*International Journal of Scientific Research and Modern Technology (IJSRMT)* *ijsrmt.com*
*Volume 2, Issue 3, 2023*
*DOI: https://doi.org/10.38124/ijsrmt.v2i3.365*

# Navigating Data Science: Current Insights and Future Perspectives in Business and Industry

Dr. Glory Sujitha Antony[1]; Amala Deepan Arulanandham[2]

[1]Senior Lecturer, School of Information Technology, IBSUniversity, Papua New Guinea
[2]Technical Expert, TwinITechnologies, Chennai

Publication Date : 2023/03/28

**Abstract**

This paper aims to frame Data Science, an increasingly fashionable and emerging topic, in the context of business and industry. A discussion is initiated on the origins of Data Science and its demanding blend of skills in data analytics, information technology, and business expertise. The discussion about data science and its requirements is vital in data analytics, information technology, and several other businesses. Data science aims to provide new or revised computational theories capable of extracting useful information from large amounts of data. This paper also involves digital data collected from images, text, audio, sensors, and traditional measurements. The different and most popular methodologies amongst the Data Science research and applications practitioners are then reviewed. Since the emerging field requires personnel with new competencies, an attempt is made to describe the Data Scientist profile, which is considered one of the most demanding jobs of the 21st Century, according to Davenport and Patil. Most people widely recognise the need to embrace Data Science but often feel intimidated by their lack of understanding and worry that their jobs will disappear. They should be encouraged: Data Science is more likely to add value to existing employment and enrich working people's lives by helping them improve by providing an informed business decision-making process. The paper concludes by presenting examples of Data Science in business and industry.

***Keywords:*** *Knowledge Discovery, Data Scientist Profile, Business Improvement, Industry 4.0, SME.*

## I. INTRODUCTION OF A DATA SCIENCE

This paper is motivated by the need to frame the research area of Data Science within the context of business and industry. Over the past three decades, Data Science has evolved and expanded, attracting interest from researchers across multiple scientific disciplines. The growing significance of Data Science is observed in both public and private organisations. For instance, in e-commerce, custom-made product recommendations and targeted advertisements can be generated if customer web activity data are accurately interpreted. Similarly, demand forecasting and logistics management can be optimised when properly analysing sales data.

In the healthcare sector, where diagnoses and treatments are increasingly digitised and recorded, the application of Data Science methodologies enables the prevention of misdiagnoses, the identification of optimal care plans for patients, and the enhancement of treatment quality. Likewise, data collected from various workflow stages in industrial production is crucial in improving product quality, detecting failures, and optimising speed and performance.

At present, decision-making processes within organisations are increasingly driven by data. With the availability of open-source software, sensors and processing tools have become accessible to small and medium enterprises (SMEs), making Big Data support essential for maintaining a competitive edge (1). The future presents vast opportunities for deeper insights into reality, whether through integrating linked datasets, as seen in Industry 4.0 (2) and new governmental initiatives aimed at connecting administrative datasets or exploring previously uncharted data sources (3).

A U.S. survey conducted by KPMG, based on a sample of 400 CEOs, revealed that approximately 77% of respondents expressed concerns regarding the quality of the data on which their decisions rely. If such concerns hold, the efforts invested in data analysis may prove ineffective. One challenge could be the lack of effective communication between those presenting and interpreting

Sujitha Antony, D. G., & Arulanandham, A. D. (2025). Navigating Data Science: Current Insights and Future Perspectives in Business and Industry. *International Journal of Scientific Research and Modern Technology*, *2*(3), 1–11. https://doi.org/10.38124/ijsrmt.v2i3.365

data. Analysts and decision-makers often operate in isolation, leading to a breakdown in the communication and development process. The full benefits of data science can only be realised if data scientists work alongside managers during decision-making and assist in interpreting data processing methods. Conversely, Managers should also engage with Data Scientists to ensure that business objectives are reflected in data analysis, leading to more informed and goal-oriented decision-making (2).

To assist researchers interested in exploring the themes and challenges of Data Science and individuals aspiring to pursue a career in this field, an overview of its origin and development is provided in Section 2. Sections 3, 4, and 5 review Data Science methodologies, key figures, and applications, respectively. Examples of Data Science projects are discussed in Section 6, and concluding remarks and future perspectives are in Section 7.

## II. THE EVOLUTION OF DATA SCIENCE: ORIGINS, GROWTH, AND DEVELOPMENT

A continuous flow of data is generated by the internet and by sensors attached to modern equipment, accumulating vast amounts of information. If data are not effectively managed and processed, companies risk being outperformed by competitors that successfully leverage these resources. This scenario led to the emergence of Data Science and continues to drive its rapid development, which has progressed in tandem with explosive technological advancements.

Data Science is recognised as a discipline that provides methodologies for processing and interpreting massive volumes of data collected through increasingly advanced devices. These analytical tasks are often challenging to accomplish using only traditional statistical methodologies. Today, Data Science is an interdisciplinary field encompassing mathematical methods, statistics, algorithm development, qualitative analysis, computer science, and a practical approach to extracting valuable insights from data. This includes structured data, such as quantitative information in predefined formats, and unstructured data, including reports, visuals, and audio.

Data Science was initially used by Turing Award winner Peter Naur in 1960 as a synonym for computer science. Later, in 1974, Naur employed the term to refer to data processing methods across a broad range of applications (4). However, it was in 1996 that the term first appeared in a public setting when the International Federation of Classification Societies organised a conference in Kobe, Japan, under the title Data Science, Classification, and Related Methods.

Since then, the international community has widely adopted the term Data Science to describe an interdisciplinary field. However, numerous debates have occurred over time regarding its distinction from or standardisation of Statistics (5). For instance, during his inaugural lecture for the H. C. Carver Professorship in Statistics at the University of Michigan in 1997(6), C. F. Jeff Wu proposed that Statistics be renamed Data Science and that Statisticians be referred to as Data Scientists(33).

Modern methodologies, however, have increasingly merged the statistics and computer science fields. This integration is evident in the interaction between computational algorithms and cognitive science in artificial intelligence, as well as in the perspective of machine learning, which is viewed as a convergence of statistics and knowledge representation (7).

