ETL and Vehicles Statistics

Ashley Sligh, Amanda Jacobson, Charlotte Coolidge, Kyle Conway, Rajiv Manglani

Vehicle production is at an all time low and prices for used cars are at an all time high. Our analysis will include and compareused car listings containing car VIN number, price listing, fuel type, etc. with a fuel efficiency analysis dataset, in order to compile a one stop shopping database withthe intent to look for the best vehicle value.

Source for fuel economy in ny state  : <https://data.world/city-of-ny/mn2p-34if> This is a report of city vehicles and actual MPG compared to EPA estimated MPG. Each line of data is a combination of all the active vehicles on the city’s telematics system broken down into year/make/model/standard type with fueling and usage data. The intent is for each line to represent the sticker MPG and the real-world MPG and how these compare to each other. The report can be found at <https://www1.nyc.gov/assets/dcas/downloads/pdf/fleet/NYC-Fleet-Newsletter-306-May-27-2020-Hybrids-Work-Even-Better-in-Reality-Than-in-Theory.pdf>.  
Source: <https://data.cityofnewyork.us/d/mn2p-34if> Last updated at <https://data.cityofnewyork.us/data.json> : 2021-11-30 Attribution is encouraged by New York City. Please visit New York City's open data portal for more information about the Terms of Use.

Source for used car: <https://www.kaggle.com/rupeshraundal/marketcheck-automotive-data-us-canada>  
Context MarketCheck's automotive data covers 8 years of inventory across the US and Canada. Each day our systems crawl and aggregate inventory from over 65k dealer websites to deliver the most comprehensive and up-to-date depictions of market activity available anywhere.  
Individual listing records show year, make, model and trim, with VIN-level histories, showing the most recent time the car showed up online back to the earliest, with every change that occurred over that time.  
Equipment breakdowns give fuel type, engine size, transmission, color, driveline and body style with links back to the VDP of the source we obtained the information from. The complete data dictionary can be found here - <https://storage.googleapis.com/marketcheck-sample-feeds/cars_data_dictionary.xlsx> In addition to our car inventory data, Marketcheck also provide several other powerful APIs for CRM marketing scrubs, basic VIN decoding, OEM build data, recalls, dealer profiles, and market activity.

One of our first steps was to manually manipulate the used-car dataset in order to parse the raw amount of data. We selected a shorter span of years in Excel, decreasing the size of our dataset from 1 million rows to approximately 405 thousand rows. We also limited the used sales listings to the makes and models contained in the fuel efficiency dataset. From our fuel efficiency dataset we deleted the “vehicle count” column which was determined to be extraneous but kept all other columns for eventual comparison.

We then loaded our csv’s into pandas dataframes to prepare for the data cleaning process. Prior to merging the dataframes, we needed to manipulate the columns in order to match the cases of make and model to ensure a successful merge. We then merged the dataframes Pandas, setting the index for each dataframe on the values ‘year,’ ‘make,’ and ‘model’ prior to performing the merge in order to join the dataframes despite the year/make/model were not unique values. Once merged, we utilized the drop column function to eliminate unnecessary data (our used car listings file had 21 columns, not all of which were necessary for our database.) Additionally, we made our join ‘inner’ so as to cut out duplicate versions of the same make, model, trim, etc. We saved our cleaned dataframe into a csv so as to easily collect all of the cleaned columns to populate our analysis ready table in SQL.

Our final result was a SQL database that could be manipulated by make, model, and year to easily compare the automobile statistics on a given vehicle type’s used car.