TOWARDS TOTAL DATA QUALITY MANAGEMENT THAT ENABLES FEASIBLE STATISTICAL MODELLING:
A CASE STUDY IN THE FINANCIAL SERVICES SECTOR

S. Strydom[1]*, I. Doyer[2] and C. Behr[3]

1 University of Pretoria, South Africa | ss.shannon.strydom@gmail.com

2 University of Pretoria, South Africa | Ilse.doyer@up.ac.za

3 University of Pretoria, South Africa | behr.carina@gmail.com

ABSTRACT

Case studies play a valuable role when researchers want to understand why or how things work in real-life settings, which are often subject to complexity not always accounted for in theoretical work. This study contributes to the literature on data quality management by examining how data deficiencies can fundamentally alter the trajectory of applied research and business projects. Instead of advancing directly to statistical modelling, the project required a structured program of data profiling, quality assessment, and change management. Findings revealed that only 19.2% of the provided data set planned to be used for statistical modelling met the data quality criteria of being unique, up-to-date, accurate, consistent, complete, and valid. Reflective journalling documented the evolution and learnings of the data project and emphasised that improving data quality demands more than technical fixes—it requires iterative, collaborative processes that address issues at their source. Equally important is fostering a culture of problem-solving, supported by tools like checklists and *Poka Yoke*, to prevent recurring errors. By documenting these challenges and responses, the study offers a replicable framework for organisations and underscores the critical role of data quality in enabling reliable analytics.

Keywords: TDQM, data quality, financial services, data cleaning

* Corresponding author

1 INTRODUCTION

In the financial services sector, the quality of data is crucial for operational efficiency, informed decision-making, and overall business success [1, 2]. With the increasing dependence on data-driven strategies, maintaining high standards of data quality has become imperative for organisations. The company of this case study, referred to as Company A, operates in the short term vehicle insurance sector in a low-middle-income country. In this industry, accurate and timely data processing—from customer policies to claims and financial transactions—is vital. This reliance highlights the necessity for robust data quality management practices.

Risk models play a crucial role in the financial sector, particularly within the insurance industry, by enabling companies to quantify and manage potential future financial losses [3]. The primary challenge in this domain is pricing premiums accurately to reflect the extent of future financial loss before it occurs. This challenge has driven the need for sophisticated statistical models that can quantify risk on a more granular level [4]. For instance, Company A aimed to redefine their vehicle insurance premium calculation framework using generalised linear modelling techniques. The initial goal was to develop models that estimate risk in terms of claim frequency and claim severity, thereby allowing for more risk-appropriate premium calculations. Such models are essential as they help in decreasing high loss ratios and improving the overall financial health of the company. However, the efficacy of these models is highly dependent on the quality of the underlying data. Discrepancies in the dataset and a lack of awareness about data quality issues can severely undermine the reliability of the resulting risk models. Consequently, addressing data quality is a critical step before any effective risk modelling can be performed.

Despite the evident importance of data quality, many companies still grapple with numerous data quality challenges. Especially where data is not automatically collected, and there is a lack of understanding and awareness of the importance of data quality, data sets can lead to missing, inconsistent and incorrect data entries. This phenomenon is demonstrated by the case study presented in this paper, which describes a project that was intended to focus on statistical modelling and data comparison, but had to take an

important detour through a comprehensive data quality investigation, data profiling, change management, and the formulation of data quality maintenance steps. This paper provides a description on the state of the real-world data, as well as the results of a systematic data cleaning process.

As to their theoretical contribution, case studies play a valuable role when researchers want to understand why or how things work in real-life settings, which are often subject to complex relationships [5]. By studying various real-life cases, researchers can extract theory on why or how certain inputs succeed or fail to produce the desired outcomes [6]. This case study contributes to the literature on data quality management by illustrating how data quality challenges can fundamentally redirect the trajectory of applied research and business projects. Instead of proceeding directly to statistical modelling and data comparison, the study highlights the necessity of first undertaking a structured process of data profiling, quality assessment, and change management. By documenting both the deficiencies in real-world data and the systematic steps taken to improve data quality, the study demonstrates the practical value of integrating data quality investigations into project workflows.

The contribution lies in showing how comprehensive data quality maintenance strategies can transform flawed datasets into reliable inputs for advanced analytics, while also offering a replicable framework for organisations facing similar challenges in contexts where data is manually captured or poorly governed. Furthermore, it provides an honest case study of the state of data in some settings that are highly reliant on the data to make decisions that will affect business profitability directly, answering the research question of what real-world data may look like. By providing an in-depth look into one case where data is still manually captured, this paper provides the opportunity for similar case studies in other settings, which can be used to create a more realistic awareness regarding the state of real-world data. Knowing what data quality levels to expect and what issues to build into system designs, further paves the way for the use of Artificial Intelligence to expedite the preparation of data to avoid the phenomenon of 'garbage in, garbage out' where there is no human-in-the-loop.

The paper is structured as follows: section 2 provides some insights from literature on data quality and its management; section 3 describes

the process followed for data quality assurance, followed by the results of the endeavor in section 4. Section 5 contains the recommended data quality assurance framework, with further research opportunities and the limitations of this study summarised in section 6.

2 LITERATURE REVIEW

2.1 The evolution of Total Data Quality Management

The concept of Total Quality Management (TQM) has its roots in the convergence of management theories and quality principles, drawing heavily on the foundational work of Deming's 14 points and Juran's quality trilogy. Over time, these ideas evolved into a holistic management philosophy aimed at embedding quality into every aspect of organisational practice [7]. TQM can be understood as a comprehensive system of practices, methodologies, and training initiatives designed to help organisations navigate dynamic environments while maintaining a clear focus on achieving and sustaining customer satisfaction [8]. TQM is thus a holistic continuous improvement approach focusing on building quality into the process to achieve customer satisfaction. It does this through employee involvement, teamwork – also with suppliers, and a long-term perspective [9]. In recent decades, the relationship between TQM and knowledge management has increasingly been strengthened leading to the adoption of many TQM principles into the sphere of knowledge and data management [10].

