IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Bastian L.>
<12-06-2024>

Outline

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

Executive Summary

Using the following methodologies:

- Data Collection through API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction

Summary of all results:

- Acquire data from open sources
- Feature selection via EDA
- Using a predictive ML model to identify key characteristics.

Introduction

Project background and context:

- Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

Problems you want to find answers:

- What factors determine if the rocket will land successfully?
- The interaction amongst various features that determine the success rate of a successful landing.
- What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Data Obtained from public sources:

  - SpaceX API (https://api.spacexdata.com/v4/rockets/)

  - Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

- Perform data wrangling

  Defining the landing outcome by analyzing key features.

- Perform exploratory data analysis (EDA) using visualization and SQL

- Perform interactive visual analytics using Folium and Plotly Dash

- Perform predictive analysis using classification models

  Compare various classification algorithms on a train-test split to identify the optimal model.

Data Collection

- Describe how data sets were collected:

  Datasets were collected from SpaceX API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

Data Collection – SpaceX API

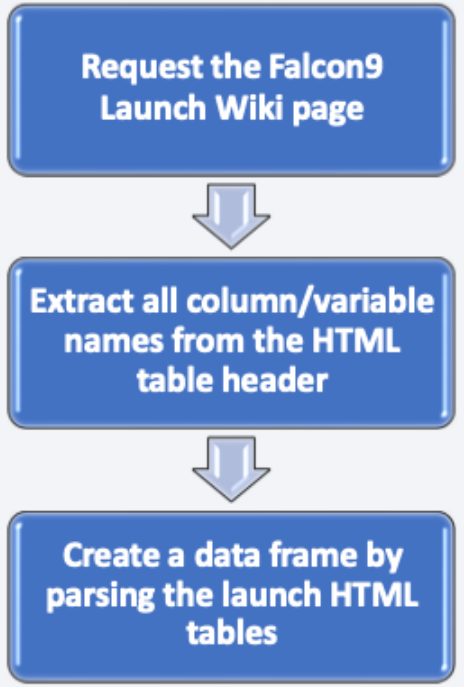
- Public API where data is stored
- This API was used according to the flowchart beside and then data is persisted.
- Code GitHub:

https://github.com/BastianLT/Ciencia-de-datos/blob/a9070d624c8fa4c3002954676587b53d84144583/Applied%20Data%20Science%20Capstone/1.jupyter-labs-spacex-data-collection-api-v2.ipynb

Data Collection - Scraping

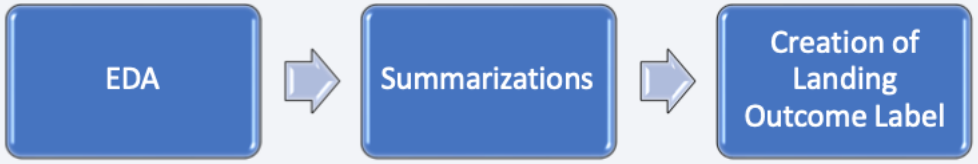
- Data from SpaceX launches obtained in Wikipedia.
- Code GitHub:

https://github.com/BastianLT/Ciencia-de-datos/blob/a9070d624c8fa4c3002954676587b53d84144583/Applied%20Data%20Science%20Capstone/2.jupyter-labs-webscraping.ipynb



Data Wrangling

- The process began with an Exploratory Data Analysis (EDA) of the dataset.
- Subsequently, calculations were performed to summarize launches by site, determine the occurrences of each orbit, and analyze mission outcomes per orbit type.
- Finally, the landing outcome label was created from Outcome column.



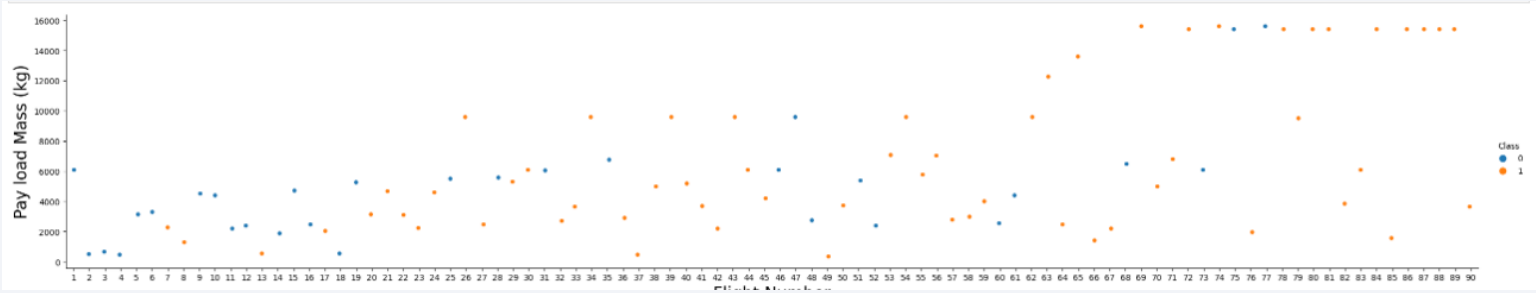
- Code GitHub: https://github.com/BastianLT/Ciencia-de-datos/blob/a9070d624c8fa4c300295467658 7b53d84144583/Applied%20Data%20Science%20Capstone/3.labs-jupyter-spacex-Data%20wrangling-v2.ipynb

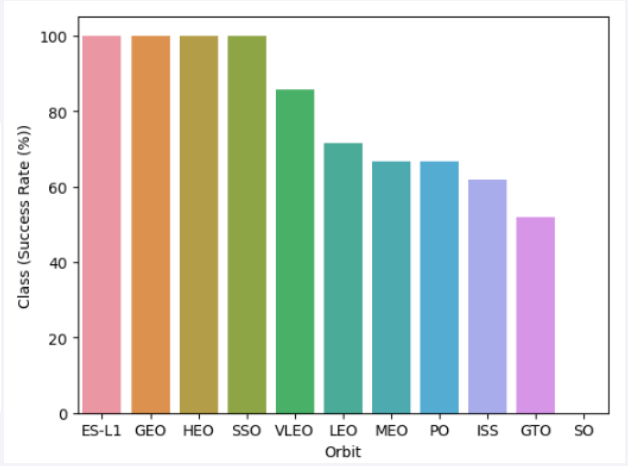
EDA with Data Visualization

- The following SQL queries were performed:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA’
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List all the booster versions that have carried the maximum payload mass. Use a subquery.
  - List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
  - Rank 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.
- Code GitHub: https://github.com/BastianLT/Ciencia-de-datos/blob/a9070d624c8fa4c3002954676587b53d84144583/Applied%20Data%20Science%20Capstone/4.jupyter-labs-eda-sql-coursera_sqllite.ipynb

EDA with SQL

- To explore data, scatterplots and bar plots were used to visualize the relationship between pair of features:
- Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit.





- Code Github: https://github.com/BastianLT/Ciencia-de-datos/blob/a9070d624c8fa4c3002954676587b53d84144583/Applied%20Data%20Science%20Capstone/5.jupyter-labs-eda-dataviz-v2.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps.
- Markers indicate points like launch sites.
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center
- Marker clusters indicates groups of events in each coordinate, like launches in a launch site and Lines are used to indicate distances between two coordinates.

- Code GitHub: https://github.com/BastianLT/Ciencia-de-datos/blob/a9070d624c8fa4c300295467658 7b53d84144583/Applied%20Data%20Science%20Capstone/6.lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

We analyzed the relationship between launch locations and payload capacity to find the most efficient options. Our analysis involved visualizing:

- The distribution of launches by site.
- The range of payloads launched from each location.

This approach enabled us to directly correlate specific sites with their ideal payload weights.

- Code GitHub: https://github.com/BastianLT/Ciencia-de-datos/blob/a9070d624c8fa4c300295467658 7b53d84144583/Applied%20Data%20Science%20Capstone/7.Interactive%20Dashboard%20with%20Ploty%20Dash.py

Predictive Analysis (Classification)

- We developed a classification model by first preparing the data with NumPy and Pandas, which included transformation and splitting into training/testing sets.
- We then built and systematically tuned several machine learning models using GridSearchCV to find the optimal hyperparameters.
- The model's performance was measured by accuracy, and we enhanced the results through iterative feature engineering and algorithm tuning to identify the top-performing model.

