IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<XU HUANCHEN>
<2025-05-30>

Outline

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

Executive Summary

- **Background**
- American private aerospace manufacturer, SpaceX, was founded in 2002 by entrepreneur Elon Musk with the goal of reducing the cost of space transportation and ultimately colonizing Mars .
- SpaceX has developed the reusable Falcon series of launch vehicles. According to its website, Falcon 9 rocket launches have a cost of only 62 million dollars each, while other providers cost upward of 165 million dollars.
- **Objectives**
- This project leverages historical launch data to predict whether the Falcon 9 first stage will land successfully, so that it can be reused.
- **Achievements**
- Developed a predictive model with 94.4% accuracy, providing data-driven decision support to save over 100 million dollars each rocket launch.

Introduction

- **Source of Data**
- Primary launch data was obtained through SpaceX's REST API (v4), supplemented by mission records scraped from Wikipedia's Falcon 9 launch tables to ensure comprehensive coverage of all attempted landings between 2010-2020.
- **Key Methodology**
- Combined data mining, wrangling, EDA, Visualization and Machine Learning.
- **Explore**

1. The overall trend of success rate
2. What are the factors that impact most on the outcome
3. Which model can predict the outcome most accurately

Section 1

Methodology

Methodology

- **Collection**: Request to the SpaceX API; Web scraping from Wikipedia
- **Wrangling**: Filter data, handle missing values and apply one-hot encoding to prepared data for analysis and modeling.
- **Explore**: Explorative Data Analysis with SQL and visualization
- **Visualize**: Build an interactive dashboard with Dash and Plotly.
- **Model development**: Logistic regression, K-Nearest Neighbors, Decision Tree and SVM are used to find the best parameters.

Data Collection

• SpaceX API Data

- Accessed SpaceX's public REST API (v4) using Python's requests library
- (https://api.spacexdata.com/v4/)

• Wikipedia Data Augmentation

- Scraped Falcon 9 mission tables from Wikipedia.
- (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

Steps

- **Request data** from SpaceX API (rocket launch data)
- **Decode response** using .json() and convert to a dataframe using .json_normalize()
- **Request information** about the launches from SpaceX API using custom functions
- **Create dictionaries** from the data
- **Create a dataframe** from the dictionaries
- **Filter the dataframe** to contain only Falcon 9 launches
- **Replace** missing values of Payload Mass with calculated .mean()
- **Export data** to csv file

- GitHub URL: https://github.com/Sywliray/IBM-Data-Science-Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api-v2.ipynb

Data Collection - Scraping

Steps

- **Request data** (Falcon 9 launch data) from Wikipedia
- **Create BeautifulSoup object** from HTML response
- **Extract column names** from HTML table header
- **Collect data** from parsing HTML tables
- **Create dictionary** from the data
- **Create a dataframe** from the dictionary
- **Export data** to csv file

- GitHub URL: https://github.com/Sywliray/IBM-Data-Science-Capstone-Project/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

Steps

- **Perform EDA** and determine data labels
- **Calculate**:
- number of launches for each site
- number and occurrence of orbit
- number and occurrence of mission outcome per orbit type
- **Transform outcome** to binary
- **Export data** to csv file

GitHub URL: https://github.com/Sywliray/IBM-Data-Science-Capstone-Project/blob/main/labs-jupyter-spacex-Data%20wrangling-v2.ipynb

EDA with Data Visualization

Charts

- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit type

Types

- **Scatter plots** show correlation between different variables, especially between continuous variables. Discovering correlations can be useful in feature selection for machine learning.
- **Bar charts** compare discrete categories and show the relationships among the categories and a measured value.

GitHub URL: https://github.com/Sywliray/IBM-Data-Science-Capstone-Project/blob/main/jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

Queries:

- Names of unique launch sites
- 5 records where launch site begins with ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1.
- Date of first successful landing on ground pad
- Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000
- Total number of successful and failed missions
- Names of booster versions which have carried the max payload
- Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015
- Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc)

GitHub URL: https://github.com/Sywliray/IBM-Data-Science-Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Markers Indicating Launch Sites

- Added orange circles at all launch sites coordinates with a popup label showing its name using its latitude and longitude coordinates

Colored Markers of Launch Outcomes

- Added colored markers of successful (green) and unsuccessful(red) launches at each launch site to show which launch site has higher success rates.

Distances Between a Launch Site to Coastline

- Added a line to show distance between a launch site and its proximity to the nearest coastline.

GitHub URL: https://github.com/Sywliray/IBM-Data-Science-Capstone-Project/blob/main/lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

- **Dropdown List with Launch Sites**
- Allow user to select all launch sites or a certain launch site

- **Slider of Payload Mass Range**
- Allow user to select payload mass range

- **Pie Chart Showing Successful Launches**
- Allow user to see the proportion of successful and unsuccessful launches

- **Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version**
- Allow user to see the correlation between Payload and Launch Success

GitHub URL: https://github.com/Sywliray/IBM-Data-Science-Capstone-Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Steps

- **Create** NumPy array from the Class column
- **Standardize** the data with Standard Scaler. Fit and transform the data.
- **Split** the data into training and testing group
- **Create** a GridSearchCV object for parameter optimization
- **Apply** the GridSearchCV object on different algorithms: logistic regression, SVM, decision tree, K-Nearest Neighbor
- **Calculate** accuracy on the test data using .score() for all models
- **Assess** the confusion matrix for all models
- **Identify** the best model using Accuracy Score

- GitHub URL: https://github.com/Sywliray/IBM-Data-Science-Capstone-Project/blob/main/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

Results

- The success rate is generally increasing with the flight number.
- Launch sites VAFB SLC 4E and KSC LC 39A have a higher success rate.
- Comprehensively, missions to VLEO have both a high success rate(nearly 90%) and a great number of launches.
- The total payload mass carried by boosters launched by NASA (CRS) is 99,980kg.
- All launch sites are located near the equator and in proximity to coastline.

