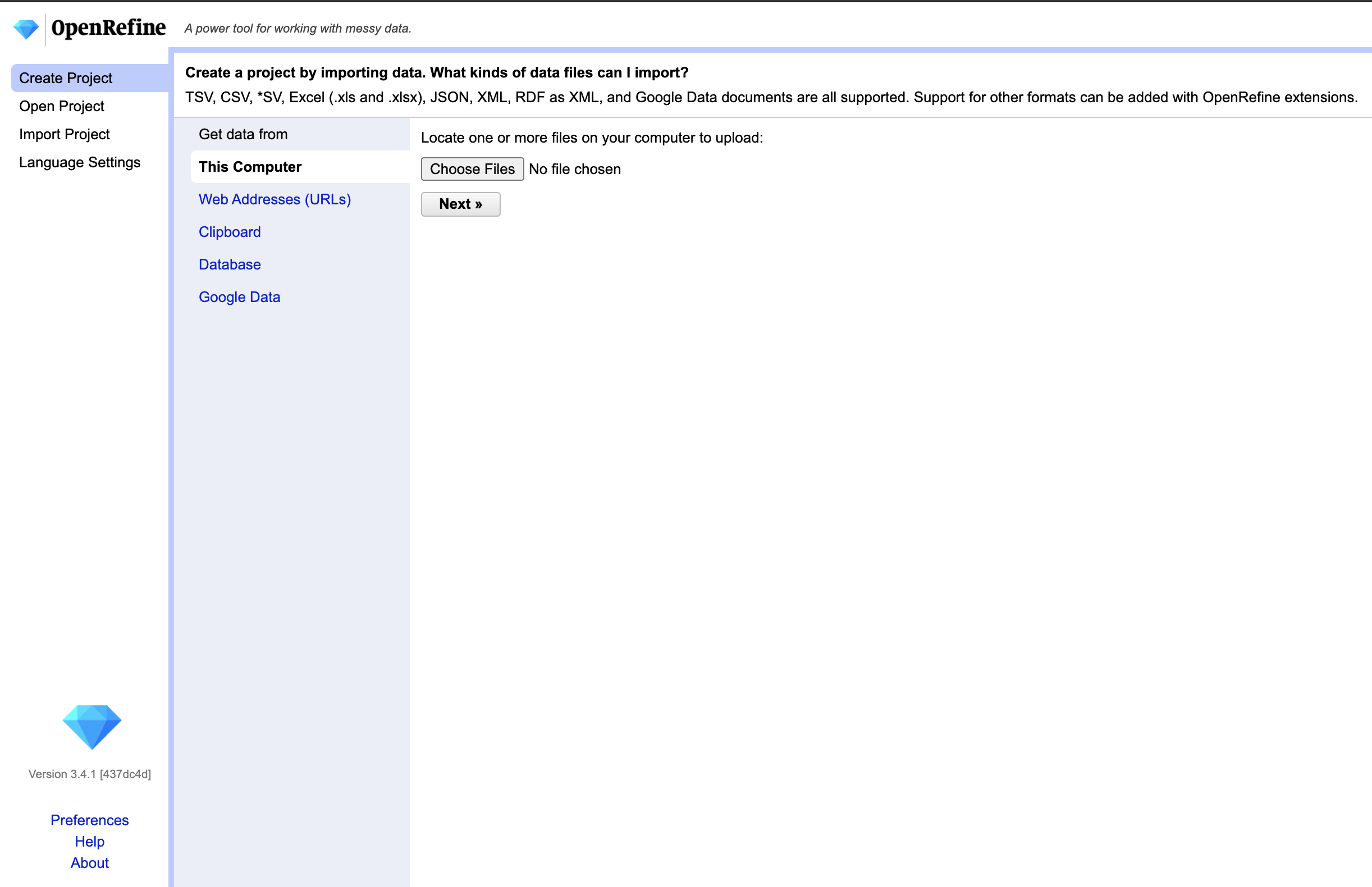
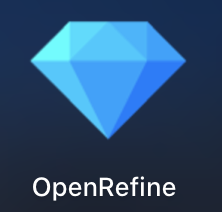
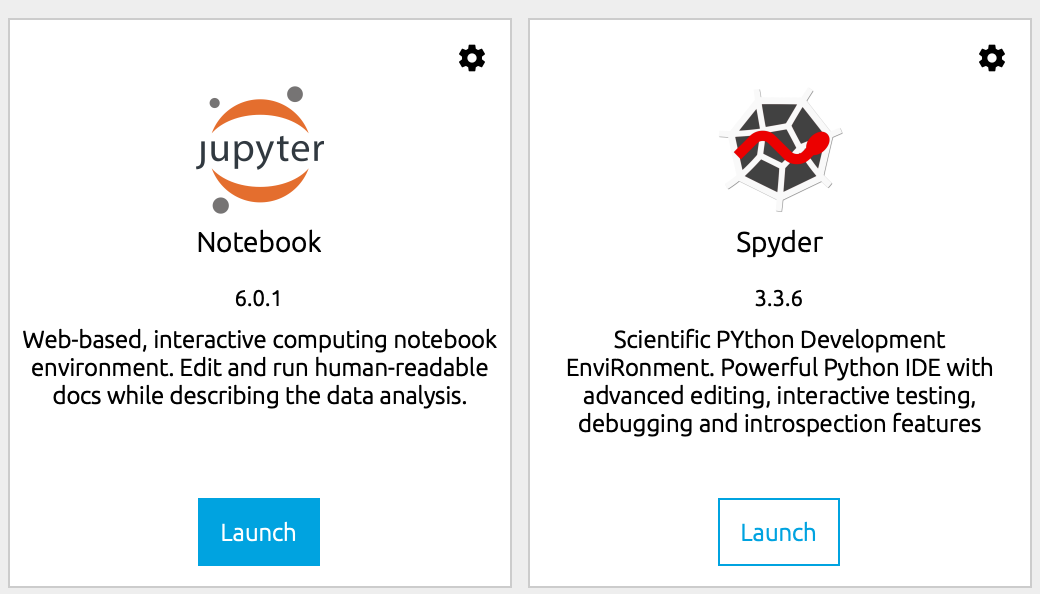
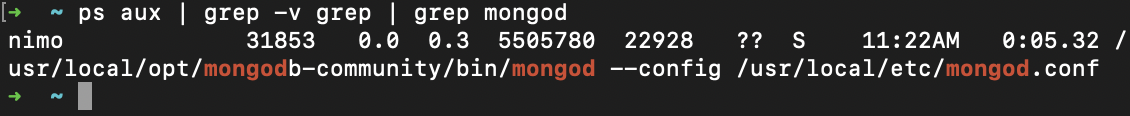
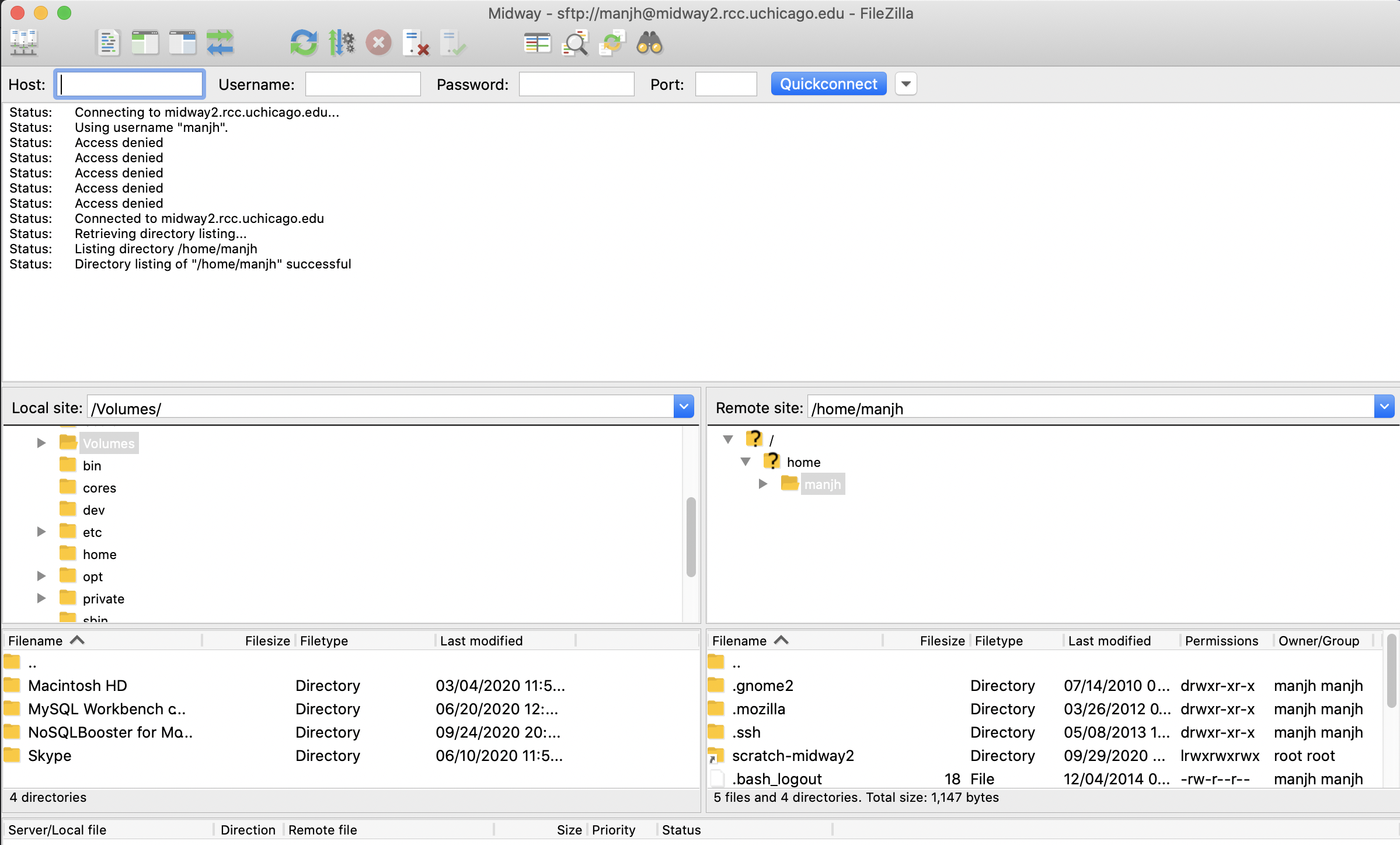
