Winning Space Race with Data Science

<Haotian Gu>
<28/07/2023>

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- **Summary of Methodologies:**
- Data Collection: The data was collected using web scraping techniques and the SpaceX API, which allowed us to gather valuable information from public sources.
- Exploratory Data Analysis (EDA): We performed data wrangling, data visualization, and interactive visual analytics to gain insights into the dataset. EDA helped identify the key features that are most relevant for predicting the success of launchings.
- **Summary of All Results:**
- Data Collection: The data collected from web scraping and the SpaceX API provided us with a comprehensive dataset to analyze.
- EDA: Through data visualization and analysis, we discovered which features have the most significant impact on the success of launchings. This information is crucial for making informed decisions.
- Machine Learning Prediction: Using machine learning models, we were able to predict the characteristics that are essential for driving successful launchings based on the collected data.
- Overall, the combination of data collection, EDA, and machine learning prediction proved to be effective in analyzing and predicting the success of launchings. The insights gained from these methodologies can be valuable for future space missions and decision-making processes.

Introduction

- The objective of this study is to evaluate the viability of the new company Space Y and assess its competitiveness with the established Space X in the space industry. In order to achieve this, we aim to address two key aspects:
- Estimating Total Launch Costs: One of the crucial factors for the success of any space company is the ability to estimate the total cost for launches accurately. To achieve this, we will focus on predicting the success of landings for the first stage of rockets. By understanding the factors that contribute to successful landings, we can make informed predictions about the total cost of launches and assess the cost-effectiveness of Space Y's operations.
- Identifying Optimal Launch Sites: Another critical aspect that determines the competitiveness of a space company is the selection of the best launch sites. We will conduct a comprehensive analysis to identify the most suitable locations for making launches. This analysis will consider various factors such as geographical conditions, regulatory environment, and accessibility to space orbits, among others.
- By pinpointing the best launch sites, Space Y can optimize its operations and enhance its competitive edge in the industry.By conducting a detailed evaluation of these two aspects, we aim to provide valuable insights to Space Y, enabling them to make strategic decisions and position themselves competitively against Space X. The findings of this study will not only help Space Y in its business operations but also contribute to advancements in the broader space industry.

Section 1

Methodology

Methodology

Executive Summary

- Data Collection:Data from Space X was collected from two sources:
- a. Space X API (https://api.spacexdata.com/v4/rockets/): This API provided valuable information about Space X rockets, including various features and outcomes.
- b. Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches): Web scraping was performed to gather additional data on Falcon 9 and Falcon Heavy launches from Wikipedia.

- Data Wrangling:The collected data was processed and cleaned to ensure consistency and accuracy. Missing values were handled, and duplicates were removed to create a clean dataset for analysis.Enriching Data:The collected data was enriched by creating a landing outcome label based on outcome data. This label indicated whether the landing of the first stage of rockets was successful or not.
- Exploratory Data Analysis (EDA):EDA was conducted to gain insights into the data through visualization and SQL queries. Various visualizations were created to understand patterns, trends, and relationships between different features.
- Interactive Visual Analytics:Folium and Plotly Dash were utilized to perform interactive visual analytics. Folium was used to create interactive maps, and Plotly Dash allowed for interactive data visualization and exploration.
- Predictive Analysis:Predictive analysis was conducted using classification models. The collected data was normalized, and a training dataset was created for model training. Four different classification models were evaluated, and their accuracy was measured using different combinations of parameters.

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

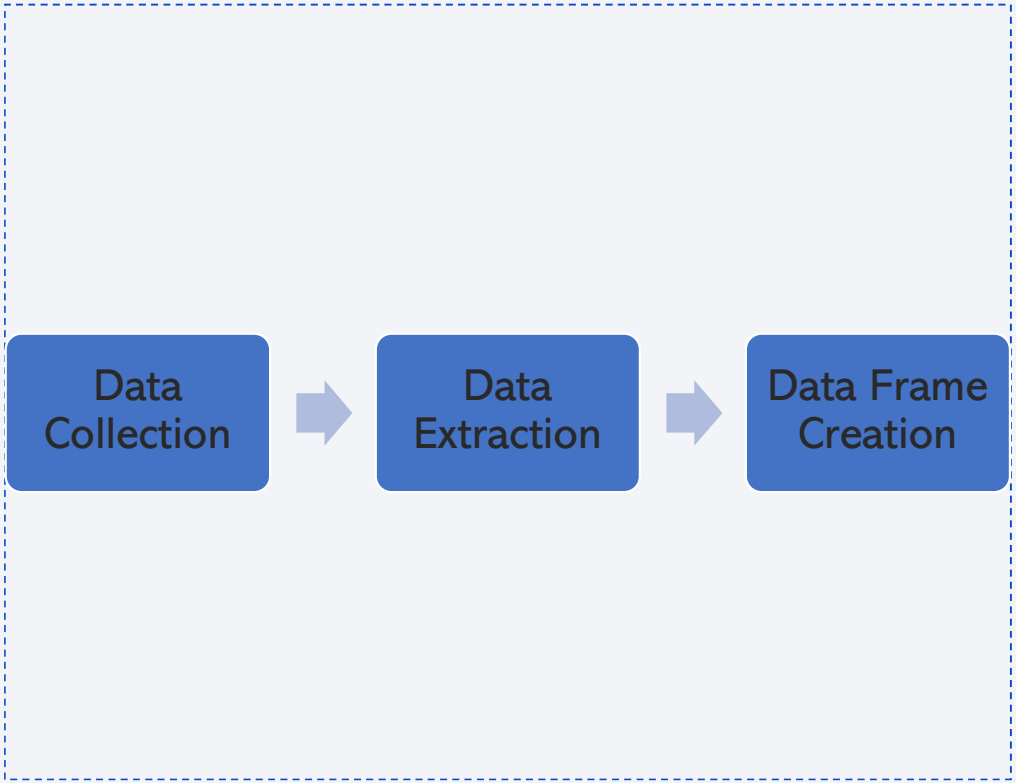
Data Collection – SpaceX API

- Data Collection:The SpaceX launch data was requested from the API (https://api.spacexdata.com/v4/rockets/) and retrieved for analysis.
- Data Filtering:The dataset was filtered to include only Falcon 9 launches, focusing specifically on this type of rocket for analysis.
- Handling Missing Values:Missing values in the dataset were addressed and appropriately dealt with to ensure data integrity and accuracy in the subsequent analysis.
- GitHub URL: https://github.com/Reper4ever/Applied_Data_Science_Capstone/blob/3f7f1cfd1690545d0ce13b2db9e23ec4cb1ee55f/week%201/Data%20Collection%20%E2%80%93%20SpaceX%20API.ipynb

Data Collection - Scraping

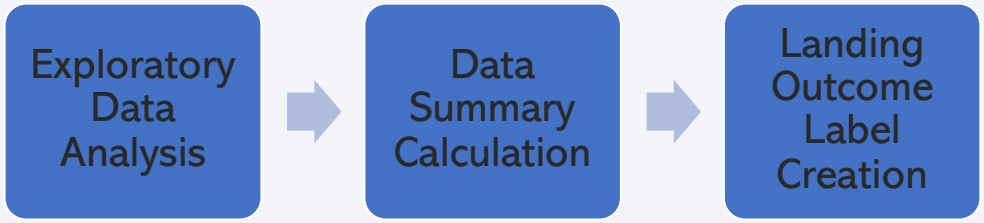
- Data Collection:The Falcon 9 Launch Wiki page was requested and retrieved for data extraction.
- Data Extraction:All column/variable names from the HTML table header were extracted to identify the relevant data fields.
- Data Frame Creation:A data frame was created by parsing the HTML tables from the Falcon 9 Launch Wiki page, organizing the data for further analysis.

