# Transform the Data

Venkata Reddy Konasani

# Contents

- Power Query
- Handling null values
- Rename
- M-Language
- Creating new columns
- Creating columns from examples
- Handling errors

# Have a Look at the data

| id | gender | age | ever_married | work_type | Residence_type |
|---|---|---|---|---|---|
| 60182 | Female | 49 | Yes | WT -Private | Urban |
| 60491 | Female | 78 | Yes | WT -Private | Urban |
| 12175 | Female | 54 | Yes | WT -Private | Urban |
| 5317 | Female | 79 | Yes | WT -Private | Urban |
| 62602 | Female | 49 | Yes | WT -Private | Urban |
| 1845 | Female | 63 | Yes | WT -Private | Urban |
| 47472 | Female | 58 | Yes | WT -Private | Urban |
| 17004 | Female | 70 | Yes | WT -Private | Urban |
| 71673 | Female | 79 | Yes | WT -Private | Urban |
| 45805 | Female | 51 | Yes | WT -Private | Urban |
| 28291 | Female | 79 | Yes | WT -Private | Urban |
| 5563 | Female | 77 | Yes | WT -Private | Urban |
| 72918 | Female | 53 | Yes | WT -Private | Urban |
| 14164 | Female | 72 | Yes | WT -Private | Urban |
| 70943 | Female | 80 | Yes | WT -Private | Urban |
| 11762 | Female | 76 | Yes | WT -Private | Urban |
| 8045 | Female | 74 | Yes | WT -Private | Urban |
| 17308 | Female | 72 | Yes | WT -Private | Urban |

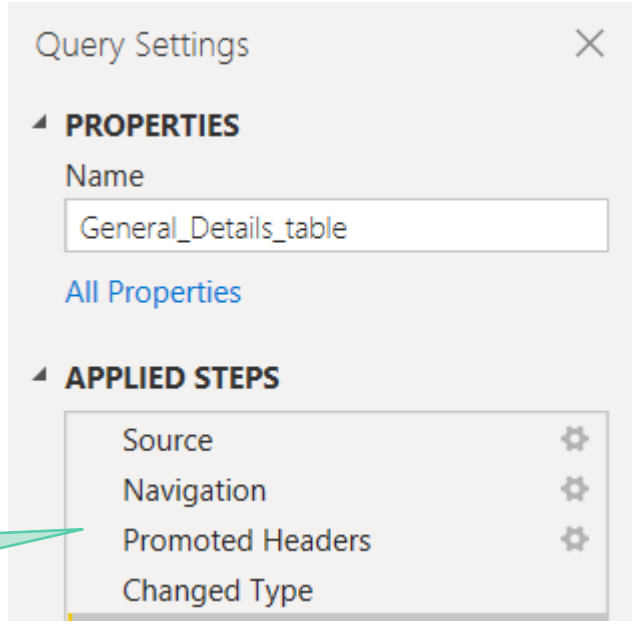
| Patient_id | hypertension | heart_disease | avg_glucose_level | bmi | smoking_status | stroke |
|---|---|---|---|---|---|---|
| 63884 | 0 | 0 | 162.96 | 39.4 | never smoked | 0 |
| 67855 | 0 | 0 | 95.04 | 42.4 | never smoked | 0 |
| 25774 | 0 | 0 | 85.37 | 33 | never smoked | 0 |
| 24447 | 0 | 0 | 82.67 | 22.5 | never smoked | 0 |
| 48588 | 0 | 0 | 109.82 | 23.7 | never smoked | 0 |
| 70336 | 0 | 0 | 60.84 | 24.5 | never smoked | 0 |
| 45801 | 0 | 0 | 97.49 | 26.9 | never smoked | 0 |
| 36275 | 0 | 0 | 206.72 | 26.7 | never smoked | 0 |
| 11577 | 0 | 0 | 214.45 | 31.2 | never smoked | 0 |
| 36811 | 0 | 0 | 94.09 | 30.9 | never smoked | 0 |
| 58261 | 0 | 0 | 141.24 | 28.5 | never smoked | 0 |
| 28526 | 0 | 0 | 203.04 | 33.6 | never smoked | 0 |
| 7282 | 0 | 0 | 81.84 | 25.1 | never smoked | 0 |
| 1686 | 0 | 0 | 71.89 | 27.6 | never smoked | 0 |
| 59368 | 0 | 0 | 243.5 | 26.1 | never smoked | 0 |
| 18051 | 0 | 0 | 91.61 | 25.2 | never smoked | 0 |
| 40840 | 0 | 0 | 138.16 | 19.4 | never smoked | 0 |
| 10449 | 0 | 0 | 75.23 | 29 | never smoked | 0 |
| 61837 | 0 | 0 | 58.95 | 24.6 | never smoked | 0 |
| 9487 | 0 | 0 | 99.92 | 25.8 | never smoked | 0 |
| 49713 | 0 | 0 | 116.23 | 26.1 | never smoked | 0 |
| 28102 | 0 | 0 | 66.3 | 27.2 | never smoked | 0 |
| 62608 | 0 | 0 | 136.8 | 37.3 | never smoked | 0 |
| 40670 | 0 | 0 | 96.57 | 34.1 | never smoked | 0 |
| 4630 | 0 | 0 | 66.42 | 23.6 | never smoked | 0 |

