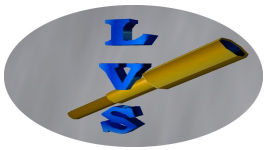
Optimizing Language Variation Analysis: Language Variation Suite

Olga Scrivner, Manuel Díaz-Campos and Rafael Orozco

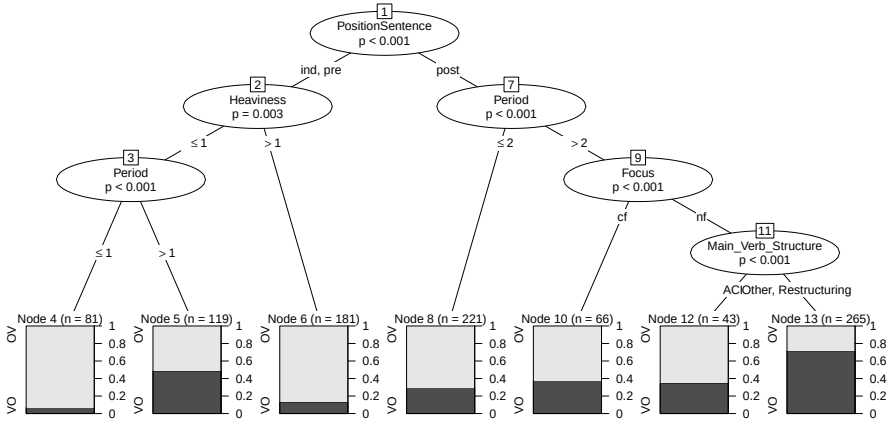


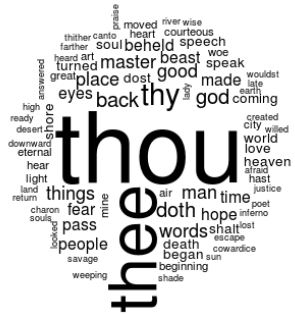
Indiana University and Louisiana State University

NWAV45, 2016

Goal

Provide researchers with a variety of quantitative methods to advance language variation studies.





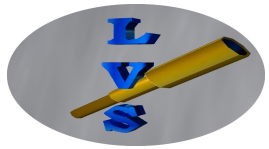
Objectives

1. Introduce a novel sociolinguistic toolkit
2. Develop practical skills
3. Understand and interpret advanced statistical models

What is LVS?

Language Variation Suite

It is a Shiny web application designed for data analysis in sociolinguistic research.



It can be used for:

- Processing spreadsheet data
- Reporting in tables and graphs
- Analyzing means, regression, conditional trees ... (and much more)

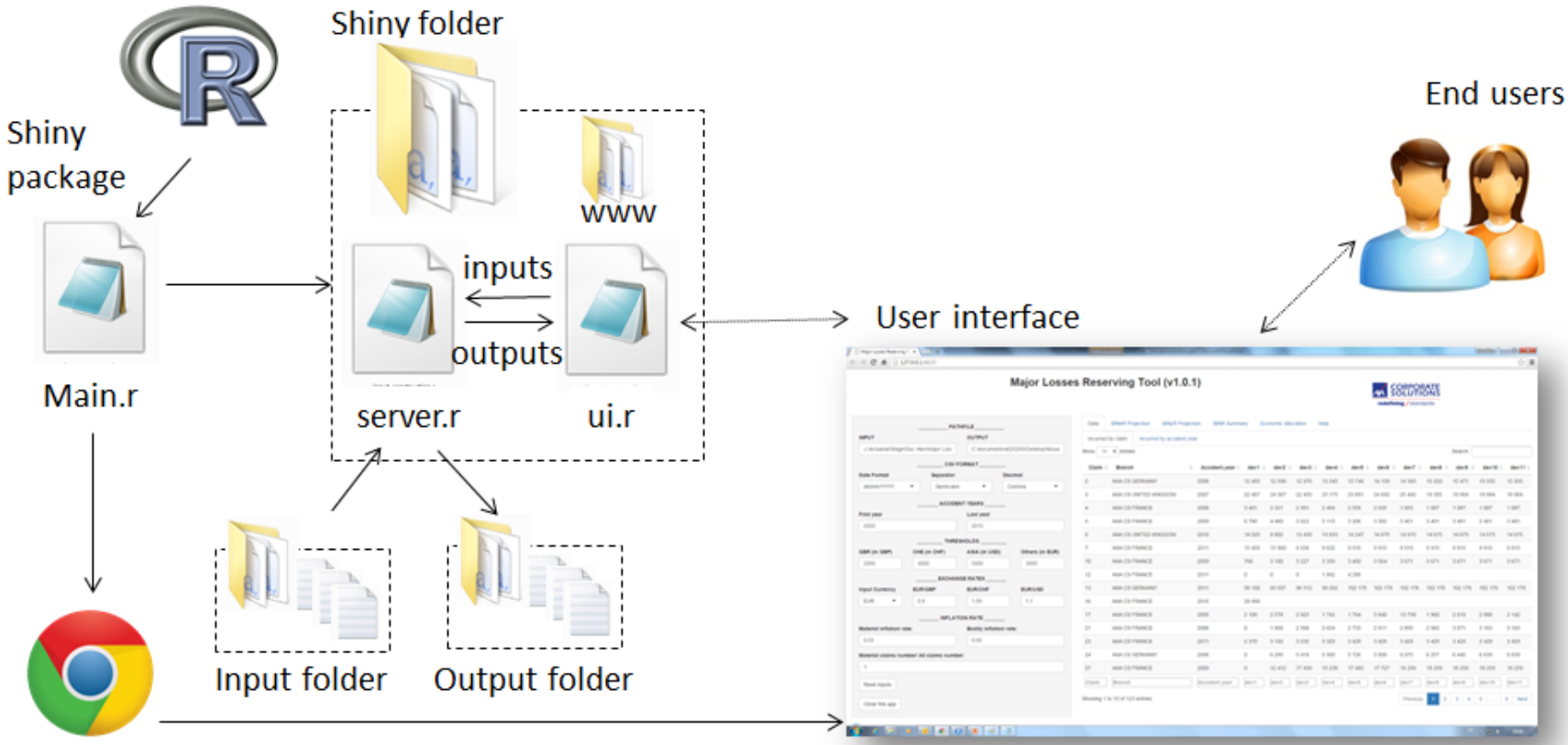
Background

LVS is built in R using Shiny package:

1. **R** - a free programming language for statistical computing and graphics
2. **Shiny App** - a web application framework for R

Computational power of R + Web interactivity

Background



http://littleactuary.github.io/blog/Web-application-framework-with-Shiny/

Data Preparation

Important things to consider before data entry:

- File format:
  - Comma separated value (CSV) - faster processing
  - Excel format will slow processing
- Column names should not contain spaces
  - Permitted: non-accented characters, numbers, underscore, hyphen, and period
- One column must contain your **dependent** variable
- The rest of the columns contain **independent** variables

A	B	C	D	E	F
Case	Number	R.Use	Lexical.Item	Style	Store
1	1	retention	Fourth	normal	Saks
1	2	retention	Fourth	normal	Saks
1	3	retention	Fourth	normal	Saks
1	4	retention	Fourth	normal	Saks
1	5	retention	Fourth	normal	Saks
1	6	retention	Fourth	normal	Saks
1	7	retention	Fourth	normal	Saks
1	8	retention	Fourth	normal	Saks

Workspace

Browser

- Chrome, Firefox, Safari - recommendable
- Explorer may cause instability issues

Accessibility

- PC, Mac, Linux
  - Data files will be uploaded from any location on your computer
- Smart Phone
  - Data files must be on a cloud platform connected to your phone account (e.g. dropbox)

Terminology Review

a. **Categorical** - non-numerical data with **two** values
   - yes - no; deletion - retention; perfective - imperfective
b. **Continuous** - numerical data
   - duration, age, chronological period
c. **Multinomial** - non-numerical data with **three or more** values
   - deletion - aspiration - retention
d. **Ordinal** - scale: currently not supported

Terminology Review

a. **Categorical** - non-numerical data with **two** values
   - yes - no; deletion - retention; perfective - imperfective
b. **Continuous** - numerical data
   - duration, age, chronological period
c. **Multinomial** - non-numerical data with **three or more** values
   - deletion - aspiration - retention
d. **Ordinal** - scale: currently not supported

ARE YOU READY?

Workshop Files

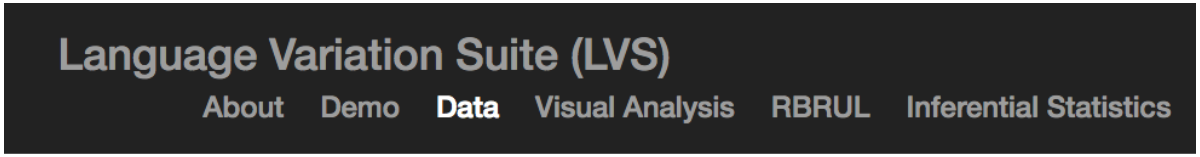
`https://languagevariationsuite.wordpress.com/`

1. **categoricaldata.csv**: categorical dependent - Labov New York 1966 study
2. **continuousdata.csv**: continuous dependent - Intervocalic /d/ in Caracas corpus (Díaz-Campos et al.)
3. **LVS web site**: `https://languagevariationsuite.shinyapps.io/Pages/`

Language Variation Suite - Structure

1. Demo
   - Brief introduction
2. Data
   - Upload file, data summary, adjust data, cross tabulation
3. Visual Analysis
   - Plotting, cluster classification
4. RBRUL
   - New version by Daniel Johnson!
5. Inferential statistics
   - Modeling, regression, varbrul analysis, conditional trees, random forest

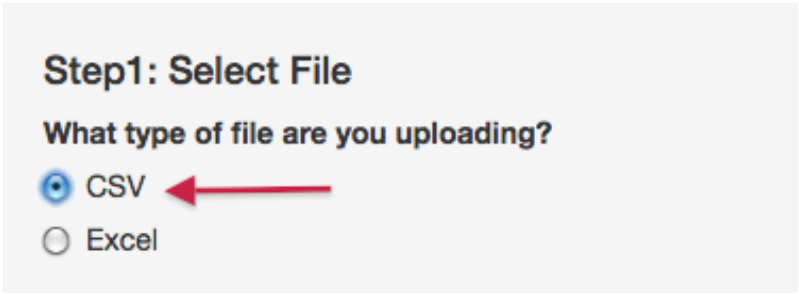
Language Variation Suite - Structure



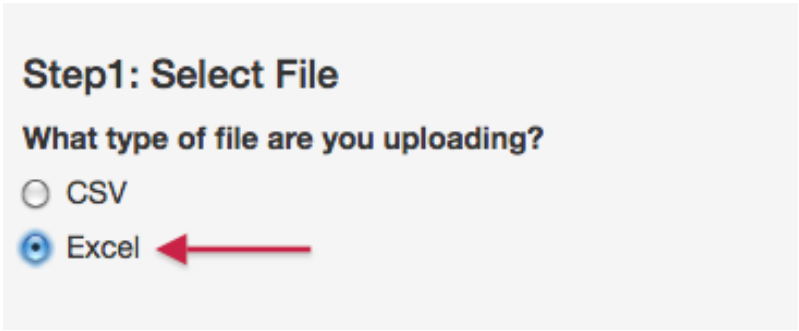
1. Demo
   - Brief introduction
2. Data
   - Upload file, data summary, adjust data, cross tabulation
3. Visual Analysis
   - Plotting, cluster classification
4. RBRUL
   - New version by Daniel Johnson!
5. Inferential statistics
   - Modeling, regression, varbrul analysis, conditional trees, random forest

