

Key Driver Techniques

Key Driver Analysis (or Relative Importance Analysis), are regression/correlation-based techniques that are used to discover which of a set of *independent variables* cause the greatest fluctuations in the given *dependent* variable, i.e., which of them have the greatest impact in determining its value. These techniques essentially assign a *weight* to each driver that represents its importance in predicting an outcome variable relative to the others.

For example, in market research satisfaction surveys the dependent variable would usually be a measure of overall satisfaction, whilst the independent variables are measures of other aspects of satisfaction, e.g., efficiency, value for money, customer service etc. The independents in this example are then often called *Drivers of Satisfaction*. By applying a suitable Key Driver technique, then ordering these variables in terms of a measure of importance, a researcher can better understand where a company should focus its attention if it wants to see the greatest impact.

Key Driver Analysis should perhaps be renamed marginal resource allocation analysis in that reallocating all of one's resources blindly based on these results often would lead to problems. Take for example, airline satisfaction. No-one would ever consider cutting safety standards to improve their food, although in Key Driver Analysis, the standard of food would almost invariably come out as more important. This is because with respect to air travel, safety is assumed and hence safety standards do not generally influence people's choice. However, the standard of food on the airline does vary between airlines and thus has a much greater influence on people's satisfaction - and thus their choice of airline. Hence Key Driver Analysis will demonstrate *where one should allocate extra resources if one wishes to see the greatest impact on their satisfaction scores*.

One of the main problems with analysing satisfaction and other similar data, is that independent variables are highly correlated with one another. This is called *multicollinearity* and can result in importance values that are derived from simple regression/correlation analysis being inaccurate and potentially highly misleading.

There are various methods to overcome this problem. Three of the most well-respected statistical techniques are

- Shapley Value Analysis (see for example, [Ulrike Grömping, The American Statistician 2007](#));
- Kruskal's Relative Importance Analysis ([William Kruskal, American Statistician 1987](#));
- Ridge Regression (also known as Tikhonov regularization; see for example https://en.wikipedia.org/wiki/Tikhonov_regularization).

Both Shapley Value Analysis and Kruskal's Relative Importance Analysis are fairly similar in concept: for each independent variable, we derive a measure of the strength of the correlation between itself and the dependent after we have “stripped” out its correlations between the other independent variables. The final importance weight for each variable is the mean of these derived correlations taken over all possible regression models between the dependent and the different possible subsets of the independents.

The difference between the techniques comes in the measure of correlation used in the procedure: *semi-partial correlations* are used in the case of Shapley whereas *partial correlations* are used in the case of Kruskal's. This simple difference leads to a huge difference in computation speeds. Shapley Values can be fairly easily computed whereas Kruskal's is more time consuming. However, since we believe that Kruskal's is more theoretically sound, we have developed an original algorithm that allows both methods to be simultaneously constructed with no extra time-cost.

In both cases, the importance weight for each variable is converted into a percentage of the total sum of the importance scores. This allows easier comparison of the importance weights without any loss of information.

Ridge Regression on the other hand, in effect, penalises the regression coefficient of the independents by a “penalty” factor to neutralise the effect of the multicollinearity, where a penalty value of 0 is equivalent to ordinary multiple linear regression. For each such factor, we define the importance weight of each independent to be the absolute value of the standardised regression coefficient, expressed as a percentage of the sum of all the absolute values of the regression coefficients.

Since the weights returned depend on the penalty factor, it is usual to compute a set of importance weights, then choose which one is most appropriate. A common way to help in your choice is to plot the (standardised) regression coefficients of the independents against the penalty factor, then choose your penalty factor when the graph appears to “flatten” (this is similar to a scree plot in factor analysis). You must be careful that the R^2 does not diminish too much though. Despite the scree plot using the actual standardised regression coefficients, for comparison with the other two methods, we still report the importance measure as a percentage of the absolute values.

JumpData Key Driver Tool

JumpData have developed an easy to use, web-based tool to conduct Key Driver analysis using the three techniques above. It allows the user to import either a .csv or Excel .xlsx file, and run both Ridge Regression and then Shapley and Kruskal's analysis simultaneously. Since standardisation of the independent variables is recommended, our tool automatically standardises the data upon upload.

