AI Mastery Capstone - Project 1: AI Programming Foundations Project

Frank Allen Motley

Udacity Institute of AI & Technology

Master of Science in Artificial Intelligence

AI Mastery Capstone

AI Programming Foundations Project: Reproducible Data Workflow

Overview

Using the Titanic dataset (Seaborn example dataset), this project builds a complete and reproducible data workflow in Python. The workflow loads the dataset, applies transparent cleaning functions, performs exploratory data analysis (EDA), produces three labeled visualizations, and summarizes findings with limitations and responsible data handling considerations. Dataset source: https://raw.githubusercontent.com/mwaskom/seaborn-data/master/titanic.csv.

Dataset Description

The Titanic dataset contains 891 rows and 15 columns describing passenger attributes and outcomes, including survival status (`survived`), passenger class (`class`/`pclass`), sex (`sex`), age (`age`), fare (`fare`), embarkation port (`embarked`), and related descriptors. The analysis focuses on survival differences across sex and passenger class, and explores the distribution of age by survival outcome.

Workflow Description

Ingestion: The notebook loads `titanic.csv` locally if present; otherwise it downloads the CSV from the published Seaborn-data repository and caches it locally.
Cleaning: Two reusable functions standardize column names and impute missing values (median for numeric columns; mode for categorical columns). Categorical columns are optionally cast to `category` to support grouping.
Exploratory analysis: A reusable `eda_summary()` function produces dataset shape, dtypes, numeric summary statistics, and grouped survival rates by key categories.
Visualizations: Three plots are generated with titles and labeled axes and saved to the `figures/` directory.
Summary: The notebook concludes with interpretations, limitations, and proposed next steps for future ML/DL integration.

Key Decisions and Assumptions

Missing-data handling uses simple imputation rules (median/mode) to keep assumptions explicit and support fast iteration in an exploratory workflow. This choice can reduce variance and affect subgroup comparisons, so results are interpreted as descriptive rather than causal (Little & Rubin, 2019). To improve cross-machine reproducibility, dependencies are pinned to NumPy 1.x (`numpy==1.26.4`) because some compiled scientific packages may not yet be compatible with NumPy 2.x in all environments. Groupby operations set `observed=` explicitly to avoid future default changes in pandas. Plot outputs are saved using relative paths (`figures/`) for portability across operating systems.

Results and Interpretation

Overall survival rate is approximately 0.384. Survival differs substantially by sex: female passengers have a much higher mean survival rate than male passengers (Figure 1). Survival also differs by passenger class, with the highest survival in First class and the lowest in Third class (Figure 2). Age distributions by survival overlap considerably (Figure 3), suggesting age effects are nuanced and may interact with class and sex.

