IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Adarsh K
2023/05/22

Outline

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

Executive Summary

- The following methodologies were used to collect, and analyze data
  - Data collection using SpaceX API and Web scraping.
  - Exploratory Data Analysis (EDA), data wrangling, data visualization, and interactive visual analytics (dashboards)
  - Machine Learning – Prediction.
- Summary of all results
  - Valuable data was collected from public sources.
  - EDA provided a way to identify features which were best to predict the success of launches.
  - Machine Learning performed on collected data revealed the best model to predict the characteristics which are important to drive opportunities in the best possible way.

Introduction

- Objective
  - Evaluate the viability of a new company SpaceY to compete with SpaceX
- Desirable results
  - Estimate the total cost for launches, by predicting successful landings of the first stage of rockets.
  - Analyze the sites used for launching rockets.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data for SpaceX Falcon9 launches were collected from 2 sources:
    - SpaceX REST API (https://api.spacexdata.com/v4/launches/past)
    - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_)
- Perform data wrangling
  - Collected data was cleaned, added a landing outcome label based on the outcome data after summarizing and analyzing features.
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

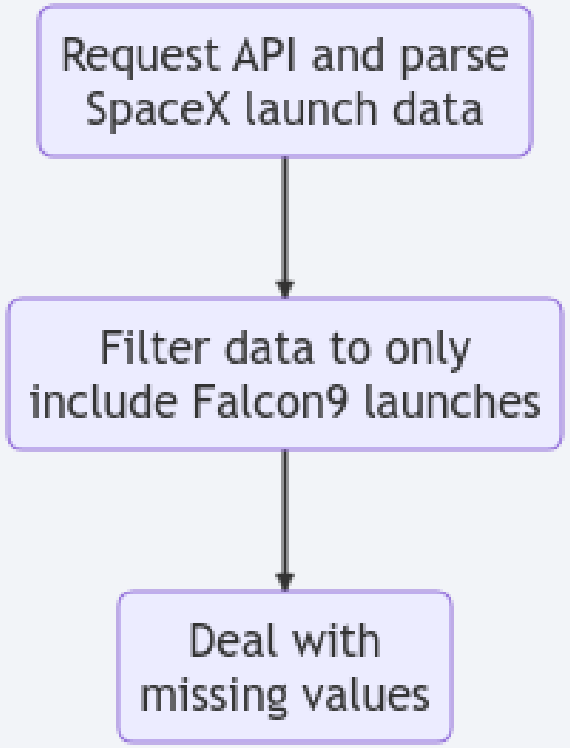
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Cleaned data collected until this stage was scaled using StandardScaler.
  - Scaled data was then split into training and testing datasets.
  - Four different classification models – Logistic Regression, Support Vector Machines (SVM), Decision Trees, and k-Nearest Neighbors (kNN) were trained using the training dataset, and tuned hyperparameters obtained using cross validation technique GridSearchCV.

Data Collection

- Data was collected from two sources
    1. SpaceX REST API (https://api.spacexdata.com/v4/launches/past) and
    2. WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_)

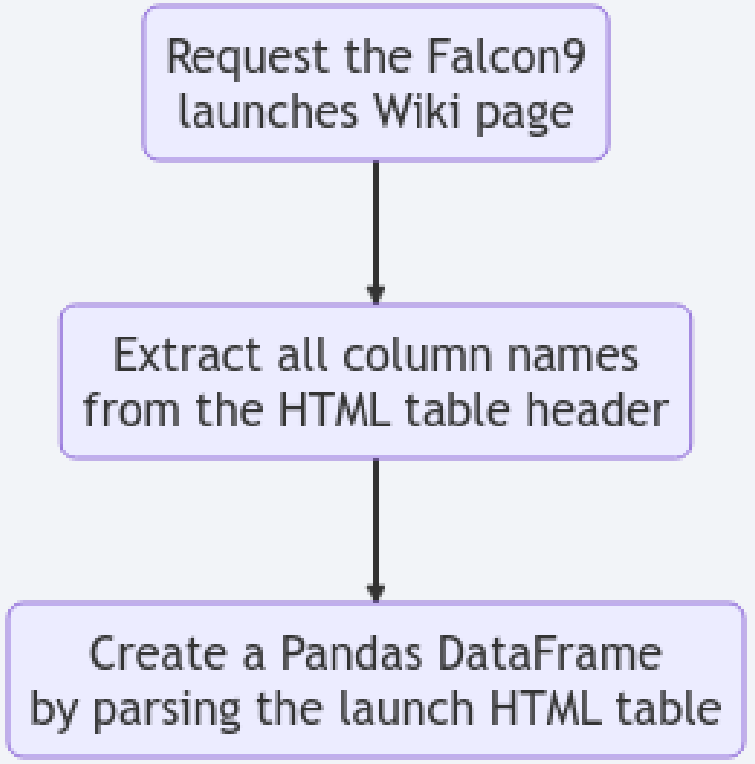
Data Collection – SpaceX API

- SpaceX offers a public API from where the data can be obtained and used.
- Refer to the flowchart to understand the data collection process.
- Source code



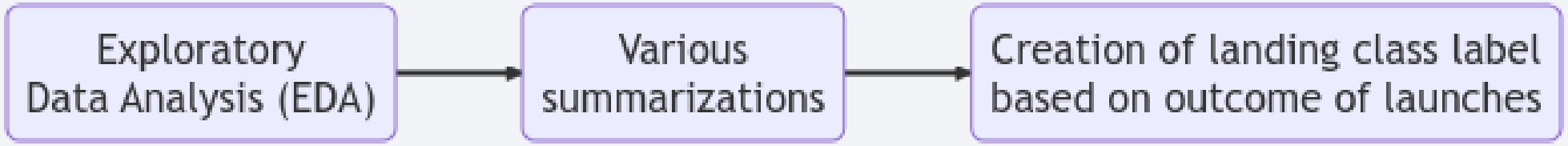
Data Collection - WebScraping

- Data of SpaceX launches can also be obtained from public sites like Wikipedia.
- Refer to the flowchart to understand the data collection process.
- Source code



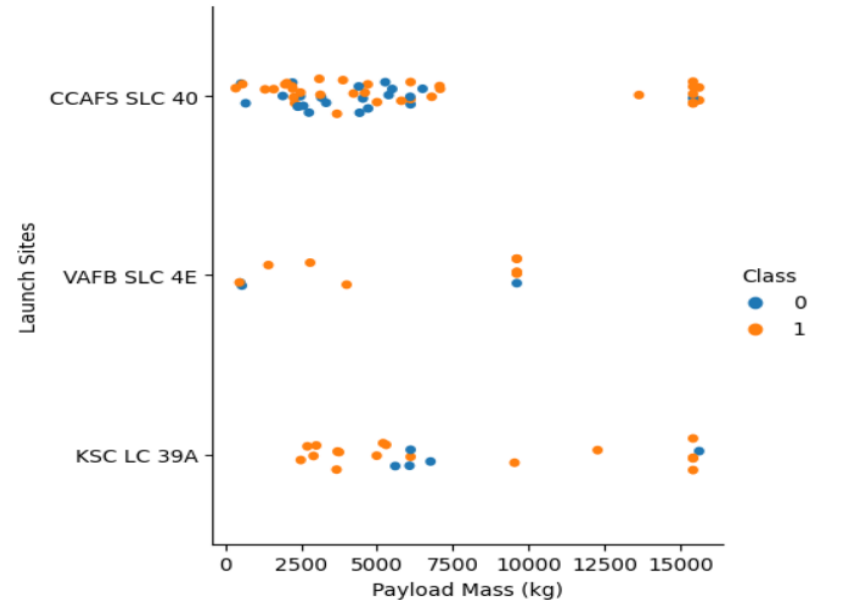
Data Wrangling

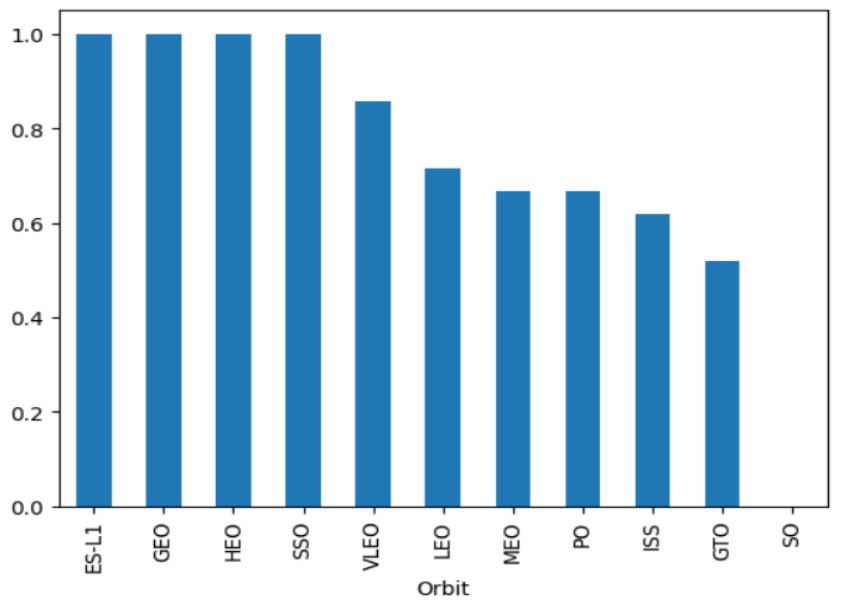
- Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summaries of launches per site, occurrence of each orbit type, occurrence of mission outcome per orbit type, and success rate of launch sites were calculated.
- Finally, a landing class label was created based on the outcome of the launch.
- Refer to the flowchart to understand the data wrangling process.
- Source code



