IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mark Tran
2/8/25

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- This project aims to find important factors that affect the success rate of SpaceX Falcon 9's booster landing and build machine learning models to predict the landing of the boosters.
- The study suggests payload mass, launching site, and launching year are among the most important indicators of landing success.
- Through the project, I have also built and optimized multiple machine learning models that are able to predict the landing success at a high accuracy. The best-performing model is Decision Tree, with a testing accuracy rate of 0.89.

Introduction

- Making space exploration affordable and available to the general public has gained a lot of interest recently. The leader in this field is SpaceX, with their reusable Falcon 9 spaceship boosters. Successful booster landings are affected by multiple factors.
- This project's goal is to identify which factors are important to the booster landing success. Another goal is to accurately predict the success of a landing based on past data.
- The results of this project could help scientists improve on the booster designs as well as launch plans to increase booster landing success.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from SpaceX REST API and web scrapping Falcon 9 Launch records from Wikipedia
- Perform data wrangling
  - Data was processed by imputing missing data and transforming to suitable formats
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Machine learning models were built, tuned, and evaluated based on accuracy scores

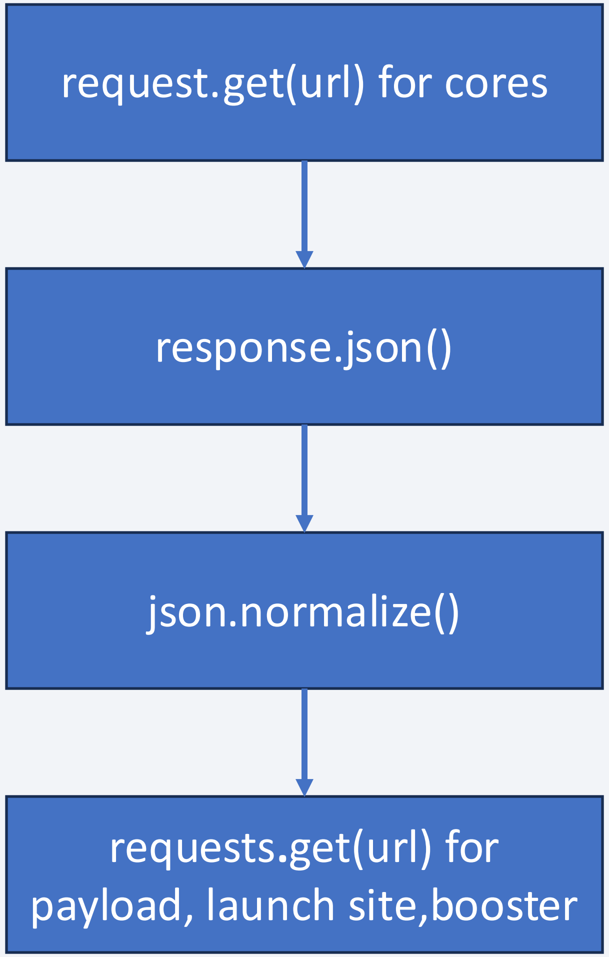
Data Collection

- The data for this project was collected from 2 sources:
  - SpaceX REST API
  - Falcon 9 Launch Records Wikipedia
- The SpaceX REST API was accessed and downloaded using the Request package in Python
- Falcon 9 Launch Record Wikipedia was downloaded by web scraping using the Request and BeatifulSoup packages in Python

Data Collection – SpaceX API

- Collected data using the Request package in Python.
- Request.get(API url)
- Reformat collected data into dataframe
- Github URL:

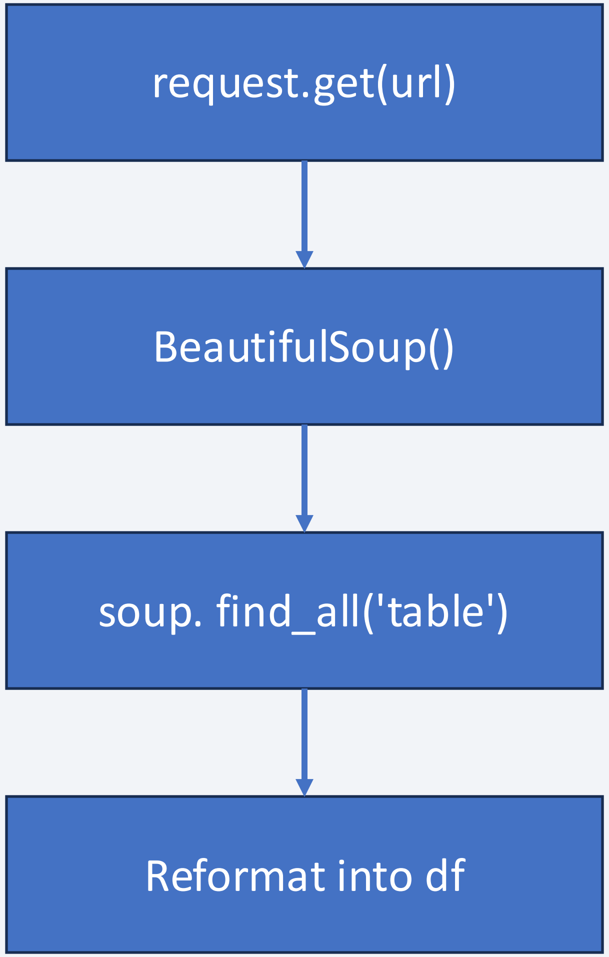
https://github.com/vietbachtran/SpaceX-IBM-Applied-Data-Science/blob/a1b3eea61cbf087612e218c73f34f9912eb014af/Codes/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Collected data from the Falcon 9 launch record Wikipedia page (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Web scrap using the BeuatifulSoup package in Python
- GitHub URL:

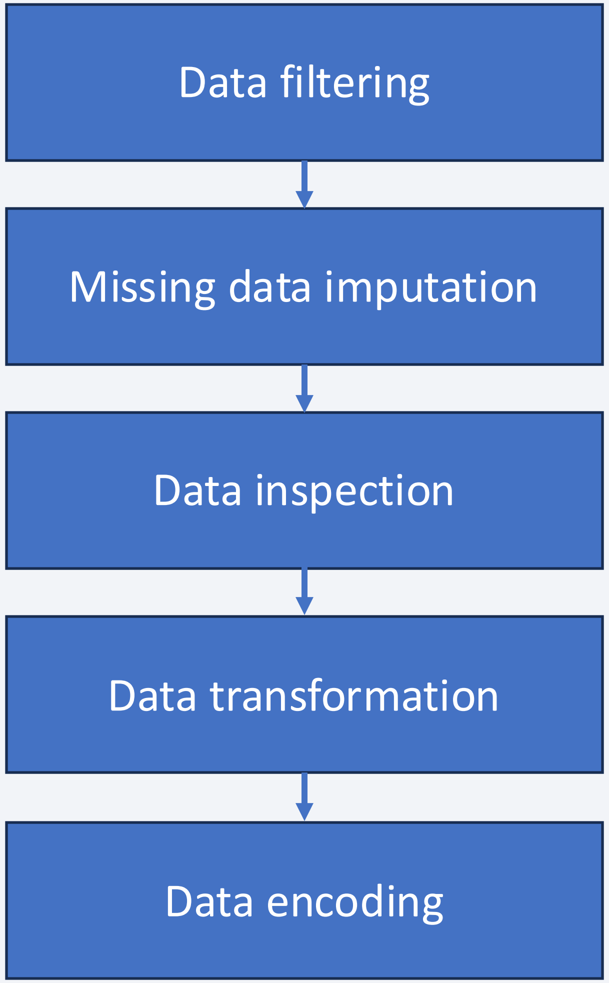
https://github.com/vietbachtran/SpaceX-IBM-Applied-Data-Science/blob/a1b3eea61cbf087612e218c73f34f9912eb014af/Codes/jupyter-labs-webscraping.ipynb



