BESPOKE ANALYTIC SOLUTIONS

2024

SERVICE OFFERING

Data Quality Assurance

Error Checking service

Technique & benefits

Error Checking Service

Contents

Introduction

Data is fast becoming the new gold of the digital age. More is produced and consumed every year, and its importance in the modern world continues to grow exponentially. This trend brings unparalleled benefits to organisations that wish to harness the power of data, but also new risks. The Error Checking service exists to ensure that any dataset you handle has been rigorously verified with state-of-the-art statistical analysis, giving you the peace of mind to proceed.

Executive Summary

Ensuring data quality is a continual challenge for all organisations. People and processes can generate errors or inconsistencies, which are often hard to detect, particularly in large datasets. Problems that are not detected at an early stage can cause larger issues down the line, resulting in financial loss, reputational harm and even possible lawsuits.

Random spot-checks and 'outlier analysis' can help to find simple errors, but they are not sufficient for interrogating a large dataset in detail. The Error Checking service goes further than any other method, generating bespoke statistical models for each of the metrics in a dataset, and flagging areas of concern in an easy-to-digest reporting framework.

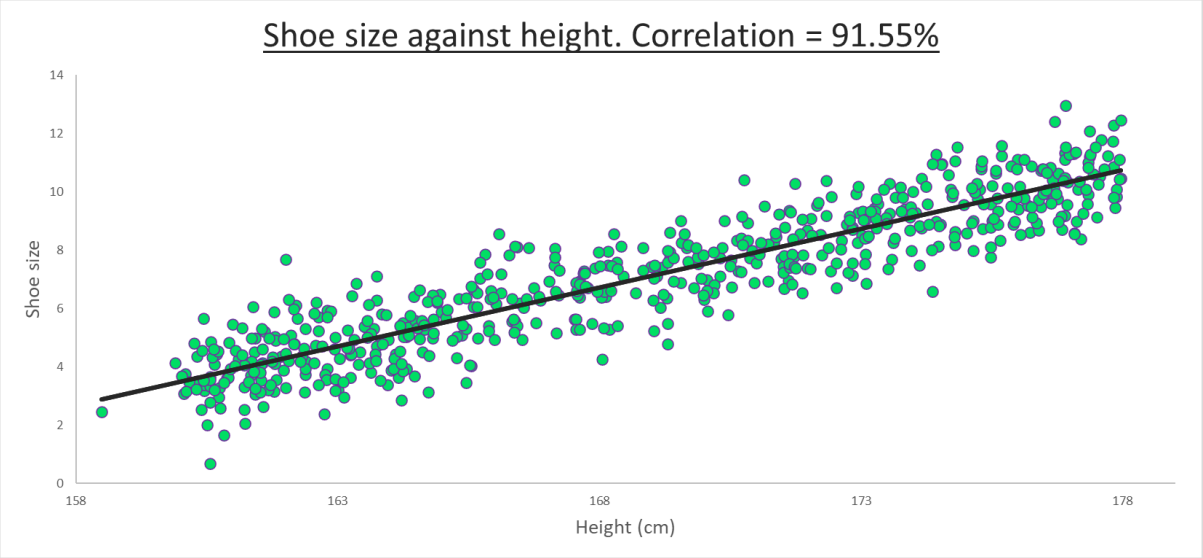
The service will not revolutionise your business, or create new opportunities for record-breaking growth. Rather, it serves to protect you from the down-stream costs and wasted time that data issues can cause. By capping your downside risk when it comes to data quality, you are free to focus on the work that really matters.

How it Works

The core of the process is 'predictive analytics'. This simply means trying to predict one metric using one or more other metrics. A simple example would be predicting a person's shoe size using their height, as shown in the graph on the following page.

