Nome: Leonardo Santos Miranda Mestrado

Relatório Final - Trabalho Prático 2

Agrupamentos das características de um carro

Base de dados e contextualização

A base de dados escolhida foi encontrada no Kaggle, denominada como Car Features and MSRP (https://www.kaggle.com/CooperUnion/cardataset). A base de dados está no formato CSV e é composta por 16 atributos e 11915 registros diferentes, fornecidos pelos websites Edmunds e Twitter, que descrevem um carro. Dentre os atributos, temos a montadora, o modelo do carro, o ano de fabricação, o tipo do motor de gasolina, o tipo de transmissão, o número de portas, o tamanho do carro e entre outros. O dataset foi escolhido por possuir uma documentação clara e com um bom número de registros e atributos de fácil entendimento, sendo possível realizar as operações de agrupamentos no custo médio de gasolina em uma rodovia, por exemplo.

Objetivo

Agrupar característica de um carro a partir do dataset definido, utilizando os algoritmos K-Means e Mistura Gaussiana. As características desejadas para o agrupamento são:

- Custo médio de gasolina na cidade e rodovia (highway e city mpg)
- Poder do motor e quantidade de cilindros (Engine hp e cylinders)

Motivação

Através das análises dos agrupamentos das características pode-se ter uma base de por exemplo, o quão agrupados ou variados estão determinados atributos com relação aos outros carros (seja da mesma montadora ou diferente). Desta forma, a montadora em questão pode levantar suposições através dos resultados dos agrupamentos e melhorar como montadora no geral.

Trabalhos Relacionados

Dentre os trabalhos relacionados com a base de dados, alguns notebooks na página do Kaggle (https://www.kaggle.com/CooperUnion/cardataset/notebooks) foram importantes para um maior entendimento dos dados, dentre eles o trabalho que mais se destacou, sendo relevante para este trabalho prático foi o:

- Exploratory Data Analysis, do usúario Tharun

Neste trabalho foi realizada uma exploração completa dos dados utilizando as bibliotecas do Python, obtendo informações relevantes como:

```
df = pd.read_csv("../input/cardataset/data.csv")
df.head(5)
```

- Uma lista completa dos atributos e seus tipos

```
df.dtypes
```

```
Make                object
Model               object
Year                 int64
Engine Fuel Type    object
Engine HP          float64
Engine Cylinders   float64
Transmission Type   object
Driven_Wheels       object
Number of Doors    float64
Market Category     object
Vehicle Size        object
Vehicle Style       object
highway MPG          int64
city mpg             int64
Popularity           int64
MSRP                 int64
dtype: object
```

- Quantidade de linhas repetidas no dataset

```
duplicate_rows_df = df[df.duplicated()]
print("number of duplicate rows: ", duplicate_rows_df.shape)
```

```
number of duplicate rows:  (989, 10)
```

- Se há algum atributo com dados nulos (quantidade)

```
print(df.isnull().sum())
```

```
Make            0
Model           0
Year            0
HP             69
Cylinders      30
Transmission    0
Drive Mode      0
MPG-H           0
MPG-C           0
Price           0
```

Metodologia

A metodologia seguida foi inspirada no CRISP-DM e ilustrada por fluxos de trabalhos realizados na plataforma Lemonade.