TQM was adapted to an information processing environment to form the Total Data Quality Management TDQM) framework [7, 11]. TDQM emphasises the accountability of all organisational members in maintaining high standards of data quality, thereby aligning data processing activities with customer expectations and operational goals. An advantage of treating an information processing system and a traditional manufacturing system in the same way is that many of the quality techniques used in TQM can be translated to suit the data environment.

Where TQM is a management approach where everyone in an organisation is accountable for maintaining high standards to produce quality products that meet customer expectations [7], TDQM views information

manufacturing as a process that converts raw data into quality information products. In product manufacturing, a processing system converts raw materials into physical products. Similarly, information manufacturing can be viewed as a processing system which converts raw data into information products. TQM was consequently adapted to an information manufacturing environment to form the TDQM framework [8].

An advantage of treating an information manufacturing system and a traditional manufacturing system in the same way is that many of the quality techniques used in TQM can be translated to suit the data environment [12]. There are however some limitations that arise due to the nature of data when compared with a physical product which should be considered when adopting established frameworks. Where a physical product is consumed, a data unit or information unit is not depleted. In addition, the duplication of a data unit is trivial when compared with the expenses incurred to copy a physical product [12].

2.2 Understanding data quality

According to Sebastian-Coleman [7], the level of data quality in TDQM is the degree to which the data meets the expectations of the data users based on their user tasks. Data quality is therefore subjective to the perceived quality of the data user. Quality thinkers have identified several quality aspects, called dimensions, to measure data quality [7].These dimensions quantify data quality in relation to a scale or metric so that data quality can be compared between data objects and over different time periods to measure the impact of data driven business processes [7].

The following fundamental concepts form part of the early stages of any TQM implementation ([13, 14] and were the focus of this case study: Standardisation, which manifests as clear and communicated quality standards, as well as a strict view of what constitutes output, namely only output that complies to the set standards, Creating awareness amongst all employees that the next downstream process is an important customer, and that every employee has a role to play in assuring quality throughout the process, and Establishing an understanding, firstly amongst management, then amongst all employees, of what it means to proactively build quality

into the process instead of reactively inspecting it in.

The six most relevant data quality dimensions identified for this case study have been compiled from the work of Wang [8], Shankaranarayanan and Cai [15], and Sebastian-Coleman [7] and were used to build a uniform awareness of what constitutes good data quality:

Uniqueness: evaluates the degree to which data cannot be mistaken for other entries,
Completeness: evaluates the extent to which all required data is present,
Consistency: evaluates whether data is uniform and reliable across different datasets,
Validity: evaluates whether data conforms to stated business rules,
Timeliness: assesses the relevance of data based on its age and the time it was captured,
Accuracy: evaluates the correctness of data in representing real-world values.

2.3 Building data quality into the process

Once these six data dimensions are being measured and monitored throughout the process, and ownership of data quality management has been established, team members should be educated on and empowered to look for and action opportunities to build quality into the process. A practical technique, highly applicable to the data environment, is that of error proofing. Poka Yoke or 'mistake proofing' is a concept originally coined by Shiego Shingo in the 1960s which has become applicable in many different environments [16, 17]. The practice aims to prevent defects from appearing in the first place by ensuring that mistakes are impossible. If this cannot be achieved then it can be used to detect and correct errors [18].

Poka Yoke consists of three sequential steps. The first step is to identify all possible errors, human and machine, that can occur in the process. The next step is to determine how the error can be detected when it occurs or when it is about to occur. The last step is to identify a specific action that can be taken to eliminate the error [19].

Poka Yoke can also be divided into three types of mistake-proofing categories namely, control, shutdown, and warning. Control is an action that corrects the error automatically. An example of a control would be auto-capitalisation at the start of a new paragraph in MSword. Shutdown occurs when the process is immediately halted upon the detection of an error. An example of a shutdown is a system that prevents submission of a data entry until all mandatory fields are filled in. Warnings inform people that an error has occurred. A drawback of warnings is that they are often ignored which is why controls and shutdowns are preferred [19].

Given the above description, it is evident that Poka Yoke can be used to reduce the reliance on people to operate free of error. One of the barriers to implementing Poka Yoke is the need for management support [20]. This is because people are set in their way of doing things and may resist changing, hence the importance of incorporating change management into the TDQM implementation. Another barrier of Poka Yoke is the lack of awareness of the concept [20]. This barrier can be remedied by creating awareness for the technique and getting employees involved in the process.

Checklists are a more common quality assurance tool used to ensure that all essential steps in the process are followed and to make it possible to confirm compliance in a standardised way [21]. It is however acknowledged that the implementation of checklists requires employee compliance and is time consuming [21], again highlighting the need for change management.

In a case study by Kalapurakal, et al. [22], the impact of checklists was evaluated in terms of the reduction of medical errors relating to dosage, patient identification and patient location. The case study highlights the use of checklists to improve the performance quality when human error needs to be addressed as opposed to system or machine error.

Although this paper focuses on the data cleaning and standardisation process that occurred at the beginning of the company's data quality management journey, the in-depth analysis of the state of the data automatically led to the further step of identifying opportunities for the implementation of Poka Yoke. The data analysis also made it clear to what extent the data quality improvement journey would rely on employees to assure quality, and thus change management was naturally an important part of the process from the start.

