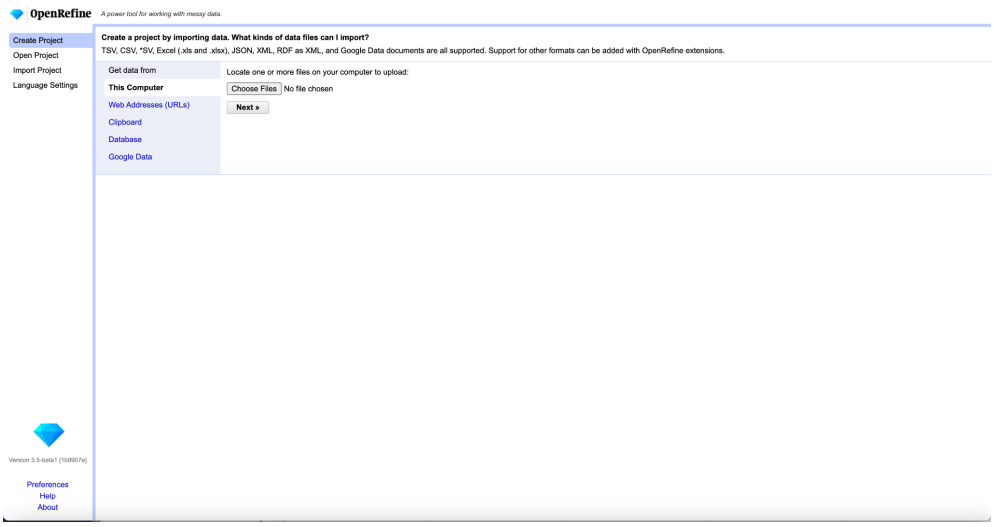
Data Engineering Platforms Assignment1

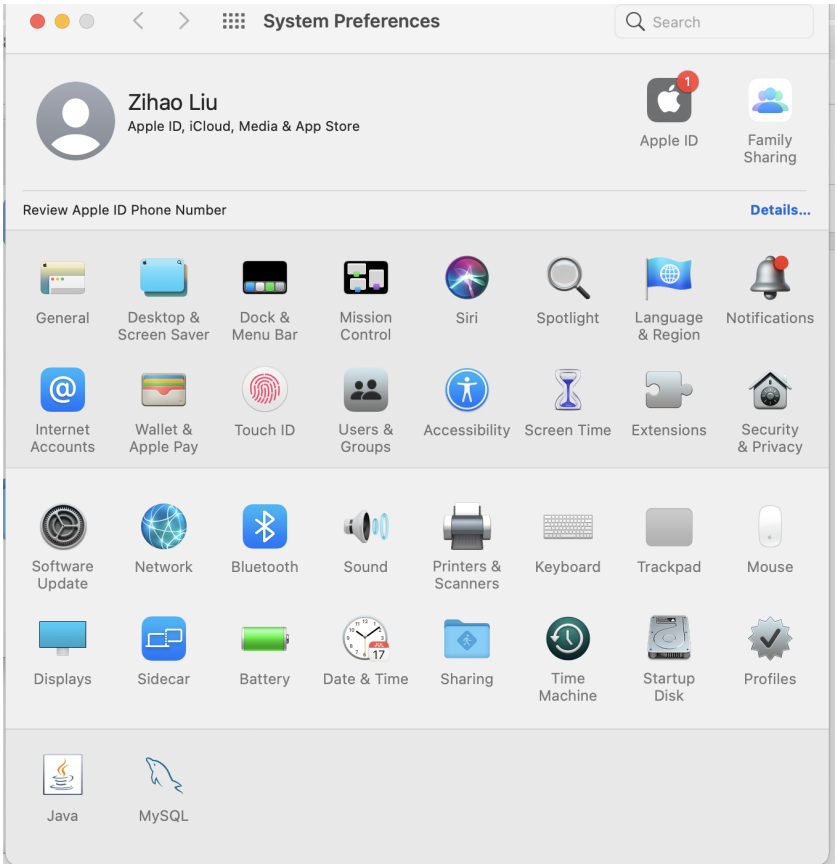
PartA

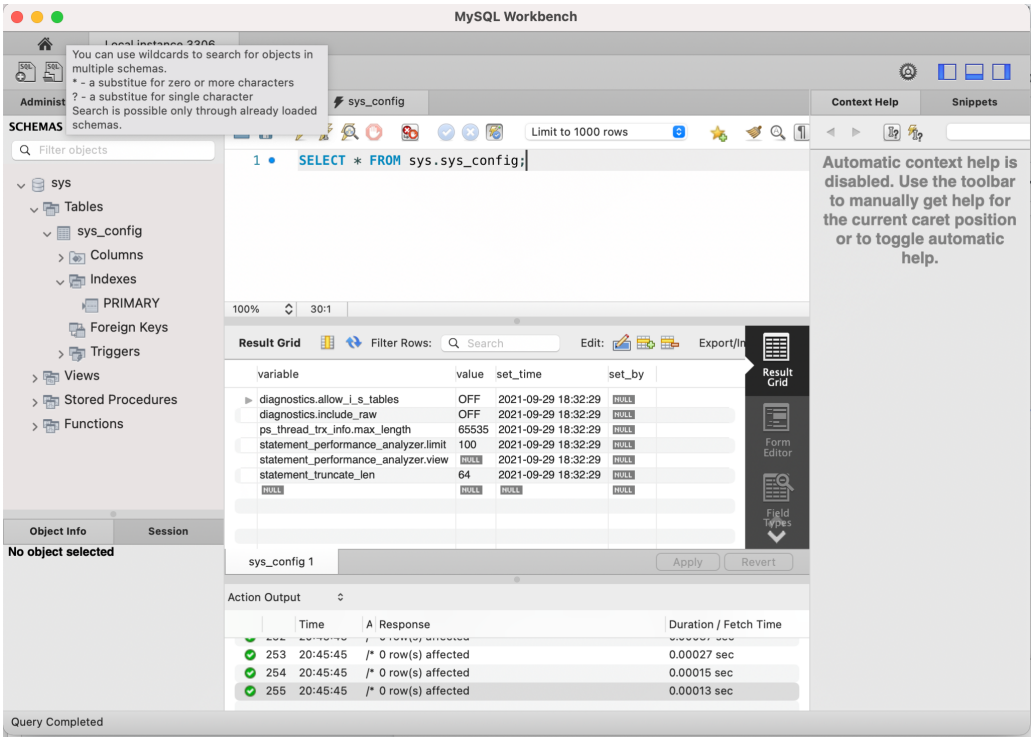
1

1) OpenRefine

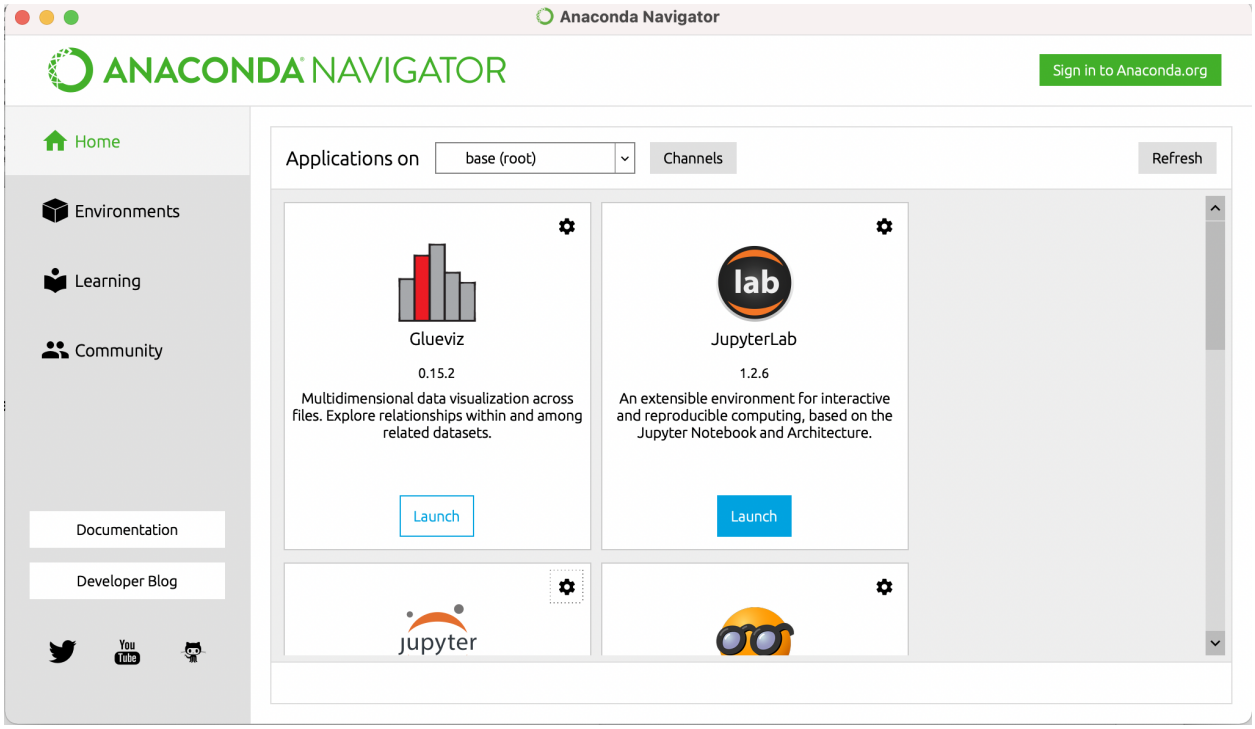


2) MySQL (server + workbench)

