000
Name: mileage, dtype: float64
```



Comparando com os Dados Filtrados

- Heatmap de correlação atualizado

Ao examinar a matriz de correlação, pode-se perceber como variáveis como o ano de fabricação, quilometragem e preço estão relacionadas.



Como Prever o Preço dos Carros?

O Random Forest é um modelo baseado em múltiplas árvores de decisão

Ele aprende padrões a partir dos dados e melhora a precisão da previsão.

```
# Inicializando o modelo de Random Forest
rf_model = RandomForestRegressor(n_estimators=50, max_depth=10, random_state=42)

# Treinando o modelo
rf_model.fit(X_train, y_train)

# Fazendo previsões no conjunto de teste
y_pred_rf = rf_model.predict(X_test)

# Avaliando o modelo
mae = mean_absolute_error(y_test, y_pred_rf)  # Erro absoluto médio
r2_rf = r2_score(y_test, y_pred_rf)  # Coeficiente de determinação R²

# Exibindo os resultados
print(f"Erro absoluto médio (MAE) com Random Forest: {mae}")
print(f"Coeficiente de determinação (R²) com Random Forest: {r2_rf}")
```

Resultados do Random Forest

- Erro absoluto médio (MAE): 3672.31
- O MAE representa o erro médio entre os valores reais e as previsões do modelo. Por exemplo, se um carro custa R$ 80.000, a previsão do modelo pode variar entre R$ 76.328 e R$ 83.672.
- Coeficiente de determinação (R^2): 0.89
- Indica a proporção da variação nos dados dependentes que pode ser explicada pelo modelo.

Optuna

Optuna é uma biblioteca de otimização automática de hiperparâmetros para machine learning.

number	value	datetime_start	datetime_complete	duration	params_max_depth	params_min_samples_split	params_n_estimators	state
0	3141.411078595406	2025-03-16 10:34:37.982334	2025-03-16 10:47:13.485655	0 days 00:12:35.503321	13	2	98	COMPLETE
1	6541.538386521133	2025-03-16 10:47:13.486653	2025-03-16 10:51:09.844524	0 days 00:03:56.357871	3	3	118	COMPLETE
2	2753.1480073444163	2025-03-16 10:51:09.846524	2025-03-16 10:55:03.957591	0 days 00:03:54.111067	17	5	25	COMPLETE
3	3664.528224755742	2025-03-16 10:55:03.959592	2025-03-16 11:04:54.730595	0 days 00:09:50.771003	10	6	98	COMPLETE
4	5651.774777552023	2025-03-16 11:04:54.732593	2025-03-16 11:07:42.988594	0 days 00:02:48.256001	4	2	64	COMPLETE
5	3013.559484936374	2025-03-16 11:07:42.989591	2025-03-16 11:17:56.141592	0 days 00:10:13.152001	14	9	77	COMPLETE
