Faculdade de Ciências - Universidade de Lisboa

Integração e Processamento Analítico de Informação

MEI - 2022/2023

# Stage 1: Data Analysis

[ Group 13 ]

Diogo Araújo - fc60997 - MEI
João Braz - fc60419 - MEI
Joel Oliveira - fc59442 - MEI
Tomás Matos - fc53438 - MI

26 de março de 2023

# Conteúdo

# 1 Introdução

Neste relatório, procuramos demonstrar o processo de negócio escolhido pelo nosso grupo e os passos que levaram à sua criação. Este processo foi escolhido como sendo a venda de carros, com base nos datasets escolhidos pelo grupo, onde procuramos encontrar informação diversa sobre carros em segunda mão que permita a sua venda e análise.

Os dados obtidos foram inicialmente procurados pelos membros do grupo com vários requisitos iniciais, incluindo mas não limitado à existência de uma coluna com uma data explicita como também um tema consistente entre eles. Isto levou ao grupo encontrar, e utilizar, várias fontes de dados relacionadas com a venda de carros. Estas foram depois analizadas pelos seus valores e erros. Após termos escolhido todos os dados relevantes, e de os termos tratado por erros e analizado, criámos um diagrama com as ligações entre estas fontes que demonstra a forma que estas se interligam. Com a base criada, foi possivel passar para a criação de um processo de negócio que conseguisse tomar vantagem dos dados obtidos. Devido à escolha ter sido focada à volta da venda de carros em segunda mão, o processo de negócio criado envolveu a venda destes mesmos. Finalmente, foi necessário definir três questões analiticas para o processo.

Este relatório irá focar-se no estado 1 do projeto, onde depois será atualizado com a continuação do seu trabalho de modo a acomodar as mudanças.

## 2 Fontes de Dados

As fontes de dados utilizadas foram obtidas através de fontes de *Open Data*. Inicialmente, foi feita uma pesquisa exploratória por vários dos sites que fornecem esse tipo de dados, sendo igualmente percebido que os dados que permitem, com maior facilidade, relacionar a um processo de negócio seriam dados relativos a algum tipo de transação.

Assim, foi decidido selecionar fontes de dados referentes à vendas de carros. Para aumentar a informação dos nossos dados foram ainda adicionadas mais duas fontes de dados, sendo uma referente a votações e reviews das compras de carros e outra referente ao preço dos combustíveis.

Por fim os dados recolhidos foram todos obtidos através do site *Kaggle*. Neste, foram selecionados 3 fontes de dados distintas sobre vendas de carros:

- *US used car sales data*;
- *Used Cars Dataset*;
- *10 million Vehicle Registrations with Prices*.

Posteriormente, foram selecionados os *datasets* referentes às *reviews* dos carros e aos preços dos combustiveis, nomeadamente:

- *Edmunds Car review*;
- *US gasoline and diesel retail prices prediction*.

De modo a manter uma consistência, todos os dados selecionados foram recolhidos nos Estados Unidos da América (EUA). Assim, foi possivel analisar os preços a partir do estado do ponto de vista da economia dos EUA apenas, sendo um exemplo disto o preço dos seus combustíveis.

# 3 Análise de Dados

Nesta secção, efetuamos uma limpeza e análise dos dados obtidos. Estes serão as nossas fontes de dados e o que iremos utilizar para criar o nosso processo de negócio.

## 3.1 Used Cars Dataset

Esta fonte de dados contèm um histórico de transações de compra de carros. Na 1, sendo esta a tabela refernte à fonte de dados, é possível verificar os campos existentes. Há vários pontos a necessitar serem trabalhados de modo a melhorar a utilidade do *dataset*. Com isto, as observações verificadas estão resumidas nos próximos parágrafos.

O campo *county* não contém nenhum valor preenchido, pelo que é completamente irrelevante. Na coluna em relação ao número de cilindros do motor, podem ser alterados os seus dados para o tipo "Numérico", sendo que o valor referente fica o número de cilindros visto sabermos a que se refere o número.

A coluna *VIN* contém vários valores em falta. Visto este valor ser uma referência externa então não há necessidade de removermos dados devido à falta deste campo. Os campos com URL's contém urls com referências a sites que já não existem. Este é também o único ficheiro com tais dados, pelo que não teriam grande utilidade. Podem ser factos associados à transação, no entanto.

Os anos de fabrico dos veículos concentram-se em anos mais recentes e os preços estão bastante concentrados num intervalo de milhares. Há, no entanto, pontuais *outliers* com valores absurdamente grandes. Os carros deste dataset contêm vendas relativamente uniformes entre as regiões mais comuns.