EDA with Data Visualization

- To explore and better understand the data, scatter plots and bar plots were used to visualize the relationships between different pairs of features.
  - Payload Mass x Flight Number, Launch Site x Flight Number, Launch Site x Payload Mass, Orbit x Flight Number, Orbit x Payload Mass
- Source code





EDA with SQL

- The following SQL queries were performed
  1. Names of unique launch sites.
  2. Five records of launch sites beginning with “CCA”.
  3. Total payload mass carried by NASA (CRS) boosters.
  4. Average payload mass carried by F9 v1.1 boosters.
  5. Date of the first successful landing outcome in ground pad.
  6. Booster names which have success in drone ship and have payload mass between 4,000 to 6,000 kg.
  7. Total number of successful and failure mission outcomes.
  8. Booster names which have carried the highest payload mass, using a subquery.
  9. Failed landing outcomes in drone ship, their booster names, month, and launch sites for the year 2015.
  10. Rank of count of landing outcomes between the dates 04-06-2010 and 20-03-2017 in descending order.
- Source code

Build an Interactive Map with Folium

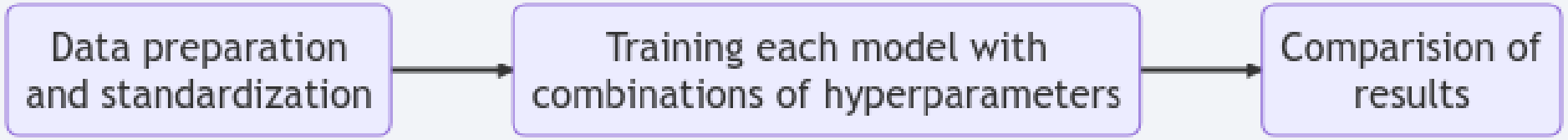
- Markers, Circles, Lines and MarkerCluster were used with Folium Maps
  - Markers indicate points like launch sites.
  - Circles highlight the area around specific coordinate.
  - MarkerCluster indicate a group of events in each coordinate, like launches in a specific launch site.
  - Lines are used to indicate distance between two coordinates.
- Source code

Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize the data.
  - Percentage of launches by all sites or a specific site (pie chart along with a dropdown to select specific sites)
  - Payload mass range (scatter plot along with a payload mass slider to select payload mass range)
- This combination allowed to quickly analyze the relationship between payload mass and launch sites, helping to identify the best site to launch the rocket.
- Source code

Predictive Analysis (Classification)

- Four classification models – Logistic Regression, Support Vector Machines (SVM), Decision Trees and k-Nearest Neighbor (kNN) – were trained on the training dataset, with tuned hyperparameters obtained using cross validation technique – GridSearchCV.
- For each model, a confusion matrix was plotted, and accuracy of the model on test dataset was calculated.
- Source code

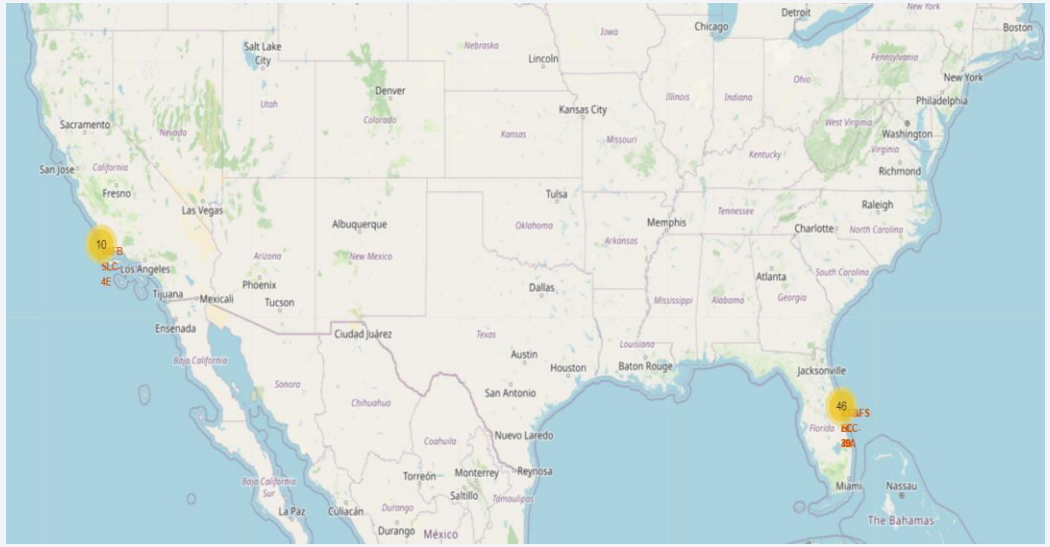


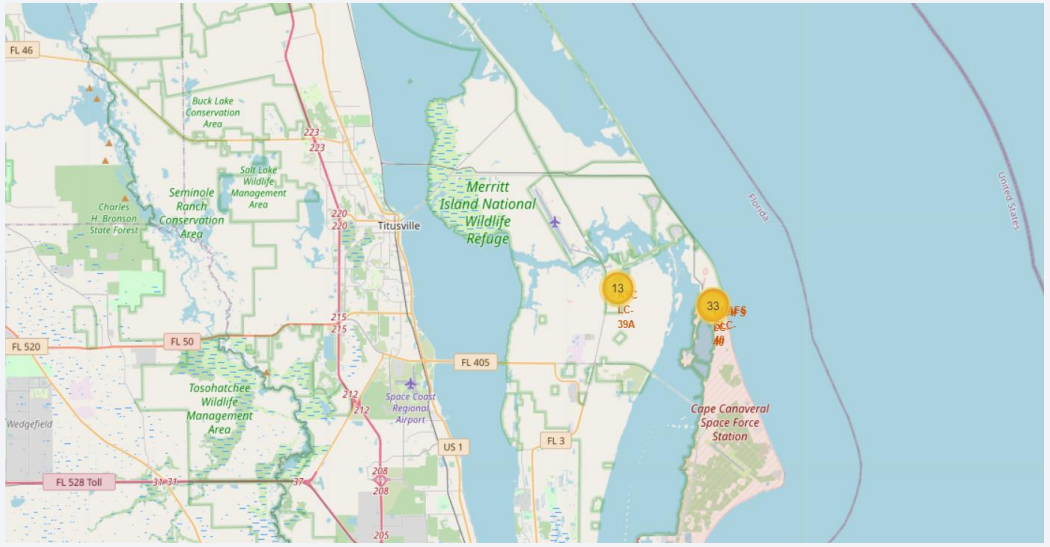
Results

- Exploratory data analysis results
  - SpaceX uses 4 different launch sites.
  - Customers for SpaceX at the beginning were SpaceX itself and NASA.
  - The average payload mass is 2,982.4 kg.
  - The first successful landing (ground pad) occurred in 2018, 8 years after the first launch.
  - Yearly success rate is trending upwards from 2013 onwards.
  - Almost all the recent launches have successfully landed.
  - Drone ship, and ground pad landings have the highest successful landing outcomes respectively.

Results

- Interactive Analytics show that the launch sites are in safe distance away from cities, closer to coastal sites and have good logistic infrastructure around.
- Most launches occur in the east coast.

