IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Deva Srirama Sai Ganesh Bandaru

07/04/2024

https://github.com/deva-04

Outline

Executive Summary

❖ Data was collected from the public SpaceX API and SpaceX Wikipedia page, and a 'class' column was added to label successful landings.

❖ The data was explored using SQL queries, visualizations, Folium maps, and dashboards. Relevant columns were selected as features, categorical variables were converted to binary using one-hot encoding, and the data was standardized.

❖ GridSearchCV was used to find the best parameters for four machine learning models: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors.

❖ The accuracy scores of all models were visualized, showing that they all achieved approximately 83.33% accuracy.

❖ However, they tended to over-predict successful landings, indicating a need for more data to improve model accuracy and reliability.

Introduction

SpaceX Falcon 9 Rocket – The Verge

Background

• The era of commercial space exploration is currently thriving, with SpaceX leading the market with its competitive pricing at $62 million compared to the industry average of $165 million USD.
• SpaceX's cost advantage is primarily attributed to its capability to recover and reuse the first stage of the rocket.
•In response, Space Y aims to enter the market and compete directly with SpaceX.

Problem Statement

• Space Y has requested our assistance in training a machine learning model to forecast the successful recovery of Stage 1 rockets.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from public sources such as the SpaceX Wikipedia page.
- Perform data wrangling
  - Addressing missing values, converting categorical variables into binary format using one-hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split data into train/test sets, train the model and evaluate for performance on test data. Tune hyperparameters using GridSearchCV.

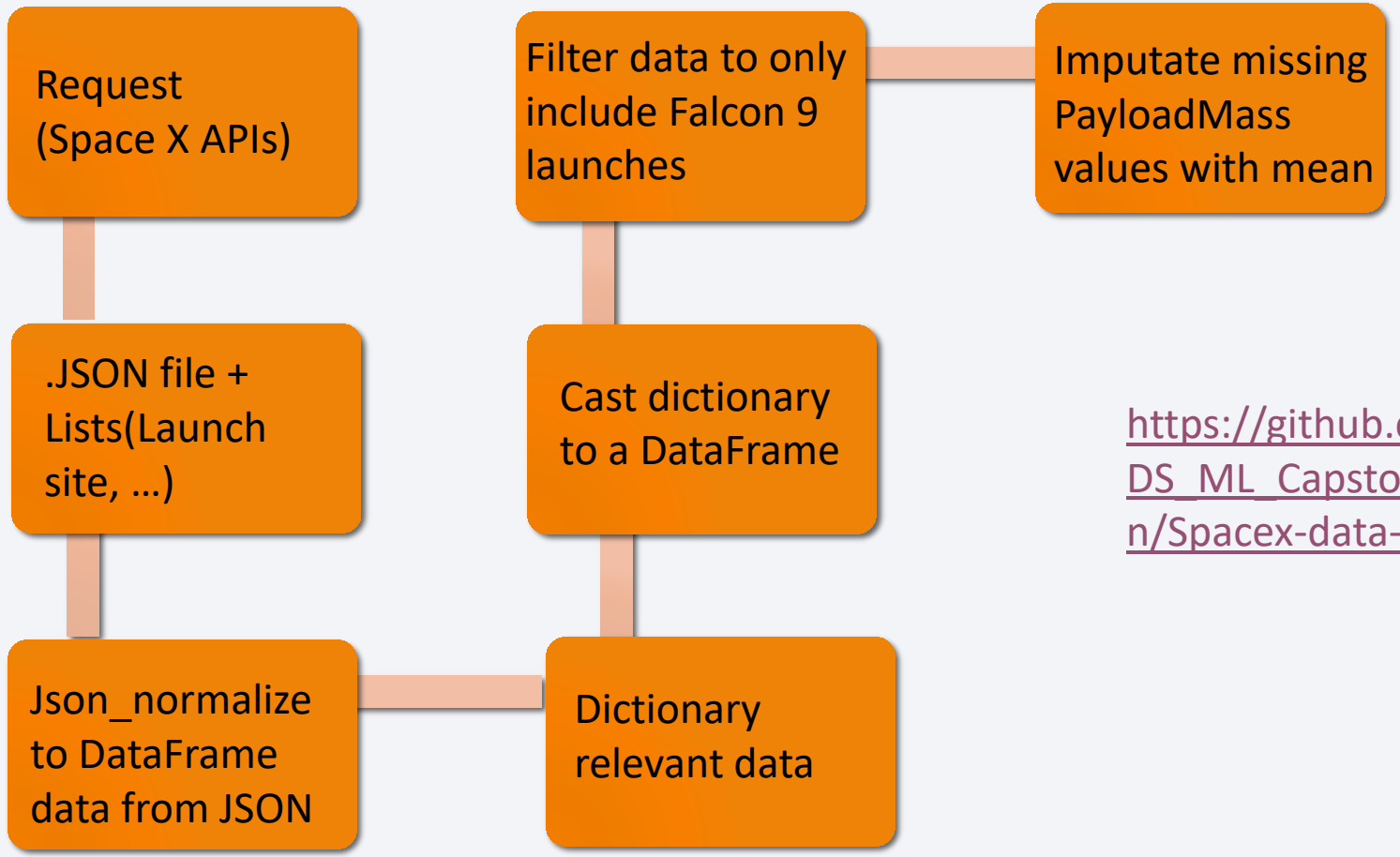
Data Collection

✓ The data collection process involved obtaining data from two sources: Space X's public API through API requests and scraping data from a specific table in Space X's Wikipedia page.

✓ The Space X API provided columns such as FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.

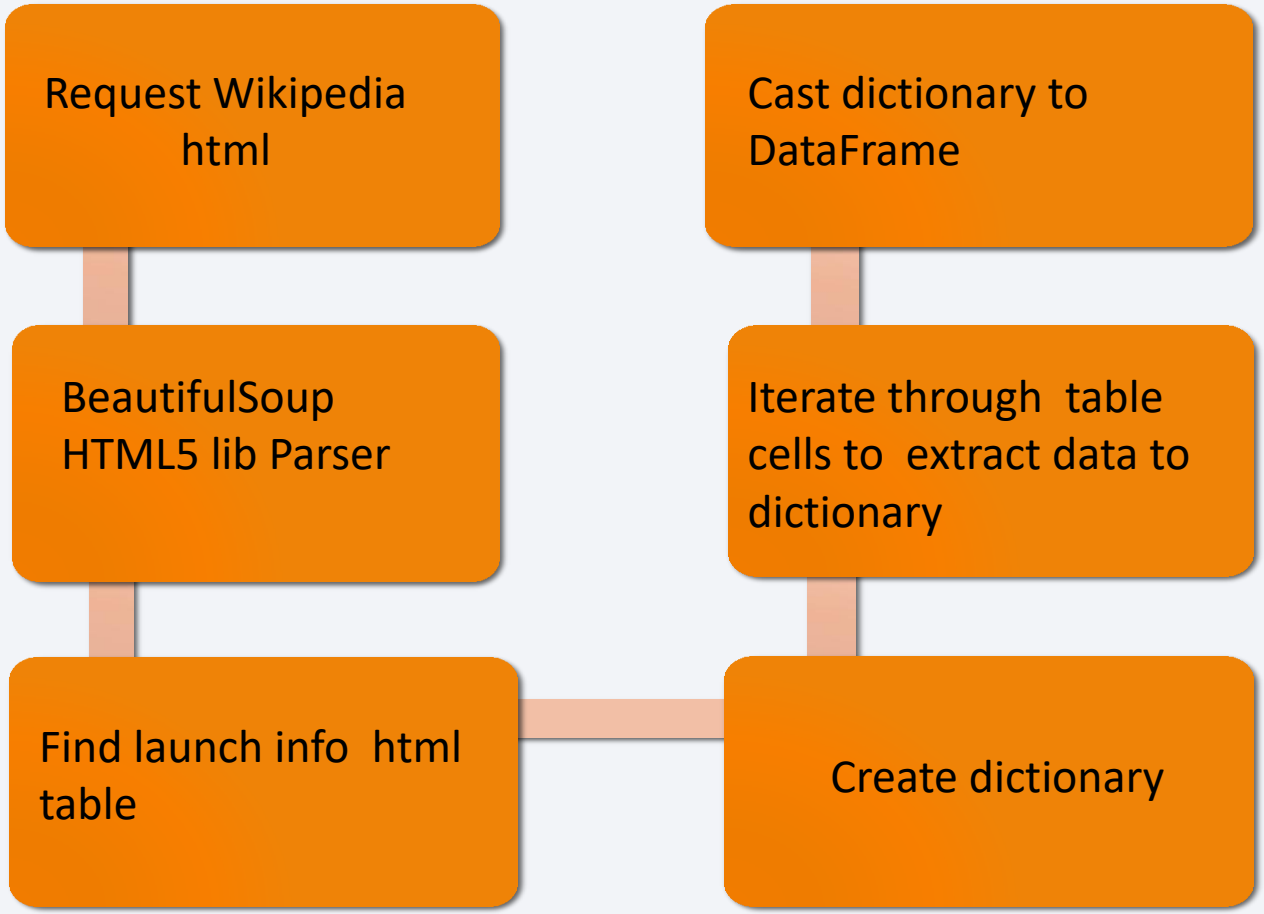
✓On the other hand, the Wikipedia web scraping yielded columns like Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.