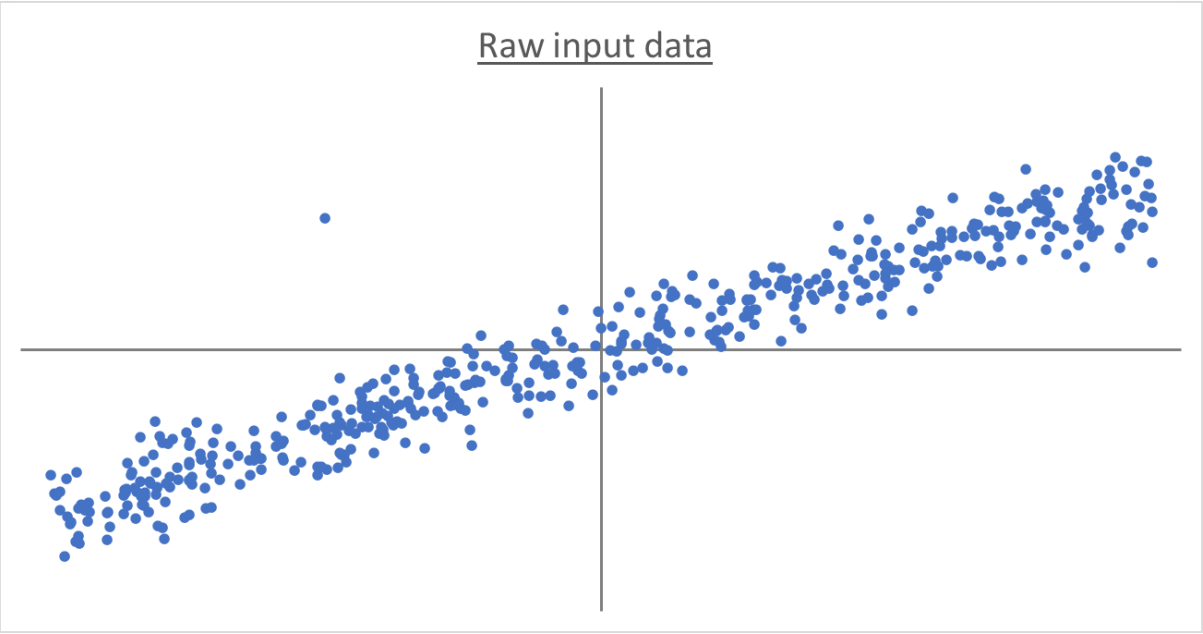
The process predicts each metric in a data table using all other metrics, and compares the real values to the ‘expected’ or ‘model’ values. Large deviations from the model values are more likely to be errors or inconsistencies, and will be automatically marked as such.

Graph 1.a: example of linear correlation



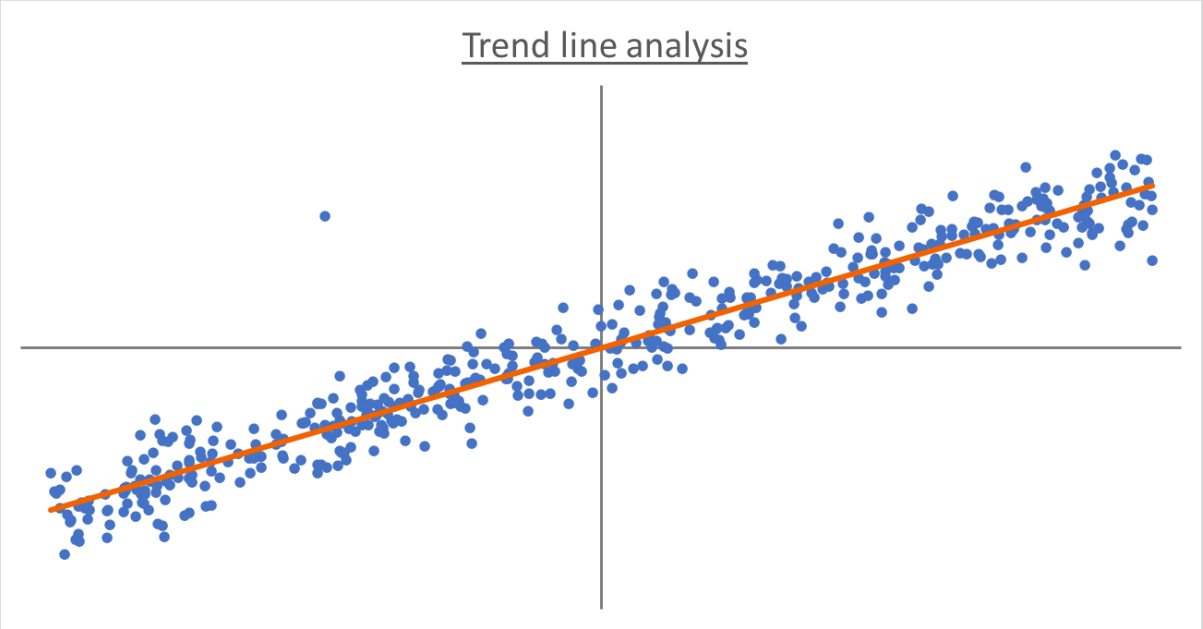
There is a strong connection between the ‘feature’ (height) and the ‘target’ (shoe size). There is also some variation around the trend line, sometimes known as the ‘line of best fit’. Observations that fall a long way away from the trend line are more likely to be errors, and are therefore deserving of further investigation. The graph below demonstrates the full analytical process, from mapping the raw data, to establishing a trend line, and finally identifying potential errors.

