IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sreerup Banerjee
04.08.2025

Outline

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

Executive Summary

❖ Summary of methodologies:

- Retrieving SpaceX Data via SpaceX API.
- Collecting SpaceX Information through Web Scraping.
- Cleaning and Structuring SpaceX Datasets.
- Performing SQL-based Exploratory Data Analysis on SpaceX Data.
- Visualizing SpaceX EDA with Python’s Pandas and Matplotlib.
- Analyzing SpaceX Launch Locations Using Folium and Plotly Dash.
- Predicting SpaceX Landings with Machine Learning Models.

❖ Summary of all results:

- Summary of EDA Findings.
- Building Interactive Data Visualizations and Dashboards.
- Conducting Classification-based Predictive Modeling.

Introduction

❖ Project background and context:

SpaceX promotes Falcon 9 rocket launches on its official website, listing the cost at $62 million per launch. In contrast, other launch providers charge upwards of $165 million. A major reason for SpaceX's lower pricing is its ability to reuse the rocket's first stage. By determining whether the first stage is likely to land successfully, we can more accurately estimate the total cost of a launch. This insight may prove valuable for competing companies looking to bid against SpaceX in the commercial launch market.

❖Problems we want to find answers:

This capstone project aims to predict the likelihood of a successful landing of the Falcon 9 first stage. We'll utilize publicly available launch data from SpaceX's website to build predictive models and derive actionable insights.

Section 1

Methodology

Methodology

Executive Summary:

- ❖ Data collection methodology:
  - Description of how data was collected.
- ❖ Perform data wrangling:
  - Description of how data was processed.
- ❖ Perform exploratory data analysis (EDA) using visualization and SQL.
- ❖ Perform interactive visual analytics using Folium and Plotly Dash.
- ❖ Perform predictive analysis using classification models:
  - Process of build, tune, evaluate classification models.

Data Collection

❖ Description of how data sets were collected:

- Falcon 9 launch data was initially gathered using the SpaceX RESTful API. This involved making GET requests to the API after defining a set of helper functions. These functions facilitated the retrieval of specific information from the launch data by referencing identification numbers.
- To standardize the JSON responses, the API data was requested and parsed via GET requests. The response content was decoded into a JSON format and subsequently converted into a Pandas DataFrame for further analysis.
- A web scraping was employed to collect historical Falcon 9 launch records from the Wikipedia page titled *List of Falcon 9 and Falcon Heavy launches*. These records were stored in an HTML table. Using the BeautifulSoup and Requests libraries, the HTML table was extracted, parsed and then transformed into a Pandas DataFrame.

Data Collection – SpaceX API

- The Falcon 9 launch data was gathered using the SpaceX RESTful API. A GET request was made to the API to retrieve launch data. The response was parsed and decoded from JSON format, which was then converted into a Pandas DataFrame for analysis.

- The complete implementation of the SpaceX API data collection process is documented in the following GitHub notebook: [SpaceX Data Collection API.]

FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins
1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False
2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False
3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False
4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False
5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False

Data Collection - Scraping

- ❖ Utilized web scraping techniques to gather historical Falcon 9 launch data from Wikipedia by employing BeautifulSoup and requests. Extracted the launch information from an HTML table on the page and constructed a data frame by parsing the HTML content.
- ❖ The following link takes to the finalized web scraping notebook on GitHub: *[Spacex Webscrapping](#)*.

Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date
Dragon Spacecraft Qualification Unit	0	LEO	[[SpaceX], \n]	Success\n	F9 v1.07B0003.18	Failure	4 June 2010
Dragon	0	LEO	[[.mw-parser-output .plainlist ol,.mw-parser-o...	Success	F9 v1.07B0004.18	Failure	8 December 2010
Dragon	525 kg	LEO	[[NASA], (, [COTS],)\n]	Success	F9 v1.07B0005.18	No attempt\n	22 May 2012
SpaceX CRS-1	4,700 kg	LEO	[[NASA], (, [CRS],)\n]	Success\n	F9 v1.07B0006.18	No attempt	8 October 2012
SpaceX CRS-2	4,877 kg	LEO	[[NASA], (, [CRS],)\n]	Success\n	F9 v1.07B0007.18	No attempt\n	1 March 2013

Data Wrangling

- After acquiring and constructing a Pandas DataFrame from the scraped dataset, the data was filtered through the BoosterVersion column to retain only Falcon 9 launch entries. Missing values in the LandingPad and PayloadMass columns were then addressed specifically, PayloadMass nulls were imputed using the column's mean.
- Conducted Exploratory Data Analysis (EDA) to uncover patterns within the dataset and identify a suitable label for training supervised learning models.
- The GitHub link for the finished web scraping notebook is : *Spacex Data wrangling*.

Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

EDA with Data Visualization

- ❖ Carried out Data Analysis and Feature Engineering using Pandas and Matplotlib, including:
  - Exploratory Data Analysis.
  - Data preparation and Feature Engineering.
- ❖ Visualized relationships through scatter plots between:
  - Flight Number and Launch Site.
  - Payload and Launch Site.
  - Flight Number and Orbit type.
  - Payload and Orbit type.
- ❖ Employed a bar chart to illustrate the success rate across different orbit types.
- ❖ Applied a line plot to show the annual trend of launch success.
- ❖ Here is the GitHub link to the finalized notebook containing EDA and data visualizations: *SpaceX EDA Data Visualization*.

EDA with SQL

❖ Executed a series of SQL queries for Exploratory Data Analysis (EDA), including:
- Retrieving names of distinct launch sites involved in the space missions.
- Displaying five records where launch sites start with the prefix 'CCA'.
- Calculating total payload mass for boosters launched under NASA's CRS missions.
- Determining the average payload mass carried by booster version F9 v1.1.
- Identifying the date of the first successful ground pad landing.
- Listing booster names that succeeded on drone ships with payloads between 4000 and 6000.
- Counting the number of missions with successful and failed outcomes.

❖ Below is the GitHub URL to the finalized EDA notebook utilizing SQL queries: *SpaceX EDA SQL*.

Build an Interactive Map with Folium

- ❖ Developed an Interactive Launch Site Map Using Folium:
  - Plotted all launch site locations on a Folium map.
  - Implemented map elements including markers, circles, and lines to visually indicate launch outcomes (success or failure).
  - Engineered a binary launch outcome dataset with 0 representing failure and 1 representing success.
- ❖ Below is the GitHub URL linking to the completed interactive visualization: *Spacex Site Location*.

Build a Dashboard with Plotly Dash

- ❖ Developed an Interactive Dashboard Application Using Plotly Dash:
  - Integrated a Launch Site drop-down input component.
  - Implemented a callback function to render a success pie chart based on the selected launch site.
  - Added a range slider to allow payload selection.
  - Incorporated a callback function to generate a success-payload scatter plot based on user input.
- ❖ Below is the GitHub URL for the finalized Plotly Dash application notebook: *Spacex Dash App*.

Predictive Analysis (Classification)

Predictive Modeling Workflow Summary:

- ❖ Loaded & preprocessed data:
  - Created Y from the Class column using NumPy.
  - Standardized features using StandardScaler.
  - Split data into training/testing (80/20).
- ❖ Trained and evaluated 4 ML models:
  - SVM, Decision Tree, KNN, Logistic Regression.
  - Used GridSearchCV (cv=10) for hyperparameter tuning.
  - Selected best params via best_score_ and best_params_.
- ❖ Final evaluation:
  - Calculated accuracy on test set using score().
  - Visualized performance with confusion matrices.
- ❖ Here is the GitHub link to the finalized notebook containing machine learning models for prediction: *Spacex Machine Learning Prediction*.