GitHub URL:
https://github.com/Reper4ever/Applied_Data_Science_Capstone/blob/5efb3d210f3fd6c9f8b82ad0895be6d704fe2dde/week%201/Data%20Collection%20-%20Scraping.ipynb



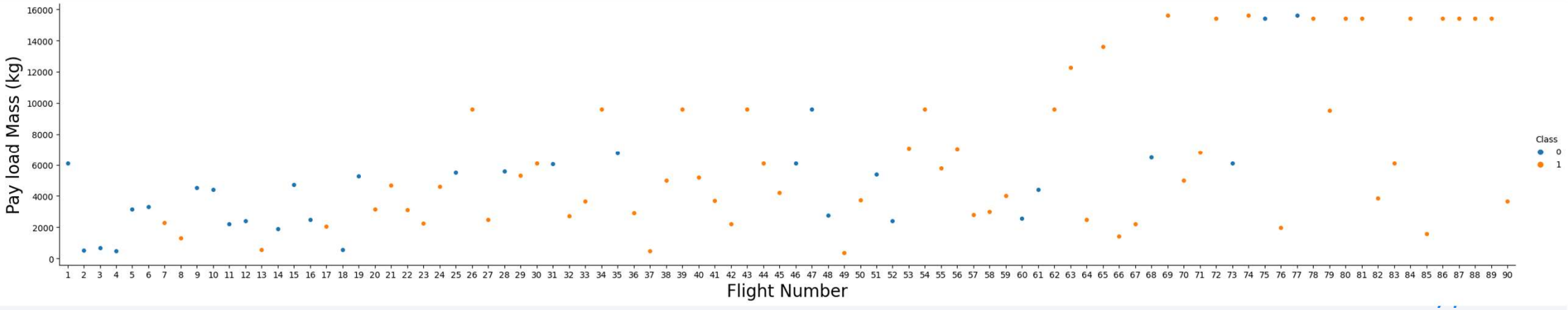
Data Wrangling

- Exploratory Data Analysis (EDA):The dataset was subjected to initial EDA to gain insights into its structure and characteristics.
- Data Summary Calculation:The number of launches per site, occurrences of each orbit, and occurrences of mission outcomes per orbit type were calculated to better understand the distribution and patterns in the data.
- Landing Outcome Label Creation:A landing outcome label was created based on the "Outcome" column to categorize the success or failure of the landing for each launch.
- GitHub URL: https://github.com/Reper4ever/Applied_Data_Science_Capstone/blob/c9ca8b71f57c61e045d69bd3bf9fd18f1bc176e5/week%201/Data%20Wrangling.ipynb



EDA with Data Visualization

- Exploratory Data Analysis (EDA) involved utilizing scatter plots and bar plots to visualize the relationships between various pairs of features. Specifically, we examined the relationships between Payload Mass and Flight Number, Launch Site and Flight Number, Launch Site and Payload Mass, Orbit and Flight Number, and Payload and Orbit. These visualizations provided valuable insights into the patterns and correlations within the dataset.
- GitHub URL: https://github.com/Reper4ever/Applied_Data_Science_Capstone/blob/518e5037cba2a9ae97f0d9bb1b1752d461dbd702/week%202/EDA%20with%20Data%20Visualization.ipynb



EDA with SQL

- The exploratory data analysis (EDA) involved employing scatterplots and barplots to effectively visualize the relationships between various features within the dataset. Specifically, the following pairs of features were analyzed: Payload Mass and Flight Number, Launch Site and Flight Number, Launch Site and Payload Mass, Orbit and Flight Number, and Payload and Orbit.
- In addition to EDA, several SQL queries were executed to gain deeper insights into the space mission data. These queries encompassed retrieving the names of unique launch sites, identifying the top 5 launch sites with names beginning with 'CCA', calculating the total payload mass carried by boosters launched by NASA (CRS), determining the average payload mass carried by booster version F9 v1.1, pinpointing the date of the first successful landing outcome in a ground pad, listing the names of boosters that successfully landed on a drone ship with a payload mass between 4000 and 6000 kg, tallying the total number of successful and failure mission outcomes, identifying booster versions that carried the maximum payload mass, retrieving information on failed landing outcomes on a drone ship along with their respective booster versions and launch site names for the year 2015, and ranking the count of landing outcomes (Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20. These queries facilitated a comprehensive understanding of the space mission data and provided valuable insights for further analysis and decision-making.
- GitHub URL: https://github.com/Reper4ever/Applied_Data_Science_Capstone/blob/e7cf2fcc5368481ff7396d9010c973159f0fdc3f/week%202/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

