IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daniel Omar Silva Navarro
December 22nd, 2023

Outline

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

Executive Summary

- This project leveraged diverse methodologies for comprehensive data analysis:
  - We gathered data through web scraping and the SpaceX API for a robust dataset.
  - Conducted Exploratory Data Analysis (EDA) encompassing data wrangling, visualization, and interactive visual analytics.
  - Employed Machine Learning Prediction techniques.

- Overview of key findings:
  - Valuable data was successfully collected from various public sources.
  - EDA highlighted the crucial features for predicting the success of launches.
  - Through Machine Learning Prediction, we identified the optimal model for predicting the key characteristics that drive success, utilizing the entire dataset we compiled.

Introduction

- The mission centers around assessing the feasibility of the emerging company Space Y in its competition with Space X.

- Key Objectives:
  - Optimal estimation of total launch costs by predicting the success of first-stage rocket landings.
  - Identification of the optimal launch locations to enhance operational efficiency and success.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from 2 sources:
    - Space X API (https://api.spacexdata.com/v4/rockets/)
    - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Enriched collected data by creating a landing outcome label based on outcome data, summarizing and analyzing relevant features.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Utilized SQL for structured analysis and visualization tools to uncover patterns, trends, and insights within the dataset.

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
  - Employed Folium and Plotly Dash to create interactive visualizations, providing a dynamic and user-friendly exploration of the data.
- Perform predictive analysis using classification models
  - Normalized data and split it into training and test datasets for evaluation using four different classification models.

Data Collection

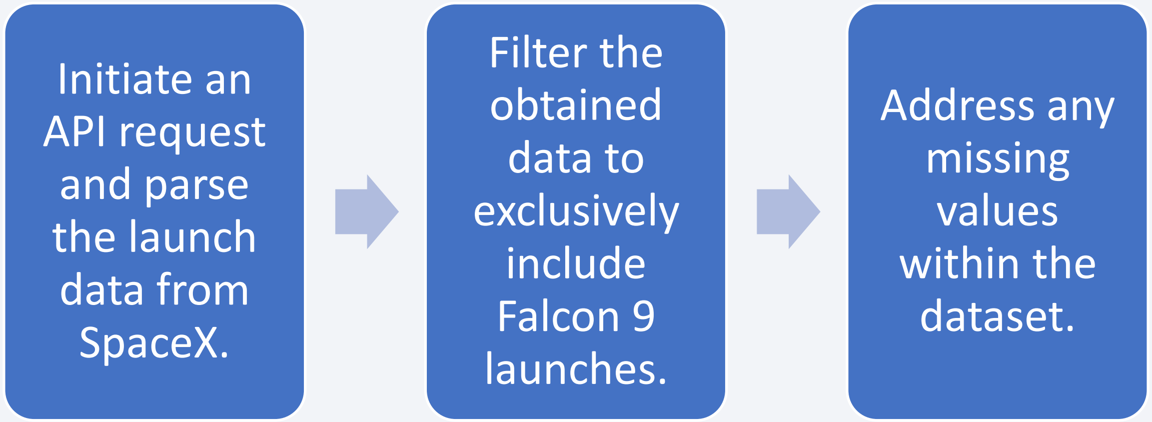
Datasets were gathered from:

- SpaceX API (https://api.spacexdata.com/v4/rockets/)
- Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches) through the application of web scraping techniques.

Data Collection – SpaceX API

- Utilizing a public API provided by SpaceX, data can be acquired and subsequently utilized.

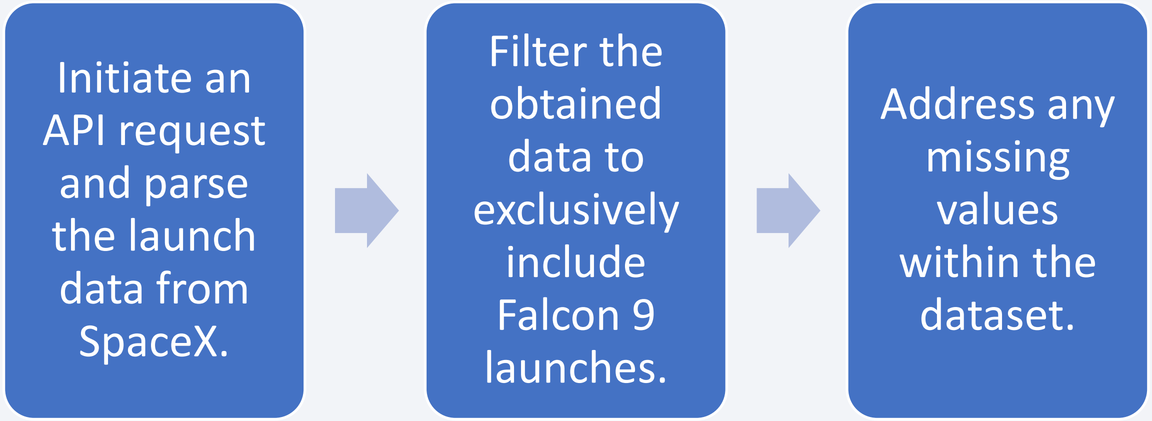
- The API, as depicted in the accompanying flowchart, was employed to retrieve data, which was then stored persistently.



Source code (GitHub):
https://github.com/daniboii201/IBMDataScience/blob/0224f761be8d65019c522982a3ffff20abd36585/01.%20Applied%20Data%20Science%20Capstone/(10.11)%20Data%20Collection%20API.ipynb

Data Collection - Scraping

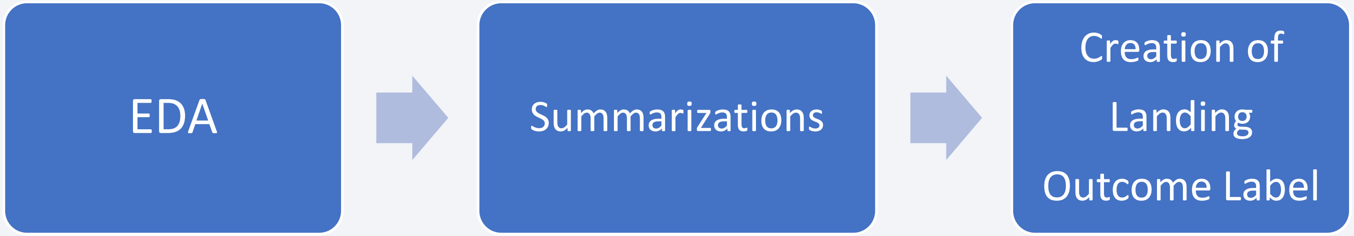
- Information on SpaceX launches is also retrievable from Wikipedia.
- Following the flowchart, data is downloaded from Wikipedia and subsequently stored persistently.



Source code (GitHub):
https://github.com/daniboii201/IBMDataScience/blob/0224f761be8d65019c522982a3ffff20abd36585/01.%20Applied%20Data%20Science%20Capstone/(10.12)%20Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

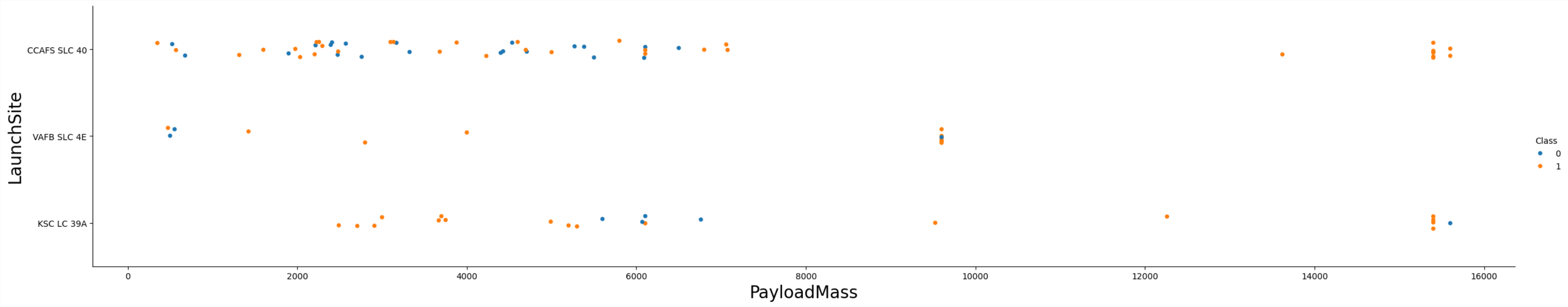
- Initially, the dataset underwent Exploratory Data Analysis (EDA).
- Subsequently, calculations were conducted to summarize launches per site, occurrences of each orbit, and occurrences of mission outcomes per orbit type.
- Finally, a landing outcome label was generated based on the information in the Outcome column.