Language Variation Suite - **Data**

1. Upload **CSV** file

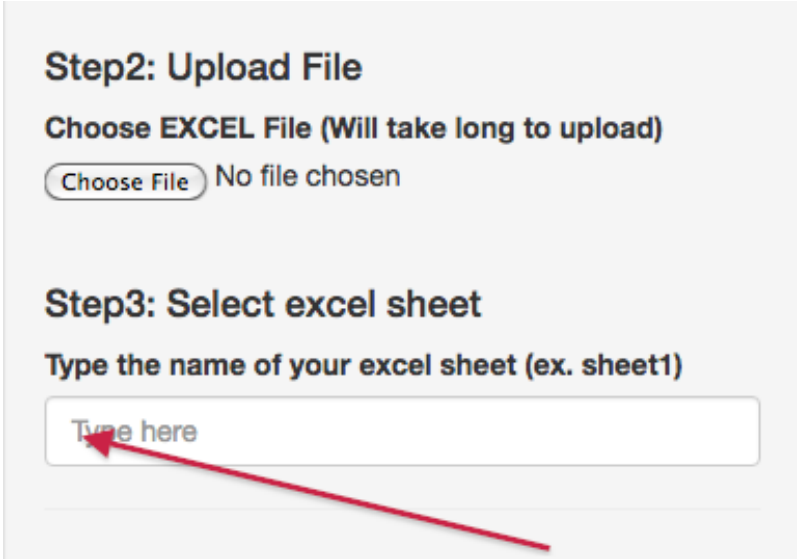


2. Upload **Excel** file

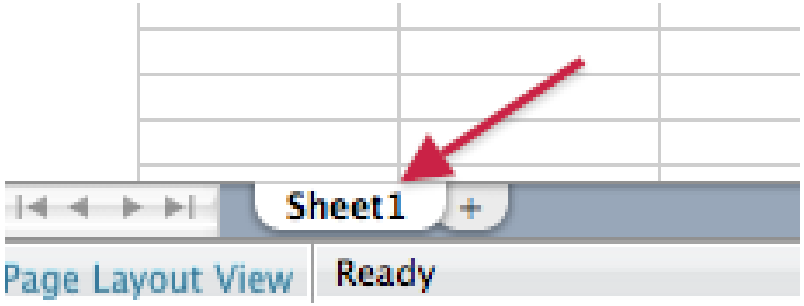


Excel Format

1. Slow processing

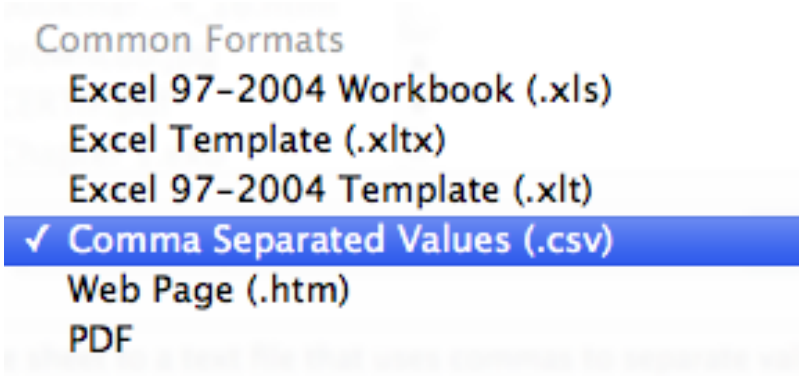


2. Requires the name of your excel sheet



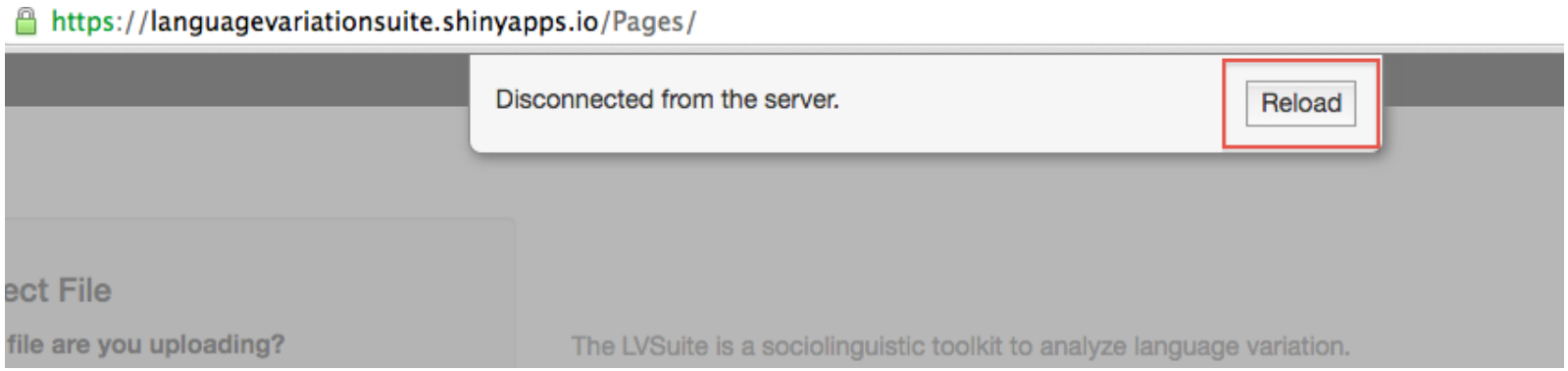
Save Excel as CSV Format

To optimize speed - **Save as CSV** prior upload



Server

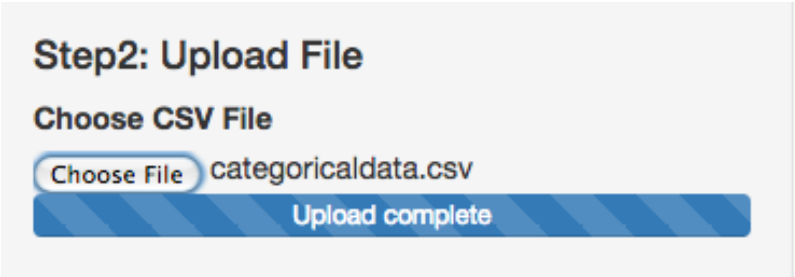
Since LVS is hosted on a server, Shiny idle time-out settings may stop application when it is left inactive (it will grey out).



Solution: Click **reload** and re-upload your csv file

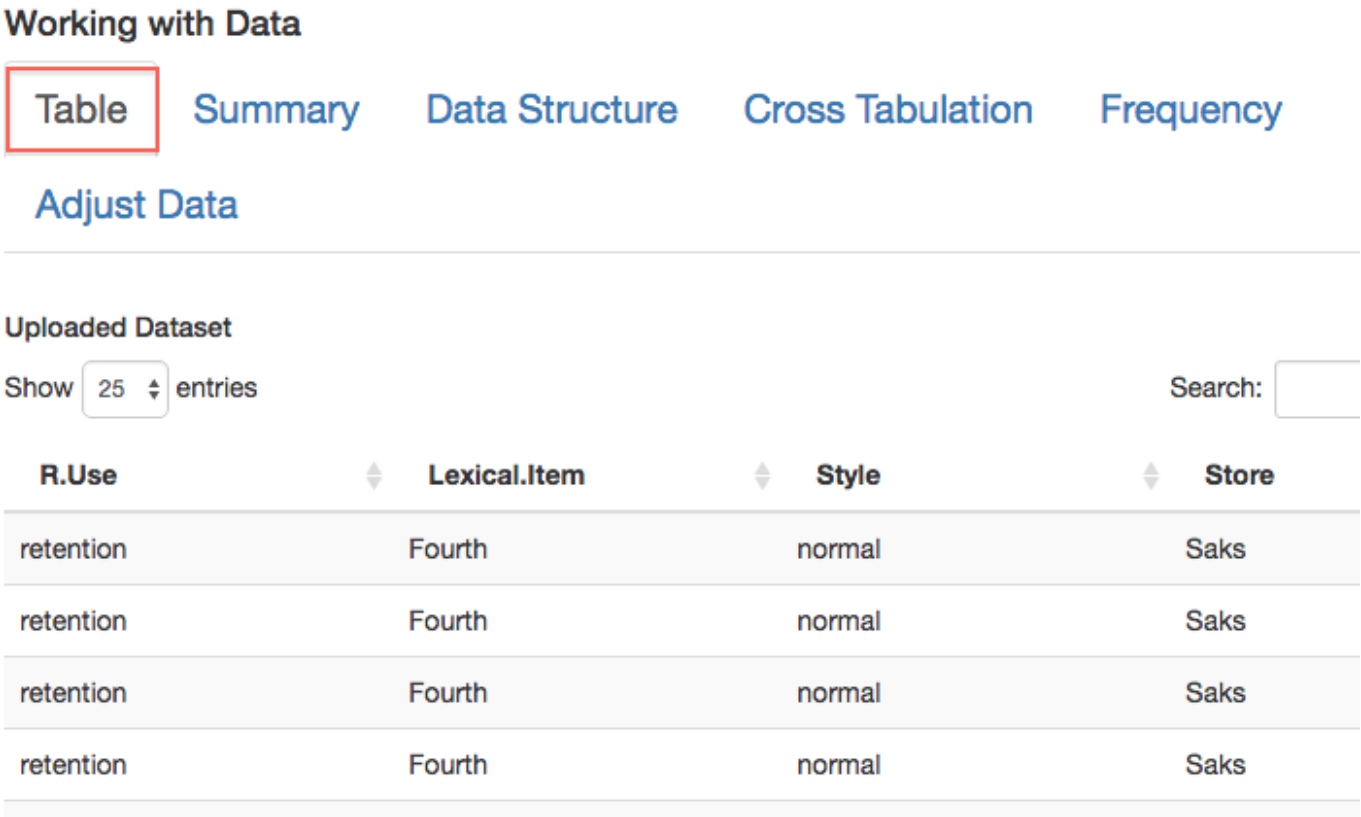
Upload File

Upload **categoricaldata.csv**



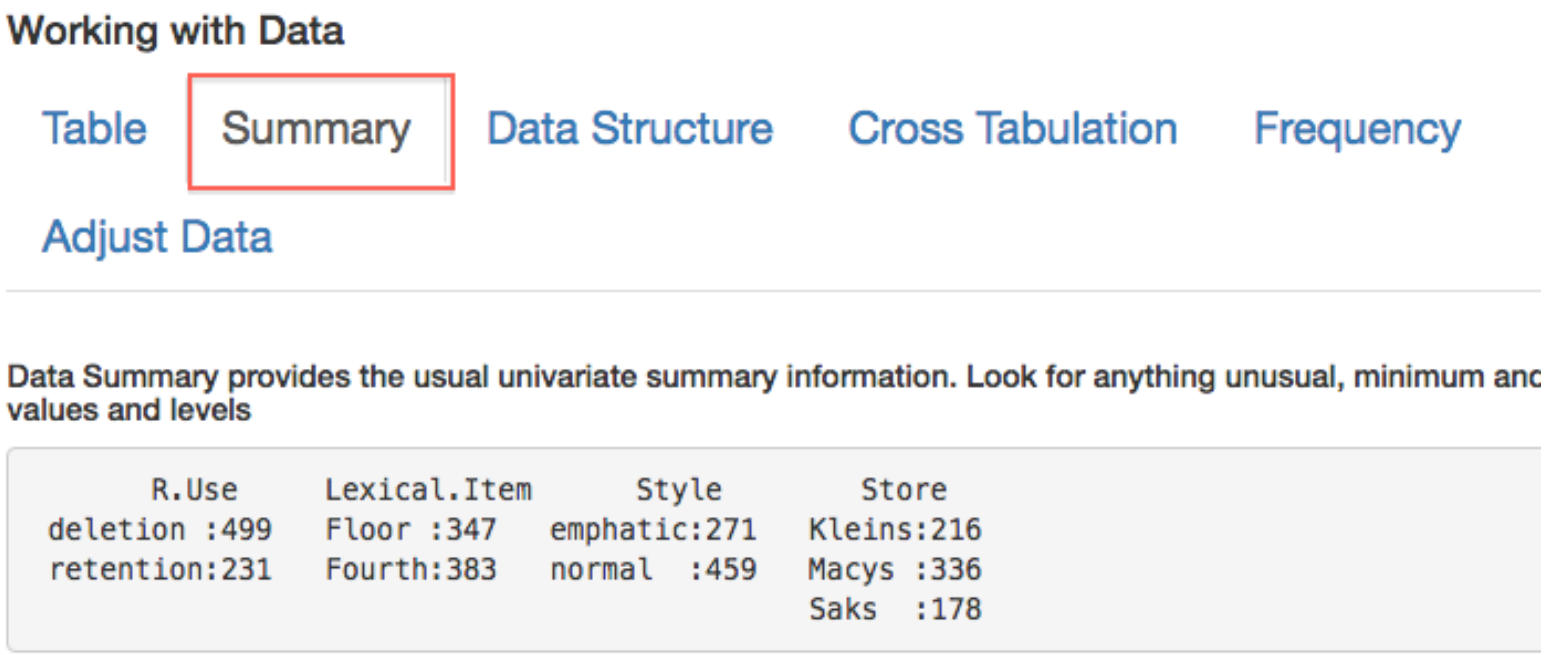
Table

Table displays our dataset and allows for sorting columns in descending/ascending order.



Summary

Summary provides a quantitative summary for each variable, e.g. frequency count, mean, median.



Data Structure

Table | Summary | Data Structure | Cross Tabulation | Frequency

Adjust Data

Variables names and types: int/num - numeric data, Factors - categorical data

```
'data.frame':   730 obs. of  4 variables:
 $ R.Use       : Factor w/ 2 levels "deletion","retention": 2 2 2 2 2 2 2 2 2 2 ...
 $ Lexical.Item: Factor w/ 2 levels "Floor","Fourth": 2 2 2 2 2 2 2 2 2 2 ...
 $ Style       : Factor w/ 2 levels "emphatic","normal": 2 2 2 2 2 2 2 2 2 2 ...
 $ Store       : Factor w/ 3 levels "Kleins","Macys",..: 3 3 3 3 3 3 3 3 3 3 ...
```

Data Structure

Table | Summary | Data Structure | Cross Tabulation | Frequency

Adjust Data

Variables names and types: int/num - numeric data, Factors - categorical data

```
'data.frame':   730 obs. of  4 variables:
 $ R.Use       : Factor w/ 2 levels "deletion","retention": 2 2 2 2 2 2 2 2 2 2 ...
 $ Lexical.Item: Factor w/ 2 levels "Floor","Fourth": 2 2 2 2 2 2 2 2 2 2 ...
 $ Style       : Factor w/ 2 levels "emphatic","normal": 2 2 2 2 2 2 2 2 2 2 ...
 $ Store       : Factor w/ 3 levels "Kleins","Macys",..: 3 3 3 3 3 3 3 3 3 3 ...
```

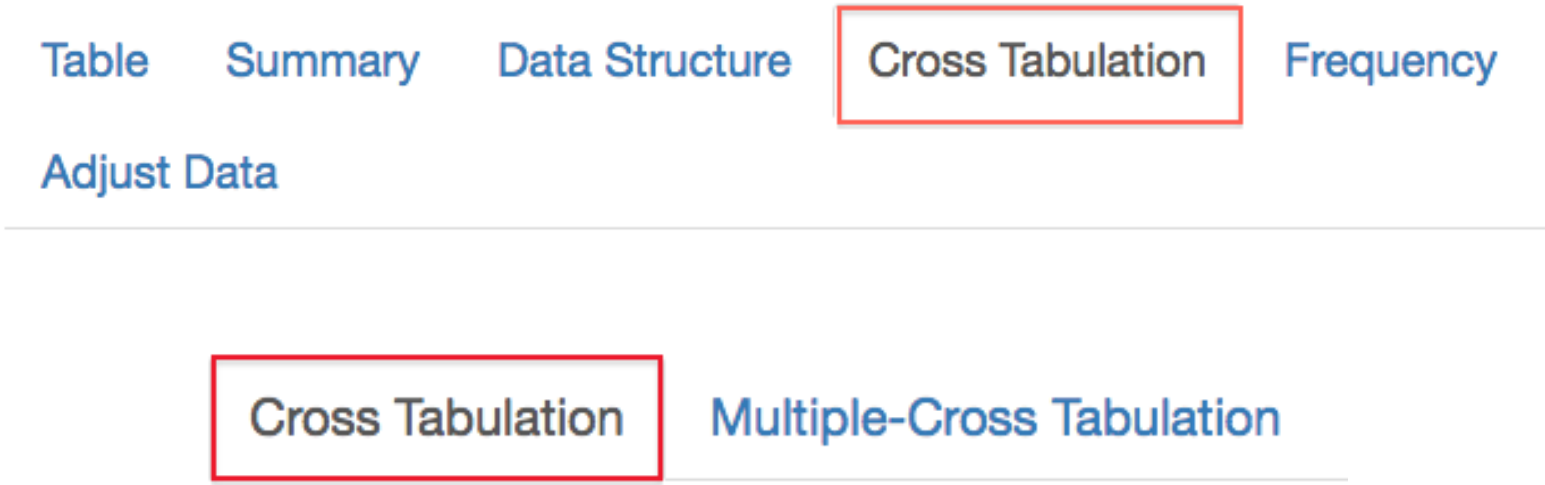
1. Total number of **observations**
2. Number of **variables**
3. Variable **types**
   - **Factor** - categorical values
   - **Num** - numeric values (0.95, 1.05)
   - **Int** - integer values (1, 2, 3)

Cross-Tabulation

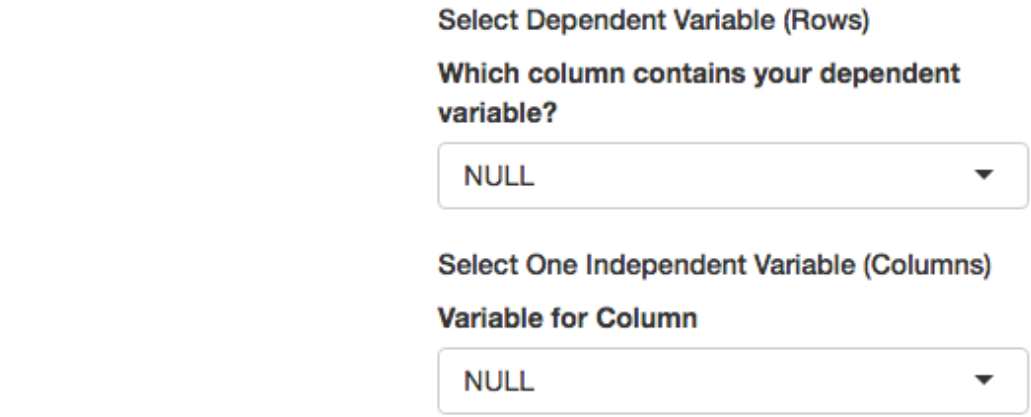
Cross-tabulation examines the relationship between two variables (their interaction).

