Statistical Analysis and Visualization of Sales Data

Zeeshan Ali [2024686] | Muhmmad Saad Ijaz [2024191] | Zain Ul Abeedin [2024671]

Abstract

This project focuses on extracting and analyzing statistical information from a sales dataset using Python. It includes computation of mean and variance, frequency distribution visualization using histograms and pie charts, and statistical analysis including confidence and tolerance intervals. Hypothesis testing is also performed to assess category-wise differences. Results validate the analysis process and showcase proficiency with statistical tools.

I. Introduction

The objective of this project is to demonstrate statistical analysis and data visualization using Python. The dataset, containing around 1,000–10,000 data points on sales transactions, was sourced online. The primary focus was to compute statistical measures and validate results through coding and visualization. The report is structured as follows: Section II details the methodology, Section III presents the results, and Section IV concludes with findings and insights.

II. Methodology

A. Basic Statistics

The mean and variance of the 'Amount' column were calculated using NumPy. Results:
Amount Mean: 5178.09
Amount Variance: 7860997.91

B. Frequency Distribution and Visualization

A histogram and pie chart were generated to visualize the distribution of sales amount and category-wise breakdown. Frequency distribution (bin center and frequency):

Bin center	frequency
982.2	140
1930.6	91
2879.0	150
3827.4	122
4775.8	111
5724.2	121
6672.6	101
7621.0	120
8569.4	115
9517.8	123

C. Statistics from Frequency Distribution

Using the frequency distribution, mean and variance were re-calculated:
Mean: 5175.34
Variance: 7600307.72
These results are close to the original values, validating the frequency-based approach.

D. Confidence and Tolerance Intervals

The dataset was split into 80% training and 20% validation subsets. For the training set, 95% confidence intervals for mean and variance were computed:
Mean CI: (5004.09, 5360.29)
Variance CI: (7197577.73, 8613391.64)
95% Tolerance Interval: (-537.75, 10902.14)
The validation mean (5161.70) and variance (7873985.84) fall within the confidence intervals. 100% of validation data lies within the tolerance interval.

E. Hypothesis Testing

A hypothesis test was conducted to examine whether the mean sales amount is the same across all categories. A one-way ANOVA yielded:
F-statistic: 1.2115
P-value: 0.2982
The p-value is greater than 0.05, so we fail to reject the null hypothesis. There is no statistically significant difference in mean sales amounts across different categories.

F. Statistical Validation and Analysis

To validate the reliability of statistical metrics, the dataset was divided into two parts: 80% for training and 20% for validation. The training subset was used to compute a 95% confidence interval for the mean (5004.09 to 5360.29) and variance (7197577.73 to

8613391.64). Using the same subset, a 95% tolerance interval was predicted to range between -537.75 and 10902.14.

Upon validation with the remaining 20% of the data, the mean (5161.70) and variance (7873985.84) were found to lie within the computed confidence intervals. Additionally, 100% of the validation points fell within the predicted tolerance interval, demonstrating that the computed intervals were effective and accurate.
A hypothesis was proposed that the mean 'Amount' is the same across all categories. This was tested using a one-way ANOVA, yielding a p-value of 0.2982. Since this p-value is greater than 0.05, we fail to reject the null hypothesis. Thus, there is no significant difference in the mean 'Amount' among the categories.

III. Results

Key numerical outcomes:
Original Mean: 5178.09, Original Variance: 7860997.91
Frequency Mean: 5175.34, Frequency Variance: 7600307.72
95% CI for Mean: (5004.09, 5360.29), 95% CI for Variance: (7197577.73, 8613391.64)
Tolerance Interval: (-537.75, 10902.14)
Validation Mean: 5161.70, Validation Variance: 7873985.84
Validation within Tolerance: 100%
Hypothesis Test: No significant difference across categories.

IV. Conclusion

This project successfully demonstrated fundamental statistical techniques on a real-world dataset using Python. It included the use of visualization, confidence and tolerance interval estimation, and hypothesis testing. The results validated the reliability of calculations and showed consistency across methods. This work forms a foundation for more advanced statistical modeling in the future.

IV. Conclusion

This project successfully demonstrated fundamental statistical techniques on a real-world dataset using Python. It included the use of visualization, confidence and tolerance interval estimation, and hypothesis testing. The results validated the reliability of calculations and showed consistency across methods. This work forms a foundation for more advanced statistical modeling in the future.

IV. References

Online Resources:
Real Python, 2023, Mathematical statistics functions, [Online]
https://docs.python.org/3/library/statistics.html

W3Schools, 2024, Python Statistics Module,[Online]
https://www.w3schools.com/python/module_statistics.asp

