Exploratory Data Analysis

This section explores the various features contained within the dataset, each representing different aspects of the rocket launches recorded between 1957 and 2020. Understanding these features is crucial for interpreting the results of the analysis and for building accurate predictive models.

Exploring Original Data Frame (15 Columns)

Analysis of Status Mission

Figure 1 highlights a clear dominance of successful missions, making up 89.1% of total launches, while failures, partial failures, and prelaunch failures only represent 7.8%, 2.3%, and 0.1%, respectively. This significant imbalance could lead predictive models to favor predicting success, thus impacting the accuracy for less common outcomes. To address this, it would be practical to switch the target variable from a multi-class to a binary classification, grouping all non-success results under a single 'Failure' category. This simplification could improve the model's effectiveness. Moreover, employing balancing techniques like oversampling or using RandomOverSampler would help create a more balanced dataset, ensuring a robust and accurate predictive model.

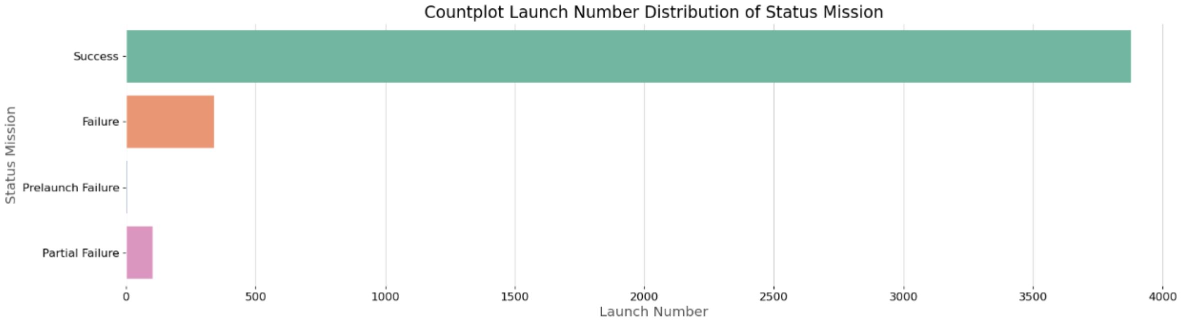

Figure 1

Analysis of Company Distribution

Figure 2 showcases the launch frequencies by 55 unique aerospace companies. RVSN USSR leads with 1,777 launches, followed by Arianespace, CASIC, and General Dynamics with 279, 256, and 251 launches respectively. Notably, newer companies like SpaceX quickly reached 100 launches, indicating shifts in industry dynamics. This data not only highlights varying capabilities and resources but also suggests correlations between launch frequency and technological advancements. To address the high cardinality issue presented by the 55 unique companies, grouping them by continents or hemispheres could simplify the dataset. This approach will streamline the analysis, allowing us to focus on regional trends and enhance predictive modeling effectiveness.

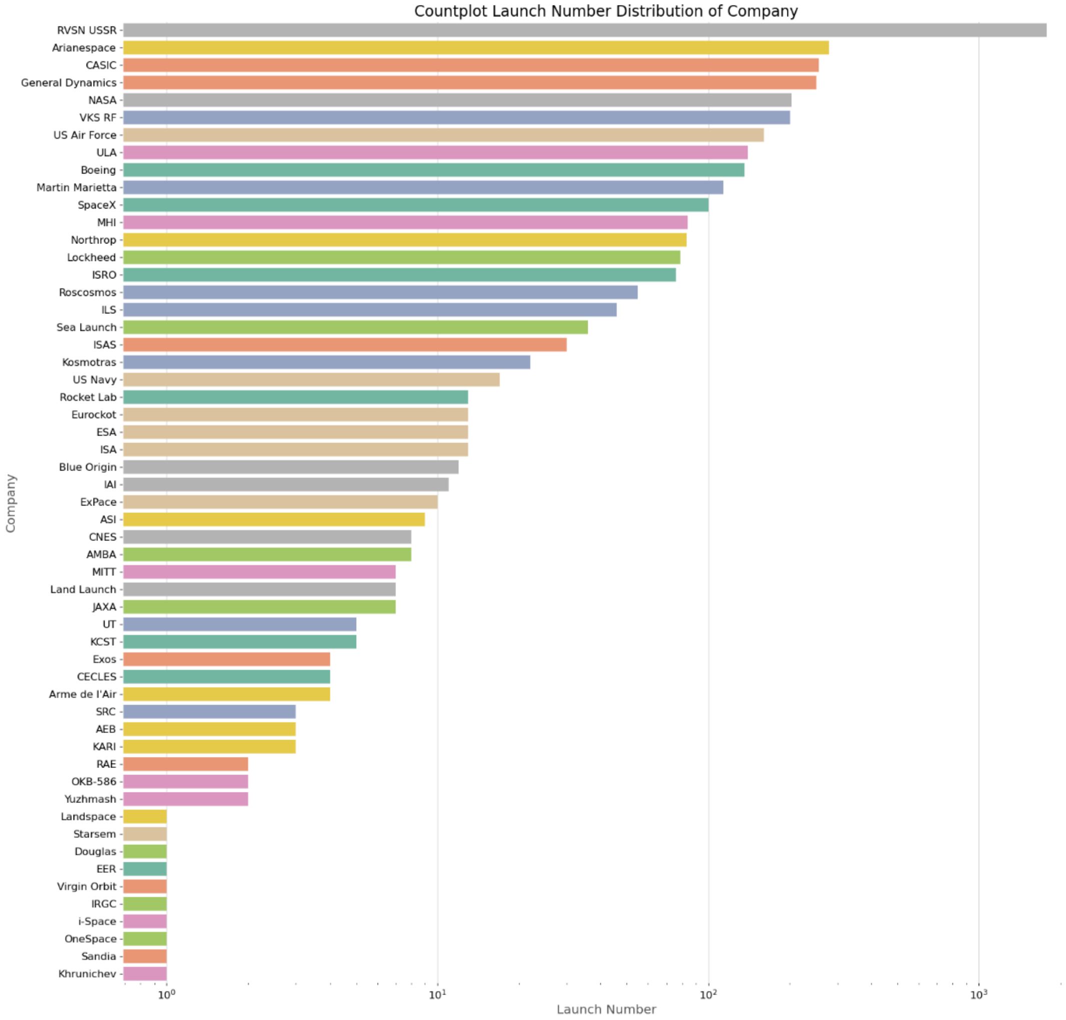

Figure 2

Analysis of Location

The "Location" feature in the dataset, characterized by 137 unique entries, presents complexities that render straightforward analysis ineffective due to the need for advanced natural language processing (NLP). The data suggest varying formats of entries, split by one, two, or three commas, each implying a different level of geographical detail. To address these inconsistencies, it is proposed that the "Location" string be decomposed into four distinct components: Pad, Center, State, and Country. This decomposition will facilitate a more structured analysis by appropriately categorizing geographical details based on the comma count, potentially revealing new geographical patterns not initially apparent. This methodological adjustment transforms a complex variable into more analyzable components, enabling further exploration with tools suited for geospatial analysis or more refined NLP techniques. This approach underscores the adaptability necessary in data science, demonstrating how redefining data representation can uncover significant trends and influence the broader understanding of spatial factors affecting launch outcomes.

