IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

AJ
03/Jan./2024

Outline

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

Executive Summary

Methodologies:

- **Data Acquisition:** Gathered valuable data from public sources using web scraping and the SpaceX API.
- **Data Preparation:**
  - Wrangled and cleaned the collected data.
  - Performed Exploratory Data Analysis (EDA).
  - Conducted EDA with SQL for deeper exploration.

Visualization:

- Built an interactive map with Folium to enhance spatial understanding.
- Created a Dashboard with Plotly Dash for clear data presentation.

Results:

- **EDA:** Identified key features most influential in predicting launch success.
- **Interactive Analytics:** Developed informative screenshots showcasing the interactive analysis capabilities.
- **Predictive Model:** Established the best model for predicting launch success based on various characteristics.

Introduction

Motivation:

The space launch industry is experiencing a renaissance, driven by competition to reduce costs and increase accessibility. SpaceX, a company known for its innovative reusable rockets, has significantly lowered launch prices compared to traditional providers. This success begs the question: can a new entrant like Space Y effectively compete with the established leader, SpaceX?

Research Objectives:

To assess Space Y's competitive viability, this study will address two key questions:

- **Estimating launch cost through successful first-stage landings:** Can we develop a reliable method to predict the total cost of a launch based on the probability of the first stage returning successfully? This requires identifying the key factors influencing landing success, such as launch site, payload orbit, rocket mass, landing pad location, and booster version.
- **Optimizing launch location:** By analyzing historical launch data, can we determine the optimal location for Space Y's launches.

Methodology

Executive Summary

- Data collection methodology:
  - Source Selection: Leveraged both the SpaceX API and web scraping techniques to gather comprehensive launch data.
    - Space X API (https://api.spacexdata.com/v4/) for direct data retrieval
    - Web Scraping data from Wikipedia (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
  - The data was processed by performing initial data cleaning which included filtering outliers, handling missing values by imputation, applying one-hot encoding to categorical variables, and creating labels for the target variable.
- Exploratory Data Analysis (EDA):
  - Data Visualization: Used various visualization techniques to explore relationships between variables and uncover patterns.
  - SQL Exploration: Employed SQL queries for deeper data exploration and manipulation.

Section 1

Methodology

Methodology

Executive Summary

- Data Collection Methodology:
  - Space X API (https://api.spacexdata.com/v4/) for direct data retrieval
  - Web Scraping data from Wikipedia (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform Data Wrangling
  - The data was processed by performing initial data cleaning which included filtering outliers, handling missing values, applying one-hot encoding to categorical variables, and creating labels for the target variable.
- Exploratory Data Analysis (EDA):
  - Data Visualization: Used various visualization techniques to explore relationships between variables and uncover patterns.
  - SQL Exploration: Employed SQL queries for deeper data exploration and manipulation.

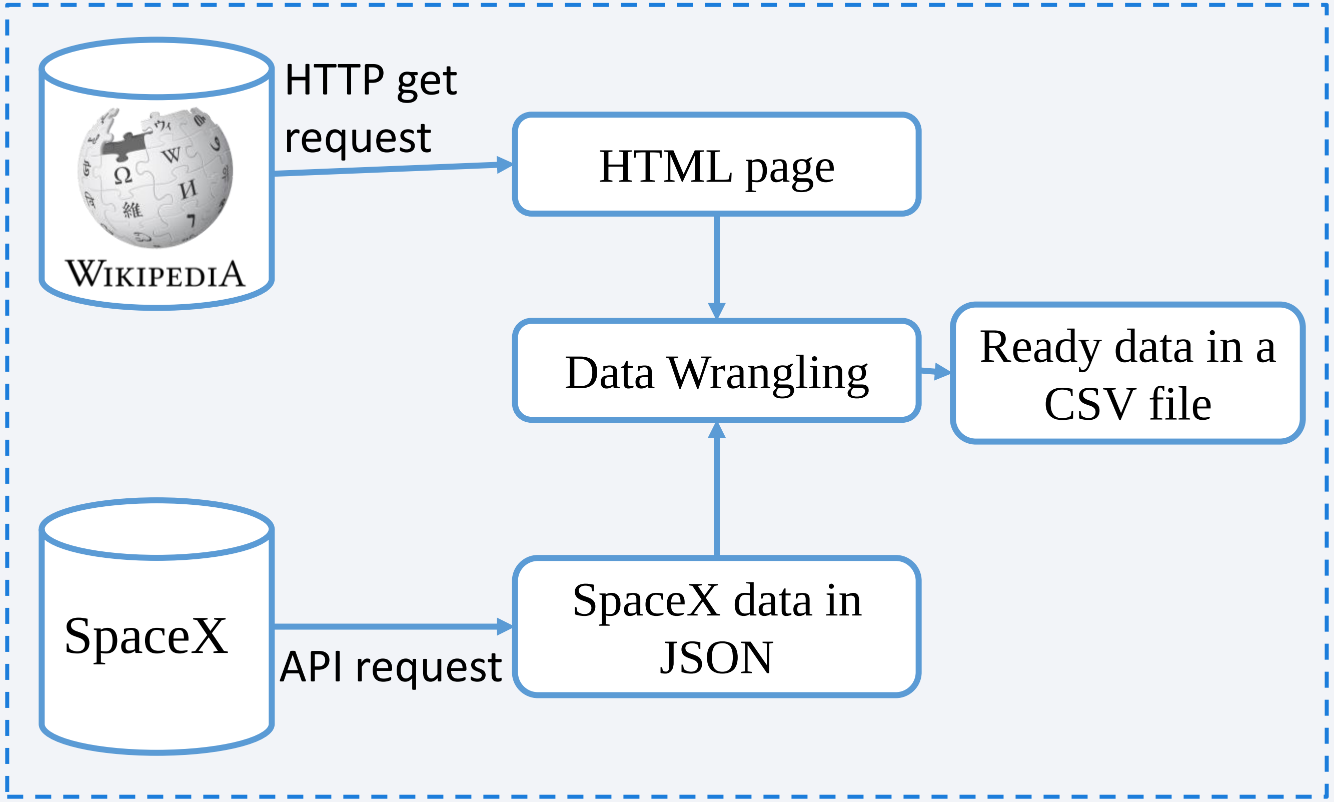
Methodology

Executive Summary

- Interactive Visualization:
  - **Folium Map:** Created an interactive map with Folium to visualize spatial patterns and trends in launch data.
  - **Plotly Dash Dashboard**: Developed an interactive dashboard using Plotly Dash for comprehensive data presentation and exploration.
- Predictive Modeling:
  - **Classification Model Selection**: Built multiple classification models to predict launch outcomes, evaluating their performance to select the best one.
  - **Model Tuning:** Optimized model hyperparameters to enhance predictive accuracy.
  - **Evaluation Metrics:** Assessed model performance using relevant metrics (e.g., accuracy, precision, recall, F1-score).

Data Collection

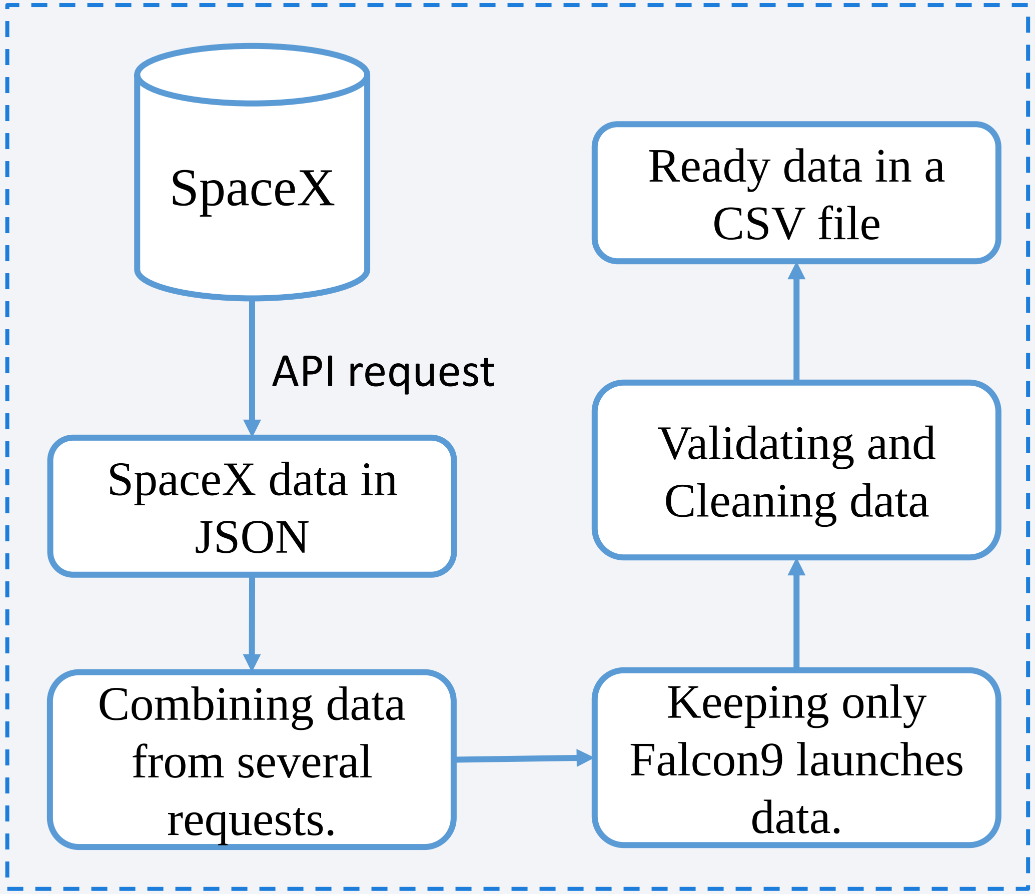
The Data collection process involved a combination of API requests from SpaceX API and web scraping data from SpaceX's Wikipedia entry. To obtain a complete dataset for detailed analysis of SpaceX launches.



Data Collection – SpaceX API

The Approach:

- **Targeted Requests:** Utilized multiple API calls, each focusing on specific data points.
- **Building the Puzzle**: Each API call returned a piece of the data puzzle. By combining these carefully, a detailed and accurate picture was constructed with a sharp focus on Falcon 9 launches for in-depth exploration.
- **Cleaning and Refining:** The raw data wasn't always perfect. data cleaning techniques were implemented to address missing values, inconsistencies, and formatting issues, ensuring high-quality data for analysis.
- The Outcome:
  - **A Rich Dataset**: Comprehensive dataset, ready to empower insightful analysis on various aspects of SpaceX's launch program.

