IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Santiago Bermeo
15/6/23

Outline

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

Executive Summary

- Summary of methodologies
  - Data was collected through web scraping and utilizing the SpaceX API.
  - Exploratory Data Analysis (EDA) was conducted, including data wrangling, data visualization, and interactive visual analytics.
  - Machine Learning techniques were employed for prediction purposes.
- Summary of all results
  - Valuable data was successfully gathered from publicly available sources.
  - EDA helped identify the key features for predicting the success of launchings.
  - Machine Learning Prediction determined the optimal model for effectively utilizing the collected data and determining the important characteristics driving this opportunity.

Introduction

- Project background and context
  - SpaceX is a leading company in the space industry that launches reusable rockets and lands them on ground pads or drone ships, achieving a high success rate and reducing launch costs.
  - Predicting rocket landing outcomes is a complex problem that involves many factors and requires reliable data and analytical tools, which are not easily accessible or available to the public.
  - SpaceY is a new startup that wants to compete with SpaceX by predicting the successful return of the first stage of rockets
- Problems you want to find answers
  - The objective is to assess the feasibility of Space Y, a new company, in competing with Space X.
  - Predicting the successful landings of the first stage of rockets to estimate total launch costs effectively and identifying the optimal launch location

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from 2 sources:
    - Space X API (https://api.spacexdata.com/v4/rockets/)
    - WebScraping

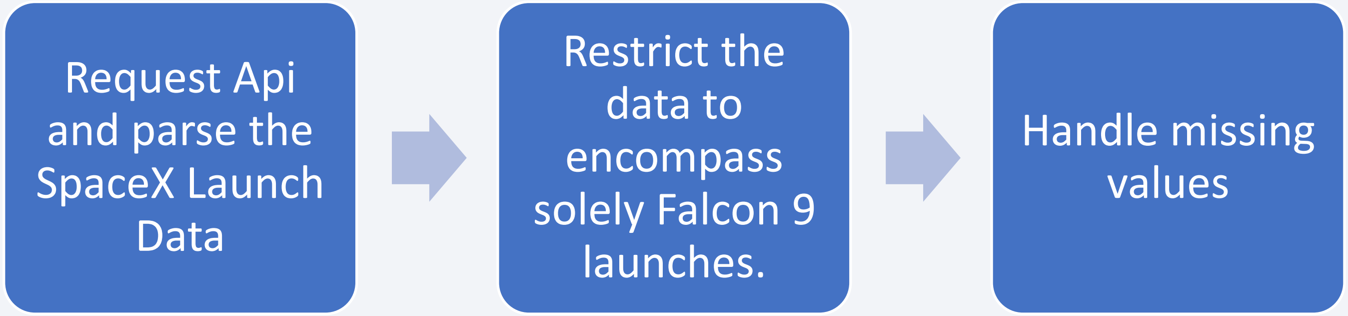
      (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - After summarizing and analyzing features, the collected data was enhanced by generating a landing outcome label using outcome data.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

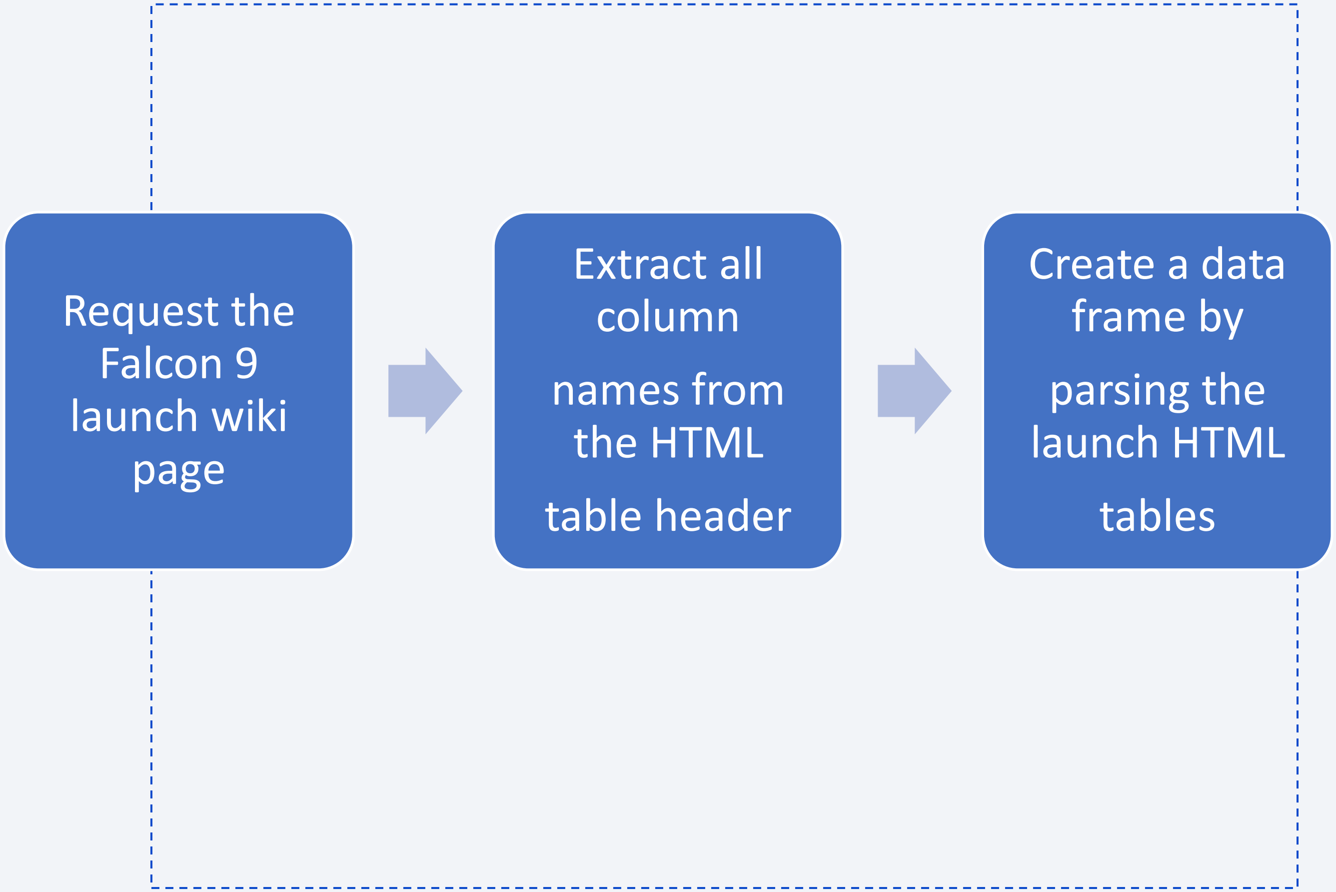
Data Collection – SpaceX API

- Data can be obtained from SpaceX's public API, which was utilized in accordance with the accompanying flowchart.
- The obtained data is then persisted for further use.
- Github URL: https://github.com/santiB73/SpaceY-FinalA/blob/main/spacex-data-collection-api.ipynb



Data Collection - Scraping

- Obtain data regarding SpaceX launches from Wikipedia as an additional source.
- Scrape the data from Wikipedia following the flowchart and store it persistently.
- Github URL: https://github.com/santiB73/SpaceY-FinalA/blob/main/webscraping.ipynb



Data Wrangling

- Conducted initial Exploratory Data Analysis (EDA) on the dataset.
- Calculated the number of launches per site, occurrences of each orbit, and occurrences of mission outcome per orbit type.
- Created the landing outcome label based on the Outcome column.

