IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anghelo Salirrosas
14/07/2025

Outline

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

Executive Summary

- This project focuses on the data analysis of SpaceX launches in order to identify patterns and make predictions about mission success. Various data science techniques were used: since data collection from APIs and web scraping, exploratory analysis with SQL, data cleansing and data visualizations through interactive maps, dashboards and finally a comparative of several predictive models for classification.
- The analysis showed that KSC LC-39A had the highest success rate among launch sites. Orbits like ES-L1 and GEO had near-perfect success, while GTO had more failures. Payloads between 4000–6000 kg and booster versions FT and B5 were most consistently associated with success. The dashboard and interactive maps effectively revealed spatial and categorical trends. All four machine learning models (Logistic Regression, SVM, Decision Tree, KNN) achieved similar performance, with ~83% accuracy in predicting landing success.

Introduction

SpaceX has conducted hundreds of launches, with success being critical for its commercial and scientific goals. Understanding what factors contribute to mission success can support decision-making and improve future outcomes.

This project aims to:

- Analyze Falcon 9 launch data to identify patterns and trends.
- Visualize launch outcomes geographically and temporally.
- Build predictive models to estimate the probability of a successful landing based on mission features.

Section 1

Methodology

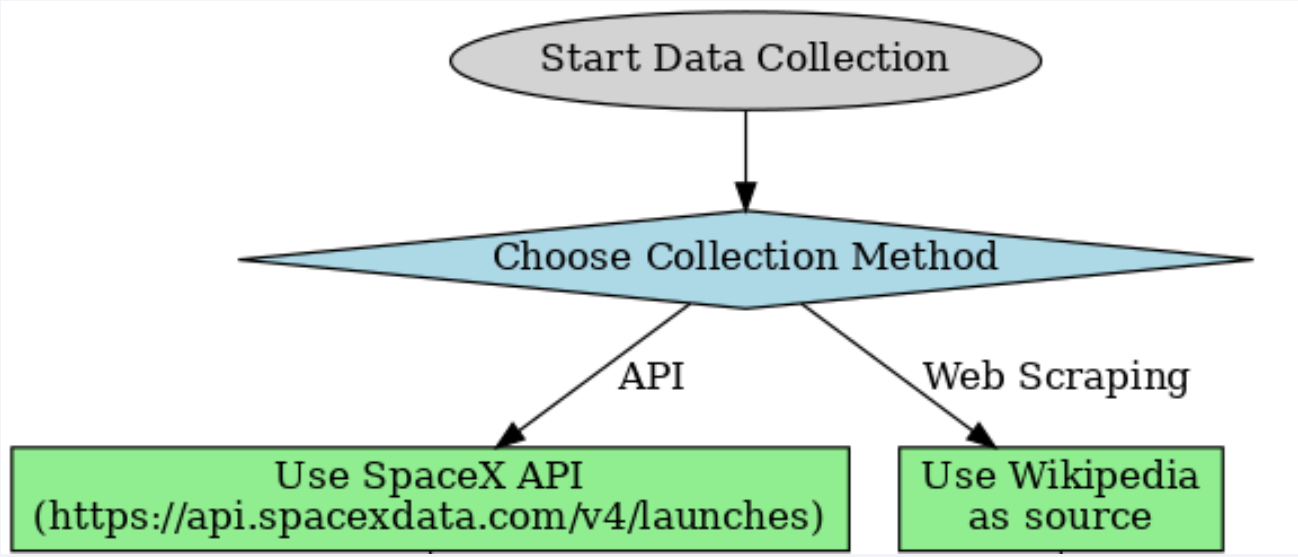
Methodology

Executive Summary

- Data collection methodology:
  - The data was extracted using the SpaceX REST API and it was used web scraping from Wikipedia.
- Perform data wrangling
  - The data was preprocessed to be able to analyze mission success numerically.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four predictive models were built.
  - The data was separated into training and testing groups.
  - The predictive models were evaluated using cross validation.

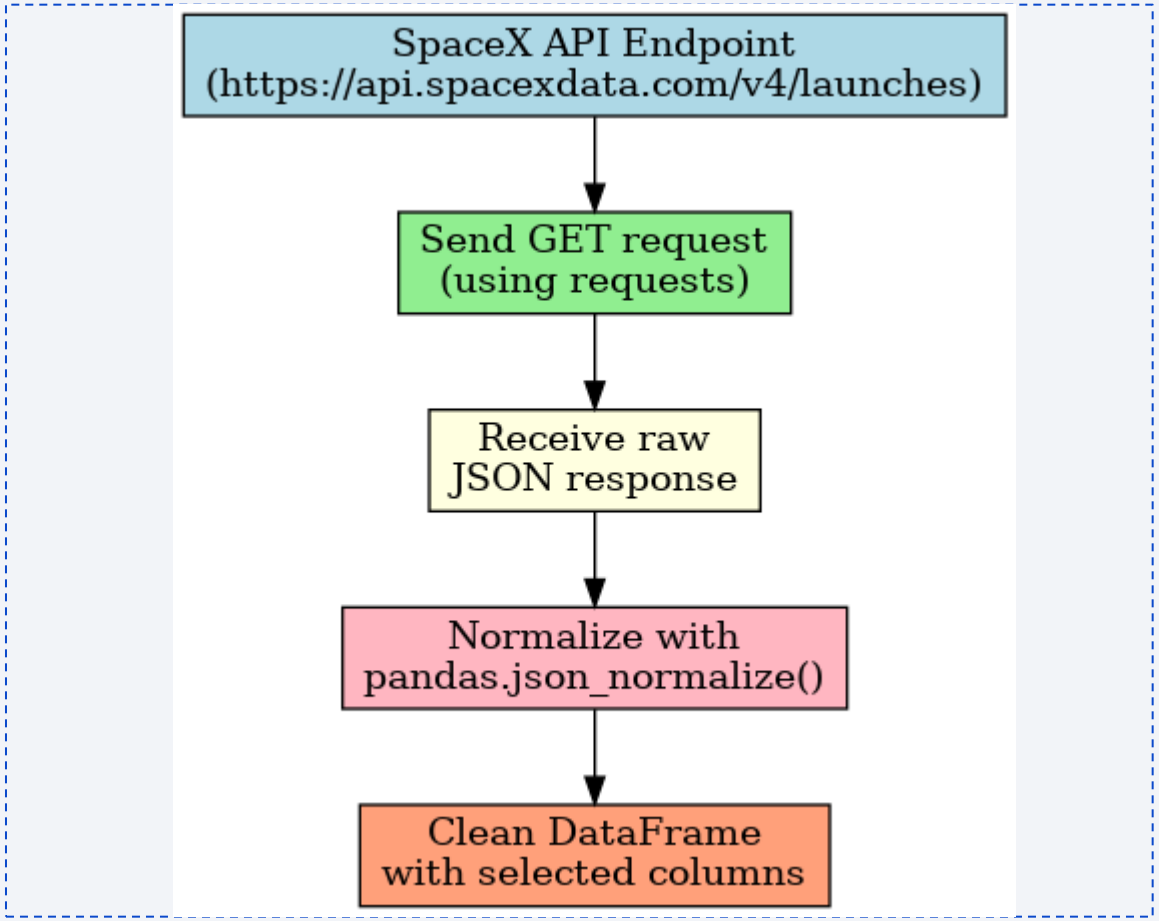
Data Collection

- We have to ways to collect data, for the first one we used the SpaceX API and for the second one we used web scraping from Wikipedia. In order to use SpaceX API we extracted the data from a JSON file and normalized it into a DataFrame. For web scraping we extracted data tables from Wikipedia in HTML format and also converted them into a Dataframe.



