Journal of Scientific & Industrial Research
Vol. 84, July 2025, pp. 770-774
DOI: 10.56042/jsir.v84i7.16992

Beyond the Iceberg: Addressing Hidden Fare Inflation in Titanic Data

Swee Chuan Tan
Singapore University of Social Sciences, 463 Clementi Rd, Singapore 599 494

Received 17 February 2025; revised 06 June 2025; accepted 17 June 2025

The Titanic-related data was originally compiled by the British Board of Trade as part of its investigation into the tragic sinking of the Royal Mail Ship Titanic. For many years, research on the Titanic disaster remained largely in the domains of historians and enthusiasts. Its popularity in the machine learning community surged after Kaggle released a curated version of the dataset for a data analysis competition. Since then, it has been widely adopted for data science education and research, including its use in teaching data preprocessing and analysis, as well as benchmarking the performance of different machine learning algorithms. However, there is a *previously overlooked* flaw in this dataset: this paper shows that the average passenger class fares computed from the Kaggle dataset differ substantially from those published by NBC Los Angeles News in June 2023. In particular, the incorrect assignment of group fares to individual passenger fares has caused systematic inflation of fare values, potentially leading to misinterpretations over the years. A methodological correction for the Fare attribute is proposed, whereby group fares are divided equally among all passengers within the same travel group. This adjustment yields a significant 15.6% improvement in Spearman's correlation between the fare and passenger class. Additionally, experimental results demonstrate that fare correction improves prediction performance in classification and regression tree. It is hoped that this correction will enhance the dataset's utility for future education and research.

Keywords: Data preprocessing, Kaggle dataset, Machine learning, Passenger survival analysis, Titanic fare

Introduction

In 1912, the Royal Mail Ship (RMS) Titanic, which its engineers deemed practically unsinkable, embarked on its maiden voyage to New York City. Unfortunately, during this inaugural journey, the ship was struck by an iceberg and suffered one of the most tragic maritime disasters in history. The sinking resulted in the deaths of about 68% of its 2207 passengers and crew.[1]

Despite the dangerous situation, a significant number of women and children were saved, as they were given priority during the rescue operations. When the first collated version of the Titanic dataset was published by *Dawson*[2] in 1995, there was considerable interest in understanding how age, sex, and socioeconomic status (suggested by passenger class) affected survival rates. Over time, additional attributes such as fare, ticket number, and cabin were included. These new attributes have made this dataset even more useful for education and research.

The dataset became extremely popular in the machine learning community after Kaggle hosted[3] it for an open and perpetual data analytics competition in year 2012. To have a sense of the popularity of the Titanic data competition, the leaderboard[3] data recorded from 13 Apr 2025 to 7 June 2025 was analysed. The analysis found a whopping 56767 submissions made by 14436 teams in just eight weeks. Fig. 1 shows a breakdown in the number of teams and submissions over these weeks.

The dataset used for the competition was a lot richer than that published[2] by Dawson, which includes additional attributes such as the number of siblings/spouses aboard, the parents/children aboard.

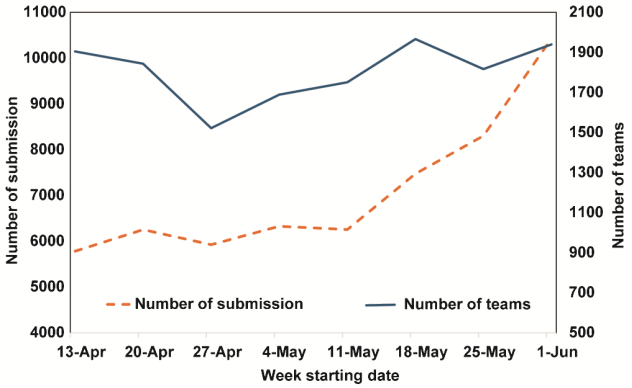

Fig. 1 — The leaderboard of Kaggle Titanic data competition website[3] shows that a total of 56767 submissions were received from 14436 teams over eight weeks

Author for Correspondence
E-mail: jamestansc@suss.edu.sg

After undergoing a series of refinements, an even more complete dataset[4] was made available on Kaggle. In 2019, Symanzik *et al.* presented a comprehensive overview[5] of the Titanic dataset and its variants. In their paper, they found over 40 articles and books featuring graphs derived from the Titanic dataset. They also identified at least 12 R packages containing 17 versions of this dataset. All of this suggests the widespread popularity of this dataset and its ongoing interest.

Despite the continual refinement and enhancement of the dataset, questions remain about the exact meaning of some attributes. In particular, it is unclear whether the Fare attribute should be treated as individual or group fares. The aim of this paper is to provide data analysis evidence that the Fare attribute should be treated as group fares, rather than individual fares assumed by most users. This difference has important implications for historical understanding and validity of data analyses.