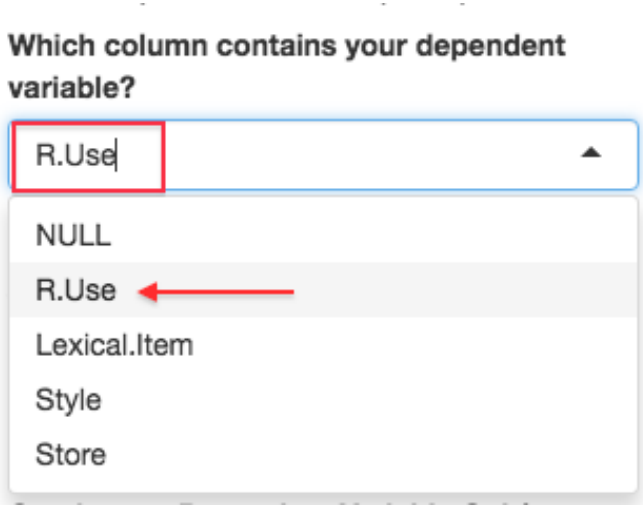
Cross-Tabulation

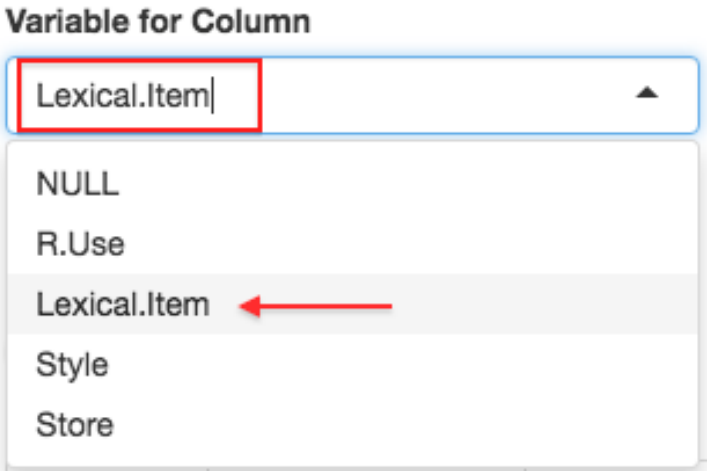
Cross-tabulation examines the relationship between two variables (their interaction).



Cross-Tabulation: One Dependent and One Independent Variables



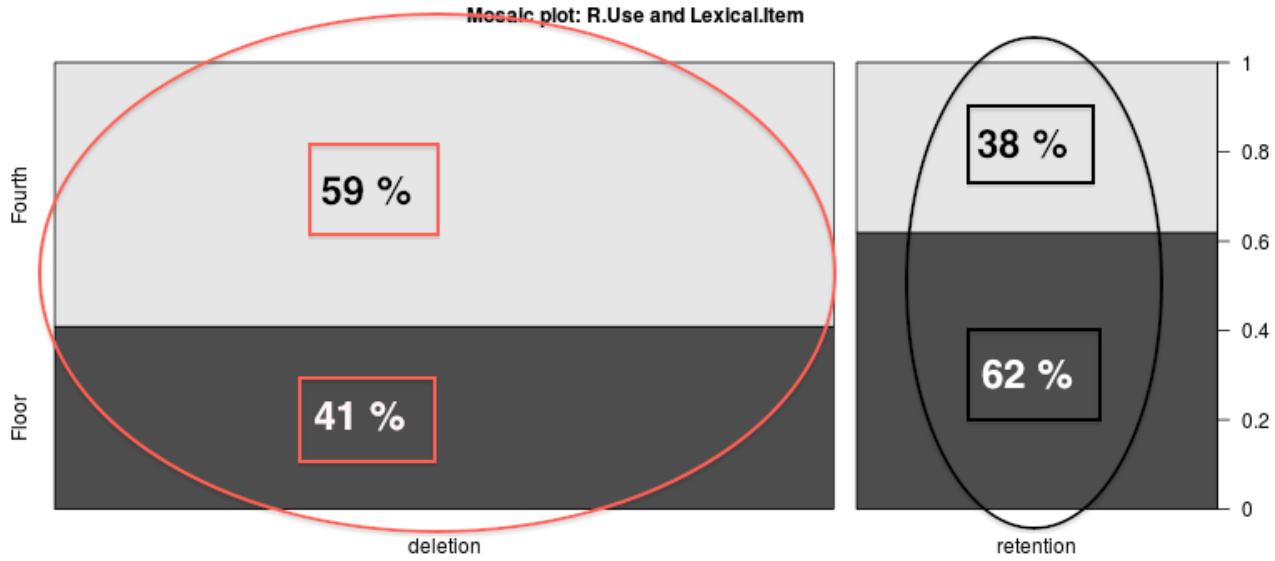




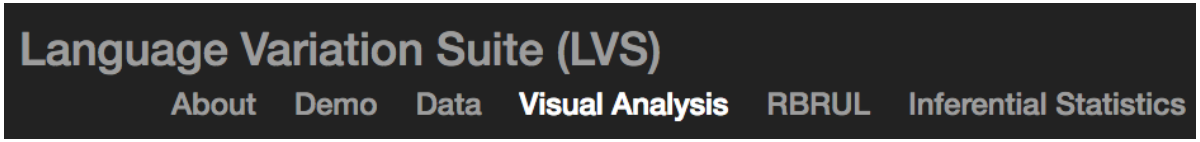
Cross-Tabulation Output

Raw frequency / Proportion by column / Proportion across row

	Floor/Col%/Row%	Fourth/Col%/Row%	RowSum
deletion	204/59/41	295/77/59	499
retention	143/41/62	88/23/38	231
ColumnSum	347	383	730



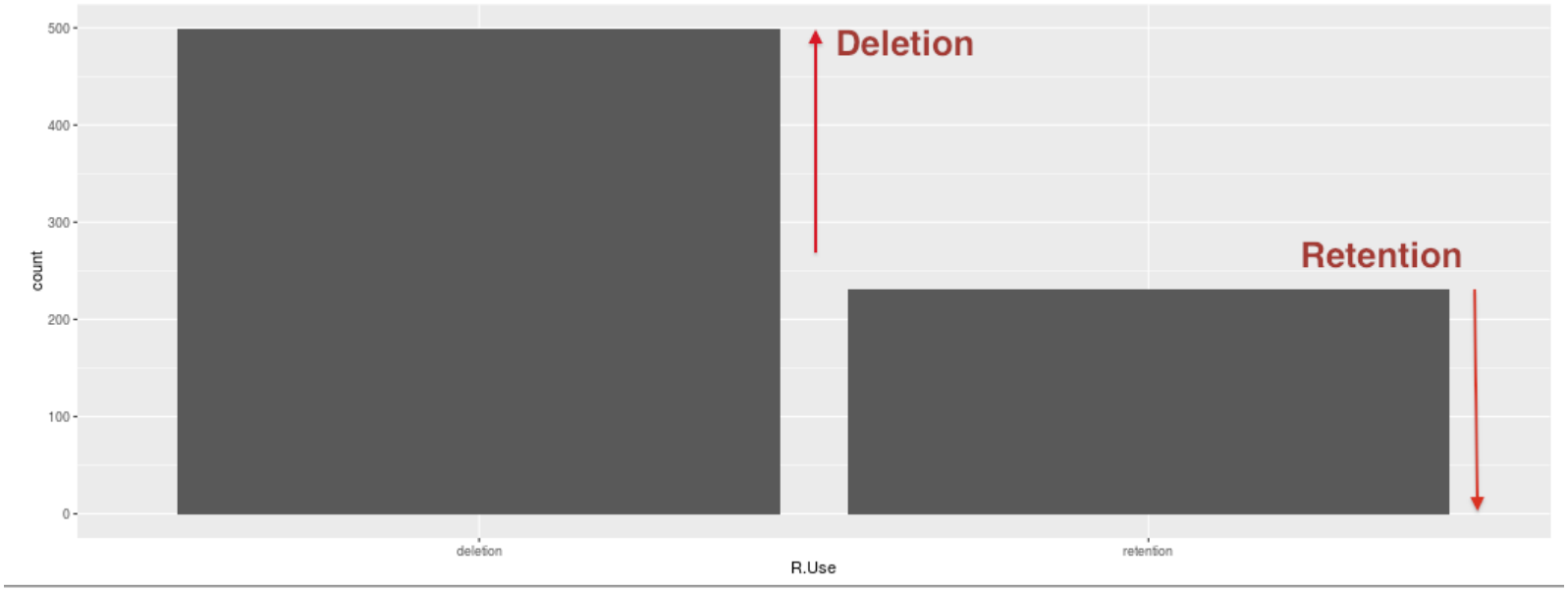
Language Variation Suite - Structure



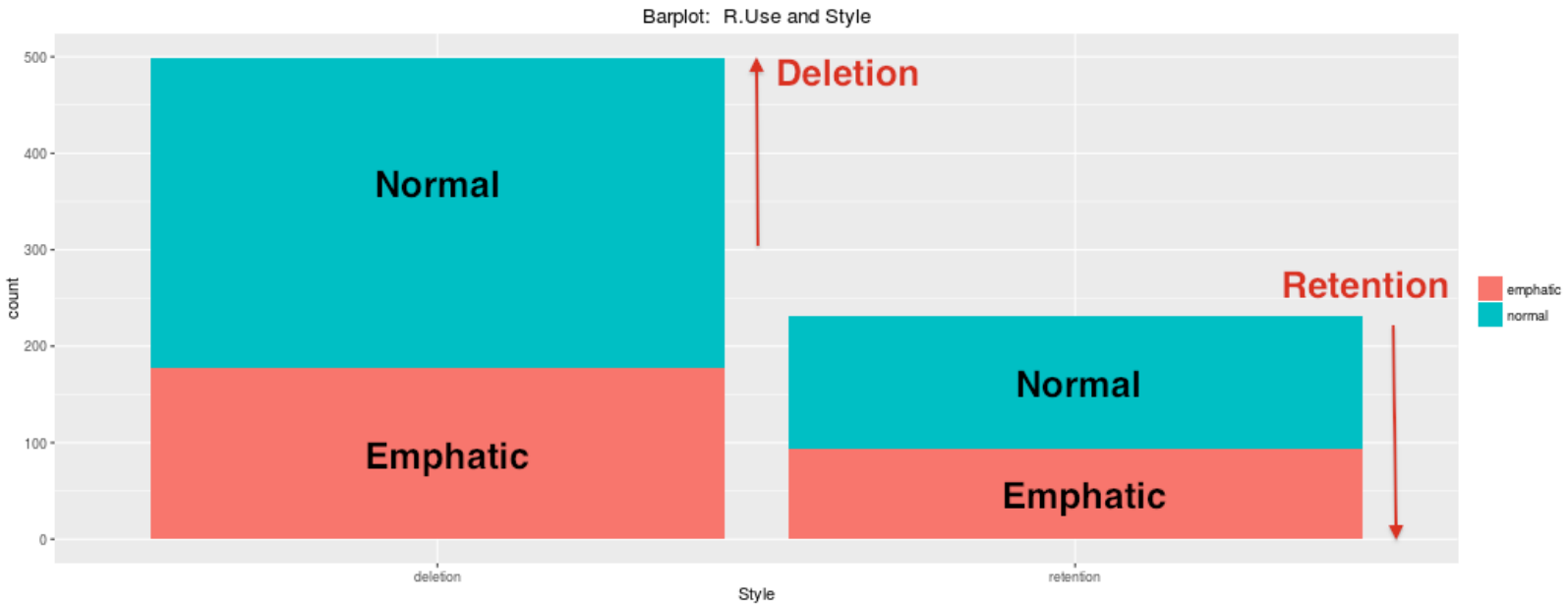
1. Demo
   - Brief introduction
2. Data
   - Upload file, data summary, adjust data, cross tabulation
3. Visual Analysis
   - Plotting, cluster classification
4. RBRUL
   - New version by Daniel Johnson!
5. Inferential statistics
   - Modeling, regression, varbrul analysis, conditional trees, random forest

Visual Analytics: "The science of analytical reasoning facilitated by visual interactive interfaces" (Thomas et al. 2005)

