**Dataset Overview:**

Total Records: 45211 (including header)

Variables: 17 columns

* 8 variables define bank client data.
* 4 variables related with the last contact of the current campaign
* 4 variables define other attributes
* 1 variable define output variables

Format: Semicolon-delimited CSV with quoted strings

Context: Direct marketing campaigns of a Portuguese banking institution customer records.

**Variable data Range/conditions:**

# Expected Domains

expected\_domains = {

'age': (18, 95),

'balance': None, # No range check on balance

'day': (1, 31),

'duration': (0, np.inf),

'campaign': (1, np.inf),

'pdays': (-1, 999),

'previous': (0, 300),

'job': ["admin.", "unknown", "unemployed", "management", "housemaid",

"entrepreneur", "student", "blue-collar", "self-employed",

"retired", "technician", "services"],

'marital': ["married", "divorced", "single"],

'education': ["unknown", "secondary", "primary", "tertiary"],

'default': ["yes", "no"],

'housing': ["yes", "no"],

'loan': ["yes", "no"],

'contact': ["unknown", "telephone", "cellular"],

'month': ["jan", "feb", "mar", "apr", "may", "jun",

"jul", "aug", "sep", "oct", "nov", "dec"],

'poutcome': ["unknown", "other", "failure", "success"],

'y': ["yes", "no"]

}

**Variable Definitions and Metadata:**

**Bank Client Details:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Name** | **Type** | **Description** | **Domain/Values** | **Notes** |
| 1 | **age** | Numeric | Client's age in years | Integer (18-95 typical range) | Age values fall between reasonable limits (e.g. 18–95). |
| 2 | **job** | Categorical | Type of occupation | "admin.", "unknown", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", "services" | There are 12 data categories.  "Unknown" means the information is missing. |
| 3 | **marital** | Categorical | Marital status | "married", "divorced", "single" | The label "Divorced" includes both divorced and widowed individuals |
| 4 | **education** | Categorical | Education level | "unknown", "secondary", "primary", "tertiary" | The "Unknown" label appears where data is not available. |
| 5 | **default** | Binary | Has credit default? | "yes", "no" | This data is important for evaluating financial risk. |
| 6 | **balance** | Numeric | Average yearly balance in euros | Real numbers (can be negative) | May contain negative values indicating debit history. |
| 7 | **housing** | Binary | Has housing loan? | "yes", "no" | Indicates mortgage/housing debit history. |
| 8 | **loan** | Binary | Has personal loan? | "yes", "no" | Indicates personal obligations |

**Last Contact Information:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Name** | **Type** | **Description** | **Domain/Values** | **Notes** |
| 9 | **contact** | Categorical | Contact communication type | "unknown", "telephone", "cellular" | "unknown" may indicate missing contact method |
| 10 | **day** | Numeric | Last contact day of the month | Integer (1-31) | Validate against month constraints |
| 11 | **month** | Categorical | Last contact month of year | "jan", "feb", "mar", "apr", "may", "jun", "jul", "aug", "sep", "oct", "nov", "dec" | 12 categories, seasonal patterns may exist |
| 12 | **duration** | Numeric | Last contact duration in seconds | Integer >=0 | Important predictor, 0 indicates very brief contact |

**Campaign History:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Name** | **Type** | **Description** | **Domain/Values** | **Notes** |
| 13 | **campaign** | Numeric | Number of contacts performed during current campaign | Integer >=1 | Includes last contact, minimum value is 1 |
| 14 | **pdays** | Numeric | Days since last contact from previous campaign | Integer; -1 = not previously contacted | -1 is a special code, not missing data |
| 15 | **previous** | Numeric | Number of contacts from previous campaigns | Integer >=0 | 0 indicates first-time contact |
| 16 | **poutcome** | Categorical | Outcome of previous marketing campaign | "unknown", "other", "failure", "success" | "unknown" may indicate no previous campaign |

**Data Analysis based on 6 pillars (accuracy, completeness, consistency, timeliness, uniqueness, and integrity):**

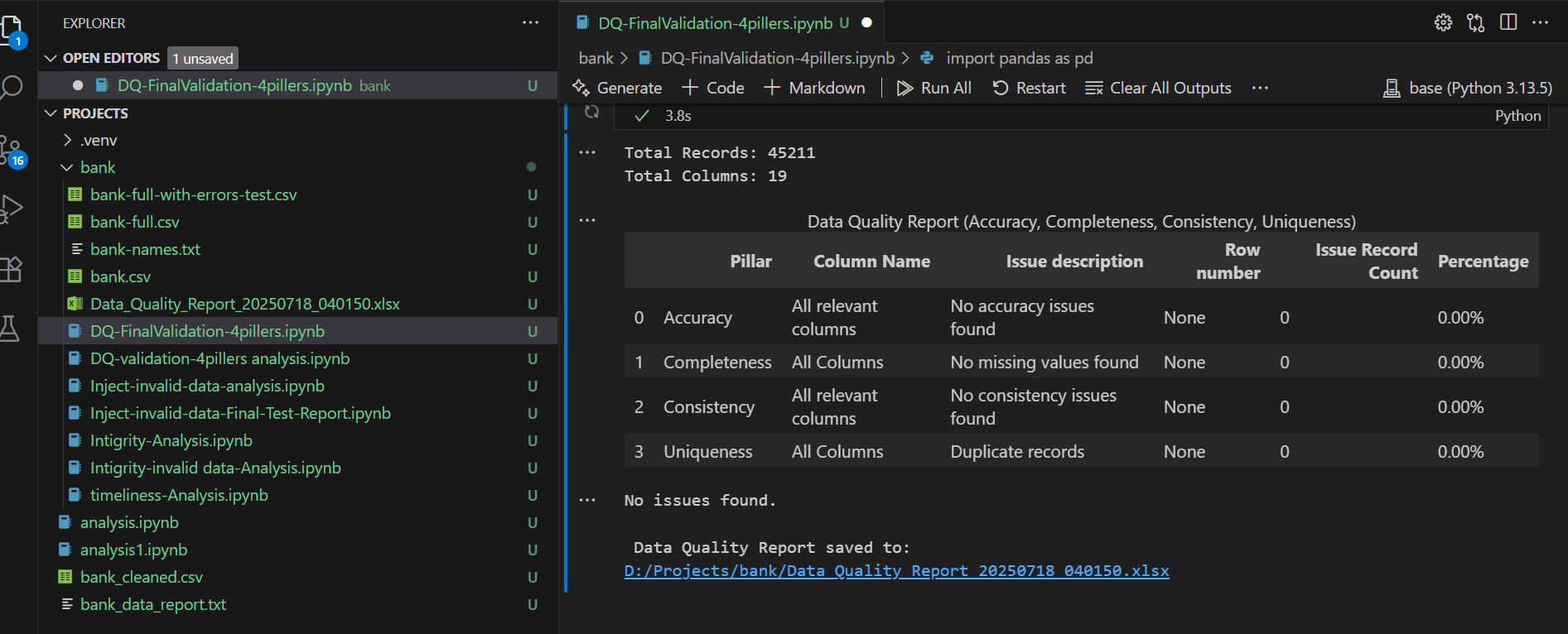
**High level overview of 6 DQ Pillars:**

|  |  |  |
| --- | --- | --- |
| **Pillar** | **Simple Definition** | **Example** |
| **Accuracy** | Data is correct and matches reality. | Age is 30 if the person is really 30. |
| **Completeness** | Nothing is missing; all required data is present. | Every record has name, phone, and address. |
| **Consistency** | Data is written the same way everywhere. | Gender is always "Male/Female", not mixed. |
| **Timeliness** | Data is up to date and current. | Using the latest customer address. |
| **Uniqueness** | No duplicate records exist. | Customer not registered twice. |
| **Integrity** | Data is properly linked and connected. | Orders are linked to real existing customers. |

