IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Zhen Hao Chong
11/22/2024

Outline

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

Executive Summary

- The primary objective of this capstone is to build a classification model to predict if Falcon 9 rocket will land successfully. The model accuracy is crucial to determine the cost of a launch.
- Historical launch data of Falcon 9 are collected via SpaceX REST API and web-scraping a Wikipedia page with Falcon 9 launch records. The data collected is processed, cleaned and transformed for further analysis.
- Exploratory data analysis is performed by SQL queries and data visualization. A Folium map is created to uncover more insights, and an interactive dashboard is built to visualize key success rate at different sites and payload mass.
- Then, 4 classification models - Logistic Regression, SVM, Decision Tree and KNN are developed, tuned and evaluated by their accuracy score to determine the model that best predicts the outcomes.
- Decision Tree achieved the highest test accuracy at 0.944, outperforming the other 3 models.

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- In this capstone, we will predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

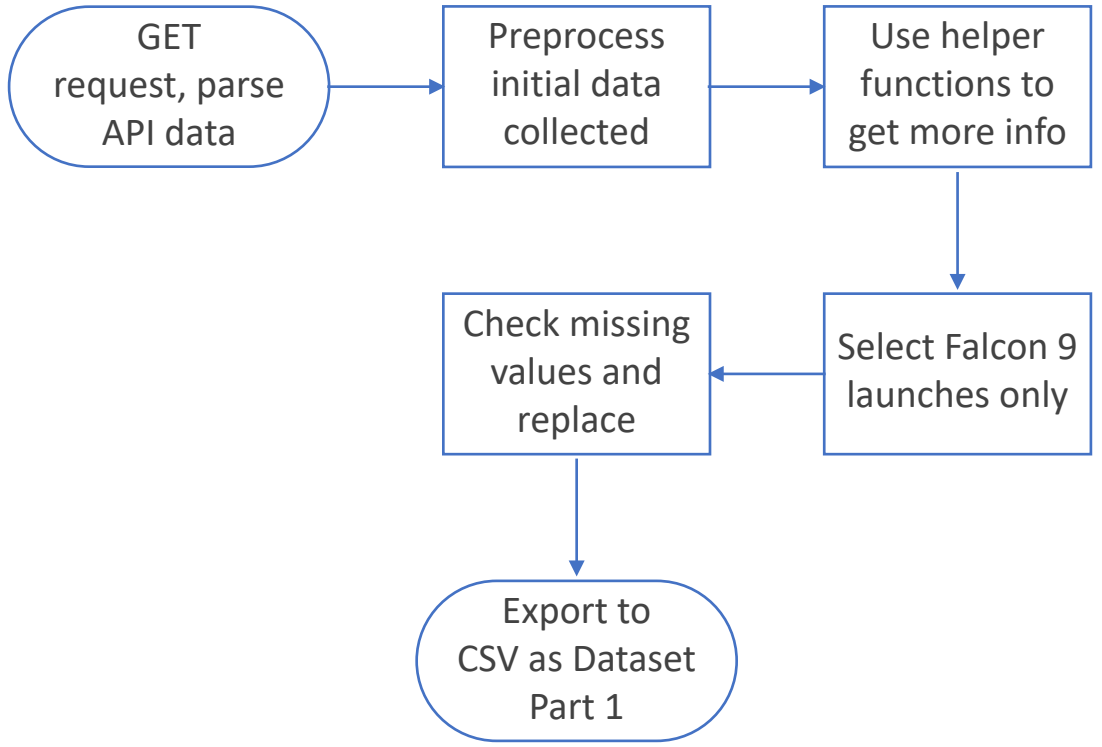
Executive Summary

- Data collection methodology
  - Collect data from SpaceX REST API and web-scraping a Wikipedia page
- Perform data wrangling
  - Check missing values and replace with suitable values
  - Assign 0 (failure) or 1 (success) to all landing outcomes
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Find patterns in the data and select features for classification models
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, tune and evaluate classification models

Data Collection – SpaceX API

- Use `requests.get` to get SpaceX API data and convert the json result into a DataFrame using `pd.json_normalize`.
- Preprocess the data loaded, keeping only required features, then use helper functions to extract more meaningful data from the API.
- Helper functions:
  - getBoosterVersion
  - getLaunchSite
  - getPayloadData
  - getCoreData
- GitHub URL:
  Data Collection - SpaceX API

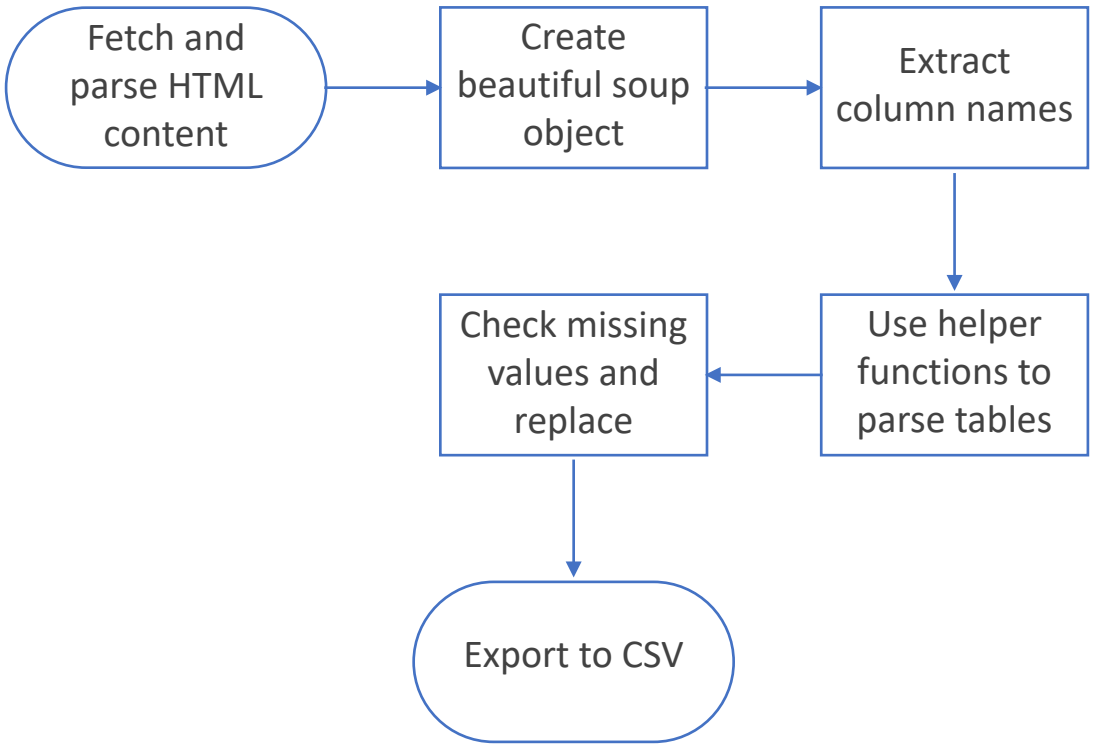
Flowchart of SpaceX API calls:



Data Collection - Scraping

- After fetching and parsing HTML content, create a beautiful soap object to extract the data that we need from the webpage.
- Use helper function to extract column names and parse the tables in Wiki page.
- Helper functions:
  - extract_column_from_hearder
  - date_time
  - booster_version
  -landing_status
  - get_mass
- GitHub URL:
  Data Collection - Scraping

