IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alessandro
D'Angelo

Outline

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

Executive Summary

Summary of Methodologies

We analysed data from SpaceX missiles launches. The methodologies utilised are the following:

- **Collection:** we collected data using web scraping techniques.
- **Data Wrangling:** we wrangled the data to get success/fail outcome variables.
- **Data Exploration:** we explored the data using data visualisation techniques, such as scatter plots and bar charts. We also analysed the data using SQL to better understand total payload, payload range for success rate and total number of success rate.
- **Geographical Data Exploration:** we used Folium to investigate the relation between success rate and locations of launch sites. We also observed the proximity of launch sites to relevant geographical markers.
- **Modelling:** we studied which predictive model was best suited for outcomes prediction using different machine learning methods.

Results

Exploratory Data Analysis

- Launch success rates have increased over time.
- KSC LC-39A is the landing site with highest success rate.
- Orbits with 100% success rate are: ES-L1, GEO, HEO, SSO.

Geo-spatial Analytics

- All launch sites are near the equator and coast lines.

Predictive Analytics

- The decision tree method performed slightly better than the others ML algorithms.

Introduction

Background

SpaceX, a leader in space exploration and commercial spaceflight, has revolutionised the aerospace industry with its reusable rocket technology and ambitious mission goals. This report delves into an analysis of SpaceX's missile launches, focusing on key factors such as payloads, launch site locations, and their correlation with the success or failure of missions. By examining historical launch data, we explored the impact of these variables on mission outcomes. Furthermore, we employed various machine learning algorithms to predict the likelihood of success in future launches. The insights gained from this analysis not only enhance our understanding of the critical factors influencing launch success but also pave the way for optimizing future missions, thereby contributing to SpaceX's goal of achieving greater efficiency and reliability in space exploration.

Problems and Goals

- We need to understand how payload mass, launch site, number of flights and orbits affects first-stage landing.
- We need to rate landing success over time.
- We need to find the best predictive model for successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection:
  - SpaceX API and web scraping
- Data wrangling
  - Filtering, missing values and one-hot encoding
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - Logistic Regression
  - SVM
  - Decision Tree
  - KNN

Data Collection

Steps:

- Request data from SpaceX API
- Convert the response into a dataframe using json_normalize()
- Create a dictionary from data
- Create dataframe from dictionary
- Filter the dataframe retaining only the Falcon9 launches
- Replace missing values of Payload Mass using the mean of the known data

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

https://github.com/ADA-Projects/IBM-SpaceX-DS-Project/blob/3221fa354de7d07b445a18d8e373a4aa6f03d909/001_SpaceX_Data_Collection_API.ipynb

Data Collection - Scraping

Steps:

- Request data of Falcon9 launches from Wikipedia
- Create a BeautifulSoup object from HTML response
- Extrapolate column names from HTML table headers
- Collect data parsing HTML tables
- Create dictionaries and dataframes from data

```
# use requests.get() method with the provided static_url
# assign the response to a object
html_data = requests.get(static_url)
html_data.status_code
```

```
200
```

Create a `BeautifulSoup` object from the HTML `response`

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(html_data.text)
```

Print the page title to verify if the `BeautifulSoup` object was created properly

```
# Use soup.title attribute
soup.title
```

https://github.com/ADA-Projects/IBM-SpaceX-DS-Project/blob/3221fa354de7d07b445a18d8e373a4aa6f03d909/002_SpaceX_Web_Scraping.ipynb

Data Wrangling

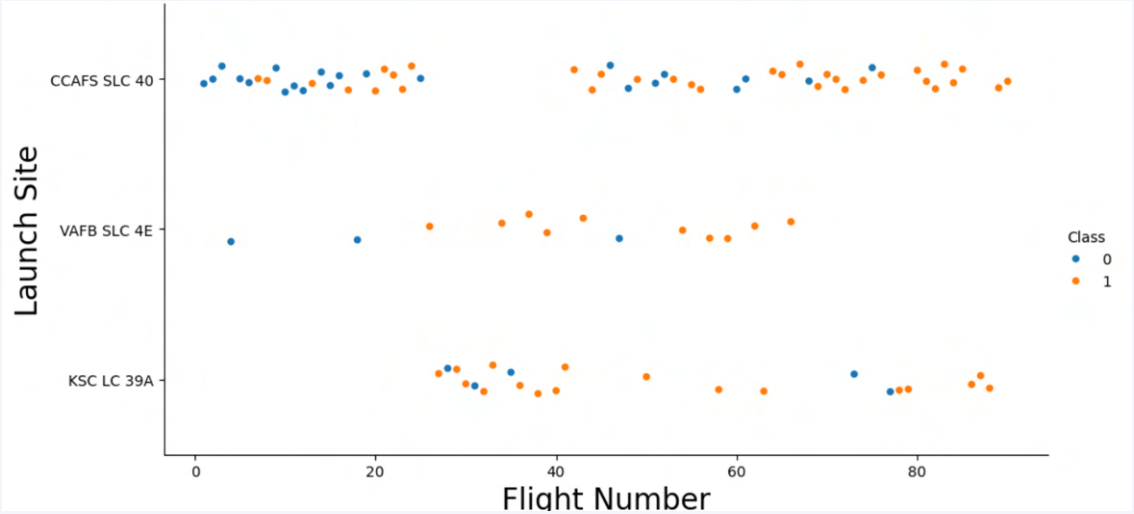
- Exploratory Data Analysis (EDA) to determine data labels
- Calculate:
  - Number of launches on each site
  - Number and occurrence of Orbits
  - Number and occurrence of mission outcomes per orbit
- Create binary categorical landing outcome attribute

