Daniel Saiz González (EIT-Health)

Giorgio Segalla (EIT-Digital)

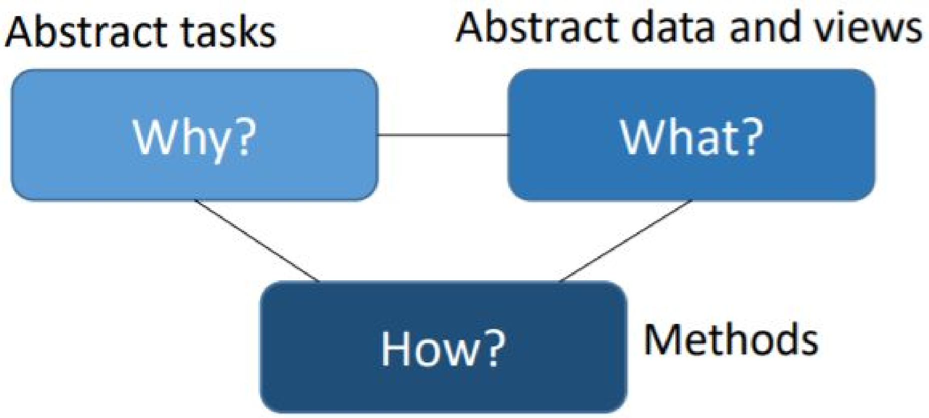
# Data Visualization practical work

## 1. Problem characterization in the application domain

The data used for this practical assignment comes from a dataset which provides detailed road safety data about the circumstances of personal injury road accident in Great Britain from 2005 to 2014. The data comes from recordings from the local authorities provided to the UK non-ministerial department called 'The National Archives', dependent from the national government. The license provided is called Open Government License and it allows the user to copy, adapt, publish and transmit the information.

## 2. Data and task abstraction

With the data provided we must answer three questions in order to create a visualization of the data. *What? Why?* and *How?*



- Data abstraction – *What?*

With this question we are trying to describe our dataset and its data types.

The dataset times is a table, with rows, corresponding to each accident; columns, to attributes of the accident and cells, defined by a pair (column, row) which contains a value of an attribute for an accident given.

The dataset is composed by 32 attributes with different types: the most common type is categorical, which applies to most of the attributes such as weather conditions, local authority and light conditions. Then, there are ordered attributes such as, longitude and latitude (quantitative), Time (sequential) and finally, day of week and month (cyclic).

In order to have a better comprehension of the location of the accidents and to cluster them, we transform using another dataset from the UK government in which shows the correspondence of each local authority code (county) with its name. In addition, the categorical attributes were represented by numbers and its value was changed into the name of the categories they represent. Regarding the dates, its format was changed into a one which can be analysed by R (%d/%m/%Y). Two additional attributes were created, in order to represent the month and year, separately. Now the date is represented by a full date and separated by day of the week, month and year to perform further analysis.

- Task abstractions – *Why?*

To answer this question, we must make a differentiation among the three visualizations of the data that are going to be produced and specify in each visualization what are the actions and the targets to get to answer the question "*why?*".

1) In which moment of the day/week are there a greater number of accidents?

Actions: we don't produce new information, so the actions refer to consume. Within this group, with the question proposed, we are looking for discovering (verification of hypothesis), to find new knowledge. The search performed will be a lookup, because we know what and where to look.

We are trying to find clusters and outliers and summarize temporally the data. We can visualize the time or day in which more accidents take place.

2) In which cities decreases the number of people injured or dead in accidents?

Actions: in this case we don't produce any new information either, so as we mentioned in the previous question, we want to find new knowledge based in the years.

We want to find trends, outliers, distribution and compare values.

3) In which counties bad weather affects more the traffic?

Actions: in this case we produce new information derived on what we already have. We created new information about bad weather gathering data from rainy, foggy and snowy days.

We want to find clusters, distribution and summarize the data.

## 3. Interaction and visual encoding – *How?*

- First question

As we want to represent two categorical keys (hour and day) in a two-dimensional array (matrix) and we want the result to be compact, to facilitate the comparison, we have chosen a heatmap to represent the number of accidents. Each cell represents a time and a day. To encode the data in each cell, we used a sequential colormap of the colour red. Lower numbers are coloured in white and highest in more intense red (Saturation). With this representation we don't want to interpolate the number of accidents for each cell, but visually represent the moments with high concentration and recognize clusters.

Regarding interaction with the plot, it is possible to filter by year and month and choose the severity of the accident or the county where it happened.

