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

Escola de Engenharia  
Universidade do Minho

4º Ano

1º Semestre

Aprendizagem Automática em Sistemas de Informação

Uma imagem com vestuário, pessoa, Cara humana, Blazer

Descrição gerada automaticamente

Francisco Miguel Pinheiro Cardoso

A79570

Índice

[1. Introdução 1](#_Toc151930568)

[2. Business Understanding 2](#_Toc151930569)

[2.1 Determine Business Objetives 2](#_Toc151930570)

[2.1.1 Background 2](#_Toc151930571)

[2.1.2 Business Objetives 3](#_Toc151930572)

[2.1.3 Business Success Criteria 3](#_Toc151930573)

[2.2 Assess Situation 4](#_Toc151930574)

[2.2.1 Inventory of Resources 4](#_Toc151930575)

[2.2.2 Requirements, Assumptions, and Constrains 4](#_Toc151930576)

[2.2.3 Risks and Contingencies 5](#_Toc151930577)

[2.2.4 Terminology 5](#_Toc151930578)

[2.2.5 Costs and Benefits 5](#_Toc151930579)

[2.3 Determine Data Mining Goals 6](#_Toc151930580)

[2.3.1 Data Mining Goals 6](#_Toc151930581)

[2.3.2 Data Mining Success Criteria 6](#_Toc151930582)

[2.4 Produce Project Plan 7](#_Toc151930583)

[2.4.1 Project Plan 7](#_Toc151930584)

[2.4.2 Initial Assessment of Tools and Techniques 9](#_Toc151930585)

[3. Data Understanding 10](#_Toc151930586)

[3.1 Collect Initial Data 10](#_Toc151930587)

[3.1.1 Initial Data Collection Report 10](#_Toc151930588)

[3.2 Describe Data 10](#_Toc151930589)

[3.2.1 Data Description Report 10](#_Toc151930590)

[3.3 Explore Data 11](#_Toc151930591)

[3.3.1 Data Exploration Report 11](#_Toc151930592)

[3.4 Verify Data Quality 23](#_Toc151930593)

[3.4.1 Data Quality Report 23](#_Toc151930594)

[4. Data Preparation 24](#_Toc151930595)

[4.1 Select Data 24](#_Toc151930596)

[4.1.1 Rationale for inclusion/exclusion 24](#_Toc151930597)

[4.2 Clean Data 24](#_Toc151930598)

[4.2.1 Data cleaning report 24](#_Toc151930599)

[4.3 Construct Data 24](#_Toc151930600)

[4.3.1 Derived Attributes 24](#_Toc151930601)

[4.3.2 Generated Records 25](#_Toc151930602)

[4.4 Intregate Data 26](#_Toc151930603)

[4.4.1 Merged Data 26](#_Toc151930604)

[4.5 Format Data 27](#_Toc151930605)

[4.5.1 Reformatted Data 27](#_Toc151930606)

[5. Conclusão 30](#_Toc151930607)

# 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.

# 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.

## 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.

## 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.

## 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

Uma imagem com texto, Tipo de letra, branco, número

Descrição gerada automaticamentePara 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.

O critério definido é:

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

## Produce Project Plan

### Project Plan

Uma imagem com texto, captura de ecrã, número, Tipo de letra

Descrição gerada automaticamente

Uma imagem com texto, captura de ecrã, número, Tipo de letra

Descrição gerada automaticamente

Uma imagem com texto, captura de ecrã, Tipo de letra, número

Descrição gerada automaticamente

Uma imagem com texto, software, file, captura de ecrã

Descrição gerada automaticamente

### 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. |

# 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”

Uma imagem com texto, file, número, Gráfico

Descrição gerada automaticamenteUma imagem com texto, file, número, Gráfico

Descrição gerada automaticamente

#### 3.3.1.2 Análise do atributo “model”

Uma imagem com texto, Gráfico, file, diagrama

Descrição gerada automaticamente

Uma imagem com texto, número, file, Tipo de letra

Descrição gerada automaticamente

#### 3.3.1.3 Análise do atributo “model\_yearl”

Uma imagem com texto, file, Gráfico, número

Descrição gerada automaticamente

Uma imagem com texto, número, Gráfico, file

Descrição gerada automaticamenteUma imagem com texto, número, file, Gráfico

Descrição gerada automaticamente

#### 3.3.1.4 Análise do atributo “milage”

Uma imagem com texto, Gráfico, file, diagrama

Descrição gerada automaticamenteUma imagem com texto, número, file, Tipo de letra

