Income and Expenditure Survey 2022-2023 v1.1 cleaning process

1. Key problem

The main problem with version 1 of the IES 2022-2023 data is that variables that should be in numeric format were sometimes coded as strings. This led to issues with merging, as the UQNO variable (household identifier) was not the same format across the different data files. Version 1.1 data files solve this issue, with variable types being assigned their appropriate formats.

2. The UQNO variable

Version 1 of the data faced problems concerning the UQNO variable. In the household data file, this variable was in string format and was uniquely identifiable. However, when using the "destring" command to convert it to numeric format, it was no longer uniquely identifiable.

This issue has been fixed in version 1.1. The problem was that using the "destring" command leads to a loss of precision, particularly with numbers that have several digits. This is because "destring" saves the resulting numeric variable in the double format. The solution was to use the "encode" command. The "encode" command converts the string to a numeric variable in the long format. Below is an illustration of this. We have UQNO (string format), testing_uqno (numeric using "destring" command) and testing2_uqno (numeric using "encode" command):

	UQNO	testing_uqno	testing2_uqno
1	160100330000009401	160100330000009408	160100330000009401
2	160100330000011901	160100330000011904	160100330000011901
3	160100630000004401	160100630000004416	160100630000004401
4	160100630000008001	160100630000008000	160100630000008001
5	160100630000011601	160100630000011616	160100630000011601
6	160100630000015201	160100630000015200	160100630000015201
7	160100630000018801	160100630000018816	160100630000018801
8	160101410000002601	160101410000002592	160101410000002601
9	160101410000009601	160101410000009600	160101410000009601
10	160101410000013101	160101410000013088	160101410000013101
11	160101410000016601	160101410000016608	160101410000016601
12	161100010000028301	161100010000028288	161100010000028301
13	161100010000048301	161100010000048288	161100010000048301
14	161100270000001101	161100270000001088	161100270000001101
15	161100270000008301	161100270000008288	161100270000008301
16	161100270000011901	161100270000011904	161100270000011901
17	161100270000015501	161100270000015488	161100270000015501
18	161100270000019101	161100270000019104	161100270000019101
19	161100270000022701	161100270000022688	161100270000022701
20	161100270000026301	161100270000026304	161100270000026301

From the image, it is clear that testing2_uqno preserves the values from UQNO while testing_uqno augments them. Using encode to convert UQNO results in the variable remaining uniquely identifiable.

3. Incorrect format of several variables across data files

The image below illustrates how some of the variables are in the wrong formats. This was taken from the person data file:

	UQNO	PERSON_ID	PERSONNO	SEX	AGE	POPULATION	LANGUAGES	HHC_RELATI~P	HHC_MARITAL	SPOUSE	SPOUSE_NAME
1	160100330000009401	1601003300000940101	01	2	46	4	1	1	3	8	88
2	160100330000011901	1601003300001190101	01	2	53	4	1	1	3	8	88
3	160100630000004401	1601006300000440101	01	2	58	2	1	1	6	8	88
4	160100630000004401	1601006300000440102	02	1	81	2	1	5	5	8	88
5	160100630000004401	1601006300000440103	03	2	38	2	1	3	7	8	88
6	160100630000004401	1601006300000440104	04	2	35	2	1	3	7	8	88
7	160100630000004401	1601006300000440105	05	2	13	2	1	7	7	8	88
8	160100630000004401	1601006300000440106	06	2	9	2	1	7	8	8	88
9	160100630000004401	1601006300000440107	07	2	7	2	1	7	8	8	88
10	160100630000004401	1601006300000440108	08	1	7	2	1	7	8	8	88
11	160100630000008001	1601006300000800101	01	2	28	2	1	1	6	8	88
12	160100630000008001	1601006300000800102	02	1	8	2	1	3	8	8	88
13	160100630000011601	1601006300001160101	01	1	44	2	1	1	1	1	2
14	160100630000011601	1601006300001160102	02	2	39	2	1	2	1	1	1
15	160100630000011601	1601006300001160103	03	1	20	2	1	3	7	8	88
16	160100630000011601	1601006300001160104	04	2	15	2	1	3	7	8	88
17	160100630000011601	1601006300001160105	05	1	14	2	1	3	7	8	88
18	160100630000015201	1601006300001520101	01	2	62	2	1	1	1	1	2
19	160100630000015201	1601006300001520102	02	1	59	2	1	2	1	1	1
20	160100630000015201	1601006300001520103	03	1	23	2	1	7	7	8	88
21	160100630000015201	1601006300001520104	04	1	14	2	1	7	7	8	88
22	160100630000018801	1601006300001880101	01	2	22	2	1	1	7	8	88
23	160101410000002601	1601014100000260101	01	2	48	2	1	1	7	8	88
24	160101410000002601	1601014100000260102	02	2	22	2	1	3	7	8	88
25	160101410000002601	1601014100000260103	03	2	16	2	1	3	7	8	88
26	160101410000009601	1601014100000960101	01	2	65	2	1	1	1	1	2
27	160101410000009601	1601014100000960102	02	1	68	2	1	2	1	1	1
28	160101410000009601	1601014100000960103	03	2	37	2	1	3	7	8	88
29	160101410000009601	1601014100000960104	04	2	13	2	1	7	7	8	88
30	160101410000013101	1601014100001310101	01	1	35	2	1	1	6	8	88
31	160101410000016601	1601014100001660101	01	1	42	2	1	1	7	8	88
32	161100010000028301	1611000100002830101	01	1	72	2	1	1	1	1	2
33	161100010000028301	1611000100002830102	02	2	75	2	1	2	1	1	1
34	161100010000028301	1611000100002830103	03	2	21	2	1	3	7	8	88
35	161100010000048301	1611000100004830101	01	1	44	2	1	1	2	1	2