- *Leo Breiman, a Statistician at the University of Berkeley, has been of a different opinion. In 2001, he said:*

*Two cultures are using statistical modelling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theories and questionable conclusions and kept statisticians from working on many interesting, current problems. Algorithmic modelling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modelling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.* (8)

- *Breiman's attitude is reminiscent of a famous phrase by George Box, who said:*

*All models are wrong, but some are useful*. (9)

The distinction between Statistics and Data Science is now widely recognised, including within university curricula, where classical statistical analysis continues to be taught in Statistics courses. In contrast, newly introduced data science courses provide a broader education incorporating algorithms and machine learning.

In 2001, Data Science was introduced as an independent discipline by William S. Cleveland (10). He identified six technical areas that define the field: multidisciplinary investigations, models and methods for data, computing with data, pedagogy, tool evaluation, and theory. Subsequently, in 2002, the International Council for Science: Committee on Data for Science and Technology (CODATA) established the first academic journal dedicated to Data Science (https://datascience.codata.org/) (9). Shortly thereafter, Columbia University launched the *Journal of Data Science* (http://www.jds-online.com/), providing a platform for professionals to share their perspectives and exchange ideas.

As a result, in recent years, a growing body of literature on Data Science has emerged, fostering information exchange and forming communities united by an interest in data analysis and its applications. For instance, modern terminology such as Fintech Data Science has been adopted in finance and financial technology. One of its proponents, Giudici (11), has suggested using the term Data Sciences to emphasise that Data Science encompasses an integrated process of activities—including defining analytical objectives, selecting and processing data, statistical modelling and interpretation, and implementing and evaluating statistical measures. This perspective highlights the field's multidisciplinary nature, given the varying objectives and terminologies across different knowledge domains.

Moreover, since the early 21st century, Data Science has expanded into various subfields, including Knowledge Discovery in Databases, Data Mining, Artificial Intelligence, Machine Learning, and Deep Learning, setting the foundation for further advancements in related areas. Established the European Association for Data Science in Europe in 2013, it has significantly promoted and popularised the terms *Data Scientist* and *Data Analyst*, particularly within the business sector.

In summary, the primary objective of Data Science is to clean, prepare, and analyse diverse datasets to extract meaningful insights. While closely related to Statistics, Data Science uses statistical methods to achieve its goals. However, it is also supported by information technology—mainly programming for big data analysis—emphasising a practical approach and a strong focus on decision-making implementation.

## III. DATA SCIENCE: KNOWLEDGE DISCOVERY IN DATA BASES

Since the early 1940s, numerous new terms have been introduced into everyday use, primarily due to the advancements in Information Technology. Therefore, distinctions need to be made between data (the essential elements, typically composed of symbols), information (the result of processing data aimed at organising, interpreting, and contextualising data to acquire meaning), and knowledge (a set of organised and processed information intended for disseminating experience, understanding, and competencies related to practical problems in industry and business). A chain of acquaintance can be envisioned, moving from data to knowledge, with the ultimate goal being formulating actions, behaviours, and decisions. The transition from data to knowledge is thus strategically important, forming the foundation of Data Science.

To achieve this objective, Data Science utilises methodologies from fields such as Mathematics, Statistics, and Computer Science to analyse and mine data for knowledge extraction. Its impact is so significant that it is now considered the fourth paradigm of science, alongside empirical, theoretical, and computational science, with Data Science being data-driven. Consequently, data science is part of a more prominent family of research domains, including academic researchers and private companies, and is known as knowledge discovery in databases (KDD). KDD is the process that extracts high-level knowledge from low-level data. Today, it is one of the most engaging, dynamic, and rapidly evolving fields, driven by continuous improvements in data warehouses and the growing use of Big Data and Big Data Analytics, which aim to convert vast, heterogeneous, and persistent data into usable information.

Big Data originates from increasingly innovative multimedia sources and is characterised by its vastness, making traditional statistical methodologies inadequate. It is possible that some proponents of Big Data claim that such an abundance of data implies that analysis no longer requires a theoretical foundation. In 2008, Chris Andersen (12) published a provocative paper titled *The End of Theory: the Data Deluge that Makes the Scientific Method Obsolete*. In this paper, he states:

"All models are wrong, but some are useful." So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behaviour, seemed to be able to consistently, if imperfectly, explain the world around us until now. Today, companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to pay for models at all. Scientists are trained to recognise that correlation is not causation and that no conclusions should be drawn based on a correlation between X and Y (it could just be a coincidence). Instead, you must understand the underlying mechanisms that connect the two. Once you have a model, you can confidently connect the data sets. Data without a model is just noise. However, faced with massive data, this approach to science — hypothesise, model, test — is becoming obsolete.

This argument sparked a vibrant debate that continues today within the scientific community. While this paper does not actively engage in the discussion either in favour of or against Andersen's view, the opportunity to express our opinion on statistical models cannot be overlooked. The valuable model referenced by Box is tied to a physical or simulated experiment with the aim of prediction. Such a model is not required to replicate the complexity of a phenomenon but rather to provide significant support to the prediction algorithm. There is a pressing need for new computational theories and tools to assist humans in extracting valuable knowledge from the ever-growing volume of digital data, as highlighted by Usama Fayyad et al. (13) and other Data Science practitioners.

Most of the work is still managed by humans, who must make the correct decisions to avoid compromising the data and, more importantly, to ensure proper data significance. Adequate prior knowledge of the problem is

necessary to correctly handle data research, preparation, selection, and cleaning, ensure appropriate data use, and interpret the mining results correctly to avoid misunderstanding the information extracted from data analysis. The KDD process in databases is essential despite being complex. It may require multiple iterations, but it will lead to success in data research problems if human interaction and participation are effectively managed. According to Usama Fayyad et al., who outlined the essential and practical steps of the KDD process (14), the steps are as follows:

- The application domain should be understood, adequate a-prior knowledge should be gained, and the goal of the KDD process should be correctly identified from the customer's perspective.
- A target data set should be selected or created on which the discovery will be performed.
- Data should be cleaned and pre-processed: possible noise should be removed by collecting sufficient information about how to identify it; strategies should be decided for missing data or missing data fields; time-sequence information and known changes should be accounted for.
- Useful features should be identified to represent the data, depending on the task's objective, trying to reduce dimensions or find invariant representations to ease research.
- A data mining method based on statistics that conforms to Step 1 (summarisation, classification, regression, clustering, etc.) should be identified.
- An exploratory analysis, model selection, and hypothesis formulation should be performed, and algorithms and methods for searching data patterns should be chosen.
- Real analysis should be done using data mining methods (including classification rules or trees, regression, and clustering).
- The identified data patterns should be explained, potentially including data visualisation to extract useful visual information.
- Actions should be taken on the discovered knowledge: it should be used directly, incorporated into other systems for further actions, or documented and reported. Potential conflicts with previous beliefs should be checked and resolved.