When conducting Key Driver analysis, we recommend Ridge Regression is run first, to provide an immediate set of results, and a guide to how the final importance weights should look. An appropriate penalty factor is chosen by inspecting the flattening of the scree plot. Following the choice of penalty factor, the more computationally expensive Shapley and Kruskal's analysis are performed, allowing the results of all three methods to be compared. The importance scores are ordinary linear regression are also outputted for completeness (these are calculated as for ridge regression).

The results of the three Key Driver techniques are likely to be fairly similar, with the Kruskal's output being favoured by JumpData as the most theoretically sound way of removing multicollinearity from the dataset. We would recommend using Ridge Regression only as a sense check of the other two techniques.

Having said that, Kruskal's technique is the most computationally expensive of the three. Time taken increases by a factor of just over two for each extra independent variable. So, with 16 independent variables, it may take around 7 seconds to execute, increasing to 35 seconds for 18 variables, 160 seconds for 20 variables, and 800 seconds – close to 13 and a half minutes for 22 variables. (These times were recorded when the program ran locally and so times may be slower on other machines). Remember though, that if you are conducting analysis with 20 variables then the *average* importance is expected to be only 5%, so you may wish to consider whether all of them are actually ever going to be reported on. We recommend, to keep the number of independents to be less than 20 for sake of ease of interpretation of the results (although we have used the toll for over this number of 22 independent variables in the past).

If you do have more than 20 independents, an alternative is pre-analyse the data using Ridge Regression, then look at the base regression important scores and pick out the (20 or less) most important from this. Then you can run both Shapley and Kruskal's on this reduced set of "more important" independents.

Quadrant maps

The results of your key driver analysis can also be visualised on a quadrant map. The x-axis is the importance weight of the driver and the y-axis is its mean, top box or other measure of value. In all cases, we split the scatterplot into the four regions below for ease of interpretation:

- High importance / Low value (*Key weaknesses*)
- High importance / High value (*Key strengthes*)
- Low importance / High value
- Low importance / Low value


Drivers that fall in the two top two quadrants (Q1 and Q2) are those that are most important in determining the value of the dependent. In the case of satisfaction surveys, this is often interpreted as the key weaknesses of the organisation (Q1), or its key strengthes (Q2). Clearly, those in Q1 are the areas where the organisation should be focusing their investment to improve satisfaction, whilst those in Q2 are areas where they need to continue to perform well, to maintain their current satisfaction level.


The bottom two quadrants contain drivers that are less important. Nonetheless, it is still a good idea to be able to visualise strengthes and weaknesses – even for less important drivers.

The tool automatically produces a quadrant map plotting mean scores against Kruskal's weights of the drivers. The data can be downloaded to change the quadrant map as suits the analysis. Often it is more appropriate to use top box as the means are clustered together. These can be easily calculated from the original data set.




We give examples of the output from the tool including the corresponding quadrant map below.

Key Driver Tool User Interface

Key Driver Tool 

 [Download CSV Sample Data](#)
[Download Excel Sample Data](#)

INSTRUCTIONS

UPLOAD FILE  keydriver_sample.csv  Cases 400
Cases in analysis  396

RUN LINEAR REGRESSION

RUN RIDGE REGRESSION

Min penalty factor	<input type="text" value="0"/>
Max penalty factor	<input type="text" value="4"/>
Ridge step	<input type="text" value="0.2"/>

