Insight from data analytics with an automotive aftermarket SME

Wayne S. Smith1,3 | Shirley Coleman1 | Jaume Bacardit2 | Syd Coxon3

1 School of Mathematics, Statistics, and Physics, Newcastle University, Newcastle upon Tyne, UK
2 School of Computing, Newcastle University, Newcastle upon Tyne, UK
3 Rain Data Ltd, UK

Correspondence
Wayne S. Smith, School of Mathematics, Statistics, and Physics, Newcastle University, Newcastle upon Tyne, UK.
Email: waynestuartsmith@hotmail.com

Abstract
There is enormous general interest in the automotive sector, and there are many sets of informative data openly available to the public. Companies involved in the automotive aftermarket sector have access to further masses of data, and this can be analysed to provide valuable insights to complement those obtainable from the internet and popular press, discovering hidden knowledge and financial benefit within the data by using statistical analysis, big data analytics, and data science. This article gives examples of insight derived from data giving great value to stakeholders in the sector. The examples are followed by a discussion of the issues involved in applying data science and ensuring that solutions are implemented and are sustainable within the business partner companies.

KEYWORDS
data science, intellectual property, monetising data, statistical process control

1 | INTRODUCTION

There is enormous general interest in the automotive sector, and there are many sets of informative data openly available to the public about preferences of vehicle manufacturer, model, and fuel type, as well as all other minutiae of customer buying habits, including colour, engine size, transmission type etc. Data, for example https://www.smmt.co.uk/vehicle-data/1, are often mentioned in the news as part of meta stories involving such issues as environmental concerns, the hazards of diesel vehicle emissions, and the resulting reduction in automotive sales because of this or the increase in sales for hybrid and electric cars as an alternative to vehicles that rely solely on fossil fuels.

The automotive aftermarket is the part of the automotive sector which deals with the replacement of vehicle parts when a vehicle is no longer under warranty from the original manufacturer. Not all companies currently involved in the automotive aftermarket sector store their data for posterity, and whilst others may periodically collect data, they are yet to discover the hidden value and financial benefit within the data that statistical analysis, big data analytics, and data science can provide. Datasets compiled from online vehicle part transactions contain a plethora of information about individual purchases and details of the parts themselves. From these types of datasets, it is possible to gain insight
into buying habits, preferences, return rates, and product attributes. Trends and patterns derived from analysis of automotive transactions can be used to inform the industry better, for example about when specific vehicle parts are bought and therefore which parts to keep in stock at different times of the year to anticipate increased sales. Data science techniques, such as association rule mining, cluster analysis, and decision tree analysis, applied to automotive aftersales data give insight that is not obvious from using traditional statistical tools.

The SME (small to medium enterprise) in this study provides analytical services to a range of different customers. Some of their research into new products and services is in response to ideas relayed to them by their customers about the sorts of knowledge they would like to gain from their datasets. Some of their research is motivated by ideas that arise from their own exploratory analysis of the datasets. These ideas often give rise to hitherto unexpected insight that the SME can also offer to its customers. In all cases, the full benefit of the data science application is realised when the results are clearly visualised and made accessible to customers. Hence, a mixture of technical skills (in analysis and visualisation), business domain knowledge, and good communication skills work together to produce results of great value.

This article gives three unique examples in which insight from large datasets within the automotive aftermarket provide benefits to the various stakeholders in the sector. An overview of each case and the business context are described in the following subsections.

1.1 | Return rates

The first example is based on data from garages buying large numbers of parts for vehicle repairs and considers the return rates of individual vehicle products. Understanding the nature of product returns is a key factor in running any successful business that supplies goods. Returns can inconvenience a company for several reasons including, the cost of re-posting, identifying if a product is fit for re-sale, the cost of future sales etc. It is worth noting however, that whilst returns cause inconvenience for a company, they can also play a considerable part in turnover and business for part dealers. When dealing with very large datasets which include hundreds of thousands of products and millions of units sold, it can be difficult to prioritise the attention to give to which product returns. One method is to consider return rates using a funnel plot as a type of statistical process control (SPC) limit chart. This type of analysis considers the return rate for each category of vehicle part according to the number of units sold. The sales and returns of any part can be considered along with the expected maximum return rate control limits to identify which return rates require immediate investigation. This gives a method to prioritise attention to those parts according to whether (and by how much) their return rates are exceptional compared with the expected data-based norms for the industry.