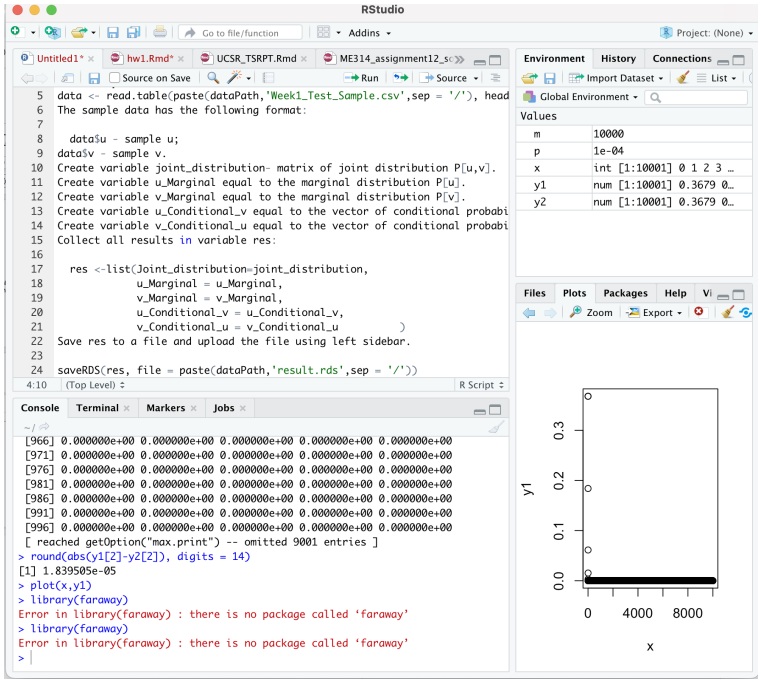

System Preferences
Search
Zihao Liu
Apple ID, iCloud, Media & App Store
Apple ID
Family Sharing
Review Apple ID Phone Number
Details...
General
Desktop & Screen Saver
Dock & Menu Bar
Mission Control
Siri
Spotlight
Language & Region
Notifications
Internet Accounts
Wallet & Apple Pay
Touch ID
Users & Groups
Accessibility
Screen Time
Extensions
Security & Privacy
Software Update
Network
Bluetooth
Sound
Printers & Scanners
Keyboard
Trackpad
Mouse
Displays
Sidecar
Battery
Date & Time
Sharing
Time Machine
Startup Disk
Profiles
Java
MySQL


MySQL Workbench
You can use wildcards to search for objects in multiple schemas.
* - a substitue for zero or more characters
? - a substitue for single character
Search is possible only through already loaded schemas.
SCHEMAS
Filter objects
sys
Tables
sys_config
Columns
Indexes
PRIMARY
Foreign Keys
Triggers
Views
Stored Procedures
Functions
Object Info
Session
No object selected
sys_config
Limit to 1000 rows
1 SELECT * FROM sys.sys_config;
Context Help
Snippets
Automatic context help is disabled. Use the toolbar to manually get help for the current caret position or to toggle automatic help.
100%
30:1
Result Grid
Filter Rows:
Search
Edit:
Export/Import:
variable
value
set_time
set_by
diagnostics.allow_i_s_tables OFF 2021-09-29 18:32:29
diagnostics.include_raw OFF 2021-09-29 18:32:29
ps_thread_trx_info.max_length 65535 2021-09-29 18:32:29
statement_performance_analyzer.limit 100 2021-09-29 18:32:29
statement_performance_analyzer.view 2021-09-29 18:32:29
statement_truncate_len 64 2021-09-29 18:32:29
Result Grid
Form Editor
Field Types
sys_config 1
Apply
Revert
Action Output
Time
Response
Duration / Fetch Time
253 20:45:45 /* 0 row(s) affected 0.00027 sec
254 20:45:45 /* 0 row(s) affected 0.00015 sec
255 20:45:45 /* 0 row(s) affected 0.00013 sec
Query Completed

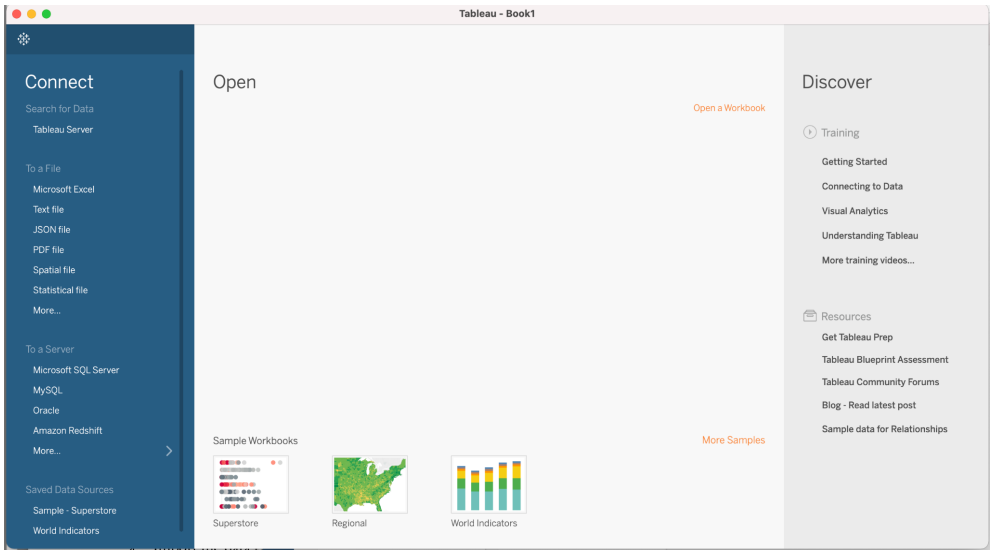
3) Anaconda (Open Data Science Platform : Python)