The red numbers mean that the variable is in the string format, while the blue numbers mean the variable is a numeric (normally, numeric variables show as black. Entries appear blue when a numeric variable has been labelled, with the blue text being the label). This leads to problems when labelling, as strings cannot be labelled. Version 1.1 fixes this issue. Below is the result after the cleaning and labelling:

	UQNO	PERSON_ID	PERSONNO	SEX	AGE	POPULATION	LANGUAGES	HHC_RELATIONSHIP
1	160100330000009401	1601003300000940101	1	Female	46	White	Afrikaans	Head/acting head
2	160100330000011901	1601003300001190101	1	Female	53	White	Afrikaans	Head/acting head
3	160100630000004401	1601006300000440101	1	Female	58	Coloured	Afrikaans	Head/acting head
4	160100630000004401	1601006300000440102	2	Male	81	Coloured	Afrikaans	Father/mother/stepfather/stepmother
5	160100630000004401	1601006300000440103	3	Female	38	Coloured	Afrikaans	Son/daughter/stepchild/adopted child
6	160100630000004401	1601006300000440104	4	Female	35	Coloured	Afrikaans	Son/daughter/stepchild/adopted child
7	160100630000004401	1601006300000440105	5	Female	13	Coloured	Afrikaans	Grandchild/great grandchild
8	160100630000004401	1601006300000440106	6	Female	9	Coloured	Afrikaans	Grandchild/great grandchild
9	160100630000004401	1601006300000440107	7	Female	7	Coloured	Afrikaans	Grandchild/great grandchild
10	160100630000004401	1601006300000440108	8	Male	7	Coloured	Afrikaans	Grandchild/great grandchild
11	160100630000008001	1601006300000800101	1	Female	28	Coloured	Afrikaans	Head/acting head
12	160100630000008001	1601006300000800102	2	Male	8	Coloured	Afrikaans	Son/daughter/stepchild/adopted child
13	160100630000011601	1601006300001160101	1	Male	44	Coloured	Afrikaans	Head/acting head
14	160100630000011601	1601006300001160102	2	Female	39	Coloured	Afrikaans	Husband/wife/partner
15	160100630000011601	1601006300001160103	3	Male	20	Coloured	Afrikaans	Son/daughter/stepchild/adopted child
16	160100630000011601	1601006300001160104	4	Female	15	Coloured	Afrikaans	Son/daughter/stepchild/adopted child
17	160100630000011601	1601006300001160105	5	Male	14	Coloured	Afrikaans	Son/daughter/stepchild/adopted child
18	160100630000015201	1601006300001520101	1	Female	62	Coloured	Afrikaans	Head/acting head
19	160100630000015201	1601006300001520102	2	Male	59	Coloured	Afrikaans	Husband/wife/partner
20	160100630000015201	1601006300001520103	3	Male	23	Coloured	Afrikaans	Grandchild/great grandchild
21	160100630000015201	1601006300001520104	4	Male	14	Coloured	Afrikaans	Grandchild/great grandchild
22	160100630000018801	1601006300001880101	1	Female	22	Coloured	Afrikaans	Head/acting head
23	160101410000002601	1601014100000260101	1	Female	48	Coloured	Afrikaans	Head/acting head
24	160101410000002601	1601014100000260102	2	Female	22	Coloured	Afrikaans	Son/daughter/stepchild/adopted child
25	160101410000002601	1601014100000260103	3	Female	16	Coloured	Afrikaans	Son/daughter/stepchild/adopted child
26	160101410000009601	1601014100000960101	1	Female	65	Coloured	Afrikaans	Head/acting head
27	160101410000009601	1601014100000960102	2	Male	68	Coloured	Afrikaans	Husband/wife/partner
28	160101410000009601	1601014100000960103	3	Female	37	Coloured	Afrikaans	Son/daughter/stepchild/adopted child
29	160101410000009601	1601014100000960104	4	Female	13	Coloured	Afrikaans	Grandchild/great grandchild
30	160101410000013101	1601014100001310101	1	Male	35	Coloured	Afrikaans	Head/acting head
31	160101410000016601	1601014100001660101	1	Male	42	Coloured	Afrikaans	Head/acting head
32	161100010000028301	1611000100002830101	1	Male	72	Coloured	Afrikaans	Head/acting head
33	161100010000028301	1611000100002830102	2	Female	75	Coloured	Afrikaans	Husband/wife/partner
34	161100010000028301	1611000100002830103	3	Female	21	Coloured	Afrikaans	Son/daughter/stepchild/adopted child
35	161100010000048301	1611000100004830101	1	Male	44	Coloured	Afrikaans	Head/acting head
36	161100010000048301	1611000100004830102	2	Female	41	Coloured	Afrikaans	Husband/wife/partner
37	161100010000048301	1611000100004830103	3	Female	20	Coloured	Afrikaans	Son/daughter/stepchild/adopted child
38	161100010000048301	1611000100004830104	4	Female	17	Coloured	Afrikaans	Son/daughter/stepchild/adopted child
39	161100010000048301	1611000100004830105	5	Male	12	Coloured	Afrikaans	Son/daughter/stepchild/adopted child
40	161100010000048301	1611000100004830106	6	Female	4	Coloured	Afrikaans	Grandchild/great grandchild
41	161100010000048301	1611000100004830107	7	Male	0	Coloured	Afrikaans	Grandchild/great grandchild

