IBM Data Science Capstone

Kingsley Chijioke

Outline

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

Executive Summary

- Methodologies Overview
  - Data collection via SpaceX API and web scraping
  - Data wrangling and preparation
  - Exploratory Data Analysis (EDA) with visualization
  - Exploratory Data Analysis with SQL
  - Interactive mapping with Folium
  - Dashboard development with Plotty Dash
  - Predictive analysis using classification
- Summary of Results
  - EDA Results and Insights
  - Interactive Analysis Demonstrations
  - Predictive Analysis Outcomes

Introduction

- Project Context
  - SpaceX, the leader in commercial space, revolutionized travel by making it more accessible. Its website showcases Falcon 9 rocket launches priced at $62 million, significantly lower than other providers' cost of over $165 million. This cost reduction stems from SpaceX's reuse of the rocket's first stage.
  - To accurately predict launch costs, we propose a predictive model using public information and machine learning techniques. We aim to predict the likelihood of first stage reuse during a launch.
- We seek answers to:
  - How payload mass, launch site, number of flights, and orbits affect first stage landing success.
  - Does the rate of successful landings increase over time?
  - What's the best algorithm for binary classification in this case?

Methodology

- Data Collection Methodology
  - Via SpaceX API
  - Web scraping Wikipedia
- Data Wrangling Process
  - Binary classification preparation
  - Missing value handling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics
  - Using Folium and Plotty Dash
- Perform predictive analysis using classification models
  - Build, tune and evaluation classification models

Data Collection

- Data collection involved API requests from SpaceX REST API and Web Scraping from a table in SpaceX’s Wikipedia entry. Both methods were used to obtain complete information for detailed analysis.

Data Collection – SpaceX API

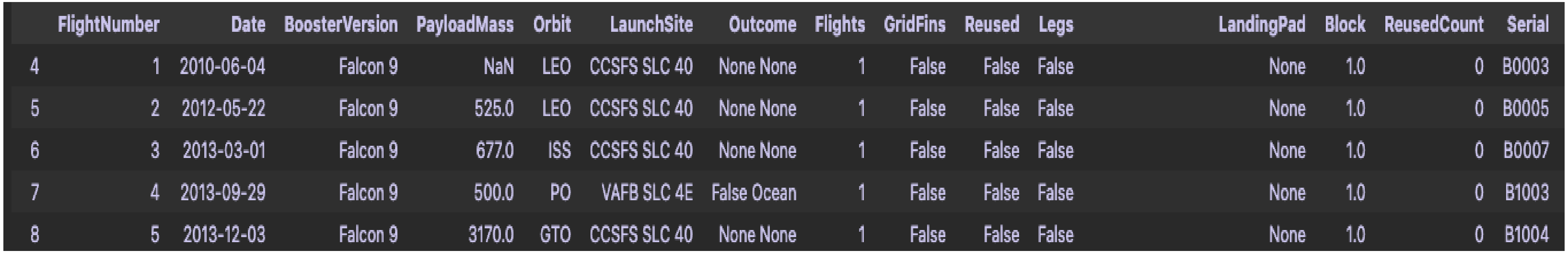
Get response from SpaceX API

Turn response into data frame and filter

Filter data frame so it only includes Falcon 9 launches

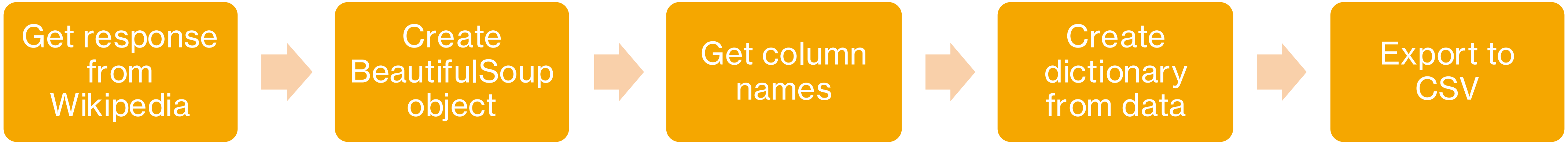
Create dictionary from data

Export to CSV



	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004

Data Collection – Web Scraping



Get response from Wikipedia

Create BeautifulSoup object

Get column names

Create dictionary from data

Export to CSV

Data Wrangling

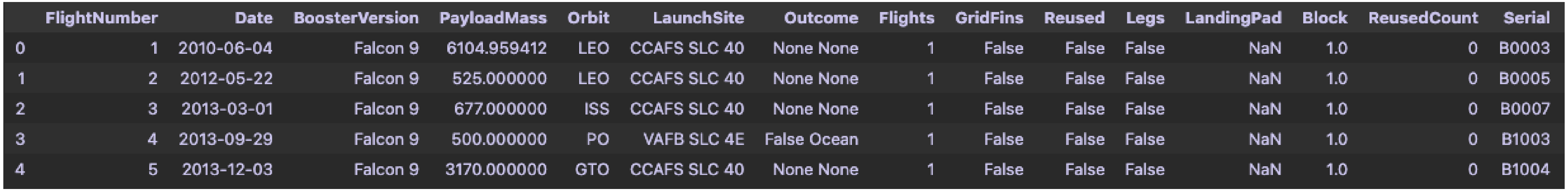
CALCULATE NUMBER OF LAUNCHES FOR EACH SITE

CALCULATE NUMBER AND OCCURRENCE FOR EACH ORBIT

CALCULATE NUMBER AND OCCURRENCE OF MISSION OUTCOME PER ORBIT TYPE

CREATE LANDING OUTCOME LABEL FROM OUTCOME COLUMN

EXPORT DATA TO CSV

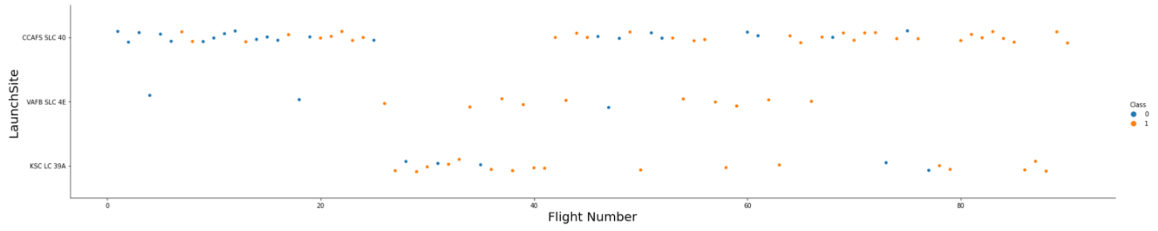


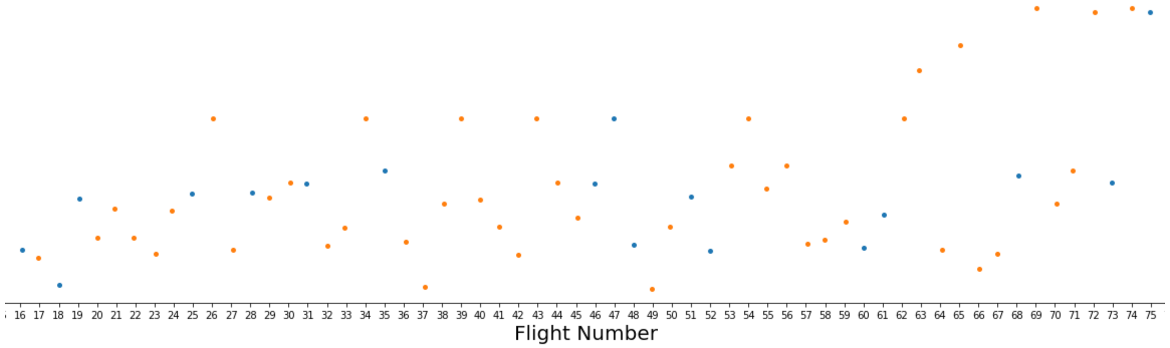
	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004

EDA with Data Visualisation

Scatter graphs drawn:

- Payload and Flight Number
- Flight Number and Launch Site
- Payload and Launch Site
- Flight Number and Orbit Type
- Payload and Orbit Type





EDA With SQL

The following SQL queries were performed:

- Names of the unique launch sites in the space mission;
- Top 5 launch sites whose name begin with the string 'CCA';
- Total payload mass carried by boosters launched by NASA (CRS);
- Average payload mass carried by booster version F9 v1.1;
- Date when the first successful landing outcome in ground pad was achieved;
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
- Total number of successful and failure mission outcomes;
- Names of the booster versions which have carried the maximum payload mass;
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

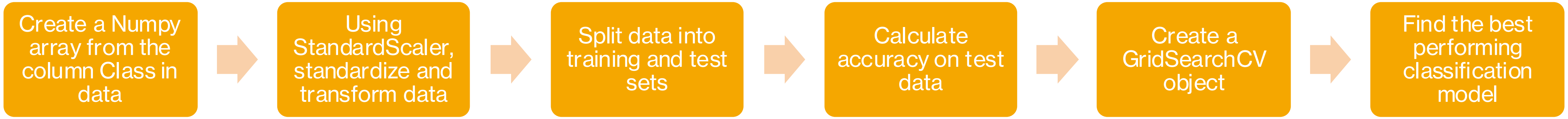
Build an Interactive with Folium

Added markers with circles, popup labels, and text labels for NASA Johnson Space Center and all launch sites using their latitude and longitude coordinates. Colored markers indicate the launch outcomes for each site: green for successful launches and red for failed launches. Colored lines show distances between launch sites and their proximities, such as railways, highways, coastlines, and closest cities.

