Aaron Olson

Artur Dzhalalov

**“US used car sales data” data set**

**Description**

For our project we used the “US used car sales data” used car dataset from Kaggle. This dataset was created by Kaggle contributor with user name “tsaustin” a.k.a TS. Its final version was uploaded on November 16, 2020. It is represented by a single file in csv format, and it is approximately 13 Mb in size. The single file contains 122145 rows representing car sales in the United States over a period of 20 months in 2019 and 2020. The aforementioned data set has a very low percentage of missing data for 9 columns out of 13 e.g. only 909 rows are missing zip code values, which means only 0.007% rows can be considered incomplete. However, four columns “trim”, “engine”, “body type”, and “drive type” are missing 40, 22, 17, 20 percent correspondingly. Given the size of the data set, we think that using data from columns with approximately 20% data missing can still yield reliable results.

**Source**

<https://www.kaggle.com/datasets/tsaustin/us-used-car-sales-data>  
**Schema**

"ID", (Primary key), "pricesold", "yearsold", "zipcode", "Mileage", "Make", "Model", "Year", "Trim", "Engine", "BodyType", "NumCylinders", "DriveType"

**Examples**

"137178","7500","2020","786\*\*","84430","Ford","Mustang","1988","LX","5.0LGas V8","Sedan","0","RWD"

"96705","15000","2019","81006","0","Replica/Kit Makes","Jaguar Beck Lister","1958","","383 Fuel injected","Convertible","8","RWD"

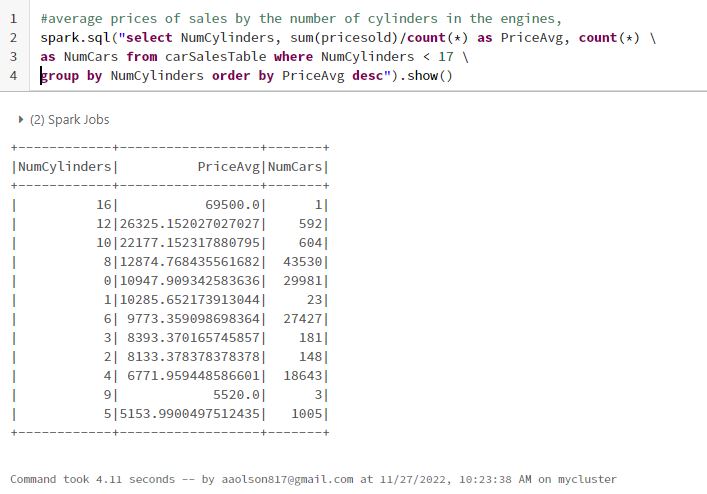
"119660","8750","2020","33449","55000","Jaguar","XJS","1995","2+2 Cabriolet","4.0L In-Line 6 Cylinder","Convertible","6","RWD"

**Questions**

Questions we are hoping to answer analyzing this dataset

* How much price difference is there between 4, 6 and 8 cylinder engine cars?
  + Is more engine power more expensive, and if so by how much?

Result

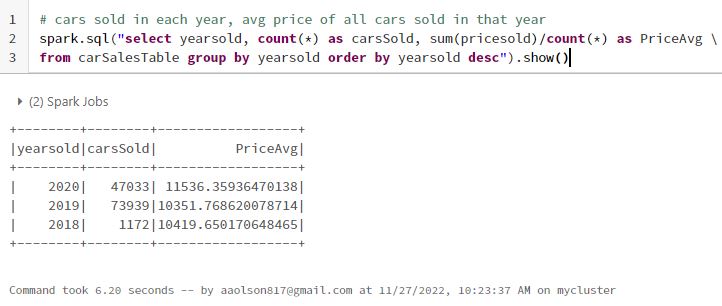


**Results**

We decided to look at all common engine types and compare all of their prices. More power costs more.