Data Collection – SpaceX API



https://github.com/deva-04/IBM-DS_ML_Capstone_Project/blob/main/Spacex-data-collection-api.ipynb

Data Collection - Scraping



https://github.com/deva-04/IBM-DS_ML_Capstone_Project/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Generate a training label based on landing outcomes, assigning a value of 1 for successful outcomes (True) and 0 for failures (False). The 'Outcome' column consists of 'Mission Outcome' and 'Landing Location' components. Create a new column named 'class' in the dataset where 'class' is set to 1 if 'Mission Outcome' is True and 0 otherwise. For mapping:
- True outcomes with ASDS, RTLS, or Ocean landings are mapped to 1.
- False outcomes with no landing or unsuccessful ASDS, Ocean, or RTLS landings are mapped to 0.

https://github.com/deva-04/IBM-DS_ML_Capstone_Project/blob/main/spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Plots included Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs. Orbit, and Success Yearly Trend.
- Various chart types such as scatter plots, line charts, and bar plots were utilized to analyze relationships between variables like Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year. These visualizations were instrumental in determining the presence of any meaningful relationships among variables, aiding in their selection for machine learning model training.

https://github.com/deva-04/IBM-DS_ML_Capstone_Project/blob/main/eda-dataviz.ipynb

EDA with SQL

- Loaded dataset into IBM DB2 Database for storage and analysis.
- Utilized SQL Python integration to perform queries on the dataset.
- Executed queries to gain insights into launch site names, mission outcomes, payload sizes of customers, booster versions, and landing outcomes.
- Extracted information about various aspects of the dataset to enhance understanding and inform further analysis.

https://github.com/deva-04/IBM-DS_ML_Capstone_Project/blob/main/SpaceX_Falcon9_EDA_SQL.ipynb

Build an Interactive Map with Folium

Map objects added to Folium Map:

- Markers for Launch Sites
- Markers for successful and unsuccessful landings
- Circles to depict proximity to key locations such as Railway, Highway, Coast, and City

This visualization strategy helps in understanding why launch sites may be located where they are, and it also visualizes successful landings relative to location.

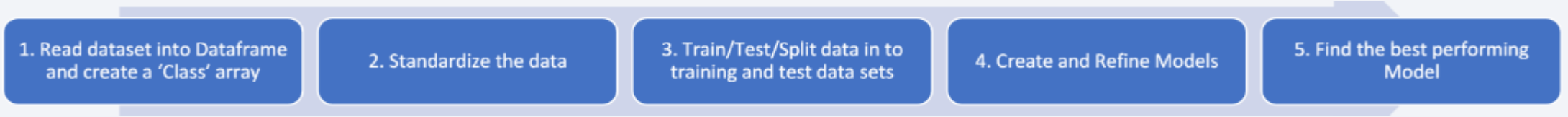
https://github.com/deva-04/IBM-DS_ML_Capstone_Project/blob/main/SpaceX_Interactive_Viz_Folium.ipynb

Build a Dashboard with Plotly Dash

- The dashboard comprises a pie chart and a scatter plot.
- The pie chart allows users to toggle between displaying the distribution of successful landings across all launch sites and showing the success rates of individual launch sites.
- The scatter plot offers flexibility by allowing users to choose between viewing data for all launch sites or a specific site, along with adjusting payload mass using a slider ranging from 0 to 10000 kg.
- The pie chart is utilized to illustrate launch site success rates, while the scatter plot aids in understanding the variability of success rates across launch sites, payload masses, and booster version categories.

https://github.com/deva-04/IBM-DS_ML_Capstone_Project/blob/main/SpaceX_Plotly_Dashboard.py

Predictive Analysis (Classification)



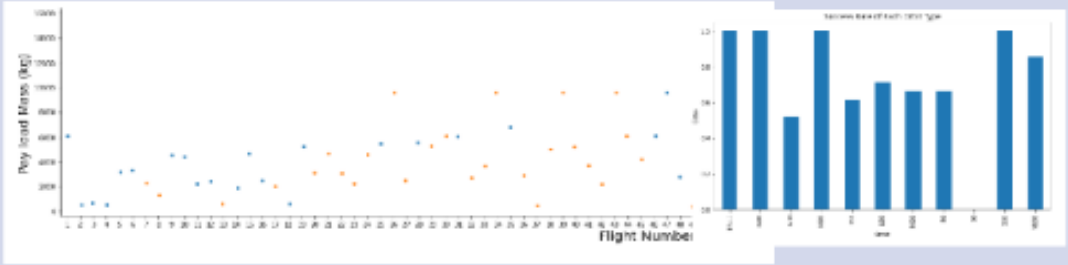
- **Data Splitting**: Split data into training and testing sets.
- **Model Training**: Trained multiple classification models (Logistic Regression, SVM, Decision Tree, KNN) on training data.
- **Model Evaluation**: Evaluated models using accuracy scores on the test set.
- **Hyperparameter Tuning**: Used GridSearchCV to find optimal hyperparameters for each model.
- **Model Selection**: Selected the best performing classification model based on evaluation metrics.
- **Model Deployment**: Deployed the selected model for predictions on new data.

https://github.com/deva-04/IBM-DS_ML_Capstone_Project/blob/main/%20SpaceX_PredictiveAnalysis.ipynb

Results

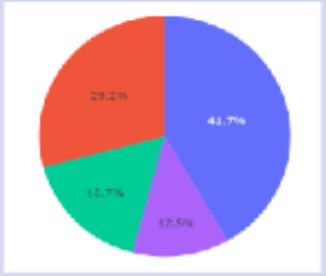
Exploratory data analysis results

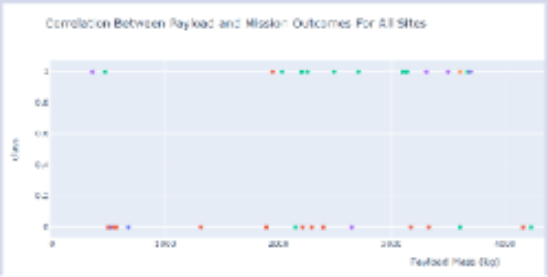
- Samples:



Interactive analytics demo in screenshots

- Samples





Predictive analysis results

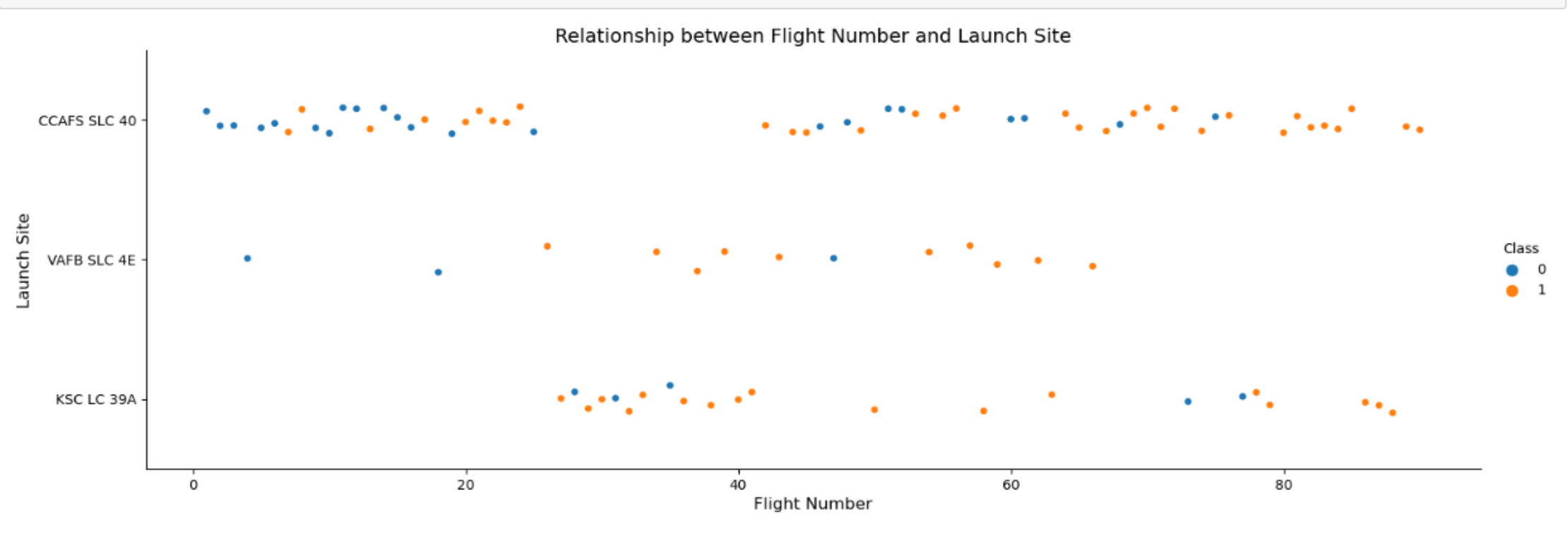
- Samples

	Algo Type	Accuracy Score
2	Decision Tree	0.903571
3	KNN	0.848214
1	SVM	0.848214
0	Logistic Regression	0.846429

