**ETL Project Proposal**

**Team Members:**

1. Brandon Tiley
2. Brett Scotland
3. Emily Dunkin
4. Waqas Zahick

**Project overview (200-300 words):**

* What is your question of interest?
  + - How does the crime rate vary throughout the UK, Canada and Australia?
* What is the expected outcome of your ETL project? (i.e. how will this data ETL project add value?) Think about how your ETL project will enrich the data and how this data might create value.
  + - The expected outcome of this project is to enrich people’s knowledge about the number of different types and amounts of crimes committed in 3 different continents / countries.
    - This will benefit people.
    - This could reveal strategies to reduce crime in other cities by investigating further the country who has the least amount of crime.

**EXTRACT - Proposed data sources (minimum two data sources):**

* Where is the data located?

On the following websites:

* + - <http://www.data.gov.au>
    - <http://www.ons/gov.uk>
    - <http://www150.statcan.gc.ca>
* What are the data set formats?
  + - Web scraping (HTML)
    - API (JSON)
    - CSV
* How will you get this data? (e.g. API, scraped data, download data)
  + - Web scraping
    - API request
    - Downloading the .csv file
* What are your assumptions about the data? What does it cover? What is it missing?
  + - We assume the data is representative of the general population and is complete.
    - CSV data doesn’t go into great detail about the type of crime’s that have been committed.
    - It covers the crimes committed over a certain timeline

**TRANSFORM - Proposed clean-up and analysis:**

* What are the transformations you will apply to the data? (e.g. filtering, aggregation, derived columns)
  + - Filter
    - Convert English data to 100,000 to match corresponding data sets (Normalise)
    - What steps will you take to clean the data and ensure its validity (e.g. messy data, duplicated data, incorrectly formatted data)
    - One data point for each month (no duplicate dates)
    - Remove text/number decoration
    - Group Australian data by month or, get total count for each month
    - Single offences -> count of offences
* How will you identify potential issues with your data sources? (e.g. exploratory data analysis, simple statistics etc)
  + - Outliers
    - Null values
* How will the data be integrated? (e.g. joins, merges)
  + - Join data sets into same table for comparisons.
* How will you apply these transformations (e.g. Jupyter notebook, Python script)
  + - Requests module to transform API (Request API to obtain web data).
    - HTML to transform scrapped data. (bs4 to pass and transform data).
    - Pandas to transform CSV.
* IMPORTANT → Why did you apply these transformations? How did this enrich your data?
  + - To integrate 3 different data sources.

**LOAD - Data storage**

* What type of database (relational, document) will you store the data?
  + - SQL Database
* Why did you choose this database over another database?
  + - Ease of access to analyse
* What are your expected tables / documents and relationships between tables / documents in your database?
  + - Table for each country, combined table with rows(dates), columns(countries).

**Potential limitations**

* What are the potential limitations of your above proposed steps?
  + - Data stored in different ways (Detailed vs Not detailed)
    - Difficult to determine what data needs to be pulled
    - Demographic difference between cities may influence the findings
    - 2015–2019-time range
* How can you control these potential issues?
  + - Normalise data (percentage as opposed to total number)
    - Find most uniform statistics to use (i.e. number of crimes)
    - Break data down into basic forms without excessive formatting to better understand

**ETL diagram:**

It may be easier to explain the process in an ETL diagram (you don’t have to create a diagram but it may help your group to formulate a plan).

Most data engineers make sure to document the ETL processes for reference. One way to do this is to use ETL diagrams. This is usually drafted at the beginning of the project and finalised at the end of the project.

Here are two examples of how an ETL diagram would look like:



