Connie Rodriguez
GEOG575
February 7th, 2020
Midterm

A brief report (~500 words) justifying why you chose the model that you did, while mentioning the challenges that you encountered, and how you overcame them. Also mention how much time you spent on this assignment (for my future reference). [15 points]

The original data was presented as two tables, banddata which describes the band observations and allpony which is an inventory of ponies. Given that SQL is a language used in relational database management, it was more appropriate to break the data into four different tables: Allpony2, Observations, FoalsProduces and Contraception. The UML diagram in Figure 1 shows the attributes and relationships for each table.

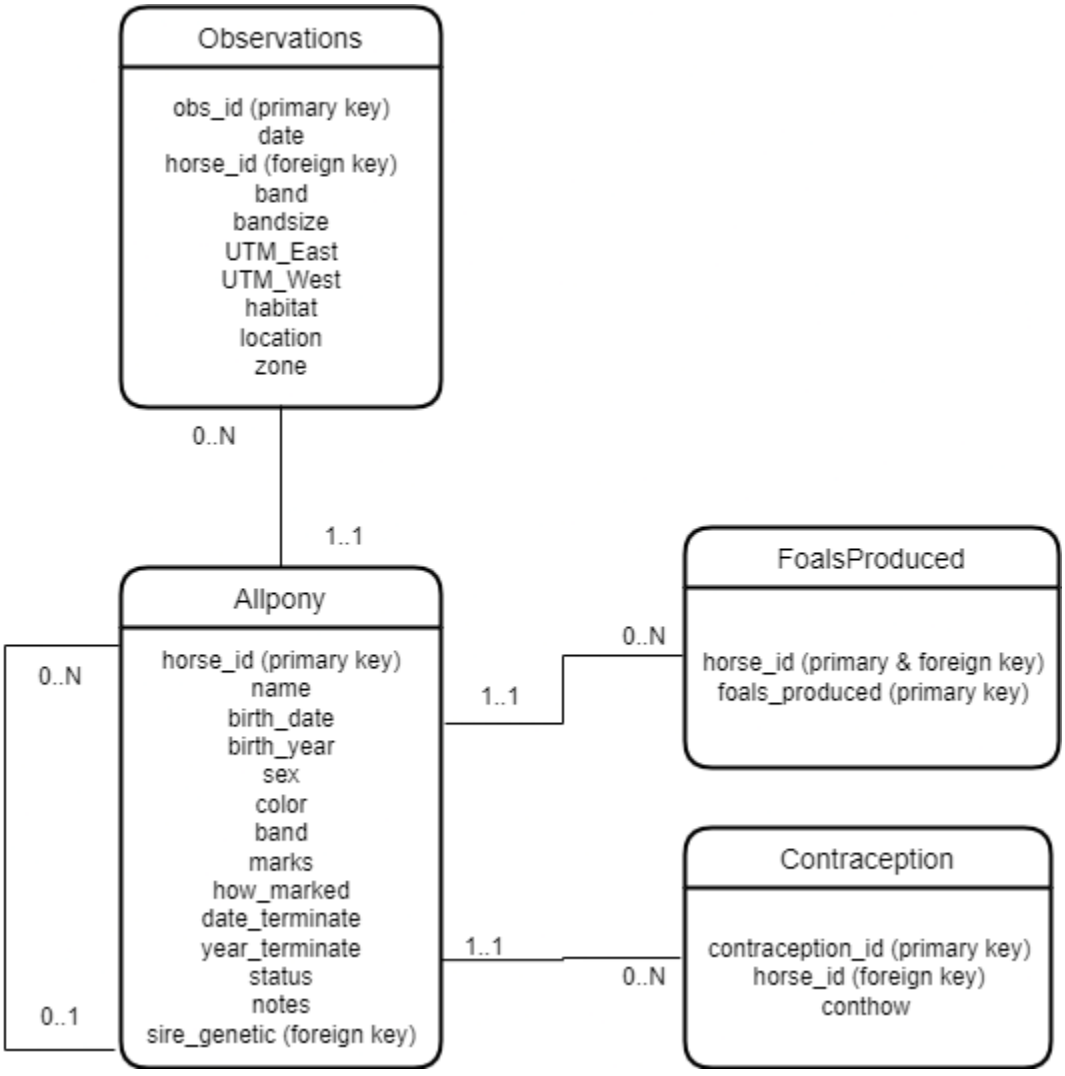

Figure 1. UML diagram showing the revised data model

The original allpony table was storing information about the horse, how often it received contraception and produced foals. The issue with the original table is that it has only one field per horse to store information about all the times that a horse received contraception or produced foals. By introducing the Contraception and FoalsProduced tables, each occurrence of an event is listed as one record which makes the data easier to analyze for the researchers. The FoalsProduced table uses a combined primary key using the horse_id and foals_produced attributes and relates to the allpony table using the horse_id attribute. In the Contraception table, it was necessary to create a synthetic key given that there was no combination of keys that would produce a candidate key. By removing the foals produced and contraception data, the Allpony2 table was left with only atomic level data for each cell. The sire_genetic functions as a self-referential foreign key which is a simplistic and efficient way to store this information in the same table.

The Observation table was derived from the banddata table. There was repeated information that are also present in the allpony table so those attributes such as sex and horse color were dropped from the Observation table because these data is also in the new Allpony2 table. The researchers are able to look up the horse information by using the horse_id from the Observations table to reference the correct horse record in the Allpony2 table.

The model conversion required some data type formatting given that some of the records such as birth_year in the allpony table and date in the banddata table, were stored as a different data type.

Perhaps the biggest challenge was that certain pony observations recorded ponies that were not originally in the allpony table causing an issue when creating and migrating data to the Observations table because horse_id was used a foreign key and the horse_key was not found in the Allpony2 table. This was resolved by first determining which ponies needed to be added to the Allpony2 table and then adding those horse_ids to the Allpony2 table. It is assumed that the researcher will take the responsibility of updating the remaining attributes for records that were added.

Another consideration was breaking the Observation table into two, one for observation events and another for horses observed that would relate to the observation events. Doing this would avoid data repetition since many horses could be observed in the same observation event so most of the data except the individual horse data would be repeated. However, given that this project has ended, there was no benefit in breaking the table further.

This assignment took me about 6 hours working on my own.