Visualització de la informació

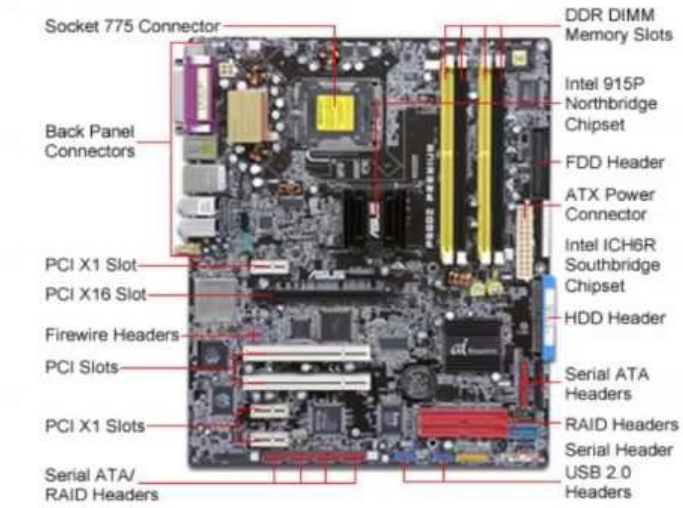
Cars Dataset

Mariano Trebino

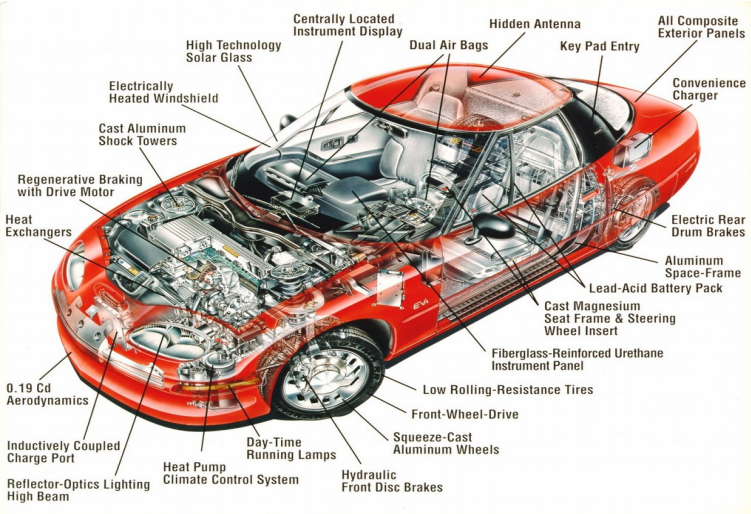
Índex de continguts

1 Motivació

Fa només una setmana vaig complir els 23 anys i ja fa cinc que hauria de tenir cotxe. De fet, no passa un dia sense que algú que no hem pregunti per què encara no tinc cotxe. La veritat és que la majoria de vegades no ho necessito ja que considero que la bicicleta és suficient per moure's per Girona. No obstant, com és lògic, hi ha una altra raó: els diners. Quan he de comprar un ordinador sóc capaç de trobar el millor ja que entenc els diferents components que el formen, com interactuen entre ells i quines característiques han de tenir (Il·lustració 1.1: Parts d'una placa base d'un ordinadors). Però, quan es tracta de cotxes no tinc ni idea. Per aquesta raó, m'agradaria estudiar un conjunt de dades de cotxes per conèixer i entendre com estan relacionades les diferents variables entre elles i, en concret, amb el preu del vehicle (Il·lustració 1.2: Parts d'un cotxe).



Il lustració 1.1: Parts d'una placa base d'un ordinadors



Il lustració 1.2: Parts d'un cotxe

2 Definició d'objectius

L'objectiu d'aquest projecte és crear una visualització que ens permeti estudiar i entendre les dades d'una manera molt fàcil. És a dir, ens ajudarà a:

- **Estudiar les inter-relacions.** Mitjançant les visualitzacions podrem veure com es relacionen les diferents variables entre elles, com per exemple que passa amb el consum quan la potència creix, com es relacionen les marques amb la potència, etc.
- **Estudiar el preu.** Concretament, podrem veure les relacions de totes les variables amb el preu, la variable amb un caràcter més important donada la motivació del projecte.

3 Descripció de les dades

El conjunt de dades que faré servir és força conegut. Es tracta d'un conjunt de dades de cotxes que podem trobar al repositori de *machine learning* de la universitat d'Irvine [1].

