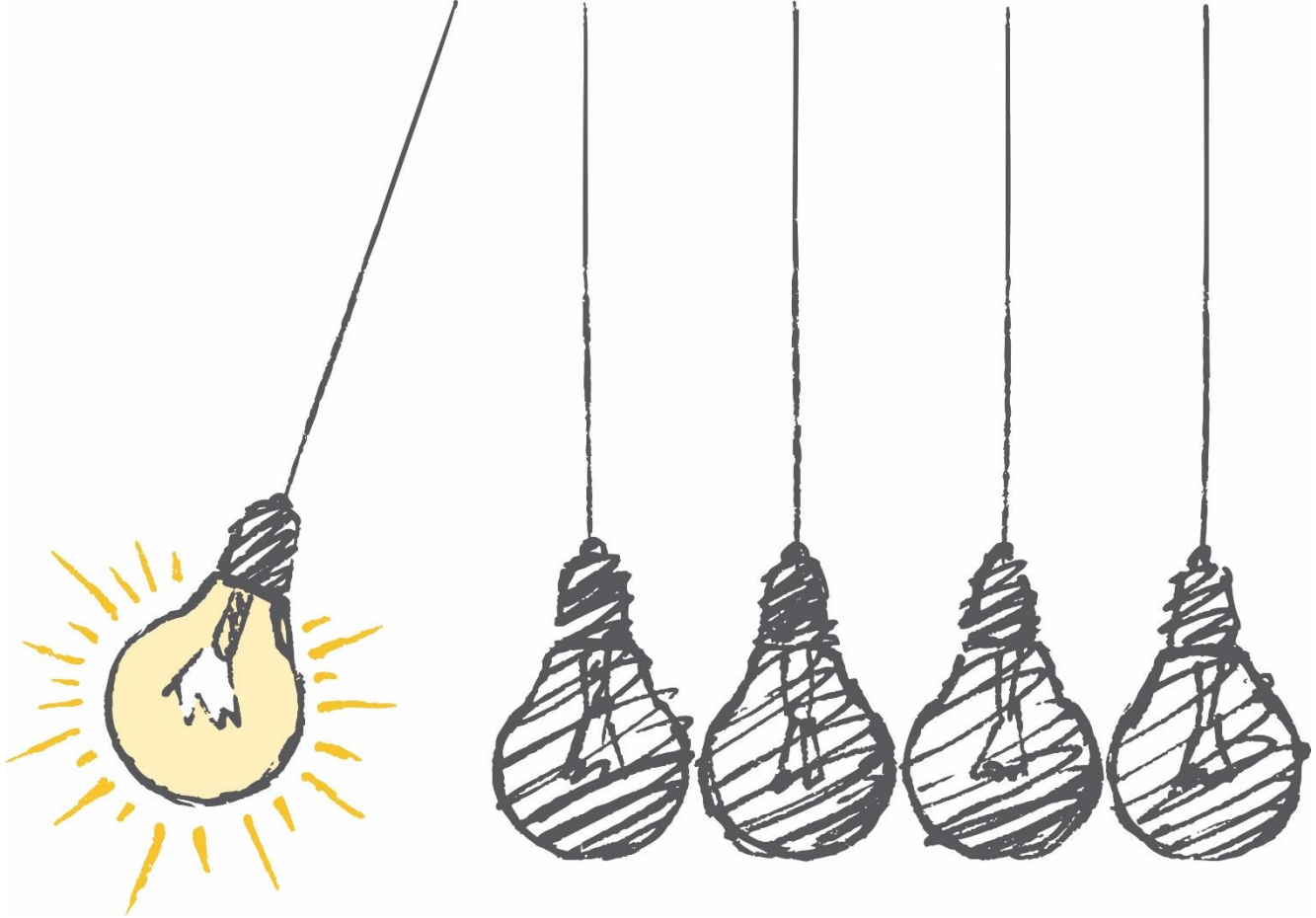
# A workflow for data documentation using Stata



Prepared by **DIME Analytics**
*dimeanalytics@worldbank.org*

Presented by **Luiza Cardoso de Andrade**
*lcardoso@worldbank.org*

WORLD BANK GROUP

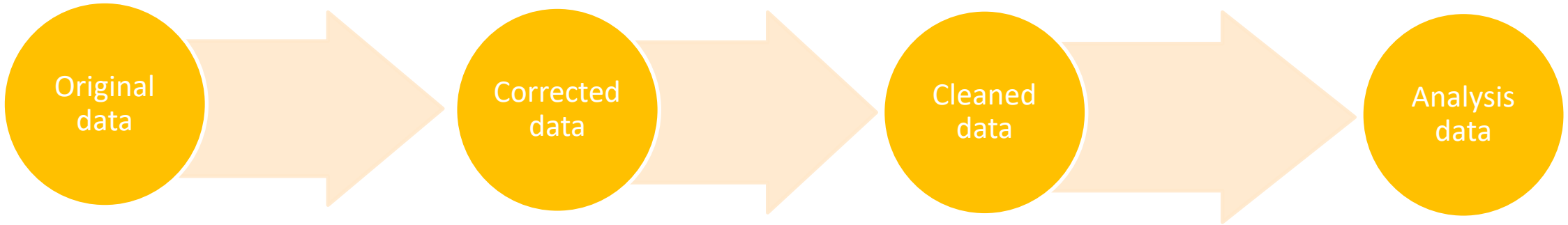
UKaid
from the British people

Norad

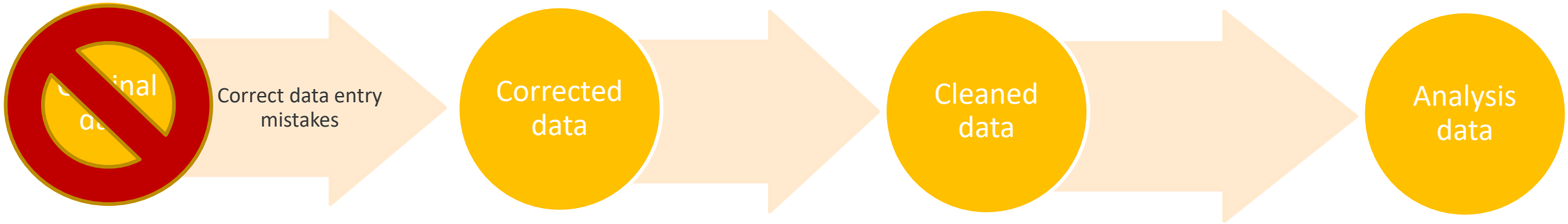
# Background

- DIME Analytics maintains two Stata packages that implement an opinionated data workflow
  - `iefieldkit`: a set of commands for data collection and processing
  - `ietoolkit`: a set of commands for data management and analysis