Source code (GitHub):
https://github.com/daniboii201/IBMDataScience/blob/0224f761be8d65019c522982a3ffff20abd36585/01.%20Applied%20Data%20Science%20Capstone/(10.13)%20Data%20Wrangling.ipynb

EDA with Data Visualization

- To investigate the data, scatterplots and barplots were employed to depict the connection between pairs of features.
  - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit



Source code (GitHub):
https://github.com/daniboii201/IBMDataScience/blob/0224f761be8d65019c522982a3ffff20abd36585/01.%20Applied%20Data%20Science%20Capstone/(10.22)%20EDA%20with%20Visualization.ipynb

EDA with SQL

The following SQL queries were executed:

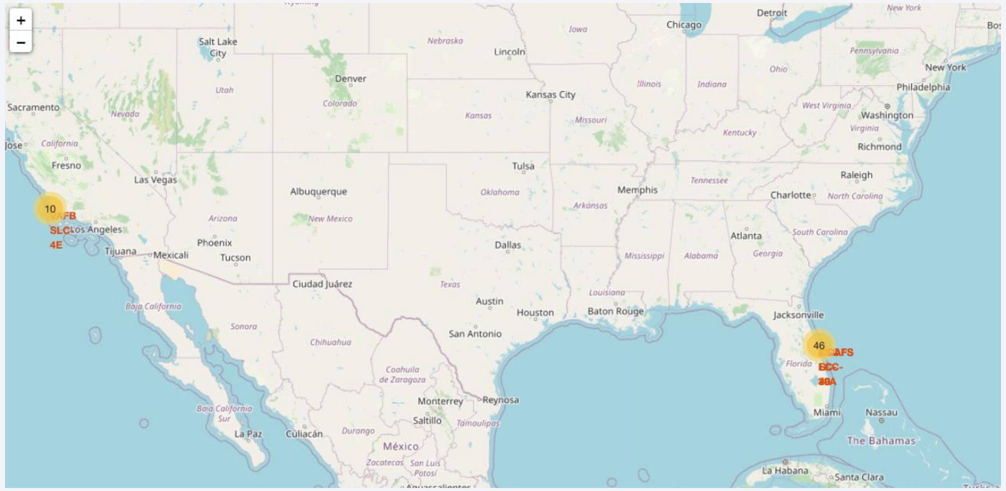
- Retrieve the names of unique launch sites in space missions.
- Identify the top 5 launch sites with names starting with 'CCA.'
- Calculate the total payload mass carried by boosters launched by NASA (CRS).
- Determine the average payload mass carried by the F9 v1.1 booster version.
- Find the date of the first successful landing outcome on a ground pad.
- Obtain the names of boosters with success in drone ship landings and payload masses between 4000 and 6000 kg.
- Count the total number of successful and failed mission outcomes.
- Identify the booster versions that carried the maximum payload mass.
- Retrieve information on failed landing outcomes on drone ships, including their booster versions and launch site names, specifically for the year 2015.
- Rank the count of landing outcomes (e.g., Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20

Source code (GitHub):
https://github.com/daniboii201/IBMDataScience/blob/0224f761be8d65019c522982a3ffff20abd36585/01.%20Applied%20Data%20Science%20Capstone/(10.21)%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Various elements such as markers, circles, lines, and marker clusters were employed on Folium Maps:

- Markers signify specific points, such as launch sites.
- Circles highlight areas around specific coordinates, exemplified by the NASA Johnson Space Center.
- Marker clusters represent groups of events at each coordinate, exemplified by launches at a particular launch site.
- Lines are utilized to denote distances between two coordinates.



Source code (GitHub):
https://github.com/daniboii201/IBMDataScience/blob/0224f761be8d65019c522982a3ffff20abd36585/01.%20Applied%20Data%20Science%20Capstone/(10.31)%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- The data was visualized using the following graphs and plots:
  - Launch percentages categorized by site.
  - Payload range distribution.
- This combination facilitated a swift analysis of the relationship between payloads and launch sites, aiding in the identification of the optimal launch locations based on payload considerations.
- Source code
  - https://github.com/daniboii201/IBMDataScience/blob/9452917ac7b27cb6d610d73ee74144a935706c97/01.%20Applied%20Data%20Science%20Capstone/(10.32)%20Interactive%20Dashboard%20with%20Ploty%20Dash.py

Predictive Analysis (Classification)

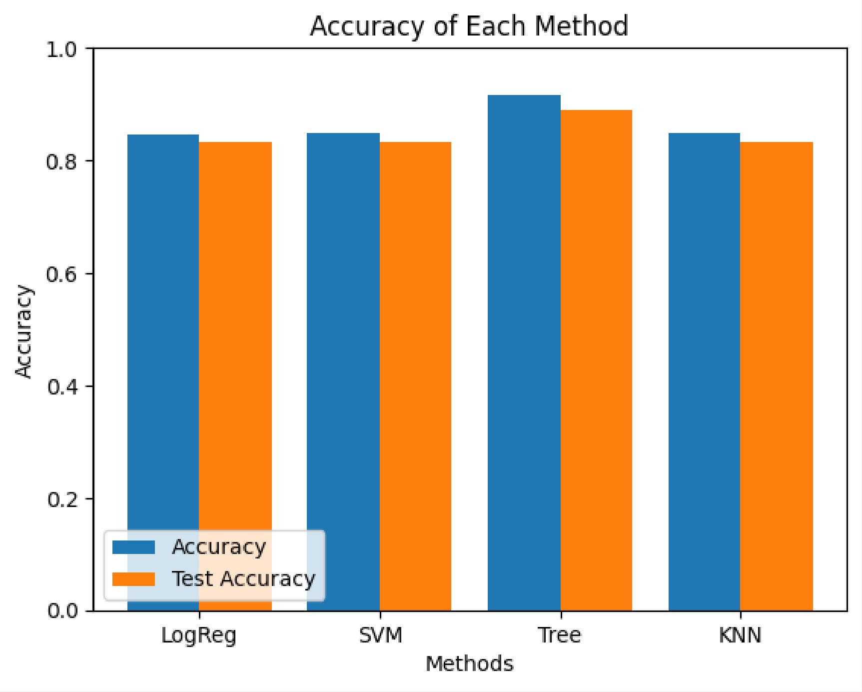
- We compared four classification models: logistic regression, support vector machine, decision tree, and k-nearest neighbors.

Data preparation and standardization → Test of each model with combinations of hyperparameters → Comparison of results

Source code (GitHub):
https://github.com/daniboii201/IBMDataScience/blob/9452917ac7b27cb6d610d73ee74144a935706c97/01.%20Applied%20Data%20Science%20Capstone/(10.41)%20Machine%20Learning%20Prediction.ipynb