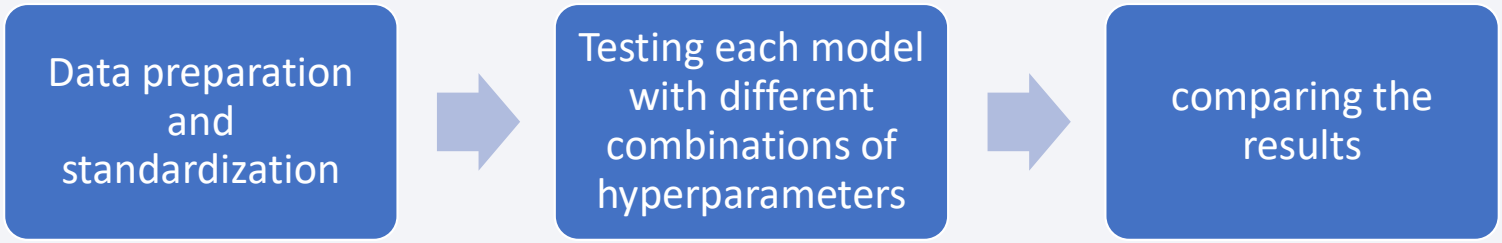
- During the visualization process, various elements were utilized with Folium Maps to enhance the representation of data. Markers were employed to denote specific points, such as launch sites, providing a clear visual indication of their locations. Additionally, circles were used to highlight specific areas around certain coordinates, for instance, NASA Johnson Space Center, drawing attention to significant regions of interest.
- To better organize and display data, marker clusters were utilized, enabling the grouping of events within each coordinate. This approach allowed for a more intuitive understanding of the distribution of launches in a particular launch site.
- Furthermore, lines were employed to represent distances between two coordinates, offering a visual representation of spatial relationships and facilitating the analysis of geographical aspects. By using these elements effectively, the Folium Maps were able to convey valuable insights and facilitate a more comprehensive exploration of the space mission data.
- GitHub URL: https://github.com/Reper4ever/Applied_Data_Science_Capstone/blob/f9b2a27888a081f0f7e45e6d7f53345767b07cbb/week%203/Build%20an%20Interactive%20Map%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- In the data visualization process, two specific graphs were employed to provide valuable insights:
- Percentage of Launches by Site: This graph showcased the distribution of launches across different sites, presenting the relative proportion of launches conducted at each site. This visualization enabled a quick analysis of the frequency of launches at different locations.
- Payload Range: Another plot utilized was the payload range graph, which displayed the variation in payload masses for different launch missions. This visualization allowed for a clear understanding of the payload distribution across various launches.
- The combination of these two visualizations proved to be particularly beneficial in analyzing the relationship between payloads and launch sites. By correlating payload ranges with specific launch locations, it facilitated the identification of the most suitable launch sites based on the payload requirements. This approach helped in determining the optimal launch site for missions with different payload masses, thereby providing valuable insights for decision-making in space missions.
- GitHub URL: https://github.com/Reper4ever/Applied_Data_Science_Capstone/blob/1e9c37caddad4b7acf743605b595b26841353710/week%203/Build%20a%20Dashboard%20with%20Plotly%20Dash.py

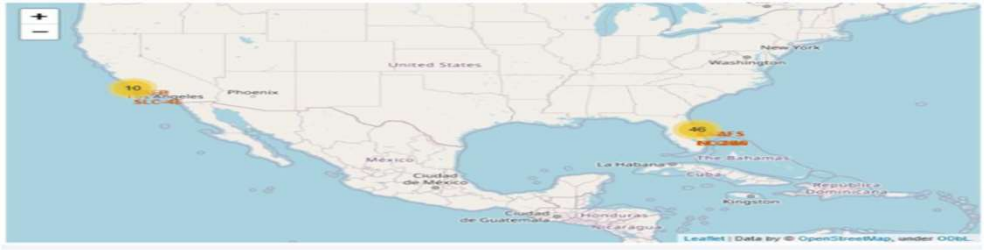
Predictive Analysis (Classification)

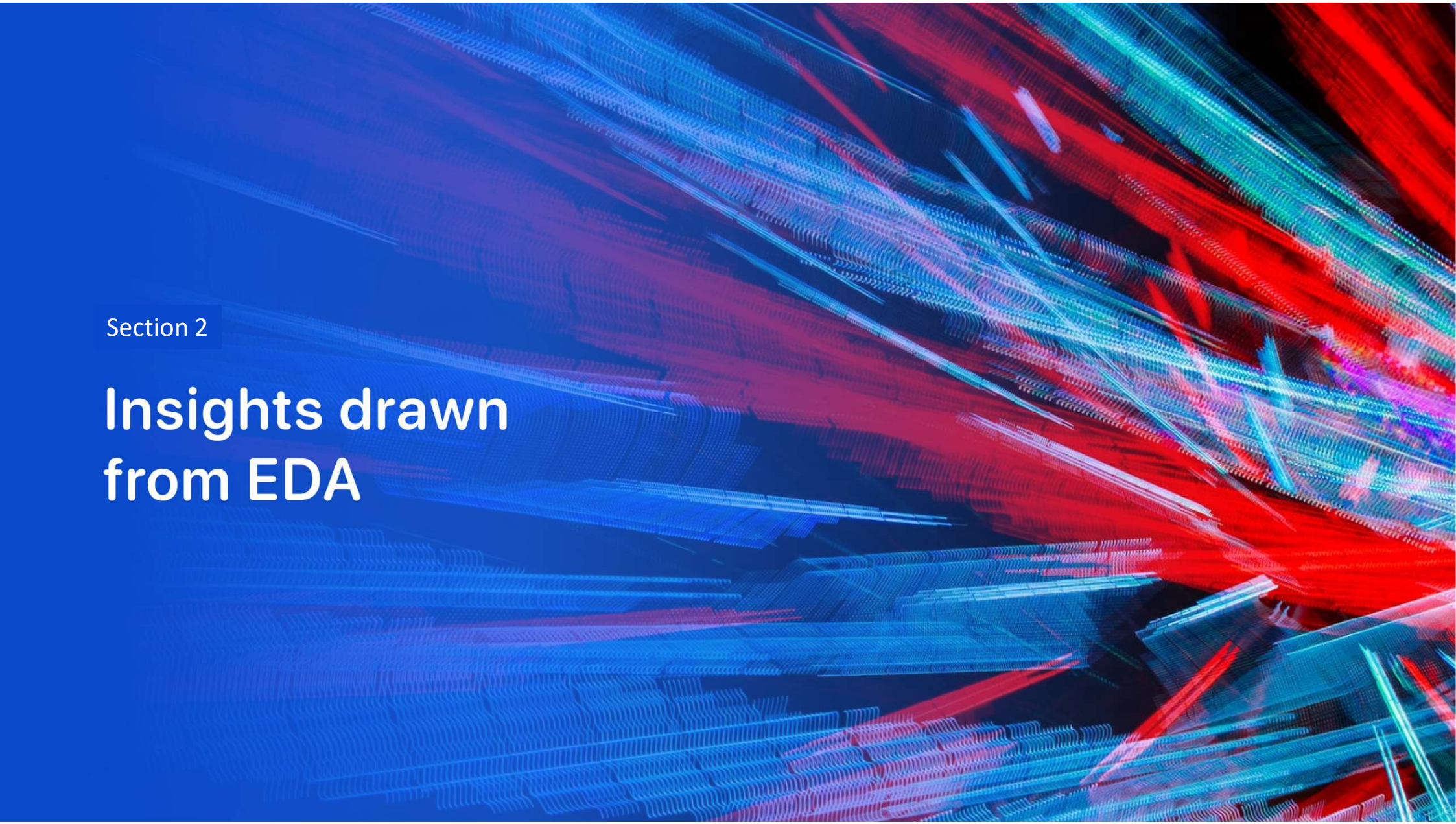
- Four classification models were compared: logistic regression, support vector machine, decision tree, and k-nearest neighbors. These models were evaluated using different combinations of parameters, and their accuracy in predicting the success of launches was assessed. The goal was to identify the best model that could effectively predict the success of launches based on the collected data.