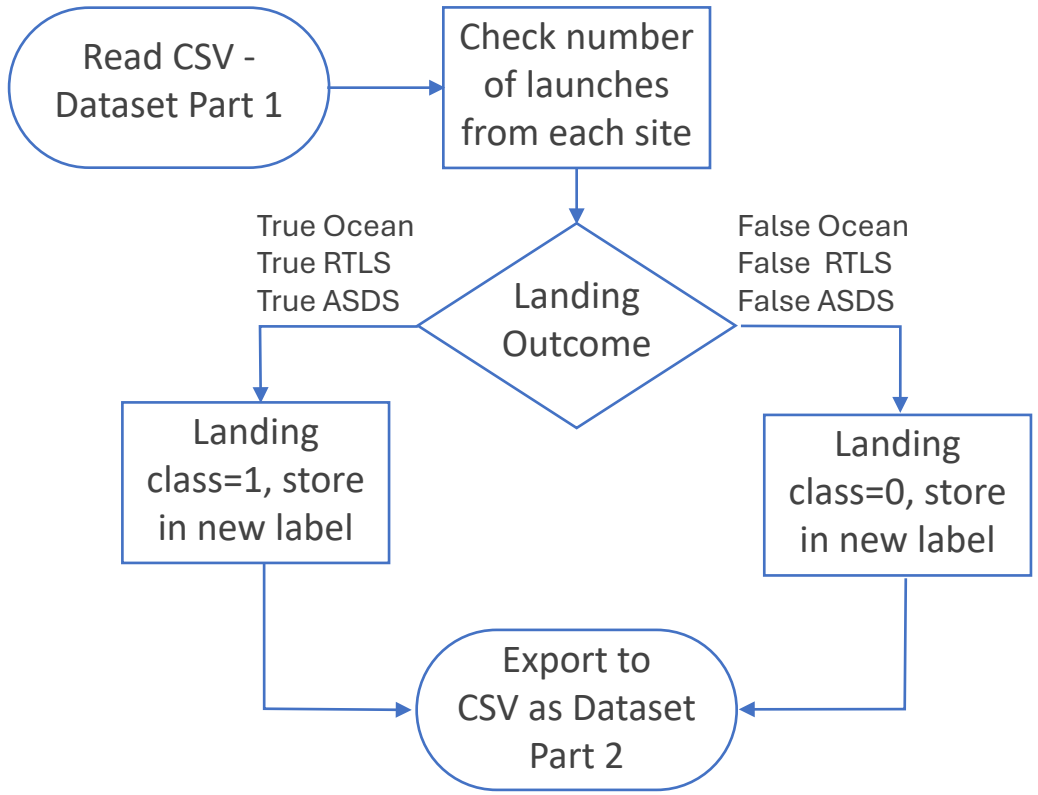
Flowchart of web scraping:



Data Wrangling

- Successful outcomes:
  True Ocean, True RTLS, True ASDS
- Unsuccessful outcomes:
  False Ocean, False RTLS, False ASDS
- Convert all landing outcomes into 1 for successful cases or 0 for unsuccessful cases, then create a new column - class.
- GitHub URL: Data Wrangling

Flowchart of data wrangling:



EDA with Data Visualization

- Perform EDA with Pandas, Matplotlib and Seaborn charts to explore correlation among different variables and find patterns in the data.
- Scatter plots visualize the landing outcome under different scenarios:
  Payload Mass vs Flight Number, Launch Site vs Flight Number, Launch Site vs Payload Mass, Orbit vs Flight Number, Orbit vs Payload Mass.
- Bar plot visualizes the relationship between success rate for each orbit type.
- Line plot visualizes the launch success yearly trend.
- Finally, select features for model building and perform one-hot encoding on categorical columns.
- GitHub URL: EDA with Data Visualization

EDA with SQL

- Perform EDA with SQL queries to understand more about the dataset:
    - Total Payload Mass
    - Average Payload Mass by F9 v1.1
    - First Successful Ground Landing Date
    - Successful Drone Ship Landing with Payload between 4000 and 6000
    - Total Number of Successful and Failure Mission Outcomes
    - Boosters Carried Maximum Payload
    - 2015 Launch Records
    - Rank Landing Outcomes Between 2010-06-04 and 2017-03-20
- GitHub URL: EDA with SQL

Build an Interactive Map with Folium

- Create a Folium map with text labels for all launch sites.
- Add map objects such as markers, circles, marker cluster and lines to the Folium map to visualize the landing outcomes at different launch sites.
- GitHub URL: Interactive Map - Folium

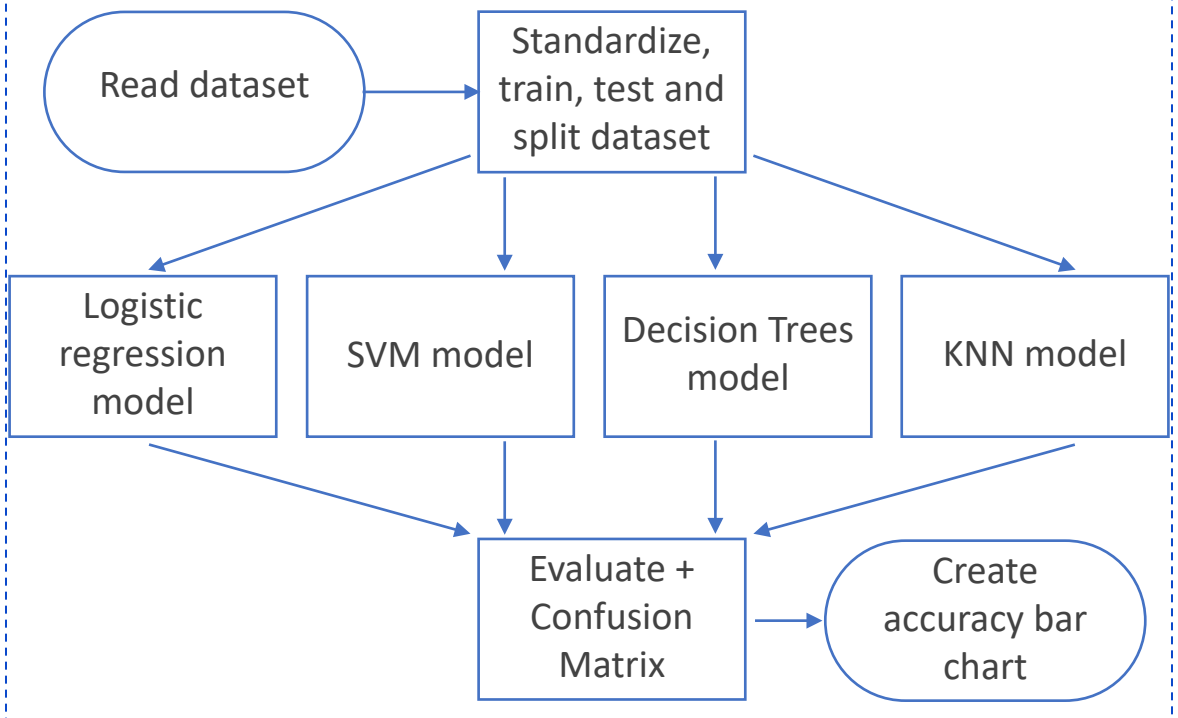
Build a Dashboard with Plotly Dash

- Create a SpaceX Launch Records Dashboard with 2 interactive charts:
    1. Pie chart showing total successful launches for each sites by selection of site in dropdown options.
    2. Scatter plot showing correlation between payload mass and success for all sites by selection of payload range in slider bar.
- GitHub URL: Interactive Dashboard - Plotly Dash

Predictive Analysis (Classification)

- Standardize / normalize the dataset, split the data into training and testing sets, then use scikit-learn library to build model for Logistic Regression, SVM, Decision Trees and KNN.
- Evaluate model and test accuracy, analyze Confusion Matrix, select best model for prediction.
- GitHub URL: Predictive Analysis - Classification

Flowchart of classification modeling:



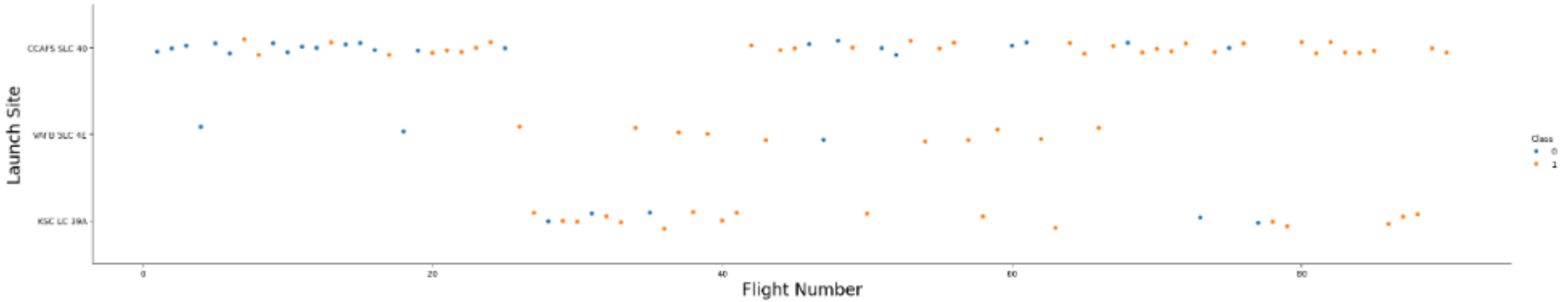
Result

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

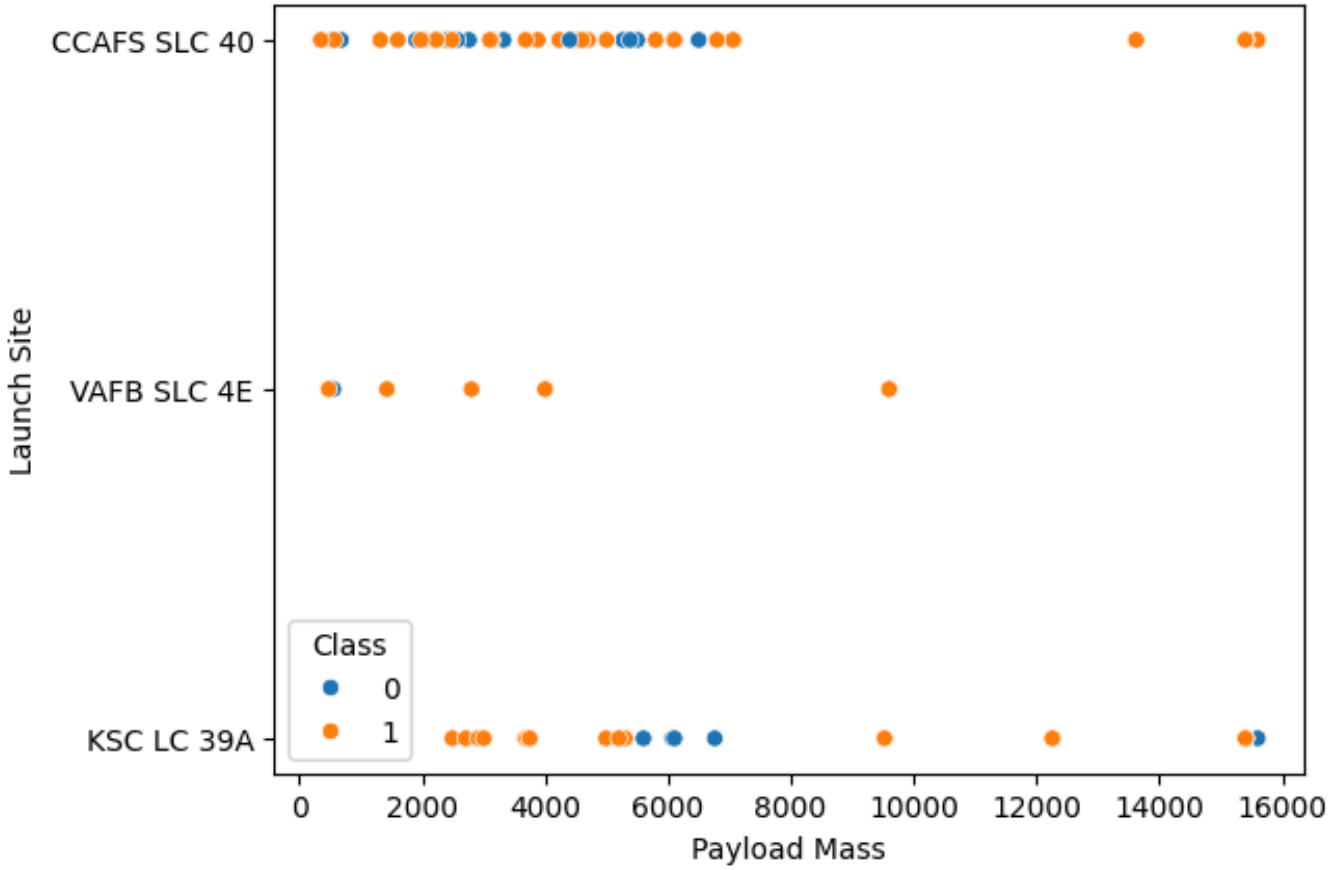
Insights drawn from EDA

Flight Number vs. Launch Site



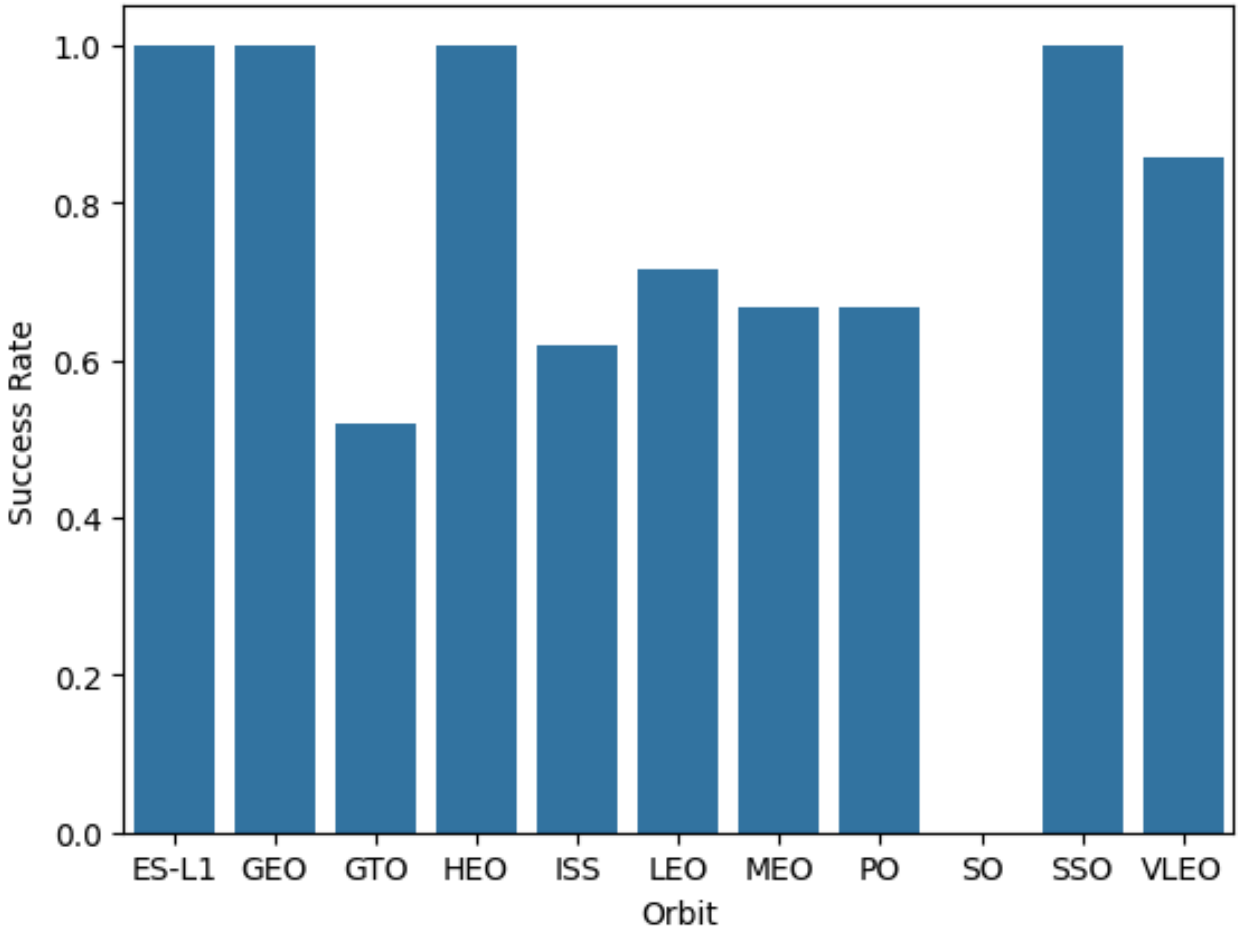
- As the flight number increases, first stage landing success rate increases at each site.