Related Work

The Titanic dataset gained popularity after Kaggle shared it with the data analytics community for an open and perpetual analytics competition.[3] Since then, the dataset gained much attention for several reasons. Firstly, the underlying pattern of the data suggests that passengers of certain genders and socioeconomic statuses had higher survival rates. This makes an interesting case study for social science discussions. Secondly, the dataset contains mixed data types and discoverable patterns, which make it well suited for educational and research purposes in machine learning and data analytics.

For educational purposes, the inherent data quality issues of the dataset make it suitable for teaching data preparation and analysis. Lindemann & Stolz used it for teaching[6] statistics and mixed methods research. For research, it is widely used to study the performance of various machine learning algorithms in predicting passenger survival. For example, Dasgupta *et al.*[7] applied Logistic Regression to understand the extent to which female passengers were more likely to survive compared to males. Another example is the work of Liang[8] in 2023, who explored the effect of replacing missing values in the Age attribute and how that affects prediction performance. Other researchers, such as Singh *et al.* studied[9] the prediction performance of Logistic Regression, Naïve Bayes, Decision tree and Random Forest in making survival predictions and identified the key predictors. Apart from machine learning, there was also an attempt to develop a survival scorecard.[10]

In most studies, the Fare attribute was either omitted or used without further analysis of its actual meaning. For example, the Fare attribute was used directly in data exploration[11] or for analyzing dataset structure.[12] From a social science perspective, the Titanic fare can also be used to study the affluent passengers on board. Thus, an accurate fare estimation will help one to gain a better understanding of socioeconomics during that period.

This paper proposes a methodological adjustment to the Fare attribute that accounts for group travel arrangements. Empirical evidence is presented through data analysis and support from relevant literature. Finally, fare normalisation is shown to significantly improve the reliability and interpretability of the analytical results.

Data and Method

The complete dataset[4] was obtained from Kaggle. This dataset contains details of 1,309 Titanic passengers, recording names, demographics (i.e., age, sex), ticket details (i.e., ticket number, fare, passenger class), group travel details (i.e., travelling with siblings, spouses, parents, children), and survival outcomes.

The fare has a range of £0 to £512, suggesting a wide range of fares. There are several examples of families with shared tickets. For example, Table 1 shows the fare details of the Allison family, which has four passengers, sharing the same ticket number of [113781], and the fare of £151.55 in total.

The investigation begins by examining the average ticket fare for each passenger class. The average fares computed from the Kaggle dataset differ substantially from those published[13] by NBC Los Angeles News in June 2023. Further investigation reveals that passengers travelling with their families often have the same ticket numbers. This suggests that the price indicated in the dataset was the total price for the

Table 1 — Fare details of the Allison family, likely a couple travelling with two very young children

Name	Sex	Age	Sibsp	Parch	Ticket	Fare (£)
Allison, Master. Hudson Trevor	male	0.9167	1	2	113781	151.55
Allison, Miss. Helen Loraine	female	2	1	2	113781	151.55
Allison, Mr. Hudson Joshua Creighton	male	30	1	2	113781	151.55
Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25	1	2	113781	151.55

cabin(s) or suite(s) they stayed in. This means that the fare needs to be divided by the number of passengers sharing the same ticket number.

Specifically, let $\mathcal{P}$ be a set of all passengers in the Titanic dataset. For each passenger $p \in \mathcal{P}$:

- Let $T(p)$ be p's ticket number
- Let $F_o(p)$ be p's *original* fare
- Let $\mathcal{G}(p) = \{q \in \mathcal{P} \mid T(q) = T(p)\}$ be the group sharing the same ticket number

Then, the corrected fare $F_c(p)$ is:

$$F_c(p) = F_o(p) / |\mathcal{G}(p)|$$

where, $|\mathcal{G}(p)|$ is the group size.

For each passenger class $k \in \{1, 2, 3\}$, the mean fare for each fare type is computed as follows:

$$\overline{F}_{type}(k) = \frac{1}{|\mathcal{P}_k|} \sum_{p \in \mathcal{P}_k} F_{type}(p)$$

where,

- $\mathcal{P}_k = \{p \in \mathcal{P} \mid \text{class}(p) = k\}$
- *type* = {*c*, *o*, *pub*} is the fare type for *corrected*, *original*, and *published* reference fare, respectively.

The above-mentioned method has the advantage of normalising the inflated fare. It assumes that passengers in the same group ticket pays the same fare. For example, the correct fare for the Allison family should have been £151.55/4 (i.e., £37.89) for each individual passenger. However, this assumption may be oversimplified if a group consists of masters and helpers. This issue will illustrated in a case example later.

