IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

René Kendler

August 30, 2024

Project:

Predict success/failure of SpaceX Falcon 9 first stage landings

Index

Executive Summary

Summary of methodologies

Data will be collected by using the SpaceX API to get information about past launches, including details about the rocket used, the payload delivered, the launch specifications, the landing specifications, and the landing outcome. Additionally, we will be doing web scraping from the Wikipedia web, in order to extract data from HTML tables containing logs of Falcon 9 rocket launches.

Data wrangling will be done upon the data collected so that clean and precise information for analysis can be archived.

Exploratory Data Analysis (EDA) applying SQL and data visualization (static and dynamic dashboards), were created to obtain insights about the data.

Predictive analytics through several machine learning techniques was performed to predict whether the Falcon 9 first stage lands successfully .

Summary of results

- Success rate since 2013 has improved.
- Different launch sites have different success rates. For example, CCAFS LC-40 has a success rate of 60%, while KSC LC-39A and VAFB SLC 4E have a success rate of around 77%.
- CCAFS LC-40 has a 60% success rate, but if the mass is greater than 10,000 kg, the success rate is 100%.
- Logistic regression, KNN, SVM all perform well on this dataset.

Introduction

Project background and context

A competing rocket launch company aims to predict the success or failure of SpaceX Falcon 9 rocket first stage landings.

SpaceX advertises Falcon 9 rocket launches on its website at a cost of $62 million, while other providers charge upwards of $165 million per launch. Much of SpaceX's cost savings come from the ability to reuse the first stage of their rockets. Therefore, accurately predicting whether the first stage will land successfully can help determine the overall cost of a launch. This information is valuable for any company looking to bid against SpaceX for rocket launches. The goal of this project is to develop a machine learning pipeline to predict the successful landing of the Falcon 9 first stage.

Problems you want to find answers

- What information is available regarding the outcomes of SpaceX Falcon 9's first stage landings?
- Which machine learning model would be most accurate for predicting the success of a Falcon 9 first stage landing in future launches?
- Is it likely that an upcoming Falcon 9 first stage will land successfully?
- How do different variables interact to affect the likelihood of a successful Falcon 9 first stage landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  We use the SpaceX API to obtain historical data from rocket launches. Additionally, we will do web scraping using the Python library called "Beautiful Soup" to extract data from HTML tables containing logs of Falcon 9 rocket launches, from a Wikipedia entry.

- Perform data wrangling

  This may include cleaning the data, removing null or inconsistent values, and transforming the data into a format suitable for modeling.

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

- Perform interactive visual analytics using Folium and Plotly Dash

- Perform predictive analysis using classification models

  We will use different algorithms and find the best hyperparameters for each of them. The accuracy of each model is evaluated using a confusion matrix

Data Collection

We will be using the SpaceX API, which to get data about rocket launches. provides us with information about past launches, including details about the rocket used, the payload delivered, the launch specifications, the landing specifications, and the landing outcome.

To get the data from the API, we will use the Python library called "requests" to make a GET request to a specific API URL. The response we will receive will be in JSON format, which is a structured data format. To convert this JSON into a table format that is easier to work with, we will use the "Json normalize" function.

We then cleaned the data, checked for missing values and fill with whatever needed.

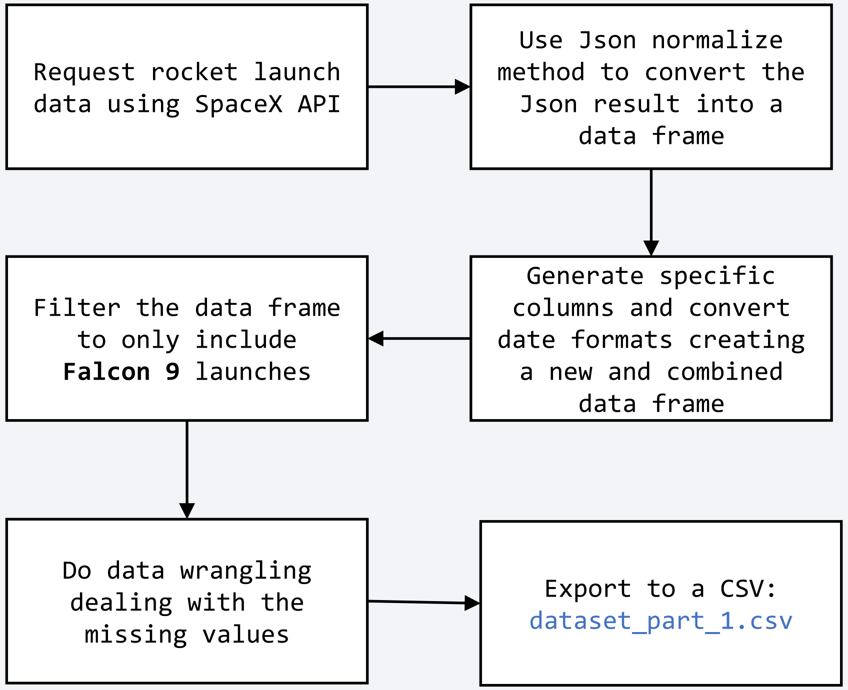
Also, we used the Python library called “Beautiful Soup” to extract data from HTML tables containing logs of Falcon 9 rocket launches from a Wikipedia entry. We will also convert this data into a table format using the “Pandas” library.

Data Collection – SpaceX API

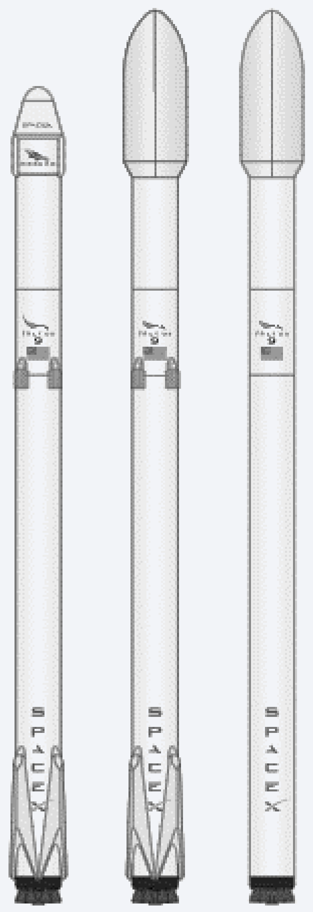
The SpaceX API provides us with information about past launches, including details about the rocket used, the payload delivered, the launch specifications, the landing specifications, and the landing outcome. Our goal is to use this data to predict whether SpaceX will attempt to land a rocket or not.

To get the data from the API, we will use the Python library called "requests" to make a GET request to a specific API URL. The response we will receive will be in JSON format, which is a structured data format. To convert this JSON into a table format that is easier to work with, we will use the "json_normalize" function.

https://github.com/ReneKendlerHering/Falcon9_Capstone/blob/main/RK-spacex-data-collection-api.ipynb



Data Collection - Scraping

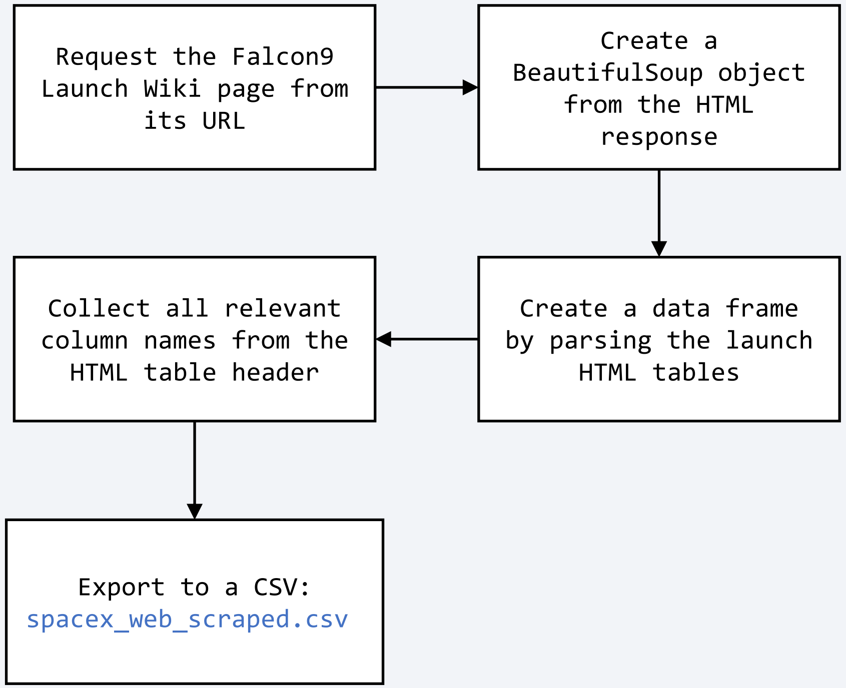


Falcon 9 v1.2 (FT)

We will search on a Wikipedia page dedicated to SpaceX (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches). It contains information such as the historical launch records about Falcon 9 rocket.