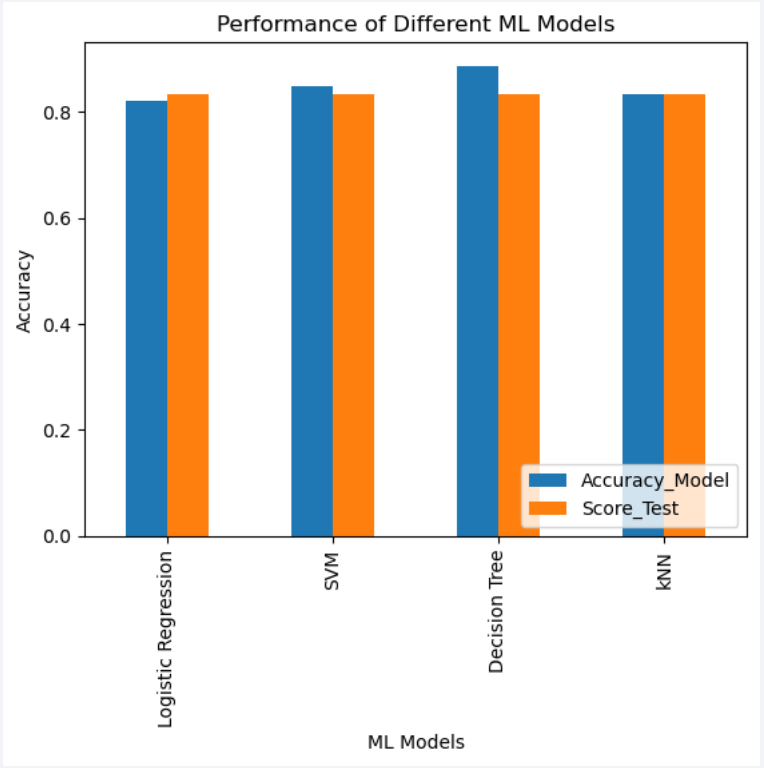


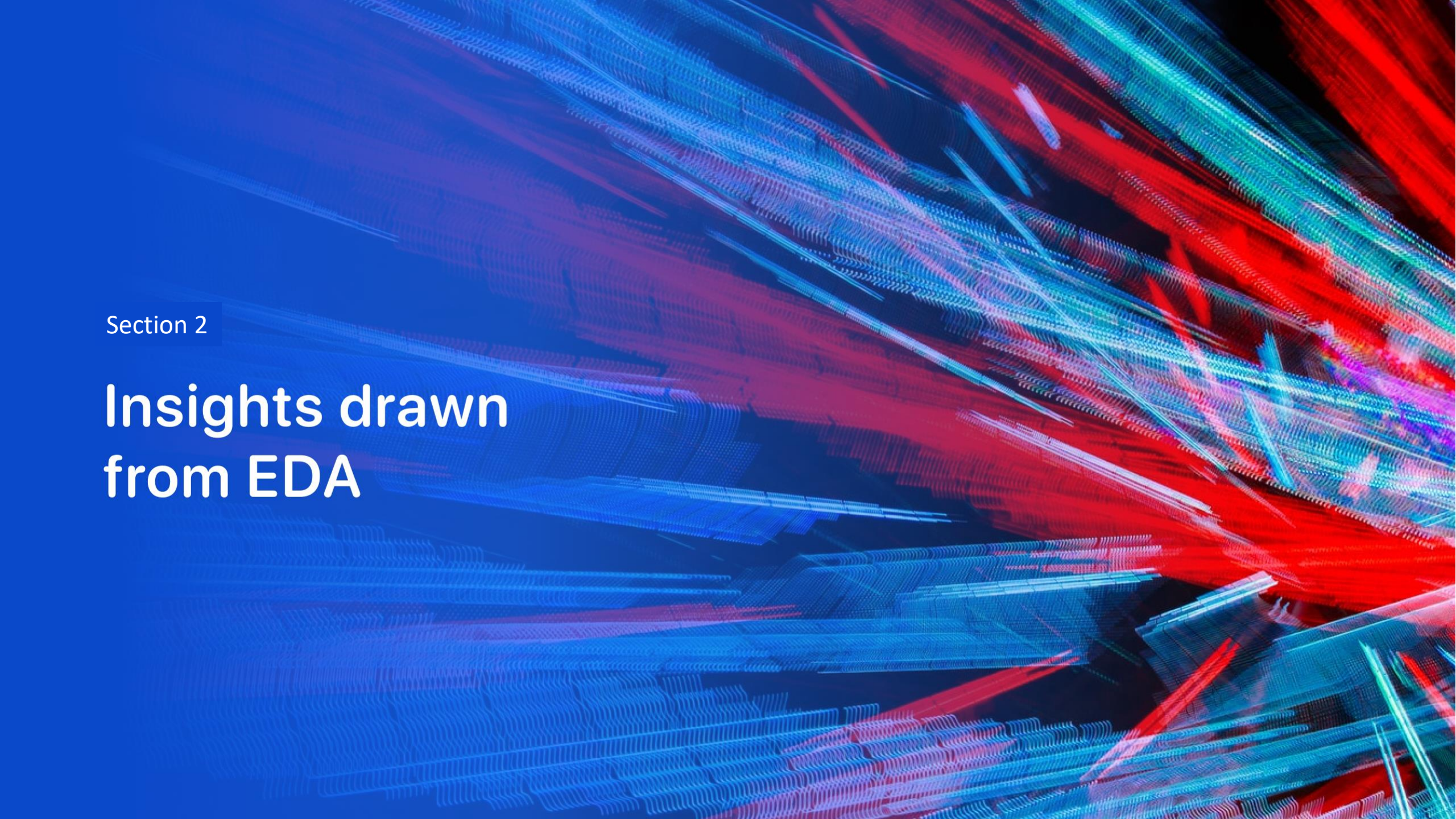


Results

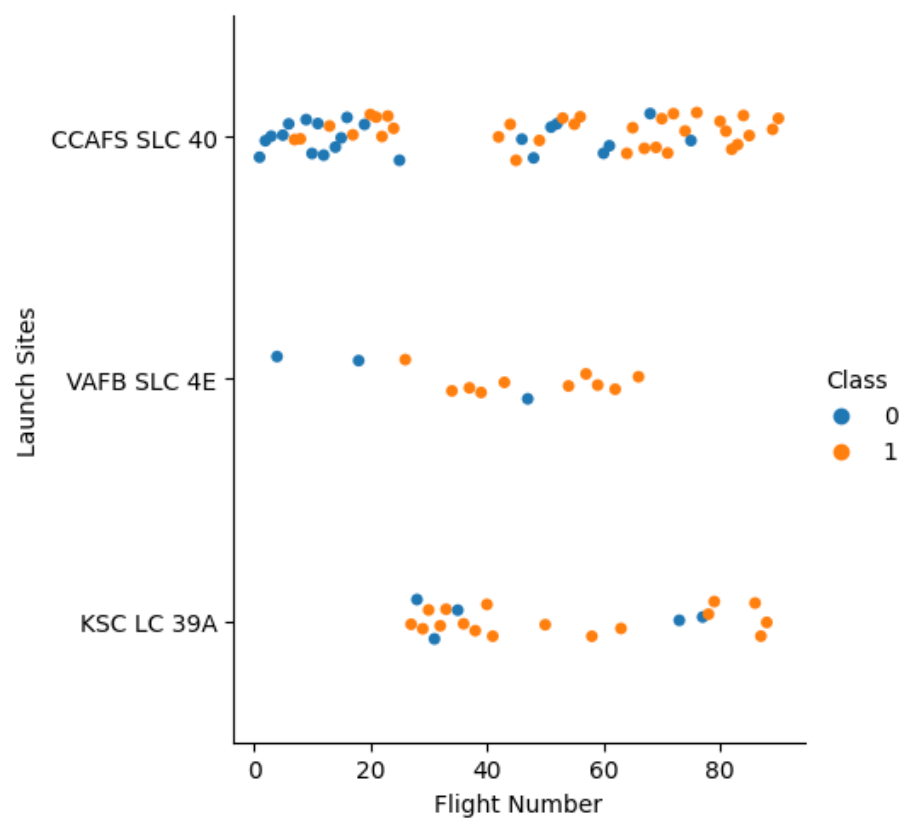
- Predictive Analysis showed that Decision Tree Classifier was the best model, having 88.75% accuracy on training set, and 83.33% accuracy on testing set.

	Accuracy_Model	Score_Test
Logistic Regression	0.821429	0.833333
SVM	0.848214	0.833333
Decision Tree	0.887500	0.833333
kNN	0.833929	0.833333



