



RESEARCHSOFTWARE

A DIVISION OF DISPLAYR

TIM BOCK PRESENTS



DIY Advanced Analysis

Session 3: Driver Analysis

Overview

- Objectives of (key) driver analysis
- Overview of techniques
- Assumptions that need to be checked when doing QA for driver analysis
- Visualization

The basic objective of (key) driver analysis

The basic objective: work out the relative importance of a series of *predictor variables* in predicting an *outcome variable*. For example:

- NPS: comfort vs customer service vs price.
- Customer satisfaction: wait time vs staff friendliness vs comfort.
- Brand preference: modernity vs friendliness vs youthfulness.

What driver analysis is not: predictive analysis (e.g., predicting sales, customer churn). Although, you can use driver analysis to make strategic predictions (e.g., if I improve, say, *fun*, then preference will increase.)

What the data looks like

1 outcome variable

Predictor variables
(Typically there will be more than 3.)

Likelihood to recommend	This brand is <i>fun</i>	This brand is <i>exciting</i>	This brand is <i>youthful</i>
6	1	1	1
9	0	1	0
7	0	0	0
6	1	1	1
9	0	1	0
7	0	0	1
7	0	0	0

This data shows 7 observations

Case study 1: Cola brand attitude

Outcome variable(s)	34 Predictor variable(s)	<i>If the brand was a person, what would its personality be?</i>	
Hate/Dislike/Neither/ Like/Love/Don't know: <ul style="list-style-type: none"> • Coke Zero • Coke • Diet Coke • Diet Pepsi • Pepsi Max • Pepsi 	Brand associations: <ul style="list-style-type: none"> • Beautiful • Carefree • Charming • Confident • Down-to-earth • Feminine • Fun • Health-conscious • Hip • Honest • Humorous 	<ul style="list-style-type: none"> • Imaginative • Individualistic • Innocent • Intelligent • Masculine • Older • Open to new experiences • Outdoorsy • Rebellious • Reckless • Reliable 	<ul style="list-style-type: none"> • Sexy • Sleepy • Tough • Traditional • Trying to be cool • Unconventional • Up-to-date • Upper-class • Urban • Weight-conscious • Wholesome • Youthful

Case study 2 (time permitting): Technology

Outcome variable(s)	Predictor variable(s)
<p>Likelihood to recommend:</p> <ul style="list-style-type: none">• Apple• Microsoft• IBM• Google• Intel• Hewlett-Packard• Sony• Dell• Yahoo• Nokia• Samsung• LG• Panasonic	<p>Brand associations:</p> <ul style="list-style-type: none">• Fun• Worth what you pay for• Innovative• Good customer service• Stylish• Easy-to-use• High quality• High performance• Low prices

The data (stacked)

From: one row per respondent

To: one row per brand per respondent

ID	Likelihood to recommend			This brand is <i>fun</i>			This brand is <i>exciting</i>		
	Apple	Microsoft	IBM	Apple	Microsoft	IBM	Apple	Microsoft	IBM
1	6	9	7	1	0	0	1	1	0
2	8	7	7	1	0	0	1	0	0
3	0	9	8	0	1	0	0	0	0
4	0	0	0	0	0	0	0	0	0



ID	Brand	Likelihood to recommend	This brand is <i>fun</i>	This brand is <i>exciting</i>
1	Apple	6	1	1
1	Microsoft	9	0	1
1	IBM	7	0	0
2	Apple	6	1	1
2	Microsoft	9	0	1
2	IBM	7	0	0
3	Apple	6	1	1
3	Microsoft	9	0	1
3	IBM	7	0	0
4	Apple	6	1	1
4	Microsoft	9	0	1
4	IBM	7	0	0

Tips for stacking

Q

- Get an SPSS .SAV data file. If you do not have an SPSS file:
 - Import your data the usual way
 - **Tools > Save Data as SPSS/CSV** and **Save as type: SPSS**
 - Re-import
- **Tools > Stack SPSS .sav Data File**
- Set the labels for the stacking variable (in Q: `observation`) in **Value Attributes**
- Delete any *None of these* data (e.g., brand associations where respondents were able to select *None of these*)

R / Displayr

The R function `reshape`

Standard “best practice” recommendation for driver analysis:

The average
improvement in R^2 that a
predictor makes across
all possible models (aka
“Shapley”)

LMG

Lindeman, Merenda, Gold (1980)

=

Kruskal

Kruskal (1987)

=

Dominance Analysis

Budescu (1993)

=

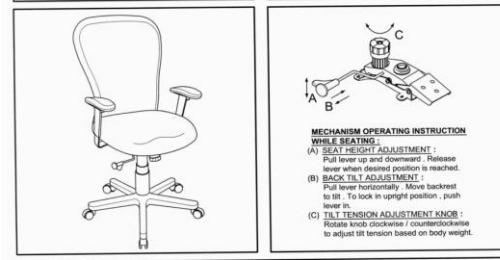
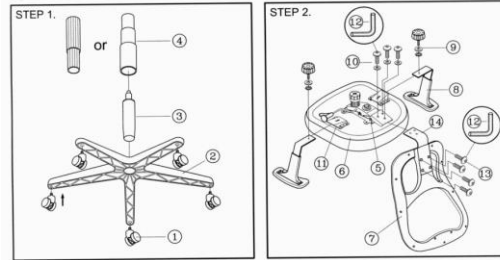
Shapley / Shapley Value

Lipovetsky and Conklin(2001)

euotech ASSEMBLY INSTRUCTION
Remove all items from the carton. Verify all pieces before assembly.