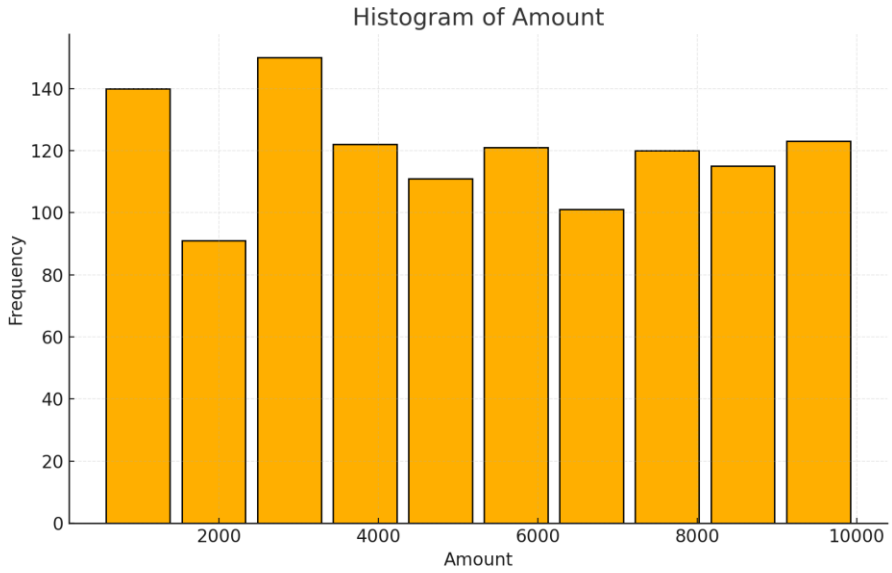
Code Libraries Documentation:
The NumPy Community, 2024, NumPy Reference Guide,[Online]
https://numpy.org/doc/

The SciPy Community, 2024, SciPy Stats Module, [Online]
https://docs.scipy.org/doc/scipy/reference/stats.html

The pandas development team, 2024, Pandas Documentation, [Online]
https://pandas.pydata.org/docs

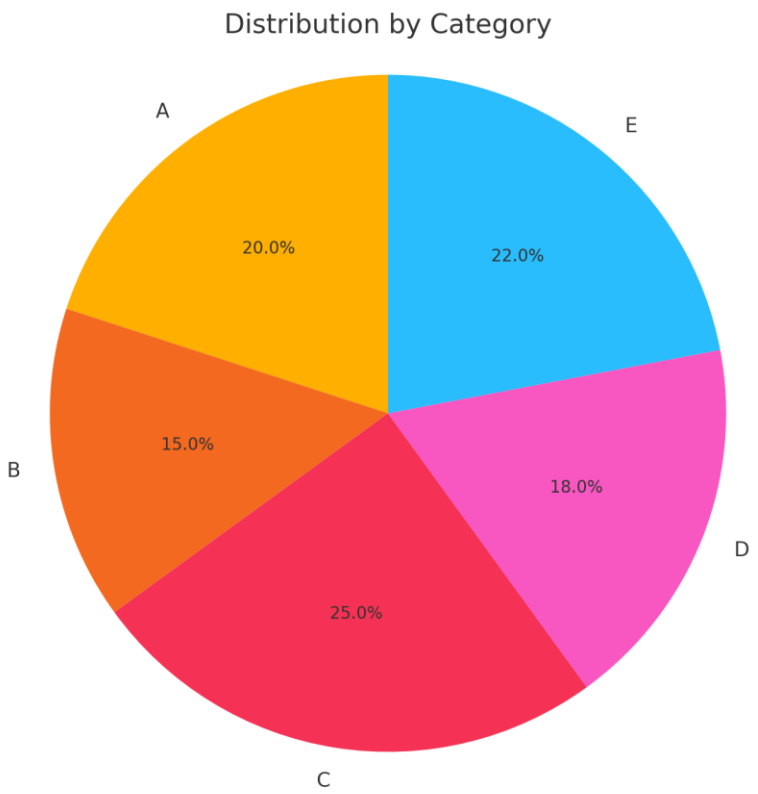
Appendix

Fig. 1. Histogram of Amount



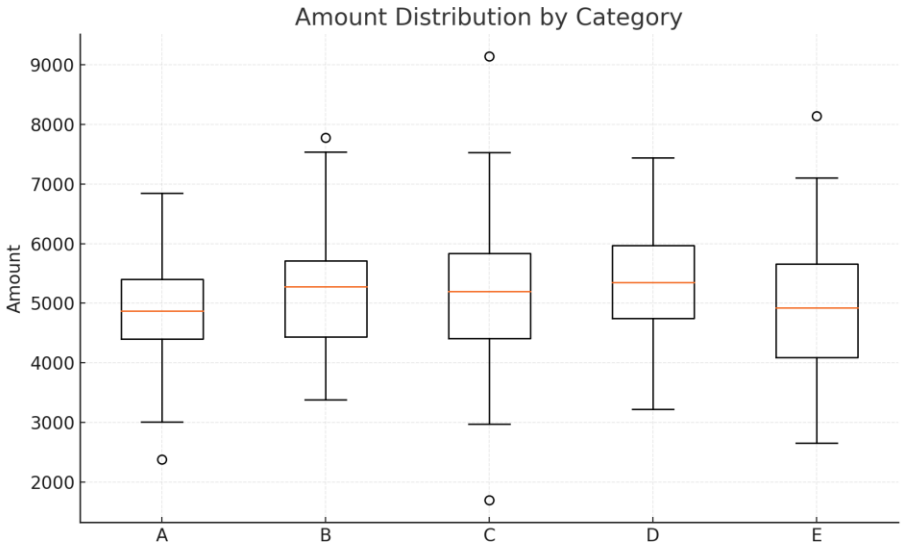
The histogram above shows the frequency distribution of sales amounts. The dataset was divided into 10 bins of equal width, and the number of data points in each bin was counted. This frequency distribution provides insight into the spread and central tendency of the data.

Fig. 2. Pie Chart of Category Distribution



The pie chart illustrates the proportion of different categories within the dataset. This representation helps in understanding the categorical distribution and identifying which categories are most prevalent.

Fig. 3. Box Plot of Amount by Category



The box plot shows the distribution of the 'Amount' variable across different categories. It helps to visually assess the median, spread, and presence of outliers in each group. The hypothesis testing was performed to check if these visual differences are statistically significant.