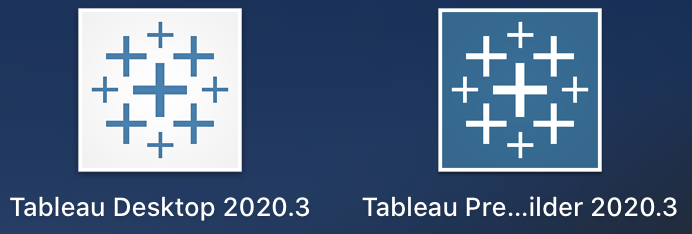
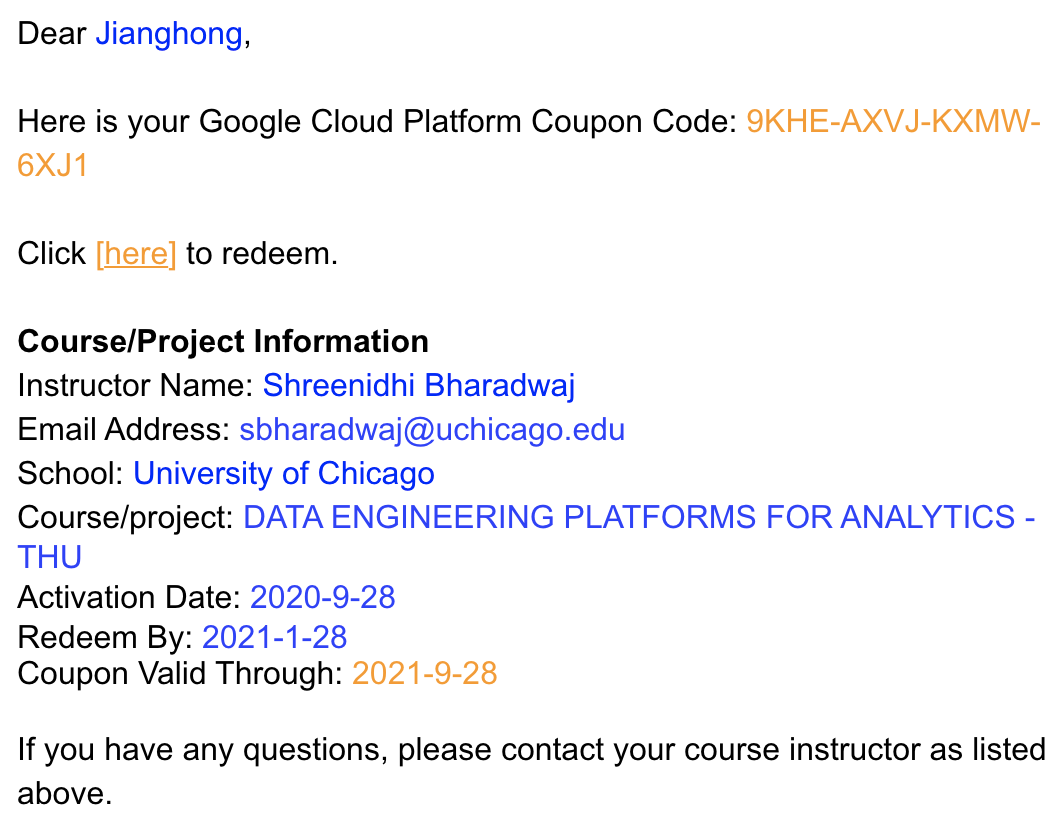
1. **Part A: Software installations, data extraction, cleaning & transformation**
   1. **Screenshot:**