1.2 | OEM

The second example deals with Original Equipment Manufacturing (OEM) standards. In the context of vehicle parts, the OEM standards refer to the manufacturer of the original equipment at the time a vehicle is made, that is, the parts assembled and installed during the construction of a new vehicle. Aftermarket parts are made by companies trying to conform to the OEM factory measurements, and there is often considerable uncertainty about which parts match to which OEM numbers. For example, several different suppliers can manufacture parts for the same specific vehicle, yet the physical dimensions and properties listed against those parts made by suppliers can vary greatly. Estimations of the factory standards can be derived via inspection of all the supplier’s parts associated with each OEM standard. Determining the correct measurements for an OEM confirms which parts should be affiliated to a vehicle and identifies those parts which may be wrongly matched to a car. In addition, supplier parts can be compared with the estimated values of the OEM factory standards to reveal those supplier parts which can be matched against other cars to which they were not originally intended to be matched. This highlights potential revenue for a part supplier, as they can then sell these parts against more cars with very little or no modification.

1.3 | Part replacement

The third example considers predicting when parts need to be replaced on a car. Comparing invoices for the replacement of vehicle parts and the mileage of a vehicle at the time of replacement highlights the potential failure rate of nonserviceable vehicle parts. Inspecting those parts that were replaced (along with the mileage at the point of
replacement) and comparing these across different vehicle makes and models also indicate the reliability of different vehicle part manufacturers. Fitting an empirical cumulative density function (ECDF) to the replacement of a vehicle part and the mileage (at replacement) aids in the process of identifying key points in a car’s history when replacement is needed and also potentially when failure may occur. This statistical technique determines the probability of an event using the observed data rather than a theoretical estimate of the probability or data published in open sources.

The rest of the manuscript consists of a data section in which the data, hardware, and software for the examples are given. Section 2 describes the datasets and how they are integrated and preprocessed ready for analysis. Section 3 describes the statistical and machine learning methods carried out in each example and gives the results of the analysis and the business value derived. Section 4 discusses the implications of such analysis and the challenges faced by the sector and the SME. We consider the issues involved in applying data science and ensuring that solutions are implemented and sustainable within the business partners.

2 | DATA

The SME receives vast quantities of data in real time from a range of sources such as look-ups in catalogues, invoices, and receipts. Some of the look-up data arise in real time and adhere to the three classic Vs of Big Data, being of high Volume, Variety, and Velocity (many other Vs have been added over time). Other data are reported on a daily or periodic basis. The data may arise from different sources and need to be integrated prior to analysis. The data are often incomplete and messy and need to be preprocessed by careful cleaning. The data are captured as comma separated values in .csv files or imported directly from a database, and analysed with:

- R (version 3.5.1) in conjunction with
- R Studio (Version 1.1.456)
- Microsoft Excel (Office 365 MSO 16.0.10827.20118)
- SQL Server Management Studio (v18.0).

The data used in each case study are interpreted as a representative sample of the entire market of the corresponding part manufacturers, but it is also possible that there are sources of bias from omission or under-representation of sub-groups. The data for each case study are considered separately in the following subsections.

2.1 | Return rates

For the analysis of product returns, data were sourced from 14,142,636 rows of credit and invoices of automotive parts covering the periods March 2008 to September 2017. From this, 12,582,464 invoice rows were extracted corresponding to the transactions of buying vehicle parts. Of these remaining rows, there were only 739,771 invoice rows which could be matched to a specific vehicle type (covering the period 2017). This final dataset was created by joining data from three sources: (a) an invoicing system that records part transactions; (b) a cataloguing system of the parts themselves; and (c) a system that matches the transactional data to individual cars (ie, matching each purchase to a vehicle registration number). Together, these sources provide a complete record of when a part was bought, the quantity, and how many if any was returned. For the example used in this article, in which air filters are considered, all instances of the part being sold (within the dataset) were selected, which in the case of air filters accounts for 1629 individual products.

