IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Aldana B. Moroni
February 25th, 2024

Outline

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

Executive Summary

- Summary of methodologies
    1. Data Collection with API & Webscrapping
    2. Exploratory Data Analysis with SQL & Data Visualization
    3. Interactive Visual Analytics & Dashboard
    4. Predictive Analysis
- Summary of all results
    1. Exploratory Data Analysis Results
    2. Interactive Folium Map & Dashboard Application
    3. Predictive Results

Introduction

Commercial space age is here!

Companies are working hard to make space travel affordable for…

EVERYONE!

***SpaceX* is perhaps the most successful nowadays, and one reason is that they have made rocket launches relatively inexpensive.**

***SpaceX* advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars, while 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 this first stage will land, we can determine the cost of a launch.

Introduction

In order for *SpaceY* to compete, there are certain questions that must be answered:

- What are the main variables that determine the success or failure in landing?
- Are there any relationships between variables?

Section 1

Methodology

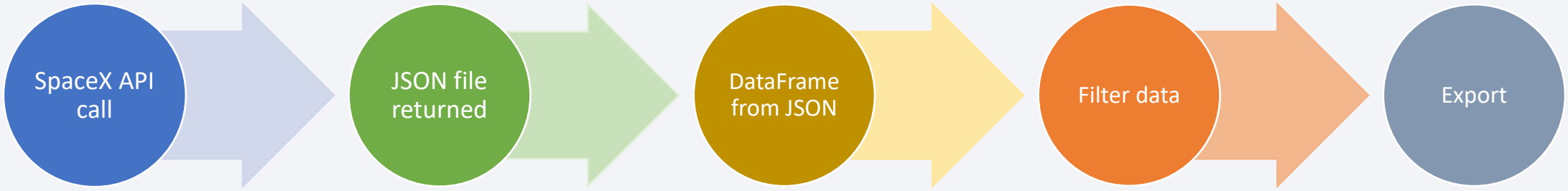
Methodology

Executive Summary

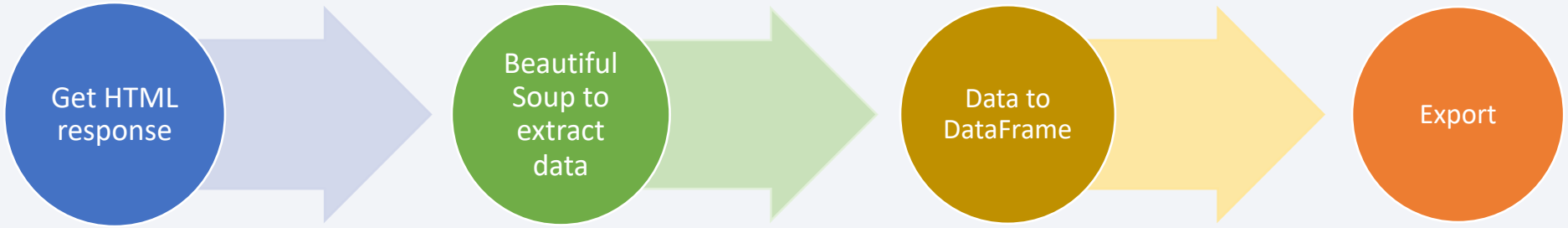
- Data collection methodology:
  - SpaceX API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Dropping unnecessary columns
  - Deal with missing values
  - One Hot Encoding for success or failed launches
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

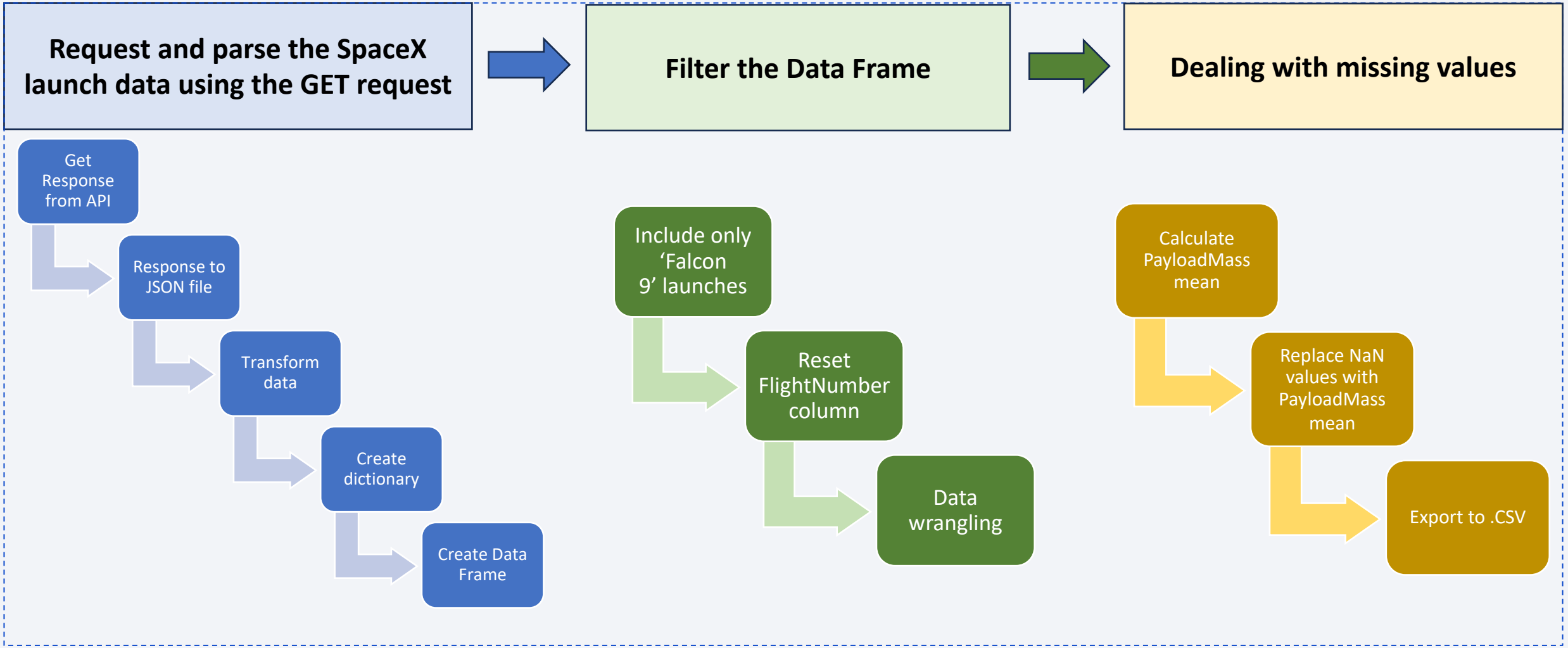
- From SpaceX API:



- From Wikipedia using Web Scrapping:

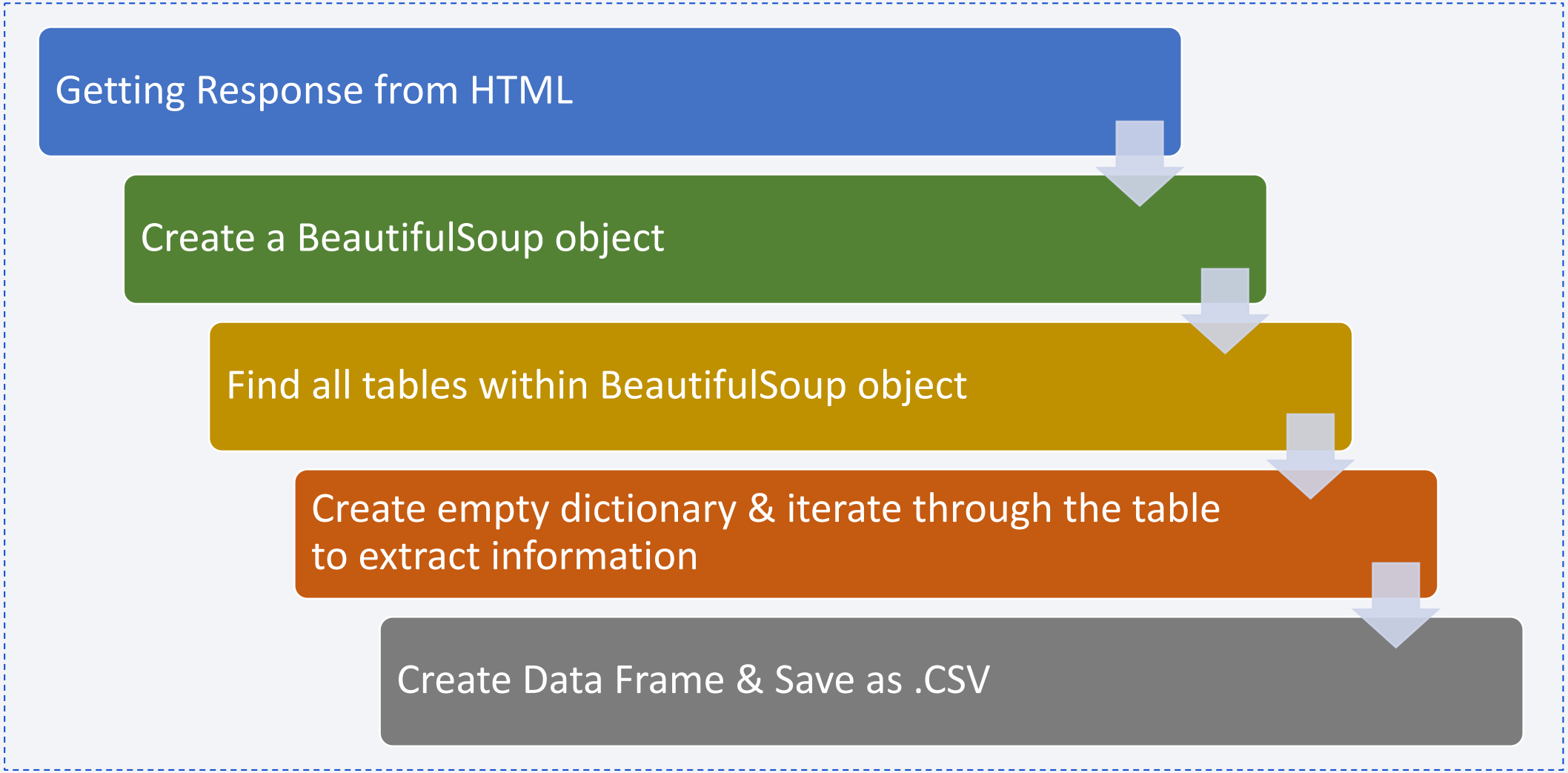


Data Collection – SpaceX API



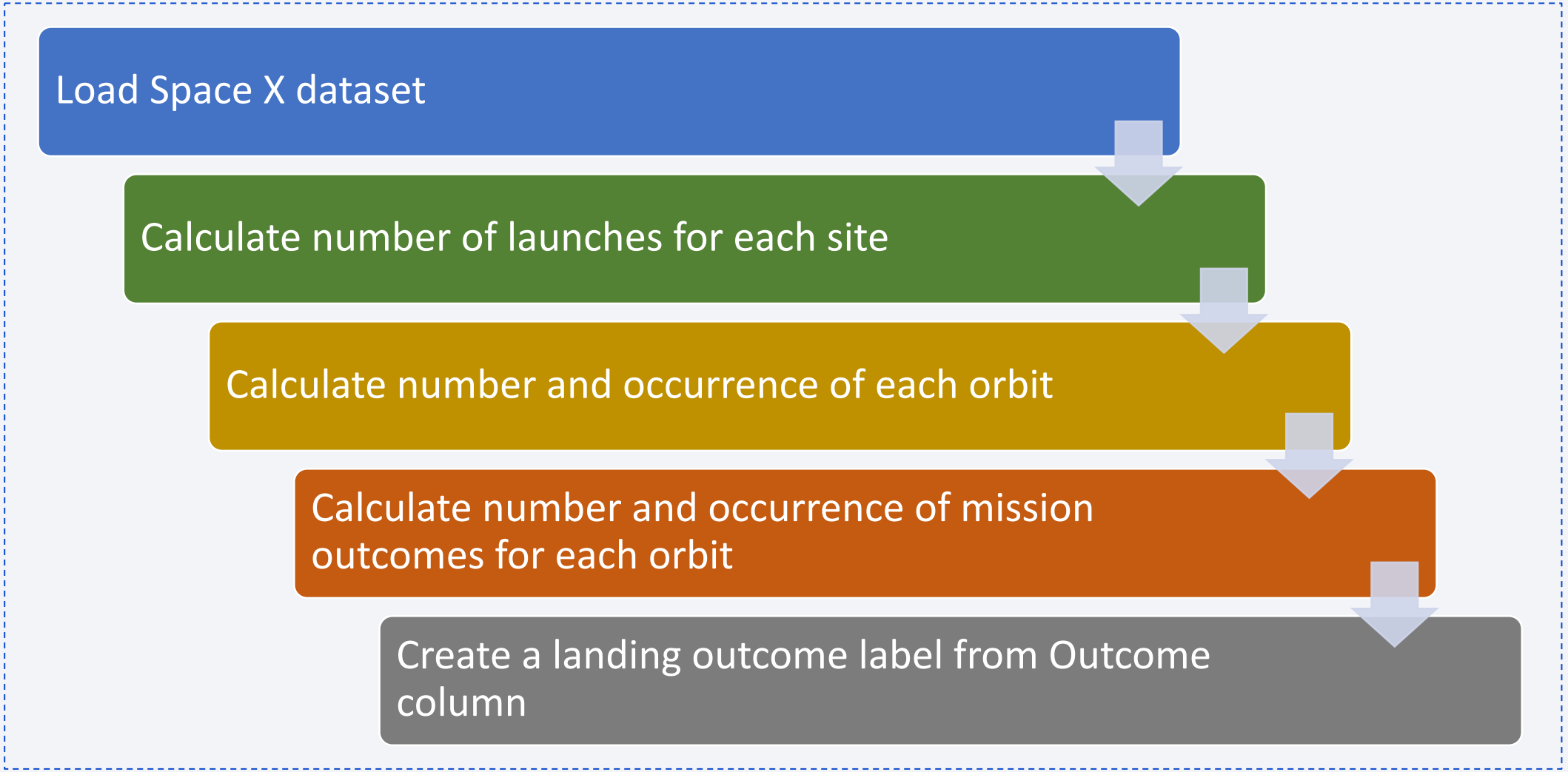
Link to Github repository

Data Collection - Scraping



Link to Github repository

Data Wrangling



Link to Github repository

EDA with Data Visualization

To perform EDA with Data Visualization, several charts were created:

- Scatter plots useful to visualize correlation between variables, like:
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Flight Number vs.Orbit Type
  - Payload vs. Orbit Type
  - Orbit Type vs. Payload Mass
- Line graph to show data trends between:
  - Success vs. Year
- Bar plot to make relationships between numerical and categorical data clearer, such as:
  - Success Rate vs. Orbit Type

Link to Github repository

EDA with SQL

Queries performed:

- Displaying the names of the unique launch sites in the space mission
- Listing 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying the average payload mass carried by booster version F9 V1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Displaying the name of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Displaying the names of the booster versions which have carried the maximum payload mass using a subquery
- Listing the records which will display month names, 'failure' landing outcomes in drone ship, booster versions and launch site in 2015.
- Rank the count of landing outcomes between 2010-06-04 and 2017-03-20, in descending order

Link to Github repository

Build an Interactive Map with Folium

A Folium map was created where the initial location was NASA Johnson Space Center (Houson, Texas). The following objects where added:

- Blue circle at NASA Johnson Space Center's coordinates with label showing its name.
- Circles at each launch site coordinates, together with a label showing each site's name.
- Cluster of markers to display multiple and different information for the same coordinates. Depending on the outcome of the landing, different colors where used: green for SUCCESSFUL and red for UNSUCCESSFUL.
- MousePosition to show the coordinates of a specific location (railway, highway, coastline).
- Polyline to draw a line between a launch site and a specific location (railway, highway, coastline).