Once the initial steps have been completed and the necessary data has been identified, the Data Mining process becomes the central core applicable to various purposes, such as verification of hypotheses, discovery of previously unseen patterns, or description, with patterns represented in a form understandable to humans. Most Data Mining methods are based on established Machine Learning, Pattern Recognition, or Statistics techniques.

Data mining is often associated with data science and is sometimes confused due to the extensive use of mathematical and computer science methods in both disciplines. However, the two fields are distinct: Data Mining is a process aimed at discovering patterns in large datasets (with prior knowledge), extracting information, and transforming it into an understandable structure for further use, typically to inform decision-making. In contrast, Data Science is a broader field that includes data cleansing, preparation, and final analysis of data sets of all types.

Lastly, the main focus of data science is data-driven decision-making, where decisions are made based on thorough and accurate data analysis rather than intuition or guesswork.

## IV. THE DATA SCIENTIST

After framing the concept of Data Science, the question arises: who is expected to be the Data Scientist? A profile of the Data Scientist is provided. The data scientist is responsible for understanding how real problems relate to available data, enabling them to use data best to create added value. Data management, integrity, and accessibility are guaranteed by the Data Scientist, who mines valuable information from data to provide knowledge, make predictions, and support decision-making.

The competencies expected from a Data Scientist are crucial. First, the Data Scientist must possess the ability to analyse problems logically while understanding the underlying business or practical issues. Familiarity with a wide range of techniques is also essential. When searching for a "data scientist job description," many websites provide detailed lists of these competencies (approximately 144,000,000 results, according to Google).

➢ *The Consensus is that the Following Statistical Methods Should be Known by any Data Scientist:*

- High-dimensional geometry, where vectors with many components represent situations in data, even when this representation is not the most natural choice for data collection.
- Singular Value Matrix Decomposition, Principal Component Analysis, and Matrix Algebra should be familiar to the Data Scientist.
- Clustering, which involves partitioning a set of elements into subsets based on specific criteria, is a standard tool in Statistics.
- Random Graphs, Random Trees, Random Walks on Directed Graphs, Markov Chains, and Neural Networks are all widely used tools.

Moreover, machine learning (ML) competence is highly desirable for data scientists. ML, an interdisciplinary field combining applied statistics and computer science, aims to estimate complex functions using algorithms that can learn from previous data. These algorithms can adapt to the analysed problem, processing large data sets in training and test phases and concepts linked to Big Data. Analytical models are usually developed on a training data set, tested on a separate test data set, and verified on additional data before being

applied to the entire data set to ensure their applicability. This approach contrasts with traditional statistical analysis, where data may be scarce, and models are assessed using leave-one-out techniques.

Algorithm accuracy is evaluated based on the task the algorithm is designed for, the performance it delivers, and the experience it gains during the learning process. The fewer data required to achieve an adequate analysis, the better the algorithm is considered.

In discussing the required competencies for Data Scientists, we also mention the distinction between supervised and unsupervised learning algorithms. Unsupervised learning algorithms experience a data set containing features and attempt to learn valuable properties, such as the probability distribution from which the data set originated or methods for de-noising it. Supervised learning algorithms, on the other hand, experience data sets where each example is associated with a label or target value. The algorithm typically uses the underlying probability distribution to predict the label or target from the features. However, the distinction between the two is unclear, as the same algorithm may be used in both cases with minimal adjustments.

Some problems, especially those in Artificial Intelligence tasks, cannot be solved by ML. Deep Learning (DL) or Deep Neural Networks has emerged as a new autonomous discipline to overcome the limitations of classical ML algorithms. Although DL is derived from ML, it can be considered a part. Various DL algorithms exist, each adapted to specific tasks such as object recognition, speech recognition, and bioinformatics. DL algorithms utilise multi-layered networks of nonlinear computational units, similar to Neural Networks, and have achieved significant success, such as outperforming humans in the ImageNet challenge (15).

For a comprehensive examination of how Statistics has evolved in recent decades, the book by Efron and Hastie (16) provides an insightful discussion. The Data Scientist is not expected to be a Computer Scientist but must possess a fair degree of familiarity with Information Technology. The Data Scientist combines Statistics, Mathematics, and Computer Science expertise. In today's digital age, the professional who can manage data and extract value from it will be highly sought after. Davenport and Patil (17) aptly referred to the Data Scientist as "the sexiest job of the 21st century."

The growing demand for Data Scientists has led universities to offer new curricula designed to train them. These courses, mainly at the postgraduate level, emphasise solid theoretical foundations in Statistics and Computer Science and practical experience in various applications. There is a need for an intense exchange between academia and industry, which can be facilitated by supervising postgraduate projects and placements as part of Data Science training (18).

Data Science Master's degrees are of great interest to university faculties due to their potential to attract students and foster collaborative projects. As the demand for Data Scientists increases, many programs have been fast-tracked into existence. Brown (19) at the University of Canterbury described a method for determining the necessary content and structure of a Data Science Master's program, resulting in a 12-month degree with modules in data analytics and an industry project.

We conclude with a statement from James Stephen Marron, Professor at the UNC-CH Department of Biostatistics, University of North Carolina: "I think it's time for Data Science to consider the concept of teamwork. Data Science problems solved by one person are mostly complete, but big challenges require a team of data scientists with diverse skill sets." The growing importance of Data Science, particularly in SMEs, demands the collaboration of experts from various fields. The T-shaped model, which combines broad knowledge and deep expertise, is essential for Data Scientists and beneficial in everyday life.

## V. THE ROLE OF DATA SCIENCE IN SHAPING BUSINESS AND INDUSTRY

The development of Data Science has been driven by the explosion of data in the digital age. It is important to note that many of the mathematical, statistical, and machine-learning techniques used in Data Science have existed for many years. What has changed is the availability of massive amounts of data, which is now stored rather than merely observed and overwritten. Recently, there has also been a realisation that profound insights and business advantages can be gained from analysing this data.

