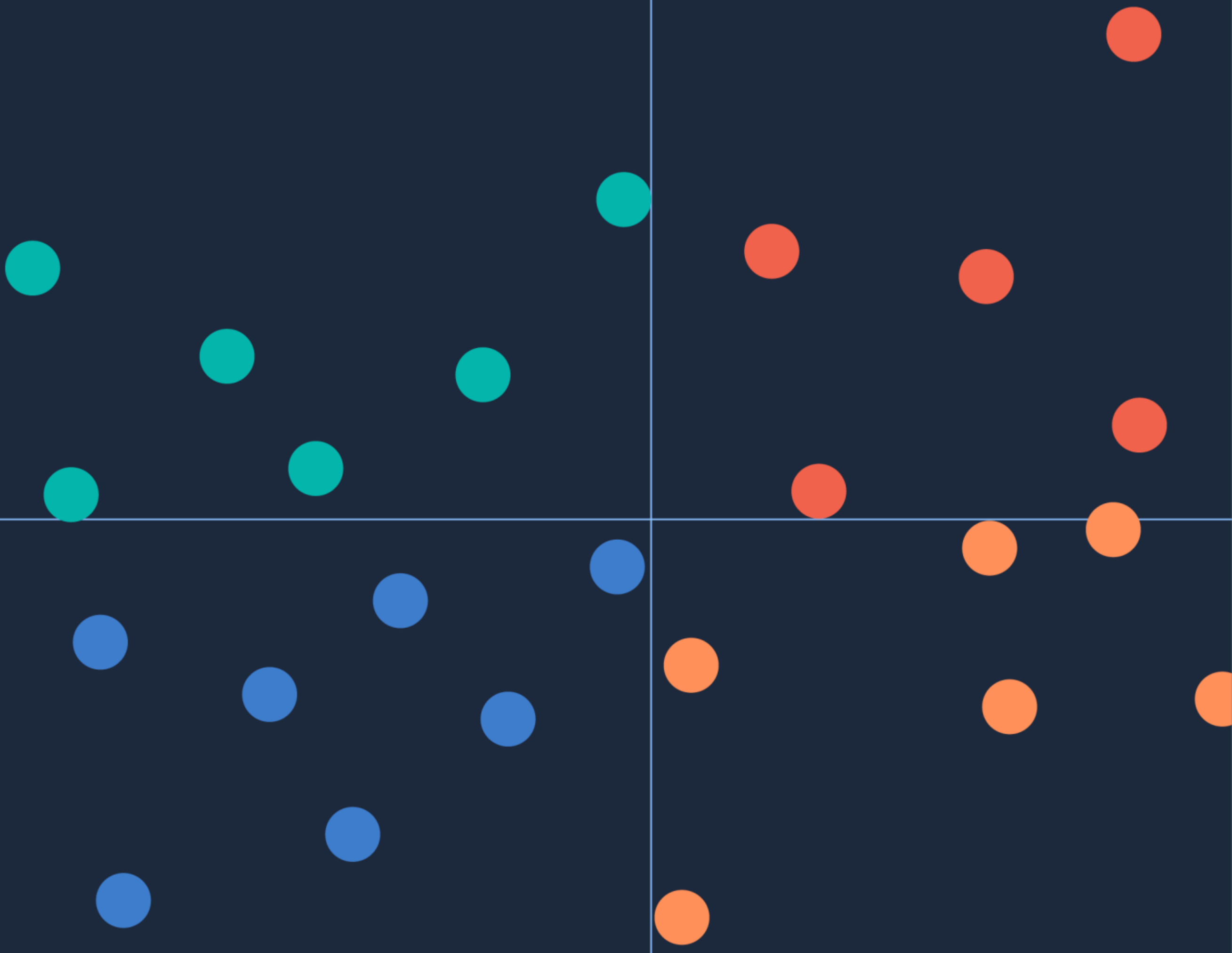
DISPLAYR

RESEARCHSOFTWARE
A DIVISION OF DISPLAYR

Driver Analysis

2nd edition



Analysis and reporting software

displayr.com

Table of Contents

What problem does driver analysis solve?

> Which area of your brand's performance should you concentrate on improving? Should you be focusing on customer service or price? Should you position as cool or competent? These are the types of questions answered by driver analysis.

Who this e-book is for

> This e-book is for people that want to do their own driver analysis.
>
> It is written for people of all levels of driver analysis expertise, from novices through to experts.
>
> The book assumes you are familiar with the basics of conducting and analyzing surveys.

10 Steps to Driver Analysis

> Driver analysis can be performed as a 10-step process, which includes:
>
> - Data preparation
> - Choosing the appropriate technique
> - Missing data
> - Model checking
> - Data visualization

To perform driver analysis:

- **a.** In Q: **Create > Regression > Driver Analysis**
- **b.** In Displayr: **Anything > Advanced Analysis > Regression > Driver Analysis**

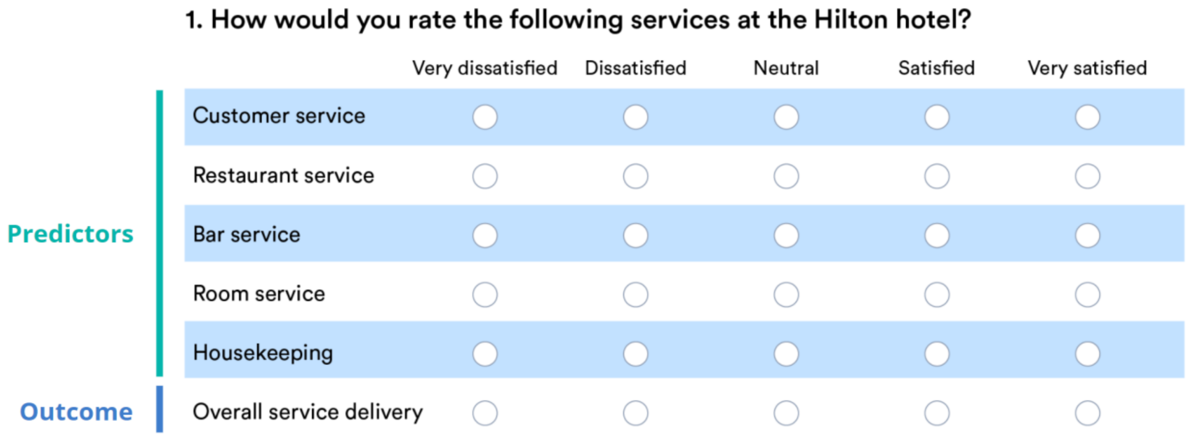
There are 10 steps to work your way through:

1. Check that driver analysis is appropriate (see **Chapter 1. Check that driver analysis is the right technique**).
2. Transform the predictors so that they are either numeric or binary (see **Chapter 2. Make predictors numeric or binary**).
3. Recode, reorder, and merge categories so that they are ordered from lowest to highest (see **Chapter 3. Assign higher values to better performance levels of the outcome and predictor variables**).
4. If you have repeated measures, choose the **Stack data** option (see **Chapter 4. Stack or use auto-stacking software (if you have repeated measures)**).
5. Choose the appropriate **Regression type** (see **Chapter 5. Choose the right regression type**).
6. If you have numeric data, choose **Output** as **Shapley Regression** (see **Chapter 6. Use Shapley Regression or Johnson's Relative Weights**).
7. Choose the appropriate **Missing data** option (see **Chapter 7.Select a proper technique for missing values**).
8. Review the various technical diagnostics, such as testing for outliers and heteroscedasticity (see **Chapter 8. Review diagnostics**).
9. Review the p-values and standard errors (see **Chapter 9. Check statistical significance**).
10. Present the results in the best way (see **Chapter 10. Data visualization**).

1. Check that driver analysis is the right technique

> Driver analysis is the right technique when you are trying to work out what's important. For example, what are the most important determinants of satisfaction.
>
> If the goal is to make quantitative predictions, and/or to use behavioral or demographic data, it is not the right technique.

The grid below is from a survey by the Hilton.[1] The last row collects data on the overall level of performance. This is an *outcome* of interest to the Hilton. The other lines measure the Hilton's performance on various attributes. Each of these attributes is a *predictor* of the outcome of *Overall Service Delivery*.

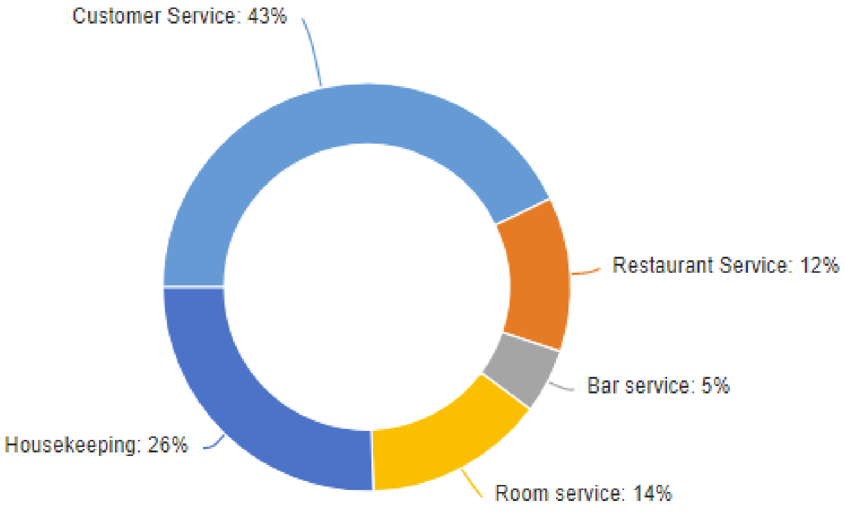


1. How would you rate the following services at the Hilton hotel?

		Very dissatisfied	Dissatisfied	Neutral	Satisfied	Very satisfied
Predictors	Customer service	○	○	○	○	○
	Restaurant service	○	○	○	○	○
	Bar service	○	○	○	○	○
	Room service	○	○	○	○	○
	Housekeeping	○	○	○	○	○
Outcome	Overall service delivery	○	○	○	○	○

Driver analysis, which is also known as *key driver analysis, importance analysis*, and *relative importance analysis*, uses the data from questions like these to work out the relative importance of each of the *predictor variables* in predicting the *outcome variable.* Each of the *predictors* is commonly referred to as a *driver.* The goal is to quantify the importance of each of the drivers. That is, the goal is to compute *importance scores,* so that we can work out which drivers are key. Importance scores are sometimes referred to as *importance weights.*

The key output from driver analysis is typically a table or chart showing the relative importance of the different drivers (predictors), such as the chart below.



---

[1] http://blog.clientheartbeat.com/customer-survey-examples/

Whereas the focus in much of data science is on prediction, with driver analysis the focus is instead on identifying the relative importance of the predictors (drivers).

The Hilton example focuses on understanding how different aspects of performance drive customer satisfaction. The other major application of driver analysis is to know how various brand associations, such as whether a brand is *Hip, Humorous,* or *Honest,* drive performance.

When not to use driver analysis

If our goal is to make quantitative predictions, such as next month’s sales, driver analysis is the wrong technique. It does not make such predictions.

Furthermore, it is only appropriate for calculating importance when the predictors are dimensions or components of performance, as in the Hilton example. It is not suitable when you have demographic or behavioral predictors.