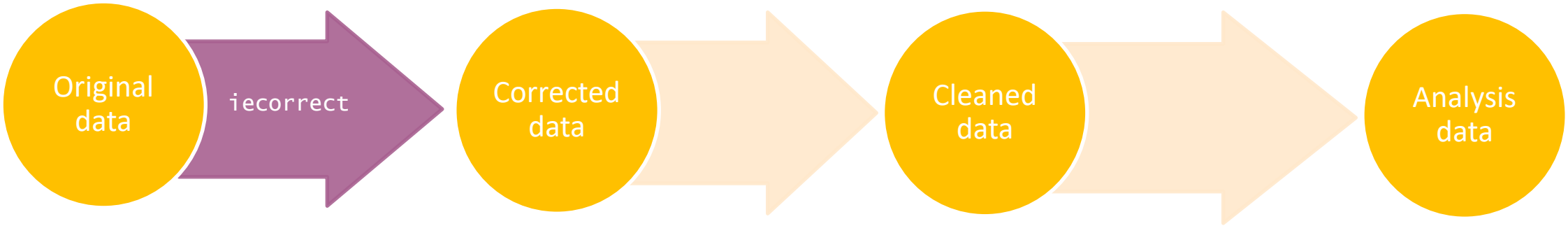
# A model of research data work



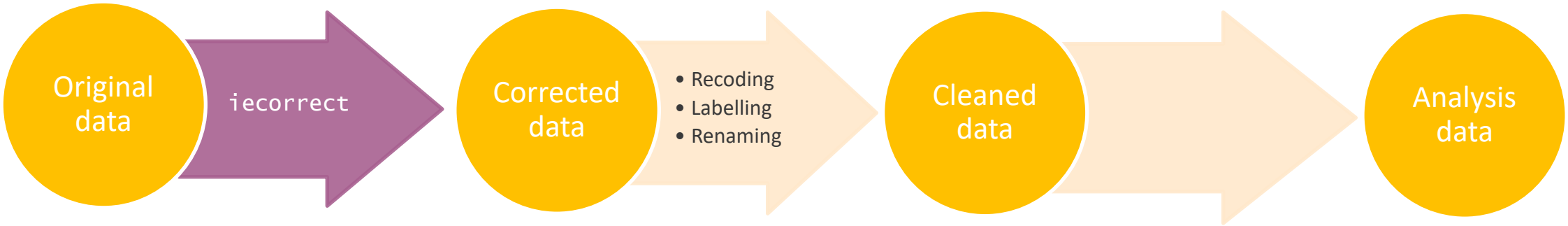
# A model of research data work



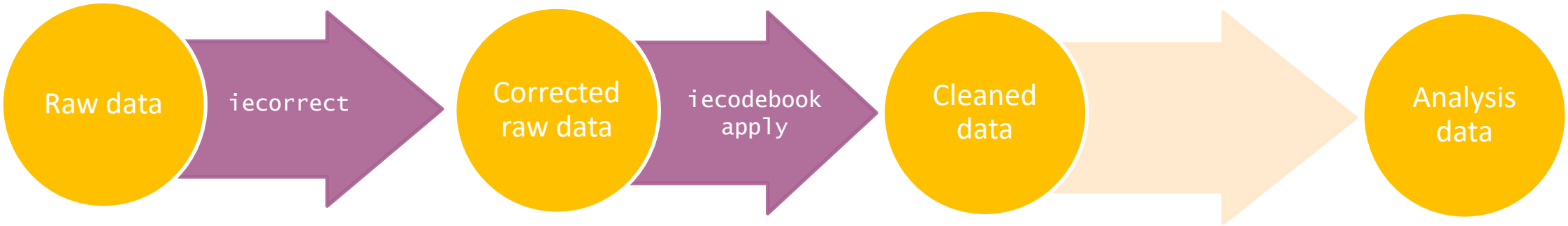
# A model of research data work



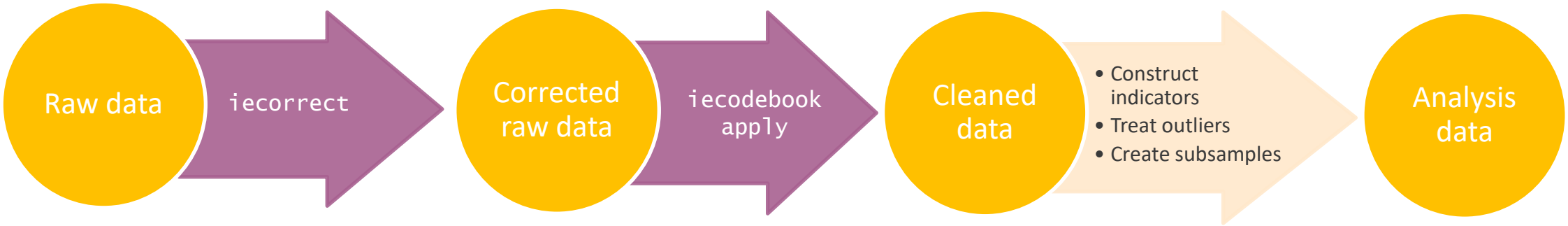
# A model of research data work



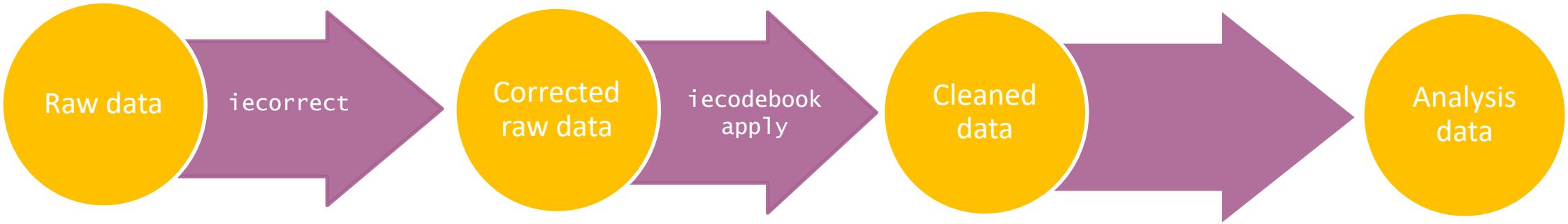
# A model of research data work



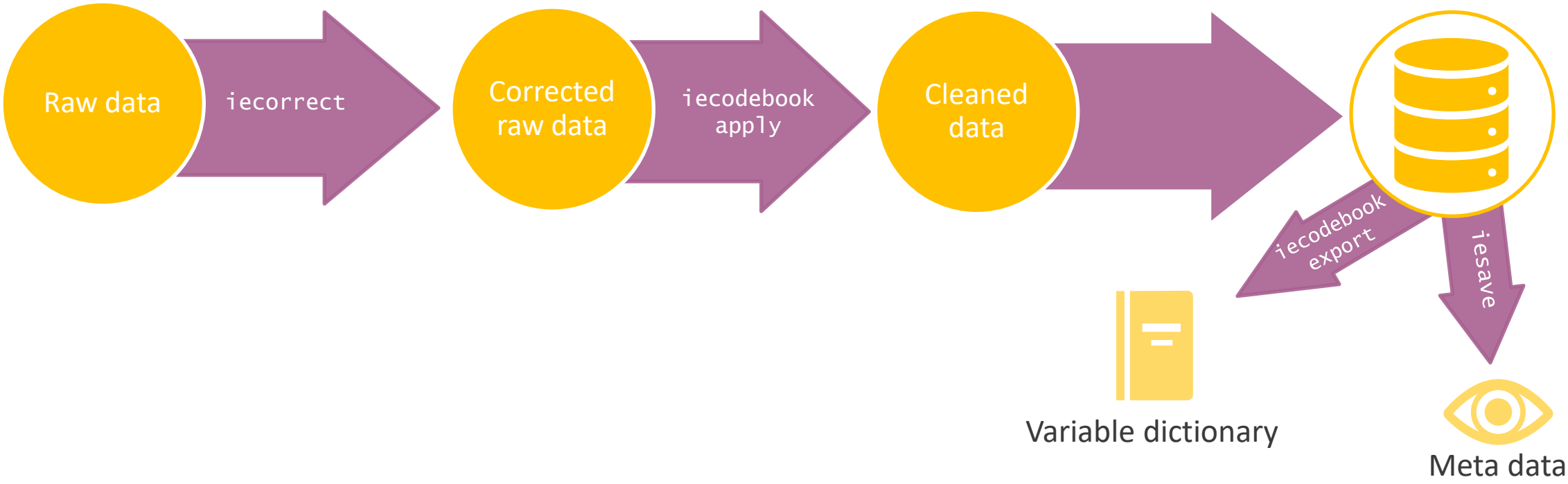
# A model of research data work



# A model of research data work



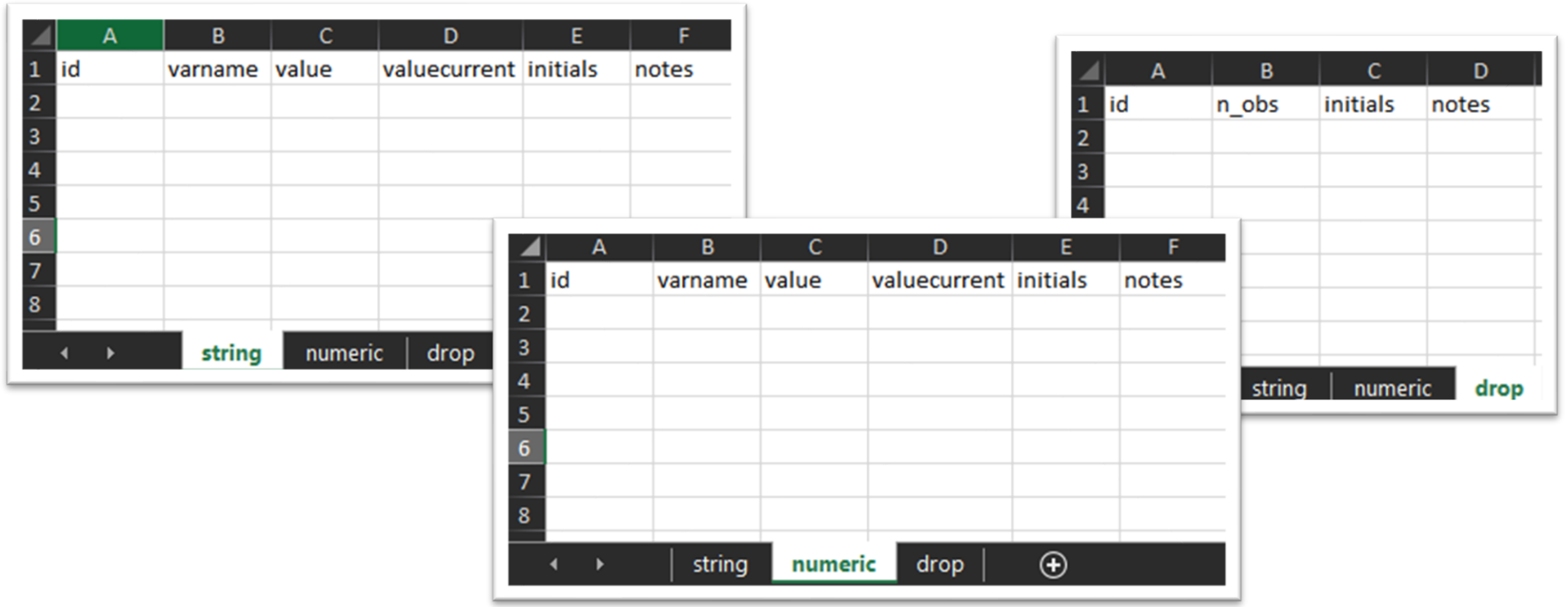
