Feature Selection - Member 3

Group K

2025-09-19

Load merged datasets

```
## Rows: 633 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Dataset, CharacteristicCategory, CharacteristicLabel, IndicatorId, ...
## dbl (5): SurveyYear, CharacteristicId, Value, DenominatorWeighted, Denominat...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 142 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Dataset, CharacteristicCategory, CharacteristicLabel, IndicatorId, ...
## dbl (5): SurveyYear, CharacteristicId, Value, DenominatorWeighted, Denominat...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 59 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Dataset, CharacteristicCategory, CharacteristicLabel, IndicatorId, ...
## dbl (5): SurveyYear, CharacteristicId, Value, DenominatorWeighted, Denominat...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 126 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Dataset, CharacteristicCategory, CharacteristicLabel, IndicatorId, ...
## dbl (5): SurveyYear, CharacteristicId, Value, DenominatorWeighted, Denominat...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

## Warning in styling_latex_scale(out, table_info, "down"): Longtable cannot be
## resized.
```

Table 1: Key columns of the dataset

SurveyYear	CharacteristicLabel	Value	DenominatorWeighted	DenominatorUnweighted
1998	Total	28.5	2871	2903
1998	Total	30.0	4122	4148
1998	Total	27.3	2010	2041
1998	Total	66.6	2871	2903
1998	Total	65.0	4122	4148
1998	Total	68.4	2010	2041
1998	Total	0.1	2871	2903
1998	Total	0.6	2871	2903
1998	Total	0.7	4122	4148
1998	Total	1.2	2871	2903

Table 2: Summary Statistics for Numeric Variables

Variable	Statistic	Value
CharacteristicId	mean	21700.135417
CharacteristicId	sd	70296.424203
CharacteristicId	min	1000.000000
CharacteristicId	max	295001.000000
DenominatorUnweighted	mean	5876.616949
DenominatorUnweighted	sd	11388.816495
DenominatorUnweighted	min	59.000000
DenominatorUnweighted	max	52465.000000
DenominatorWeighted	mean	5820.471783
DenominatorWeighted	sd	11325.270089
DenominatorWeighted	min	68.000000
DenominatorWeighted	max	52007.000000
SurveyYear	mean	2009.081250
SurveyYear	sd	8.760613
SurveyYear	min	1998.000000
SurveyYear	max	2016.000000
Value	mean	961.535625
Value	sd	4379.877231
Value	min	-1.100000
Value	max	52465.000000

Table 3: Missing Values Per Column

	x
Dataset	0
SurveyYear	0
CharacteristicId	0
CharacteristicCategory	0
CharacteristicLabel	0
IndicatorId	0
IndicatorType	0
Value	0

	x
DenominatorWeighted	74
DenominatorUnweighted	75

```
## No duplicates found.
```