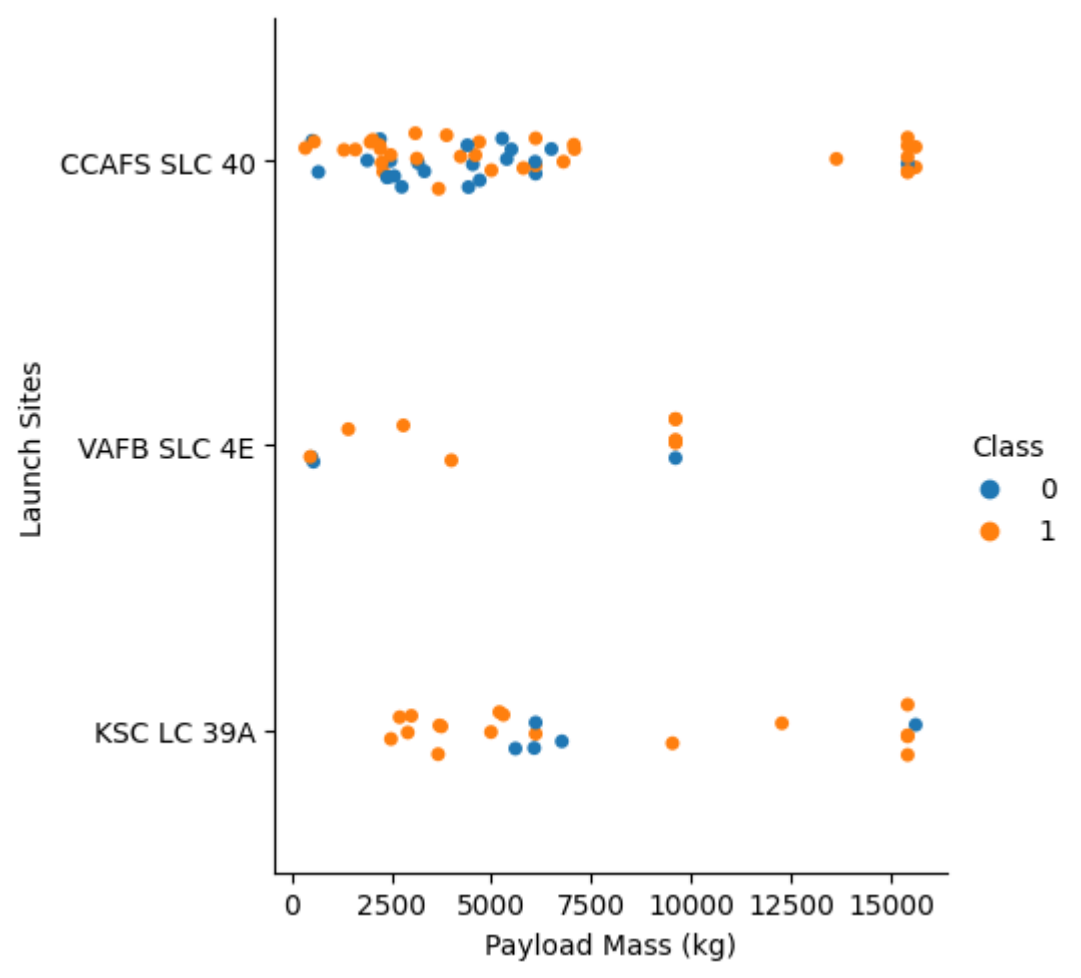

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



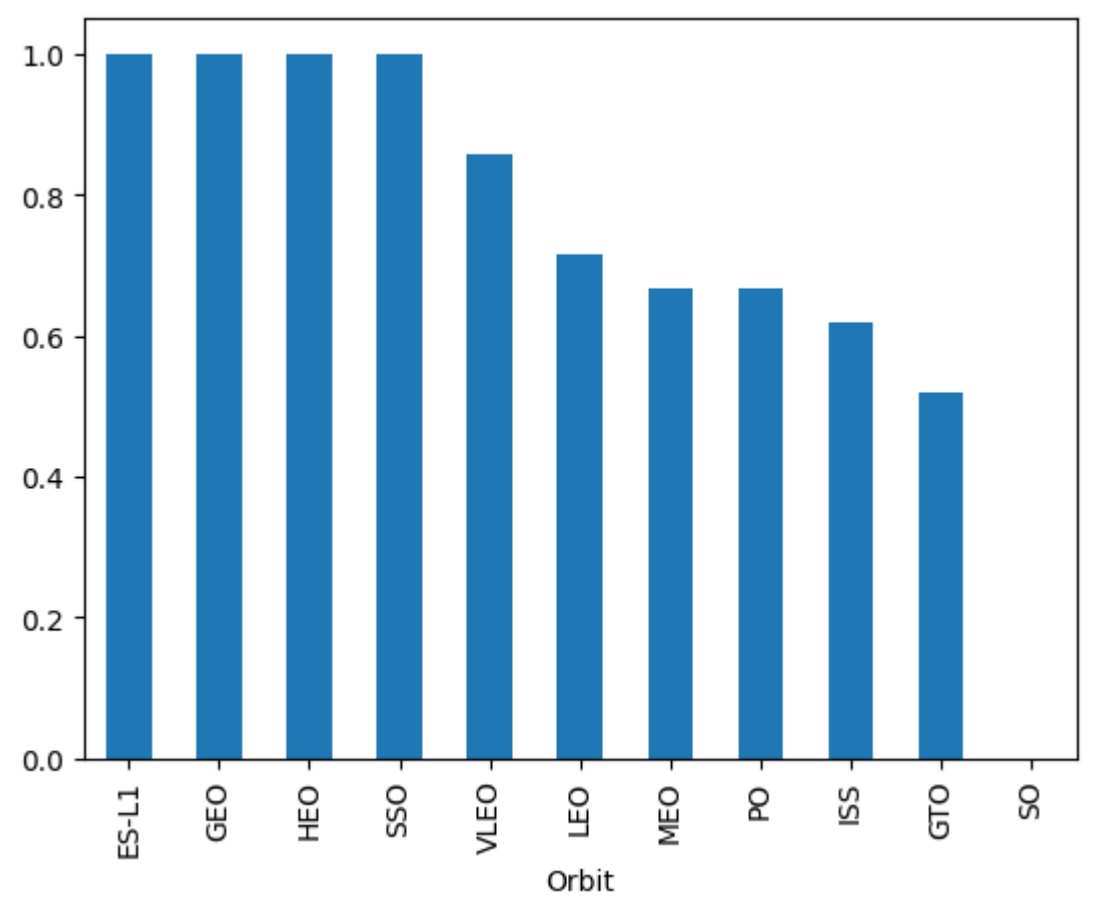
- From the plot, we can verify that most tests happened at CCAFS SLC 40 site.
- The landings failed in the beginning but started to show success over time. Recently all the rockets have been successfully landed.
- Tests have been paused at the site VAFB SLC 4E.
- KSC LC 39A doesn't have any records in the beginning, indicating that SpaceX started using the site later on in their space venturing journey.

Payload vs. Launch Site



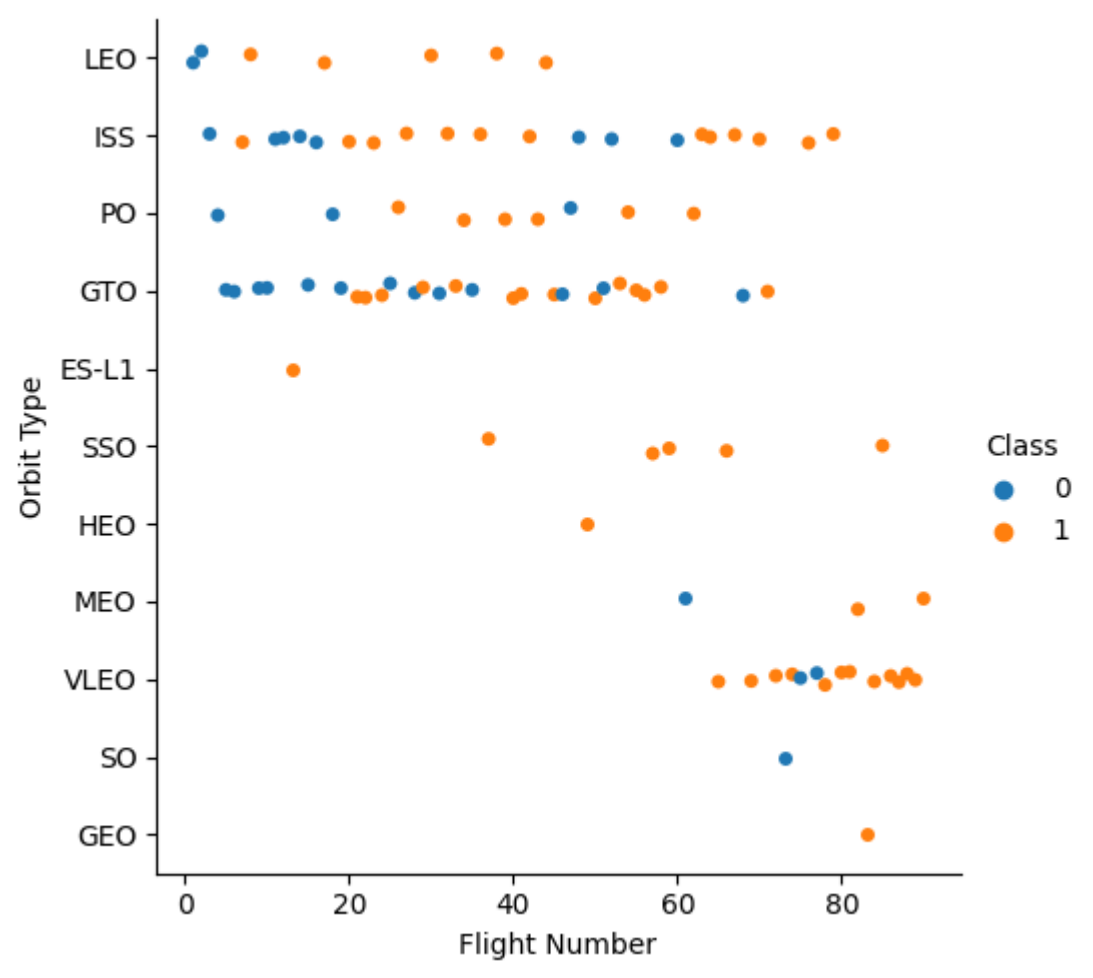
- Payloads over 9,000 kg have excellent success rate.
- Launching payloads over 12,000 kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A sites.