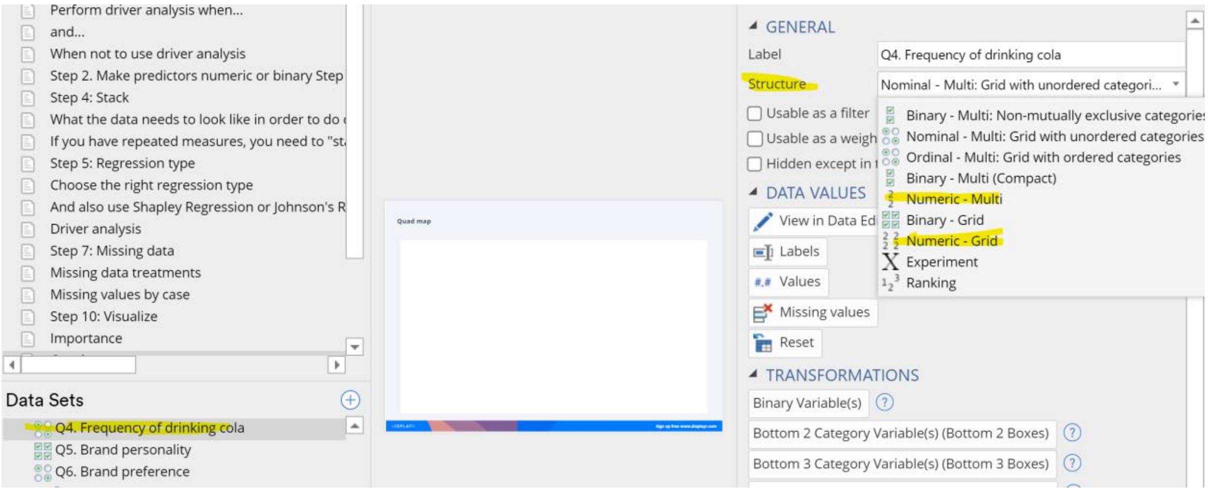
2. Make predictors numeric or binary

> The standard *driver analysis* techniques assume that the predictor variables are numeric rather than categorical variables.

Making variables numeric in Displayr

All *variable sets* in Displayr are stored with a default structure. You can override this by:

1. Selecting the variable set.
2. Changing **Structure**.
3. For driver analysis, choosing one of the three numeric structures:
    a. **Numeric**
    b. **Numeric – Multi**
    c. **Numeric – Grid**



Label: Q4. Frequency of drinking cola

Structure: Nominal - Multi: Grid with unordered categori...

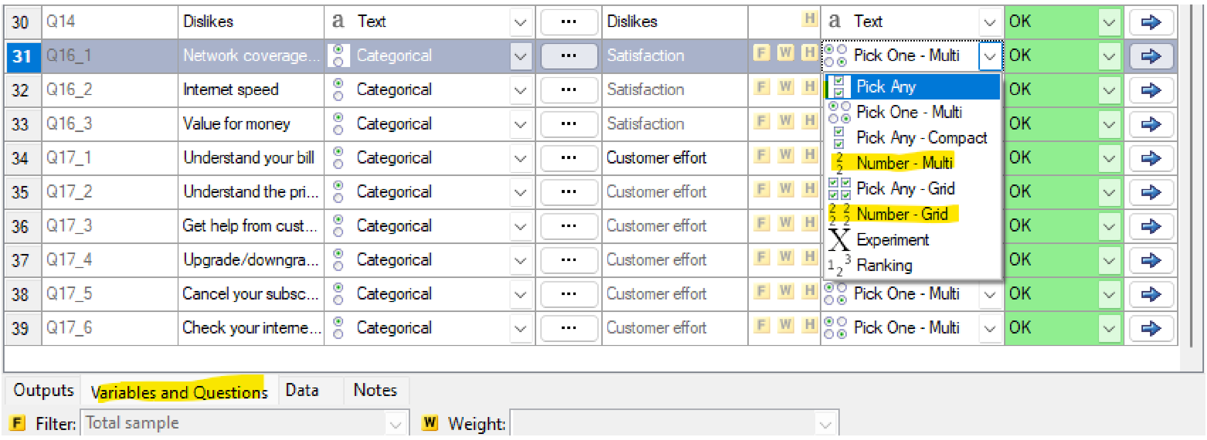
Perform driver analysis when...
and...
When not to use driver analysis
Step 2. Make predictors numeric or binary Step
Step 4: Stack
What the data needs to look like in order to do
If you have repeated measures, you need to "st
Step 5: Regression type
Choose the right regression type
And also use Shapley Regression or Johnson's R
Driver analysis
Step 7: Missing data
Missing data treatments
Missing values by case
Step 10: Visualize
Importance
Data Sets
Q4. Frequency of drinking cola
Q5. Brand personality
Q6. Brand preference
Quad map
GENERAL
Usable as a filter
Usable as a weigh
Hidden except in
Binary - Multi: Non-mutually exclusive categories
Nominal - Multi: Grid with unordered categories
Ordinal - Multi: Grid with ordered categories
Binary - Multi (Compact)
Numeric - Multi
Binary - Grid
Numeric - Grid
Experiment
Ranking
DATA VALUES
View in Data Ed
Labels
Values
Missing values
Reset
TRANSFORMATIONS
Binary Variable(s)
Bottom 2 Category Variable(s) (Bottom 2 Boxes)
Bottom 3 Category Variable(s) (Bottom 3 Boxes)

.

Making variables numeric in Q

All *questions* in Displayr are stored with default structure. You can override this by:

1. Going to the Variables and Questions tab.
2. Changing the **Question Type** setting.
3. For driver analysis, choosing one of the three numeric structures:
    a. **Number**
    b. **Number – Multi**
    c. **Number – Grid**



30	Q14	Dislikes	Text	···	Dislikes	H	Text	OK
31	Q16_1	Network coverage...	Categorical	···	Satisfaction	F W H	Pick One - Multi	OK
32	Q16_2	Internet speed	Categorical	···	Satisfaction	F W H		OK
33	Q16_3	Value for money	Categorical	···	Satisfaction	F W H		OK
34	Q17_1	Understand your bill	Categorical	···	Customer effort	F W H		OK
35	Q17_2	Understand the pri...	Categorical	···	Customer effort	F W H		OK
36	Q17_3	Get help from cust...	Categorical	···	Customer effort	F W H		OK
37	Q17_4	Upgrade/downgra...	Categorical	···	Customer effort	F W H		OK
38	Q17_5	Cancel your subsc...	Categorical	···	Customer effort	F W H	Pick One - Multi	OK
39	Q17_6	Check your interne...	Categorical	···	Customer effort	F W H	Pick One - Multi	OK

Filter: Total sample

Weight:

3. Assign higher values to better performance levels of the outcome and predictor variables

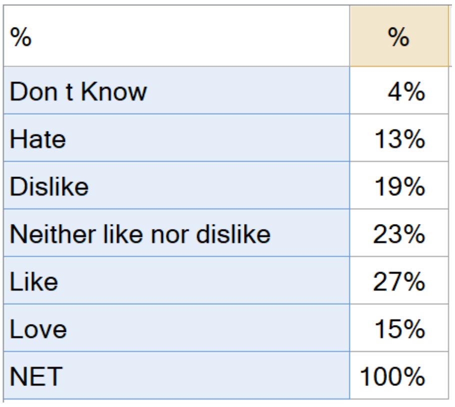
> The standard *driver analysis* techniques assume that the outcome and predictor variables are ordered from lowest to highest, where higher levels indicate more positive attitudes.

There are two different ways of reordering:

- Reordering categories when the outcome variable is categorical.
- Recoding values for numeric outcomes and predictors.

Reordering categories

The table below shows the percentage of people indicating different levels of preference for the various cola brands. Ignoring the NET category, which is just a total, are the other categories ordered meaningfully? Only partly. Five categories are in a sensible order: *Love* > *Like* > *Neither* > *Dislike* > *Hate*. However, the ordering of the categories in the table below implies that people who have said *Don't know* like the brands even less than those that said *Hate.* This is unlikely to be true.



%	%
Don t Know	4%
Hate	13%
Dislike	19%
Neither like nor dislike	23%
Like	27%
Love	15%
NET	100%

In situations where the categories are not ordered from lowest to highest, there are a variety of solutions, including:

- Treating some of the categories as *missing data*. In this case, this is the appropriate solution for *Don't know.*
- Merge categories. For example, if we had a poorly worded scale that included *Good* and *OK,* we would best merge them into a combined category, as their order is unclear.
- Reorder the categories so that their order is consistent with the ordering in terms of performance. There's a little trap for beginners here: in the world of survey research, in the case of a table like the "lower" values are at the top of the table and the higher values at the bottom.[2]
- Restructure the data. For example, if our outcome variable records which brand each person preferred, we can recode the data to a series of binary outcome variables, with one for each brand indicating whether a person chose it or not. Then, we can stack the

---

[2] The logic of this is that just about all data analysis software sorts the categories, from lowest to highest, when creating tables.

data again and proceed with the standard techniques (i.e., so it becomes doubly-stacked).[3]

Reordering the outcome variable in Displayr

1. Create a table of the variable set
2. Reorder by dragging and dropping the categories.
3. For any Don't know categories, right-click on the categories and select **Delete**, which will recode them as missing data.

Reordering the outcome variable in Q

1. Create a table of the question
2. Reorder by dragging and dropping the categories.
3. For any Don't know categories, right-click on the categories and select **Remove**, which will recode them as missing data.

Recoding numeric values

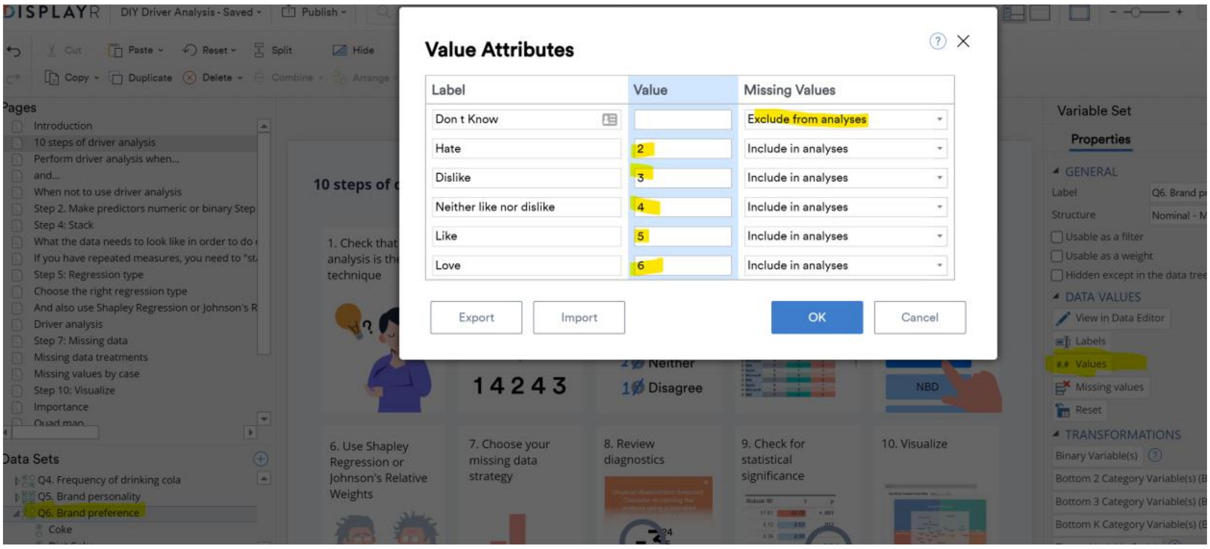
Survey data is stored with numeric values associated with each category. For example, a person may have chosen the option of “0 Not at all likely”, but this may be stored in the underlying data with a value of 1. This underlying value is is the value used when numeric variables are used in driver analysis. The process of changing the values is referred to as *recoding.*

Recoding values in Displayr

1. Select the variable sets.
2. Click **Values** in the object inspector.
3. Change the values in the **Values** column.
4. Set any categories that should be treated as missing data to **Exclude from analysis**.

