Data Cleaning and Visualization Analysis Based on Pandas and Matplotlib a Case Study of the Titanic Dataset

Kaiwen Zuo

School of Science and Technology, Beijing Normal-Hong Kong Baptist University, Zhuhai, 519087, China
t330026244@mail.bnbu.edu.cn

Abstract. In the big data time, data cleaning and visualization become essential for valuable information from raw materials. The very heady challenge of such work is to know what works for whom and how it does. It needs systematic exercises for data preprocessing and exploratory visualization. The classic Titanic passenger's dataset is used in this work to fill this gap. The Pandas library was applied for systematic data cleaning. This involved dealing with missing values in 'Age' and 'Embarked', dropping the 'Cabin' column, as well as creating new features such as 'Family Size'. Then, the paper plotted a sequence of charts in Matplotlib. These plots looked at the relationship between passenger survival rates and important factors such as gender, class, age. Social-demographic factors influencing survival were indicated in the results. This validates the efficacy of pairing thorough data cleaning with deep visualization. Further research may include additional external variables and more sophisticated visual tools. This would enhance the depth and explicative power of analysis.

Keywords: Data Cleaning, Data Visualization, Survival Analysis.

1 Introduction

With the Further advance of science and technology, people have entered the era of big data. Data is very significant this age. In business analysis, political decision making, scientific exploration or even daily life, data has been playing critical roles in the promotion of social advancement and the advance of human knowledge. In recent years, and with the production of large volumes of data as well as raw data often presenting missing or abnormal values, this aspect has become primordial- knowing how to clean, process and analyze large volumes of data. As McKinney claims, cleaning and transforming the data is one of the most onerous tasks in any data work flow and accounts for the majority of time spent working with data [1].

In the age of big data, programming languages like Python can be incredibly useful and supportive when working with large quantities of data. Of these, the Pandas library is provides powerful functionality for preparation and analysis of data as well as an excellent flexible Data Frame data-structure [2]. For data visualization, Matplotlib offers a robust base with its developed 2D graphics platform [3].

© The Author(s) 2026
K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_85

Methodologically, on the other hand, there is an advanced state of research. McKinneys book is a comprehensive instruction on handling, scrubbing and mining data with Python [1]. The article "Challenges in benchmarking stream learning algorithms with real-world data" it is a nice summary of data preprocessing pipelines [4]. Munzner's theoretical work provides valuable insight to researcher sand practitioners in visual data analysis [5].

Although existing research provides a rich theoretical and technical basis for data cleaning and visualization, most studies remain at the methodological level. They do not combine data preprocessing with exploratory visualization and lack systematic practical research based on real datasets. Therefore, studying the complete analysis process is very important.

In this paper, the Titanic passenger data is utilized as an essential one. It performs data fixing and visual analysis on this dataset in Pandas and Matplotlib. This study will initially check the quality of data and clean the dataset using Pandas tools. It includes some operations like: filling the NaN data in Age column, filling up the missing values of column Embarked, dropping the feature Cabin and finally creating new features like Family Size. The approach to use these methods will be based on the systematic data cleaning framework by de Jonge et al. [6].

From the sanitized data, the paper will proceed to draw a collection of visual charts using Matplotlib. This will answer questions about the dynamics of survival rates and what might have been contributing to this. The paper tries to adhere as much as possible with Munzner's principles of visualization design, in order to ensure optimized information transmit on by the charts [5]. The goal of this paper is to capture a replicable data analysis procedure as an example for subsequent exercises in data science.

The organization of the paper is as follows: In Chapter 2, it focuses on some research tools with their related theoretical basis such as Pandas, Matplotlib and basic processing of data. The process of data cleaning and preprocessing is described in chapter 3. The findings of the data visualization analysis are presented and discussed in Chapter 4. Finally, the research conclusions and future development area presented in Chapter 5.

2 Research Tools and Theoretical Basi

2.1 Pandas and Data Preprocessing

This work is written in the combination of Python data analysis ecosystem, which uses package Pandas for cleaning, preprocessing. Pandas is the go-to tool for working with structured data in Python. Its fast and navigable representation of data is particularly useful for preprocessing. Moreover, Pandas Data Frame can be used to efficiently store different types of data and execute rapid data filters, transformations, and joins - providing an established base for further analyses [1]. Similarly, VanderPlas also highlighted the efficiency of Pandas to process data and create features including missing values imputation and type conversion [7]. Given these features, the paper will use Pandas to perform an in-depth examination of data types and variables distributions as well as missing proportion, and design cleaning policies. Meanwhile, similar

research inspiration, the paper coded some common feature such as Family Size, Is Alone and Title to enhance the interpretability of following data analysis process.

