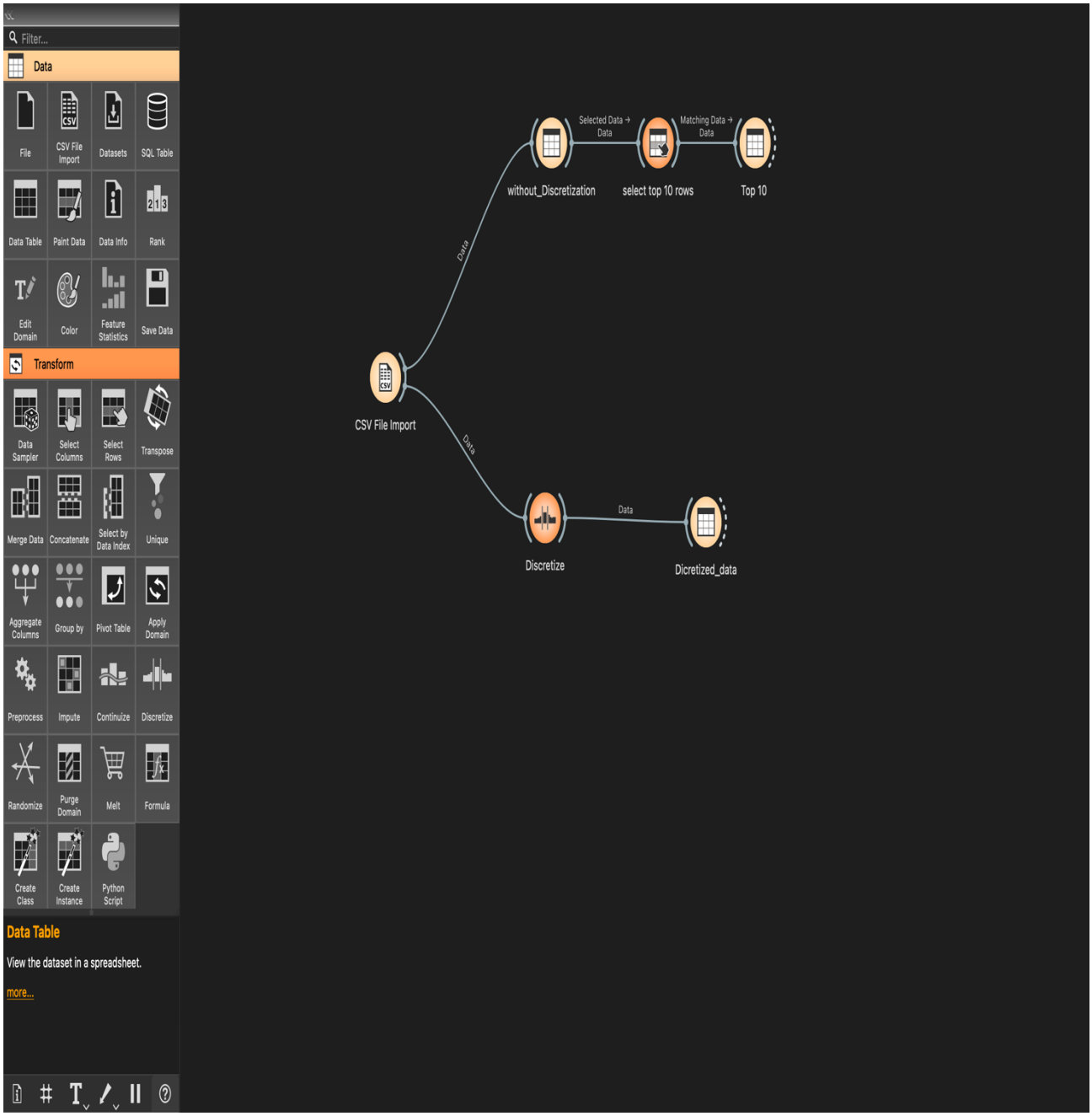
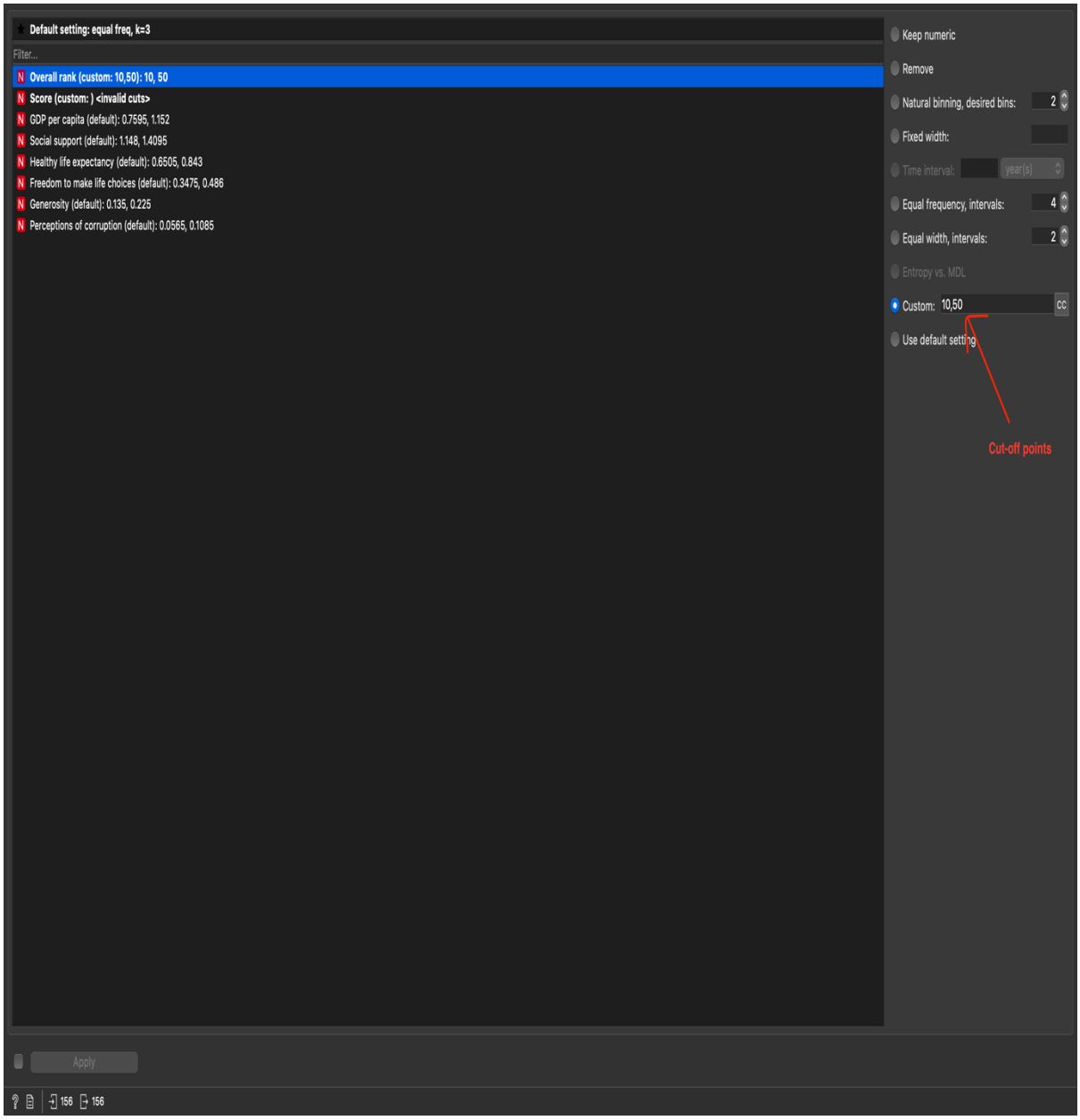
# Assignment I – Data Type Portability: Discretization