- Github URL: https://github.com/santiB73/SpaceY-FinalA/blob/main/spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Utilized scatterplots and barplots to visually explore the relationships between pairs of features.
    - Explored the relationships between Payload Mass and Flight Number, Launch Site and Flight Number, Launch Site and Payload Mass, Orbit and Flight Number, and Payload and Orbit.
- Github URL: https://github.com/santiB73/SpaceY-FinalA/blob/main/eda-dataviz%20(1).ipynb

EDA with SQL

- 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 (ground pad)) between the date 06-04-2010 and 03-20-2017.
- GitHub URL: https://github.com/santiB73/SpaceY-FinalA/blob/main/eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Employed various visual elements, including markers, circles, lines, and marker clusters, on Folium Maps.
- Markers were utilized to represent points such as launch sites.
- Circles were used to highlight specific areas around coordinates, for instance, NASA Johnson Space Center.
- Marker clusters were employed to group events within each coordinate, such as launches at a particular launch site.
- Lines were utilized to indicate distances between two coordinates.
- GitHub URL: https://github.com/santiB73/SpaceY-FinalA/blob/main/launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Employed various graphs and plots to visualize the data, including:
  - Percentage of launches by site.
  - Payload range.
- This combination of visualizations facilitated a rapid analysis of the relationship between payloads and launch sites, aiding in the identification of the optimal launch site based on payload requirements.
- GitHub URL: https://github.com/santiB73/SpaceY-FinalA/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.

Data preparation and standardization → Model testing using different combinations of hyperparameters → Results comparison

- Github URL: https://github.com/santiB73/SpaceY-FinalA/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite%20(1).ipynb

Results

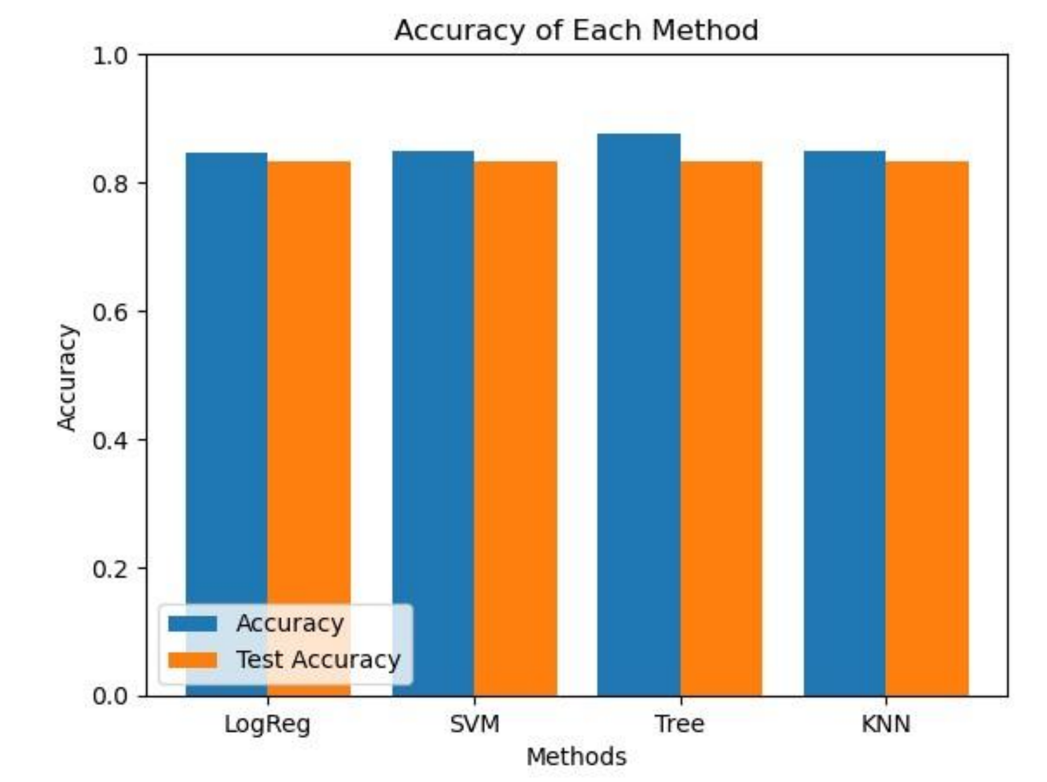
- Key findings from exploratory data analysis:
  - SpaceX operates from 4 different launch sites.
  - The initial launches were conducted by SpaceX itself and NASA.
  - The average payload of the F9 v1.1 booster is 2,928 kg.
  - The first successful landing outcome occurred in 2015, five years after the first launch.
  - Many Falcon 9 booster versions successfully landed on drone ships with payloads above the average.
  - Almost 100% of mission outcomes were successful.
  - Two booster versions (F9 v1.1 B1012 and F9 v1.1 B1015) failed to land on drone ships in 2015.
  - The number of successful landing outcomes improved over the years.

Results

- Insights from interactive analytics:
  - Launch sites are strategically located in safe areas, often near the sea, and are supported by robust logistic infrastructures.
  - The majority of launches occur at launch sites along the east coast.