---

[3] This violates a technical assumption of all the models we have considered which is that the residuals are independently and identically distributed. There are ways of avoiding this (e.g., estimating a multinomial logit model), but they are both beyond the scope of this e-book and are often problematic in that such models have yet to be developed that deal with correlations among the predictors).



Label	Value	Missing Values
Don t Know		Exclude from analyses
Hate	2	Include in analyses
Dislike	3	Include in analyses
Neither like nor dislike	4	Include in analyses
Like	5	Include in analyses
Love	6	Include in analyses

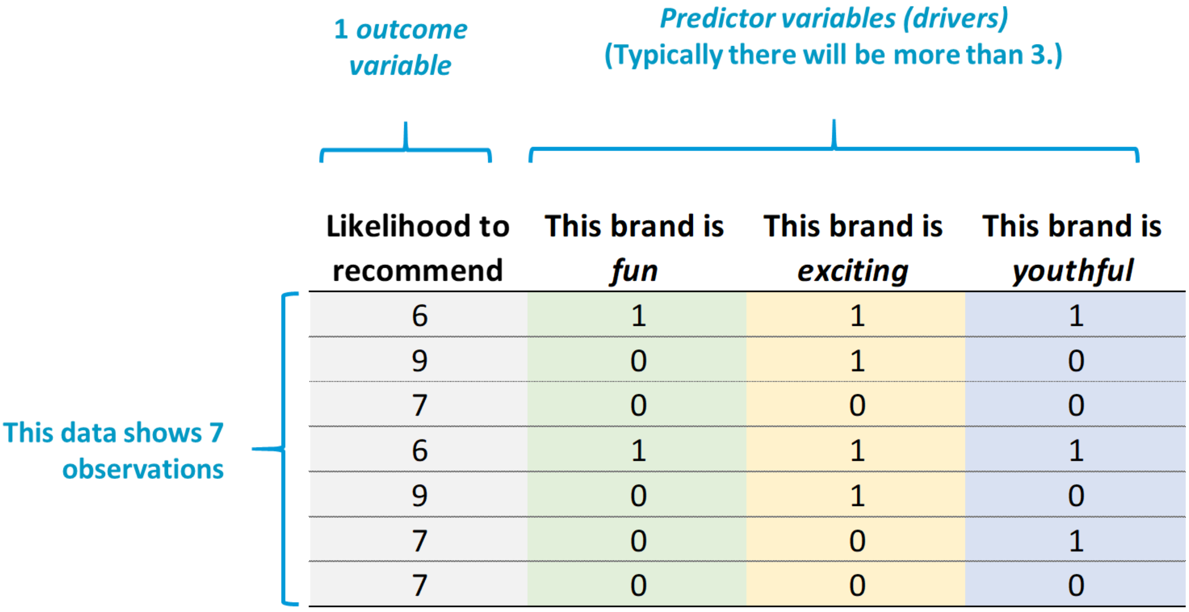
Recoding value in Q

1. Go to the **Variables and Questions tab**.
2. Press the … button for the variable you wish to recode.
3. Modify the values.
4. Check **Missing Data** if appropriate.

4. Stack or use auto-stacking software (if you have repeated measures)

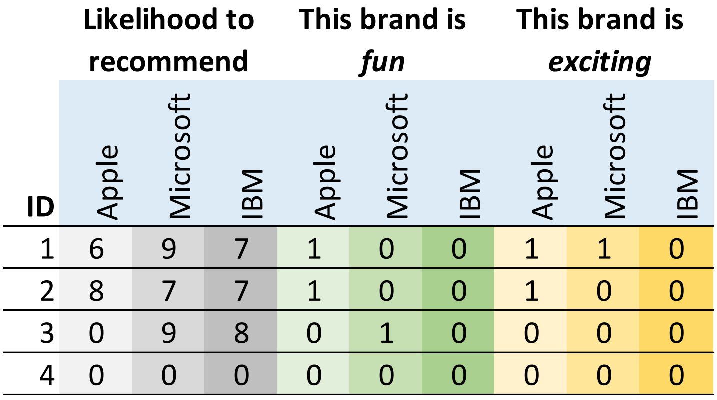
> Where each respondent has provided outcome data and predictor data for two or more brands, we need to *stack* the data to perform the driver analysis. In Displayr and Q this can be automated as a part of the driver analysis itself.

The underlying math of all the main techniques for performing driver analysis assumes that there is one variable (i.e., column) containing the outcome variable and one variable for each predictor variable. There is a single outcome variable in the example below, which is a rating of how likely people were to recommend a brand on a scale of 0 to 10. There are three predictor variables, measuring whether the survey respondent indicated the brands were *fun*, *exciting*, or *youthful* (a 1 means they considered the brand to be like this, and a 0 indicates they did not).


1 outcome variable
Predictor variables (drivers)
(Typically there will be more than 3.)
This data shows 7 observations

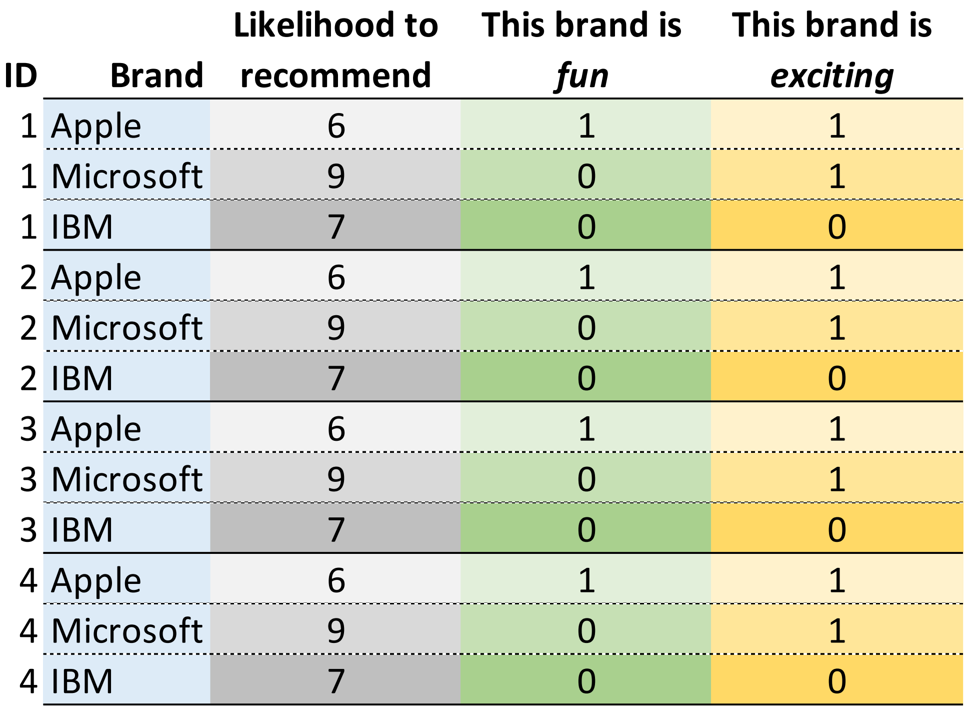
Likelihood to recommend	This brand is *fun*	This brand is *exciting*	This brand is *youthful*
6	1	1	1
9	0	1	0
7	0	0	0
6	1	1	1
9	0	1	0
7	0	0	1
7	0	0	0

However, often when we conduct driver analysis, we wish to perform the analysis using the data for multiple brands at the same time, as then the driver scores can be interpreted as indicating what is important for the whole market. Typically, if we have the data for multiple brands, it will be in a *wide* format, where we have one row of data for each survey respondent. As an example, the table to the right shows data for three brands.



	Likelihood to recommend			This brand is *fun*			This brand is *exciting*		
ID	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

Survey data is usually in the wide format due to the convention of recording each respondent's data in a separate row of the data file. The required fix is to rearrange the data by *stacking* the data for each brand on top of each other brand. For the example above, this means that we end up with three rows of data for each survey respondent, as shown in the table below.



ID	Brand	Likelihood to recommend	This brand is *fun*	This brand is *exciting*
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

Auto-stacking

Both Q and Displayr will automatically stack data at the same time as performing the driver analysis. To do this, we need to:

2. Ensure that the outcome and predictor variables are set up appropriately (see the previous chapter).
3. Choose to do a driver analysis.
    a. In Q: **Create > Regression > Driver Analysis**
    b. In Displayr: **Anything > Advanced Analysis > Regression > Driver Analysis**
4. Check **Stack data**[4]**.**

Stacking

An alternative to auto-stacking is to create a new data file that is stacked, and then use this in the analyses. Both Q and Displayr have tools that can be used to do this, but in general auto-stacking is preferable as it is faster and there's less of a chance of user error.

---

[4] The stacking performed on Displayr and Q is equivalent to the stacking shown in the table above. It only uses a different ordering that gives mathematically equivalent values for all computed outputs. It orders by Brand first and then by ID instead of ID first and Brand that is shown in the table.

5. Choose the right regression type

> Many potential outcome variables can be used when conducting a driver analysis, from five-point rating scales to utilities from conjoint studies. To perform a valid driver analysis, we need to select an appropriate *regression type* for our data.

The table below lists the various ways of measuring outcomes, and the correct regression type[5] for each, as well as how to create these models in Displayr and Q.