PART LIST

KEY QTY	DESCRIPTION	KEY QTY	DESCRIPTION	KEY QTY	DESCRIPTION
1	5	7	1	11	2
2	1	8	2	12	1
3	1	9	2	13	4
4	1	10	3	14	1
5	1				
6	1				



MECHANISM OPERATING INSTRUCTION
WHILE SEATING:
(A) **SEAT HEIGHT ADJUSTMENT:** Pull lever up and downward. Release lever when desired position is reached.
(B) **BACK TILT ADJUSTMENT:** Pull lever horizontally. Move backrest to tilt. To lock in upright position, push lever in.
(C) **TILT TENSION ADJUSTMENT KNOB:** Rotate knob clockwise / counterclockwise to adjust tilt tension based on body weight.

PREVENTIVE MAINTENANCE AND WARNING!
• USE THIS PRODUCT ONLY FOR SEATING ONE PERSON AT A TIME.
• DO NOT USE THIS CHAIR AS A STEP STOOL, LADDER, OR TO STAND ON.
• DO NOT SIT ON ANY PART OF THE CHAIR EXCEPT THE SEAT.
• DO NOT USE CHAIR ON UNEVEN FLOOR SURFACES.
• DO NOT USE CHAIR UNLESS ALL BOLTS, SCREWS AND KNOBS ARE TIGHT. AT LEAST EVERY SIX MONTHS, CHECK ALL BOLTS, SCREWS AND KNOBS TO BE SURE THEY ARE TIGHT.
• IF ANY PARTS ARE MISSING, BROKEN, DAMAGED OR WORN, STOP USE OF THE PRODUCT UNTIL REPAIRS ARE MADE USING FACTORY AUTHORIZED PARTS.
• DISPOSE OF PACKAGING PROPERLY. RECYCLIC BAG IS NOT A TOY. DO NOT USE PLASTIC BAG AS HEAD COVERING. IT MAY CAUSE SUFFOCATION.
• FAILURE TO FOLLOW THESE WARNINGS COULD RESULT IN SERIOUS INJURY.

MM9500



Much too hard
Best practice:
Bespoke models
(e.g., Bayesian multilevel model)

Too hard
GLMs
(e.g., linear regression)

Too Soft
Bivariate metrics
E.g., Correlations, Jaccard Coefficients

Just Right
Shapley, Relative Importance Analysis

What makes bespoke models and GLMs too hard?

To estimate an OK bespoke model, you need to have a few weeks, and know lots of things, including:

- Joint interpretation of parameter estimates, the predictor covariance matrix, and the parameter covariance matrix
- Conditional effects
- Multicollinearity
- Confounding (e.g., suppressor effects)
- Estimation (ML, Bayesian)
- Specification of informative priors
- Specification of random effects

To understand importance in a GLM (e.g., linear regression), you need to know quite a lot about:

- Joint interpretation of parameter estimates, the predictor covariance matrix, and the parameter covariance matrix
- Conditional effects
- Multicollinearity
- Confounding (e.g., suppressor effects)

Shapley and similar methods allow us to be less careful when interpreting results



Bespoke models
& GLMs



Relative Importance
Analysis

AKA Relative Weight: Johnson (2000)



Random Forest
(for importance analysis)



Shapley



Shapley

With coefficient adjustment
Lipovetsky and Conklin(2001)



Kruskal's Squared
partial correlation
Called **Kruskal** in Q



Proportional
Marginal Variance
Decomposition

Creating Shapley analysis in Q

- Open `Initial.Q`. This already contains the cola data.
- **File > Data Sets > Add to Project > From File > Stacked Technology**
- **Create > Regression > Driver (Importance) Analysis > Shapley**
- Dependent variable: **Q3. Likelihood to recommend [Stacked Technology]**
- Dependent variable: **Q4** variables from `Stacked Technology`
- **No** when asked about confidence intervals (clicking Yes is **OK** as well)
- *Note that High Quality is the most important, with a score of 18.2*
- Right-click: **Reference name:** `shapley`

