

Exploratory Data Analysis



THE BREAD AND BUTTER OF STATISTICS

Getting to Know your Variables



- **Continuous Variables**

- ✦ Look at descriptive statistics (Sample > Descriptives)
- ✦ Plot histograms (Plots > Quick > Histogram)

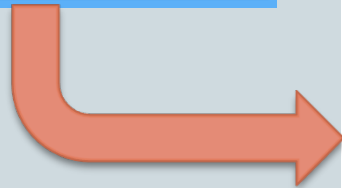
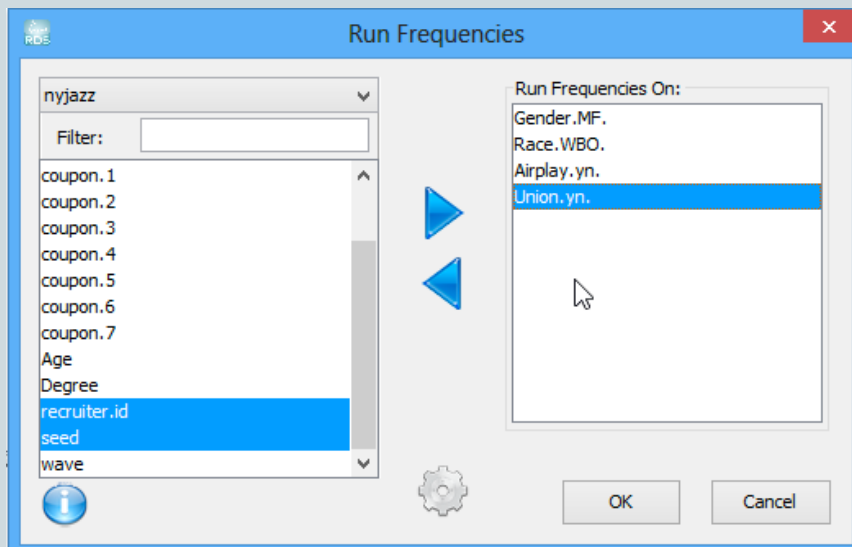
- **Categorical Variables**

- ✦ Look at Frequencies (Sample > Frequencies)
- ✦ Make bar charts (Plots > Quick > bar)

- **Relationships**

- ✦ Crosstabs (Sample > Contingency Tables)
- ✦ Descriptives with strata (Sample Descriptives)

Getting to Know your Categorical Variables



Console View | Element View

Frequencies

Frequencies (Gender.MF.)

	Value	# of Cases	%	Cumulative %
1	1	191	73.70	73.70
2	2	68	26.30	100.00

Case Summary (Gender.MF.)

	Valid	Missing	Total	% Missing
1	259.00	5.00	264.00	1.90

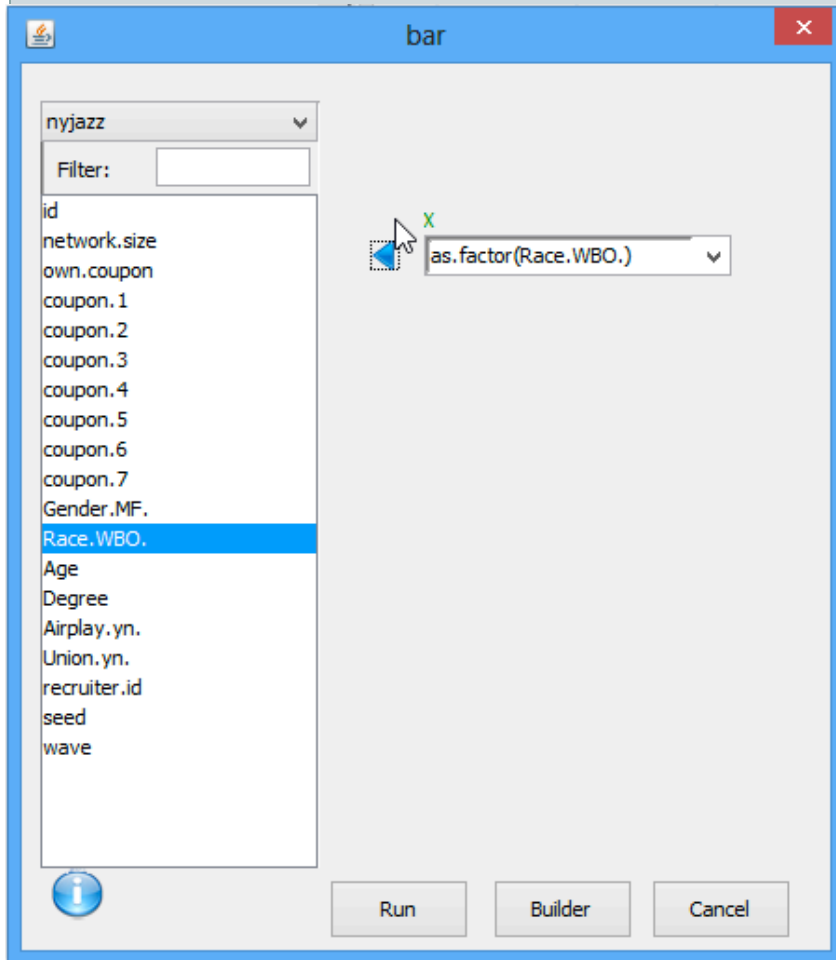
Frequencies (Race.WBO.)