Results

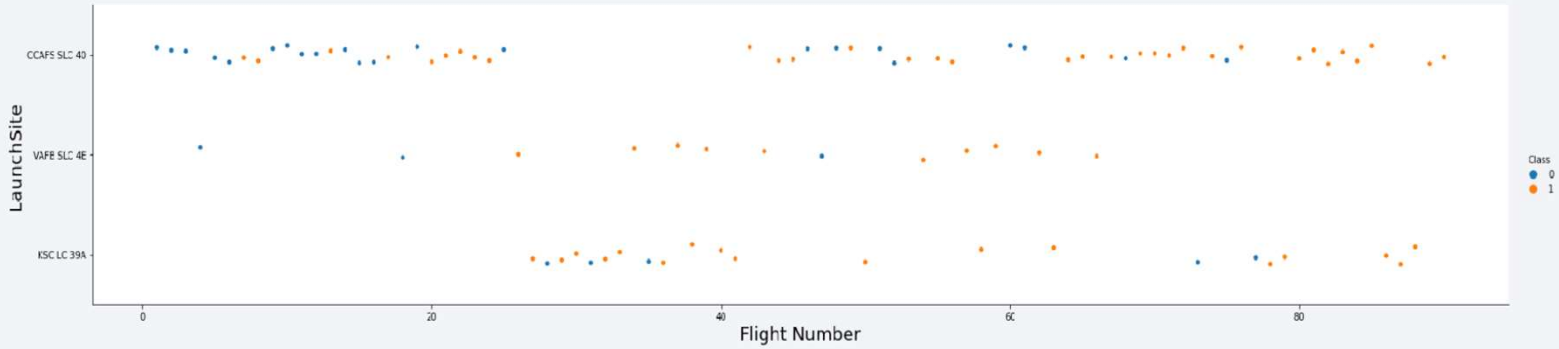
- SpaceX utilizes four distinct launch sites.
- The initial launches were conducted for SpaceX itself and NASA.
- The average payload of the F9 v1.1 booster is approximately 2,928 kg.
- The first successful landing outcome occurred in 2015, five years after the first launch.
- Numerous Falcon 9 booster versions achieved successful landings on drone ships with payloads exceeding the average.
- The overall mission success rate is nearly 100%. In 2015, two booster versions, namely F9 v1.1 B1012 and F9 v1.1 B1015, failed to land on drone ships.
- Over the years, the number of successful landing outcomes has improved significantly.




Section 2
Insights drawn from EDA

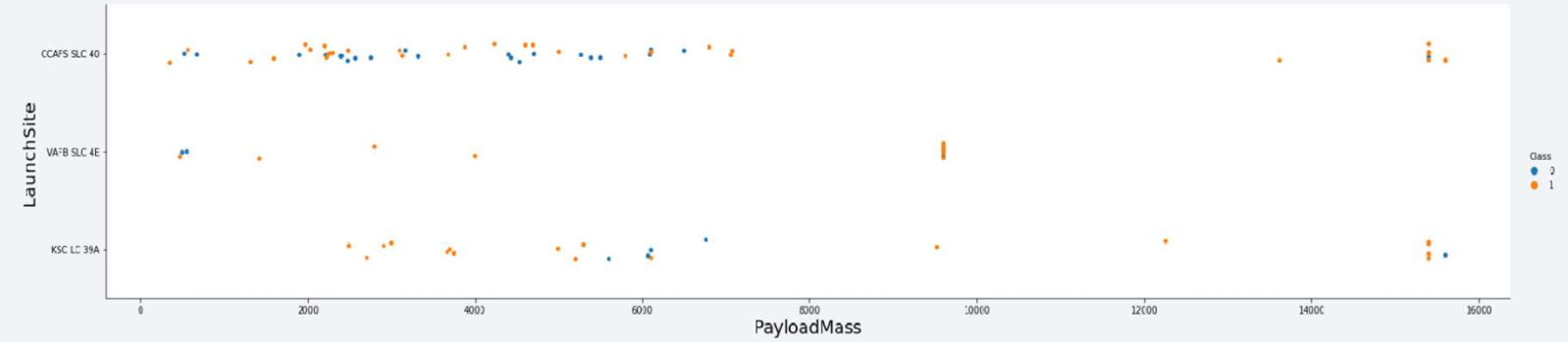
Flight Number vs. Launch Site

- It is evident that the best launch site currently is CCAFS SLC 40, as it has recorded the highest number of recent successful launches. VAFB SLC 4E ranks second, and KSC LC 39A ranks third in terms of successful launches. Furthermore, the plot also shows a positive trend in the overall success rate, indicating an improvement over time.



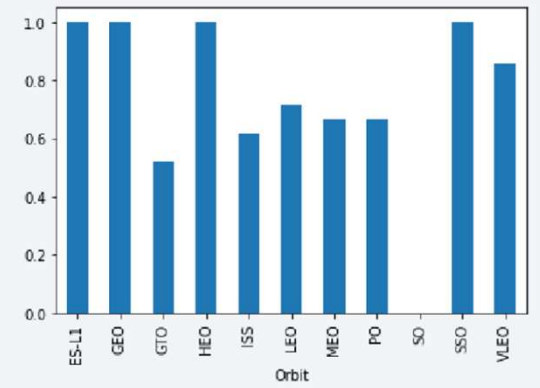
Payload vs. Launch Site

- Payloads exceeding 9,000kg, roughly equivalent to the weight of a school bus, exhibit an excellent success rate in the launches. Moreover, payloads surpassing 12,000kg appear to be feasible only at the CCAFS SLC 40 and KSC LC 39A launch sites. This information suggests that these two launch sites have the capability and infrastructure to handle heavier payloads successfully.