Graph 1.b.i. Raw data mapped on a 2D plane



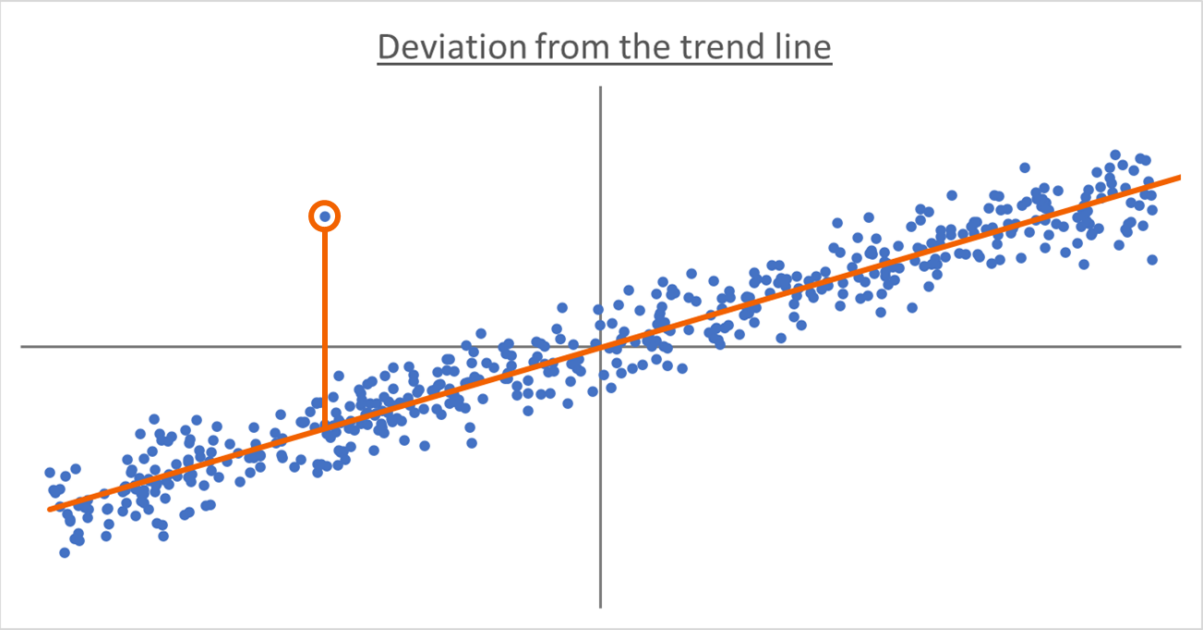
In the error-checking process, each metric is mapped against every other metric in a 2D grid.

Graph 1.b.ii. Data with the line of best fit added



Linear regression is used to establish the trend line, which will sit as close as possible to as many data points as possible.

Graph 1.b.iii. Deviations from trend line are flagged

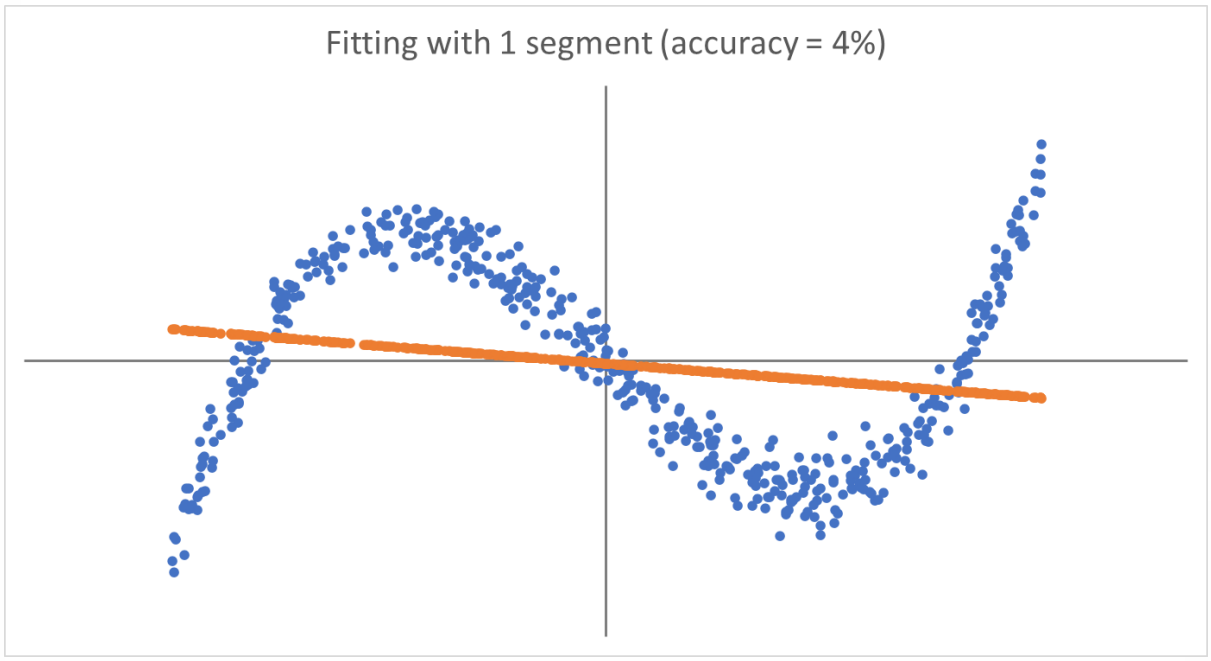


Each data point has an actual value (shown by the blue dots), and a 'model' value, indicated by the position of the orange line. Subtracting one from the other gives the deviations from the trend line.