Section 2

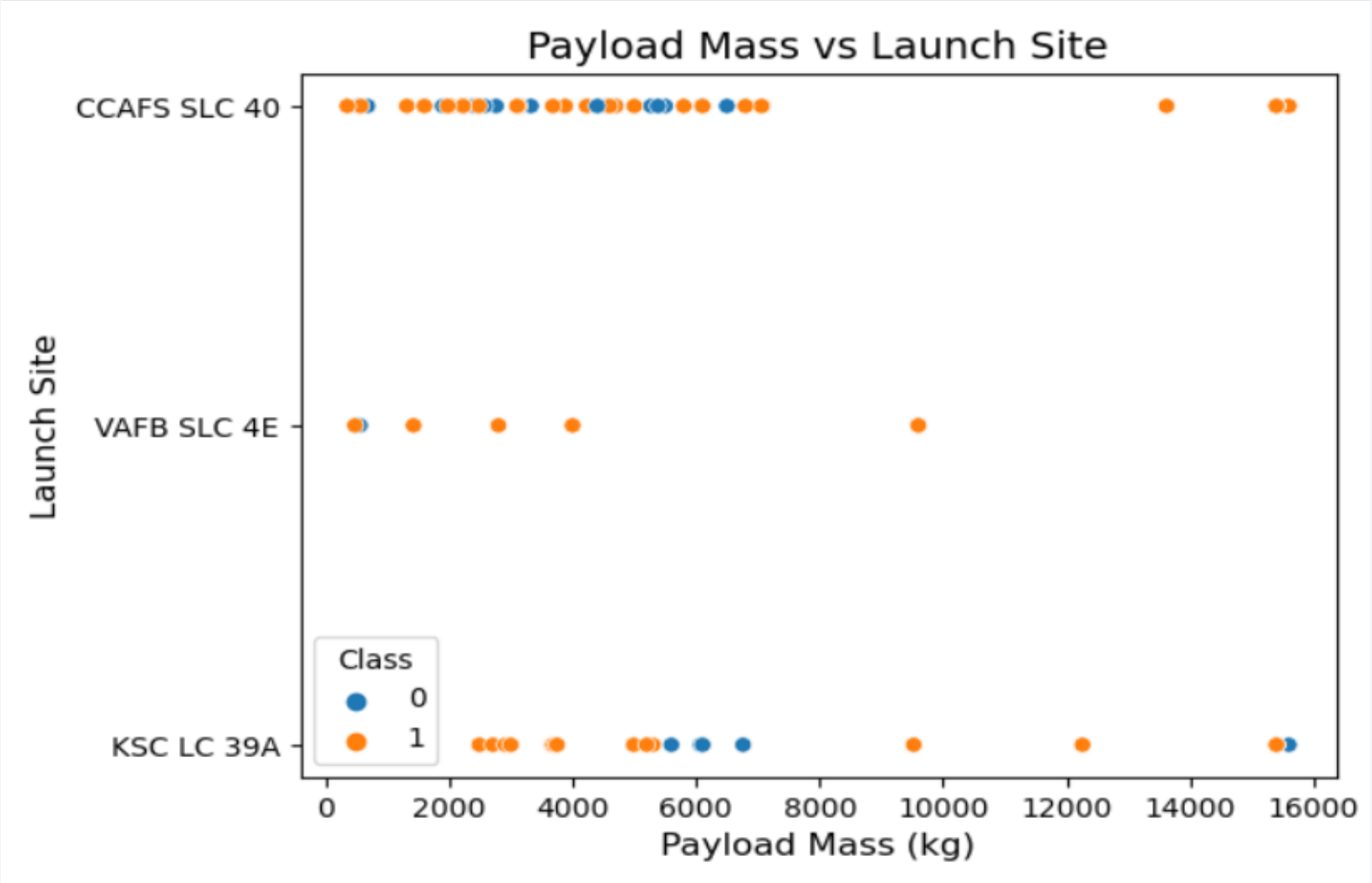
Insights drawn from EDA

Flight Number vs. Launch Site



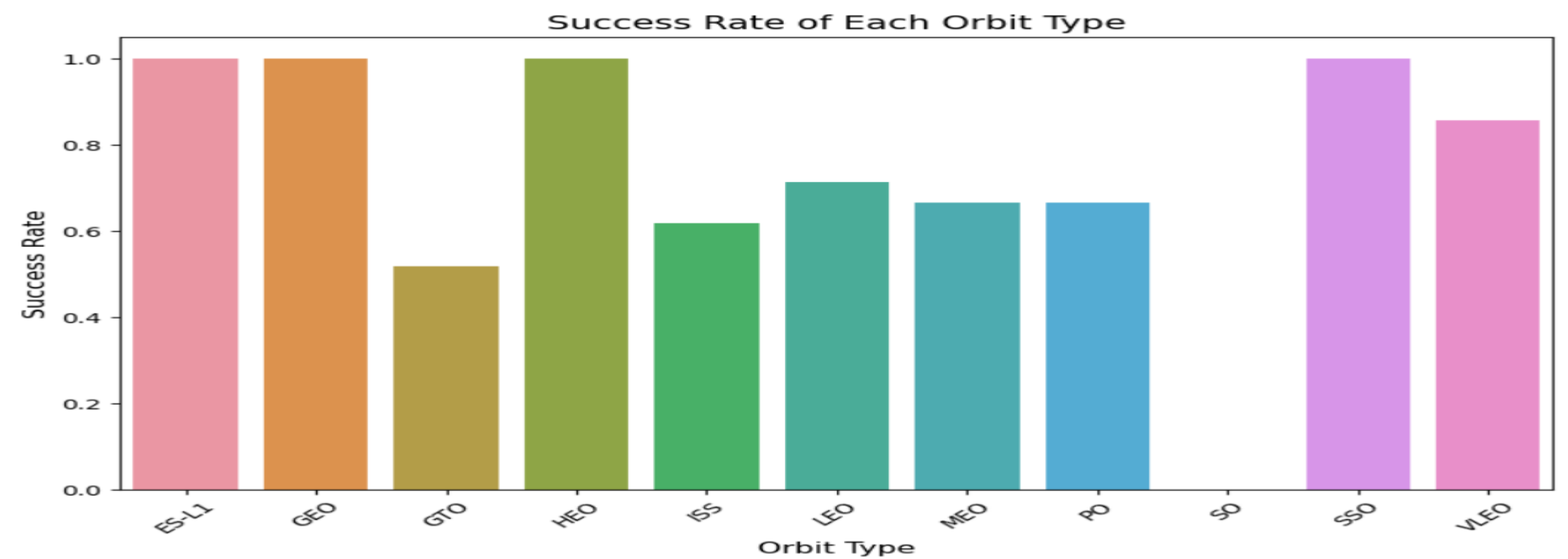
- The data shows increasing success rates over time, especially around flight number 20, indicating a significant breakthrough.
- CCAFS appears to be the main launch site due to its higher launch volume.

