IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

L Stuurman

01-2026

Outline

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

Executive Summary

- ➢Summary of methodology
- ➢Data collection
- ➢Data wrangling
- ➢Exploratory Data Analysis with Data Visualisation and SQL
- ➢Building an interactive map with Folium and a Dashboard with Plotly
- ➢Predictive analysis (Classification)
- ➢Summary of results
- ➢Exploratory Data Analysis results
- ➢Interactive analytics visuals
- ➢Predictive analysis results

Introduction

- Background

SpaceX is a successful space exploration company that is able to continuously top the market with their cost conscious approach. This is why we feel instead of using rocket science to determine if the first stage will land, we can use machine learning models and public information, by learning patterns in data to determine the cost of each launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Using SpaceX rest API
- Perform data wrangling
    - Filtering data
    - Replacement/removal of missing values
    - Binary classification of data using hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Building, tuning and evaluating classification of models, ensuring the best result

Data Collection

- Data collection was done using the SpaceX Rest API as well as Web Scrapping data from tables on the SpaceX Wikipedia page.

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
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 – SpaceX API

https://github.com/Lee-AnnS/DataScienceEcosystem/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

- Requesting Rocket launch data from SpaceX API
- Decoding the response content with .json() and turning it into a data frame using .json_normalize()
- Requesting needed information about the launches by applying custom functions
- Constructing data we have obtained into a dictonary and creating a data frame from it
- Filtering the data frame for only falcon 9 launches
- Replacing missing values in Payload Mass column with ,mean()
- Exporting the data to csv

Data Collection - Scraping

https://github.com/Lee-AnnS/DataScienceEcosystem/blob/main/jupyter-labs-webscraping.ipynb

- Requested falcon 9 launches from Wikipedia
- Created BeautifulSoup object from HTML response
- Extracted all the column names from HTML table header
- Collected the data by parsing HTML tables
- Constructed data obtained into a dictionary and created a data frame from the dictionary
- Exported the data to csv

Data Wrangling

https://github.com/Lee-AnnS/DataScienceEcosystem/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship. We simply converted these outcomes into Training labels with 1 meaning successfully landed and 0 meaning unsuccessful.

EDA with Data Visualization

- https://github.com/Lee-AnnS/DataScienceEcosystem/blob/main/edadataviz.ipynb
- The following plots can be visualised:
- 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 Success Rate Yearly Trend Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model. Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value. Line charts show trends in data over time (time series).

EDA with SQL

- https://github.com/Lee-AnnS/DataScienceEcosystem/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb
- Displayed the names of the unique launch sites in the space mission
- Displayed 5 records where launch sites begin with the string 'CCA’
- Displayed the total payload mass carried by boosters launched by NASA (CRS)
- Displayed the average payload mass carried by booster version F9 v1.1
- Listed the date when the first succesful landing outcome in ground pad was acheived.
- Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listed the total number of successful and failure mission outcomes
- Listed all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.
- Listed the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Ranked the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

Build an Interactive Map with Folium

https://github.com/Lee-AnnS/DataScienceEcosystem/blob/main/lab_jupyter_launch_site_location.ipynb

Markers of all Launch Sites:

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location. Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Coloured Markers of the launch outcomes for each Launch Site:

- Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches:

- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range:

- Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

- Added a scatter chart to show the correlation between Payload and Launch Success

Predictive Analysis (Classification)

https://github.com/LeeAnnS/DataScienceEcosystem/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

- Creating a NumPy array from the column "Class" in data
- Standardizing the data with StandardScaler, then fitting and transforming it
- Splitting the data into training and testing sets with train_test_split function
- Creating a GridSearchCV object with cv = 10 to find the best parameters
- Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models
- Calculating the accuracy on the test data using the method score() for all models
- Examining the confusion matrix for all models
- Finding the method performs best by examining the Jaccard_score and FI_score metrics

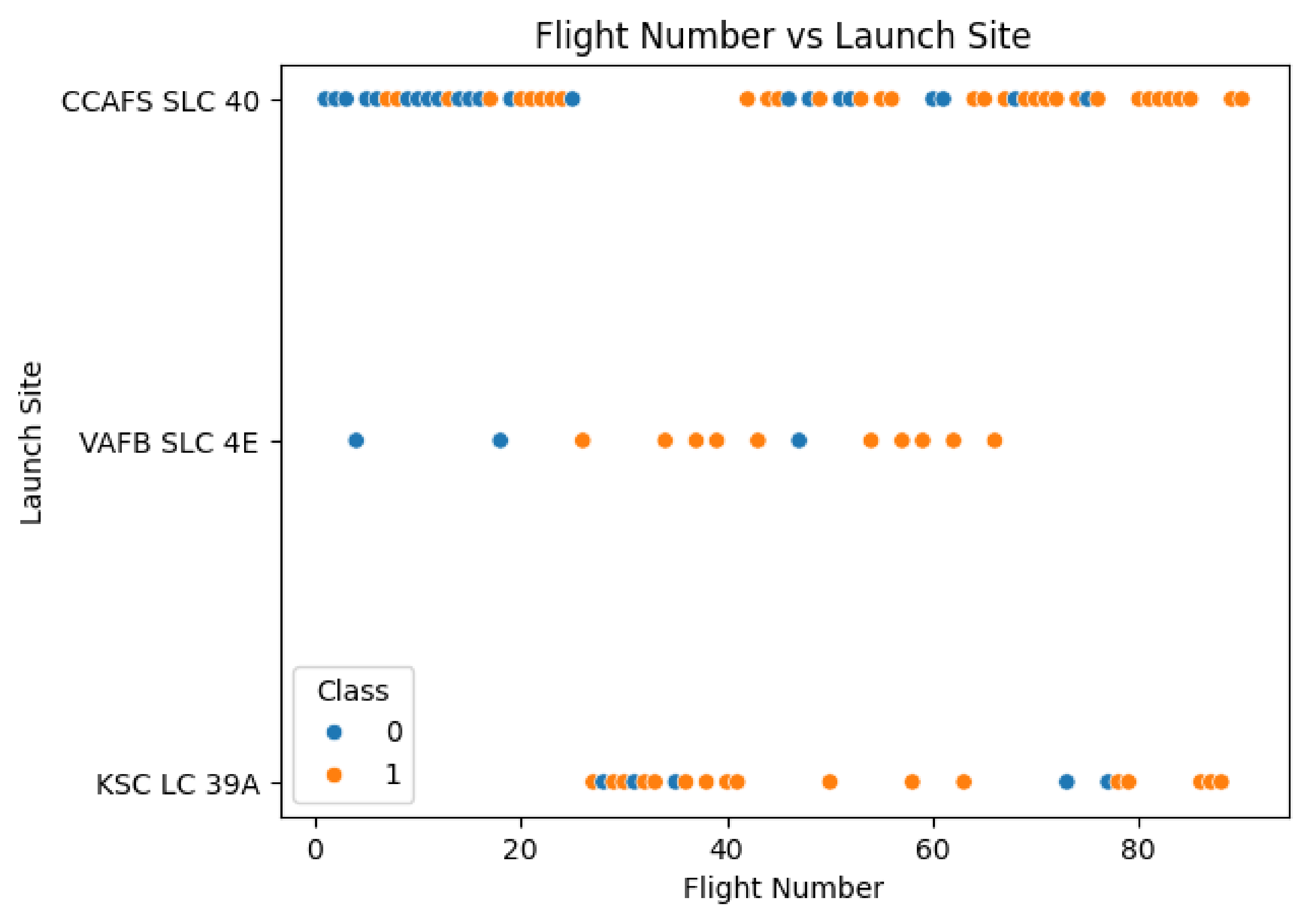
Results

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

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

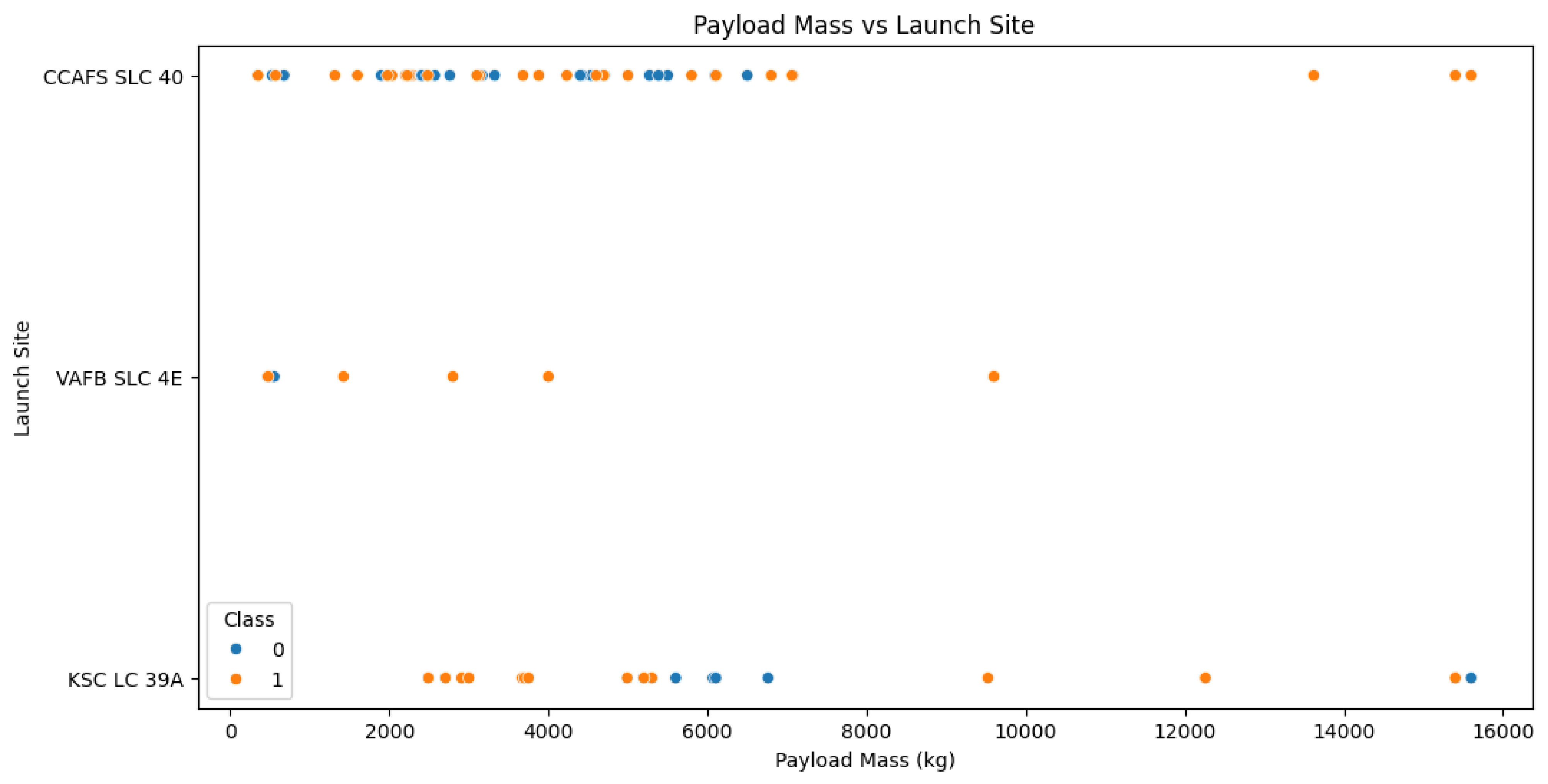


The earliest flights all failed while the latest flights all succeeded.

