IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jurre Knoest
14-05-2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection with REST API's and web scraping
  - Data wrangling – replace missing values with either mean or max, one-hot-encoding
  - Exploratory data analysis – SQL, pandas, numpy, matplotlib, seaborn
  - Data visualization – Folium, Plotly, Dash
  - Predictive Analysis – modelling: Logistic Regression, Decision Tree, K-Nearest Neighbors, Support Vector Machine
- Summary of all results
  - The success rate for landing outcomes has shown a consistent improvement over time and with each launch.
  - Launch sites are strategically located near the equator, close to a coast, and sufficiently distant from other proximities to avoid hindrances, yet still convenient for material transportation.
  - Payloads within the range of 2,000 to 6,000 kilograms had the highest success rate, with the FT Booster version being predominantly associated with successful landings.
  - KSC LC-39A emerged as the launch site with the highest success rate.
  - Among the models tested, the Decision Tree Model exhibited the highest accuracy, achieving an 89% accuracy rate in predicting landing outcomes.

Introduction

BACKGROUND

- Rocket launches are a costly undertaking ($165 million per launch), and if our space company wants to compete with the likes of SpaceX we need to find ways to reduce the costs. One way of doing this is to reuse part of the rocket after a successful launch and landing ($62 million per launch). With insight into the factors associated with a successful landing we can make predictions about the outcome of a launch, and thus make competitive budget assessments. In doing so we look at data from the competition regarding the Falcon 9 Rocket

EXPLORE

- The data and its uses
- The success factors and their correlations
- The best predictive model for a successful landing

Section 1

Methodology

Methodology

- **COLLECTION:** The data was collected Using SpaceX REST API's and web scraping
- **WRANGLING:** The data was filtered based on what we needed - The data was parsed in the correct formats - missing values were replaced by either the mean or the max, depending on the type - One Hot Encoding was applied to make the data suitable for modelling
- **EXPLORATION: Exploratory Data Analysis:** The Data was explored using Statistics, SQL and Visualization techniques, such as Folium and Dash Plotly
- **PREDICTIVE DATA ANALYSIS:** The Data was fed to various existing models – evaluated and fine tuned to find the best prediction model

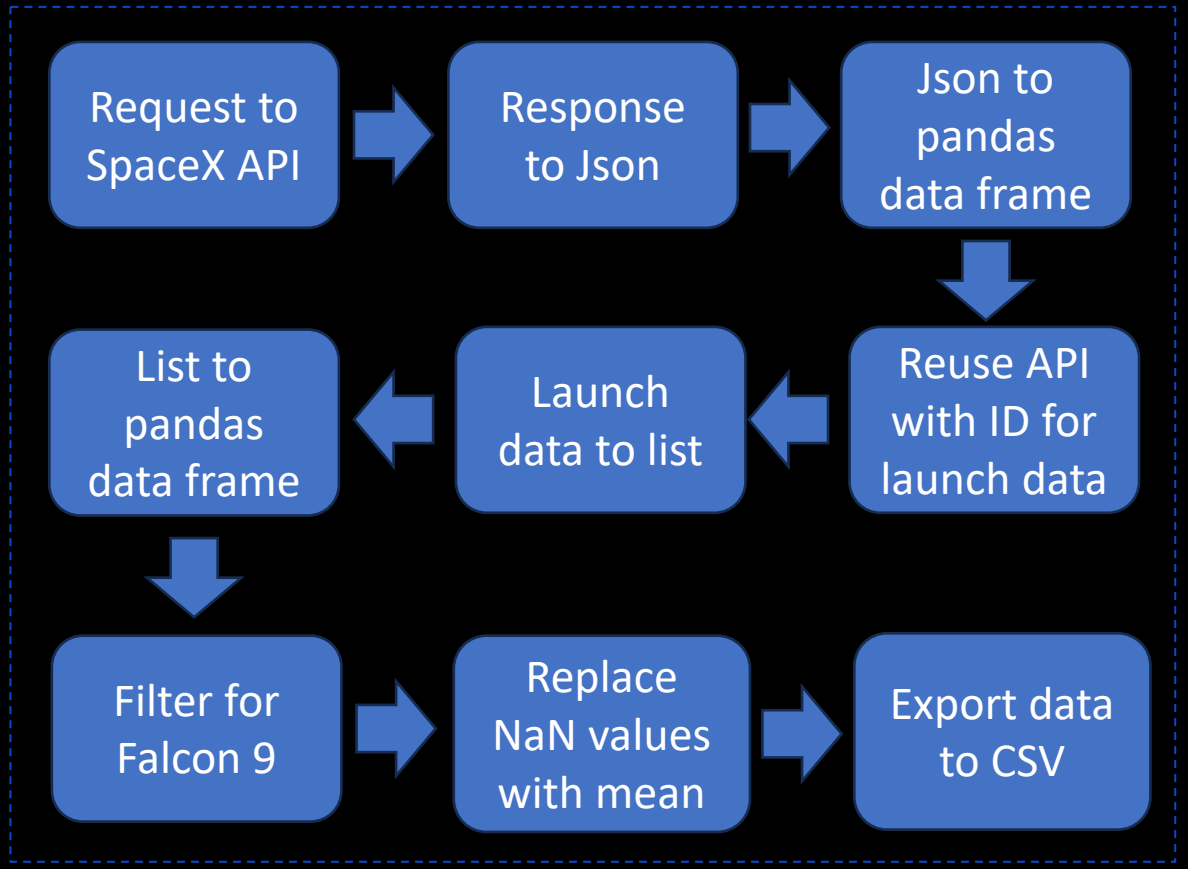
Data Collection

- The first part of the data was collected through an IBM call to a public SpaceX API to launch data in JSON format
- The Second part of the Data was collected through web scraping, extracting launch data from HTML tables on Wikipedia (9-6-2021)

Data Collection – SpaceX API

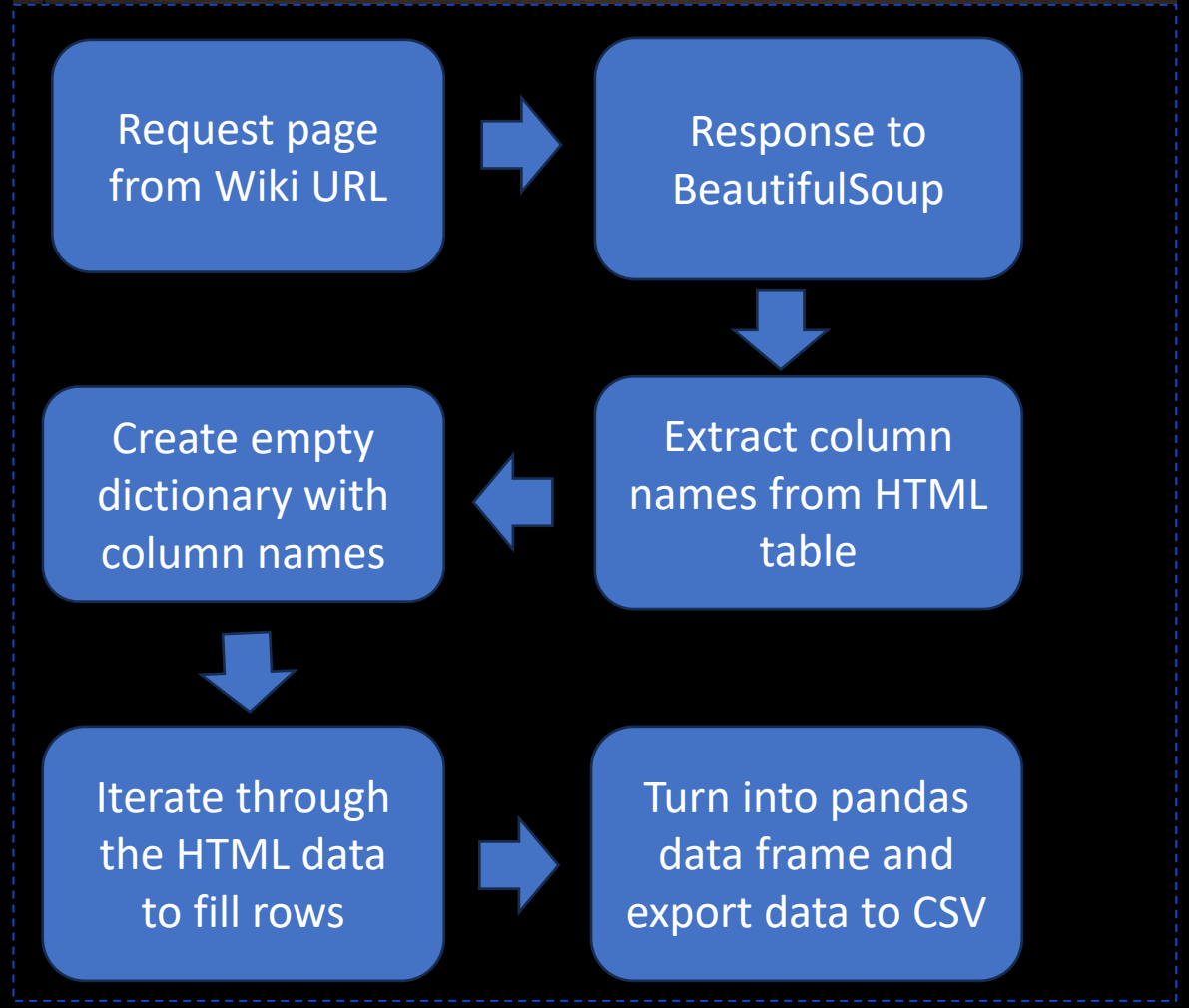
- The flowchart to the right explains how the data was collected through SpaceX REST API's.
- After the first collection we had the ID's necessary to get the actual data we needed
- The actual launch data was filtered for Falcon 9 Booster Versions
- Missing values for Payload Mass were replaced with the mean
- This cleaned data were then exported to an CSV
- GitHub URL: (Click Here)

```
response = requests.get(static_json_url)
response_json = response.json()
data = pd.json_normalize(response_json)
```