**DQ scenarios:**

|  |  |
| --- | --- |
| **#** | **Scenarios:** |
| 1 | Accuracy check - Evaluate whether all 45,211 records across 17 columns meet the expected value ranges and defined conditions. |
| 2 | Data Completeness Assessment - Assess the dataset for missing or null values to confirm full data availability across all 17 columns. |
| 3 | Data Consistency Verification -Check that all the data follows the same format, categories, and makes sense throughout the whole dataset. |
| 4 | Data Uniqueness Validation - Identify any repeated entries or values to make sure important fields are unique across all 45,211 record |
| 5 | Data Integrity Validation - Check that the columns are correctly linked and follow the rules, so the data stays accurate and well-organized. |
| 6 | Data Timeliness Evaluation - Verify time-related data to show the right dates or times based on business rules. |
| 7 | Data Type Compliance - Verify that each column has the right type of data, like numbers, words, or dates, and follows the rules |
| 8 | Special Character Inspection - any unusual or special characters that could cause problems when using or reading the data. |

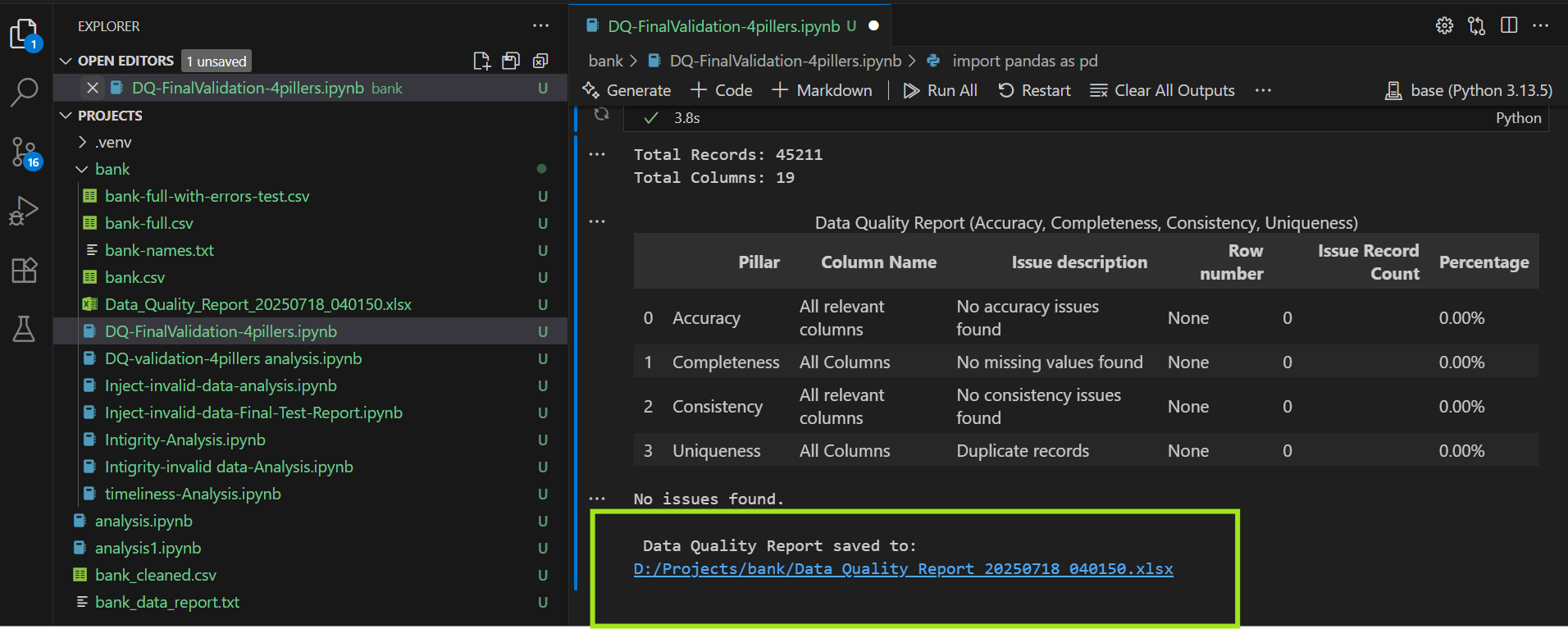
**For valid data:**

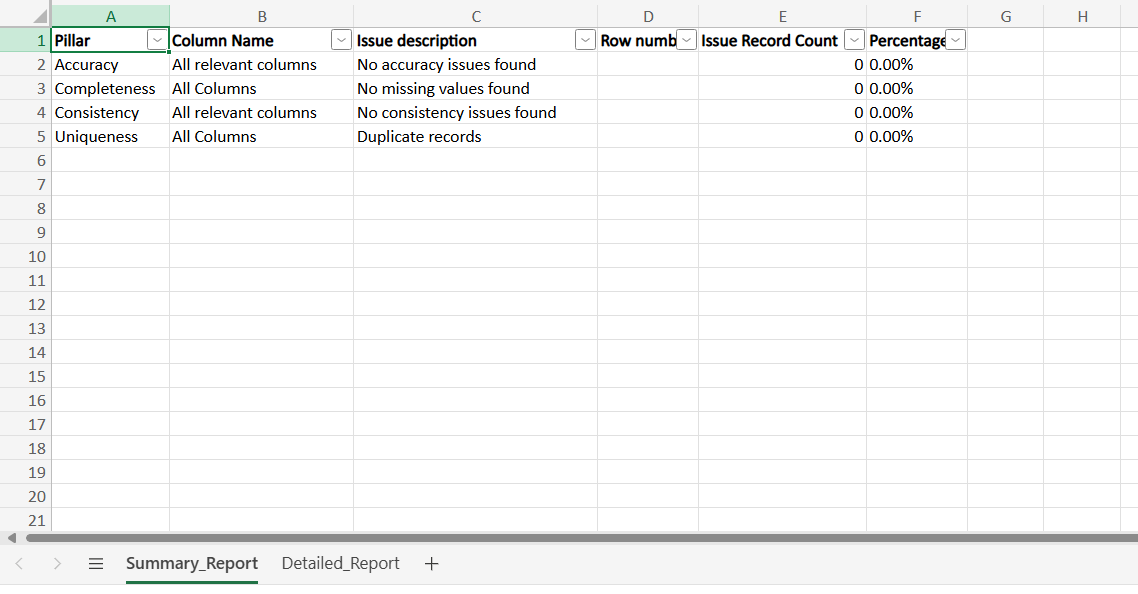


Data Output: No issues found

**Test Report: (In Excel)**

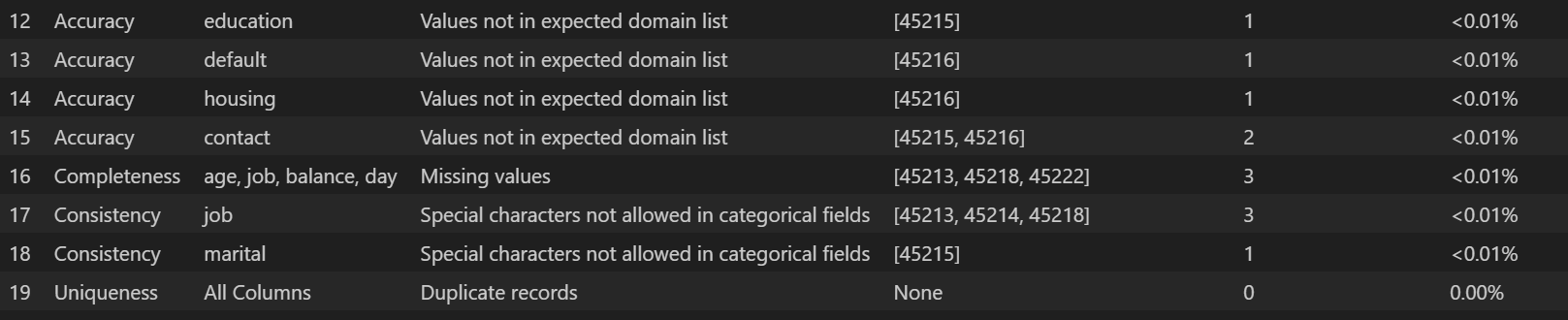
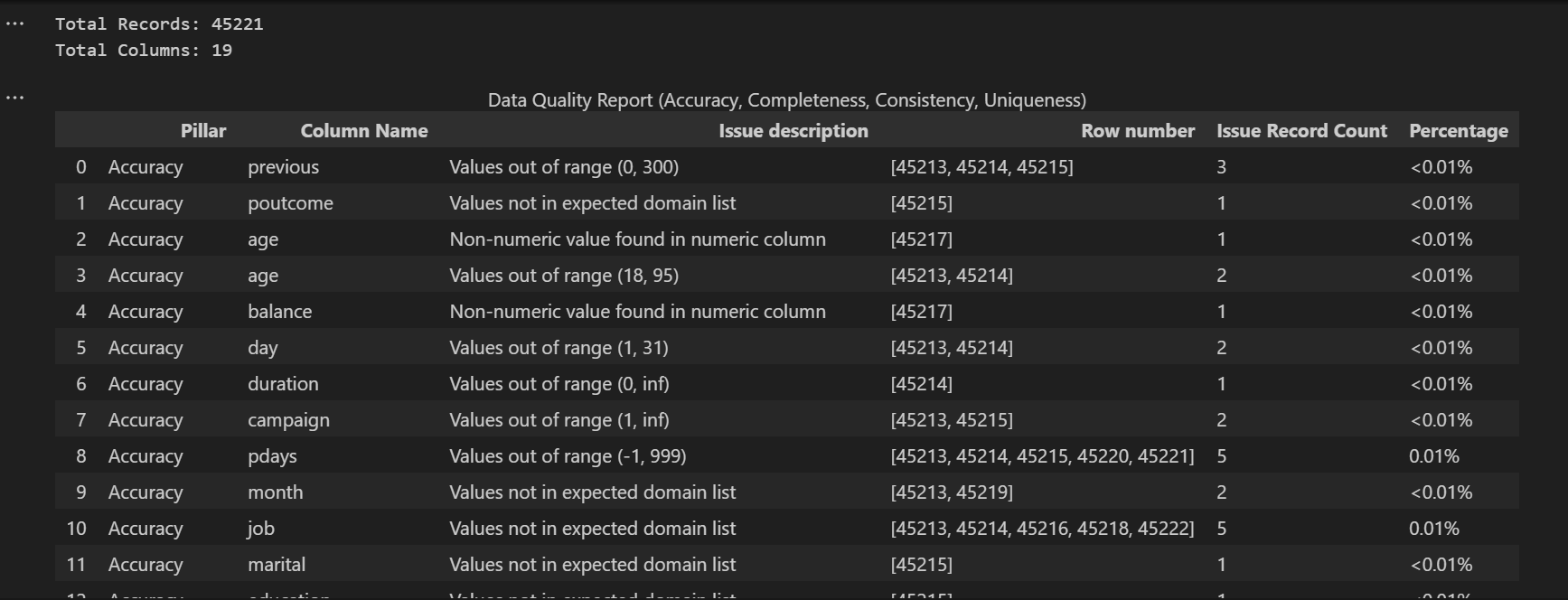
Test result generated as report with updated date and timestamp