Section 2

Insights drawn from EDA

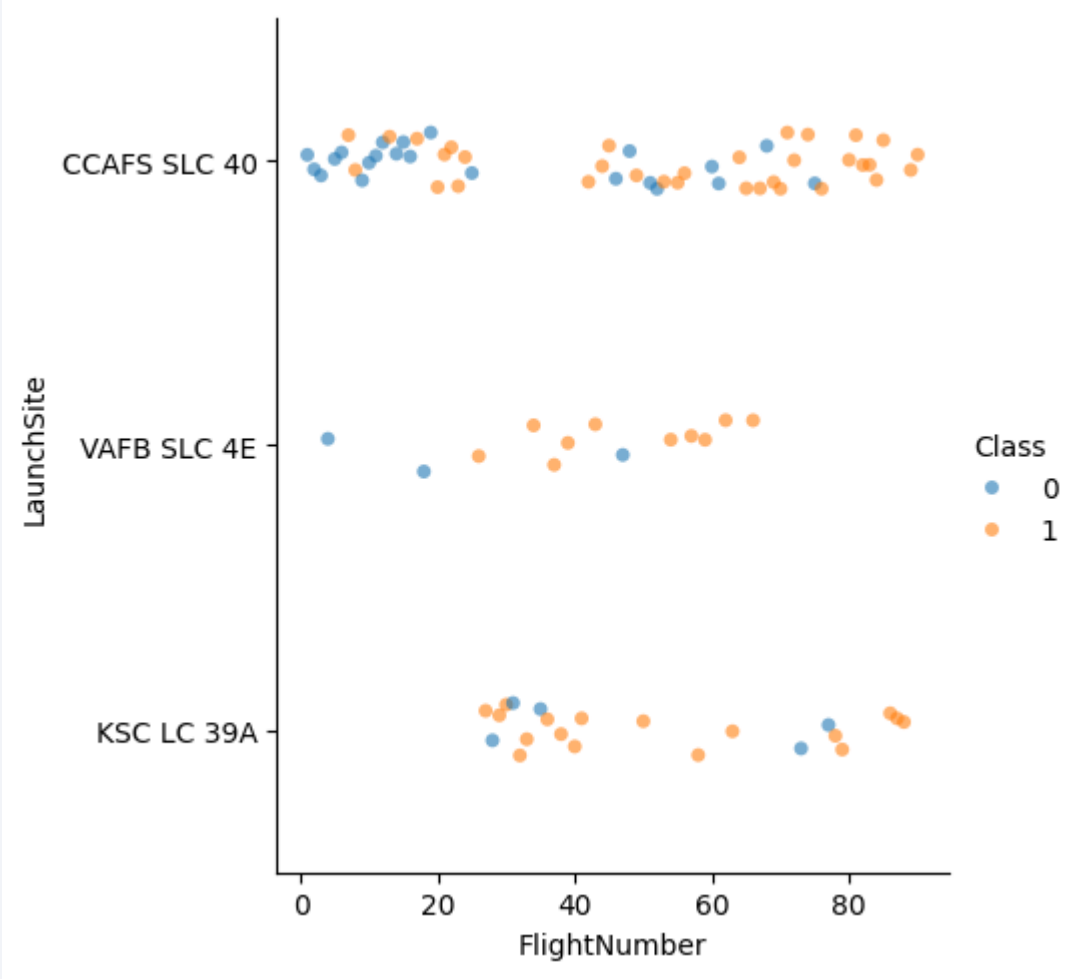
Flight Number vs. Launch Site

Earlier flights have a lower success rate.

Later flights have a higher success rate.

VAFB SLC 4E and KSC LC 39A have a higher success rate.

About half of flights were from CCAFS SLC 40.



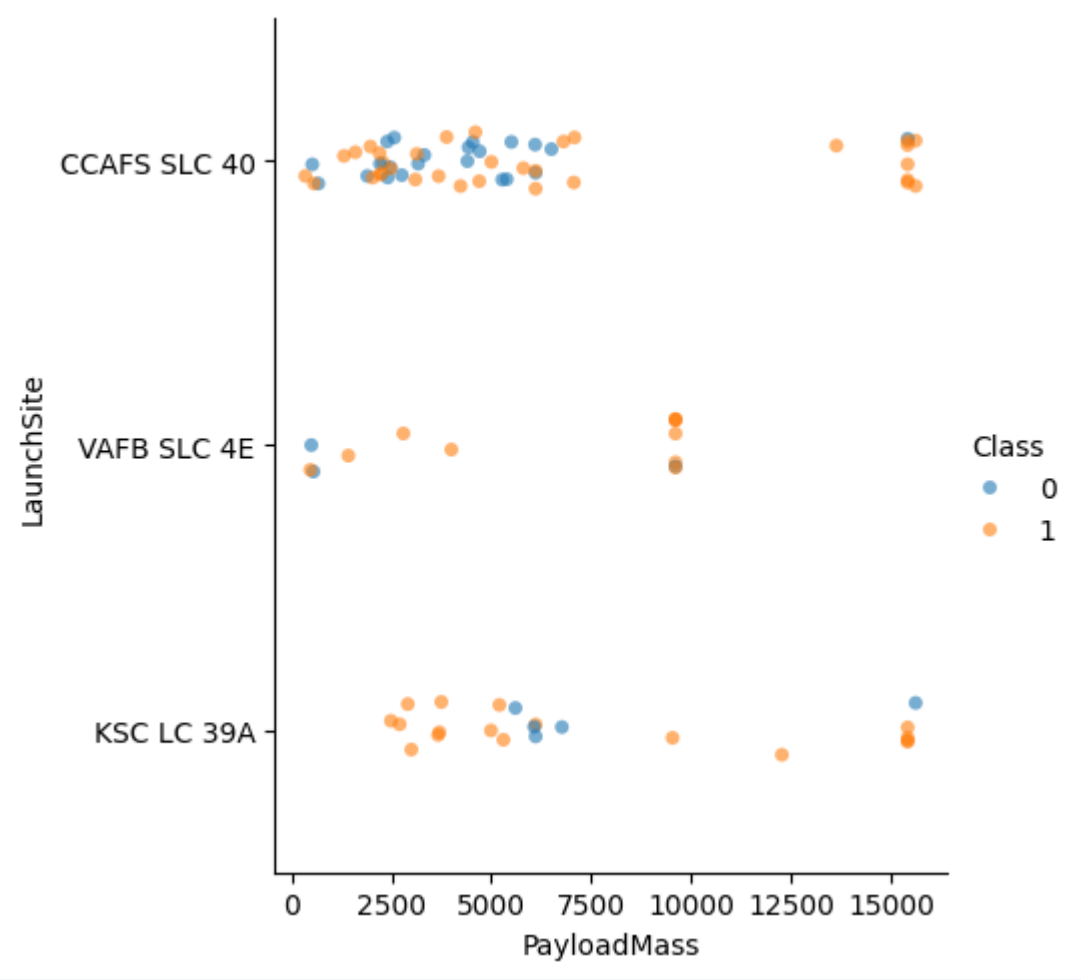
Payload vs. Launch Site

KSC LC 39A has 100% success rate for payload under 5,000kg.

All missions from VAFB SLC 4E were under 10,000 kg.

CCAFS SLC 40 and KSC LC 39A both had heavy-lift missions over 12,000kg and failed once respectively.

CCAFS SLC 40 shows a mixed success rate for payload under 7,500kg.



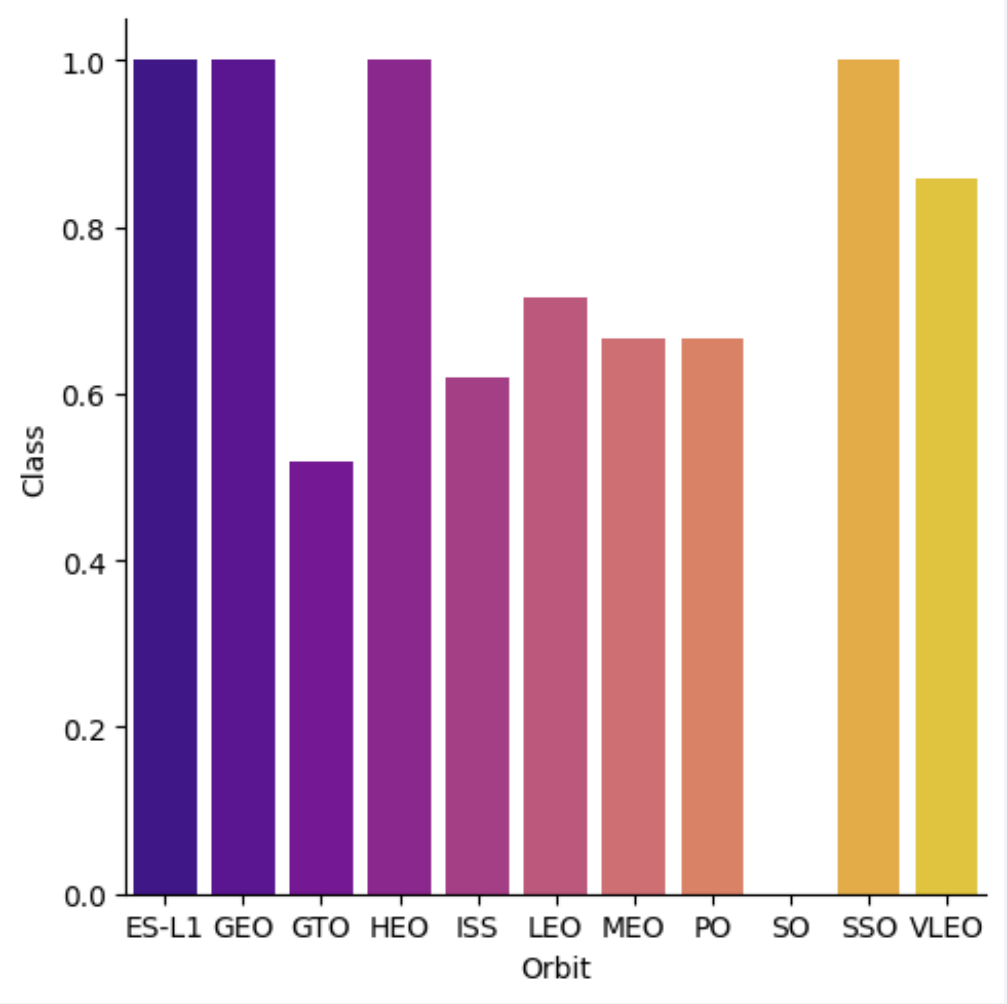
Success Rate vs. Orbit Type

Missions to ES-L1, GEO, HEO, SSO all have 100% success rate.