2.4 An overview of change management models

Kotter's 8-Step Change Model is an expansion of Lewin's three step change management model [9]. According to Galli [9], Kotter's model provides more direction on how to implement change and also incorporates more of the human component of change by delving deeper into preparing employees for change and supporting employees during change. The explicit steps in this model also make it easy to implement [9].

One major drawback of this model is that employees may experience a top-down approach [23]. Employees are only involved after the change vision has already been developed, which may create more resistance to change. If employees do not get on board with the newly developed strategic vision, then this may result in wasted effort. Kotter's model may be enhanced if used in conjunction with a model which is more people centric [9].

The ADKAR model consists of five sequential goals that focus on the acceptance of change. The acronym stands for Awareness, Desire, Knowledge, Ability and Reinforcement [9]. Awareness occurs when employees are informed of the need for change. Desire occurs when employees decide that they want to partake in the change. This is followed by the Knowledge of what the change demands and how to drive the change. Ability is what is required to implement the change. Lastly, Reinforcement is required to ensure that the change is sustained [24].

The ADKAR model is useful in identifying the relevant areas of resistance which are prohibiting change from happening [24]The goal-orientated structure of this model helps to facilitate change by working towards smaller, seemingly more attainable milestones as opposed to Lewin's model which focuses on the overall process. According to Galli [9], since this model focuses primarily on the human component of change, it is better suited to smaller environments and project teams as opposed to large scale organisations with complex business processes. Galli [9] postulates that in practice the ADKAR model should be paired with another model which deals with the human emotions that accompany change.

The Kubler-Ross change curve, also referred to as the grief curve or the five stages of grief, was introduced by psychiatrist Elisabeth Kubler-

Ross in 1969 as a product of her work with terminally ill patients. This model originally described how people handle grief when confronted with a terminal illness. The five stages of grief are defined as denial, anger, bargaining, depression and ultimately acceptance [25]. The change curve was subsequently translated to a business environment to describe how people experience disruptive change and how this influences their productivity levels.

The Kübler-Ross model is a people-centric model that focuses on the human experience to change and highlights the importance of considering an individual's experience to change as opposed to that of the team as a whole [25]. The framework allows team leaders to anticipate and understand their employees' reactions to change. Furthermore, team leaders are empowered to proactively support their employees so that the change is more likely to be successful. This model does however fail to provide an actionable framework for how the change should be implemented. It is therefore recommended that this model be paired with another change management model that addresses these shortcomings. However, by using this model, greater consideration can be given to how the information can be relayed so that employees feel supported during the change initiative. This model addresses the shortcomings of the ADKAR model and Kotter's 8-Step Model so that a holistic approach can be taken in creating a case for change.

Additionally, the McKinsey's 7S framework provides insight as to the seven interrelated key elements that companies should consider when implementing change in order to produce a synergistic outcome. These elements are further divided into hard and soft elements. The hard elements, namely strategy, structure and systems, can be controlled by management. Strategy is the proposed plan to transform the organisation from its current state to the desired state. Structure refers to the organisational structure such as the reporting hierarchy, roles and responsibilities and systems that are defined by the formal procedures and day-to-day activities [9, 26]. The soft elements are more often governed by intrinsic culture and include shared values, style, staff and skills. Shared values refer to the engrained organisational culture and company values. Style emphasises how the leadership style influences the resistance to change. Staff refers to the

capabilities of the employees, and skills refers to the core competencies of those employees [27].

Unlike the other change management models, the McKinsey 7S model highlights the strengths and weaknesses that exist in different dimensions of an organisation, which allows managers to identify where there is a need for change. The McKinsey model does however not explore how employees experience change or how change should be implemented. Another disadvantage of this model is that proper implementation can be time consuming and therefore not cost effective for large organisations [9].

There are many other concepts within the TDQM framework that address various aspects of the quality assurance journey and only the ones relevant to the case study presented in this paper have been discussed here. How the reviewed concepts were integrated to address real-world quality issues is presented next.

3 THE CASE STUDY: REAL-WORLD DATA QUALITY

As mentioned in the introduction of this paper, the initial objective of this project was to utilise historical data to develop a statistical model to estimate the risk associated with a given policy. The existing premium framework would be replaced with the statistical model such that more risk appropriate premiums could be charged to clients on an individual basis as opposed to the existing fixed percentage approach. The project sponsor requested that generalised linear modelling be used to develop two statistical models in order to quantify the risk in terms of claim frequency and claim severity separately. Before implementation, the impact of the models would be evaluated by comparing the actual losses to the losses incurred if the statistical models were used to determine the premium charged. In addition, the available data was to be compared to other insurance models to identify if there is additional information that is currently not recorded by Company A which would benefit future risk models.

When a new policy is activated the policy information is stored on the system in the policy dataset. If a client submits a claim, the claim is entered and stored on the system in the claims dataset. Not all policies are expected

to have corresponding claims as a client may have been claim free, but all claims are expected to correspond to a policy. These two datasets are then aggregated to calculate the loss ratio and evaluate the performance of the portfolio as well as the individual policies. The project sponsor provided the aggregated dataset containing both policy and claims information for the development of the statistical models.

After the application of generalised linear modelling was investigated in literature, the data was used to develop a frequency model with a Poisson distribution. The development of the frequency model had little success as the model's predictions were severely inaccurate and many of the variables available were not significant in predicting claim frequency. This raised the question whether the data being used for the model was of sufficient quality and contained sufficient variables to develop an accurate model. Upon inspection of the dataset, several discrepancies were identified. These discrepancies included:

- A significant number of empty data fields,
- Claims information for which there was no corresponding policy,
- Extreme vehicle values,
- Negative vehicle values,
- Negative vehicle ages, and
- Negative premiums.

