Winning Space Race with Data Science

Andrey Pomortsev

27/12/2023

Outline

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

Executive Summary

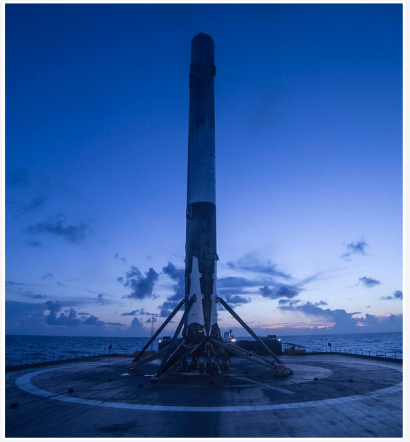
Methodologies:

- Data collection using SpaceX API, and web scrapping
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis using Classification

Results:

- Valuable data was successfully collected from public sources.
- Exploratory Data Analysis identified key features for predicting the success of launches.
- Machine Learning Prediction determined the optimal model for forecasting important characteristics that drive this opportunity effectively, utilizing all collected data.

Introduction



The project aims to predict the successful landing of the Falcon 9 first stage, crucial for determining the cost savings associated with SpaceX's reusable rocket technology and aiding potential competitors in bidding for rocket launches against SpaceX.

Questions I want to find answers:

- Impact of variables (payload mass, launch site, flights, orbits) on first stage landing success.
- Trend in successful landings over the years.
- Optimal algorithm for binary classification in this scenario.

Section 1

Methodology

Methodology

- Utilized SpaceX REST API for structured data acquisition.
- Employed Web Scraping techniques on Wikipedia for supplementary data collection.
- Data Wrangling Procedures:
  - Filtered the acquired data for relevance.
  - Addressed missing values through appropriate strategies.
  - Applied One Hot Encoding to prepare the data for binary classification.
- Exploratory Data Analysis (EDA):
  - Conducted EDA using a combination of visualization techniques and SQL queries.
- Interactive Visual Analytics:
  - Utilized Folium and Plotly Dash for interactive and dynamic visualizations.
- Predictive Analysis with Classification Models:
  - Developed, fine-tuned, and evaluated classification models to achieve optimal results.

Data Collection

The data collection process was meticulously orchestrated by amalgamating SpaceX REST API requests and Web Scraping techniques from a dedicated table in SpaceX's Wikipedia entry. This dual approach was imperative to ensure a comprehensive dataset, facilitating a nuanced and thorough analysis of the launches.

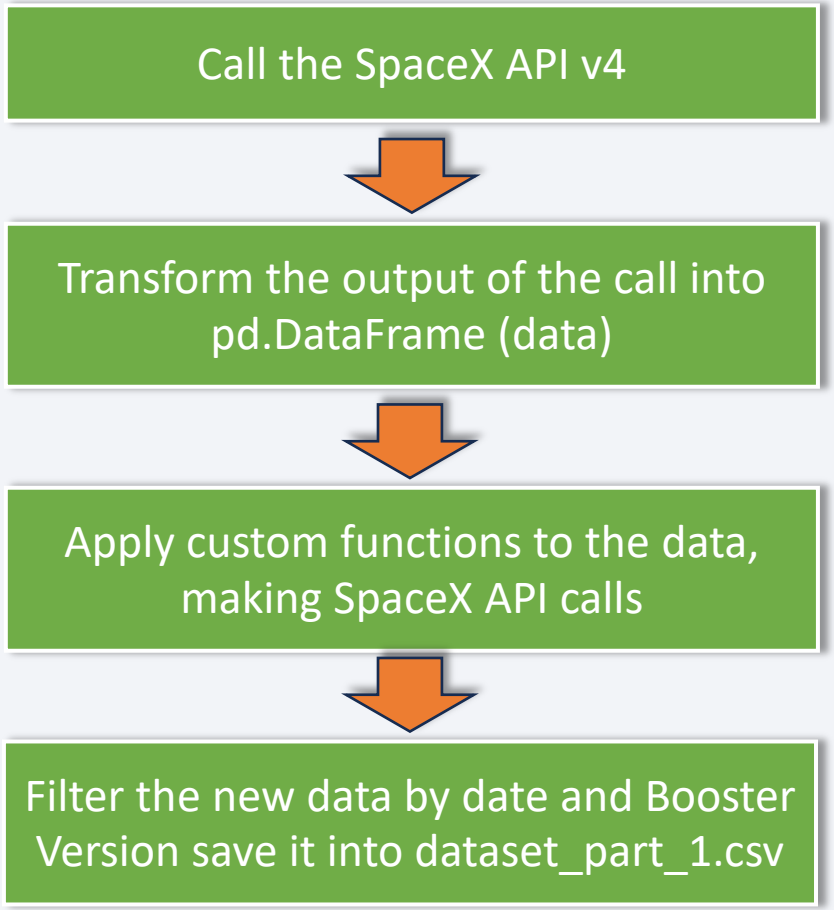
The SpaceX REST API yielded essential columns such as FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude. Meanwhile, the Web Scraping from Wikipedia enriched the dataset with columns including Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.

This meticulous data aggregation process enabled us to achieve a holistic and detailed understanding of the launches, setting the stage for a comprehensive analysis.

Data Collection – SpaceX API

- I used SpaceX API v4 to get the data
- `api.spacexdata.com/v4/launches/past`

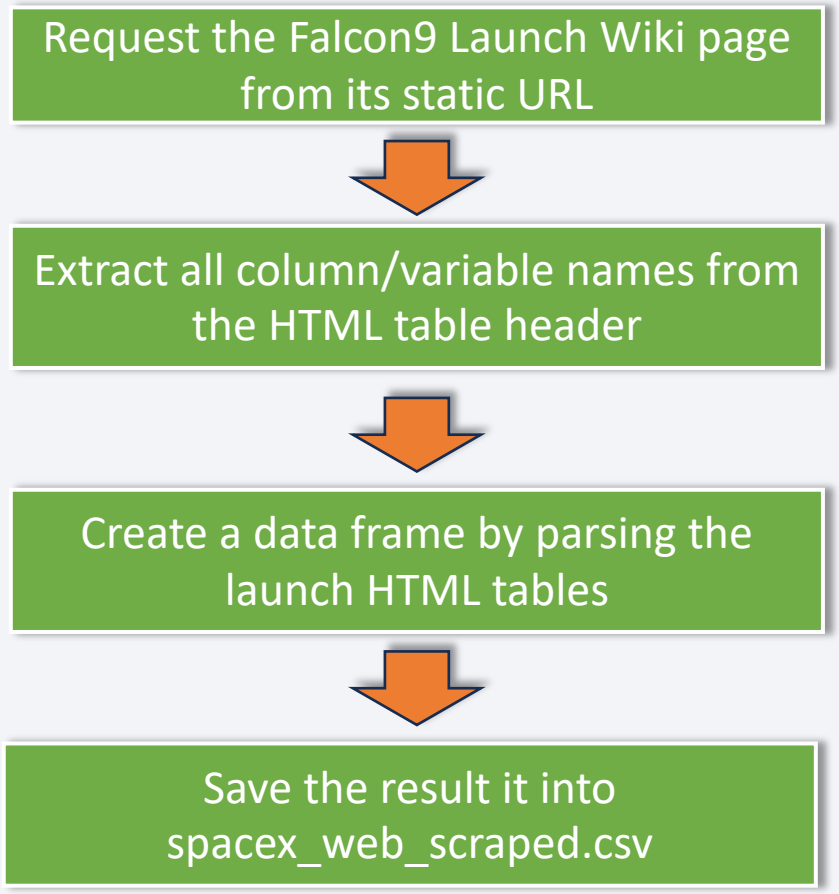
- Calls from the custom functions:
  - `…/v4/rockets/ + value`
  - `…/v4/launchpads/ + value`
  - `…/v4/payloads/ + value`



The completed SpaceX API calls notebook.
The result: dataset_part_1.csv

Data Collection - Scraping

- To keep the lab results reproducible, I used the data from a snapshot of the `List of Falcon 9 and Falcon Heavy launches` Wikipage updated on`9th June 2021`.
- BeautifulSoup version 4.12.2 was utilized for web scraping.
- The scraped data was stored at first in a dictionary (hash map), then turned into a pd.DataFrame, and finally saved in csv format.



The completed web scraping notebook.
The result: space_web_scraped.csv

Data Wrangling

Objective: Convert diverse landing outcomes into binary training labels.

- Outcome Types:
  - Successful ocean landing denoted by True Ocean.
  - Unsuccessful ocean landing denoted by False Ocean.
  - Successful ground pad landing denoted by True RTLS.
  - Unsuccessful ground pad landing denoted by False RTLS.
  - Successful drone ship landing denoted by True ASDS.
  - Unsuccessful drone ship landing denoted by False ASDS.
- Conversion Process:
  - Assign "1" for successful landings.
  - Assign "0" for unsuccessful landings.
- Goal: Simplify and standardize outcomes for efficient training label creation.

The completed data wrangling notebook

EDA with Data Visualization

1. Flight Number vs. Launch Site:
   - Use scatter plot to explore patterns in the relationship between flight numbers and launch sites.
2. Payload vs. Launch Site:
   - Utilize categorical scatter plot to compare payload characteristics across launch sites.
3. Success Rate of Each Orbit Type:
   - Present data using bar chart to visualize success rates across different orbit types.
4. Flight Number vs. Orbit Type:
   - Use categorical scatter chart to examine the distribution of flight numbers across orbit types.
5. Payload vs. Orbit Type:
   - Illustrate payload characteristics across orbit types with categorical scatter plot .
6. Launch Success Yearly Trend:
   - Track the annual trend of launch success using a line chart.

The completed data visualization notebook

EDA with SQL

- Unique Launch Sites:
  - Display names of launch sites.
- Launch Sites Starting:
  - Show 5 records with launch sites starting 'CCA'.
- Total Payload by NASA:
  - Display total payload by NASA (CRS) boosters.
- Average Payload of a booster:
  - Show average payload of F9 v1.1 boosters.
- First Successful Ground Pad Landing:
  - Retrieve date of first successful ground pad landing.

The completed EDA with SQL notebook

- Successful Drone Ship Landings:
  - List boosters with drone ship success and payload between 4000 and 6000 kg.
- Total Successful and Failed Missions:
  - Summarize total successful and failed mission outcomes.
- Boosters with Max Payload:
  - Identify booster_versions with maximum payload using a subquery.
- 2015 Month-wise Records:
  - Display records for 2015 with month names, drone ship failures, booster versions, and launch sites.
- Rank Landing Outcomes:
  - Rank landing outcomes count between June 4, 2010, and March 20, 2017, in descending order.

Build an Interactive Map with Folium

- Markers of all Launch Sites:
  - Added Marker with Circle, Popup Label, and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
  - Added Markers with Circle, Popup Label, and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to the Equator and coasts.
- Coloured Markers of the launch outcomes for each Launch Site:
  - Added color Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Distances between a Launch Site to its proximities:
  - Added color Lines to show distances between the Launch Site CCAFS CLS-40 and closest Railway, Highway, Coastline, and Closest City (Melbourne, FL).

The completed interactive map with Folium notebook

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:
  - Simplify user interaction by implementing a dropdown list for easy Launch Site selection. This enhances user experience, streamlining the exploration of launch data.
- Pie Chart displaying Success Launches:
  - Provide a visual overview of launch success rates. The pie chart offers a quick insight into the overall success-to-failure ratio. When selecting a specific Launch Site, it reveals site-specific success patterns, aiding a more focused analysis.
- Payload Mass Range Slider:
  - Empower users to customize their analysis with a payload mass range slider. This feature allows users to refine data exploration based on specific payload criteria, adding a personalized touch to the investigation.
- Scatter Chart portraying Payload Mass vs. Success Rate for various Booster Versions:
  - Enhance data understanding by visually representing the relationship between Payload Mass and Launch Success. The scatter chart facilitates quick pattern identification across different Booster Versions, enabling users to make informed decisions and derive meaningful insights.

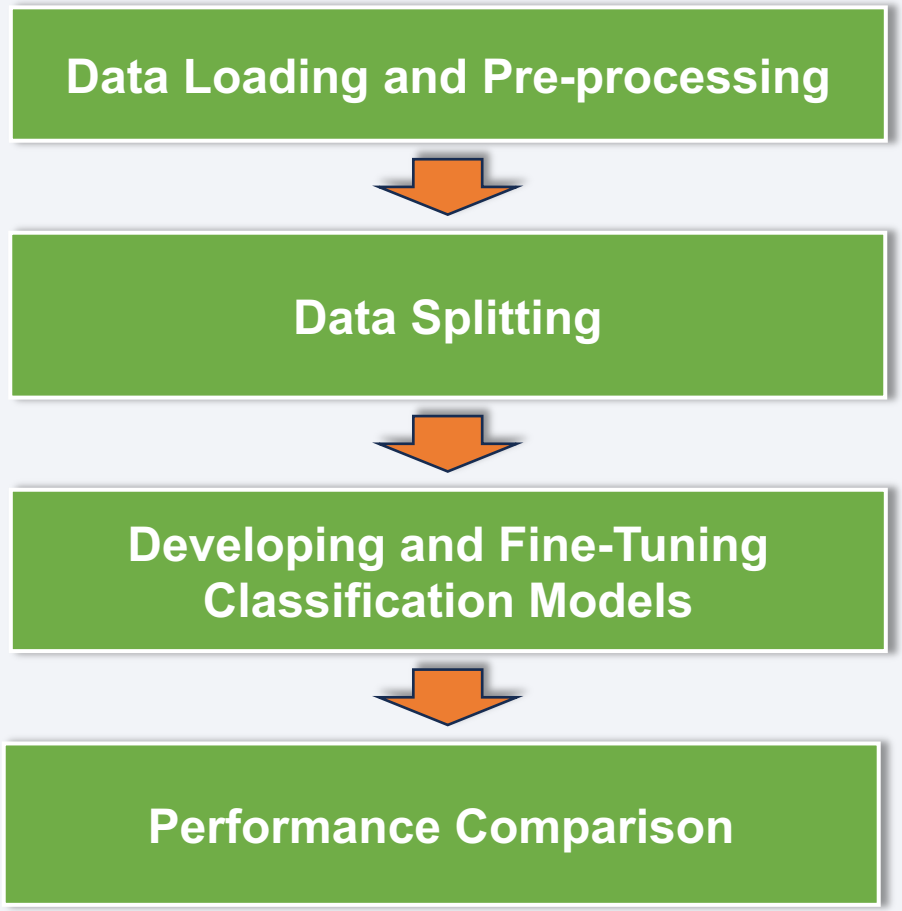
The completed Plotly Dash lab

Predictive Analysis (Classification)

1.Data Handling:
- Loaded and standardized the dataset.
- Split data into training and test sets.

2.Model Evaluation:
- Applied Logistic Regression, SVM, Decision Tree, and KNN classifiers and tuned hyperparameters using GridSearchCV for each, evaluating their performance through accuracy and confusion matrices.

3.Comparison and Selection:
- Compared models based on accuracy, precision, recall, and training time.
- Identified the best-performing model for the given dataset.



The completed predictive analysis lab

Results

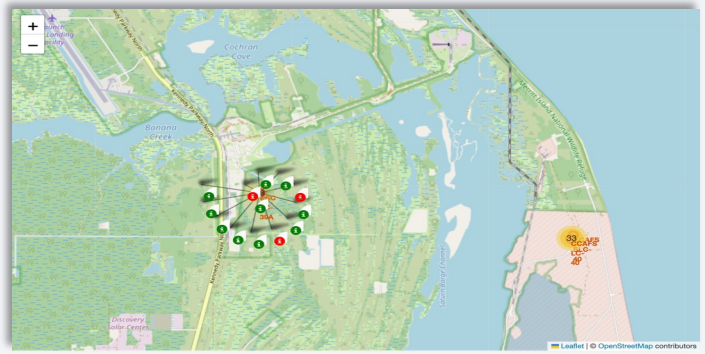
Exploratory Data Analysis results:

- Space X utilizes 4 distinct launch sites.
- Initial launches were directed towards Space X itself and NASA.
- The average payload of the F9 v1.1 booster is 2,928 kg.
- The initial successful landing outcome occurred in 2015, five years after the first launch.
- Several Falcon 9 booster versions demonstrated successful landings on drone ships with payloads surpassing the average.
- Nearly 100% of mission outcomes achieved success.
- In 2015, two booster versions (F9 v1.1 B1012 and F9 v1.1 B1015) experienced failures at landing on drone ships.
- The success rate of landing outcomes improved with each passing year.

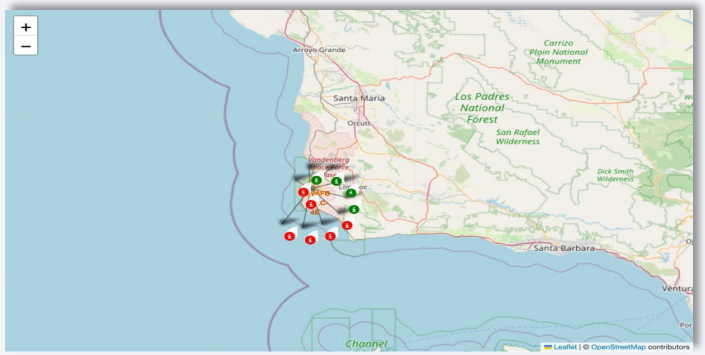
Results

Interactive analytics demo in screenshots:

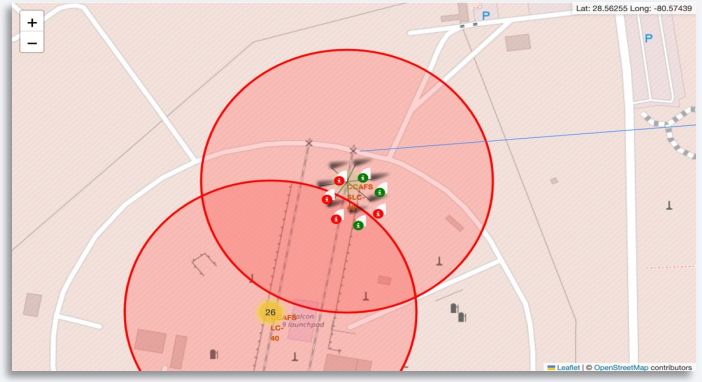
KSC LC-39A



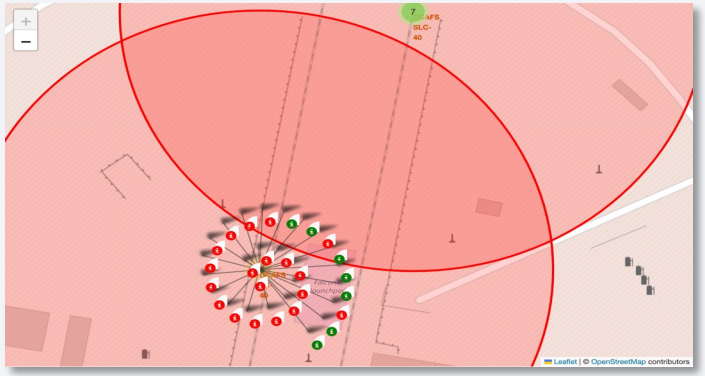
VAFB SLC-4E



CCAFS SLC-40



CCAFS LC-40

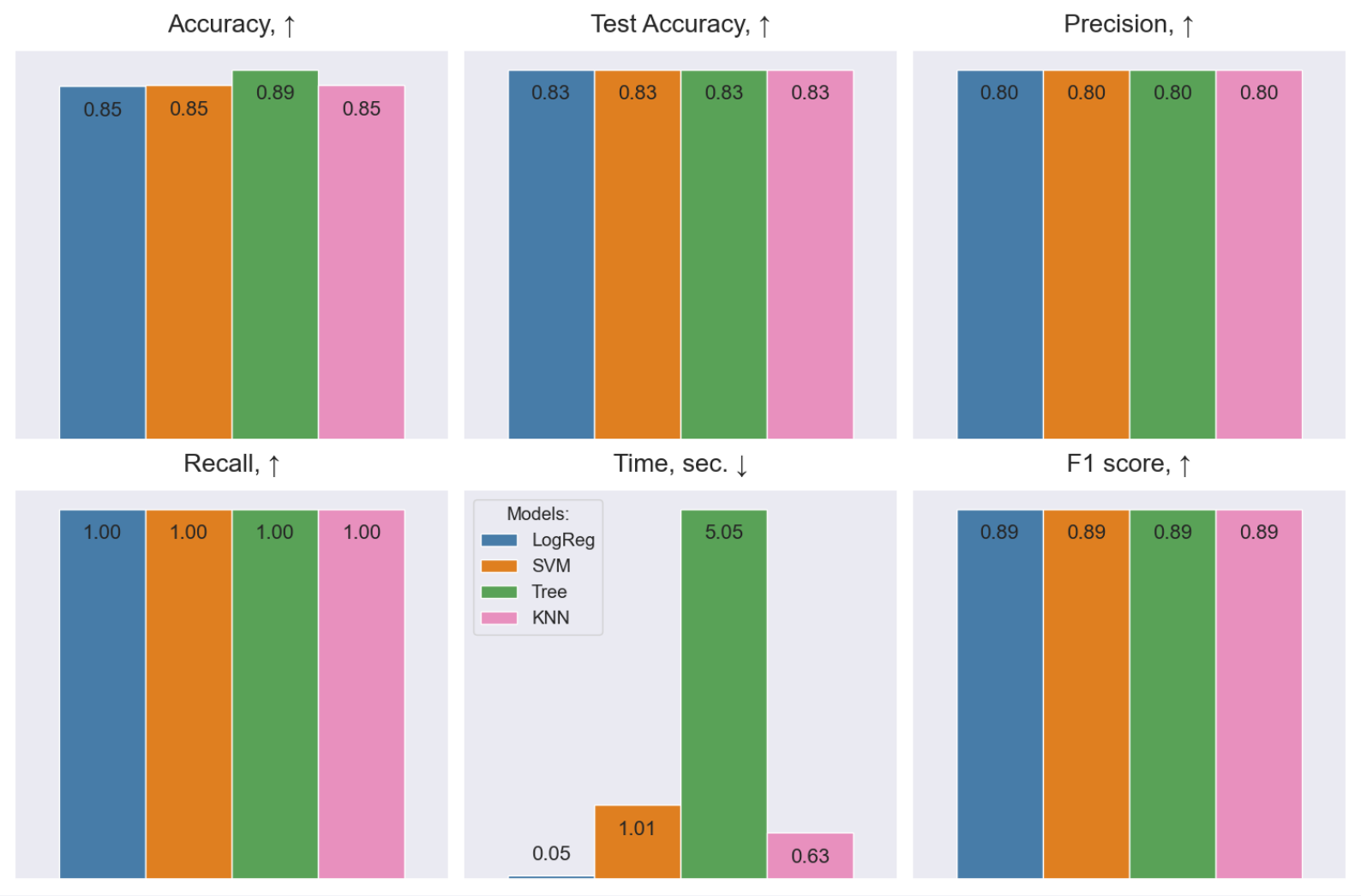


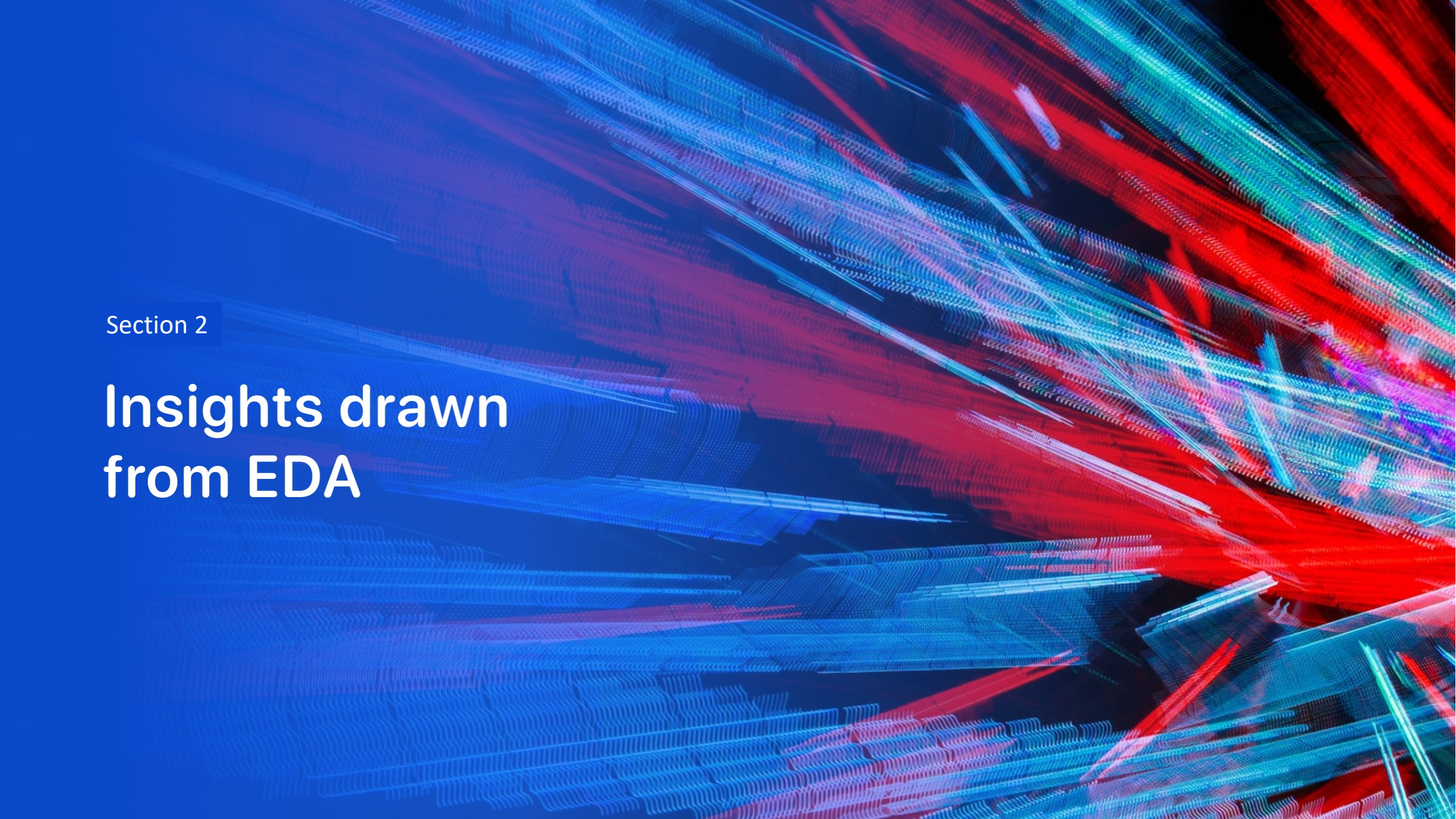
Results

There exists a trade-off between a model's performance and its speed, influenced by the choice of algorithms: LogisticRegression, SVM, KNN, and DecisionTrees.

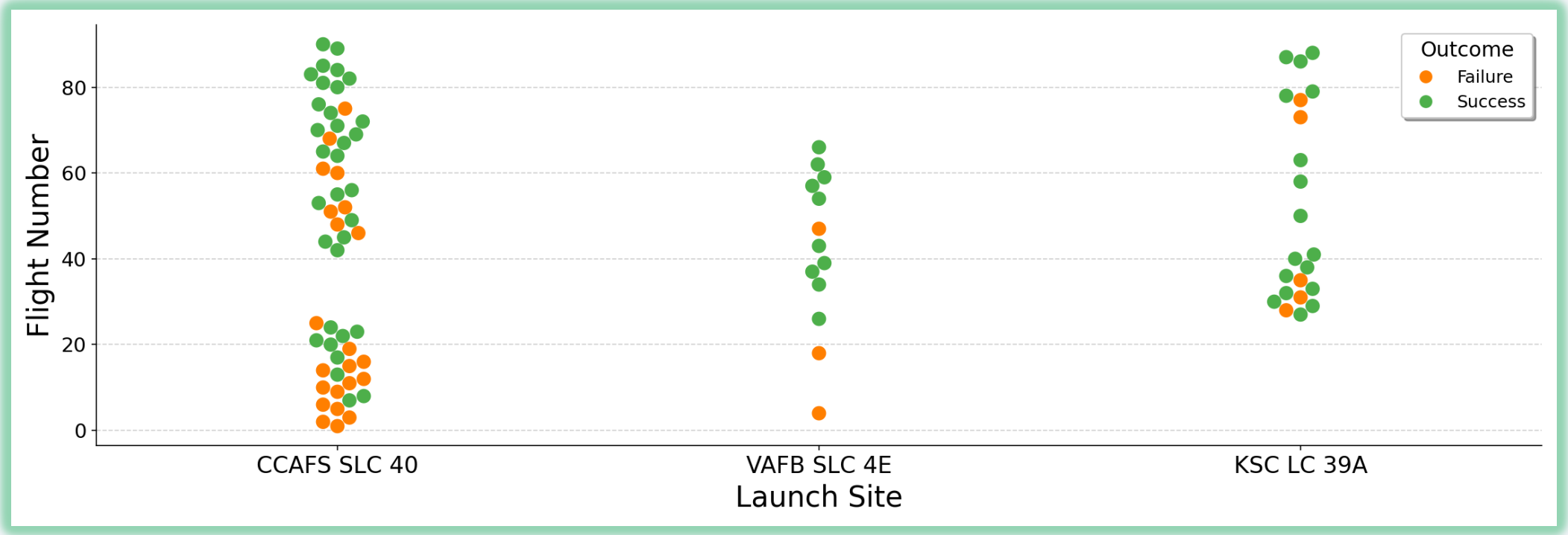
- **Logistic Regression:**
  - **Performance:** Generally robust and efficient for linearly separable data. Works well when assumptions align with the data.
  - **Speed:** Fast training and prediction due to its simplicity. Suitable for large datasets.
- **Support Vector Machines (SVM):**
  - **Performance:** Effective in high-dimensional spaces; versatile with different kernel functions for complex relationships.
  - **Speed:** Training time can be slower, especially with large datasets or complex kernels.
- **K-Nearest Neighbors (KNN):**
  - **Performance:** Adapts well to data with non-linear patterns. Simple and effective for classification tasks.
  - **Speed:** Slower, particularly as the dataset size increases, as it involves computing distances for predictions.
- **Decision Trees:**
  - **Performance:** Can capture complex relationships, non-linearity, and interactions in the data.
  - **Speed:** Training is fast, but prediction speed may vary based on the depth and complexity of the tree.

Results


Accuracy, ↑
0.85
0.85
0.89
0.85
Test Accuracy, ↑
0.83
0.83
0.83
0.83
Precision, ↑
0.80
0.80
0.80
0.80
Recall, ↑
1.00
1.00
1.00
1.00
Time, sec. ↓
Models:
LogReg
SVM
Tree
KNN
0.05
1.01
5.05
0.63
F1 score, ↑
0.89
0.89
0.89
0.89