Results

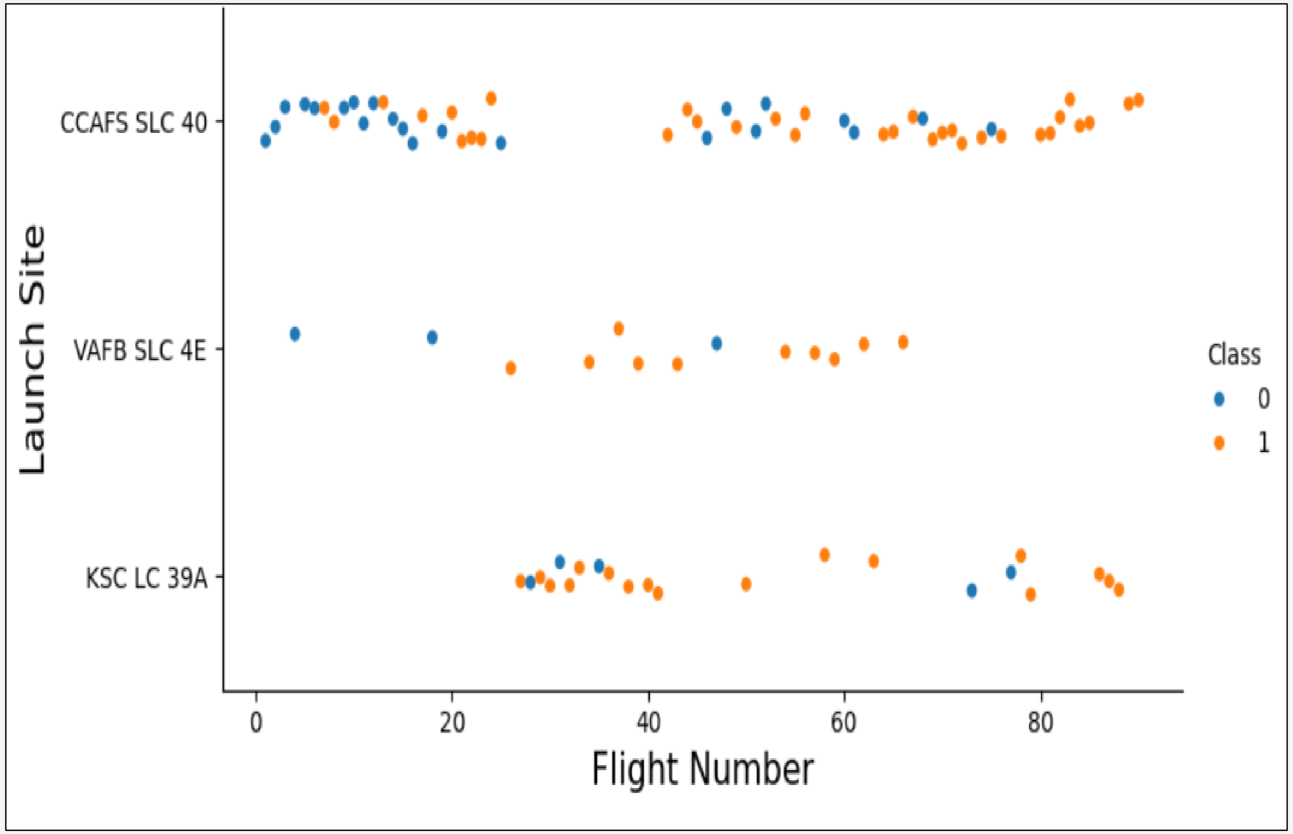
- ❖ Exploratory data analysis results.
- ❖ Interactive analytics demo in screenshots.
- ❖ Predictive analysis results.
- ❖ All the project works are uploaded in my Github Repository: *SpaceX Falcon9 Landing*.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

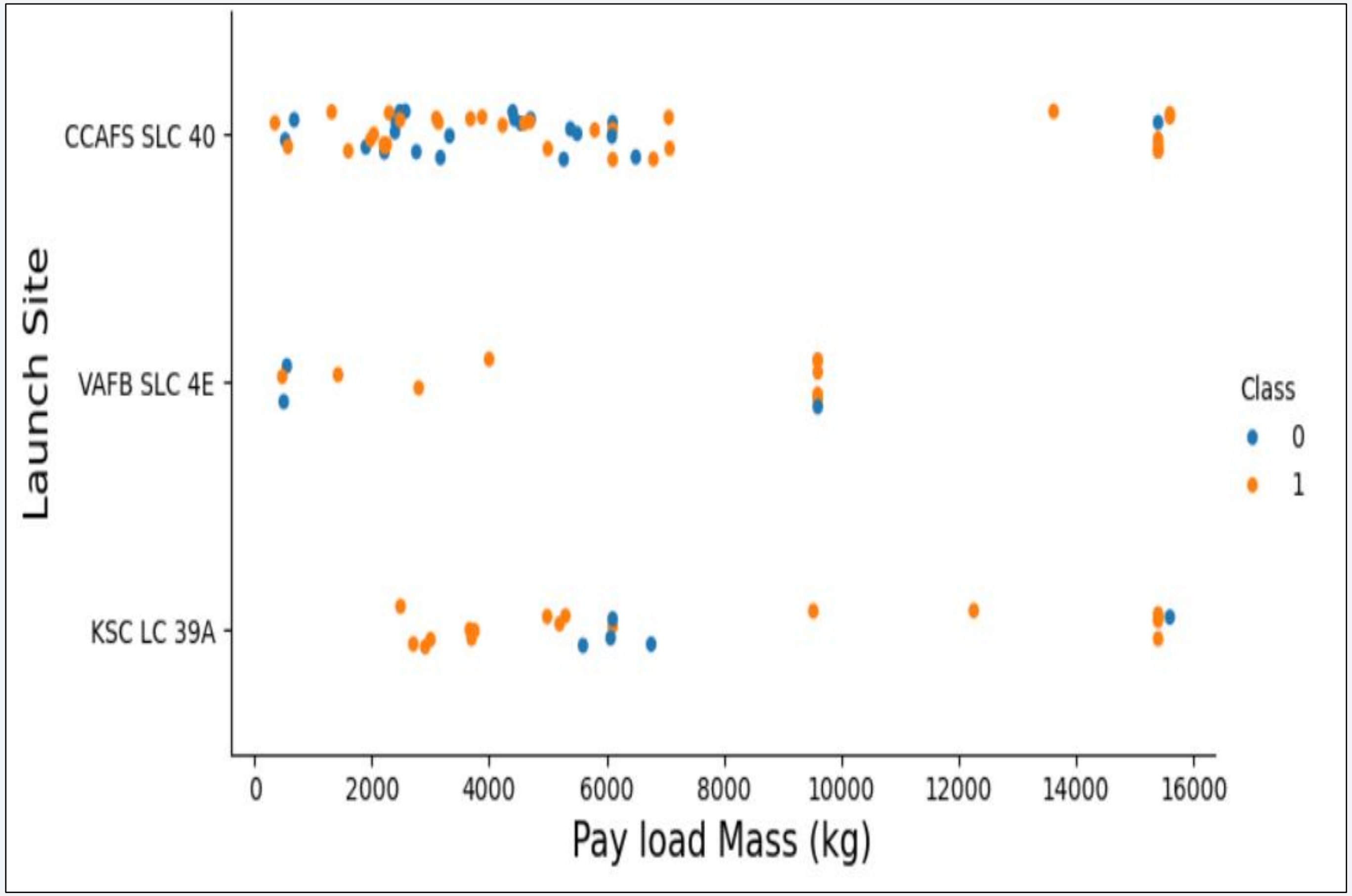
❖ Scatter plot of Flight Number vs. Launch Site:



- Launch Site CCAFS SLC 40 records the highest number of launches.
- Class 0 (Failed) is frequently observed in early flight numbers.
- Falcon 9 first stage failures are represented by Class '0' (● blue markers), while successful landings are shown by Class '1' (● orange markers).

