IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ionatan Ben Shalom
30/10/2024

Outline

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

Executive Summary

Summary of Methodologies

The research aims to determine the factors contributing to a successful rocket landing. To achieve this, the following methodologies were employed:

- **Data Collection**: Used SpaceX REST API and web scraping.
- **Data Wrangling**: Cleaned and transformed data to create success/fail outcomes.
- **Exploratory Data Analysis**: Explored factors like payload size, launch site, and trends using SQL. Visualized success rates with charts.
- **Interactive Visual Analytics**: Used Folium for mapping and Plotly Dash for interactive dashboards.
- **Predictive Analysis**: Built models (logistic regression, SVM, decision tree, KNN) to predict outcomes. Decision tree model performed best.

Summary of Results

- **Exploratory Data Analysis:**
  - Launch success rates have increased over time
  - KSC LC-39A has the highest success rate among landing sites
  - Orbits ES-L1, GEO, HEO, and SSO have achieved a 100% success rate
- **Visualization\Analytics**: Most launch sites are located near the equator, and all are close to the coast.
- **Predictive Analytics**: All models performed similarly on the test set, with the decision tree model slightly outperforming the others

Introduction

Background

SpaceX, a leader in the space industry, aims to make space travel accessible and affordable for everyone. Its achievements include sending spacecraft to the International Space Station, deploying a satellite constellation to provide internet access, and conducting manned missions to space. SpaceX can do this because its rocket launches are relatively inexpensive ($62 million per launch) due to the reuse of the first stage of its Falcon 9 rocket. In contrast, other providers that cannot reuse the first stage face launch costs exceeding $165 million each. By predicting whether the first stage will land successfully, we can estimate the launch cost. To achieve this, we leverage public data and ML models to predict if SpaceX, or a competitor, can reuse the first stage.

Research

How payload mass, launch site, number of flights, and orbits influence first-stage landing success

Trends in the rate of successful landings over time

Identifying the best model for predicting successful landings

Section 1

Methodology

Methodology

- Data collection methodology:
  - Data was collected using SpaceX REST API and web scraping techniques
- Perform data wrangling
  - By filtering the data, handling missing values, and applying one-hot encoding – I prepared the data for analysis and modelling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building models (logistic regression, SVM, decision tree, KNN) to predict outcomes by selecting, tuning, and evaluating each model's performance.

Data Collection

The Process

- Request data from SpaceX API (rocket launch data)
- Decode response using .json() and convert to a dataframe using .json_normalize()
- Request information about the launches from SpaceX API using custom functions
- Create dictionary from the data
- Create dataframe from the dictionary
- Filter dataframe to contain only Falcon 9 launches
- Replace missing values of Payload Mass with calculated .mean()
- Export data to CSV file

Data Collection – SpaceX API

Data Collection with SpaceX REST API Steps

Data Retrieval: Use SpaceX API endpoints to obtain various launch details, such as booster information, payload, launch sites, and core data.

Extract Key Information: Specific functions are defined to pull targeted details:

- Booster: Type of booster used (e.g., Falcon 1, Falcon 9).
- Payload: Mass and orbit details.
- Launch Site: Name, longitude, and latitude.
- Core Data: Landing outcome, landing type, number of flights, grid fins, reuse status, core block version, and serial.

Variable Assignment: Initialize global variables to store each extracted detail. Populate variables by calling the extraction functions.

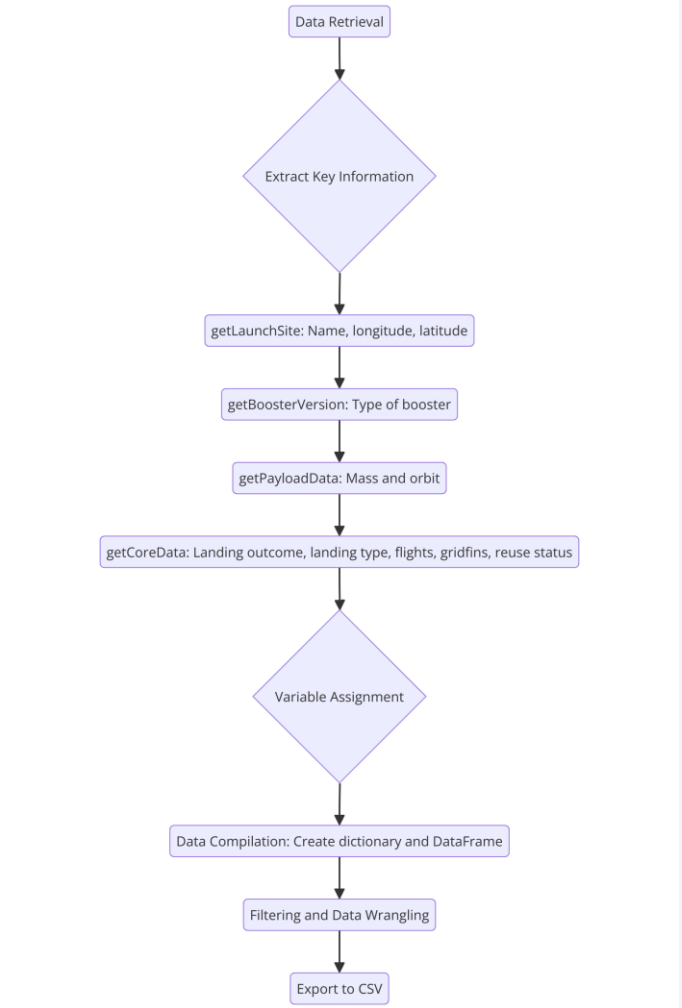
Data Compilation: Combine the variables into a dictionary, launch_dict, to structure the data and then create a pandas DataFrame, launch_data, from this dictionary.

Filtering and Data Wrangling:

- Filter the DataFrame to keep only Falcon 9 launches.
- Reset FlightNumber for sequential ordering.
- Handle missing values:
  - Replace NaN values in the PayloadMass column with the mean payload mass.
  - Allow None for missing values in the LandingPad column.

Export to CSV: Save the cleaned data to a CSV file (dataset_part_1.csv) for further analysis.

GitHub URL – Click here



Data Collection - Scraping

Web Scraping Steps

Send HTTP Request: I used requests to get the HTML content of the Wikipedia page and ensured it was successful.

Parse HTML: I parsed the HTML using BeautifulSoup to navigate the page easily.

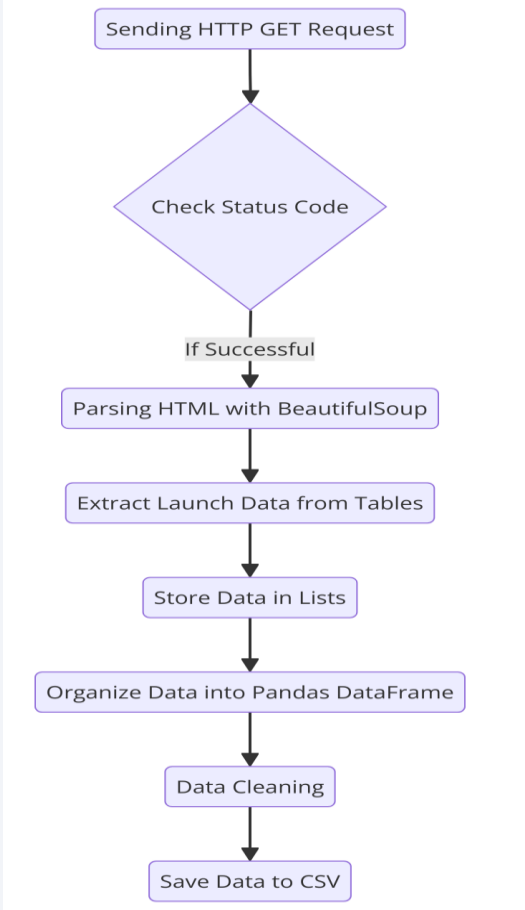
Extract Data: I identified the relevant <table> elements and iterated through rows to extract data like mission name, date, and landing success.

Store Data: I temporarily stored the data in lists and then organized it into a pandas DataFrame.

Clean Data: I handled missing values and standardized formats.

Save Data: Finally, I saved the cleaned data to a CSV file.

GitHub URL – Click here



Data Wrangling

Data Wrangling Steps

Library Installation and Import: I Installed and imported essential libraries: pandas for data manipulation and numpy for numerical operations.

Data Loading: I Loaded the dataset from a provided URL using pd.read_csv() and displayed the first 10 rows to inspect the data.

Data Cleaning: Made checks for missing values by calculating the percentage of null values per column.

Data Exploration: I Examined data types of each column with df.dtypes and Counted unique values in specific columns like LaunchSite and Orbit to understand distributions.

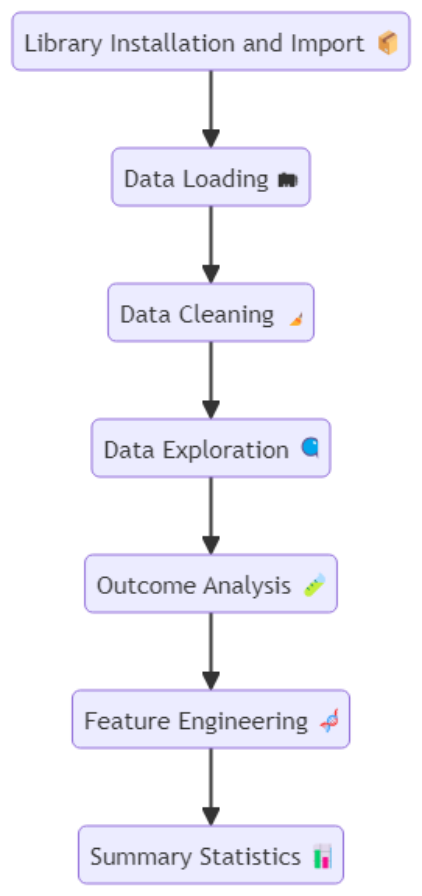
Outcome Analysis: I Analysed the Outcome column to classify landing outcomes and Defined "bad" outcomes using a specific subset of values from the Outcome column.

Feature Engineering: I Created a new binary Class column:

- Assigns 1 for successful (non-bad) outcomes.
- Assigns 0 for bad outcomes.

Summary Statistics: I Computed the mean of the Class column to gauge the success rate.

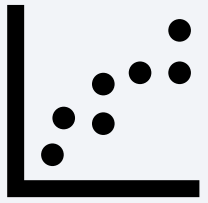
GitHub URL – Click here



EDA with Data Visualization

I used charts to analyse rocket launch data:

- **Line chart** shows launch success rates from 2013 to 2020, highlighting improvement over time.
- **Bar chart** visualizes relationships between variables to identify those impacting success rates.
- **Scatter plots** explore how payload mass and orbit types relate to launch success.



These visualizations guide feature selection for building a predictive model on launch outcomes.

GitHub URL – Click here

EDA with SQL

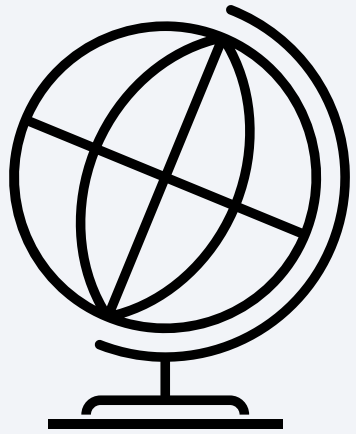
SQL queries performed

- **Table Operations**: Dropped an existing table (SPACEXTABLE) if it existed, then created a new table (SPACEXTABLE) from SPACEXTBL where the Date column is not null.
- **Data Retrieval**: Queried distinct launch_site values, selected the first 5 records for LAUNCH_SITE starting with "CCA", and retrieved specific columns for launches in 2015 with failed landing outcomes.
- **Aggregations**: Calculated total sum, average, and minimum values for payload mass and landing dates based on specific conditions.
- **Filtering:** Retrieved payloads with successful landings on a drone ship and payload mass between 4000 and 6000.
- **Grouping and Counting**: Counted occurrences of different MISSION_OUTCOME and Landing_Outcome values, grouping and ordering by count where necessary. Selected booster version with the maximum payload mass.

GitHub URL – Click here

Build an Interactive Map with Folium

Using Folium for the interactive map, I implemented various features including markers to pinpoint launch locations (such as the NASA JSC space station launch site), circular highlights with text labels at specific coordinates, and connecting lines to show the proximity between different launch sites.



GitHub URL – Click here

Build a Dashboard with Plotly Dash

Within my Plotly Dash application, I incorporated interactive elements including a dropdown menu and range slider, enabling users to manipulate and explore both the pie chart and scatter plot visualizations.

GitHub URL – Click here

Predictive Analysis (Classification)

Machine Learning Steps:

- Importing the required libraries.
- Loading the cleaned data.
- Standardizing the data to prevent bias.
- Splitting the data into 20% for testing data and 80% training data.
- Initializing 4 different classification algorithms: Logistic Regression, Support Vector Machine (SVM), Decision Tree, K Nearest Neighbours (KNN)
- Using Grid Search technique to find the best parameters.
- Using evaluation techniques including Confusion Matrix, F1 Score, and Jaccard Score for the purpose of using the best model among the algorithms above.

GitHub URL – Click here

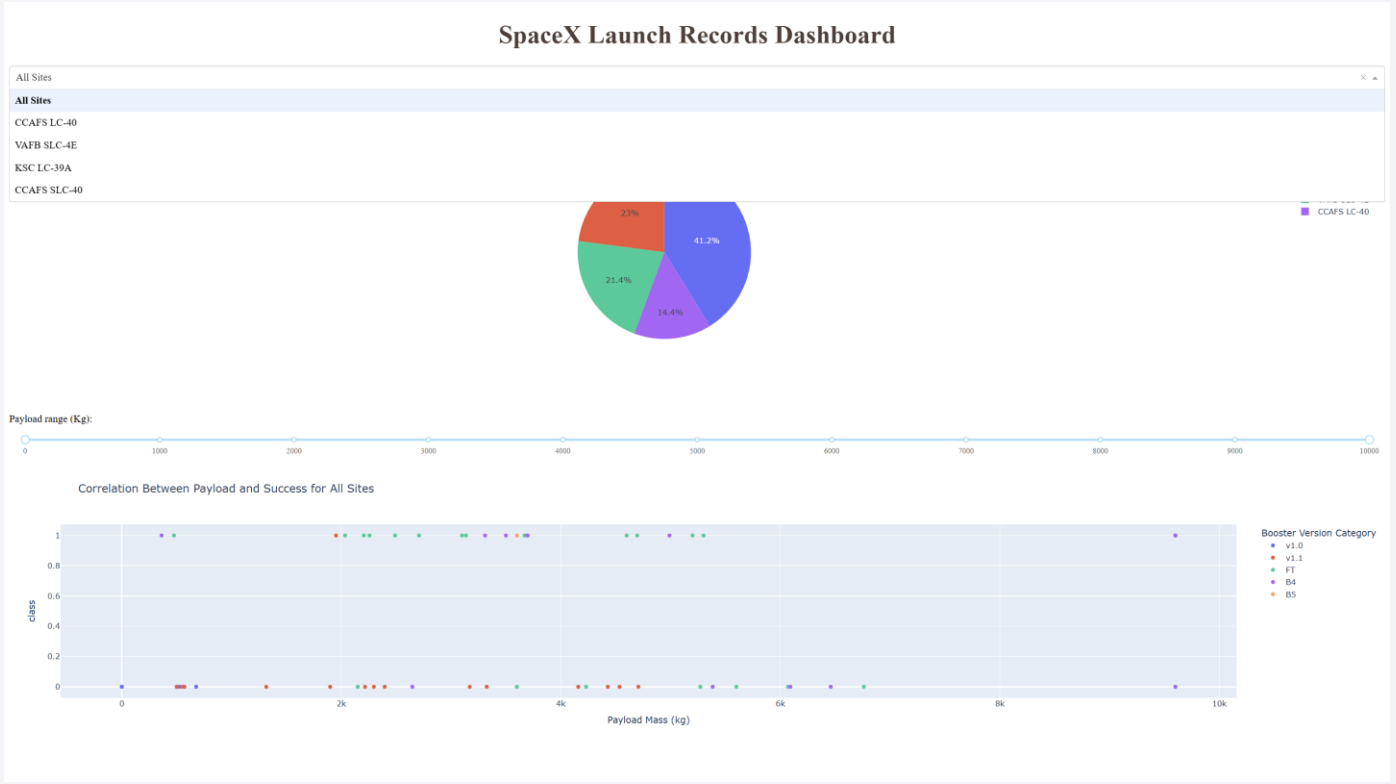
Results

The exploratory analysis revealed notable findings:

- NASA (CRS) had a total 'payload mass' of 45,596, while the average payload_mass for 'booster version F9 v1.1' was 2928.4.
- The first successful ground pad landing was recorded on '2015-12-22', with up to 5 Booster_versions achieving successful landing_outcome on 'drone_ship' carrying payload_mass between (4000,6000).
- SpaceX's Falcon 9 achieved a total of 99 successful mission_outcomes.