Payload vs. Launch Site



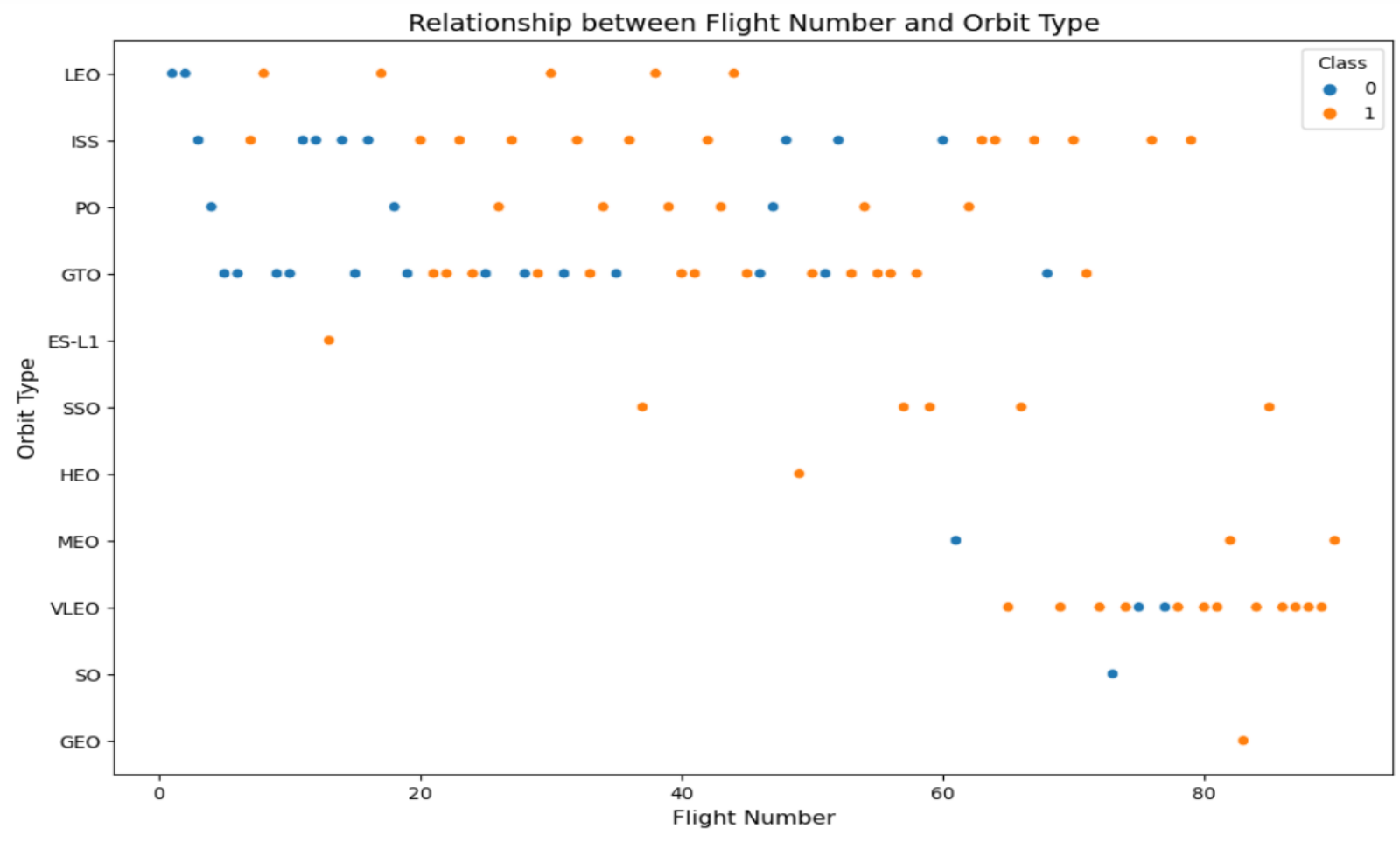
Payload mass is predominantly within the range of 0-6000 kg, and various launch sites exhibit distinct preferences for different payload masses.

Success Rate vs. Orbit Type



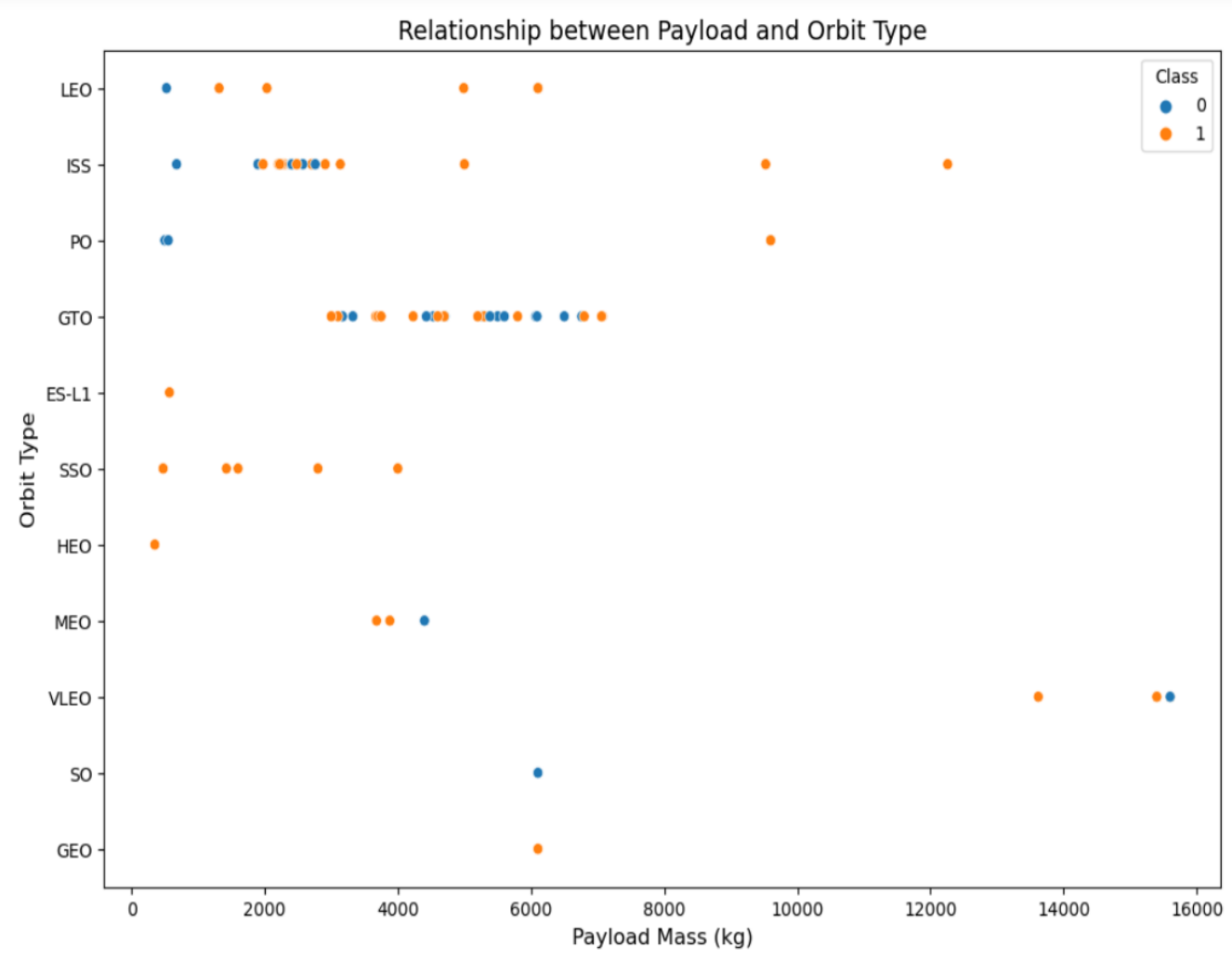
- ES-L1, GEO, HEO, and SSO had a 100% success rate, with varying sample sizes.
- VLEO had decent success rates across multiple attempts, while GTO showed a 50% success rate with the largest sample size. SO had a 0% success rate.

Flight Number vs. Orbit Type



- SpaceX's orbit preferences shifted over time with flight number, correlating with launch outcomes.
- Initially focusing on LEO orbits with moderate success, SpaceX later transitioned back to VLEO in recent launches.
- The data suggests that SpaceX performs better in lower orbits or Sun-synchronous orbits.

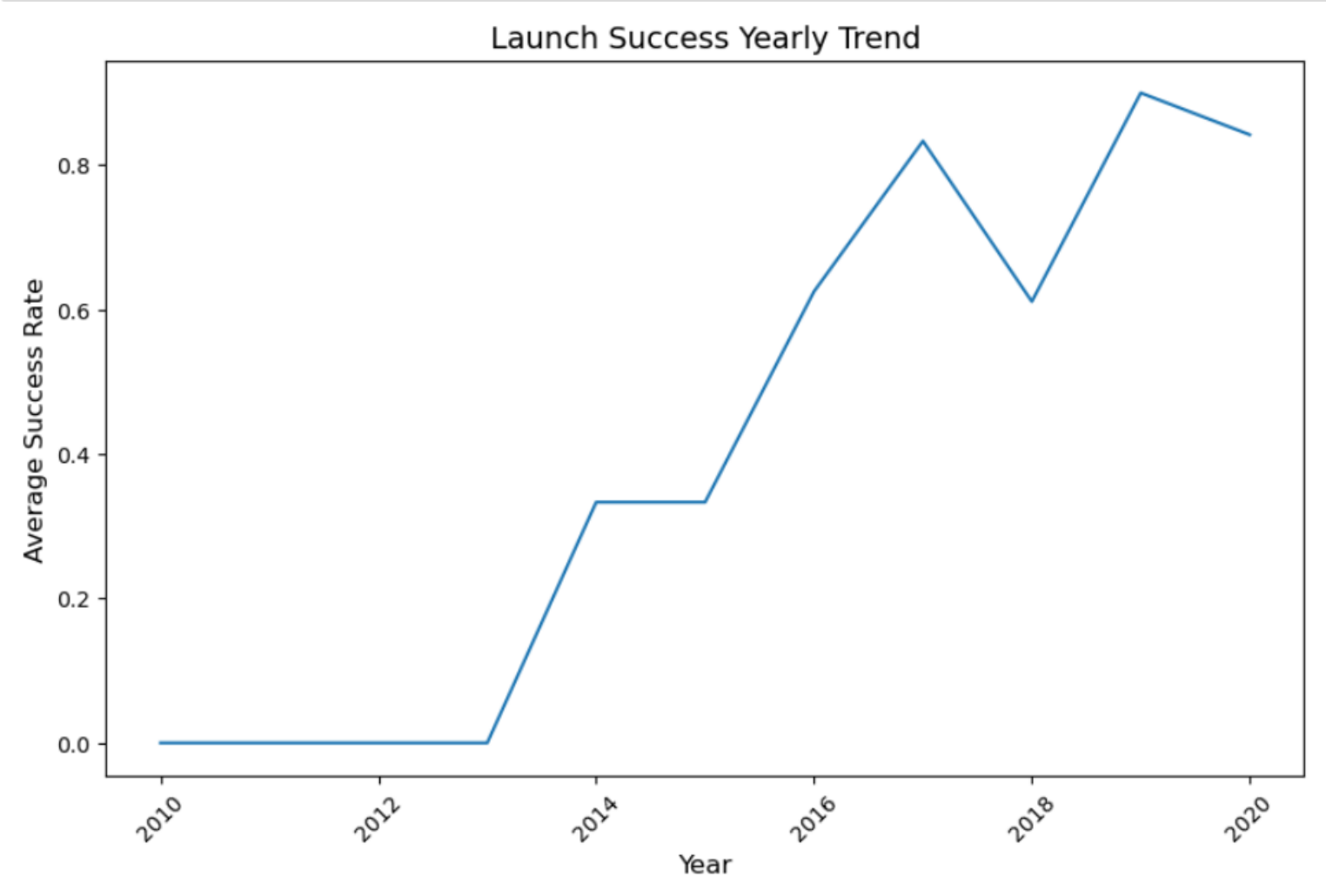
Payload vs. Orbit Type



- There is a correlation between payload mass and orbit type, with LEO and SSO showing lower payload masses.