Analysis of Status Rocket

Figure 3 shows 3,534 launches with retired rockets and 790 with active rockets. This significant disparity suggests that the majority of launches historically have used rockets that are now retired. Analyzing the reasons behind the prevalence of retired rockets could provide insights into rocket lifecycle, technological advancements, and changes in safety standards. Such data are vital for understanding trends in rocket usage and informing future rocket design and launch strategies. This disparity underscores the evolving nature of aerospace technology.

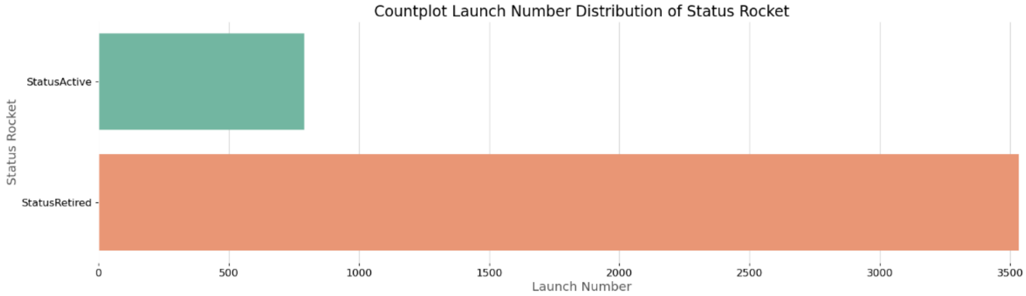

Figure 3

Analysis of Rocket Cost

Figure 4's boxplot, shown on a log scale, doesn't fully capture the distribution's outliers, crucial for understanding the full range of rocket cost extremes. Significant outliers include rocket costs at $450 million (136 launches), $1.16 billion (13 launches), and an exceptional $5 billion (2 launches). These outliers represent substantial investments well above the typical range, with a recalculated practical minimum of $5.3 million and a maximum at $350 million, as determined by the IQR.

Figure 5's histogram highlights a pronounced peak at $50 million, indicative of a common expenditure for many launches, and a secondary peak at $450 million. This aligns with one of the outlier groups from Figure 4, suggesting a specific tier of mission complexity requiring significantly higher funding.

This analysis provides a clear depiction of the financial landscape within aerospace ventures, emphasizing both typical mission costs and exceptional financial commitments for more ambitious projects.

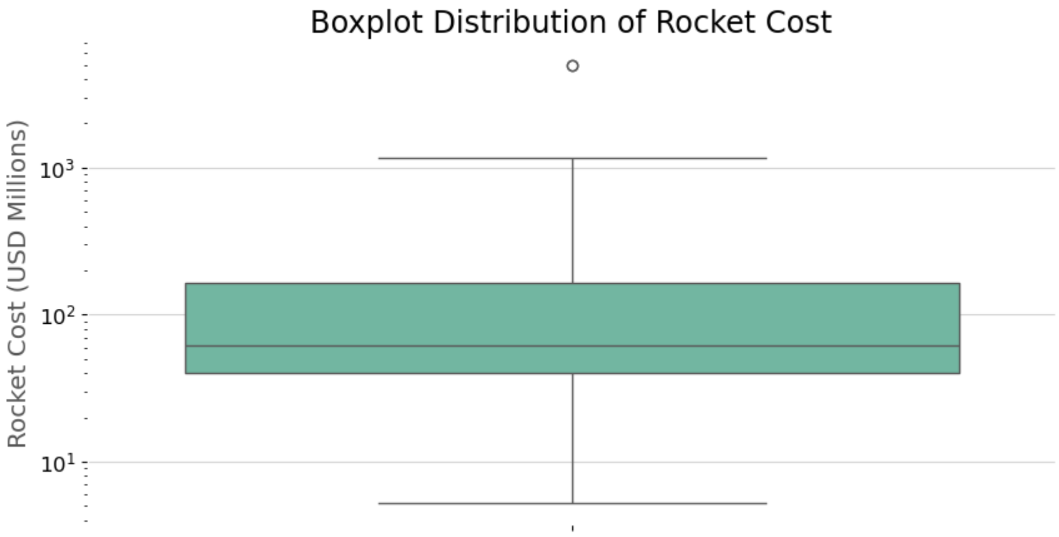

Figure 4

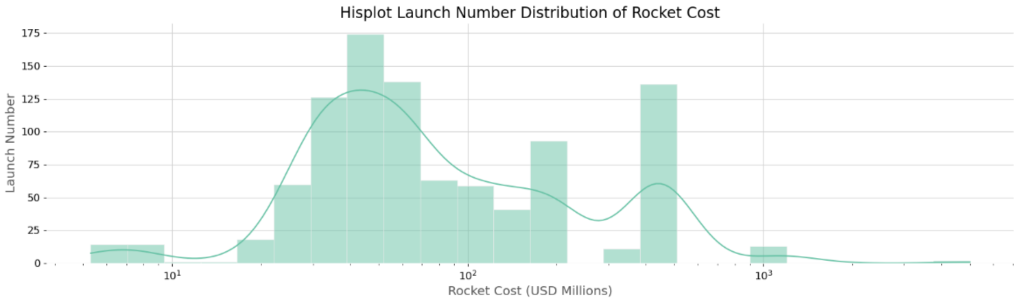

Figure 5

Analysis of Launch Country

Figure 6 shows a count plot of launches by company origin, highlighting Russia as the dominant player with 2,064 launches, followed by the USA with 1,374 and China with 269. This distribution emphasizes the extensive involvement of these countries in global aerospace activities, reflecting their longstanding commitment and leading roles in space exploration. The data suggests a concentration of aerospace capabilities and resources within a few key nations, underlining their strategic importance in the space sector.