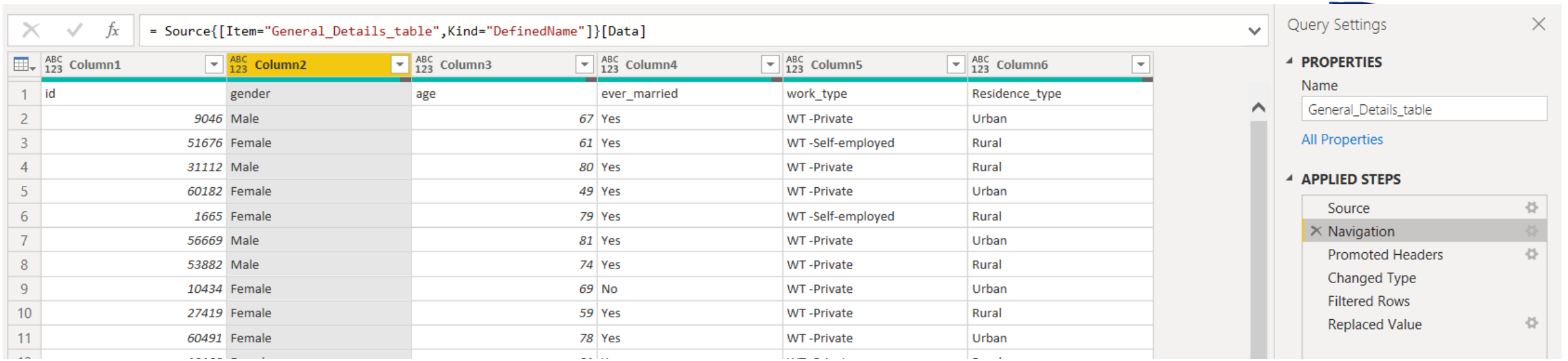
# Stroke Case study

# Step-3: Transform the data; Prepare it for analysis

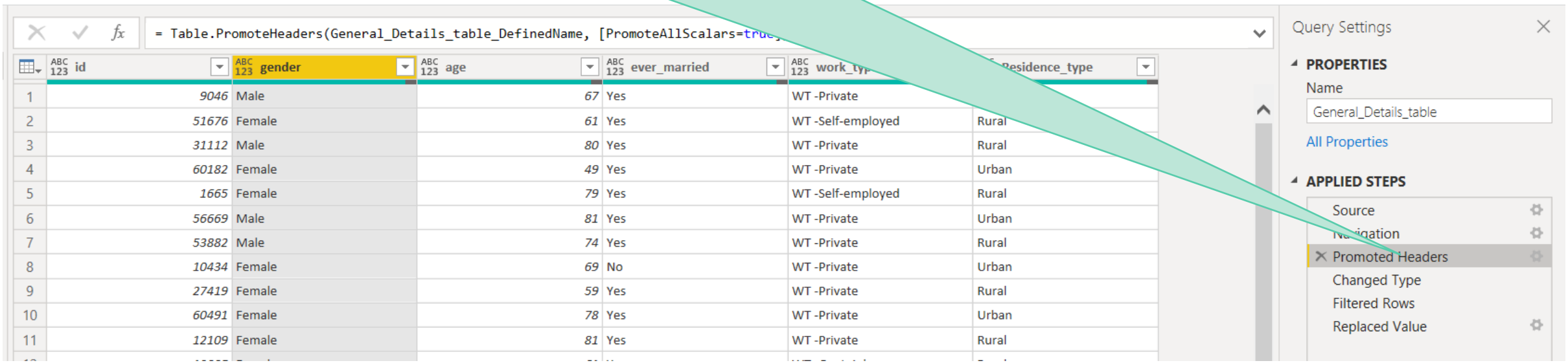
# Power Query Editor

- Start with General Details table data transformation
- Click on Transform data, it will open the power query editor
- Check the query settings, Some of the transformations are already applied on the data



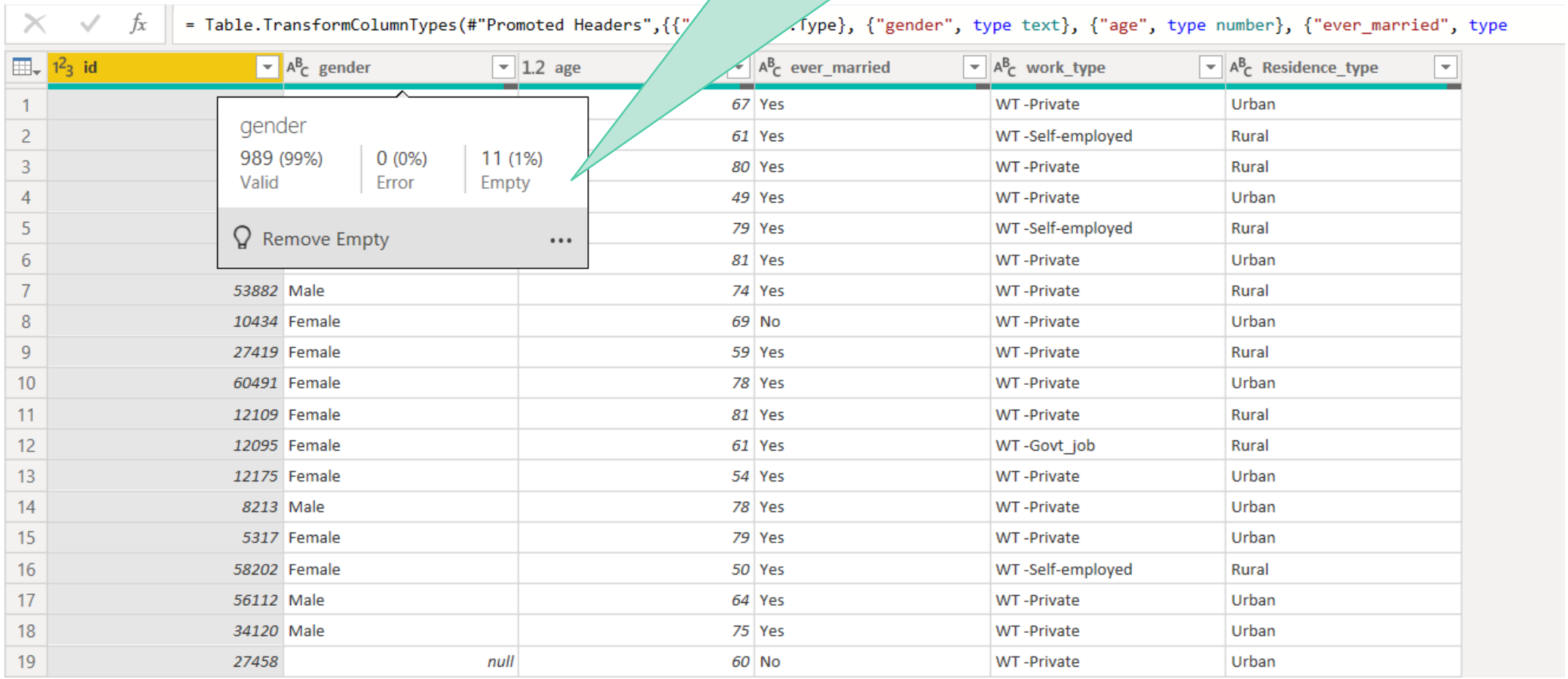



You can check the effect of a particular transformation step



# Gender variable

- 1% null values in gender variable
- Can be replaced with “NA”