All these objects provide a better understanding, not only of the problem at hand, but of the the data that we are working with.

Visualizing the surrounding areas of the launch sites, as well as the outcome of each launch, can provide key information that would be harder to decipher otherwise.

Link to Github repository

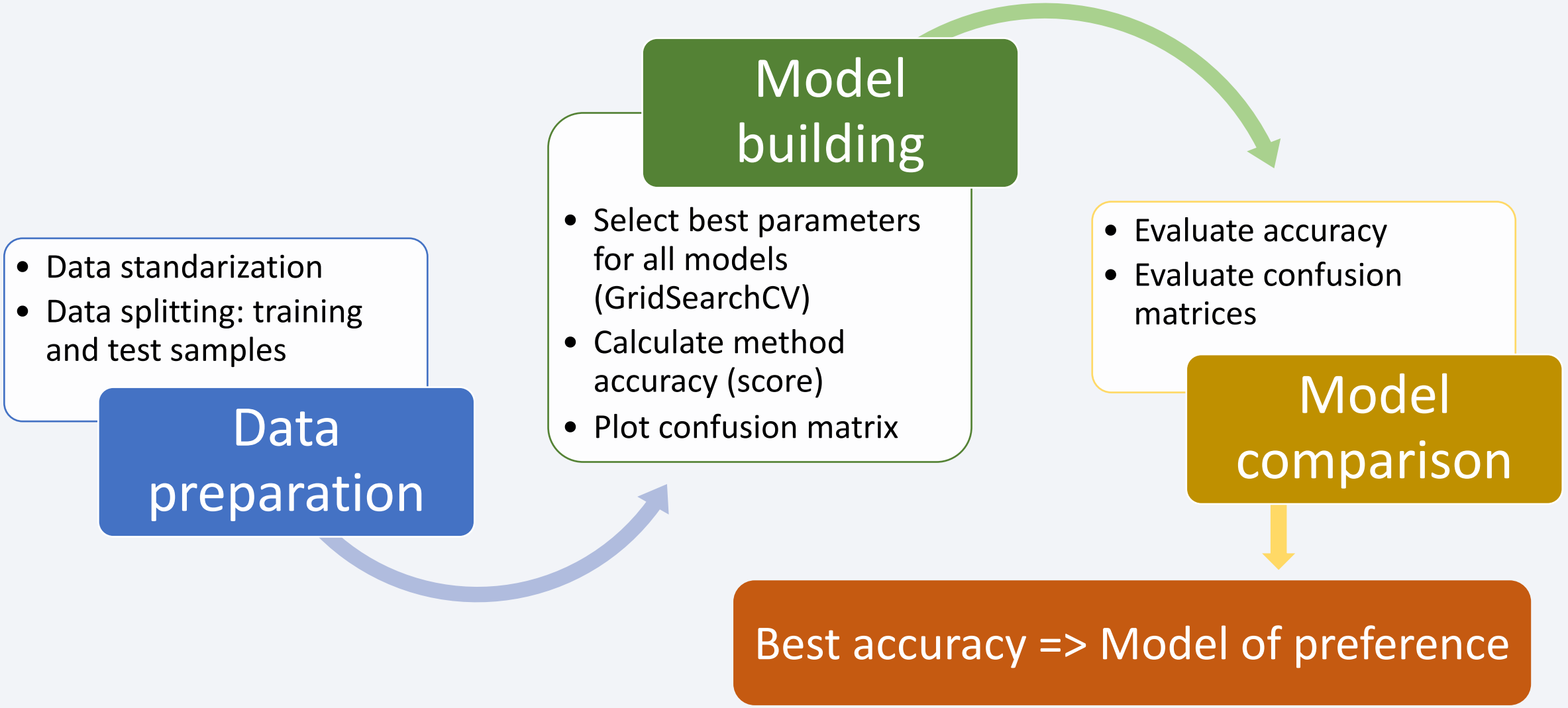
Build a Dashboard with Plotly Dash

A Dashboard was created using Plotly Dash, and several features were added for a better visualization:

- A dropdown that allows to choose between a specific launch site or all sites at once
- If all sites are selected, a pie chart was included to display the launch site success count for every place. However, if a specific site is selected in the dropdown, the pie chart will display success and failure counts for that place.
- A rangeslider allows to select payload mass in a specific range
- Depending on the range selected, a scatter chart was included to show the relationship between the payload mass and the launch outcome by booster version.

Link to Github repository

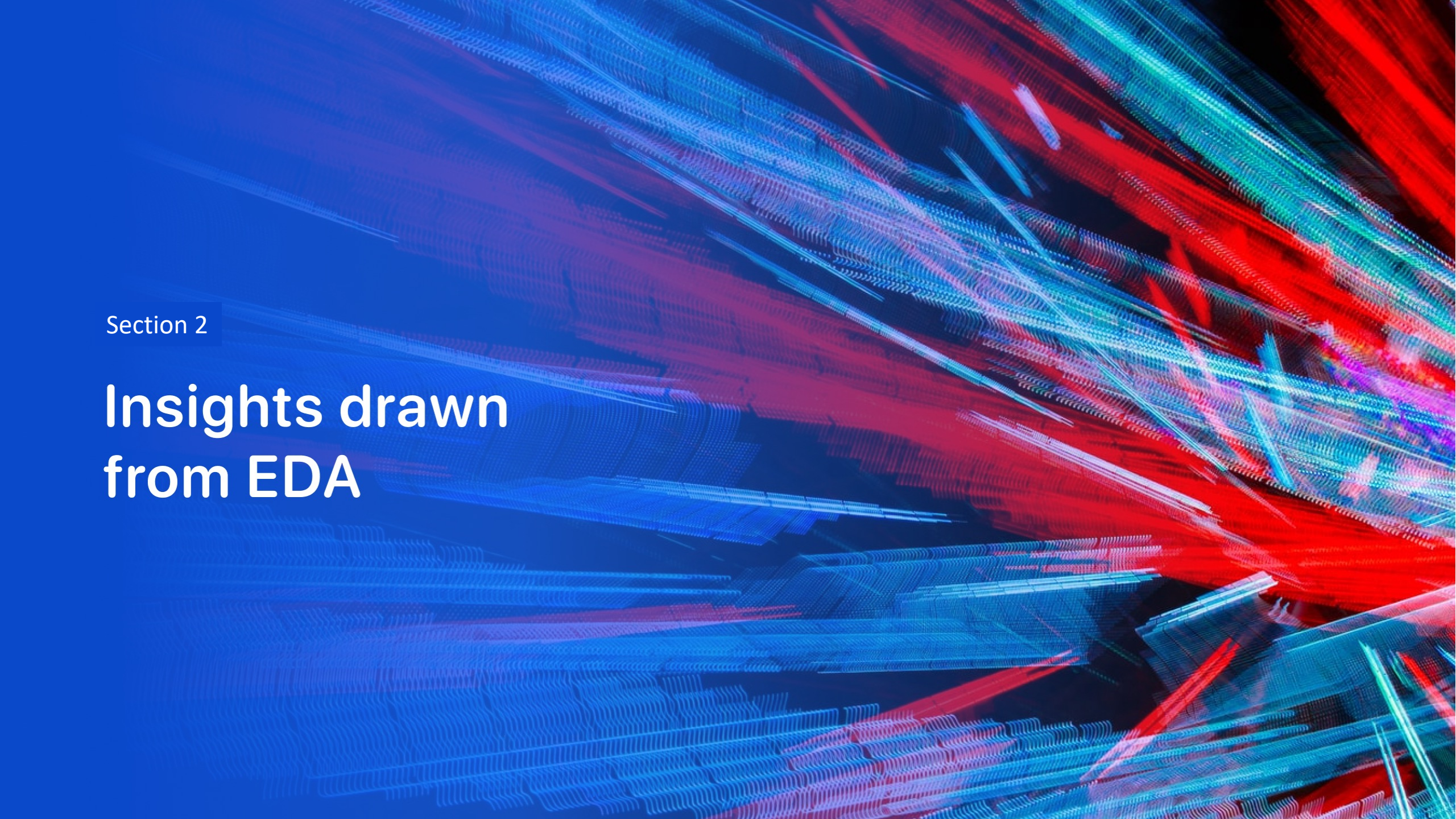
Predictive Analysis (Classification)



