IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sabrina Pinto
10/02/2025

Outline

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

Executive Summary

Summary of methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashbord with Plotly Dash
- Predictive Analysis (Classification)

Summary of all results

- Exploratory Data Analysis results
- Interactive Analytics demo
- Predictive Analysis results

Introduction

- Project background and context

SpaceX is one of the most successful companies of the commercial space age, making space travel affordable. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars where other providers cost upwards of 165 million dollars each; much of the savings is because SpaceX can reuse the first stage.

Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. We are going to predict if SpaceX will reuse the first stage, using publicly available information to train a machine learning model.

- Problems you want to find answers

How do specific variables (such as payload mass, launch site location, type of orbit) affect the outcome of the first stage landing?

Has the number of successful landings increased over the years?

What Machine Learning Algorithm is better suited for classification in this specific case?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  The data was collected using SpaceX Rest API and using Web Scraping on Wikipedia.

- Perform data wrangling

  The data was filtered, formatted and processed to deal with missing values. Categorical features were converted with one-hot-encoding.

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

  Four different Machine Learning models where trained on the processed data, then evaluated on their results.

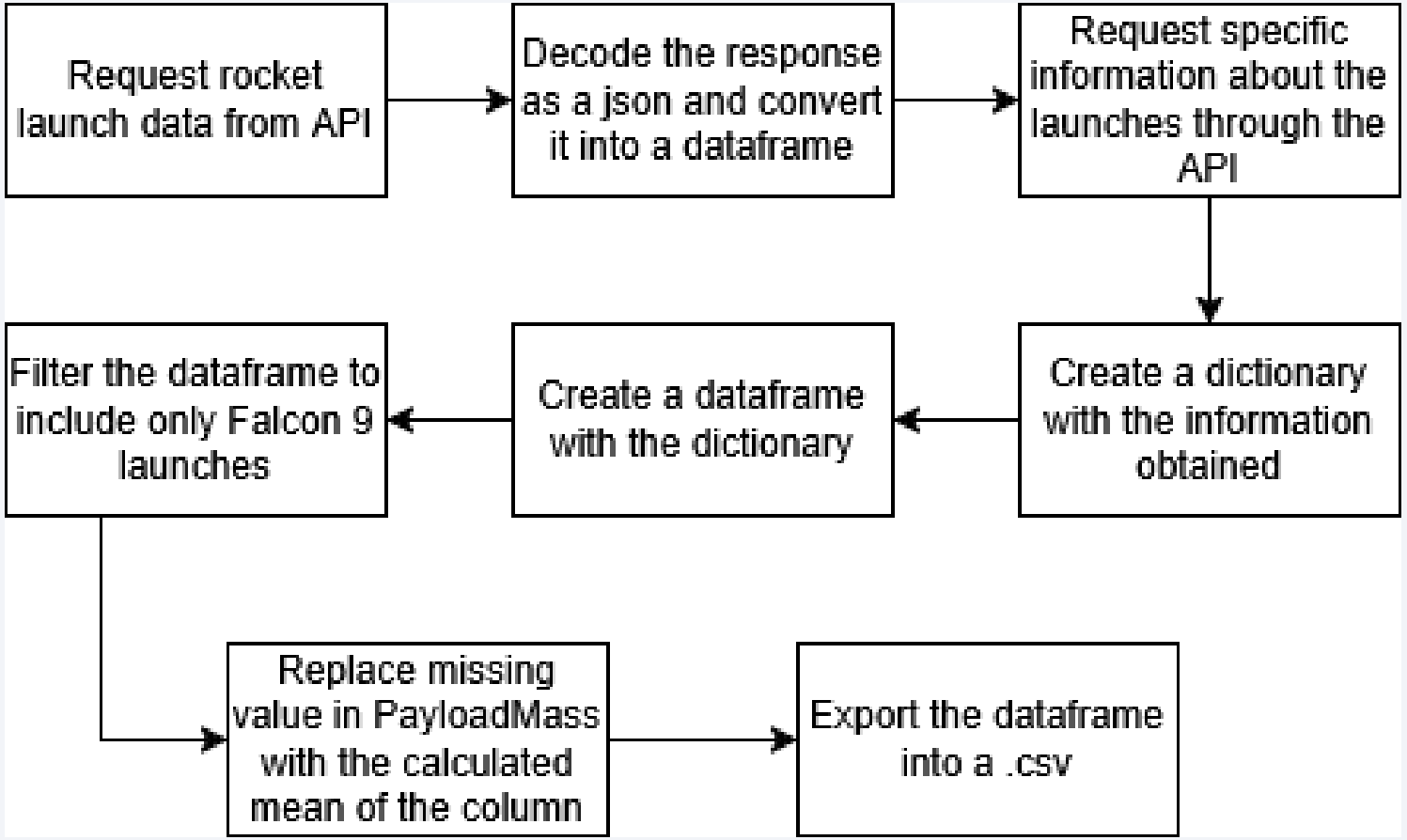
Data Collection

The data collection process was done through a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

Both processes were necessary to ensure the collection of all the necessary data for the analysis.