Data Wrangling

- Select only launches for Falcon 9
- Inspect Missing values and impute them with column mean
- Inspect data types and change to the correct types
- Inspect values in columns to make sure no unusual occurrences
- Transform categorical landing outcome into binary
- GitHub URL:

https://github.com/vietbachtran/SpaceX-IBM-Applied-Data-Science/blob/a1b3eea61cbf087612e218c73f34f9912eb014af/Codes/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Scatter plots between 2 features with color encoding landing outcome. This helps reveal the correlation between 2 features and how they affect the outcome.
- Bar plots and line plots showing the landing success rates of different categorical data. This helps reveal the effect of single features on the outcome.
- EDA helped identify important features for subsequent analyses.
- GitHub URL:

https://github.com/vietbachtran/SpaceX-IBM-Applied-Data-Science/blob/a1b3eea61cbf087612e218c73f34f9912eb014af/Codes/edadataviz.ipynb

EDA with SQL

- Find unique launch sites
- Find 5 records where launch sites begin with `CCA`
- Find the total payload mass carried by NASA (CRS)
- Find the average payload mass carried by booster version F9 v1.1
- Find when the first successful landing in a ground pad happened
- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Find the total number of successful and failure mission outcomes
- List the names of the booster which have carried the maximum payload mass
- List the failed landing outcomes in drone ship, their booster versions, and launch site names for 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
- GitHub URL:

https://github.com/vietbachtran/SpaceX-IBM-Applied-Data-Science/blob/a1b3eea61cbf087612e218c73f34f9912eb014af/Codes/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Create a folium world map and add the following objects:

- Markers and circles for all launch sites. This helps viewers visualize the location of sites.
- Icons marking individual launch outcomes for each site. Red: failed, green: successful. This helps viewers notice which sites have a higher success rate than others.
- Lines and distances of launch sites to the closest coastlines, railroads, highways, and cities. This could help explain the success rate of a site and for better launch planning.
- GitHub URL:

https://github.com/vietbachtran/SpaceX-IBM-Applied-Data-Science/blob/a1b3eea61cbf087612e218c73f34f9912eb014af/Codes/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

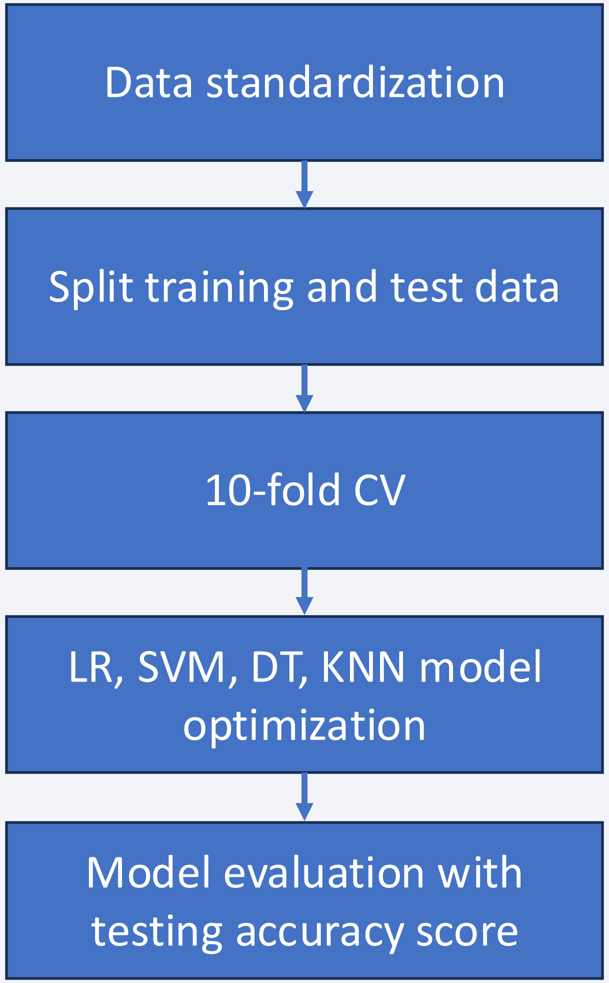
- Users can select all launch sites or a single launch site to further investigate
- Pie chart showing the number of total landing successes for the selected launch site. This would help users understand the success rate of the specific launch sites and make it easy to compare landing successes across sites.
- Users can select the payload range to inspect the correlation between payload mass and landing outcome.
- Scatter plot showing the correlation between payload mass within the selected range and landing outcome. This would help users understand how payload mass affects the landing outcome in different ranges.
- GitHub URL:

https://github.com/vietbachtran/SpaceX-IBM-Applied-Data-Science/blob/a1b3eea61cbf087612e218c73f34f9912eb014af/Codes/spacex_dash_app.py

Predictive Analysis (Classification)

- Standardize the data
- Split data into training and test data with test data size of 0.2 of the original dataset
- Fit models with GridCVSearch 10-fold cross-validation of training data to tune hyperparameters
- Optimize Logistic regression, Support vector machine, decision tree, and K-nearest neighbor models
- Evaluate model performance with accuracy score and confusion matrix of test data
- Select the model with the highest test accuracy score as the best-performing model
- GitHub URL:

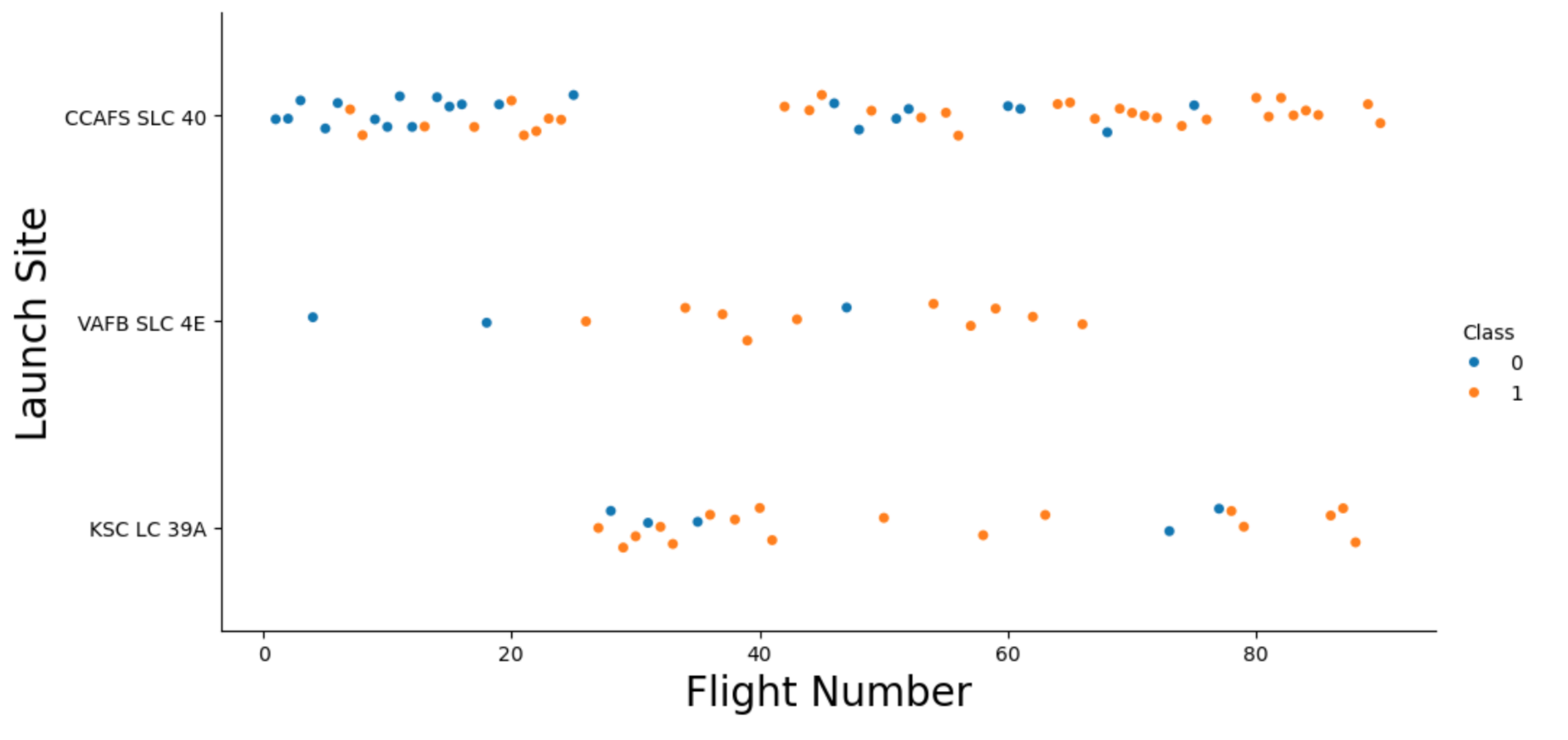
https://github.com/vietbachtran/SpaceX-IBM-Applied-Data-Science/blob/5ae7e2a57870d2b7f131c8ac071abf3e88fa0053/Codes/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Section 2