2.2 Matplotlib and Data Visualization

On the data visualization, as the most classical data visualization library in Py-thon, Matplotlib gives strong support for graph drawing of this study. Hunter provided an explanation of its design philosophy, which focuses on providing a MATLAB-style syntax for 10 easy to use plotting functions [3]. The library has rich set of graph types and is very customizable. Moreover, as noted by Munzner that visual design is the critical component for communicating data [5], and Matplotlib, due to its flexibility and extensibility de- signs can easily support different types of visual encoding. With respect to these benefits, the present paper uses Matplotlib to plot different statistical charts such as bar diagrams, scatter plots and heatmaps respectively. It aims to illustrate the contrasted survival of classes, sex groups and age categories, and uses heatmaps to display the correlation structure between variables.

2.3 Pandas and Data Preprocessing

In general, the research process consists of four steps: data exploration, data cleaning, feature engineering and visualization analysis. In addition to the two main tools, Pandas and Matplotlib, the study also extensively capitalizes on the computational benefits of NumPy arrays. Van der Walt et al. that NumPy serves as an effective foundation for numerical computation, thus enabling Pandas and Matplotlib to have good performance throughput even with large data [8]. This approach still stands in the construction of Pandas, the library that, as McKinney himself expressed, aims at bringing flexible and easy-to-use data structures to Python [2]. In the specific realization process, the data exploration stage uses Pandas to conduct descriptive statistical analysis and distribution analysis to have a preliminary understanding of our data. The data cleaning step is based on the data cleaning method recommended by van der Loo and de Jonge fora systematic treatment of missing values and outliers [6]. During feature engineering, the paper refers to the general Titanic analysis pipeline on Kaggle and build a set of derived variables [9, 10]. Finally, visualization of the data is performed with Matplotlib in multi-dimensional analysis.

3 Preprocessing methods and procedures for the Titanic dataset

3.1 Preliminary Data Analysis

The Titanic dataset comes from Kaggle. It contains 891 records and 12 feature fields. Initial checks showed that Age, Cabin, and Embarked have missing values to varying degrees. Cabin has the highest missing rate, reaching 77%. Therefore, different completion methods need to be developed based on variable characteristics.

3.2 Missing Value Handling

Three different strategies are applied in this work for missing values. First, Cabin was removed all together because it was greatly under-accounted for and impossible to accurately estimate. Second, the null rate of Embarked is rather small, and the paper used mode for imputation. Age is also strongly correlated with passenger social characteristics, thus to prevent loss of accuracy in inference, it was inferred using the combination "Sex+ Class" median value.

3.3 Feature Engineering

Feature engineering is conducted with the purpose of data expressiveness. The paper constructed features such as Family Size, Is Alone and Title based on previous studies. Family Size is about how many people there are in one family and can assist in predicting whether size of family would matter to the survival rate. The passenger 'WasAlone' attribute is able to indicate if a passenger was not accompanied. Name is drawn from the title and serves as an indicator of societal status or culture. For the less frequent titles, they were included in the Rare category so as not to jeopardize feature stability.

3.4 Categorical Variable Encoding

To facilitate the analysis of data, text variables must be changed into numeric variables. The paper applies a mapping and recording method to certain variables such as Sex, Embarked, and Title so that the model can analyze them. The structure of the dataset is then being regularized after encoding and fits this processing stage better.

3.5 Cleaning Results Summary

After complete cleaning, there are no missing values in the dataset, and multiple new feature variables have been added, laying a good foundation for subsequent visualization analysis.

4 Data Visualization and Analysis

This chapter draw multiple sets of statistical charts based on the data cleaning results. These charts show the survival rate differences among different groups. Each chart is accompanied by a brief description and analysis to help readers understand the data relationships.

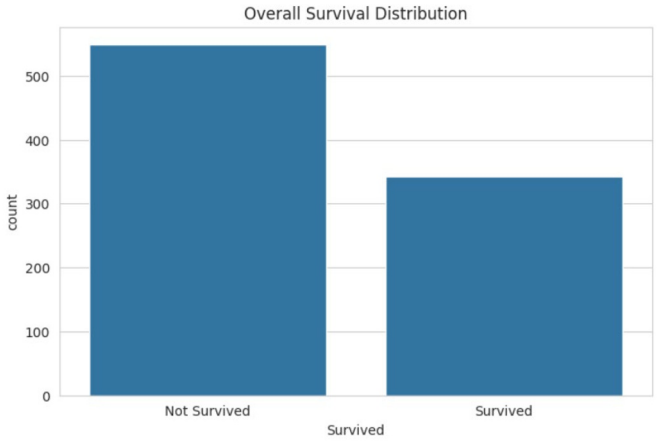

Fig. 1. Overall Survival Distribution

As shown in Figure 1, the overall survival rate is about 38%, indicating that most passengers were not rescued in the accident. This result provides a benchmark for subsequent analysis of survival rates in different subgroups.