Once we have identified the relevant web page, we will use the BeautifulSoup library to parse the HTML code of those pages and extract the data we need. BeautifulSoup allows us to search for specific elements in the HTML code using CSS selectors or search methods.

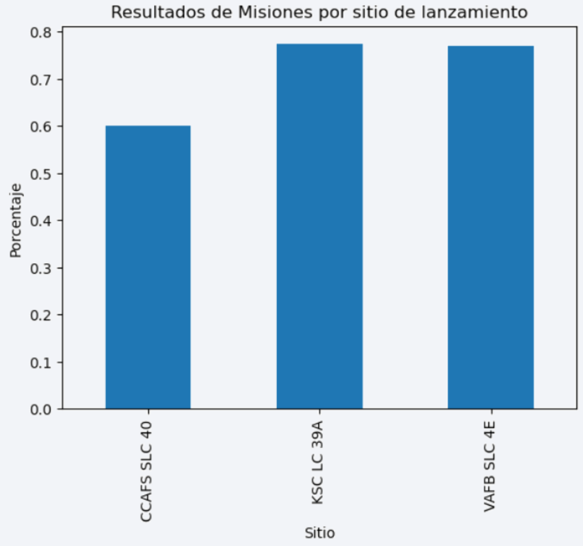
https://github.com/ReneKendlerHering/Falcon9_Capstone/blob/main/RK-spacex-data-webscraping.ipynb



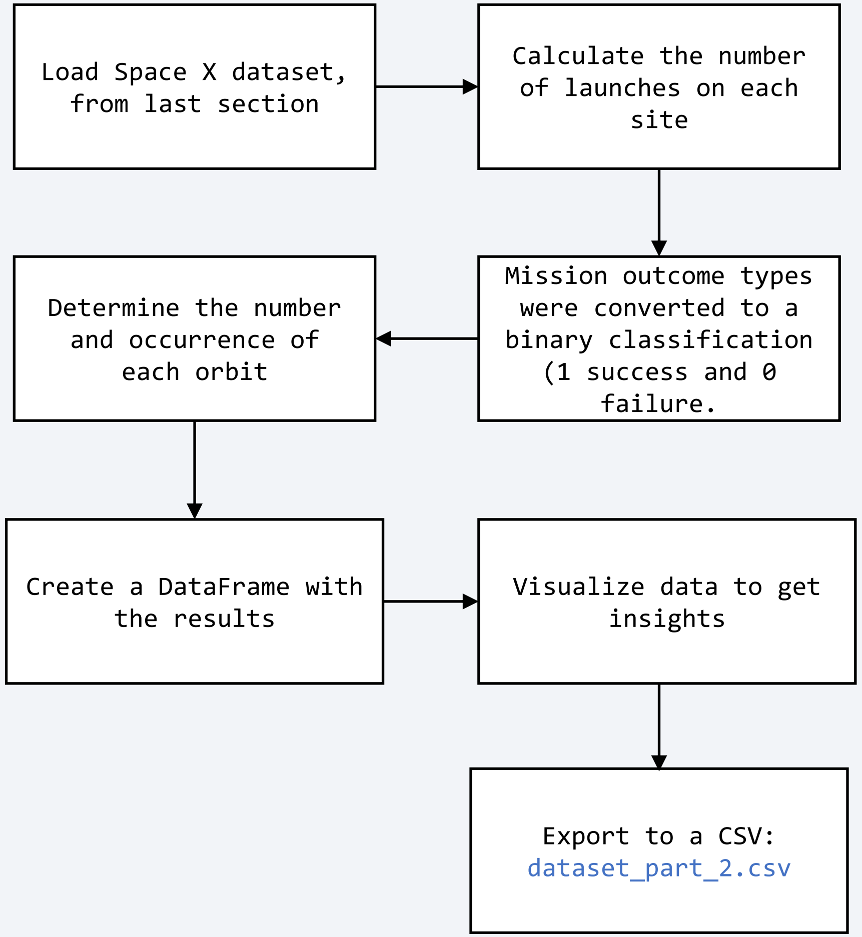
Data Wrangling

```
Número de misiones por sitio:
LaunchSite
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
dtype: int64
```

```
Proporción de misiones exitosas por sitio:
LaunchSite
CCAFS SLC 40     60.0%
KSC LC 39A      77.27%
VAFB SLC 4E     76.92%
```



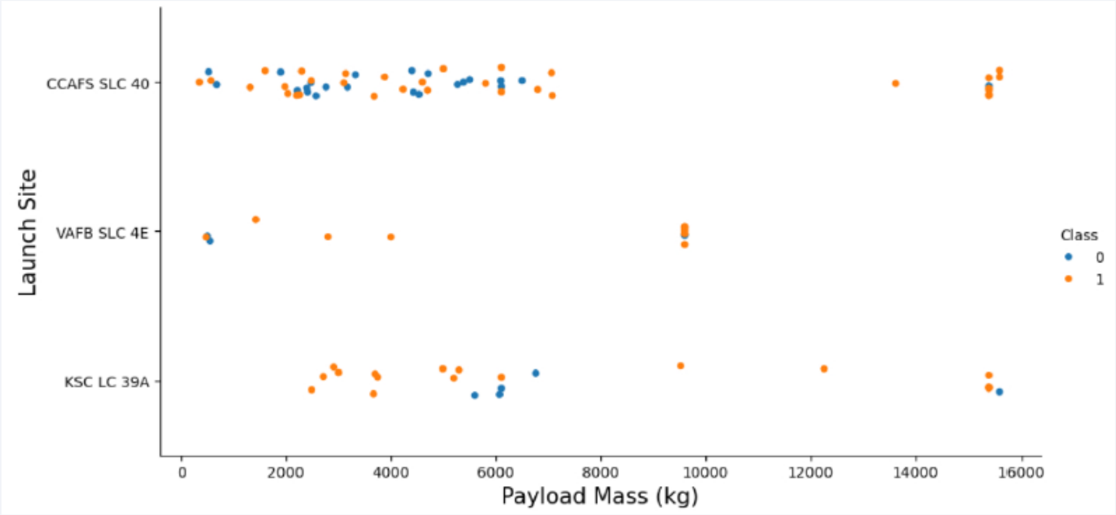
https://github.com/ReneKendlerHering/Falcon9_Capstone/blob/main/RK-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

Scatterplots: helps us to identify a possible relationship between two different sets of variables on the observed mission outcome.

Look at:	Variable X	Variable Y
Launch site trends	Launch site	Flight number
Launch site trends	Launch site	Payload
Orbit type trends	Orbit type	Flight number
Orbit type trends	Orbit type	Payload

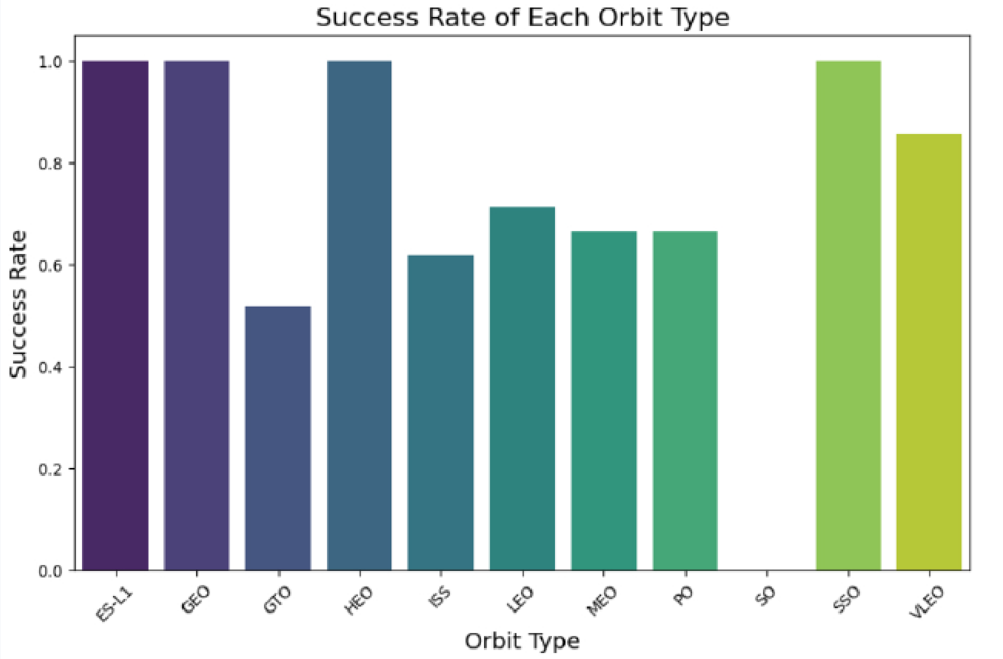


https://github.com/ReneKendlerHering/Falcon9_Capstone/blob/main/RK-spaceX-EDA-datavisualization.ipynb

EDA with Data Visualization

Barchart: From a bar chart, we can see which groups are highest or most common, and how other groups compare against the others.