Results

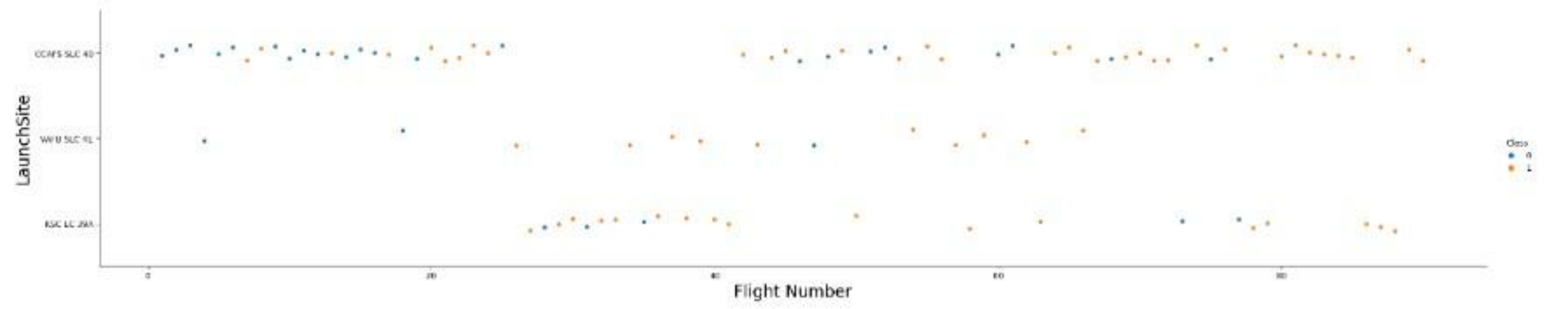
- Predictive Analysis Findings:
- The predictive analysis revealed that all models achieved a test accuracy of 83%. However, the Decision Tree Classifier outperformed the others with an impressive accuracy score of 87.5%. Therefore, the Decision Tree Classifier is considered the most reliable model for predicting successful landings.



Section 2

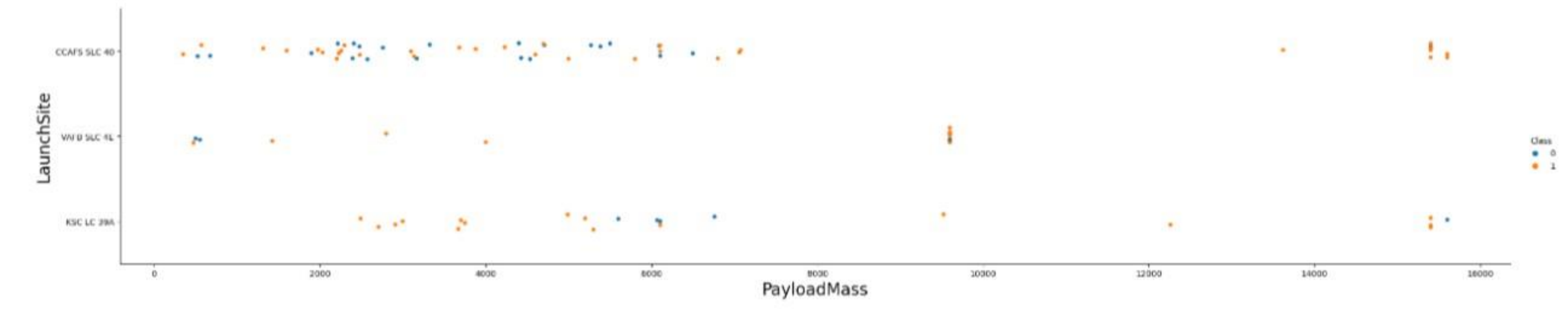
Insights drawn from EDA

Flight Number vs. Launch Site



- Based on the provided plot, it is evident that the current top-performing launch site is CCAF5 SLC 40, with the highest number of recent successful launches.
- Following closely is VAFB SLC 4E in second place, and KSC LC 39A in third place.
- Additionally, the plot illustrates an overall improvement in the success rate over time.