```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise
landing_class=[]
for outcome in df['Outcome']:
    if outcome in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)
```

https://github.com/ADA-Projects/IBM-SpaceX-DS-Project/blob/3221fa354de7d07b445a18d8e373a4aa6f03d909/003_SpaceX_Data_Wrangling.ipynb

EDA with Data Visualization

- To explore the data we used scatter plots and bar charts, in order to look for possible correlations. We analysed the following:
  - Payload Mass and Flight Number,
  - Launch Site and Flight Number
  - Launch Site and Payload Mass,
  - Orbit and Flight Number,
  - Payload and Orbit



https://github.com/ADA-Projects/IBM-SpaceX-DS-Project/blob/3221fa354de7d07b445a18d8e373a4aa6f03d909/005_SpaceX_EDA_Data_Visualization.ipynb

EDA with SQL

- We performed the following SQL queries:
  - 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;
  - Rank of the count of landing outcomes (such as Failure, drone ship, or Success, groundpad) between the date 2010-06-04 and 2017-03-20.

https://github.com/ADA-Projects/IBM-SpaceX-DS-Project/blob/3221fa354de7d07b445a18d8e373a4aa6f03d909/004_SpaceX_EDA_SQL.ipynb

Build an Interactive Map with Folium

Markers Indicating Launch Sites:

- Added blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates.
- Added red circles at all launch sites coordinates with a popup label showing its name using its name using its latitude and longitude coordinates.

Colored Markers of Launch Outcomes:

- Added colored markers of *successful* (green) and *unsuccessful* (red) *launches* at each launch site to show which launch sites have high success rates.

Distances from Launch Sites and Points of Interest:

- Added colored lines to show distance between launch site CCAFS SLC-40 and its proximity to the nearest coastline, railway, highway, and city.