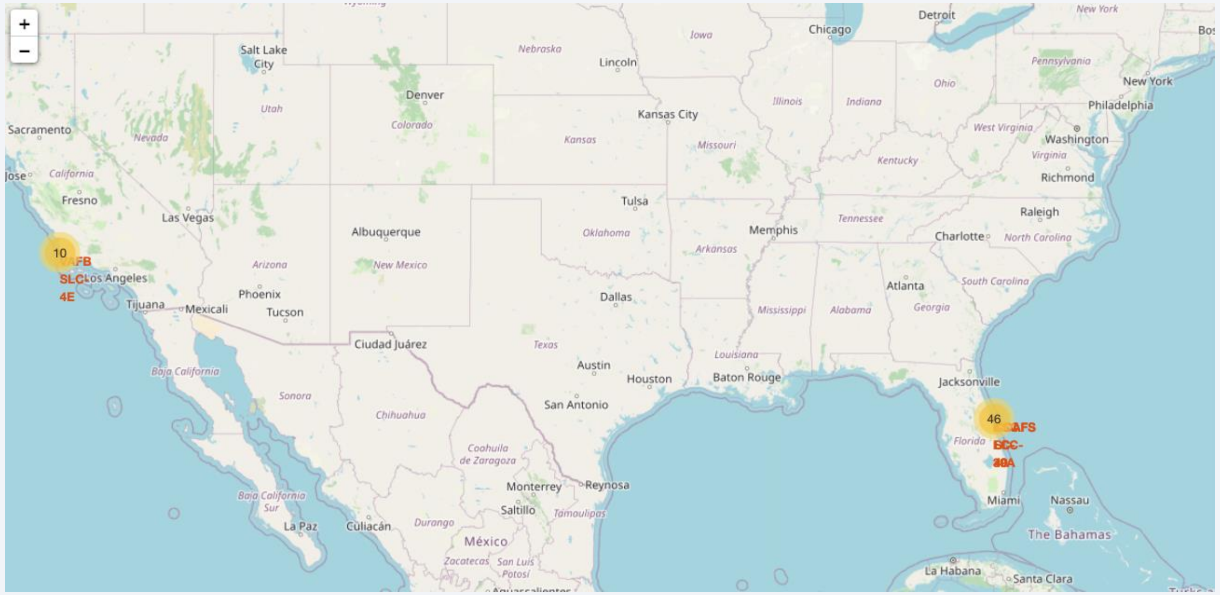
Results

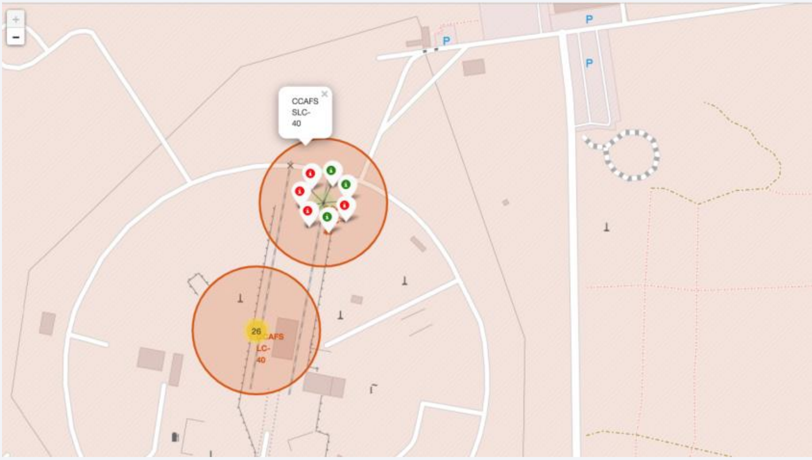
- Exploratory data analysis results: Space X uses 4 different launch sites;
  - The first launches were done to Space X itself and NASA; The average payload of F9 v1.1 booster is 2,928 kg;
  - The first success landing outcome happened in 2015 fiver year after the first launch; Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
  - Almost 100% of mission outcomes were successful;
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The number of landing outcomes became as better as years passed.



Results

- Employing interactive analytics enabled the identification of launch sites typically located in secure areas, often near the sea, with well-developed logistic infrastructure.
- The majority of launches occur at launch sites on the east coast.

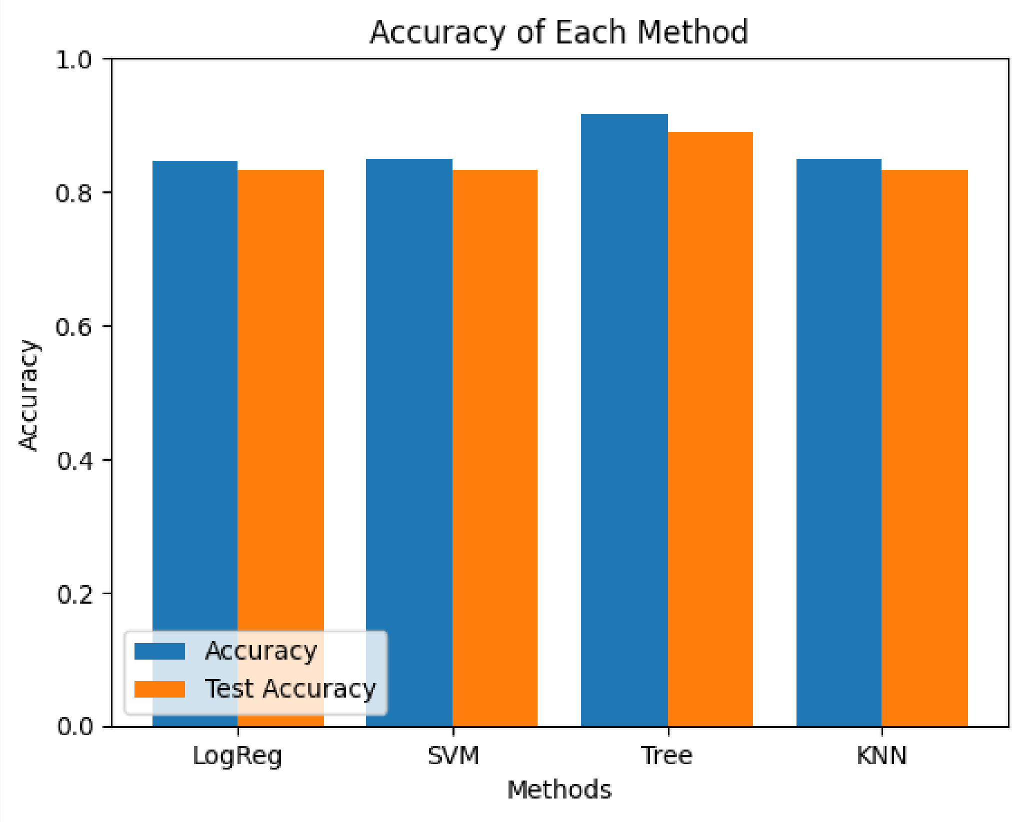




Results

- The Predictive Analysis revealed that the Decision Tree Classifier stands out as the optimal model for forecasting successful landings, attaining an accuracy surpassing 87% and demonstrating a test data accuracy exceeding 94%.

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.91607	0.88889
KNN	0.84821	0.83333

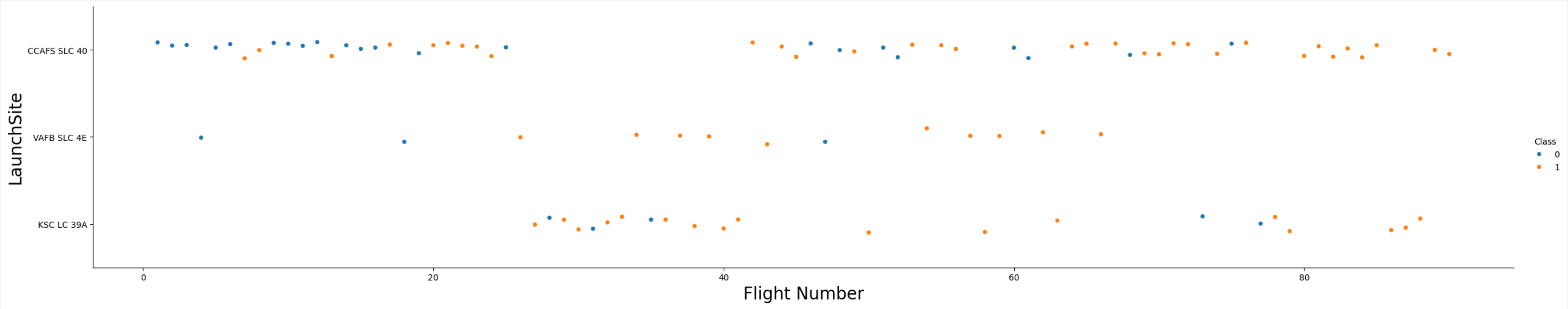


Section 2

Insights drawn from EDA

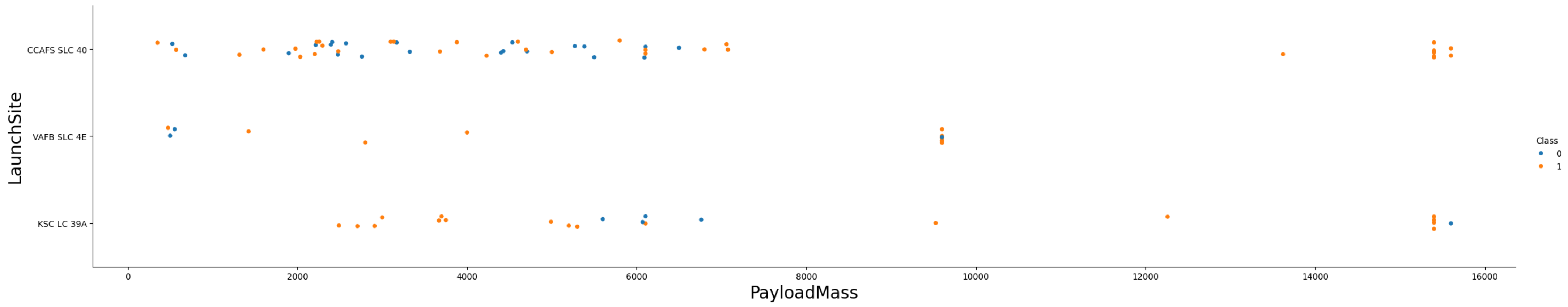
Flight Number vs. Launch Site

- The graph indicates that presently, CCAF5 SLC 40 is the most effective launch site, hosting the majority of recent successful launches.
- Following closely in second place is VAFB SLC 4E, and in third place is KSC LC 39A.
- Moreover, there is a noticeable improvement in the overall success rate over time.



