First of all I will use ydata-profiling Approach for car\_prices.parquet:

And for car.csv I will take manual checking

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

Parquet files are typically smaller than csv because they are stored in a columnar format and utilize compression. It efficiently handles complex data and compresses it better, while **CSV** stores data in a row-wise format with no compression.

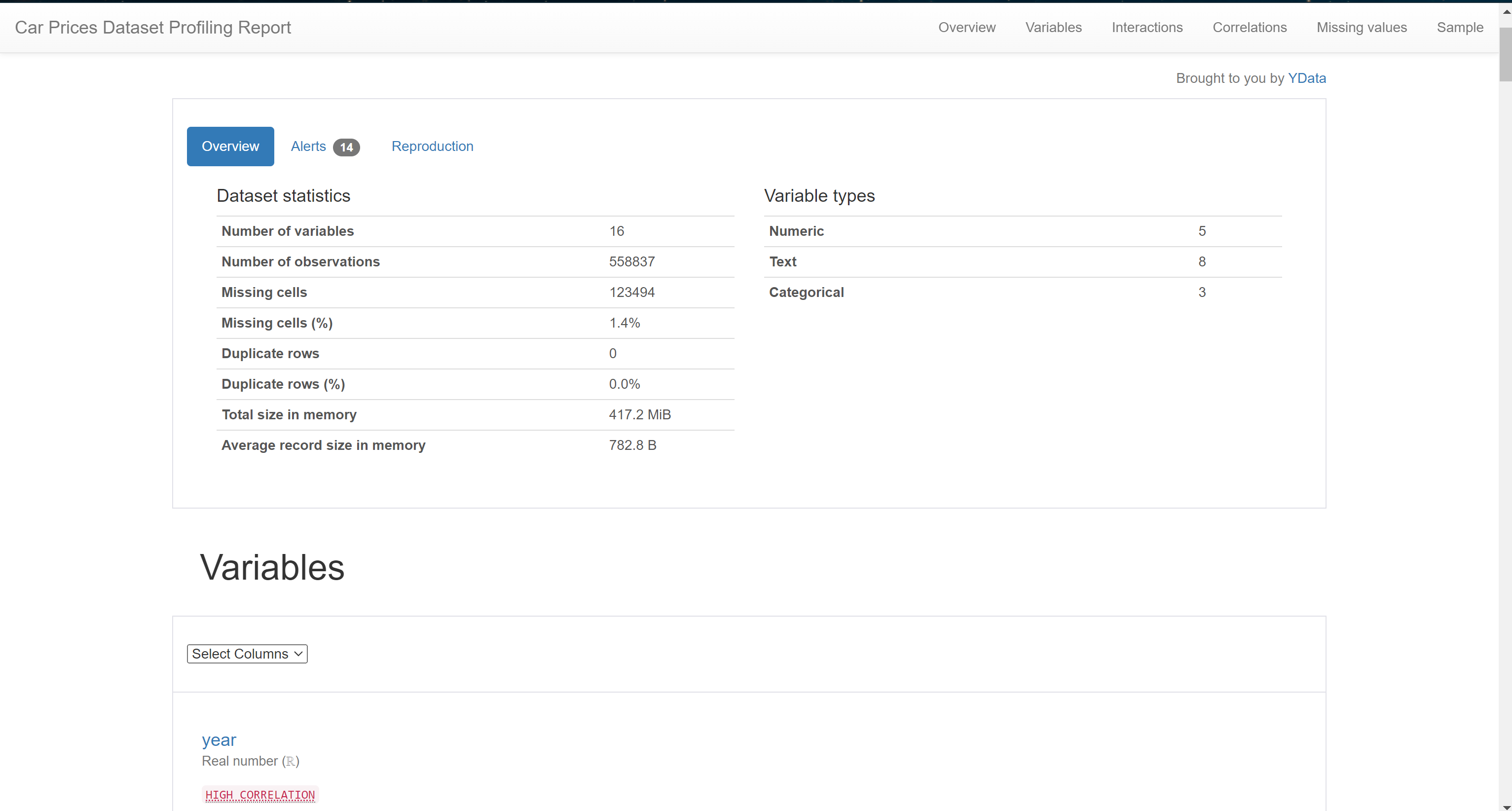
Section 2 – Describe bank and car\_prices dataset in several sentences, what information is contained in the datasets and how it could be used for business.

**Bank. Csv -** contains age, job, marriage status, education, balance, if they have loan, contact info and duration, probably in the bank.