- Code GitHub: https://github.com/BastianLT/Ciencia-de-datos/blob/a9070d624c8fa4c3002954676587b53d84144583/Applied%20Data%20Science%20Capstone/8.SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

Results

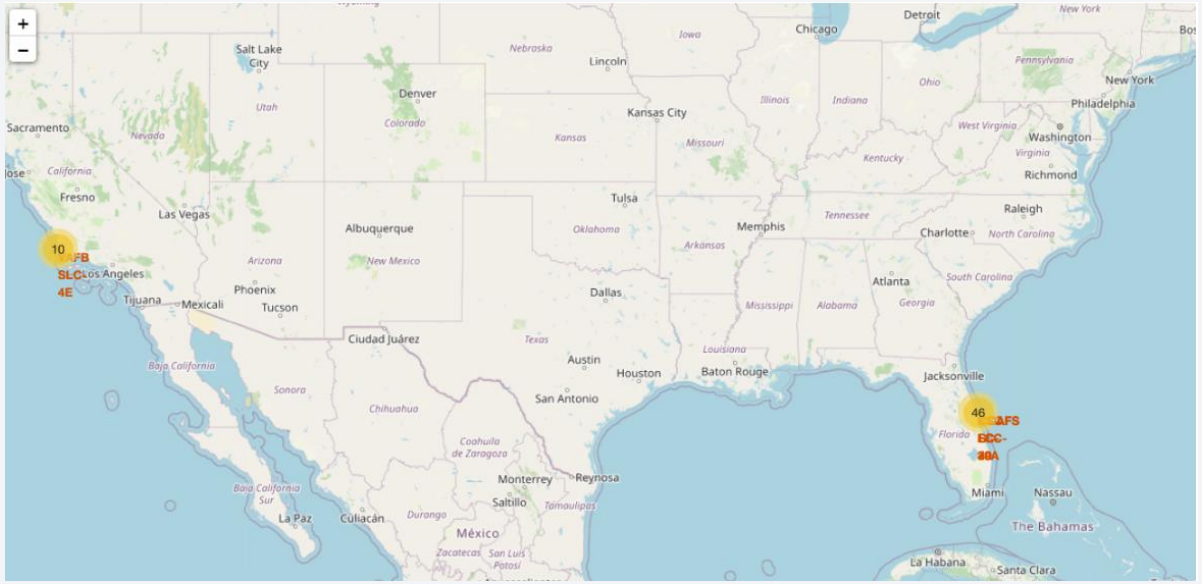
The EDA results:

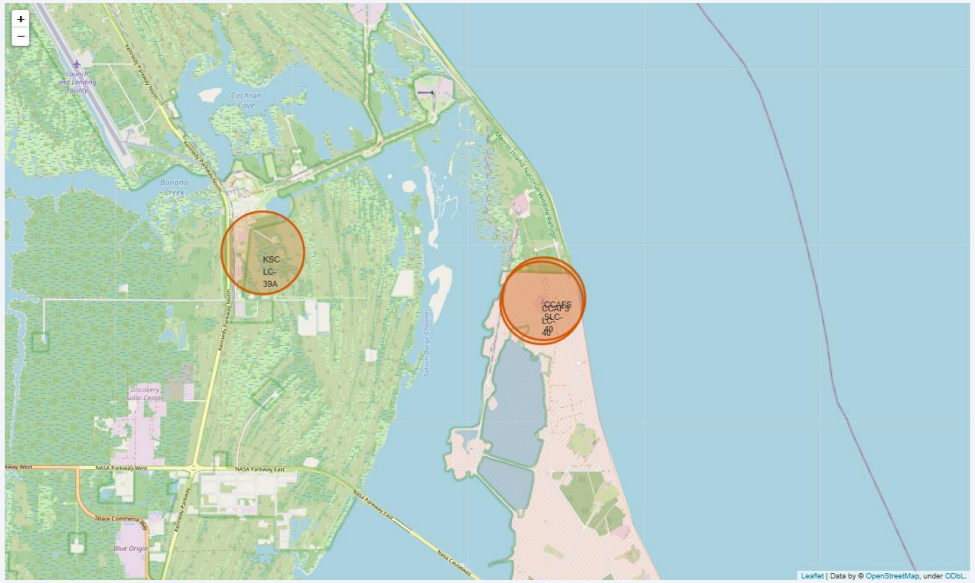
- Space X uses 4 different launch sites.
- The first launches were done to Space X itself and NASA.
- The first success landing outcome happened in 2015 fiver year after the first launch.
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015.
- The number of landing outcomes became as better as years passed.

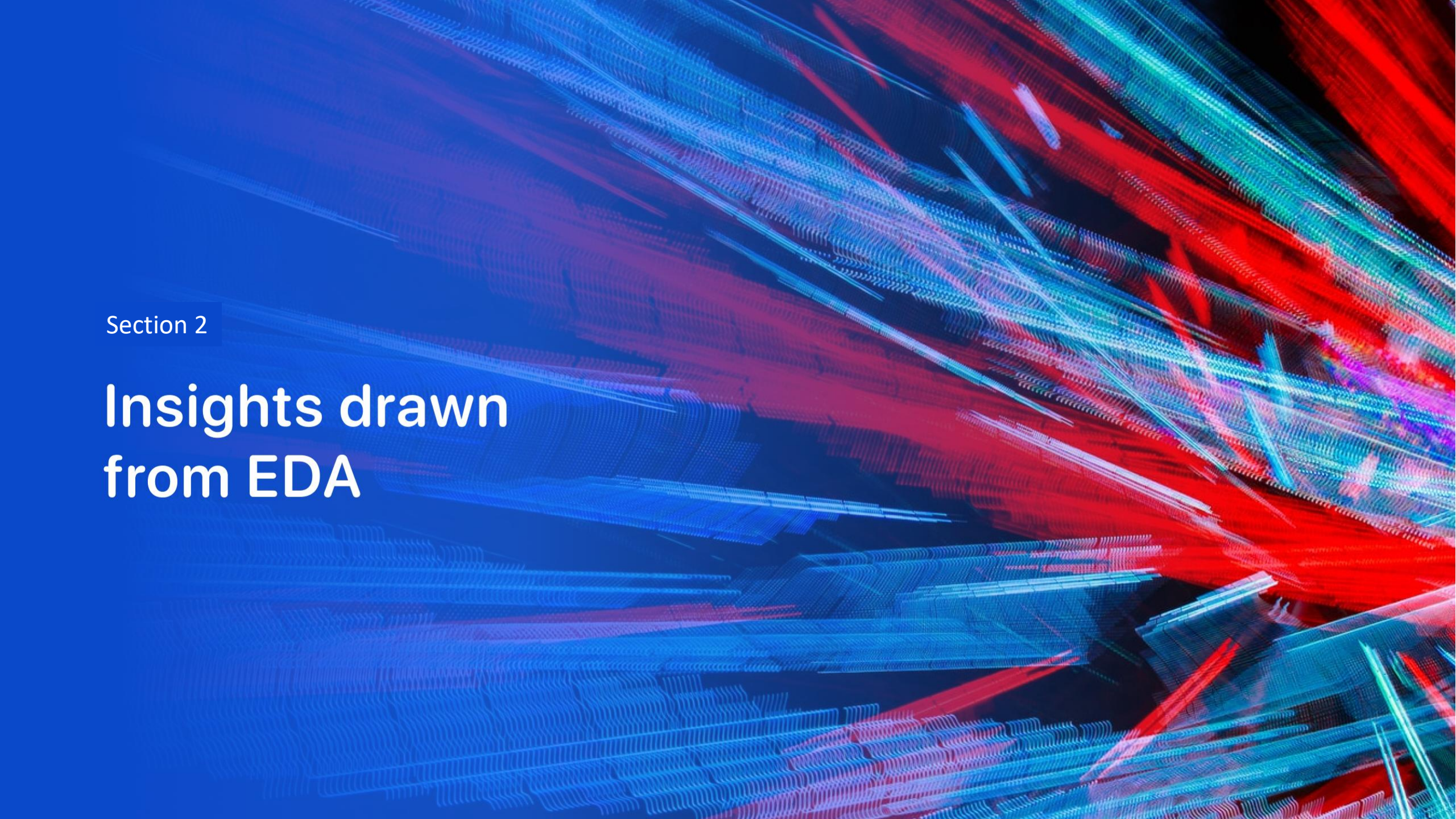
Results

Using interactive analytics, we identified that optimal launch sites share common characteristics: they are situated in secure locations (often coastal) and have well-developed logistical infrastructure.