Insights drawn from EDA

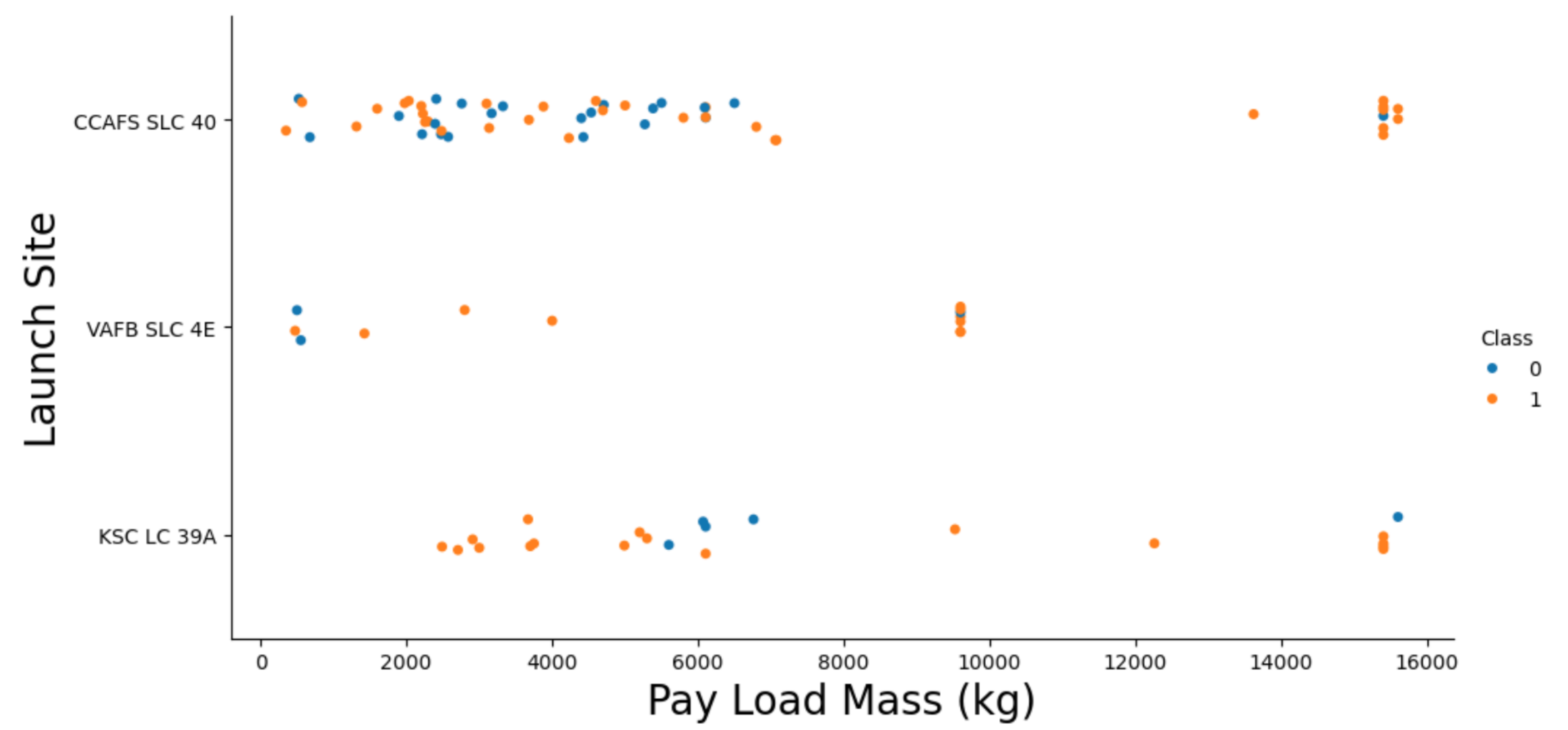
Flight Number vs. Launch Site



The overall trend in all launch sites is higher flight numbers are more likely to be successful.

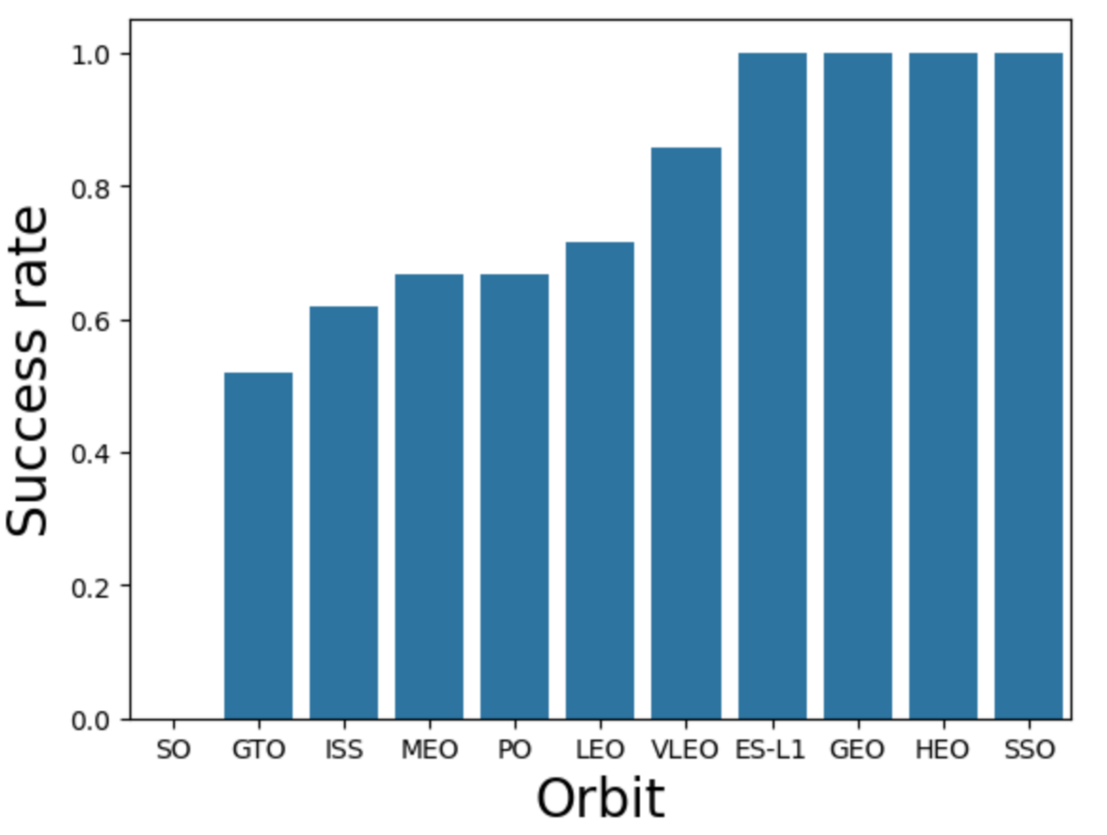
This suggests that later flights have a higher success rate than earlier flights.

Payload vs. Launch Site



Flights with heavier payloads were launched from CCAFS SLC 40 and KSC LC 39A. The flights with heavier pay load mass were more likely to be successful.