Build a Dashboard with Plotly Dash

- Launch Site Selection: Added a dropdown list to select launch sites.
- Launch Success Visualization: Added a pie chart to display successful launch counts for all sites and a specific site.
- Payload Range Selection: Added a slider to select payload range.

Predictive Analysis (Classification)



Create a Numpy array from the column Class in data

Using StandardScaler, standardize and transform data

Split data into training and test sets

Calculate accuracy on test data

Create a GridSearchCV object

Find the best performing classification model

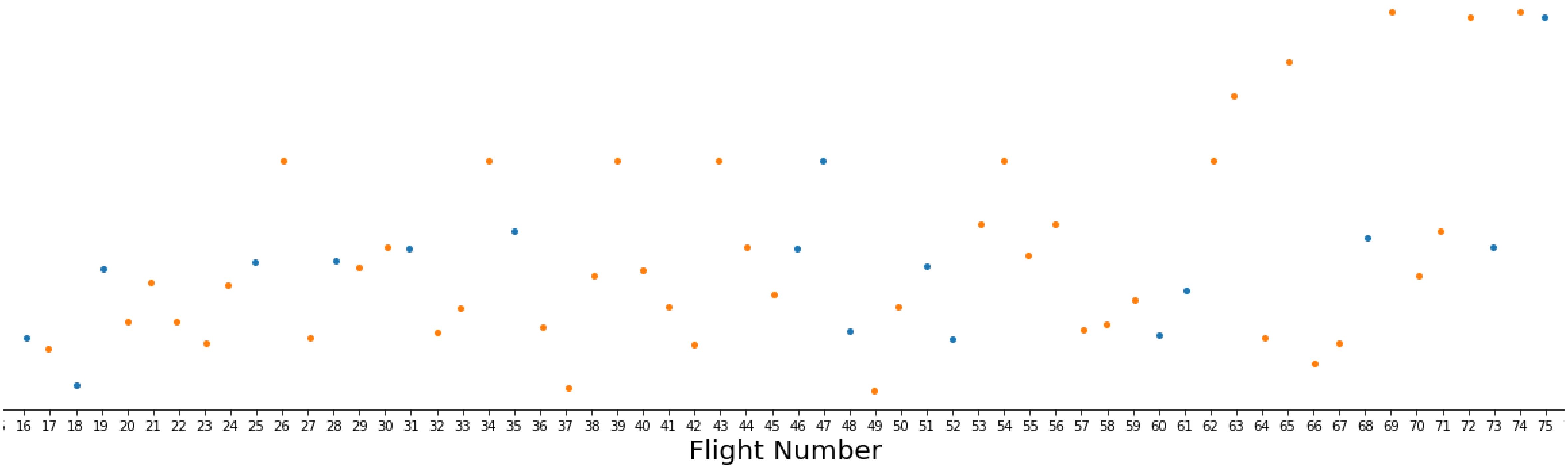
Results

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

EDA with Visualisation

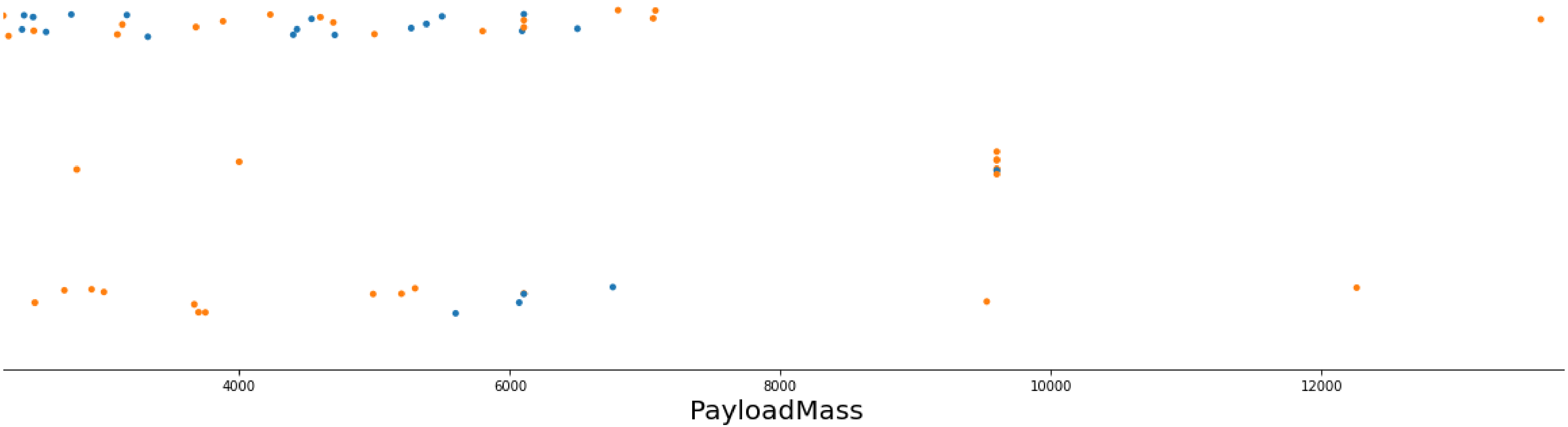
Flight Number vs Launch Site

- Early flights failed; later flights succeeded. Launch sites have varying success rates, with newer launches having higher success rates.



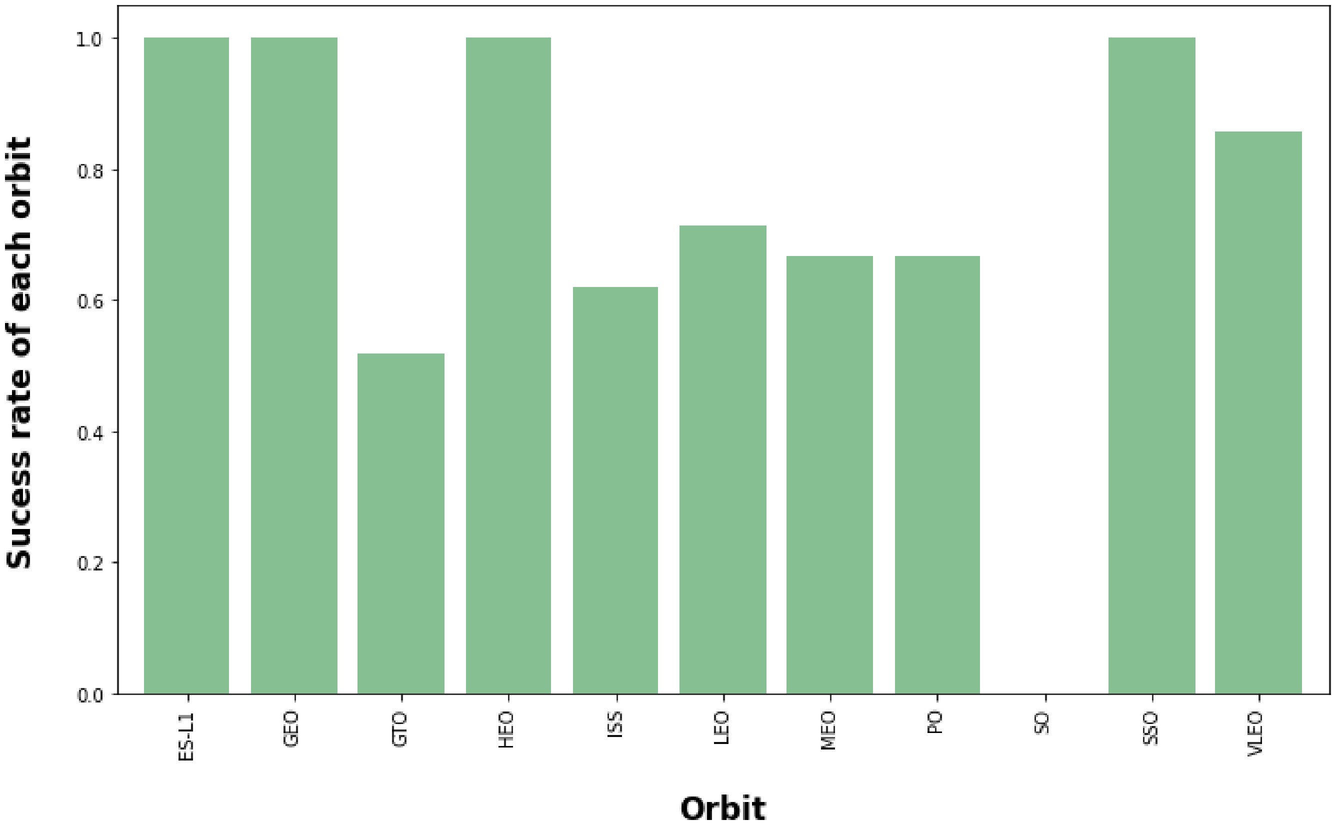
Payload vs Launch Site

- Higher payload mass correlates with higher launch success rates. KSC LC 39A boasts a 100% success rate for payloads under 5500 kg.

