IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ka Wai Lawrence Wong
14th August 2024

Outline

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

Executive Summary

To predict the true cost of a launch it is necessary to accurately predict the successful landing of booster rockets. To achieve this data was collected from SpaceX's API and scrapped from wiki to study the historic performance of all Falcon 9 launches.

After preparing the data, exploratory data analysis was performed and a dashboard was made. It was found it took SpaceX 5 years to achieve a successful booster landing, the Falcon 9 carries the heaviest payloads, the most successful launch site is in Florida at KSC LC-39A with a success rate of 76.9%.

Based on the predictive analysis, the data used to train the Logistic regression model showed it can accurately predict landing outcomes with 83% accuracy. The model predicts a positive landing outcome 83% of the time and of all positive landing outcomes, 20% will be false positives.

Introduction

This project focuses on predicting the successful landing of the Falcon 9 rocket's first stage. SpaceX offers Falcon 9 launches at a competitive price of $62 million, significantly lower than the $165 million charged by other providers, primarily due to their ability to reuse the first stage. By predicting whether the first stage will land successfully, the true cost of a launch can be estimated. This analysis provides valuable insights for any company looking to compete with SpaceX in the rocket launch market. The project involves gathering and formatting data from an API to support this analysis.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology: requested JSON content from SpaceX API using .get. Scrapped data from Wikipedia
- Perform data wrangling: performed study on summary statistics of the data, prepared and exported data for further 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: employed GridSearchCV to test 4 models: Logistic regression, SVM, Tree and K-nearest neighbour. Log. Regression yields best result for substantially lower computation cost (time).

Data Collection

Data were collected from two sources using two different methods.

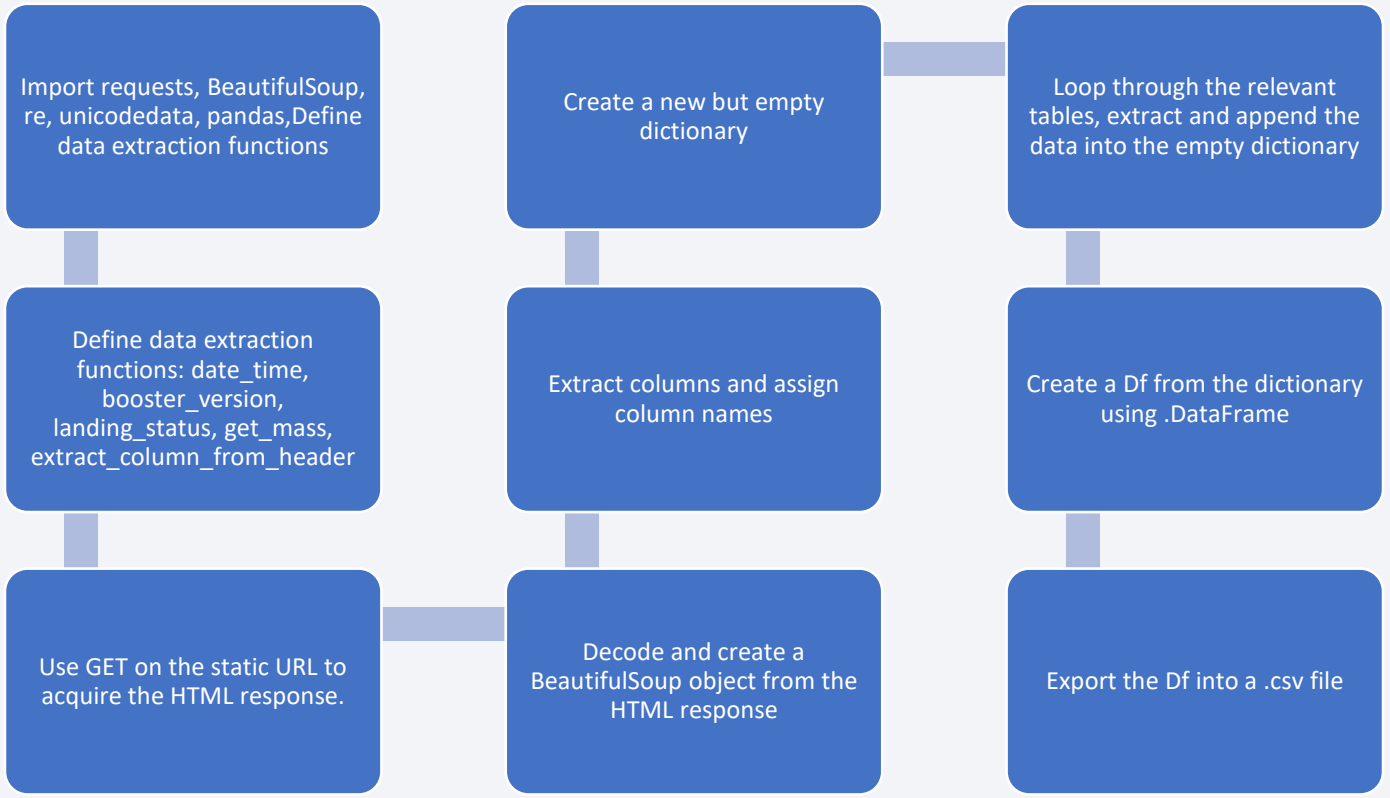
1- Some SpaceX Falcon 9 data were collected using GET request to fetch JSON data from a static URL. Then relevant data frames were created. They were then filtered and data relevant to the Falcon9 extracted.

2- Other data related to the Falcon 9 were collected using web scraping. GET request was used to fetch HTML data that then were read using BeautifulSoup. A data frame was created by parsing the HTML tables and Falcon 9 data was extracted.

Data Collection – SpaceX API

GITHUB URL: https://github.com/Razc90/submissions/blob/main/Capstone/1_jupyter-labs-spacex-data-collection-api.ipynb

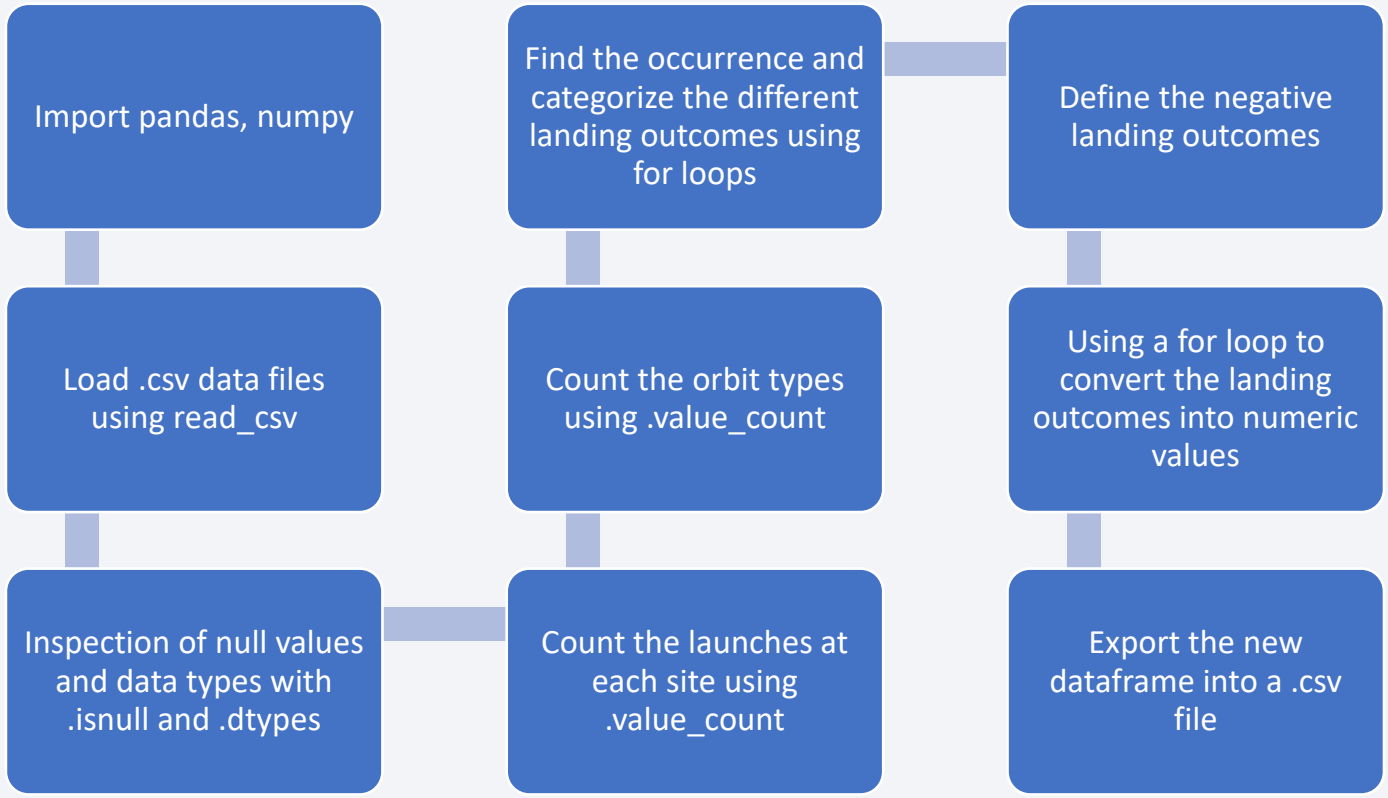
Data Collection - Scraping



GITHUB URL: https://github.com/Razc90/submissions/blob/main/Capstone/1_jupyter-labs-webscraping.ipynb

Data Wrangling

The data wrangling process performed focused on study of summary statistics and further preparation of data through transforming features into formats suitable for later use.



