Escola de Engenharia
Universidade do Minho

Mestrado Integrado em Engenharia de
Gestão e Sistemas de Informação
2023/2024

4º Ano

1º Semestre

Aprendizagem Automática em Sistemas de Informação

Francisco Miguel Pinheiro Cardoso

A79570

Índice

1. Introdução

O presente relatório realiza-se no âmbito da unidade curricular "Aprendizagem Automática em Sistemas Empresariais", lecionada no 1º ano do Mestrado em Engenharia e Gestão de Sistemas de Informação, tendo como intuito a aplicação da metodologia CRISP-DM "Cross Industry Standard Process for Data Mining" para compreensão dos conceitos, princípios e recursos associadas ao Data Mining.

A partir da base de dados facultada pelo docente - Used Car Price Prediction Dataset -, a qual têm informação relacionada com a indústria automóvel (tendências, preferências do consumidor, etc.), procurei encontrar soluções que permitem aumentar o volume de negócio.

De acordo com a metodologia CRISP-DM, a análise a ser efetuada divide-se em 6 fases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evalution e Deployment.

O presente relatório apenas abrange o Business Understanding, centrado em entender os objetivos, requisitos e critérios de sucesso do negócio, a fim de desenvolver um plano inicial que permita satisfazer os clientes. Assim sendo, subdividi o relatório em quatro subseções. Na primeira subseção, enquadra-se os objetivos do negócio e refere-se qual o critério de sucesso. Na segunda subseção, expõe-se quais os recursos disponíveis para a elaboração do presente projeto, os requisitos que têm de ser cumpridos e os pressupostos que permitem validar os resultados e as restrições que podem ocorrer no desenvolvimento do projeto. Na terceira subseção, apresenta-se os objetivos de data mining e o critério de sucesso do mesmo. Por último, clarifica-se o plano de projeto e quais as ferramentas a serem utilizadas para o seu desenvolvimento.

2. Business Understanding

A Compreensão do Negócio é a fase inicial da metodologia CRISP-DM, que se foca em compreender os objetivos do projeto e requisitos de uma perspetiva de negócio, para mais tarde converter este conhecimento numa definição de um problema de data mining e um plano preliminar desenhado para alcançar os objetivos.

2.1 Determine Business Objetives

2.1.1 Background

Cars.com é líder no mercado digital e fornecedor de soluções para a indústria automóvel que interliga compradores e vendedores de automóveis. A empresa fornece aos compradores dados, recursos e ferramentas digitais necessárias para tomarem decisões de compras informadas e estabelecerem uma ligação com os retalhistas. Num mercado em constante mudança, a Cars.com dispõe de soluções técnicas inovadoras e de informações baseadas em dados para conseguirem alcançar e influenciar as pessoas prontas a comprar, aumentarem a circulação de inventário e ganharem quota de mercado.

Em 2018, a Cars.com adquiriu Dealer Inspire, uma empresa de tecnologia inovadora que desenvolve soluções que preparam os concessionários para o futuro com operações mais eficientes, um processo de compra de automóveis mais rápido e fácil, assim como experiências digitais conectadas que vendem e fazem manutenção de mais veículos.

Cars.com inventou a pesquisa de automóveis. O seu website e as soluções inovadoras fazem o contacto entre o comprador e o vendedor. A empresa conta com colaboradores espalhados pelos Estados Unidos da América. Ao fim de muitos anos, continuam com uma cultura de start-up com inovação e paixão pelos colaboradores no centro do negócio.

A Cars.com é uma marca premiada, uma equipa de liderança que conta com os melhores e mais brilhantes funcionários da indústria. Foram considerados um dos melhores locais para trabalhar peço The Chicago Tribune, Built in Chicago e Chicago Innovation.

2.1.2 Business Objetives

De forma a atingir a solução pretendida, é necessário definir objetivos de negócio, para tal, defini os seguintes:

- Melhorar a experiência no website, tanto do vendedor como do comprador;
- Expandir os serviços oferecidos relacionados com o negócio automóvel;
- Apostar na sustentabilidade e promover o negócio com foco nos veículos elétricos;
- Aumentar a base de clientes, utilizar técnicas para adquirir novos clientes e manter os existentes.

2.1.3 Business Success Criteria

Com o intuito de aumentar os resultados da empresa na perspetiva do negócio, é necessário definir critérios de sucesso. Com os dados fornecidos foi possível definir um critério de sucesso:

- Aumentar as vendas dos carros mais antigos no website.

Para tal, utilizarei os dados fornecidos para saber quais os carros com melhores condições, melhores preços, para que possam ser mais facilmente promovidos no website.

2.2 Assess Situation

2.2.1 Inventory of Resources

Os recursos disponíveis para a realização deste projeto, incluem:

- Pessoal: A realização do projeto dispõe de um aluno, do Mestrado Integrado em Engenharia de Gestão de Sistemas de Informação, com alguma experiência em data mining, análise de negócio e dados.
- Dados: Os dados são fornecidos no website kaggle, onde contêm as características sobre cada viatura disponível no website, num ficheiro em formato .csv.
- Hardware: O aluno possui dois computadores para a análise e tratamento de dados.
- Software: O aluno tem disponível para utilizar neste projeto, ferramentas como Jupyter para programação em Python e RapidMiner para a realização de datamining. Para o processamento de dados tem disponível o Talend, para a modelação dos modelos e dashboards o Tableau e por fim, o Microsoft Word e Excel para documentação e arquivo de dados respetivamente.

2.2.2 Requirements, Assumptions, and Constrains

Neste projeto, existem requisitos que têm de ser cumpridos, assim como pressupostos que permitem validar os resultados e por fim, restrições que podem ocorrer no desenvolvimento do projeto.

Restrições:

- Realizar as entregas nos prazos estipulados;
- Apresentar resultados compreensíveis e com qualidade;
- Utilizar a metodologia CRISP-DM;
- Utilizar as devidas ferramentas projetadas.

Pressupostos:

- Os datasets têm de apresentar dados reais;
- Os dados não podem apresentar erros;
- Os datasets têm de ser suficientes para responder aos requisitos do projeto.

Restrições:

- Pouca experiência em Data Mining;
- Sobrecarga de trabalho de grupo, sendo um só aluno a realizar;
- Pouca experiência na metodologia CRIPS-DM;
- Pouca experiência nas ferramentas a ser utilizadas.

2.2.3 Risks and Contingencies

Riscos	**Consequência**	**Impacto**	**Contingência**
Inexperiência na utilização das ferramentas	Fraca evolução no desenvolvimento do projeto	4	Pesquisa e prática na utilização das ferramentas; solicitar apoio ao docente
Elevada sobrecarga de trabalho	Fraca demonstração de resultados	4	Boa gestão e organização de tempo
Inexperiência da metodologia CRISP-DM	Incorreto desenvolvimento do projeto	4	Revisão constante da metodologia e solicitar apoio ao docente

2.2.4 Terminology

O projeto dispõe de terminologias relevantes compostas por duas componentes, entre elas de negócio e data mining.

- CRISP-DM: "Cross Industry Standard Process for Data Mining" é uma metodologia utilizada para estruturar projetos de data mining. Fornece uma estrutura abrangente para planear, implementar e avaliar o processo de data mining, constituído por seis fases.
- Data Mining: Processo de descobrir informações, padrões e conhecimentos em grandes conjuntos de dados. Envolve a utilização de técnicas computacionais para analisar dados e extrair informação significativa. Utilizada para tomadas de decisões, identificação de tendências, previsões e otimizações.
- DataSet: Conjunto de dados estruturados por colunas e linhas, onde cada coluna representa uma variável e cada linha corresponde a um determinado conjunto de dados.