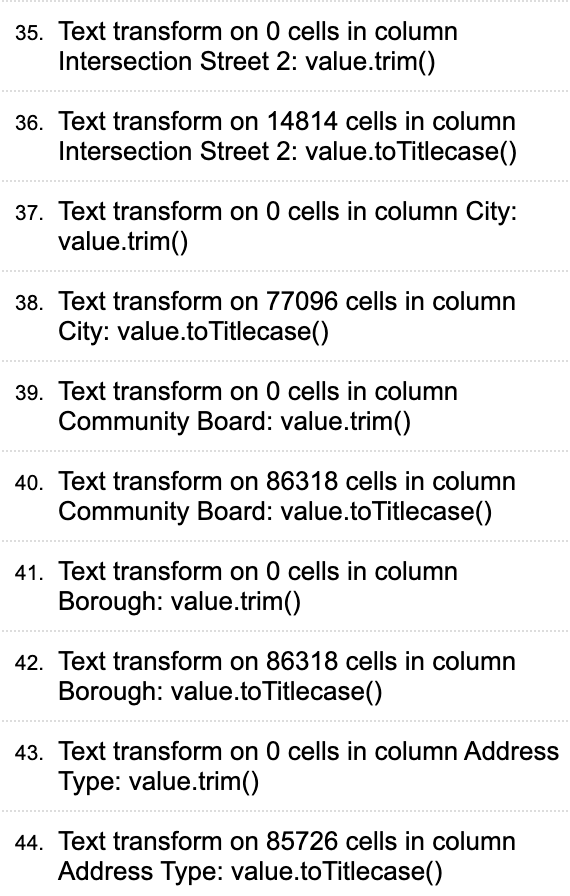
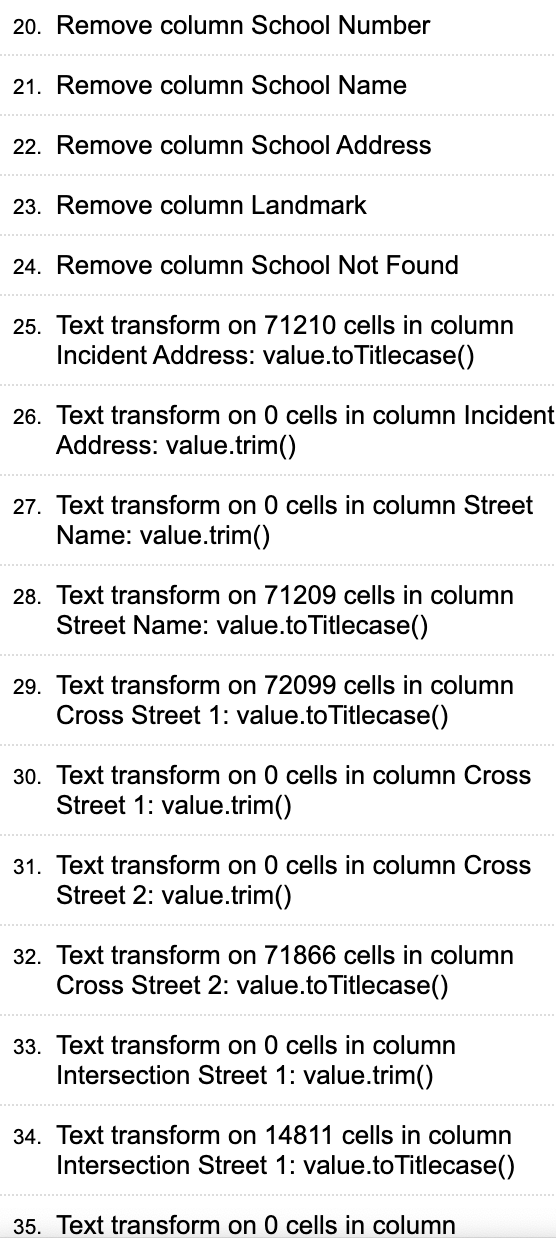
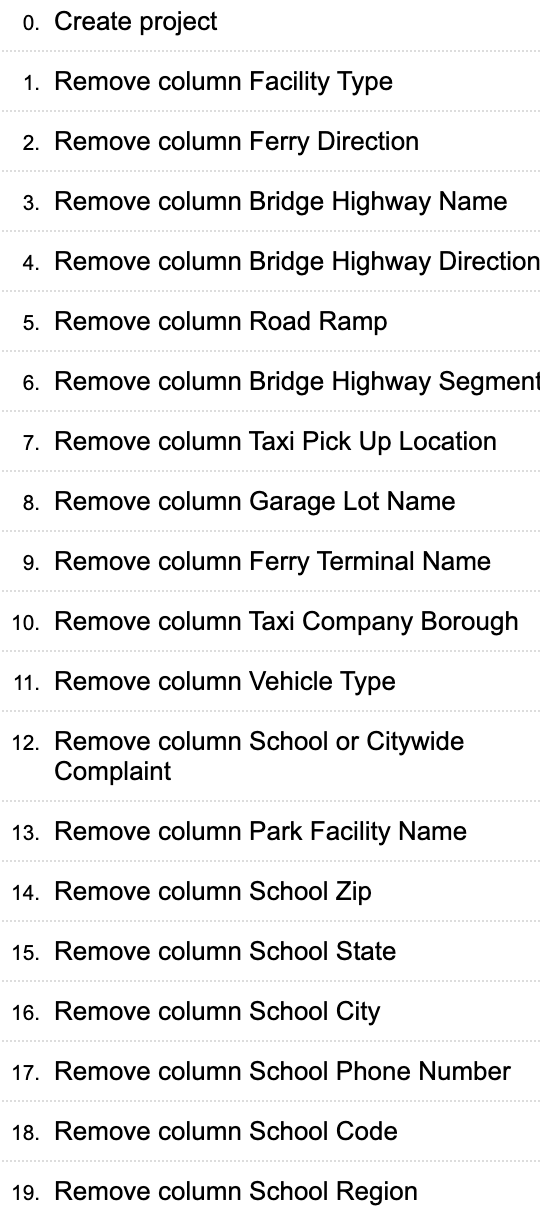