Look at:	Related to:
Success rate	Orbit type

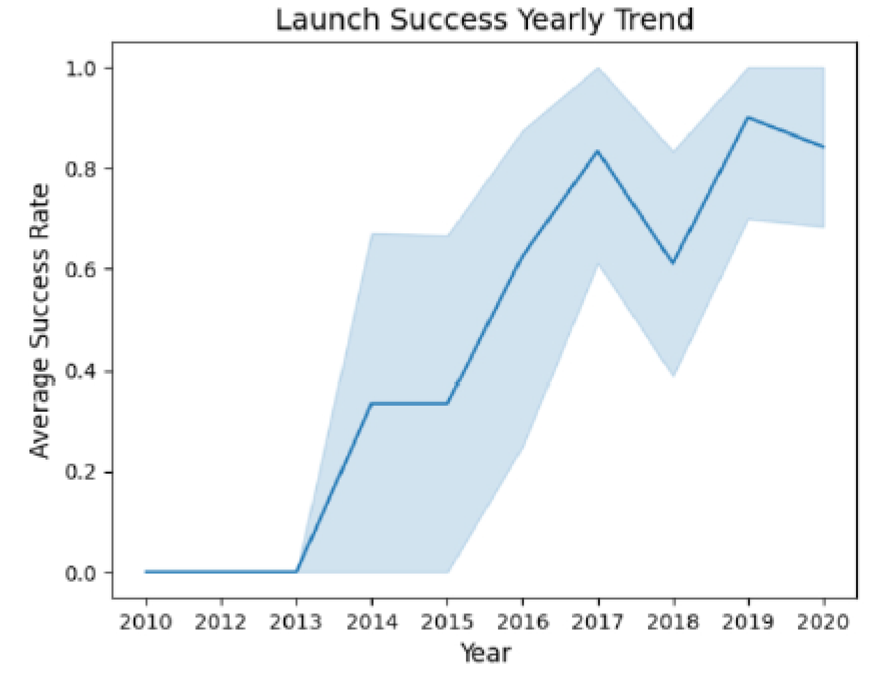


https://github.com/ReneKendlerHering/Falcon9_Capstone/blob/main/RK-spaceX-EDA-datavisualization.ipynb

EDA with Data Visualization

Linechart: These charts are useful for tracking change over time. A line chart can help make predictions about what might happen next.

Look at:	Over:
Launch success	Year



https://github.com/ReneKendlerHering/Falcon9_Capstone/blob/main/RK-spaceX-EDA-datavisualization.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display the total payload mass carried by boosters launched by NASA (CRS)
- 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 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

https://github.com/ReneKendlerHering/Falcon9_Capstone/blob/main/RK-spaceX-EDA-SQL.ipynb

FirstSuccessfulLandingDate
2015-12-22

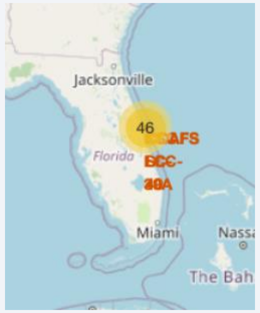
TotalPayloadMass
45596

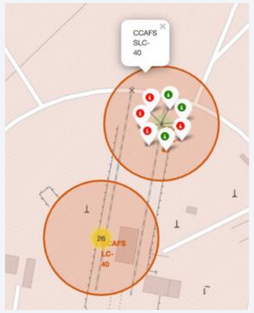
Landing_Outcome	OutcomeCount	OutcomeRank
No attempt	10	1
Success (drone ship)	5	2
Failure (drone ship)	5	2
Success (ground pad)	3	4
Controlled (ocean)	3	4
Uncontrolled (ocean)	2	6
Failure (parachute)	2	6
Precluded (drone ship)	1	8

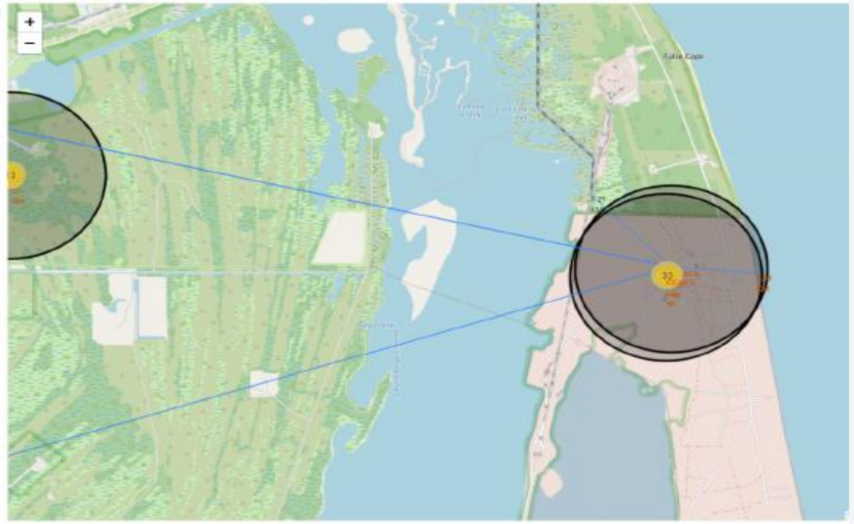
Build an Interactive Map with Folium

- General markers, were used to identify site coordinates. Also colored markers, where used to identify which launch sites have relative high success rates.
- Circle were added to show location
- Lines were added to explore and analyze the proximity of launch sites to nearby features like cities, railroads and highways.

https://github.com/ReneKendlerHering/Falcon9_Capstone/blob/main/RK-spaceX-launch-site-location-analysis-FOLIUM.ipynb





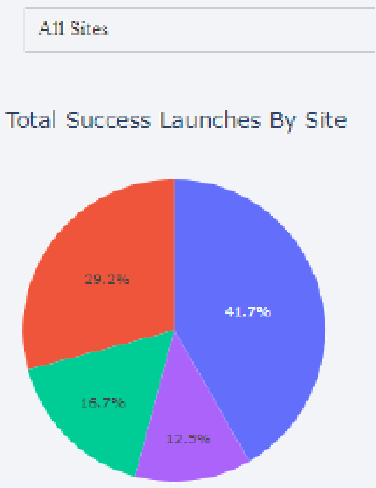



Build a Dashboard with Plotly Dash

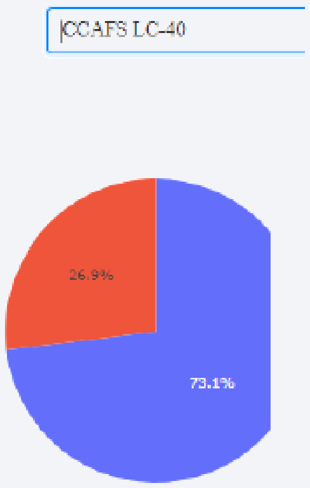
- A launch site Drop-down input was created allowing to choose between all sites or a specific one.
- To select payload amount, a ranger slider was implemented
- A pie chart was used to show success launches based on the selected site dropdown
- A scatter plot was used to show correlation between the payload and success for all sites

https://github.com/ReneKendlerHering/Falcon9_Capstone/blob/main/RK-spacex-Plotly-Dash-app.py

- Pie chart for all sites are selected



- Pie chart for is selected



Predictive Analysis (Classification)

1. Building the model

- Preprocessing (loading and standardizing)
- Train split the data (80% train, 20% test)
- The algorithms used in the model construction were: Logistic Regression, Support Vector Machines, Decision Tree Classifier and Nearest Neighbors.

2.Training and testing the models

- Train the different models and perform a Grid Search, which will allow us to find the hyperparameters that allow a given algorithm to perform best.
- Calculate the accuracy on the test data using the method score.
- Evaluate the accuracy of each model using a confusion matrix.