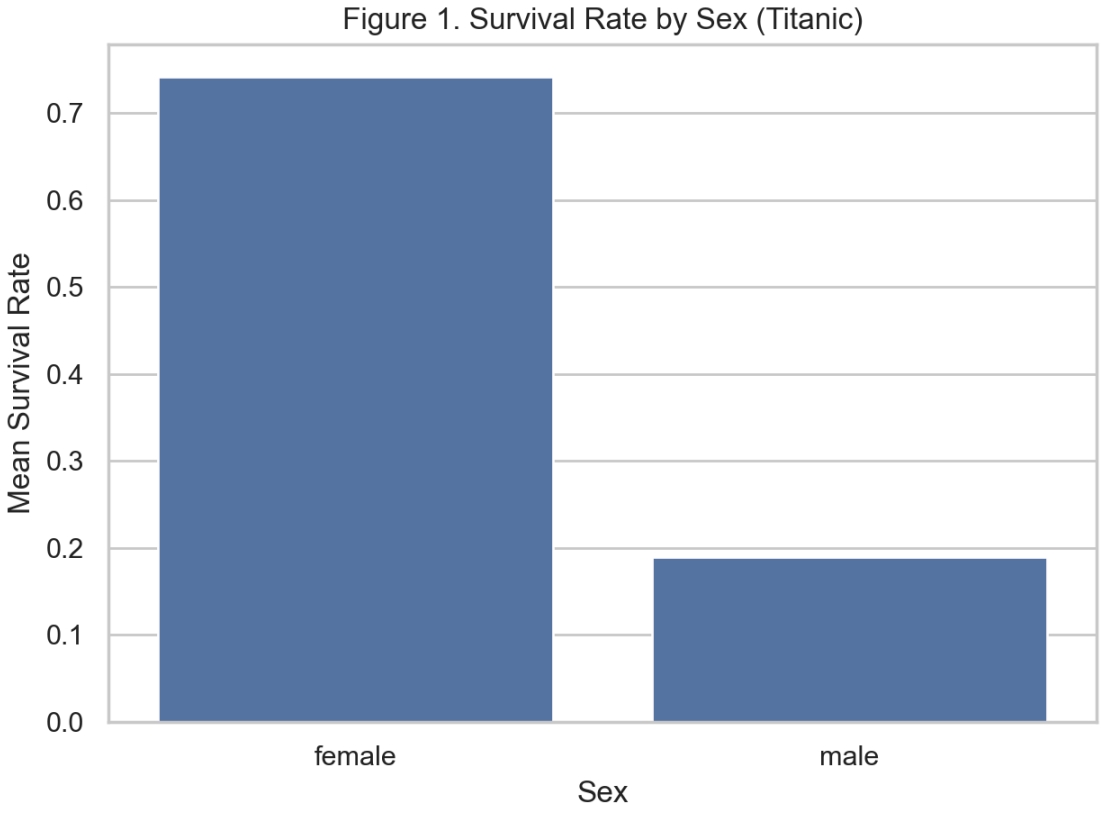


Figure 2. Survival Rate by Passenger Class (Titanic)