GITHUB URL: https://github.com/Razc90/submissions/blob/main/Capstone/1_labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Several plots were produced during this process:

1. Catplot- Payload mass vs flight number (hue = Class: landing outcome): shows relationship between ayload and landing outcomes as more launches were performed.
2. Catplot- Launch site vs flight number (hue = Class: landing outcome): shows relationship between launch site usage and landing outcomes over time as more launches were performed.
3. Catplot- Launch site vs Payload mass (hue = Class: landing outcome): shows relationship between launch site and payload and landing outcomes
4. Bar graph- Orbit type vs Landing success rate: shows the relationship between the type of orbit the craft goes into and the rate at which the booster makes a successful landing
5. Catplot- Orbit type vs flight number (hue = Class: landing outcome): shows relationship between the type of orbit the craft gets into and landing outcomes over time as more launches were performed. It also shows the orbit requirements change over time.
6. Catplot- Orbit type vs payload mass (hue = Class: landing outcome): shows relationship between the type of orbit the craft gets into and payload mass.
7. Line plot- Annual average landing success rate from 2010 to 2020: shows the evolution of the landing success rate over time.

GITHUB URL: https://github.com/Razc90/submissions/blob/main/Capstone/2_jupyter-labs-eda-dataviz.ipynb

EDA with SQL

SQL queries made:

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was achieved.
6. List the names of the boosters which have success in drone ship landings and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failed mission outcomes
8. List the names of the booster versions which have carried the maximum payload mass via using a subquery
9. List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
10. Rank 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.

GITHUB URL: https://github.com/Razc90/submissions/blob/main/Capstone/2_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

The follow objects were added to the Folium map:

- Circle: used to highlight an area on the map
- Marker: used to show a specific coordinate on the map
- Cluster: used to simplify a map containing many markers having the same coordinate.
- Mouse position: used to display longitude and latitude data of the location the computer mouse is point at.
- PolyLine: used to display a line between two coordinates. Useful at showing distance between two points.

GITHUB URL:
https://github.com/Razc90/submissions/blob/main/Capstone/spacex_dash_app.py

Build a Dashboard with Plotly Dash

The dashboard contain the following interactive features:

- Launch site drop-down selection: allows the user to select whether they want to see data of all the site at once or individual sites
- Payload range slider: allows the user to filter which payload range they want displayed in their Landing outcome (class) vs Payload mass scatter plot.

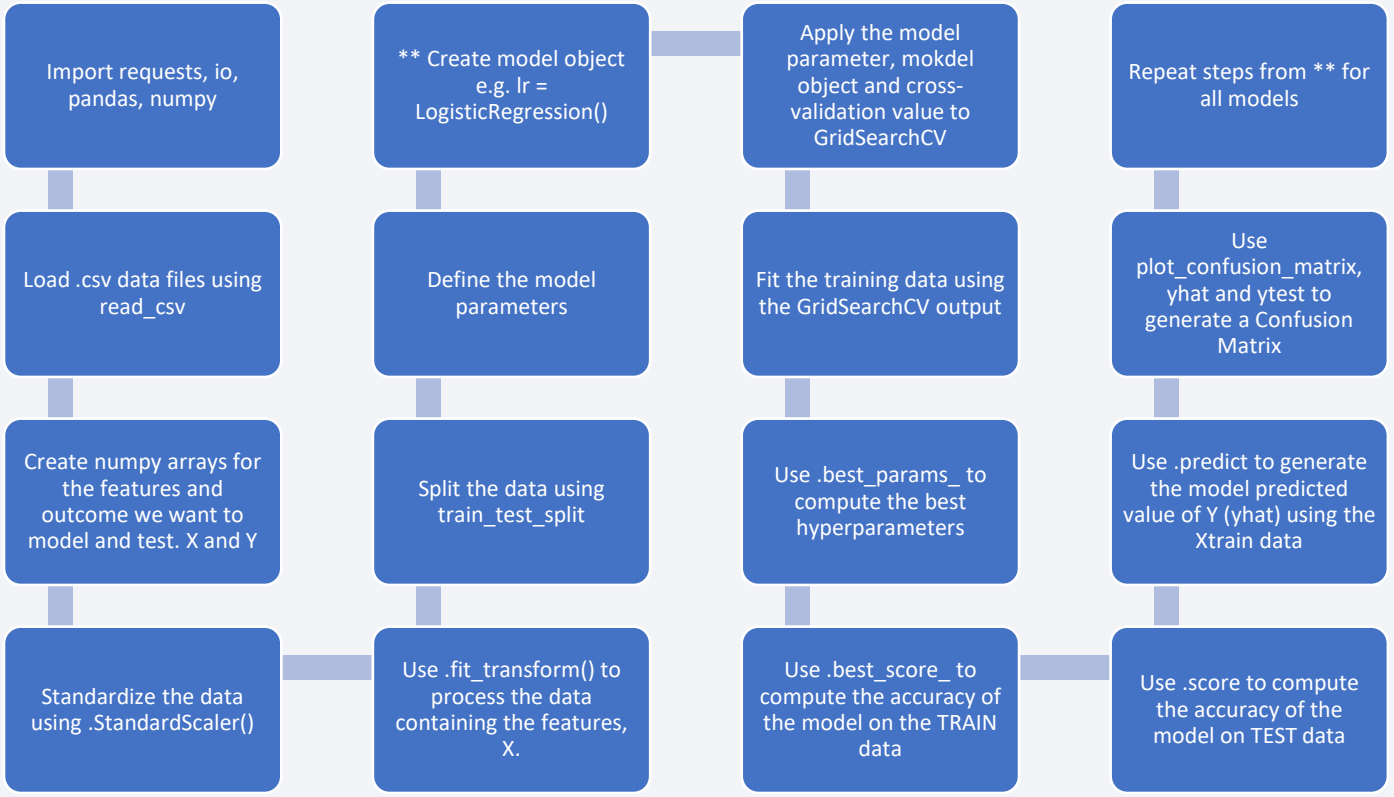
The dashboard contain the following plots:

- Pie chart (all): displays the proportion of successful outcomes by launch site
- Pie chart (individual): displays the proportion of successful and failure outcomes by each launch site
- Scatterplot Displays all the Landing outcome (Class) vs Payload mass categorized by the mission launch site. The x-range adjusted via the Payload range slider.

GITHUB URL:
https://github.com/Razc90/submissions/blob/main/Capstone/4_SpaceX_Machine_Learning_Prediction_Part_5.ipynb

Predictive Analysis (Classification)

The predictive models were built using processes. The data was first split into training and test sets. These data sets were then passed through GridSearchCV until its accuracy and confusion matrix is evaluated. This process was repeated for different models (logistics regression, SVM, tree, K-nearest neighbour). Performance of the model is evaluated based on a combination of its train and test accuracy scores as its confusion matrix values.