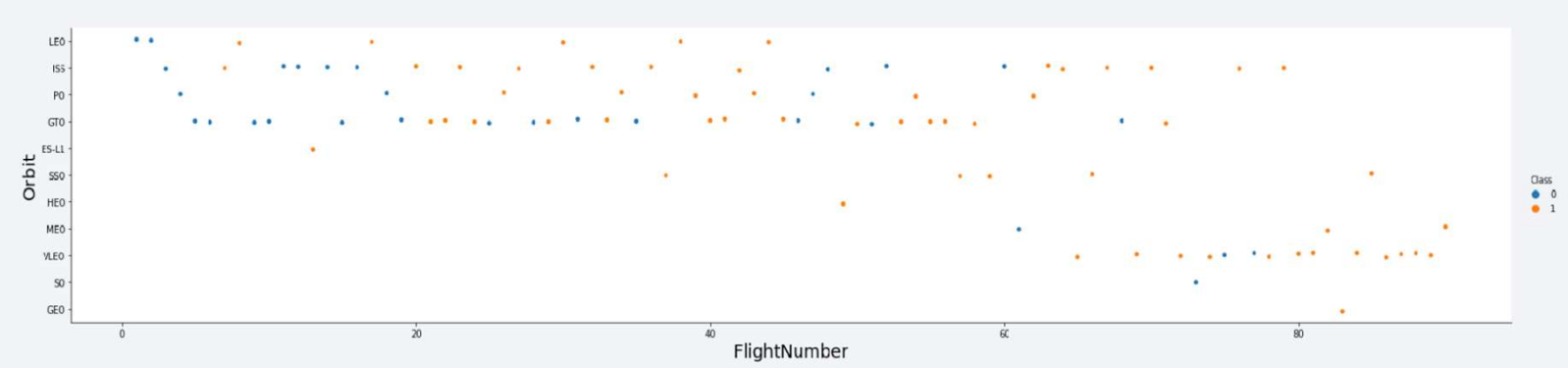
Success Rate vs. Orbit Type

- The orbits with the highest success rates are ES-L1, GEO, HEO, and SSO. These orbits have achieved significant success in their launches. Following closely are VLEO, with a success rate above 80%, and LFO, with a success rate above 70%. These findings indicate that ES-L1, GEO, HEO, and SSO orbits are particularly reliable and successful in their launch missions.



Flight Number vs. Orbit Type

- Indeed, the success rate for all orbits has shown improvement over time, indicating that Space Y has been consistently enhancing its launch capabilities and achieving better outcomes across different orbits. Particularly, the increase in the frequency of VLEO (Very Low Earth Orbit) launches presents a new business opportunity for Space Y. The recent surge in VLEO missions suggests a growing demand for this orbit, making it a promising area for Space Y to explore and potentially capitalize on in the future.

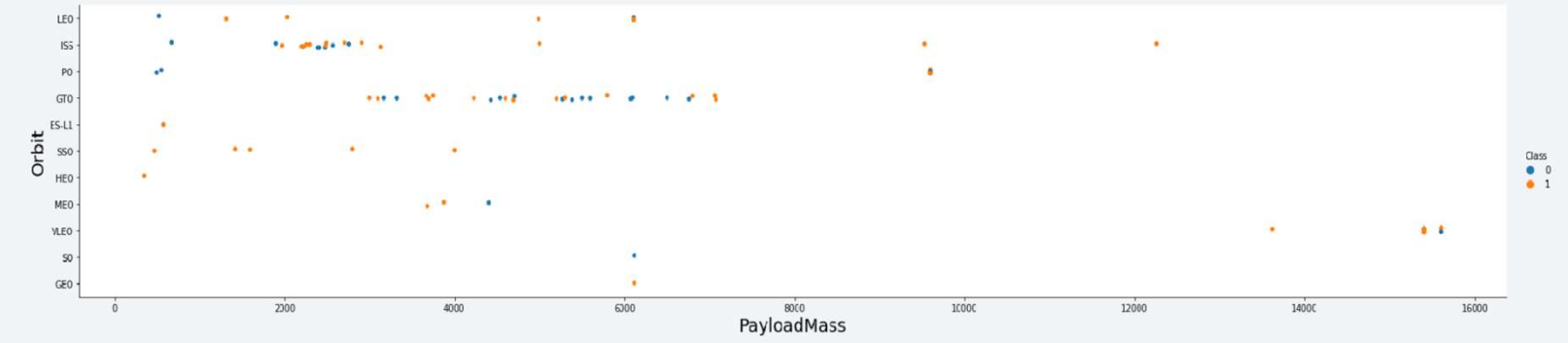


Payload vs. Orbit Type

- It appears that there is no clear relationship between the payload and success rate for the GTO (Geostationary Transfer Orbit) orbit. While GTO has the widest range of payload masses, it still maintains a good rate of success.

- Additionally, it is worth noting that there have been relatively few launches to the SO (Sun-Synchronous Orbit) and GEO (Geostationary Orbit) orbits. This could be due to various factors, such as the specific requirements and demand for these orbits, which may result in a lower number of missions compared to other orbits.

- Overall, the data suggests that different orbits have unique characteristics and varying levels of demand, which may influence the number of launches and success rates observed for each orbit.

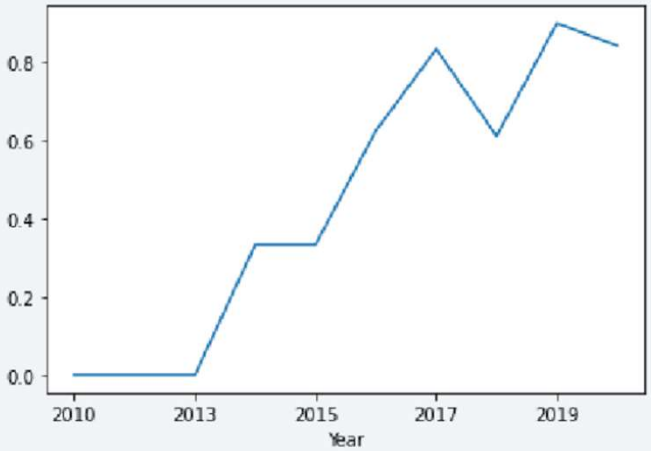


Launch Success Yearly Trend

- Indeed, the success rate of SpaceX launches started increasing in 2013 and continued to improve until 2020. The initial three years (2010-2012) might have been a period of adjustments and technological improvements for SpaceX as they were relatively new to the space industry.

- During these years, SpaceX was likely fine-tuning their launch processes, developing more advanced technologies, and gaining valuable experience with their Falcon 9 rockets. As a result of continuous efforts and enhancements, they managed to achieve a higher success rate in subsequent years.

- This trend of improvement in success rates demonstrates SpaceX's commitment to safety, reliability, and continuous innovation in their space missions. It also reflects their ability to learn from previous missions and make necessary adjustments to enhance the success of future launches.



All Launch Site Names

- The success rates mentioned earlier are obtained by calculating the percentage of successful launches for each unique "launch_site" value in the dataset. By selecting unique occurrences of "launch_site" values, we can analyze the success rates specific to each launch site independently.

- To calculate the success rate for a particular launch site, we first count the total number of launches from that site and then count the number of successful launches. Dividing the number of successful launches by the total number of launches and multiplying by 100 gives us the percentage of successful launches for that site.

- This process is repeated for each unique "launch_site" value in the dataset, allowing us to compare the success rates among different launch sites and identify which sites have higher or lower success rates. It provides valuable insights into the performance of each launch site and helps in making informed decisions regarding future launch operations.