Payload vs. Launch Site



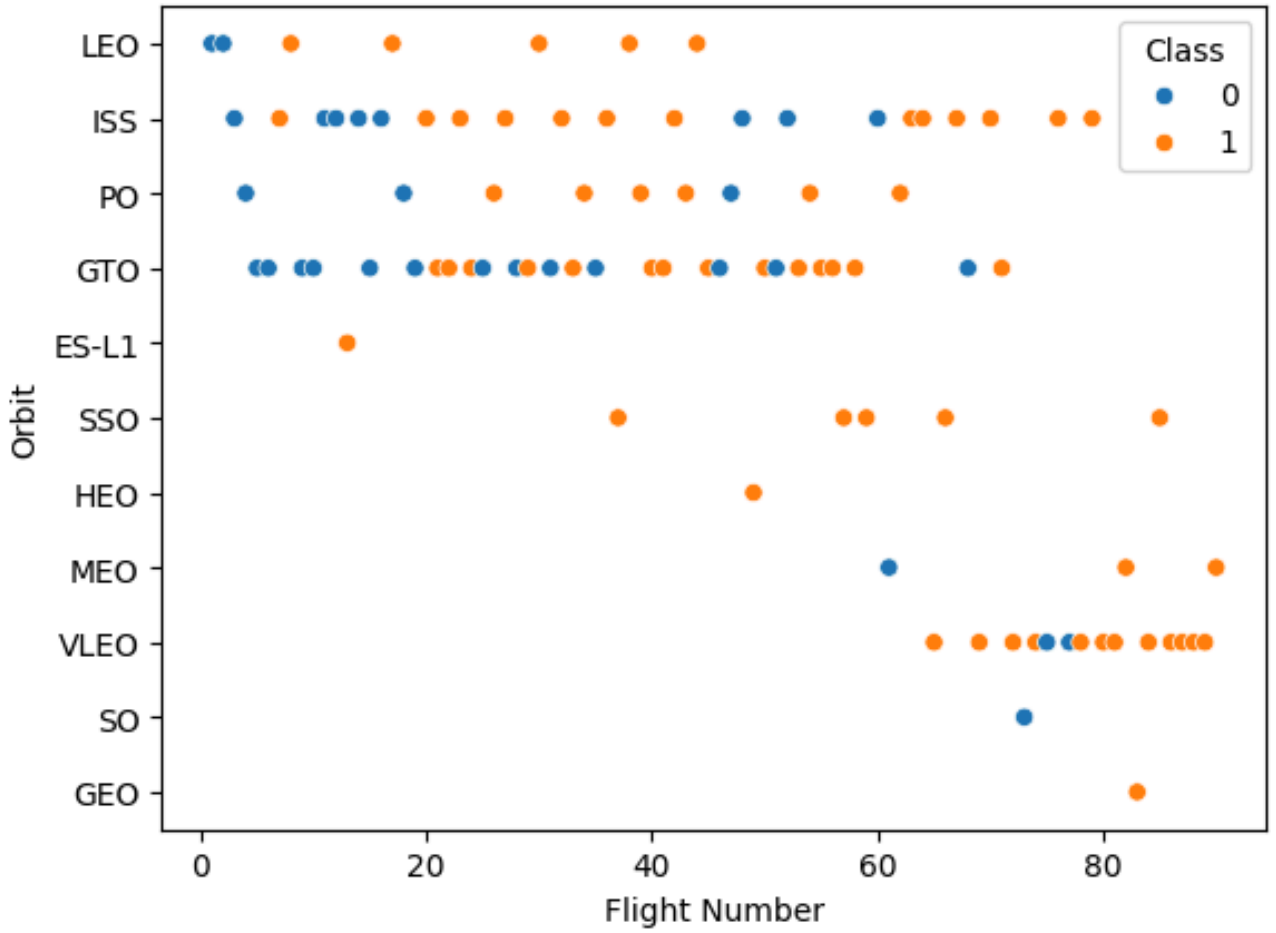
- CCAFS SLC 40: For >10000kg payload, CCAFS SLC 40 has 100% success rate, mixed outcomes for payload <10000kg.
- VAFB SLC 4E: No rockets launched for heavy payload mass (greater than 10000kg), high success rate below 10000kg.
- KSC LC 39A: High success rate for high and low payload, except 5000 - 7000 kg range.

Success Rate vs. Orbit Type



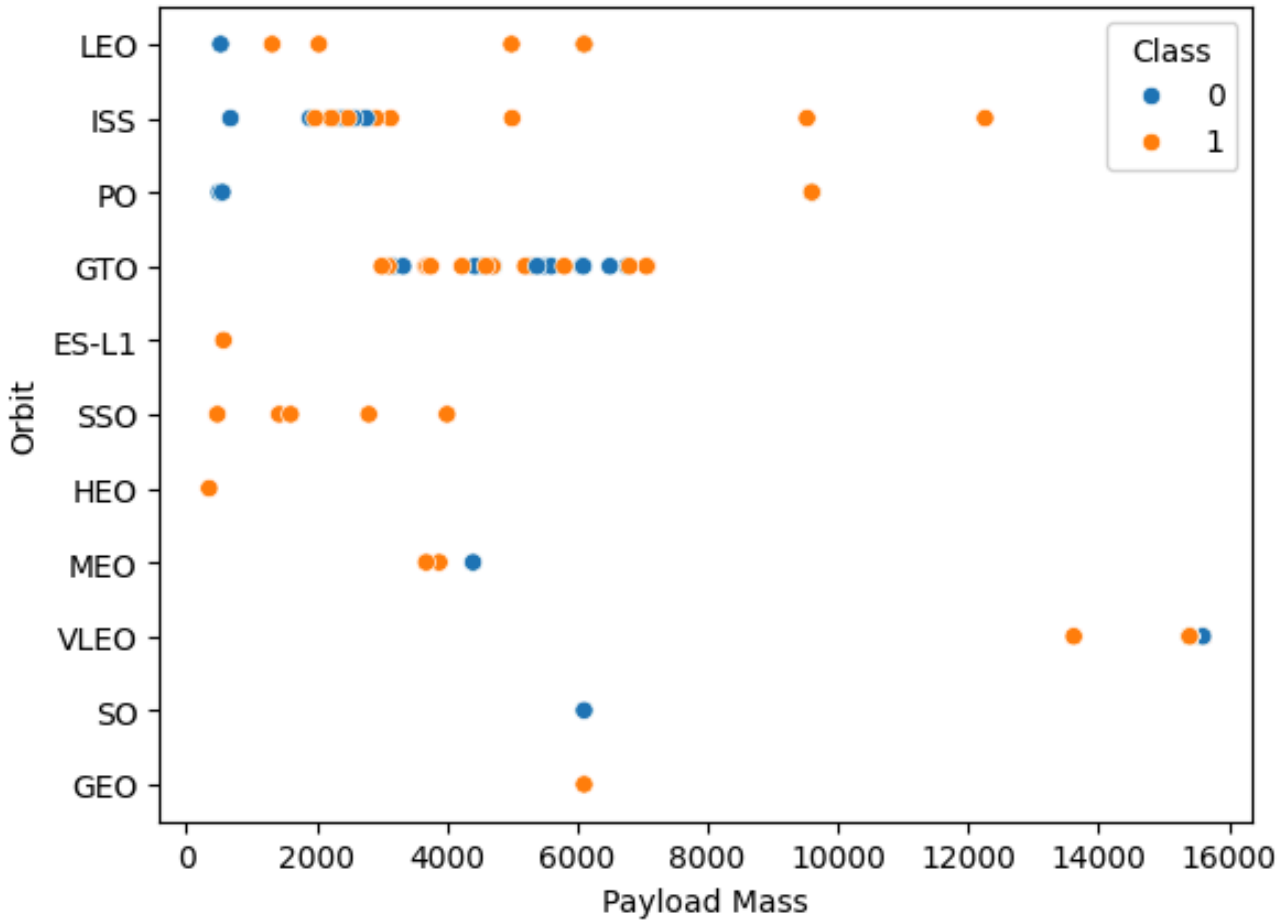
- Without considering number of occurrence, Orbit Types: ES-L1, GEO, HEO, SSO have highest success rate, while GTO lowest.
- ISS, LEO, MEO and PO all have >50% success rate.

