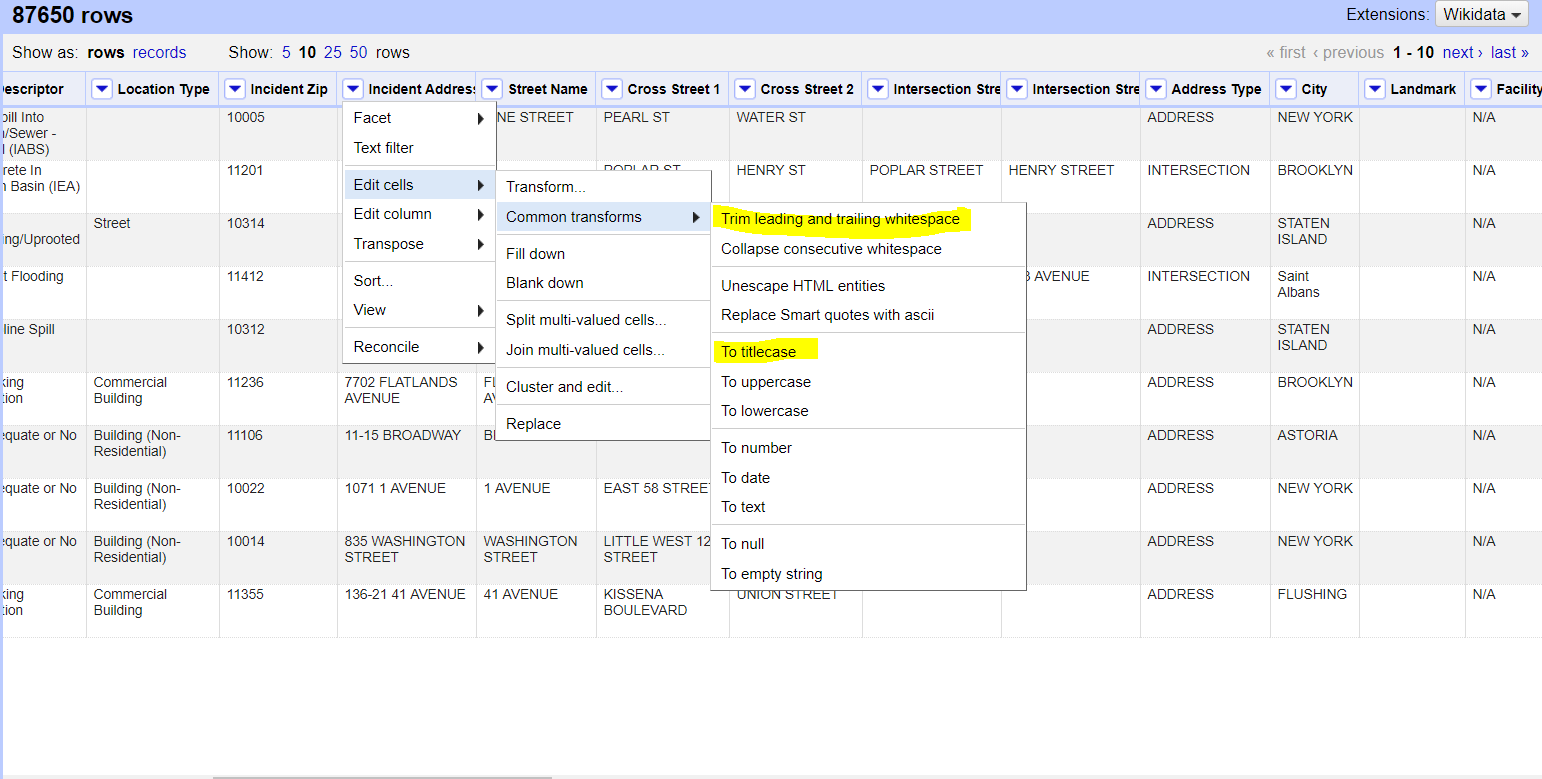
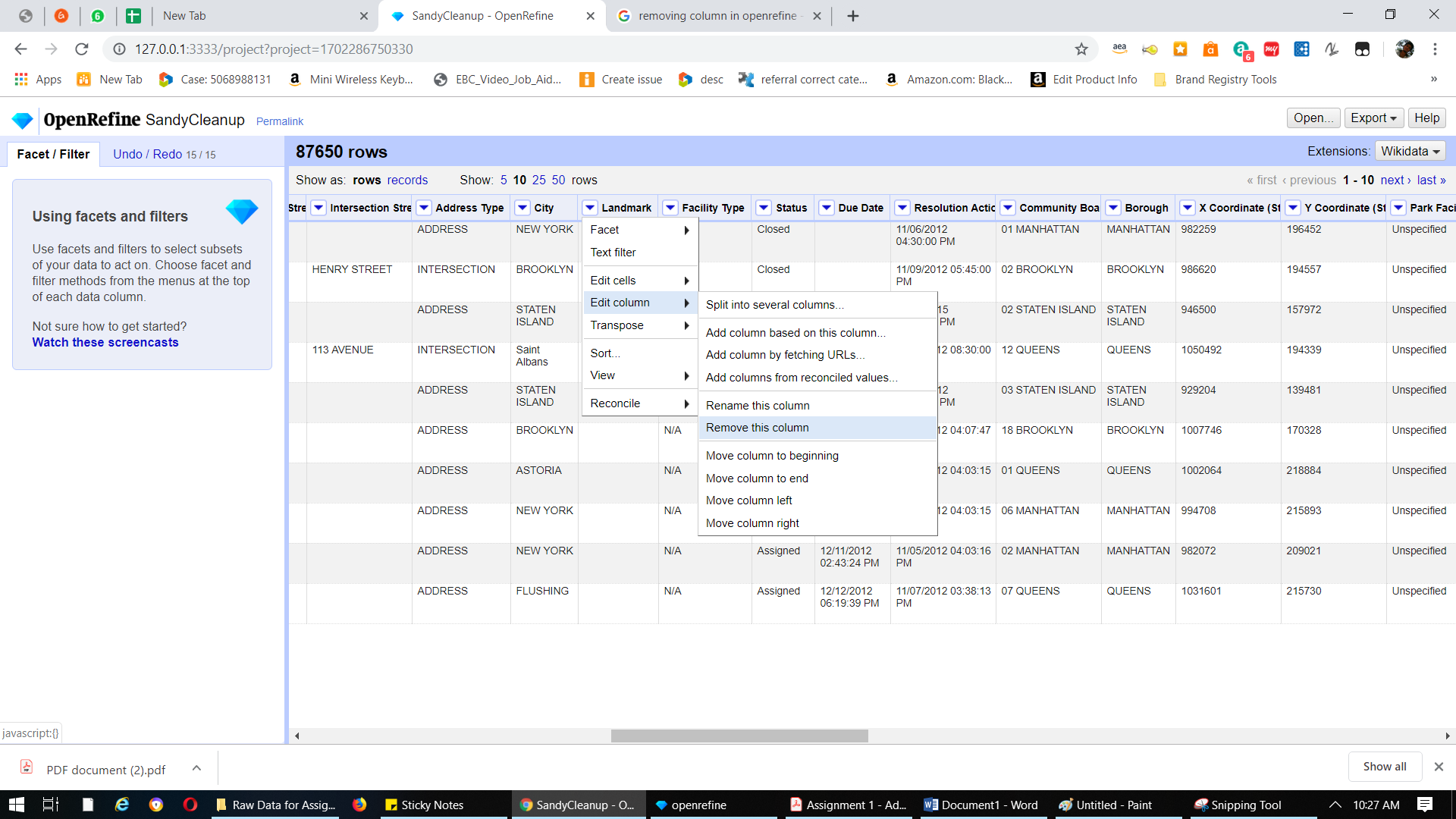
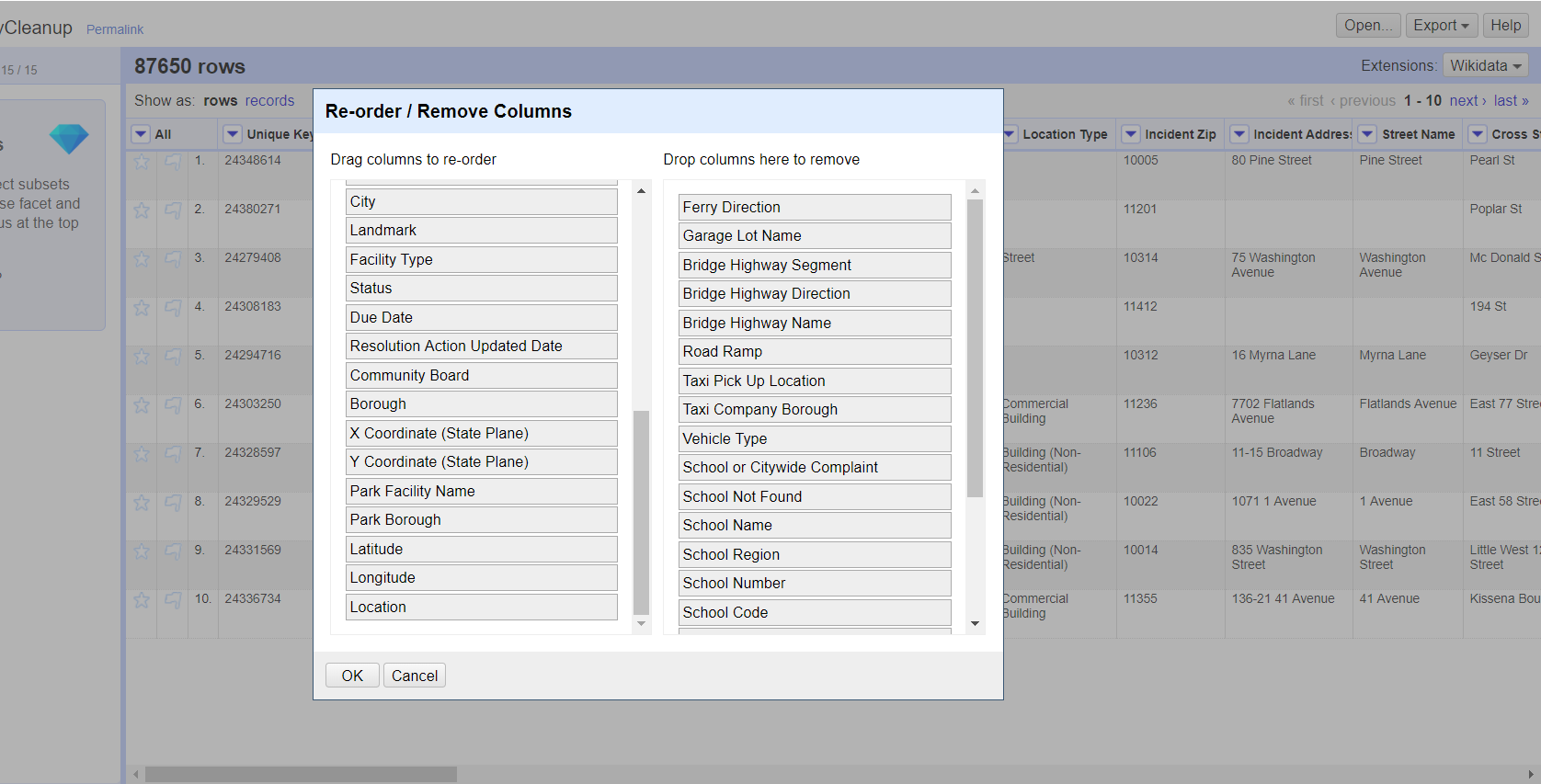
**1. Data Wrangling Exercise:**