2.2.5 Costs and Benefits

Este projeto não dispõe de custos, pois encontra-se associado a uma unidade curricular de nível académico. A nível de benefícios, encontra-se o aproveitamento da utilização da metodologia CRIPS-DM, assim como a utilização de ferramentas tecnológicas.

Caso fosse necessário, avaliar os custos no caso de projeto ser real, seriam identificados custos de recursos humanos, custos de software e hardware, assim como infraestruturas necessárias, entre outros. Já os benefícios seriam aplicados à empresa, já que a mesma disponibilizaria do output de projeto, numa maior transação de vendas associadas ao website.

2.3 Determine Data Mining Goals

2.3.1 Data Mining Goals

A fim de determinar os objetivos de data mining, é preciso realizar a análise e tratamento dos dados, entender e compreender de que forma é possível explorar as características de cada veículo de forma que os entusiastas, compradores e investigadores interessados em análises, tomem decisões de compras informadas a nível da indústria automóvel e preferências do consumidor.

2.3.2 Data Mining Success Criteria

Para definir o critério de sucesso data mining, defini um valor mínimo (valor desejável) para um critério baseado na fórmula que irá ser utilizada para a avaliação conhecida como Root Mean Squared Error (RMSE). Esta fórmula representa uma medida que calcula a raiz quadrática média dos erros entre valores e reais e possíveis.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

O critério definido é:

- Valor de RMSE de < 0.2, na atribuição dos preços desejados para os veículos.

2.4 Produce Project Plan

2.4.1 Project Plan

Business Understanding

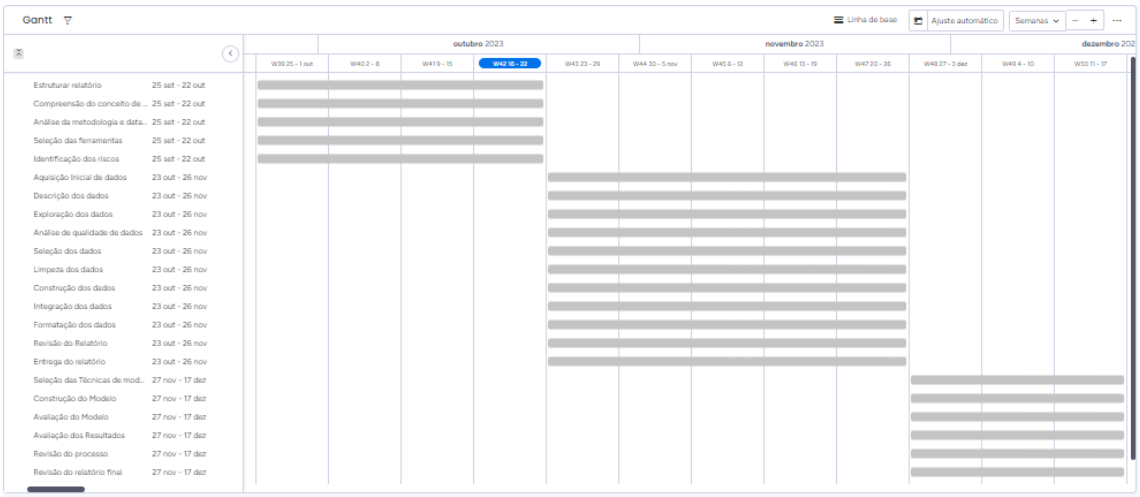
Tarefa	Data	Responsável
Estruturar relatório	25 set - 22 out	Francisco Cardoso
Compreensão do conceit...	25 set - 22 out	Francisco Cardoso
Análise da metodologia e...	25 set - 22 out	Francisco Cardoso
Seleção das ferramentas	25 set - 22 out	Francisco Cardoso
Identificação dos riscos	25 set - 22 out	Francisco Cardoso
Revisão do relatório	25 set - 22 out	Francisco Cardoso
Entrega do relatório	25 set - 22 out	Francisco Cardoso

Data Understanding + Data Preparation

Tarefa	Data	Responsável
Aquisição Inicial de dados	23 out - 26 nov	Francisco Cardoso
Descrição dos dados	23 out - 26 nov	Francisco Cardoso
Exploração dos dados	23 out - 26 nov	Francisco Cardoso
Análise de qualidade de d...	23 out - 26 nov	Francisco Cardoso
Seleção dos dados	23 out - 26 nov	Francisco Cardoso
Limpeza dos dados	23 out - 26 nov	Francisco Cardoso
Construção dos dados	23 out - 26 nov	Francisco Cardoso
Integração dos dados	23 out - 26 nov	Francisco Cardoso
Formatação dos dados	23 out - 26 nov	Francisco Cardoso
Revisão do Relatório	23 out - 26 nov	Francisco Cardoso
Entrega do relatório	23 out - 26 nov	Francisco Cardoso

Modeling + Evaluation

	Tarefa	Data	Responsável
☐	Seleção das Técnicas de ...	27 nov - 17 dez	Francisco Cardoso
☐	Construção do Modelo	27 nov - 17 dez	Francisco Cardoso
☐	Avaliação do Modelo	27 nov - 17 dez	Francisco Cardoso
☐	Avaliação dos Resultados	27 nov - 17 dez	Francisco Cardoso
☐	Revisão do processo	27 nov - 17 dez	Francisco Cardoso
☐	Revisão do relatório final	27 nov - 17 dez	Francisco Cardoso
☐	Entrega do relatório final	27 nov - 17 dez	Francisco Cardoso



2.4.2 Initial Assessment of Tools and Techniques

Neste projeto, irei utilizar diferentes ferramentas para o seu desenvolvimento. Na tabela a seguir, estão representadas as mesmas consoante as suas funcionalidades.

Ferramenta	Funcionalidade
Microsoft Word	Elaboração dos relatórios.
Microsoft Excel	Visualização dos dados utilizados no projeto
Talend	Análise e tratamento de dados.
Rapid Miner	Criar modelos de dados para os requisitos de negócio.
Jupyter Notebook	Executar scripts em Python, para manipulação de dados.
Tableau	Visualização de modelos e dashboards.

3. Data Understanding

Nesta segunda etapa, e segundo a metodologia Crisp-DM o principal objetivo é colecionar dados e posteriormente descrevê-los. Inicialmente temos de verificar se os dados se enquadram com as necessidades do projeto e verificar a qualidade dos mesmos.

3.1 Collect Initial Data

3.1.1 Initial Data Collection Report

Na realização deste projeto foi utilizado 1 dataset denominado de Train, fornecido pelo docente da unidade curricular. O dataset apresenta 3207 exemplos de veículos e 12 atributos sobre os detalhes dos mesmos (brand, model, model_year, mileage, fuel_type, engine, transmission, ext_col, int_col, accident, clean_title, price).