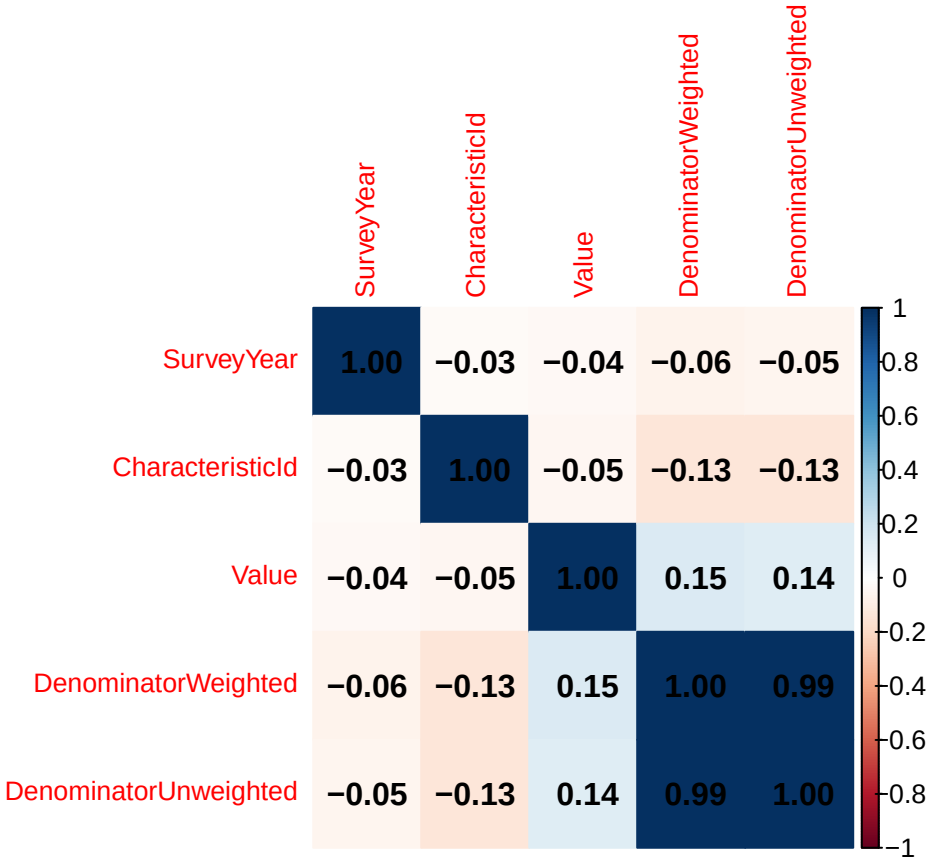


Table 4: Correlation Matrix for Numerical Features

	SurveyYear	CharacteristicId	Value	DenominatorWeighted	DenominatorUnweighted
SurveyYear	1.0000000	-0.0269827	-0.0369511	-0.0605797	-0.0522899
CharacteristicId	-0.0269827	1.0000000	-0.0463360	-0.1327941	-0.1328899
Value	-0.0369511	-0.0463360	1.0000000	0.1515455	0.1396777
DenominatorWeighted	-0.0605797	-0.1327941	0.1515455	1.0000000	0.9917636
DenominatorUnweighted	-0.0522899	-0.1328899	0.1396777	0.9917636	1.0000000

```
##  [1] "Dataset"                "SurveyYear"             "CharacteristicId"
##  [4] "CharacteristicCategory" "CharacteristicLabel"    "IndicatorId"
##  [7] "IndicatorType"          "Value"                  "DenominatorWeighted"
## [10] "DenominatorUnweighted"
```

```
## [1] 28.5 30.0 27.3 66.6 65.0 68.4
```

```
## [1] 2871 4122 2010 2871 4122 2010
```

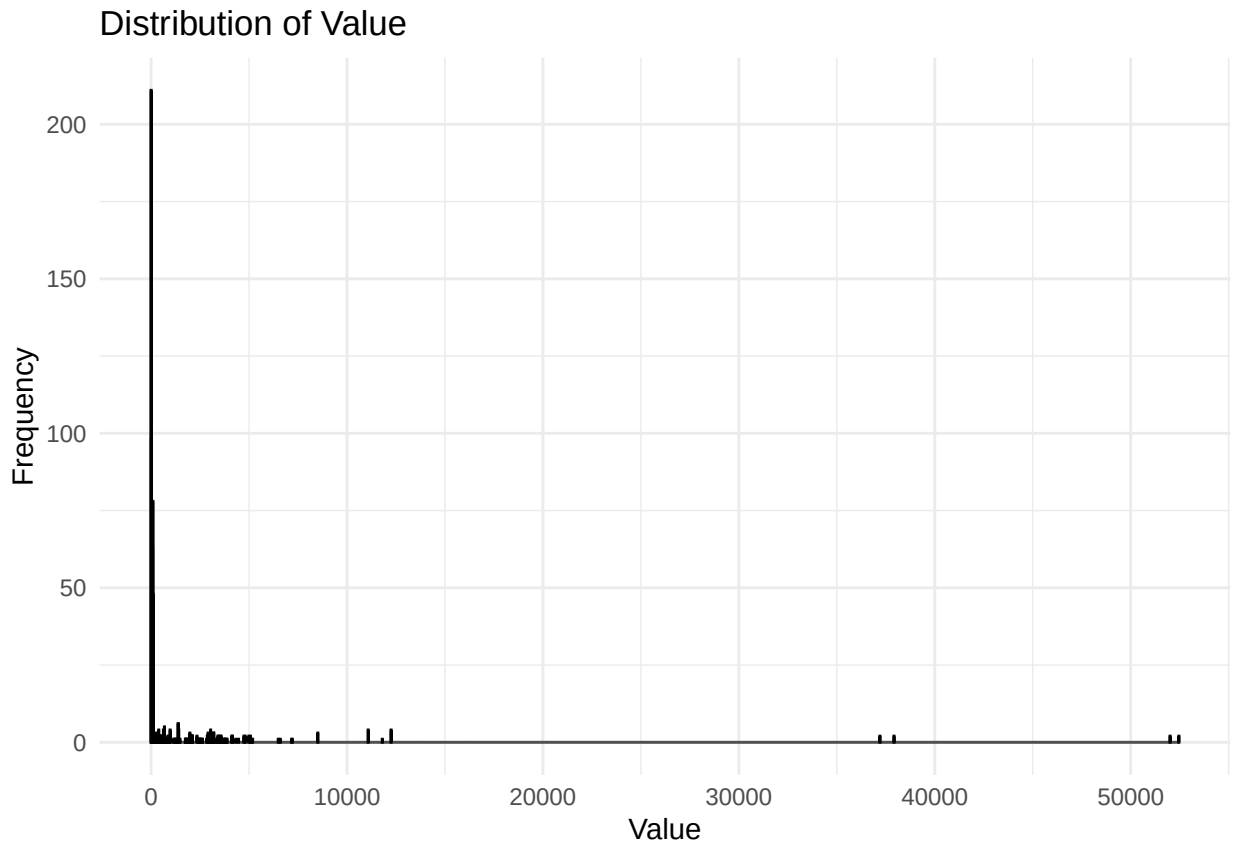
```
##
##  Pearson's product-moment correlation
##
## data:  cleaned_data$Value and cleaned_data$DenominatorWeighted
## t = 6.6072, df = 884, p-value = 6.757e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1532610 0.2788127
## sample estimates:
##       cor
## 0.2169338
```

```
##
##  Pearson's product-moment correlation
##
## data:  cleaned_data$Value and cleaned_data$DenominatorUnweighted
## t = 4.1822, df = 879, p-value = 3.177e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.07431399 0.20384635
## sample estimates:
##       cor
## 0.1396777
```

```
##
##  Pearson's product-moment correlation
##
## data:  cleaned_data$DenominatorWeighted and cleaned_data$DenominatorUnweighted
## t = 229.57, df = 879, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9906042 0.9927806
## sample estimates:
##       cor
## 0.9917636
```

Table 5: Selected Features for Modeling

Value	DenominatorWeighted	DenominatorUnweighted
28.5	2871	2903
30.0	4122	4148
27.3	2010	2041
66.6	2871	2903
65.0	4122	4148
68.4	2010	2041




```
## Feature selection was performed based on correlation analysis and statistical tests. The selected fea
```