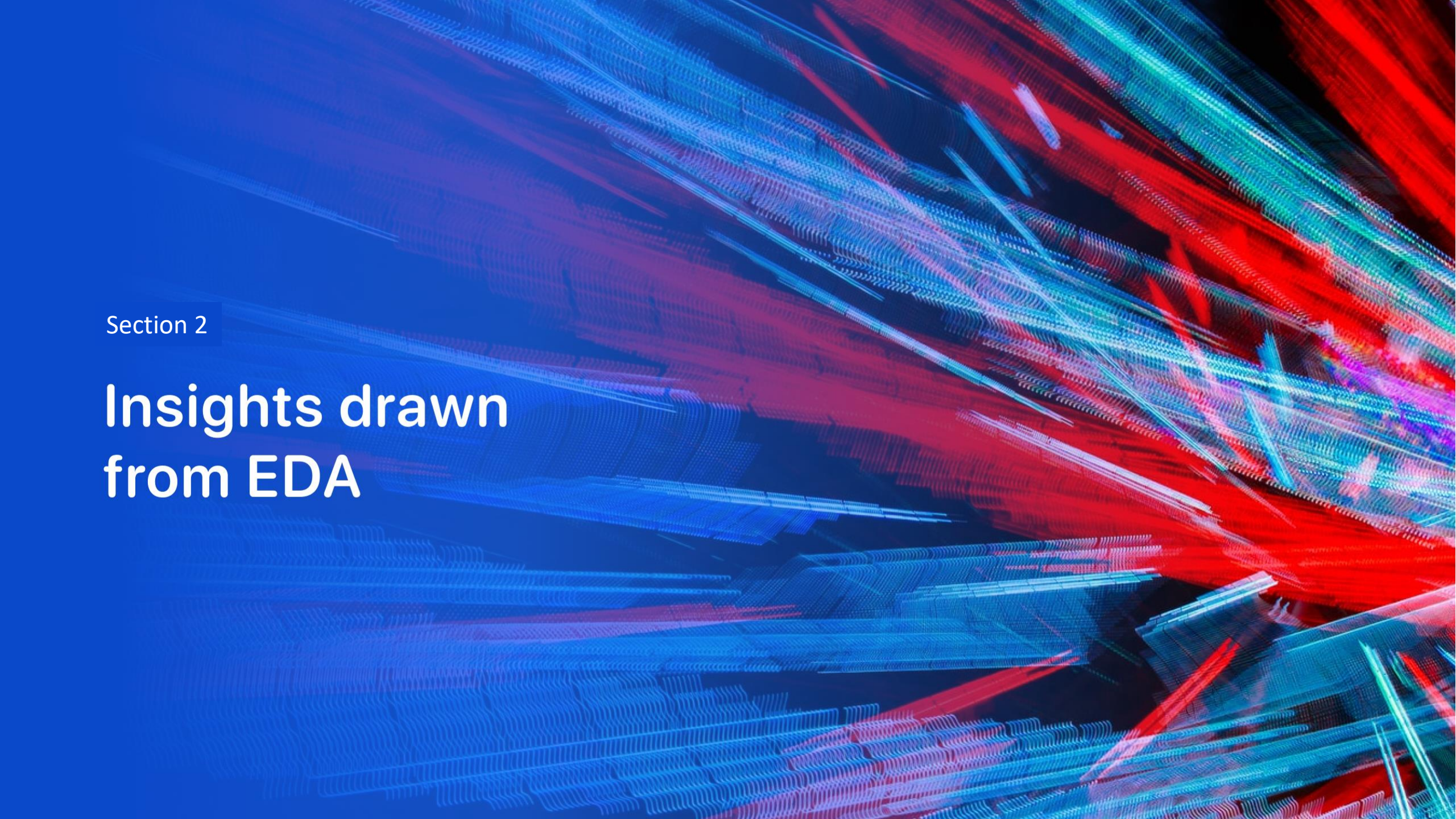


GITHUB URL:

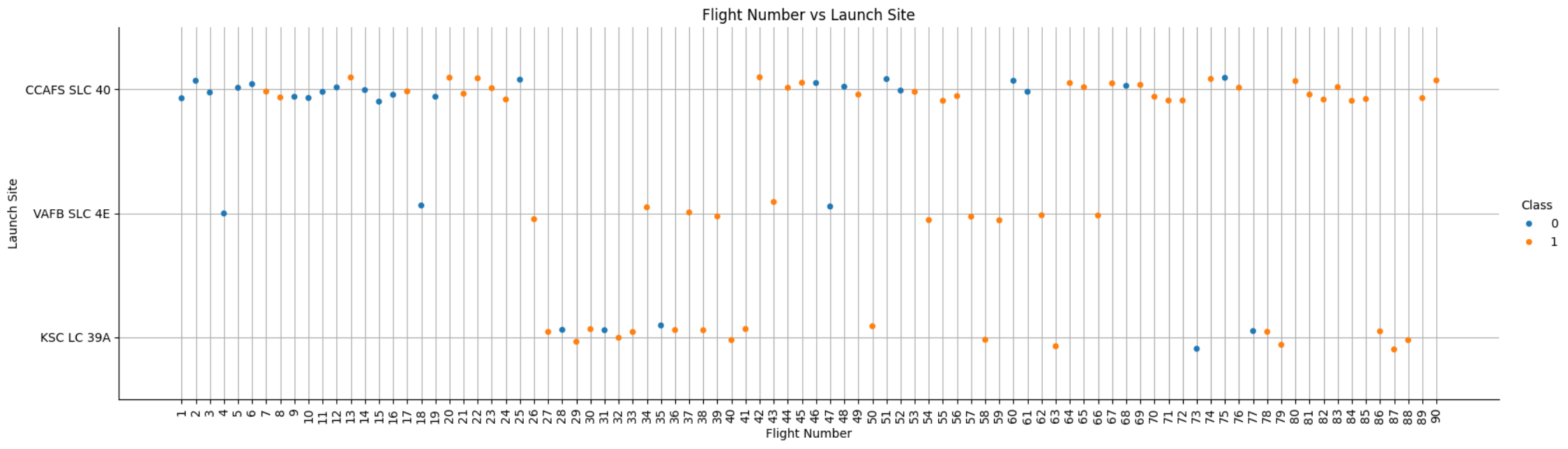
Results

The results of the project are show cased in the following sections:

- Section 2: Insights from exploratory data analysis
- Section 3: Launch sites proximities analysis
- Section 4: Interactive analytics demo in screenshots
- Section 5: Predictive analysis results

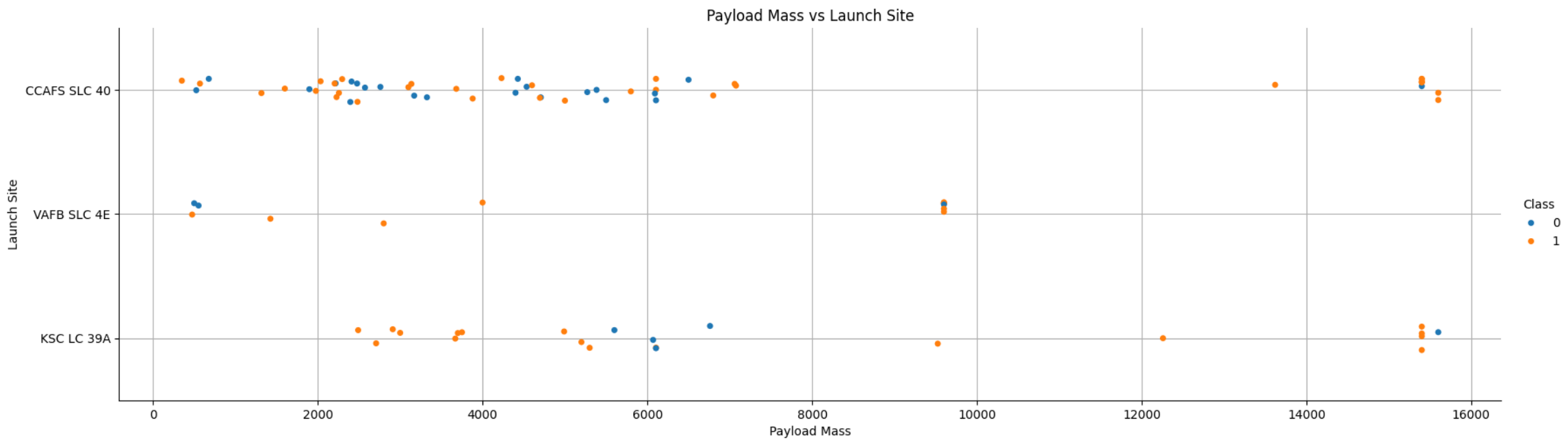

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



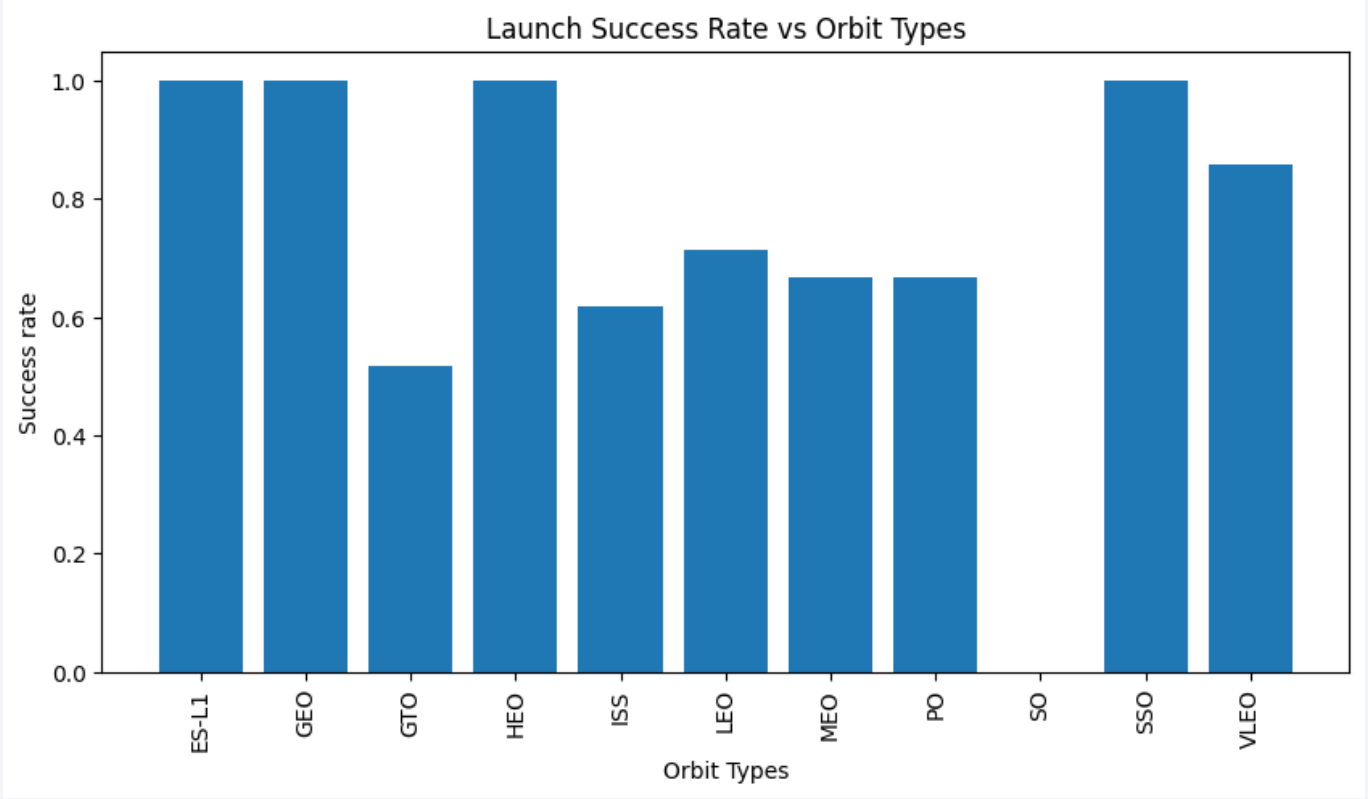
This plot shows Launch Site usage with Flight Number and its respective Landing Outcome (Class). It is evident most launch were at CCAFS except for a period (#25 – 42) where launches took place consistently at KSC. Landing failures are also higher during earlier launches as the rockets are developed.