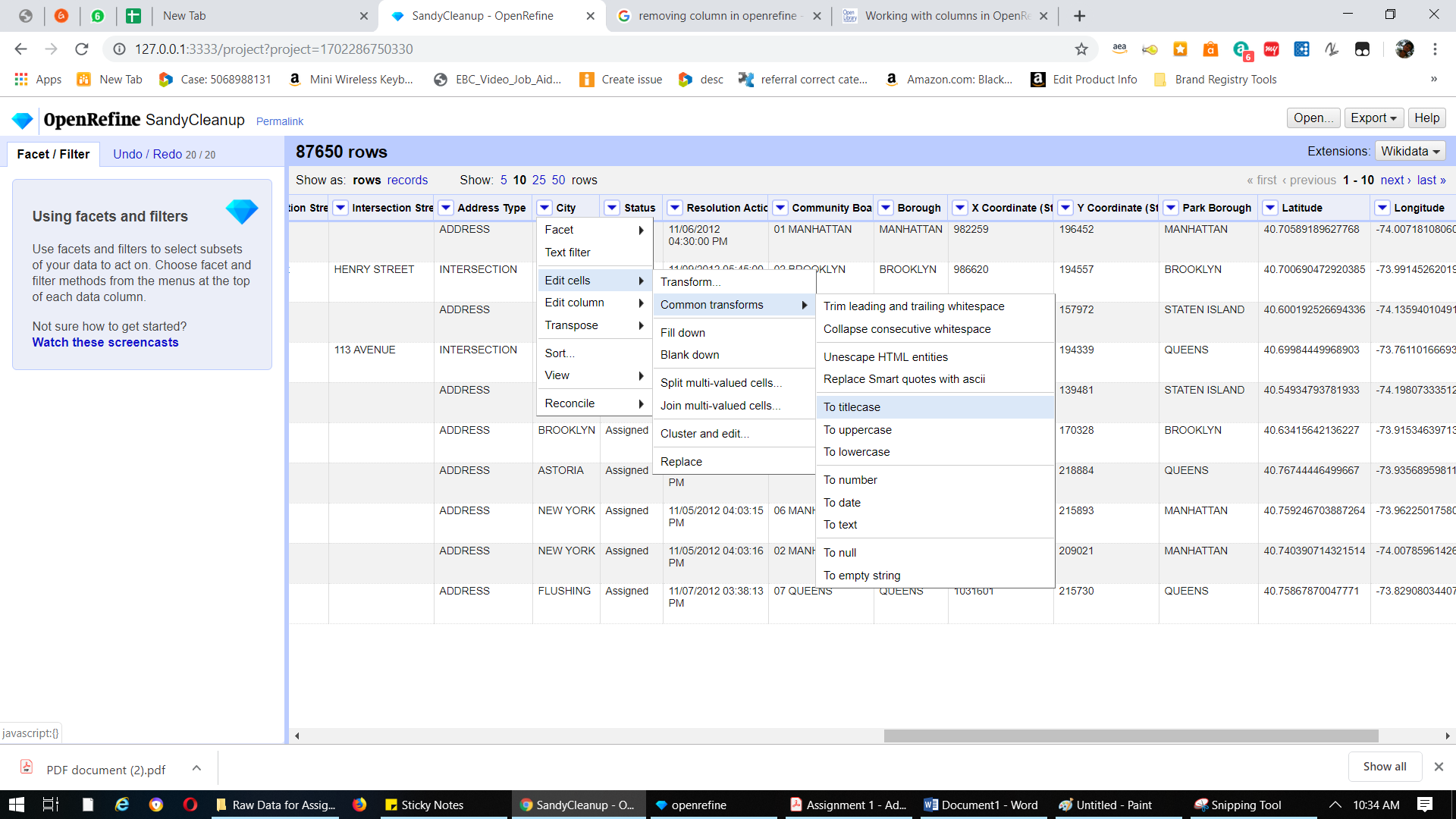
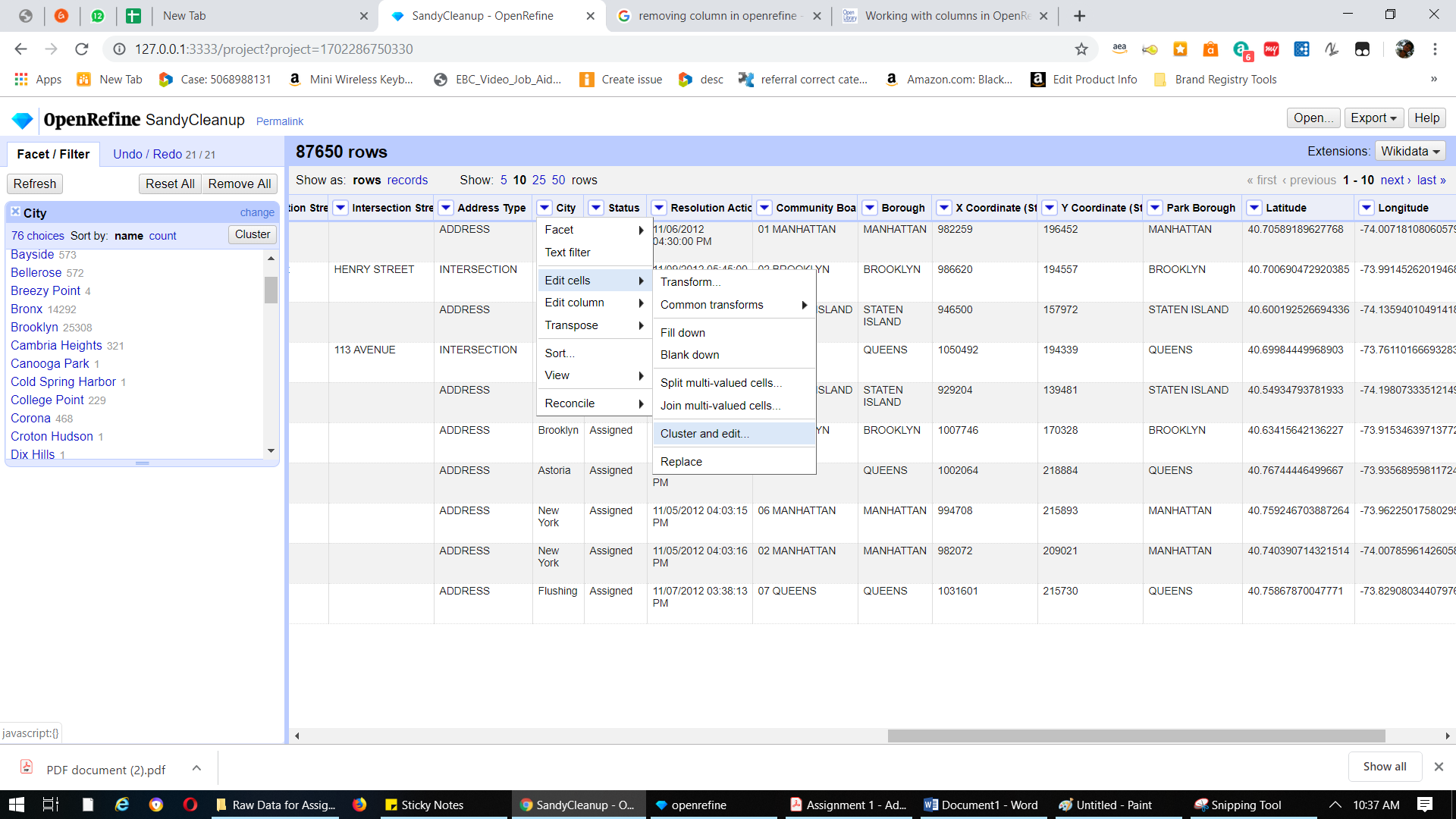
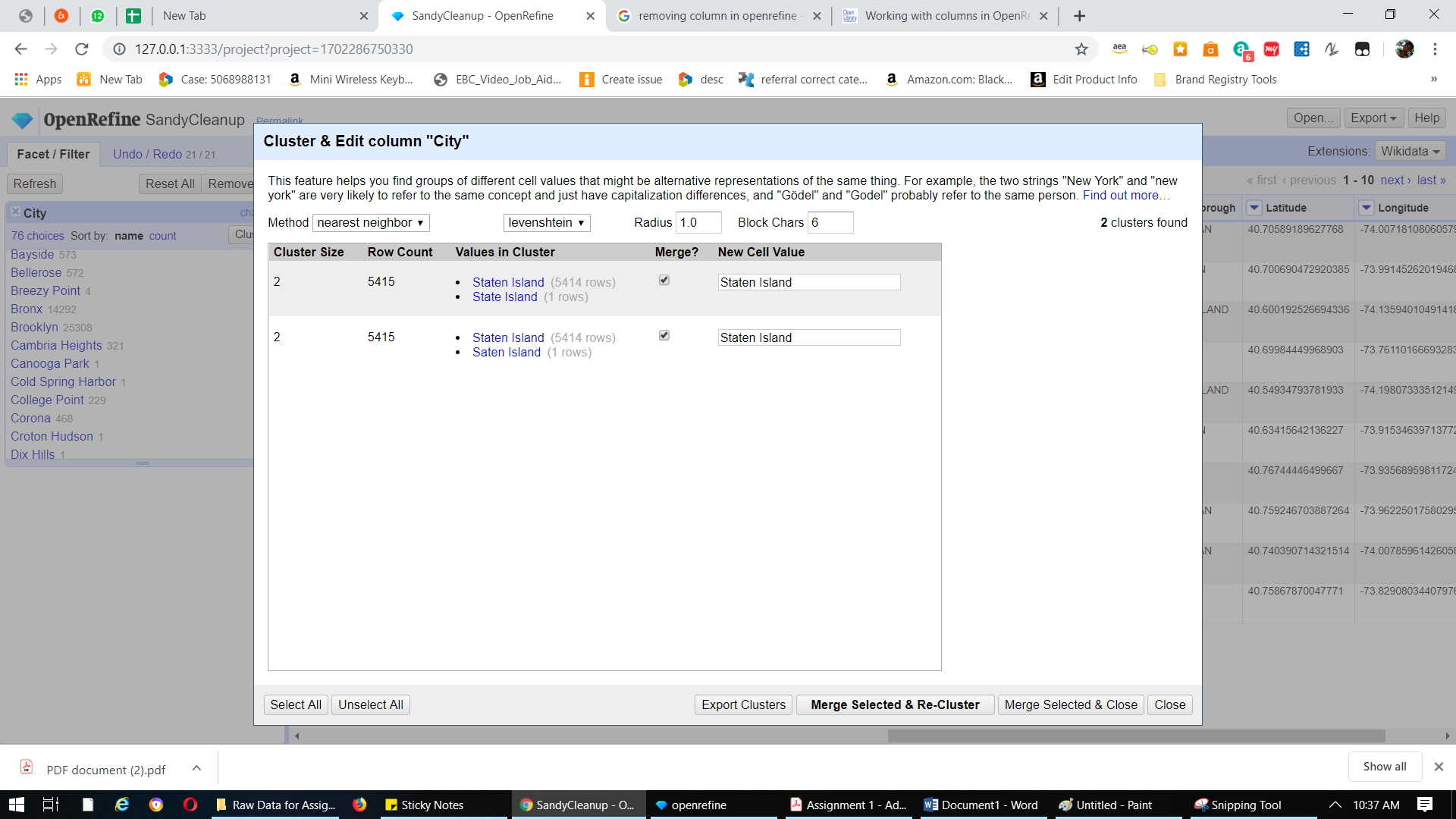
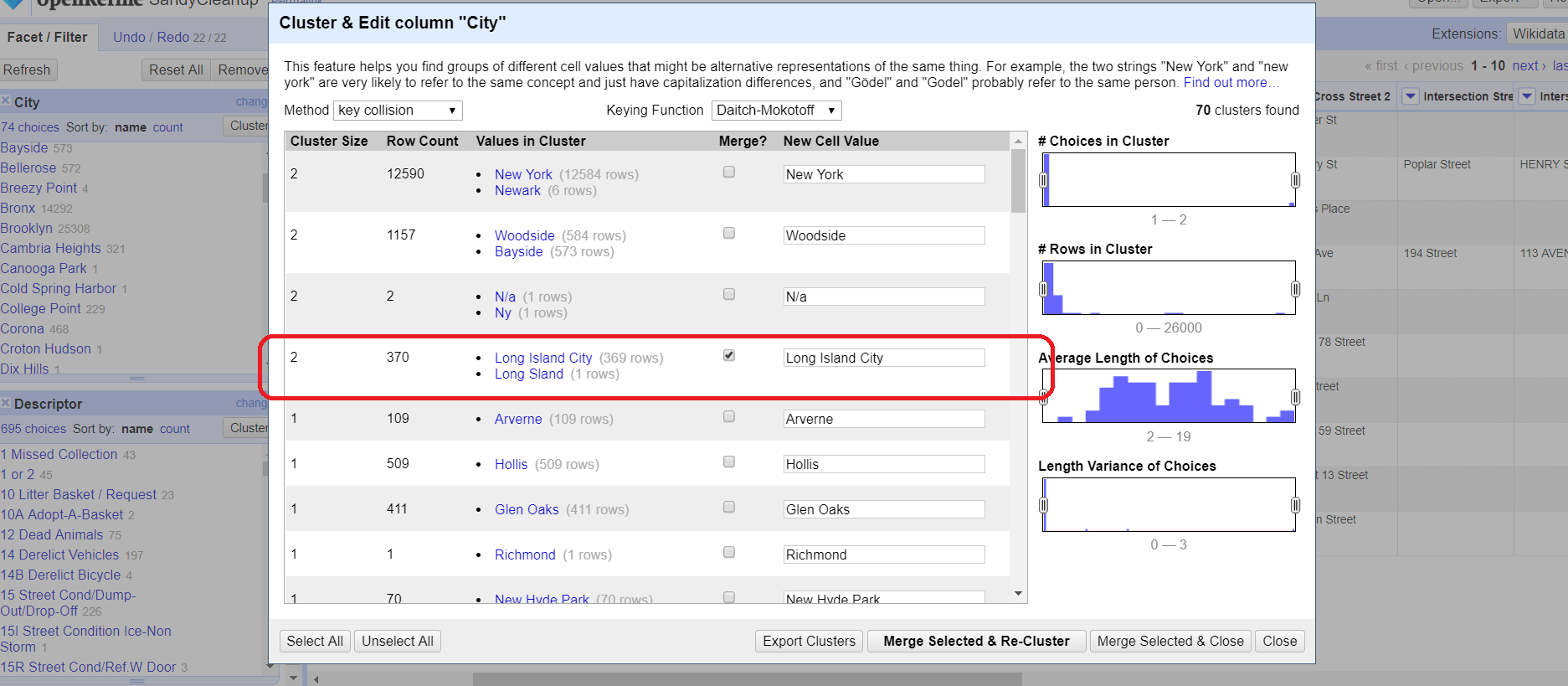
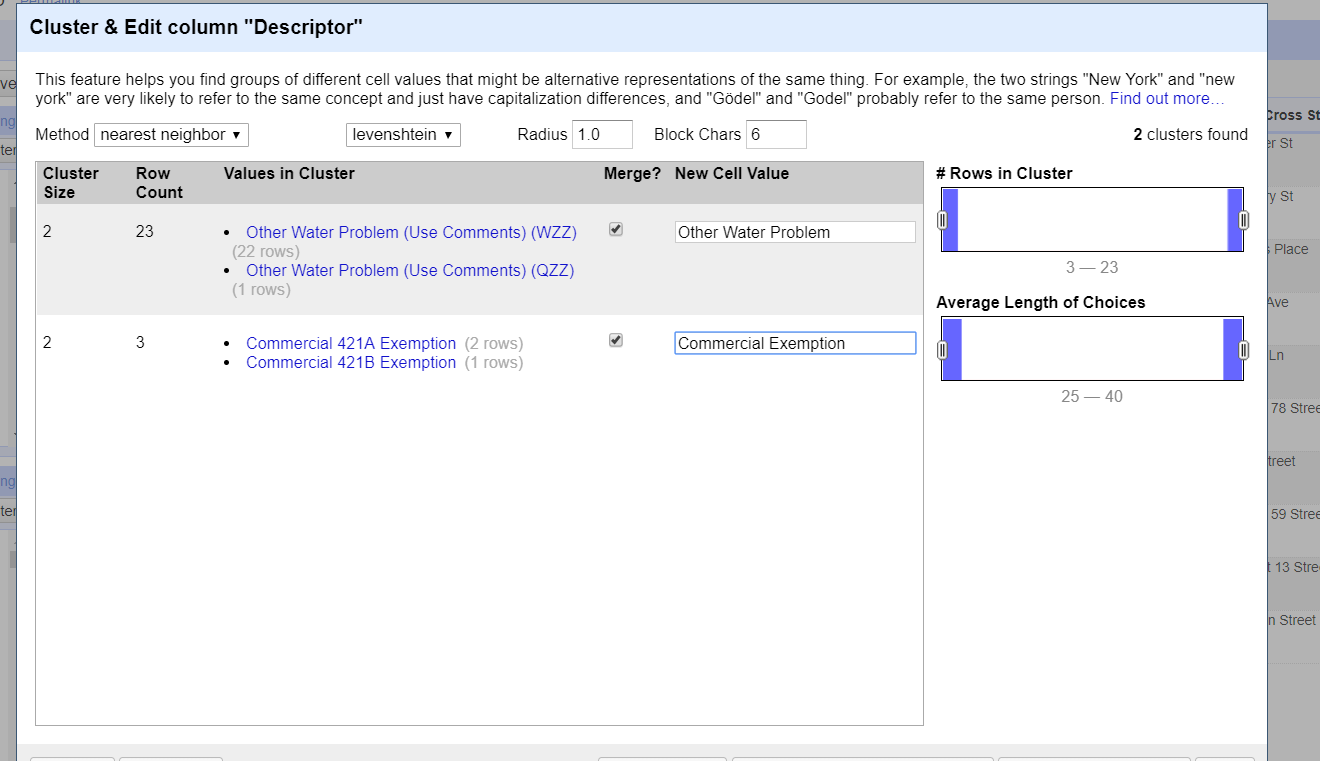
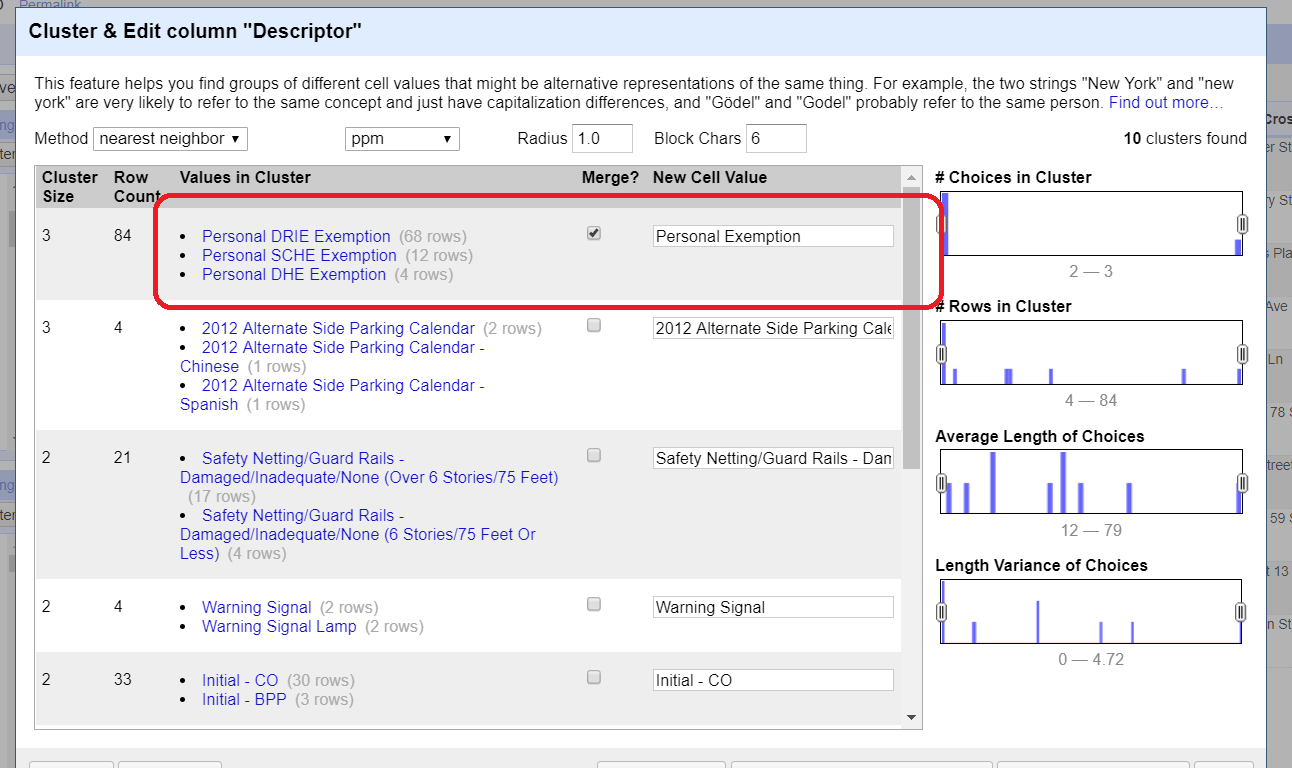
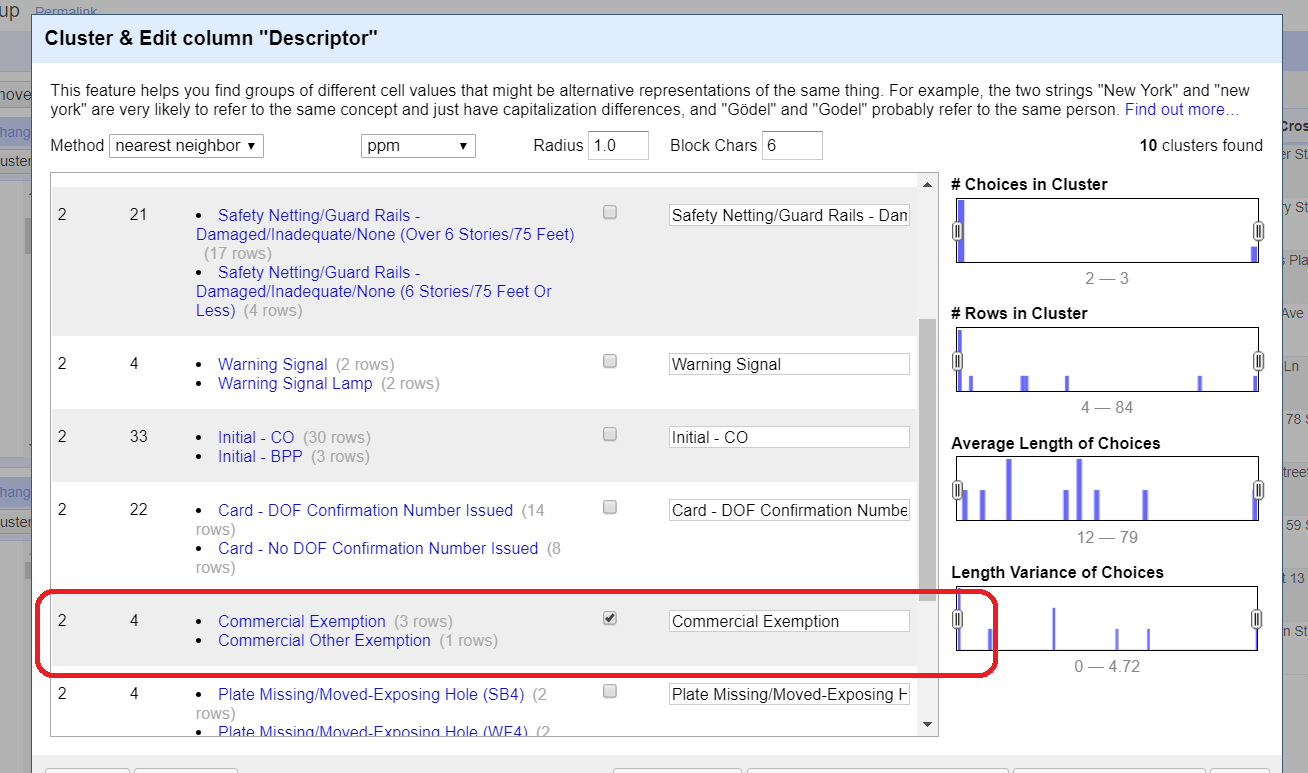
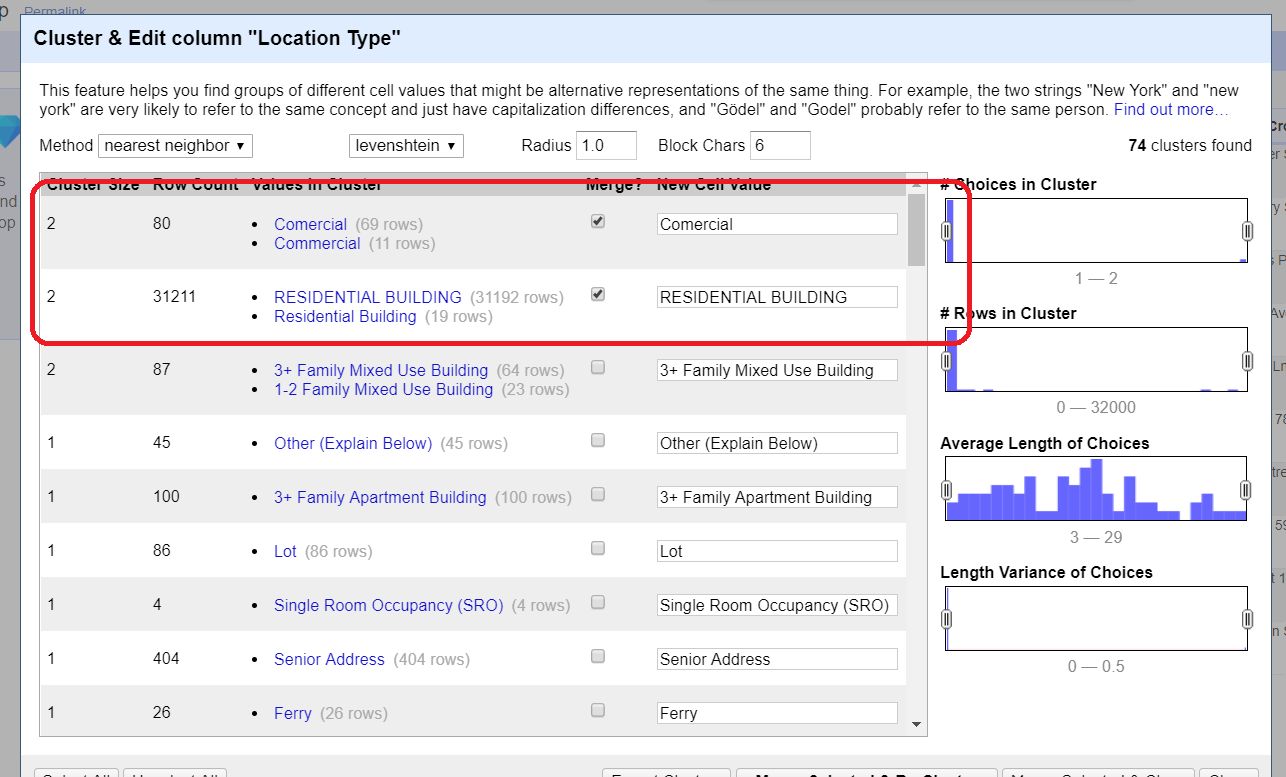
* 1. b. Trim white spaces on all address related columns and transform addresses into title case
* To remove the whitespace, chose **Edit cells > Common transforms > Trim leading and trailing whitespace** first and then used **“To titleCase”** common transfer to change the case for all address related columns (Incident address, street name, cross street1, cross street 2, intersection street)

