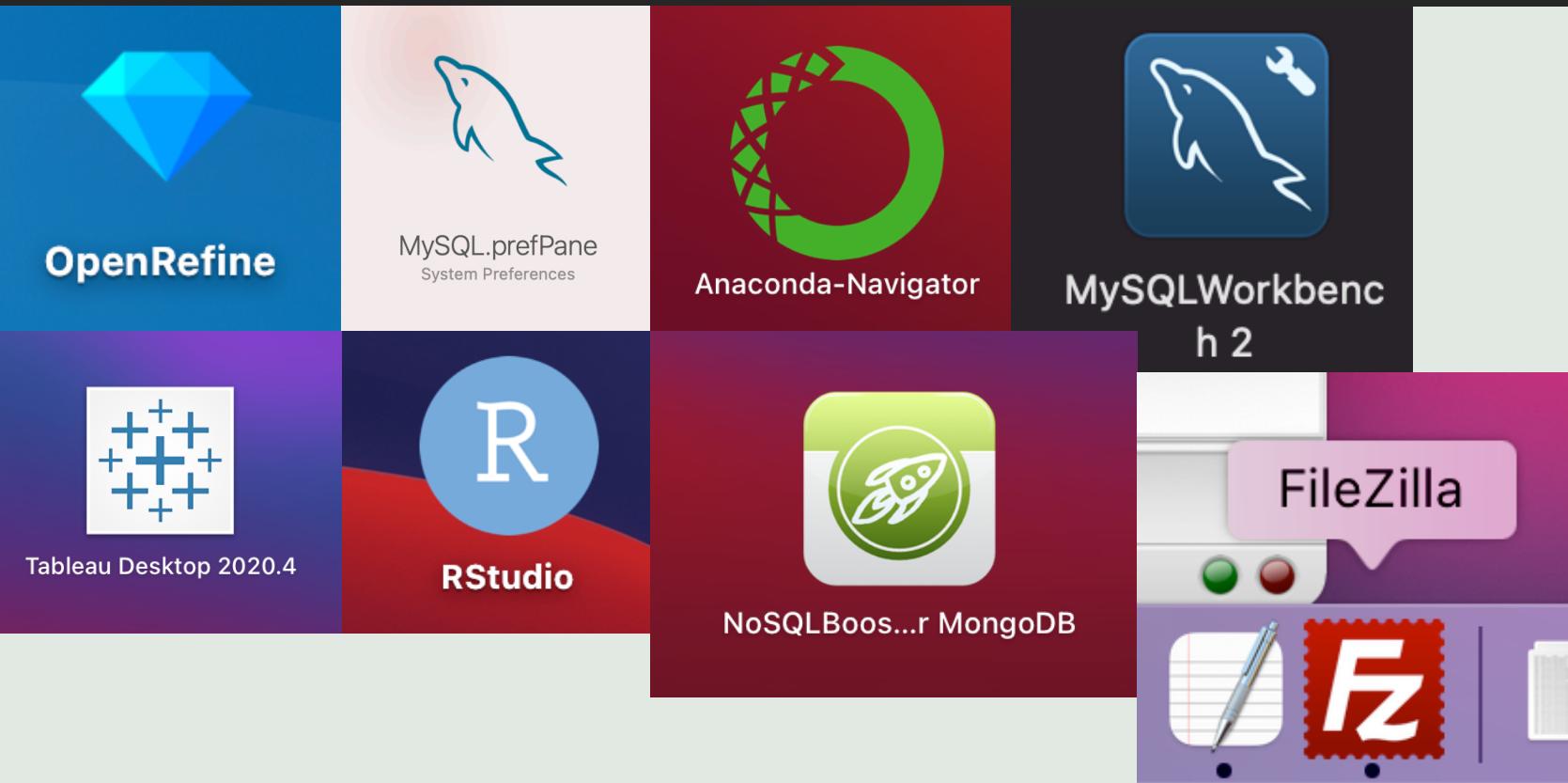


Data Engineering Platform

Assignment 1
Fiona Fei



Part A: 1. Install Software



Note :