# A model of research data work



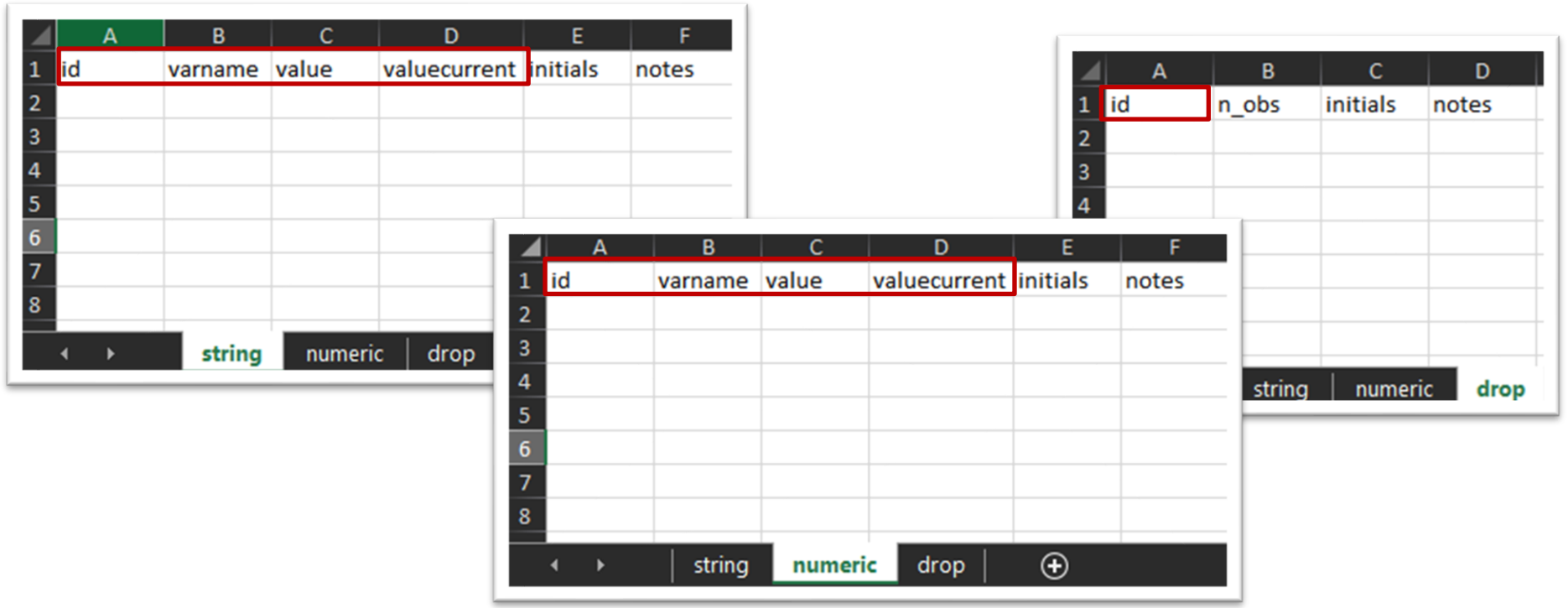
# `iecorrect:`

Modify data points in a dataset using an external human-readable changelog (spreadsheet) and maintain non-code documentation for all manual data point edits

```
iecorrect template using ///
    "/path/to/corrections/file.xlsx", ///
    idvar(varlist)
```

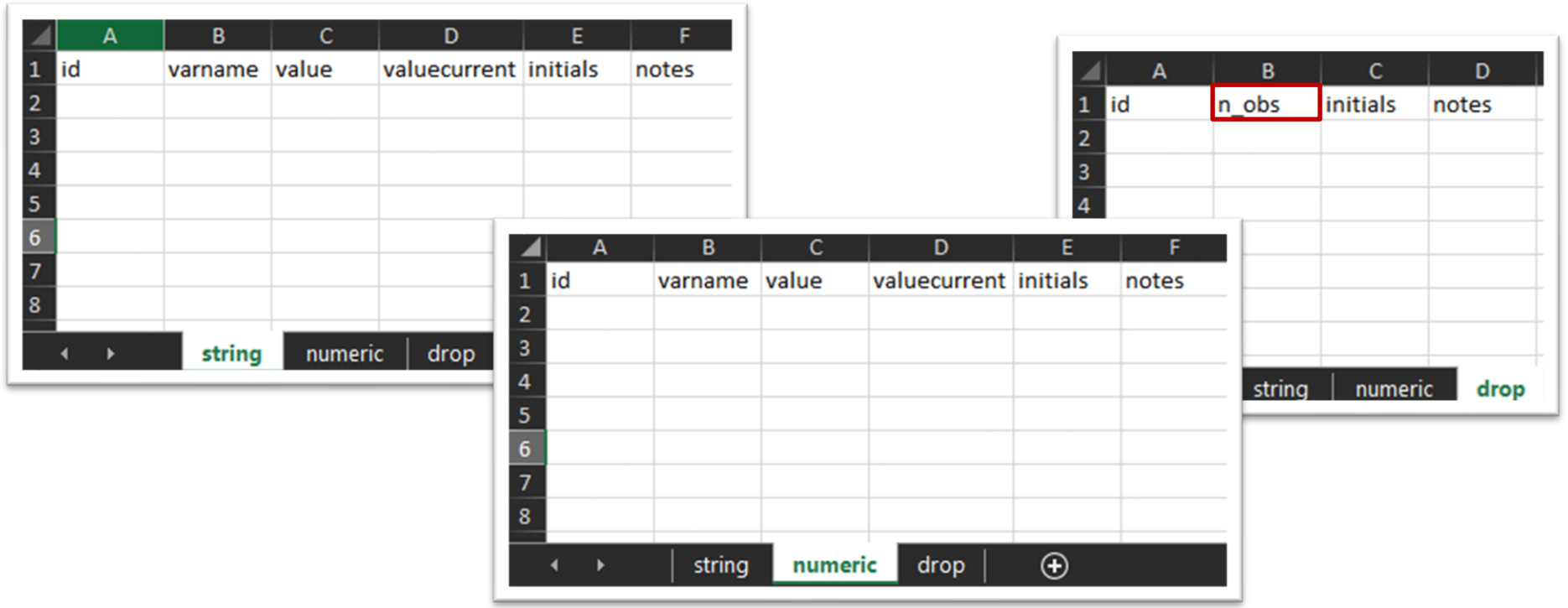


```
iecorrect template using ///
    "/path/to/corrections/file.xlsx", ///
    idvar(varlist)
```