Flight Number vs. Orbit Type



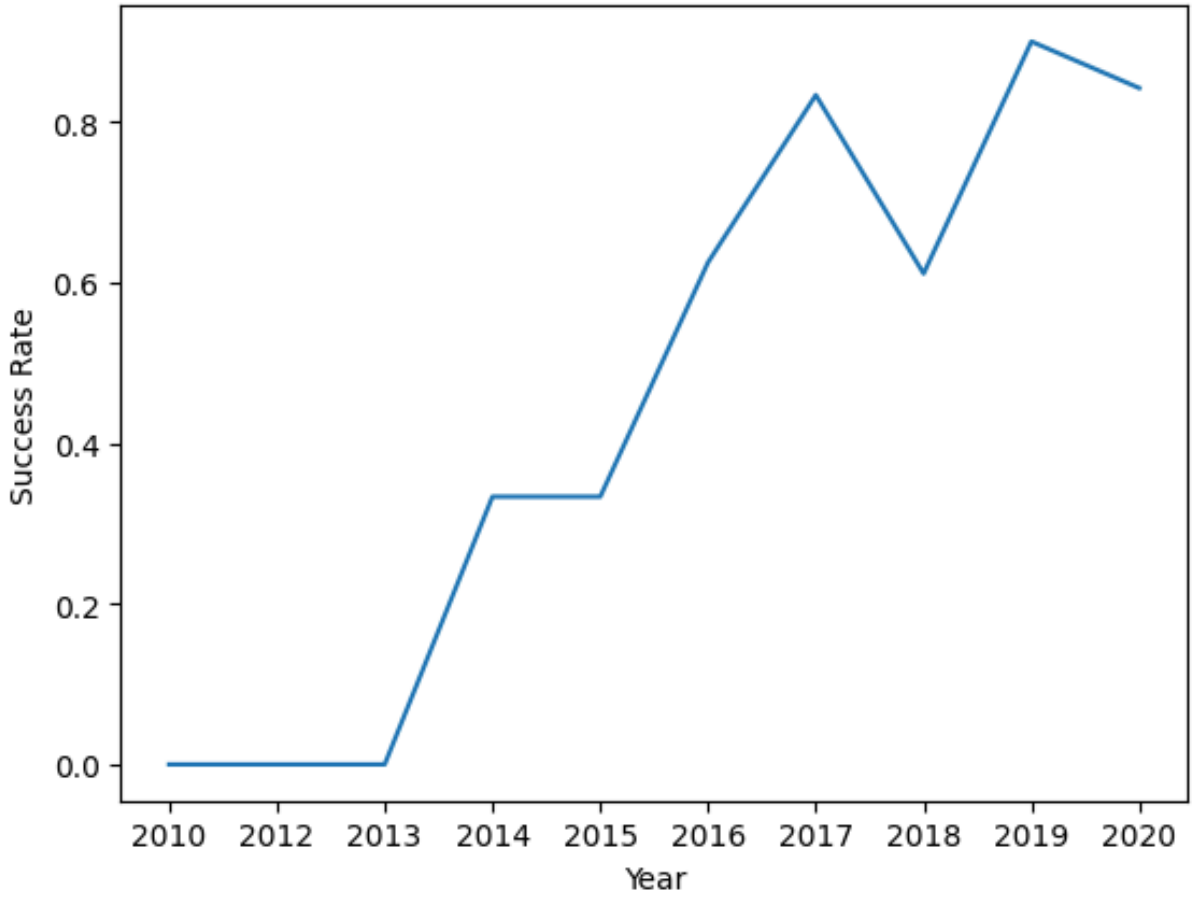
- Generally more success cases as flight number increases.
- GTO has the lowest success rate, improved with higher flight numbers.
- SSO has 100% success rate, VLEO gives quite high success rate at high flight number.
- Other orbit types – LEO, ISS, PO worth mentioned too for their success rate.

Payload vs. Orbit Type



- GTO: 3000-7000kg payload range, mixed outcomes.
- SSO has 100% success rate below 5000kg payload, VLEO and ISS give high success rate at high payload.

Launch Success Yearly Trend



- Since 2013, the success rate consistently increases. A little bit of plunge from 2017 - 2018, but has picked up since.

All Launch Site Names

- Unique launch sites:

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

- There are a total of 4 unique launch sites, namely CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40.

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with ‘CCA’

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Total payload carried by boosters from NASA = 45596kg

Customer	SUM(PAYLOAD_MASS__KG_)
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 = 2928.4kg

Booster_Version	AVG(PAYLOAD_MASS__KG_)
F9 v1.1	2928.4

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad: 2015-12-22

Landing_Outcome	MIN(Date)
Success (ground pad)	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000: F9 FT B1020

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1020	5271

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes:

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

- 100 successful outcomes, 1 failure outcome.

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015:

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

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

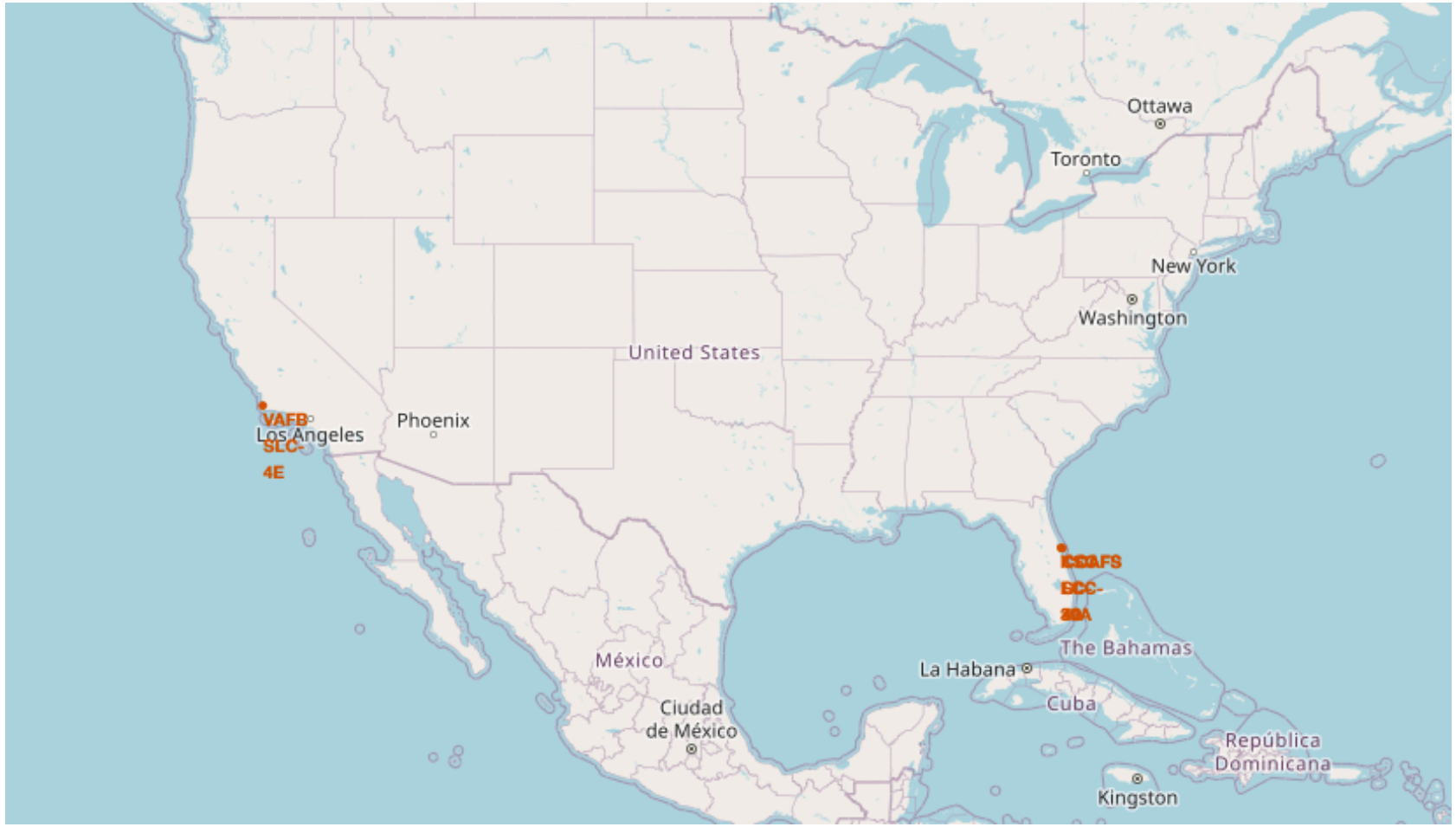
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order:

Date	Landing_Outcome	COUNT(Landing_Outcome)
2017-03-16	No attempt	1
2017-02-19	Success (ground pad)	1
2017-01-14	Success (drone ship)	1
2016-08-14	Success (drone ship)	1
2016-07-18	Success (ground pad)	1
2016-06-15	Failure (drone ship)	1
2016-05-27	Success (drone ship)	1
2016-05-06	Success (drone ship)	1
2016-04-08	Success (drone ship)	1
2016-03-04	Failure (drone ship)	1
2016-01-17	Failure (drone ship)	1
2015-12-22	Success (ground pad)	1
2015-06-28	Precluded (drone ship)	1
2015-04-27	No attempt	1
2015-04-14	Failure (drone ship)	1
2015-03-02	No attempt	1
2015-02-11	Controlled (ocean)	1
2015-01-10	Failure (drone ship)	1
2014-09-21	Uncontrolled (ocean)	1
2014-09-07	No attempt	1
2014-08-05	No attempt	1