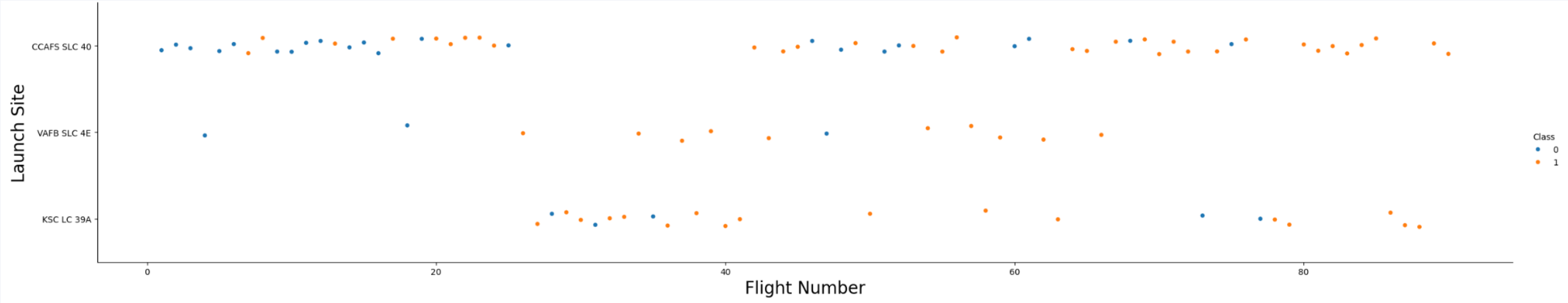
Link to Github repository

Results

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

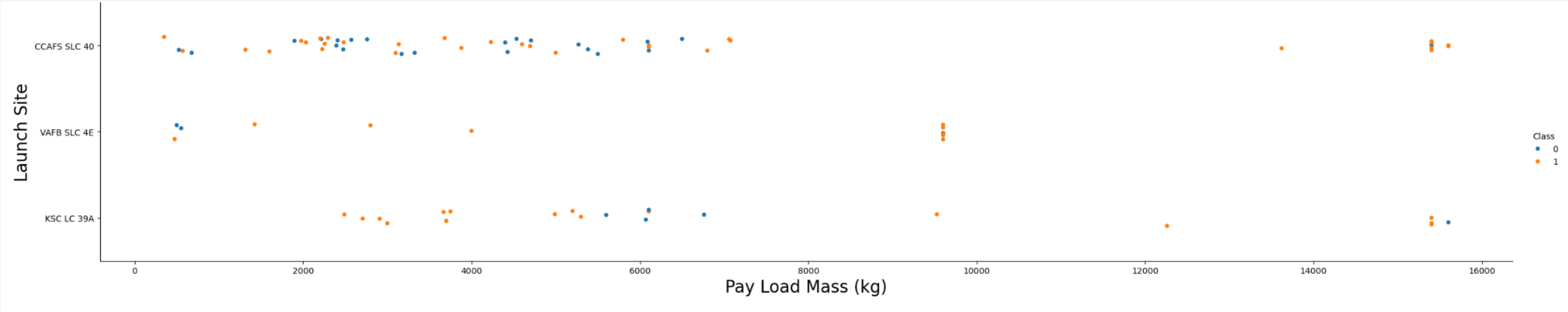

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



For each site, when the flight number is higher, the more likely is the first stage to land successfully

Payload vs. Launch Site

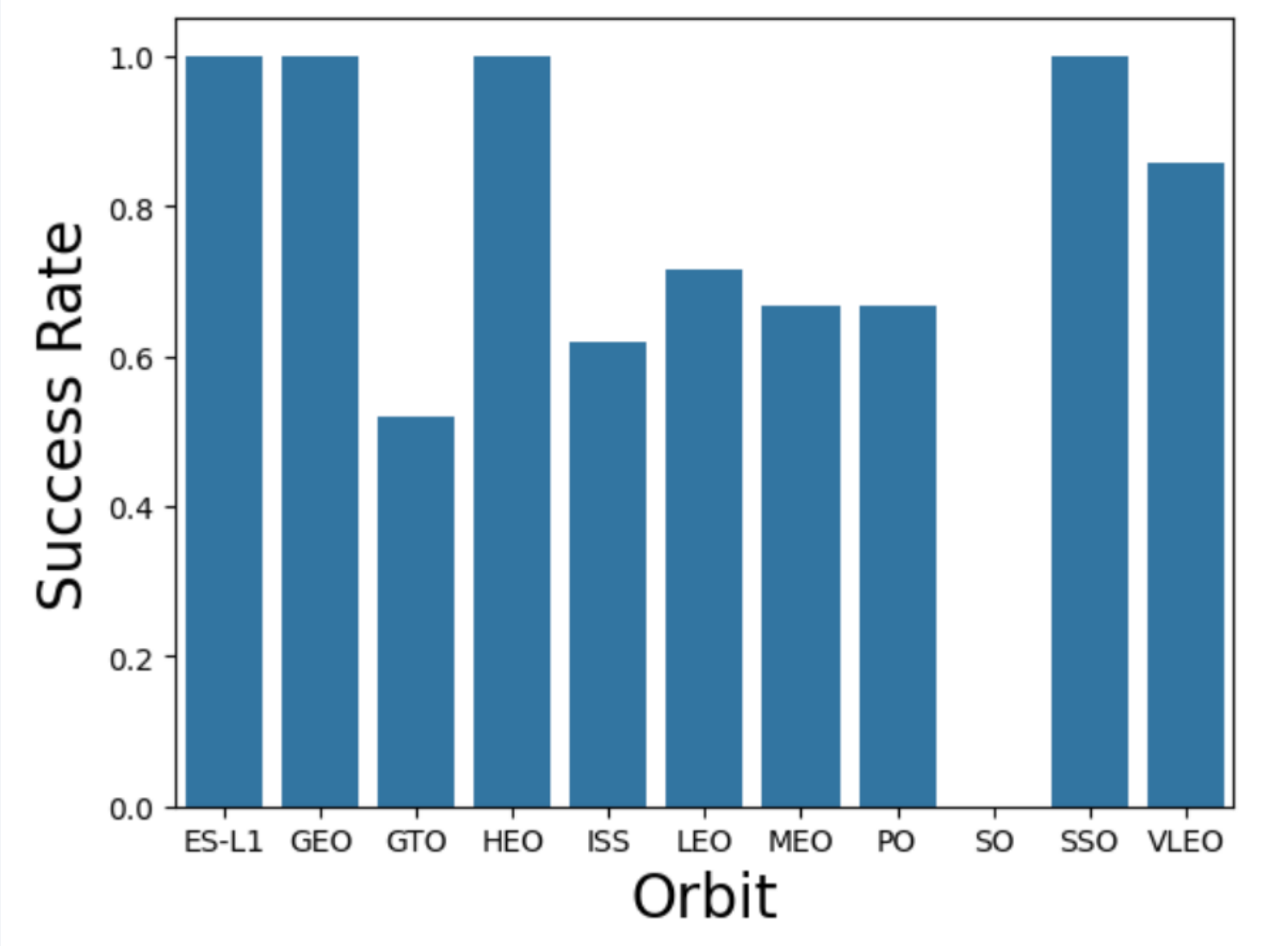


The data indicates that the higher the payload, the more likely the first stage is to return.

