IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sebastien Olive

2023-08-23

Outline

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

Executive Summary

Summary of Methodologies

The research attempts to identify the factors for a successful rocket landing. To make this determination, the following methodologies where used:

- **Collect** data using SpaceX REST API and web scraping techniques
- **Wrangle** data to create success/fail outcome variable
- **Explore** data with data visualization techniques, considering the following factors: payload, launch site, flight number and yearly trend
- **Analyze** the data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total # of successful and failed outcomes
- **Explore** launch site success rates and proximity to geographical markers • Visualize the launch sites with the most success and successful payload ranges
- **Build** Models to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor (KNN)

Results

Exploratory Data Analysis:
- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES -L1, GEO, HEO, and SSO have a 100% success rate

Visualization/Analytics:
- Most launch sites are near the equator, and all are close to the coast

Predictive Analytics:
- All models performed similarly on the test set. The decision tree model slightly outperformed

Introduction

- Our company would like to copy SpaceX’s business model, as they manage to provide space flights for 37.5% of the price of other companies. The main reason is because SpaceX can reuse the first stage of their Falcon 9 rockets.

- Therefore, if we can determine if the first stage of our rockets will land, we can determine the cost of a launch.

Section 1

Methodology

Methodology Part 1

Executive Summary

- Data collection methodology:
  - We will be working with SpaceX launch data that is gathered from an API, specifically the SpaceX REST API. This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome
- Perform data wrangling
  - Use of other API to replace IDs by actual values
  - Filter only Falcon 9 data
  - Dealing with Null values (Replaced the Payload Mass missing values with the mean
  - Encoding of all categorical data

Methodology Part 2

Executive Summary

- 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

- We collected and made sure the data is in the correct format from an API. It included:
  - Requesting to the SpaceX API
  - Cleaning the requested data

- We also scrapped from the "List of Falcon 9 and Falcon Heavy launches" wikipedia page using BeautifulSoup by:
  - Extracting a Falcon 9 launch records HTML table:
  - Parsing the table and convert it into a Pandas data frame

Data Collection – SpaceX API

Json to dataframe → **Values replacement** → **Filtering** → **Null values**

Json to dataframe

To collect our data, we requested rocket launch data from SpaceX API, decoded the response content as a Json and turned it into a dataframe.

Values replacement

We used the API a second time to replace ID values with actual data values.

Filtering

We filtered to only include Falcon 9 launches.

Null values

We replaced missing values such as in the Payload Mass columns by the mean value.

Notebook reference: https://github.com/OlSeb/SpaceX-Landing-Prediction/blob/main/1-jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

HTTP Request → **Table Extraction** → **Data frame**

HTTP Request

We requested the Falcon9 Launch Wiki page from its URL by performing an HTTP GET method as an HTTP response, and created a BeautifulSoup object from it.

Table Extraction

We extracted all column/variable names from the HTML table header after finding all the tables, then we iterated through the elements and extracted column names one by one.

Data frame

We created a data frame by parsing the launch HTML tables.

Notebook reference: https://github.com/OlSeb/SpaceX-Landing-Prediction/blob/main/2-jupyter-labs-webscraping.ipynb

Data Wrangling

Launches sites

We found out there were 55 launches on CCAFS SLC 40, 22 on KSC LC 39A, and 13 on VAFB SLC 4E.

Orbits

The 3 orbits targeted the most were GTO, ISS and VLEO.

Landing outcomes

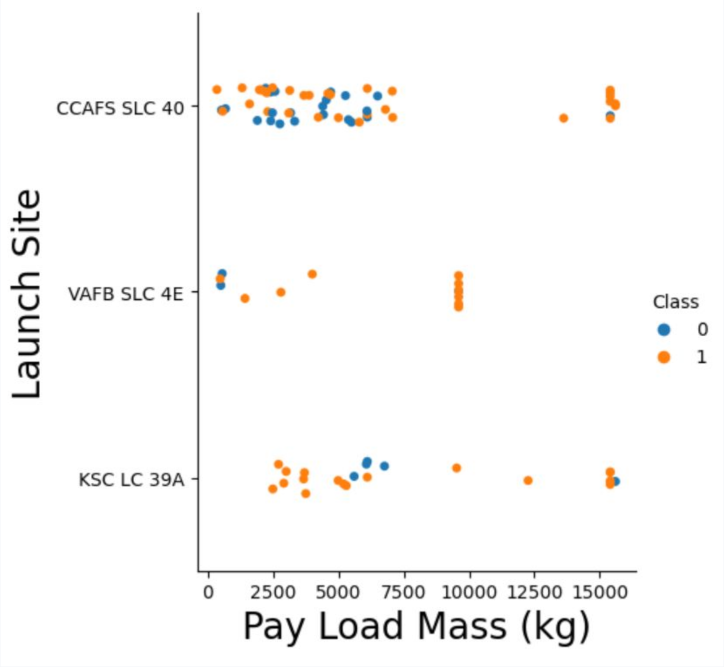
Most landings were made on drone ships and ground pads. However, 22 landings failed.

Success Rate

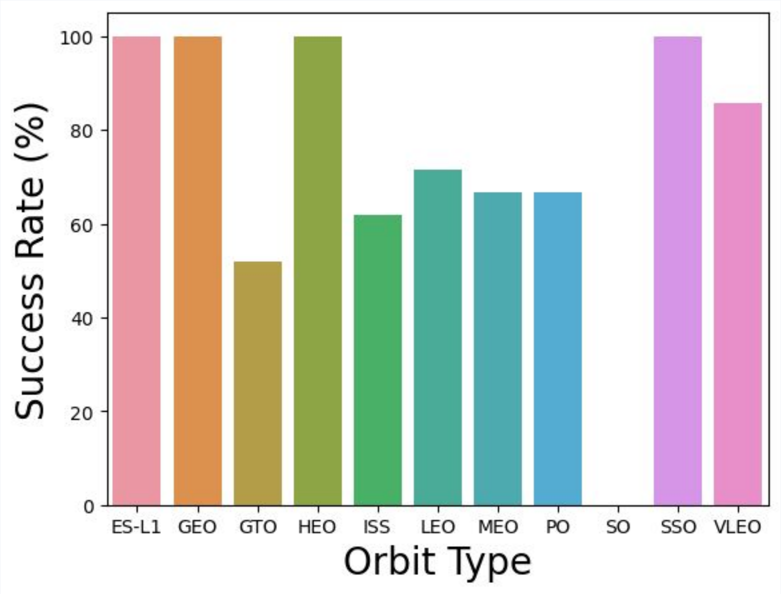
After manipulating the possible outcomes, we were able to assess that 2 out of 3 launches are successful.

Notebook reference: https://github.com/OlSeb/SpaceX-Landing-Prediction/blob/main/3-IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

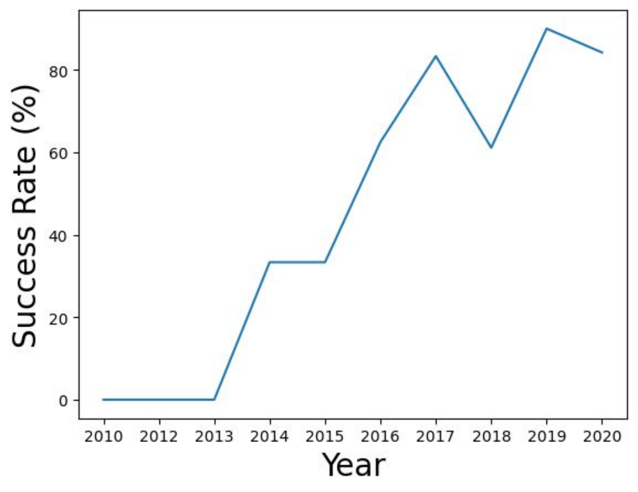
EDA with Data Visualization



The VAFB SLC 4E launch site has the highest success rate butdoes not have heavy payload mass.



- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO



The success rate increase heavily from 2013, with a peak at 85% in 2019.

