

UNIVERSIDADE DE AVEIRO

DEPARTAMENTO DE ELECTRÓNICA, TELECOMUNICAÇÕES E
INFORMÁTICA

MESTRADO EM ENGENHARIA INFORMÁTICA

Trabalho Prático nº2 : D3



Autores:

David FERREIRA, Nº 72603

Andreia MACHADO, Nº 76501

VISUALIZAÇÃO DE INFORMAÇÃO

8 de Janeiro de 2018

Conteúdo

Introdução	4
Descrição do dataset considerado	5
Processos de limpeza de dados	6
Descrição da Aplicação	7
World Map Sightings Representation	7
Desenvolvimento Prático Associado	7
Sightings Per Year	11
Desenvolvimento Prático Associado	11
Problemas durante o desenvolvimento	14
Visualização de World Map Sightings Representation	14
Gráfico de Barras	15
Avaliação heurística	17
Heurísticas de Nielsen	17
Heurísticas de Forsell	18
Heurísticas de Zuk	19
Conclusão	20

Introdução

O trabalho desenvolvido foi para o âmbito da unidade curricular de Visualização de Informação.

Este projeto baseou-se no desenvolvimento de duas visualizações distintas em *D3* [1] que, com base em informação contida num *dataset*, escolhido pelo grupo, permitam aos utilizadores obterem um *insight* sobre os dados apresentados, exigindo-se o mínimo esforço cognitivo possível.

Como tema do trabalho foi escolhido um *dataset* de *UFO Sightings* [2] que resulta de um conjunto de registos de avistamentos de objetos voadores não identificados (OVNIs) ao longo do último século.

Neste relatório será descrito o desenvolvimento do trabalho, os problemas com os quais o grupo se defrontou, o modo de utilização e a informação possível de extrair das visualizações e ainda as decisões tomadas em grupo para o correto desenvolvimento das mesmas.

Descrição do *Dataset*

O *dataset* considerado pelo grupo para este trabalho resulta de um aglomerar de registos de avistamentos de OVNI's ao longo do ultimo século [2].

Este *dataset* contém informações relativas ao local do avistamento, ao formato do objeto, à duração da sua presença, à sua data de avistamento e ainda comentários gerais efetuados pelos sujeitos, Figura 1.

Associados a estes dados recolhidos encontram-se disponíveis para *download* dois *datasets*, um completo e um *scrubbed*. A diferença entre ambos é que o *scrubbed* apenas contém relatórios completos enquanto que, no *complete*, as coordenadas dos locais de avistamentos associados encontram-se geo-codificadas.

Embora fosse do interesse do grupo os relatórios estarem completos, foi optado pela utilização do *dataset complete* uma vez que era necessário que as coordenadas se encontrassem geo-codificadas para implementar a visualização desenvolvida.

datetime	city	state	country	shape	duration (seconds)	duration (hours/min)	comments	date posted	latitude	longitude
10/10/1949 20:30	san marcos	tx	us	cylinder	2700	45 minutes	This event took place in early fall around 1949-50. It occurred after a Boy Scout meeting in the Baptist Church. The Baptist Church ait	4/27/2004	29.8830556	-97.9411111
10/10/1949 21:00	lackland afb	tx		light	7200	1-2 hrs	1949 Lackland AFB, TX. Lights racing across the sky & making 90 degree turns on a dime.	12/16/2005	29.38421	-98.581082
10/10/1955 17:00	chester (uk/england)		gb	circle	20	20 seconds	Green/Orange circular disc over Chester, England	1/21/2008	53.2	-2.916667
10/10/1956 21:00	edna	tx	us	circle	20	1/2 hour	My older brother and twin sister were leaving the only Edna theater at about 9 PM,...we had our bikes and I took a different route home	1/17/2004	28.9783333	-96.6458333
10/10/1960 20:00	kaneohe	hi	us	light	900	15 minutes	AS a Marine 1st Lt. flying an FJ4B fighter/attack aircraft on a solo night exercise, I was at 50ꯠ' in a "clean" aircraft (no ordian	1/22/2004	21.480556	-157.8036111
10/10/1961 19:00	bristol	tn	us	sphere	300	5 minutes	My father is now 89 my brother 52 the girl with us now 51 myself 49 and the other fellow which worked with my father if he's still livi	4/27/2007	36.5950000	-82.1888889
10/10/1965 21:00	penarth (uk/wales)		gb	circle	180	about 3 mins	penarth uk circle 3mins stayed 30ft above me for 3 mins slowly moved of and then with the blink of the eye the speed was unreal	2/14/2006	51.434722	-3.18