Success Rate vs. Orbit Type



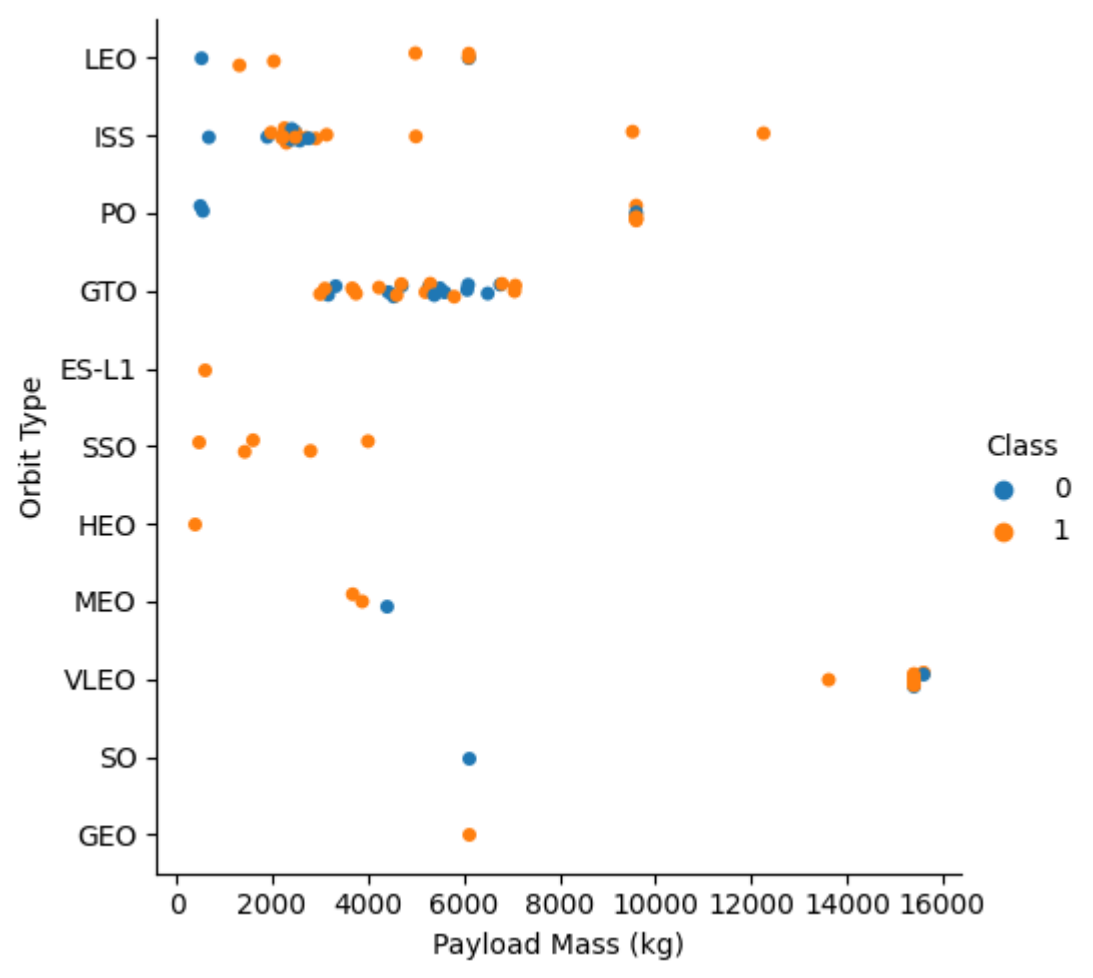
- ES-L1, GEO, HEO, SSO orbit types have almost 100% success rates.
- Followed by VLEO and LEO.

Flight Number vs. Orbit Type



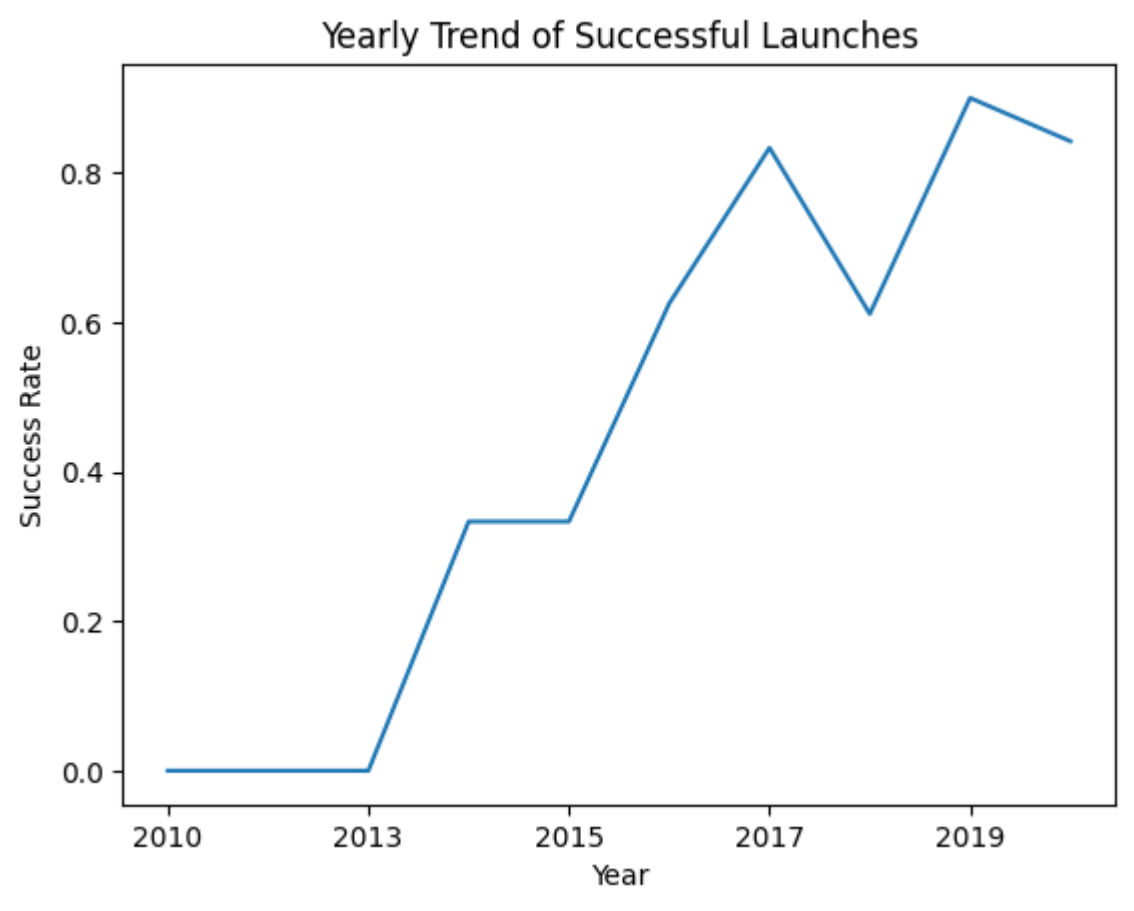
- Success rates improved overtime for all orbit
- VLEO orbit type seems like a new business opportunity, due to its recent increase in frequency.

Payload vs. Orbit Type



- There is no apparent relationship between payload and orbit for GTO.
- VLEO is used to carry high payload masses.
- SSO has seen highest number of successes in lower payload region.
- ISS has widest range of payload masses.

Launch Success Yearly Trend



- Success rate keeps increasing until 2020.
- First three years were a period for testing and adjustments.

All Launch Site Names

- According to the data, there are four launch sites.

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

- They were obtained by selecting unique occurrences of launch sites.

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

- Used a condition to specify the structure of launch site string and to display only top 5 records

Total Payload Mass

- Total payload carried by boosters from NASA (CRS)

TOTAL_PAYLOAD_MASS
45596.0

- Used the aggregate function SUM to calculate the total payload of records where customer is “NASA (CRS)”

Average Payload Mass by F9 v1.1

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

AVERAGE_PAYLOAD_MASS
2928.4

- Used the aggregate function AVG to calculate the average payload mass for the records where booster version is “F9 v1.1”

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

FIRST_SUCCESSFUL_LANDING
01/08/2018

- Used the aggregate function MIN to select the record with the oldest date where the landing outcome was a “Success (ground pad)”

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- Used a WHERE and BETWEEN clause to specify the above-mentioned condition to get the result.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

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

- This query was obtained after grouping mission outcomes and counting the records within each group.

Boosters Carried Maximum Payload

- List 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

- Used a subquery with MAX payload mass to get the records.

