IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

OUTLINE

EXECUTIVE SUMMARY

The following methodologies were applied for data analysis:

- Gathering data through web scraping methods and utilization of the SpaceX API;
- Carrying out Exploratory Data Analysis (EDA), encompassing data wrangling, data representation, and interactive visual analytics;
- Application of Machine Learning for predictions.
- Comprehensive recapitulation of all findings

The EDA process facilitated the identification of optimal features for predicting launch successes;

Through Machine Learning, the most effective model was determined for predicting the vital characteristics that best leverage this opportunity, utilizing all the gathered data.

INTRODUCTION

The goal is to assess the competitive potential of the emerging company: **Space Y, in relation to Space X.**

This analysis aims to find the **optimal strategy to predict the total expense for launches**, by forecasting the successful return of rocket's and t**he most favorable location for initiating launches.**

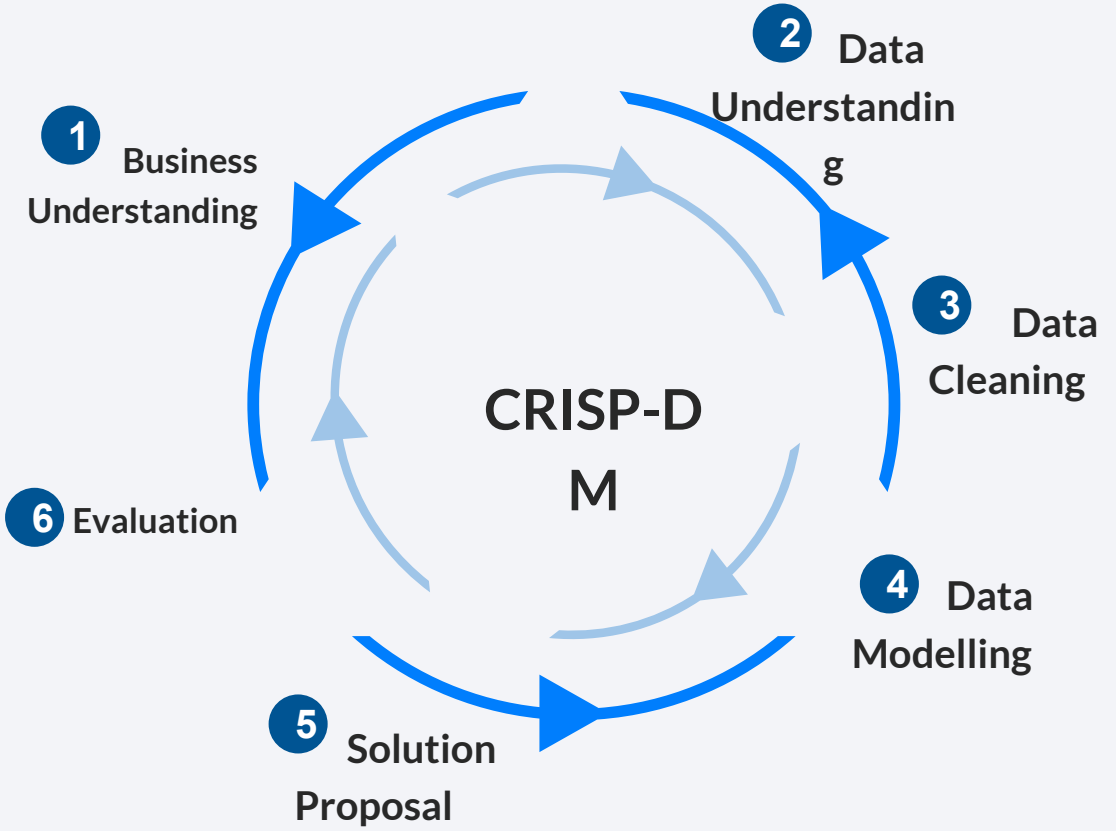
Section 1

Methodology

METHODOLOGY

The CRISP-DM (Cross-Industry Standard Process for Data Mining) was used to develop this project.

CRISP methodology in data analysis is a structured approach to planning and implementing a data mining project. It includes six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.



METHODOLOGY

01 | **Data Collection**

Collect Data using API and Web Scraping

02 | **Data Wrangling**

Filtering data, handling missing values and prepare data for modelling

03 | **Data Exploratory**

SQL for EDA and visualization methods

04 | **Interactive Visual Analytics**

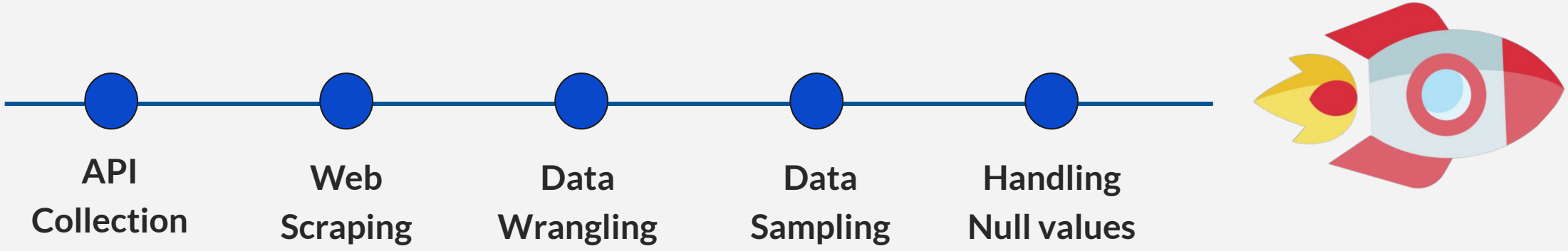
Maps and interactive dashboards using Folum and Plotly Dash

05 | **Predictive Analysis**

Classification model to predict landing outcomes

DATA COLLECTION PROCESS

The diagram illustrates the sequential process of data collection, beginning with API sourcing and web scraping, followed by data cleaning, sampling, and finally, handling of null values.



API DATA COLLECTION

- The SpaceX REST API was utilized to gather data about SpaceX launches.
- The API provided data such as rocket details, payload delivery, launch and landing specifications, and landing outcomes.
- Data was collected from the specific endpoint 'api.spacexdata.com/v4/launches/past'.
- A 'get' request was sent using the requests library to retrieve the launch data.
- The received response was in the form of JSON objects.
- The json_normalize function was used to convert the structured JSON data into a flat table.

API DATA COLLECTION

To collect the data, we used SpaceX API. To manipulate the data, we followed the flowchart below:

Source Code

https://github.com/daianeklein/ibm-data-data-science-professional-certificate/blob/main/01_data_collection_API.ipynb

WEB SCRAPING

- Used the Python BeautifulSoup package to web scrape Falcon 9 launch records from related Wiki pages.

- Extracted data from HTML tables and converted it into a Pandas DataFrame for further analysis.

WEB SCRAPING

The launches data can be collected on Wikipage page as well. This data collection followed the following chart:

Source Code

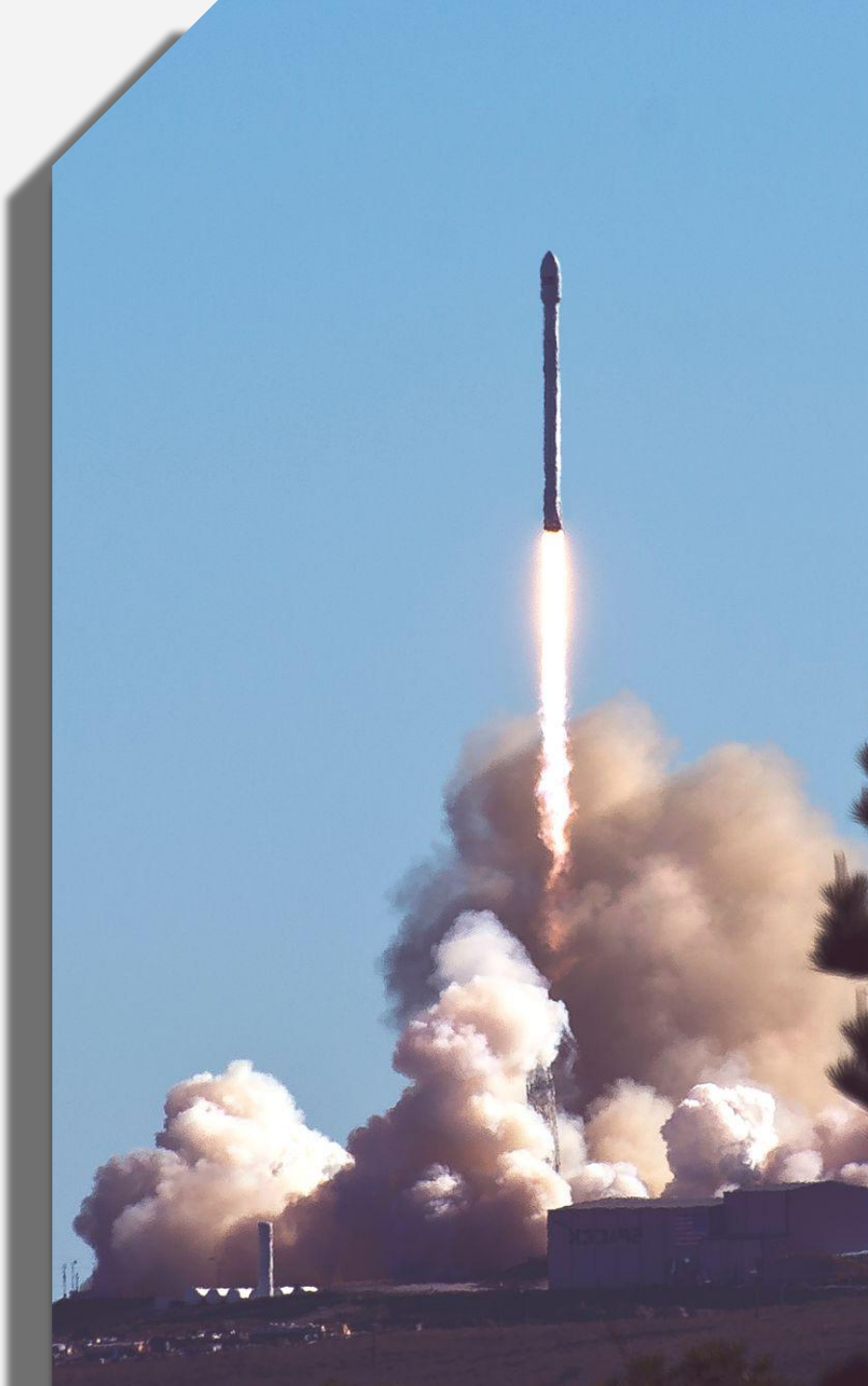
https://github.com/daianeklein/ibm-data-data-science-professional-certificate/blob/main/01_data_collection_API.ipynb

DATA WRANGLING

- Identified columns such as 'rocket' that contained identification numbers instead of actual data.
- Re-targeted the API to retrieve specific data for each identification number using already created functions: Booster, Launchpad, payload, and core.
- Data stored in lists and utilized to create the final dataset.