3.2 Describe Data

3.2.1 Data Description Report

Atributo	Descrição	Formato	Quantidade	Exemplos
brand	Marca do veículo	String	3207	Jeep
model	Modelo do veículo	String	3207	Wrangler Sport
model_year	Ano do veículo	Integer	3207	2014
mileage	Distância em miles percorrida pelo veículo	Integer	3207	71,000 mi.
fuel_type	Tipo de combustível do veículo	String	3207	Gasoline
engine	Especificações do motor do veículo	String	3207	285.0HP 3.6L V6 Cylinder Engine Gasoline Fuel
transmission	Tipo de transmissão do veículo	String	3207	5-Speed A/T
ext_col	Cor exterior do veículo	String	3207	Gray
int_col	Cor interior do veículo	String	3207	Black
accident	Histórico de acidentes do veículo	String	3207	None Reported
clean_title	Especificação sobre perda total do veículo	String	3207	Yes
price	Preço do veículo	Integer	3207	22000

3.3 Explore Data

3.3.1 Data Exploration Report

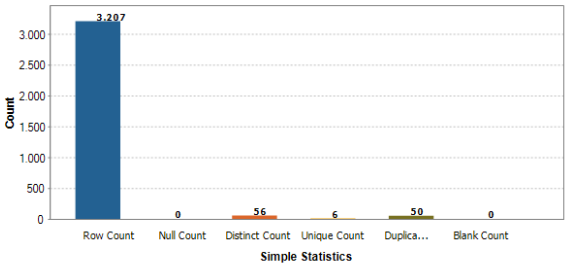
Para a exploração dos dados foi utilizado o Talend Data Quality, de forma a fazer uma análise preliminar dos dados. Esta serviu para encontrar características particulares das variáveis, tais como, padrões, outliers e possíveis anomalias.

3.3.1.1 Análise do atributo "brand"

Column: metadata.brand

Simple Statistics

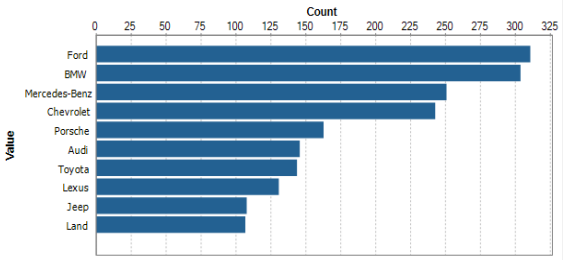
Label	Count	%
Row Count	3207	100.00%
Null Count	0	0.00%
Distinct Count	56	1.75%
Unique Count	6	0.19%
Duplicate Count	50	1.56%
Blank Count	0	0.00%



Column: metadata.brand

Value Frequency

Value	Count	%
Ford	311	9.70%
BMW	304	9.48%
Mercedes-Benz	251	7.83%
Chevrolet	243	7.58%
Porsche	163	5.08%
Audi	146	4.55%
Toyota	144	4.49%
Lexus	131	4.08%
Jeep	108	3.37%
Land	107	3.34%

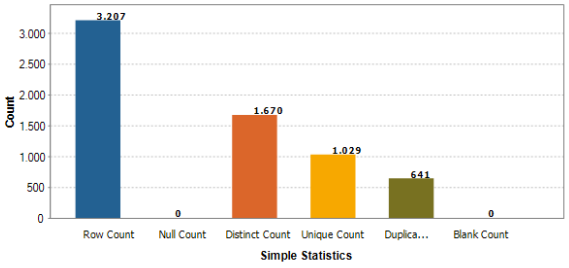


3.3.1.2 Análise do atributo "model"

Column: metadata.model

Simple Statistics

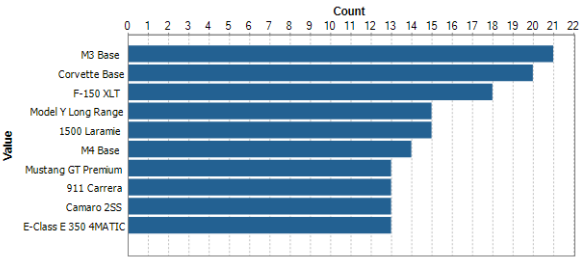
Label	Count	%
Row Count	3207	100.00%
Null Count	0	0.00%
Distinct Count	1670	52.07%
Unique Count	1029	32.09%
Duplicate Count	641	19.99%
Blank Count	0	0.00%



Column: metadata.model

Value Frequency

Value	Count	%
M3 Base	21	0.65%
Corvette Base	20	0.62%
F-150 XLT	18	0.56%
Model Y Long Range	15	0.47%
1500 Laramie	15	0.47%
M4 Base	14	0.44%
Mustang GT Premium	13	0.41%
911 Carrera	13	0.41%
Camaro 2SS	13	0.41%
E-Class E 350 4MATIC	13	0.41%

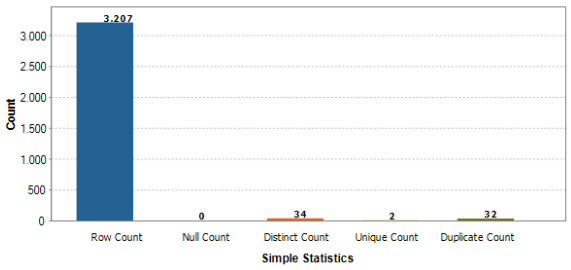


3.3.1.3 Análise do atributo "model_yearl"

Column: metadata.model_year

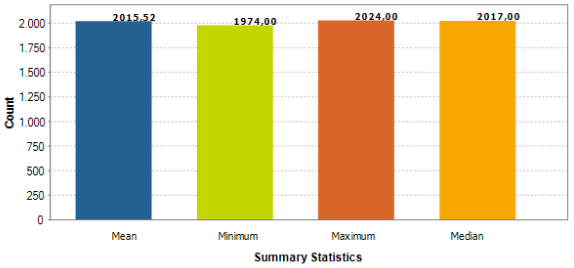
Simple Statistics

Label	Count	%
Row Count	3207	100.00%
Null Count	0	0.00%
Distinct Count	34	1.06%
Unique Count	2	0.06%
Duplicate Count	32	1.00%



Summary Statistics

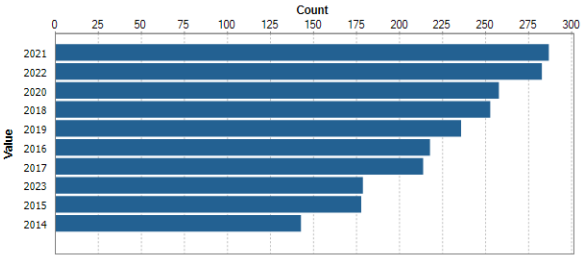
Label	Value
Mean	2015.517929529155
Median	2017.0
Range	50.0
Minimum	1974
Maximum	2024



Column: metadata.model_year

Value Frequency

Value	Count	%
2021	287	8.95%
2022	283	8.82%
2020	258	8.04%
2018	253	7.89%
2019	236	7.36%
2016	218	6.80%
2017	214	6.67%
2023	179	5.58%
2015	178	5.55%
2014	143	4.46%