Launch sides:

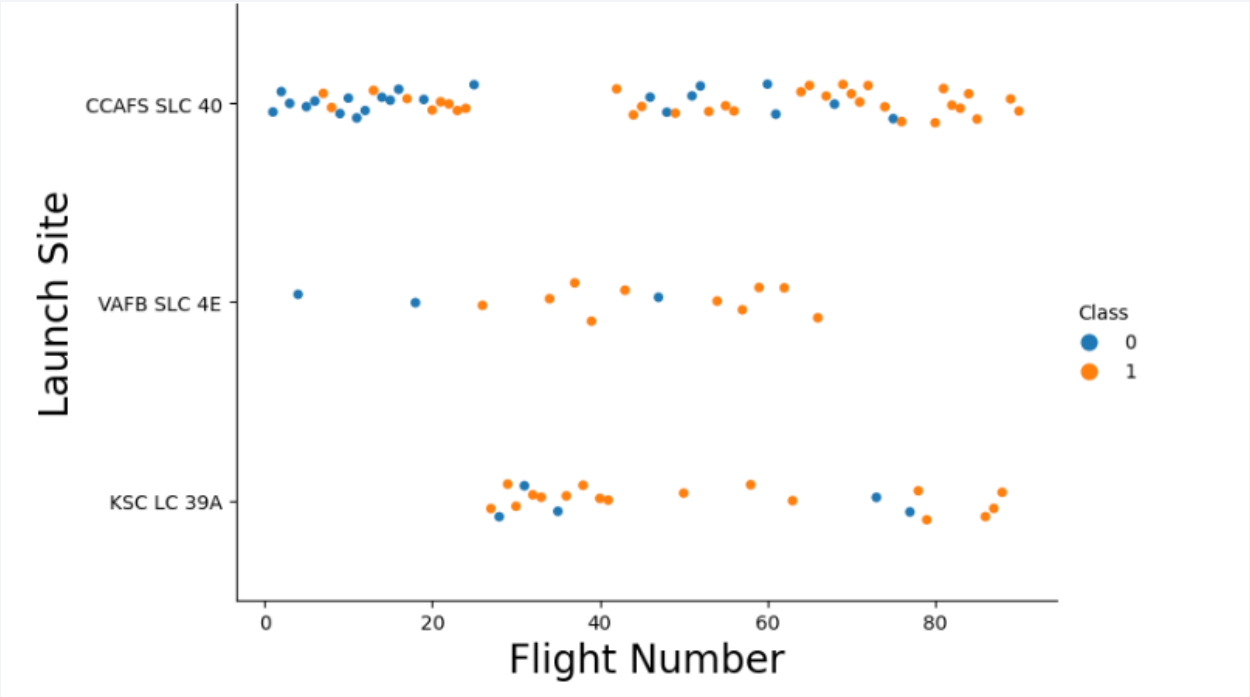





Section 2
Insights drawn from EDA

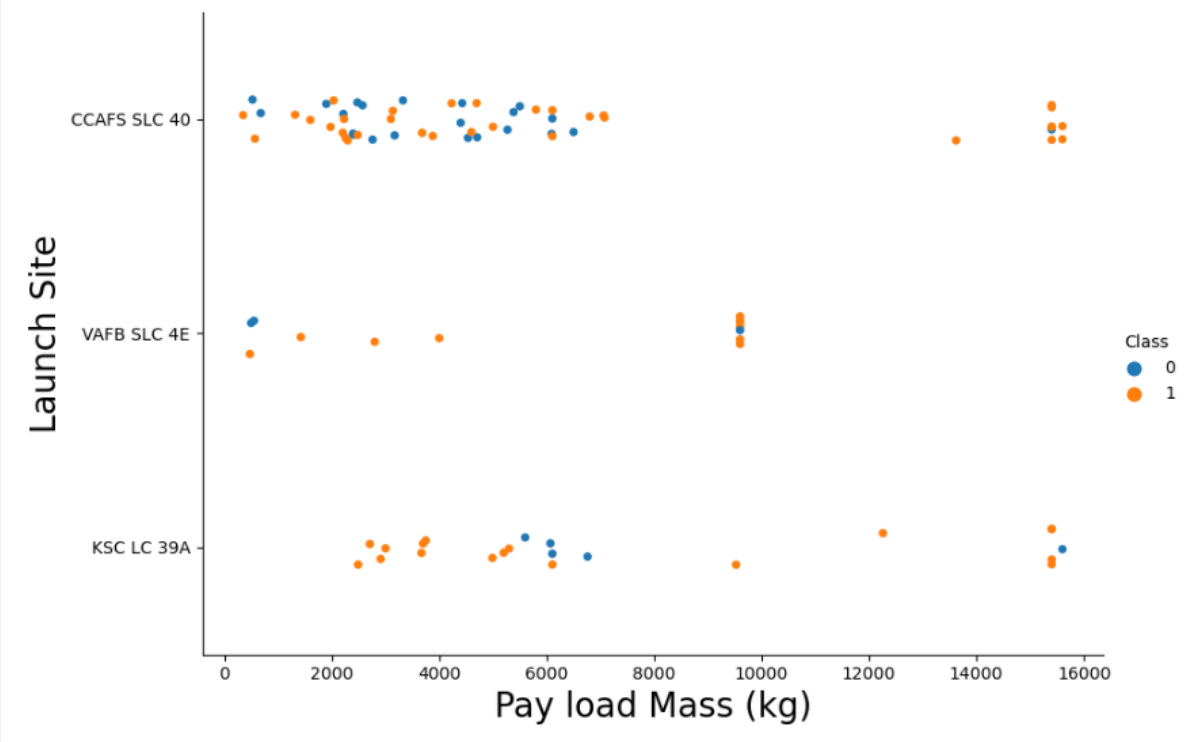
Flight Number vs. Launch Site

- The polt shows that the launch success rate improves as the number of flights increases, demonstrating a clear learning curve. The experience gained over time has allowed the company to significantly increase the reliability of its recent missions.



