# ASSIGNMENT 1

**Assignment Objectives**

- Data Preparation
- Relational Algebra
- Relational Modeling
- Database Normalization

**Software Installation**

Follow the installation guides in the course material and install the following software on your computer: OpenRefine, MySQL database, MySQL workbench

**Data**

- Download the sandyrelated.csv dataset which can be found under the data folder on the course portal.
- Create database using sql scripts located under **Files>Data>Sakila** folder in the course material
- First run sakila-schema.sql and then sakila-data.sql
- Further documentation : https://dev.mysql.com/doc/sakila/en/

**Submissions**

- Solutions and screenshots should be submitted as a single PDF or Microsoft Word document.
- Submit other artifacts such as Excel or text files as required.
- Please do not need to submit the cleaned up dataset(s) or the OpenRefine project

1. **Data Wrangling Exercise:**
   Run the following data preparation steps on the dataset below and submit screenshots for questions d-h.
   a. Import the dataset into OpenRefine and create a new project "SandyCleanup"
   b. Trim white spaces on all address related columns and transform addresses into title case
   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.
   d. Convert the City column to title case, then Cluster and Merge the column
   e. Clean up Descriptor Column - Cluster and Merge 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"
      4. "Personal DRIE Exemption", "Personal SCHE Exemption", "Personal DHE Exemption" as "Personal Exemption"
   f. Clean up Location Type - Cluster and Merge following text categories:
      1. "Comercial", "Commercial", "Store/Commercial" as "Commercial"
      2. "RESIDENTIAL BUILDING", "Residential Building", "Residence" as "Residential"
      3. "Club/Bar/Restaurant", "Bar/Restaurant", "Restaurant" as "Club/Bar/Restaurant"
      4. "3+ Family Apt. Building", "3+ Family Apartment Building" as "3+ Family Apartment"
      5. "Street/Sidewalk", "Street and Sidewalk" as "Street/Sidewalk"

**Optional Practice Questions**

g. Look for at least two other clean up opportunities and execute using OpenRefine
h. Export final project into a CSV file on your local computer (you do not need to submit this file)
i. Online web services such as the following can be used to fetch the address given a geocode:
   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

a. Download Sakila dataset and unzip sakila-db.zip file
b. Open MySQL workbench and execute the sakila-schema.sql script
c. Reverse Engineer the database and generate the EER model
d. Modify the EER model to add a new lookup table : **payment_type**
   - This table will have a 1 to Many relationship with the Payment table.
   - Attributes of payment_type table:
     - payment_type_id (Primary Key) : SMALLINT(6)
     - method - varchar (10)
     - description – varchar (45)
   - Add payment_type_id as a foreign key in the Payment table as follows:
     - payment_type_id (Foreign Key) : SMALLINT(6)

Note: Submit the screenshot for the above change in the EER model. You do not need to make changes to the physical tables or add data in the database.

e. For the Payment table fill out the form below:

**Table Name:** Payment

| Field ( Attributes) | Primary Key (Y/N) | Foreign Key (Y/N) | Related Table(s) and Cardinality between tables |
|---|---|---|---|
| | | | |
| | | | |
| | | | |

## 3. Relational Algebra

For the Sakila dataset, provide the relational algebra syntax (only) for the following queries :

a. List all payments greater than and equal to 2$ and less than equal to 7$
b. List all the movies with title and description that are rated PG-13
c. Replace the word “film” with “movie” for all attributes and relations starting with the word “film”
d. List all customer names who have returned their rentals in the current month

**4. Normalization**

a. For the table below, provide examples of insertion, deletion, and modification anomalies.

b. Normalize this data to 3NF and list any assumptions made during the normalization process. Submit an Excel workbook and with a separate tab for each normal form

| Physician Name | Patient Name | Patient Address | Appointment Date | Surgery |
|---|---|---|---|---|
| Helen Pearson | Joe Korn | Randolph Street, Chicago | 3/7/2017 | Tendon Repair |
| Helen Pearson | Gillian White | Illinois Street, Chicago | 3/22/2017 | Skin Graft |
| Olga Kay | Joe Korn | Randolph Street, Chicago | 6/13/2016 | Sentinel Node Biopsy |
| Robert Smith | Jill Bell | Huron Street, Chicago | 6/13/2017 | Tendon Repair |
| Robert Smith | Jill Bell | Huron Street, Chicago | 6/14/2017 | Skin Graft |
| Wei Jing | Mike Li | Lake Street, Chicago | 6/13/2017 | Knee Arthroscopy |
| Ashish Patel | Gillian White | Dearborn Street, Chicago | 8/15/2017 | Sentinel Node Biopsy |
| Ashish Patel | Iam MacKay | Dearborn Street, Chicago | 1/4/2016 | Hepatic Resection |
| Ashish Patel | Iam MacKay | Dearborn Street, Chicago | 1/5/2018 | Liver Transplant |
| Helen Pearson | Sheela Nupur | Monroe Street, Chicago | 1/4/2016 | Knee Arthroscopy |
| Wei Jing | Joe Korn | Randolph Street, Chicago | 2/12/2016 | Skin Graft |
| Wei Jing | Mike Li | Lake Street, Chicago | 4/15/2018 | Skin Graft |