The success rate for VLEO is nearly 90%.

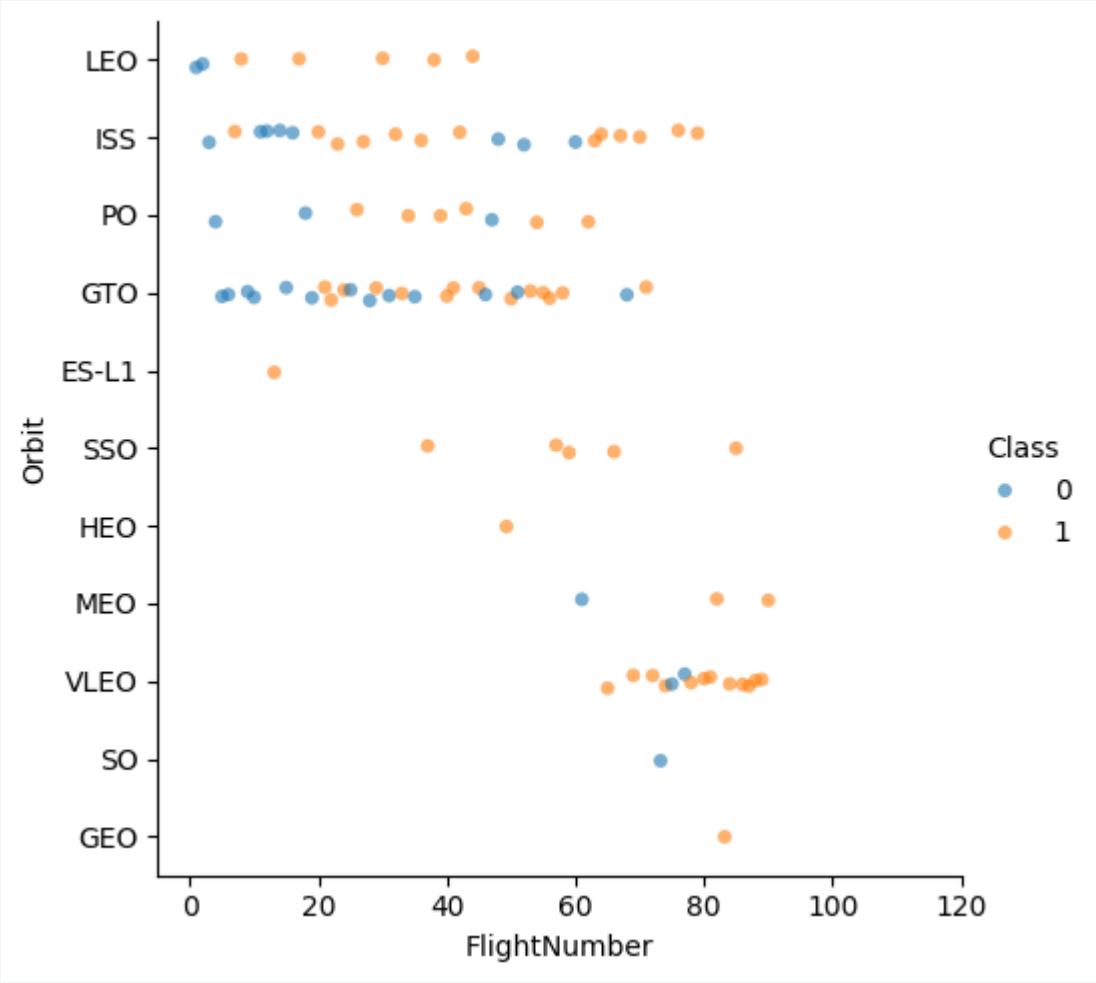
Missions to GTO, ISS, LEO, MEO, PO have 50% to 70% success rate.

The success rate for SO is zero.



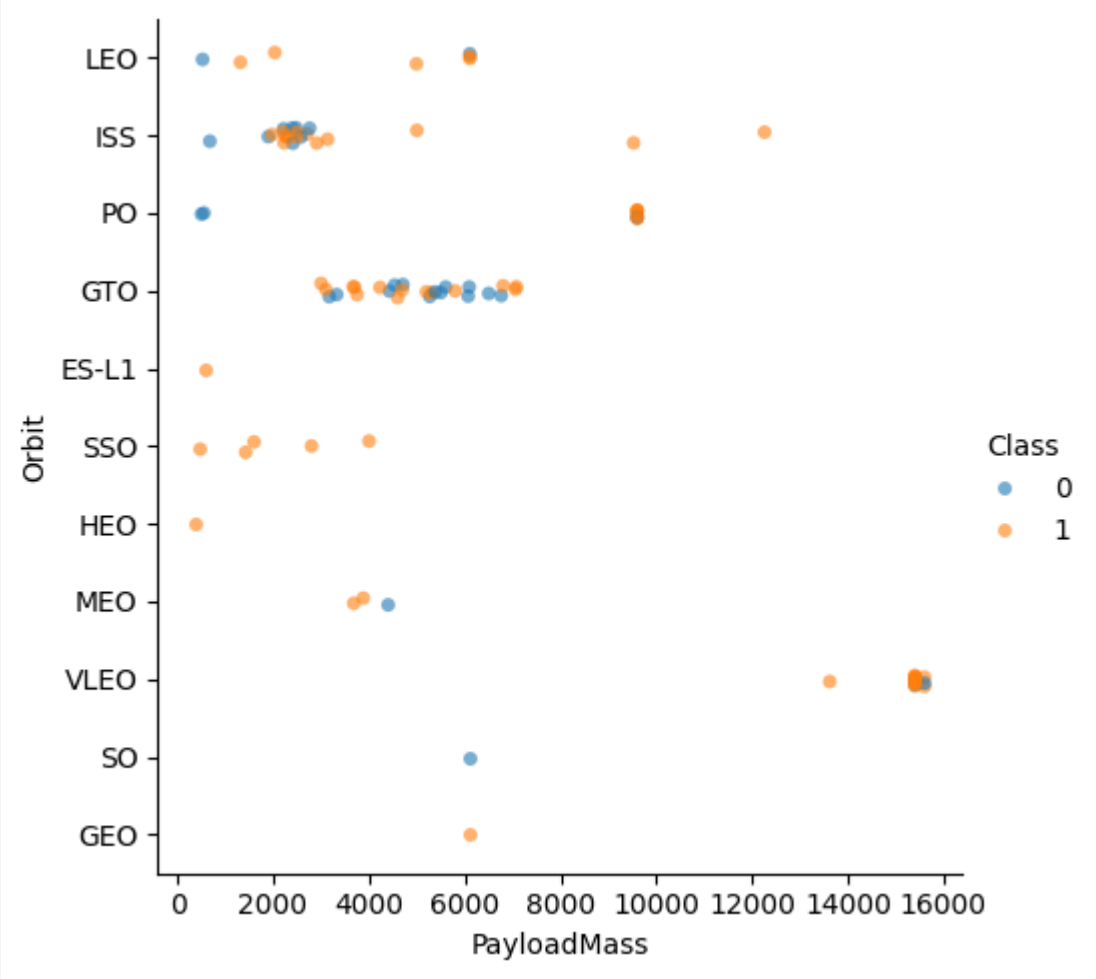
Flight Number vs. Orbit Type

- The 100% success rate for ES-L1, GEO, HEO is because there's one and only successful mission for each.
- 5 missions to SSO all landed successfully.
- Comprehensively, VLEO owns both a high success rate(nearly 90%) and a great number of launches.
- The zero-success of SO is because there's only one launch to it and it failed.
- Typically, success rate increases with flight number, which is clear in the case of LEO. However, GTO does not follow this pattern.



