Step 1: Saved each spreadsheet as a CSV

- The CSVs are what was entered into the DB

- The CSVs are located in the asCSV folder. Many tables needed to be cleaned before they could be entered into the DB. Tables were cleaned in the clean.R script. Cleaned csv files are saved as adj<TableName>.csv. If a file had to be cleaned, then the csv entered in the DB for that table is the file that has the adj prefix in the asCSV folder.

- In the schema for the DB below, if you see additional notes for a given table, then it was cleaned in the clean.R script. These notes describe the cleaning that was done in Clean.R

Step 2: Schema Creation [In sqlite syntax]

AGENCY TABLE

CREATE TABLE AGENCY

( id INT,

Name TEXT NOT NULL,

City TEXT NOT NULL,

State TEXT NOT NULL,

PRIMARY KEY (id))

COMPANY TABLE

CREATE TABLE COMPANY

( id INT,

UnderwritingAgencyName TEXT NOT NULL,

City TEXT NOT NULL,

State TEXT NOT NULL,

PRIMARY KEY(id))

POLICY TABLE

\*Note that Sqlite3 does not have a date data type. Hence dates are text. [Yet sqlite3 does support date manipulation with strftime]

CREATE TABLE POLICY

( number INT,

StartDate TEXT NOT NULL,

EndDate TEXT NOT NULL,

AgencyID INT NOT NULL,

CompanyID INT NOT NULL,

CancelDate TEXT,

PRIMARY KEY(number),

FOREIGN KEY (AgencyID) REFERENCES AGENCY,

FOREIGN KEY (CompanyID) REFERENCES COMPANY)

DRIVER TABLE

\*Vehicle Id comes from the risk table. In R, I do a join between the driver table and the risk table to get the vehicle id from the risk table and put it into the driver table. The join variable is Driver.ID. Driver.ID is unique in both tables. Note that there are more drivers than risks [7,940]. Hence there are 7,940 vehicle Ids that are NULL in the driver table.

\*In driver table, many columns have empty values for MilesToWork and DOB. I changed these cells to NULL using R

CREATE TABLE DRIVER

(id INT,

Fname TEXT NOT NULL,

Lname TEXT NOT NULL,

Violations INT NOT NULL,

Accidents INT NOT NULL,

MaritalStatus TEXT NOT NULL,

Gender TEXT NOT NULL,

MilesToWork INT,

PrimaryVehicleUsage TEXT NOT NULL,

DateOfBirth TEXT,

VehicleID INT,

FOREIGN KEY (VehicleID) REFERENCES VEHICLE,

PRIMARY KEY(id))

VEHICLE TABLE

CREATE TABLE VEHICLE

(id INT,

Make TEXT NOT NULL,

Model TEXT NOT NULL,

AvgPrice INT NOT NULL,

PRIMARY KEY(id))

CLAIM TABLE

\*Assuming that claimant refers to a driver. Not 100% sure that this is true

CREATE TABLE CLAIM

(id INT,

PolicyID INT NOT NULL,

DriverID INT NOT NULL,

VehicleID INT NOT NULL,

Amount INT NOT NULL,

Description TEXT NOT NULL,

PRIMARY KEY(id),

FOREIGN KEY (PolicyID) REFERENCES POLICY,

FOREIGN KEY (DriverID) REFERENCES DRIVER,

FOREIGN KEY (VehicleID) REFERENCES VEHICLE)

LOCATION TABLE

Population Distribution and Location excel files were joined using the zip code. This was done in R, so that we don’t have to complete this join everytime we want to query population distribution information from SQL

CREATE TABLE LOCATION

( ZipCode INT NOT NULL,

id INT,

City TEXT NOT NULL,

State TEXT NOT NULL,

County TEXT NOT NULL

Population INT NOT NULL,

Percent0to15 REAL NOT NULL,

Percent15to25 REAL NOT NULL,

Percent25to40 REAL NOT NULL,

Percent50 REAL NOT NULL,

PRIMARY KEY(`id`) )

DRIVERVEH TABLE

The vehicle model year is oddly placed in the risk table. Instead of putting it here, I create a Driver-Vehicle table that associates a driver with a particular vehicle. The model year for the vehicle is stored in this table

CREATE TABLE DRIVERVEH

(

DriverID INT,

VehicleID INT,

VehicleModelYear INT,

PRIMARY KEY (DriverID,VehicleID),

FOREIGN KEY (DriverID) REFERENCES DRIVER,

FOREIGN KEY (VehicleID) REFERENCES VEHICLE

)

RISK TABLE

Model.Year has been taken out.

\*There was one cell with no value at all. I inserted a NULL

CREATE TABLE RISK

(

id INT,

PolicyID INT NOT NULL,

DriverID INT NOT NULL,

VehicleID INT NOT NULL,

LocationID INT,

Premium INT,

PRIMARY KEY (id),

FOREIGN KEY (PolicyID) REFERENCES POLICY,

FOREIGN KEY (DriverID) REFERENCES DRIVER,

FOREIGN KEY (VehicleID) REFERENCES VEHICLE,

FOREIGN KEY (LocationID) REFERENCES LOCATION)

Notes: In my asCSV folder:

If a sheet has an adj file, then the adj file was inserted into the DB. Otherwise, the file without the adj was inserted into the db.

In general: To load a file into the db

.headers on

.mode csv

.import <FILE> <TABLE>

\*For some reason, there was an issue with loading the DriverVeh table into the DB. So I had to manually delete the first row (which was the header) when using the above method for import