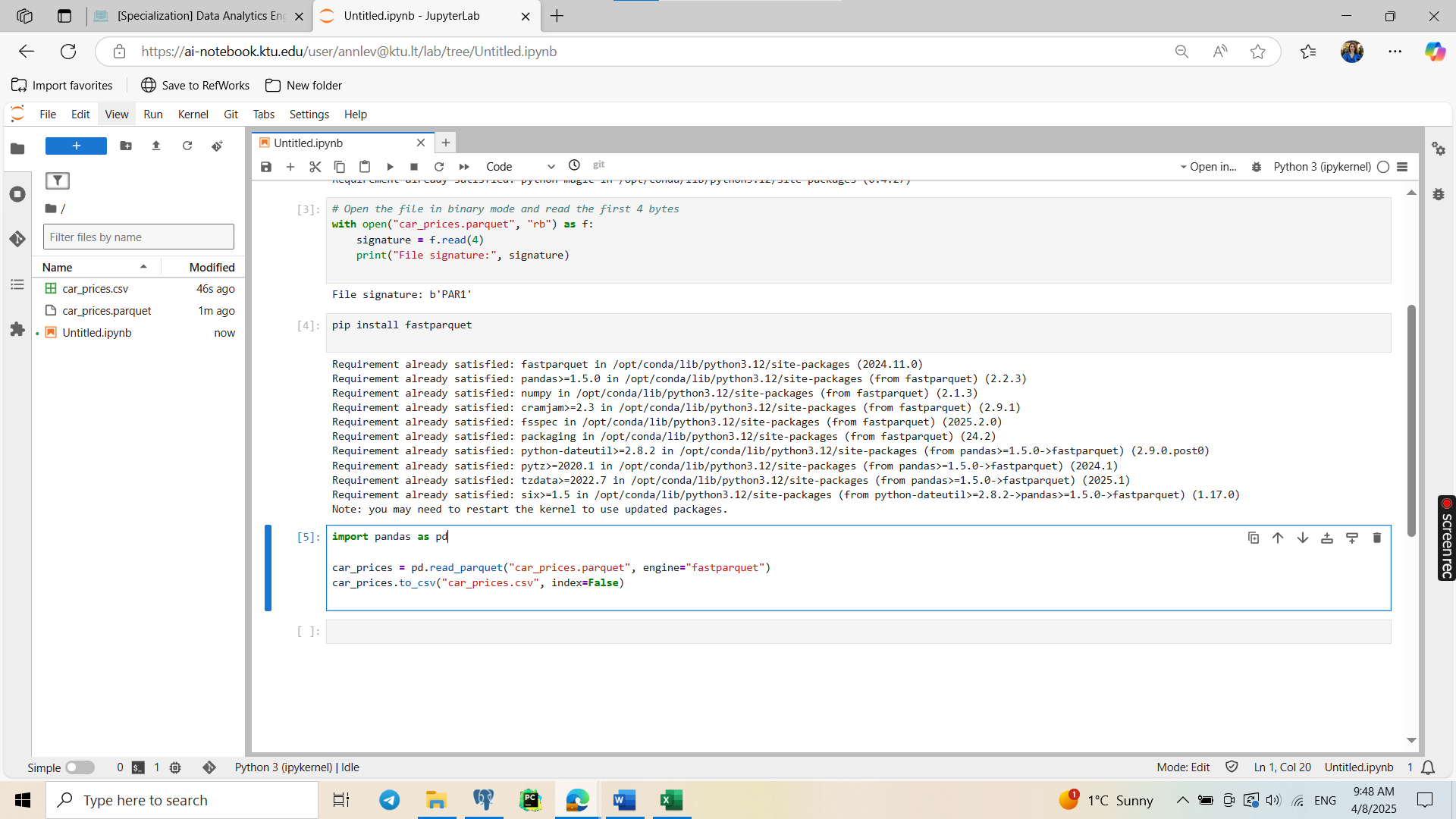
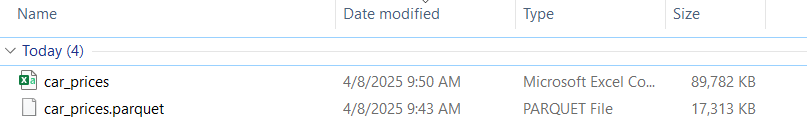
**Task 1**





The .parquet file is 17.313 KB while the .csv version is 89.782 KB. This is because Parquet is a columnar, binary format that is optimized for efficient storage and retrieval, especially when reading only specific columns. CSV, being plain text, is more verbose and lacks compression.