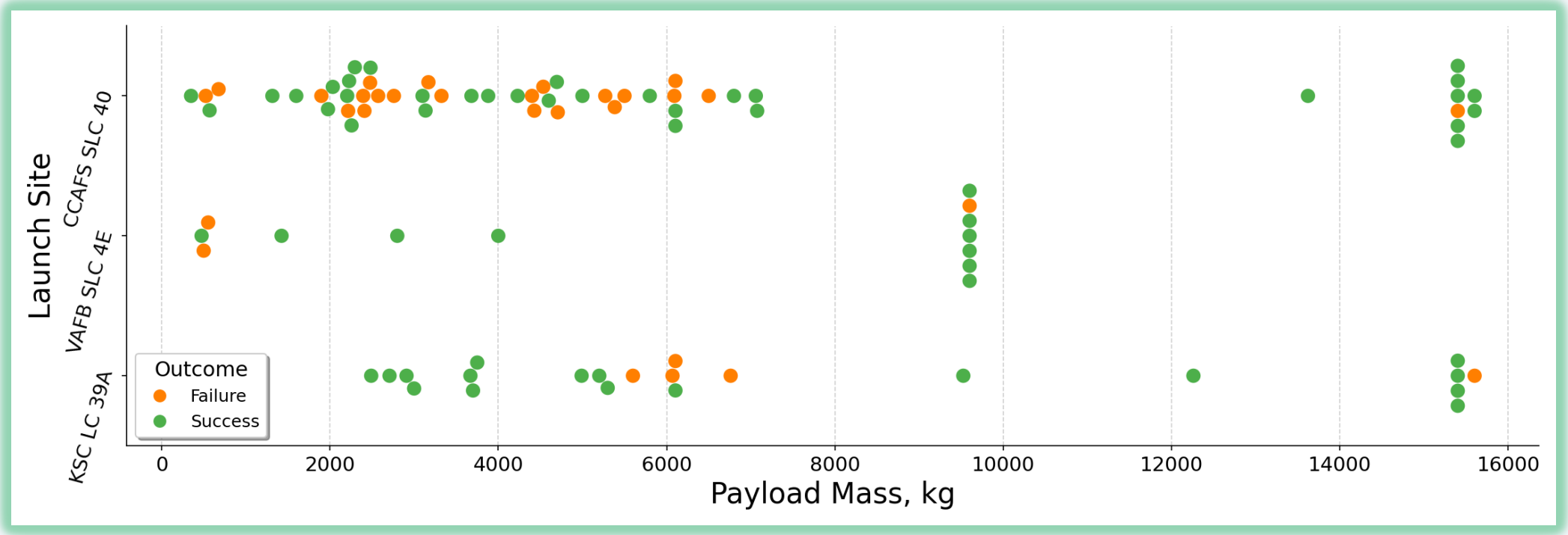

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



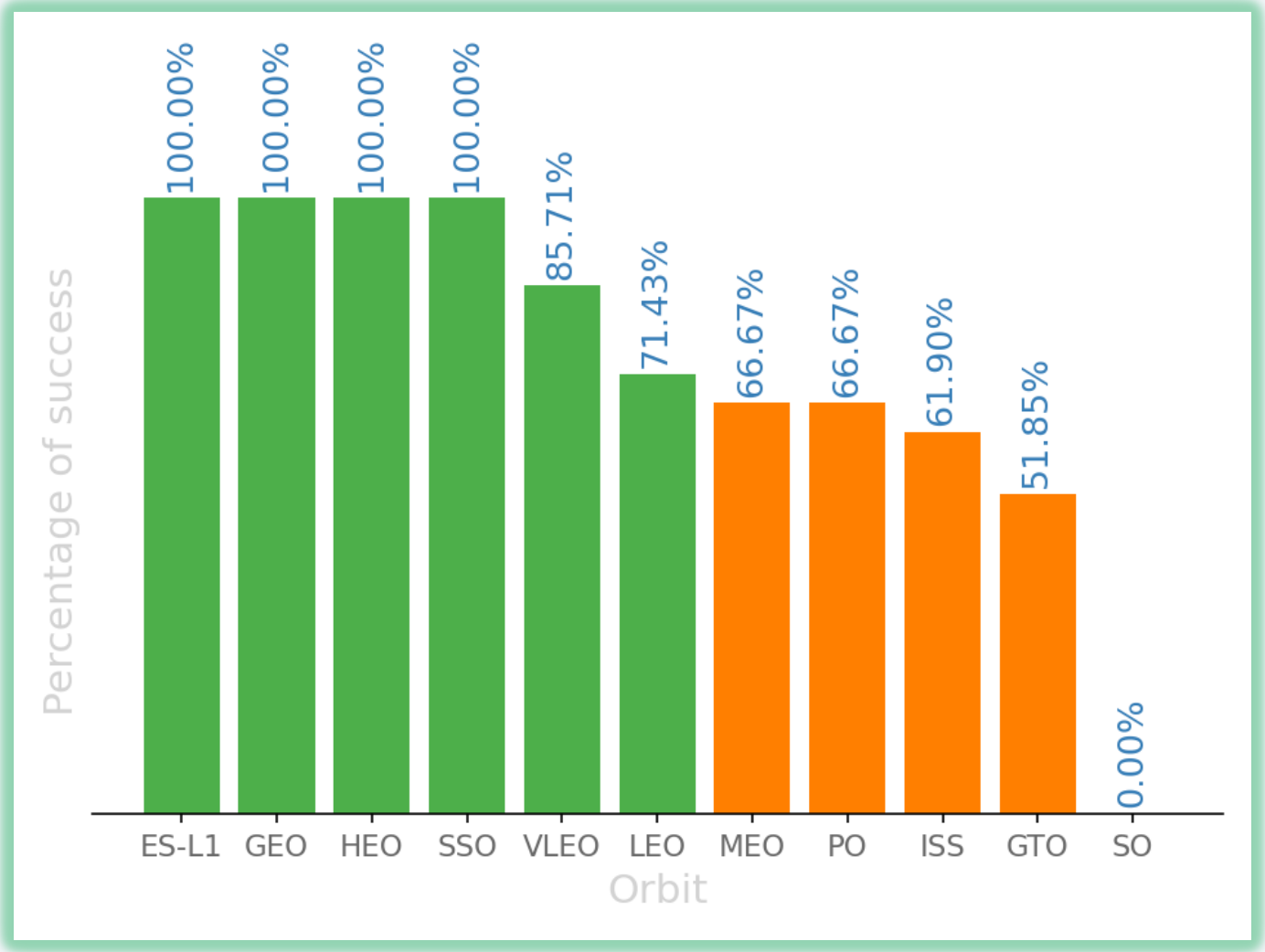
- Initial flights predominantly failed, contrasting with recent successes.
- CCAFS SLC 40 has the highest launch frequency.
- VAFB SLC 4E and KSC LC 39A show superior success rates.
- The plot suggests a progressive increase in success rates for new launches.
- Currently, CCAFS SLC 40 leads in success, followed by VAFB SLC 4E and KSC LC 39A.
- Overall success rates have improved over time, evident in the plot.

Payload vs. Launch Site



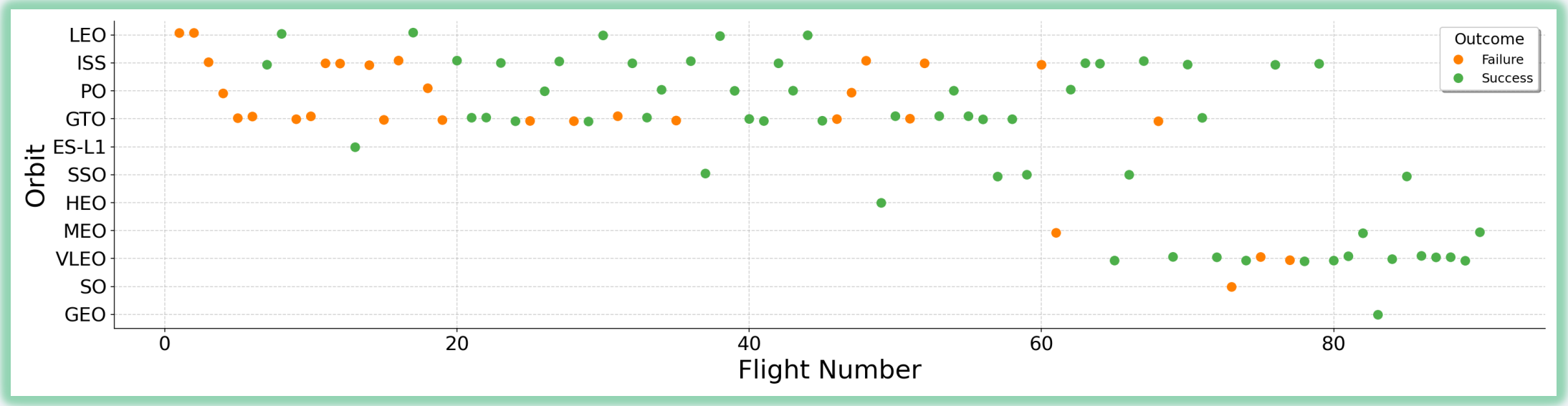
The success rate correlates positively with higher payload mass across 2/3 launch sites. Notably, payloads exceeding 7000 kg mostly achieved success. KSC LC 39A exhibited a 100% success rate for payloads under 5500 kg. Exceptional success rates were observed for payloads exceeding 9000 kg. Furthermore, payloads exceeding 12,000 kg were launched only from CCAFS SLC 40 and KSC LC 39A sites.

Success Rate vs. Orbit Type



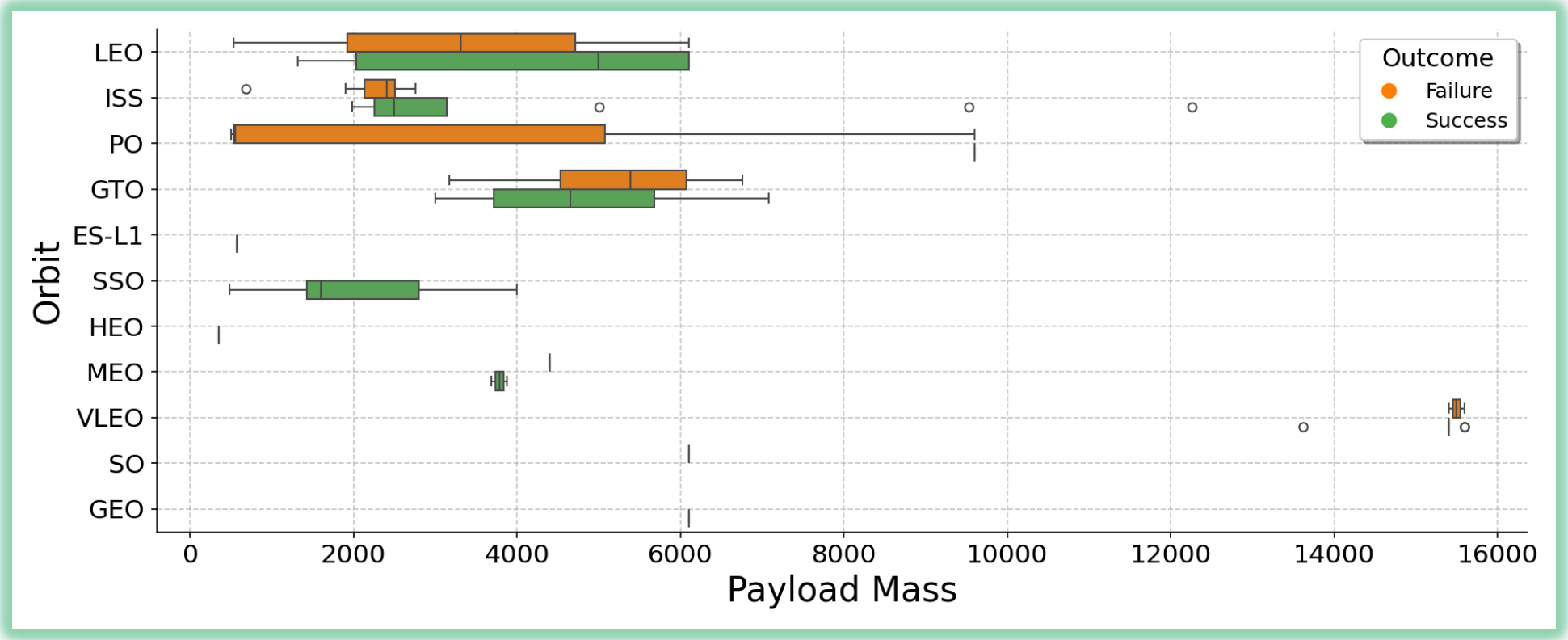
- The orbits marked in green exhibit the highest success rates, while the orange ones demonstrate an acceptable success rate.
- However, the SO orbit stands out with a 0% success rate, attributed to the sole failed attempt to reach it.

Flight Number vs. Orbit Type



The scatterplot illustrating the number of flights to different orbits reveals a continuous increase over time. Following early launches, a discernible shift occurs where subsequent launches predominantly achieve success across various orbits. Notably, a significant portion of the later launches is directed towards reaching the Very Low Earth Orbit (VLEO).