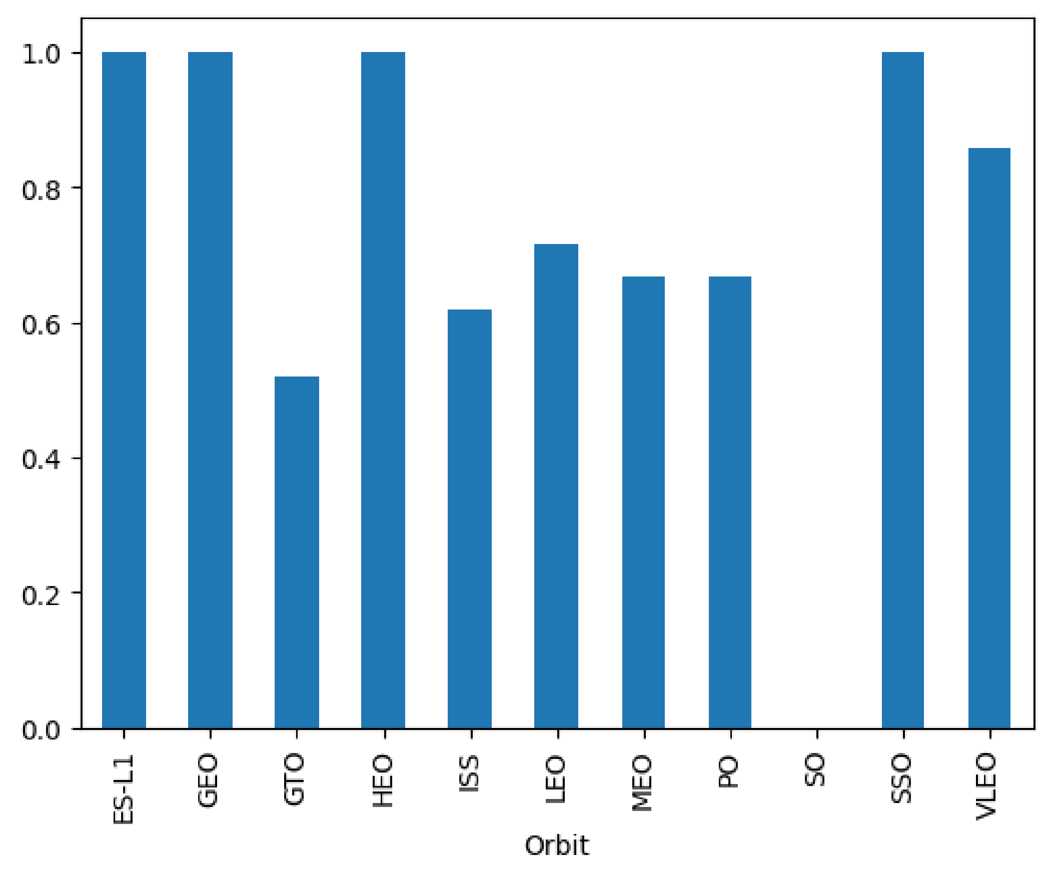
Payload vs. Launch Site

- Payloads exceeding 9,000kg, approximately equivalent to the weight of a school bus, exhibit a high success rate.
- Payloads surpassing 12,000kg appear achievable primarily at the CCAFS SLC 40 and KSC LC 39A launch sites.



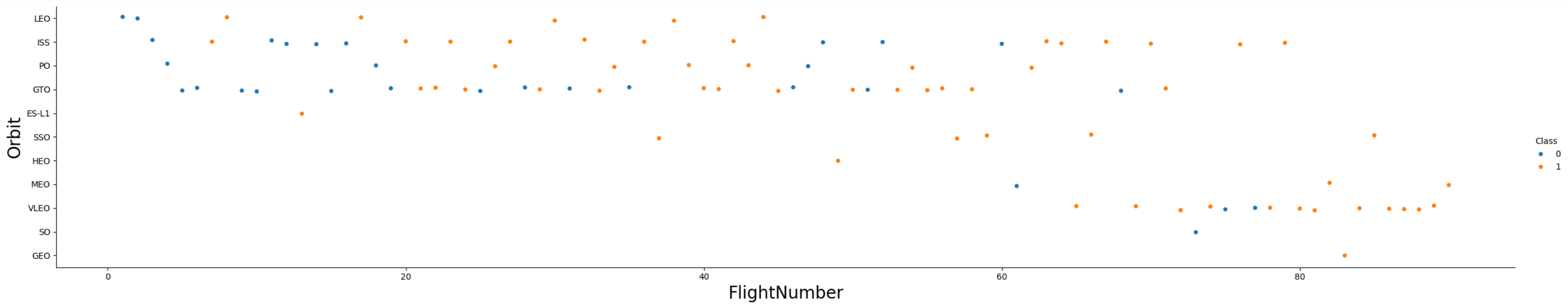
Success Rate vs. Orbit Type

- The biggest success rates happens to orbits:
  - ES-L
  - GEO
  - HEO
  - SSO
- Followed by:
  - VLEO (above 80%)
  - LFO (above 70%)



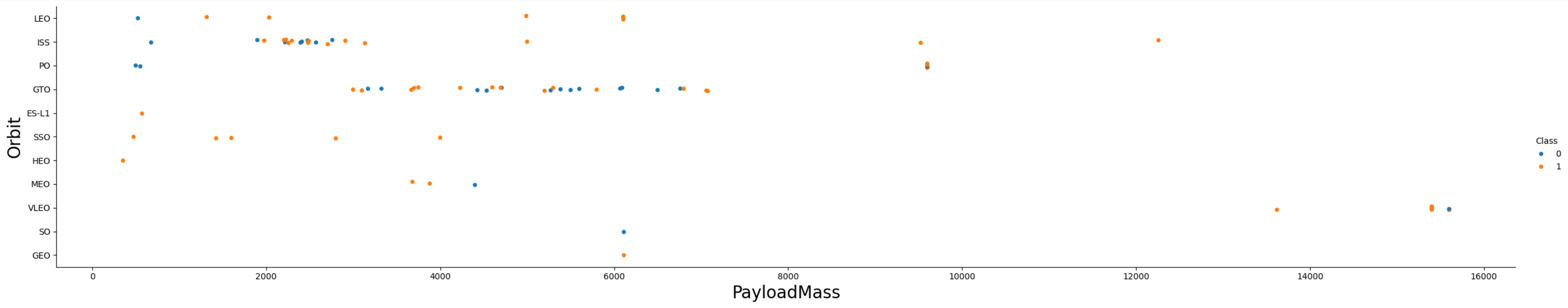
Flight Number vs. Orbit Type

- Evidently, the success rate has shown improvement across all orbits over time.
- The recent uptick in the frequency of VLEO orbit launches suggests a new business opportunity.