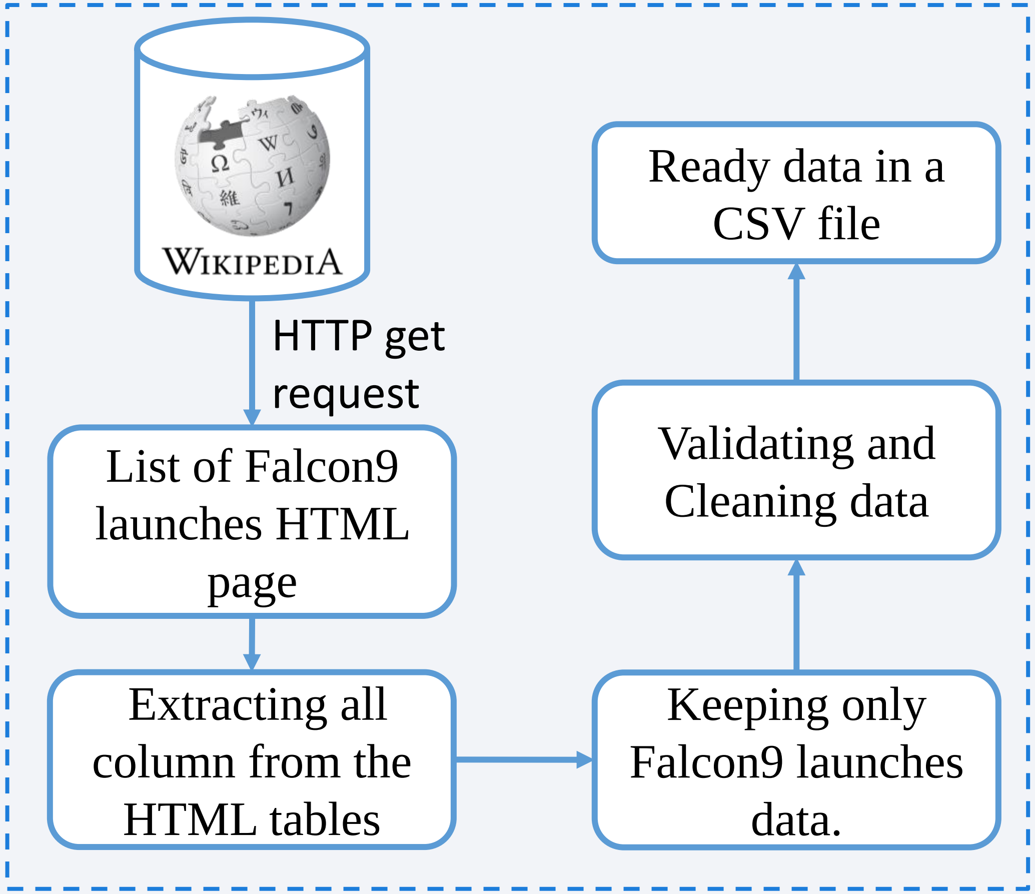

SpaceX
API request
SpaceX data in JSON
Combining data from several requests.
Keeping only Falcon9 launches data.
Validating and Cleaning data
Ready data in a CSV file

Related Notebook

Data Collection – Scraping

The Approach:

- **Mining Wikipedia's Knowledge:** Extracted a wealth of Falcon 9 launch data from Wikipedia.
- **Parsing the Launch Record**: Employing web scraping techniques, the HTML structure of the page was parsed, carefully extracting key details about each Falcon 9 launch.
- **Cleaning and Refining:** Implemented data cleaning techniques to address missing values, inconsistencies, and formatting issues, ensuring high-quality data for analysis.
- The Outcome:
  - **A Rich Dataset**: Comprehensive dataset of Falcon 9 launches, open for exploration and discovery.


WIKIPEDIA
HTTP get request
List of Falcon9 launches HTML page
Extracting all column from the HTML tables
Keeping only Falcon9 launches data.
Validating and Cleaning data
Ready data in a CSV file

Related Notebook

Data Wrangling

Exploratory Data Analysis (EDA):

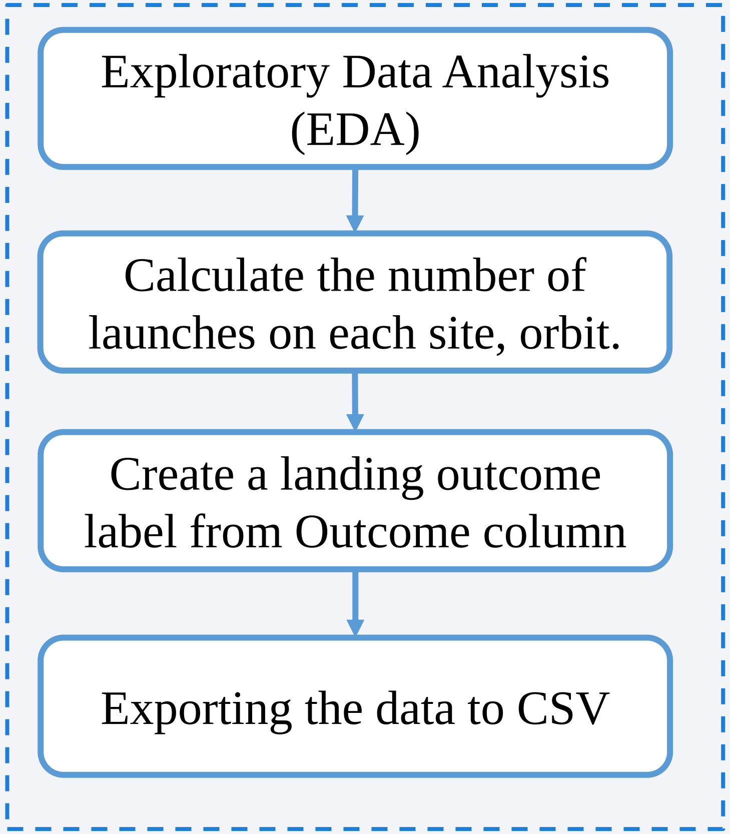
- Analyzed the overall data distribution and characteristics
- Identified missing values and inconsistencies.

Feature Engineering:

- Derived Landing Outcome Label
- **Calculated Feature Counts:** Determined the frequency of launches per launch site, occurrences of each orbit type, and mission outcome distribution within each orbit.

Data Wrangling:

- **Data Cleaning:** Addressed missing values and inconsistencies as identified in the EDA.
- **Data Transformation:** Applied further transformations based on EDA findings (e.g., encoding categorical variables)



Exploratory Data Analysis (EDA)

Calculate the number of launches on each site, orbit.

Create a landing outcome label from Outcome column

Exporting the data to CSV

Related Notebook

EDA with Data Visualization

In the Exploratory Data Analysis (EDA) stage, various charts were utilized to visualize relationships between key features within the SpaceX launch data.

Scatter Plots:

- **Flight Number vs. Launch Site**: This could indicate if heavier payloads require specific launch configurations or experience different outcomes.
- **Payload Mass vs. Launch Site :** This could reveal site limitations or preferences based on payload size.
- **Flight Number vs. Orbit Type:** This could uncover patterns or trends in mission planning.
- **Payload Mass vs. Orbit Type:** This could inform future launch planning based on payload characteristics.

Bar Charts:

- **Orbit Type vs. Success Rate:** This insights into potential challenges associated with specific orbits.

Bar Charts:

- **Success Rate Yearly Trend:** This could reveal potential temporal trends or improvements in SpaceX's operations.

Related Notebook

EDA with SQL

Summary of the SQL queries performed :

- All Launch Site Names
- Launch Site Names Begin with 'CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- First Successful Ground Landing Date
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
- Total Number of Successful and Failure Mission Outcomes
- Names of the booster which have carried the maximum payload mass
- 2015 Failed Launch Records
- Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Related Notebook

Build an Interactive Map with Folium

Map Objects

Markers:

- **Launch Sites:** Pinpoint the exact locations of all launch sites, providing a visual overview of their geographical distribution.
- **Launch Outcomes:** Visually display launch successes (green markers) and failures (red markers) at each site, enabling quick identification of sites with higher success rates.

Marker Clusters: Group multiple launch markers at each site, preventing visual clutter and ensuring clarity when numerous launches have occurred from a single location.

Circles: Subtly highlight the surrounding areas of launch sites, aiding in visual distinction and proximity assessment.

Proximity Lines: Depict distances between a specific launch site (VAFB SLC-4E in the example) and nearby transportation infrastructure (railway, highway), coastline, and the closest city. This aids in understanding the site's accessibility and potential logistical considerations..

Related Notebook

Build a Dashboard with Plotly Dash

The following graphs and plots were used to visualize data:

1. Total Successful Launches by Site:

   This pie chart highlights the percentage of successful launches originating from each site. It helps viewers identify which sites have the highest track record of success, potentially revealing factors like operational expertise or favorable launch conditions.

2. Correlation between Payload and Success:

   his scatter plot tackles a different angle, exploring the potential relationship between payload weight and launch success rates

Related Notebook

Predictive Analysis (Classification)

Model Development Process

Data Preparation:

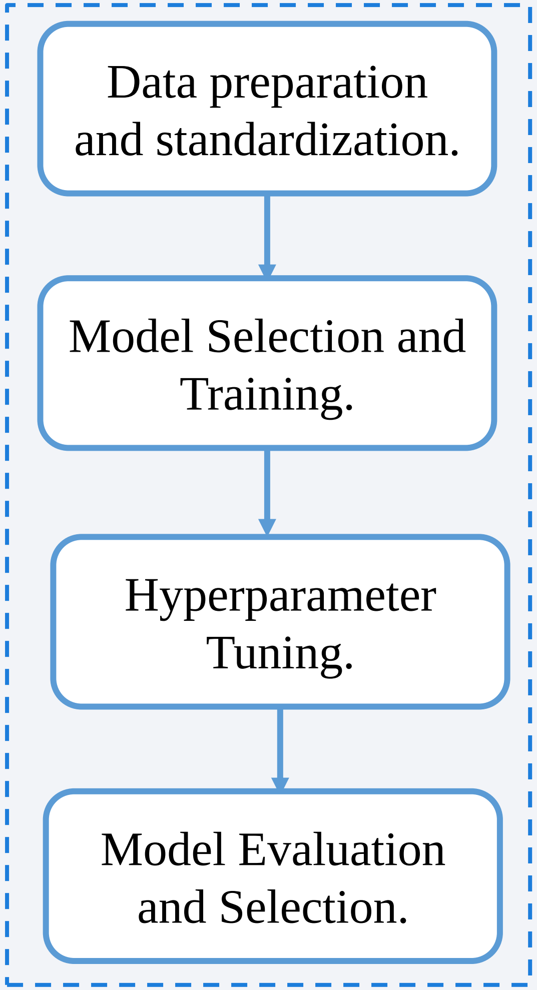
- **Standardization:** Normalized features using StandardScaler to ensure consistency in model training.
- **Train-Test Split:** Divided data into training and testing sets using 80/20 split ratio.

Model Selection and Training:

- **Model Candidates:** Explored four diverse classification models: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).
- **Hyperparameter Tuning:** Employed Grid Search with cross-validation to systematically identify optimal hyperparameters for each model, ensuring robust performance.

Model Evaluation and Selection:

Assessed model accuracy using a comprehensive set of metrics: (Accuracy, Confusion Matrix, F1 Score, etc.) Based on these metrics, selected the model demonstrating the most consistent and reliable performance



Data preparation and standardization.

Model Selection and Training.

Hyperparameter Tuning.

Model Evaluation and Selection.