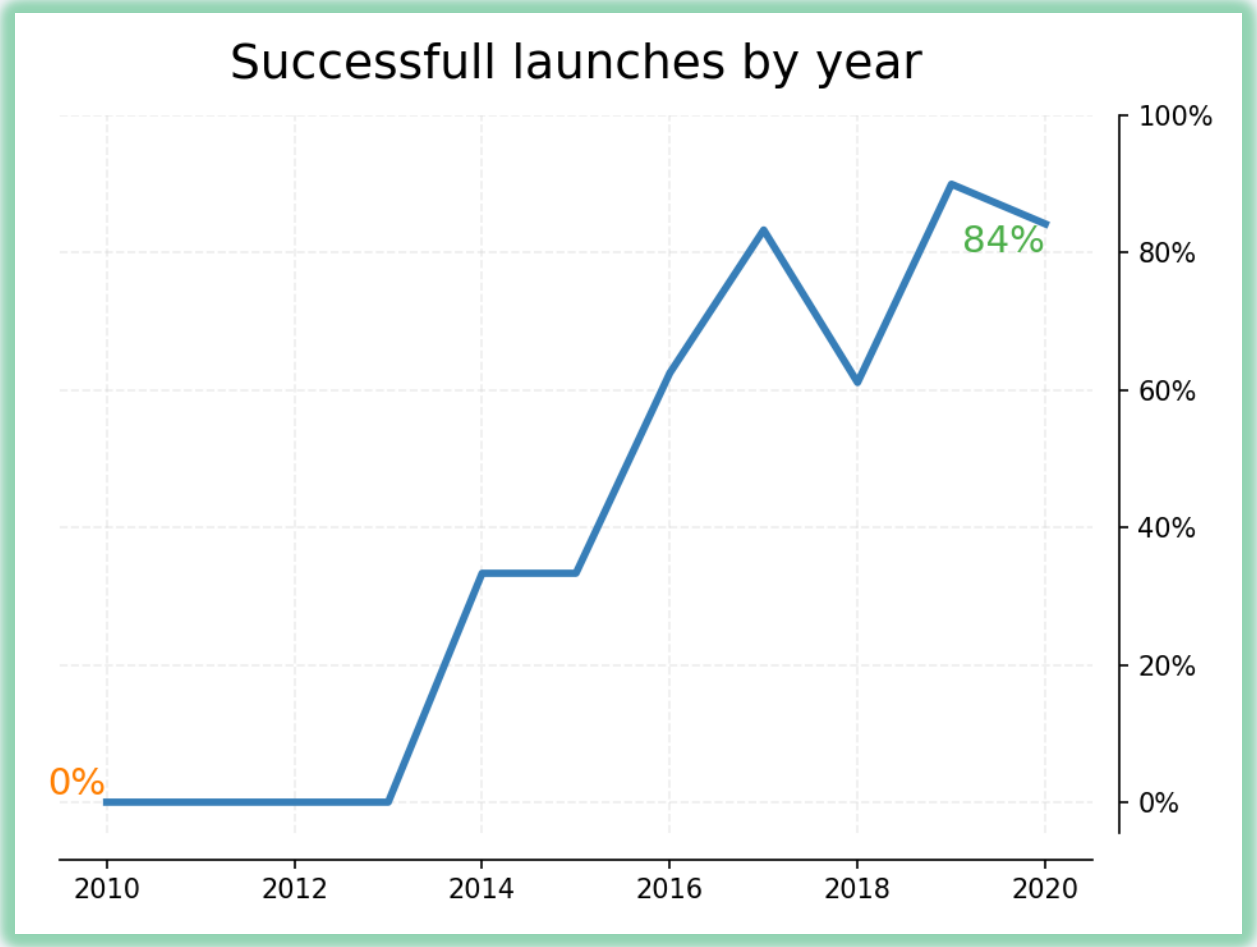
Payload vs. Orbit Type



VLEO orbit exhibit favorable characteristics when dealing with large payload masses, SSO demonstrates a notable success rate for payloads up to 4,000 kg. In contrast, the ISS is versatile and capable of accommodating both scenarios.

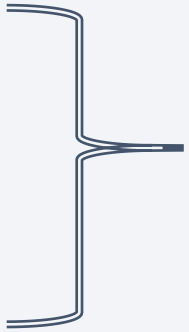
Launch Success Yearly Trend

- The initial success was achieved after 3 years of trial and error.
- The success rate surpassed 50% after 5 years.
- From 2017 to 2018 there were some failures.



All Launch Site Names

- The names of the unique launch sites in the space mission



Launch Sites
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	Teim (UTC)	Booster Version	Launch Site	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	7:44:00	F9 v1.0 B0005	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COST)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (COST)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (COST)	Success	No attempt

Total Payload Mass

The total payload mass carried by boosters launched by NASA (CRS)

```
# '\' is used to break line for readability
%sql\
SELECT\
  sum(PAYLOAD_MASS__KG_) AS Payload_from_NASA\
FROM SPACEXTBL\
WHERE Customer = "NASA (CRS)";
```

Payload from NASA
45,596

Average Payload Mass by F9 v1.1

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

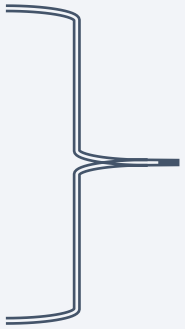
Average Payload F9 v1.1
2928.4

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

Orbit	Average Payload F9 v1.1
GTO	3676.67
LEO (ISS)	2296.0
LEO	1316.0

First Successful Ground Landing Date

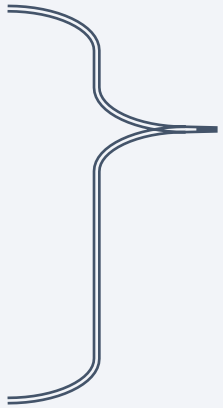
Listing the date of the initial successful ground pad landing outcome.



First Successful Landing Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

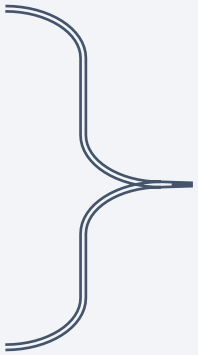
- The names of boosters which have successfully landed on drone ship and had payload mass between 4000 and 6000.



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

Total Number of Successful and Failure Mission Outcomes

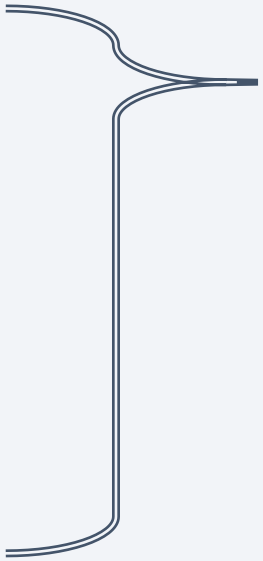
- The total number of successful and failure mission outcomes



Success	Failure
98	3

Boosters Carried Maximum Payload

The booster names that have transported the highest 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

The unsuccessful landing outcomes on drone ships, along with their booster versions and launch site names for the 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

The landing outcomes' count between the dates, ranked in descending order.

Landing Outcome	Rank
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

Location Markers for All Launch Sites on a Map

- **Proximity to Equator:**

  Most launch sites are situated near the Equator line. The Earth's surface at the equator moves faster than any other location, already at a speed of 1,670 km/hour.

- **Inertia and Space Launch:**

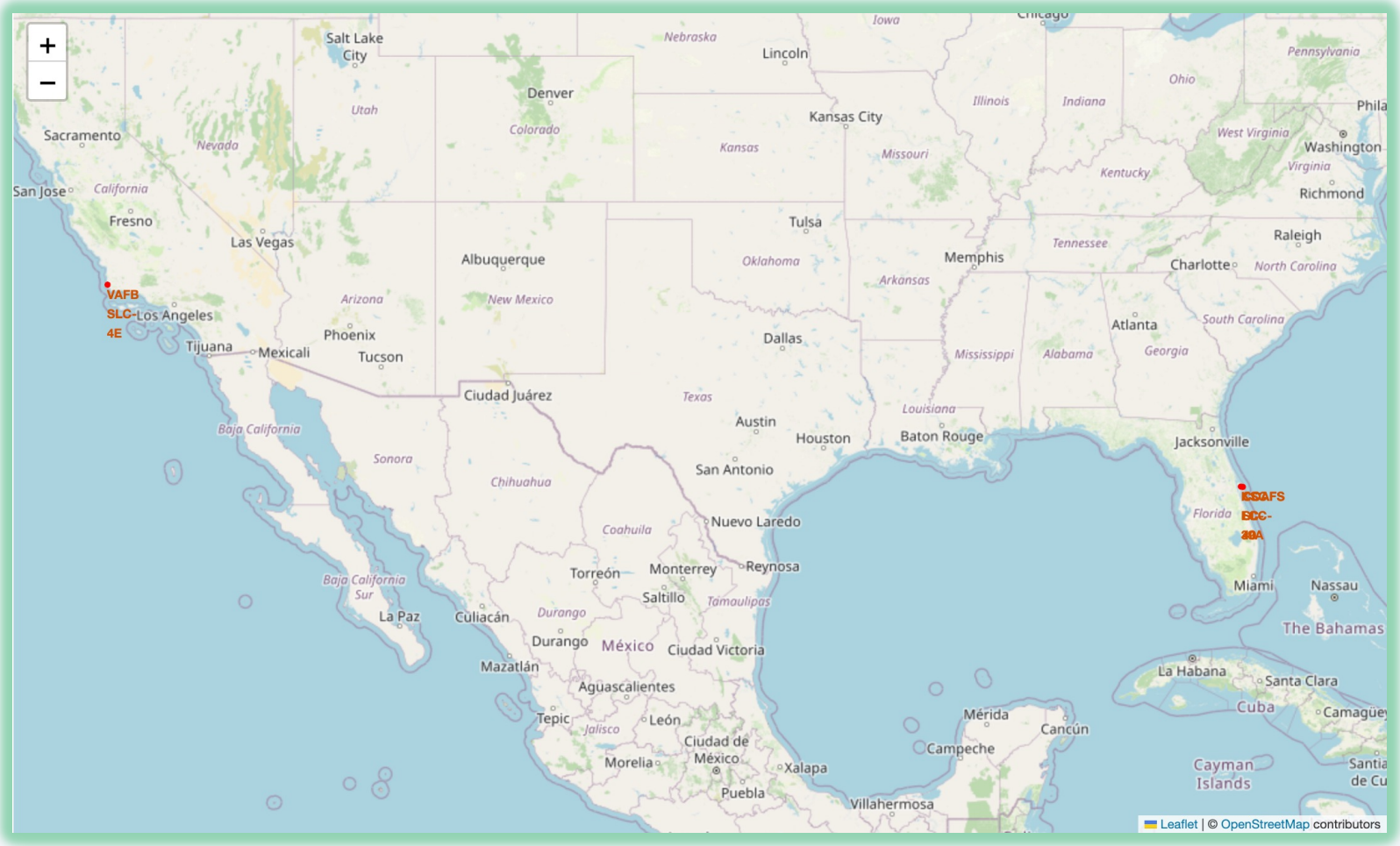
  Anything launched from the equator into space maintains the Earth's rotational speed.

  Due to inertia, the spacecraft retains the speed it had before launch, aiding in achieving and maintaining orbit.

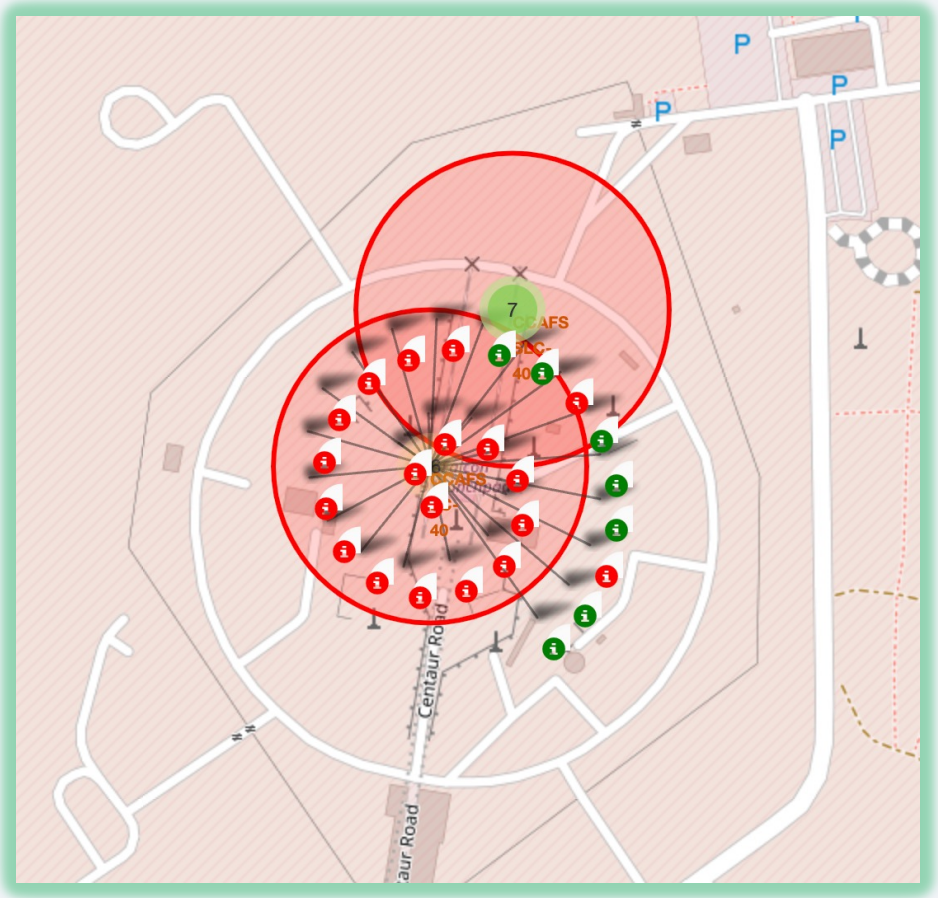
- **Coastal Locations:**

  Many launch sites are strategically located very close to the coast.