Payload vs. Launch Site



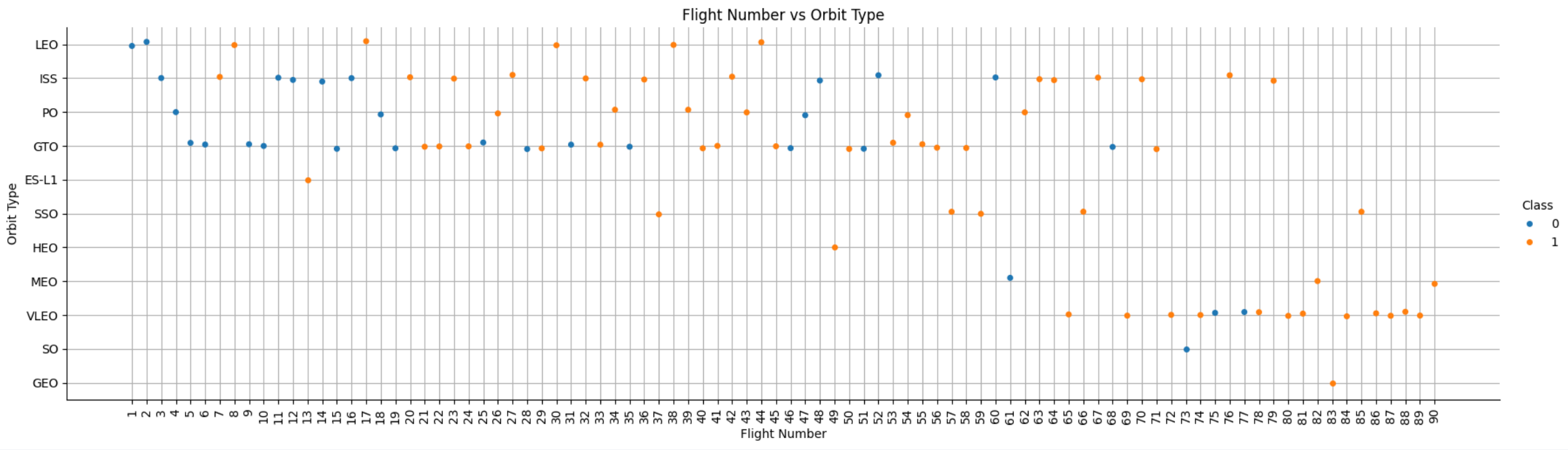
This plot shows the relationship between Launch Sites and Landing Outcomes(Class) with respect to Payload Mass. It is Observed CCAFS and KSC hosts launches of a spectrum of payload mass from sub-2000kg to over 14000kg. However, VAFB only manages launches with payload up to around 9500kg. Landing success rate is also comparatively higher with high payloads. It is possible that this is likely due to those being later launches as earlier launches would test with smaller payloads.

Success Rate vs. Orbit Type



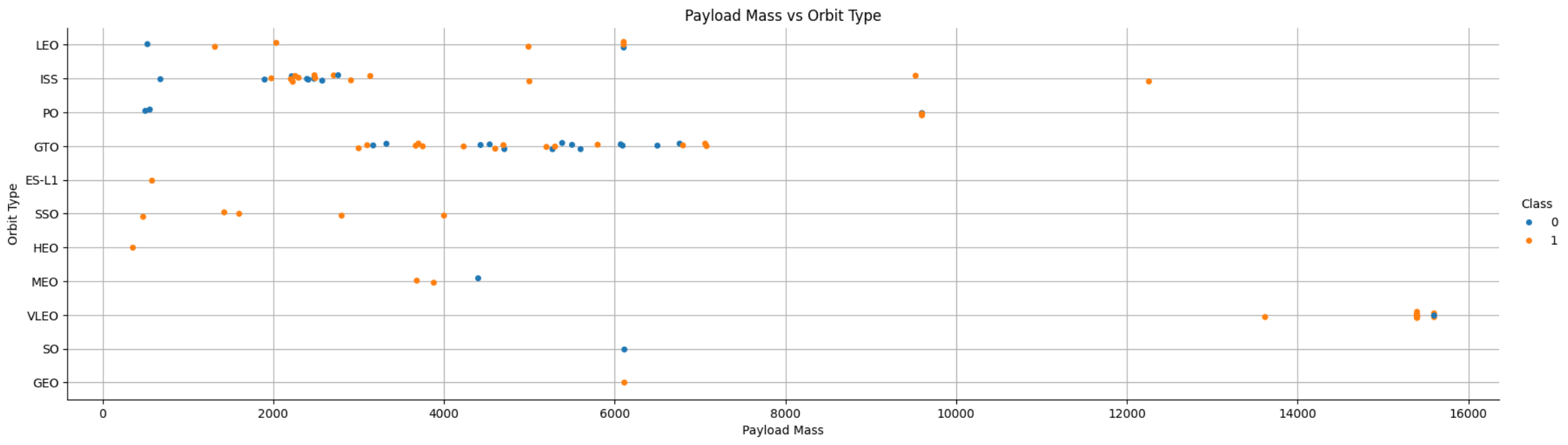
This bar chart shorts the Landing success rate for each mission orbit type. It can be seen ES-L1, GEO, HEO, SSO and VLEO all have very high success rates. To further verify any relationship between orbit and landing success rate it is important to examine the Booster Version as well as the Flight Number as these can have significant influence on mission success.

Flight Number vs. Orbit Type



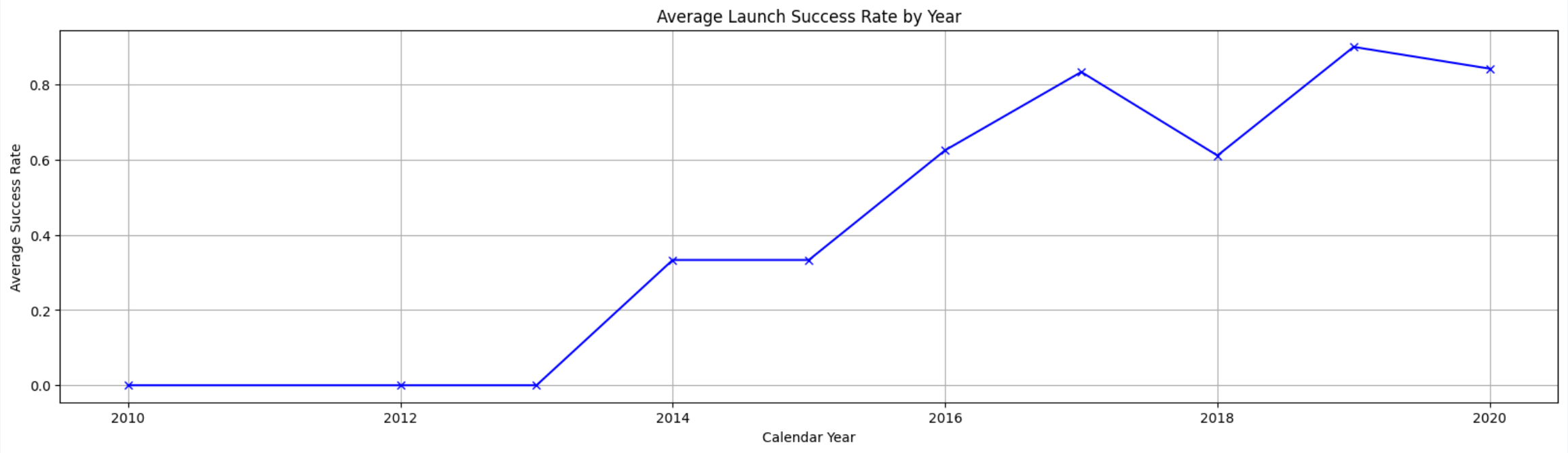
This plot shows the relationship between Orbit Type, Landing Outcome(Class) and Flight Number. It is evident GTO and ISS has the highest number of launches. The orbit of many later launches involve VLEO instead of GTO orbit.

Payload vs. Orbit Type



This plot shows the relationship between the Orbit Type, Landing Outcome (Class) with Payload Mass. Most launches to the ISS involves a payload of around 2500kg while payload in launches to the GTO orbit ranges from 2500 to 7000kg. The heaviest payload are those launched to VLEO orbits.