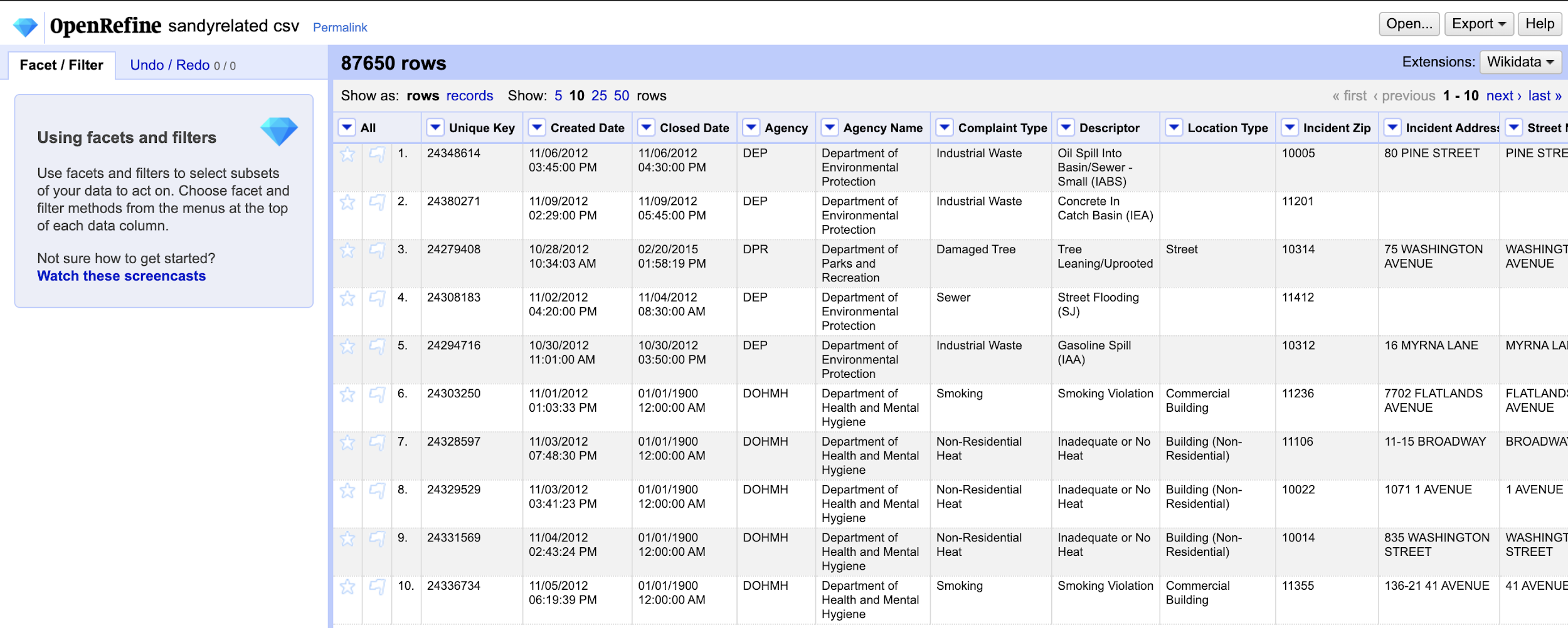


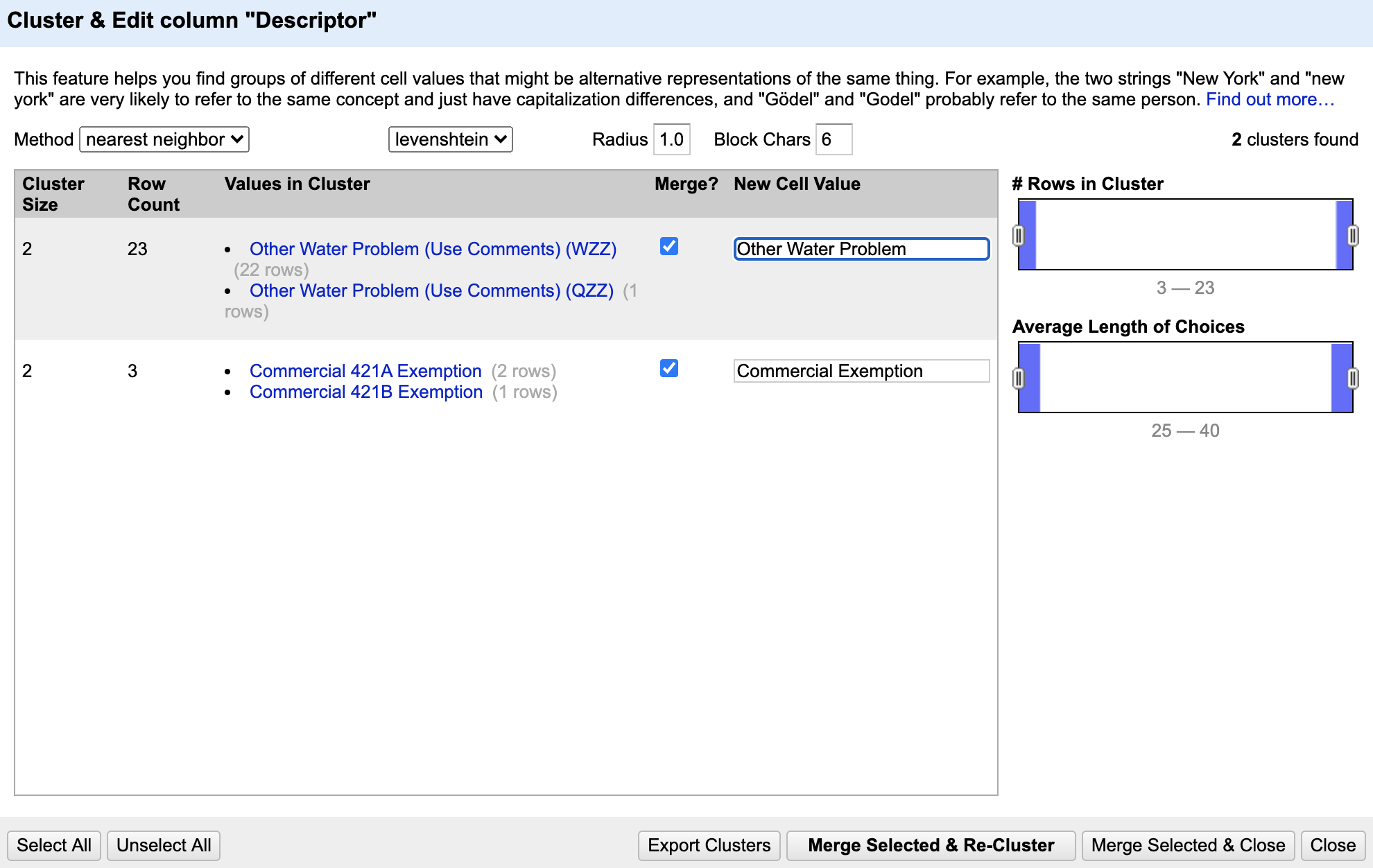


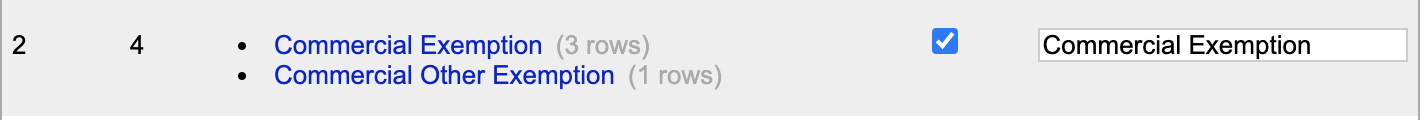


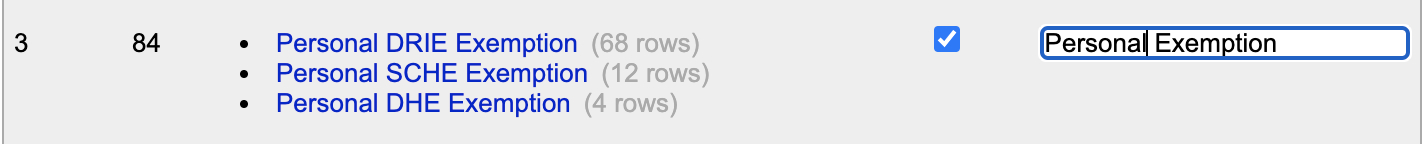
* 1. **OpenRefine screenshot**

