IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tommy Bell
20/08/2025

Outline

- EXECUTIVE SUMMARY
- INTRODUCTION
- METHODOLOGY
- RESULTS
- CONCLUSION
- APPENDIX

Executive Summary

SpaceX is a leading innovator in commercial spaceflight, with an extensive record of orbital launches.Their launch history provides a rich dataset for understanding the factors driving mission success.

This project collected and cleaned data from SpaceX’s API and Wikipedia to build a reliable analysis base.

Interactive dashboards (Folium & Plotly Dash) and EDA were used to uncover patterns in launch outcomes. With predictive model accurately determining the success of launches.

Insights derived from this research can inform strategic decisions for a new space technology company, guiding design, planning, and investment.

Introduction

- Founded in 2002 by Elon Musk, SpaceX has become a global leader in commercial spaceflight and reusable rocket technology.
- The company has completed hundreds of launches, ranging from satellite deployments to International Space Station (ISS) resupply missions.
- SpaceX pioneered reusability, drastically reducing the cost of access to space compared to traditional aerospace firms.
- Its diverse missions provide a rich dataset covering payload types, orbits, launch sites, and booster versions.
- Analyzing SpaceX’s past launch performance offers strategic insights that can inform planning, risk assessment, and technology development for new entrants in the space sector.

Section 1

Methodology

Methodology

Data collection was performed using SpaceX's API and web scraping from Wikipedia.

Data wrangling involved making non-binary classifications into binary ones, as well as dealing with Nan values where applicable.

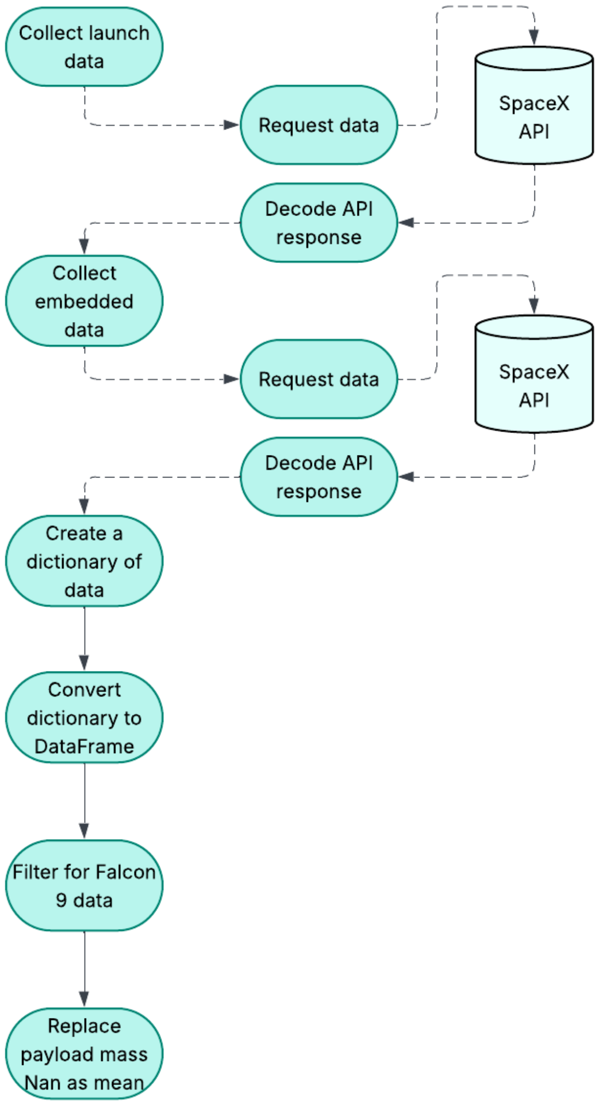
Exploratory data analysis (EDA) was used to derive preliminary insights into the relationship between variables in the dataset and understand more about the data collected, using visualization and SQL queries.

Folium and Plotly Dash were used to create interactive visuals to clearly display data which can aid in the development of a new space technology company.

Predictive analytics using classification models allowed the development of highly precise models to determine which launches would be successful.

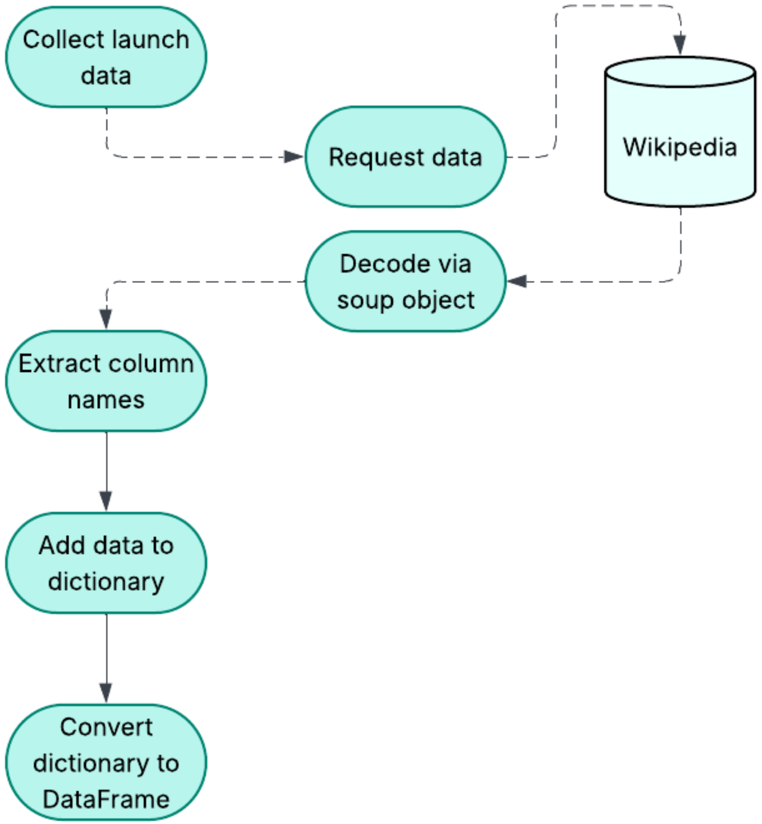
Data Collection – SpaceX API

- The workflow for collecting data with the SpaceX API is presented here.
- The notebook carrying out this work can be found here: [IBM-Data-Science-Course/Capstone-Project/jupyter-labs-spacex-data-collection-api.ipynb at main · TommyEBell/IBM-Data-Science-Course]



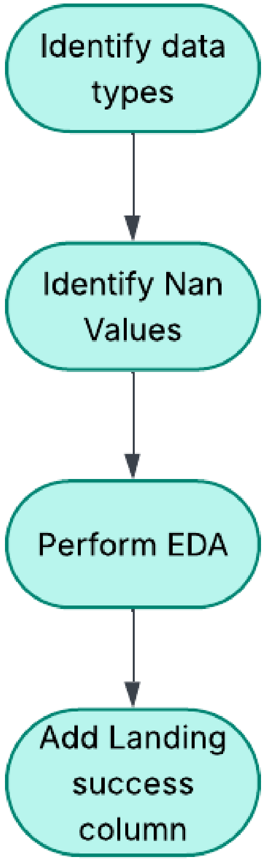
Data Collection - Scraping

- The workflow for data collection via web scraping of Wikipedia pages is displayed.
- The notebook where this process was carried out is found here: [IBM-Data-Science-Course/Capstone-Project/jupyter-labs-webscraping.ipynb at main · TommyEBell/IBM-Data-Science-Course]



Data Wrangling

- The data wrangling process is shown here.
- In this process EDA included:
  - Calculating launches counts from each site.
  - Viewing the mission outcome classifications.
- The notebook where this wrangling was carried out can be found here: [IBM-Data-Science-Course/Capstone-Project/labs-jupyter-spacex-Data wrangling.ipynb at main · TommyEBell/IBM-Data-Science-Course]