Figura 1: Excerto do *dataset complete*

Processos de limpeza de dados

Tendo sido escolhido o *dataset complete* foi necessário o desenvolver de processos de limpeza de dados de forma a tornar o *dataset* consistente e livre de dados nulos, permitindo assim o desenvolver de visualizações consistentes que abordem da forma mais direta possível os dados reais e a sua correspondente representação visual.

Para isso foi utilizada uma biblioteca de *Python* [3] denominada de *Pandas* [4] que efetua a leitura de ficheiros *CSV* permitindo uma manipulação dos seus dados.

As ações de limpeza consideradas foram a eliminação de linhas cujo os valores associados fossem nulos ou caso os valores de latitude e longitude a serem utilizados fossem incoerentes.

Para além destes passos de limpeza foram também efetuadas alterações ao *CSV* pois havia 198 linhas que continham uma virgula em excesso o que fazia com que uma dimensão adicional não existente fosse considerada.

Descrição das Visualizações

Como foi já dito na introdução, o projeto desenvolvido baseou-se no construir de duas visualizações distintas em *D3* [1].

Para este projeto, foi desenvolvido um *site* para apresentar as visualizações desenvolvidas num dado contexto (introdução, descrição dos dados e das visualizações, apresentação das visualizações e conclusões obtidas). Para o site foi usado o *template Unika* [5].

Nesta secção serão abordadas as estratégias utilizadas para o desenvolver de cada uma das mesmas sendo dada uma introdução a cada e ainda efetuada uma análise à informação representada e ao *insight* possível de se obter sobre a sua visualização.

World Map Sightings Representation

Para esta visualização foi considerado um mapa mundo em que é demonstrado a localização dos avistamentos de OVNI's ao longo do último século.

O resultado da visualização pode ser verificado na Figura 2 e a descrição da sua construção será efetuada de seguida.

Desenvolvimento Prático Associado

Inicialmente foi criado e definido alguns parâmetros do *slider* [6] para que o utilizador possa utiliza-lo de forma a filtrar os dados a mostrar no mapa, sendo os valores do mínimo e máximo definidos após a leitura do *dataset*.

De seguida é efetuada a leitura do *dataset* para memória e filtra-se os dados necessários para desenvolver a visualização, que são o ano, a hora, a forma, a cidade, a latitude e a longitude. Esses dados filtrados são então armazenados num dicionário no qual a chave é o ano e o valor correspondente é uma lista de todos os avistamentos de OVNI's associados ao mesmo. Durante a