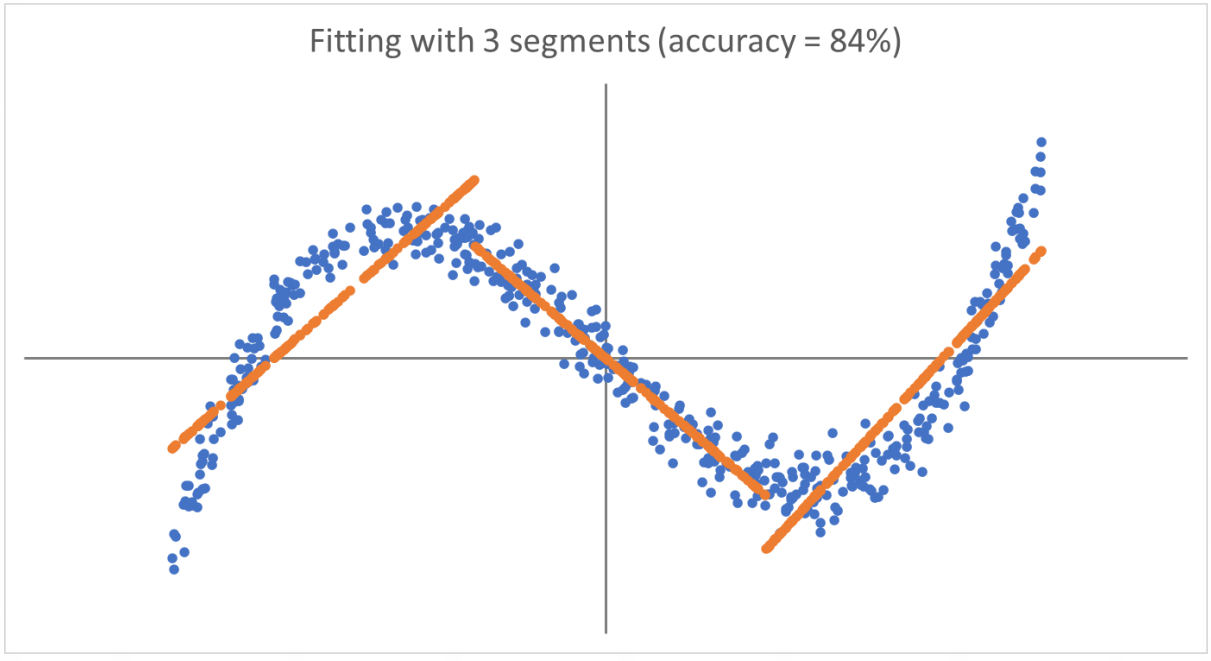
The above process is repeated across all pairs of metrics in a given data table. For example, if a table has 5 metrics then the process will run for a total of 20 times. 10 metrics would mean 90 correlations, and 15 metrics would mean 210 correlations. The process combines all of these relationship to generate a balanced and sophisticated model of the entire dataset.

The process handles categorical as well as numerical data, and is also able to go beyond simple straight-line correlations to model complex non-linear dynamics. The graphs below illustrate the advantage of using multiple segments, when modelling numerical data.

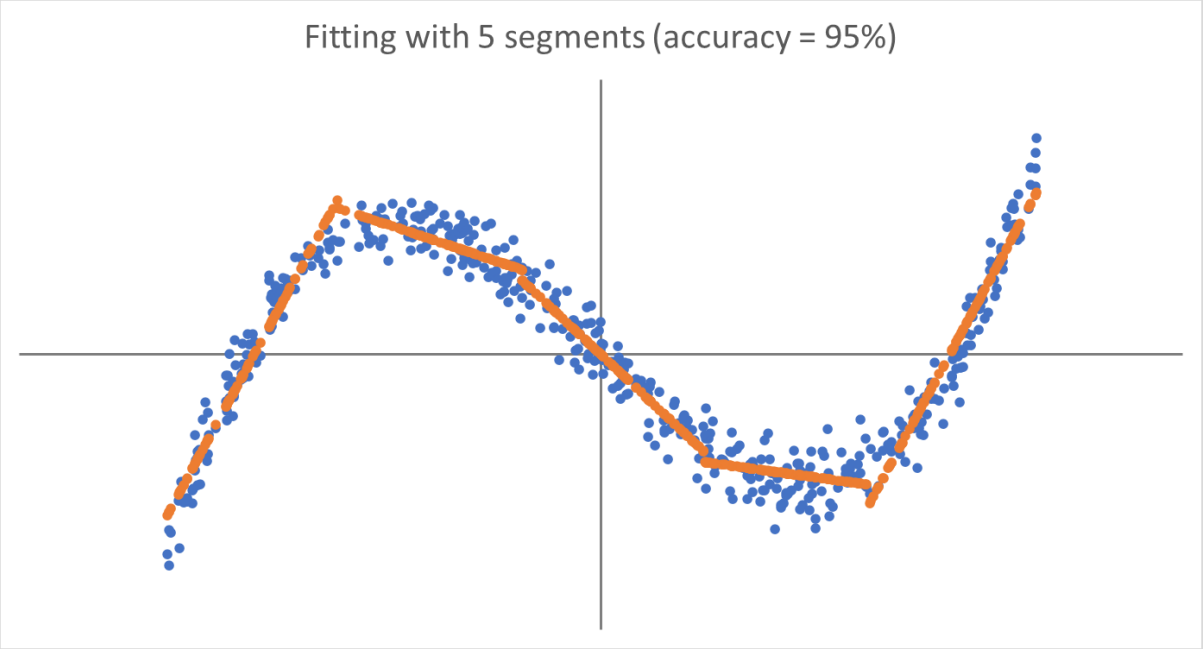
Graph 2.a. A simple linear correlation to model data