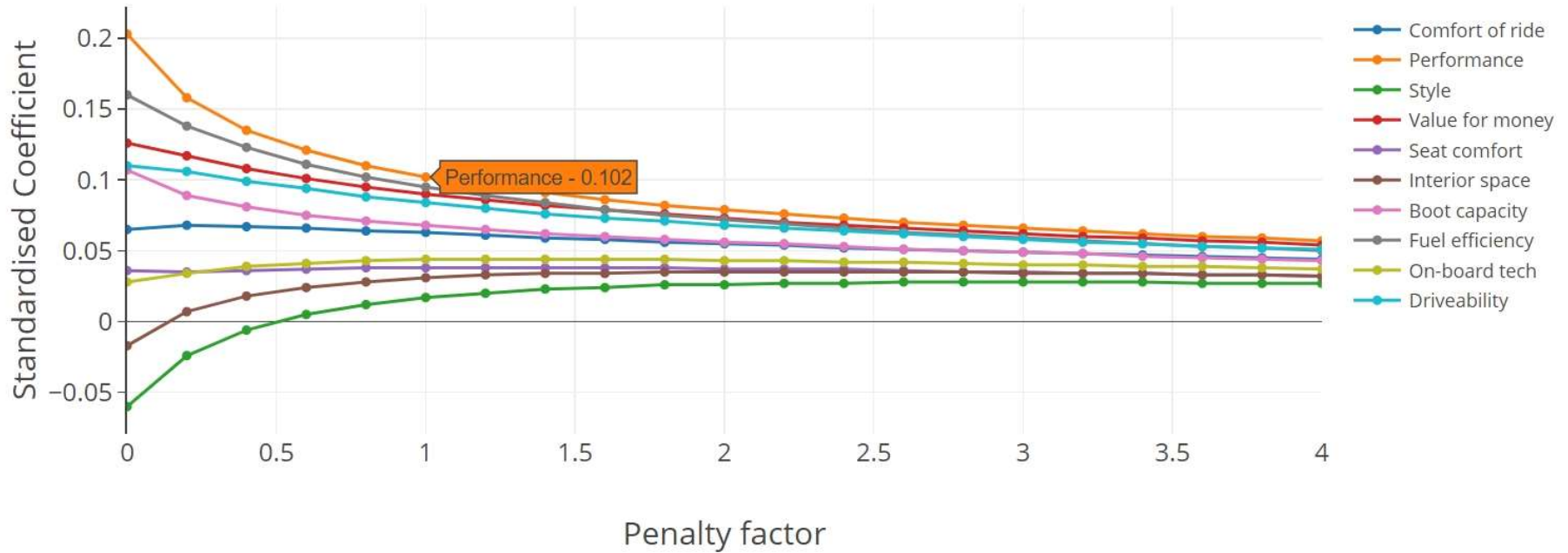
RUN KEY DRIVER ANALYSIS

Example 1: Satisfaction survey of cars in the domestic car market

Here, we use our web-based Key Driver tool, to conduct analysis of a satisfaction survey of various cars in the UK. First, we run the Ridge Regression analysis and choose a penalty factor, using the scree plot below.



Ridge Scree Plot



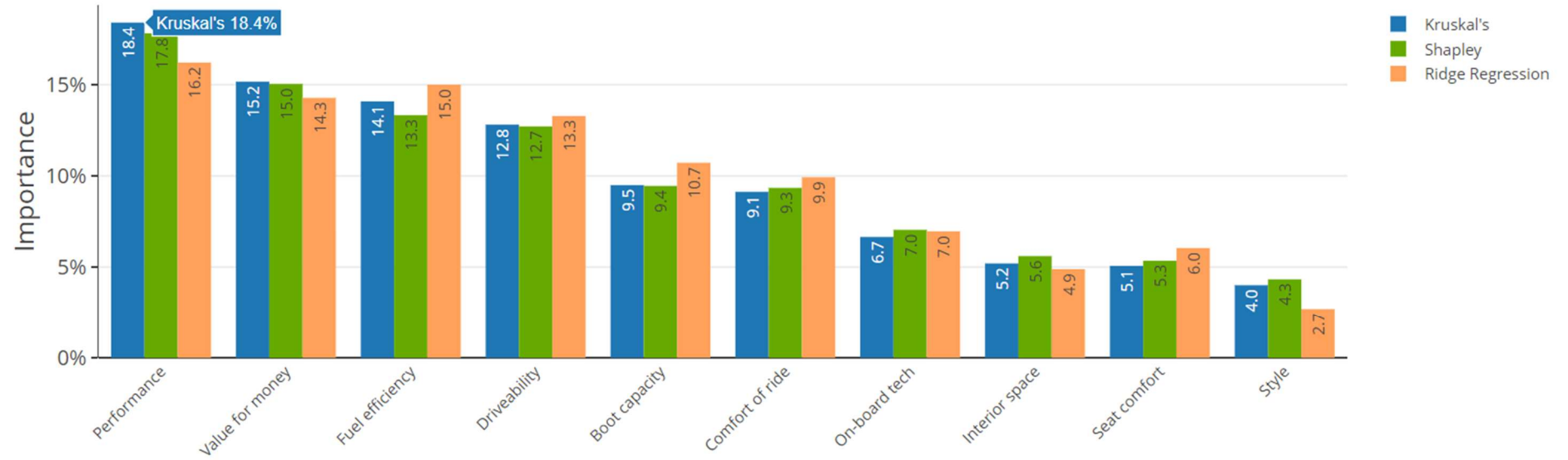
Note that when the penalty factor is zero, i.e., if as in *ordinary multiple linear regression*, the coefficients of two of the drivers are negative. This is a demonstration of the problem of multicollinearity – neither of those drivers are actually negatively correlated with overall satisfaction. However, since they all highly correlated with each other, the coefficients in a regression can be extremely unreliable. By penalising the regression coefficients, this issue can be somewhat alleviated. We will discuss this further with the second example below.