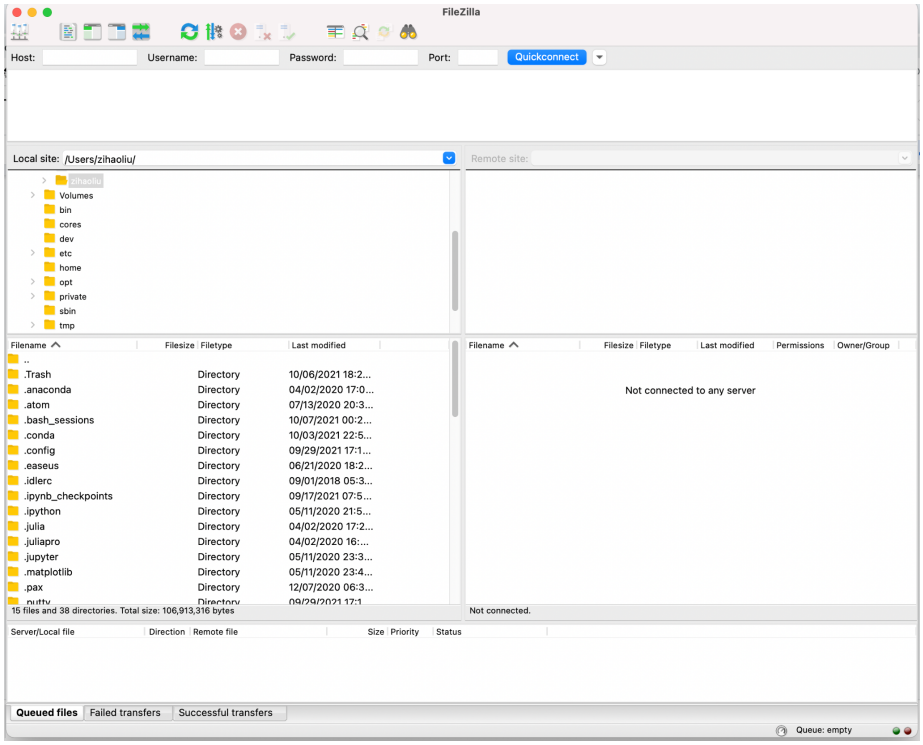
4) R-studio



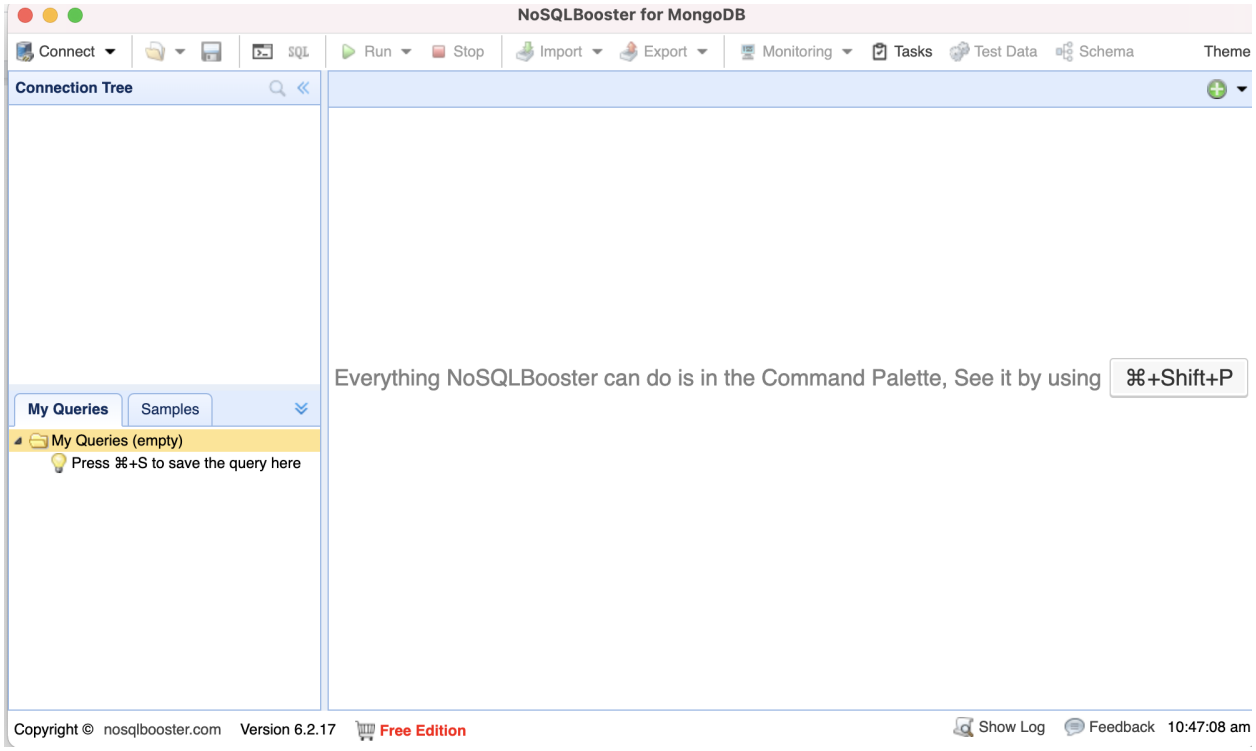
5) Tableau (https://www.tableau.com/academic/students)



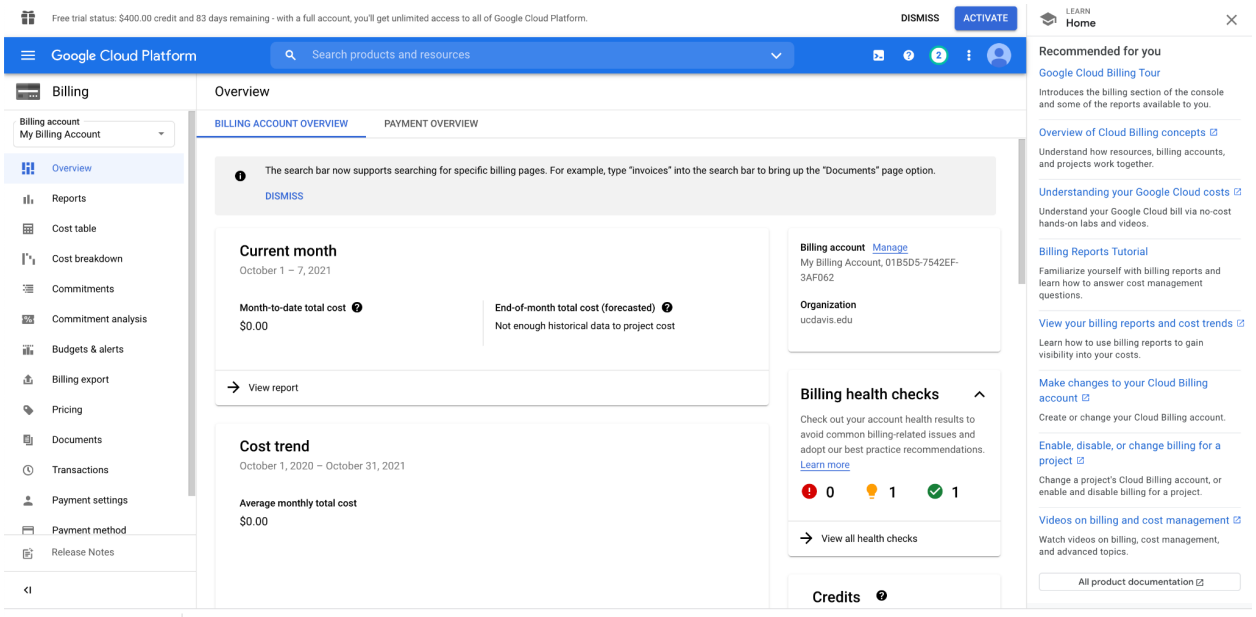
6) FileZilla Or CyberDuck


FileZilla
Host:
Username:
Password:
Port:
Quickconnect
Local site: /Users/zihaoliu/
Remote site:
zihaoliu
Volumes
bin
cores
dev
etc
home
opt
private
sbin
tmp
Filename
Filesize
Filetype
Last modified
Permissions
Owner/Group
..
.Trash Directory 10/06/2021 18:2...
.anaconda Directory 04/02/2020 17:0...
.atom Directory 07/13/2020 20:3...
.bash_sessions Directory 10/07/2021 00:2...
.conda Directory 10/03/2021 22:5...
.config Directory 09/29/2021 17:1...
.easeus Directory 06/21/2020 18:2...
.idlerc Directory 09/01/2018 05:3...
.ipynb_checkpoints Directory 09/17/2021 07:5...
.ipython Directory 05/11/2020 21:5...
.julia Directory 04/02/2020 17:2...
.juliapro Directory 04/02/2020 16:...
.jupyter Directory 05/11/2020 23:3...
.matplotlib Directory 05/11/2020 23:4...
.pax Directory 12/07/2020 06:3...
.putty Directory 09/29/2021 17:1
Not connected to any server
15 files and 38 directories. Total size: 106,913,316 bytes
Not connected.
Server/Local file
Direction
Remote file
Size
Priority
Status
Queued files
Failed transfers
Successful transfers
Queue: empty