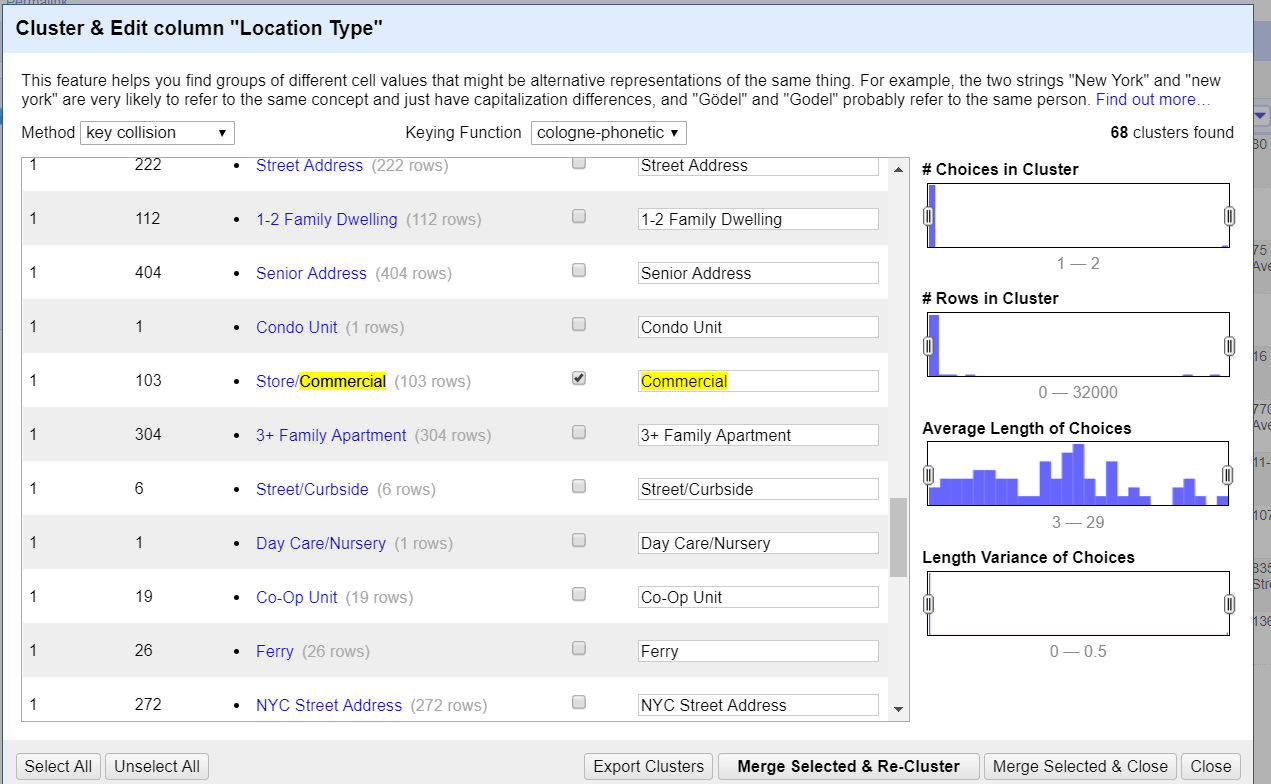


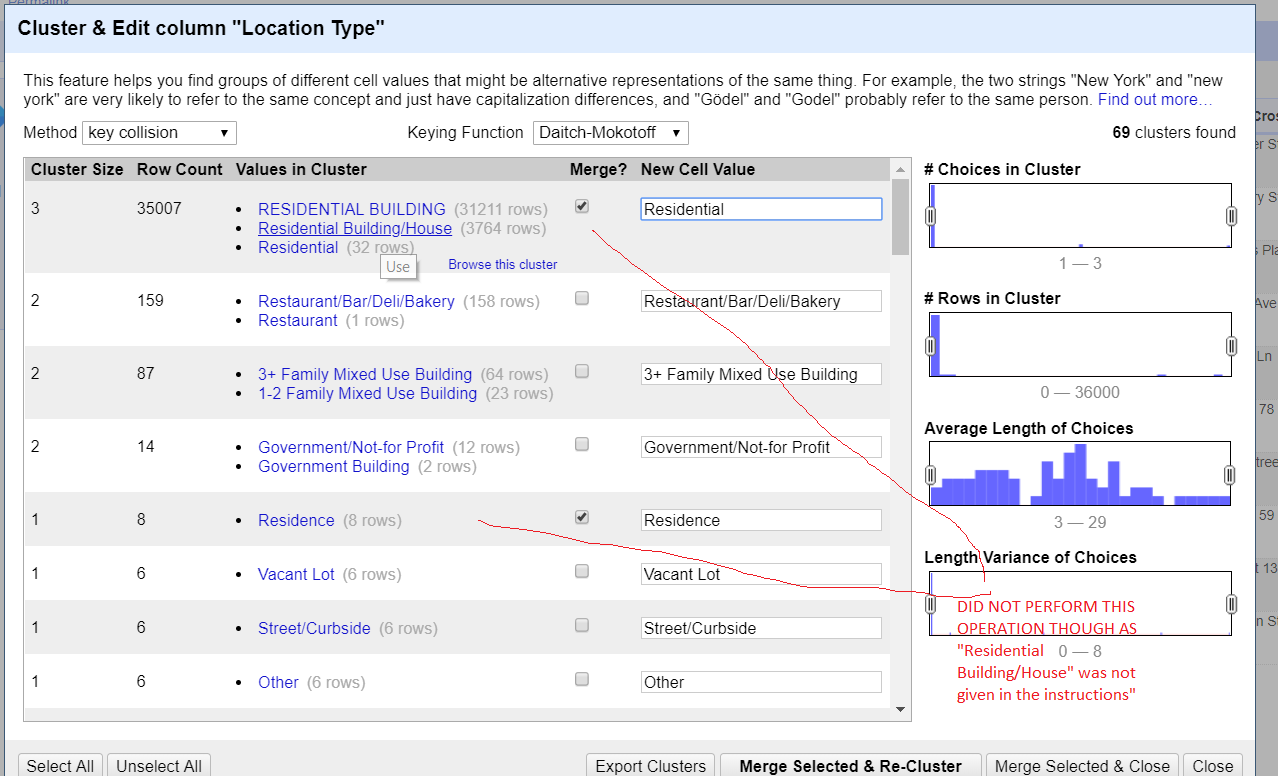
* 1. c. Remove columns where majority of the cells are empty or have "Unspecified" or “NA” values. Do not remove columns which are being used in subsequent questions.
* Used **re-order/remove columns** from “**All>Re-order/remove columns**” for majority of the columns that were empty starting with “School name till Lattitude” Column and removed columns manually for rest of the columns by “**Edit column>Remove this column**”, columns such as “Park Facility Name, Due Date, Facility Type, Landmark”.