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

Date	Time UTC	Booster Version	Launch Site	Payload	Payload Mass kg	Orbit	Customer	Mission Outcome	Landing Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

Total Payload Mass

Total Payload (kg)
111.268

Average Payload Mass by F9 v1.1

Avg Payload (kg)
2.928

First Successful Ground Landing Date

Min Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

Booster Version (...)
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3

Booster Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

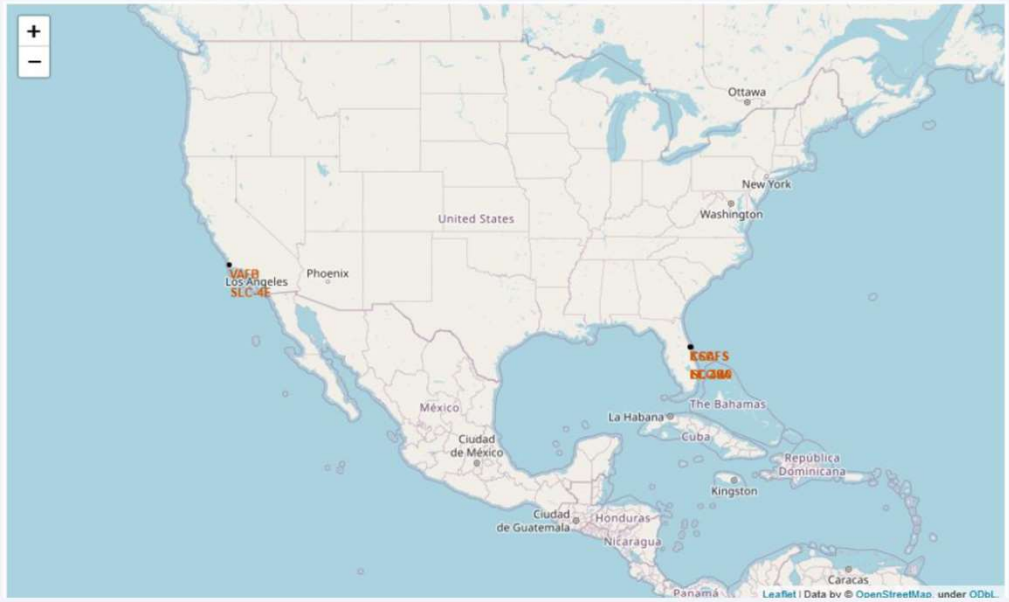
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

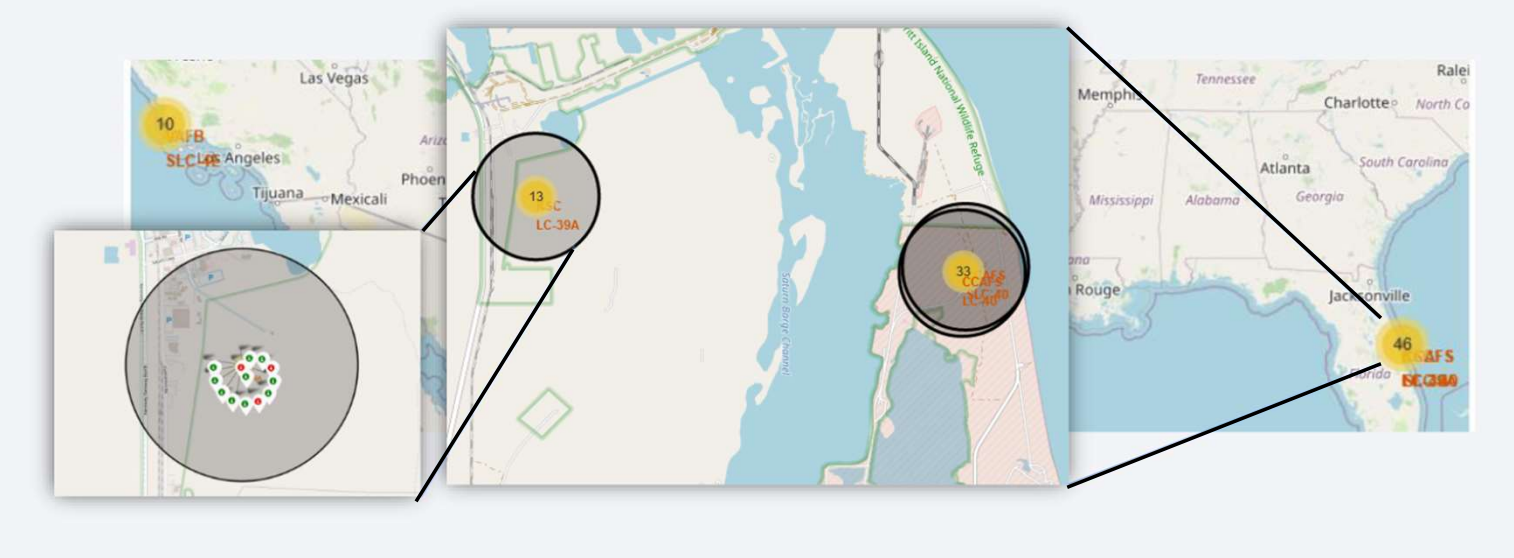
launch sites

The selection of launch sites for space missions often prioritizes proximity to the sea for safety considerations, allowing for safer abort options in case of emergencies. Additionally, launch sites are strategically located not too far from roads and railroads to facilitate transportation of equipment, payloads, and personnel to the launch site efficiently. This balance between proximity to the sea and accessibility to transportation infrastructure ensures the smooth execution of space missions while maintaining safety standards.