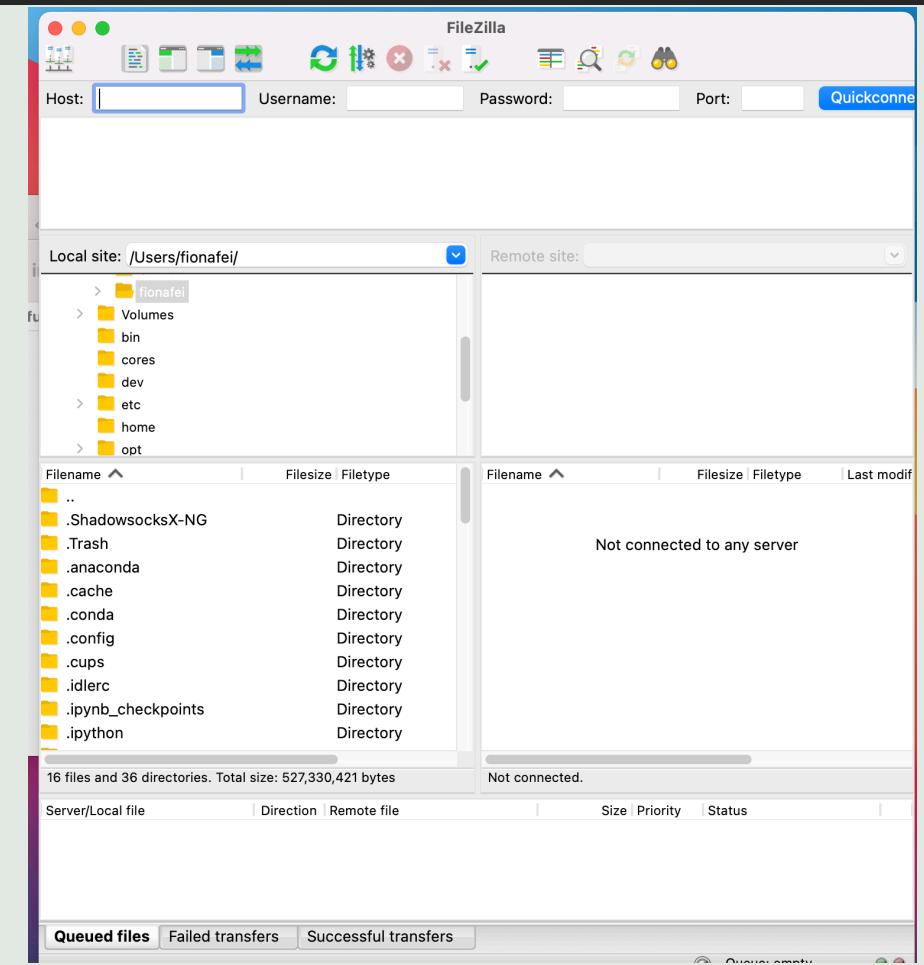
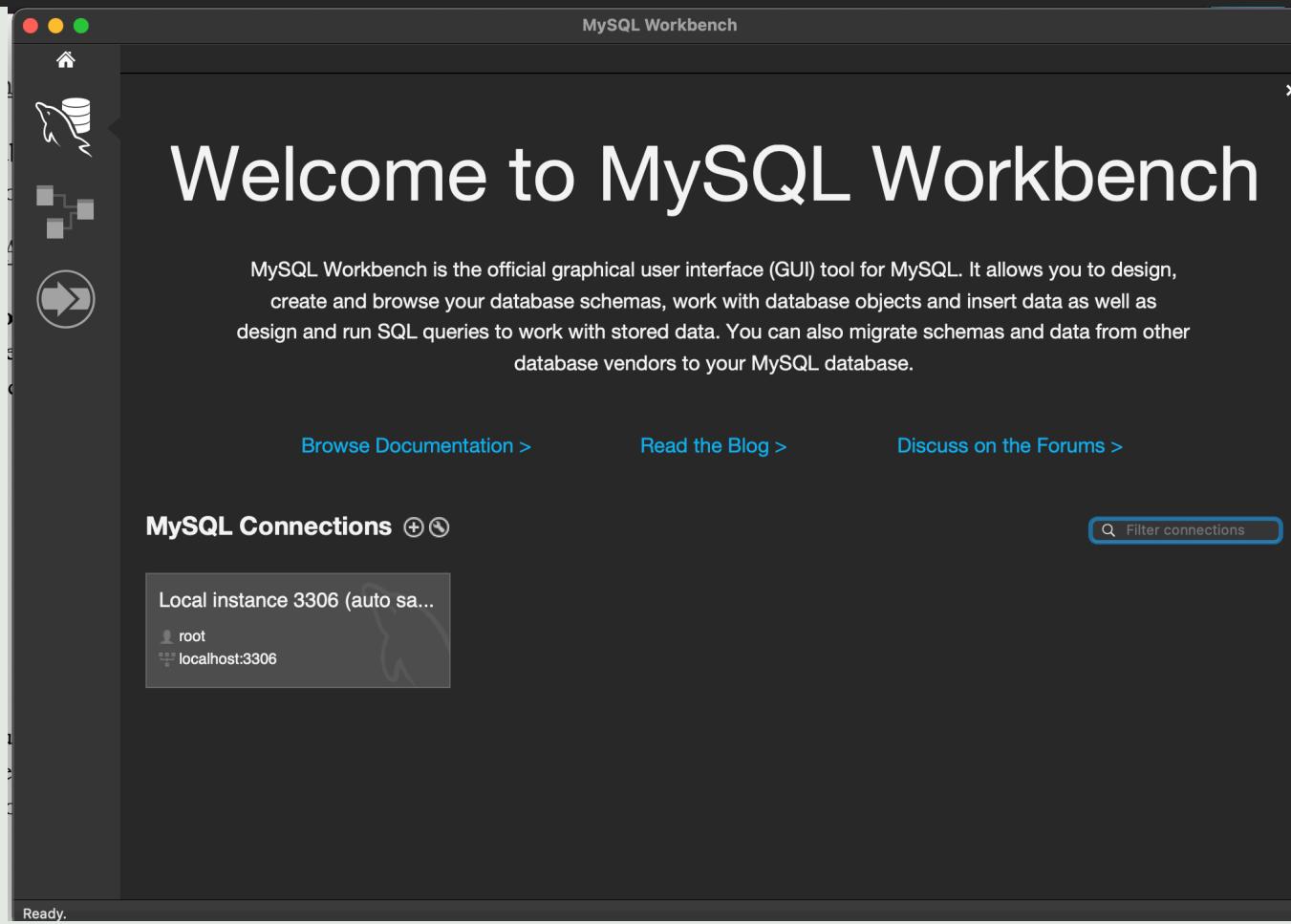
The applications are in different places on my laptop: the desktop, dock tool bar, application files, etc. I couldn't take a screenshot that contains all of them, so I took screenshots of them separately, but they are all in my laptop.