2.2 | OEM

For the analysis on OEM factory standards, data from an online electronic cataloguing system of aftermarket vehicle parts were used. For a category of part, such as fuel filters, all instances of those parts were collated and grouped according to the specific cars for which they were recommended as suitable. Comparing the dimensions across these parts (for a specific part for a specific car) provides a method of determining the attributes of the OEM. Two separate algorithms were developed from this: (a) in which all measurements for a part are compared so that anomalies and outliers can be identified—detecting parts that do not match a typical measurement, and (b) identifying parts that can potentially fit other cars which were not considered previously. For the example used in this article, in which fuel filters are considered, there are 1,042,252 combinations of part type to specific vehicle type (OEM standard).
2.3 | Part replacement

For the analyses on predicting part replacement, the data were sourced from an online garage management system that records transactions of aftermarket parts. For each transaction recorded, the mileage of the car at the time of purchasing a replacement part was recorded, as well as the make and model of car. The data for this analysis comprised 5,913,350 rows of invoice data covering the period January 2016 to July 2018. From this dataset, a specific vehicle make/model was extracted for a given part that was replaced, along with the mileage at the time of replacement. For a popular make/model/part replacement, typically a few thousand examples would be available.

3 | PROCEDURES AND RESULTS

Once the datasets are in a suitable form for analysis and have been checked for anomalies using data visualisations and summary statistics, they are ready for further analysis. Each case is described separately in the subsections below.

3.1 | Return rates

In this example, the aim is to give a clear picture of the parts whose return rates require further investigation. Funnel plots were developed in health care and have been used to monitor data from clinics to identify areas of concern. Funnel plots are particularly useful when studying proportions where the denominator can vary widely. Funnel plots can be used as a type of SPC chart to show when to “leave the process alone” and when to take action. Control limits are calculated as the mean return rate plus and minus 3 times its standard deviation and intend to contain nearly all the data in a stable system. Points within the control limits represent products whose return rates conform to the expected returns level. Any points outside of the control limits are investigated as they may represent exceptional circumstances. Resolving these exceptions brings business advantage and produces a more stable process.

For a given vehicle part, such as air filter, there were 1,629 unique product transactions within the dataset. For each product, the return rate is calculated by dividing the number of returns by the number of sales for each product. The overall return rate was calculated as

\[
P = \frac{\text{the average return rate of all 1,629 air filter products.}}{
\sqrt{\text{Sample Size}}}
\]

The funnel plot control limits were calculated as follows:

\[
P \pm 3 \sqrt{\frac{P(1-P)}{\text{Sample Size}}}
\]

where:

\[
\sqrt{\frac{P(1-P)}{\text{Sample Size}}}
\]

is the standard deviation of a proportion, and \(P\) is the mean proportion, which for this dataset is \(= 0.11\).

Figure 1 shows the return rates for 1,629 air filter products. The x-axis is the number of products bought, and the y-axis is the return rate. The curved lines represent the funnel plot control limits calculated according to the formula above. Each dot represents the return rate for a specific air filter applied to a specific vehicle type. The red dotted horizontal line is the average return rate of 0.11, for the 1,629 air filters. The dots above the upper control limit of the funnel chart (some of the extreme cases are highlighted in green) provide clear examples of air filters whose return rates are higher than expected, and these air filters (as applied to specific cars) were investigated to find the cause of the high return rate. There is a hyperbolic pattern of the return rates at small quantities, caused by similar return rates for various quantities bought. This could be due to parts being batched in specific quantities prior to returning.

3.2 | OEM

The attribute measurements that were used in the estimation of OEM factory standards were sparse. That is, of a possible 90 measurements for each part, only a few measurements were available to describe that part. This sparseness was visualised as a dendrogram and heatmap as follows.
Using fuel filters as an example, 30 000 samples of supplier parts were extracted. From this sample, a binary matrix was generated, where a value of 1 denoted an entry for the attribute to the part, and 0 denoted absence of any attribute measurement. A dendrogram was generated from this binary matrix which revealed which attributes are more likely to co-occur. This dendrogram was also used as the ordering for a heatmap.