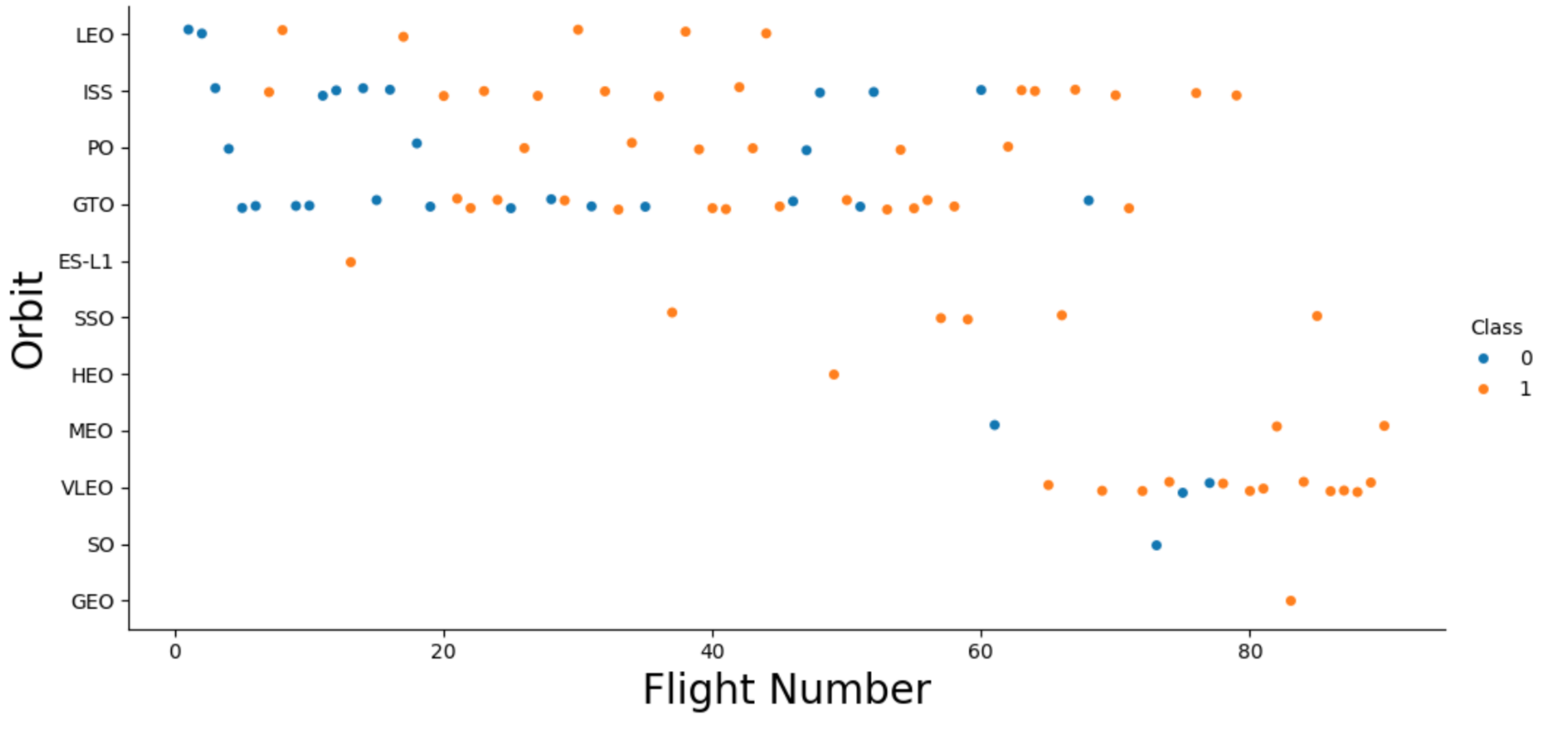
Success Rate vs. Orbit Type



SO orbit has the lowest success rate of 0%. This suggests that none of the flights orbiting SO ever landed successfully.

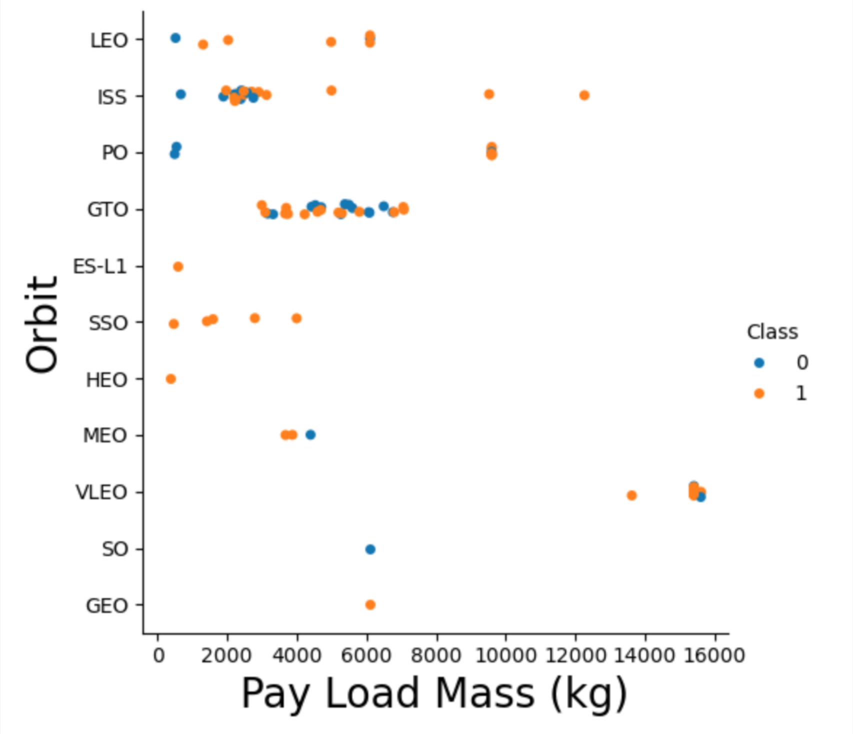
ES-L1, GEO, HEO, and SSO orbits have the highest success rates of 1, indicating all flights with these orbits landed successfully.

Flight Number vs. Orbit Type



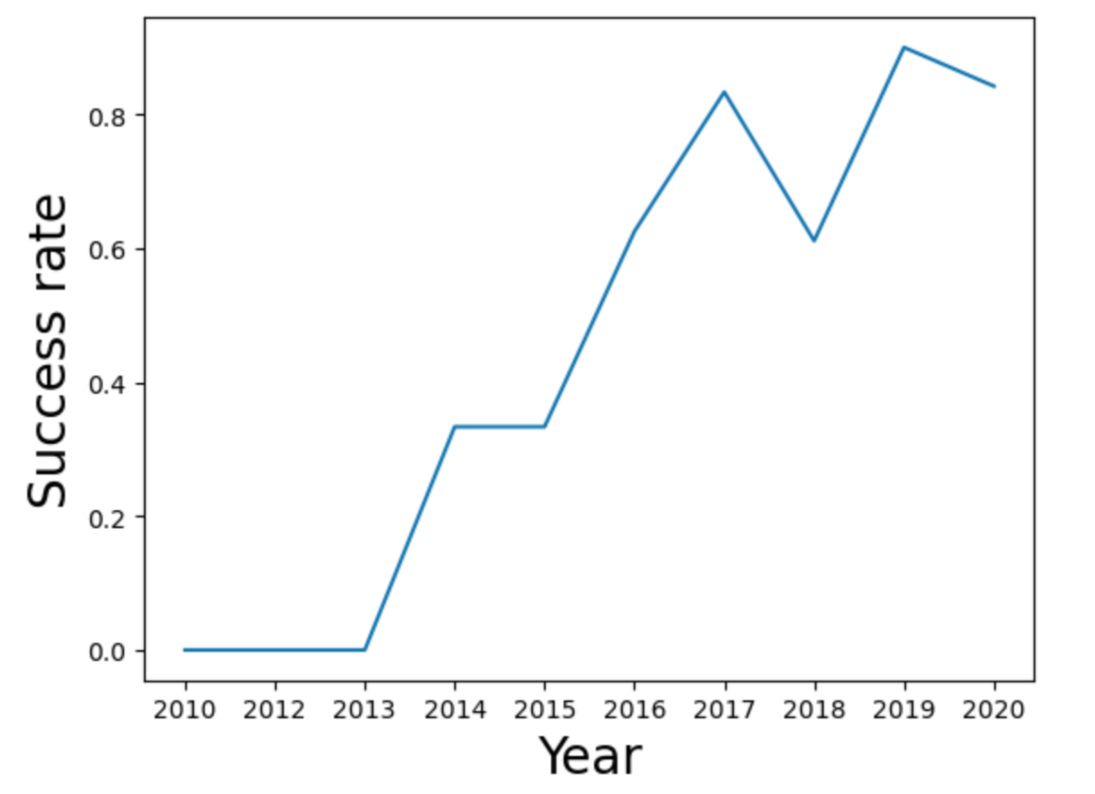
Earlier flights had orbits LEO, ISS, PO, GTO, and ES-L1, while more recent flights orbited SSO, HEO, MEO, VLEO, SO, and GEO.