Inquiries into these discrepancies were met with resistance from the individuals responsible for managing the data system, highlighting the need early on for change management and involving the team members in the data analysis to create awareness of the need for change. In addition, not only were there no relevant policyholder characteristics in the data but there were very few recorded vehicle characteristics. This was concerning as the data provided for this project was also used by Company A to calculate the loss ratio and to make business decisions. Since the quality of the data was observed to be inadequate, these internal evaluations may have motivated business decisions which were in fact not beneficial.

The reliability of any statistical model is directly related to the reliability of the data used to develop the model [28]. Due to the considerable prevalence of discrepancies in the data and the lack of available data, the

modelling component of the project became secondary and the new problem that emerged was one concerning the quality of the data. It also became apparent that an appropriate change management framework would be required to motivate the project findings to the team and encourage productive change. After consultation with the project sponsor, the scope of the project was updated to address the data quality issues and include a change management component.

After the discrepancies were discovered in the dataset, the claims data and policy data were requested as stored in their raw form on Company A's system. A preliminary data analysis was carried out on the requested data to better understand the data environment and ultimately produce a clean dataset for modelling purposes.

The first task of this project was thus a comprehensive data quality investigation, which informed the data cleaning and preparation process. Once the data had been cleaned and prepared, it could be used for the originally planned statistical modelling. This paper will however focus on describing the data quality improvement journey, which consists of three phases: data collection, data cleaning and preparation, and data quality assurance initiation.

3.1 Data collection

Given the significant number of discrepancies found in the original merged dataset provided by the project sponsor, the raw claims dataset and policy dataset were requested. The policy data and the claims data were imported into RStudio for further analysis. There were 25 961 entries in the policy dataset and 2 381 entries in the claims dataset before analysis, covering three financial years. An important principle extracted from the experience was thus that data cleaning should start from the source data. This is similar to the concept of Genshi Genbutsu in lean management where problem solvers are encouraged to 'go to the floor' – 'the scene of the crime' – to start problem solving at the source.

3.2 Data preparation and cleaning

The data cleaning and preparation was done in sequential steps as follows:

Removal of duplicates: Firstly, the duplicates in the claims dataset and the policy dataset were removed. These were entries for which the values in the columns for a specific row were exactly the same as for another row. This means that data were entered into the system twice and the duplicate line-items should not have been included in the respective datasets. No duplicates were found in the policy data set, but the claims data set contained 59 (2.48%) duplicate entries. This provided the current state of one of the quality dimensions, namely uniqueness.

Removal of test run data: In both datasets there were entries which were clearly used for testing purposes since the word 'test' appeared once or more in a single line-item. In the claims data set, three (0.13%) test run entries were removed, and 134 (0.52%) were removed in the policy data set.

Aggregating of data within the data sets: The claims data contained a unique entry for every single claim incurred. However, some vehicles incurred more than one claim for the duration of the policy. Here a renewal of a policy is considered as a new vehicle entry so that the claim count can be assessed for every vehicle every year. Since more than one vehicle could belong to a specific policy number and the policy numbers are often carried over to the next year, the policy number could not be used as a unique identifier. The policy inception date, which is the date the policy was started, and the vehicle registration were combined to create a new variable that could be used to uniquely identify a claim for a specific vehicle in a specific year. Due to the lack of standardisation in how the data was entered, a further preparation step had to be included here: the vehicle registration numbers had to all first be capitalised and any spaces in the characters had to be removed. The claims data was then aggregated so that for every vehicle and every year the claim count and the total claim severity could be seen. This resulted in the removal of 177 (7.63%) entries from the claims data set.

Similarly, the same unique key was created for the policy data set. Some of the entries in the policy dataset were created to update existing entries. For example: say a vehicle was registered to a policy with a sum insured of USD15 000 and a premium of USD750. A week after the policy inception date, it was discovered that the true value of the vehicle was actually USD20 000. An adjustment entry was then made with a sum insured value of +USD5 000. Several of these adjustments also existed for the premium variable. The updated entries were aggregated with the original entries such that the new entries reflected the total sum insured and total premium for every vehicle in every year. During this step 5308 (20.55%) of the policy entries were removed.

The waterfall charts in Figure 1 (claims data) and Figure 2 (policy data) visualise the effect of the removal of duplicates and test data, as well as the aggregation of fields relating to the same claim or policy. After this initial cleaning process, the policy and claims data sets were merged to ensure that each claim is associated with the correct policy and to further assess and correct data quality issues.

