DATA ENGINEERING PLATFORMS (MSCA 31012)

ASSIGNMENT 1

Submissions (via Canvas)

- Submit solutions in PDF, PPT, Excel or MS Word document (as applicable). Do not submit zip files.
- Do not submit the cleaned up dataset for the OpenRefine project.

Part A : Software installations, data extraction, cleaning & transformation

1. **Follow the installation guides uploaded** (or search google for installation instructions) and install the following software on your local computer (submit a **screenshot of your desktop** with shortcuts and validations). **– { 20 Points }**
    1) OpenRefine
    2) MySQL (server + workbench)
    3) Anaconda (Open Data Science Platform : Python)
    4) R-studio
    5) Tableau (https://www.tableau.com/academic/students)
    6) FileZilla Or CyberDuck
    7) MongoDB
    8) GCP (credits added to your account)

2. Run the following data preparation steps on the dataset below and submit **relevant screenshots** for steps d-g as word or pdf document. **– { 20 Points }**
   Note: Dataset sandyrelated.csv is uploaded as part of this assignment.
    a. Import the data into OpenRefine and create a new project "SandyCleanup"
    b. Remove columns where majority of the cells are empty or have "Unspecified" or "NA" values (Do not remove the columns that are needed to complete the rest of this exercise)
    c. Trim white spaces on all address related columns and transform addresses into title case
    d. Convert City to title case, then Cluster and Merge the column
    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"
        3. "Street/Sidewalk", "Street and Sidewalk" as "Street/Sidewalk"
    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. Please follow the best practices for file naming.

Part B : Relational data model and design principles

Data (Sakila dataset)

- We will use the Sakila database schema which can be found at: http://dev.mysql.com/doc/index-other.html
- Full documentation: http://dev.mysql.com/doc/sakila/en/

1. Relational Data Modeling **– { 20 Points }**

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 the following 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. 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) (only enter this for foreign key fields) & Type of relationship between tables

2. Normalization : For the table below: **– { 10 Points }**

a. Provide examples of insertion, deletion, and modification anomalies.
b. Normalize this table to 3NF and list any assumptions.

Physician Name	**Physician's Office**	**Patient Name**	**Patient Address**	**Appointment Date**	**Surgery**
Helen Pearson	Chicago Ave, Chicago	Joe Korn	Randolph Street, Chicago	3/7/2017	Tendon Repair
Helen Pearson	Chicago Ave, Chicago	Gillian White	Illinois Street, Chicago	3/22/2017	Skin Graft

Olga Kay	Clark Street, Chicago	Joe Korn	Randolph Street, Chicago	6/13/2016	Sentinel Node Biopsy
Robert Smith	Madison Street, Chicago	Jill Bell	Huron Street, Chicago	6/13/2017	Tendon Repair
Robert Smith	Madison Street, Chicago	Jill Bell	Huron Street, Chicago	6/14/2017	Skin Graft
Wei Jing	Adams Street, Chicago	Mike Li	Lake Street, Chicago	6/13/2017	Knee Arthroscopy
Jay Patel	Monroe Street, Chicago	Gillian White	Illinois Street, Chicago	8/15/2017	Sentinel Node Biopsy
Jay Patel	Monroe Street, Chicago	Ian MacKay	Dearborn Street, Chicago	1/4/2016	Hepatic Resection
Jay Patel	Monroe Street, Chicago	Ian MacKay	Dearborn Street, Chicago	1/5/2018	Liver Transplant
Helen Pearson	Chicago Ave, Chicago	Sheela Nupur	Monroe Street, Chicago	1/4/2016	Knee Arthroscopy
Wei Jing	Adams Street, Chicago	Joe Korn	Randolph Street, Chicago	2/12/2016	Skin Graft
Wei Jing	Adams Street, Chicago	Mike Li	Lake Street, Chicago	4/15/2018	Skin Graft

3. Data Modeling: **– { 10 Points }**

Design a data model that can be used for property management and monitoring of single-family homes for investors and owners. Consider data for the following entities/attributes that need to be captured by business:

a. Home location
b. Age of the house
c. Construction material used
d. Type of residence (apt, condo, etc.)
e. Home layout (number of roomes, sq footage, etc.)
f. Number and Types of Appliances (Heating, Fridge etc.)
g. Name and other details of the renters/leasers/resident (s)
h. Rental Payments made against the house
i. Add other entities (and/or collection of attributes) that you think could add insights for the investors and business users

Please submit a PPT with 4 slides that details the Entity Relationship Diagram (tables/relationships/cardinality/datatypes), short summary of Design considerations (which database, how many users , need for distributed databases, data security, privacy and integrity).

Part C: Data Collection & Preparation

1. This assignment is related to data collection and transformation. **– { 20 Points}**
    i. Using Public APIs : Choose any data provider (such as Twitter/YouTube… etc). to collect data and transform it to a clean structured tabular data (sample size of 50 records) using Python
    j. WebScraping : Choose a website you want to scrape. Collect some of the data from the website and transform it to a clean structured tabular data (sample size of 50 records) using Python

References:

- https://medium.com/pew-research-center-decoded/using-apis-to-collect-website-data-b7fc340d59e3
- https://towardsdatascience.com/getting-started-with-apis-in-python-to-gather-data-1185796b1ec3
- https://www.dataquest.io/blog/python-api-tutorial/
- https://www.dataquest.io/blog/web-scraping-tutorial-python/
- https://likegeeks.com/python-web-scraping/

Another useful site : https://lmgtfy.com/