Project Update – 29/07/2021

# To do’s

* Finalise database schema, correcting issues and typos.
* Start organising data by the schema
* Complete preliminary analysis
* Create outline of dissertation document and start writing

# What you did and created?

* Completed database schema and implemented on the data
* Data Analysis:
  + Completed creating the questions and the tables and plots required for the answers
  + Reviewed the theming of plots and ensured a consistent style was used.
* Thesis:
  + Created an outline of thesis – reviewed already written sections and project proposal
  + Reviewed documents and papers on GDPR Compliance for Web Scraping and

# What decision have you made and how they were made?

* Decided on the graphs to use (e.g. geom\_smooth, geom\_density with ..count.. identifier calls)
* Decided on themes to use (default with enforced colouring to ensure consistency)

# Relevant References

|  |  |  |
| --- | --- | --- |
| **Name** | **Link** | **Notes** |
| Advanced R | http://adv-r.had.co.nz/Exceptions-Debugging.html#condition-handling | Used to develop error handling functionality |
| Stack Overflow: Advanced TryCatch statements | <https://stackoverflow.com/questions/12193779/how-to-write-trycatch-in-r> | Used to develop complex error handling functionality |
| GtR-1-API-v3.3 | <https://gtr.ukri.org/resources/GtR-1-API-v3.3.pdf> | Used to develop API access expressions |
| Progress Package Readme | https://github.com/r-lib/progress#readme | Used vinettes/tutorial to develop progress bar functionality |
| Rvest | https://rvest.tidyverse.org/ | Used to develop scaping from HTML documents functionality |
| JSONLITE | <https://cran.r-project.org/web/packages/jsonlite/vignettes/json-aaquickstart.html> | Used to request and convert JSON files into dataframes |
| Subset list based on a filter | https://stackoverflow.com/questions/6941506/subset-elements-in-a-list-based-on-a-logical-condition | Used to filter the elements of a list based on a condition and explain the use of the Filter function |
| rvest Error in open.connection(x, “rb”) | https://stackoverflow.com/questions/33295686/rvest-error-in-open-connectionx-rb-timeout-was-reached | Used to solve error with html document retreval and develop more stable method. |
| [How to convert variable (object) name into String](https://stackoverflow.com/questions/14577412/how-to-convert-variable-object-name-into-string) | https://stackoverflow.com/questions/14577412/how-to-convert-variable-object-name-into-string | Used in logging functionality to return the variable name parsed into a function |
| Regex101 | https://regex101.com/ | Used to develop regex expressions |
| DrawSQL | <https://drawsql.app/> | Used to create database schema |
| Nas Introduced by Coercion | <https://statisticsglobe.com/warning-message-nas-introduced-by-coercion-in-r> | Used to solve an issue when coercing data to the correct data type where lots of warnings would be thrown. |
| Remove periods/dots in entire data frame | https://stackoverflow.com/questions/55026965/remove-periods-dots-in-entire-data-frame | Used to develop method to remove thousand seperators from values in dataframes |
| Change column order | <https://dplyr.tidyverse.org/reference/relocate.html> | Used to have a more elegant method of reordering columns than using a select |
| Remove all special characters from a string in R | https://intellipaat.com/community/15034/remove-all-special-characters-from-a-string-in-r | Used to clean first names from person dataframe |
| genderizeR | https://github.com/kalimu/genderizeR | Package used to access the genderize API easily and process names - detailed citation available on the page |
| R for Data Science - Data Import - Parsering a vector | https://r4ds.had.co.nz/data-import.html#parsing-a-vector | Used to solve an encoding issues with names |
| How to plot multiple data series in GGPLOT | https://www.sixhat.net/how-to-plot-multpile-data-series-with-ggplot.html | Used to plot two data series to the same graph and have a legend show |
| Relational Data | https://r4ds.had.co.nz/relational-data.html | Used to plan out dataset schema and idenify appropriate primary and foreign keys |
| GGPlot Legend Title, Position and Labels | <https://www.datanovia.com/en/blog/ggplot-legend-title-position-and-labels/> | Used to customise plot legends |
| *How to rank within groups in R* | *https://stackoverflow.com/questions/31859175/how-to-rank-within-groups-in-r/43666202#43666202* | *Used to add a rank column that had the rank of cases within groups* |
| *How to import multiple .csv files at once* | [*https://stackoverflow.com/questions/11433432/how-to-import-multiple-csv-files-at-once*](https://stackoverflow.com/questions/11433432/how-to-import-multiple-csv-files-at-once) | *Used to load a folder of CSV files into a named list of dataframes.* |
| *Change density plot to count and scaled axis* | [*https://stackoverflow.com/a/51386116*](https://stackoverflow.com/a/51386116) | *Used to change a denisty plot to a smoothed count plot* |
| *How to deal with zero in log plot* | *https://stackoverflow.com/questions/40219639/how-to-deal-with-zero-in-log-plot* | *Used to handle log transformations on a plot when the dataset includes zeros that need plotting* |
| *GDPR Compliance for Web Scraping* | *https://www.zyte.com/blog/web-scraping-gdpr-compliance-guide/* | *Used to review and establish GDPR compliance of project and web scraping* |
| *GDPR and Research – An Overview for Researchers* | *https://www.ukri.org/wp-content/uploads/2020/10/UKRI-020920-GDPR-FAQs.pdf* | *Used to review and establish GDPR compliance of project and web scraping* |
| [*Remove space between plotted data and the axes*](https://stackoverflow.com/questions/22945651/remove-space-between-plotted-data-and-the-axes) | [*https://stackoverflow.com/questions/22945651/remove-space-between-plotted-data-and-the-axes*](https://stackoverflow.com/questions/22945651/remove-space-between-plotted-data-and-the-axes) | *Used to removed gaps between axis and plotted data - cleaning up axis* |
| [*Editing legend (text) labels in ggplot*](https://stackoverflow.com/questions/23635662/editing-legend-text-labels-in-ggplot) | *https://stackoverflow.com/questions/23635662/editing-legend-text-labels-in-ggplot* | *Used to edit legend labels for plots* |

# To do items for coming weeks

* Finish disso