Results

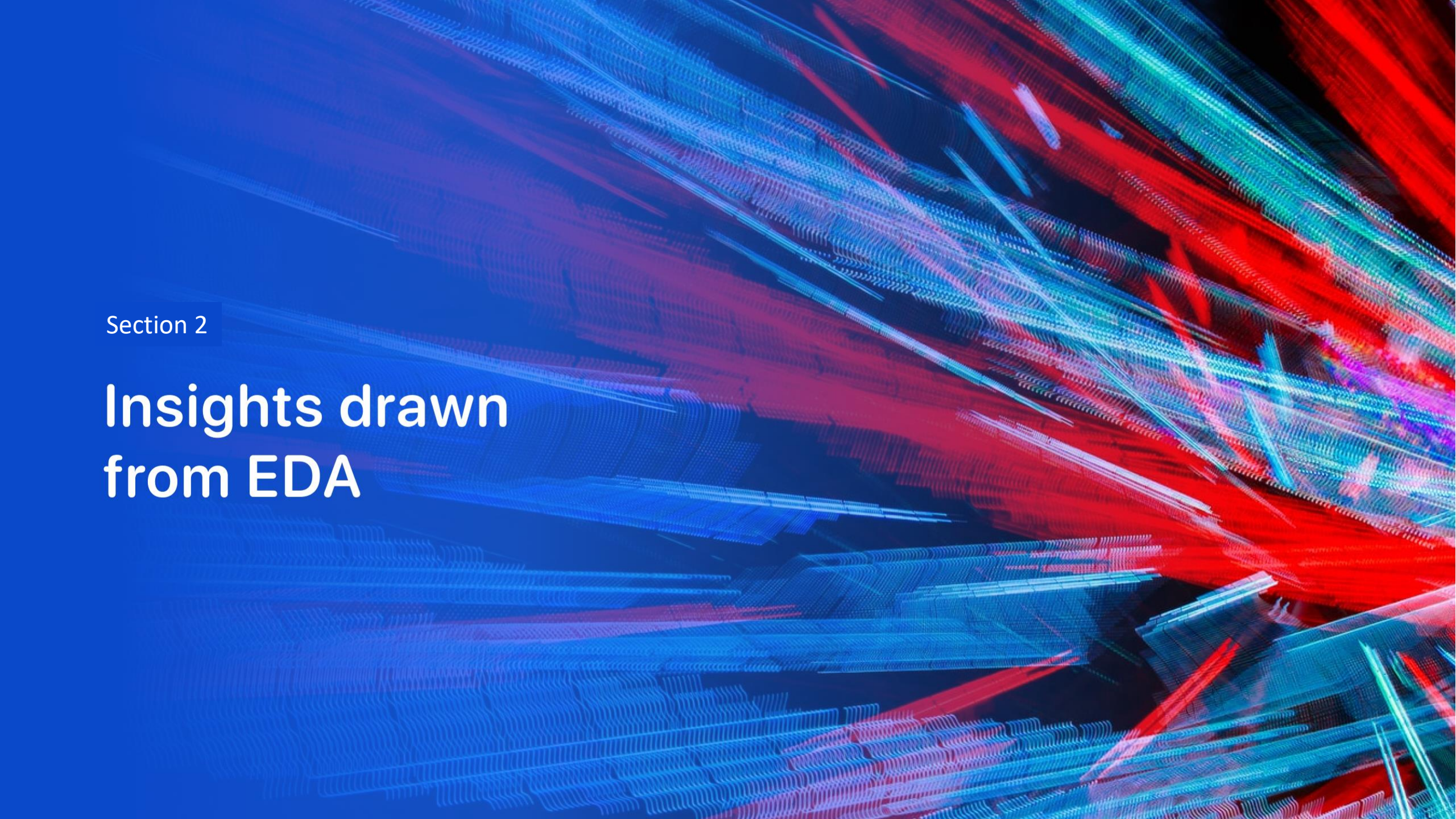
The visualizations highlighted additional insights, showing improved success rates for CCAFS SLC 40 with increasing flight numbers, and perfect success rate (100%) for Falcon 9 landings with orbit types 'ES-L1, SSO, HEO and GEO'. Additionally, Falcon 9's first stage landing demonstrated remarkable improvement in success rates from 2010 to 2020."



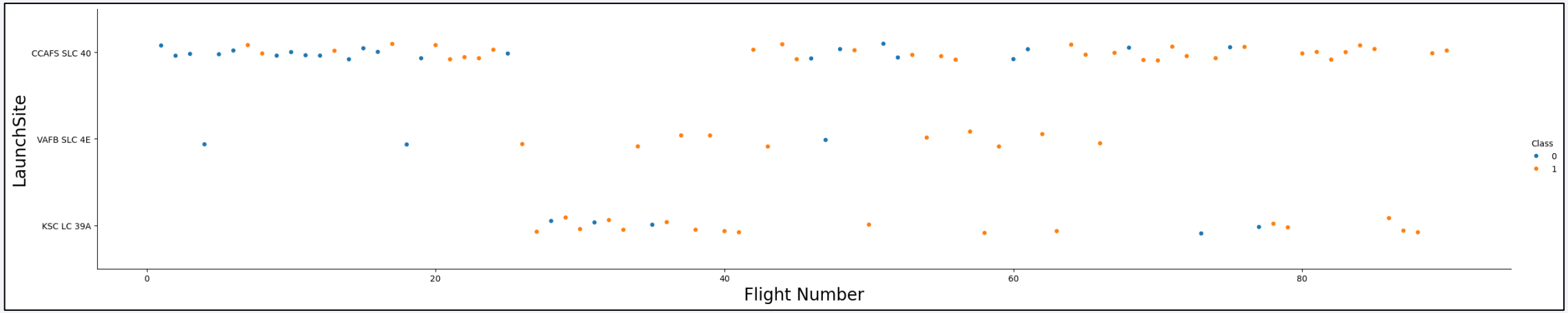
Results

All the models used yielded good performance with and accuracy above 83%, while Decision Tree with Grid Search CV achieved the highest accuracy of 88.89%.

We can be sure that 88.89% of the time our model will be able to predict whether Falcon 9 first stage landing will be successful.

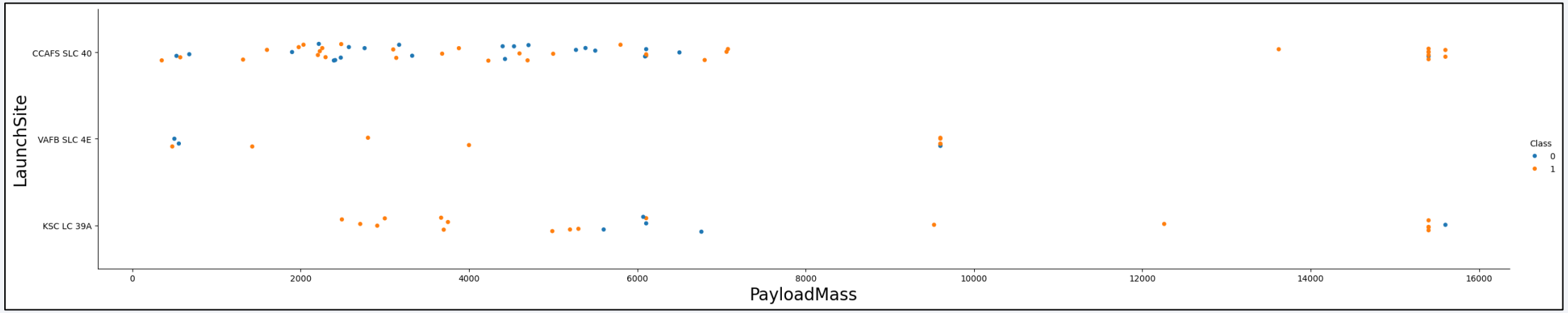

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



Based on the relationship between Flight Number and Launch Site shown above, we can observe that with a lower flight number (20), there are no success or failure metrics for "KSC LC 39A," while "VAFB SLC 4E" has two recorded failures. However, as the flight number goes beyond 80, "CCAFS SLC 40" shows more successful launches, with no failures recorded for "VAFB SLC 4E."

