Shruthi Kalasapura Ramesh

B00822766

**Assignment 1**

**Database Management & Warehousing Analysis**

1. **Feasibility Analysis**
2. Do you have enough datasets to start building the CMS? List all the datasets, you find critical for the project.

Yes, I have sufficient datasets to start building the CMS for “Analytics 5408”.

1. List all the datasets, you find critical for the project. Note: You need to provide URLs of the

actual datasets and write reason of selection in single sentence.

**Education/Schools**

|  |  |
| --- | --- |
| Dataset1:  *Combined public and private expenditure on educational institutions* | URL: <https://open.canada.ca/data/en/dataset/e82807fd-b7e8-42cd-89d1-da70a07d2a72>  Reason: *This dataset provides the combined public and private expenditure on educational institutions as Indices of change and percentage distribution, based on level of education.* |

**Motor vehicles**

|  |  |
| --- | --- |
| Dataset1:  *Length of track* | URL: [*https://open.canada.ca/data/en/dataset/8f99b0b8-ba66-4b8a-969b-43a4512c4c76*](https://open.canada.ca/data/en/dataset/8f99b0b8-ba66-4b8a-969b-43a4512c4c76)  Reason: *This is a R*ailway industry *dataset that provides information of* length of track operated at the end of the year, by company. |
| Dataset 2:  *Fuel Consumption* | URL: [*https://open.canada.ca/data/en/dataset/9fc7c3d3-47e5-4a7a-a839-15df20c91960*](https://open.canada.ca/data/en/dataset/9fc7c3d3-47e5-4a7a-a839-15df20c91960)  Reason: The dataset provides information of diesel fuel consumption by Railway industry, by year |

**Economics & Industry**

|  |  |
| --- | --- |
| Dataset1:  *Cannabis industry production account* | URL: [*https://open.canada.ca/data/en/dataset/7fdcee18-317a-48e6-9bbc-cd481a714881*](https://open.canada.ca/data/en/dataset/7fdcee18-317a-48e6-9bbc-cd481a714881)  Reason: *This dataset helps to understand production information of Cannabis Industry.* |
| Dataset 2:  *Cannabis Income account* | URL: [*https://open.canada.ca/data/en/dataset/86a5c29c-0871-47ad-8da6-8a6b3992aea1*](https://open.canada.ca/data/en/dataset/86a5c29c-0871-47ad-8da6-8a6b3992aea1)  Reason: The dataset gives an insight of income by cannabis industry. |
| Dataset 3:  *Cannabis supply & use* | URL: [*https://open.canada.ca/data/en/dataset/7bd8c167-0df1-416b-929f-d572ca91bbcf*](https://open.canada.ca/data/en/dataset/7bd8c167-0df1-416b-929f-d572ca91bbcf)  Reason: This dataset helps us to comprehend supplies and uses of cannabis by year. |

1. From your selected datasets, identify entity sets and their attributes. You can create your own

entity set by omitting certain fields/attributes from a dataset, or by merging more than one

datasets. Consider fields/attributes that you find crucial to design the CMS.

Tools and programming languages used to clean and organize data are Microsoft Excel, MySQL workbench and Python.

**Education**

The dataset *Indices of change and percentage distribution of combined public and private expenditure on educational institutions, by level of education (D1)* is useful in this domain. This dataset has 16 attributes out of which, I feel, 8 are useful to get the clear picture of the dataset.

Entity sets of D1:

***Expenditure\_edu*** (REF\_DATE, DGUID, GEO, Level of education, Statistics, UOM, UOM\_ID, SCALAR\_FACTOR, SCALAR\_ID, VALUE)

**Motor vehicles**

The datasets identified in this domain are *Railway industry length of track operated at the end of the year, by company (D1)* and *Railway industry diesel fuel consumption(D2).* Both D1 and D2 have 16 entities. I have selected 10 attributes for my design from each dataset.

Entity sets of D1:

***TracksLengthsOperated*** (REF\_DATE, DGUID, GEO, Companies, Length of track, UOM, UOM\_ID, SCALAR\_FACTOR, SCALAR\_ID, VALUE)

Entity sets of D2:

***FuelConsumption*** (REF\_DATE, DGUID, GEO, Diesel fuel consumption, UOM, UOM\_ID, SCALAR\_FACTOR, SCALAR\_ID, VALUE)

**Economics & Industry**

I have chosen 3 datasets viz. *Cannabis industry production account(D1), Cannabis supply, use and gross domestic product (D1)* and*Cannabis income account(D3)*. After omitting the unnecessary fields. From D1, D2, D3 I have chosen11, 10 the following entity sets:

Entity sets of D1:

***CannabisProdution*** (REF\_DATE, DGUID, GEO, Indicator, Industry, Authority, UOM, UOM\_ID, SCALAR\_FACTOR, SCALAR\_ID, VALUE)

Entity sets of D2:

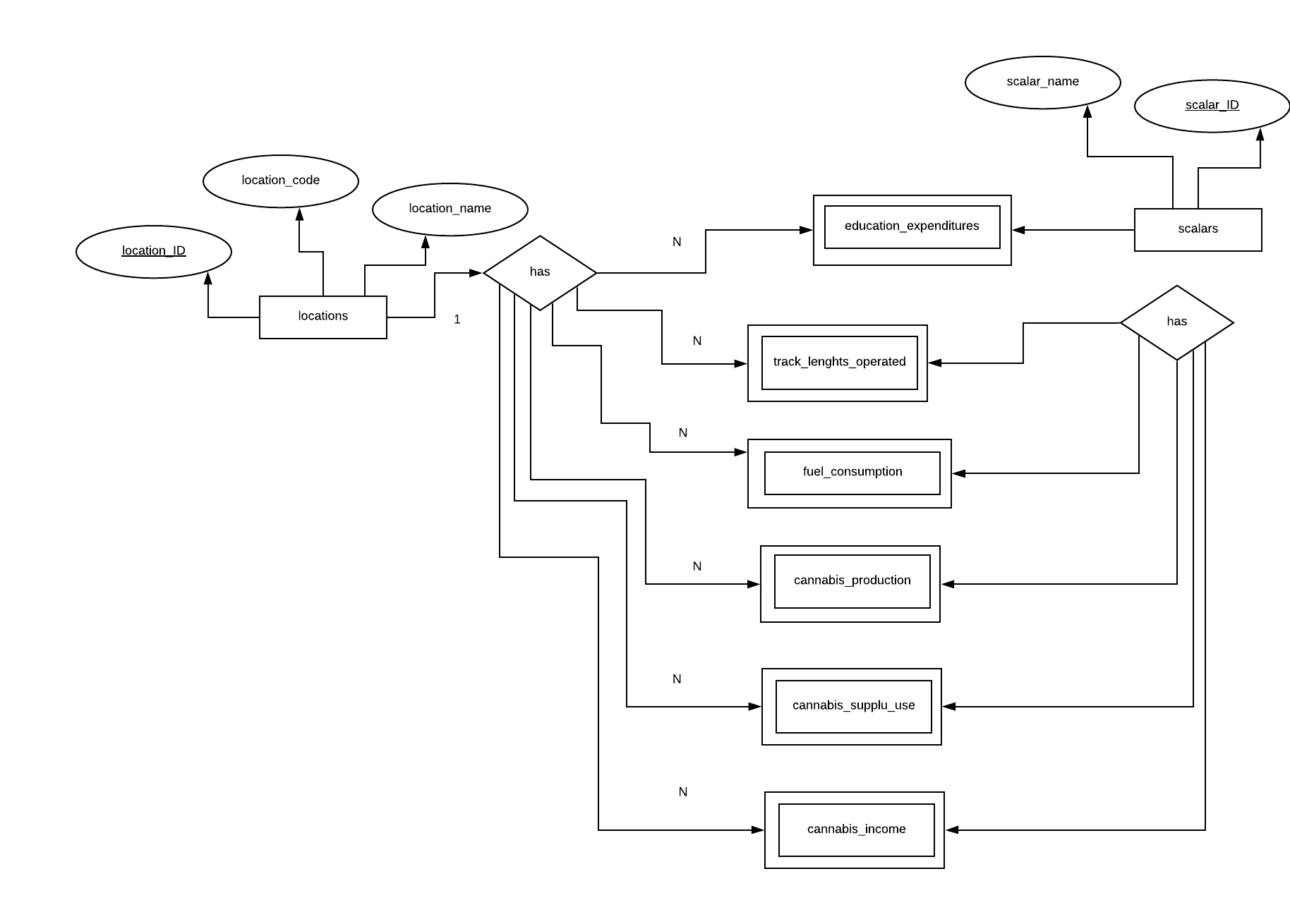
***CannabisSupplyandUse*** (REF\_DATE, DGUID, GEO, Indicator, Type of use, UOM, UOM\_ID, SCALAR\_FACTOR, SCALAR\_ID, VALUE)

Entity sets of D3:

***CannabisIncome*** (REF\_DATE, DGUID, GEO, Estimate, Industry, UOM, UOM\_ID, SCALAR\_FACTOR, SCALAR\_ID, VALUE)

1. List the strong/weak entity sets in your system. In addition, display the total participation.

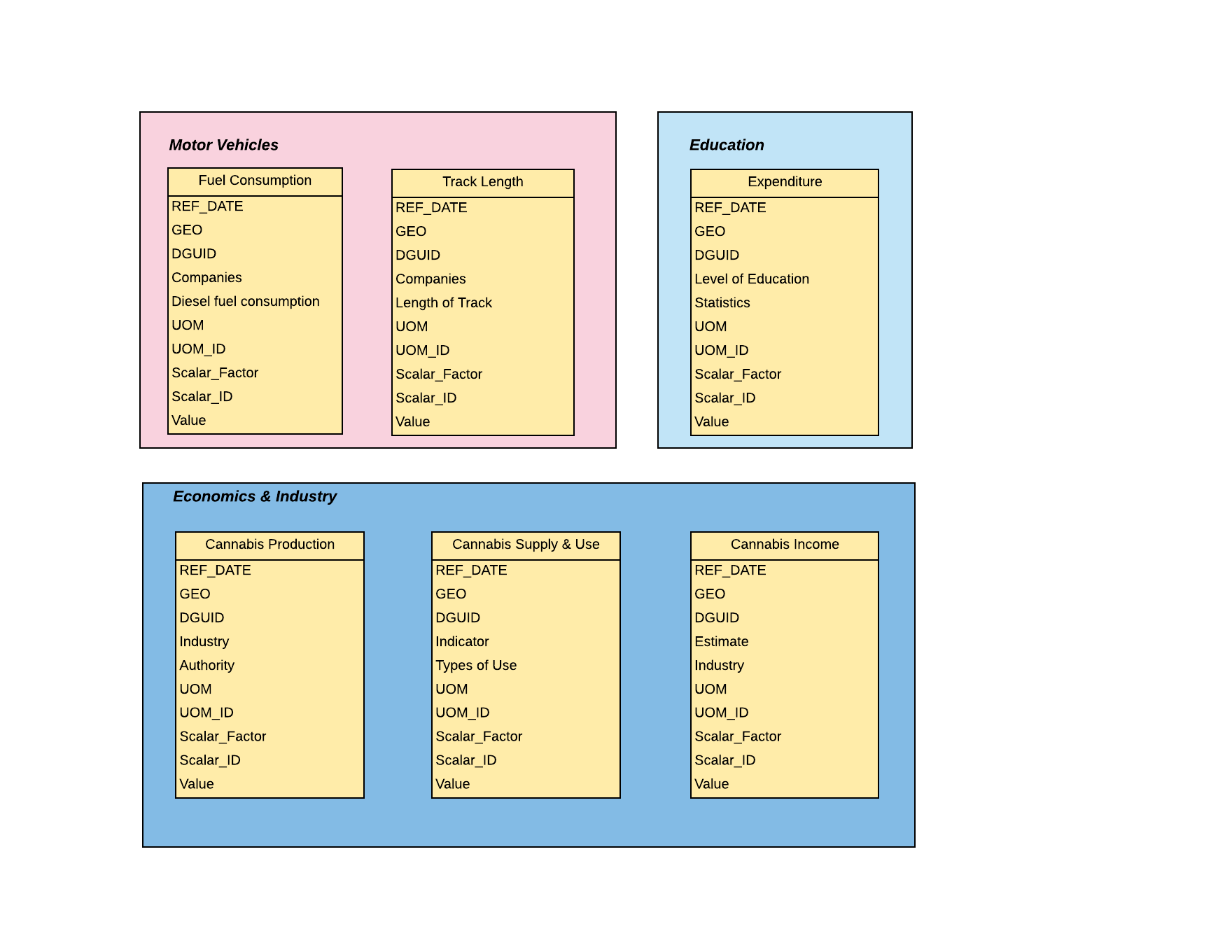
Weak entities are not identified in the initial design. Normalizing the entities facilitates the identification of weak and strong entities. After normalization, I have recognized few entities such as Locations, Measures and Scalars as strong entities. As show in the figure locations, measures, and scalars have 1 to many relationships with education\_expenditure, track\_lengths\_operated, fuel\_consumption, cannabis\_production, cannabis\_supply\_use, cannabis\_income.



1. **Data Modelling**

1. Is your initial sketch/design free from any design issues? (Yes/No) Provide justification to support your answer.

No. This is my initial model. The design is not normalised, relationships between the entities are not determined, and the entities do not have unique primary keys. This database has redundant values and hence it is not efficient.



There are no disjoints and overlapping subtypes in the normalized ERD, hence no extended ERD has been designed.

1. **Database Creation and Data Insertion & Normalization:**

I have created two different databases for initial and final designs. All the files used for initial design and final design are in ***InitialFiles*** and ***FinalFiles*** folders respectively.

Contents of ***InitialFiles:*** All CSV files before cleaning the data, image and ERD of initial design, SQL text file that has create and insert statements for all the tables.

Contents of ***FinalFiles*** ***:*** All CSV files after cleaning and organizing the data, image and ERD of final design, SQL text file that has create and insert statements for all the final tables.

References

[1]"MySQL :: MySQL Workbench Manual :: 6.5.2 SQL Data Export and Import Wizard", *Dev.mysql.com*, 2019. [Online]. Available: https://dev.mysql.com/doc/workbench/en/wb-admin-export-import-management.html. [Accessed: 31- May- 2019].

[2]"The 3 Types of Relationships in Database Design | Database.Guide", *Database.guide*, 2019. [Online]. Available: https://database.guide/the-3-types-of-relationships-in-database-design/. [Accessed: 31- May- 2019].