DATA WRANGLING

- Filtered the gathered data to remove records for the Falcon 1 booster, focusing solely on Falcon 9.
- Handling Null Values:
- Encountered NULL values in certain columns, such as 'PayloadMass'.
- Calculated the mean of the PayloadMass data and replaced the null values in 'PayloadMass' with this mean.
- Null values in 'LandingPad' were left as is to indicate instances when a landing pad was not used, to be dealt with using one hot encoding later on.



DATA WRANGLING

After Exploratory Data Analysis (EDA) was conducted on the dataset, summaries of launches per site, occurrences of each orbit, and occurrences of mission outcome per orbit type were then calculated.

Source Code

https://github.com/daianeklein/ibm-data-data-science-professional-certificate/blob/main/03_data_wrangling.ipynb

EDA WITH VISUALIZATION

Data exploration was facilitated through the use of scatterplots and barplots to visualize relationships between pairs of features.

Bar charts were employed to compare discrete categories. These charts illustrate the relationships between categories and a measured value.

Source Code

https://github.com/daianeklein/ibm-data-data-science-professional-certificate/blob/main/05_data_visualization.ipynb

EDA WITH SQL

Queries were performed:

- Unique launch sites
- Launch site that begins with 'CCA'
- Total Payload mass carried by booster launched by NASA
- AVG payload mass carried by booster version F9 V1.1
- Date of the first successful landing on the ground pad.
- Names of boosters that successfully landed on a drone ship and had a payload mass between 4,000 and 6,000.
- Total number of successful and failed missions.
- Names of booster versions that have carried the maximum payload.
- Failed landing outcomes on drone ship, their booster version, and launch site for the months in the year 2015.
- Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc).

Source Code

https://github.com/daianeklein/ibm-data-data-science-professional-certificate/blob/main/04_EDA_SQL.ipynb

17

MAP WITH FOLIUM

Markers indicating launch sites were added:

A **blue** circle was added at NASA Johnson Space Center's coordinates with a popup label showing its name, utilizing its latitude and longitude coordinates.

Red circles were added at all launch site coordinates with popup labels displaying their names, using their respective latitude and longitude coordinates.

Colored markers of launch outcomes were incorporated:

Successful launches were indicated with **green** markers and unsuccessful ones with **red** markers at each launch site to visualize success rates.

Colored lines were added to illustrate the distance between launch site CCAFS SLC-40 and its nearest coastline, railway, highway, and city.

Source Code

https://github.com/daianeklein/ibm-data-data-science-professional-certificate/blob/main/06_interactive_visual_analytics_folium.ipynb

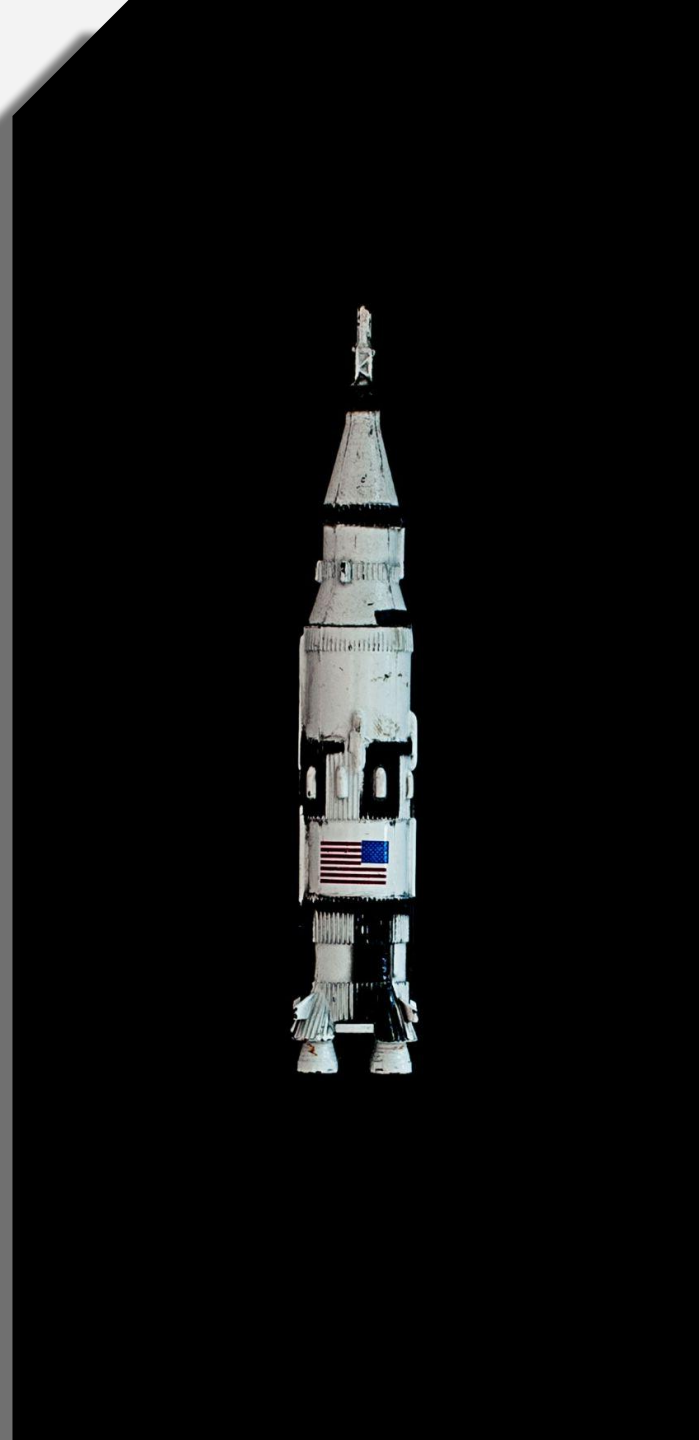
DASHBOARD - PLOTLY DASH

Various graphs and plots were utilized to visualize the data, including:

Percentage of launches by site

Payload range

This combination of visuals enabled rapid analysis of the relationship between payloads and launch sites, assisting in identifying the optimal launch locations based on payloads.



Source Code

https://github.com/daianeklein/ibm-data-data-science-professional-certificate/blob/main/07_interactive_visual_analytics_plotly.py

PREDICTIVE ANALYSIS

Four classification models were compared: logistic regression, support vector machine, decision tree, and k-nearest neighbors. The steps taken are shown below:

1. Numpy array fro class column
2. StandardScaler to transform the data
3. Split data into train and test
4. GridSearch optimization
5. Different algorithms applied: Logistic Regression, Support Vector Machine, Decision Tree, KNN
6. Calculate the accuracy
7. Analyze the Confunsion Matrix
8. Identity the best model based on Jaccard score, F1 score and accuracy

Source Code

https://github.com/daianeklein/ibm-data-data-science-professional-certificate/blob/main/08_predictive_analytics.ipynb

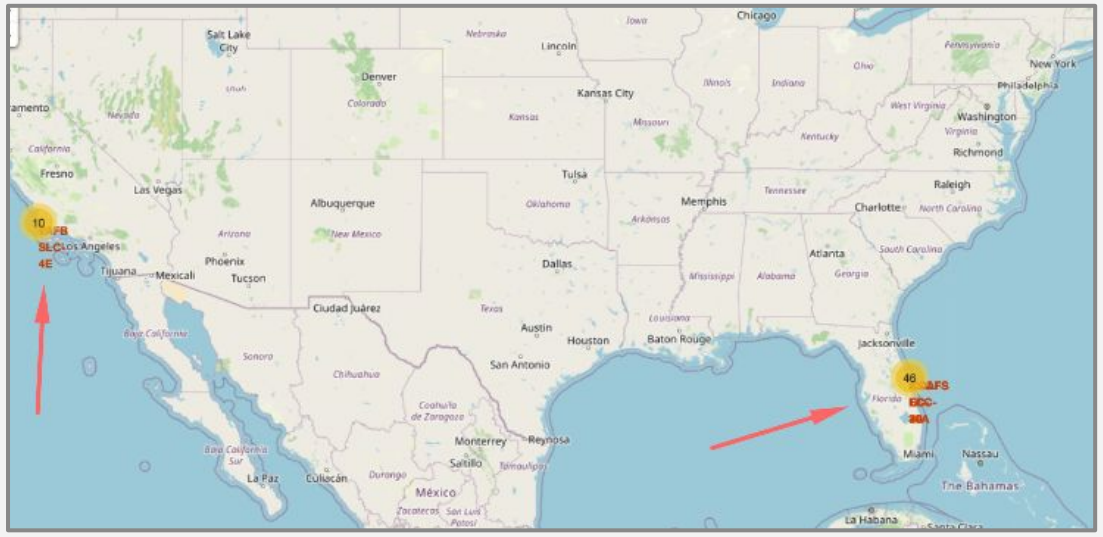
RESULTS

EXPLORATORY DATA ANALYSIS

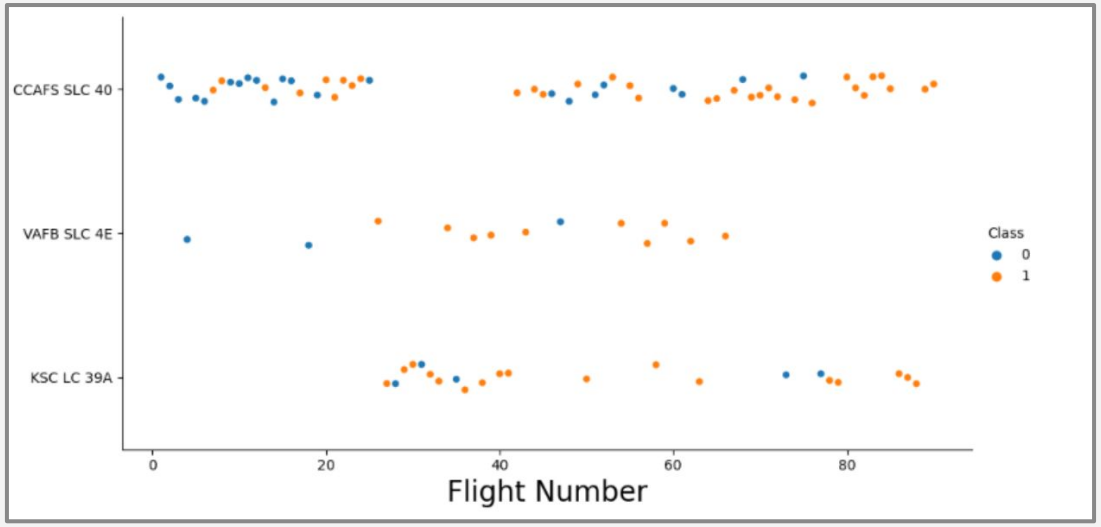
- Space X has four launch sites, mostly near the equator and coast.
- Initial launches were for Space X and NASA, with a F9 v1.1 booster average payload of 2,928 kg.
- Successful landings started in 2015, with increasing success rates over time.
- KSC LC-39A has the highest landing success rate.
- 100% success rate for ES-L1, GEO, HEO, and SSO orbits.
- Decision Tree is the best predictive model for the dataset.