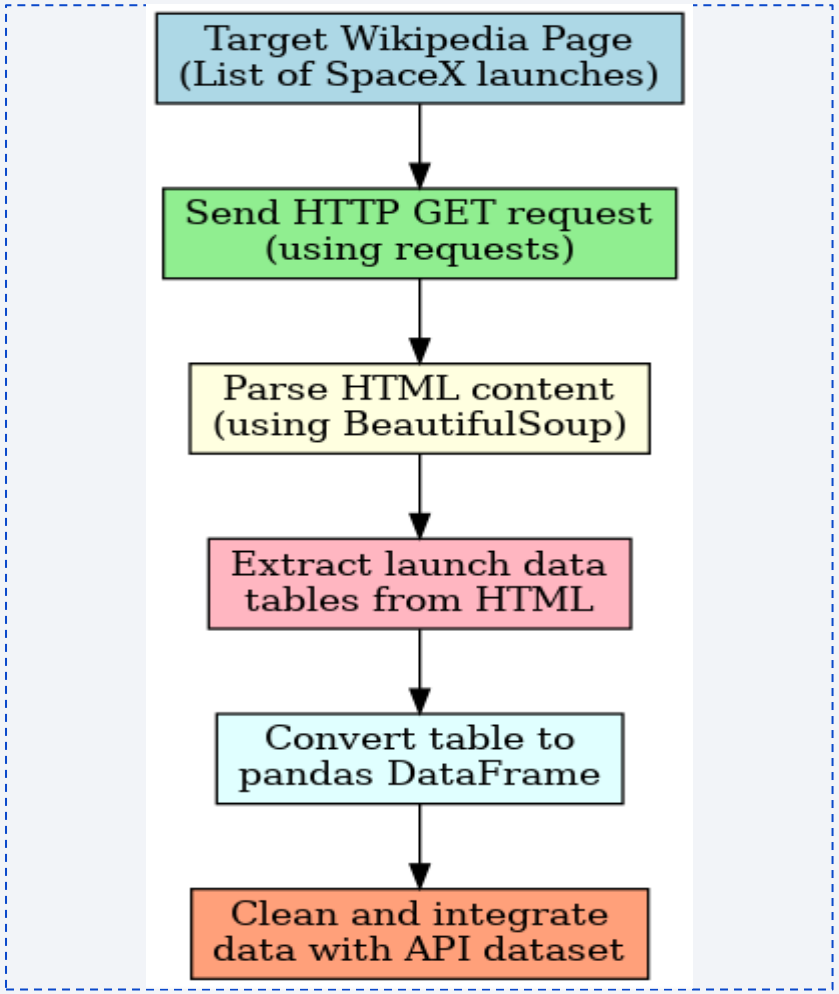
Data Collection – SpaceX API

- The launch data was collected from the official SpaceX REST API using a GET from **request** library and normalize it to get a **pandas** DataFrame. Then clean the DataFrame and select the data keeping only features we need from the launches and select only launches from Falcon 9 version, making sure there are no missing values.
- https://github.com/Anghelo-Salirrosas/DS-Repository/blob/2bcc512975eab9fe4f3f3351d02238429f01704c/1.1.DataCollectionAPI.ipynb



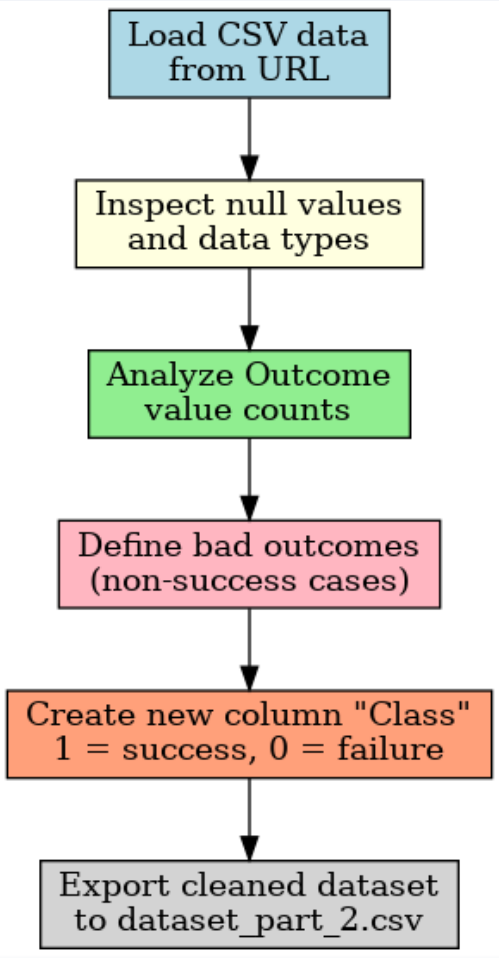
Data Collection - Scraping

- The launch data was also collected from Wikipedia using web scraping. An HTTP GET request was sent using the requests library to retrieve the HTML content of the page. Then, BeautifulSoup was used to parse the HTML and extract the launch table. The extracted data was converted into a pandas DataFrame, cleaned to remove irrelevant or malformed entries, and matched with the Falcon 9 launch records. Missing values were handled and only relevant columns were kept for further analysis.
- https://github.com/Anghelo-Salirrosas/DS-Repository/blob/2bcc512975eab9fe4f3f3351d02238429f01704c/1.2.DataCollectWithWebScrapping.ipynb



Data Wrangling

- The SpaceX launch data collected was cleaned and transformed using pandas and converted into a CSV file. The CSV dataset was loaded and explored using pandas. A new binary classification column (Class) was created by mapping launch outcomes to 1 (success) or 0 (failure) based on predefined failure conditions. The cleaned and labeled data was saved for further analysis.
- https://github.com/Anghelo-Salirrosas/DS-Repository/blob/2bcc512975eab9fe4f3f3351d02238429f01704c/1.3.DatWrangling.ipynb



EDA with Data Visualization

A visual exploratory analysis was performed using statistical graphs to study Falcon 9 launch performance.

- Flight Number vs. Payload Mass (Catplot):To explore whether heavier payloads are associated with mission outcomes and how performance evolves over time.
- Flight Number vs. Launch Site (Catplot):To observe if certain launch sites have higher frequencies of successful missions over time.
- Payload Mass vs. Launch Site (Catplot):To analyze which sites support heavier payloads and if that affects success rates.
- Success Rate by Orbit (Barplot):To compare the average success rate across different orbit types.
- Flight Number vs. Orbit (Catplot):To examine how orbit types are distributed over time and their impact on success.
- Payload Mass vs. Orbit (Catplot):To study how different payload sizes perform across various orbit types.
- Yearly Launch Success Rate (Lineplot):To identify trends in success rates over the years and see if SpaceX has improved.

https://github.com/Anghelo-Salirrosas/DS-Repository/blob/2bcc512975eab9fe4f3f3351d02238429f01704c/2.2.EDAwithVisualization.ipynb

EDA with SQL

An exploratory analysis of the data obtained was carried out to obtain some information.

- Retrieved unique launch sites using DISTINCT(Launch_Site).
- Filtered 5 records where launch sites start with 'CCA'.
- Calculated total payload mass carried by NASA (CRS) missions.
- Computed average payload mass for booster version 'F9 v1.1'.
- Identified the date of the first successful ground pad landing.
- Selected booster versions that landed successfully on drone ship with payload between 4000 and 6000 kg.
- Counted mission outcomes (successes and failures).
- Found booster versions that carried the maximum payload mass.
- Retrieved failed drone ship landings in 2015, including booster version, launch site, and month.
- Ranked landing outcomes (Success (ground pad) and Failure (drone ship)) between 2010-06-04 and 2017-03-20.

https://github.com/Anghelo-Salirrosas/DS-Repository/blob/2bcc512975eab9fe4f3f3351d02238429f01704c/2.1.EDAwithSQL.ipynb

Build an Interactive Map with Folium

- **Circles**: Red circles were added at the coordinates of each launch site and at NASA's Johnson Space Center.
  Purpose: To highlight each launch site's location with a popup label for easy identification.
- **Text Markers** (DivIcon):Custom markers using DivIcon were placed at each launch site and NASA JSC to display site names directly on the map.
  Purpose: To provide a clear, always-visible label without needing to click.
- **Clustered Markers**: A MarkerCluster was created for all historical launch points. Marker color represents mission outcome: green for success, red for failure.
  Purpose: To visualize launch outcome distribution and manage overlapping markers in dense areas.
- **Popups:** Each marker includes a popup with the launch site name and mission result (e.g., "VAFB SLC-4E Success").
  Purpose: To give immediate context about the launch upon clicking a marker.
- **Mouse Position Tool**: A dynamic mouse position feature displays real-time latitude and longitude when hovering over the map.
  Purpose: To enable quick extraction of coordinates and improve geographic analysis.

https://github.com/Anghelo-Salirrosas/DS-Repository/blob/2bcc512975eab9fe4f3f3351d02238429f01704c/3.1.VisualAnalyticswithFolium.ipynb

Build a Dashboard with Plotly Dash

For the dashboard 2 selection inputs and 2 output graphs were built.

- Launch Site Dropdown (Dropdown menu)
  Allows users to select either all launch sites or a specific one. To filter the visualizations based on the user's interest in a particular site.

- Success Pie Chart (dcc.Graph):
  - Total successful launches by site (when "All Sites" is selected)
  - Success vs. failure count for the selected site.
  To provide an overview of mission outcomes per site and identify which sites perform best.

- Payload Range Slider (RangeSlider)
  Allows users to select a range of payload mass (0–10,000 kg). To analyze how payload size correlates with mission success.