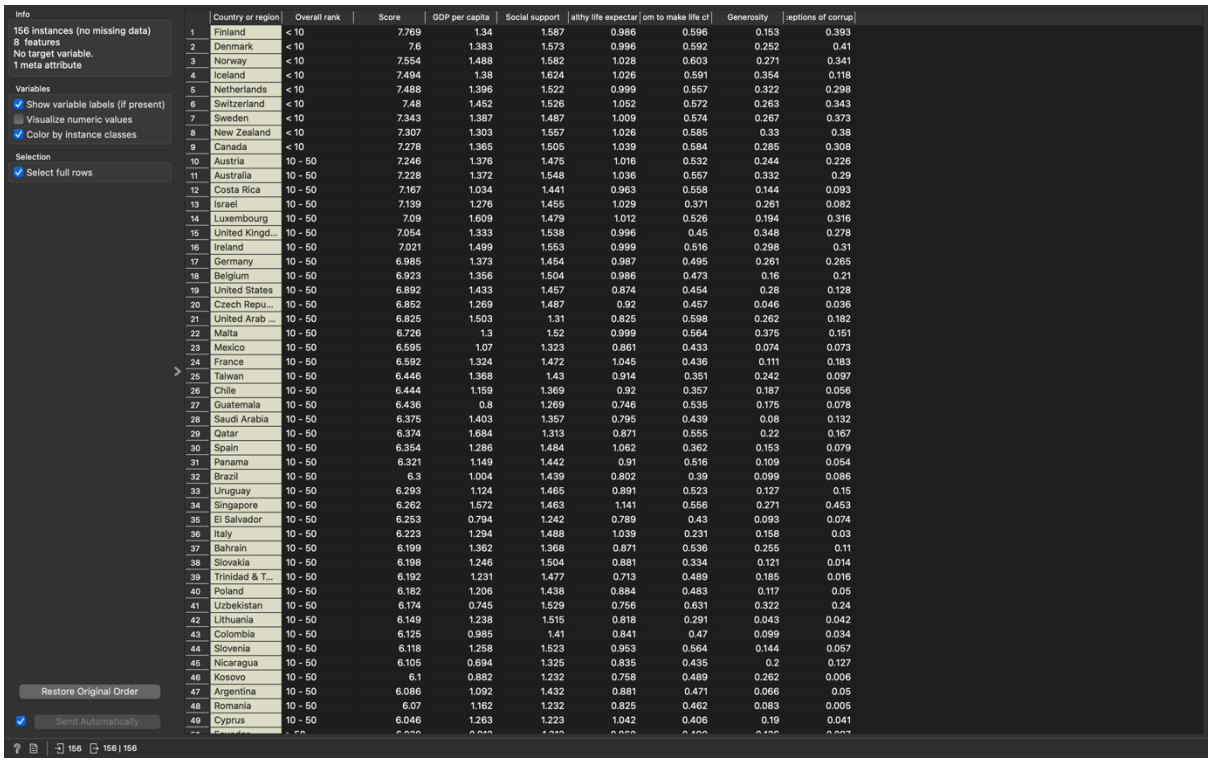
- Screenshot of the complete workflow:



- Discretization step

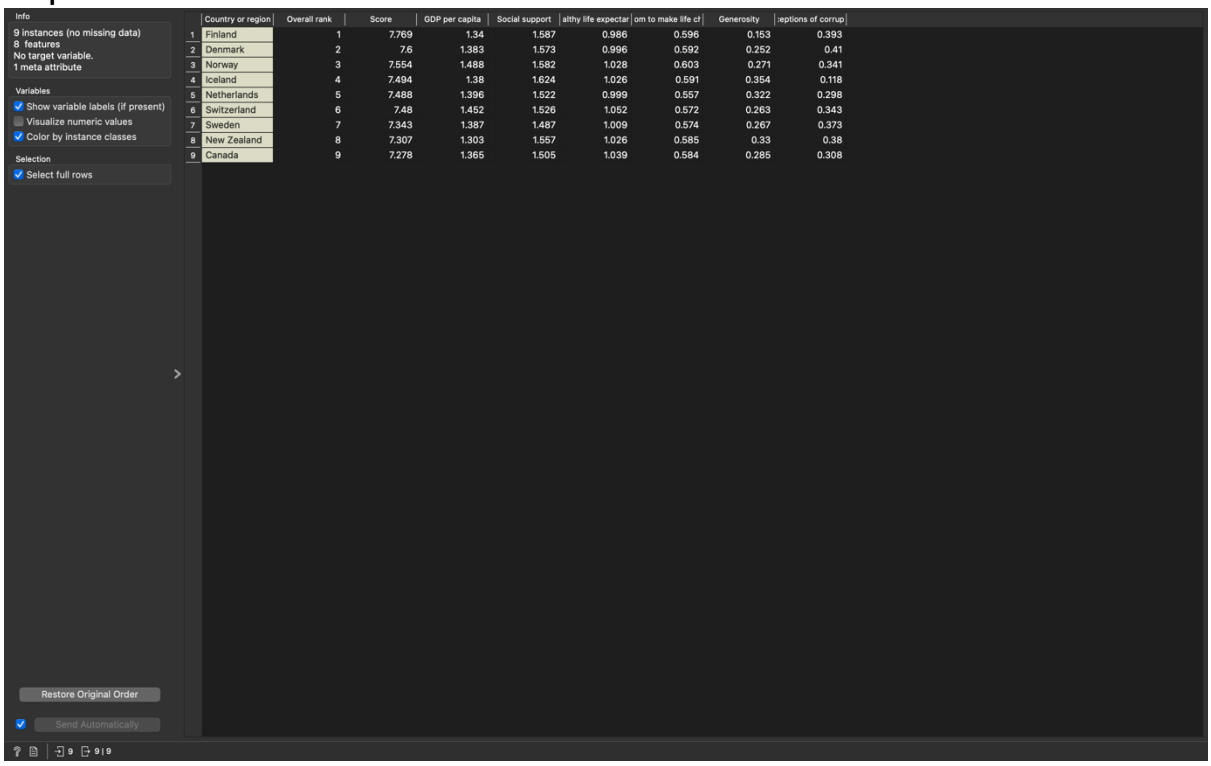


- Resulting data



| | Country or region | Overall rank | Score | GDP per capita | Social support | althy life expectar | om to make life ch | Generosity | :eptions of corrup |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Finland | < 10 | 7.769 | 1.34 | 1.587 | 0.986 | 0.596 | 0.153 | 0.393 |
| 2 | Denmark | < 10 | 7.6 | 1.383 | 1.573 | 0.996 | 0.592 | 0.252 | 0.41 |
| 3 | Norway | < 10 | 7.554 | 1.488 | 1.582 | 1.028 | 0.603 | 0.271 | 0.341 |
| 4 | Iceland | < 10 | 7.494 | 1.38 | 1.624 | 1.026 | 0.591 | 0.354 | 0.118 |
| 5 | Netherlands | < 10 | 7.488 | 1.396 | 1.522 | 0.999 | 0.557 | 0.322 | 0.298 |
| 6 | Switzerland | < 10 | 7.48 | 1.452 | 1.526 | 1.052 | 0.572 | 0.263 | 0.343 |
| 7 | Sweden | < 10 | 7.343 | 1.387 | 1.487 | 1.009 | 0.574 | 0.267 | 0.373 |
| 8 | New Zealand | < 10 | 7.307 | 1.303 | 1.557 | 1.026 | 0.585 | 0.33 | 0.38 |
| 9 | Canada | < 10 | 7.278 | 1.365 | 1.505 | 1.039 | 0.584 | 0.285 | 0.308 |
| 10 | Austria | 10 - 50 | 7.246 | 1.376 | 1.475 | 1.016 | 0.532 | 0.244 | 0.226 |
| 11 | Australia | 10 - 50 | 7.228 | 1.372 | 1.548 | 1.036 | 0.557 | 0.332 | 0.29 |
| 12 | Costa Rica | 10 - 50 | 7.167 | 1.034 | 1.441 | 0.963 | 0.558 | 0.144 | 0.093 |
| 13 | Israel | 10 - 50 | 7.139 | 1.276 | 1.455 | 1.029 | 0.371 | 0.261 | 0.082 |
| 14 | Luxembourg | 10 - 50 | 7.09 | 1.609 | 1.479 | 1.012 | 0.526 | 0.194 | 0.316 |
| 15 | United Kingd... | 10 - 50 | 7.054 | 1.333 | 1.538 | 0.996 | 0.45 | 0.348 | 0.278 |
| 16 | Ireland | 10 - 50 | 7.021 | 1.499 | 1.553 | 0.999 | 0.516 | 0.298 | 0.31 |
| 17 | Germany | 10 - 50 | 6.985 | 1.373 | 1.454 | 0.987 | 0.495 | 0.261 | 0.265 |
| 18 | Belgium | 10 - 50 | 6.923 | 1.356 | 1.504 | 0.986 | 0.473 | 0.16 | 0.21 |
| 19 | United States | 10 - 50 | 6.892 | 1.433 | 1.457 | 0.874 | 0.454 | 0.28 | 0.128 |
| 20 | Czech Repu... | 10 - 50 | 6.852 | 1.269 | 1.487 | 0.92 | 0.457 | 0.046 | 0.036 |
| 21 | United Arab ... | 10 - 50 | 6.825 | 1.503 | 1.31 | 0.825 | 0.598 | 0.262 | 0.182 |
| 22 | Malta | 10 - 50 | 6.726 | 1.3 | 1.52 | 0.999 | 0.564 | 0.375 | 0.151 |
| 23 | Mexico | 10 - 50 | 6.595 | 1.07 | 1.323 | 0.861 | 0.433 | 0.074 | 0.073 |