The CCAFS SLC 40 has majority of launches

The VAFB SLC 40 and KSC LC 39A have higher success rates

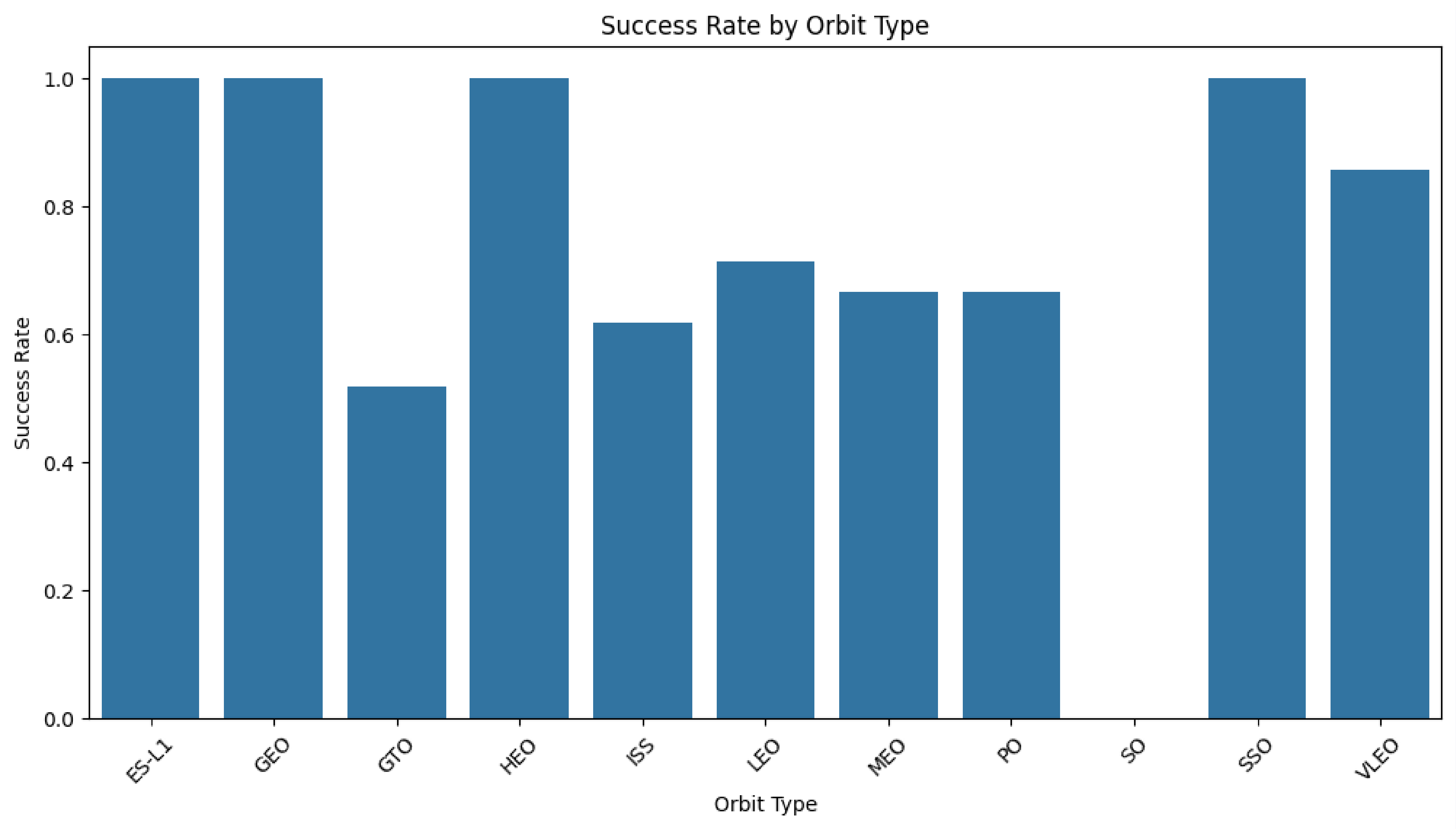
Payload vs. Launch Site



The higher the Payload Mass the higher the success rate

All launches with a payload over 8000kg were successful

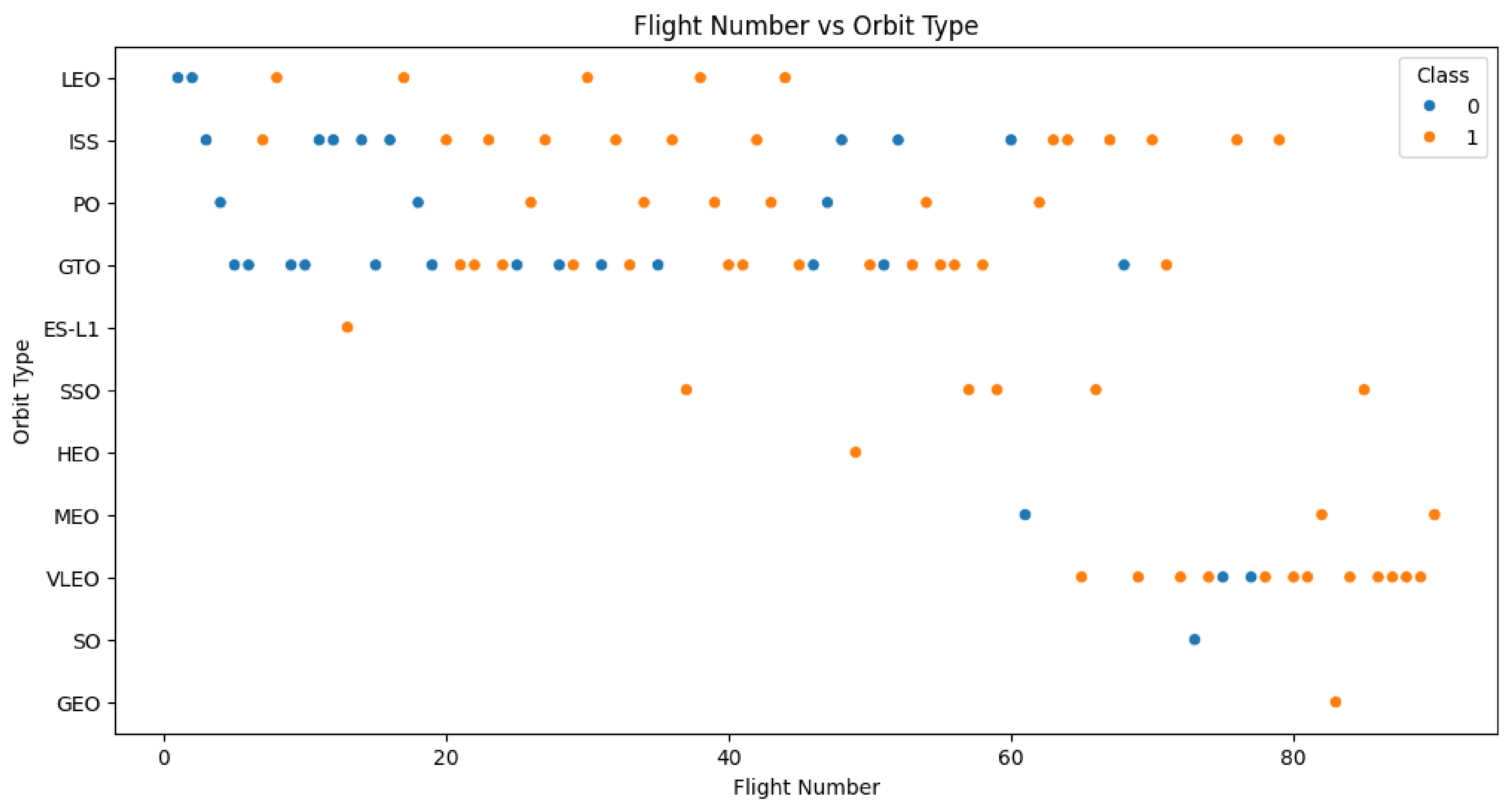
Success Rate vs. Orbit Type



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

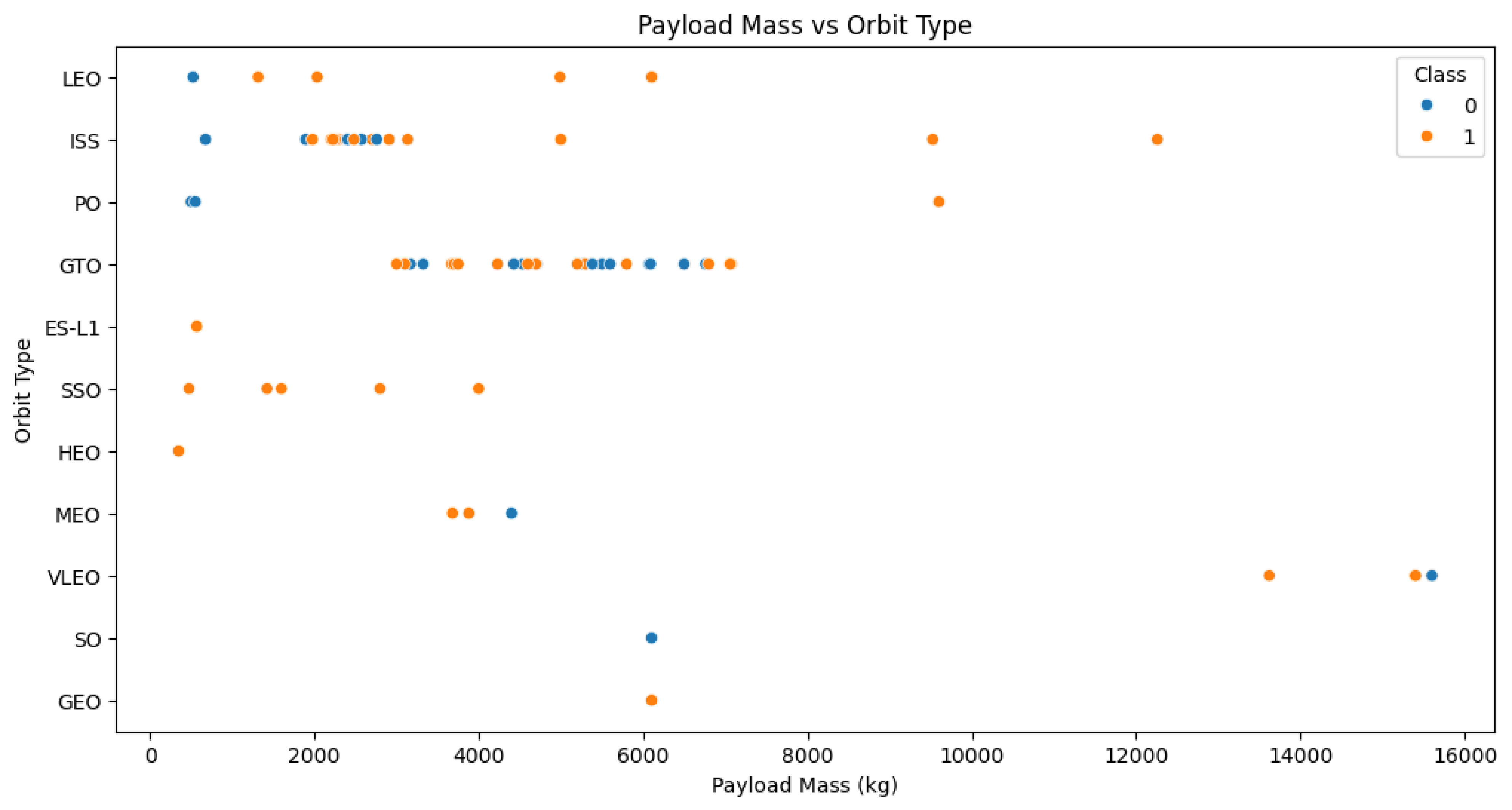
SO had a 0% success rate

Flight Number vs. Orbit Type



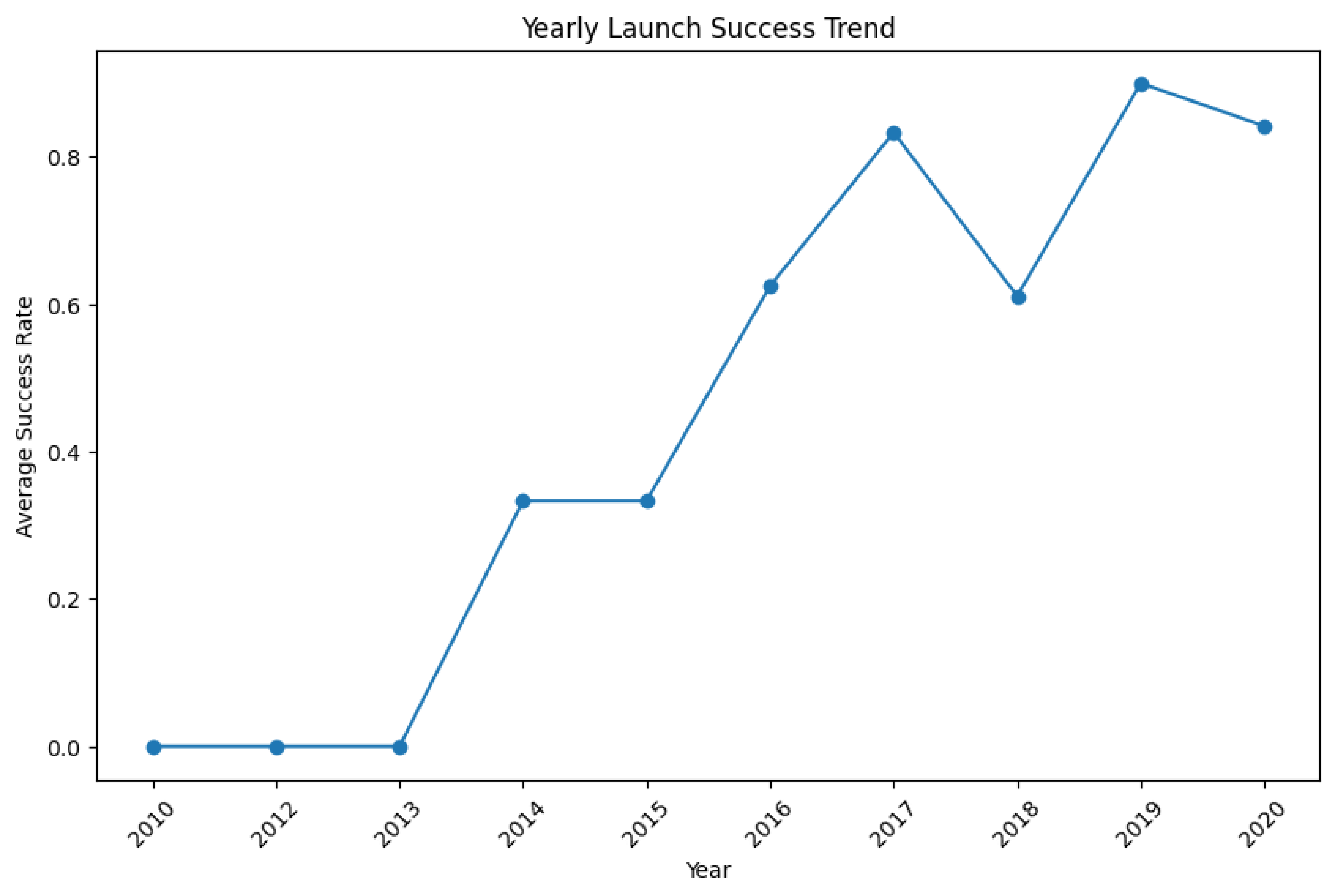