- Scatter Plot of Payload vs. Outcome (dcc.Graph)
  Plots payload mass on the x-axis and mission outcome (success/failure) on the y-axis. Points are color-coded by booster version. To visually explore the relationship between payload mass, booster version, and success rate.

https://github.com/Anghelo-Salirrosas/DS-Repository/blob/2bcc512975eab9fe4f3f3351d02238429f01704c/3.2.spacex-dash-app.py

Predictive Analysis (Classification)

Predictive analysis was separated in three phases:

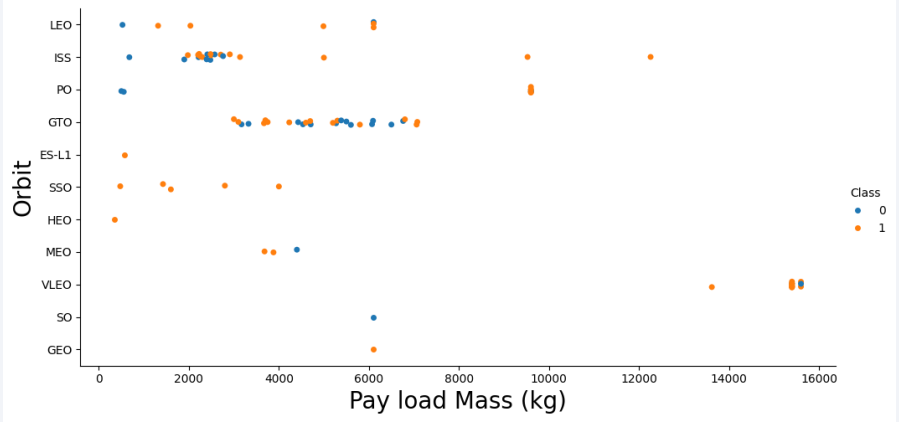
- Data Preparation:
  - Loaded the cleaned target labels from dataset_part_2.csv and the one-hot encoded features from dataset_part_3.csv.
  - Standardized the feature data using StandardScaler.
  - Split the dataset into training and testing sets (80/20 split).
- Model Building & Evaluation:
  - Built and tuned four classification models using GridSearchCV with 10-fold cross-validation:
    - Logistic Regression
    - Support Vector Machine (SVM)
    - Decision Tree
    - K-Nearest Neighbors (KNN)

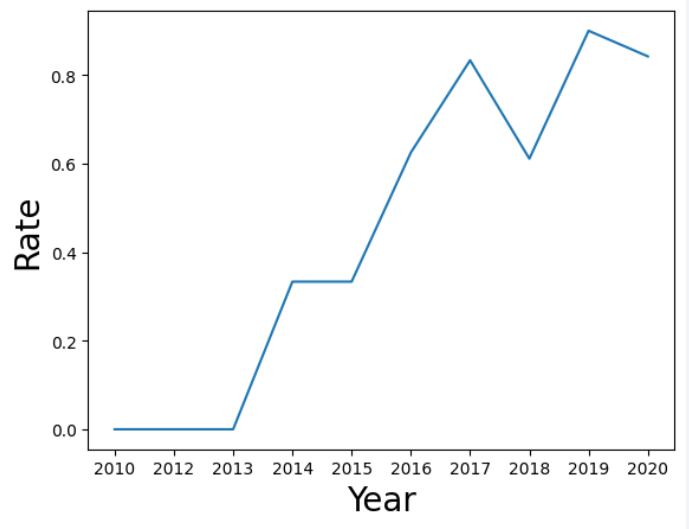
    For each model:Best hyperparameters were selected, accuracy on the training and test sets was computed and a confusion matrix was plotted to assess performance visually.
- Model Comparison:
  Accuracy scores from all four models were compared.The best performing model was selected based on highest test accuracy and balanced confusion matrix.

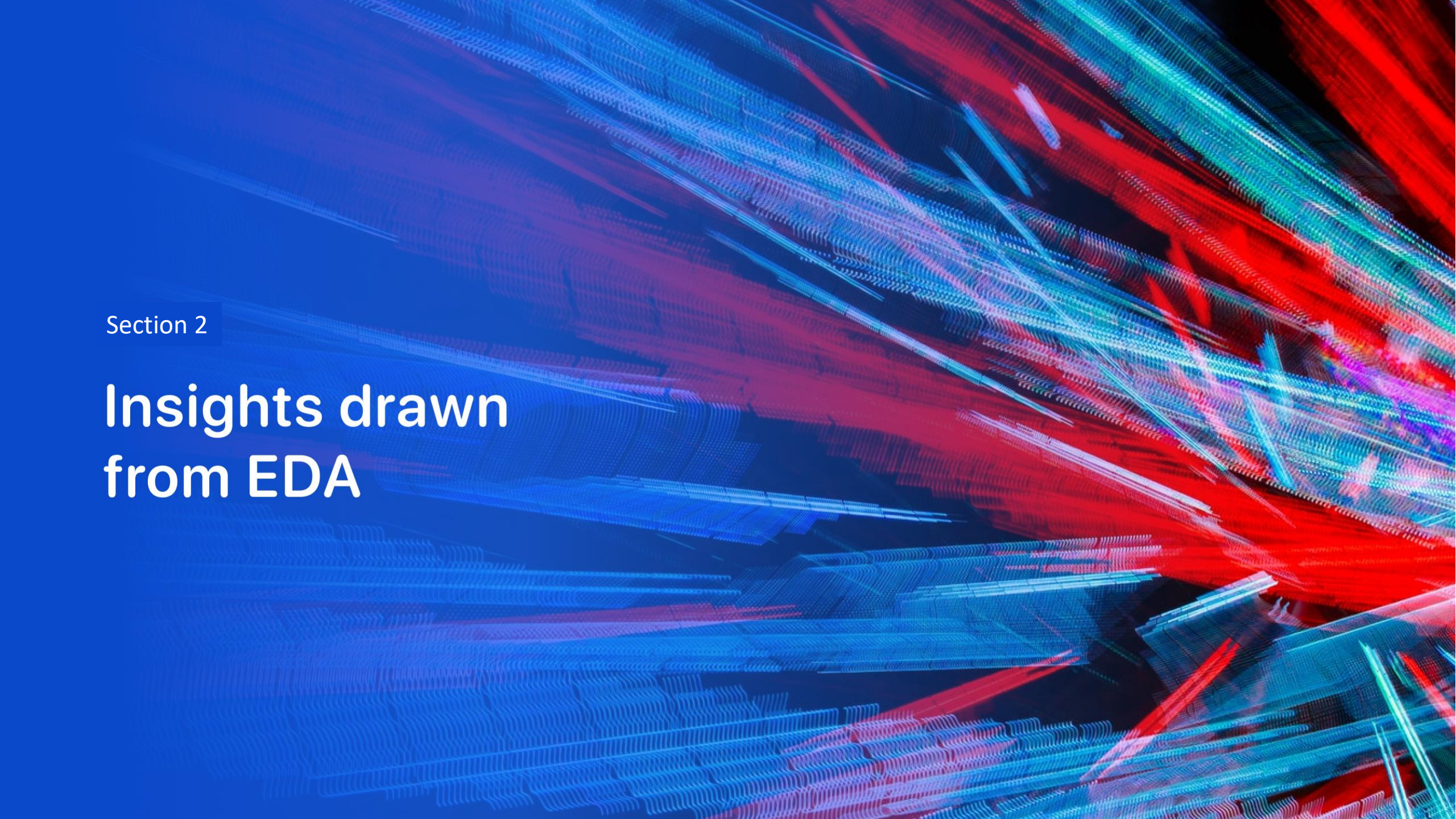
https://github.com/Anghelo-Salirrosas/DS-Repository/blob/2bcc512975eab9fe4f3f3351d02238429f01704c/3.3.MachineLearningPrediction.ipynb

Results

- Exploratory data analysis results:
  - Launches with payload mass bigger than 8000 kg have about 90% success rate
  - ES-L1, GEO, HEO, SSO orbits have 100% success rate
  - Over the years since 2010, the success rate has continued to increase.
- Four classifiers—Logistic Regression, SVM, Decision Tree, and KNN—were trained using GridSearchCV with cross-validation. Each model was evaluated using test set accuracy and confusion matrices. Evaluation across all of them yielded a similar accuracy of 84%, so we can choose any of them to predict mission success.