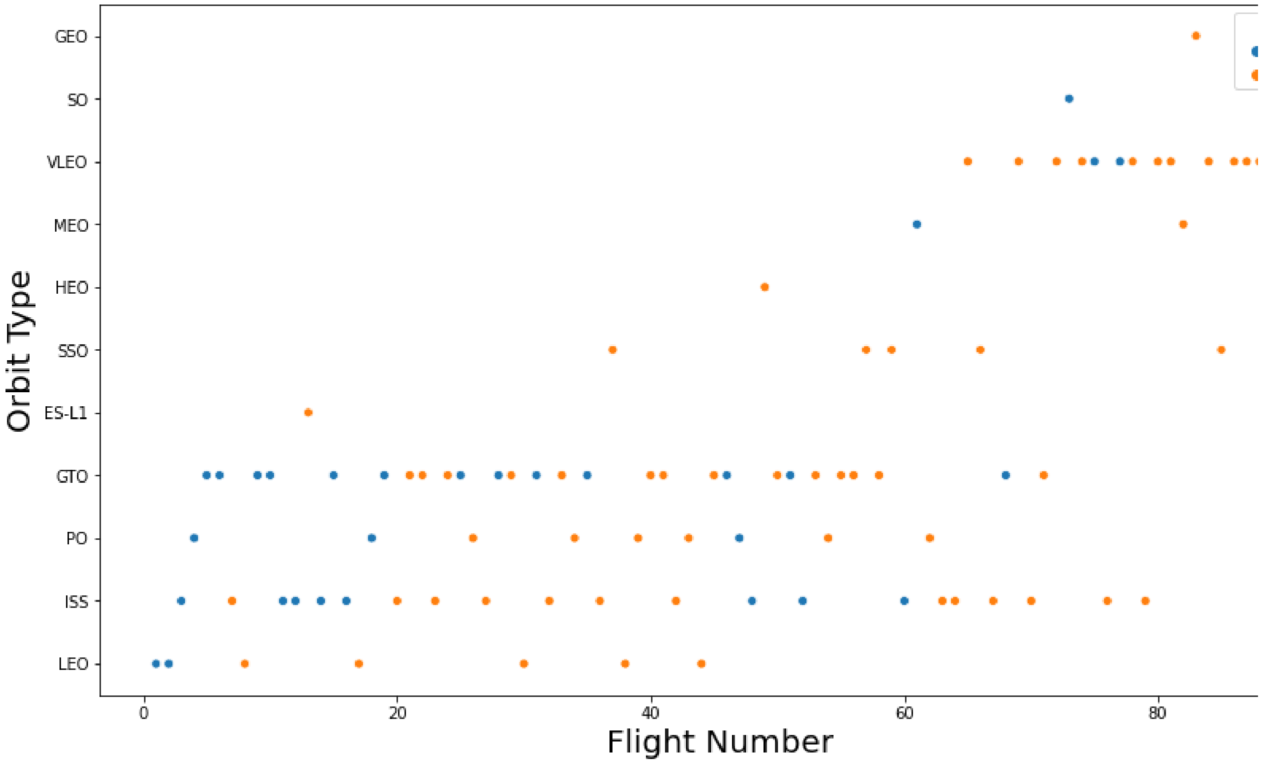


Success Rate vs Orbit Type

Orbits with 100% success rate include ES-L1, GEO, HEO, and SSO.

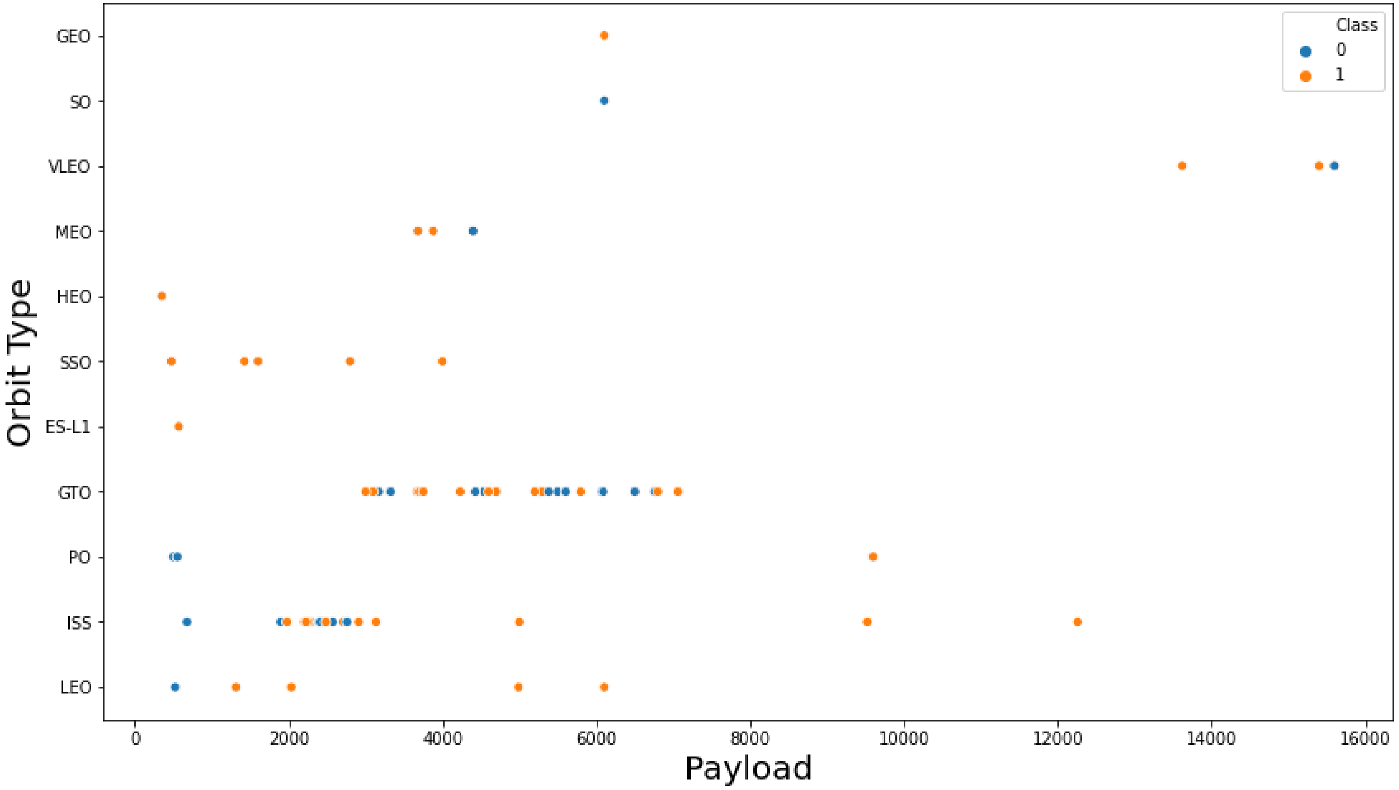


Flight Number vs Orbit Type



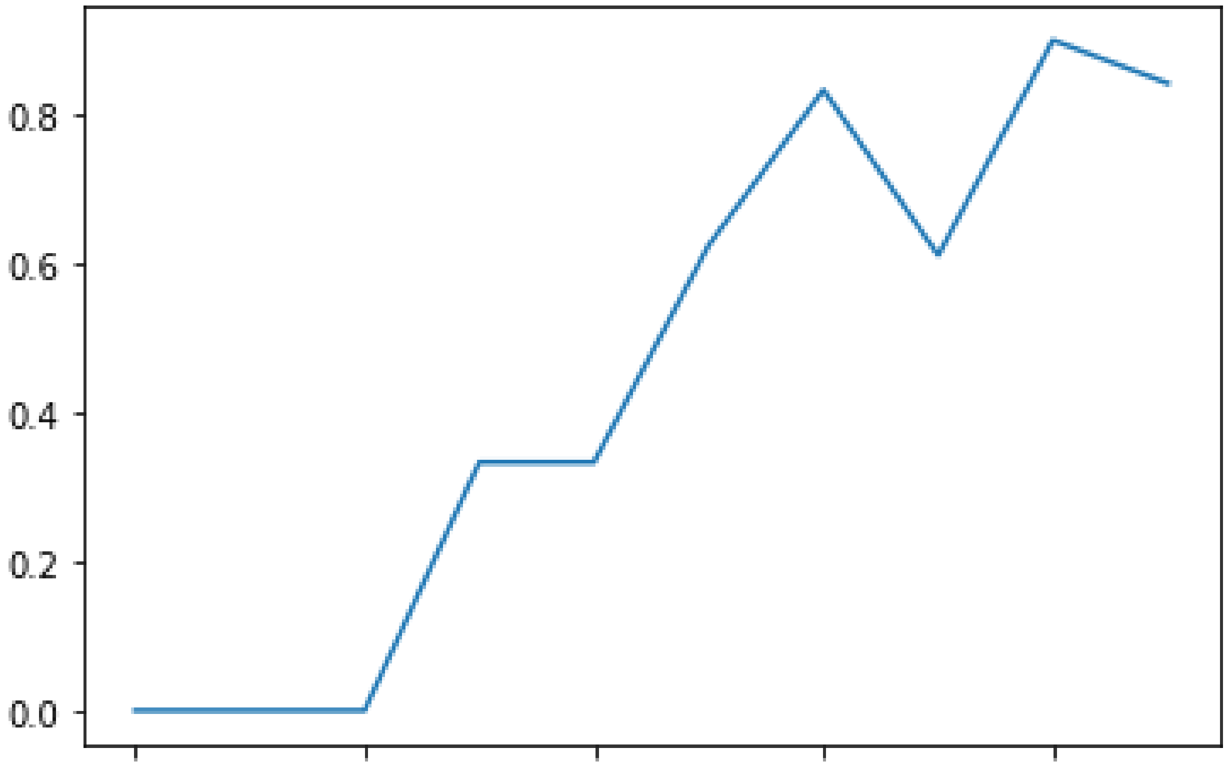
In LEO orbit, the Success is related to the number of flights; in GTO orbit, there's no such relationship.

