IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Asad Moosvi
11/08/2025

Outline

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

Executive Summary

Methodologies Summary
- Collected data
- Performed data wrangling
- Conducted exploratory data analysis using visualizations
- Conducted exploratory data analysis using SQL
- Created an interactive map with Folium
- Developed a dashboard with Plotly Dash
- Performed predictive analysis using classification models

Results Summary
- Findings from exploratory data analysis
- Interactive analytics demonstration (screenshots)
- Outcomes of predictive analysis

Introduction

Project Background and Context
SpaceX is the leading company in the commercial space era, making space travel more affordable. Falcon 9 rocket launches are advertised at 62 million dollars, compared to over 165 million dollars from other providers. Much of this cost advantage comes from reusing the first stage. Predicting the likelihood of a first-stage landing allows for estimating launch costs. Using public data and machine learning models, this project aims to predict whether SpaceX will reuse the first stage.

Questions to be Answered
- How do factors such as payload mass, launch site, number of flights, and orbit type influence first-stage landing success?
- Has the rate of successful landings improved over time?
- Which algorithm is most effective for binary classification in this scenario?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Web Scraping Wikipedia
- Perform data wrangling
  - Filtered the dataset
  - Handled missing values
  - Applied One-Hot Encoding to prepare data for binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

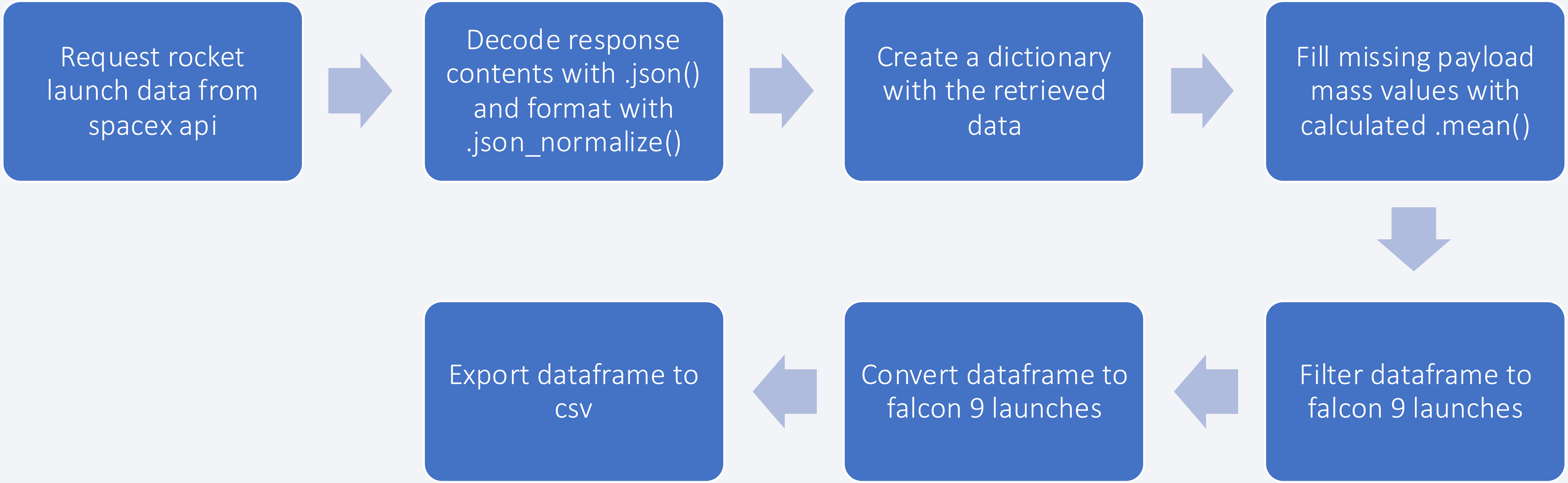
- The data collection process combined API requests from the SpaceX REST API with web scraping of a table from SpaceX's Wikipedia page.

  Both methods were required to gather complete launch information for a more detailed analysis.

  Data columns obtained via SpaceX REST API:
  FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