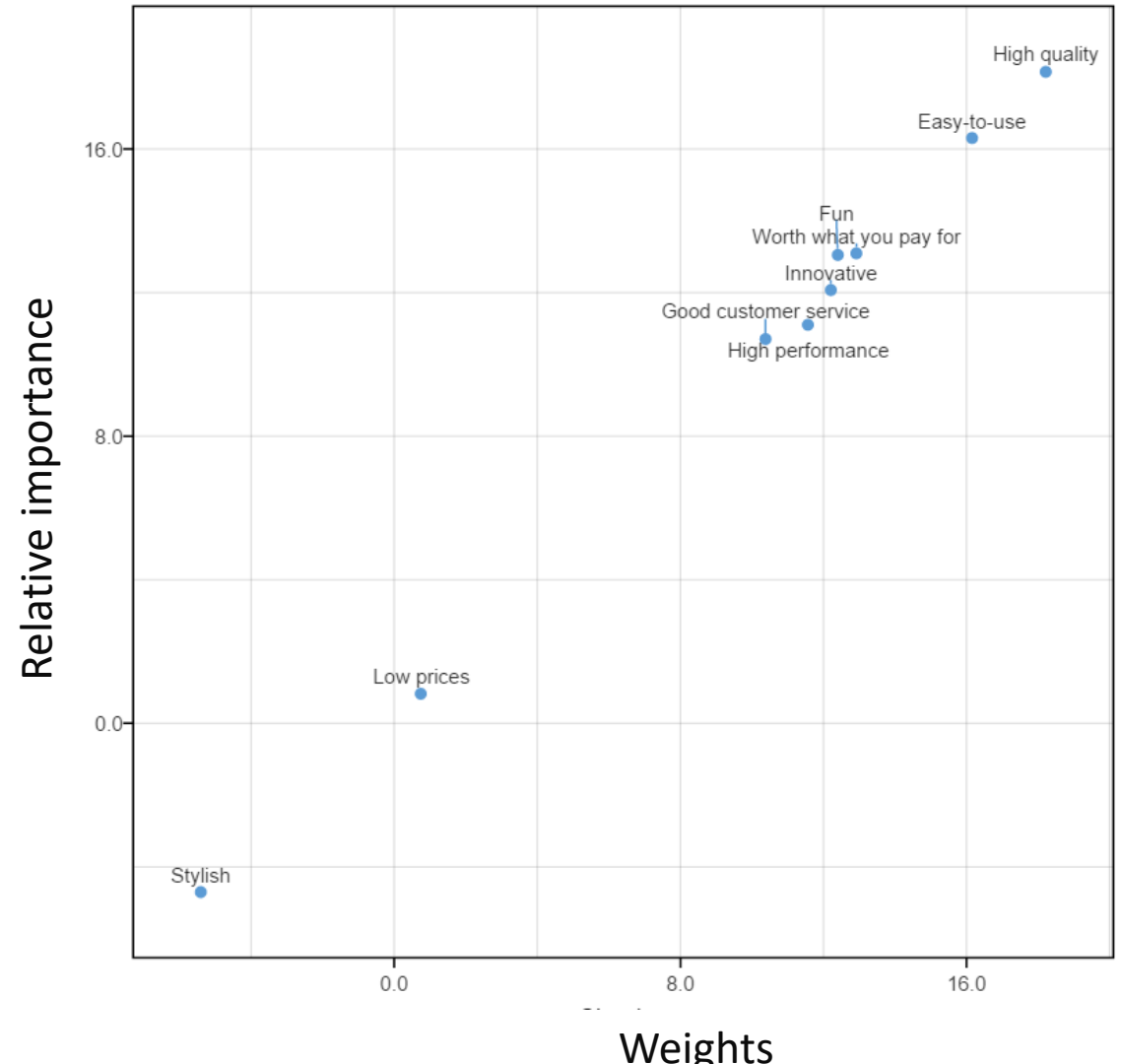
Everything I demonstrate in this webinar is described on a slide like this. The rest of them are hidden in this deck, but you can get them if you download the slides. So, there is no need to take detailed notes.

Shapley and Relative Importance Analysis give very similar results (Case Study 2)

The plot on the right shows that we get very similar results from performing driver analysis using Shapley and Relative Importance Analysis.

Please see the following blog posts for more on this:

- *4 reasons to compute importance using Relative Weights rather than Shapley Regression*
- *The difference between Shapley Regression and Relative Weights*



