Envirocar visualization exploring an environmental and traffic data set

Rodrigo Claro Zembrzuski
Programa de Pos Graduacao em Ciencia da Computacao
Universidade Federal do Rio Grande do Sul - UFRGS
Porto Alegre, Brasil
Web page: http://inf.ufrgs.br/∼rczembrzuski

***Abstract*—Envirocar eh um projeto que coleta e armazena dados ambientais e dados de trafego de automoveis. Sao dados de sensores implantados dentro dos veiculos que trazem informacoes precisas sobre navegacao dos carros e dados relativos a conservacao do meio-ambiente, como emissao de gas-carbonico e consumo do veiculo.**

Esse artigo tem por objetivo apresentar uma serie de visualizacoes para analise exploratoria desse banco de dados. Com isso, teremos uma intuicao sobre se esse banco de dados pode servir para alguma coisa, ou nao.

***Keywords*-envirocar; smartcars; kibana; geolocation; data visualization;**

I. Introduction

Envirocar eh um projeto da Universidade de Munster. O objetivo do projeto eh oferecer uma plataforma simples para coleta, armazenamento e distribuicao de dados ambientais e dados de logistica de trafego [ref].

A plataforma une dados gerados por sensores comuns, que vem na maioria dos automoveis atuais com sensores de geolocalizacao, presente na maioria dos smartphones. Com isso, cria-se uma base de dados relevante para se medir o fluxo de automoveis nas avenidas e quantidade de emissao de poluentes nas cidades.

A. Motivacao

Um dos grandes desafios da sociedade nos nossos dias eh preservar e desenvolver a mobilidade urbana de maneira sustentavel, isto eh, mitigar o impacto do transporte de passageiros meio-ambiente [ref].

Alem disso, observar padroes de comportamento nos motoristas e nas cidades ao longo do mundo pode ser interessante tanto para ajuda-los em alguma coisa, quanto para ver diferencas culturais de paises e culturas diferrentes evidenciadas no modo em que dirigem.

B. Objetivos

O objetivo desse trabalho eh fazer uma exploracao basica pela base dados para ganhar intuicao sobre ela. Para observar se padroes basicos de comportamento dos usuarios podem ser detectados conforme ideias basicas que se tem sobre isso e, por fim, para saber se a base de dados pode ser utilizada como base para experimentos mais complexos, como treinamento de modelos de aprendizagem para um ganho no mundo relevante, como melhorias no trafego, detecao de anomalias em comportamentos e outras cositas mais.

Contributions: Acredito que esse trabalho possa dar um insight interessante sobre o quao relevante esse dataset pode ser para a observacao de comportamento dos motoristas.

C. Resultados esperados

In order to produce this application, we start with this processing, followed by this technique. In order to cope with this challenge, we introduce this formulation to produce this intermediate result. The formulation leads to this type of system, which is efficiently solved by adapting this technique. The final result is produced by this transform. The whole process is schematized in Fig. **??**.

II. Caracterizacao dos dados

Os dados sao obtidos atraves de sensores comuns em diversos automoveis (OBD-II) e enriquecidos com informacoes de geolocalizacao que estao presentes em smartphones. Para preservar a identidade dos usuarios, suas informacoes pessoais nao sao disponibilizadas e tambem sao retiradas as informacoes dos primeiros e ultimos 200m de deslocamento, com objetivo que o usuario nao seja identificado atraves das informacoes de origem e de destino.

Uma vez que os automoveis estao em deslocamento, suas informacoes mudam rapidamente. Desse modo, informacoes populadas no dataset a cada 2 segundos.

Os dados originais do servidor sao persistidos num banco de dados NoSQL – MongoDB – e oferecidos abertos ao publico atraves de um RESTful web service. Os dados sao oferecidos em um formato JSON, que eh simples de entender e processar.