  Launching rockets towards the ocean helps minimize the risk of debris dropping or exploding near populated areas.

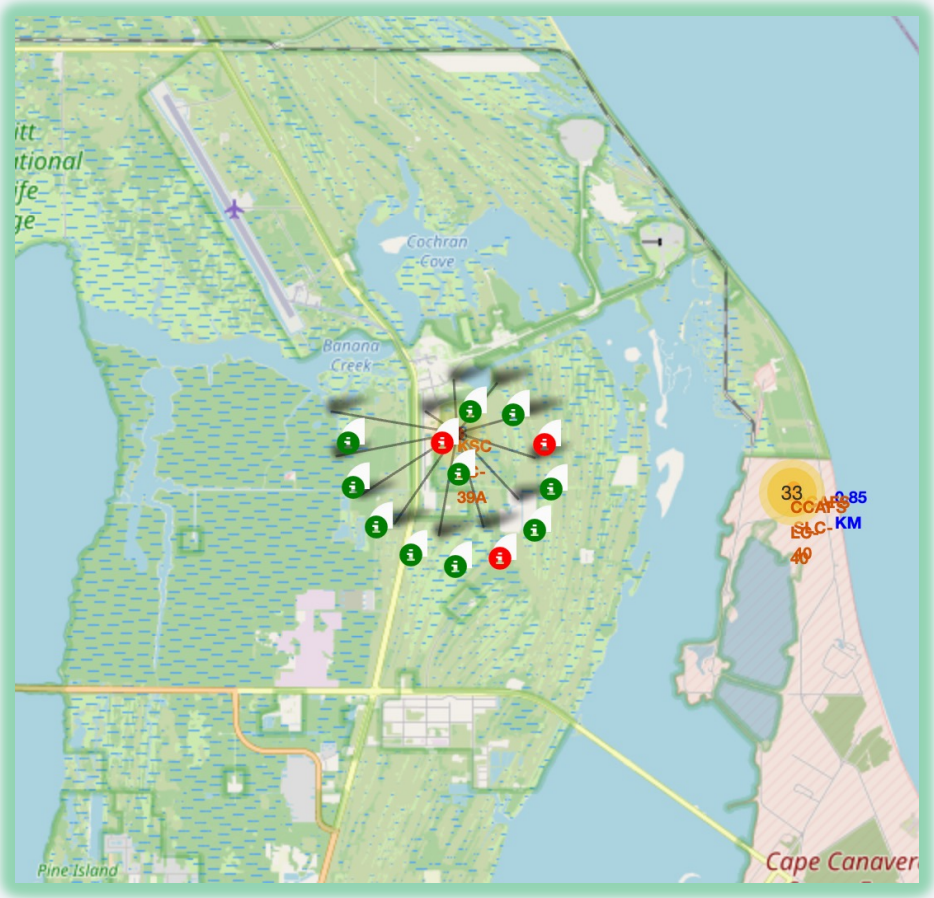


Colour-Labeled Launch Records on the Map



We should be able to easily identify launch sites with relatively high success rates based on the color-labeled markers.

- Success
- Failure

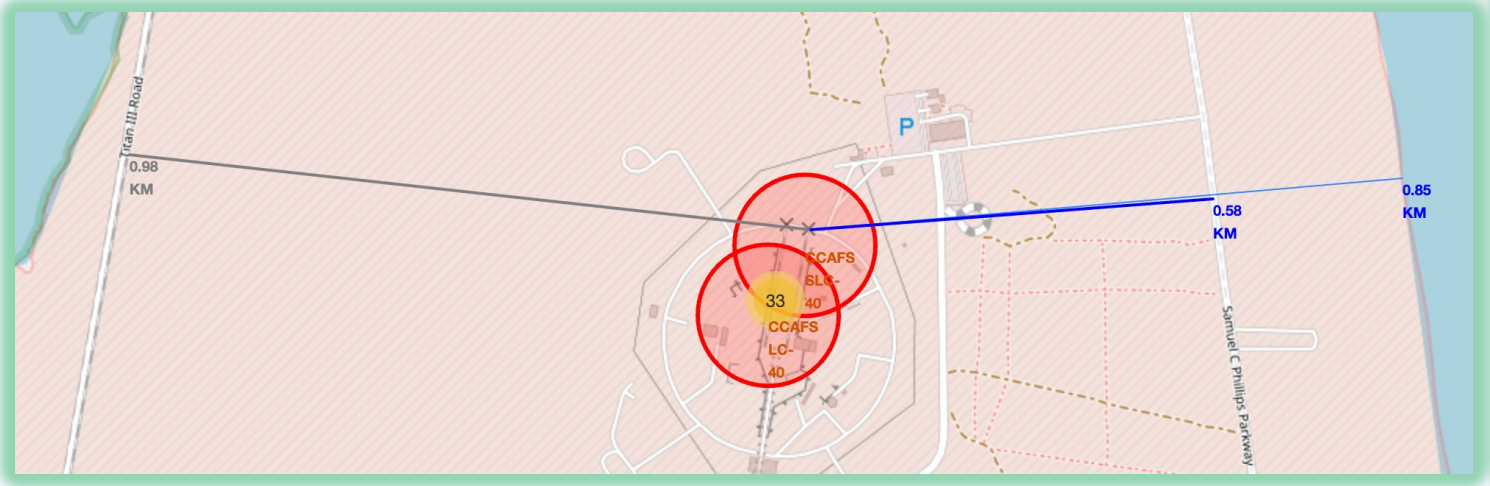


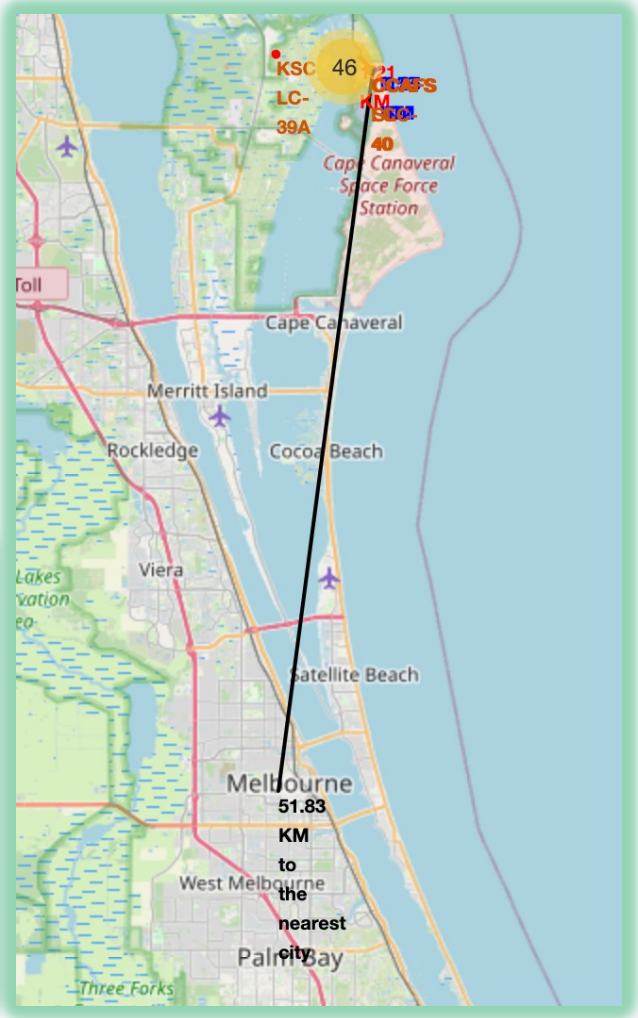
Distance from the launch site CCAFS SLC-40 to its proximities

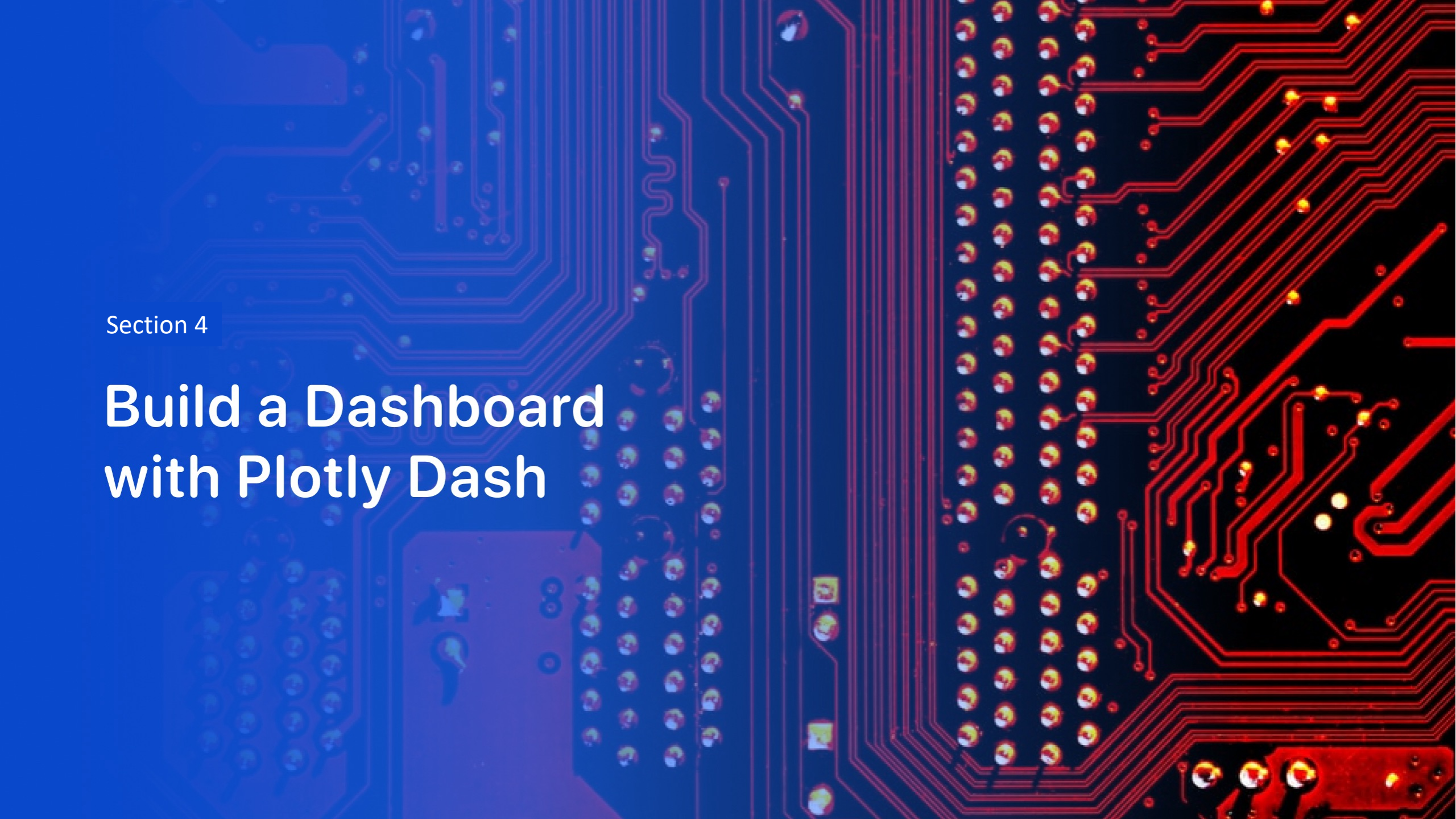
- Upon visually analysing Launch Site CCAFS SLC-40, it is evident that it is in close proximity to the following:

  Railway: 0.98 km | Highway: 0.58 km | Coastline: 0.85 km

- Additionally, the launch site is relatively close to the city of Melbourne, approximately 51.82 km away.