A marca mais comum nas vendas é a marca "Ford"e o combustível mais comum é a gasolina. Relativamente aos campos de texto, como *manufacturer* ou *model*, todos os dados já estão em *lowercase* e uniformizados. Desse modo, não é necessário tratamento adicional.

Não foram encontrados dados duplicados. Por fim, foram mantidos apenas dados com todas as informações relevantes. Para isto, foram só removidas as colunas não necessárias à análise e as linhas com dados em falta (à excessão da coluna *VIN*, cujo valor não é relevante). No fim ficamos com cerca de 45% dos dados originais, o que em números absolutos resultou num total de cerca de 175 mil transações neste dataset.

| Campo | Descrição | Tipo de Dados | Exemplo |
|---|---|---|---|
| id | Identificador Único | Numérico | 7301591192 |
| url | URL do artigo | Texto | "https://prescott.craig[...].html" |
| region | Região do site Craigslist | Texto | "prescott" |
| region_url | URL da Região do site Craigslist | Texto | "https://prescott.craigslist.org" |
| price | Preço do Carro | Numérico | 6000 |
| year | Ano do Carro | Numérico | 2019 |
| manufacturer | Marca do Carro | Texto | "ford" |
| model | Modelo do Carro | Texto | "f-150 xlt" |
| condition | Condição do Carro | Texto | "excellent" |
| cylinders | Nº de Cilindros do Motor do Carro | Texto | "6 cylinders" |
| fuel | Combustível utilizado pelo Carro | Texto | "Gas" |
| odometer | Milhas percorridas pelo Carro | Numérico | 128000 |
| title_status | Estado do Carro | Texto | "clean" |
| transmission | Transmissão do Carro | Texto | "manual" |
| VIN | Identificador Único Externo | Texto | "1J4RS4GG8BC644367" |
| drive | Tração do Carro | Texto | "rwd" |
| size | Tamanho do Carro | Texto | "full-size" |
| type | Genérico do Carro | Texto | "sedan" |
| paint_color | Cor do Carro | Texto | "silver" |
| image_url | URL da imagem do Carro | Texto | "https://images.craigslist[...].jpg" |
| description | Descrição do Carro | Texto | "White Toyota 4-Runner[...]" |
| county | Coluna Vazia | NULL | NULL |
| state | Estado onde se situa a venda | Texto | "al" |
| lat | Latitude do Estado | Numérico | 33.209789 |
| long | Longitude do Estado | Numérico | -86.783493 |
| posting_data | Data da transação | Data e Hora | 2021-05-02T08:48:09-0500 |

Tab. 1: Descrição da tabela "vehicles.csv"

### 3.2 US used car sales data

Apesar de uma redução considerável na quantidade de campos neste ficheiro, o tratamento deste ficheiro implicou ainda uma quantidade elevada de trabalho. A partir da tabela 2, é possível verificar, por exemplo, que os campos de texto não se encontram em letra pequena, entre outros cenários necessários de formatar.

Verificou-se que estes dados não contêm informações sobre qual o tipo de combustível utilizado por cada veículo. No entanto, como pode ser visualizado que nos EUA a maior parte dos carros consomem gasolina, verificou-se que é possível inferir o tipo de combustível do carro, através de descritores do modelo ou do motor, pois quando os veículos não são a gasolina estes campos contêm *keywords* como 'diesel' ou 'hybrid'. Relativamente à localização, só há referência ao zipcode. Este campo, no entanto, fornece informações relativas ao estado, latitude e longitude. Assim são obtidos três campos adicionais, à semelhança da fonte de dados anterior. Estes dados foram obtidos através da livraria de python *pyzipcode*.

Relevante à análise dos dados, no campo *Year* foram encontrados valores irregulares. Neste foram encontrados 12 veículos cujo ano de fabrico se encontrava na ordem dos 10 milhões. Verificámos que o problema se devia a um preenchimento de zeros à direita, exemplificando:

[ Ano real = 1999 → Ano nos dados = 19990000 ]

Assim, a solução foi dividir os respetivos valores por 10000. Houve um caso em que o ano era 2914, sendo que este foi retirado dos dados.

Ao verificar a distribuição do número de cilindros, verificaram-se vários cenários inesperados. Os valores, à primeira vista irregulares foram os seguintes: 1) 118 cilindros; 2) 123 cilindros; 3) 350 cilindros; 4) 440 cilindros; 5) 2147483647 cilindros. Estes casos foram assim inspecionados um a um. Nos exemplos em que foi possível inferir o número de cilindros, quer pelo nome do motor (e.g V8, 8 cilindros) quer por carros com as mesmas caracteristicas, os valores foram corrigidos. Nos cenários onde tal não foi possível, as linhas foram descartadas.