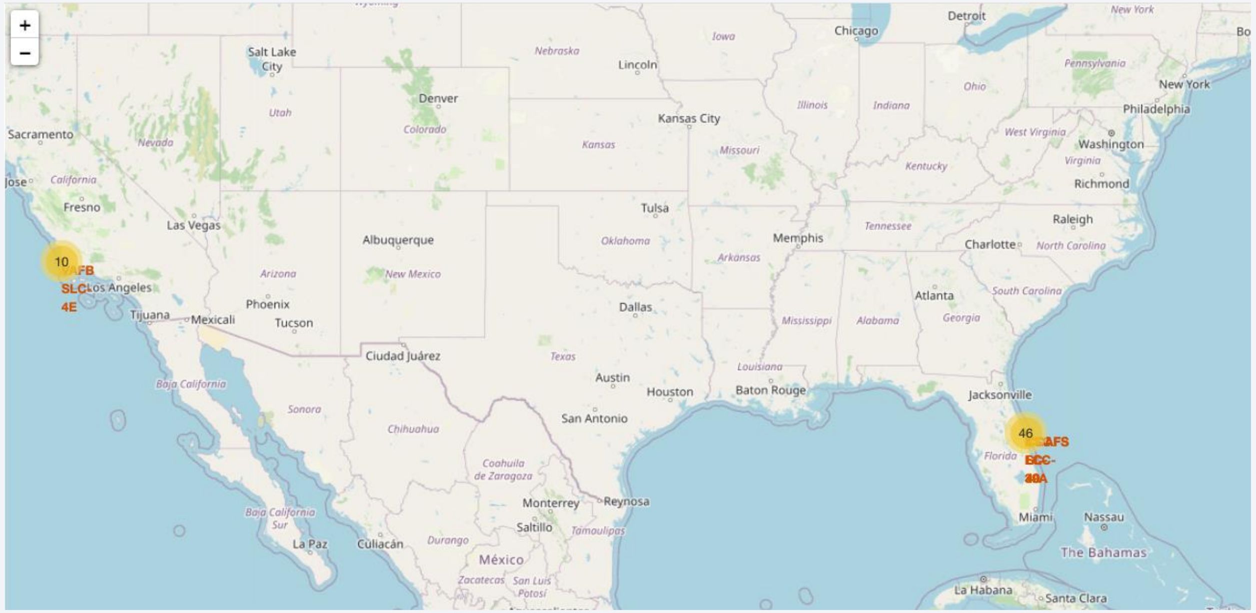
Related Notebook

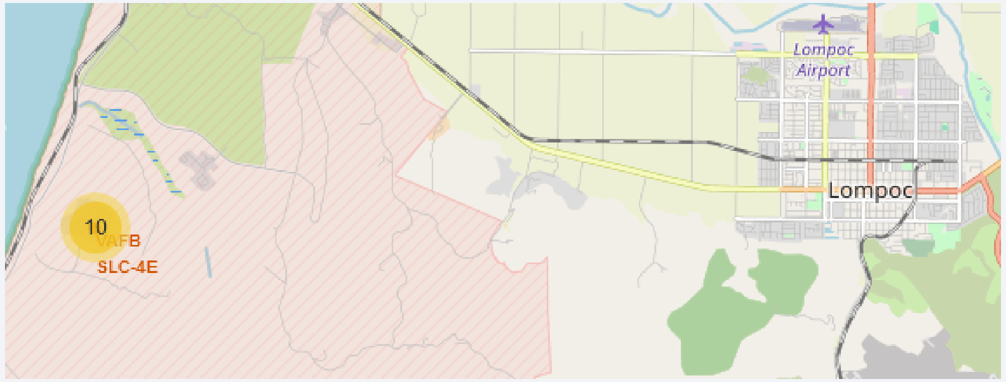
Results

- Exploratory data analysis results
  - Space X uses 4 different launch sites;
  - The average payload of F9 v 1.1 booster is 2,928 kg;
  - The first success landing outcome happened in 2015 fiver year after the first launch;
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
  - Two booster versions failed at landing in drone ships in 201 5: F9 v 1.1 BIOI 2 and F9 v 1.1 BIOI 5;
  - The number of landing outcomes became as better as years passed.

Results

- Exploratory data analysis results
  - Using interactive analytics was possible to identify that launch sites use to be in
  - safety places, near sea, for example and have a good logistic infrastructure around.





Results

Predictive analysis results

Decision Tree Classifier Excels in Overall Accuracy:

- Achieved the highest accuracy on the test data, correctly predicting landing outcomes 94% of the time.
- Effectively captures complex relationships within the dataset, leading to robust performance.

Support Vector Machine (SVM) Classifier Demonstrates Strength in Handling Class Imbalance:

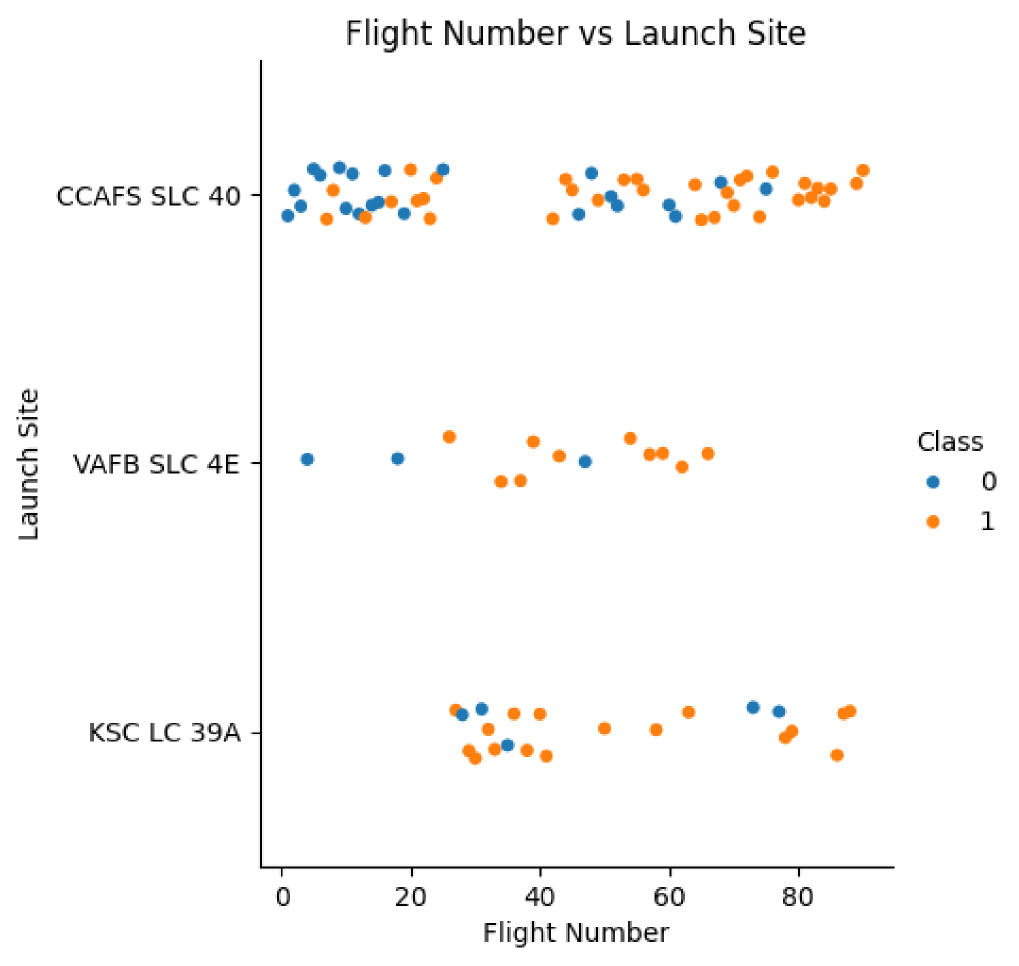
- Boasts the highest F1-score, indicating a superior balance between precision and recall, even when dealing with imbalanced classes (e.g., more successful landings than unsuccessful ones).
- Effectively differentiates between classes while minimizing both false positives and false negatives.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

1. Dominance of CCAF5 SLC 40:
   - This launch site stands out with a high concentration of successful launches in recent times, suggesting strong performance and potentially favorable conditions for launches.
2. VAFB SLC 4E and KSC LC 39A follow in terms of success rates, demonstrating their capability for reliable launches.
3. Overall, Success Rate Increase:
   - The upward trend in the plot across all sites indicates a general improvement in SpaceX's overall launch success rate over time.

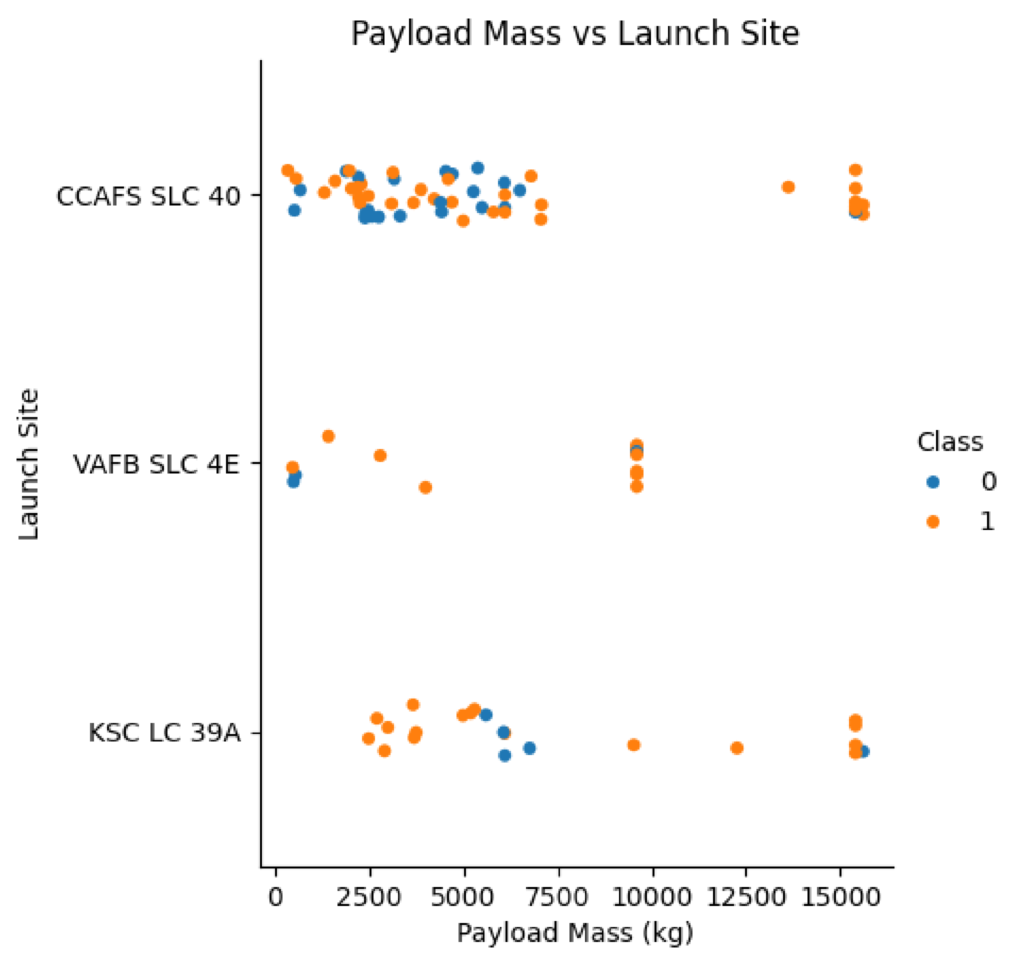


Payload vs. Launch Site

1. Success Rate and Payload Weight:
   - Payloads heavier than 9,000kg show a significantly higher success rate across all launch sites.

