Dishes dataset:

Used various clustering techniques to cluster the name column in order to condense dishes into groups. Many dishes seemed to be the same, but had variations in spelling and ordering of words.

Before clustering, I title-cased the field, removed double spaces, and trailing whitespace. The thought here is that, “If I have ‘Steak Dinner’ and ‘Steak dinner’ (notice the double space and capitalization), that by doing the basic transformations to the text would result in both becoming “Steak Dinner” without even having to cluster.

After this, I did perform clustering to catch cases like “Honey Toast” and “Toast with Honey” to cluster to “Honey Toast” for example.

Next, I transformed the menus\_appeared and times\_appeared fields to be numeric. I first had created a facet to validate that the values were valid. Here, I found many had value of 0. This was a bit suspicious to me, but I figured, “Well, a dish can exist that nobody has placed on a menu”. To know for sure if 0 is valid or not, I would need to understand the domain application better. So, I will count 0 as a valid rather than converting to null.

For times\_appeared, I followed a similar approach as I used for menus\_appeared. This time, I found negative values in time\_appeared, though. These I converted to nulls.

Next, for first\_appeared and last\_appeared, I understand these fields to be years.

Looking at the text facet, I could see the distribution mostly landed between 1800 and 2012. That said, there were a lot of records in year 0,1 and 2998. These, I converted to nulls as they seem to be data errors.

For lowest\_price and highest\_price, I found many 0 values. For this use-case, I left them as 0 and did not filter them out. When querying the data to answer questions, the data requester can decide whether to keep or exclude 0’s in the query. I don’t want to filter them out at this point which would remove other valuable, valid, data. Example:  
If I remove a menu item that has price 0, first\_appeared 1900, last\_appear 1950, times\_appeared of 100, I lose those valid fields if I filter out anything with a price of 0. So, I will leave it in the data and allow the “querying user” to decide for themselves.