| 24 | France | 10 - 50 | 6.592 | 1.324 | 1.472 | 1.045 | 0.436 | 0.111 | 0.183 |
| 25 | Taiwan | 10 - 50 | 6.446 | 1.368 | 1.43 | 0.914 | 0.351 | 0.242 | 0.097 |
| 26 | Chile | 10 - 50 | 6.444 | 1.159 | 1.369 | 0.92 | 0.357 | 0.187 | 0.056 |
| 27 | Guatemala | 10 - 50 | 6.436 | 0.8 | 1.269 | 0.746 | 0.535 | 0.175 | 0.078 |
| 28 | Saudi Arabia | 10 - 50 | 6.375 | 1.403 | 1.357 | 0.795 | 0.439 | 0.08 | 0.132 |
| 29 | Qatar | 10 - 50 | 6.374 | 1.684 | 1.313 | 0.871 | 0.555 | 0.22 | 0.167 |
| 30 | Spain | 10 - 50 | 6.354 | 1.286 | 1.484 | 1.062 | 0.362 | 0.153 | 0.079 |
| 31 | Panama | 10 - 50 | 6.321 | 1.149 | 1.442 | 0.91 | 0.516 | 0.109 | 0.054 |
| 32 | Brazil | 10 - 50 | 6.3 | 1.004 | 1.439 | 0.802 | 0.39 | 0.099 | 0.086 |
| 33 | Uruguay | 10 - 50 | 6.293 | 1.124 | 1.465 | 0.891 | 0.523 | 0.127 | 0.15 |
| 34 | Singapore | 10 - 50 | 6.262 | 1.572 | 1.463 | 1.141 | 0.556 | 0.271 | 0.453 |
| 35 | El Salvador | 10 - 50 | 6.253 | 0.794 | 1.242 | 0.789 | 0.43 | 0.093 | 0.074 |
| 36 | Italy | 10 - 50 | 6.223 | 1.294 | 1.488 | 1.039 | 0.231 | 0.158 | 0.03 |
| 37 | Bahrain | 10 - 50 | 6.199 | 1.362 | 1.368 | 0.871 | 0.536 | 0.255 | 0.11 |
| 38 | Slovakia | 10 - 50 | 6.198 | 1.246 | 1.504 | 0.881 | 0.334 | 0.121 | 0.014 |
| 39 | Trinidad & T... | 10 - 50 | 6.192 | 1.231 | 1.477 | 0.713 | 0.489 | 0.185 | 0.016 |
| 40 | Poland | 10 - 50 | 6.182 | 1.206 | 1.438 | 0.884 | 0.483 | 0.117 | 0.05 |
| 41 | Uzbekistan | 10 - 50 | 6.174 | 0.745 | 1.529 | 0.756 | 0.631 | 0.322 | 0.24 |
| 42 | Lithuania | 10 - 50 | 6.149 | 1.238 | 1.515 | 0.818 | 0.291 | 0.043 | 0.042 |
| 43 | Colombia | 10 - 50 | 6.125 | 0.985 | 1.41 | 0.841 | 0.47 | 0.099 | 0.034 |
| 44 | Slovenia | 10 - 50 | 6.118 | 1.258 | 1.523 | 0.953 | 0.564 | 0.144 | 0.057 |
| 45 | Nicaragua | 10 - 50 | 6.105 | 0.694 | 1.325 | 0.835 | 0.435 | 0.2 | 0.127 |
| 46 | Kosovo | 10 - 50 | 6.1 | 0.882 | 1.232 | 0.758 | 0.489 | 0.262 | 0.006 |
| 47 | Argentina | 10 - 50 | 6.086 | 1.092 | 1.432 | 0.881 | 0.471 | 0.066 | 0.05 |
| 48 | Romania | 10 - 50 | 6.07 | 1.162 | 1.232 | 0.825 | 0.462 | 0.083 | 0.005 |
| 49 | Cyprus | 10 - 50 | 6.046 | 1.263 | 1.223 | 1.042 | 0.406 | 0.19 | 0.041 |