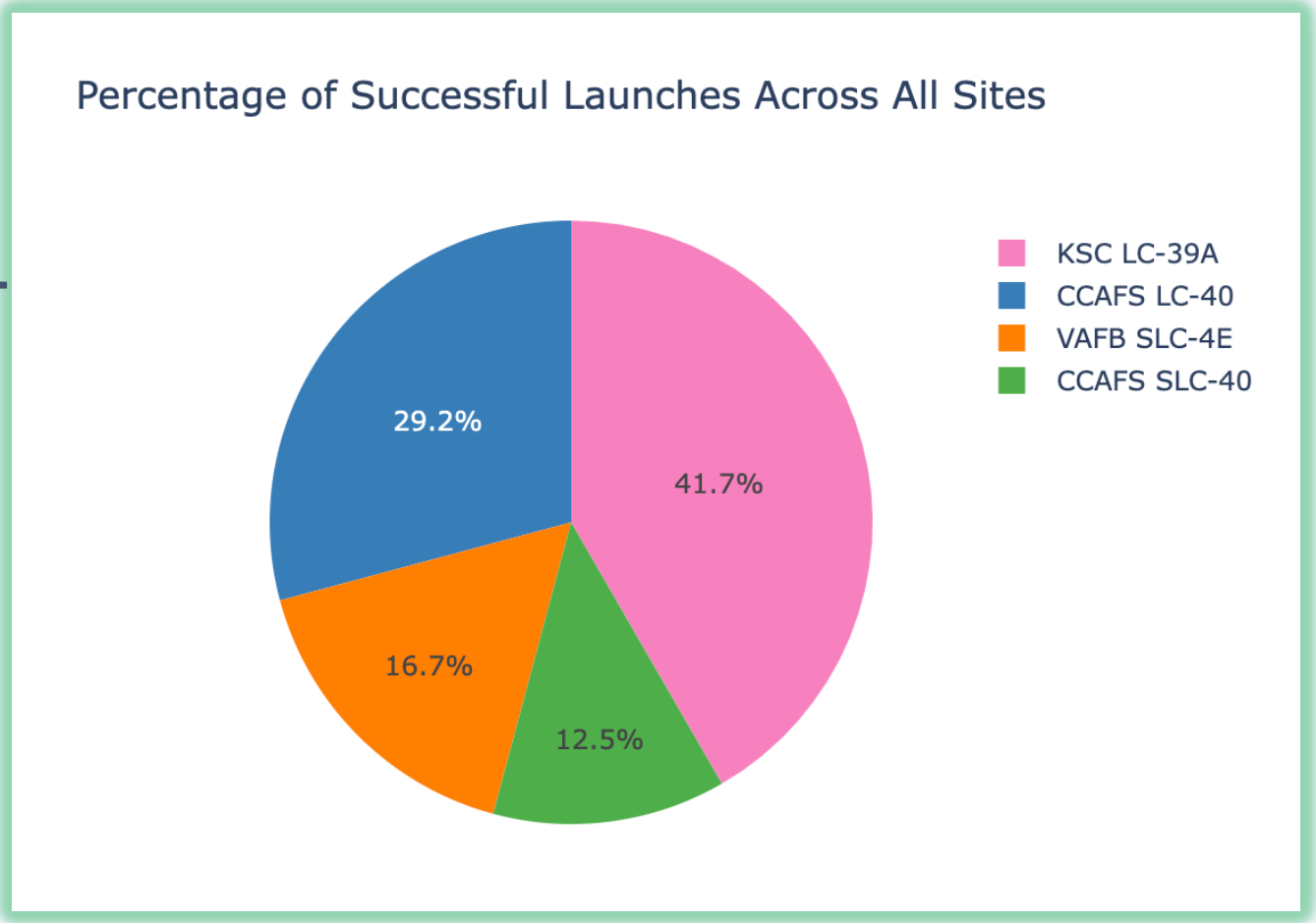





Section 4
Build a Dashboard with Plotly Dash

Total Count of Successful Launches Across All Sites

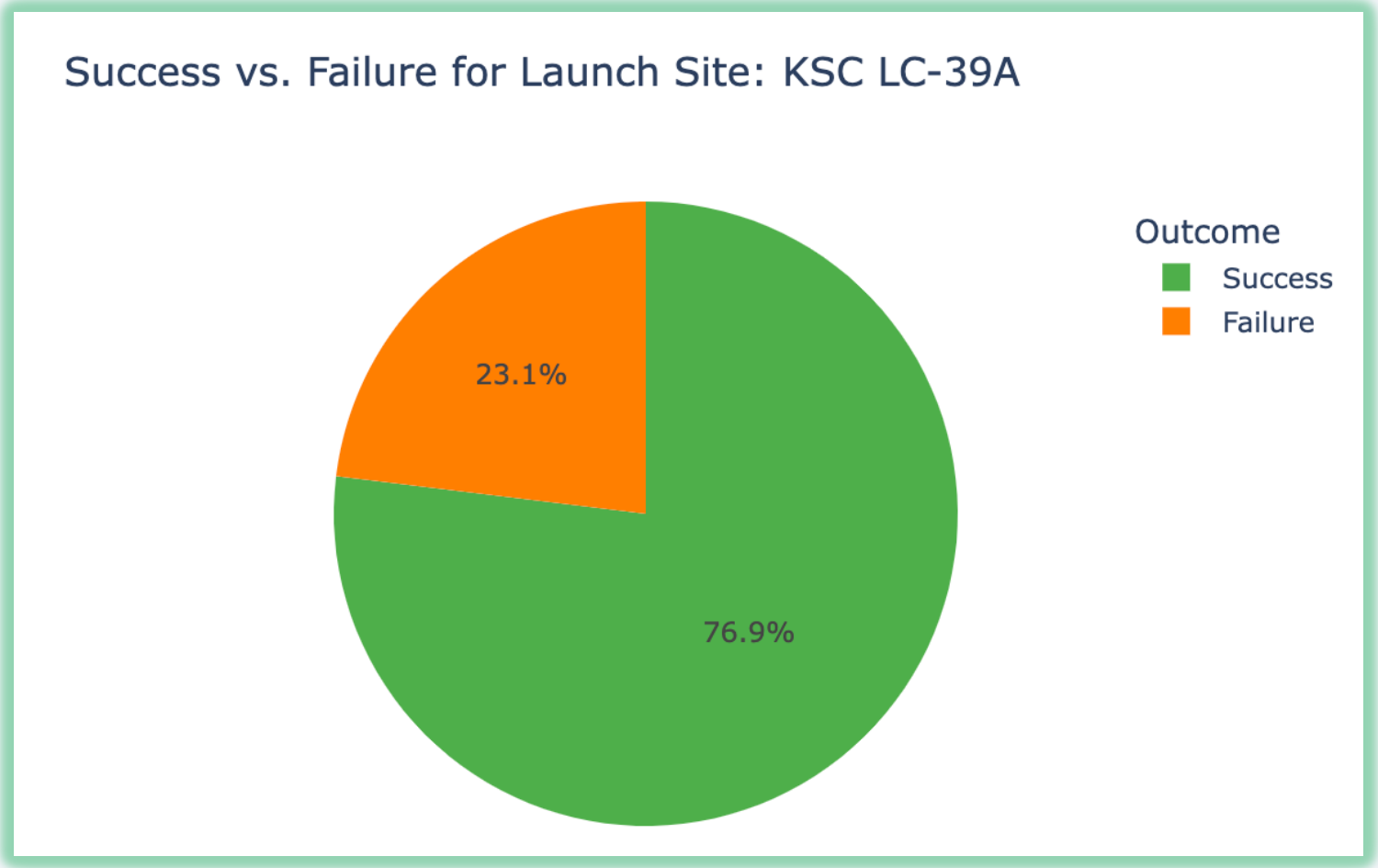
The identification of the launch site with the highest count of successful launches among the four different launch sites.



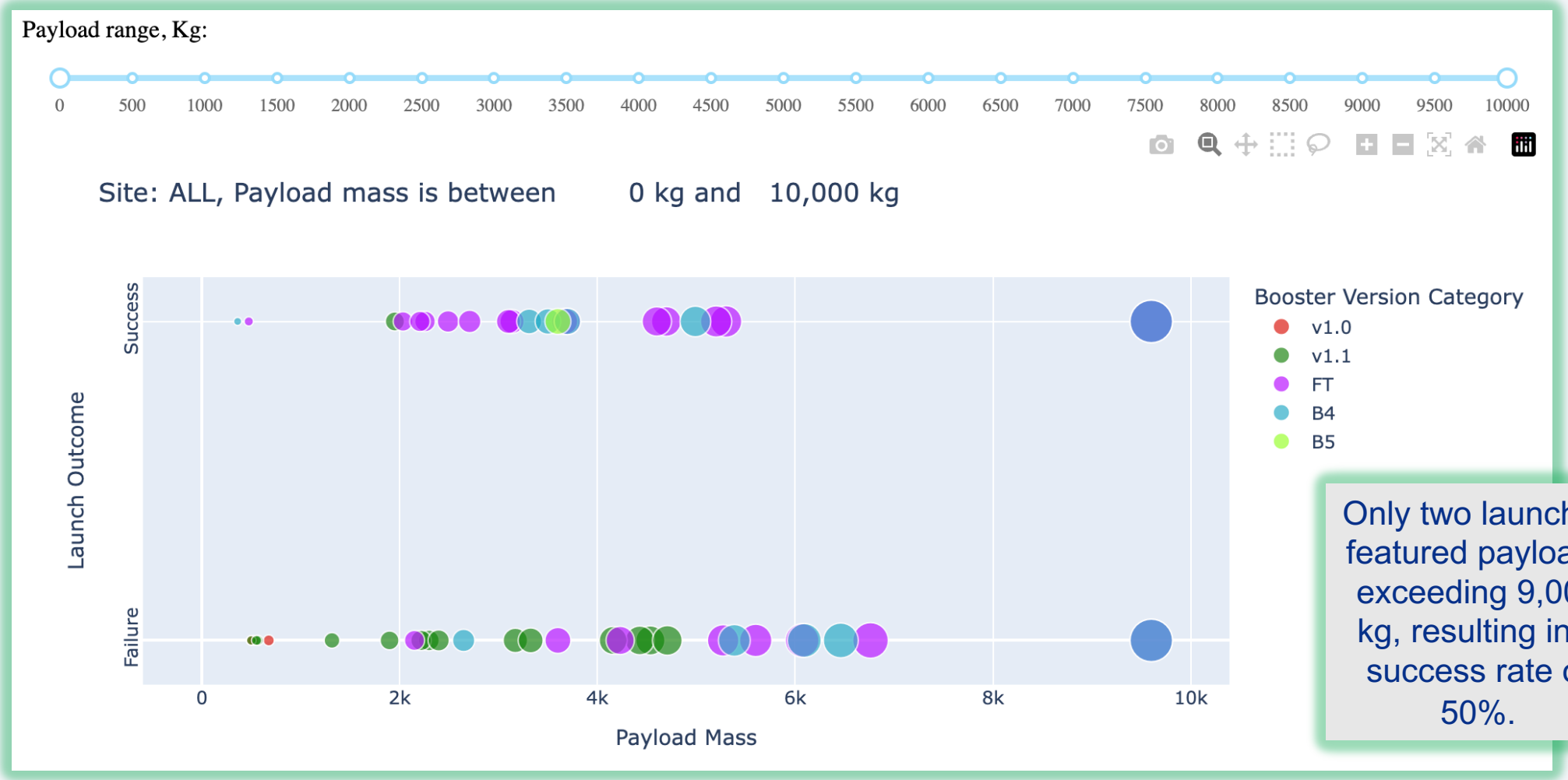
Launch Site with Highest Launch Success Ratio

