

Coursework 1 – COMP6214 Open Data Innovation

Kang Cheng

Host of Visualization

This project is hosted on GitHub.

A live demo is at <http://chengkang.me/coursework/comp6214/cw1/>.

Source code at <https://github.com/cheng-kang/coursework/tree/master/comp6214/cw1>.

Error in Data

1. Different names for the same agency.
 - E.g. grid 16D & grid 17D.
 - Unify them to the most common one.
2. Different date format.
 - E.g. grid 5K & grid 6K.
 - Unify them to the most common format.
3. Empty data.
 - E.g. grid 5G.
 - Multiple ways to deal with empty data:
 - i. Discard the whole row if columns being used to do visualization include empty grid.
 - ii. Mark as unknown.
4. Numbers contain unit string.
 - E.g. grid 26N.
 - Parse them to numbers. E.g. `parseFloat(data)` in JavaScript.
5. Redundant “total data”.
 - E.g. line 177 – line 452
 - Delete them.
6. Misplace of columns.
 - E.g. line 650 after column G. Line 178 to line 452.
 - Manually merge wrongly separated columns and place columns into right places.
7. Wrong data.
 - E.g. in the column of Unique Project ID, there are multiple identical ids.
 - Discard the whole column since there is no way to correct those data.

Intended Audience & Visualization

The intended audiences are the US government (mostly) and US citizens.

Visualizations include:

1. Average completion time of all departments
2. Investment distribution among departments
3. Investment completion condition of all departments

Those visualizations should help the US government to get an overview of the distribution of money spent on different departments, and the efficiency of different departments. They will help the US government to decide future financial allocation. Also, they could help US citizens better understand government investments on different aspects.