Part A: 1. Install Software

The image displays three separate screenshots illustrating software installation and cloud computing concepts:

- Anaconda Navigator:** A screenshot of the Anaconda Navigator interface, showing a grid of data analysis and machine learning tools. The visible applications include JupyterLab (2.1.5), Jupyter Notebook (6.0.3), IPython (4.7.5), RStudio (1.0.44), Spyder (4.1.4), Glueviz (1.0.0), and Orange 3 (3.26.0). Each application has a "Launch" or "Install" button.
- Google Cloud Platform:** A screenshot of the Google Cloud Platform homepage. It features a navigation bar with "My First Project". The main content area is titled "Welcome, Fiona" and "Get started with Google Cloud". Below this, there's a section titled "Begin with the basics" with a "GO TO CHECKLIST" button.
- Welcome Screen:** A general welcome screen with a dark background. It features a large "Welcome, Fiona" message and a "Get started with Google Cloud" call-to-action.

Part A: 1. Install Software



Part A: 2. Data Preparation Step d-g

Screenshot of a data preparation interface showing various steps (d-g) for preparing a dataset.

Step 1: Cross Street 2

	Cross Street 2	Address Type	City	Status	Due Date	Resolution Action
Water St	Address	Facet			11/06/2012 04:30:00 P	
Henry St	Intersection	Text filter				
Niles Place	Trim leading and trailing whitespace	Edit cells	Transform...			
113 Ave	Collapse consecutive whitespace		Common transforms			
Lisa Ln	Unescape HTML entities		Fill down			
East 78 Street	Replace Smart quotes with ascii		Blank down			
12 Street	To titlecase		Split multi-valued cells...			
	To uppercase		Join multi-valued cells...			
	To lowercase		Cluster and edit...			
	To number		Replace			

Step 2: Descriptor

- 692 choices Sort by: name count Cluster
- Other Billing Issue 29
- Other Hazmats 1
- Other Housing Options 6
- Other Issue 1
- Other School Condition 10
- Other Water Problem** 23
- Out of Order 25
- Over Capacity 3
- Overcharge 623

Step 3: Complaint Type

- 166 choices Sort by: name count Cluster
- MANHATTAN DOF Property - Owner Issue 127
- DOF Property - Payment Issue 71
- MANHATTAN DOF Property - Reduction Issue 105
- DOF Property - Request Copy 72
- DOT Literature Request 8
- DPR Literature Request 6
- Drinking 15
- Drinking Water 3
- EAP Inspection - F59 14
- Elder Abuse 15

Step 4: Location Type

- 69 choices Sort by: name count Cluster
- Sidewalk 687
- Single Room Occupancy (SRO) 4
- Street 27758
- Street Address 222
- Street/Curbside 6
- Street/Sidewalk 2360
- Tattoo Parlor 2
- Tenant Address 2
- Terminal 12
- Utility 4