	Value	# of Cases	%	Cumulative %
1	1	142	54.80	54.80
2	2	85	32.80	87.60
3	3	32	12.40	100.00

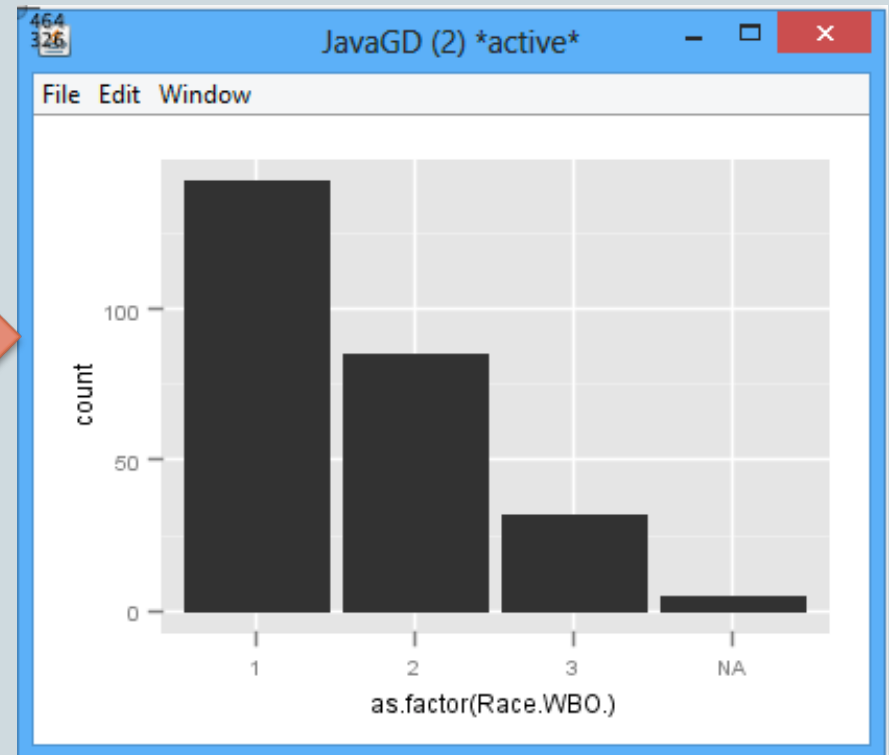
Case Summary (Race.WBO.)