**Task 2**

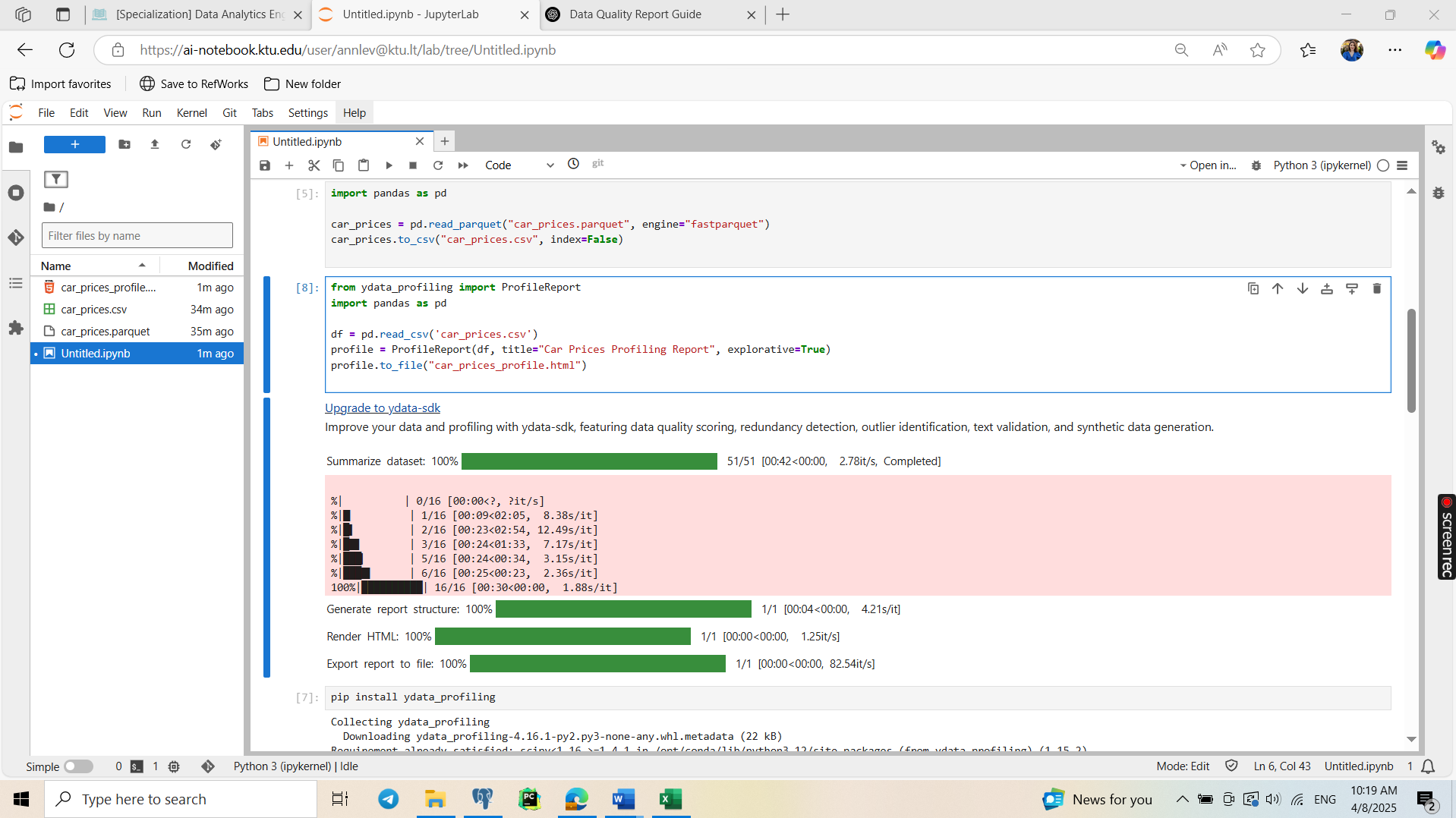
Bank Dataset: Contains customer demographic and marketing data, including age, job, marital status, and outcome of a marketing campaign. Businesses can use this data to segment customers and optimize targeted marketing efforts.