Payload vs Orbit Type

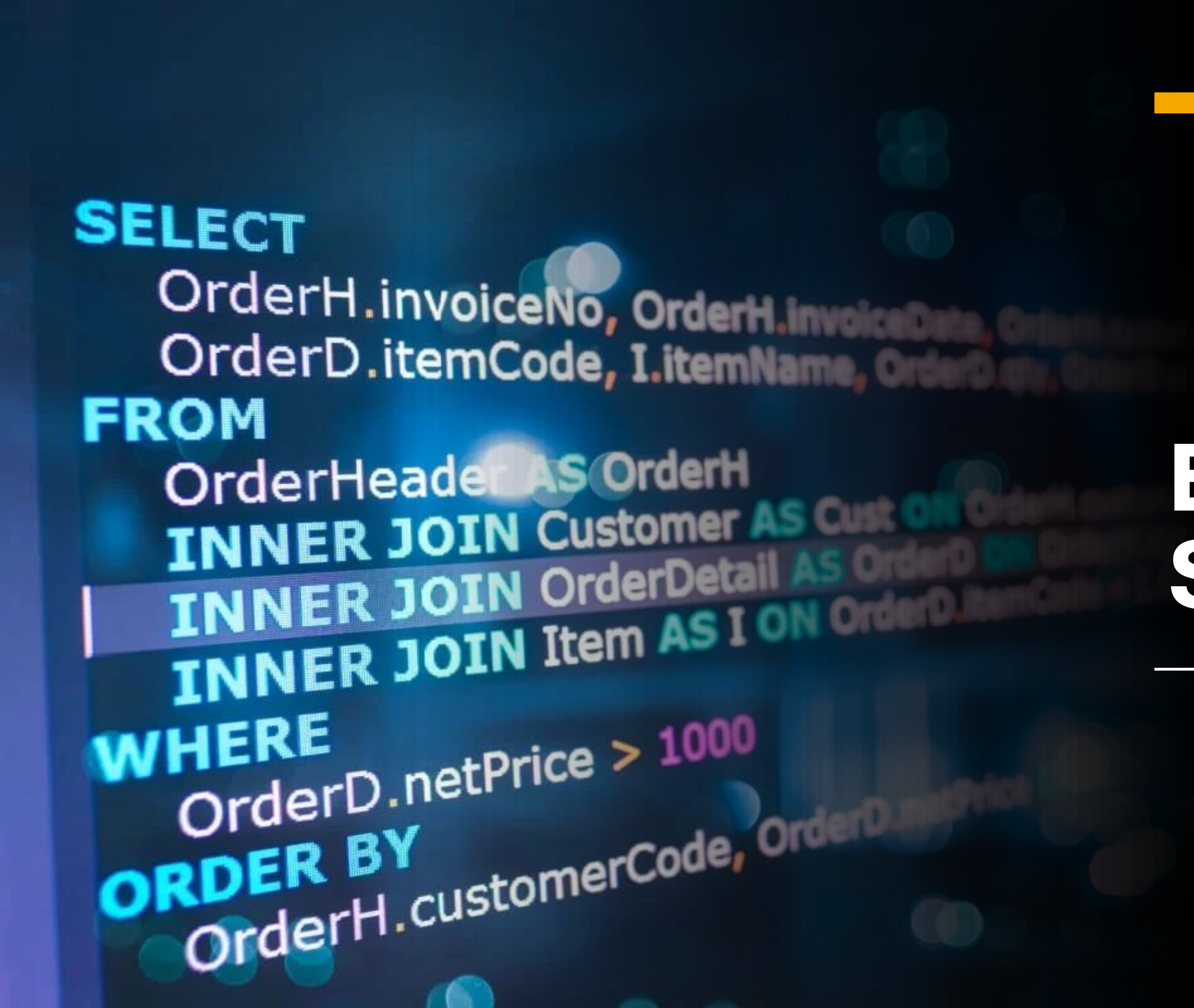


With heavy payloads, successful landing rates are higher for Polar, LEO, and ISS. However, for GTO, both positive landing rates and negative landing (unsuccessful missions) occur.

Launch Success Yearly Trend



The success rate has been increasing since 2013 until 2020.

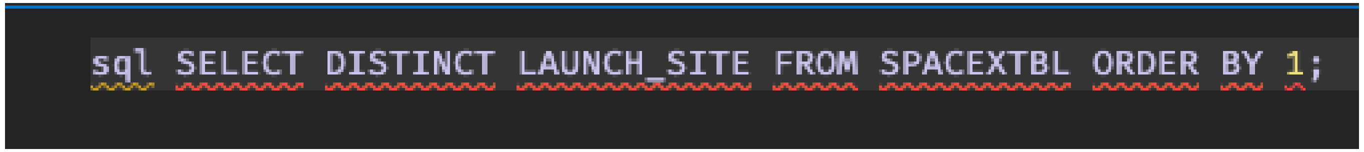


```
SELECT
    OrderH.invoiceNo, OrderH.invoiceDate,
    OrderD.itemCode, I.itemName, OrderD.qty,
FROM
    OrderHeader AS OrderH
    INNER JOIN Customer AS Cust ON
    INNER JOIN OrderDetail AS OrderD ON
    INNER JOIN Item AS I ON OrderD.itemCode =
WHERE
    OrderD.netPrice > 1000
ORDER BY
    OrderH.customerCode, OrderD.netPrice
```

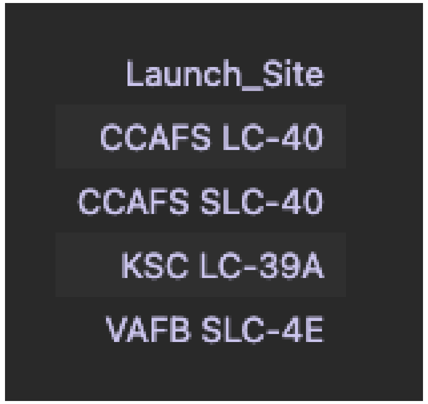
EDA WITH SQL

All Launch Site Names

Fetches distinct LAUNCH_SITE from SPACEXTBL table



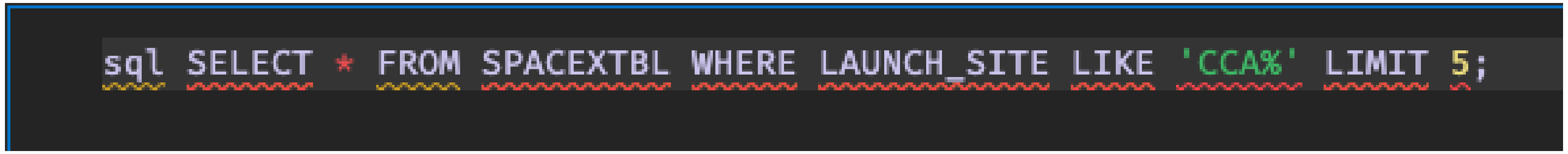
```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```



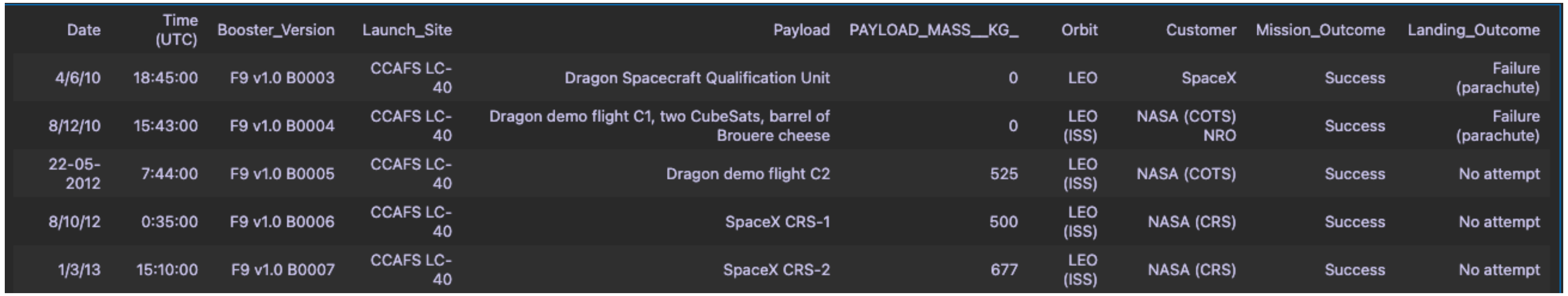
```
 Launch_Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E
```

Launch Site Names begin with ‘CCA’

Select 5 records from SPACEXTBL where LAUNCH_SITE begins with ‘CCA’



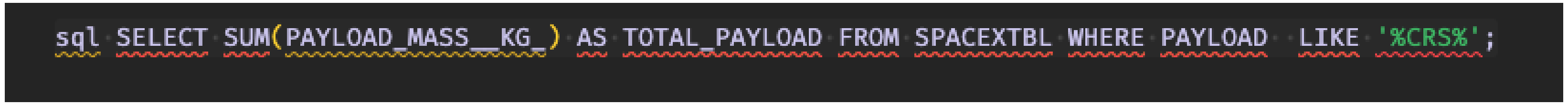
```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```



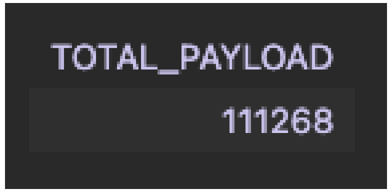
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
4/6/10	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
8/12/10	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
8/10/12	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
1/3/13	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Calculates the total payload mass for SpaceX missions containing ‘CRS’ in their payload name.