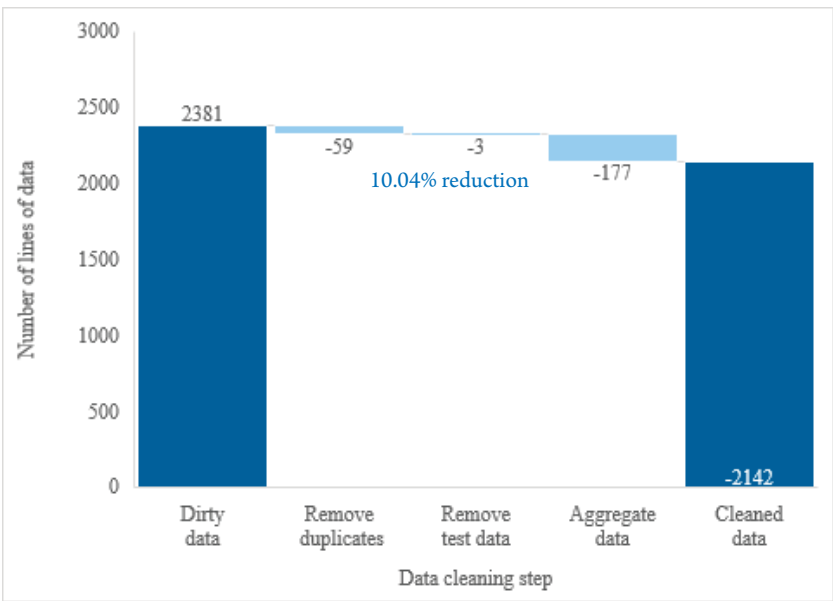


Figure 1: Claims data cleaning process results

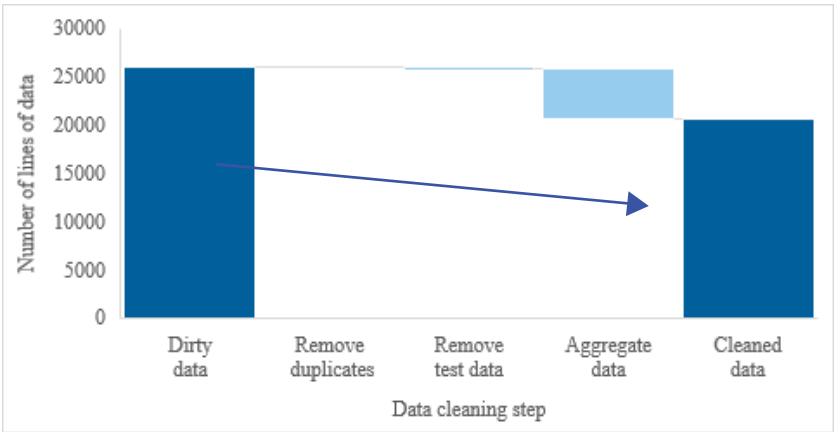


Figure 2: Policy data cleaning process results

Merging data sets: The next step was to merge the policy data and the claims data so that for every year and every vehicle the total claim severity and claim count were known. The unique key created in both datasets was used to merge the corresponding entries.

Correction of data: Upon merger, it became evident that there were some claim entries that did not correspond to a policy. Since it is impossible to make a claim without a corresponding policy, this indicated that there may be data errors contained within the policy inception date and the vehicle registration number that formed the unique key used for merger. The data thus had to be screened for the 115 unmatched claims to correct the data errors that could be identified so that these entries could be included in the aggregated dataset.

Creation of essential input data: The vehicle age of each vehicle was calculated using the vehicle manufacture date and the policy inception data. The exposure of the policy was calculated using the policy inception date and the policy expiration date whilst also accounting for policies which ran over into the next financial year.

Removal of irrelevant information: Irrelevant columns were then excluded from the dataset so that the data contained only vehicle characteristics, driver characteristics and the metrics needed for the risk model, such as claim count, claim severity and policy exposure.

Assessment of missing values: the driver characteristics, as well as the vehicle characteristics, were checked for missing values. The sum insured, vehicle make, and age data sets were complete as a result of some of the corrective actions already mentioned. More than half of the entries were missing for the engine size, fuel type and body type and no driver characteristics had been captured at all.

Exploration of available data: the structure of each variable was explored further by observation and by calculating summary statistics such as mean, median, minimum and maximum.

All of the actions mentioned up to this point resulted in the line items being reduced from 28453 raw line-items to 20238 line-items, a 28.59% reduction in the data set (Figure 3).

Up until this point, the data preparation, cleaning and correction judgement calls were relatively obvious, for example, duplicates, test run data, missing values etc. However, further data cleaning steps required frequent consultation with various stakeholders, including the management team. No data were removed without confirmation from the team that there was indeed no other option to deal with the discrepancy. The data cleaning steps are described next.

Detail data cleaning: During this step, the data was screened for the following:

- Outliers in the different variables such as sum insured and vehicle age,
- Missing values,
- Inconsistent entries,
- Spelling errors,
- Ambiguous entries,
- Incorrect entries,
- Inconsistencies, and
- Sparsely populated rows or columns.

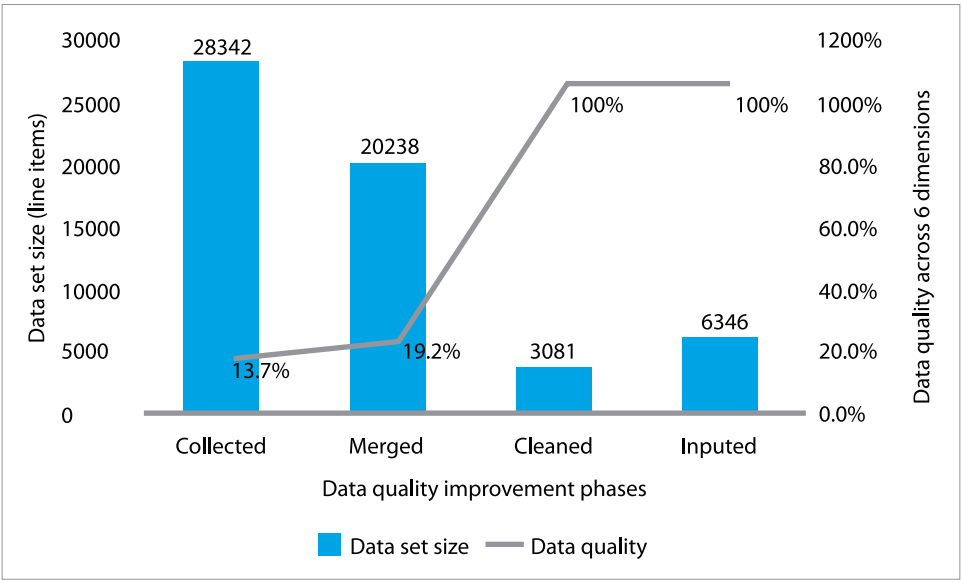
As mentioned, the relevant stakeholders were consulted throughout this process so that no corrections or removals were made without careful consideration. This cleaning process revealed the true state of the data and

reduced the data set to only 3 881 line-items, 19.18% of the merged data sets, that were full and reliable enough to use in statistical modelling.