In the LEO orbit the success appears to be related to the number of flights

Payload vs. Orbit Type



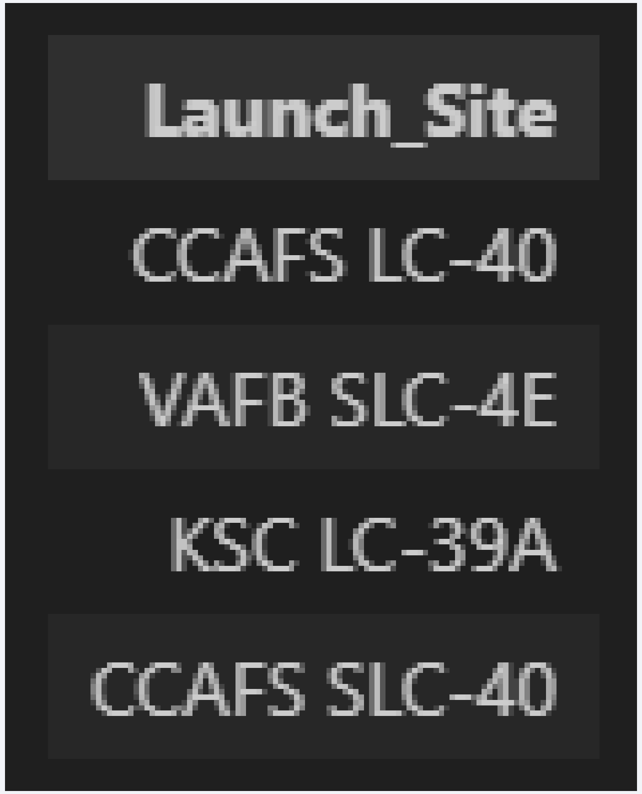
Heavy payloads have a negative influence on GTO orbits and positive on ISS and LEO orbits

Launch Success Yearly Trend



The has been a launch success rate increase since 2013, which dropped in 2020

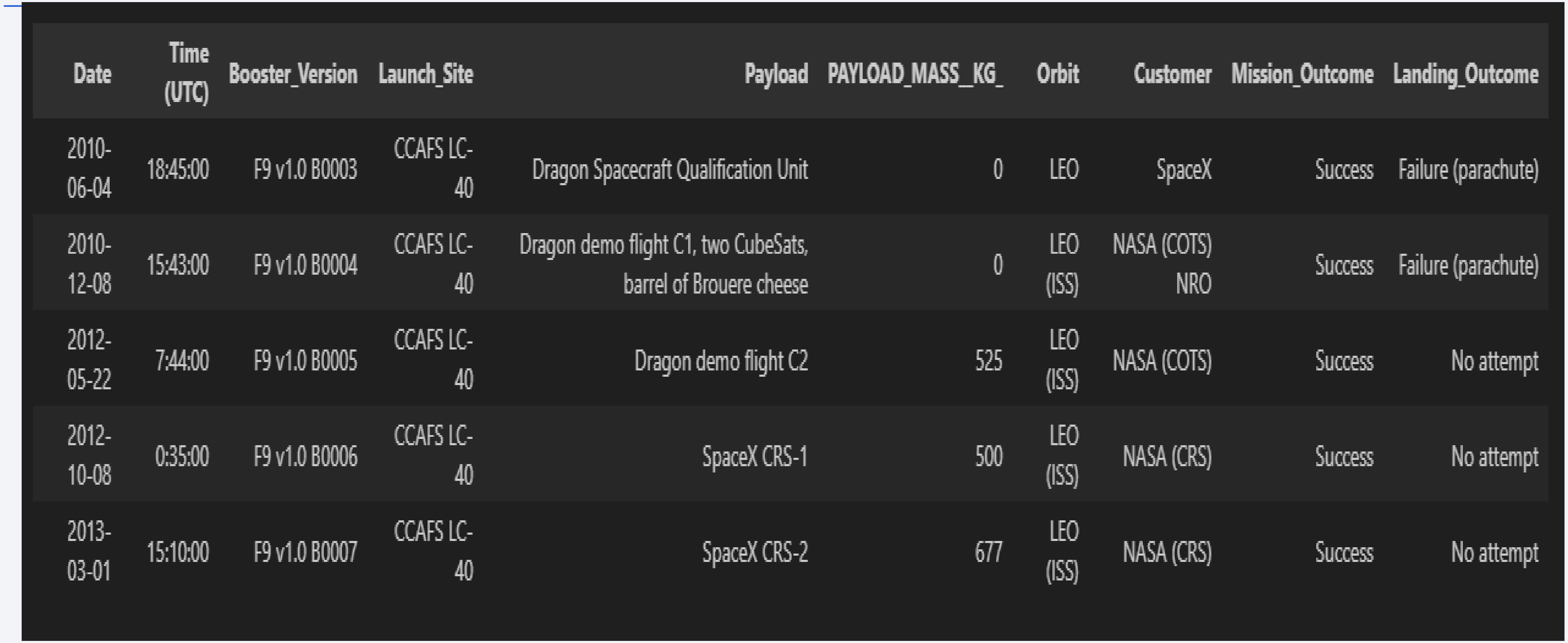
All Launch Site Names



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

Here are the names of the unique launch sites displayed

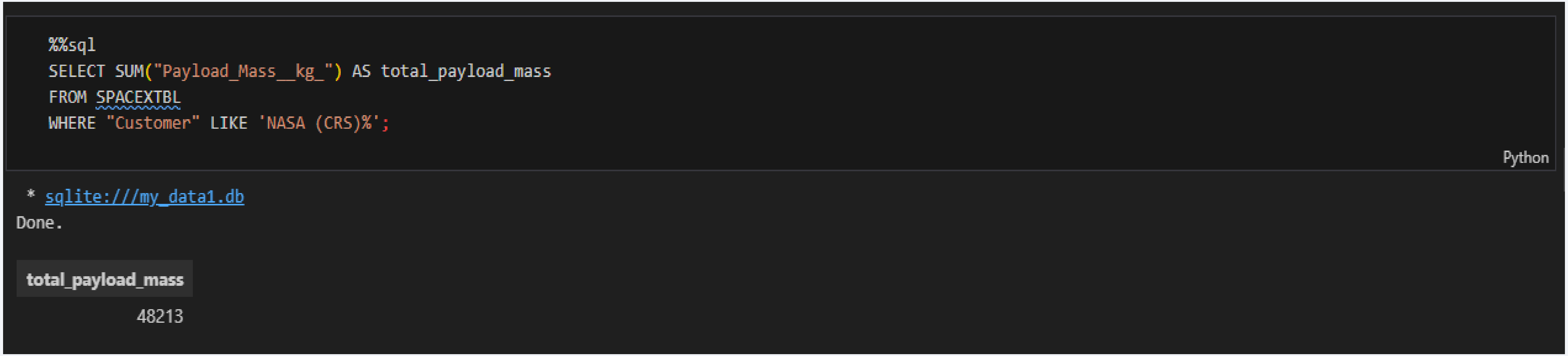
Launch Site Names Begin with 'CCA'