Payload vs. Launch Site

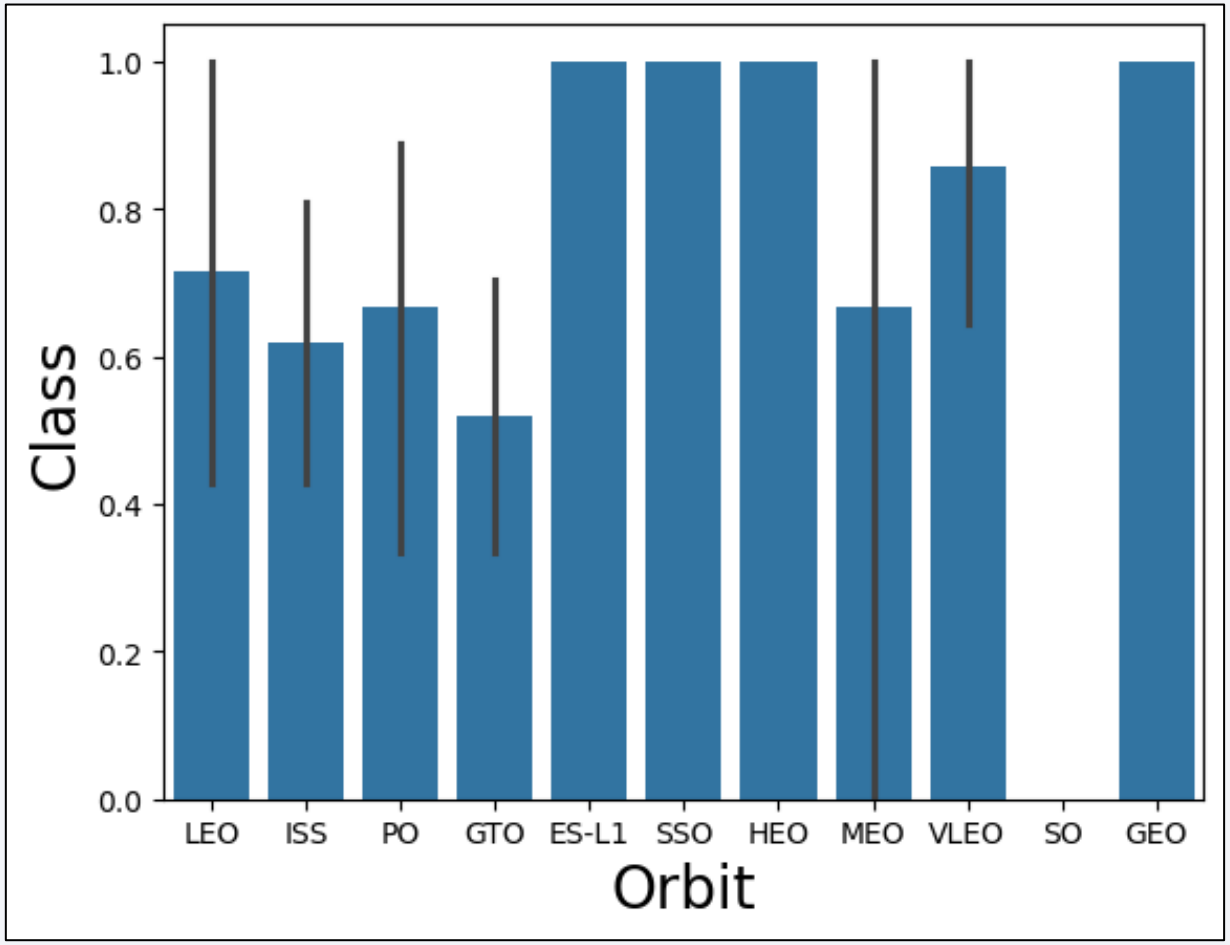


The plot above tells us that there is no strong correlation between payload mass and the success of the first stage return, as the distribution of failed and successful trials remains fairly balanced across various payload masses.

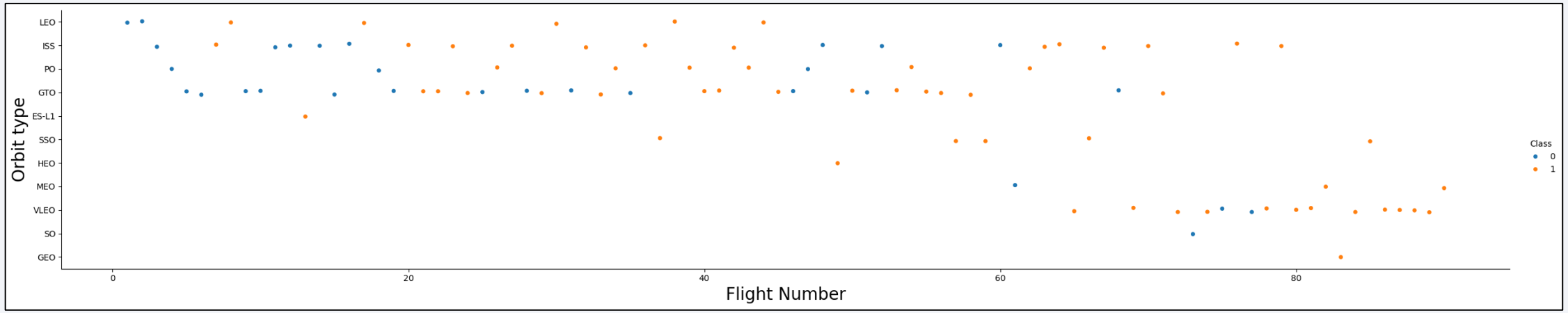
Success Rate vs. Orbit Type

Four orbit types—ES-L1, SSO, HEO, and GEO—achieve the highest success rate of 100%

The 'GTO' orbit has the lowest success rate, making it essential to understand the factors contributing to its performance to minimize the risk of first stage return failures.

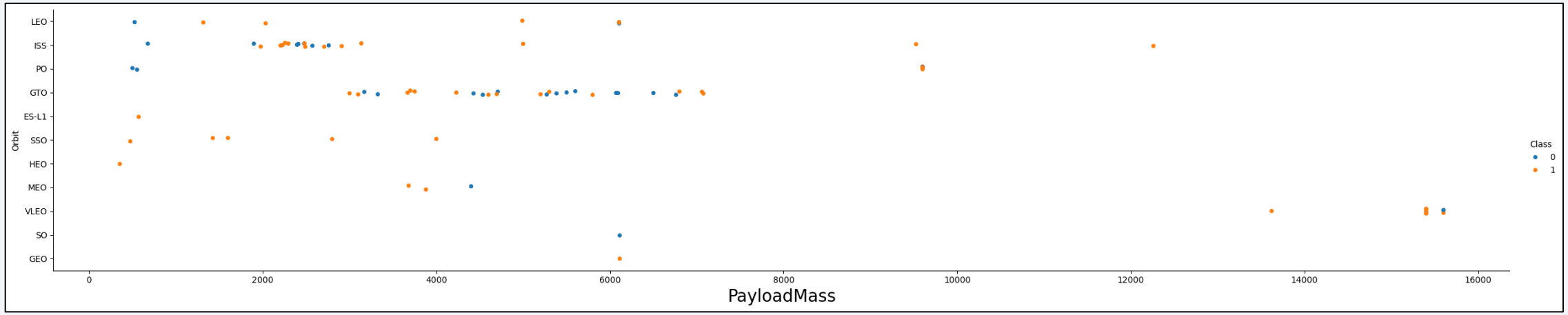


Flight Number vs. Orbit Type



In the GTO orbit, there seems to be no clear relationship between success and flight number.

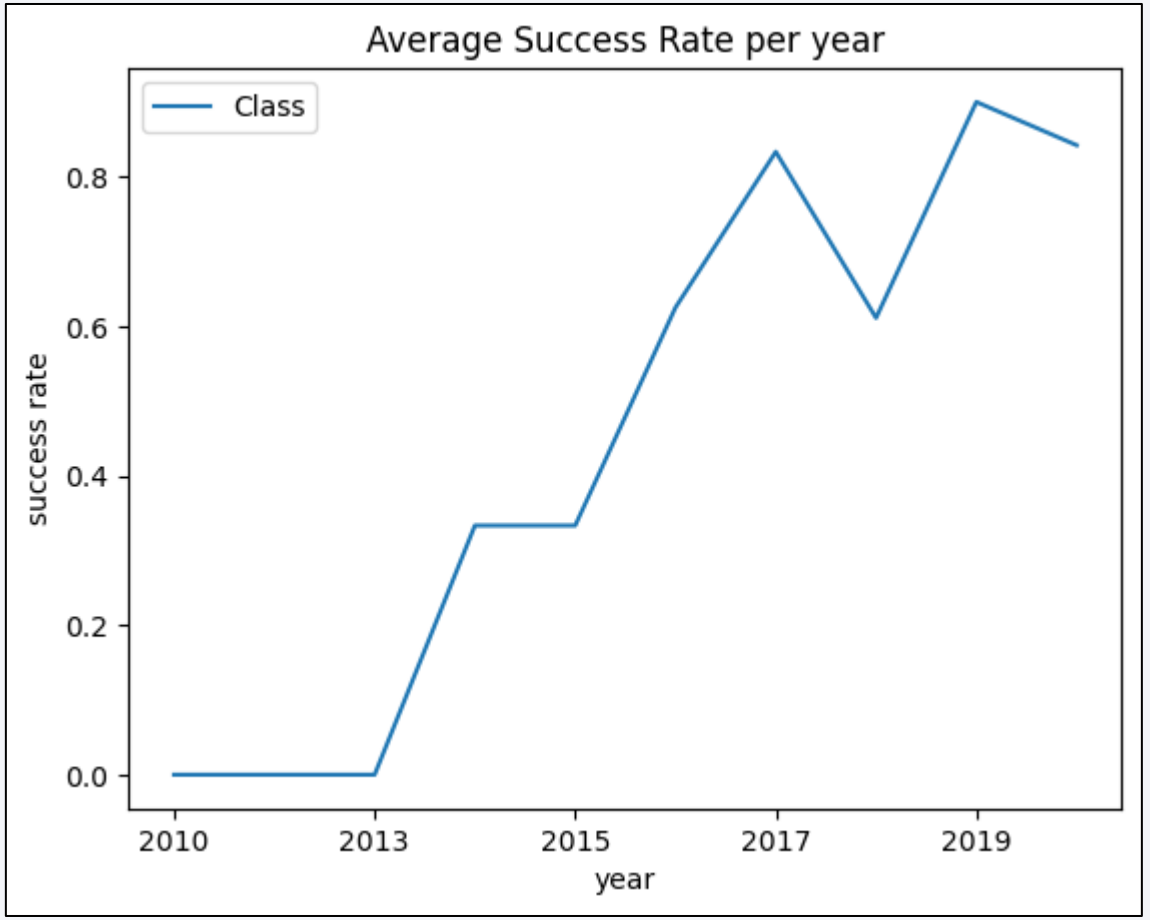
Payload vs. Orbit Type



There is more positive landing rate with heavy payloads for PO, LEO and ISS, however for GTO there's the presence of both positive and negative landing.