- Top 10 countries



| | Country or region | Overall rank | Score | GDP per capita | Social support | althy life expectar | om to make life ch | Generosity | :eptions of corrup |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Finland | 1 | 7.769 | 1.34 | 1.587 | 0.986 | 0.596 | 0.153 | 0.393 |
| 2 | Denmark | 2 | 7.6 | 1.383 | 1.573 | 0.996 | 0.592 | 0.252 | 0.41 |
| 3 | Norway | 3 | 7.554 | 1.488 | 1.582 | 1.028 | 0.603 | 0.271 | 0.341 |
| 4 | Iceland | 4 | 7.494 | 1.38 | 1.624 | 1.026 | 0.591 | 0.354 | 0.118 |
| 5 | Netherlands | 5 | 7.488 | 1.396 | 1.522 | 0.999 | 0.557 | 0.322 | 0.298 |
| 6 | Switzerland | 6 | 7.48 | 1.452 | 1.526 | 1.052 | 0.572 | 0.263 | 0.343 |
| 7 | Sweden | 7 | 7.343 | 1.387 | 1.487 | 1.009 | 0.574 | 0.267 | 0.373 |
| 8 | New Zealand | 8 | 7.307 | 1.303 | 1.557 | 1.026 | 0.585 | 0.33 | 0.38 |
| 9 | Canada | 9 | 7.278 | 1.365 | 1.505 | 1.039 | 0.584 | 0.285 | 0.308 |

## Assignment II – Pre-Processing of Text Data

1. a) First, remove all preprocessing steps except "Transformation" and transform the corpus into lowercase. This avoids to have multiple versions ("Large" and "large") of the same word. If you checked the box "Apply Automatically" on the lower left, the changes are directly ap- plied. You can use the "Word Cloud" widget to observe the effect of the transformation. Please note that also punctuation symbols are also counted. Save the word cloud clicking on the disk symbol.

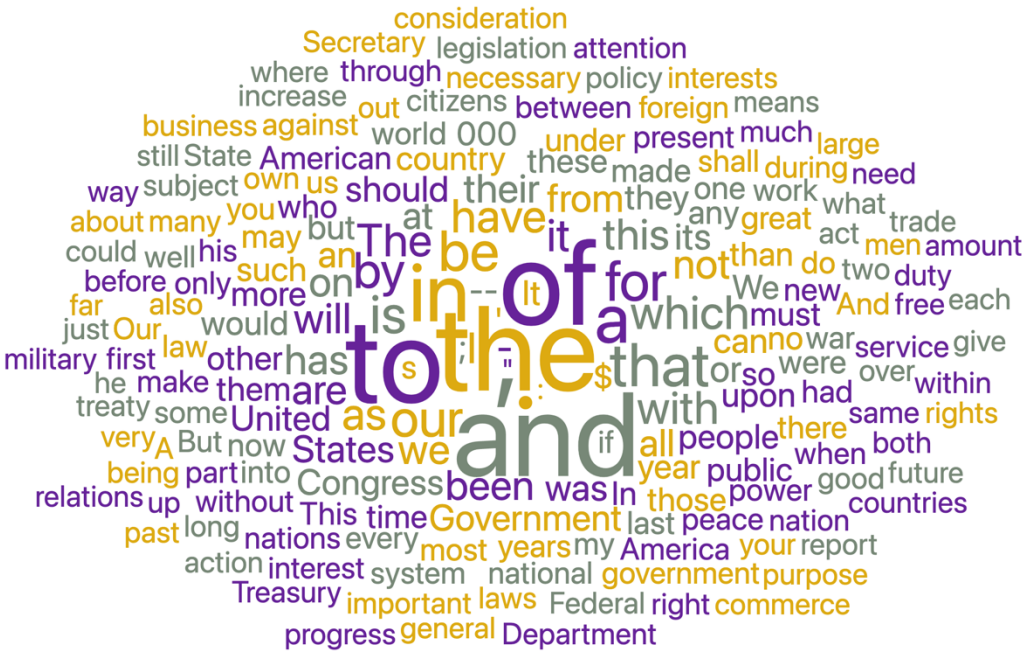



Fig: Word Cloud – Before lower case conversion

We can see that one of the most frequent words are 'the' and 'THE'.