Data Science as a profession is increasingly seen as high-paying, glamorous, and in high demand. Previously, the creative "Madison Avenue" professionals were responsible for selling most effectively. Behavioural profiling and customer segmentation techniques employed by mathematicians and statisticians have proven that scientifically targeted advertising is far more profitable than even the most elegant and subtle advertising campaigns. New digital displays and fast, interactive processing of customer profiles allow promotions to be presented to people who are most likely to be susceptible. As a result, the "math men" overtake the "mad men."

Data Scientists are now recognised as valuable and desirable professionals. A lively debate has occurred regarding the relationship between Statistics and Data Science (20). Statistics plays a crucial role in Data Science, and it is worth considering how statisticians should manage that role. Many data analysis practices involve opaque solutions, where data is fed into a "black box," calculations are made, and an answer is provided. Black box techniques conceal the algorithms being used and only report the outcome. While there may be valid reasons for using a black box—such as complexity or preventing

interference—the downside is that these techniques can alienate statisticians and foster a careless attitude toward statistical details. This approach may be damaging in the long term.

Black box data analytical solutions for prediction often lack robustness against changes in influential variables within data sets. These solutions are often based on assemblies of models, and predictions are averaged using various methods, not all of which are suitable for the type of data being input. Determining which predictors have had the most significant impact on the prediction is difficult. One predictor may be easier to collect than others and have higher quality. Still, its advantage may not be recognised unless further checks are made to examine the effect of swapping variables. Black box users are not forced to check their data before analysis, and there is often little emphasis on residual analysis, meaning variables with gross outliers are not detected. Users may overlook data errors and opportunities to identify key subsets or obvious explanations for apparent patterns. However, the black box user tends to overlook these issues as long as their prediction appears better than before and works in the short term. In summary, the problems with relying on black boxes are as follows:

- Algorithms, tuning parameters, subtle effects, and assemblies are not accessible;
- Black boxes may suffice for short-term solutions, but robustness to change is uncertain;
- The skills required to understand the analytical methods are not valued or developed.

Business people prefer black-box approaches as they are superficially more straightforward to understand. Statisticians must reclaim the field rather than let core black box services take over. They must stand up for the importance of checking models and not simply accepting the solutions offered. Statisticians must advocate for their vital roles, such as checking data quality, evaluating the costs of obtaining variables and using proxies, conducting sensitivity experiments, and constructing and validating models.

Companies have a love-hate relationship with data science based on a fundamental fear of numbers and a strong desire for the benefits of analysis (21). It is essential to encourage staff by emphasising the many positive outcomes of Data Science, such as reducing wasted time and effort, allowing more focus on core work, and enhancing profitability and job security.

The growth of data science positively affects academia, business, and industry. It draws attention to statistical techniques and raises awareness of data in various aspects of life, similar to how Six Sigma, Total Quality Management, and other business improvement initiatives did (22). Data Scientists need to be numerate, and the growth of Data Science fosters enthusiasm for education in STEM subjects at all levels. Besides undergraduate and postgraduate degrees, opportunities for vocational training are also emerging. Thus, data science boosts interest in mathematics and statistics from high school to tertiary education and professional development. An example of government interest in this field is the Knowledge Transfer Partnerships funded by Innovate UK. These partnerships provide up to two-thirds of the costs for one-to-three-year projects in which a postgraduate research associate is embedded in a company to develop new expertise supervised by a university academic. These projects offer an effective way for SMEs to grow their business with state-of-the-art guidance and reduced financial risk (23).

In some respects, we are in a golden era, with great opportunities offered by the expansion of Data Science before data analytics becomes so entrenched that fewer creative opportunities remain for developing bespoke solutions. It is vital to ensure that the quality and integrity of Statistics are maintained, not only for professional reasons but more importantly because only when Statistics is applied correctly and sensitively can the full benefits of Data Science be realised.

## VI. DATA SCIENCE IN PRACTICE: SUCCESS STORIES

Some examples demonstrate the collection of specialist skills required for data science to be practical. The business need is clearly stated in each case study, the data integration is explained, and the statistical analysis and outcomes are described.

➢ *Automotive Retail Sector*

A massive quantity of data is generated by the automotive after-sales business. Each day, catalogues are consulted by vehicle owners and service engineers to find the right parts for vehicle repairs. This case study involves a small to medium enterprise (SME) handling big data from catalogue look-ups and other data associated with the automotive after-sales market sector. The business motivation is to increase service offerings to customers, focusing on issues identified by the customers themselves and other opportunities that emerged after exploratory statistical analysis and data visualisation were applied to the data. Empirical data analysis can be used to explore various scenarios. By applying IT skills to amalgamate diverse sources of information followed by statistical analysis to identify patterns, valuable business insight can be generated. This enables the SME's current business to be improved, and new products and revenue streams may potentially be identified (24). Three examples of mileage, return rates, and original equipment manufacturing are considered.

Firstly, the vast amount of available data highlights distinct differences across different vehicle types when looking at the mileage of vehicles coming for repair. Data on the vehicle and the parts fitted are integrated and checked in preparation for analysis. Data dimensions include the making and model of the car, type of repair, vehicle age, date of visit to the garage, mileage at the time

of repair, and parts fitted. The data are analysed by constructing empirical cumulative distribution curves for mileage for specific makes and models of vehicles and types of repair. Figure 1 compares the curves of mileage in vehicles coming to the garage for brake disk replacement for three popular saloon cars (33). It is shown that cars of type A go to the garage with much lower mileage than cars of types B and C (33). In this dataset, 50% of vehicles of type A have mileage of 65,000 or less, whereas for cars of types B and C, the percentage is closer to 25%. This suggests that cars of type A require brake disc replacement earlier and, in this sense, are considered less desirable.