https://github.com/ADA-Projects/IBM-SpaceX-DS-Project/blob/3221fa354de7d07b445a18d8e373a4aa6f03d909/005_SpaceX_EDA_Data_Visualization.ipynb

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites:

- Allow user to select all launch sites or a specific launch site.

Pie Chart Showing Successful Launches:

- Allow user to see successful and unsuccessful launches as a percentage of the total.

Slider of Payload Mass Range:

- Allow user to select payload mass range.

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version:

- Allow user to see the correlation between Payload and Launch Success.

https://github.com/ADA-Projects/IBM-SpaceX-DS-Project/blob/3221fa354de7d07b445a18d8e373a4aa6f03d909/005_SpaceX_EDA_Data_Visualization.ipynb

Predictive Analysis (Classification)

Steps:

- Create NumPy array from the Class column
- Standardize the data with StandardScaler. Fit and transform the data.
- Split the data using train_test_split
- Create a GridSearchCV object with cv=10 for parameter optimization
- Apply GridSearchCV on different algorithms: logistic regression, support vector machine, decision tree, K-Nearest Neighbor
- Calculate accuracy on the test data using .best_score() for all models
- Assess the confusion matrix for all models
- Identify the best model using R2 scores

https://github.com/ADA-Projects/IBM-SpaceX-DS-Project/blob/3221fa354de7d07b445a18d8e373a4aa6f03d909/008_SpaceX_Predictive_Analytics.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.

Visual Analytics:

- Most launch sites are near the equator, and all are close to the coast.
- Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities.

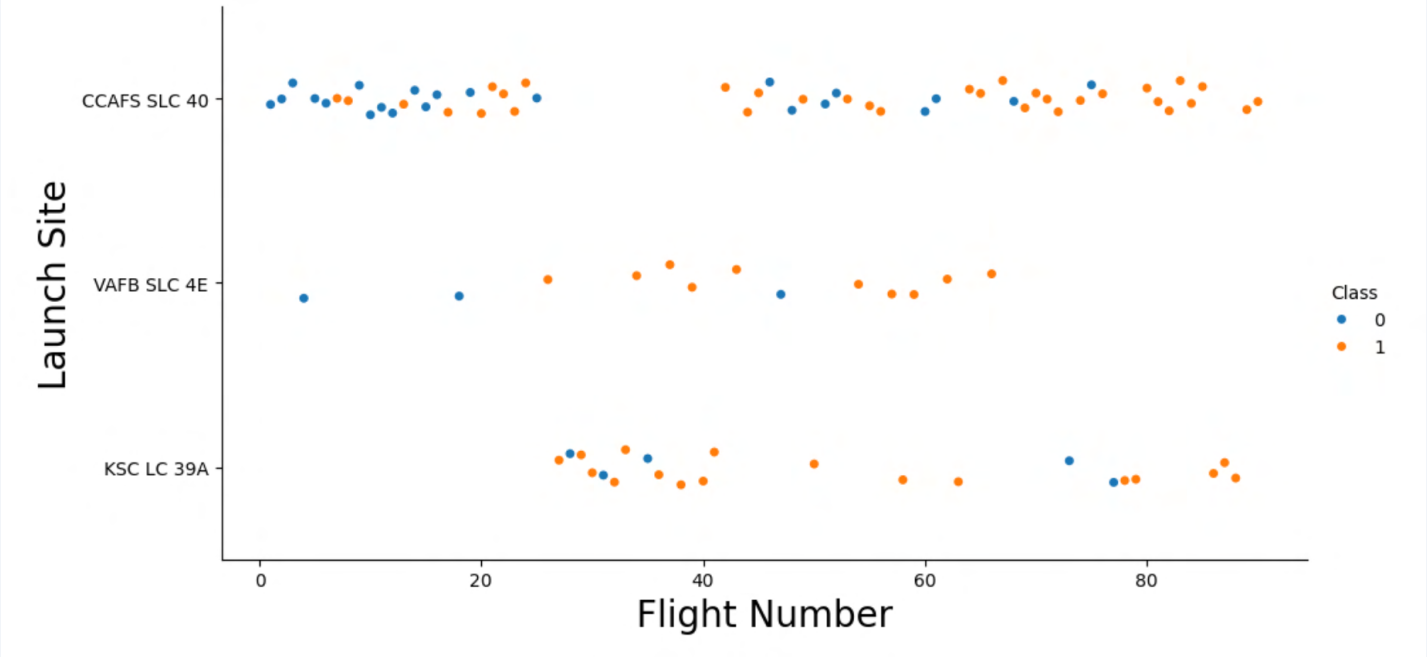
Predictive Analytics:

- Decision Tree model is the best predictive model for the dataset.