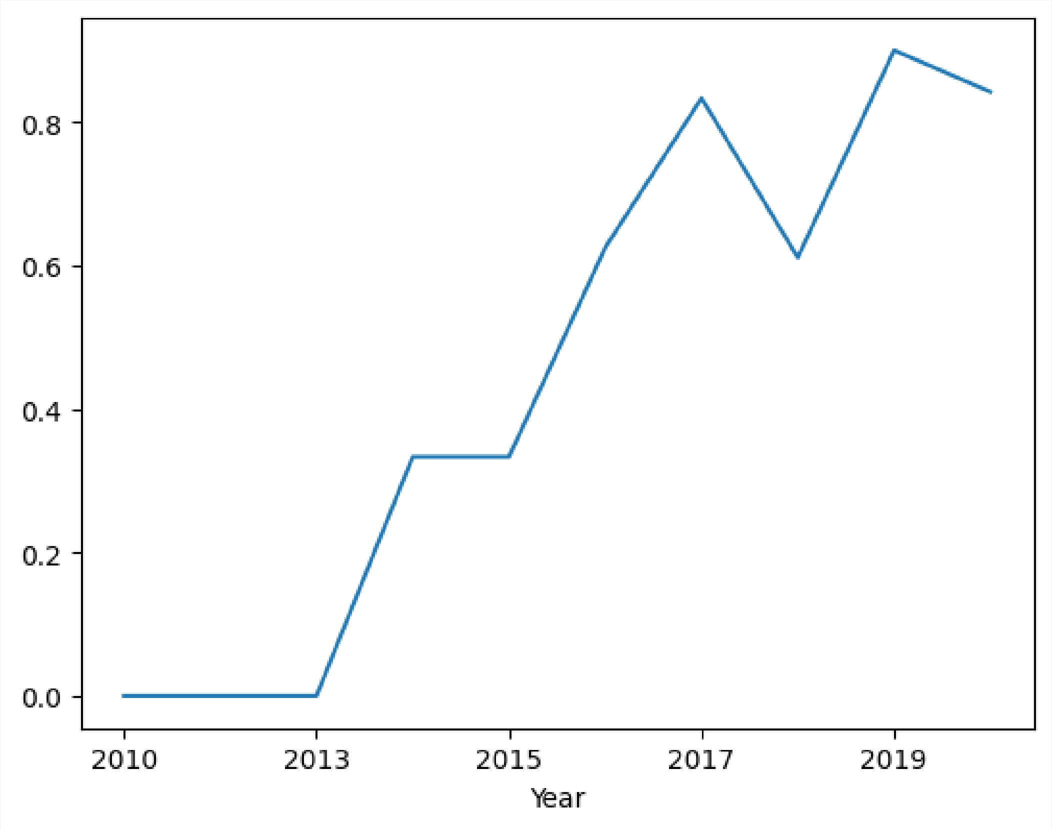
Payload vs. Orbit Type

- It seems there is no correlation between payload and success rate for GTO orbit.
- The ISS orbit exhibits a broad range of payloads and maintains a commendable success rate.
- There have been limited launches to the SO and GEO orbits.



Launch Success Yearly Trend

- The success rate began to rise in 2013 and continued until 2020.
- It appears that the initial three years marked a phase of adjustments and technological improvements.



All Launch Site Names

- The data indicates the presence of four launch sites.

- These sites are identified by selecting distinct instances of "launch_site" values from the dataset.

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745

Launch Site Names Begin with 'CCA'

- There are 5 entries for launch sites starting with CCA:
- Here, we can observe five instances of launches from Cape Canaveral.

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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 attemp

Total Payload Mass

Cumulative payload transported by boosters affiliated with NASA:

- The overall payload is computed by summing the payloads associated with codes containing 'CRS,' indicative of NASA missions.

Total Payload (kg)
111.268

Average Payload Mass by F9 v1.1

The mean payload mass transported by the F9 v1.1 booster version:

- By filtering the data for the mentioned booster version and computing the average payload mass, we arrived at a value of 2.928 kg.

Avg Payload (kg)
2.928

First Successful Ground Landing Date

The initial successful landing outcome on a ground pad:

- By filtering the data for successful landings on ground pads and extracting the earliest date, we can pinpoint the first occurrence, which took place on 12/22/2015.

Min Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters that have achieved successful landings on a drone ship and carried a payload mass exceeding 4000 but less than 6000.
- Applying the specified filters to identify unique booster versions, the outcome comprises these four results.

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

Total Number of Successful and Failure Mission Outcomes

Count of successful and failed mission outcomes:

- By grouping mission outcomes and tallying the records for each category, we derived the summarized information presented above.

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

Boosters Carried Maximum Payload

- Boosters that transported the maximum payload mass.
- These are the boosters identified as having transported the highest payload mass recorded in the dataset.

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

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

2015 Launch Records

- Instances of unsuccessful landings on a drone ship, including their respective booster versions and launch site names, specifically for the year 2015.
- The list provided below comprises the only two instances meeting these criteria.

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

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

Ranking of all landing outcomes from June 4, 2010, to March 20, 2017:

- This data presentation emphasizes the importance of considering instances labeled as "No attempt."

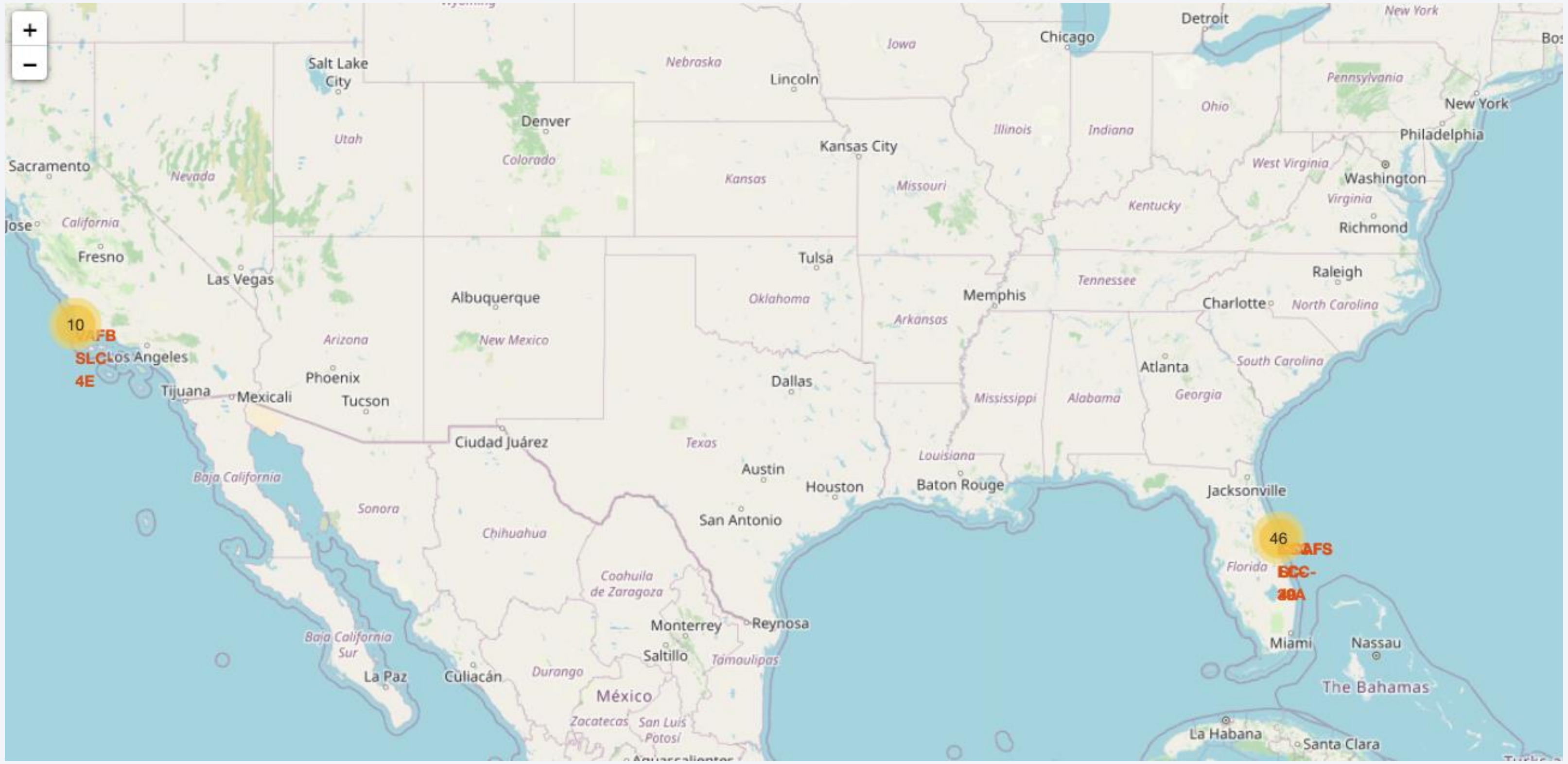
Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

