

| Business Template  **Introduction to Data Quality** |
| --- |
|  |

1. DATA PROFILING

As you already know, there are various activities and areas we cover in data quality (DQ). One of the key actions we take to familiarize ourselves with the data and identify dependencies or insights for our future DQ framework is Data Profiling.

Data profiling helps us understand the structure, patterns and anomalies within the dataset, enabling better decision-making in data validation and quality checks. This process can be performed using various tools and Python libraries, such as pandas-profiling, ydata-profiling, AWS Deequ, Great Expectations and other, alongside manual SQL analysis to extract meaningful insights.

In this task, you will apply data profiling techniques to analyze your dataset, uncover dependencies, and gather useful observations for improving data quality be based on bank.csv and car\_prices.parquet files.

You need to analyze these files using 2 of 3 approaches listed below:

1. Manual reading of data in Excel.

2. Uploading file as External table in the DBeaver and then analyzing using SQL.

3. Using ydata-profiling library

Choose the most suitable options for you, 1 approach per file, so one of the approaches will not be used.

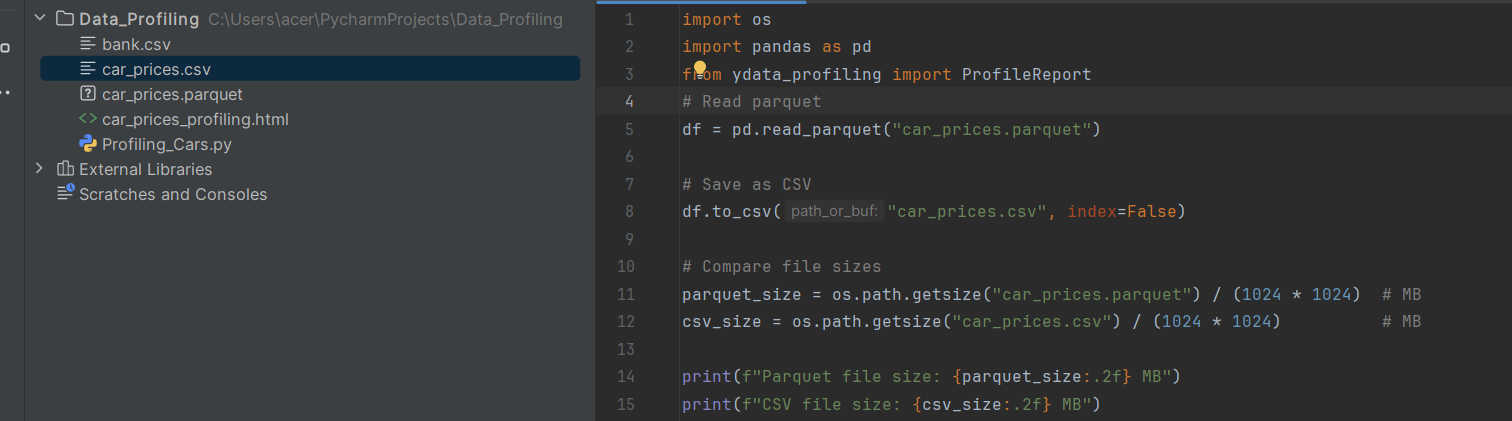
One of the provided files is in the parquet format – please read about parquet files here and use Python and pandas library to transform parquet file into .csv file.

In this report, I applied the second and third approaches—using DBeaver and the *ydata-profiling* library in Python—because they were more straightforward in this context. Although SQL could be used directly on the bank CSV file, it would require running multiple separate queries, making the process less convenient. In contrast, *ydata-profiling* handles most of the analysis automatically, providing a comprehensive overview with minimal manual effort. With the car prices dataset, which was originally in Parquet; since we first needed to compare dimensions, it was converted to CSV before performing the analysis in Python.

2. REPORT

Please, provide report in the Word format named as DQ\_Introduction\_Name\_Surname.docx Report should contain following information:

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



The dataset was stored in two formats: **parquet** and **CSV**.