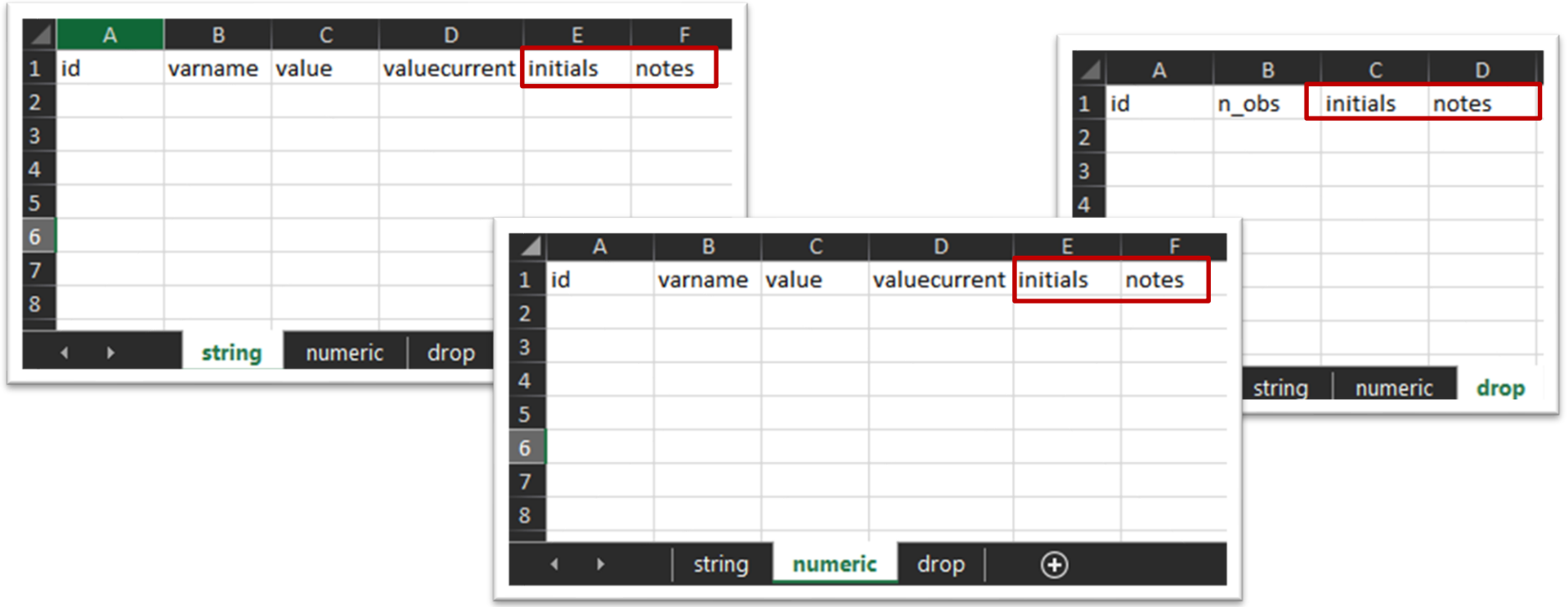

Changes to be made

```
iecorrect template using ///
    "/path/to/corrections/file.xlsx", ///
    idvar(varlist)
```



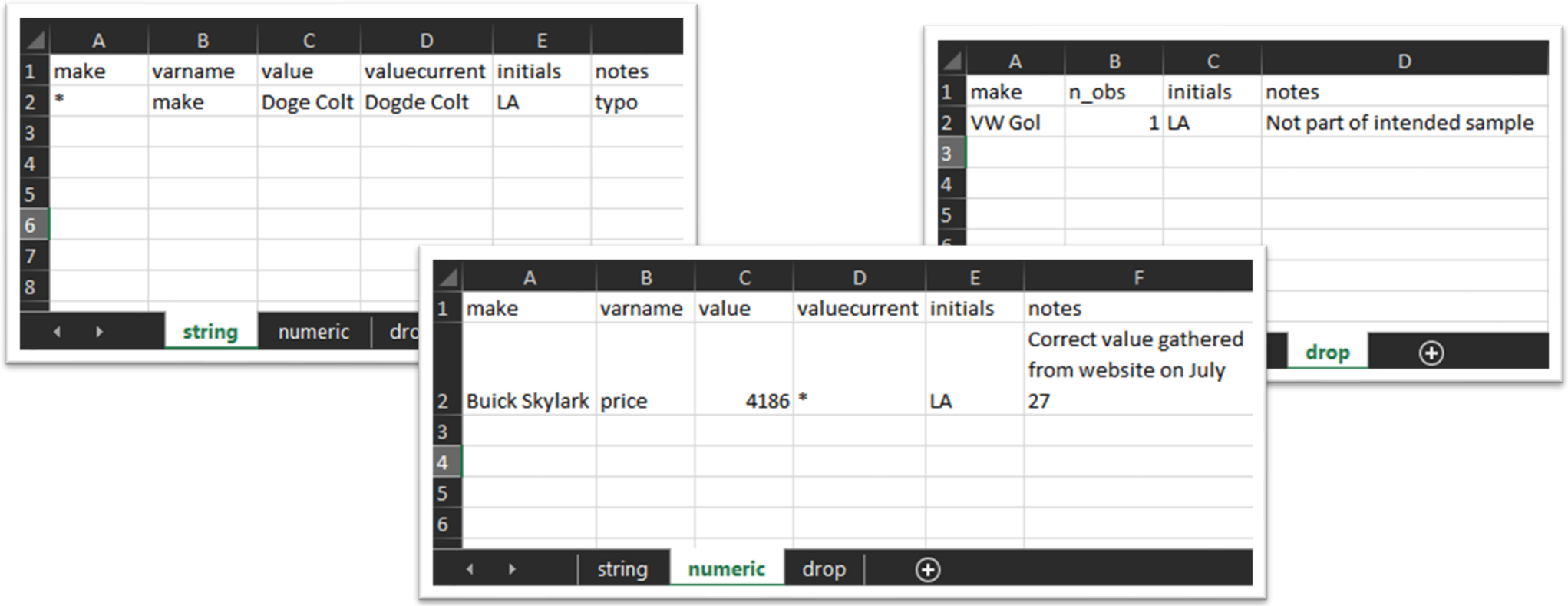

Prevention of unintentional changes

```
iecorrect template using ///
    "/path/to/corrections/file.xlsx", ///
    idvar(varlist)
```