Launch Success Yearly Trend



This line plot shows the annual average successful landing rate between 2010 and 2020. The success rate increased from 2013 to 2017. The success rate peaked at 2019.

All Launch Site Names

This query returns a result showing the 4 launch sites used by SpaceX.

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

```
%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

This query returns data where the craft was launched at the CCAFS LC-40 site

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

This query returns the result of the cumulated total payload mass of 45596kg for the SpaceX customer, NASA.

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_version like 'F9 v1.1%'
```

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

This query returns the result showing the average payload mass of 2534.67kg carried by booster rocket version F9 V1.1

First Successful Ground Landing Date

```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
```

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

min(Date)
2015-12-22

This query returns the result of the date of the first successful ground pad landing, which was 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select distinct(Booster_Version) from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_<6000
```

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

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

This query returns the different booster rocket versions that carried a payload mass between 4000kg and 6000kg and have successfully landed on a drone ship.

Booster Version:

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

```
%sql select count(case when Mission_Outcome like 'Suc%' then 1 end) as "Successful",count(case when Mission_Outcome like 'Fail%' then 1 end) as "Failed"
from SPACEXTABLE;
```

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

Successful	Failed
100	1

This query returns the number of successful and failed mission outcomes. Note, this is not the same as landing outcomes.

Successful	Failed
100	1

Boosters Carried Maximum Payload

This query returns a list of booster rocket versions that have carried the highest payload mass to date.

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

```
%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

2015 Launch Records

```
%sql select substr(Date, 6,2) as "Month", Booster_Version
from SPACEXTABLE where Landing_Outcome = 'Failure (drone ship)'and substr(Date,0,5)='2015'
```

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

Month	Booster_Version
01	F9 v1.1 B1012
04	F9 v1.1 B1015

This query retrieves a table showing booster rocket versions and the month of launches in 2015 where the rocket failed to land on a drone ship.

Month	Booster_Version
01	F9 v1.1 B1012
04	F9 v1.1 B1015

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

The query returns a table ranking the landing outcomes and each outcome's occurrence. Excluding No attempts, successful drone ship landings has the highest number of occurrences

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
- Precluded (drone ship) 1
- Failure (parachute) 1

```
%sql select Landing_Outcome, count(*) as outcome_count
from SPACEXTABLE where Date>'2010-06-04' and Date<'2017-03-20' group by Landing_Outcome order by outcome_count desc
```

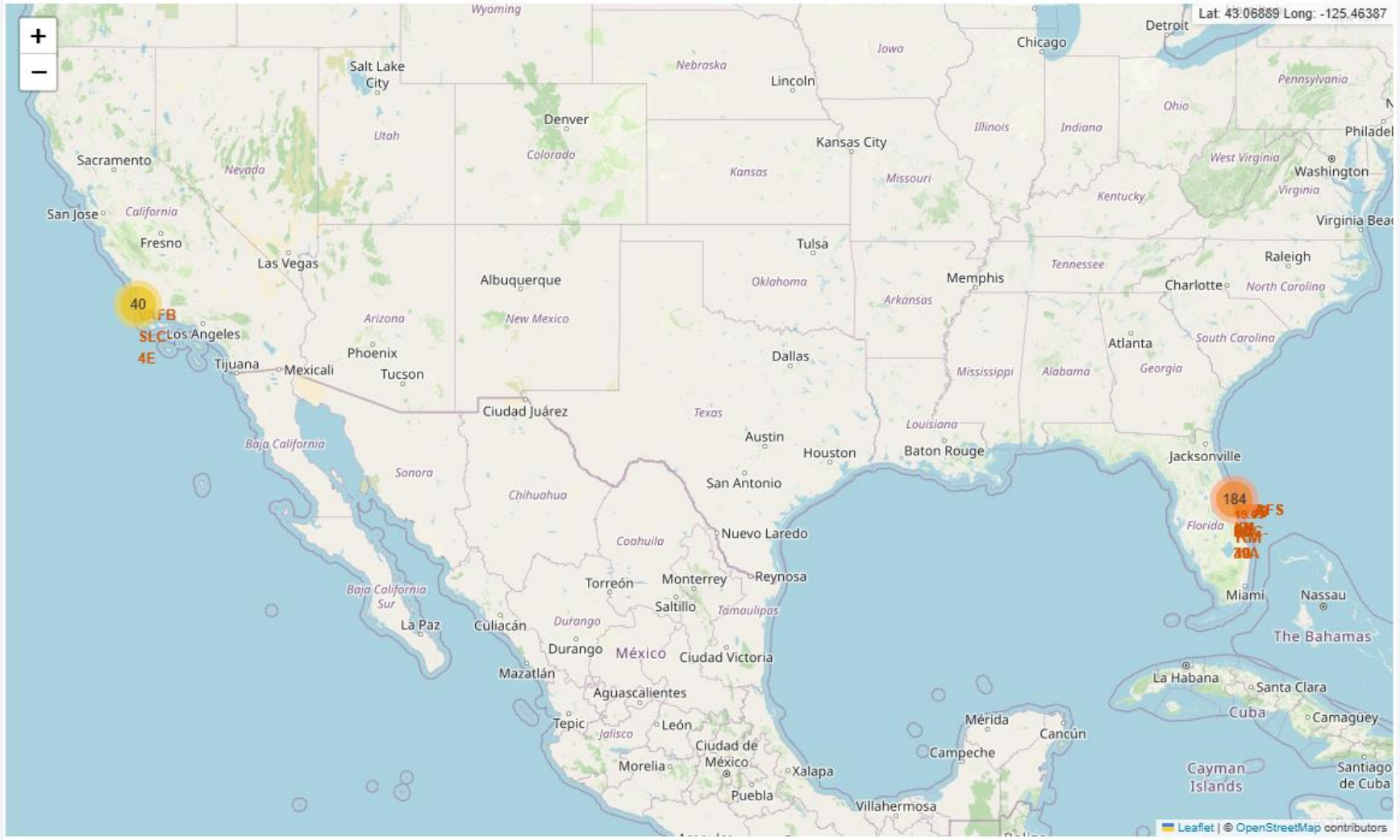
```
 * sqlite:///my_data1.db
Done.
```

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
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

