**Problem Set 4**

*You may work on this in groups if you want, but each person should submit their own answers.*

Find a dataset that is currently in a table that you think could be turned into a multi-table database. This can be any kind of data that interests you.

**1)** Describe the data. Write a metadata document that describes the source of the data, what the data are, what each column is, etc. See for example, [this (Links to an external site.)](https://github.com/CompTools/Class_Files/blob/master/data/flights_metadata.md) or [this (Links to an external site.)](https://github.com/CompTools/Class_Files/blob/master/data/CO_OPS__wl_file.md). (4 pts.)

**Dairy cattle project in New Guinea.csv Metadata**

**Information about the Dairy cattle project in New Guinea.csv file.**

This file contains some farmer’s personal information as well as farming activities from collected through a survey in a liaison with Swiss NGO. The data are from the department (state) of New Guinea, Nicaragua for starting a project in order to enhance dairy cattle production and management in this area.

After conducting surveys in 20 counties Data were collected from 95 farmers in total and available March 30, 2014.

The columns in the dataset are in the table below.

| **Column header** | **Description** |
| --- | --- |
| Farm\_ful\_nam | Farmer’s full names |
| Pers\_ID | This stands for the personal identification obtained of farmers all over the department of New Guinea. |
| Geogra\_locat | Geographic location , the county where the farms are located |
| Farm\_siz\_ha | This is the farm size in hectares per farmer |
| Annua\_incom$ | Annual incomes in U.S dollars |
| Annua\_expnse$ | Annual expenses in U.S dollars |
| Famil\_expnse$ | Annual family expenses in U.S dollars |
| Steers\_num | The total number of steers owned by farmers |
| Heifers\_num | Number of heifers raised per farmer |
| Bulls\_num | Total number of bulls being bred per farmer |
| Cows\_num | The total amount of cows per rancher |
| Tot\_herd\_num | This column stands for the total number of herds considered that every farmer possesses. |

**2)** Normalize the data. Write descriptions (words, drawings, or tables are fine--not code at this point) of the tables you could use to normalize these data. Describe why you chose the normalize the data in this manner. Include the column names, data types for each column, primary and foreign keys you would use. (8pts.)

I normalized the dataset called “Farmer’s info” in the **Dairy cattle project in New Guinea database** by adding a primary key called **“TraineeID”** since I had the full names of the farmers before. Simultaneously, I added another column called **“Training\_received”** and for this one I am using a foreign key since I created another table called **“Trainings”** in a different spreadsheet listing and describing the training courses that are offered by the NGO that is conducting the project in that specific geographic location.

The reason I decided to normalize these data is because I want to use my first table called “Farmer’s info” and link it with the information, and data I have in other worksheets. For instance, by adding a specific number or code to the farmer’s names not only will facilitates any sort of data analysis, but also will make my data more consistent and accurate, with not repeated values or names that could cause inconsistency.

Table 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TraineeID** | **Farm\_ful\_nam** | **Training\_received** | **Geogra\_locat** | **Farm\_siz\_ha** |
| 1 | AMIN NOE MEJIA URBINA | 9AI | PUERTO PRINCIPE | 120 |
| 2 | ABELINO ARAGON HERNANDEZ | 6GM | KURINWAS | 36 |
| 95 | ABELINO LOPEZ LOPEZ | 7AN | LA ESPERANZA | 20 |

**Data type**

**Trainee ID:** this is an INTEGER (20) data type and I am using here as a primary key

**Farmer’s full name:** this one is a VARCHAR/TINYTEXT (100) type where I give around 100 character for a full name description.

**Training received:** this is also another INTEGER (20) data type representing the kind of training every farmer has taken. Thus, the numbers display stand for the training’s names created that I show later.

**Geographic location:** I will be using the CHAR (100) type by describing the counties that will be selected for this project in the state of New Guinea.

**Farm’s size in hectares:** this column represents anINTEGER (20) data type to portray the size of the farmer’s farm by the number of hectares they have.