EDA with Data Visualization

- The visual exploratory data analysis carried out allowed initial insights into the relationships between variables in the data set.
- The relationships plotted included:
  - How payload mass varied with flight number and mission success
  - How flight number, launch site and mission success were related.
  - How payload mass, launch site and mission success varied.
  - The relation between mission success and orbit type.
  - The relation between payload mass, orbit type and mission success.
  - The success rate per year.
- Modification of the DataFrame, converting multivariable columns into binary variables was also carried out.
- This work can be found here: [IBM-Data-Science-Course/Capstone-Project/edadataviz.ipynb at main · TommyEBell/IBM-Data-Science-Course]()

EDA with SQL

- SQL queries were used to infer more information from the dataset, including:
  - Distinct launch pads
  - Launch site records beginning with ‘CAA’
  - Total payload launched by NASA
  - Average payload using booster F9 v1.1
  - First successful landing
  - Boosters with successful landings carrying payload >4000kg, <6000kg
  - Success vs failure counts of mission outcomes
  - Booster versions which have carried maximum payload
  - Failed landing outcomes from 2015
  - Counts of landing outcomes between 2010-06-04 and 2017-03-20
- This work can be found here: [IBM-Data-Science-Course/Capstone-Project/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · TommyEBell/IBM-Data-Science-Course]()

Build an Interactive Map with Folium

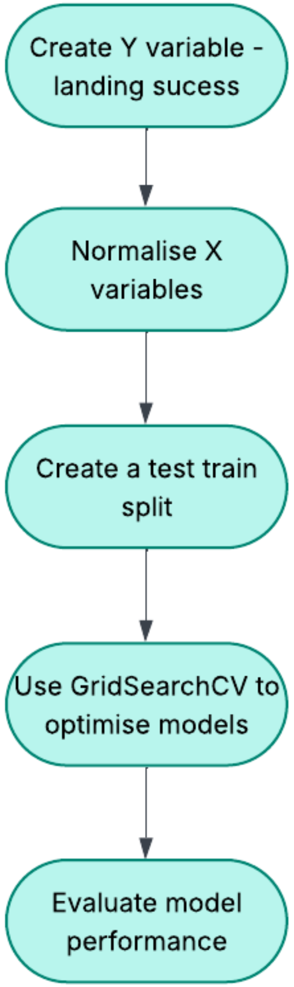
- The interactive map for SpaceX launchpads included:
    - Locations of the Launchpads
    - Each record of a launch from each pad coloured for success/failure.
    - Distance to coastline.
    - Distance to city.
    - Distance to railway.
- These additions help indicate which locations are best suited for building new launchpads for the new company.
- The notebook displaying this work can be found here: [IBM-Data-Science-Course/Capstone-Project/lab_jupyter_launch_site_location.ipynb at main · TommyEBell/IBM-Data-Science-Course]

Build a Dashboard with Plotly Dash

- The dashboard created contains a dropdown menu to select analysis of each launch site or all launch sites combined.
- The analyses shown are a pie chart of the success rate of the launchpad as well as a scatter graph of the success rate varying with payload mass.
- These graphs allow for easy interpretation of the most successful launch sites and payload masses.
- The script for this work can be found using this link: IBM-Data-Science-Course/Capstone-Project/spacex-dash-app.py at main · TommyEBell/IBM-Data-Science-Course

Predictive Analysis (Classification)

- The models evaluated included:
  - Logistic regression
  - SVM
  - Decision Tree
  - K-nearest neighbour
- The link for this notebook is provided here: [IBM-Data-Science-Course/Capstone-Project/SpaceX_Machine Learning Prediction_Part_5 (1).ipynb at main · TommyEBell/IBM-Data-Science-Course]



Section 2

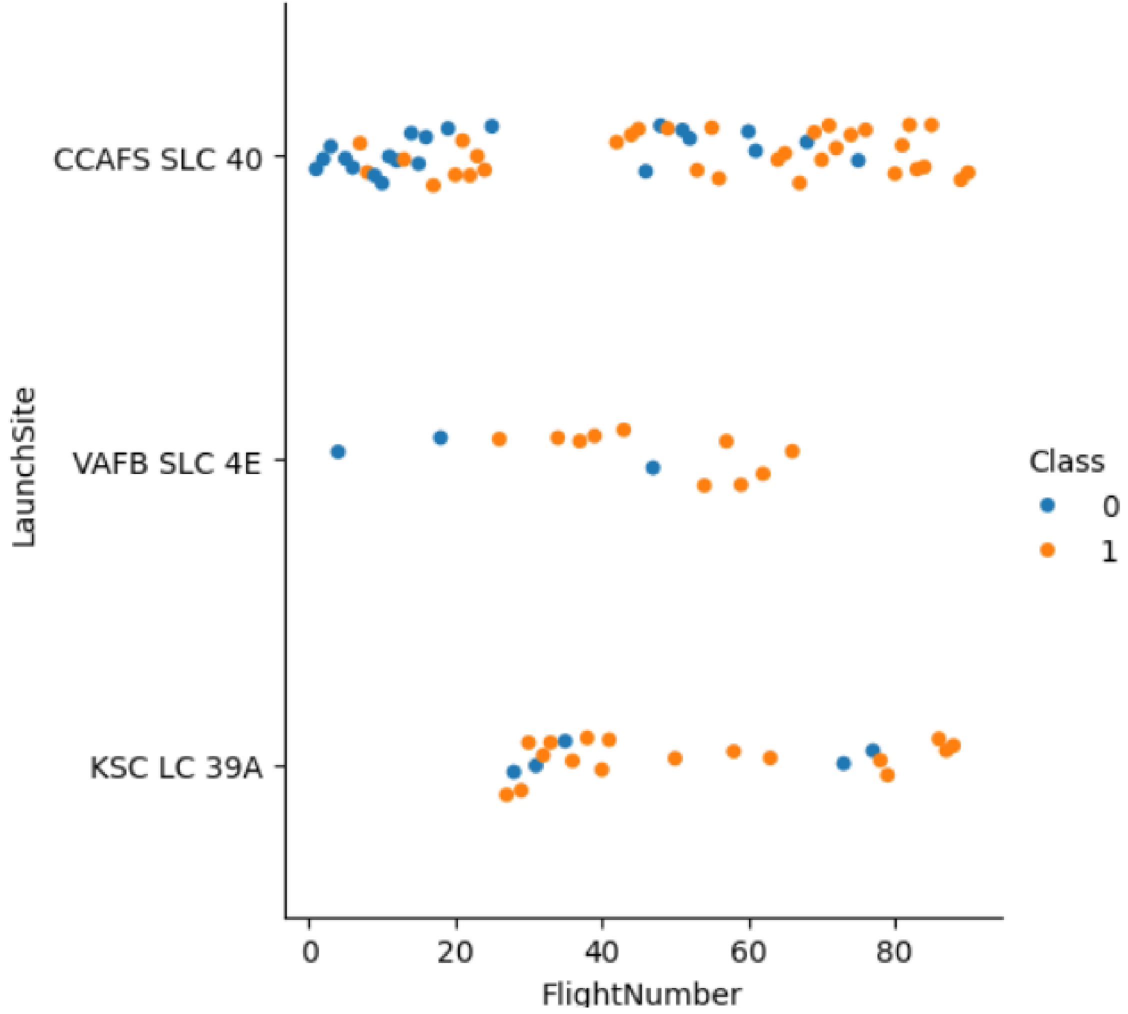
Insights drawn from EDA

Flight Number vs. Launch Site

'0' or blue represent a failure to land the rocket, with '1' or orange representing success.

This graph shows that CCAFS SLC 40 was primarily the first launch site used and remains the most used.

There were a lot more failures in the beginning than towards the end, showing increased success.