* **Parquet file size:** 16.91 MB
* **CSV file size:** 88.21 MB

The parquet file is significantly smaller than the CSV file, even though both contain the same number of rows and columns. This difference is caused by:

* **Compression:** Parquet applies efficient column-wise compression, while CSV stores data as plain text without compression.
* **Columnar format:** Parquet stores data column by column, which reduces redundancy and makes it more efficient for analytics. CSV is row-based and repeats values more often.
* **Efficient encoding:** Parquet uses binary encoding for numbers and strings, while CSV stores everything as text, which takes up more space.

As a result, parquet is more storage-efficient and better suited for large datasets, while CSV is easier for manual reading but less efficient.

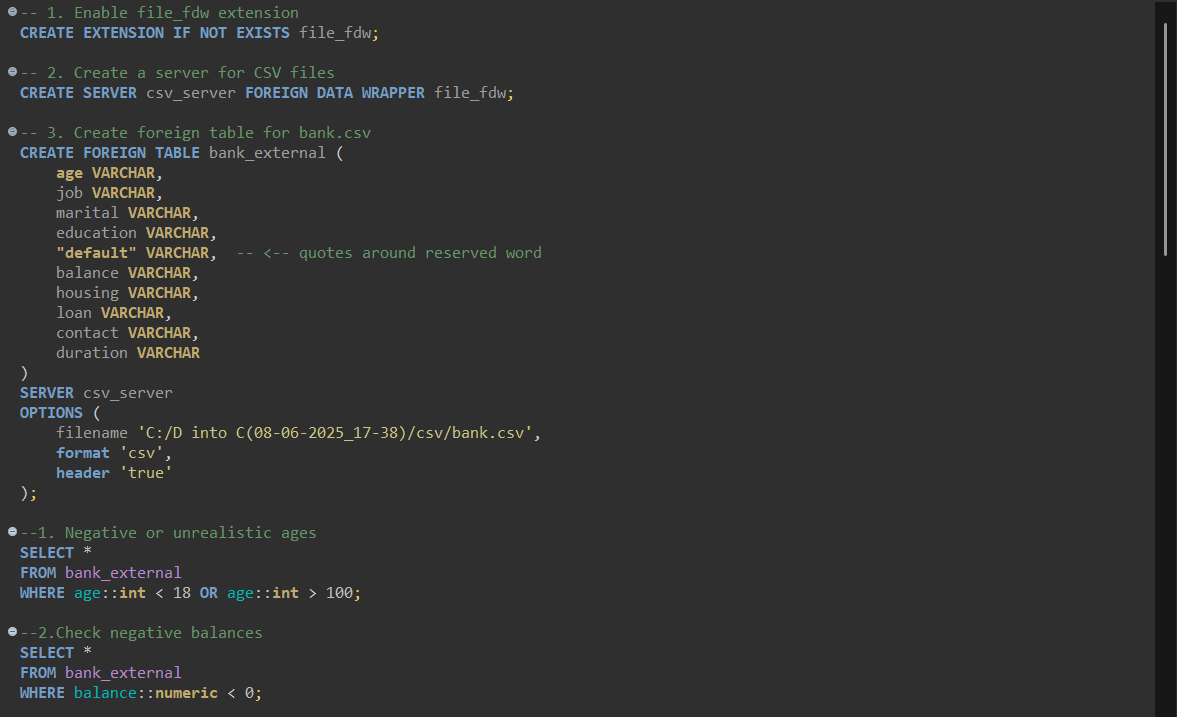
Section 2 – Describe bank and car\_prices dataset in several sentences, what information is contained in the datasets and how it could be used for business.