- Conversely, the successful orbit type VLEO typically has higher payload mass values.

Launch Success Yearly Trend



Success rates have generally risen since 2013, with a minor decline noted in 2018. In recent years, success rates have stabilized around 80%.

All Launch Site Names

- Query:

```
select distinct Launch_Site from spacextbl
```

- Description:
  - 'distinct' returns only unique values from the queries column (Launch_Site)
  - There are 4 unique launch sites

- Result:

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

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

sum_payload_mass_kg
45596

- This query calculates the total payload mass in kilograms for payloads where NASA was the customer.
- The payloads associated with CRS (Commercial Resupply Services) were intended for delivery to the International Space Station (ISS).

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-8
Done.
```

avg_payload_mass_kg
2928

- This query computes the average payload mass for launches utilizing the booster version F9 v1.1.
- The average payload mass for F9 1.1 launches falls towards the lower end of our payload mass range.

First Successful Ground Landing Date

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

first_success
2015-12-22

- This query provides the date of the initial successful ground pad landing, which occurred towards the end of 2015.
- Successful landings, in general, began to occur around 2014.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

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

- This query lists the four booster versions that achieved successful drone ship landings with payloads ranging between 4000 and 6000 kg (excluding the upper limit).

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-
Done.
```

mission_outcome	no_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- This query provides a breakdown of each mission outcome. SpaceX consistently achieves its mission outcome almost 99% of the time.
- This indicates that most landing failures are intentional.
- Interestingly, one launch has an ambiguous payload status, and unfortunately, one launch failed during flight.

Boosters Carried Maximum Payload

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
Done.
```

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

- This query identifies the booster versions that carried the highest payload mass of 15600 kg.
- These booster versions share similarities and are all categorized as F9 B5 B10xx.x.
- This suggests a correlation between payload mass and the specific booster version utilized.

2015 Launch Records

```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

- This query retrieves data on launches from 2015 where the stage 1 failed to land on a drone ship.
- The results include information such as the Month, Landing Outcome, Booster Version, Payload Mass (in kg), and Launch site.
- There were two instances of such occurrences during that year.