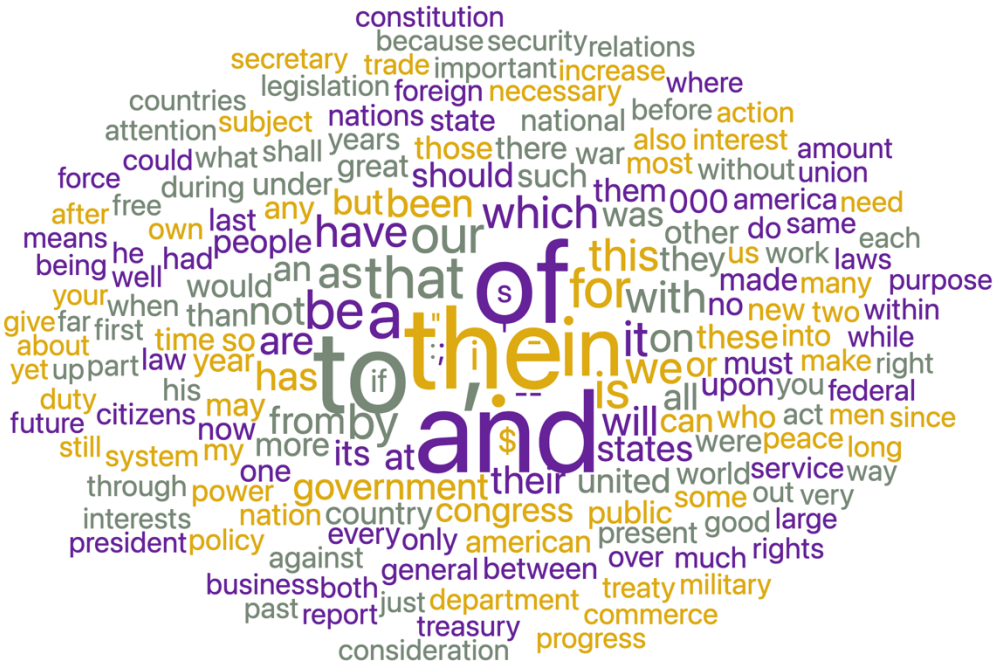



Fig: Word Cloud – After lower case conversion

Now we don’t have multiple versions of same word.

1. b) Now apply “Tokenization”. This is used to break the text into smaller pieces like sentences, words, ... Please have a look at the aforementioned documentation to understand the different tokenization methods. Use Regular Expressions to split the text by words without keeping punctuation. This is quite a common way to break down text. What are the top 3 words? Save the word cloud clicking on the disk symbol.

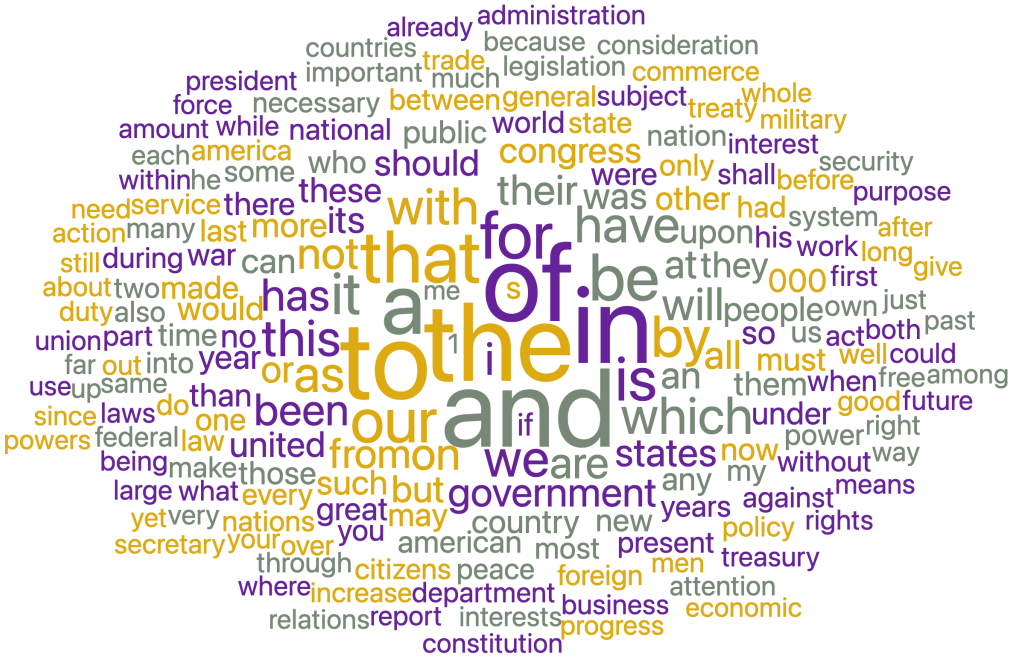



Fig: Tokenization without keeping punctuations

The top 3 words are ‘the’, ‘of’ and ‘and’.