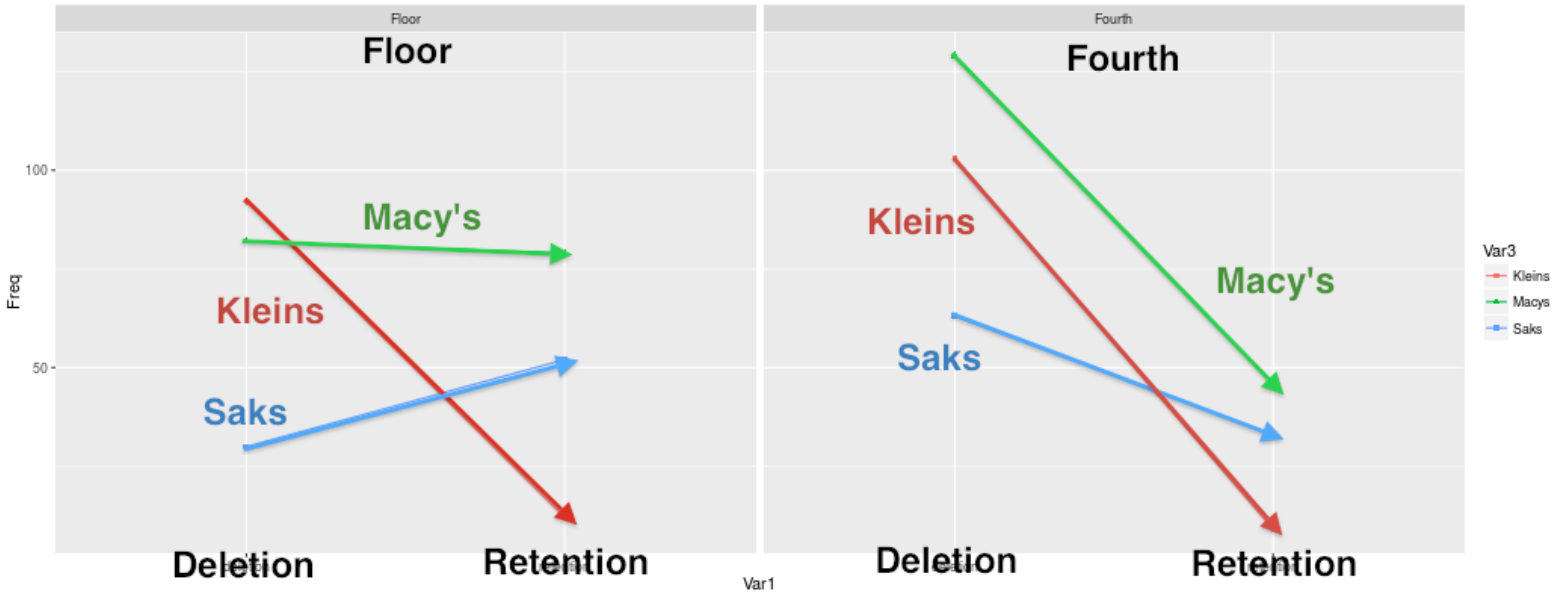
One Variable Plot



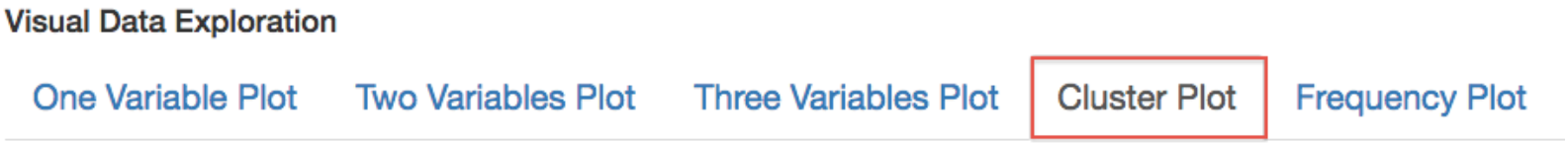
Two Variables Plot



Three Variables Plot

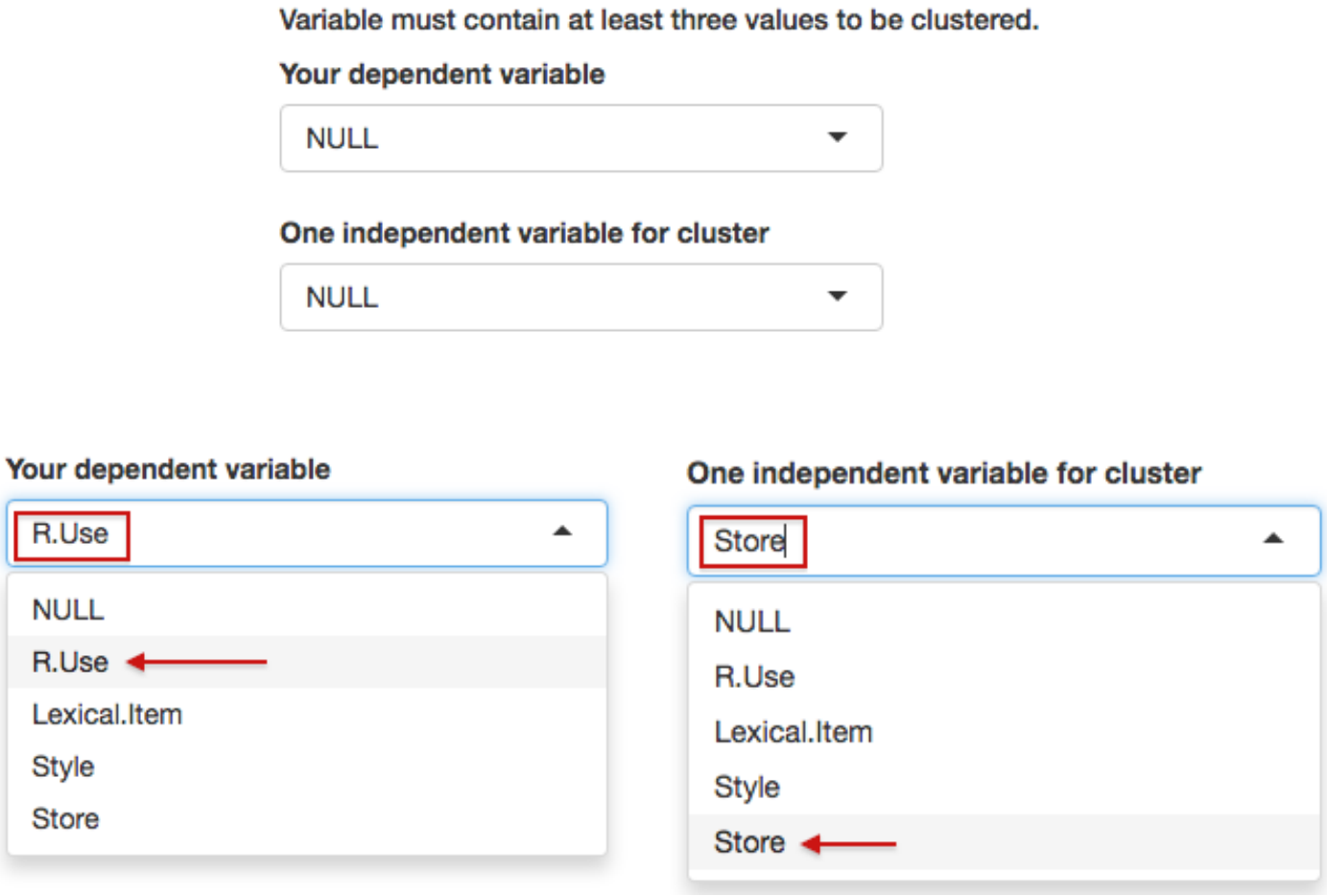


Cluster Plot

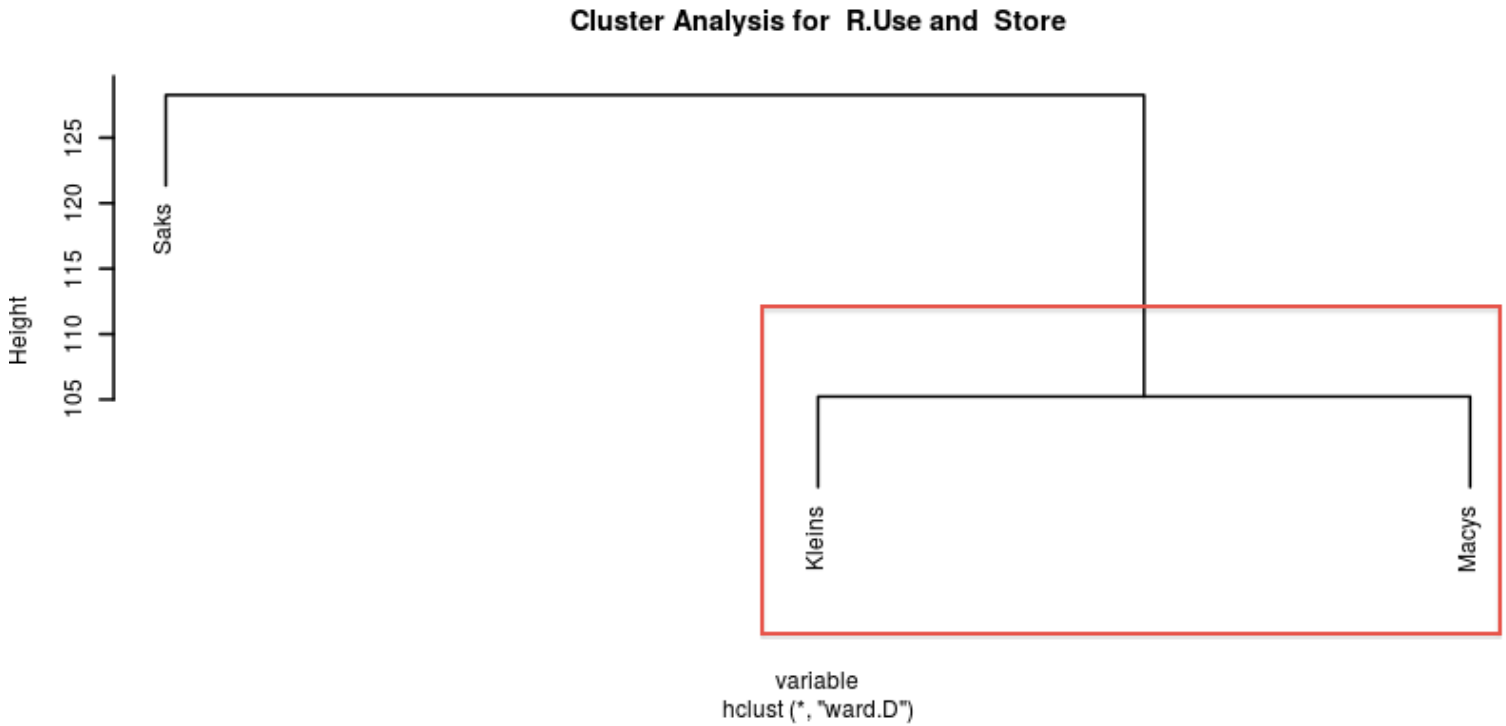


- Classification of data into **sub-groups** is based on **pairwise similarities**
- Groups are clustered in the form of a **tree-like dendrogram**

Cluster Plot



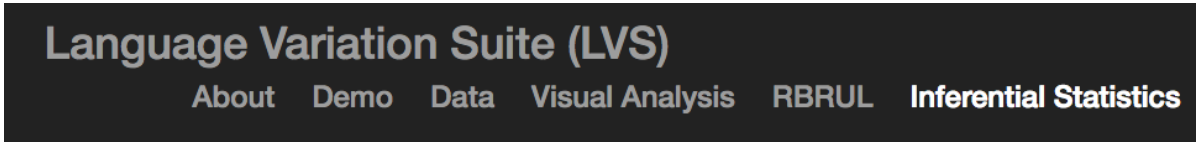
Cluster Plot



Saks (upper middle-class store), **Macy's** (middle-class store), **Kleins** (working-class)