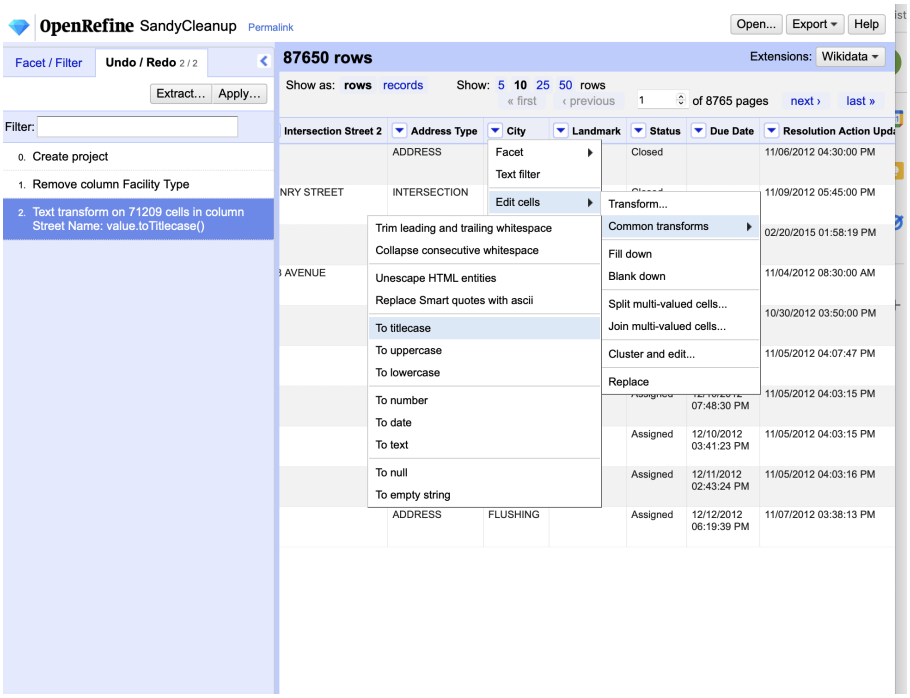
7) MongoDB

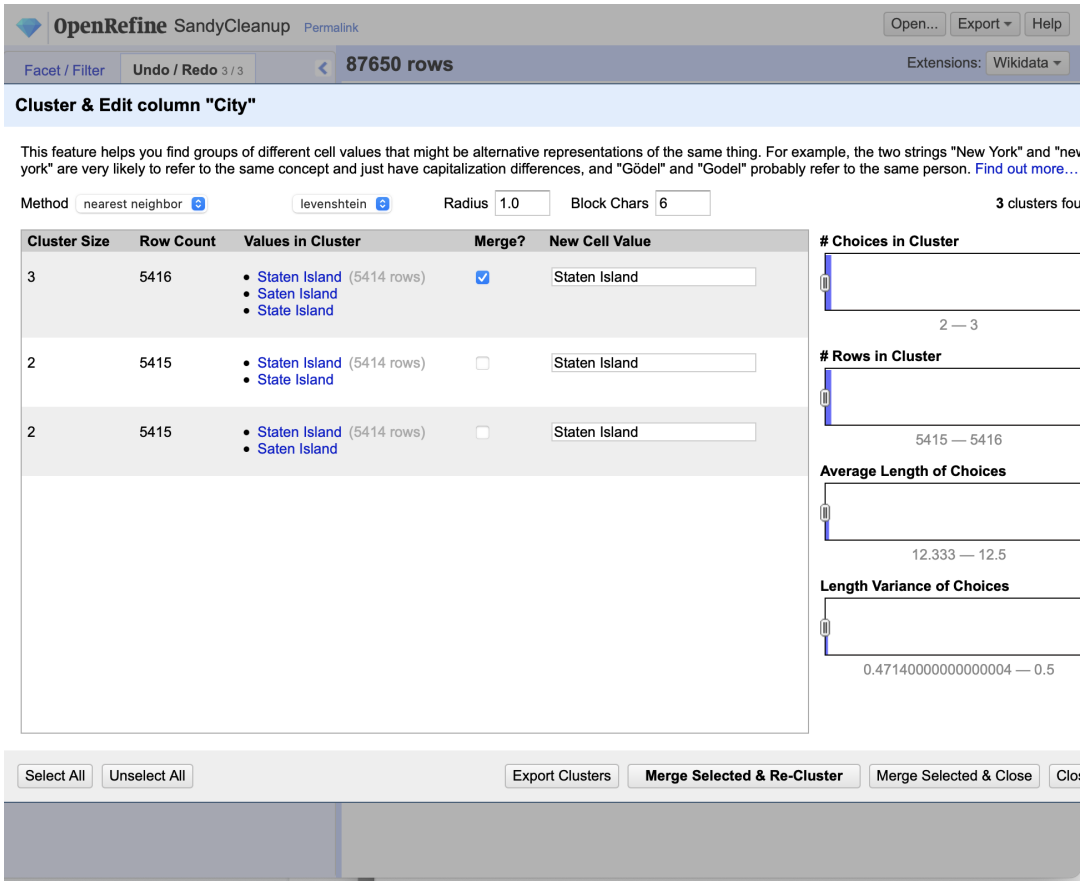


8) GCP (credits added to your account)

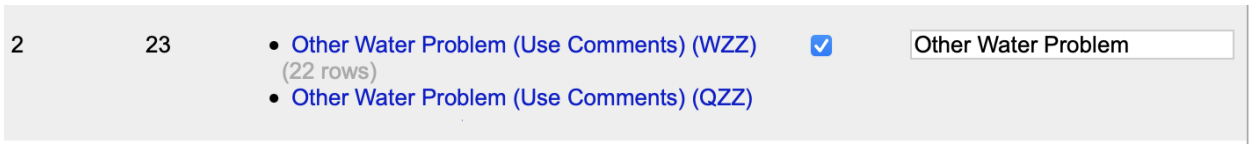


d.Convert City to title case, then Cluster and Merge the column

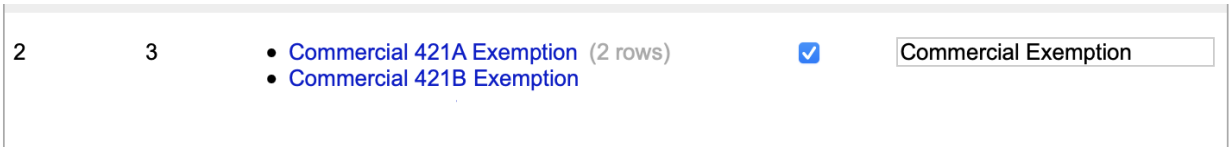

OpenRefine SandyCleanup Permalink
Open... Export Help
Facet / Filter
Undo / Redo 2 / 2
87650 rows
Extensions: Wikidata
Extract... Apply...
Show as: rows records
Show: 5 10 25 50 rows
« first ‹ previous 1 of 8765 pages next › last »
Filter:
0. Create project
1. Remove column Facility Type
2. Text transform on 71209 cells in column Street Name: value.toTitlecase()
Intersection Street 2
Address Type
City
Landmark
Status
Due Date
Resolution Action Upd
Facet
Text filter
Edit cells
Transform...
Common transforms
Trim leading and trailing whitespace
Collapse consecutive whitespace
Fill down
Blank down
Unescape HTML entities
Replace Smart quotes with ascii
Split multi-valued cells...
Join multi-valued cells...
To titlecase
To uppercase
To lowercase
Cluster and edit...
Replace
To number
To date
To text
To null
To empty string


OpenRefine SandyCleanup Permalink
Open... Export Help
Facet / Filter
Undo / Redo 3 / 3
87650 rows
Extensions: Wikidata
Cluster & Edit column "City"
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "nev
york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more…
Method nearest neighbor levenshtein Radius 1.0 Block Chars 6
3 clusters fou
Cluster Size
Row Count
Values in Cluster
Merge?
New Cell Value
3 5416 Staten Island (5414 rows) Saten Island State Island Staten Island
2 5415 Staten Island (5414 rows) State Island Staten Island
2 5415 Staten Island (5414 rows) Saten Island Staten Island
Choices in Cluster
2 — 3
Rows in Cluster
5415 — 5416
Average Length of Choices
12.333 — 12.5
Length Variance of Choices
0.47140000000000004 — 0.5
Select All
Unselect All
Export Clusters
Merge Selected & Re-Cluster
Merge Selected & Close

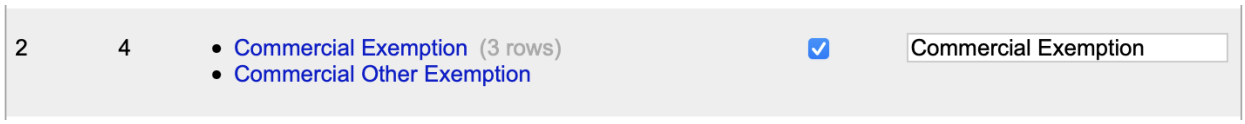
e.Clean up the Descriptor Column - Cluster and Merge the following text categories:
1. "Other Water problem(WZZ)", "Other Water problem(QZZ)" as "Other Water Problem”



2."Commercial 421 A/B Exemptions" as "Commercial Exemption"

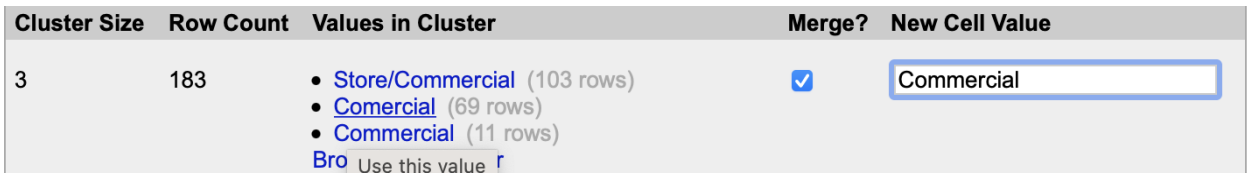


3."Commercial Exemption" "Commercial Other Exemption" as "Commercial Exemption"