Step 5: Community

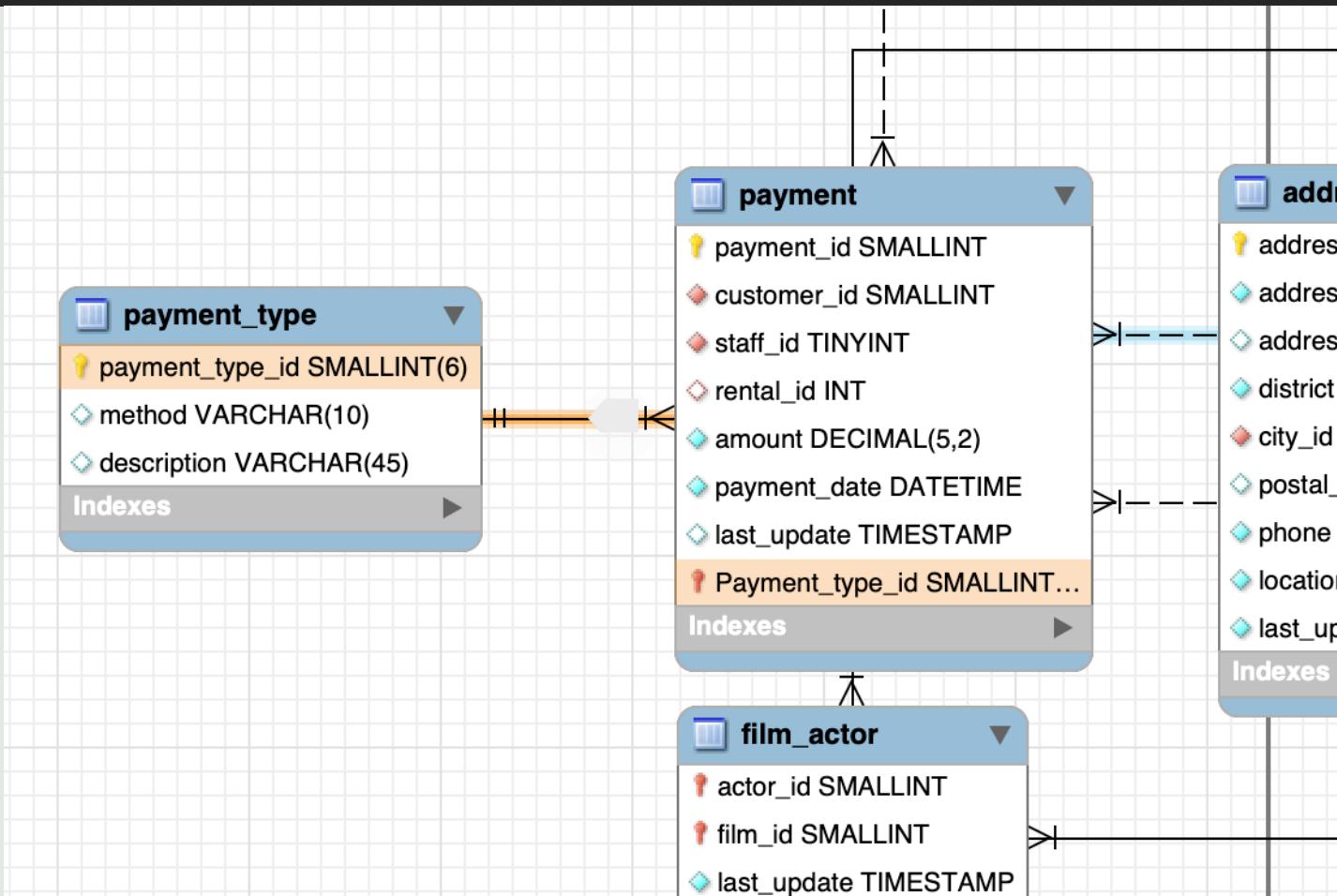
- 01 MANHATTAN
- 02 BROOKLYN
- 02 STATEN ISLAND
- 12 QUEENS
- Drinking

Part A: 2. Data Preparation Step d-g

The image shows a data preparation interface with four main panels:

- City Selection Panel (Left):** Displays a list of cities with a count of incidents. The list includes Rockville Center, Rosedale, Saint Albans, Sioux Falls, South Ozone Park, South Ozone Pk, South Richmond Hill, Springfield Gardens, Staten Island, Sunnyside, Trevose, Club/Bar/Restaurant, Co-Op Unit, Commercial, Commercial Building, Condo Unit, Construction Site, Crosswalk, and Curb.
- Incident Address Selection Panel (Top Center):** Shows two address fields: "80 Pine Street" and "Pine St". Below them is a dropdown menu with "South Ozone Park" selected. At the bottom are "Apply" and "Cancel" buttons.
- City Details Panel (Middle Center):** Displays a list of cities with a count of incidents. The list includes Maspeth, Middle Village, N/a, Natick, New Hyde Park, New York, Newark, Ny, Oakland Gardens, Ozone Park, Passaic, and Queen.
- Summary Table Panel (Bottom Right):** A table showing incident counts for various categories. The categories and counts are: Public/Unfenced Area (9), Recreation Center (3), Residential (31251), Residential Building/House (3764), Restaurant/Bar/Deli/Bakery (159), Roadway (4), School (46), Senior Address (404), and Smoking (Damaged Tree).