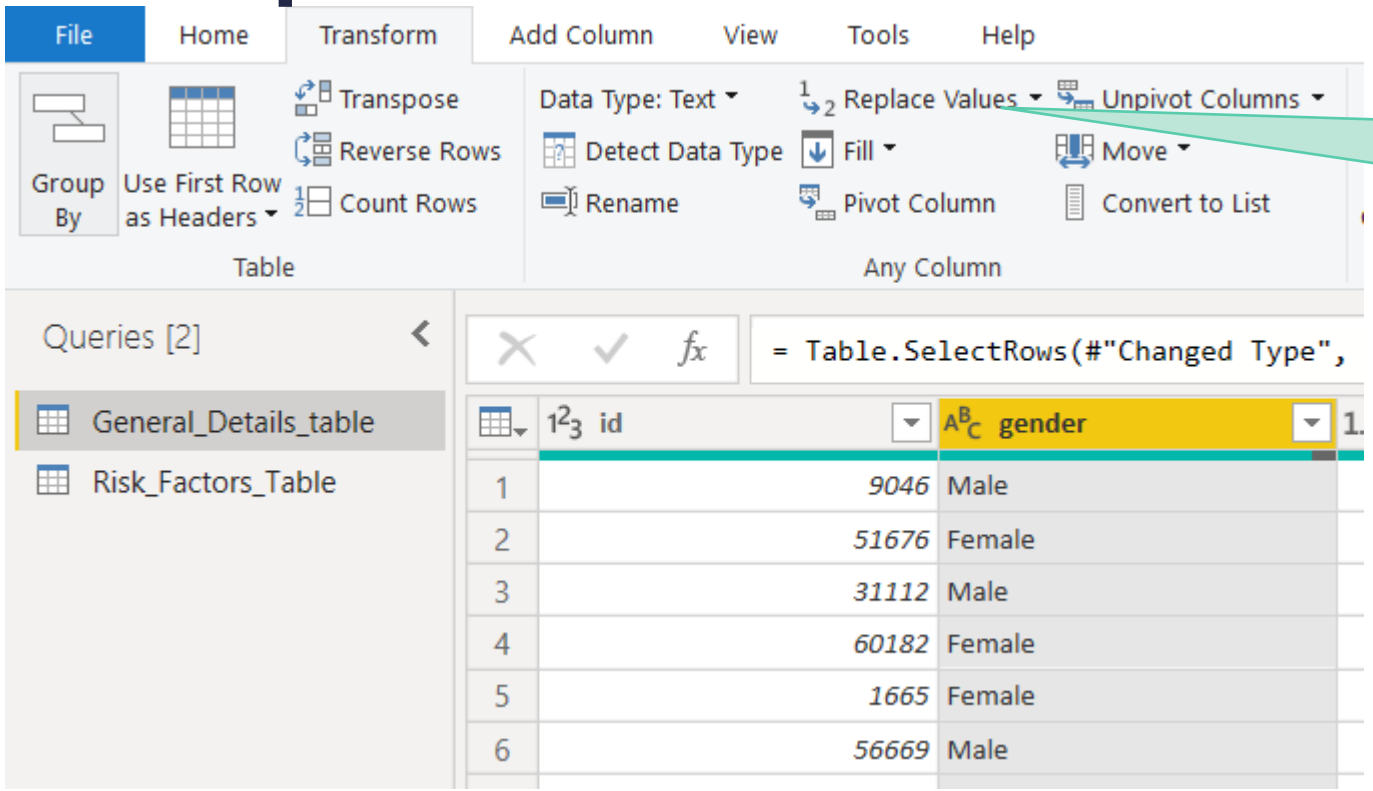


= Table.TransformColumnTypes(#"Promoted Headers",{{" .Type}, {"gender", type text}, {"age", type number}, {"ever_married", type

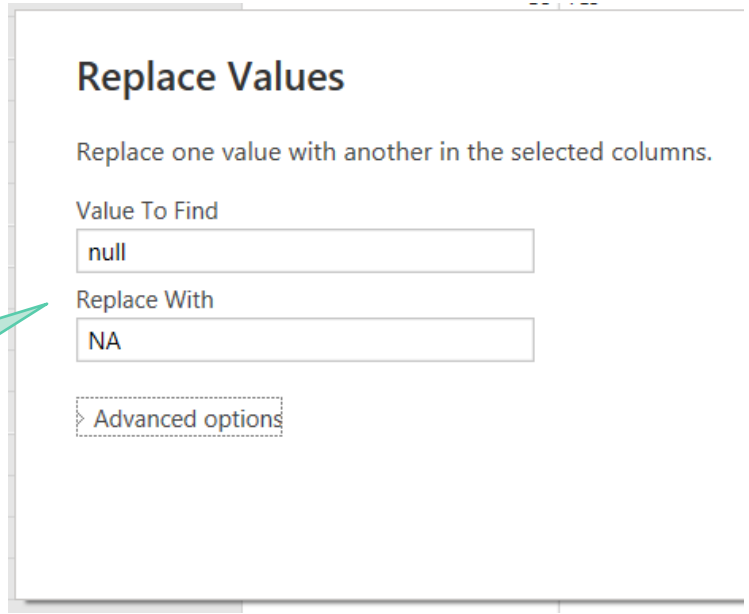
gender
989 (99%) Valid | 0 (0%) Error | 11 (1%) Empty
Remove Empty

| | id | gender | age | ever_married | work_type | Residence_type |
|---|---|---|---|---|---|---|
| 1 | | | 67 | Yes | WT -Private | Urban |
| 2 | | | 61 | Yes | WT -Self-employed | Rural |
| 3 | | | 80 | Yes | WT -Private | Rural |
| 4 | | | 49 | Yes | WT -Private | Urban |
| 5 | | | 79 | Yes | WT -Self-employed | Rural |
| 6 | | | 81 | Yes | WT -Private | Urban |
| 7 | 53882 | Male | 74 | Yes | WT -Private | Rural |
| 8 | 10434 | Female | 69 | No | WT -Private | Urban |
| 9 | 27419 | Female | 59 | Yes | WT -Private | Rural |
| 10 | 60491 | Female | 78 | Yes | WT -Private | Urban |
| 11 | 12109 | Female | 81 | Yes | WT -Private | Rural |
| 12 | 12095 | Female | 61 | Yes | WT -Govt_job | Rural |
| 13 | 12175 | Female | 54 | Yes | WT -Private | Urban |
| 14 | 8213 | Male | 78 | Yes | WT -Private | Urban |
| 15 | 5317 | Female | 79 | Yes | WT -Private | Urban |
| 16 | 58202 | Female | 50 | Yes | WT -Self-employed | Rural |
| 17 | 56112 | Male | 64 | Yes | WT -Private | Urban |
| 18 | 34120 | Male | 75 | Yes | WT -Private | Urban |
| 19 | 27458 | null | 60 | No | WT -Private | Urban |