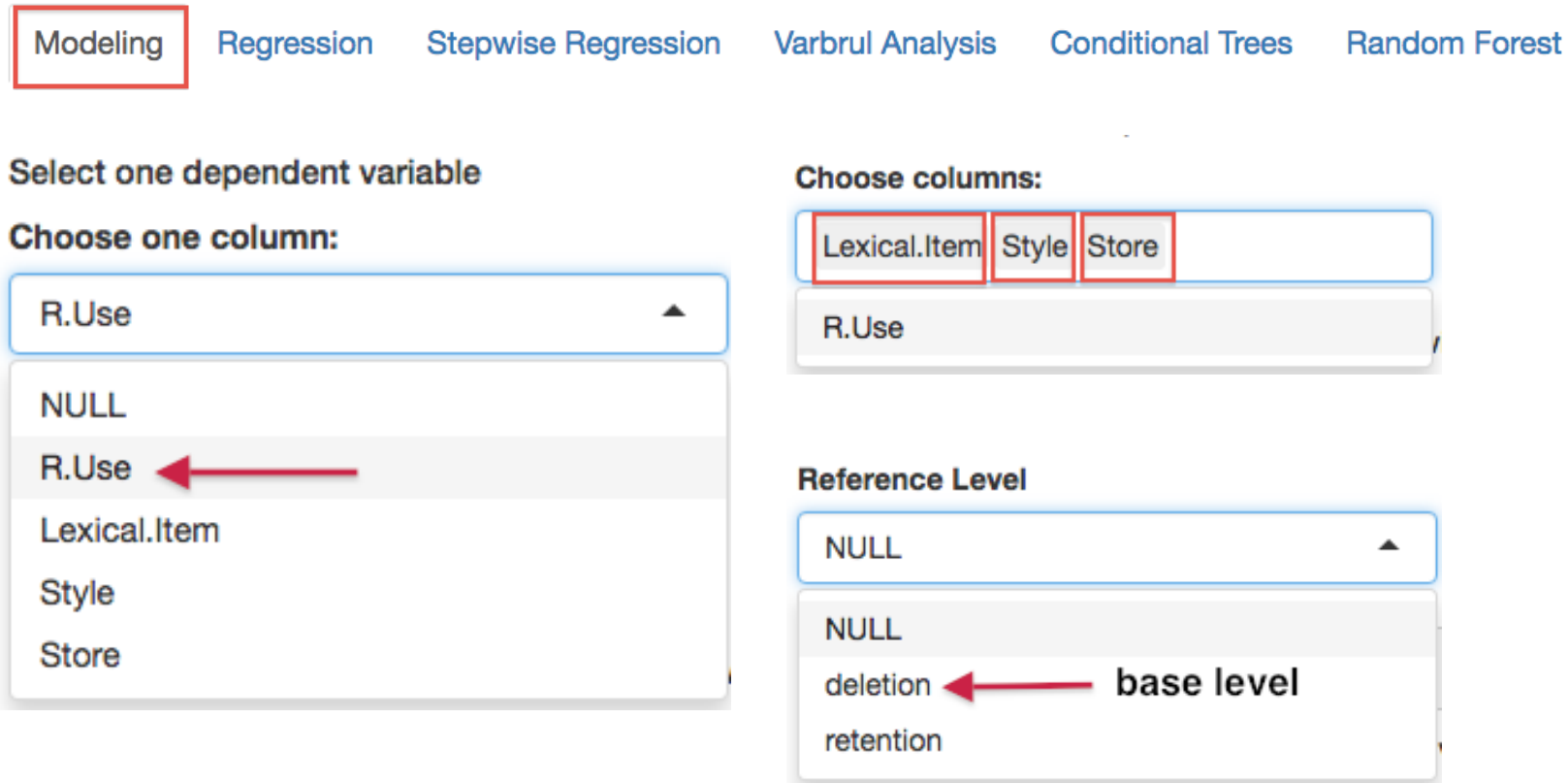
Inferential Statistics

Language Variation Suite - Structure

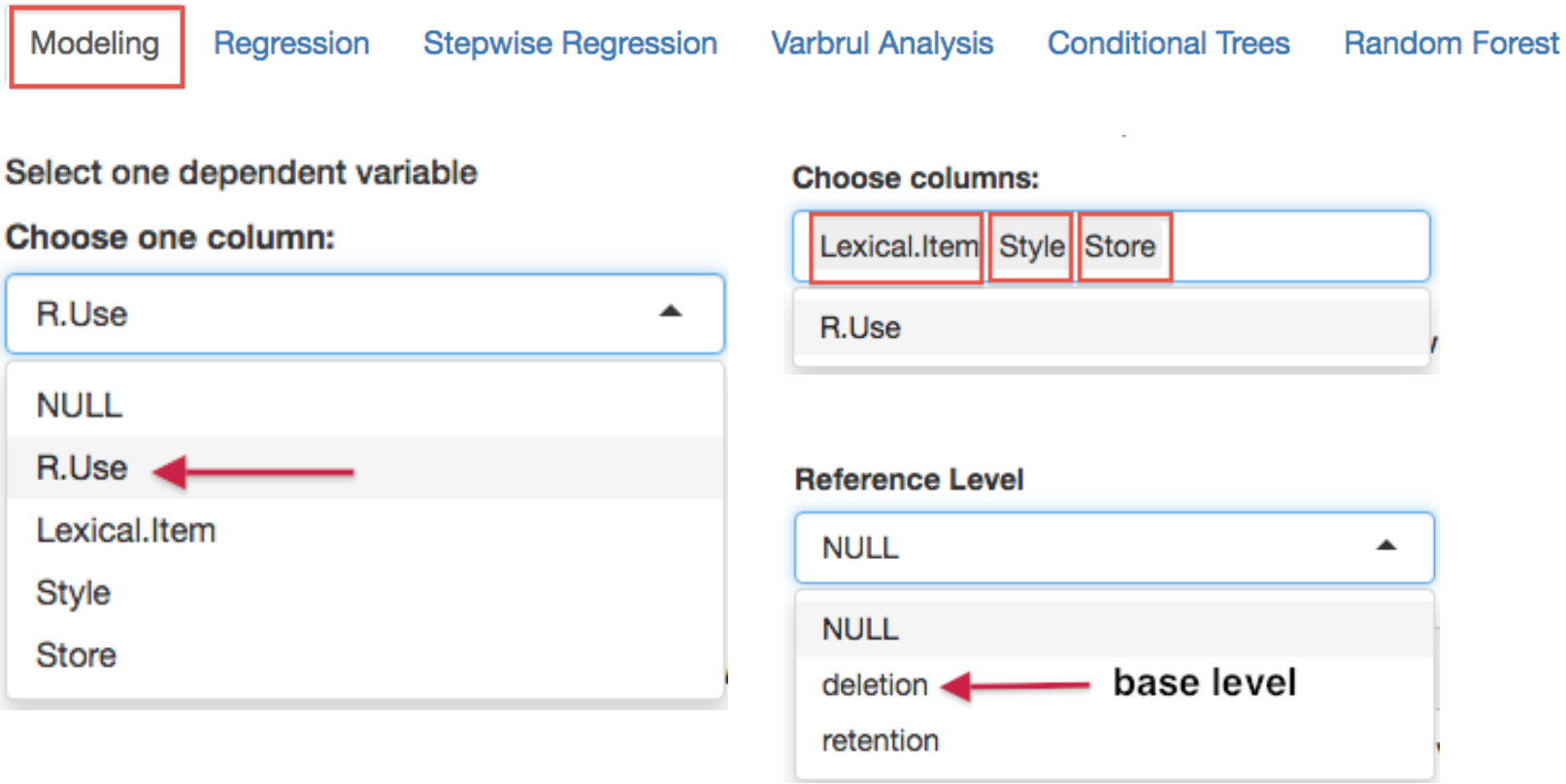


1. Demo
   - Brief introduction
2. Data
   - Upload file, data summary, adjust data, cross tabulation
3. Visual Analysis
   - Plotting, cluster classification
4. RBRUL
   - New version by Daniel Johnson!
5. Inferential statistics
   - Modeling, regression, varbrul analysis, conditional trees, random forest

Modeling


Modeling
Regression
Stepwise Regression
Varbrul Analysis
Conditional Trees
Random Forest
Select one dependent variable
Choose one column:
R.Use
NULL
R.Use
Lexical.Item
Style
Store
Choose columns:
Lexical.Item
Style
Store
R.Use
Reference Level
NULL
NULL
deletion
base level
retention

Modeling

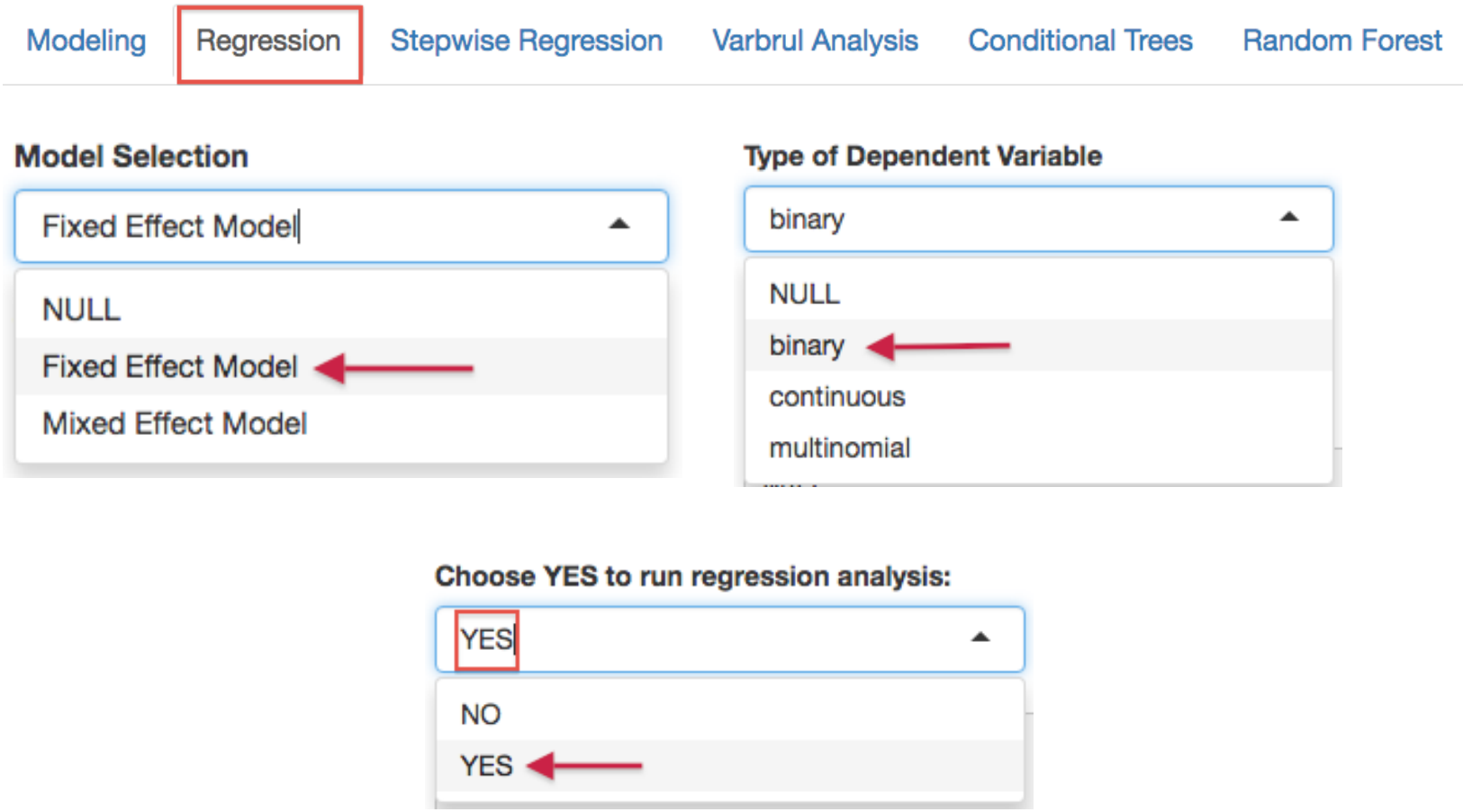


We are interested in RETENTION = Application

Regression Types

- **Model**
  a.) Fixed effect
  b.) Mixed effect - individual speaker/token variation (within group)
- **Type of Dependent Variable**
  a.) Binary/categorical (only two values)
  b.) Continuous (numeric)
  c.) Multinomial - categorical with more than two values

Regression


Modeling
Regression
Stepwise Regression
Varbrul Analysis
Conditional Trees
Random Forest
Model Selection
Fixed Effect Model
NULL
Fixed Effect Model
Mixed Effect Model
Type of Dependent Variable
binary
NULL
binary
continuous
multinomial
Choose YES to run regression analysis:
YES
NO
YES

Model Output


```
Call:
glm(formula = as.formula(paste(y, paste(listfactors, collapse = "+"),
    sep = "~")), family = binomial, data = plotData(), na.action = na.omit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4534  -0.8549  -0.5164   1.0493   2.4455

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.6276     0.2596  -6.269 3.64e-10 ***
Lexical.ItemFourth -0.9912     0.1749  -5.666 1.46e-08 ***
Stylenormal        -0.3197     0.1787  -1.789   0.0736 .
StoreMacys          1.8004     0.2615   6.884 5.81e-12 ***
StoreSaks           2.2564     0.2817   8.011 1.13e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 911.27  on 729  degrees of freedom
Residual deviance: 791.82  on 725  degrees of freedom
AIC: 801.82

Number of Fisher Scoring iterations: 5
```

Model Output

```
Call:
glm(formula = as.formula(paste(y, paste(listfactors, collapse = "+"),
    sep = "~")), family = binomial, data = plotData(), na.action = na.omit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4534  -0.8549  -0.5164   1.0493   2.4455

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.6276     0.2596  -6.269 3.64e-10 ***
Lexical.ItemFourth -0.9912     0.1749  -5.666 1.46e-08 ***
Stylenormal        -0.3197     0.1787  -1.789   0.0736 .
StoreMacys          1.8004     0.2615   6.884 5.81e-12 ***
StoreSaks           2.2564     0.2817   8.011 1.13e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 911.27  on 729  degrees of freedom
Residual deviance: 791.82  on 725  degrees of freedom
AIC: 801.82

Number of Fisher Scoring iterations: 5
```

Interpretation

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.6276     0.2596  -6.269 3.64e-10 ***
Lexical.ItemFourth -0.9912     0.1749  -5.666 1.46e-08 ***
Stylenormal        -0.3197     0.1787  -1.789   0.0736 .
StoreMacys          1.8004     0.2615   6.884 5.81e-12 ***
StoreSaks           2.2564     0.2817   8.011 1.13e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Lexical item **Fourth** has a negative effect on **retention** and is significant
- **Normal** style has a slightly negative effect on **retention** but its coefficient is not significant
- **Macy's** and **Saks** have a positive and significant effect on **retention**. Saks (upper middle class store) is more significant than Macy's (middle class store)