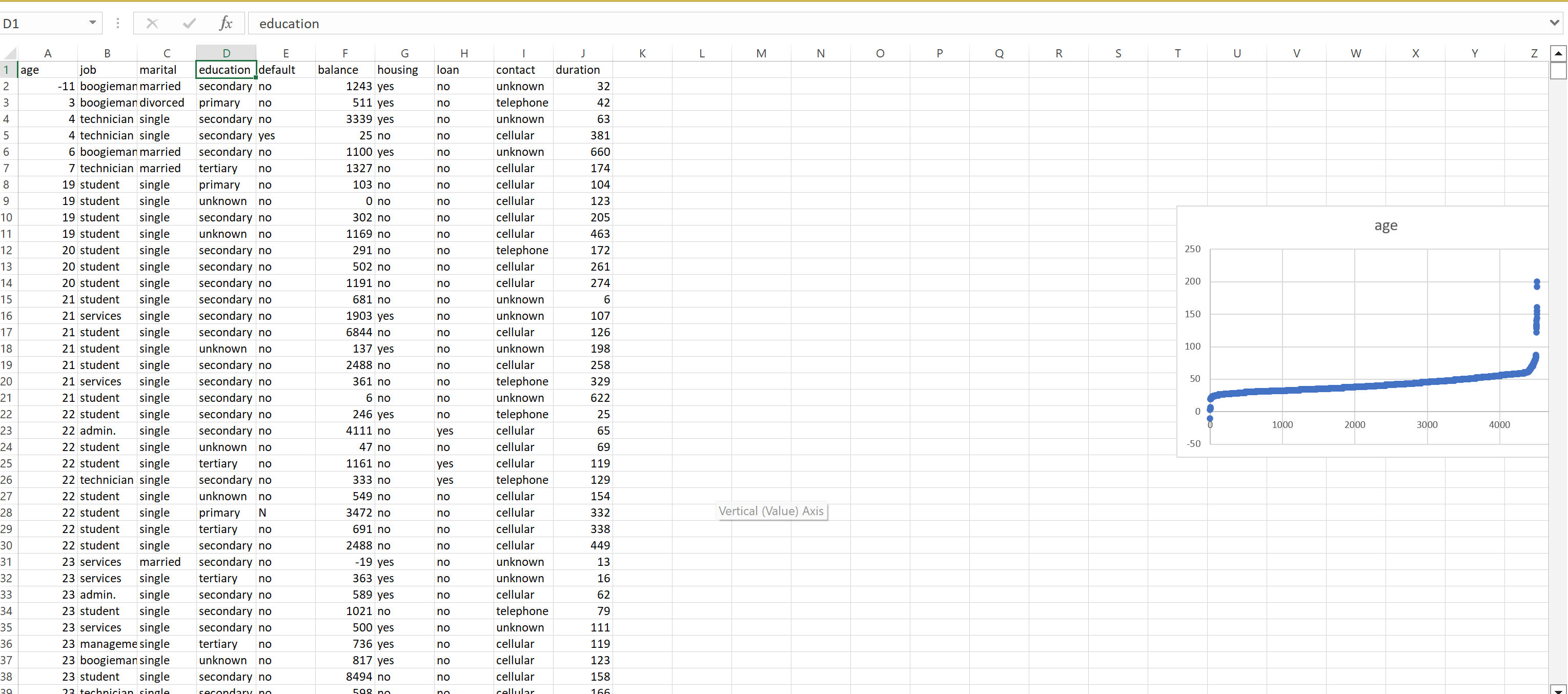
This dataset can help in understanding **customer profiles**, trends and some correlation about age and loan, or marriage status and loan, or balance, as well as education and loan and so on.