KSC LC-39A stands tall as the epitome of triumph, boasting an impressive launch success rate of 76.9%.

It has orchestrated 10 triumphant launches, facing only 3 instances of fleeting setbacks.



Payload vs. Launch Outcome Scatter Plot for All Sites



Only two launches featured payloads exceeding 9,000 kg, resulting in a success rate of 50%.

Payload vs. Launch Outcome Scatter Plot for All Sites

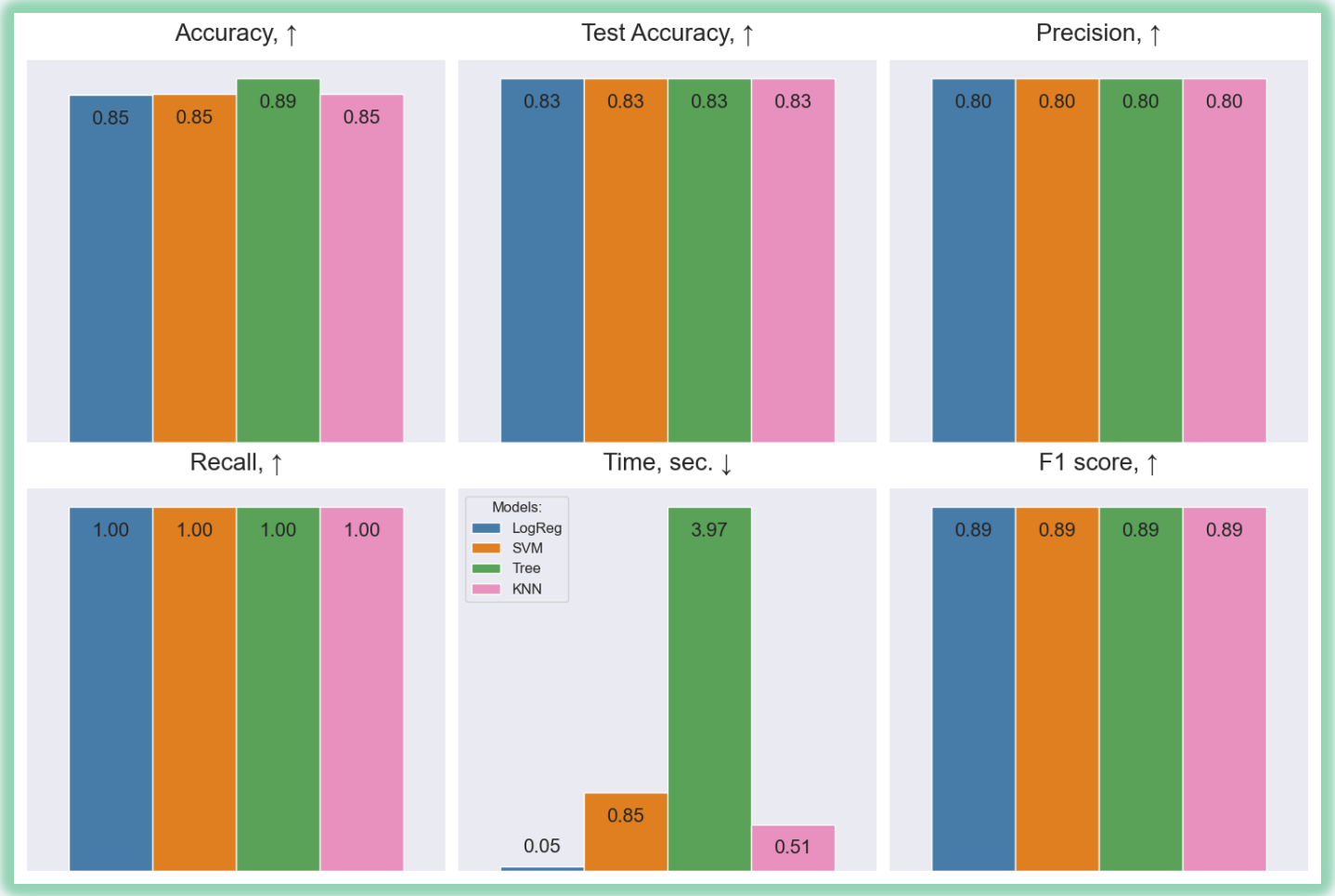
The FT series booster boasts the highest success rate for payloads under 5,500 kg.

Section 5

Predictive Analysis (Classification)

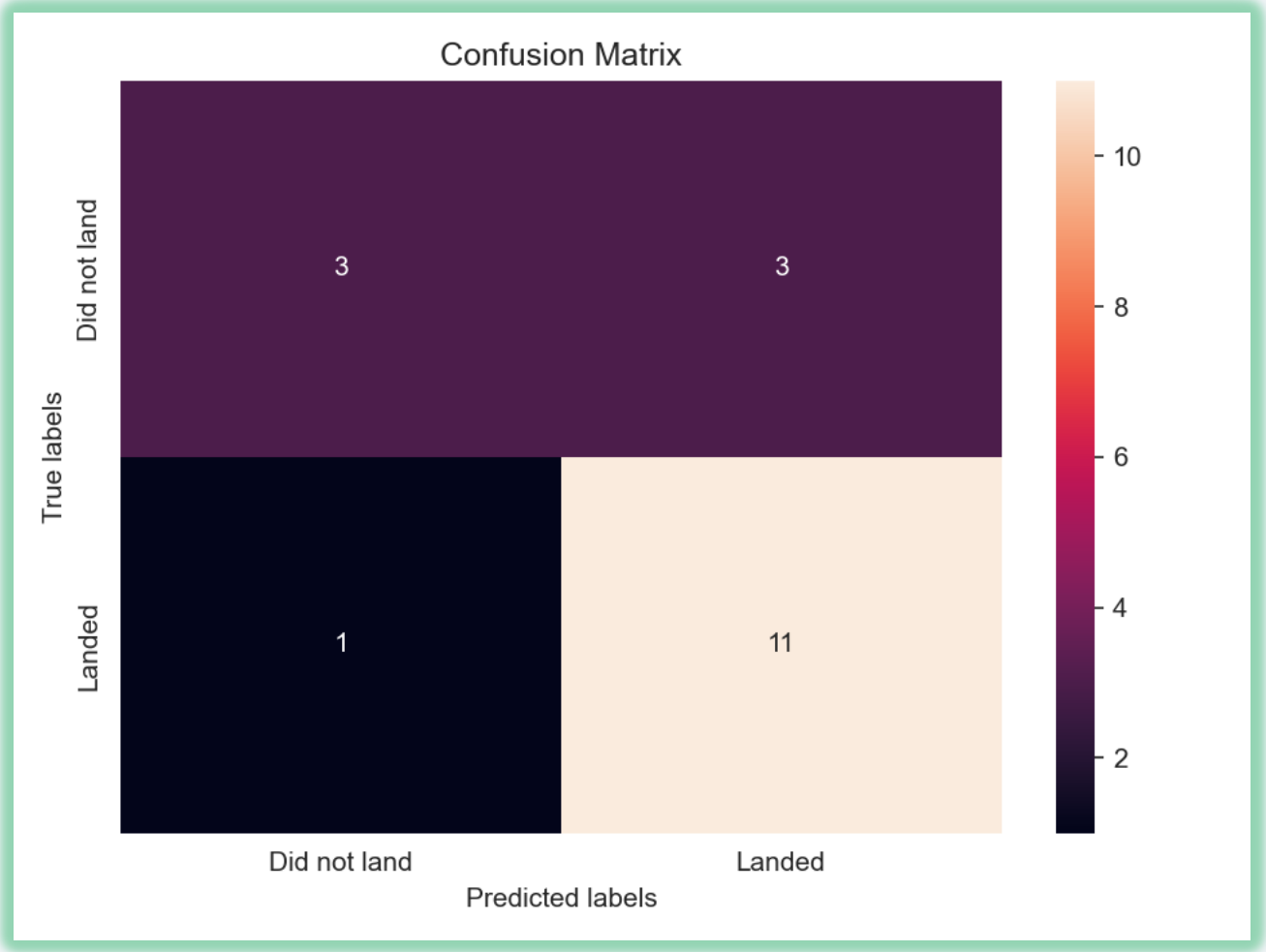
Classification Accuracy

The bar chart reveals that Decision Tree exhibits a slightly superior Accuracy, although the significance of this for classification problems may be marginal. However, this seemingly advantageous Accuracy comes at the cost of the highest execution time among the models. Considering that all other metrics are equal across the models, a crucial question arises: is it justifiable to allocate additional resources for a modest gain in Accuracy, especially when confronted with an increase in execution time?



Confusion Matrix

The Confusion Matrix for the Decision Tree Classifier serves as a testament to its accuracy, prominently displaying significant counts of true positives and true negatives when juxtaposed with the occurrences of false outcomes.



Conclusions

The analysis of the dataset reveals key insights:

- Launches with lower payload masses consistently yield better results than those with larger payload masses.
- Most launch sites are strategically situated near the Equator line, and all sites are in close proximity to the coast.
- The success rate of launches demonstrates a positive trend over the years.
- KSC LC-39A emerges as the top-performing launch site, boasting the highest success rate.
- Orbits ES-L1, GEO, HEO, and SSO achieve a perfect 100% success rate.
- The Decision Tree Model stands out as the most effective algorithm.

Further observations:

- The optimal launch site is identified as KSC LC-39A.
- Launches exceeding 7,000 kg exhibit lower risk.
- While overall mission outcomes tend to be successful, there is a notable improvement in successful landing outcomes over time, possibly attributed to the evolution of processes and rocket technology.

Appendix

Thank you!