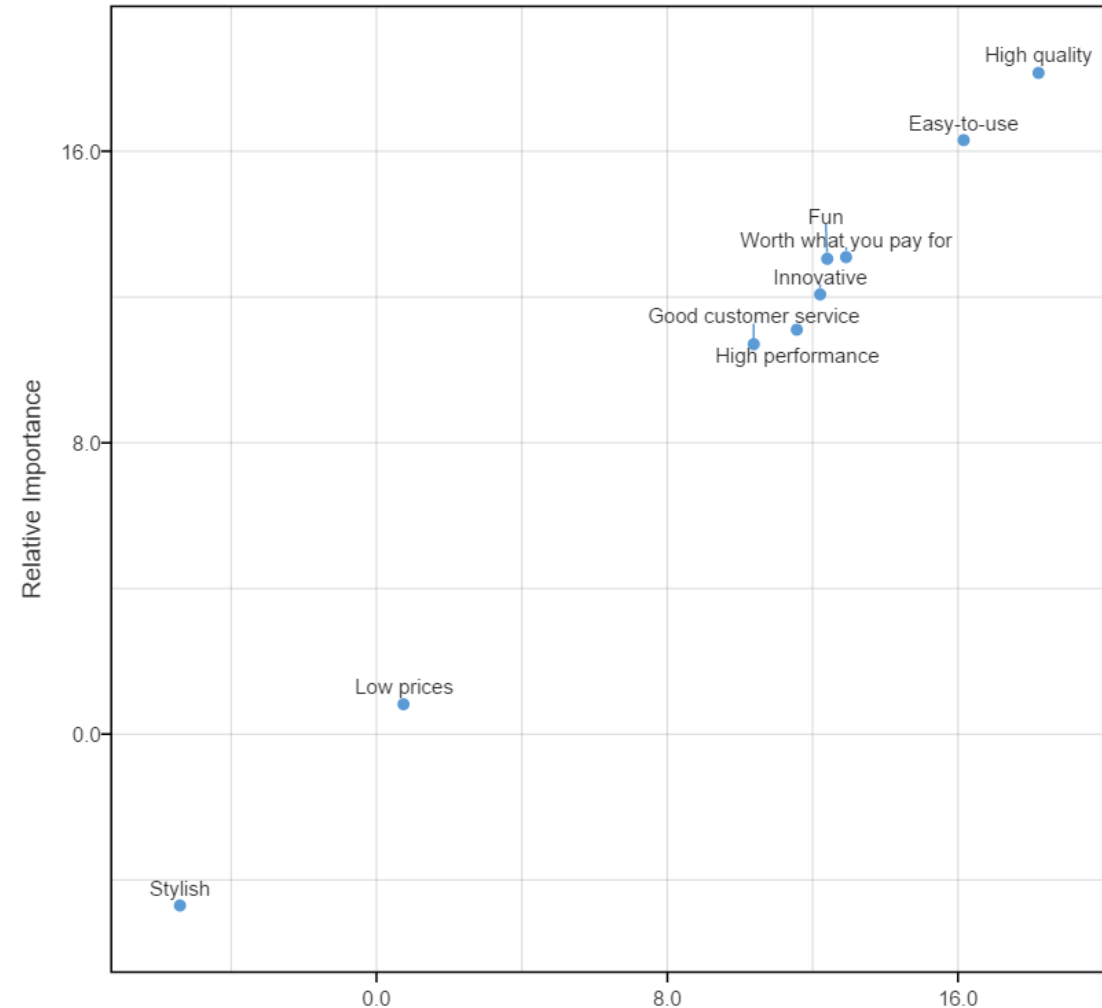
Basic process for driver analysis

1. Import *stacked data*
2. Start with a linear regression model
3. Check the assumptions



2: There are 15 or fewer predictors (if using Shapley)

- *With the cola study, we have 34 variables, and that will take an infinite amount of time to compute, so using Shapley is not an option and we have to use Relative Importance Analysis.*
- *We can use the technology data set, which only has 9 predictors, to explore how similar the techniques are.*
- **Create > Regression > Linear Regression**
 - **Reference name:** `relative.importance`
 - **Select variables**
 - **Output:** Relative importance analysis
 - Check **Automatic Note that High Quality is again most important**
- **Right-click: Add R Output:**

```
comparison = cbind(shapley = shapley[-10],
  "Relative Importance" =
  relative.importance$relative.importance$importance)
```
- **Calculate**
- Change `shapley` to `shapley[-10]`
- **Calculate**
- **Right-click: Add R Output:** `correlation = cor(comparison)`
- Increase number of decimal places. Note the correlation is 0.999
- Rename output: **Correlation**
- **Insert > Charts > Visualization > Labeled Scatterplot,**
 - **Table:** `comparison`
 - **Automatic**