Payload vs. Orbit Type



Flights with VLEO orbit can take a high pay load mass, while other orbits usually have a lower pay load mass.

Launch Success Yearly Trend



Over the years, the success rate has increased steadily. This suggests that the newer technologies and updates have significantly increased the success rate.

However, the success rate dropped in 2018. This calls for further investigation in 2018 launches.

All Launch Site Names

- The SQL query shows that there are 4 distinct launch sites in our dataset.

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

Launch Site Names Begin with 'CCA'

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

This SQL query shows the first 5 launches from a launch site that begins with ‘CCA’.

Total Payload Mass

- The total payload mass carried by NASA (CRS) mission is calculated using a group by SQL query.

Total Payload Mass by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was found using SQL query.

Average Payload mass by F9 v1.1 booster
2534.6666666666665

First Successful Ground Landing Date

- The first successful landing outcome on ground pad in our data dates back to 2015.

First successful landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The SQL query shows a full list of booster names that have successfully landed on a drone ship and had a payload mass greater than 4000 but less than 6000 kg.

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

- The SQL query calculates the total number of successful and failure mission outcomes
- There are 1 failed mission and 100 successful missions.

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

Boosters Carried Maximum Payload

- The SQL query lists the names of the booster which have carried the maximum payload mass

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

Month	failure landing_outcomes in drone ship	Booster_Versions	Launch_Site
01	1	F9 v1.1 B1012	CCAFS LC-40
02	0	F9 v1.1 B1013	CCAFS LC-40
03	0	F9 v1.1 B1014	CCAFS LC-40
04	1	F9 v1.1 B1015,F9 v1.1 B1016	CCAFS LC-40
06	0	F9 v1.1 B1018	CCAFS LC-40
12	0	F9 FT B1019	CCAFS LC-40

The SQL shows the 2015 launch records by the month. It shows the total number of failure landing in drone ship, booster version names, and launch sites for each month of 2015.

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

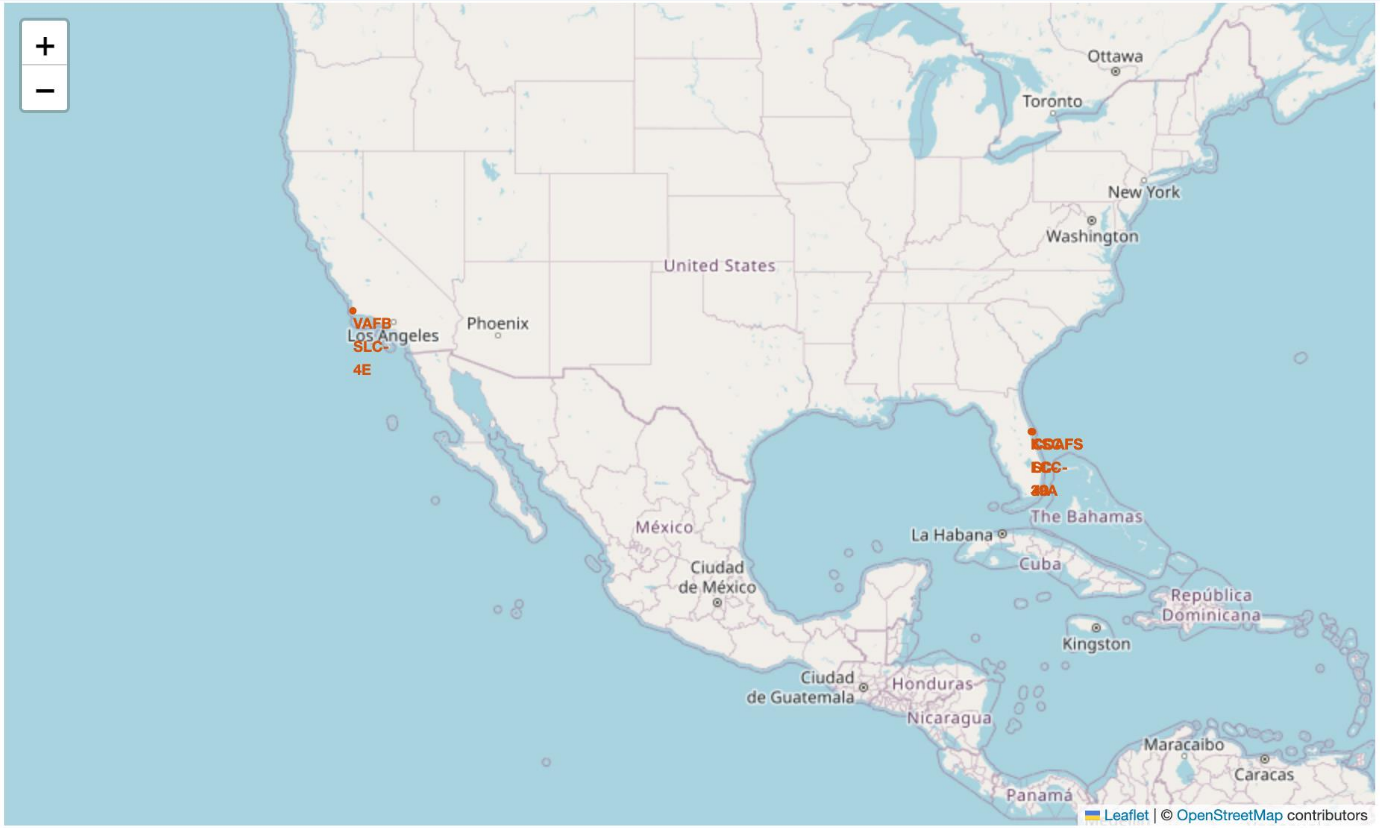
- The SQL query counts the number of landings outcomes between the date 2010-06-04 and 2017-03-20, in descending order
- During this period, the most popular landing outcome is no attempt.

