

Lab 0: Looking at the Titanic

Abstract

This lab uses the supplied Titanic passenger data to describe how survival varied across passenger characteristics. I loaded 891 records, filled missing ages with the median, filled two missing embarkation values with the mode, and compared group survival rates. Survival was highest in first class (63.0%) and among female passengers (74.2%), while third class (24.2%) and male passengers (18.9%) had much lower rates. Small-to-medium family groups also had higher observed survival rates than people travelling alone or in very large groups. These comparisons identify strong patterns in this dataset, but they cannot show that any one characteristic caused survival.

Background and Motivation

The Titanic dataset records passenger information such as ticket class, sex, age, family relationships, fare, embarkation port, and survival outcome. It is a useful introductory dataset because it combines numeric and categorical variables and has realistic missing values. The main motivation is to practice a small, reproducible analysis workflow: inspect data quality, make a limited set of transformations, summarize groups, and communicate results with figures.

Survival during a shipwreck is not determined by a single variable. Passenger class may be connected to cabin location, ticket resources, and access to crew instructions; sex and age may be related to evacuation practices. Therefore, this lab treats the data as observational. The goal is to describe associations rather than make claims about what definitely caused individual outcomes. It is important to note that correlation does not always imply causation, especially when there are multiple variables in play.

Problem Statement

The question is: how did observed Titanic survival rates differ by passenger class and other available passenger characteristics? This is challenging because the data are incomplete and the variables are related to each other. For example, class and fare both reflect socioeconomic position, so a simple class comparison may also reflect other unmeasured differences. I use grouped counts and mean survival outcomes to give an interpretable baseline rather than attempting a predictive or causal model.

System Design

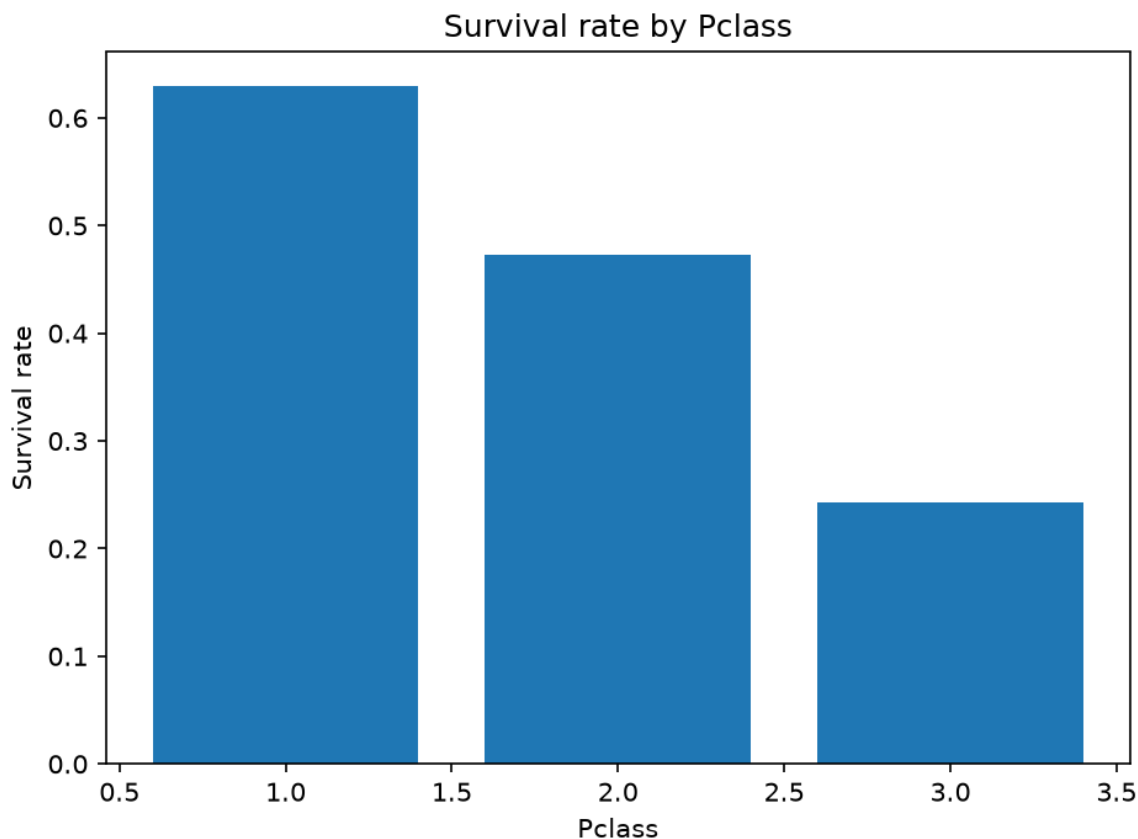
I read `data/titanic.csv` into a pandas `DataFrame` without modifying the source file. The