There is a significant cost increase from 8 -> 10 (~100%), 6 -> 8 is a good jump too (~25%)

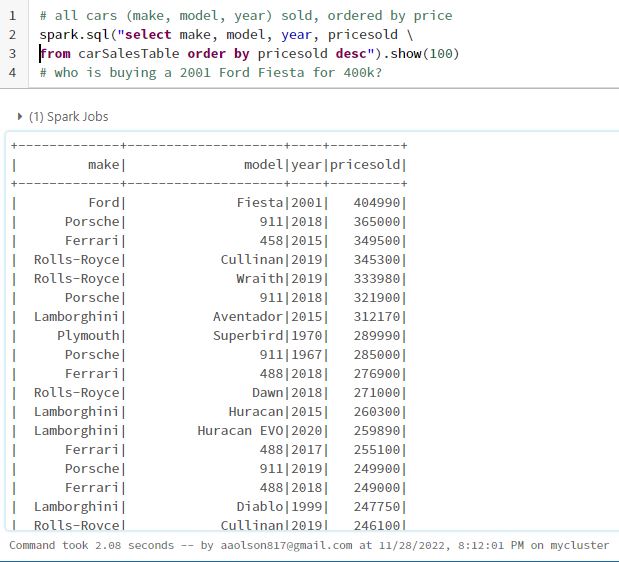
* In what years were the most cars sold? By number of cars.
  + How did COVID pandemic affect car sales?



**Results**

There were not as many cars sold according to this dataset, and the average price was over 10% higher in the year 2020 vs 2019.

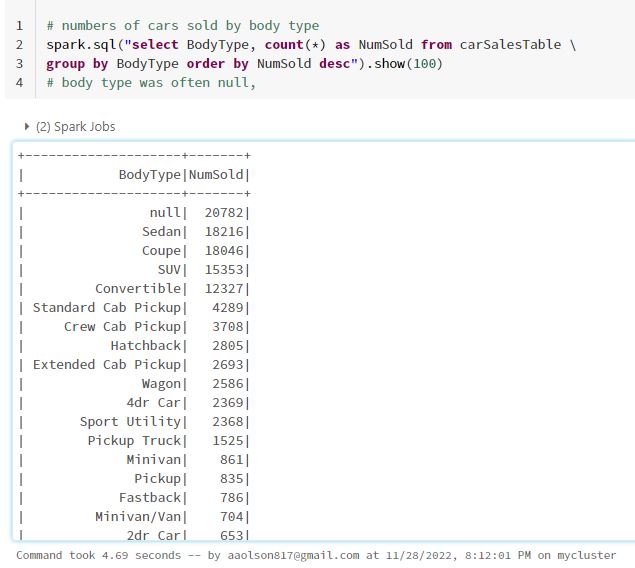
* **What are the most expensive cars? Ranked from most expensive to least**
  + This could be helpful information for car collectors and luxury car enthusiasts.



**Results**

The most expensive cars were high end sports cars, luxury cars, or old collector cars, with the exception of the Ford Fiesta. These cars seem to hold or increase their value over time.

* **Most popular cars sold by vehicle type e.g. SUV, Sedan, Truck, etc.**
  + Valuable input for car sales teams which will help them to maintain their inventory more efficiently.