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

Displaying the 5 records where launch sites begin with the string ‘CCA’

Total Payload Mass



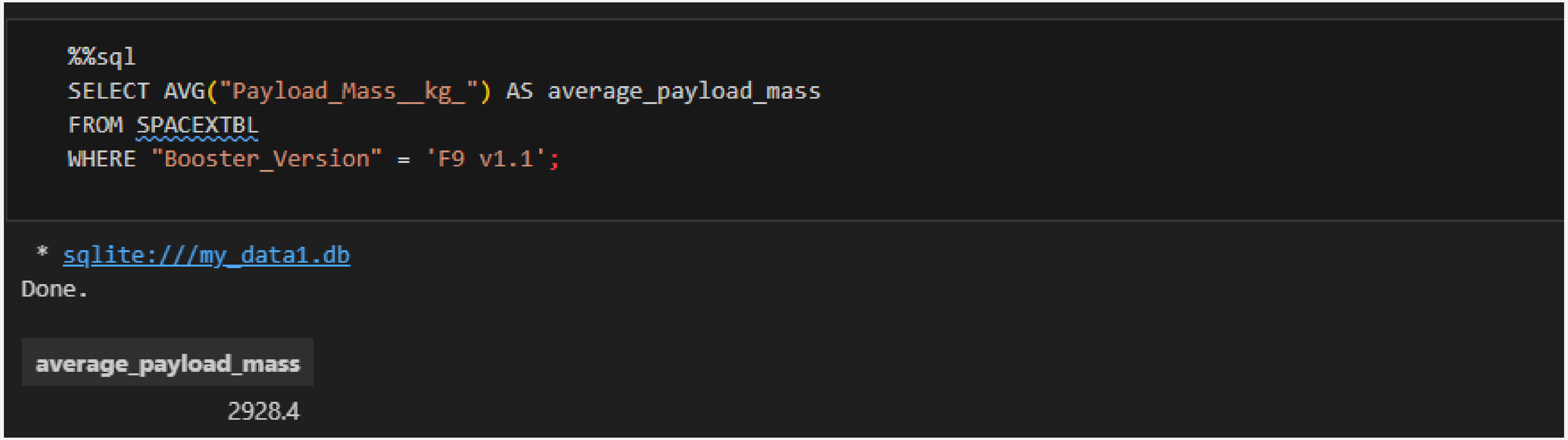
```
%%sql
SELECT SUM("Payload_Mass__kg_") AS total_payload_mass
FROM SPACEXTBL
WHERE "Customer" LIKE 'NASA (CRS)%';
```

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

total_payload_mass
             48213
```

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

Average Payload Mass by F9 v1.1



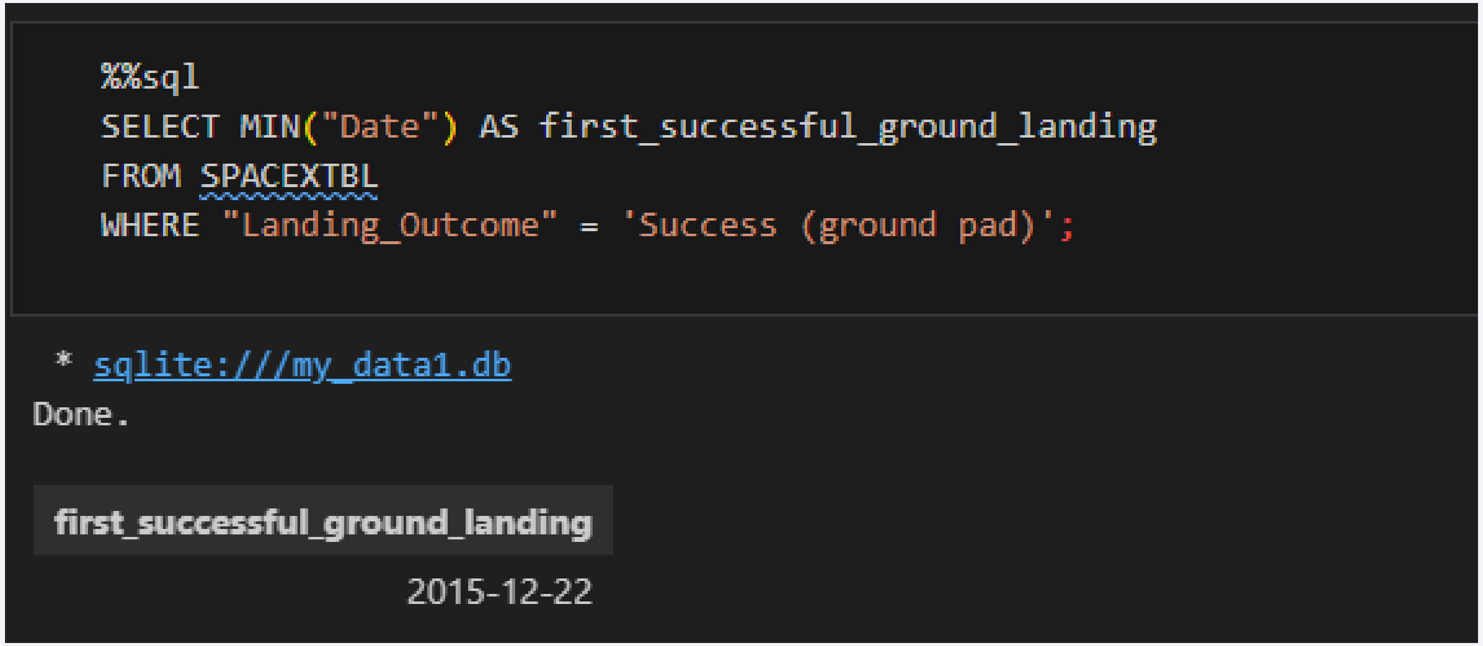
```
%%sql
SELECT AVG("Payload_Mass__kg_") AS average_payload_mass
FROM SPACEXTBL
WHERE "Booster_Version" = 'F9 v1.1';
```

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

average_payload_mass
2928.4

Displaying the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date



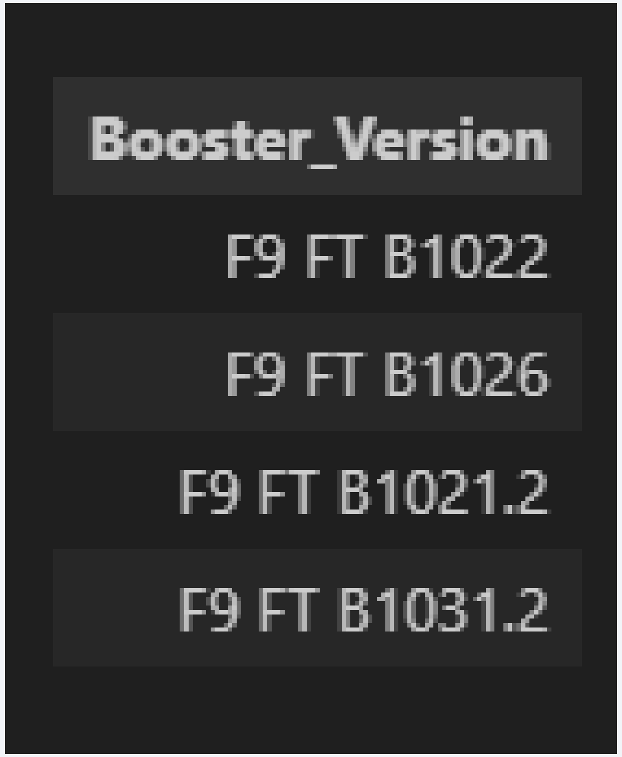
```
%%sql
SELECT MIN("Date") AS first_successful_ground_landing
FROM SPACEXTBL
WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