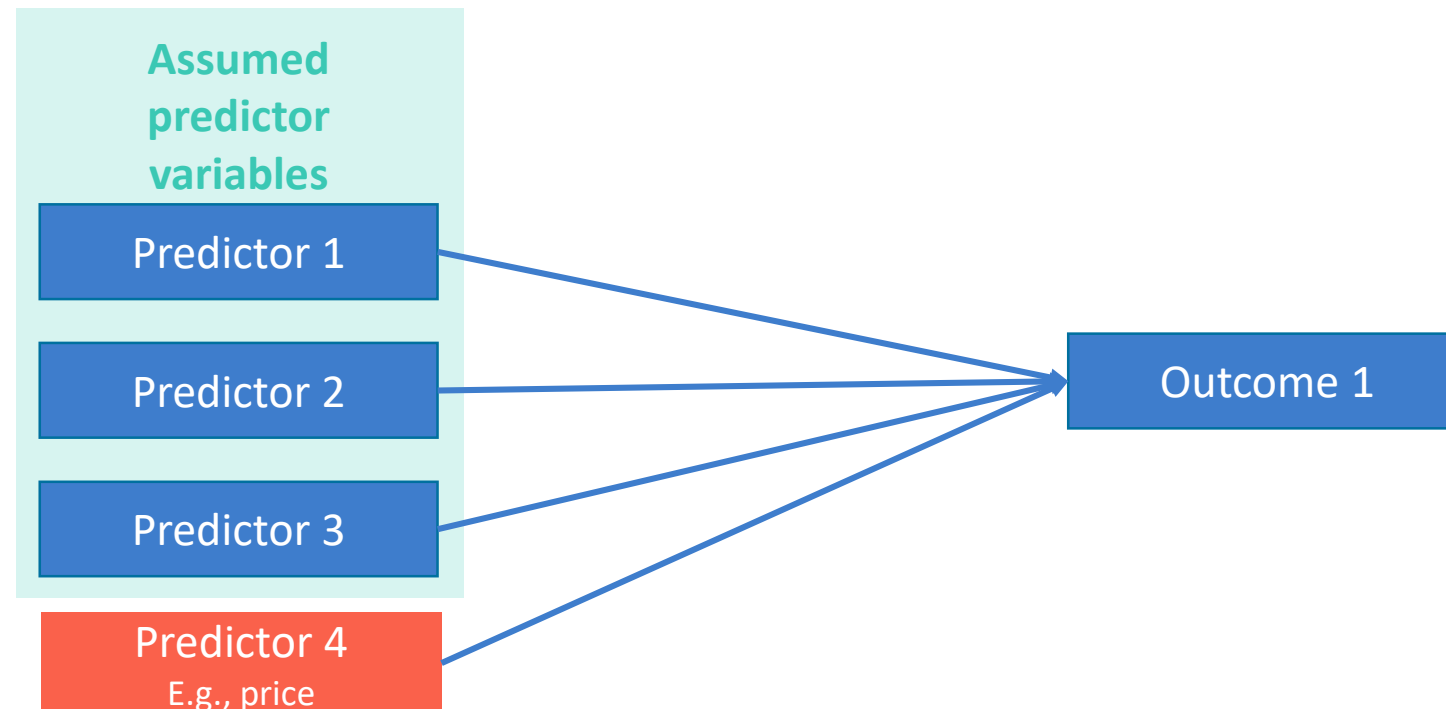


7: The causal model is plausible

	Options (not mutually exclusive) 	Comments 
<p>Issue</p> <p>All driver analysis techniques assume that the analysis is a plausible explanation of the causal relationship between the predictor variables and the outcome variable.</p> <p>This assumption is never true.</p>	<p>Build a bespoke model</p>	<p>This is usually too hard</p>
<p>How to test</p> <p>Common sense. Four common examples are shown on the next slides.</p>	<p>Include all the relevant (non-outcome) variables and cross your fingers (if you have not collected the data, you cannot magic it into existence)</p>	<p>Rightly-or-wrongly, this is how 99.9%* of all modelling is done</p> <p>* Made-up number</p>

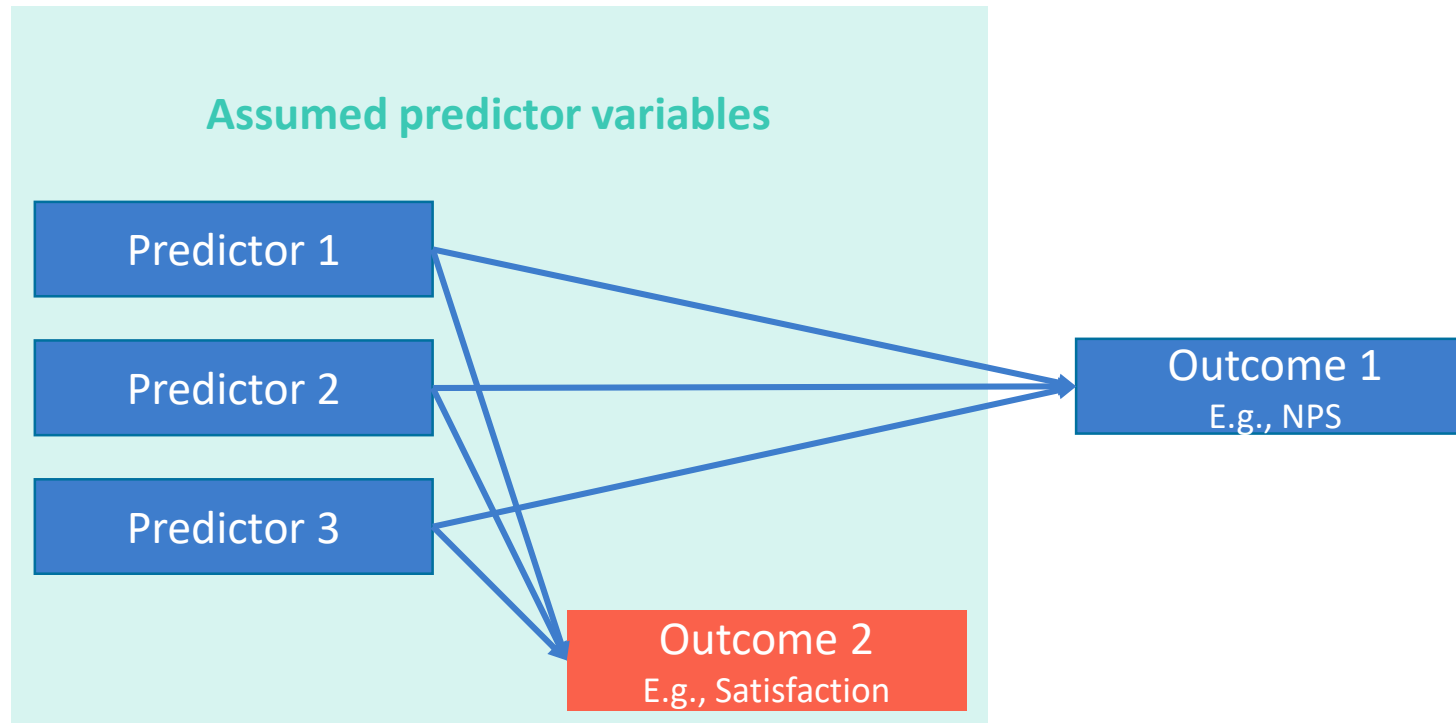
Example causality problem: Omitted variable bias

If we fail to include a relevant predictor variable, and that variable is correlated with the predictor variables that we do include, the estimates of importance will be wrong. If your R-square is less than 0.9, you may have this problem (a typical R-square is closer to 0.2 than 0.9).