3.Finding the best model

- Create a report in which we can compare the results and find which methods performs best

https://github.com/ReneKendlerHering/Falcon9_Capstone/blob/main/RK-spaceX-Machine-Learning-Prediction.ipynb

Predictive Analysis (Classification)

Method	Test Data Accuracy
Logistic_Reg	0.833333
SVM	0.833333
Decision Tree	0.722222
KNN	0.833333

All models demonstrated strong performance, with the exception of the Decision Tree, which underperformed compared to the others

https://github.com/ReneKendlerHering/Falcon9_Capstone/blob/main/RK-spaceX-Machine-Learning-Prediction.ipynb

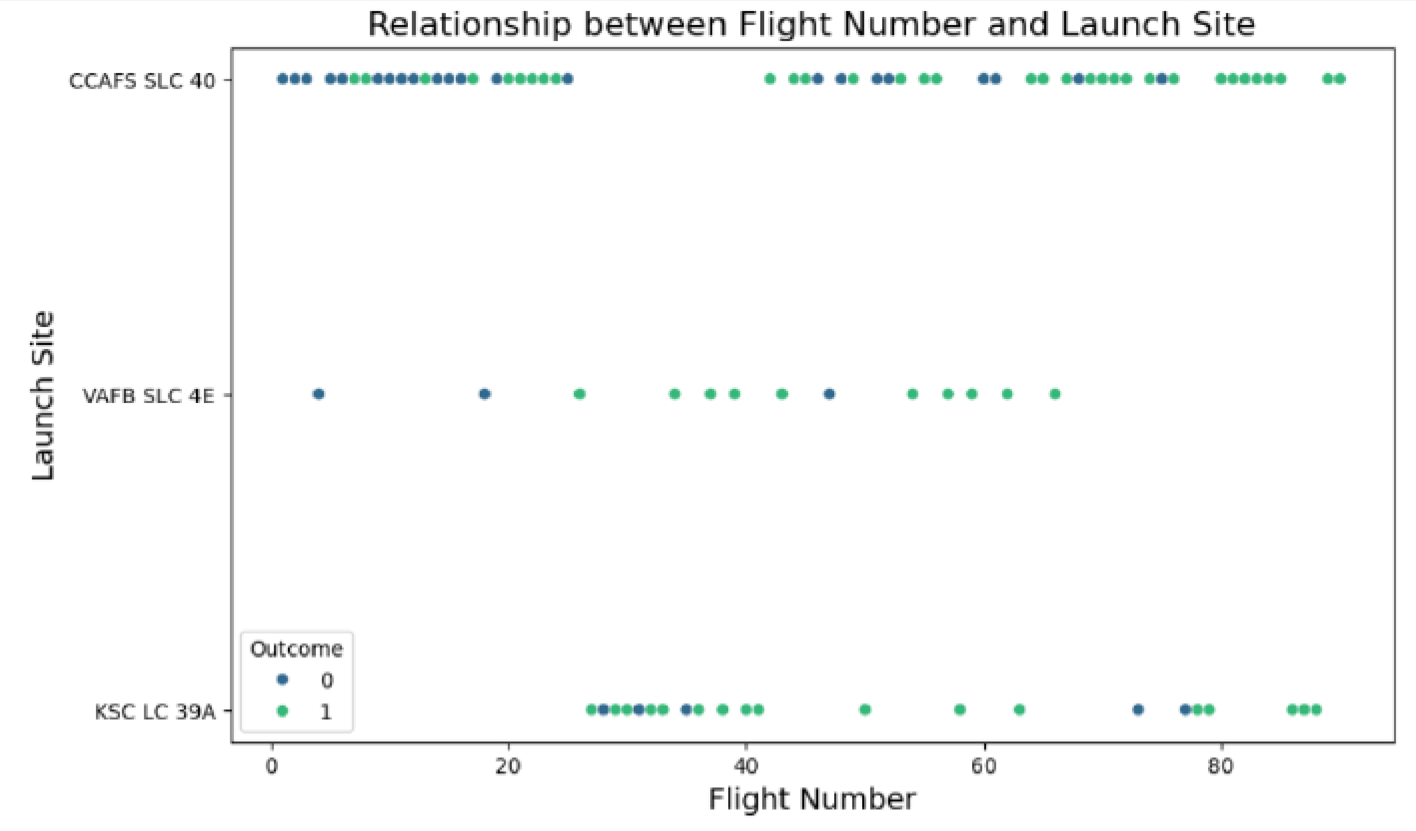
Results

- Section 2 - Insight's drawn from EDA
- Section 3 – Launch sites proximity analysis
- Section 4 – Dashboard with plotly Dash
- Section 5 - Predictive analysis

Section 2

Insights drawn from EDA

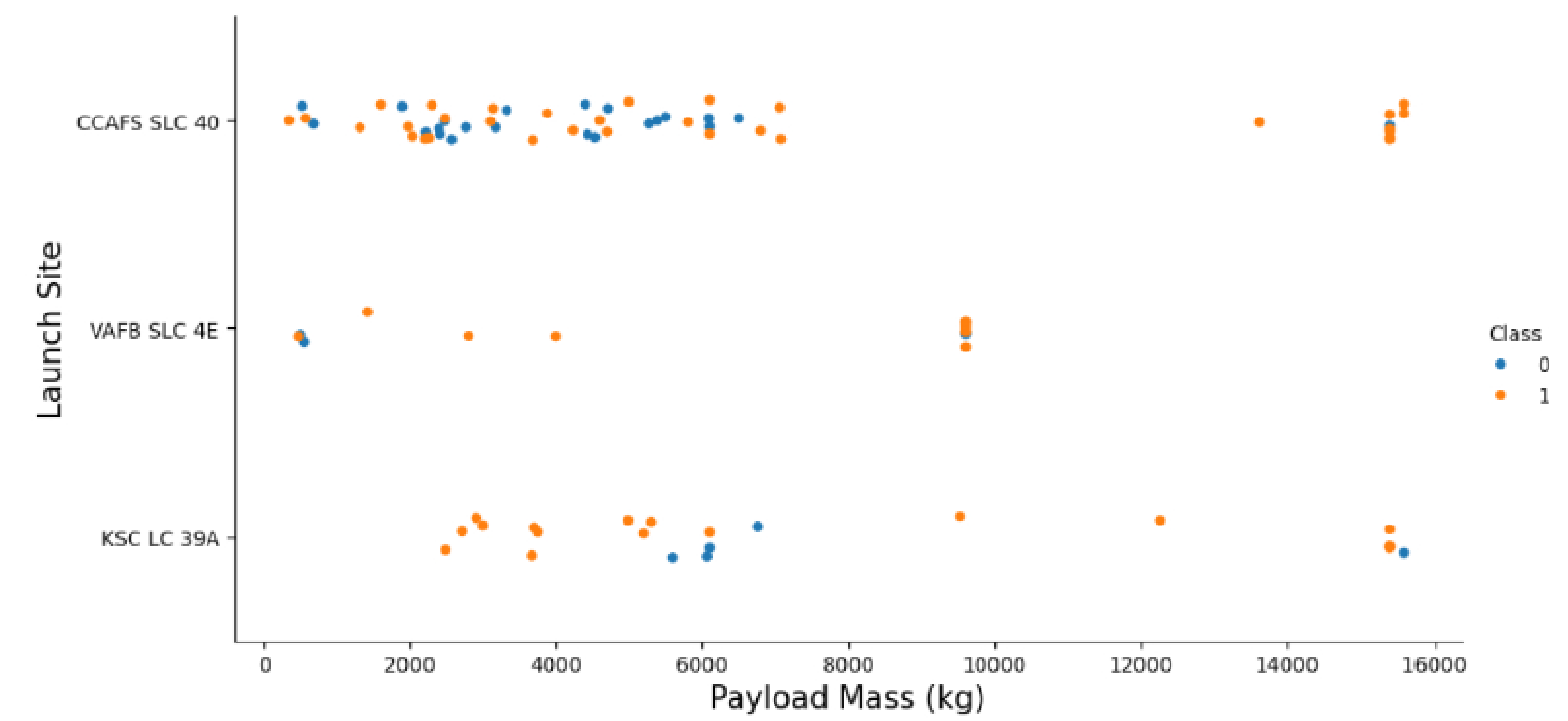
Flight Number vs. Launch Site



As the number of flights increased, the success ratio also was becoming better.