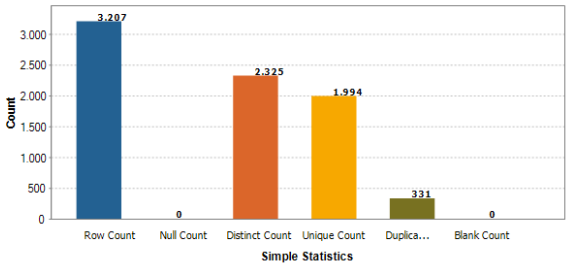


3.3.1.4 Análise do atributo "milage"

Column: metadata.milage

Simple Statistics

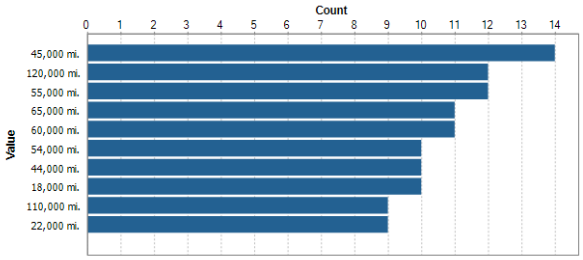
Label	Count	%
Row Count	3207	100.00%
Null Count	0	0.00%
Distinct Count	2325	72.50%
Unique Count	1994	62.18%
Duplicate Count	331	10.32%
Blank Count	0	0.00%



Column: metadata.milage

Value Frequency

Value	Count	%
45,000 mi.	14	0.44%
120,000 mi.	12	0.37%
55,000 mi.	12	0.37%
65,000 mi.	11	0.34%
60,000 mi.	11	0.34%
54,000 mi.	10	0.31%
44,000 mi.	10	0.31%
18,000 mi.	10	0.31%
110,000 mi.	9	0.28%
22,000 mi.	9	0.28%

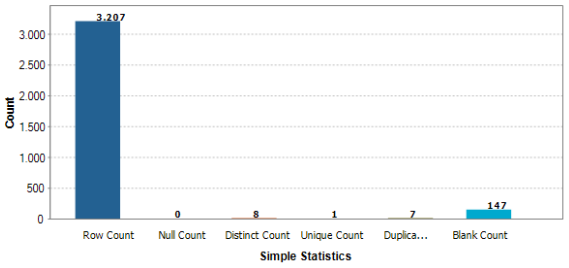


3.3.1.5 Análise do atributo "fuel_type"

Column: metadata.fuel_type

Simple Statistics

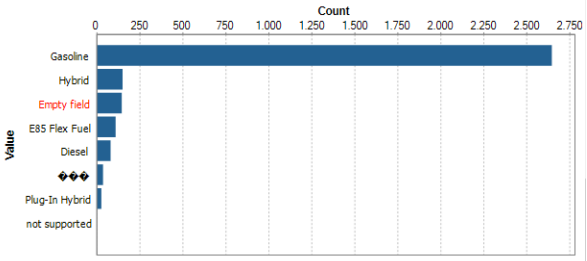
Label	Count	%
Row Count	3207	100.00%
Null Count	0	0.00%
Distinct Count	8	0.25%
Unique Count	1	0.03%
Duplicate Count	7	0.22%
Blank Count	147	4.58%



Column: metadata.fuel_type

Value Frequency

Value	Count	%
Gasoline	2648	82.57%
Hybrid	152	4.74%
Empty field	147	4.58%
E85 Flex Fuel	111	3.46%
Diesel	82	2.56%
��?	38	1.18%
Plug-In Hybrid	28	0.87%
not supported	1	0.03%

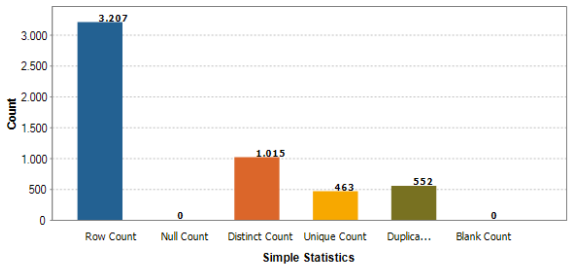


3.3.1.6 Análise do atributo "engine_type"

Column: metadata.engine

Simple Statistics

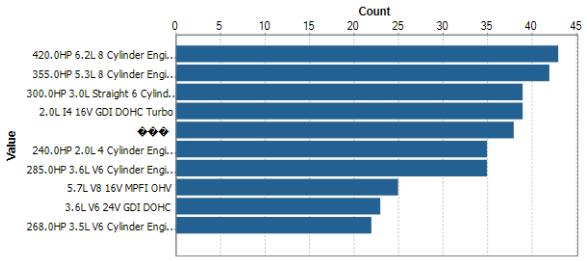
Label	Count	%
Row Count	3207	100.00%
Null Count	0	0.00%
Distinct Count	1015	31.65%
Unique Count	463	14.44%
Duplicate Count	552	17.21%
Blank Count	0	0.00%



Column: metadata.engine

Value Frequency

Value	Count	%
420.0HP 6.2L 8 Cylinder Engine Ga...	43	1.34%
355.0HP 5.3L 8 Cylinder Engine Ga...	42	1.31%
300.0HP 3.0L Straight 6 Cylinder E...	39	1.22%
2.0L I4 16V GDI DOHC Turbo	39	1.22%
???	38	1.18%
240.0HP 2.0L 4 Cylinder Engine Ga...	35	1.09%
285.0HP 3.6L V6 Cylinder Engine G...	35	1.09%
5.7L V8 16V MPFI OHV	25	0.78%
3.6L V6 24V GDI DOHC	23	0.72%
268.0HP 3.5L V6 Cylinder Engine G...	22	0.69%

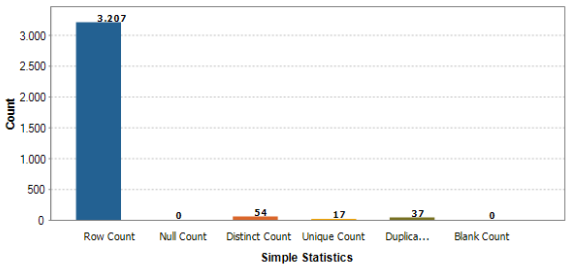


3.3.1.7 Análise do atributo "transmission"

Column: metadata.transmission

Simple Statistics

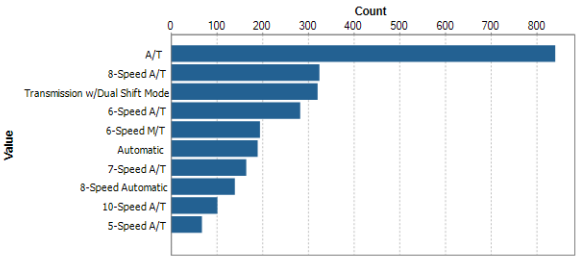
Label	Count	%
Row Count	3207	100.00%
Null Count	0	0.00%
Distinct Count	54	1.68%
Unique Count	17	0.53%
Duplicate Count	37	1.15%
Blank Count	0	0.00%



Column: metadata.transmission

Value Frequency

Value	Count	%
A/T	841	26.22%
8-Speed A/T	325	10.13%
Transmission w/Dual Shift Mode	321	10.01%
6-Speed A/T	283	8.82%
6-Speed M/T	195	6.08%
Automatic	190	5.92%
7-Speed A/T	165	5.14%
8-Speed Automatic	140	4.37%
10-Speed A/T	102	3.18%
5-Speed A/T	68	2.12%