Car Prices Dataset: Includes car listing data such as make, model, year, price, transmission type. Useful for dealerships or price comparison platforms to predict fair car pricing or detect market trends.

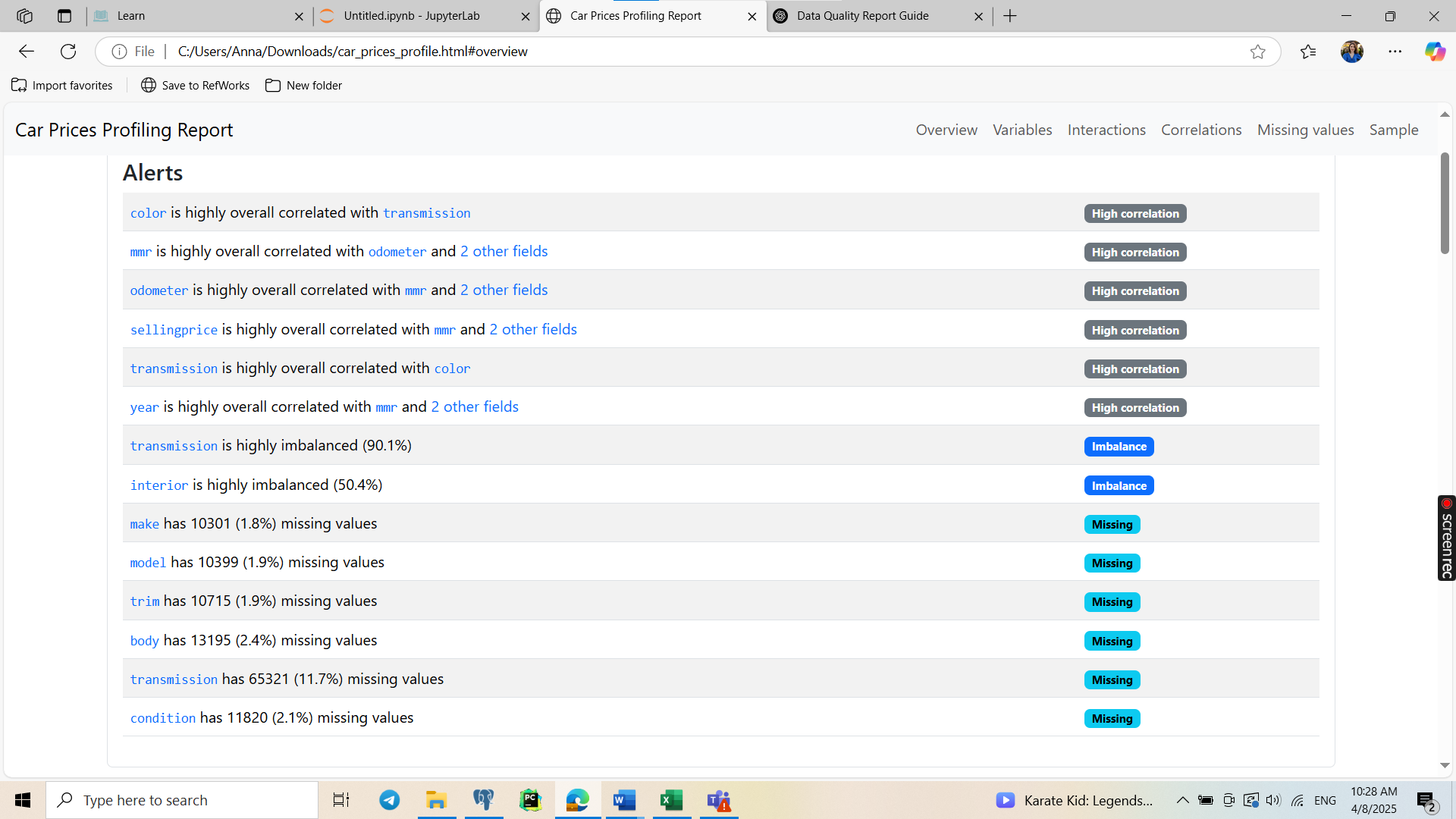
**Task 3**

**Analyzing with SQL bank.csv**

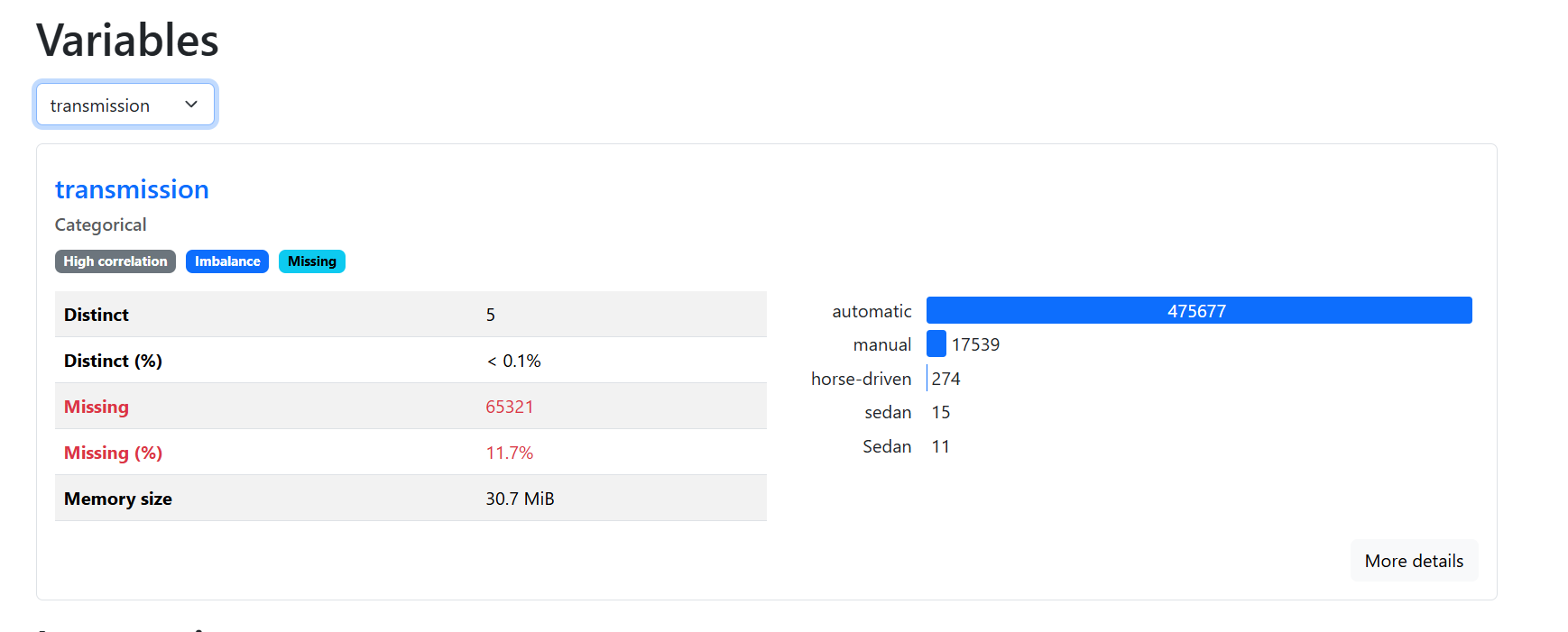
|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Issue Description** | **Example / Query Used** | **Why It's a Problem** |
| age | Negative or zero ages | SELECT \* FROM ext\_bank WHERE age <= 0; | Invalid age — age cannot be 0 or negative. |
| age | Unusually high ages (over 100) | SELECT \* FROM ext\_bank WHERE age > 100; | May be outliers or data entry errors. |
| job | Inconsistent formats (e.g., admin vs admin.) | SELECT DISTINCT job FROM ext\_bank; | Causes inconsistent grouping and classification. |
| default\_status, housing, loan | Unexpected values (should be yes/no) | SELECT DISTINCT default\_status FROM ext\_bank;  SELECT DISTINCT housing FROM sa\_v\_bank\_data.ext\_bank;  SELECT DISTINCT loan FROM sa\_v\_bank\_data.ext\_bank; | Could lead to incorrect segmentation or misinterpretation. |
| contact | Null or blank values | SELECT \* FROM ext\_bank WHERE contact IS NULL; | Contact info is crucial for marketing outreach. |
| duration | Values <= 0 | SELECT \* FROM ext\_bank WHERE duration < 0; | Duration likely refers to call time — too short or too long can signal bad data. |
| All columns | Duplicate rows | SELECT age, job, marital, education, default\_status, balance, housing, loan, contact, duration, COUNT(\*)  FROM sa\_v\_bank\_data.ext\_bank  GROUP BY age, job, marital, education, default\_status, balance, housing, loan, contact, duration  HAVING COUNT(\*) > 1; | Duplicates distort statistics and analysis results. |