Launch Success Yearly Trend

The annual success rates of Falcon 9 landing kept increasing from 2013 to 2020.



All Launch Site Names - Unique Launch Sites

```
%sql select distinct launch_site from SPACEXTBL;
```

Launch_Site	Site Description
CCAFS LC-40	Space Launch Complex 40 (formerly known as Launch Complex 40, or LC-40) is an orbital launch pad situated in the northern region of Cape Canaveral, Florida.
VAFB SLC-4E	Space Launch Complex 4 at Vandenberg Space Force Base in California, U.S., is a launch and landing site with two pads. Both are used by SpaceX for Falcon 9 operations: one pad is designated for launches, while the other, known as Landing Zone 4, is used for landings.
KSC LC-39A	Launch Complex 39A is the first of three launch pads within Launch Complex 39, located at NASA's Kennedy Space Centre on Merritt Island, Florida (Wikipedia).
CCAFS SLC-40	Space Launch Complex 40 (formerly known as Launch Complex 40, or LC-40) is an orbital launch pad situated in the northern region of Cape Canaveral, Florida.

Launch Site Names Begin with 'CCA'

Here below is a list of the first five Launch Site whose names begin with “CCA”:

```
%sql SELECT * \
    FROM SPACEXTBL \
    WHERE LAUNCH_SITE LIKE'CCA%' LIMIT 5;
```

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

Total Payload Mass

The total payload mass of boosters from NASA is 45596.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2928.4

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

The dates of the first successful landing outcome on ground pad is:

```
%sql SELECT MIN(DATE) \
FROM SPACEXTBL \
WHERE Landing_Outcome = 'Success (ground pad)'
```

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

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass between 4000 to 6000 are:

```
%sql SELECT PAYLOAD \
FROM SPACEXTBL \
WHERE Landing_Outcome = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failed mission outcomes are 98 for success, one for Failure (in flight) and 1 for success (payload status unclear).

```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
FROM SPACEXTBL \
GROUP BY MISSION_OUTCOME;
```

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

Mission_Outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

List of the booster which have carried the maximum payload mass:

```
%sql SELECT BOOSTER_VERSION \
FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

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

2015 Launch Records

There were two failed landing in 2015 on a drone ship which both in the same site, CCAFS LC-40 and with same booster version F9 v1.1:

```
%sql SELECT substr(Date,6,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing_Outcome] \
FROM SPACEXTBL \
where [Landing_Outcome] = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