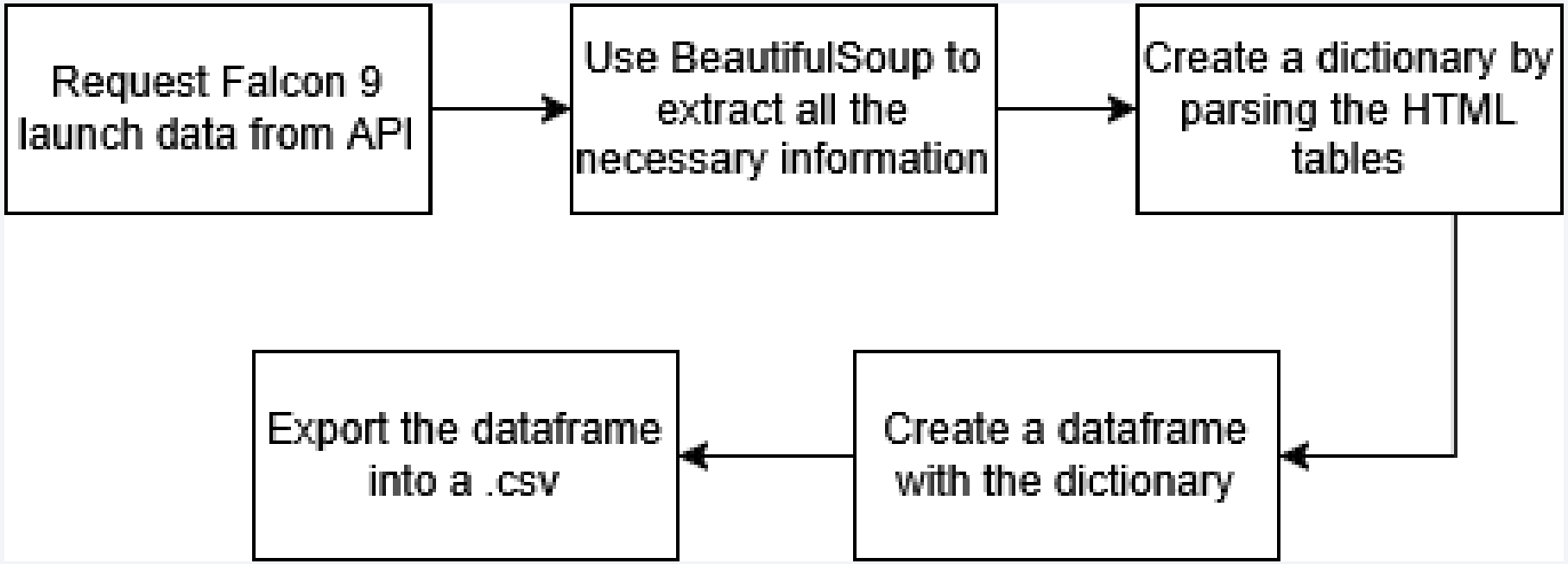
Data Collection – SpaceX API

GitHub URL: Data Collection API



Data Collection - Scraping

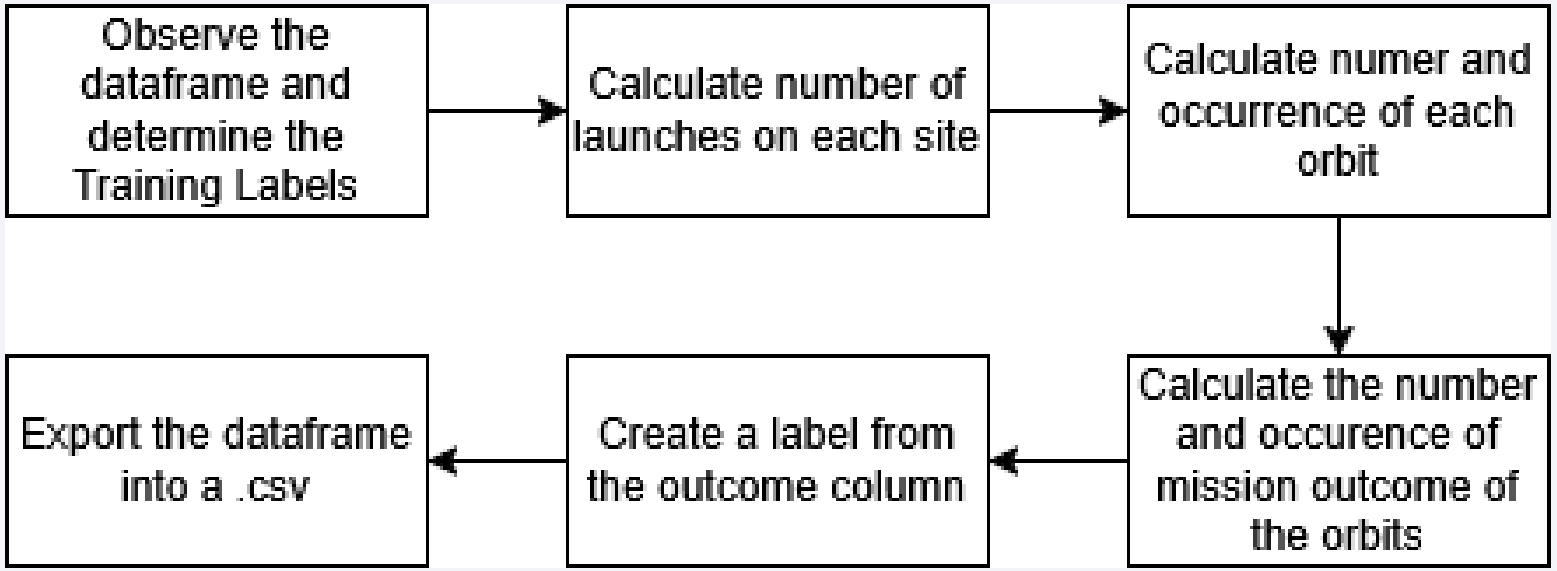
GitHub URL: Data Collection with Web Scraping



Data Wrangling

GitHub URL: Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident. These values where converted into binary values as Training Labels, with 1 representing a successful landing and 0 representing an unsuccessful landing.



EDA with Data Visualization

GitHub URL: EDA with Data Visualization

On the collected and processed data, the following charts were plotted:

Scatter Plots, used to show the relationships between:

- FlightNumber and PayloadMass
- FlightNumber and LaunchSite
- PayloadMass and LaunchSite
- FlightNumber and Orbit

Bar Chart, used to show the comparison of categorical values, namely of:

- Success Rate of each Orbit type

Line Chart, used to show the trends in data over time, specifically of:

- Success Yearly Trend

EDA with SQL

GitHub URL: EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass between 4000 and 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the month of failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order

Build an Interactive Map with Folium

GitHub URL: Interactive Visual Analytics with Folium

- Markers of all Launch Sites:

Added Marker with Circle, Popup Label and Text Label for each launch site in the dataframe.

- Coloured Markers of the launch outcomes for each Launch Site:

Added coloured Markers of success and failed launches using Marker Cluster to visualize the success rates of each launch site.

- Distances between a Launch Site to its proximities:

Added coloured Lines to show distances between a launch site and the closest coastline, city, railway and highway.

Build a Dashboard with Plotly Dash

GitHub URL: Interactive Dashboard with Ploty Dash

- Launch Sites Dropdown List:

Added a dropdown list to enable Launch Site selection.

- Pie Chart of Success Launches:

Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

- Slider of Payload Mass Range:

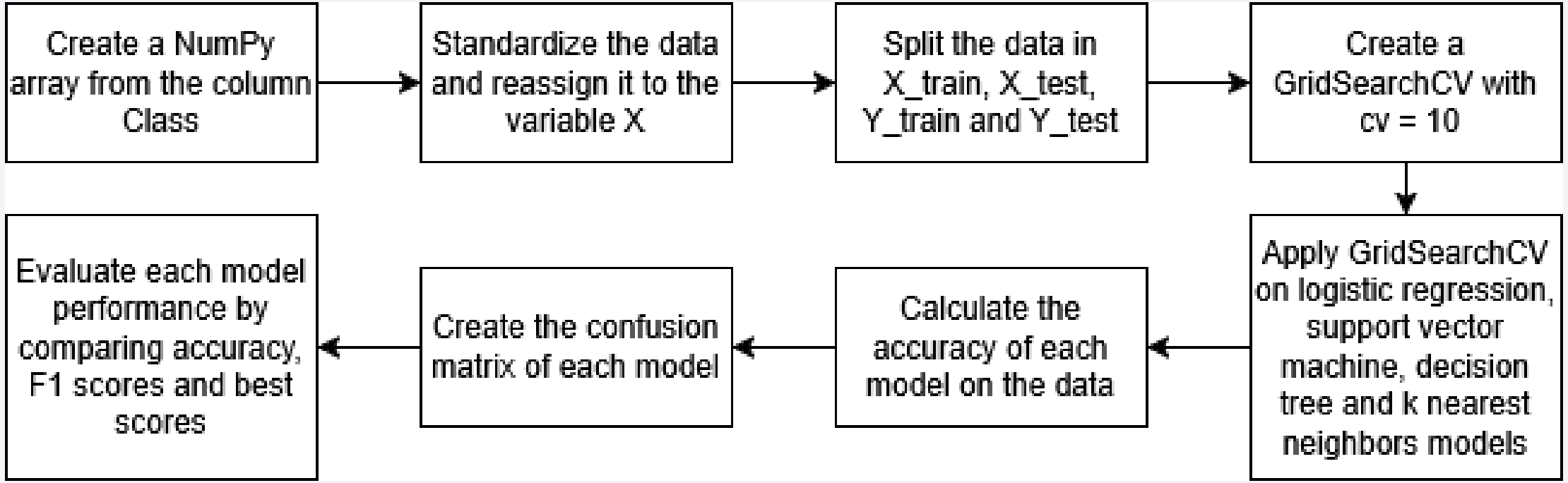
Added a slider to select Payload range.

- Scatter Chart of Payload Mass and Success Rate:

Added a scatter chart to show the correlation between Payload and Launch Success.

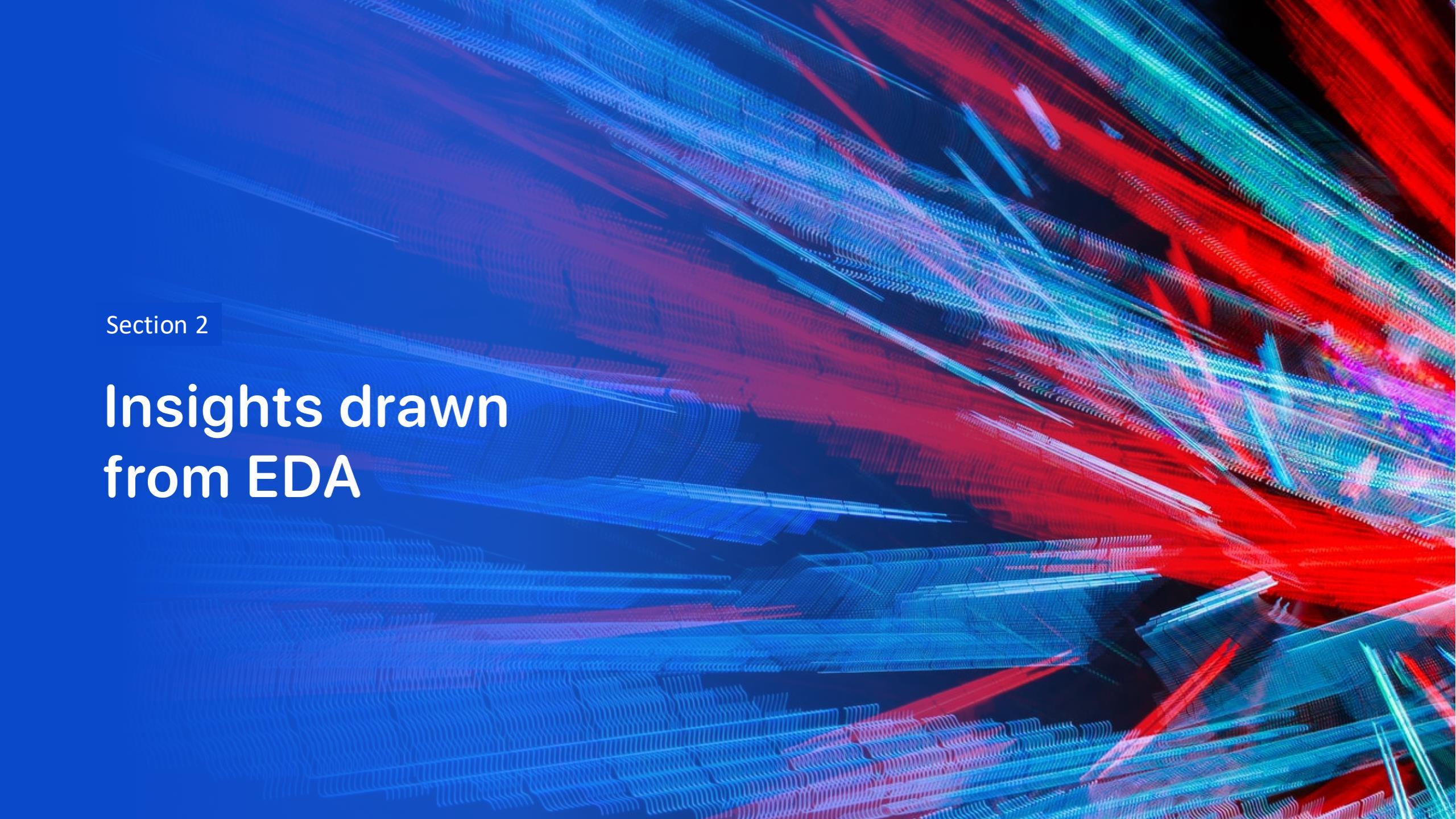
Predictive Analysis (Classification)