This table **Trainings** used to be linked with the farmer’s info table in this first part is shown below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CourseID** | **Course name** | **Description** | **Frecuency\_Month** | **Hours\_Month** |
| 9AI | Artificial insemination | Techniques to better use studs with high quality semen derived from top notch breeds | 2 | 4 |
| 6GM | Grassland management | improved grasses to optimize biomass yield | 3 | 8 |
| 7AN | Animal nutrition | Strategies to feed dairy cows specially during the dry season | 6 | 10 |

This table is called “Trainings” and the idea is to generate information about the type of trainings farmers have taken before

Moreover, I have also generated two more tables named **Farmersherd** and **Trainingsforfarmers**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **TraineeID** | **Steers\_num** | **Heifers\_num** | **Bulls\_num** | **Cows\_num** | **Tot\_herd\_num** |
| 1 | 3 | 5 | 1 | 6 | 15 |
| 2 | 12 | 10 | 1 | 25 | 46 |
| 3 | 180 | 240 | 10 | 70 | 500 |
| 95 | 3 | 4 | 0 | 6 | 700 |

In this last table **Trainingsforfarmers** as you can see below, I linked the information I have in my first three previous tables into a new integrated table (farmer’s info, trainings and farmer’s herd)

|  |  |  |  |
| --- | --- | --- | --- |
| **TraineeID** | **Training\_received** | **Next\_Course ID** | **Tot\_herd\_num** |
| 1 | 9AI | 3MPT | 15 |
| 2 | 6GM | 5T&R | 46 |
| 3 | 7AN | 8AR | 500 |
| 4 | 6GM | 4ABG | 470 |
| 5 | 8AR | 5T&R | 920 |
| 6 | 3MPT | 3MPT | 560 |
| 95 | 7AN | 8AR | 700 |

This table is normalized by utilizing the foreign keys from the first three databases such as Trainee ID, Training received , Next\_Course ID, Tot\_herd\_numwhich as the rest of the columns also have INTEGER data types. The purpose of normalizing the data is to have information that can show us with just a number/code personal information of every farmer, what kind of trainings they have received to better decide what are the courses that we should provide to all famers as well as it is important to track the number of the herd they have.

**3)** Either directly in sqlite or in Python with SQLAlchemy, write the code needed to define the tables above. (6pts.)

sqlite> **CREATE** **TABLE** `Farmersinfo` (

`TraineeID` integer **NOT** **NULL** **PRIMARY** **KEY** AUTOINCREMENT

, `Farm\_ful\_nam` CHAR(100) **NOT** **NULL** **DEFAULT** ''

, `Training\_received` CHAR(20) **NOT** **NULL DEFAULT** '0'

, `Geogra\_locat` CHAR(100) **NOT** **NULL** **DEFAULT** ''

, `Farm\_siz\_ha` integer **NOT** **NULL** **DEFAULT** '0'

, **CONSTRAINT** `Farmersinfo\_ibfk\_1` **FOREIGN** **KEY** (`Training\_received`) **REFERENCES `** Trainings**` (`**CourseID**`)**

);

sqlite> .schema Farmersinfo (to see the table’s command written above)

sqlite> **CREATE** **TABLE** `Trainings` (

`CourseID**`** CHAR(20) **NOT** **NULL** **PRIMARY** **KEY**

, `Course name` CHAR(100) **NOT** **NULL** **DEFAULT** ''

, `Description` CHAR(1000) **NOT** **NULL** **DEFAULT** '0'

, `Frecuency\_Month` integer **NOT** **NULL** **DEFAULT** ''

, `Hours\_Month` integer **NOT** **NULL** **DEFAULT** ''

);

sqlite> **CREATE** **TABLE** `Farmersherd` (

`TraineeID**`** integer **NOT** **NULL** **DEFAULT** ''

, `Steers\_num` integer **NOT** **NULL** **DEFAULT** ''