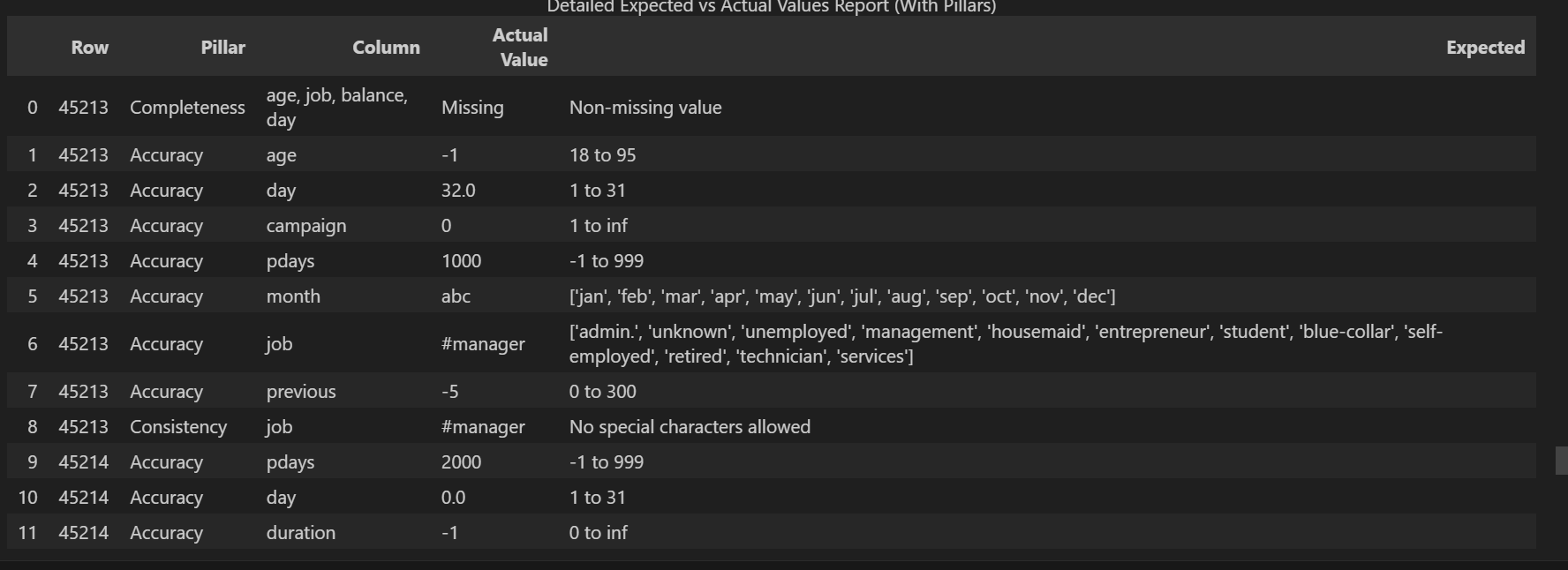


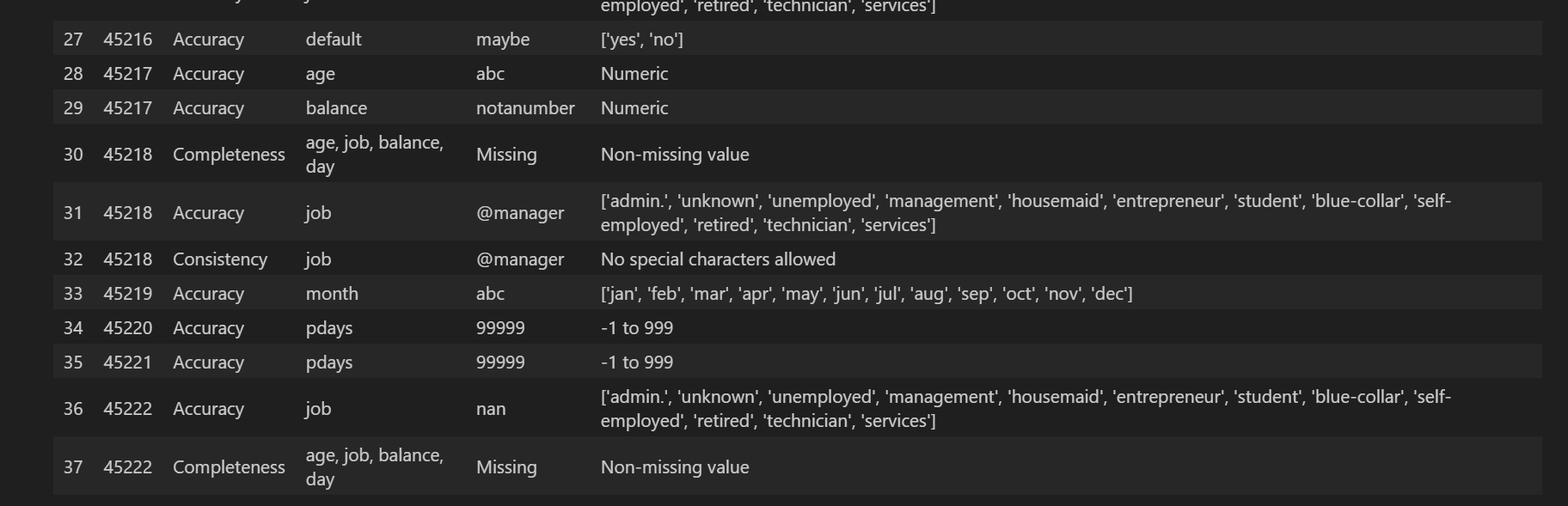
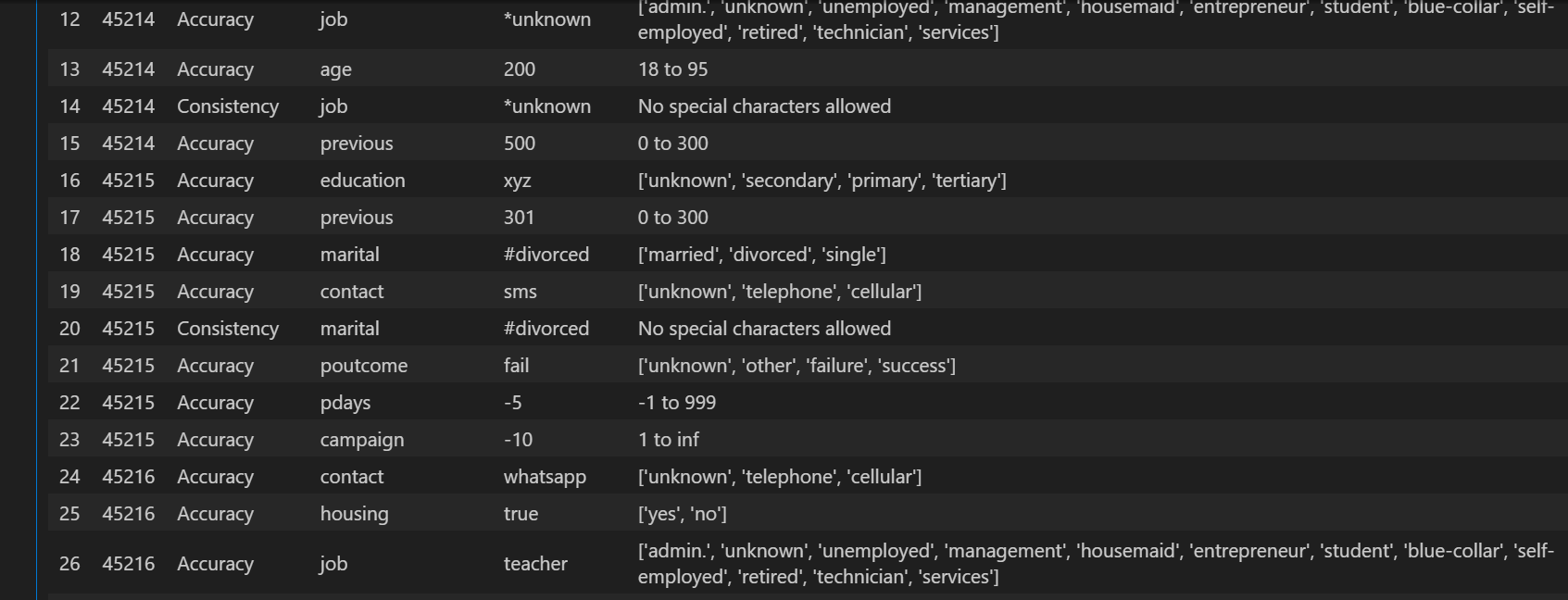
**Inject invalid data in data set to verify DQ issue:**