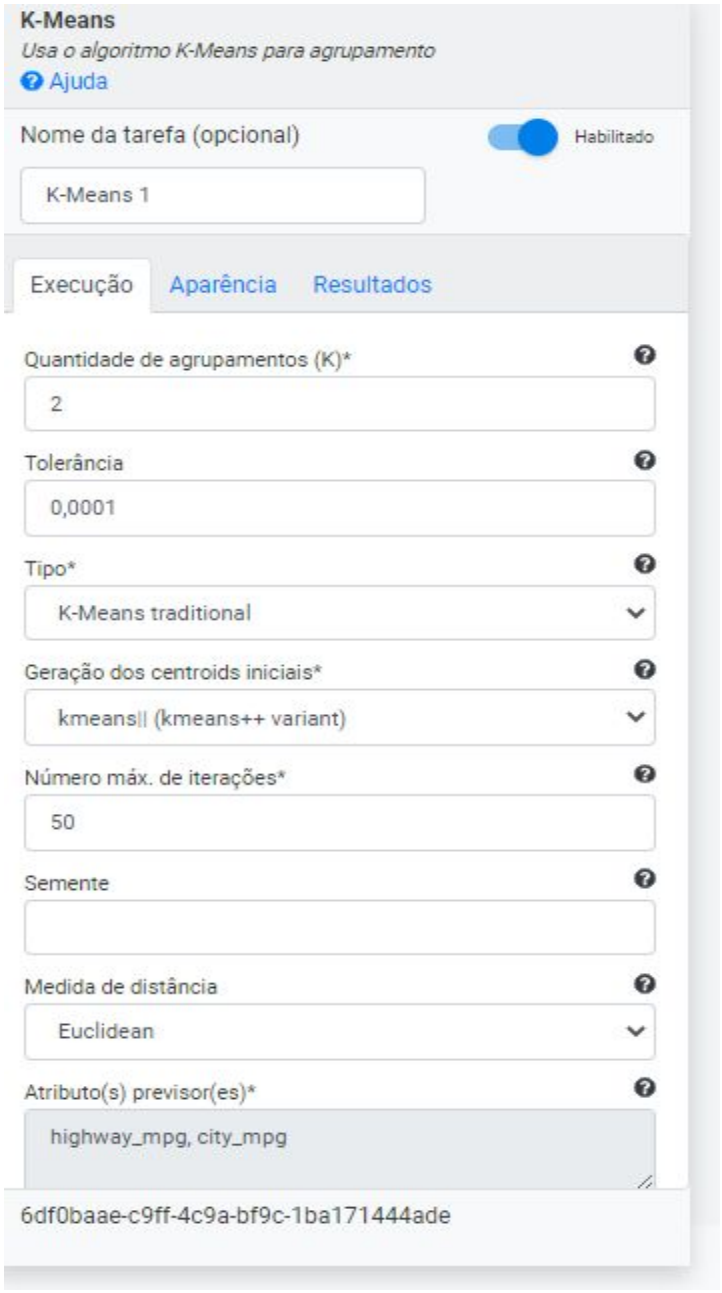
Resultados experimentais e análise

Nesta seção haverá imagens contendo ilustrações dos resultados obtidos através dos fluxos de trabalhos realizados na plataforma Lemonade, com aos agrupamentos descritos na seção de objetivo.

- **Agrupamento 1:** Custo médio de gasolina na cidade e rodovia (highway e city mpg)

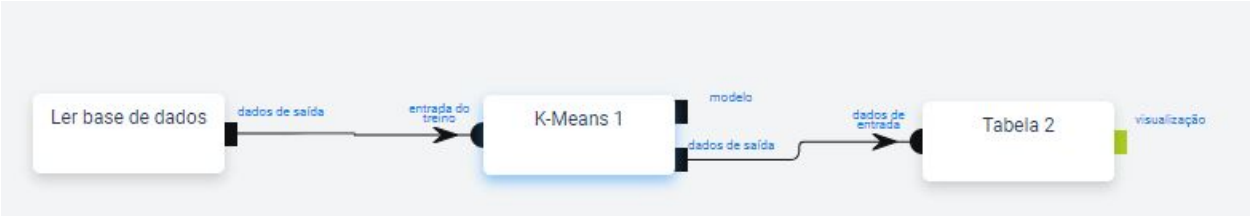
Utilizando o agrupamento K-Means, os seguintes parâmetros foram utilizados:

Figura 1 - Parâmetros utilizados para o agrupamento 1, no algoritmo K-Means:



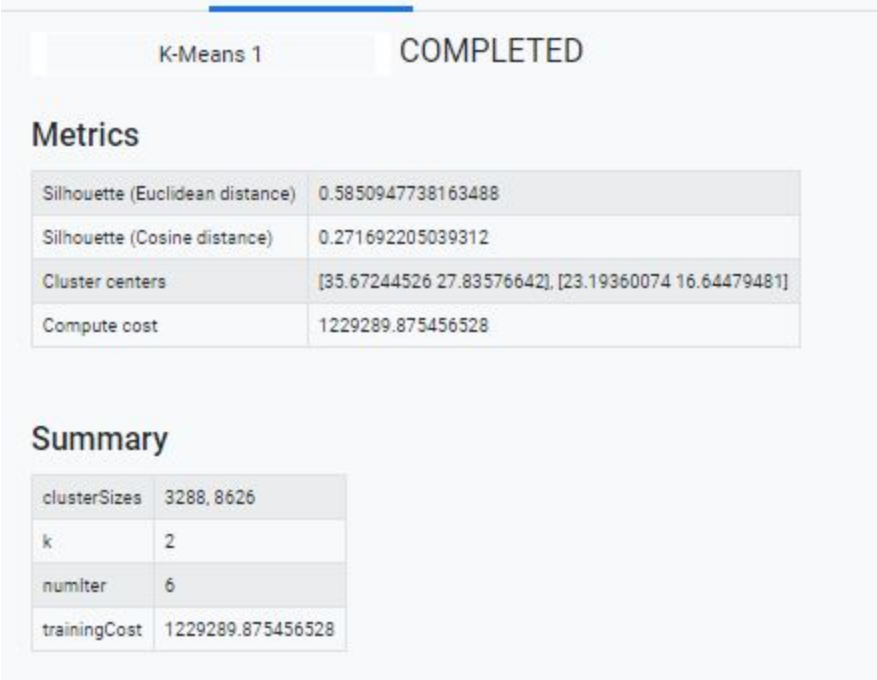
- Fluxo de trabalho:

Figura 2 - Fluxo de trabalho para o agrupamento 1



Executando o fluxo, os dois clusters foram definidos separando a base de dados em dois diferentes grupos. Os valores dos clusters para o agrupamento do custo médio de gasolina na cidade e rodovia (highway e city mpg) foram [23,35] respectivamente, como pode ser visto em maiores detalhes na figura 3:

Figura 3 - Detalhes da execução do fluxo de trabalho



E visualizando os resultados que foram gerados na tabelas, a coluna “Prediction” descreve em qual agrupamento o dado foi definido:

Figura 4 - Visualização dos resultados gerados pela tabela do agrupamento 1.

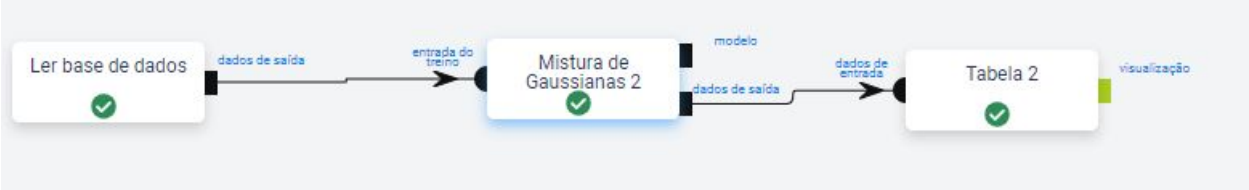
Results

Make	Model	Year	Engine fuel type	Engine hp	Engine cylinders	Transmission type	Driven wheels	Number of doors	Market category	Vehicle size	Vehicle style	Highway mpg	City mpg	Popularity	Msrp	Predicl
Mercedes-Benz	190-Class	1993	regular unleaded	158	6	MANUAL	rear wheel drive	4	Luxury	Compact	Sedan	25	17	617	2000	1
BMW	2 Series	2016	premium unleaded (required)	240	4	AUTOMATIC	rear wheel drive	2	Luxury,Performance	Compact	Coupe	35	23	3916	32850	0
BMW	2 Series	2016	premium unleaded (required)	240	4	AUTOMATIC	rear wheel drive	2	Luxury	Compact	Convertible	34	23	3916	38650	0
BMW	2 Series	2016	premium unleaded (required)	320	6	AUTOMATIC	rear wheel drive	2	Factory Tuner,Luxury,High-Performance	Compact	Convertible	31	20	3916	48750	1
BMW	2 Series	2016	premium unleaded (required)	240	4	AUTOMATIC	all wheel drive	2	Luxury,Performance	Compact	Coupe	35	23	3916	34850	0
BMW	2 Series	2016	premium unleaded (required)	240	4	AUTOMATIC	all wheel drive	2	Luxury	Compact	Convertible	34	22	3916	40650	0
BMW	2 Series	2016	premium unleaded (required)	320	6	AUTOMATIC	rear wheel drive	2	Factory Tuner,Luxury,High-Performance	Compact	Coupe	31	20	3916	44150	1
BMW	2 Series	2016	premium unleaded (required)	240	4	MANUAL	rear wheel drive	2	Luxury,Performance	Compact	Coupe	34	22	3916	32850	0
BMW	2 Series	2016	premium unleaded (required)	320	6	AUTOMATIC	all wheel drive	2	Factory Tuner,Luxury,High-Performance	Compact	Coupe	30	20	3916	46150	1
BMW	2 Series	2016	premium unleaded (required)	320	6	AUTOMATIC	rear wheel drive	2	Factory Tuner,Luxury,High-Performance	Compact	Convertible	30	20	3916	50750	1