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

```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

- This query provides a list of successful landings between June 4, 2010, and March 20, 2017, inclusive.
- Successful landings during this period include both drone ship and ground pad landings, totaling 8 successful landings overall.

Section 3

Launch Sites Proximities Analysis

Launch Sites Map

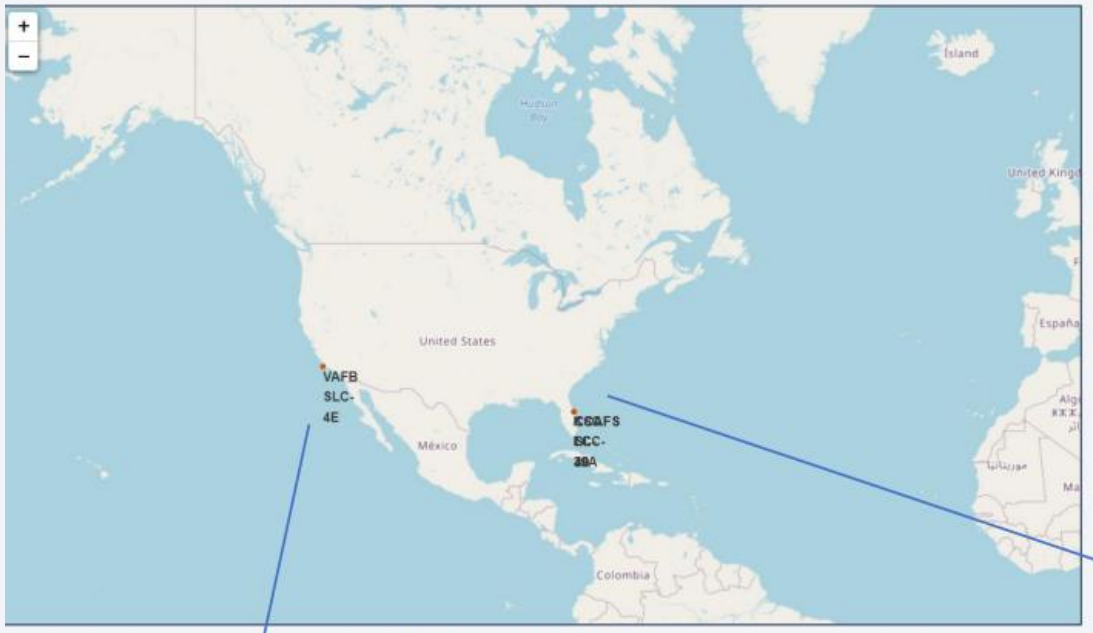


Fig 1 – Global Map

Fig 2 – Zoom 1

Figure 1 on left displays the Global map with Falcon 9 launch sites that are located in the United States (in California and Florida). Each launch site contains a circle, label, and a popup to highlight the location and the name of the launch site. It is also evident that all launch sites are near the coast.

Figure 2 and Figure 3 zoom in to the launch sites to display 4 launch sites:

- VAFB SLC-4E (CA)
- CCAFS LC-40 (FL)
- KSC LC-39A (FL)
- CCAFS SLC-40 (FL)

Fig 3 – Zoom 2

Success/Failed Launch Map for all Launch Sites

Fig 1 – US map with all Launch Sites

- Figure 1 is the US map with all the Launch Sites. The numbers on each site depict the total number of successful and failed launches
- Figure 2, 3, 4, and 5 zoom in to each site and displays the success/fail markers with green as success and red as failed
- By looking at each site map, KSC LC-39A Launch Site has the greatest number of successful launches

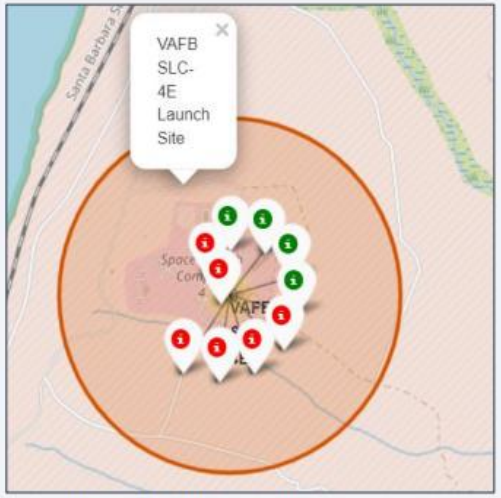


Fig 2 – VAFB Launch Site with success/failed markers

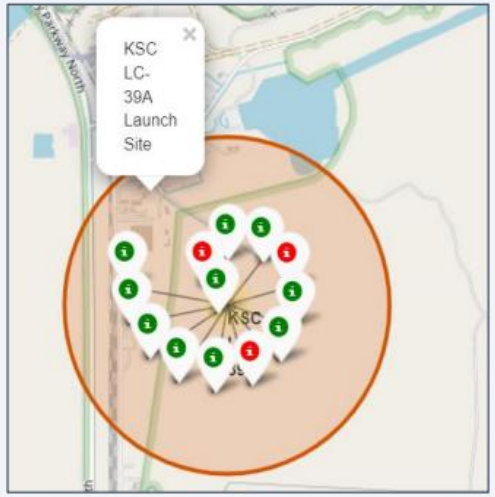


Fig 3 – KSC LC-39A success/failed markers

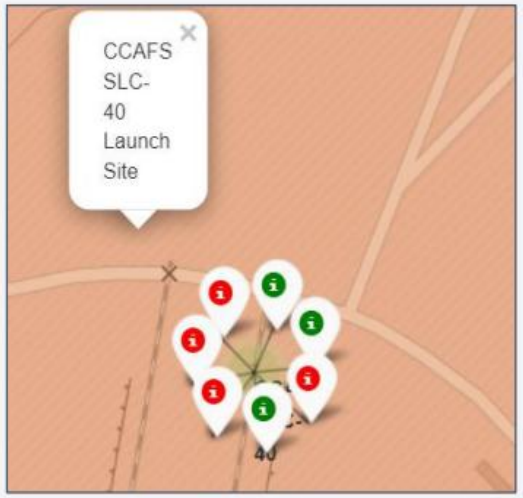


Fig 4 – CCAFS SLC-40 success/failed markers

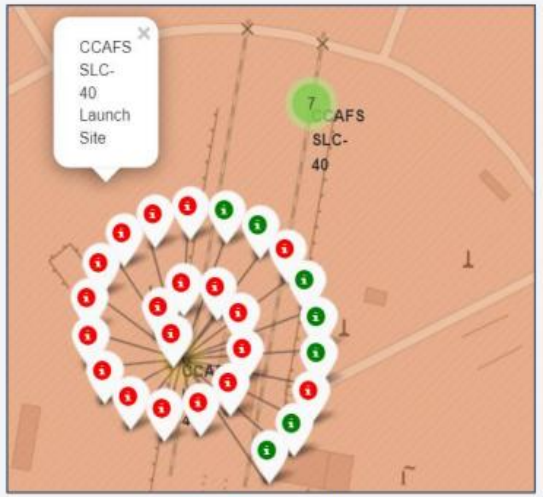


Fig 5 – CCAFS SLC-40 success/failed markers

Launch Site to proximity Distance Map

Fig 1 – Proximity site map for VAFB SLC-4E

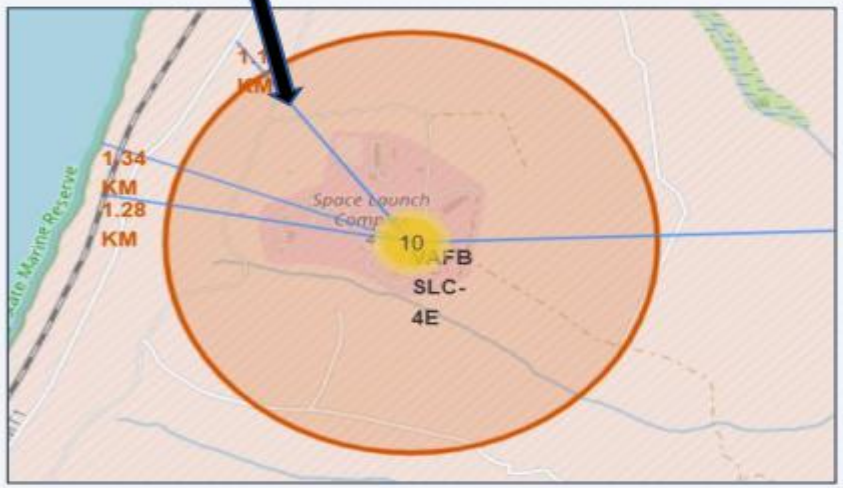
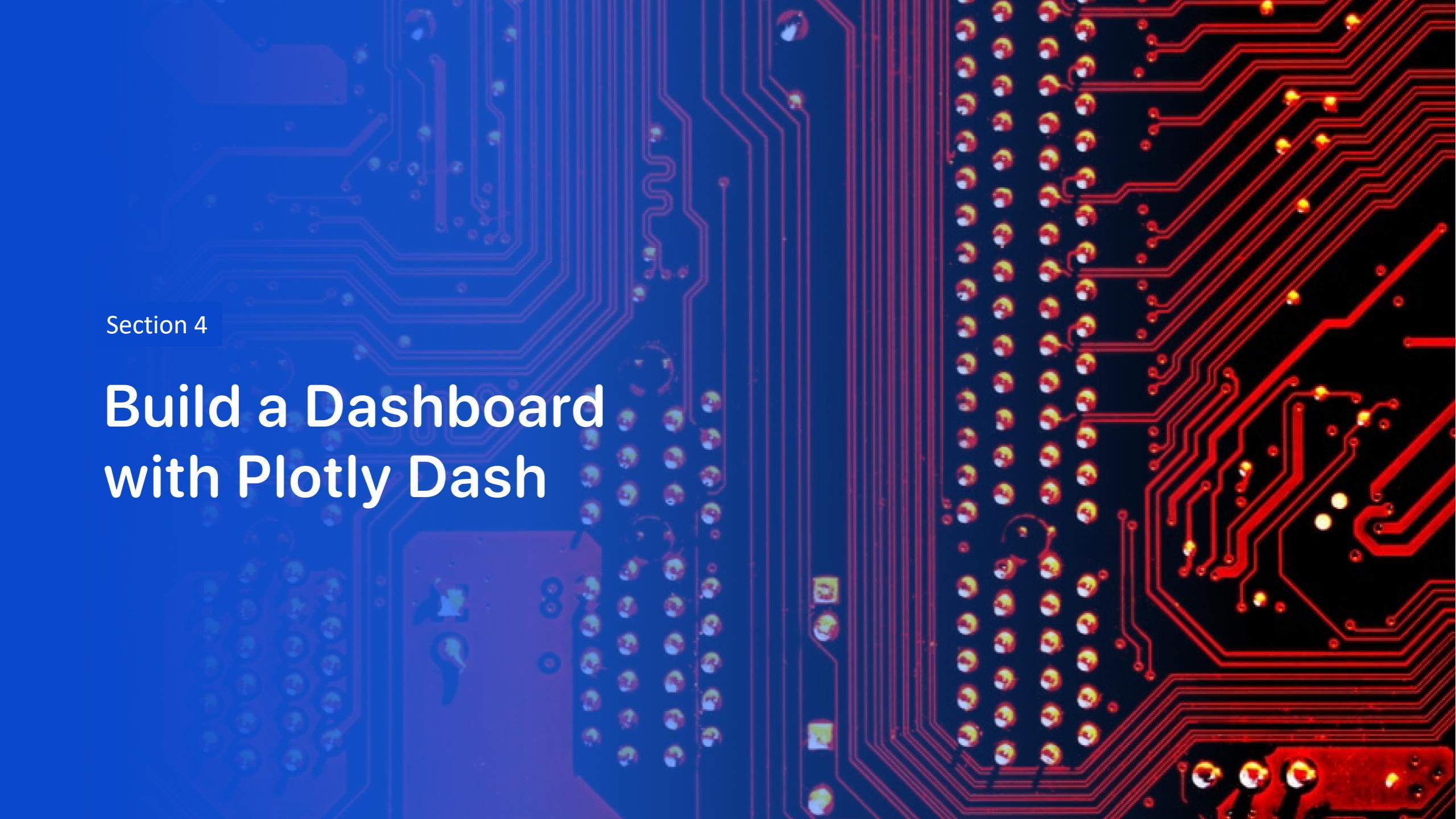