Notebook reference: https://github.com/OlSeb/SpaceX-Landing-Prediction/blob/main/5-IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- SELECT DISTINCT Launch_Site FROM SPACEXTBL;
  -> Launch Sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40

- SELECT AVG(PAYLOAD_MASS__KG_), Booster_Version FROM SPACEXTBL
  WHERE Booster_Version == "F9 v1.1";
  -> Average payload mass of the F9 v1.1 booster: 2928.4 kgs

- SELECT MIN(Date), * FROM SPACEXTBL
  WHERE Landing_Outcome = "Success (ground pad)";
  -> First successful landing on a ground pad on 2015-12-22

- SELECT Landing_Outcome, COUNT(Landing_Outcome)
  FROM SPACEXTBL
  GROUP BY Landing_Outcome HAVING date
  BETWEEN "2010-06-04" AND "2017-03-20"
  ORDER BY COUNT(Landing_Outcome) DESC

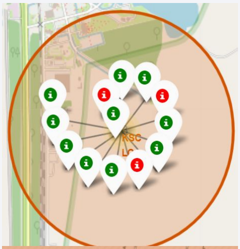
Landing_Outcome	COUNT(Landing_Outcome)
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Notebook reference: https://github.com/OlSeb/SpaceX-Landing-Prediction/blob/main/4-jupyter-labs-eda-sql-coursera_sqllite.ipynb

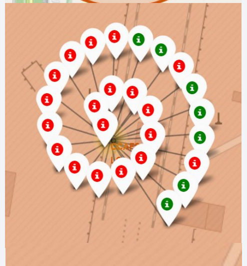
Build an Interactive Map with Folium

All launch sites are on the coastlines of the USA, each within only few kilometers from the ocean. There are however apart from cities and highways. They are also near the equator.

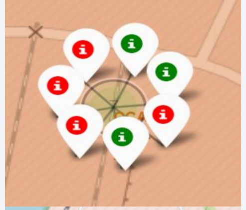
Success/Fail launches per site



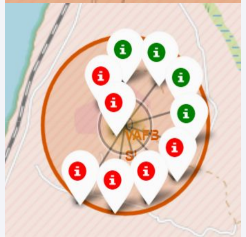
KSC
LC-39A



CCAFS LC-40



CCAFS
SLC-40

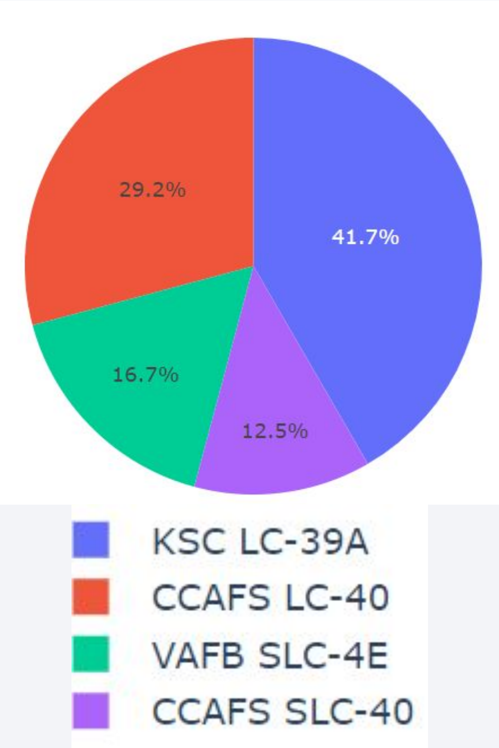


KSC
LC-39A

Notebook reference: https://github.com/OlSeb/SpaceX-Landing-Prediction/blob/main/6-IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

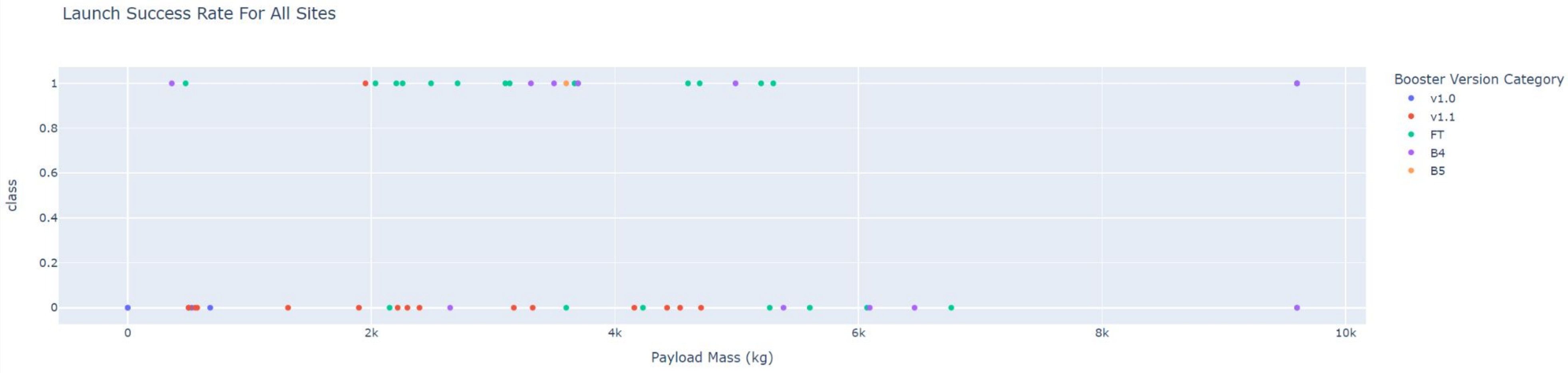
Build a Dashboard with Plotly Dash

Success Count
for all launch sites



As seen below, We note that The FT rocket version has the best success/failure proportion

Above 6 tons, the chance of success is small.



Notebook reference: https://github.com/OlSeb/SpaceX-Landing-Prediction/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Standardized the data

Train/Test Split

GridSearch for 4 models
Tree model gave us 90% accuracy.

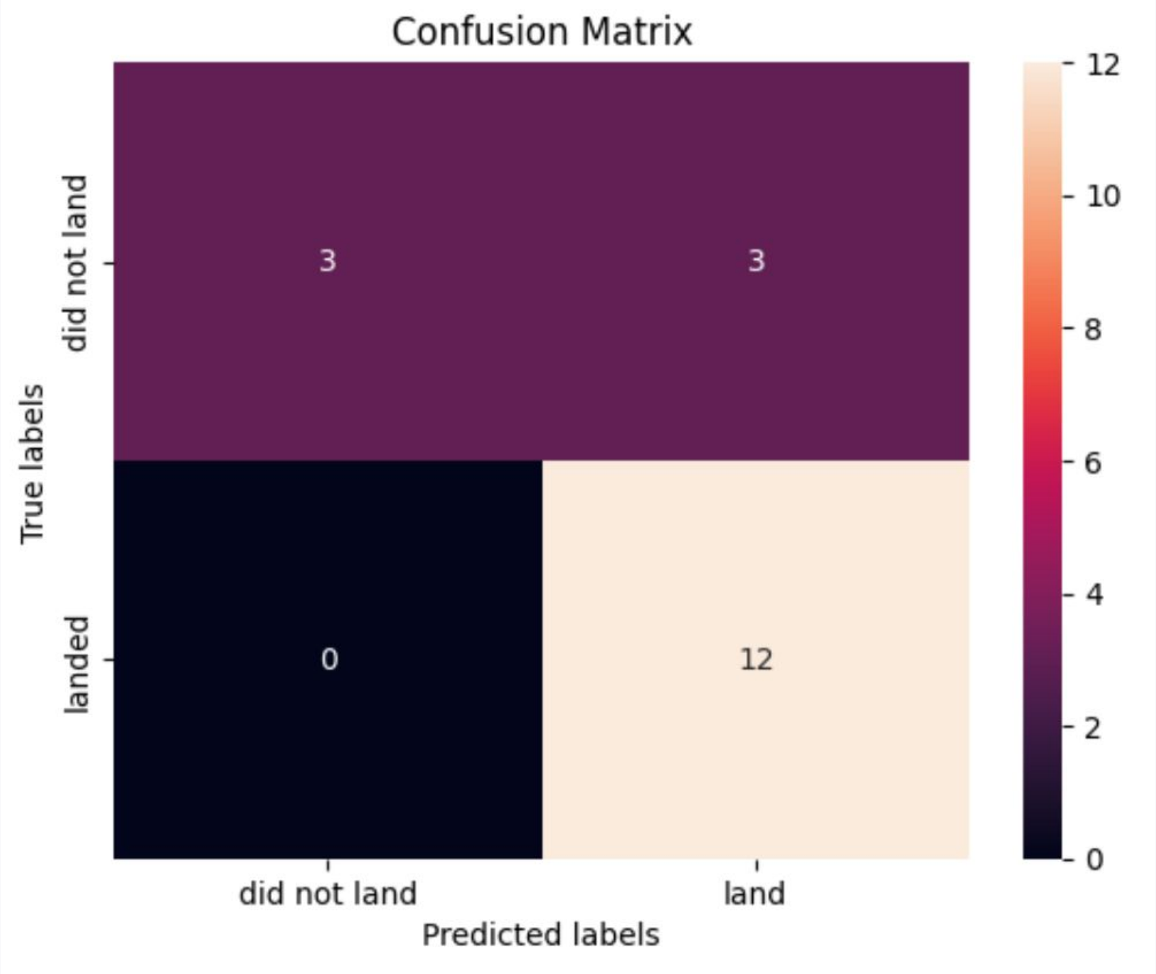
Evaluate the accuracy
All the models performed similarly. The decision tree was however slightly better.