Penalty factors

	0	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6	1.8	2
Comfort of ride	0.065	0.068	0.067	0.066	0.064	0.063	0.061	0.059	0.058	0.056	0.055
Performance	0.203	0.158	0.135	0.121	0.11	0.102	0.096	0.091	0.086	0.082	0.079
Style	-0.060	-0.024	-0.006	0.005	0.012	0.017	0.020	0.023	0.024	0.026	0.026
Value for money	0.126	0.117	0.108	0.101	0.095	0.090	0.086	0.082	0.079	0.076	0.073
Seat comfort	0.036	0.035	0.036	0.037	0.038	0.038	0.038	0.038	0.038	0.038	0.037
Interior space	-0.017	0.007	0.018	0.024	0.028	0.031	0.033	0.034	0.034	0.035	0.035
Boot capacity	0.107	0.089	0.081	0.075	0.071	0.068	0.065	0.062	0.060	0.058	0.056
Fuel efficiency	0.160	0.138	0.123	0.111	0.102	0.095	0.089	0.084	0.079	0.075	0.072
On-board tech	0.028	0.034	0.039	0.041	0.043	0.044	0.044	0.044	0.044	0.044	0.043
Driveability	0.110	0.106	0.099	0.094	0.088	0.084	0.08	0.076	0.073	0.071	0.068
R-squared	33.5%	33.2%	32.8%	32.3%	31.9%	31.4%	30.9%	30.5%	30.1%	29.6%	29.2%



Key Driver Importance Weights



Key Driver Results with Ridge penalty factor of 1

	Kruskal's	Shapley	Ridge	Mean
Performance	18.4%	17.8%	16.2%	5.540
Value for money	15.2%	15.0%	14.3%	5.543
Fuel efficiency	14.1%	13.3%	15.0%	5.280
Driveability	12.8%	12.7%	13.3%	5.657
Boot capacity	9.5%	9.4%	10.7%	5.571
Comfort of ride	9.1%	9.3%	9.9%	5.412
On-board tech	6.7%	7.0%	7.0%	5.351
Interior space	5.2%	5.6%	4.9%	5.369
Seat comfort	5.1%	5.3%	6.0%	5.182
Style	4.0%	4.3%	2.7%	5.354