Para realizacao desse trabalho, foram acessadas as apis rest e persistidas as informacoes de cada deslocamento no banco de dados ElasticSearch, por motivos que serao esclarecidos em seguida. Entretanto os dados originais foram mantidos e enriquecidos com algumas outras informacoes para que ficassem mais faceis de processa-los.

A. Caracterizao dos dados originais

Cada item do dataset eh composto por dois atributos principais, que sao subdivididos em diversas partes e serao mais detalhados na sequencia: 'properties' e 'features'.

TABLE I
PROPERTIES: PROPRIEDADES GERAIS DE CADA VEICULO

Variavel	Tipo da variavel	Descricao
type	categorica	tipo de veiculo
constructionYear	discreta	ano de construcao
model	categorica	modelo do veiculo
fuelType	categorica	tipo de combustivel
engineDisplacement	discreta	cilindrada do motor
manufactuerer	categorica	fabricante

- properties: possui as caracteristicas gerais do veiculo em questao. Esta subdividido nos itens apreseentados na tabela 1.
- features: possui as caracteristicas do deslocamento – ou da 'viagem' em si. A viagem eh caracterizada pela composicao, a cada dois segundos, do seguinte conjunto de dados:
  - coordinates – geoespacial – lat/long
  - speed – continuo – km/h
  - rpm – continuo – u/min
  - gps accuracy – nao sei – nao sei
  - maf – nao sei – nao sei
  - engine load – nao sei – nao sei
  - gps pdop – nao sei – nao sei
  - o2 lambda voltage – nao sei – nao sei
  - throttle position – nao sei – nao sei
  - consumption – continuo – l/h
  - gps vdop – nao sei – nao sei
  - gps speed – continuo – km/h
  - gps bearing – nao sei – nao sei
  - intake pressure – nao sei – nao sei
  - co2 – atributo continuo – kg/h
  - time – atributo temporal – timestamp

B. Questoes a serem respondidas

Existe uma serie de questoes que me vem rah cabeca nesse instante que, inclusive, ja tomei notas.

- quais marcas/modelos consomem mais gasolina?
- quais marcas/modelos poluem mais o meio ambiente?
- qual o perfil de pilotagem de cada automovel?
- muita gente usa esse dataset?
- de que lugares sao as pessoas que utilizam esse sistema? no Brasil, no mundo?
- existe muita gente utilizando esse sistema hoje em dia?
- quais os modelos e fabricantes mais comuns?
- os modelos e fabricantes mais comum na Europa sao os mesmos do Brasil?
- podemos descobrir quais sao as regioes mais poluidas e regioes menos poluidas?

III. PREPARACAO DOS DADOS PARA VISUALICAO

O envirocar eh uma base aberta de dados para quem quiser usar. Simples assim. Entretanto, eles nao colocam seu banco de dados para que seja feito download de forma simples. Eles expoem um servico REST que, de forma simplificada, oferece um acesso facil aos dados.

Para que eu pudesse fazer a visualizacao de maneira geral, primeiramente eu fiz um script que baixava cada uma das viagens para minha maquina e as indexava no elasticsearch. O script rodou por horas, porque eu fiz download de maneira sequencial, ate para nao acocar o servidor dos caras.

A base de dados ate nao eh muito grande. Sao 15 mil viagens, o que nao parece muito. Entretanto, o volume de dados eh relativamente grande, pois para cada viagem, eh enviado para o banco de dados uma serie de informacoes a cada 2 segundos. Entao, embora o numero de viagens seja relativamente pequeno, o tamanho de cada viagem eh grande. Para se ter uma ideia, minha base indexada no elasticsearch ocupou 6 Gb de disco.

Alem disso, para poder facilitar as visualizacoes, eu fiz uma serie de metricas. Por exemplo, para comparar quais carros emitem mais gas carbonito, por exemplo, para cada viagem eu computei a media do gas carbonico emitido naquela viagem, pois a base de dados nao oferece issos. A base de dados oferece uns dados longos e, se eu quiser manter isso no artigo, vou ter que dar um jeito de explicar muito bem, porque ficou uma bosta, no fim das contas.