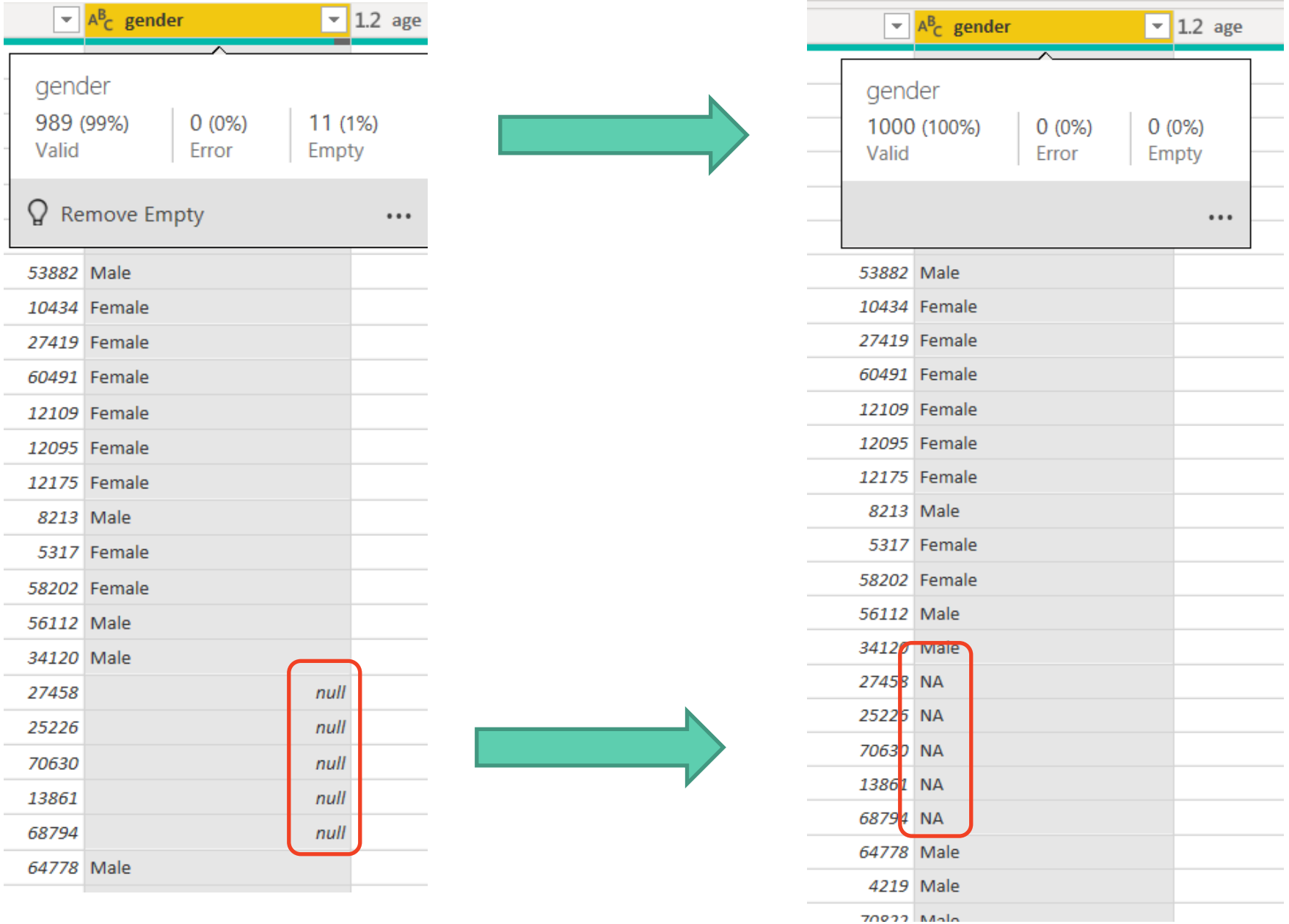
# Replace null values




Menu options >> Transform >> Replace Values

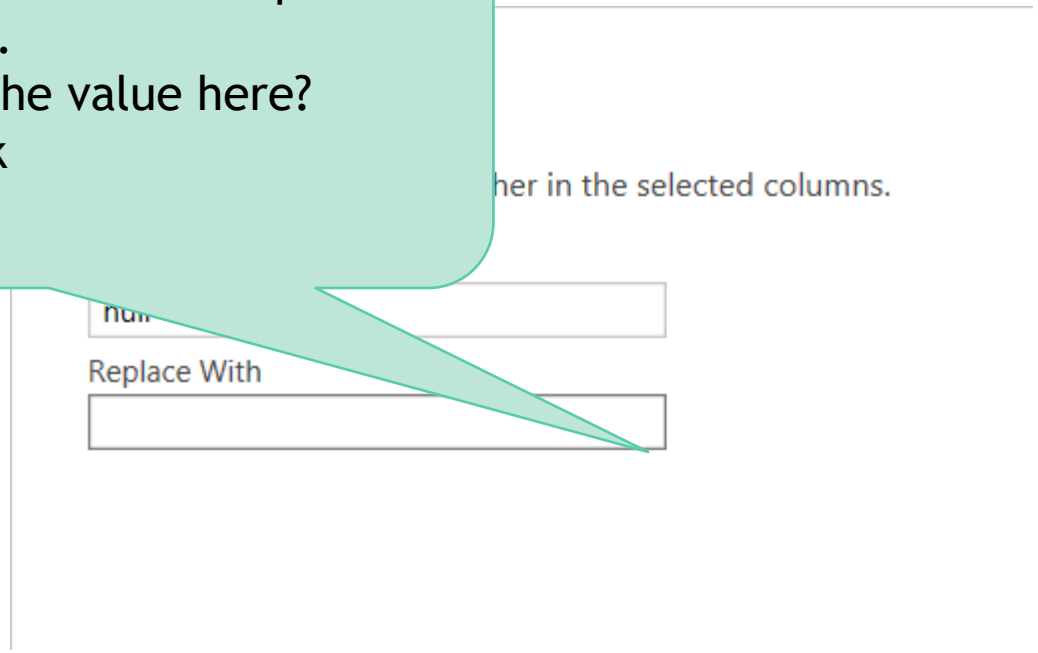



Value to find >> null
Value to replace it with >> NA


gender
age
gender
989 (99%)
Valid
0 (0%)
Error
11 (1%)
Empty
Remove Empty
53882 Male
10434 Female
27419 Female
60491 Female
12109 Female
12095 Female
12175 Female
8213 Male
5317 Female
58202 Female
56112 Male
34120 Male
27458 null
25226 null
70630 null
13861 null
68794 null
64778 Male
gender
1000 (100%)
Valid
0 (0%)
Error
0 (0%)
Empty
53882 Male
10434 Female
27419 Female
60491 Female
12109 Female
12095 Female
12175 Female
8213 Male
5317 Female
58202 Female
56112 Male
34120 Male
27458 NA
25226 NA
70630 NA
13861 NA
68794 NA
64778 Male
4219 Male

# Age - Replace null values with mean

- Age is a numeric column we can replace null values with mean.
- But, how to mention the value here?
- We need to use a trick