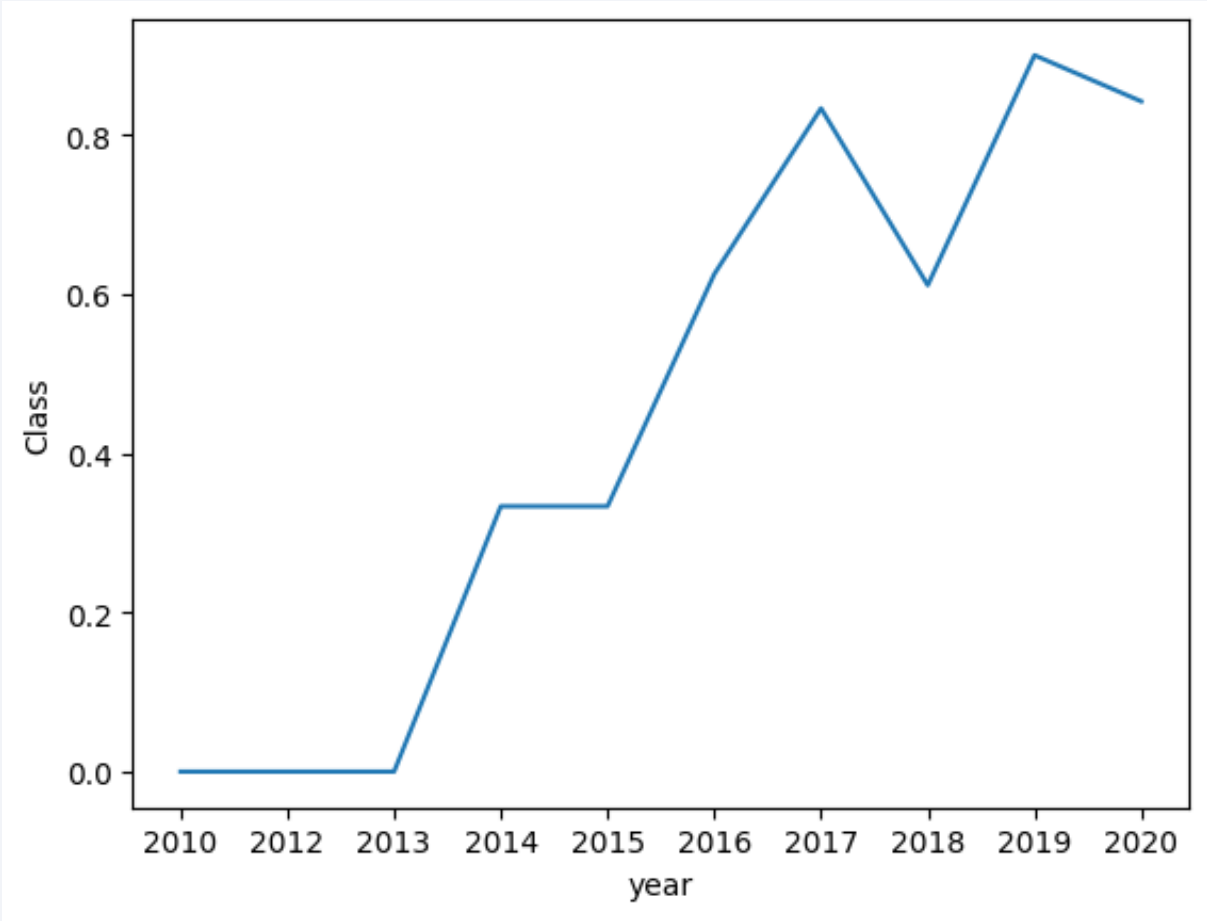
Payload vs. Orbit Type

- Missions to ES-L1, SSO, HEO, MEO with a payload under 4,000kg have 100% success rate.
- Missions to ISS with a payload above 4,000 all landed successfully.
- All missions to VLEO were above 13,000kg.



Launch Success Yearly Trend

- The line chart indicates an increasing trend in success rate over years.
- The success rate increases from 0 to about 90% from 2013 to 2019, with a level-off in 2014 and a dip in 2018 and 2020.



All Launch Site Names

- Falcon 9 has 4 launch sites.

```
%sql SELECT DISTINCT launch_site FROM SPACEXTABLE
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA` are displayed below:

```
%sql SELECT * FROM SPACEXTABLE WHERE launch_site LIKE 'CCA%' LIMIT 5
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Payload Mass

The total payload mass carried by boosters launched by NASA (CRS) is 99,980kg.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass FROM SPACEXTABLE WHERE Customer LIKE 'NASA%'
```

```
 * sqlite:///my_data1.db
Done.
```

total_payload_mass
99980

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2,534.67kg.

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS avg_payload_mass FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%'
```

```
 * sqlite:///my_data1.db
Done.
```

avg_payload_mass
2534.6666666666665

First Successful Ground Landing Date

The date of the first successful landing outcome on ground pad is 2015-12-22.

```
%sql SELECT MIN(DATE) FROM SPACEXTABLE WHERE LANDING_OUTCOME = 'Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

MIN(DATE)
2015-12-22

Successful Drone Ship Landing

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4,000 but less than 6,000 are as shown below:

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE LANDING_OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

```
 * sqlite:///my_data1.db
Done.
```

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 mission outcomes are 100

with 1 failure in flight.

```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) AS Count FROM SPACEXTABLE GROUP BY Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

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

Boosters Carried Maximum Payload

The names of the booster which have carried the maximum payload mass.

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

```
 * sqlite:///my_data1.db
Done.
```

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 failed landing outcomes in drone ship are both launched from CCAFS LC-40 in January and April 2025 respectively, shown as follows:

```
%sql SELECT substr(Date, 6,2) AS Month,Landing_Outcome,Booster_Version,launch_site  FROM (SELECT
```

```
 * sqlite:///my_data1.db
Done.
```

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

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

- Rank of landing outcomes between 2010-06-04 and 2017-03-20 is as follows. Landing on ground pads has a 100% success rate.

