MJ Carnahan

Brittany Clemmons

Meagan Colley

ETL PROJECT REPORT

### Extract:

* We searched Kaggle to find our databases. We settled on 2 databases that observed honey production in the United States.One was just an average overview while the other controlled for pesticides used. It was a nice pairing because many columns overlapped.
* The data was presented in multiple csv files.

### Transform:

* The first step was to load the proper CSV files in a jupyter notebook and make them into Pandas dataframes.
* Next, we combined the CSV files from the same database. In order to execute this we used the ‘.append’ function from Pandas.
* We then edited the dataframes to include the same time frame. One dataframe included the years 1998-2007 and the other included years 1998-2012. So, we got rid of 2008-2012 in the larger data set using this line of code: “honey\_df = honeydf[honeydf['year']<2007]”
* The last step of transforming the data was to drop irrelevant columns, such as repeating ones, and rename columns so that they played nicely with SQL.

### Load:

* Although we originally intended to use SQL we actually settled on MongoDB because it seemed simpler.
* One error we ran into in SQL is that some of our column names (“state” for instance) actually called a function in postgres. By time we had fixed the problem by renaming our columns MJ had already figured out MongoDB.
* In order to successfully utilize MongoDB we imported PyMongo and from PyMongo we imported MongoClient.
* The next step was to set the connection and client.
* Next we named our database in Honey\_db in MongoDB, and named our collections honey\_db and honey2\_db.
* We then converted the dataframes into dictionaries in order for MongoDB to read them in. To do so we used this line of code (interchanging the names of the dataframes) “data\_dict = pesticidedf.to\_dict("records")”
* We the ninserted both collections into MongoDB using this line ‘collection2.insert\_one({"index":"state","data":data\_dict2})’
* Lastly, we imported the dataframes as CSV files which isn’t necessary when using MongoDB but we did so in anticipation of using SQL