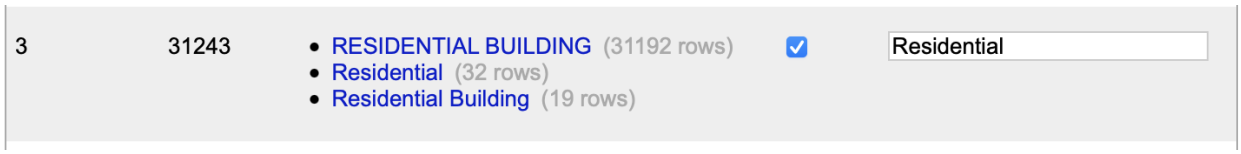


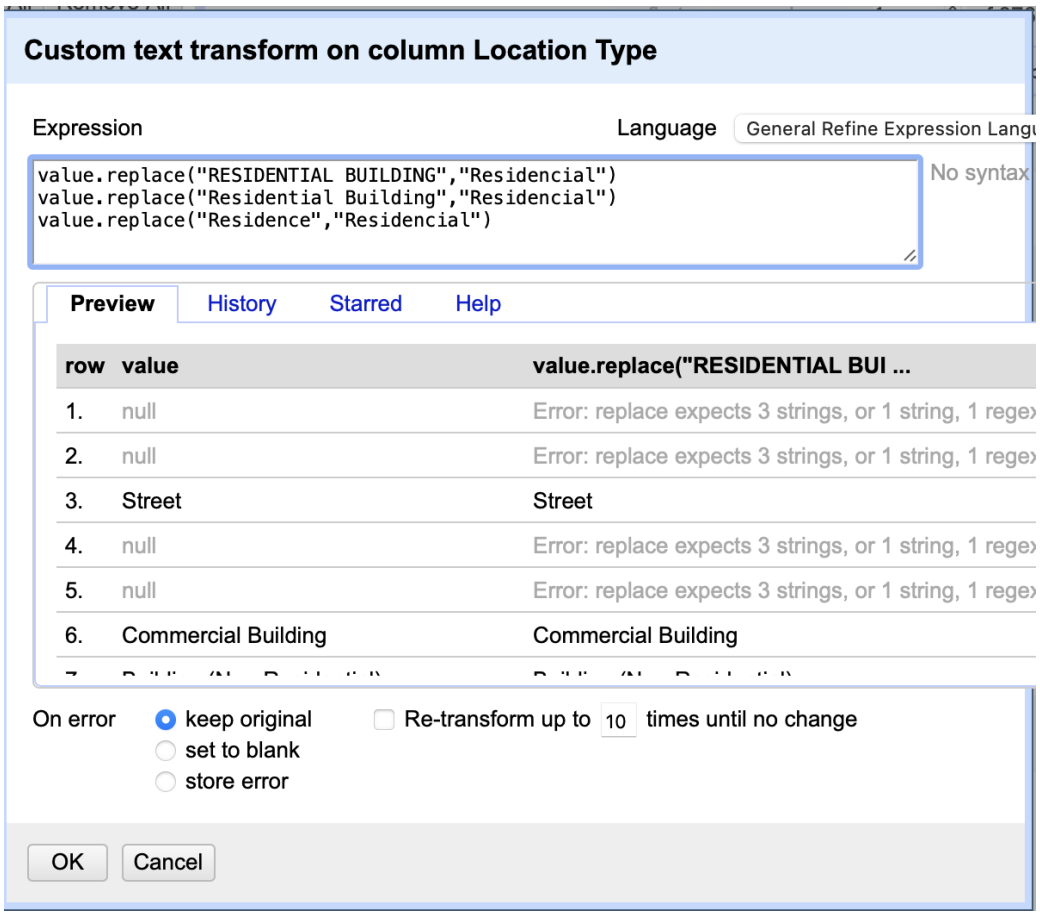
f.Clean up the Location Type - Cluster and Merge the following text categories:

1."Comercial", "Commercial", "Store/Commercial" as "Commercial"



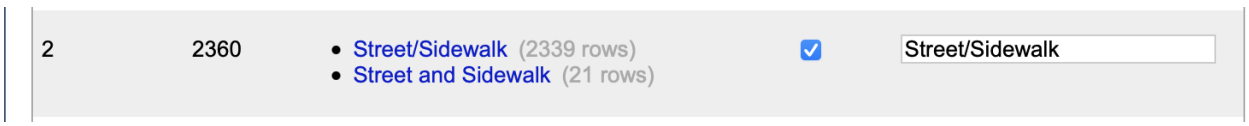
2."RESIDENTIAL BUILDING", "Residential Building", "Residence" as "Residential"





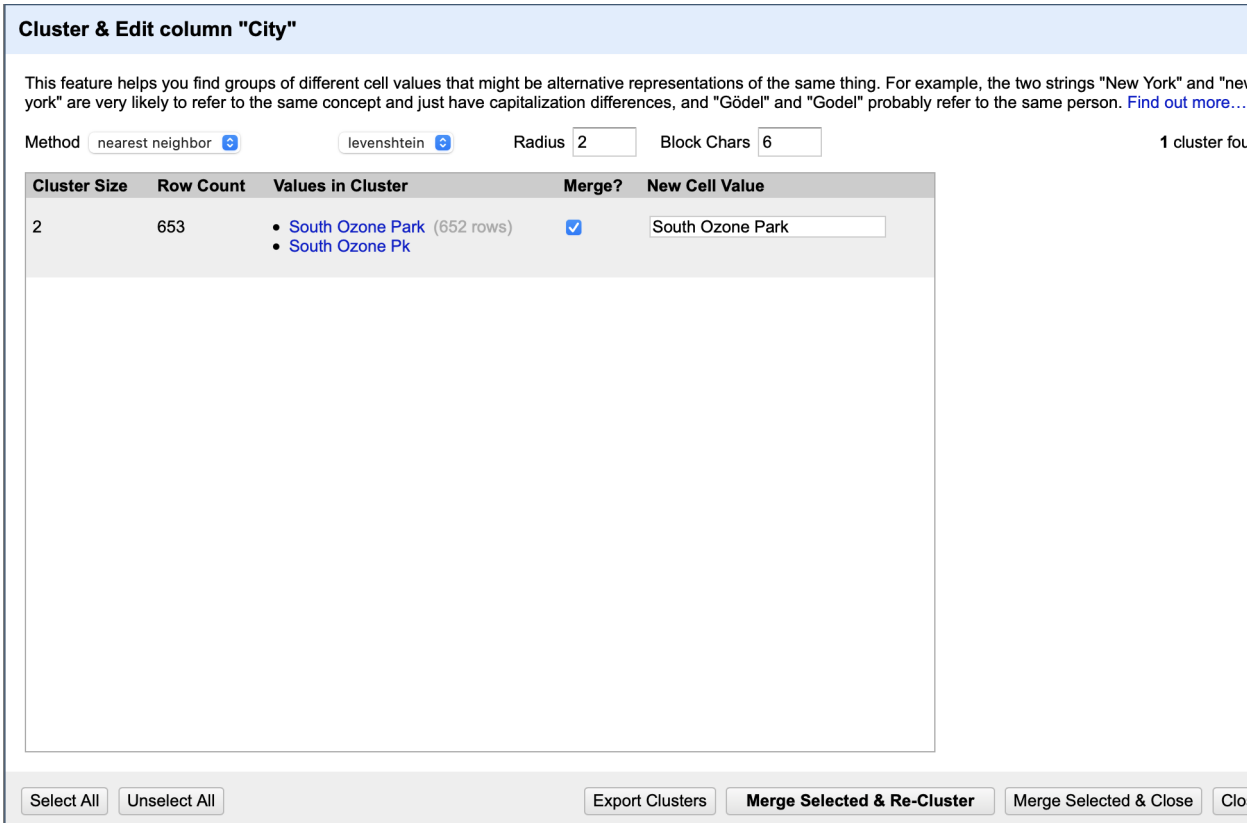
Since nearest neighbor cluster's setting does not exactly capture these three, I use transform instead

3."Street/Sidewalk", "Street and Sidewalk" as "Street/Sidewalk"

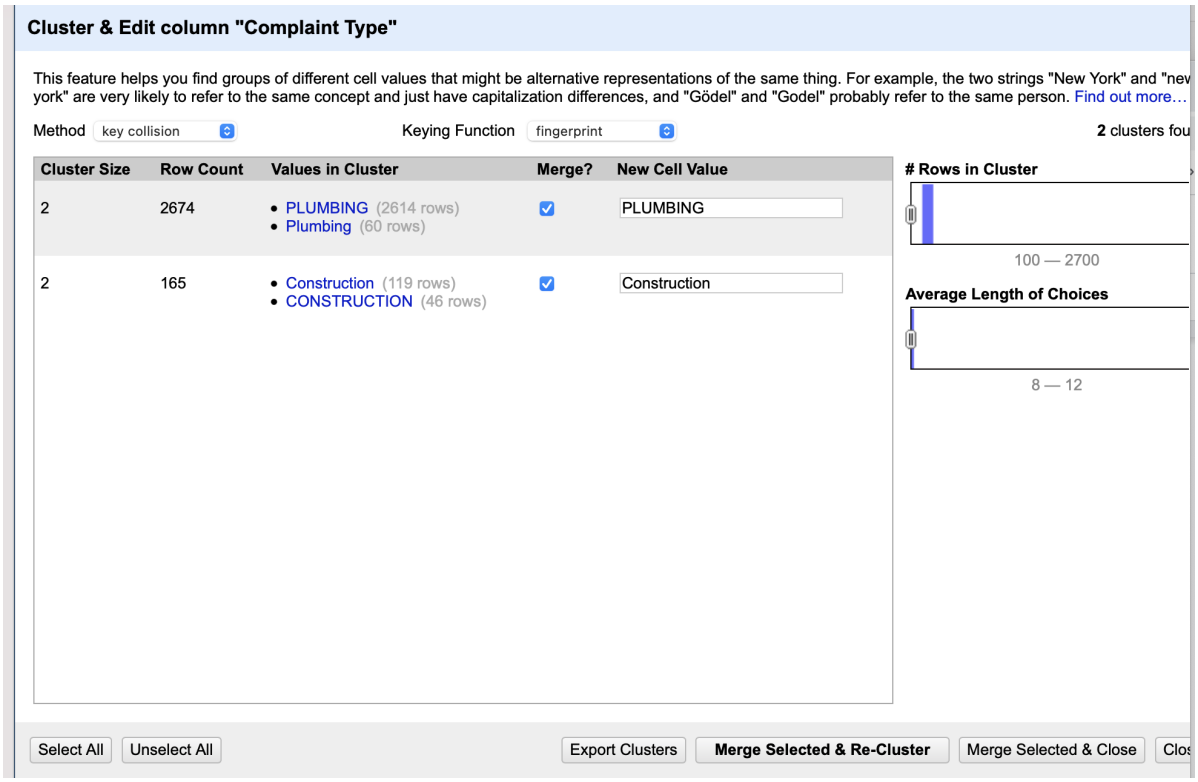


g.Look for at least two other clean up opportunities and execute using OpenRefine

(a) In City column,Merge this city since it’s the same place.