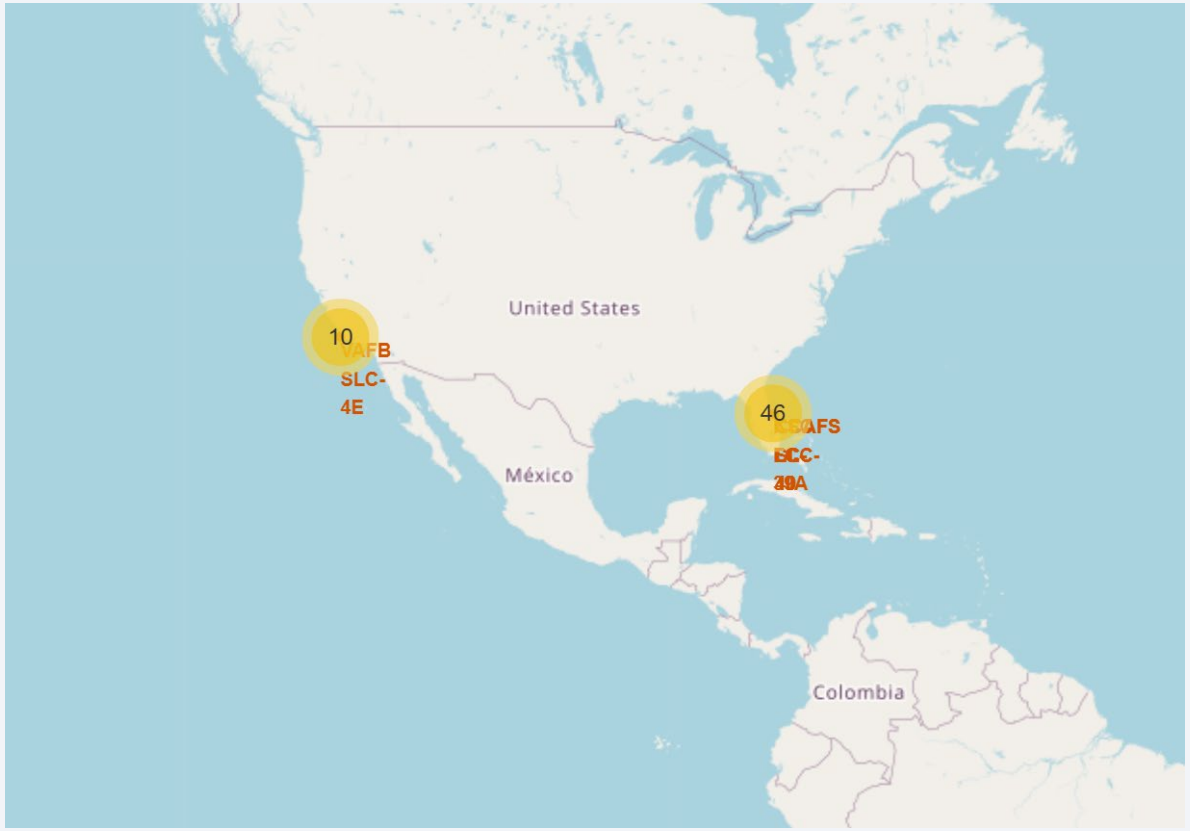
```
%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) AS Count FROM SPACEXTABLE WHER
```

```
 * sqlite:///my_data1.db
Done.
```

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


Section 3
Launch Sites
Proximities Analysis

Launch Sites



All launch sites are located near the equator.

Reasons:

- The Earth rotates fastest at the equator (~1,670 km/h), providing rockets with an extra “boost”.

- Launching near the equator minimizes the energy needed to reach a certain orbit.

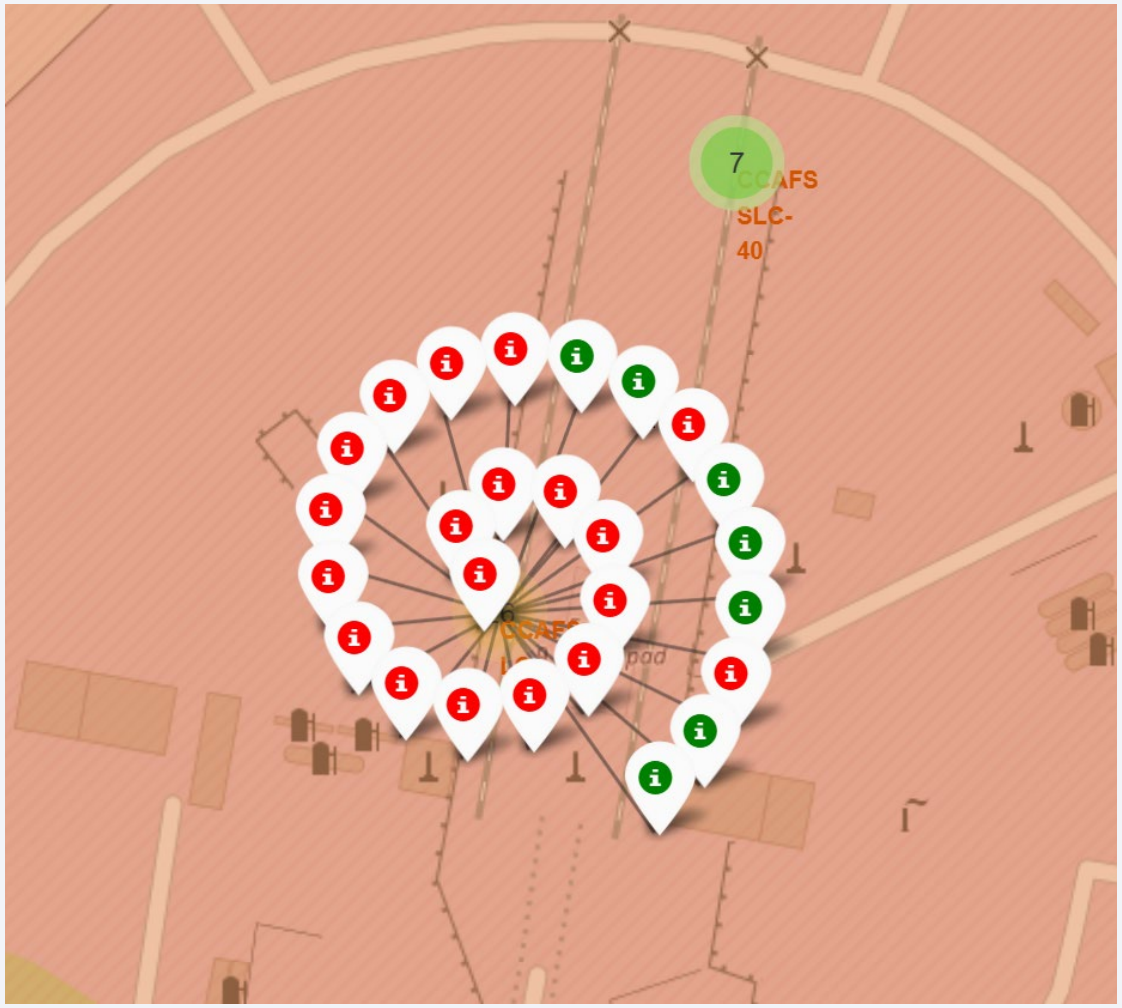
- Proximity to the equator = fuel savings + higher payload capacity, making it the optimal choice for most space missions.

Launch Outcomes

Markers

- Green indicates successful landing.
- Red indicates failed landing.

- CCAFS LC-40 has a 26.9% success rate.