Payload vs. Launch Site

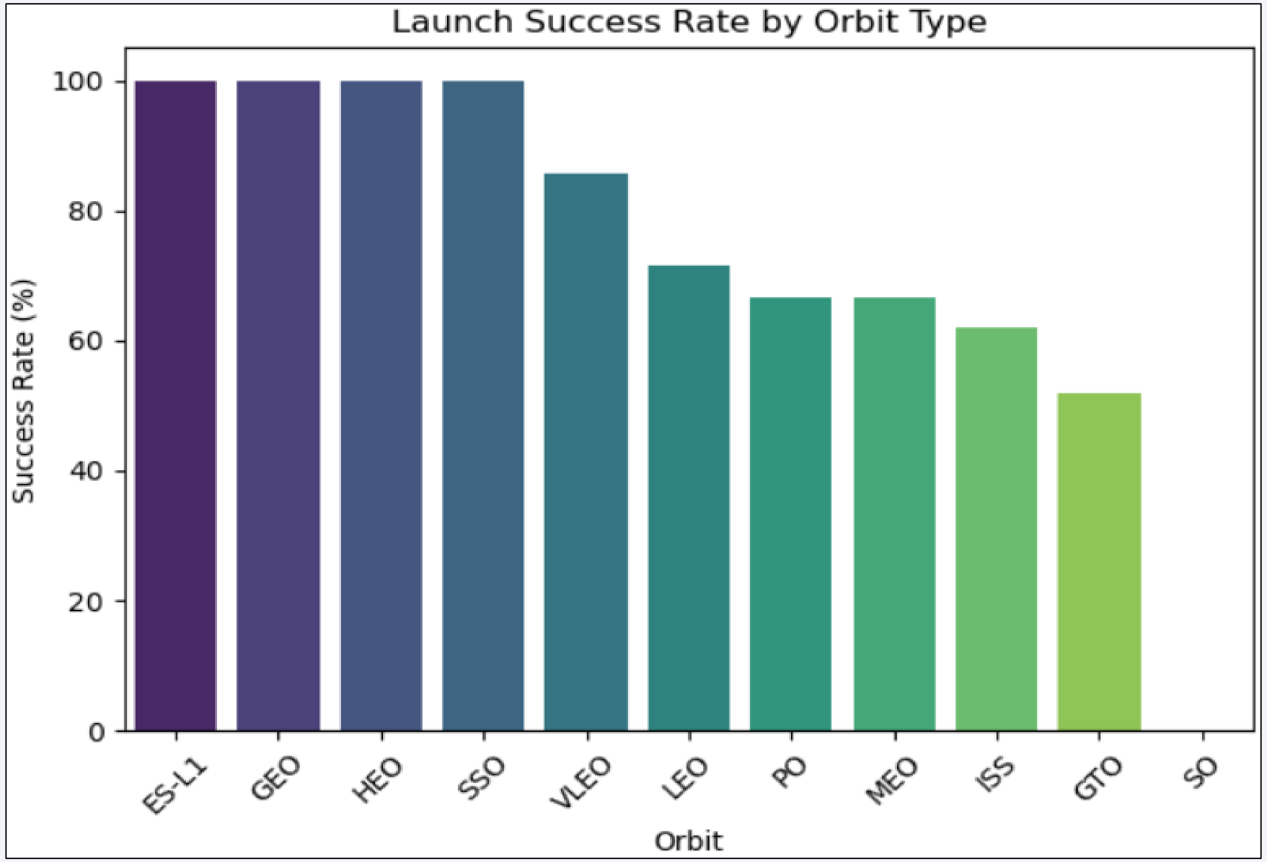
❖ Scatter plot of Payload vs. Launch Site:



- Launch Site VAFB SLC 4E does not feature any launches with payload masses exceeding 10,000 kg.
- At the KSC LC 39A launch site, failed landings are concentrated within a narrow range of payload masses.
- Falcon 9 first stage failures are represented by Class '0' (● blue markers), while successful landings are shown by Class '1' (● orange markers).

Success Rate vs. Orbit Type

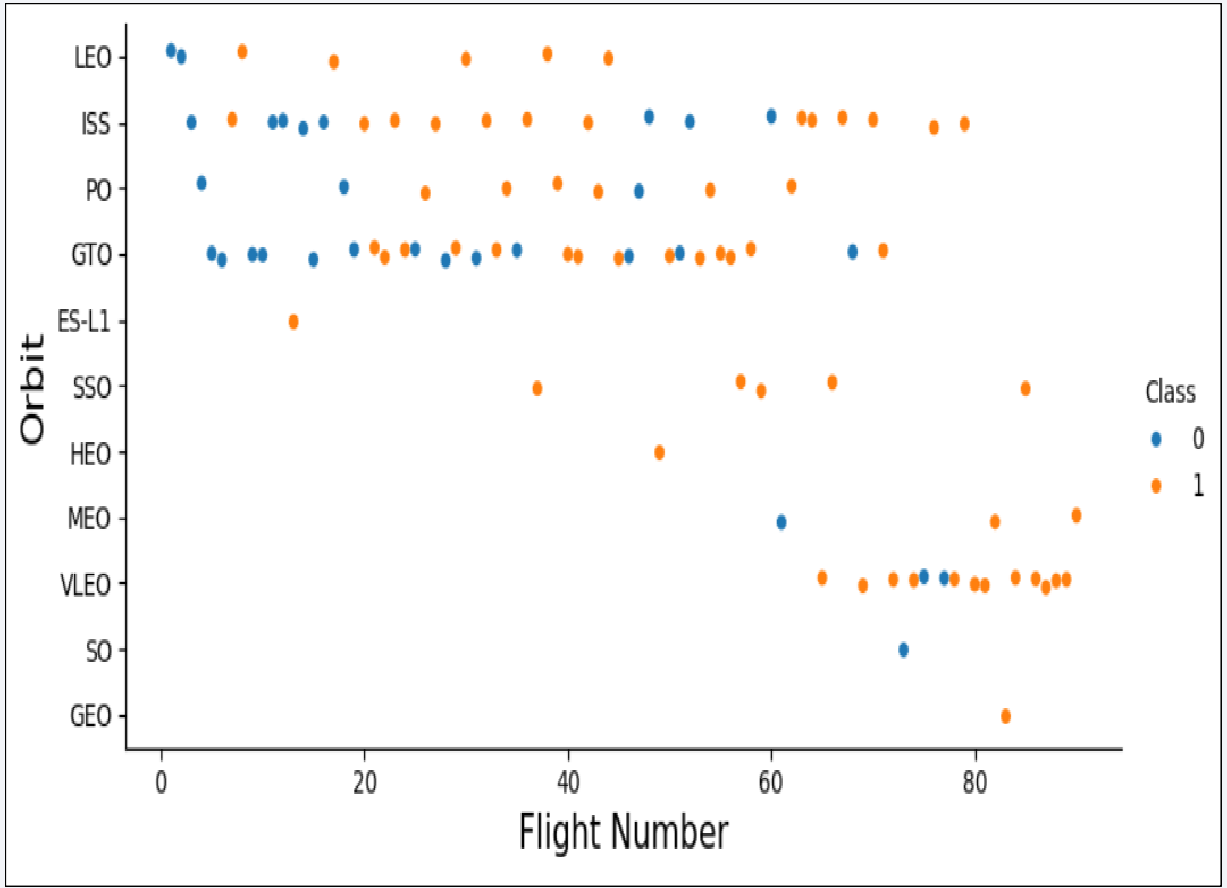
❖ Bar chart for the success rate of each orbit type:



- Orbits ES-L1, GEO, HEO, and SSO demonstrate the highest success rates.
- No successful first stage landings have occurred in SO orbits.
- In the bar chart of Falcon 9 First Stage Landing, from Purple (●) shows 100% success to white (●) shows 0% success.

