Chandler ETL project Report

Extract: Our original data sources were demographic car data from <https://www.kaggle.com/minaboulos/fuel-economy> and car sales data from <https://www.kaggle.com/dmi3kno/newcarsalesnorway?select=norway_new_car_sales_by_model.csv> . Both sources were csv’s.

Transform (Specific to my Part the Norway sales data): There was an issue initially with hidden characters that started with the csv upload requiring an additional encoding='latin1’ added to the read in. I also singled and edited out all hidden characters with a .str.lstrp(). Next I used a .str.rsplit() to remove the car make from the car model column. It already exists in the make column and will allow me to link up with my partners data on a compound key later. Next I used a group by to aggregate my data from a monthly, to a yearly level. I then transformed my data by building out a sales rating column that used the quantity of sales to grade the success for a years sales per make and model as high, medium, or low.

Load: After creating a table query in postgres to model my pandas data as well as my partners I utilized SQL Alchemy to load my data into postgres. Our compound key was chosen based on the two tables same unique values. (Model and Year). We chose this data and this direction to set ourselves up for an analysis involving what car characteristics drive car sales in Norway and how that may evolve over time.