Data imputation: Using fields that had been populated allowed for other fields, such as vehicle models, to be deduced from the available data. Through this data imputation process an additional 2465 entries could be included in the modelling dataset. However, the data in some columns were so diversely different in terms of the type and level of data captured that even a pattern extractor was unable to salvage the data.

Data set finalisation: Finally, the final state of the clean data was presented to the management team, who then decided on the set of variables deemed fit for statistical analysis.

The data preparation and cleaning phase thus resulted in only 19.18% of the merged data set being usable for statistical analysis and modelling. The data imputation step increased this to 28.00%. Figure 3 summarises the results of the data cleaning.



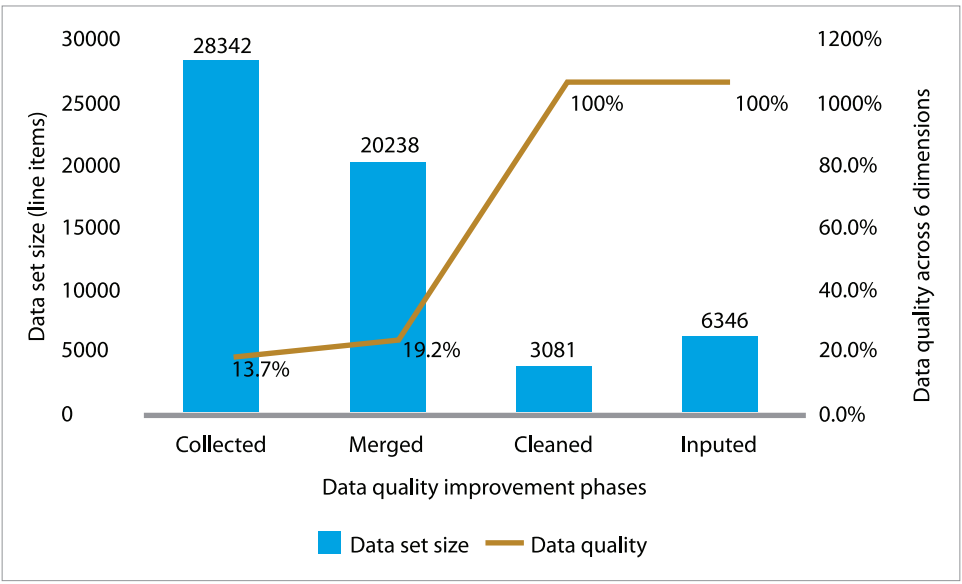


Figure 3: Results of data cleaning phases

To complete the current state investigation, the data analysis was summarised into the six data quality dimensions as follows:

- Uniqueness: The percentage of duplicates in the dataset was used as the metric for this dimension. Ideally, the percentage of duplicates should be 0%. The analysis revealed that the percentage of duplicate entries was 2.48% for the claims dataset and 0.52% for the policy dataset;
- Timeliness: The percentage of out-of-date entries was used as the metric for this dimension. Ideally, there should be no outdated data entries. However, the investigation found that 20.53% of the policy data was out-of-date, highlighting a significant issue in maintaining current data;
- Validity: The percentage of invalid data was used as the metric for this dimension, with an ideal value of 0%. The results showed that 1.72% of the data was invalid, indicating the presence of data entries that did not meet the required standards;
- Consistency: The actual-to-ideal ratio was used as the metric, with an ideal ratio of 1:1. The investigation revealed inconsistencies in various attributes: the vehicle registration had a 9:1 ratio, vehicle make had a 1.44:1 ratio, vehicle body type had a 1.43:1 ratio, and vehicle model had a 28.55:1 ratio;
- Completeness: The percentage of non-null data entries was used as the metric, with an ideal value of 100%. The investigation found that only 53.85% of the data was non-null, indicating substantial gaps in the data that needed to be filled to enhance its overall quality; and,
- Accuracy: The percentage of incorrect data was used as a metric, with an ideal value of 0%. The analysis showed that 5.37% of vehicle registration data and 8.24% of overall data were incorrect.

3.3 Data quality assurance initiation

The in-depth investigation into, and preparation and cleaning of, the data set raised awareness of the data quality issues amongst the stakeholders that had been involved during the process. However, improving the data quality in both the short and long term would require a holistic approach involving all employees.

Kotter's eight steps, infused with the learnings from the ADKAR and Kubler-Ross models, were used to guide the change management process to engage the team throughout the process. The goal of the first step of the change management initiative was to convince the data team of the urgency to implement change by highlighting the severity and impact of the data quality issues. A collaborative approach ensured that the team could raise questions and concerns regarding the findings and/or process and make suggestions on how quality could be built into the process going forward.

Change leaders who could be entrusted to inspire the desire to change and lead the change initiative were identified, after which the focus shifted to creating a shared vision on what constitutes data quality. For this the six data dimensions and their customised metrics were introduced to monitor data quality in future. To empower teams to attain the vision, the concept of proactively building quality into the process was introduced.