Section 3

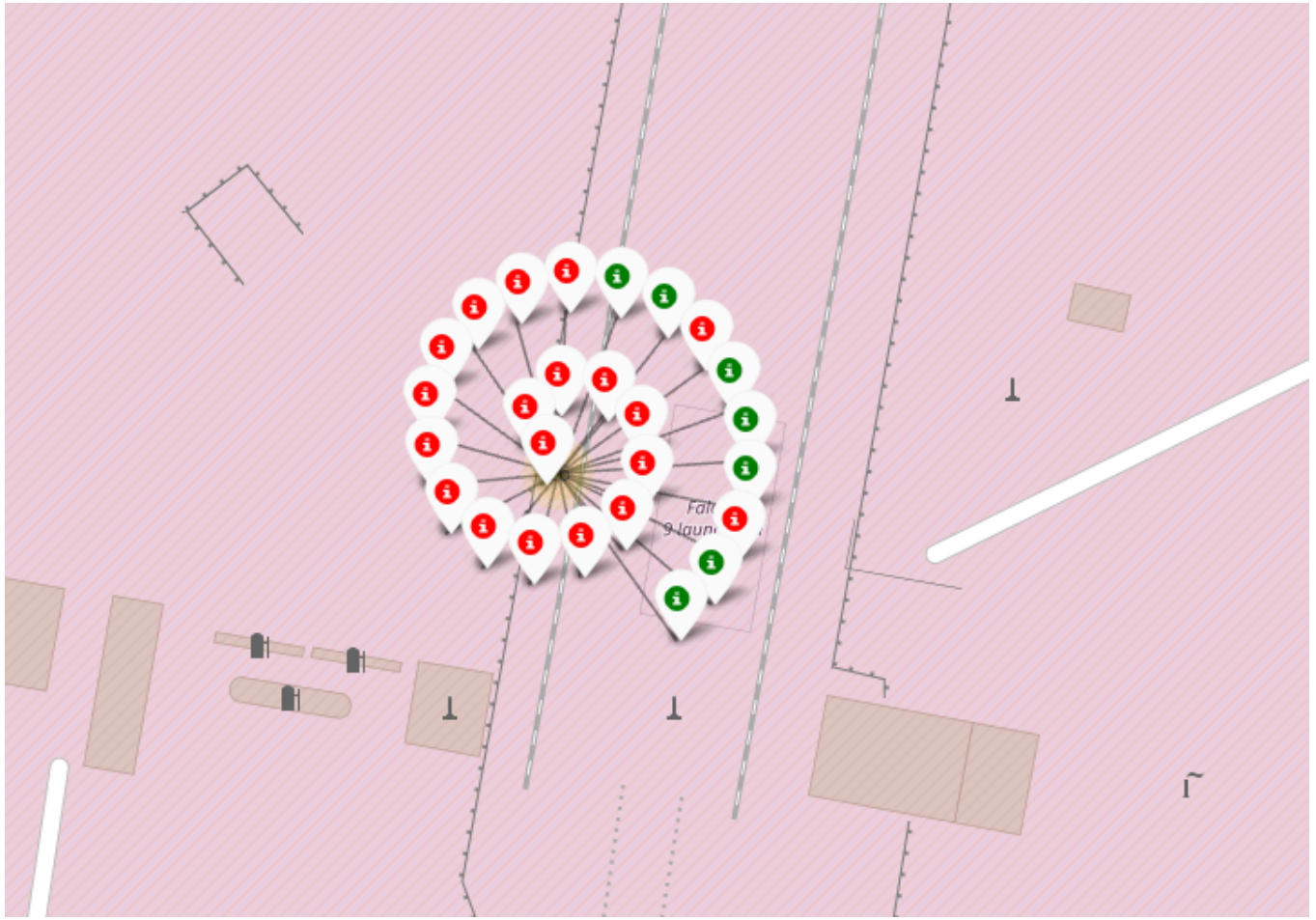
Launch Sites Proximities Analysis

Launch Sites Location



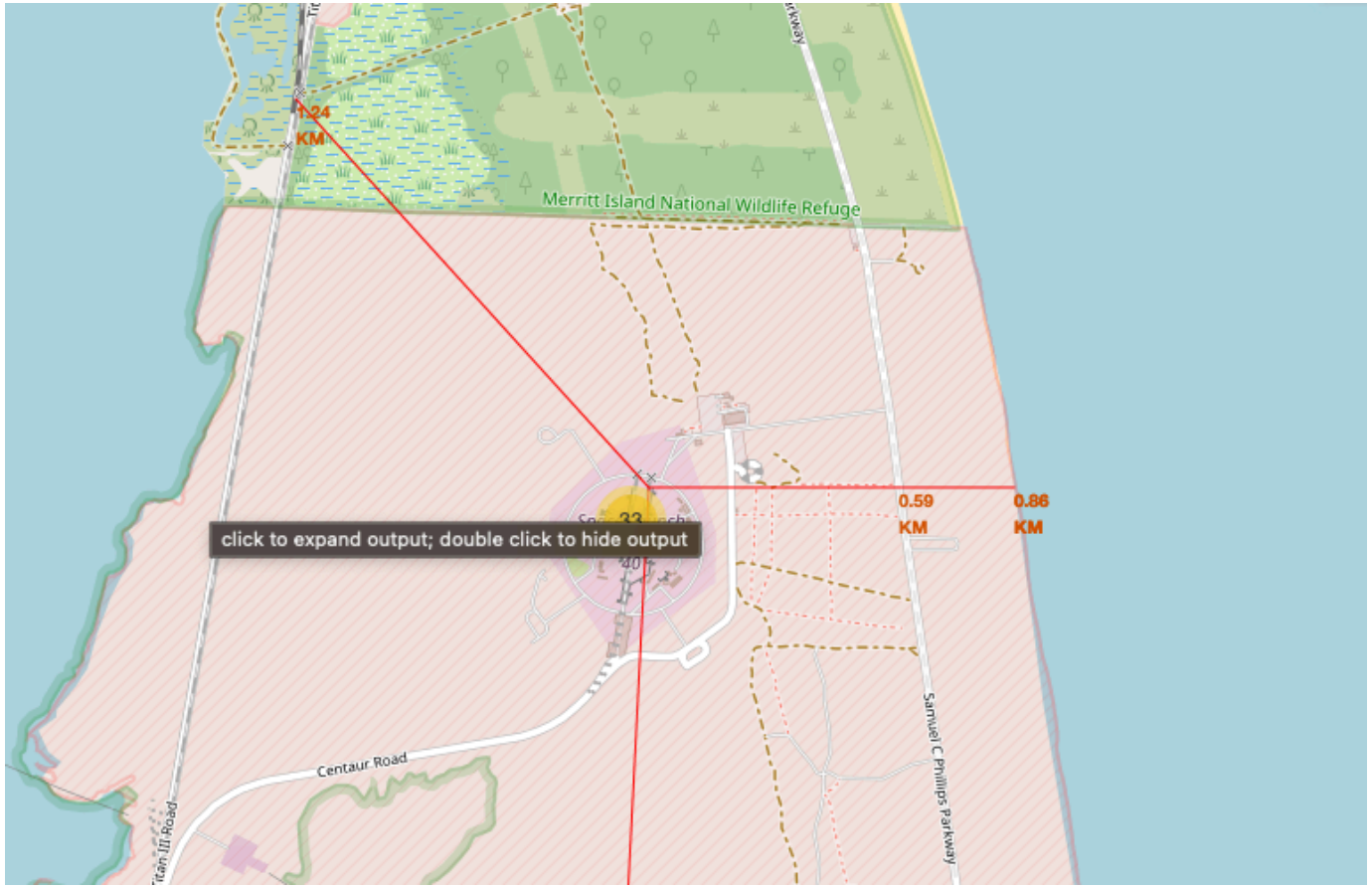
- 1 Launch Site in California, another 3 in Florida.

Launch Outcome at a Site



- Launch Site: CCAFA LC-40 has more unsuccessful launches than successful.