first_successful_ground_landing
2015-12-22

Listing the date of the first successful ground landing

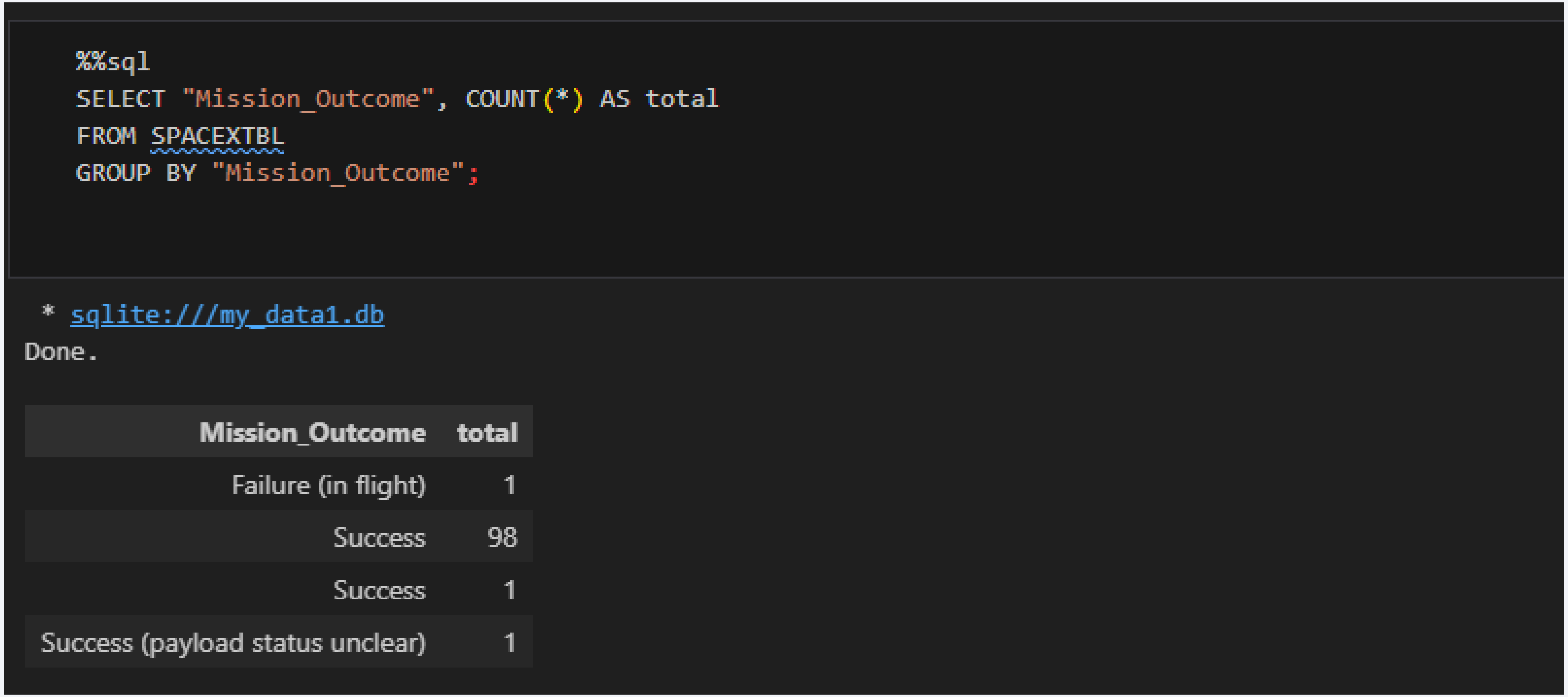
Successful Drone Ship Landing with Payload between 4000 and 6000



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

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

Total Number of Successful and Failure Mission Outcomes



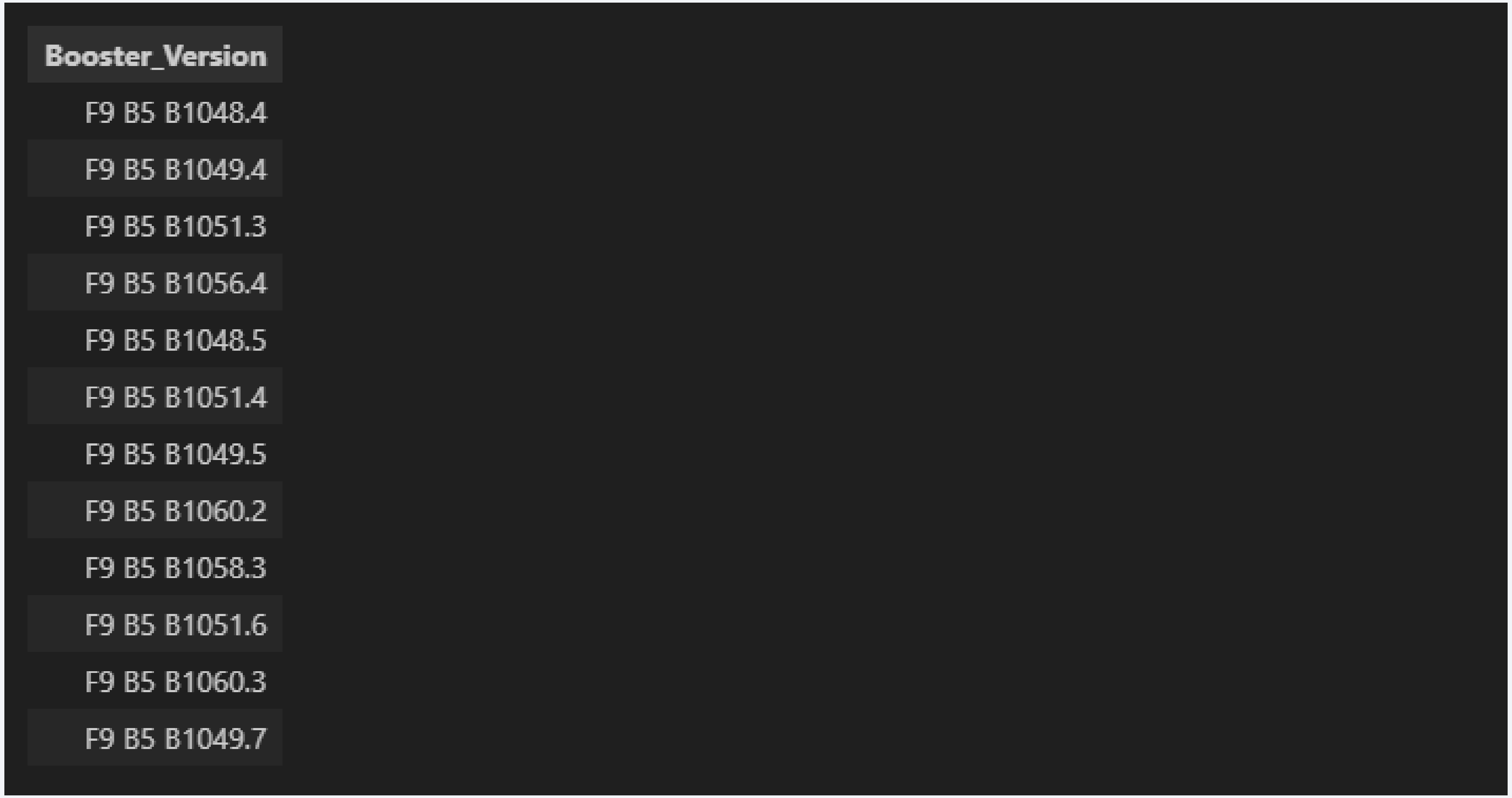
```
%%sql
SELECT "Mission_Outcome", COUNT(*) AS total
FROM SPACEXTBL
GROUP BY "Mission_Outcome";
```

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

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

Listing the total number of successful and failure mission outcomes

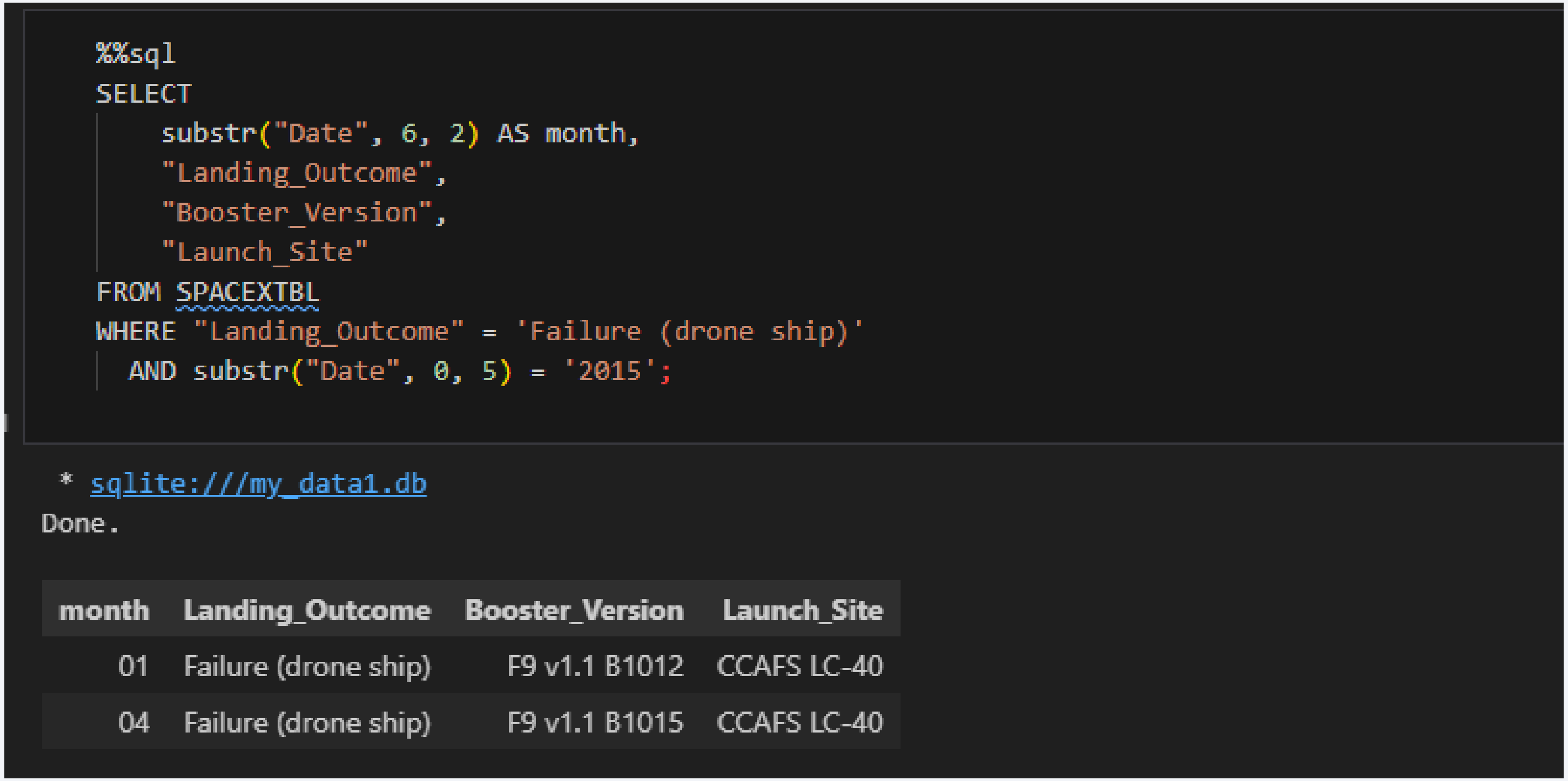
Boosters Carried Maximum Payload



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

Listing the names of the booster versions which have carried the maximum payload masses