Graph 2.b. A more sophisticated correlation that performs better



Graph 2.c. A correlation with 5 segments, which accurately fits the underlying pattern



As can be seen, even small increases in the flexibility of the system can lead to much higher accuracy. This is analogous to the process of generating a predictive AI model with machine learning.

Once the models has been developed, the 'trend' values are compared against the real values, and the deviations recorded. This reveals exactly where the data table behaves differently from expected, and hence the observations that are most likely to be errors or anomalies.

Reporting

A report is generated based on your data, and is split into 6 sections. The data used to generate the following analysis comes from records of UK local government expenditure and income: https://www.gov.uk/government/statistics/local-authority-revenue-expenditure-and-financing-england-2022-to-2023-individual-local-authority-data-outturn.

1. Results Summary

Table 3.a. The spread of deviations from trend, summarised at a high level

Row item deviations	
Count	414
Min	0
25th percentile	0.02
Median	0.07
Average	0.1
75th percentile	0.14
Max	0.69
Spread	0.11

Column metric deviations	
Count	9
Min	0
25th percentile	0.055
Median	0.1
Average	0.1
75th percentile	0.145
Max	0.19
Spread	0.05

Data point deviations	
Count	3,726
Min	0
25th percentile	0
Median	0.03
Average	0.1
75th percentile	0.09
Max	2.28
Spread	0.2

The first part of the report shows high-level statistical summaries, of the deviations from expected values. Particularly under “Row item deviations”, you will be able to see if it is a minority of entries that are driving the bulk of the deviations.

2. Main Results

A. Row items

Table 3.b.i. Average deviation from trend for each row item

Row items	
Item	Deviation
Adur	0.01
Allerdale	0.02
Amber Valley	0.01
Arun	0.01
Ashfield	0.01
Ashford	0.01
Avon & Somerset Police	0.14
Avon Combined Fire	0.13
Babergh	0.02
Barking & Dagenham	0.16
Barnet	0.13
Barnsley	0.09
Barrow-in-Furness	0.01
Basildon	0.03
Basingstoke & Deane	0.11
Bassetlaw	0.01
Bath & North East Somerset	0.12

Adjacent is an example of the first results section, which concerns the average for row items. For each row item, an average is taken across all of the metrics in the table, combining the differences between real and expected values. Higher values indicate more unusual results, and these are coloured in yellow/red. Where data was close to the expected values, it is shaded in green.

This and all other parts of the Results tab can be turned into Excel tables, offering further scope for filtering and analysis.

The average values in the “Deviation” column do not vary systematically with the size or content of the original data table. This allows you to make meaningful comparisons between the results from the analysis of different tables.

B. Row items – core

Table 3.b.ii. Most significant row items in terms of average deviation

Row items - core	
Item	Deviation
Greenwich	0.69
Westminster	0.64
West Midlands Police	0.61
Leeds	0.6
Cheshire East	0.53
Bristol	0.48
Greater Manchester Combined Authority	0.46
Richmond upon Thames	0.46
Greater Manchester Police	0.44
Birmingham	0.43
Manchester	0.41
Thames Valley Police	0.41
West Yorkshire Police	0.41
Southwark	0.4
Liverpool City Region Combined Authority	0.37

The next section returns the same data as the first, but sorted and filtered to focus on the minority of entries that produced the most significant deviations from trend values. In this example there are 414 row items, of which 71 are displayed in the ‘core’ section.

This focussed list makes it easy to see which items are more likely to contain errors, or might be deserving of further investigation.

By using advanced analytics and focussing on the most significant outliers, the service saves a great deal of time and effort that would be spent manually spot-checking data tables.

C. Column metrics

Table 3.b.iii. Average deviations of column metrics and their predictability

Column metrics		
Metric	Deviation	Predictability
Class	0	100.00%
Detailed Class	0.12	86.71%
Employees	0.15	91.00%
Running Expenses	0.1	96.53%
Total Expenditure	0.04	99.57%
Sales, Fees and Charges	0.19	81.11%
Other Income	0.14	92.56%
Total Income	0.1	95.96%
Net Current Expenditure	0.07	98.63%

The third section shows the average deviation from expected values, for each of the metrics. It also displays the final accuracy of the model, attempting to predict each of the metrics using all other metrics.

The 'Predictability' column indicates the extent to which the data table is internally well-connected. Bear in mind that there may be more complex connections between metrics that the process does not pick up.