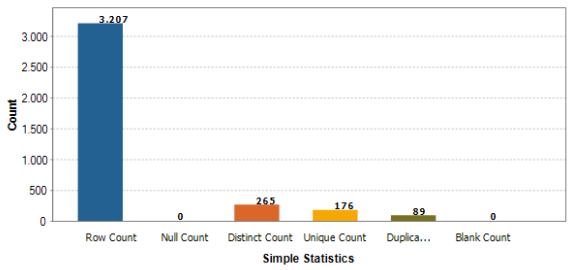


3.3.1.8 Análise do atributo "ext_col"

Column: metadata.ext_col

Simple Statistics

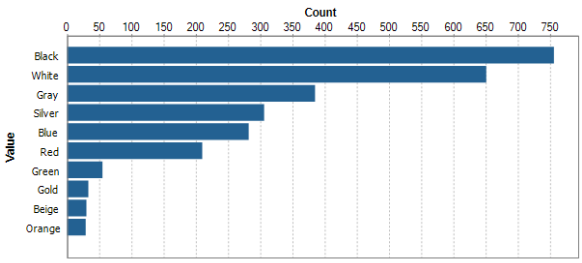
Label	Count	%
Row Count	3207	100.00%
Null Count	0	0.00%
Distinct Count	265	8.26%
Unique Count	176	5.49%
Duplicate Count	89	2.78%
Blank Count	0	0.00%



Column: metadata.ext_col

Value Frequency

Value	Count	%
Black	756	23.57%
White	651	20.30%
Gray	385	12.00%
Silver	306	9.54%
Blue	282	8.79%
Red	210	6.55%
Green	55	1.71%
Gold	33	1.03%
Beige	30	0.94%
Orange	29	0.90%

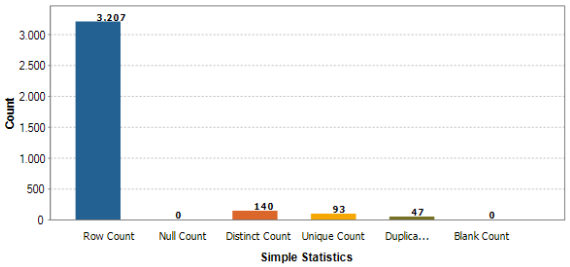


3.3.1.9 Análise do atributo "int_col"

Column: metadata.int_col

Simple Statistics

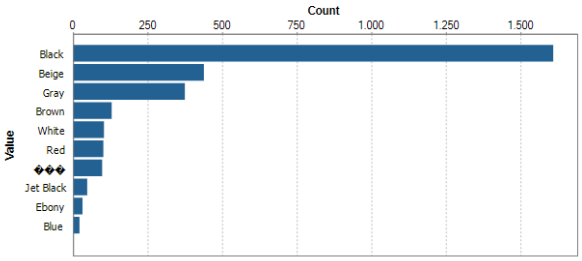
Label	Count	%
Row Count	3207	100.00%
Null Count	0	0.00%
Distinct Count	140	4.37%
Unique Count	93	2.90%
Duplicate Count	47	1.47%
Blank Count	0	0.00%



Column: metadata.int_col

Value Frequency

Value	Count	%
Black	1610	50.20%
Beige	439	13.69%
Gray	375	11.69%
Brown	130	4.05%
White	104	3.24%
Red	102	3.18%
◆◆◆	98	3.06%
Jet Black	48	1.50%
Ebony	32	1.00%
Blue	22	0.69%

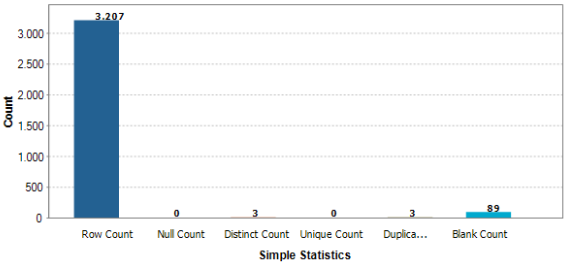


3.3.1.10. Análise do atributo "accident"

Column: metadata.accident

Simple Statistics

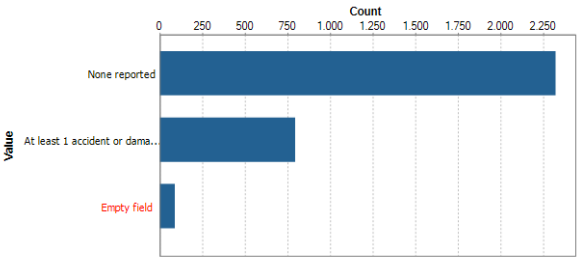
Label	Count	%
Row Count	3207	100.00%
Null Count	0	0.00%
Distinct Count	3	0.09%
Unique Count	0	0.00%
Duplicate Count	3	0.09%
Blank Count	89	2.78%



Column: metadata.accident

Value Frequency

Value	Count	%
None reported	2323	72.44%
At least 1 accident or damage rep...	795	24.79%
Empty field	89	2.78%

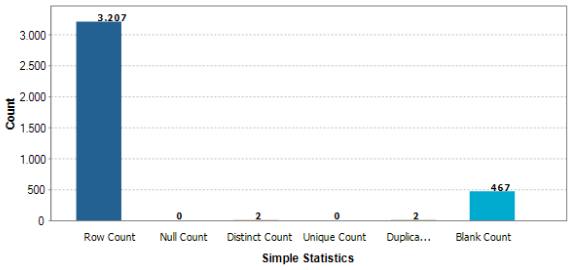


3.3.1.11. Análise do atributo "clean_title"

Column: metadata.clean_title

Simple Statistics

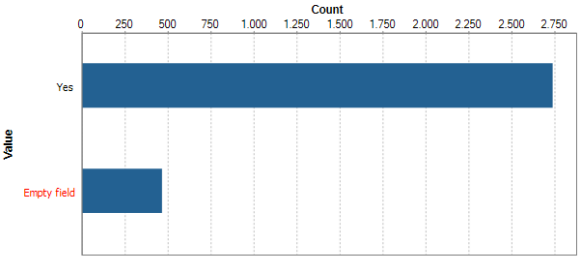
Label	Count	%
Row Count	3207	100.00%
Null Count	0	0.00%
Distinct Count	2	0.06%
Unique Count	0	0.00%
Duplicate Count	2	0.06%
Blank Count	467	14.56%



Column: metadata.clean_title

Value Frequency

Value	Count	%
Yes	2740	85.44%
Empty field	467	14.56%

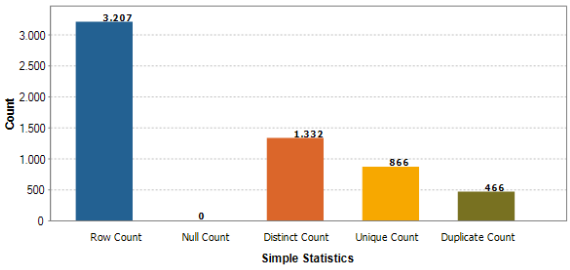


3.3.1.12. Análise do atributo "price"

Column: metadata.price

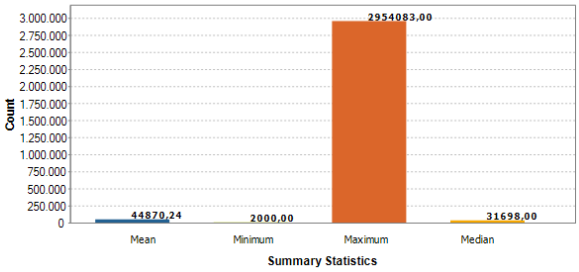
Simple Statistics

Label	Count	%
Row Count	3207	100.00%
Null Count	0	0.00%
Distinct Count	1332	41.53%
Unique Count	866	27.00%
Duplicate Count	466	14.53%