**Data Quality Report (Accuracy, Completeness, Consistency, Uniqueness)**

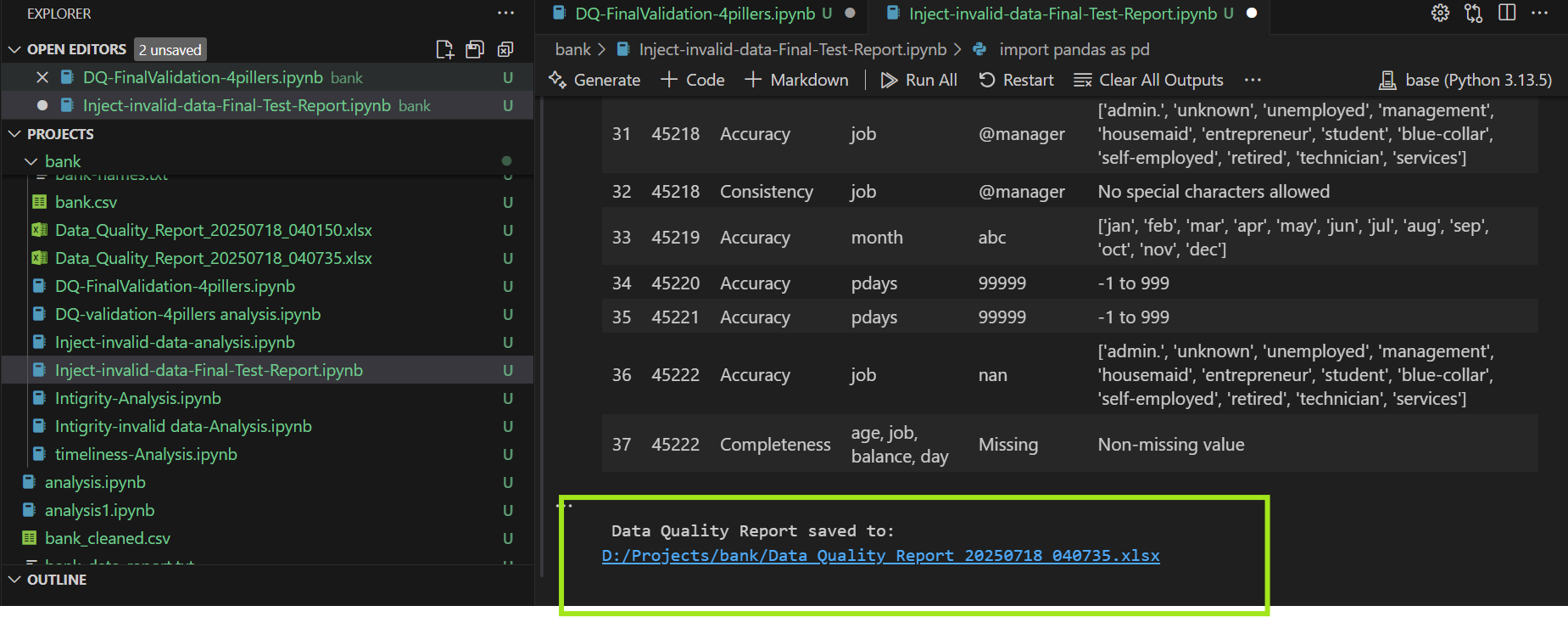


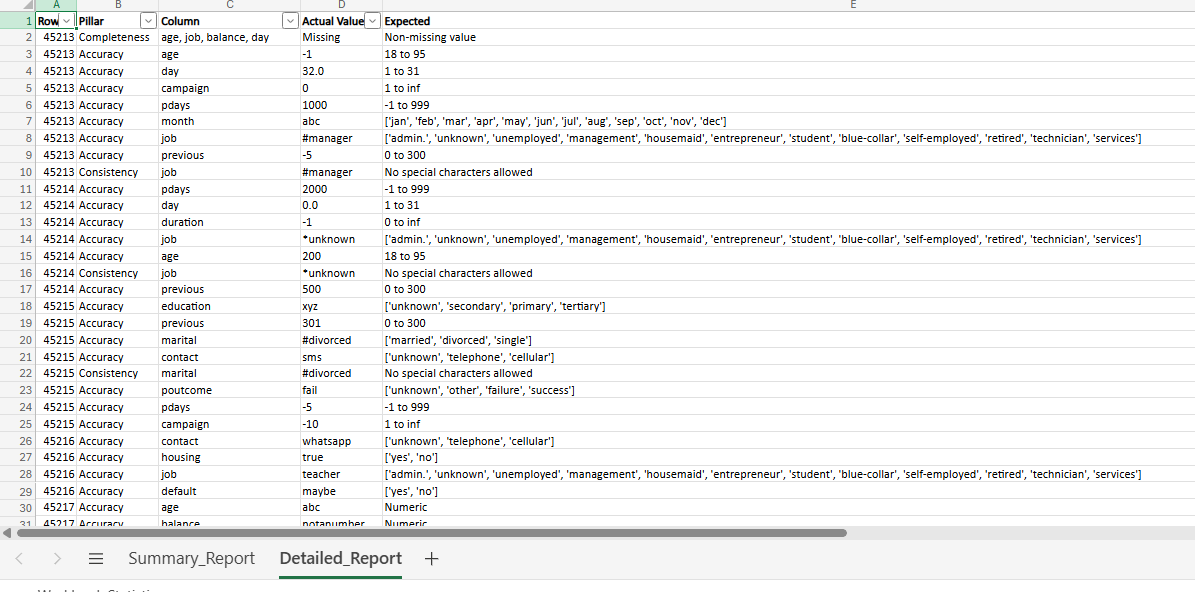