Annotations on who made the changes and why

```
iecorrect apply using ///
    "Documentation/corrections/auto.xlsx", ///
    idvar(make)
```



| | A | B | C | D | E | |
|---|---|---|---|---|---|---|
| 1 | make | varname | value | valuecurrent | initials | notes |
| 2 | * | make | Doge Colt | Dogde Colt | LA | typo |
| 3 | | | | | | |
| 4 | | | | | | |
| 5 | | | | | | |
| 6 | | | | | | |
| 7 | | | | | | |
| 8 | | | | | | |

string | numeric | dro

| | A | B | C | D |
|---|---|---|---|---|
| 1 | make | n_obs | initials | notes |
| 2 | VW Gol | 1 | LA | Not part of intended sample |
| 3 | | | | |
| 4 | | | | |
| 5 | | | | |

drop

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | make | varname | value | valuecurrent | initials | notes |
| 2 | Buick Skylark | price | 4186 | * | LA | Correct value gathered from website on July 27 |
| 3 | | | | | | |
| 4 | | | | | | |
| 5 | | | | | | |
| 6 | | | | | | |

string | numeric | drop

# Why was `iecorrect` created?

- Implement multiple data corrections in one line of code, avoiding repetitive code
- Changelog is more accessible to non-coders
- Creates incentives to write documentation at the same time as corrections are being made
- Checks are built into the command to ensure changes are only made when relevant

# iedupliates:

Identify duplicates in ID variable and export them in an Excel file that also can be used to correct the duplicates.

```
ieduplicates idvarname ///
        using "/path/to/duplicates/report.xlsx", ///
        uniquevars(varlist)
```

| A | B | C | D | E | F | G | H | I | J | |
|---|---|---|---|---|---|---|---|---|---|---|
| hhid | duplistid | datelisted | datefixed | correct | drop | newid | initials | notes | key | listofdiffs |
| 2658 | 1 | 14Jun2019 | | | | | | | uuid:04d07103-0d93-4df9-a009-563f3e8c6a9f | submissiondate starttime endtime enumeratc |
| 2658 | 2 | 14Jun2019 | | | | | | | uuid:1b71f9fb-c1fb-484e-b56d-15d28e6cf580 | submissiondate starttime endtime enumeratc |
| 5000 | 3 | 14Jun2019 | | | | | | | uuid:13990178-9437-482a-acc7-be4b89ecc684 | submissiondate starttime endtime deviceid su |
| 5000 | 4 | 14Jun2019 | | | | | | | uuid:03d46bda-2f57-405f-accf-362287d1a362 | submissiondate starttime endtime deviceid su |
| 6498 | 5 | 14Jun2019 | | | | | | | uuid:1ac93e91-005c-4eef-accf-f0729f864eea | submissiondate key |
| 6498 | 6 | 14Jun2019 | | | | | | | uuid:as289ki0-772b-3247-accf-al38lnaap714 | submissiondate key |
| 9856 | 7 | 14Jun2019 | | | | | | | uuid:2435b795-693d-43b7-9596-ee517719fc61 | submissiondate starttime endtime grandma ic |
| 9856 | 8 | 14Jun2019 | | | | | | | uuid:7530d987-f688-403f-9948-a3c0dcfebcaa | submissiondate starttime endtime grandma ic |

```
ieduplicates hhid ///
        using "project/documentation/duplicates.xlsx", ///
        uniquevars(key)
```

| | A | B | C | D | E | F | G | H | I | J | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | uuid | duplistid | datelisted | datefixed | correct | drop | newid | initials | notes | key | listofdiffs |
| 2 | 2658 | 1 | 14Jun2019 | 14Jun2019 | yes | | | MK | Household | uuid:04d07103-0d93-4df9-a009-563f3e8c6a9f | submissiondate starttime endtime enumerat |
| 3 | 2658 | 2 | 14Jun2019 | 14Jun2019 | | yes | | MK | First inter | uuid:1b71f9fb-c1fb-484e-b56d-15d28e6cf580 | submissiondate starttime endtime enumerat |
| 4 | 5000 | 3 | 14Jun2019 | 14Jun2019 | yes | | | MK | Survey frc | uuid:13990178-9437-482a-acc7-be4b89ecc684 | submissiondate starttime endtime deviceid s |
| 5 | 5000 | 4 | 14Jun2019 | 14Jun2019 | | | 5001 | MK | Wrong ID, | uuid:03d46bda-2f57-405f-accf-362287d1a362 | submissiondate starttime endtime deviceid s |
| 6 | 6498 | 5 | 14Jun2019 | 14Jun2019 | | yes | | LT | Submitted | uuid:1ac93e91-005c-4eef-accf-f0729f864eea | submissiondate key |
| 7 | 6498 | 6 | 14Jun2019 | 14Jun2019 | yes | | | LT | Submitted | uuid:as289ki0-772b-3247-accf-al38lnaap714 | submissiondate key |
| 8 | 9856 | 7 | 14Jun2019 | | | | | | | uuid:2435b795-693d-43b7-9596-ee517719fc61 | submissiondate starttime endtime grandma i |
| 9 | 9856 | 8 | 14Jun2019 | | | | | | | uuid:7530d987-f688-403f-9948-a3c0dcfebcaa | submissiondate starttime endtime grandma i |

# Why was `ieduplicates` created?

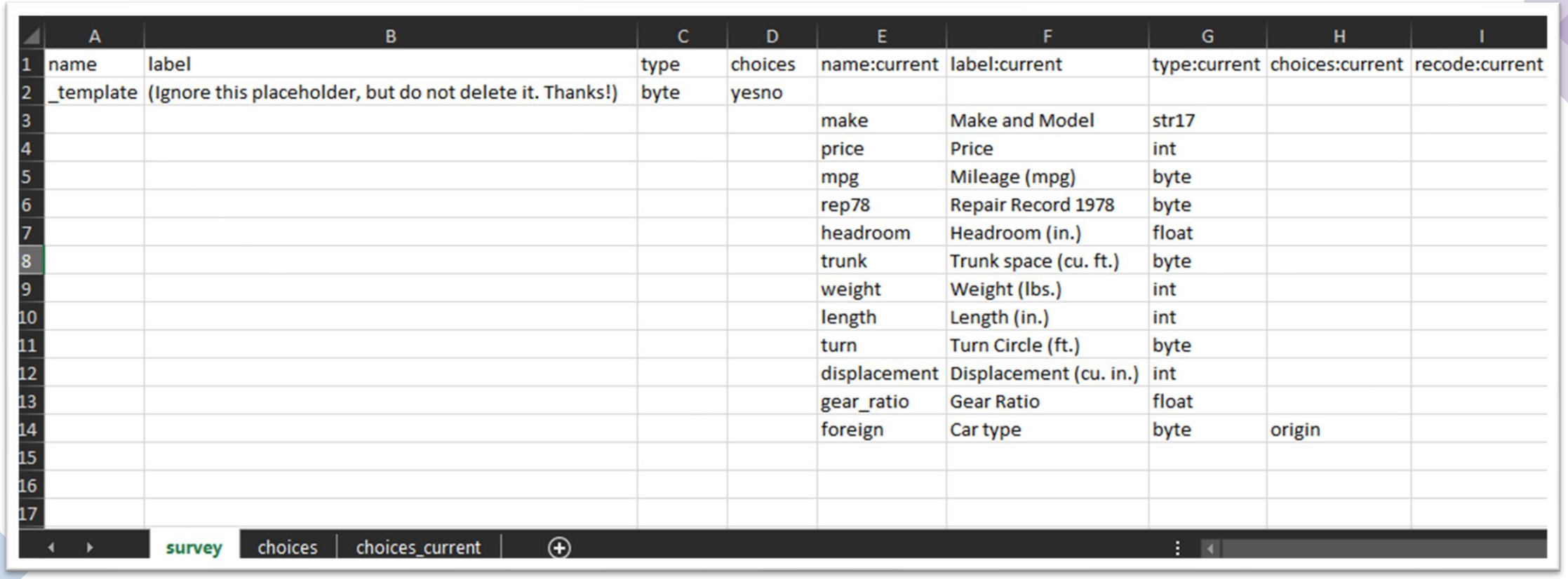
- Help identify and solve duplicate entries by exporting an easy to read report
- Implement multiple corrections in one line of code
- Creates a record of how duplicates were solved