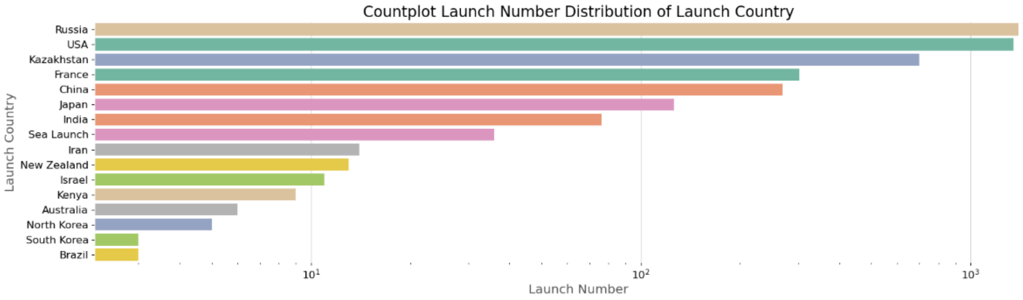

Figure 6

Analysis of Company Origin

Figure 7 presents a count plot of launches by country, where Russia and the USA are nearly tied at the top with 1,398 and 1,351 launches, respectively. The significant count of 701 launches from Kazakhstan is primarily due to the operations at the Baikonur Cosmodrome, a major launch facility used historically by Russia. This plot illustrates the impact of geographic and geopolitical factors on launch activities, showing how certain locations become pivotal due to their infrastructure and historical significance in space launch history.

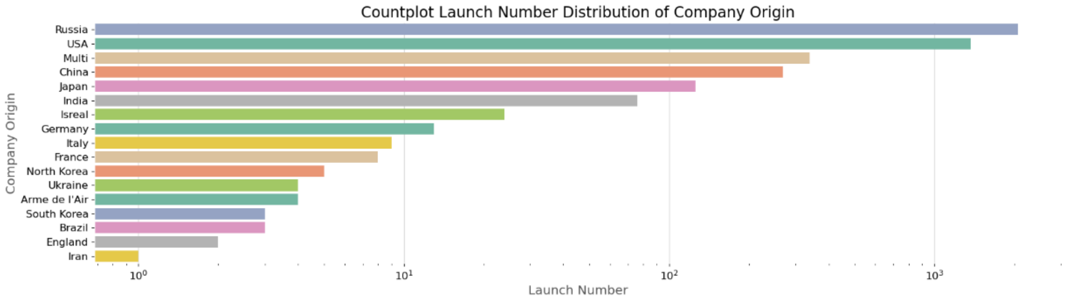

Figure 7

Analysis of Ownership

Figure 8, depicting the distribution of rocket launches based on ownership, shows state-owned entities leading with 2,931 launches compared to 1,393 by private companies. This highlights the significant role of government-backed programs in space exploration, traditionally supported by extensive resources and infrastructure. However, the notable presence of private sector launches illustrates their growing influence and the shifting dynamics within the aerospace industry, signaling an increasing trend towards privatization and innovation in space technology. This trend underscores the evolving partnership and competition between public and private entities in the space race.

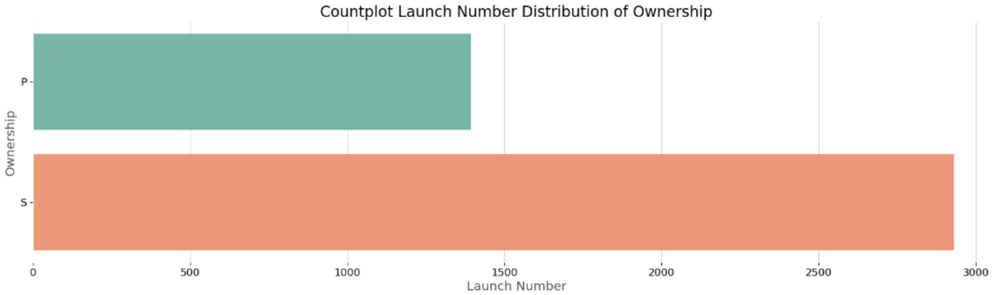

Figure 8

Analysis of Year

Figure 9 charts the rocket launch activities from 1957 to 2020, highlighting significant peaks in the mid-1970s and 2018 with 119 and 117 launches, respectively. These peaks reflect periods of heightened space exploration efforts, possibly tied to major technological advancements, or increases in space mission funding. The graph also illustrates dips during periods like the early 2000s suggesting times when space exploration may have been impacted by economic downturns or shifts in space policy. This visual representation helps to identify key trends in space activity over the decades, providing a clear view of how historical, technological, and political factors have influenced launch frequencies.

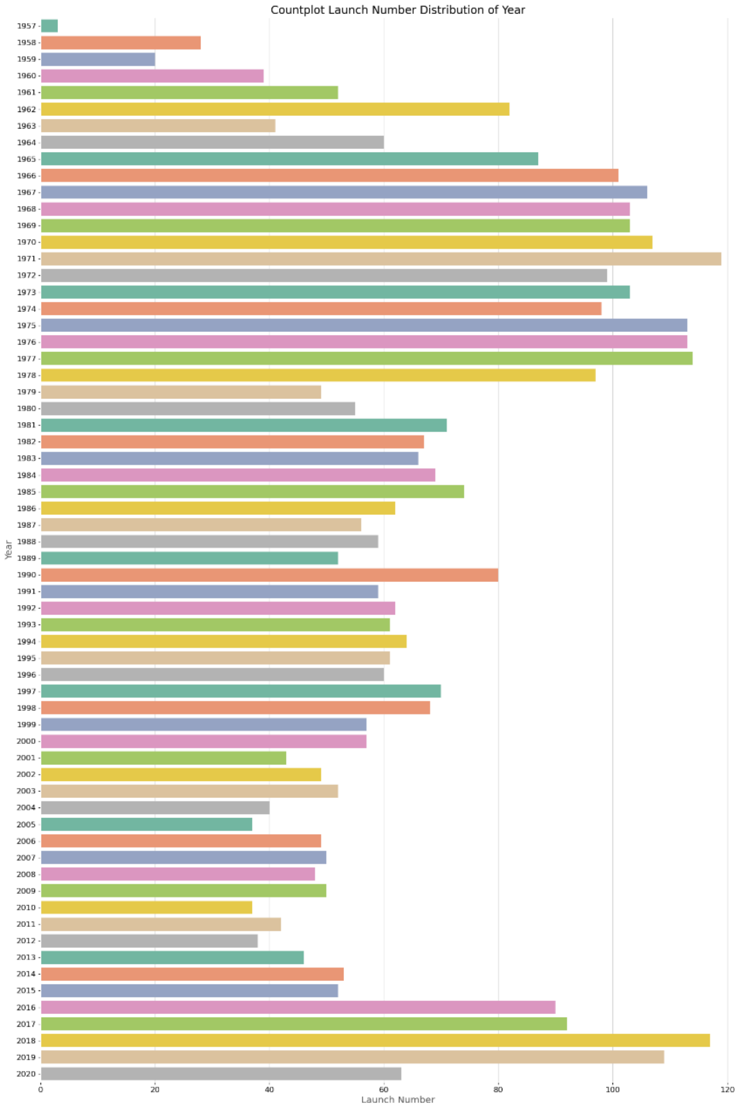

Figure 9

Analysis of Month

Figure 10 reveals the monthly distribution of rocket launches, showing December as the most active month with 450 launches, followed closely by June with 402 launches. This pattern suggests a seasonal trend in launch activities, possibly influenced by end-of-year budget cycles and favorable weather conditions in key launch locations during these months. The lowest number of launches occurs in January with 268, which may reflect operational slowdowns after the peak activity in December. The graph provides insights into how operational, environmental, and financial factors align to influence the timing of launches throughout the year, offering a nuanced understanding of the strategic planning in the aerospace sector.