Flight Number vs. Orbit Type

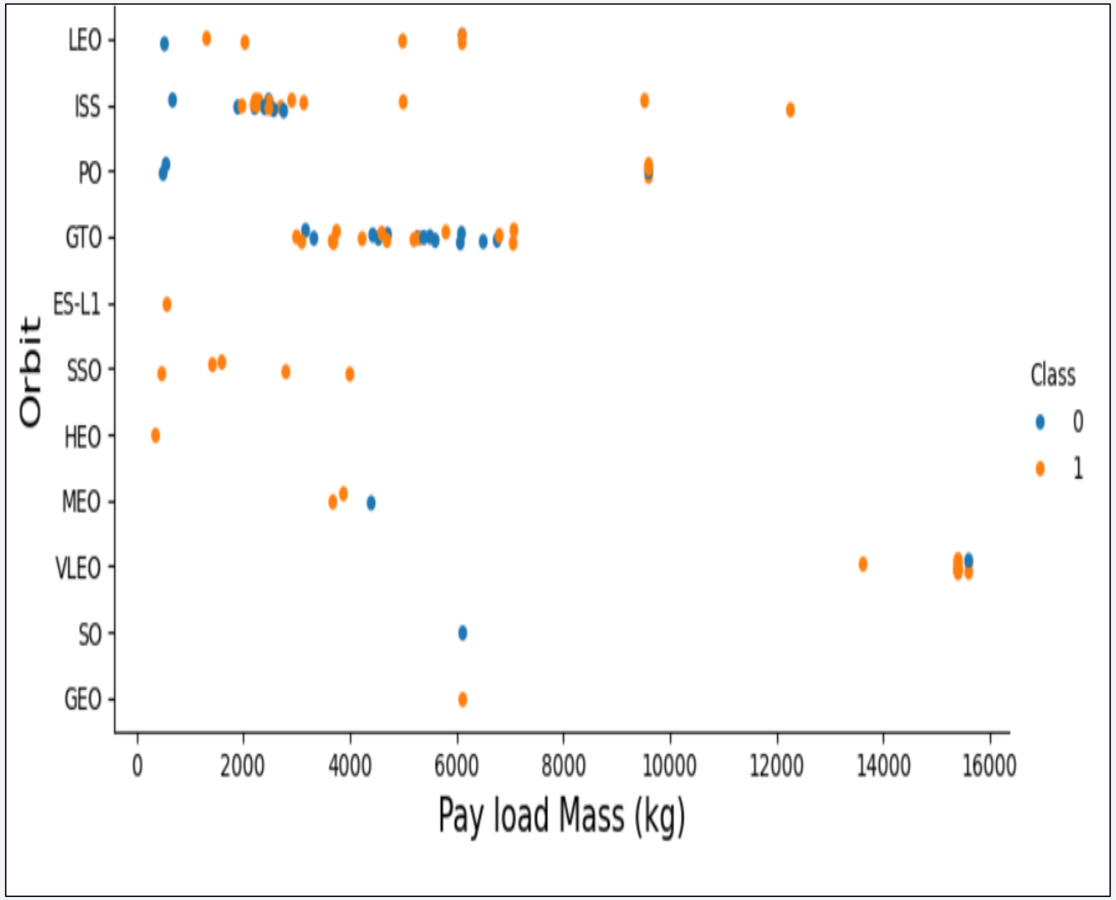
❖ Scatter plot of Flight Number vs. Orbit Type:



- A correlation exists between flight number and success rate, with higher flight numbers linked to increased success rates.
- Orbits MEO, VLEO, SO, and GEO are observed beginning at flight number 60.
- Falcon 9 first stage failures are represented by Class '0' (● blue markers), while successful landings are shown by Class '1' (● orange markers).

Payload vs. Orbit Type

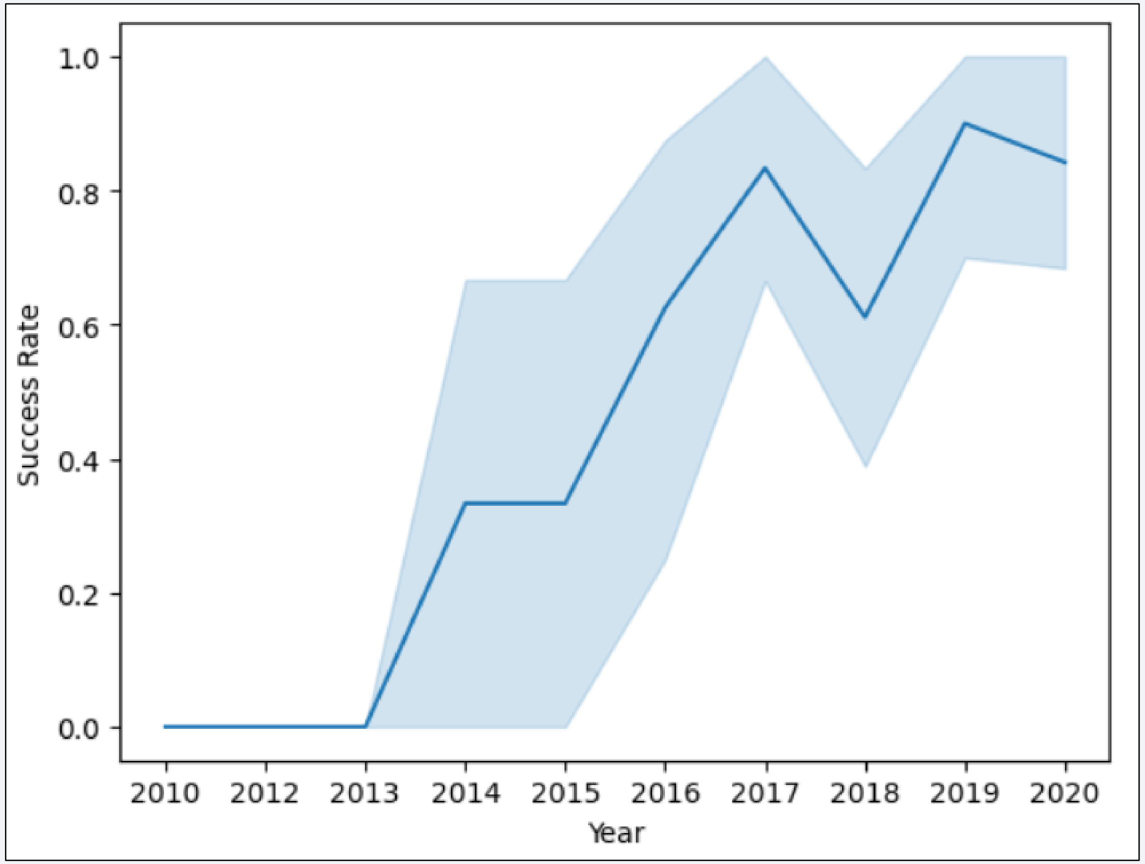
❖ Scatter plot of Payload vs. Orbit Type:



- Certain orbit types exhibit higher success rates than others.
- There is no clear correlation between success rate and payload mass.
- VLEO (Very Low Earth Orbit) utilizes the highest payload mass among the observed orbit types.
- Falcon 9 first stage failures are represented by Class '0' (● Blue markers), while successful landings are shown by Class '1' (● Orange markers).

