**Databases for Data Science**

**Term Project**

The first project consists of two parts. In the first part, you will create a database called “Portal Project Ecological Database.” This is a database from a scientific project for long-term monitoring and experimental manipulation of a Chihuahuan desert ecosystem near Portal, Arizona. Hence, they are real scientific data gathered over a long period of time. You will be installing this database in your SQLite database system instance that you have already created and will be using it for performing many data manipulation operations.

In the second part, you will create a database called “Northwind,” which consists of a schema for small businesses. This is a popular database used for learning about relational databases. You will use the SQLite system to install this database and perform data querying using SQL.

1. Portal Project: The data for the Portal Project can be found at the figshare site:

<https://figshare.com/articles/Portal_Project_Teaching_Database/1314459>

Create a folder called PortalProject, and download the Download All zip file (called 1314459.zip). Unzip it and you will see a set of .csv files in there as well as a .sqlite file.

I don’t want you to use the .sqlite file. Instead, create three tables (species, plots, surveys) using CREATE statements, and then upload the files in the .csv file into the database. Save the database in a file called YourName\_PortalProject.db, and upload that also as part of the project submission. You will also do some SQL queries and data manipulation as mentioned later.

1. Northwind: The data for Northwind is found in GitHub at the following location:

<https://github.com/jpwhite3/northwind-SQLite3>

Create a folder called Northwind, and then clone or download the zip file in there and unzip it. The website contains a READMe.md file, which shows the schema for the database. The file is also found in your directory as Northwind\_ERD. Unzip the Northwind\_large.sqlite.zip file. This contains the database that you want to upload into SQLite. In this case, you can open the .sqlite file as your database (no need for creating tables or uploading CSV files). You will do some SQL queries as mentioned next as part of the project submission.

**Queries to Be Answered**

Portal Project:

1. Find all genus and species of the rodent taxa.
2. Find all species (genus, species, and taxa) that were found by survey in 1987.
3. Print hindfoot length and weight of all species found in “Long-term Krat Exclosure.”
4. Find the dominant species in “Spectab exclosure.” What if you restrict to only 1998?

Northwind Project:

1. Find all suppliers who supply a beverage (category).
2. Find suppliers (print supplier companyname and region) who supplied both condiments and confections (categories).
3. Find all orders (print only orderId, orderdate, Product Name) by customer “Piccolo und mehr” through the employee named “Michael Suyama”.
4. Count the number of shipments handled by each shipping company. What if you group by year?

Portal Project:

CREATE TABLE species (

species\_id VARCHAR(2) PRIMARY KEY,

genus VARCHAR(20),

species VARCHAR(20),

taxa VARCHAR(20)

);

CREATE TABLE plots (

plot\_id INTEGER PRIMARY KEY,

plot\_type VARCHAR(20)

);

CREATE TABLE surveys (

record\_id INTEGER PRIMARY KEY,

month INTEGER,

day INTEGER,

year INTEGER,

plot\_id INTEGER,

species\_id INTEGER,

sex VARCHAR(1),

hindfoot\_length INTEGER,

weight INTEGER,

FOREIGN KEY (plot\_id) REFERENCES plots(plot\_id),

FOREIGN KEY (species\_id) REFERENCES species(species\_id)

);

1. Find all genus and species of the rodent taxa.

select distinct genus, species

from species

where taxa = "Rodent";

1. Find all species (genus, species, and taxa) that were found by survey in 1987.

select distinct p.genus, p.species, p.taxa

from species as p, surveys as u

where p.species\_id = u.species\_id

and u.year = 1987;

1. Print hindfoot length and weight of all species found in “Long-term Krat Exclosure.”

select distinct(s.species\_id) as species\_id, s.hindfoot\_length, s.weight

from surveys as s, plots as p

where s.plot\_id = p.plot\_id

and p.plot\_type = "Long-term Krat Exclosure"

order by species\_id;

1. Find the dominant species in “Spectab exclosure.” What if you restrict to only 1998?

select sp.genus, sp.species, sp.taxa, count(\*) AS num\_observations

from surveys as s, plots as p, species as sp

where s.plot\_id = p.plot\_id

and s.species\_id = sp.species\_id

and p.plot\_type = "Spectab exclosure"

order by num\_observations

limit 1;

select sp.genus, sp.species, sp.taxa, count(\*) AS num\_observations

from surveys as s, plots as p, species as sp

where s.plot\_id = p.plot\_id

and s.species\_id = sp.species\_id

and p.plot\_type = "Spectab exclosure"

and s.year = 1998

order by num\_observations

limit 1;

Northwind Project:

1. Find all suppliers who supply a beverage (category).

select distinct s.CompanyName

from products as p, categories as c, suppliers as s

where p.CategoryID = c.CategoryID

and p.SupplierID = s.SupplierID

and c.CategoryName = "Beverages";

1. Find suppliers (print supplier companyname and region) who supplied both condiments and confections (categories).

select distinct s.CompanyName, s.region

from products as p, categories as c, suppliers as s

where p.CategoryID = c.CategoryID

and p.SupplierID = s.SupplierID

and c.CategoryName = "Condiments"

and s.CompanyName in

(select s2.CompanyName

from products as p2, categories as c2, suppliers as s2

where p2.CategoryID = c2.CategoryID

and p2.SupplierID = s2.SupplierID

and c2.CategoryName = "Confections");

1. Find all orders (print only orderId, orderdate, Product Name) by customer “Piccolo und mehr” through the employee named “Michael Suyama”.

select o.OrderID, o.OrderDate, p.ProductName

from Orders as o, Customers as c, Employees as e, OrderDetails as od, Products as p

where o.CustomerID = c.CustomerID

and o.EmployeeID = e.EmployeeID

and o.OrderID = od.OrderID

and od.ProductID = p.ProductID

and c.CompanyName = "Piccolo und mehr"

and e.FirstName = "Michael"

and e.LastName = "Suyama";

1. Count the number of shipments handled by each shipping company. What if you group by year?

select s.CompanyName, count(o.OrderID) as numShipments

from shippers s, orders o

where o.ShipVia = s.ShipperID

group by s.CompanyName;

select s.CompanyName, strftime('%Y', o.OrderDate) as Year, count(o.OrderID) as numShipments

from shippers s, orders o

where o.ShipVia = s.ShipperID

group by s.CompanyName, strftime('%Y', o.OrderDate);