     This could be due to:

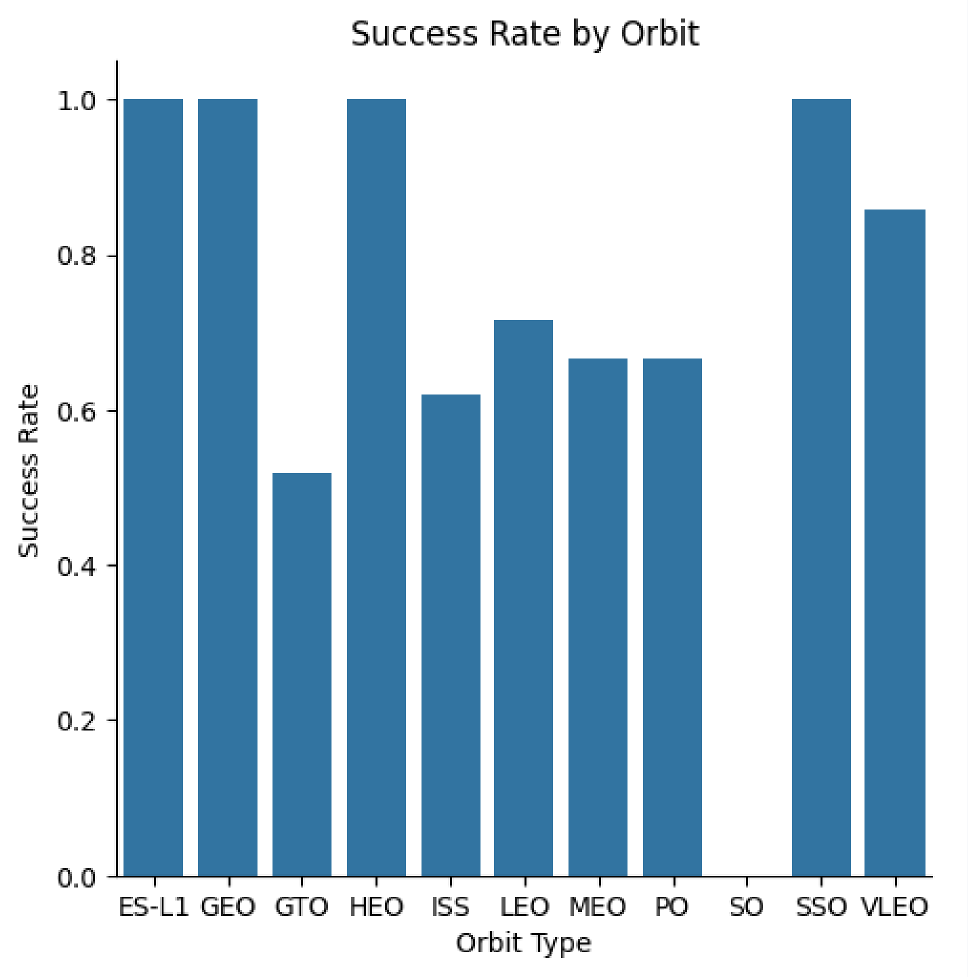
     - Optimized launch configurations for heavier payloads.
     - Improved capabilities and technologies used for handling heavier weights.
     - Statistical bias, with successful launches of heavier payloads attracting more attention or being featured more prominently in the data.
2. Payload Restrictions by Launch Site:
   - Launching payloads exceeding 12,000kg appears restricted to certain sites, notably CCAFS SLC 40 and KSC LC 39A.



Success Rate vs. Orbit Type

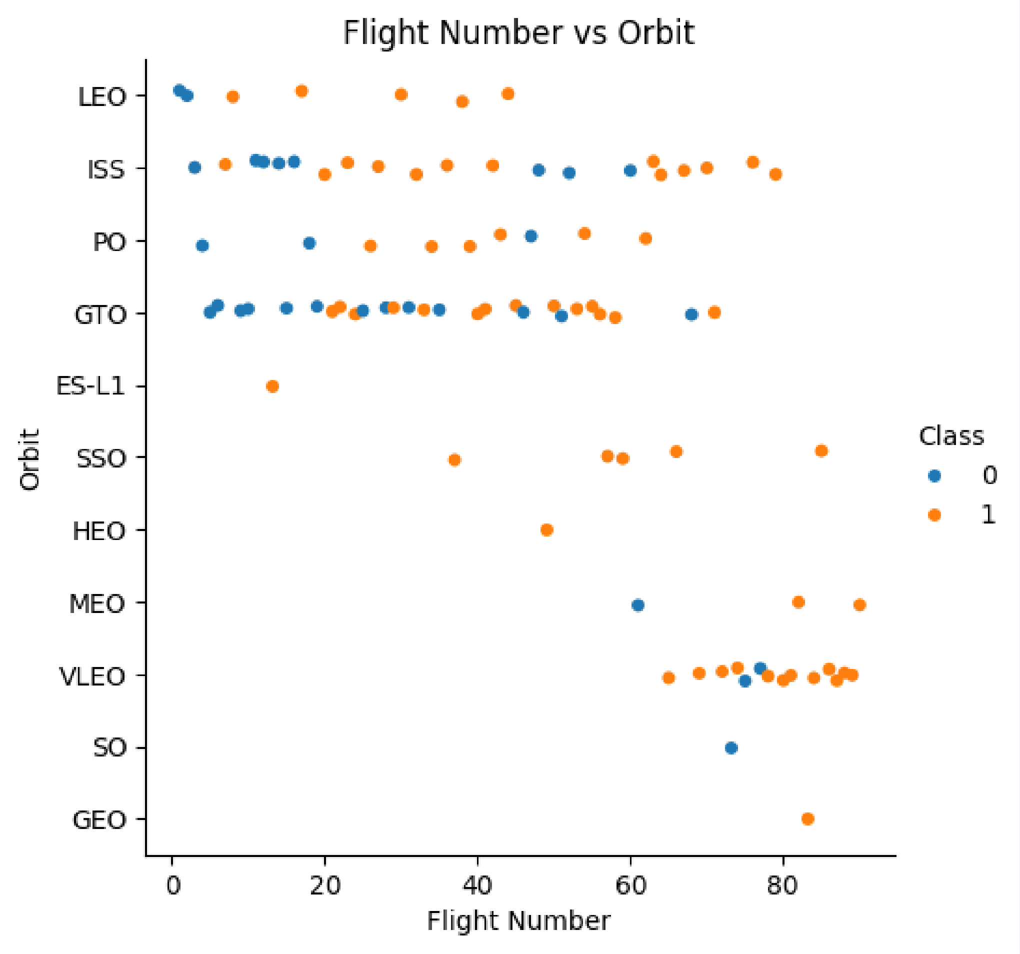
1. Orbits with the highest success rates include:
   - ES-LI (Earth-Space Low Inclination).
   - GEO (Geostationary)
   - HEO (Highly Elliptical Orbit)
   - SSO (Sun-Synchronous Orbit)

2. Mid-Tier Success:
   - Orbits like VLEO (Very Low Earth Orbit) and LEO (Low Earth Orbit) exhibit good success rates, exceeding 80% and 70%, respectively.



Flight Number vs. Orbit Type

1. Overall, Success Rate Improvement:
   - The general upward trend across all orbit types suggests a consistent improvement in SpaceX's launch success rate over time.
   - This aligns with earlier observations pointing towards technological advancements, operational refinements, and accumulated experience contributing to overall success
2. VLEO as a Potential Business Opportunity:
   - The increasing frequency of launches to VLEO (Very Low Earth Orbit) signifies a growing interest in this emerging orbital region.
   - This could be driven by:
     - Growing demand for applications like satellite constellations for internet connectivity, Earth observation, and scientific research.

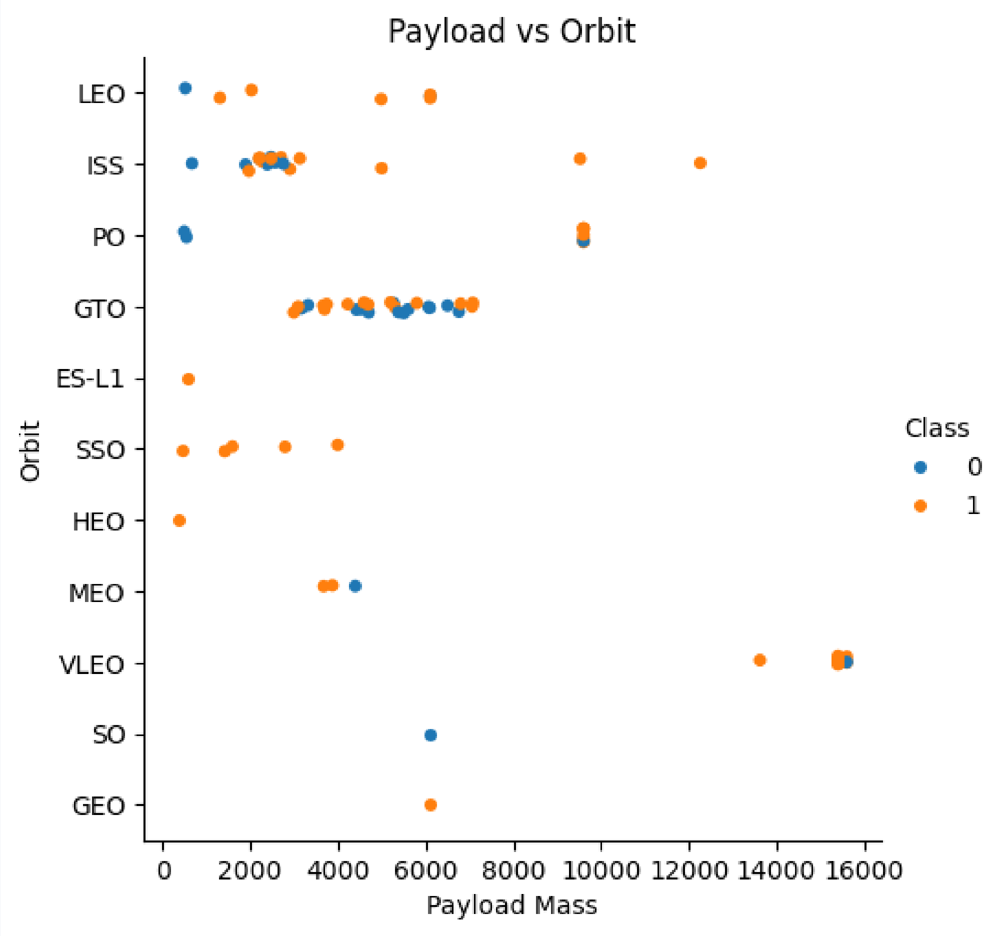


Payload vs. Orbit Type

1. No Correlation for GTO:
   - There is no apparent relationship between payload size and success rate for launches to GTO (Geostationary Transfer Orbit).

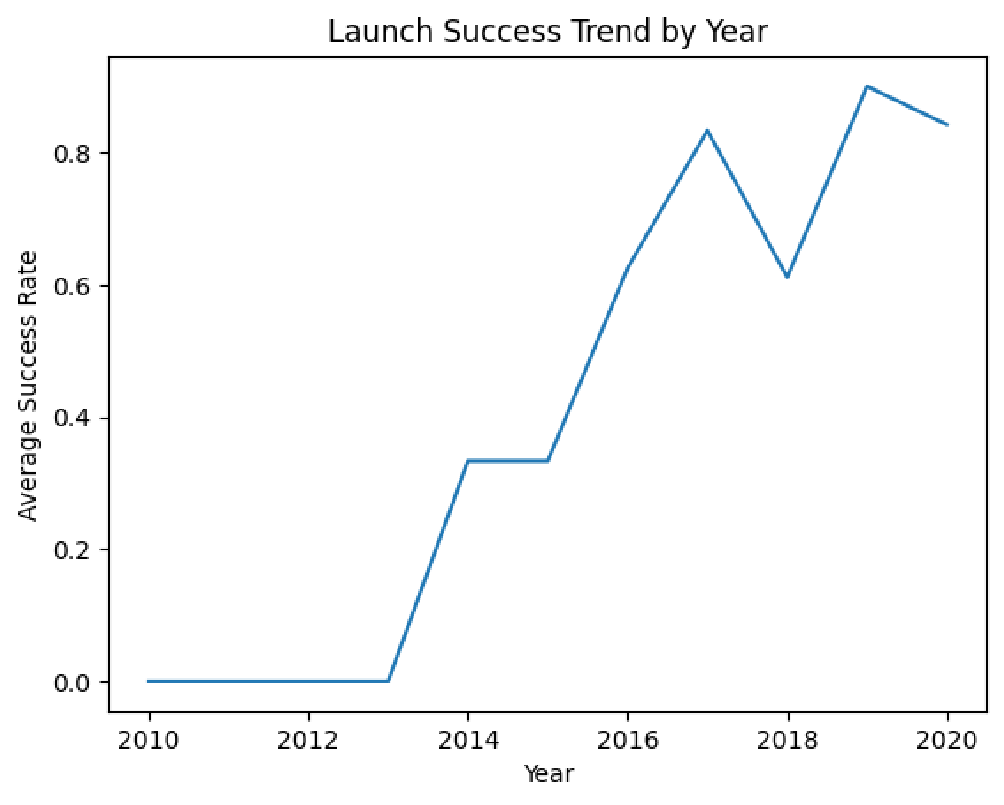
     This could indicate::

     - SpaceX utilizes similar launch configurations and procedures for payloads of various sizes destined for GTO, regardless of individual weight.
     - Any potential influence of payload weight on GTO launch success might be masked by other factors.
2. ISS with Wide Payload Range and Good Success:
   - This suggests successful launches of diverse payload sizes to the International Space Station (ISS) orbit, highlighting SpaceX's flexibility and capability in handling varied orbital cargo.