LAUNCH SITES

- Launch sites are strategically located in safe places, usually near the sea, with robust logistical infrastructure.
- Majority of launches occur at east coast sites.

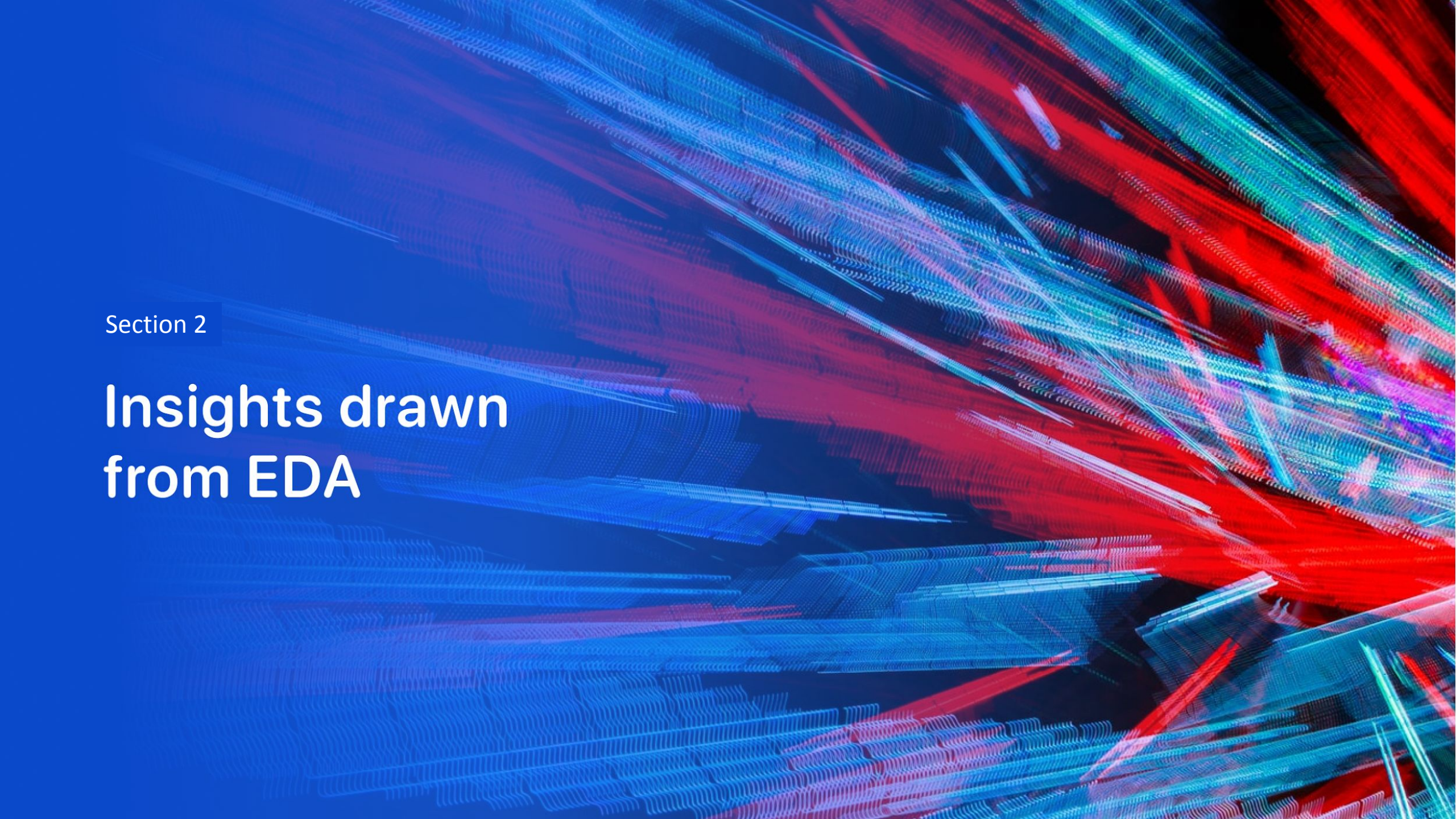


- Early flights had lower success rates, while later flights showed improved success.
- Approximately half of the launches were from CCAFS SLC 40.
- VAFB SLC 4E and KSC LC 39A demonstrated higher success rates.
- Newer launches generally show a higher success rate.



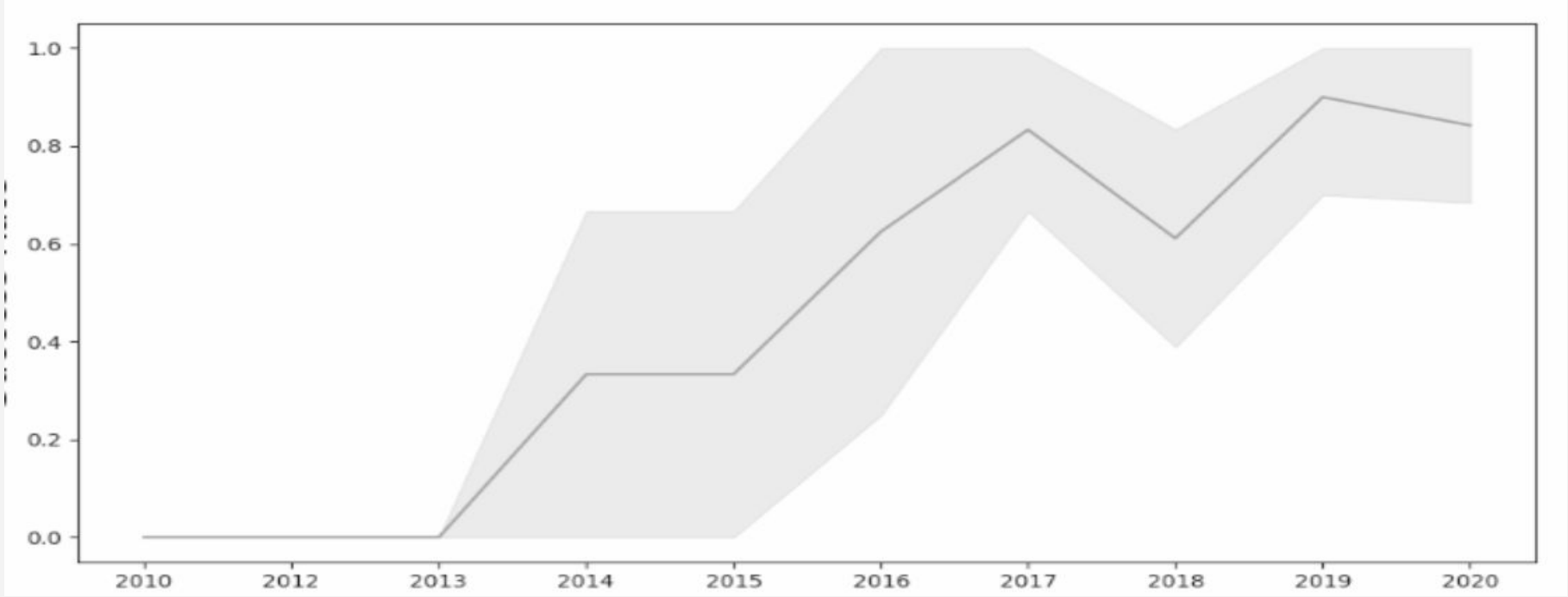
Section 2

Insights drawn from EDA



LAUNCH SUCCESS OVER TIME

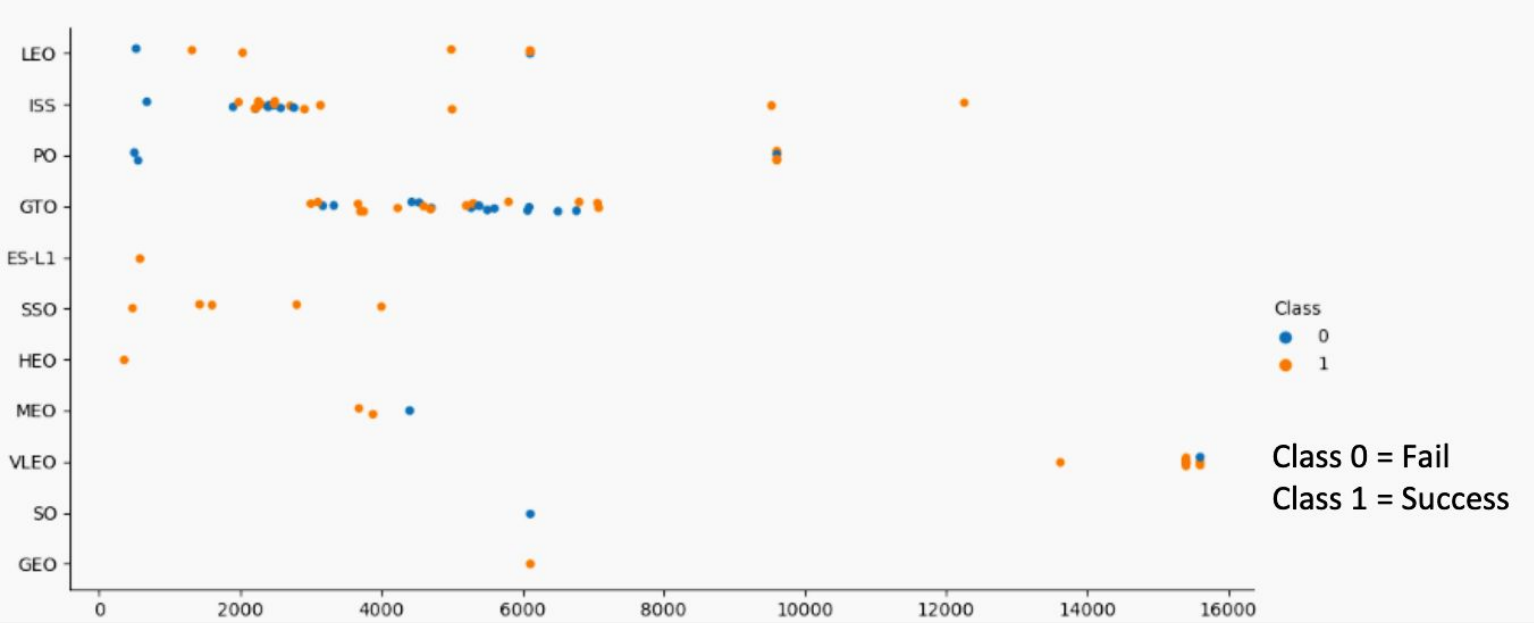
- Success rate showed improvement from 2013-2017 and 2018-2019.
- A decrease was observed from 2017-2018 and 2019-2020.
- Overall, success rate has shown a positive trend since 2013.