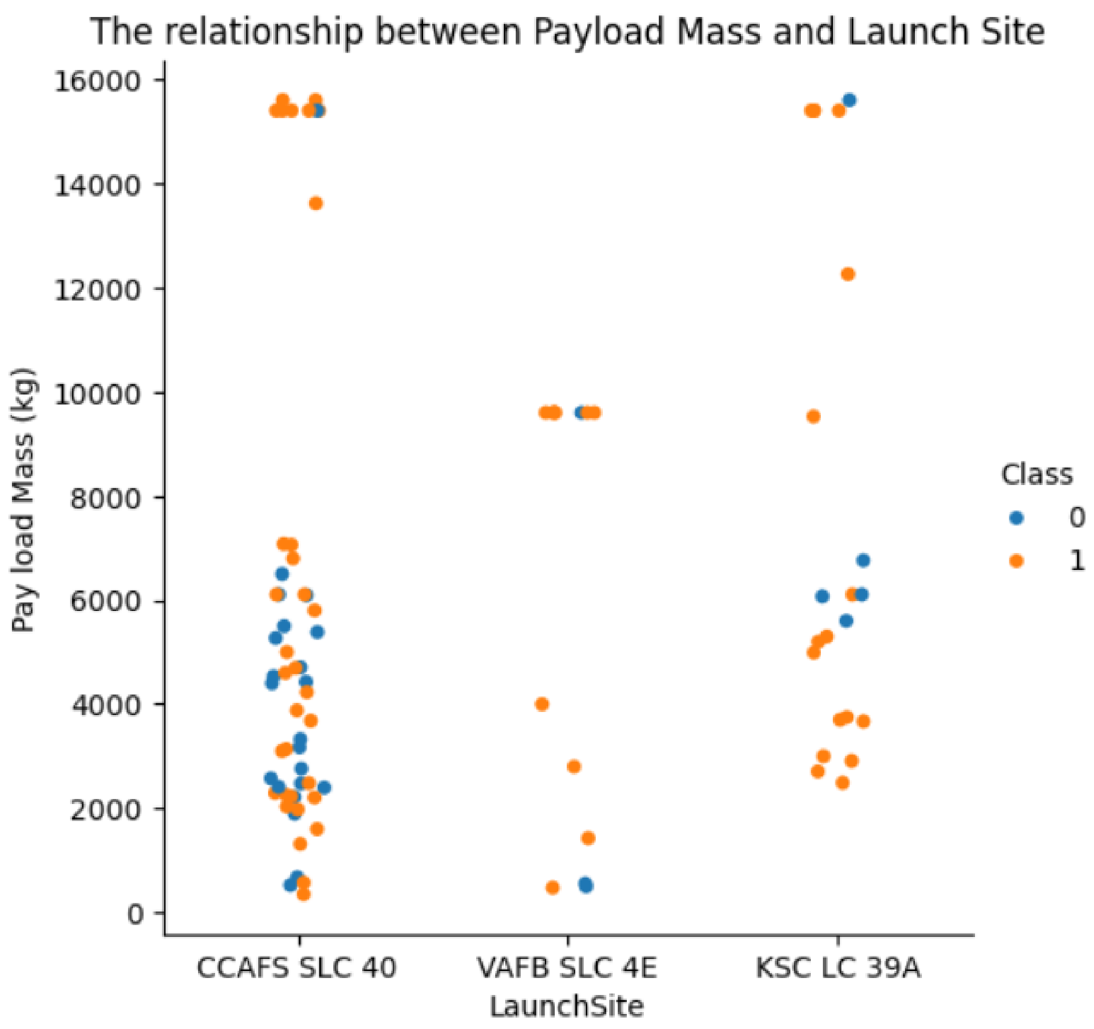


Payload vs. Launch Site

'0' or blue represent a failure to land the rocket, with '1' or orange representing success.

The VAFB SLC 4E launch site appears to have a much lower payload limit than the other two.

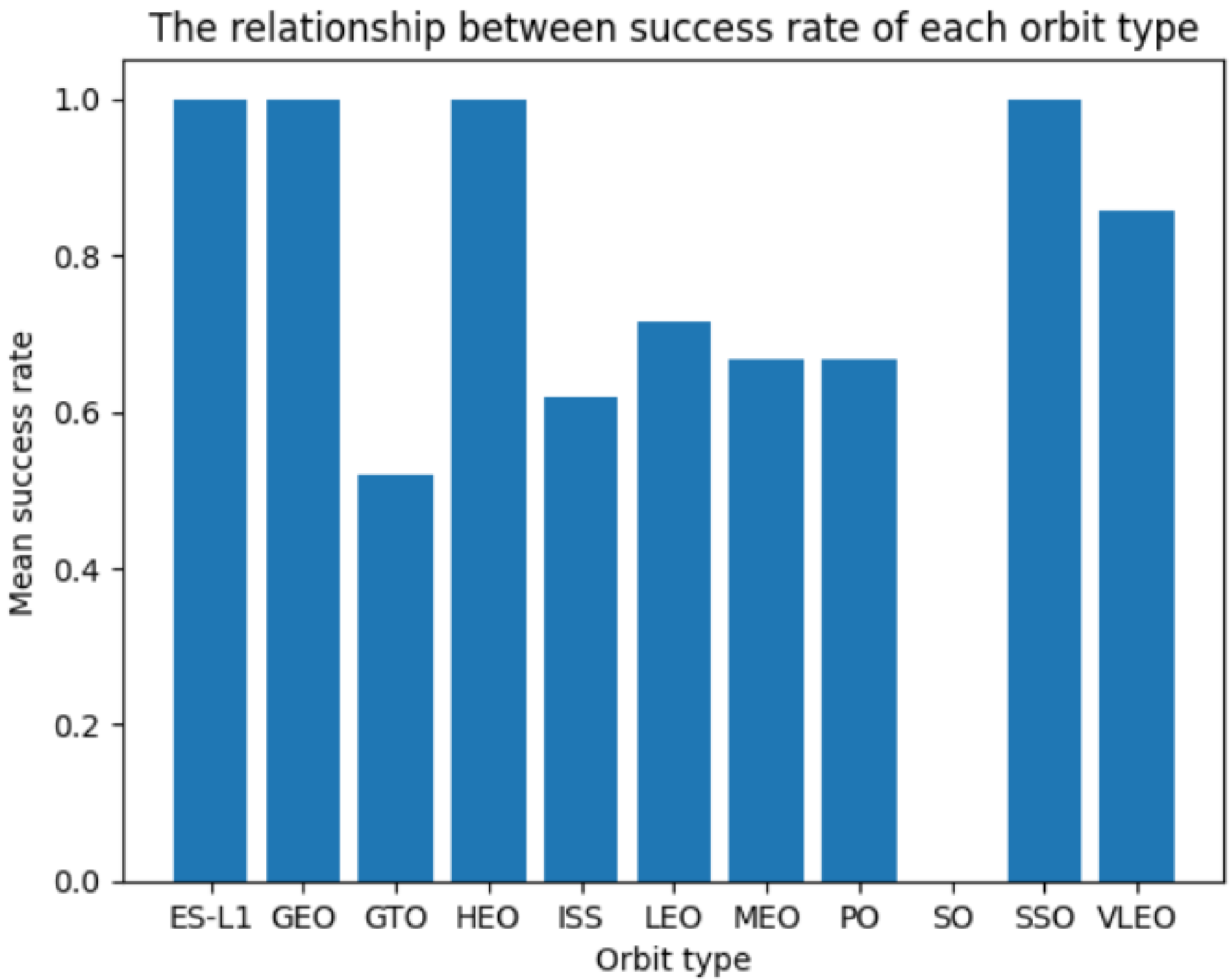
Generally, higher payload masses appear to have a greater success rate, however there is likely a temporal aspect not represented here.



Success Rate vs. Orbit Type

Most orbit types have a high success rate with only a Semi-Synchronous Orbit (SO) having a 0% success rate.