D. Individual items – core

Table 3.b.iv. Specific and detailed summary of deviations from trend lines

Individual items - core						
Row item	Metric	Value	Expected	Deviation	Row deviation	Metric deviation
Greenwich	Sales, Fees and Charges	£256,210	£134,593	2.28	0.69	0.19
Westminster	Sales, Fees and Charges	£235,778	£121,328	2.15	0.64	0.19
West Midlands Police	Employees	£707,582	£358,505	1.96	0.61	0.15
Cheshire East	Sales, Fees and Charges	£202,298	£99,337	1.93	0.53	0.19
Bristol	Sales, Fees and Charges	£182,246	£91,116	1.71	0.48	0.19
Southwark	Sales, Fees and Charges	£5,291	£89,475	1.58	0.4	0.19
Richmond upon Thames	Sales, Fees and Charges	£166,553	£83,569	1.56	0.46	0.19
Greenwich	Other Income	£159,514	£273,424	1.52	0.69	0.14
Greater Manchester Police	Employees	£614,213	£357,802	1.44	0.44	0.15
Greenwich	Total Income	£415,724	£258,902	1.43	0.69	0.1
Westminster	Other Income	£123,497	£230,956	1.43	0.64	0.14
Cheshire East	Other Income	£45,138	£146,816	1.36	0.53	0.14
West Midlands Police	Running Expenses	£161,727	£553,313	1.34	0.61	0.1
Westminster	Total Income	£359,276	£217,386	1.3	0.64	0.1
Newham	Sales, Fees and Charges	£165,733	£98,071	1.27	0.33	0.19
Leeds	Sales, Fees and Charges	£118,701	£185,696	1.26	0.6	0.19
Bristol	Other Income	£18,962	£112,036	1.24	0.48	0.14
Birmingham	Sales, Fees and Charges	£237,154	£173,912	1.19	0.43	0.19

In the last part of the results report, a summary of individual data points is produced. As shown in the table, the highlighted data points come from the original 2-dimensional grid. The table shows the minority that are significantly different from their expected values. This makes it even easier to track down potential anomalies and errors, by tracking them down to the level of individual cells.

3. Score grid

Each metric can be used to predict every other metric, with varying levels of accuracy. This section is set up so that row items (on the left-hand side) predict column items (on the top). For example "Employees" predicts "Running Expenses" with an accuracy of 74.29%.

Table 4. Summary of predictive power for each pair of metrics

Metric	Class	Detailed Class	Employees	Running Expenses	Total Expenditure	Sales, Fees and Charges	Other Income	Total Income	Net Current Expenditure
Class		86.71%	58.21%	74.69%	71.96%	62.94%	57.45%	67.03%	70.75%
Detailed Class	100.00%		67.23%	75.40%	73.79%	63.15%	57.70%	67.30%	73.21%
Employees	62.08%	62.08%		74.29%	88.92%	57.35%	61.47%	68.03%	90.37%
Running Expenses	69.32%	65.22%	76.87%		96.71%	75.85%	77.58%	88.03%	94.11%
Total Expenditure	66.43%	64.49%	89.51%	96.53%		74.28%	76.71%	86.00%	98.63%
Sales, Fees and Charges	64.98%	59.90%	56.63%	75.26%	72.78%		56.91%	81.20%	65.80%
Other Income	62.08%	58.45%	63.39%	78.29%	77.57%	58.48%		92.96%	68.28%
Total Income	64.49%	60.87%	67.27%	86.55%	84.56%	80.84%	92.56%		74.77%
Net Current Expenditure	71.26%	66.67%	91.00%	93.45%	98.67%	67.44%	68.48%	77.69%	

This supplementary analysis will help to understand where the strongest relationships are in your data. This will help you to make accurate predictions based on limited information, and understand where certain metrics can and cannot be predicted based on the available data.

4. Source data

This tab reproduces the original data that was fed into the process, but adds a coloured overlay to summarise cells that were different from expected. With this information, you can see which values in the table were higher or lower than expected, and by how much.

Table 5.a. Original data, with coloured overlay

Local authority	Class	Detailed Class	Employees	Running Expenses	Total Expenditure	Sales, Fees and Charges	Other Income	Total Income	Net Current Expenditure
Bracknell Forest	UA	UA	£111,094	£160,618	£271,712	£20,661	£36,591	£57,252	£214,460
Bradford	MD	MD	£452,765	£628,430	£1,081,195	£78,505	£142,839	£221,344	£859,851
Braintree	SD	SD	£19,136	£27,286	£46,422	£6,267	£21,127	£27,394	£19,028
Breckland	SD	SD	£11,211	£30,659	£41,871	£5,894	£9,791	£15,684	£26,187
Brent	LB	LB	£278,319	£429,802	£708,121	£118,461	£27,804	£146,265	£561,856
Brentwood	SD	SD	£11,543	£26,436	£37,980	£7,970	£21,851	£29,821	£8,158
Brighton & Hove	UA	UA	£301,556	£420,860	£722,416	£109,948	£83,228	£193,176	£529,240
Bristol	UA	UA	£296,355	£641,539	£937,894	£182,246	£18,962	£201,208	£736,686
Broadland	SD	SD	£9,661	£11,866	£21,527	£4,249	£4,286	£8,535	£12,992
Bromley	LB	LB	£103,882	£453,043	£556,926	£51,353	£123,496	£174,849	£382,077
Broxbourne	SD	SD	£7,941	£25,247	£33,188	£10,070	£10,754	£20,824	£12,364
Broxtowe	SD	SD	£11,951	£15,384	£27,335	£3,634	£12,614	£16,248	£11,087
Buckinghamshire & Milton Ke	O	FR	£23,631	£6,609	£30,240	£804	£201	£1,005	£29,235
Buckinghamshire Council	UA	UA	£461,139	£652,089	£1,113,228	£126,286	£135,239	£261,525	£851,703