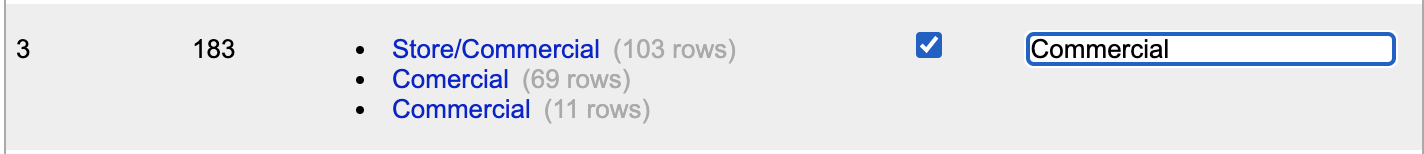


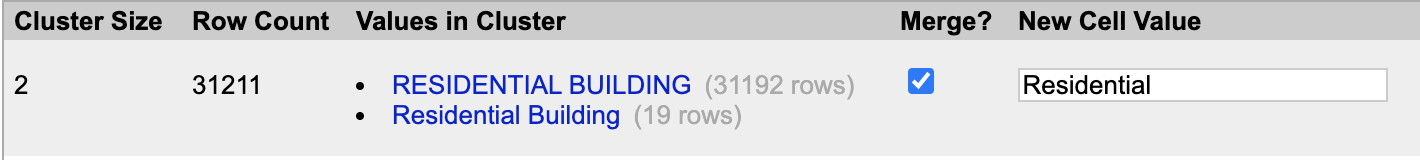


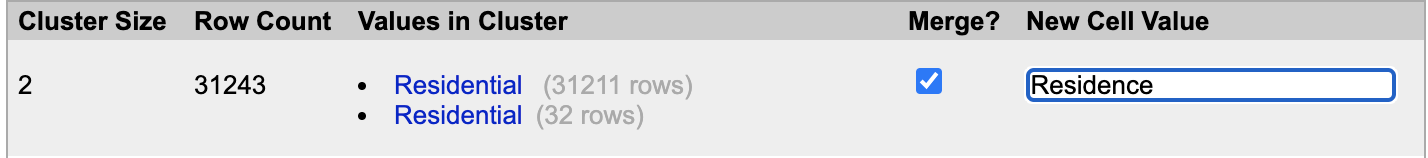


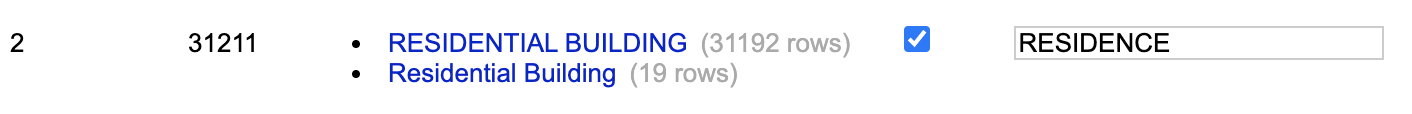


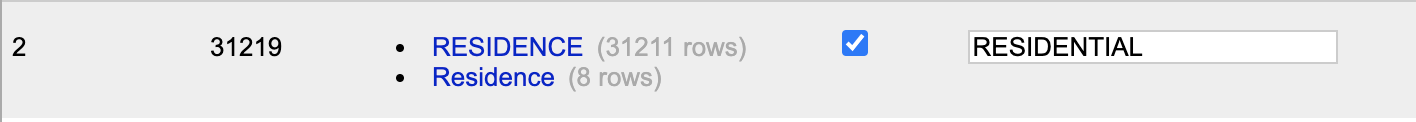