Outcome	Examples	Regression type (i.e., the model for the *generalized linear model, GLM)*	Displayr/Q instructions
Two categories	1: This is a brand I buy 0: This is a brand I do not buy	*Binary logit* (also known as logistic regression)	**Anything > Advanced Analysis/Create > Regression > Binary Logit**
Three to 11 ordered categories	1: Hate 2: Dislike 3: Acceptable 4: Like 5: Love	*Ordered logit*	**Anything > Advanced Analysis /Create > Regression > Ordered Logit**
12 or more ordered categories	How would you rate your happiness on a scale of 0 to 100 ______	*Linear regression*	**Anything > Advanced Analysis /Create > Regression > Linear Regression**
Net Promoter Score (NPS)	-100: Detractor 0: Passive/Neutral 100: Promoter	*Linear regression*	**Anything > Advanced Analysis /Create > Regression > Linear Regression**
Purchase or usage quantities	Number of cans of coke consumed per week	*NBD* (or, if you get a weird message, *quasi-Poisson regression*)[6]	**Anything > Advanced Analysis /Create > Regression > NBD/Quasi-Poisson Regression**
Utilities from a conjoint or MaxDiff study	That is, a variable that contains the estimated utilities for each respondent	*Linear regression*	**Anything > Advanced Analysis /Create > Regression > Linear Regression**
Probabilities or shares from conjoint and MaxDiff studies	That is, a variable that contains the probabilities for each respondent	Use the utilities instead of the preference shares, and then use linear regression[7]	**Anything > Advanced Analysis /Create > Regression > Linear Regression**

Choosing the regression type in Q and Displayr

As discussed in the previous chapter, if you need the da

1. Choose to do a driver analysis.
    - **a.** In Q: **Create > Regression > Driver Analysis**
    - **b.** In Displayr: **Anything > Advanced Analysis > Regression > Driver Analysis**

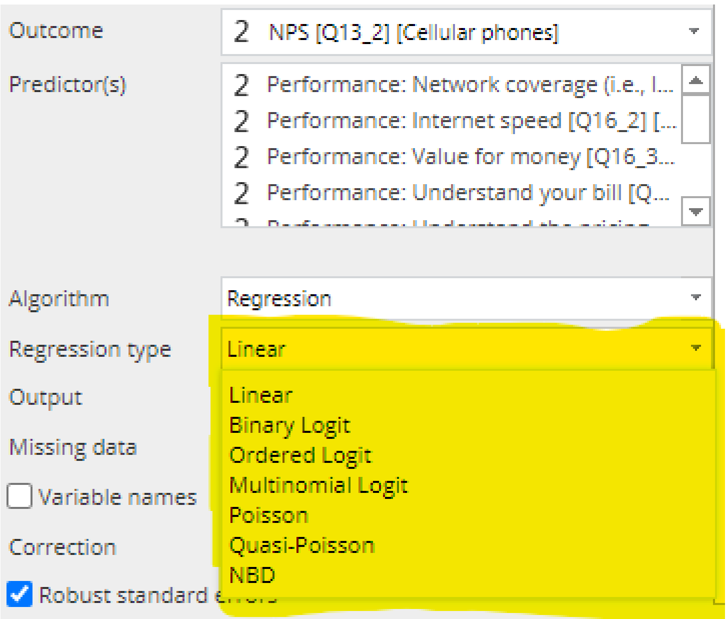
---

[5] Or, to be more precise, the statistical model being referred to is the assumed distribution and link function of the outcome variable in a generalized linear model (GLM), and the NBD and ordered logit are not technically GLMs.

[6] Poisson Regression will give you correct importance scores, but the assumptions of the significance tests are virtually never met with the type of data used in driver analyses.

[7] Due to the way that preference shares and probabilities from MaxDiff and conjoint studies are computed, driver analysis models fitted to such data suffers from higher-order serial correlation in the residuals.

2. Choose the **Regression type:**



Outcome: NPS [Q13_2] [Cellular phones]

Predictor(s):
Performance: Network coverage (i.e., l...
Performance: Internet speed [Q16_2] [...
Performance: Value for money [Q16_3...
Performance: Understand your bill [Q...

Algorithm: Regression

Regression type: Linear

6. Use Shapley Regression or Johnson’s Relative Weights

> Traditional statistical models, such as linear regression, suffer from various problems when applied to driver analysis. These problems are addressed by using either Shapley Regression or Johnson’s Relative Weights.

The standard statistical models,[8] such as linear regression, binary logit, and ordered logit modes, can give the wrong answers when applied to driver analysis in a number of different situations:

- Where there are correlations between predictors. In particular, if there are very high corrrelations, there is a problem known as *multicollinearity.*
- When there are different scales or scale usage in the predictor variables.
- When there is noise in the data, they can be overly sensitive to it, causing the results to be quite inconsistent.
- When there is uncertainty about which predictors to use in the model.

Two related techniques, Shapley Regression[9] or Johnson's Relative Weights,[10] correct for these problems, so it is standard practice to use these techniques when performing driver analysis.

High correlations between predictor variables

Traditional statistical models, like linear regression, are difficult to interpret correctly when predictor variables are highly correlated. The basic problem can be understood with an analogy.

If you follow sport, you will know that people of African descent tend to dominate most athletics. Why? Perhaps they have genetic advantages. Maybe it is because people from African backgrounds have fewer opportunities in other areas, focusing on sport instead. Maybe it is some combination. It is hard to know because there are correlations between these things, so disentangling them becomes problematic.

Statistical techniques of all kinds struggle to work out answers to questions like this. It is no fault of the techniques. It is complex to disentangle such relationships from data, and no amount of complex math can easily resolve such questions. This is why experiments, rather than statistical analysis, are the gold standard in most areas of science.

Traditional statistical methods are particularly problematic when the predictors are highly correlated. The greater the correlation between the predictors, the more these techniques pick up random patterns in the data (hence the possibility of incorrect signs, as discussed in the previous chapter). What this means in practice is that these GLMs become highly unreliable when there are strong correlations between predictors. An analysis based on the estimated coefficients from one data set may end up drawing radically different conclusions to another done on a near-identical data set.

If you are an experienced statistician with a lot of time, you can take all this into account by looking at the correlations between the predictors and the standard errors of the estimated coefficients.

---

[8] When we refer to *traditional statistical models* we are referring to the Generalized Linear Model and similar related models, such as Multinoial Logit, Ordered Logit, and NBD Regression.

[9] Lipovetsky, S. and Conklin, M. (2001). Analysis of regression in game theory approach. Applied Stochastic Models in Business and Industry, 17(4):319–330.

[10] Johnson, J.W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. Multivariate behavioral research 35, 1-19, and, Johnson JW, Lebreton JM (2004). "History and Use of Relative Importance Indices in Organizational Research." Organizational Research Methods, 7, 238–257.

However, the more practical solution is usually to use techniques that are more reliable when there are high correlations, such as Shapley Regression and Johnson’s Relative Weights.

These techniques do not solve the problem of correlated predictors. Rather, they ensure that you get stable results in the presence of correlated predictors. At a conceptual level, the way they do this is by computing importance scores as a weighted average of predictors, where the weights are determined by the extent of intercorrelation between the predictors.

The scale of predictor variables

When we use a traditional statistical model, it implicitly assumes that the predictors are on the same scale. This issue is a little more nuanced than it may first appear.

Consider the case where one predictor is on a 10-point scale, and the other on a 5-point scale. If this occurs, the estimated coefficients generally shouldn’t be directly compared.

A straightforward remedy to this problem is just to recode the predictors so that they have the same range (e.g., recode the 5-point scale, so that a 1 → 1, 2 → 3.25, 3 → 5.5, 4 → 7.75, 5 →10).

However, this won’t necessarily solve the problem, as even if the data is transformed so that the possible ranges of the predictors are equivalent, the usage of the scales may not be consistent. Consider the following two predictors, where they are rated on a scale of Strongly Disagree (1) to Strongly Agree (5):

- “The prices for AT&T are the best value in the world”
- “Call quality for AT&T is satisfactory”.

It is very likely that the first of these statements is going to elicit a lot of ratings of 1 and 2, whereas we would expect a greater spread of the ratings for call quality. The consequence of this is that even though the same scales were shown to the respondents in the questionnaire, in reality the data is on different scales and steps need to be taken to fix this.

A common attempt to solve this problem is to divide each of the values by their standard deviation (i.e., to *normalize* or *standardize* the variables). This also doesn’t necessarily solve the problem. Consider instead if the questions were worded so that they were not so obviously different in the scale they were encouraging, such as:

- “The prices are acceptable”
- The call quality is acceptable”

If with this data, there is a much larger variation in the price predictor this would likely reflect that in the market there is much more variation in perceptions about price, so if we standardize the data by dividing by the standard deviation, we have the unfortunate consequence of changing the estimated importance of the variables.

One solution to this problem is to not use GLMs, and to instead use models that seek to quantify how well each of the predictors explains the variation in the outcome variable.[11] The two main ways of doing this are to use *Shapley Regression* and *Johnson's Relative Weights.*

Sensitivity to noise in the data

Even when predictors are not correlated, the conclusions of standard statistical models can be quite inconsistent due to whatever other noise is in the data. In particular, sampling error leads to situations where one survey can get a result that cannot be replicated in future studies.

One solution to this problem is to use *shrinkage estimators,* which explicitly take into account that it is likely that the estimates of a traditional model will include some over and under estimates of importance.

Shapley Regression and Johnson's Relative Weights are shrinkage estimators.

Uncertainty about which predictors to use in the model

Sometimes it is not clear when predictors should and should not be used in an analysis. As discussed in the next section, Shapley Regression and Johnson's Relative Weights also address this issue.

How Shapley Regression works

Shapley Regression has been invented many times by different people and is known by a variety of other names, including LMG,[12] Kruskal analysis,[13] and dominance analysis[14], and incremental r-squared analysis.

The first step with Shapley Regression is to compute linear regressions using all possible combinations of predictors[15], with the R-squared statistic being calculated for each regression. For example, if we have three predictors, A, B, and C, then 8 linear regressions are estimated with the following combinations of predictors (hypothetical R-Squared statistics are shown in brackets):

---

[11] A common misunderstanding is that this what GLMs see to do. However, the coefficients estimated by GLMs are the marginal effect of a 1 unit increase in the predictor variables.

[12] After the authors of Lindeman RH, Merenda PF, Gold RZ (1980). Introduction to Bivariate and Multivariate Analysis. Scott, Foresman, Glenview, IL.

[13] Kruskal, William. (1987). Relative Importance by Averaging Over Orderings. American Statistician, 41, 6-10. Kruskal describes two techniques, one of which is the same as Shapley. Unfortunately, it is not uncommon for researchers to say they used "Kruskal" without being clear about which of the two techniques they refer to.

[14] Budescu, DV (1993), Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression, Psychological bulletin 114 (3), 542-551.

[15] In practice, a clever mathematical shortcut is used so there is no need to compute all the regressions. See Kruskal, William. (1987). Relative Importance by Averaging Over Orderings. American Statistician. 41. 6-10.

- No predictors (other than the intercept; R-Squared: 0)
- A only (R-Squared = .4)
- B only (R-Squared = .2)
- C only (R-Squared = .1)
- A and B (R-Squared = .4)
- A and C (R-Squared = .5)
- B and C (R-Squared = .3)
- A, B, and C (R-Squared = .5)

Before looking at the computations of Shapley, take a second to think about a simple approach to computing importance, which is to just compare the regressions with a single predictor. This would lead to the conclusion that A is twice as important as B, which is twice as important as C.

Back to Shapley. For each predictor, we compute the average improvement that is created when adding that variable to a model. In the case of predictor A:

- It adds .4 when added on its own (i.e., the A only model minus the No Predictors model).
- .4 - .2 = .2 when added to B (i.e., model A and B less model B)
- .5 - .1 = .4 when added to C
- .5 - .3 = .2 when added to the regression with B and C.

Shapley Regression uses a weighted average of these numbers: 1/3 (.4) + 1/6 (.2) + 1/6 (.4) + 1/3 (.2) = .3. Thus, we can say that the average effect of adding A to a model is that it improves the R-Squared statistic by .3.