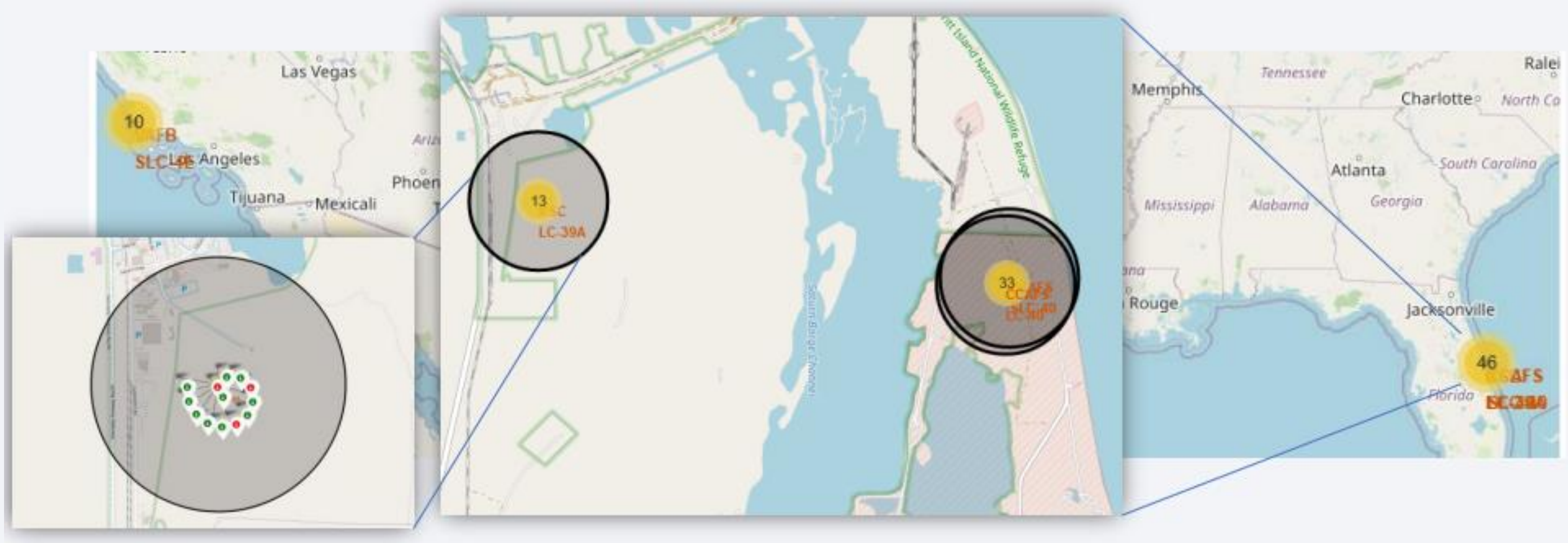
All launch sites

- Launch sites are situated in proximity to the sea, likely for safety reasons, yet remain conveniently accessible to roads and railroads.



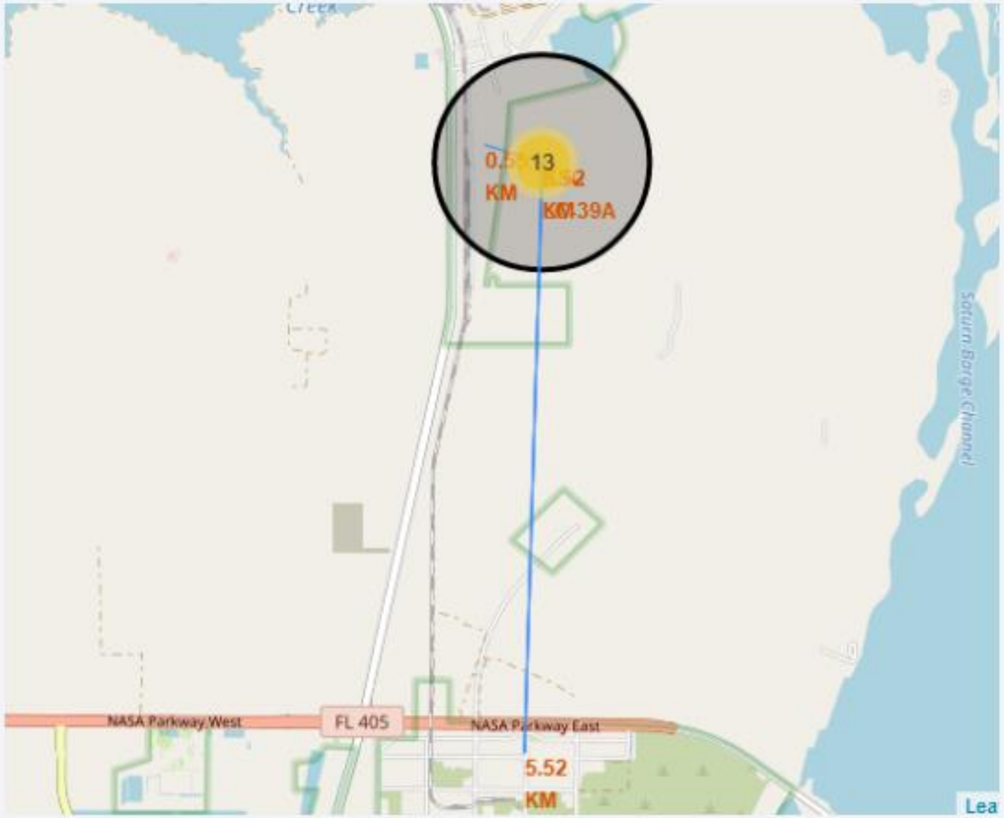
Launch Outcomes by Site

- Instances of launch outcomes for the KSC LC-39A launch site.
- Successful outcomes are represented by green markers, while failures are denoted by red markers.



Logistics and Safety

- The KSC LC-39A launch site possesses favorable logistical characteristics, situated in close proximity to both a railroad and a road, and is relatively distant from populated areas.