2015 Launch Records



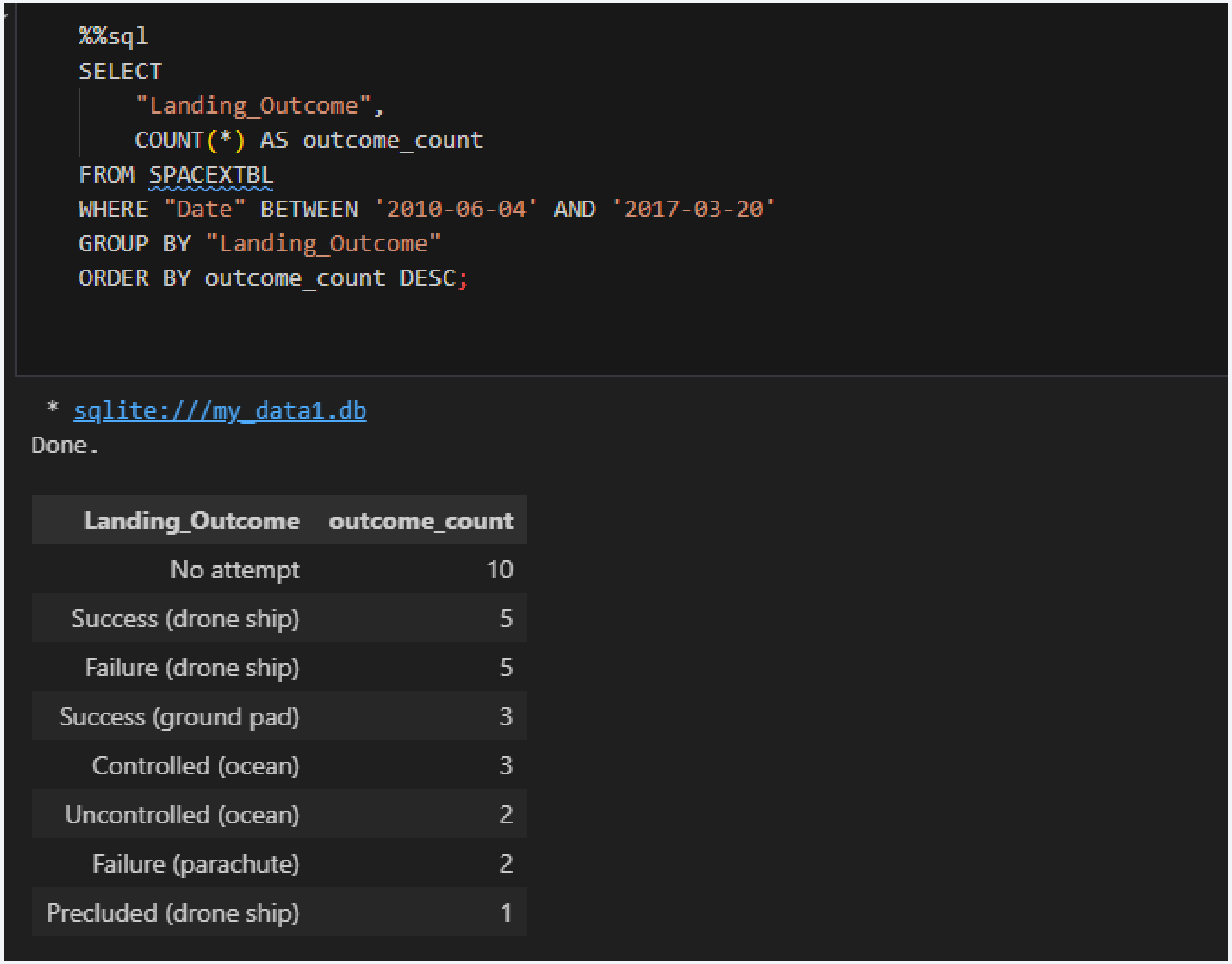
```
%%sql
SELECT
    substr("Date", 6, 2) AS month,
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM SPACEXTBL
WHERE "Landing_Outcome" = 'Failure (drone ship)'
  AND substr("Date", 0, 5) = '2015';
```

```
 * 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

Listing the failed landing outcomes in the year 2015

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