Aquest conjunt està format per 205 observacions amb 26 atributs i conté alguns valors. D'aquests 26 atributs, 16 són numèrics i 10 categòrics (veure Taula 3.1: Descripció de les dades).

Atribut	Tipus	Possibles valors
Symboling	*Numèric discret*	*[-3, 3]*
Normalized-losses	*Numèric continu*	*[65, 256]*
Make	*Categòric*	*alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercuri, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo*
Fuel-type	*Categòric*	*Diesel, gas*
Num-of-doors	*Categòric*	*Four, two*
Aspiration	*Categòric*	*Std, turbo*
Body-style	*Categòric*	*Hardtop, wagon, sedan, hatchback, convertible.*
Drive-wheels	*Categòric*	*4wd, fwd, rwd.*
Engine-location	*Categòric*	*Front, rear*
Engine-type	*Categòric*	*Dohc, dohcv, l, ohc, ohcf, ohcv, rotor*
Num-of-cylinders	*Categòric*	*Eight, five, four, six, three, twelve, two*
Fuel-system	*Categòric*	*1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi*
Wheel-base	*Numèric continu*	*[86.6, 120.9]*

Atribut	Tipus	Possibles valors
Length	*Numèric continu*	*[141.1, 208.1]*
Width	*Numèric continu*	*[60.3, 72.3]*
Height	*Numèric continu*	*[47.8, 59.8]*
Curb-weight	*Numèric continu*	*[1488, 4066]*
Engine-size	*Numèric continu*	*[61, 326]*
Bore	*Numèric continu*	*[2.54, 3.94]*
Stroke	*Numèric continu*	*[2.07, 4.17]*
Compression-ratio	*Numèric continu*	*[7, 23]*
Horsepower	*Numèric continu*	*[48, 288]*
Peak-rpm	*Numèric continu*	*[4150, 6600]*
City-mpg	*Numèric continu*	*[13, 49]*
Highway-mpg	*Numèric continu*	*[16, 54]*
Price	*Numèric continu*	*[5118, 45400]*

Taula 3.1: Descripció de les dades

4 Disseny

Com hem vist anteriorment, les meves dades tenen una quantitat elevada d'atributs i això fa difícil representar-ho correctament en la majoria de gràfics. L'únic gràfic que em permetia visualitzar moltes variables i veure com es relacionaven entre elles era el gràfic de **coordenades paral·leles** ja que em permet utilitzar múltiples dimensions.

No obstant, no és possible (ni recomanable) posar tota la informació en únic gràfic ja que provoca problemes de oclusió i cluttering. Per aquesta raó, he decidit incorporar gràfics menys importants que suportin d'alguna manera al gràfic de coordenades paral·leles, tal i com es fa en aquest exemple [2]. Tot i això, en certes ocasions pot haver-hi masses línies en el gràfic de coordenades paral·leles i, per aquest motiu, he cregut oportú incorporar eines de filtratge.

Així doncs, exposats els problemes anteriors, he desenvolupat el meu projecte regint-me per un conjunt de principis per aconseguir el millor resultat possible:

- **Gràfic principal.** La visualització girara al voltant d'un gràfic principal, en aquest cas, les coordenades paral·leles que atraurà la major part de l'atenció.
- **Gràfics auxiliars.** A més del gràfic principal, incorporaré diversos gràfics auxiliars que ajudin i suportin el gràfic principal afegint nova informació.
- **Visualització + Exploració**. A més a més dels gràfics, la visualització també incorporarà eines per que l'usuari poguí navegar i interactuar amb les dades per tal de filtrar-les i extreure informació d'una manera més fàcil.
- **Layout..** Òbviament, tot això haurà d'estar contingut en un layout atractiu visualment.

4.1 Gràfic principal

Com ja he avançat abans, el gràfic principal és el de coordenades paral·leles. Tota la visualització girara entorn a aquest gràfic ja que serà el més visual i el que transmetrà més informació.