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

- Used SUBSTR function to collect the month and constrained the output for the year 2015 and where landing outcome is a “Failure (drone ship)”

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

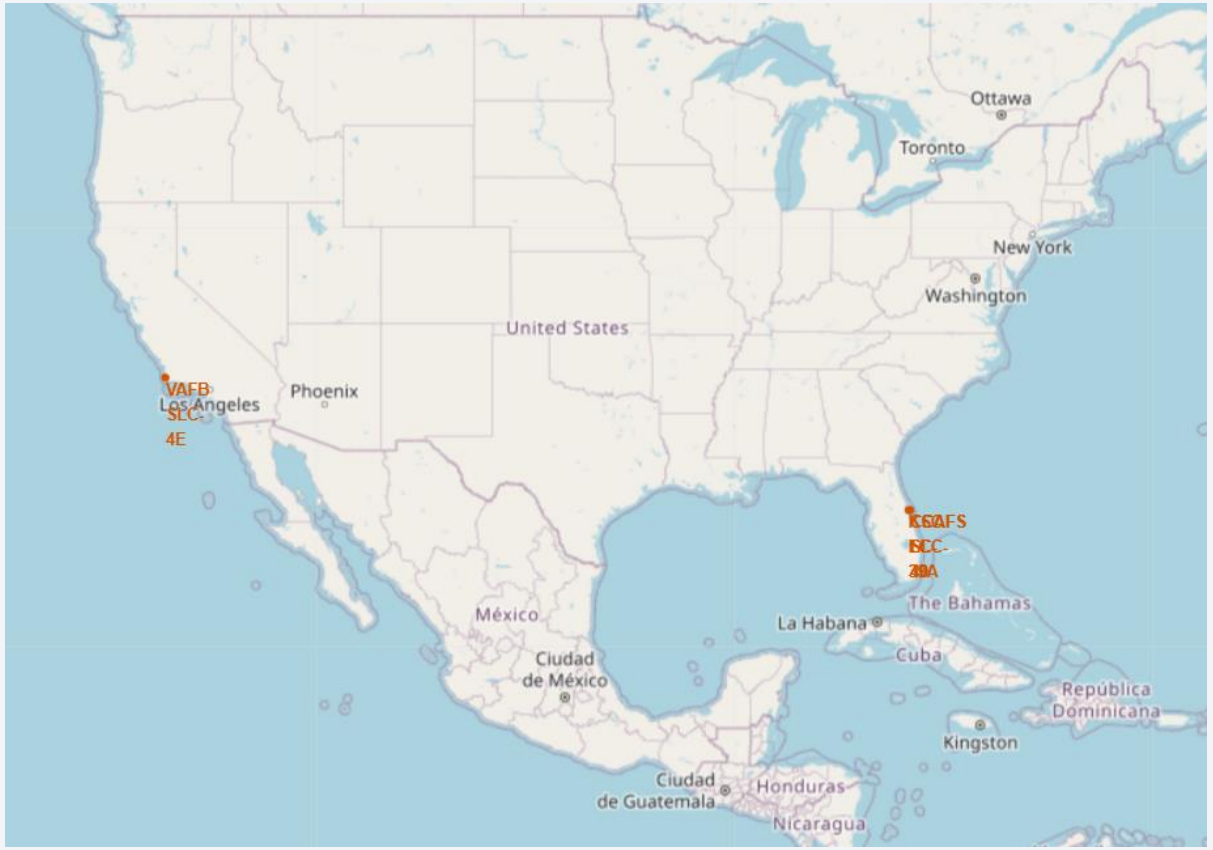
Landing_Outcome	RANK
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

- Grouped the landing outcome and counted the rows of the groups to calculate rank.

Section 3

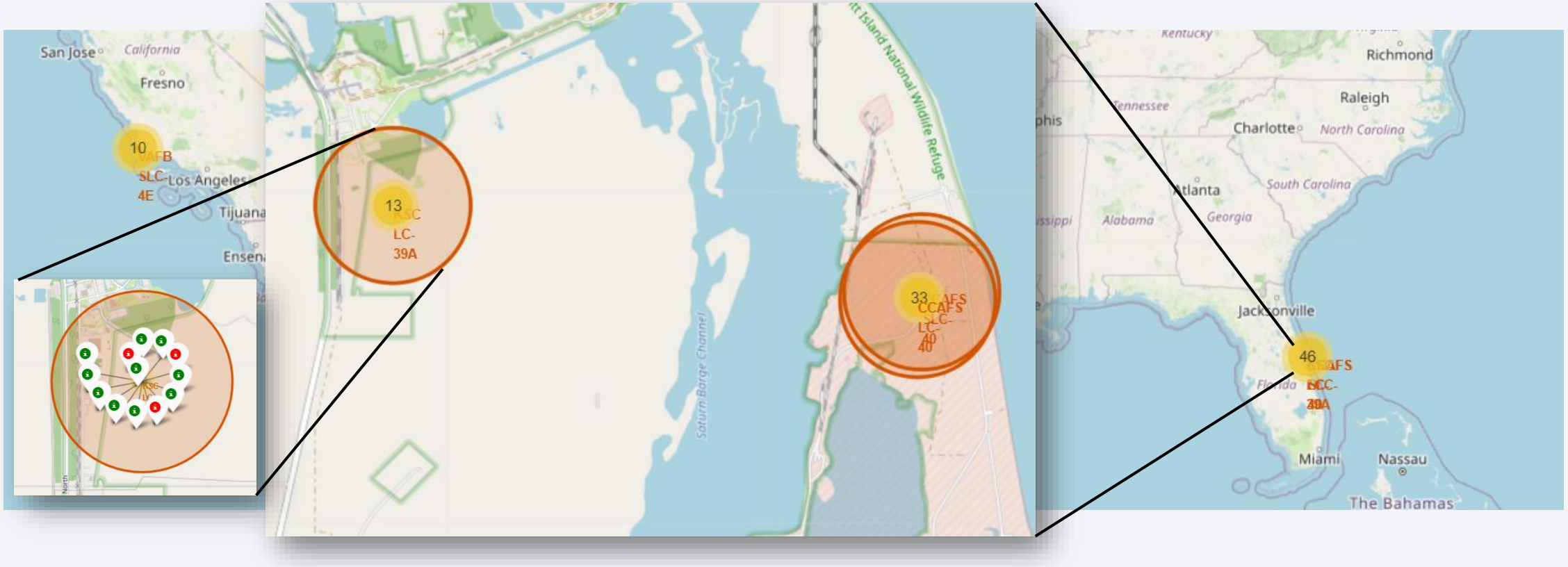
Launch Sites Proximities Analysis

All Launch Sites



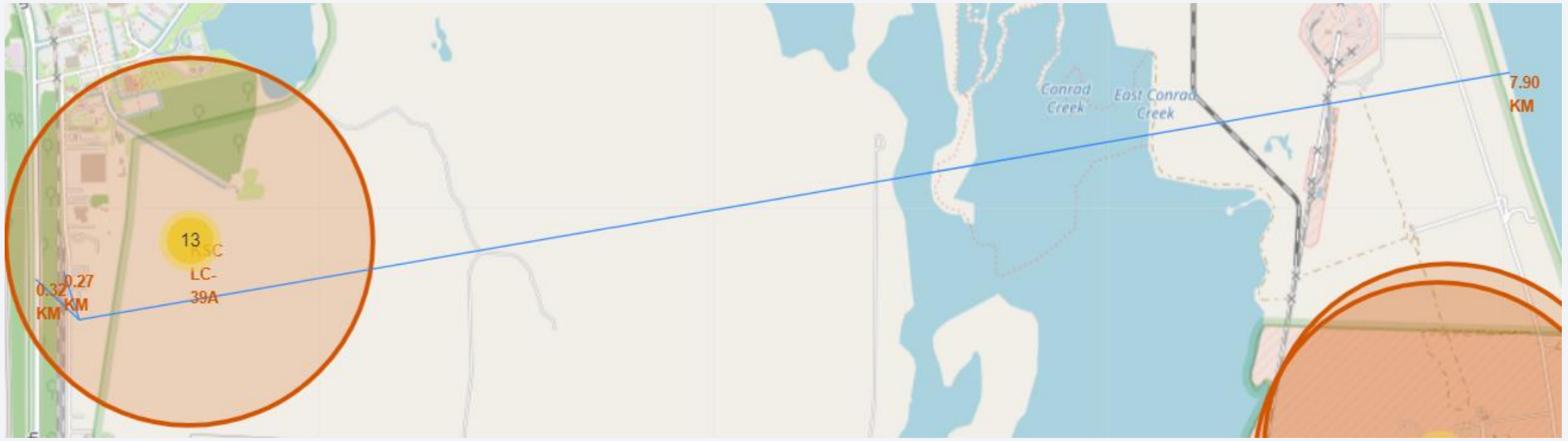
All launch sites are near the coast, and relevant logistic infrastructure. Far away from cities, because of safety concerns.