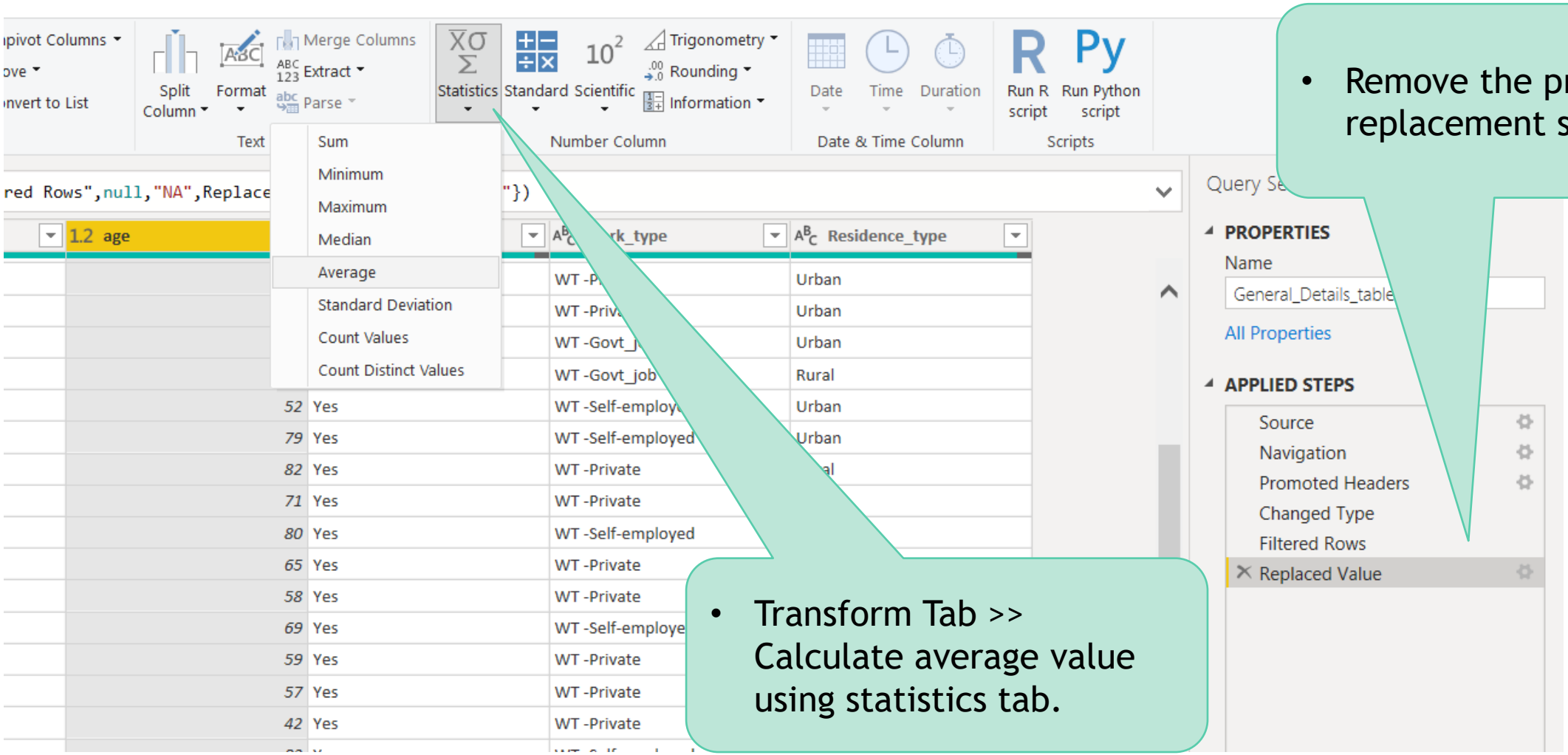
# Trick - Replace null values with mean

First replace by “0” get the formula

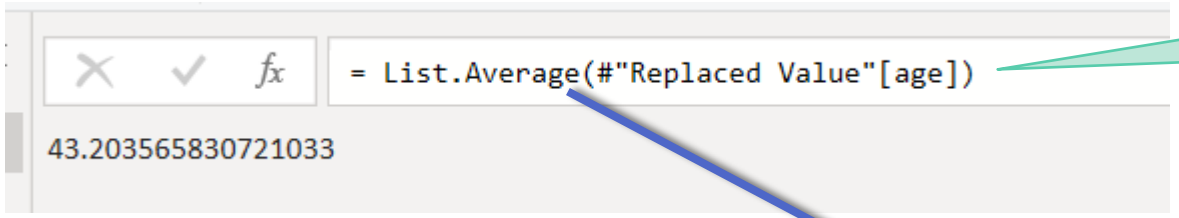
```
= Table.ReplaceValue(#"Replaced Value",null,0,Replacer.ReplaceValue,{"age"})
```

- This “0” need to be replaced by average value
- Now lest calculate the average value of age column and get the formula for it

# Trick - Replace null values with mean



# Trick - Replace null values with mean



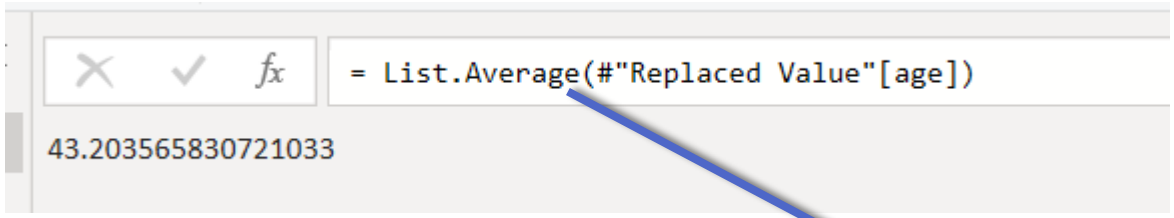

- Take this formula and replace it in the place of “0”

```
= Table.ReplaceValue(#"Replaced Value",null,0,Replacer.ReplaceValue,{"age"})
```

```
= Table.ReplaceValue(#"Replaced Value",null,List.Average(#"Replaced Value"[age]),Replacer.ReplaceValue,{"age"})
```

# Video- Replace null values with mean

# Power Query M Formula Language




- This code is known as the Power Query M Formula Language

```
= Table.ReplaceValue(#"Replaced Value",null,0,Replacer.ReplaceValue,{"age"})
```

```
= Table.ReplaceValue(#"Replaced Value",null,List.Average(#"Replaced Value"[age]),Replacer.ReplaceValue,{"age"})
```

# M- language

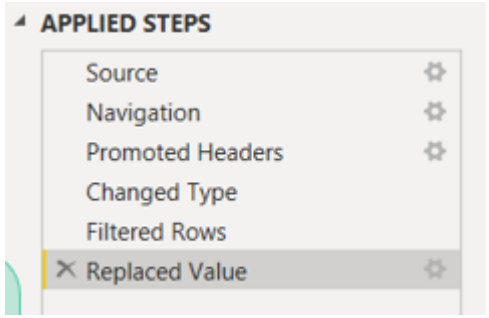
- The M stands for data Mash-up, as power query is all about connecting to various different data sources and “Mashing” them up.
- M code is the language behind the scenes of power query.
- It's a functional, case sensitive language
- When you create a data transformation in the power query editor UI, Power BI is writing the corresponding M code for the query.
- M code comes with a very large library of predefined functions available and you can also create your own.