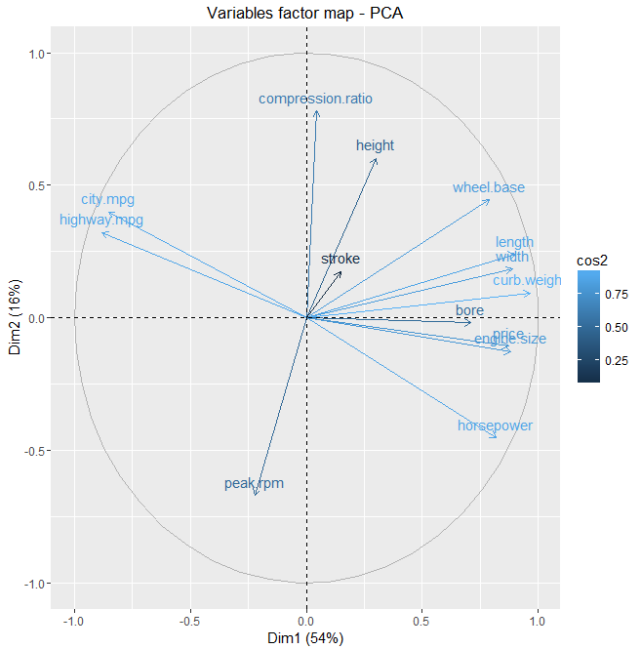
En aquest gràfic, hi ha quatre aspectes a tenir en compte:

1. **Variables.** Quines variables farem servir com coordenades paral·leles.
2. **Ordre**. Com estaran ordenades aquestes coordenades paral·leles. Això pot influir molt a l'hora de trobar patrons interessants en les dades.
3. **Escala.** Les diferents variables han d'estar escalades per ser representades al gràfic.
4. **Rotació.** Avaluar la possibilitat de rotar 180º algun eix.

4.1.1 Variables

Donat que prèviament he fet un anàlisi PCA i MCA per una altra assignatura, he pogut veure quines variables contínues (veure Il·lustració 4.1: Projecció de les variables continues sobre els dos primers eixos factorials utilitzant el color per mostrar la representació de cada variable en els eixos) i categòriques (veure Il·lustració 4.2: Coeficient de correlació quadrat entre les variables continues i categòriques i les diferents components principals) aporten major quantitat d'informació i són les que he escollit per representar en el gràfic.

Així doncs, les variables que he escollit són: *make, fuel.type,, city.mpg, price, curb.weight, compression.ratio, engine.size, length, horsepower, num-of-cylinders i width.* He decidit descartar *fuel.system* per què guarda molta similitud amb *fuel.type* al igual que *city.mpg* i *highway.mpg.*



Il lustració 4.1: Projecció de les variables continues sobre els dos primers eixos factorials utilitzant el color per mostrar la representació de cada variable en els eixos

$eta2

	Dim.1	Dim.2
make	0.68916923	0.4332268599
fuel.type	0.01271703	0.6117680597
aspiration	0.08545552	0.0375814388
num.of.doors	0.04029174	0.1664422051
body.style	0.12929465	0.1583343505
drive.wheels	0.45792081	0.0004800835
engine.location	0.01149021	0.0920221008
engine.type	0.25174035	0.2219856322
num.of.cylinders	0.48429972	0.1085485552
fuel.system	0.48675468	0.6887202196

Il lustració 4.2: Coeficient de correlació quadrat entre les variables continues i categòriques i les diferents components principals

Per ultim, atès que la variable *make* pren molts valors diferents, a més a més d'incloure com una coordenada paral·lela, he decidit utilitzar-la per codificar el color. Per tal d'escollir colors ben diferenciats he utilitzat colors vius generats a partir d'un algoritme [4].

4.1.2 Ordre

Per determinar la disposició i l'ordre de les variables al gràfic, he analitzat la Il·lustració 4.1: Projecció de les variables continues sobre els dos primers eixos factorials utilitzant el color per mostrar la representació de cada variable en els eixos i he posat de manera contigua les variables relacionades entre elles de forma molt clara, com per exemple *horsepower* i *city.mpg,* o *price* i *engine.size,* etc.

4.1.3 Escala

Per escalar els diferents valors de les variables de manera adient, ho faré amb una escala lineal per les variables continues i amb una escala ordinal per les categòriques.

4.1.4 Rotació

No he cregut necessari fer cap rotació dels eixos. Sembla que la informació ja es veu de manera molt clara.

4.2 Gràfics auxiliars

Per tal d'ajudar al gràfic principal, aportant nova informació sense dificultar la visualització de les dades, he incorporat tres gràfics auxiliars.

- **Donut selected data.** Aquest donut informa de les dades que es mostren actualment a les visualitzacions (per què han sigut filtrades) respecte del total, per saber en tot moment amb quines dades estàs treballant.
- **Donut makes breakdown.** Mostra, de les dades actuals que estàs fent servir, com estan distribuïdes respecte la variable *make.*
- **Data table.** Mostra la informació i el valor concret de cada variable per cada vehicle.