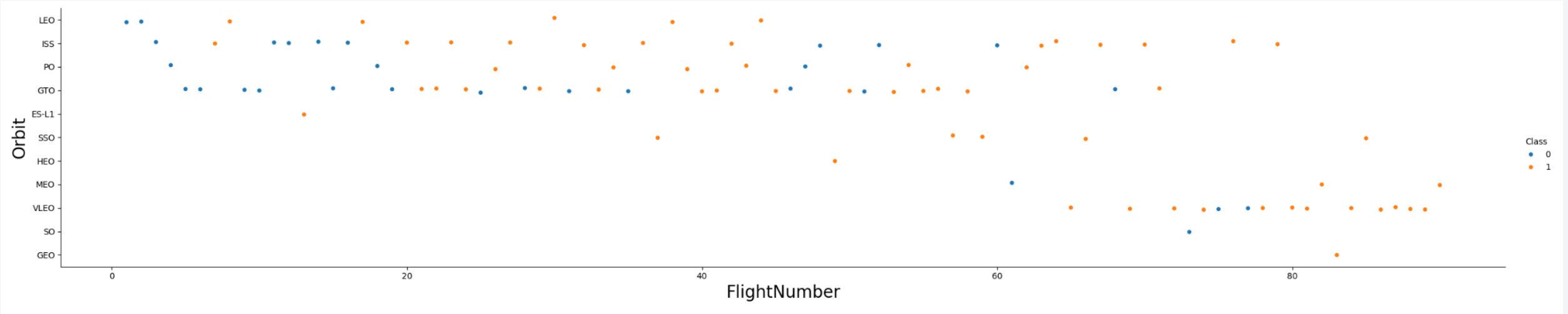
Note: for the VAFB-SLC 4E launch site there are no rockets launched with payload mass greater than 10000 for some reason.

Success Rate vs. Orbit Type

The Orbit types with the higher success rate are ES-L1, GEO, HEO and SSO



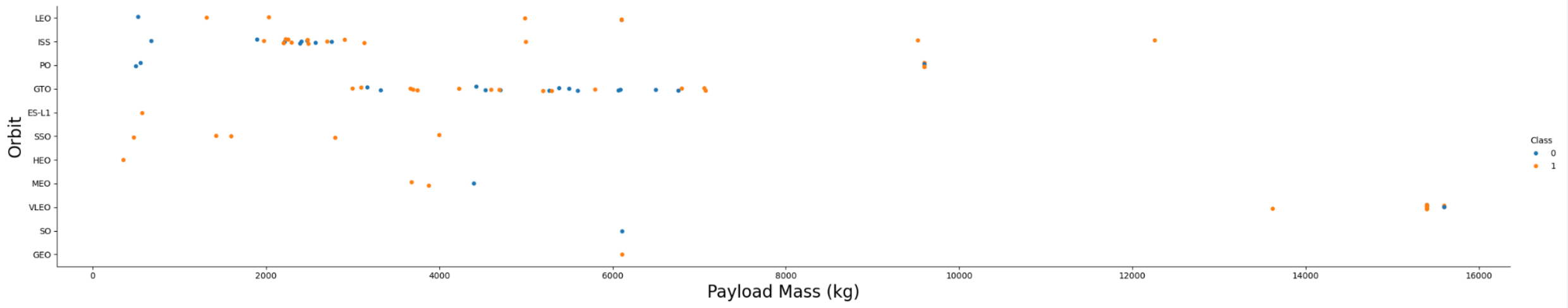
Flight Number vs. Orbit Type



For most orbit types, as the flight number increases, the success rate also gets higher.

However, the same relationship doesn’t seem to apply when in GTO orbit.

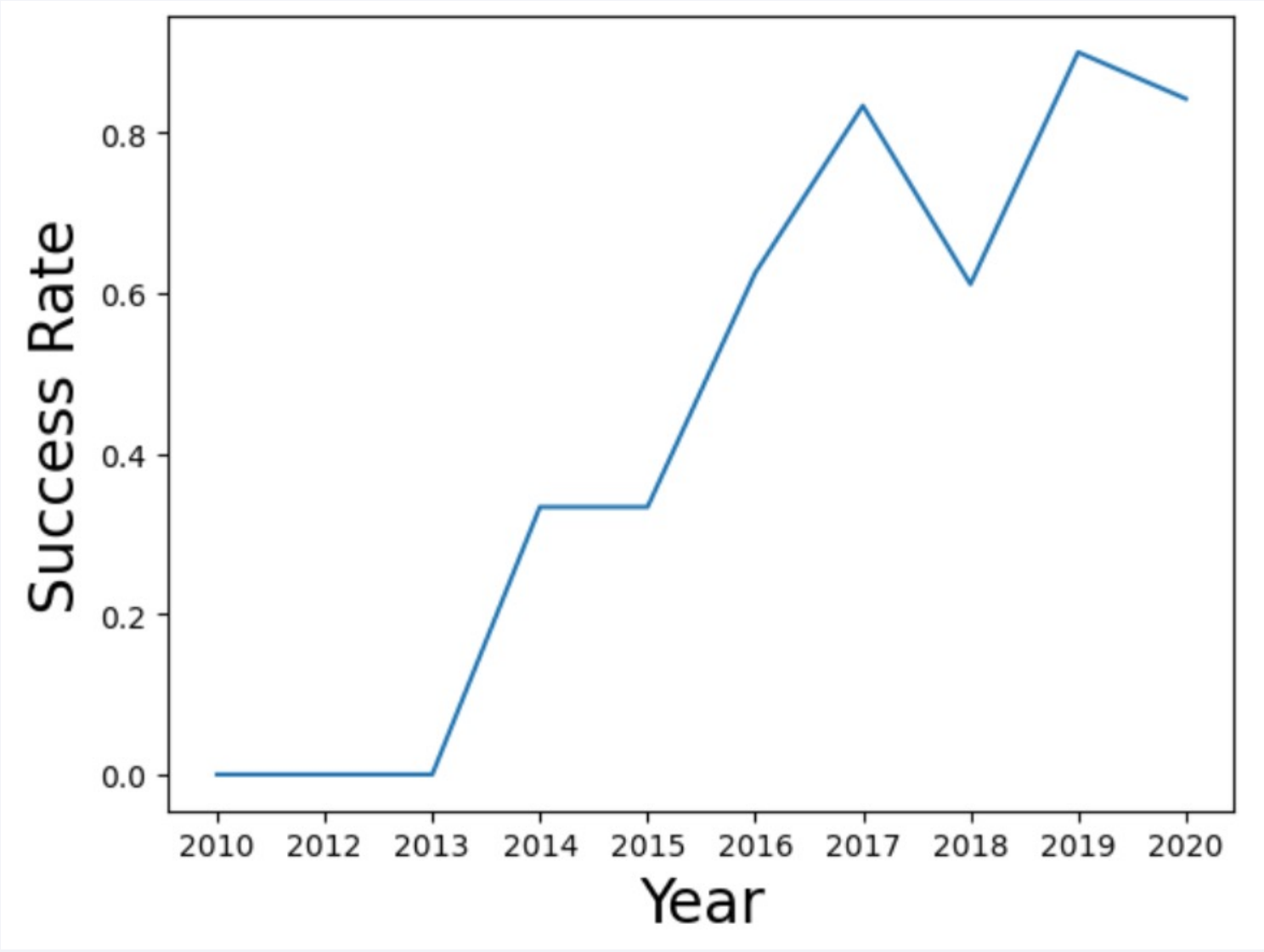
Payload vs. Orbit Type



For LEO, IS, and PO, as the payload get heavier, the successful rate increases. However, for GTO, this relationship does not apply.

Launch Success Yearly Trend

Since 2013, the success rate kept increasing untill 2020, with a slight decrease in 2018.



All Launch Site Names

```
SELECT DISTINCT Launch_Site FROM SPACEXTABLE order by Launch_Site desc
```

The use of “DISTINCT” allows for unique Launch Sites to be listed.

