Statistical Analysis Report

S. Palis

Udacity Master's Degree in AI Capstone Conduct A Statistical Analysis Using Python

February 8, 2026

GitHub repository:

https://github.com/MinervaRose/conduct-a-statistical-analysis-using-python

Overview

This report presents a statistical analysis of the Space Mission Launche**s** dataset from Kaggle, which contains historical records of orbital and suborbital launch missions across multiple organizations and countries. The goal was to explore launch-cost variability and mission outcomes to support an aerospace "launch risk analysis" perspective, using descriptive statistics, visualization, and hypothesis testing.

Dataset Description

The dataset includes 4,324 rows (missions) and 9 original columns. Variables used in this analysis include Mission_Status (categorical outcome such as Success/Failure), Organisation (launch provider), Date (launch time stamp), and Price (numeric launch cost when available). The dataset also includes contextual text fields such as Location and Detail, which provide descriptive mission information. Because the Price column contains a substantial amount of missing data, all cost-based analyses were performed using only missions with recorded prices.

Methods

Initial data screening and reproducibility (IDA). Before hypothesis testing, the workflow included structured data screening—checking data types, identifying missingness, and validating key variable distributions—to ensure the dataset was appropriate for inference and to document decisions transparently. This structured data-screening workflow follows recommendations that systematic initial data analysis improves transparency and reproducibility in statistical research (Lusa et al., 2024).

Descriptive statistics. Summary statistics (count, mean, median, quartiles, and spread) were computed to describe the distribution of launch prices and to identify variability and skew. Categorical frequency tables were computed for mission outcomes and organizations to characterize class imbalance and launch concentration.

Visualizations. Five plots were created to support the launch risk narrative:

1. Histogram of Price to examine cost variability and skew.
2. Boxplot of Price by Mission_Status to compare cost distributions across outcomes.
3. Count plot of Mission_Status to quantify outcome imbalance.
4. Bar chart of top organizations by mission count to show concentration of launch activity.
5. Scatter plot of Price over time to explore whether launch cost changes systematically across years.

Hypothesis test selection**.** The primary inferential question was whether average launch price differs between successful and failed missions (Success vs Failure). Because this is a comparison of means between two independent groups with unequal sample sizes and potentially unequal variances, Welch's two-sample *t*-test was used (implemented via SciPy using `equal_var=False`). Welch's two-sample *t*-test was selected because it is more robust when group variances and sample sizes are unequal, which is a common condition in observational datasets (Ruxton, 2006).

Assumptions. The Welch t-test assumes independent observations within groups and that the sampling distribution of the mean difference is approximately normal (often reasonable with moderate sample sizes). Because the dataset contains many missing prices, the analysis also assumes that the recorded Price values are not systematically missing in a way that would invalidate comparisons (a limitation discussed below). Transparent reporting of missing data is essential because incomplete observations can bias inference when missingness is systematic (Lee et al., 2021).

Results

Descriptive statistics show that recorded launch prices vary widely, indicating substantial economic variability across missions. The histogram in Figure 1 shows a strongly right-skewed distribution, with most missions clustered at lower price levels and a small number of extremely high-cost launches.

Mission outcome counts reveal that successful missions dominate the dataset. This imbalance is clearly visible in Figure 3, where successes vastly outnumber failures. The boxplot in Figure 2 compares launch prices across mission outcomes and shows wide variability within each category.

Launch activity is highly concentrated among a small number of organizations. As shown in Figure 4, a few agencies account for a disproportionately large share of missions, suggesting uneven distribution of operational risk.

The scatter plot in Figure 5 illustrates substantial price variation over time without a simple linear trend. Launch costs fluctuate considerably across years, indicating that pricing is influenced by mission-specific factors rather than steady time-based growth.

The Welch *t*-test comparing successful and failed missions produced a test statistic of 3.77 and a *p*-value of 0.00052, indicating a statistically measurable difference in average launch price between the two groups.

Figures referenced correspond to visualizations presented in the accompanying Jupyter notebook.

Interpretation for a Non-Technical Audience

This dataset shows that most rocket launches recorded were successful, and only a small fraction failed. Launch prices, when reported, range from relatively low-cost missions to extremely expensive ones, so “typical” mission costs vary dramatically. When we compared prices for successful missions to prices for failed missions, we found evidence that the average price differs between the two groups. This does not mean that spending more guarantees success, but it suggests that mission cost and mission outcome are not completely unrelated in the historical data.

Limitations and Potential Bias

A major limitation is missing pricing data: most missions have no recorded Price, so the cost-based analysis reflects only the subset with available pricing. If the presence or absence of pricing information is related to organization, era, mission type, or outcome, results may not generalize to all launches. Missing data can introduce bias when the absence of observations is related to the variables under study, making transparent documentation of missingness a statistical best practice (Lee et al., 2021).

Another limitation is class imbalance (many more successes than failures). The failure group is much smaller, which can increase uncertainty about failure-related estimates even when a test is statistically significant. Finally, the dataset is observational, not experimental: many confounding factors (vehicle type, payload complexity, era, reporting practices, etc.) could influence both Price and Mission_Status. Therefore, results should be interpreted as associations, not causal effects.

References

Lusa, L., Proust-Lima, C., Schmidt, C. O., Lee, K. J., le Cessie, S., Baillie, M., … Huebner, M. (2024). *Initial data analysis for longitudinal studies to build a solid foundation for reproducible analysis*. PLOS ONE, 19(5), e0295726. https://doi.org/10.1371/journal.pone.0295726

Lee, K. J., Tilling, K. M., Cornish, R. P., Little, R. J. A., Bell, M. L., Goetghebeur, E., … Carpenter, J. R. (2021). *Framework for the treatment and reporting of missing data in observational studies: The TARMOS framework*. Journal of Clinical Epidemiology, 134, 79–88. https://doi.org/10.1016/j.jclinepi.2021.01.008

Ruxton, G. D. (2006). *The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test*. Behavioral Ecology, 17(4), 688–690. https://doi.org/10.1093/beheco/ark016

Sefercanapaydn. (n.d.). *Mission launches* [Data set]. Kaggle.

https://www.kaggle.com/datasets/sefercanapaydn/mission-launches