Section 2

Insights drawn from EDA

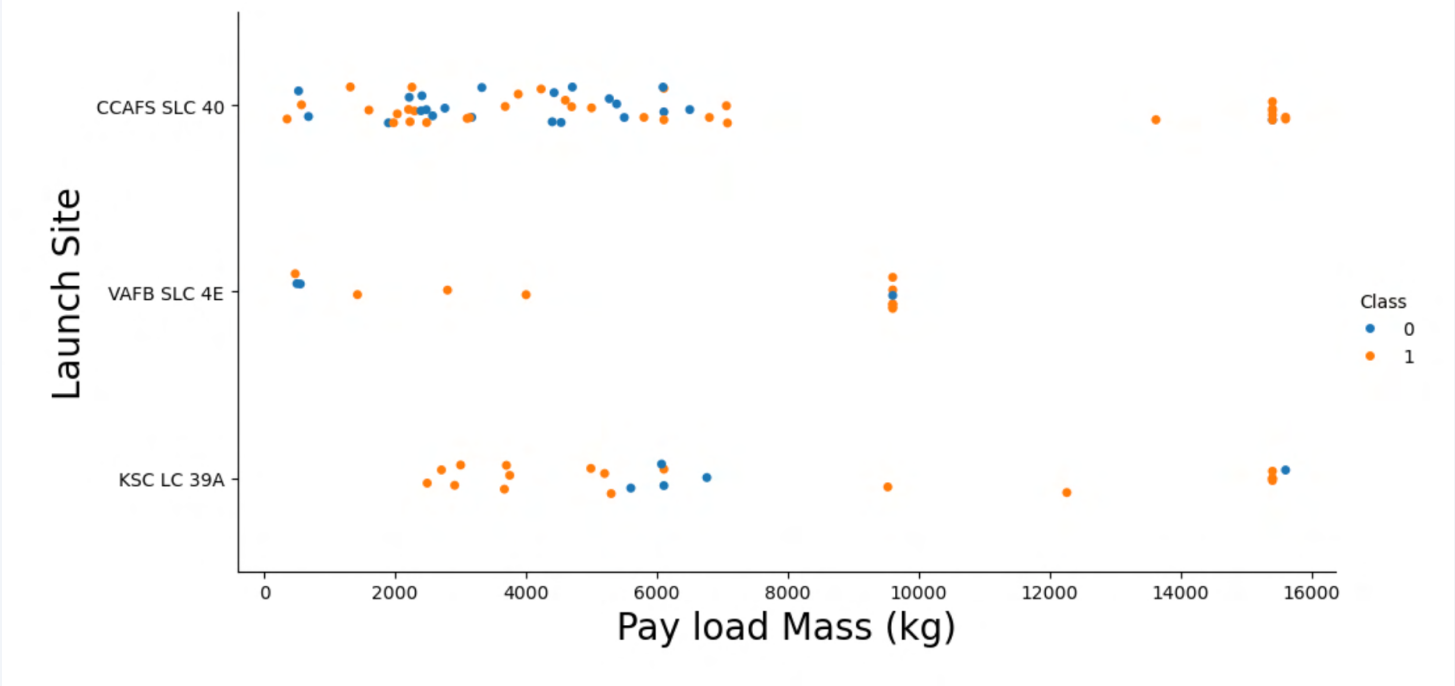
Flight Number vs. Launch Site



Steps:

- Earlier flights had a lower success rate (blue = fail).
- Later flights had a higher success rate (orange = success).
- Around half of launches were from CCAFS SLC 40 launch site.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- We can deduce that new launches have a higher success rate.

Payload vs. Launch Site



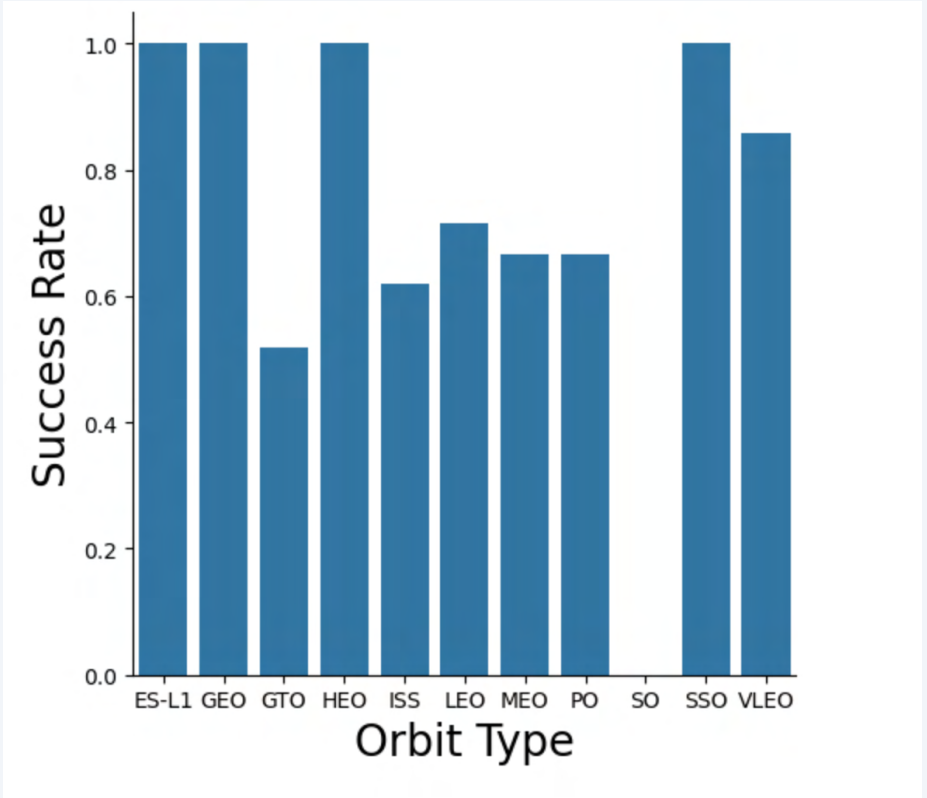
Steps:

- For launch sites CCAFS SLC 40 and VAFB SLC 4E, the higher the payload mass (kg), the higher the success rate.
- Most launches with a payload greater than 7,000 kg were successful.
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg.
- VAFB SKC 4E has not launched anything greater than 10,000 kg.

Success Rate vs. Orbit Type

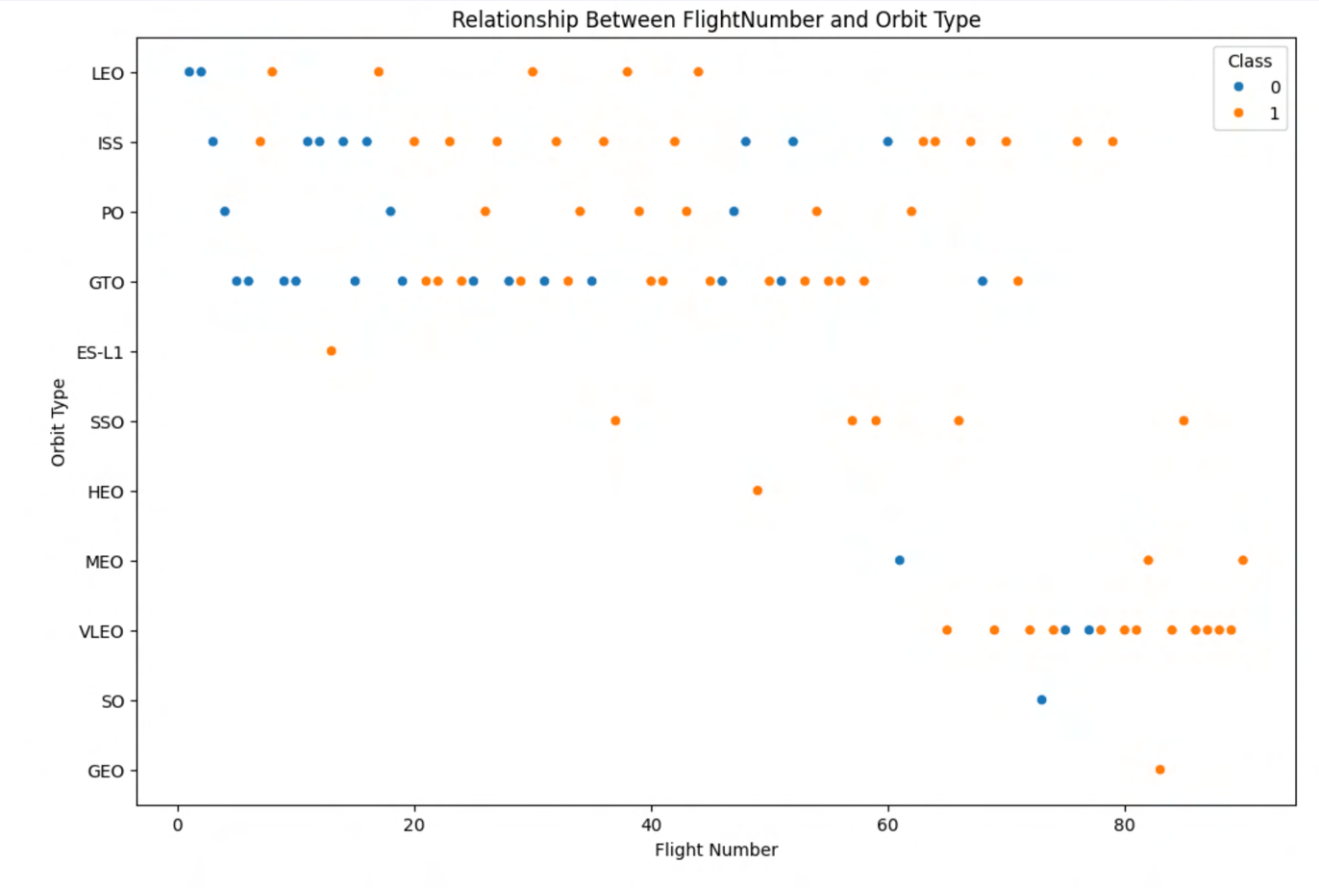
Success Rate:

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



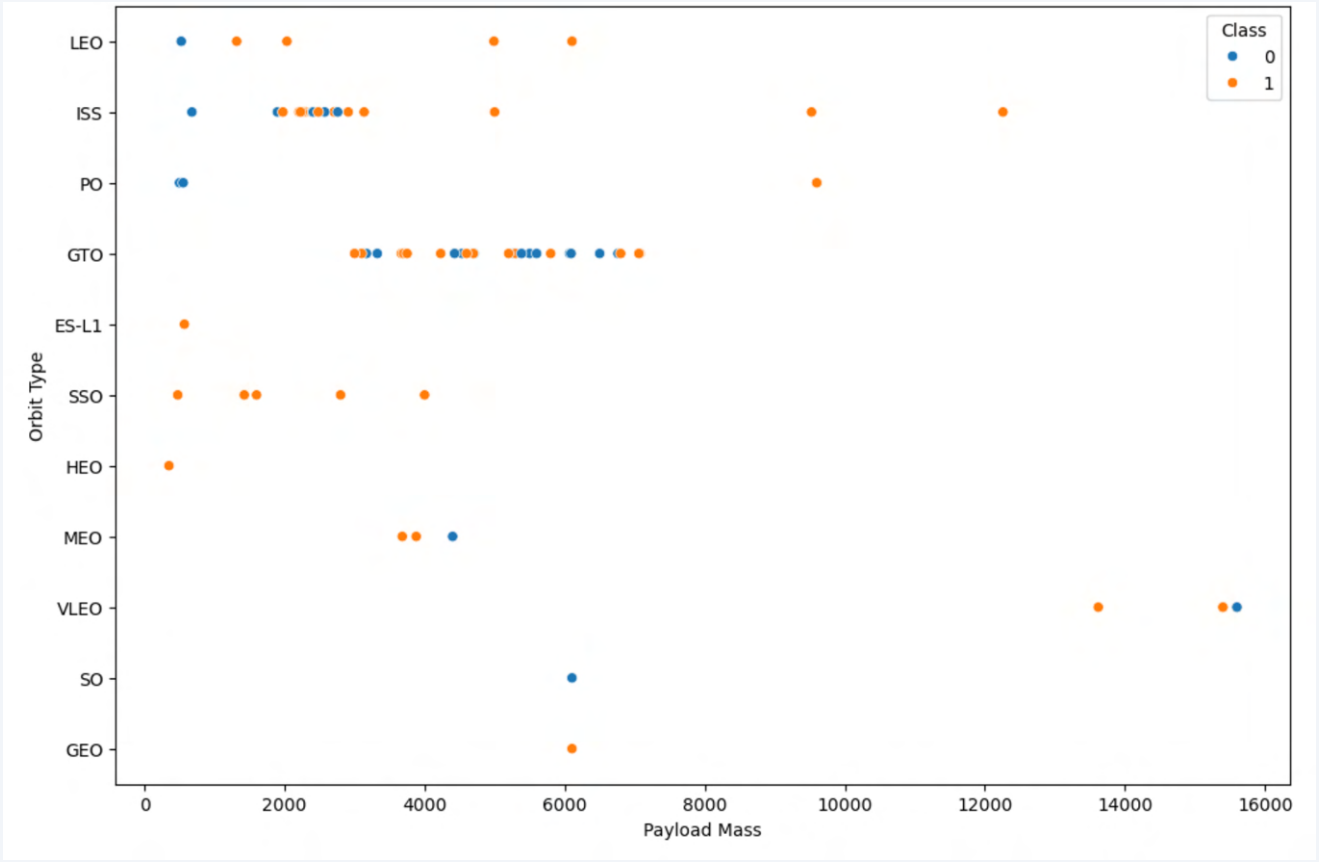
Flight Number vs. Orbit Type

- Success rate improved with the increase of number of flights for each orbit.
- VLEO orbit seems a new good business opportunity, due to recent increase of its success rate.



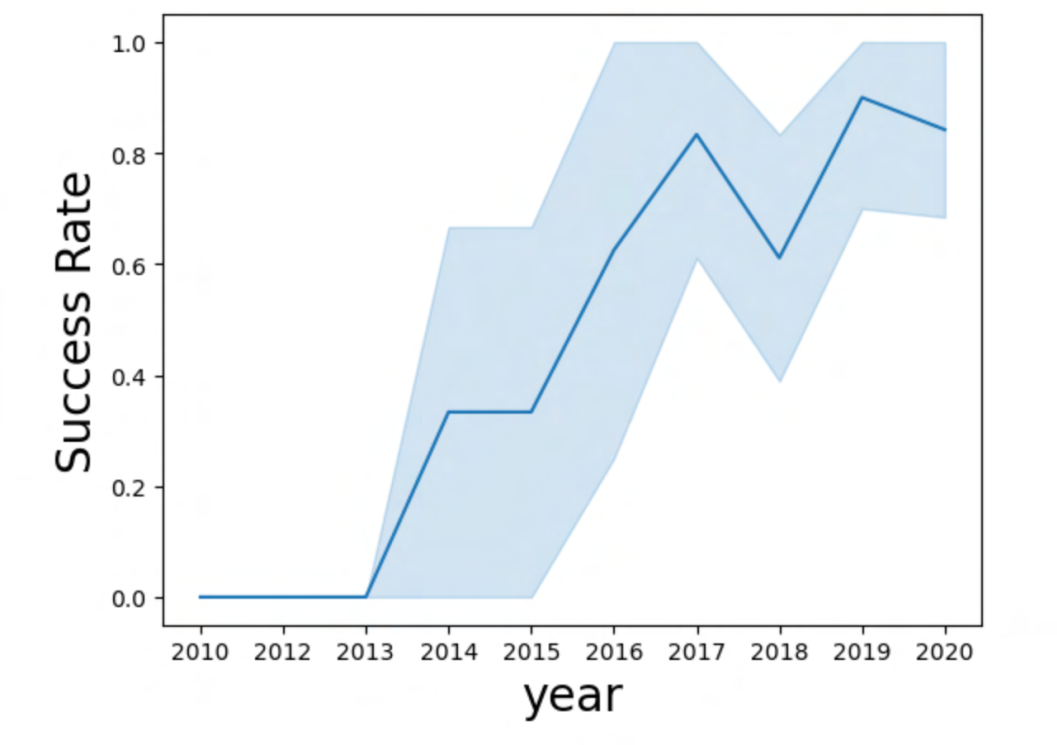
Payload vs. Orbit Type

- LEO, ISS, PO have better success rate with higher payloads.
- Success rate on the GTO orbit does not seem to depend of the payload.



Launch Success Yearly Trend

- The overall success rate has improved significantly from 2013 to 2020.
- Success rate had a steep increase in the periods 2013-2017 and 2018-2019.
- Success rate decreased only during 2017-2018 and 2019-2020.



All Launch Site Names

Launch Sites:

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

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;
```