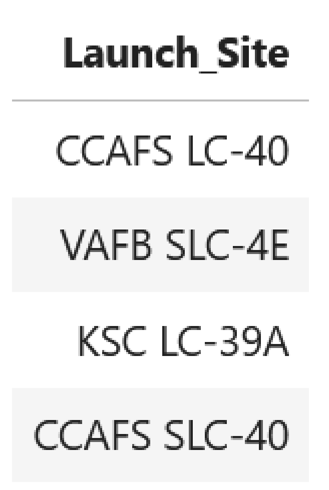
Launch Success Yearly Trend

1. **Overall Upward Trend:**
   - The line chart clearly shows a general increase in success rate from 2013 onwards, reaching a peak in 2020.
2. **Temporary Dip in 2017-2018:** The noticeable dip in success rate between 2017 and 2018 warrants further investigation.
3. **Overall Upward Trend:** The first three years (all failures) as a period of adjustments and technology improvement.



All Launch Site Names

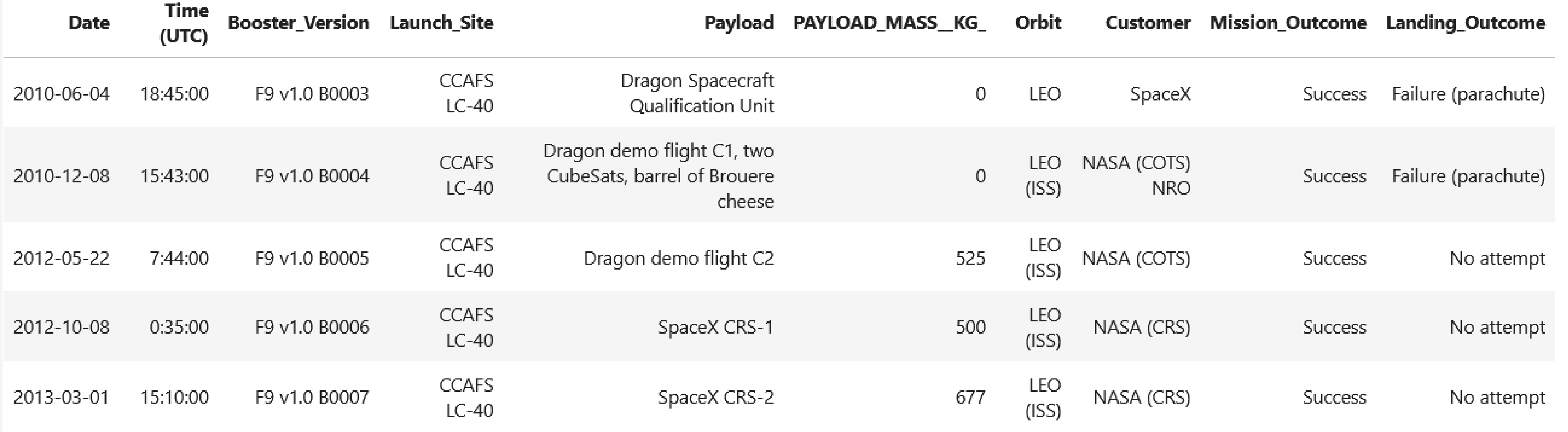
- According to data, there are four launch sites:



Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

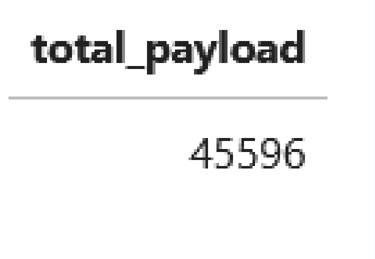


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Here we can see five samples of Cape Canaveral launches.

Total Payload Mass

- Total payload carried by boosters from NASA

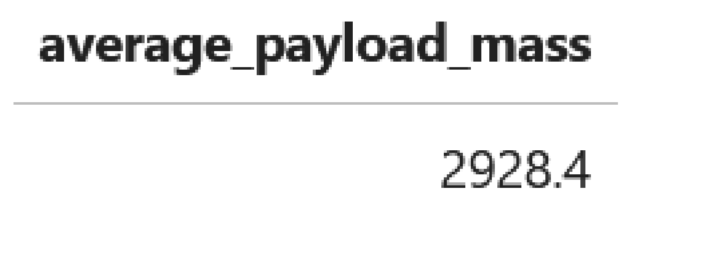


total_payload
45596

- Total payload calculated above, by summing all payloads whose codes contain 'NASA (CRS)'.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1

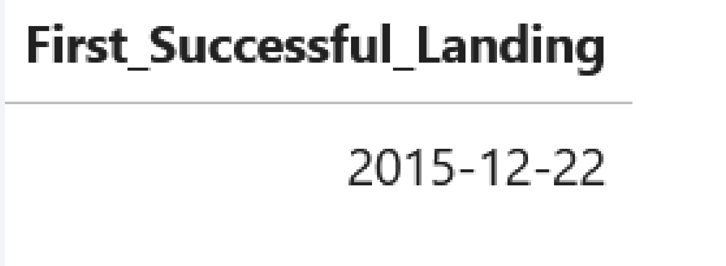


average_payload_mass
2928.4

- The value was obtained by filtering the data by the booster version (F9 v1.1) and calculating the average.

First Successful Ground Landing Date

- The first successful landing outcome on ground pad

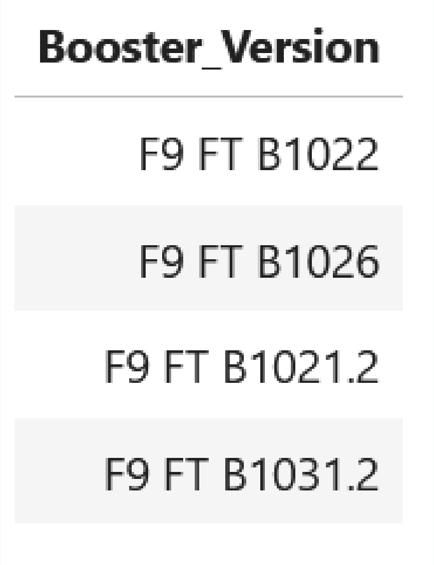


First_Successful_Landing
2015-12-22

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

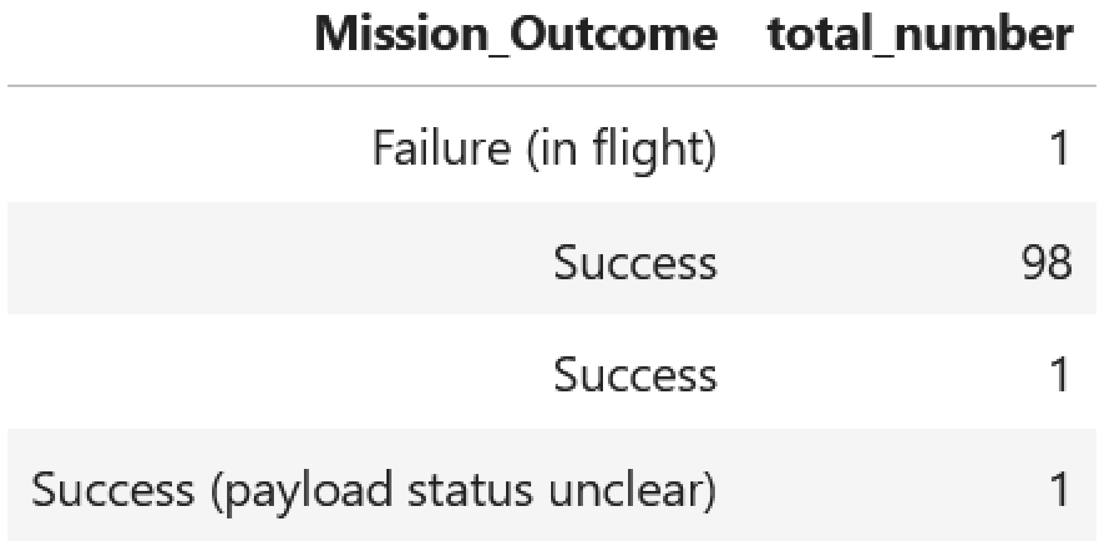


Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- Selecting distinct booster versions with Payload between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

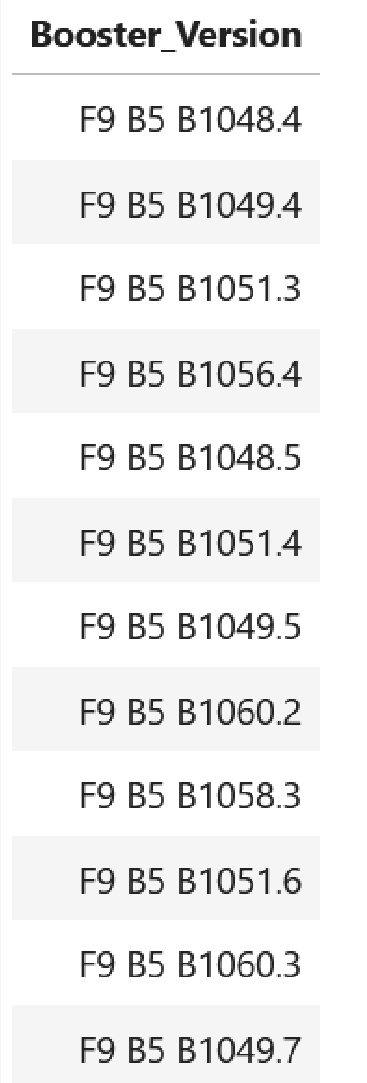


Mission_Outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- rouping mission outcomes and counting records for each group.