http://www.free-online-calculator-use.com/scientific-notation-converter.html

Interpretation


```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.6276     0.2596  -6.269 3.64e-10 ***
Lexical.ItemFourth -0.9912     0.1749  -5.666 1.46e-08 ***
Stylenormal       -0.3197     0.1787  -1.789   0.0736 .
StoreMacys         1.8004     0.2615   6.884 5.81e-12 ***
StoreSaks          2.2564     0.2817   8.011 1.13e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

exponential notation:

1.48e-8

.0000000146

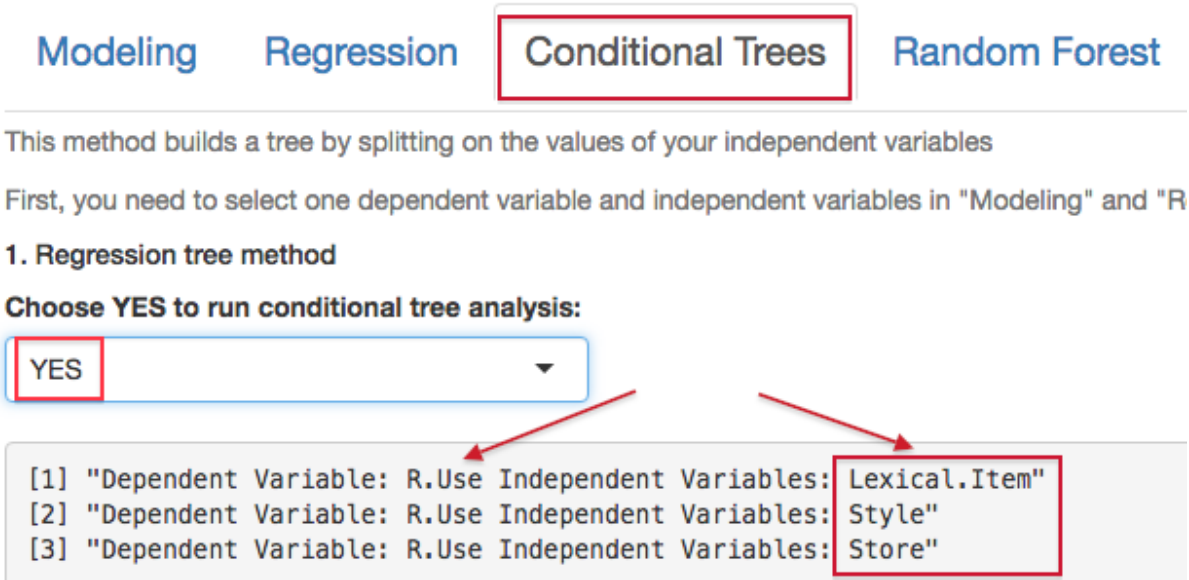
87654321

0.0000000148

- Lexical item **Fourth** has a negative effect on **retention** and is significant
- **Normal** style has a slightly negative effect on **retention** but its coefficient is not significant
- **Macy's** and **Saks** have a positive and significant effect on **retention**. Saks (upper middle class store) is more significant than Macy's (middle class store)

http://www.free-online-calculator-use.com/scientific-notation-converter.html

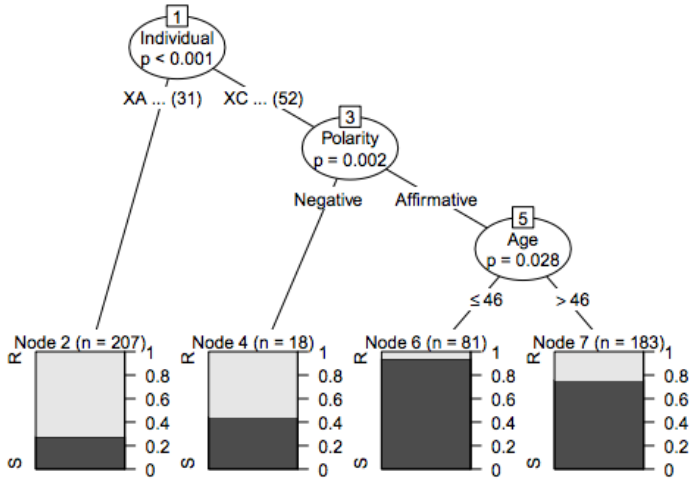
Conditional Tree



Conditional Tree

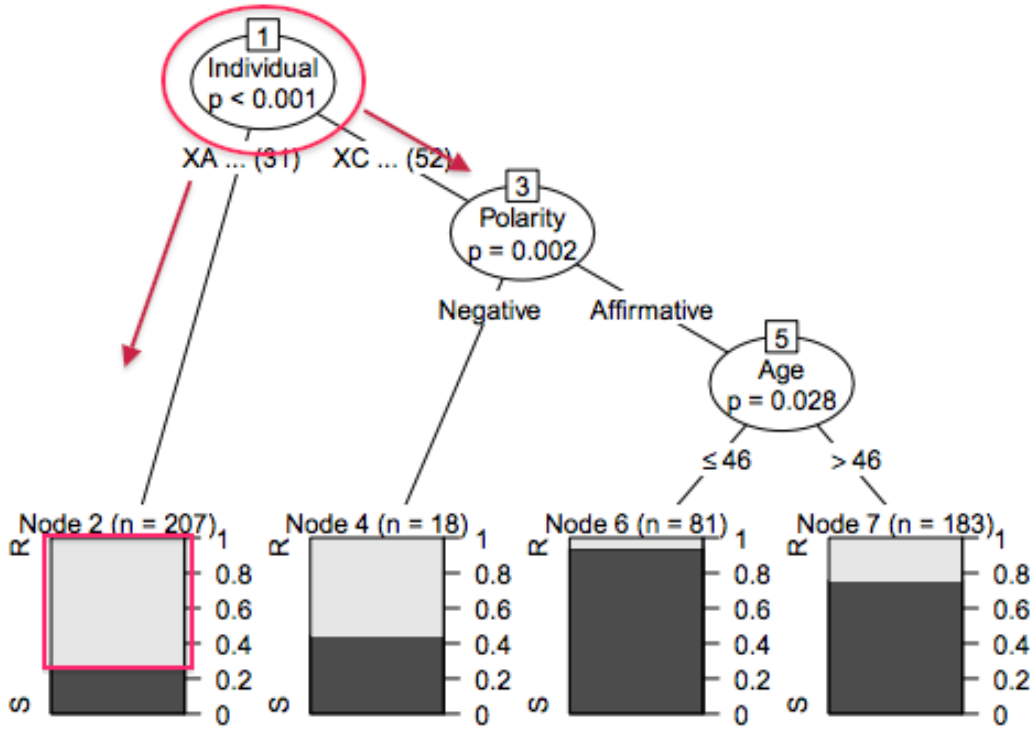
Conditional tree: a simple non-parametric regression analysis, commonly used in social and psychological studies

- Linear regression: all information is combined linearly
- Conditional tree regression: visual splitting to capture interaction between variables



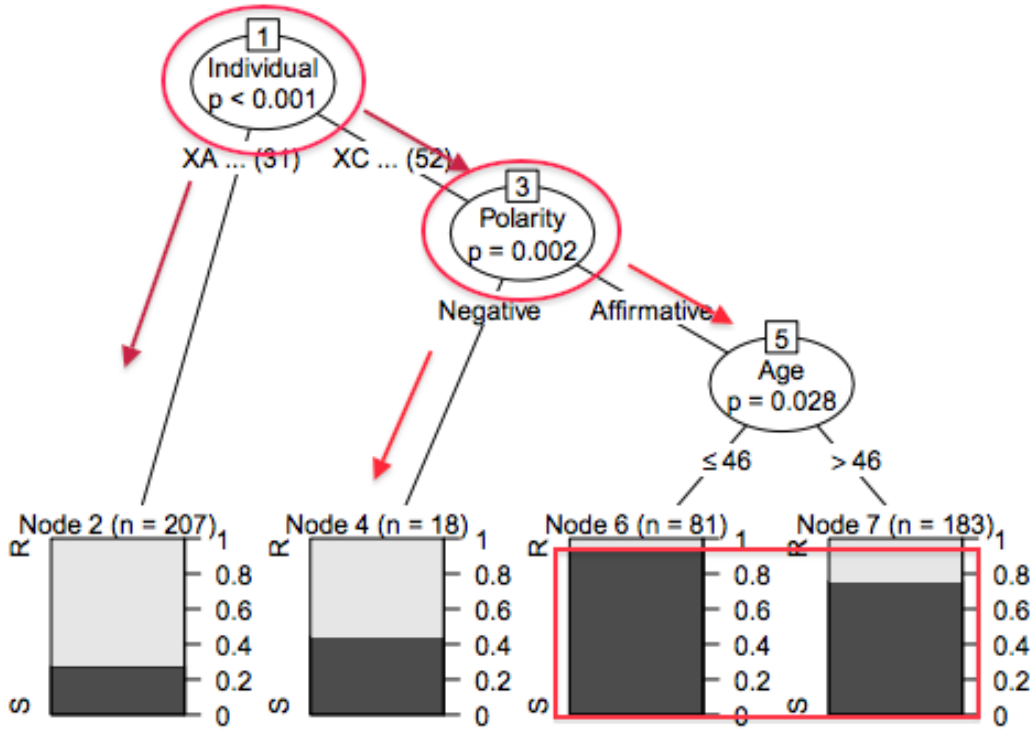
Recursive splitting (tree branches)

Conditional Tree - Tagliamonte and Baayen 2012



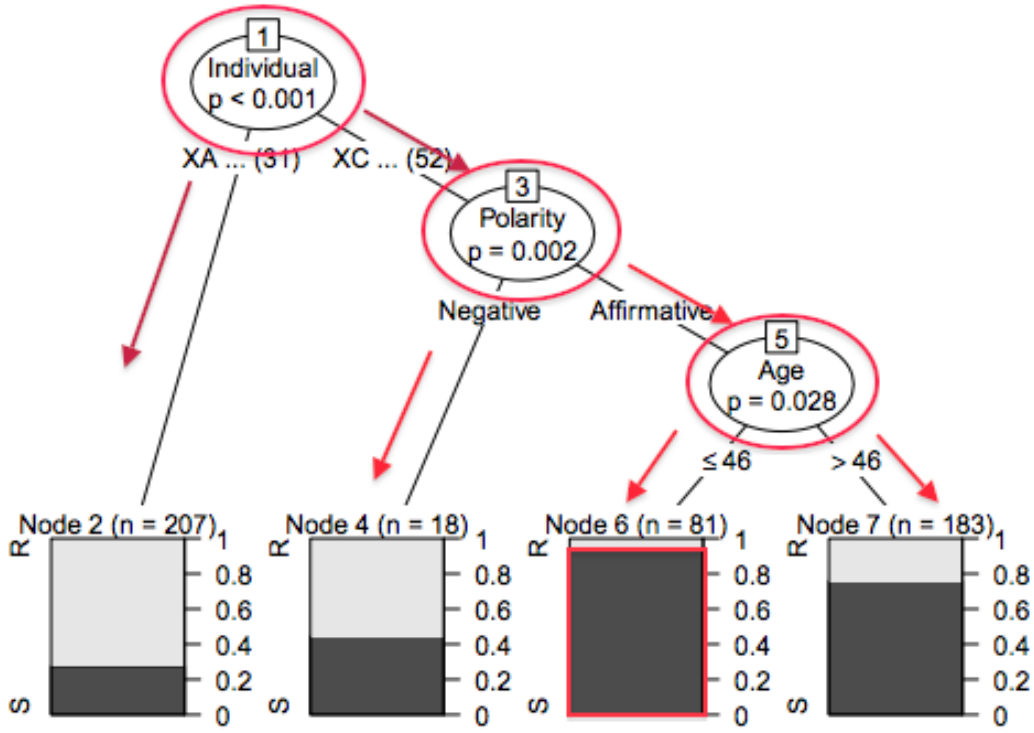
1. The distribution of **was/were** is split in two groups by individuals.
2. The variant **were** occurs significantly more frequently with the first group.

Conditional Tree - Tagliamonte and Baayen (2012)



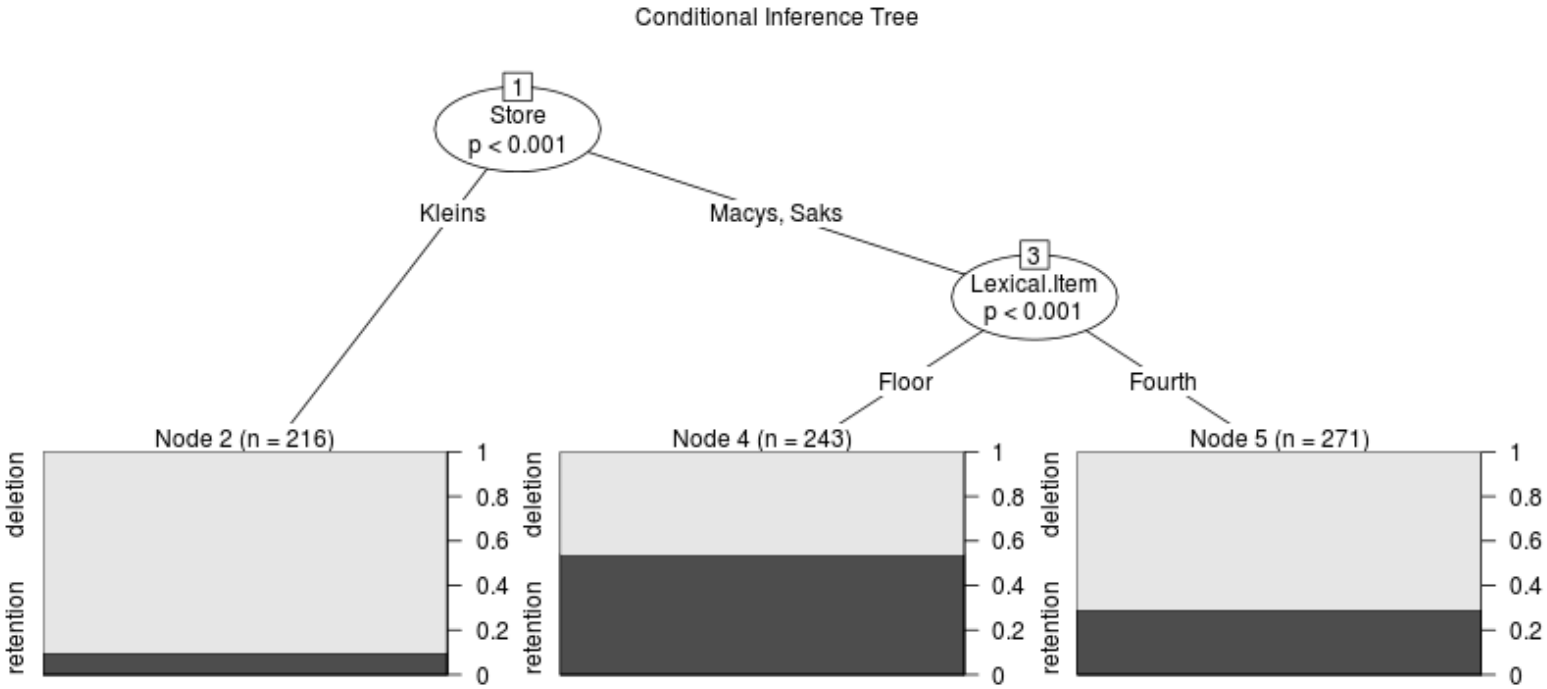
1. **Polarity** is relevant to the second group of individuals.
2. The variant **were** occurs significantly more often with **negative** polarity

Conditional Tree - Tagliamonte and Baayen (2012)



1. **Affirmative Polarity** is conditioned by **Age**.
2. The variant **was** is produced significantly more often by Individuals of 46 and younger.

Conditional Tree

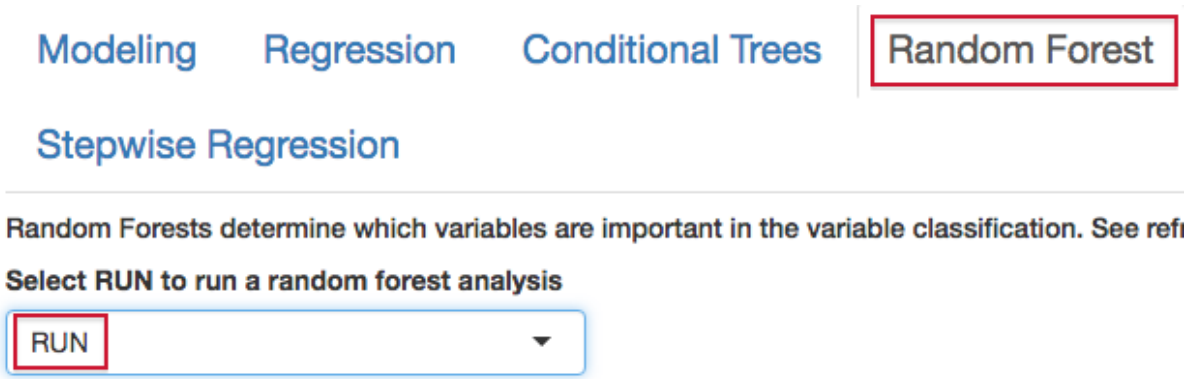


1. **Store** is the most significant factor for R-use
   - **Kleins** (working class store) - more R-deletion
2. R-use in Macy's and Saks is conditioned by **lexical item**:
   - **Floor** shows more R-retention than **Fourth**
3. **Style** is not significant

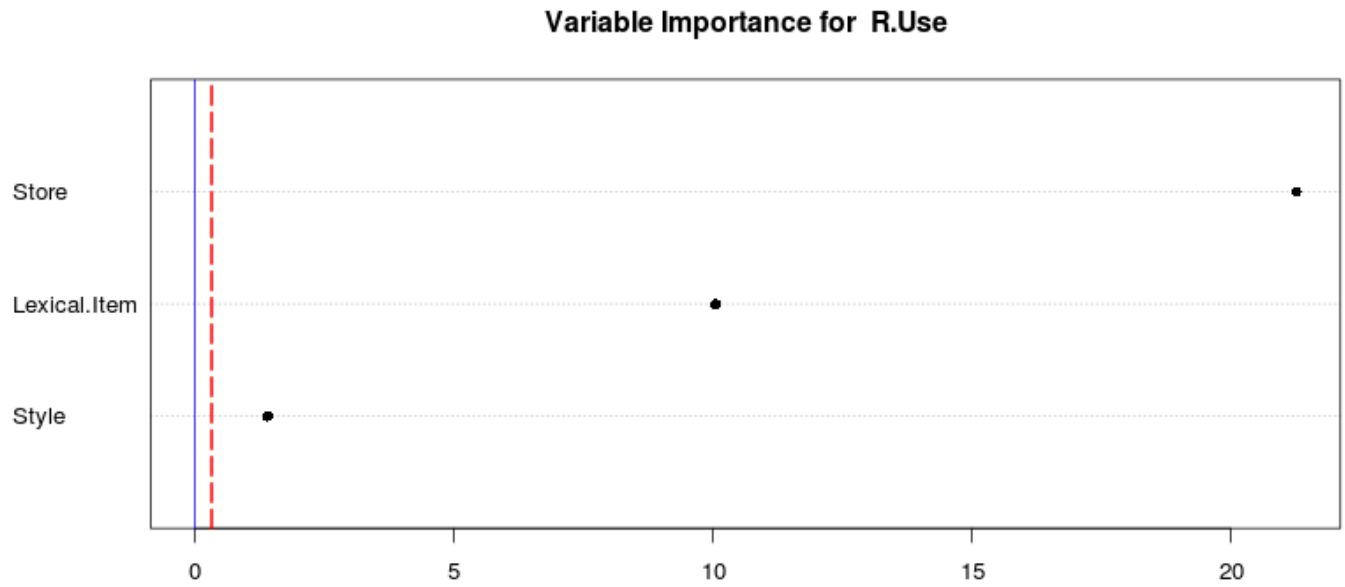
Random Forest

1. Variable importance for predictors
2. Robust technique with *small* **n** *large* **p** data
3. All predictors considered jointly (allows for inclusion of correlated factors)

Random Forest



Random Forest



1. **Store** is the most important predictor
2. **Lexical Item** is the second predictor
3. **Style** is irrelevant: close to zero and red dotted line (cut-off value).

Let's Have a Short Break

Fixed and Mixed Models

Fixed Effects Model : All predictors are treated independent.

Underlying assumption - no group-internal variation between speakers or tokens

Mixed Effects Model : Allows for evaluation of individual- and group-level variation

Fixed and Mixed Models

Fixed Regression Model - ignoring individual variations (speakers or words) may lead to Type I Error: "a chance effect is mistaken for a real difference between the populations"

Mixed Regression Model - prone to Type II Error: "if speaker variation is at a high level, we cannot discern small population effects without a large number of speakers" (Johnson 2009, 2015)

Mixed Effect Regression

Mixed Model = fixed effects + random effects

Fixed-effect factor - "repeatable and a small number of levels"

Random-effect factor - "a non-repeatable random sample from a larger population" (Wieling 2012)

- **walk, sleep, study, finish, eat, etc**
- **event verb, stative verb**
- **speaker1, speaker3, speaker3, etc**
- **male, female**

Mixed Effect Regression

Mixed Model = fixed effects + random effects

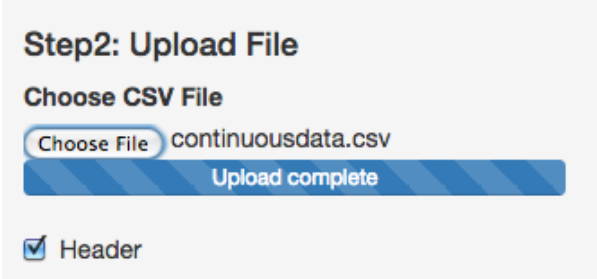
Fixed-effect factor - "repeatable and a small number of levels"

Random-effect factor - "a non-repeatable random sample from a larger population" (Wieling 2012)

- **walk, sleep, study, finish, eat, etc**
- **event verb, stative verb**
- **speaker1, speaker3, speaker3, etc**
- **male, female**

Preparing for Mixed Model

1. Download **continuousdata.csv**
2. Upload this file on LVS



Table

Table | Summary | Data Structure | Cross Tabulation | Frequency

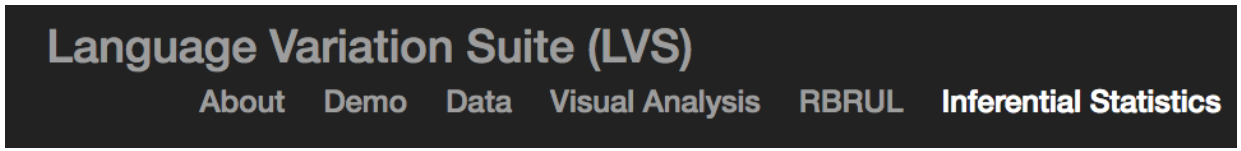
Uploaded Dataset

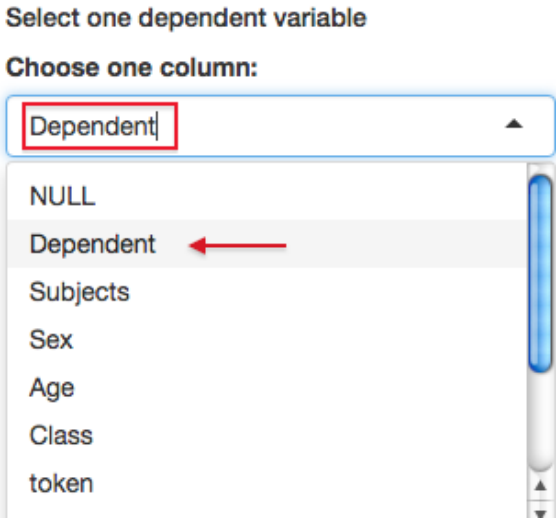
RANDOM

Show 25 entries Search:

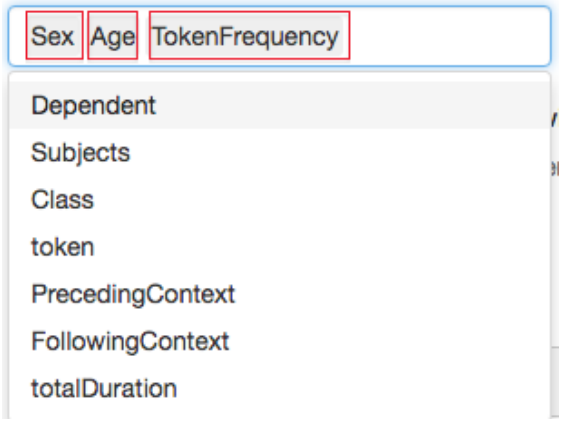
Dependent	Subjects	Sex	Age	Class	token	TokenFrequency
0.97	CA1HA	m	20-34	1	mudamos	45
0.98	CA1HA	m	20-34	1	edad	149
0.96	CA1HA	m	20-34	1	florida	20
0.95	CA1HA	m	20-34	1	edad	149
0.98	CA1HA	m	20-34	1	distanciados	2
0.98	CA1HA	m	20-34	1	cada	331

Mixed Effect Modeling



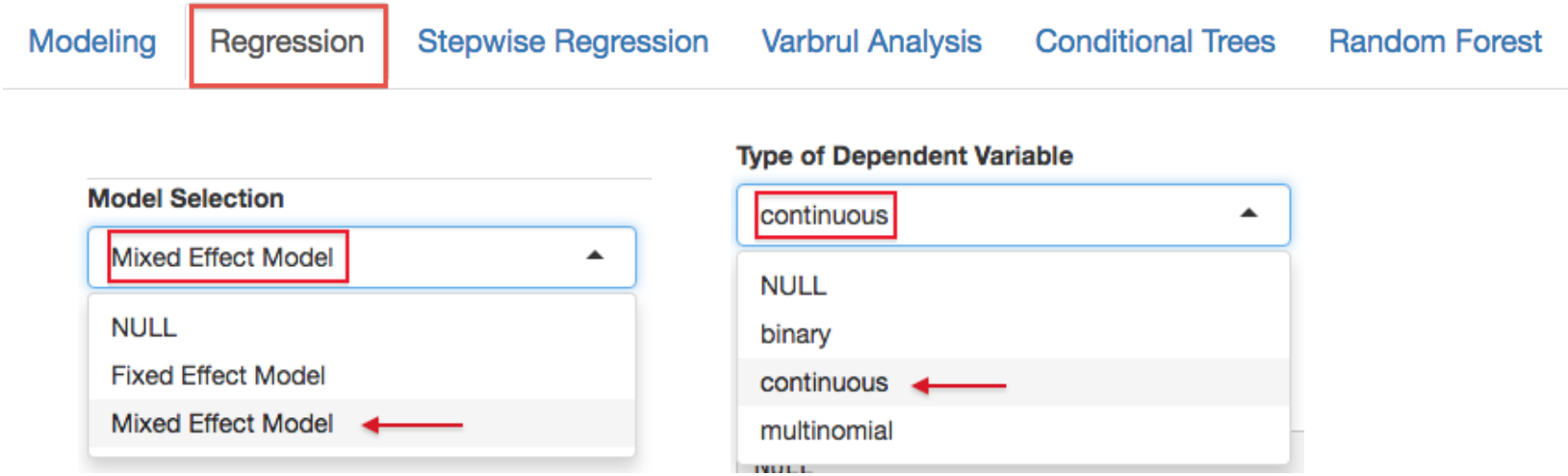


Fixed Effects - independent variables

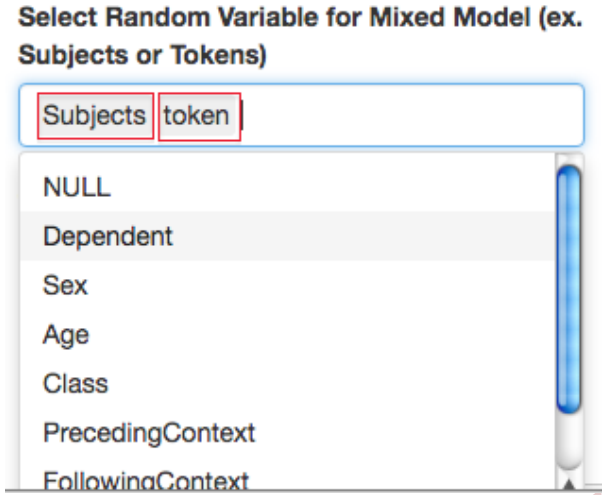


NULL when the dependent variable is continuous

Mixed Effect Modeling



Mixed Effects - group-internal variation



Regression Results