```
%sql SELECT [Landing_Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '2010-06-04' and '2017-03-20' group by [Landing_Outcome] order by count_outcomes DESC;
```

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

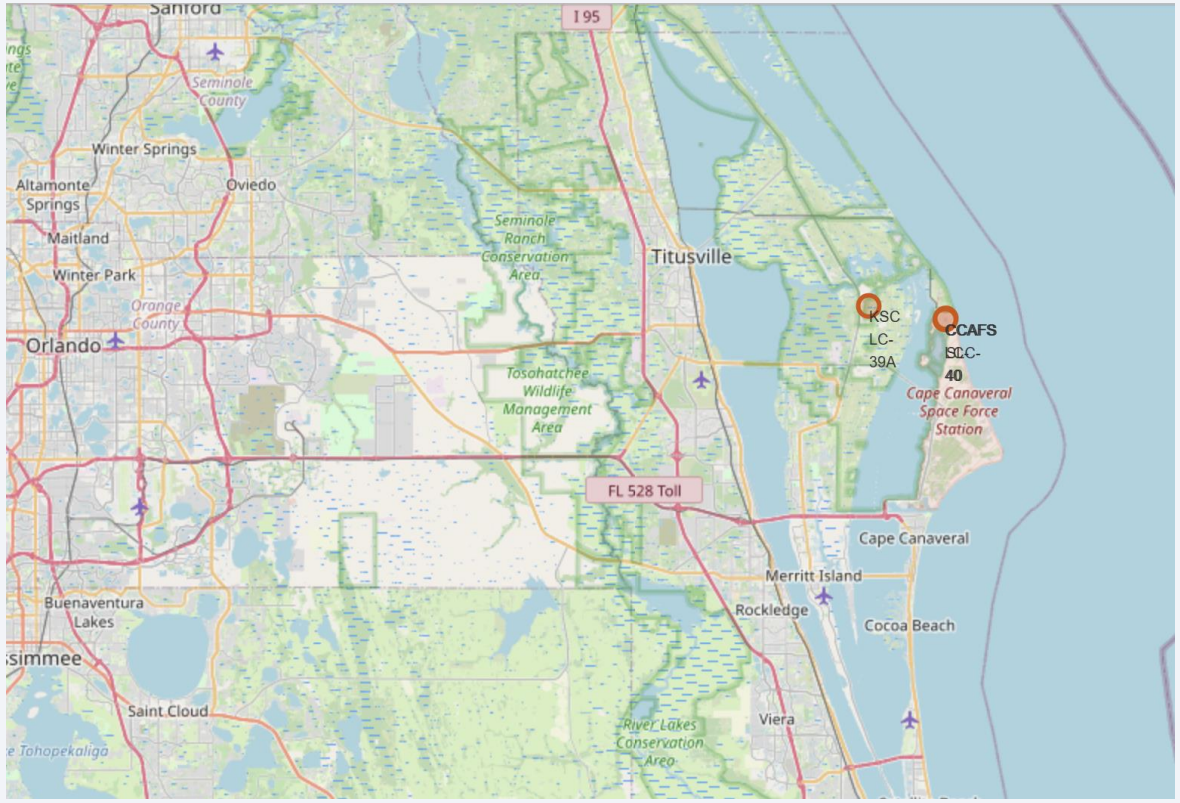
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

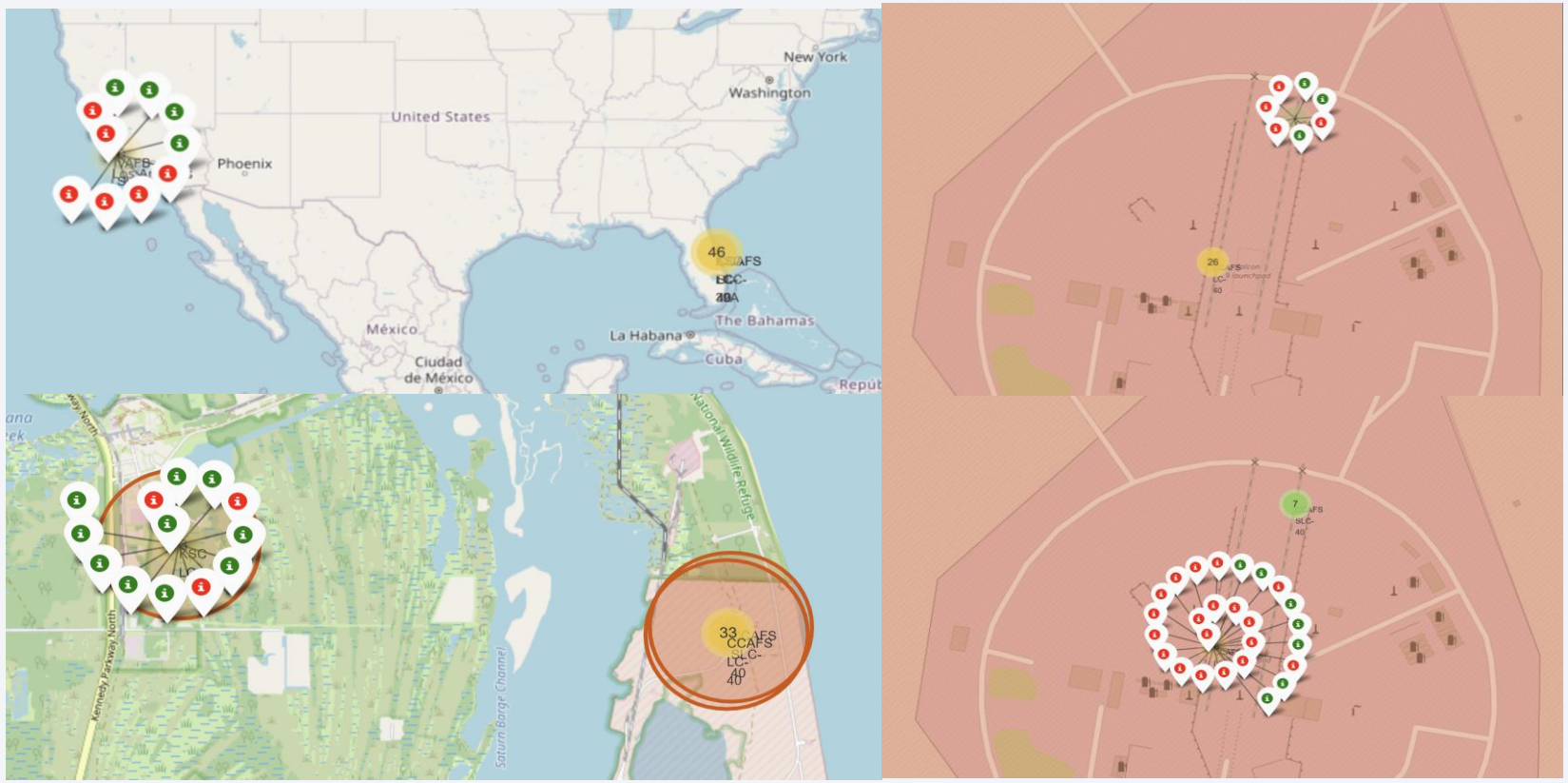
Lunch Sites Locations

All lunch sites are in the US and are above the equator line and are in very close proximity to the coast.



Folium Map: Success Rate For Each Launch Location

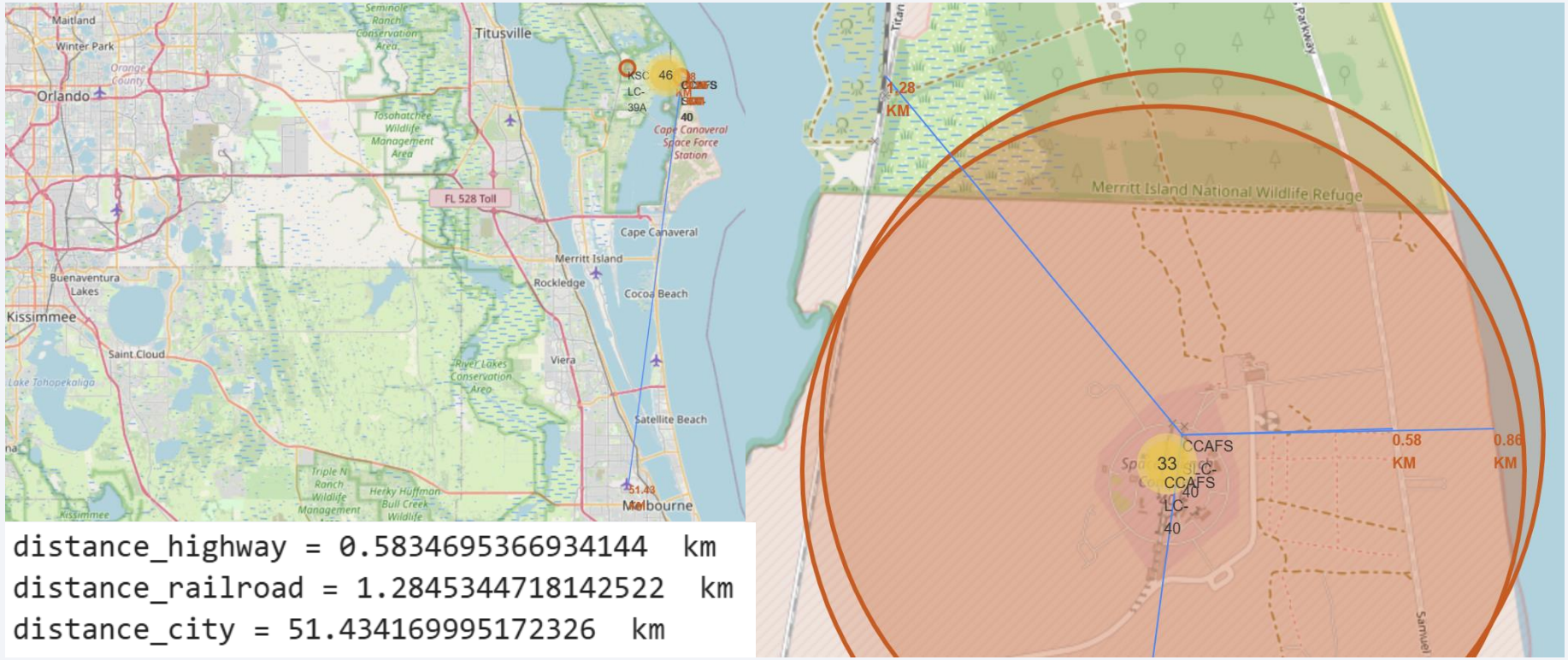