The use of ”desc” after the “order by” delivers the Launch Sites listed in descending order

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

Launch Site Names Begin with 'CCA'

```
SELECT * FROM SPACEXTABLE 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

To find 5 records of launch sited beginning with ‘CCA’, the ”WHERE” clause needs to be used along with the “LIKE” clause, followed by the string ‘CCA%’ that indicates that the Launch Site name need to begin with ‘CCA’.

Finally, to limit the records to a total of 5, the “LIMIT” clause also needs to be used

Total Payload Mass

```
SELECT SUM(PAYLOAD_MASS__KG_) AS "TOTAL_PAYLOAD_MASS" FROM SPACEXTABLE GROUP BY Customer HAVING Customer LIKE "NASA (CRS)"
```

TOTAL_PAYLOAD_MASS
45596

This query returns the total payload mass (in Kg) carried by boosters that have been launched by NASA (CRS)

Average Payload Mass by F9 v1.1

```
SELECT AVG(PAYLOAD_MASS__KG_) AS "AVERAGE_PAYLOAD_MASS" FROM SPACEXTABLE GROUP BY Booster_version HAVING Booster_Version Like "F9 v1.1"
```

AVERAGE_PAYLOAD_MASS
2928.4

This query returns the average payload mass (in Kg) carried only by booster version F9 v1.1

First Successful Ground Landing Date

```
SELECT MIN(Date) FROM SPACEXTABLE GROUP BY Landing_Outcome HAVING Landing_Outcome LIKE "Success (ground pad)"
```

MIN(Date)
2015-12-22

With this query, we can find the date when the first succesful landing outcome in ground pad was acheived

Successful Drone Ship Landing with Payload between 4000 and 6000

```
SELECT Booster_Version FROM SPACEXTABLE GROUP BY Booster_Version HAVING Landing_Outcome LIKE "Success (drone ship)" AND 4000 < PAYLOAD_MASS__KG_ < 6000
```

Booster_Version
F9 B4 B1041.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1031.2
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1036.1
F9 FT B1038.1

This is the way to list the names of the booster versions which have successfully landed on drone ship and had payload mass greater than 4000 Kg but less than 6000 Kg

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT (SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Success%') AS Sucess,
(SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Failure%') AS Failure
```

Sucess	Failure
100	1

For this query, a total of two subqueries must be used:

- The first one counts the number of successful missions,
- The second subquery counts the total of unsuccessful missions.

In order to filter the mission outcome, a WHERE clause must be followed by a LIKE clause.

To sum all missions of a specific type, the COUNT clause has to be used.