Durante a análise, verificou-se um número grande de veículos com 0 cilindros. À primeira vista notou-se que poderia ser referente a veículos que eram vendidos sem motor. Na correção dos dados anterior, foi notório que existiam amostras cujo numero de cilindros era referido no nome do motor, mas o valor não se refletia no campo "NumCylinders"que se encontrava a 0. Esta correção foi também efetuada tanto quanto possível.

| Campo | Descrição | Tipo de Dados | Exemplo |
|---|---|---|---|
| ID | Identificador Único | Numérico | 137178 |
| pricesold | Preço de Venda do Carro | Numérico | 7500 |
| yearsold | Ano de Venda do Carro | Numérico | 2020 |
| datesold | Data de Venda do Carro | Data | 2020-11-23 |
| zipcode | Código Postal dos EUA | Numérico | 81006 |
| Mileage | Milhas percorridas pelo Carro | Numérico | 84430 |
| Make | Marca do Carro | Texto | "Ford" |
| Model | Modelo do Carro | Texto | "Mustang" |
| Year | Ano do Carro | Numérico | 1988 |
| Trim | Versão do Modelo | Texto | "Lx" |
| Engine | Motor do Carro | Texto | "5.0L Gas V8" |
| BodyType | Genérico do Carro | Texto | "Sedan" |
| NumCylinders | Nº de Cilindros do Motor do Carro | Numérico | 8 |
| DriveType | Tração do Carro | Texto | "RWD" |

Tab. 2: Descrição da tabela "used_car_sales.csv"

Por fim foi verificado no campo "Make"que existiam várias nomenclaturas para a mesma marca, dando como exemplo "ford"e "ford 4x4", entre muitas outras. Estes cenários foram também corrigidos um a um. Existiam inúmeros casos onde as características do carro (o modelo ou aspetos como "equipado para pessoas com deficiência") se encontravam no campo errado. Assim, alguns dos valores nulos do campo "Model"foram preenchidos. Neste *dataset* também terminámos com cerca de 50% dos dados, um total de cerca de 60 mil observações.

### 3.3 10 million Vehicle Registrations with Prices

Neste dataset, sendo possivel verificar os seus valores na tabela 3, conseguimos visualizar os diferentes tipos de veículos registrados em Tenesse entre 2018 e 2022 bem como o preço do carro e a localização em que este foi vendido. Para esta fonte de dados, foi necessário primeiro efetuar uma limpeza dos dados seguido por uma formatação e análise deles.

Durante a limpeza, foi possivel visualizar a existência de vários valores nulos em várias das colunas do dataset, incluindo o preço do veiculo e o VIN. Devido à quantidade elevada de dados nesta tabela, e o número de valores nulos ser relativamente pequeno, decidimos remover todas as linhas que tinham um valor não existente devido a não impactar muito os dados finais.

Depois de estes dados serem removidos, efetuámos a formatação da tabela de modo a facilitar a sua visualização. Isto envolveu renomear as suas colunas para nomes mais perceptiveis, sendo que neste caso alterámos para ficarem em lower case. Foi também efetuada uma alteração dos tipos das colunas, sendo que todas as colunas numéricas foram alteradas para "int32"enquanto as colunas textuais foram alteradas para "category". As datas foram igualmente alteradas para estar no formato dos EUA, sendo esse um formato que demonstra inicialmente o ano seguido pelo mes e o dia.

Com a formatação efetuada, podémos passar para a sua análise. Este dataset contém um total de 13 colunas e 9498752 linhas / dados após a sua limpeza. Os tipos de variáveis presentes em cada coluna são divididos entre numéricos, textos (categorias) e datas, sendo que estes tipos são utilizados para representar valores presentes num carro, tal como o seu preço, o seu modelo e a sua data de compra. Ao visualizarmos estes dados, conseguimos perceber que uma parte dos preços estavam iguais a zero. Estes não foram removidos do dataset de modo a ficarmos com um número mais elevado de exemplos para as outras colunas. No entanto, estes valores não serão utilizados em queries relevantes ao preço. Esta foi a única coluna com problemas, sendo que nas outras colunas não conseguimos encontrar nada.

