IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tom Scherbluk
March 23, 2024

Outline

Executive Summary

Problem Statement
The success of SpaceX hinges on the reusability of their rockets. For another firm to bid competitively against SpaceX, it is imperative to accurately estimate their rockets' performance. To do this, an analysis must be undertaken to build a model that predicts the landing outcome of a rocket after it is launched.

Solution
By collecting data on SpaceX's Falcon 9 rocket launches and using it to train Machine Learning models, we were able to predict whether the first stage of the rocket will land successfully. This was done via:

- Data Collection using API requests and Web Scraping
- Data Analysis
- Predictive Modeling using Machine Learning Algorithms

Value
The predictive models enabled accurate estimates of the success of a rocket landing. Since a failed landing could result in a loss of tens of millions of dollars, the accuracy of the model will enable a competitive company to determine launch costs and make competitive bids against SpaceX.

Assessment and Next Steps
The space launch market is projected to reach USD 26.16 billion by 2027.
Powerful Analytics and Machine Learning tools improve the ability of competitors to accurately predict space launch outcomes and confidently bid for launch contracts. competitiveness – they can also increase customer confidence in our offerings.

For more information, visit the full project repository: https://github.com/Vision-City/SpaceX-Project

Introduction

Project background and context

With the explosion of the commercial space age, companies are developing technology in remarkable ways to make space launches more affordable. SpaceX is one of the most successful companies thanks to their innovative reusable boosters. Their publicly available data will form the basis of an analysis to understand the viability of reusing a rocket's first stage. Through this analysis, we'll use this benchmark for reusable rockets to determine how cost-effective this kind of performance level can be for a competitive company.

What problems we want answered

This project's goal was to predict whether a SpaceX Falcon 9 first stage will land successfully. Falcon 9 rocket launches cost 62 million dollars. Other providers cost upward of 165 million dollars each. Much of the savings is because SpaceX can reuse the first stage.

If an alternate company wants to bid against SpaceX for a launch, it is crucial to estimate if the first stage will land, since that will help us determine the cost of a launch. With the help of the Data Science findings and models, accurately informed bids against SpaceX can be made for rocket launch contracts.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Launch data was collected from two sources: the SpaceX REST API and scraped from Wiki pages that contained Falcon rocket launch information.
- Perform data wrangling
  - The JSON data from the SpaceX API was converted into a dataframe for analysis whereas the information from the Wiki scraped through BeautifulSoup was converted into a Pandas dataframe for analysis.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The data was split into training data and test data to find the best Hyperparameter for Support Vector Machines, Classification Trees, and Logistic Regression. A comparison of the results of these methods yielded the one that performed best using the test data.

Data Collection

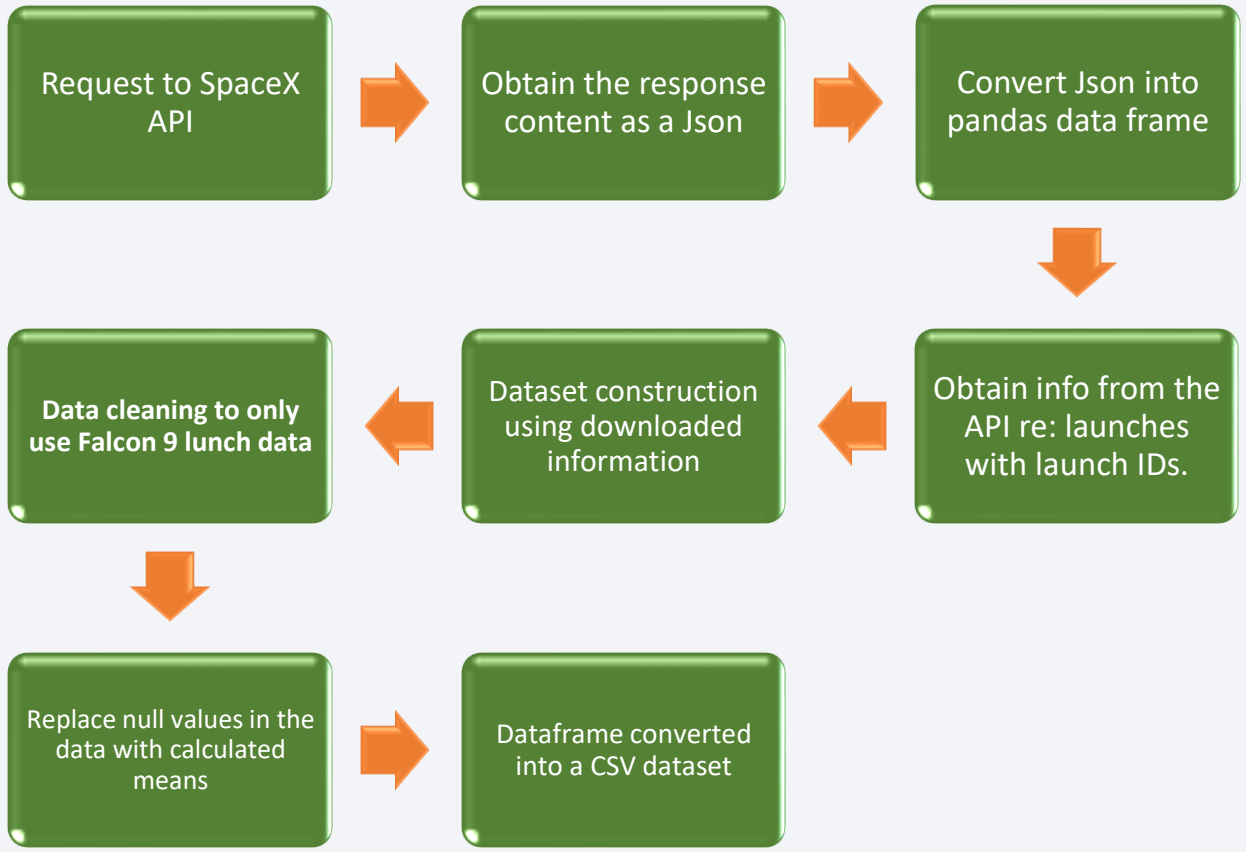
The success of this project hinges on the data collection stage. Without accurate collection, machine learning model training will yield inaccurate results on which to make business decision from.

The two methods used to collect data were:

- Data collection via the SpaceX API.
- Data collection by Web Scraping from a Wiki page of SpaceX launch information

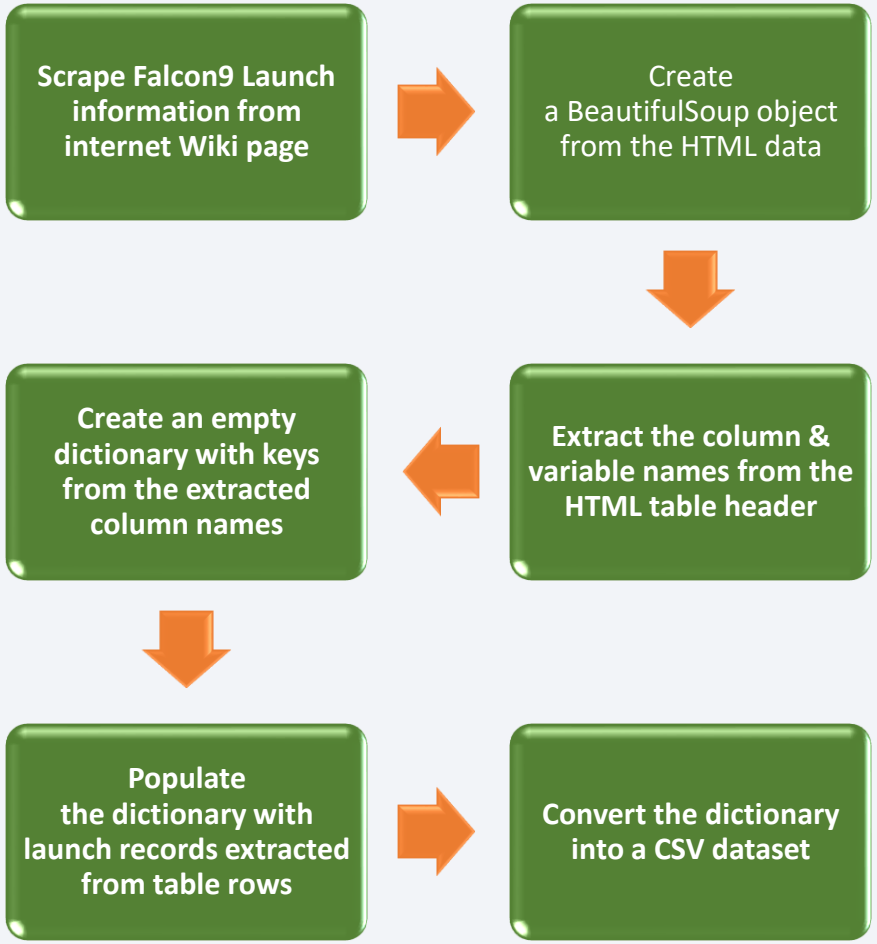
Data Collection – SpaceX API

- A request was made to SpaceX API and the data was checked to ensure it was in the correct format.
- Some basic data wrangling was performed in order to clean the gathered data.
- The resulting data frame was converted into a CSV dataset.
- URL link: https://github.com/Vision-City/SpaceX-Project/blob/main/notebooks/Week1-Data-Collection.ipynb



Data Collection - Scraping

- Using BeautifulSoup, a web scraping was performed on the Wiki page with title: "*List of Falcon 9 and Falcon Heavy Launches*"
- The launch records were stored in an HTML table.
- The table was parsed and converted into a CSV dataset.
- URL link: https://github.com/Vision-City/SpaceX-Project/blob/main/notebooks/Week1-Data-Scraping.ipynb