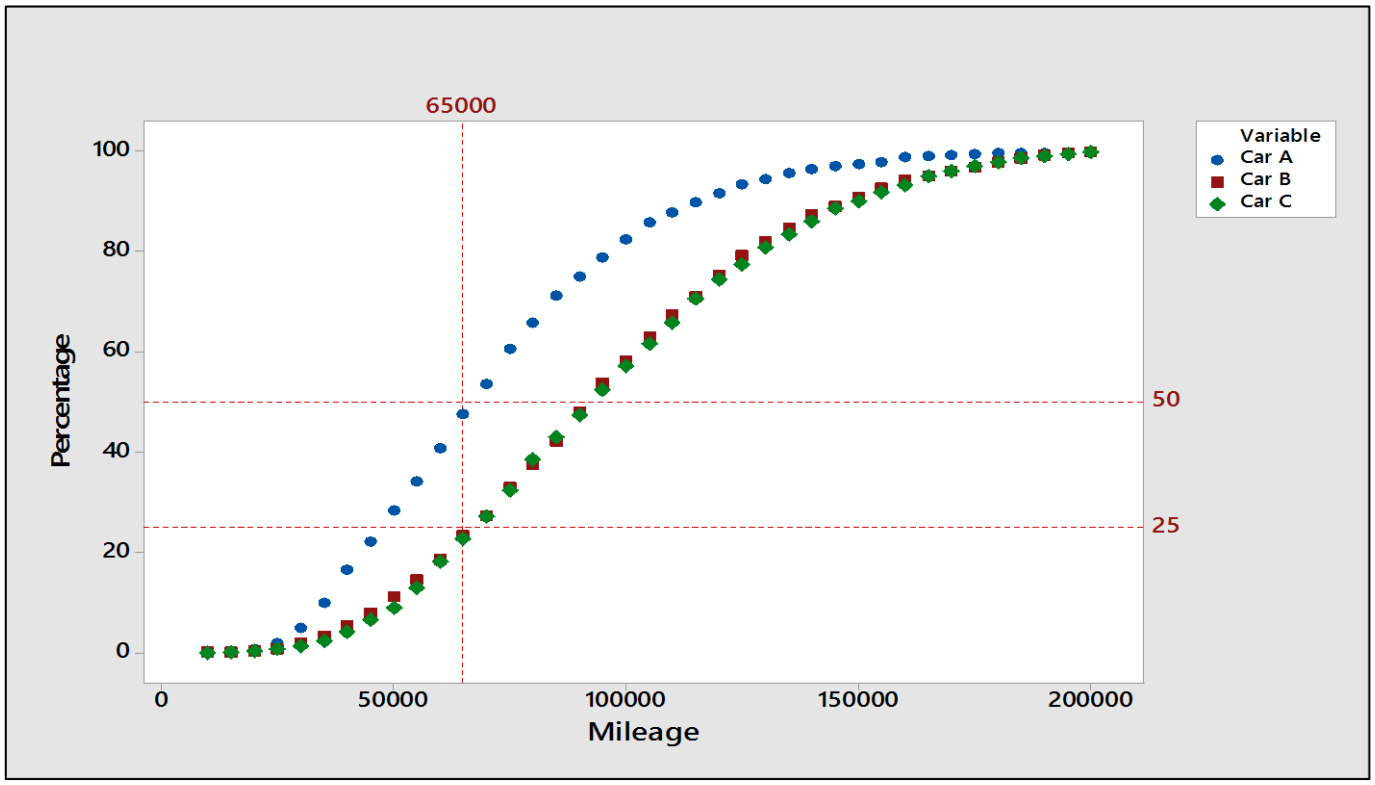


Fig 1 Brake Disc Replacement for Three Popular Saloon Cars (1)

The outcome of the Data Science analysis is an empirical tool that can be updated as more data arise and can be used by the garage to determine necessary stock levels, anticipate co-morbidities when vehicles come in for a particular repair, and provide a service to customers considering the purchase of a used vehicle.

Secondly, return rates for specific items bought from a catalogue are shown to differ widely; those items exceeding control limits in a funnel plot of return rates warrant further investigation, which may include examining how the parts are presented in the catalogue.

Thirdly, the original manufacturer is the only party that knows precisely which original equipment was fitted to a specific vehicle, and their parts are typically more expensive than the generic copies made by other companies. When buyers choose between alternative products, uncertainty exists, presenting business opportunities for data analytics. Suppliers of a particular spare part for a vehicle were interested in knowing whether their part would fit vehicles other than the one it was designed for. A search through the database of part dimensions and a matching process led to identifying several cars that could also use the part (24).

This case study illustrates the value added to administrative and operational data from various company sources. It demonstrates improvements in garage service offerings, catalogue clarity and usability, and supplier reach.

➢ *Shipping Sector*

The shipping industry, the nucleus of global trade, is susceptible to fuel prices. Fuel costs represent over 50% of the total operating costs of a vessel (26) and contribute significantly to the overall transportation cost of cargo. Fluctuations in crude oil prices and stricter environmental regulations on the emission of noxious and greenhouse gases are influential factors in the operation of the shipping industry. Data Science has been employed to develop new products offered by an SME dealing with shipping performance data.

Sensors attached to a ship's engines record fuel consumption, which is displayed in real-time on the ship's bridge. In addition to providing a check on fuel theft and gross errors in engine function, this data can be utilised for other purposes. The business motivation behind this case study is to provide a decision support tool to aid the scheduling of ferry services.

Fuel consumption is linked to the ship's speed over the ground. Company operational data becomes especially useful when enhanced with open data on weather and tides. The weather components that most affect fuel

consumption are headwind and crosswind. Tidal data is also essential in specific shipping routes, and in this case study, tides around the UK are quantified using algorithms and data from freely available websites.

Company data on fuel consumption across the journey of a ferry undertaking a daily return service on a specific route is integrated with weather and tide data after adjustments for location and time granulation issues. The data are analysed using multivariate statistical analysis to create a predictive model for fuel consumption based on known tides and predicted weather.