| Campo | Descrição | Tipo de Dados | Exemplo |
|---|---|---|---|
| VIN | Identificador Único do Veiculo | Texto | "1GCHK23123F115066" |
| price | Preço de Venda | Numérico | 500.0 |
| odometer_type | Tipo de Odometro | Numérico | 1 |
| mileage | Milhagem | Numérico | 25818 |
| county | Condado | Texto | Lewis |
| zip | Zip do Condado | Numérico | 38462 |
| model_year | Ano do Modelo | Numérico | 2003 |
| make | Marca do Veiculo | Texto | "CHEV" |
| model | Modelo do Veiculo | Texto | "SIL" |
| vehicle_type | Tipo de Veiculo | Texto | "AUTO" |
| new_used | Utilizado / Novo | Texto | "U" |
| title_issue_date | Data de Titulo | Data | 2020-10-19 |
| purchase_date | Data de Compra | Data | 2020-08-08 |

Tab. 3: Descrição da tabela "tn_mvr-2018-2022.csv"

### 3.4 Edmunds car review

Este dataset foi feito através de um *webscrapping* do *site edmunds.com* que encontrou várias *reviews* encontradas entre 2000 até 2019. Este dataset pode ser verificado na tabela 4 com os seus campos existentes e com um exemplo dos seus valores.

A presente fonte de dados continha 310 valores (linhas) repetidos, sendo por isso descartadas. Relativamente aos valores em falta, sendo estes cerca de 150 valores, estes também foram descartados. O dataset ficou reduzido de 299045 a 9859 valores, havendo por isso apenas uma perda de 3% do dataset anterior ao pré-processamento.

| Campo | Descrição | Tipo de Dados | Exemplo |
|---|---|---|---|
| company | Marca do veículo | Texto | "ford" |
| model | Modelo do veículo | Texto | "model-x" |
| year | Ano do veículo | Numérico | 1995 |
| reviewer | Nome de utilizador da publicação | Texto | "boudouris$_3$" |
| date | Data da publicação | Data | 2012/5/1 |
| titulo | Título da publicação | Texto | "An Excellent Car!" |
| rating | Rating 0-5 | Numérico | 4 |
| review | Descrição da análise | Texto | "Not a trouble with maintenance ..." |

Tab. 4: Descrição da tabela "edmundscar.csv"

### 3.5 US gasoline and diesel retail prices prediction

Esta fonte de dados, que é possivel visualizar na tabela 5, contém um histórico do preço do gasóleo (diesel) e da gasolina ao longo do tempo onde cada elemento do dataset é relativo ao preço numa dada semana. Esta fonte de dados contém 1361 entradas relativas ao preço dos combustivéis desde 1995 até ao ano de 2021.

Neste dataset, não foi possivel encontrar / visualizar valores duplicados nem valores em falta. A data dos preços dos combustivéis foi convertida de um formato *String* para o formato *datetime64*, ou seja, ano-mês-dia. Relativamente, os restantes dados tem o tipo Numérico devido a serem essencialmente os valores de gasolina. Desse modo, não são demonstrados os exemplos na próxima tabela devido a serem todos formatados igualmente.

| Campo | Descrição | Tipo de Dados |
|---|---|---|
| data | Data de coleta do preço | Data |
| a1 | Gasolina de todos os graus e formulações | Numérico |
| a2 | Gasolina convencional de todos os graus | Numérico |
| a3 | Gasolina reformulada de todos os graus | Numérico |
| r1 | Gasolina regular de todas as formulações | Numérico |
| r2 | Gasolina regular convencional | Numérico |
| r3 | Gasolina regular reformulada | Numérico |
| m1 | Gasolina média de todas as formulações | Numérico |
| m2 | Gasolina média convencional | Numérico |
| m3 | Gasolina média reformulada | Numérico |
| p1 | Gasolina premium de todas as formulações | Numérico |
| p2 | Gasolina premium convencional | Numérico |
| p3 | Gasolina premium reformulada | Numérico |
| d1 | Diesel No 2 | Numérico |

Tab. 5: Descrição da tabela "OilPrices.csv"

# 4 Diagrama

Nesta secção, representamos um esquema que relaciona as várias fontes de dados e que as interliga. Este está apresentado na Figura 1. Para os *datasets* referentes às vendas de carros, é possível verificar que dados têm em comum. Para os restantes, podemos verificar que dados é que permitem a relação com os factos.