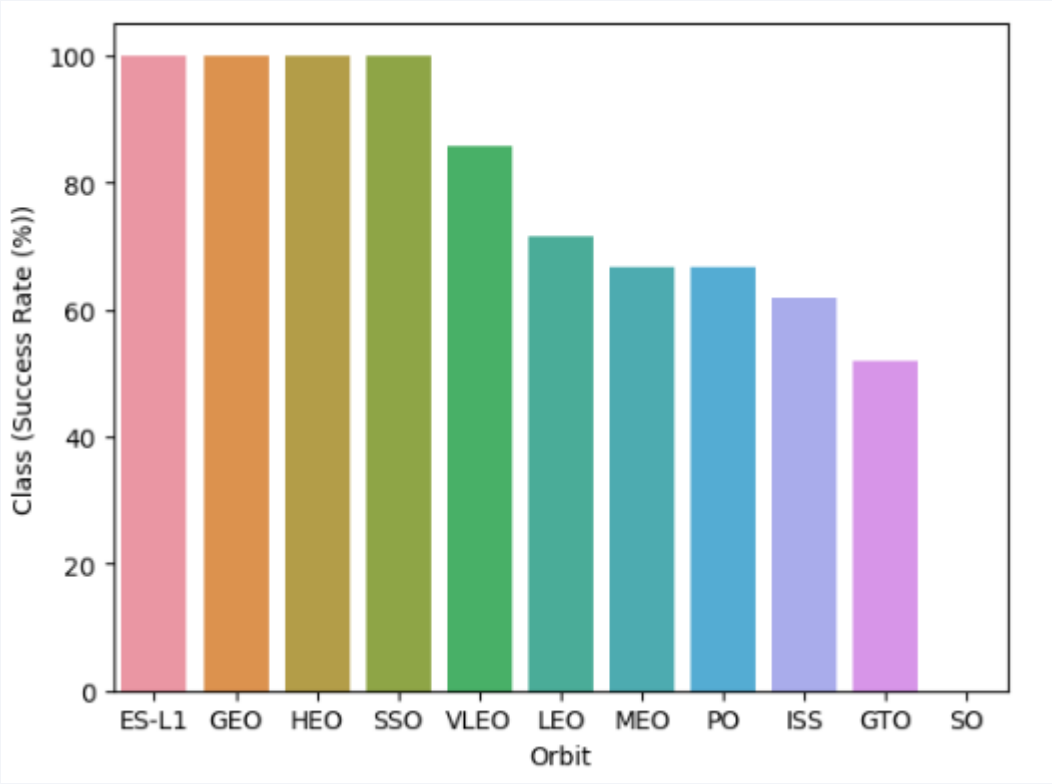
Payload vs. Launch Site

- Launch site selection is directly tied to payload mass. Instead of being a risk, the launch system is highly reliable for heavy-lift missions, which are handled by specialized sites like KSC LC 39A and show a high probability of success.



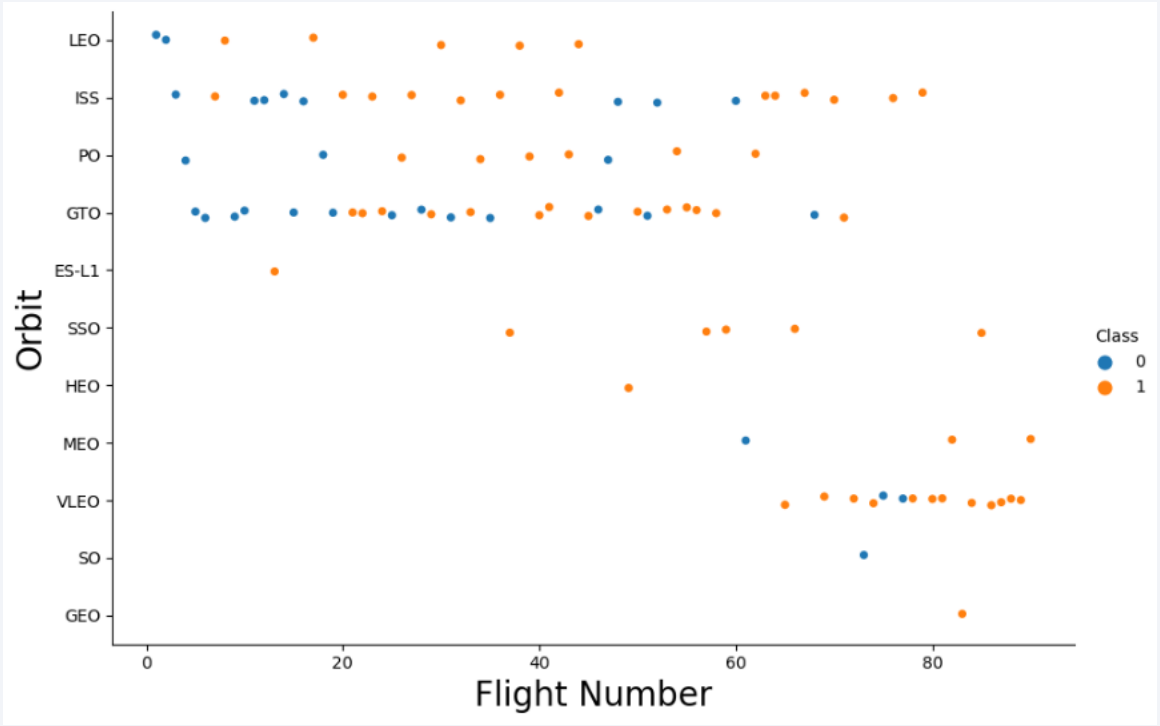
Success Rate vs. Orbit Type

- From the plot, we can see that ES-L1, GEO, HEO, SSO, had the most success rate.



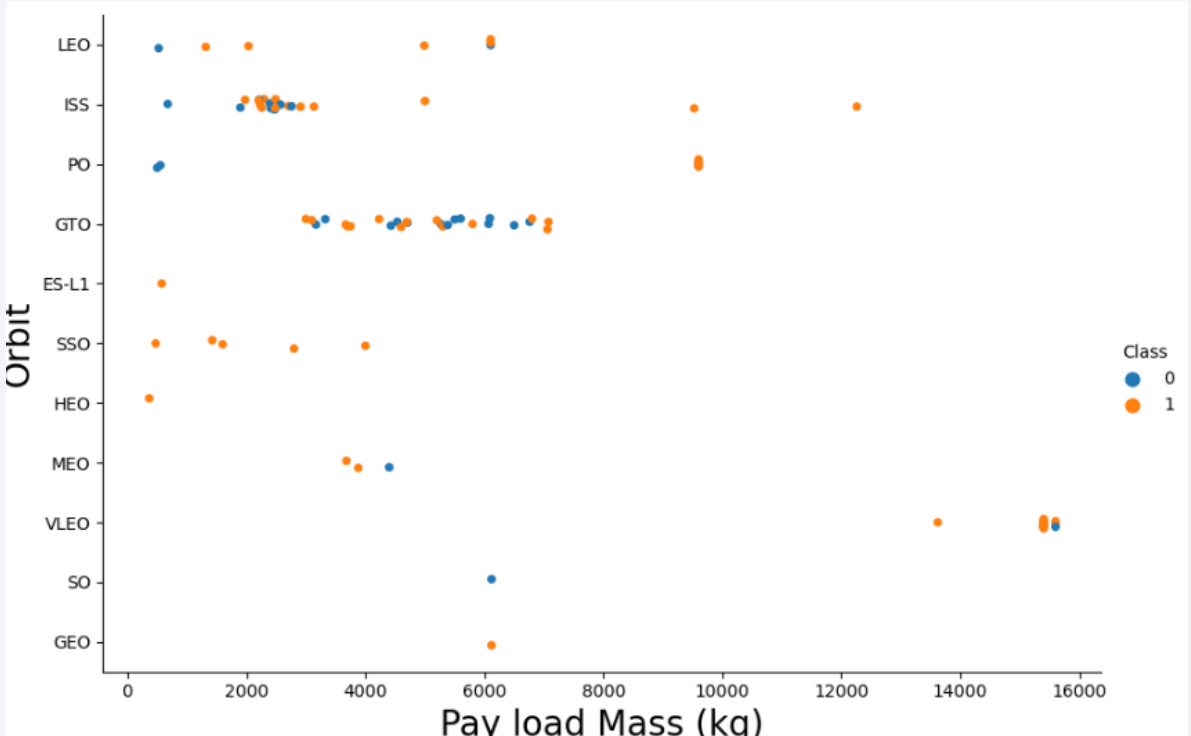
Flight Number vs. Orbit Type

- The plot shows the relationship between Flight Number and Orbit. We observe that for the LEO orbit, the success rate is related to the flight number, whereas for the GTO orbit, there is no clear relationship between the flight number and **mission success.**