	Valid	Missing	Total	% Missing
1	259.00	5.00	264.00	1.90

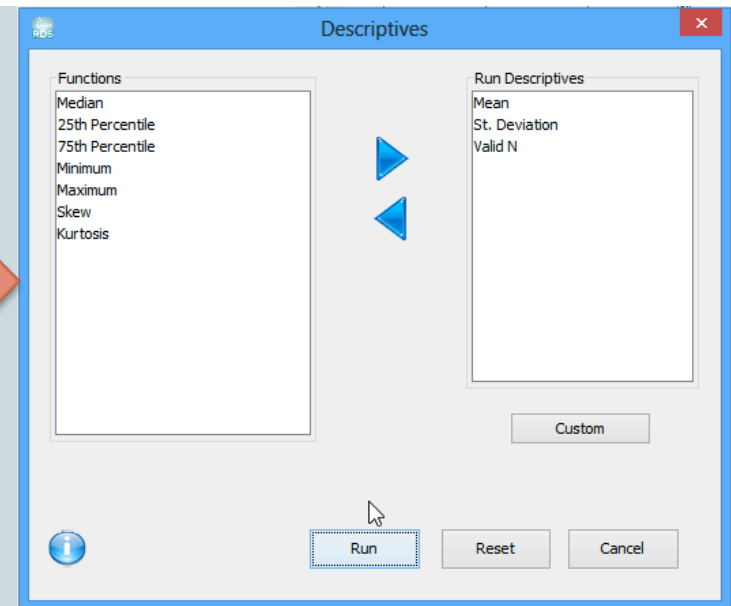
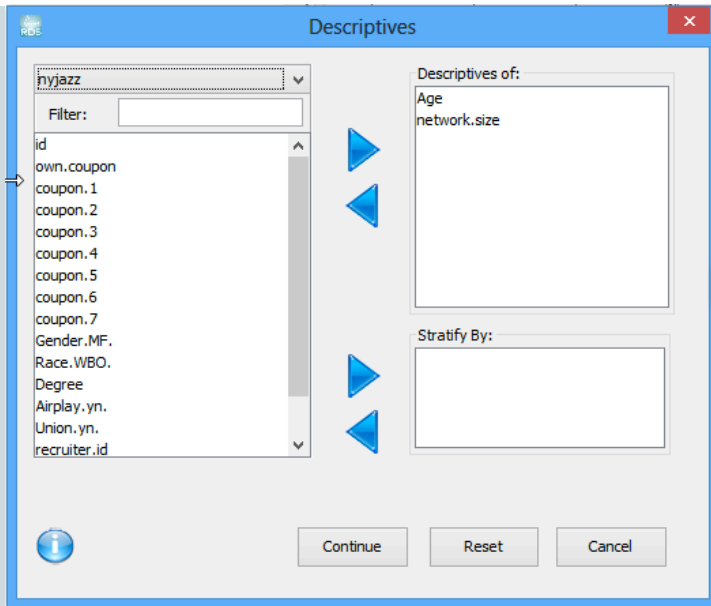
Getting to Know your Categorical Variables



The screenshot shows the 'bar' window in RStudio. On the left, a list of variables is displayed, with 'Race.WBO.' selected. The 'Filter:' field is empty. On the right, the plot type is set to 'as.factor(Race.WBO.)'. The window has 'Run', 'Builder', and 'Cancel' buttons at the bottom.



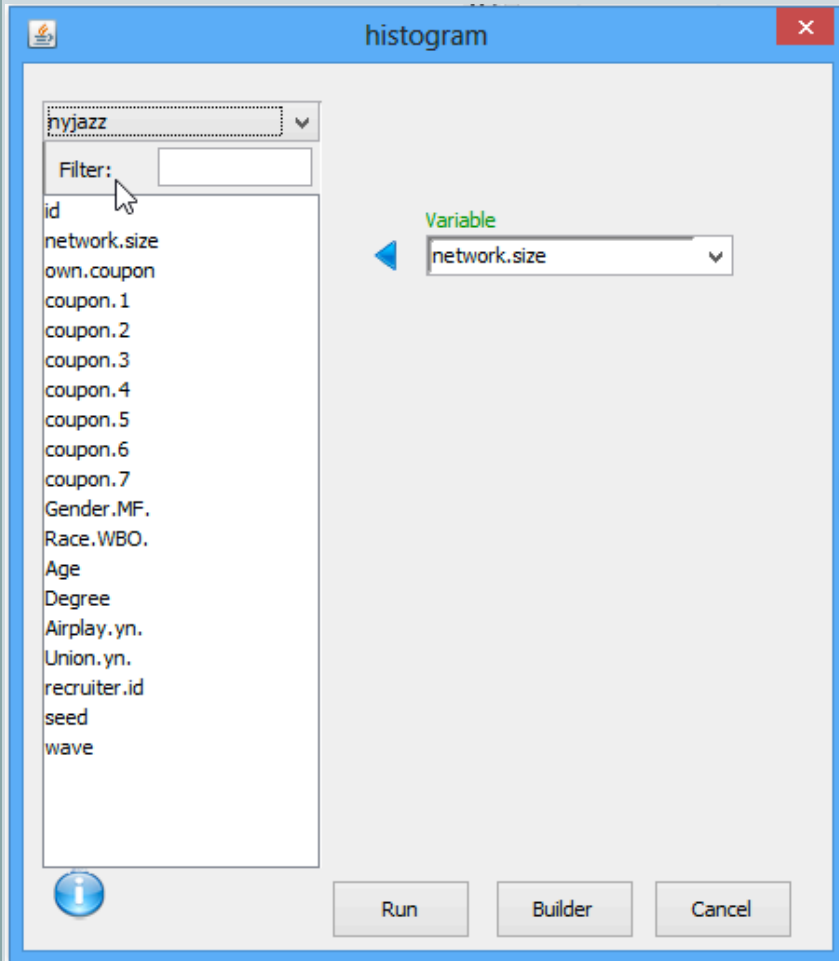
Getting to Know your Continuous Variables