Boosters Carried Maximum Payload

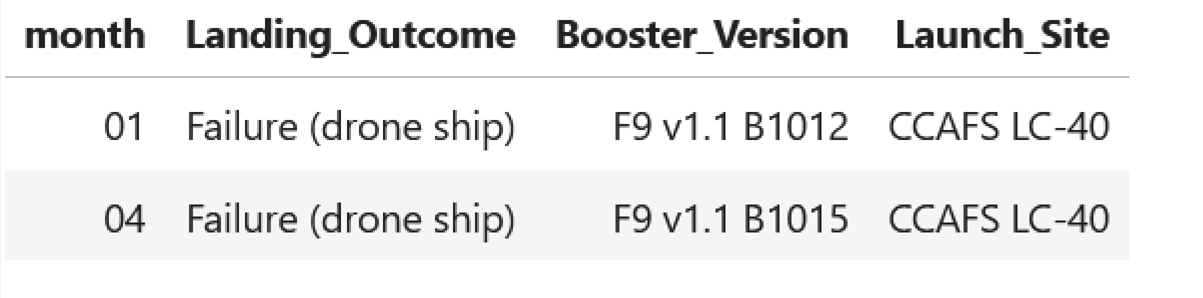
- Boosters which have carried the maximum payload mass



Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

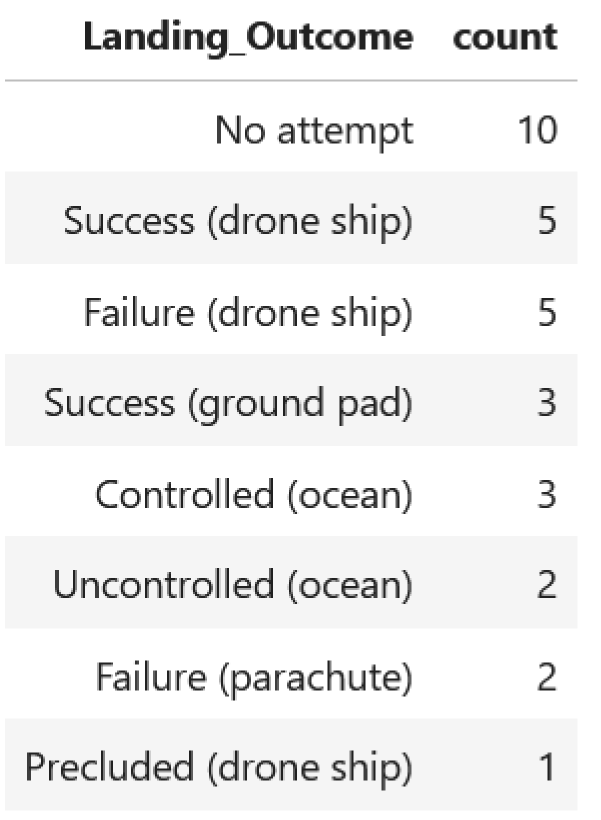
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015



month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	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

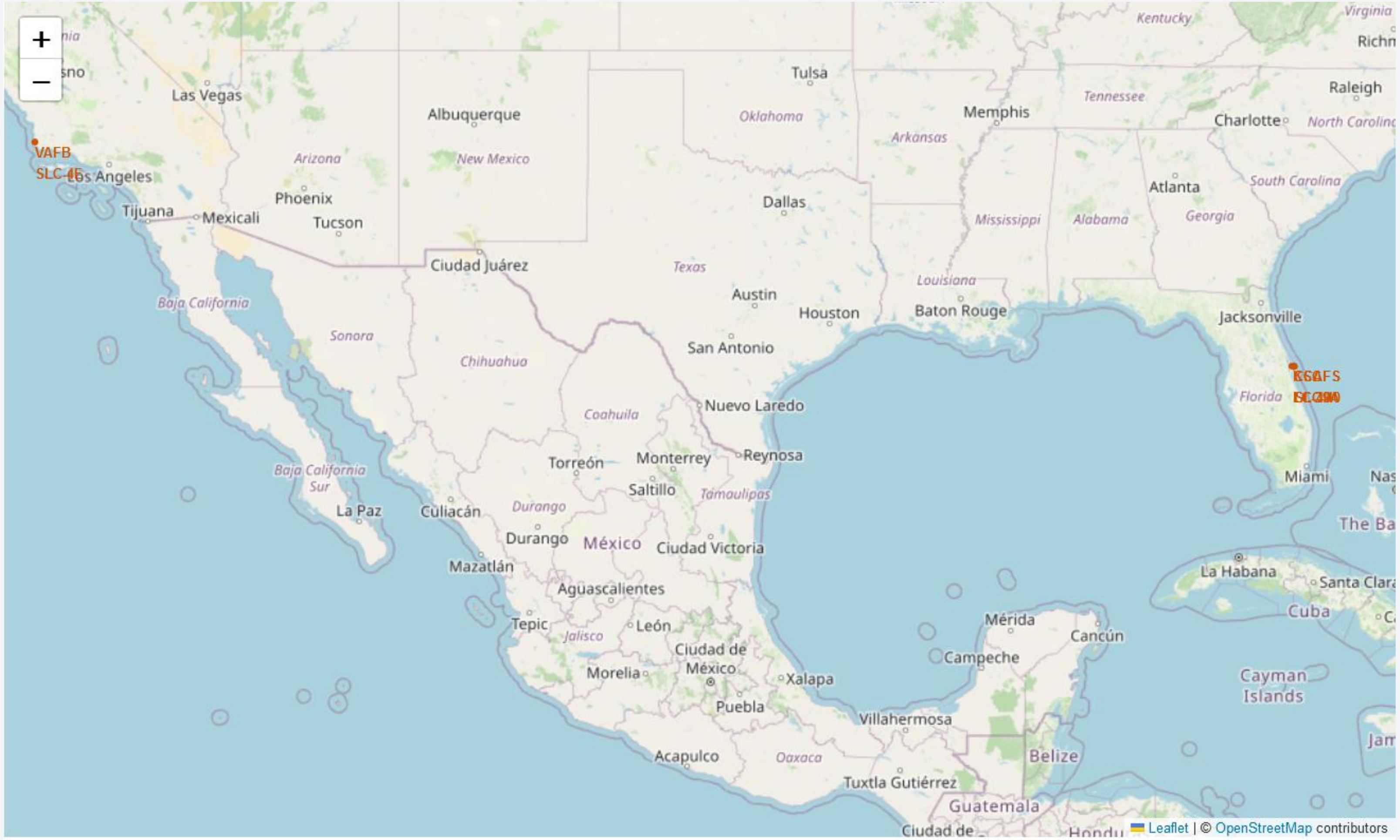


Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

All launch sites



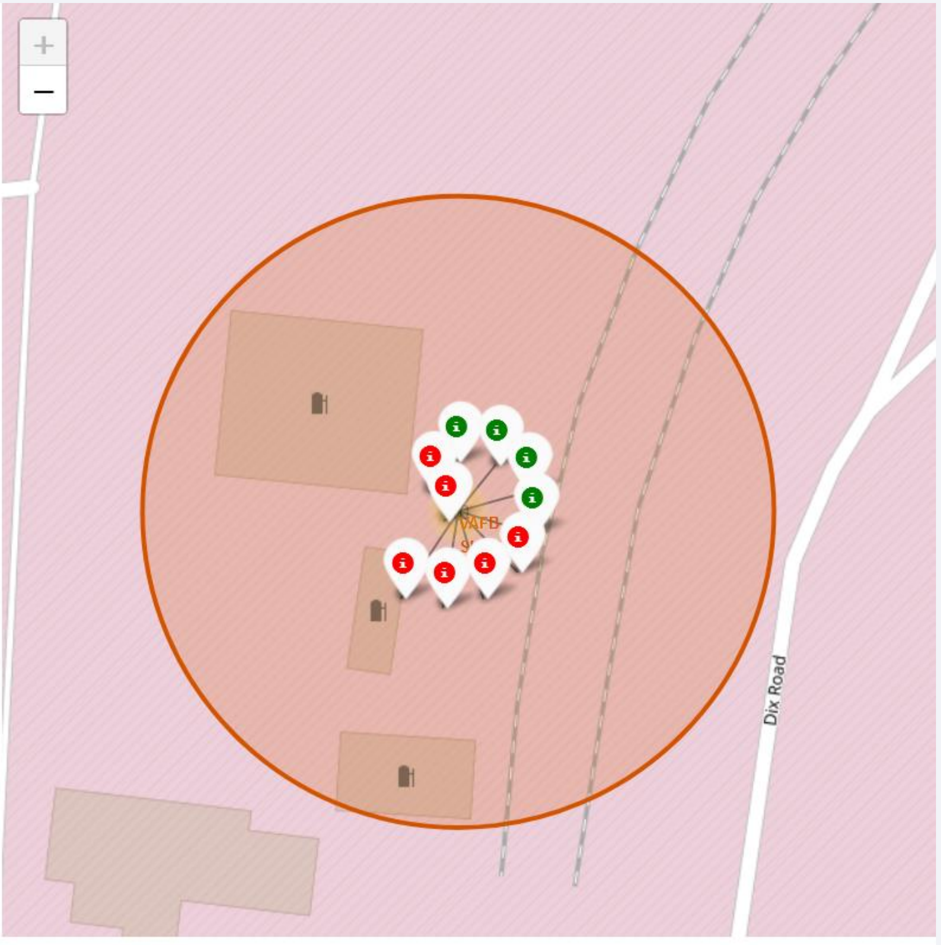
- Launch sites are near sea, but not too far from roads and railroads.

Color-labeled launch records

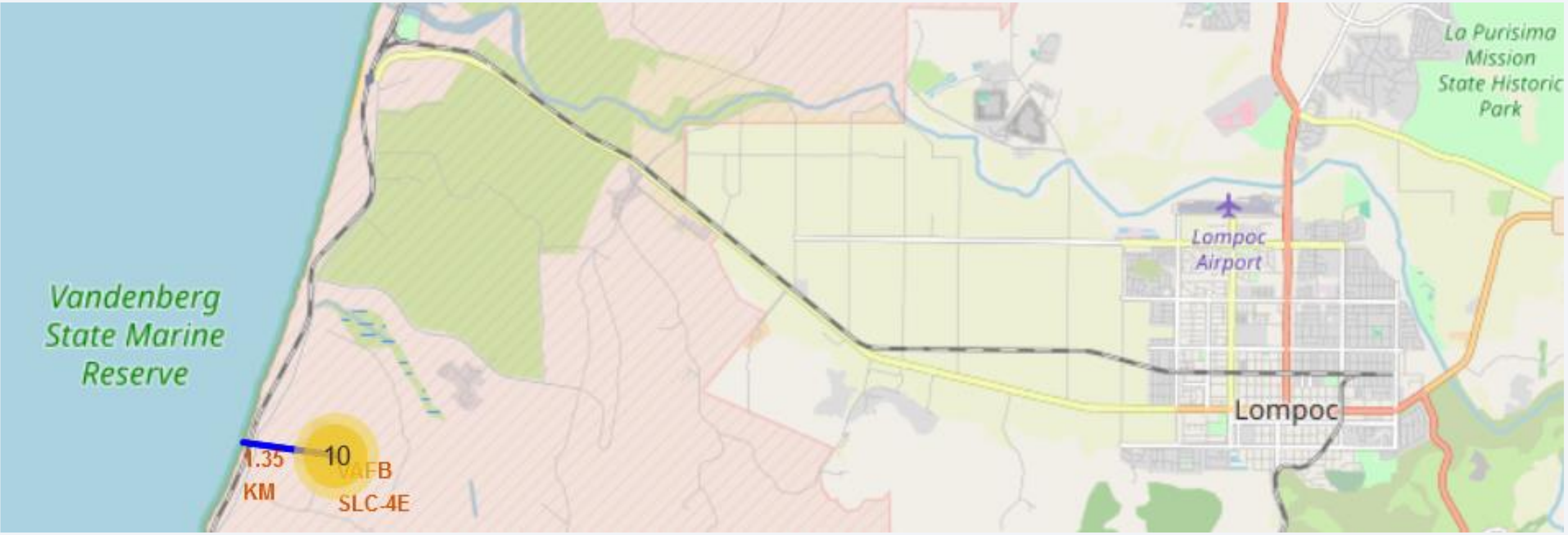
Explanation:

From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

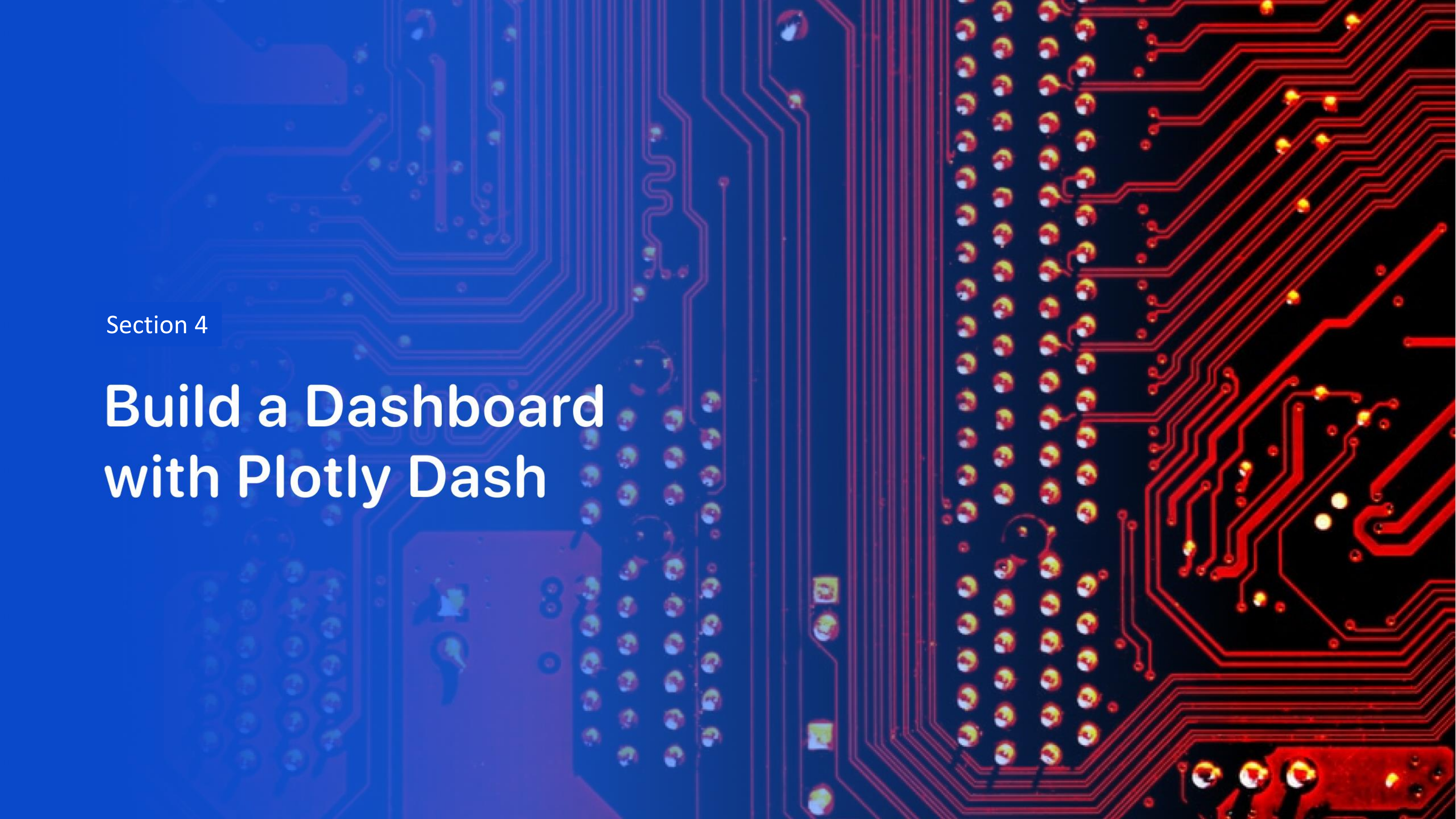
- Green Marker = Successful Launch
- Red Marker = Failed Launch
- Launch Site VAFB SLC-4E has a very low Success Rate.



Distance from the launch site VAFB SLC-4E to its proximities


La Purisima Mission State Historic Park
Lompoc Airport
Vandenberg State Marine Reserve
Lompoc
1.35 KM
10
VAFB SLC-4E

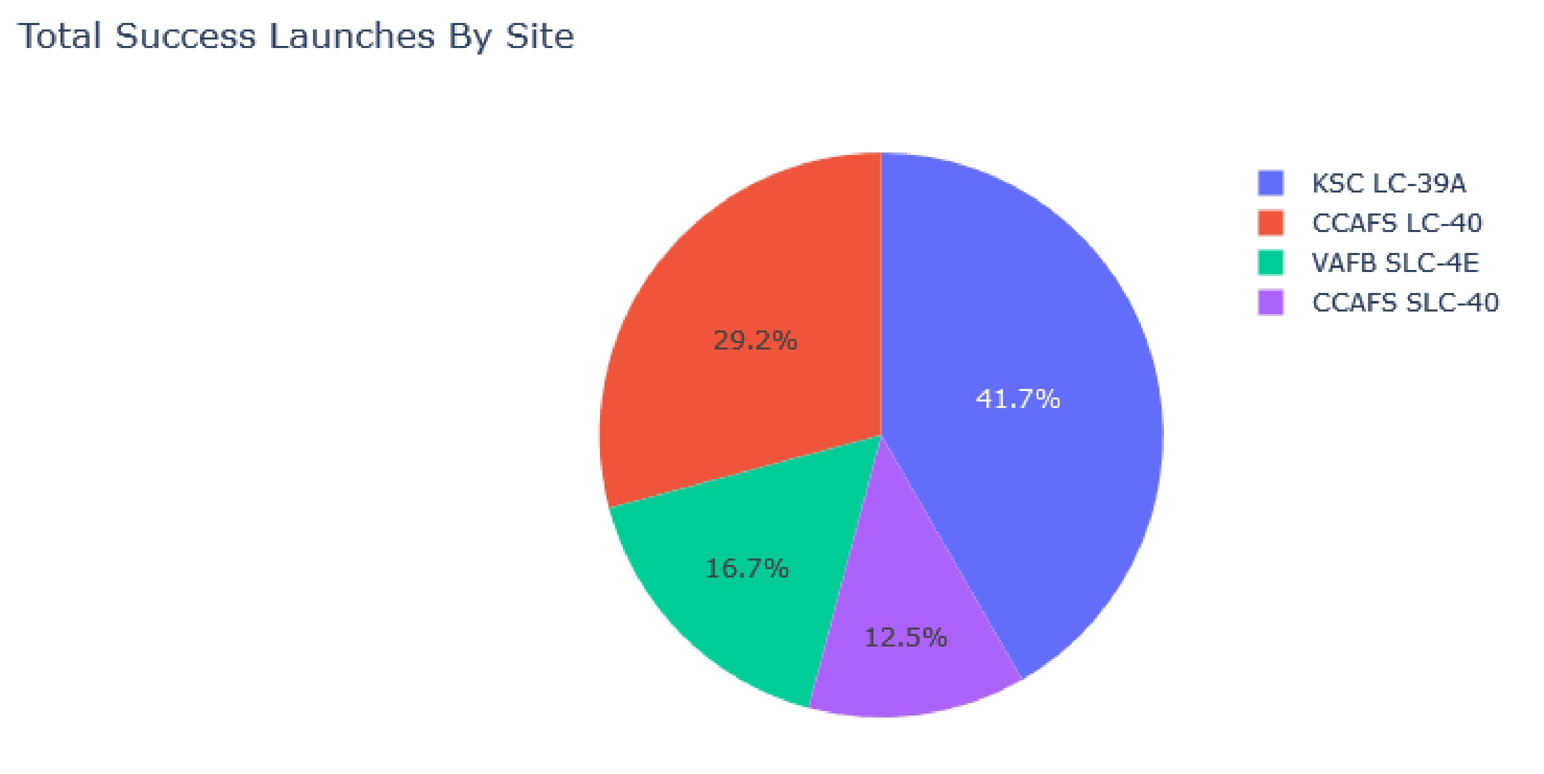
- Launch site VAFB SLC-4E has good logistics aspects, being near coastline, railroad and road and relatively far from inhabited areas.



Section 4

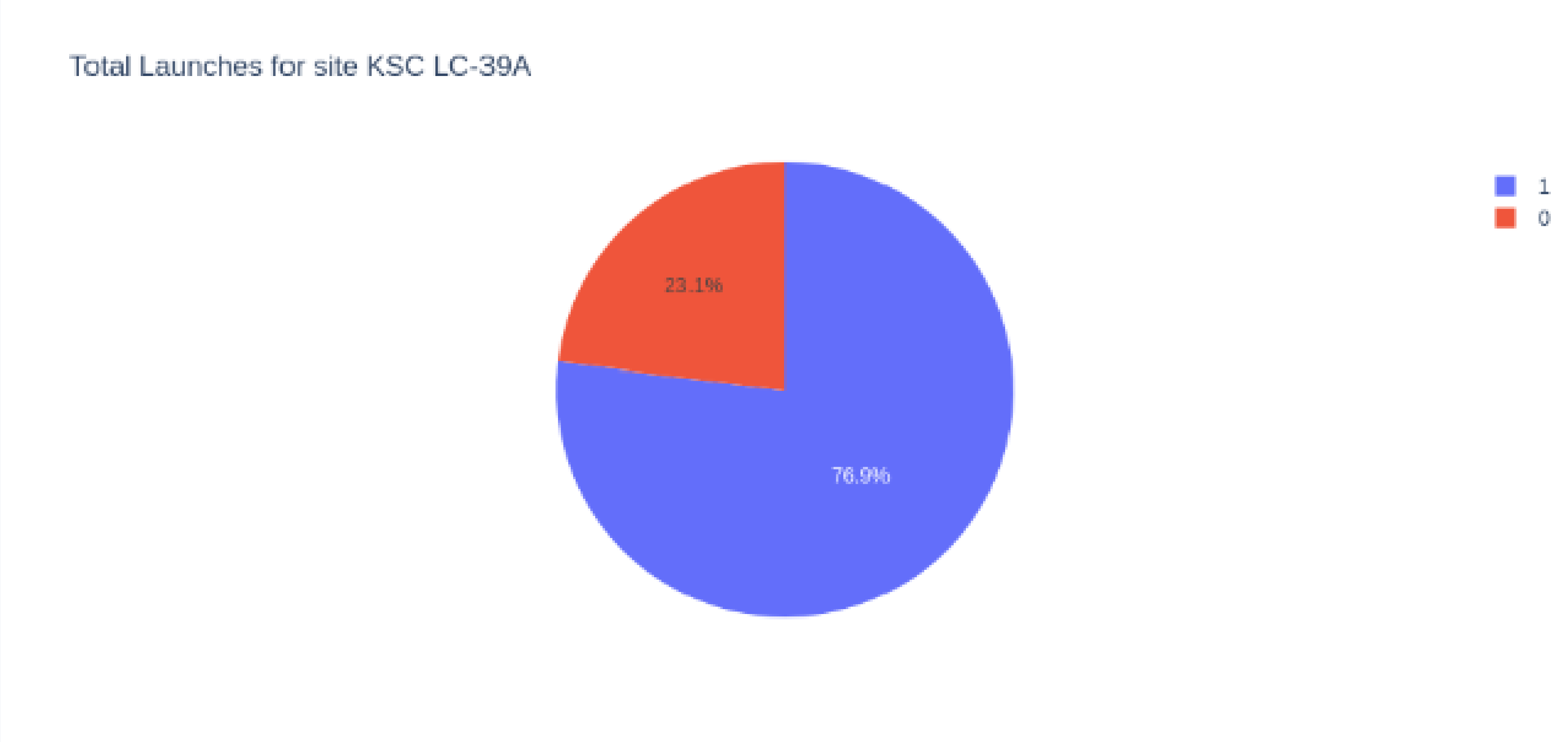
Build a Dashboard with Plotly Dash

Successful Launches by Site



- The place from where launches are done seems to be a very important factor of success of missions, with KSC LC-39A having the most successful launches.