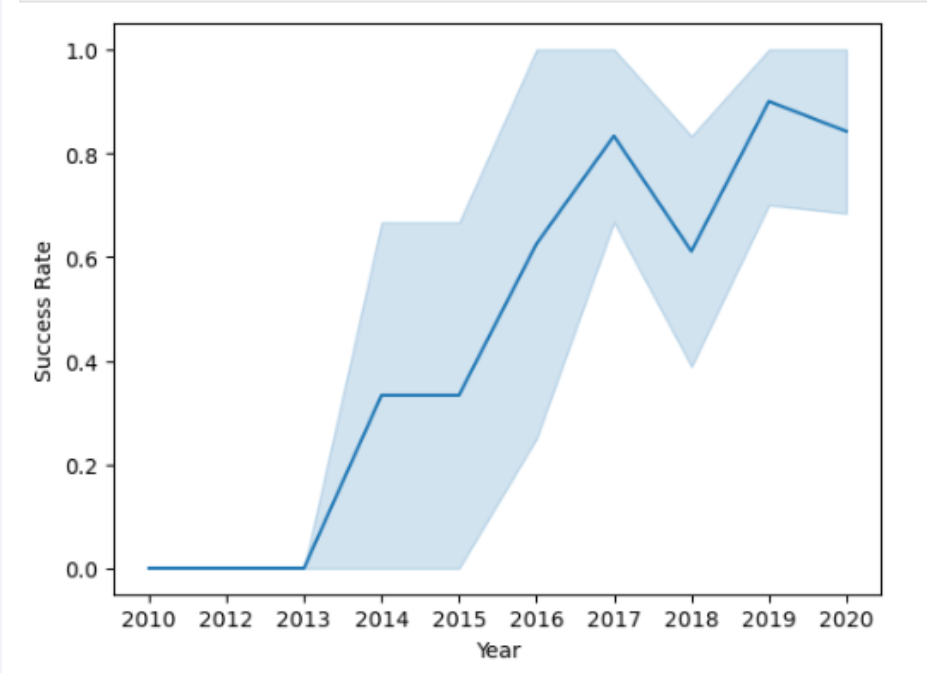
Payload vs. Orbit Type

- Mission success isn't dependent on a light payload. The results demonstrate that heavy-lift missions are highly successful when targeting specific orbits, whereas GTO missions pose the greatest challenge, irrespective of the specific payload mass



Launch Success Yearly Trend

- This plot shows a clear learning curve, starting with a 0% success rate until 2013, followed by a rapid improvement to over 80% by 2017. After a significant dip in 2018, the success rate recovered to its highest point, finishing the decade with high but variable reliability. This demonstrates a mature system that overcame initial failures to achieve consistent success.



All Launch Site Names

- These are all the launch sites to show only unique launch sites from the SpaceX data.

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

Launch Site Names Begin with 'CCA'

- We used the query to display 5 records where the launch sites begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * FROM 'SPACEXTBL' 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 carried by NASA's boosters is 45,596 kg.

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass(Kgs)", Customer FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)';
```

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

Total Payload Mass(Kgs)	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

- The average payload of F9 v1,1 is 2928.4 kgs-

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) as "Payload Mass Kgs", Customer, Booster_Version FROM 'SPACEXTBL' WHERE Booster_Version LIKE 'F9 v1.1';
```

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

Payload Mass Kgs	Customer	Booster_Version
2928.4	SES	F9 v1.1

First Successful Ground Landing Date

First Landing is in 22 of December of 2015

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (ground pad)";
```

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

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Successful landing of a drone with a payload of between 4000 and 6000 kg. The F9 B5 booster achieved this with a payload of 5800 kg.

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

```
]: %sql SELECT "Booster_Version", "Payload", "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD
 * sqlite:///my_data1.db
Done.
```

]:

Booster_Version	Payload	PAYLOAD_MASS__KG_
F9 B5 B1046.2	Merah Putih	5800

Total Number of Successful and Failure Mission Outcomes

- The total numbers of missions Successful and Failure.

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

Boosters Carried Maximum Payload

- These are the boosters that have carried the maximum payload mass.

```
%sql SELECT "Booster_Version",Payload, "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL);
```

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

Booster_Version	Payload	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Starlink 1 v1.0, SpaceX CRS-19	15600
F9 B5 B1049.4	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600
F9 B5 B1051.3	Starlink 3 v1.0, Starlink 4 v1.0	15600
F9 B5 B1056.4	Starlink 4 v1.0, SpaceX CRS-20	15600
F9 B5 B1048.5	Starlink 5 v1.0, Starlink 6 v1.0	15600
F9 B5 B1051.4	Starlink 6 v1.0, Crew Dragon Demo-2	15600
F9 B5 B1049.5	Starlink 7 v1.0, Starlink 8 v1.0	15600
F9 B5 B1060.2	Starlink 11 v1.0, Starlink 12 v1.0	15600
F9 B5 B1058.3	Starlink 12 v1.0, Starlink 13 v1.0	15600
F9 B5 B1051.6	Starlink 13 v1.0, Starlink 14 v1.0	15600
F9 B5 B1060.3	Starlink 14 v1.0, GPS III-04	15600
F9 B5 B1049.7	Starlink 15 v1.0, SpaceX CRS-21	15600

2015 Launch Records

- These are the results of failed in landing outcomes in drone ship, their booster versions, and launch site names for 2015.

```
%sql SELECT substr(Date,6,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, Landing_Outcome FROM SPACEXTBL WHERE Landing_Outcome='Failure (drone ship)' AND substr(Date,0,5)='2015';
```

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

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)

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

- These are the results of landings between June 4, 2010, and March 20, 2017.

Rank 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.