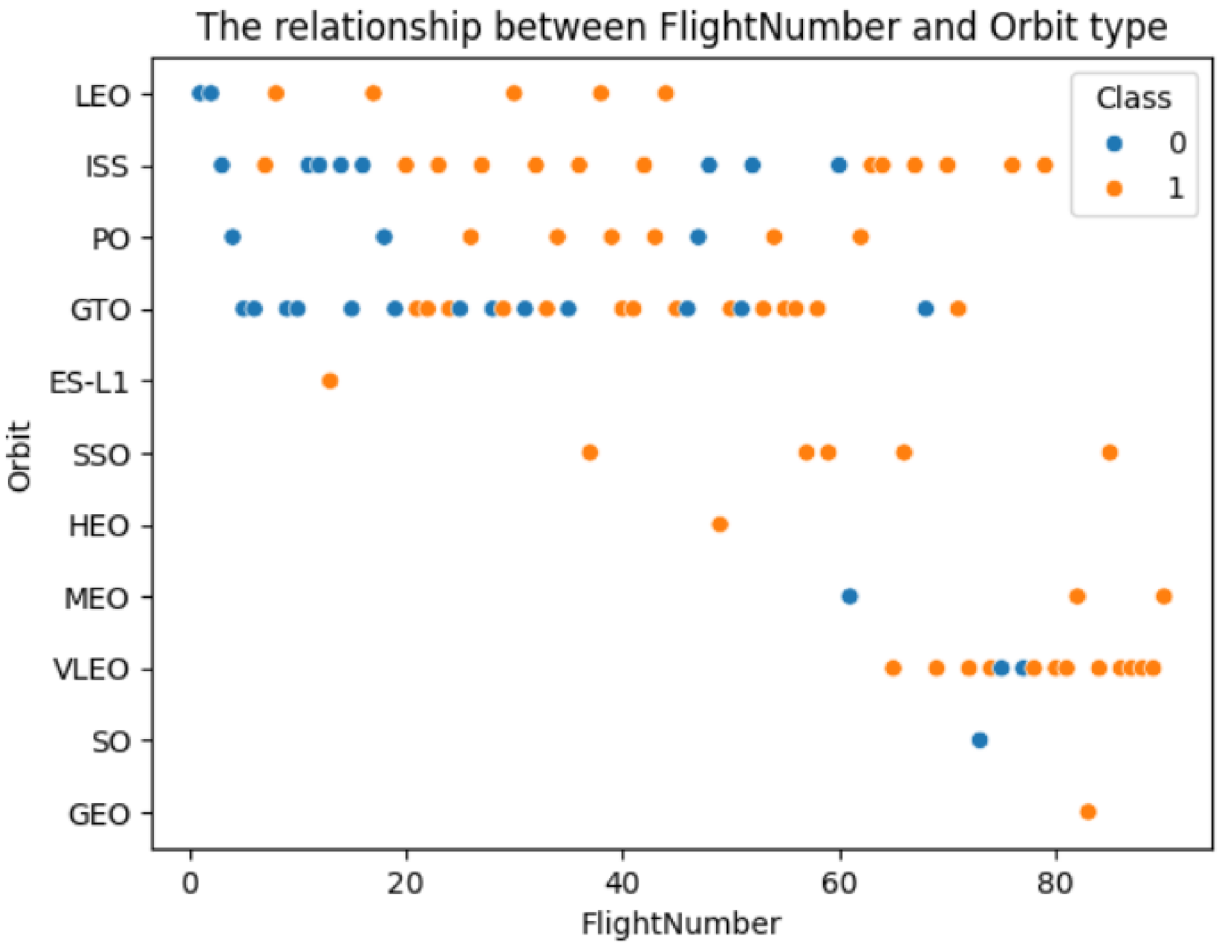
ES-L1, GEO, HEO, and SSO all have a 100% success rate, indicating an easier launch plan to begin testing.



Flight Number vs. Orbit Type

‘0’ or blue represent a failure to land the rocket, with ‘1’ or orange representing success.

This graph highlights the counts of each orbit type tested, highlighting accuracy bias, including SO which only has 1 test.

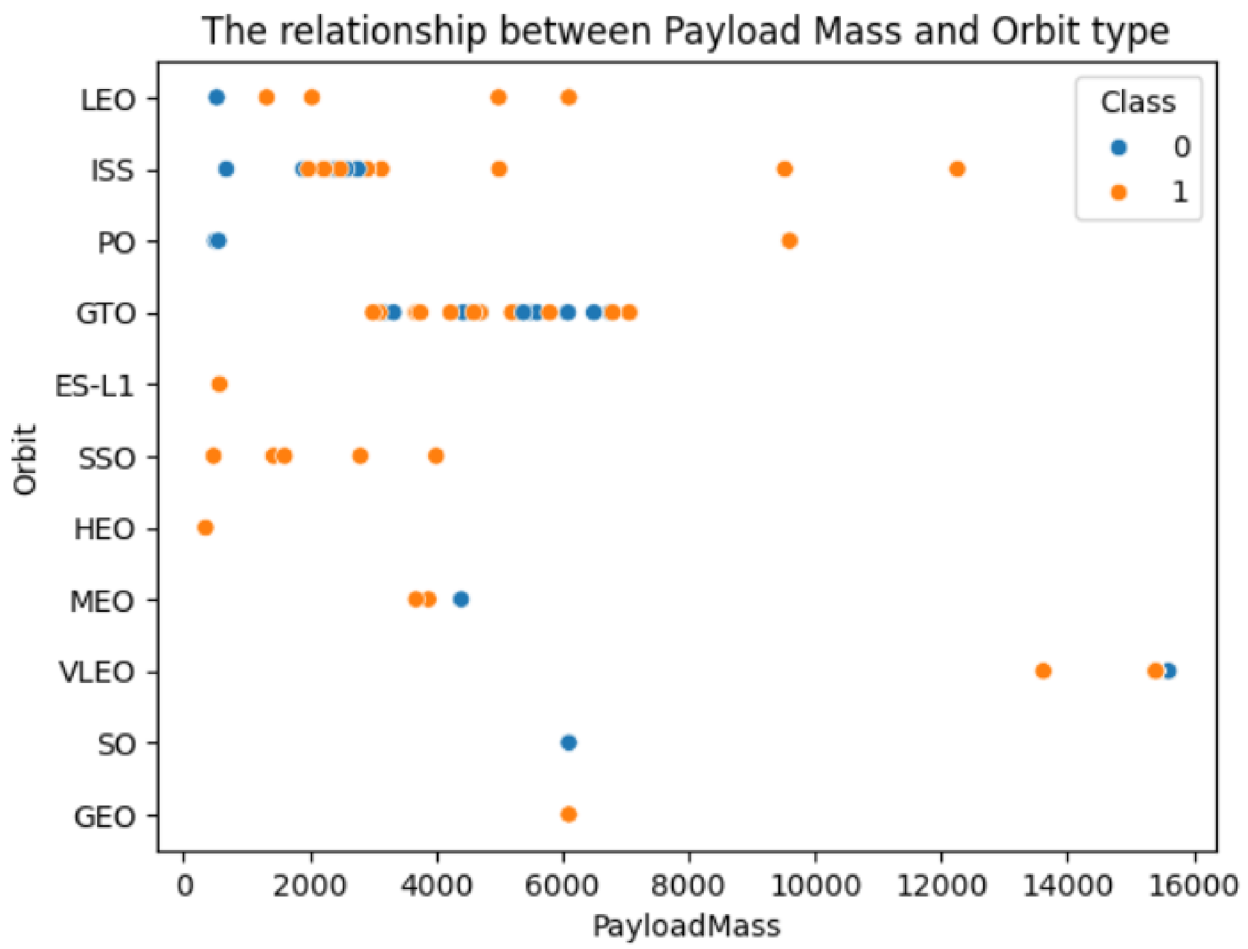


Payload vs. Orbit Type

‘0’ or blue represent a failure to land the rocket, with ‘1’ or orange representing success.

Some orbit types, such as MEO, GTO and ISS, appear to mostly have a tight range of payload masses, possibly due to limitations in technology or repeated tests.

