**ETL PROJECT: ZOMBIE DEER APOCALYPSE**

**GROUP GIT REPOSITORY:** <https://github.com/sorlowek/ETL-Project>

**OVERVIEW:** Our team found data about recorded interactions with deer affected by chronic wasting disease from 2010-2013 in Canada and merged it with Canadian census data from 2006-2011. We thought that with further exploration of the data one could see if there is a relationship between the number of affected deer sightings and the density of human population in different areas.

**EXTRACT:** Using Google and [Google Dataset search](https://toolbox.google.com/datasetsearch) we found the data listed below. We downloaded the CSV and PDF files to our computer and loaded them into jupyter notebooks.

* **Deer Data**: <https://datadryad.org/resource/doi:10.5061/dryad.521ng>
  + CWD Surveillance Data from 2000 to 2010
  + CWD Surveillance Data in Alberta from 2011 to 2013
* **Census Data**:<https://www12.statcan.gc.ca/census-recensement/2011/dp-pd/prof/index.cfm?Lang=E>
  + Actual PDF data in git repository

**TRANSFORM:**

* Imported data files and renamed column titles so they would be uniform amongst each dataset
* Dropped unnecessary columns
* The “CWD Surveillance Data from 2011 to 2013” dataset containing information about the deer contained cartesian coordinates for where each deer was found. We found code that would transform these coordinates into the more usable Long Lat format.
* From a [second website](https://www.geeksforgeeks.org/python-reverse-geocoding-to-get-location-on-a-map-using-geographic-coordinates/) we found code that calculated address locations based on Long Lat coordinates.
* We then selected the province location of each deer from this dataset and added it as a column to the “CWD Surveillance Data from 2011 to 2013” dataframe.
  + We exported this dataset as a csv due to the length of time it took to run the code. We then commented out the original code and read the exported csv back into the jupyter notebook for subsequent runs of the code.
  + This involved a secondary transformation of the dataframe by dropping an “Unnamed: 0” index column.
* We then appended the CWD Surveillance Data from 2000 to 2010 to the transformed CWD Surveillance Data in Alberta from 2011 to 2013 datatset to create one dataframe containing all relevant deer information.
* The census dataset was manually manipulated in Excel as it was originally in a .pdf format and was not amenable to reading directly into the jupyter notebook.

**LOAD:** There were two steps main steps to loading the data frames into MySql.

* Step one was to create engine connection through sqlalchemy in the jupyter notebook
* Step two was to create the database, tables and query in MySql
  + We created a database called ‘deer\_db’
  + We created two tables for our the deer data and the census data called deercwd and census
  + Through a query we merged the two tables on ‘province’ and got the following final output:

|  |
| --- |
|  |