**ALT 2 – Data Analytics – Data is the new Oil**

**What is data analytics?**

Data analytics is the science of analysing raw data in order to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption.

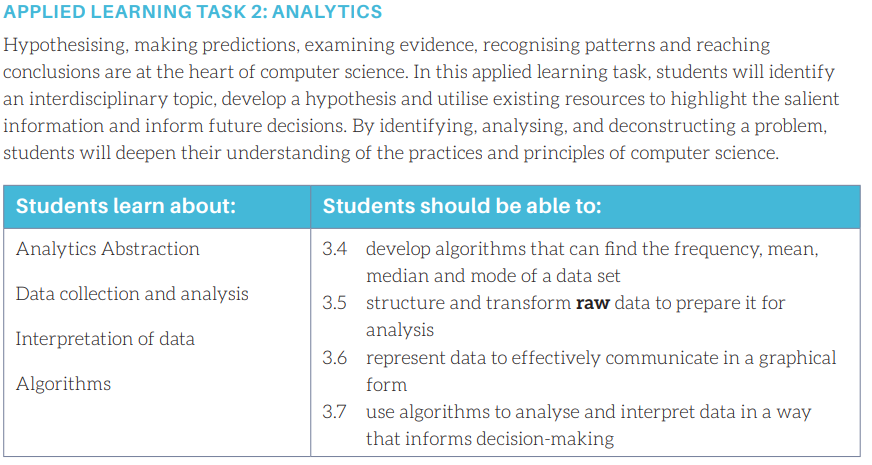
Data analytics techniques can reveal trends and metrics that would otherwise be lost in the mass of information. This information can then be used to optimize processes to increase the overall efficiency of a business or system.

For example – Tesco Clubcard holders have their data monitored every time they use the card. This data is analysed by algorithms to check for spending patterns, dips in spending amounts etc. This data can then be used to encourage the costumer to visit more often (with vouchers) or encourage the costumer to buy new products they normally wouldn’t.

**Your task:**

Below is an extract from the curriculum outlining the task. You essentially must:

1. Come up with a Hypothesis to investigate using a dataset (see example at end)
2. Use python coding language to read in a data set of your choice. You will need the “statistics” package and the “pandas” package to make this easier.
3. Use python to prepare or examine raw (not checked over for errors) data. You may need to edit this data to make it useable.
4. Analyse this data to find the:
   1. Mean (average)
   2. Frequency
   3. Mode (occurs most often)
   4. Median (middle)
5. Represent this information in graph form using python.
6. How would this analysis of your data inform decision making (Tesco Clubcard, Target pregnancy story)?



**Where to start:**

Firstly, you will need a dataset. You should decide on what type of data you would like to analyse. E.G – FIFA statistics, Earthquakes, Restaurant Rankings etc.

Then follow the task outline as above. Use your book to help guide you through the process. Remember CSV files are Comma Separated Values.

Use pages 72 and 73 to help with the planning of this topic. It runs through the stages of analytics:

1. Accessing data – getting the raw dataset.
2. Pre-Processing the data – any errors (Removing data P. 78)
3. Sorting the data – min and max, mode, mean, median, frequency (P.80-83)
4. Visualisation – (P. 84-88)
5. Analysis

**Structure:**

**What does your project do? And not do?**

Describe what the project does.

**Aims? / Any Limitations?**

What does the project set out to achieve (what are the limitations of your code/project)?

**Are there any ethical issues?**

Will your project cross any lines regarding ethics? Will it offend or can the data be used freely?