4. Compromise

One thing to note from the previous image is that the UQNO variable displays as blue, when we would expect it to display as black. This is because the “encode” command generates a number per string and then turns the string into the label. Thus, in the household data file, there are 19940 unique observations and so the UQNO variable ranges from 1 to 19940:

```
. sum UQNO
```

Variable	Obs	Mean	Std. dev.	Min	Max
UQNO	19,940	9970.5	5756.327	1	19940

The image below shows this clearly, with the first ten entries of the UQNO variable being displayed with their corresponding labels (taken from the household data file):

```
. tab UQNO if UQNO <= 10
```

Household identifier	Freq.	Percent	Cum.
1. 160100330000009401	1	10.00	10.00
2. 160100330000011901	1	10.00	20.00
3. 160100630000004401	1	10.00	30.00
4. 160100630000008001	1	10.00	40.00
5. 160100630000011601	1	10.00	50.00
6. 160100630000015201	1	10.00	60.00
7. 160100630000018801	1	10.00	70.00
8. 160101410000002601	1	10.00	80.00
9. 160101410000009601	1	10.00	90.00
10. 160101410000013101	1	10.00	100.00
Total	10	100.00	

This is not an issue because the encoding is consistent across the data files. For instance, this same table will be the same across all the other data files that have the UQNO variable. Thus, merging is possible:

Geography file:

```
. tab UQNO if UQNO <= 10
```

Household identifier	Freq.	Percent	Cum.
1. 160100330000009401	1	10.00	10.00
2. 160100330000011901	1	10.00	20.00
3. 160100630000004401	1	10.00	30.00
4. 160100630000008001	1	10.00	40.00
5. 160100630000011601	1	10.00	50.00
6. 160100630000015201	1	10.00	60.00
7. 160100630000018801	1	10.00	70.00
8. 160101410000002601	1	10.00	80.00
9. 160101410000009601	1	10.00	90.00
10. 160101410000013101	1	10.00	100.00
Total	10	100.00	

Person file:

```
. tab UQNO if UQNO <= 10
```

Household identifier	Freq.	Percent	Cum.
1. 160100330000009401	1	3.33	3.33
2. 160100330000011901	1	3.33	6.67
3. 160100630000004401	8	26.67	33.33
4. 160100630000008001	2	6.67	40.00
5. 160100630000011601	5	16.67	56.67
6. 160100630000015201	4	13.33	70.00
7. 160100630000018801	1	3.33	73.33
8. 160101410000002601	3	10.00	83.33
9. 160101410000009601	4	13.33	96.67
10. 160101410000013101	1	3.33	100.00
Total	30	100.00	

Note: There are multiple frequencies per UQNO because multiple people may belong to a single household.

This is consistent across any cross section of the data files. Thus, the encoding is consistent across data files[1]. One possible option for future cleaning is to remove these value labels entirely.

5. The cleaning process

Steps taken to obtain version 1.1:

1. Download csv files from Stats SA Isibalo data portal
2. Import data files into Stata, ensuring id variables imported as strings and numbers are imported as numeric
3. Convert id variables to numeric using “encode” command.
4. Label data files using “20260420- ies20222023-curation-dofile” do file
5. After labelling, save the data files as version 1.1

[1] In STATA, encoding can only take place when a variable has 65536 unique values or less. There are 70339 unique person ID’s. Hence, the person ID variable was left as a string across data files.