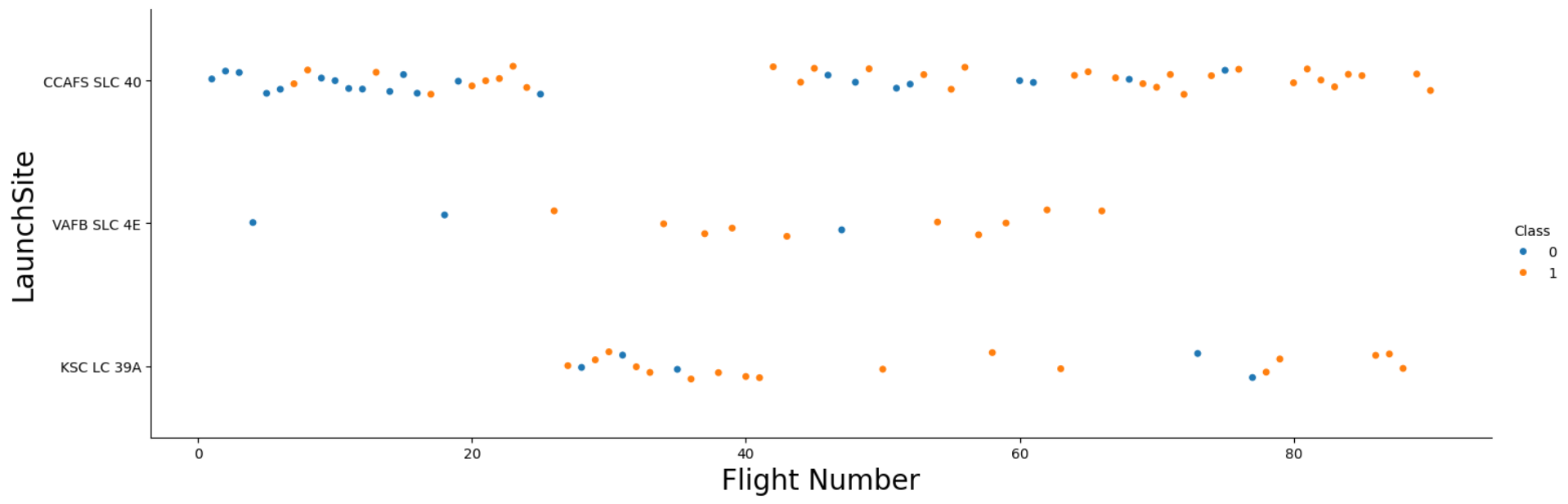





Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

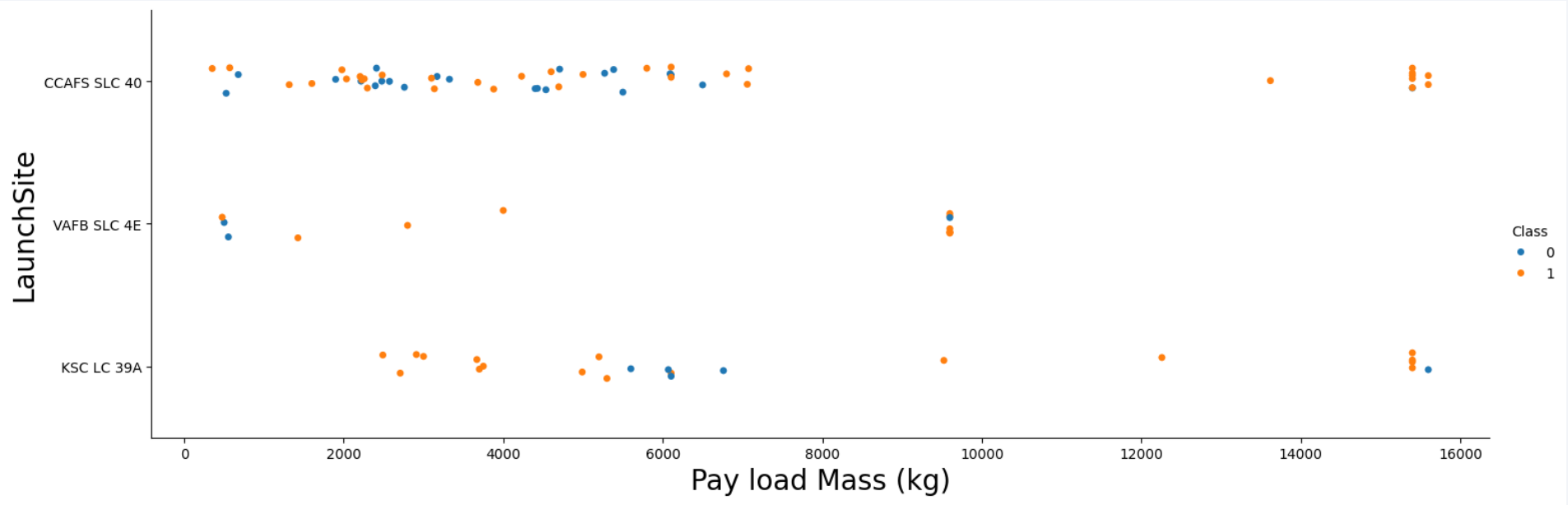
- Flight Number vs. Launch Site



The plot shows the relationship between flight number and launch site, with color indicating mission success (orange) or failure (blue). It reveals that recent launches, especially from KSC LC 39A and VAFB SLC 4E, tend to have higher success rates.

Payload vs. Launch Site

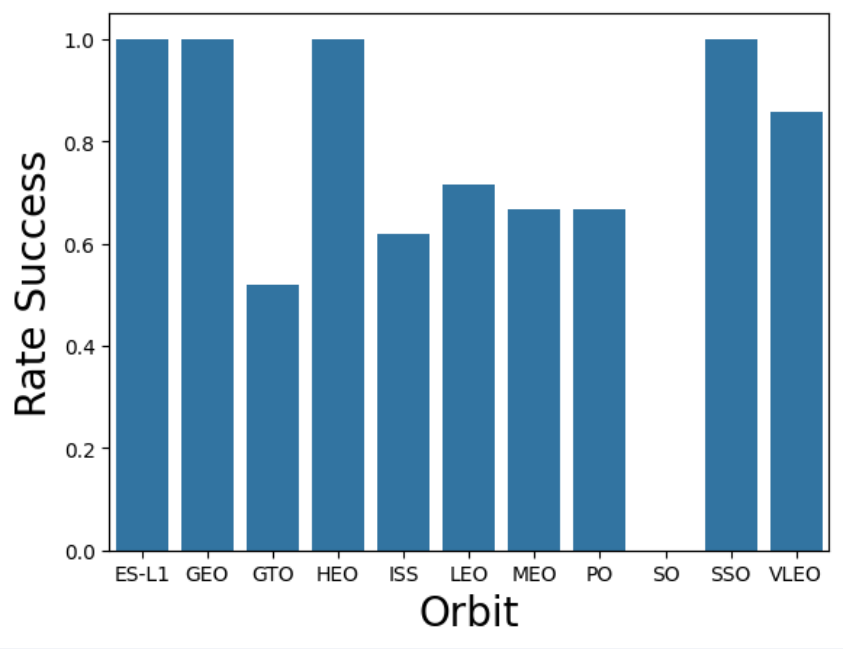
- Payload vs. Launch Site



This plot shows the relationship between payload mass (kg) and launch site, with color indicating mission outcome: orange for success and blue for failure. It reveals that successful launches occurred across all payload ranges, especially at CCAFS SLC 40. However, KSC LC 39A and VAFB SLC 4E show fewer failures with mid to high payloads, suggesting better performance under heavier loads.

Success Rate vs. Orbit Type

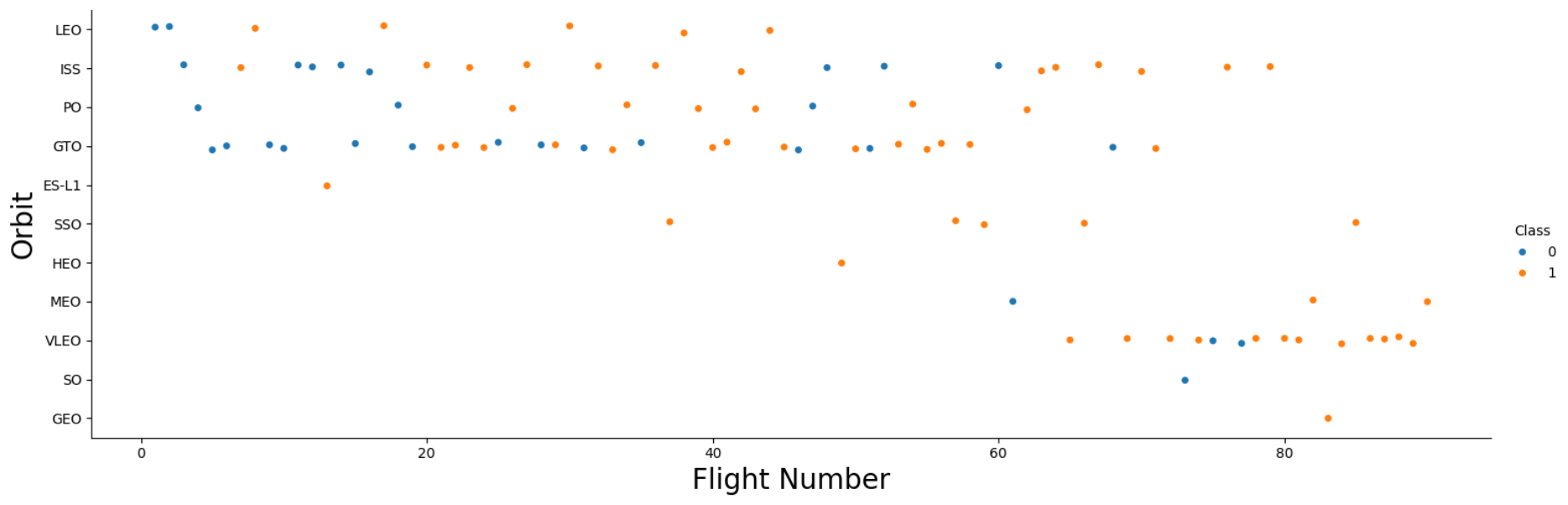
- Success Rate vs. Orbit Type



This bar plot shows the success rate of launches by orbit type. Most orbits, such as ES-L1, GEO, HEO, and SSO, have near-perfect success rates, indicating high reliability. In contrast, GTO (Geostationary Transfer Orbit) shows a significantly lower success rate, suggesting it's more challenging or risk-prone. Also, the SO orbit shows a success rate of no success.