Descriptive Statistics

	Mean	St. Deviation	Valid N
Age	46.46	13.20	263
network.size	223.78	176.17	243

Getting to Know your Continuous Variables



histogram

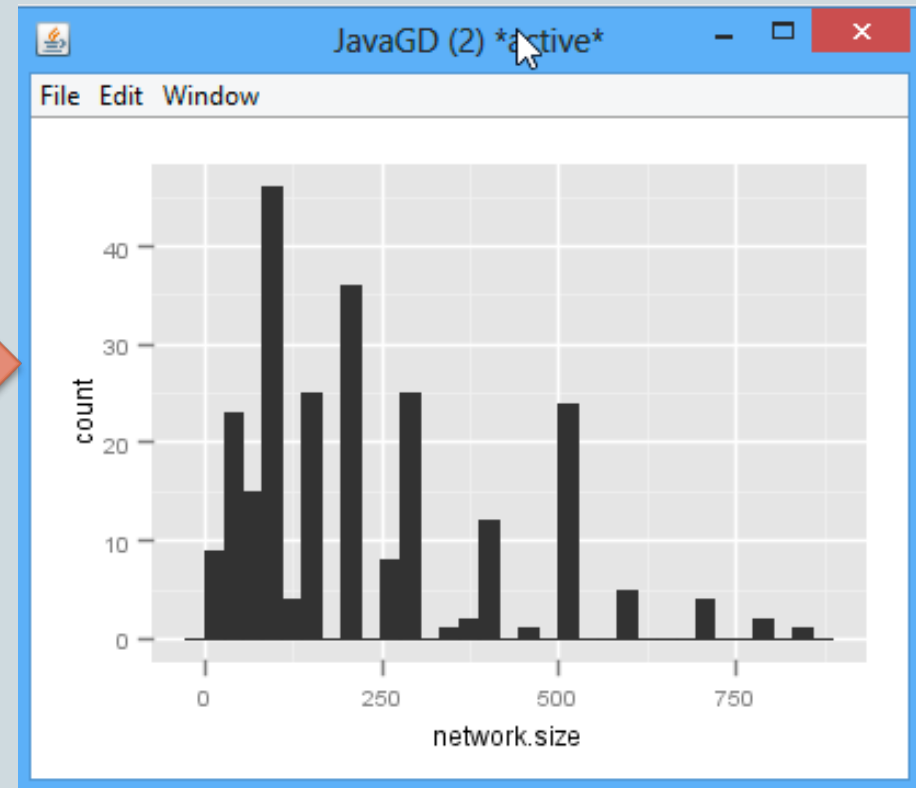
nyjazz

Filter:

id
network.size
own.coupon
coupon.1
coupon.2
coupon.3
coupon.4
coupon.5
coupon.6
coupon.7
Gender.MF.
Race.WBO.
Age
Degree
Airplay.yn.
Union.yn.
recruiter.id
seed
wave

Variable
network.size

Run Builder Cancel



Describing relationships: Crosstabs



Contingency Tables

nyjazz

Filter:

id
network.size
own.coupon
coupon.1
coupon.2
coupon.3
coupon.4
coupon.5
coupon.6
coupon.7
Gender.MF.
Age
Degree
Union.yn.
recruiter.id
seed
wave

Row
Race.WBO.

Column
Airplay.yn.