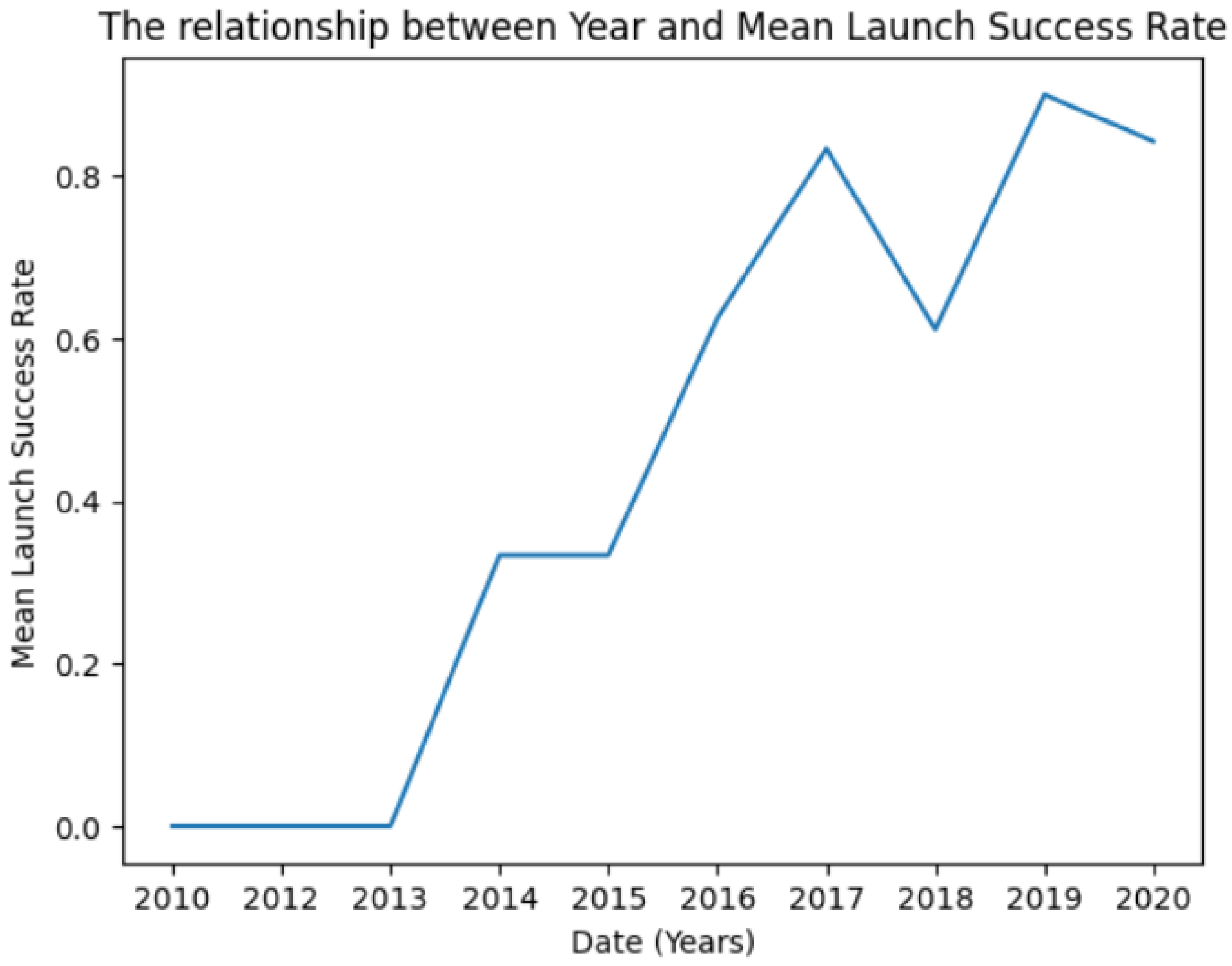
Again, the lack of data in some orbit types highlights the potential inaccuracies in quantitative analyses.



Launch Success Yearly Trend

The clear increase in success per year highlights improved understanding of launches by SpaceX, a baseline which a new company could use to compare progress.

This graph may suffer from large uncertainties with years which have few launch attempts.



All Launch Site Names

The unique launch sites used by SpaceX are:

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

This insight was obtained from the following SQL query:

```
%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTBL
```

Launch Site Names Begin with 'CCA'

The first 5 records with launch sites whose name began with ‘CCA’ are all from CCAFS LC-40.

This insight was gained from the following query:

```
%sql SELECT Launch_Site FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

The total payload carried by boosters from NASA was 45,596kg.

This value was obtained using the following query:

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer IS 'NASA (CRS)'
```

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 was 2928.4kg.

The query used to derive this insight was:

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version IS 'F9 v1.1'
```

The date of the first successful landing outcome on a ground pad was 22/12/2015.

This was obtained using the following query:

```
%sql SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome is 'Success (ground pad)'
```

Successful Drone Ship Landing with Payload between 4000 and 6000kg

The boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:

F9 FT 81022

F9 FT 81026

F9 FT 81081.2

F9 FT 81031.2

This insight was obtained via the following SQL query:

```
%sql SELECT DISTINCT(Booster_Version) FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

```
AND Landing_Outcome IS 'Success (drone ship)'
```

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes was 100 to 1.

This was extracted using the following query:

```
%%sql
SELECT
  CASE
    WHEN Mission_Outcome LIKE '%Success%' THEN 'Success'
    ELSE 'Other'
  END AS Outcome_Group,
  COUNT(*) AS Total
FROM SPACEXTBL
GROUP BY Outcome_Group
```

Boosters Carried Maximum Payload

The boosters which have carried maximum payload mass are given in the table on the right.

The query to obtain this results is shown below:

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

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

2015 Launch Records

- The List of failed landing outcomes in drone ship with their booster versions and launch site names for 2015 is shown below. Where the first coloumn represents the month in which the record occurred.

substr(Date, 6,2)	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

- The query used to extract this information is given below;

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

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

- The ranked counts of landing outcomes between the dates 2010-06-04 and 2017-03-20 is shown in the table on the right.

Landing Outcome	Count
No attempt	10
Success (Ground Pad)	5
Success(Drone Ship)	4
Failure (Drone Ship)	3
Precluded (Drone Ship)	3
Controlled (Ocean)	2
Uncontrolled (Ocean)	1
Failure (Parachute)	1

- The query to extract this information is given below:

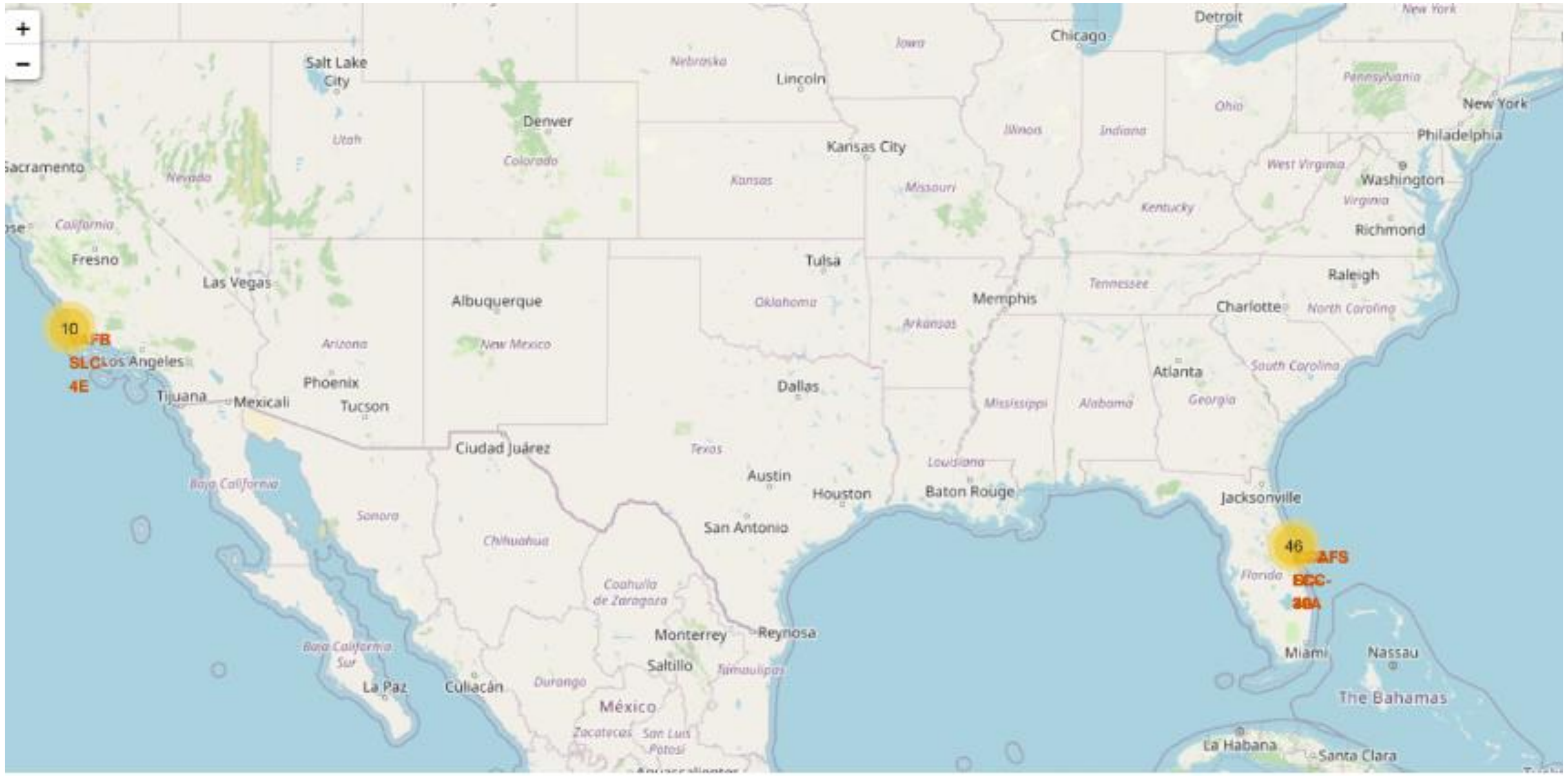
```
sql SELECT Landing_Outcome FROM SPACEXTBL WHERE substr(Date, 0, 11)  BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY Date DESC
```

