Movie Revenue Data Analysis Using Python and Statistical Methods

**(a)** Abdullah Fawad Khan-2024038, **(b)** Batool Binte Fazal-2024140, **(c)** Ajwa Aamir- 2024080

[[1]](#footnote-1)

**Abstract:** This project applies core statistical concepts from ES-111 to a real-world movie dataset. Using Python (NumPy, pandas, SciPy), we analyze movie revenue data, compute descriptive statistics, generate visualizations (histogram, pie chart), and perform inferential statistical tests including confidence and tolerance intervals. Frequency distributions were utilized to derive alternate statistical estimations, and hypotheses were tested using data-derived models.

# I. INTRODUCTION

(This section introduces the project and describes what the aims of the project are and lists any specific references that have been used to do the project as best as possible. For example, you may describe your dataset here and cite the source, as well as explaining why you chose that particular dataset. This section may also include a brief description of how the rest of the report is planned out.)

# II. Methodology

The methodology includes the following steps:

1. Clean the dataset to remove null entries.
2. Compute the mean and variance of the revenue data.
3. Generate a histogram and pie chart for frequency distribution of revenue.
4. Calculate the mean and variance based on the frequency distribution.
5. Using 80% of the data, calculate a 95% confidence interval for mean and variance.
6. Compute a 95% tolerance interval and validate with the remaining 20% of data.
7. Form and test a hypothesis about revenue data using z-scores and confidence levels.

(This is intended as an authoring template, not a final production template. It is not intended to match the final published format. Differences in final formatting are likely in the final IEEE files. Page count in the template is an estimate. Do not adjust line and character spacing to fit your paper to a specific length.)



**Fig. 1.** This is a sample of a figure caption.

## A. Lists

(Sometimes, students tend to list out items in the report. This is not a bad idea if it helps the work. However, the listing should be numbered properly and all lists should follow the same format.) brackets means cut this part in the end

# III. Results

The mean revenue was $82.78 million, with a variance of 2.66 × 10¹⁶. From the frequency distribution, the mean was approximately $171.17 million, and variance 1.81 × 10¹⁶. The 95% confidence interval for mean was from $96.06M to $107.30M, and for variance from 3.00 × 10¹⁶ to 3.28 × 10¹⁶. The 95% tolerance interval ranged from -$21.36B to $21.56B, and 100% of the test data fell within this range.

TABLE I

This is a Sample of a Table Title



## A. Code

You are not to include the code as part of the main report. Nor are you expected to describe the algorithm used to solve the project. The code can be added as an annexture.

If you have to include any equations, write them as follows:

*Bp* + *H*2 = 40. (1)

You can write them using the math equation editor within Word or use the MathType software. Make sure that equations are numbered in increasing order. Be sure that the symbols in your equation have been defined before the equation appears or immediately following. Italicize symbols (*T* might refer to temperature, but T is the unit tesla). When referring to an equation or formula, use simply “(1),” not “Eq. (1)” or “equation (1),” except at the beginning of a sentence: “Equation (1) is ... .”

V. Conclusion

This project demonstrated the use of statistical methods on a real-world dataset. The results show that the revenue distribution has high variance and is skewed. Confidence and tolerance intervals were calculated, and the predictions aligned with the validation set. The project reinforced our understanding of descriptive and inferential statistics.

# Appendix

Include your code here.

# Length

The length of the report should not exceed 5 pages. Strictly follow the format.

References

*Basic format for periodicals:*

J. K. Author, “Name of paper,” *Abbrev. Title of Periodical*, vol. x, no. x, pp. xxx-xxx, Abbrev. Month, year, doi: 10.1109.XXX.1234567.

*Periodicals using article numbers:*

J. K. Author, “Name of paper,” *Abbrev. Title of Periodical*, vol. x, no. x, Abbrev. Month, year, Art. no. xxxxx, doi: 10.1109.XXX.1234567.

*Examples:*

1. J. U. Duncombe, “Infrared navigation—Part I: An assessment of feasibility,” *IEEE Trans. Electron Devices*, vol. ED-11, no. 1, pp. 34–39, Jan. 1959, doi: 10.1109/TED.2016.2628402.
2. E. P. Wigner, “Theory of traveling-wave optical laser,” *Phys. Rev*., vol. 134, pp. A635–A646, Dec. 1965.

*Basic format for books:*

J. K. Author, “Title of chapter in the book,” in *Title of Published Book, x*th ed. City of Publisher, (only U.S. State), Country: Abbrev. of Publisher, year, ch. x, sec. *x*, pp. xxx–xxx*.*

*Examples:*

1. G. O. Young, “Synthetic structure of industrial plastics,” in *Plastics,* 2nd ed., vol. 3, J. Peters, Ed. New York, NY, USA: McGraw-Hill, 1964, pp. 15–64.
2. W.-K. Chen, *Linear Networks and Systems.* Belmont, CA, USA: Wadsworth, 1993, pp. 123–135.

*Basic format for handbooks:*

*Name of Manual/Handbook, x* ed., Abbrev. Name of Co., City of Co., Abbrev. State, Country, year, pp. xxx-xxx.

