**ETL PROJECT TECHNICAL REPORT**

**Team members:** Misael Obregonand XuanCong Tran

**Project description:**

The project consists of creating a database that was combined from two sources. The first source was accessed through Kaggle.com named top\_selling\_car.csv. This source provides information on the highest selling used cars for year 2018. Data on this file include information such as brand, model, year and other variables. The table below provides data provided from the csv file and the second column are data used for this project.

|  |  |
| --- | --- |
| **Top\_Selling\_Car\_CSV from Kaggle.com** | **Data Used For ETL Project** |
| Ranking | Ranking |
| Car brand | Car brand |
| Car model | Car model |
| Release year | Release year |
| Models sold, in thousands |  |
| Maximum speed, km/h | Maximum speed, km/h |
| Percent of sells among the brand |  |
| Sales in 2018, in thousands |  |
| Price, $ |  |

The second source of data used is from Edmunds.com website. Using information from the above-mentioned file, such as car brand, model, year, and ranking, this information was extracted and loaded into pandas jupyter notebook for the code and for scraping of data (second source) from the Edmonds.com website.

For the code we used SQLAlchemy, BeautifulSoup, Splinter, and PostgresSQL to scrape and join the databases.

**2. Data Cleanup & Analysis:**

Initial cleanup of data was to use only needed data from the original csv file on list of cars (see above). The data is then used to look up and scrape information from Edmunds.com on the list of cars. Using python code (in Pandas environment) was created a search for each car on the list provided the source for new data which included model, description, vin, mileage, and price on each car found on the website.

Any insufficient information (such as none, null, missing) on car searches is not collected and is part of a function built into the code to avoid missing data. Once all data have been scraped, the new data is appended, and a new data frame is created in Pandas.

**3. Storing the data:**

Final step is to store data. We connected to a local database using PostgresSQL to create the tables for new database. Then Pandas is used to load dataframe created into the new SQL database. Once tables are loaded and new database are created, we can join the tables and query the data as needed in PostgresSQL or in Pandas as needed.