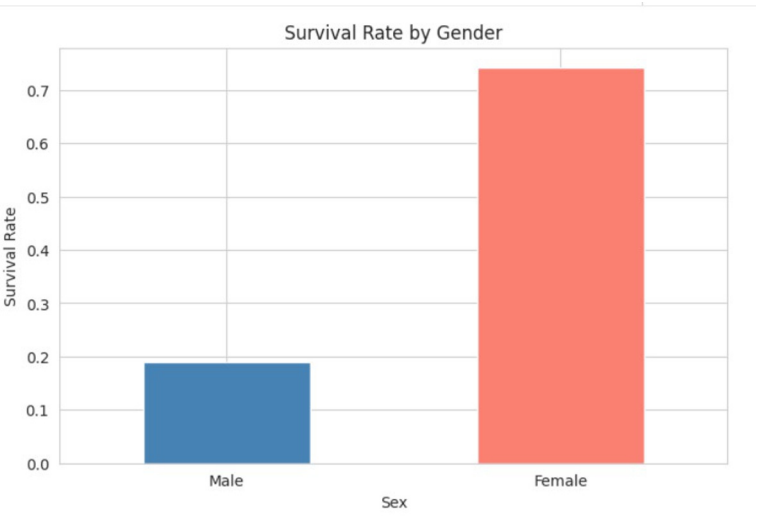

Fig. 2. Survival Rate by Gender

Figure 2 shows gender difference is significant. The survival rate of females is significantly higher than that of males. Considering the historical background, it can be inferred that the "women and children first" policy played a key role in the rescue, reflecting the important influence of gender roles in emergencies.

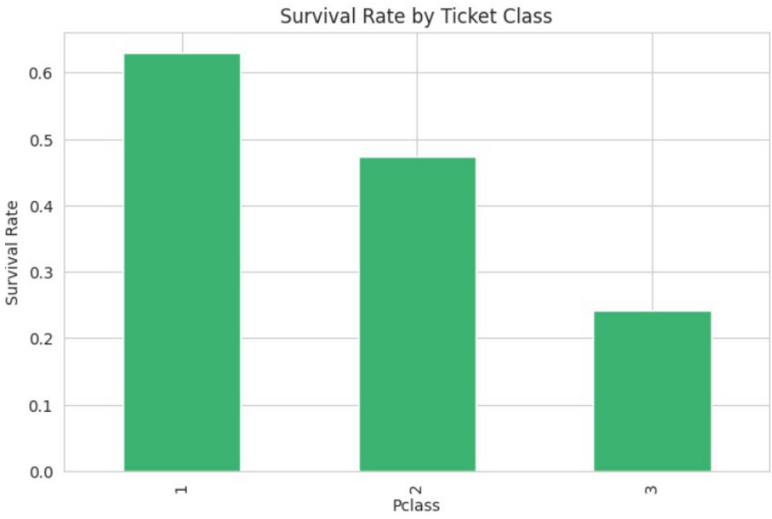

Fig. 3. Survival Rate by Ticket Class

As illustrated in Figure 3, higher-class passengers had a higher survival rate. Because higher-class cabins were often located on the upper decks, these passengers were closer to the lifeboats and could get information earlier. This also reflects the impact of social class differences in emergencies.

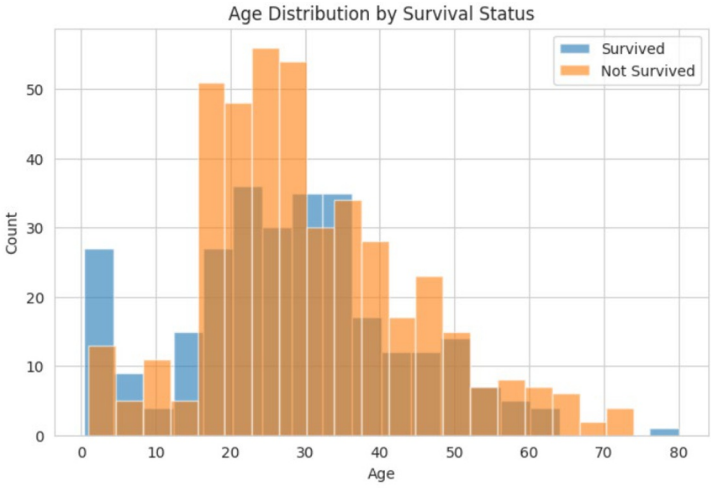

Fig. 4. Age Distribution by Survival Status

As shown in Figure 4, the survival rate of children was significantly higher than that of adults. This phenomenon may be related to their physical factors and the priority rescue

policy. The differences between age groups further verify the important role of age in emergency events.