Confusion matrix to visualize and compare
Tree model only gave us 2 negatives.

Model Selection
Tree model shows better performance

Notebook reference: https://github.com/OlSeb/SpaceX-Landing-Prediction/blob/main/7-SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Predictive Analysis (Classification)



- All the confusion matrices were identical.
- The fact that there are false positives (Type 1 error) is not good.
- Confusion Matrix Outputs:
  - 12 True positive,
  - 3 True negative,
  - 3 False positive,
  - 0 False Negative.

Notebook reference: https://github.com/OlSeb/SpaceX-Landing-Prediction/blob/main/7-SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

Exploratory Data Analysis:
• Launch success has improved over time
• KSC LC-39A has the highest success rate among landing sites
• Orbits ES -L1, GEO, HEO, and SSO have a 100% success rate

Visualization/Analytics:
• Most launch sites are near the equator, and all are close to the coast

Predictive Analytics:
• All models performed similarly on the test set. The decision tree model slightly outperformed

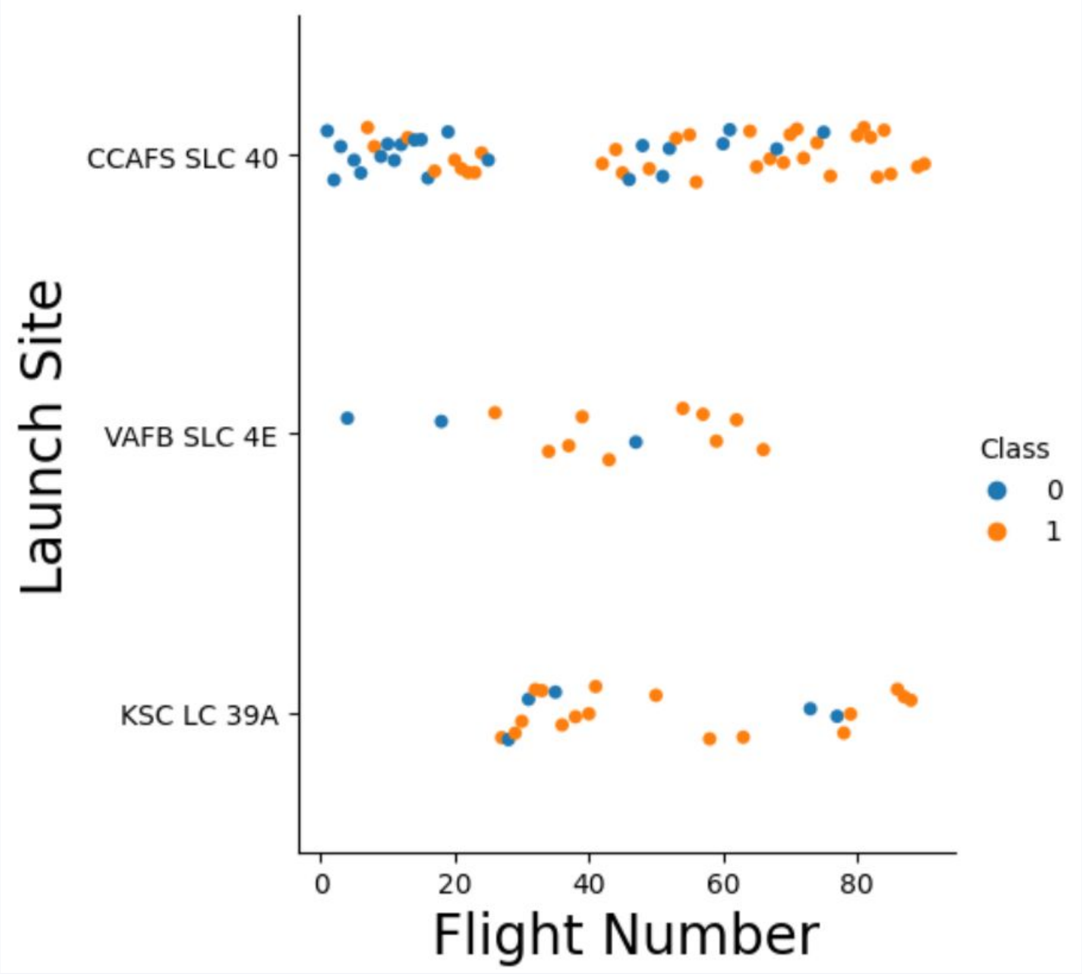
Section 2

Insights drawn from EDA

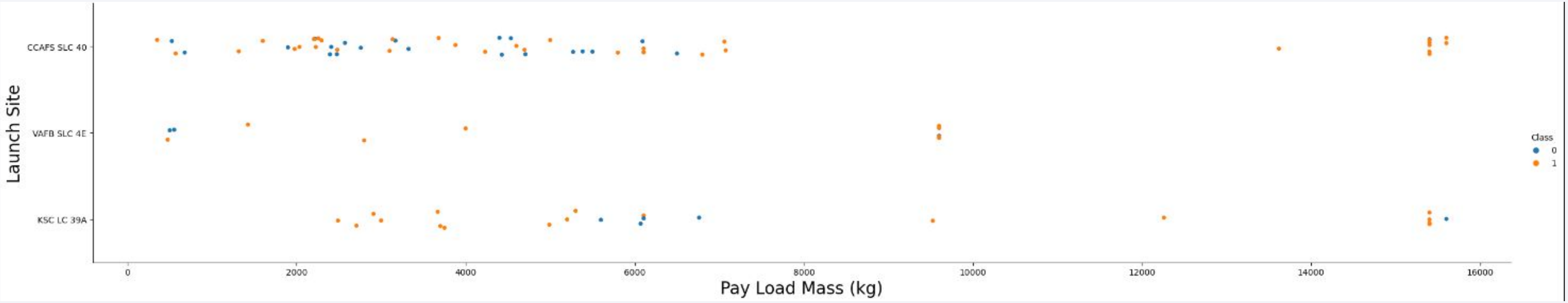
Flight Number vs. Launch Site

We note that between launch ~25 and ~42, the main site was the Kennedy Space Center instead of Cape Canaveral.

The success rate improved from the ~18th launch.