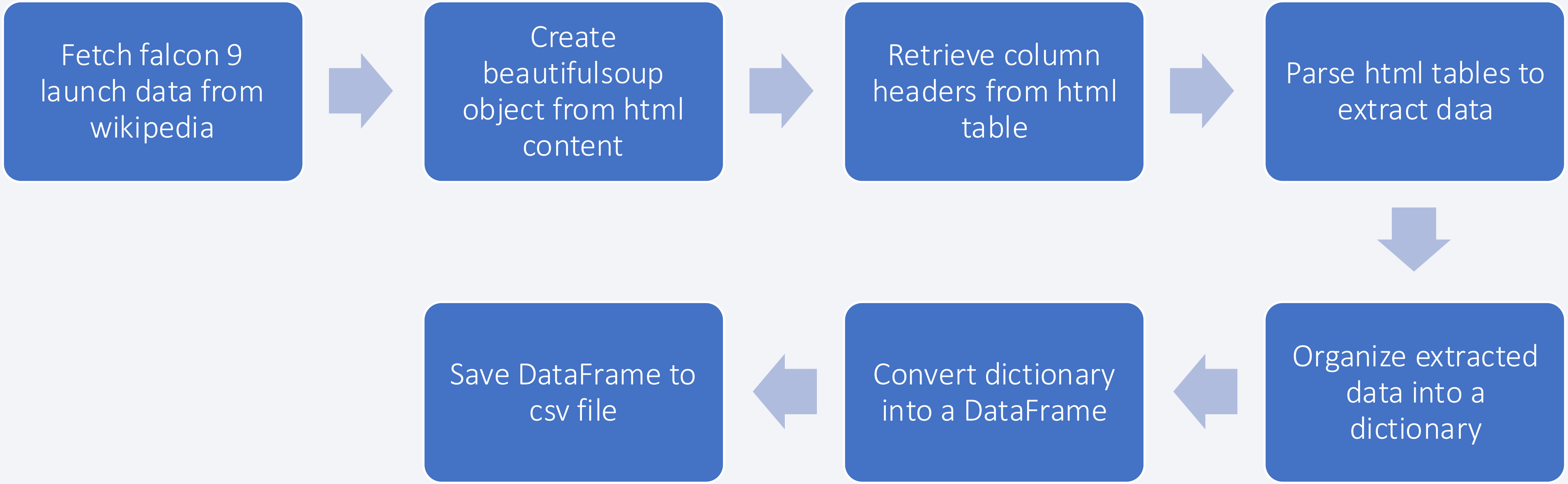
  Data columns obtained via Wikipedia web scraping:
  Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



https://github.com/asadmoosvi/datascience-capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

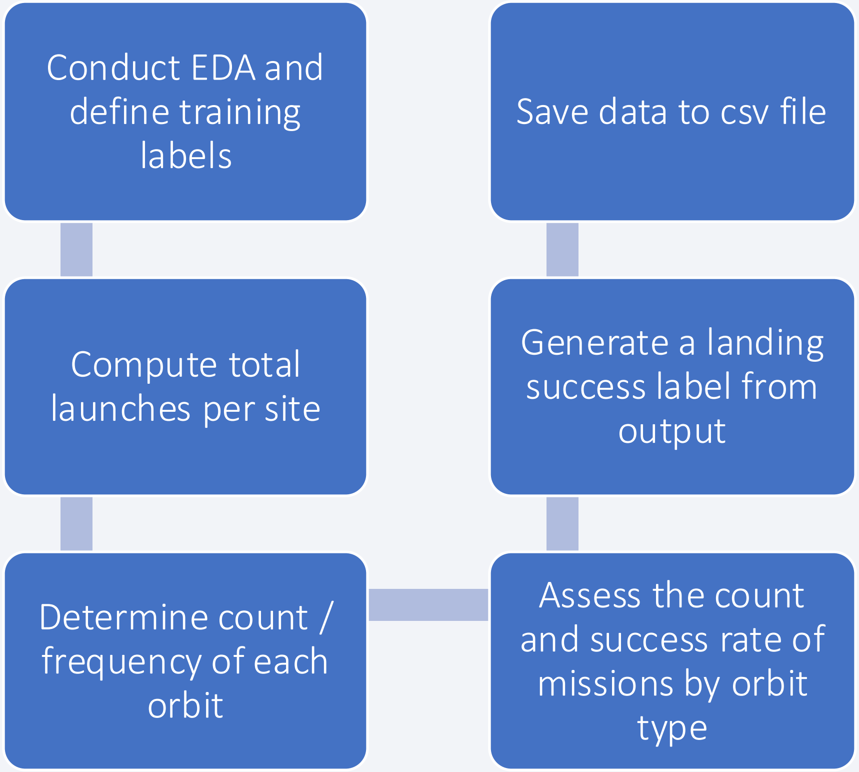
Data Collection - Scraping



https://github.com/asadmoosvi/datascience-capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

In the dataset, there are multiple scenarios where the booster did not land successfully. In some cases, a landing was attempted but failed due to accidents. For instance:
- True Ocean: Successfully landed in a designated ocean area.
- False Ocean: Attempted landing in a designated ocean area but failed.
- True RTLS: Successfully landed on a ground pad.
- False RTLS: Attempted landing on a ground pad but failed.
- True ASDS: Successfully landed on a drone ship.
- False ASDS: Attempted landing on a drone ship but failed.



https://github.com/asadmoosvi/datascience-capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Charts created include:
Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs. Orbit Type, and Yearly Trend of Success Rate.