The launch site with less flights is the VAFB SLC 4E

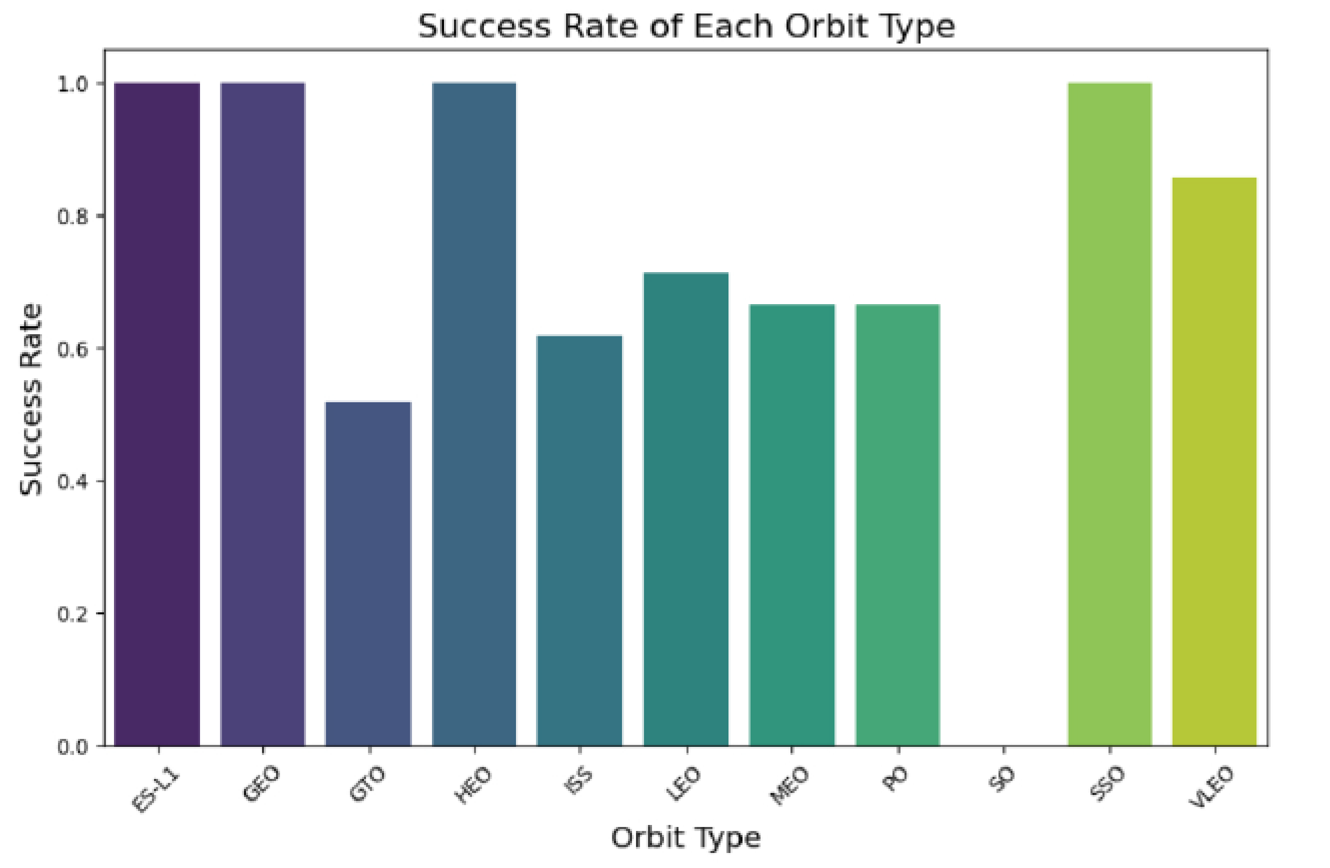
Payload vs. Launch Site



It seems that for heavy load mass (greater than 10.000), the success is good.

The launch site VAFB SLC 4E, has no rocket launched for payloads greater than 10.000.

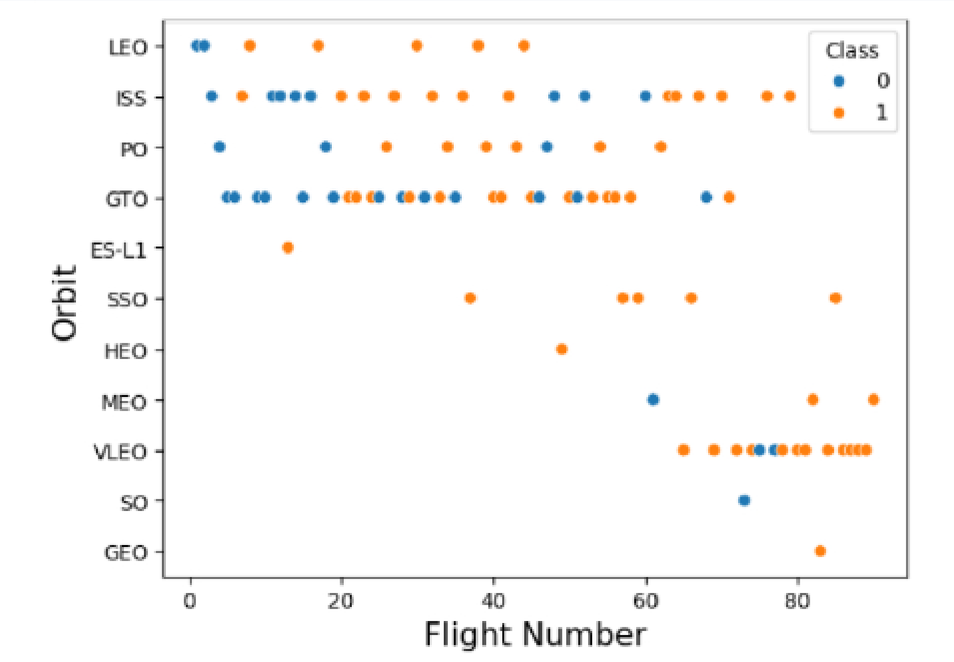
Success Rate vs. Orbit Type



The orbits with higher success rates are:
ES-L1, GEO, HEO and SSO.

This is not conclusive. Further analysis should be done.

Flight Number vs. Orbit Type

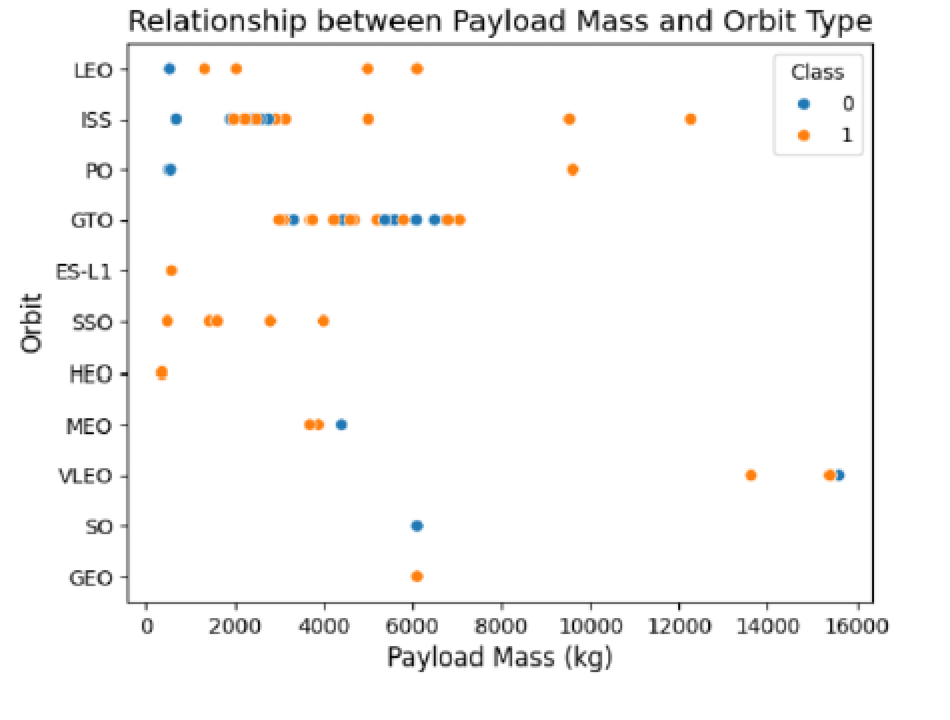


Here we can observe that ES-L1, GEO and HEO only had 1 rocket launch which was successful. More testing should be done in this orbits to draw a conclusion.

On the contrary, SSO orbit had some successful flights with no failure.

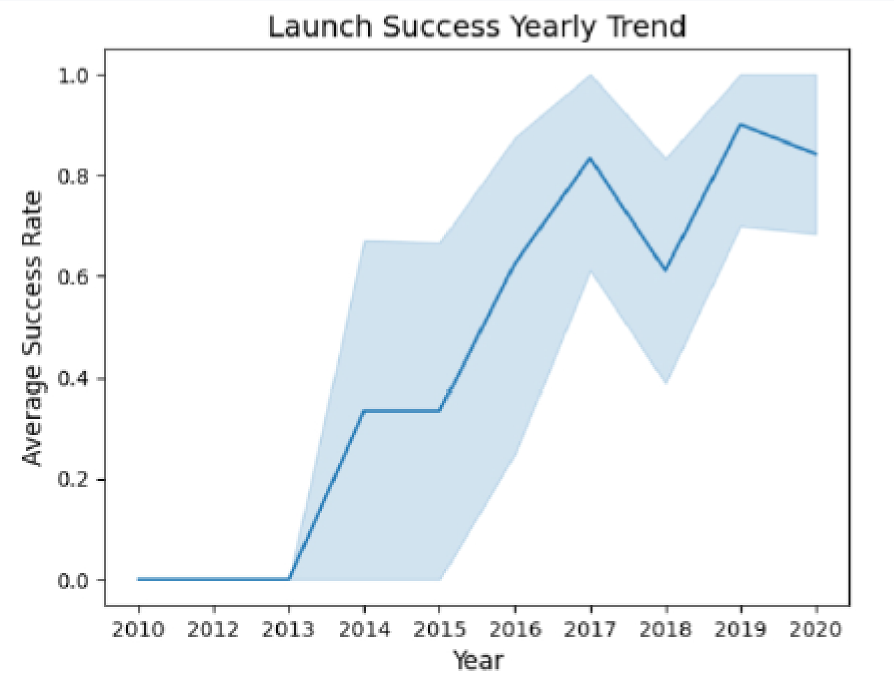
Also, LEO, ISS, PO and VLEO have some good success/failure ratios.