Descrição gerada automaticamente

#### 3.3.1.5 Análise do atributo “fuel\_type”

Uma imagem com texto, file, Gráfico, número

Descrição gerada automaticamente

Uma imagem com texto, número, Tipo de letra, file

Descrição gerada automaticamente

#### 3.3.1.6 Análise do atributo “engine\_type”

Uma imagem com texto, Gráfico, file, diagrama

Descrição gerada automaticamenteUma imagem com texto, Tipo de letra, file, número

Descrição gerada automaticamente

#### 3.3.1.7 Análise do atributo “transmission”

Uma imagem com texto, file, Gráfico, número

Descrição gerada automaticamenteUma imagem com texto, número, file, Tipo de letra

Descrição gerada automaticamente

#### 3.3.1.8 Análise do atributo “ext\_col”

Uma imagem com texto, file, Gráfico, número

Descrição gerada automaticamenteUma imagem com texto, número, file, Gráfico

Descrição gerada automaticamente

#### 3.3.1.9 Análise do atributo “int\_col”

Uma imagem com texto, file, Gráfico, número

Descrição gerada automaticamenteUma imagem com texto, número, file, Gráfico

Descrição gerada automaticamente

#### 3.3.1.10. Análise do atributo “accident”

Uma imagem com texto, file, Gráfico, captura de ecrã

Descrição gerada automaticamenteUma imagem com texto, file, Tipo de letra, número

Descrição gerada automaticamente

#### 3.3.1.11. Análise do atributo “clean\_titile”

Uma imagem com texto, file, número, Gráfico

Descrição gerada automaticamenteUma imagem com texto, file, Gráfico, diagrama

Descrição gerada automaticamente

#### 3.3.1.12. Análise do atributo “price”

Uma imagem com texto, Gráfico, file, diagrama

Descrição gerada automaticamente

Uma imagem com texto, captura de ecrã, file, Gráfico

Descrição gerada automaticamenteUma imagem com texto, captura de ecrã, número, file

Descrição gerada automaticamente

#### 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.

Uma imagem com texto, captura de ecrã, software

Descrição gerada automaticamente

Uma imagem com diagrama, Gráfico, captura de ecrã, file

Descrição gerada automaticamente

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]

Uma imagem com texto, captura de ecrã, diagrama, file

Descrição gerada automaticamente

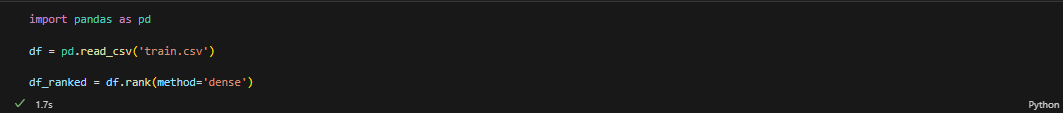
Outliers: [1.0, 1.0, 5.0, 1.0, 5.0, 7.0, 7.0, 1.0, 5.0, 5.0, 1.0, 1.0, 1.0, 5.0, 1.0, 1.0, 1.0, 5.0, 1.0, 1.0, 1.0, 5.0, 1.0, 1.0, 1.0, 1.0, 7.0, 7.0, 1.0, 1.0, 1.0, 1.0, 7.0, 1.0, 1.0, 7.0, 1.0, 5.0, 1.0, 1.0, 5.0, 7.0, 1.0, 7.0, 1.0, 7.0, 7.0, 1.0, 1.0, 7.0, 5.0, 1.0, 1.0, 7.0, 1.0, 5.0, 7.0, 1.0, 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.

Uma imagem com captura de ecrã, padrão, quadrado, Retângulo

Descrição gerada automaticamente



Uma imagem com texto, software, Software de multimédia, captura de ecrã

Descrição gerada automaticamente

## 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 | - |

# 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.

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

#### 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.



#### 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.



## 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 |

# 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.