4.3 Visualització + Exploració

Per aconseguir lligar aquesta visualització i exploració incorporaré diverses eines interactives que permetin a l'usuari filtrar la informació. A més a més, aquestes eines estan sincronitzades en els diferents gràfics auxiliars, de manera que qualsevol modificació es transmet amb caràcter immediat a les altres vistes.

Les operacions que he incorporat als diferents gràfics són:

- **Brush and linking**. Permet aplicar filtres als gràfic de coordenades paral·leles i això provoca una actualització dels gràfics auxiliars, modificant el nombre de vehicles seleccionats, la partició per marques i les fileres de la taula.
- **Hover donut.** Permet d'una banda, obtenir més detalls de les dades actualment seleccionades a més de ressaltar aquestes en el gràfic principal.
- **Hover table.** Permet ressaltar la fila seleccionada al gràfic de coordenades paral·leles.
- **Sort table.** Permet ordenar la taula per la variable escollida de manera ascendent o descendent.

4.4 Layout

Per tal que tota la visualització estigui emmarcada dintre d'un layout i un disseny atractiu en general, he utilitzat una plantilla de bootstrap [3] i he intentat mantenir una certa coherència utilitzant els mateixos colors a les diferents parts de la pàgina, així com el tipus de lletra i les mides.

5 Implementació

Per implementar aquest projecte, he començat per les parts més fàcils primer, per adaptar-me i entendre el funcionament de D3 i Javascript abans de fer les parts més complicades.

Llavors he dividit el projecte en tres fases de desenvolupament:

1. Desenvolupament dels diferents gràfics sense interaccions.
2. Desenvolupament de les interaccions en els gràfics i les sincronitzacions.
3. Millores generals.

5.1 Fase 1

Durant aquesta primera fase del projecte m'he encarregat de dissenyar i implementar cada element per separat, sense cap mena de interacció amb els altres. De manera que primer de tot vaig implementar el més fàcil, els donuts, sense interacció ni res més. Només mostraven dades estàtiques. Després vaig fer la llegenda, les coordenades paral·leles i la taula de dades [4], de la mateixa manera, només mostrant les dades, sense poder interactuar amb elles.

En aquesta primera fase no vaig tenir gaires problemes ja que la majoria d'aquests gràfics són fàcils de fer i hi ha molts exemples per la xarxa.

L'únic problema que vaig tenir al començament del projecte va ser que D3 no funciona amb Chrome perquè hi ha un error amb el *cors-origin*, però es soluciona fent servir Firefox.

5.2 Fase 2

Un cop tenia fets i funcionant els diferents gràfics vaig implementar la interacció entre ells. Aquesta va ser la part complicada del projecte. L'idea era que al fer qualsevol mena de modificació en alguns dels gràfics, aquesta s'hauria de transmetre als altres gràfics de manera immediata.

Això em va suposar dos problemes molt inter-relacionats, d'una banda la gestió de les dades i l'altre la gestió dels events.

Per tal de solucionar la gestió de les dades i aconseguir una sincronització, vaig decidir, treballar amb un conjunt de dades original i un conjunt més per cada gràfic, de manera que cada gràfic tingués el seu conjunt propi únicament amb les dades que cal mostrar, però a la vegada sincronitzats, mostrant diferent informació del mateix objecte. No obstant, com era d'esperar, això em va provocar molts problemes d'integritat i coherència, ja que era complicat mantenir diferents conjunts amb diferent informació però dels

mateixos objectes.

D'altra banda, per gestionar els events que modificaven aquestes dades (eines d'exploració) vaig intentar-ho utilitzant una mena de MVC però no em va acabar d'agradar la gestió dels events (ja que es feien a la Vista en comptes de al Controlador), a més a més que afegia força complexitat a la simplicitat de la visualització.

Finalment, vaig fer una implementació molt simplista utilitzant una classe central que gestiona tant les dades com els callbacks dels gràfics i manté en tot moment un conjunt amb les dades actualitzades a través dels events. Llavors, cada gràfic sempre fa servir aquest mateix conjunt de dades i té la responsabilitat d'extreure la informació necessària per mostrar. D'aquesta manera vaig aconseguir simplificar molt el problema, treballant i actualitzant un únic conjunt, traslladant la responsabilitat de mostrar la informació adient a cada gràfic, i la sincronització a una classe central.