Launch Success Yearly Trend

❖ Line chart of yearly average success rate:



- The success rate of Falcon 9's first-stage landings has shown a significant upward trend over the years.
- Notably, 2018 saw a dip in performance, with the success rate dropping to 60%, compared to 80% in 2017.
- In the year-by-year analysis of Falcon 9 First Landing Success Rate, the trend represents the percentage of successful landings in Sky blue (●) .

All Launch Site Names

- ❖ Query: %sql SELECT DISTINCT LAUNCH_SITE as Launch_Site FROM SPACEXTABLE
- ❖ Result:

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

- ❖ Explanation: There are four distinct launch sites.

Launch Site Names Begin with 'CCA'

❖ Query: %sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5

❖ Result:

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

❖ Explanation: A simple sampling method to assess database content.

Total Payload Mass

- Query: %sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass(Kgs)", Customer FROM SPACEXTBL WHERE Customer = 'NASA (CRS)’
- Result:

Total Payload Mass(Kgs)	Customer
45596	NASA (CRS)

- Explanation: Boosters used for NASA’s CRS missions have transported a total payload of 45,596 kg.

Average Payload Mass by F9 v1.1

- ❖ Query: %sql SELECT AVG(PAYLOAD_MASS__KG_) as "Average Payload Mass(Kgs)", Customer, Booster_Version FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1%’

- ❖ Result:

Average Payload Mass(Kgs)	Customer	Booster_Version
2534.6666666666665	MDA	F9 v1.1 B1003

- ❖ Explanation: The average payload mass transported by the booster version Falcon 9 v1.1 is 2,928 kg.

First Successful Ground Landing Date

❖ Query: %sql SELECT MIN(DATE) as "First Successful Landing" FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)’

❖ Result:

First Successful Landing
2015-12-22

❖ Explanation: The inaugural successful landing on a ground pad took place on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query: %sql SELECT Booster_Version, Payload FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000

- Result:

Booster_Version	Payload
F9 FT B1022	JCSAT-14
F9 FT B1026	JCSAT-16
F9 FT B1021.2	SES-10
F9 FT B1031.2	SES-11 / EchoStar 105

- Explanation: Four booster versions have successfully landed on a drone ship while delivering payloads exceeding 4,000 kg but remaining under 6,000 kg.

Total Number of Successful and Failure Mission Outcomes

- Query: %sql SELECT Mission_Outcome, COUNT("Mission_Outcome") as Total FROM SPACEXTBL GROUP BY "Mission_Outcome“
- Result:

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

- Explanation: A total of 101 missions were conducted, with 61 successes and 40 failures recorded.

Boosters Carried Maximum Payload

- Query: %sql SELECT Booster_Version,Payload, PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
- Result:

Booster_Version	Payload	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Starlink 1 v1.0, SpaceX CRS-19	15600
F9 B5 B1049.4	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600
F9 B5 B1051.3	Starlink 3 v1.0, Starlink 4 v1.0	15600
F9 B5 B1056.4	Starlink 4 v1.0, SpaceX CRS-20	15600
F9 B5 B1048.5	Starlink 5 v1.0, Starlink 6 v1.0	15600
F9 B5 B1051.4	Starlink 6 v1.0, Crew Dragon Demo-2	15600
F9 B5 B1049.5	Starlink 7 v1.0, Starlink 8 v1.0	15600
F9 B5 B1060.2	Starlink 11 v1.0, Starlink 12 v1.0	15600
F9 B5 B1058.3	Starlink 12 v1.0, Starlink 13 v1.0	15600
F9 B5 B1051.6	Starlink 13 v1.0, Starlink 14 v1.0	15600
F9 B5 B1060.3	Starlink 14 v1.0, GPS III-04	15600
F9 B5 B1049.7	Starlink 15 v1.0, SpaceX CRS-21	15600

- Explanation:
  The maximum payload mass recorded in this dataset is 15,600 kg, which was delivered by twelve individual Falcon 9 boosters.

2015 Launch Records

❖ Query: %sql SELECT substr(Date, 6, 2) as MONTH, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(Date,1, 4) = '2015' AND Landing_Outcome = 'Failure (drone ship)’

❖ Result:

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

❖ Explanation: In 2015, two unsuccessful drone ship landings were recorded, both originating from CCAFS LC-40—one in January, and the other in April.

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

- Query: %sql SELECT Landing_Outcome, COUNT(*) as count_outcomes FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY count_outcomes DESC
- Result:
- Explanation:

  The most frequent landing outcome was "no attempt made."

Landing_Outcome	count_outcomes
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