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

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

Launch Site Names Begin with 'CCA'

First 5 entries of Launch Sites with ‘CCA’:

- We selected from out table only launch sites that contained the sought string and we limited the SQL query search to the five top most entries.

```
%%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
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass and Average Payload Mass

Total Payload NASA:

- The total payload carried by boosters from NASA is **45596 kg**.

```
%%sql SELECT SUM(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL
    WHERE CUSTOMER = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload F9 v1.1:

- The average payload carried by boosters version F9 v1.1 is **2928,4 kg**.

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The first successful landing date (on ground pad) was the **2015-12-22**.

```
%%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

Payload Names with boosters mass in between 4000 kg and 6000 kg are:

- JCSAT-14
- JCSAT-16
- SES-10
- SES-11/EchoStar 105

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

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

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

Success and Failure Mission Outcomes:

- There was 1 failure (in flight).
- 99 Successes.
- 1 Success outcome has an unclear payload status.

```
%%sql SELECT Mission_Outcome, COUNT(*) as total_number
FROM SPACEXTBL
GROUP BY Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

Boosters carrying maximum payload are:

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

The query used is:

```
%%sql SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

2015 Launch Records

The failed landing records in 2015 with Date, Booster Version, Launch Site are:

```
%%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

Ranking of all landing outcomes between the date 2010-06-04 and 2017-03-20 (in descending order):

```
%%sql SELECT [Landing_Outcome], count(*) as count_outcomes
FROM SPACEXTBL
WHERE DATE between '2010-06-04' and '2017-03-20' group by [Landing_Outcome] order by count_outcomes DESC;
```

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

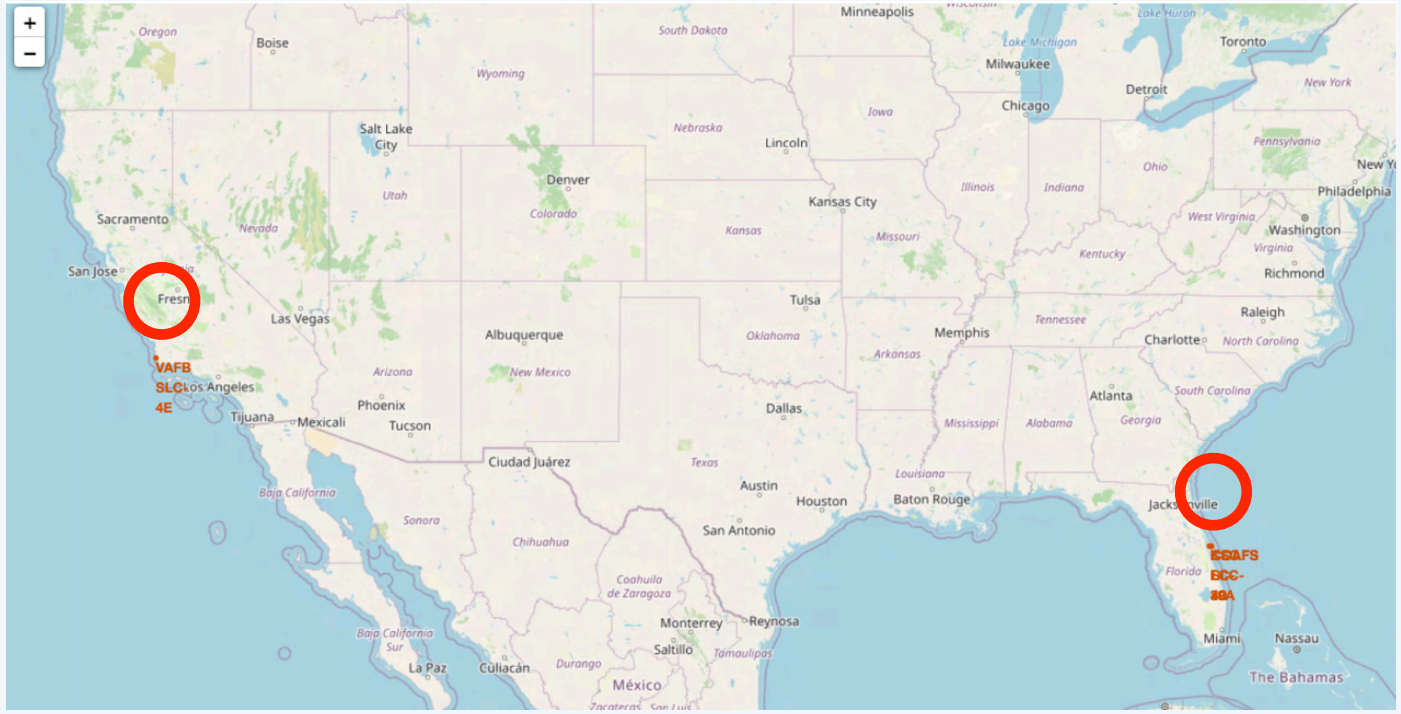
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

N.B.: The tables shows that “No attempts” must be taken into account.