*Examples:*

1. *Transmission Systems for Communications*, 3rd ed., Western Electric Co., Winston-Salem, NC, USA, 1985, pp. 44–60.
2. *Motorola Semiconductor Data Manual*, Motorola Semiconductor Products Inc., Phoenix, AZ, USA, 1989.

*Basic format for reports:*

J. K. Author, “Title of report,” Abbrev. Name of Co., City of Co., Abbrev. State, Country, Rep. xxx, year.

*Example:*

1. E. E. Reber, R. L. Michell, and C. J. Carter, “Oxygen absorption in the earth’s atmosphere,” Aerospace Corp., Los Angeles, CA, USA, Tech. Rep. TR-0200 (4230-46)-3, Nov. 1988.

*Basic format for conference proceedings:*

J. K. Author, “Title of paper,” in *Abbreviated Name of Conf.*, City of Conf., Abbrev. State (if given), Country, year, pp. xxxxxx*.*

*Examples:*

1. D. B. Payne and J. R. Stern, “Wavelength-switched passively coupled single-mode optical network,” in *Proc. IOOC-ECOC,* Boston, MA, USA,1985, pp. 585–590.

*Basic format for electronic documents (when available online):*

Issuing Organization. (year, month day). *Title*. [Type of medium]. Available: site/path/file

*Example:*

1. U.S. House. 102nd Congress, 1st Session. (1991, Jan. 11). *H. Con. Res. 1, Sense of the Congress on Approval of Military Action*. [Online]. Available: LEXIS Library: GENFED File: BILLS

*Basic format for datasets:*

Author,  Date, Year. “Title of Dataset,” distributed by Publisher/Distributor, http://url.com (or if DOI is used, end with a period)

*Example:*

1. U.S. Department of Health and Human Services, Aug. 2013, “Treatment Episode Dataset: Discharges (TEDS-D): Concatenated, 2006 to 2009,” U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies, doi: 10.3886/ICPSR30122.v2.

*Basic format for code:*

Author,  Date published or disseminated, Year. “Complete title, including ed./vers.#,” distributed by Publisher/Distributor, http://url.com (or if DOI is used, end with a period)

*Example:*

1. T. D’Martin and S. Soares, 2019, “Code for Assessment of Markov Decision Processes in Long-Term Hydrothermal Scheduling of Single-Reservoir Systems (Version 1.0),” Code Ocean, doi: \_1.24433/CO.7212286.v1