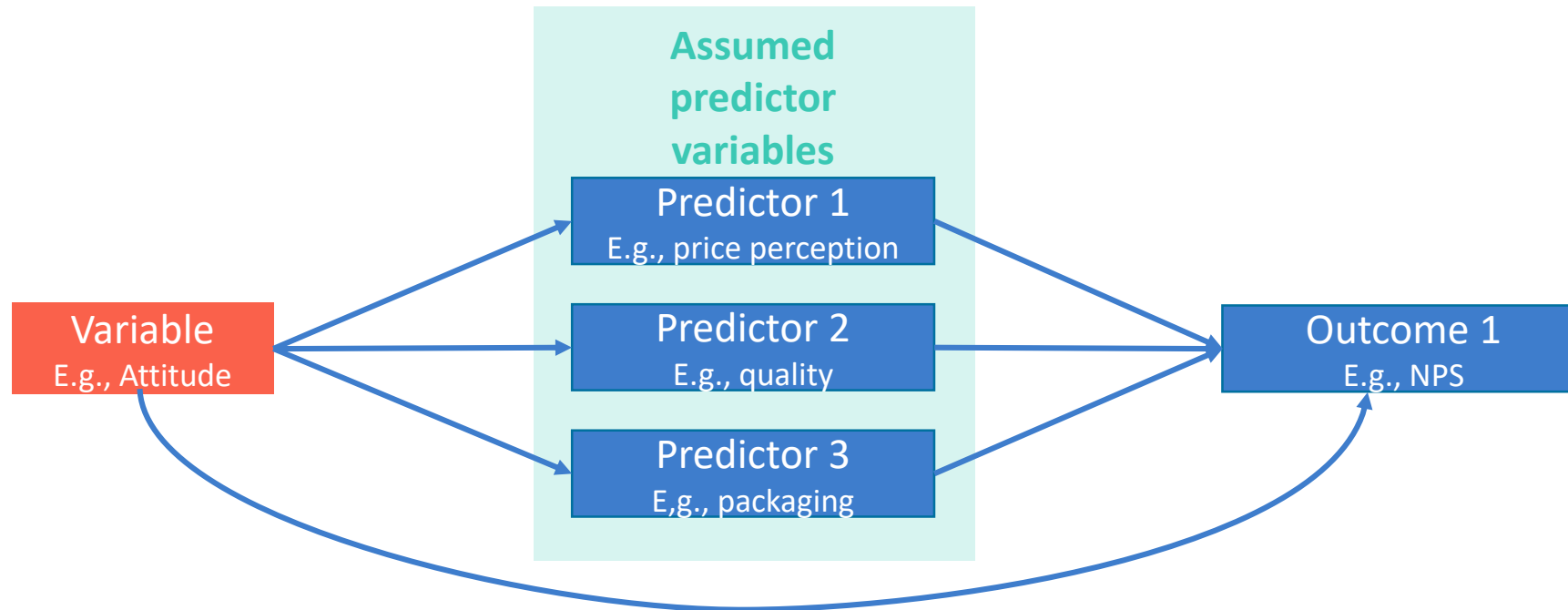
Example causality problem: Outcome variable included as a predictor

If we include a predictor variable that is really an outcome variable, the estimates of importance will be wrong.



Example causality problem: Backdoor path

If *backdoor path* exists from the predictors to the outcome variable, the estimates of importance will be wrong (*spurious*).



Example causality problem: Functional form

If we have the wrong functional form (i.e., assumed equation), the estimates of importance will be wrong.