Flight Number vs. Orbit Type

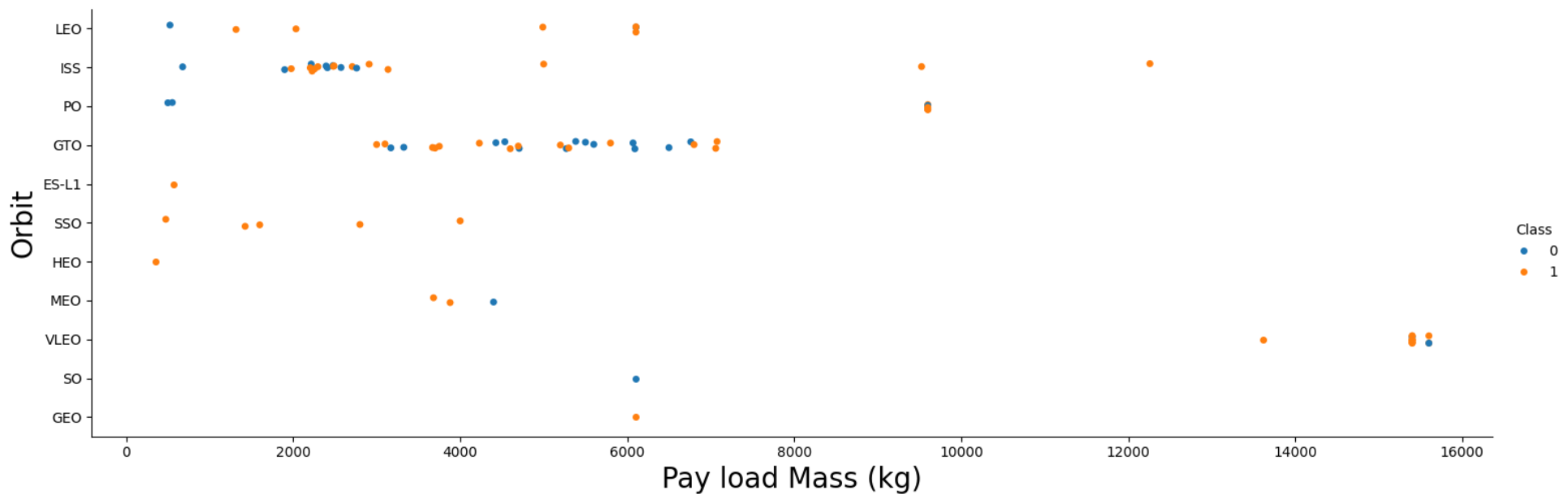
- Flight Number vs. Orbit Type



This plot shows the relationship between flight number and orbit type, with color indicating mission success (orange) or failure (blue). It reveals that certain orbits like VLEO, SSO, and HEO have a high concentration of successful missions, especially in later flights. In contrast, GTO exhibits more variability, with both successful and failed attempts across different flight numbers. The plot suggests that launch outcomes tend to improve over time for most orbits.

Payload vs. Orbit Type

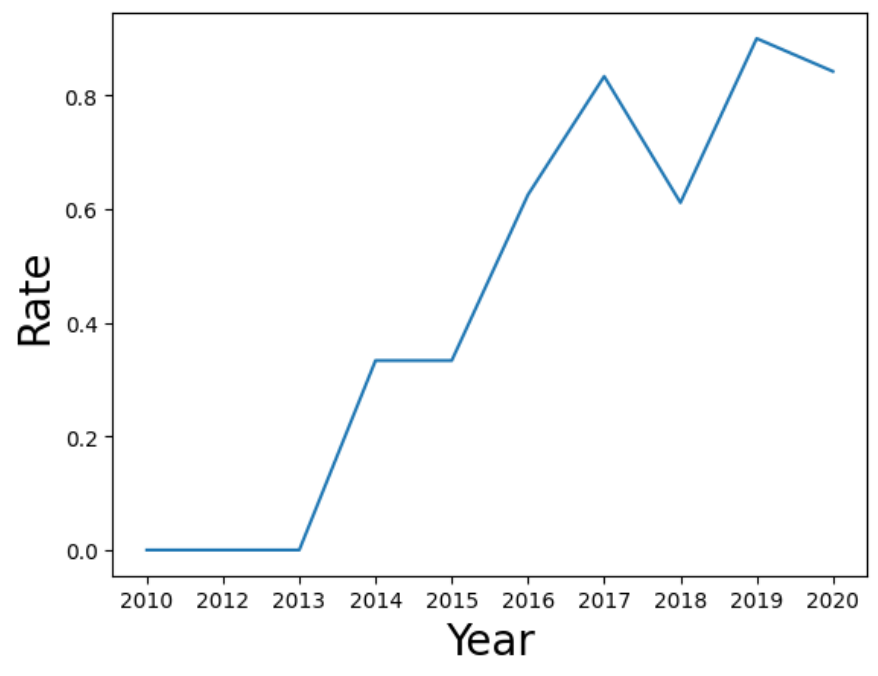
- Payload vs. Orbit Type



This plot shows the relationship between payload mass (kg) and orbit type, with mission success shown in orange and failure in blue. It reveals that heavier payloads are often associated with PO, ISS and VLEO orbits, and most of those missions were successful. Orbits like GTO and ISS appear with smaller payloads and more mixed outcomes. Overall, the success rate is generally high across different orbits and payload sizes, with a visible trend of successful missions even at higher payloads.

Launch Success Yearly Trend

- Launch Success Yearly Trend



This line plot shows the yearly trend of mission success rates from 2010 to 2020. It reveals a clear improvement over time, with success rates increasing sharply after 2015 and peaking in 2019. This indicates that SpaceX has consistently enhanced its launch reliability over the years.

All Launch Site Names

- %sql select distinct("Launch_Site") from SPACEXTABLE;

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

This SQL query retrieves the unique launch sites from the dataset. It helps identify the different locations SpaceX has used for missions, which is crucial for performance analysis by site.

Launch Site Names Begin with 'CCA'

- %sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5;

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

This query filters the launch records to only those where the launch site starts with "CCA", which corresponds to Cape Canaveral facilities. The LIMIT 5 ensures that only the first five matching rows are returned for quick inspection.

Total Payload Mass

- %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer LIKE 'NASA%';

sum(PAYLOAD_MASS__KG_)
99980

This query calculates the total payload mass carried in missions where NASA was the customer. It filters records using Customer LIKE 'NASA%' to include all NASA-related entries and returns the sum of the payload mass column.

Average Payload Mass by F9 v1.1

- %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version like 'F9 v1.1%';

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

This query calculates the average payload mass for missions launched using the Falcon 9 v1.1 booster version. It helps assess the typical carrying capacity of that specific rocket version based on historical data.

First Successful Ground Landing Date

- %sql select min(Date) from SPACEXTABLE where Landing_Outcome like 'Success (ground pad)%';

This query returns the earliest date when a launch successfully landed on a ground pad. It uses MIN(Date) to find the first occurrence where the Landing_Outcome includes "Success (ground pad)".

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql select Booster_Version from SPACEXTABLE where Landing_Outcome like 'Success (drone ship)%' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000;

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

This query retrieves the names of booster versions that successfully landed on a drone ship and carried a payload mass between 4000 and 6000 kg. It could be helpful to identify which boosters handled moderate-to-heavy payloads and still achieved successful drone landings, helping evaluate booster reliability under load.

Total Number of Successful and Failure Mission Outcomes

- %sql select Mission_Outcome, count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome;

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

This query counts how many times each mission outcome occurred by grouping the data by the Mission_Outcome column. To get a summary of how often missions succeeded, failed, or were partially successful.

Boosters Carried Maximum Payload

- %sql select booster_version, payload_mass__kg_ from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

This query finds the booster(s) that carried the maximum payload mass by using a subquery to get the highest payload value in the dataset. To identify which booster version achieved the highest payload capacity.

2015 Launch Records

- %sql select Landing_Outcome, Booster_Version, Launch_Site, substr(Date, 6,2) as Month from SPACEXTABLE where substr(Date,0,5)='2015' and Landing_Outcome like 'Failure (drone ship)%';

Landing_Outcome	Booster_Version	Launch_Site	Month
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	01
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	04

This query filters for failed drone ship landings in 2015 and returns the landing outcome, booster version, launch site, and the month in which the failure occurred.