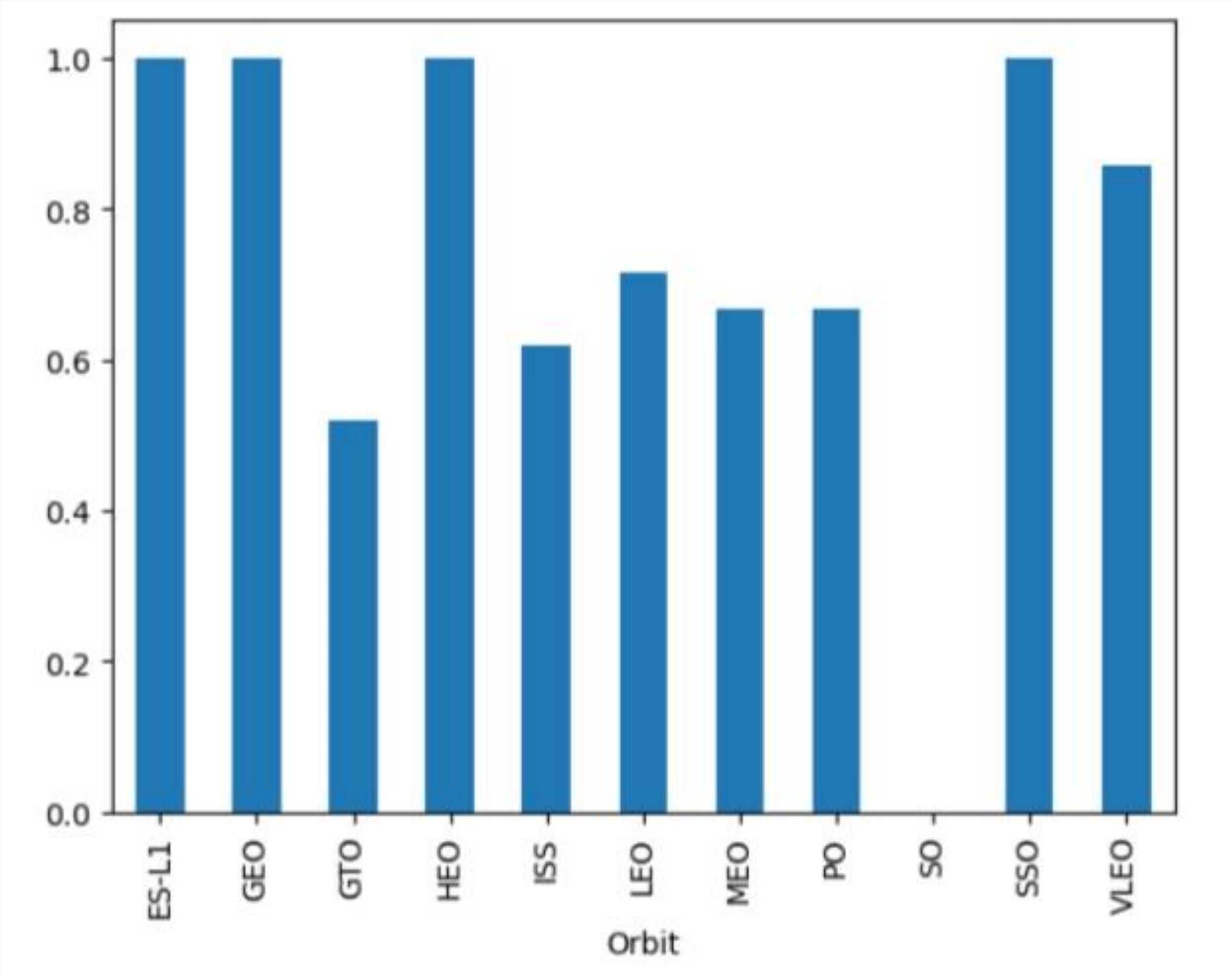
Payload vs. Launch Site



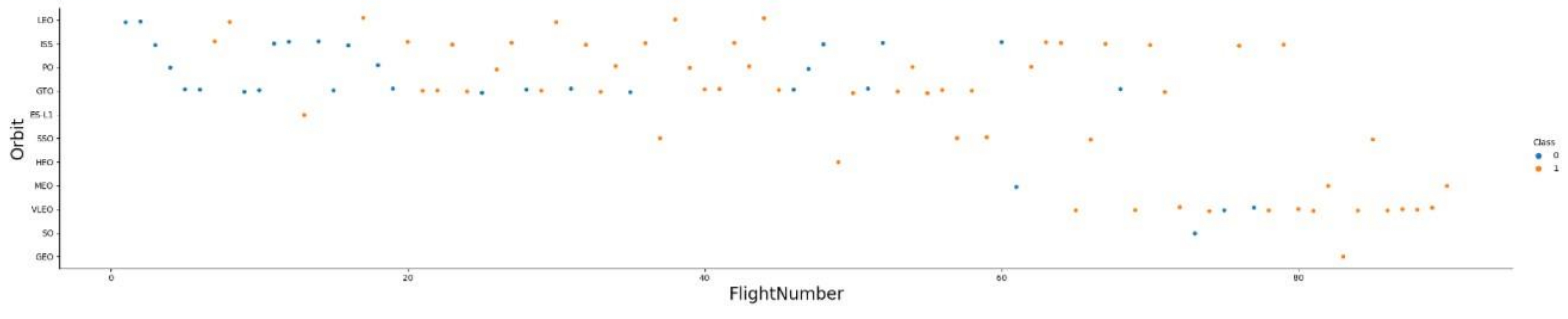
- Payloads weighing over 9,000kg (approximately the weight of a school bus) exhibit a high success rate.
- It appears that payloads exceeding 12,000kg are only feasible at the CCAFS SLC 40 and KSC LC 39A launch sites.

Success Rate vs. Orbit Type

- The highest success rates are observed for the following orbits:
  - ES-L1
  - GEO
  - HEO
  - SSO
- This is followed by:
  - VLEO with a success rate above 80%
  - LFO with a success rate above 70%

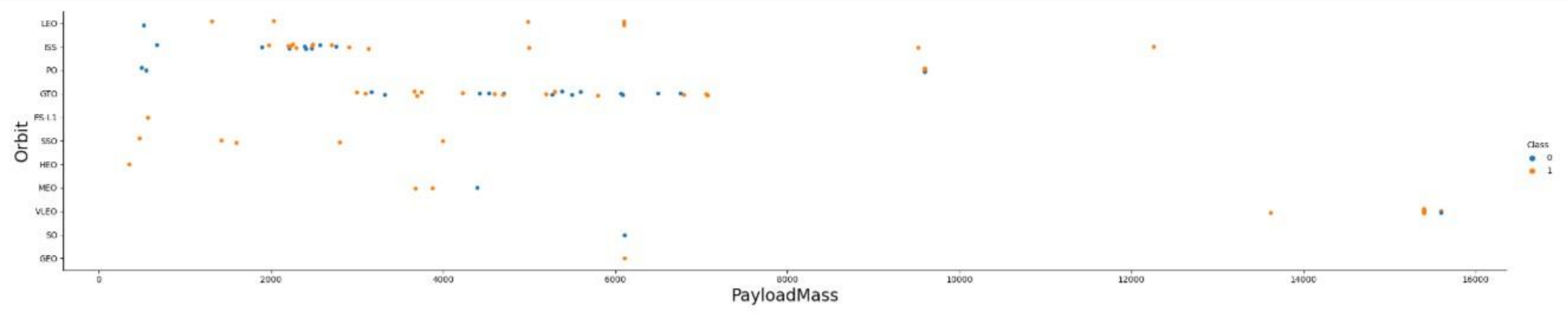


Flight Number vs. Orbit Type



- There appears to be an improvement in success rates for all orbits over time.
- The increasing frequency of VLEO orbit presents a new business opportunity.