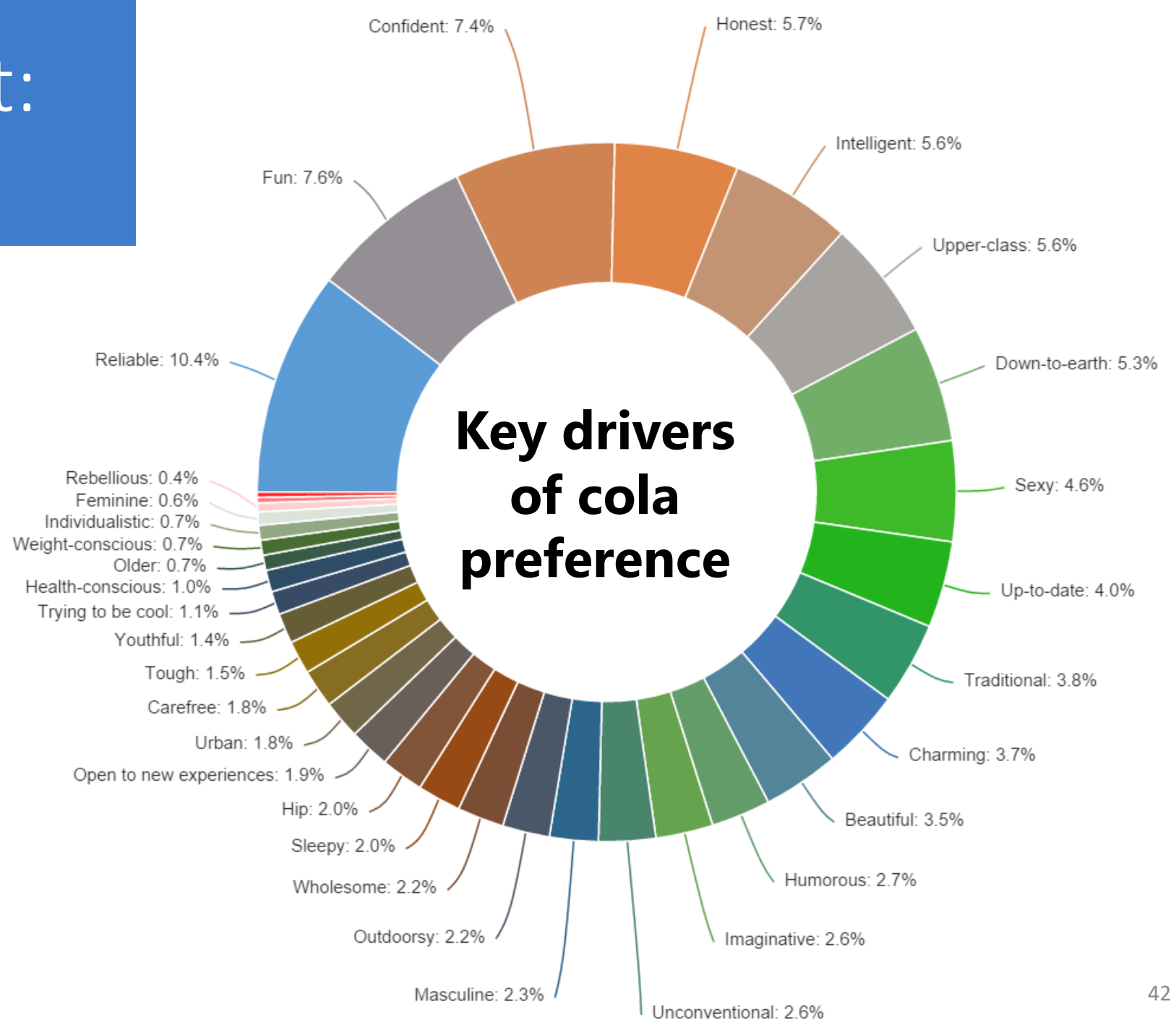
Assumed functional form

$$\text{Outcome} = \text{Predictor 1} + \text{Predictor 2} + \text{Predictor 3}$$