- Second question

In this case we want to represent the number of accidents with a fatal end (dead or injured) in the main cities in GB across the years. This could give an idea if cities are making efforts in decreasing this number and see a clear decrease across the years. Two continuos attributes are going to be represented, sum of fatal accidents (Y) and time (X). Since it is very difficult to represent every day or month of the 10 years of accidents recorded in the dataset, we want to group the accidents by years, and this can be done with a histogram. It is easy to see patterns since its visual channel, length, it is the more effective and expressive.

Regarding interaction, we can select the city among the most important ones, the severity of the accident(in this case we can select other severities apart from fatal, which is the one of interest) and the weather conditions, to check if it has something to do with the number of fatal accidents.

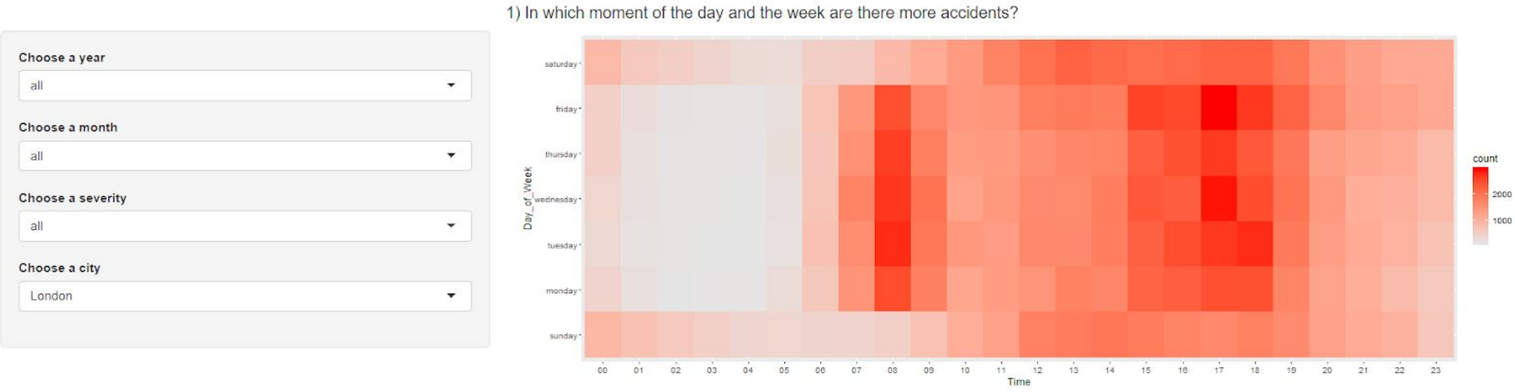
- Third question

In this case we want to encode the quantitative attributes in regions limited with two-dimensional marks. To create the map, we used a geoJSON file of the local authorities of the UK, which are commonly called counties. To give values to the regions in the map, we modified the geoJSON file, including the value of interest, in this case, percentage of accidents caused by bad weather (snow, rain and fog). As channel to represent the percentages, we used a sequential colormap of red, as the one used in the first question. When assigning the values to the regions, we realised that there are many counties with no data about accidents, we don't know whether this means that no accidents happened there or there is no data recorded, so we assigned a NA value to those regions and coloured in grey (so they don't interfere with the sequential red). The interaction with this plot resides mainly in manipulate choices, such as selecting the elements (and see the value assigned to it) and navigating to change the visualization.

## 4. Algorithmic implementation

The library to represent the histogram and the heatmap was *ggplot2* and to represent the map was *Leaflet*, which is an open source library for mapping applications. In order to give interaction, the library *shiny* was implemented. The representation is composed by two parts, the *ui* and the *server*. For the interaction with the graph, different ways are provided, such as selectors and radio buttons (for a single choice).