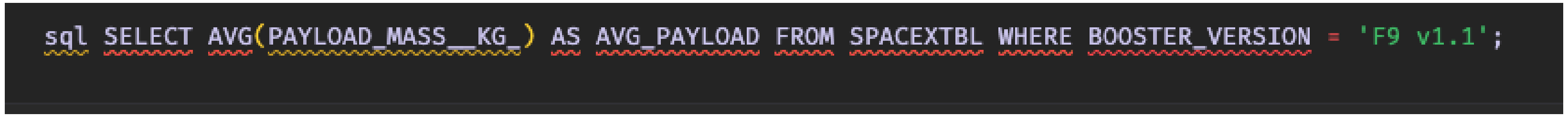
```
sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';
```



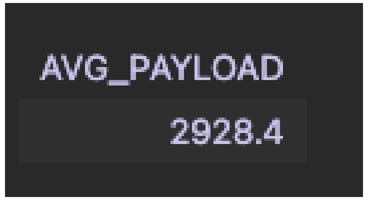
```
TOTAL_PAYLOAD
       111268
```

Average Payload Mass by F9 v1.1

Calculates average payload mass where booster version is “F9 v1.1”



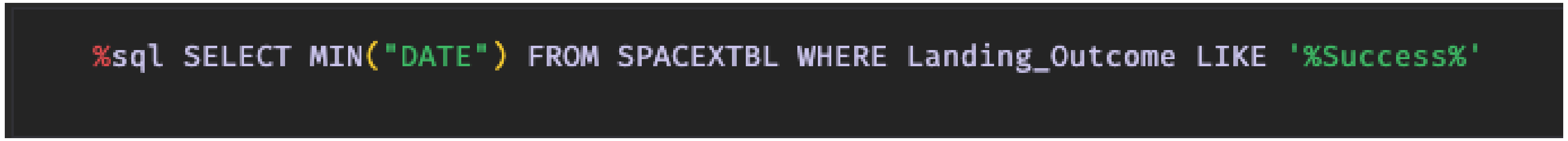
```
sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```



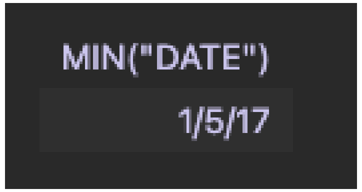
```
AVG_PAYLOAD
     2928.4
```

First Successful Ground Landing Date

Uses the MIN function to calculate the first successful landing outcome



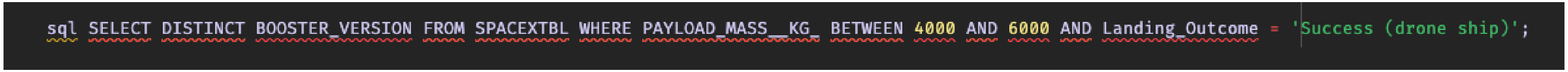
```
%sql SELECT MIN("DATE") FROM SPACEXTBL WHERE Landing_Outcome LIKE '%Success%'
```



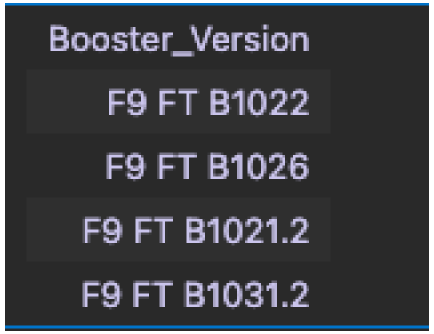
```
MIN("DATE")
     1/5/17
```

Successful Drop Ship Landing with Payload between 4000 and 6000

Returns the booster version of successful drone ship where the payload mass is between 4000 and 6000



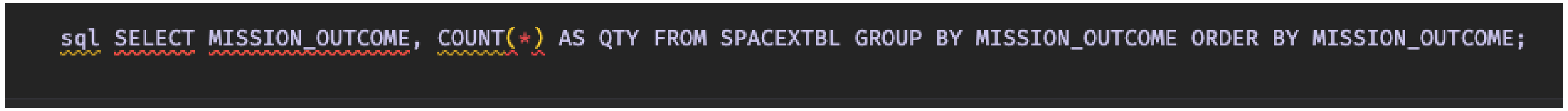
```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND Landing_Outcome = 'Success (drone ship)';
```



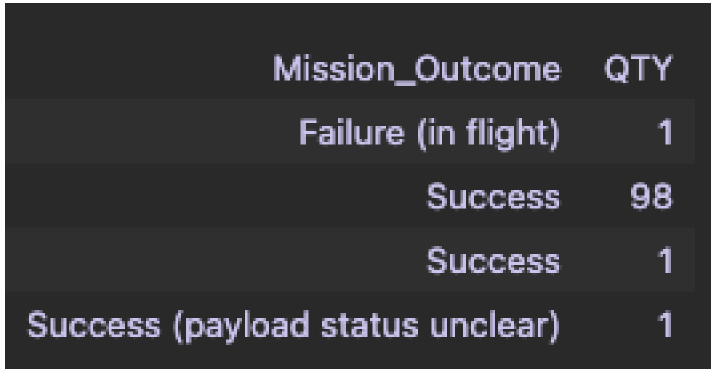
Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

Returns the total number of successful and failure mission outcomes



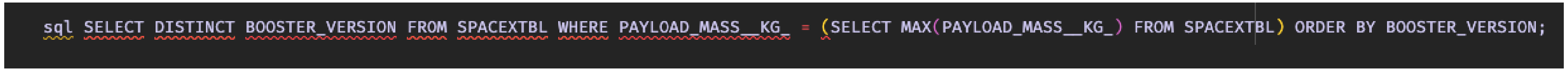
```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```



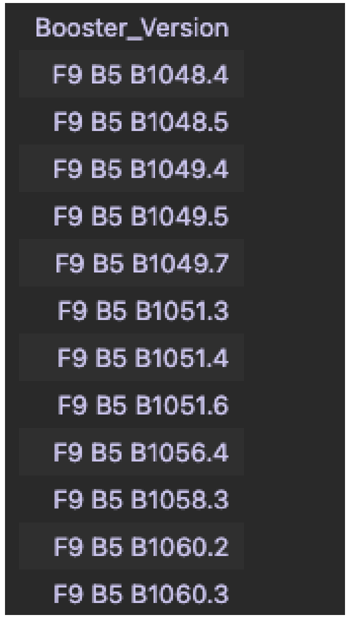
Mission_Outcome	QTY
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Returns the names of booster version that have carried the maximum payload mass



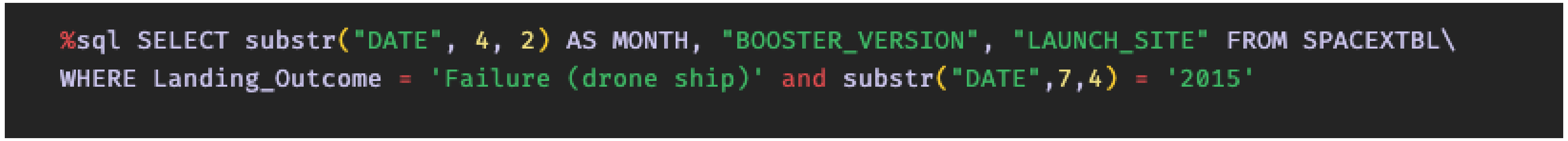
```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;
```



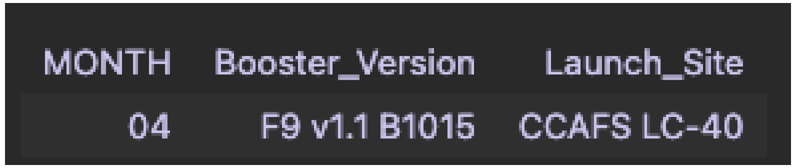
Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