1.c) Now filter the stopwords. What are the top 3 words now? Save the word cloud clicking on the disk symbol.


Fig: Word Cloud – Stop words filtered

Now the top 3 words are ‘government’, ‘states’ and ‘congress’.

1. d) Apply the standard normalization (Porter Stemmer). This does, to put it simple, convert words to their base form, like e.g. “the boy's cars are different colors” à “the boy car be differ color”. What do you observe?


Fig: Word Cloud – After Normalization

Now we observe that word are converted to their base forms for ex: ‘government’ is converted to ‘govern’.

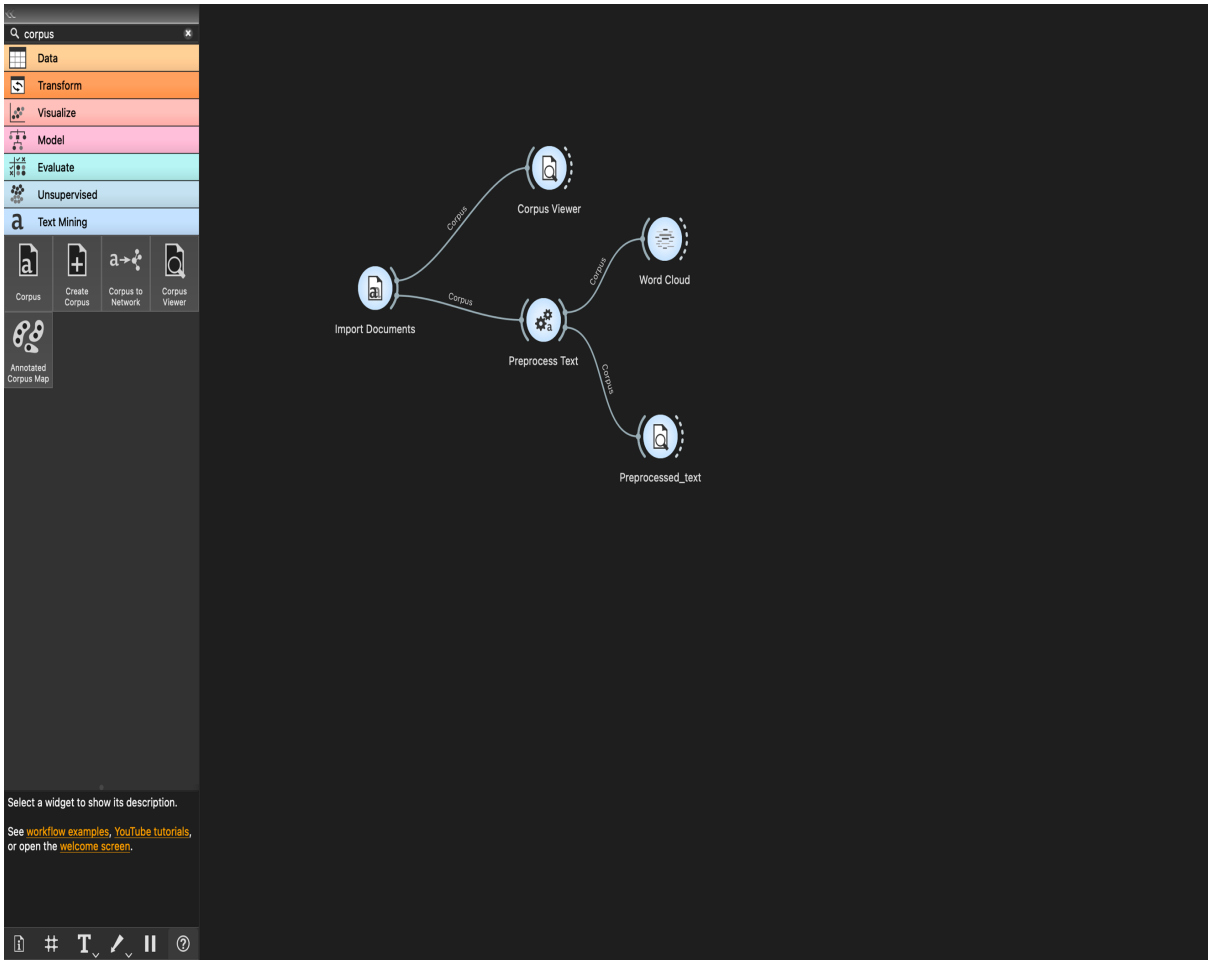



Fig: Complete Workflow