Payload vs. Orbit Type



We can observe that with heavy payloads, the successful landing are more for Polar, LEO and ISS orbits.

Launch Success Yearly Trend



Since 2013, success rate kept increasing.

All Launch Site Names

```
In [16]: %%sql
SELECT DISTINCT Launch_Site FROM SPACEXTABLE;

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

Out[16]:

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

Displaying the unique launch sites in the space mission

Launch Site Names Begin with 'CCA'

In [18]:
```
%%sql
SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

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

Out[18]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO
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)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)

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

Total Payload Mass

```
In [19]: %%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS TotalPayloadMass
FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)';

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

Out[19]:

TotalPayloadMass
45596

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

Average Payload Mass by F9 v1.1

In [20]:
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD
FROM SPACEXTABLE
WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

Out[20]:

AVG_PAYLOAD
2928.4

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

First Successful Ground Landing Date

In [31]:
```
%%sql
SELECT MIN(DATE) AS FirstSuccessfulLandingDate
FROM SPACEXTABLE
WHERE Landing_Outcome = "Success (ground pad)";
```

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

Out[31]:

FirstSuccessfulLandingDate
2015-12-22

Listing the date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [30]: %%sql
SELECT DISTINCT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome = "Success (drone ship)"
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;

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

Out[30]:

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

WHERE clause is used to filter for boosters which have successfully landed on drone ship and applied the **AND** condition to determine successful landing with payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

In [27]:
```
%%sql
SELECT
  COUNT(CASE WHEN Mission_Outcome LIKE 'Suc%' THEN 1 END) AS TotalSuccessful,
  COUNT(CASE WHEN Mission_Outcome LIKE 'Fai%' THEN 1 END) AS TotalFailed
FROM SPACEXTABLE;
```

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

Out[27]:

TotalSuccessful	TotalFailed
100	1

CASE command is used to select between successful and failed missions.

Boosters Carried Maximum Payload

```
In [28]: %%sql
         SELECT Booster_Version
         FROM SPACEXTABLE
         WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);

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

Out[28]:

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 mass using a subquery.

2015 Launch Records

In [37]:
```
%%sql
SELECT
  SUBSTR(Date,6,2) AS Month,
  Landing_Outcome,
  Booster_Version,
  Launch_Site
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Failure (drone ship)'
  AND SUBSTR(Date,0,5) = '2015';
```

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

Out[37]:

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 records which will display the month names, failure landing outcomes in drone ship, booster versions and launch site for the months in year 2015.

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

In [38]:
```
%%sql
SELECT
    Landing_Outcome,

    COUNT(*) AS OutcomeCount,
    RANK() OVER (ORDER BY COUNT(*) DESC) AS OutcomeRank
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome;
```

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

Out[38]:

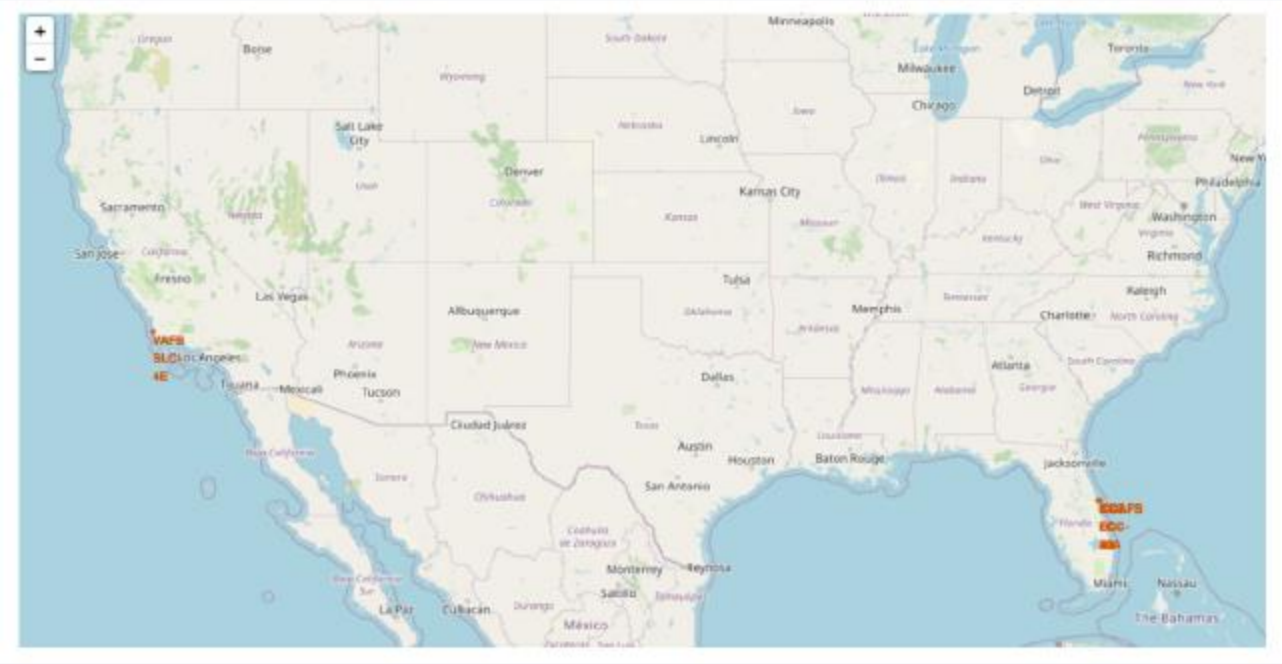
Landing_Outcome	OutcomeCount	OutcomeRank
No attempt	10	1
Success (drone ship)	5	2
Failure (drone ship)	5	2
Success (ground pad)	3	4
Controlled (ocean)	3	4
Uncontrolled (ocean)	2	6
Failure (parachute)	2	6
Precluded (drone ship)	1	8

The count of landing outcomes between the dates 2010-06-04 2017-03-20 were shown in descending order.

Section 3

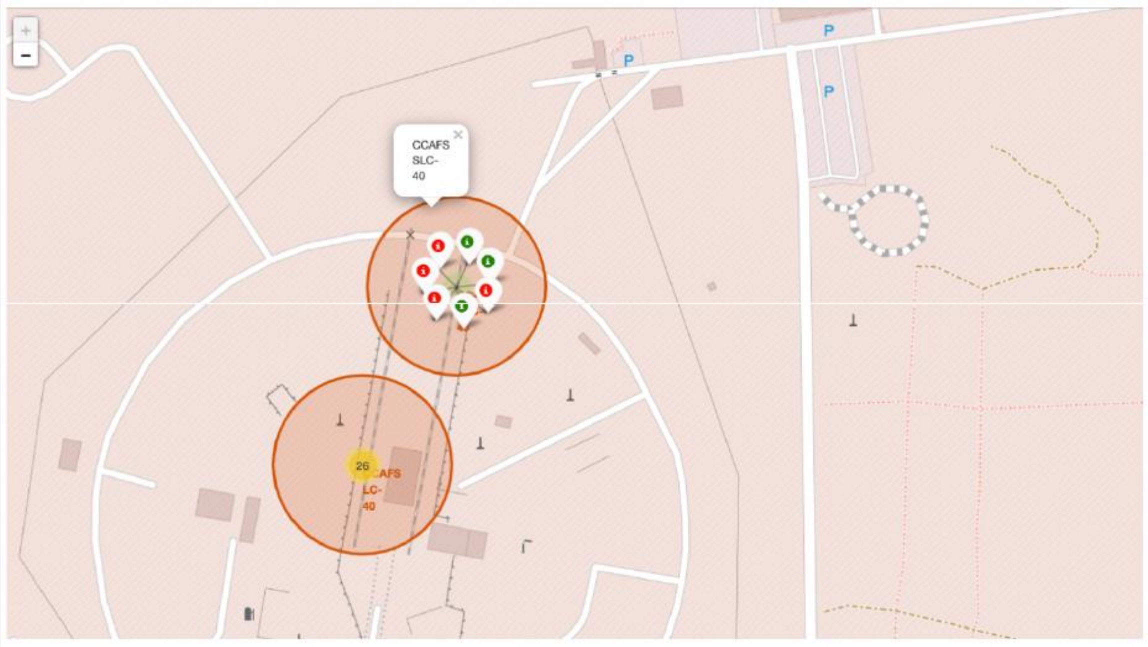
Launch Sites Proximities Analysis

United States Launch sites



We can observe launch places on both sides of the country. All of them are in proximity to the coast, providing better safety.