We can see that VAFB SLC-4E and CCAFS SLC-40 have relatively high success rates. Green Marker = Successful Return Red Marker = Failed Return



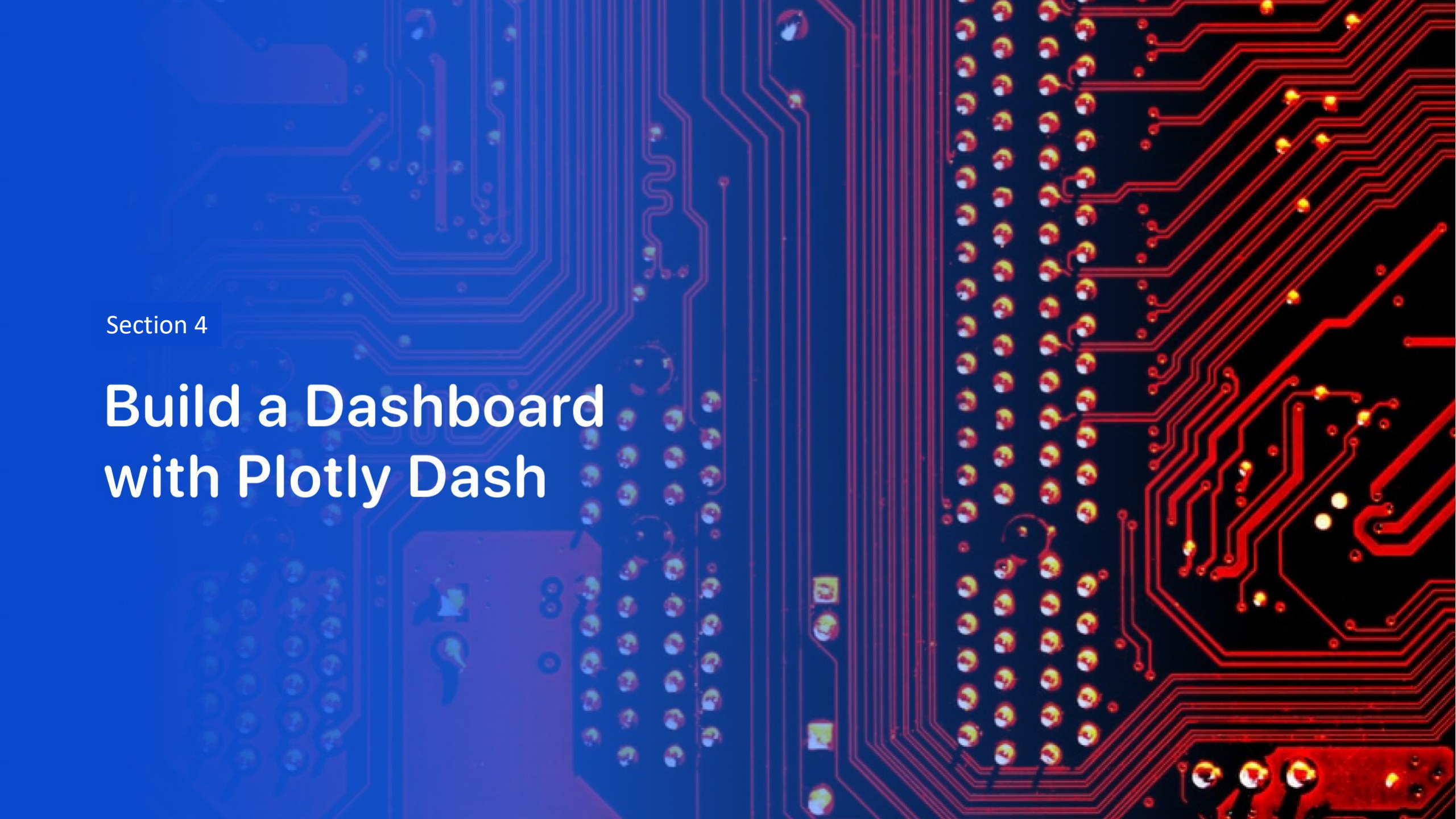
Folium Map: Closest Proximities to CCAFS LC-40

The distances between the launch site (CCAFS LC-40) to its proximities were calculated.

Launch sites are typically near railways and highways for transport, close to coastlines for safe over-water launches, and distanced from cities to minimize risk in case of accidents.



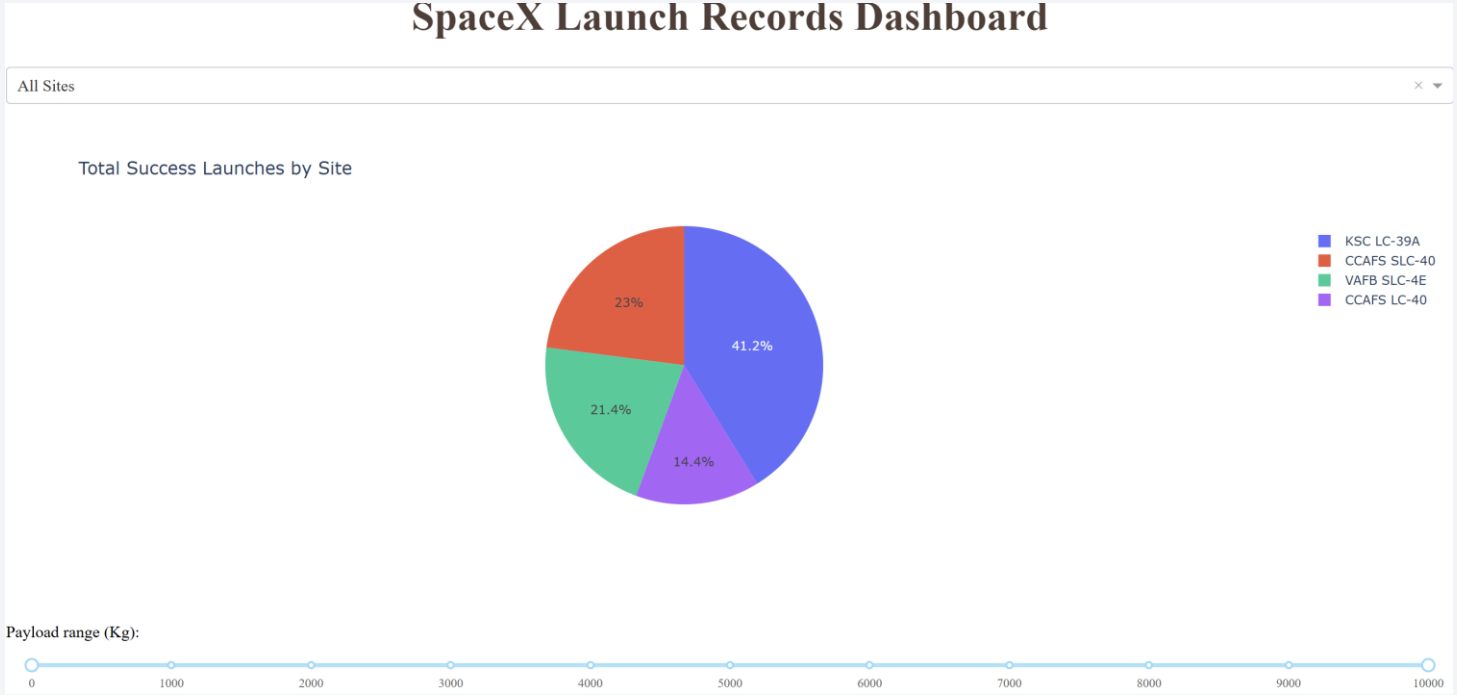