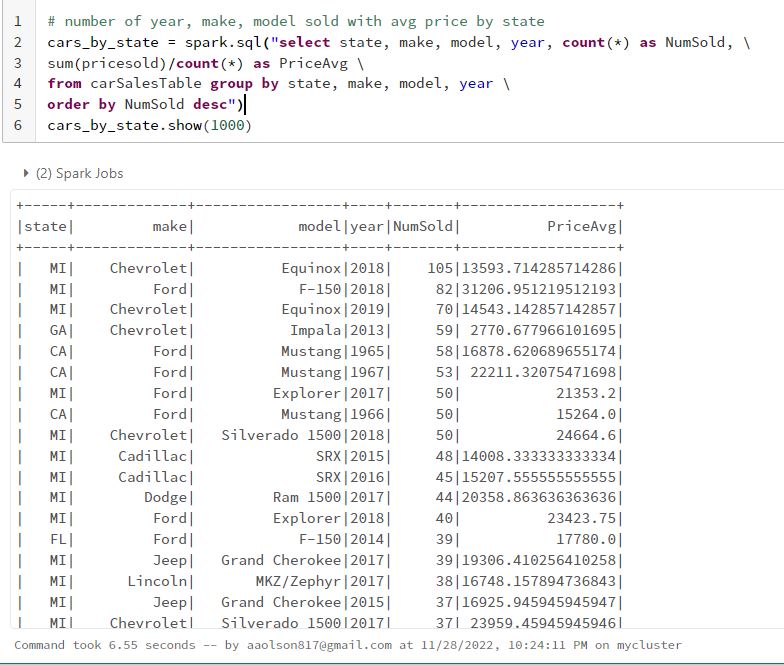
**Results**

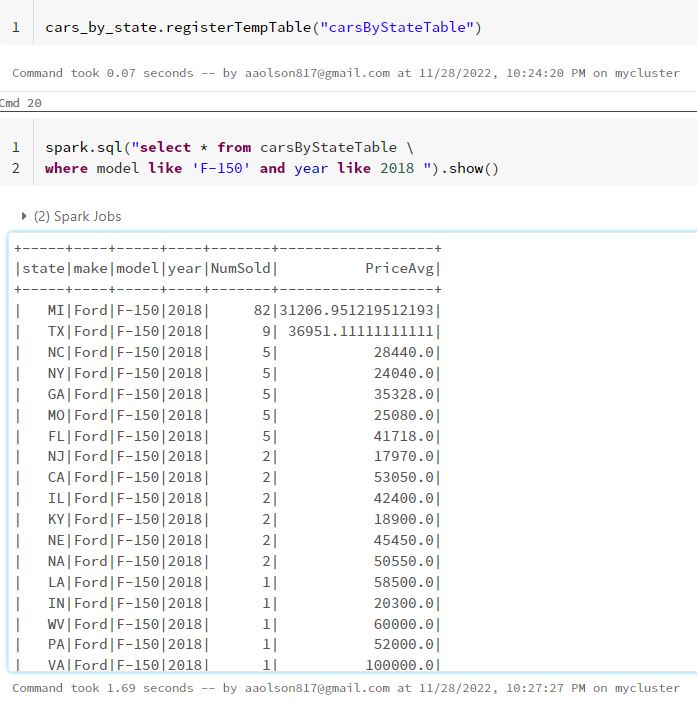
Sedans were the most popular. There were quite a few (20782) cars sold where the body type was not specified. Not listing the body type on the sale is fairly common

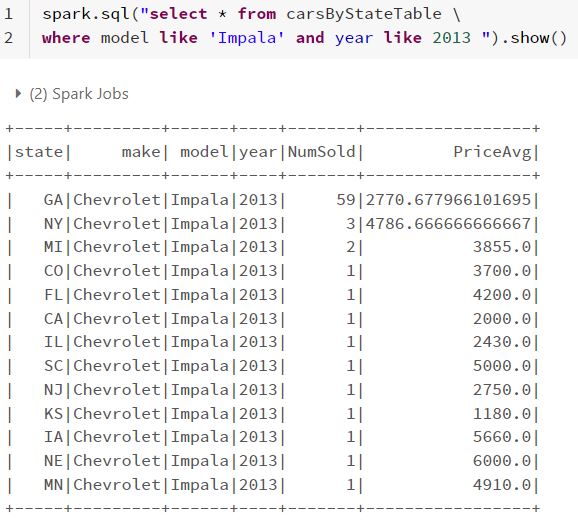
There are many non-standard body type descriptions

Sedans, Coupes, SUVs, Convertibles are very popular

* **Price of a same car model by the State?**
  + This could be helpful for finding bargains





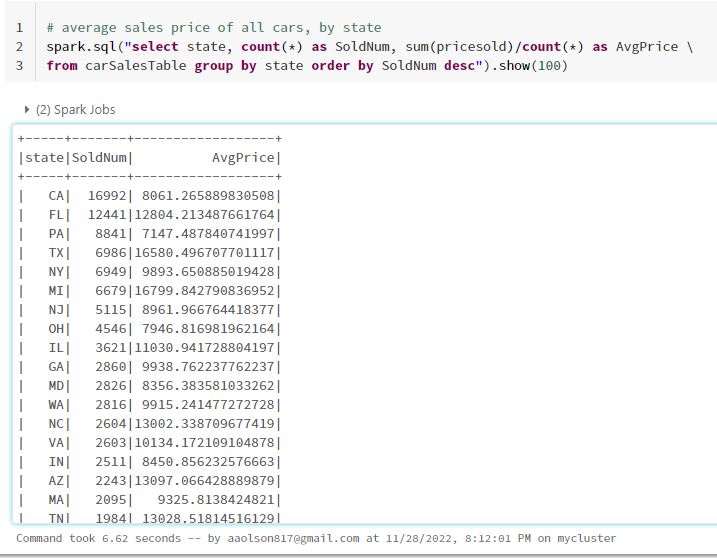


**Results**

I included just two examples. This particular part of the dataset has a lot of information to explore.