Part B: 1.d



Part B: 1.e Relational Data Modeling

Field (Attributes)	Primary Key (Y/N)	Foreign Key (Y/N)	Related Tables(s)
Rental_id		Y	Rental (M:N)
Paymentd_id	Y	Y	
Customer_id	N	Y	Customer (1:N)
Staff_id	N	Y	Staff (1:N)

Part B: 2.a Examples of anomalies

- Insertion anomalies:

If the surgery is defined and null values are not accepted, a new patient without assigned surgery won't be able to insert into the database.
- Deletion anomalies:

If Jay Patel quit the physician job and deleted him from the database, the Liver Transplant surgery will also be ceased to exist because it's only assigned to Jay Patel.
- Modification anomalies:

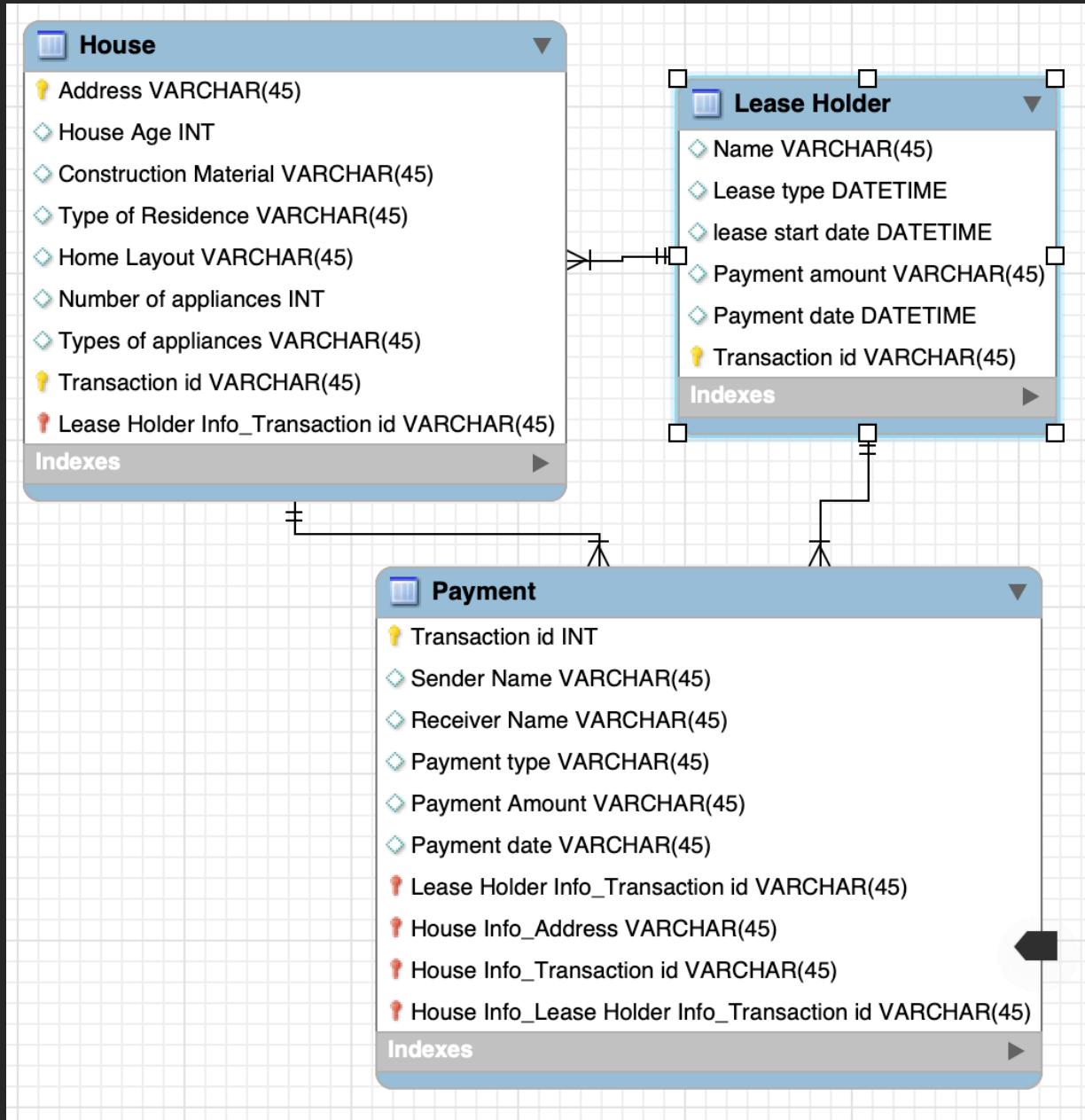
If Robert Smith's address is incorrect and needs to be updated, it needs to be updated for two times or there will be inconsistent data in the database.