Proximity to the coastline

The distance between VAFB SLC-4E and its nearest coastline is **1.38km.**

Reasons

Safety: Minimizing Risk to Populated Areas

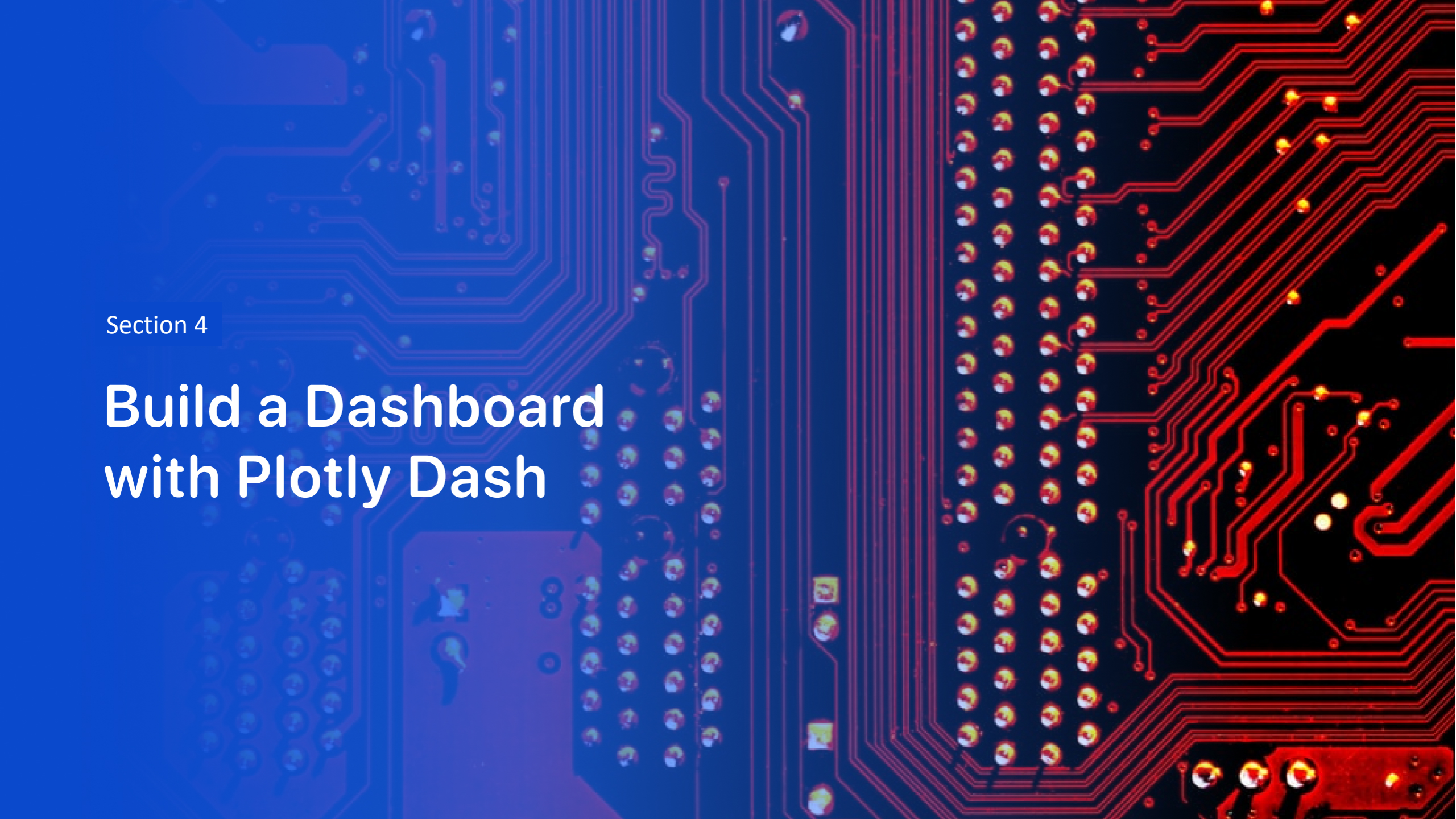
Rockets sometimes fail during launch or have discarded stages that fall back to Earth.

Launching over the ocean ensures that debris or failed rockets fall into water rather than populated areas.

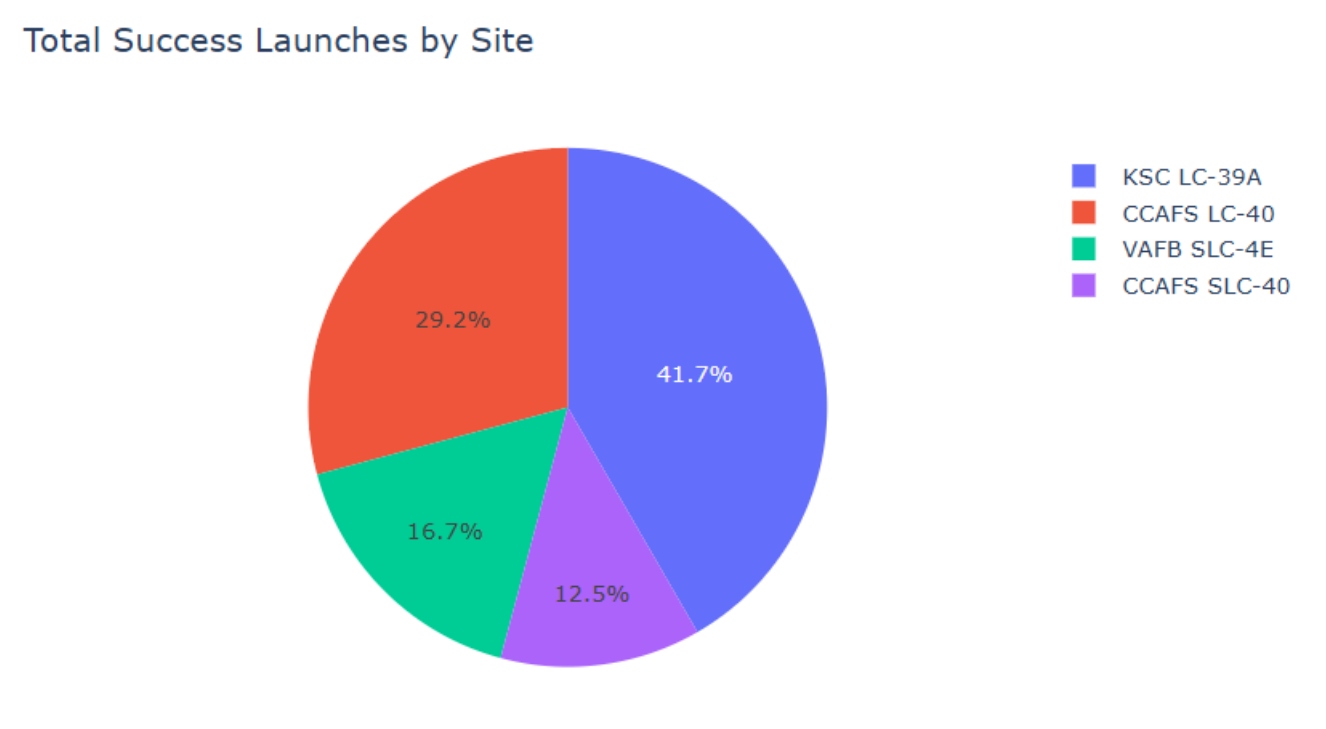
Logistics: Easy Transport of Large Rocket Components

Heavy rockets are too large for road/rail transport.

Coastal sites allow delivery by ship.