Figure 2 shows both the dendrogram and the corresponding heatmap. Rows within the heatmap represent individual part descriptions and columns represent the attributes, such as height, length, and diameter. The heatmap colour gradient was generated according to

FIGURE 1 Funnel plot of return rates for 1629 air filter products [Colour figure can be viewed at wileyonlinelibrary.com]
Positive numerical values up to 500 (constrained to a maximum of 500) are shown as a yellow gradient. Text only values are fixed at −500 and shown as dark blue. Decimal values contained within a text/character string are extracted and multiplied by −100; these are shown as a blue gradient. Absent values (ie, there was no attribute measurement for the part) are shown as grey.

The level of sparseness shown in Figure 2 is evident, because all grey areas within this sampled data show where there was no data present, and grey areas constitute most of the heatmap. The dendrogram in Figure 2 and associated heatmap also indicates which attributes co-occur (for this part type, fuel filter). This co-occurrence is revealed by looking at the lowest branches of the diagram, ie, when two attributes are paired together. The first two pairings (from left-to-right) of “Outer Diameter” and “Height” are two attributes with both high frequency (populated often) and co-occurrence. Other points of interest are when the attributes shaded in blue co-occur (which is a decimal number contained within a string), such as “Thread Measurement 1” and “Thread Measurement 2.”

An example of the variation typically found for a given part across suppliers is demonstrated in Figure 3. Figure 3 shows the frequency distribution of the height attribute for 29 suppliers of fuel filters affiliated to a particular Ford car. The x-axis represents height (mm), and the y-axis is frequency. The plot demonstrates a pervasive problem in the automotive aftermarket—trying to determine the correct measurements of a part for a given car. In the graph, 29 height measurements of fuel filters (all listed against the same Ford car) are plotted as a frequency histogram. In this instance, 113 mm appears to be the correct measurement, but there is also range of other values (from 110.5 to 114.5 mm). The extent to which these different measurements are important will depend on how critical the height measurement is for this filter on this particular Ford car.

Estimations of the factory standards can be derived via inspection of all the supplier’s parts associated with each OEM standard. Using fuzzy matching to compare supplier parts with these estimated values of the OEM factory standards reveals those supplier parts which can be matched against other cars to which they were not originally intended to be matched, and in doing so highlights potential revenue for a part supplier, as they can sell these parts against more cars with very little modification. This is only possible due to the analysis of vast quantities of supplier data.

3.3 | Part replacement

The data for this analysis comprised 5,913,350 rows of invoice data covering the period January 2016 to July 2018. This dataset was then filtered according to a vehicle part (eg, brake disc), a specific make of vehicle and model (labelled as “small car A” in Figures 4, 5, 7, and 8). The car is referred to as “small car A” for reasons of business confidentiality. The data could also be filtered according to a specific year that a vehicle was manufactured, but for the example below, all

![Frequency of Height Measurements for Ford Fuel Filters](image)

**FIGURE 3** The frequency distribution of the height attribute for 29 suppliers of fuel filters affiliated to a particular Ford car. The x-axis represents the height (mm), and the y-axis is frequency [Colour figure can be viewed at wileyonlinelibrary.com]
cars with the desired make/model, manufactured from 2006 to present, were considered for the analysis. From a filtered dataset, a frequency histogram of the mileage (at the time of replacement) was generated, as shown in Figure 4.

From the above graph, an ECDF was generated, as shown in Figure 5. The ECDF was calculated according to the “ecdf” function within R from the library package “Hmisc.” Figure 5 shows the probability of vehicles arriving at a garage for a replacement/repair (y-axis) with at most the given mileage (x-axis). Plotting the ECDF to the replacement

FIGURE 4  The above figure shows a density plot of the mileage at the time of replacement for small car A. The x-axis is mileage, and the y-axis represents density.

FIGURE 5  The above figure shows the ECDF of the density data shown in Figure 4. The probability of a replacement/repair is represented by the y-axis, and the mileage at the time of replacement is represented by the x-axis. The sigmoidal curve in the graph is the ECDF for small car A—From Figure 4 [Colour figure can be viewed at wileyonlinelibrary.com]
of a vehicle part and the mileage (at replacement) aids in the process of identifying key points in a car's history when replacement is needed and potentially when failure may occur. This statistical technique determines the probability of an event using the observed data rather than a theoretical estimate of the probability.

This type of analysis extends to comparisons across vehicle make and model. For example, Figure 6 shows the ECDF for brake disc replacement for three large saloon cars (from different manufacturers of similar value). The graph in Figure 6 shows that at the 50% probability of part replacement, the vehicle represented by the top ECDF requires a part replacement sooner (≈ 66 000 miles) compared with the other two cars (≈ 90 000 miles).