Returns failed drone ship in the year 2015



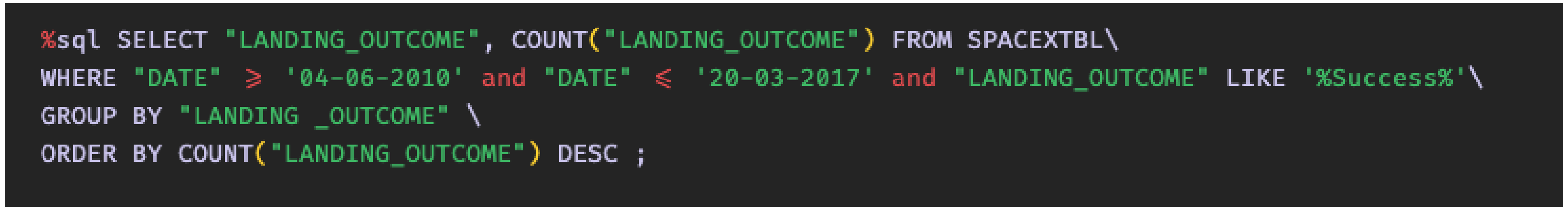
```
%sql SELECT substr("DATE", 4, 2) AS MONTH, "BOOSTER_VERSION", "LAUNCH_SITE" FROM SPACEXTBL\
WHERE Landing_Outcome = 'Failure (drone ship)' and substr("DATE",7,4) = '2015'
```



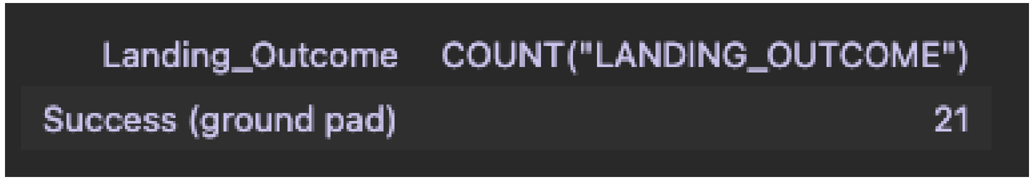
MONTH	Booster_Version	Launch_Site
04	F9 v1.1 B1015	CCAFS LC-40

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

Rank the count of landing outcomes between 2010-06-04 and 2017-03-20



```
%sql SELECT "LANDING_OUTCOME", COUNT("LANDING_OUTCOME") FROM SPACEXTBL\
WHERE "DATE" ⩾ '04-06-2010' and "DATE" ⩽ '20-03-2017' and "LANDING_OUTCOME" LIKE '%Success%'\
GROUP BY "LANDING _OUTCOME" \
ORDER BY COUNT("LANDING_OUTCOME") DESC ;
```

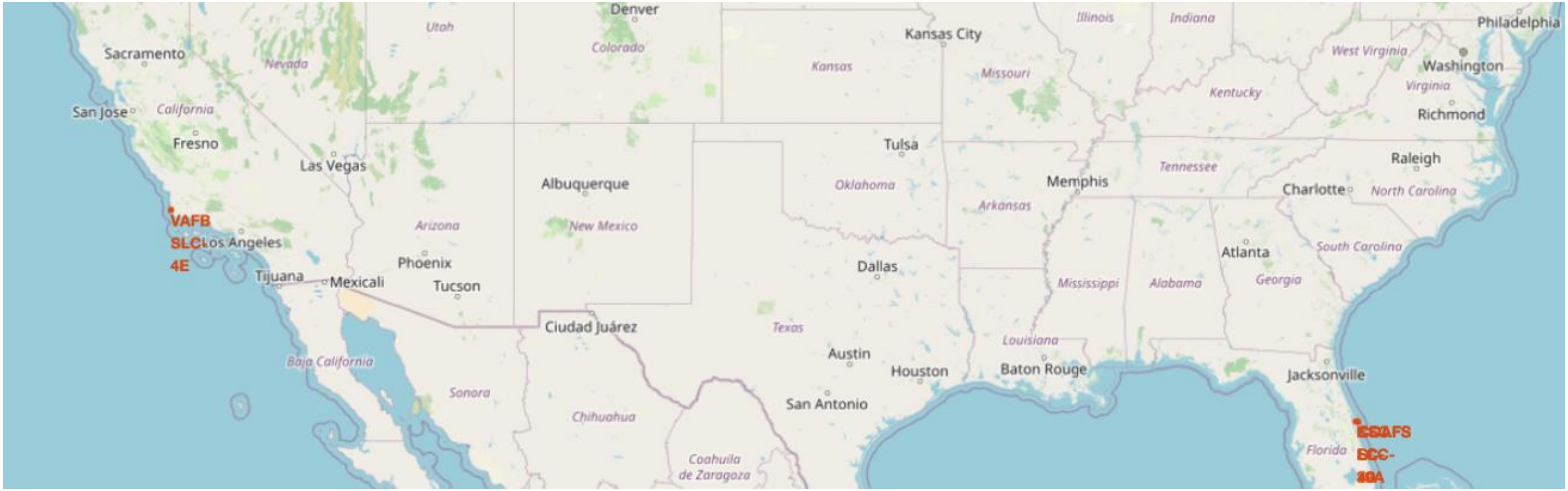


Landing_Outcome	COUNT("LANDING_OUTCOME")
Success (ground pad)	21

Interactive Map with Folium

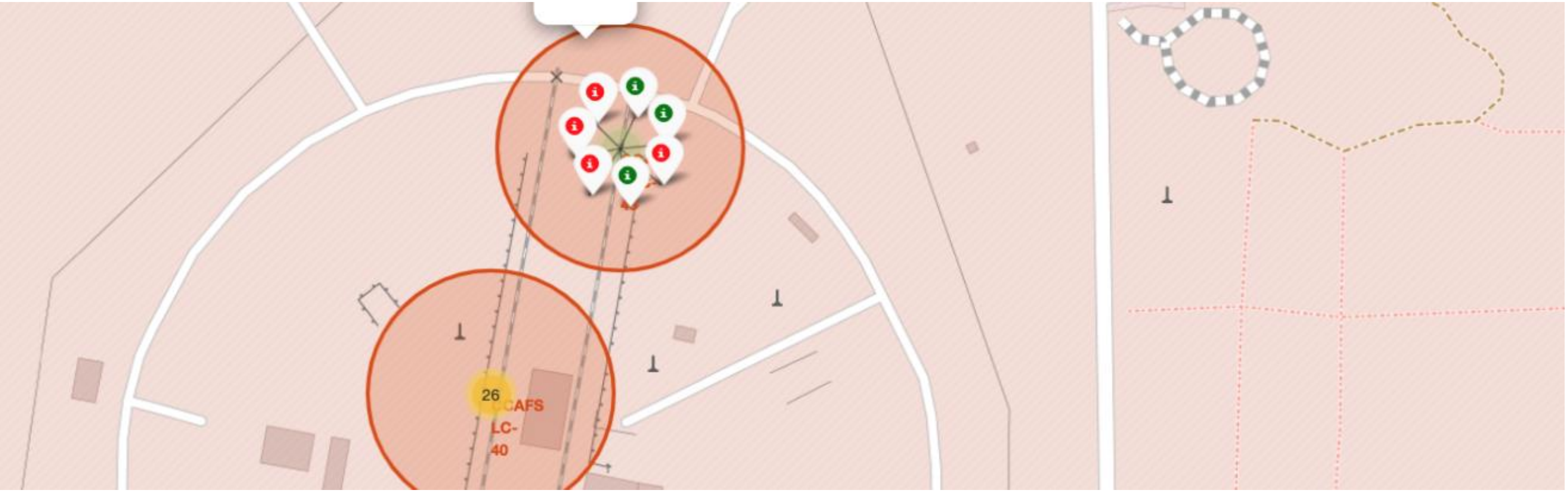
All Launch Sites on Folium Map

Most launch sites are near the equator, where the land moves fastest. Anything on the equator’s surface moves at 1670 km/h. When a ship launches from the equator, it goes into space and continues moving at the same speed. This speed helps the spacecraft stay in orbit due to inertia. Launch sites are also near the coast to minimize the risk of debris or explosions near people when launching rockets towards the ocean.