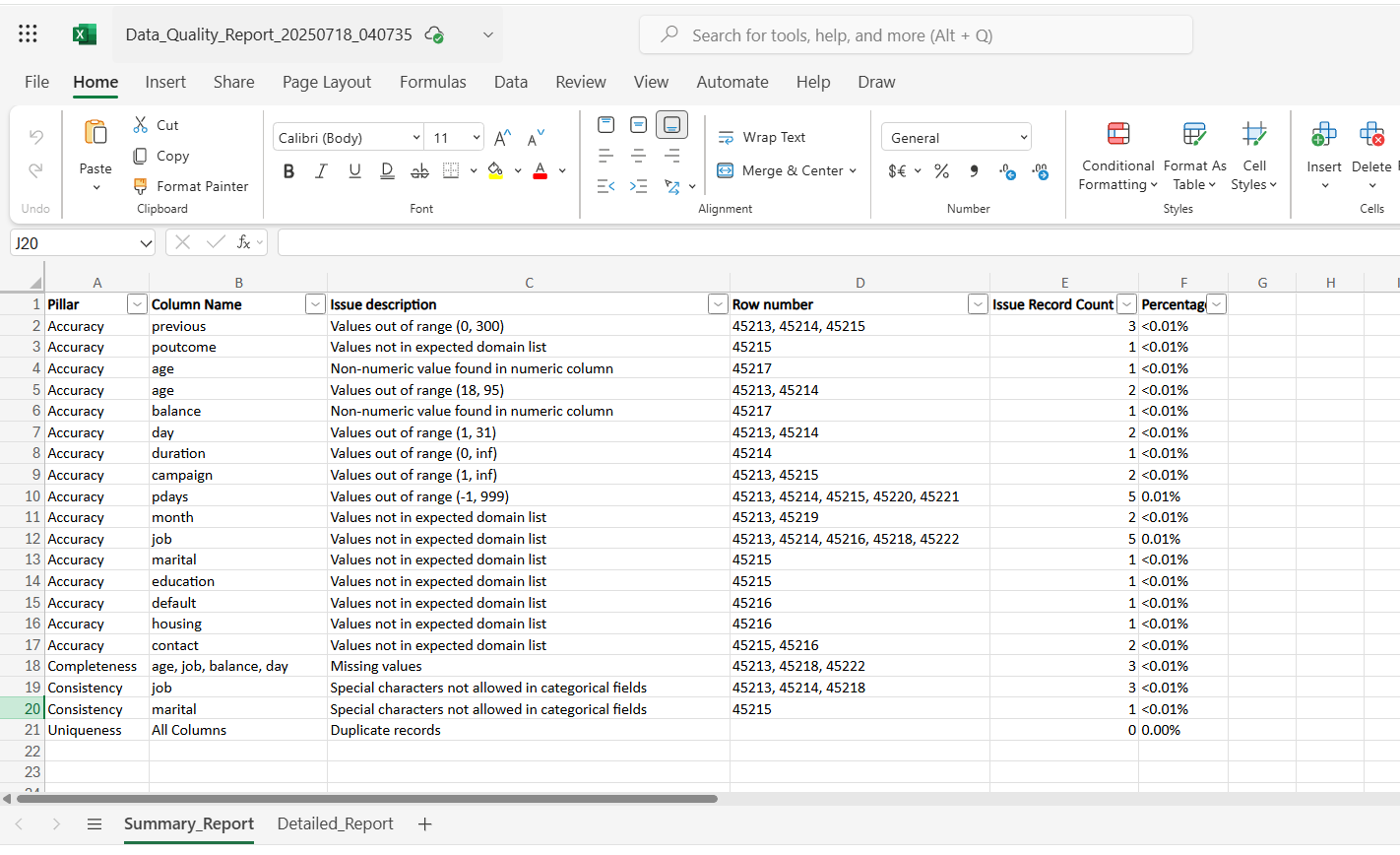
**Detailed Expected vs Actual Values Report (With Pillars)**



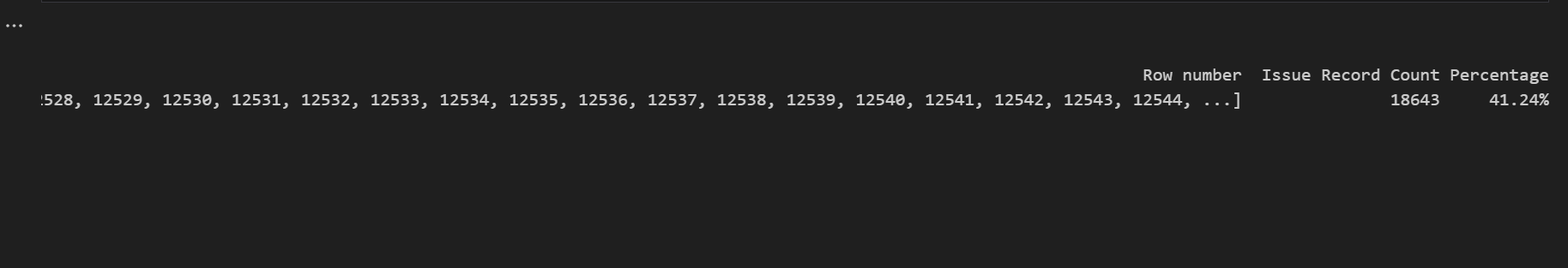
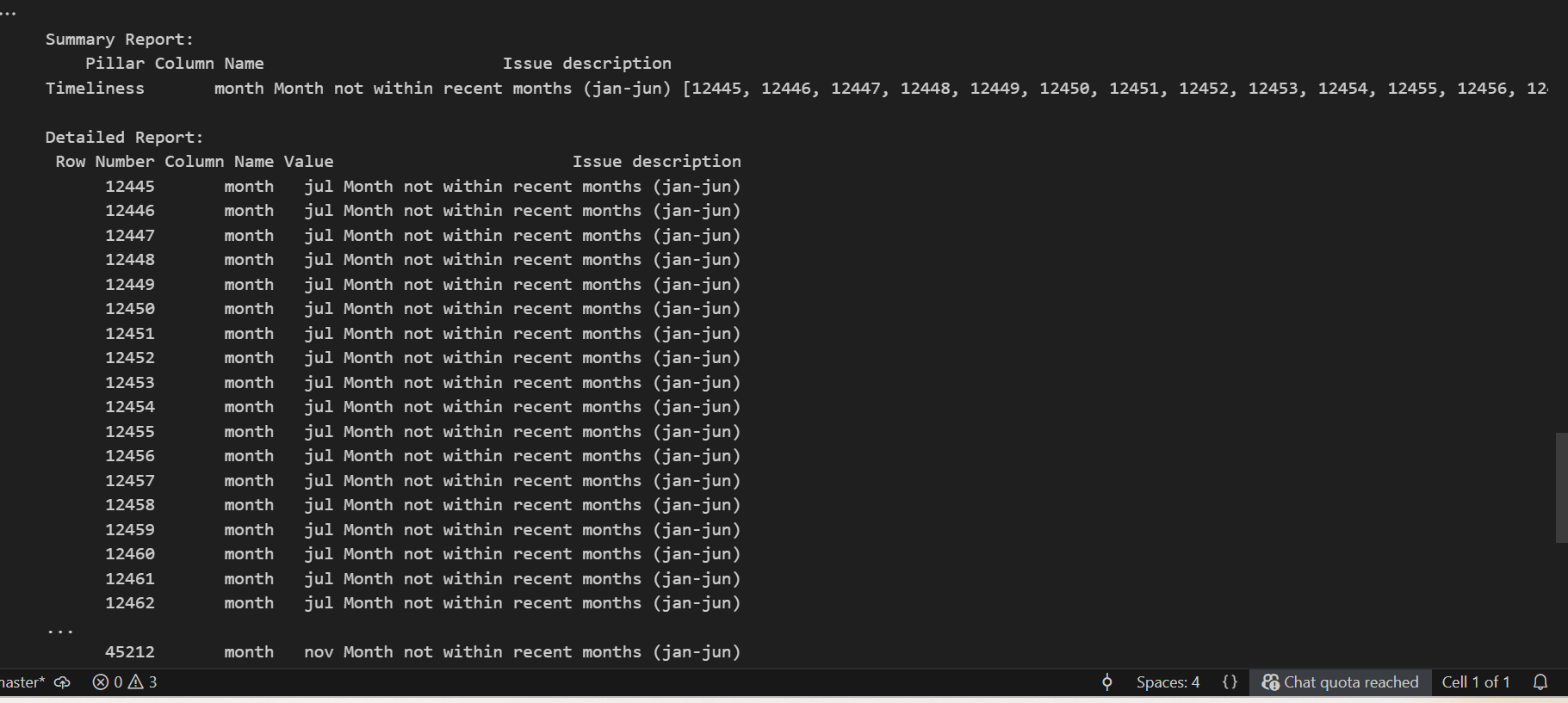


