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

File Size Comparison

After converting the original car\_prices.parquet file to car\_prices.csv using Python with library pandas, we observed a significant difference in file sizes:

* **car\_prices.parquet:** 17 313 KB
* **car\_prices.**csv: 90 328 KB
* The .csv file is approximately 5 times larger than the .parquet file.

Reasons for the Size Difference

The difference in file sizes is due to the nature of the file formats:

* **Parquet** is a **columnar storage format** that stores data by columns rather than rows. It is designed for **efficient data compression** and performance optimization, especially for analytics workloads.
* **CSV** is a **plain-text, row-based format**, which:
  + Repeats field names in each row.
  + Does **not compress data**.
  + Stores all data as **strings**, even for numerical and date values.

As a result, Parquet files are much **more compact**, **faster to read**, and better at preserving **data types**, which makes them highly suitable for data profiling, machine learning, and large-scale data analysis.

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

**bank.csv Dataset**

The bank.csv dataset contains information collected from a direct marketing campaign by a bank. The dataset includes attributes such as:

* **Client details** – age, job, marital status, education, etc.
* **Banking information** – balance, housing loan, personal loan.
* **Campaign interaction** – contact type, duration.

**Business Use Case:**

* Analyze customer behavior and preferences.
* Improve targeting in marketing campaigns.
* Identify customer segments likely to respond positively.
* Build predictive models to increase conversion rates.

**car\_prices.parquet Dataset**

The car\_prices.parquet dataset contains records of car sales and includes data such as:

* **Vehicle details** – make, model, trim, color, year, VIN, condition, body type, transmission, odometer, interior.
* **Market-related info** – selling price, location (state).
* **Valuation** – various MMR (Manheim Market Report) prices for auction and trade-in values.
* **Sale metadata** – sale date and seller details.

**Business Use Case:**

* Pricing analysis and market trend forecasting.
* Valuation of used vehicles based on condition and history.
* Inventory management for car dealerships.
* Identifying regions with high sales volume or price fluctuations.

**Section 3 – For each dataset find and list data anomalies that should be raised. Add brief description of each data anomaly (like dataset, attribute, issue description, example of corrupted values, how did you find it, why you think that this data is incorrect and etc).**

1. bank.csv Dataset (analysis with DBeaver)

Anomaly 1: Invalid Job Entries

* Attribute: job
* Issue: "unknown" is not a valid job category.
* Example: job = "unknown"
* Detection SQL code:

SELECT job, count(\*) as count

FROM bank

WHERE job = 'unknown' or job is null

GROUP BY job;

* Impact: Reduces accuracy in client segmentation.
* Root Cause: Missing data during collection or system errors

Anomaly 2: Age Outliers

* Attribute: age
* Issue: Values >90 and values<16 years unlikely for banking products
* Example: age = 95
* Detection code:

SELECT \* FROM bank WHERE age > 90 or age < 16;

* Impact: Skews marketing ROI calculations.
* Root Cause: Data entry errors or system defaults

Anomaly 3: Mismatched Contact/Duration

* Attribute: contact, duration
* Issue: Illogical combinations of contact status and call duration
* Example: contact = "unknown", duration = 226
* Detection code:

SELECT age, job, contact, duration

FROM bank

WHERE

(contact = 'unknown' and duration > 0)

or

(contact = 'unknown' and duration = 0)

or

(contact != 'unknown' and duration = 0);

* Impact: Distorts call center metrics and campaign effectiveness analysis
* Root Cause: System synchronization issues or incomplete data recording

2. car\_prices.parquet Dataset (analysis with Python)

Anomaly 1: Missing Values in Key Columns

* Attributes: make, model, sellingprice
* Issue: Some entries have null or missing values in important fields like make, model, or sellingprice.
* Example: make = None, sellingprice = None
* Detection Method: df.isnull().sum() and profiling report.
* Impact: Prevents accurate pricing analysis and inventory management
* Root Cause: Extraction errors or incomplete data sources

Anomaly 2: Negative or Zero Prices

* Attribute: sellingprice
* Issue: Some records show selling price ≤ 0, which is invalid.
* Example: sellingprice = 0
* Detection Method: Conditional filtering: df[df["sellingprice"] <= 0]
* Impact: Skews market price analysis and valuation models
* Root Cause: Data entry mistakes or system placeholder values

Anomaly 3: Sale Date Formatting Issues

* Attribute: saledate
* Issue: Some saledate entries may have inconsistent formats or invalid dates.
* Example: saledate = "32/13/2022".
* Detection Method: Attempting to parse with pd.to\_datetime()
* Impact: Prevents proper time-series analysis and trend detection
* Root Cause: Multiple source systems with different date formats