Results

The overall average ticket fare computed from the original Kaggle dataset is about £33.30 as given in Table 2. However, for the fares published[13] by NBC Los Angeles News, the overall mean fare is approximately £13.73. This discrepancy suggests that the Fare in the original dataset could have been misinterpreted and incorrectly used. On the other hand, the corrected mean fare is £15.50, which is a lot closer to the mean published fare. In addition, the corrected fares also have smaller spreads as compared to the fares in the original dataset.

Apart from studying the overall means, the analysis also examines the breakdown of the fares between passengers travelling alone versus those travelling with one or more individuals (denoted as *In Group*). Table 3 shows that the fares computed from the original data for those who travelled *alone* are approximately half of those travelling in groups (which may include parents, children, spouse, siblings, and helpers). This fare pattern is consistent across all the passenger classes, which is unusual because the fares should be similar regardless of whether a passenger travelled alone or in groups.

It is clear from Table 4 that the corrected fares are similar between solo travellers and those travelling in groups, a logical pattern that further validates the proposed hypothesis.

Finally, the correlation between pclass and fare is computed, where pclass is treated as an ordinal number having values of 1 (First class), 2 (Second class), and 3 (Third class). The resulting Spearman's correlation is −0.718. However, when the corrected fare is used, its Spearman's correlation with pclass improves to −0.83, which is a 15.6% improvement.

Titanic Passenger Fares: A Case Example

A case example can be examined here, using the fare paid by passengers who stayed in the most expensive suite—Charlotte Drake Cardeza, a wealthy American. She paid £512 for a group ticket that included a maid, a valet, and her son, Thomas.

According to the dataset, Charlotte and Thomas stay in the luxurious suites B51, B53, and B55,

Table 3 — Fares computed from the original Kaggle dataset show discrepancies between individuals who travelled alone and those travelling in groups; All fares are in £

	Pclass		
In Group	First	Second	Third
No	65.47	15.25	9.10
Yes	109.14	29.06	21.67

Table 4 — Corrected fares are highly similar between individuals who travelled alone and those travelling in groups; All fares are in £

	Pclass		
In Group	First	Second	Third
No	35.92	11.64	7.77
Yes	38.04	11.11	6.39

Table 2 — Mean fares across three passenger classes (pclass), computed from the Kaggle dataset, *published* by NBC Los Angeles News, and the corrected version proposed by this paper; The standard deviation is after the ± sign of each mean fare; The last row shows the mean fares weighted by record counts ($|\mathcal{P}_k|$); All fares are in £

pclass	Mean fare (Kaggle) (F_o)	NBC LA Published Fare (F_{pub})	Corrected Mean Fare (F_c)	Count ($|\mathcal{P}_k|$)
1 (First)	87.51±80.45	From ~30	36.99±21.54	323
2 (Second)	21.18±13.61	~12	11.41±2.63	277
3 (Third)	13.30±11.49	~7	7.31±2.71	709
Mean	33.30±51.76	13.73	15.50±16.78	

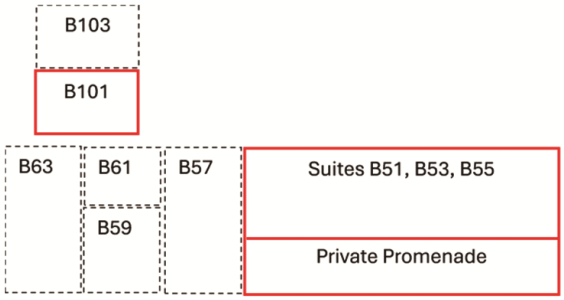

Fig. 2 — The B101 cabin occupied by the servants was at most one eighth the size of the suite of their masters (Adapted from source: Encyclopedia Titanica[15]

while the valet stayed in cabin B101. The Cave list also suggests that the maid could have stayed in B101 too. The Cave list is a name list held by a Titanic staff member named Cave[14], which was found after the tragedy.

From the deck plan[15], one can visually estimate that cabins B51, B53, B55 and its private promenade combined are at least eight times larger than Cabin B101, as shown in Fig. 2. Thus, it is incorrect for the Titanic dataset to indicate a fare of £512 in the valet's and maid's records. It would be more accurate to treat £512 as a group fare and divide it by four for individual fare estimation. Specifically, the individual fare can be calculated using the group fare divided by the number of individuals sharing the same ticket number, which is the average fare of £128 per passenger for this group.

As mentioned earlier, the proposed fare correction method assumes an equal distribution of group fares among all passengers, which may be oversimplified when travel groups include individuals with different accommodations—such as this case example, where masters residing in large suites while helpers occupying smaller cabins. Fortunately, such cases are rare because of very few suites on Titanic. Therefore, the violation of this assumption is not serious.