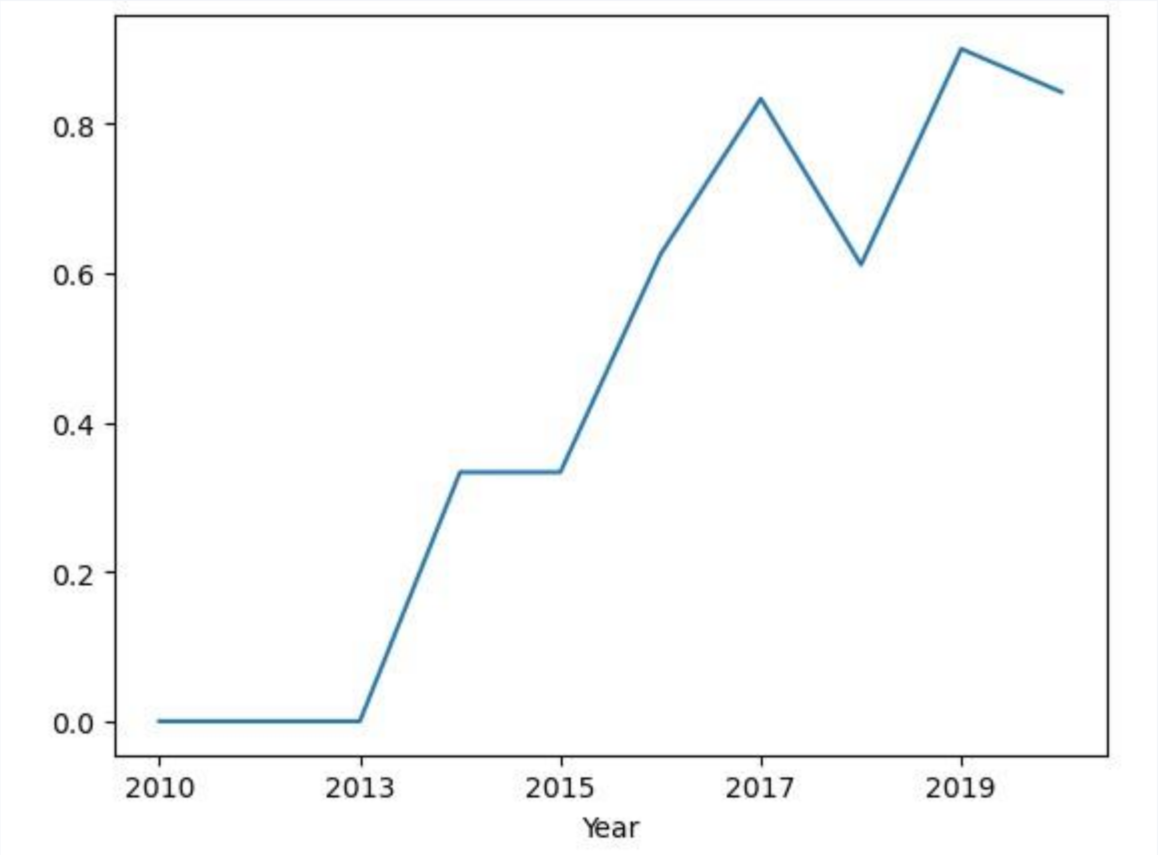
Payload vs. Orbit Type



- There doesn't seem to be a correlation between payload and success rate for the GTO orbit.
- The ISS orbit has a wide range of payloads and a favorable success rate.
- There are only a few launches to the SO and GEO orbits.

Launch Success Yearly Trend

- The success rate began to rise in 2013 and continued to increase until 2020.
- The initial three years appear to have been a period of adjustments and technological improvements.



All Launch Site Names

- According to data, there are four launch sites:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- The unique occurrences of "launch_site" values from the dataset are obtained by selecting them.

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

- Here we can see five samples of Cape Canaveral launches.

Total Payload Mass

- Total payload carried by boosters from NASA: 111.268 KG

- The total payload calculated above corresponds to NASA, as it includes the sum of all payloads with codes containing 'CRS'.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1: 2928.4 KG

- By filtering the data using the mentioned booster version and calculating the average payload mass, we obtained a value of 2928 kg.

First Successful Ground Landing Date

- First successful landing outcome on ground pad: 01/08/2018

- By filtering the data based on successful landing outcomes on the ground pad and retrieving the earliest date, it is possible to identify the first occurrence, which occurred on 01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:
- F9 FT B1022
- F9 FT B1026
- F9 FT B1021.2
- F9 FT B1031.2

- Based on the applied filters, the following four distinct booster versions were selected.

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

Mission_Outcome	QTY
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- By grouping mission outcomes and tallying the number of records for each group, we arrived at the aforementioned summary.

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass:
- F9 B5 B1048.4
- F9 B5 B1048.5
- F9 B5 B1049.4
- F9 B5 B1049.5
- F9 B5 B1049.7
- F9 B5 B1051.3
- F9 B5 B1051.4
- F9 B5 B1051.6
- F9 B5 B1056.4
- F9 B5 B1058.3
- F9 B5 B1060.2
- F9 B5 B1060.3

2015 Launch Records

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

month	Landing_Outcome	Booster_Version	Launch_Site
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40

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

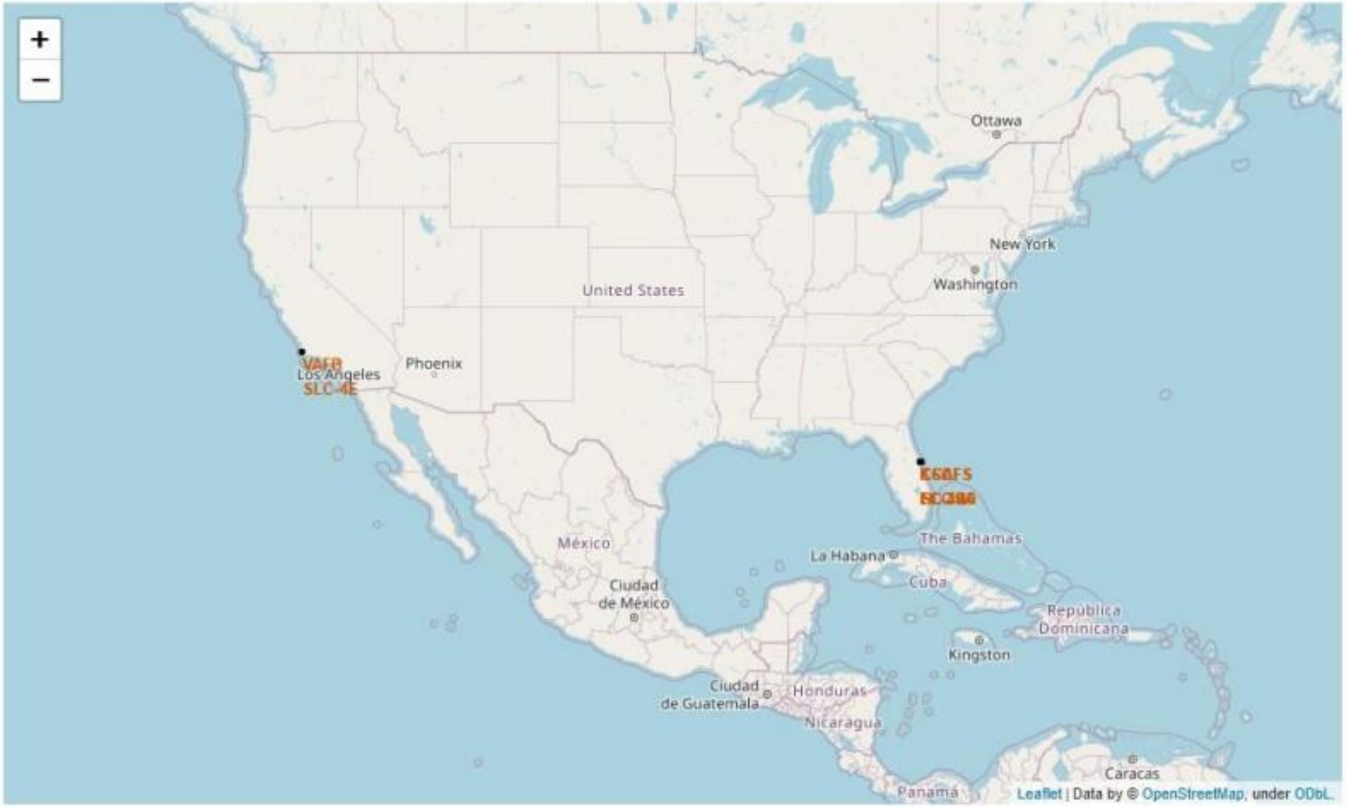
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order:

Landing_Outcome	QTY
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Section 3