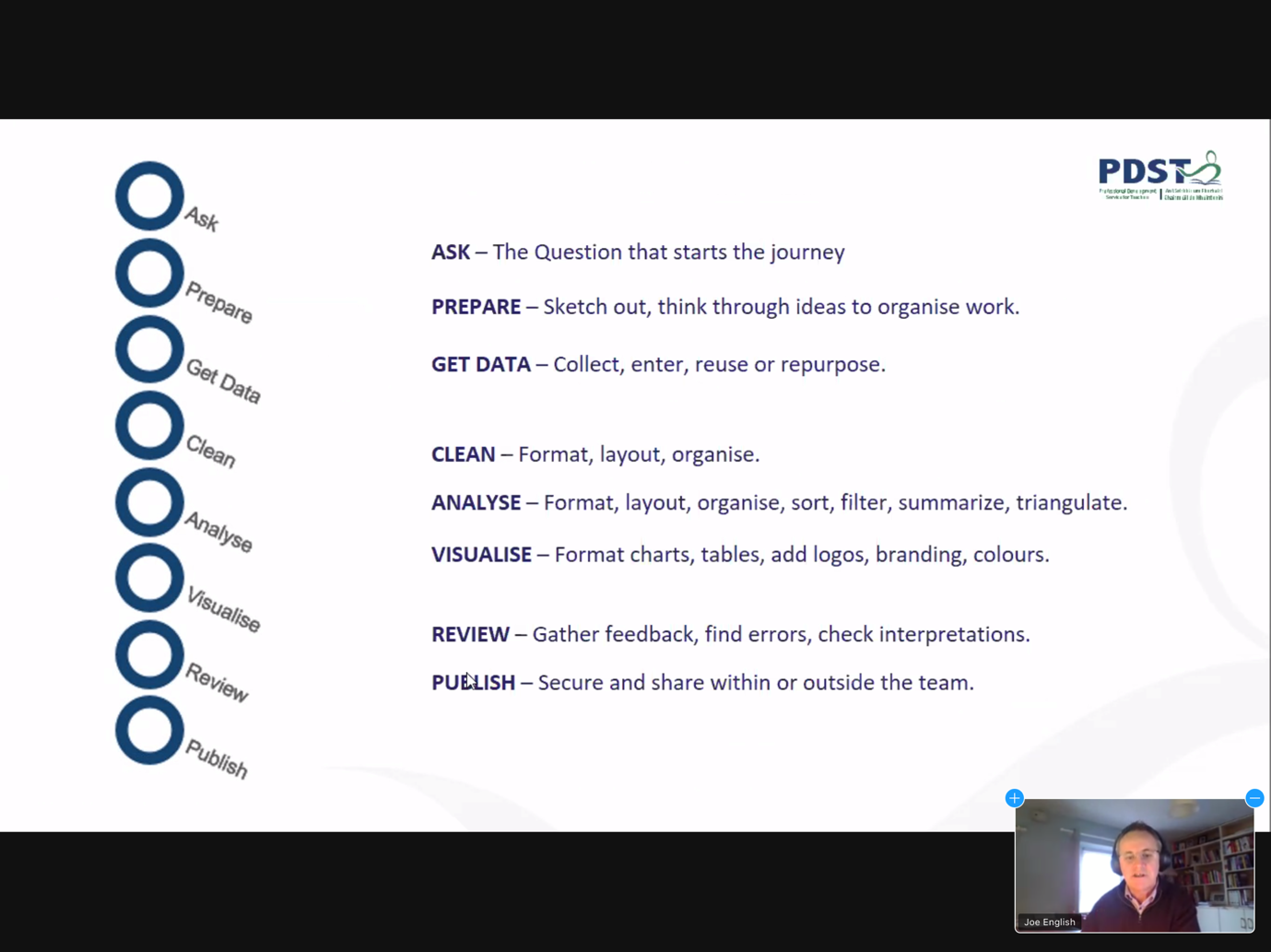
**Who are the end users?**

Who/what will/may use the result of your project?

**Tools / Materials required?**

What will you use to complete the task?

**What are the roles and responsibilities?**

Examples: programmer, researcher, graph plotter, record keeper, presenter, team leader, project manager, timekeeper.

**What you will/may need:**

Kaggle.com – dataset library

statistics library – Python

panda library – Python

matplotlib.pyplot

**Useful resources –**

[NCCA Curriculum Online ALT2 Support](https://www.curriculumonline.ie/Senior-cycle/Senior-Cycle-Subjects/Computer-Science/CS-Support-for-Teaching-and-Learning/Support-Material-for-Teaching-and-Learning/2-ALT-Resources/ALT2-Support/)

\*\*\*\*\*\* <https://www.curriculumonline.ie/getmedia/4f4f0cbe-ca22-4a16-a53f-026a4e68cf64/Generic-Example.zip> \*\*\*\*\*\*\*

**Very Brief Example (Getting Started):**

**Significant Earthquakes 1965-2016**

**Data Source:**

|  | **Hypothesis** |
| --- | --- |
| **1** | **There are more earthquakes in recent years.** |
| **2** | **There are more earthquakes in Winter than Summer in N. Hemisphere** |
| **3** | **There are more earthquakes in N. Hemisphere than Southern** |
| **4** | **The earthquake's intensity has increased.** |
| **5** | **More earthquakes occur before noon.** |
| **6** | **Most earthquakes do not register above 7.** |
| **7** | **There is no correlation between depth and magnitude** |
| **8** | **USA exploded more Nuclear bombs than Russia** |
| **9** | **China was the main cause of nuclear explosions in the 1990s** |
| **10** | The majority of earthquakes come from the top 3 locations |
| **11** | High magnitude earth quakes come in clusters of 2s and 3s |
| **12** | Earthquakes are stronger in norther hermisphere (over southern) |
| **13** | Majority of earthquakes happen around PRF |
| **14** | Do earthquakes happen in close proximity to rock bursts |
| **15** | Check are locations along fault lines |

**Chosen Hypothesis:**

??

***What does your project do? And not do?***

It examines the earthquake intensity between 2000 and 2015. It takes the average intensity of each year to graph any changes. It doesn’t compare the intensity with years outside the specified range.

***Aims? / Any Limitations?***

To investigate if the average intensity is truly increasing. With this dataset we cannot explore any potential causes of the increase.

***Are there any ethical issues?***

None, as the data is scientific, not personal data.

***Who are the end users?***

Researchers, academia, students/teachers of Geography

***Tools / Materials required?***

Python IDE (Thonny**)**

MS Excel to clean data manually

MS Excel/Plotly for graphs

Datasets from Kaggle