Scatter plots illustrate relationships between variables, which, if present, can be leveraged in machine learning models.

Bar charts compare discrete categories, aiming to highlight the relationship between each category and its associated measured value.

Line charts display data trends over time (time series).

https://github.com/asadmoosvi/datascience-capstone/blob/main/edadataviz.ipynb

EDA with SQL

Executed SQL queries to:

- Retrieve the names of unique launch sites in the dataset.
- Display 5 records with launch sites starting with ‘CCA’.
- Calculate the total payload mass carried by boosters launched for NASA (CRS).
- Determine the average payload mass for booster version F9 v1.1.
- Identify the date of the first successful ground pad landing.
- List boosters with successful drone ship landings and payload mass between 4000 and 6000.
- Count the total number of successful and failed mission outcomes.
- Identify booster versions that carried the maximum payload mass.
- Retrieve failed drone ship landings in 2015, along with booster versions and launch site names.
- Rank landing outcomes (e.g., Failure (drone ship), Success (ground pad)) between 2010-06-04 and 2017-03-20 in descending order.

https://github.com/asadmoosvi/datascience-capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Markers for All Launch Sites:
- Placed a marker with a circle, popup label, and text label for NASA Johnson Space Center using its latitude and longitude as the starting location.
- Placed markers with circles, popup labels, and text labels for all launch sites based on their latitude and longitude to illustrate their geographic positions and proximity to the equator and coastlines.

Coloured Markers for Launch Outcomes:
- Used green markers for successful launches and red markers for failed launches, grouped with Marker Cluster to highlight launch sites with higher success rates.

Distances from a Launch Site to Nearby Features:
- Drew coloured lines showing the distances from the KSC LC-39A launch site (example) to nearby features such as railways, highways, the coastline, and the nearest city.

https://github.com/asadmoosvi/datascience-capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:
- Implemented a dropdown menu to allow selection of a specific launch site.

Pie Chart of Successful Launches (All Sites or Selected Site):
- Created a pie chart displaying the total count of successful launches for all sites, or success vs. failure counts when a specific site is selected.

Payload Mass Range Slider:
- Added a slider to filter launches by payload mass range.

Scatter Plot of Payload Mass vs. Success Rate by Booster Version:
- Created a scatter plot illustrating the relationship between payload mass and launch success across different booster versions.

https://github.com/asadmoosvi/datascience-capstone/blob/main/spacex-dash.py

Predictive Analysis (Classification)

1. Generating a NumPy array using the "Class" column from the dataset.
2. Normalizing the data using StandardScaler, followed by fitting and transformation.
3. Dividing the dataset into training and testing sets with the train_test_split function.
4. Constructing a GridSearchCV object with cv=10 to identify optimal parameters.
5. Identifying the top-performing method by evaluating Jaccard score and F1-score metrics.
6. Analyzing the confusion matrix across all models.
7. Measuring the accuracy on the test data using the method.score() for each model.
8. Implementing GridSearchCV on Logistic Regression, SVM, Decision Tree, and KNN models.

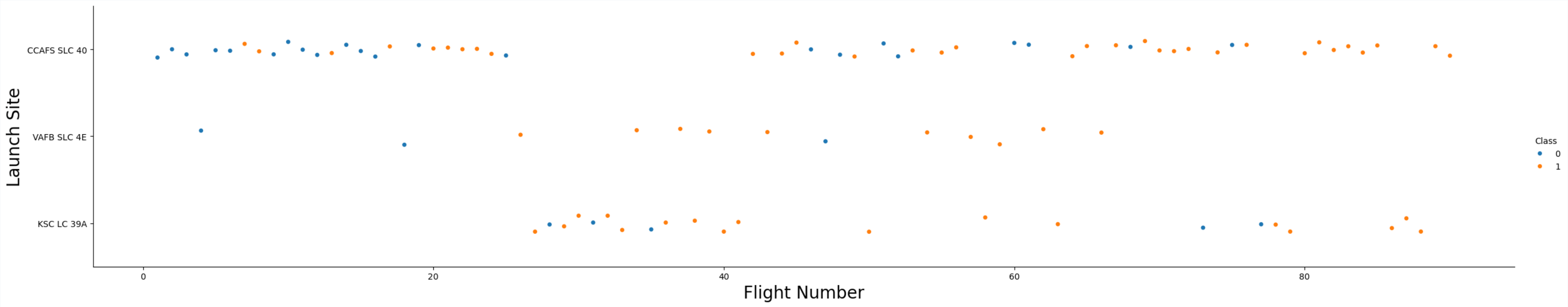
https://github.com/asadmoosvi/datascience-capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