Purple shading on category metrics shows where the item belonged to a different category than was predicted. Blue shading shows where values are lower than expected, with bolder shades indicating a larger difference. The same logic applies to red shading for values that were higher than expected.

5. Model values

This section presents the idealised "model values" for each of the row items and each of the metrics. Following on from the example in the earlier graph, these are the values given by the lines of best fit. The overlay of colours is identical – switching back and forth between these tabs will reveal why each of the cells are shaded as they are.

Table 5.b. Fitted data, with coloured overlay

Local authority	Class	Detailed Class	Employees	Running Expenses	Total Expenditure	Sales, Fees and Charges	Other Income	Total Income	Net Current Expenditure
Bracknell Forest	UA	UA	£120,680	£137,599	£294,196	£25,541	£31,601	£61,692	£219,737
Bradford	MD	MD	£410,854	£692,145	£1,051,418	£95,433	£127,185	£236,629	£817,409
Braintree	SD	SD	£16,012	£28,320	£49,536	£9,609	£17,659	£33,031	£23,985
Breckland	SD	SD	£20,653	£26,033	£54,813	£7,675	£8,385	£19,762	£20,306
Brent	LB	LB	£293,026	£447,699	£713,495	£62,229	£89,860	£112,807	£504,934
Brentwood	SD	SD	£8,007	£22,498	£33,399	£10,064	£19,397	£33,026	£19,127
Brighton & Hove	UA	UA	£280,130	£457,065	£714,366	£85,775	£105,994	£164,577	£516,906
Bristol	UA	UA	£362,154	£598,249	£949,736	£91,116	£112,036	£123,980	£697,382
Broadland	SD	SD	£11,121	£12,703	£31,361	£4,303	£4,263	£11,185	£11,377
Bromley	LB	LB	£206,718	£340,005	£567,003	£68,105	£108,341	£193,583	£410,336
Broxbourne	SD	SD	£10,716	£18,151	£35,690	£8,038	£12,592	£17,682	£17,677
Broxtowe	SD	SD	£9,894	£16,793	£31,905	£7,083	£8,806	£17,160	£14,307
Buckinghamshire & Milton Key	O	P	£21,548	£18,839	£43,435	£998	£671	£8,535	£15,773
Buckinghamshire Council	UA	UA	£407,631	£713,133	£1,063,538	£103,773	£157,415	£235,442	£844,238

6. Model deviations

The final part of the report shows the underlying differences between the actual and modelled data. As before, the overlay of colouring is the same. This tab makes it easy to run your own statistical analyses of where the deviations from trend values have occurred.

Table 5.c. Deviations from trend, with coloured overlay

Local authority	Class	Detailed Class	Employees	Running Expenses	Total Expenditure	Sales, Fees and Charges	Other Income	Total Income	Net Current Expenditure
Bracknell Forest	0.00	0.00	-0.05	0.08	-0.05	-0.09	0.07	-0.04	-0.02
Bradford	0.00	0.00	0.24	-0.22	0.07	-0.32	0.21	-0.14	0.12
Braintree	0.00	0.00	0.02	0.00	-0.01	-0.06	0.05	-0.05	-0.01
Breckland	0.00	0.00	-0.05	0.02	-0.03	-0.03	0.02	-0.04	0.02
Brent	0.00	0.00	-0.08	-0.06	-0.01	1.05	-0.83	0.31	0.17
Brentwood	0.00	0.00	0.02	0.01	0.01	-0.04	0.03	-0.03	-0.03
Brighton & Hove	0.00	0.00	0.12	-0.12	0.02	0.45	-0.30	0.26	0.04
Bristol	0.00	0.00	-0.37	0.15	-0.03	1.71	-1.24	0.71	0.12
Broadland	0.00	0.00	-0.01	0.00	-0.02	0.00	0.00	-0.02	0.00
Bromley	0.00	0.00	-0.58	0.39	-0.02	-0.31	0.20	-0.17	-0.08
Broxbourne	0.00	0.00	-0.02	0.02	-0.01	0.04	-0.02	0.03	-0.02
Broxtowe	0.00	0.00	0.01	0.00	-0.01	-0.06	0.05	-0.01	-0.01
Buckinghamshire & Milton Key	0.00	0.87	0.01	-0.04	-0.03	0.00	-0.01	-0.07	0.04
Buckinghamshire Council	0.00	0.00	0.30	-0.21	0.12	0.42	-0.30	0.24	0.02

The Advantages

This service has a broad range of applications. It can be used to help detect fraud, by finding transactions that do not fit the normal pattern within a ledger. It can detect potential data entry errors in a set of survey responses, working with correlations between metrics to flag up inconsistencies. Additionally, the results could help to show which people in an organisation deviate from normal work patterns, and may benefit from a tailored employee-management strategy.