**Analyzing with ydata-profiling car\_prices.csv**

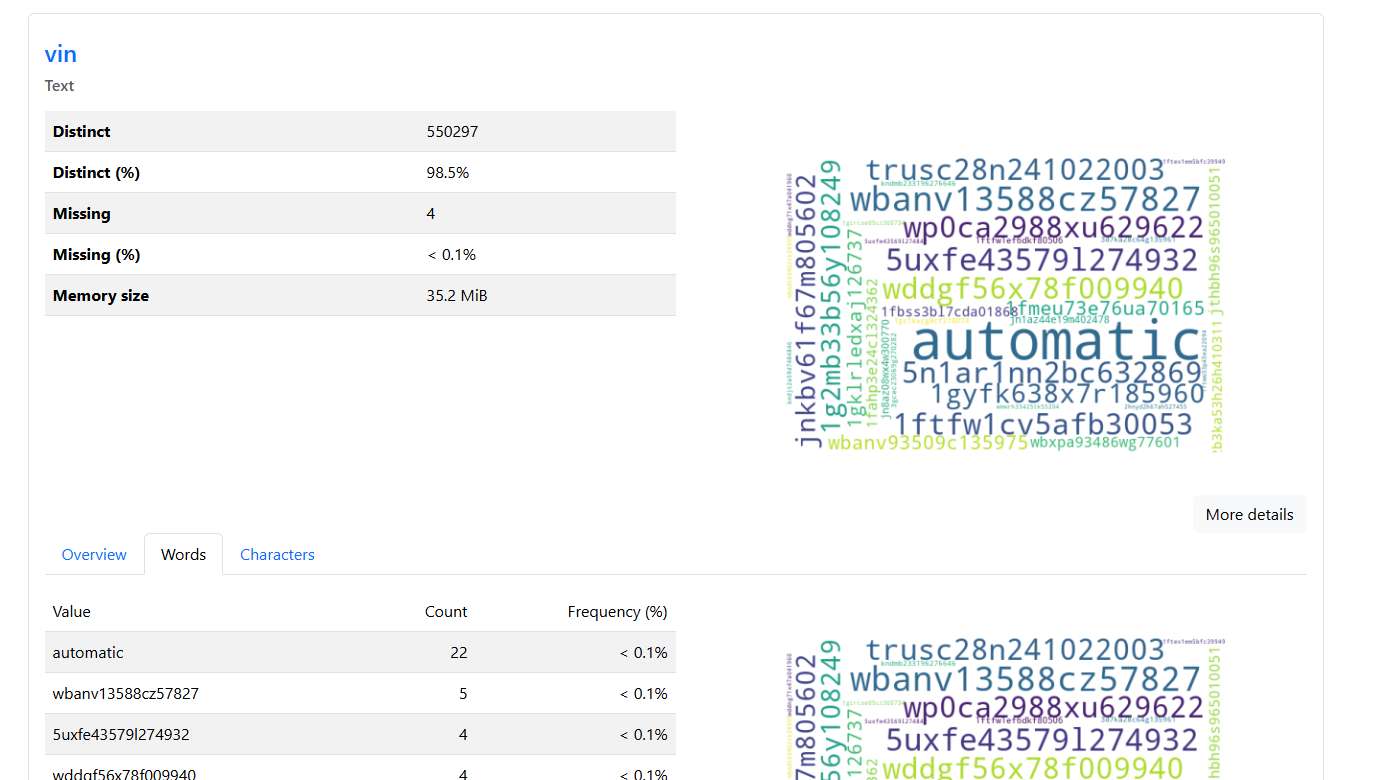
A lot of missing values.