# `iecodebook:`

Automate repetitive data cleaning tasks

1. **`apply:`** bulk rename, recode, and label variables
2. **`append:`** harmonize two or more datasets to have the same variable names, labels, and value; then append them
3. **`export:`** create a document listing all variable names, variable labels and value labels

```
iecodebook template using "/path/to/codebook.xlsx", ///
    replace
```



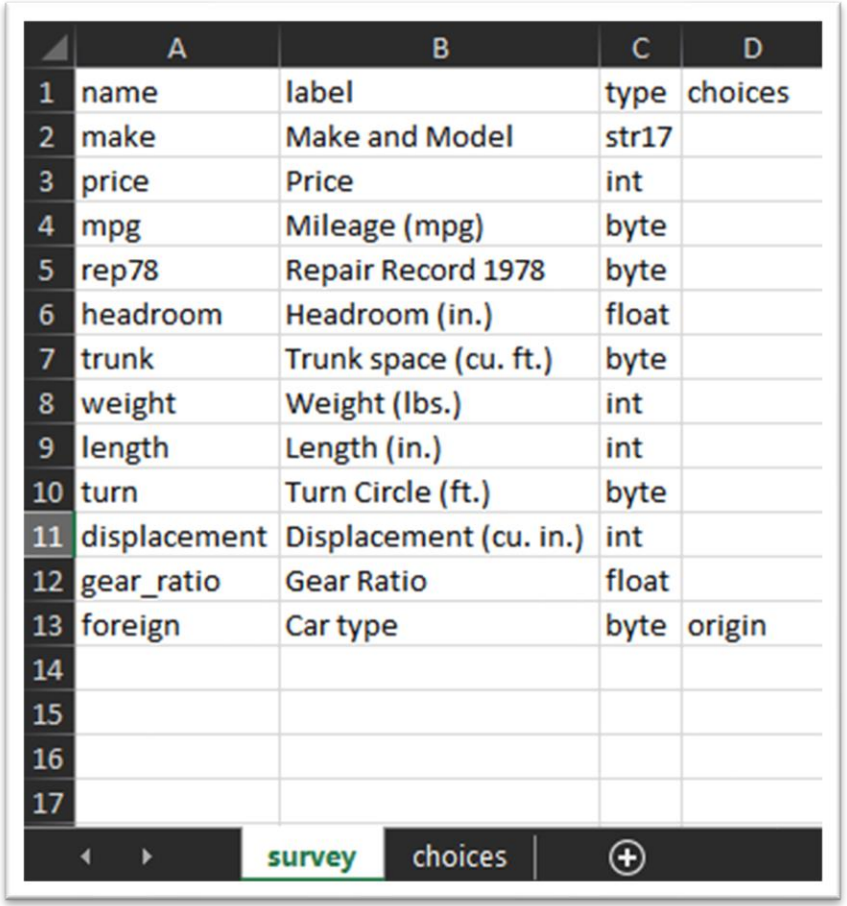
| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | name | label | type | choices | name:current | label:current | type:current | choices:current | recode:current |
| 2 | _template | (Ignore this placeholder, but do not delete it. Thanks!) | byte | yesno | | | | | |
| 3 | | | | | make | Make and Model | str17 | | |
| 4 | | | | | price | Price | int | | |
| 5 | | | | | mpg | Mileage (mpg) | byte | | |
| 6 | | | | | rep78 | Repair Record 1978 | byte | | |
| 7 | | | | | headroom | Headroom (in.) | float | | |
| 8 | | | | | trunk | Trunk space (cu. ft.) | byte | | |
| 9 | | | | | weight | Weight (lbs.) | int | | |
| 10 | | | | | length | Length (in.) | int | | |
| 11 | | | | | turn | Turn Circle (ft.) | byte | | |
| 12 | | | | | displacement | Displacement (cu. in.) | int | | |
| 13 | | | | | gear_ratio | Gear Ratio | float | | |
| 14 | | | | | foreign | Car type | byte | origin | |
| 15 | | | | | | | | | |
| 16 | | | | | | | | | |
| 17 | | | | | | | | | |

survey | choices | choices_current

# iecodebook apply using “project/documentation/auto.xlsx”

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | name | label | type | choices | name:current | label:current | type:current | choices:current | |
| 2 | survey | Data Source (do not edit this row) | byte | yesno | | | | | |
| 3 | make | Make and Model | str17 | | make | Make and Model | str17 | | |
| 4 | price | Price | int | | price | Price | int | | |
| 5 | mpg | Mileage (mpg) | byte | | mpg | Mileage (mpg) | byte | | |
| 6 | rep78 | Repair Record 1978 | byte | | rep78 | Repair Record 1978 | byte | | |
| 7 | headroom | Headroom (in.) | float | | headroom | Headroom (in.) | float | | |
| 8 | trunk | Trunk space (cu. ft.) | byte | | trunk | Trunk space (cu. ft.) | byte | | |
| 9 | weight | Weight (lbs.) | int | | weight | Weight (lbs.) | int | | |
| 10 | length | Length (in.) | int | | length | Length (in.) | int | | |
| 11 | turn | Turn Circle (ft.) | byte | | turn | Turn Circle (ft.) | byte | | |
| 12 | displacement | Displacement (cu. in.) | int | | displacement | Displacement (cu. in.) | int | | |
| 13 | gear_ratio | Gear Ratio | float | | gear_ratio | Gear Ratio | float | | |
| 14 | foreign | Car type | byte | origin | foreign | Car type | byte | origin | |

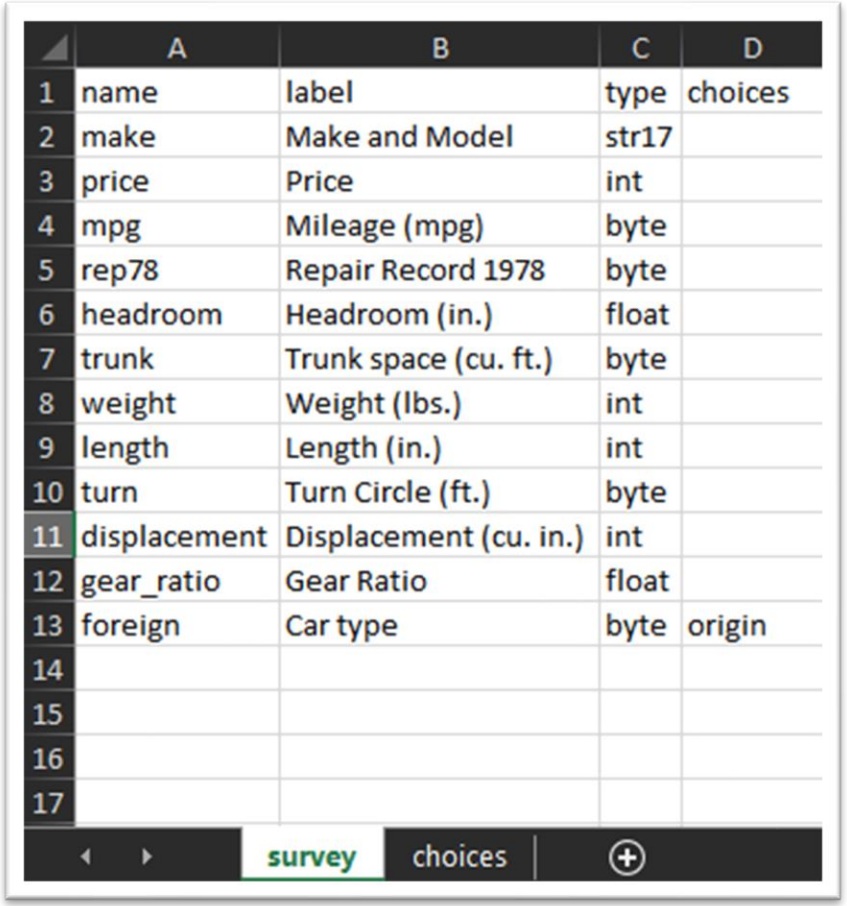
# iecodebook export using “project/data/auto.xlsx”



