Qualitative Coding

To code the characteristics of table changes, we first execute the scripts to obtain intermediate tables for each line of wrangling code. Then we compare and record the differences between each pair of input and output tables in detail from various aspects, including their changes in columns, rows, and cells. Here we show three examples of how we code table differences:

Example 1:

InstanceNo: 44

Code: all_by_state.sort_values("per_1k_residents", ascending=False, inplace=True)

Table Differences: The position order of rows has changed. In the input table, the column "state" is in ascending order and the column "per_1k_residents" is in disorder, while in the output table, the column "state" is in disorder and the column "per_1k_residents" is in descending order.

Example 2:

InstanceNo: 212

Code: id_name = id_name[df['online_fac_name'].notnull()]

Table Differences: The number of rows has decreased from 13032 to 5187. The column "online_fac_name" in the input table has empty cells, while the column "online_fac_name" in the output table does not. All rows with empty cells in column "online_fac_name" have been removed.

Example 3:

InstanceNo: 549

Code: df["year"] = df["time"].str[:4]

Table Differences: The number of columns has increased from 6 to 7. The cell values in the created column "year" of the output table are a subset of the cell values in the column "time" of both the input and output tables. Specifically, the cell values in the column "year" are the first 4 characters of the cell values in the column "time".

Next, we use these table differences as our qualitative data and apply open coding and axial coding to them. Specifically, during the open coding process, we turn these differences into small and discrete parts and create codes to label them. For instance, we break the table differences in Example 2 into two parts and rewrite them into two codes, respectively, i.e., the number of rows in the output table is less than the number of rows in the input table and there are no missing values in column *i* of the output table and there are missing values in column *i* of the input table. Here *i* is a parameter that refers to a column. In this case, *i* is "online_fac_name", and we regard codes with different parameter values as the same code. We maintain a codeset

including all unique open codes and update it when encountering a new open code.
In the axial coding process, we draw connections and relationships between these open codes. Based on our observations and wrangling experiences, we adopt the codes which are conducive to distinguishing between different transformations. After frequent discussions with two data scientists, we group and condense related codes into broader categories. Finally, we derive five properties of data changes (i.e., Number, Order, Relation, Value, and Type) and four data objects (i.e., Tables, Columns, Rows, and Cells) from these categories.

The Design Space

To construct a design space for table changes, we employ the properties of data changes and the types of data objects as two primary dimensions in our design space and take the open codes in the codeset as characteristics. Moreover, we extend several additional characteristics which reflect unchanged semantics under the two dimensions for completeness, such as the number of columns in the output table is equal to the number of columns in the input table and the index order of the rows in the output table is the same as the index order of the rows in the input table. Ultimately, there are a total of 103 characteristics in our design space. The description of each characteristic is listed as follows:

Number

Table

1. the number of output tables is greater than zero and the number of input tables is zero
2. the number of output tables is zero and the number of input tables is greater than zero
3. the number of output tables is greater than zero and the number of output tables is equal to the number of input tables
4. the number of output tables is greater than zero and the number of output tables is greater than the number of input tables
5. the number of output tables is greater than zero and the number of output tables is less than the number of input tables

Column

6. the number of columns in the output table is equal to the number of columns in the input table
7. the number of columns in the output table is less than the number of columns in the input table

8. the number of columns in the output table is greater than the number of columns in the input table
9. the minimum number of columns in the output tables is greater than the number of columns in the input table
10. the minimum number of columns in the output tables is equal to the number of columns in the input table
11. the minimum number of columns in the output tables is less than the number of columns in the input table and the maximum number of columns in the output tables is greater than the number of columns in the input table
12. the maximum number of columns in the output tables is equal to the number of columns in the input table
13. the maximum number of columns in the output tables is less than the number of columns in the input table and the sum of the number of columns in the output tables is greater than the number of columns in the input table
14. the sum of the number of columns in the output tables is equal to the number of columns in the input table
15. the sum of the number of columns in the output tables is less than the number of columns in the input table
16. the number of columns in the output table is less than the minimum number of columns in the input tables
17. the number of columns in the output table is equal to the minimum number of columns in the input tables
18. the number of columns in the output table is greater than the minimum number of columns in the input tables and the number of columns in the output table is less than the maximum number of columns in the input tables
19. the number of columns in the output table is equal to the maximum number of columns in the input tables
20. the number of columns in the output table is greater than the maximum number of columns in the input tables and the number of columns in the output table is less than the sum of the number of columns in the input tables
21. the number of columns in the output table is equal to the sum of the number of columns in the input tables
22. the number of columns in the output table is greater than the sum of the number of columns in the input tables

Row

23. the number of rows in the output table is equal to the number of rows in the input table
24. the number of rows in the output table is less than the number of rows in the input table

25. the number of rows in the output table is greater than the number of rows in the input table
26. the minimum number of rows in the output tables is greater than the number of rows in the input table
27. the minimum number of rows in the output tables is equal to the number of rows in the input table
28. the minimum number of rows in the output tables is less than the number of rows in the input table and the maximum number of rows in the output tables is greater than the number of rows in the input table
29. the maximum number of rows in the output tables is equal to the number of rows in the input table
30. the maximum number of rows in the output tables is less than the number of rows in the input table and the sum of the number of rows in the output tables is greater than the number of rows in the input table
31. the sum of the number of rows in the output tables is equal to the number of rows in the input table
32. the sum of the number of rows in the output tables is less than the number of rows in the input table
33. the number of rows in the output table is less than the minimum number of rows in the input tables
34. the number of rows in the output table is equal to the minimum number of rows in the input tables
35. the number of rows in the output table is greater than the minimum number of rows in the input tables and the number of rows in the output table is less than the maximum number of rows in the input tables
36. the number of rows in the output table is equal to the maximum number of rows in the input tables
37. the number of rows in the output table is greater than the maximum number of rows in the input tables and the number of rows in the output table is less than the sum of the number of rows in the input tables
38. the number of rows in the output table is equal to the sum of the number of rows in the input tables
39. the number of rows in the output table is greater than the sum of the number of rows in the input tables