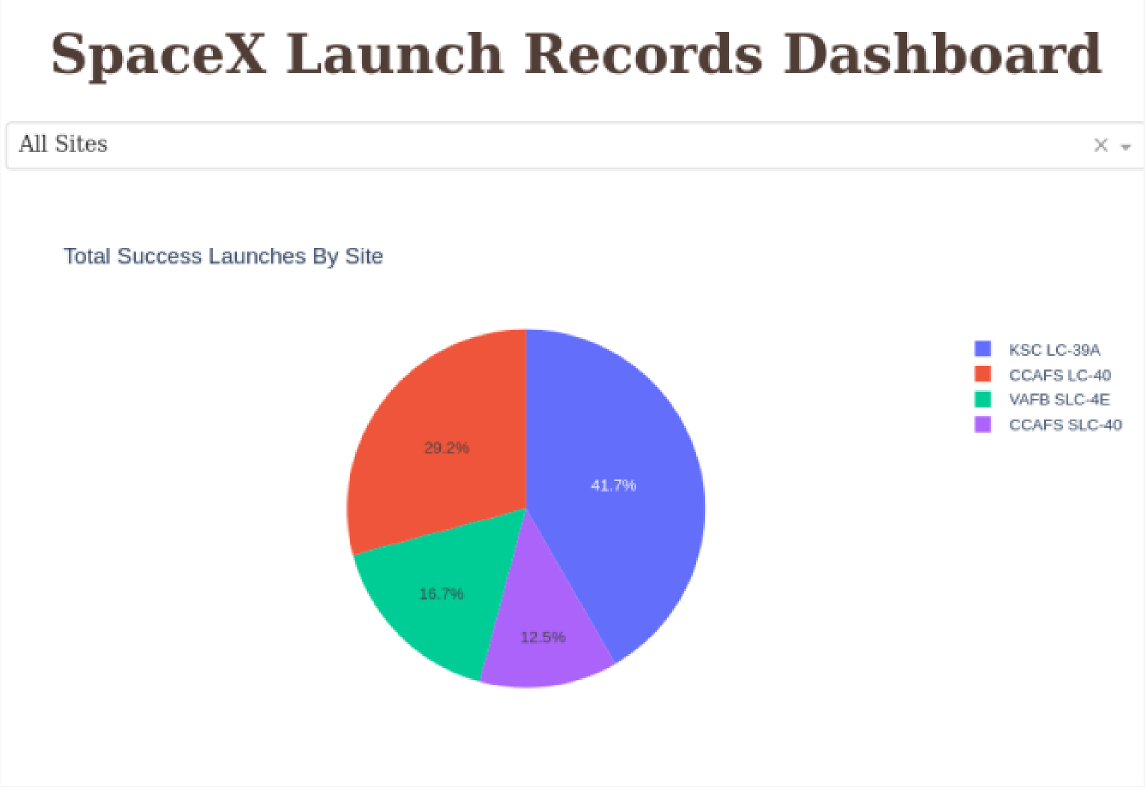


Section 4

Build a Dashboard with Plotly Dash

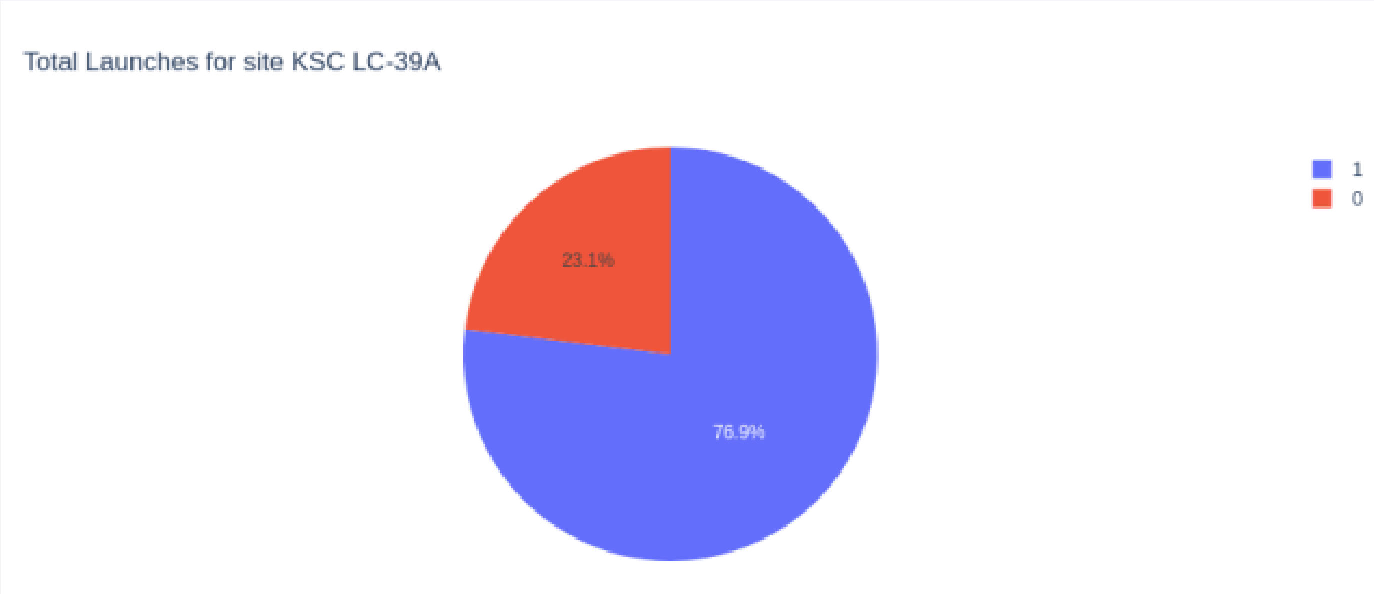
Successful Launches by Site

- The location from which launches are conducted appears to be a crucial determinant of mission success.



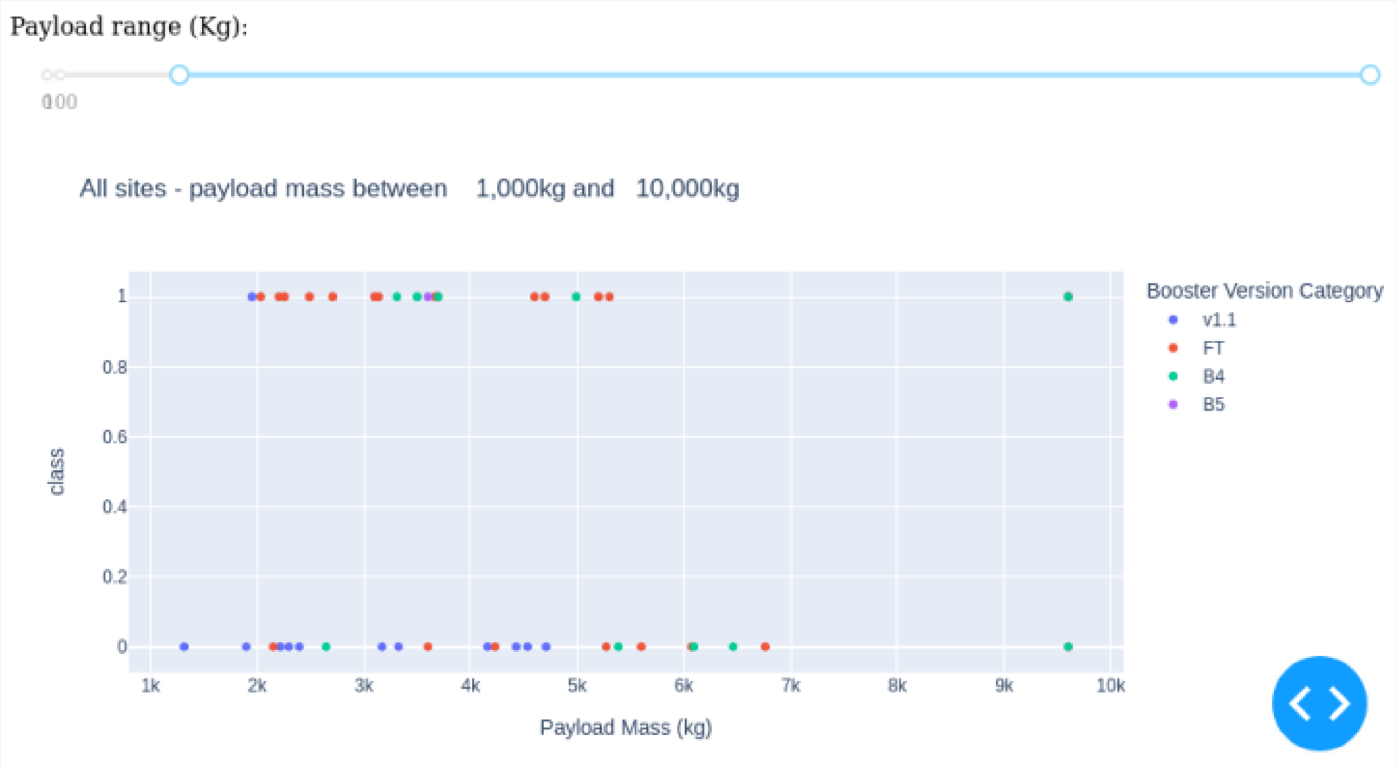
Launch Success Ratio for KSC LC-39A

- The success rate at this site is 76.9%.