**Test Report: (In Excel)**

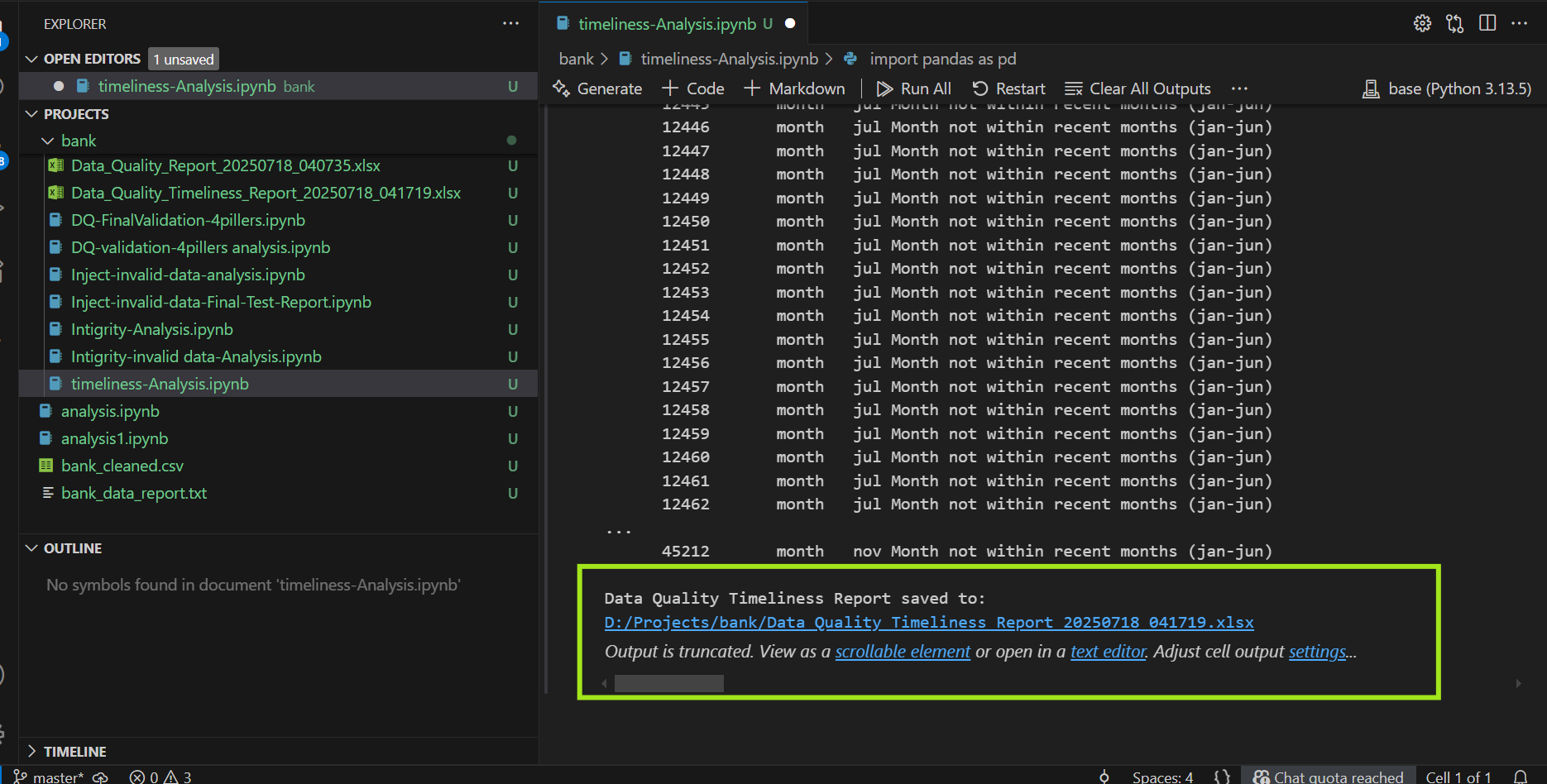
Test result generated as report with updated date and timestamp