Section 4
Build a Dashboard with Plotly Dash

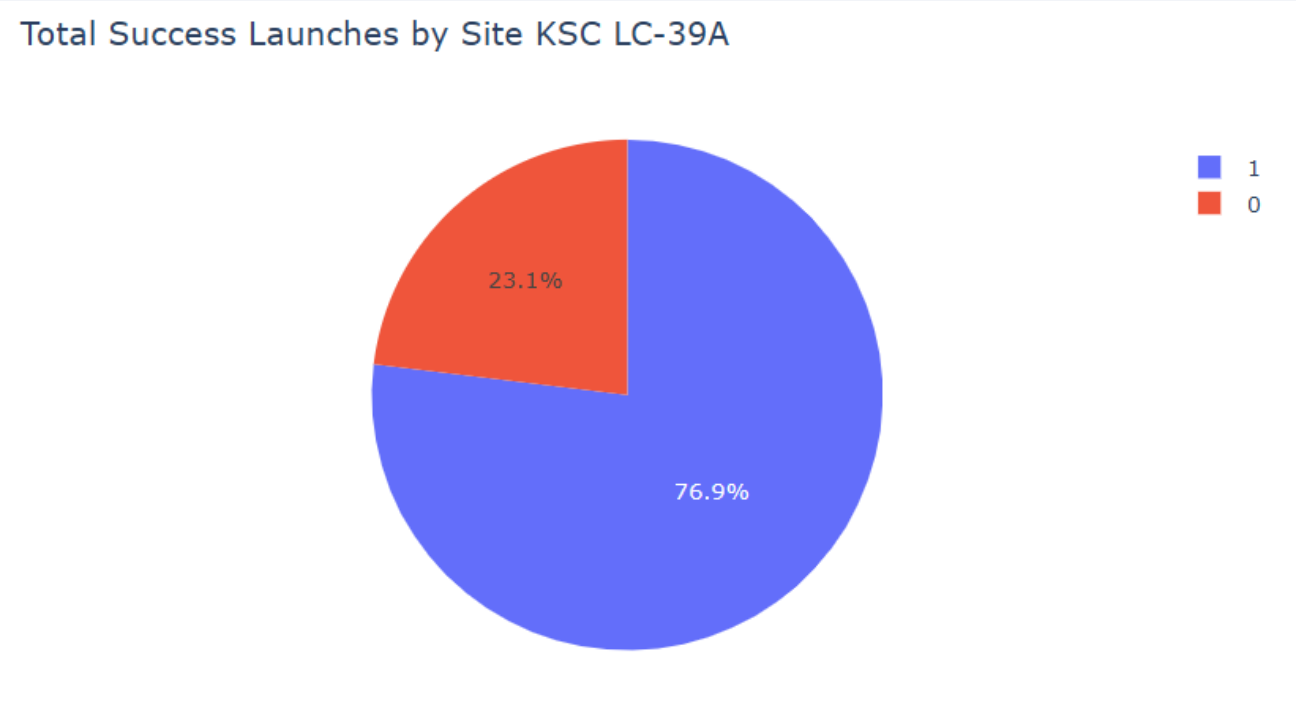
Total Success Launches by Site



- KSC LC-39A **ranked 1st** in count of success among all sites, contributing 41.7% of all successful landings.
- CCAFS LC-40 contributed nearly 30 % of all success.
- CCAFS SLC-40 covered 12.5%, the least among all sites.

Launch Outcomes by KSC LC-39A

- The launch site with highest launch success ratio is **KSC LC-39A**.
- **76.9%** of its launches landed successfully.



Payload vs. Launch Outcome

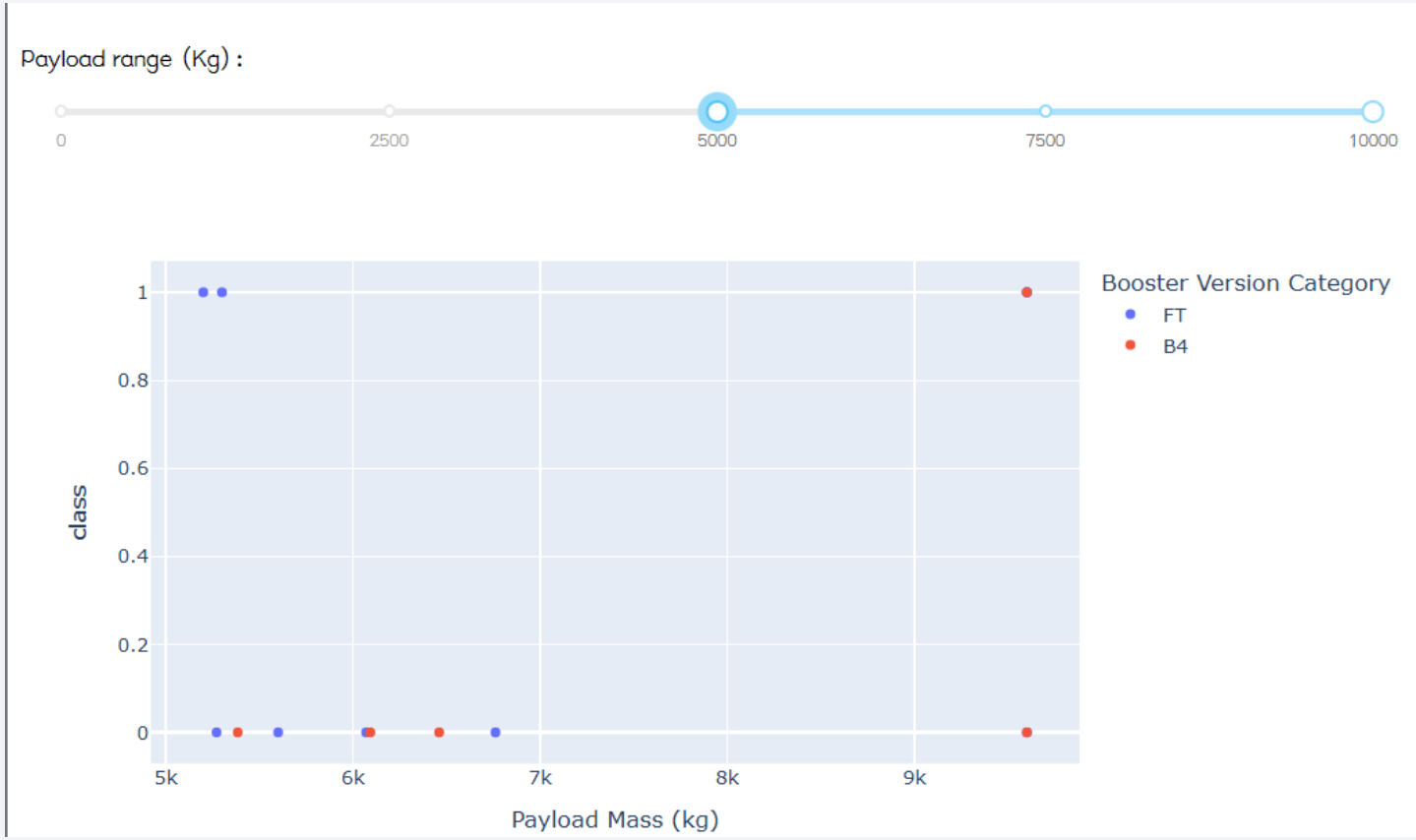
Payload below 5,000kg

- Booster version **FT** and **B4** have the highest success rate and tend to be used in heavier missions in this payload range.
- Booster version **v1.0** and **v1.1** failed in most of their missions, with **v1.1** failed regardless of the payload.

Payload vs. Launch Outcome

Payload from 5,000kg to 10,000kg

- Only Booster Version FT and B4 were used in this payload range.
- Both boosters have a low success rate, with FT slightly better.
- Most of the missions were below 7,000kg.