analysis makes a copy before cleaning so the raw data remains available for checking missingness. Missing `Age` values are replaced with the median age and missing `Embarked` values with its most frequent value. I used these choices because they are common ways to fill in missing data; I did not compare them with alternatives. `Cabin` is retained but not used in the comparisons because 687 of 891 values are missing. I also added `FamilySize = SibSp + Parch + 1`, where the final one counts the passenger.

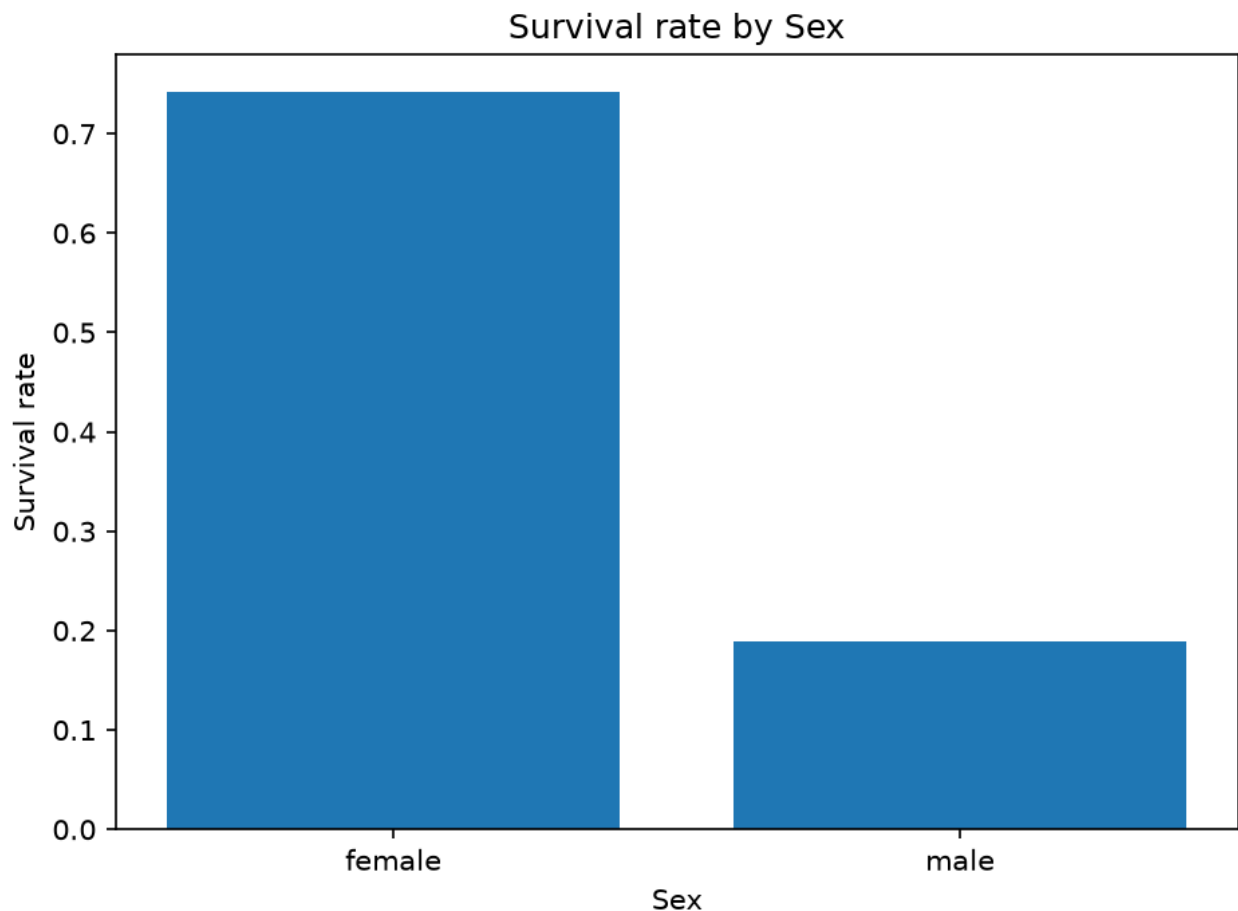
For each grouping variable, I calculated the number of passengers and the mean of the binary `Survived` column. That mean is the survival rate. I selected the grouping variables because they looked interesting in the CSV, rather than from a pre-specified hypothesis. Bar charts use the same 0--1 vertical scale when comparing rates. This is a simple but effective approach for a first pass.