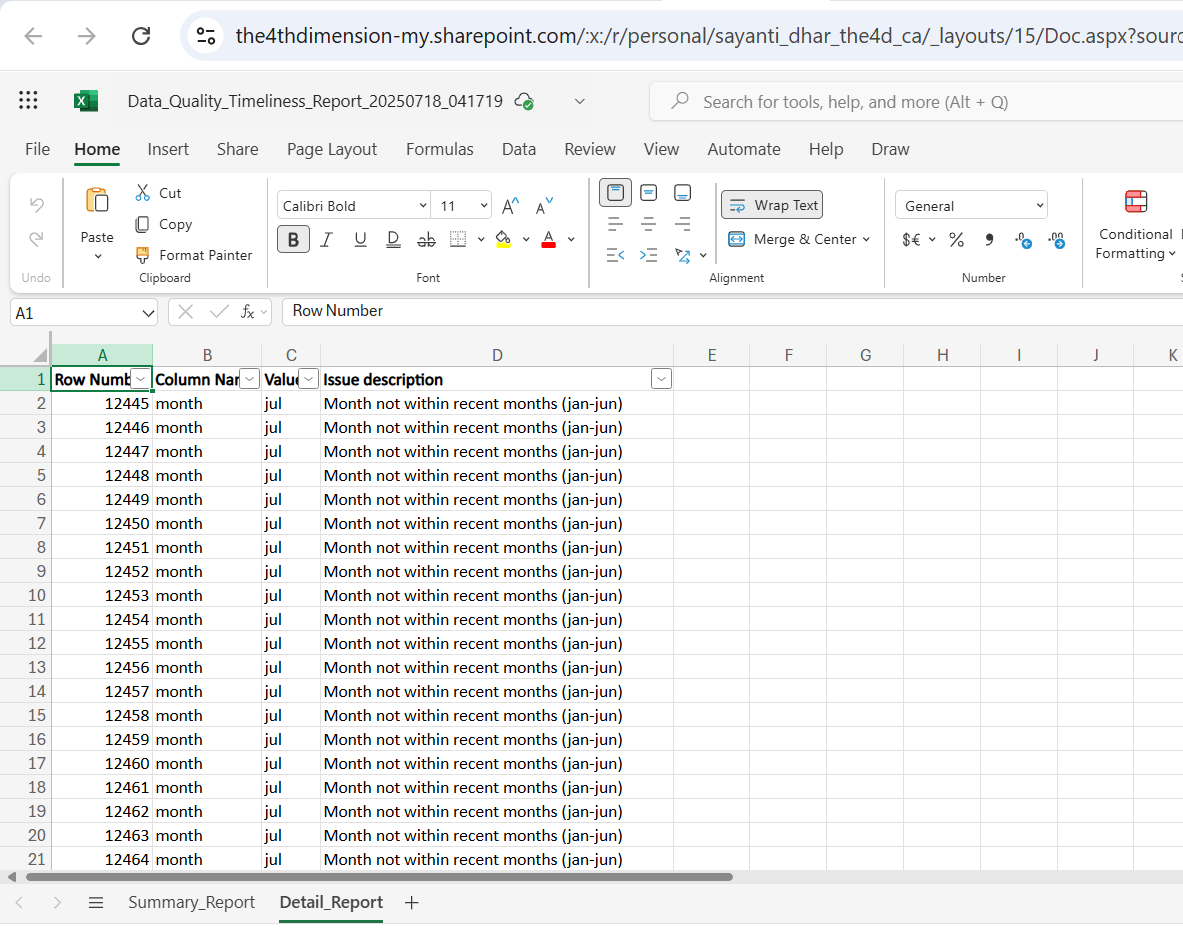
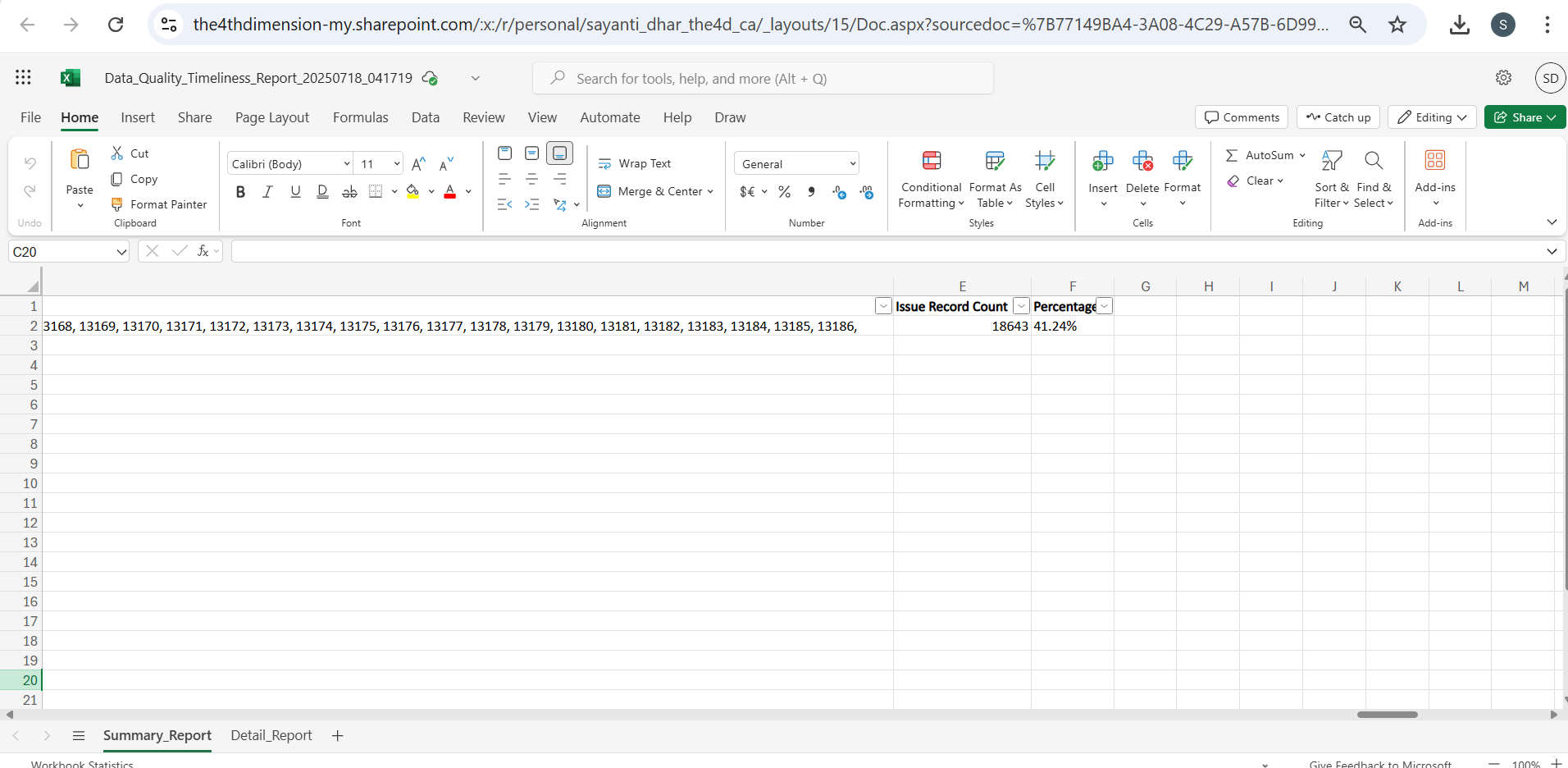
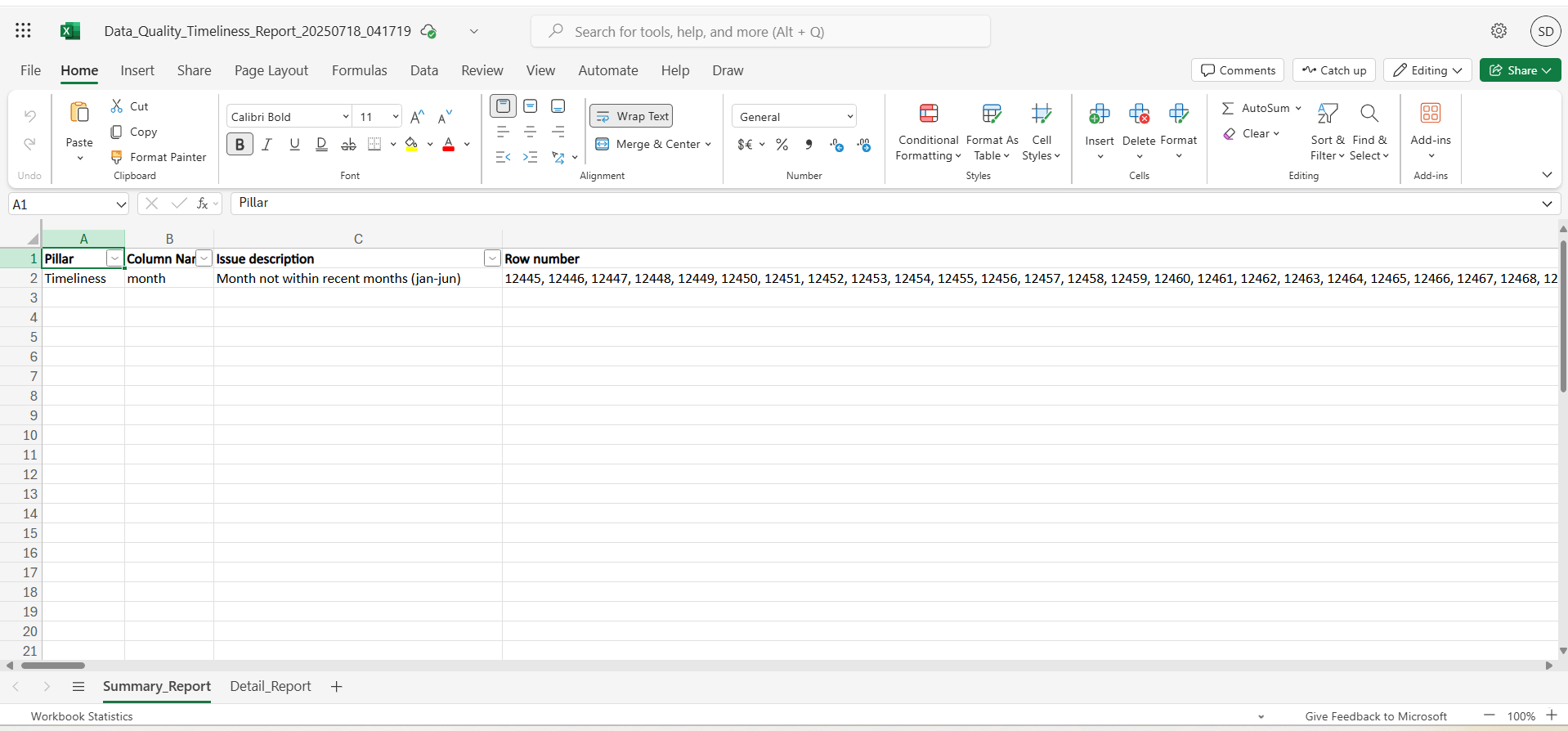


**Data validation based on Timeliness:**

**Test Report: (In Excel)**

Test result generated as report with updated date and timestamp.