Landing_Outcome	Count()
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

Section 3

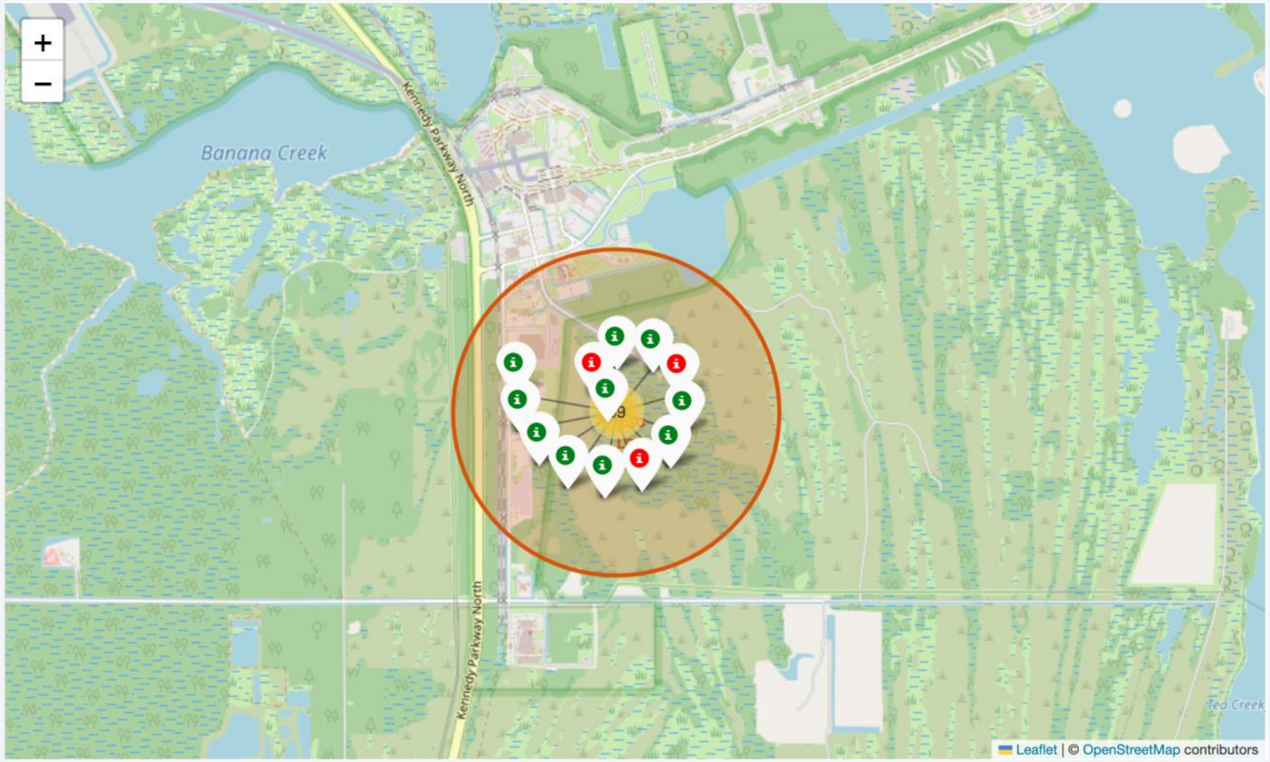
Launch Sites Proximities Analysis

Maps for all launch sites



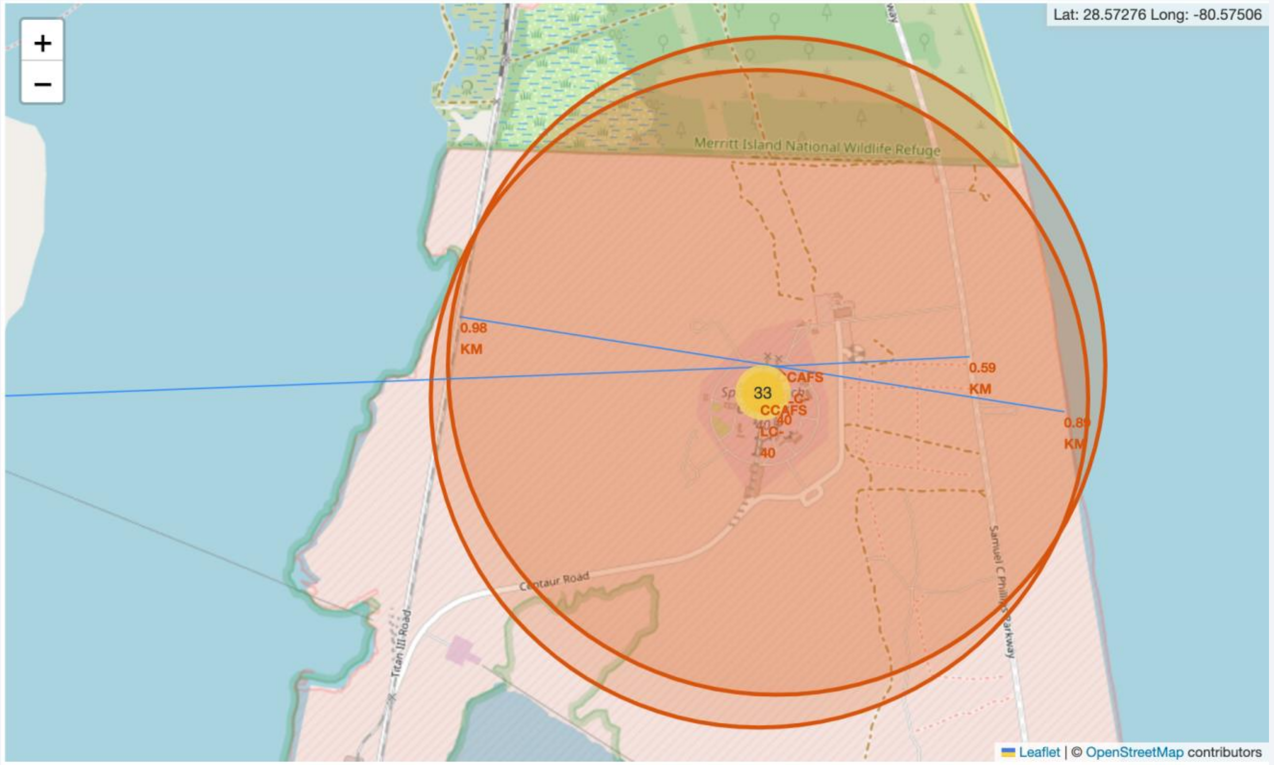
- The circle and name markers show where all the launch sites are located.
- The launch sites are either in Cape Canaveral Space Force Station in Florida or Vandenberg Space Force Base in California.

Landing outcomes at KSC LC-39A



- The circle and marker show the KSC LC-39A launch site. The icons mark the individual flights from the launch site with red encoding failed landing and green encoding successful landing.
- The launch site has a high success landing rate. This interactive map helps viewers explore the success rate of different sites.

CCAFS SLC-40's proximities

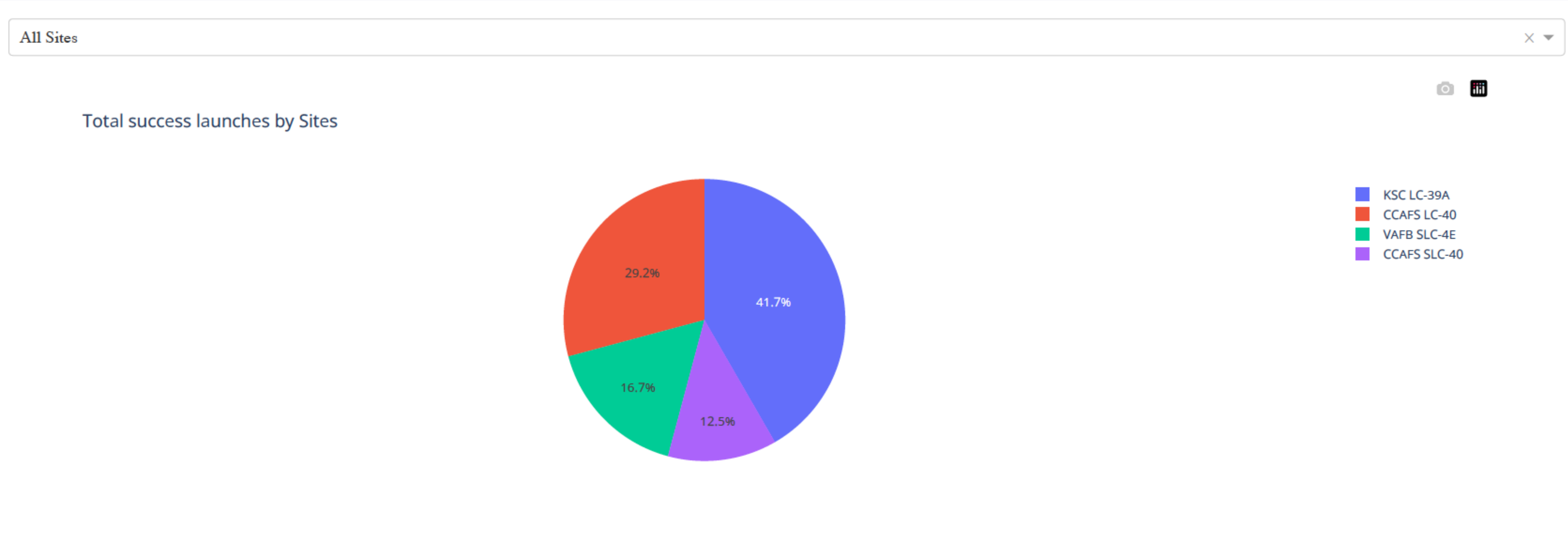


- The map shows the different proximities including coastline, railway, and highway to the CCAFS SLC-40 launch site.
- The blue lines connect the launch site to proximities. The labels show the distances in Km from the site to the proximities.
- The site is close proximities to a coastline, highway, and railway can increase its success rate.