The ECDF plot can be used to compare vehicle makes and models; however, any such comparison requires considerable assumptions to be made. We assume that the proportion of first and subsequent replacements is the same for each make and model. The assumption is reasonable for this data because each of the three makes and models have been in existence for a similar time period and a similar number of instances were used to generate each ECDF. The number of instances used to generate each of the ECDF curves are top plot 2596, and lower plots 2901 and 2515.

The histogram of mileage is a cross section of vehicles and includes those having their first and potentially subsequent part replacements. The natural distribution of mileage for first replacement is likely to be positively skewed. Combining mileage for several different subsets (first replacement, second replacement, etc.) has the effect of shifting the mode to the right. Hence, the results are conservative in that the vehicles are likely to have a lower mileage than that shown in the ECDF.

To demonstrate this, an Expectation-Maximisation (EM) algorithm for the inference of a normal mixture model was applied to the brake disc replacement data from Figure 4 (small car A). It was implemented in R via the “normalmixEM” function (from the “mixtools” library). This R function permits a range of inputs to specify the initial parameters for the centre (of the distributions), standard deviations, and lambdas for each normal distribution prior to running. However, many iterations showed that these starting values made little difference to the convergence of the output values, which are shown in the table below in Table 1:

These two distributions (from Table 1) are displayed below in Figure 7, overlaid on the density plot from Figure 4. The two curves in Figure 7 show the results of a finite mixture model applied to the density data (brake disc replacement for small car A), extracting two normal distributions. Assuming that there is more than one distribution contained within the underlying data, the left-hand curve in Figure 7 (red plot) represents the first replacement of brake disc for small car A, and the right-hand curve (blue/green plot) represents the second time the part was replaced.

The left-hand curve in Figure 7 was used as the basis for estimating the first time a brake disc was replaced on small car A. A total of 100 000 samples were generated from a normal distribution with a mean of 40 583 and a standard
deviation of $13\ 181$ (see Table 1, for distribution 1); this is shown as a density curve in Figure 8. Overlaid on the density plot of Figure 8, is the ECDF for this distribution (red curve). The red curve represents (potentially) the probability of the first time that a brake disc was replaced on small car A. This red curve is a left-ward shifted version of the curve shown in Figure 5, suggesting that the comparable probability of replacement as read from Figure 8, is sooner (50% chance of replacement at $\approx 40\ 000$ miles instead of at $\approx 55\ 000$ miles) than that suggested by Figure 5.

One point worth noting, however, is that the table of values shown above displays the lambda values for each distribution as 0.45 (for distribution 1) and 0.55 (for distribution 2). The lambda values for this model are the mixing proportions of each normal curve. This suggests that the red curve from Figure 7 has less area than the blue/green curve, which in turn suggests that there are more second time replacements recorded than first. There is no obvious reason why the data shown in Figure 4 would contain more instances of second time of replacement than first, and moreover, given the life expectancy of a car and the part, it would make more sense for there to be more first-time replacements than second.

The density data in Figures 4 and 7 show the part replacement for small car A, which could also be considered a measure of part failure (in this case for brake discs). The distribution of the mileage data is skewed to the right which is not uncommon for failure data, and a log transformation is indicated. The Expectation-Maximisation (EM) algorithm for the inference of a normal mixture model was applied to the log of the mileage data to discover the nature of the underlying distributions, given this transformation; however, the results gave similar but more extreme results to that of the untransformed data.

It is possible that the largest distribution (the blue/green shaded area) in Figure 7 contains other distributions corresponding to third time (or more) of part replacement. For this reason, another normal mixture model was applied to the brake disc data in which three distributions were extracted; the results are listed below in Table 2.
Table 2 shows that even with three distributions, the second distribution still has the largest contribution with a lambda value of 0.62. Re-running the analysis with log transformed data did not converge consistently on lambda or centre values.

A parametric stochastic EM (St-EM) algorithm for a univariate finite mixture of Weibull distributions was also applied to the same mileage data, using a two-component mixture. It was implemented in R via the “weibullRMM_SEM” function (from the “mixtools” library). The lambda values for the first and second components are lambda 46% and 54%, respectively, again suggesting that there are more second time replacements recorded than first.