Summary Statistics

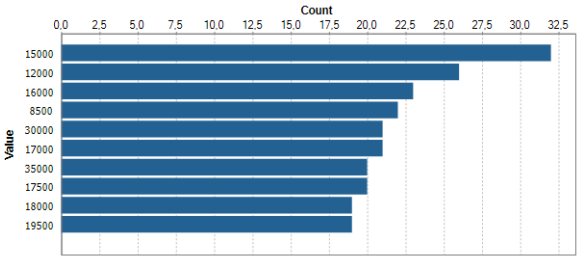
Label	Value
Mean	44870.24415341441
Median	31698.0
Range	2952083.0
Minimum	2000
Maximum	2954083



Column: metadata.price

Value Frequency

Value	Count	%
15000	32	1.00%
12000	26	0.81%
16000	23	0.72%
8500	22	0.69%
30000	21	0.65%
17000	21	0.65%
35000	20	0.62%
17500	20	0.62%
18000	19	0.59%
19500	19	0.59%



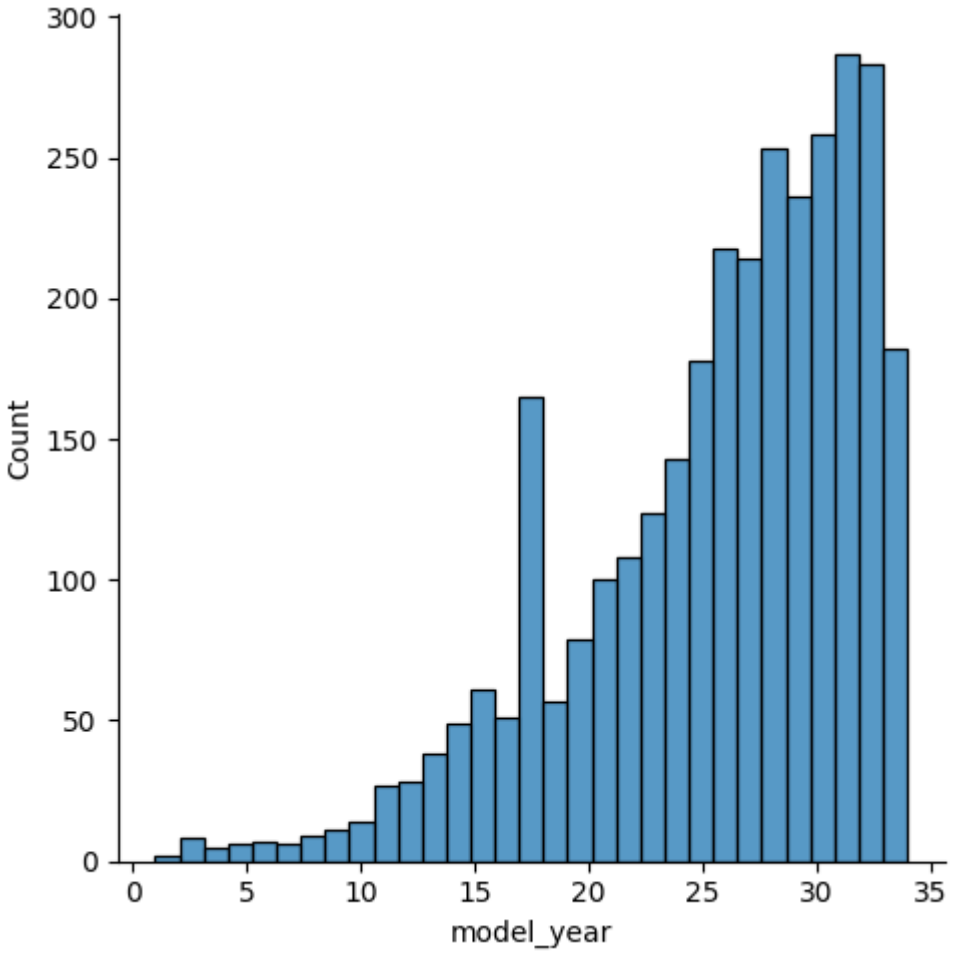
3.3.1.12 Outliers

Os outliers são valores fora do normal encontradas nos dados das tabelas. Para investigar a sua existência foi utilizado o método Z-score.

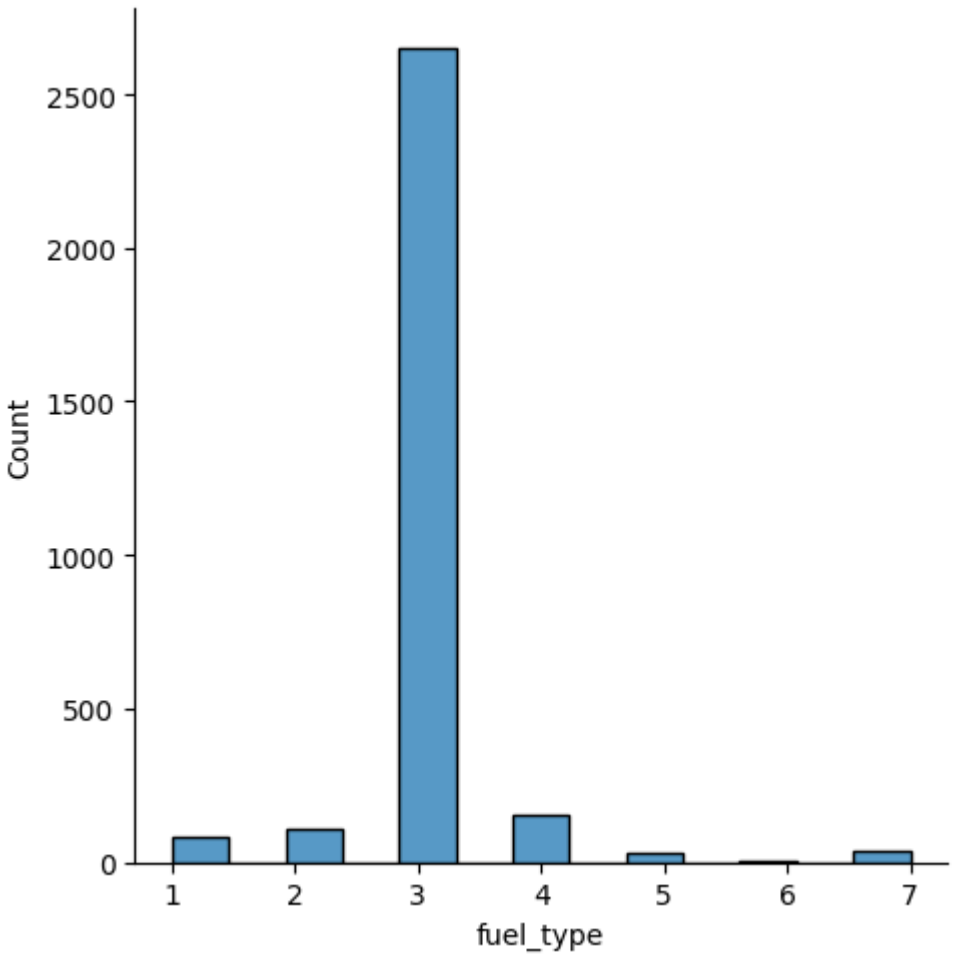
O método Z-score em python consiste em calcular a média de cada coluna e somar o desvio padrão para calcular o maior valor admissível e subtrair o desvio padrão para obter o menor valor admissível. De notar que o desvio padrão é ainda multiplicado por um fator k, que deve ser decidido em função do número de linhas do dataset. Assim, todos os valores que estiverem fora do intervalo entre o menor valor e o maior valor admissíveis, são considerados outliers.