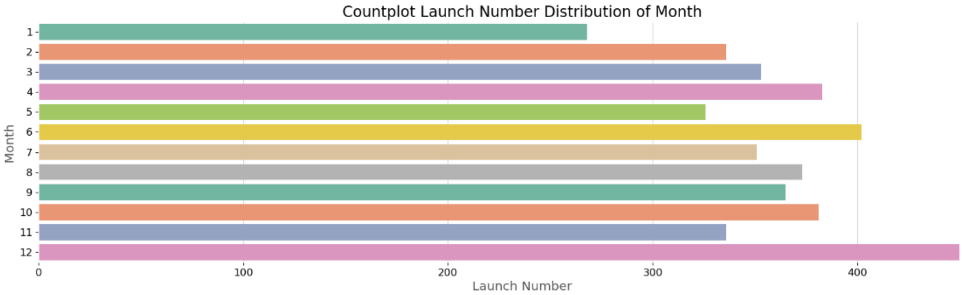

Figure 10

Analysis of Day

Figure 11 illustrates the distribution of rocket launches by day of the month, revealing a relatively even spread across most days with a slight increase towards the end of the month. The most launches occur on the 28th with 187 launches, followed by the 25th and 24th with 175 and 173 launches, respectively. The lowest numbers are seen on the 31st, with only 66 launches, likely due to fewer months containing this day. Overall, the distribution does not show significant variations or distinct patterns that might suggest strategic preferences for specific launch days, indicating that day-to-day timing decisions are likely influenced more by operational readiness and external scheduling factors than by any date preference.

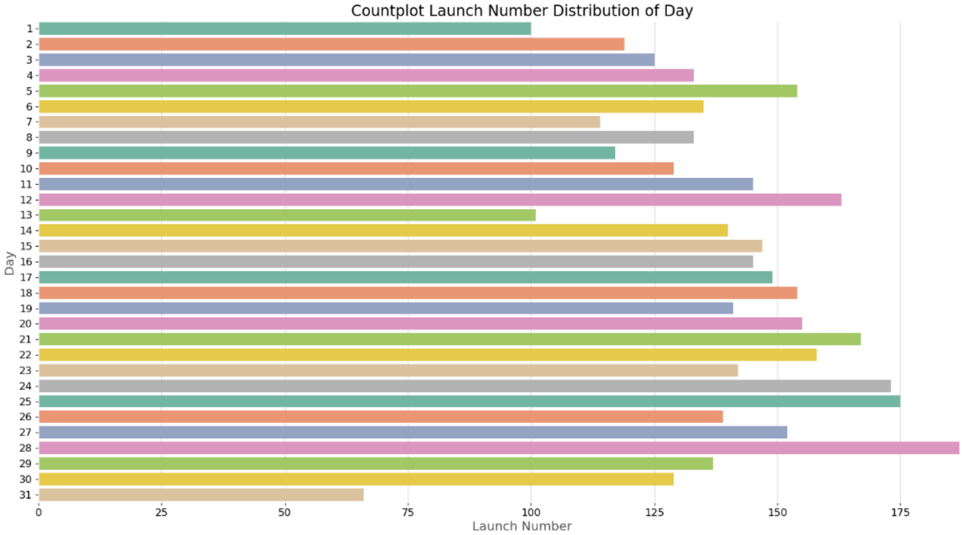

Figure 11

Analysis of DateTime

This feature has 4,319 unique timestamps for each rocket launch. Its detailed precision makes simple plotting unfeasible. A practical approach is to break down this timestamp into year, month, and day to better capture and analyze trends over time.

Analysis of Date

With 3,922 unique entries, the Date features similarly tracks when each launch occurred. To make sense of this vast data, grouping it by year or even by month could help highlight key periods of activity and historical trends in space launches.

Analysis of Time

Capturing the exact time of each launch with 1,273 unique times, this data might reveal preferred launch windows. Analyzing it could tell us if certain times of the day are favored for launches, possibly due to logistical reasons or specific orbital dynamics.

Analysis of Detail

This feature is a text-based description with 4,278 unique entries, making standard visual analysis challenging. Applying text analysis techniques like keyword extraction could unveil common themes or frequently mentioned details about the launches, such as payload types or rocket models.

Exploring New Features

Analysis of Launch Country and Company Origin Latitude/Longitude

Figures 12 and 13 show scatterplots of geographical distribution for company origins and launch countries, highlighting significant launch activity clusters in the US and Russia, primarily located in the northern and eastern hemispheres. Reference Figure 14, a global hemisphere map, contextualizes these findings, showing that while most launches occur close to their origin country, some from the southern and eastern hemispheres launch from distant sites due to logistical and geopolitical factors.

