Cleaning/data collection

To advise UoW on how best to increase their global ranking, we consulted three data sources of independent rankings. All data was for 2020 as it is the most recent full year which maximises the probability of our data being complete and relevant. The three sources of data we used were "Times Higher Education World University Rankings" (THE), "Center for World University Rankings" (CWUR), and "QS Top Universities" (QS). Two of these sources were available for direct CSV downloads while the last was copied from a web table to a text file with tab as a delimiter (Times Higher Education, n.d.; CWUR, n.d.; QS, n.d.).

To counteract bad characters and mismatched titles in the datasets such as foreign letters (e.g., É), or alternate names (e.g., University of Waikato vs Waikato University), we found a clean list of complete university names that could be downloaded as a csv and imported into Power BI as UTF-8 (endSly, 2015). After importing the three source files as tables in Power BI, some preliminary cleaning such as removing abbreviated names off the ends of university names which would interfere with fuzzy matching, and replacing value ranges such as 200-300 with flat numbers, was done using before-delimiter extraction. The tables were then merged by university names using fuzzy matching with a threshold of 0.7, starting with a right-outer merge of the biggest data set (CWUR) with the university names list. From here the other two tables were left-outer merged by university name with the same fuzzy threshold.

There were some occasional mismatches with the matching such as Tokyo and Toyo, and a lot of data was not detected across the three datasets resulting in a much smaller pool of data. However, the data we had left was more complete and as a result would give more accurate insights. From here, final cleaning such as column renaming, replacing zeros with nulls for incomplete data, and removing arbitrary columns and white space using find and replace, was carried out.

Data Transformation

To construct our dashboard a range of new columns and measures were created, both to create new insight, and to assist with countering limitations of built-in visuals, these can be seen in **figure 1**. Some notable transformations were creating the "average rank" and "rank of average rank" columns which averaged the rank across our three sources for each university, then was ranked again to create the latter column. Some transformations for ease of display included: creating a duplicate table with some pivoted columns to assist in comparing across scored categorical groups, creating default card measures when filters were not applied, and creating some adjusted measures to allow for viewing on the same scale.

The only user-created data used in our dashboards was regression outputs from excel which were imported into Power BI then related to the table by the university column. This allowed the display of regression values (e.g., coefficients) not natively supported by Power BI.

Figure 1.

Name	Type	Description
Average Rank	Column	Takes average of rank from three data sources
Rank of Average Rank	Column	Rank by average rank column to create new rank incorporating all three data sources
Master 2	Table	New table using some pivoted columns for freedom of analysis, related by university name for filtering
Attribute	Pivoted Column	Converts teaching, research, citations, industry income, and international outlook columns names to an attribute column for easier group comparison
Value	Pivoted Column	Converts teaching, research, citations, industry incomes, and international outlook column values to a values column for easier group comparison
Global Average	Measure	Takes average for all universities for value column which can be filtered by attribute
Waikato Values	Measure	Filters value column where university is Waikato University which can be filtered by attribute to avoid being impacted by chart filters
Default Average Rank	Measure	Checks to see if card is filtered by university, otherwise set to Waikato University average rank
Default CWUR Rank	Measure	Checks to see if card is filtered by university, otherwise set to Waikato University CWUR rank
Default THE Rank	Measure	Checks to see if card is filtered by university, otherwise set to Waikato University THE rank
Default QS Rank	Measure	Checks to see if card is filtered by university, otherwise set to Waikato University QS rank
Default Attribute	Measure	Checks to see if card is filtered by university, otherwise set to Waikato University's highest scoring attribute
Default Uni Name	Measure	Checks to see if card is filtered by university, otherwise set to Waikato University
Research Adjusted	Measure	Adjusts research column by -70 constant to match scale of top 20 ranks for ease of relationship viewing
Teaching Adjusted	Measure	Adjusts teaching column by -70 constant to match scale of top 20 ranks for ease of relationship viewing
Regression	Table	Contains single variable linear regression values for attributes against rank of average rank sourced from excel

Categories	Column	Contains list of categories related to attribute column
Coefficients	Column	Contains coefficients of each single variable regression
Intercept	Column	Contains intercepts of each single variable regression
R^2	Column	Contains r^2 of each single variable regression
Default Location	Measure	Checks if location is filtered by location, otherwise set to global