Poka Yoke solutions were developed to alleviate the cognitive load of the data team as well as to build quality into the process throughout. The following control and shutdown error-proofing opportunities existed:

- Auto-formatting fields to capitalise, exclude or reject spaces;
- Making filling of fields essential for analysis mandatory;
- Preventing users from submitting data which do not comply with the defined business rules, are invalid, duplicates or missing;
- Implementing drop-down lists to ensure the entry of only valid, correctly spelled data;
- Implementing specified ranges for continuous variables to immediately flag unrealistic and incorrect values;
- Formatting fields to only accept a certain character type or number of characters;
- Cross checking between categorical variables such as driver age and driver experience level to ensure validity;
- Auto-filling fields such as vehicle make and model used to prefill the body type; and,
- Creating checklists of required information to be collected during telephonic conversations with potential policy holders.

4 CLEANING REAL-WORLD DATA SETS: KEY STEPS AND PRINCIPLES

The preliminary data analysis and the data quality investigation successfully quantified the extent of the data quality issues and provided a strong case for change, and the initiation of a journey towards zero-defect when it comes to data quality. The process, as discovered through the experience of cleaning a real-world and very messy data set, and key learnings are discussed next.

4.1 Case study process flow

The steps followed during the data cleaning process were described in detail in section 3.2. The process can be summarised more generically as follows: the process begins with data gathering, followed by data preparation and cleaning, which includes removing duplicates, correcting errors, merging datasets, addressing missing values, and finalising inputs. Where needed, imputation and detailed cleaning ensure data completeness and reliability. The final step is building quality into the process, including embedding continuous checks and improvements for long-term accuracy. Figure 4 shows a comprehensive version of the process flow.

Figure 4: Data quality improvement process flow

4.2 Key learnings

Apart from working through a messy set of real-world data systematically, some key learnings can be extracted and are summarised here.

1. Start at the Source: Where possible, start by analysing the raw data instead of a derivative of the data set where errors might already be concealed or aggregated;
2. Assume the Worst: Adopting a cautious mindset helps ensure that all potential flaws are considered. Data should not be trusted at face value until it has been validated through structured checks;
3. Apply Validity and Logic Checks from Multiple Angles: Comprehensive reviews, for example, using statistical tools, logic tests, and business rules, help uncover inconsistencies. Techniques such as R programming and pattern recognition enhance this process. Typical checks include:

    - Duplicates
    - Test-run or trial data
    - Outdated or obsolete records
    - Outliers
    - Invalid entries (for example, negative vehicle ages)

4. Consider Data Imputation: Where feasible, imputation can supplement existing valid data and prevent unnecessary data loss;
5. Embrace Continuous Improvement: Data quality is not a one-off task. The process must be iterative, with regular monitoring and adaptation to meet evolving business needs [29];
6. Build Awareness Early: Communicating findings as they emerge ensures transparency and helps establish momentum for change. Early engagement builds awareness of the need for cleaner data and drives behavioural change through early adopters as well as the Hawthorne Effect;
7. Foster Collaboration and Problem Solving: Since most data errors originate from human activity, the cleaning process must be

collaborative. Team members should be encouraged to identify and flag discrepancies without fear of blame. A constructive, problem-solving culture is essential;

8. Avoid Blame, Promote Accountability: The guiding coalition should lead by example—rewarding problem reporting, linking reported issues to visible problem solving, and engaging data owners in developing solutions; and,
9. Empower Employees through Practical Tools: Introducing employees to error-proofing methods such as Poka Yoke and checklists provides simple, effective mechanisms to prevent recurring errors. These tools give staff a 'way out' of persistent issues, reducing resistance and building ownership of the process.

5 CONCLUSION

The data quality investigation emerged as the pivotal focus of this project, highlighting the critical importance of data integrity for subsequent statistical modelling efforts. Significant discrepancies within the data necessitated a comprehensive approach to identify and rectify these issues.

Utilising the TDQM framework, the six key dimensions of data quality were evaluated: uniqueness, completeness, consistency, validity, timeliness, and accuracy. This rigorous assessment revealed substantial deficiencies, prompting an expanded project scope to include targeted data quality recommendations and a change management initiative. Quantitative metrics provided clear insights into the extent of the problems. Only 19.2% of the original data set met the six data quality dimension criteria.

The project's findings indicated that data cleaning and preparation were crucial steps. The process involved removing duplicates, handling missing values, and correcting inconsistencies in order to ensure that the data met the defined quality metrics. For instance, the uniqueness dimension revealed a 2.48% duplication rate in the claims dataset, necessitating further cleaning to meet the ideal standard of 0%. Timeliness was another critical dimension, with 20.53% of the policy data found to be out-of-date, indicating the need for more timely data updates. A high incompleteness of

data, with more than 46% of data points being incomplete, meant that a risk model was infeasible.

The data cleaning project demonstrated that improving data quality requires more than technical fixes—it demands a structured, collaborative, and iterative approach. Key principles include addressing issues at the source, applying rigorous validity checks, and considering methods such as imputation to preserve useful information. Continuous improvement is vital, with findings communicated early and often to build awareness. Most importantly, success depends on fostering a culture of problem solving rather than blame, empowering employees with practical tools like checklists and Poka Yoke to prevent recurring errors. Together, these practices form the foundation for sustainable progress toward zero-defect data quality.

The formation of a dedicated data quality improvement team was a direct outcome of this initiative, highlighting the project's success in fostering organisational change. Developed solutions focused on sustaining the data quality and were categorised into mistake-proofing actions, emphasizing prevention, detection, and correction. The iterative nature of data quality improvement was emphasised, advocating for continuous re-evaluation and adaptation to meet evolving business needs.

In conclusion, the project has laid a robust foundation for ongoing data quality improvement within Company A. By addressing the significant data quality issues, the project has ensured that future statistical models developed using this data will be more reliable and accurate. The methodologies and frameworks employed are generalisable to other researchers and organisations within the financial services sector, providing a structured approach to tackling similar data quality challenges. This comprehensive focus on data quality and change management has set the stage for more dependable data-driven decision making and enhanced business outcomes for Company A.