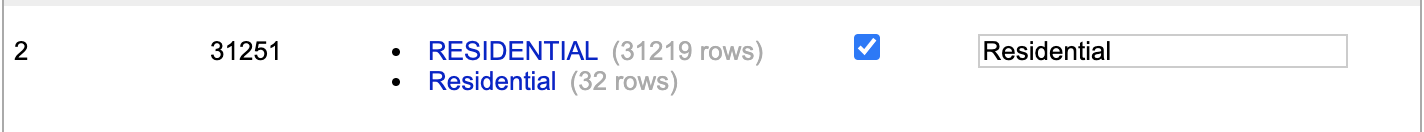


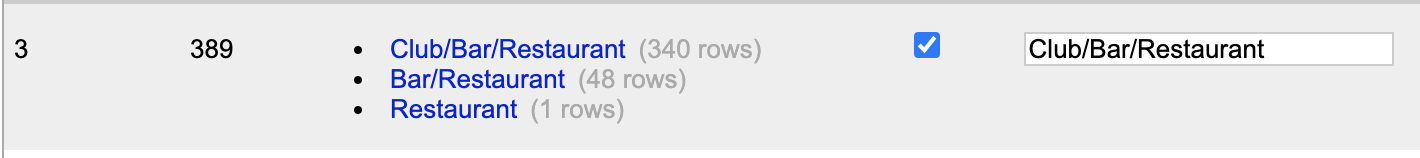


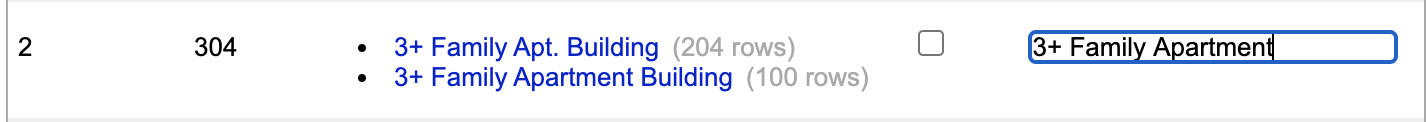


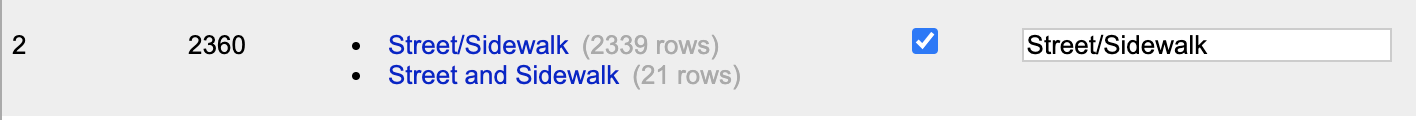
Three steps for cluster “residential”