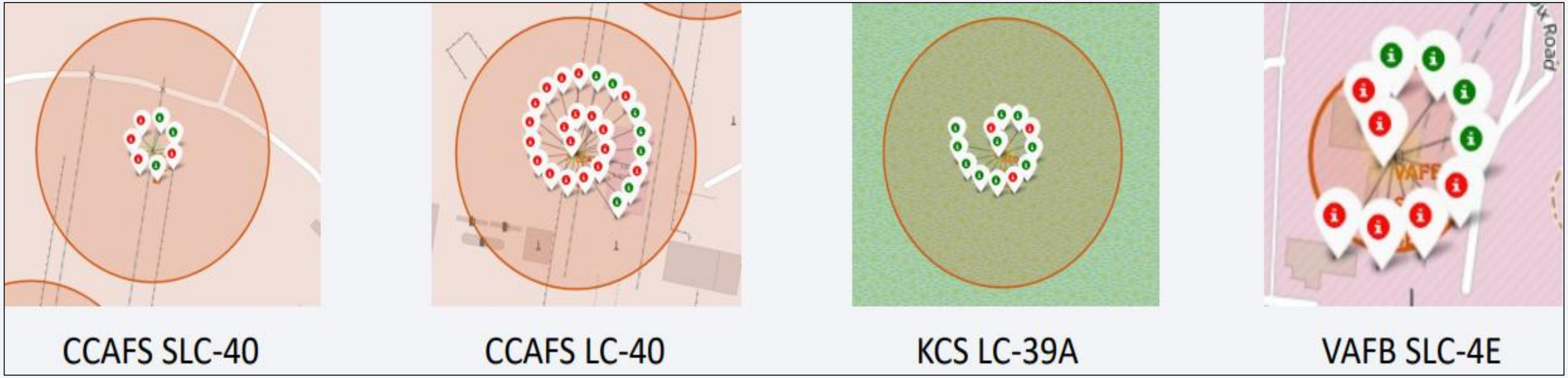
Section 3

Launch Sites Proximities Analysis

Falcon 9 Launch Pads and Facilities

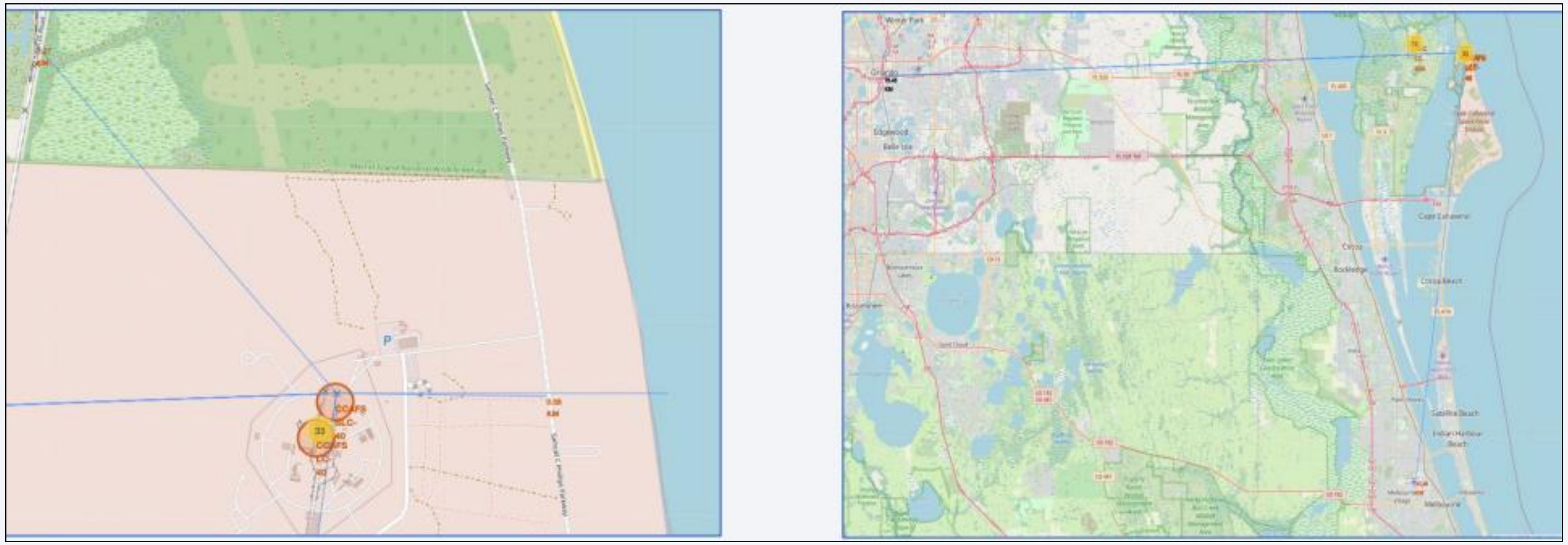
- VAFB SLC-4E (California, USA): Vandenberg Air Force Base, Space Launch Complex 4E.
- KSC LC-39A (Florida, USA): Kennedy Space Center, Launch Complex 39A.
- CCAFS LC-40 (Florida, USA): Cape Canaveral Air Force Station, Launch Complex 40.
- CCAFS SLC-40 (Florida, USA): Cape Canaveral Air Force Station, Space Launch Complex 40.

Success and Failure Indicators for Falcon 9 Landings on Map



- ❖ Map markers indicate the outcomes—successful or failed—of Falcon 9 first stage landings, linked to the geographical coordinates of each launch site.
- ❖ The distribution of green (success) and red (failure) markers provides insight into the relative landing success rate for each location.

Proximity Measurements Around Rocket Launch Facilities



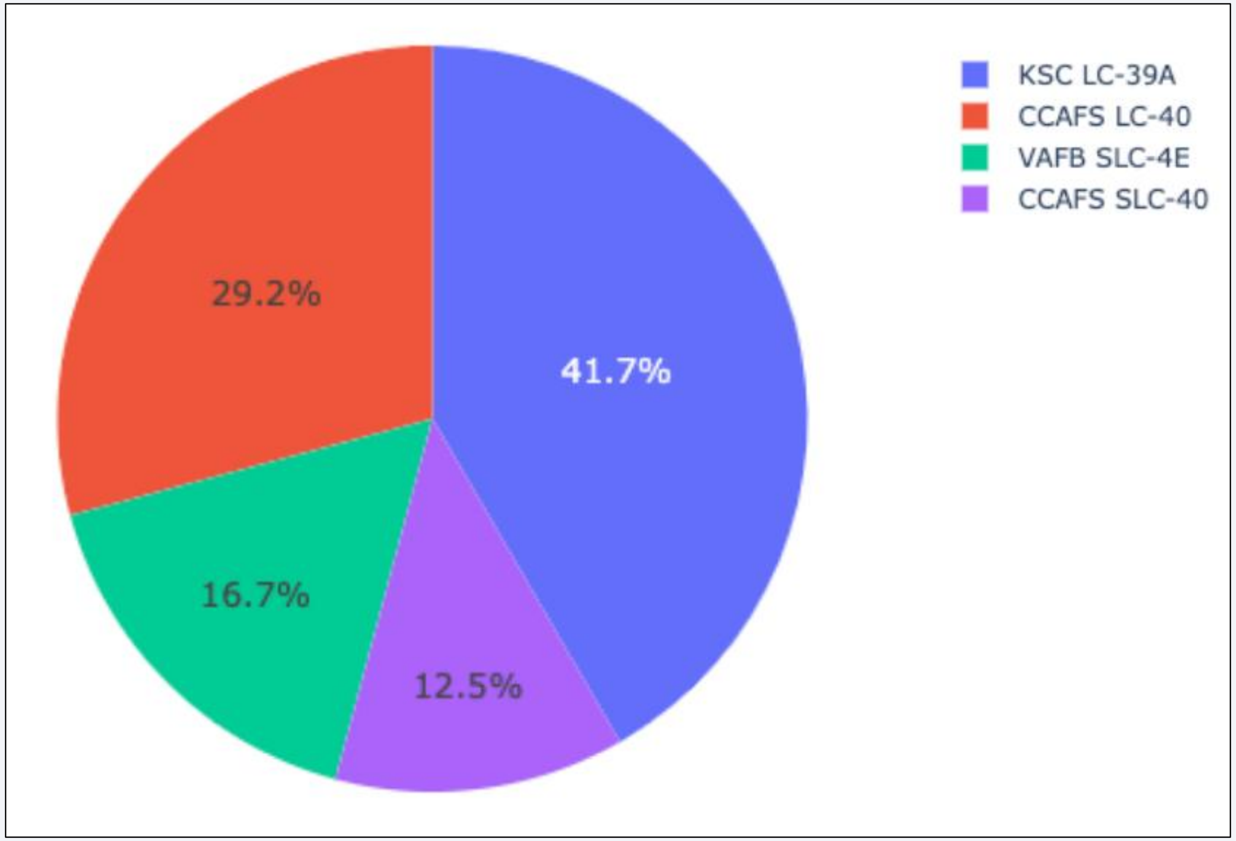
- Although CCAFS LC-40 and CCAFS SLC-40 have nearly identical coordinates, they are not precisely co-located.
- The perimeter road lies approximately 0.19 km from LC-40's coordinates.
- The nearest coastline is around 0.92 km away.
- A rail line is situated roughly 1.33 km from the launch site.