```
%%sql
SELECT
    "Landing_Outcome",
    COUNT(*) AS outcome_count
FROM SPACEXTBL
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY outcome_count DESC;
```

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

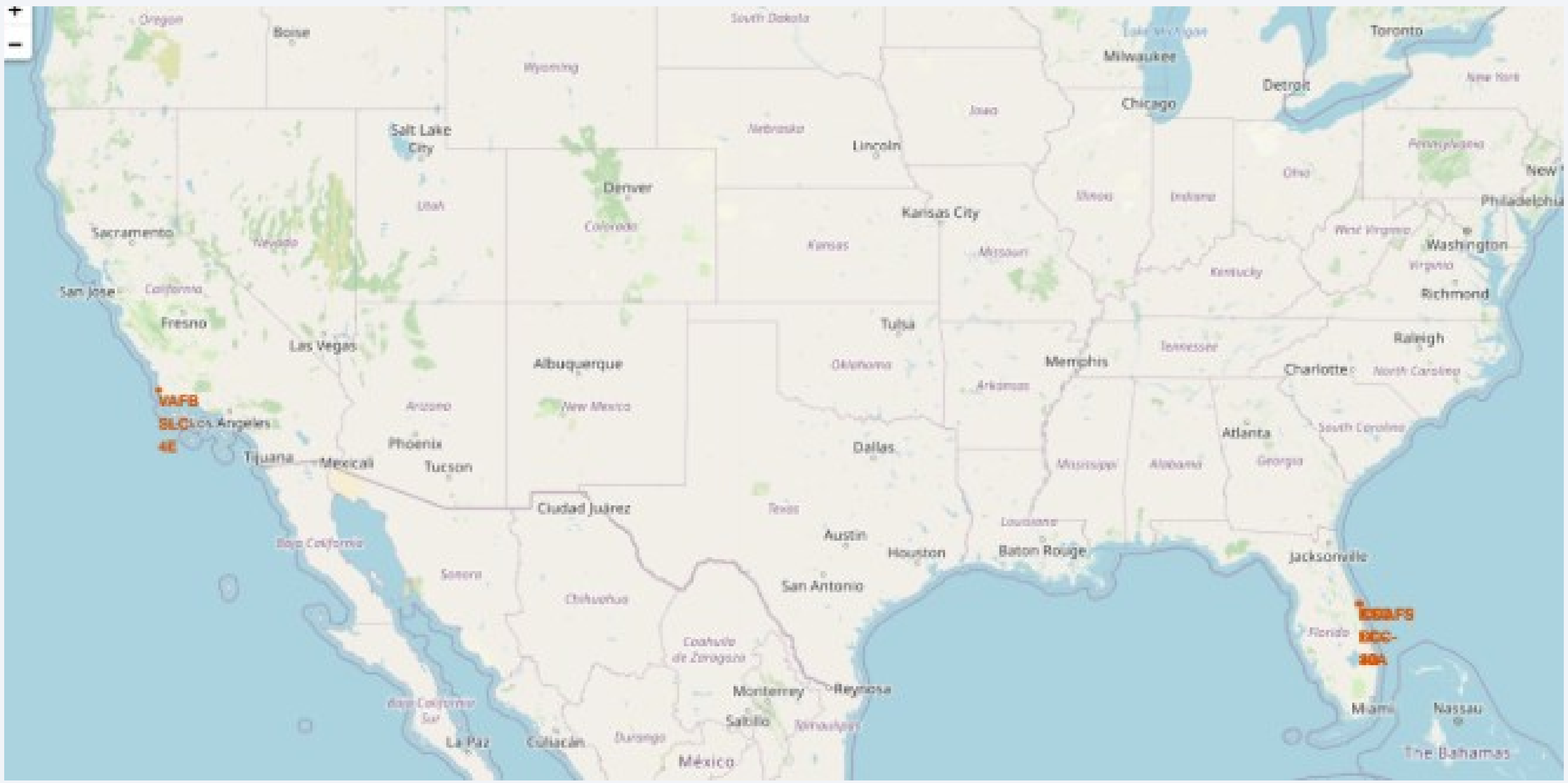
Landing_Outcome	outcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order

Section 3

Launch Sites Proximities Analysis

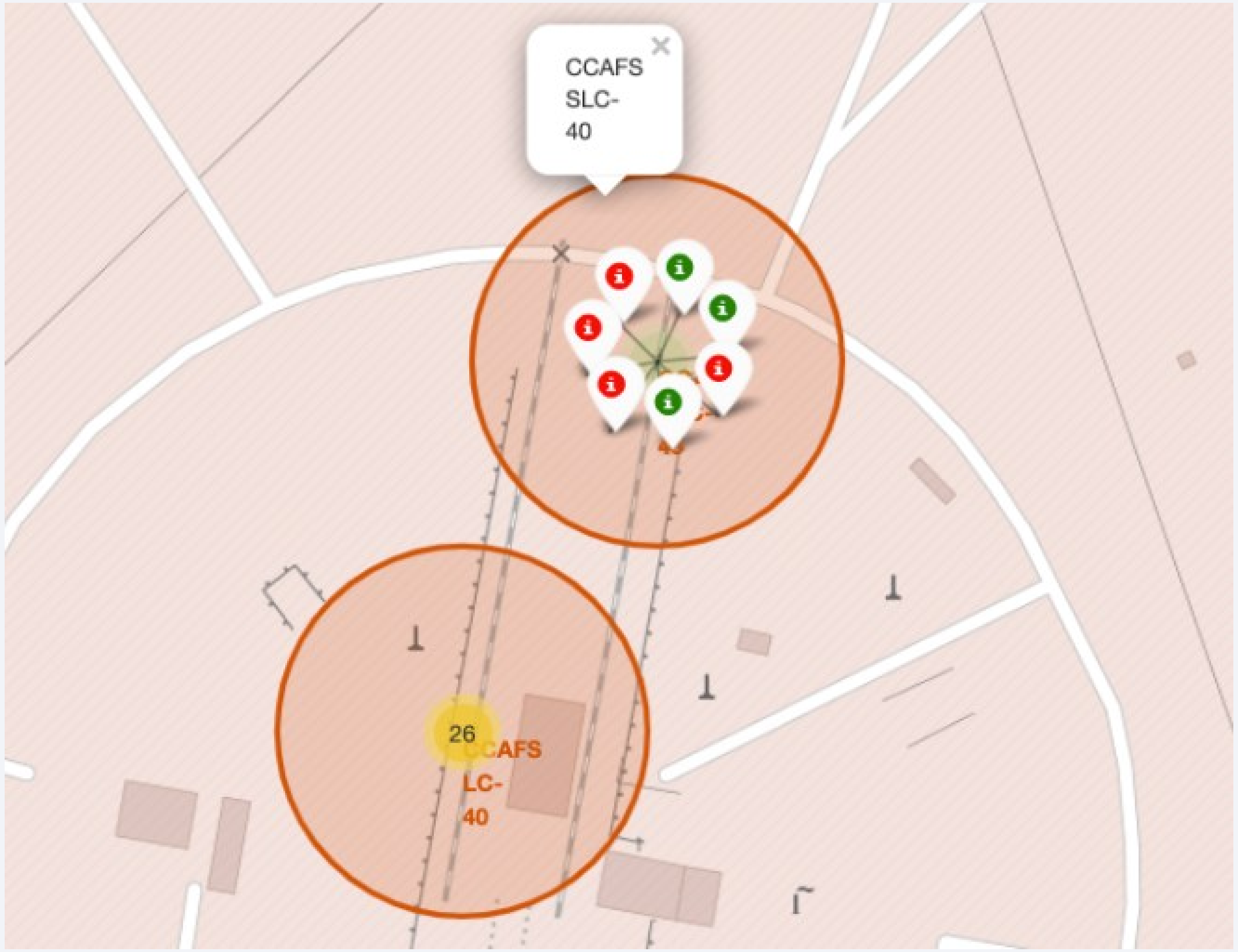
Launch Site location marker map



Most of Launch sites are in proximity to the Equator line
because the land is moving faster at the equator than any other place on the surface of the Earth.

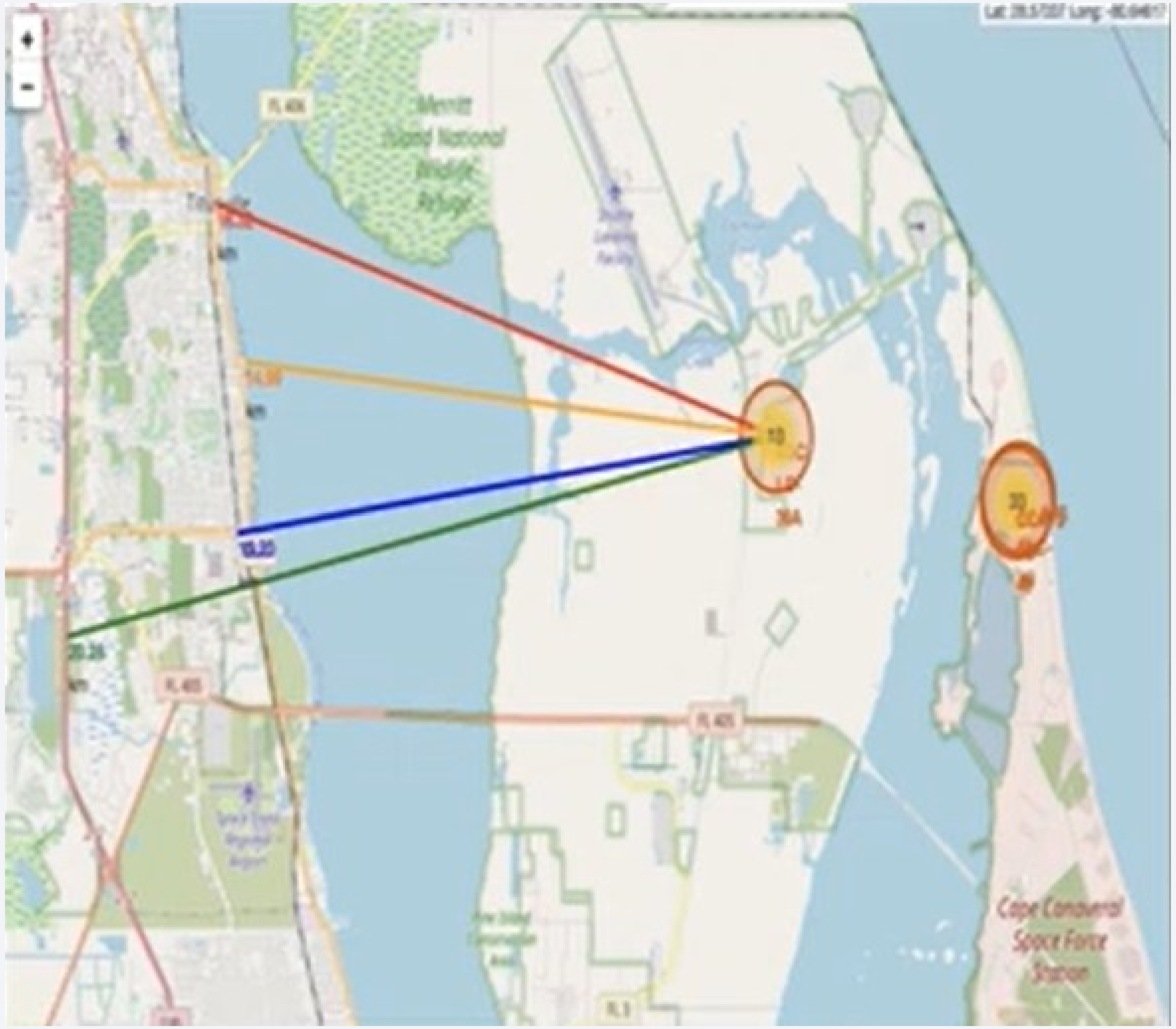
All launch sites are in very close proximity to the coast; while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people.

Colour labeled launch records map



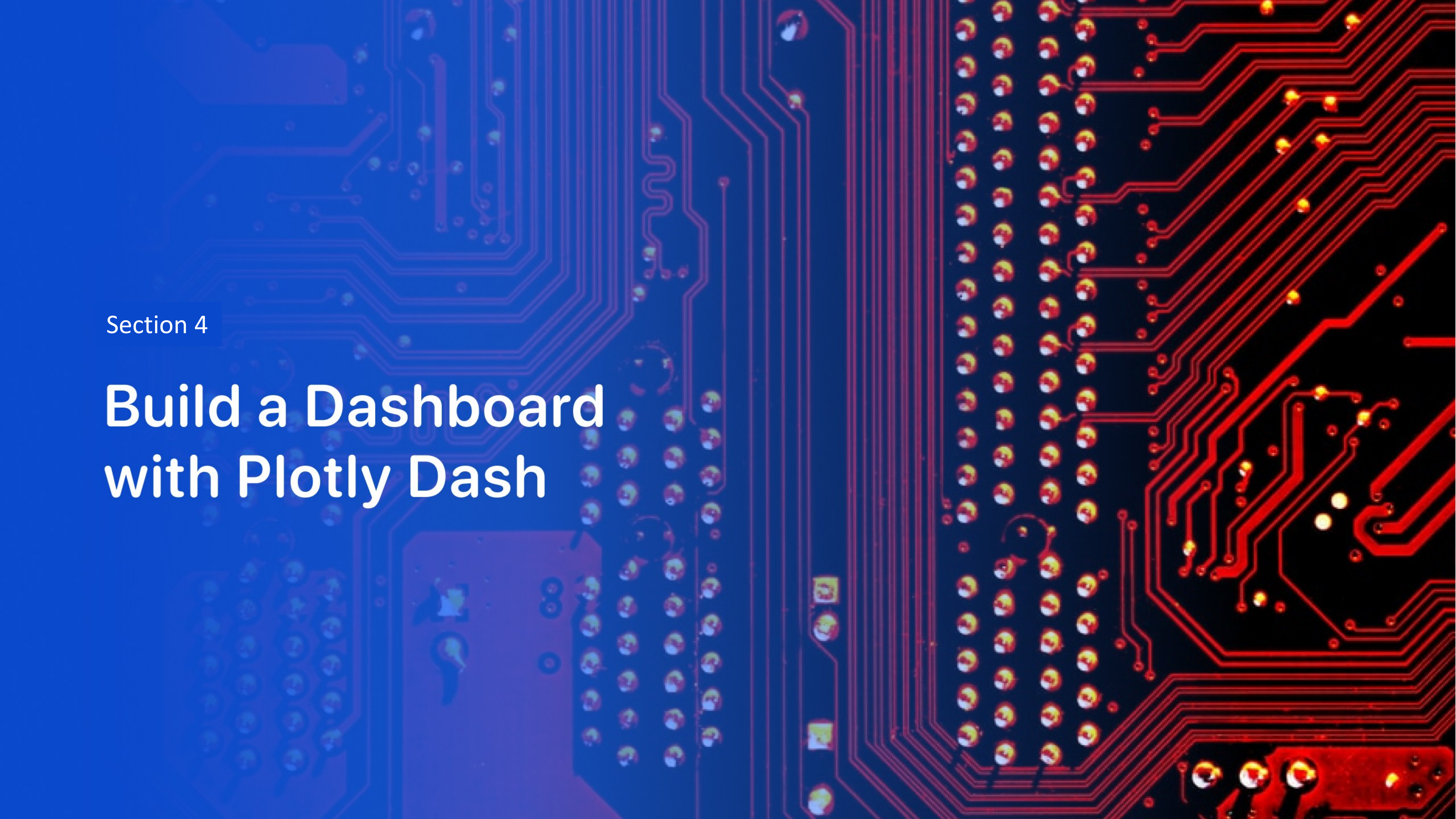
From The coloured labelled markers we can easily see which launch sites have relatively high success rates (green markers) and low success rates (red markers)

From KSC LC-39A launch site to proximities



The launch site is close to the coastline, railway and highway

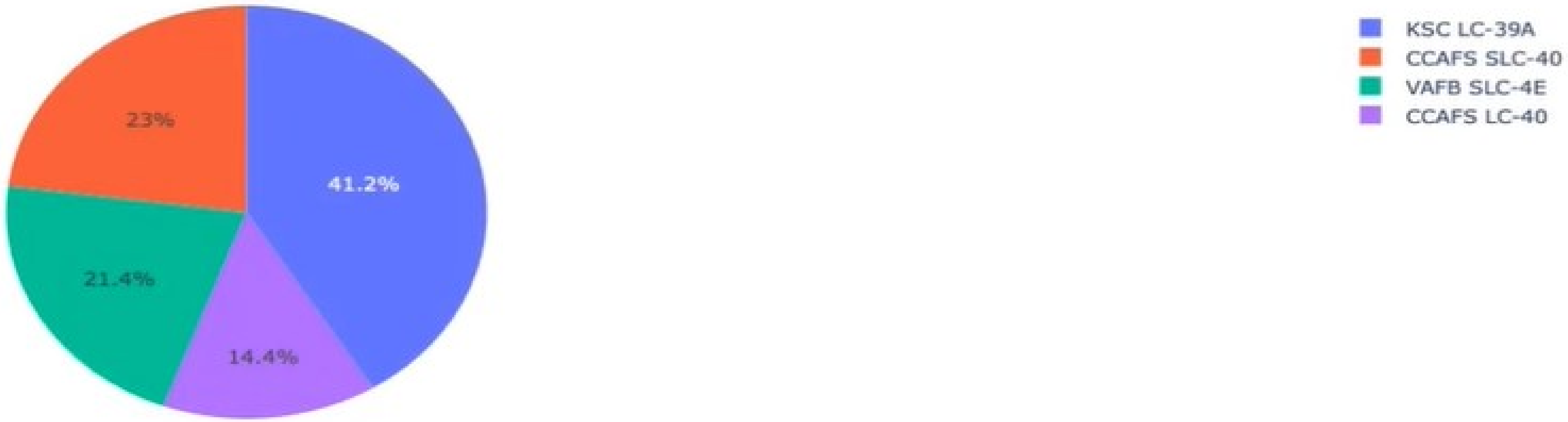