Evaluation Using Predictive Models

Finally, the effect of fare correction on passenger class prediction performance was evaluated using Classification and Regression Tree (CART). The dataset (1309 instances) was split into 80% for training and 20% for testing. To account for experimental variability, the model evaluation was repeated 30 times, with each iteration using a unique random seed for training and testing data partitioning.

Across 30 experimental runs, the CART model attained an average prediction accuracy of 87.13% (±33.38%) using the original dataset. However, after applying fare correction, the average accuracy is improved by 8.6 percentage points, to 95.73% (±19.09%). These results suggest higher precision and greater stability (reduced standard deviation). This improvement is also confirmed by conducting a matched-pairs t-test, with $t = 13.96$ and $p < 0.001$. These improvements in results underscore the importance of applying the fare correction method.

Conclusions

This study demonstrates how rigorous data validation can rectify a long-standing error that remained hidden for more than a decade. Specifically, a critical error in the Titanic dataset's Fare attribute was identified and corrected by accounting for group ticket sharing. This correction results in more accurate data analyses and improved performance of machine learning models. Additionally, this fare correction improves the usefulness of the dataset for social science research. It facilitates a deeper socioeconomic analysis based on more accurate individual fare data, which is crucial for understanding how passengers' wealth might influence their survival rates during the disaster.

The enhanced dataset will be made available on Kaggle, offering researchers a more accurate benchmark for algorithm testing and a valuable case study in data preprocessing and validation for data science educators. It is also hoped that this work can lead to more reliable and insightful analyses across various fields, enhancing understanding of both data science and historical events.

References

1 Howells R, *The Myth of the Titanic,* Palgrave Macmillan, (1999).

2 Dawson R J M, The "Unusual Episode" data revisited, *J Stat Educ*, **3(3)** (1995).

3 Kaggle, *Titanic − Machine Learning from Disaster,* (2012), https://www.kaggle.com/c/titanic [Accessed 17 June 2025].

4 Vinicius B P, *The Complete Titanic Dataset*, (2020), https://www.kaggle.com/datasets/vinicius150987/titanic3 [Accessed 17 June 2025].

5 Symanzik J, Friendly M & Onder O, The unsinkable Titanic data, *Joint Stat Meetings*, (2019).

6 Lindemann A & Stolz J, Teaching mixed methods: Using the Titanic datasets to teach mixed methods data analysis, *European J Res Methods Behavioral Social Sci*, **17(3)** (2021) 231–249, https://doi.org/10.5964/meth.4241.

7 Dasgupta A, Mishra V P, Jha S, Singh B & Shukla V K, Predicting the likelihood of survival of Titanic's passengers

by machine learning. *2021 Int Conf Computat Intell Knowl Econ (ICCIKE)*, 52–57 (2021), https://doi.org/10.1109/ICCIKE51210.2021.9410751.

8 Liang W, (n.d.). *Disaster Prediction based on Machine Learning Algorithms*, Department of Statistics, University of Toronto.

9 Singh A, Saraswat S & Faujdar N, Analyzing Titanic disaster using machine learning algorithms, *2017 Int Conf Comput, Commun Automat (ICCCA)*, 406–411 (2017), https://doi.org/10.1109/CCAA.2017.8229850.

10 Ligot D V, *Developing a Titanic Survival Scorecard: Risk Analysis of Populations Through Statistical Scoring Methods*, SSRN, (2022), http://dx.doi.org/10.2139/ssrn.4015684.

11 Lam E & Tang C, *CS229 Titanic – Machine Learning from Disaster,* Stanford University, (2012), https://cs229.stanford.edu/proj2012/LamTang-TitanicMachineLearningFrom Disaster.pdf.

12 Buzmakov A, Kuznetsov S O, Makhalova T & Napoli A, Exploring the dataset structure by means of delta-classes of equivalence: The case of the Titanic dataset, *The 9th Int Workshop "What Can FCA Do for Artificial Intelligence?"* (2021).

13 Gavin M, *History of the Titanic: 10 Questions about the Ill-Fated Ship.* NBC Los Angeles, (2023, June 20), https://www.nbclosangeles.com/news/national-international/history-of-the-titanic-10-questions-about-the-ill-fated-ship/3173692/ [Accessed 17 June 2025].

14 Encyclopedia Titanica, Cabin allocations, (n.d.-a), https://www.encyclopedia-titanica.org/cabins.html [Accessed 17 June 2025].

15 Encyclopedia Titanica, RMS Titanic: Plan of B deck, (n.d.-b), https://www.encyclopedia-titanica.org/titanic-deckplans/b-deck.html [Accessed 17 June 2025].