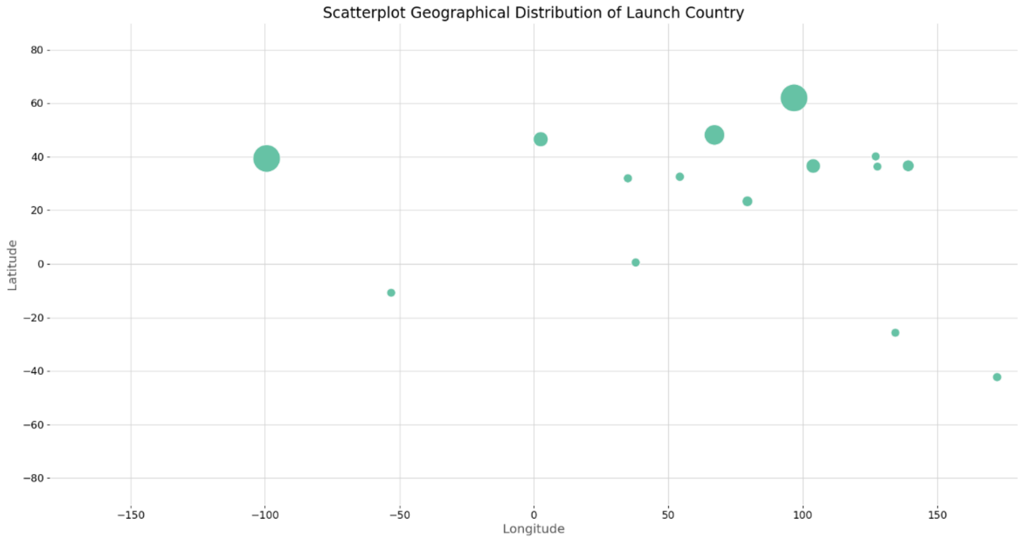

Figure 12

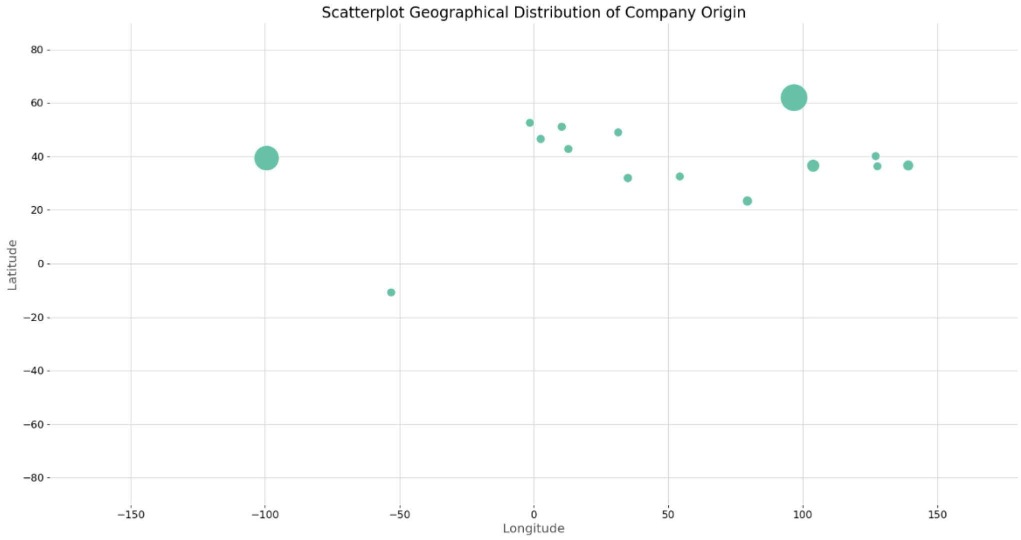

Figure 13

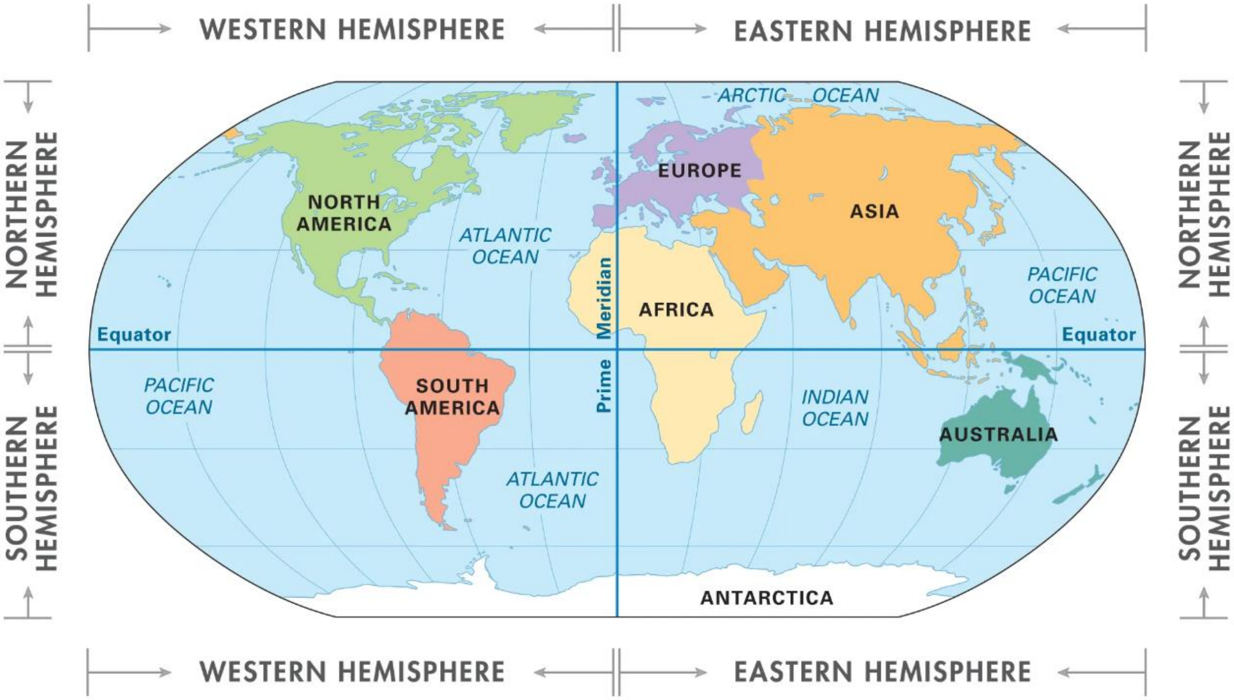
WESTERN HEMISPHERE
EASTERN HEMISPHERE
NORTHERN HEMISPHERE
SOUTHERN HEMISPHERE
ARCTIC OCEAN
EUROPE
ASIA
NORTH AMERICA
ATLANTIC OCEAN
PACIFIC OCEAN
AFRICA
Prime Meridian
Equator
Equator
PACIFIC OCEAN
SOUTH AMERICA
INDIAN OCEAN
AUSTRALIA
ATLANTIC OCEAN
ANTARCTICA
NORTHERN HEMISPHERE
SOUTHERN HEMISPHERE
WESTERN HEMISPHERE
EASTERN HEMISPHERE

Figure 14

Analysis of Hour, Day, and Year Cosine/Sine

The "Hour" feature in the dataset, along with "Day" and "Year", has been transformed into cosine and sine values to represent time cyclically. This conversion is essential for accurately capturing the periodic nature of these time-related features. By plotting "Hour" on a unit circle, the correctness of this transformation is verified, ensuring that the data points are uniformly distributed and confirming the integrity of the temporal patterns. Although similar transformations were applied to "Day" and "Year", only the "Hour" visualization is displayed in Figure 15, efficiently demonstrating the effectiveness of this method in exploratory data analysis.