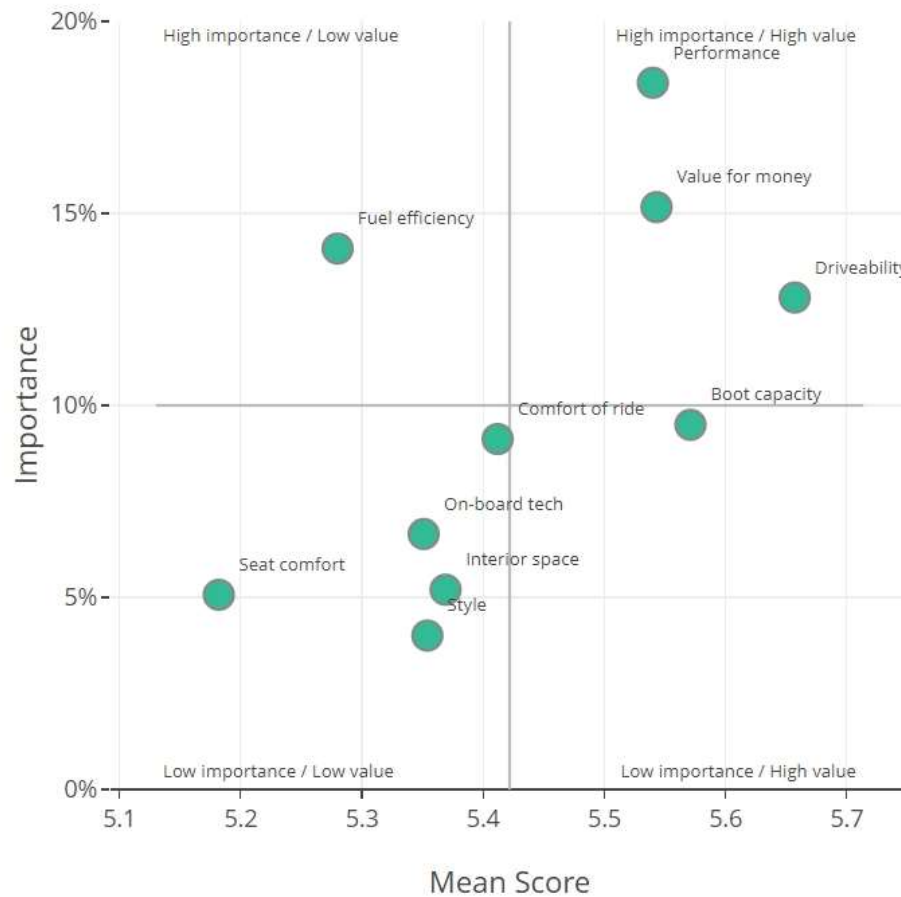
The results are somewhat as expected. “Performance”, “value for money”, “driveability” and “fuel efficiency” are the most importance factors whilst “style”, “seat comfort” and the “interior space” in a car, less so. We note that we the scree plot for the Ridge Regression flattens when the penalty factor is closer to 2, however we are mindful of the low R^2 value.

In fact, the R^2 is quite low overall here, which is somewhat unusual for this sort of data. So, there are probably some other factors that are important in determining satisfaction that not been surveyed that were not asked in the survey. In addition, that the mean scores are very clustered together, which again can be common with this type of data. We should therefore be careful about the scale when plotting on a quadrant map. Perhaps a *Top Box percentage* might be more appropriate in this case.

Having said that, the above quadrant map shows that performance, value for money and driveability areas of **strength**, with only fuel efficiency an apparent area of **weakness**. This would give the manufacturer an area to focus on. The other attributes where the cars score very poorly on (curiously mainly to do with the interior of a car) are less important in determining satisfaction.