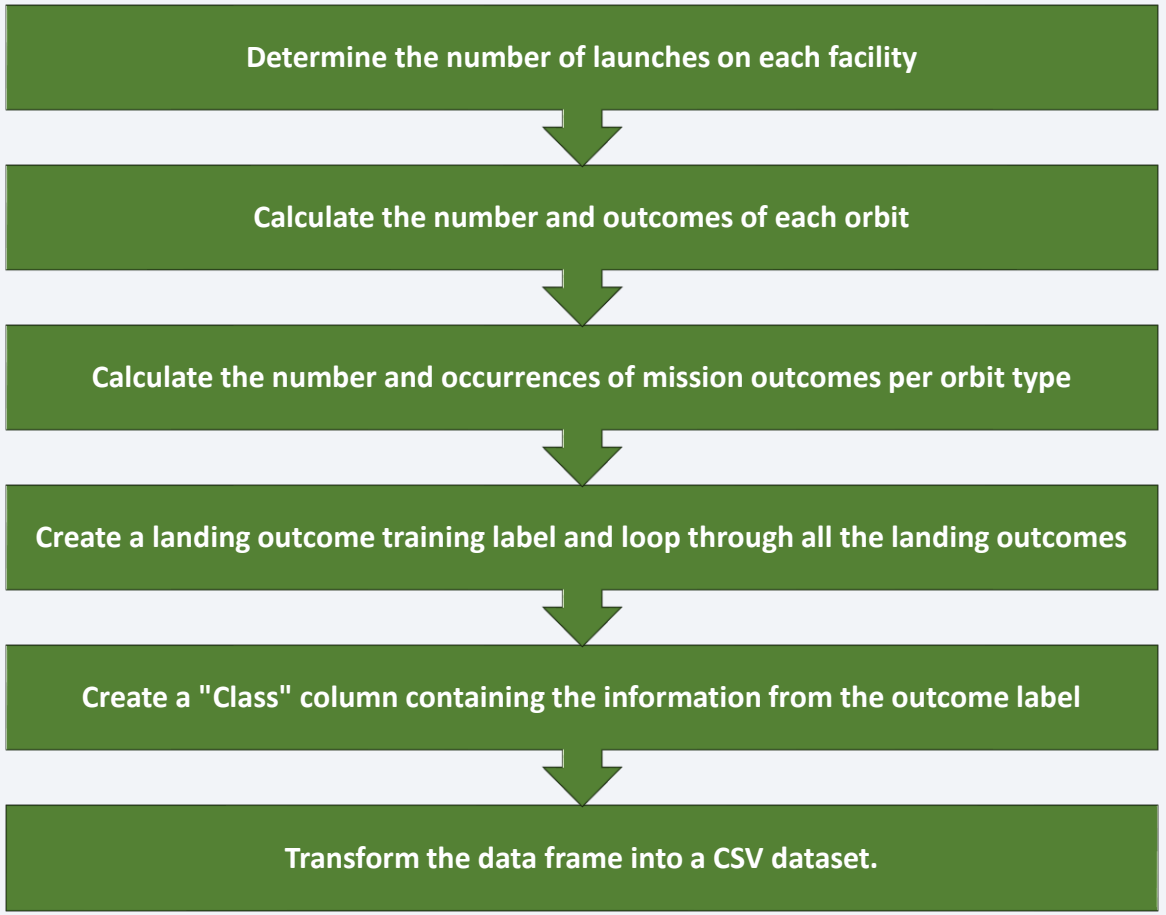
Data Wrangling

At this point, it was necessary to find patterns in the data that would determine the parameters for training supervised machine learning models.

The data set contained several different cases where the first stage rocket did not land successfully.

This descriptive information was converted into Training Labels, with ‘1’ denoting the rocket landed successfully while ‘0’ denoted an unsuccessful landing.

URL link: https://github.com/Vision-City/SpaceX-Project/blob/main/notebooks/Week1-Data-Wrangling.ipynb

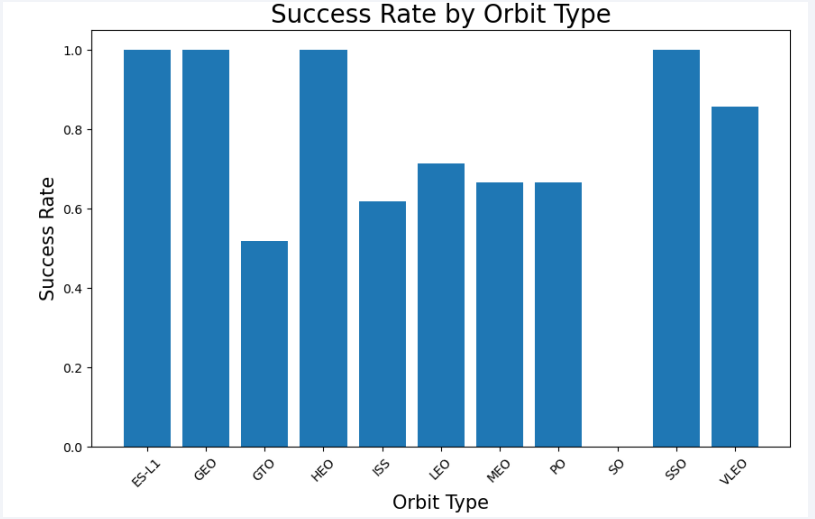


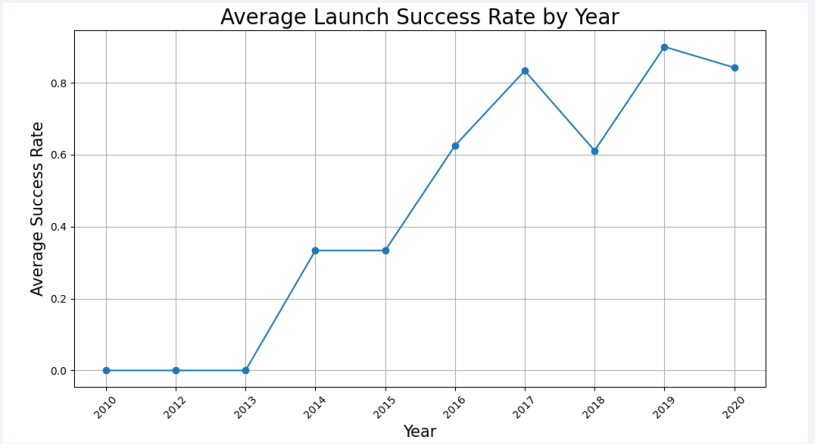
EDA with Data Visualization

Data visualization is essential for understanding data by organizing it into a form that's easier to understand, so that trends and outliers can be understood.

- Cat plots and scatter plots were used to view the relationships of categorical variables like Launch Site and Orbit.
- A bar chart was used to visualize the success rate of each orbit type.
- A line chart was used to visualize the launch success yearly trend.

URL link: https://github.com/Vision-City/SpaceX-Project/blob/main/notebooks/Week2-EDA-with-visualization.ipynb





EDA with SQL

To evaluate the data, several SQL queries were used to perform the following:

- Display the names of the unique launch sites in the space mission
- Assess the payload mass with boosters launched by NASA (CRS)
- Determine average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing occurred
- List the names of the boosters which had successful drone ship landings with specific payload ranges
- Determine the dates of successful and failed landing outcomes
- Rank landing outcomes

Total_Payload_Mass
45596

Average_Payload_Mass
2928.4

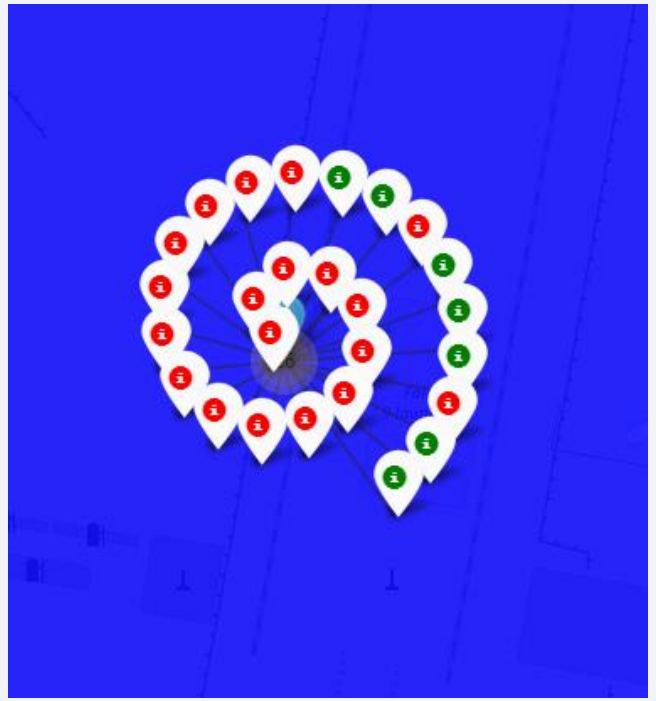
First_Successful_Ground_Pad_Landing
2015-12-22

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

SQL queries can be seen at this URL link:
https://github.com/Vision-City/SpaceX-Project/blob/main/notebooks/Week2-EDA-with-SQL.ipynb

Build an Interactive Map with Folium

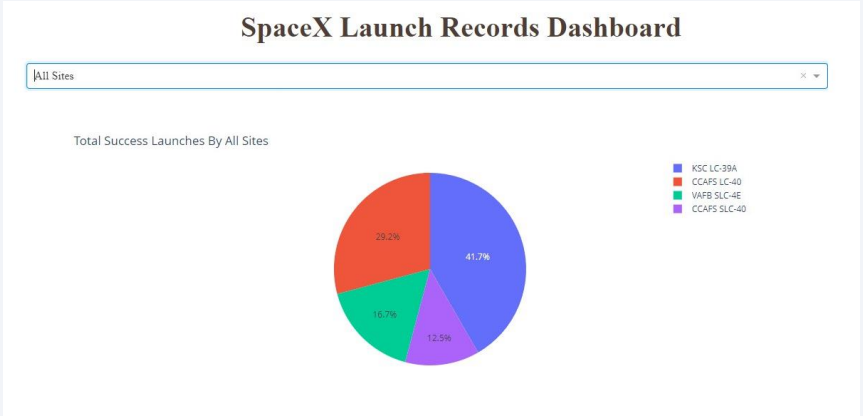
Launch success rate may depend on the elements like location and proximity of a launch site. **Folium Interactive Map was used for visualizing and analyzing SpaceX Launch Sites.**