Section 3
Launch Sites
Proximities Analysis

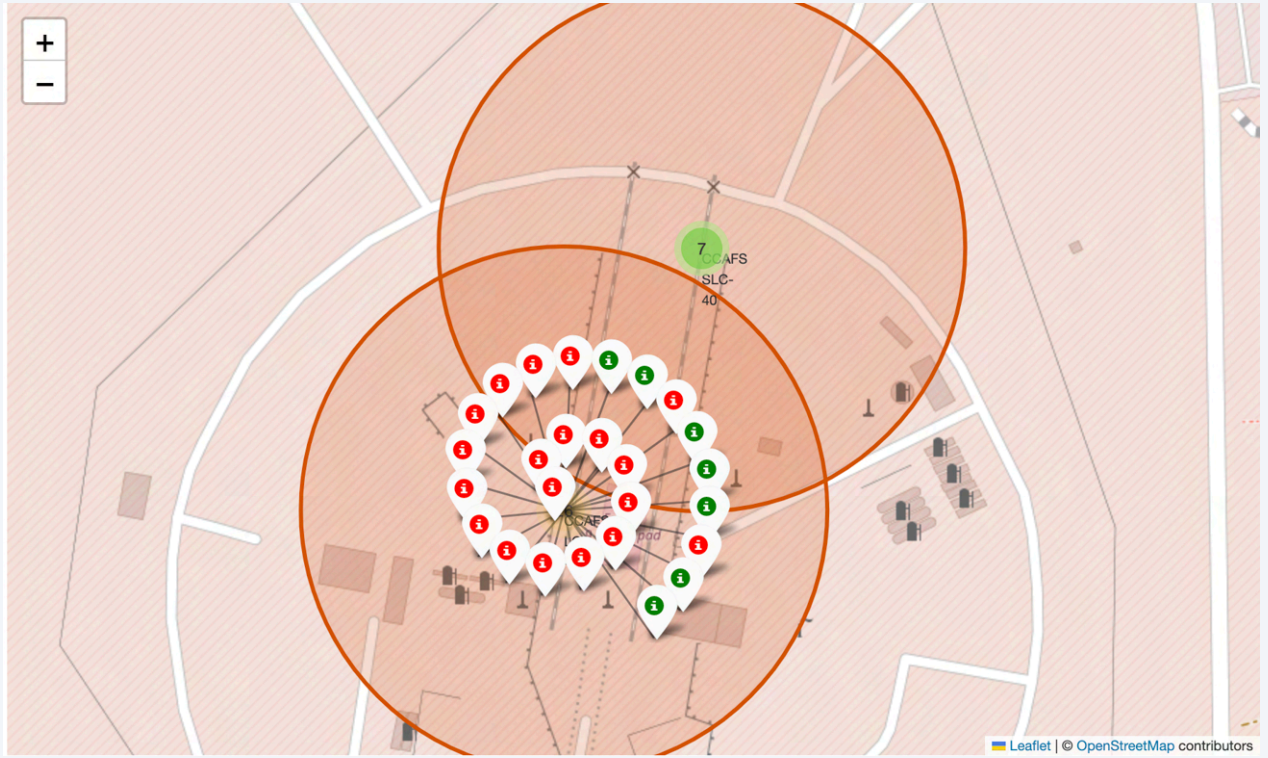
Launch Sites

The closer the launch site to the **equator**, the easier it is to launch to equatorial orbit, and the more help one gets from Earth's rotation. Indeed, rockets launched from sites near the equator get an additional natural boost- due to the rotational speed of earth - that helps save the cost of putting in extra fuel and boosters.



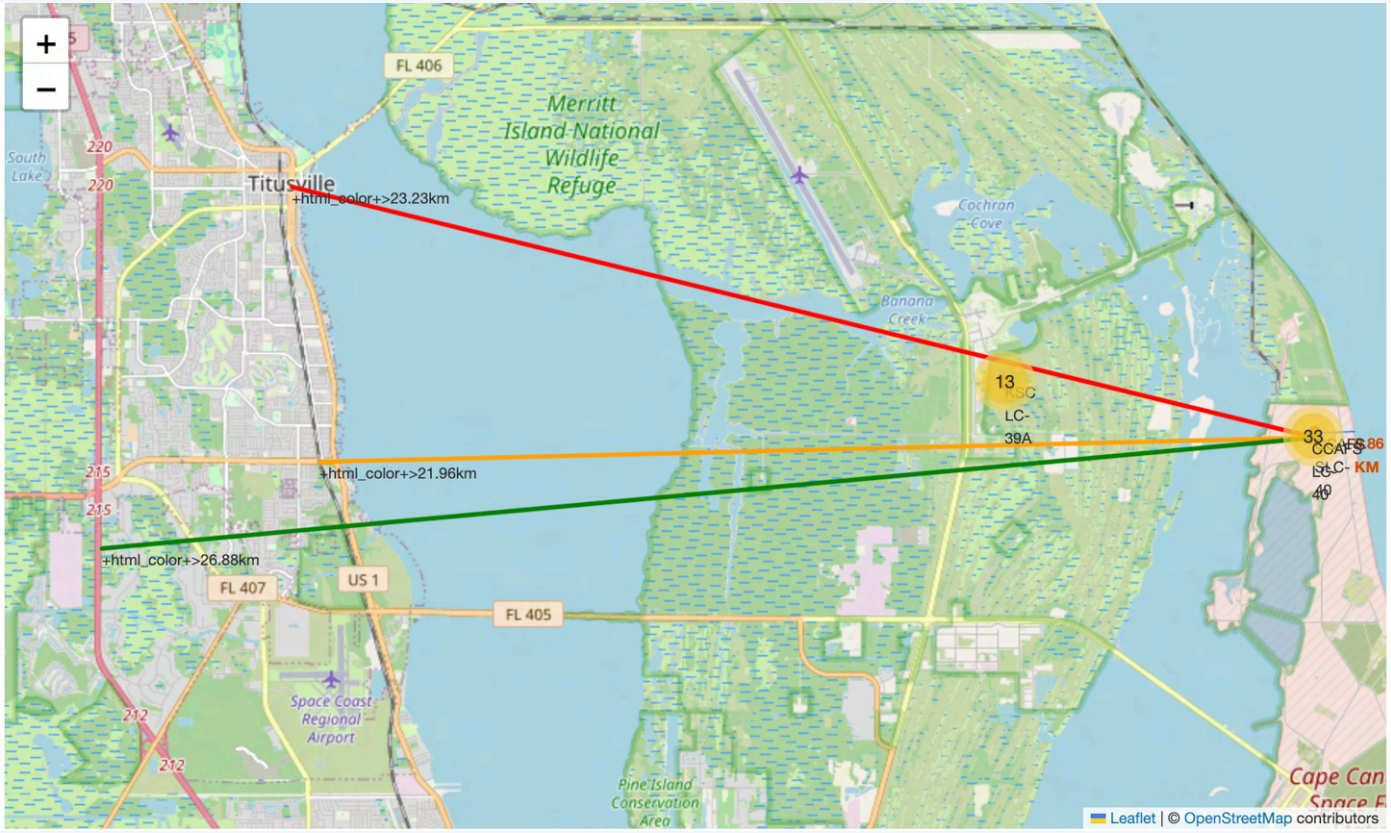
Launch Outcomes

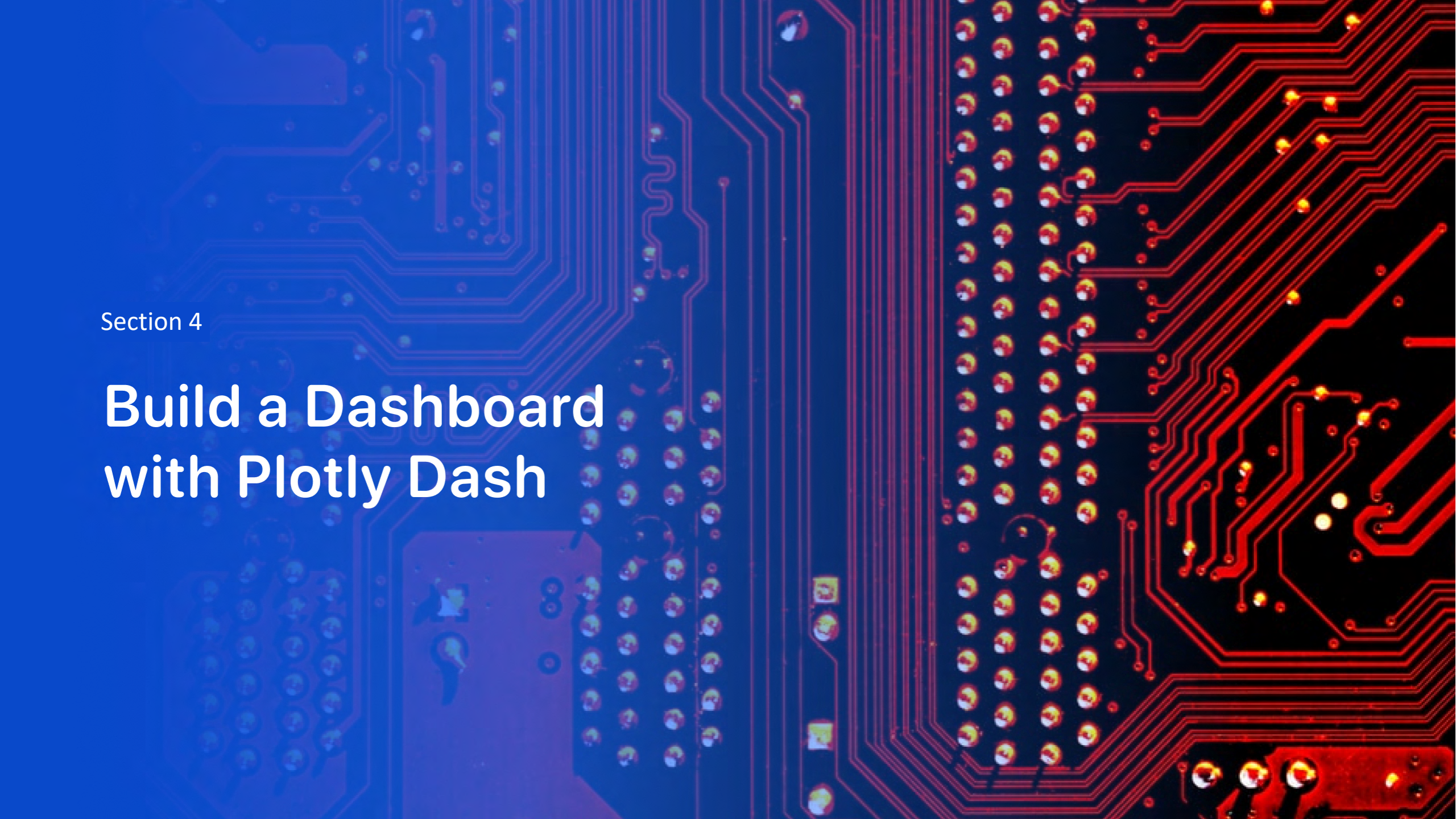
The picture on the right shows an examples of coloured markers used to denote successful or failure launches at each location. Green makers represent successful launches, while red markers represent failed ones



