

|  |
| --- |
| ETL Project  2020 |
|  |
| September 10  DU Bootcamp  Authored by: Curtis Caile, Dave Borowski, Chris Joncha and Brooke Crofts |

# ETL Project Introduction

|  |
| --- |
| *Extract, Transform, Load* – DU Bootcamp 2020 The main goals of this project were to use two data sources to extract, transform and load data scraped from the web and load it into a relational or non-relational database for further use. The final database should have tables or collections making the scraped data easy to query or create easy-to-understand visual outputs.  This report will describe the data used, the type of transformation chosen, relational production database used and the final tables produced. |
|  |
| **SUMMARY:**  The two sources of data chosen for the scrape were sailboatlistings.com and boattrader.com. HTML code for both websites was studied and a Jupyter Notebook was created with parameters set for *for\_loops* to parse through the websites and extract the title, price, length, year and description of boats for sale in and around the Denver area. Both sources of data scraped were then transformed into final production state using Python code in a Jupyter Notebook. After transformation, the data was then written to csv files and loaded as a collection in MongoDB.  **BODY:**   * Dependencies used: Beautiful Soup, requests, splinter, pymongo and pandas * Extract:   Splinter (chosen by the team for its ease of use by a Python jupyter notebook and easily let us automate browser action when visiting URL’s and interacting with them)/Beautiful Soup/requests.get, pandas, try-except blocks for attribute errors, import time for website mngmt, next button functions for multi-page scraping   * Transform:   For\_loops, list.append, pandas DF with joins/concatenation for column management and try/except blocks for parsing management   * Load:   Chose the flexible, document-oriented database, MongoDB and loaded both csv files as a collection.  **CONCLUSION:**  After some initial challenges with the extraction portion of the project, parsing through poorly written websites, we rose to the occasion and made it work. The transformation was made fairly easy with Jupyter notebook and then using the to\_csv function, the load was made smoothly into a MongoDB collection. |

Bibliography:

Craigslist.com

Boattrader.com

Acknowledgements:

Curtis Caile, Dave Borowski, Chris Joncha, and Brooke Crofts

DU Bootcamp Instructor Kevin Lee and teaching assistants

Appendices:

A screenshot of a cell phone

Description automatically generated