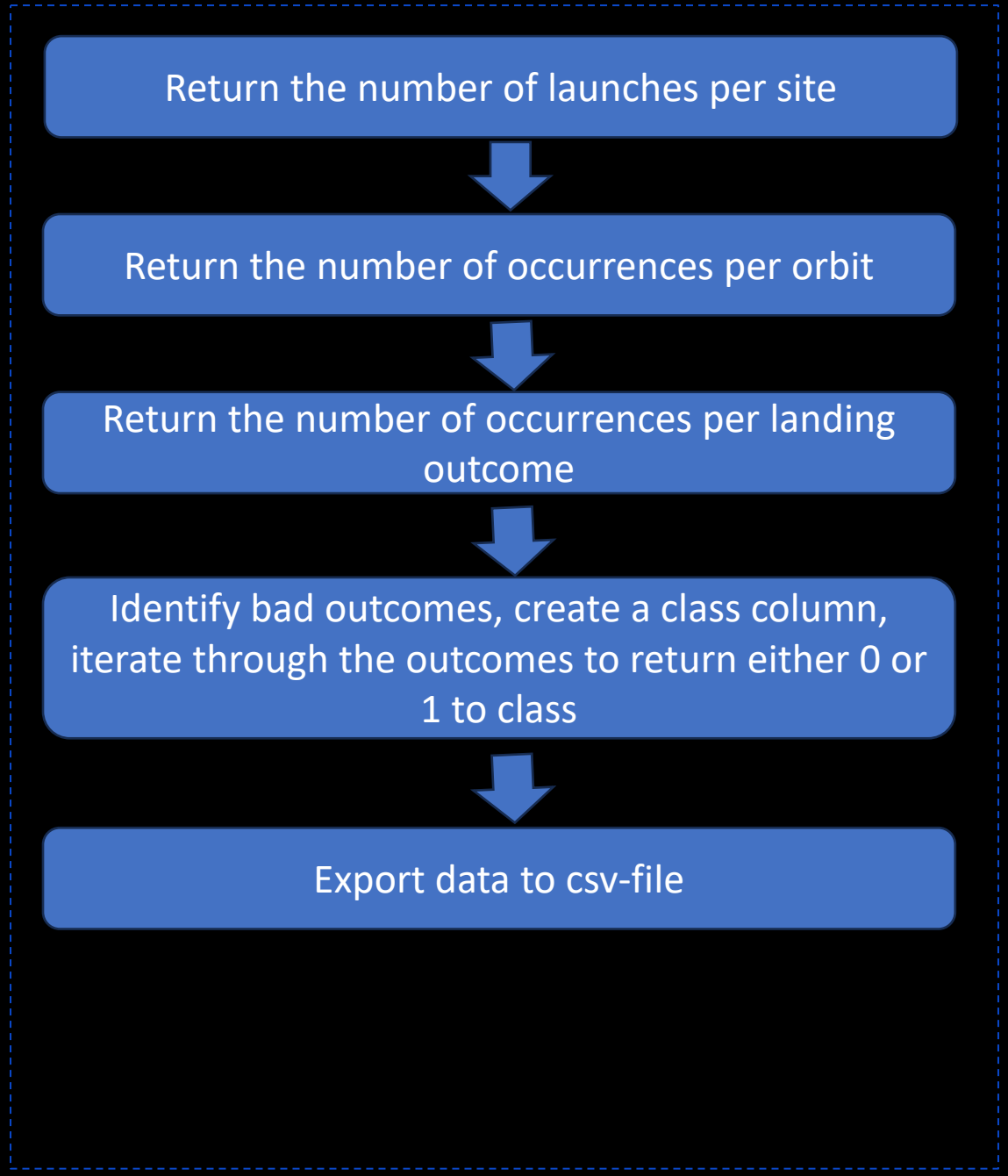
Data Collection - Scraping

- Scraping Wikipedia for a list of Falcon 9 launches, using BeautifulSoup to extract html-tables
- Remove unnecessary columns and create an empty dictionary for storing the column data, iterating through the soup to fill the dictionary with rows
- Convert the dictionary to a pandas data frame and export to a csv-file
- GitHub URL: (Click Here)



Data Wrangling

- In this stage we had to prepare the data related to the landing outcomes for modelling
- Various landing outcomes were converted to a binary classification of 0 for failed landings and 1 for successful landings
- These results were stored in a new column called Class
- This new classification was added to the data frame
- GitHub URL: (Click here)



EDA with Data Visualization

- **SCATTERPLOT CHARTS**
  - Correlations between Launch Site and Flight Number
  - Correlations between Launch Site and Payload Mass
  - Correlations between Orbit Type and Flight Number
  - Correlations between Orbit Type and Payload Mass
- **BAR CHARTS**
  - The success rate per Orbit
- **LINE PLOT**
  - Trend line to gauge the success rate over the years
- **GitHub URL:** (Click Here)

- **ANALYSIS**
  - The scatterplots show a possible relationship between variables, which in turn could be used in choosing and fine tuning a predictive model
  - The bar chart is used for making comparisons between distinct variables of a category (in this case the success rate for each orbit)
  - The Line plot is used to gauge the overall success rate of launches and landings throughout the years and possibly spot a trend

EDA with SQL

- Unique Launch Sites
- 5 records where the launch site begins with ‘CCA’
- Total Payload Mass carried by boosters launched by NASA (CRS)
- Average Payload Mass carried by Booster version F9 v1.1
- Date of the first successful landing on ground pad
- Names of boosters carrying a Payload Mass between 4000 and 6000 and successfully landing on a drone ship
- The total number of successful and failure mission outcomes
- Names of booster versions that carried the max Payload Mass
- Failed landing outcomes on drone ship, booster version, and launch site for the months in 2015
- Landing outcomes between 4-6-2010 and 20-3-2017
- GitHub URL: (Click Here)

Build an Interactive Map with Folium

- **MARKERS**
  - Launch Site Names
  - Circles
  - Landing outcomes per site
  - Lines to proximities (e.g. city, coast, railway, highway)
- **ANALYSIS**
  - These markers were placed to get a feel for the locations. These markers could reveal certain characteristics that could be of influence on the landing outcomes
- GitHub URL: (Click here)

Build a Dashboard with Plotly Dash

- Dropdown Menu for Individual Launch Sites
  - Allows the user to select an individual or all launch Sites
- Pie Chart Showing Landing Outcomes
  - Shows Successful and Failed Landing Outcomes per site
- Slider for Payload Mass Range
  - Allows the user to adjust the Payload Mass
- Scatterplot Success Rate Chart Payload Mass vs Booster Version
  - Shows a correlation between Payload Mass and Booster Version in regard to the landing outcome success rate
- GitHub URL: (Click Here) (Click on the SpaceX Dash App link in the Notebook to open the dashboard in a browser)

Predictive Analysis (Classification)

- Create a Numpy Array from the Data Frame (Class)
- Use StandardScaler to Standardize the Data
- Split the Data in a Training Set and a Testing Set
- Create a GridsearchCV for Parameter Optimization
- Apply on the Following Models: Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbors
- Accuracy Testing for all the Models
- Assess the Confusion Matrices
- Identify the Most Accurate Model

- GitHub URL: (Click Here)