GitHub URL: Machine Learning Prediction



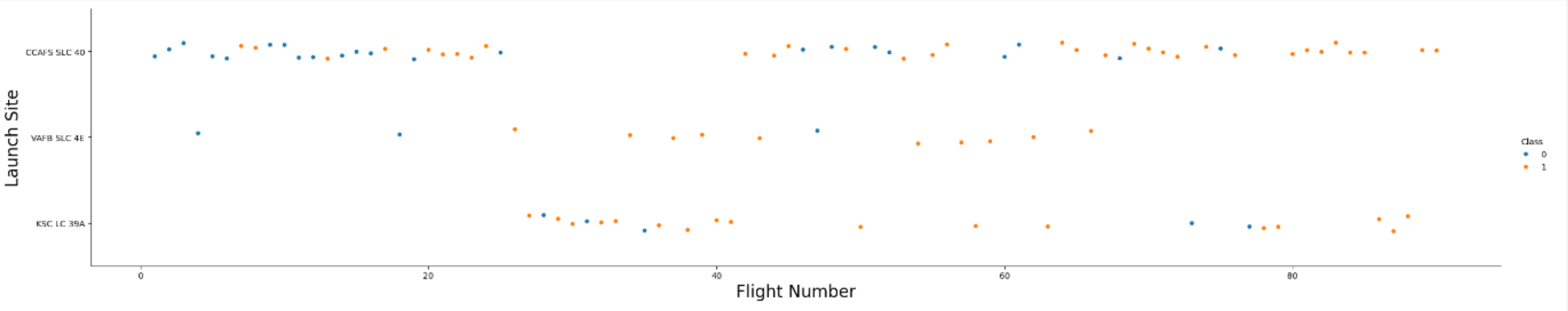
Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

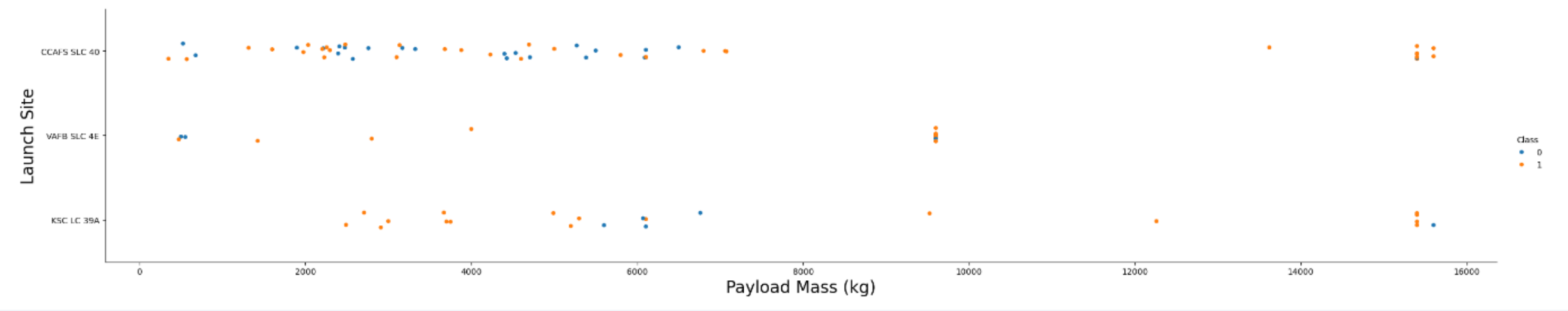
The plot shows how most of the earliest launches failed while the latest mostly succeeded. Additionally, the plot shows how the CCAFS SLC 40 is the launch site of almost half the total launches, but the VAFB SLC 4E and the KSC LC 39A sites have more successful launches.



Payload vs. Launch Site

The plot shows how for each launch with higher payload mass, the success rate is higher, in fact most of the launches with a payload higher than 7000kg is successful.

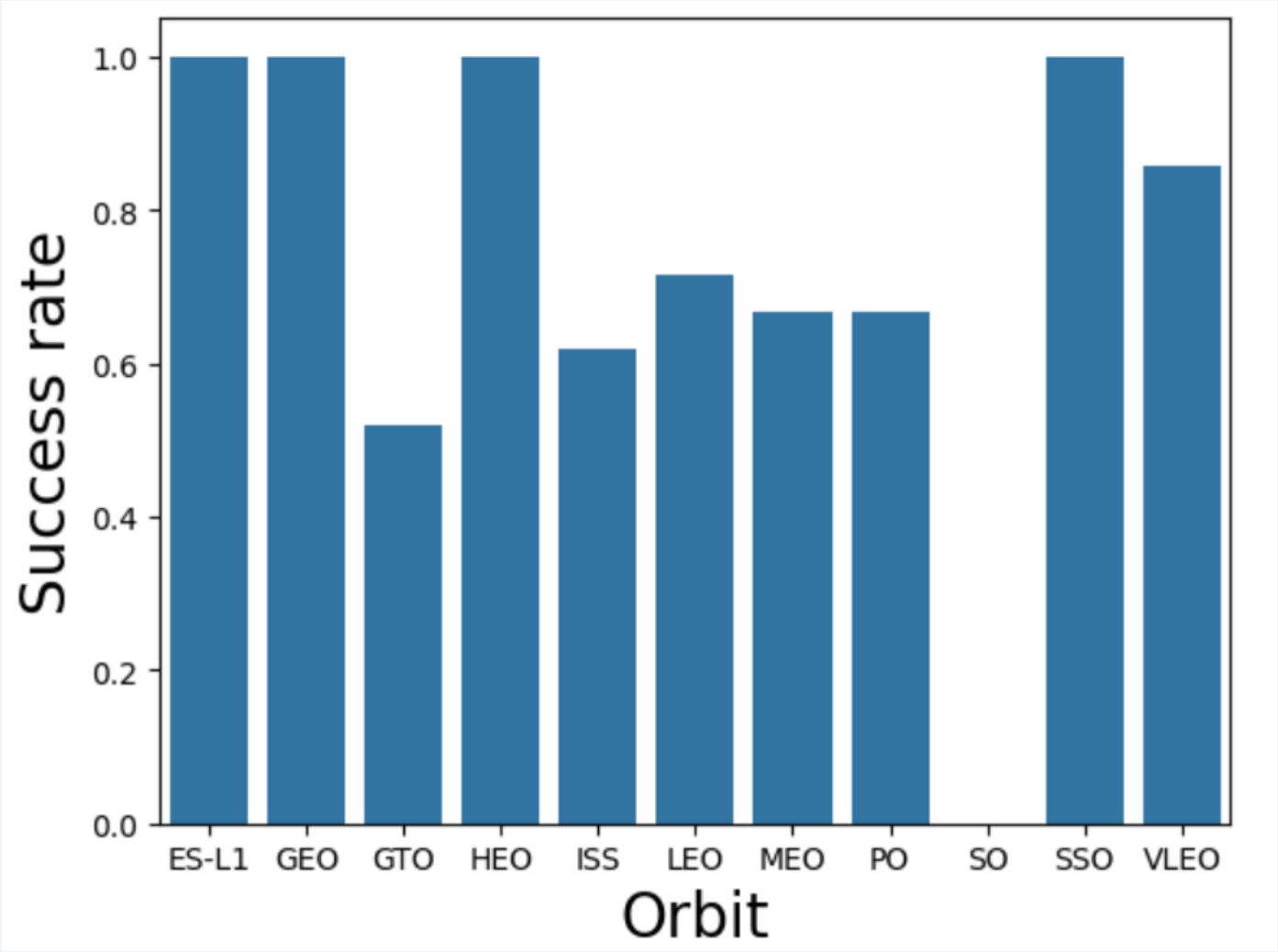
Additionally, the plot shows how the KSC LC 39A site has a success rate of 100% for all launches where the payload mass was lower than 5500kg.