Payload vs. Launch Site



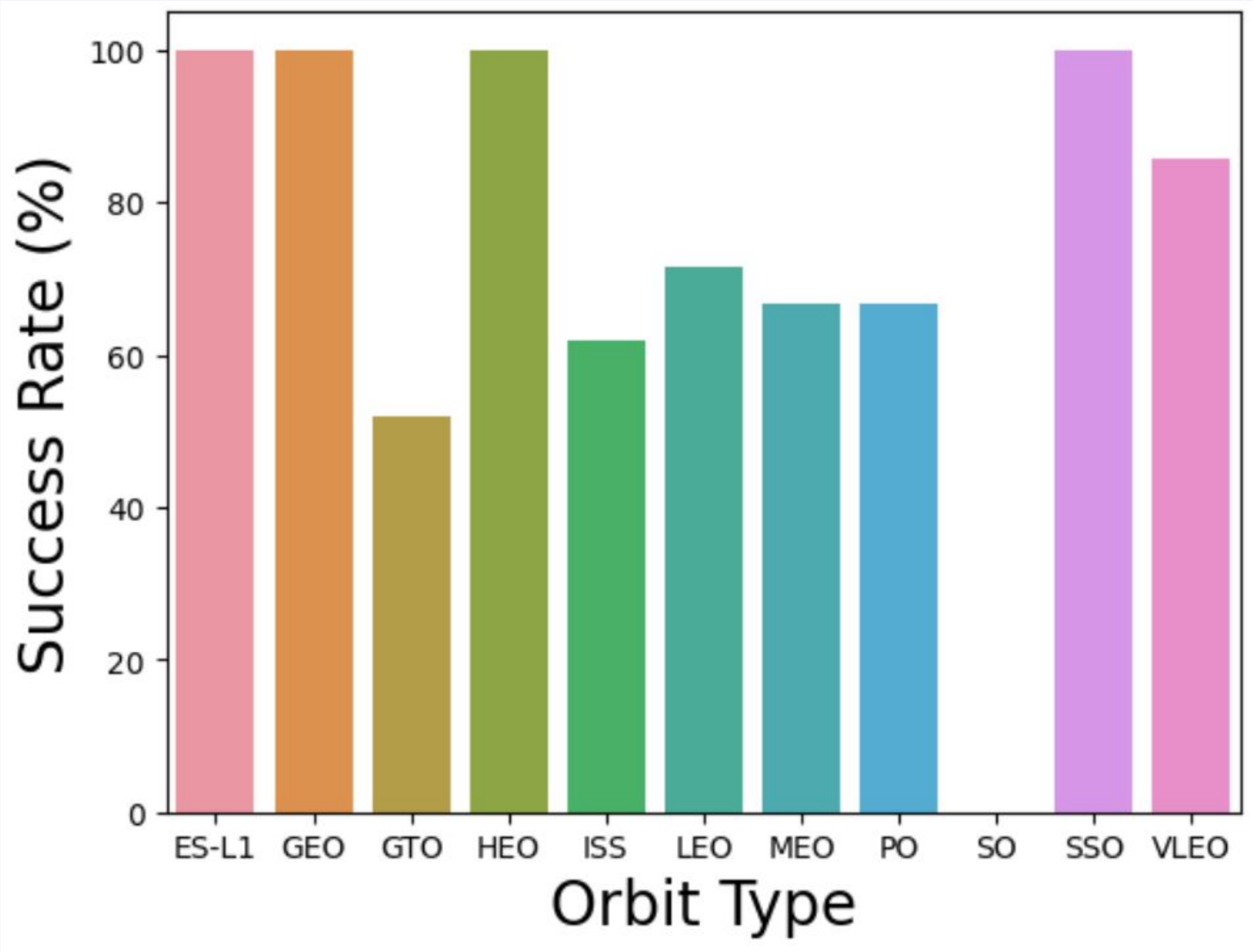
- The Vandenbrg Space Launch Complex doesn't have any launch for payload above 10 000 kg.
- The success rate is extremely high for payload of 7000 kg and higher.

Success Rate vs. Orbit Type

- 100% Success Rate: ES-L1, GEO, HEO and SSO

- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO

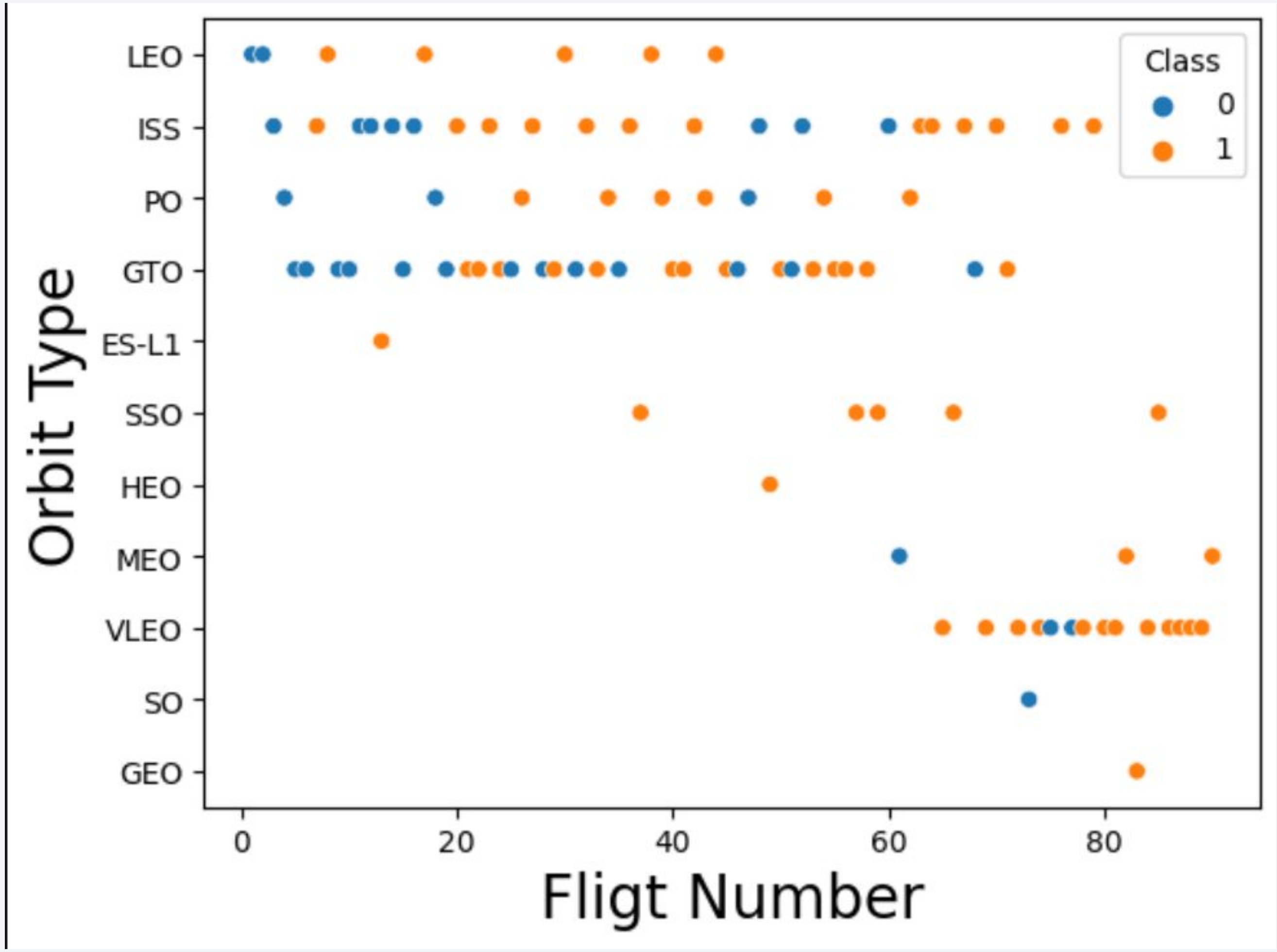
- 0% Success Rate: SO



Flight Number vs. Orbit Type

- The main orbit types were ISS and GTO until flight 65 were we mainly launched for VLEO orbit.

- Only the GTO orbit doesn’t show improvement in succes.