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

- %sql select Landing_Outcome,Date from (select * from SPACEXTABLE where Landing_Outcome like 'Success (ground pad)%' or Landing_Outcome like 'Failure (drone ship)%') where Date between '2010-06-04' and '2017-03-20' order by Date desc;

Landing_Outcome	Date
Success (ground pad)	2017-02-19
Success (ground pad)	2016-07-18
Failure (drone ship)	2016-06-15
Failure (drone ship)	2016-03-04
Failure (drone ship)	2016-01-17
Success (ground pad)	2015-12-22
Failure (drone ship)	2015-04-14
Failure (drone ship)	2015-01-10

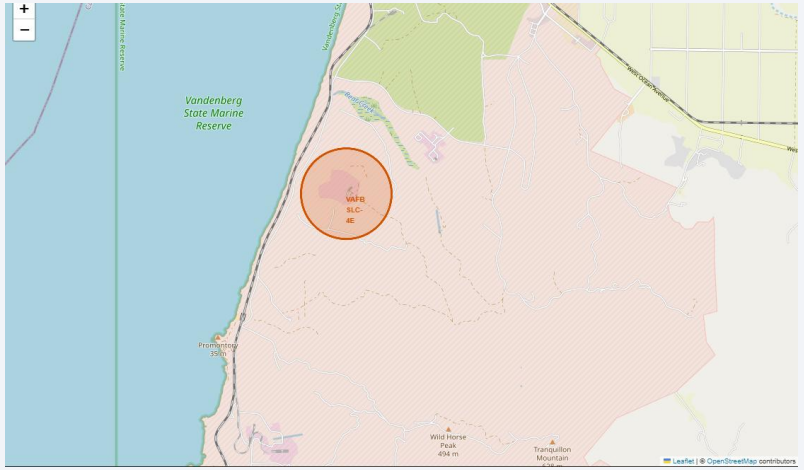
This query retrieves records of landings that were either successful on ground pads or failed on drone ships, limited to the date range between 2010-06-04 and 2017-03-20, and sorted from most recent to oldest.

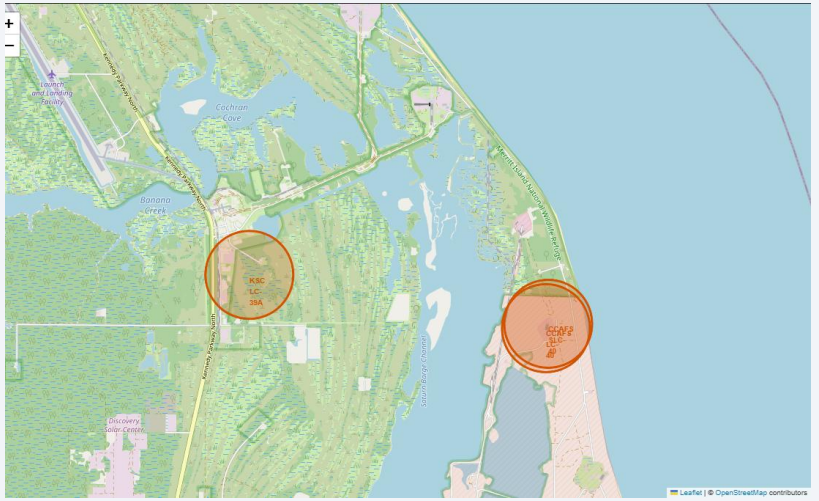
Section 3

Launch Sites Proximities Analysis

Launches Site on Map

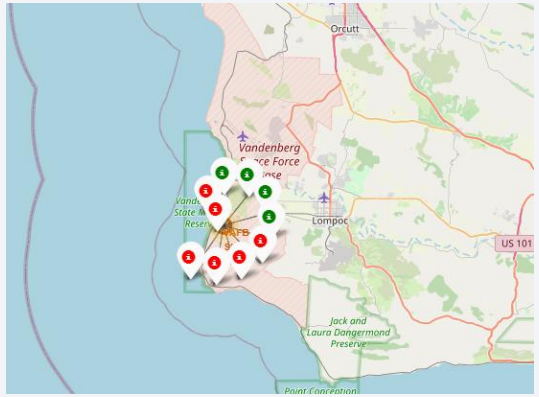
- The Folium map displays the geographic locations of SpaceX launch sites across the United States, specifically Florida and California.
- Each launch site is represented by:
  - A circle (orange, 1000-meter radius) centered on its latitude and longitude.
  - A text label marker (DivIcon) with the site's name, shown directly on the map for clarity.
- The Florida cluster includes:
  - CCAFS SLC-40
  - CCAFS LC-40
  - KSC LC-39A
- The California site is:
  - VAFB SLC-4E

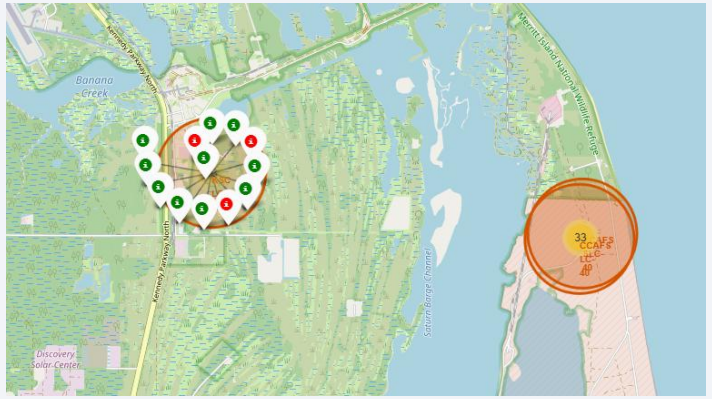




Launches classified by success

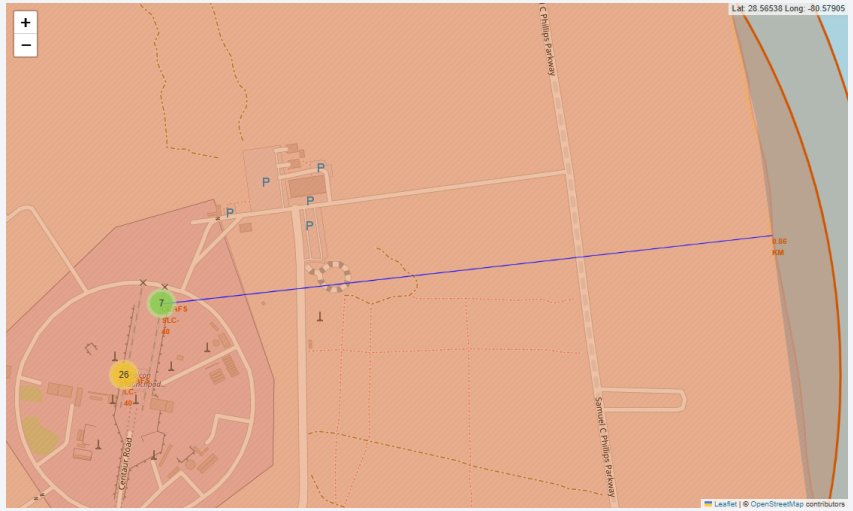
- The interactive map uses clustered markers to classify SpaceX launches by success or failure. Green markers represent successful missions, while red ones indicate failures. Zoomed-in views of VAFB and KSC reveal the launch concentration and outcome distribution per site, providing visual insight into spatial launch performance.
    - First image: A global view showing clustered launch markers over Florida and California, indicating areas with high launch activity.
    - Second image: A zoomed-in view of VAFB SLC 4E (California) showing individual launches marked as green (success) or red (failure).
    - Third image: A close-up of KSC LC 39A (Florida), displaying both clustered outcomes and a location circle for geographic context.

