

# Envirocar visualization

## exploring an environmental and traffic data set

Rodrigo Claro Zembruski  
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/~rczembruski>

**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 diferentes 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 coisas 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.

#### A. Caracterizacao geral dos dados

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'.

- **properties:** possui as caracteristicas gerais do veiculo em questao. Esta subdividido nos itens apresentados na tabela 1. Veja um exemplo de propriedades extraido do banco de dados.

TABLE I  
PROPERTIES: PROPRIEDADES GERAIS DE CADA VEICULO

Variavel	Descricao
type	categorica
constructionYear	discreta
model	categorica
fuelType	categorica
engineDisplacement	discreta
manufacturer	categorica

TABLE II  
FEATURES: INFORMACOES DE CADA TIMESTAMP DO DESLOCAMENTO

Variavel	Unidade de medida
coordinates	geoespacial
speed	continuo
rpm	continuo
gps accuracy	continuo
maf	continuo
engine load	continuo
gps pdop	continuo
o2 lambda voltage	continuo
throttle position	continuo
consumption	continuo
gps vdop	continuo
gps speed	continuo
gps bearing	continuo
intake pressure	continuo
co2	continuo
time	temporal

```

"properties": {
  "sensor": {
    "type": "car",
    "properties": {
      "constructionYear": 2011,
      "model": "Avensis",
      "fuelType": "gasoline",
      "id": "574e78cbe4b09078f97bbb4a",
      "engineDisplacement": 1800,
      "manufacturer": "Toyota"
    }
  }
}

```

- features: possui as características do deslocamento – ou da ‘viagem’ em si. A viagem é caracterizada pela composição, a cada dois segundos, do seguinte conjunto de dados:

```

{
  "geometry": {
    "coordinates": [
      6.443663779195363,
      51.20348336793408
    ],
    "type": "Point"
  },
  "type": "Feature",
  "properties": {
    "phenomenons": {
      "Speed": {
        "value": 34.647194623947144,
        "unit": "km/h"
      },
      "Rpm": {
        "value": 1584.3050694465637,
        "unit": "u/min"
      }
    }
  }
}

```

```

},
"GPS Accuracy": {
  "value": 2.999999910593033,
  "unit": "%"
},
"MAF": {
  "value": 9.437765815629945,
  "unit": "l/s"
},
"Engine Load": {
  "value": 43.96216858346017,
  "unit": "%"
},
"GPS PDOP": {
  "value": 1.4999999776482582,
  "unit": "precision"
},
"O2 Lambda Voltage": {
  "value": 3.2762881521175586,
  "unit": "V"
},
"Throttle Position": {
  "value": 21,
  "unit": "%"
},
"Consumption": {
  "value": 3.102402130874109,
  "unit": "l/h"
},
"GPS VDOP": {
  "value": 1.1779116287827491,
  "unit": "precision"
},
"GPS Speed": {
  "value": 33.47147411240462,
  "unit": "km/h"
},
"GPS HDOP": {
  "value": 0.899999986588955,
  "unit": "precision"
},
"Intake Pressure": {
  "value": 44.14216932654381,
  "unit": "kPa"
},
"GPS Bearing": {
  "value": 304.7174273121018,
  "unit": "deg"
},
"Intake Temperature": {
  "value": 13.000000387430191,
  "unit": "c"
},
"CO2": {
  "value": 7.290645007554156,
  "unit": "kg/h"
},
"O2 Lambda Voltage ER": {
  "value": 0.9988191702782387,
  "unit": "ratio"
},
"GPS Altitude": {
  "value": 104.90115790988267,
  "unit": "m"
}
},
"id": "590ad752268d1b08a47f18d4",

```

```
    "time": "2017-03-27T04:51:05Z"
  }
}
```

### B. Questões a serem respondidas

Existe uma série de questões que me vem na cabeça nesse instante que, inclusive, já tomei notas.

- quais marcas/modelos consomem mais gasolina?
- quais marcas/modelos poluem mais o meio ambiente?
- qual o perfil de pilotagem de cada automóvel?
- muita gente usa esse dataset?
- de que lugares são 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 comuns na Europa são os mesmos do Brasil?
- podemos descobrir quais são as regiões mais poluídas e regiões menos poluídas?
- é possível utilizar este dataset para extrair padrões de comportamento de cada usuário?

### III. PREPARAÇÃO DOS DADOS PARA VISUALIZAÇÃO

O enviócar é uma base aberta de dados para quem quiser usar. Simples assim. Entretanto, eles não colocam seu banco de dados para que seja feito download de forma simples. Eles expõem um serviço REST que, de forma simplificada, oferece um acesso fácil aos dados.