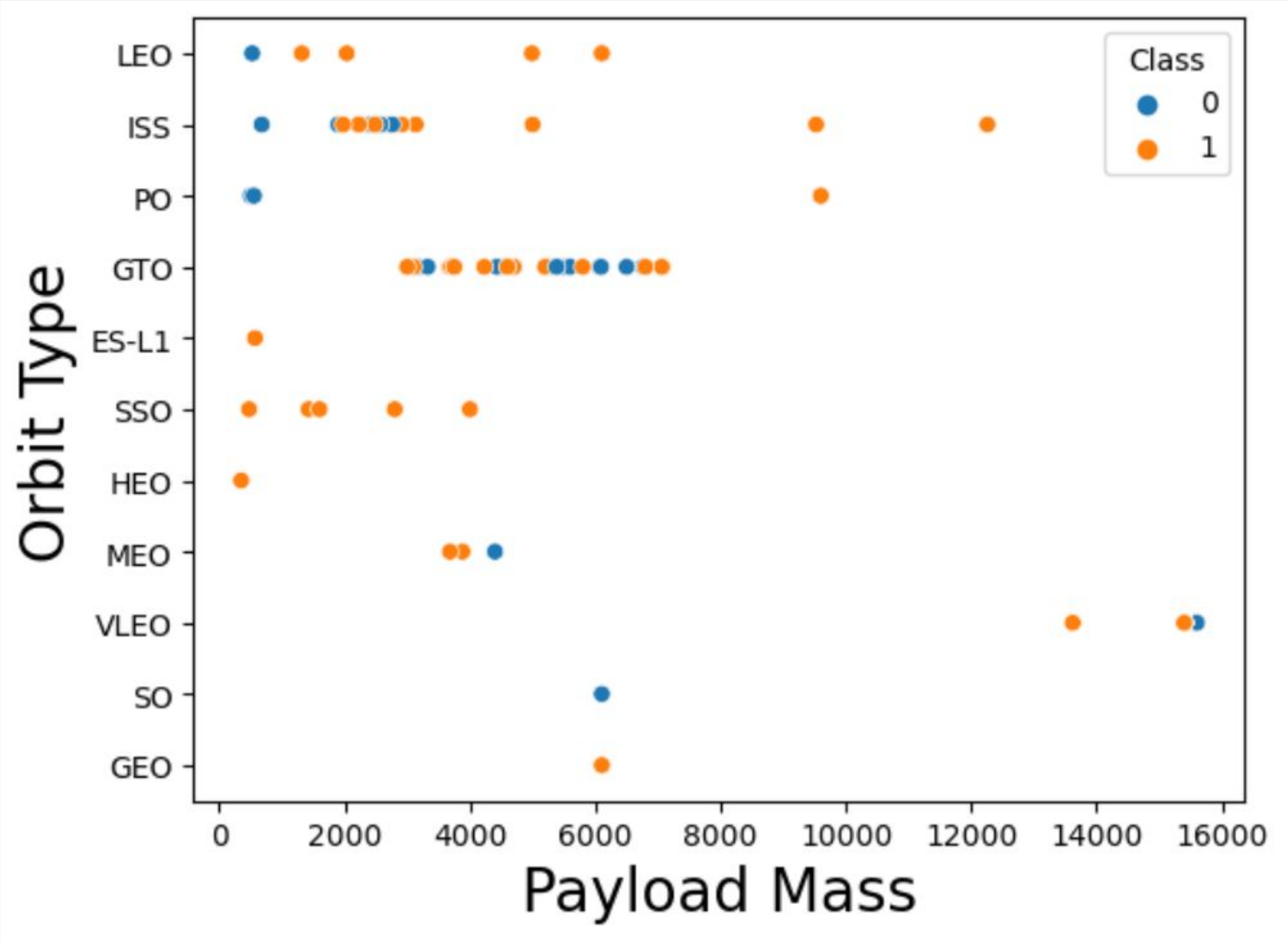


Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

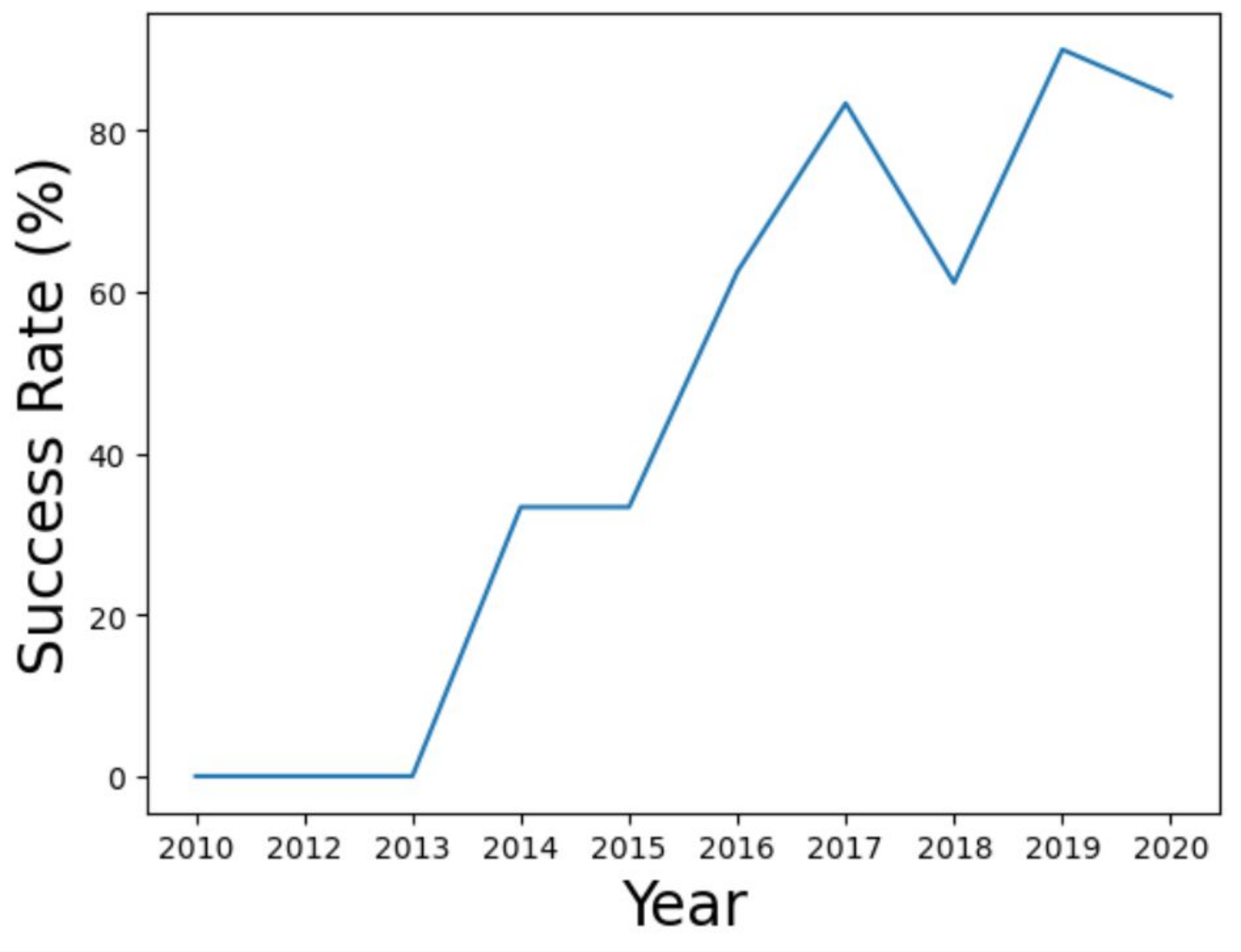
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

-



Launch Success Yearly Trend

The success rate since 2013 kept increasing untill 2020.



All Launch Site Names

The following SQL query give us the below launch site names:

SELECT DISTINCT Launch_Site FROM SPACEXTBL;

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

Launch Site Names Begin with 'CCA'

```
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
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass, Customer
FROM SPACEXTBL
WHERE Customer == "NASA (CRS)"
;
```

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

Total_Payload_Mass	Customer
45596	NASA (CRS)

All of the launches from the NASA (CRS) site had a combined mass of 45,596 kg.

Average Payload Mass by F9 v1.1

```
SELECT AVG(PAYLOAD_MASS__KG_), Booster_Version
FROM SPACEXTBL
WHERE Booster_Version == "F9 v1.1"
;
```

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

AVG(PAYLOAD_MASS__KG_)	Booster_Version
2928.4	F9 v1.1

The average payload mass for a F9 v1.1 booster version is 2,928 kg.