The report comes as an Excel document, making it well-adapted for sharing within and between organisations. The interactive nature of MS Excel makes it easy to scrutinise the results, and further investigation can be carried out in a convenient analytical environment. Presenting the results in this way also provides the facility to export individual tabs, each of which focuses on a specific aspect of the error-checking process.

The service has the potential to save significant amounts of time, that would otherwise be spent manually checking for errors and inconsistencies. Thus, it can free up your team to focus on value-adding activities. It also reduces the likelihood of using problematic data; by knowing in advance where the potential errors are, you can undertake quicker, targeted investigations. This leads to far more effective validation than spot-checks or basic statistical analysis.

Finally, the process is set up to work with any combination of numerical and text-based (categorical) metrics. This gives the service an unparalleled flexibility, helping you to detect inconsistencies and areas for further investigation, without having to spend time on pre-processing.

Privacy

The service comes with a supplementary tool, for anonymising data prior to egress, and decrypting the results upon receipt. This ensures that no sensitive data is ever made available to a 3rd party, while still delivering the full benefits of the error checking service.

Encoding

The tool automatically encodes categorical (text-based) data as “V_1”, “V_2” etc. for all unique values, in all columns. Numerical data is “normalised”, that is to say, adjusted to fit roughly in the range of -2 to +2. This obscures sensitive data, without impacting the quality of the analysis. An example below demonstrates the process, using data on a species of mollusc called “abalones”.

Abalone	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
Abl_1	M	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15
Abl_2	M	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7
Abl_3	F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9
Abl_4	M	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10
Abl_5	I	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7
Abl_6	I	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8
Abl_7	F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20
Abl_8	F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16
Abl_9	M	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9
Abl_10	F	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19
Abl_11	F	0.525	0.38	0.14	0.6065	0.194	0.1475	0.21	14

In the original data table, the row items and column headers can be seen, as well as the specific values in cells. Compare this to the same data below, after applying the “privacy mask”.

Items	Metric_1	Metric_2	Metric_3	Metric_4	Metric_5	Metric_6	Metric_7	Metric_8	Metric_9
Item_1	V_1	-0.57456	-0.43215	-1.06442	-0.6419	-0.60769	-0.72621	-0.63822	1.571544
Item_2	V_1	-1.44899	-1.43993	-1.18398	-1.23028	-1.17091	-1.20522	-1.21299	-0.91001
Item_3	V_2	0.050033	0.12213	-0.10799	-0.30947	-0.4635	-0.35669	-0.20714	-0.28962
Item_4	V_1	-0.69948	-0.43215	-0.3471	-0.63782	-0.64824	-0.6076	-0.60229	0.020571
Item_5	V_3	-1.61554	-1.54071	-1.42309	-1.27209	-1.21597	-1.28734	-1.32076	-0.91001
Item_6	V_3	-0.82439	-1.08721	-1.06442	-0.97331	-0.98392	-0.94063	-0.85376	-0.59982
Item_7	V_2	0.050033	0.071741	0.250672	-0.10451	-0.55136	-0.35669	0.655017	3.122516
Item_8	V_2	0.174951	0.172519	-0.3471	-0.12388	-0.29453	-0.2837	0.152092	1.881738
Item_9	V_1	-0.408	-0.38176	-0.3471	-0.65108	-0.64373	-0.62129	-0.53045	-0.28962
Item_10	V_2	0.216591	0.323686	0.250672	0.134109	-0.20216	-0.27001	0.58317	2.812322
Item_11	V_2	0.008394	-0.28098	0.011563	-0.45325	-0.74511	-0.30195	-0.20714	1.261349

All sensitive information has been removed from the table. Unique items in “Metric_1” (“Sex”) respect the original pattern, and the relative values of entries in numerical fields are preserved. However, there is no way to extract sensitive information from the “masked” dataset.

Decoding

This process can be automatically reversed in a matter of seconds, by using the original dataset as a “key” to remove the mask. This puts you in complete control of the process, making it possible to undertake sophisticated analysis of sensitive data without ever risking a leak or GDPR breach.

Simply select the dataset you would like to use as the key, choose the masked analysis file, and run the system to un-mask your results. These tools are included as standard for all users.

Contact

If this service may be useful for your organisation, please email ba.solutions@tutanota.com for a no-obligation sample analysis. Other services are also available, such as bespoke analytics, automation and presentation of data – please enquire for details.