Evaluation

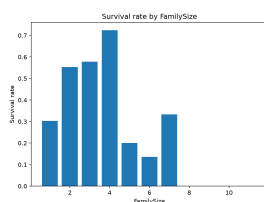
I explored survival by passenger class, sex, embarkation port, family size, and an optional set of age ranges. All comparisons use the 891 cleaned passenger records. The class result in Figure 1 is a clear gradient: first-class passengers had a 63.0% survival rate (136/216), second-class passengers had a 47.3% rate (87/184), and third-class passengers had a 24.2% rate (119/491). The lower the class number, the higher the survival rate, which is a pretty clear trend.



Sex showed the largest descriptive separation: 74.2% of female passengers survived (233/314), compared with 18.9% of male passengers (109/577), as shown in Figure 2. This makes sex the most important variable for predicting survival. The embarkation comparison was less direct: rates were 55.4% for Cherbourg, 39.0% for Queenstown, and 33.9% for Southampton. Passengers who boarded at Cherbourg were therefore safer, perhaps because that port gave them better access to lifeboats. These port groups may differ in class composition, so the graph should not be read as an effect of where a person boarded.



Family size produced a non-linear pattern. Passengers travelling alone had a 30.4% survival rate, groups of two through four ranged from 55.3% to 72.4%, and many larger groups had low rates. Figure 3 makes the pattern visible, but the largest sizes contain only 6--22 people, so those values are unstable. The age-group check found 58.0% survival among the 69 children age 12 or younger, but age was median-imputed for 177 passengers, which adds uncertainty.



Overall, class, sex, and family size are useful variables for further analysis. Together they explain why some passengers survived and others did not. A next step would stratify class results by sex and use confidence intervals or a logistic regression to account for related predictors.

Future Directions & Challenges

The main challenge was missing data: 77.1% of `Cabin` values are absent and 19.9% of ages are absent. Median and mode imputation allow the introductory plots to run, but they can hide real variation and are not necessarily the best choice for a later model. The data also omit important details about location, lifeboat access, and the timing of evacuation.

Future work could compare class and sex together, report uncertainty for each group, and test whether family size still matters after controlling for class and sex. It would also be worth checking results with a missing-age indicator rather than treating every imputed age as equally certain. In other words, more data does not automatically mean more truth; the data needs to be good too.

Conclusion

The Titanic data show substantial survival-rate differences by passenger class and sex, with a more complicated association for family size. The lab also established a repeatable workflow for loading, cleaning, plotting, and reporting data. These results are descriptive, not causal, but they provide a solid baseline for more careful analysis.