These counter-intuitive findings suggest that more modelling may be required on the density data to derive probability curves for the first time of part replacement and a more detailed investigation to find more explanatory data around the mileage figures.

4 | DISCUSSION

Data science requires a combination of IT skills, application of statistical analysis, and machine learning techniques and business know-how. We have demonstrated that considerable, valuable insight can be obtained from applying data science techniques to automotive aftermarket sales data.
The ideas for analysis arise in two ways. SMEs set up to capitalise on the growth of data, benefit from the ample ideas from customers giving a pull to the data analytics carried out by the SME. In addition, there is a push aspect to the data analytics in which ideas that arise as a result of the analysis can be offered as an additional new service. One such example was highlighted previously in the results section when it was suggested that during the process of determining OEM measurements, it is possible to find parts which can be applied to cars not previously considered, by matching measurements of those parts. The current limitations of data collection and analysis should serve as motivation for further analyses and collection methods. For example, a complete dataset of all part measurements (as described in case study 2) would improve the certainty of OEM estimates. Another example was highlighted in the results section which described how probability functions (ECDF) can be derived for part replacement. It was noted that in addition to considering a probability of part replacement for a single vehicle type, ie, make/model, a comparison can be made across make/models. From this, it is possible to provide a service to advise the second-hand vehicle market how expensive one particular vehicle make/model may be to repair and service during the car’s life compared with other make/models.

The analytics is only part of the story as the results then have to be presented in an accessible form for the customer. This step is nontrivial and again uses the IT skills so vital in integrating the data and preparing it for analysis. User interaction has to be made clear and straightforward. It has to be decided what the customer wants and needs, and what can be offered. The data visualisation then has to be made available using a full screen display and including as much flexibility as possible.

The SME needs to ensure a revenue stream from the work, and this can arise from refreshing the results that the customer pays for on a regular basis, eg, a 3-monthly update. The SME’s intellectual property (IP) has to be protected and one way to address this issue is to combine the SME’s domain knowledge with the data analytics so that they have a unique selling point and can offer a product that is not easily copied by their competitors.

Embarking on monetising data requires initial effort before there are any financial gains. SMEs are typically challenged by the difficulty of obtaining finance, expert staff, and consultancy facilities, and are hampered by lack of expertise in data science. The UK government helps selected SMEs by providing up to 2/3 funding via Innovate UK to set up a Knowledge Transfer Partnership (KTP) to support a 1 to 3-year project in which a post graduate research associate is employed by a partner university but works full-time in the SME on a substantial project likely to have demonstrable impacts and improve the profitability of the SME. The KTP project needs to provide the academics with new research experience and useful ideas and data for teaching as well as embedding a new capability in the SME and improving the productivity of the UK. Other options to help SMEs embrace the benefits of data science include SMEs providing work-based projects for graduate students supervised by academics or encouraging staff to educate themselves via online courses and continuing professional development.

A key to successful analyses in any sector is data quality, both in terms of completeness of the sample records and also regarding representativity for the underlying population. Improving these whenever possible will in turn always improve the quality of the analyses. For the studies mentioned here, historic data were often not compatible with current datasets (meaning database tables could not be joined together). Better quality would come from a tighter integration between data sources. Data harmonisation techniques can contribute to this task but are sometimes labour-intensive to set up (eg, building a whole ontology of potential variables and their possible relationships). On the other hand, if data sources are designed independently, often only a small fraction of the variables can effectively be harmonised, with the consequence of having to drop large amounts of the original data. From the perspective of automotive aftermarket data analyses, it would be advantageous combining supplier and competitor records. However, this is overly optimistic because there is often poor harmonisation even within a company. That said, aiming to prospectively agree on a common data format and standard will help to minimise poor compatibility.