Quadrant map

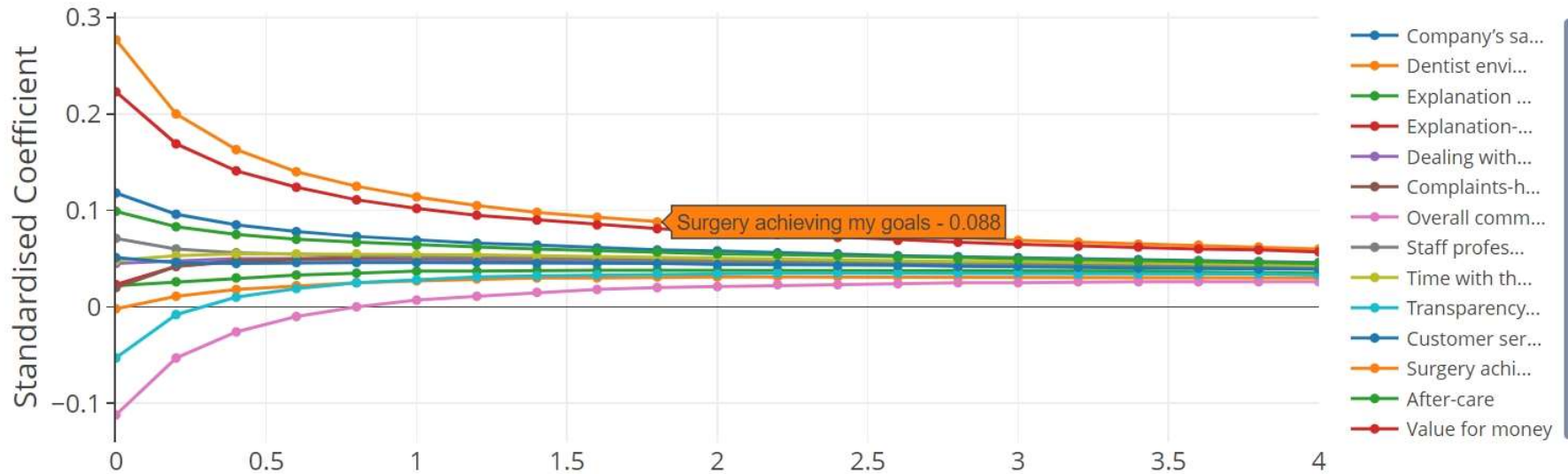


Example 2: Satisfaction with dental treatment

Here, we again use our web-based Key Driver tool, to conduct analysis an analysis of a satisfaction survey of various dental practices in the UK. First, we run the Ridge Regression analysis and choose a penalty factor, using the scree plot below.



