# 2. Data Understanding

## 2.1 Collect Initial Data

### 2.1.1 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).

## 2.2 Describe Data

### 2.2.1 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 |

## 2.3 Explore Data

### 2.3.1 Data Exploration Report

#### 2.3.1.1 Análise do atributo “brand”

Uma imagem com texto, file, número, Gráfico

Descrição gerada automaticamente

Uma imagem com texto, file, número, Gráfico

Descrição gerada automaticamente

#### 2.3.1.2 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

#### 2.3.1.3 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

#### 2.3.1.4 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

#### 2.3.1.5 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

#### 2.3.1.6 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

#### 2.3.1.7 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

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

Uma imagem com texto, file, Gráfico, número

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

Descrição gerada automaticamente

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

Uma imagem com texto, file, Gráfico, número

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

Descrição gerada automaticamente

#### 2.3.1.10. 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

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

Uma imagem com texto, file, número, Gráfico

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

Descrição gerada automaticamente

#### 2.3.1.12. 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

#### 2.3.1.12 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.

Uma imagem com texto, captura de ecrã, software

Descrição gerada automaticamenteDe 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 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

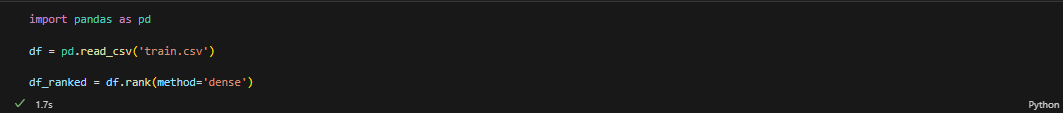
Descrição gerada automaticamente

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]

#### 2.3.1.13 The Pearson Coeficient

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

## 2.4 Verify Data Quality

### 2.4.1 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 | - |

# 3. 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.

## 3.1 Select Data

### 3.1.1 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.

## 3.2 Clean Data

### 3.2.1 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.

## 3.3 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.

### 3.3.1 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.

### 3.3.2 Generated Records

## 3.4 Intregate Data

### 3.4.1 Merged Data

## 3.5 Format Data

### 3.5.1 Reformatted Data