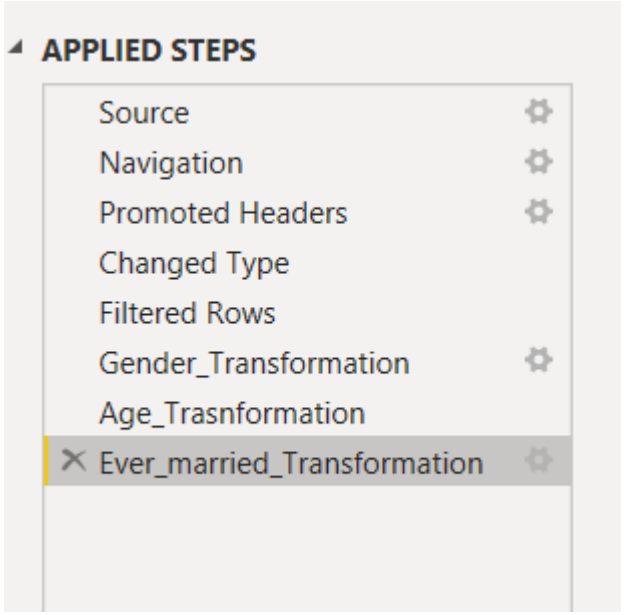
# Transform – “ever married”

- Replace missing values with “NA”

# Before you proceed

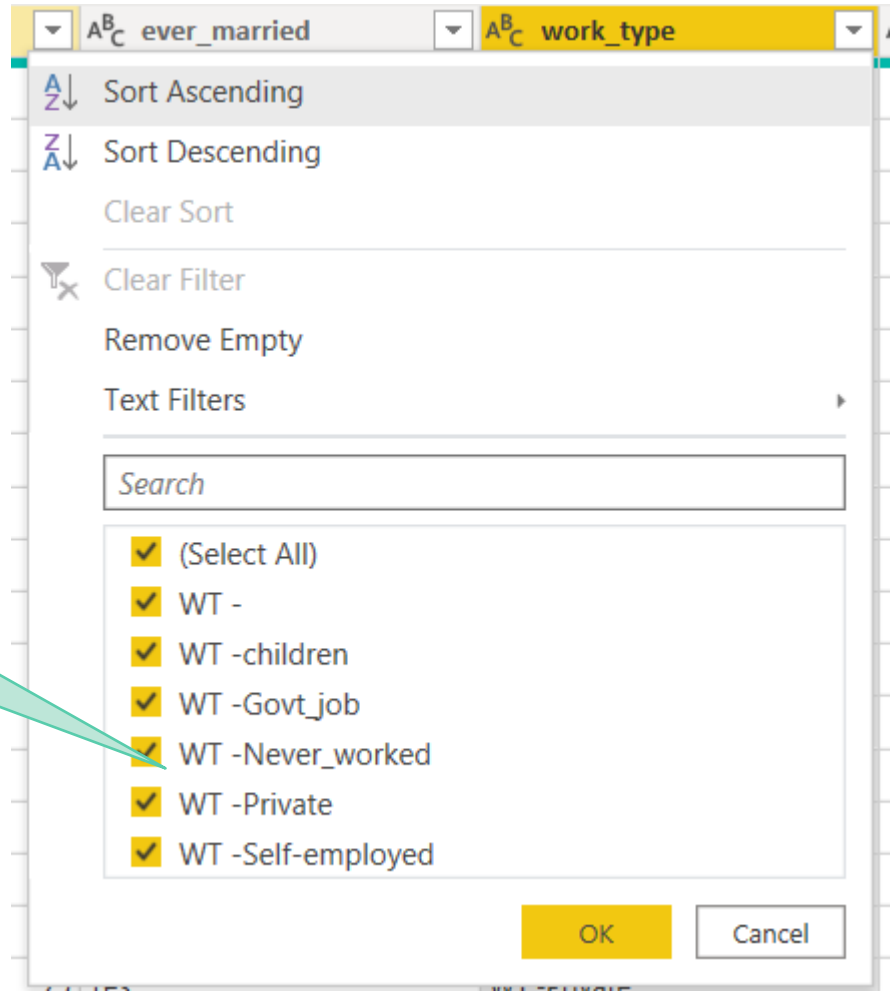
- Rename each transformation step



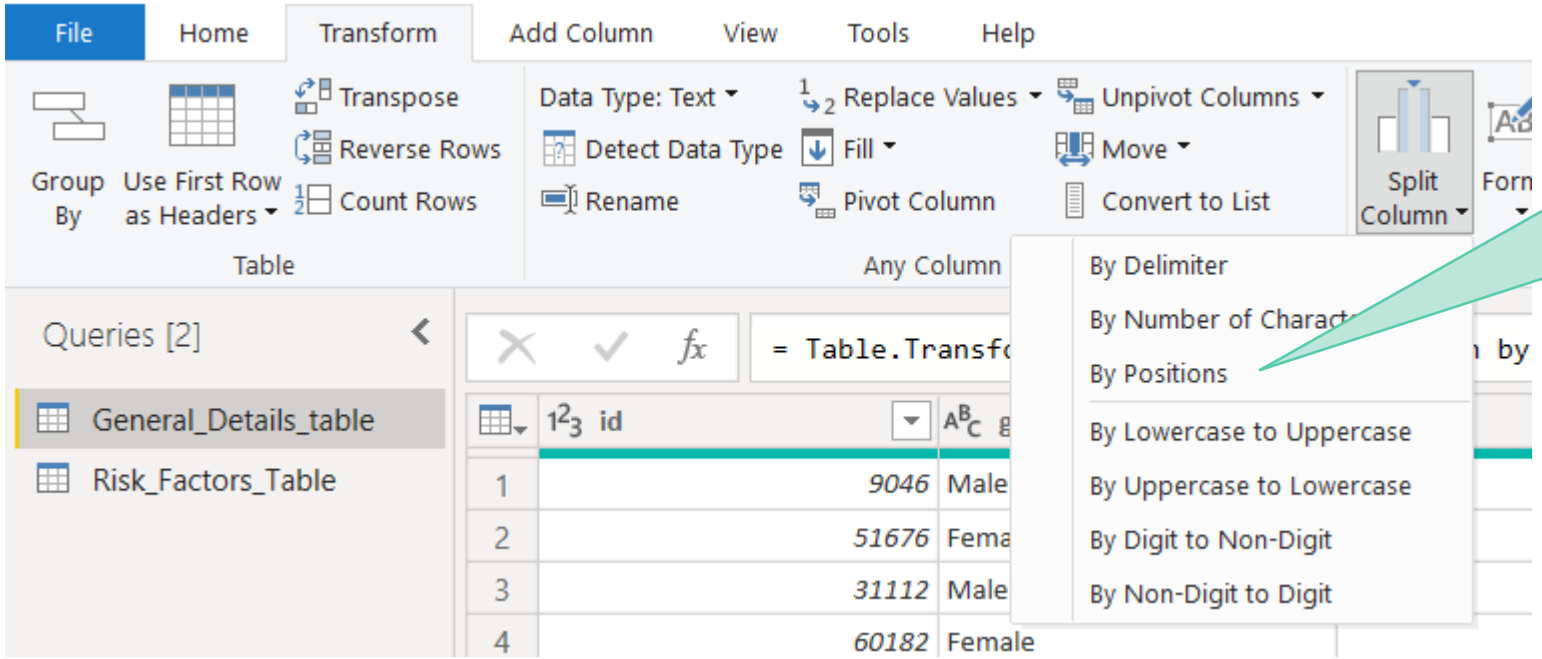



# Transform – work_type

- Remove WT from work_type column values
- There are a couple of ways to do this
- Create a duplicate column to apply different methods