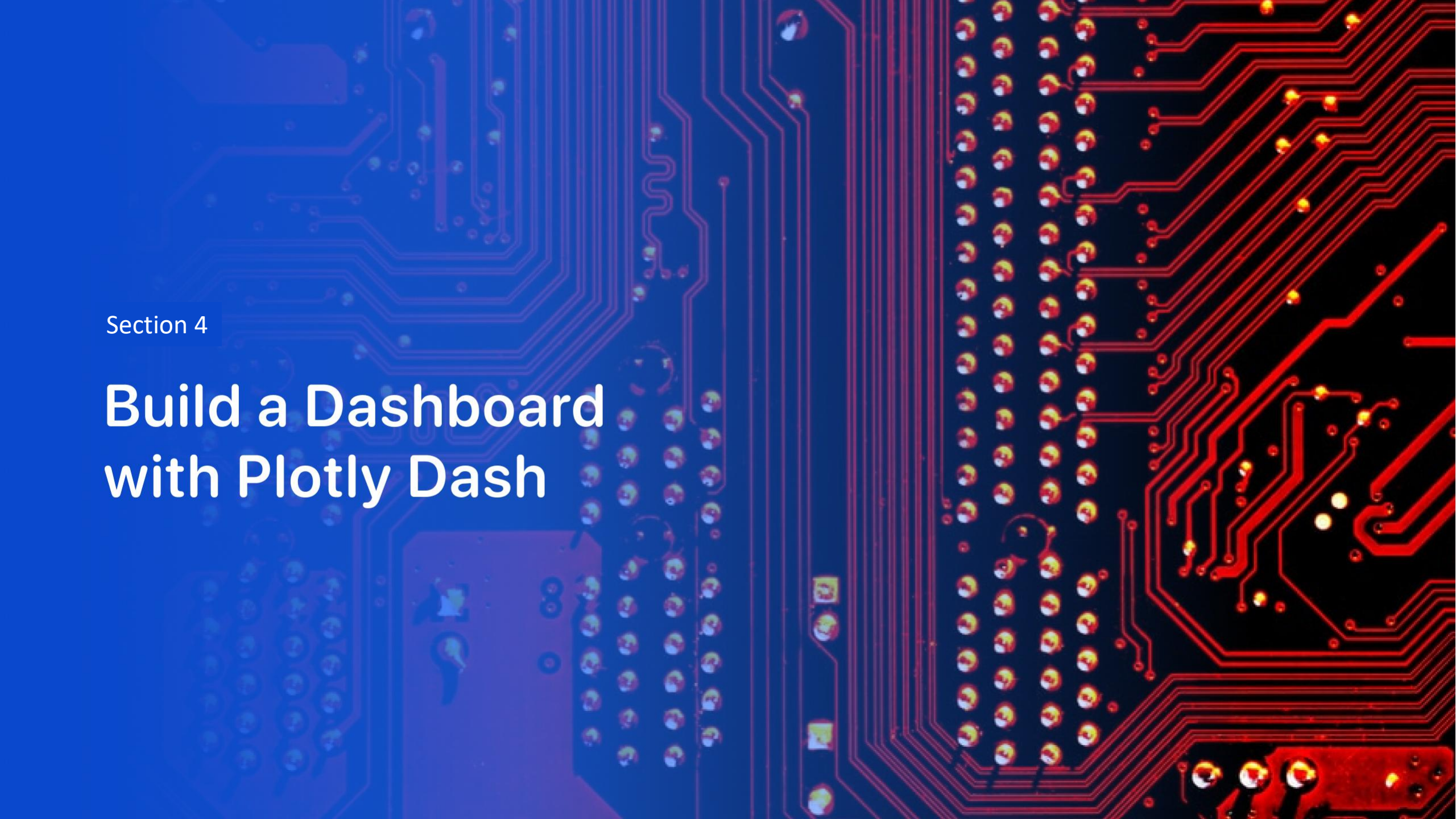




Launches site proximities

- The map shows the calculated distance from CCAFS SLC-40 to the nearest coastline using a blue line and a labeled marker. This helps visualize spatial proximity for potential safety or logistical analysis.
- In the map we can see:
  - A blue line connects the launch site (CCAFS SLC-40) to the nearest point on the coastline, indicating proximity.
  - A text marker labeled "0.86 KM" shows the exact distance, calculated dynamically using geospatial coordinates.

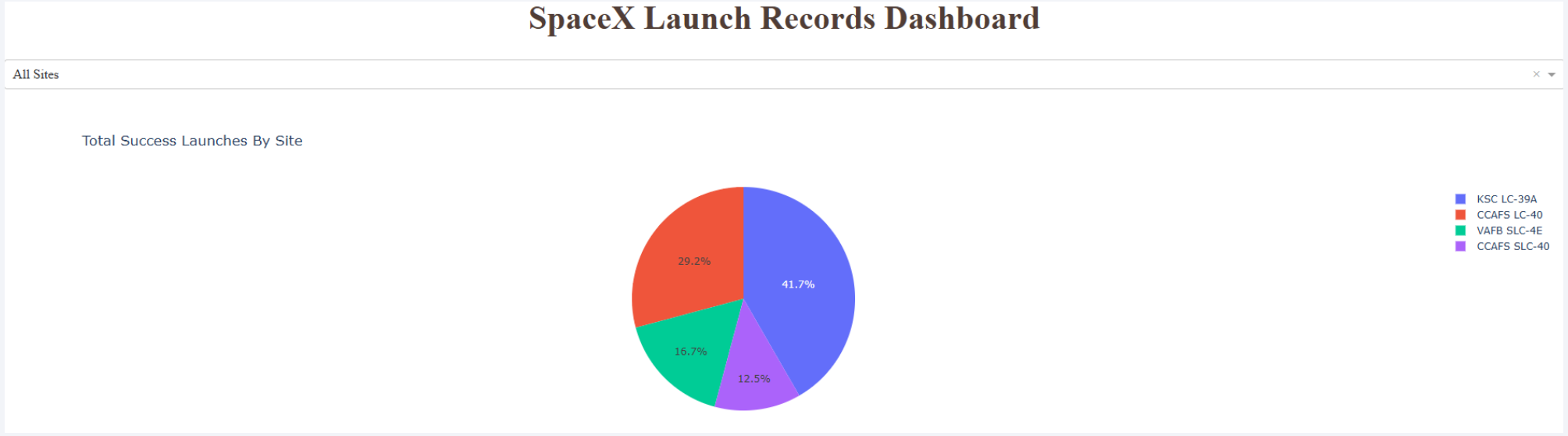



Section 4
Build a Dashboard with Plotly Dash

Total launch success by site

This is a pie chart from the SpaceX Launch Records Dashboard showing the total number of successful launches by launch site. It includes:

- A dropdown menu allowing users to filter by site (currently set to "All Sites").
- The pie chart visualizes the distribution of successful launches, with KSC LC-39A leading at 41.7%, followed by CCAFS LC-40, VAFB SLC-4E, and a second CCAFS SLC-40 entry.
- The chart helps identify which launch sites contribute the most to SpaceX's success, indicating KSC LC-39A as the most reliable site overall.

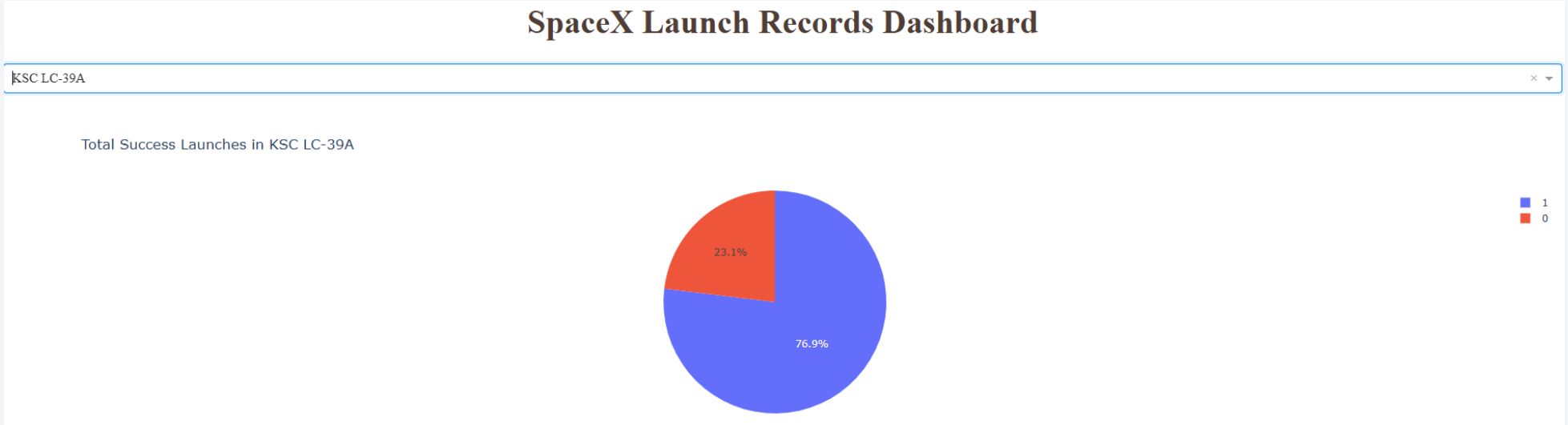


Highest launch success ratio

This dashboard pie chart focuses on launch outcomes at KSC LC-39A, the site with the highest success ratio. The chart reveals:

- 76.9% of launches were successful (blue segment).
- 23.1% were failures (red segment).