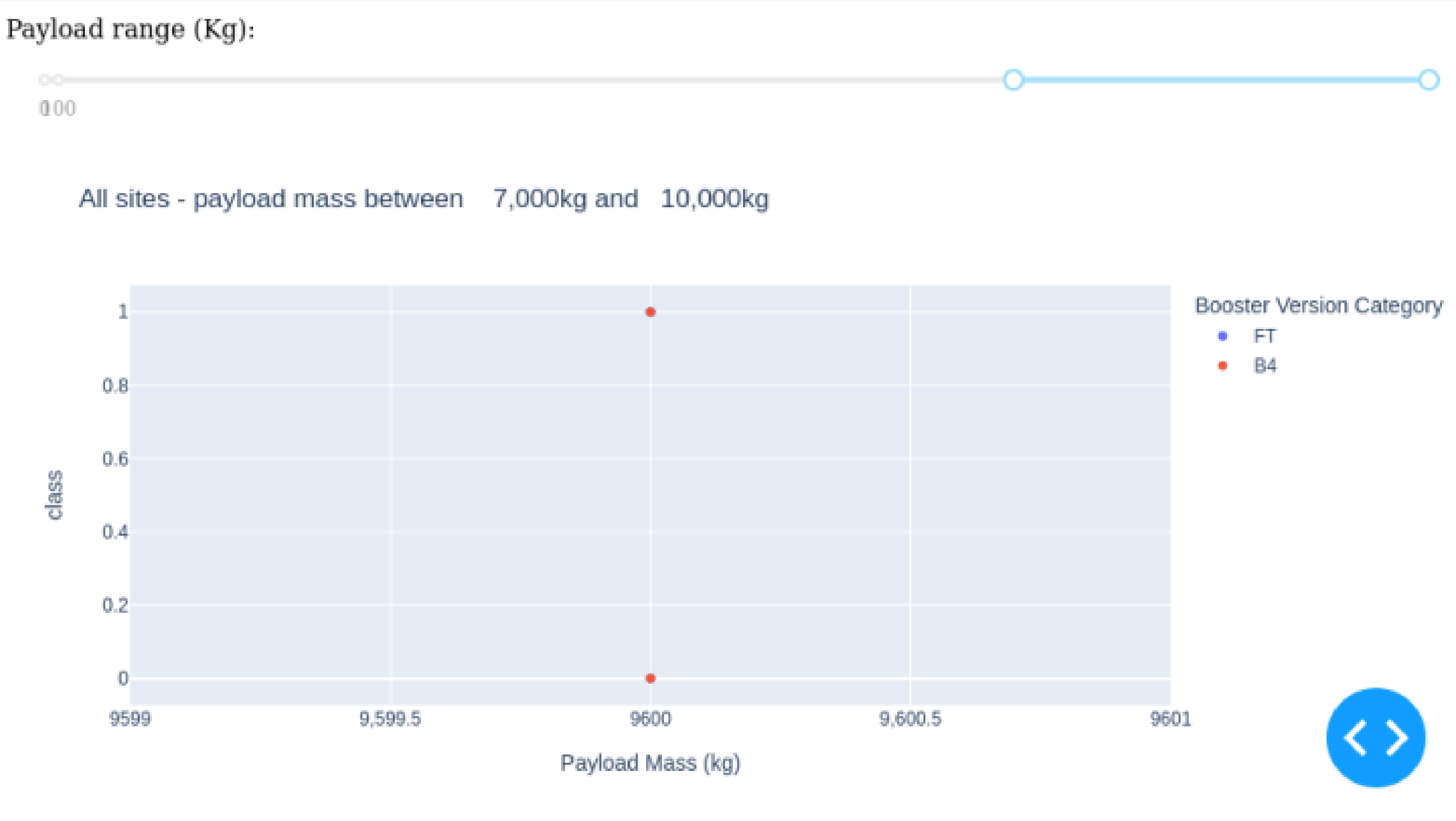
Payload vs. Launch Outcome

- The combination of payloads under 6,000kg and FT boosters yields the highest success rate.



Payload vs. Launch Outcome (continuation)

- Insufficient data is available to assess the risk of launches exceeding 7,000kg.

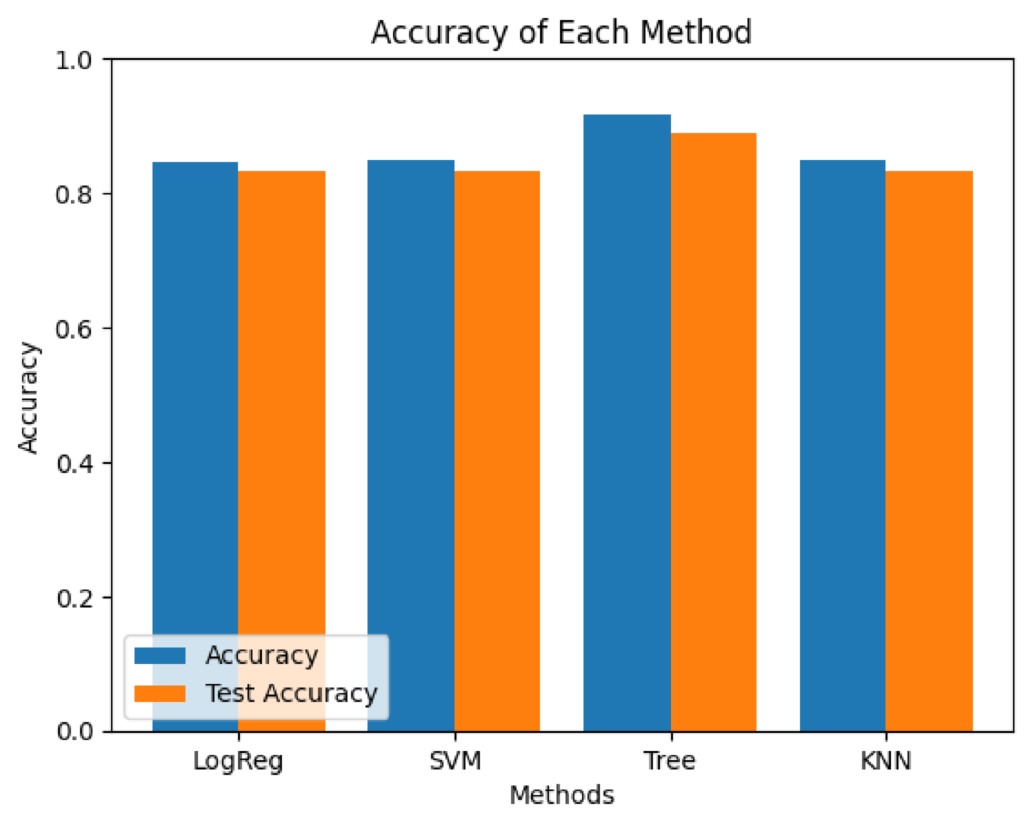


Section 5

Predictive Analysis (Classification)

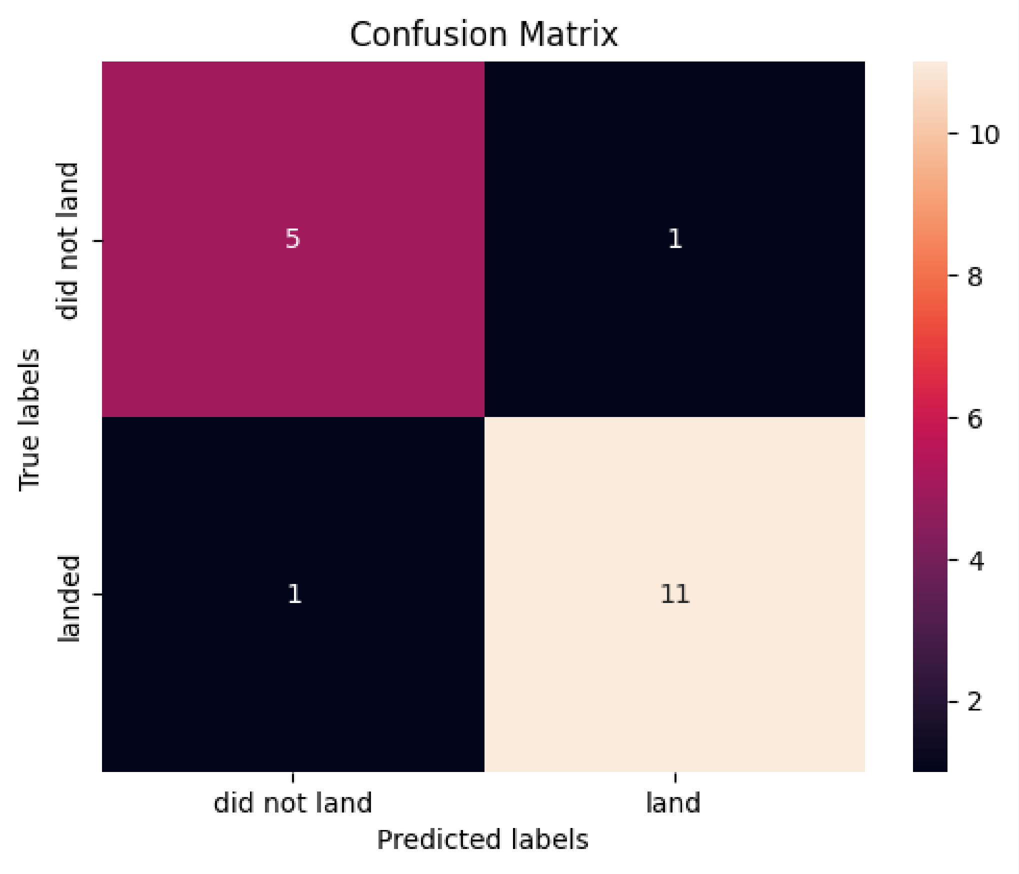
Classification Accuracy

- Four classification models were experimented with, and their accuracies are displayed.
- The Decision Tree Classifier emerged as the model with the highest classification accuracy, exceeding 87%.



Confusion Matrix

- The confusion matrix of the Decision Tree Classifier validates its accuracy, evident in the substantial counts of true positive and true negative values in comparison to the false values.



Conclusions

- Various data sources were scrutinized, refining conclusions throughout the process.
- KSC LC-39A emerges as the optimal launch site.
- Launches surpassing 7,000kg exhibit lower risk.
- While the majority of mission outcomes are successful, the success of landing outcomes appears to improve over time, aligning with the evolution of processes and rockets.
- The Decision Tree Classifier serves as a valuable tool for predicting successful landings, potentially contributing to increased profits.

Appendix

- Folium didn't show maps on GitHub, so I took screenshots.

Thank you!