**TUTORIALS VISION ZERO INNOVATION LAB**

Here, a series of practical demos and tutorials have been compiled for the Vision Zero Innovation Lab on Wednesday April 27th in Leeds as part of the Leeds Digital Festival.

To access more information on the Vision Zero Lab, visit the [Github page](https://github.com/StudentDataLabs/VisionZeroInnovationLab) or check out [eventbrite](https://www.eventbrite.co.uk/e/vision-zero-innovation-lab-leeds-digital-festival-tickets-21587524868).

**Demo 1 – Asking Questions**

**Demo 2 – Collecting Data**

**Demo 3 – Exploratory Data Analysis - Excel**

**Demo 4 - Data Exploration, Analysis, Modeling and Visualisation - R**

**Demo 5 – Mapping - QGIS**

**DEMO 1 - Asking Questions**

This first demo will be focused on how to ask questions of data. Determining the problem is one of the most important parts of the data science process. It governs the rest of the project

This Lab is concerned with exploring ways to reduce the number of road deaths to zero in Leeds.

I’d like you to come up with two questions of your own. You could explore:

* A particular area or road
* A section of the population – such as gender or age
* Cycling or pedestrian casualties

I’d like you to think about who is your target audience, why you are undertaking the project and what the effect is intended to be.

Our main question for the Innovation Lab is how can we understand road casualties better in Leeds and what explore what we do to reduce them.

Explainer:

# This is an explainer to let you know what the code is doing

This is the code – you can cut and paste this into the R console

**DEMO 2 - Collecting Data**

The next stage is to see what data we have and what we can do to get more. If you go into our public Github files, you’ll find a series of datasets.

We’ll predominantly be working with open data from Leeds Data Mill, a pioneering portal that has achieved international recognition.

Now, we’ll be going online to explore what open data we can find. It takes a little sleuthing to find what you are after but there’s a wealth of useful information we can gather. Remember: “Data is the new oil”.

Process:

* Open the Leeds Data Mill site on your browser
* Search “diabetes”
* Explore the options
* Download the data

I’d like you to be critical of the open data. Be aware of how the data is collected and bear in mind that no dataset is perfect. Many will be quite messy and will need to be cleansed so that we can analyse it.

It is useful to know the context of the data and what each value and variable means. Thankfully, there is an option on Leeds Data Mill to download ‘Guidance’. This will help us understand the data better. Please spend some time looking at this as it will help later on.

**DEMO 3 - Exploratory Data Analysis - Excel**

I’m going to show you how to do some exploratory data analysis in Excel and R. We’ll learn about basic formatting, removing duplicates, using pivot tables and functions to explore our data, and some quick data visualizations. We will also use Excel to produce some summary statistics.

As with all the Demos, these are intended to be practical exercises that will show you how to tackle real world data problems and conduct projects of your own. However, they are by no means exhaustive. Please use these tutorials as stepping stones to continue your learning.

In this demo we’ll be 1 open dataset – accidents\_2009-2014.csv

First, I’m going to tell you some more about Excel. Excel is the go-to spreadsheet tool for many data analysts. It’s extremely popular, fairly powerful and easy to use. The Health and Social Care Information Centre (HSCIC) in Leeds, for example, uses Excel to analyse health data.

Here, we’ll perform some common data exploration and cleaning exercise: removing duplicates, using Excel’s pivot tables and creating some quick data visualisations.

Removing Duplicates

You’ll find that many datasets contain duplicate rows. In our dataset, where there has been more than one casualty, a number of rows have been created. If we were looking at the personal characteristics of the casualties (gender, age) then we would leave the duplicates. However, since we are interested in the Number of Casualties, it is necessary to remove these duplicates.

* First, we open our CSV file dataset in Excel
* Identify cells, columns, rows, tables; top ribons
* First row should contain the columns
* Select ‘Data’ from the top ribbon
* Highlight all our columns and select ‘Remove Duplicates’
* You will see a list of columns. Since the duplicates can be seen from the Reference Number column, leave this checked as ‘Column A’ and uncheck all others. Also, leave ‘Select All’ checked
* Select ‘Remove Duplicates’
* Save our file as .CSV

Pivot Tables

Pivot tables are the go-to point for exploratory data analysis. They can be used to explore the data and provide summary statistics.

* Insert > Pivot Table
* Select the columns we are interested in (i.e. all except the Reference Number, Easting and Northing columns)
* A new sheet will be created. From here, we can drag and drop the variables we are interested in.

Summary Statistics

We might also want to count the number of occurances of a piece of text, say how many 20-30 year olds there were in a selection

* To do this we go back to our original sheet and use the formula: =COUNTIF(range, “20-30”)
* We may also want to find some summary statistics. To find the average of a range of values in Excel, enter the function

=AVERAGE(range)

* You can also find the MEDIAN, MODE and many of other insights

**DEMO 4 - Exploration, Analysis, Modeling and Visualisation – R**

In this demo we’ll be using R to explore, analyse and visualize road casualty data. We’ll also cover some machine learning exercises. This practical guide will show you how to install packages, read in data and explore and visualise data. This code is open source and can additionally be found on Github. This means it can be applied to other data projects. As mentioned earlier, these tutorials are designed to get you up and running on practical data science projects but they are by no means exhaustive.

Background

R is an extremely popular open source statistical programming language. It’s used by most companies with a strong element of data science in what they do: Google, Facebook but also the New York Times, the Financial Times and many more. It is the most popular tool for data science, with great visualization, data manipulation and modeling capabilities.

R can seem quite intimidating at first – don’t worry! The best way to get used to R is to simply get stuck into new projects, get your hands dirty work on real world data sets and open source projects. It has a fantastic online community that you should definitely check out.

Download R Studio by visiting this site:

<https://www.rstudio.com/products/rstudio/download/>

The code and demos have been published online to RPubs. The links can be found below:

1. [Data Exploration and Cleaning](http://rpubs.com/StudentDataLabs/Vision-Zero-Innovation-Lab-1)
2. [Mapping Cycling Crashes in R](http://rpubs.com/lbuk/Vision-Zero-Innovation-Lab_Mapping-Cycling-Crashes-R)
3. [Statistical Analysis and Modeling](http://rpubs.com/VisionZeroLab/Statistical-Analysis-Modeling)

**DEMO 5 – Mapping in QGIS**

In this section, we’ll be using QGIS – a free and open source Geographic Information System (GIS) – to identify road casualty hotspots in Leeds.

QGIS is a powerful tool used by cartographers, data scientists, journalists and so on. The Heatmap plugin is one of the most powerful ways to perform dot density analysis and create a beautiful map.

# Download QGIS

To download QGIS for free, visit: <https://www.qgis.org/en/site/forusers/download.html>

# Add the .csv file

First, add file – accidents\_2009-2014\_duplicates\_mapping.csv

Add Layer > Delimited Text File

In the window the appears, make sure x = Longitude and y = Latitude and check the Spatial Index box.

Set the map projection to Google Mercator. There are many options to choose here but since Google Maps is a good reference point, it will do for this purpose.

# Create the Heatmap, customize design

What you’ll see is a map showing all our points. To create a heatmap:

Properties > Style > Heatmap

From here, you can choose the colour and set the density. Since we are interested in finding hotspots, it is generally better to lower the density.

To export the map, follow [this link](http://docs.qgis.org/2.0/en/docs/user_manual/print_composer/print_composer.html) to find out more about Print Composer