| | A | B | C | D |
|---|---|---|---|---|
| 1 | name | label | type | choices |
| 2 | make | Make and Model | str17 | |
| 3 | price | Price | int | |
| 4 | mpg | Mileage (mpg) | byte | |
| 5 | rep78 | Repair Record 1978 | byte | |
| 6 | headroom | Headroom (in.) | float | |
| 7 | trunk | Trunk space (cu. ft.) | byte | |
| 8 | weight | Weight (lbs.) | int | |
| 9 | length | Length (in.) | int | |
| 10 | turn | Turn Circle (ft.) | byte | |
| 11 | displacement | Displacement (cu. in.) | int | |
| 12 | gear_ratio | Gear Ratio | float | |
| 13 | foreign | Car type | byte | origin |
| 14 | | | | |
| 15 | | | | |
| 16 | | | | |
| 17 | | | | |

survey choices

```
iecodebook export using “project/data/auto.xlsx”,
      [replace] [save] [verify] [signature][reset]
      [trim("/path/to/dofile1.do" ["/path/to/dofile2.do"] [...])]
```



| | A | B | C | D |
|---|---|---|---|---|
| 1 | name | label | type | choices |
| 2 | make | Make and Model | str17 | |
| 3 | price | Price | int | |
| 4 | mpg | Mileage (mpg) | byte | |
| 5 | rep78 | Repair Record 1978 | byte | |
| 6 | headroom | Headroom (in.) | float | |
| 7 | trunk | Trunk space (cu. ft.) | byte | |
| 8 | weight | Weight (lbs.) | int | |
| 9 | length | Length (in.) | int | |
| 10 | turn | Turn Circle (ft.) | byte | |
| 11 | displacement | Displacement (cu. in.) | int | |
| 12 | gear_ratio | Gear Ratio | float | |
| 13 | foreign | Car type | byte | origin |
| 14 | | | | |
| 15 | | | | |
| 16 | | | | |
| 17 | | | | |

survey | choices

# Why was `iecodebook` created?

- Automate repetitive tasks, creating a standardized input and avoiding repetitive code
- Excel sheet is more accessible to non-coders
- Record of changes is easier to read than the code that implements the changes
- Automates the creation of variable dictionary templates

# iesave:

Automate common checks to the data before saving a dataset. Optionally, save a plain text metadata report describing the dataset saved.

# Basic usage

```
iesave using "${project}/data/auto.dta",   ///
    idvars(make) version(13) replace
```

- Check that data is fully and uniquely identified by ID variables
- Optimize storage on disk using `compress`
- Save the data set in the desired .dta version
- Save metadata to data set characteristics

# Variable reports

```
iesave using "${project}/data/auto.dta",   ///
    idvars(make) version(13) replace       ///
    report("${project}/data/auto.md", replace)
```

- Check that data is fully and uniquely identified by ID variables
- Optimize storage on disk using `compress`
- Save the data set in the desired .dta version
- Save metadata to data set characteristics
- Export a plain text report describing the data set saved and its variables

# Variable reports

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Number of observations: | 74 | | | | | | | | | |
| 2 | Number of variables: | 16 | | | | | | | | | |
| 3 | ID variable(s): | make | | | | | | | | | |
| 4 | Data signature: | 74:16(110350):932916212:1889387683 | | | | | | | | | |
| 5 | Last saved by: | wb501238 | | | | | | | | | |
| 6 | Last saved at: | 8/2/2022 11:39 | | | | | | | | | |
| 7 | | | | | | | | | | | |
| 8 | Variable type: String | | | | | | | | | | |
| 9 | Name | Label | Type | Complete | Number of levels | | | | | | |
| 10 | make | Make and Model | str17 | 74 | 74 | | | | | | |
| 11 | | | | | | | | | | | |
| 12 | Variable type: Continuous | | | | | | | | | | |
| 13 | Name | Label | Type | Complete | Mean | Std Dev | p0 | p25 | p50 | p75 | p100 |
| 14 | displacement | Displacement (cu. in.) | int | 74 | 197.3 | 91.84 | 79 | 119 | 196 | 250 | 425 |
| 15 | gear_ratio | Gear Ratio | float | 74 | 3.015 | 0.4563 | 2.19 | 2.73 | 2.955 | 3.37 | 3.89 |
| 16 | headroom | Headroom (in.) | float | 74 | 2.993 | 0.846 | 1.5 | 2.5 | 3 | 3.5 | 5 |
| 17 | length | Length (in.) | int | 74 | 187.9 | 22.27 | 142 | 170 | 192.5 | 204 | 233 |
| 18 | mpg | Mileage (mpg) | byte | 74 | 21.3 | 5.786 | 12 | 18 | 20 | 25 | 41 |
| 19 | price | Price | int | 74 | 6165 | 2949 | 3291 | 4195 | 5007 | 6342 | 15906 |
| 20 | rep78 | Repair Record 1978 | byte | 69 | 3.406 | 0.9899 | 1 | 3 | 3 | 4 | 5 |
| 21 | trunk | Trunk space (cu. ft.) | byte | 74 | 13.76 | 4.277 | 5 | 10 | 14 | 17 | 23 |
| 22 | turn | Turn Circle (ft.) | byte | 74 | 39.65 | 4.399 | 31 | 36 | 40 | 43 | 51 |
| 23 | weight | Weight (lbs.) | int | 74 | 3019 | 777.2 | 1760 | 2240 | 3190 | 3600 | 4840 |

CSV

- **Number of observations:** 74
- **Number of variables:** 16
- **ID variable(s):** make
- **Data signature:** 74:16(110350):932916212:1889387683
- **Last saved by:** wb501238
- **Last saved at:** 11:39:57 2 Aug 2022