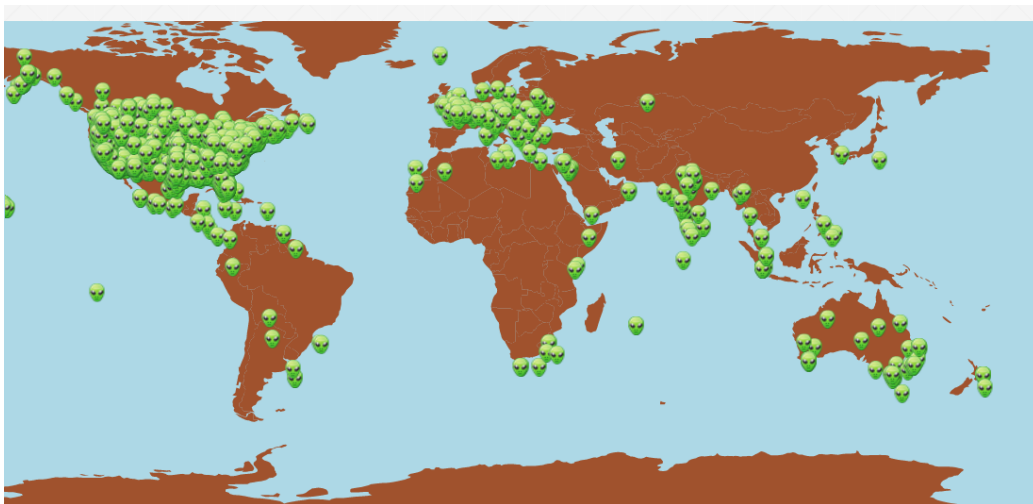


Figura 2: World Map Sightings Representation para o ano 2013

criação desta estrutura de dados são também criadas outras duas estruturas adicionais. Uma, é uma lista que guarda os dados que vão ser utilizados para criar os *markers* no mapa. Devido ao facto de o *dataset* possuir cerca de 88000 mil registos, optou-se por inicialmente apenas mostrar os avistamentos de OVNI's no ano de 2013, sendo o ano mais recente que se encontra no *dataset*. A outra estrutura resulta num dicionário no qual a chave é o ano e o valor associado uma lista que guarda o número de avistamento de OVNI's durante a manhã, tarde e noite desse respetivo ano.

Após o preenchimento de todas as estruturas é criada uma projeção de um mapa mundo [7] e com base num ficheiro *GeoJson* [8] é efetuado o desenho das regiões no mesmo. Assim que o mapa e as respetivas áreas são projetadas na visualização, por cada par *Latitude, Longitude* dos registos dos avistamentos correspondentes aos anos considerados é efetuado o *plot* de um marcador no mapa com base numa transformação em relação à própria projeção do mapa, de forma a mapear as coordenadas geo-codificadas no mapa a ser visualizado. De notar que inicialmente, sendo que o ano considerado é apenas o de 2013, apenas os marcadores respetivos a esse ano serão *plotted*.

Assim que os marcadores são representados são também definidos eventos de *MouseOver* em que sempre que o utilizador passa por cima de um, a sua *tooltip* associada é gerada indicando qual a data e a hora, a *shape* do OVNI e a *city* onde ocorreu o avistamento (Figura 3).



Figura 3: Exemplo de *Tooltip* no mapa mundo

O utilizador pode alterar/filtrar os dados, que são usados para mostrar no mapa a localização dos OVNIIs com um *marker*, usando dois tipos de filtros que são (Figura 4) :

1. um *slider*, para definir o *range* de anos
2. três *checkbox*, para definir se o utilizador pretende ver os avistamentos de OVNIIs que aconteceram de manhã, de tarde e/ou de noite



Figura 4: Filtros possíveis para filtrar os dados a apresentar no mapa

O utilizador ao usar qualquer um dos filtros referidos acima é definida uma nova lista que irá conter apenas os dados que satisfazem o pedido do utilizador. De seguida, são eliminados todos os *markers* que se encontram no mapa, e são criados novos segundo a lista criada. Após a criação dos *markers* no mapa é verificado se no *range* de anos definidos no *slider* existem registo de avistamentos de OVNIIs de manhã, de tarde e de noite desses anos. Isto serve para bloquear os filtros de *checkbox* quando não tem qualquer tipo



Figura 5: Impossibilidade de usar o filtro quando este não modifica a visualização

de influencia dos *markers* visualizados no mapa. Por exemplo, na figura 5 no range de anos 1906 - 1915, não existe avistamentos de OVNI's durante a manhã pois a *checkbox* correspondente está inacessível.

Para eliminar os filtros que entretanto o utilizador fez basta apenas clicar no botão *Eliminar Filtros*, no qual os valores de *default* serão novamente repostos.

Por fim, com esta visualização pode se concluir que usando todos os anos que se encontram no *dataset* (Figura 6), o número de avistamentos é muito mais acentuado na América do Norte e Central e na Europa Central.