IV. TECNICA DE VISUALIZAO DESENVOLVIDA

A exploracao dos dados foi feito com com o uso de duas ferramentas: Kibana e R.

Kibana, tipicamente, eh utilizada para analise por inspeo manual e visualizao de informaes presentes no ElasticSearch (MARINO, 2015; VAARANDI; PIHELGAS, 2014). Dessa forma, eh ela que ira apresentar os dados armazenados pelo Logstash no ElasticSearch, em uma interface, via browser, altamente customizvel com histogramas e outros painis que propiciam uma viso geral sobre os dados. O Kibana possibilita transformar os logs em informaes teis (valor) atravs de Dashboards, pois permite realizar correlao de eventos, filtrar logs por origem, hospedeiros, entre outras combinaes (VAARANDI; NIZINSKI, 2013).

R eh uma linguagem muito utilizada para analises estatisticas e tambem possue uma gama vasta para visualizacao de dados. Foi utilizada para produzir graficos menos interessantes ao olho; mais toscas em questoes de boniteza, mas com uma representatividade muito boa. Na verdade, tenho que escrever um texto legal aqui sobre como o R eh legal para visualizacao de dados. Ah, o que eu tenho que falar eh que usei o R quando eu achei que o Kibana era insuficiente para fazer uma analise mais profunda. Acho que posso concluir, la no final, que o R tem muito mais flexibilidade, maleabilidade e estensibilidade que o Kibana, porque nao consegui fazer matriz de correlacoes nem boxplots com o Kibana.

A maior vantagem do kibana sobre o R eh que ele casa muito bem com o elasticsearch. Mas eu ja sei tanto de elasticsearch que nao me importo com isso.

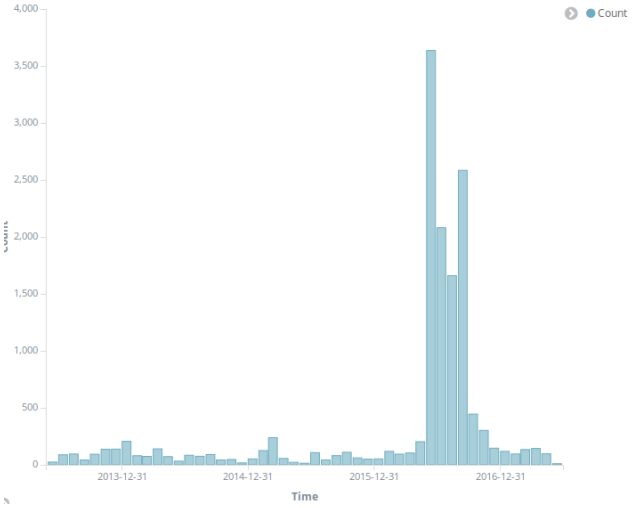
A. Histograma

Com o intuito de saber se muita gente ou se pouca gente utiliza a plataforma envirocar, foi plotado um histograma. Com

ele, alem de saber se muita gente ou se pouca gente o utiliza, podemos ver a evolucao de sua utilizacao ao longo do tempo.

Observando o grafico abaixo, eu vejo uma coisa meio ruim em relacao a ferramenta. Ela foi muito utilizada no final de 2006, chegando a 3600 viagens no periodo de um mes, mas foi somente naquele periodo. Em novembro de 2006, ja passou a acontecer muito poucas viagens de automovel, e em 2017 temos cerca de 100 viagens por mes.

Nao eh muito mas, ainda, acredito que possa ser utilizada com o proposito que eu quero, que eh de tentar agrupar os motoristas e identificar seus padores de comportamento.



No eixo x, observamos a quantidade de uso da plataforma ao longo do tempo.