Stratify By

Subset

Cells
Statistics
Results
Run
Reset
Cancel



Console View Element View

Contingency Tables

Race.WBO. by Airplay.yn. across levels of

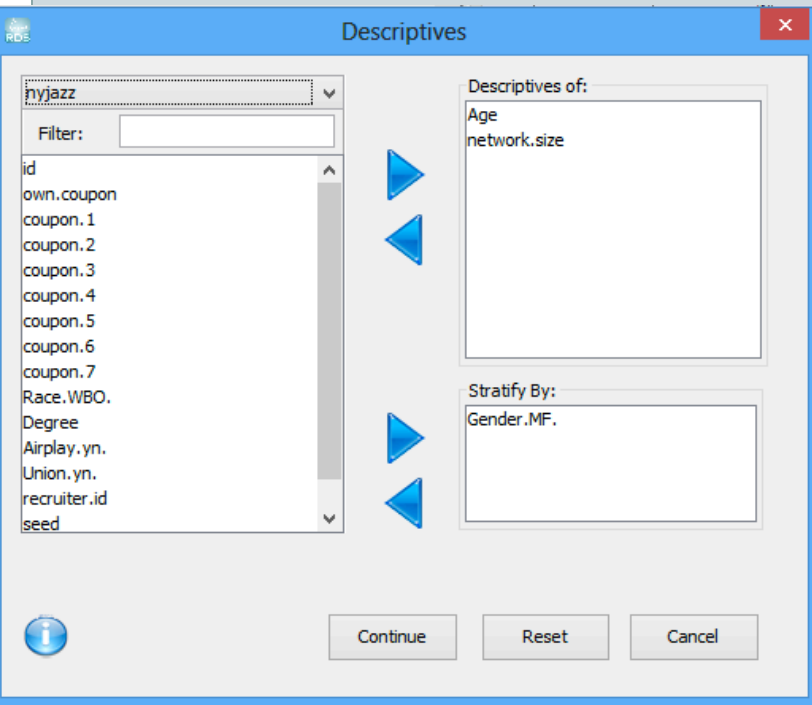
		Airplay.yn.		
Race.WBO.		1	2	Row Total
1	Count	109	27	136
	Row %	80.15%	19.85%	54.84%
	Column %	53.43%	61.36%	
2	Count	69	11	80
	Row %	86.25%	13.75%	32.26%
	Column %	33.82%	25.00%	
3	Count	26	6	32
	Row %	81.25%	18.75%	12.90%
	Column %	12.75%	13.64%	
Column Total		204	44	248
Column %		82.26%	17.74%	

Contingency Table Tests

Tests for Race.WBO. by Airplay.yn. across levels of

	statistic	df	asymptotic p-value
Chi Squared	1.31	2	0.519

Describing relationships: Descriptives



Descriptive Statistics

Variable: Age

	<i>Gender.MF.</i>	<i>Mean</i>	<i>St. Deviation</i>	<i>Valid N</i>
1	1	46.12	12.96	191
2	2	47.88	14.16	67

Variable: network.size

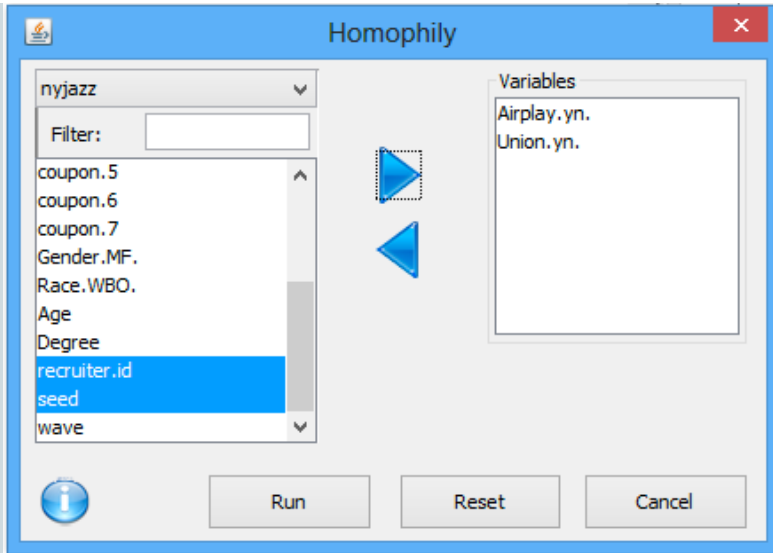
	<i>Gender.MF.</i>	<i>Mean</i>	<i>St. Deviation</i>	<i>Valid N</i>
1	1	222.99	171.78	174
2	2	232.23	190.16	64

Recruitment Information



- **Homophily (Sample > Sample Homophily)**
 - Tendency of like to recruit like.
 - Lots of homophily → High Variance
- **Recruitment tree (Plots > Plot Recruitment Tree)**
 - Are the chains long?
 - Do most of the subjects come from the same seed?
 - Visualize homophily.
- **Other Diagnostics (Plots > Plot Recruitment Tree)**
 - Does network size change over the course of sampling?
 - How many recruits are in each wave?
 - How many recruits originate from each seed?