First Successful Ground Landing Date

```
SELECT MIN(Date), *
FROM SPACEXTBL
WHERE Landing_Outcome = "Success (ground pad)"
;
```

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

MIN(Date)	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-12-22	2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

The first successful landing on ground pad was on December 22nd, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
SELECT Booster_Version
FROM SPACEXTBL
WHERE Landing_Outcome = "Success (drone ship)"
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
;
```

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

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

Total Number of Successful and Failure Mission Outcomes

```sql
SELECT Mission_Outcome, COUNT(Mission_Outcome)
FROM SPACEXTBL
GROUP BY Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

```
SELECT DISTINCT Booster_Version
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ == (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

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

List of the names of the booster which have carried the maximum payload mass

2015 Launch Records

```
SELECT substr(Date, 6, 2) as Month, substr(Date,1,4) AS Year, Landing_Outcome, Booster_Version
FROM SPACEXTBL
WHERE substr(Date,1,4)='2015' AND Landing_Outcome = "Failure (drone ship)"
;
```

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

Month	Year	Landing_Outcome	Booster_Version
10	2015	Failure (drone ship)	F9 v1.1 B1012
04	2015	Failure (drone ship)	F9 v1.1 B1015

List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

```
SELECT Landing_Outcome, COUNT(Landing_Outcome)
FROM SPACEXTBL
GROUP BY Landing_Outcome
HAVING date BETWEEN "2010-06-04" AND "2017-03-20"
ORDER BY COUNT(Landing_Outcome) DESC
```

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

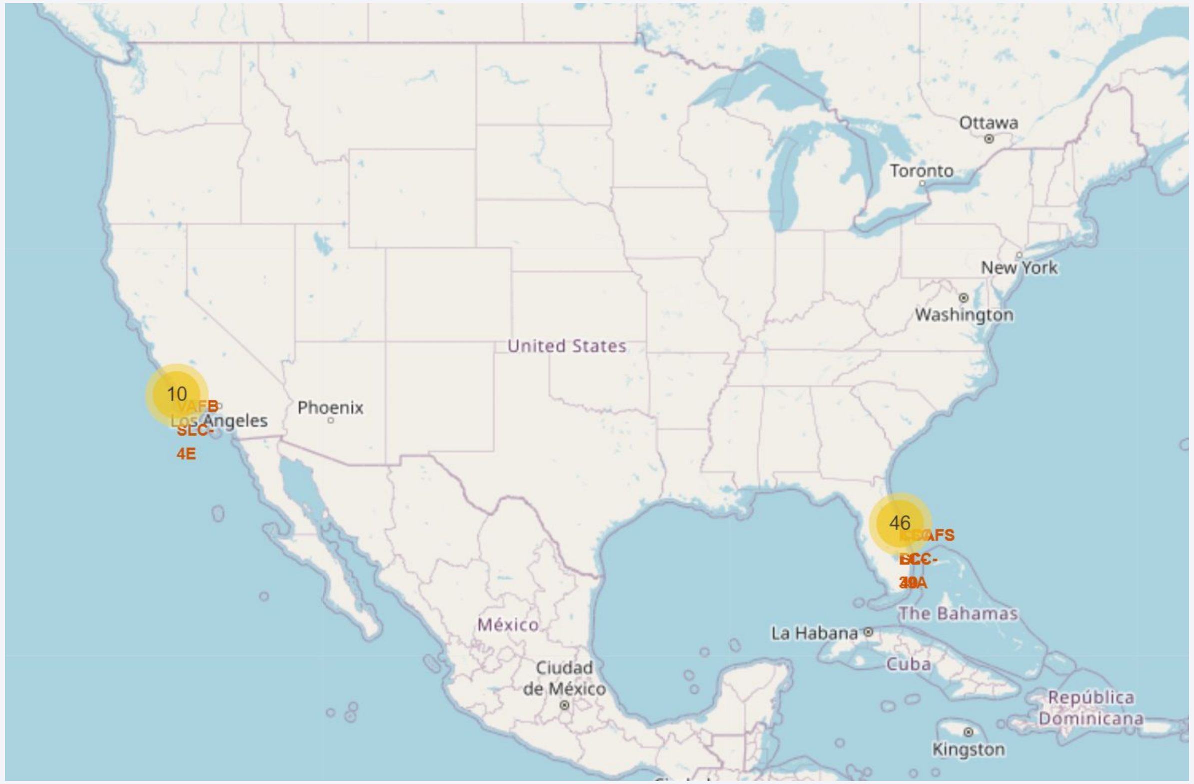
Landing_Outcome	COUNT(Landing_Outcome)
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Count of landing outcomes between the date 2010-06-04 and 2017-03-20.

Section 3

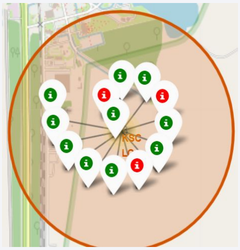
Launch Sites Proximities Analysis

Launch Sites Map

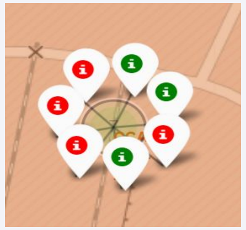


All launch sites are on the coastal regions on the USA, with the Vandenberg Space Force Base on the West coast that launched 10 rockets and the 3 other sites in Florida that launched 46 rockets.

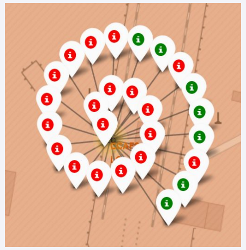
Success per Site



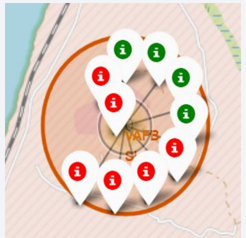
KSC
LC-39A



CCAFS
SLC-40



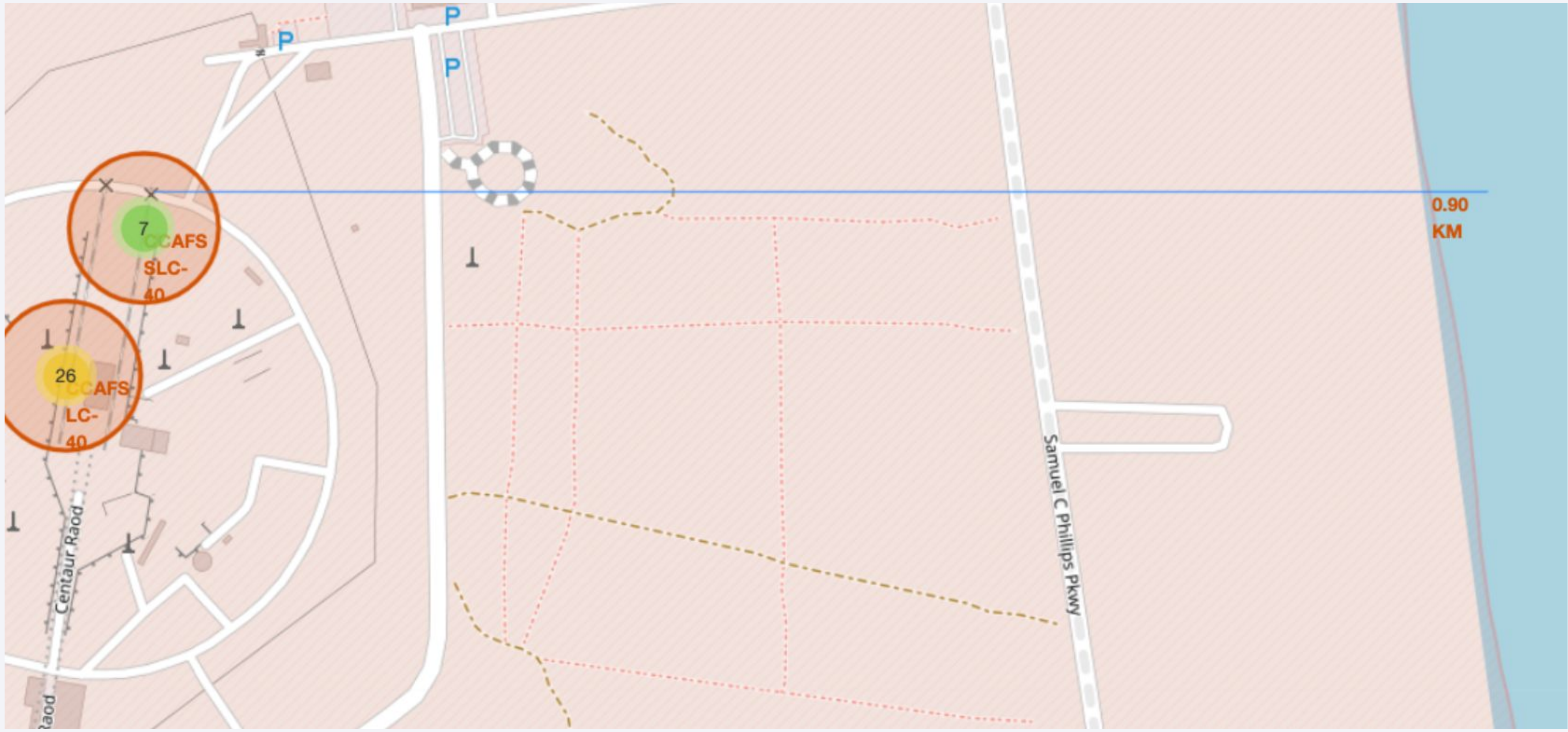
CCAFS LC-40



KSC
LC-39A

The Kennedy Space Center Launch Complex shows a better success rate than the other sites.

Site proximity

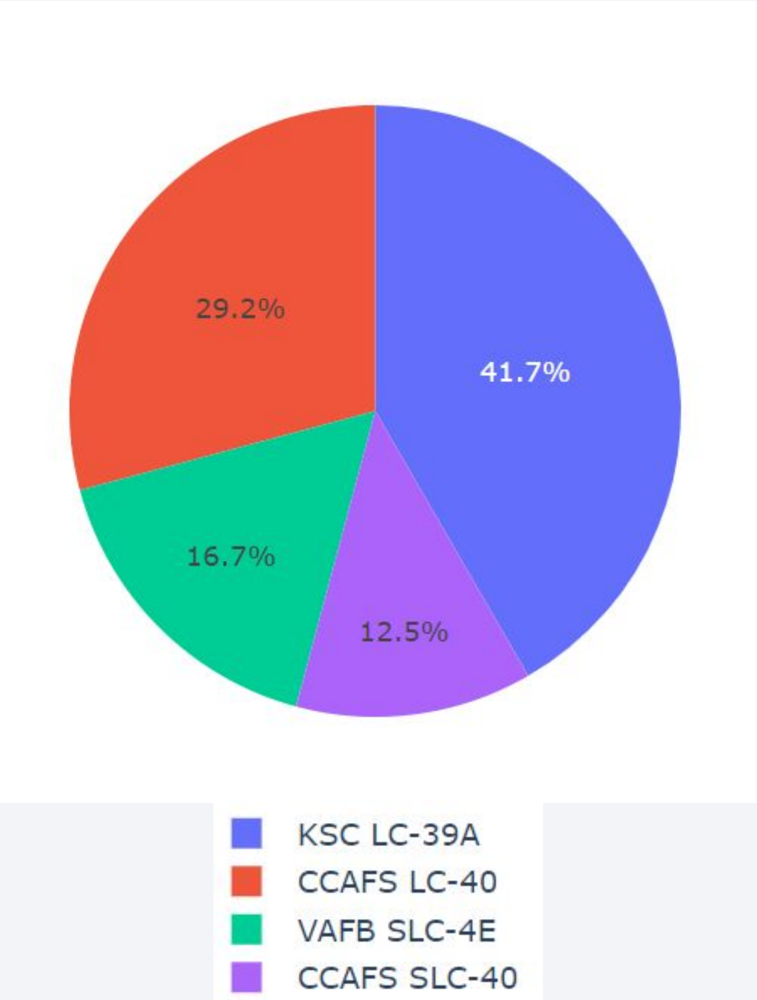


The Cape Canaveral Space Launch Complex is within 1 kilometer from the ocean.

Section 4

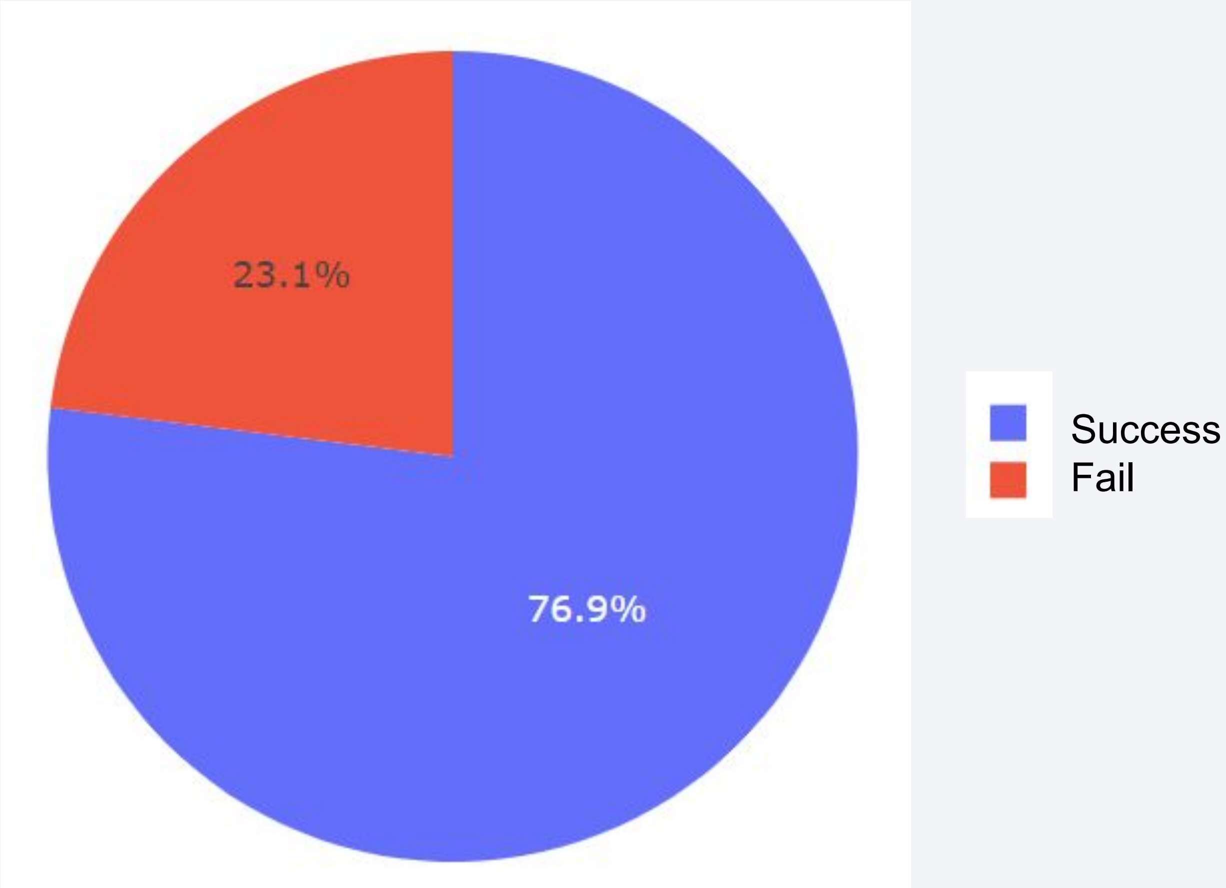