Ridge Scree Plot



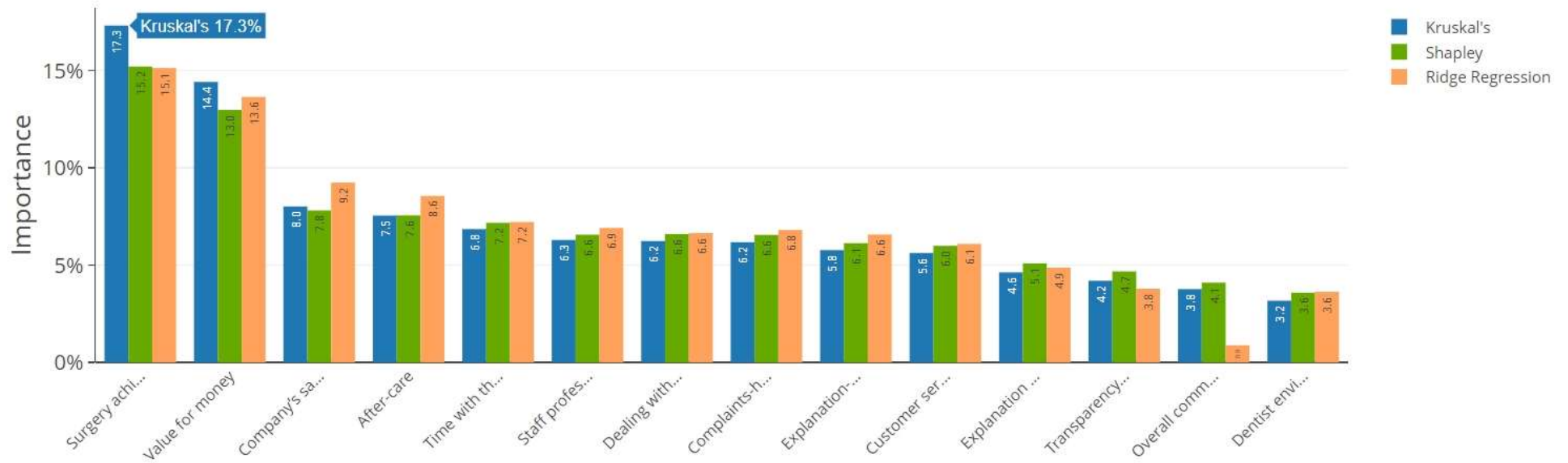
Penalty factors

	0	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6	1.8	2
Overall communication											
Company's sales practice	0.118	0.096	0.085	0.078	0.073	0.069	0.066	0.064	0.061	0.059	0.058
Dentist environment / facilities	-0.002	0.011	0.018	0.022	0.025	0.027	0.029	0.030	0.030	0.031	0.031
Explanation - during procedure	0.022	0.026	0.030	0.033	0.035	0.037	0.037	0.038	0.038	0.038	0.038
Explanation- after procedure	0.023	0.042	0.047	0.049	0.049	0.049	0.049	0.049	0.048	0.047	0.047
Dealing with any concerns	0.045	0.048	0.050	0.050	0.050	0.050	0.050	0.049	0.048	0.048	0.047
Complaints-handling	0.020	0.042	0.048	0.050	0.051	0.051	0.051	0.050	0.050	0.049	0.048
Overall communication	-0.112	-0.053	-0.026	-0.010	0.000	0.007	0.011	0.015	0.018	0.020	0.021
Staff professionalism	0.071	0.060	0.056	0.054	0.053	0.052	0.051	0.050	0.049	0.049	0.048
Time with the dentist	0.048	0.053	0.055	0.055	0.055	0.054	0.054	0.053	0.052	0.051	0.050
Transparency of pricing	-0.053	-0.008	0.010	0.019	0.025	0.028	0.031	0.032	0.033	0.034	0.034
Customer service	0.051	0.046	0.045	0.046	0.046	0.046	0.046	0.045	0.045	0.045	0.044
Surgery achieving my goals	0.277	0.200	0.163	0.140	0.125	0.114	0.105	0.098	0.093	0.088	0.084
After-care	0.099	0.083	0.075	0.070	0.067	0.064	0.062	0.060	0.058	0.057	0.055
Value for money	0.223	0.169	0.141	0.124	0.111	0.102	0.095	0.090	0.085	0.081	0.078
R-squared	50.2%	49.6%	48.9%	48.3%	47.7%	47.2%	46.8%	46.4%	46.0%	45.7%	45.3%

This example perfectly demonstrates the problem with multicollinearity with this type of data. Look at the coefficient for “Overall communication”. With no penalty factor, this coefficient is large **but** with a negative sign. In fact, its absolute value is the fourth largest overall, and significant, implying that better overall communication leads to *lower* overall satisfaction. This is clearly incorrect and is a direct result of the multicollinearity in the dataset - the minimum correlation between all the factors was over 0.5. (If we ran a regression with “Overall communication” as the only independent variable it would have a positive coefficient of 0.471.)