Using markers and circles

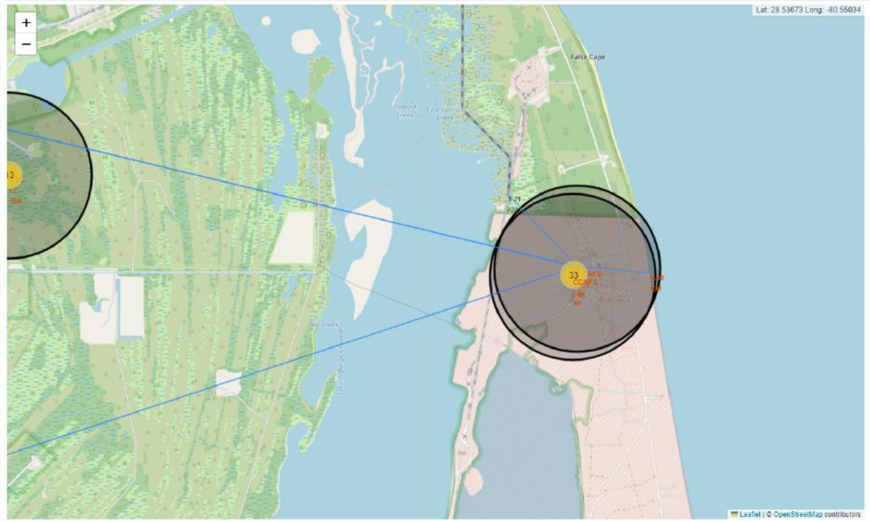


Using circles, we can pinpoint in the map the location of the launch sites.

The markers in this example, are used to:

- Show the launch pad ID
- Show the number of launches
- Show if the launch was successful (green) or unsuccessful (red)

Explore the proximities of the launchpad



A Polyline was drawn for a launch site, to the nearest coastline, city, railroad and highway coordinates.

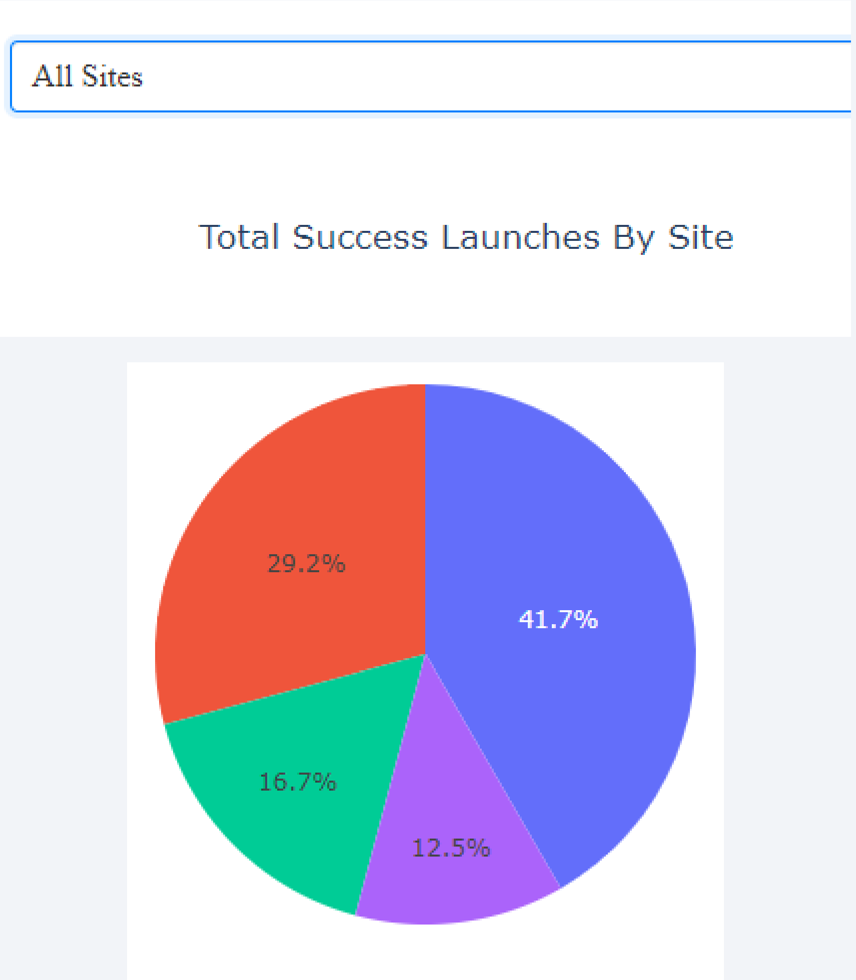
The marker, indicated the distance from this points to the launch site.

In this example, we can observe that the coastline and the railroad facilities are near the launch site, and as expected, the nearest city and highway are further away for security reasons.

Section 4

Build a Dashboard with Plotly Dash

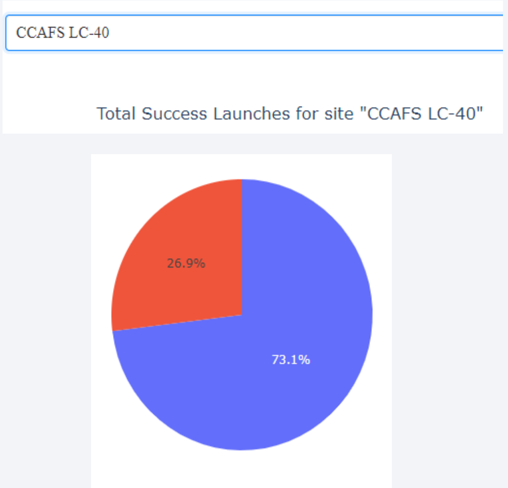
Launch success percentage by each launch site

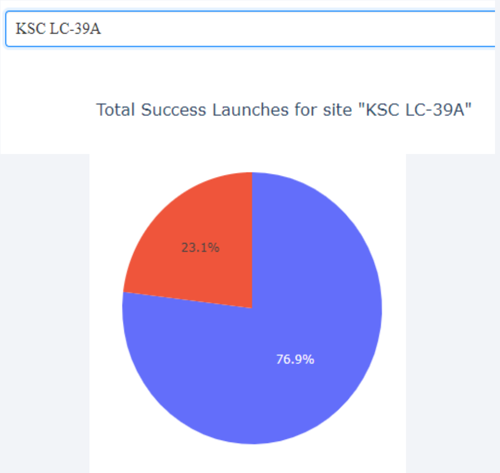


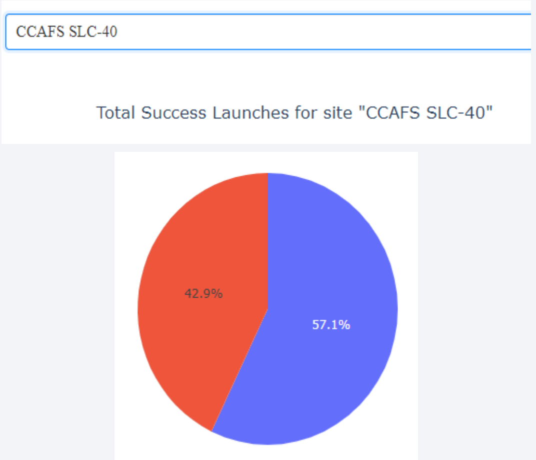
With all the launch sites selected, the pie chart, displays the distribution of success for the Falcon 9 first stage landing, among the different launch sites.

The most successful site was the KSC LC-39A with 41.7% of success.