Certain make/model/years seem to cluster heavily in certain states in this dataset. Going to certain states for certain car models could be advantageous.

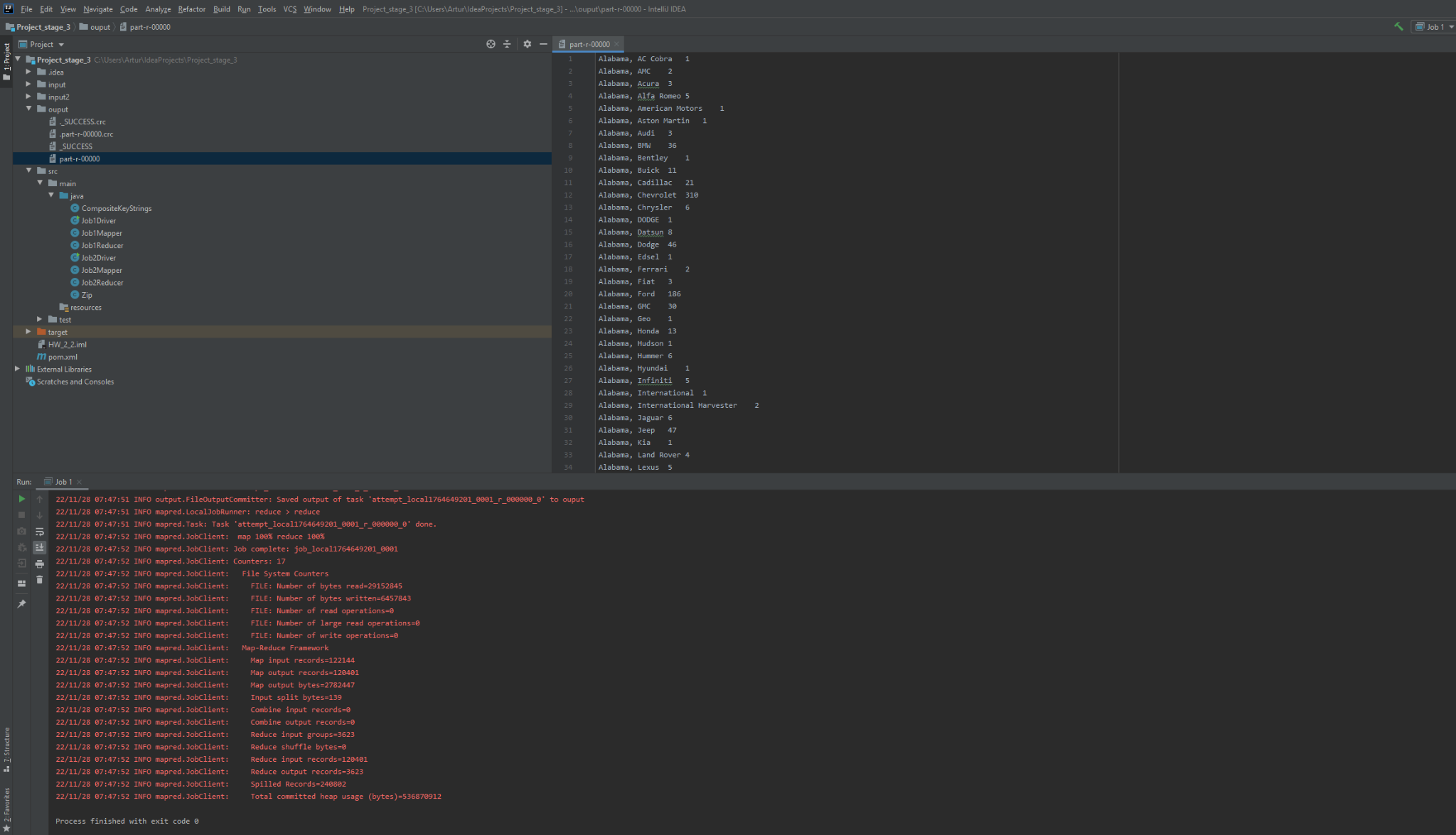
* **Number of cars sold by state and average price**
  + Demographic information about financial well being of country’s population by state