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT "BOOSTER_VERSION" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT max("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
```

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

To answer which boosters carried the maximum payload, a subquery might be used to filter data, where a MAX clause returns the the heaviest payload mass.
The SELECT DISTINCT in the main query takes the subquery results for unique booster versions.

2015 Launch Records

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

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

This query allows to list the failed landing outcomes in drone ship, along with their booster versions, and launch site names in a specific month of 2015

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

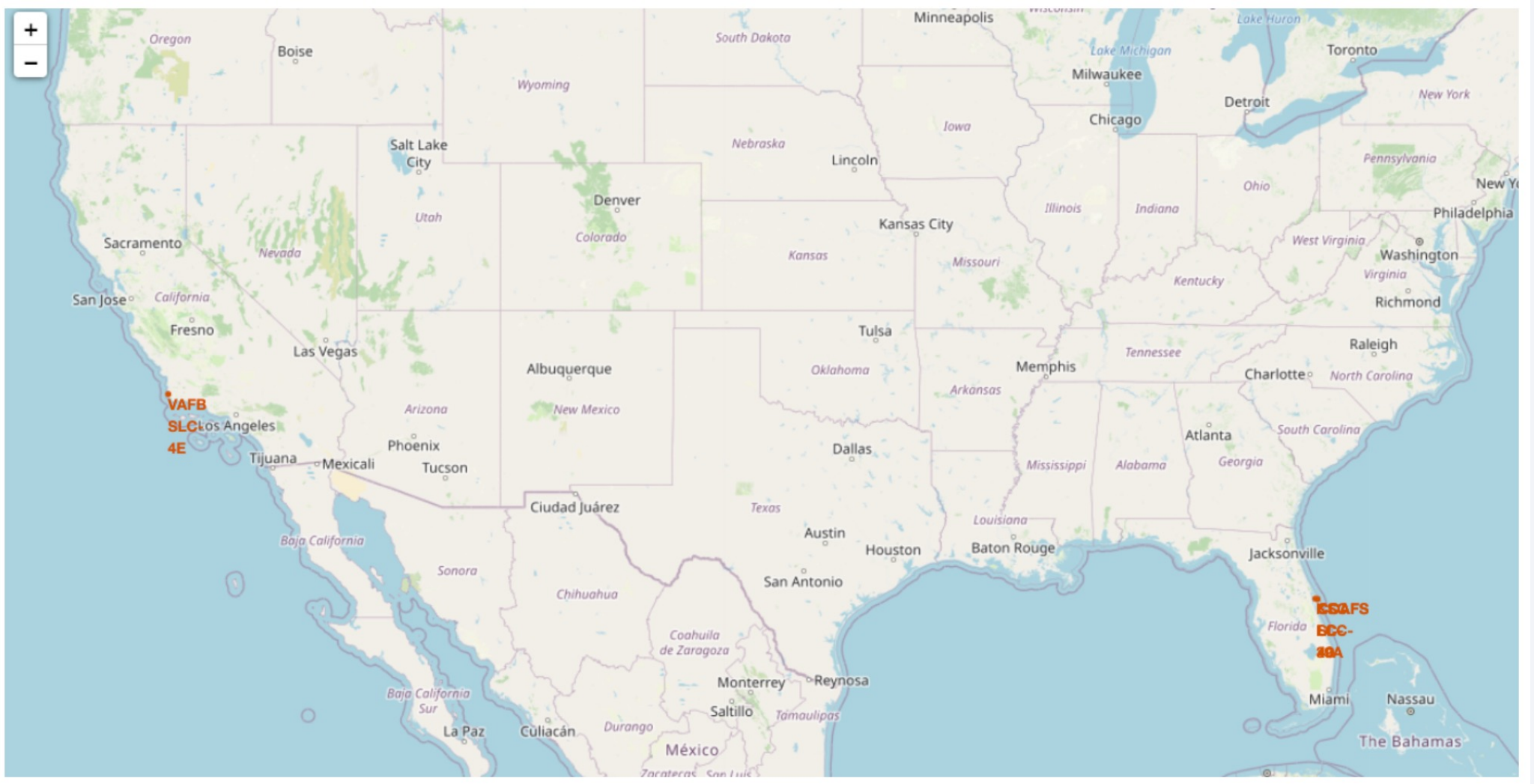
```
SELECT 2010-06-04>= Date <=2017-03-20 AS Date, Landing_Outcome, COUNT(Landing_Outcome) FROM SPACEXTABLE GROUP BY Landing_Outcome order by COUNT(Landing_Outcome) DESC
```

With this query, we can rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between 2010-06-04 and 2017-03-20, in descending order.