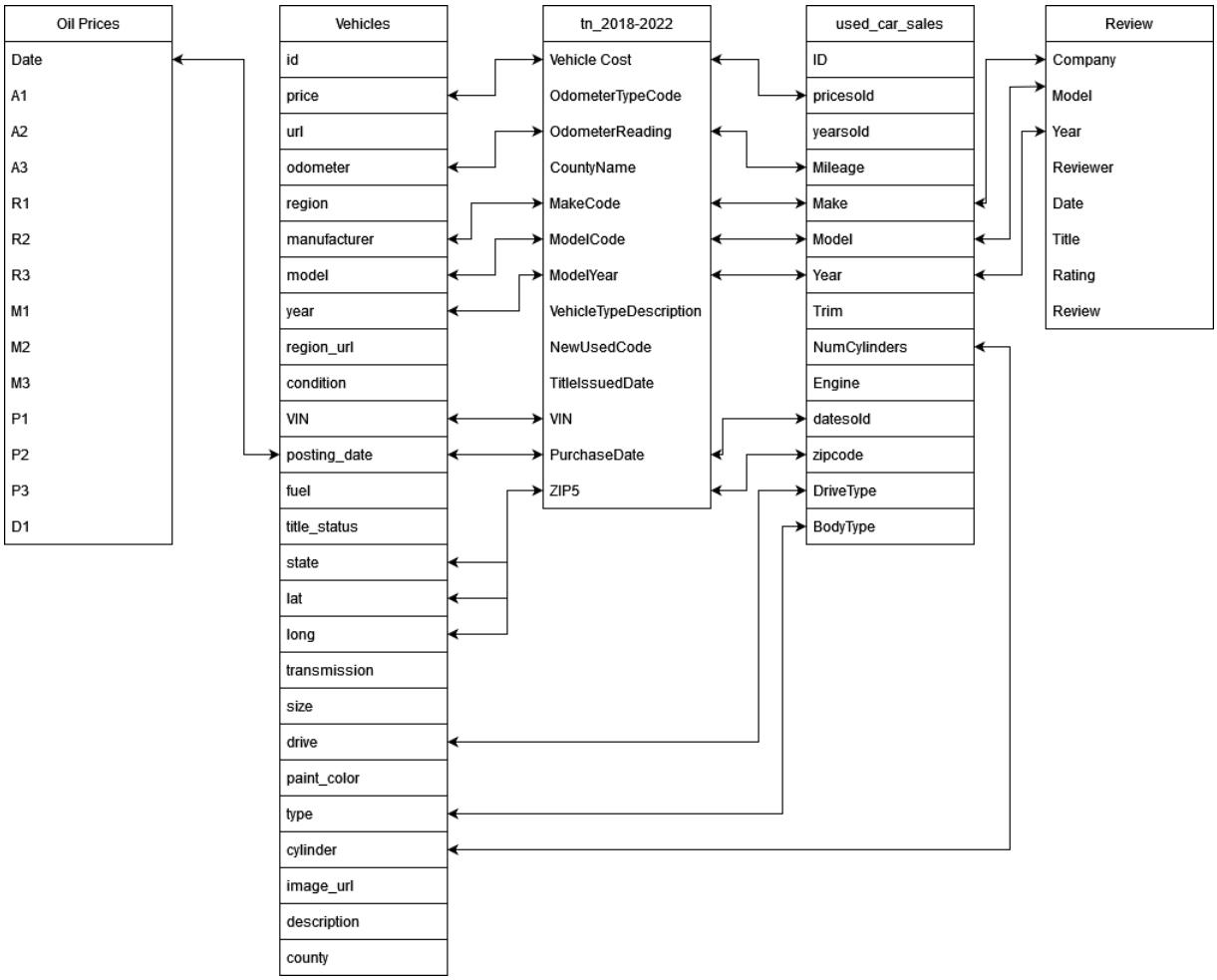


Fig. 1: Diagrama de relações entre os dados

# 5 Processo de Negócio

O nosso processo de negócio, como foi dito anteriormente, baseia-se na venda de carros em segunda mão. Este negócio foi criado com base nos datasets obtidos, e escolhidos, que foram demonstrados nas secções anteriores. Com estes apresentado no diagrama, é possível efetuar várias comparações relativamente ao preço dos carros. As perguntas de negócio que criámos são as seguintes:

- *O tipo de combustível do carro faz diferença no preço? Um carro a gasolina é mais caro que um carro a gasóleo? Que tipo de combustível vende mais*?
- *A diferença de preço da gasolina para o gasóleo influencia os hábitos dos consumidores?*
- *Os carros mais baratos mas com menos qualidade vendem mais*?
- *A região do veículo influencia o preço do carro? Será que uma certa região vende mais*?

Com as respostas às questões realizadas anteriormente, é possível obter o melhor lucro possível consoante as variáveis instroduzidas nas questões. Isto é feito de forma a que se minimize o custo da fabricação do veículo e que se efetua uma boa venda do mesmo. Com isto, é possível criar um modelo de carro que, quando derivado o seu tipo de combustível, o seu custo e a sua região, tenha mais sucesso nas vendas.