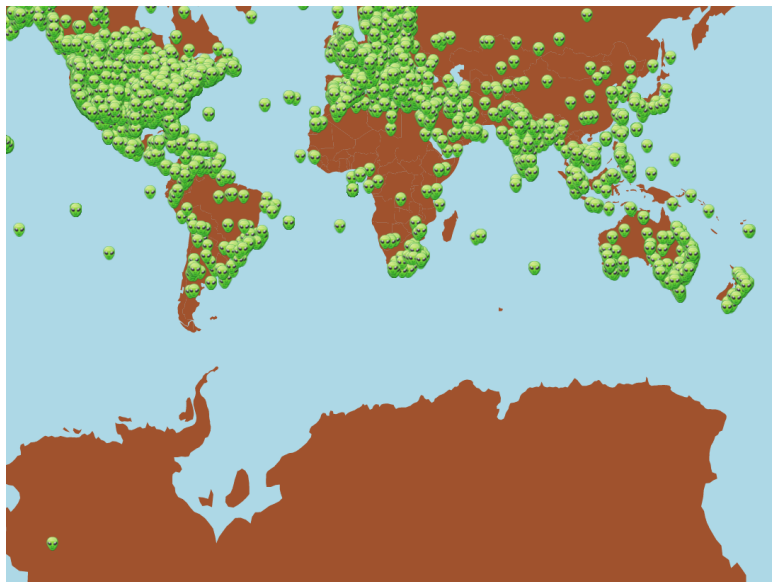


Figura 6: World Map Sighting Representation para todos os anos do *dataset*

Sightings Per Year

Para esta visualização foi considerado um gráfico de linhas em que é demonstrado o crescimento do número de registos de avistamentos de OVNI's ao longo do último século.

O resultado da visualização pode ser verificado na Figura 7 e a descrição da sua construção será efetuada de seguida.

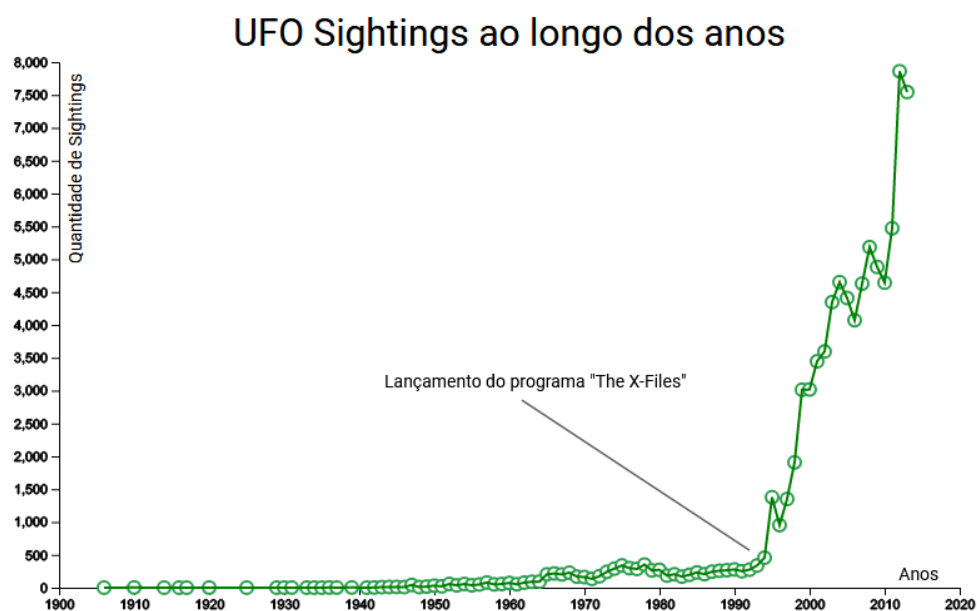


Figura 7: Sightings per Year

Desenvolvimento Prático Associado

Inicialmente foi efetuada a leitura do *dataset* para memória e filtra-se os dados necessários para o desenvolver da visualização, o ano e o número de *sightings* correspondentes. Esses dados filtrados são armazenados num dicionário de forma a ser possível separar o número de avistamentos por ano. Após preencher a estrutura de dados com a informação necessária do *dataset* é efetuada a inicialização dos eixos *X* e *Y*, sendo que a *X* o domínio é definido com base no *range* dos anos considerados no registo dos avistamentos e

Y é definido com base no número de *sightings* máximo em todos os anos considerados. Dito isto, o eixo X representa os anos e o Y representa o número de registos de avistamentos associados (Figura 8).