Section 4

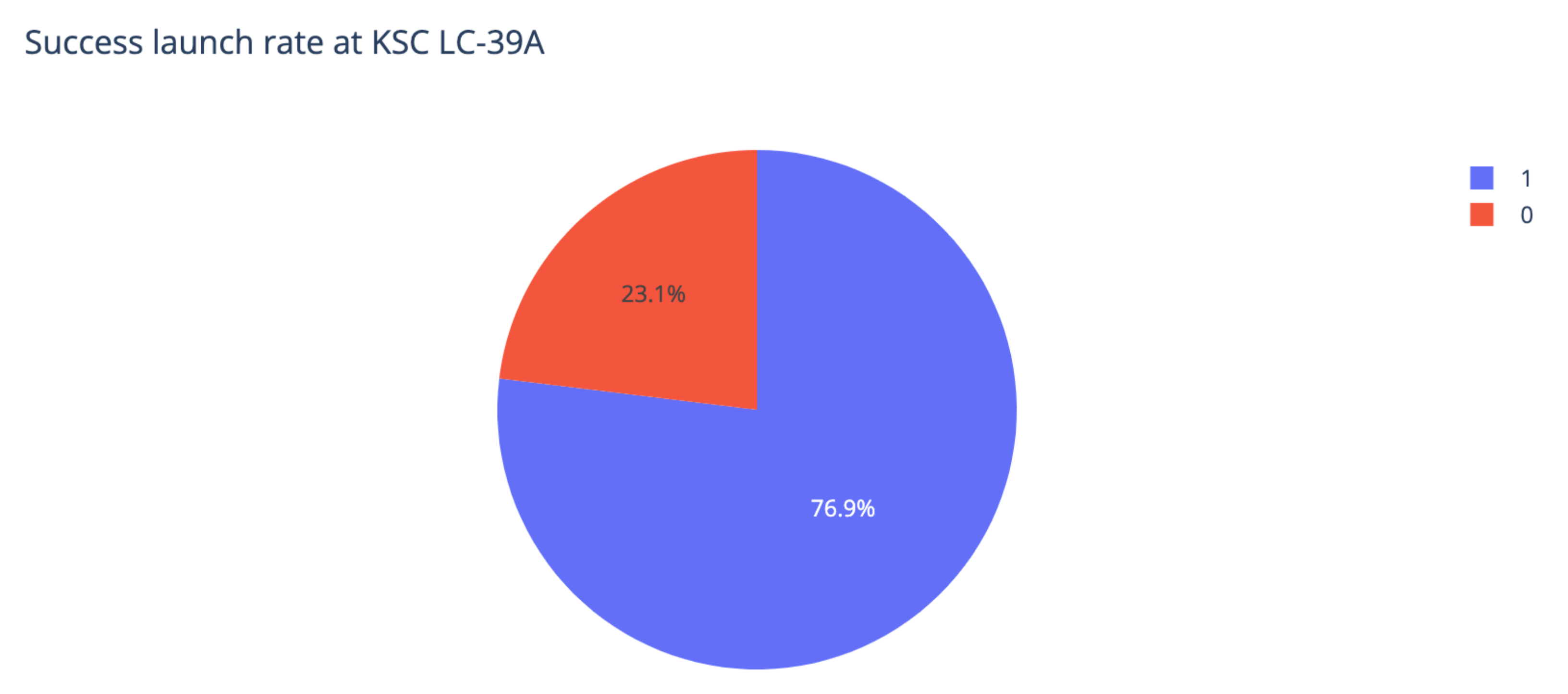
Build a Dashboard with Plotly Dash

Launch success count for all sites



- The launch success count for all launch sites Is shown in the pie chart above.
- KSC LC-39A has the highest number of successes, while CCAFS SLC-40 has the lowest landing successes.

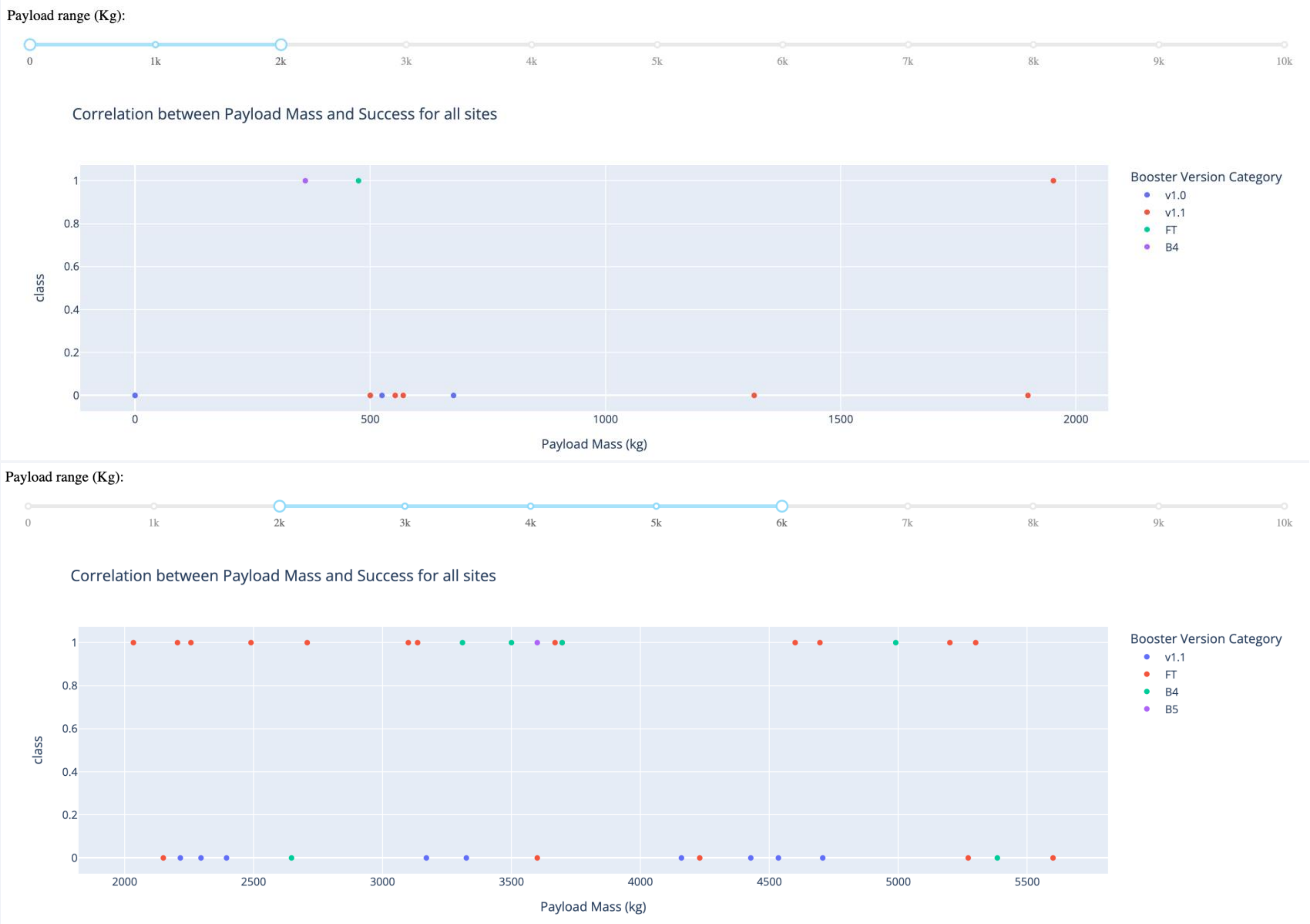
Launch success ratio for KSC LC-39A



- KSC LC-39A site has the highest success ratio among all launch sites. The success rate is 76.9%.
- The pie chart above shows the landing success ratio with 1 encoding successful landing, while 0 encoding failed landing.