Launch Outcomes by Site

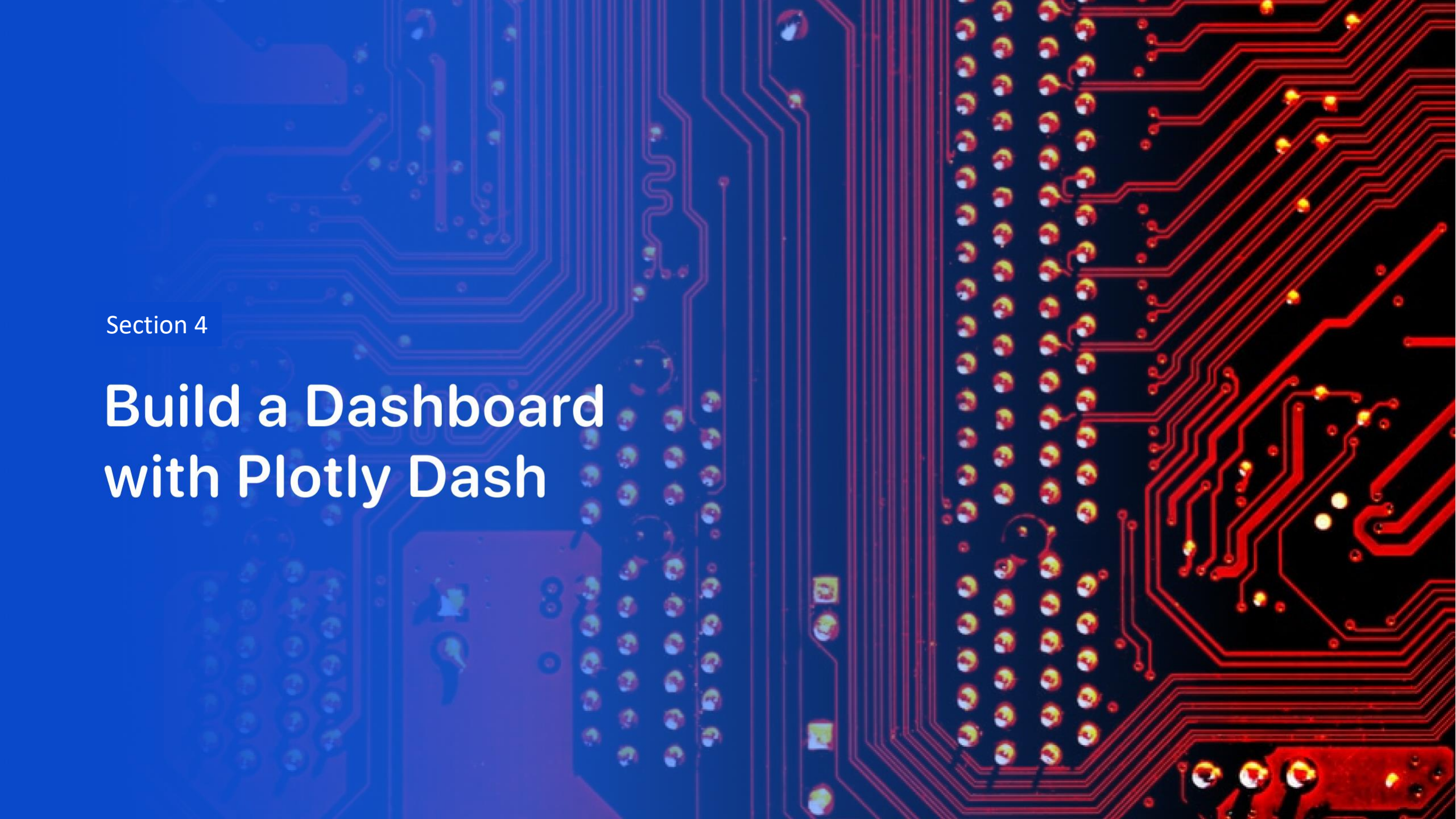


- Green Markers indicate successful landings and red indicates Markers indicate failed landings.

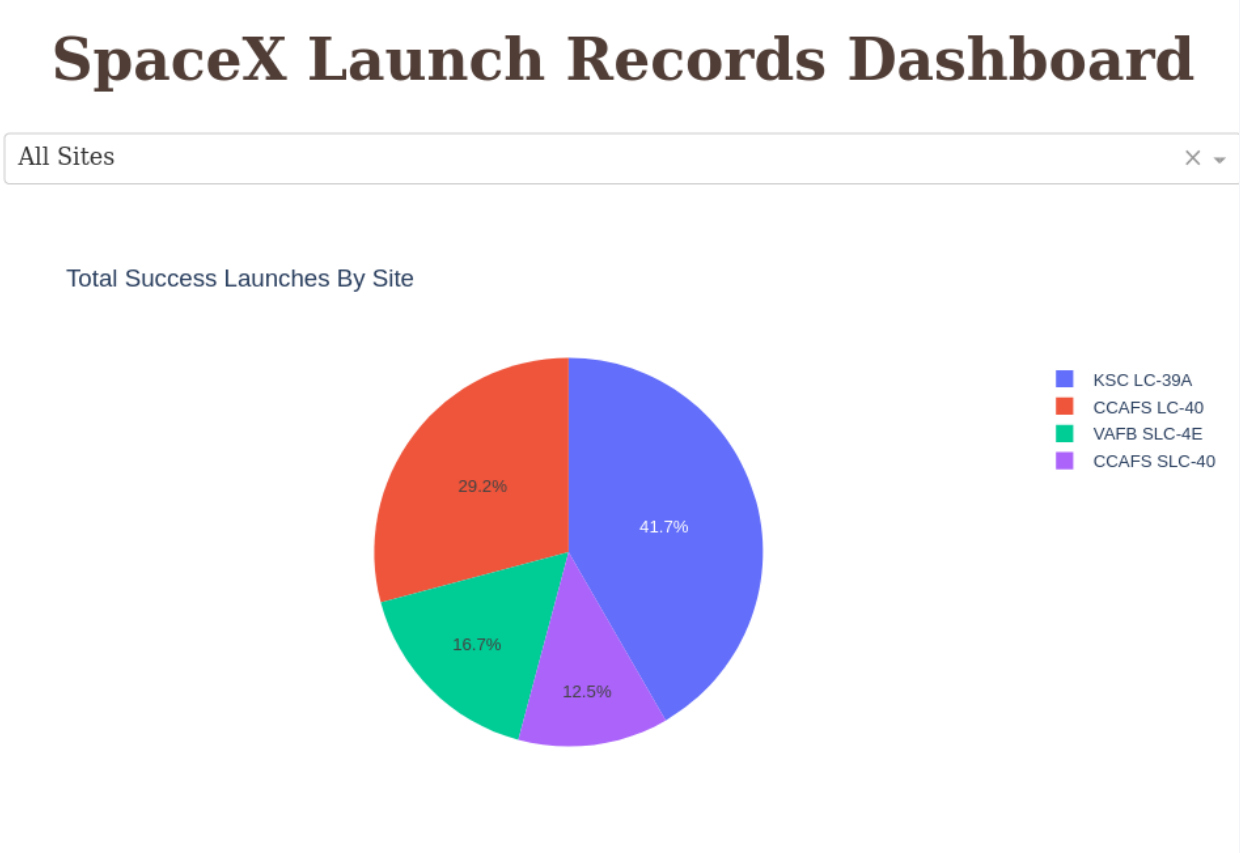
Logistics and Safety



- Launch site KSC LC 39A is 7.9km away from coastline, 0.27km away from railroad, and 0.32km away from highway and relatively far away from residential cities.