```
distance_highway = 0.5834695366934144  km
distance_railroad = 1.2845344718142522  km
distance_city = 51.434169995172326  km
```


Section 4
Build a Dashboard with Plotly Dash

Dashboard: Launch Success Count For All Sites

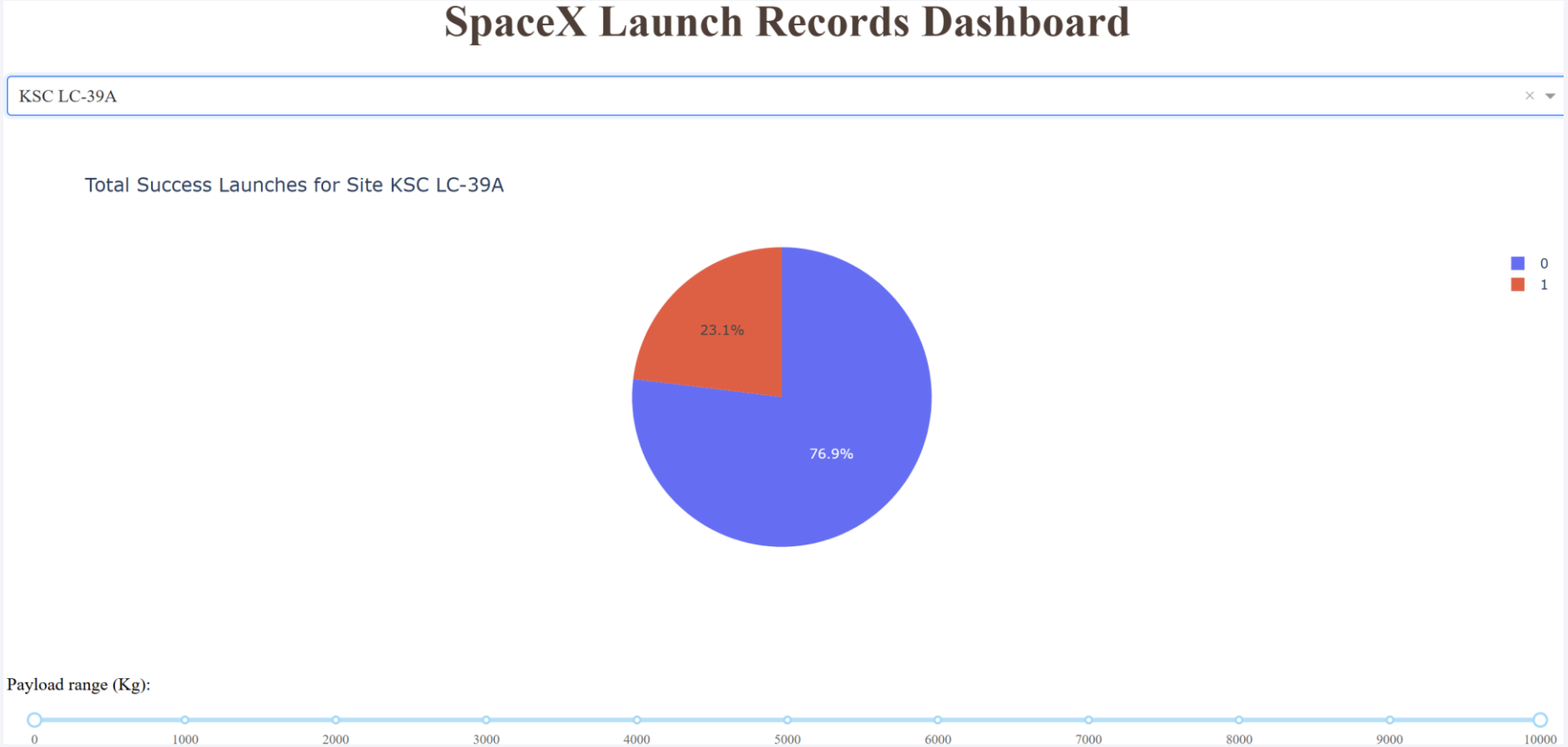
This graph shows the success percentage for every single site in terms of first stage return:

- The best site is KSC LC-39A with 41.2%. of total successful rocket launch among all sites
- The least site for successful rocket launch: CCAFS SLC-40 with only14.4%.



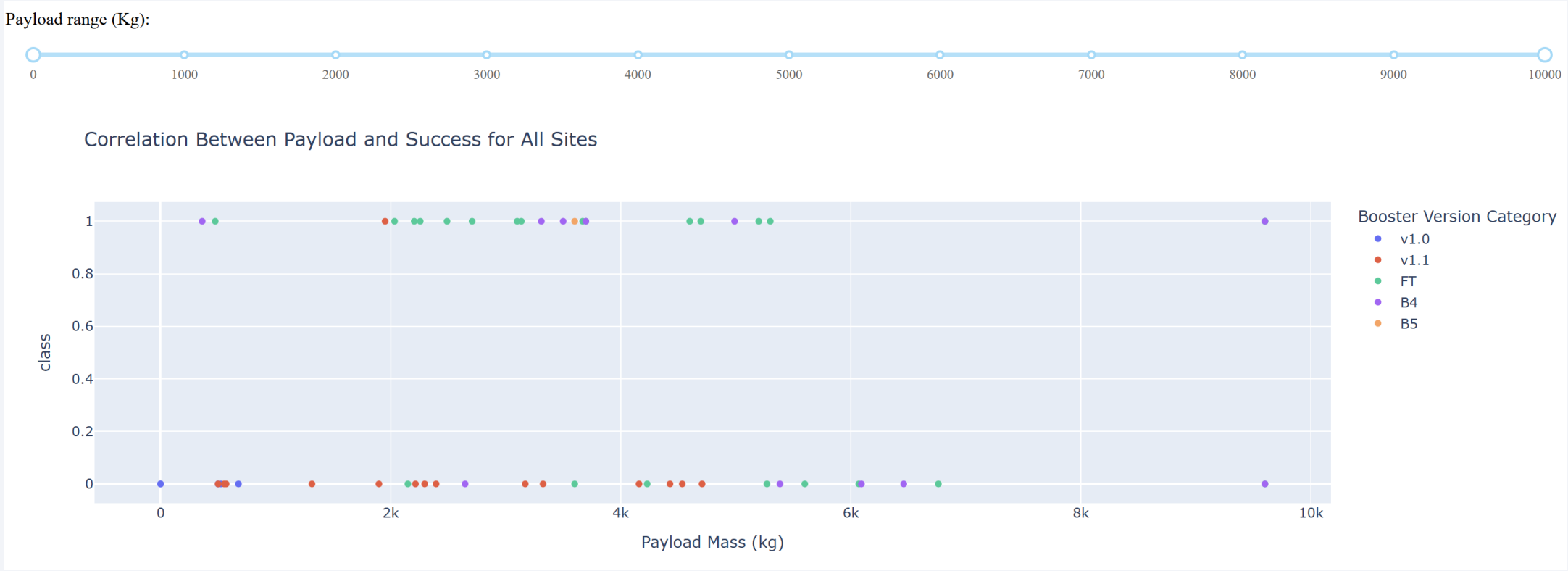
Dashboard: Launch Success For KSC LC 39A

Total Success Launches for site KSC LC-39 :



Dashboard: Scatter Plot Payload Mass And Launch Outcomes

This interactive scatter plot illustrates the relationship between payload mass and launch outcomes, highlighting how payloads under 4000 kg tend to have higher success rates, particularly for certain booster versions

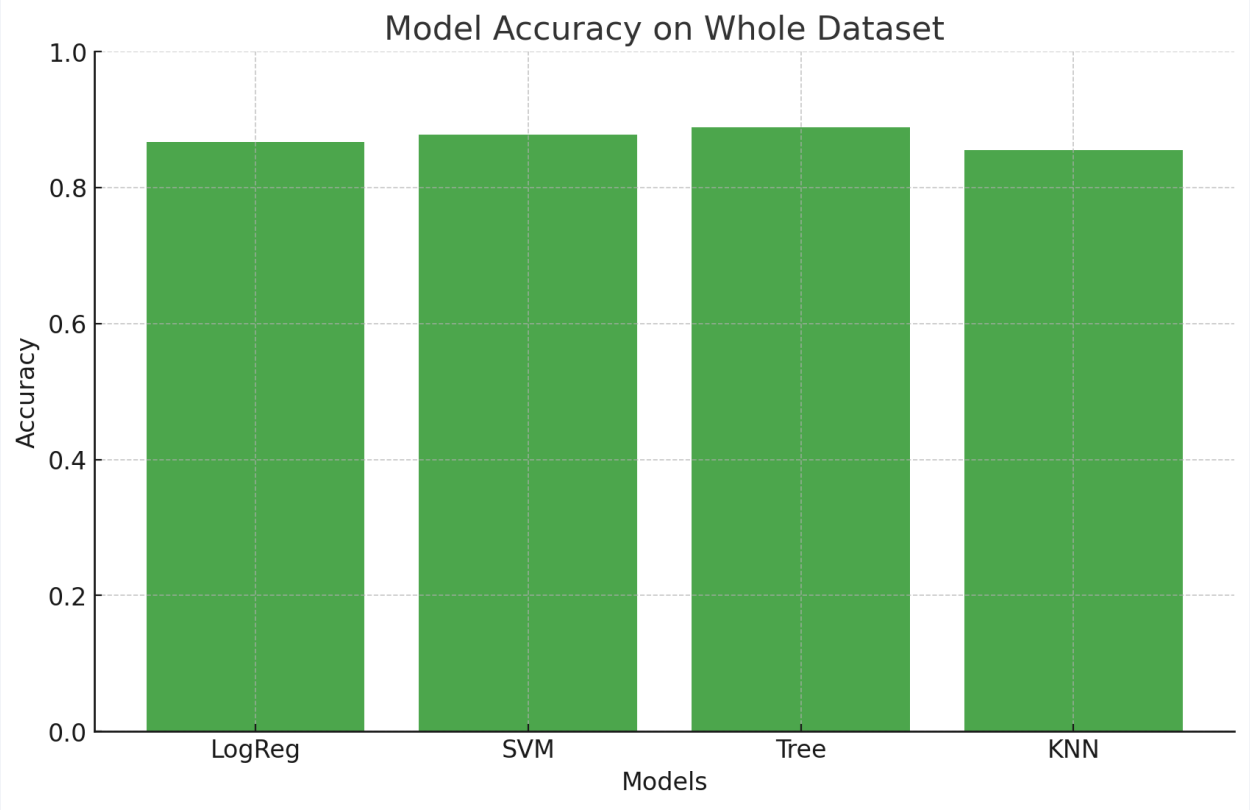


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The Decision Tree model achieved the highest accuracy on the whole dataset (0.8889), while all models performed equally on the test sets with an accuracy of 0.8333.

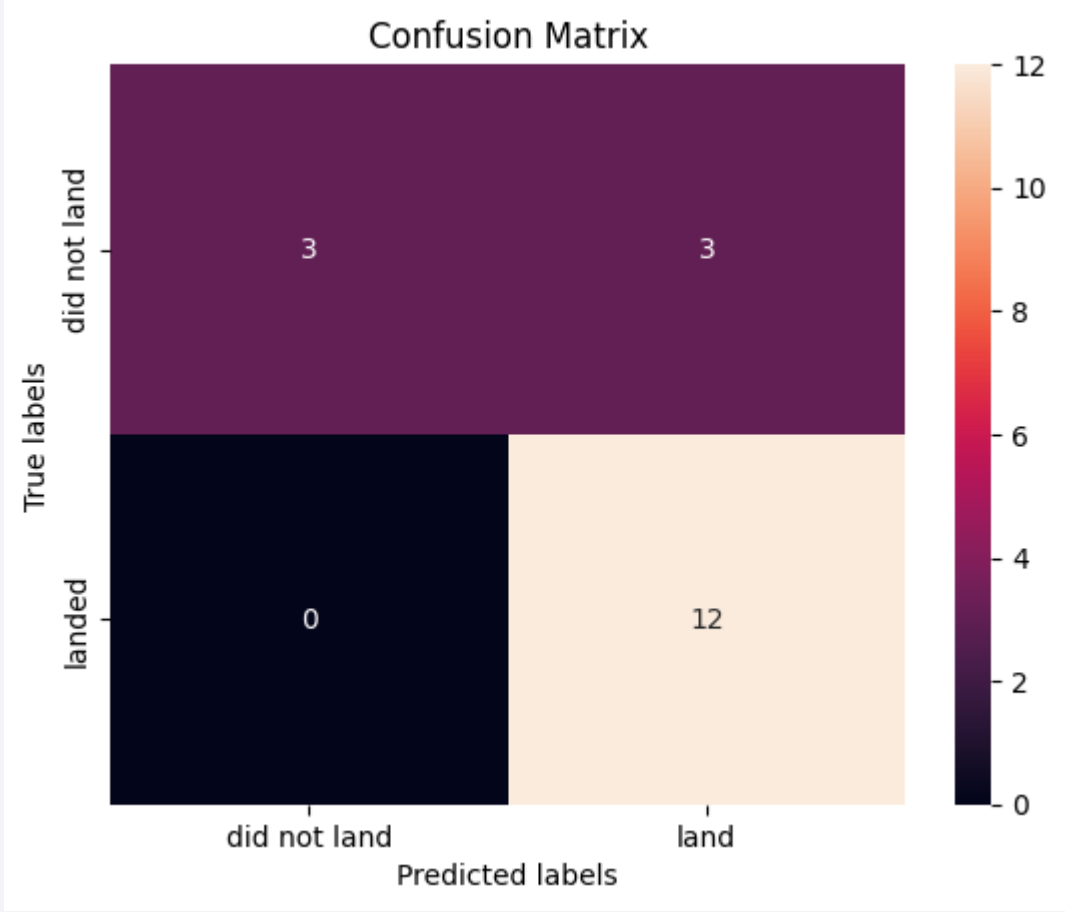


Confusion Matrix

The confusion matrix shown here evaluates the performance of the Decision Tree model on the test set by comparing the predicted labels to the true labels.

- **True Negatives (Top-left, 3)**: The model correctly predicted "did not land" for 3 cases.
- **False Positives (Top-right, 3)**: The model incorrectly predicted "land" when it actually "did not land" for 3 cases.
- **False Negatives (Bottom-left, 0)**: There were no cases where the model predicted "did not land" when it should have predicted "land."
- **True Positives (Bottom-right, 12)**: The model correctly predicted "land" for 12 cases.

This matrix indicates that while the model performed well in identifying "landed" instances, it had some errors predicting "did not land," resulting in 3 false positives. The absence of false negatives suggests strong sensitivity in detecting "landed" instances.



Conclusions

- Successful first-stage returns result in significant cost savings for rocket launches.
- Our model which considers 83 attributes can impact the likelihood of a successful first-stage return.
- SpaceX Falcon 9 launch sites are strategically located near highways, railways, and coastlines to reduce transportation costs.
- Success rates are affected by orbit type and booster version.
- SpaceX's success rate has improved over time, with the KSC LC-39A site achieving the highest success rate, though further investigation is required.

Thank you!