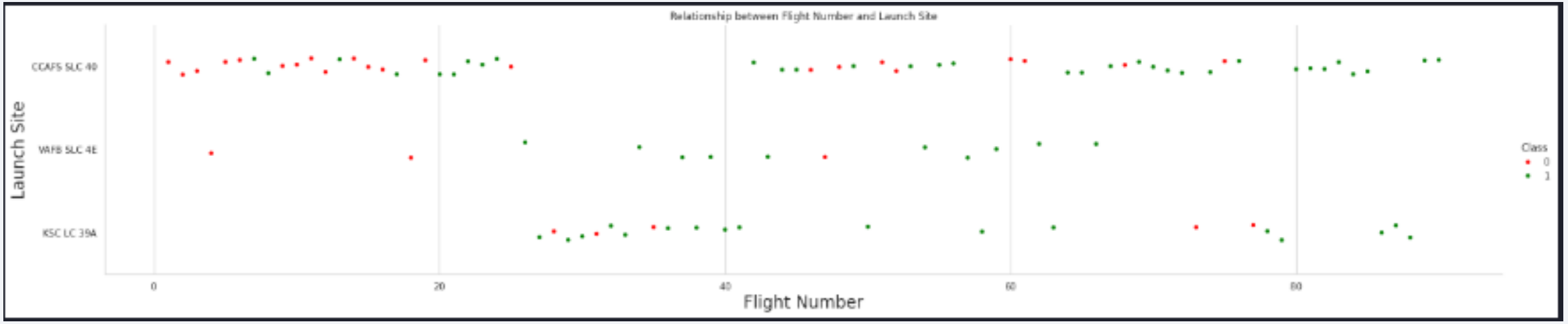
Results

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

Section 2

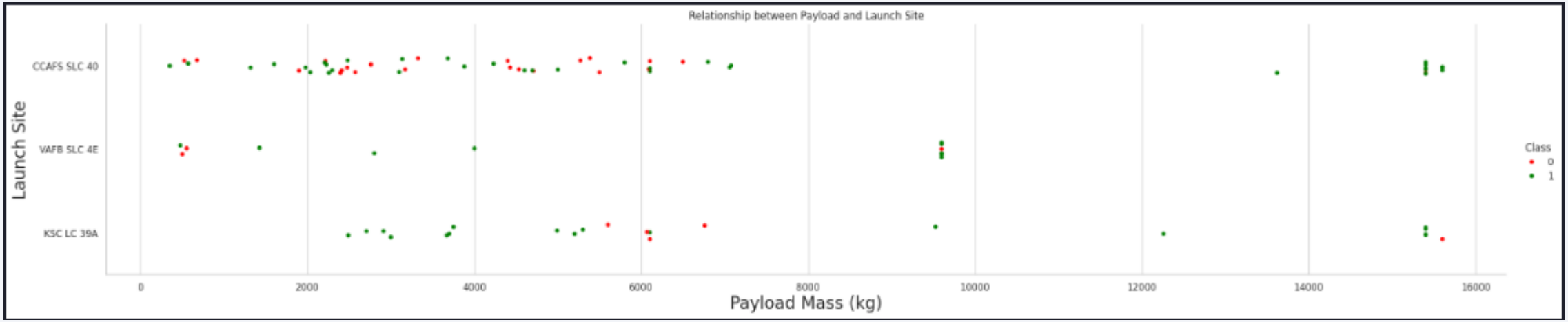
Insights drawn from EDA

Flight Number vs. Launch Site



- GREEN = SUCCESS and RED = FAILURE
- Up until ca. flight number 25, launches occurred mostly from launch site CCAFS SLC 40
- The success rate of these early flights was not very high.
- After flight number 25 other sites became popular for launches as well
- After flight 25 the success rate for all launch sites went up