PAYLOAD & ORBIT

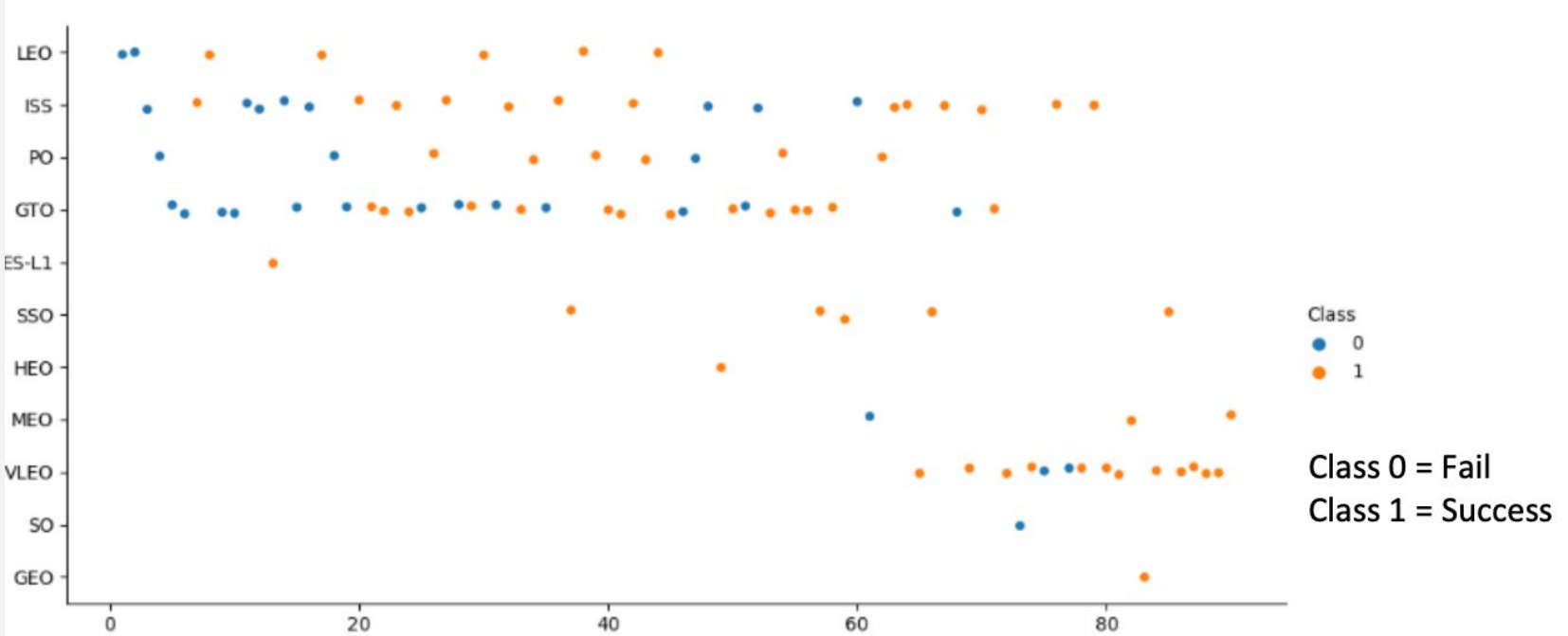
- No clear relationship between payload and success rate for GTO orbit.
- ISS orbit handles a wide range of payloads with a high success rate.
- Fewer launches to SO and GEO orbits.

- Heavy payloads fare better with LEO, ISS and PO orbits.
- GTO orbit has inconsistent success with heavier payloads.



FLIGHT NUMBER & ORBIT

- Success rates improved over time across all orbits.
- VLEO orbit shows potential as a new business opportunity due to its increasing frequency.



LAUNCH SITE NAMES

There are four launch sites:

1. CCAFS LC-40
2. CCAFS SLC-40
3. KSC LC-39A
4. VAFB SLC-4E

Launch Site names - Begin with 'CCA'

Date	Time	Booster_Version	Launch_Site	Payload	PAYLOA	Orbit	Customer	Mission_Outcome	Landing_Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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 (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

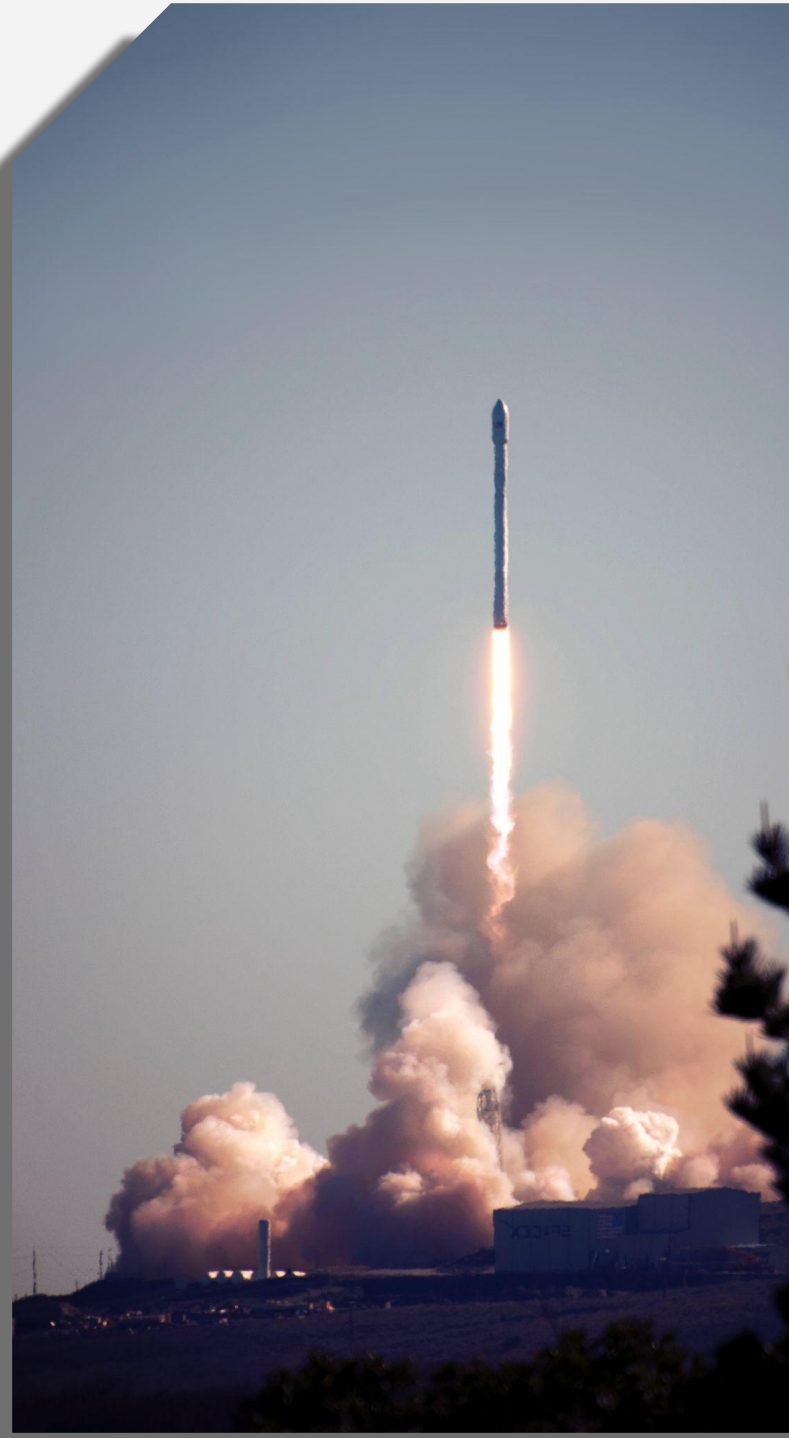
TOTAL PAYLOAD MASS

Total payload carried by NASA boosters:

111.268

AVG payload carried by NASA boosters:

2.928



LANDING DATES

1st Successful landing in Ground

2015/12/22

Failure in Flight

1

Success

99

BOOSTERS

Max Payload Carrying

- 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 RECORDS

Month, date, booster version, launch site and landing outcome

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

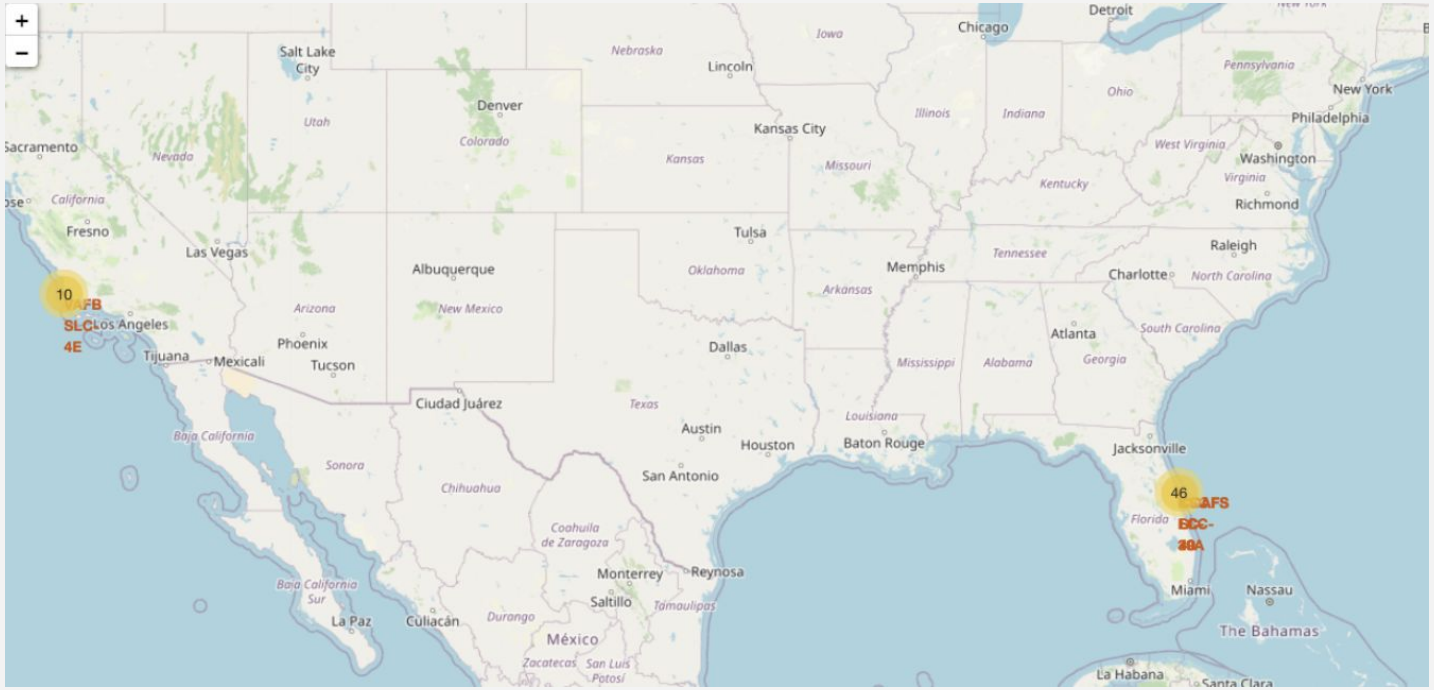
All landing outcomes from 2010/06/04 to 2017/03/20

Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

LAUNCH SITES

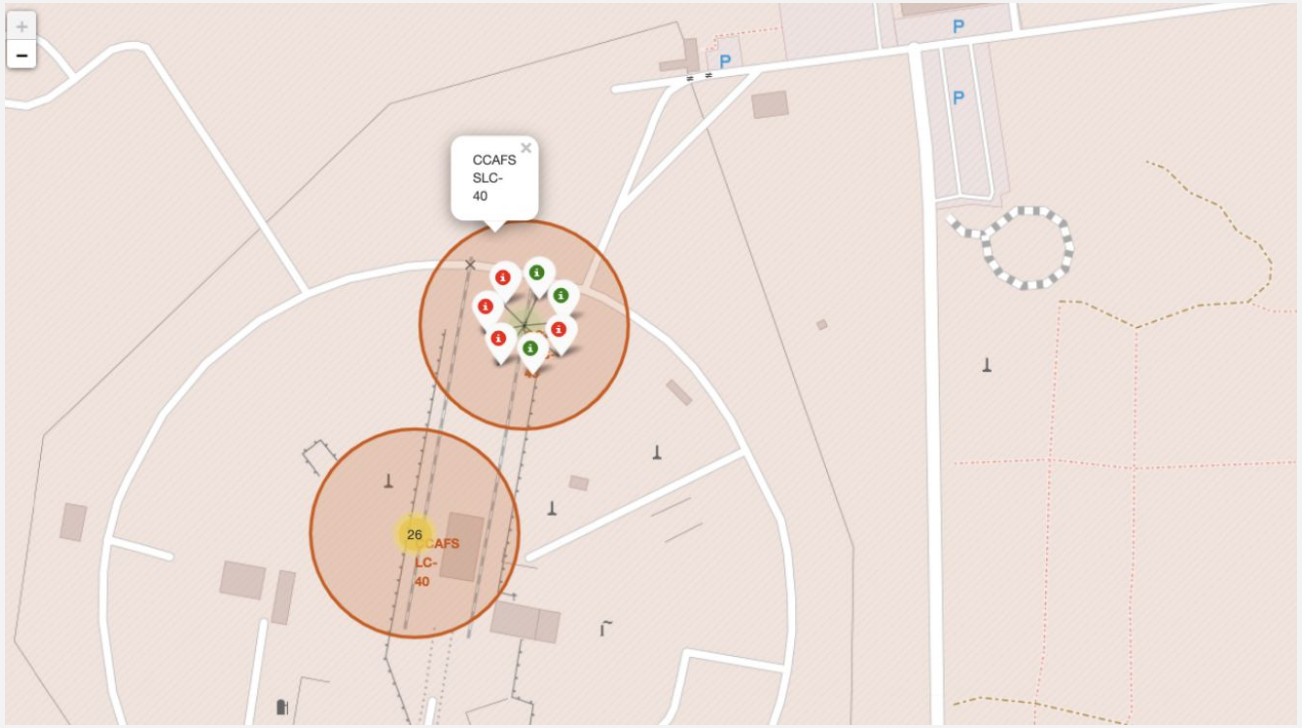
Launch sites are often located near the sea for safety reasons, however, these sites are also not too far from roads and railroads for logistical reason.



- Being close to the Equator provides an advantage for launching rockets into equatorial orbit.
- The Earth's rotational speed gives rockets an additional natural boost when launched near the Equator.
- This natural boost reduces the need for extra fuel and boosters, leading to cost savings.

LAUNCH OUTCOMES

Launch site CCAFSSLC-40 has success rate of 42.9%.