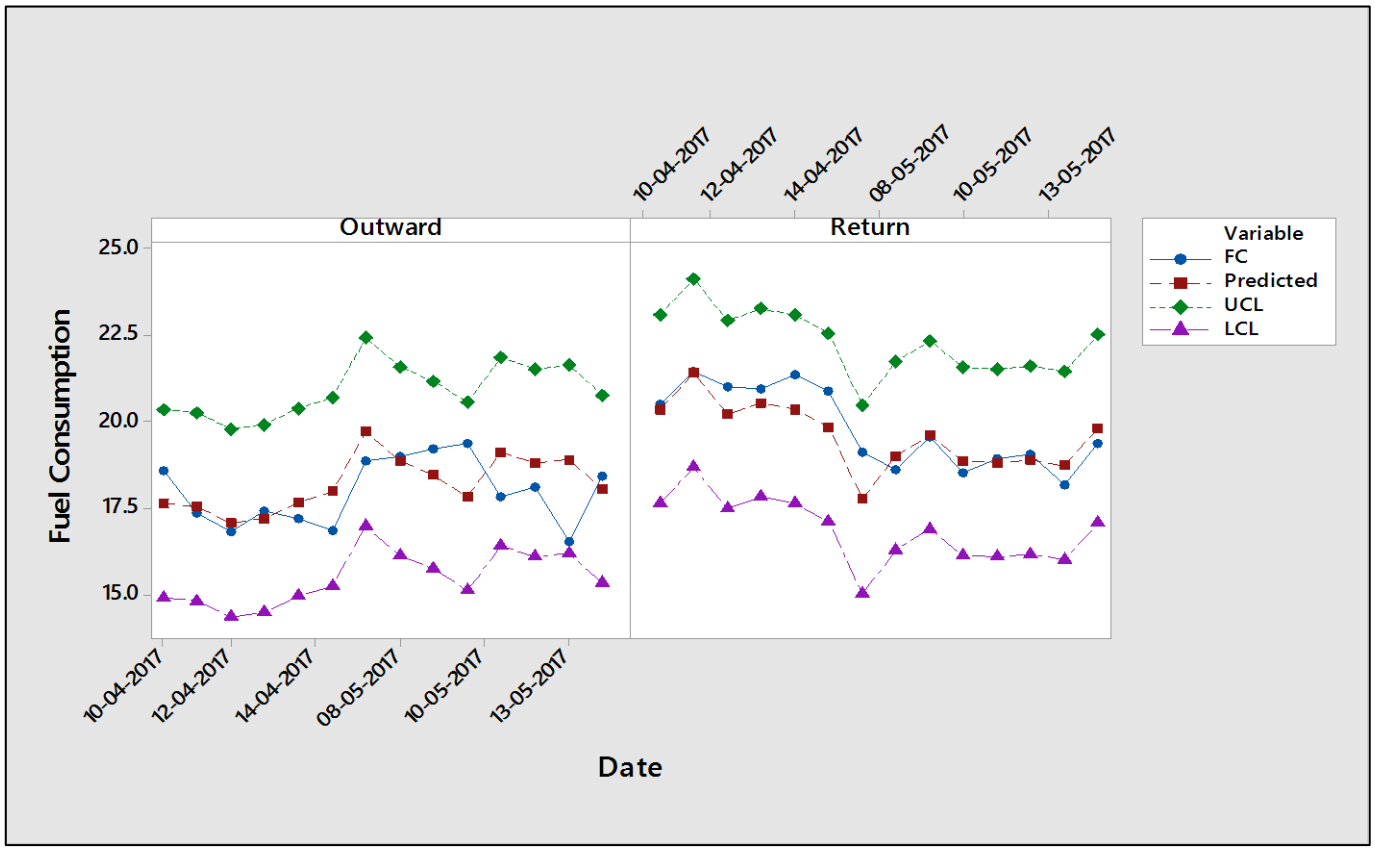


Fig 2 SPC Chart for Fuel Consumption

Figure 2 SPC chart for fuel consumption. UCL and LCL are upper and lower control limits. Lines for Predicted and actual fuel consumption (FC) are shown (1)

Source:

Figure 2 illustrates a statistical process control chart constructed using residual fuel consumption after fitting a regression model. It is evident that during the time shown in the control chart, the outward journey generally requires less fuel than the return journey due to the tide. Journeys within the upper and lower control limits for specific weather and tide conditions are considered satisfactory. Journeys exceeding either control limit merit further investigation into the causes of fuel consumption. In this case, all journeys remained within the control limits. Using control limits prevents unnecessary checks from being made when there is no real change in the vessel's performance.

Data Science allows the shipping company to understand the performance of its ships and the variation in fuel consumption costs and corresponding emissions due to weather and tides. Where flexibility exists, journeys can be scheduled to minimise expenses. This way, the analysis provides a valuable management decision-making tool based on collected and available data.

Another application of data science in the SME estimates the most economical speed at which the vessel can travel. Ships undergo dock trials during construction, builder trials when completed, and sea trials when they begin service. These trials indicate the economical speed for travel. However, a ship's performance changes over time due to fouling, minor collisions, etc. The most economical speed can be reassessed by applying statistical thinking to design trials during the ship's service, leading to improved business outcomes.

This example demonstrates the added value of fuel consumption data for monitoring, developing new products, and widening the SME's customer base. The work involves detailed knowledge of the shipping scenario, handling vast sets of fast-moving data, statistical analysis, and business acumen to convert findings into a valuable product.

- *Social Housing Sector*

Social housing, accounting for about 50% of rental properties in the UK, aims to provide affordable accommodation and is governed by strict governmental

regulations. Certain social housing providers have tens of thousands of properties, but rent collection, repairs, and maintenance are usually outsourced to bespoke software providers. The business motivation of this case study is to identify tenants at risk of falling into arrears before their debts become too challenging to manage. These tenants can then be assisted in making their payments, potentially reducing the arrears caseloads and account processing time, thus increasing income collection.

In this example, an SME software house uses Data Science to maximise insights from a vast reservoir of data on rent balances, property repairs, empty properties, or voids (26). Social housing data is complex, containing information about the property, the tenant, and the payment details during each tenancy. Confidential data must be handled with strict adherence to data protection laws. Extensive data amalgamation from multiple tables, including property characteristics, tenancy information, and payment interactions, is required for analysis. Weekly arrears data are analysed using time series and machine learning methods, including cluster analysis. Figure 3 shows the main pattern of rent balance profiles over different periods. When rent has been paid, balances are negative, indicating the tenant is in credit, while a positive balance indicates a debt. Cluster analysis identifies a significant tenancy cluster and multiple exceptional clusters. In Figure 3, the left-hand side central cluster illustrates temporal patterns. The top plot shows that rent arrears follow a downward trend, the middle plot shows an increase throughout the month, and the lower plot shows a peak in winter. The plots on the right-hand side of Figure 3 display typical outliers, with an overall upward trend in the top plot, no clear monthly pattern in the middle plot, and a more variable yearly pattern in the lower plot.