Para que eu pudesse fazer a visualização de maneira geral, primeiramente eu fiz um script que baixava cada uma das viagens para minha máquina e as indexava no Elasticsearch. O script rodou por horas, porque eu fiz download de maneira sequencial, até para não acocar o servidor dos caras.

A base de dados até não é muito grande. São 15 mil viagens, o que não parece muito. Entretanto, o volume de dados é relativamente grande, pois para cada viagem, é enviado para o banco de dados uma série de informações a cada 2 segundos. Então, embora o número de viagens seja relativamente pequeno, o tamanho de cada viagem é grande. Para se ter uma ideia, minha base indexada no Elasticsearch ocupou 6 Gb de disco.

Além disso, para poder facilitar as visualizações, eu fiz uma série de métricas. Por exemplo, para comparar quais carros emitem mais gás carbônico, por exemplo, para cada viagem eu computei a média do gás carbônico emitido naquela viagem, pois a base de dados não oferece isso. 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. TÉCNICA DE VISUALIZAÇÃO DESENVOLVIDA

A exploração dos dados foi feita com o uso de duas ferramentas: Kibana e R.

Kibana é uma ferramenta utilizada para análise por inspeção manual e visualização de informações que funciona de maneira

natural com o Elasticsearch. Dessa forma, é ela que irá apresentar os dados armazenados no Elasticsearch, em uma interface, via browser, altamente customizável com histogramas, mapas e outros painéis que propiciam uma visão geral sobre os dados. O Kibana possibilita transformar os logs em informações (valor) através de Dashboards, pois permite realizar correlação de eventos, filtrar logs por origem, hospedeiros, entre outras combinações (VAARANDI; NIZINSKI, 2013).

Foi utilizado o Elasticsearch como base primária dos dados, pois além do suporte à persistência, vem junto com uma série de mecanismos e algoritmos de recuperação de informação. A ferramenta permite combinar geolocalização com full-text search, structured text, and analytics.

Outra ferramenta para visualização de informações utilizada foi a linguagem R. A linguagem também possui uma variada gama de técnicas para visualização de dados, além de possuir um suporte muito grande a questões de estatística e de probabilidade.

O R não é uma ferramenta de tão alto nível de abstração quanto o Kibana em que, uma vez que os dados estão persistidos, se montam visualizações com alguns cliques. Aqui, é necessário que se escreva código para que as visualizações apareçam. Os gráficos não aparecem de uma forma tão elegante quanto o Elasticsearch, mas possui muito mais flexibilidade, maleabilidade e extensibilidade que o Kibana. Além disso, existe um conjunto muito mais vasto de visualizações que o Kibana, como boxplots e matrizes de correlações.

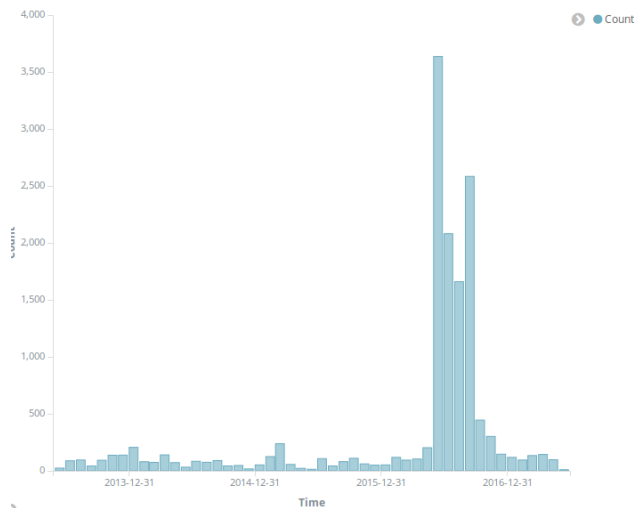
Dessa maneira, a combinação de ferramentas se fez necessária, para atingir requisitos um pouco diferentes. Essa combinação trará uma experiência legal para o usuário.

#### A. Histograma

Com o intuito de saber se muita gente utiliza a plataforma enviócar, foi desenvolvido um histograma em que o eixo x corresponde à variável temporal e o eixo y corresponde à quantidade de viagens naquele período de tempo. Com isso, poderá ser inferido se bastante gente vem utilizando a ferramenta e, principalmente, se seu uso vem crescendo ou decrescendo com o passar do tempo. É um estudo de tendências, portanto.

Pode-se observar o gráfico abaixo, que não existe uma tendência de aumento ou de diminuição ao longo do tempo. O que se observa é um 'boom' de uso no período em volta do ano de 2016, mas que não se mantém no restante do tempo. No período de novembro de 2016, por exemplo, tivemos 3600 viagens naquele mês. Mas no início de 2017, não passavam de 100 viagens por mês.