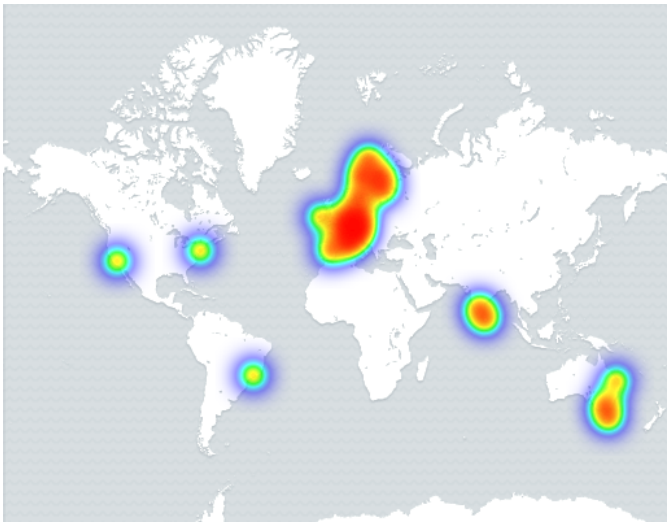
B. Heatmaps

Outra duvida que eu tinha no inicio era a distribuicao geografica das pessoas que utilizam a plataforma. Como a plataforma foi desenvolvida na Alemanha, supus que a maioria das pessoas estavam utilizando por l. Estava realmente curioso para saber se existem pessoas em outros lugares do mundo a utilizando tambem. Para isso, um heatmap veio a calhar.

Como se pode observar no grafico de heatmaps, eu estava supondo corretamente. A maioria das pessoas que utilizam essa plataforma se encontra na Europa. Entretando, se observa algumas coisas nao imaginadas observando esse grafico. Observa-se pontos isolados nos Estados Unidos, Brasil, India e Australia.

Essa visualizacao eh bem importante pois, para algumas outras visualizacoes que vamos ver a seguir, foi assim que eu dividi o globo. Em 5 regioes – ou serao 6 – bem definidas, baseados a partir desse heatmap.

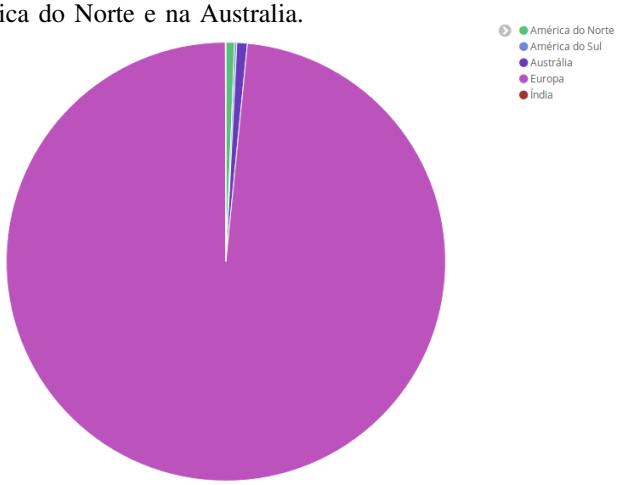
Outra coisa legal de notar eh que o heatmap deixou claro na nossa mente os pontos do globo em que mais se usa essa base de dados, mas nao deixou muito clara a proporcao de pessoas que a utilziam nas diferentes partes do globo.



Aqui, podemos observar que a maioria dos usuarios, de fato, se encontra, na Europa. Contudo, existem pessoas utilizando em diversos outros pontos do globo tambm.

C. Grafico em pizza

Dada minha duvida sobre as proporcoes das pessoas que utilizam o envirocar em diferentes partes do globo, acreditei que um siples grafico de pizza iria mostrar claramente as proporcoes. Para isso, dividi baseado no heatmap o globo nas 6 regioes que la apareceram: America do Norte, America do Sul, Australia, Europa, India. Com isso, criei um grafico de pizza, que mostra que, de fato, a maioria esmagadora das pessoas esta na Europa. Existe alguma representatividade tambem na America do Norte e na Australia.