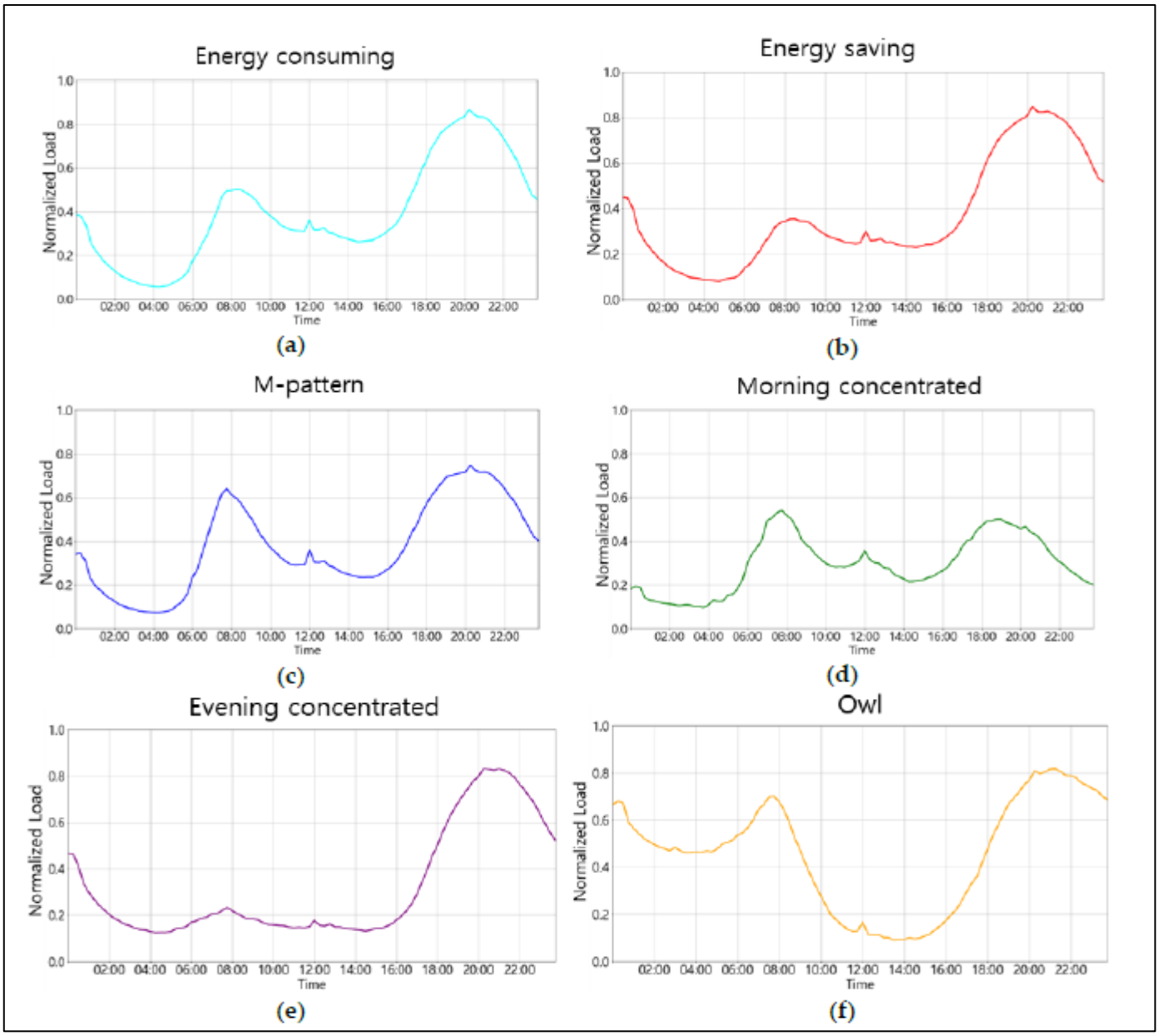


Figure 3 Results of Residential Customer Clustering
Source: https://www.mdpi.com/electronics/electronics-10-00290/article_deploy/html/images/electronics-10-00290-g005-550.jpg

Statistical and machine learning analysis have been used to predict arrears, model weekly, monthly, and yearly rent balance patterns, and identify tenants' clusters while understanding each cluster's essential features. Additionally, data visualisation has proven a powerful tool for providing social housing providers insights into their business. This example illustrates Data Science applied in a sector unfamiliar with data analytics, revealing enormous benefits from using vast, routinely collected operational data sensitively.

## VII. CONCLUSION AND FUTURE VIEW

A representative episode from a 2004 New York Times article (Hays, 2004) is discussed by Foster Provost F. et al. (28). The article describes how Wal-Mart's executives, as Hurricane Frances threatened Florida, saw an opportunity to utilise predictive technology. A week before the storm landed, Linda M. Dillman, Wal-Mart's chief information officer, urged her team to use data from Hurricane Charley, which had struck earlier, to create forecasts. Supported by extensive shopper history stored in Wal-Mart's data warehouse, she believed the company could begin predicting outcomes rather than waiting for them to happen.

The authors raise questions about the utility of data-driven predictions in this context: is it for anticipating the need for increased water bottles? Or identifying a specific product that sold out due to the hurricane? Or learning from previous hurricanes to identify particular needs? While these questions can be addressed quickly, the focus should shift towards more sophisticated, targeted models that can analyse the vast amount of data Wal-Mart has stored before events like Hurricane Florence. The New York Times article continues by revealing the findings of the data analysis:

The experts mined the data and identified products that would be in demand—beyond the typical flashlights. "We didn't know in the past that strawberry Pop-Tarts increase in sales, like seven times their normal sales rate, ahead of a hurricane," said Ms. Dillman in a recent interview. "And the pre-hurricane top-selling item was beer."

This anecdote illustrates the strength of Data Science: Data-Driven Decision Making, where decisions are based on rigorous data analysis rather than mere intuition. It has been shown that data dependency increases company productivity by at least 6%. Businesses require in-depth analysis and detailed insights into the current situation and potential changes.

However, as businesses transition to more data-driven models, they face a shortage of qualified personnel, i.e., Data Scientists. There are not enough Data Scientists to meet the rapidly changing demands of the business world. This has led universities to develop academic and post-academic programs with curricula designed to address this gap. This trend reflects the promising job prospects for Data Scientists and emphasises the importance of data analysis skills for aspiring managers.

To ensure good practice, specialists in various areas of Data Science must collaborate and embrace diverse thinking styles. The importance of a business focus, emphasised by Six Sigma, is even more crucial in Data Science. Solutions must address strategic business issues to avoid reducing Data Science to an academic exercise, which could limit the potential for innovative applications. Data Scientists must be creative, business-centric, and always mindful of the advantages of collaborating with specialists from different fields. Effective communication and collaboration are essential for good Data Science.