Build a Dashboard with Plotly Dash

Launch Success per Site



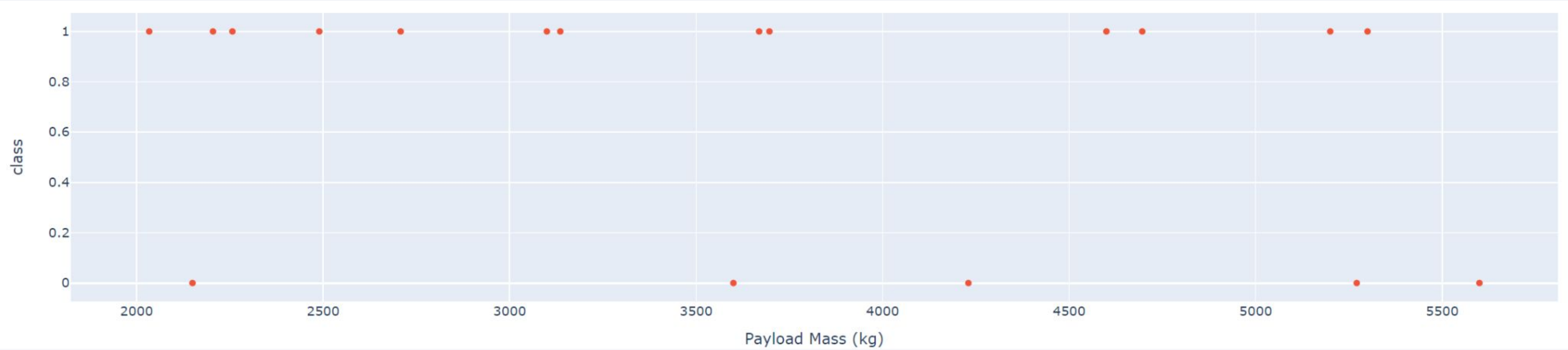
KSC LC-39A has the most successful launches amongst launch sites (41.2%)

Most Successful Launch Site



The Kennedy Space Center Launch Complex 39A in Florida has the highest success rate.

Payload Mass and Version Success



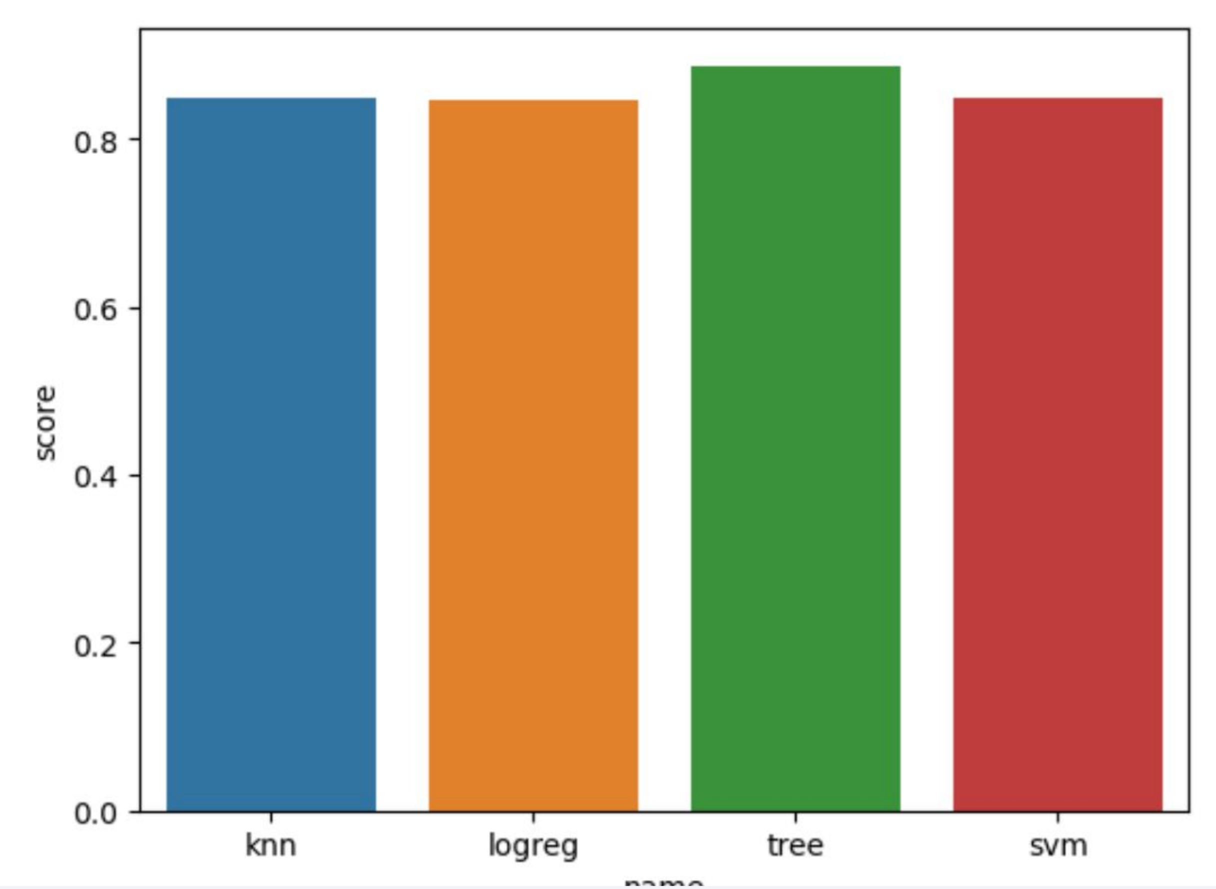
The FT rocket version has the best success/failure proportion

Above 6 tons, the chance of success is small.

Section 5

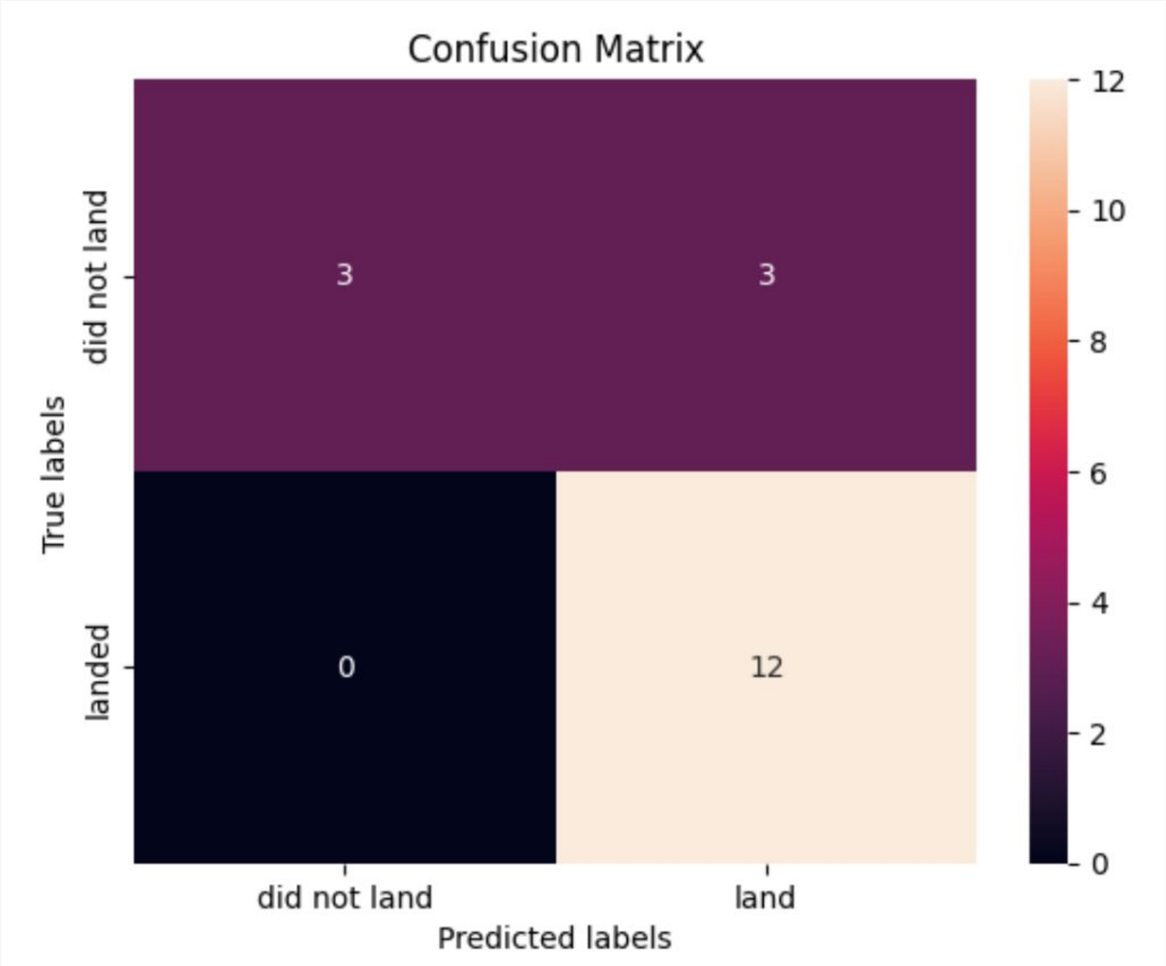
Predictive Analysis (Classification)

Classification Accuracy



Our decision tree model has a slightly higher accuracy.

Confusion Matrix



- All the confusion matrices were identical.
- The fact that there are false positives (Type 1 error) is not good.
- Confusion Matrix Outputs:
  - 12 True positive,
  - 3 True negative,
  - 3 False positive,
  - 0 False Negative.

Conclusions

- **Model Performance**: The models performed similarly on the test set with the decision tree model slightly outperforming.
- **Equator**: Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters.
- **Coast**: All the launch sites are close to the coast.
- **Launch Success**: Increases over time.
- **KSC LC-39A**: Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg.
- **Orbits**: ES-L1, GEO, HEO, and SSO have a 100% success rate.
- **Payload Mass**: Across all launch sites, the heavier the payload mass (kg), the higher the success rate.