Order

Column

40. the index order of the columns in the output table is the same as the index order of the columns in the input table

41. the index order of the columns in the output table is different from the index order of the columns in the input table
42. column *i* in the output table is in disorder
43. column *i* in the output table is in ascending order
44. column *i* in the output table is in descending order

Row

45. the index order of the rows in the output table is the same as the index order of the rows in the input table
46. the index order of the rows in the output table is different from the index order of the rows in the input table
47. row *i* in the output table is in disorder
48. row *i* in the output table is in ascending order
49. row *i* in the output table is in descending order

Relation

Column

50. the contents of column *i* in the output table are a subset of the contents of column *j* in the input table
51. the contents of column *i* in the output table are a superset of the contents of column *j* in the input table
52. the contents of column *i* in the output table are equal to the contents of column *j* in the input table
53. the contents of column *i* in the output table have other relations to the contents of column *j* in the input table
54. there are no duplicate columns in the output table and there are duplicate columns in the input table
55. there are duplicate columns in the output table and there are no duplicate columns in the input table
56. there are no duplicate columns in the output table and there are no duplicate columns in the input table
57. there are duplicate columns in the output table and there are duplicate columns in the input table

Row

58. the contents of row *i* in the output table are a subset of the contents of row *j* in the input table
59. the contents of row *i* in the output table are a superset of the contents of row *j* in the input table
60. the contents of row *i* in the output table are equal to the contents of row *j* in the input table

61. the contents of row *i* in the output table have other relations to the contents of row *j* in the input table
62. there are no duplicate rows in the output table and there are duplicate rows in the input table
63. there are duplicate rows in the output table and there are no duplicate rows in the input table
64. there are no duplicate rows in the output table and there are no duplicate rows in the input table
65. there are duplicate rows in the output table and there are duplicate rows in the input table

Cell

66. the cell content of row *i* and column *j* in the output table is a subset of the cell content of row *p* and column *q* in the input table
67. the cell content of row *i* and column *j* in the output table is a superset of the cell content of row *p* and column *q* in the input table
68. the cell content of row *i* and column *j* in the output table is equal to the cell content of row *p* and column *q* in the input table
69. the cell content of row *i* and column *j* in the output table has other relations to the cell content of row *p* and column *q* in the input table
70. there are no duplicate cells in column *i* of the output table and there are duplicate cells in column *i* of the input table
71. there are duplicate cells in column *i* of the output table and there are no duplicate cells in column *i* of the input table
72. there are no duplicate cells in column *i* of the output table and there are no duplicate cells in column *i* of the input table
73. there are duplicate cells in column *i* of the output table and there are duplicate cells in column *i* of the input table
74. there are no duplicate cells in row *i* of the output table and there are duplicate cells in row *i* of the input table
75. there are duplicate cells in row *i* of the output table and there are no duplicate cells in row *i* of the input table
76. there are no duplicate cells in row *i* of the output table and there are no duplicate cells in row *i* of the input table
77. there are duplicate cells in row *i* of the output table and there are duplicate cells in row *i* of the input table

Value

Table

78. there are no missing values in the output table and there are missing values in the input table
79. there are missing values in the output table and there are no missing values in the input table
80. there are no missing values in the output table and there are no missing values in the input table
81. there are missing values in the output table and there are missing values in the input table
82. there are specific values *val* in the output table
83. there are no specific values *val* in the output table

Column

84. there are no missing values in column *i* of the output table and there are missing values in column *i* of the input table
85. there are missing values in column *i* of the output table and there are no missing values in column *i* of the input table
86. there are no missing values in column *i* of the output table and there are no missing values in column *i* of the input table
87. there are missing values in column *i* of the output table and there are missing values in column *i* of the input table
88. there are specific values *val* in column *i* of the output table
89. there are no specific values *val* in column *i* of the output table

Row

90. there are no missing values in row *i* of the output table and there are missing values in row *i* of the input table
91. there are missing values in row *i* of the output table and there are no missing values in row *i* of the input table
92. there are no missing values in row *i* of the output table and there are no missing values in row *i* of the input table
93. there are missing values in row *i* of the output table and there are missing values in row *i* of the input table
94. there are specific values *val* in row *i* of the output table
95. there are no specific values *val* in row *i* of the output table

Cell

96. the cell content of row *i* and column *j* in the output table is not a missing value and the cell content of row *i* and column *j* in the input table is a missing value
97. the cell content of row *i* and column *j* in the output table is a missing value and the cell content of row *i* and column *j* in the input table is not a missing value

98. the cell content of row *i* and column *j* in the output table is not a missing value and the cell content of row *i* and column *j* in the input table is not a missing value
99. the cell content of row *i* and column *j* in the output table is a missing value and the cell content of row *i* and column *j* in the input table is a missing value
100. the cell content of row *i* and column *j* in the output table is *val1* and the cell content of row *i* and column *j* in the input table is *val2*
101. the cell content of row *i* and column *j* in the output table is not *val1* and the cell content of row *i* and column *j* in the input table is *val2*

Type

Column

102. the data type of column *i* in the output table is the same as the data type of column *i* in the input table
103. the data type of column *i* in the output table is different from the data type of column *i* in the input table