Section 3

Launch Sites Proximities Analysis

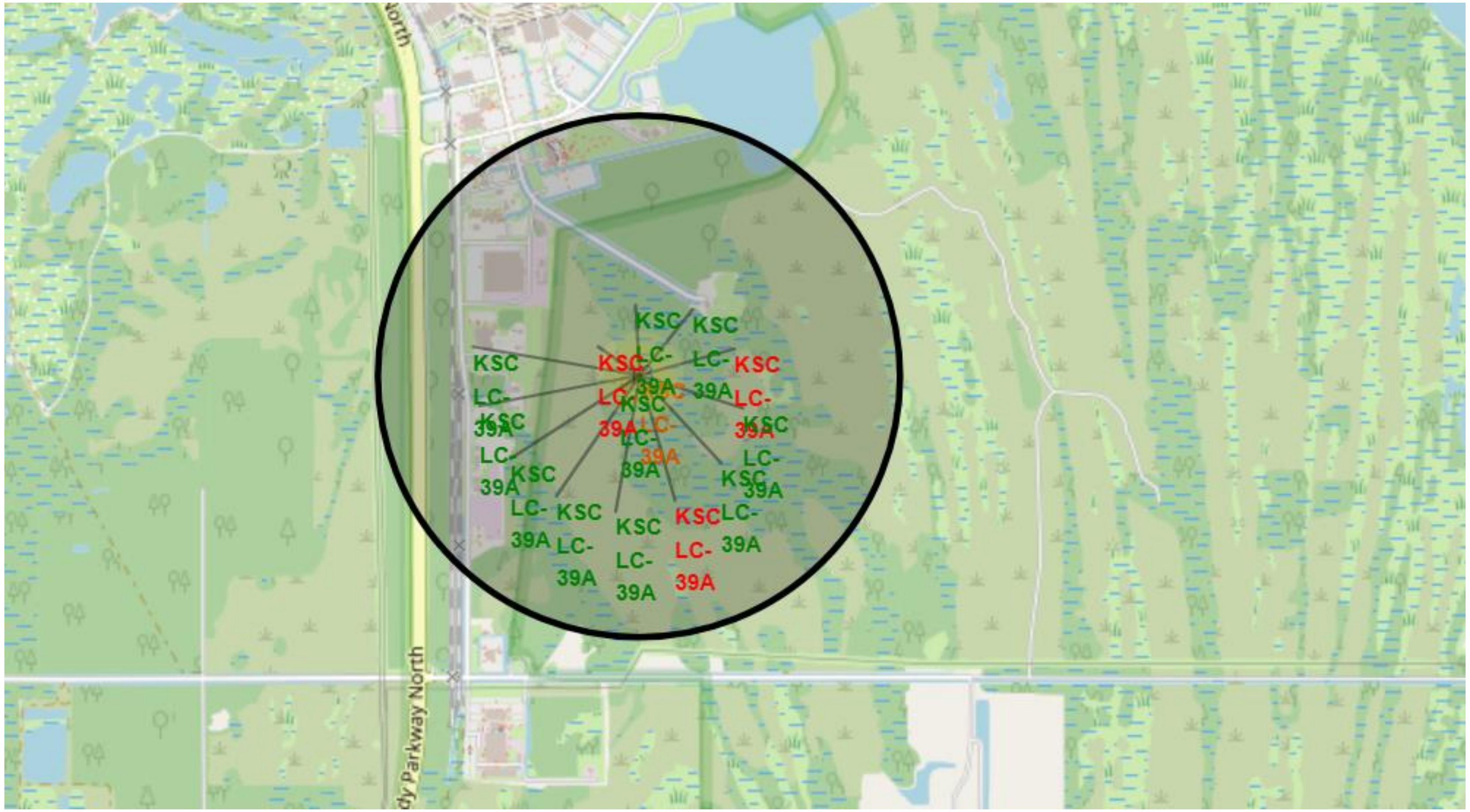
Launch site use and location

- The map shows the launch sites along with a count of the number of launches carried out from each site.



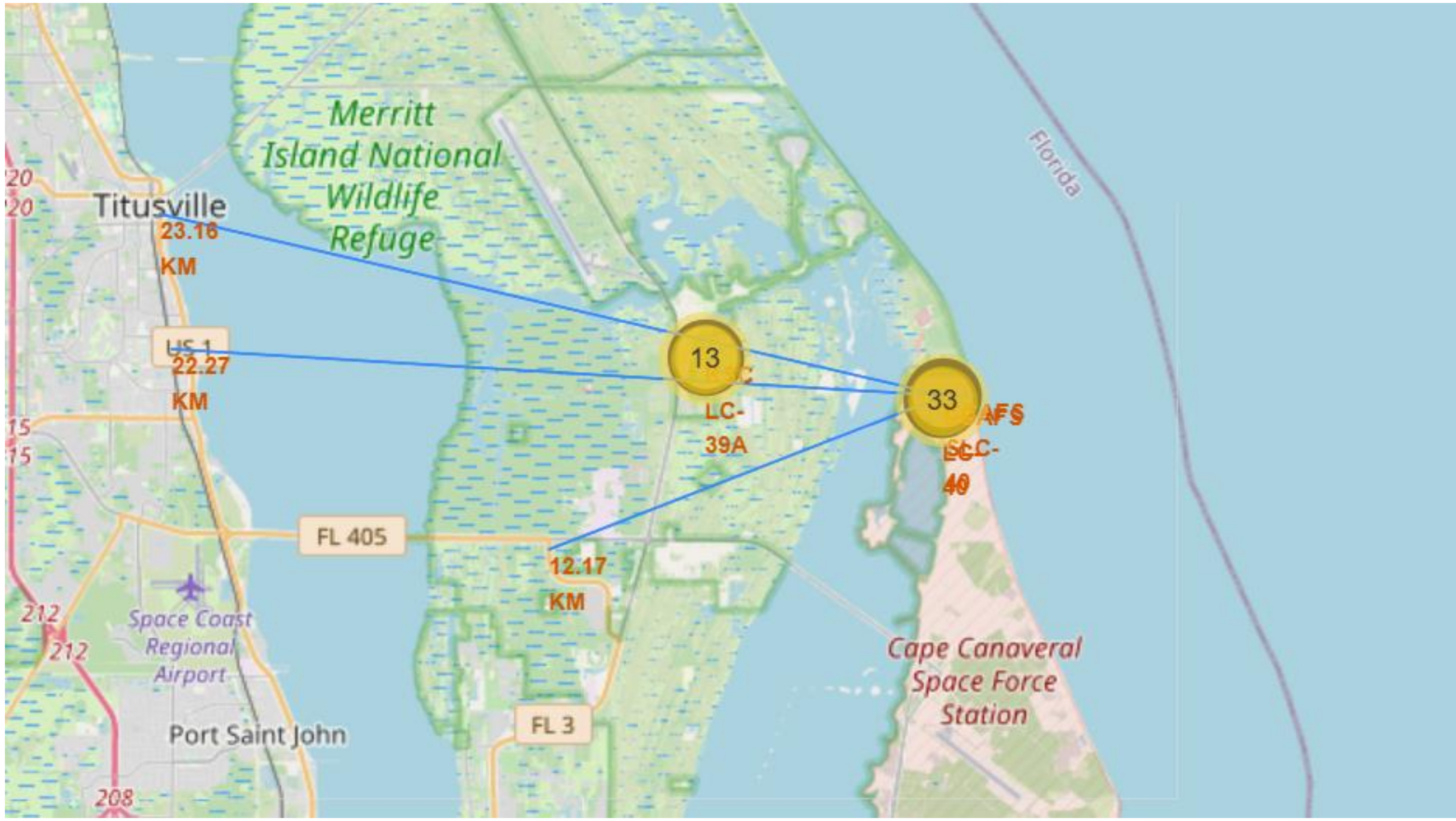
Launch site success

- Selecting on each launch site displays the successful and failed landing records.