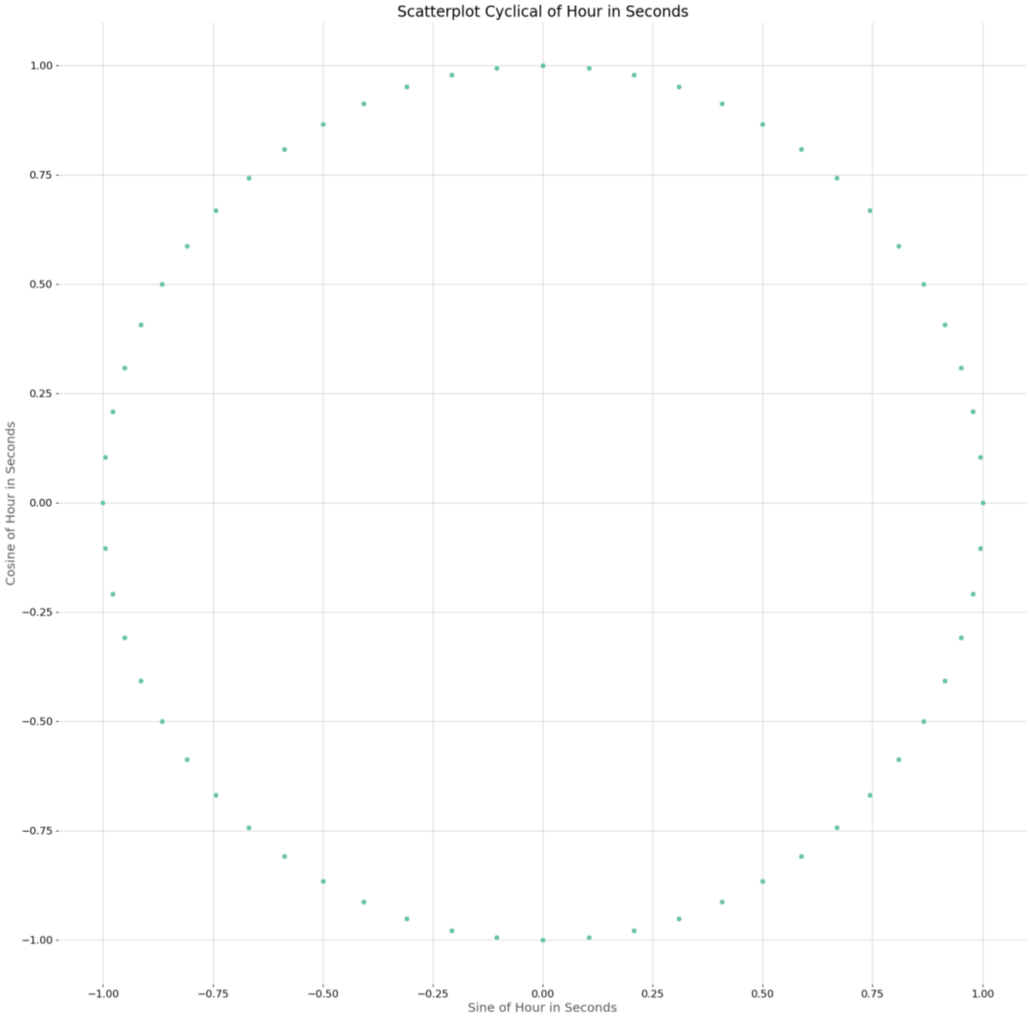

Figure 15

Analysis of Center, Pad, State and Country

The "Location" feature has been broken down into Center, Pad, State, and Country for focused analysis. The Center feature, with its 42 unique values, exhibits high cardinality, suggesting a potential benefit in categorizing these into broader geographical or operational groups to clarify trends at major launch sites. The Pad, having 124 unique entries, presents a challenge for straightforward analysis, indicating a need for advanced NLP or aggregation techniques to extract useful insights. State data, heavily marked with 'missing' but prominently featuring locations like Florida and California, underscores their significance in space activities and warrants a closer look at regional distributions. The Country component has been thoroughly examined in the "Launch Country" section and requires no further analysis.

Analysis of Mission and Rocket

The "Detail" feature, containing text descriptions of launches, was broken down into "Mission" and "Rocket." Given their high cardinality, Count Vectorizing followed by Truncated Singular Value Decomposition (SVD) was used for dimensionality reduction. For "Mission," analyzed in Figure 16, over 3,000 components were initially identified, but using the elbow method, this was reduced to 150 components capturing about 0.5 of the variance ratios. This high number of components makes it impractical to use as features due to complexity. Conversely, "Rocket" data in Figure 17 was effectively reduced to 20 components capturing 0.7 of the variance ratios, which can be integrated as new features in the dataset. This analysis shows that while "Rocket" components can enhance the dataset, "Mission" requires further NLP refinement due to its complexity and the high number of components needed to capture significant variance.

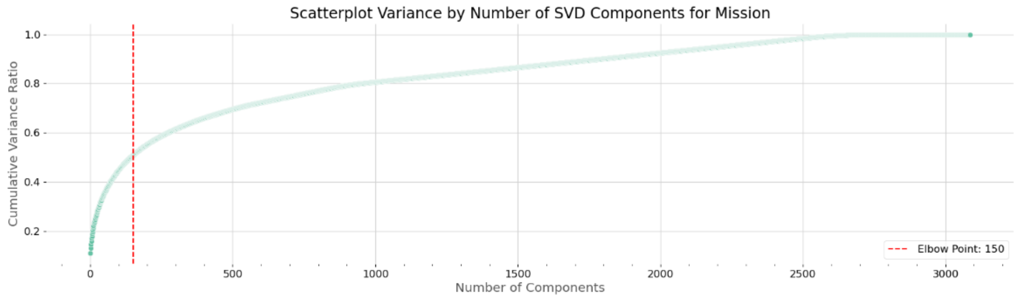

Figure 16

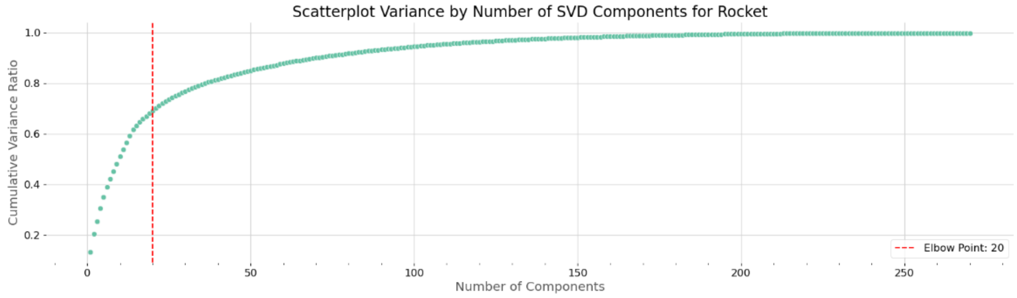

Figure 17

Exploring Features with Target Variable

Analysis of Company with Status Mission

Figure 18 illustrates the success and failure percentages of various aerospace companies using a stacked bar plot. Companies like ASI, Blue Origin, Douglas, IRGC, Khrunichev, OKB-586, Starsem, Yuzhmash, and i-Space boast a 100% success rate, while AEB, EER, Exos, Landspace, OneSpace, Sandia, and Virgin Orbit experienced a 100% failure rate. Additionally, companies such as CELEC, ISA, KARI, KCST, US Navy, and UT demonstrate relatively higher failure rates than successes. This bar plot provides a clear comparison of mission outcomes, revealing stark differences in company performance and emphasizing the varying success levels within the aerospace sector. Understanding these variations is crucial for assessing reliability, investment potential, and identifying areas for technological improvement.