Fig 2 – Zoom in for sites – coastline, railroad, and highway

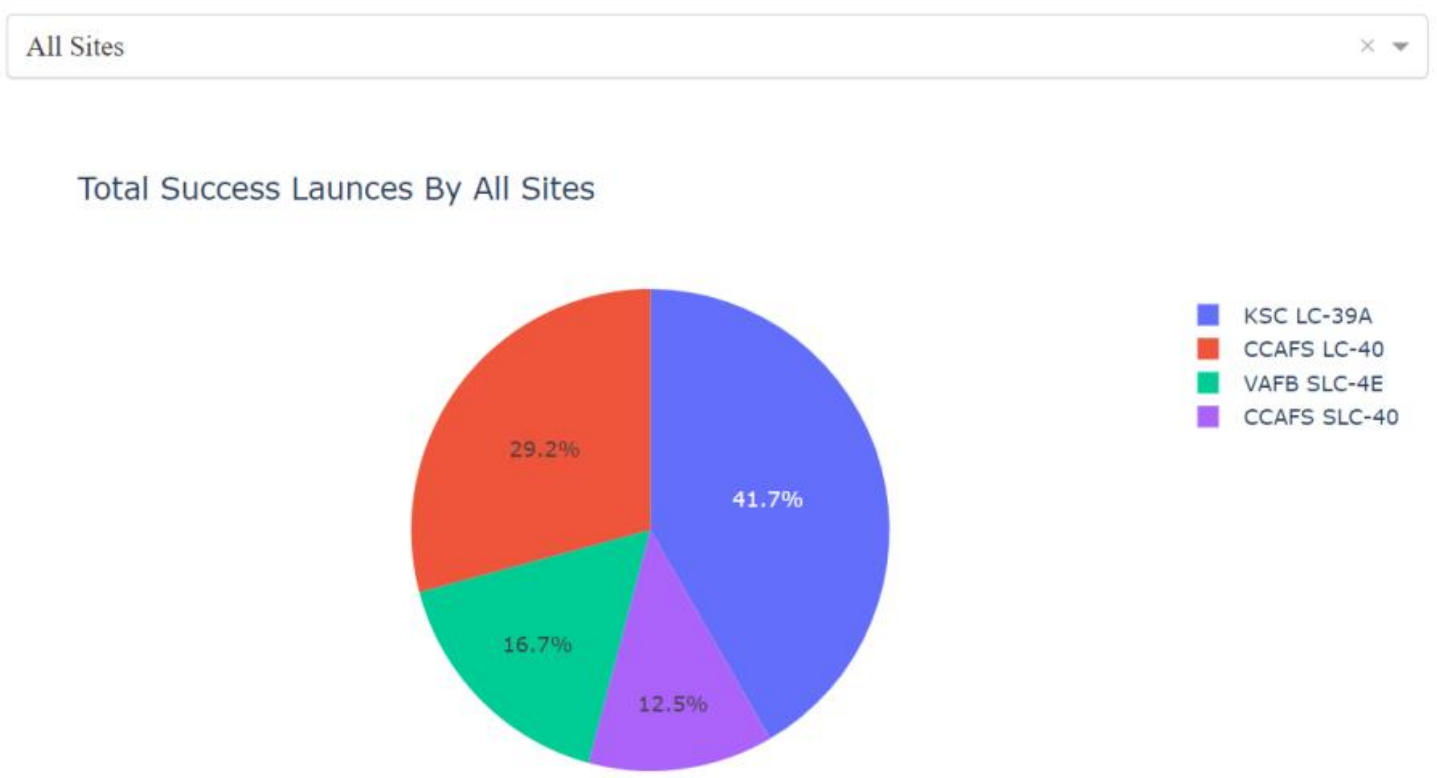
Figure 1 displays all the proximity sites marked on the map for Launch Site VAFB SLC-4E. City Lompoc is located further away from Launch Site compared to other proximities such as coastline, railroad, highway, etc. The map also displays a marker with city distance from the Launch Site (14.09 km)

Figure 2 provides a zoom in view into other proximities such as coastline, railroad, and highway with respective distances from the Launch Site

In general, cities are located away from the Launch Sites to minimize impacts of any accidental impacts to the general public and infrastructure. Launch Sites are strategically located near the coastline, railroad, and highways to provide easy access to resources.

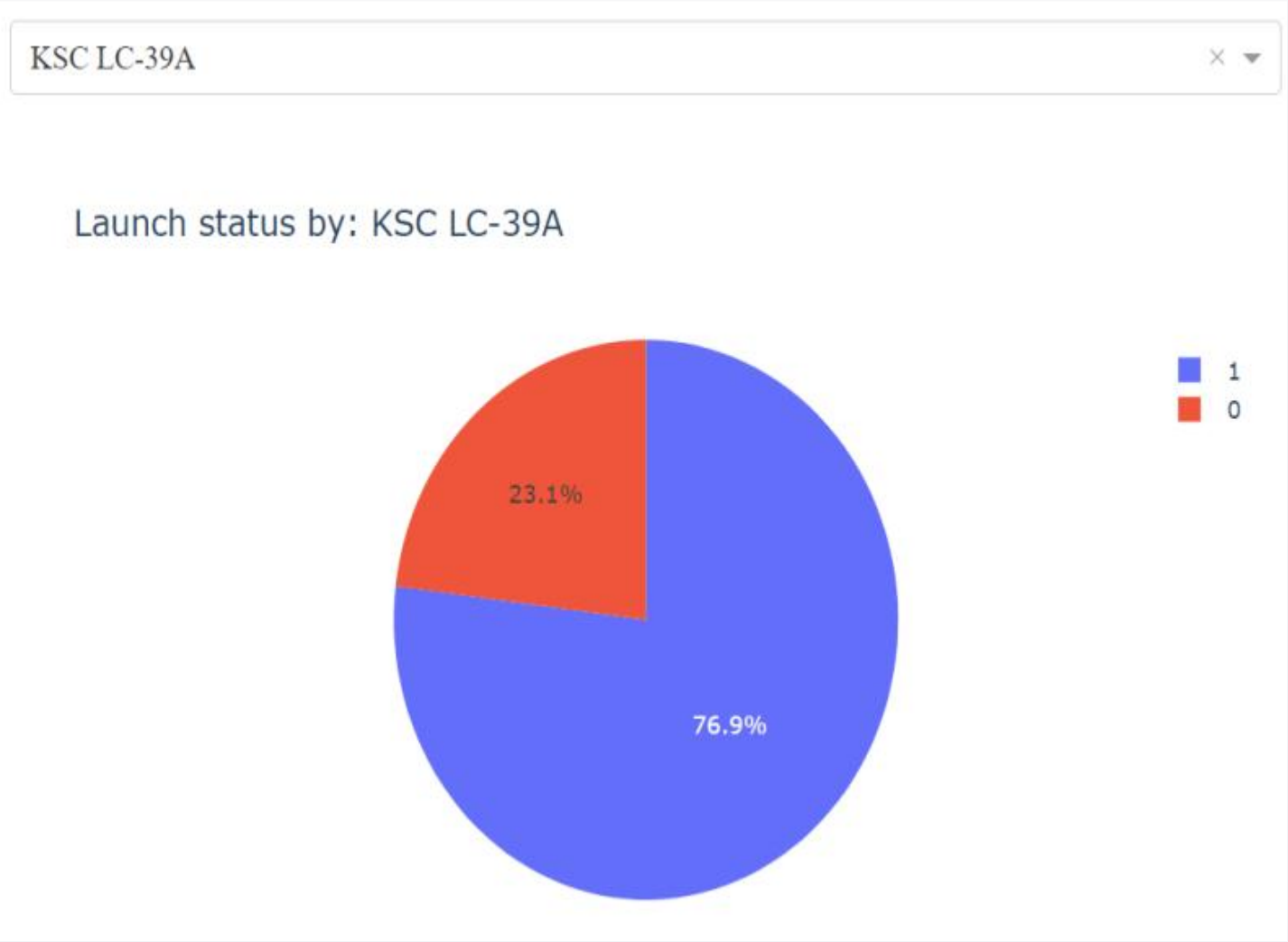

Section 4
Build a Dashboard with Plotly Dash

Successfull Launches Across Launch Sites



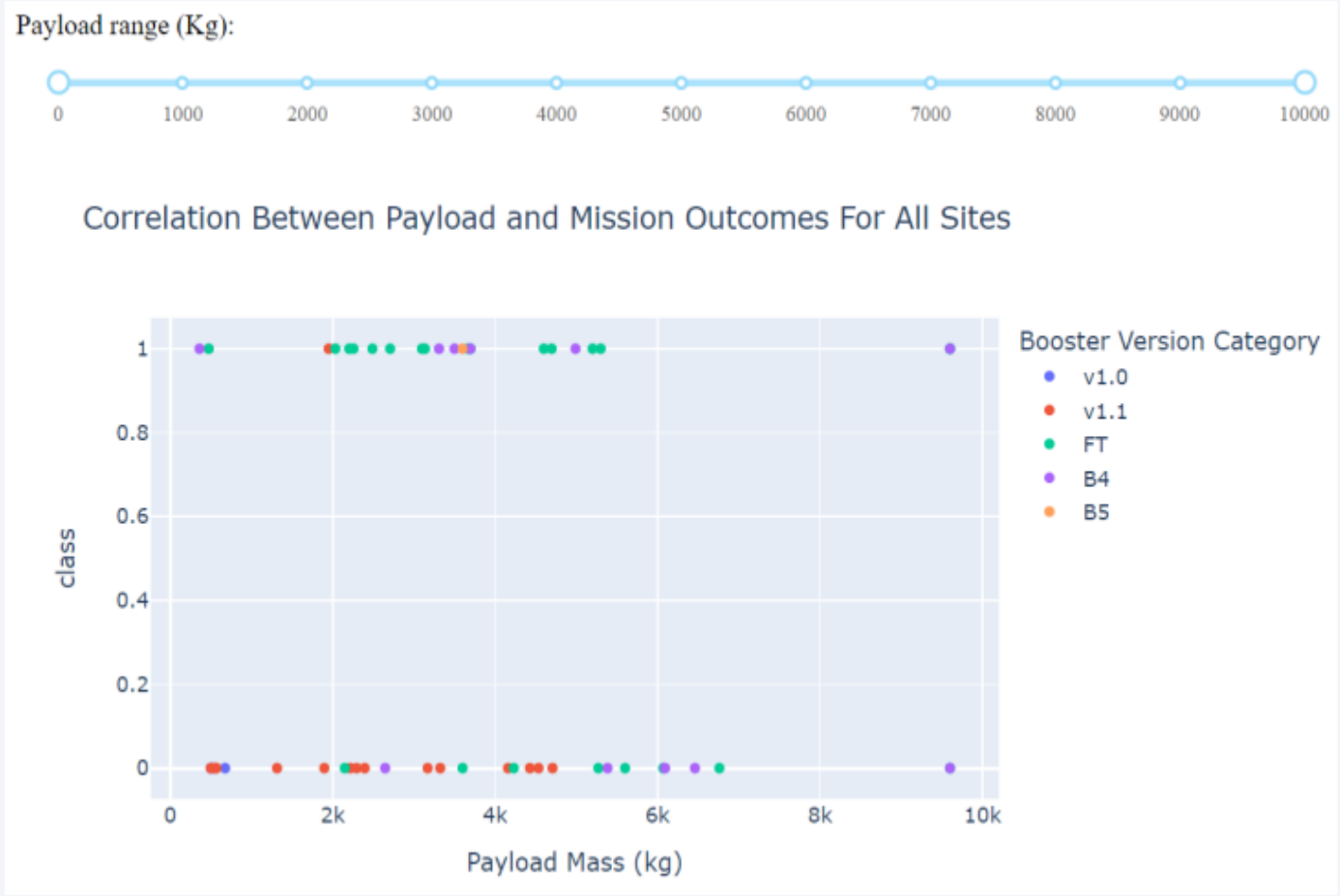
- Launch Site ‘KSC LC-39A’ has the highest launch success rate .
- Launch Site ‘CCAFS SLC□40’ has the lowest launch success rate

Launch Site with Highest Launch Success Ratio



- KSC LC-39A Launch Site has the highest launch success rate and count.
- Launch success rate is 76.9%
- Launch success failure rate is 23.1%

Payload vs. Launch Outcome Scatter Plot for All Sites

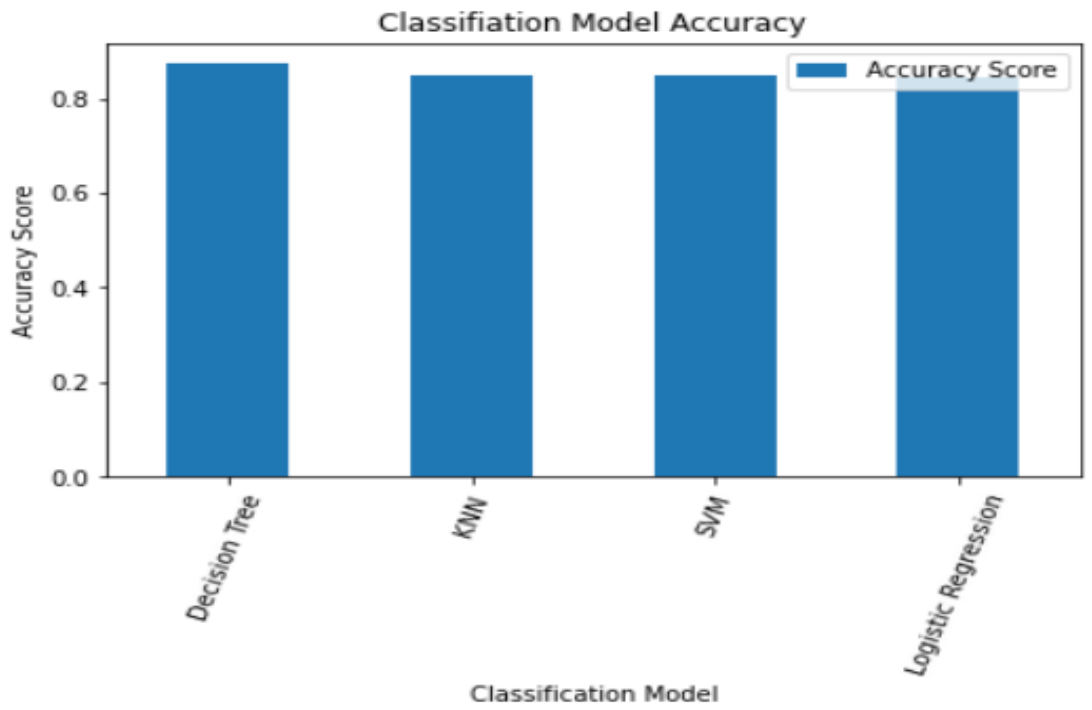


- Most successful launches are in the payload range from 2000 to about 5500.
- Booster version category ‘FT’ has the most successful launches.
- Only booster with a success launch when payload is greater than 6k is ‘B4’.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



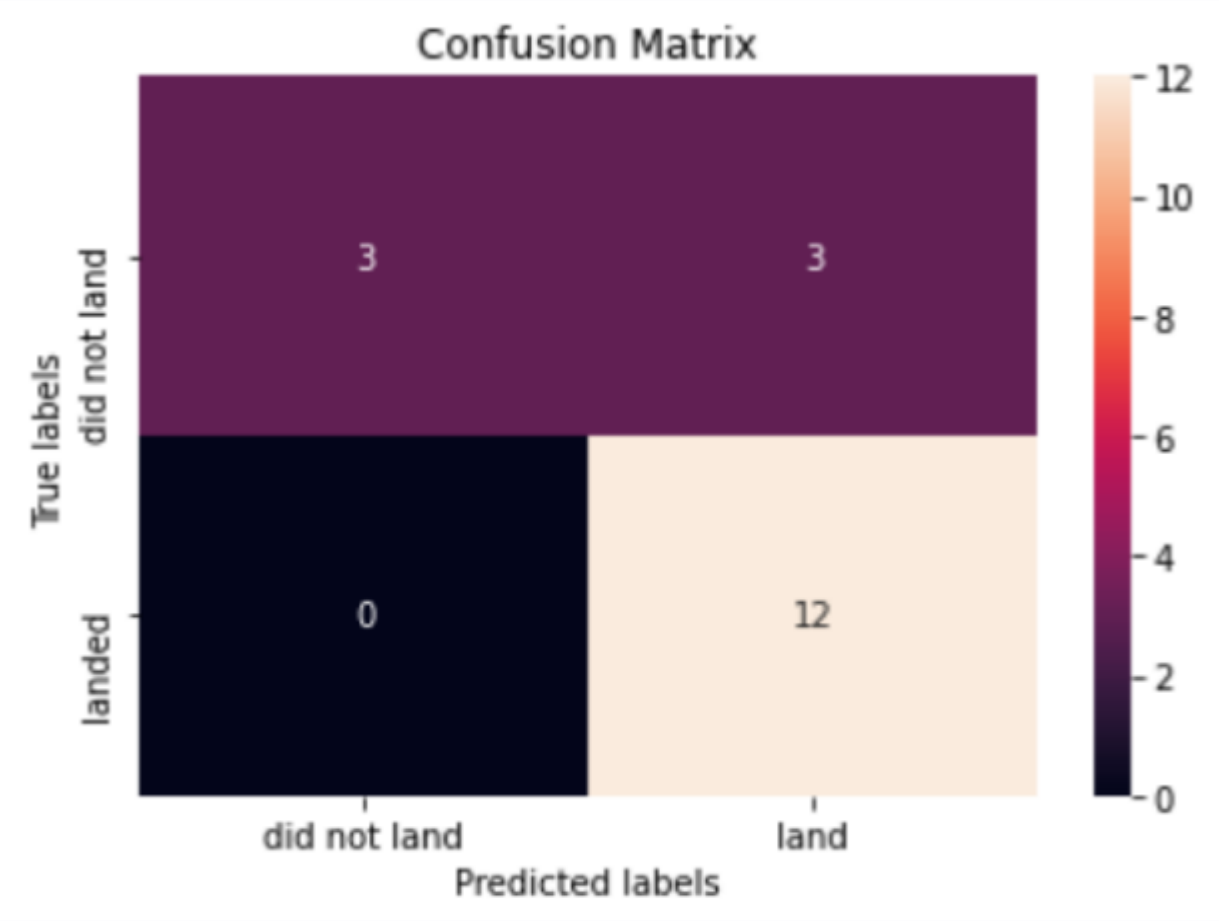
	Algo Type	Accuracy Score	Test Data Accuracy Score
2	Decision Tree	0.875000	0.833333
3	KNN	0.848214	0.833333
1	SVM	0.848214	0.833333
0	Logistic Regression	0.846429	0.833333

