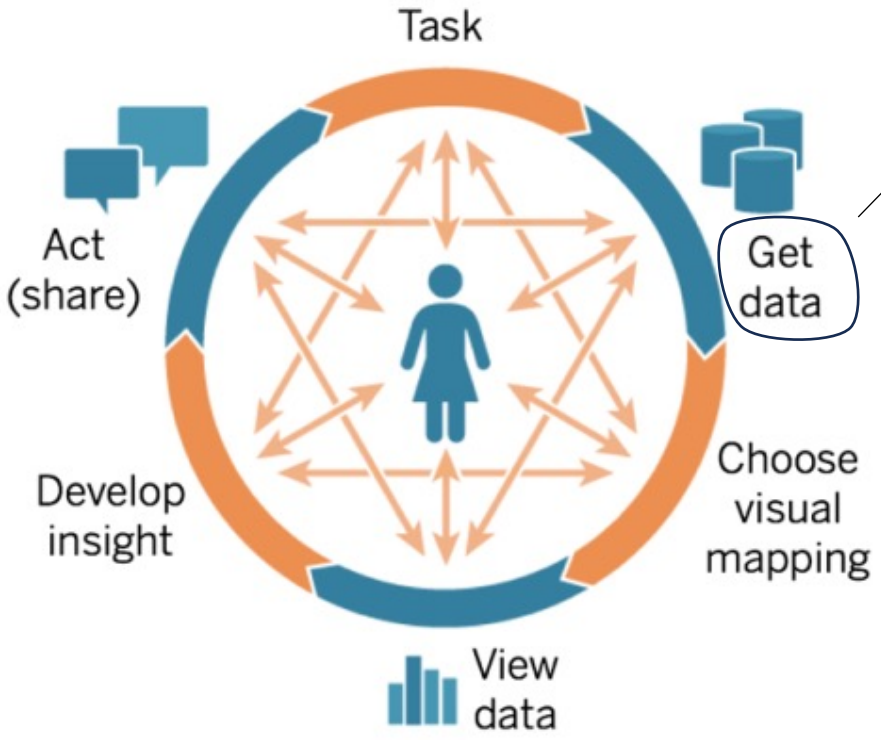
Lifecycle



Three sets of data was provided:

1. **Trees Data:** downloaded from the Council website (Excel)
2. **Environmental Data:** extracted from our council assets database (Csv)
3. **Common Names Data:** scraped from a horticultural website using coding (Json)

Identifier	0
Number Of Trees	115
Site Name	0
Contract Area	0
Scientific Name	0
Inspection Date	401
Inspection Due Date	401
Height In Metres	610
Spread In Metres	715
Diameter In Centimetres At Breast Height	712
Ward Code	226
Ward Name	226
Easting	56
Northing	56
Longitude	56
Latitude	56
Location	56

Identifier	0
Maturity	409
Physiological Condition	472
Tree Set To Be Removed	0
Removal Reason	23331
Capital Asset Value For Amenity Trees	710
Carbon Storage In Kilograms	2860
Gross Carbon Sequestration Per Year In Kilograms	2866
Pollution Removal Per Year In Grams	2860

Scientific Name	0
Common Name	24

In our analysis, we removed missing values (null and 0 values) from ‘Easting’ and ‘Northing’ columns in the Trees dataset. We executed this by creating a copy of the original dataset and applying the necessary functions to remove the values. This could have been directly applied to the original dataset, but if there was a mistake with this step then the original data would need to be loaded again.

The Trees and Common Names datasets are intended for public use hence the data is unrestricted, whereas the Environment data was extracted from the Council’s asset database which is internal, and so is sensitive information which is safeguarded.

Requirements

- The data requirements stated that the missing values and unmatching data would be required to complete the initiatives.

- The initiatives can still be carried out if this requirement was missed but it will not be respective of all the trees recorded.

Any limitations ?

- Not all missing values can be retrieved; a null value may have been recorded due to that data simply not existing. Same applies to unmatching data (e.g. a tree may not have a common name).

Classification ambiguity

- Inspection due date was ambiguous to classify as the column consisted of numeric data but had a datatype of string (object) which was confusing. After some research into variables, applying operations to dates does not make sense, therefore, it is qualitative data and not quantitative.

Quality

Evaluation data quality

- As missing values can indicate data quality issues, missing values for each dataset were identified. Percentages and total counts of null and zero values were returned respectively in the form of tables.

- Outliers of certain columns were also identified in the Trees dataset and shown in boxplots. We used these boxplots to make a judgement on whether the outliers were just extreme values or outliers.

- Having missing values and outliers affects the validity of the data in respect to the trees in Camden, therefore, it was important to identify and analyse these. Poor data quality can lead to inaccurate, unreliable, or misleading results.

- When identifying unmatched data, we found that there are 23 trees that were not in the environmental dataset and 76 trees not in the common names so these trees will not show up when the tables are joined. What we realised from this analysis is that when acquiring data from different sources, we need to make sure they have a common identifier so the datasets can be combined successfully.