6	4683.390333136438	2025-03-16 11:17:56.142590	2025-03-16 11:30:46.630535	0 days 00:12:50.487945	6	7	197	COMPLETE
7	2813.3078897658656	2025-03-16 11:30:46.631537	2025-03-16 11:38:29.091714	0 days 00:07:42.460177	16	9	52	COMPLETE
8	2672.6769774182912	2025-03-16 11:38:29.092716	2025-03-16 11:48:05.530918	0 days 00:09:36.438202	18	7	59	COMPLETE
9	5651.3653883922525	2025-03-16 11:48:05.531918	2025-03-16 11:52:26.429539	0 days 00:04:20.897621	4	7	98	COMPLETE
10	2597.0926017754055	2025-03-16 11:52:26.430538	2025-03-16 12:17:53.756125	0 days 00:25:27.325587	20	10	149	COMPLETE
11	2596.7579185116274	2025-03-16 12:17:53.758129	2025-03-16 12:44:09.815682	0 days 00:26:16.057553	20	10	154	COMPLETE
12	2596.2514752375655	2025-03-16 12:44:09.816685	2025-03-16 13:11:09.085052	0 days 00:26:59.268367	20	10	158	COMPLETE
13	2595.9018589461807	2025-03-16 13:11:09.087055	2025-03-16 13:39:40.085766	0 days 00:28:30.998711	20	10	167	COMPLETE
14	3663.712005650104	2025-03-16 13:39:40.087763	2025-03-16 15:20:19.941336	0 days 01:40:39.853573	10	9	183	COMPLETE
15	2893.4262570591413	2025-03-16 15:20:19.943339	2025-03-16 15:42:14.509732	0 days 00:21:54.566393	15	8	157	COMPLETE
16	2672.9791549017314	2025-03-16 15:42:14.512735	2025-03-16 16:02:30.003591	0 days 00:20:15.490856	18	10	127	COMPLETE
17	3298.0873071335045	2025-03-16 16:02:30.005593	2025-03-16 16:21:41.267715	0 days 00:19:11.262122	12	4	165	COMPLETE
18	2590.4391095423375	2025-03-16 16:21:41.271769	2025-03-16 16:46:44.009999	0 days 00:25:02.738230	20	8	137	COMPLETE
19	4385.3073382072425	2025-03-16 16:46:44.011995	2025-03-16 16:56:28.769777	0 days 00:09:44.757782	7	8	135	COMPLETE

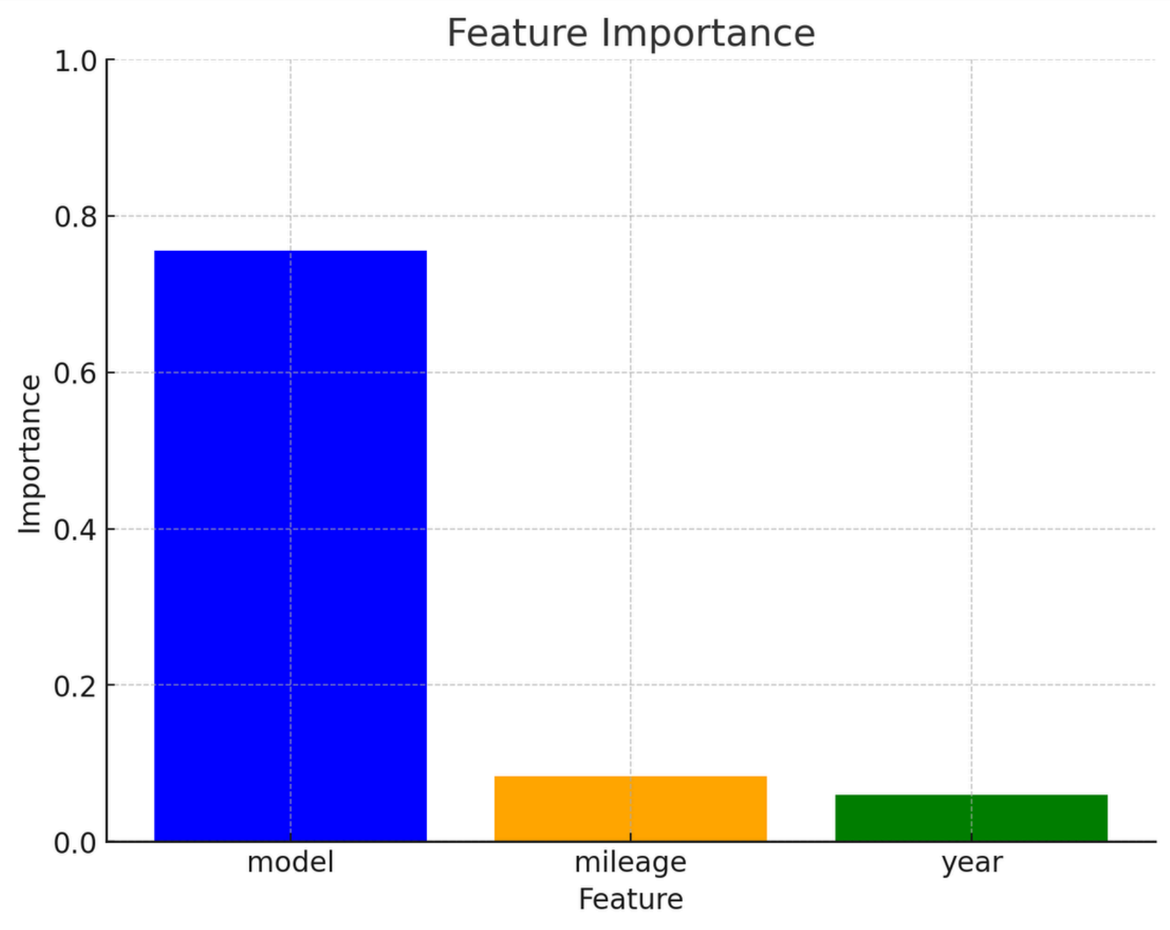
O Modelo Foi Preciso?

Métricas de desempenho:
Erro absoluto médio (MAE): 2665.68

O MAE representa o erro médio entre os valores reais e as previsões do modelo. Por exemplo, se um carro custa R$ 80.000, a previsão do modelo pode variar entre R$ 77.335 e R$ 82.665.

Coeficiente de determinação (R^2): 0.93
indica a proporção da variação nos dados dependentes que pode ser explicada pelo modelo.

O Que Mais Impacta no Preço?



Distribuição dos dados

✓ Conseguimos prever preços de carros usados com boa precisão
✓ O modelo identificou que o modelo do carro, a quilometragem e ano são fatores essenciais

Obrigado!