Não se pode observar, portanto, um uso massivo da plataforma. Entretanto, temos uma amostra que parece razoável para que se possa tentar responder as outras perguntas do artigo.



No eixo x, observamos a quantidade de uso da plataforma ao longo dos meses.

## B. Heatmaps

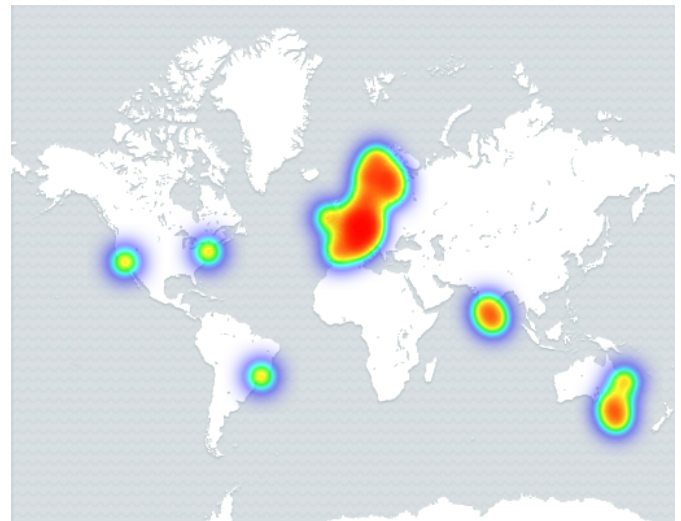
Uma vez visualizada a distribuicao temporal da utilizacao da plataforma, ira ser observada a distribuicao geografica dos usuarios. Como ja se sabe que eh uma plataforma desenvolvida na Alemanha, pode-se supor que muita gente na Europa ira utilizar a ferramenta, em relacao as outras regioes do globo.

A visualizacao por calor, utilizando como metrica a quantidade de viagens dispostas no globo terrestre, dara um senso de quais regioes sao mais utilizadas a ferramenta.

Observando o grafico a seguir, observamos aquilo que ja era suposto de antemao: de fato, na Europa, a plataforma eh mais utilizada. Alem disso, conseguimos obter alguns dados que nao eram imaginados. Vemos pontos isolados de uso nos Estados Unidos, Brasil, India e Australia.

Com as informacoes geoespaciais, ganha-se confianca de que poderao ser respondidas outras questoes, como o comportamento dos motoristas em diferentes regioes do globo. Alem disso, pode-se utilizar esse grafico para clusterizar de maneira macro o globo terrestre de acordo com essa base de dados: America do Norte, America do Sul, Europa, India e Australia.

A visualizacao de heatmap traz uma ideia muito clara sobre os pontos que mais utilizam a ferramenta. Traz-nos a ideia de que em diversos pontos do globo ela esta sendo utilizada, com enfase na Europa. Traz confianca de que poderemos observar padroes de comportamento em diversos pontos do globo. Ela traz uma ideia basica, mas nao definitiva, da quantidade de pessoas que utiliza a ferramenta ao longo 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

O heatmap deu uma ideia basica sobre a quantidade de pessoas que utilizam a plataforma ao redor do globo. Entretanto, nao trouxe uma ideia definitiva sobre as proporcoes de cada regioao.

Para isso, sera utilizado um grafico em pizza, que traz claramente a ideia de proporcoes ao usuario, ao dispor em fatias cada segmento analisado.

Nesse grafico, sera utilizado o cluster inferido visualmente do heatmap com as seguintes regioes: Europa, America do Norte, America do Sul, India e Australia.

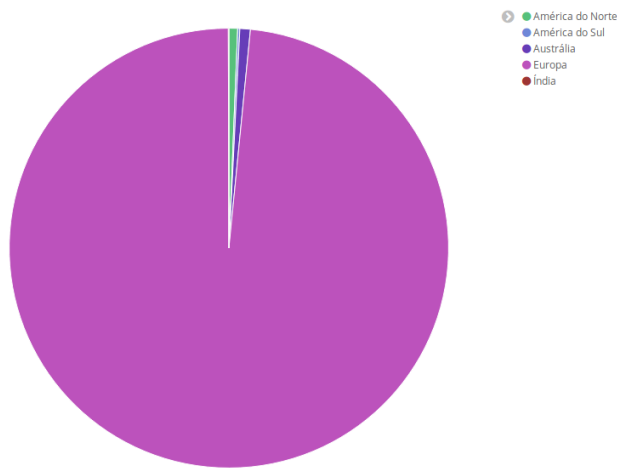
Aqui, pode-se observar a imensa maioria, de fato, utilizando na Europa: 97% dos usuarios esta la. Temos alguma representatividade, com 1% dos Usuarios na America do Norte e 1% na Australia.