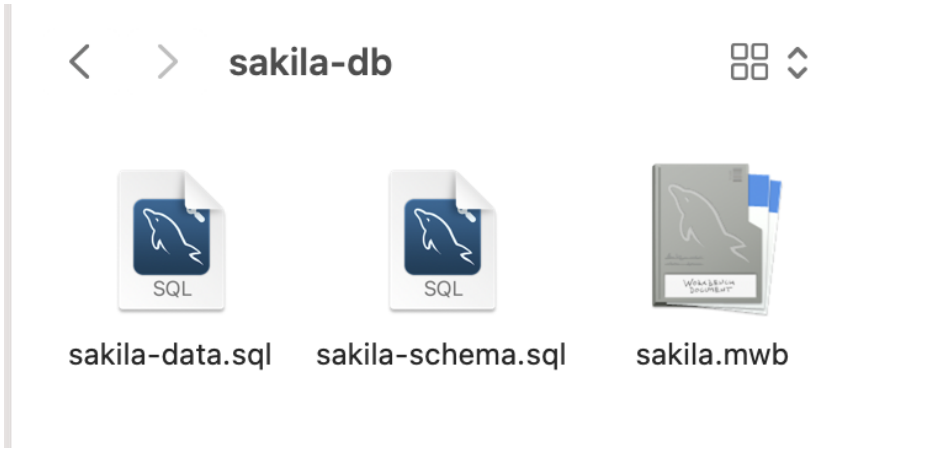
(b) In column Complaint Type, they describe the same thing but use capitalized letters.



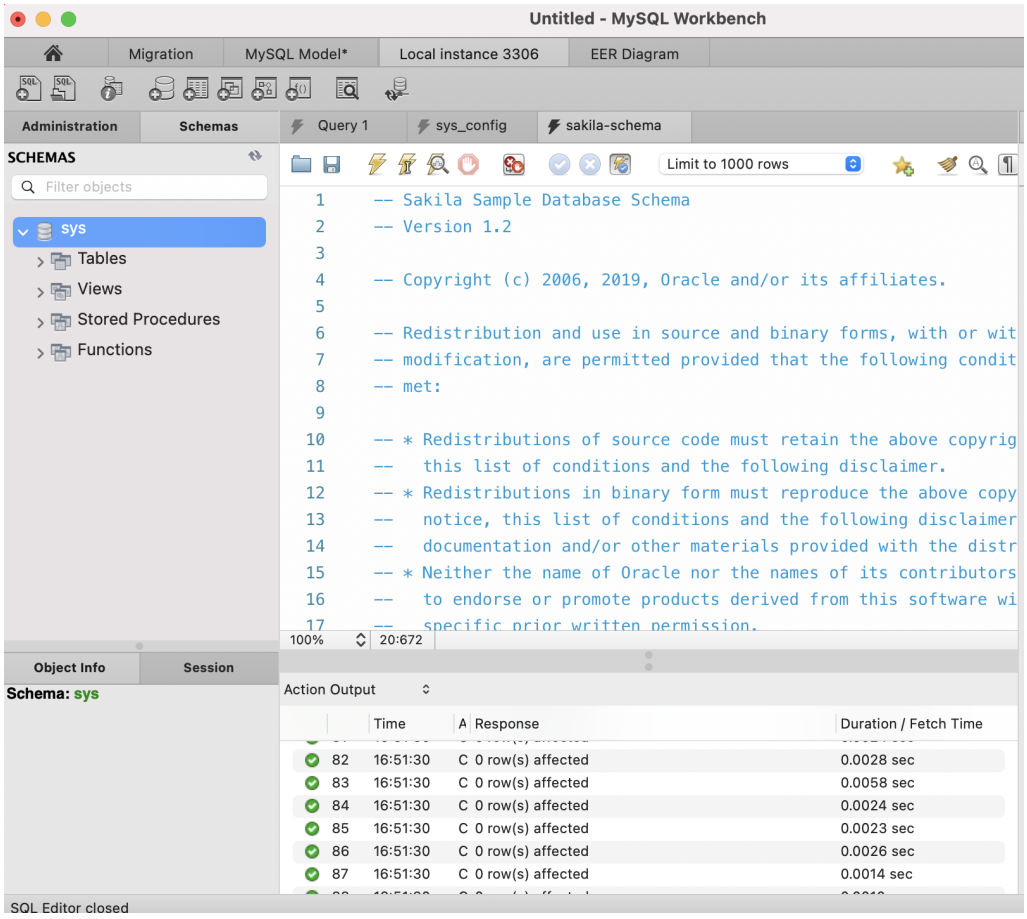
Part B : Relational data model and design principles

1.Relational Data Modeling

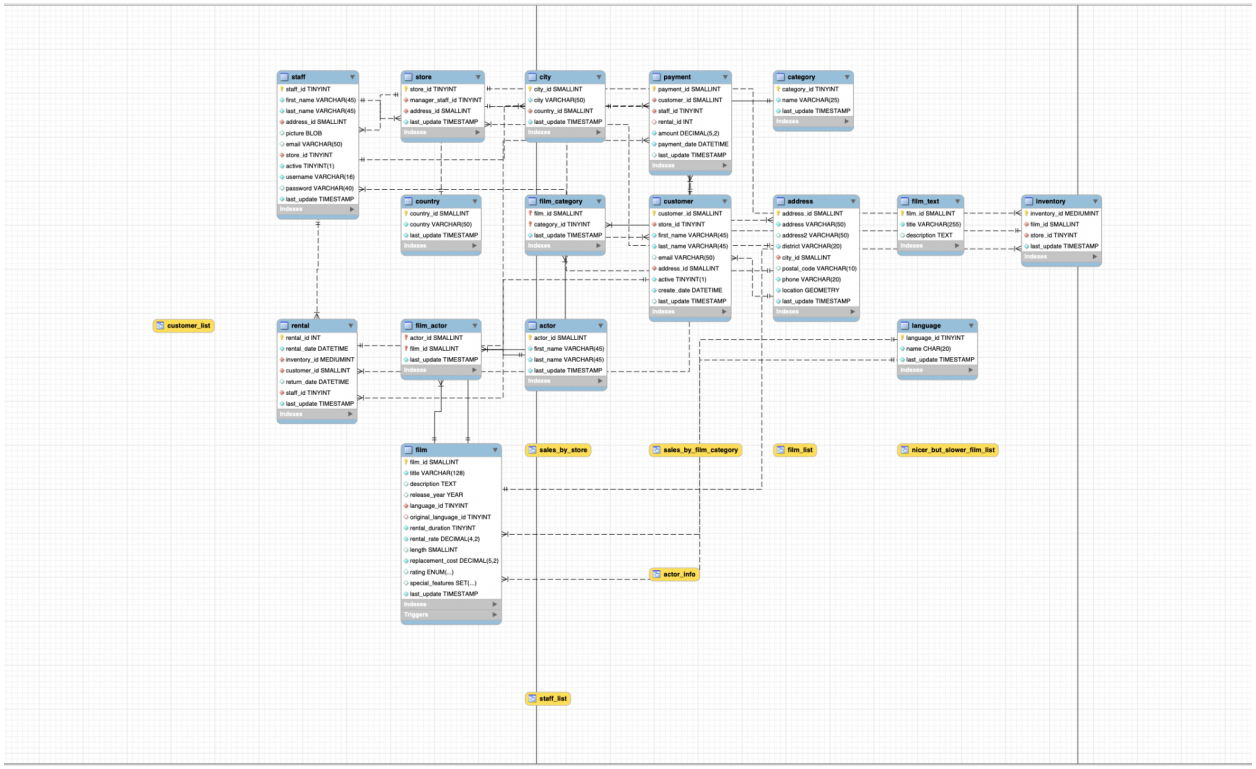
a.Download Sakila dataset and unzip sakila-db.zip file from the URL listed above.



b.Execute sakila-schema.sql file in the SQL workbench



c.Reverse Engineer the database and generate the EER diagram using the MySQL workbench.



d.Add a new lookup table: payment_type (1 to Many relationship with payment entity) with thefollowing attributes:

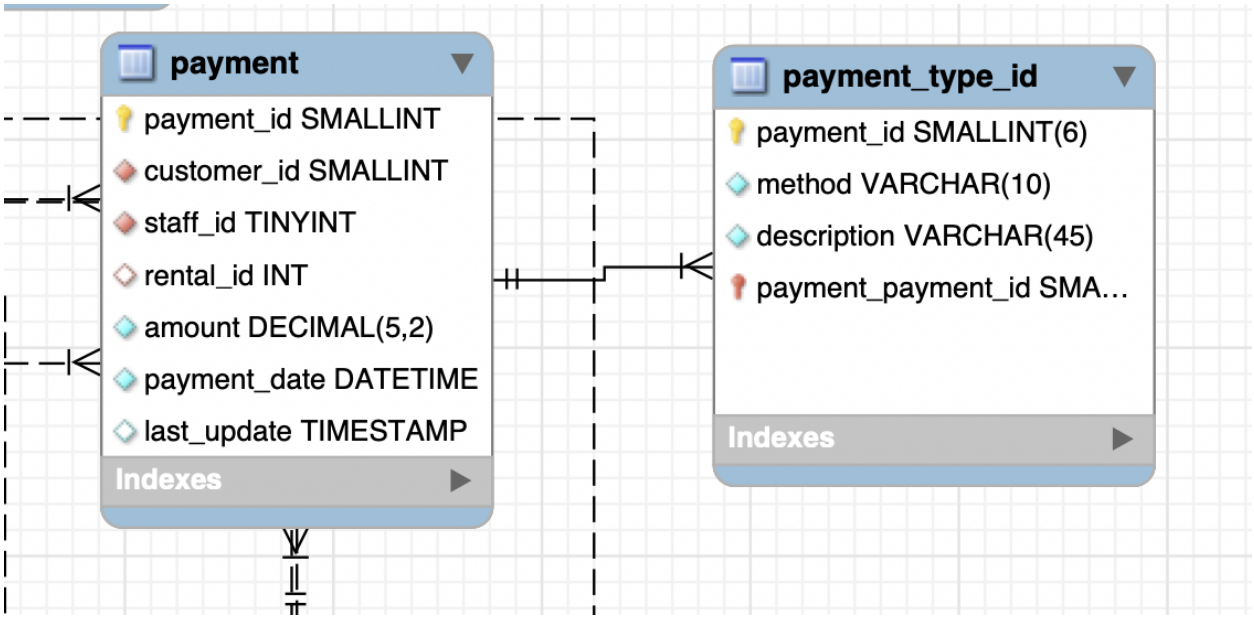
➢ payment_type_id (Primary Key) : SMALLINT(6)

➢ method - varchar (10)

➢ description – varchar (45)

Add the foreign key payment_type_id in the Payment entity with the following attributes:

➢ Payment_type_id (Foreign Key) : SMALLINT(6)



E

Field	Primary Key	Foreign Key	
payment_id	Y	N	
customer_id	N	N	
staff_id,rental_id	N	N	
amount	N	N	
payment_date	N	N	

last_update	N	N	
fk_payment_customer	N	Y	customer;1:n
fk_payment_rental	N	Y	rental;1:n
fk_payment_staff	N	Y	staff;1:n

2.Normalization : For the table below:
a.Provide examples of insertion, deletion, and modification anomalies.

Generally, the table is not well structured, unnormalized with redundant data.

Insertion anomalies:
In the table, there is no unique patient ID or doctor ID,so it’s impossible for us to create an appointment when a new patient has the same name as an earlier patient or a new doctor with the same name as an existing doctor. Because we cannot distinguish them. In other words, a certain attribute cannot be inserted without the presence of other attributes.we can't add a new patient unless we create an appointment.
Deletion anomalies:
If we want to delete a patient with multiple visits, many rows need to be deleted. Also, if wn want to delete a doctor, multiple rows need to be deleted for data integrity.Meanwhile, if we delete a doctor’s record, the patient’s detail related to him or her are also lost from the database.

Modification anomalies:
If we are interested in changing a column of a particular patient or doctor. For example, when we change a doctor’s name, we need to change all rows related to this doctor to avoid inconsistency.

b.Normalize this table to 3NF and list any assumptions.

Physician ID	First Name	Last Name	Physician's Offic	City
1	Helen	Pearson	Chicago Ave	Chicago
2	Olga	Kay	Clark Street	Chicago
3	Robert	Smith	Madison Street	Chicago
4	Wei	Jing	Adams Street	Chicago
5	Jay	Patel	Monroe Street	Chicago

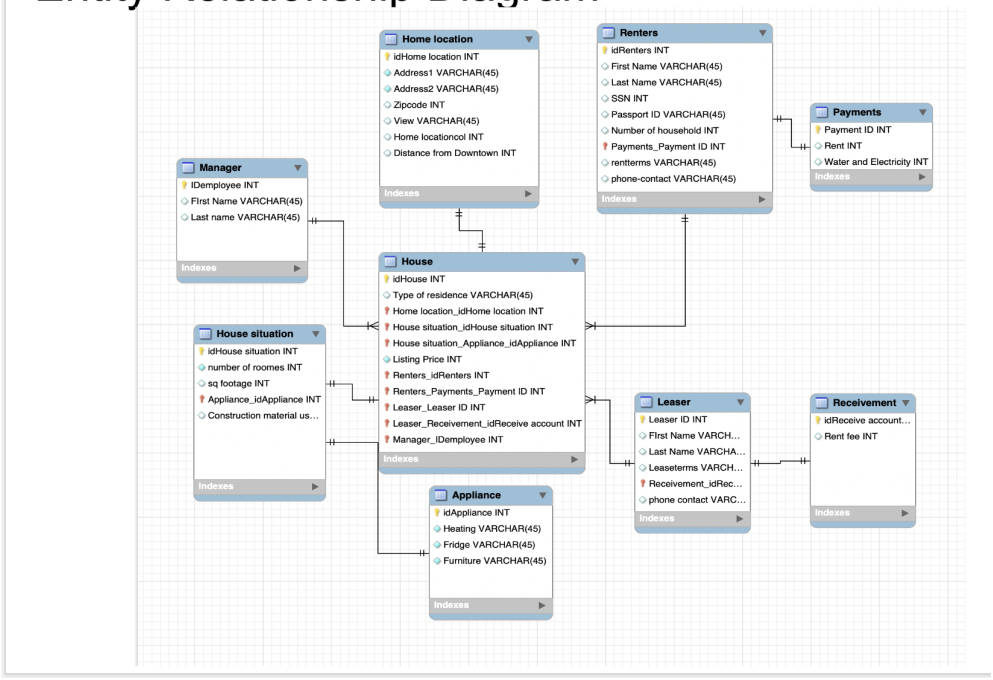
Patient ID	Patient Name	Last Name	Patient Address	City
1	Joe	Korn	Randolph Street	Chicago
2	Gillian	White	Illinois Street	Chicago
3	Jill	Bell	Huron Street	Chicago
4	Mike	Li	Lake Street	Chicago
5	Ian	Mackay	Dearborn Street	Chicago
6	Sheela	Nupur	Monroe Street	Chicago

Appointment Da	Surgery	Physician ID	Patient ID
3/7/2017	Tendon Repair	1	1
3/22/2017	Skin Graft	1	2
6/13/2016	Sentinel Node B	2	1
6/13/2017	Tendon Repair	3	3
6/14/2017	Skin Graft	3	3
6/13/2017	Knee Arthroscop	4	4
8/15/2017	Sentinel Node B	5	2
1/4/2016	Hepatic Resectic	5	5
1/5/2018	Liver Transplant	5	5
1/4/2016	Knee Arthroscop	1	6
2/12/2016	Skin Graft	4	4
4/15/2018	Skin Graft	4	4

Assumptions:The appointment date and surgery are fixed. Patients names and Doctors' names are unique,so a patient number(primary key) can be assigned to the patient and physician Id can be assigned to physician names as primary key.

3. Data Modeling:

Entity Relationship Diagram



Tables/relationships/Cardinality/Datatypes

Entities are: house, house location,manager,house situation, appliance,renters,leaser,payments,recivements

When creating Cardinailty, most relationships are 1 to. I make the assumptions that renters and leasers can rent or lease more than one property,and managers can manage multiple properties, so there would be 1:m relationship. Managers can

For data types, most are INT or VARCHAR.Since for database, we store mostly numeric values or description attributes.

Design Considerations

The database is a relational database. OLTP is needed since it needs to capture transactions(who rents the room now, payments) in real time.When creating this model, I consider 3 parts: People's information, House's information,and a "main" table with a lot of foreign keys as a sum.The volume of data depends on the number of properties and people managed by this database.

The number of users will be large(we need faster processing speed) and monetary transaction and customer info are sensitive and happen in this database.So,we need a distributed databases to take advantage of its data recoverability,improved performance, and transparency.