Part B: 2.b Normalize to 3NF

A	B	C	D	E
Physician			Patient:	
Physician Name	Physician Office		Patient Name	Patient Address
Helen Pearson	Chicago Ave, Chicago		Joe Korn	Randolph Street, Chicago
Olga Kay	Clark Street, Chicago		Gillian White	Illinois Street, Chicago
Robert Smith	Madison Street, Chicago		Jill Bell	Huron Street, Chicago
Wei Jing	Adams Street, Chicago		Mike Li	Lake Street, Chicago
Jay Patel	Monroe Street, Chicago		Ian MacKay	Dearborn Street, Chicago
			Sheela Nupur	Monroe Street, Chicago
Appointment				
Appointment Date	Patient Name	Physician Name	Surgery Type	
3/7/17	Joe Korn	Helen Pearson	Tendon Repair	
3/22/17	Gillian White	Helen Pearson	Skin Graft	
6/13/16	Joe Korn	Olga Kay	Sentinel Node Biopsy	
6/13/17	Jill Bell	Robert Smith	Tendon Repair	
6/14/17	Jill Bell	Robert Smith	Skin Graft	
6/13/17	Mike Li	Wei Jing	Knee Arthroscopy	
8/15/17	Gillian White	Jay Patel	Sentinel Node Biopsy	
1/4/16	Ian MacKay	Jay Patel	Hepatic Resection	
1/5/18	Ian MacKay	Jay Patel	Liver Transplant	
1/4/16	Sheela Nupur	Helen Pearson	Knee Arthroscopy	
2/12/16	Joe Korn	Wei Jing	Skin Graft	
4/15/18	Mike Li	Wei Jing	Skin Graft	

Assumptions:

- Assume each physician has an assigned unique office
- Assume each patient have an assigned address and appointment.
- Assume each appointment has complete information about the date, patient name, physician name, and surgery type.



Part B: 3 Data Modeling

- Tables(3):
House, Lease Holder, Payment
- Relationships:
House is connected to the lease holder by sharing the same transaction id, lease holder is connected to payment by sharing the same transaction id, payment date, payment amount. House is connected to the payment sharing the same transaction id.

Part B: 3. Data Modeling

- Cardinality

House (N:M) Lease Holder

House (1:N) Payment

Lease Holder (1:N) Payment

- Datatypes:

VARCHAR(45): Address, Construction Material, Type of Residence, Home Layout, Types of appliances, transaction id, Name, lease type, payment amount, sender name, receiver name

INT: House Age, Number of Alliances

Date: Lease start date, payment date

Part B: 3. Data Modeling

Summary of Design Considerations:

This database is designed as a relational database. The database could be very large depends on the number of houses that are available and the number of customers that have the intention to purchase the houses.

The number of users and the number of houses will determine whether centralized will be sufficient or distributed, and therefore determine the critical entities and store them centrally.

Part B: 3. Data Modeling

Summary of Design Considerations:

There are some fields that needs security storage such as the lease holder's name and payment amount. They should belong to high privacy components that are not openly to the public.

As for data integrity, it is necessary to restrict the users that are allowed to get access to this database because it contains the personal information of lease holders and their payments. The information should be only be accessed by limited number of users. And the 3NF format will help with keeping data private and reduce the data anomalies scenarios.