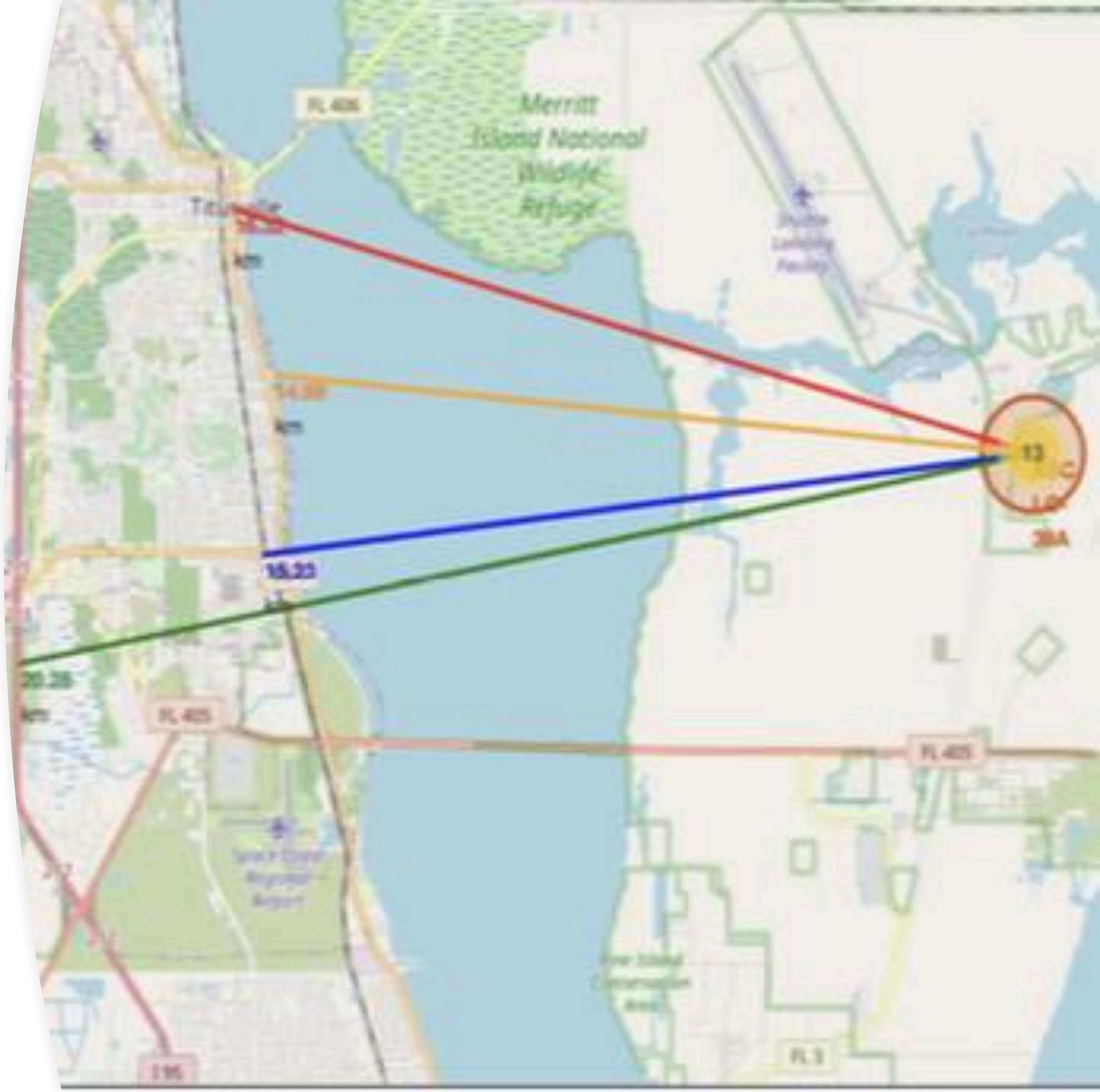
Color labeled launch records

We can easily identify which launch sites have high success rates by using the color-labeled markers: green for successful launches and red for failed launches. Launch Site KSC LC-39A stands out with an exceptionally high success rate.



Distance from the launch site KSC LC-39A to its proximities

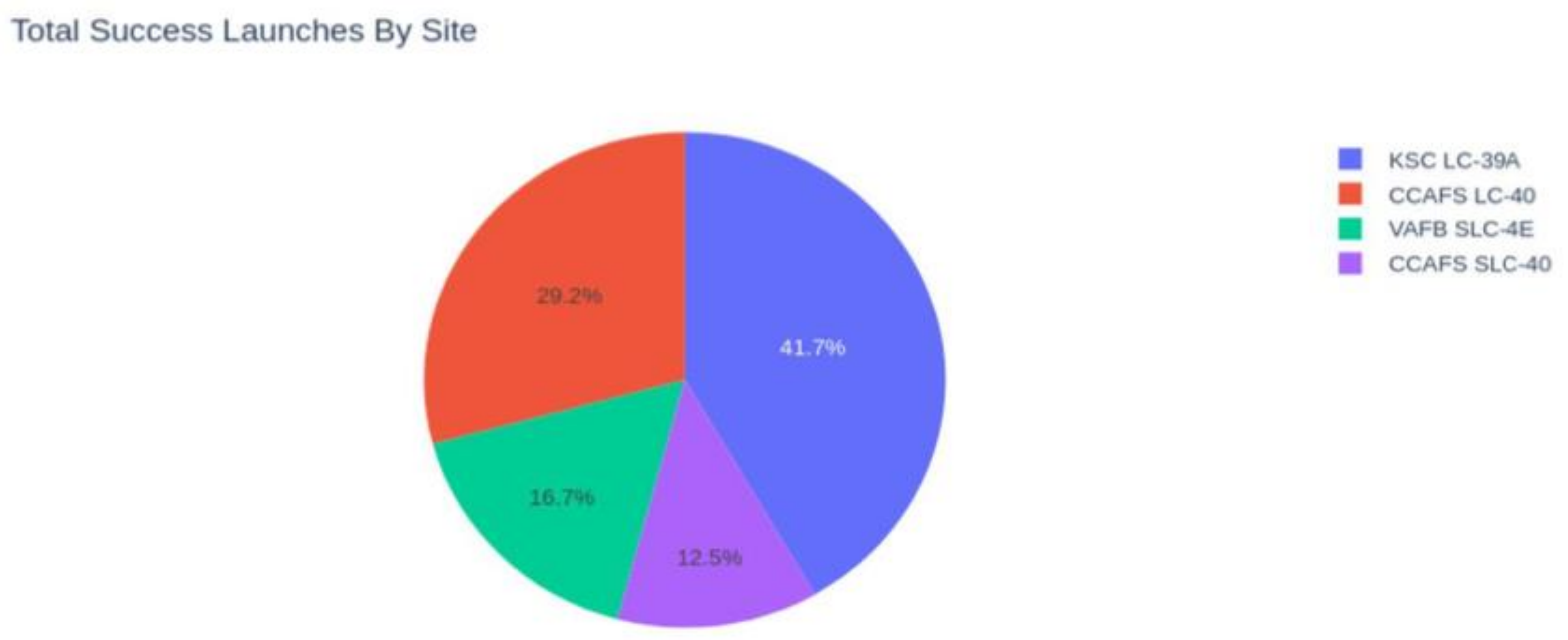
- The launch site KSC LC-39A is close to several major infrastructure:
  - Railway (15.23 km)
  - Highway (20.28 km)
  - Coastline (14.99 km)
  - Its closest city, Titusville (16.32 km)
- A high-speed failed rocket can cover distances of 15-20 km in a few seconds, posing a potential danger to populated areas.



Build a Dashboard with Plotly Dash

Launch success count for all sites

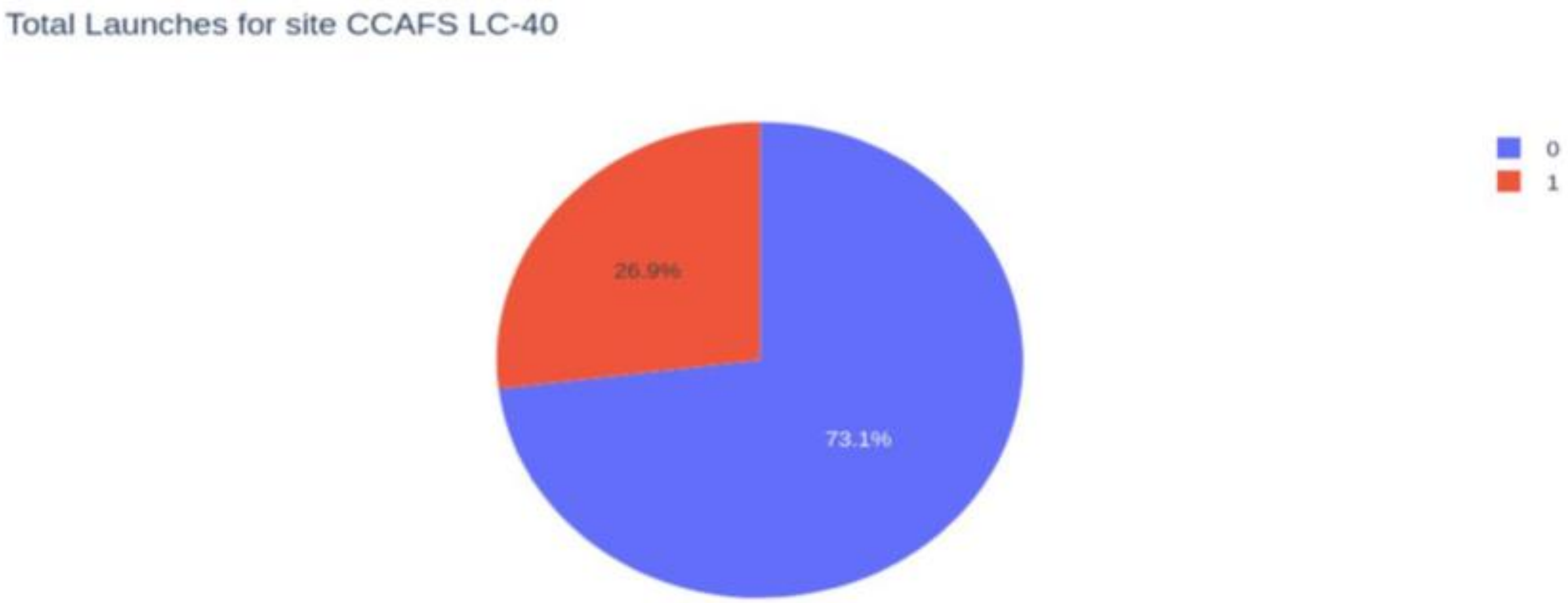
KSC LC-39A has the most successful launches among all sites.