EDA with Visualization

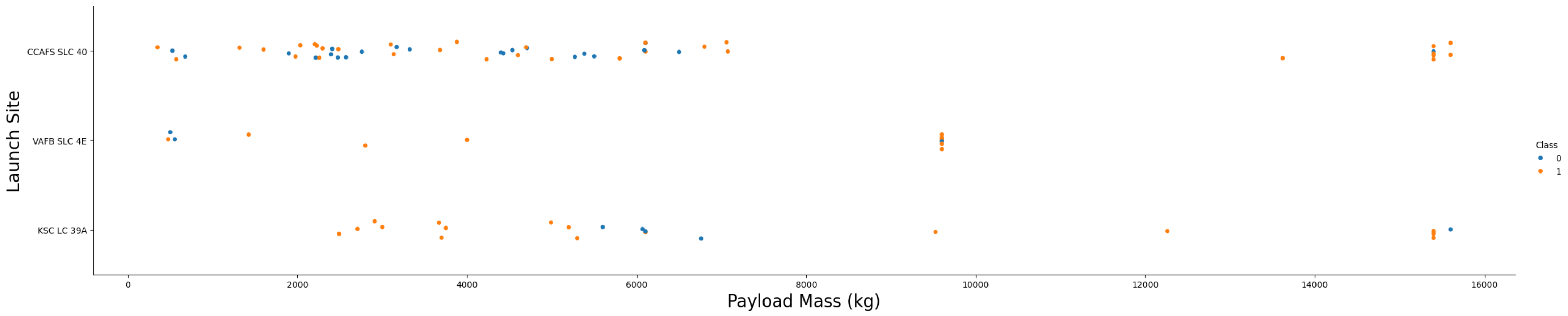
Flight Number vs Launch Site



Explanation:
- Early flights mostly failed, while recent flights have all been successful.
- The CCAFS SLC 40 site accounts for roughly half of all launches.
- VAFB SLC 4E and KSC LC 39A show higher success rates.
- Overall, newer launches tend to have a higher likelihood of success.

EDA with Visualization

Payload vs Launch Site



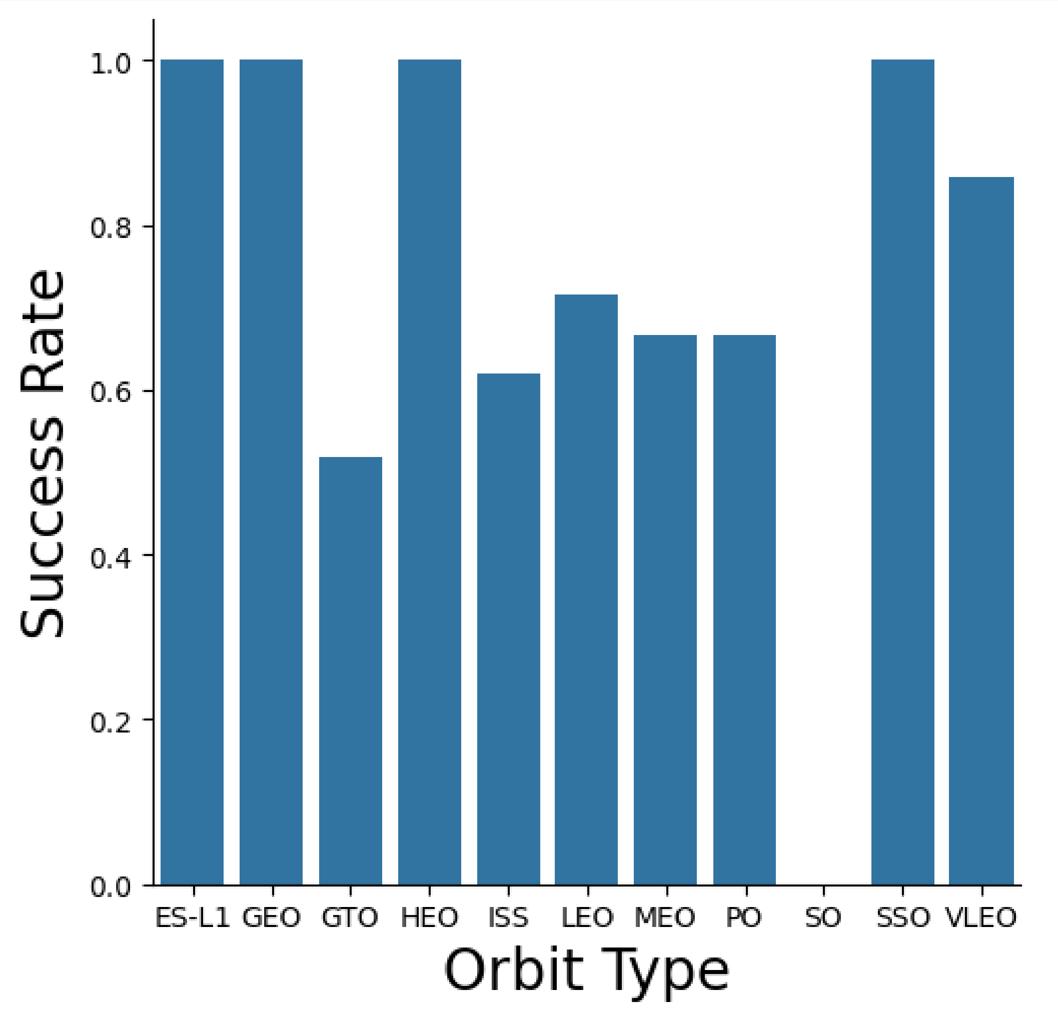
Explanation:

- At all launch sites, higher payload mass is generally associated with higher success rates.
- Most launches with payloads exceeding 7000 kg were successful.
- KSC LC 39A achieved a 100% success rate for payloads under 5500 kg as well.

EDA with Visualization

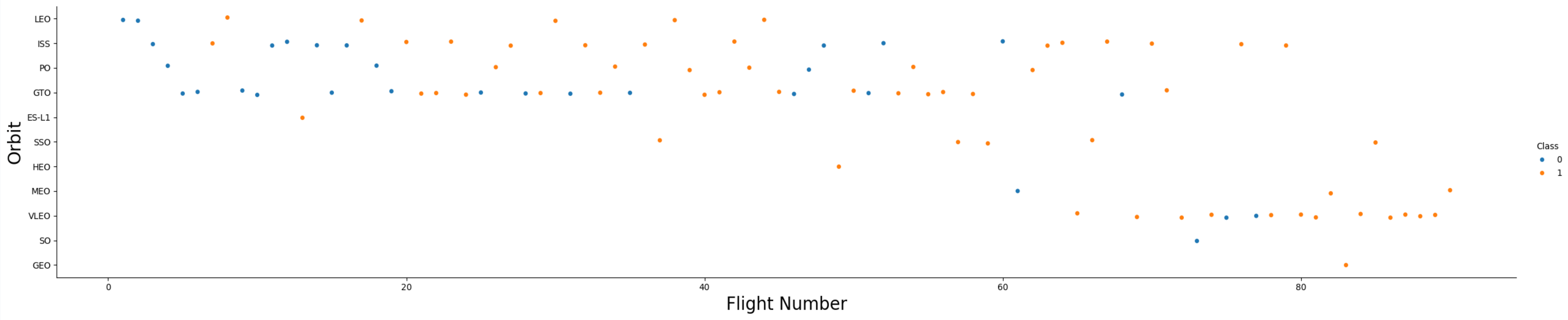
Success Rate vs Orbit Type

Explanation:
- Orbits with a 100% success rate: ES-L1, GEO, HEO, SSO
- Orbits with a 0% success rate: SO
- Orbits with success rates between 50% and 85%: GTO, ISS, LEO, MEO, PO



EDA with Visualization

Flight Number vs Orbit Type



Explanation:
- In the LEO orbit, success appears to be linked to the number of flights.
- In contrast, no clear relationship is observed between flight number and success in the GTO orbit.