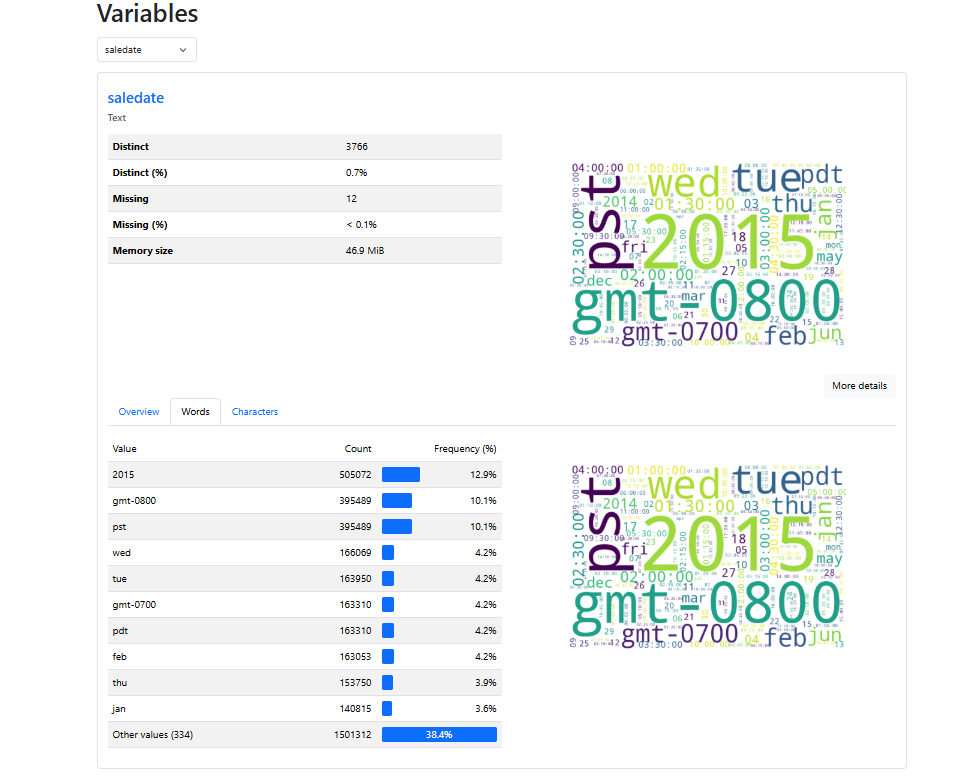
“sedan” and “Sedan” in transmission, it seams to be incorrect even logically, but also upper and lower case as well.



In VIN word “automatic” can’t be



Not understandable values (gmt-0800, gmt-0700) for sales date

****