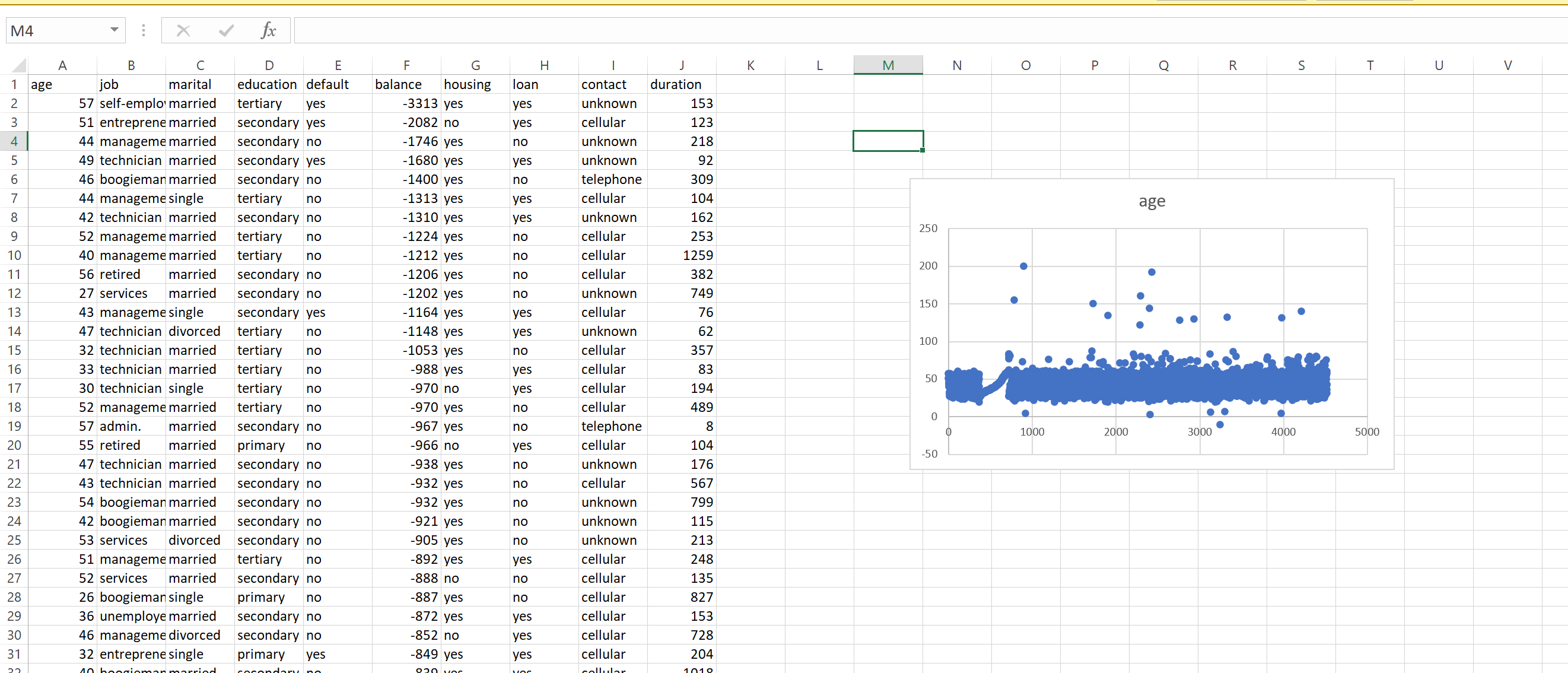
In the **car\_price** dataset, there is information about year, model, transmittion, state, model, interior, selling price and seller

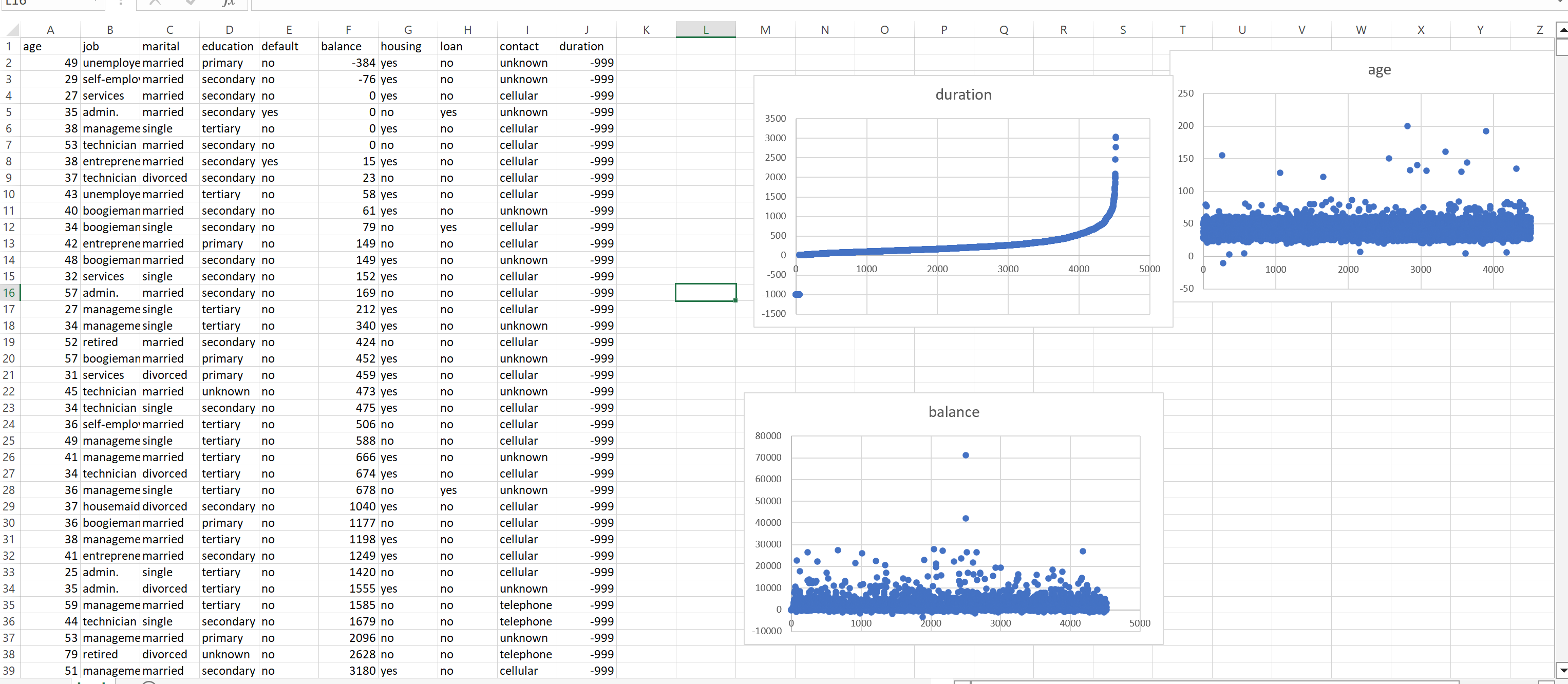
This dataset can help in pricing analysis, market trend prediction, and comparison of car models based on price or features.



