# **DQ\_Introduction\_Fatkhullakh\_Turakhonov**

## **Section 1 – Parquet vs CSV comparison**

* **car\_prices.parquet** size: ~17 MB
* **car\_prices.csv** size: ~90 MB
* **Reason:** Parquet is a columnar format with built-in compression and encoding. It stores data more efficiently than CSV, which is plain text and row-oriented. That’s why the CSV version is about 5× larger.

## **Section 2 – Dataset descriptions**

### **bank.csv**

This dataset contains banking and demographic information about clients, such as age, job, marital status, loan history, and campaign response.  
 **Possible business uses:**

* Customer segmentation (e.g., by age, marital status, occupation).
* Marketing effectiveness analysis (response rates by demographic).
* Creditworthiness and risk analysis.
* Churn prediction and retention campaigns.

### **car\_prices**

This dataset contains historical car listings, including make, model, year, trim, body type, transmission, mileage (odometer), condition, color, interior, seller, wholesale MMR price, final selling price, and sale date.  
 **Possible business uses:**

* Market price analysis and valuation of used cars.
* Residual value modeling for leasing.
* Fraud detection (spotting impossible prices or odometer readings).
* Inventory optimization and demand forecasting for dealers.

## **Section 3 – Data anomalies**

### **bank.csv (analyzed manually in Excel)**

* **Negative or zero ages** → invalid customer records.
* **Blank job or marital fields** → missing critical demographic info.
* **Category typos** (e.g., Maried instead of Married) → need normalization.
* **Inconsistent flags** (e.g., default=yes but balance very high) → possible data entry issue.
* **Duplicate clients** (same id/email/phone) → needs deduplication.  
   *(Insert Excel filter/sort screenshots here.)*

### **car\_prices (profiling with ydata-profiling)**

* **Missing values:**
  + make ~1.8%, model ~1.9%, trim ~1.9%, body ~2.4%, transmission ~11.7%, condition ~2.1%.
* **Invalid categories:**
  + transmission contains nonsense values (horse-driven, sedan, Sedan). These belong in body, not transmission.
  + interior includes “—” instead of NULL.
* **Class imbalance:**
  + 90% of cars are automatic transmission → may bias models.
  + Over 50% interior = black/gray.
* **Outliers / invalid numbers:**
  + condition max = 982 (expected 0–5 scale).
  + sellingprice min = 1 → invalid.
  + odometer = 0 for many used cars → unlikely unless new cars.
* **Correlations:** suspicious high correlation between color and transmission, likely caused by mis-labeled categories.
* **Keys:** VIN should be unique, but ~1.5% duplicates → repeated sales or data duplication.
* **Dates:** saledate is text, not datetime, 12 missing values. Needs type conversion.  
   *(Insert screenshots from profiling HTML: Alerts, Missing values, Variable stats.)*

# **Conclusion**

Both datasets contain useful business information but show typical data quality issues: missing values, invalid categories, typos, and outliers. These need to be cleaned and standardized before downstream analytics or machine learning can be reliable.