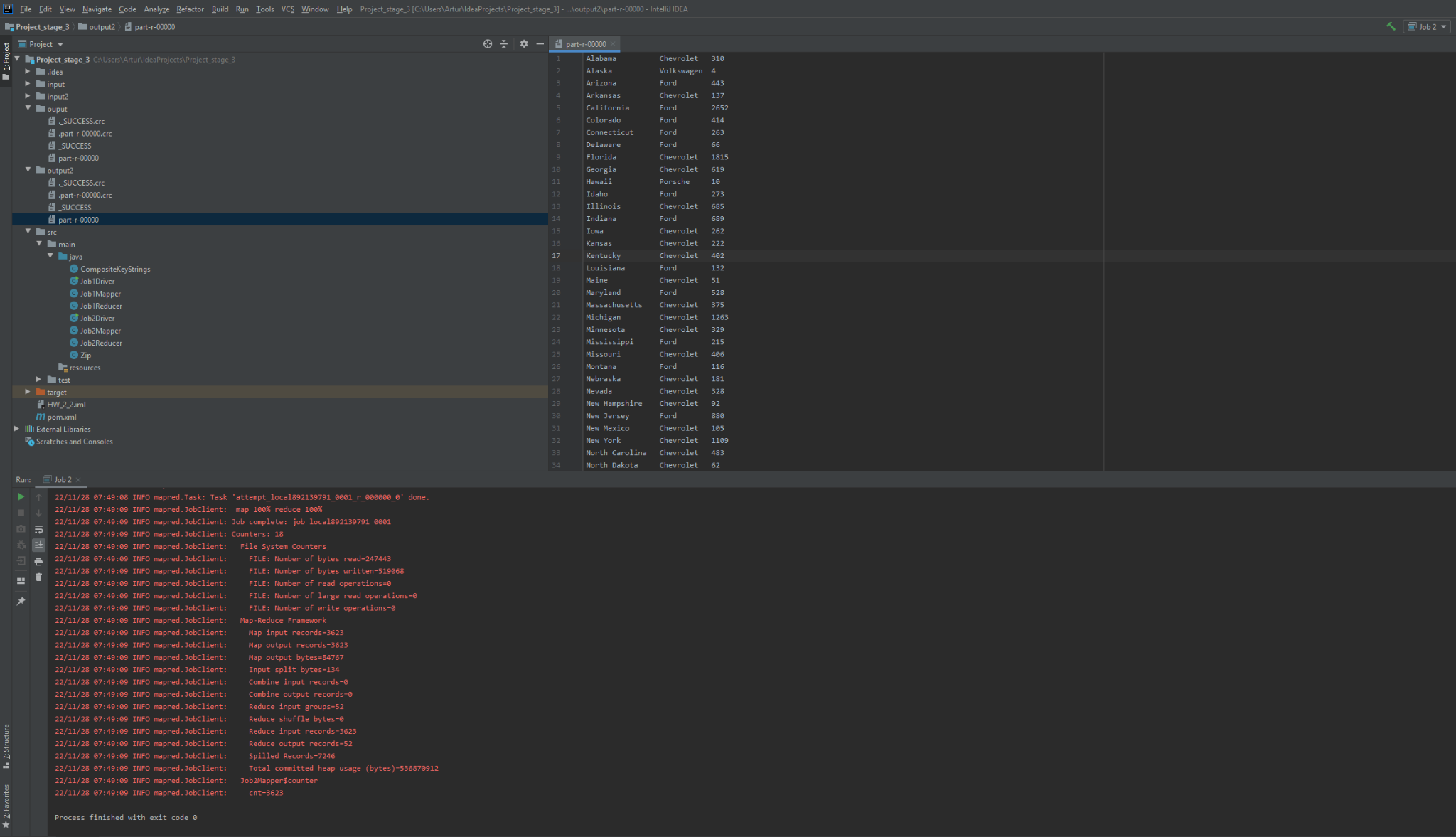


**Results**

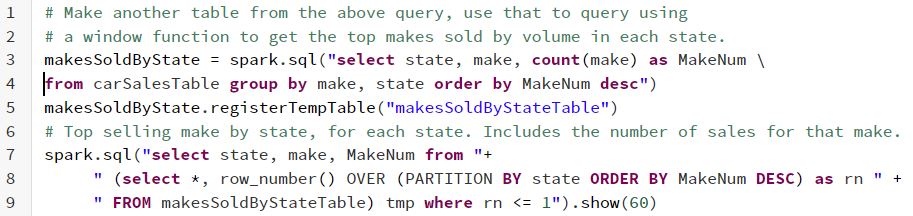
FL, TX, and MI have a lot of expensive (relatively) used car sales volume. Going from 1 state to another could result in a price delta of ~$5000. Given the results from the cars by state, going to certain states for certain models could be advantageous.

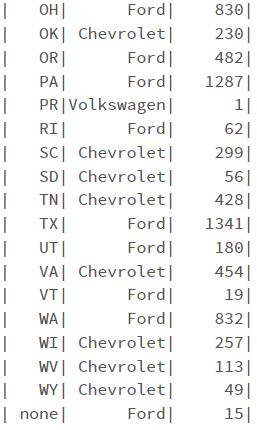
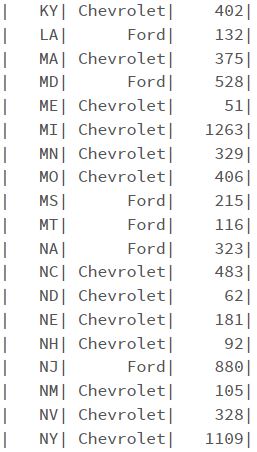
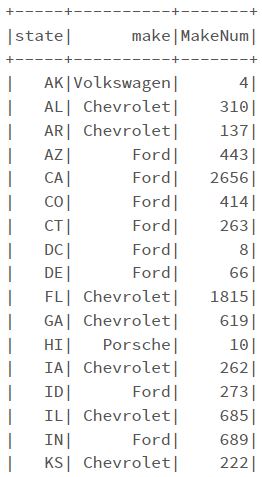
* **Number of cars of the same manufacturer sold in each state**
  + What car manufacturer is the most preferred by state and nation wide





This calc was done both in pyspark and Hadoop. The results were the same, but it was easier to view in spark.