Após a inicialização e definição do domínio de cada eixo, é então passado à criação de um círculo que representa uma coordenada no plano 2D correspondente ao ano e ao número de registos de avistamentos a ele associados. De forma aos círculos ficarem contidos no plano de visualização foi necessário mapear os dados por eles representada segundo o domínio definido para os eixos considerados através de um método *scale()* que efetua de forma dinâmica o *resize* dos dados.

Com os círculos definidos e representados, foi utilizado uma *line* que de forma iterativa vai unir cada um deles de forma ao gráfico obter um aspeto similar ao de um gráfico de linhas.

De forma a oferecer uma melhor e mais detalhada visualização dos dados foi implementado um método que sempre que o rato passe por cima de cada círculo seja apresentado ao utilizador uma *tooltip* com informação relativa ao ano e numero de *sightings* associados a cada um dos mesmos (Figura 9).

Para além do desenvolvimento de todos estes elementos, foram também definidos 3 *labels* com carácter informativo para que o utilizador possa ter noção do que esperar ao visualizar o gráfico e que dados são representados por cada um dos eixos.

Por fim, foi também adicionado uma referência no mínimo curiosa relativamente ao crescimento do número de avistamentos de OVNI's, pois desde que a série televisiva *X-Files* [9] estreou (1993) o número de registos aumentou exponencialmente nos anos seguintes.

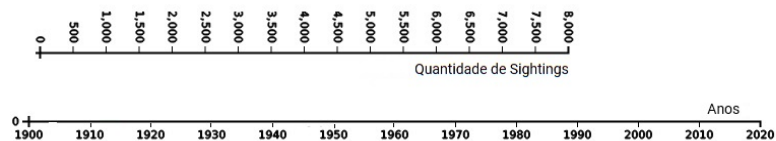


Figura 8: Eixos X e Y do gráfico que mostra o crescimento do número de registos de avistamentos de OVNI's ao longo dos anos



Figura 9: Exemplo de *Tooltip* no gráfico

Problemas durante o desenvolvimento

Durante a implementação das visualizações e do planeamento da estratégia de desenvolvimento das mesmas o grupo deparou-se com vários problemas, sendo esses aqui listados e discutidos de seguida.

Visualização de World Map Sightings Representation

Para o desenvolvimento desta visualização o grupo foi deparado com dois problemas para os quais infelizmente não conseguiu arranjar solução. Embora seja possível integrar uma API de mapas como por exemplo o *Google Maps API* [10] ou mesmo o *Leaflet* [11] tal não foi utilizado sendo que o trabalho desenvolvido deveria de abordar ao máximo o *D3* [1] e não uma API externa, usando apenas *JavaScript*.

Como tal, foram utilizadas ferramentas suportadas pelo D3 como por exemplo, *projections* [7], para efetuar uma representação do mapa mundo e posterior desenho das áreas dos continentes segundo um file *TopoJson* [8].

O primeiro problema com o qual o grupo se deparou foi que o *plotting* dos *markers* no mapa é efetuado segundo uma *scale* e um *mapping*, e ao ter sido desenvolvido um método para efetuar zoom sobre os continentes do mapa, os marcadores não eram dinamicamente alterados ficando com a posição anterior. Tal pode ser verificado na Figura 10.

Embora fosse possível resolver este problema, seria necessário fazer *unplot* a todos os *markers* e fazer *plot* novamente a todos eles, o que iria levar a um sobrecarregamento do *browser* devido ao enorme número de *plots* a terem que ser efetuados de forma repetida.



Figura 10: Problema com método de *zooming* desenvolvido

Por fim, outro problema associado foi a sobreposição de marcadores quando o seu número é muito elevado, o que faz com que o *tooltip* desenvolvido para cada um, sempre que ocorre um *MouseOver* sobre ele seja apenas gerada para o marcador mais “a cima” da sobreposição, isto é, para o marcador que foi *plotted* em último lugar na sobreposição.