Além da utilização do algoritmo K-Means, também queria testar utilizando a mistura Gaussiana para ver o quanto os resultados iriam divergir. Dito isso, segue abaixo as figuras contendo os detalhes da execução:

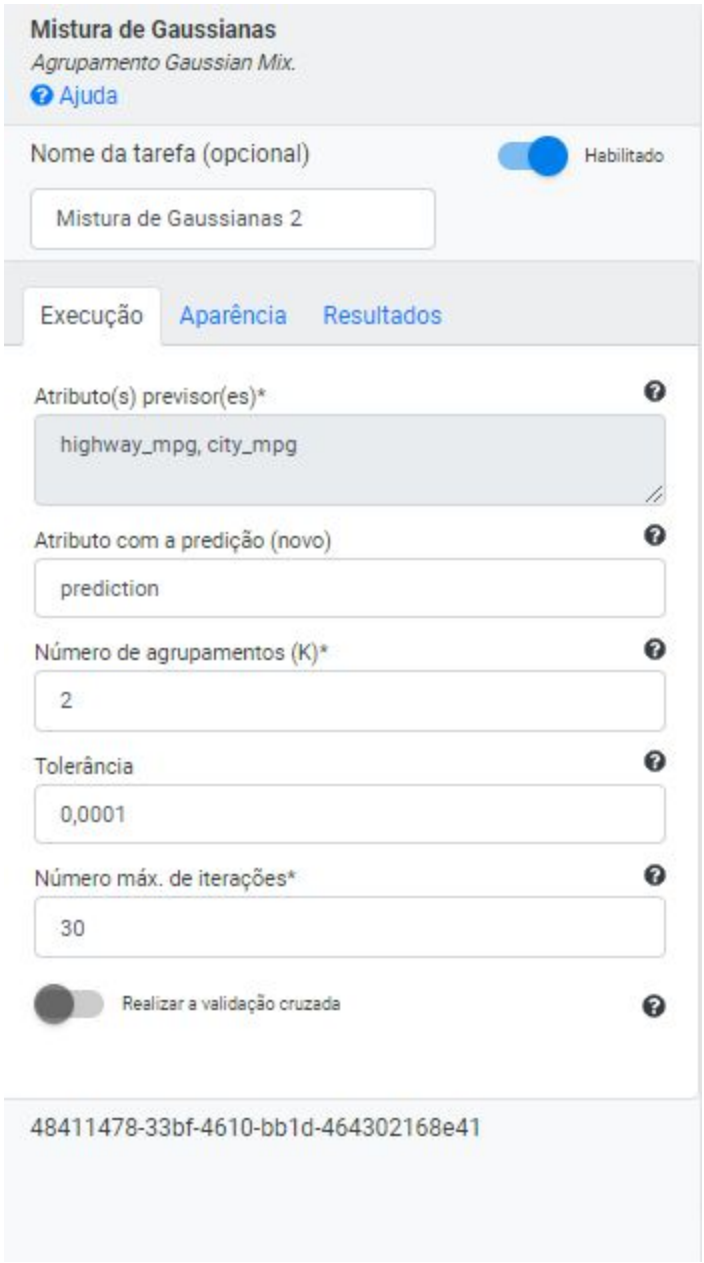
- Fluxo de trabalho:

Figura 5 - Fluxo de trabalho para o agrupamento 1



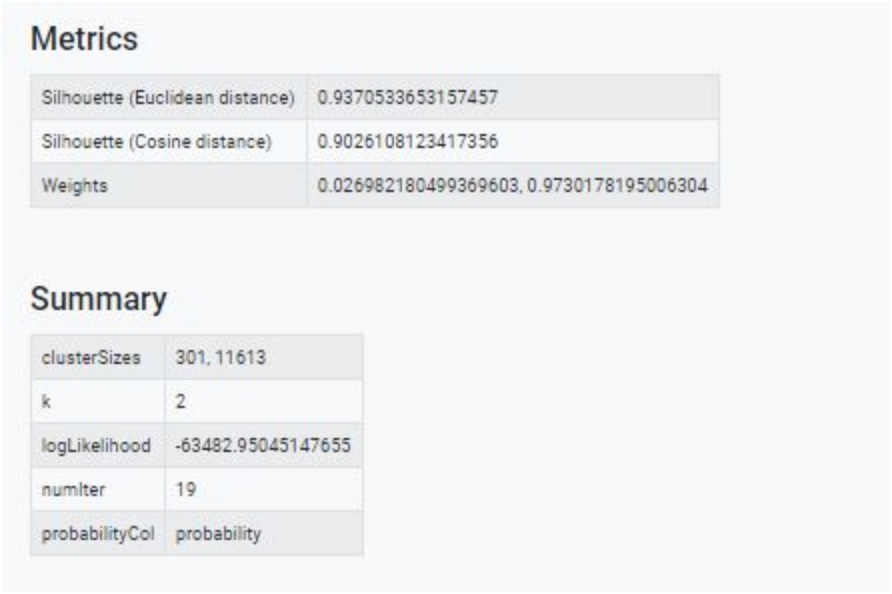
- Os seguintes parâmetros foram utilizados no algoritmo de mistura de Gaussiana:

Figura 6 - Parâmetros utilizados para o agrupamento 1, na mistura de Gaussiana



Executando o fluxo acima, o modelo foi treinado conforme os parâmetros e os clusters foram definidos:

Figura 7 - Detalhes da execução do fluxo de trabalho



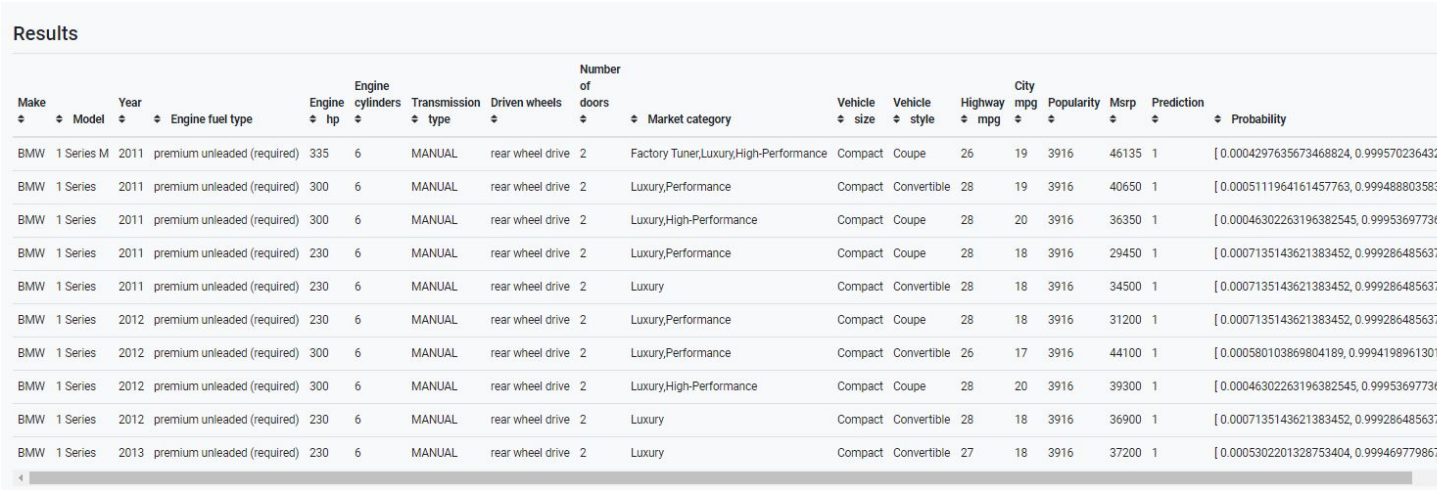
Metrics

Silhouette (Euclidean distance)	0.9370533653157457
Silhouette (Cosine distance)	0.9026108123417356
Weights	0.026982180499369603, 0.9730178195006304

Summary

clusterSizes	301, 11613
k	2
logLikelihood	-63482.95045147655
numIter	19
probabilityCol	probability

Figura 8 - Visualização dos resultados gerados pela tabela do agrupamento 1.