```
%sql SELECT Landing_Outcome, COUNT(*) FROM SPACEXTBL WHERE DATE BETWEEN "2010-06-04" AND "2017-03-20" GROUP BY Landing_Outcome HAVING Landing_Outcome="Success (ground pad)" OR Landing_Outcome="Fail
```

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

Landing_Outcome	COUNT(*)
Success (ground pad)	3
Failure (drone ship)	5

Section 3

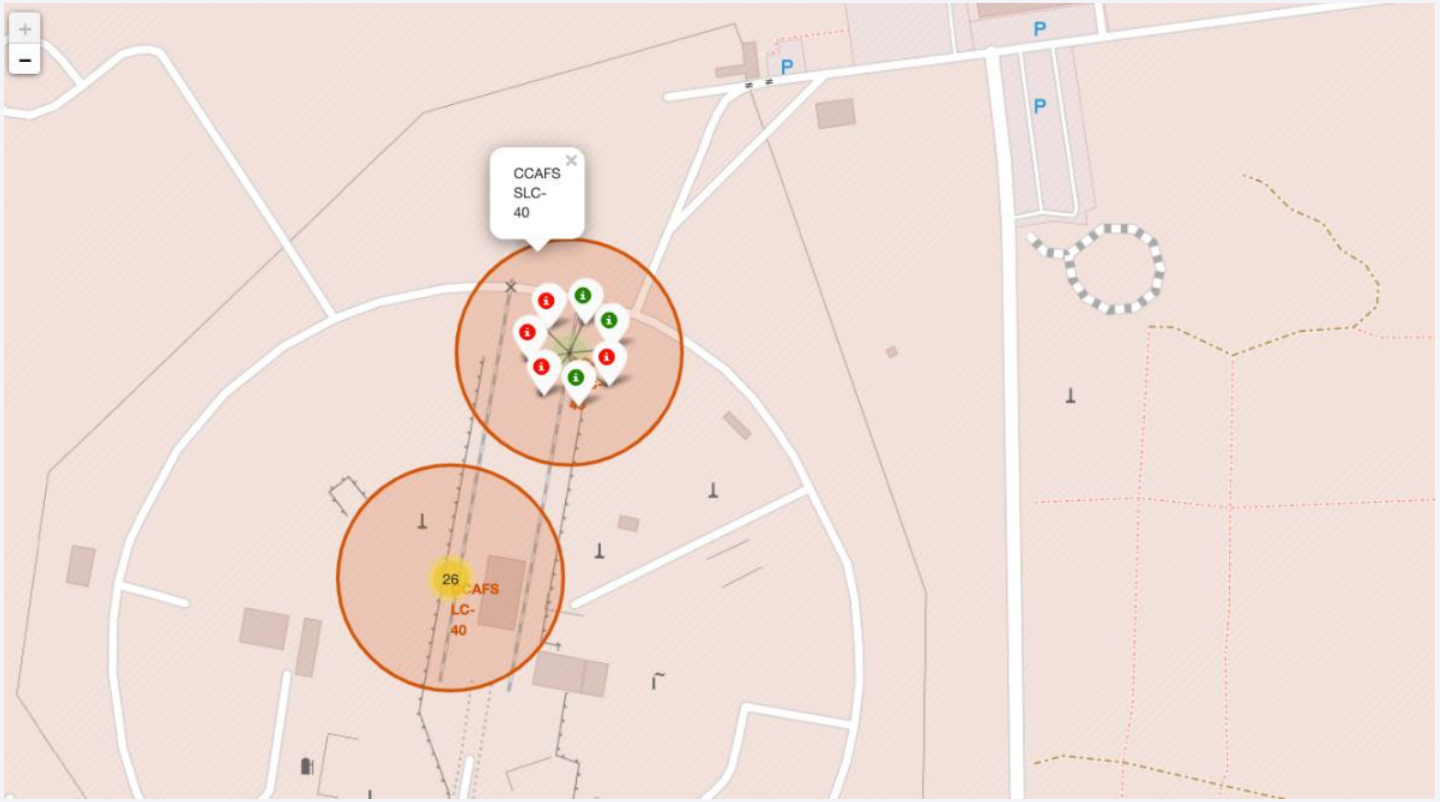
Launch Sites Proximities Analysis

All launch sites global map markers

- All launch sides of SpaceX in united states

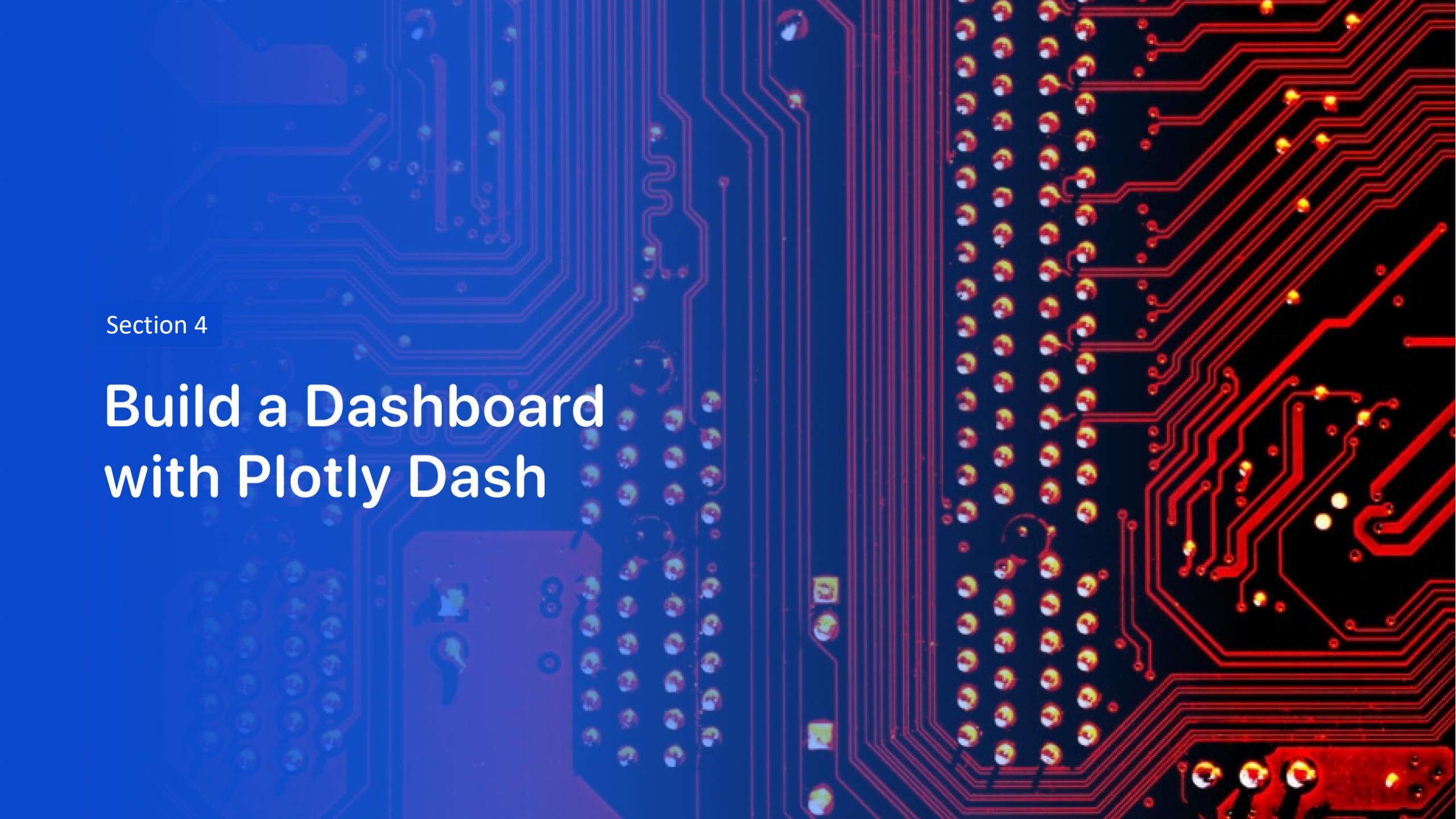
Markers showing launch sites with color labels

- Florida launch sides green makers shows successful launchers and red marker shows failures.



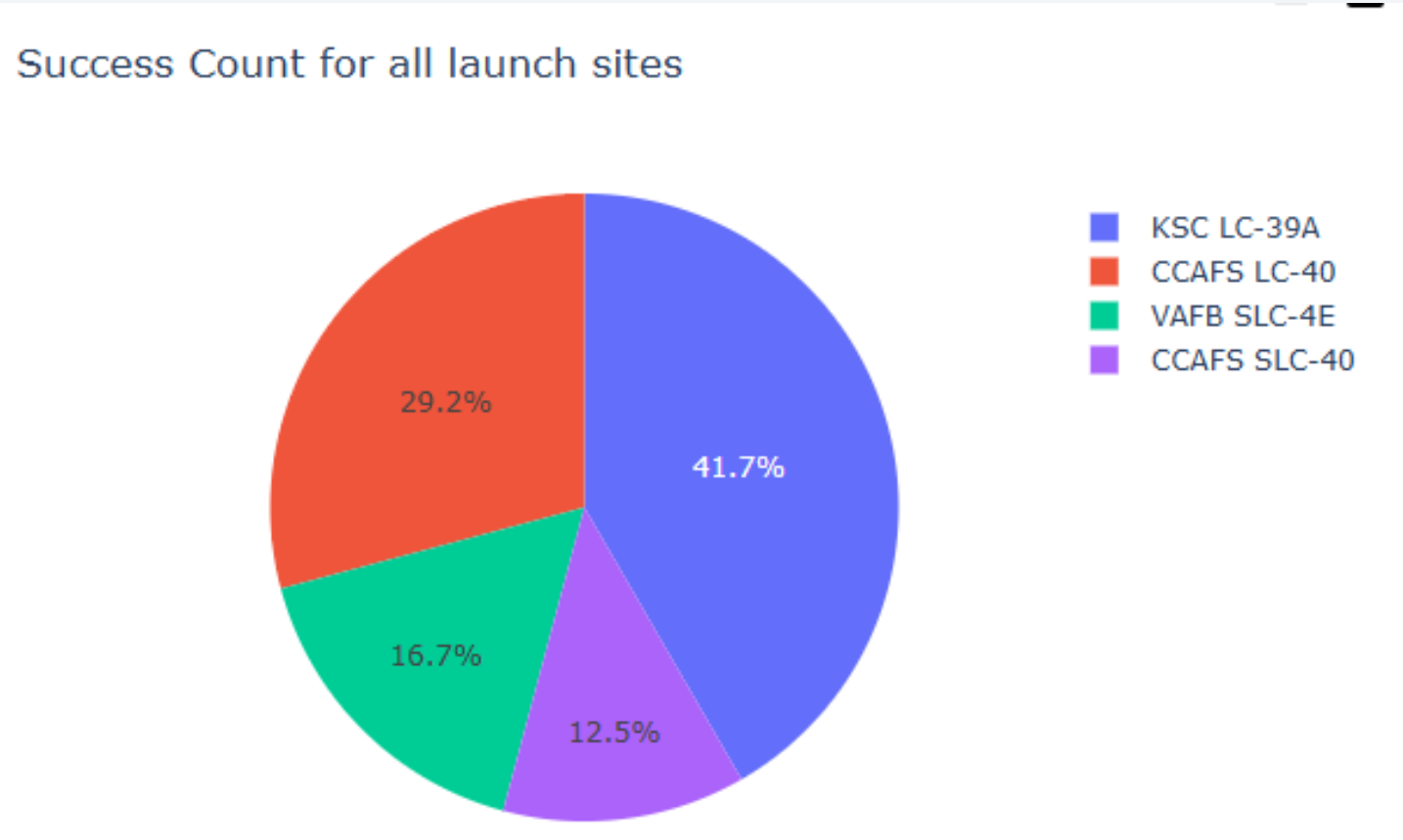
Launch Site distance to landmarks

- Distant to coast.