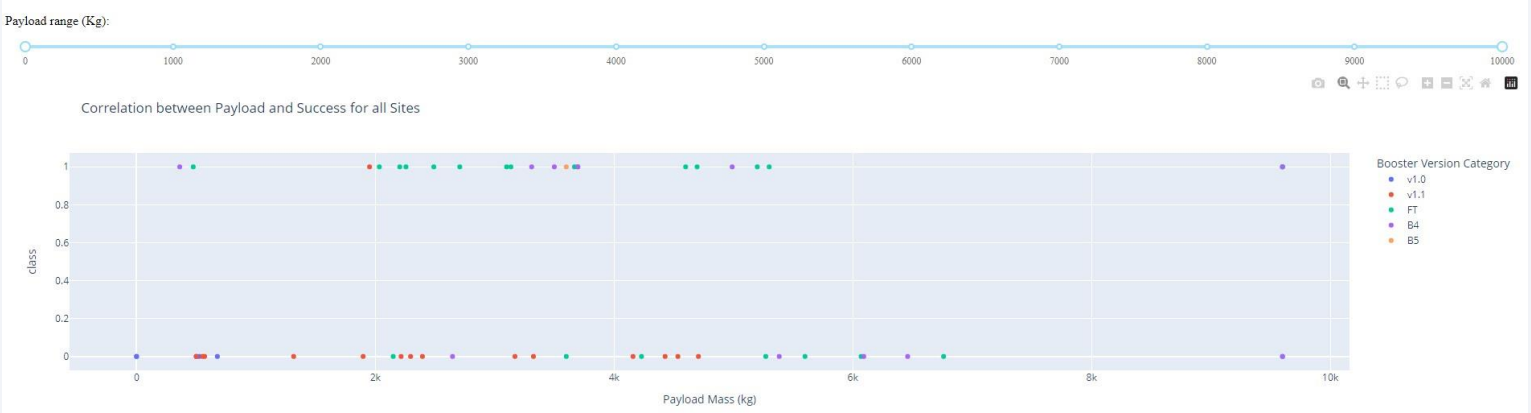


- Folium Markers were plotted to show the SpaceX launch sites and nearby important landmarks like railways, highways, cities and coastlines.
- Polylines were used to connect the launch sites to landmarks.
- Folium Circles were used to highlight launch sites areas.
- Marker clusters were used on the map to mark the successful or failed launches for each launch site
- URL link: https://github.com/Vision-City/SpaceX-Project/blob/main/notebooks/Week3-Folium-Visual-Analytics.ipynb

Build a Dashboard with Plotly Dash

- Pie charts and scatter charts were used to visualize the launch records of SpaceX.
- These charts displayed the rocket launch success rate per launch site. This helped visualize and understand the factors that influence success rates at each site, like payload mass and booster versions.





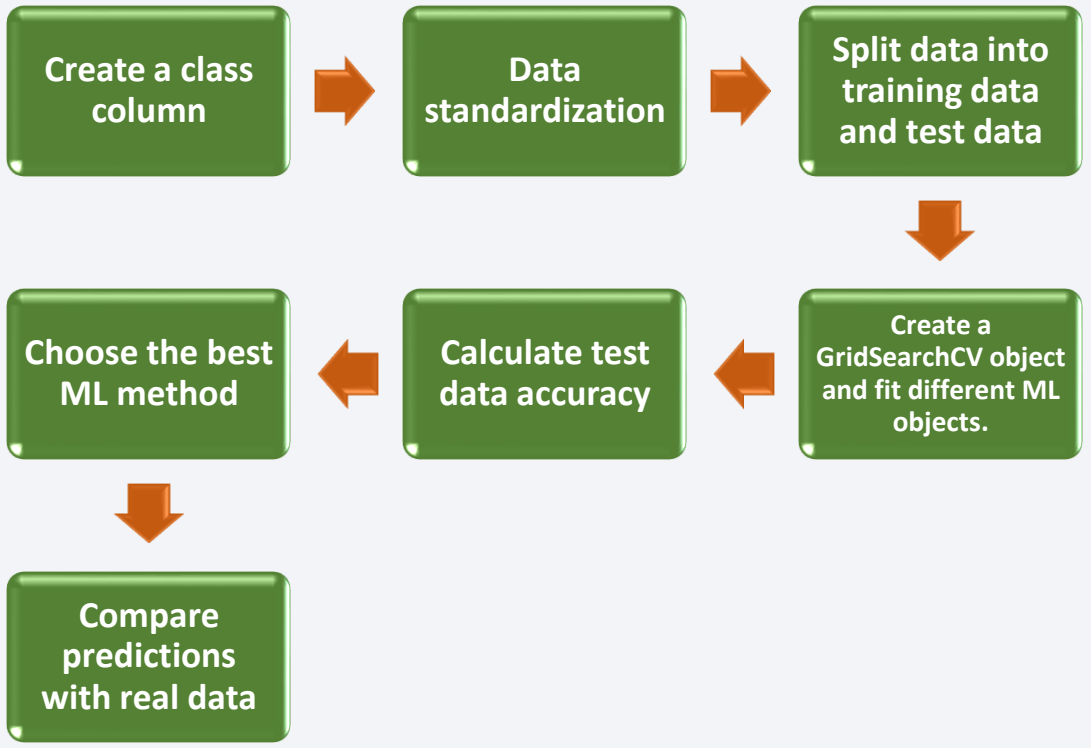
URL link: https://github.com/Vision-City/SpaceX-Project/blob/main/notebooks/Week3-Dashboard-with-Plotly.py

Predictive Analysis (Classification)

Scikit-learn machine learning library was used for predictive analysis. The following steps were followed:

- Created a machine learning pipeline to predict if the first stage will land given the data.
- The best ML method was determined using *GridSearchCV.*
- Predictions were compared with real data.
- The Decision Tree model scored the best accuracy of 87.5%

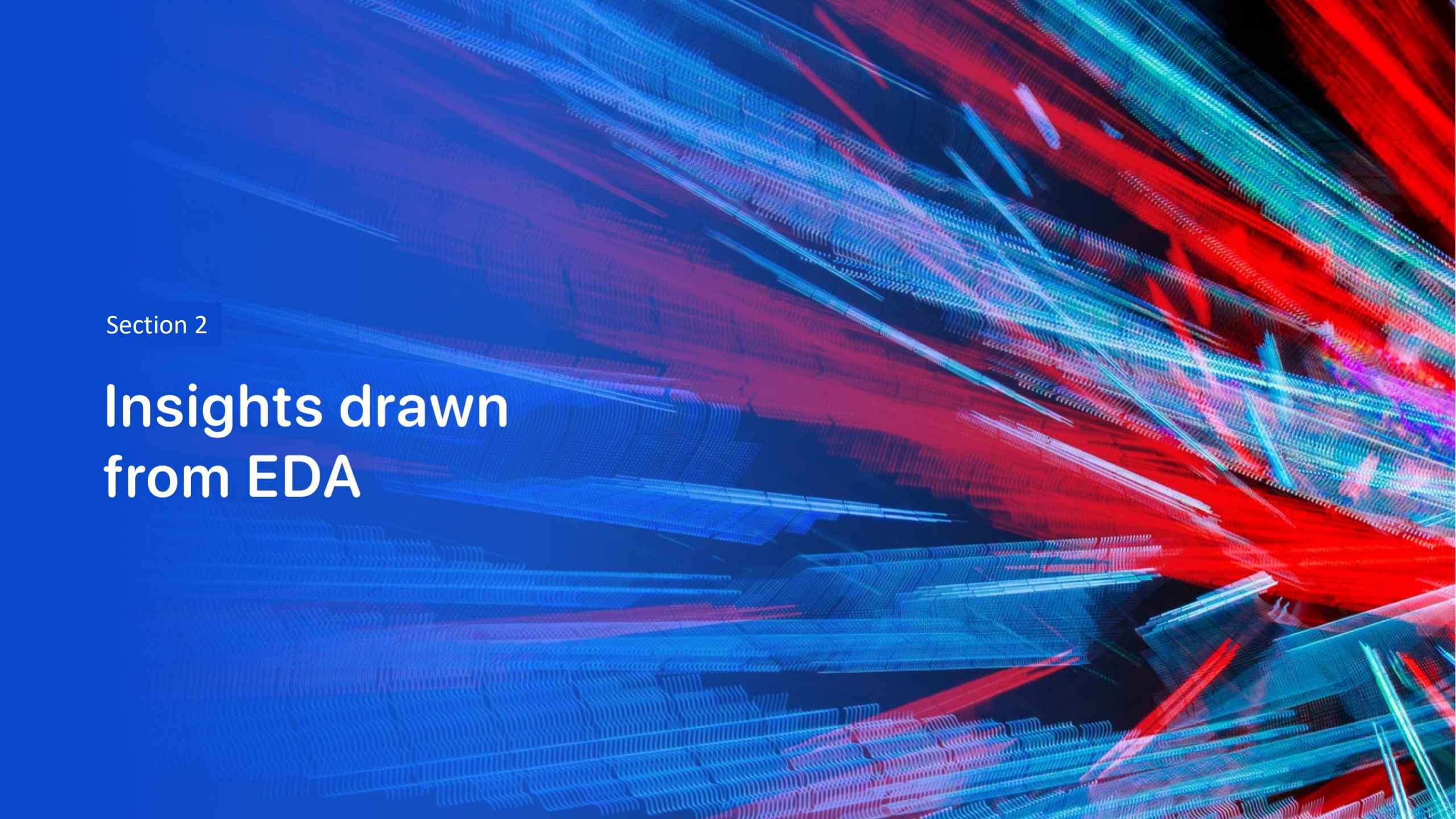
URL link: https://github.com/Vision-City/SpaceX-Project/blob/main/notebooks/Week4-Machine-Learning.ipynb



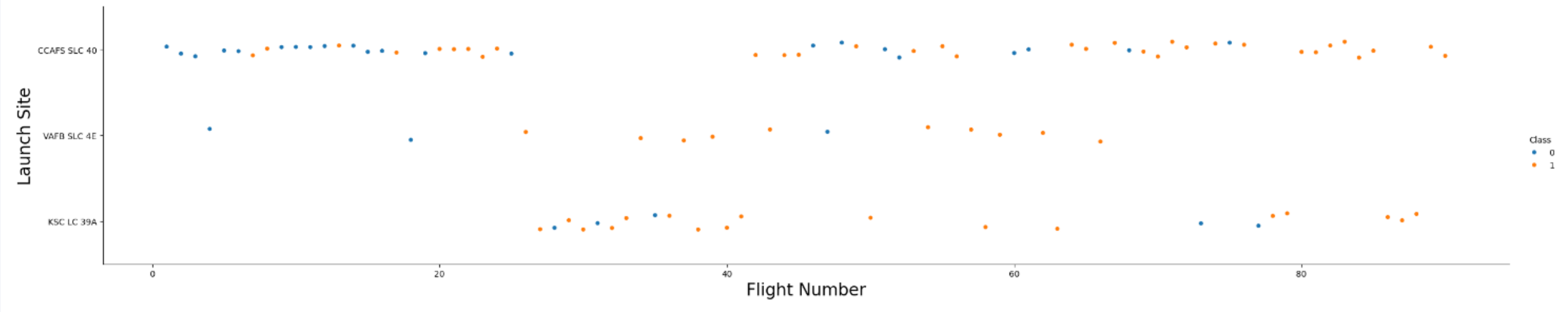
Results

On the following pages, we will go into greater detail regarding:

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

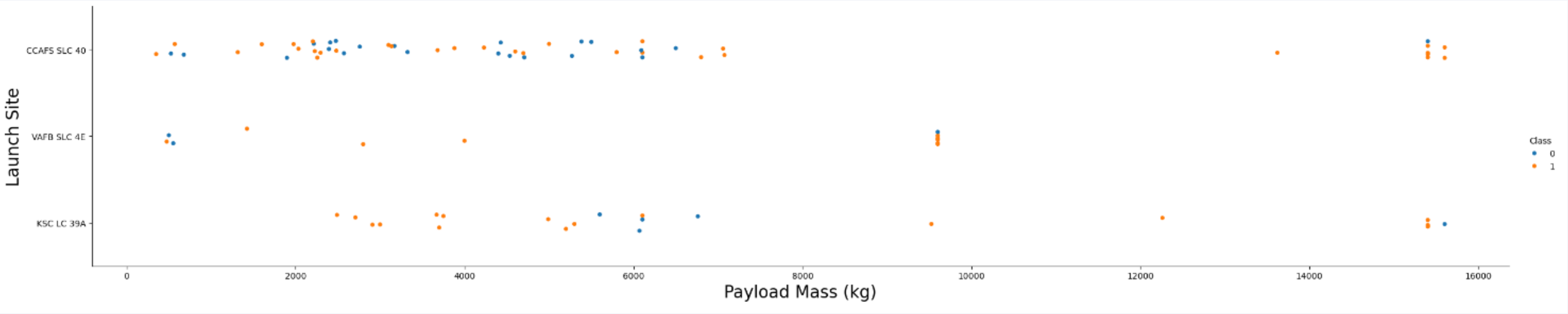

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- For all scatter plots, the blue dots (0) represent failed flights. The orange dots (1) represent successful flights.
- There were more successful flights as the flight numbers increased for all launch sites.
- Launch site **CCAFS SLC 40** had the greatest number of successes while the site **VAFB SLC 4E** had the least number of successes.

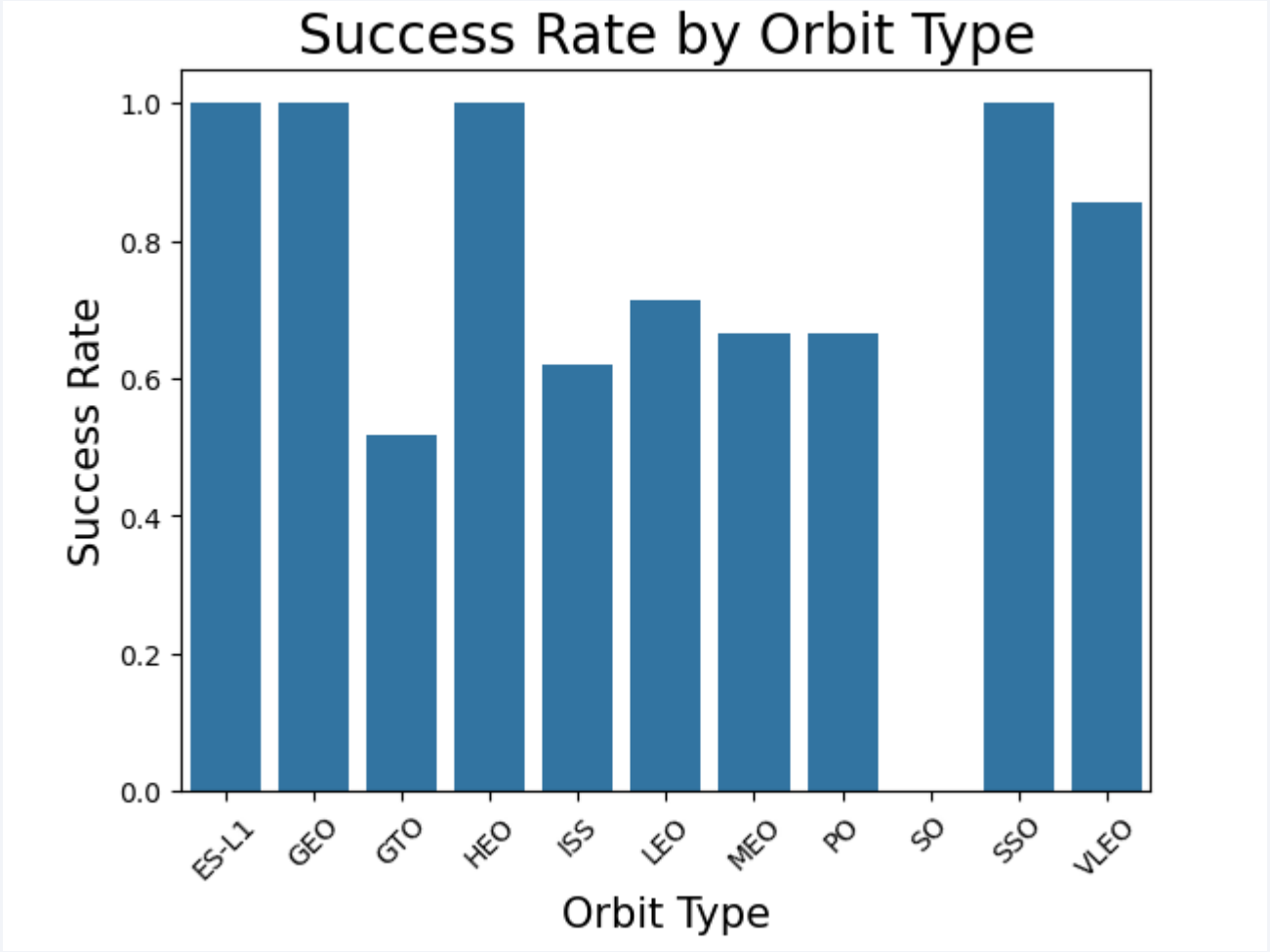
Payload vs. Launch Site



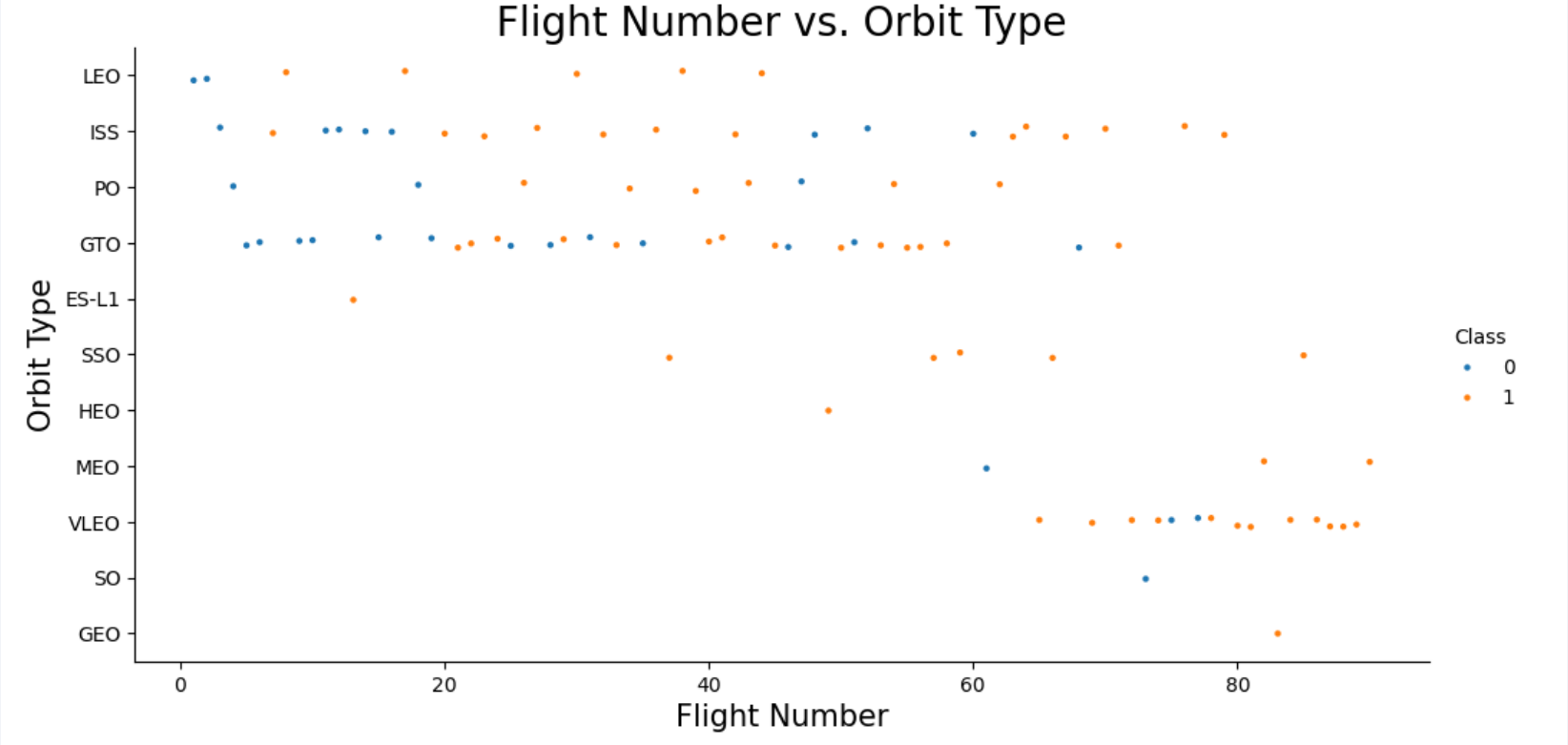
- For Falcon9 launches, heavy payloads (> 10,000 kg) are sent to low/medium orbits only.
- It looks like the percentage of failures is lower for heavy payloads, indicating that low orbits are less risky to the success of the mission (recovery of booster).

Success Rate vs. Orbit Type

- The orbit types ES-L1, GEO, HEO and SSO had the highest success rates.