Payload mass vs Launch outcome

- The scatter plots show the correlation between payload mass and landing outcome in all launch sites.
- The slider can be used to select the range of payload mass plotted.
- The first plot shows the payload range 0-2000kg. In this payload mass range, there is a low success rate.
- The second plot shows the payload mass range 2000-6000kg. In this range, there is a higher success rate.

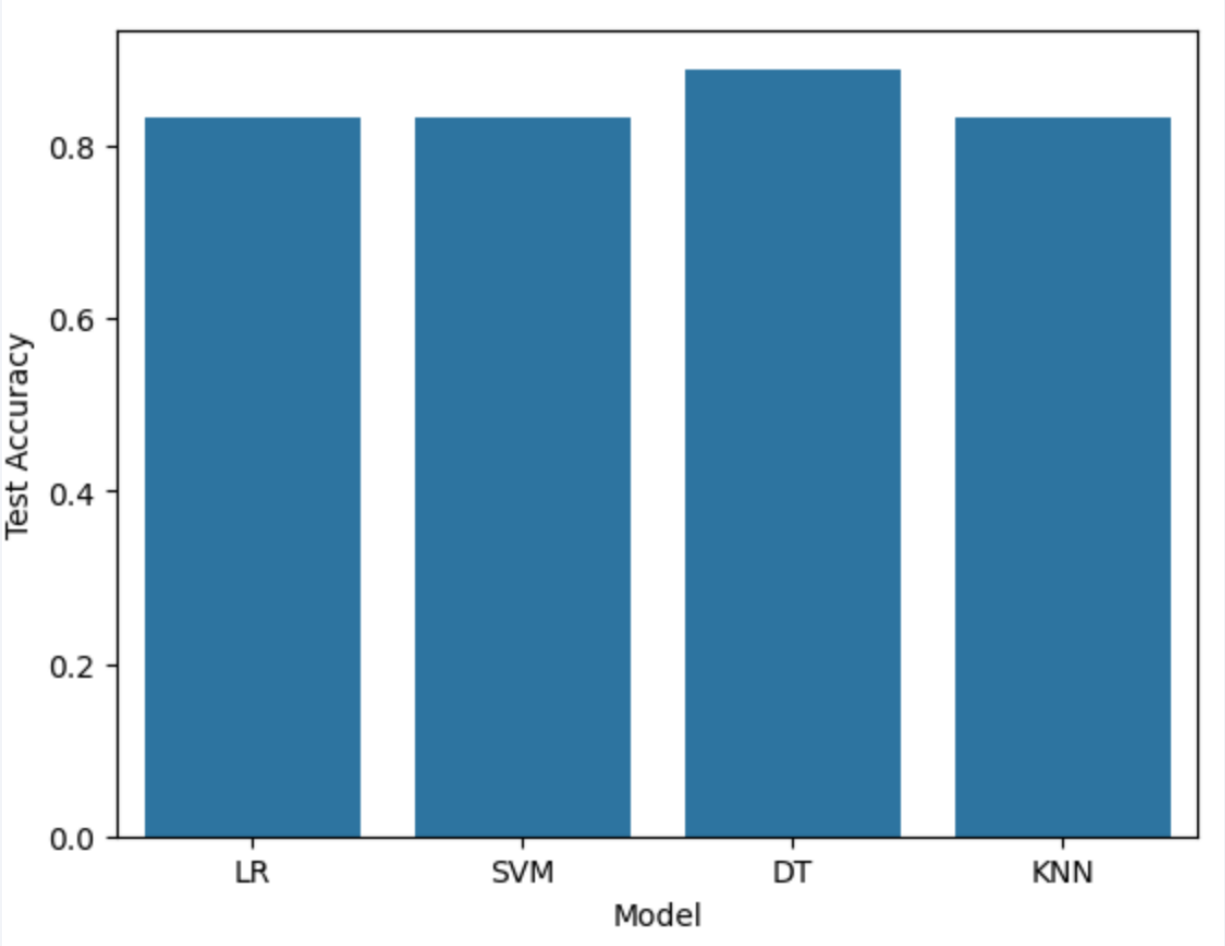


Section 5

Predictive Analysis (Classification)

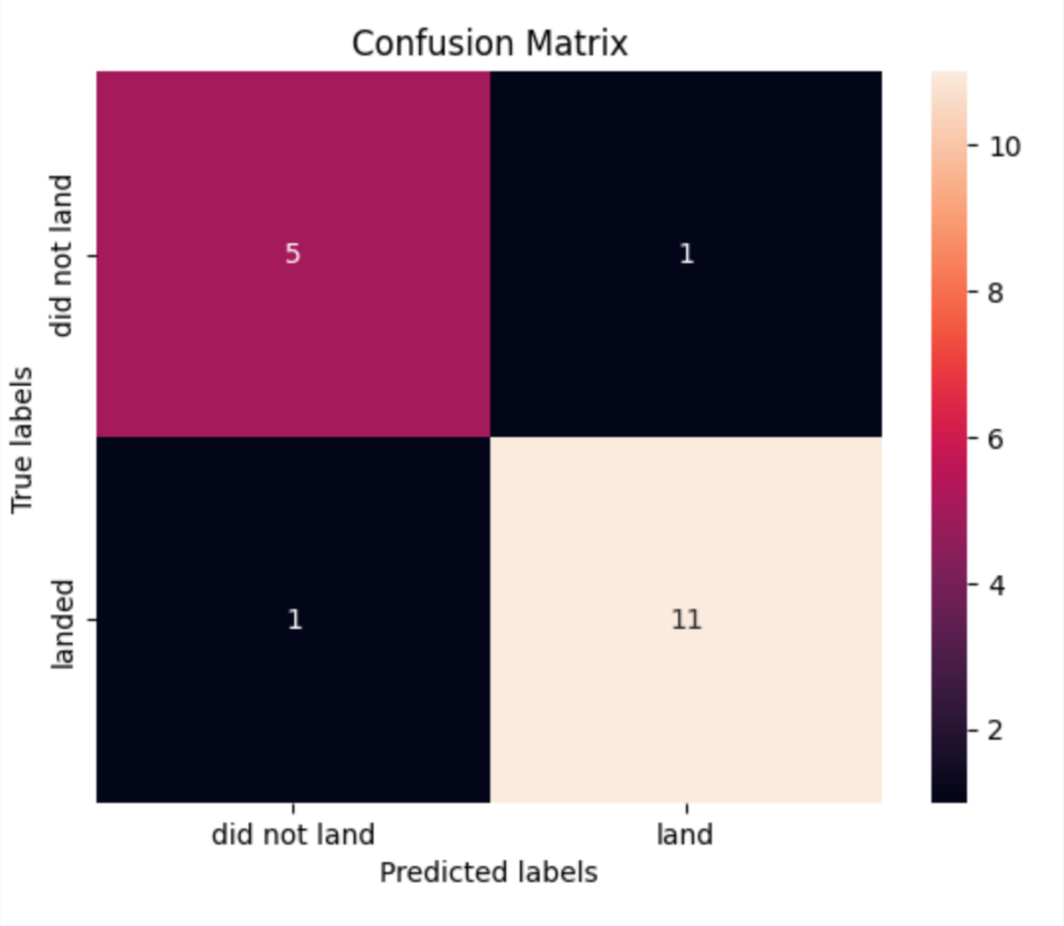
Classification Accuracy

- The bar chart shows the test accuracy score of the different models (LR: logistic regression, SVM: support vector machine, DT: decision tree, KNN: k-nearest neighbor).
- Decision tree has the highest test accuracy among the models. This suggests this is the best-performing model for this data set.



Confusion Matrix

- This is the confusion matrix of the decision tree model.
- The model misclassified 2 flight outcomes. The model made 1 type 1 error and 1 type 2 error.
- This suggests that the model is high performing with high accuracy.



Conclusions

- Among other features, Payload mass, orbit, flight number, launch year, and launch site are important features that affect the landing outcome.
- The launch sites' locations and proximities are also good indicators of the launch site's success rate.
- Among the machine learning models, Decision tree is the best-performing model to predict the launch outcome for this dataset.
- The trained and optimized decision tree model achieved a high test accuracy score of 0.89.
- This project has identified important factors to launch outcomes and built a model to predict launch success with high accuracy.

Thank you!