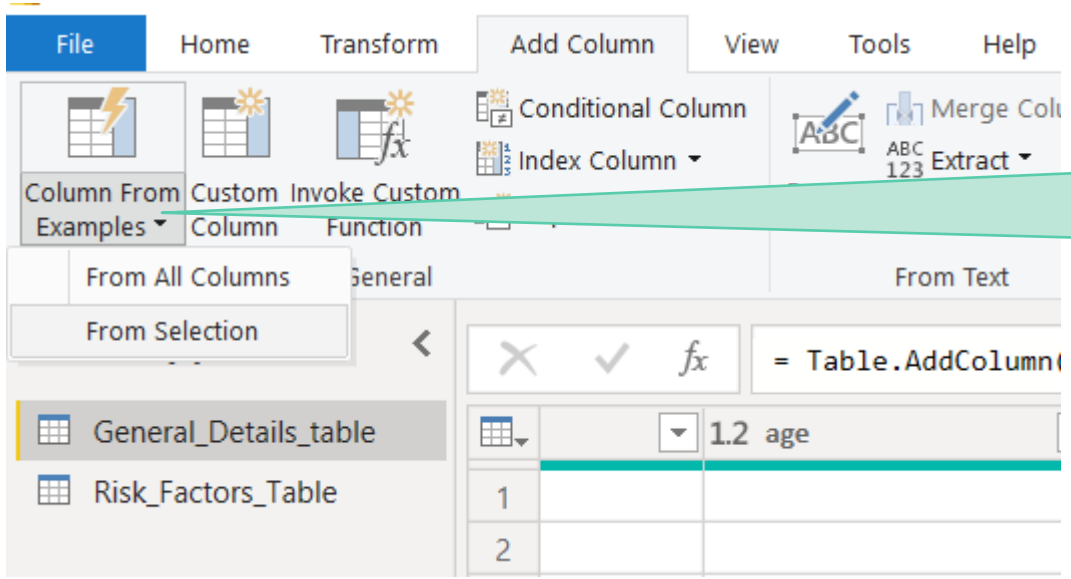


# Transform – work_type - Using split column



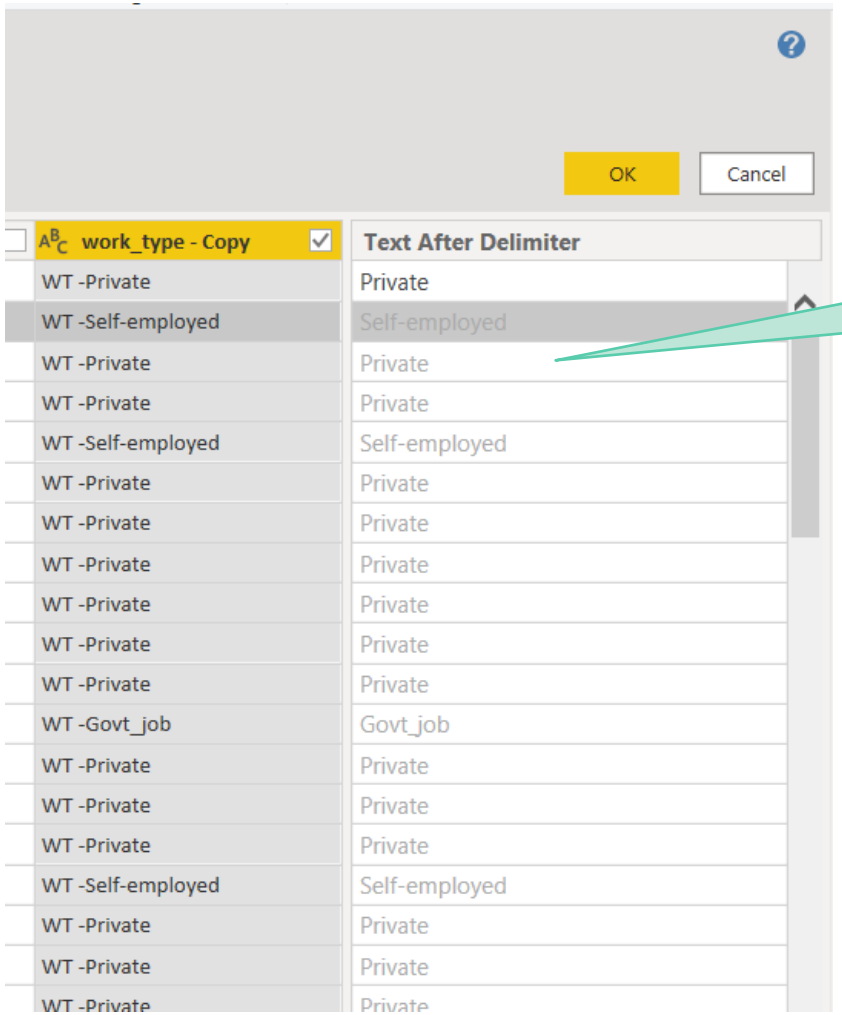

- Transform >> Split Column >> By Position >> 0 - 4

# Transform – work_type - Using Column from Examples




- Select the column >> Add Column >> Column From Examples >> From Selection

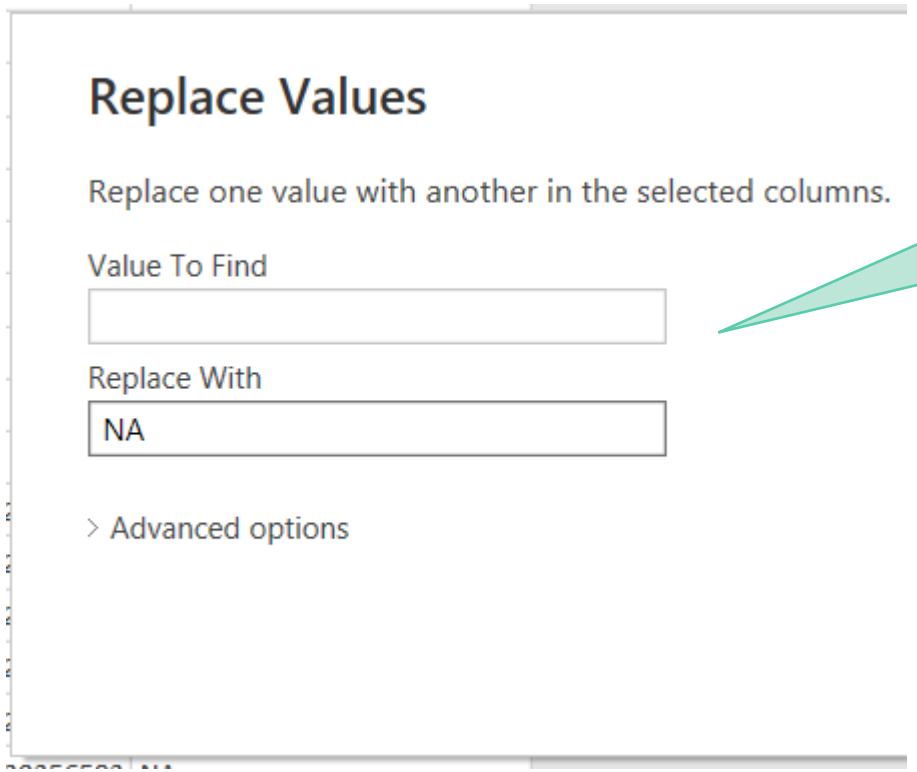
# Transform – work_type - Using Column from Examples




- Enter a sample value. Rest of the values will be filled automatically.

# Transform – work_type- Further Cleaning

- Replace <<space>> values with “NA”




- We have to replace the blank spaces

# Transform – Residence_type

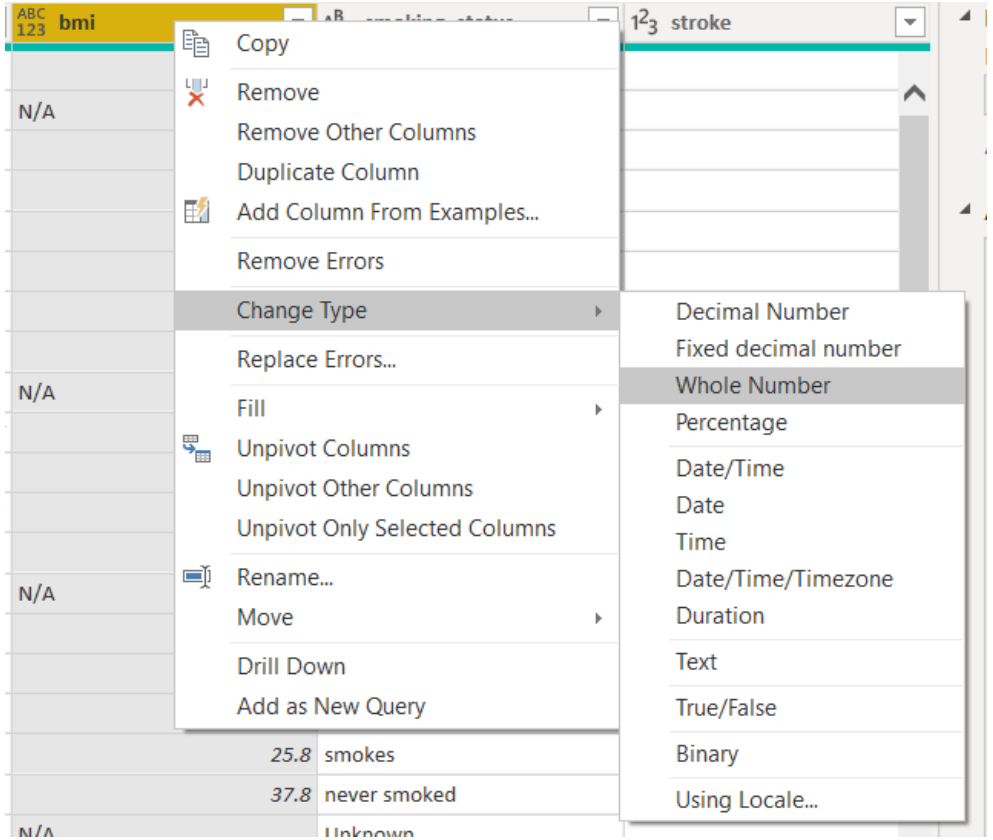