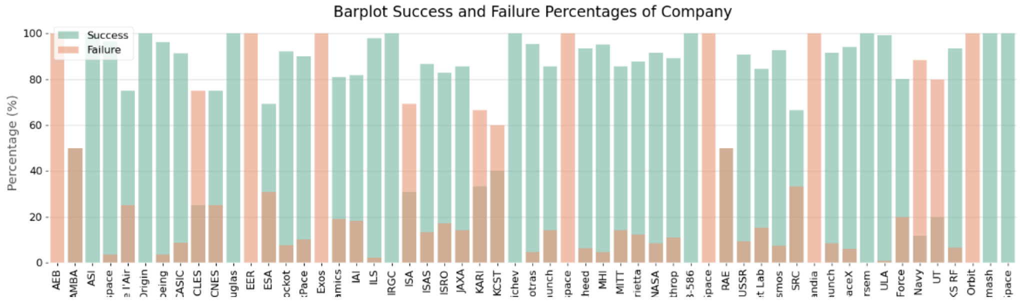

Figure 18

Analysis of Launch Country and Company Origin with Status Mission

Figure 19 shows a stacked bar plot comparing success and failure percentages across launch countries. Kenya stands out with a 100% success rate, while Brazil has a 100% failure rate. Other countries like Iran, North Korea, and South Korea exhibit higher failure rates than success, suggesting challenges in their space programs.

Figure 20, another stacked bar plot, illustrates mission success and failure percentages by company origin. Iran, Italy, and Ukraine have a perfect 100% success rate, while Brazil again registers a 100% failure rate. North Korea and South Korea display higher failure rates than successes, indicating technological or strategic challenges.

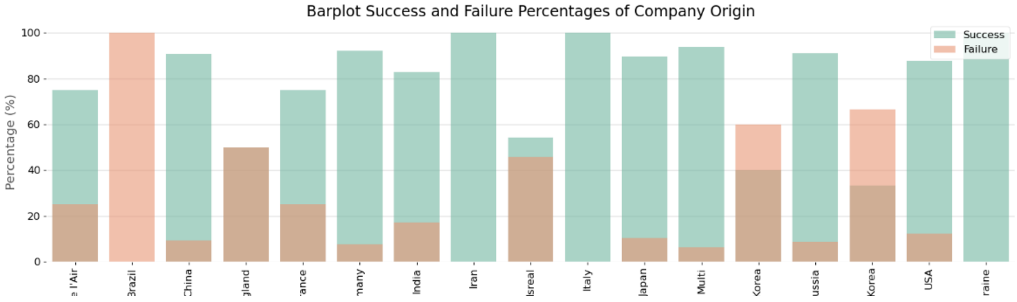

Figure 19

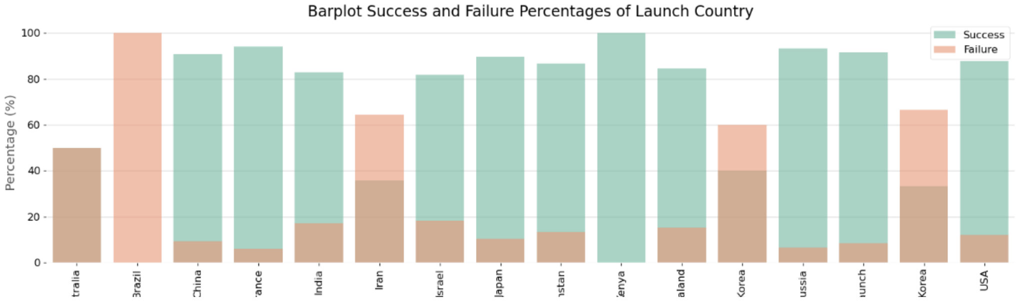

Figure 20

Analysis of Year with Status Mission

Figure 21 shows a stacked bar plot comparing success and failure percentages across different years. In 1958, the failure rate was the highest compared to success, and this trend continued in 1959. By 1960, failure and success rates were relatively even. From 1971 onward, the failure rate gradually declined, stabilizing at or below the mean failure rate of 12.11%. Simultaneously, the success rate increased, remaining at or above the mean success rate of 87.89% from 1971 to 2020. This analysis reveals a historical improvement in space mission outcomes, highlighting advancements in technology, strategic planning, and international collaboration.

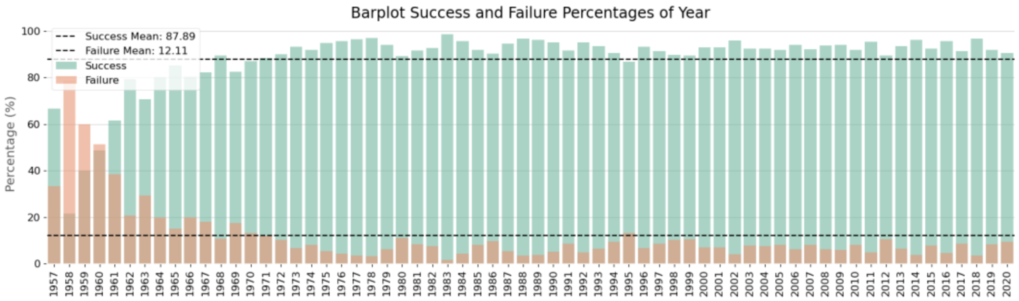

Figure 21

Analysis of Rocket Status and Ownership with Status Mission

Figure 22 uses a heatmap to illustrate the relationship between Rocket Status and Mission Status, revealing significant differences between active and retired rockets in their mission outcomes. StatusActive rockets have 54 failures and 736 successes, while StatusRetired rockets have 391 failures and 3,143 successes. This visual clearly shows that retired rockets are predominantly linked with higher failure rates, while active rockets exhibit higher success rates.

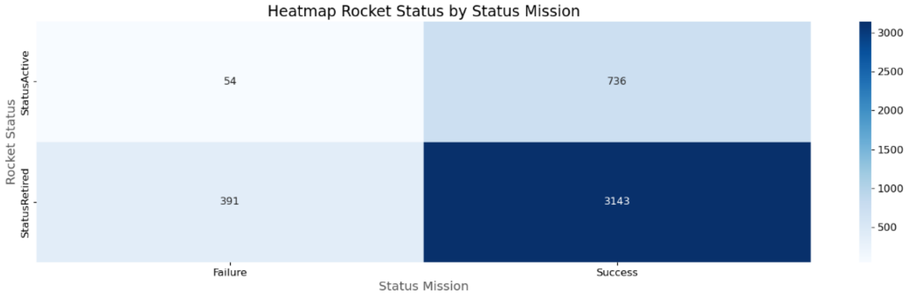

Figure 22

Figure 23, another heatmap, examines the relationship between Ownership and Mission Status. Private ownership (P) is associated with 118 failures and 1,275 successes, whereas state ownership (S) has 327 failures and 2,604 successes. The gradient illustrates that while both ownership types have higher success rates than failures, state-owned missions have a slightly higher success-to-failure ratio compared to private missions.

