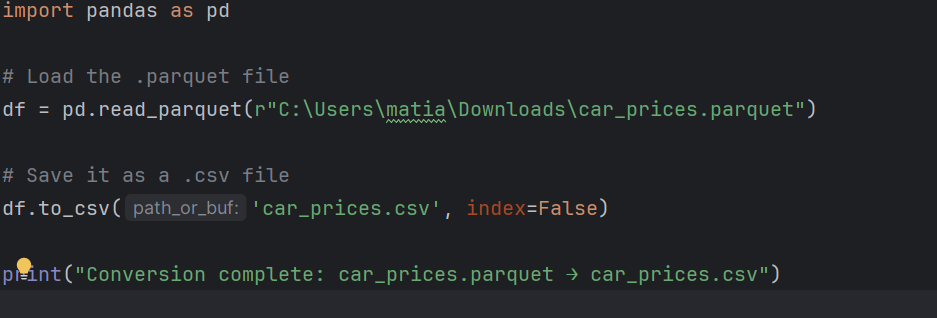
I created external table for bank.csv file and used second approach for that file. And I just turned car\_prices\_parquet file into a csv file. And just read data from there.

This is the python code of how I turned it into .csv file:



Now I will answer the questions:

Section 1-Compare car\_prices.parquet with car\_prices.csv file (size). Briefly describe what is causing such a difference.

Answer:

When we compare the two versions of the same dataset, there's a huge difference in file size. The .parquet file is only 17 KB, while the .csv file is 88 MB. At first glance, this might seem surprising, but it's actually expected due to how these file formats work.

The .csv format is plain text and stores all the data row by row, without any built-in compression. Every value, even if it repeats across thousands of rows (like car makes or models), is written out in full. This adds up quickly and leads to a large file size.

On the other hand, the .parquet file is a modern, columnar storage format. It stores data column by column and compresses it very efficiently. Repeated values are grouped and stored more smartly, and the format also understands data types, which helps reduce size even further.

So, even though both files contain the same data, the .parquet version is way smaller because it's optimized for storage and performance, especially when working with large datasets.

Section 2-Describe bank and car\_prices dataset in several sentences, what information is contained in the datasets and how it could be used for business.

Answer:

The bank.csv dataset contains customer information collected during a marketing campaign by a bank. It includes details like age, job, marital status, education, account balance, loan history, and whether the customer subscribed to a term deposit. It also logs campaign-related data, such as the contact method, number of contacts, and the outcome of previous campaigns.

This dataset can help the bank analyze customer profiles, segment audiences, and evaluate the effectiveness of marketing strategies. It can also be used to build predictive models to identify which types of customers are more likely to respond positively to future campaigns.

2.car\_prices.csv Dataset:

The car\_prices.csv dataset provides detailed information about used car listings. It includes fields like the make, model, year, mileage (odometer), condition, body type, transmission, color, and price. It also contains market-related data such as MMR (Manheim Market Report value), sale date, and VIN numbers.

This dataset can be useful for businesses in the automotive industry, such as dealerships or online marketplaces. It can support pricing analysis, inventory management, demand forecasting, and help in identifying trends in customer preferences based on car features and market behavior.

3-

While examining both datasets, I came across several data issues that should be flagged and cleaned before any serious analysis.

Dataset: bank.csv

This dataset contains information from a marketing campaign,mostly personal and financial details of customers.

Anomaly 1: Inconsistent categorical values

In column education there are multiple versions of what appears to be the same category. For example, both "basic.4y" and "basic.6y" exist, which likely refers to different levels of basic education.

Anomaly 2: Questionable values in default column

Some rows have unknown values instead of yes or no.

Dataset: car\_prices.csv

Anomaly 1: Illogical year values

Some values are unrealistically old (like before 1950), and some even look like future years.

Anomaly 2: Text in numeric fields

Some entries contain non-numeric characters or are empty. mileage = 'N/A' or price = 'unknown'