ETL Project

Jasmeet Aujla and Andrew Maximos

Extract

The information was obtained by web scraping kijiji ads , particularly subaru impreza vehicles; utilizing selenium webdriver and options . Additional information tied to individual ads was scraped off carquery, and combined with the aforementioned information.

Transformation

All the information scraped off the kijiji website was placed within a pandas dataframe, any rows that were not functional were removed. The next step was adding the secondary sources of data in the same dataframe. Once again any rows that contained unfunctional data were removed from the dataframe; resulting in a shorter more concise dataframe.

Loading

Lastly the database that was utilized was mongodb, due to the nature of the information structure, and group members preference. A client was generated with the localhost (default port 27017) , database and collection were constructed ; the dataframe was transformed to a dictionary and pushed to the collection.