98% dos deslocamentos esta na Europa. 1% esta na America do Norte e 1% esta na Australia. Outras regioes sao sao insignificantes, que nem visualizamos no grafico .

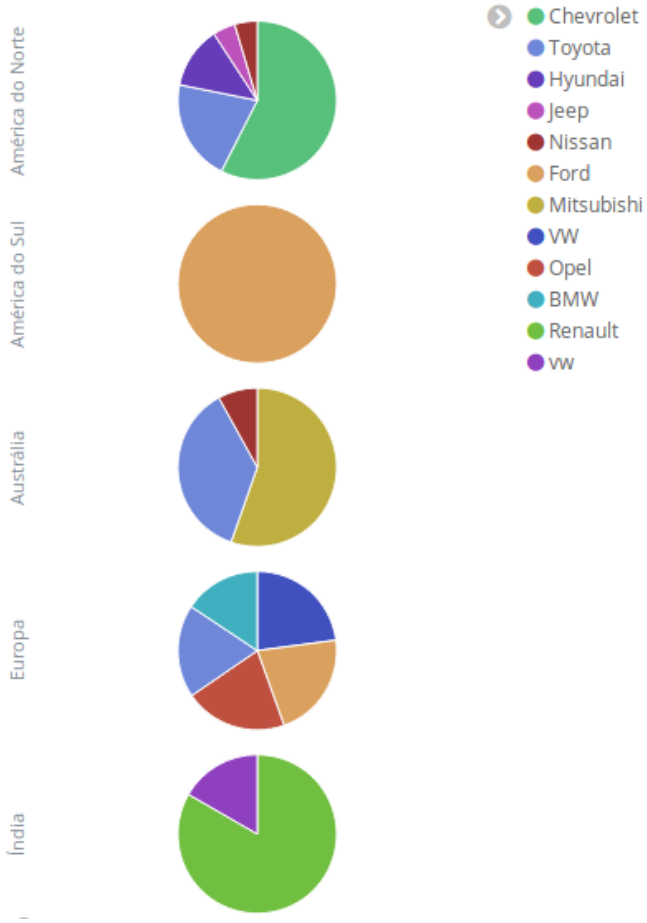
D. Graficos em pizza

Os graficos em pizza sao tao simples e interessantes, que resolvi fazer outro. Estava na duvida sobre quais sao as marcas de automovel mais utilizadas em diferentes partes do globo, que resolvi fazer um grafico de pizza separado em linhas. Achei que a divisao por sunburst ficava um pouco complicada de visualizar; entao, fiz n graficos em pizza, em que cada

grafico corresponde a uma regiao do globo e eh exibida a fatia de marcas por regiao.

Pensei em fazer essa visualizacao para ver se eu conseguiria ver algum padrao de comportamento diferente dependendo da regiao. Por exemplo, se eh verdade que americanos gostam de carros maiores. Se londrinos andam de carros pequenos que nem o mr bean, ou se eh tudo uma bobagem.

Baseado nessa amostra de dados, nao consegui chegar a uma conclusao definiva para isso. Alias, achei parecida a distribuicao dos estados unidos com a da europa. Nas demais regioes, acredito que a amostra simplesmente seja insuficiente para se ter um grafico bacana.



Estados Unidos e Europa apresentam resultados parecidos. America do Sul supoe que somente um cara utilizou a plataforma.