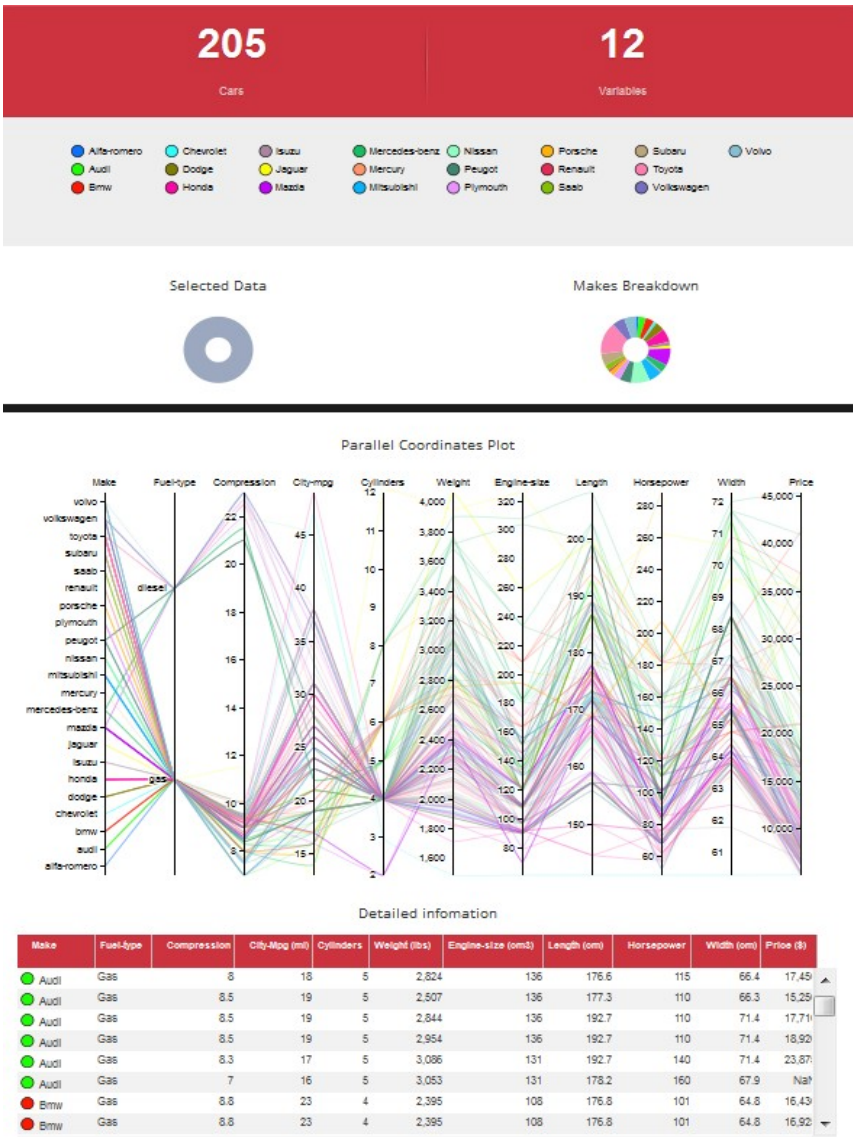
5.3 Fase 3

Per últim, vaig modificar estils, incloure un template, re-factoritzar el codi, per deixar-ho tot el més ben presentat possible i llest per entregar.

6 Resultats

En la meva opinió, els resultats (veure Il·lustració 6.1: Visualització resultant) són força gratificants per dues raons. La primera és el fet d'incloure un gràfic principal que atrau tota la atenció i que permet visualitzar la major part de la informació ajudant-se d'altres gràfics menys importants per explicar els detalls als quals ell no arriba. La segona raó i molt important és el fet de que no només incloc una visualització, si no que també hi ha eines d'exploració que l'usuari pot fer servir per filtrar les dades i crear un infinit nombre de visualitzacions, facilitant l'enteniment i l'estudi de les dades.

D'altra banda, també crec que els gràfics que he fet servir són molt adients per les dades que tracto i permeten entendre i extreure molta informació de manera molt fàcil. El fet d'haver analitzat personalment les dades mitjançant un PCA i un MCA m'ha ajudat molt a entendre les dades i trobar la manera més adient per visualitzar-les.