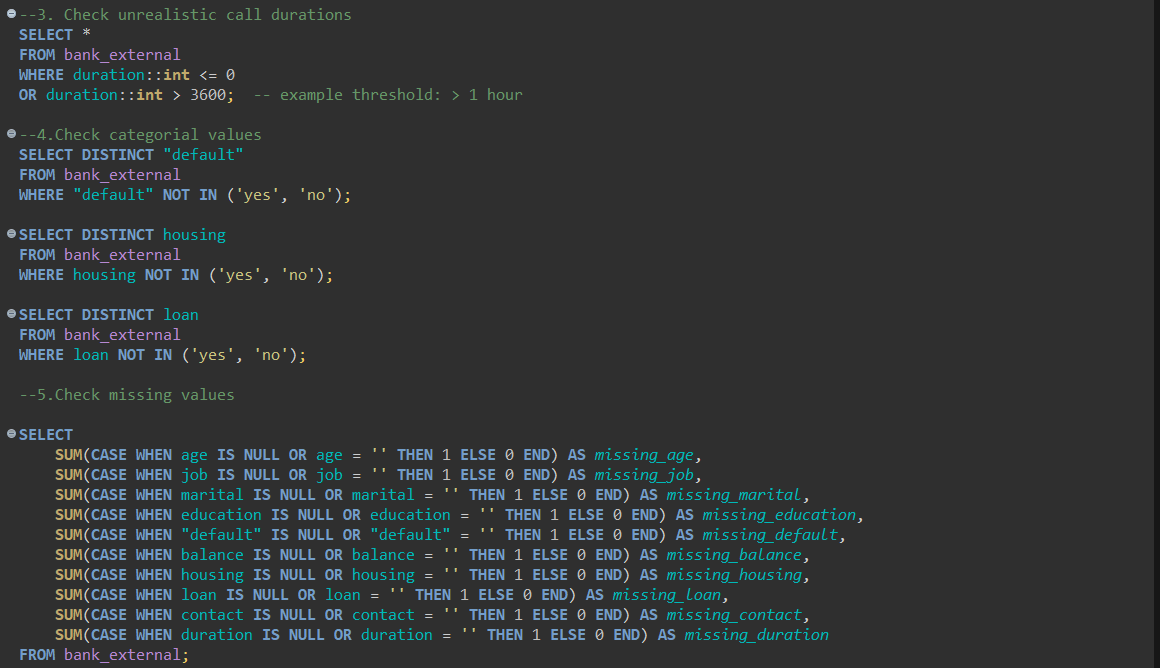
**Bank Dataset:** The dataset contains demographic (age, job, marital status, education), financial (balance, loan, default), and communication details. This data can be used for customer segmentation, credit risk assessment, and optimizing marketing campaigns.

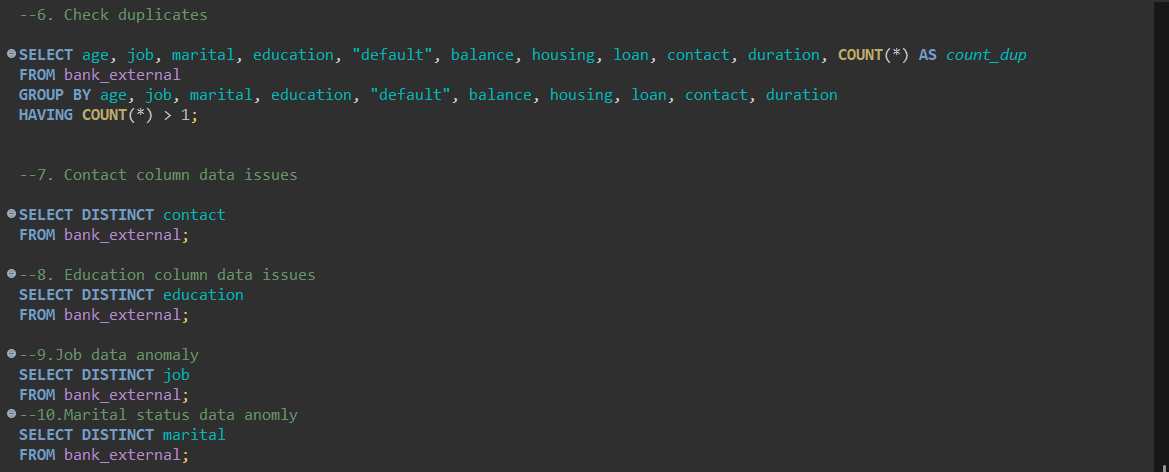
**Car Prices Dataset:** The dataset contains car attributes such as make, model, year, fuel type, mileage, and selling price. This can be used for pricing analysis, building valuation models, detecting fraudulent listings, and understanding market trends.