E. Sunburst

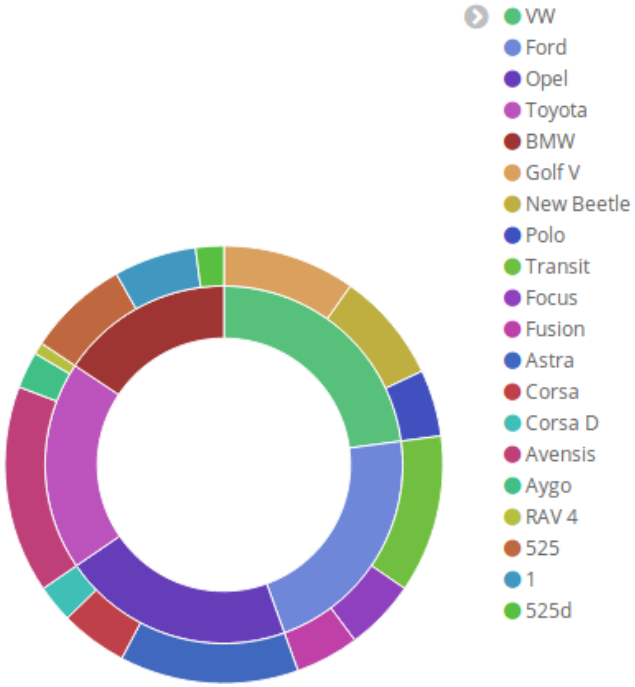
Uma das tecnicas legais vista em aula e que eu gostei muito foi o sunburst. Tambem vi que o kibana oferece uma forma bem simples de implementa-lo. Entao, tinha que achar uma forma legal de utiliza-lo.

Pensei, entao, em fazer o sunburst para ver quais sao os modelos de automovel mais utilizados por marca. Como sao muitas marcas e modelos, achei prudente limitar a 5 o numero de marcas e em 3 o numero de modelos, para nao ficar um grafico muito sobrecarregado.

E achei que ficou um grafico aprazivel de se ver e bem informativo. Ali, podemos ver os 3 modelos mais comuns por fabricante. Vemos que marcas como VW, Ford, Opel (GM) e Toyota possuem praticamente a mesma proporcao. Um pouco atras esta BMW. Achei isso um padrao de comportamento particularmente curioso, pois as principais marcas da europa sao as mesmas do brasil. A diferenca eh que toyota eh bem comum por la, ao contrario daqui e a grande diferenca mesmo eh o BMW. Aqui eu nunca vejo BMW pelas ruas.

Depois, tambem preciso fazer alguma observacao sobre os modelos dos automoveis. Ate porque, se eu nao fizer essa observacao, nao teria razao para ser um sunbsurst. Poderia ser um grafico de pizza. Entretanto, o sunburst eu quero utilizar.

Alem disso, achei o grafico de donuts mais bonito que o grafico de pizza.



Observamos que as marcas mais comuns na Europa sao as mesmas do brasil, exceto pelo BMW. Nunca vemos BMWs pelo Brasil.

F. Radar

Fiz um grafico de radar, primeiramente, porque achei que era um grafico que eu teria que colocar, devido ao fato de eu ter feito a cadeira de visualizacao de informacoes. Eu deveria, pois, colocar uns graficos avancados, em vez de colocar somente graficos simples e toscos.

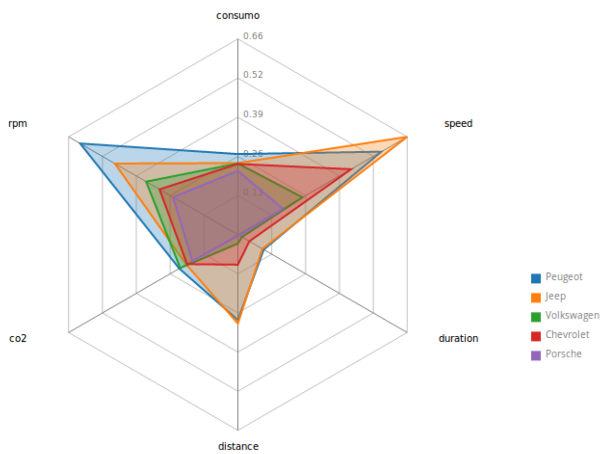
Optei, entao, por fazer uma media dos principais atributos que pude encontrar: co2, rpm, consumo, speed, duration e distance. Entao, calculei a media de cada um desses atributos selecionado por fabricante.