Logistic and Safety

Launch sites are usually close to coast lines due to safety reasons, and not too close to highly densely populated spots. But at the same time, they are placed not too far from connections to cities and main highways to allow easy connections for workers and for receiving and sending equipment.

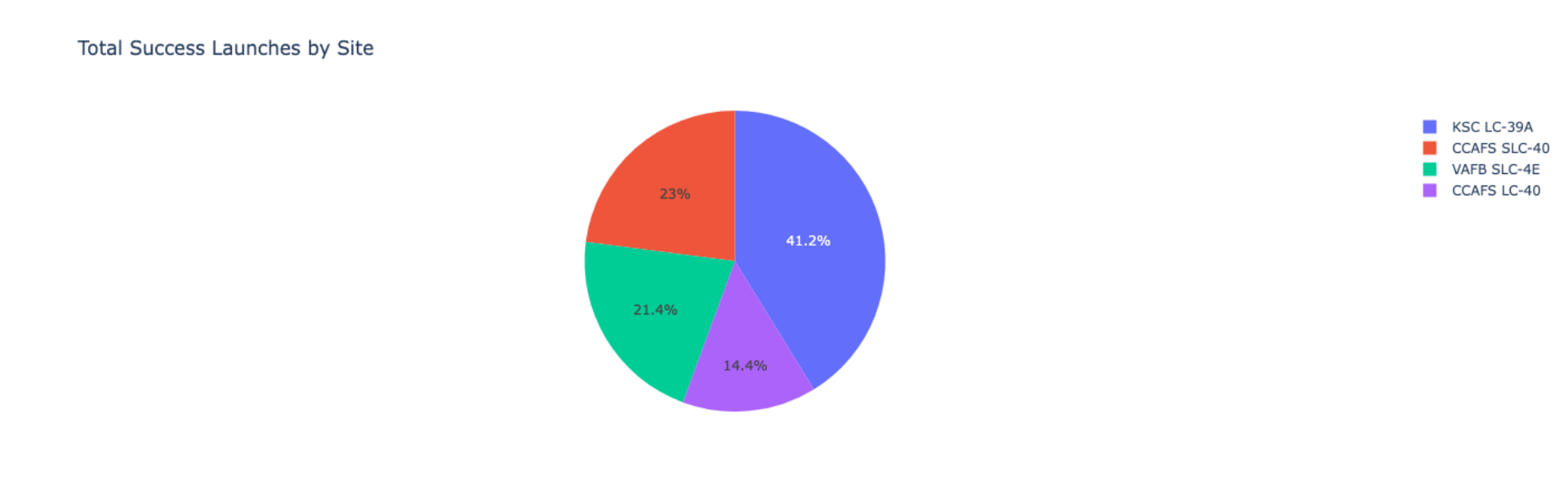




Section 4

Build a Dashboard with Plotly Dash

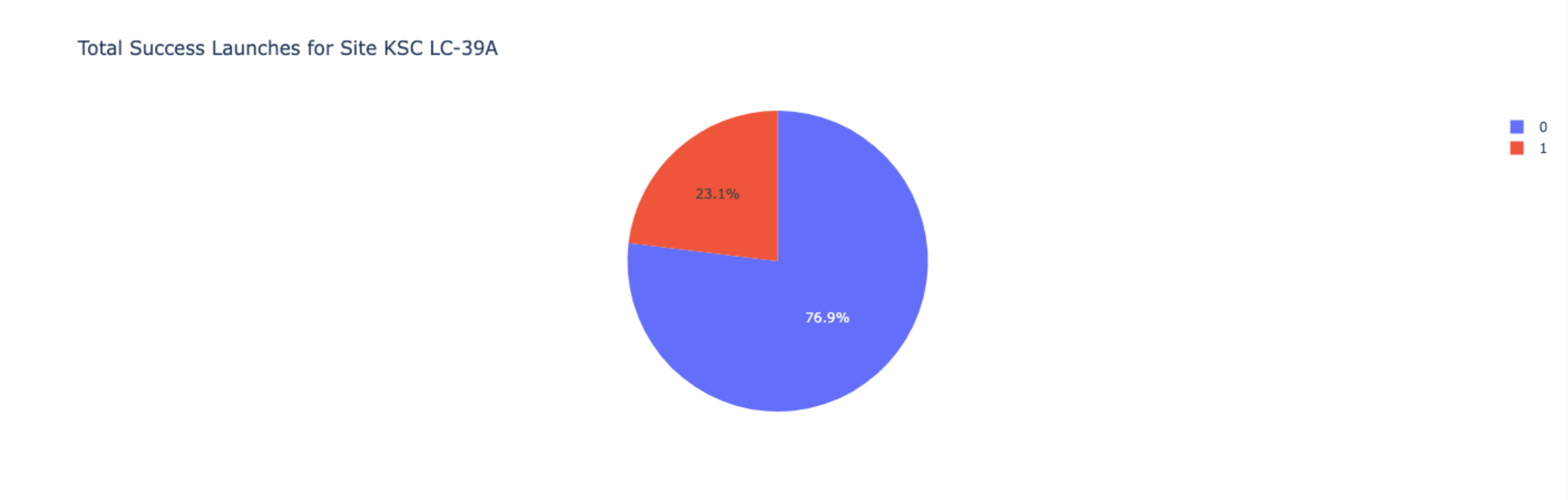
Successful Launches by Site



Launch sites play a major role when it comes to success rates.