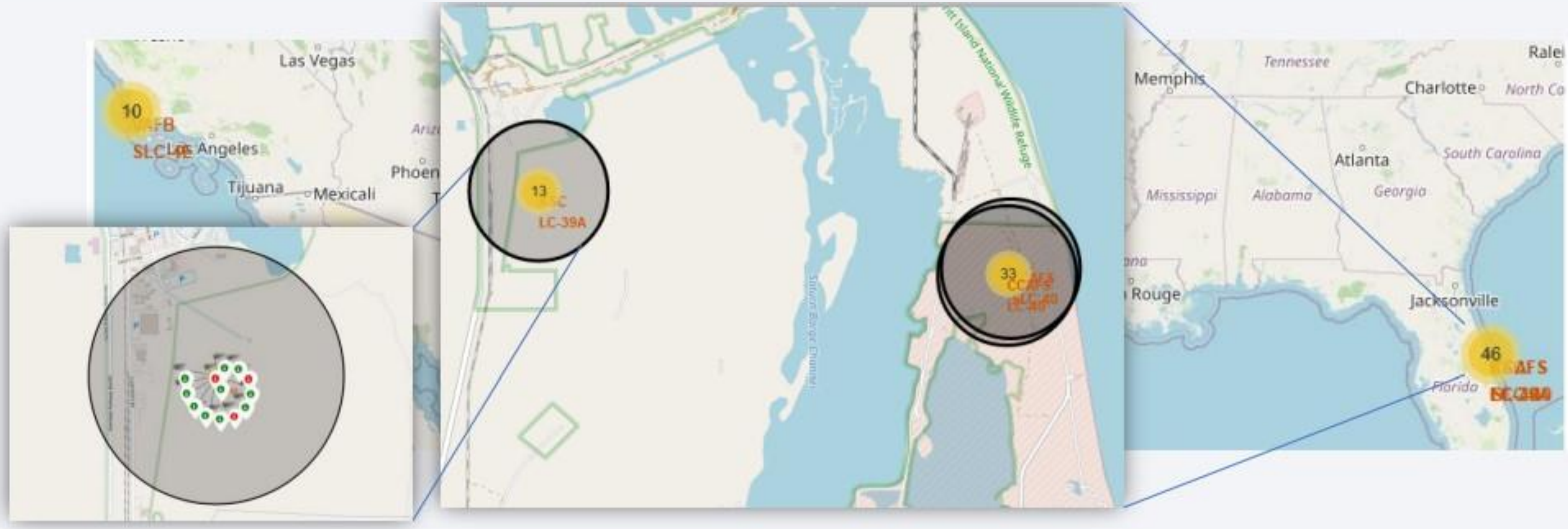
Launch Sites Proximities Analysis

All launch sites



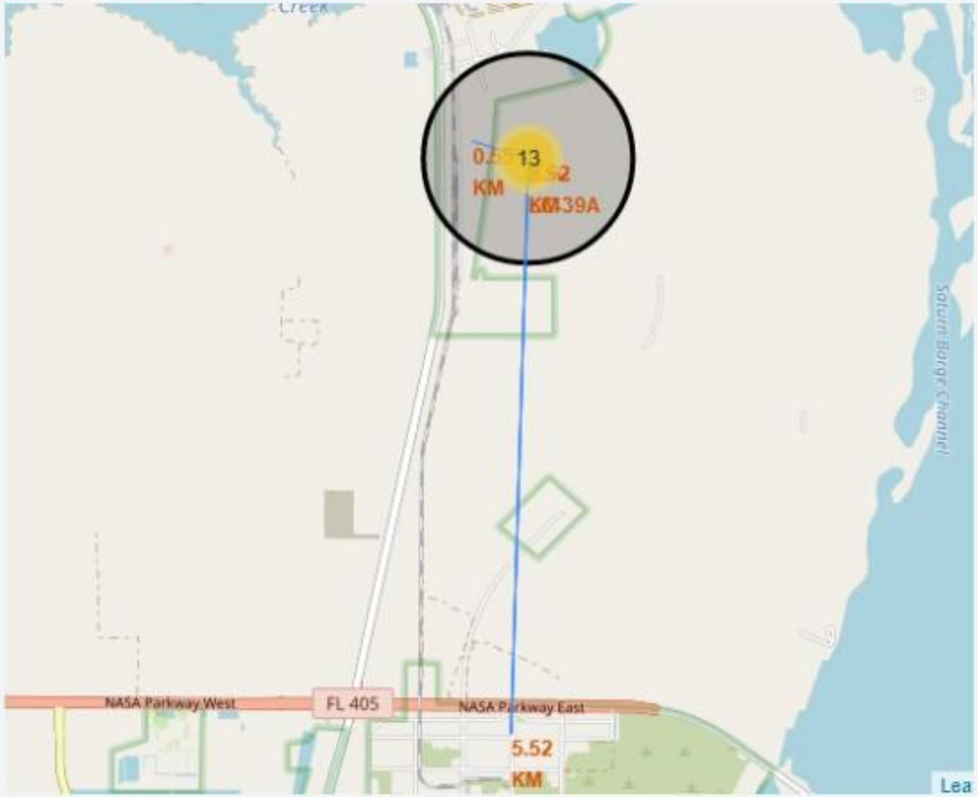
- Launch sites are strategically located in close proximity to the sea, likely for safety reasons, while also maintaining convenient access to roads and railroads.

Launch Outcomes by Site



- Successful outcomes are denoted by green markers, while red markers indicate failures.

