

|  |
| --- |
| 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 we chose to use, relational production database used and the final tables produced. |
|  |
| **SUMMARY:**  The two sources of data chosen for the scrape were Craigslist.com and boattrader.com. HTML code for both websites were studied and a Jupyter Notebook was created with parameters set for *for\_loops* to parse through the website and collect the title, price and link to boats for sale in and around the Denver area. The results from Craigslist.com were generated in list format and written to a collection in MongoDB. The results from boattrader.com were generated in list format and then placed into two Pandas dataframes and joined on the boats year, make and model. Both sources of data scraped were felt to be at final production, so they were combined into a collection within MongoDB.  **BODY:**   * Dependencies * 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 for column management   * Load:   Chose the flexible, document-oriented database, MongoDB and loaded as a collection.  **CONCLUSION:**  Initial problems we ran into during the Extract portion was when it was determined that sailboatlistings.com was a poorly written website and extremely hard to scrape the data with any efficiency. After more research, boattrader.com proved to be well written and easy to scrape so it became our second source of data. |

Bibliography:

Craigslist.com

Boattrader.com

Acknowledgements:

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

DU Bootcamp Instructor Kevin Lee and teaching assistants

Appendices:

Any graphs of diagrams we would like to use