Section 4
Build a Dashboard with Plotly Dash

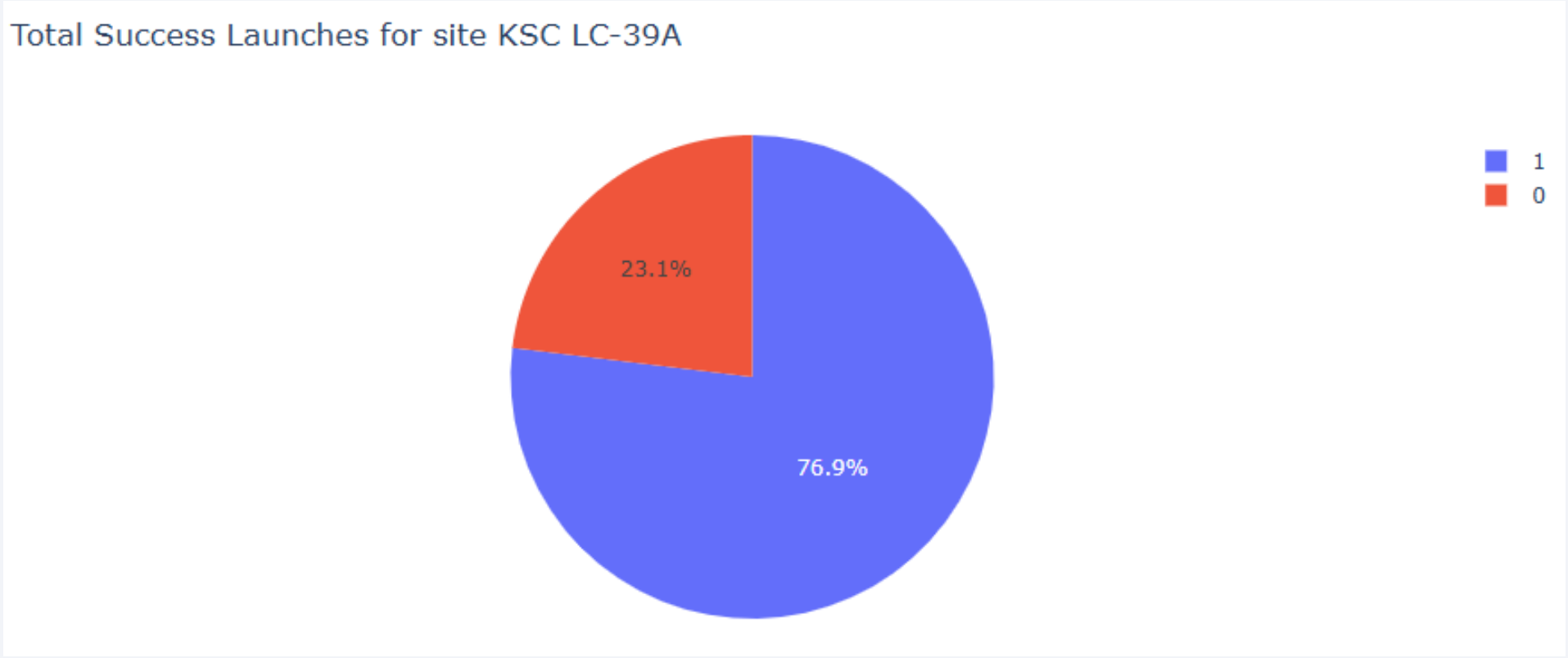
Pie chart showing the success percentage achieved by each launch site

- Total Success launches by all sites



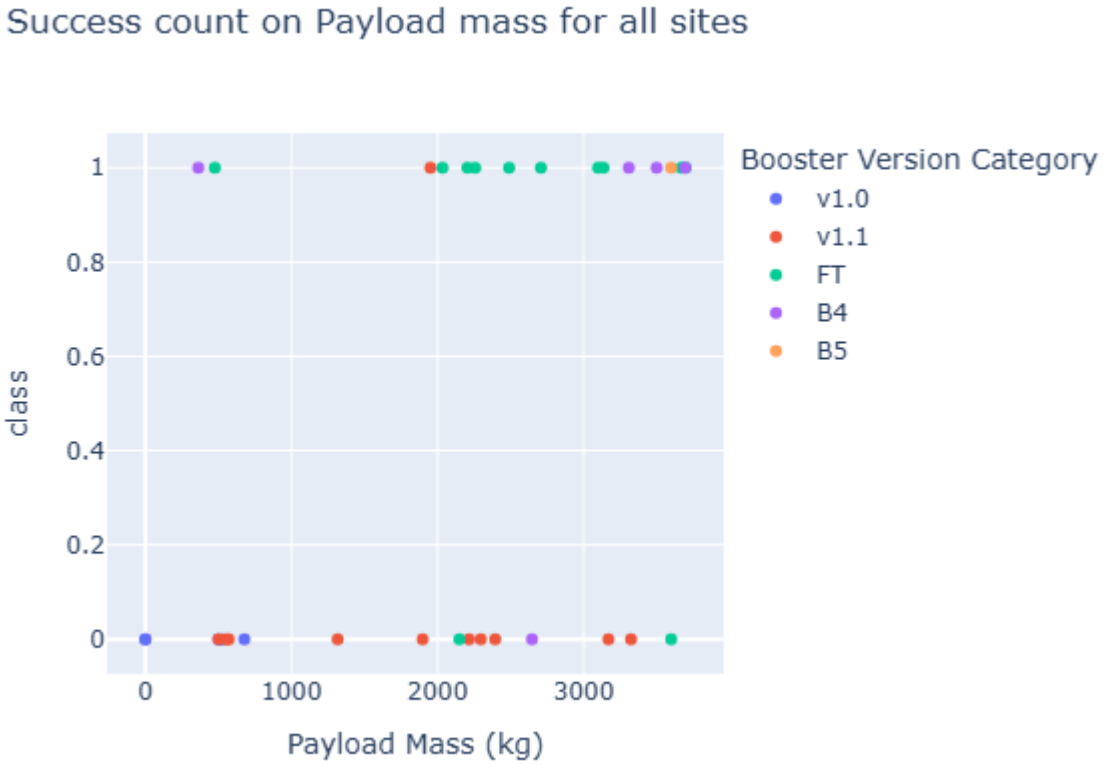
Pie chart showing the Launch site with the highest launch success ratio

- The KSC LC 39A achieved a 76.9% success rate while getting a 23,1% failure rate.

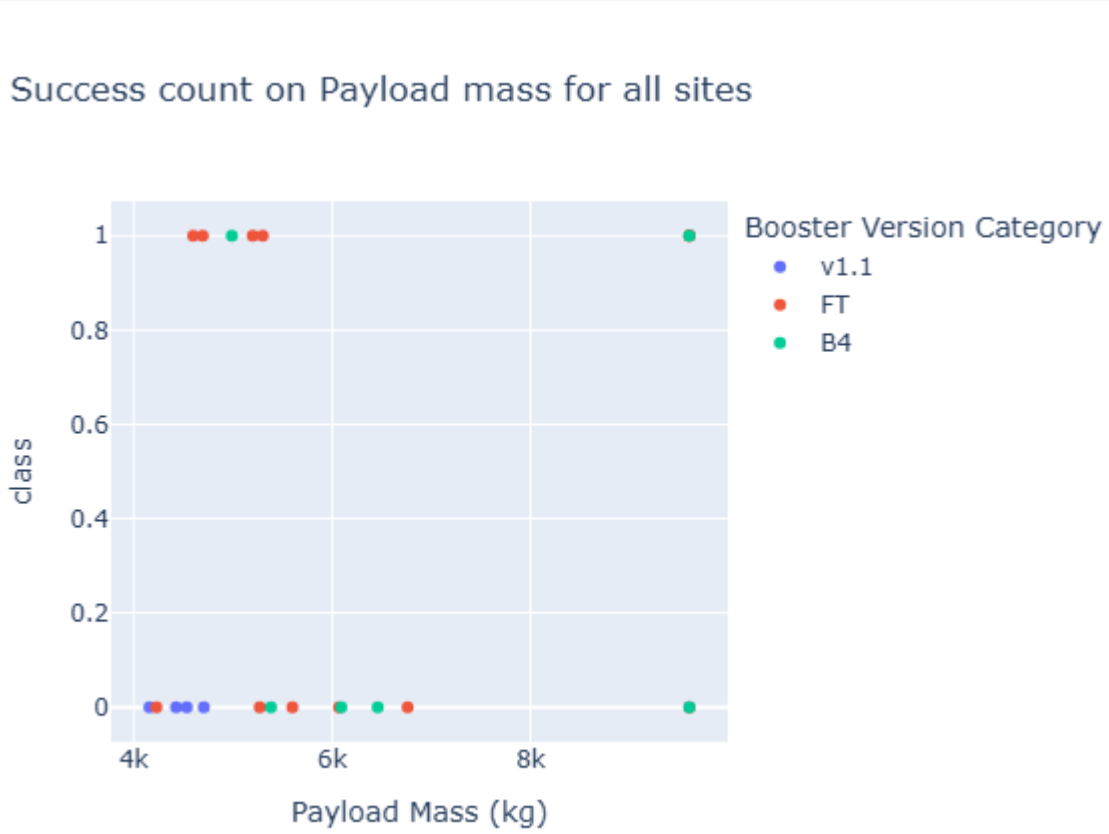


Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider

- Low Weighted Payload 0kg- 4000kg



- Heavy Weighted Payload 4000kg- 10000kg



Section 5

Predictive Analysis (Classification)

Classification Accuracy

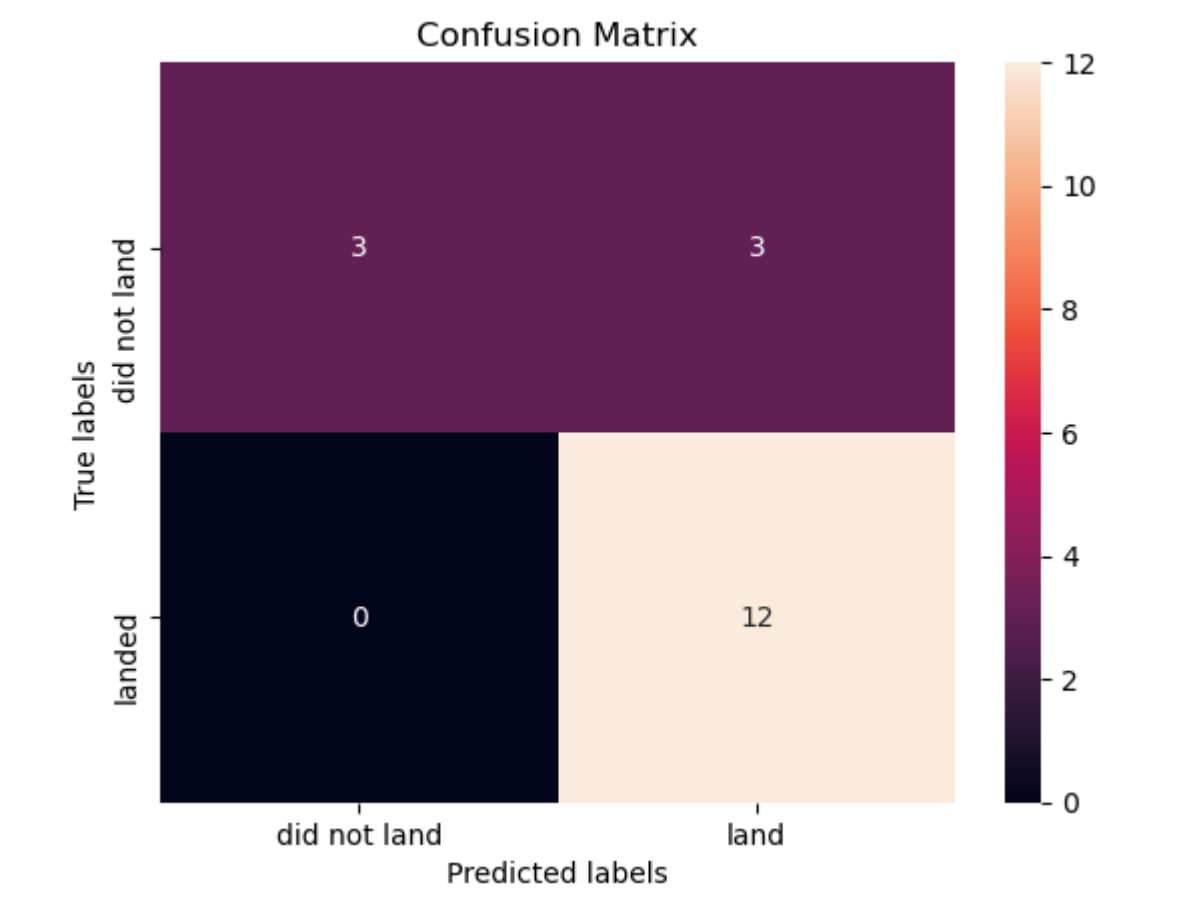
- The best models are from the SVM method because it is the one with the most precision.