Launch site with highest success ratio



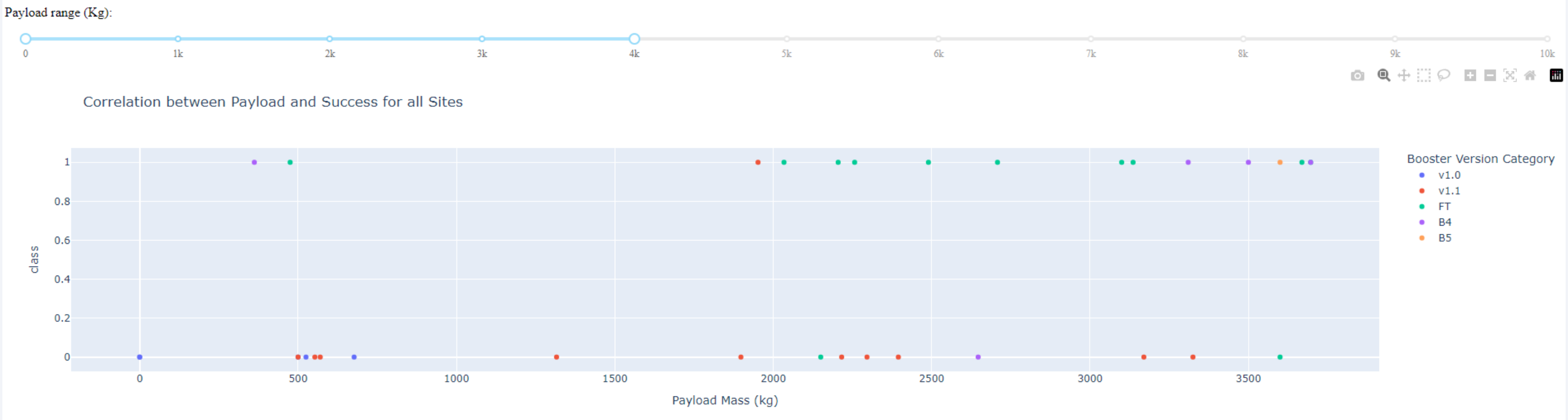




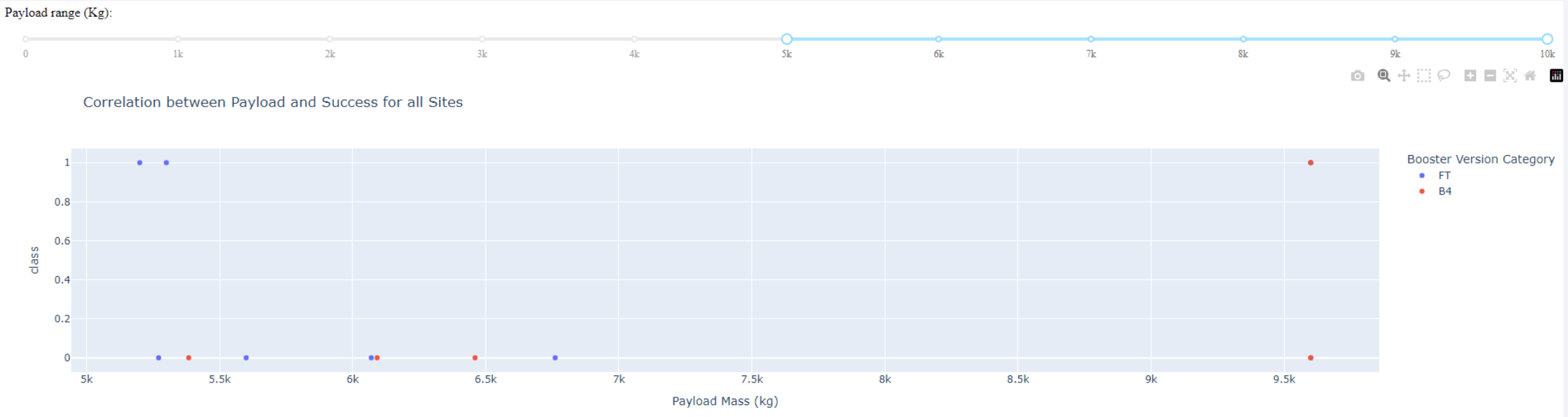
We can observe, that the launch site with the highest success ratio is the KSC LC-39A with a 76.9% of success.

Payload vs launch outcome

Less than 4k Payload



From 5k onward



The FT booster has the largest success rate.

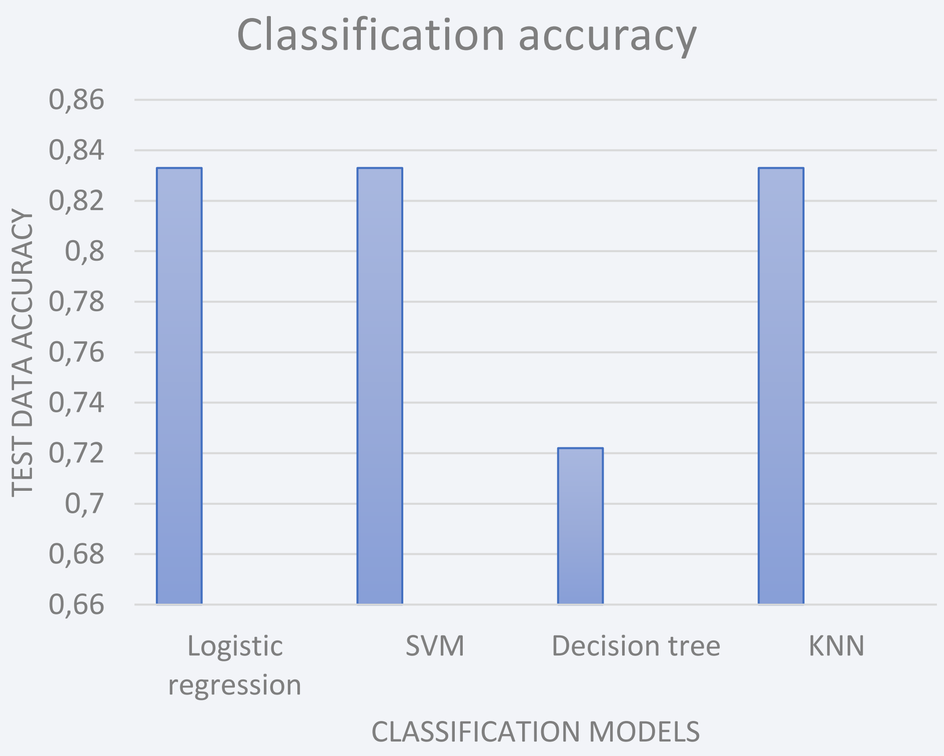
Most launches were made with the cargo loads below 4k.

Success rate is higher in this payload range.

Section 5

Predictive Analysis (Classification)

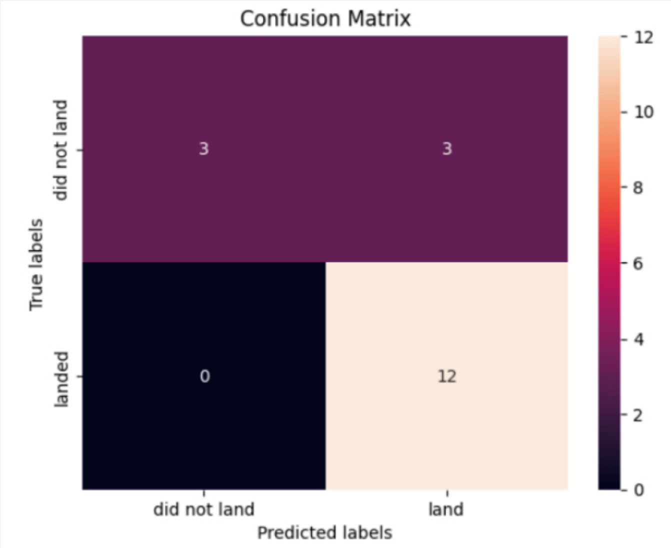
Classification Accuracy



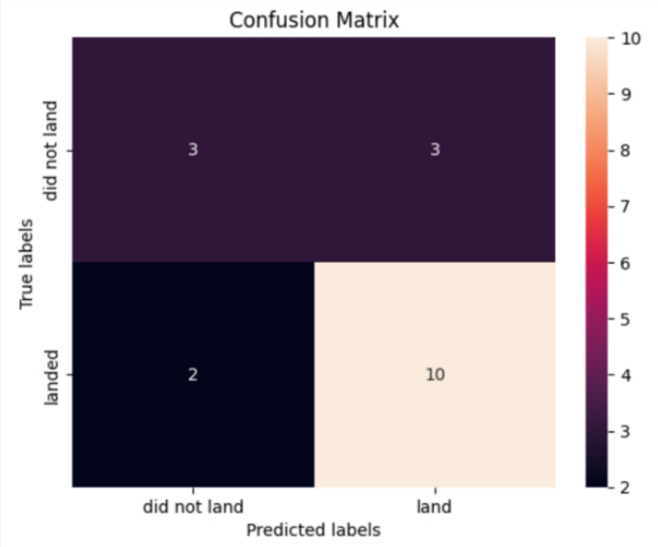
While most models exhibited satisfactory results, the Decision Tree's performance was notably inferior.

Confusion Matrix

Logistic regression, KNN and SVC



Decision tree



True negative	False positive
False negative	True positive

- Best model according to the confusion matrix are the logistic regression, KNN and SVC.
- 12 true positives and 3 true negatives
- 3 false positives and 0 false negatives

Conclusions

Conclusions

- SpaceX's Falcon 9 first state landing outcomes shows an improving success trend starting from 2013.
- KSC LC-39A has the most successful launch rate among launching sites.
- Machine learning algorithms can be used to predict first stage landing outcomes. 3 Algorithms, showed similar outcomes.
- The low weighted payloads (<4k) performed better than the heavier payloads (>4k).
- The KSC LC-39A has the highest success rate with a 76.9%.

Thank you!