```
Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.7906 -0.4281  0.1440  0.6619  1.8390

Random effects:
 Groups   Name        Variance  Std.Dev.
 token    (Intercept) 7.436e-06 0.002727
 Subjects (Intercept) 1.455e-04 0.012064
 Residual             9.616e-04 0.031010
Number of obs: 517, groups:  token, 301; Subjects, 12

Fixed effects:
                 Estimate Std. Error         df t value Pr(>|t|)
(Intercept)     9.591e-01  7.495e-03  8.050e+00 127.964 1.31e-14 ***
Sexm            4.018e-03  7.490e-03  8.030e+00   0.537   0.6061
Age35-54        6.121e-04  9.167e-03  8.007e+00   0.067   0.9484
Age55+         -1.643e-02  9.172e-03  8.024e+00  -1.791   0.1110
TokenFrequency  1.082e-05  3.853e-06  6.046e+00   2.807   0.0306 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression Results


```
Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.7906 -0.4281  0.1440  0.6619  1.8390

Random effects:
 Groups   Name        Variance  Std.Dev.
 token    (Intercept) 7.436e-06 0.002727
 Subjects (Intercept) 1.455e-04 0.012064
 Residual             9.616e-04 0.031010
Number of obs: 517, groups:  token, 301; Subjects, 12

Fixed effects:
                 Estimate Std. Error         df t value Pr(>|t|)
(Intercept)     9.591e-01  7.495e-03  8.050e+00 127.964 1.31e-14 ***
Sexm            4.018e-03  7.490e-03  8.030e+00   0.537   0.6061
Age35-54        6.121e-04  9.167e-03  8.007e+00   0.067   0.9484
Age55+         -1.643e-02  9.172e-03  8.024e+00  -1.791   0.1110
TokenFrequency  1.082e-05  3.853e-06  6.046e+00   2.807   0.0306 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation - Random Effects


```
Random effects:
 Groups   Name        Variance  Std.Dev.
 token    (Intercept) 7.436e-06 0.002727
 Subjects (Intercept) 1.455e-04 0.012064
 Residual             9.616e-04 0.031010
Number of obs: 517, groups:  token, 301; Subjects, 12
```

1. **Standard Deviation**: a measure of the variability for each random effect (speakers and tokens)
2. **Residual**: random variation that is not due to speakers or tokens (residual error)

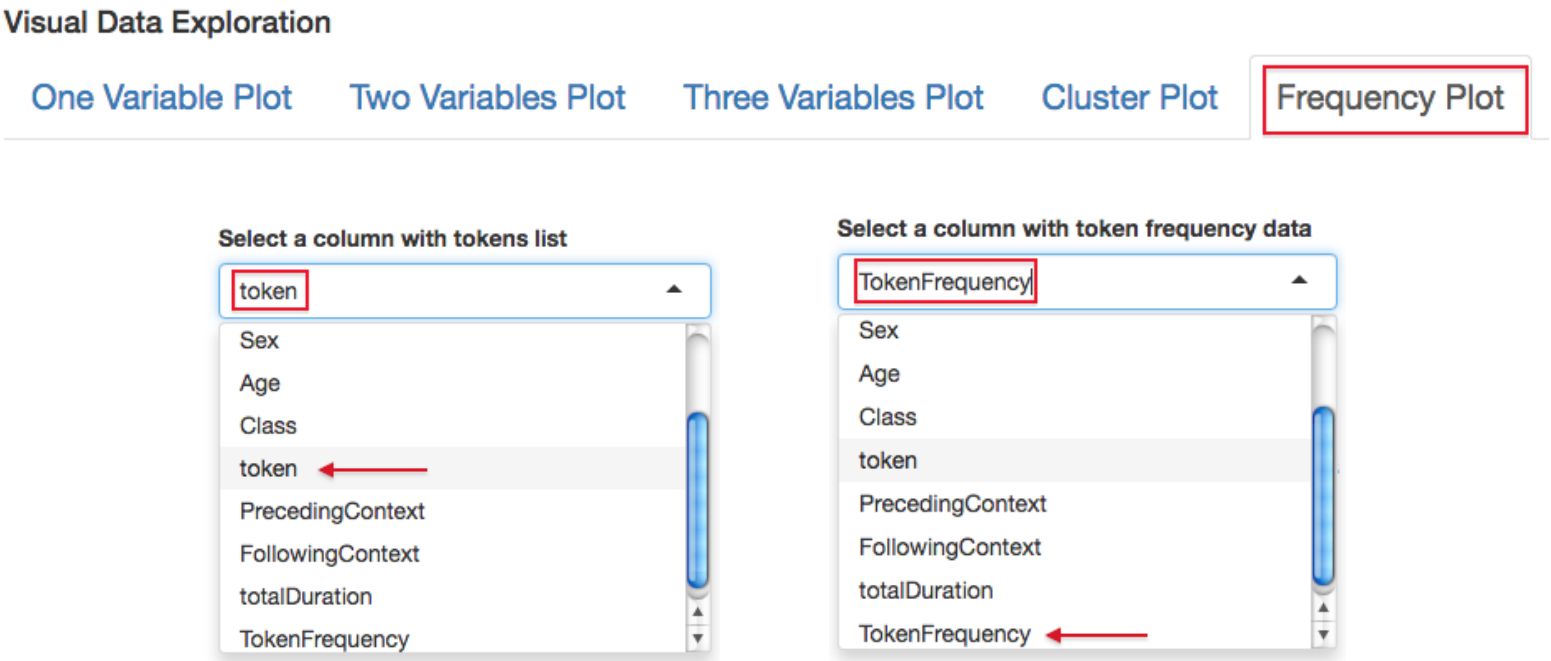
Interpretation - Fixed Effects


```
Fixed effects:
                 Estimate Std. Error         df t value Pr(>|t|)
(Intercept)     9.591e-01  7.495e-03  8.050e+00 127.964 1.31e-14 ***
Sexm            4.018e-03  7.490e-03  8.030e+00   0.537   0.6061
Age35-54        6.121e-04  9.167e-03  8.007e+00   0.067   0.9484
Age55+         -1.643e-02  9.172e-03  8.024e+00  -1.791   0.1110
TokenFrequency  1.082e-05  3.853e-06  6.046e+00   2.807   0.0306 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