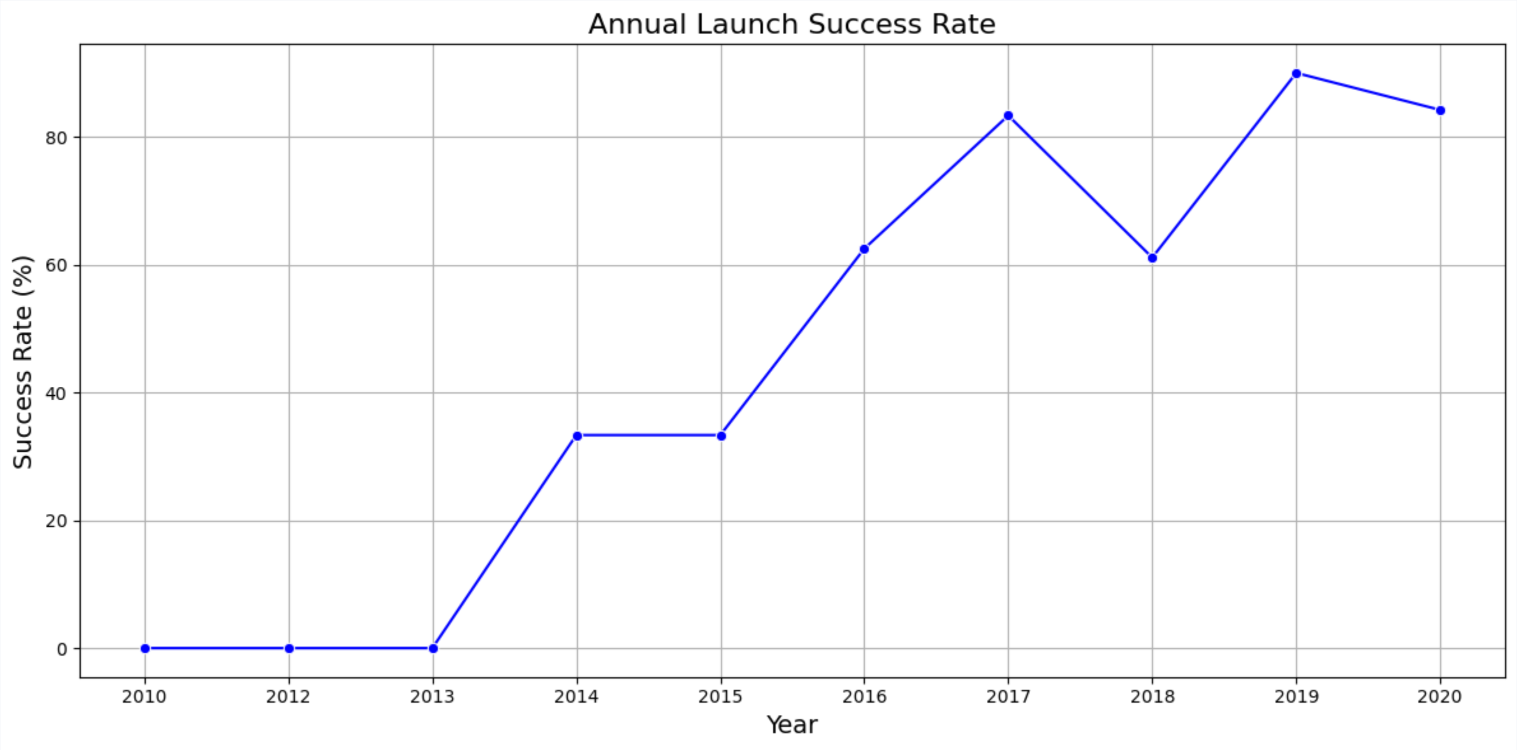
EDA with Visualization

Payload Mass vs Orbit Type

Explanation:
- Heavy payloads tend to negatively affect success in GTO orbits, but have a positive impact in Polar LEO (ISS) orbits.

EDA with Visualization

Launch Success Yearly Trend



Explanation:
The success rate steadily kept increasing from 2013 up until 2020. There was a dip in 2018, but overall it kept going up.

EDA with SQL

All launch site names

Task 1

Display the names of the unique launch sites in the space mission

In [12]:
```
%sql select distinct launch_site from SPACEXTABLE;
```

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

Out[12]:

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

EDA with SQL

Launch site names that begin with CCA

Out[13]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Ou
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	(para
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	(para
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No a
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No a
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No a

EDA with SQL

Total payload mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
In [14]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';

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

Out[14]:

total_payload_mass
45596

EDA with SQL

Average payload mass

Task 4

Display average payload mass carried by booster version F9 v1.1

In [15]:
```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%
```

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

Out[15]:

average_payload_mass
2534.6666666666665

EDA with SQL

First successful ground landing date

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [17]:
```
%sql select min(date) as first_successful_landing from SPACEXTABLE where landing_outcome = 'Success (ground pad)'
```

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

Out[17]:

first_successful_landing
2015-12-22

EDA with SQL

Successful drone ship landing with payload b/w 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [20]:

```
%sql select booster_version from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and payload_mass__kg_
```

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

Out[20]:

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

EDA with SQL

Total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

In [21]:
```
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

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

Out[21]:

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

EDA with SQL

Boosters carried max payload

Task 8

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

In [23]:
```
%sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEX
```

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

Out[23]:

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

EDA with SQL

2015 Launch Records

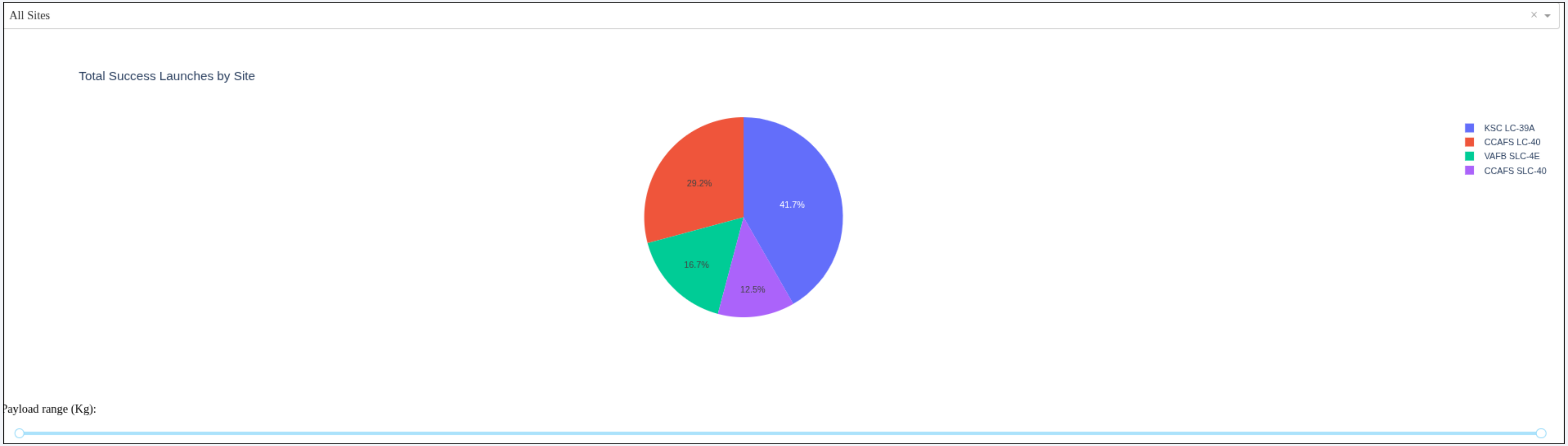
```
In [33]: %sql select substr(Date, 6, 2) as month, date, booster_version, launch_site, landing_outcome from SPACEXTABLE whe