Il·lustració 6.1: Visualització resultant

De fet, fent servir aquesta visualització podem extreure diverses conclusions:

- La majoria de vehicles i marques prefereixen gasolina a dièsel.
- El vehicles dièsel tenen un rati de compressió dos vegades més gran que els gasolina.
- Els vehicles dièsel sembla que tinguin un millor rendiment, encara que no està molt clar.
- La major part dels vehicles utilitzen motors amb quatre cilindres.
- Els vehicles més pesats, llargs i amples tenen un motor més gran amb més cilindres i per tant més potencia.
- A mida que augmenta el preu, augmenta la mida del motor així com les dimensions (amplada, llargada i pes) del vehicle.
- Donada la premissa anterior, les marques més cares com Bmw, Mercedes-benz, Porsche i Jaguar (marques europees) són també les que tenen cotxes amb més potencia i solen utilitzar més de quatre cilindres.
- La majoria de marques, especialment les nord-americanes i asiàtiques són les més barates.
- I moltes més

Així doncs, veiem la gran utilitat d'aquests gràfics i la quantitat d'informació que en podem extreure de manera molt rapida i fàcil.

7 Conclusions

Personalment, haig de dir que mai m'ha agradat Javascript i en aquest projecte no ha sigut menys. Reconec que és un llenguatge molt utilitzat i potent però a mi no m'agrada. No m'agrada la programació web ni m'agrada com funciona el llenguatge ni com està estructurat.

Pel que fa a D3, haig d'admetre que és molt potent i permet fer visualitzacions molt maques i útils. A més a més, hi ha moltíssima documentació. Personalment, jo no creia que tanta gent ho fes servir.

No obstant, per fer algunes coses que hauria de ser simple es complica molt (com per exemple mostrar un *tooltip* al fer *hover* sobre un gràfic). També m'he sorprès que no incorporés cap eina per crear taules de dades i he hagut de fer servir una llibreria externa (*SlickGrid*) [5,6]. I aquest és un altre problema. Al final del projecte, has de fer servir moltes llibreries diferents que tenen conflictes entre elles i, al final, no saps ni què estàs fent servir. Per fer qualsevol cosa has d'afegir una nova llibreria amb dues dependències.

Per altra banda, encara que la part de programació sí que pot ser feina d'un enginyer, jo ho passo molt malament durant el disseny i la distribució dels gràfics, escollint colors, etc. Crec que aquesta és la part que menys m'agrada, en la que més temps inverteixo i que pitjor em queda.

Pel que fa al treball futur, com he comentat, la visualització és molt completa. Pot ser podríem afegir més variables al plot, o intentar evitar la superposició de les línies en algunes zones de molt trànsit per millorar la visibilitat.

Crec que el punt de millora més important és el disseny general, la distribució dels gràfics, els colors, espais buits, etc. Així com intentar fer un disseny responsiu, encara que és força complicat utilitzant un gràfic de coordenades paral·leles ja que s'allarguen horitzontalment.

8 Temporalització

En un principi, havia dissenyat un esquema temporal que no he seguit ja que he tingut altres treballs que fer i tot s'ha descontrolat força. Llavors, com he comentat abans, he dividit el projecte en una fase de disseny i les fases d'implementació descrites anteriorment:

1. El **disseny** l'he portat a terme principalment a l'inici del projecte per trobar els millors grafies i durant tot el projecte per anar fent les modificacions que feien falta.
2. El **desenvolupament** dels diferents gràfics sense interaccions. Aquesta fase no m'ha portat molt de temps ja que ha sigut ràpid d'implementar, aproximadament una setmana.
3. El **desenvolupament** de les interaccions en els gràfics i les sincronitzacions. Aquesta ha sigut la part complicada probablement entre dos i tres setmanes ja que m'he trobat amb varis problemes.
4. **Millores generals.** Arreglar el codi, les visualitzacions, el layout m'ha portat entre tres i quatre dies aproximadament.

9 Bibliografia

[1] UCI Machine Learning Repository, Automobile Data Set, Disponible en http://archive.ics.uci.edu/ml/datasets/Automobile

[2] Nutrient Contents – Parallel Coordinates, Disponible en http://exposedata.com/parallel/

[3] Domainer – Bootstrap template for people selling domains, Disponible en http://www.gettemplate.com/info/domainer/

[4] Quite: Select different colors for categories when drawing graphs, Disponible en http://jnnnnn.blogspot.com.es/2015/10/selecting-different-colours-for.html

[5] GitHub, mleibman/SlickGrid repository, Disponible en https://github.com/mleibman/SlickGrid

[6] Pildoras JS, SlickGrid DataView (I) – Introducción y uso, Disponible en http://pildorasjs.blogspot.com.es/2015/06/slickgrid-dataview-introduccion.html