Statisticians must be confident in the value and significance of their work. Companies require guidance in navigating the new data-centric world, and presenting case studies demonstrating the benefits of accurate analysis—over simple "black box" solutions—can help. Reports on successful and less successful Data Science projects should be readily available in academic publications and the trade press, where business professionals are most familiar.

This need is being met through a growing body of case studies and opinions in popular media, such as the article by Shan (29) and the paper by Giudici (10). While textbooks are slower to incorporate Data Science, some excellent publications, such as James et al. (30) and the comprehensive Data Science Handbook (31), are leading the way. We are optimistic about the continued growth of statistical involvement in Data Science and the increasing importance of Data Science as a key operational strategy for companies. The substantial financial benefits ensure ample funding for research, and new ideas and methods continue to emerge.

## REFERENCES


[1]. Vicario, Grazia & Coleman, Shirley. (2019). A review of data science in business and industry and a future view. Applied Stochastic Models in Business and Industry. DOI: http://dx.doi.org/10.1002/asmb.2488

[2]. Coleman, SY. Data science in Industry 4.0. [Book auth.] ECMI. ECMI conference, Budapest June 2018, in ECMI book subseries of Mathematics in Industry. s.l.: Springer, 2019.

[3]. Ahlemeyer-Stubbe, A and Coleman, SY. Monetising data – how to uplift your business, Wiley. London: Wiley, 2018.

[4]. Naur, Peter. "Concise Survey of Computer Methods". Lund, Sweden: Studentlitteratur. 1974. Retrieved from: http://www.naur.com/Conc.Surv.html

[5]. Joseph, Hugh A. Chipman and V. Roshan. A Conversation with Jeff Wu. Statistical Science. 2016, Vol. 31, 4, pp. 624–636.

[6]. Wu, C.F.J. “Statistics = Data Science?” 1997.
[7]. Flach, P. Machine Learning: The Art and Science of Algorithms that Make Sense of Data. s.l.: Cambridge University Press, 2012.
[8]. Breiman, L. (2001) Statistical Modelling: The Two Cultures. Statistical Science. 16(3), 199-231.
[9]. Box, G. E. P. (1976). Science and Statistics. Journal of the American Statistical Association, 71: 791–799. DOI:10.1080/01621459.1976.10480949.
[10]. Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. International Statistical Review / Revue Internationale de Statistique, 21–26.
[11]. Giudici, P. Financial data science. Statistics & Probability Letters. May 2018, Vol. 36, pp. 160-164.
[12]. Wired. , Chris Andersen (2008) The end of Theory: the Data Deluge makes the Scientific method Obsolete.
[13]. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. From Data Mining to Knowledge Discovery in Databases. Al Magazine, 17(3). 1996. Retrieved from: https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131.
[14]. Brachman, R., and Anand, T. 1996. The Process of Knowledge Discovery in Databases: A Human-Centered Approach. In Advances in Knowledge Discovery and Data Mining, 37–58, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Menlo Park, Calif.:.
[15]. The Great Expectations of the ImageNet Challenge. Soft., Science. s.l.: https://www.scnsoft.com/blog/imagenet-challenge-2017-expectations, 2017.
[16]. Efron, B and Hastie, T. Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. s.l.: Cambridge University Press, 2016.
[17]. Patil, d and Davenport, TH. Getting control of Big Data. Harvard Business Review. 2012.
[18]. Coleman, SY and Kenett, RS (2017) The Information Quality Framework for Evaluating Data Science Programs. Available at SSRN: https://ssrn.com/abstract=2911557.
[19]. Brown, J. (2018). Leading change: Developing a new Applied Data Science programme. ICOTS10, Kyoto, http://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_1H1.pdf?1531364187.
[20]. kdnuggets discussion. [Online] 2013. https://www.kdnuggets.com/2013/04/data-science-end-statistics-discussion.html.
[21]. Coleman, SY, Gob, R, Manco, G, Pievatolo, A, Tort-Martorell, X, Reis, M (2016) How Can SMEs Benefit from Big Data? Challenges and a Path Forward. Journal of Quality and Reliability Engineering Int., http://onlinelibrary.wiley.com/doi/10.1002/qre.2008/full.
[22]. Coleman, S.Y. Six Sigma – an opportunity for statistics and for statisticians. 2008, Vol. 5, pp. 94-96.
[23]. Mustafazade, F. (2018). Using social science data to solve a social housing problem. https://blog.esrc.ac.uk/2018/10/19/using-social-science-data-to-solve-a-social-housing-problem/.
[24]. Smith, W, Coleman, S, Bacardit, J, Coxon, S. Insight from Data Analytics with an Automotive Aftermarket SME. Quality and Reliability Engineering International. 2019, pp. 1-12.
[25]. Smith, W, Coleman, S, Bacardit, J. and Coxon, S. (2018) How data can change the automotive aftermarket. Focus, p30-32, October, www.ciltuk.org.uk.
[26]. Zaman, I, Pazouki, K, Norman, R, Younessi, S, Coleman, SY. Development of automatic mode detection system by implementing the statistical analysis of ship data to monitor the performance. Int. J. Maritime Engineering, RINA Trans A3. 2017, Vol. 159, pp. 225-35.
[27]. Mustafazade, F, Coleman S, Bacardit, J (2018). Application of machine learning for decision support in social housing. Statistics and Data Science - new Developments for Business and Industrial Applications conference, Turin, http://www.sds2018.polito.it.
[28]. Provost, F., Fawcett, T. Data Science for Business - What you need to know about Data Mining and Data-Analytic Thinking. s.l.: O’Reilly Media, USA. , 2013.
[29]. Shan, Carl. What-are-good-examples-of-using-data-science-for-development-and-or-social-good. quora.com. [Online] 2014. [Cited: 2 August 2019.] https://www.quora.com/What-are-good-examples-of-using-data-science-for-development-and-or-social-good.
[30]. James, G, et al. The Introduction to Statistical Learning with Applications in R. s.l. : Springer, 2017.
[31]. Cady, Field. The Data Science Handbook, Wiley. s.l. : Wiley, 2017. ISBN: 978-1-119-09294-0.
[32]. G, V., & S, C. A Review of Data Science in Business and Industry and a Future View. https://core.ac.uk/download/429246501.pdf
[33]. Grazia, V., & Shirley, C. (2020). A review of data science in business and industry and a future view. https://doi.org/10.1002/asmb.2488