KSC LC-39A Launch Success

The most successful launch site is **KSC LC-39A**, where we had 10 successful launches and 3 failed ones.



Success/Failure VS Payload

From the plot below we can see that on the whole range of Payload mass, we have significantly better outcomes when restricting to the range in between **2000 kg** and **6000 kg**.

Payload between 2000kg and 6000kg

Using the slider bar, we can clean up some noise and concentrate in between 2000 kg and 6000 kg. What we can see is that the **Booster version FT** has better landing outcomes, and even *better success rates* if we further restrict to a payload in the range **2000 kg - 3500 kg**.

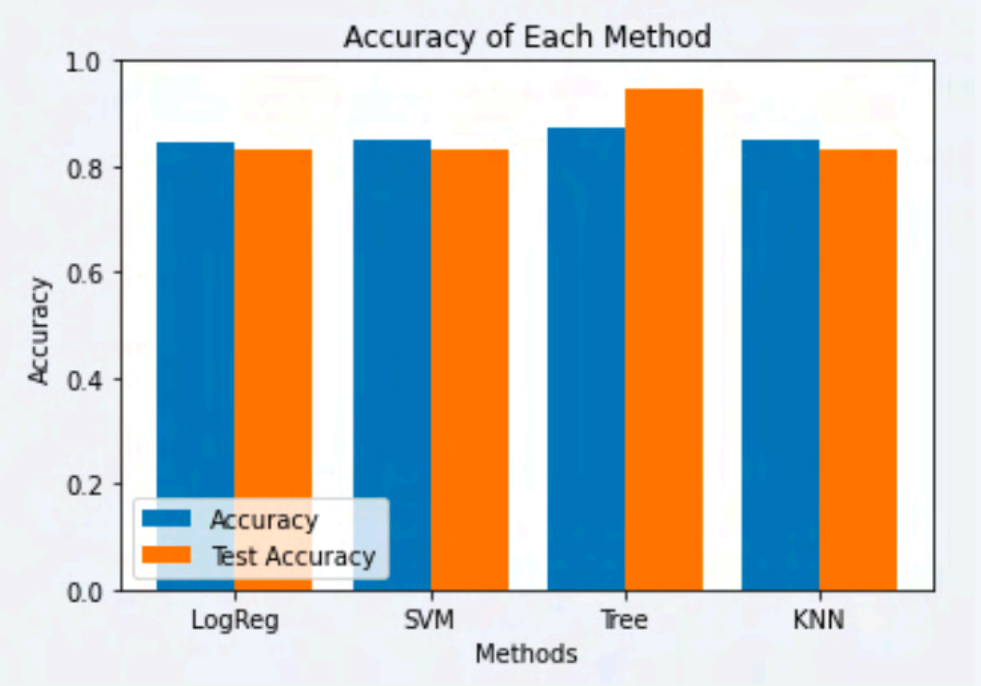
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were used: Logistic Regression (LogReg), Support Vector Machine (SVM), Decision Tree (Tree) and K-Nearest Neighbourhood (KNN).

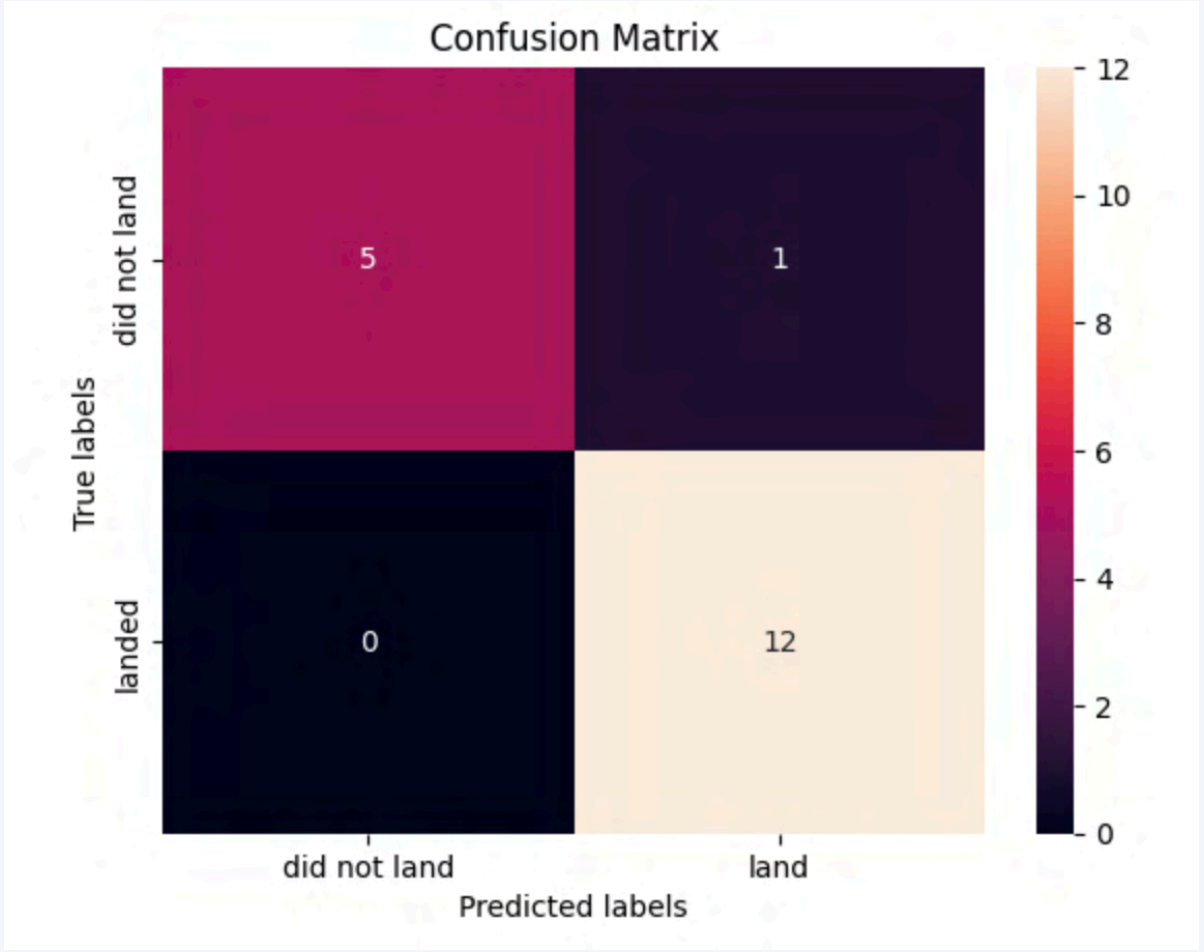
- While all the methods performed similarly, the decision tree model stood out with a better accuracy.



Confusion Matrix

Decision Tree Confusion Matrix:

- We had 12 true positives.
- We had 0 false negatives.
- We had 1 false positive.
- We had 5 true negatives



Conclusions

- **Model Performance**: The *decision tree* model was the most accurate.
- **Launch Sites Location**: 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.
- **Launch Success**: Increased 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 range in between 2000kg and 5000kg was the most successful one.

Thank you!