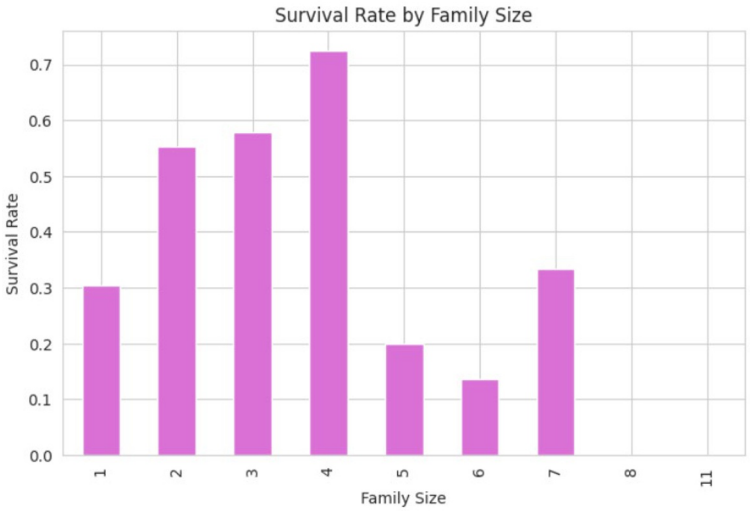

Fig. 5. Survival Rate by Family Size

Figure 5 shows that the relationship between family size and survival rate is inverted U-shaped. When the number of family members is moderate, mutual assistance may increase the chance of survival. However, too large or too small family sizes may have disadvantages during escape.

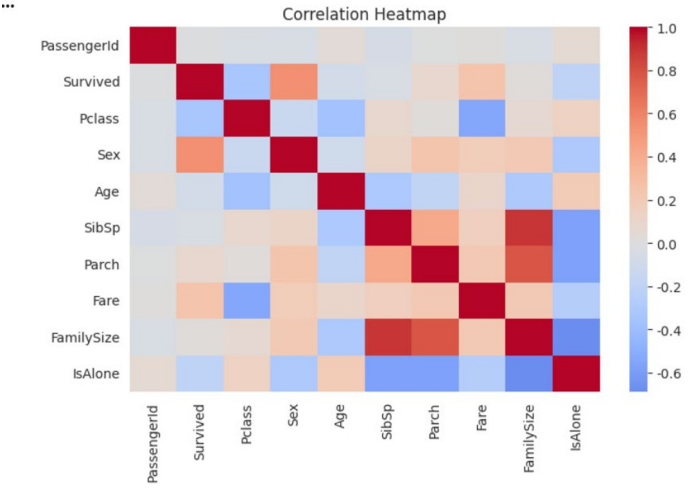

Fig. 6. Correlation Heatmap

As shown in Figure 6, the heatmap shows the correlation structure between main variables. Gender, class, and family size show strong correlations with survival rate, helping to comprehensively understand the influence of different factors.

5 Conclusion

This article systematically used Pandas and Matplotlib to preprocess the Titanic dataset, then perform some visual based analysis on it. It revealed important trends in regard to the survival rate of passengers. The findings indicate that gender, passenger class, age and family size were significant predictors of survival. Women, children, and those of a higher class had significantly better chances of survival. This is consistent with accounts of the "women and children first" saving rule as well as the influence of class in emergencies.

A clear and reproducible data analysis pipeline could be compiled out of the aforementioned systematic (e.g., median imputation and mode filling) cleaning combined with specific visualization. This work is restricted by the data scale and the simplicity of feature engineering though. Further studies may incorporate additional external features, e.g., ticket fare patterns and social network structures of passengers, to extend the dimensions of analysis. Also, connecting with soft wares that enable interactive visualization or ML libraries can be advantageous to generate deeper insights and more dynamic data exploration. These enhancements will continue to close the gap between data pre-processing and analytic findings that are actionable. They will enable more sophisticated real world data science applications.

References

1. McKinney, W.: Python for Data Analysis. O'Reilly Media, Location (2017)
2. McKinney, W.: Data Structures for Statistical Computing in Python. In: Proceedings of the 9th Python in Science Conference, pp. 56–61 (2011)
3. Hunter, J. D.: Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering 9(3), 90–95 (2007)
4. Anonymous: Challenges in Benchmarking Stream Learning Algorithms with Real-world Data. Data Mining and Knowledge Discovery 34(1), 1–30 (2020)
5. Munzner, T.: Visualization Analysis and Design. CRC Press, Location (2014)
6. Van der Loo, M., De Jonge, E.: Introduction to Data Cleaning with R. Statistics Netherlands, Location (2013)
7. VanderPlas, J.: Python Data Science Handbook. O'Reilly Media, Location (2016)
8. Van der Walt, S., Colbert, S.C., Varoquaux, G.: The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering 13(2), 22–30 (2011)
9. StartupSci: Titanic Data Science Solutions. Kaggle. https://kaggle.com, last accessed (2023)
10. Raschka, S., & Mirjalili, V.: Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2 (3rd ed.). Packt Publishing. (2019)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.