Section 3 – For each dataset find and list data anomalies that should be raised. Add brief description of each data anomaly (what is it, how do you find it, why you think that this data is incorrect, some screenshots will be beneficial.







First in Bank.csv there is obvious anomaly of age -11 and 200. And duration -999, that can be logical.

I found anomalies by sorting data in excel.

Anomalies in car\_prices -

Missing Values:

year 0

make 10301

model 10399

trim 10715

body 13195

transmission 65321

vin 4

state 0

condition 11820

odometer 94

color 749

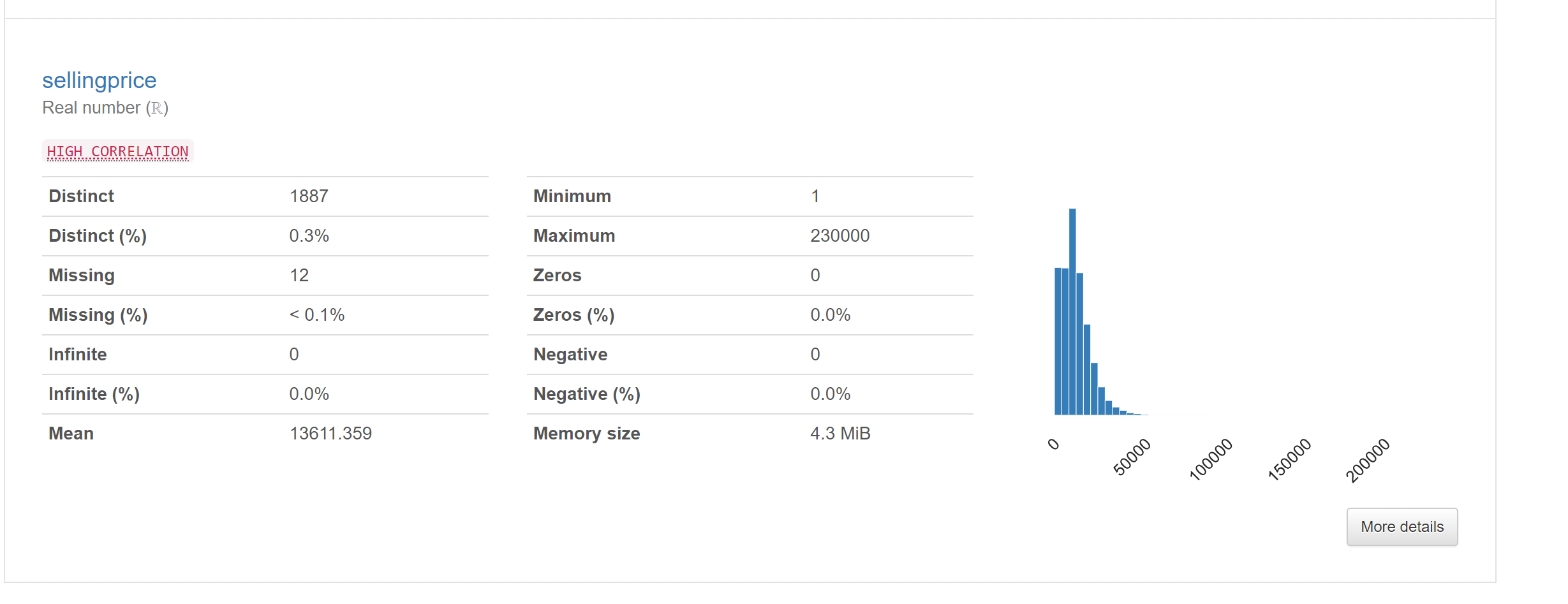
interior 749

seller 0

mmr 123

sellingprice 12

saledate 12



Selling price is 1 min, which is too less, so its anomaly.