The logic of the weights of 1/3, 1/6, 1/6, and 1/3, is that we equally weight the regressions based on the number of possible models. That is, we have two regressions where A is used with one other variable, but only one model where A is estimated on its own, and only one where A is used with two other variables, so this weighting scheme takes this into account, so that the weighting for models with a single predictor are 1/3, with two predictors are 1/3, and with 3 predictors are 1/3. For all three predictors we then have their average incremental improvement being:

- A: .3
- B: .1
- C: .1

Lastly, we re-base these so that they sum to 100%;

- A: 60%
- B: 20%
- C: 20%

This analysis has changed our conclusions about the relative importance of the variables. The earlier simple analysis found that B was twice as important as C, but the Shapley Regression shows they are equally important. Why? Whenever C is added, it increases the R-Squared, whereas for B, it has no effect when A is already in the model.

Shapley Regression versus Johnson's Relative Weights

Both Shapley Regression and Johnson's Relative Weights address the same problems.

The theoretical basis and math of the two methods are entirely different. Shapley Regression calculates lots of linear regressions with varying subsets of predictor variables. Johnson's Relative Weights is an orthonormal rotation of the predictor variables.

Nevertheless, the two methods give essentially identical results, so there is no need to understand the intricacies of the two methods in order to choose one.

There are, however, two practical differences between the methods:

- Shapley Regression is only applicable for linear regression.[16] Johnson's Relative Weights is applicable for any GLM.
- Johnson's Relative Weights is much faster to compute. Once you get past 30 predictors, it can be impractical to use Shapley Regression.

For these reasons, our preference is always to use Johnson's Relative Weights. However, as Shapley Regression is better known and some academic papers are critical of Johnson's Relative Weights, the safer option is to use Shapley Regression whenever your data is linear.

Estimating in Q and Displayr

By default, Q and Displayr will perform Johnson's Relative Weights whenever you select the option of completing a driver analysis.

To change to Shapley Regression, change **Output** from **Relative Importance Analysis** to **Shapley Regression.**

Caveat: It's not always the case that you should use these techniques

Shapley Regression and Johnson's Relative Weights both provide solutions to the problems of correlation between the predictors and the issue of the predictors being on different scales. Are they the best approach and should they always be applied? Probably not. But, they do seem to be the best approach that can be relatively safely applied, with more sophisticated approaches (e.g.,

---

[16] It is possible to use the same type of logic of Shapley Regression to non-linear regressions, using a pseudo-R-square (e.g., McFadden's Rho Square) in place of the R-Square. However, there is no theoretical basis for doing this, as these pseudo-r-squares do not have the property that when a variable is added they always increase.

regularization via Bayesian methods) being considerably more complex and time consuming to implement.

7.Select a proper technique for missing values

> if there are missing values in the predictors, it’s necessary to diagnose why they are missing and then choose the appropriate missing value treatment.

The table below describes the key ways of dealing with missing data. You can choose most of these from the **Missing data** settings with the regression and driver analyses in Q and Displayr.

It is important to appreciate that some of the ways of dealing with missing data, such as mean replacement, random forest, hot-decking, and standard imputation methods, are in widespread use because they are helpful in many different contexts. However, these methods are never valid when the goal is to perform regression-related methods, such as driver analysis.[17]

Missing data treatment	How it works	Key assumption(s)	Limitations
Use partial data (pairwise correlations)	Where the outcome and predictor variables are all numeric, it is possible to calculate the correlations between all the variables and compute the regression from these correlations	The data is believed to be missing due to essentially random events (e.g., randomization is used so that people are only asked a subset of questions). To use jargon, the data is *Missing Completely At Random (MCAR)*	• The key assumption is rarely true in practice. • Only works when the outcome variable and predictor variables are all numeric • Makes some complicated mathematical assumptions that may not be satisfied (if they are not satisfied, you will get a weird error, so you don't need to understand the assumptions)
Exclude cases with missing data (also known as *complete case analysis* and *listwise deletion*)	Any observations (rows) in the data with missing values on any of the predictors or outcome variable are automatically excluded from the analysis	Same as above	Same as above and can result in very small and sample sizes, further reducing the reliability of analyses. This approach is rarely appropriate in in driver analysis.
Replace missing data with means	The mean value is computed for each predictor. Any respondent with a missing value for that predictor is assigned the mean value.	NA	This method is invalid. A key determinant of the importance of a predictor in driver analysis is how much variation there is in a predictor variable and whether this variation is correlated with the variation in the outcome variable. This method reduces the variation in the predictor variables and its correlation with other variables in a way that is different for each predictor variable (thereby biasing the estimates of importance)

---

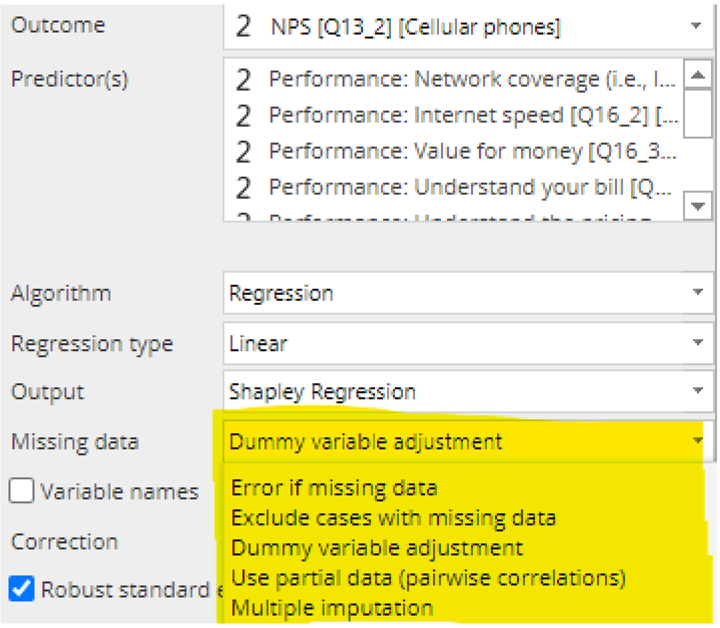
[17] These methods all reduce the variance of the variables used in the regression, and the variance is a key determinant the importance of the predictors.

Missing data treatment	How it works	Key assumption(s)	Limitations
Imputation using GLMs (e.g., linear regression) and the best-fit estimates from other models (e.g., CART, Random Forest)	A model is used to predict the missing values or each outcome variable. The predicted value is used in place of the missing value.	• The non-missing data contains enough information to adequately predict the missing values. The jargon for this is that the data is *Missing At Random (MAR)* • There is no noise in the data (i.e., the missing values can be predicted with complete accuracy)	These assumptions are never met in practice. As a result, this method suffers from the same limitation as *Replace missing data with means,* although will typically be not quite as wrong
Stochastic imputation (e.g., using GLMs, hot decking, latent class analysis, CART, random forest)	A model is used to predict the distribution of the missing values (e.g., if predicting a 5-point rating scale, it may predict that there is a 1% chance of the value being a 1, a 15% chance of a 2, and so on). A random number generator is used to choose the value to replace the missing data, where the probability of selection is based on the estimated probabilities.	• The data is *Missing At Random* (see above) • The sample size is sufficiently large so that the random selection does not add significant noise to the estimated importance scores	• Testing these assumptions is very time consuming and complicated (e.g., conduct the driver analysis 10 times, each time with different randomly generated results and see how different the results are). • All resulting statistical tests are invalid, as they fail to account for the randomness used in imputing the missing values For these reasons *multiple imputation* is always preferable.
Multiple imputation	This is the same as stochastic imputation, except that the process is repeated multiple times, each with different randomness being used to replace the missing values. The driver scores that are estimated are then based on the average of the results of each of the analyses.	The data is *Missing At Random* (see above)	The *Missing At Random* assumption may not be correct.
Dummy variable adjustment	Each missing value is replaced with some arbitrary value (e.g., its mean), and a binary variable is added to the model which takes a value of 1 where the data is missing, and a value of 0 where it is not. Then, the model is estimated and the importance scores are calculated from the predictors (and the coefficients of the binary variables are ignored).	The data is missing because it cannot exist. For example, if the predictor is satisfaction with telephone banking, and the person has no data because they have never used telephone banking.	It is often hard to verify why data is missing.

It is recommended to use dummy variable adjustment if it’s assumptions are met and, failing that use partial data/pairwise correlations if it’s assumptions are met, and, failing that, use multiple imputation.[18]

Choosing the missing data setting in Q and Displayr

These options are selectable from the **Missing data** option:



Outcome: 2 NPS [Q13_2] [Cellular phones]

Predictor(s):
2 Performance: Network coverage (i.e., l...
2 Performance: Internet speed [Q16_2] [...
2 Performance: Value for money [Q16_3...
2 Performance: Understand your bill [Q...

Algorithm: Regression

Regression type: Linear

Output: Shapley Regression

Missing data: Dummy variable adjustment

☐ Variable names

Correction

☑ Robust standard e

Error if missing data
Exclude cases with missing data
Dummy variable adjustment
Use partial data (pairwise correlations)
Multiple imputation

---

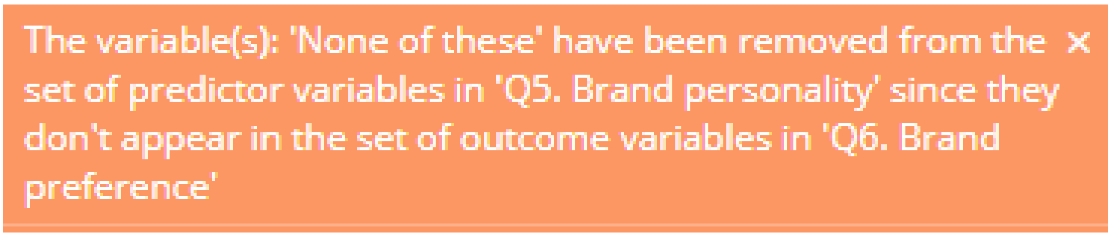
[18] Paul Allison (2001), *Missing Data,* Sage; Rubin, Donald B. (1996) “Multiple Imputation after 18+ Years (with discussion).” *Journal of the American Statistical Association* 91: 473-489, Judea Pearl (2009), Causality, 2nd Edition.

8. Review diagnostics

> There are a series of standard diagnostic tests for reviewing a driver analysis:
>
> - Checking the regression type
> - Studying the signs of coefficients
> - Identifying and investigating the effects of outliers
> - Testing and correcting for multicollinearity
> - Testing and correcting for heteroskedasticity

Techniques like driver analysis make many theoretical assumptions. It can be useful to verify that these assumptions are appropriate.

Displayr and Q automatically perform common checks. Where the checks do not pass, they are shown in orange boxes. For example:



The variable(s): 'None of these' have been removed from the set of predictor variables in 'Q5. Brand personality' since they don't appear in the set of outcome variables in 'Q6. Brand preference'

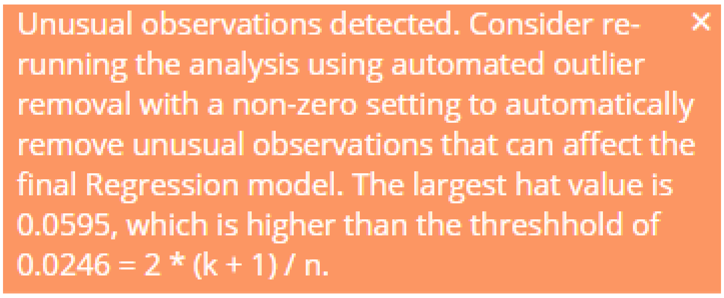
Checking the regression type

Chapter 5. Choose the right regression type discussed how to choose your regression type. Both Q and Displayr have expert systems that will alert you if your choice may be wrong.