Launch site hyperparameters

- The map also shows the distance to the nearest main road, beach, city, and railway.
- This helps to identify suitable locations for launch sites.

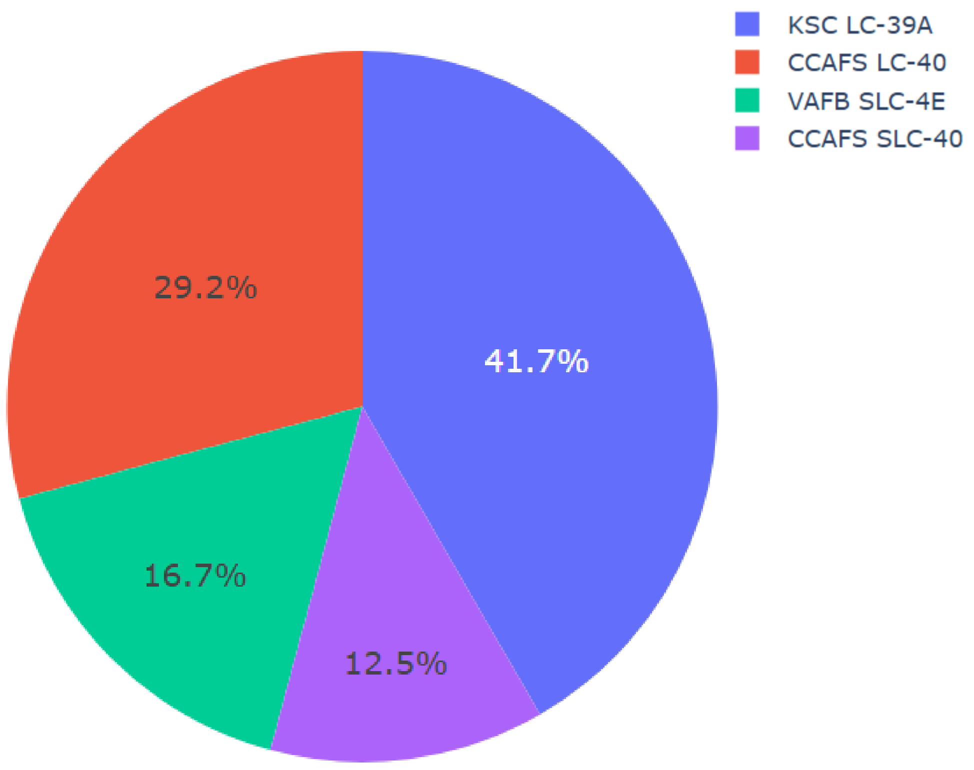


Section 4

Build a Dashboard with Plotly Dash

Success of Launches for All Sites

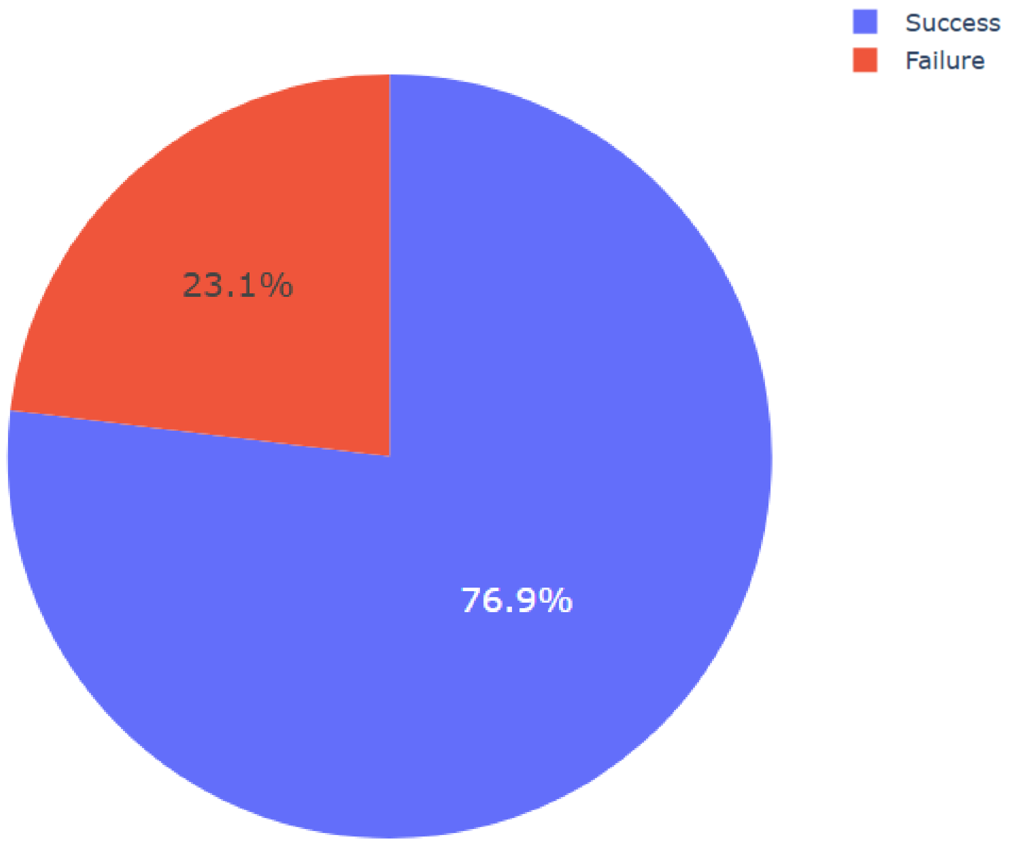
- The number of successes for each site is displayed in a pie chart in the dashboard.
- This displays in a visually simplistic way how KSC LC-39A has the greatest number of successes compared to the rest.



Success rate of a launch site

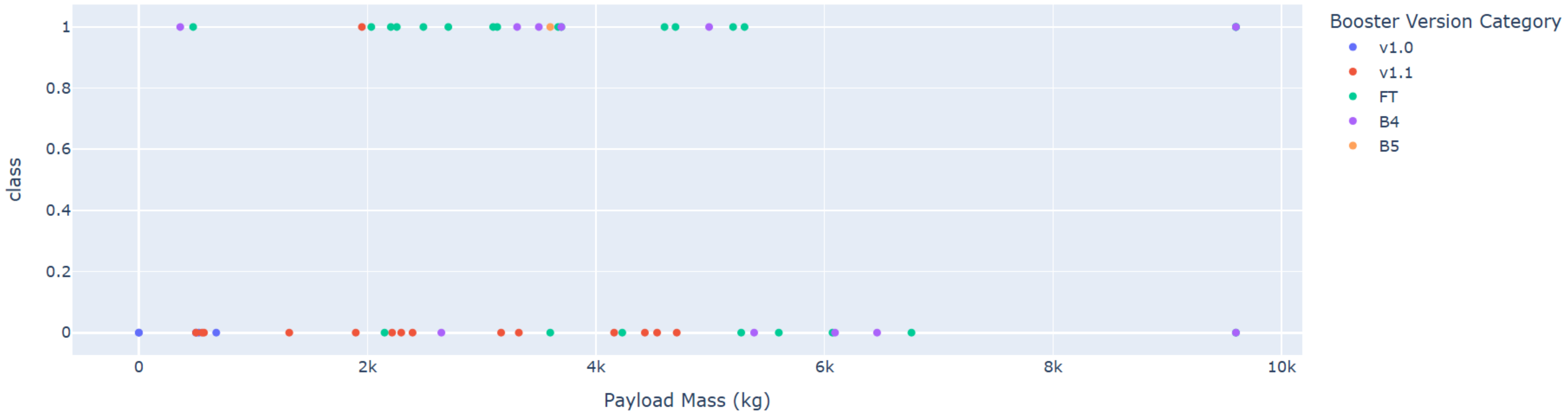
- KSC LC-39A also has the highest success rate of all launch sites, at 76.9%.