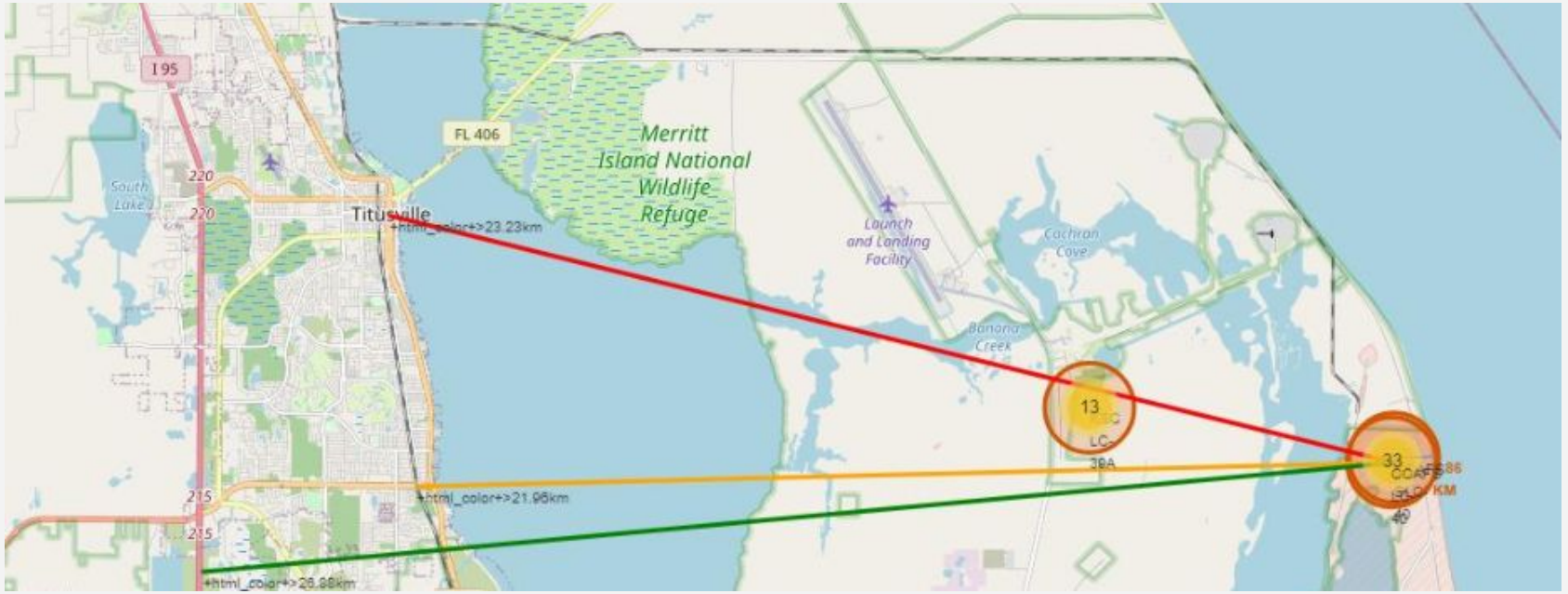


Green marks → Successful

Red marks → Unsuccessful

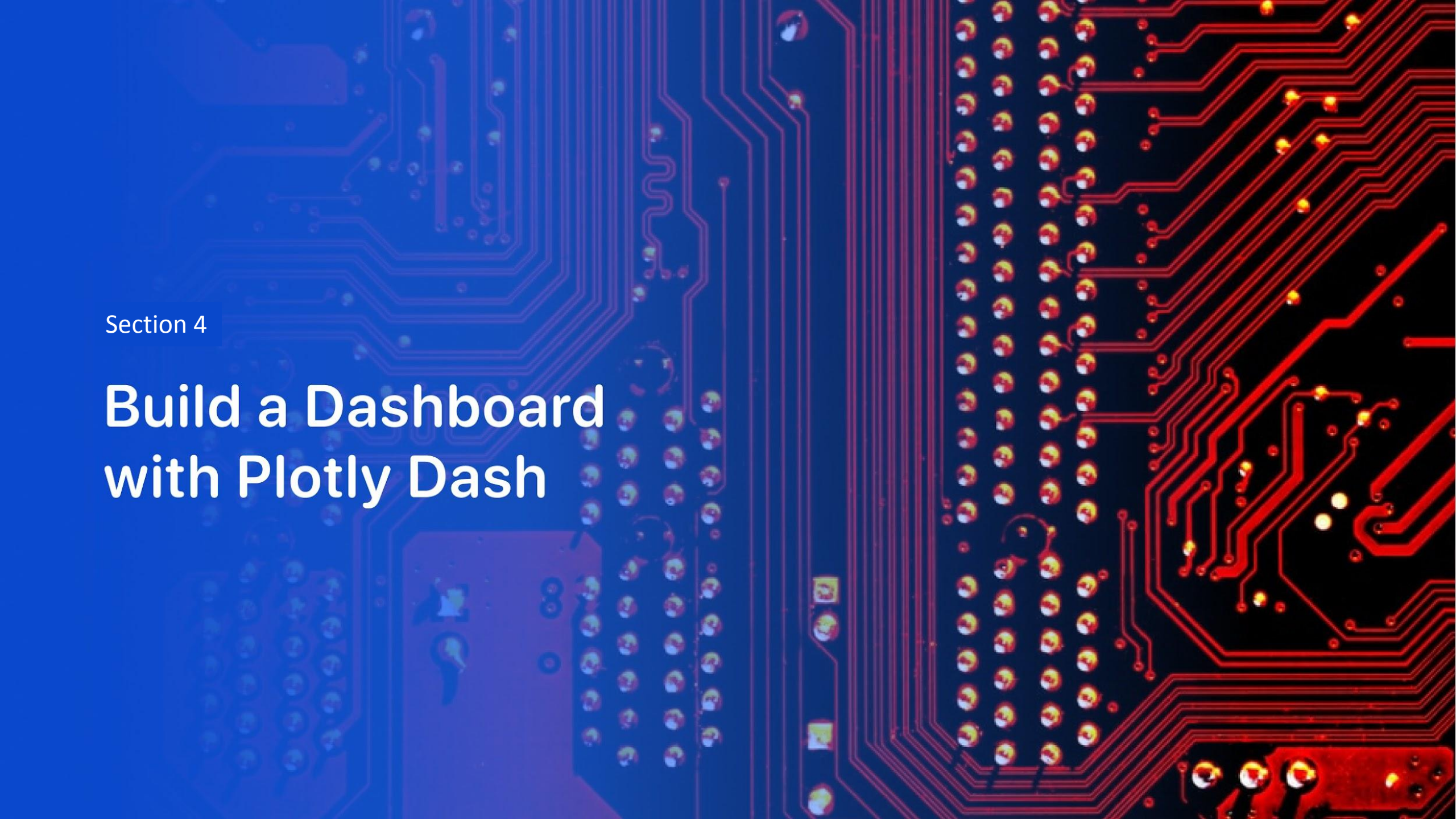
LOGISTICS

- Coastal launch sites minimize risks associated with spent stages or failed launches.
- Safety zones are required around launch sites for public safety and security.
- Launch sites should be distanced from potential damage zones, yet near transport links for logistics.



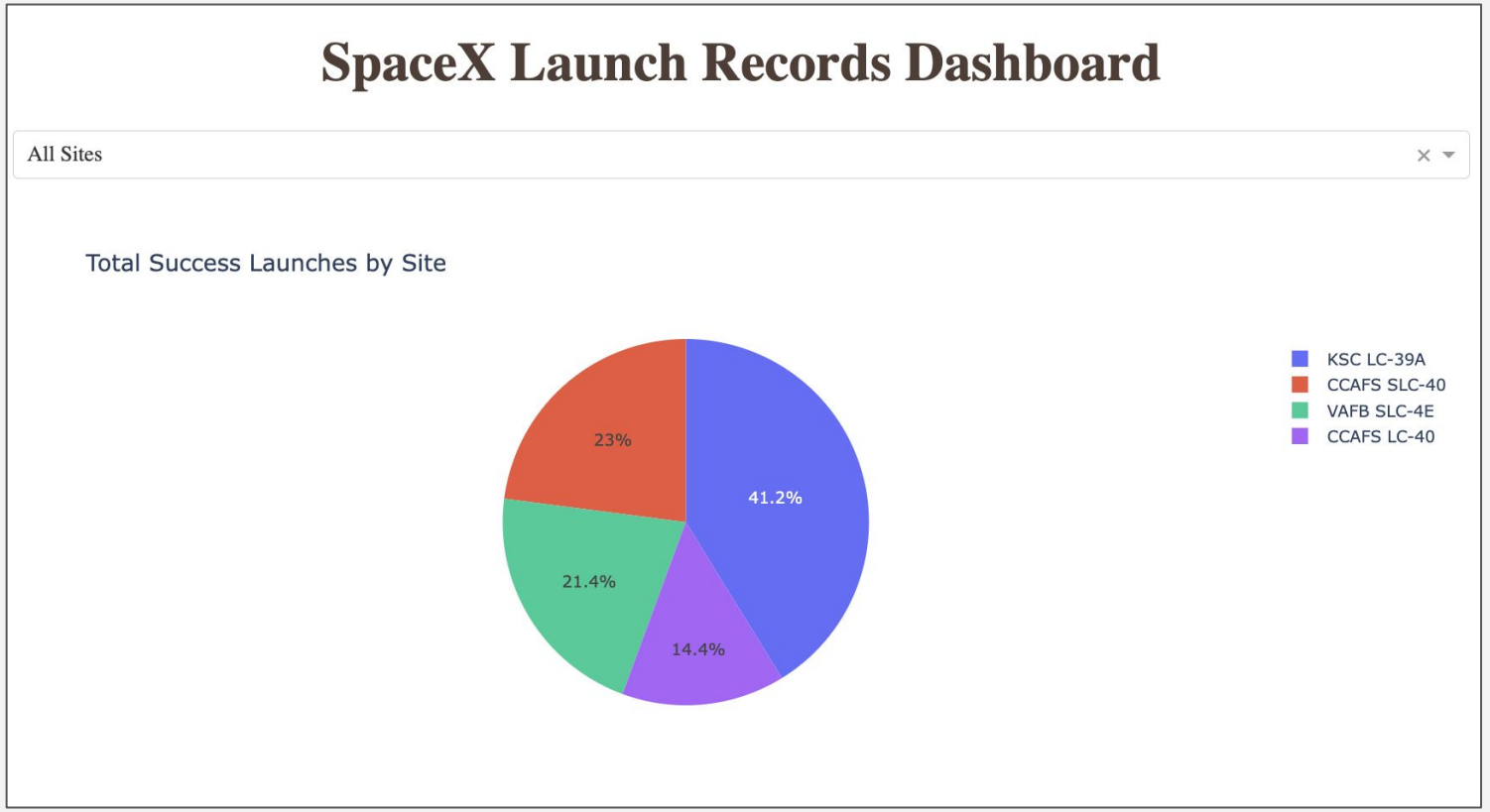
Distances

- 0.86 km from nearest coastline
- 21.96 km from nearest railway
- 23.23 km from nearest city
- 26.88 km from nearest highway


Section 4
Build a Dashboard with Plotly Dash

SUCCESSFUL LAUNCHES BY SITE

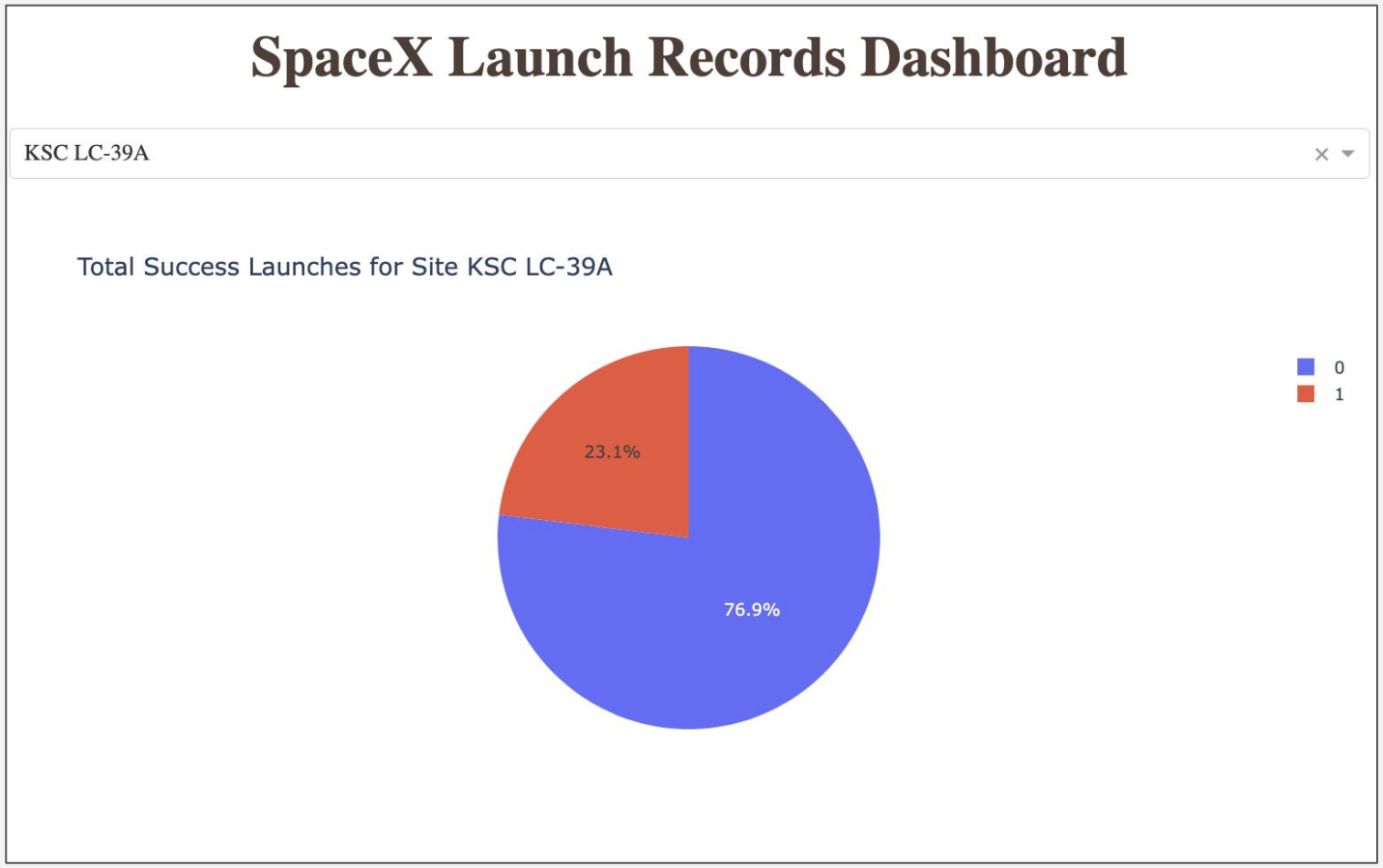
- The location of the launch site plays a significant role in the success of space missions.
- Kennedy Space Center Launch Complex 39A (KSC LC-39A) records the highest success rate among launch sites, with 41.2% of launches being successful.



