**Group Scaling-Octo-Waddle Project Report**

Brandon Alexander, Brian Pate, Esther Wanjiru, Forrest McDougal

**Extract**

* Landofamerica.com
  + We used beautiful soup and splinter to scrape each of the land listings on a webpage. This data included the price, zipcode, land size, and oftentimes more. We iterated through 1000 search result pages and obtained roughly 18,000 unique listings
* Openweathermap API
  + For each listing, if there was a zipcode provided, we then used the zipcode with the open weather map API by creating a function that returned basic weather information including temperature, wind speed, humidity etc.
* Census API
  + For each listing we again used the zipcode and passed that through a function that used a census API. From this we were able to get basic census data including information on population, median age, income level, etc.
* Data World CSV
  + From data world we found a CSV that contained information about a zipcode’s land including the total land area, number of housing units, percent of total county population within that zipcode, etc. We loaded this using pandas.

**Transform**

* Web Scraping
* Weather API
  + We studied the owm wrapper and found the appropriate way to use the API with zipcodes. From there we passed our API key in using a config file and wrote a simple function that took a zipcode as an input.
* Census API
  + Like the weather API, we simply had to study the wrapper and pass in our API key. We also had to study the different key IDs to find what data we wanted to use. We then wrote a simple function accepting a zipcode as the parameter
* Data World CSV
  + We loaded the CSV into a jupyter notebook and made it a dataframe. Then we set the zipcode as the index and wrote a simple function to get the land data when a zipcode was passed as a paramenter

**Load**

* For loading the data we created a local connection to mongoDB, created the dataframe, and started a for loop that iterated through all of the posts, calling each of our functions for each post. Once all of the data was obtained and in dictionary format the data was inserted into the database. The for loop then repeated and continued storing for 1000 pages.