De seguida, encontra-se o programa em python criado bem como os resultados obtidos, ou seja, os outliers de cada coluna. Abaixo de cada gráfico apresentamos um array com os valores identificados como outliers do respetivo gráfico.

```
plt.figure(figsize=(16,5))
plt.subplot(1,2,1)
sns.displot(df_ranked['price'])

plt.show()

def detect_outlier(data):
    threshold = 3
    outliers = []

    mean = np.mean(data)
    std = np.std(data)

    for y in data:
        z_score = (y - mean) / std
        if np.abs(z_score) > threshold:
            outliers.append(y)
    return outliers

def unique(list1):
    unique_list = []

    for x in list1:
        if x not in unique_list:
            unique_list.append(x)
    return unique_list

dados = df_ranked['price']

# Detectar outliers
outliers = detect_outlier(dados)
print("Outliers:", outliers)
```



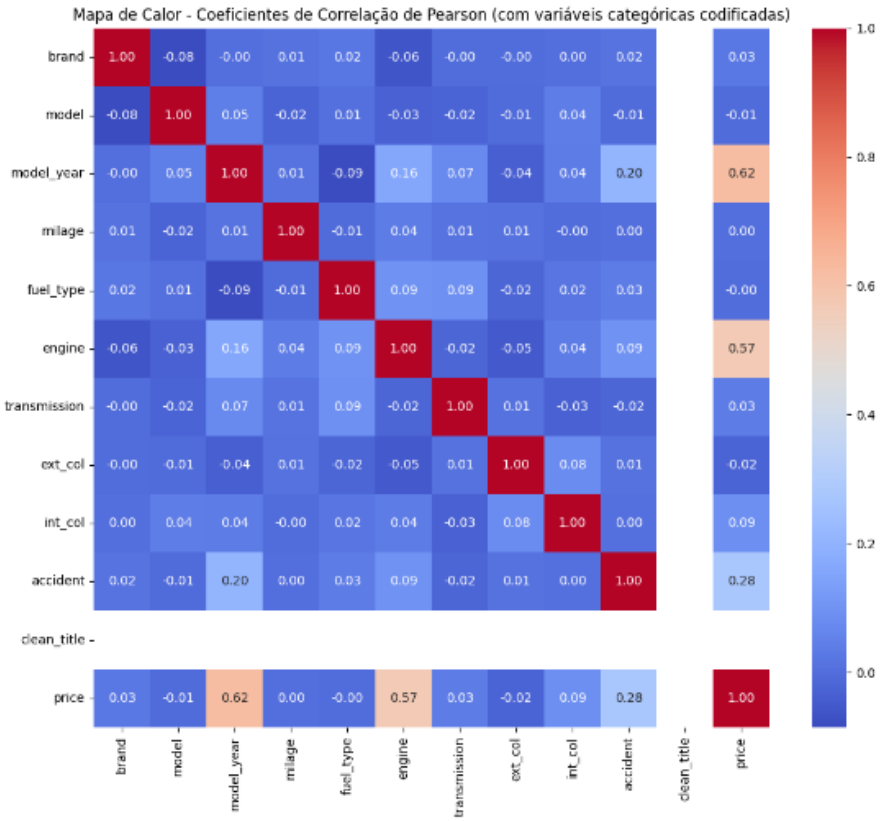
Outliers: [6.0, 6.0, 6.0, 3.0, 5.0, 3.0, 6.0, 4.0, 3.0, 2.0, 7.0, 6.0, 5.0, 7.0, 5.0, 5.0, 3.0, 7.0, 5.0, 6.0, 5.0, 6.0, 7.0, 3.0, 4.0, 4.0, 7.0, 3.0, 3.0, 7.0, 4.0, 4.0, 3.0, 1.0]



Outliers: [1.0, 1.0, 5.0, 1.0, 5.0, 7.0, 7.0, 1.0, 5.0, 5.0, 1.0, 1.0, 1.0, 5.0, 1.0, 1.0, 1.0, 5.0, 1.0, 1.0, 1.0, 5.0, 1.0, 1.0, 1.0, 1.0, 7.0, 7.0, 1.0, 1.0, 1.0, 1.0, 7.0, 1.0, 1.0, 7.0, 1.0, 5.0, 1.0, 1.0, 5.0, 7.0, 1.0, 7.0, 1.0, 7.0, 7.0, 1.0, 1.0, 7.0, 5.0, 1.0, 1.0, 7.0, 1.0, 5.0, 7.0, 1.0, 5.0, 7.0, 1.0, 1.0, 5.0, 7.0, 5.0, 1.0, 6.0, 7.0, 1.0, 1.0, 7.0, 1.0, 1.0, 7.0, 7.0, 5.0, 5.0, 7.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 7.0, 1.0, 1.0, 1.0, 7.0, 1.0, 5.0, 5.0, 1.0, 7.0, 1.0, 5.0, 7.0, 1.0, 1.0, 1.0, 5.0, 7.0, 1.0, 1.0, 1.0, 7.0, 1.0, 7.0, 1.0, 1.0, 5.0, 1.0, 7.0, 1.0, 5.0, 1.0, 7.0, 5.0, 7.0, 7.0, 7.0, 5.0, 1.0, 7.0, 1.0, 7.0, 1.0, 1.0, 5.0, 1.0, 5.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 5.0, 7.0, 7.0, 7.0, 7.0, 1.0, 5.0, 1.0]

3.3.1.13 The Pearson Coeficient

Nesta análise vamos relacionar as colunas para tentar perceber a influência que as mesmas possuem entre elas como podemos verificar no modelo a seguir. Também podemos verificar o código python para a modelação do mesmo.




```
import pandas as pd

df = pd.read_csv('train.csv')

df_ranked = df.rank(method='dense')
```

✓ 1.7s Python

```
correlation_matrix = df_ranked.corr()

import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Mapa de Calor - Coeficientes de Correlação de Pearson (com variáveis categóricas codificadas)')
plt.show()
```

✓ 3.7s Python

3.4 Verify Data Quality

3.4.1 Data Quality Report

Atributo	Anomalias	Solução
brand	N/A	-
model	N/A	-
model_year	N/A	-
mileage	N/A	-
fuel_type	Espaços Vazios	Eliminar linhas com valores nulos
engine	Valores Mal Definidos	Eliminar linhas com valores mal definidos
transmission	N/A	-
ext_col	N/A	-
int_col	Valores Mal Definidos	Eliminar linhas com valores mal definidos
accident	Espaços Vazios	Eliminar linhas com valores nulos
clean_title	Espaços Vazios	Eliminar linhas com valores nulos
price	N/A	-

4. Data Preparation

De forma a tornar os dados mais adequados e relevantes para o estudo dos mesmos, terá de ser efetuada a preparação dos mesmos nesta fase.

4.1 Select Data

4.1.1 Rationale for inclusion/exclusion

Considerei que todos os atributos do dataset train.csv irão ser incluídos no estudo deste projeto por serem relevantes para uma maior taxa de critério de sucesso.

4.2 Clean Data

4.2.1 Data cleaning report

Nesta fase, com base nas anomalias identificadas no Data Quality Report, e sendo que o dataset tem uma vasta quantidade de valores, decidi eliminar linhas com valores nulos e mal definidos.