Section 4

Build a Dashboard with Plotly Dash

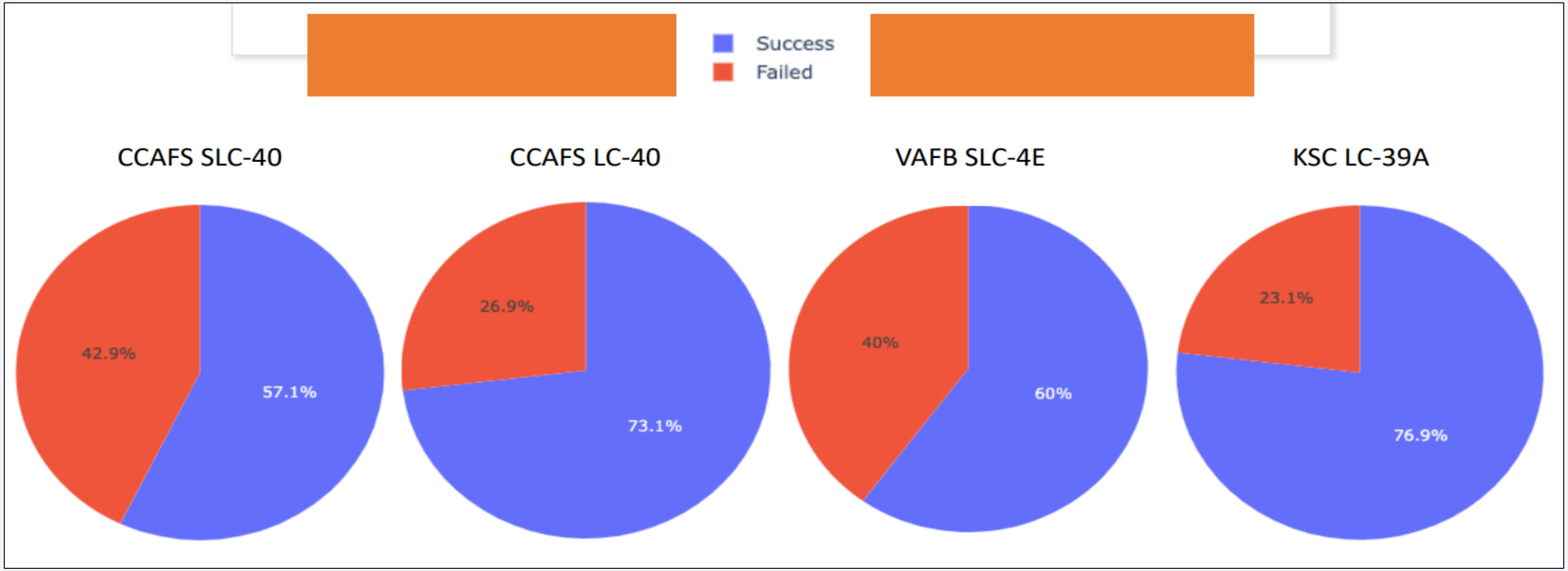
Falcon 9 Launch Success for All Sites

- ❖ Kennedy Space Center Launch Complex 39A (KSC LC-39A) recorded the highest number of successful Falcon 9 first stage landings, representing 41.7% of the total successes.
- ❖ When data from all launch sites is considered, the pie chart visualizes how these successful landings are distributed across the various locations.

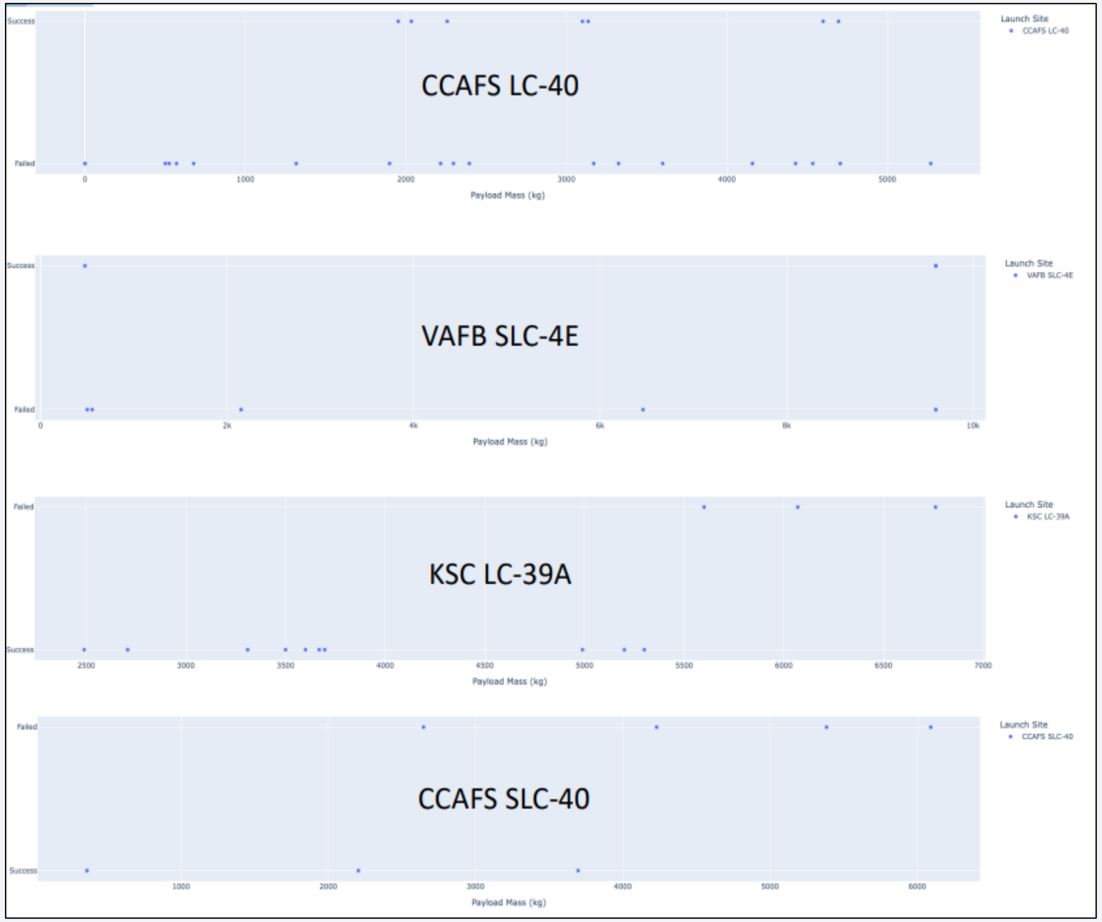


Falcon 9 Launch Success for All Sites (Individually)

❖ CCAFS SLC-40 achieved the highest success rate for Falcon 9 first stage landings, accounting for 42.9% of all successful landings.