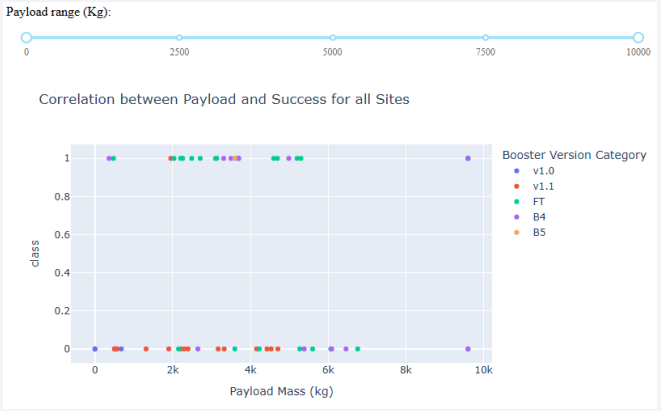
By filtering the dropdown to KSC LC-39A, the dashboard highlights its strong performance, justifying its selection as the most reliable launch site.

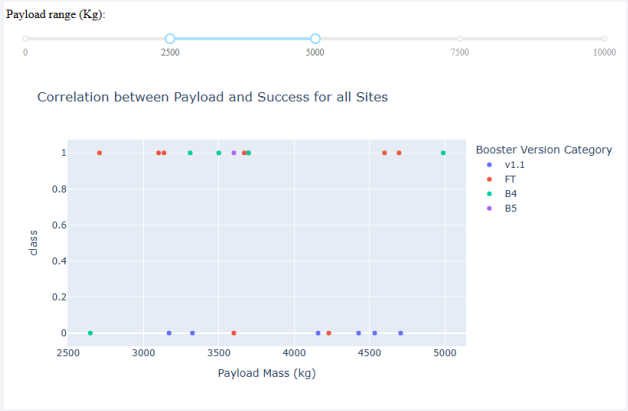


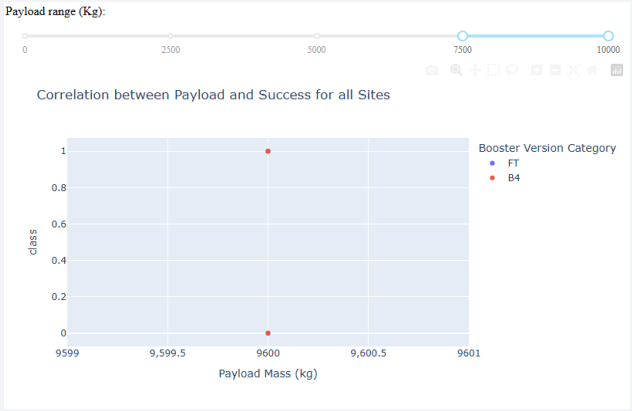
<Dashboard Screenshot 3>

- Full Payload Range (0–10,000 kg):
  Most booster versions (especially FT and B5) have high success rates.
  Failures are scattered across all versions, but older versions like v1.0 show more failures.
- Medium Payload Range (2500–5000 kg):Majority of launches are successful.FT and B4 dominate this range with solid performance.
- High Payload Range (~9600 kg):Very few launches, but they show a 100% success rate.All performed by FT and B4 boosters, indicating reliability at high payloads.

Then, we can say Booster Versions FT and B4 demonstrate the highest success rate, especially at medium and high payload ranges. Old versions (v1.0, v1.1) are linked with more failures. The data suggests a strong positive correlation between advanced booster versions and mission success, even at higher payload capacities.





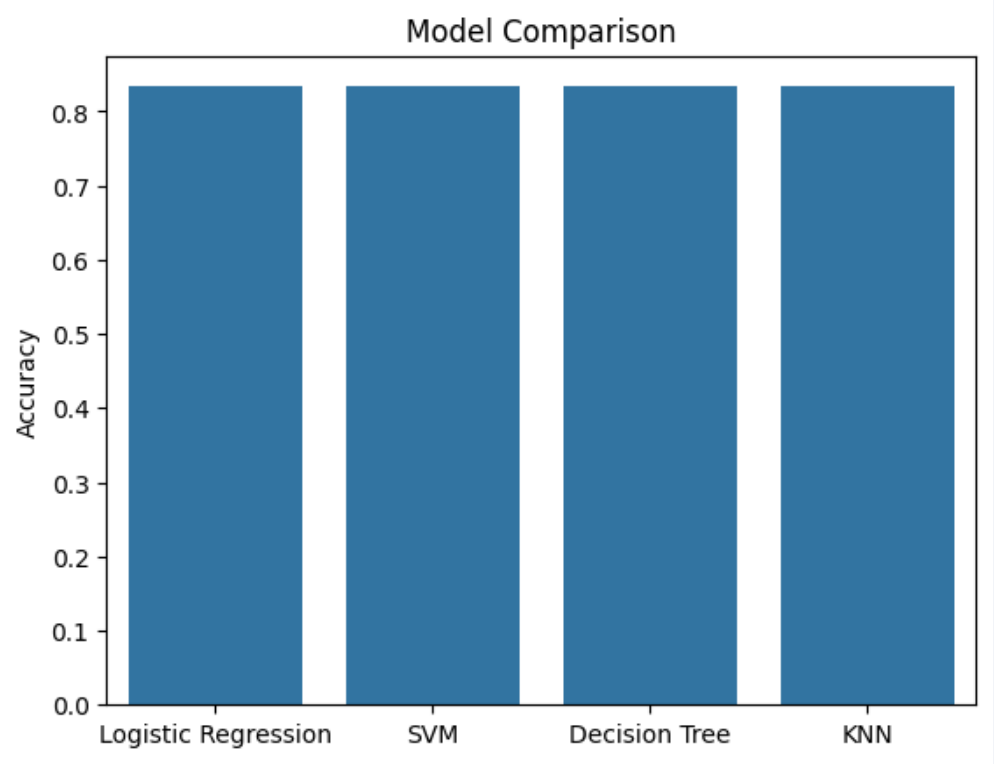


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The bar chart compares the classification accuracy of four machine learning models: Logistic Regression, SVM, Decision Tree, and KNN. All models show approximately the same accuracy level, around **0.83**, indicating that none of them significantly outperforms the others on this dataset. Therefore, model selection may depend on other factors such as interpretability or training time.



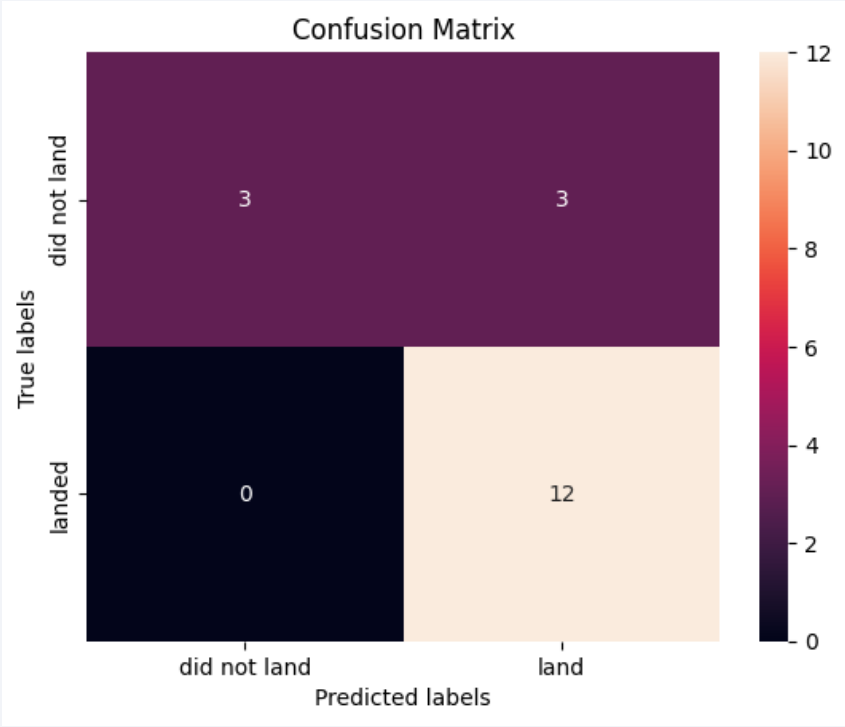
Confusion Matrix

The confusion matrix shows the performance of the four trained models on the test set:

- True Positives (landed predicted as landed): 12
- True Negatives (did not land predicted as did not land): 3
- False Positives (did not land predicted as landed): 3
- False Negatives (landed predicted as did not land): 0

This means the models correctly classified all landings, but it misclassified 3 non-landings as landings.

Overall, the models has a strong bias toward predicting landings and performs well in identifying successful landings.



Conclusions

The analysis revealed key patterns in SpaceX launches.

- Certain orbits (e.g., ES-L1, GEO) and higher payloads (>8000 kg) had higher success rates.
- KSC LC-39A was the most reliable launch site.
- Recent booster versions like FT and B5 demonstrated superior performance.
- All four predictive models showed similar accuracy (~83%), with Logistic Regression being slightly more interpretable.

This pipeline demonstrates the value of combining data engineering, EDA, visualization, and machine learning to generate insights and support decision-making in aerospace applications.

Appendix

- Repository: https://github.com/Anghelo-Salirrosas/DS-Repository

Thank you!