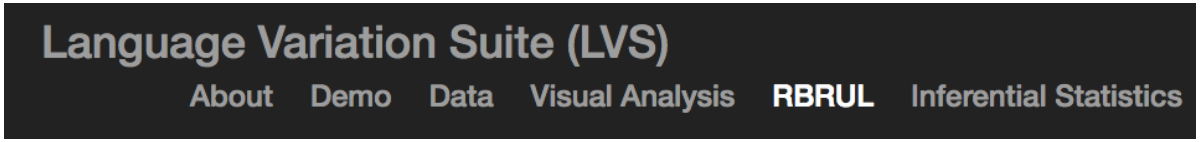
1. **Estimate/coefficient**: reported in log-odds (negative or positive)
2. **P-value**: tells you if the level is significant

Bonus - Word Clouds

Frequency Plot

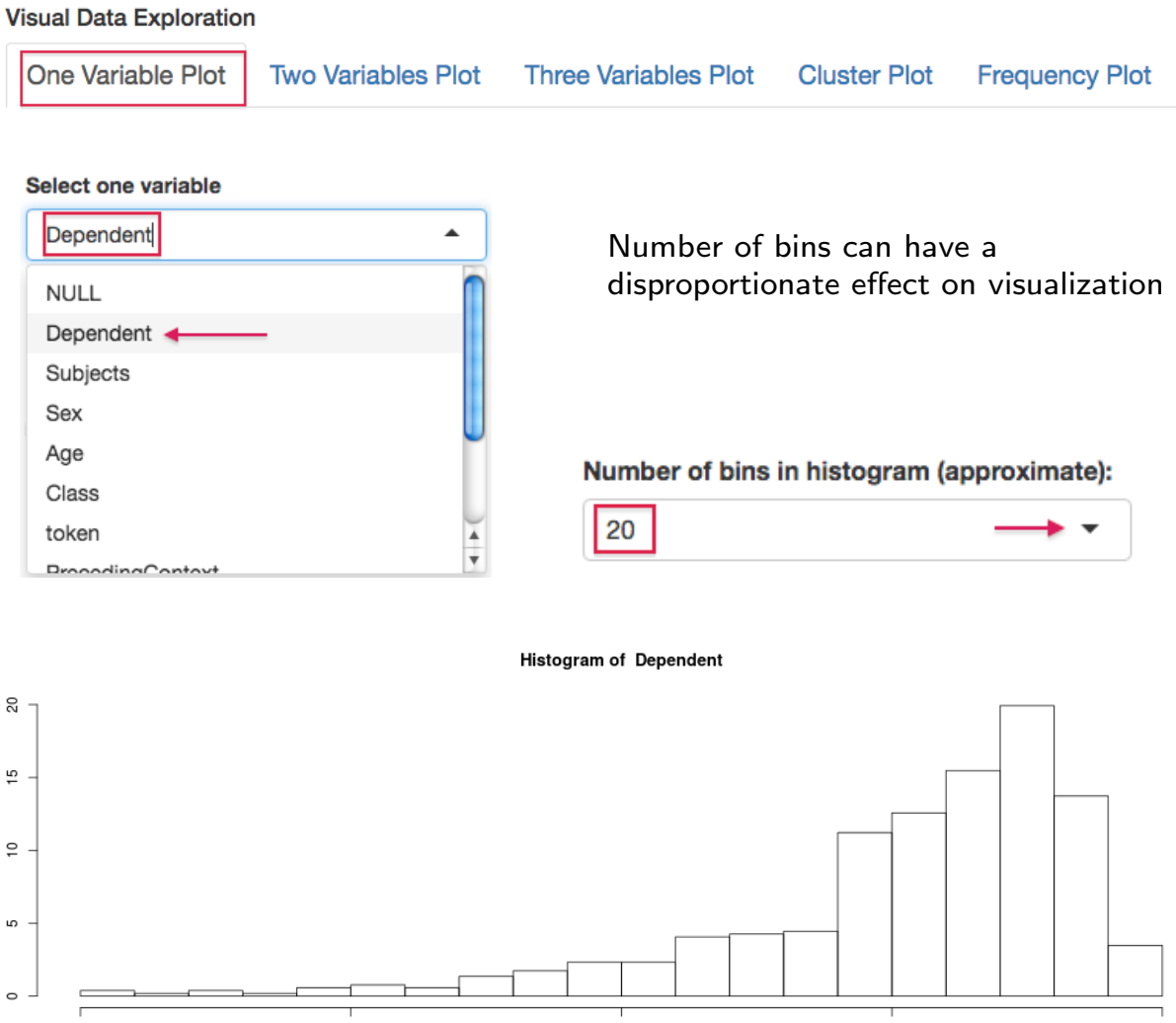


Language Variation Suite - Structure



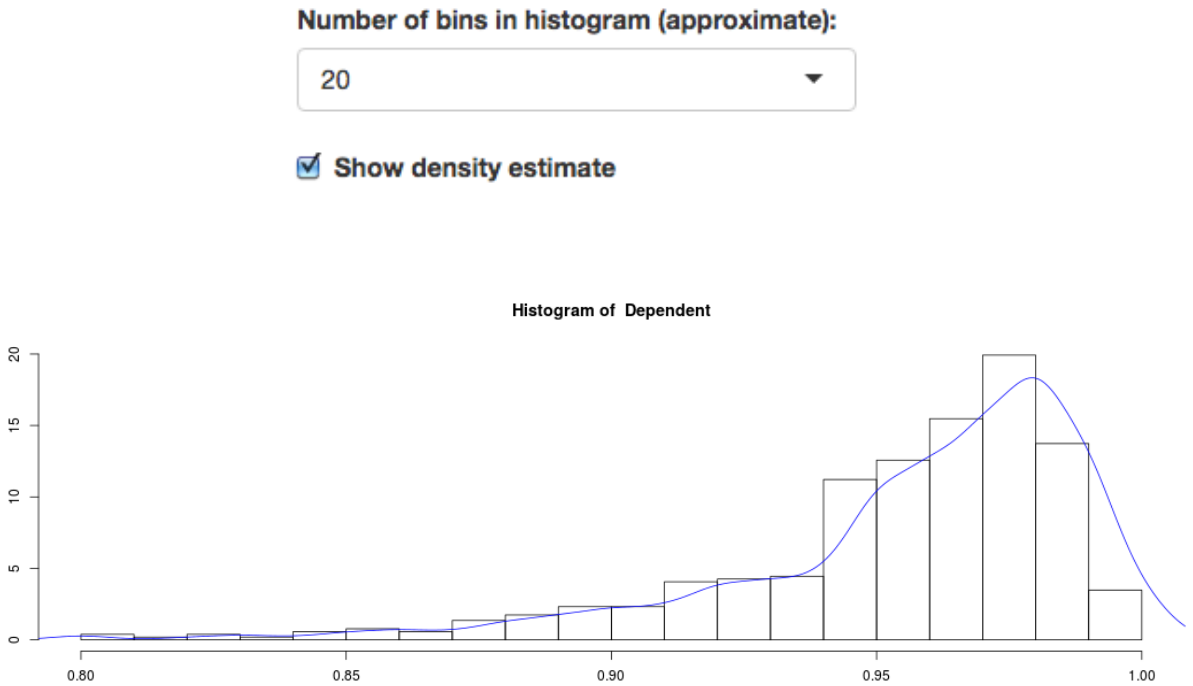
1. Demo
   - Brief introduction
2. Data
   - Upload file, data summary, adjust data, cross tabulation
3. Visual Analysis
   - Plotting, cluster classification
4. RBRUL
   - New version by Daniel Johnson!
5. Inferential statistics
   - Modeling, regression, varbrul analysis, conditional trees, random forest

Appendix 1: Density



Histogram

Density: a non-parametric model of the distribution of points based on a smooth density estimate



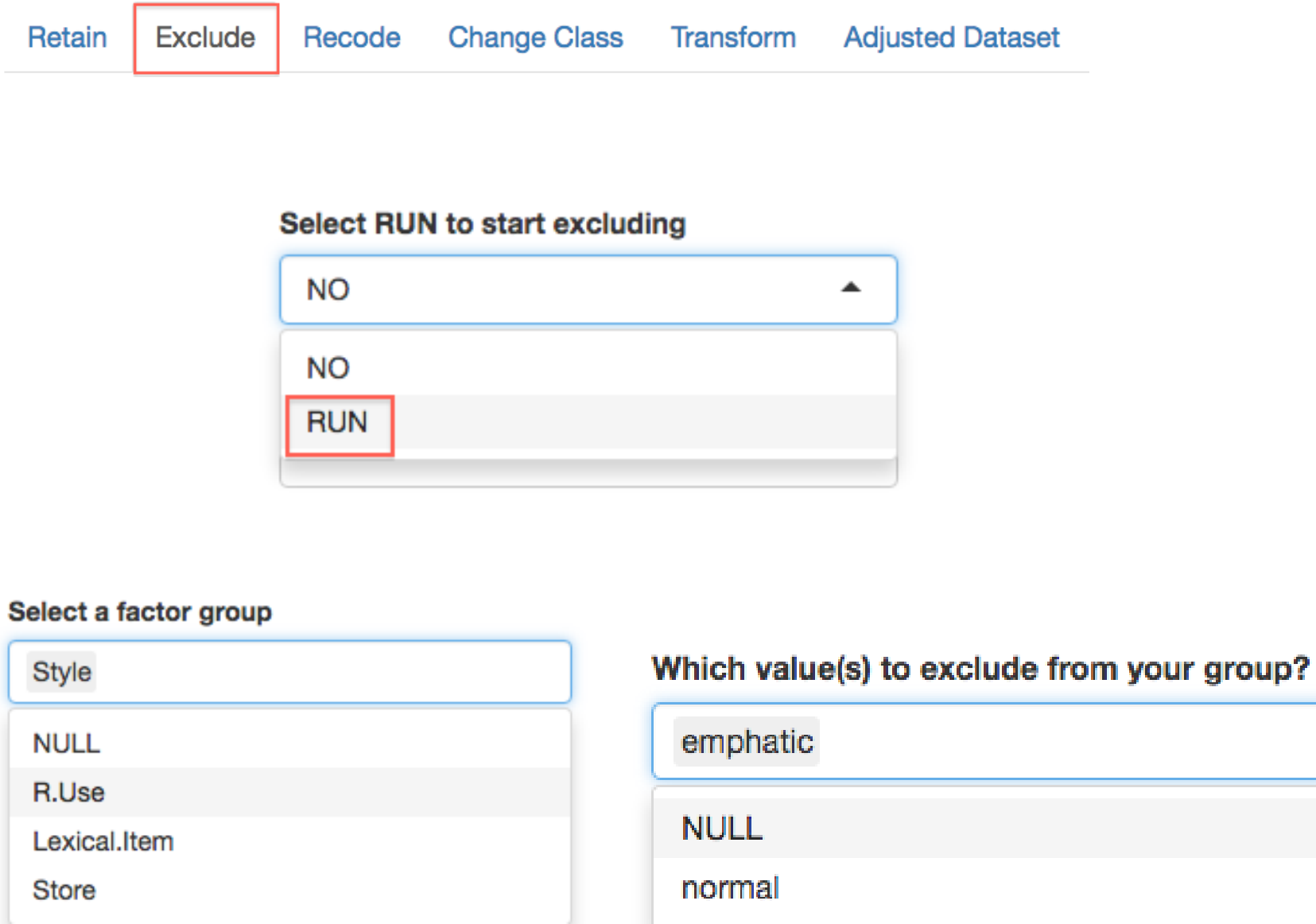
http://scikit-learn.org/stable/modules/density.html

Appenix 2 - Data Modification

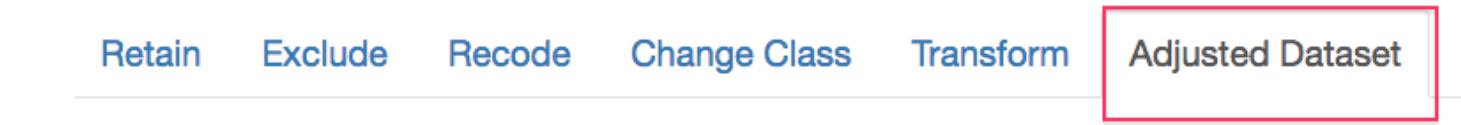
Adjust Data

- **Retain**: Select data subset
- **Exclude**: Exclude variables from a factor group
- **Recode**: Combine and rename variables
- **Change class**: Numeric $\rightarrow$ factor; factor $\rightarrow$ numeric
- **Transform**: Apply log transformation to a specific column
- **ADJUSTED DATASET**:
  - **Run** - to apply all above changes
  - **Reset** - to reset to the original dataset

Exclude: Emphatic Style



Adjusted Dataset



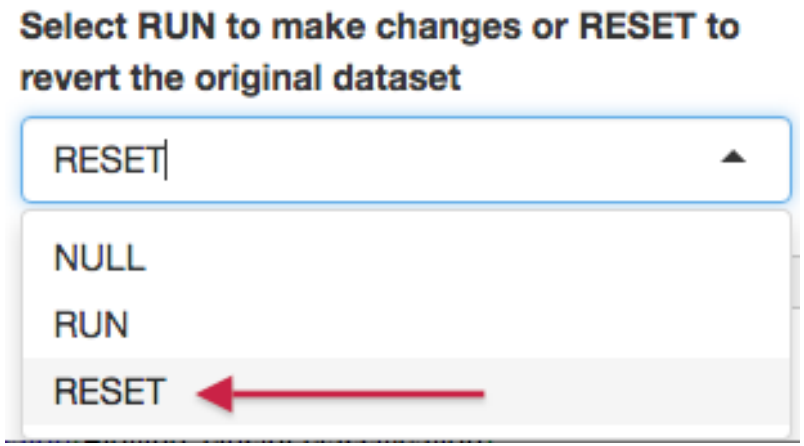
Select RUN to make changes or RESET to revert the original dataset

RUN

```
     R.Use      Lexical.Item     Style          Store
deletion :322   Floor :223    normal:459   Kleins:130
retention:137   Fourth:236                 Macys :224
                                           Saks  :105
```

Adjusting Dataset

To revert to the original data, select **RESET**:



Appendix 3 - Model Comparison

References I

[1] Baayen, Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics*. Cambridge: Cambridge University Press

[2] Bentivoglio, Paola and Mercedes Sedano. 1993. Investigación sociolingüística: sus métodos aplicados a una experiencia venezolana. Boletín de Lingüística 8. 3-35

[3] Gries, Stefan Th. 2015. *Quantitative designs and statistical techniques*. In Douglas Biber Randi Reppen (eds.), The Cambridge Handbook of English Corpus Linguistics. Cambridge: Cambridge University Press

[4] Labov, W. 1966. The Social Stratification of English in New York City. Washington: Center for Applied Linguistics

[5] Schnapp, Jeffrey, and Peter Presner. 2009. Digital Humanities Manifesto 2.0.

[6] http://gifsanimados.espaciolatino.com/x_bob_esponja_8.gif

[7] https://daniellestolt.files.wordpress.com/2013/01/are-you-ready1.gif

[8] http://www.martijnwieling.nl/R/sheets.pdf