Success Rate vs. Orbit Type

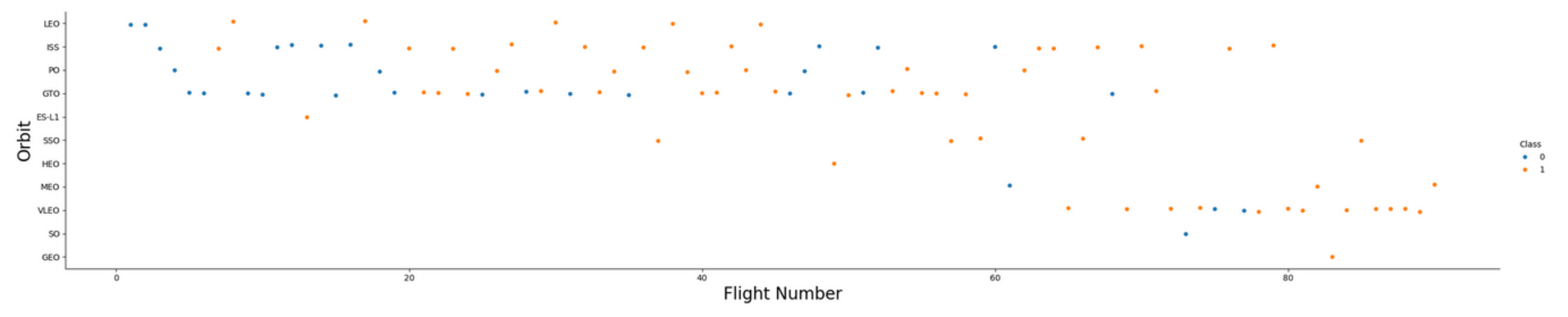
The chart shows that:

- Four orbit types have a success rate of 100% (ES-L1, GEO, HEO and SSO)
- One orbit type has a success rate of 0% (SO)
- The rest of the types have a success rate between 50% and 85%



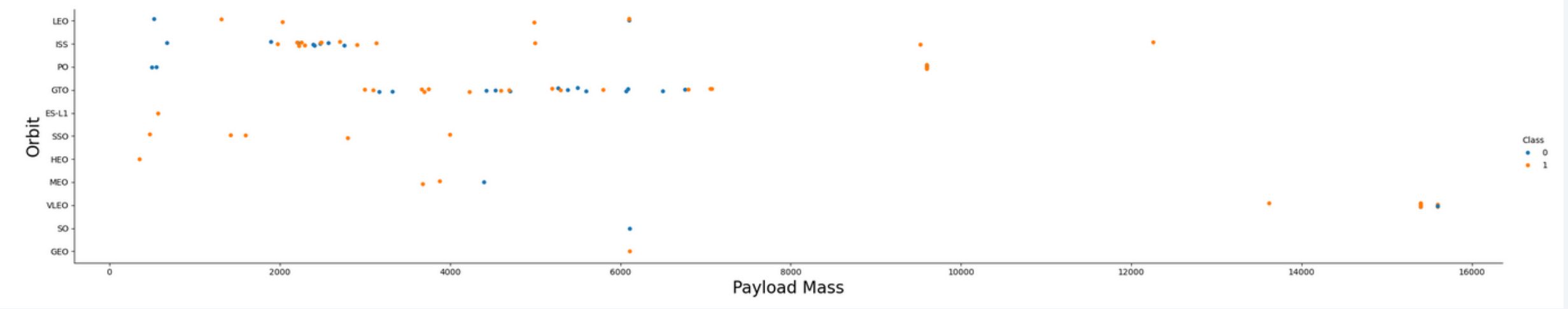
Flight Number vs. Orbit Type

The plot shows that the success rate for the LEO orbit type is correlated to the number of flights, but there also seems to be no correlation between the number of flights and the GTO orbit type.



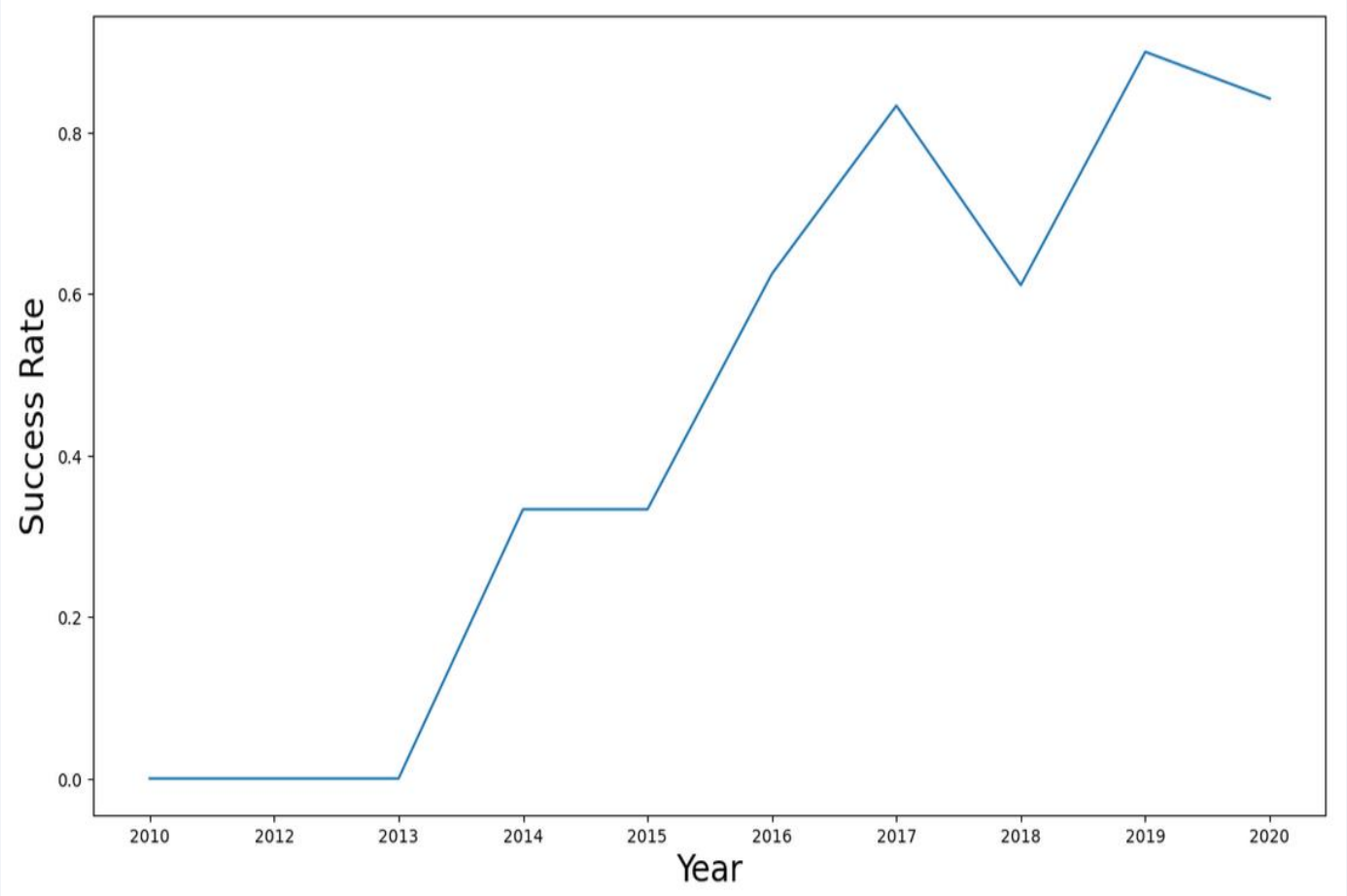
Payload vs. Orbit Type

The plot shows how payloads between 3500kg and 6500kg affect negatively launches for the GTO orbit type, while weights of between 2000kg and 4000kg seem to affect positively launches for the ISS orbit type.



Launch Success Yearly Trend

The chart clearly shows how the success rate of the launches has been increasing since 2013, with a slight decline between 2017 and 2018, followed by further increase.



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

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	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

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

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

total_payload_mass_NASA_CRS
45596

Average Payload Mass by F9 v1.1

```
%%sql SELECT AVG(PAYLOAD_MASS__KG_) AS average_payload_mass_F9v11 FROM SPACEXTABLE
    WHERE BOOSTER_VERSION LIKE 'F9 v1.1%';
```

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

average_payload_mass_F9v11
2534.6666666666665

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

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql SELECT BOOSTER_VERSION AS success FROM SPACEXTABLE
    WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

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

Total Number of Successful and Failure Mission Outcomes

```
%%sql SELECT total_successes, total_failures FROM
    (SELECT COUNT(*) AS total_successes FROM SPACEXTABLE
    WHERE MISSION_OUTCOME LIKE 'Success%') success_table,
    (SELECT COUNT(*) total_failures FROM SPACEXTABLE
    WHERE MISSION_OUTCOME LIKE 'Failure%') failure_table;
```

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

total_successes	total_failures
100	1

Boosters Carried Maximum Payload

```
%%sql SELECT 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

2015 Launch Records