, `Heifers\_num` integer **NOT** **NULL** **DEFAULT** '0'

, `Bulls\_num` integer **NOT** **NULL** **DEFAULT** '0'

, `Cows\_num` integer **NOT** **NULL** **DEFAULT** ''

, `Tot\_herd\_num` integer **NOT** **NULL** **DEFAULT** ''

, **CONSTRAINT** ` Farmersherd\_ibfk\_1` **FOREIGN** **KEY** (`TraineeID`) **REFERENCES `** Farmersinfo**` (`**TraineeID**`)**

);

sqlite> **CREATE** **TABLE** ` **Trainingsforfarmers**` (

`TraineeID**`** integer **NOT** **NULL** **DEFAULT** ''

, `Training\_received` CHAR(20) **NOT** **NULL DEFAULT** '0'

, `Next\_Course ID` CHAR(20) **NOT** **NULL DEFAULT** '0'

, `Tot\_herd\_num` integer **NOT** **NULL** **DEFAULT** '0'

, **CONSTRAINT** ` **Trainingsforfarmers**\_ibfk\_1` **FOREIGN** **KEY** (`TraineeID`) **REFERENCES `** Farmersinfo**` (`**TraineeID**`) FOREIGN** **KEY** (`Training\_received`) **REFERENCES `** Farmersinfo**` (`**Training\_received**`) FOREIGN** **KEY** (`Tot\_herd\_num`) **REFERENCES `** Farmersherd**` (`**Tot\_herd\_num**`)**

);

**4)** Write the code to load the data into the database. (2pts) **Note this is only 2 points! Don't spend a ton of time on this. If it turns into a headache, don't worry about it.**

**5)** Find another dataset, briefly describe the dataset and columns. Briefly describe the tables you would use and how you would normalize these data. (2pts.)

**Goat kids’ birthday’s records from 2012 to 2015**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tag ID** | **Gender** | **Breed\_crossbred** | **Birthdaydate** | **Parturitiontype** | **Birthweight** | **Motherstagnumber** |
| 181 | Male | saanen/nubia | 2/13/12 | double | 6 lbs | 104 |
| 230 | Female | saanen | 2/13/12 | double | 5 lbs | 104 |
| 183 | Male | saanen | 2/19/12 | double | 7 lbs | 48 |
| 232 | Female | saanen | 2/19/12 | double | 6 lbs | 48 |
| 234 | Female | saanen | 2/19/12 | unique | 6 lbs | 148 |
| 185 | Male | saanen | 2/19/12 | unique | 7 lbs | 120 |
| 236 | Female | nubia | 2/19/12 | unique | 7 lbs | 136 |
| 187 | Male | nubia | 2/21/12 | unique | 5 lbs | 138 |
| 411 | Male | manch | 1/27/15 | double | 4.5 lbs | 170 |
| 416 | Female | nud/saan | 1/26/15 | unique | 2.5 lbs | 314 |

**Tag ID:** this column will collect information related to the kid’s ear unique tag number.

data type: INTEGER (20).

**Gender:** clearly stating the gender of the kids that will become the next does or bucks in the herd

data type: CHAR(10).

**Breed\_crossbred:** here I would like to display pure breeds as well as the existent crossbreeding

data type: CHAR(30).

**Birthdaydate:** this one stands for the date when the kids were born. Data type: date

**Parturitiontype:** it is common that this specie can have from one to even three babies in a single parturition. Data type: CHAR(20)

**Birthweight:** this column is also important to monitor the average weight gain per animal to select the future young generation of rams or goats as replacement depending on their performance. Data type: INTEGER(10)

**Motherstagnumber:** this is a foreign key from another data base called **Rams&Goats** the idea is to track the performance of the kids as well as to observe and select what are the most productive goats in the heard. For instance, which ones are having double or triple parturition, meat and milk yield for their offspring’s, and so on. Data type: INTEGER(15).