The goal in creating these data visualizations was to demonstrate interesting correlations between information on the artist and information on their work. Does being born in a particular area correlate with art being located in a particular borough, or the use of a particular material? By comparing artist dates of birth with information about their works, is it possible to see trends in material usage change over time? Unfortunately, the data did not provide the sort of insights I was hoping for, though they did illuminate some of the limitations of linked data as it exists today. More explanation will be provided below.

Note: In the following graphs, the data set has already been filtered to exclude artist list items that were obviously typos, repeats of other artists, or lists of artists, to avoid representing one person multiple times.

Sheet 1: This packed bubble visualization plotted artist birthplace against the borough their works were located in, to try to divine whether artists from a particular area tended to have works in a particular borough. In a certain sense this visualization does expose such a relationship; it is clear that the majority of monuments from artists not born in New York City are located in Manhattan, followed by Brooklyn, then Queens, the Bronx, and finally Staten Island. However, this is not an especially interesting statement, especially considering that artists born in New York City had essentially the same distribution (this could probably be attributed to the higher density of monuments in Manhattan and Brooklyn overall). Furthermore, the overall sample size is so small (the only pairing with more than three items is NYC/Manhattan, with eight) that trying to draw conclusions from the dataset is not a good idea.

Sheet 2: This visualization plots artist birthplace with the materials they used. The main issue with this visualization is that there was no controlled vocabulary of material terms used in the csv, so there are a huge number of separate items that actually represent the same material (ex: Bronze, bronze, bronze (?), bas-relief -- bronze). Thus, while the visualization as it is indicates that the most commonly used material from NYC natives and non-natives is Bronze, it is possible (though unlikely) that another material could take the lead if material terms were joined together properly. The materials could be manually renamed, but given the small size of the dataset (again the largest pairing only has six items with no other pairing greater than three) it wouldn’t be worth the effort, as there would still be too little data to make meaningful observations.

Sheet 3: This visualization exposes why the datasets in the first two sheets are so small. This is a less filtered version of sheet 2, that doesn’t filter out artists with no birthplace or artists with no listed materials used (or both). The n/a bubbles (indicating no birthplace recorded) are by far the largest. Since most people with a wikidata page have their birthplace listed, this indicates that the majority of the artist names on the list could not be resolved with a page in wikidata, either because the artist isn’t listed on wikidata, or because their names didn’t match closely enough (probably because the artist is only listed under one name in wikidata, with no aliases to try to match the list name with).

Though these visualizations did not provide the sort of information I was hoping for, they did demonstrate something useful: the limitations of linked open data as it exists today. While it is good that csv files like the one used in this project are made available, their utility is limited due to the lack of care in their production, with frequent typos and no use of controlled vocabularies to aid in data analysis. As for wikidata, while it is already an impressive tool (it is shocking to me how easily I was able to access such a large amount of data) it is unfortunately still incomplete. There are simply too many artists that have no representation in the data, and too many of the existing pages are not detailed enough (especially in terms of aliases, which are crucial for name resolution). The potential value of the data I used makes its limitations all the more frustrating, though wikidata at least has potential to improve over time.