Key Driver Importance Weights

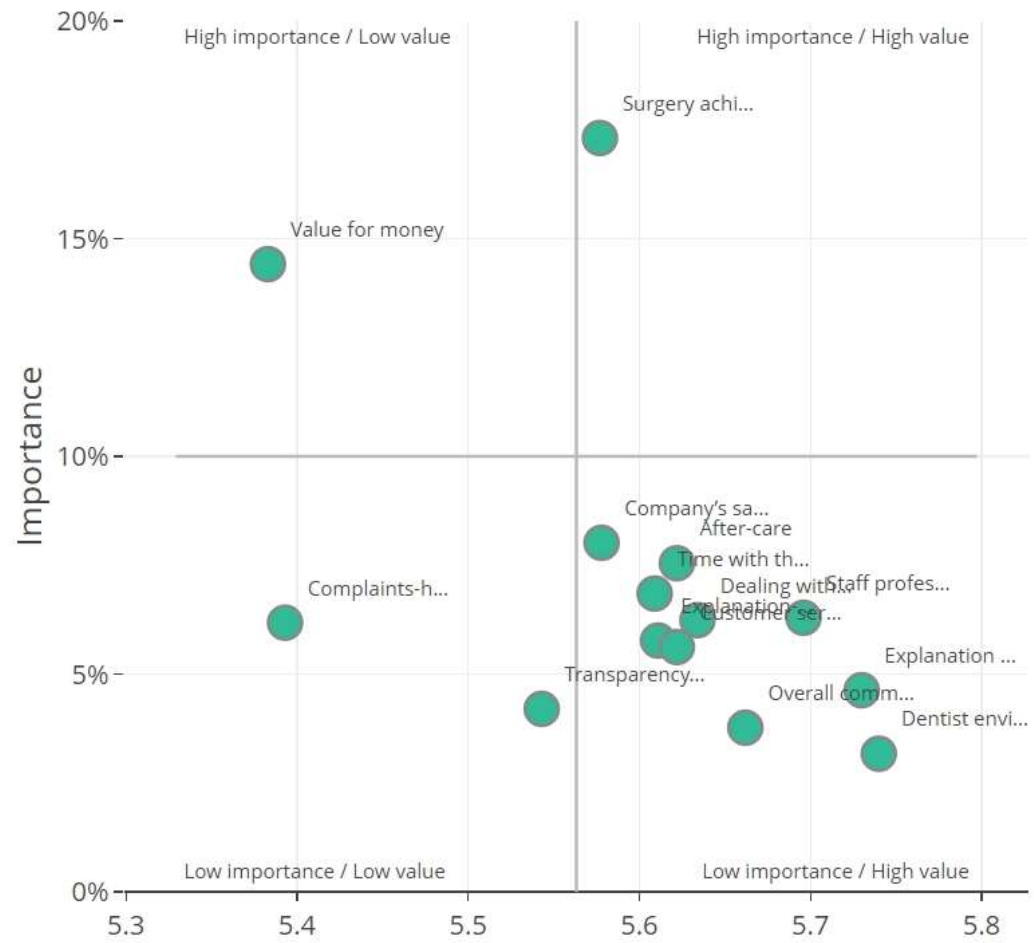


Key Driver Results with penalty factor of 1.6

	Kruskal's	Shapley	Ridge	Mean
Surgery achieving my goals	17.3%	15.2%	13.1%	5.577
Value for money	14.4%	13.0%	12.0%	5.383
Company's sales practice	8.0%	7.8%	8.7%	5.578
After-care	7.5%	7.6%	8.2%	5.622
Time with the dentist	6.8%	7.2%	7.3%	5.609
Staff professionalism	6.3%	6.6%	7.0%	5.696
Dealing with any concerns	6.2%	6.6%	6.8%	5.634
Complaints-handling	6.2%	6.6%	7.0%	5.393
Explanation- after procedure	5.8%	6.1%	6.8%	5.611
Customer service	5.6%	6.0%	6.3%	5.622
Explanation - during procedure	4.6%	5.1%	5.4%	5.730
Transparency of pricing	4.2%	4.7%	4.7%	5.543
Overall communication	3.8%	4.1%	2.5%	5.662
Dentist environment / facilities	3.2%	3.6%	4.2%	5.740

With the multicollinearity accounted for, we have two clear factors: "Surgery achieving my goals" and "Value for money". Notice that "Overall communication" is now relatively unimportant. Also, note the R^2 is much higher in this example than the cars example above, showing that the independent variables in this example capture more of the information than in the previous one.

Equally importantly, there is a clear difference between the mean scores of the two most important variables, despite again the means being generally clustered together (so a top-box measure may see differentiation). This is reflected in the tool's quadrant map. Coupled with the relatively high scores for the cluster of attributes in the bottom right quadrant, this implies that price of its services is something the surgeries should look at. In addition, the complaints procedure, although not as important as value for money, should have some attention to paid to it.



Summary

JumpData's online Key Driver tool provides a fast and inexpensive way of running three of the most popular Key Driver techniques. Even the computationally expensive Kruskal's analysis runs in a remarkably short length of time.

The tool allows market research companies and other organisations to derive better insights from their data, than classical linear regression will provide. This drives value for the client, better informing their decision-making process.