Outliers

Just a few weird observations can, in theory, cause a driver analysis to be highly misleading. One solution to this is to use a *robust* method, such as *robust GLMs* and *random forest.* Unfortunately, there are no robust versions of Shapley Regression nor of Johnson's Relative Weights, so in practice we usually need to instead first identify outliers/unusual observations and conduct a sensitivity analysis to assess their impact.

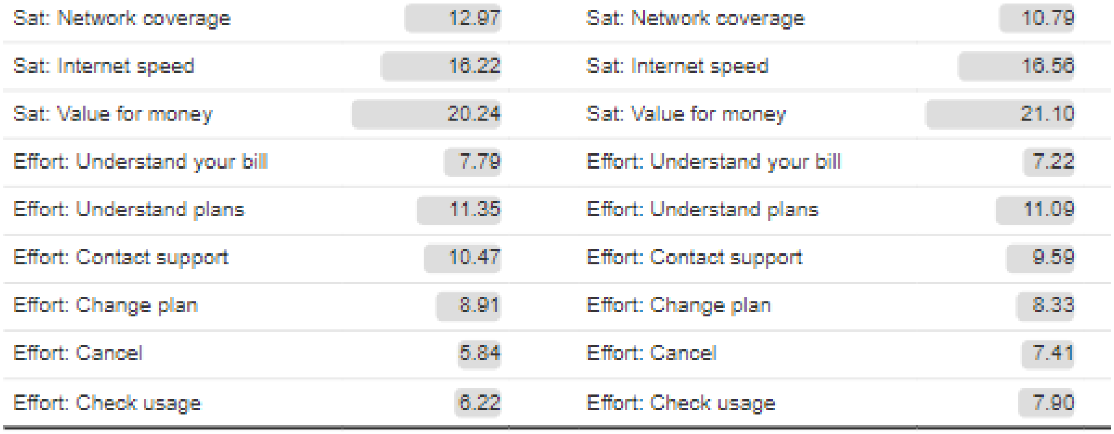
If any unusual observations/outliers are flagged, we can re-estimate the importance scores with these variables excluded and compare the two sets of importance scores. If the results are substantially different, we need to investigate the outliers/unusual observations to get a better feeling about what is going on. Fortunately, it tends to be the case with driver analysis that either the results do not change, or, when they do change it is because the outliers contained obvious data problems (e.g., missing values recorded as -99). This is because driver analysis is typically based on categorical variables, so no aberrant observation can normally ever have a high level of influence over a model.



Unusual observations detected. Consider re-running the analysis using automated outlier removal with a non-zero setting to automatically remove unusual observations that can affect the final Regression model. The largest hat value is 0.0595, which is higher than the threshhold of 0.0246 = 2 * (k + 1) / n.

Q and Displayr automatically look for unusual observations when estimating GLMs and driver analyses, showing a warning like the one to the right.

When this warning appears, the appropriate next step is usually to remove the outliers and see if the results change. You can automatically remove the outliers using the **Automatic outlier removal percentage** setting. For example, the output on the left below is for the full sample, and the one on the right is with 20% of observations removed. While there are some differences, the key conclusions remain unchanged, which tells us we can ignore the issue of outliers.

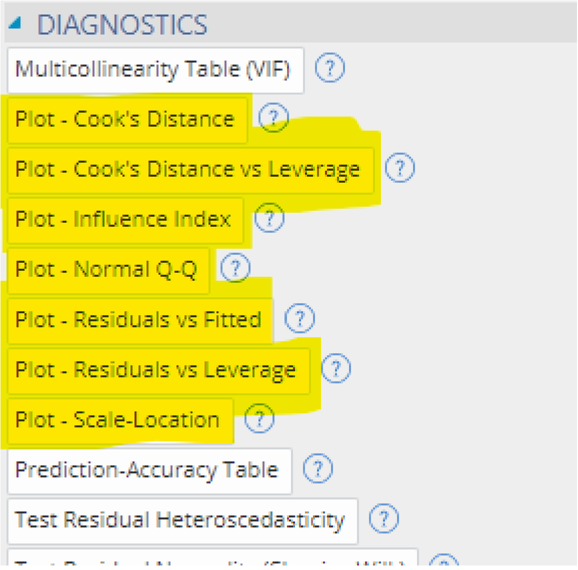


Sat: Network coverage	12.97	Sat: Network coverage	10.79
Sat: Internet speed	16.22	Sat: Internet speed	16.56
Sat: Value for money	20.24	Sat: Value for money	21.10
Effort: Understand your bill	7.79	Effort: Understand your bill	7.22
Effort: Understand plans	11.35	Effort: Understand plans	11.09
Effort: Contact support	10.47	Effort: Contact support	9.59
Effort: Change plan	8.91	Effort: Change plan	8.33
Effort: Cancel	5.84	Effort: Cancel	7.41
Effort: Check usage	6.22	Effort: Check usage	7.90

A more thorough approach is to review the standard diagnostic plots for detecting outliers.

Reviewing diagnostic plots in Displayr

Scroll to the bottom of the driver analysis form (i.e., the place where you choose the data and select your options), and click one of the plot options, and then review the specific observations in the data edtitor.


DIAGNOSTICS
Multicollinearity Table (VIF)
Plot - Cook's Distance
Plot - Cook's Distance vs Leverage
Plot - Influence Index
Plot - Normal Q-Q
Plot - Residuals vs Fitted
Plot - Residuals vs Leverage
Plot - Scale-Location
Prediction-Accuracy Table
Test Residual Heteroscedasticity

Reviewing diagnostic plots in Q

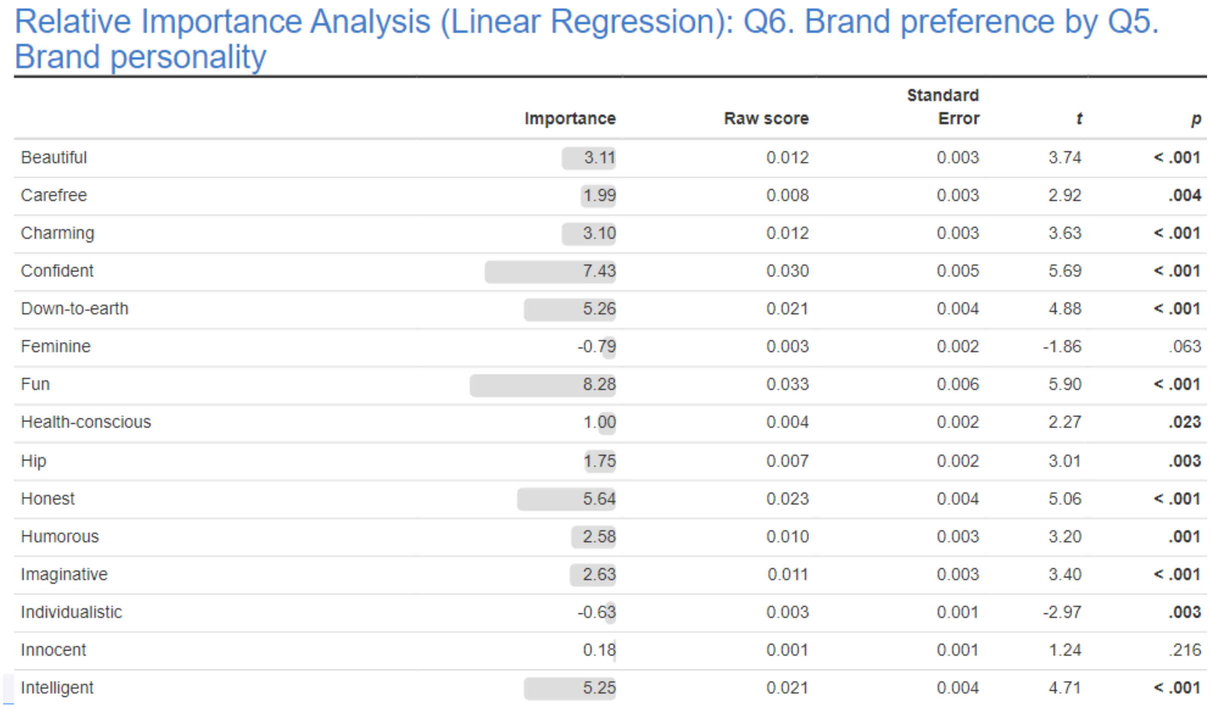
1. Select your model.
2. **Create > Regression > Diagnostic > Plot** and choose your desired plot.

3. Review specific observations/cases in the **Data tab**.

Review the signs of the importance scores

When examining the outputs from a driver analysis, it is natural to assume that if a variable is important, it means that improving performance on that variable will lead to an improvement on the outcome variable. For example, if customer service is important, we interpret this as meaning that improving customer service will lead to an improvement in overall satisfaction.

Sometimes this interpretation will be wrong. If this interpretation is bad, it can lead to severe problems.[19] As an example, the output below is from a driver analysis predicting cola preference based in predictors that measure brand personality. For example, the second highest coefficient is for Confident, telling us that if people consider a brand to be confident, that is a relatively good indicator they will have a stronger preference for the brand. However, there are negative signs for the coefficients of two of the predictors, Feminine and Individualistic, suggesting that these brand associations lead to a reduction in preference. This is potentially a problem; because the more modern techniques all assume a positive effect, we may draw the wrong conclusions.



Relative Importance Analysis (Linear Regression): Q6. Brand preference by Q5. Brand personality

	Importance	Raw score	Standard Error	*t*	*p*
Beautiful	3.11	0.012	0.003	3.74	**< .001**
Carefree	1.99	0.008	0.003	2.92	**.004**
Charming	3.10	0.012	0.003	3.63	**< .001**
Confident	7.43	0.030	0.005	5.69	**< .001**
Down-to-earth	5.26	0.021	0.004	4.88	**< .001**
Feminine	-0.79	0.003	0.002	-1.86	.063
Fun	8.28	0.033	0.006	5.90	**< .001**
Health-conscious	1.00	0.004	0.002	2.27	**.023**
Hip	1.75	0.007	0.002	3.01	**.003**
Honest	5.64	0.023	0.004	5.06	**< .001**
Humorous	2.58	0.010	0.003	3.20	**.001**
Imaginative	2.63	0.011	0.003	3.40	**< .001**
Individualistic	-0.63	0.003	0.001	-2.97	**.003**
Innocent	0.18	0.001	0.001	1.24	.216
Intelligent	5.25	0.021	0.004	4.71	**< .001**

---

[19] The underlying math of both Shapley Regression and Johnson's relative weight ensures that both techniques always report a positive importance score. However, the correct interpretation may often be negative. For example, if you have a study of airlines with an attribute of Delays, it will be shown to have a positive importance score, but the meaning is that the it is important to have fewer delays. In the case of Delays, the correct interpretation is obvious. This is not always the case.