Logistics and Safety

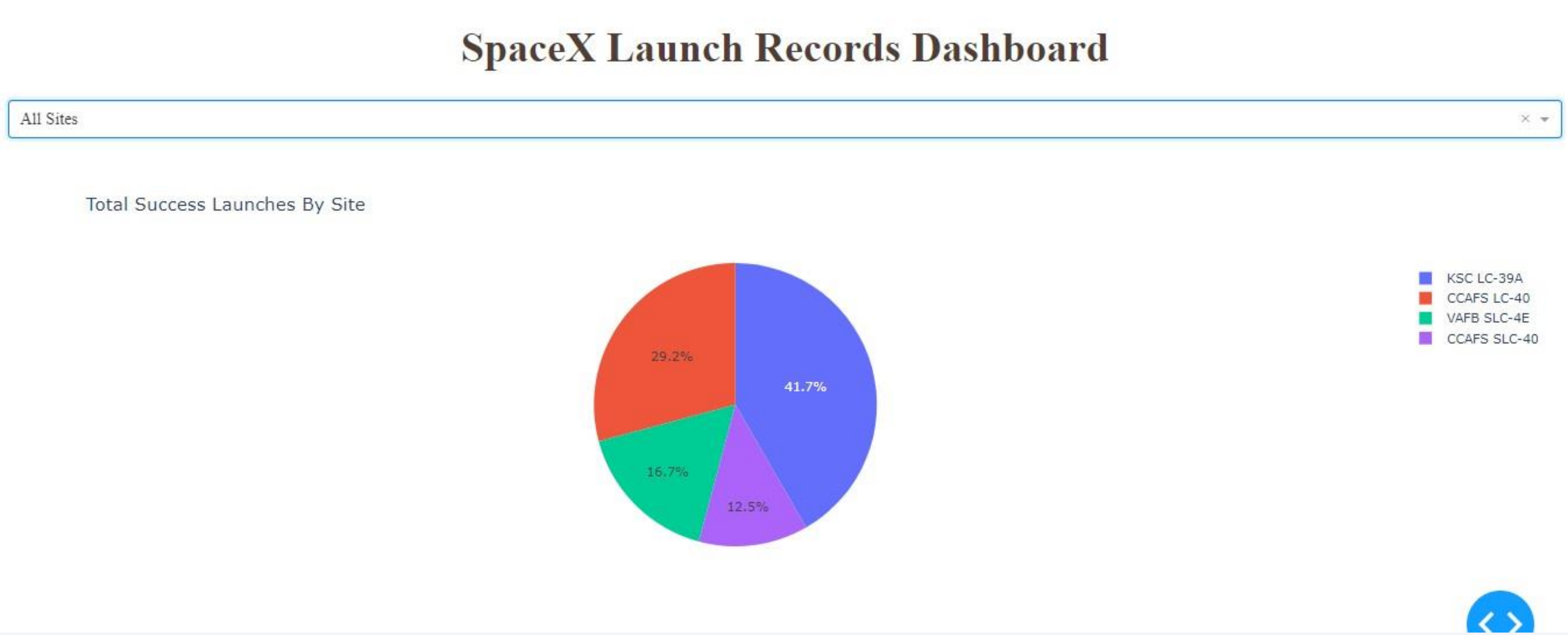


- KSC LC-39A launch site exhibits favorable logistical characteristics as it is located in proximity to both rail and road infrastructure, while also being relatively distant from populated areas.

Section 4

Build a Dashboard with Plotly Dash

Successful Launches by Site

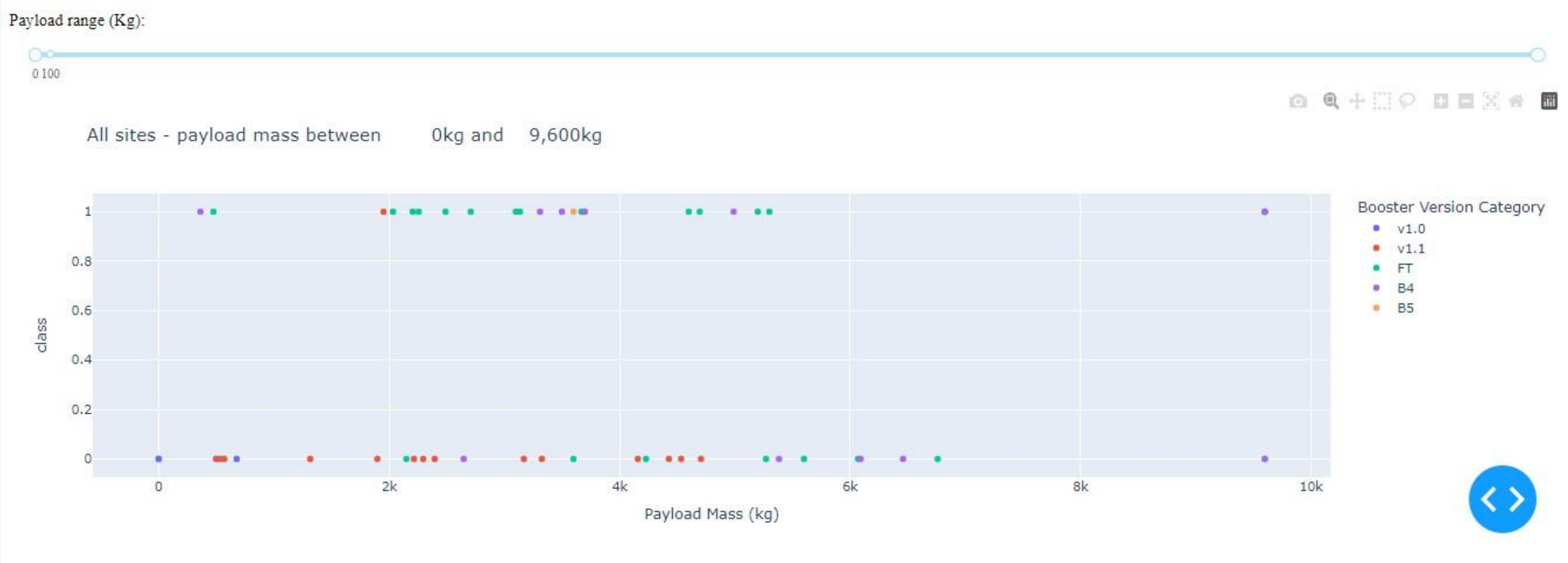


- The location of launch sites appears to be a critical factor influencing the success of missions.