ORCID
Wayne S. Smith https://orcid.org/0000-0002-9344-6444
Jaume Bacardit https://orcid.org/0000-0002-2692-7205

REFERENCES
1. See https://www.smmt.co.uk/vehicle-data/.
2. Ahlemeyer-Stubbe A, Coleman S. Monetising Data—How to Uplift Your Business. London, UK: Wiley; 2018.
3. Smith W, Coleman S, Bacardit J, Coxon S. Optimising the automotive aftermarket using data science. Poster presented at Automechanika, Frankfurt, 11-15th September. 2018. https://automechanika.messefrankfurt.com/frankfurt/en.html.
4. Smith W, Coleman S, Bacardit J Coxon S. How data can change online publication. 2018. https://www.warrantyweek.com/archive/ww20141204.html the automotive aftermarket. Focus, p30–32, October, www.ciltuk.org.uk.

5. Smith W, Coleman S, Bacardit J. Insight into aftermarket automotive sales, factory standards and predicting auto part replacement. Abstract presented at ENBIS 2018 in Nancy, France. 2018.

6. Spiegelhalter DJ. Funnel plots for comparing institutional performance. Stat Med. 2004;24(8):1185-1202. https://doi.org/10.1002/sim.1970

7. Coleman SY, Gob R, Manco G, Piegatolo A, Tort-Martorell X, Reis M. How can SMEs benefit from big data? Challenges and a path forward. Qual Reliab Eng Int. 2016;32(6):2151-2164. https://doi.org/10.1002/qre.2008/full

8. Mustafazade F. Using social science data to solve a social housing problem. ESRC Case Study. 2018. https://blog.esrc.ac.uk/2018/10/19/using-social-science-data-to-solve-a-social-housing-problem/

**AUTHOR BIOGRAPHIES**

**Wayne S. Smith** is employed by Newcastle University (School of Mathematics, Statistics and Physics) as a KTP Research Associate to work for Rain Data Ltd based in Northumberland, UK. His work focuses on providing insight into the automotive aftermarket from online transactional data, using both traditional statistical methods and machine learning techniques. Wayne gained his PhD at Newcastle University in 2005 in Psychology/Neuroscience and has since worked for Newcastle University and Durham University in various research based and consultancy roles.

**Shirley Coleman** is Technical Director of the Industrial Statistics Research Unit, School of Mathematics, Statistics and Physics, Newcastle University, and a visiting scholar at the Faculty of Economics, Ljubljana University, Slovenia. She works on data analytics in small and medium enterprises (SMEs) and contributed a highly ranked impact case study to Newcastle University’s Research Excellence Framework. She is the academic lead on Innovate UK-funded Knowledge Transfer Partnerships developing data science capabilities and specialises in statistical and machine learning techniques applied to company data. She publishes in trade and academic journals and is co-editor of several books. She is a past President of the European Network for Business and Industrial Statistics (ENBIS), an elected member of the International Statistics Institute, and a Chartered Statistician of the Royal Statistical Society, instrumental in mentoring early career statisticians and developing relationships with business and industry.

**Jaume Bacardit** has received a BEng, MEng in Computer Engineering, and a PhD in Computer Science from Ramon Llull University, Spain, in 1998, 2000, and 2004, respectively. He is currently Reader in Machine Learning at Newcastle University in the United Kingdom. Bacardit’s research interests include the development of machine learning methods for large-scale problems, the design of techniques to extract knowledge and improve the interpretability of machine learning algorithms, and the application of these methods to a broad range of problems, mostly in biomedical domains.

**Syd Coxon** has had a 25-year-long and varied career in IT, starting out in technical support, progressing through to development, management, and consultancy roles. His passion for data and learning through data meant that the consultancy and development roles he took on usually involved database design and integration of large volumes of data. This experience lead him to work for Euro Car Parts–the largest parts supplier in the UK automotive aftermarket–to create an electronic car parts catalogue for their branch network as well as Ecommerce websites. The catalogue became widely recognised as best in class and was an essential part in the growing success of Euro Car Parts. Euro Car Parts has many large sister companies throughout Europe, all owned by their parent company LKQ. Syd has been heavily involved in catalogue data integration across the whole LKQ group. In September 2016, Syd set up his own company, Rain Data, which specialises in data conversion and data analytics for a variety of clients predominantly from the automotive aftermarket. Establishing Rain Data was the fulfillment of a long-held ambition–to have a company specialising in data.

**How to cite this article:** Smith WS, Coleman S, Bacardit J, Coxon S. Insight from data analytics with an automotive aftermarket SME. Qual Reliab Engng Int. 2019;35:1396–1407. https://doi.org/10.1002/qre.2529