Data security,privacy,and integrity

Data security is very important to us and requires us to "protect data from compromise by external attackers and malicious insiders." Payment and SSN are very sensitive information, and they cannot be leaked. SO we need high level of data security.We need a great web firewall.

Data privacy is about how data is collected,shared and used.So we need a front-end applications to only allow customers manage their own account rather than entire database.Also,we will not have public api.Employees should sign an agreement for properly use database.

Data integrity is also important to us as the accuracy and consistency of data stored in our database are very important. If we make mistakes or typos on payment amount, customers will go mad. We also need to often backup our database

Part C

i.Using Public APIs:

```
##API
!pip install requests
!conda install requests
```

```
import requests
import pandas as pd
import json
```

In [2]:
```
response = requests.get("https://api.artic.edu/api/v1/artworks?page=2&limit=100")
print(response.status_code)
```

```
200
```

In [3]:
```
response_data=json.loads(response.text)
df=pd.DataFrame(response_data["data"])
print(df.head())
```

```
       id api_model                                        api_link  is_b
oosted  \
0  260125  artworks  https://api.artic.edu/api/v1/artworks/260125
False
1  224591  artworks  https://api.artic.edu/api/v1/artworks/224591
False
2    5983  artworks    https://api.artic.edu/api/v1/artworks/5983
False
3     201  artworks     https://api.artic.edu/api/v1/artworks/201
False
4    7001  artworks    https://api.artic.edu/api/v1/artworks/7001
False

                                               title alt_titles  \
0                                      Serie Gamines       None
1  Polychrome Standing Figure with Exaggerated He...       None
2                                          Birdstone       None
3  Dipper or Ladle with Interlocking Zigzag and S...       None
4  Bowl Depicting Rows Containing Repeated Geomet...       None
```

In [4]:
```
columns=["id",'title',"place_of_origin"]
df=df[columns]
```

Out[97]:

	id	title	place_of_origin

0	260125	Serie Gamines	Colombia
1	224591	Polychrome Standing Figure with Exaggerated He...	Nayarit state
2	5983	Birdstone	Illinois
3	201	Dipper or Ladle with Interlocking Zigzag and S...	New Mexico
4	7001	Bowl Depicting Rows Containing Repeated Geomet...	Nazca Valley
5	6391	Vessel in the Form of a Severed Trophy Head	South Coast
6	25059	Disc from an Earflare Inscied with a Figure in...	Veracruz Llave
7	16042	Standing Male Figure Holding a Plate	Colima state
8	44817	Nose Ornament with Lateral Extensions in Sugge...	Miraflores
9	35603	Pendant in the Form of an Abstract Bird with O...	Panama
10	51182	Bowl with Mirror Pattern of Birds Framed by Ge...	Southwest
11	51176	Bowl	Southwest
12	59390	Plate with Feather Motif	San Ildefonso Pueblo
13	52787	Polychrome Bowl with Abstract Geometric Motifs	Arizona
14	61535	Bowl with Interior Four-Part Design with Hatch...	Arizona
15	90907	Bowl Depicting Birds	Nazca Valley

16	83313	Cup plate	Sandwich
17	91497	Miniature Ceremonial Vessel (Aryballos)	South Coast
18	91111	One of a Pair of Matched Bowls Depicting Costu...	Peru
19	91532	Vessel Depicting Feline Figures	North Coast
20	91546	Ceremonial Vessel with Masked Deities	North Coast
21	91543	Stirrup Spout Vessel with Fineline Image of a ...	North Coast
22	91655	Blackware Vessel with Relief Depicting a Scene...	North Coast
23	91556	Vessel Depicting a Composite Feline Figure	North Coast
24	120519	Female Figure	México
25	120080	Eccentric Flint	México
26	126468	Miniature Bowl	South Coast
27	125777	One of a Pair of Matched Bowls Depicting Costu...	Peru
28	146930	Mosaic Disk with a Mythological and Historical...	Oaxaca state
29	126469	Miniature Bowl	South Coast

30	150057	Standing Figure of a Mother and Child	Jalisco
31	146942	Ceremonial Knife	Colima
32	212967	Shell Mosaic Ritual Mask	Teotihuacán
33	224587	Standing, "Smiling" Figure with Hands Raised	Veracruz state
34	241729	Eshu (The Trickster)	None
35	241045	Truro Series #1	None
36	202486	Pumpkin	None
37	194725	Untitled	United States
38	146908	Untitled	United States
39	15790	Folding Screen (Biombo)	China
40	192336	Deeps and Skies	United States
41	146904	Low Piece (Bench)	United States
42	118769	Bronze Chair	United States
43	117190	Family of Robot: Baby	Korea

44	185075	Highway Junction 14-5	Japan
45	102599	Unidentified Sitter	United States
46	50154	Chez Mondrian	United States
47	158462	Untitled	United States
48	109915	Clothespin	United States
49	76907	Beach Scene	United States

j.WebScraping:

In [98]:
```
## Websraping
import re
import csv
import requests
from bs4 import BeautifulSoup
```

In [99]:
```
url = 'https://en.m.wikipedia.org/wiki/List_of_current_United_States_senators#List_of_senators'
headers = {'User-Agent': "Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko)"}
response = requests.get(url, headers=headers)
html = response.content
```

In [100]:
```
soup = BeautifulSoup(html, 'lxml')
```

In [101]:
```
senator_table = soup.find('table', attrs={'id': 'senators'})
```

In [102]:
```
senators = open('us_senators.csv', 'w', encoding='utf-8', newline='')
writer = csv.writer(senators)
```

In [103]:
```
for idx, tr in enumerate(senator_table.find_all('tr')):
    row = []
    for td in tr:
        value = re.findall('>([^><]+)<', str(td))
        if len(value) > 0:
            row.append(''.join(value).strip())
    if idx > 0:
        if idx % 2 == 0:
            del row[2]
            row = [''] + row
        else:
            del row[3]
    writer.writerow(row)
```

In [104]:
```
senators.close()
```

In [105]:
```
import pandas as pd
senators_1 = pd.read_csv("us_senators.csv")
```

In [106]:
```
senators_1=senators_1.drop(columns=['Portrait',"Education","Assumed office",
                                    'Occupation(s)',"Born"])
```

senators_1.head(50)

	State	Senator	Party	Term up
0	Alabama	Richard Shelby	Republican[2]	2022
1	Alabama	Tommy Tuberville	Republican	2026
2	Alaska	Lisa Murkowski	Republican	2022
3	Alaska	Dan Sullivan	Republican	2026
4	Arizona	Kyrsten Sinema	Democratic	2024

5	Arizona	Mark Kelly	Democratic	2022
6	Arkansas	John Boozman	Republican	2022
7	Arkansas	Tom Cotton	Republican	2026
8	California	Dianne Feinstein	Democratic	2024
9	California	Alex Padilla	Democratic	2022
10	Colorado	Michael Bennet	Democratic	2022
11	Colorado	John Hickenlooper	Democratic	2026
12	Connecticut	Richard Blumenthal	Democratic	2022
13	Connecticut	Chris Murphy	Democratic	2024
14	Delaware	Tom Carper	Democratic	2024
15	Delaware	Chris Coons	Democratic	2026
16	Florida	Marco Rubio	Republican	2022
17	Florida	Rick Scott	Republican	2024
18	Georgia	Jon Ossoff	Democratic	2026
19	Georgia	Raphael Warnock	Democratic	2022

20	Hawaii	Brian Schatz	Democratic	2022
21	Hawaii	Mazie Hirono	Democratic	2024
22	Idaho	Mike Crapo	Republican	2022
23	Idaho	Jim Risch	Republican	2026
24	Illinois	Dick Durbin	Democratic	2026
25	Illinois	Tammy Duckworth	Democratic	2022
26	Indiana	Todd Young	Republican	2022
27	Indiana	Mike Braun	Republican	2024
28	Iowa	Chuck Grassley	Republican	2022
29	Iowa	Joni Ernst	Republican	2026
30	Kansas	Jerry Moran	Republican	2022
31	Kansas	Roger Marshall	Republican	2026
32	Kentucky	Mitch McConnell	Republican	2026
33	Kentucky	Rand Paul	Republican	2022

34	Louisiana	Bill Cassidy	Republican	2026
35	Louisiana	John Kennedy	Republican	2022
36	Maine	Susan Collins	Republican	2026
37	Maine	Angus King	Independent[a]	2024
38	Maryland	Ben Cardin	Democratic	2024
39	Maryland	Chris Van Hollen	Democratic	2022
40	Massachusetts	Elizabeth Warren	Democratic	2024
41	Massachusetts	Ed Markey	Democratic	2026
42	Michigan	Debbie Stabenow	Democratic	2024
43	Michigan	Gary Peters	Democratic	2026
44	Minnesota	Amy Klobuchar	Democratic	2024
45	Minnesota	Tina Smith	Democratic	2026
46	Mississippi	Roger Wicker	Republican	2024
47	Mississippi	Cindy Hyde-Smith	Republican	2026

48	Missouri	Roy Blunt	Republican	2022
49	Missouri	Josh Hawley	Republican	2024