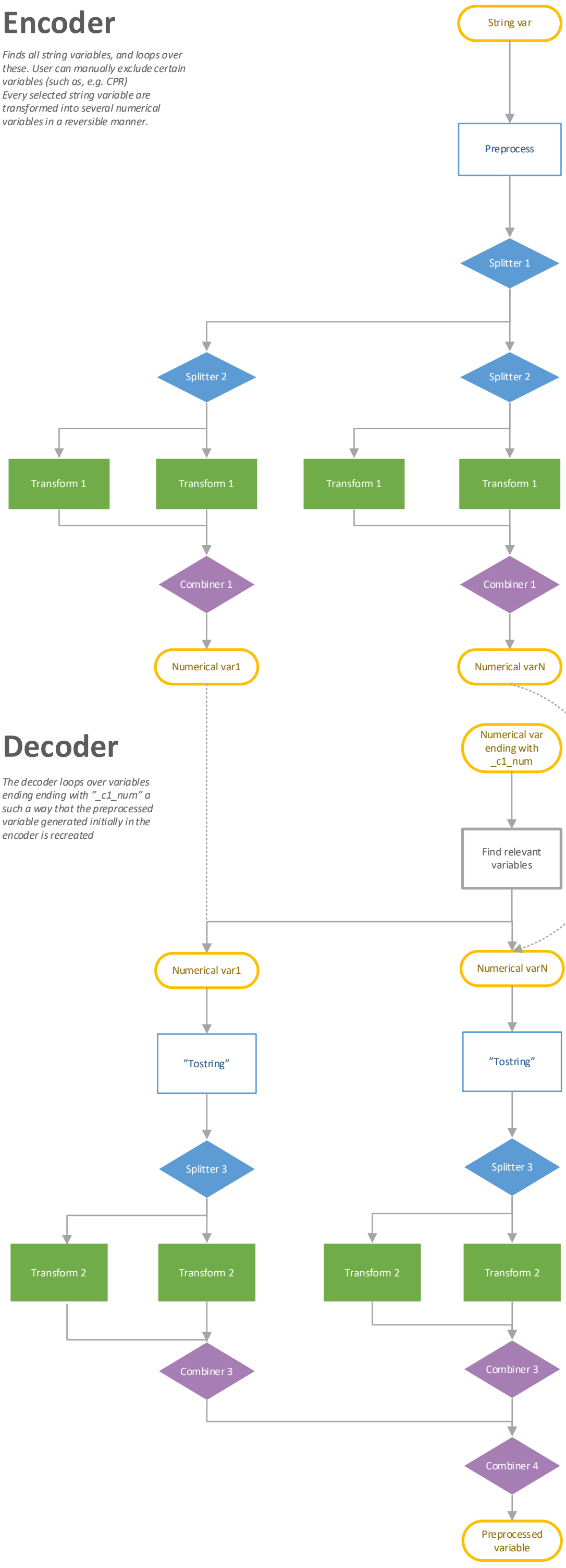
Encoder

Finds all string variables, and loops over these. User can manually exclude certain variables (such as, e.g. CPR)
Every selected string variable are transformed into several numerical variables in a reversible manner.



- Should not have a name that is "too long"
- Makes lower case
- Danish letters to e.g. "ae"
- Finds maximum length of strings
- Splits string variable into several string vars containing substrings of parent
- Recursively takes first 7 symbols of parent string
- Splits parent into several string vars containing at most one symbol
- Turns the a single symbol (or the empty string) into a number between 0 and 99
- Combines the individual numbers into one number
- Does so by summing them as terms a_i * 10^(2(i-1))
- These variables are doubles strictly less than 10^14.
- Variable names are of the form "original name" _c*_num

Decoder

The decoder loops over variables ending ending with "_c1_num" a such a way that the preprocessed variable generated initially in the encoder is recreated

- Does this by reducing the variable name to "original name"_c and looping over variables starting with this
- These are the variables found. They correspond to the endpoints of the encoder
- Starts with 'tostring'-ing the variables
- Replaces "." with ""
- If the length of the string is uneven, a "0" is added to the beginning
- Note that these strings are at most 14 symbols long
- Splits into several string variables containing two or zero symbols
- Output corresponds (roughly) to the output of *Transform 1*
- Uses (roughly) the same encoding scheme as *Transform 1*, but in the opposite direction
- Only "roughly" because the input "numbers" are actually strings, and if they are less than 10, they are on the form "0x".
- The outputs are string variables containing a single symbol
- Combines the (at most 7) string variables into one string variable.
- Combines the substrings into one final string variable which is equal to the preprocessed variable
- *Exception*: sometimes there are weirdness in the character encoding in Stata and there are cases where characters such as ´ can not be treated properly. In such cases the particular cell is generally useless. Only the particular cell is effected by this.