Total Success Launches By Site

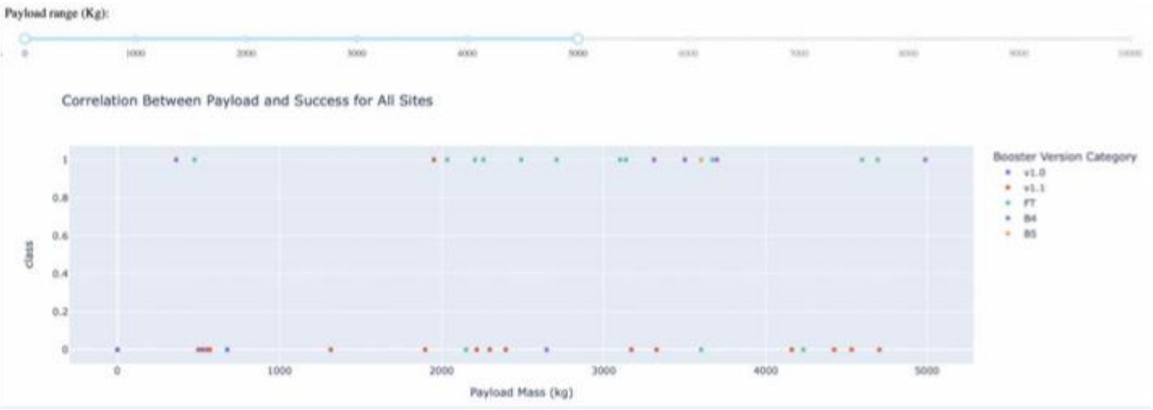
Launch site with highest launch success ratio

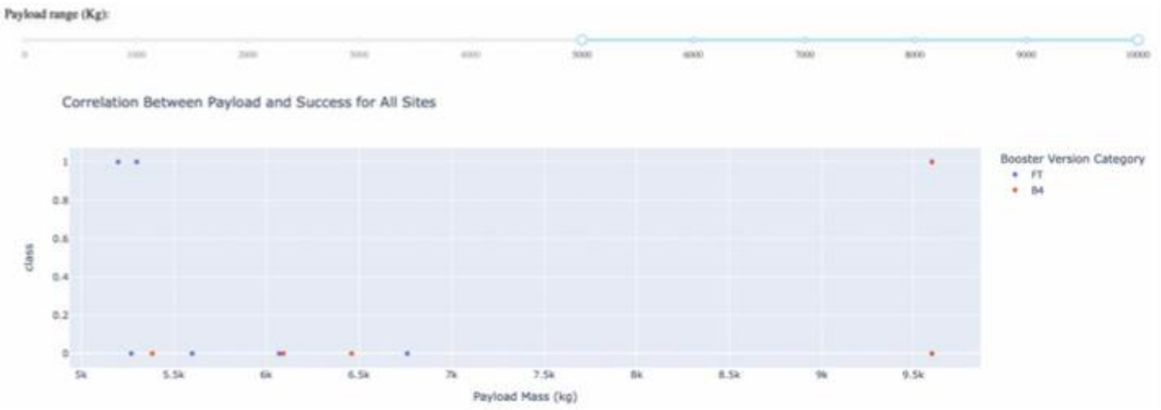
KSC LC-39A has the highest launch success rate (73.1%) with 10 successful and 3 failed landings.



Payload Mass vs Launch Outcome for all sites

Payloads between 2000 and 5500 kg have the highest success rate.

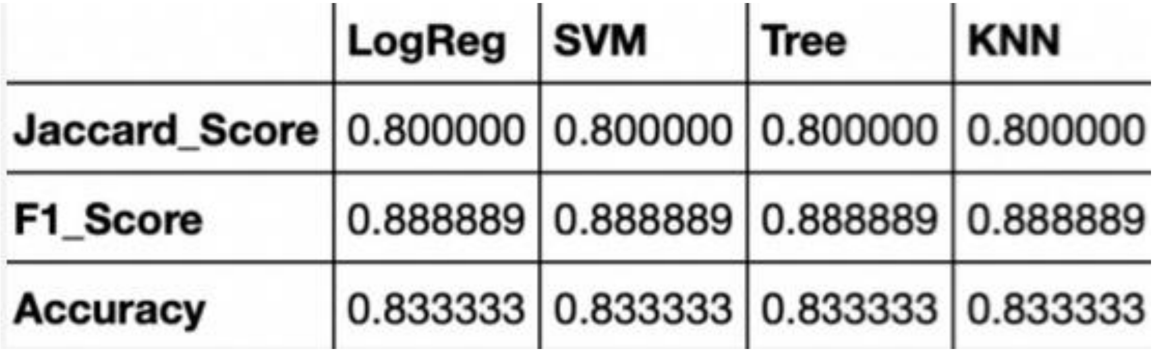




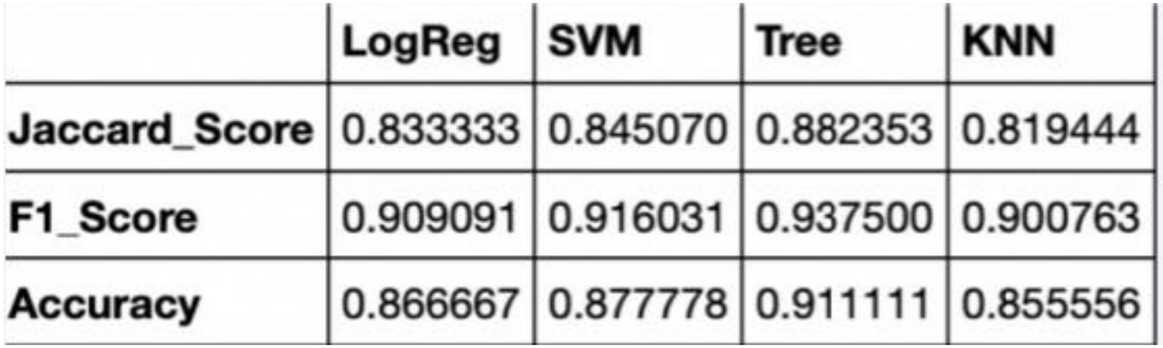
Predictive Analysis (Classification)

Classification Accuracy

We can’t confirm which method performs best based on the Test Set scores due to the small test sample size (18 samples). Therefore, we tested all methods on the whole Dataset. The Decision Tree Model has the highest scores and accuracy.



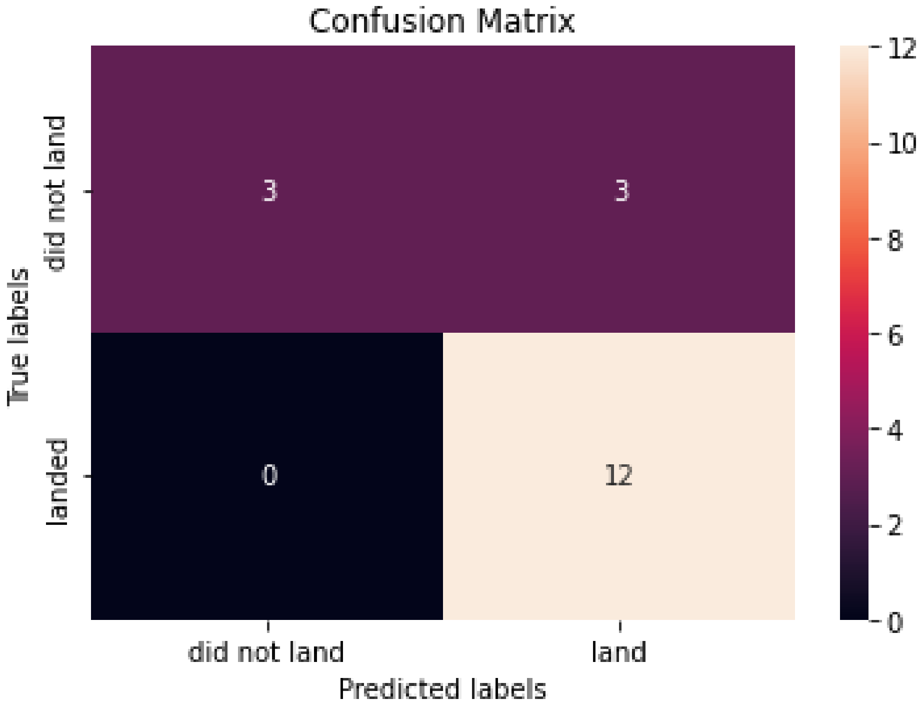
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

Logistic regression can distinguish between classes, but it faces a major problem with false positives.



Conclusion

Decision Tree Model is the best algorithm for this dataset. Low payload mass launches show better results than larger payload mass launches. Most launch sites are near the Equator and the coast. Launch success rates increase over time. KSC LC-39A has the highest success rate. ES-L1, GEO, HEO, and SSO orbits have a 100% success rate.

Thank you!