```
%%sql SELECT SUBSTR(Date, 6,2) AS Landing_Month, LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
    FROM SPACEXTABLE
    WHERE LANDING_OUTCOME = 'Failure (drone ship)' AND SUBSTR(DAte,0,5) = '2015';
```

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

Landing_Month	Landing_Outcome	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

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

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

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

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

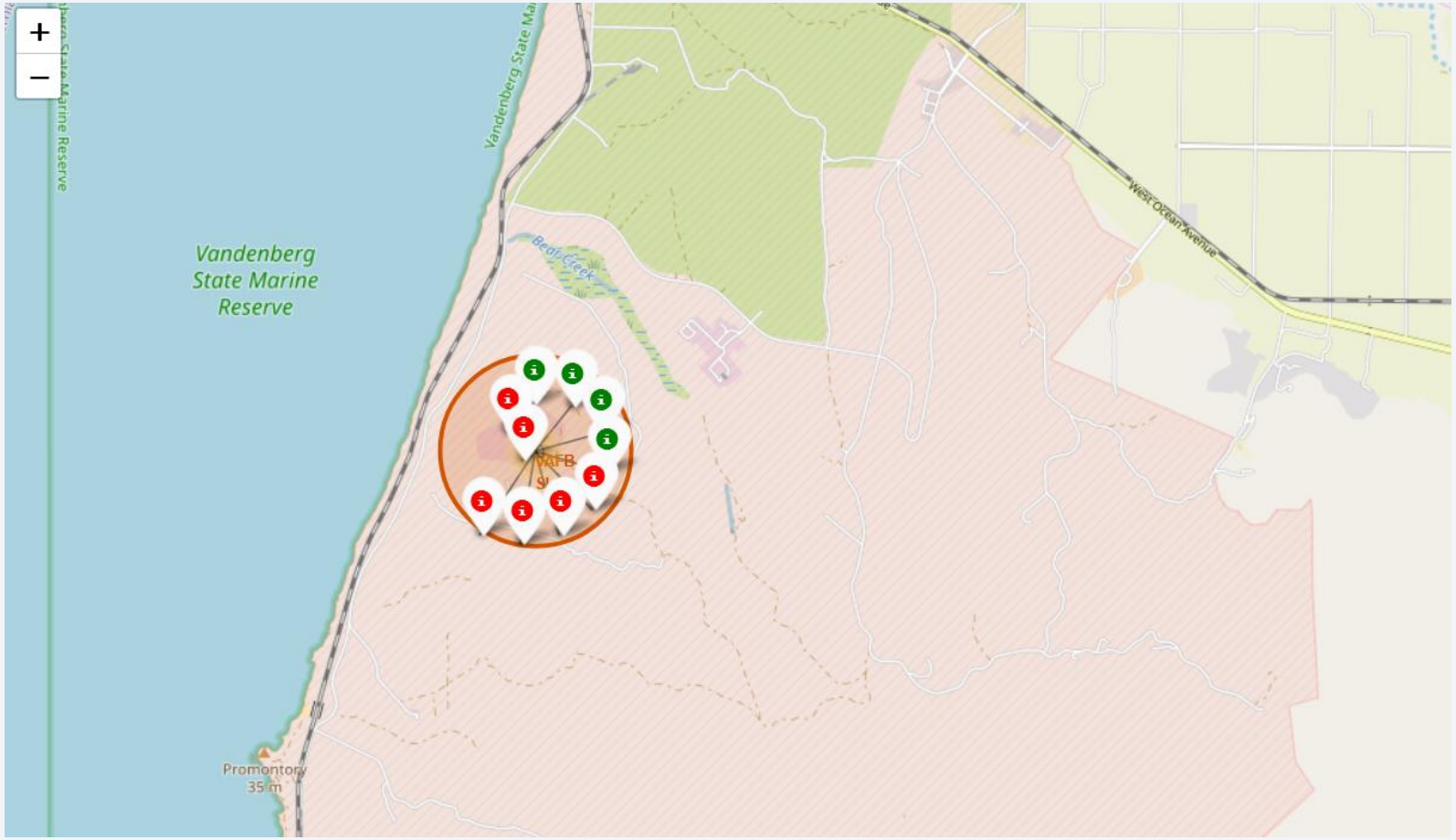
Launch Sites Proximities Analysis

All Launch Sites on Map

The map shows markers for all Launch Sites

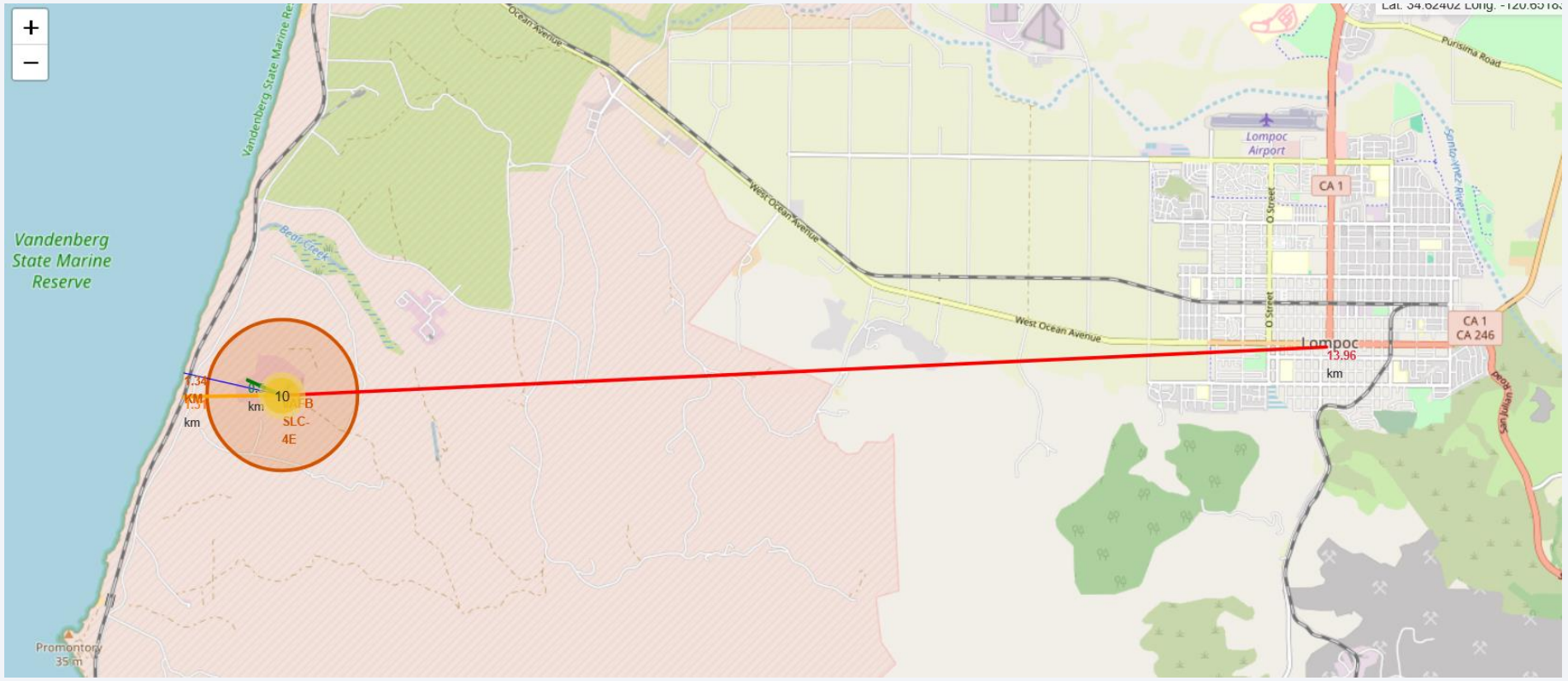
Succeeded and failed launches on Map

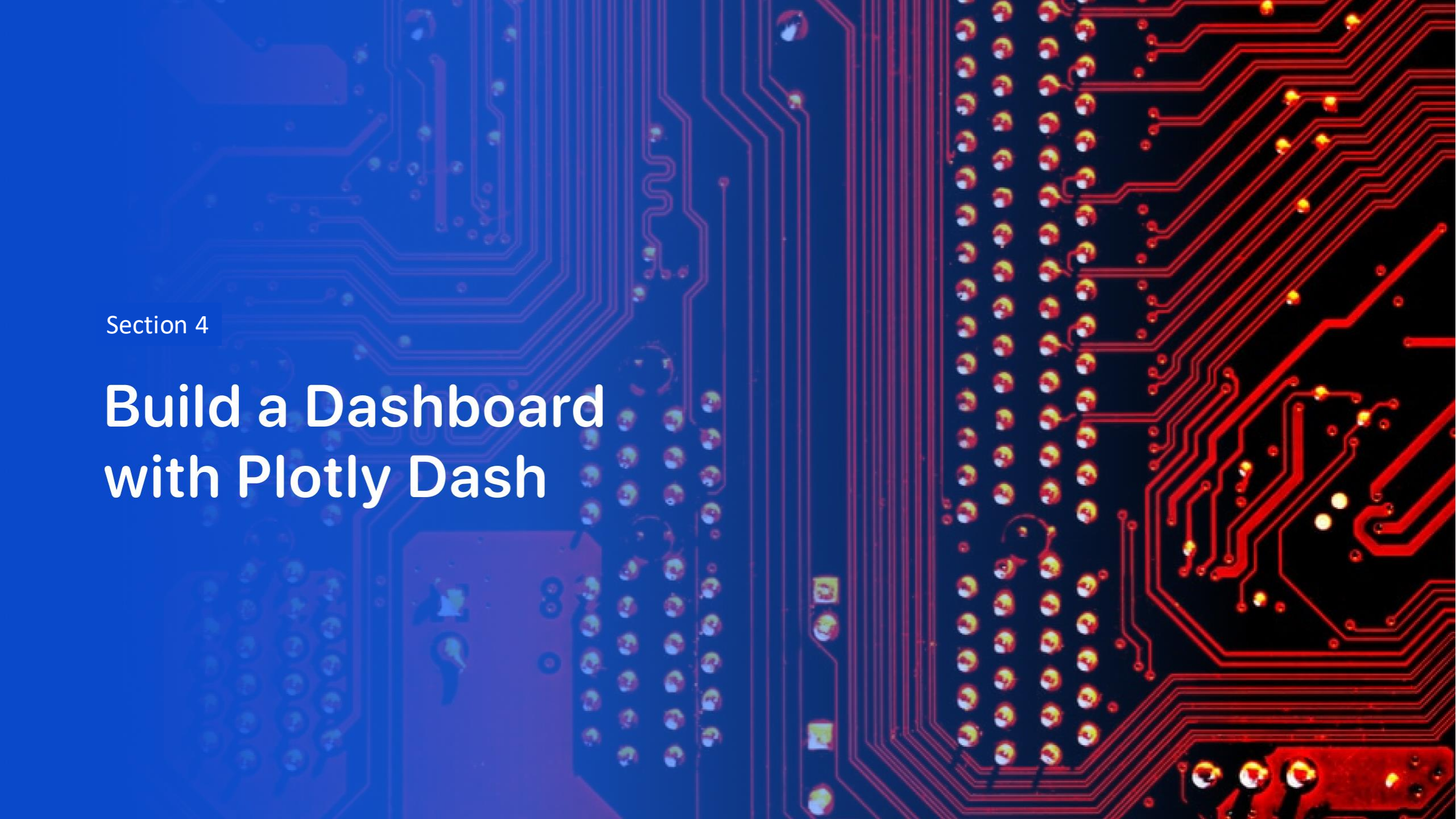
When zoomed in on each Launch Site, the map shows green and red colored markers to identify successful and failed launches.



Distance between Launch Site and proximities