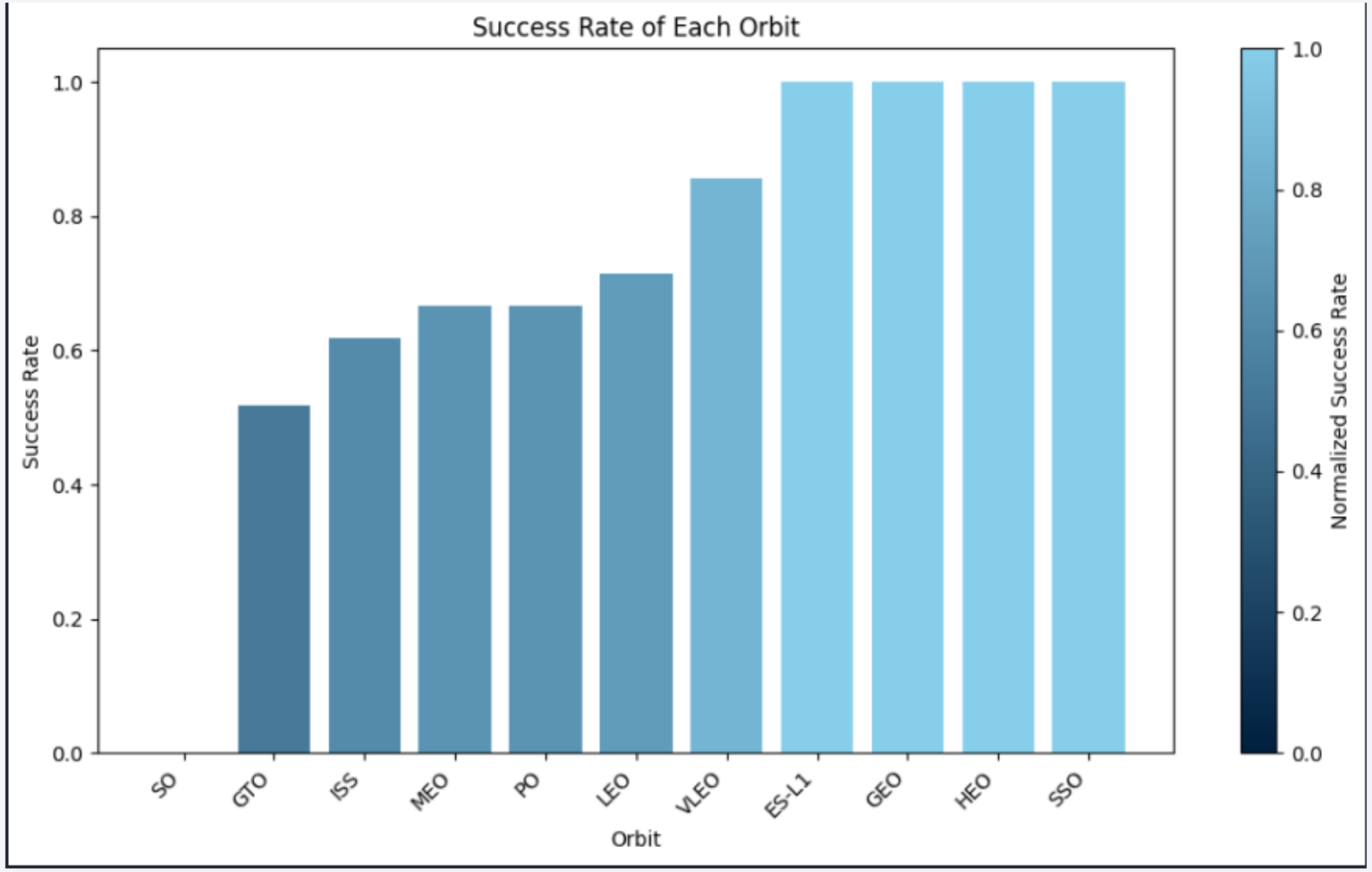
Payload vs. Launch Site



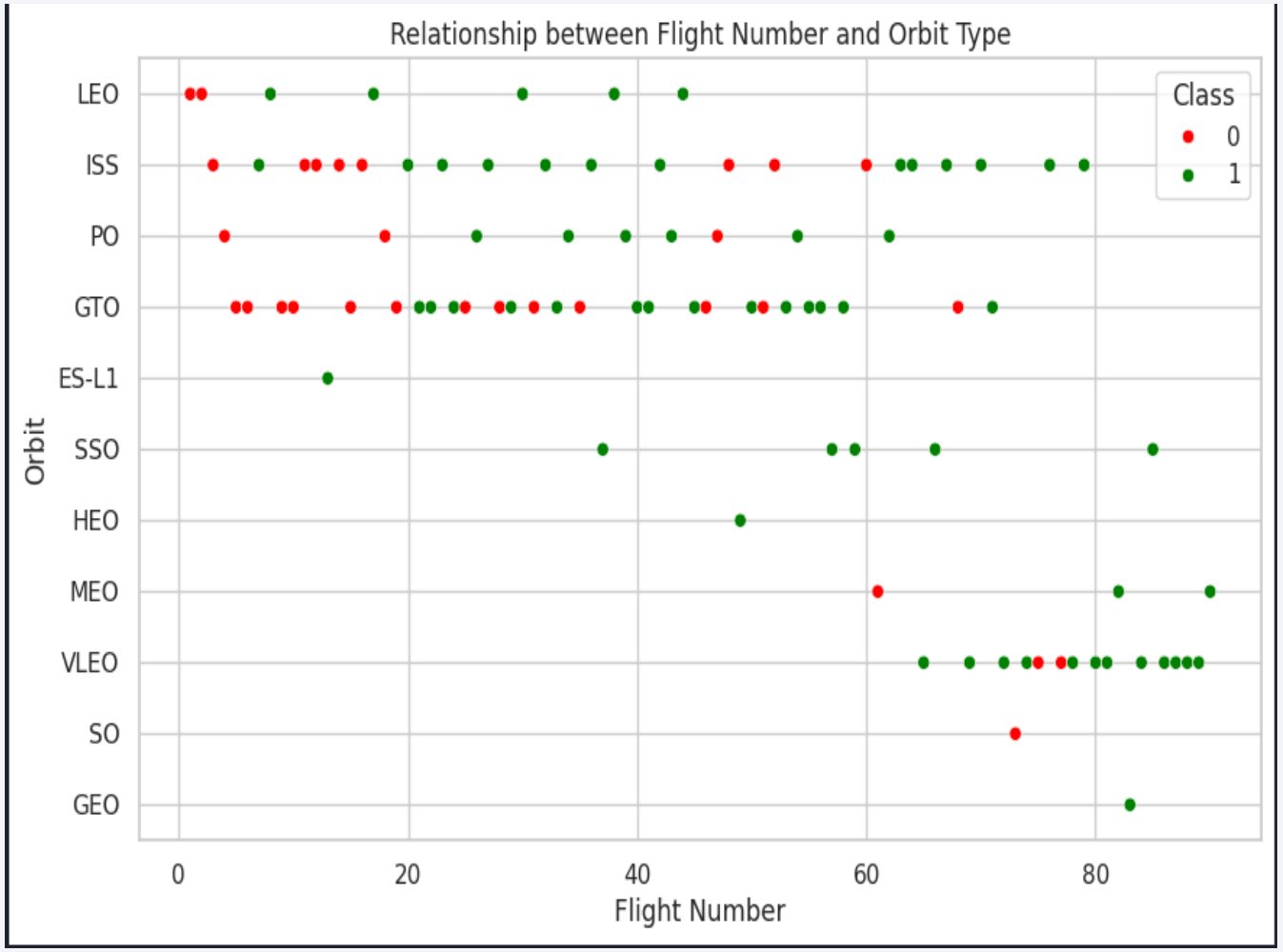
- GREEN = SUCCESS and RED = FAILURE
- The higher the payload the higher the success rate
- Launches with a payload > 7000 KG are very successful
- KSC LC 39A has a success rate of 100% for payloads under 5.500 KG
- VAFB SLC 4E did not launch anything with a payload > 10.000 KG

Success Rate vs. Orbit Type

- Launches that go into Orbits **SSO**, **HEO**, **GEO**, and **ES-L1** have 100% success rates
- Launches into Orbits **VLEO**, **LEO**, **PO**, **MEO**, **ISS**, and **GTO** have success rates between 50%-90%
- Launches into **SO** have a success rate of 0%

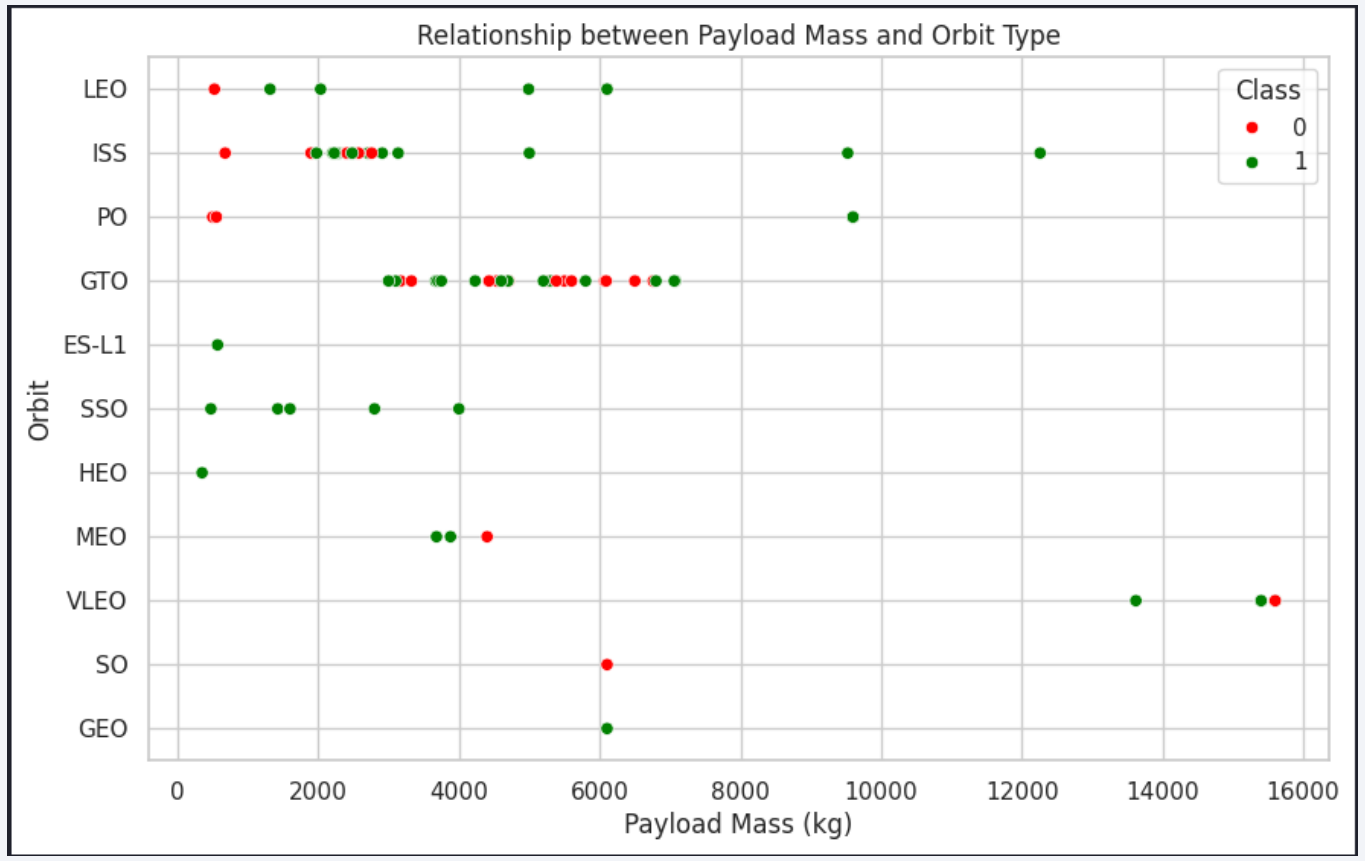


Flight Number vs. Orbit Type



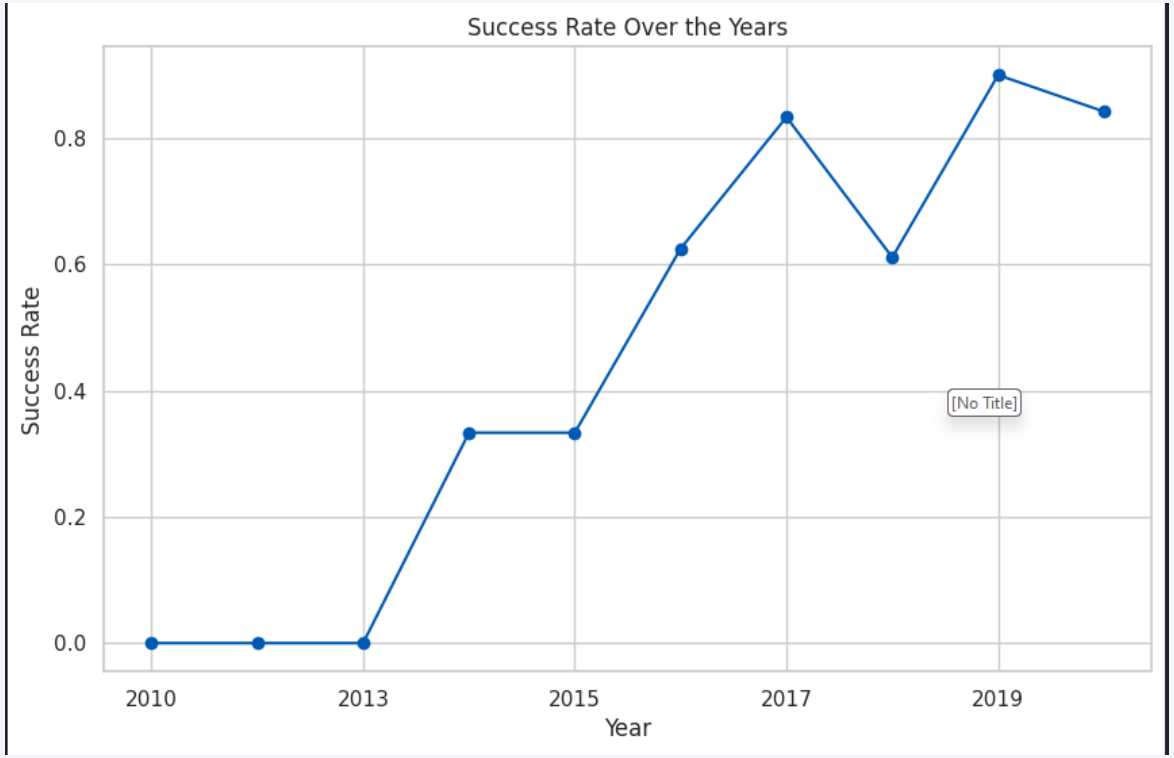
- GREEN = SUCCESS and RED = FAILURE
- In general with an increase in flights so increases the success rate for each orbit
- Orbit **SO** is the exception
- The **GTO** Orbit has a slightly lower success rate even after multiple flights