Homophily



```
$Airplay.yn.  
Recruitment Homophily for Airplay.yn.
```

```
Homophily = 1.019792
```

```
                                     Airplay.yn. of recruit  
Airplay.yn. of respondent   1   2  
1 155 33  
2  40 11
```

```
Number of cases in table: 239
```

```
Number of factors: 2
```

```
Test for independence of all factors:
```

```
Chisq = 0.4306, df = 1, p-value = 0.5117
```

```
$Union.yn.  
Recruitment Homophily for Union.yn.
```

```
Homophily = 1.209508
```

```
                                     Union.yn. of recruit  
Union.yn. of respondent   1   2  
1  48 38  
2  51 113
```

```
Number of cases in table: 250
```

```
Number of factors: 2
```

```
Test for independence of all factors:
```

```
Chisq = 14.409, df = 1, p-value = 0.0001471
```

❖ Homophily near 1 means no homophily

❖ p-values assume a simple random sample, so only use them as rough guides.

❖ Here we see little to no homophily in Airplay.yn. And lots in Union.yn.

Recruitment Tree



Plot Recruitment Tree

nyjazz

Filter:

- id
- network.size
- own.coupon
- coupon.1
- coupon.2
- coupon.3
- coupon.4
- coupon.5
- coupon.6
- coupon.7
- Gender.MF.
- Race.WBO.
- Age
- Degree
- Airplay.yn.
- recruiter.id**
- seed
- wave

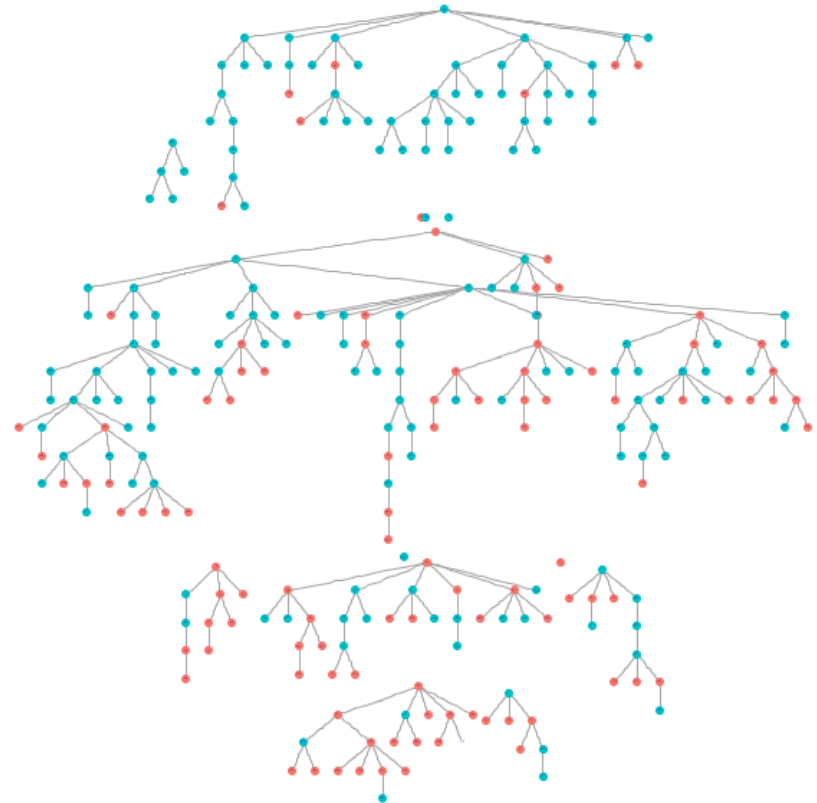
Node Color (optional)

Node Size (optional)

Node Label (by default the id)

Node Label Size

Output
 Graphics windows
 PDF Report



Union.yn.
• 1 • 2

Diagnostics



Diagnostic Plots

nyjazz

Filter:

- id
- network.size
- own.coupon
- coupon.1
- coupon.2
- coupon.3
- coupon.4
- coupon.5
- coupon.6
- coupon.7
- Gender.MF.
- Race.WBO.
- Age
- Degree
- Airplay.yn.
- Union.yn.
- recruiter.id
- seed
- wave

Stratify by (optional)

Plots

- Recruitment tree
- Network size by wave
- Recruits by wave
- Recruits per seed
- Recruits per subject

Output

- Graphics windows
- PDF Report

Run **Reset** **Cancel**

Diagnostics