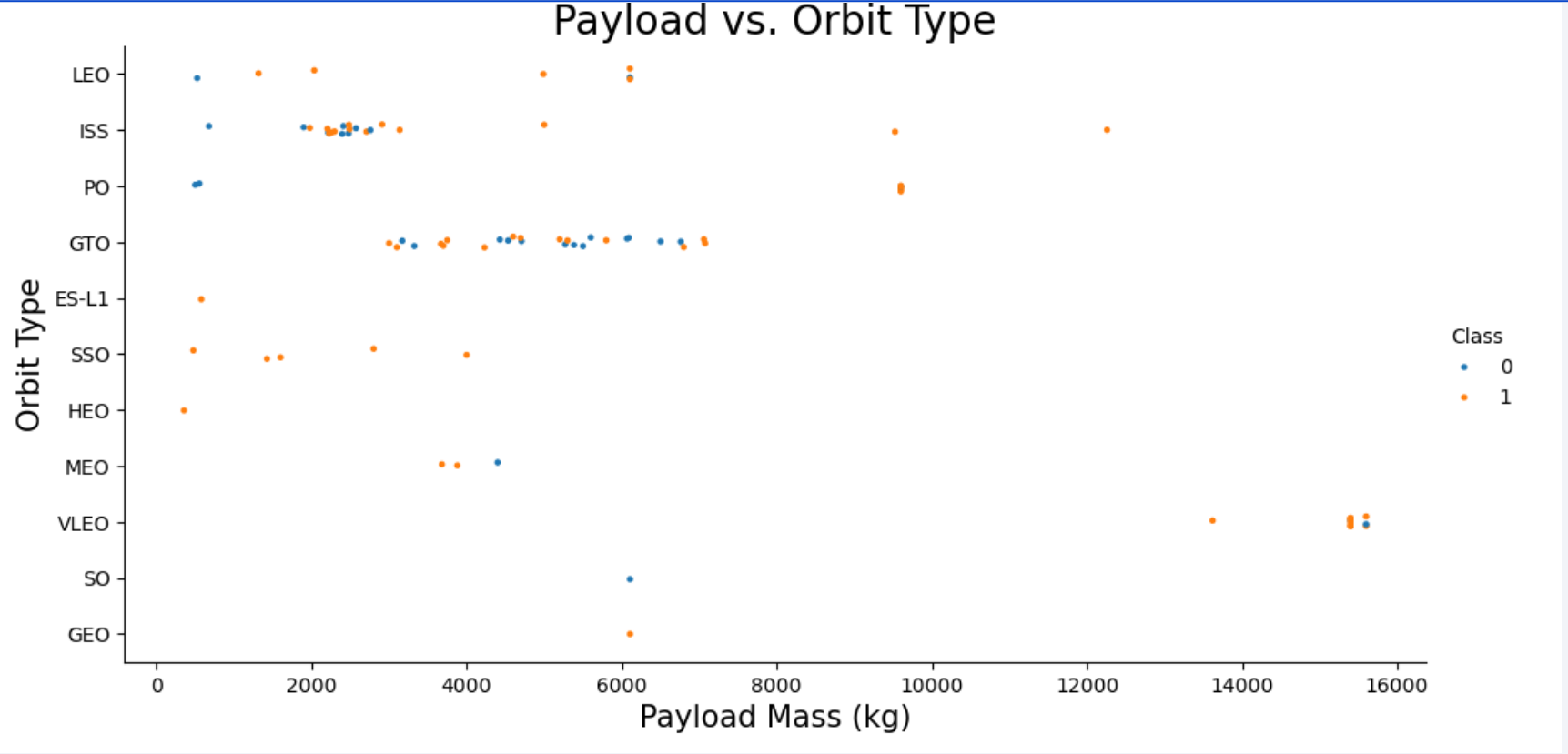


Flight Number vs. Orbit Type



- The plot above shows the Flight Number vs. Orbit type.
- For the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there appears to be no relationship between flight number and the orbit.

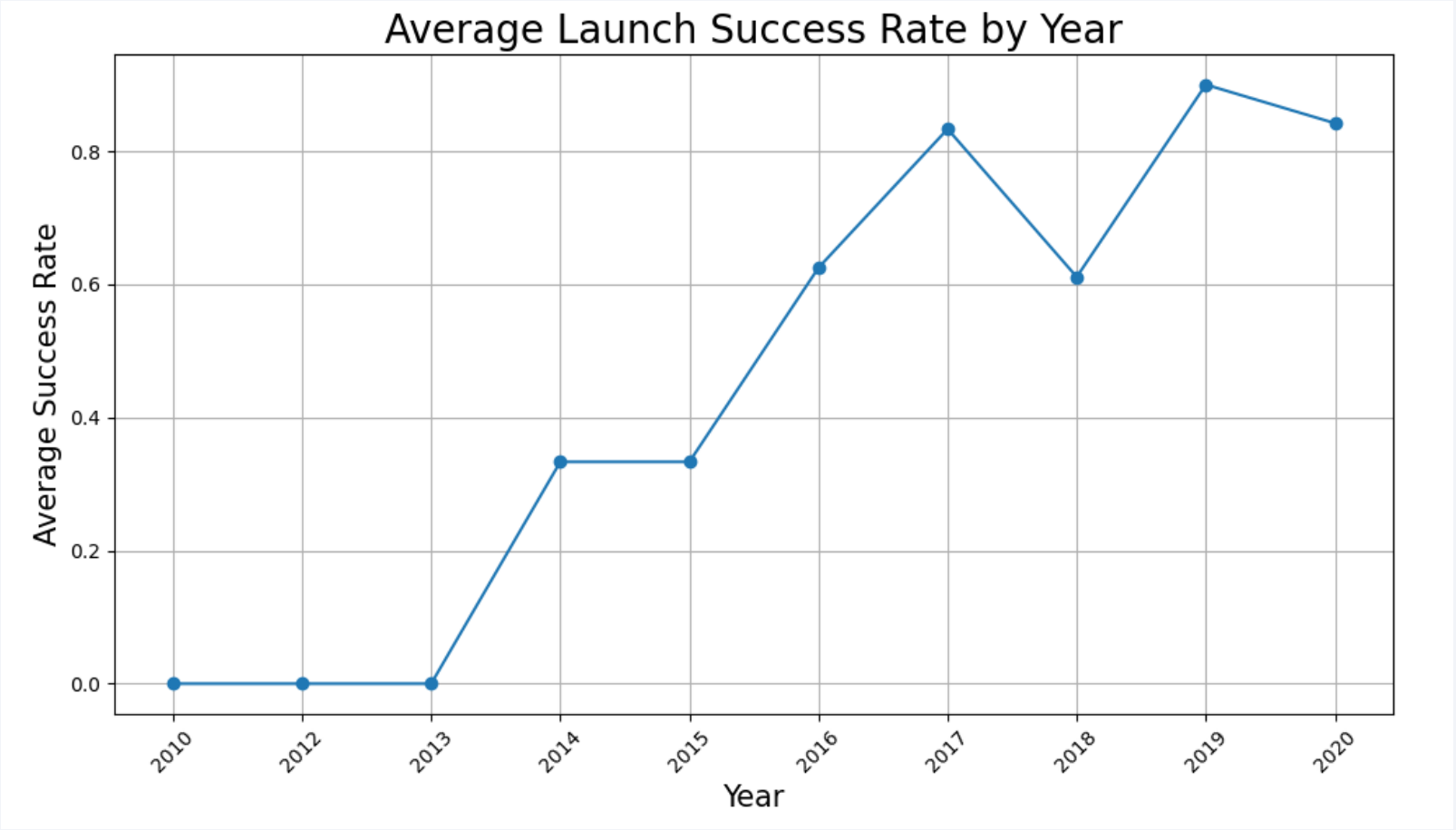
Payload vs. Orbit Type



- When we look at heavy payloads (> 10,000 kg), there are more successful landings for PO, LEO and ISS orbits.
- For GTO orbits, there is no observable pattern to successful landings.

Launch Success Yearly Trend

- This chart illustrates the launch success rate has increased from 2013 to 2020



All Launch Site Names

- We used the key word DISTINCT to filter out only unique launch sites from the SpaceX data.

```
[12]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;
 * sqlite:///my_data1.db
Done.
[12]: ,,,,,,,,,,,,,,,,,,,,
```

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

Launch Site Names Begin with 'CCA'

```
[13]: %%sql
SELECT * FROM SPACEXTBL
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5;
```

```
 * sqlite:///my_data1.db

Done.
```

[13]: ''

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- The query shown above was used to display 5 records where launch sites begin with ‘CCA’

Total Payload Mass

```
[14]: %%sql
      SELECT SUM(PAYLOAD_MASS__KG_) as Total_Payload_Mass
      FROM SPACEXTBL
      WHERE Customer = 'NASA (CRS)';

       * sqlite:///my_data1.db

      Done.
[14]: ,,,,,,,,,,
```

Total_Payload_Mass
45596

- Using the above query, the total payload carried by boosters for NASA (CRS) was determined to be 45596 kg

Average Payload Mass by F9 v1.1

```
[15]: %%sql
      SELECT AVG(PAYLOAD_MASS__KG_) as Average_Payload_Mass
      FROM SPACEXTBL
      WHERE Booster_Version = 'F9 v1.1';

       * sqlite:///my_data1.db

      Done.
[15]: ,,,,,,,,,,
```

Average_Payload_Mass
2928.4

- Using the above query, the average payload mass carried by booster F9 v1.1 was found to be 2928.4 kg

First Successful Ground Landing Date

```
[16]: %%sql
      SELECT MIN(Date) as First_Successful_Ground_Pad_Landing
      FROM SPACEXTBL
      WHERE Landing_Outcome = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

```
[16]: ''
```

First_Successful_Ground_Pad_Landing
2015-12-22

- The above query was used to determine the first successful ground landing date as December 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- In this query, the WHERE clause was used to filter for boosters which successfully landed on drone ships and applied the AND condition to determine successful landing with payload mass greater than 4000 kg but less than 6000 kg
- This payload range seemed to be ideal for successful drone ship landings

[17]:
```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ > 4000
AND PAYLOAD_MASS__KG_ < 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

[17]:

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

Total Number of Successful and Failed Mission Outcomes

[18]:
```
%%sql
SELECT
    SUM(CASE WHEN Mission_Outcome LIKE 'Success%' THEN 1 ELSE 0 END) AS Total_Success,
    SUM(CASE WHEN Mission_Outcome LIKE 'Failure%' THEN 1 ELSE 0 END) AS Total_Failure
FROM SPACEXTBL;
```

```
 * sqlite:///my_data1.db

Done.
```

[18]:

Total_Success	Total_Failure
100	1

- Using the above query, it was determined that 100 missions were successful and only 1 was a failure. Because of the nature of success and failure entries in the data, the '%' wildcard was needed to focus on only the phrases "Success" and "Failure" while eliminating any other descriptive elements.