LAUNCH SUCCESS RATIO - KSC LC-39A

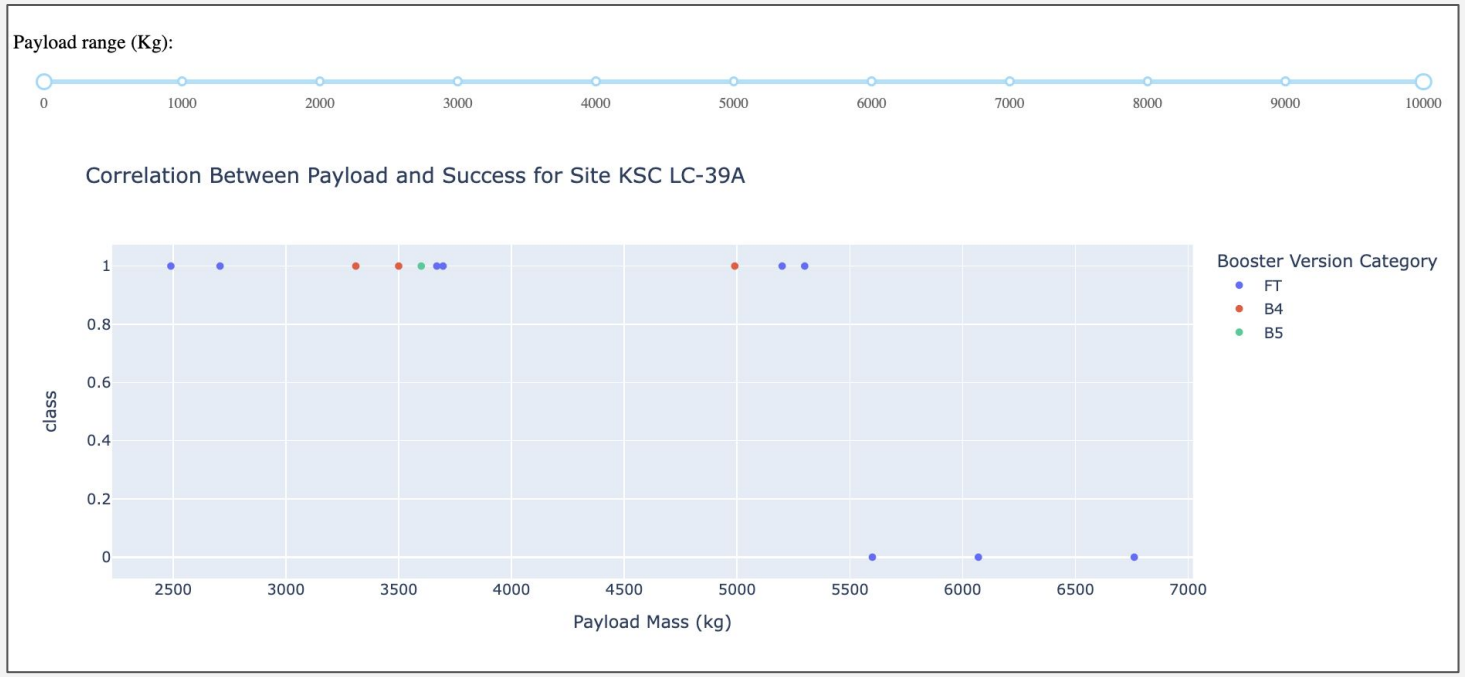
Successful Launches in this site

76.9%



PAYLOAD & LAUNCH OUTCOME

- Payloads weighing between 2,000 kg and 5,000 kg demonstrate the highest success rate.
- A score of 1 represents a successful mission outcome, while a score of 0 indicates an unsuccessful outcome.
- The most successful mission configuration consists of payloads under 6,000 kg combined with Falcon Thrust (FT) boosters.



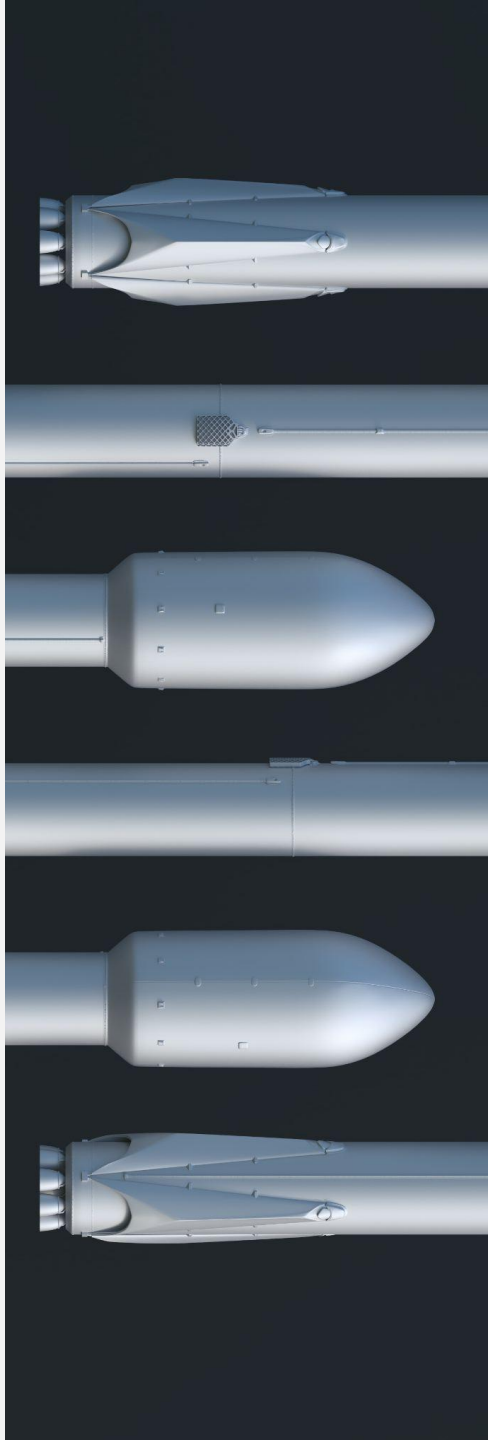
Section 5

Predictive Analysis (Classification)

Machine Learning Process

Machine learning pipeline to predict the successful landing of Falcon 9's first stage.

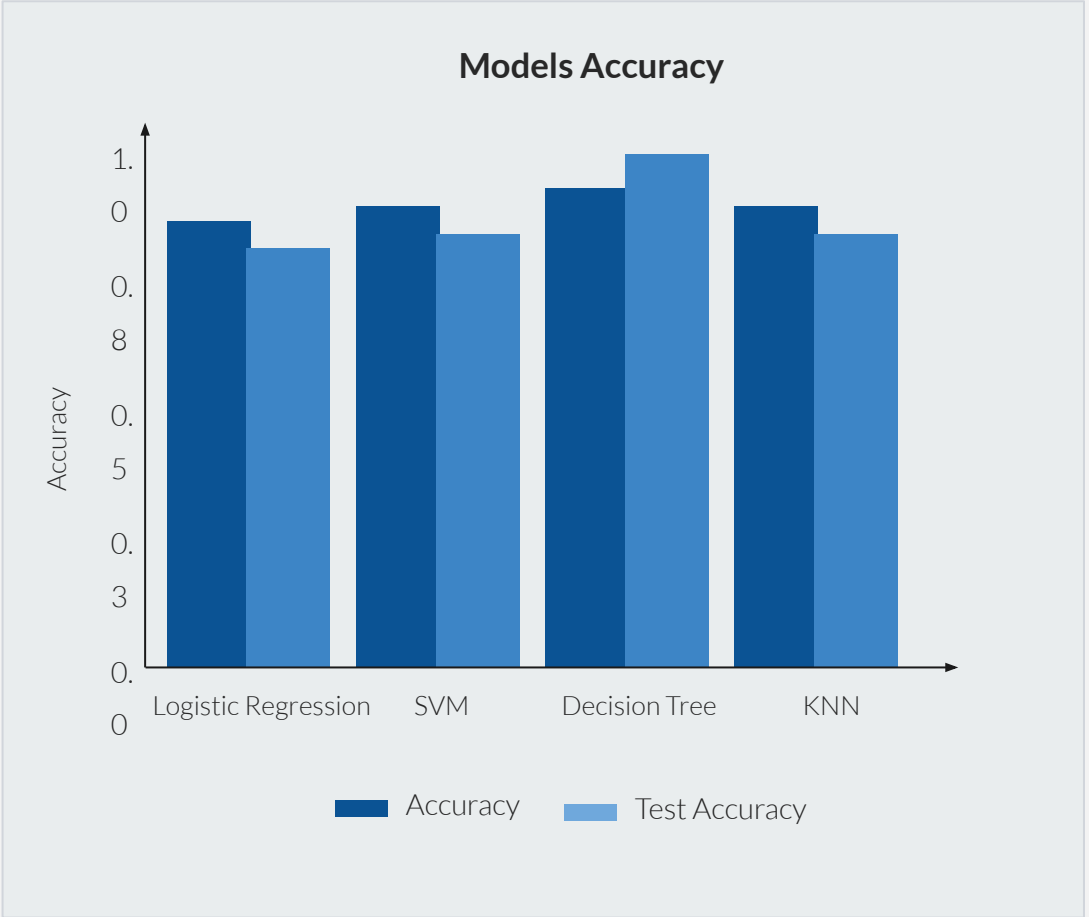
1. The process began with **Preprocessing**, which was used to standardize the data.
2. Then, the **Train_test_split** function was employed to divide the data into separate sets for training and testing.
3. A **Grid Search** was performed after training the model to identify the best-performing hyperparameters.
4. Using these **optimal hyperparameters**, the model that provided the highest accuracy using the training data was determined.
5. Different models were tested including **Logistic Regression, Support Vector Machines, Decision Tree Classifier**, and **K-nearest Neighbors**.
6. The final step of the analysis was the generation of a **confusion matrix** to evaluate the performance of the model.



ML MODELS ACCURACY

- Four different classification models were tested and their accuracy levels were compared.

- The model that yielded the highest classification accuracy was the Decision Tree Classifier, which achieved accuracy rates of over 87%.