Payload vs. Orbit Type



- GREEN = SUCCESS and RED = FAILURE
- Payloads > 5.000 KG tend to have a higher success rate per orbit (with the exception of **GTO**)
- Orbits **ES-L1**, **SSO**, **HEO**, and **MEO** have high success rates if the payload is < 4.000 KG

Launch Success Yearly Trend



- The trend line shows a rapid improvement in success rate from 2013 onwards
- In 2018 there was a slight dip in success rate
- In 2017 and again in 2019-2020 the success rate is > 80%

All Launch Site Names

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
```

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

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

- The query on the left returns all the distinct launch sites
- There are four unique launch sites
- These data are useful if we want to locate the sites on the map later on

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT *
FROM SPACEXTABLE
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40

- The query on the left returns 5 records where launch sites begin with `CCA`
- This query is used as a sample to get a feel for the data

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS TotalPayloadMass
FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

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

TotalPayloadMass
45596

- The query on the left calculates the total payload carried by boosters from NASA
- The total payload mass carried by boosters from NASA is 45.596 KG

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVGPayloadMass
FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

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

AVGPayloadMass
2928.4

- The query on the left calculates the average payload mass carried by booster version F9 v1.1
- On average the Falcon 9 v1.1 Booster version carried a load of 2.928,4 KG

First Successful Ground Landing Date

```
%%sql
SELECT MIN(Date) AS First_Successful_Landing
FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
```

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

First_Successful_Landing
2015-12-22

- The query on the left returns the date of the first successful landing outcome on a ground pad
- The first successful landing occurred on December 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT DISTINCT Booster_Version AS Successful_Boosters
FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

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

- The query on the left returns the names of boosters which have successfully landed on drone ship and had a payload mass between 4000 and 6000
- There are 4 Booster types with a payload mass between 4000 and 6000 that made a successful landing on a drone ship

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT
    SUM(CASE WHEN Mission_Outcome LIKE '%uccess%' THEN 1 ELSE 0 END) AS Success,
    SUM(CASE WHEN Mission_Outcome NOT LIKE '%uccess%' THEN 1 ELSE 0 END) AS Failure
FROM
    SPACEXTABLE;
```

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

Success	Failure
100	1

```
%%sql
SELECT
    COUNT(CASE WHEN LOWER(Landing_Outcome) LIKE '%success%' THEN 1 END) AS Success,
    COUNT(CASE WHEN LOWER(Landing_Outcome) NOT LIKE '%success%' THEN 1 END) AS Failure
FROM
    SPACEXTABLE;
```

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

Success	Failure
61	40

- The query above (left) shows the total number of successful and failure mission outcomes
- Of the 101 missions only 1 outcome was a failure
- (Be aware that the mission outcome is not the same as the landing outcome (right), which has 61 successful landings and 40 failures)

Boosters Carried Maximum Payload

```
%%sql
SELECT DISTINCT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_)
FROM SPACEXTABLE);
```

```
 * 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

- The query on the left returns the names of the booster which have carried the maximum payload mass
- There are 12 Booster Versions that carried the maximum payload of 15.600 KG

2015 Launch Records

```
%%sql
SELECT SUBSTR(Date, 6, 2) AS Month, Landing_Outcome AS Failure, Booster_Version, Launch_Site
FROM SPACEXTABLE
WHERE SUBSTR(Date, 0, 5) = '2015' AND Landing_Outcome = 'Failure (drone ship)';
```

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

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

- The query above returns the failed landing outcomes on a drone ship, their booster versions, and their launch sites for the year 2015
- There were only 2 failures (1 in January and 1 in April) with 2 Booster versions, both launched from CCAFS LC-40

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