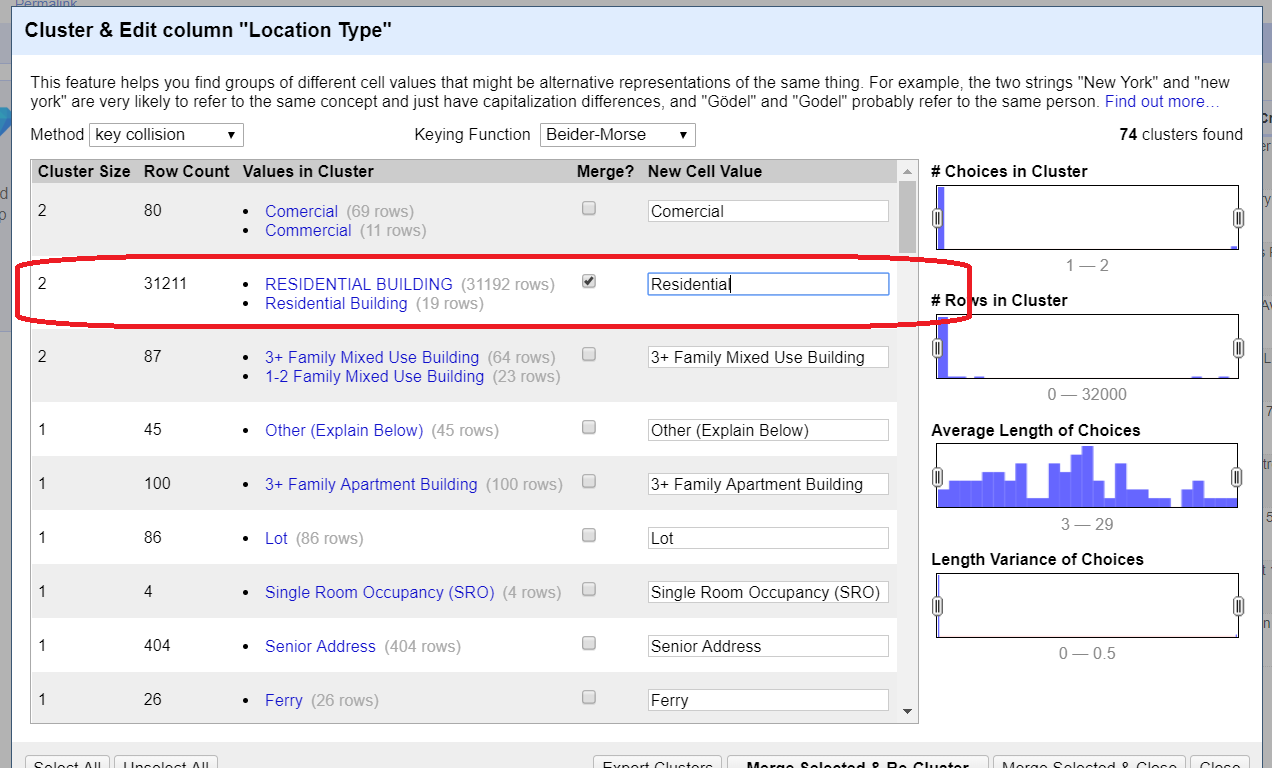




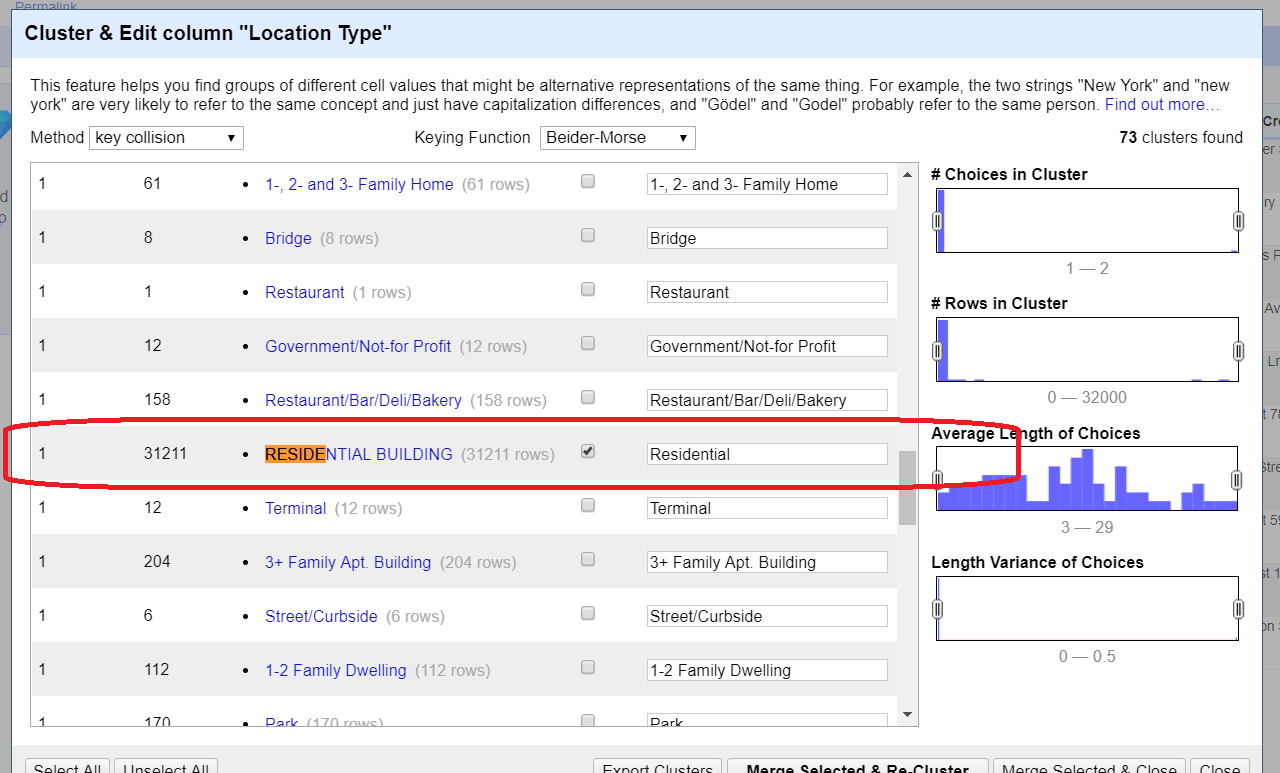
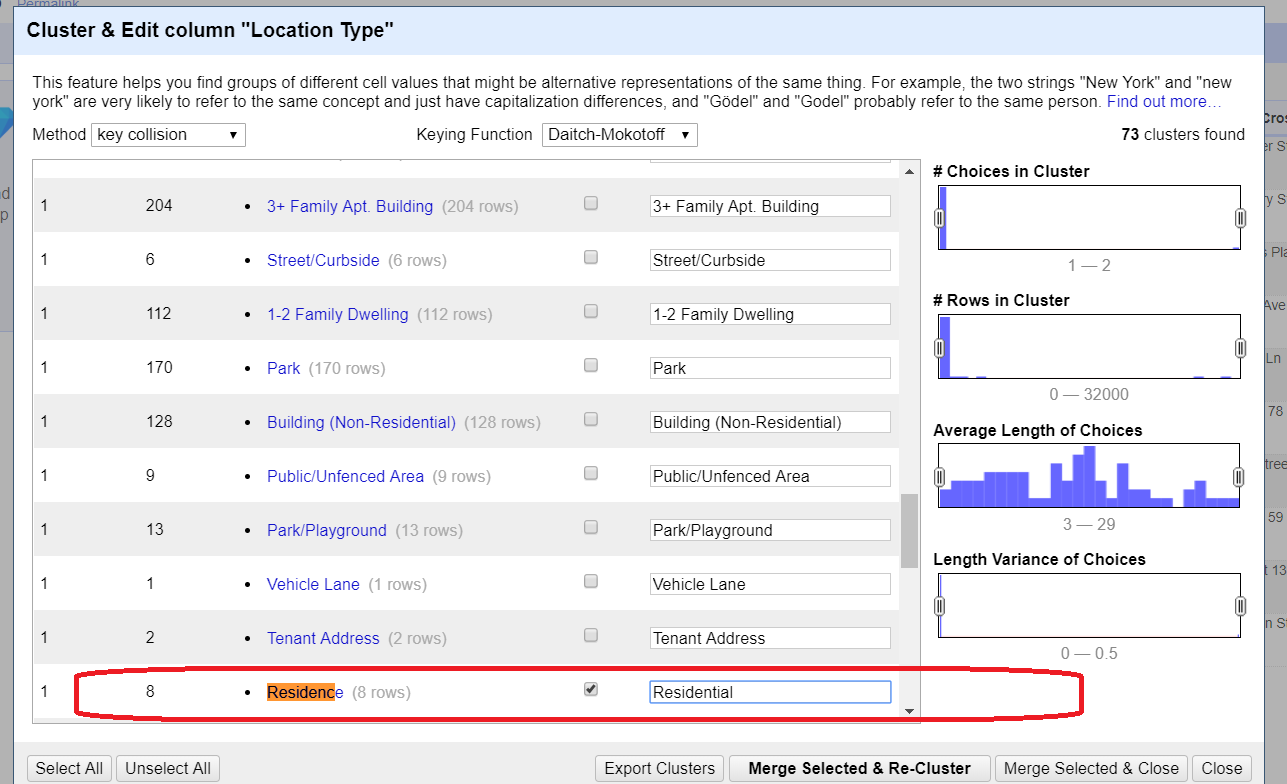
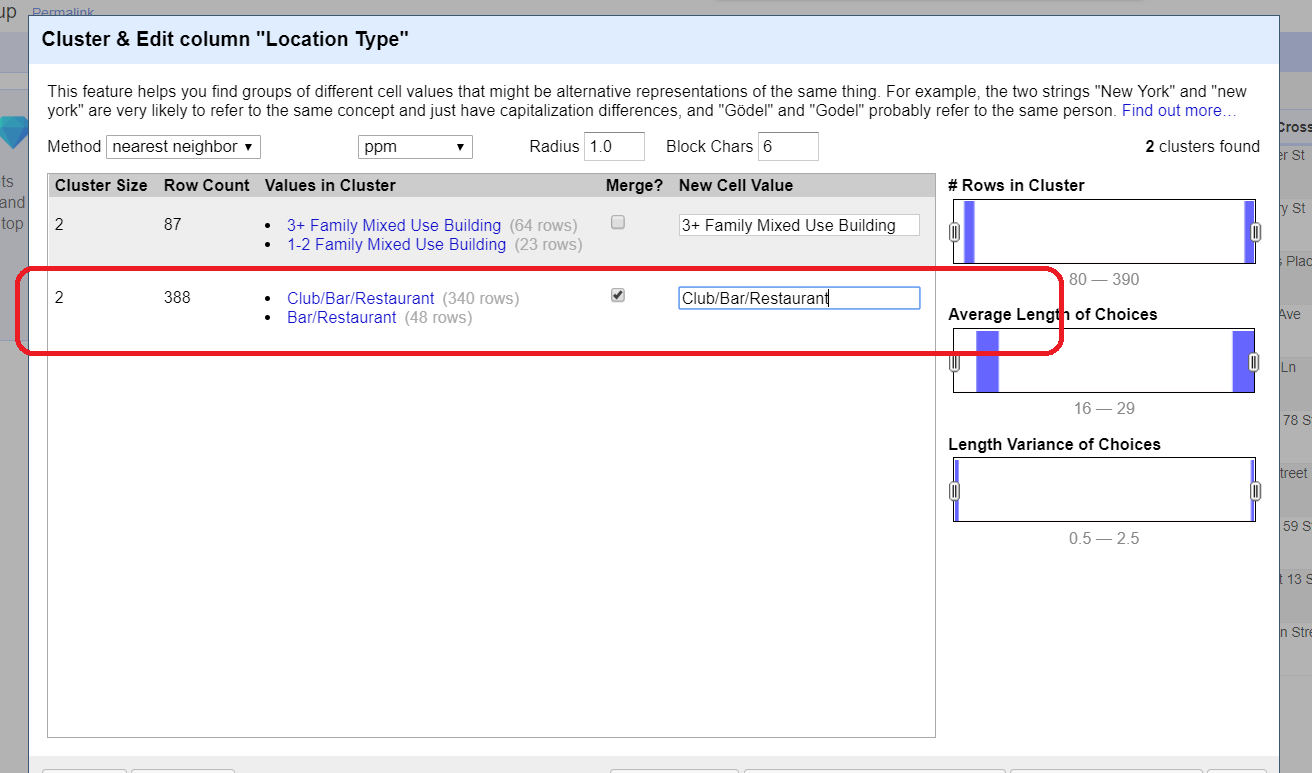
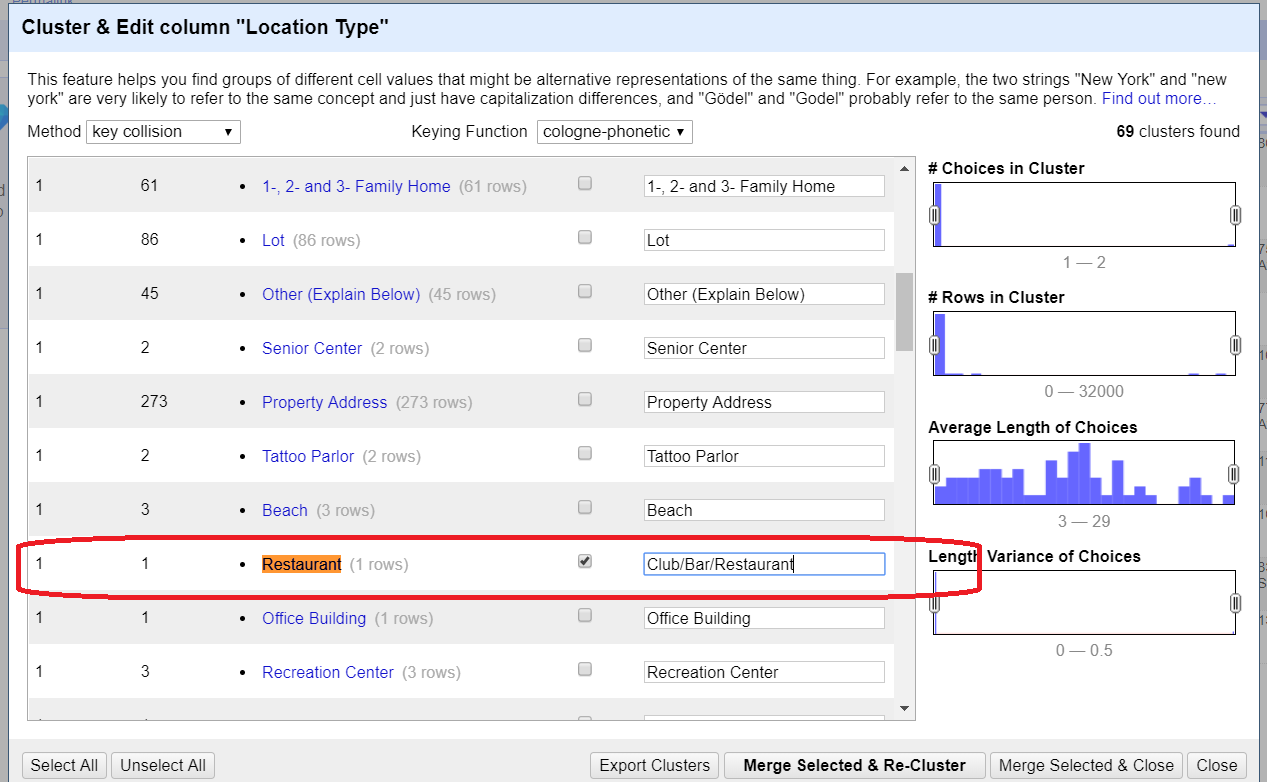
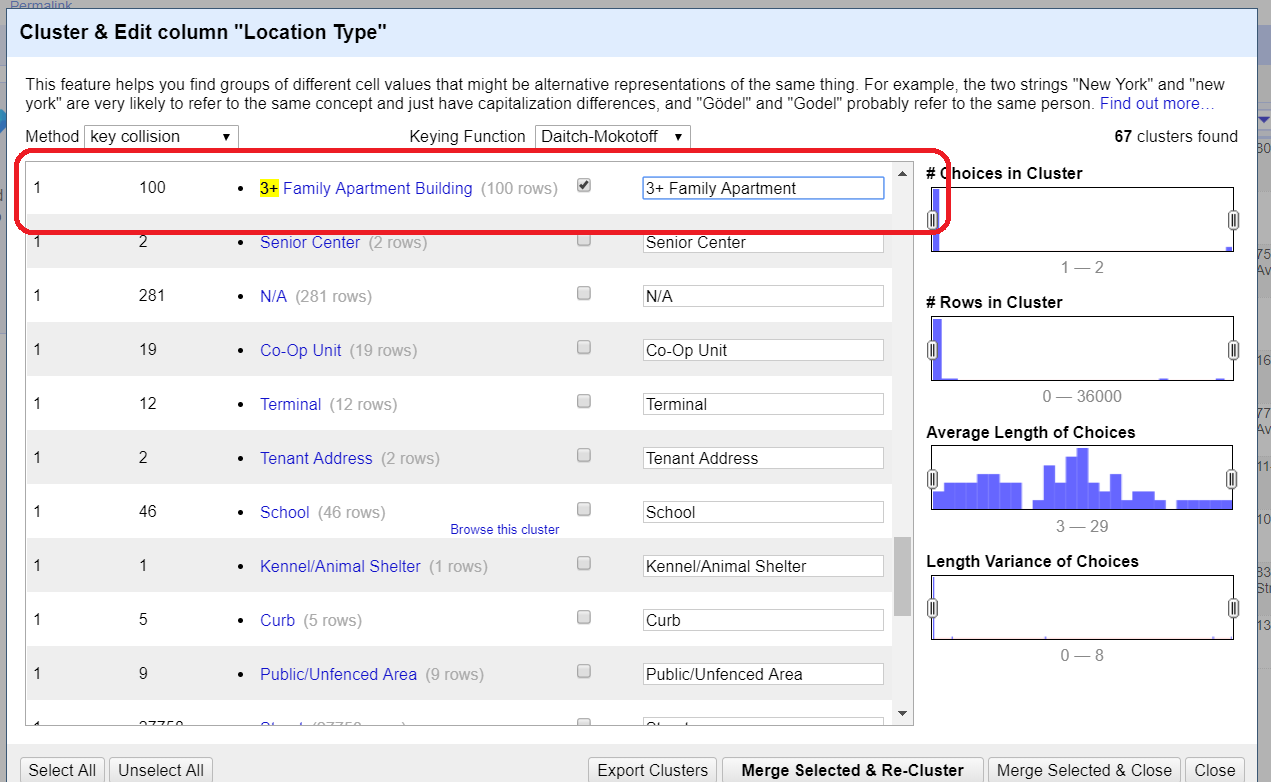
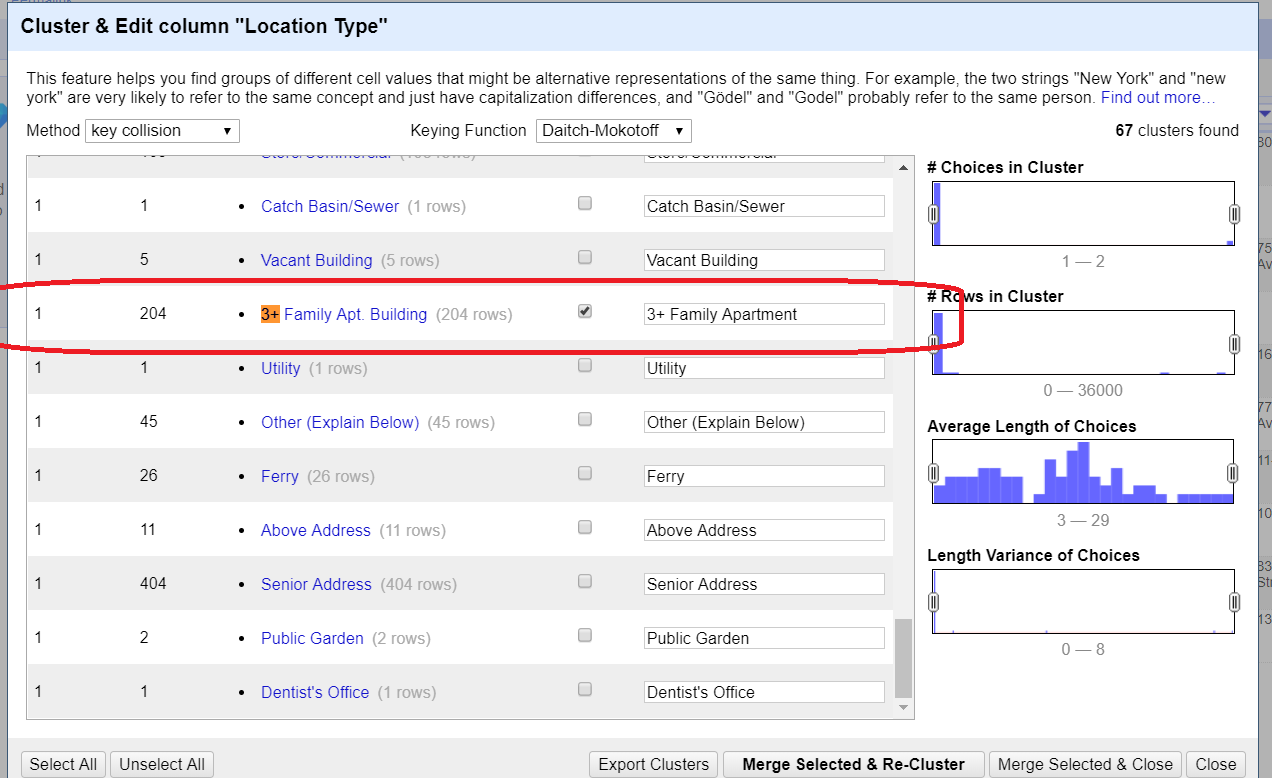
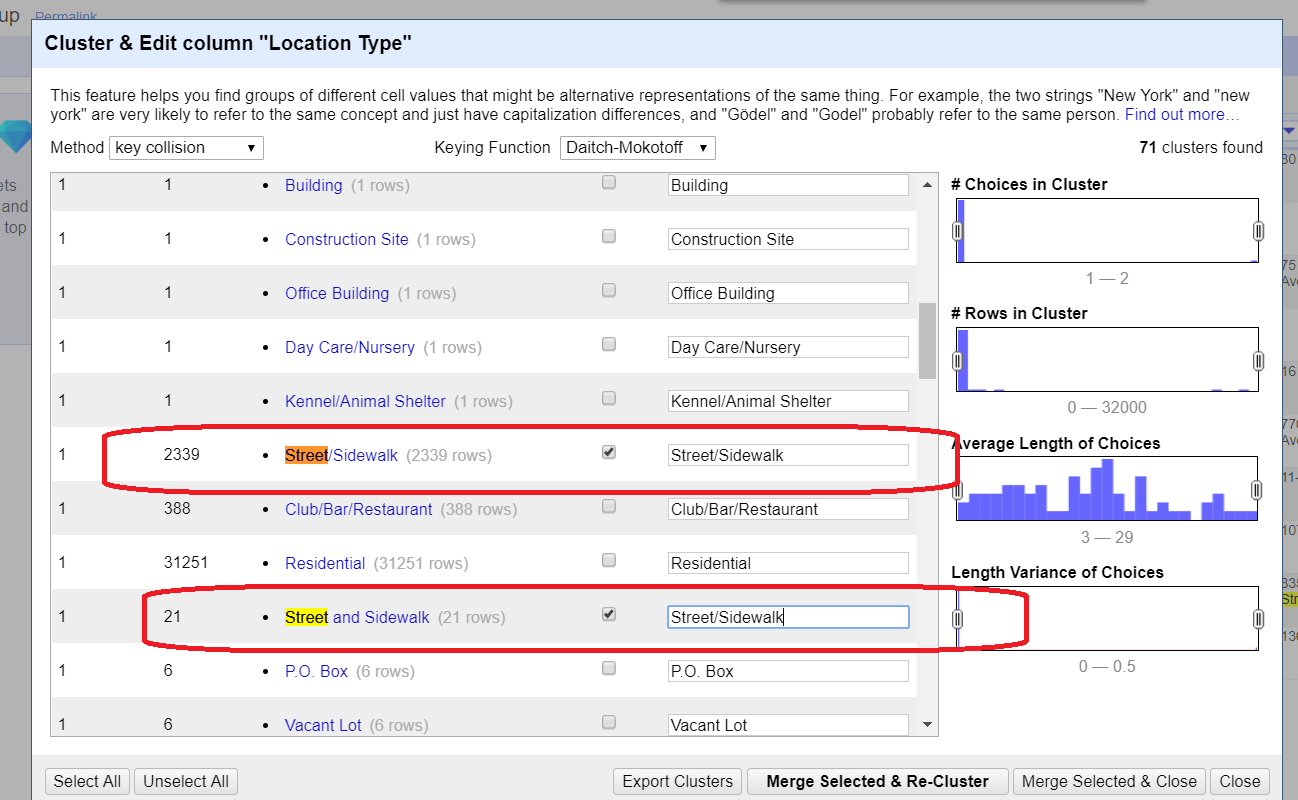
* 1. d. Convert the City column to title case, then Cluster and Merge the column
  2. 
  3. 
  4. 
  5. 
  6. e. Clean up Descriptor Column - Cluster and Merge following text categories:
  7. 1. "Other Water problem(WZZ)", "Other Water problem(QZZ)" as "Other Water Problem"
  8. 2. "Commercial 421 A/B Exemptions" as "Commercial Exemption"
  9. 
  10. 3. "Commercial Exemption" "Commercial Other Exemption" as "Commercial Exemption"
  11. 4. "Personal DRIE Exemption", "Personal SCHE Exemption", "Personal DHE Exemption" as "Personal Exemption"
  12. 
  13. 
  14. f. Clean up Location Type - Cluster and Merge following text categories:
  15. 1. "Comercial", "Commercial", "Store/Commercial" as "Commercial" - Overall 183 rows of Commercial at the end
  16. 2. "RESIDENTIAL BUILDING", "Residential Building", "Residence" as "Residential" – Overall 31251 rows of Residential at the end
  17. 