Boosters Carried Maximum Payload

- Using the query and subquery below, the results on the right illustrate the booster versions carrying the maximum payload.

```
[20]: %%sql
SELECT Booster_Version, PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL
);

 * sqlite:///my_data1.db
```

[20]:

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

- From the query on the right, it was determined that two boosters failed in the first 4 months of 2015.
- From previous data, we also know that in December of 2015 the first successful landing took place.

[21]:
```
%%sql
SELECT
    SUBSTR(Date, 6, 2) AS month,
    Booster_Version,
    Launch_Site,
    Landing_Outcome
FROM SPACEXTBL
WHERE
    SUBSTR(Date, 1, 4) = '2015'
    AND Landing_Outcome LIKE 'Failure (drone ship)%';
```

```
 * sqlite:///my_data1.db

Done.
```

[21]:

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

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

- The code shown to the right created a table of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
- The GROUP BY clause was used to group the landing outcomes, and the ORDER BY clause was used to order the grouped landing outcomes in descending order of frequency

[22]:
```
%%sql
SELECT Landing_Outcome, COUNT(Landing_Outcome) AS Outcome_Count
FROM SPACEXTBL
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Outcome_Count DESC;
```

```
 * sqlite:///my_data1.db

Done.
```

[22]:

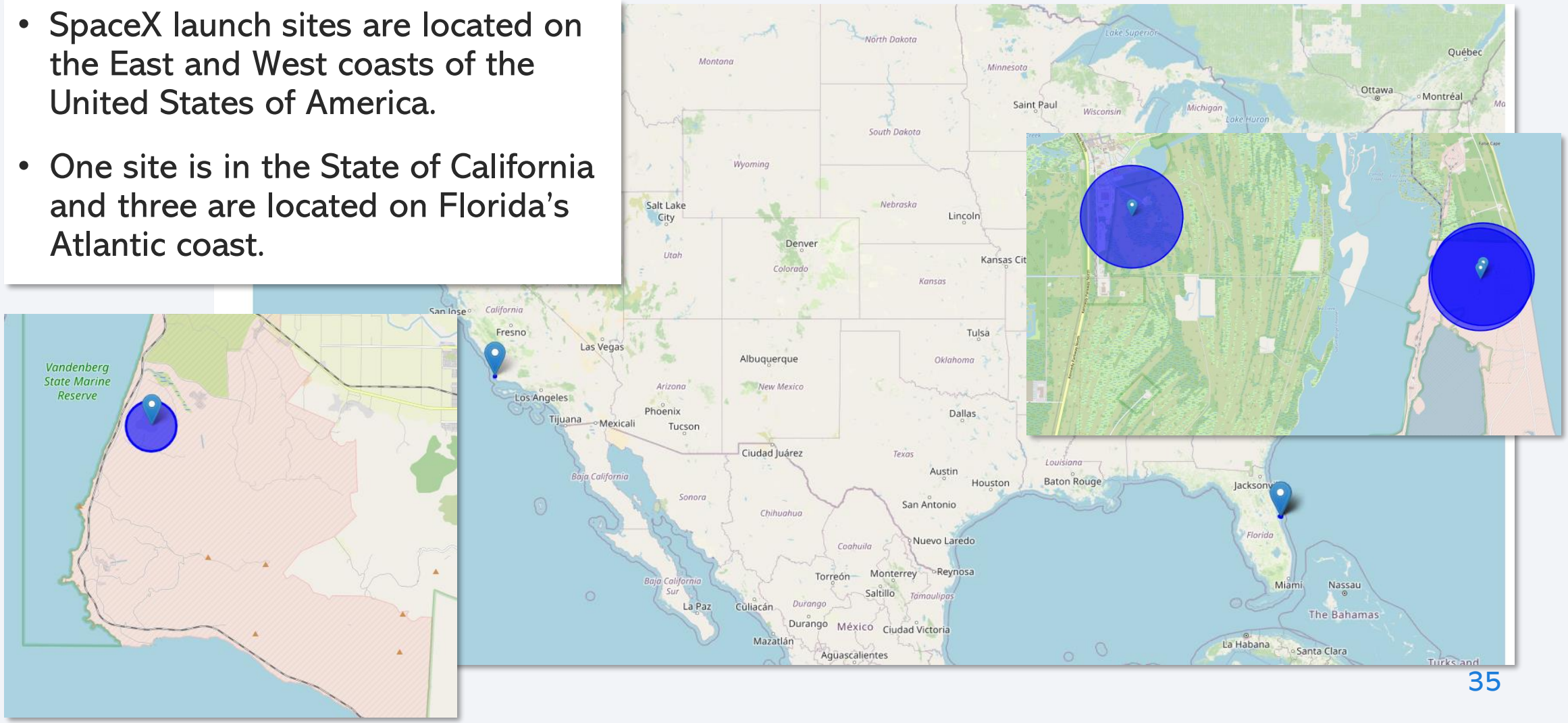
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

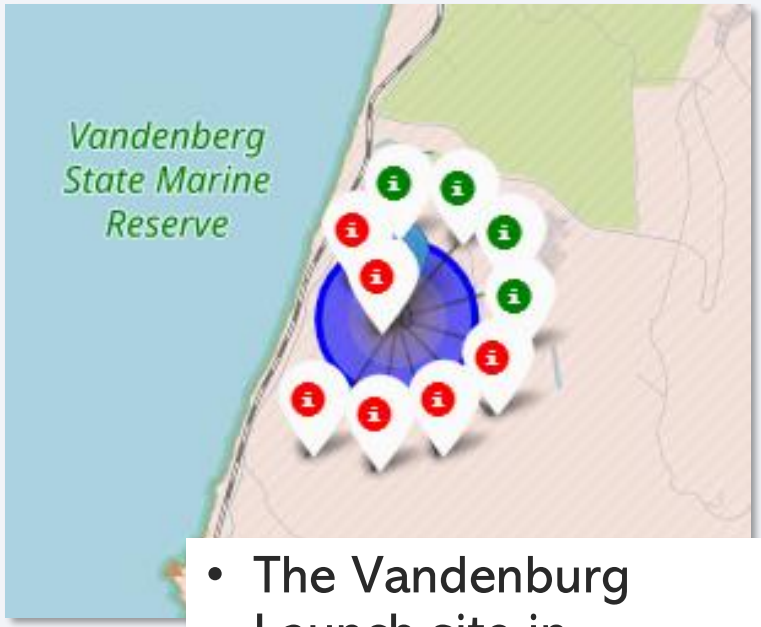
Launch Sites Proximities Analysis

All Launch Sites Global Map Markers

- SpaceX launch sites are located on the East and West coasts of the United States of America.
- One site is in the State of California and three are located on Florida's Atlantic coast.



Marker Clusters Showing Successes & Failures per Site

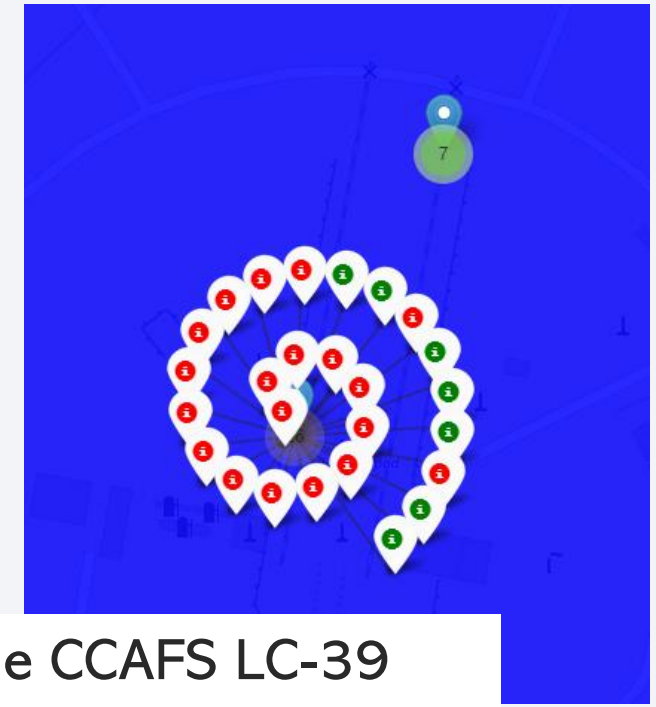


- The Vandenburg Launch site in California shows the 6 failures and 4 successful launches with marker clusters

- The KSC LC-39A Launch site in Florida shows the 3 failures and 10 successful launches with marker clusters

Successful launches are represented by green markers while failed launches are represented by the red markers