Gráfico de Barras

Inicialmente a ideia do grupo seria o desenvolver de uma visualização com base num gráfico de barras onde fosse possível verificar para cada ano o número de avistamentos correspondentes. Assim, cada barra representaria um ano e o seu valor segundo *Y*, isto é, a sua *height* representaria o número de ocorrências associadas. Parte do resultado pode ser observado nas Figuras 11 e 12.

Um dos problemas que fez o grupo desistir desta ideia inicial foi o número de anos considerados nos dados o que levaria à representação de um número bastante elevado de colunas. Tal poderia ser resolvido com a utilização de um simples *scroller* horizontal. No entanto, outro dos problemas associados foi que nos anos iniciais o número de avistamentos correspondente é muito baixo comparativamente com os anos seguintes, o que fez com que as funcionalida-

des de visualização que o grupo queria demonstrar tivessem que ser divididas pelas outras duas visualizações futuramente consideradas e desenvolvidas.



Figura 11: Resultado do Gráfico de Barras - 1

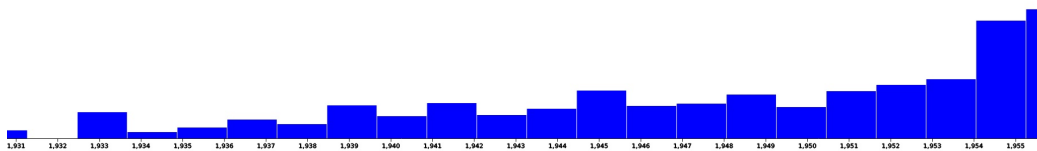


Figura 12: Resultado do Gráfico de Barras - 2

Avaliação heurística

Para este trabalho também foi feita uma avaliação heurística ao site desenvolvido usando as heurísticas de *Nielsen*, de *Forsell* e de *Zuk*. Nas secções seguintes encontram-se os resultados dessas mesmas avaliações.

Heurísticas de Nielsen

Heurística	Problema	Rating
Visibility of system status	Tal como é visível quando se faz <i>hover</i> nos círculos na segunda visualização, deveria de ser também visível quando se passa no <i>marker</i> da primeira	2
Recognition rather than recall	Em ambas as visualizações deveria de haver uma forma de clicar no marcador ou no círculo e fazer com que as <i>tooltip</i> se tornassem permanentes	2
Aesthetic and minimalist design	Embora necessário, os marcadores sobrepõem-se a si próprios criando problemas ao nível da geração de <i>tooltips</i> e a tornar o <i>loading</i> do mapa um processo demorado	2
Help and documentation	Embora a aplicação seja exploratória e de uso fácil poderia ser efetuado um tutorial do uso das visualizações para auxiliar ainda mais o utilizador	1

Tabela 1: Resultados da avaliação heurística da aplicação usando as heurísticas de *Nielsen*

Heurísticas de Forsell

Heurística	Problema	Rating
Orientation and help	Embora a aplicação seja exploratória e de uso fácil poderia ser efetuado um tutorial do uso das visualizações para auxiliar ainda mais o utilizador	1
Spatial organization	Embora necessário, os marcadores sobrepõem-se a si próprios criando problemas ao nível da geração de <i>tooltips</i> e a tornar o <i>loading</i> do mapa um processo demorado	2
	Há zonas no mapa que nunca têm marcadores, como tal, poderia ser considerado um mapa que excluísse zonas sem avistamentos	1
Recognition rather than recall	Em ambas as visualizações deveria de haver uma forma de clicar no marcador ou no círculo e fazer com que as <i>tooltip</i> se tornassem permanentes	2
Remove the extraneous	A curiosidade da segunda visualização poderia ser opcional, porque no fundo distrai o utilizador do seu objetivo que é ver o crescimento do número de avistamentos ao longo dos anos	2
	Há zonas no mapa que nunca têm marcadores, como tal, poderia ser considerado um mapa que excluísse zonas sem avistamentos	1