```
Report = pd.DataFrame({'Method' : ['Test Data Accuracy']})
knn_accuracy=knn_cv.score(X_train, Y_train)
Decision_tree_accuracy=tree_cv.score(X_train, Y_train)
SVM_accuracy=svm_cv.score(X_train, Y_train)
Logistic_Regression=logreg_cv.score(X_test, Y_test)
Report['Logistic_Reg'] = [Logistic_Regression]
Report['SVM'] = [SVM_accuracy]
Report['Decision Tree'] = [Decision_tree_accuracy]
Report['KNN'] = [knn_accuracy]
Report.transpose()
```

	0
Method	Test Data Accuracy
Logistic_Reg	0.833333
SVM	0.888889
Decision Tree	0.861111
KNN	0.861111

Confusion Matrix

- The decision tree classifier's confusion matrix reveals that its performance is significantly hampered by false positives. The main issue is the model's tendency to identify unsuccessful landings as if they were successful.



Conclusions

We can conclude that:

- Flight Volume and Success Correlation: A positive correlation exists between a launch site's flight volume and its success rate; more launches correlate with higher reliability.
- Annual Improvement Trend: The overall launch success rate showed a consistent upward trend from 2013 through 2020.
- Most Reliable Orbits: Missions targeting the ES-L1, GEO, HEO, SSO, and VLEO orbits demonstrated the highest success rates.
- KSC Site Performance: The KSC LC-39A launch site stood out with the highest number of successful launches.
- Optimal Algorithm: For this classification task, the Decision Tree model was identified as the best-performing machine learning algorithm.

Thank you!