Negative coefficients in Shapley Regression and Johnson’s Relative Weights

The theoretical assumptions of Johnson’s Relative Weights and Shapley Regression assumes a positive importance score. Our experience in working with many customers is that often they make mistakes which mean this assumption is not met, and their entire analyses are invalid. For this reason, both Q and Displayr always compute an appropriate generalized linear model and use the signs from this analysis as the signs for the Shapley Regression and Johnson’s Relative Weights. Consequently, they can show negative signs. The fix for this is:

- Read the previous chapter and follow its instructions.
- Check the **Absolute importance scores** option.

Causes and solutions for negative signs

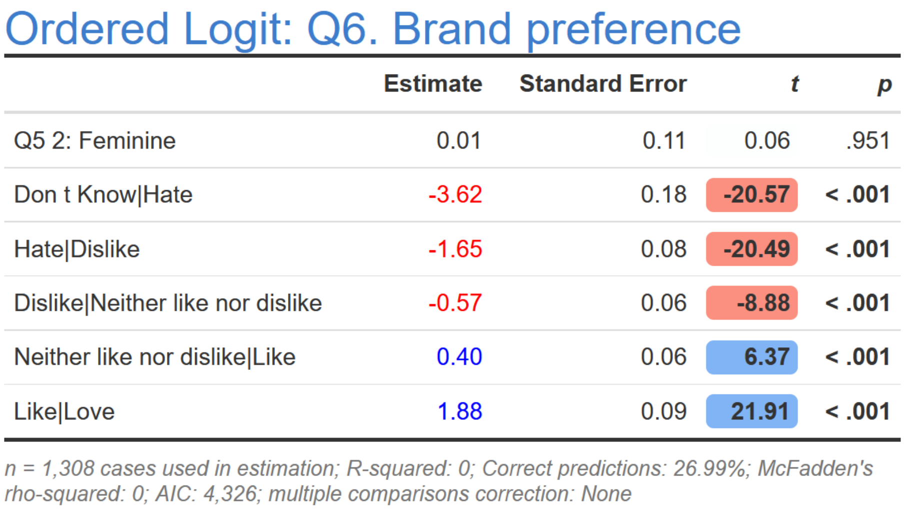
There are several possible different explanations for the negative signs.

The signs are “wrong”

The signs themselves are estimates, and they may be unreliable estimates. There are two reasons why this can occur. The first is that the predictor is unimportant, and whether or not it was negative was something of a coin toss (random event). The second is that the predictor variable is highly correlated with other predictor variables.

The good news is that if either of these explanations is correct, it will show up as a predictor *not* being highly statistically significant. If this occurs, the simplest solution is usually just to either ignore the sign (more about this in **Chapter 6. Use Shapley Regression or Johnson’s Relative Weights**) or treat the coefficient as being 0.

We can gain additional support for the intuition that the estimate of the sign may be wrong by estimating a GLM with only one predictor and checking that it has the correct sign or computing the correlation. For example, the ordered logit coefficient for Feminine is shown below. Its coefficient is essentially zero, consistent with the conclusion that it is not a reliable predictor of brand.



Ordered Logit: Q6. Brand preference

	Estimate	Standard Error	*t*	*p*
Q5 2: Feminine	0.01	0.11	0.06	.951
Don t Know|Hate	-3.62	0.18	**-20.57**	**< .001**
Hate|Dislike	-1.65	0.08	**-20.49**	**< .001**
Dislike|Neither like nor dislike	-0.57	0.06	**-8.88**	**< .001**
Neither like nor dislike|Like	0.40	0.06	**6.37**	**< .001**
Like|Love	1.88	0.09	**21.91**	**< .001**

n = 1,308 cases used in estimation; R-squared: 0; Correct predictions: 26.99%; McFadden's rho-squared: 0; AIC: 4,326; multiple comparisons correction: None

If the data is being auto-stacked, it means you will have selected all the variables together at once, so cannot just select the one predictor of interest. The workaround to this is to:

- Duplicate the predictors.
- Create a table of the grid question.
- Hide all the predictors except the one of interest
- Select the modified question/variable set in the GLM.

The variables have been incorrectly coded

The predictor or outcome variable have been coded incorrectly, and higher values indicate worse performance. If this has occurred, the remedy is to recode the data correctly (see **1.** and Error! Reference source not found.).

It is not appropriate to perform driver analysis

Driver analysis assumes that the predictors cause the outcome variable and that each has an independent and separate linear effect. This may not be true. (If it is not true, there is no ready fix for the problem.)

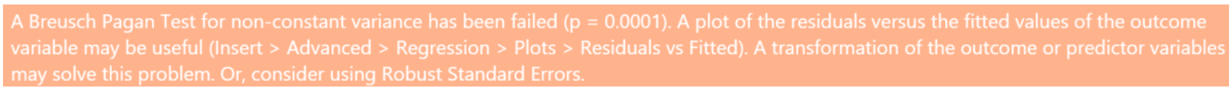
Testing for and correcting for multicollinearity

Displayr and Q calculate VIFs and GVIFs to assess multicollinearity. The remedy is to apply Shapley Regression or Johnson's Relative Weights. As discussed in **Chapter 6. Use Shapley Regression or Johnson's Relative Weights**, these techniques should be routinely used anyway, regardless of whether there is or is not multicollinearity.

Heteroskedasticity

When **Regression type** is set to **Linear**, an assumption is made that there is no pattern in the residuals from the analysis (i.e., the difference between the observed and predicted values). This assumption is known as *homoscedasticity* or *constant variance.* If this assumption is not true, the standard significance test results will be misleading.

Displayr and Q automatically perform a *Breusch Pagan Test* for non-constant variance, and alert you if the test is failed:



A Breusch Pagan Test for non-constant variance has been failed (p = 0.0001). A plot of the residuals versus the fitted values of the outcome variable may be useful (Insert > Advanced > Regression > Plots > Residuals vs Fitted). A transformation of the outcome or predictor variables may solve this problem. Or, consider using Robust Standard Errors.

The warning is alerting you that the fitted model seems to have residuals that have non-constant variance which violates the standard linear regression assumption that the residuals have constant variance.

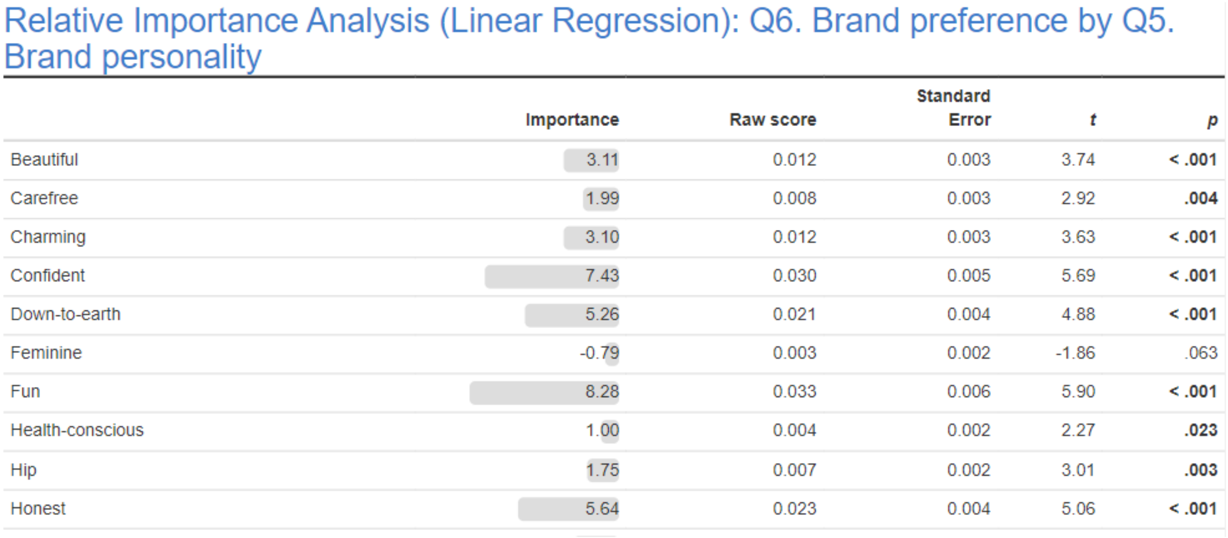
A solution to this is to fit a more general model that allows non-constant variance in the residuals and recompute the statistical tests with non-constant variance in mind. To do this in Q or Displayr when estimating a Linear Regression, with or without Johnson's Relative Weights, select **Robust standard errors.** This allows non-constant variance in the residuals and re-computes the standard errors and resulting significance tests.

9. Check statistical significance

> It is useful to verify that key conclusions are statistically significant.

In the previous chapter, we investigated the negative signs for Feminine and Individualistic. If we look at the *p* column, we can see that neither of them is significantly different to 0 at the 0.05 level, telling us that we probably should not be too concerned by the negative coefficients.

It can also be useful to examine confidence intervals before drawing too strong conclusions based on the rank ordering of coefficients. For example, the output below tells us that Fun is most important, followed by Confident. When we look at the raw scores (which add up to the R-Square), we see their raw scores are .033 and .030, and the difference between these is thus 0.003. This difference is smaller than their standard errors (0.005 and 0.006), telling us that the difference is not even close to being statistically significant at the 0.05 level.[20]



Relative Importance Analysis (Linear Regression): Q6. Brand preference by Q5. Brand personality

	Importance	Raw score	Standard Error	*t*	*p*
Beautiful	3.11	0.012	0.003	3.74	**< .001**
Carefree	1.99	0.008	0.003	2.92	**.004**
Charming	3.10	0.012	0.003	3.63	**< .001**
Confident	7.43	0.030	0.005	5.69	**< .001**
Down-to-earth	5.26	0.021	0.004	4.88	**< .001**
Feminine	-0.79	0.003	0.002	-1.86	.063
Fun	8.28	0.033	0.006	5.90	**< .001**
Health-conscious	1.00	0.004	0.002	2.27	**.023**
Hip	1.75	0.007	0.002	3.01	**.003**
Honest	5.64	0.023	0.004	5.06	**< .001**

---

[20] To test more precisely, see https://wiki.q-researchsoftware.com/wiki/Independent_Samples_t-Test_-_Comparing_Two_Coefficients.