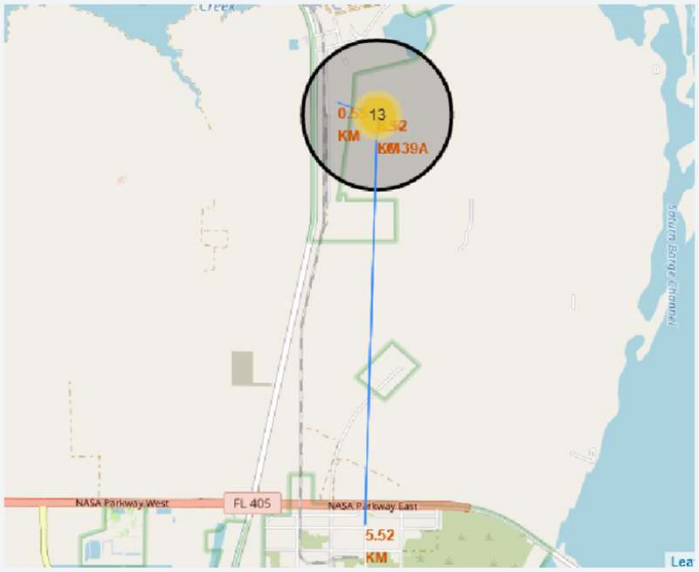
outcomes of space missions

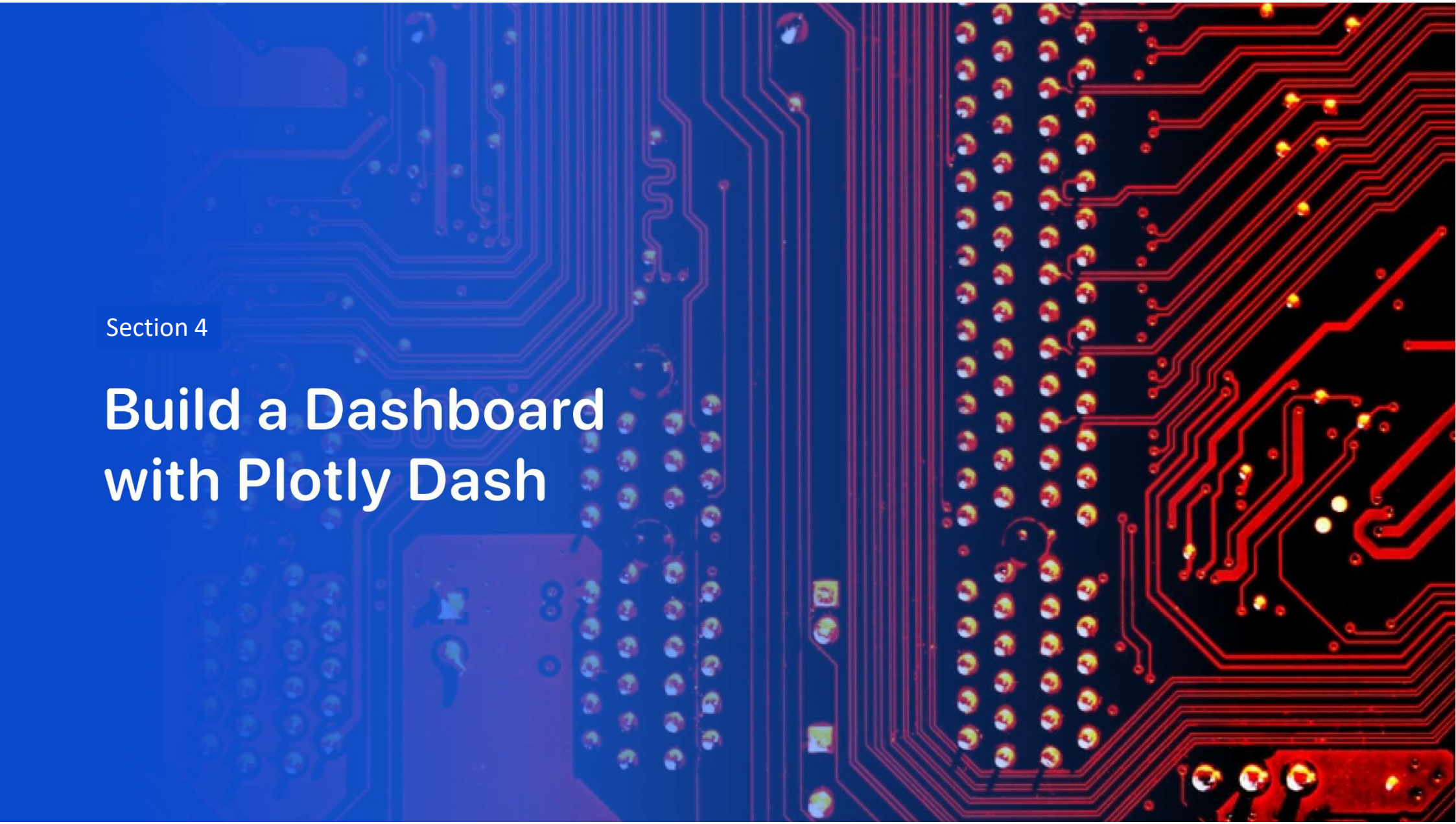
The visualization uses green markers to indicate successful outcomes of space missions and red markers to indicate mission failures. This color coding system helps easily distinguish between successful and unsuccessful launches, providing a clear and intuitive representation of the data.



logistics

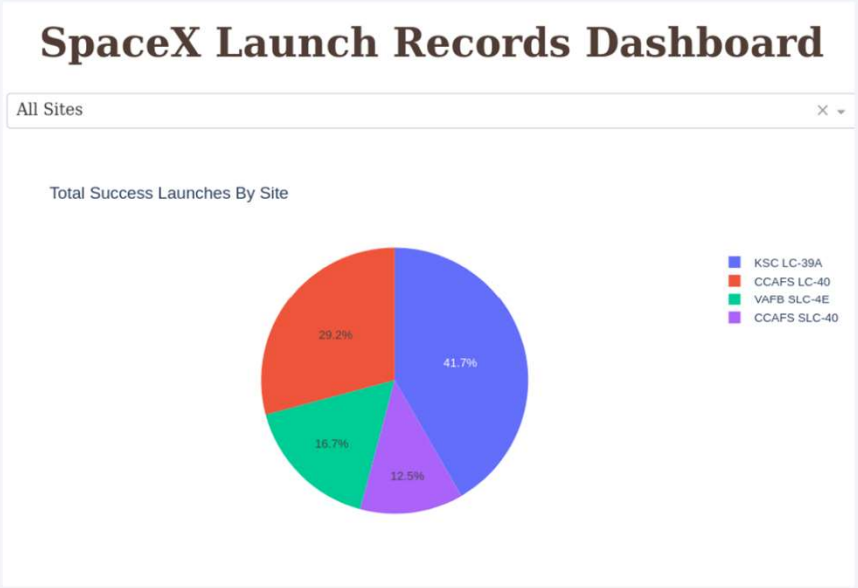
The launch site KSC LC-39A exhibits favorable logistics aspects as it is relatively far from inhabited areas, reducing potential risks to populated regions. Additionally, its proximity to both railroad and road infrastructure enhances transportation efficiency and accessibility for logistical operations. These factors contribute to KSC LC-39A's suitability as a launch site for space missions.




Section 4
Build a Dashboard with Plotly Dash

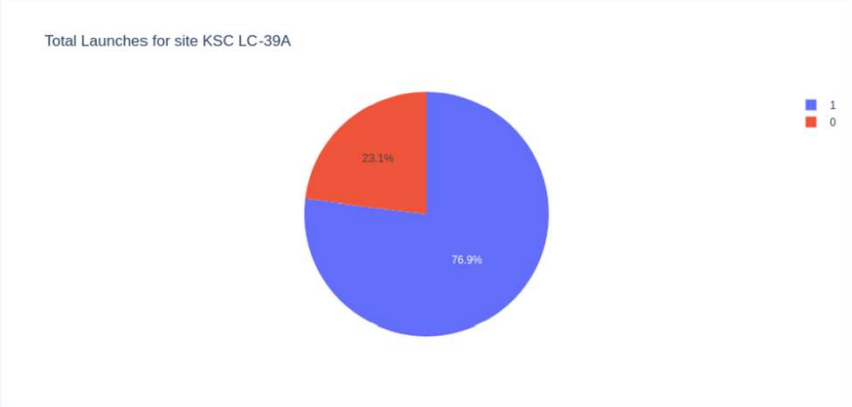
launch site -- crucial role in the success of space missions

- Indeed, the launch site plays a crucial role in the success of space missions. Factors such as geographic location, weather conditions, proximity to populated areas, and available infrastructure can significantly impact the outcome of a mission. Choosing the right launch site can optimize the trajectory, payload capacity, and safety of the launch, ultimately increasing the chances of a successful mission. It is essential for space agencies and companies like SpaceX to carefully consider and select the most suitable launch site based on various operational and environmental factors to ensure mission success.