Results

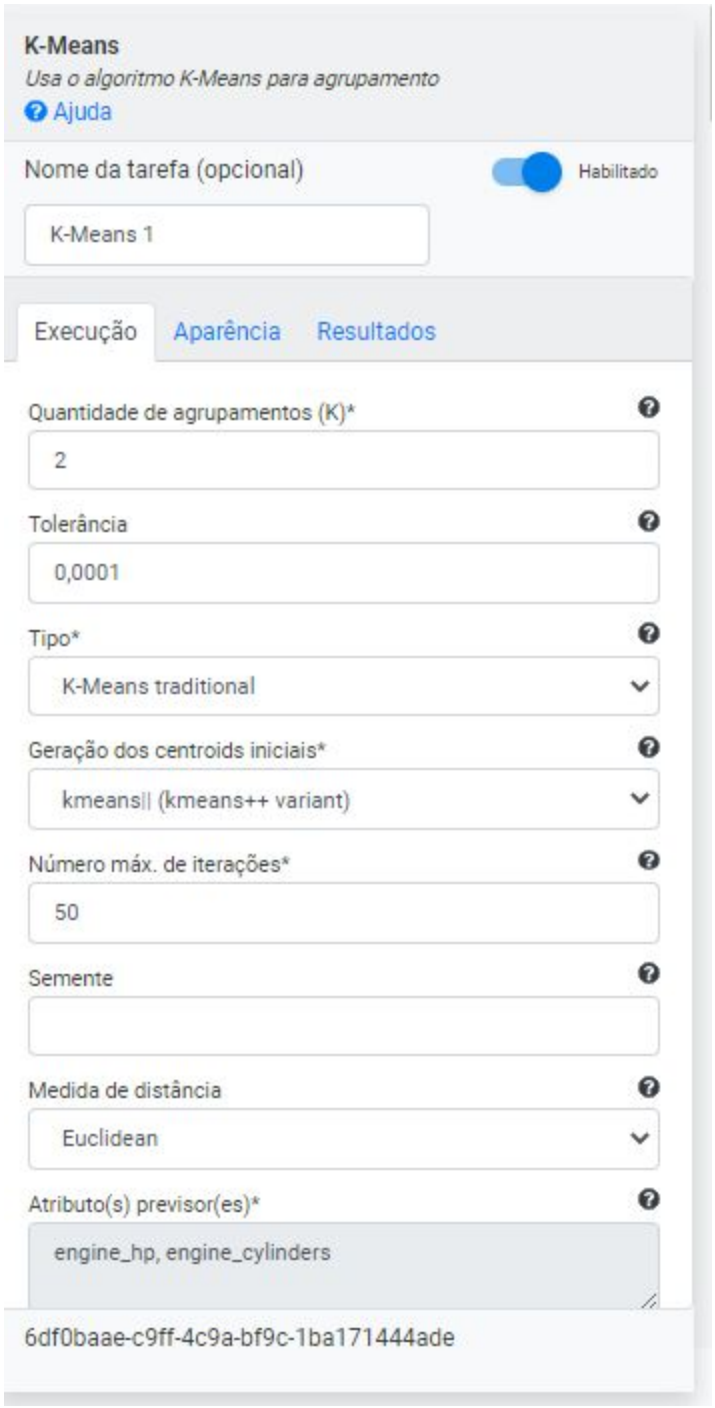
Make	Model	Year	Engine fuel type	Engine hp	Engine cylinders	Transmission type	Driven wheels	Number of doors	Market category	Vehicle size	Vehicle style	Highway mpg	City mpg	Popularity	Msrp	Prediction	Probability
BMW	1 Series M	2011	premium unleaded (required)	335	6	MANUAL	rear wheel drive	2	Factory Tuner,Luxury,High-Performance	Compact	Coupe	26	19	3916	46135	1	[0.000429763567346882
BMW	1 Series	2011	premium unleaded (required)	300	6	MANUAL	rear wheel drive	2	Luxury,Performance	Compact	Convertible	28	19	3916	40650	1	[0.000511196416145776
BMW	1 Series	2011	premium unleaded (required)	300	6	MANUAL	rear wheel drive	2	Luxury,High-Performance	Compact	Coupe	28	20	3916	36350	1	[0.000463022631963825
BMW	1 Series	2011	premium unleaded (required)	230	6	MANUAL	rear wheel drive	2	Luxury,Performance	Compact	Coupe	28	18	3916	29450	1	[0.000713514362138345
BMW	1 Series	2011	premium unleaded (required)	230	6	MANUAL	rear wheel drive	2	Luxury	Compact	Convertible	28	18	3916	34500	1	[0.000713514362138345
BMW	1 Series	2012	premium unleaded (required)	230	6	MANUAL	rear wheel drive	2	Luxury,Performance	Compact	Coupe	28	18	3916	31200	1	[0.000713514362138345
BMW	1 Series	2012	premium unleaded (required)	300	6	MANUAL	rear wheel drive	2	Luxury,Performance	Compact	Convertible	26	17	3916	44100	1	[0.000580103869804189
BMW	1 Series	2012	premium unleaded (required)	300	6	MANUAL	rear wheel drive	2	Luxury,High-Performance	Compact	Coupe	28	20	3916	39300	1	[0.000463022631963825
BMW	1 Series	2012	premium unleaded (required)	230	6	MANUAL	rear wheel drive	2	Luxury	Compact	Convertible	28	18	3916	36900	1	[0.000713514362138345
BMW	1 Series	2013	premium unleaded (required)	230	6	MANUAL	rear wheel drive	2	Luxury	Compact	Convertible	27	18	3916	37200	1	[0.000530220132875340