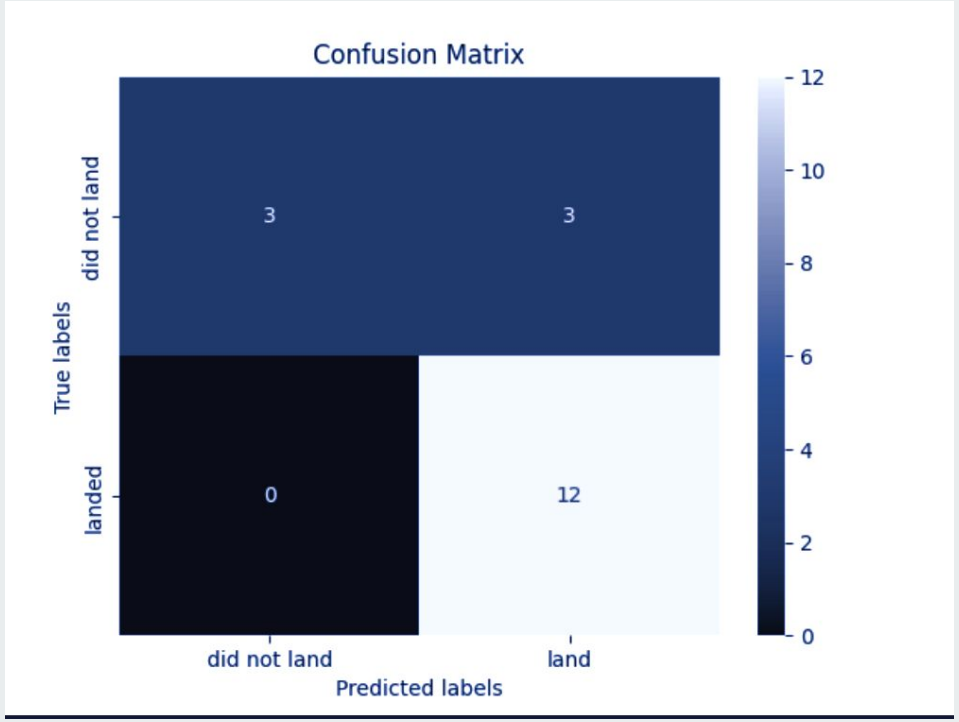
CONFUSION MATRIX

A confusion matrix is a table in machine learning that shows the accuracy of a classification model's predictions by detailing correct and incorrect outcomes.

Outputs:

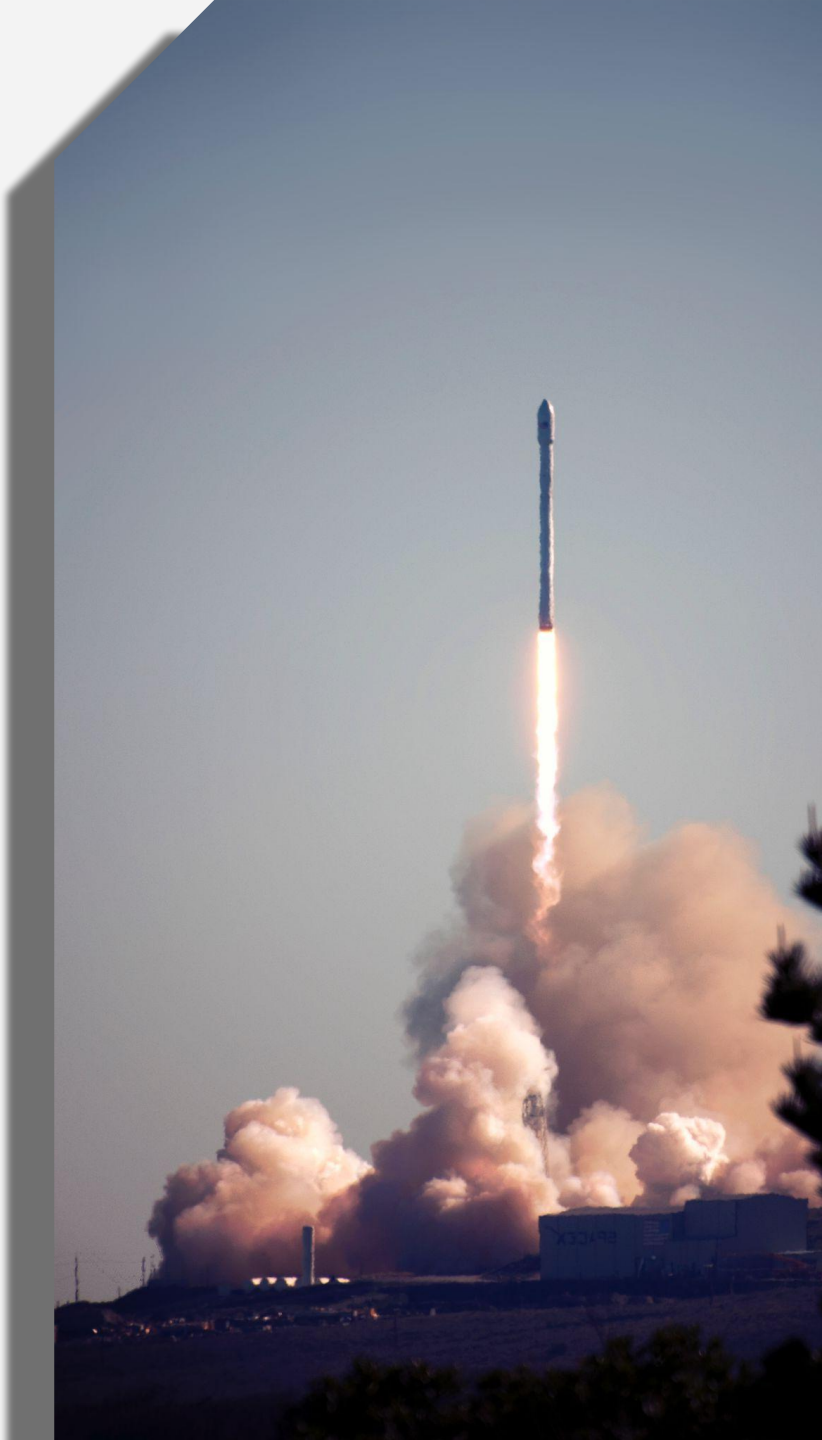
- 12 True positive
- 3 True negative
- 3 False positive
- 0 False Negative

Confusion Matrix



CONCLUSION

- Data from various sources was analyzed and refined throughout the process.
- The most optimal launch site was identified as KSC LC-39A.
- Launches with payloads above 7,000kg were found to be less risky.
- Most mission outcomes were successful, and successful landing outcomes appear to improve over time with advancements in processes and rockets.
- The Decision Tree Classifier emerged as a reliable model for predicting successful landings, aiding in profit increase.



CONCLUSION

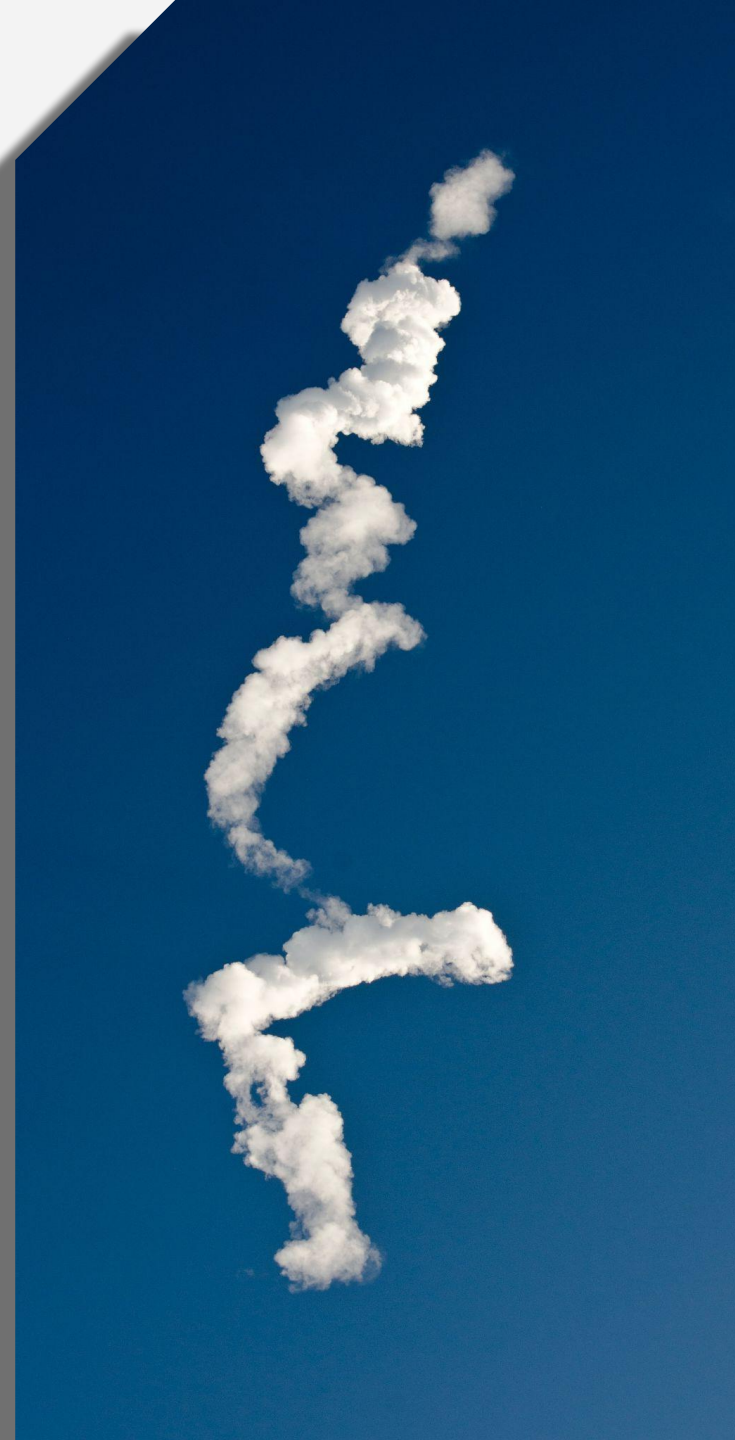
KEY FINDINGS

Data Point	Findings
Experience and Success Rate	Success rates at launch sites increased with more experience. Early flights were mainly unsuccessful.
Time and Success Rate	Between 2010 and 2013, all landings failed. Success rates increased post-2013, with more than 50% chance of success after 2016.
Orbit Types	ES-L1, GEO, HEO, and SSO had the highest (100%) success rates. The success rate in SSO was particularly impressive with 5 successful flights.
Payload and Orbit	Heavy payloads had more success with PO, ISS, and LEO orbits. VLEO launches were associated with heavier payloads.
Launch Site Success	KSC LC-39 A had the most successful launches (41.7% of total) and the highest rate of successful launches (76.9%).
Payload Success	Success for heavy payloads (over 4000kg) was lower than for lighter payloads.
Model Performance	The best performing model was the Decision Tree model, with 94.44% accuracy.

CONCLUSION

RECOMMENDATIONS FOR BETTER DATA ANALYSIS AND PREDICTIONS

- Collect larger datasets for more generalizable predictive analytics results.
- Conduct further feature analysis or Principal Component Analysis (PCA) to improve accuracy.
- Test the XGBoost model, as it's a powerful model not utilized in this study, to see if it outperforms other classification models.



Thank you!