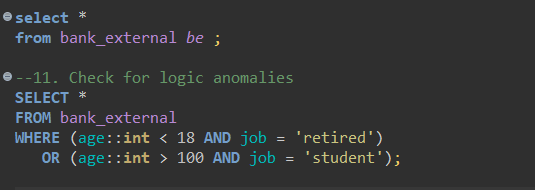
Section 3 – For each dataset find and list data anomalies that should be raised. Add brief description of each data anomaly (like dataset, attribute, issue description, example of corrupted values, how did you find it, why you think that this data is incorrect and etc).

***Bank Dataset:***

******

******

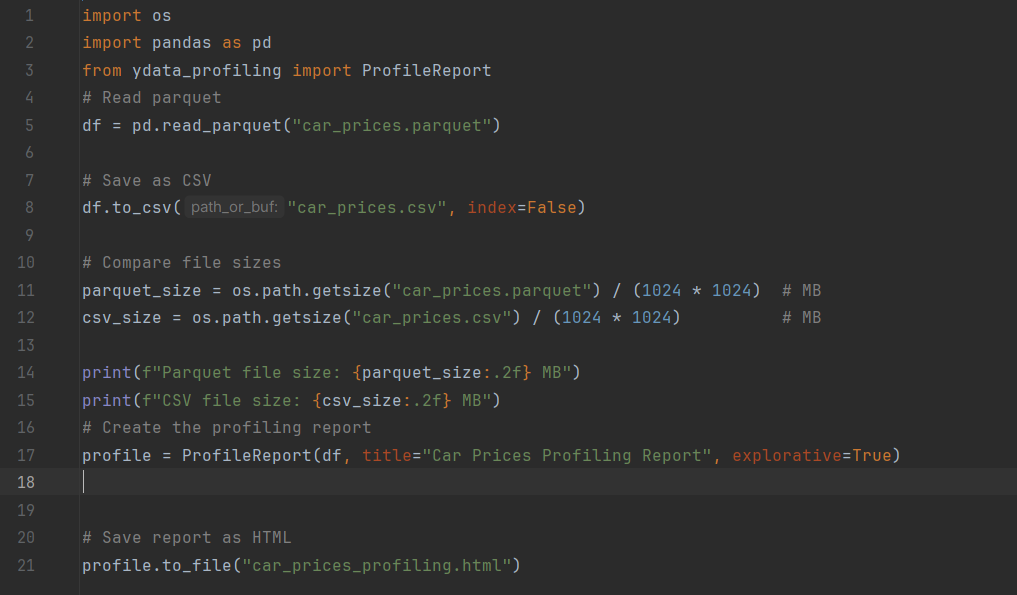
******

******

| **Dataset** | **Attribute** | **Issue Description** | **Example** | **How Found** | **Why Incorrect** |
| --- | --- | --- | --- | --- | --- |
| Bank | age | Negative or unrealistic values | -5, 120 | SQL: WHERE age::int < 18 OR age::int > 100 | Customers should be adults; negative or extreme ages indicate data entry errors |
| Bank | balance | Negative balances | -5000 | SQL: WHERE balance::numeric < 0 | Extreme negative balances are unlikely in this context; may indicate errors |
| Bank | duration | Unrealistic call duration (likely placeholder for missing) | -9999 | SQL: WHERE duration::int <= 0 OR duration::int > 3600 | Duration cannot be negative; likely a placeholder for missing or invalid data |
| Bank | default | Unexpected categorical values | "N" | SQL: SELECT DISTINCT "default" WHERE "default" NOT IN ('yes','no') | Only "yes" or "no" are valid; other values indicate data entry mistakes |
| Bank | loan | Inconsistent casing | "YES", "NO" | SQL: SELECT DISTINCT loan | Values should be standardized to lowercase "yes"/"no" for consistency |
| Bank | contact | Unexpected categorical values / placeholders | "NA", "Unkn", "unknown" | SQL: SELECT DISTINCT contact / WHERE contact IS NULL OR contact = '' | Only "cellular" or "telephone" expected; inconsistent or placeholder values reduce data quality |
| Bank | education | Unexpected categorical values / placeholders | "unknown", "Duck" | SQL: SELECT DISTINCT education | Only "primary", "secondary", "tertiary" expected; other values indicate missing data or entry errors |
| Bank | job | Unexpected categorical values / placeholders | "unknown", "boogieman" | SQL: SELECT DISTINCT job | Only valid job categories expected; other values indicate missing data or entry errors |
| Bank | marital | None detected | – | SQL: SELECT DISTINCT marital | All values are valid; no anomalies in this column |
| Bank | housing | None detected | – | SQL: SELECT DISTINCT housing | All values are valid; no anomalies in this column |
| Bank | Missing Values | None detected | – | SQL: Null count queries | No exact nulls detected; data is complete |
| Bank | Duplicates | None detected | – | SQL: GROUP BY COUNT(\*) > 1 | No duplicate rows found |