Payload vs. Launch Outcome

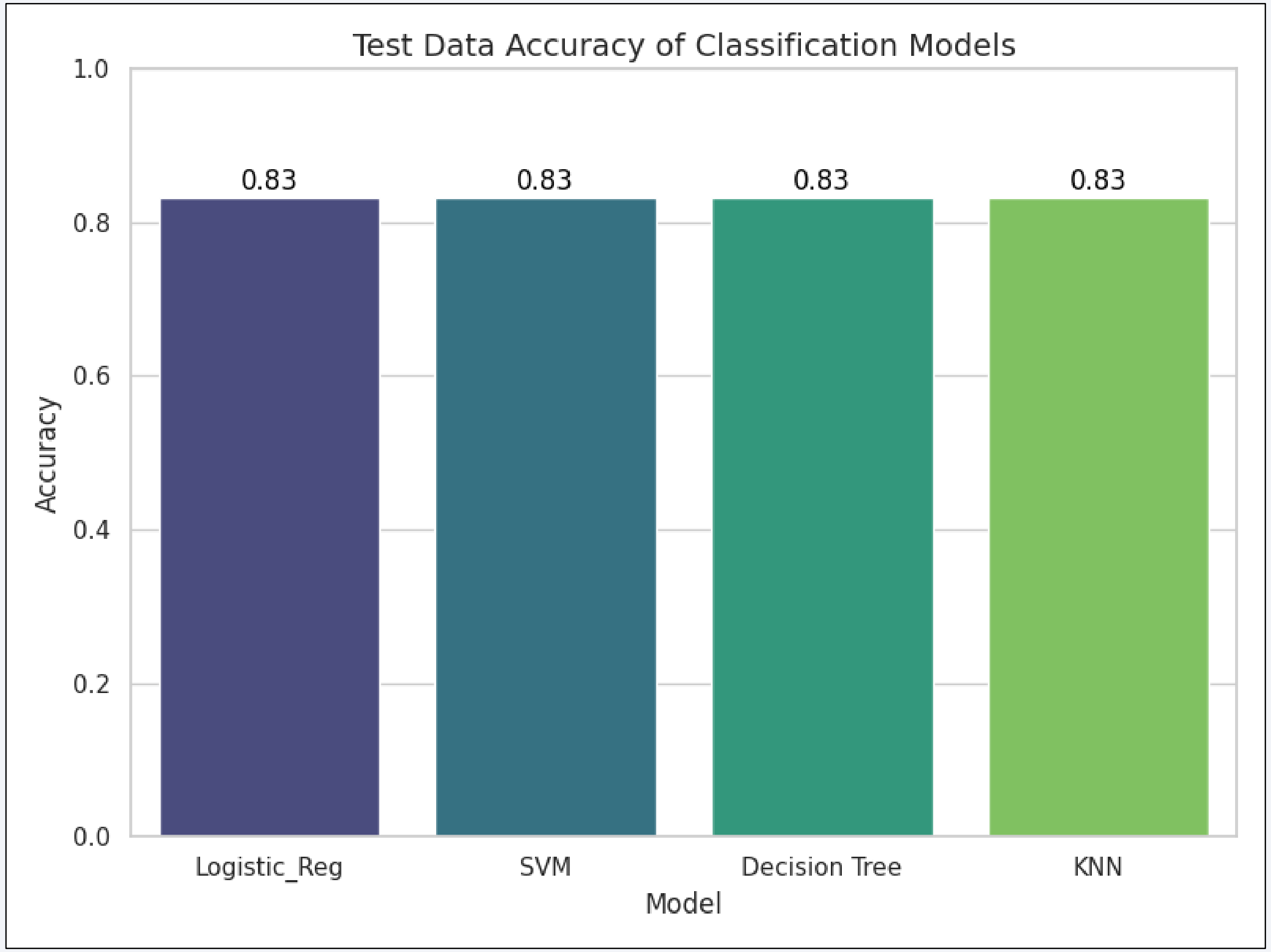


- ❖ These scatter plots illustrate the relationship between payload mass and launch outcomes across all launch sites, with payload values filtered using a range slider.
- ❖ Launches carrying payloads between approximately 2,000 kg and 5,000 kg demonstrate the highest success rate.
- ❖ Among booster types, the ‘FT’ version stands out as the most successful, contributing to the largest proportion of successful launches.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



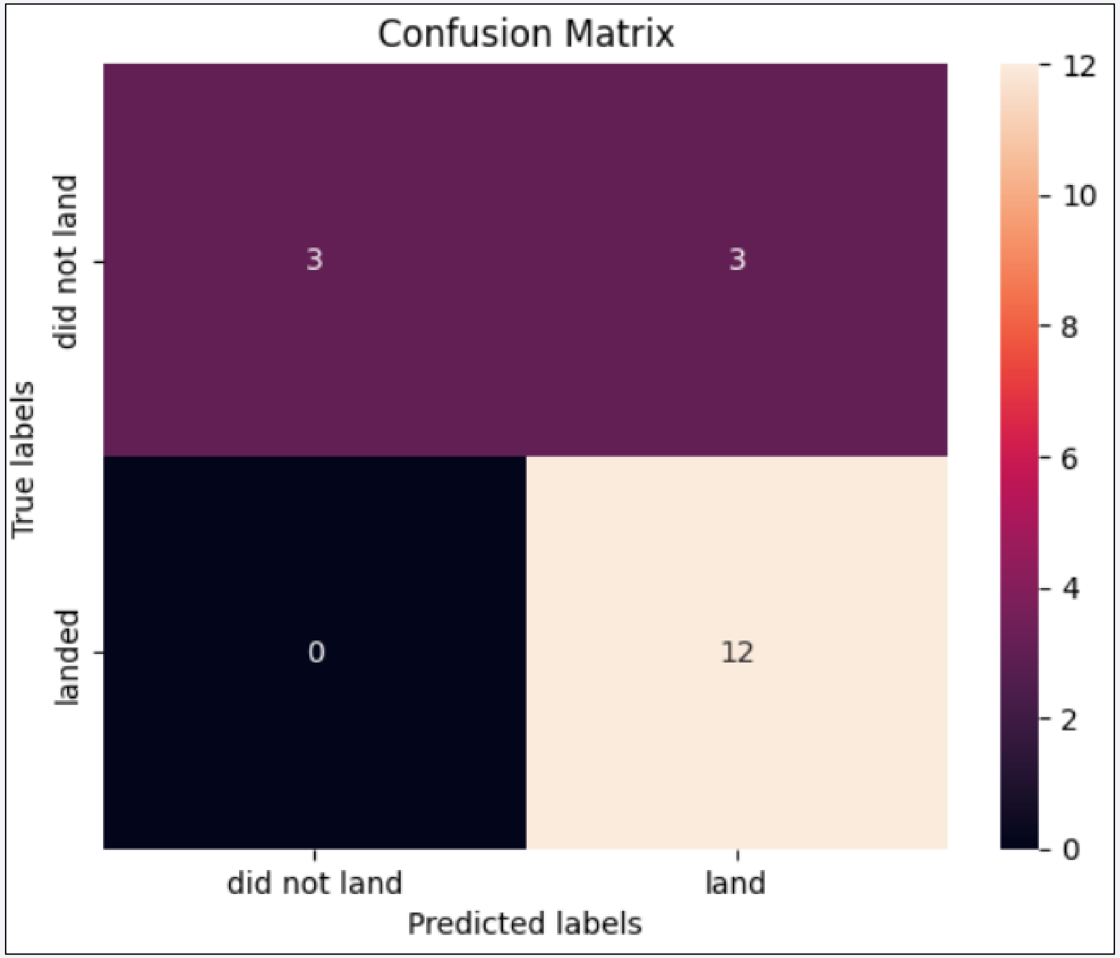
- ❖ All four models—Logistic Regression, SVM, Decision Tree, and KNN—achieved an identical test accuracy of 0.833333.
- ❖ Since the accuracy values are the same, no single model outperforms the others based on this metric alone.

Confusion Matrix

❖ The confusion matrix for the Logistic Regression model is plotted.

❖ Prediction Summary:

- 12 instances correctly classified as positives (True Positives).
- 3 instances correctly classified as negatives (True Negatives).
- 3 instances incorrectly classified as positives (False Positives).
- No instances misclassified as negatives (False Negatives: 0).



Conclusions

- ❖ Imperfection in Launch Outcomes: SpaceX's Falcon 9 first stage landing attempts have not consistently resulted in success, highlighting the complexity and challenges of reusable rocket technology.
- ❖ Positive Trend in Performance: Despite early setbacks, there is a clear upward trend in successful landings over time, indicating improvements in design, technology, and operational expertise.
- ❖ Predictive Modeling Potential: With the accumulation of launch data, machine learning models can be trained to forecast future landing outcomes based on variables such as payload weight, launch site, booster version, and more.

Appendix

❖ Initial Data Sources for Analysis:

- *SpaceX API (JSON format)*
  [Access API Data](#).
- *Wikipedia Launch Records (Webpage)*
  Falcon 9 and Falcon Heavy Launches.
- *SpaceX Launch Dataset (CSV)*
  CSV Dataset.
- *Geospatial Launch Data (CSV)*
  Launch Geo CSV.
- *Dashboard Data for Launches (CSV)*
  Launch Dash CSV.

❖ GitHub Repository Links (CSV Files and Web Scraping):

- *Dataset Part 1 (CSV) .*
- *Web Scraped SpaceX Data .*
- *Dataset Part 2 (CSV) .*
- *SpaceX Launches (CSV) .*
- *Dataset Part 3 (CSV) .*
- *Launch Geo (CSV) .*
- *Launch Dashboard Data (CSV).*

Thank you!