Its closest city is Titusville which is 16,32km away



Section 4

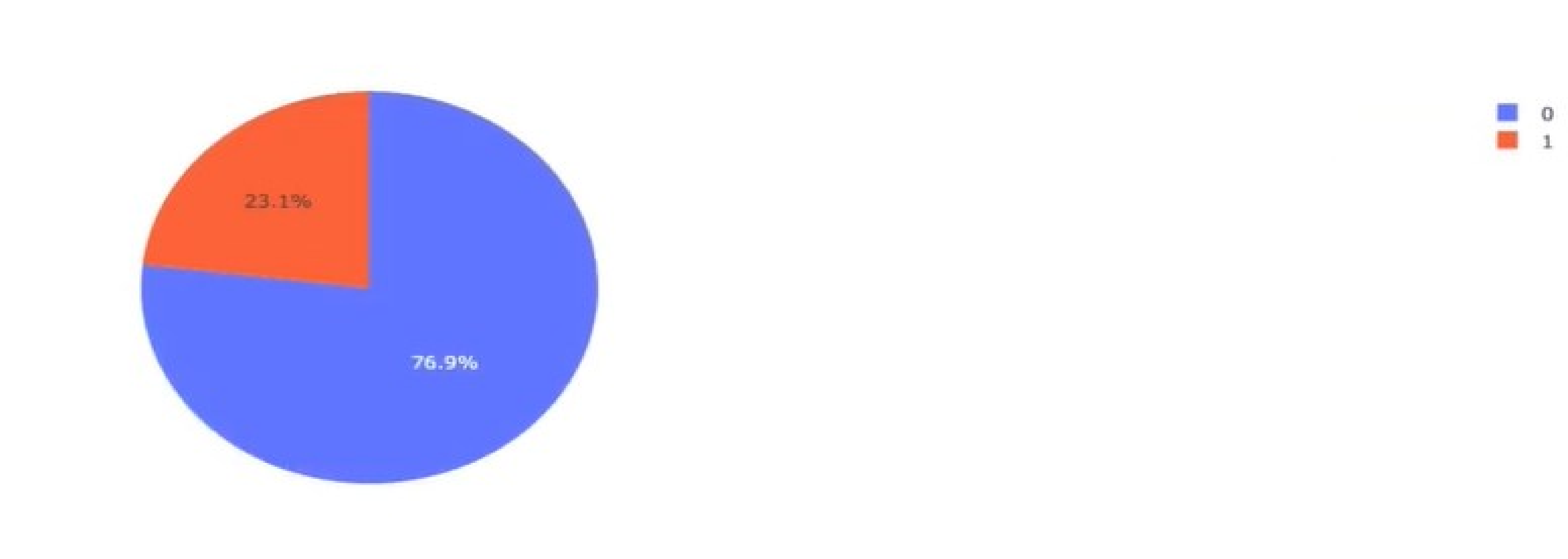
Build a Dashboard with Plotly Dash

Total successful launches by site



KSC LC-39A is clearly shown to be the most successful site and CCAFS LC-40 is the least

Total successful launches for the KSC LC-39A site

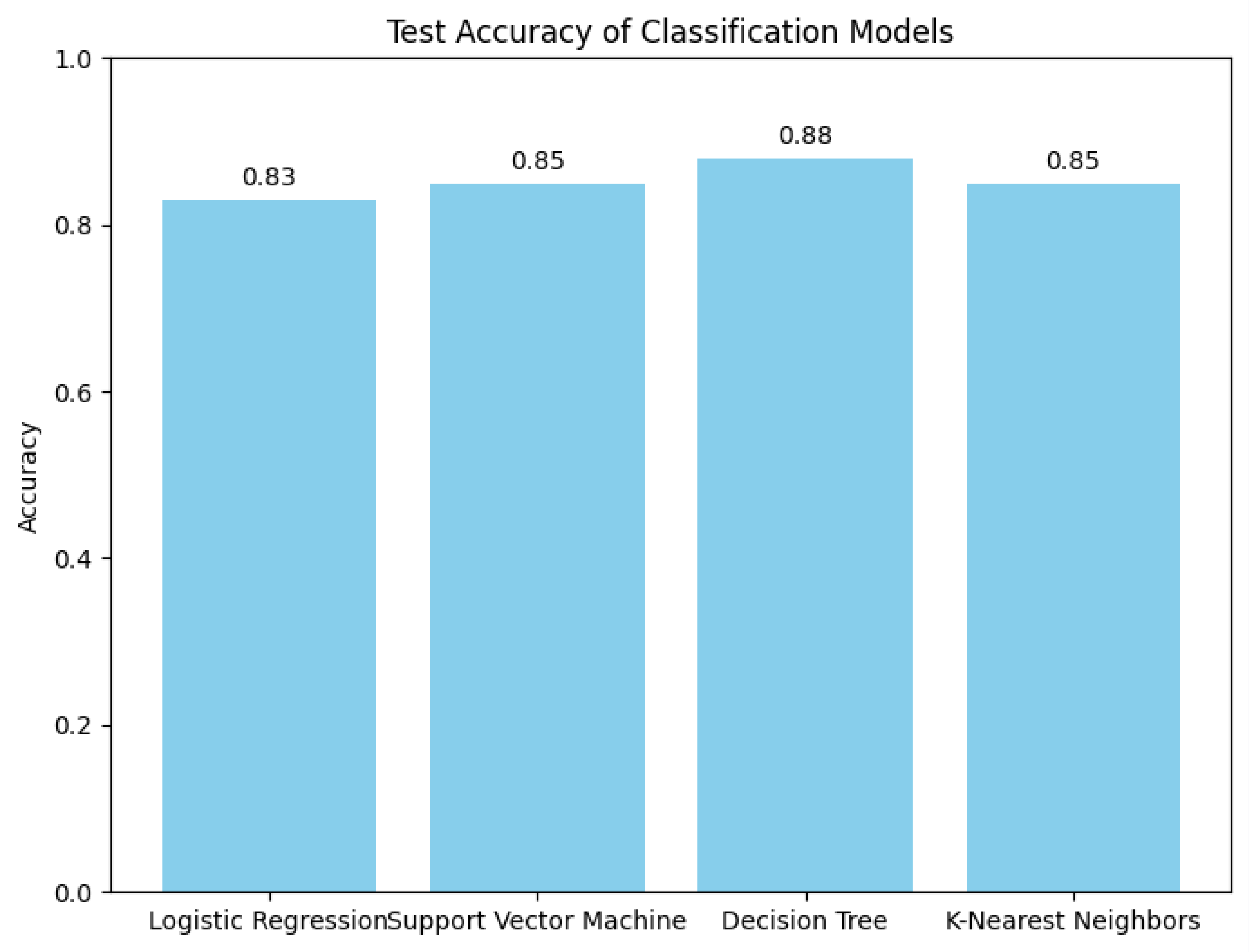


Shows a high success rate with 10 successes and 3 failed landings

Section 5

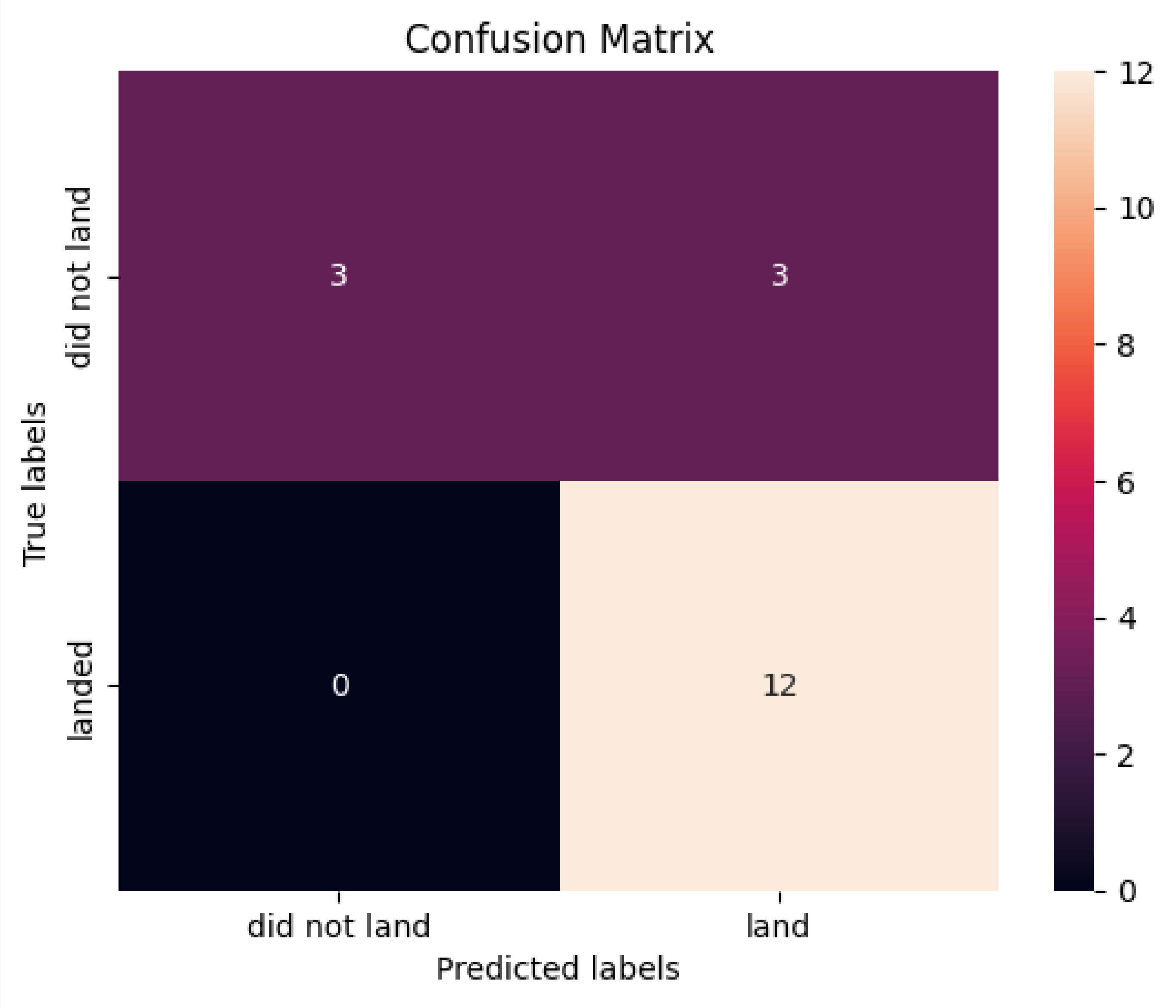
Predictive Analysis (Classification)

Classification Accuracy



The bar graph confirms that the Decision Tree model was the best performing with the highest accuracy

Confusion Matrix



Logistic Regression performed the best as it had the most true positives. It had a few number of false positives meaning it can distinguish between classes

Conclusions

- Decision Tree is the best performing model for this dataset
- Launches with low payload mass performed better than those with higher payload masses
- Most launch sites are in proximity to the equator and the coastline
- Launch success rate increases over the years
- KSC LC-39A has the highest launch success rate

Thank you!