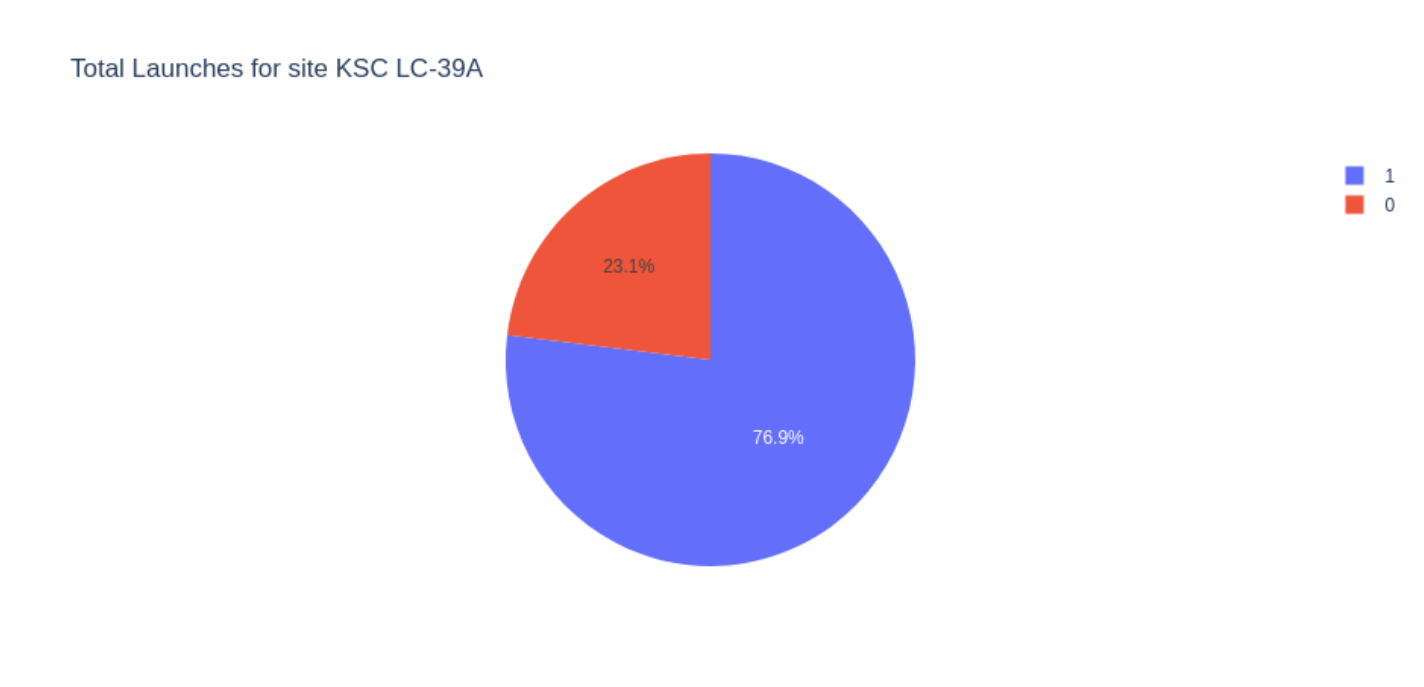

Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Site



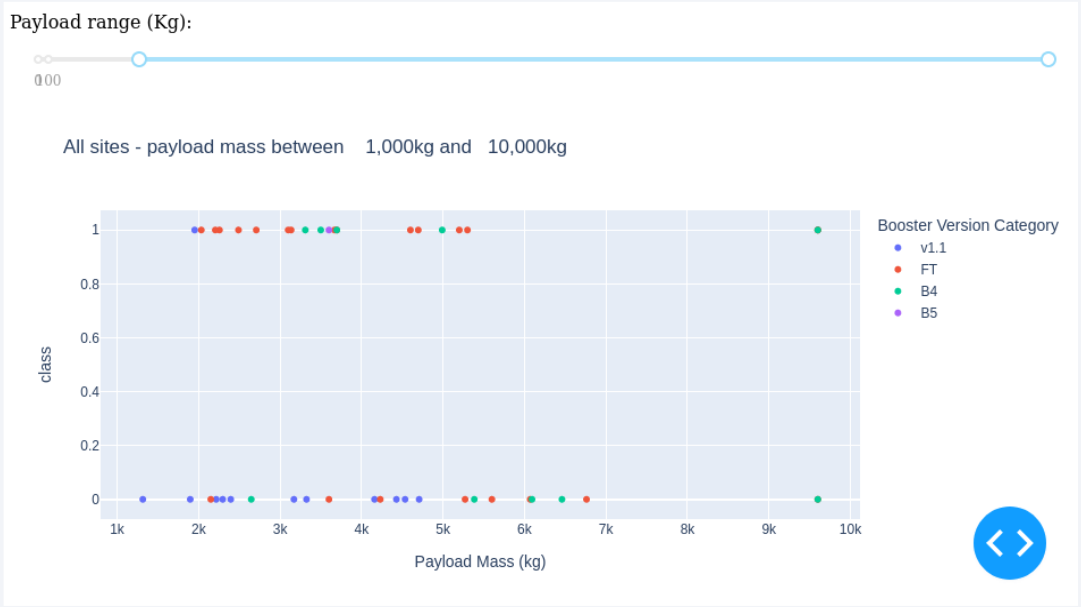
- Selection of launch sites seems to be an important factor for success.

Launch Success Ratio for KSC LC 39A



- 76.9% of launches are successful in this site.

Payload vs. Launch Outcome for All Sites



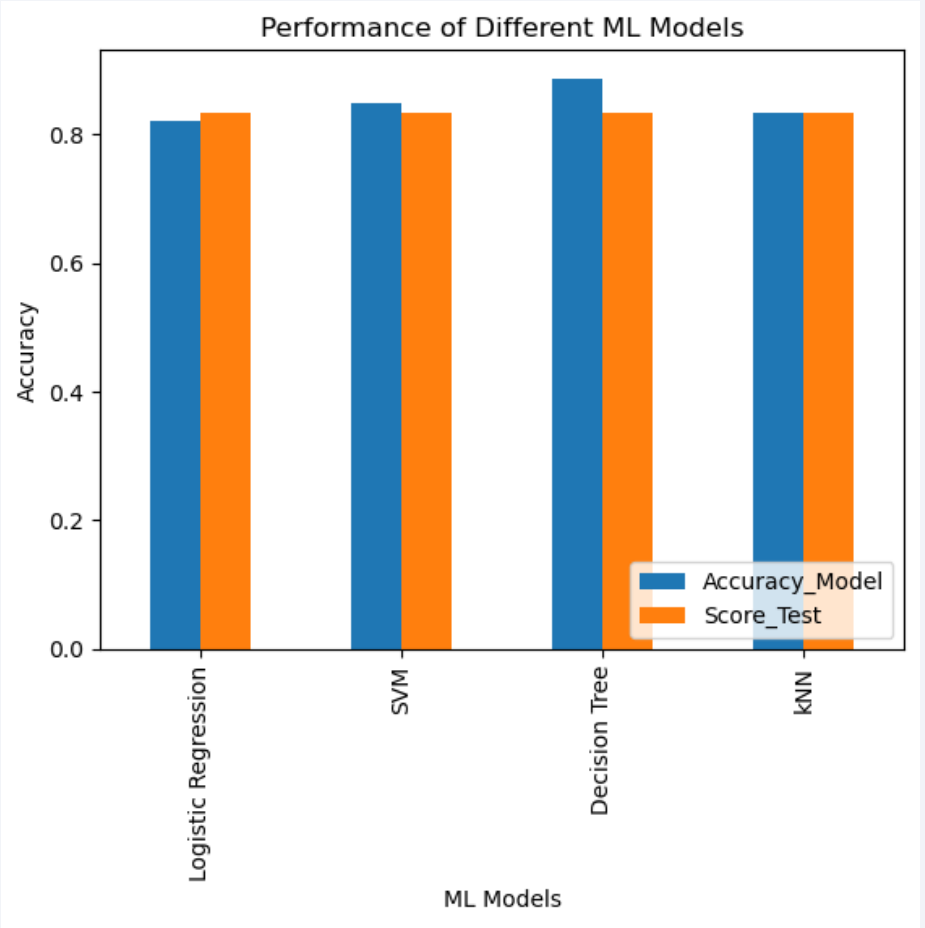
- Payloads under 6,000 kg and FT boosters are the most successful combination.
- There's not enough data to estimate risk of launches over 7,000 kg

Section 5

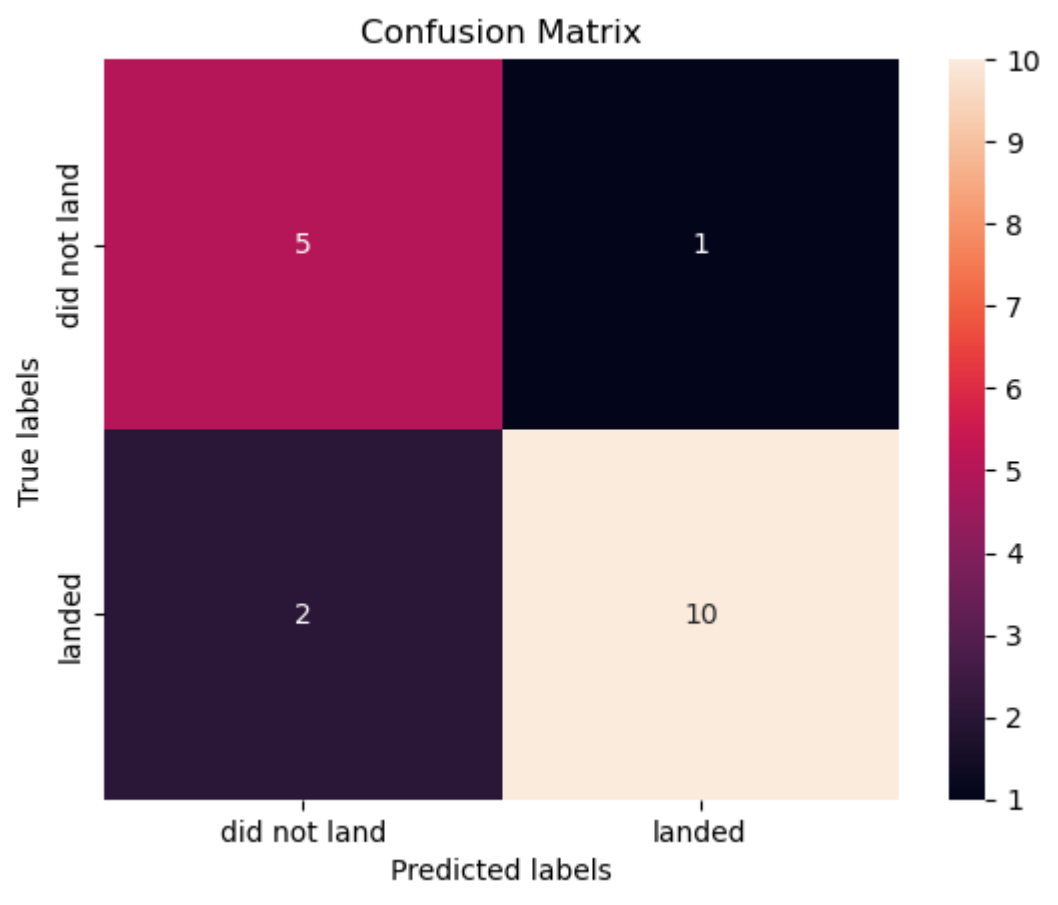
Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were trained, and their accuracies were plotted.
- Decision Tree model had the highest accuracy of 88.75%



Confusion Matrix of Decision Tree Classifier



- Confusion Matrix of Decision Tree Classifier proves the accuracy of the model is high by showing high values for True Positives and True negatives.

Conclusions

- Data were procured from different sources. They were cleaned and scaled to provide statistically accurate results.
- The best launch site is KSC LC 39A with 77.27% landing success rate.
- Launches above 7,000 kg are less risky.
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets.
- VLEO missions seem to be a business growth opportunity, having both high success rate and increased customers in the recent period.
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Thank you!