Launch Site to it Proximities

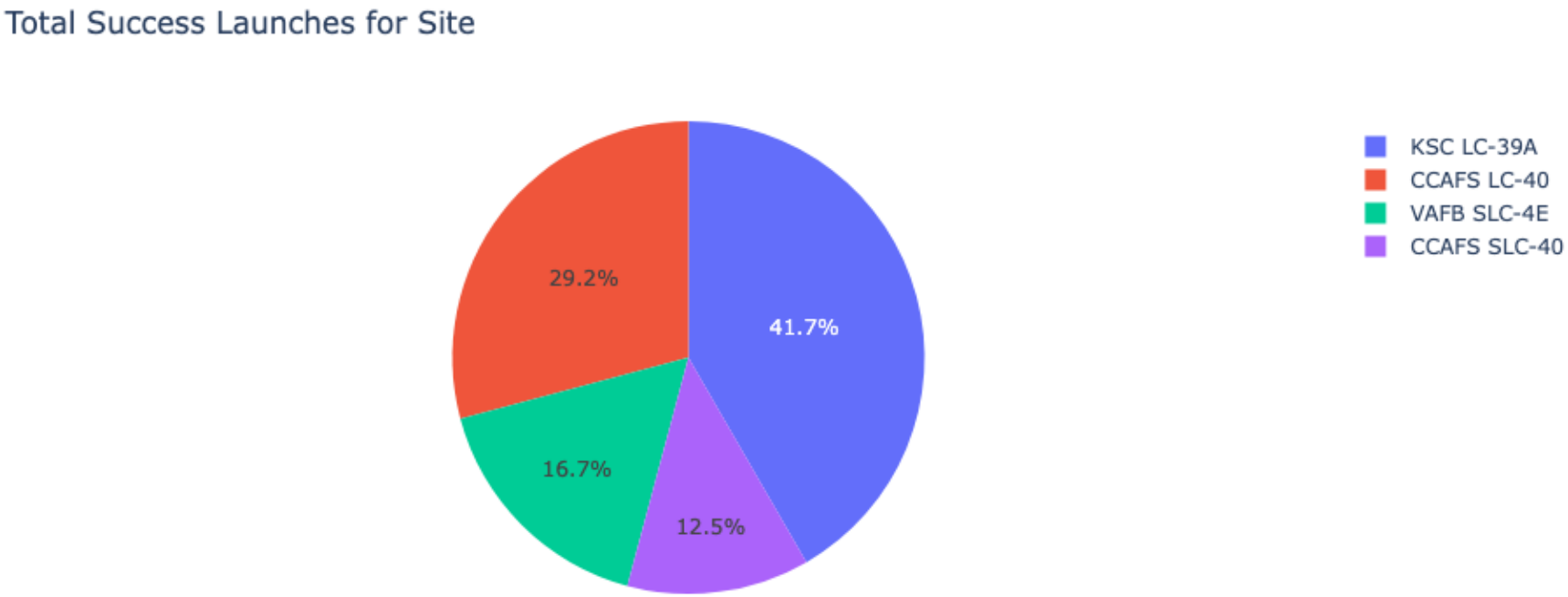


- Proximities such as railway, highway and coastline do not seem to have obvious impact on launch successes.

Section 4

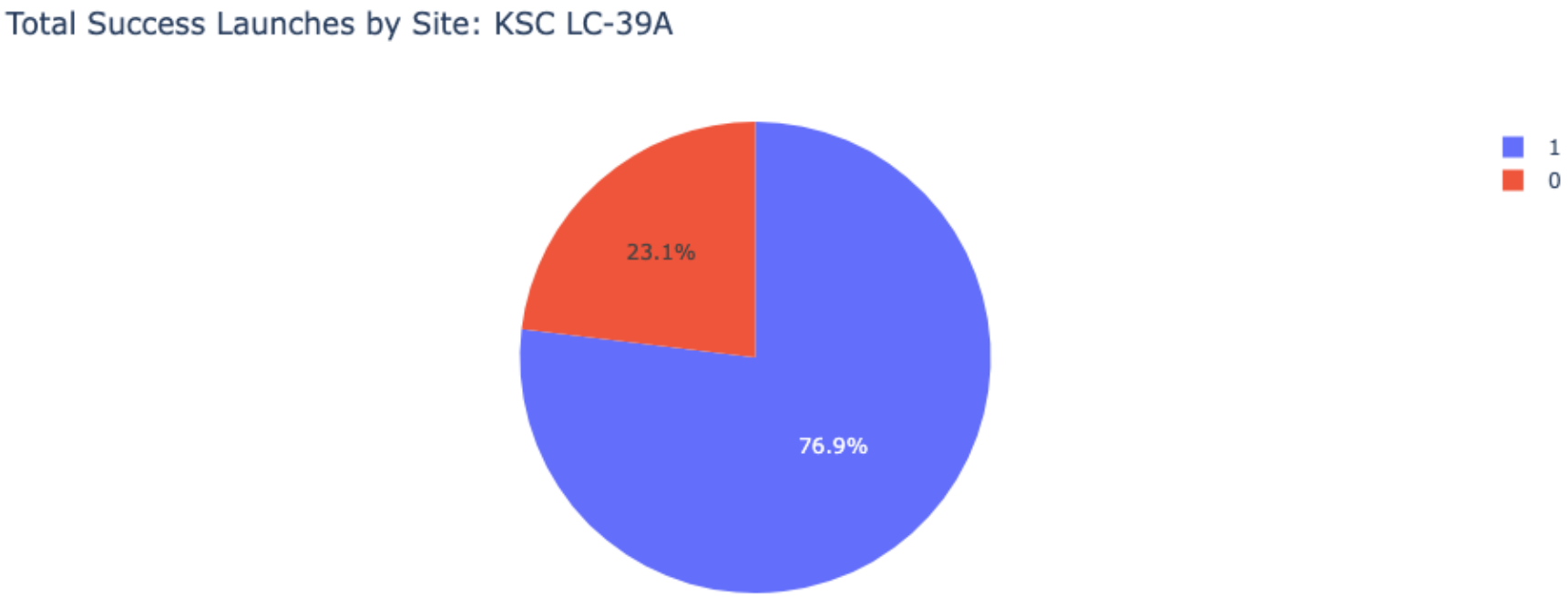
Build a Dashboard with Plotly Dash

Total Launch Success for All Sites



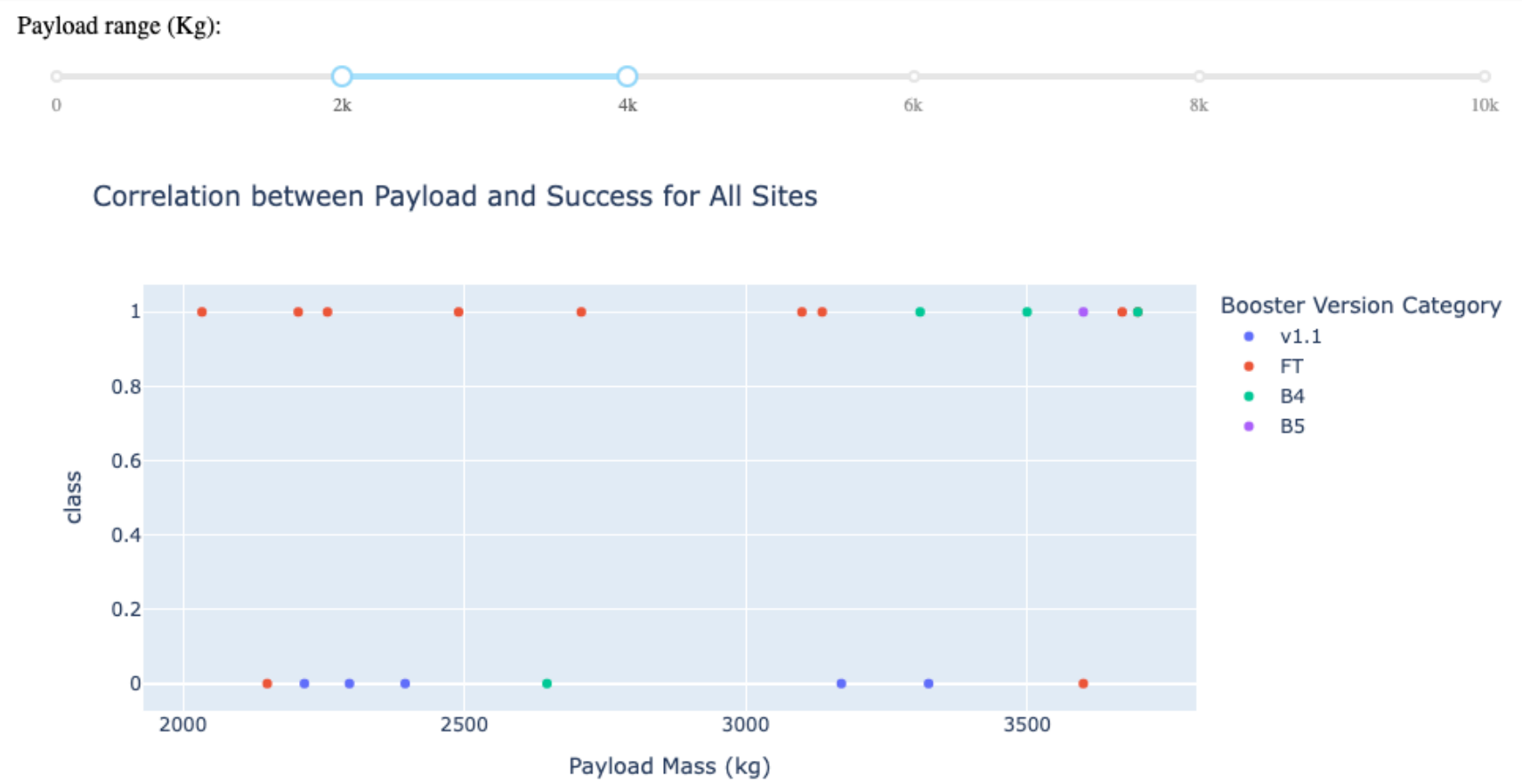
- Launch Site: KSC LC-39A has the highest launch success, followed by CCAFS LC-40, VAFB SLC-4E and CCAFS SLC-40.