Com os resultados obtidos após a execução da mistura de Gaussiana eu pude comparar com os que eu obtive através do K-Means. Essa comparação me fez entender mais como ambos os algoritmos se comportam e como se diferenciam entre si.

- **<u>Agrupamento 2</u>:** Poder do motor e quantidade de cilindros (Engine hp e cylinders)

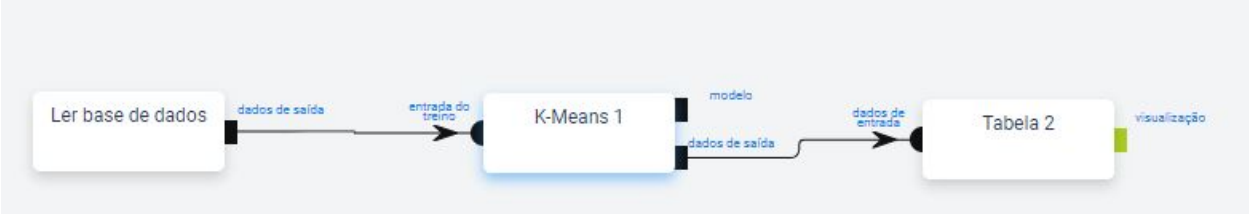
Utilizando o agrupamento K-Means, os seguintes parâmetros foram utilizados:

Figura 9 - Parâmetros utilizados do agrupamento 2



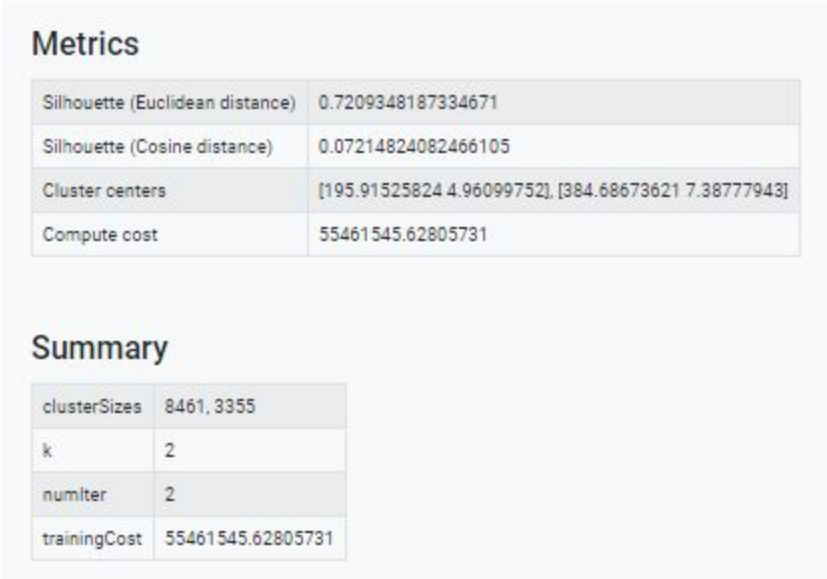
Com o fluxo de trabalho:

Figura 10 - Fluxo de trabalho para o agrupamento 2



Utilizei os mesmos parâmetros no algoritmo K-Means para testar como seria o comportamento dos clusters quando a variação dos números são diferentes - geralmente os cilindros são entre 4~~10 e o poder do motor variam de 100 até 500. Executando o fluxo, os dois clusters foram definidos separando a base de dados em dois diferentes grupos. Os valores dos clusters para o agrupamento do número de cilindros do motor e poder do motor (engine_hp e engine_cylinders) foram [195,384] respectivamente, como pode ser visto em maiores detalhes na figura 3:

Figura 11 - Detalhes da execução do fluxo de trabalho para o agrupamento 2



Metrics

Silhouette (Euclidean distance)	0.7209348187334671
Silhouette (Cosine distance)	0.07214824082466105
Cluster centers	[195.91525824 4.96099752], [384.68673621 7.38777943]
Compute cost	55461545.62805731

Summary

clusterSizes	8461, 3355
k	2
numIter	2
trainingCost	55461545.62805731

E Visualizando os resultados da tabela:

Figura 12 - Visualização dos resultados gerados pela tabela do agrupamento 1.

Results

Make	Model	Year	Engine fuel type	Engine hp	Engine cylinders	Transmission type	Driven wheels	Number of doors	Market category	Vehicle size	Vehicle style	Highway mpg	City mpg	Popularity	Msrp	Prediction
BMW	1 Series M	2011	premium unleaded (required)	335	6	MANUAL	rear wheel drive	2	Factory Tuner,Luxury,High-Performance	Compact	Coupe	26	19	3916	46135	1
BMW	1 Series	2011	premium unleaded (required)	300	6	MANUAL	rear wheel drive	2	Luxury,Performance	Compact	Convertible	28	19	3916	40650	1
BMW	1 Series	2011	premium unleaded (required)	300	6	MANUAL	rear wheel drive	2	Luxury,High-Performance	Compact	Coupe	28	20	3916	36350	1
BMW	1 Series	2011	premium unleaded (required)	230	6	MANUAL	rear wheel drive	2	Luxury,Performance	Compact	Coupe	28	18	3916	29450	0
BMW	1 Series	2011	premium unleaded (required)	230	6	MANUAL	rear wheel drive	2	Luxury	Compact	Convertible	28	18	3916	34500	0
BMW	1 Series	2012	premium unleaded (required)	230	6	MANUAL	rear wheel drive	2	Luxury,Performance	Compact	Coupe	28	18	3916	31200	0
BMW	1 Series	2012	premium unleaded (required)	300	6	MANUAL	rear wheel drive	2	Luxury,Performance	Compact	Convertible	26	17	3916	44100	1
BMW	1 Series	2012	premium unleaded (required)	300	6	MANUAL	rear wheel drive	2	Luxury,High-Performance	Compact	Coupe	28	20	3916	39300	1
BMW	1 Series	2012	premium unleaded (required)	230	6	MANUAL	rear wheel drive	2	Luxury	Compact	Convertible	28	18	3916	36900	0
BMW	1 Series	2013	premium unleaded (required)	230	6	MANUAL	rear wheel drive	2	Luxury	Compact	Convertible	27	18	3916	37200	0

Conclusões e perspectivas

No desenvolvimento deste trabalho comecei a entender como os algoritmos de agrupamento se comportam na prática, não somente na teoria como nas vídeo aulas que assisti. Também percebi que na proposta de trabalho prático que eu entreguei havia tentativas de agrupamento para características do tipo categórico, o que eu não consegui realizar através do agrupamento K-Means ou na Mistura de Gaussiana, fornecidos pelo Lemonade. Espero que os resultados que obtive através das diversas execuções realizadas no Lemonade sejam satisfatórios e tenham relevância para este trabalho prático. Vou anexar juntamente com este pdf uma pasta com todas as imagens dos resultados obtidos, caso sejam difíceis de visualizar.