City Analytics: Data confidence index – scoring framework

1. Data collection method	2. Data processing methodology	3. Accuracy	4. Representativeness	5. Privacy	6. Frequency	7. Data verification	Overall score
How does the data get collected?	How does the data get stored, processed, cleansed and transformed?	How well does the data represent the real-world domain?	How much of a population/sample size does the data represent?	How well does the data deidentify or mask personally identifiable attributes of people/ subjects recorded in the data?	How often does new data become available?	Can the data be verified using other sources?	The value that will be used as the index to rank datasets used by Economic Development
1-5 score	1-5 score	1-5 score	1-5 score	1-5 score	1-5 score	1-5 score	Average of the 7 sections
1. Data is manually captured and entered	1. Data is handled manually by people to fix or transform source data. Process is undocumented and not transparent to end users/analysts	1. Data does not accurately represent the real-world domain	1. Data represents 0-5% of population	1. Dataset contains personally identifiable attributes, or is collected in breach of privacy rules	1. Data is updated annually or less	1. Data does not correlate with other sources or no form of correlation/ verification has been attempted	
2. Data is sourced in a semi-automated way	2. Data is handled manually by people to fix/ transform source. Process is semi documented, but tacit knowledge is undocumented and untransparent	2. Data inconsistently represents real world domain, and largely consists of null values or erroneous values	2. Data represents 6-24% of population	2. Data is collected with user consent, but contains personally identifiable attributes	2. Data is updated quarterly	2. Data shows inconsistent correlation with other data sources	
3. Data is captured in an automated way, but is prone to errors or incorrectly capturing records	3. Process is manual, but data methodology is documented	3. Data represents real world domain but is prone to regular errors/ anomalies that reduce confidence	3. Data represents 25-50% of population	3. Data is collected with user consent and contains de-identified attributes. However, data can be linked to other datasets to re-identify individuals (through the use of a unique identifier)	3. Data is updated monthly	3. Data shows matching patterns and trends with other third party-data sources (even if raw numbers differ)	
4. Data is captured in an automated way, and is mostly free from errors	4. Process is semi-automated and documentation on methodology, including calculations, transformations and other assumptions is documented	4. Data is error free and complete but is unable to be verified through an alternative data source or QA method	4. Data represents 51-75% of population	4. Data is collected with user consent and contains no personally identifiable attributes. Data could potentially be linked to other datasets to re-identify individuals, though the process to do this is difficult	4. Data is updated weekly	4. Data shows matching patterns and trends with other Council data sources, such as counters, sensors, or system data (even if raw numbers differ)	
5. Data is captured in an automated way, and is entirely free from data capture errors	5. Data transformation/ cleansing process is automated or pre-processed, documentation exists, and methodology has been shared with end users, and QA'd by other analysts	5. Patterns/trends are accurately represented, and can be cross-checked or validated through other methods or data sources	5. Data represents 76-100% of population	5. Data is collected with user consent and contains no personally identifiable attributes. Data cannot be linked to other datasets to reidentify individuals	5. Data is updated in daily or in real-time	5. Data numbers, patterns and trends correlates with reputable data sources	BRISBANE CITY *Dedicated to a better Brisbane*