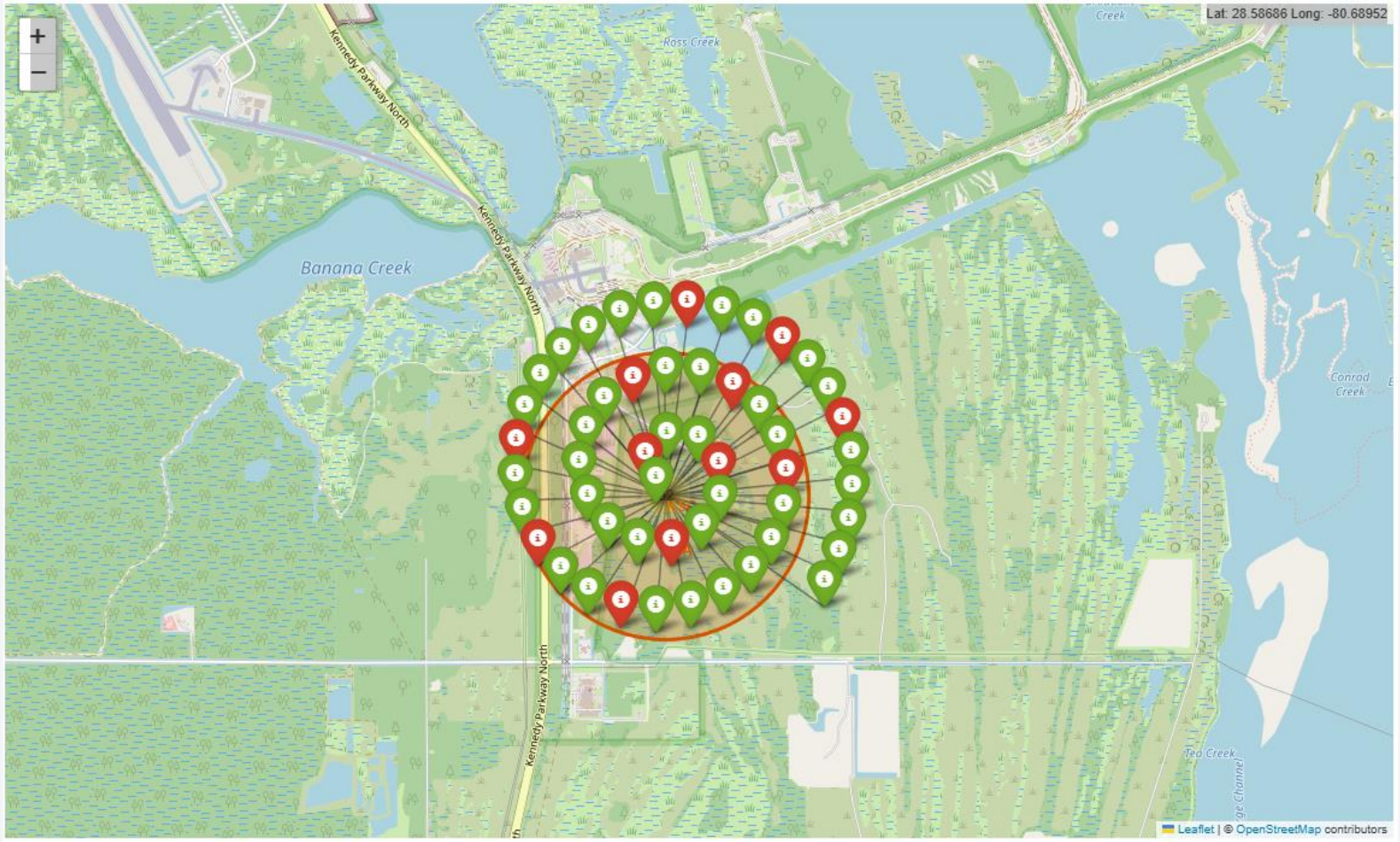
Launch Sites Proximities Analysis

SpaceX Launch Site Locations



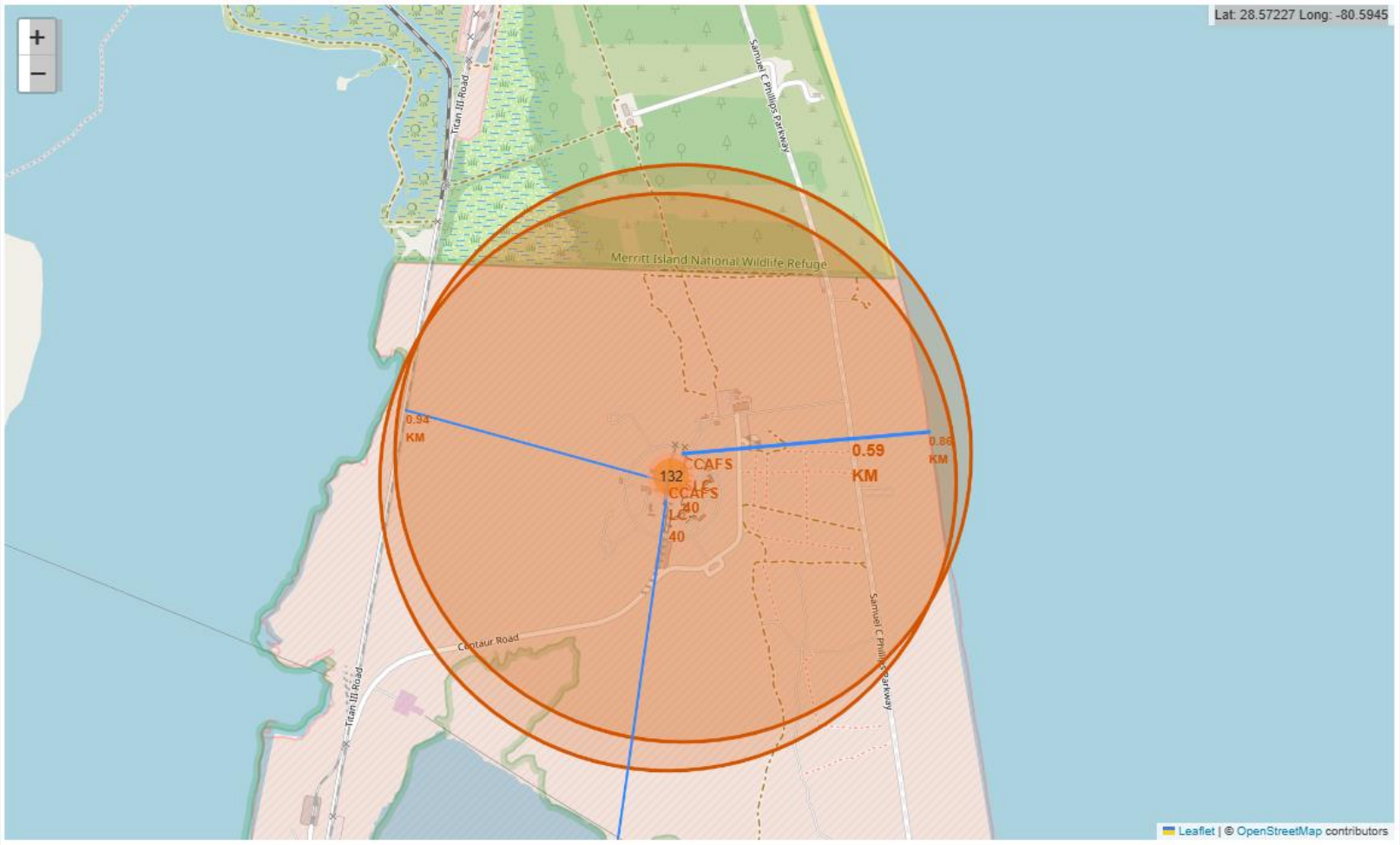
This is a regional map showing all the launch sites used by SpaceX. The coloured group shows the number of launches took place at the launch sites in that geographical area. It is evident 184 launches took place in Florida while 40 launches took place in California

Launches at Site RSC LC-39A

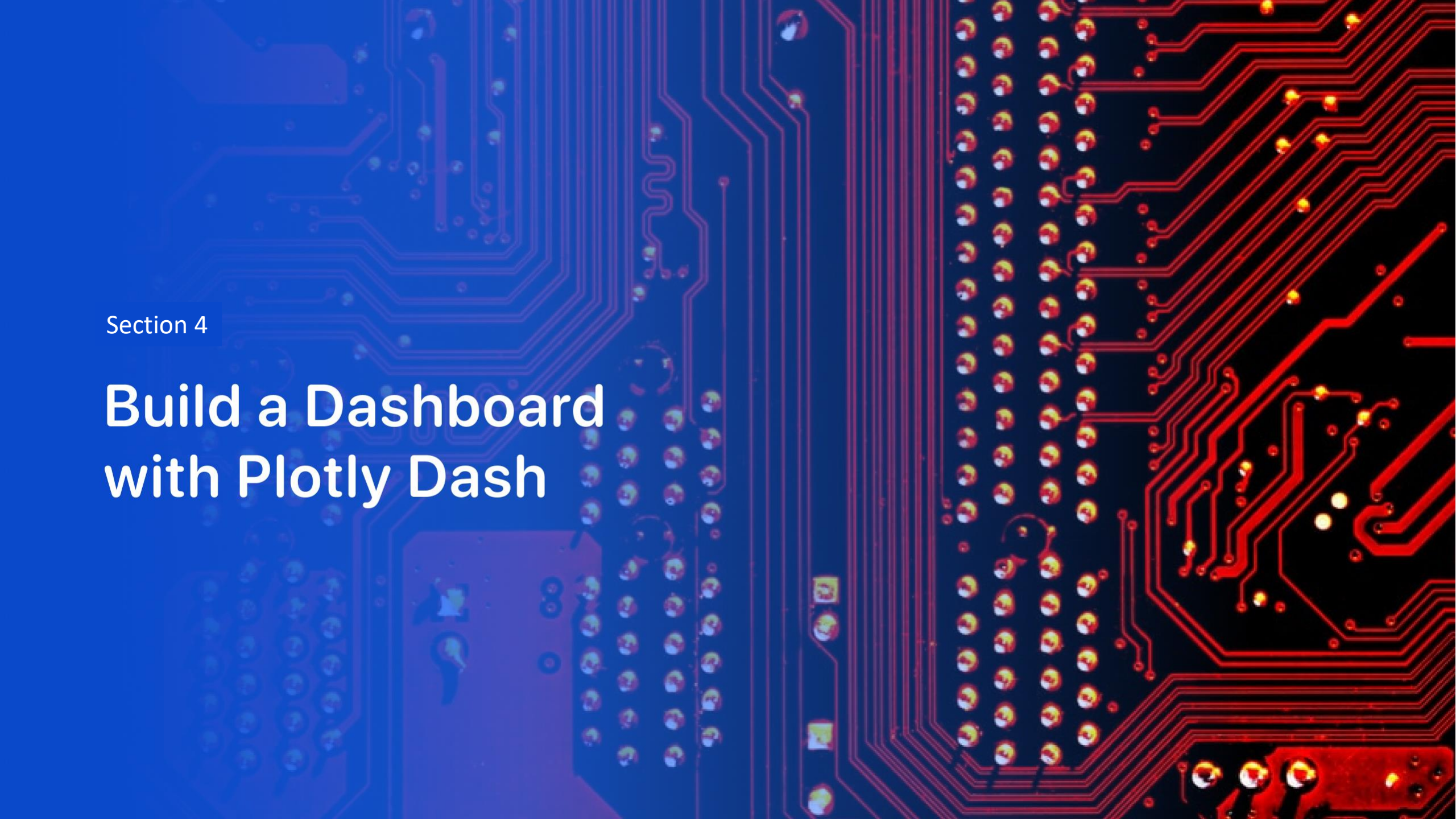


The marker cluster shows the individual launches and landing outcome at RSC LC-39A. Through the marker colour, green shows a successful landings and red for failed ones.