- Replace null values with “NA”

# Remaining cleaning

- Finally delete the un-necessary columns related to work type.

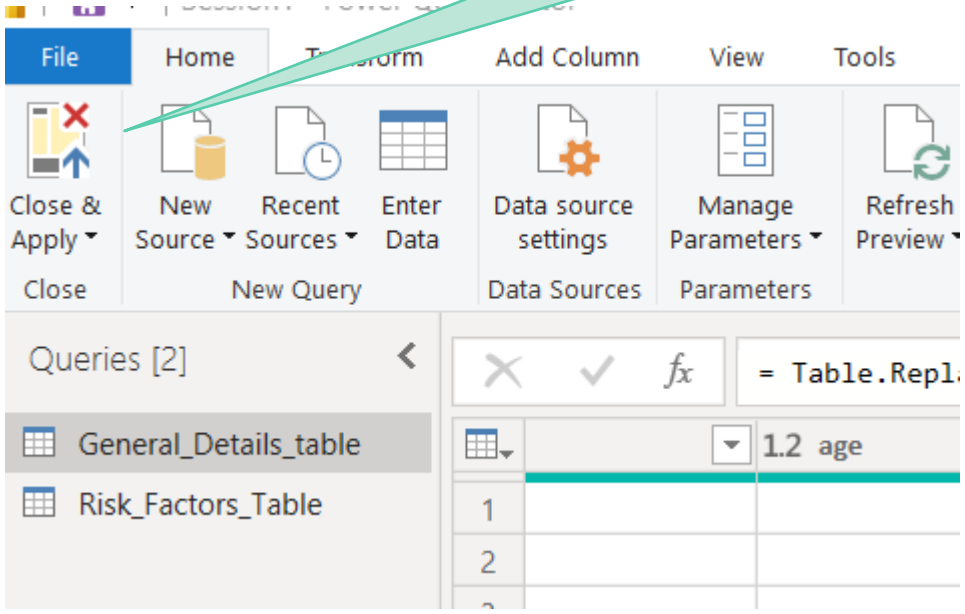
# Transforming the Risk factors table

1. Change the bmi column type to numeric
2. Transform >> Replace Values >> Replace Errors >> null
3. Calculate the average >> Take the formula >> Replace with zero and then include the formula



# Important step before you proceed...

- You have to click on Close and Apply
- Without this, if you close the query tab, then the steps will not be applied.



# Next Step –
# Step4: Model the data and create derived columns