- The Decision Tree algorithm achieved the highest classification score of 0.8750, as depicted in the bar chart and confirmed by the accuracy scores.

- All classification algorithms had the same accuracy score of 0.8333 on the test data.

- Since the accuracy scores are very close and consistent across algorithms, there is a need for a more extensive dataset to fine-tune the models further.

Confusion Matrix



•The confusion matrix is consistent across all models (LR, SVM, Decision Tree, KNN).

• According to the confusion matrix, the classifier made a total of 18 predictions.

• It correctly predicted 12 scenarios as successful landings (True positive) and 3 scenarios as failed landings (True negative).

• However, it incorrectly predicted 3 scenarios as successful landings when they were not (False positive).

Conclusions

- As the number of flights increases, there is a higher likelihood of successful first stage landings.
- Although success rates tend to rise with increasing payload mass, no clear correlation exists between payload mass and success rates.
- The launch success rate notably increased by approximately 80% from 2013 to 2020.
- 'KSC LC-39A' has the highest launch success rate among launch sites, while 'CCAFS SLC-40' has the lowest.
- Orbits such as ES-L1, GEO, HEO, and SSO boast the highest launch success rates, whereas GTO orbits exhibit the lowest.
- Launch sites are strategically positioned away from cities and closer to coastlines, railroads, and highways.
- The Decision Tree model is the top-performing machine learning classification model, achieving an accuracy rate of approximately 87.5%. However, the accuracy scores for all models were around 83% on the test data, suggesting a need for more data to further refine and potentially improve model performance.

Appendix

GitHub respository url:

https://github.com/deva-04/IBM-DS_ML_Capstone_Project

Special Thanks to the *edX Team*

Thank you!