## 5. Results



Choose a year: all

Choose a month: all

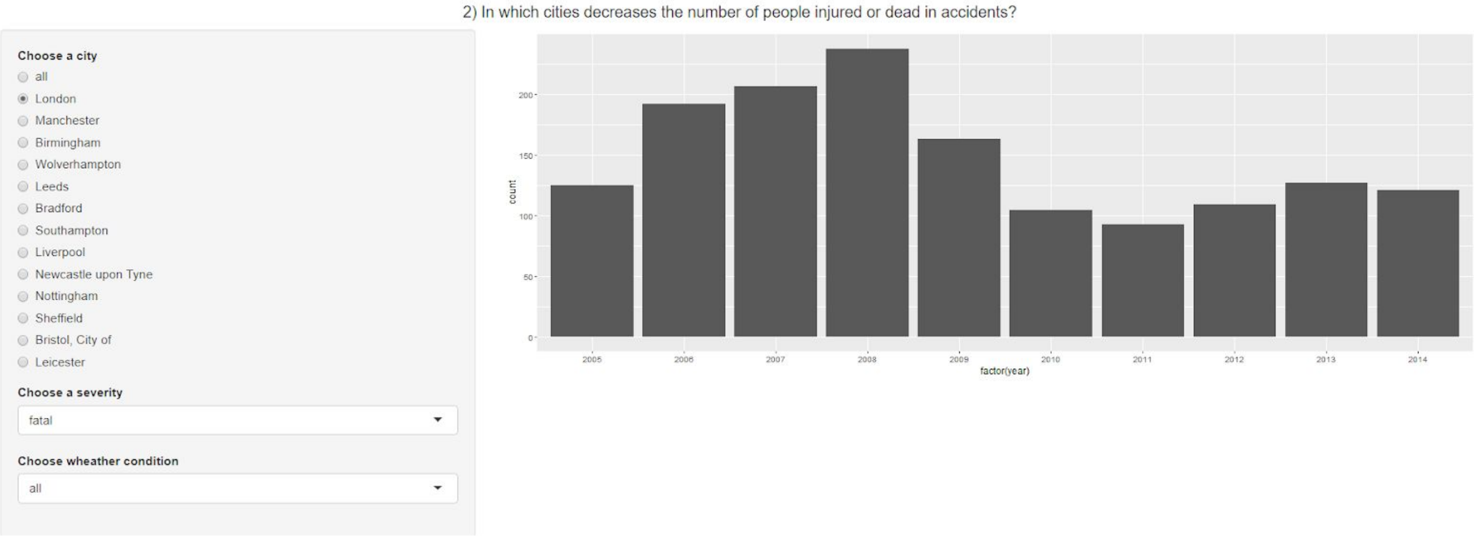
Choose a severity: all

Choose a city: London

1) In which moment of the day and the week are there more accidents?

| Idiom (1) | Heatmap |
|---|---|

| | |
|---|---|
| What: data | Table: two categorical key attributes (time, day of week), one quantitative value attribute (count of accidents) |
| How: encode | 2D matrix alignment of area marks, sequential colormap |
| Why: task | Find clusters, outliers, summarize |
| Scale | Items: 1.641.000<br>Categorical and quantitative attributes |



Choose a city

○ all<br>
● London<br>
○ Manchester<br>
○ Birmingham<br>
○ Wolverhampton<br>
○ Leeds<br>
○ Bradford<br>
○ Southampton<br>
○ Liverpool<br>
○ Newcastle upon Tyne<br>
○ Nottingham<br>
○ Sheffield<br>
○ Bristol, City of<br>
○ Leicester

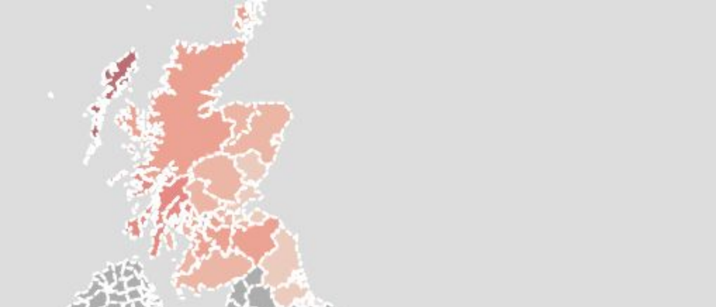
Choose a severity: fatal

Choose wheather condition: all

2) In which cities decreases the number of people injured or dead in accidents?

| Idiom (2) | Histogram |
|---|---|
| What: data | Table: two quantitative attributes, one as factor. |
| How: encode | Bar chart with length as channel |
| Why: task | Find trends, outliers, distribution and compare values |
| Scale | Items: 1.641.000<br>Categorical and quantitative attributes |

| Idiom (3) | Choropleth map |
|---|---|
| What: data | Geographic geometry data. Table with two quantitative attribute per region (only one is displayed) |
| How: encode | Space: use given geometry for area mark boundaries. Colour: sequential colormap |
| Why: task | Find clusters, distribution and summarize the data |
| Scale | Items: 1.641.000<br>Categorical and quantitative attributes |



Daniel Saiz González (EIT-Health)

Giorgio Segalla (EIT-Digital)

**How to run?**

First of all in the moodle is only consented to send a archive with less than 50mb, but our data are quite big i've sent also the info with an email(giorgio.segalla@alumnos.upm.es)

I let the dataset in my google drive, you can download it by the link:

https://drive.google.com/open?id=1OvenBA1e51VeysNhgnCv9INE9OtZA30A

Copy,past and uncompress in the principle folder.

How to execute?

use the function of shiny runApp from the directory

example

```
shiny::runApp("1question")
```

the operation can take some minutes the first time, for the operation of wrangling

We have tried to put the different questions in the shiny server, but the file are too big for the free version.