**U5 Case Studies on Data Investigations**

**Kazil, J & Jarmul, K. (2016) Data Wrangling with Python. O'Reilly. Media Inc.**

**Chapter 6**

***Case Studies: Example Data Investigations***

**Ebola Crisis**

Example flow of questioning and investigating if looking into Ebola in West Africa:

* Google Ebola, find many international organisations tracking and providing access to tools.
* Find WHO site with latest stats, interactive maps, KPIs all updated weekly. They provide data in CSV and JSON – reliable source verifiable and updated.
* Find Github repositories with raw data from many sources and a contract to speak to about data updating and collection formats. Data in formats that can be parsed.
* Decide on a question around burials being safe and find a report with a named person to try and speak to.

Results: Good sources, from verifiable organisations and someone to speak to about the research

**Train Safety**

Looking at negative factors affecting train safety in US:

* Look at previous research, find Federal Railroad Admin (DRA), reports say most due to poor track conditions or human error.
* Looking at human side, identify that sleep and issues relating to drugs and alcohol could contribute.
* Narrow questions down to looking at these factors, you have some initial datasets from research and contacts to ask for more.

**Football Salaries**

Looking at EPL salaries

* Find a series of website with player salaries and endorsements listed, contact the authors to find out more and if data is updated.
* Find a range of playing stats on he official EPL website, can obtain these by web scraping.
* So can start to compare the player salary data vs the playing stats data.

**Child Labour**

Focus on international crisis in child labour.

* Find UNICEF’s open data site dedicated to child labour.
* This has entire datasets on women’s wellbeing and child wellbeing / status
* Allows us to consider question: “Does early marriage affect child labour rates?”
* Government data from the US on rates across the world allows us to cross reference vs UNICEF data.
* Also find ILO trend report on child labour. This has links to many different historic dataset.

So obtained many datasets from reputable sources that can be compared and cross referenced and considered how to refine questions and find sources to answer them.

**Reflections**

Keys here are setting out to find sources of data, research and contacts to help you answer the questions. Questioning the reliability and data collection sources to understand if they are suitable for the analysis. Finding contacts and information about those data sources.