Tabela 2: Resultados da avaliação heurística da aplicação usando as heurísticas de *Forsell*

Heurísticas de Zuk

Heurística	Problema	Rating
Ensure visual variable has sufficient length	Na segunda visualização os primeiros anos do último século parecem que não tiveram qualquer tipo de ocorrência, algo que não é verdade	2
Put the most data in the least space	Há zonas no mapa que nunca têm marcadores, como tal, poderia ser considerado um mapa que excluísse zonas sem avistamentos	1
Remove the extraneous (ink)	A curiosidade da segunda visualização poderia ser opcional, porque no fundo distrai o utilizador do seu objetivo que é ver o crescimento do número de avistamentos ao longo dos anos	2
	Há zonas no mapa que nunca têm marcadores, como tal, poderia ser considerado um mapa que excluísse zonas sem avistamentos	1
Integrate text wherever relevant	A curiosidade da segunda visualização poderia ser opcional, porque no fundo distrai o utilizador do seu objetivo que é ver o crescimento do número de avistamentos ao longo dos anos	2

Tabela 3: Resultados da avaliação heurística da aplicação usando as heurísticas de *Zuk*

Conclusão

Com o desenvolvimento das visualizações foi possível aprofundar os conhecimentos lecionados na aula teórica da unidade curricular tendo-os aplicado na prática.

Para além de termos aprofundado o nosso conhecimento em representação e apresentação de informação, foi também possível ter uma ideia dos problemas subjacentes às mesmas sendo que foi assumido pelo grupo um espírito crítico em relação ao próprio trabalho desenvolvido.

Foi também possível de verificar o quão poderoso e útil pode ser uma boa visualização sobre dados de forma a diminuir em grande parte o esforço cognitivo que seria necessário sem a utilização da mesma.

No fundo, duas visualizações foram definidas em que não só é possível verificar o crescimento do número de registo de avistamento de OVNI's ao longo do último século como também os locais onde mais avistamentos foram registados.

Bibliografia

- [1] Mike Bostock. *D3.js - Data-Driven Documents*. 2011. URL: <https://d3js.org/> (acedido em 03/01/2018).
- [2] National UFO Reporting Center (NUFORC). *UFO Sightings — Kaggle*. 2016. URL: <https://www.kaggle.com/NUFORC/ufo-sightings> (acedido em 03/01/2018).
- [3] Guido van Rossum. *Welcome to Python.org*. 1995. URL: <https://www.python.org/> (acedido em 03/01/2018).
- [4] Phillip Cloud Et Al. Tom Augspurger Chris Bartak. *Python Data Analysis Library — pandas: Python Data Analysis Library*. 2011. URL: <http://pandas.pydata.org/> (acedido em 03/01/2018).
- [5] Veronika Poljakova & Imransdesign. *Unika - Responsive One Page HTML5 Template*. 2015. URL: <http://demo.imransdesign.com/unika/> (acedido em 03/01/2018).
- [6] Dave Methvin e Kris Borchers. *Slider — jQuery UI*. 2018. URL: <https://jqueryui.com/slider/> (acedido em 03/01/2018).
- [7] Mike Bostock. *d3/d3-geo-projection: Extended geographic projections for d3-geo*. 2017. URL: <https://github.com/d3/d3-geo-projection> (acedido em 03/01/2018).
- [8] Johan Sundström. *johan/world.geo.json: Annotated geo-json geometry files for the world*. 2016. URL: <https://github.com/johan/world.geo.json> (acedido em 03/01/2018).
- [9] *Arquivo X (TV Series 1993–) - IMDb*. 2017. URL: <http://www.imdb.com/title/tt0106179/> (acedido em 03/01/2018).
- [10] Google e Google Services. *Google Maps JavaScript API — Google Developers*. 2017. URL: <https://developers.google.com/maps/documentation/javascript/> (acedido em 04/01/2018).

- [11] Vladimir Agafonkin. *Leaflet - a JavaScript library for interactive maps*. 2017. URL: <http://leafletjs.com/> (acedido em 04/01/2018).