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

Out[33]:

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)
01	2016-01-17	F9 v1.1 B1017	VAFB SLC-4E	Failure (drone ship)
03	2016-03-04	F9 FT B1020	CCAFS LC-40	Failure (drone ship)
06	2016-06-15	F9 FT B1024	CCAFS LC-40	Failure (drone ship)

Dashboard with Plotly and Dash

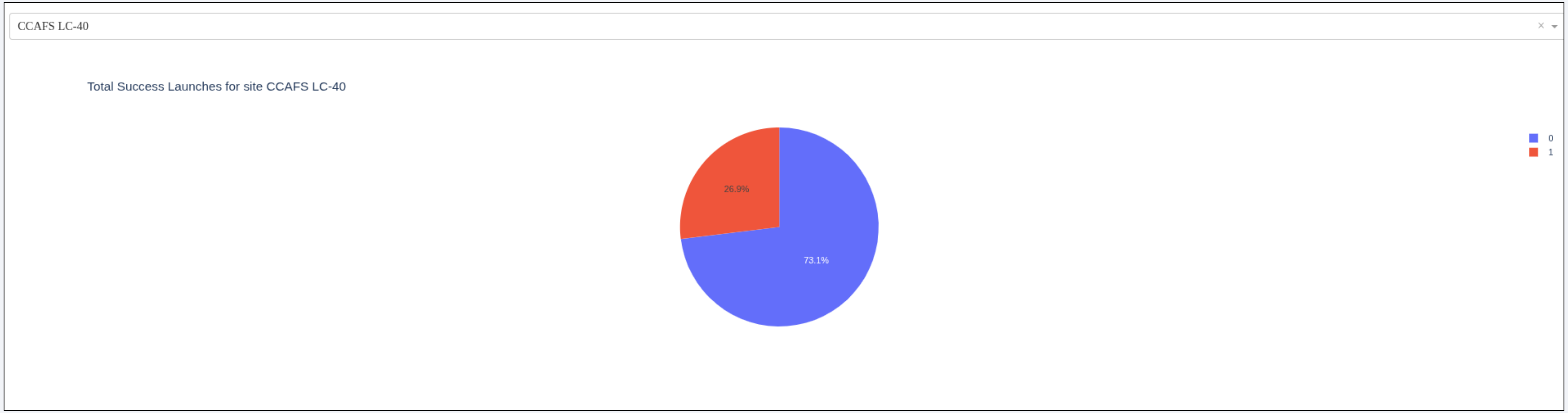
Launch Success Count for All Sites



The chart indicates that KSC LC-39A has the highest number of successful launches among all sites.