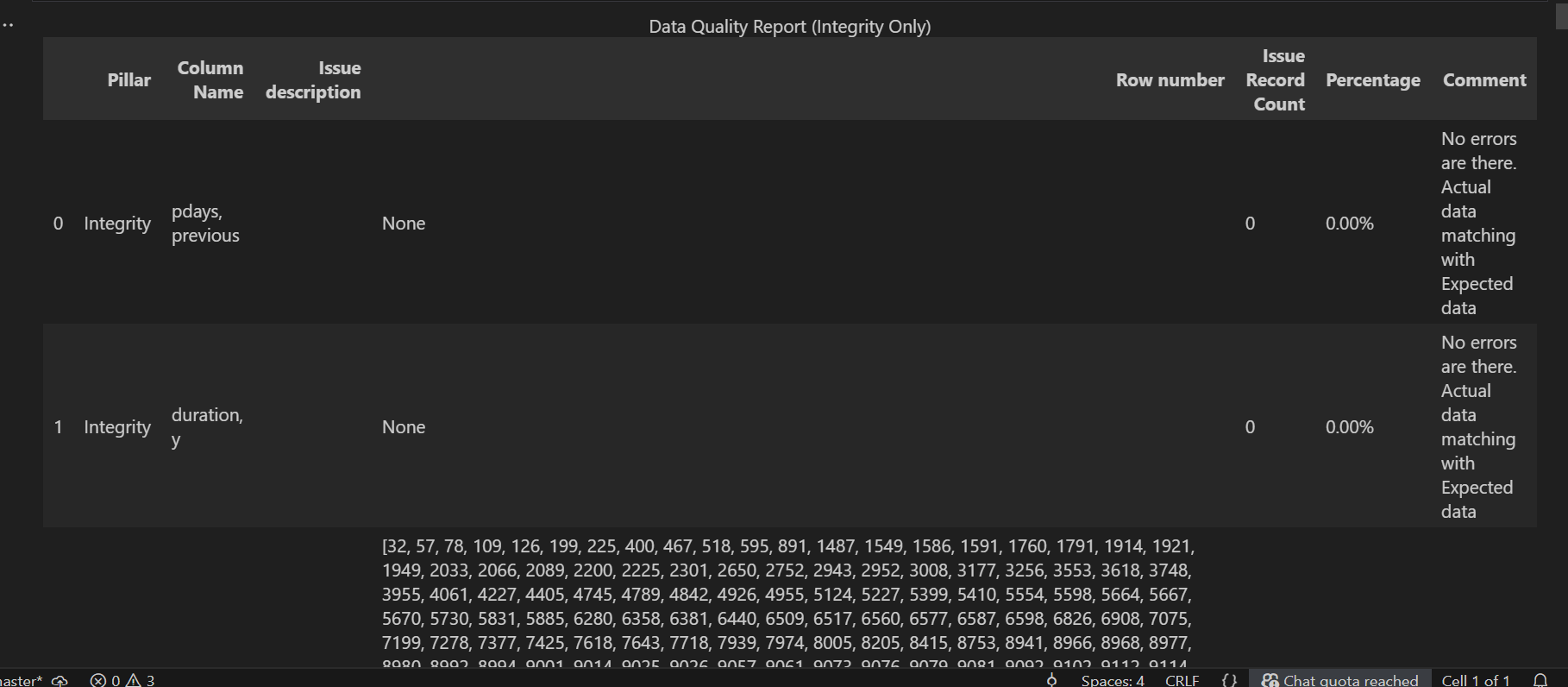


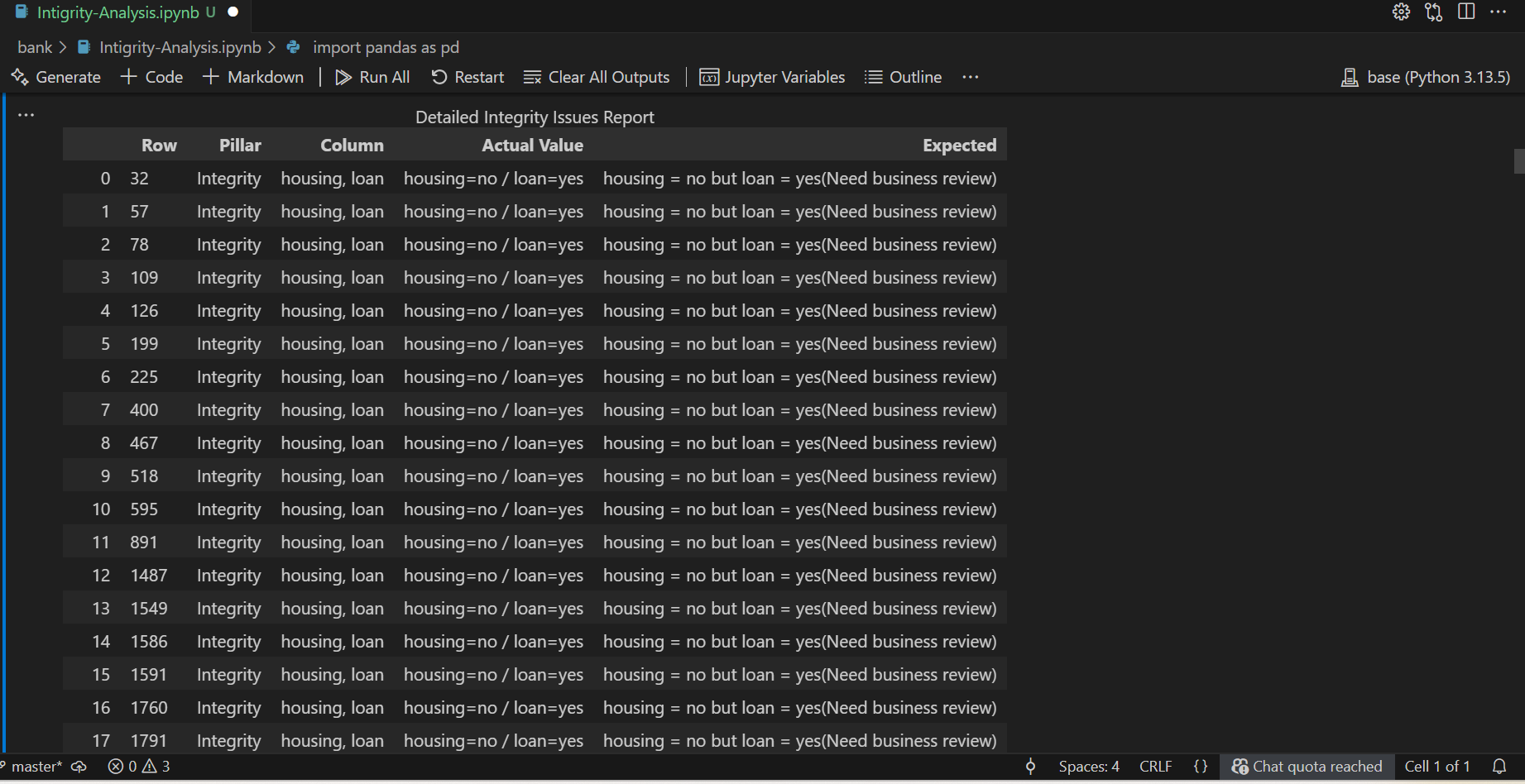


**Data validation based on Integrity:**

**Integrity Rule:**

* **If pdays = -1, previous must be 0**
* **If duration = 0 then y should be no**
* **housing = no and loan = yes [flagged for business review]**





**Test Report: (In Excel)**

Test results generated as a report with updated date and timestamp.