Launch Site with Highest Launch Success



- Launch Site: KSC LC-39A has a launch success rate of 76.9%.

Payload vs Launch Outcome for all Sites

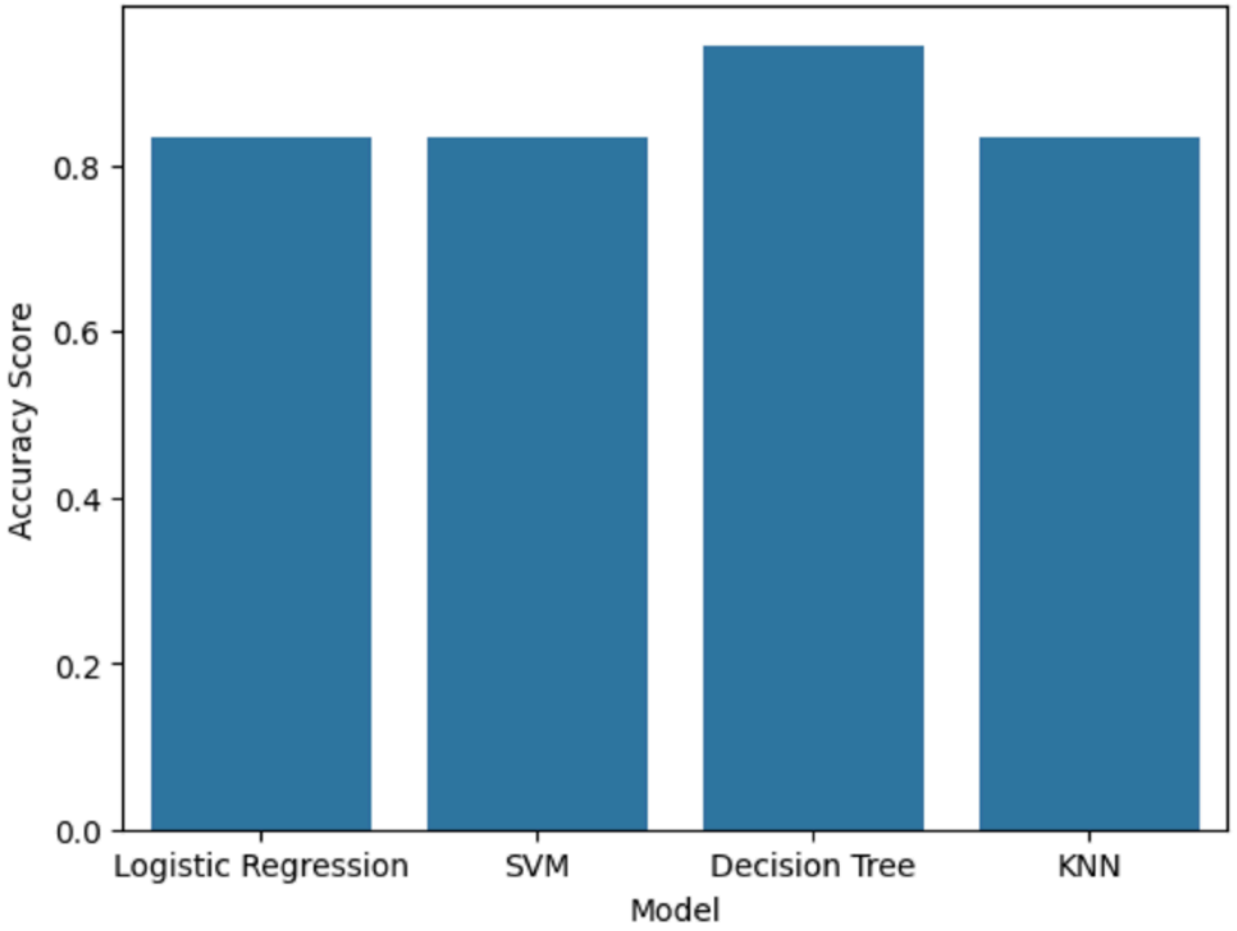


- Payload Range: 2000 - 4000kg and Booster Version: FT have the largest success rate.

Section 5

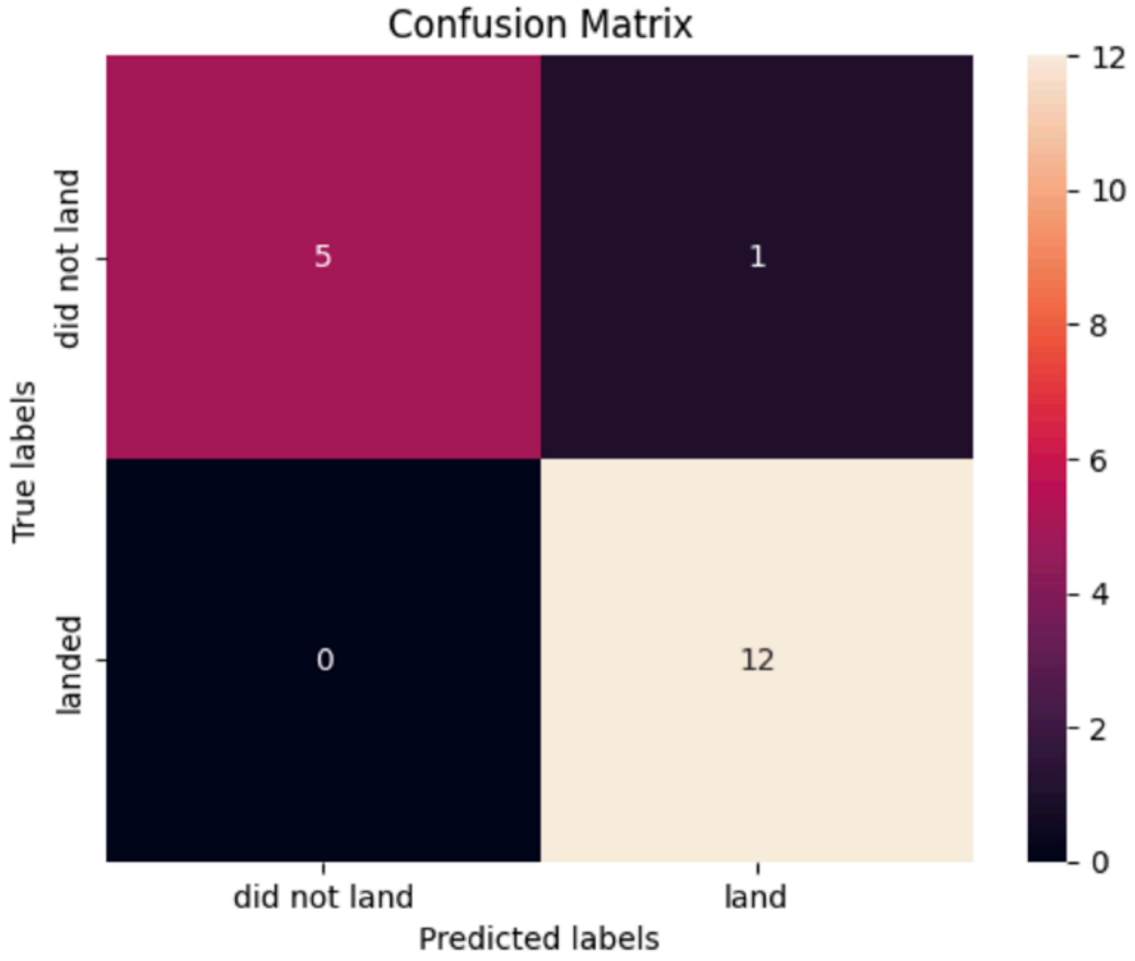
Predictive Analysis (Classification)

Classification Accuracy



- Decision Tree has the highest accuracy score among all models tested.
  Accuracy of validation data: 0.875
  Accuracy of test data: 0.944

Confusion Matrix



Decision Tree model:

- Out of 6 unsuccessful landing outcomes, the model able to predict 5 of them, miss out 1.
- For the total of 12 successful landing outcomes, the model predicts 13, wrongly label 1 outcome only.
- In this case, Decision Tree model gives the best performance compared to the other models.

Conclusions

- Among the 4 classification models developed and tested- Logistic Regression, SVM, Decision Tree and KNN, the Decision Tree model achieved the highest test accuracy at 0.944, outperforming the other models. Hence, the Decision Tree model is selected to predict the landing success of future launches.
- By accurately predicting landing outcomes, we can better manage and optimize future launches, paving the way for more sustainable and economical space exploration.

Thank you!