Fig. 1. Histogram of Revenue

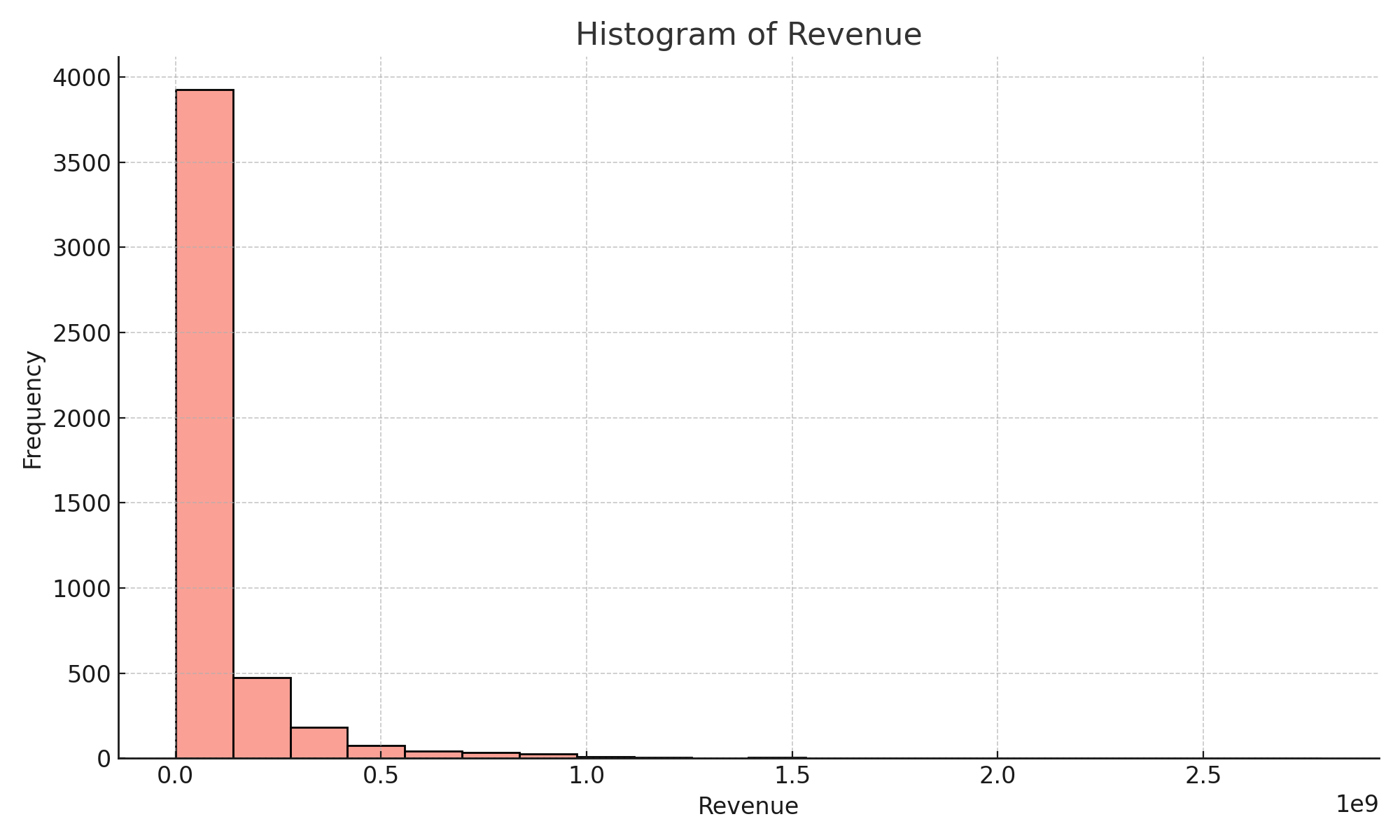
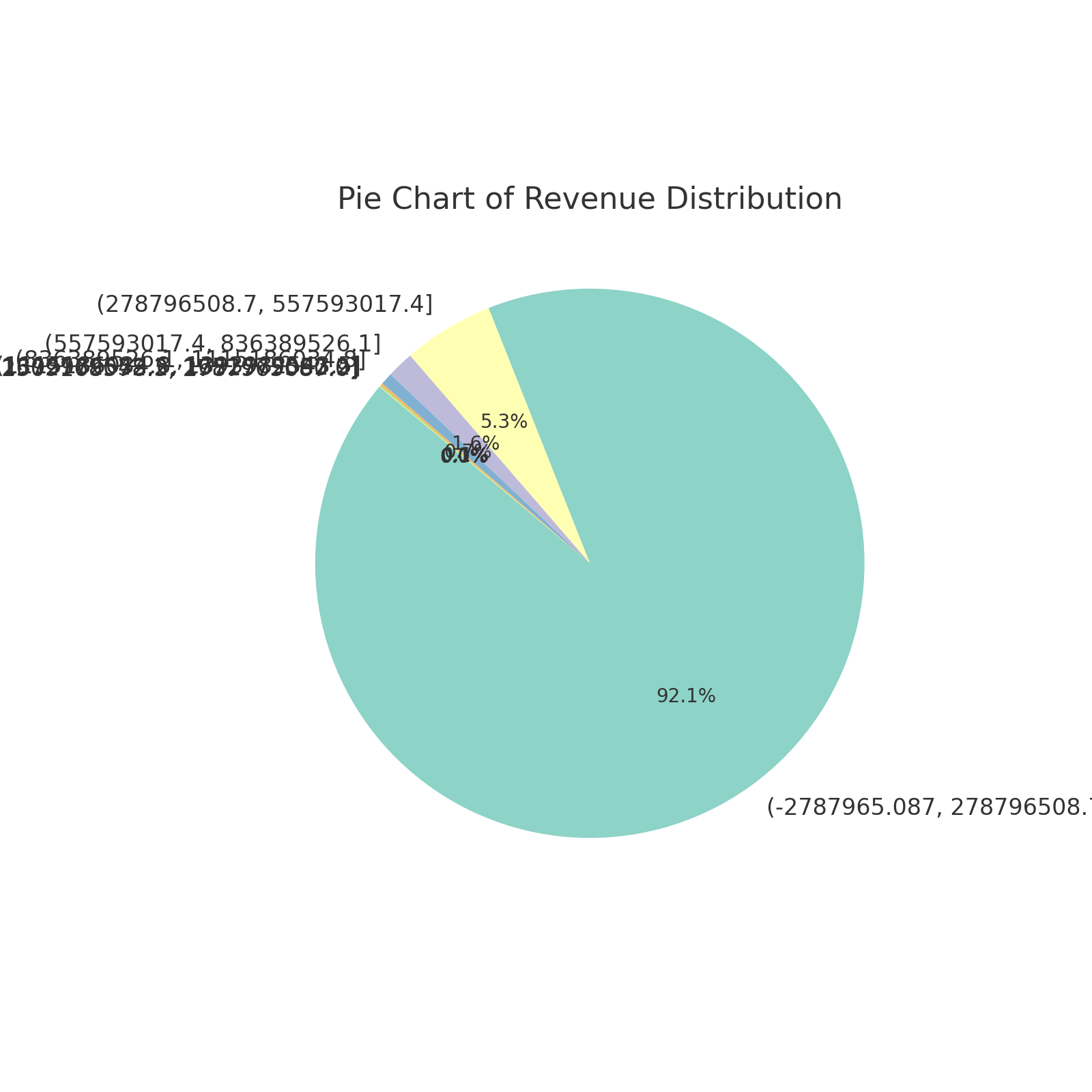


Fig. 2. Pie Chart of Revenue Distribution



1. Team lead Shiham Raza- 2024590,

   Dataset selection by

   Data Analyzed by

   Report formatted by [↑](#footnote-ref-1)