### Variable type: String

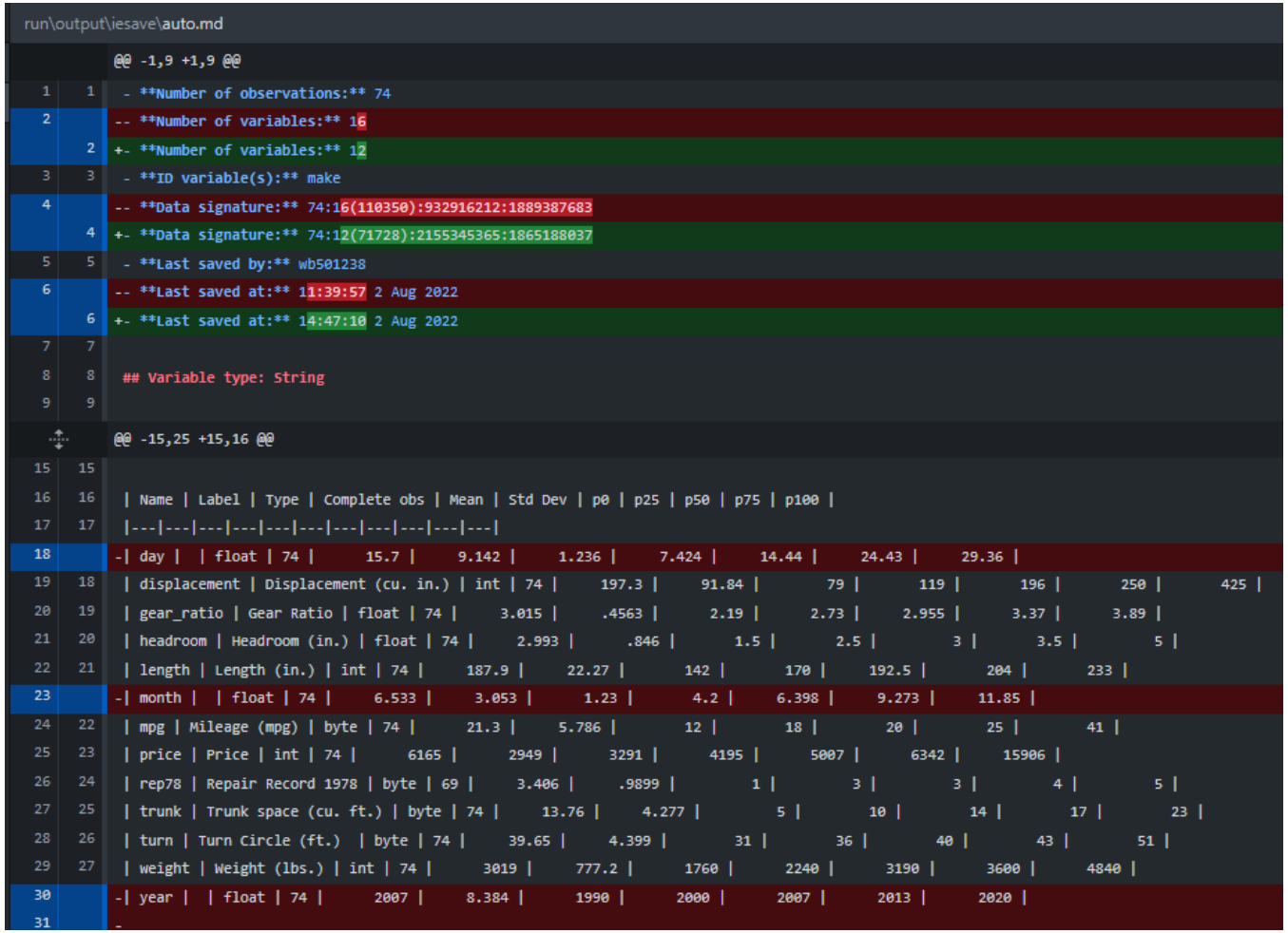
| Name | Label | Type | Complete obs | Number of levels |
|---|---|---|---|---|
| make | Make and Model | str17 | 74 | 74 |

### Variable type: Continuous

| Name | Label | Type | Complete obs | Mean | Std Dev | p0 | p25 | p50 | p75 | p100 |
|---|---|---|---|---|---|---|---|---|---|---|
| displacement | Displacement (cu. in.) | int | 74 | 197.3 | 91.84 | 79 | 119 | 196 | 250 | 425 |
| gear_ratio | Gear Ratio | float | 74 | 3.015 | .4563 | 2.19 | 2.73 | 2.955 | 3.37 | 3.89 |
| headroom | Headroom (in.) | float | 74 | 2.993 | .846 | 1.5 | 2.5 | 3 | 3.5 | 5 |
| length | Length (in.) | int | 74 | 187.9 | 22.27 | 142 | 170 | 192.5 | 204 | 233 |
| mpg | Mileage (mpg) | byte | 74 | 21.3 | 5.786 | 12 | 18 | 20 | 25 | 41 |
| price | Price | int | 74 | 6165 | 2949 | 3291 | 4195 | 5007 | 6342 | 15906 |
| rep78 | Repair Record 1978 | byte | 69 | 3.406 | .9899 | 1 | 3 | 3 | 4 | 5 |
| trunk | Trunk space (cu. ft.) | byte | 74 | 13.76 | 4.277 | 5 | 10 | 14 | 17 | 23 |
| turn | Turn Circle (ft.) | byte | 74 | 39.65 | 4.399 | 31 | 36 | 40 | 43 | 51 |
| weight | Weight (lbs.) | int | 74 | 3019 | 777.2 | 1760 | 2240 | 3190 | 3600 | 4840 |

Markdown

# Variable reports



run\output\iesave\auto.md

```
@@ -1,9 +1,9 @@
 - **Number of observations:** 74
-- **Number of variables:** 16
+- **Number of variables:** 12
 - **ID variable(s):** make
-- **Data signature:** 74:16(110350):932916212:1889387683
+- **Data signature:** 74:12(71728):2155345365:1865188037
 - **Last saved by:** wb501238
-- **Last saved at:** 11:39:57 2 Aug 2022
+- **Last saved at:** 14:47:10 2 Aug 2022

 ## Variable type: String

@@ -15,25 +15,16 @@

 | Name | Label | Type | Complete obs | Mean | Std Dev | p0 | p25 | p50 | p75 | p100 |
 |---|---|---|---|---|---|---|---|---|---|---|
-| day |  | float | 74 |  15.7 |  9.142 |  1.236 |  7.424 |  14.44 |  24.43 |  29.36 |
 | displacement | Displacement (cu. in.) | int | 74 |  197.3 |  91.84 |  79 |  119 |  196 |  250 |  425 |
 | gear_ratio | Gear Ratio | float | 74 |  3.015 |  .4563 |  2.19 |  2.73 |  2.955 |  3.37 |  3.89 |
 | headroom | Headroom (in.) | float | 74 |  2.993 |  .846 |  1.5 |  2.5 |  3 |  3.5 |  5 |
 | length | Length (in.) | int | 74 |  187.9 |  22.27 |  142 |  170 |  192.5 |  204 |  233 |
-| month |  | float | 74 |  6.533 |  3.053 |  1.23 |  4.2 |  6.398 |  9.273 |  11.85 |
 | mpg | Mileage (mpg) | byte | 74 |  21.3 |  5.786 |  12 |  18 |  20 |  25 |  41 |
 | price | Price | int | 74 |  6165 |  2949 |  3291 |  4195 |  5007 |  6342 |  15906 |
 | rep78 | Repair Record 1978 | byte | 69 |  3.406 |  .9899 |  1 |  3 |  3 |  4 |  5 |
 | trunk | Trunk space (cu. ft.) | byte | 74 |  13.76 |  4.277 |  5 |  10 |  14 |  17 |  23 |
 | turn | Turn Circle (ft.) | byte | 74 |  39.65 |  4.399 |  31 |  36 |  40 |  43 |  51 |
 | weight | Weight (lbs.) | int | 74 |  3019 |  777.2 |  1760 |  2240 |  3190 |  3600 |  4840 |
-| year |  | float | 74 |  2007 |  8.384 |  1990 |  2000 |  2007 |  2013 |  2020 |
-
```

# Other data quality checks

Additional data quality checks to be performed before the data is saved

- All variables should be labeled
- There should be no standard missing values (“.”)
- There should be no unlabeled levels in categorical variables

# Why was `iesave` created?

- Implement an opinionated best practice add-on to `save`
- Automate best practices such as exporting meta data, compressing the data, and testing ID variables
- Allows users to track changes to the data using git
- Variable report is easier to read than simply outputting codebook as .txt

# In short

# In short

- Data documentation is a key component of research transparency
  - However, it can be a time-consuming task to comment scripts and create reports on the data
  - This means documentation is often neglected
- The commands presented here make it easier to create data documentation by
  - Creating standardized and easy to read documents outlining data changes
  - Simplifying code through automation of repetitive tasks
  - Combining the implementation of data changes and its documentation
  - Automating the creation of data documentation such as variable dictionaries and codebooks
  - Allow team members who do not code directly contribute to the process

Thank you!

# Useful links

## iefieldkit

- GitHub repository
- Full documentation
- Stata Journal article
- 2019 Stata conference presentation

## ietoolkit

- GitHub repository
- Full documentation
- 2019 Stata conference presentation