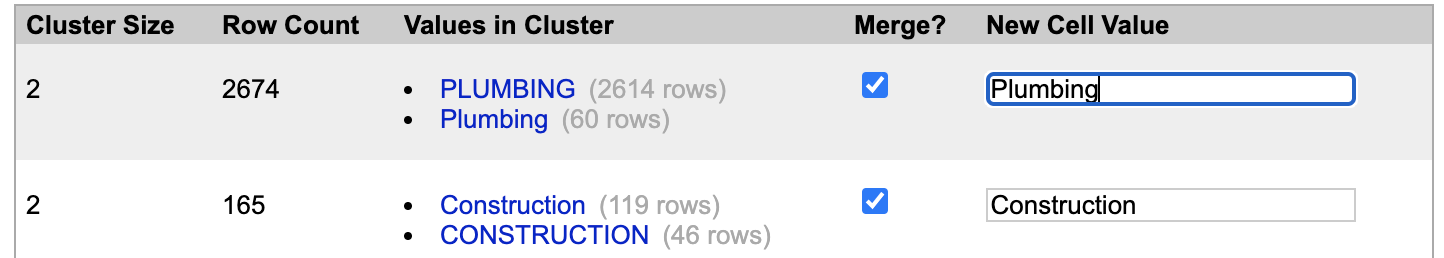








**Other clean up opportunities:**



* 1. **Memorandum**

**MEMORANDUM**

Date: 10/15/2020

To: Executive Management

From: Jianghong Man

Subject: Modern Data Management Requires a Balance Between Collecting Data and Connecting to Data

Nowadays, data-driven decision making is the gold standard in the world of business. Regardless of the field of business or preference for defining data (quantitative, qualitative), accurate data collection is essential to help business make constructive decisions. However, the need for integragation increasing amount of real-time (or "near-real-time") data makes the method of “collecting all the data in a central location prior to taking action” nearly impossible. Also, some regulations that protect personally information also make it difficult to collect data. In this case, business should consider balance connecting and collecting data for modern data management.

According to Gartner, the data management capabilities include data describing, data organizing, data integration, data sharing, data governing, and data implemention. Data analytics should think of these six data management processes as "common capabilities", which means to use and reuse those methods in a mixed way when dealing with data-related cases. Thus, the mixed usage of the method can help organizations achieve efficiency and effectiveness in morden data management.

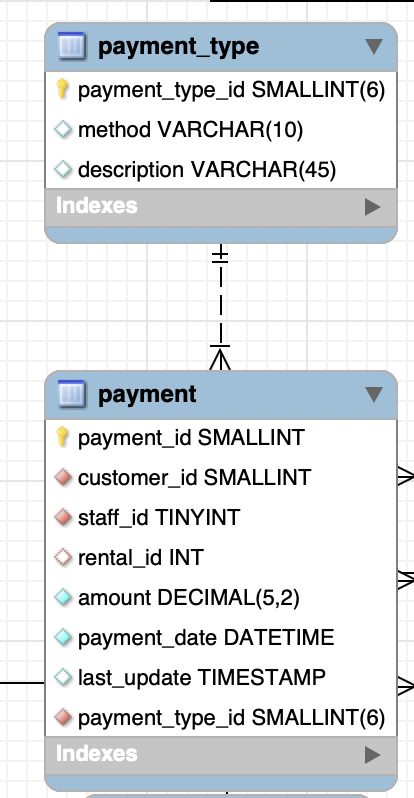
There are also two approaches of managing data. Mode 1 use to perform better when describing, organizing, integrating, sharing, governing and implementing data, while mode 2 have a more-relaxed approach as flexibility and ease of access to data are paramount. In this case, mode 1 use cases can often still rely on a more "collect" style of architecture while mode 2 use cases will align better with the "connect" style of data management. Within the organization, we should address both modes in conjunction thus to support a combination of standards-based solutions to business issues.

Also, data transparency and trust are two topics business should pay attention to. Business users should determine the method to manage data while considering to expose transparency and protect privacy. Last but not least, in order to meet the need of some nontechnical roles, it requires a strong focus on metadata management to support describing and organizing the data. Providing transparency about who is using the data and how the data is being used is what business organizer should consider.

All in all, combining data that belongs to the various levels of governance requires both connecting to and collecting data as appropriate. Business should balance the usage of both approach in order to achieve better decision making.

Please let me know if you have any questions.

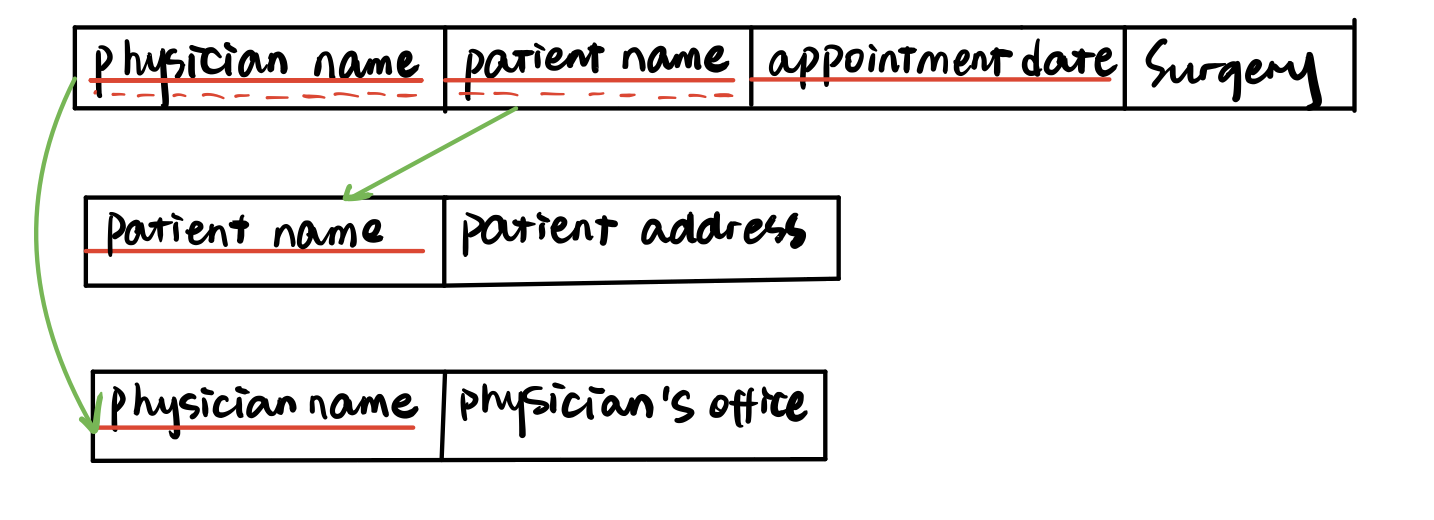
1. **Part B : Relational data model and design principles**
   1. **EER screen shot and payment table**