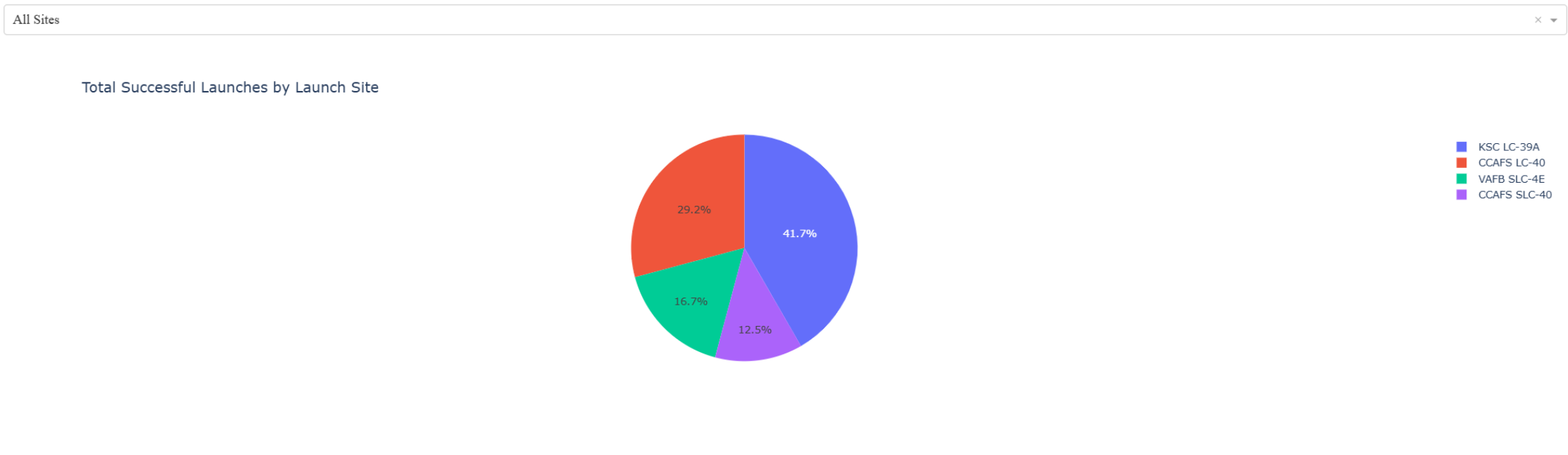
CCAFS Sites Proximity to Key Geographical Structures



This map shows the CCAFS launch sites' approximate distance to key geographical features nearby. Approximately: 0.59km to the nearest highway, 0.86km to the nearest coastline and 0.94km to the nearest railroad. It is not shown on the map here but it is also 19.59km to the nearest city, which is Cape Canaveral.

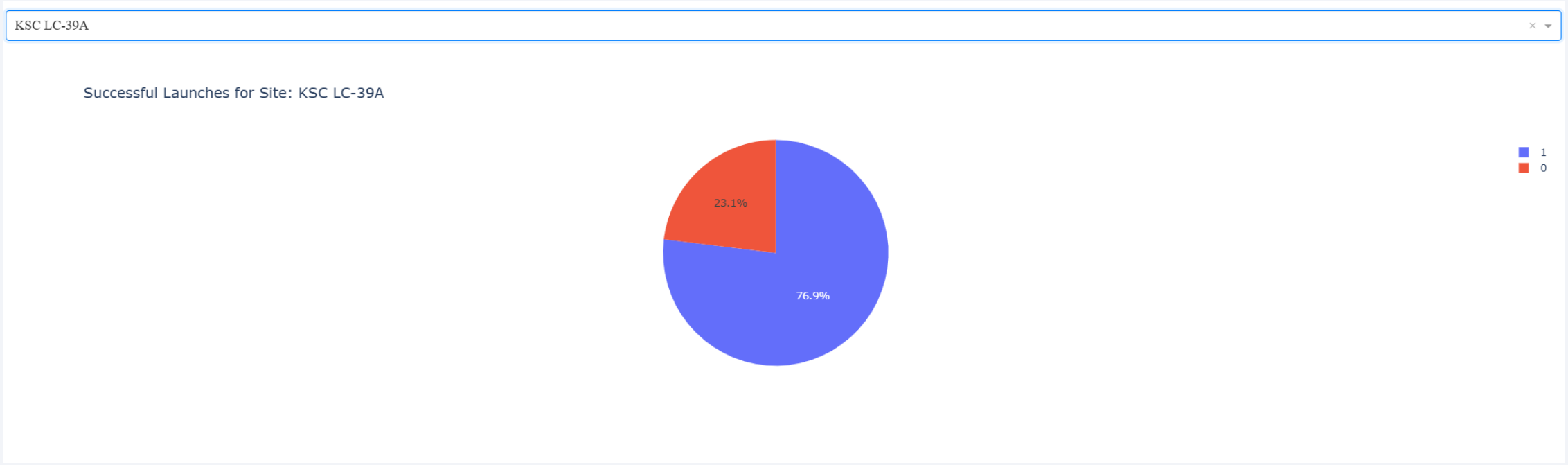

Section 4
Build a Dashboard with Plotly Dash

Total Successful Launches by Site



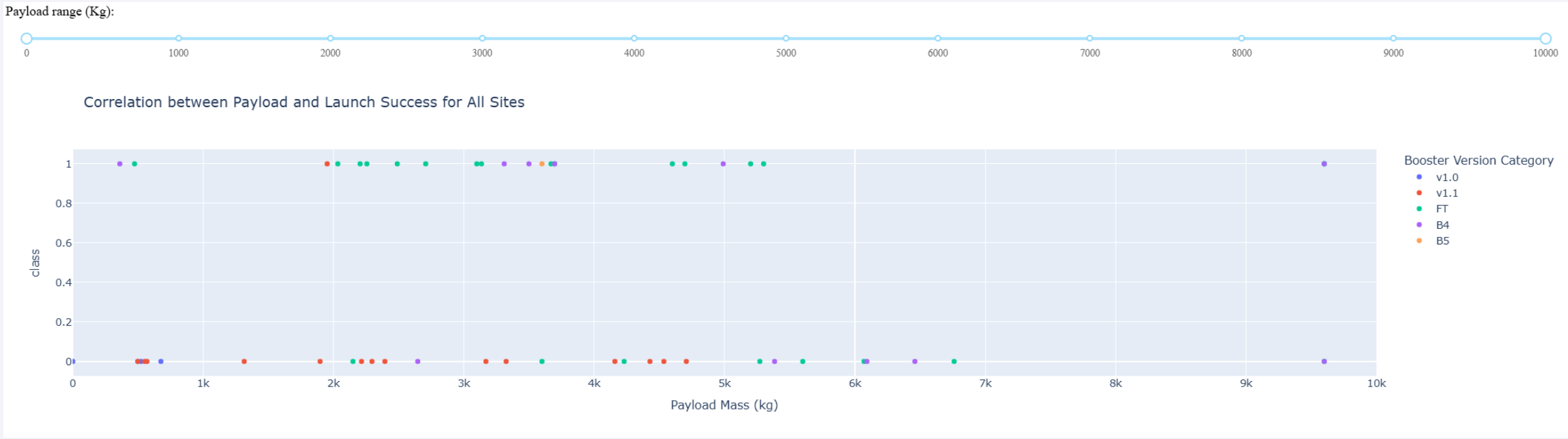
This pie chart illustrates the proportion of total successful launches by each launch site. It can be observed that KSC LC-39A has the highest percentage of successful launches while CCAFS SLC-40 has the least

Launch Success Rate at KSC LC-39A



This pie chart illustrates the launch success rate of site KSC LC-39A. It has the highest success rate of all SpaceX launch site at 76.9% success rate.

Relationship Between Launch Success and Payload for Various Boosters



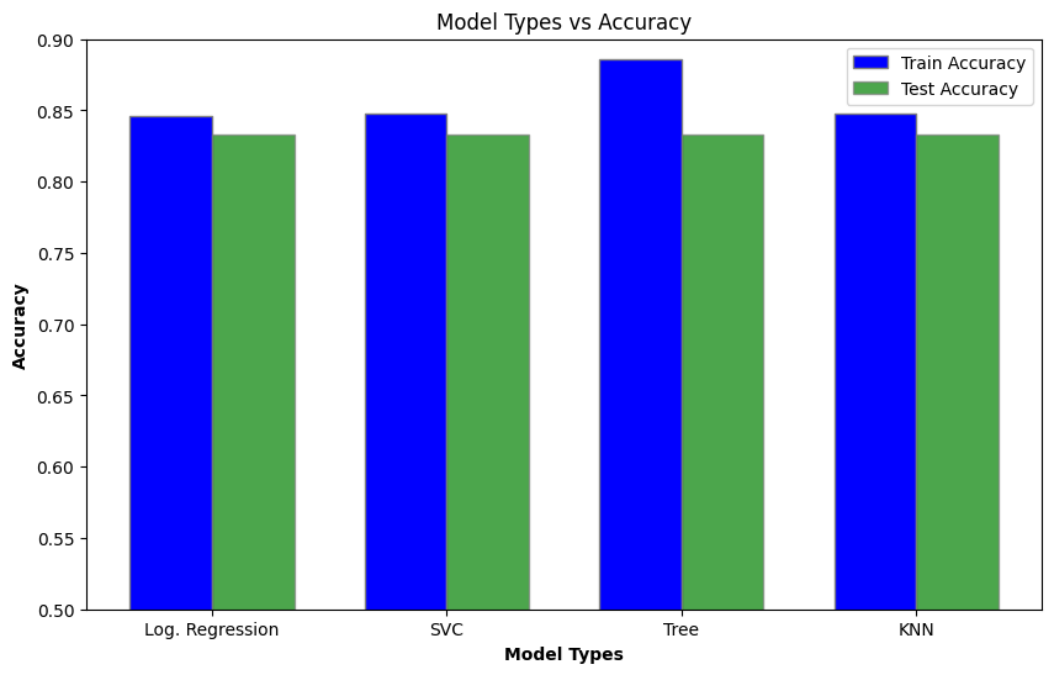
This scatter plot illustrates the relationship between the launch success, payload and the version of booster rocket used for the launch. From this plot, it can be seen booster version V1.0 and V1.1 failed their launches. This is inline with other data showing SpaceX did not yield a successful launch until after 2013 and these are earlier booster versions. FT is successful at carrying payloads up to 5000kg. The B4 booster yields mixed results as it demonstrates both success and failed launches across the payload range. The single B5 launch successfully carried approximately 3500kgs. While this plot do show some correlation between launch success and payload for various boosters, it is essential to examine how the mission orbit also affects the launch success.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