Na America do Sul e na India, nao chega a 1% de usuarios. Parece ter sido realmente um usuario curioso que comecou a utilizar a plataforma em sua terra natal.

Essa informacao podera ser utilizada de uma maneira que achei incrivel: para comparacao de dados globais, como estatisticas e tentativas de detecao de comportamento em massa, devera ser dado foco nos usuarios da Europa.

Entretanto, essa falta de representatividade em outras partes do globo podera favorecer um olhar mais individualista. Por exemplo, pode-se acreditar que na America do Sul e na India, eh sempre a mesma pessoa que utiliza a ferramenta. Desse modo, poderemos observar algum comportamento individual. Do mesmo modo pode ser feito na America do Norte e na Australia, com um pouco mais de criterio, pois la houve muito mais uso.

Entretanto, na Europa, nao podemos fazer isso, pois eh muita gente usando. Na Europa, serao feitas analises mais genericas.



#### D. Graficos em pizza

Observadas algumas questoes temporais e espaciais, irao ser analisados alguns padroes de comportamento em diferentes regioes do globo. Sera utilizado aquele cluster inferido do primeiro heatmap para fazer uma analise sobre quais sao os fabricantes mais utilizados em diferentes partes do globo.

Sobre padroes de comportamento, espera-se verificar se supostos padroes culturais podem de uma determinada regio do globo pode ser inferida atraves deste banco de dados. Por exemplo, eh verdade que americanos gostam de carros espacosos? Eh verdade que na Europa e Estados Unidos os carros sao muito melhores do que os da America do Sul?

Para responder essa pergunta, algumas visualizacoes foram feitas algumas suposicoes de tecnicas de visualizacao: sunburst e nugget.

A tecnica de sunburst nao pareceu muito adequada devido a quantidade de informacoes – muito grande para um espaco pequeno – e o biscoito nao transpareceu de maneira clara o que estava sendo perguntado.

Para isso, foi utilizada novamente a tecnica de visualizacao de graficos em pizza. Entretanto, agora, em vez de plotar somente um grafico, o foram plotados 5 graficos em pizza, em que cada pizza corresponde a uma regio do cluster. Essa visualizacao foi escolhida em detrimento do sunburst pois a segmentacao trouxe mais evidencia para respoder aquilo que estava sendo perguntado. O sunburst trouxe muita informacao em muito pouco espaco fisico, o que pareceu meio confuso.

Observando os fabricantes, vemos que as principais marcas na America do Norte, Australia e Europa sao as mesmas.

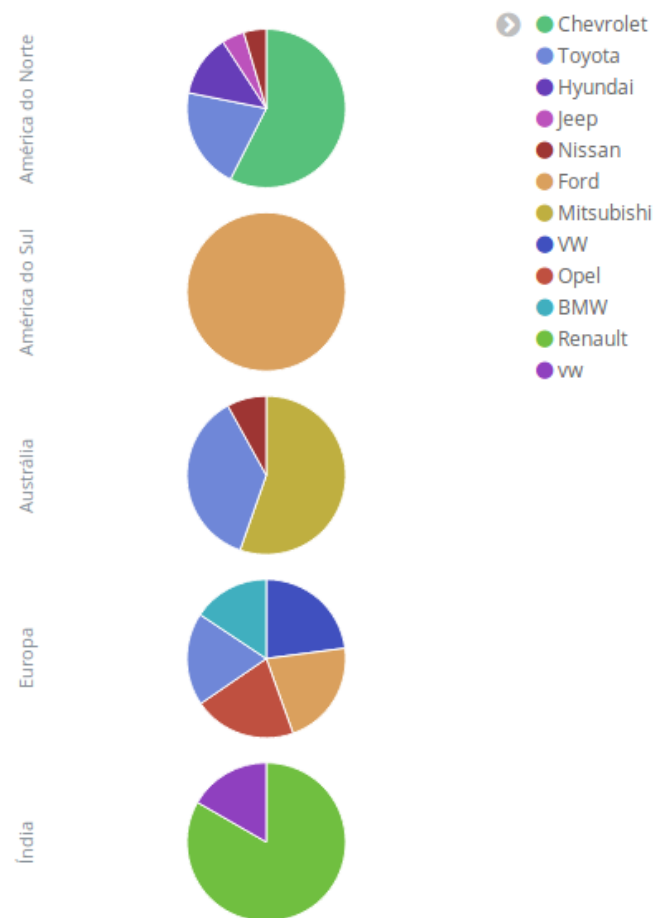
America do Norte: Chevrolet, Toyota e Hiunday Australia: Ford, Toyota e Opel Europa: Toyota, Ford, Opel, VW, BMW

Ja na America do Sul e India, parecem ser regioes mais humildes tambem, pois aqui nao observamos Hiunday e BMW.