```
%%sql
SELECT Landing_Outcome, COUNT(*) AS Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Count DESC;
```

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

Landing_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

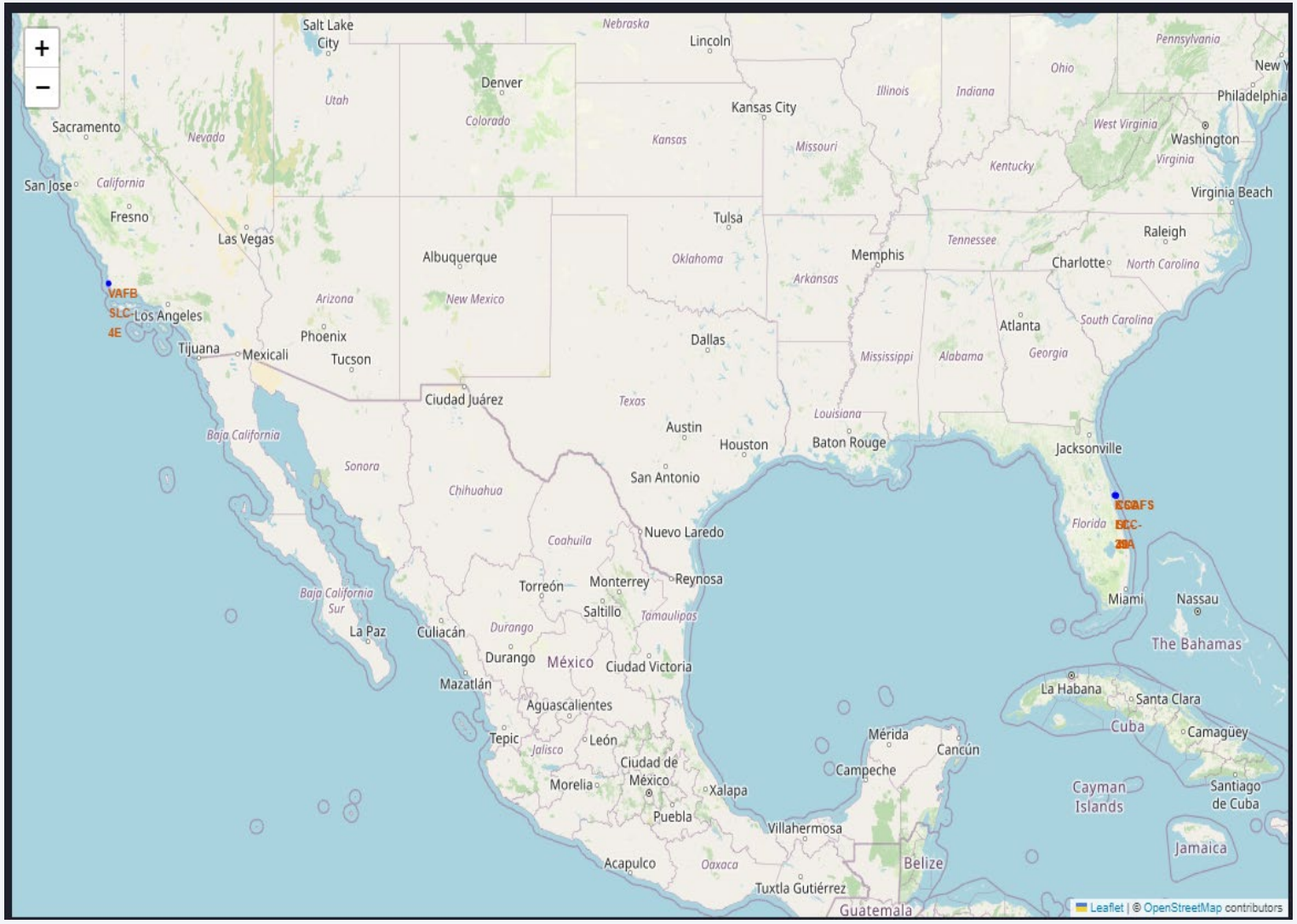
- The query on the left returns the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

- In 1/3 of the cases between 2010 and 2017 no landing was attempted (landing_outcome ‘No attempt’ with 10 counts)

Section 3

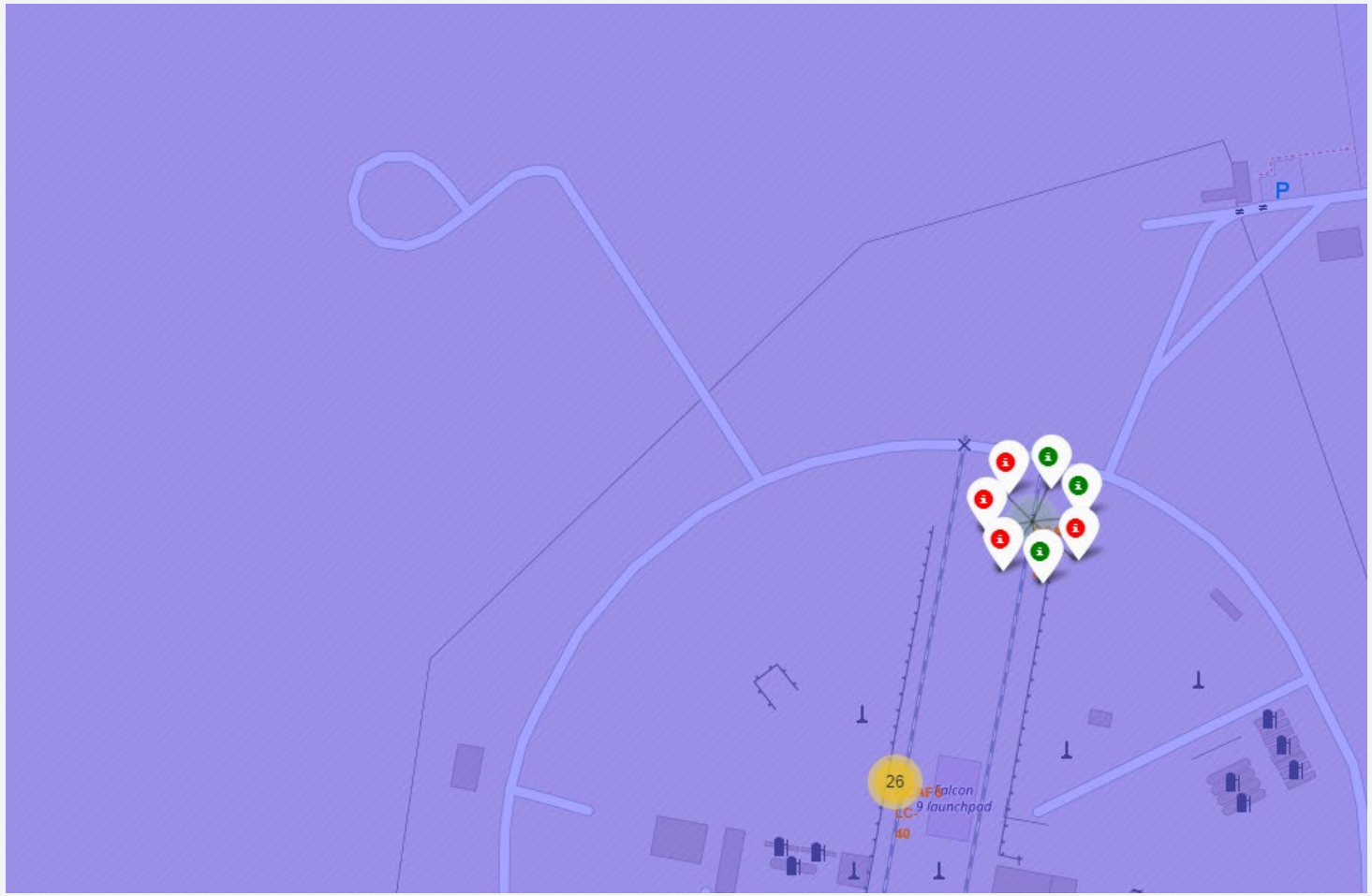
Launch Sites Proximities Analysis

Falcon 9 Launch Site Locations



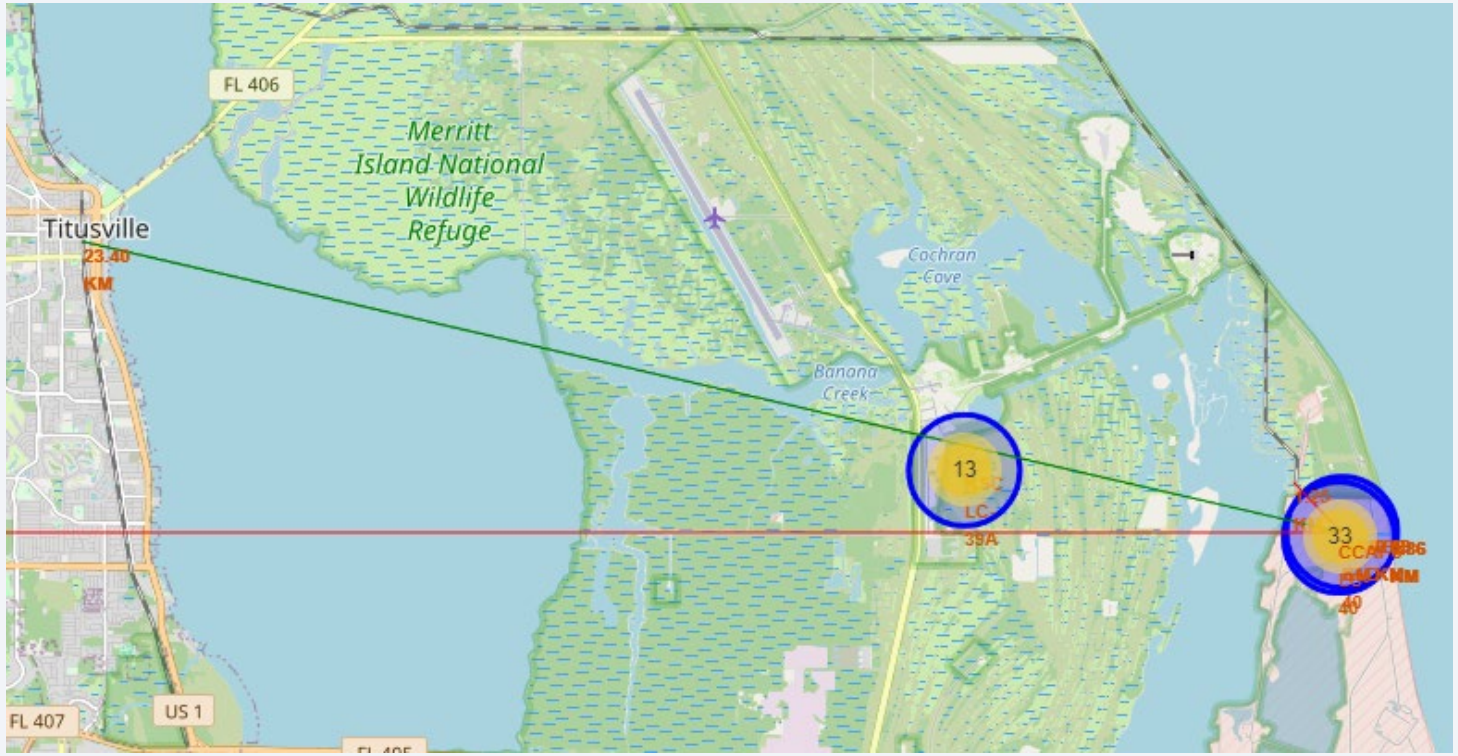
- The launch Sites are located near the coast and only a short distance from the equator
- Being close to the equator gives the rockets a natural boost from the Earth's rotation
- This is cost efficient since there is less fuel necessary to get into orbit
- Source: NASA Explanations

Launch Outcomes

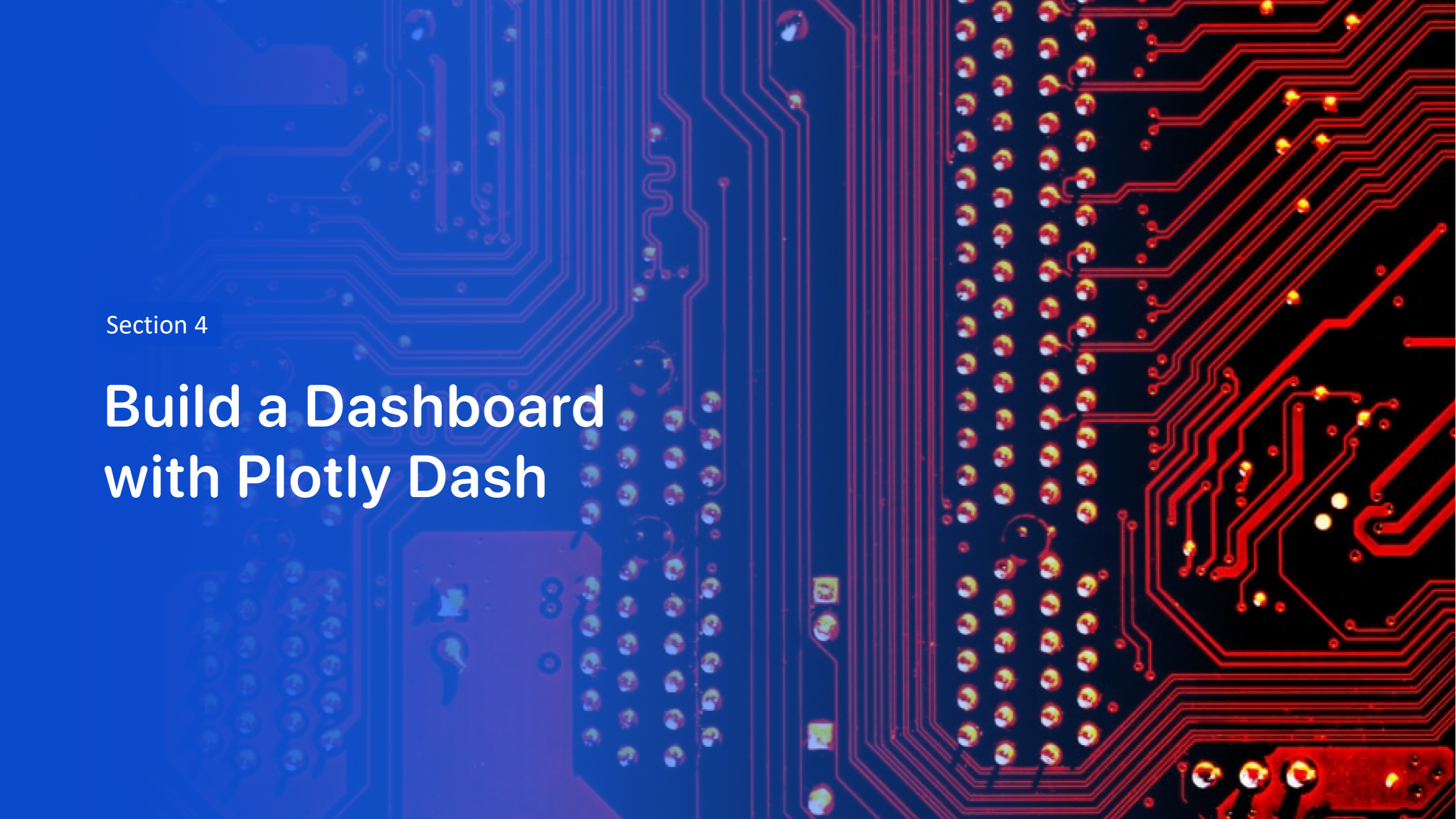


- GREEN = SUCCESS
- RED = FAILURE
- We are looking at CCAF SLC-40 with 3 successful and 7 unsuccessful launches
- Success rate of (42,9%)

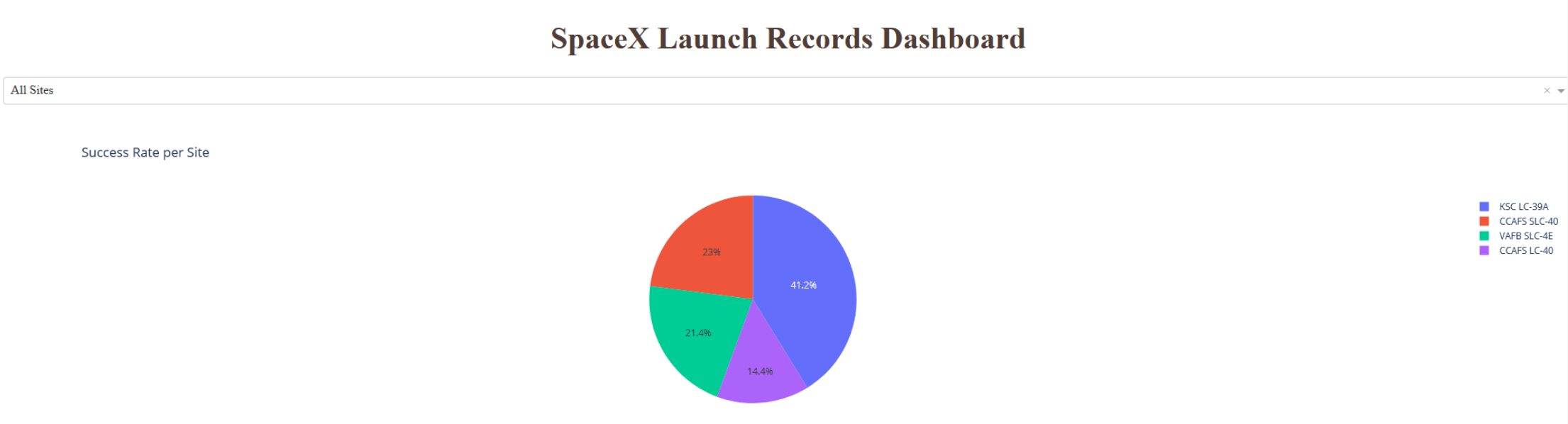
Distance from Launch Site to Proximities



- Site: CCAFS SLC-40
- Proximities:
  - Nearest Coast Line: 0,86 KM
  - Nearest Railway: 1,25 KM
  - Nearest Highway: 0,59 KM
  - Nearest City (Titusville): 23,40 KM

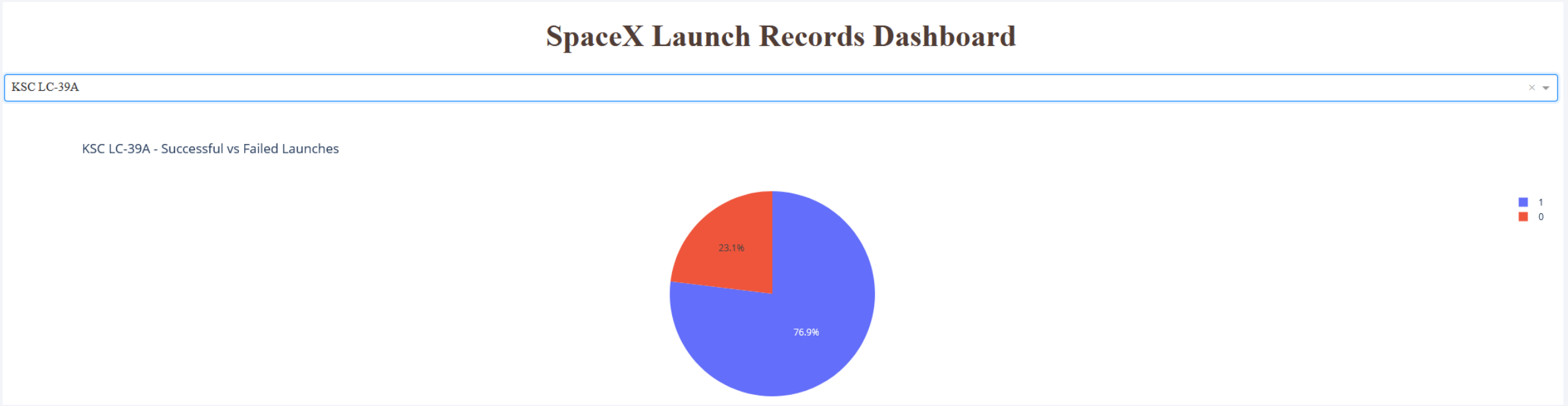

Section 4
Build a Dashboard with Plotly Dash

Success by Launch Site



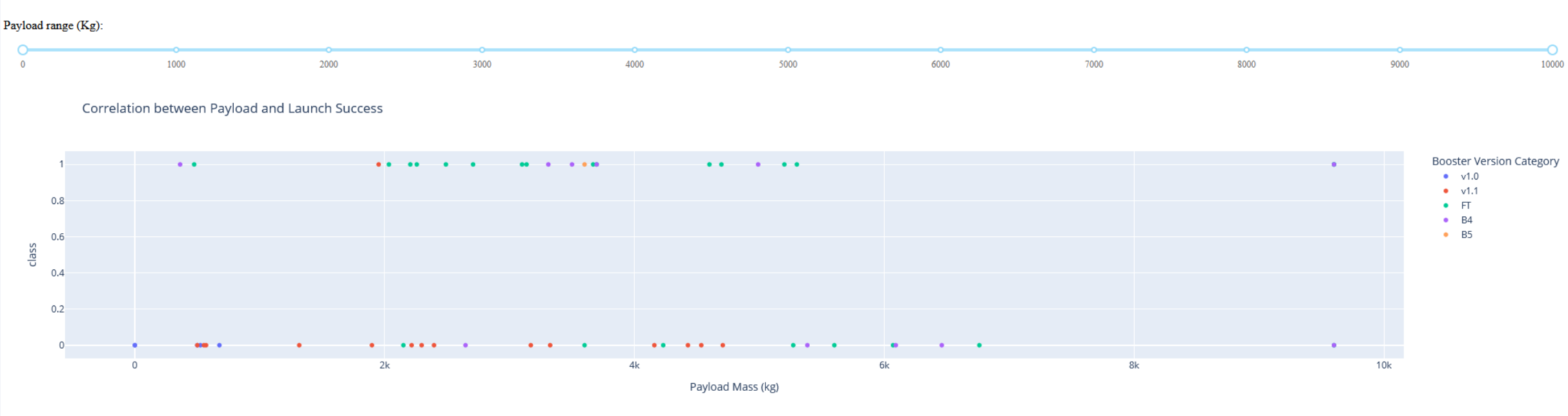
- The default page shows the distribution of successful launches by site
- KSC CL-39A has the largest contribution to the overall success rate
- You can select each of the 4 launch sites to check their individual success rate

Launch Success for KSC LC-39A



- BLUE = SUCCESS and RED = FAILURE
- KSC LC-39A has the highest success rate (76,9%)
- There were 10 successful and 3 failed launches

Outcomes: Payload Mass vs Booster version



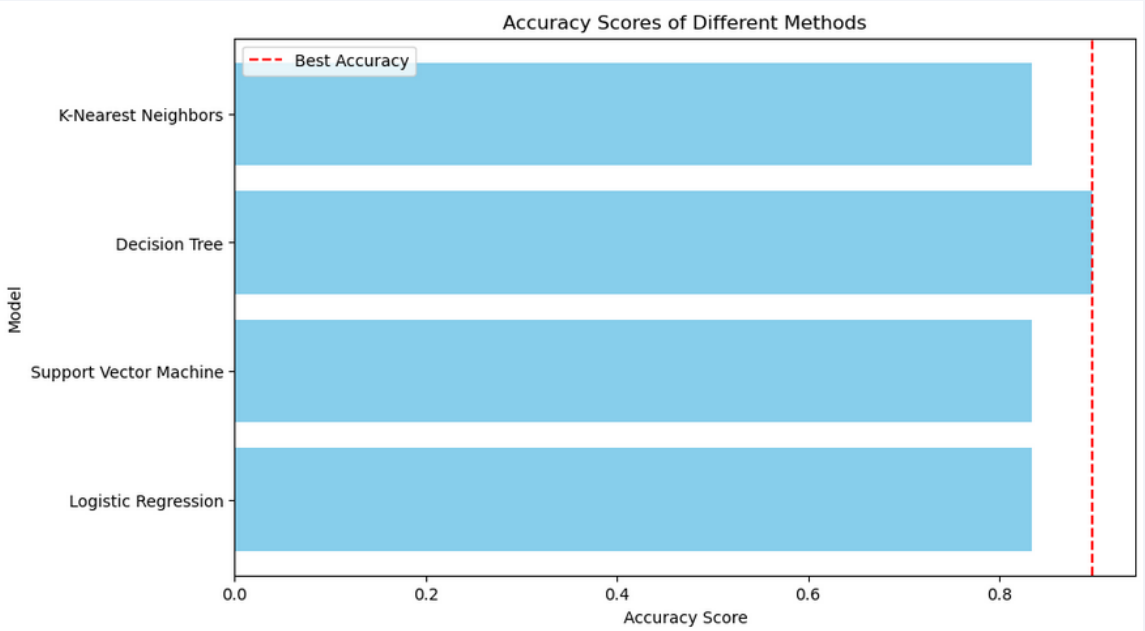
- 1 = Success and 0 = Failure
- No matter the payload mass, Booster v1.1 has predominantly failures
- Most successes occurred when the payload mass was between 2.000 and 6000 KG, and the FT Booster version accounted for most of these successes

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- All the models performed with an accuracy > 80%
- The Decision Tree outperformed the other models by a small margin (89%)

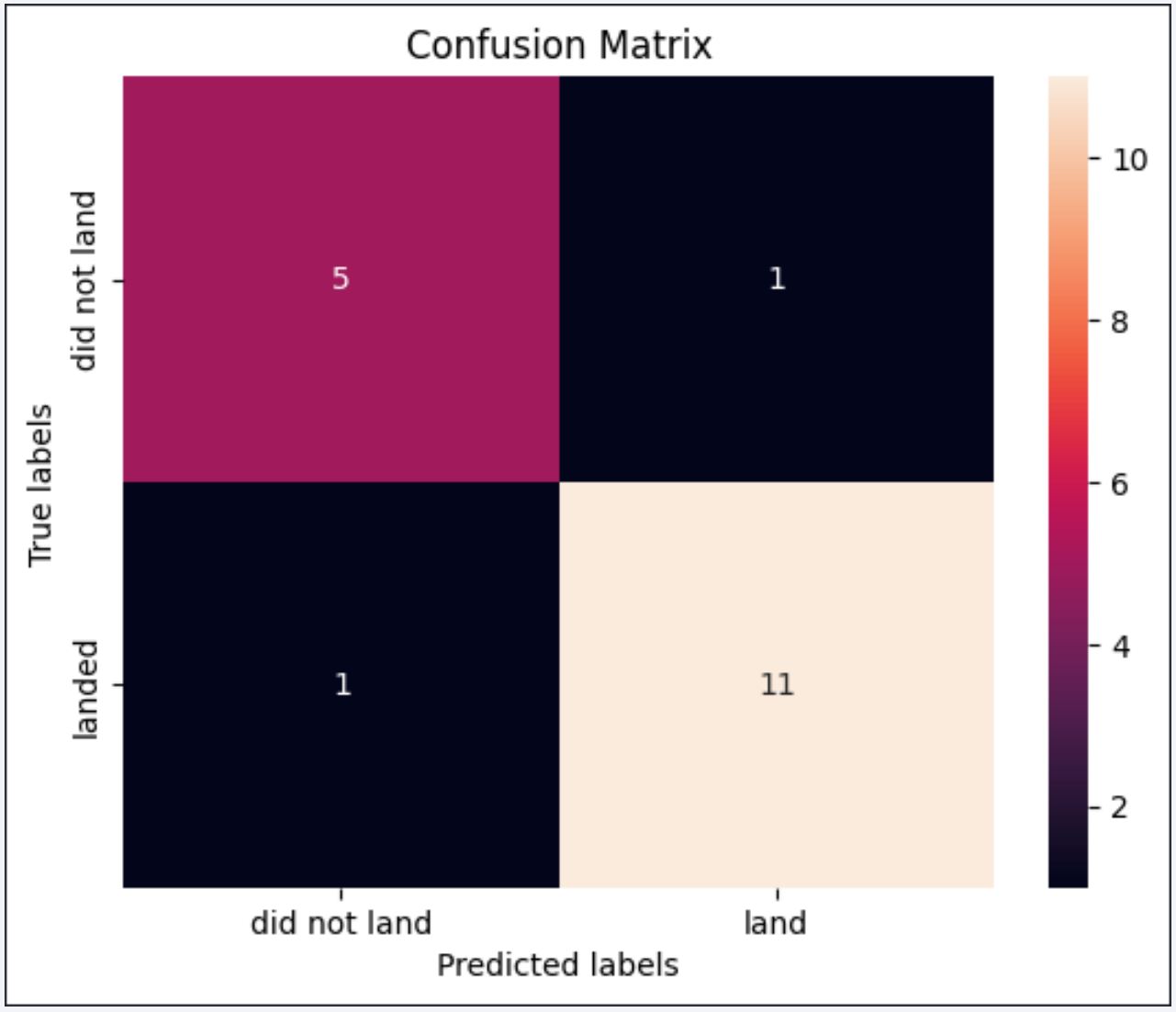