Achei o resultado bastante curioso: Nele se pode observer-var que Peugeot eh o carro que normalmente se anda com a rpm mais alta e eh o cara que tem maior consumo, mas nao eh o cara que anda mais rapido.

O cara que anda mais rapido eh o Jeep, e ele consome menos que os Peugeot, normalmente.

Curioso observar tambem, que o Porshe eh a marca que se anda com rpm mais baixa – ao contrario do que normalmente se imagina – e cuja velocidade eh mais baixa tb. Acredito que isso se de porque o Porshe deve ter uma representatividade muito baixa na amostra. Mas o fato eh que os caras de porshe nao andam rapido.

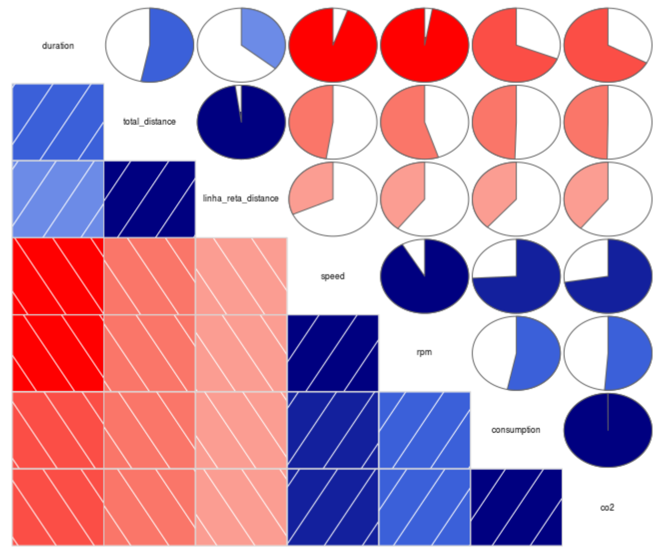
Uma coisa que achei bem interessante eh a chevrolet. Chevrolet parece ser um carro que polui pouco e que anda com a rpm nao muito alta.



Diversos atributos seccionado por fabricante.

G. *Matriz de correlacao*

Aqui eu resolvi plotar, utilizando R, uma matriz de correlacao para que se possa ver como alguns dos principais atributos medidos pelos sensores se relacionam. Nessa representacao, quanto mais azul e quanto mais cheio esta o circulo superior, maior eh a correlacao entre os atributos. Observamos que quanto maior o consumo de combustivel, maior eh a quantidade de gas carbonico emitido. Observamos tambem que existe grande correlacao entre rpm e velocidade, o que me surpreendeu um pouco, ate. Outra coisa que me surpeendeu um pouco foi que parece nao haver correlacao entre velocidade e duracao de uma viagem. Me surpreendeu porque normalmente eu ando em velocidade mais alta quando faco viagens mais longas. Mas acho que na Europa nao eh assim que acontece. Existem outras coisas que se pode observar por ai tb.



Diversos atributos seccionado por fabricante.

V. Resultados

Achei interessante que na Europa, a Toyota aparece como uma marca comum. BMW segue sendo top por la. Fazer mais resultados. Fazer mais resultados. Fazer mais resultados. Fazer mais resultados. Fazer mais resultados. Fazer mais resultados. Fazer mais resultados. Fazer mais resultados. Fazer mais resultados. Fazer mais resultados. Fazer mais resultados. Fazer mais resultados. Fazer mais resultados. Fazer mais resultados. Fazer mais resultados. Fazer mais resultados.

VI. Conclusao

Aqui vai minha conclusao. Aqui vai minha conclusao.

Acknowledgment

The authors would like to thank this colleague and this financing institute.