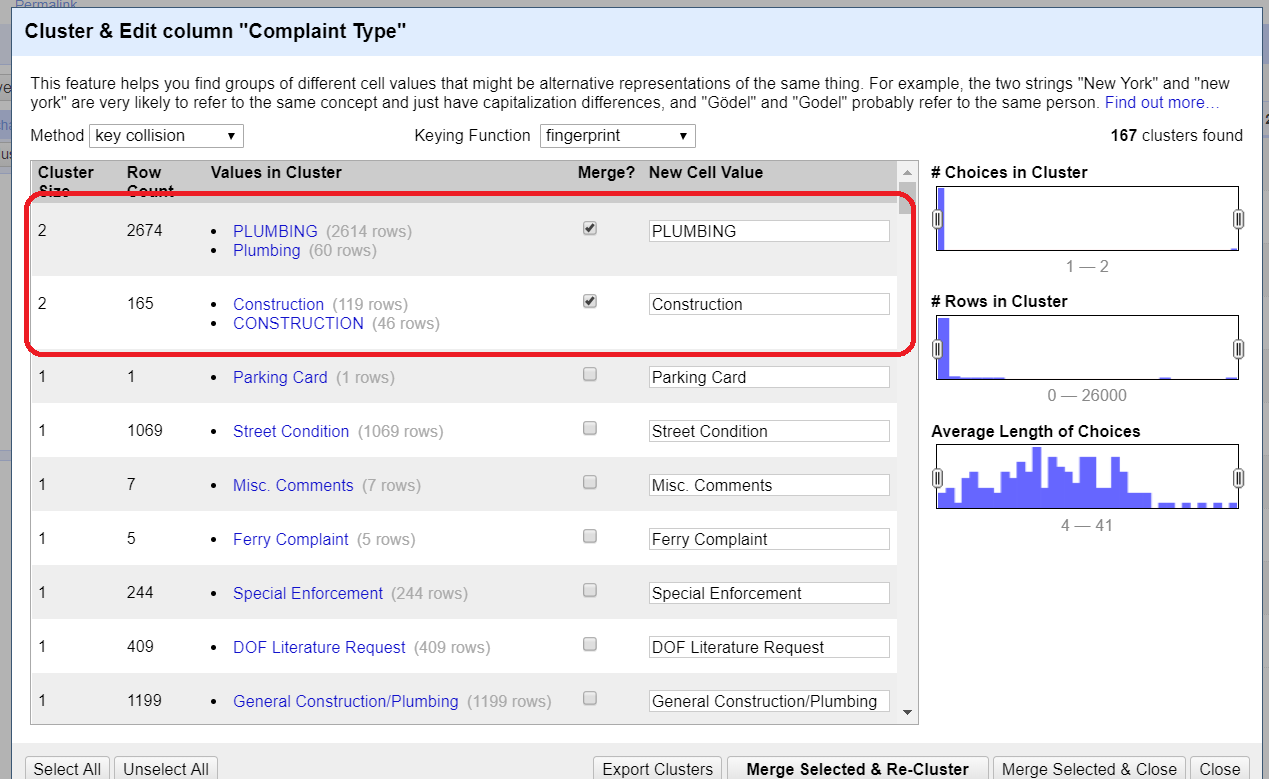


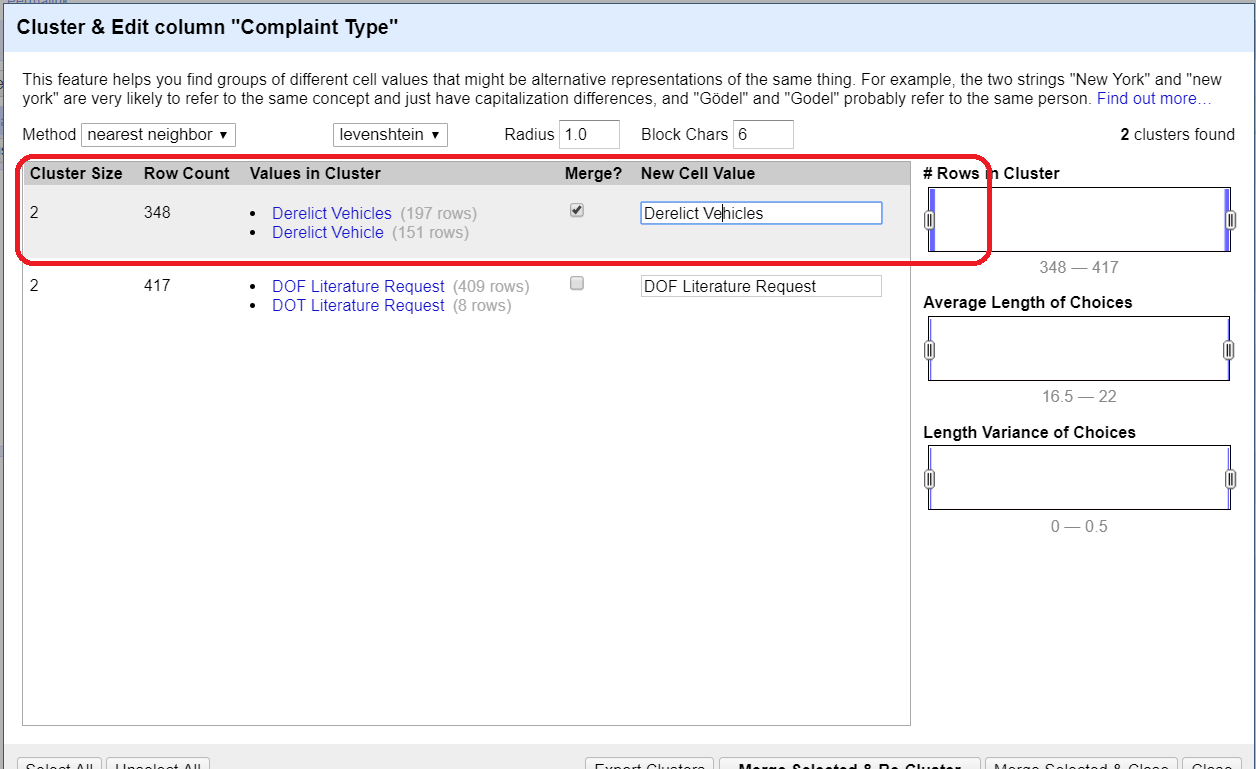
* 1. 

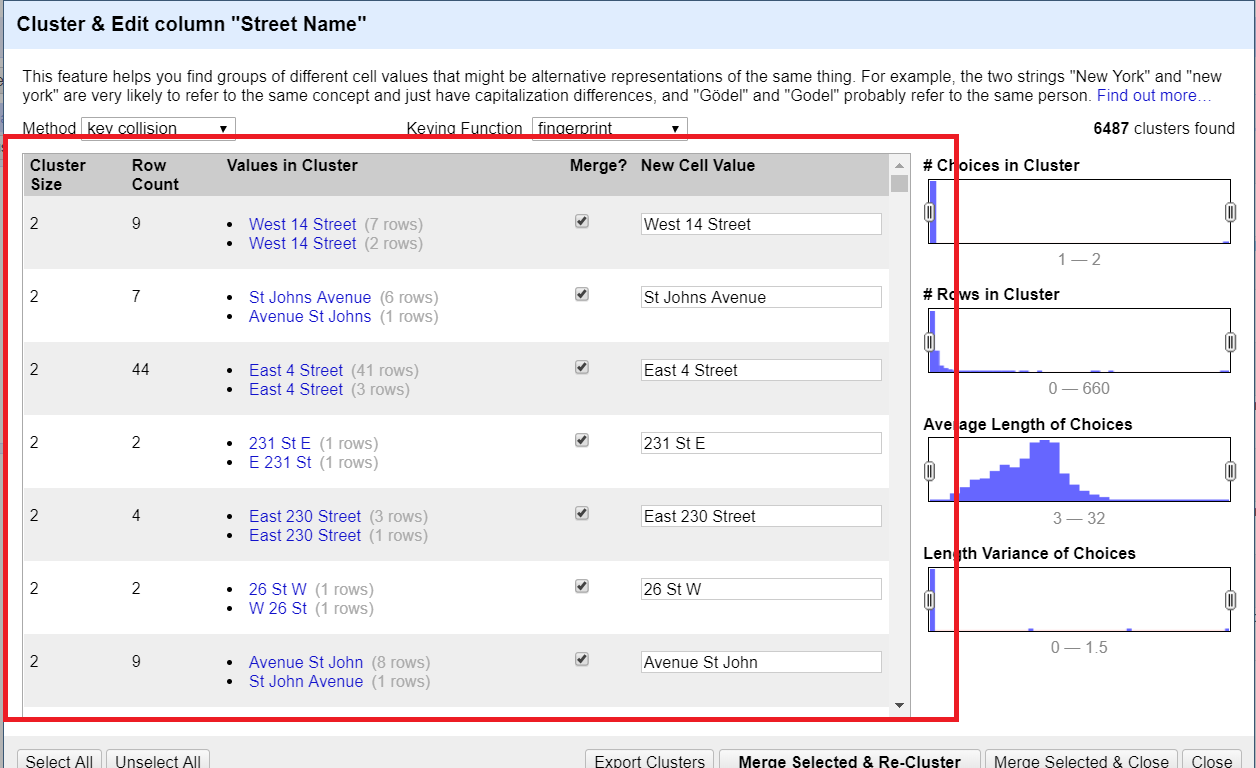
In below two screenshots, I found the RESIDENTIAL BUILDING first, checked it and renamed it to “Residential”. Similarly, found residence and renamed it to “Residential”.

* 1. 
  2. 
  3. 3. "Club/Bar/Restaurant", "Bar/Restaurant", "Restaurant" as "Club/Bar/Restaurant" - Overall 389 rows of Commercial at the end
  4. 
  5. In below screenshot, found the “restaurant”, renamed it “Club/Bar/Restaurant” and merged with existing “Club/Bar/Restaurant”.
  6. 
  7. 4. "3+ Family Apt. Building", "3+ Family Apartment Building" as "3+ Family Apartment" - Overall 304rows of Commercial at the end
  8. 
  9. 
  10. 5. "Street/Sidewalk", "Street and Sidewalk" as "Street/Sidewalk" - Overall 2360 rows of Commercial at the end
  11. 

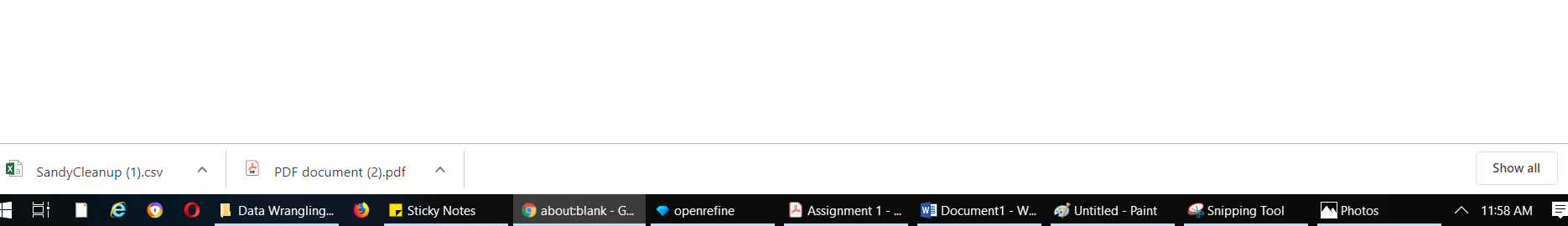
g. Look for at least two other clean up opportunities and execute using OpenRefine







* 1. h. Export final project into a CSV file on your local computer (you do not need to submit this file)



* 1. i. Online web services such as the following can be used to fetch the address given a geocode:

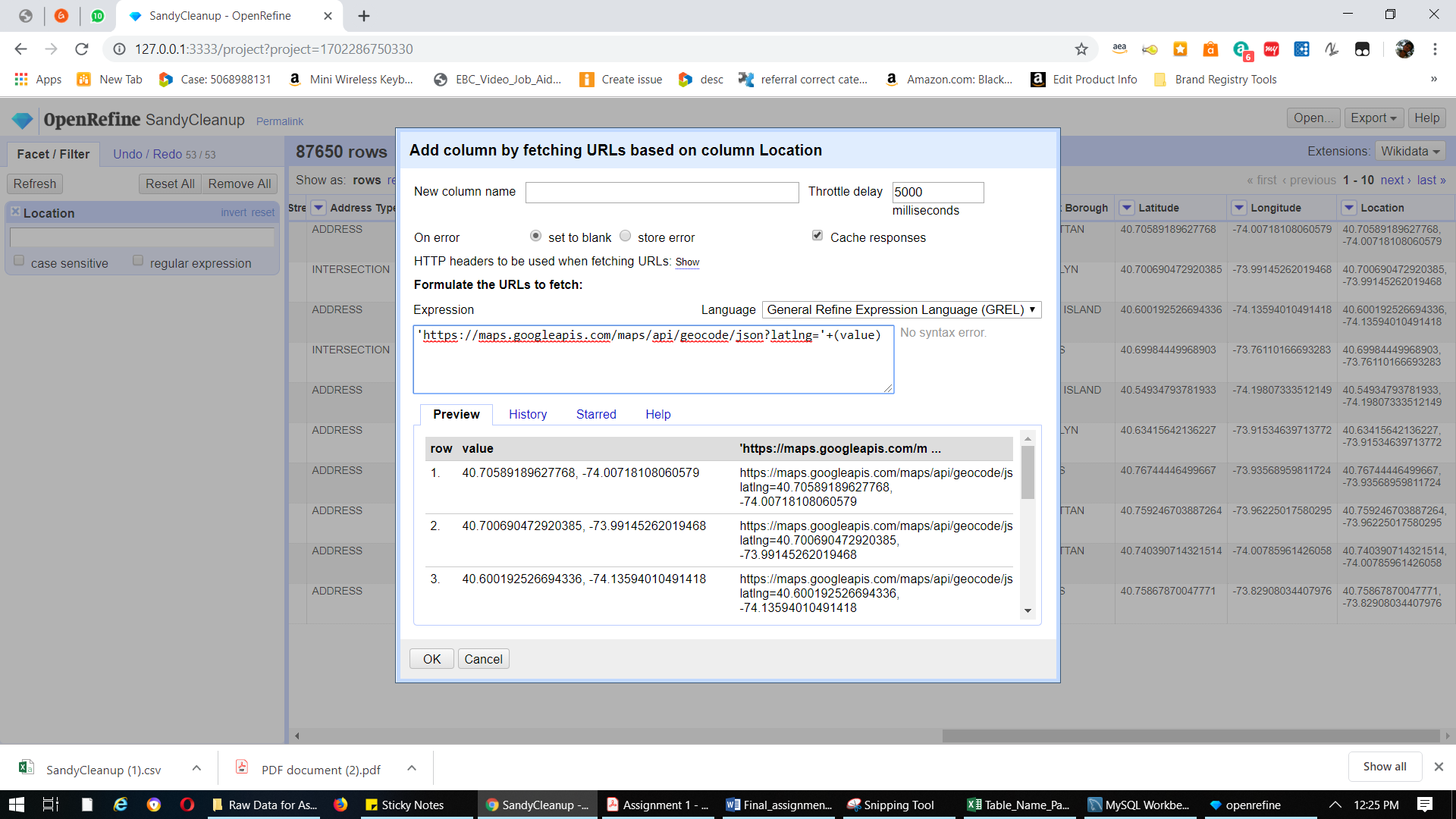
**Optional Practice Questions**

Google Reverse Geocoding for a Latitude/Longitude

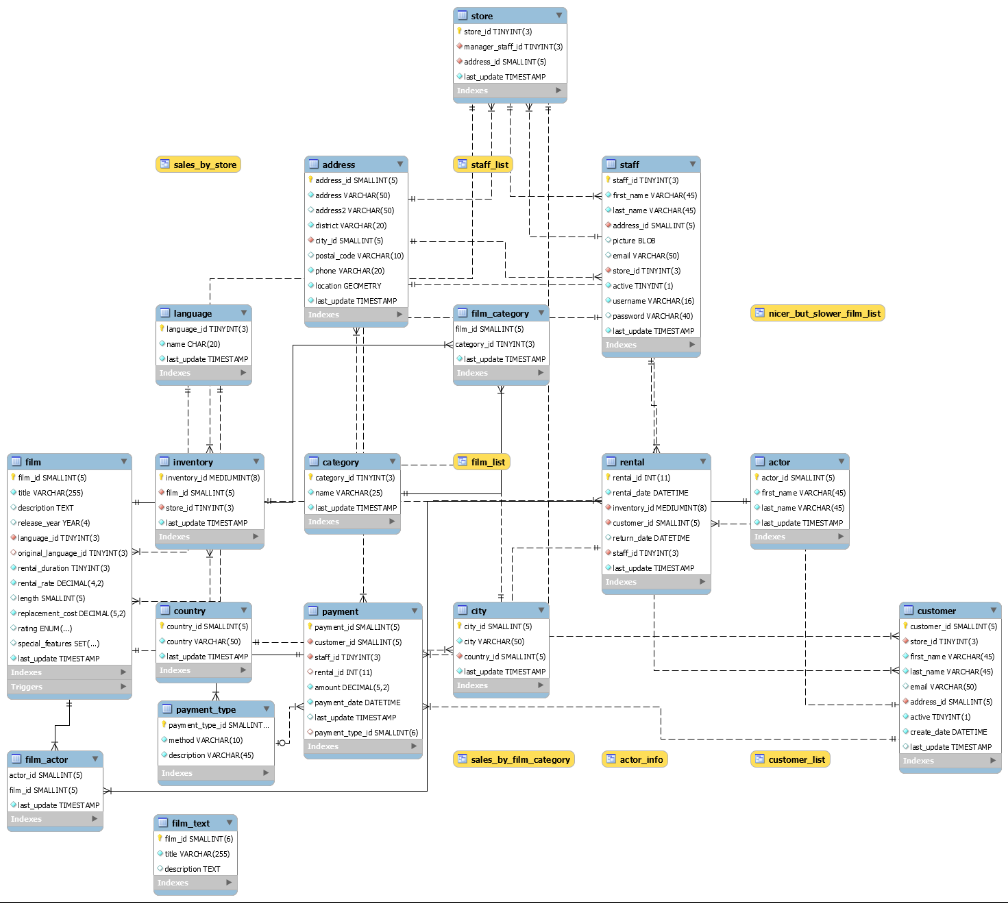
Web Service API Example:

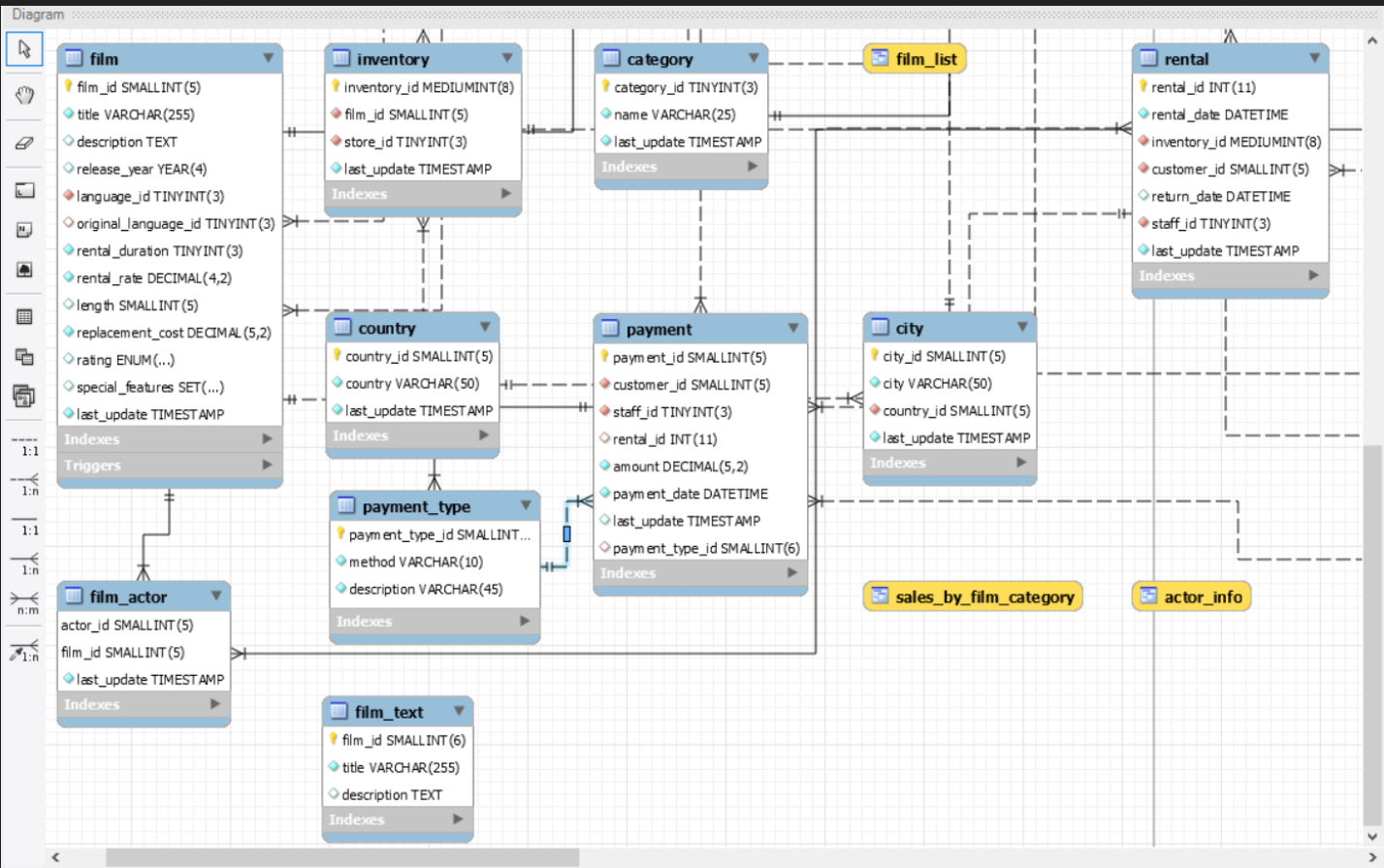
https://maps.googleapis.com/maps/api/geocode/json?latlng=40.714224,-73.961452

Formulate the URL expression in OpenRefine that would fetch the complete JSON results from this web service API (You do not need to invoke the API or download the results of the web service call)



**2. Relational Modeling**





**Table Name:** Payment

|  |  |  |  |
| --- | --- | --- | --- |
| Field ( Attributes) | Primary Key (Y/N) | Foreign Key (Y/N) | Related Table(s) and Cardinality between tables |
| payment\_id | Y | N |  |
| customer\_id | N | Y | customer (Many to one) |
| staff\_id | N | Y | staff(Many to one) |
| rental\_id | N | Y | rental (Many to one) |
| amount | N | N |  |
| payment\_date | N | N |  |
| last\_update | N | N |  |
| payment\_type\_id | N | Y | payment\_type (Many to one) |

**3. Relational Algebra**

a. List all payments greater than and equal to 2$ and less than equal to 7$

σ*amount >=2 and <=7*(Payment)

* 1. b. List all the movies with title and description that are rated PG-13

∏title, description (σ*rating = “PG-13”*(film))

* 1. c. Replace the word “film” with “movie” for all attributes and relations starting with the word “film”

*ρ*  movie (film)

* 1. d. List all customer names who have returned their rentals in the current month

1. ∏first\_name, last\_name (rental⋈ month (rental.return\_date)=month(current\_date())customer

**4. Normalization**

Anomalies:

**a. Insert Anomaly:**

If new physician joins but he has not yet assigned or handled a patient yet, then we will not be able to insert a new record for physician.

**b. Deletion Anomaly:**

Let’s say, we need to remove the appointment date 1/4/2016 for “Sheela Nupur”. On doing so, we will lose the corresponding details about Iam MacKay.

**c. Update Anomaly:**

Patient Jill Bell has 2 records and we need to be careful while updating the address value in all 3 places as any spacing or spelling difference will hamper data integrity.

**EXCEL WHILE WITH ALL THREE FORMS IS ATTACHED SEPARATELY**