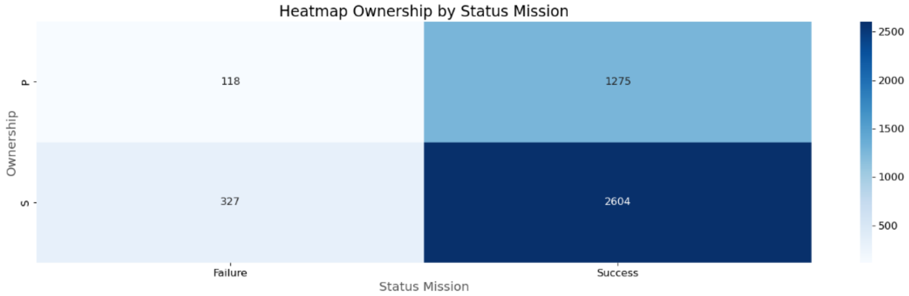

Figure 23

Analysis of Rocket Cost with Status Mission

Figure 24's boxplot, shown on a log scale, doesn't fully capture the distribution's outliers, crucial for understanding the full range of rocket cost extremes. For successful missions, significant outliers include rocket costs at $450 million (134 launches), $1.16 billion (12 launches), and an exceptional $5 billion (2 launches). The practical minimum is nearly zero, with a recalculated maximum at $440 million, as determined by the IQR. For failed missions, notable outliers include rocket costs at $200 million (2 launches), $450 million (2 launches), $350 million (1 launch), $136.6 million (1 launch), and $1.16 billion (1 launch). The recalculated practical minimum is nearly zero, and the maximum is around $113.36 million. This analysis underscores that even high-cost missions are not immune to failure, highlighting the importance of thorough risk assessment in space missions.

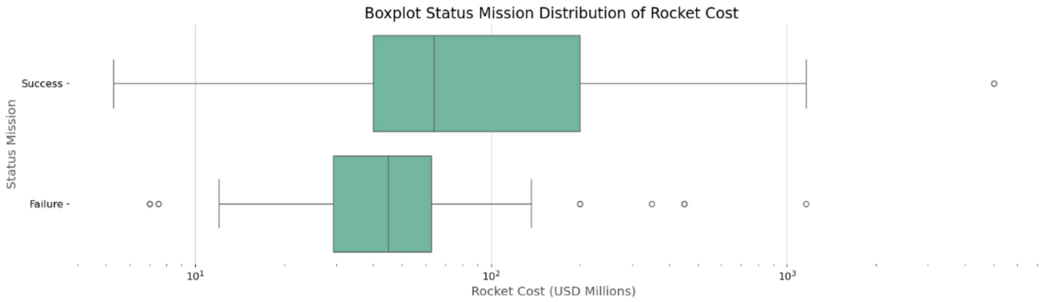

Figure 24

Analysis of States with Status Mission

Figure 25's bar plot illustrates the distribution of mission success and failure across states, measured by percentage. Alaska and Texas stand out with 100% success rates, while Hawaii, Mahanhao, and New Mexico exhibit 100% failure rates. Notably, no state has a higher failure rate than success, and most launch states show a higher proportion of successful missions. These trends are not confined to the U.S., as other launch states worldwide also display higher success rates.

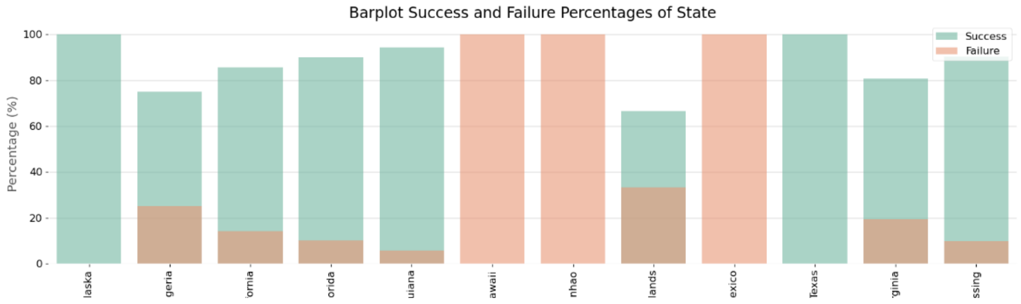

Figure 25

Analysis of Unix Time with Status Mission

Figure 26 shows a histogram plot of mission status (success and failure) over Unix Time. The x-axis represents Unix Time in seconds, ranging from -0.375e9 (1950) to 1.50e9 (2020), and the y-axis shows the number of launches. KDE lines distinguish between successful and failed missions. Successful missions peak around 1970s but decline until a secondary peak at 2020s. Meanwhile, failed missions initially peak around 1950s to 1970s, with about 50 failed launches per year, and then steadily decline to below 10 failures per year by the 2020s. Despite the consistent number of failures from the 1970s to the early 2000s, the decreasing number of successful launches resulted in a reduced success-to-failure ratio. However, this ratio improves in the 2020s as successful launches rise and failures continue to decrease. Figure 26 provides a clearer picture than Figure 9 by highlighting both successful and failed launches over time, emphasizing the importance of strategic planning and risk assessment in improving the success-to-failure ratio for space missions.

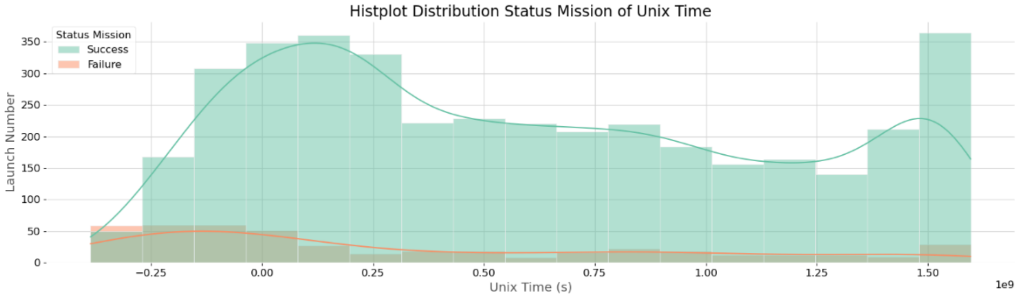

Figure 26