Dashboard with Plotly and Dash

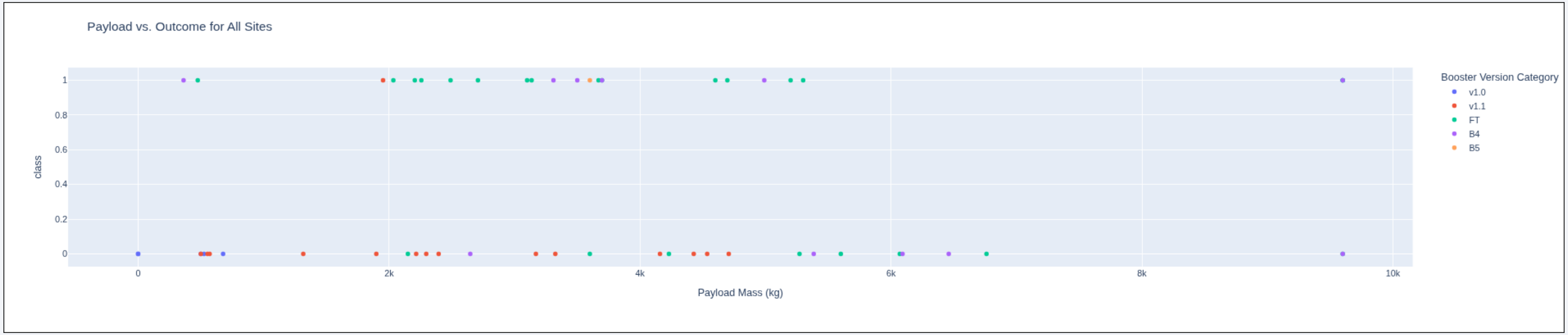
Launch site with highest success ratio



- KSC LC-39A achieves the highest launch success rate at 76.9%, with 10 successes and only 3 failures.

Dashboard with Plotly and Dash

Payload Mass vs Launch Outcome for All Sites

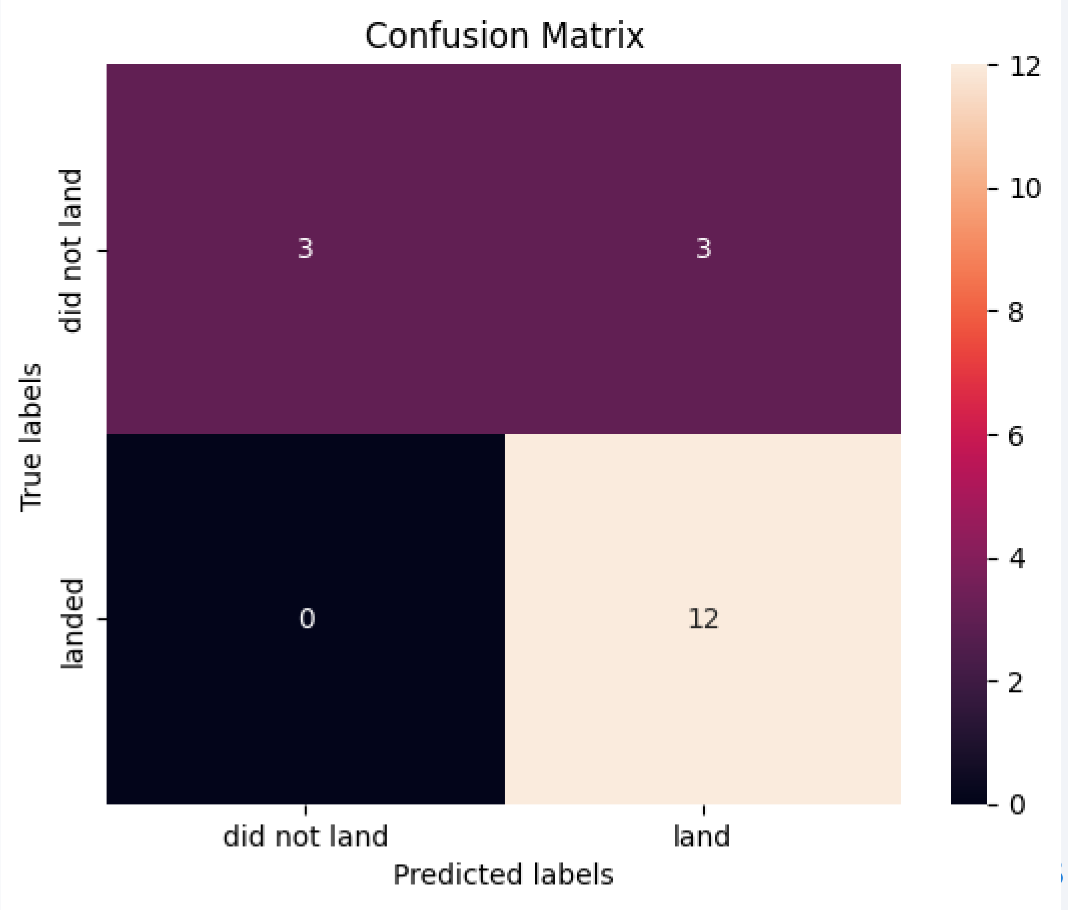


The chart shows that payloads ranging from 2000 to 5500 kg have the highest success rate.

Predictive Analysis (Classification)

Confusion Matrix

Examining the confusion matrix reveals that logistic regression can differentiate between the classes, with the main issue being false positives.



Conclusions

- Launches with lower payload mass generally have better outcomes than those with higher payload mass.
- Most launch sites are near the equator, with all located close to the coast.
- Launch success rates have increased over time.
- KSC LC-39A has the highest success rate among all sites.
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate.

Thank you!