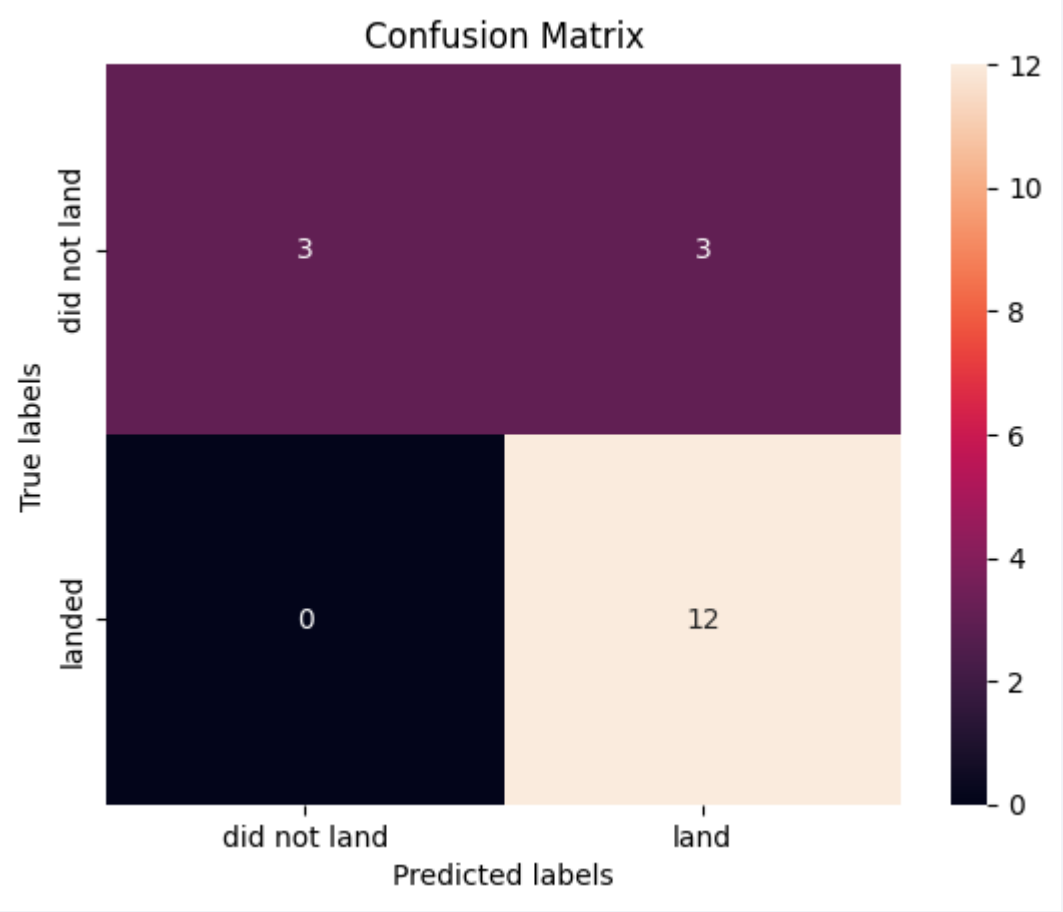
This bar graph shows the train and test accuracy of the 4 models tested. All 4 models show identical test accuracies and near-identical training accuracies. Statistically, Tree model has the highest training accuracy, however, training accuracy is not as significant as testing accuracy and the Tree model take significantly longer to process compare to Logistic regression. The best performing model after weighting up the computational cost and accuracy benefits is Logistic regression as it completed the computation in 0.068s and is close to 30 times faster than its nearest model, SVC, at 1.72s.



	label	train_acc	test_acc	test_tm
0	Log. Regression	0.846429	0.833333	0.067908
1	SVC	0.848214	0.833333	1.723315
2	Tree	0.885714	0.833333	5.617139
3	KNN	0.848214	0.833333	2.597537

Confusion Matrix

This Logistic regression confusion matrix shows the model can predict mission outcomes with good accuracy. 83% of predictions will be positive. With 3 false positive predictions, this translates to a 20% false positive outcome of all predicted positive outcomes.



Conclusions

- Based on SpaceX's record, it will likely take 5 years to achieve successful booster landings
- The Falcon9 carries the heaviest payloads for SpaceX
- Most SpaceX launches take place in Florida and the KSC LC-39A site has the highest launch success rate at 76.9%
- Logistic regression is the best performing model and is able to predict landing outcomes with 83% accuracy using test-data sets.
- 83% of all landing predictions will output a successful landing with 20% of these predictions being a false positive.

Appendix

Log. Reg.:

	precision	recall	f1-score	support
0	1.00	0.50	0.67	6
1	0.80	1.00	0.89	12
accuracy			0.83	18
macro avg	0.90	0.75	0.78	18
weighted avg	0.87	0.83	0.81	18

SVM:

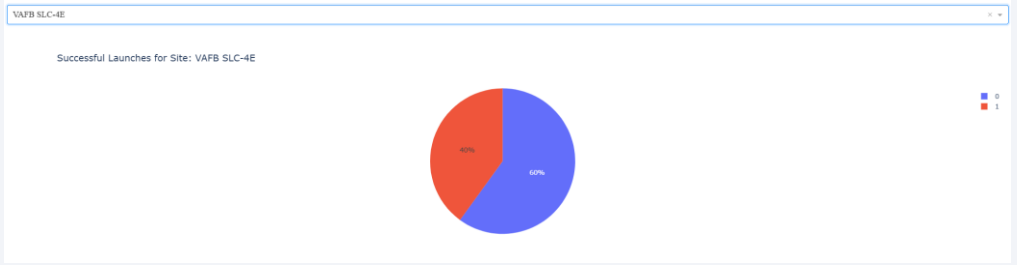
	precision	recall	f1-score	support
0	1.00	0.50	0.67	6
1	0.80	1.00	0.89	12
accuracy			0.83	18
macro avg	0.90	0.75	0.78	18
weighted avg	0.87	0.83	0.81	18

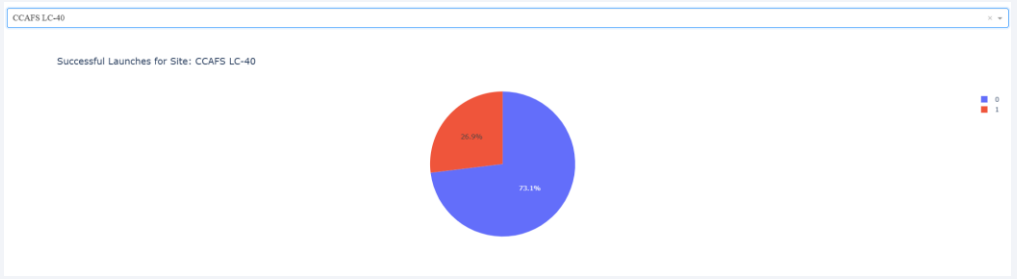
Tree:

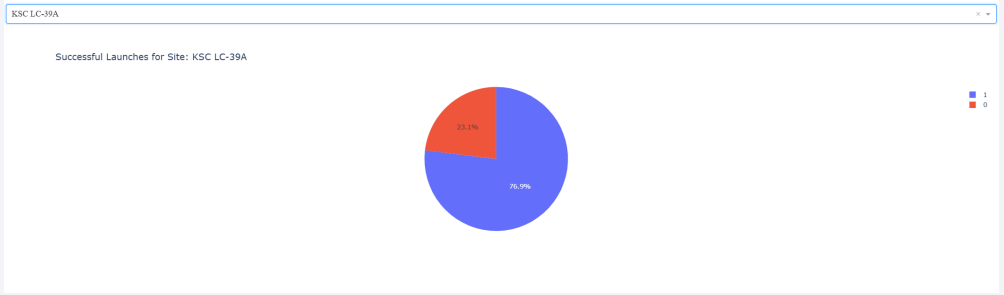
	precision	recall	f1-score	support
0	1.00	0.50	0.67	6
1	0.80	1.00	0.89	12
accuracy			0.83	18
macro avg	0.90	0.75	0.78	18
weighted avg	0.87	0.83	0.81	18

KNN:

	precision	recall	f1-score	support
0	1.00	0.50	0.67	6
1	0.80	1.00	0.89	12
accuracy			0.83	18
macro avg	0.90	0.75	0.78	18
weighted avg	0.87	0.83	0.81	18







Thank you!