Launch Site With Highest Success Ratio

- KSC LC-39A has the highest success ratio with a ratio of 76.9%

Payload vs. Launch Outcome



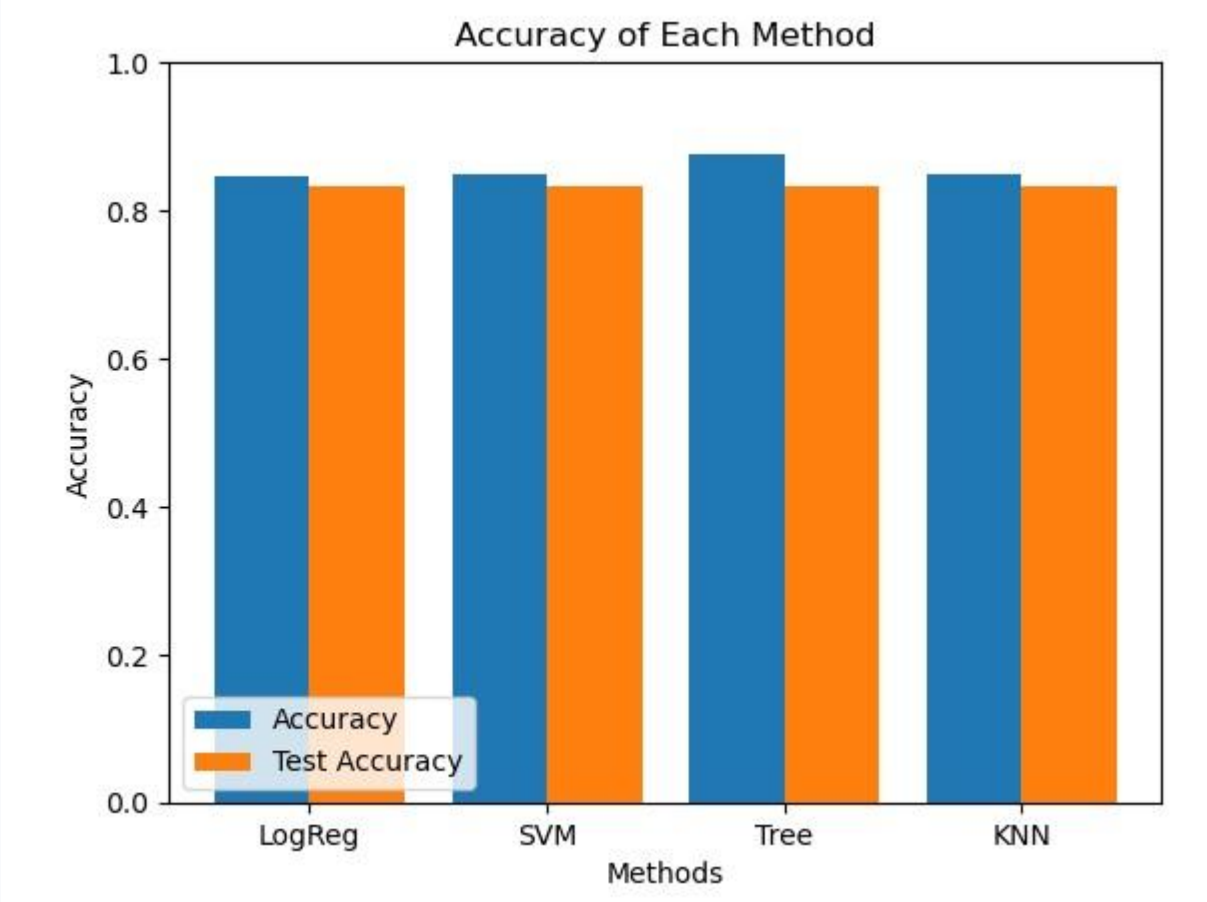
- The combination of payloads under 6,000kg and FT boosters yields the highest success rate.

Section 5

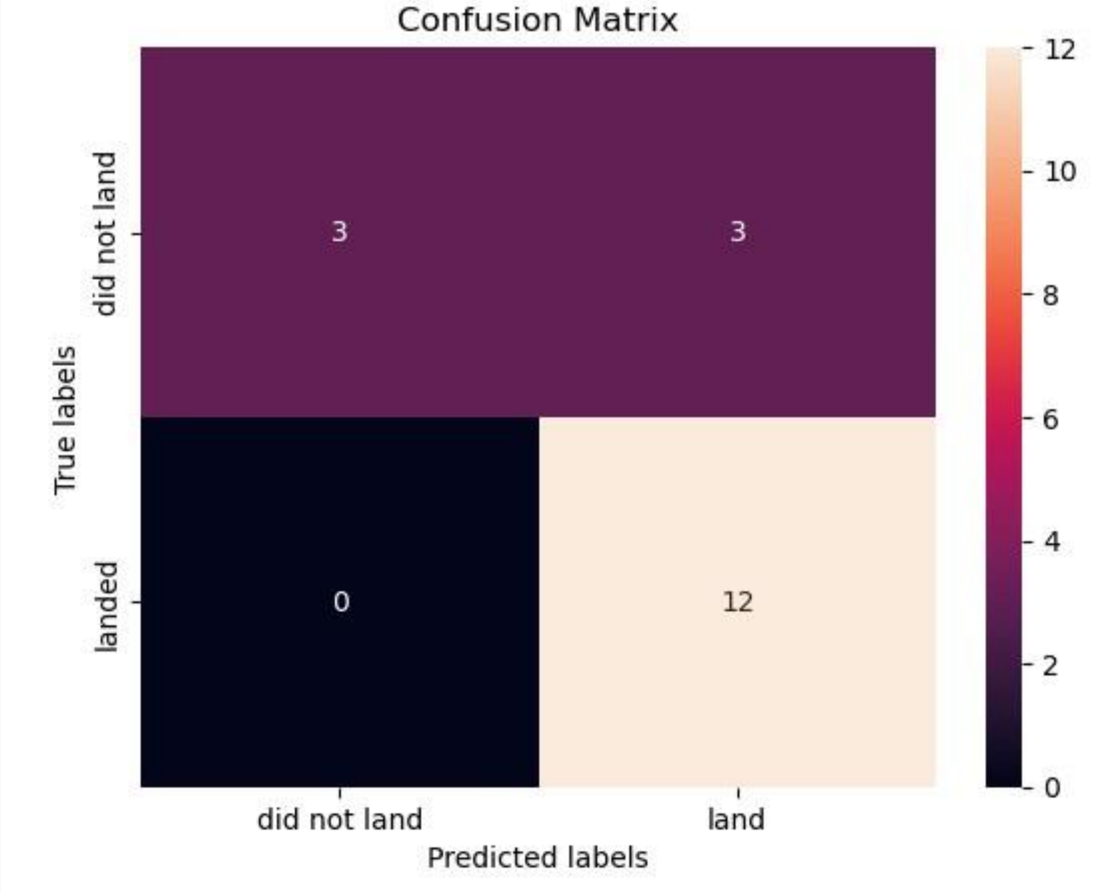
Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were evaluated and their accuracies are shown in the plot beside.
- All models achieved a test accuracy of 83%. However, the Decision Tree Classifier outperformed the others with an impressive accuracy score of 87.5%. Therefore, the Decision Tree Classifier is considered the most reliable model for predicting successful landings.



Confusion Matrix



- The confusion matrix shows that the model is great predicting when the first stage landed but not when it did not land

Conclusions

- Multiple data sources were examined, enhancing the conclusions throughout the analysis.
- KSC LC-39A emerged as the optimal launch site.
- Launches with payloads exceeding 7,000kg exhibit lower risks.
- While the majority of mission outcomes are successful, successful landing outcomes demonstrate improvement over time, reflecting advancements in processes and rockets.
- The Decision Tree Classifier is a valuable tool for predicting successful landings and enhancing profitability.

Appendix

- SQL Queries can be found in the following link: https://github.com/santiB73/SpaceY-FinalA/blob/main/eda-sql-coursera_sqllite.ipynb
- Data sets and all the code used in this project can be found in the following link: https://github.com/santiB73/SpaceY-FinalA

Thank you!