Marker Clusters Showing Successes & Failures per Site



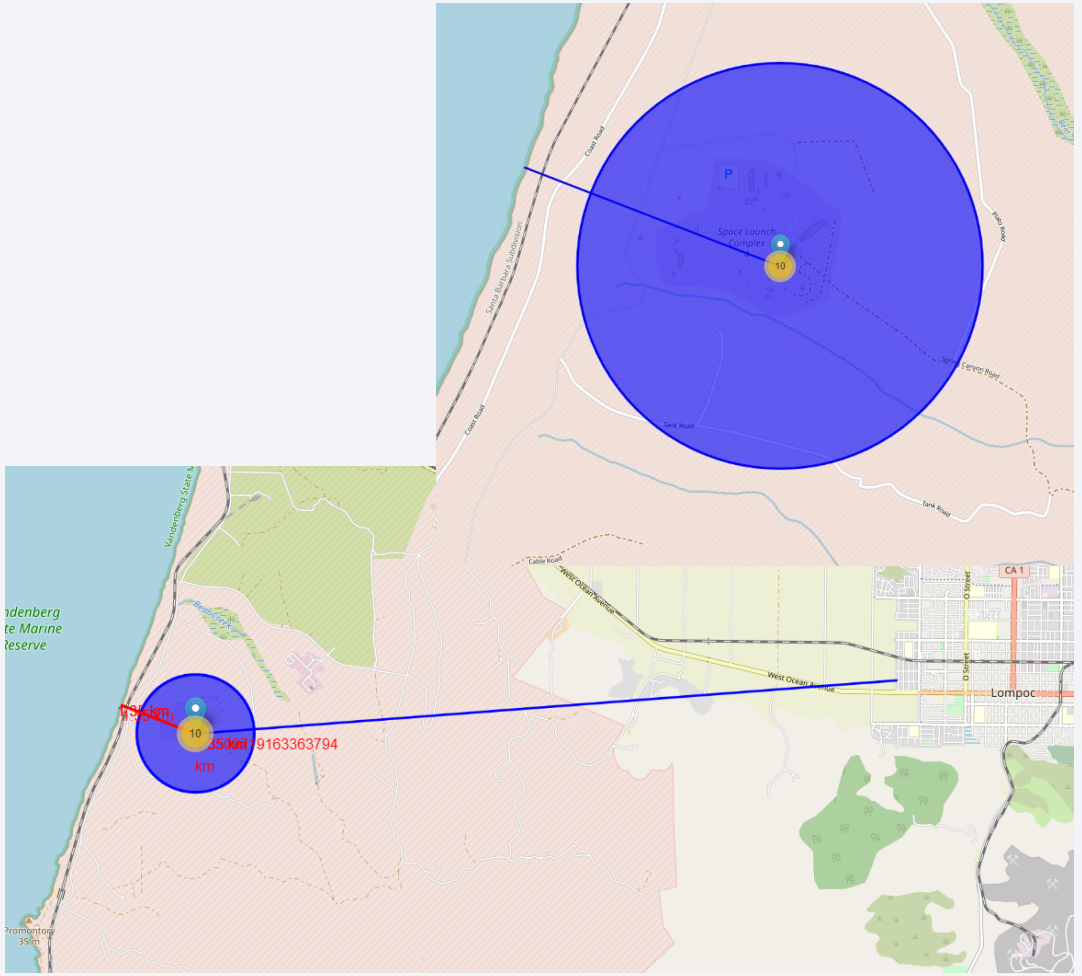
- The CCAFS LC-39 Launch site shows the 19 failures and 7 successful launches with marker clusters

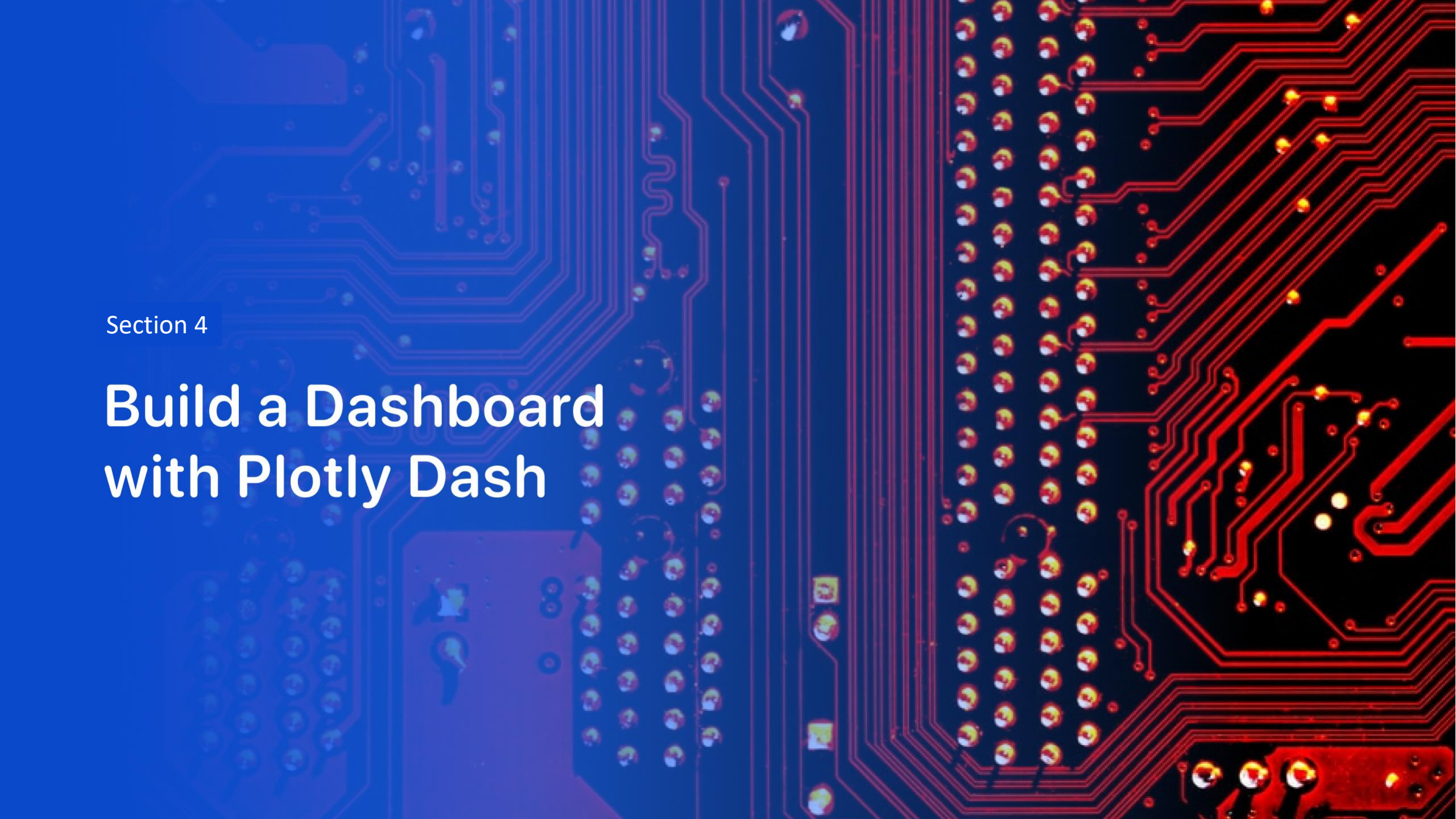
- The CCAFS SLC-40 Launch site shows the 4 failures and 3 successful launches with marker clusters

Successful launches are represented by green markers while failed launches are represented by the red markers

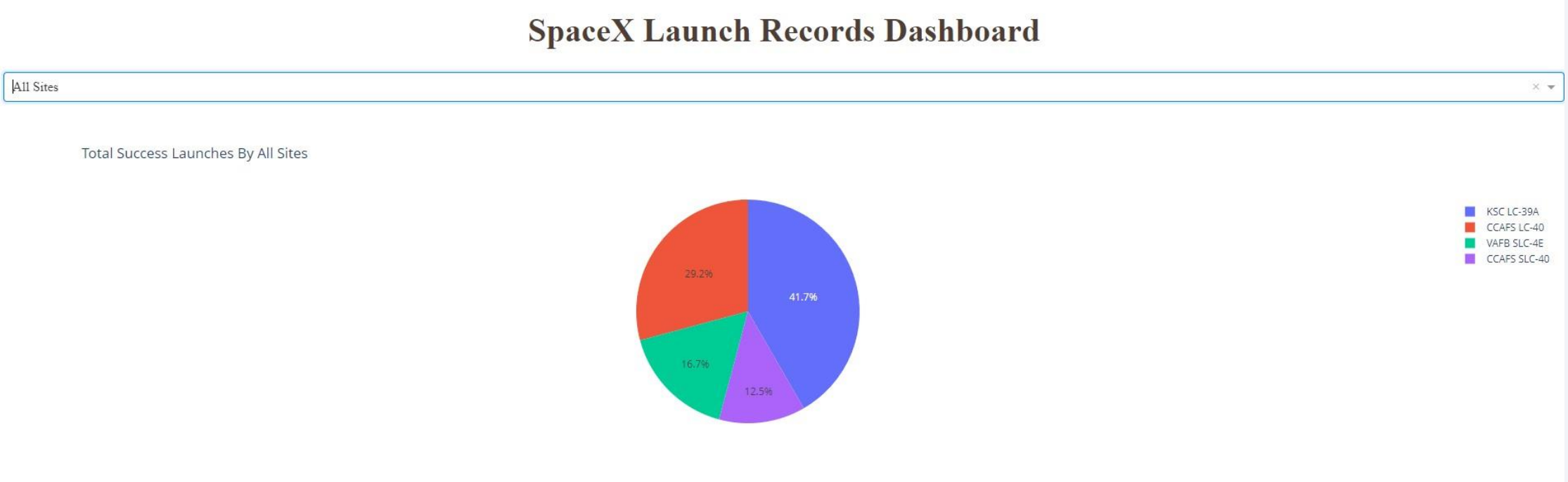
Launch Site Distance to Landmarks

- Upon examining launch sites, it seems that they are close to coastlines and railways. Launching towards the ocean is a safety precaution in case a rocket needs to be terminated in flight. Railways are helpful for moving large components to the launch site.
- The launch locations are more removed from highways and cities as safety precautions.



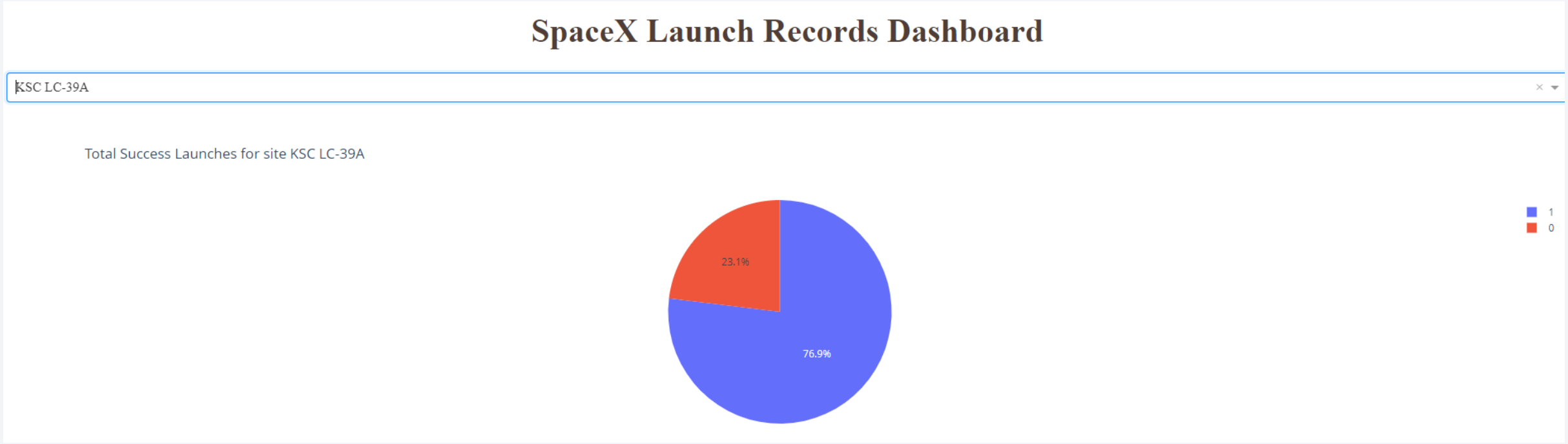

Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Site



- All successful launches by site are shown in the above pie chart
- KSC LC39A has the largest number of successful launches as well as the highest launch success rate.
- More analysis may be needed to understand this correlation.