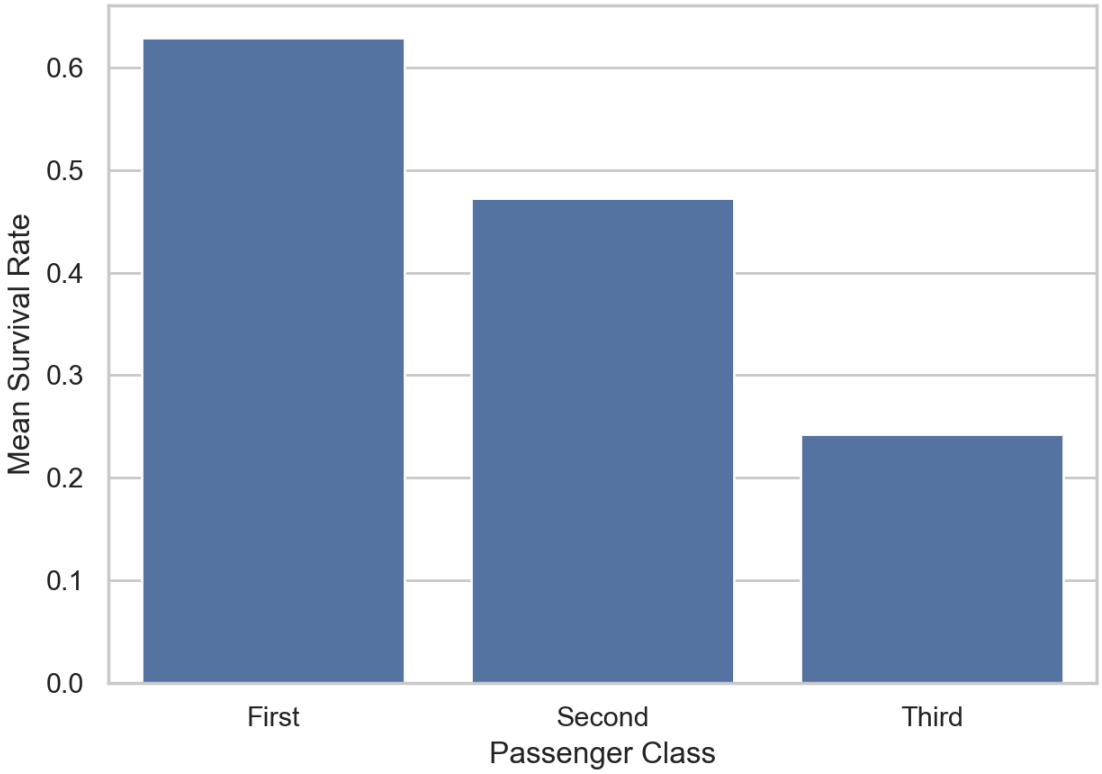
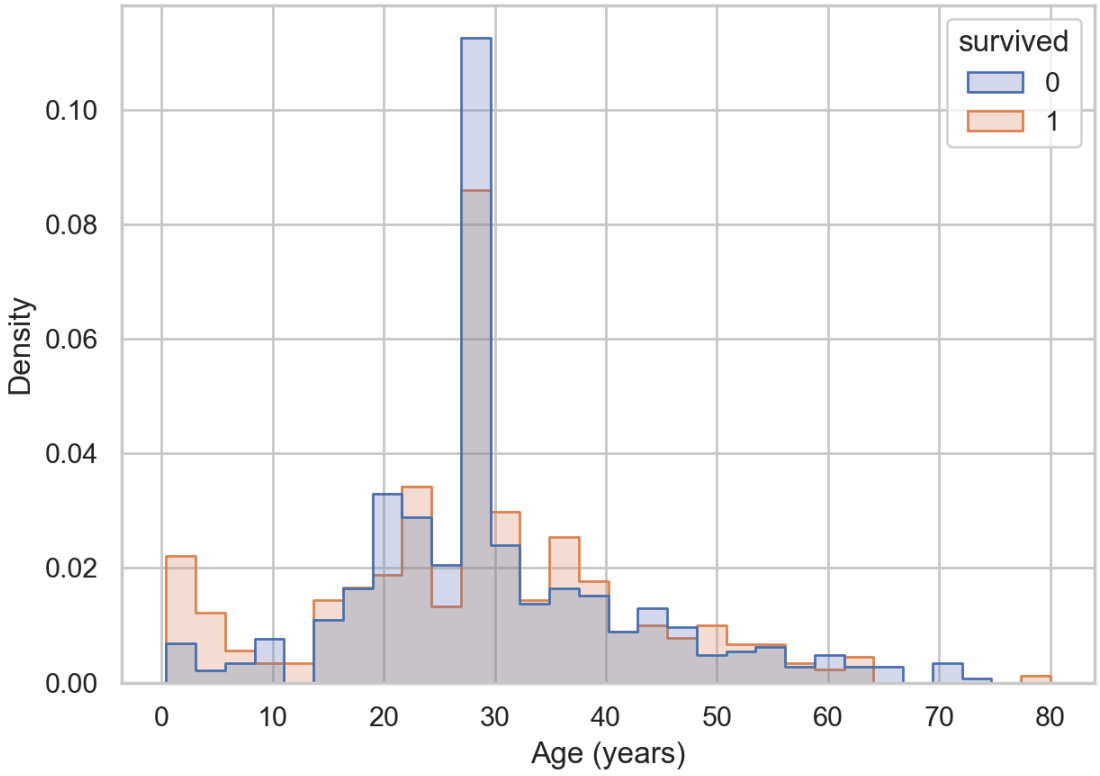


Figure 3. Age Distribution by Survival Outcome (Titanic)



Responsible Practice (Bias and Data Quality)

Several factors could introduce bias or misleading interpretations. Median/mode imputation can change distributions and attenuate real variability, especially for age. The dataset reflects a specific historical context and passenger population; it should not be generalized to other settings. To reduce risk, a next iteration would include sensitivity checks (e.g., comparing results with and without imputation), clearer missingness diagnostics, and stronger dataset documentation practices (Gebru et al., 2021).

Reproducibility

A reviewer can reproduce results by installing dependencies from `requirements.txt` and running `data_workflow.ipynb` top-to-bottom. The notebook uses relative file paths and creates the `figures/` directory automatically. Version control is demonstrated via multiple commits and an additional development branch in Git/GitHub (Chacon & Straub, 2014).

References

Chacon, S., & Straub, B. (2014). Pro Git (2nd ed.). Apress. https://git-scm.com/book/en/v2

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723

Little, R. J. A., & Rubin, D. B. (2019). Statistical analysis with missing data (3rd ed.). Wiley.

The pandas development team. (n.d.). Working with missing data. pandas documentation. https://pandas.pydata.org/docs/user_guide/missing_data.html

Matplotlib developers. (n.d.). Matplotlib documentation. https://matplotlib.org/stable/