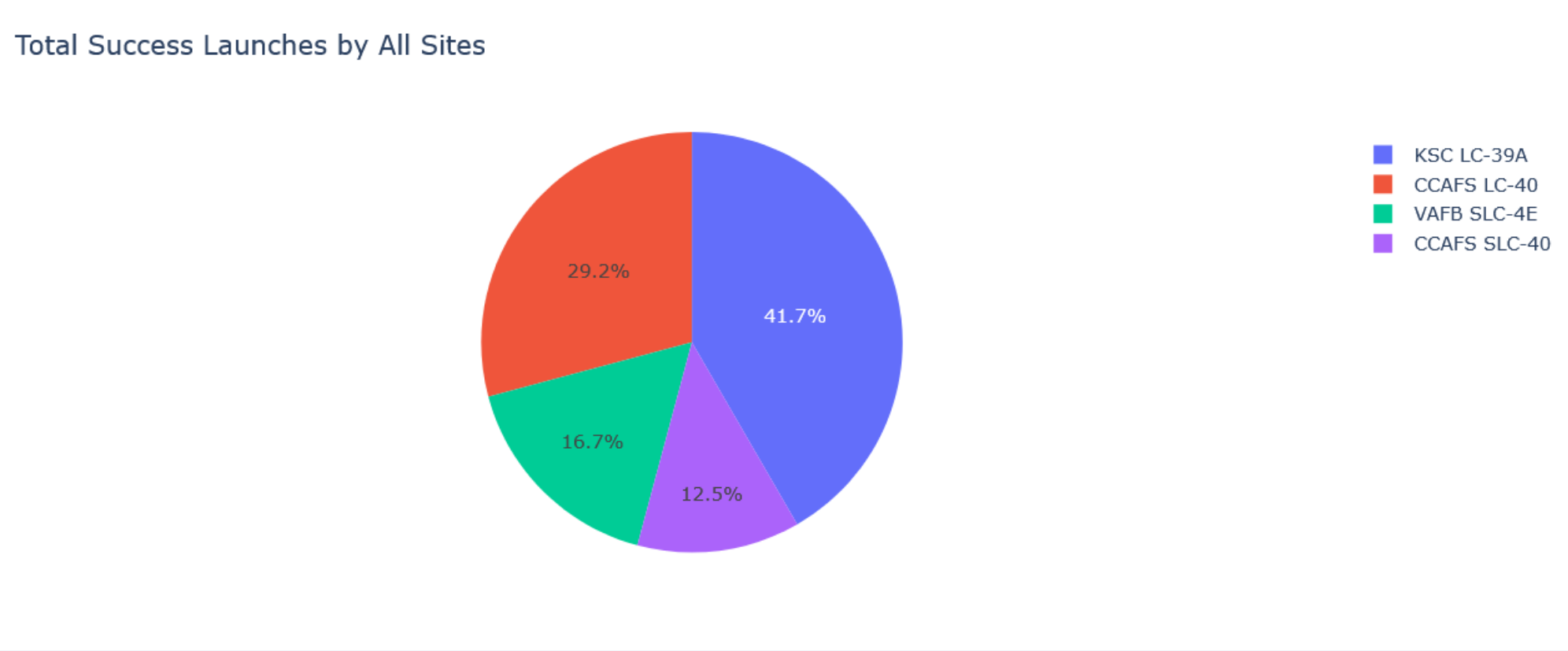
For each Launch Site, a coloured line is drawn between each site and its relative significant proximities, such as: closest city, closest coastline, closest railway and closest highway.




Section 4
Build a Dashboard with Plotly Dash

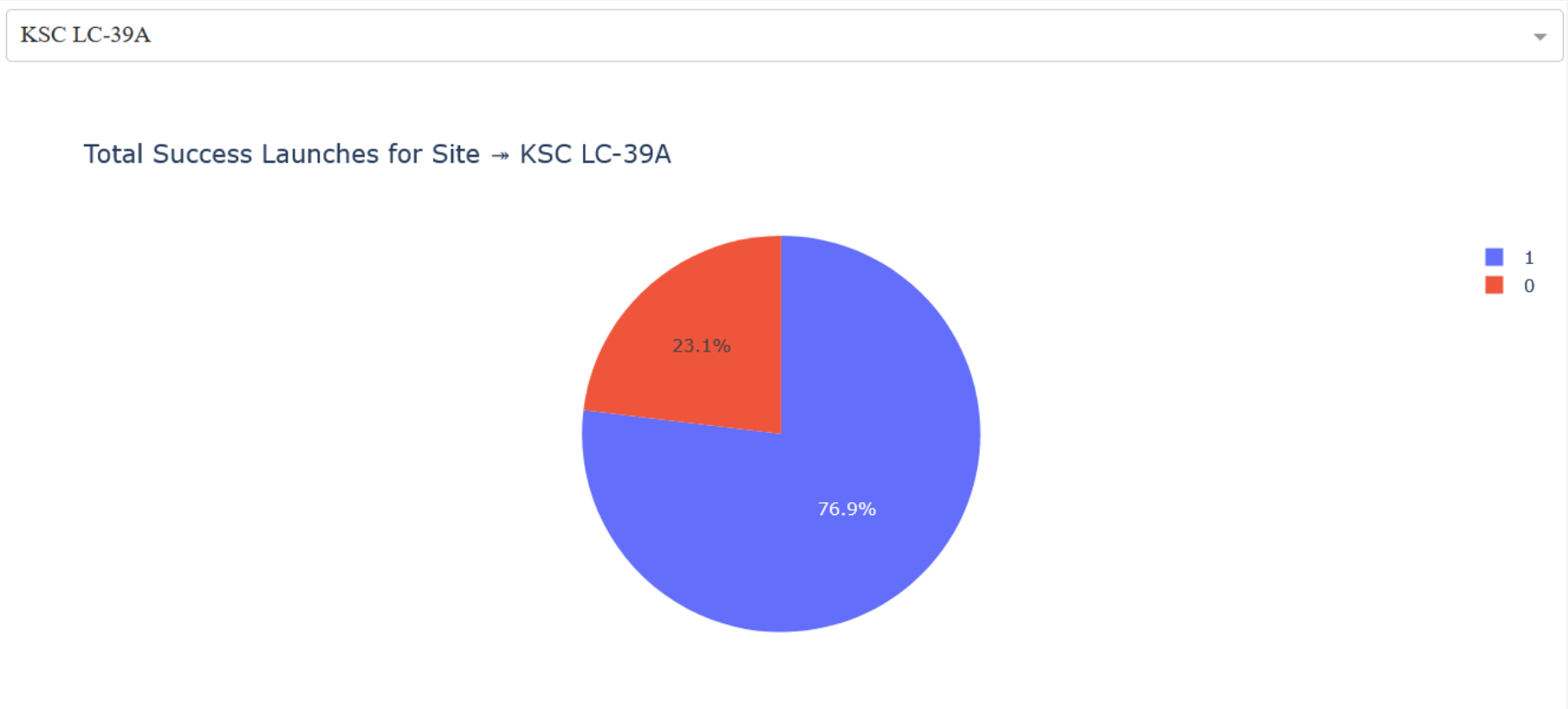
Launch Success for all Sites

The chart shows a comparison of success rates for all Launch Sites. The chart shows how KSC LC-39A clearly has a highest success rate.



Launch Site with highest Launch Success Ratio

The chart shows the rate of successful (in blue) and failed (in red) specifically for the KSC LC-39A which was shown to be the site with highest success rate. The chart shows how it has a 76.9% success ratio.



Payload Mass vs. Launch Outcome for all sites

The two plots show two examples of correlation between Payload Mass and Launch Outcomes, for all sites, with payload weight ranges selected from 2000kg to 8000kg in the first plot, and from 4000kg to 10000kg in the second plot.