America do Sul: Ford India: Renault e VW

Baseado nessa amostra de dados, nao se pode chegar a uma conclusao definiva para as questoes levantadas. A distribuicao entre Europa, Estados Unidos e Australia ficaram razoavelmente parecidas. Ja na America do Sul e India, realmente

parecem regioes menos desenvolvidas. Entretanto, a amostra dos dados nao parece ser significativa o suficiente para uma conclusao assertiva. A observacao do grafico parece corroborar a ideia inicial, utilizada para fazer as pergutnas, mas nao ha evidencias que, de fato, elas acontecem.



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

#### E. Sunburst

A Sunburst visualization is a radial space-filling visualization technique for displaying tree like structures. There are other space-filling visualization methods that use other visual encodings for describing hierarchies.

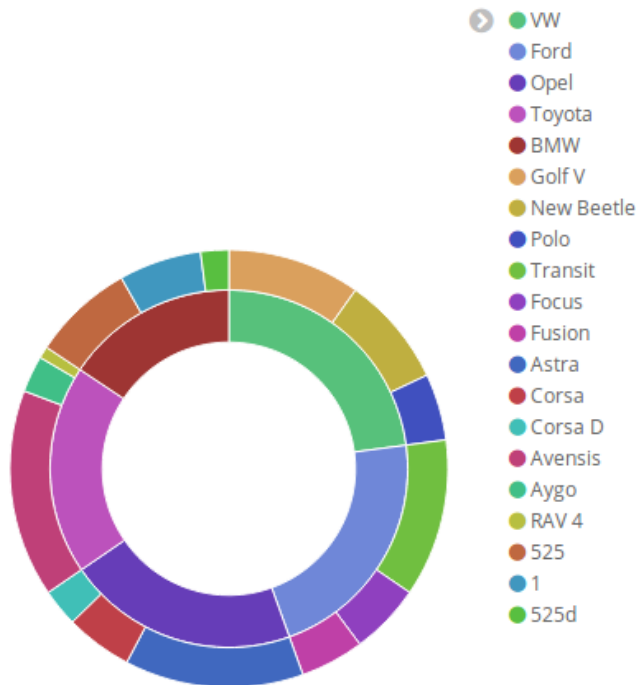
Essa tecnica mostra-se muito adequada para condensar informacoes hierarquicas, como para responder a pergunta: para cada fabricante, qual eh o modelo de automovel mais utilizado.

Como na Europa existe uma amostra muito mais significativa do que em outras regioes do globo, a pergunta fica limitada a regio da Europa. Alem disso, para que a visualizacao nao acontecesse de maneira muito confusa, para que ela ficasse clara, foram escolhidos os 5 fabricantes mais utilizados e os 5 modelos mais utilizados por cada fabricante. Se nao houvesse essa limitacao, o grafico ficaria muito poluido, o que gera um desconforto ao usuario.

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 sunburst. 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.

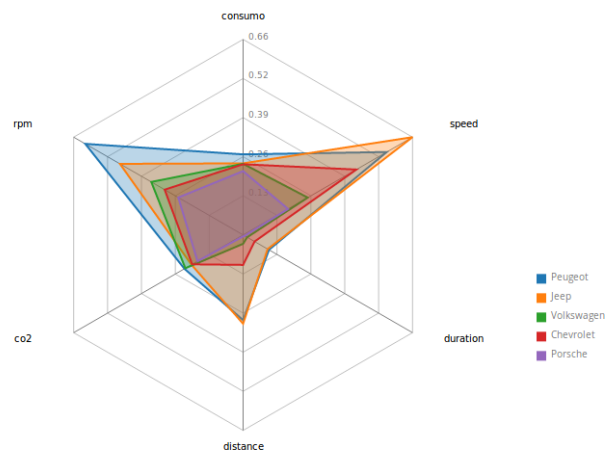
Optei, entao, por fazer uma media dos principais atributos que pude encontrar: co2, rpm, consumo, speed, duration e distance. Entao, calculei a media de cada um desses atributos selecionado por fabricante.

Achei o resultado bastante curioso: Nele se pode observar 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 Porsche 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 Porsche deve ter uma representatividade muito baixa na amostra. Mas o fato eh que os caras de porsche 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 surpreendeu 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.