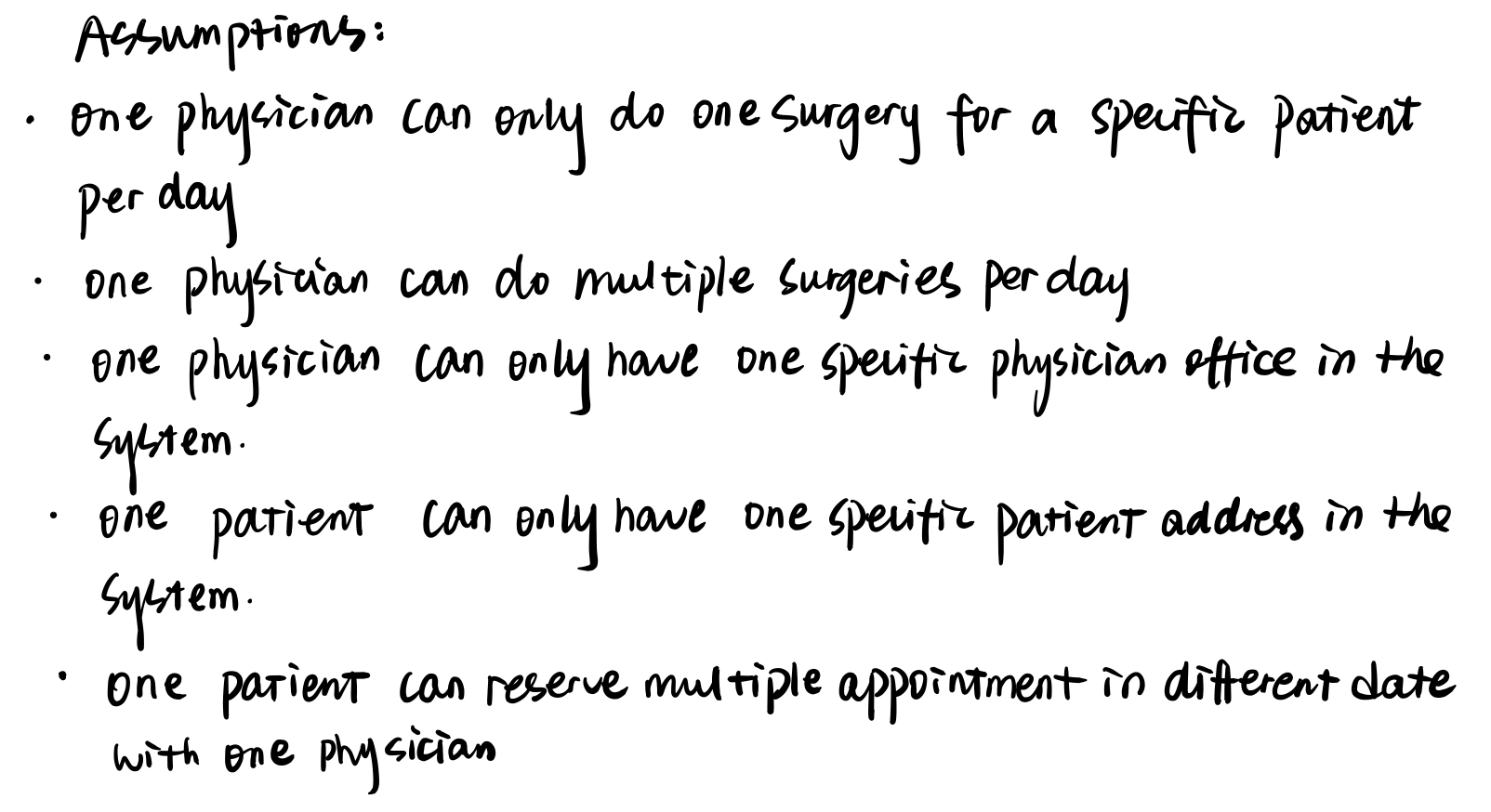


|  |  |  |  |
| --- | --- | --- | --- |
| Field (Attributes) | Primary Key (Y/N) | Foreign Key (Y/N) | Related Table(s)  (only enter this for foreign key fields) & Type of relationship between tables |
| payment\_id | Y | N |  |
| customer\_id | N | Y | Customer - 1:n |
| staff\_id | N | Y | Staff - 1:n |
| rental\_id | N | Y | Rental - 1:n |
| amount | N | N |  |
| payment\_date | N | N |  |
| last\_update | N | N |  |
| Payment\_type\_id | N | Y | Payment\_type - 1:n |

* 1. **Normalization**

1. Insertion anomalies: in the second row, when the physician, Helen Pearson take another patient different from that of the first row, we will add duplicate data such as “physician name” and “physician’s office” into the database.
2. Deletion anomalies: if we delete the third row, the physician name “Olga Kay”, we will lose the data which represent “Olga Kay”’s physician’s office, the patient, the surgery and corresponding appointment date this physician had in the database.
3. Modification anomalies: change the physician’s name for “Helen Pearson” will require us to change multiple rows related to this specific physician’s name.





* 1. **Model design (refer to the PPT)**

1. **Optional question (refer to the “YouTube\_API\_Scraping.ipynb” and “Web\_Scraping.ipynb”, and the output file “meet\_arnold.json”)**