Data Analysis Process

1. Read the project outline. Decided that SQL would facilitate the mining of all data needed. R could then be used to create effective visualisation.
2. Created an empty database in SQL Server Management Studio.
3. Imported CSV file into database (Only the ‘Data’ sheet of excel file).
4. Some queries were written to investigate data quality and look for errors.
5. Blank cells from excel sheet come through as NULL rows and needed to be deleted.
6. Used the coding information excel sheet to rename the icdsub4 categories, for easy interpretation.
7. Found that the total cancer category was not matching the excel data sheets. This was due to the D45-D47 category. The D45-D47 category values must also be included in the “C00-C96, D45-D47” category as a double-up.
8. When D45-D47 is excluded the data values match what they should as per the excel sheet.
9. D45-D47 category was deleted.
10. **Editing finished, now time to query.**
11. Various queries were created to mine the data needed from the database (Sub-Set and Tabulate). The results of these queries were used to tabulate data and import these datasets into R for visualisation. The requirements of the report will not require me to analyse the raw data further in R.

* Death rate/time
* Death rate/cancer type
* Registration rate/cancer type
* Variations between the sexes (registration and death)
* Total amounts (registrations and deaths)

1. Mined data were imported into R and saved as datasets.
2. Various plots were created using ggplot2 to allow easy visualisation of trends. This will likely be the most important information needed in the briefing.
3. **Report written.**