Total Successful Launches for Site KSC LC-39A



- The launch site with highest launch success ratio is KSC LC-39A
- The data shows 76.9% of the total launches at this location were successful, making this the highest-success launchpad of all sites.

Payload Mass vs. Launch Success for All Sites

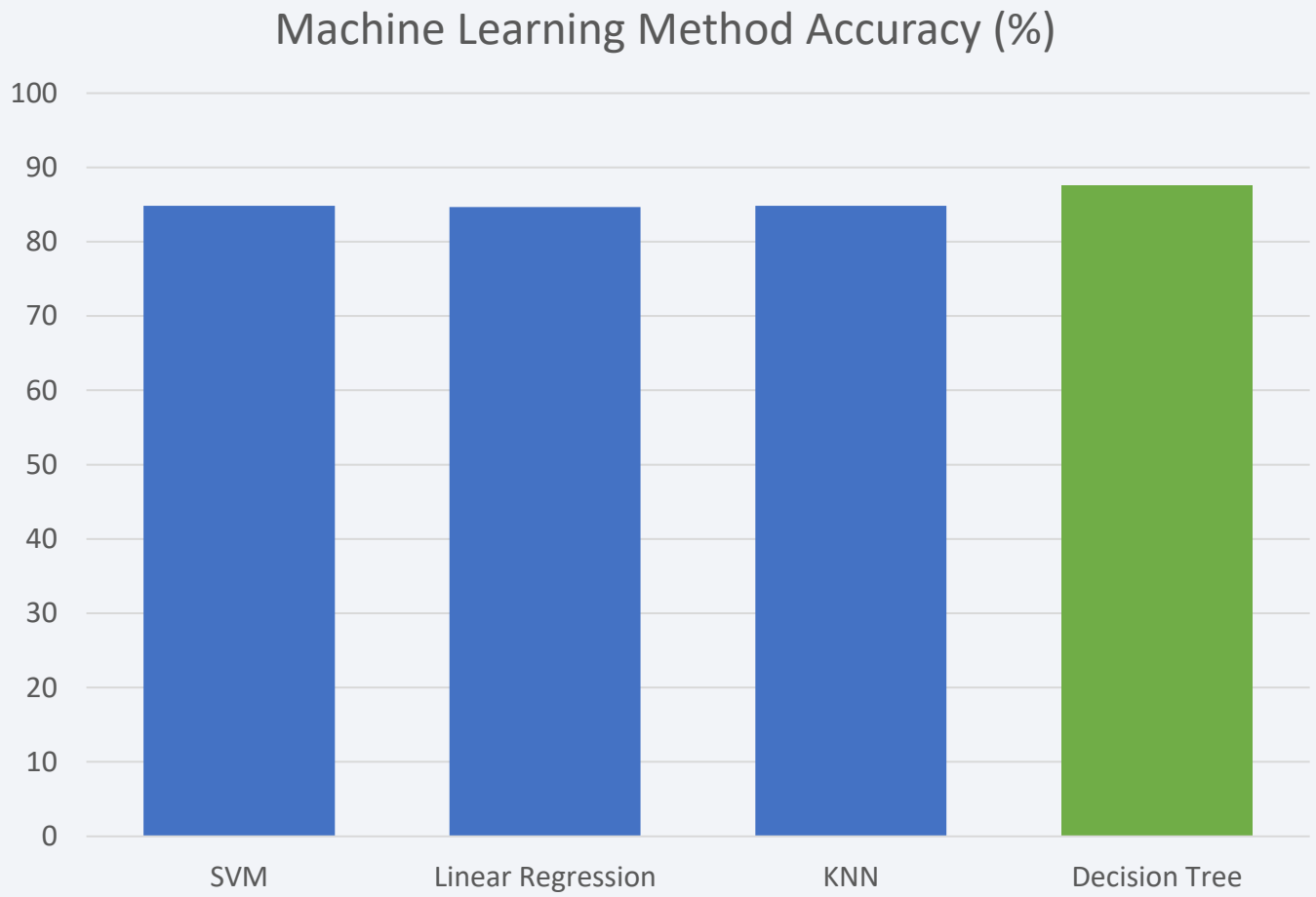
- The payload range between 2000 kg and 4000 kg has the highest success rate.
- The launch success rate was very low between the payload range of 0kg and 2500kg. It seems to indicate that very low masses lower launch success.
- The booster version **FT**, represented by green dots, has a higher success rate than other boosters

Section 5

Predictive Analysis (Classification)

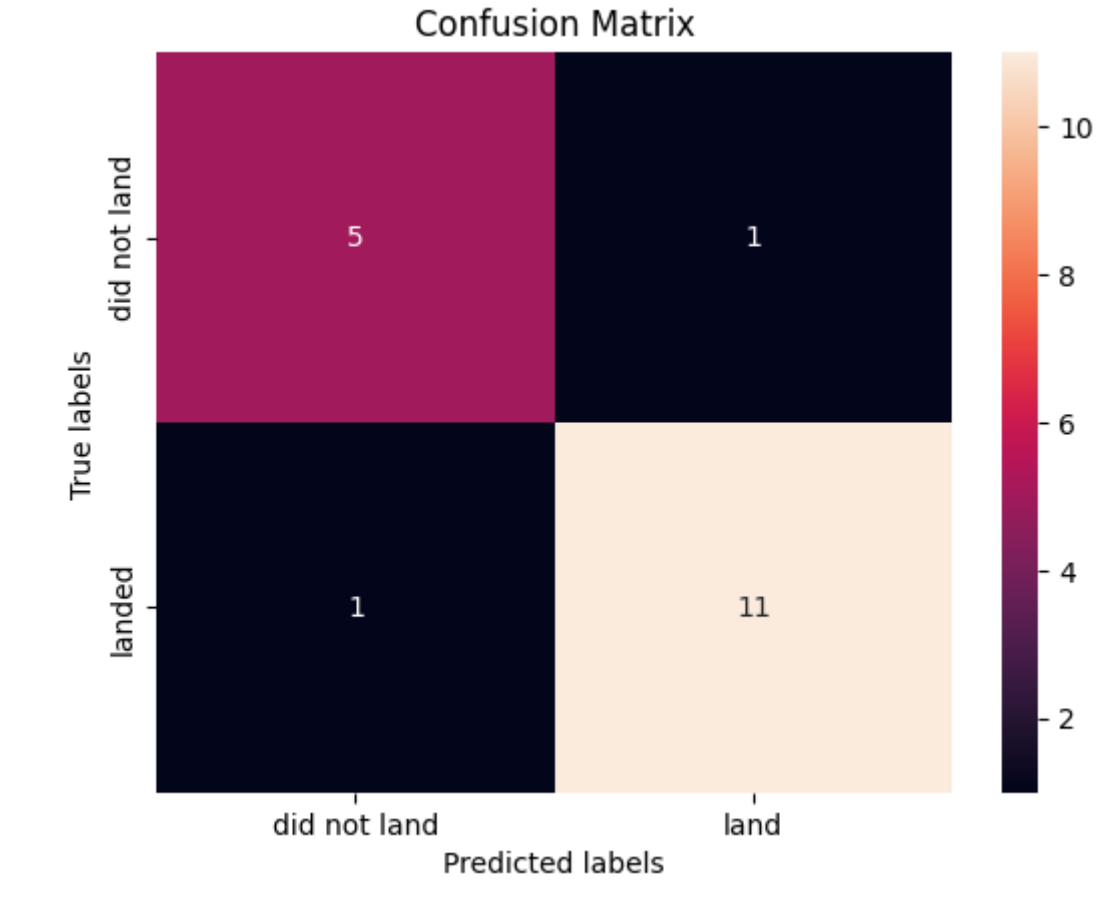
Classification Accuracy

- Of the four methods evaluated, the Decision Tree had the best classification accuracy score of 87.5% and was selected.



Confusion Matrix

Below is the confusion matrix of the best performing model – the decision tree classifier. This resulted in the best output – failing only once to make an accurate prediction.



Decision trees are highly interpretable and can handle complex, non-linear relationships by splitting data based on feature thresholds. They excel at modeling decision rules, which aligns well with the structured and rule-based nature of launch success criteria.

Rocket launches are impacted by a multitude of complex factors. Decision trees can capture these interactions through hierarchical feature splits, whereas logistic regression assumes independence among features and KNN relies on local similarity, which might not capture the global decision boundaries as effectively.

Another factor that was important in this case was that datasets containing many categorical variables or missing values, can be handled by decision trees without extensive preprocessing required by other models.

Conclusions

In order to understand the competitive nature of SpaceX, it was essential to analyze their launch data. Through this process, a general understanding of their success emerged.

- All their launch sites are located near the coast, away from nearby cities. This enables to them to test their rocket landings without much interference. The safety of launching over water and away from populated areas is also an essential element to rocket launches.
- Site KSC LC-39A had the highest launch success rate out of all the launch sites.
- From 2015 onwards, the success rate of rocket landings significantly increased. This indicated that SpaceX had been learning from every launch, whether it was a success or failure.

All this data was used to train a Decision Tree classifier that can predict the landing outcome of rocket launches with 87.5 % accuracy.

The knowledge gained through this analysis will allow our company to make launch offers that are competitive with SpaceX, due to the strong certainty to the outcomes of the model we have developed. This is a winning business advantage and a distinct benefit for our investors and customers.

Appendix

- IBM. *Data Science Professional Certificate.* *https://www.coursera.org/professional-certificates/ibm-data-science*
- Tom Scherbluk. *Applied Data Science Capstone Project* : https://github.com/Vision-City/SpaceX-Project/tree/main
- *Space.com. SpaceX Lands Orbital Rocket Successfully in Historic First.* SpaceX. *REST API. https://api.spacexdata.com/v4/*
- Wikipedia. *List of Falcon 9 and Falcon Heavy launches. https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922*
- Markets and Markets. *Market Report - Satellite Launch Vehicle Market.* https://www.marketsandmarkets.com/Market-Reports/satellite-launch-vehicle-market-115959224.html


Thank You!