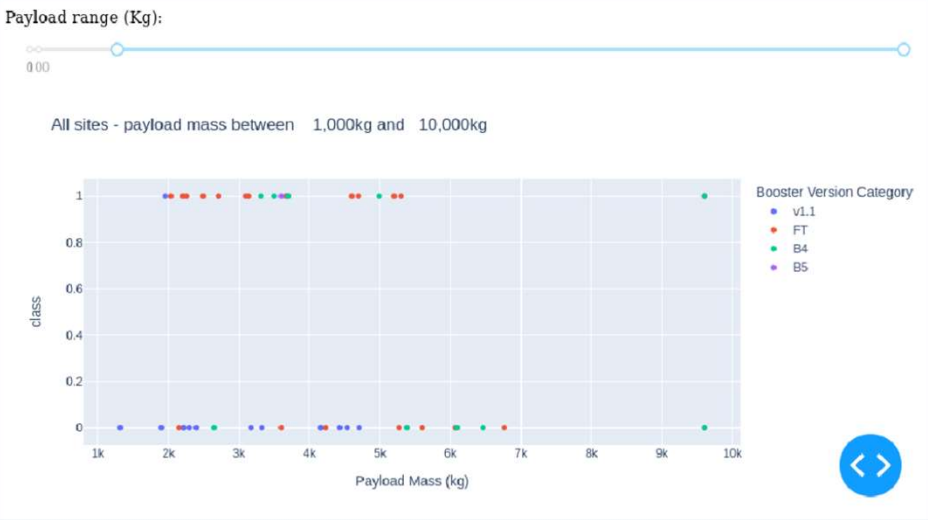
KSR LC-39A

- At the specific launch site mentioned, approximately 76.9% of the launches have been successful. This success rate indicates that the site has been effective in facilitating successful space missions, with the majority of launches achieving their objectives. The high success rate may be attributed to various factors, including meticulous planning, advanced technology, and experienced teams working at the site. Maintaining a high success rate is crucial for space agencies and companies to ensure the efficient and reliable execution of space missions.



Payloads vs. launch outcomes

- According to the data analysis, payloads under 6,000kg paired with FT boosters have shown the highest success rate among various payload and booster combinations. This finding suggests that the FT boosters are well-suited for handling payloads within this weight range, resulting in a higher likelihood of mission success. The successful combination of payload and booster is crucial in ensuring the efficient and reliable execution of space missions while optimizing resources and achieving mission objectives effectively.

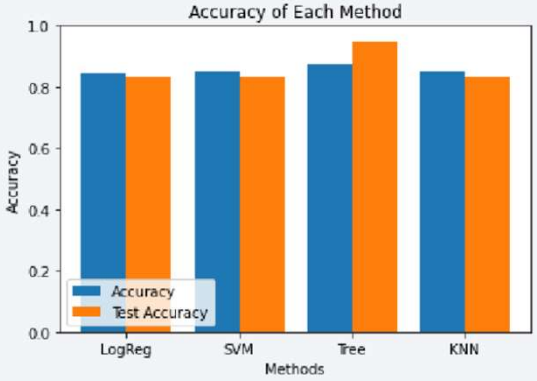


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- During the analysis, four classification models were evaluated, and their respective accuracies were plotted for comparison. Among these models, the Decision Tree Classifier exhibited the highest classification accuracy, consistently achieving accuracies over 87%. This result indicates that the Decision Tree Classifier is the most effective model for predicting the outcomes of space mission launches. With its high accuracy, this model can provide valuable insights and aid in making informed decisions, enhancing the success rate of space missions and contributing to the overall success of the space exploration endeavors.

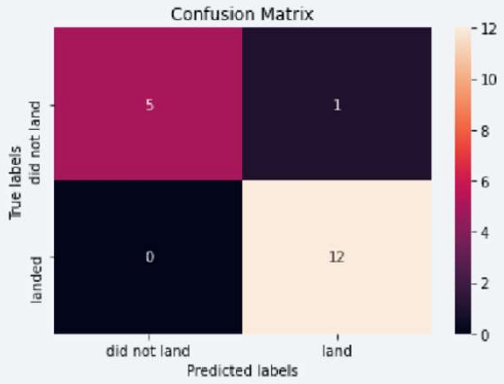


Confusion Matrix

- The Confusion Matrix of the Decision Tree Classifier indeed provides valuable insights into its accuracy by showcasing the counts of true positives, true negatives, false positives, and false negatives. The true positives and true negatives represent the correct predictions made by the model for successful and unsuccessful launches, respectively. These values are typically larger in a well-performing model, indicating that it correctly identifies successful and unsuccessful outcomes.

- On the other hand, false positives and false negatives represent the incorrect predictions made by the model, where it mistakenly identifies an unsuccessful launch as successful or vice versa. In a highly accurate model, the numbers of false positives and false negatives are usually relatively smaller.

- By analyzing the Confusion Matrix, decision-makers can gain a deeper understanding of the model's performance, its ability to correctly classify different outcomes, and its overall accuracy in predicting launch results. This information is crucial for making data-driven decisions and improving the success rate of space missions.



Conclusions

- Throughout the analysis process, various data sources were examined and conclusions were iteratively refined to arrive at meaningful insights. The investigation revealed that the optimal launch site for successful missions is KSC LC-39A, offering favorable logistics aspects and proximity to crucial transportation routes.

- Moreover, it was observed that launches with payloads exceeding 7,000kg present lower risks and a higher probability of success. The data also indicated an overall improvement in mission outcomes over time, particularly in terms of successful landings, attributed to advancements in processes and rocket technology.

- In order to predict successful landings and enhance profitability, the Decision Tree Classifier emerged as a reliable model with accuracies surpassing 87%, providing valuable guidance for informed decision-making and strategic planning.

- By combining these findings from various data sources and analytical techniques, valuable insights have been obtained, leading to a deeper understanding of factors influencing successful space missions and driving opportunities for further advancements in the space industry.

Appendix

- No more

Thank you!