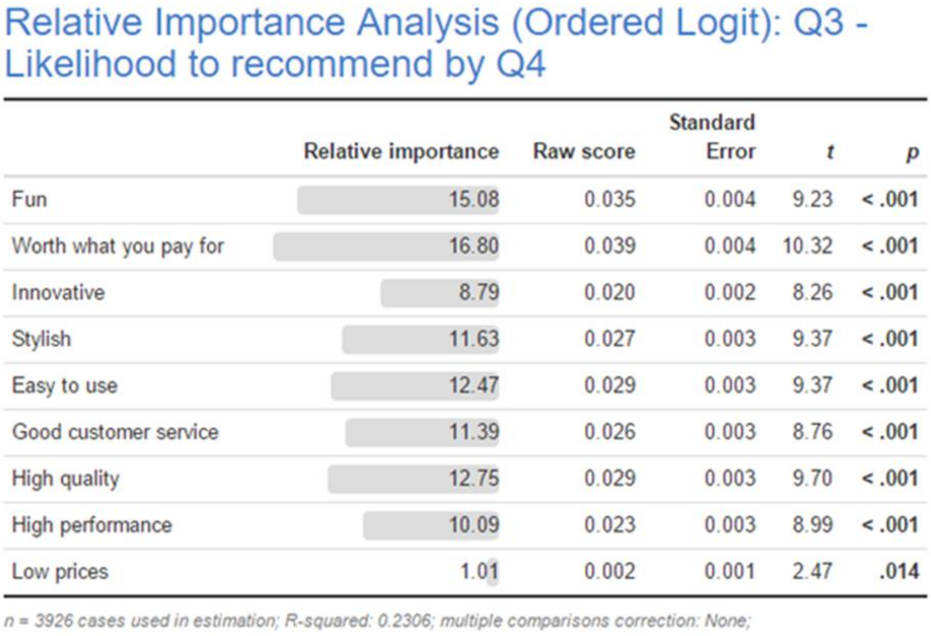
10. Data visualization

> There are typically three main types of data visualizations used in driver analysis:
>
> - Visualizations of importance scores
> - Importance by sub-group
> - Quad-maps / performance-importance scatterplots

Importance scores

The essential output of a driver analysis is a set of importance scores, which are typically made to add up to either the R-Squared Statistic, 1, or 100.

The default output of Q and Displayr, shown below, is intended to show this in a relatively attractive way. However, you can instead plot this as a bar, pie chart, or any other standard visualization by using the **Visualization** menu in Displayr, or, **Create > Charts > Visualization** in Q, and then selecting the driver analysis in **Inputs > DATA SOURCE > Output**.



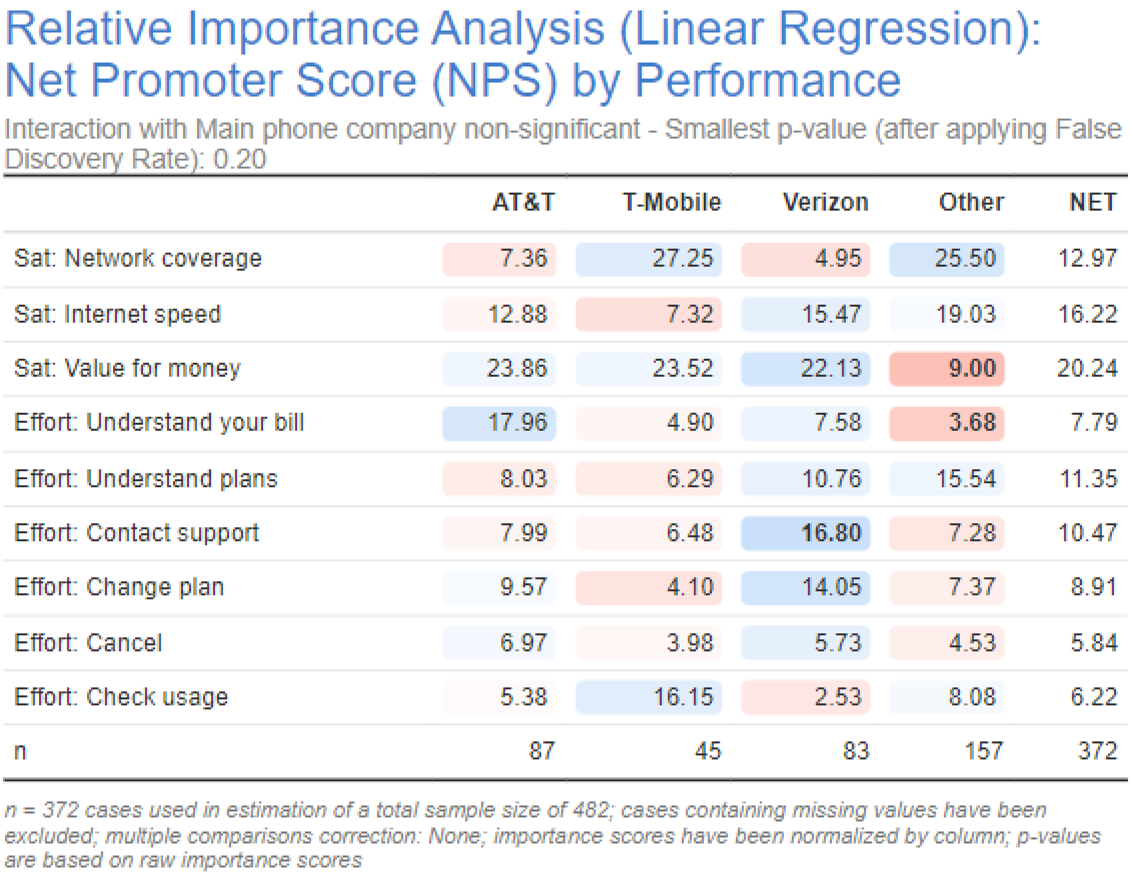
Relative Importance Analysis (Ordered Logit): Q3 - Likelihood to recommend by Q4

	Relative importance	Raw score	Standard Error	*t*	*p*
Fun	15.08	0.035	0.004	9.23	< .001
Worth what you pay for	16.80	0.039	0.004	10.32	< .001
Innovative	8.79	0.020	0.002	8.26	< .001
Stylish	11.63	0.027	0.003	9.37	< .001
Easy to use	12.47	0.029	0.003	9.37	< .001
Good customer service	11.39	0.026	0.003	8.76	< .001
High quality	12.75	0.029	0.003	9.70	< .001
High performance	10.09	0.023	0.003	8.99	< .001
Low prices	1.01	0.002	0.001	2.47	.014

n = 3926 cases used in estimation; R-squared: 0.2306; multiple comparisons correction: None;

Crosstabs

A second approach to visualizing driver analysis is to use crosstabs to compare sub-groups. This approach is also automated in Q and Displayr, where the column variable is selected by choosing the **Crosstab interaction** setting in driver analysis and GLMs. Color-coding highlights differences between sub-groups.



Relative Importance Analysis (Linear Regression): Net Promoter Score (NPS) by Performance

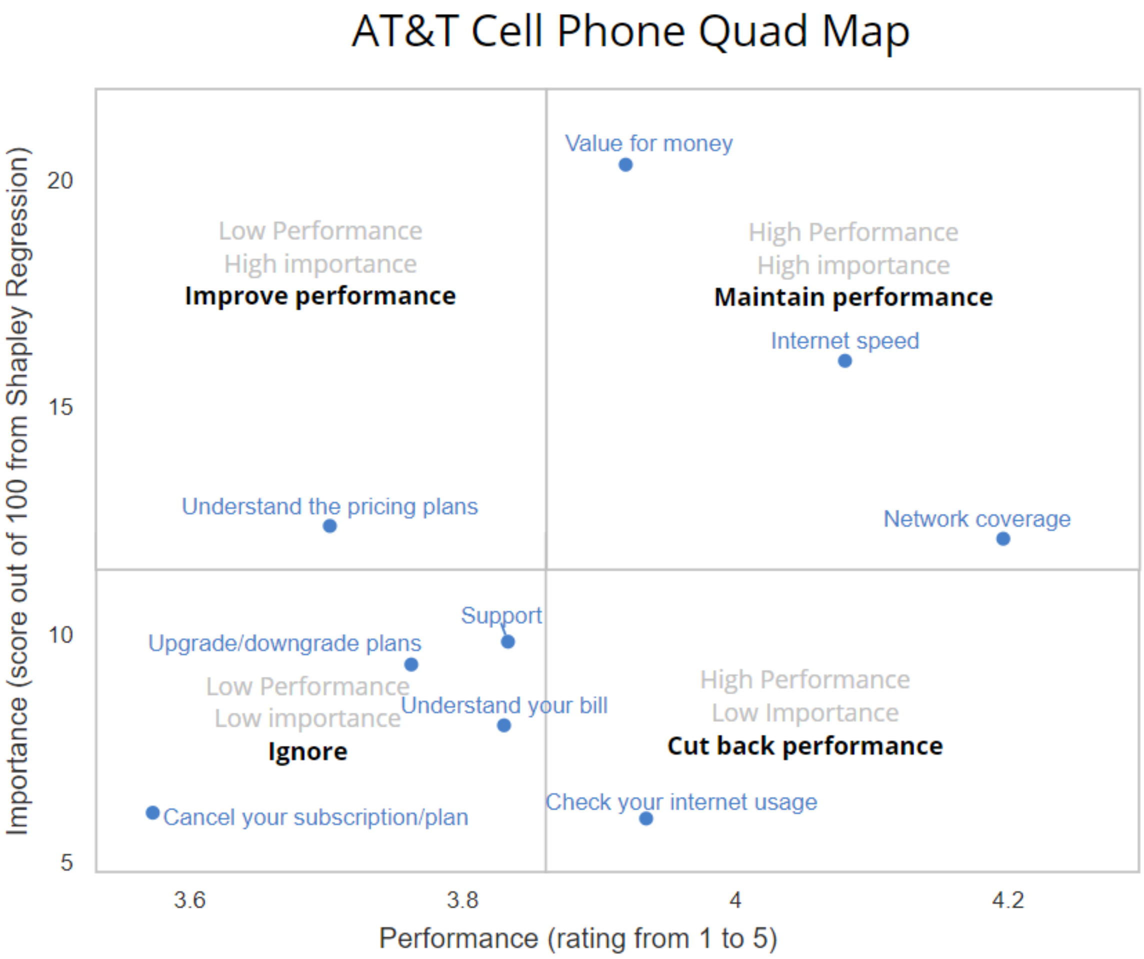
Interaction with Main phone company non-significant - Smallest p-value (after applying False Discovery Rate): 0.20

	AT&T	T-Mobile	Verizon	Other	NET
Sat: Network coverage	7.36	27.25	4.95	25.50	12.97
Sat: Internet speed	12.88	7.32	15.47	19.03	16.22
Sat: Value for money	23.86	23.52	22.13	**9.00**	20.24
Effort: Understand your bill	17.96	4.90	7.58	**3.68**	7.79
Effort: Understand plans	8.03	6.29	10.76	15.54	11.35
Effort: Contact support	7.99	6.48	**16.80**	7.28	10.47
Effort: Change plan	9.57	4.10	14.05	7.37	8.91
Effort: Cancel	6.97	3.98	5.73	4.53	5.84
Effort: Check usage	5.38	16.15	2.53	8.08	6.22
n	87	45	83	157	372

n = 372 cases used in estimation of a total sample size of 482; cases containing missing values have been excluded; multiple comparisons correction: None; importance scores have been normalized by column; p-values are based on raw importance scores

Quad maps / Performance-Importance scatterplots

A *performance-importance* chart, also known as a *quad map,* plots the importance scores by performance. It is known as a quad chart because it is common to overlay a management consulting-style two-by-two matrix on top, breaking it into four quadrants.



Creating quad maps in Q and Displayr

1. Create a SUMMARY table of the performance data (if necessary, combine it into a single question or variable set first).
2. In:
    a. Q: **Create > Charts > Visualization > Scatterplot**
    b. Displayr: **Visualization > Scatterplot**
3. In **X coordinates** select the performance table (or, select the importance table; either one is fine).
4. In **Y coordinates** select the importance table.
5. In **Chart > APPEARANCE** set **Show Labels** to **On chart**.
6. Format it as you want. In the example above I've drawn boxes and text over the top of the visualization.