In addition to the main anomalies identified above, a review of value distributions for categorical attributes (job, education, marital, contact, default, loan, housing) and numeric attributes (age, balance, duration) was performed. Aside from the anomalies already flagged, all other values appear within expected ranges, with no rare or unexpected categories detected. This indicates that the dataset is generally consistent and reliable for analysis.

***Car Prices Dataset:***

******

| **Dataset** | **Attribute** | **Issue Description** | **Example** | **How Found** | **Why Incorrect** |
| --- | --- | --- | --- | --- | --- |
| Car Prices | year | None detected | – | Profiling min/max | All values are within realistic car manufacturing years; no anomalies |
| Car Prices | make | Missing values | NA / blank | Profiling missing values | Missing data reduces completeness; may affect analysis of car makes |
| Car Prices | model | Missing values | NA / blank | Profiling missing values | Missing entries reduce completeness; could affect analysis or grouping by model |
| Car Prices | trim | Missing values | NA / blank | Profiling missing values | Missing entries reduce completeness; could affect analysis or grouping by trim |
| Car Prices | body | Missing values | NA / blank | Profiling missing values | Missing entries reduce completeness; may affect analysis by car body type |
| Car Prices | transmission | Missing values and unexpected categories | Missing: 65,321; Rare: horse-driven, sedan, Sedan | Profiling missing values & distinct values | Missing values reduce completeness; rare categories are invalid and may indicate data entry errors |
| Car Prices | vin | Minor missing values | 4 missing entries | Profiling missing values | Missing VINs reduce ability to uniquely identify vehicles; otherwise data is clean |
| Car Prices | state | None detected | – | Profiling distinct & missing values | All values are valid; no anomalies |
| Car Prices | condition | Extreme numeric values; minor missing | Max: 982; Missing: 11,820 | Profiling min/max & missing values | Maximum is unrealistically high compared to typical car condition; missing values reduce completeness |
| Car Prices | odometer | Extreme numeric values; minor zeros | Max: 999,999; Zeros: 152 | Profiling min/max & zero counts | Values outside typical car mileage range may be data entry errors or placeholders; zeros reduce data quality |
| Car Prices | color | Rare / unexpected categories; minor missing | Other 41 rare colors; Missing: 749 | Profiling distinct & missing values | Rare or inconsistent color values may indicate typos or data entry errors; missing values reduce completeness |
| Car Prices | interior | Rare / unexpected categories; minor missing | — (17,077); Other 12 rare values | Profiling distinct & missing values | Placeholder “—” and rare values may indicate inconsistent labeling; missing data reduces completeness |
| Car Prices | seller | None detected | – | Profiling distinct & missing values | All values are present and valid; no anomalies |
| Car Prices | mmr | Extreme numeric outliers; minor missing | Max: 182,000; Missing: 123 | Profiling min/max & missing values | Maximum value is unusually high compared to typical market values; missing data reduces completeness |
| Car Prices | sellingprice | Extreme numeric outliers; minor missing | Min: 1; Missing: 12 | Profiling min/max & missing values | Minimum value 1 is unrealistic; missing data reduces completeness |
| Car Prices | saledate | Minor missing values | Missing: 12 | Profiling missing values | Missing sale dates reduce completeness for analysis and tracking |

Overall, the **bank dataset** is generally clean, with minor inconsistencies in categorical values and numeric ranges that need correction. The **car prices dataset** contains several minor missing values and a few extreme numeric outliers, but most columns are complete and valid. Both datasets are suitable for business analysis after addressing the identified anomalies.