- This figure can be displayed for any of the launch sites by selecting the launch site of interest from the dropdown menu at the top of the dashboard.



The relationship between payload mass and success of launches

- This figure shows the success of the rockets with different payload masses, launched from all sites.
- The range of payload masses displayed can be altered using the adjustable slider above the figure.

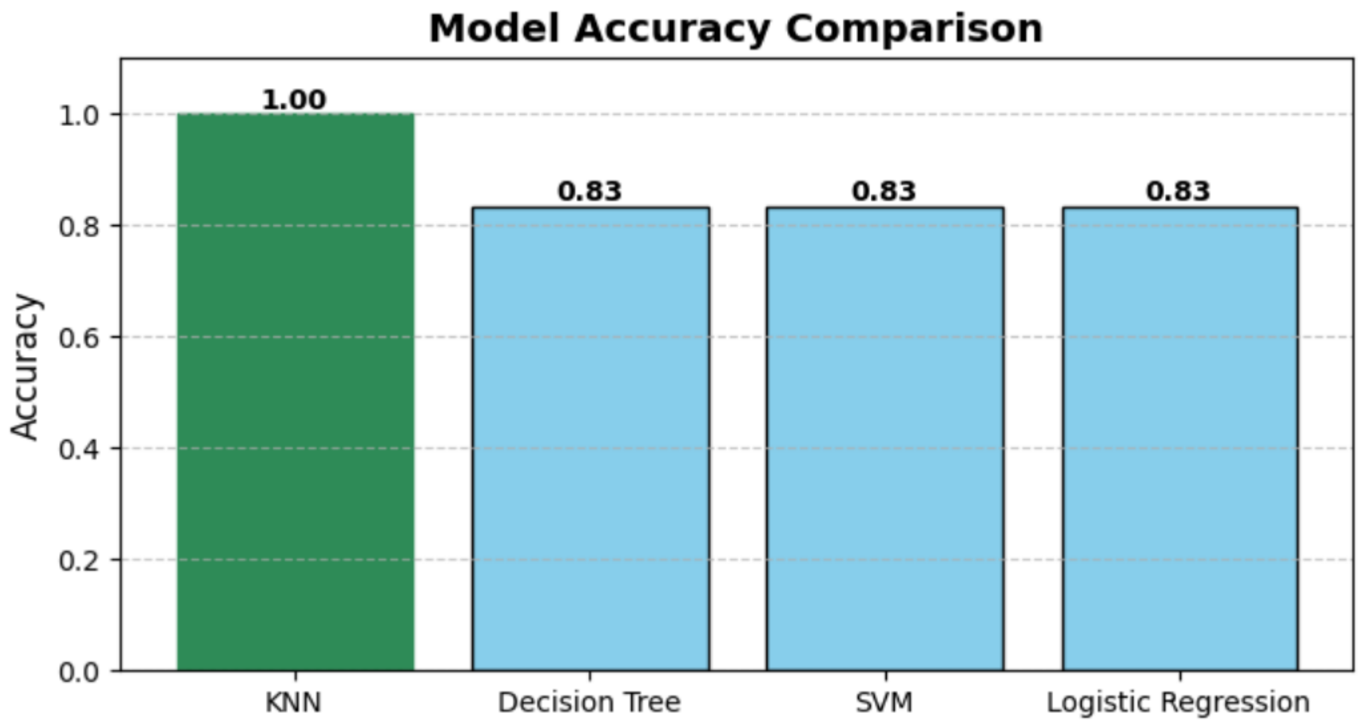


Section 5

Predictive Analysis (Classification)

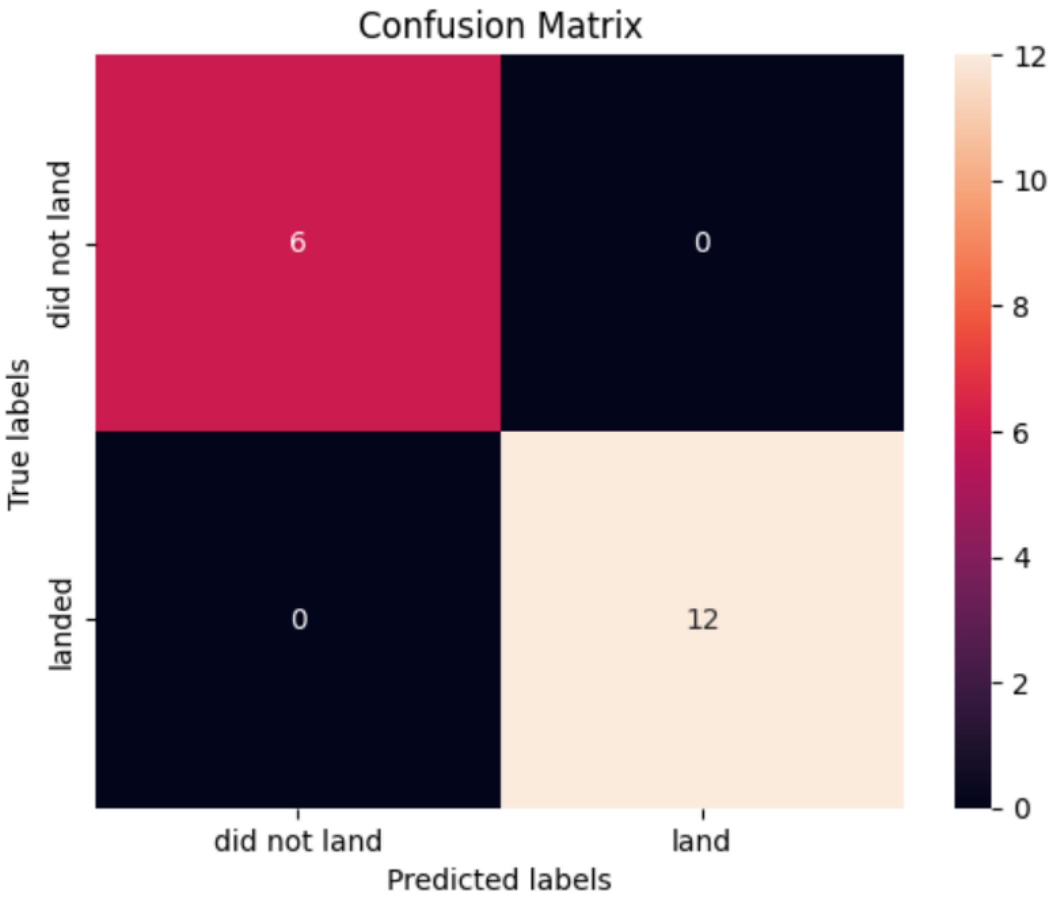
Classification Accuracy

- The highest accuracy model that was train was k-nearest neighbours.
- All other models had a similar score ~83%.



Confusion Matrix

- The confusion matrix for KNN shows perfect classification with no false positive or negatives.



Conclusions

- Interactive dashboards (Folium & Plotly Dash) provided clear visual insights into launch performance.
- EDA revealed key patterns and relationships influencing launch success.
- KNN achieved high accuracy in forecasting successful launches, showing potential for use in the future to steer the design of launch plans.
- All these findings can guide decision-making for new space technology ventures.
- This these insights are informative, some aspects such as calculating orbit type launch successes fall victim to low data counts, leading to unreliable conclusions.

Thank you!