6 REFERENCES

[1] M. Al-Okaily and A. Al-Okaily, "Financial data modeling: an analysis of factors influencing big data analytics-driven financial decision quality," *Journal of Modelling in Management,* vol. 20, no. 2, pp. 301-321, 2025.

[2] Z. B. Yusof, "The Role of High-Quality Data in Risk Assessment: Strategies for Ensuring Accuracy, Completeness, and Timeliness in Financial Predictive Analytics," *International Journal of Advanced Computational Methodologies and Emerging Technologies,* vol. 15, no. 2, pp. 8-16, 2025.

[3] A. Magri, A. Farrugia, F. Valletta, and S. Grima, "An analysis of the risk factors determining motor insurance premium in a Small Island State: The case of Malta," 2019.

[4] M. David, "Auto insurance premium calculation using generalized linear models," *Procedia Economics and Finance,* vol. 20, pp. 147-156, 2015.

[5] R. K. Yin, *Case study research: Design and methods*. sage, 2009.

[6] G. R. Baker, "The contribution of case study research to knowledge of how to improve quality of care," *BMJ quality & safety,* vol. 20, no. Suppl 1, pp. i30-i35, 2011.

[7] L. Sebastian-Coleman, *Measuring data quality for ongoing improvement: a data quality assessment framework*. Newnes, 2012.

[8] R. Y. Wang, "A product perspective on total data quality management," *Communications of the ACM,* vol. 41, no. 2, pp. 58-65, 1998.

[9] B. J. Galli, "Change management models: A comparative analysis and concerns," *IEEE engineering management review,* vol. 46, no. 3, pp. 124-132, 2018.

[10] D. Marchiori and L. Mendes, "Knowledge management and total quality management: foundations, intellectual structures, insights regarding evolution of the literature," *Total Quality Management & Business Excellence,* vol. 31, no. 9-10, pp. 1135-1169, 2020.

[11] B. Krol, "Towards a data quality management framework for digital soil mapping with limited data," in *Digital soil mapping with limited data*: Springer, 2008, pp. 137-149.

[12] D. Ballou, R. Wang, H. Pazer, and G. K. Tayi, "Modeling information manufacturing systems to determine information product quality," *Management Science,* vol. 44, no. 4, pp. 462-484, 1998.

[13] D. Kiran, "Total quality management: An overview," *Total Quality Management,* pp. 1-14, 2017.

[14] I. Kobayashi, *20 keys to workplace improvement*, Rev. and expanded. ed. Portland, Ore: Productivity Press (in English), 1995.

[15] G. Shankaranarayanan and Y. Cai, "Supporting data quality management in decision-making," *Decision support systems,* vol. 42, no. 1, pp. 302-317, 2006.

[16] P. M. Senge, "The fifth discipline," *Measuring business excellence,* vol. 1, no. 3, pp. 46-51, 1997.

[17] A. Zhang, "Quality improvement through Poka–Yoke: from engineering design to information system design," *International Journal of Six Sigma and Competitive Advantage,* vol. 8, no. 2, pp. 147-159, 2014.

[18] S. Shingo, *Zero quality control: Source inspection and the poka-yoke system*. Routledge, 2021.

[19] A. Puvanasvaran, N. Jamibollah, and N. Norazlin, "Integration of poka yoke into process failure mode and effect analysis: A case study," *American Journal of Applied Sciences,* vol. 11, no. 8, p. 1332, 2014.

[20] M. Lazarevic, J. Mandic, N. Sremcev, D. Vukelic, and M. Debevec, "A systematic literature review of Poka-Yoke and novel approach to theoretical aspects," *Strojniski Vestnik/Journal of Mechanical Engineering*, vol. 65, no. 7-8, pp. 454-467, 2019.

[21] P. J. G. Seoane, "Use and limitations of checklists. Other strategies for audits and inspections," *The Quality Assurance Journal: The Quality Assurance Journal for Pharmaceutical, Health and Environmental Professionals,* vol. 5, no. 3, pp. 133-136, 2001.

[22] J. A. Kalapurakal *et al.*, "A comprehensive quality assurance program for personnel and procedures in radiation oncology: value of voluntary error reporting and checklists," *International Journal of Radiation Oncology* Biology* Physics,* vol. 86, no. 2, pp. 241-248, 2013.

[23] S. H. Appelbaum, S. Habashy, J. L. Malo, and H. Shafiq, "Back to the future: revisiting Kotter's 1996 change model," *Journal of Management development,* vol. 31, no. 8, pp. 764-782, 2012.

[24] J. Hiatt, *ADKAR: a model for change in business, government, and our community*. Prosci, 2006.

[25] C. A. Corr, "Elisabeth Kübler-Ross and the "five stages" model in a sampling of recent American textbooks," *OMEGA-Journal of Death and Dying,* vol. 82, no. 2, pp. 294-322, 2020.

[26] D. F. Channon and A. A. Caldart, "McKinsey 7S model," *Wiley encyclopedia of management,* pp. 1-1, 2015.

[27] T. Peters and R. Waterman Jr, "McKinsey 7-S model," *Leadership Excellence,* vol. 28, no. 10, p. 2011, 2011.

[28] M. Goldburd, A. Khare, D. Tevet, and D. Guller, "Generalized linear models for insurance rating," *Casualty Actuarial Society, CAS Monographs Series,* vol. 5, p. 77, 2016.

[29] A. Ibrahim, M. Ibrahim, and N. S. M. Satar, "Factors influencing master data quality: A systematic review," *International Journal of Advanced Computer Science and Applications,* vol. 12, no. 2, 2021.