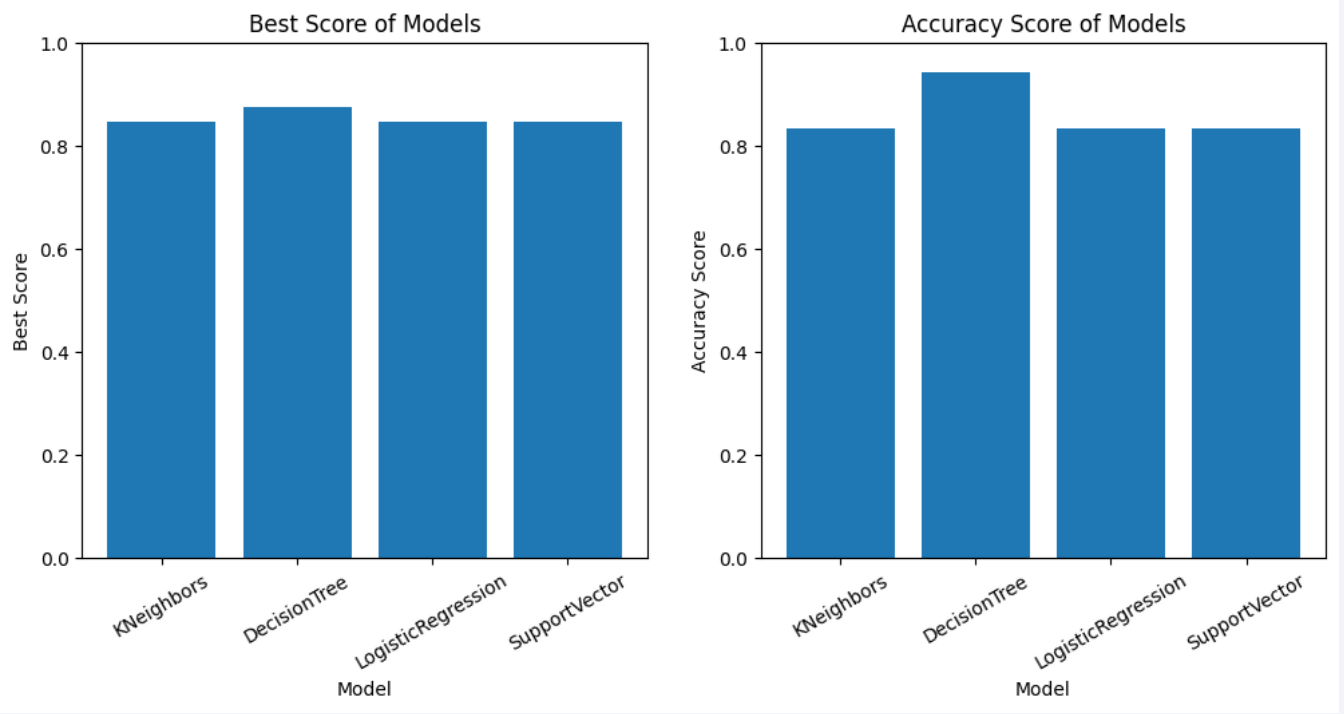
Section 5

Predictive Analysis (Classification)

Classification Accuracy

Models:

- K-Nearest Neighbors
- Decision Tree
- Logistic Regression
- Support Vector Machine (SVM)



Decision Tree performs the best on test data with an accuracy score of **94.4%**.

The accuracy scores of the other 3 models on training and test data are the same, 84.8% and 83.3% respectively.

Confusion Matrix

Confusion matrix of Decision Tree:

Total Samples: 18

True Positive: 12

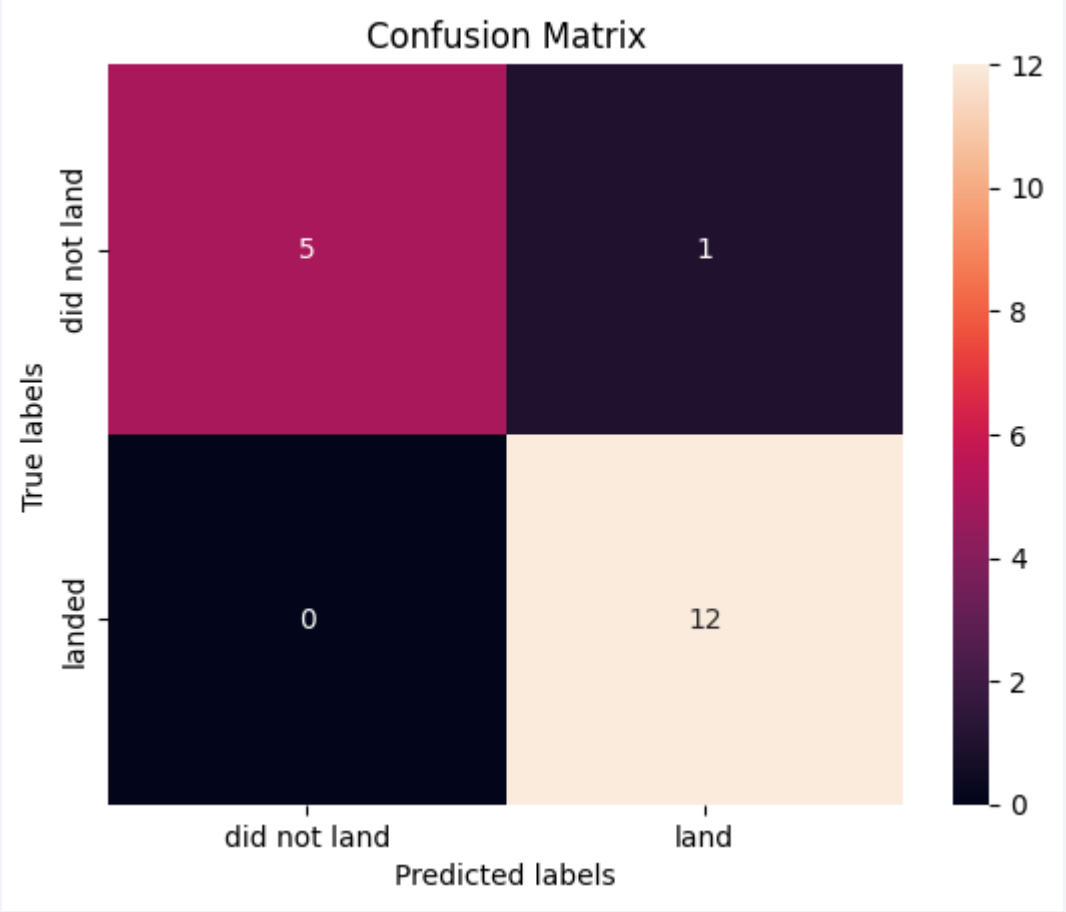
True Negative: 5

False Positive: 1

False Negative: 0

Summary:

- The Decision Tree model performs great on the test data with only 1 Type I Error (False Positive).



Classification Report

The Decision Tree model has a precision rate of **92%**, and a recall rate of **1**.

	precision	recall	f1-score	support
0	1.00	0.83	0.91	6
1	0.92	1.00	0.96	12
accuracy			0.94	18
macro avg	0.96	0.92	0.93	18
weighted avg	0.95	0.94	0.94	18

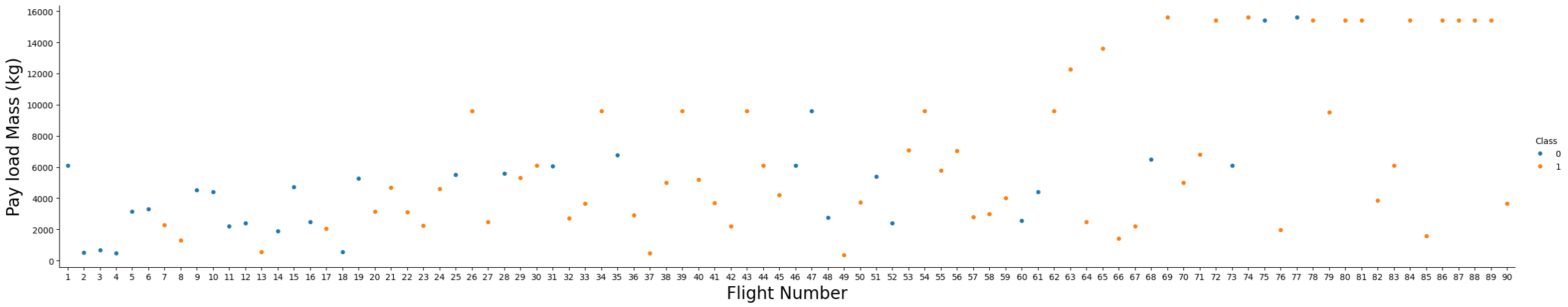
Conclusions

- The success rate is generally increasing with the flight number.
- Launch site KSC LC 39A has the highest success rate among all, thus are more recommended for future launches.
- All launch sites are located near the equator and in proximity to coastline, for both safety and transportation reasons.
- As payload increase, the success rate tends to be higher.
- Missions targeting ES-L1, GEO, HEO, SSO all have 100% success rate.
- Decision Tree model is the most suitable model for predicting outcomes, and has an accuracy of 94.4% on test data.
- **Improvements**:
- The model can be better tested and tuned to be more generalized if the dataset is bigger.

Appendix

Correlation between Payload and Flight Number

- In overall, SpaceX is upgrading from lower payload mass to higher ones above 14,000kg, and the success rate is increasing over time.
- From this trend, we expect launches with higher payload mass and higher success rate in the upcoming future.



Thank you!