Section 5

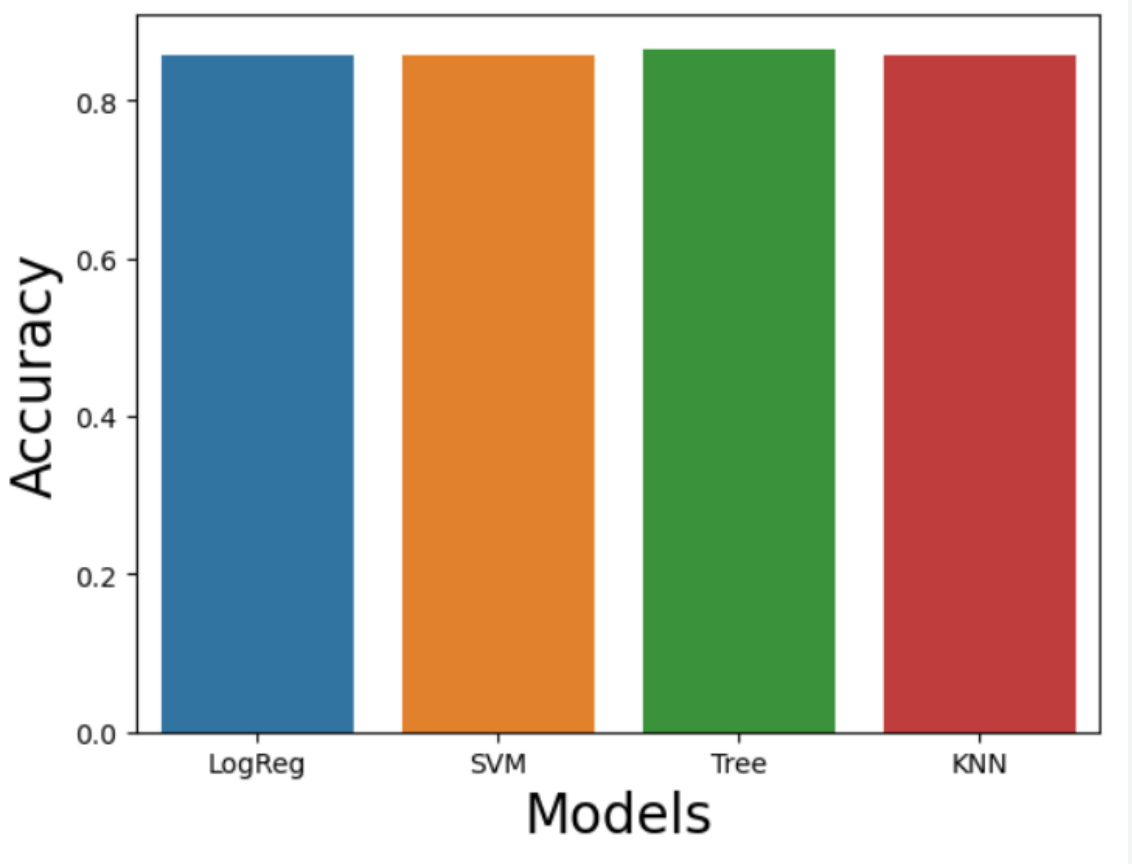
Predictive Analysis (Classification)

Classification Accuracy

Based on the scores obtained for the test sets, it is not possible to confirm which model performs best on accuracy alone.

This might be caused by the size of the sample size (18 samples).

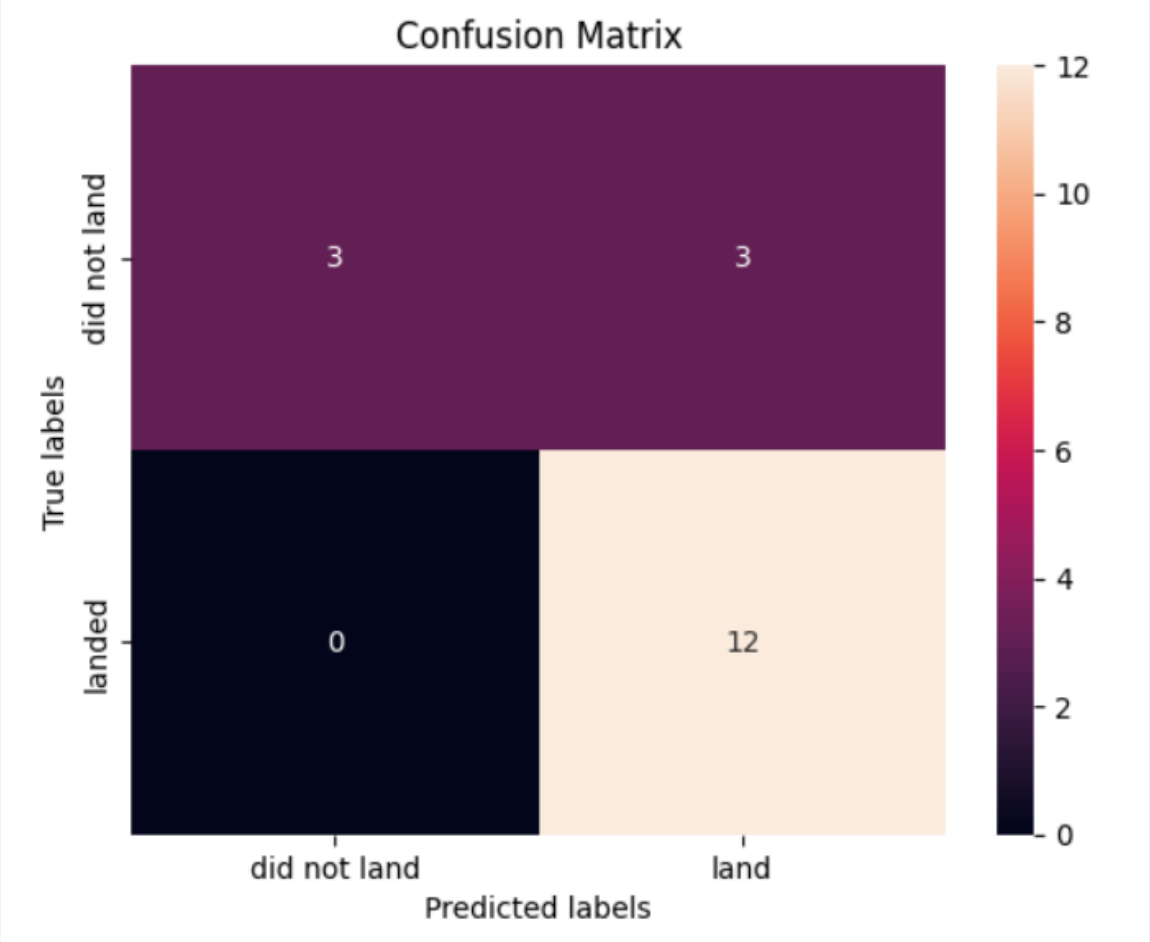
LogReg	0.833333
SVM	0.833333
Tree	0.833333
KNN	0.833333



Confusion Matrix – Logistic Regression

Logistic Regression

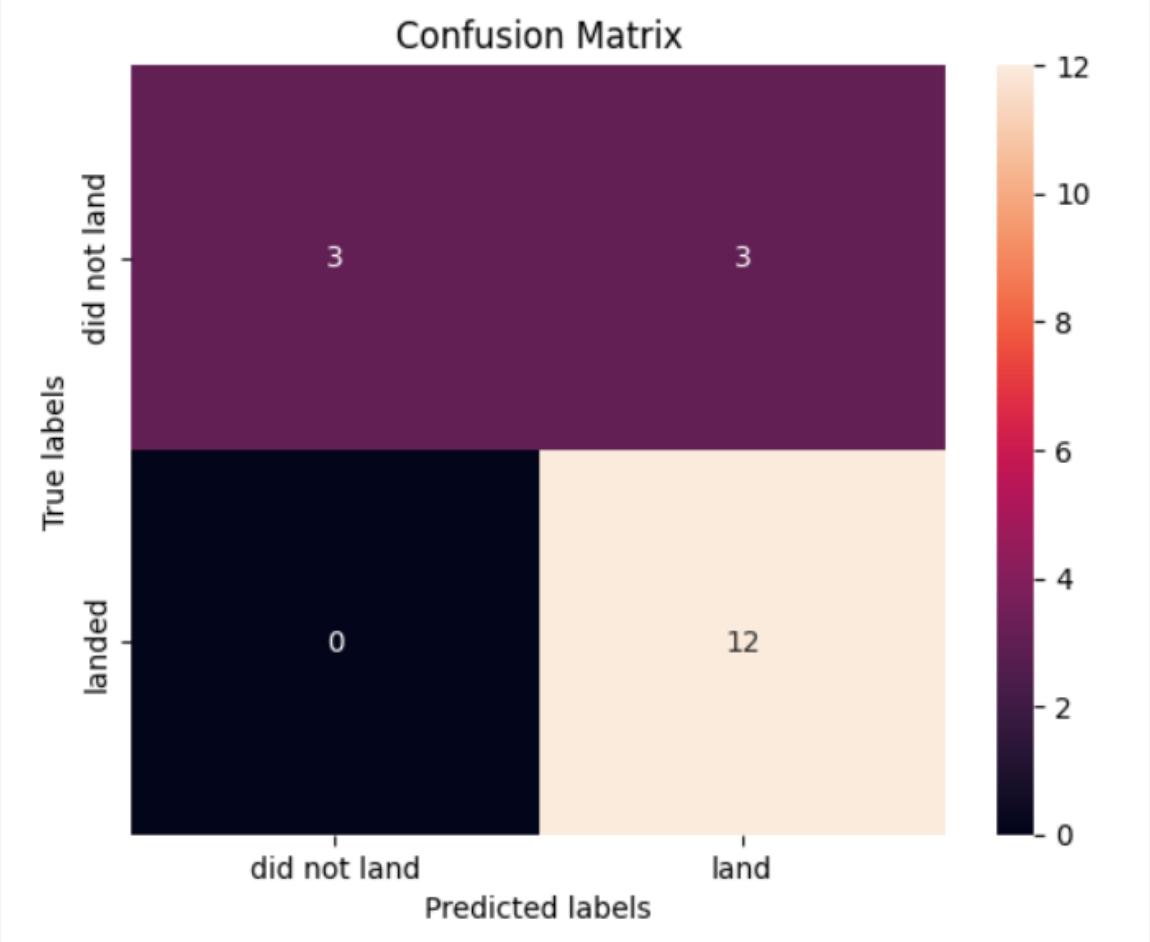
- F1 score: 0.888889
- Accuracy score: 0.833333
- GridSearchCV best score: 0.846429



Confusion Matrix – SVM

Support Vector Machine

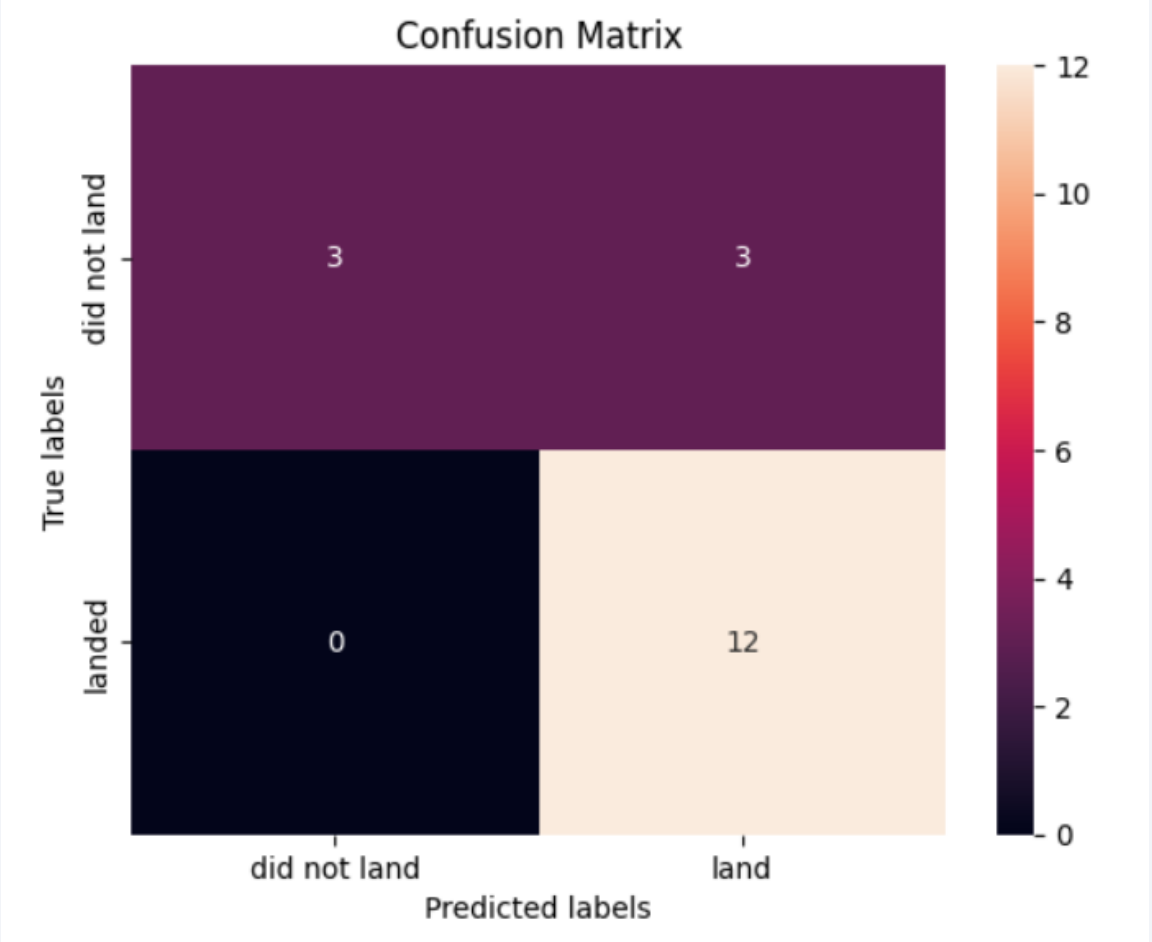
- F1 score: 0.888889
- Accuracy score: 0.833333
- GridSearchCV best score: 0.848214



Confusion Matrix – Decision Tree

Decision Tree

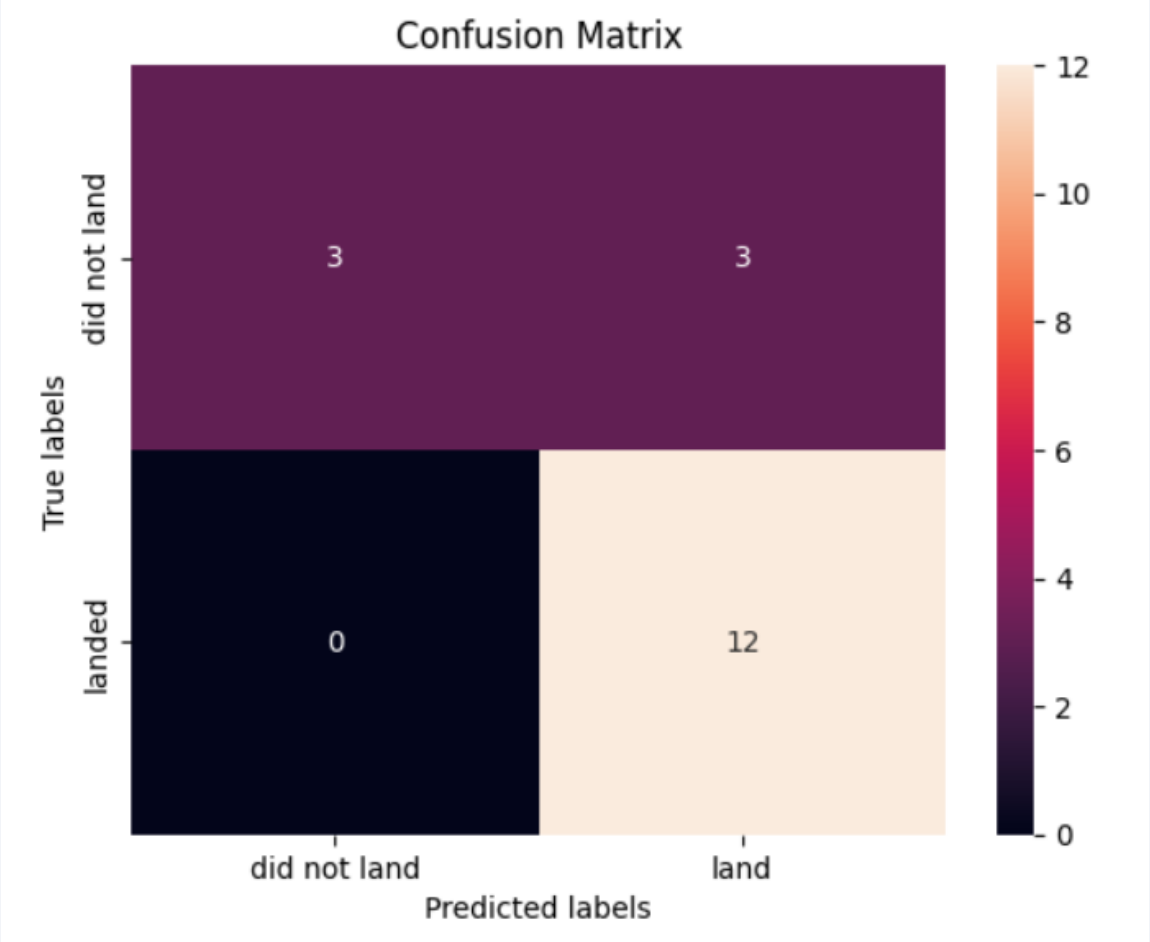
- F1 score: 0.888889
- Accuracy score: 0.833333
- GridSearchCV best score: 0.873214



Confusion Matrix – KNN

K-Nearest Neighbors

- F1 score: 0.888889
- Accuracy score: 0.833333
- GridSearchCV best score: 0.848214



Model Evaluation

Comparing the four models it is clear that they all have the same accuracy score and confusion matrix when tested on the set. Because of this, additional scores where necessary when evaluating which model performs best. Specifically, the F1 score and the respective GridSearchCV best scores.

	LogReg	SVM	Tree	KNN
F1 Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333
Best Score	0.846429	0.848214	0.873214	0.848214

Considering those values, it seems that the Decision Tree Model is the best algorithm for this dataset.

Conclusions

- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increased over the years.
- KSC LC-39A lauch site has the highest success rate of the launches from all the sites.
- Orbits types ES-L1, GEO, HEO and SSO have 100% success rate.
- The Decision Tree Model is the best machine learning algorithm for this dataset.

Thank you!