4.3 Construct Data

Esta tarefa inclui operações construtivas de preparação de dados, tais como a produção de atributos derivados, novos registos completos ou valores transformados para atributos existentes.

4.3.1 Derived Attributes

Os atributos derivados são considerados novos atributos que vão ser desenvolvidos através de atributos já existentes no mesmo registo de dados. No entanto, decidi não o fazer por falta de necessidade e relação entre duas colunas para originar uma nova.

4.3.2 Generated Records

Os registos gerados são registos completamente novos, com novas representações de dados e novos conhecimentos de dados.

4.3.2.1 Cenário 1

Este dataset remete ao dataset original sem qualquer tipo de modificação das colunas ou valores.

4.3.2.2 Cenário 2

O dataset desenvolvido, contem todos os atributos iniciais sem as linhas que contêm vazios ou mal definidos, como identificados anteriormente. De seguida, encontra-se o código utilizado para originar o mesmo.

```python
import pandas as pd

# Carregar o conjunto de dados do CSV
df = pd.read_csv('train.csv')

# Substituir valores por NaN
df = df.replace('not supported', pd.NA)
df = df.replace('-', pd.NA)

# Eliminar linhas com valores vazios em qualquer coluna
df_sem_nulos = df.dropna()

# Exibir o DataFrame resultante
print("DataFrame sem linhas com valores vazios:")
print(df_sem_nulos)

# Salvar o DataFrame resultante em um novo arquivo CSV
df_sem_nulos.to_csv('cenário1.csv', index=False)
```

4.3.2.3 Cenário 3

Neste cenário eliminei as colunas que não são referentes aos aspetos que considerei não básicos num automóvel como “engine”, “transmission”, “clean_title” e “accident”. De seguida, encontra-se o código utilizado para originar o mesmo.

```
df1 = df[['brand', 'model', 'model_year', 'milage', 'fuel_type', 'ext_col', 'int_col', 'price']]

df1.to_csv(os.path.join(os.getcwd(), "cenário3.csv"), index=False)
```

4.3.2.4 Cenário 4

Neste cenário vou utilizar o método de Pearson e eliminar as colunas que apresentam menos de 0 como coeficiente, deixando a marca e o modelo como identificação.

```
df1 = df[['brand', 'model', 'model_year', 'engine', 'transmission', 'int_col', 'accident','price']]

df1.to_csv(os.path.join(os.getcwd(), "cenário4.csv"), index=False)
```

4.4 Intregate Data

4.4.1 Merged Data

Visto que só existe um dataset para análise e não existe possibilidade de adicionar novos atributos por falta de dados, não se efetuou integração de dados.

4.5 Format Data

4.5.1 Reformatted Data

A principal formatação dos dados foi realizada em Construct Data, onde foi efetuado o tratamento de dados que se considerou relevante, dando origem aos 4 datasets representados nas seguintes tabelas.

Cenário 1

Atributo	Descrição	Formato	Quantidade	Exemplos
brand	Marca do veículo	String	3207	Jeep
model	Modelo do veículo	String	3207	Wrangler Sport
model_year	Ano do veículo	Integer	3207	2014
mileage	Distância em miles percorrida pelo veículo	Integer	3207	71,000 mi.
fuel_type	Tipo de combustível do veículo	String	3207	Gasoline
engine	Especificações do motor do veículo	String	3207	285.0HP 3.6L V6 Cylinder Engine Gasoline Fuel
transmission	Tipo de transmissão do veículo	String	3207	5-Speed A/T
ext_col	Cor exterior do veículo	String	3207	Gray
int_col	Cor interior do veículo	String	3207	Black
accident	Histórico de acidentes do veículo	String	3207	None Reported
clean_title	Especificação sobre perda total do veículo	String	3207	Yes
price	Preço do veículo	Integer	3207	22000

Cenário 2

Atributo	Descrição	Formato	Quantidade	Exemplos
brand	Marca do veículo	String	3207	Jeep
model	Modelo do veículo	String	3207	Wrangler Sport
model_year	Ano do veículo	Integer	3207	2014
mileage	Distância em miles percorrida pelo veículo	Integer	3207	71,000 mi.
fuel_type	Tipo de combustível do veículo	String	3207	Gasoline
engine	Especificações do motor do veículo	String	3207	285.0HP 3.6L V6 Cylinder Engine Gasoline Fuel
transmission	Tipo de transmissão do veículo	String	3207	5-Speed A/T
ext_col	Cor exterior do veículo	String	3207	Gray
int_col	Cor interior do veículo	String	3207	Black
accident	Histórico de acidentes do veículo	String	3207	None Reported
clean_title	Especificação sobre perda total do veículo	String	3207	Yes
price	Preço do veículo	Integer	3207	22000

Cenário 3

Atributo	Descrição	Formato	Quantidade	Exemplos
brand	Marca do veículo	String	3207	Jeep
model	Modelo do veículo	String	3207	Wrangler Sport
model_year	Ano do veículo	Integer	3207	2014
mileage	Distância em miles percorrida pelo veículo	Integer	3207	71,000 mi.
fuel_type	Tipo de combustível do veículo	String	3207	Gasoline
ext_col	Cor exterior do veículo	String	3207	Gray
int_col	Cor interior do veículo	String	3207	Black
price	Preço do veículo	Integer	3207	22000

Cenário 4

Atributo	Descrição	Formato	Quantidade	Exemplos
brand	Marca do veículo	String	3207	Jeep
model	Modelo do veículo	String	3207	Wrangler Sport
model_year	Ano do veículo	Integer	3207	2014
engine	Especificações do motor do veículo	String	3207	285.0HP 3.6L V6 Cylinder Engine Gasoline Fuel
transmission	Tipo de transmissão do veículo	String	3207	5-Speed A/T
int_col	Cor interior do veículo	String	3207	Black
accident	Histórico de acidentes do veículo	String	3207	None Reported
price	Preço do veículo	Integer	3207	22000

5. Conclusão

Durante a primeira etapa de Business Understanding baseada na metodologia CRISP-DM, concluo o sucesso não só da sua realização, como da importância na idealização de todo o negócio para a dinâmica e estudo da unidade curricular. Foi possível identificar os objetivos do negócio e de Data Mining, apesar de algumas dificuldades. Nos critérios de sucesso de data mining tive dificuldade na definição dos mesmos, assim como na produção do plano de projeto, pois recorri a uma alternativa ao MS Project para a realização do Diagrama de Gantt.

Na segunda etapa, o Data Understanding, descrevi o dataset que iria utilizar para a realização do projeto. De seguida, analisei cada coluna do dataset com recurso ao Talend, de forma a entender os dados que o dataset continha. Averiguei a existência de outliers utilizando o método z-score em python. De seguida, analisei a relação entre as colunas através do coeficiente de Pearson. Verifiquei a qualidade dos dados e as anomalias que o mesmo apresentava, bem como as soluções para resolver as mesmas.

Na terceira etapa, Data Preparation, efetuei a preparação dos dados. Criei quatro cenários diferentes baseados em análise feitas anteriormente, como a remoção de anomalias nos valores do dataset, avaliações básicas de escolha de carro e a relação entre as colunas observadas no coeficiente de Pearson. Obtive alguma dificuldade nesta fase para a definição de possíveis cenários, para que o mesmo não sofresse repercussões futuras.