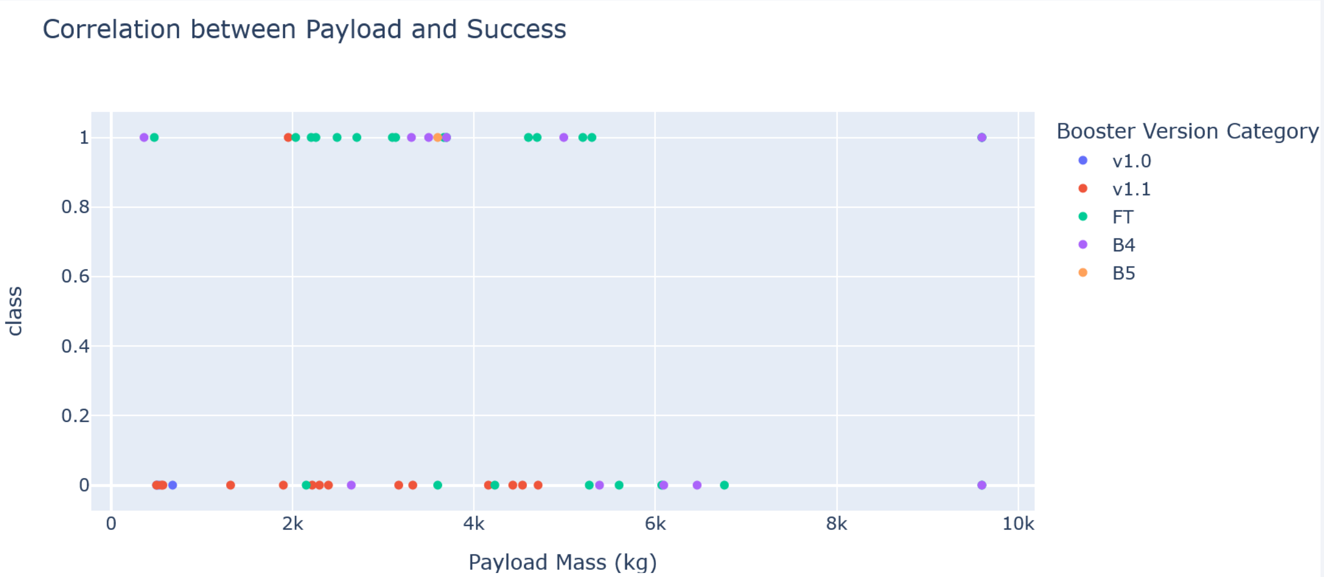
Launch site with highest launch success ratio

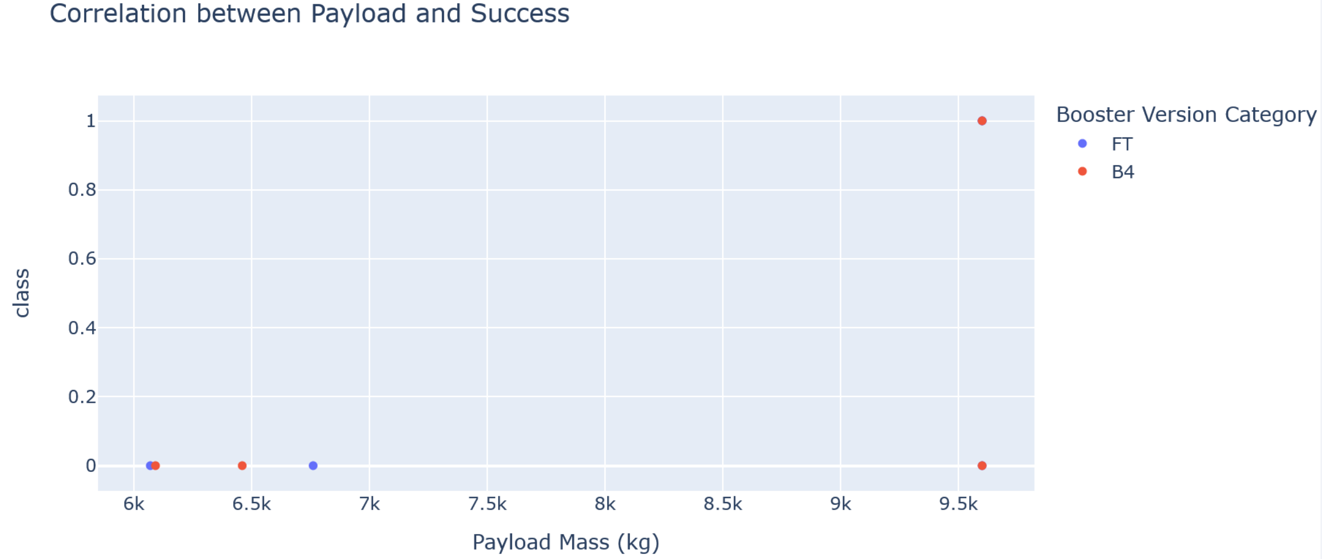


- KSC LC-39A has the highest launch success rate with 10 successful and only 3 failed landings.

<Dashboard Screenshot 3>

- Payloads between 2000 and 5500 kg have the highest success rate.
- There's not enough data to estimate risk of launches over 7,000kg


Correlation between Payload and Success

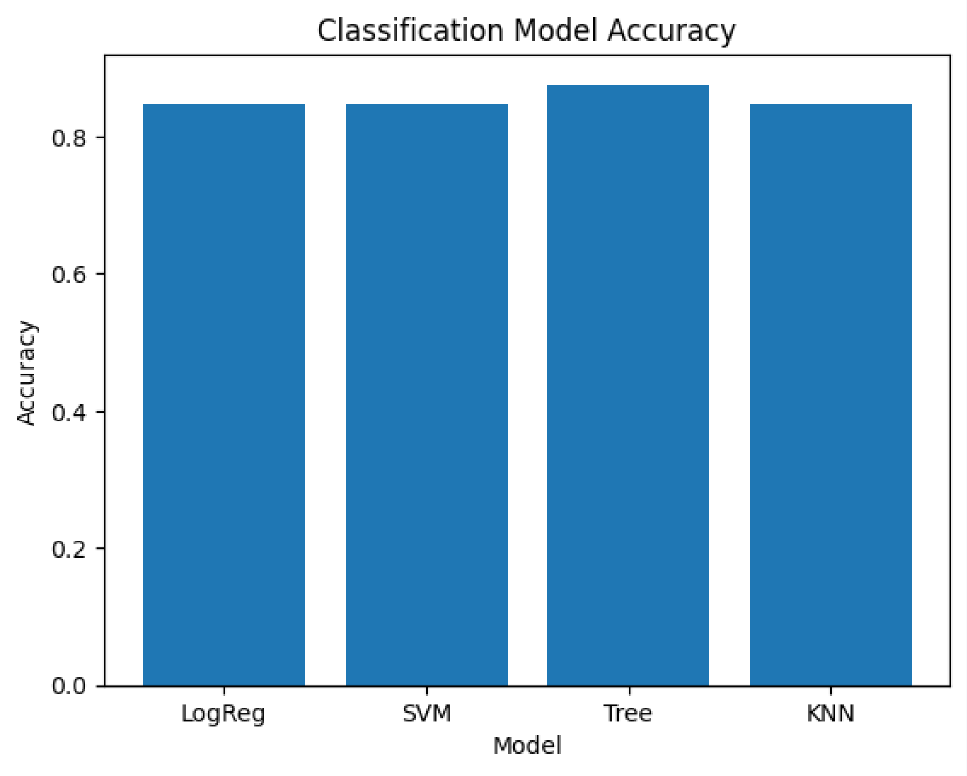

Correlation between Payload and Success

Section 5

Predictive Analysis (Classification)

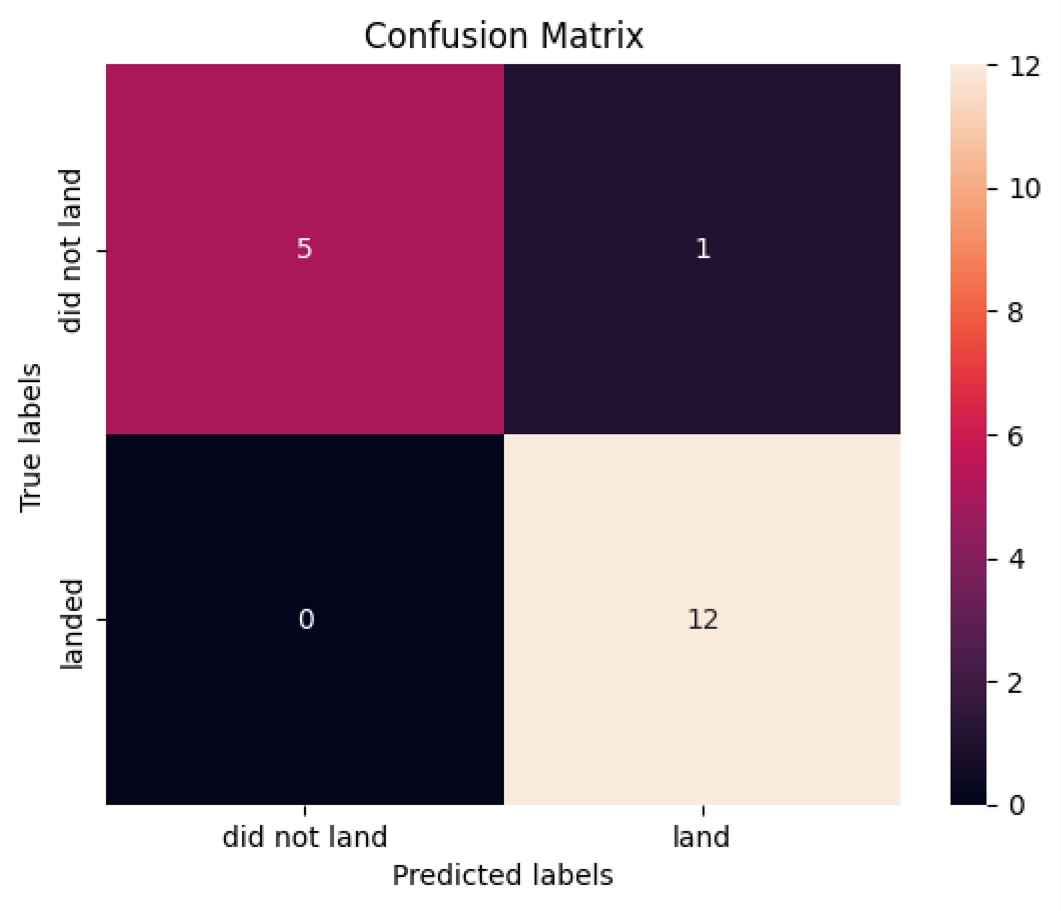
Classification Accuracy

- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 94%.



Confusion Matrix

- High Accuracy: The matrix demonstrates the Decision Tree Classifier's strong performance.
- Excellent Recall: Notably, the model correctly identified all 12 positive cases (True Positives), resulting in a perfect recall of 100%. This means it didn't miss any actual positive instances.
- Solid Precision: The absence of any False Positives (0 FP) indicates a precision of 100%, ensuring that all instances classified as positive were genuinely positive.
- Minor False Negative: While the model correctly identified most negative cases (5 True Negatives), one False Negative (FN) occurred, suggesting a slight under-prediction of negative instances.



Conclusions

Data Analysis:

- Launch sites primarily concentrate near the equator and coastline.
- **Best Launch Site:** KSC LC-39A exhibits the highest success rate.
- **Payload Weight and Risk:** Launching payloads between 2000 and 5500 kg appears less risky.
- Certain orbits like ES-LI, GEO, HEO, and SSO have achieved the highest success rates.
- **Landing Outcomes:** Success rates for landing appear to be rising over time, potentially due to process and rocket advancements.

Predictive Modeling:

- Decision Tree Classifier emerged as the optimal model for predicting successful landings, potentially leading to increased profits.

Thank you!