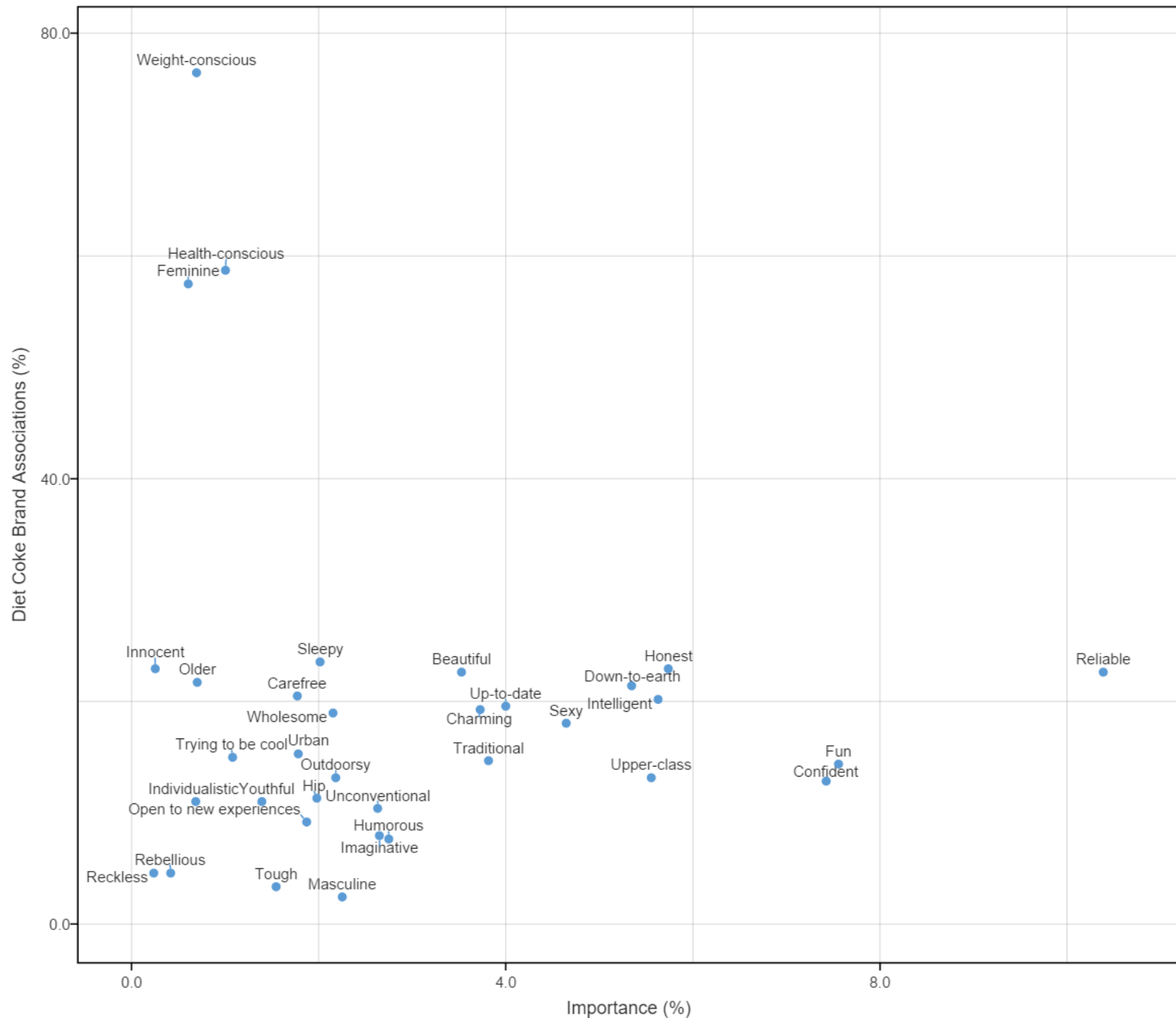
True functional form

$$\text{Outcome} = \text{Predictor 1} \times \text{Predictor 2} + \text{Predictor 3}$$

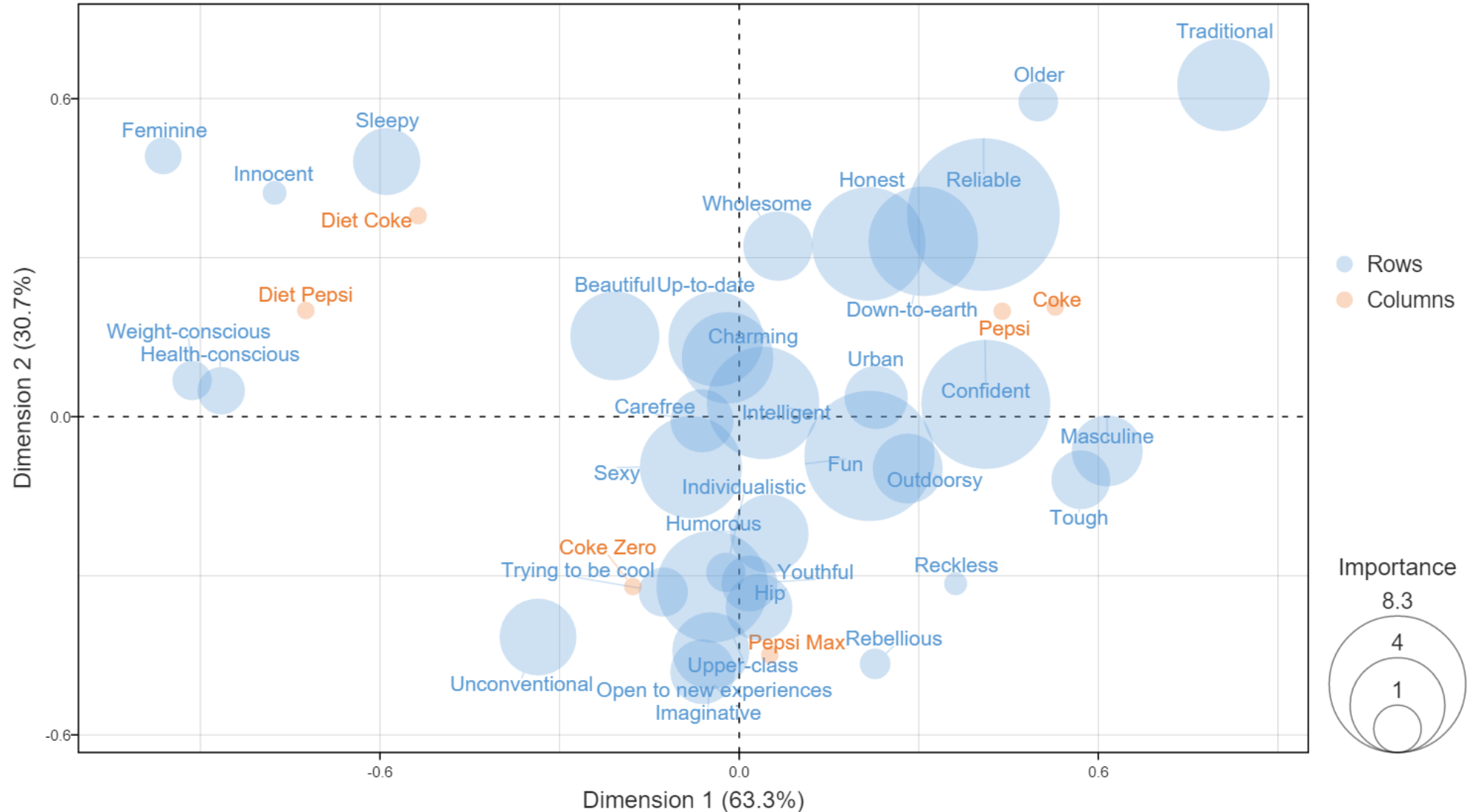
Example output: Importance scores



Example output: Performance- Importance Chart (aka Quad Chart)



Example output: Correspondence Analysis with Importance





RESEARCH SOFTWARE

A DIVISION OF DISPLAYR

TIM BOCK PRESENTS



Q&A Session