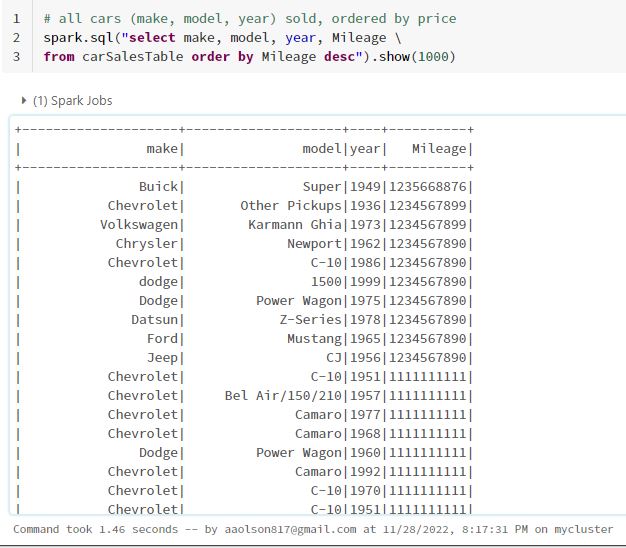




**Results**

Populous states do a lot of sales volume. Michigan also does a lot of volume in comparison to the state population, at least according to this dataset. Might be due to the large amount of manufacturers in the state.

* What cars are still being sold on the market with the highest mileage, grouped by manufacturer
  + What cars and what manufacturers have and produce the longest lasting engines.



**Results**

These numbers do not seem real, not sure they can be believed. Due to inaccuracy, it is unclear how useful this is.

**Appendix**

Source Data - <https://www.kaggle.com/datasets/tsaustin/us-used-car-sales-data>

Source Code - <https://github.com/aaronjolson/ICS-613-bigdata>