```
accuracy_scores = {
    'Logistic Regression': logreg_cv.score(X_test, Y_test),
    'Support Vector Machine': svm_cv.score(X_test, Y_test),
    'Decision Tree': tree_cv.score(X_test, Y_test),
    'K-Nearest Neighbors': knn_cv.score(X_test, Y_test)}

best_method = max(accuracy_scores, key=accuracy_scores.get)
best_accuracy = accuracy_scores[best_method]

print(f"The method that performs best is {best_method} with an accuracy of {best_accuracy:.2f}.")
```

```
The method that performs best is Decision Tree with an accuracy of 0.89.
```

Confusion Matrix

- Decision Tree Model Confusion Matrix Explanation
  - True Positive (TP): 11
  - True Negative (TN): 5
  - False Positive (FP): 1
  - False Negative (FN): 1
- With only 1 FP and 1 FN this model is highly accurate with an almost 90% accuracy



Conclusions

- The success rate for landing outcomes has shown a consistent improvement over time and with each launch.
- Launch sites are strategically located near the equator, close to a coast, and sufficiently distant from other proximities to avoid hindrances, yet still convenient for material transportation.
- Payloads within the range of 2,000 to 6,000 kilograms had the highest success rate, with the FT Booster version being predominantly associated with successful landings.
- KSC LC-39A emerged as the launch site with the highest success rate.
- Among the models tested, the Decision Tree Model exhibited the highest accuracy, achieving an 89% accuracy rate in predicting landing outcomes.

Appendix

- For any Python code snippets, SQL Queries, charts, Notebook outputs or data sets I refer you to my GitHub repository (Click Here)

Thank you!