Date	Landing_Outcome	COUNT(Landing_Outcome)
1	Success	38
1	No attempt	21
1	Success (drone ship)	14
1	Success (ground pad)	9
1	Failure (drone ship)	5
1	Controlled (ocean)	5
1	Failure	3
1	Uncontrolled (ocean)	2
1	Failure (parachute)	2
1	Precluded (drone ship)	1
1	No attempt	1


Section 3
Launch Sites
Proximities Analysis

Folium Map with Launch Sites

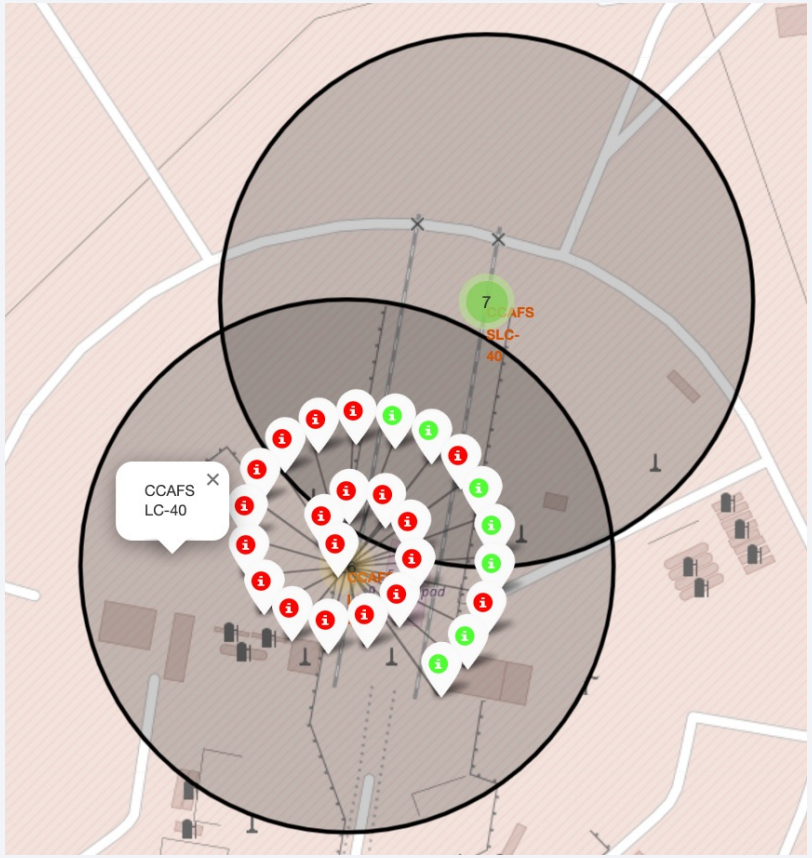


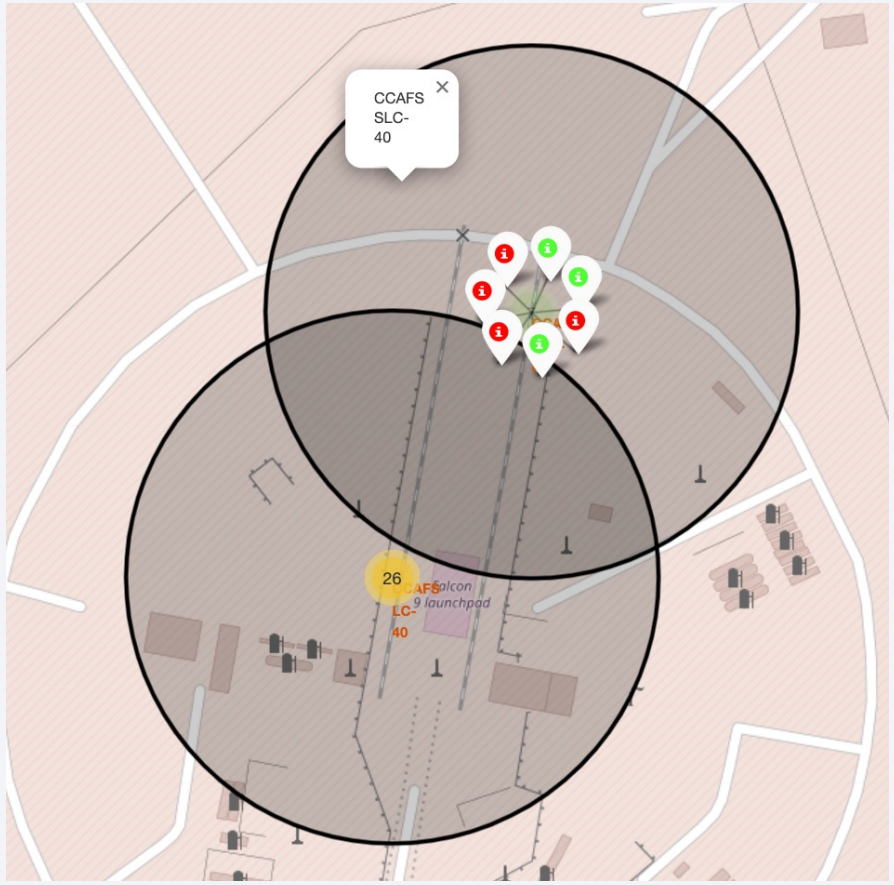
Launch sites are shown in this Folium map, and it is evident that they locate on opposite coasts of the United States

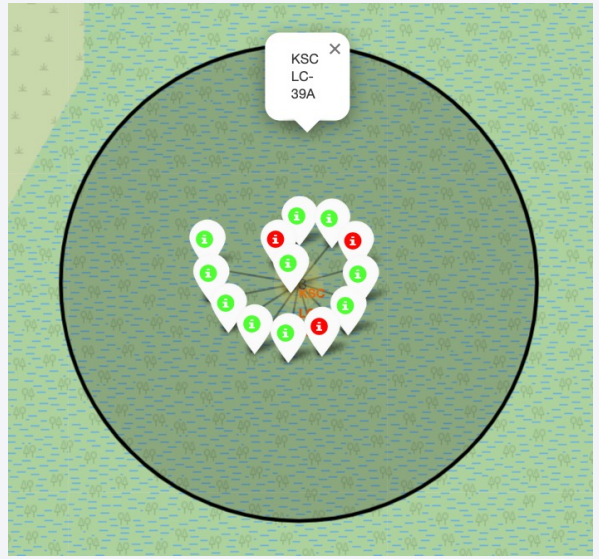
Folium Map with color-labeled launch outcomes

Successful launch outcomes are marked with green, while the failed ones are red.

The site with the highest success rate appears to be KSC LC-39A

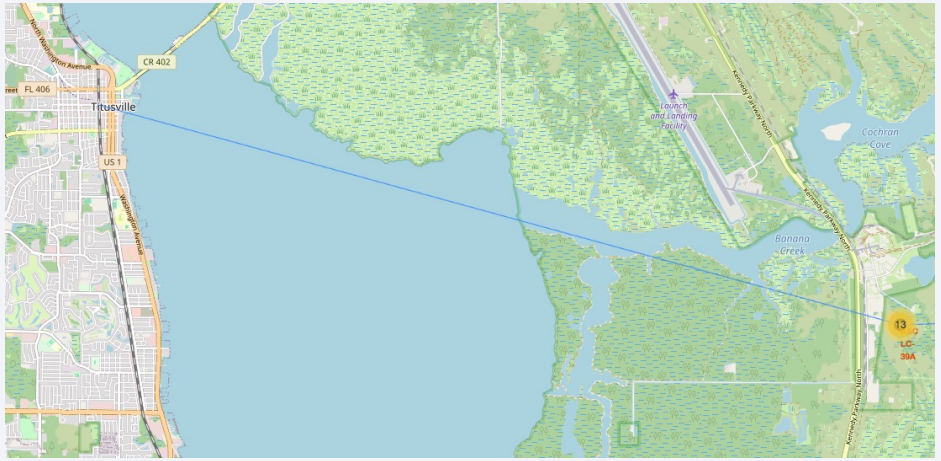




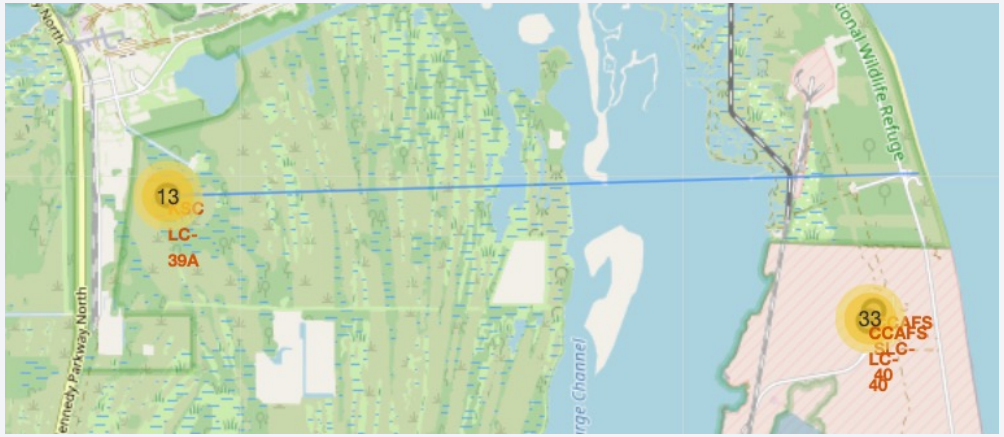


KSC LC-39A and its proximities

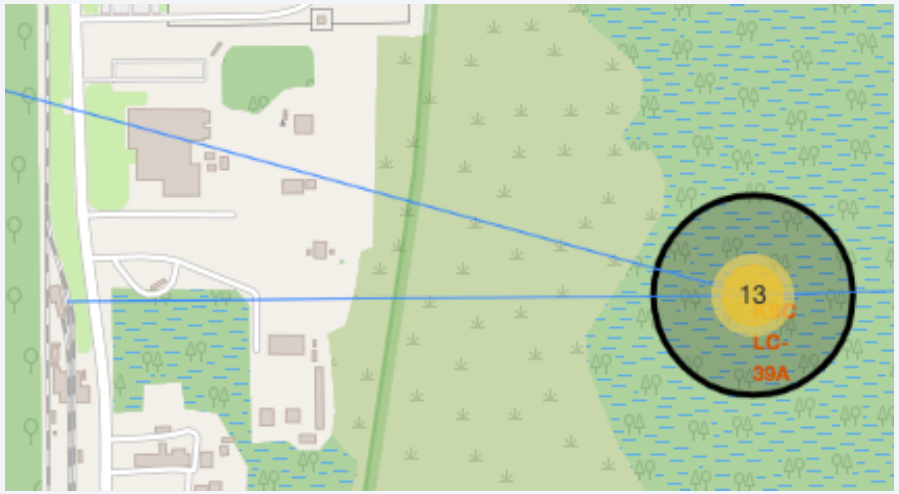
KSC LC-39A is relatively far from any city



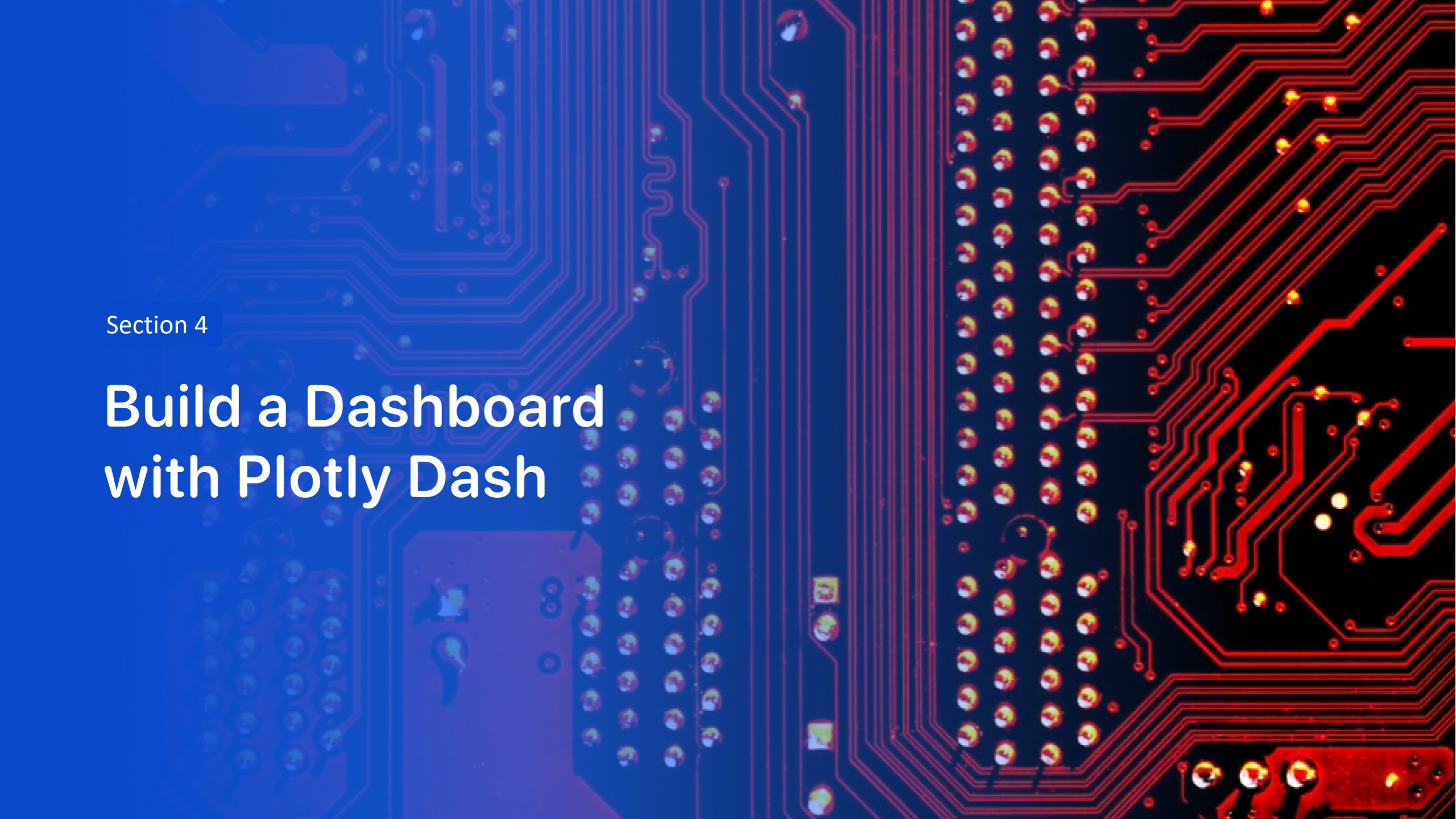
KSC LC-39A is relatively close to the Atlantic ocean's coastline



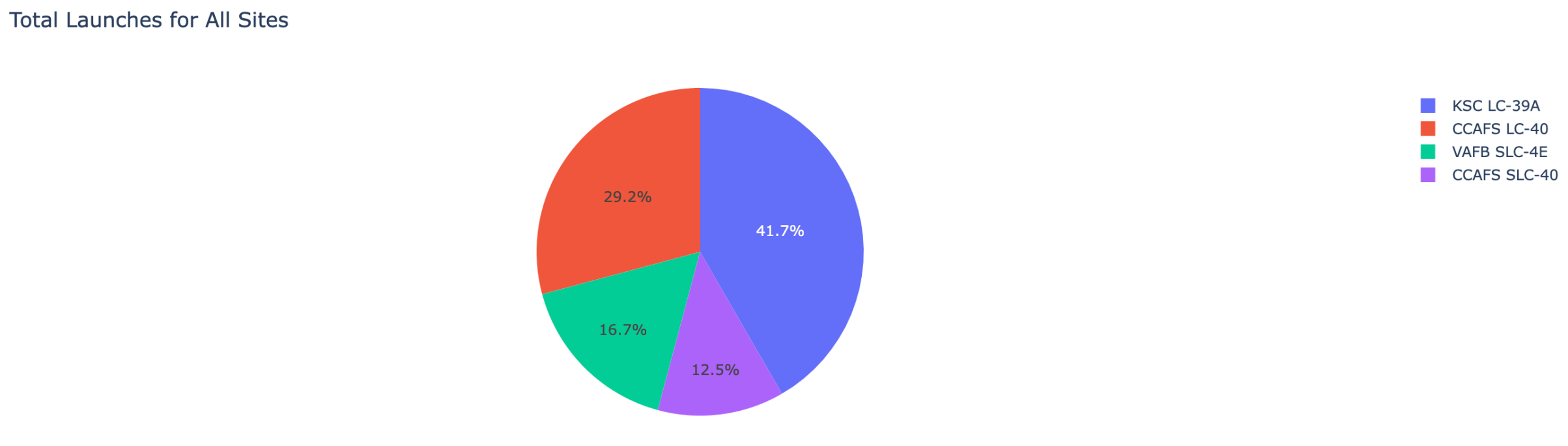
KSC LC-39A is close to a railway



KSC LC-39A is close to a highway

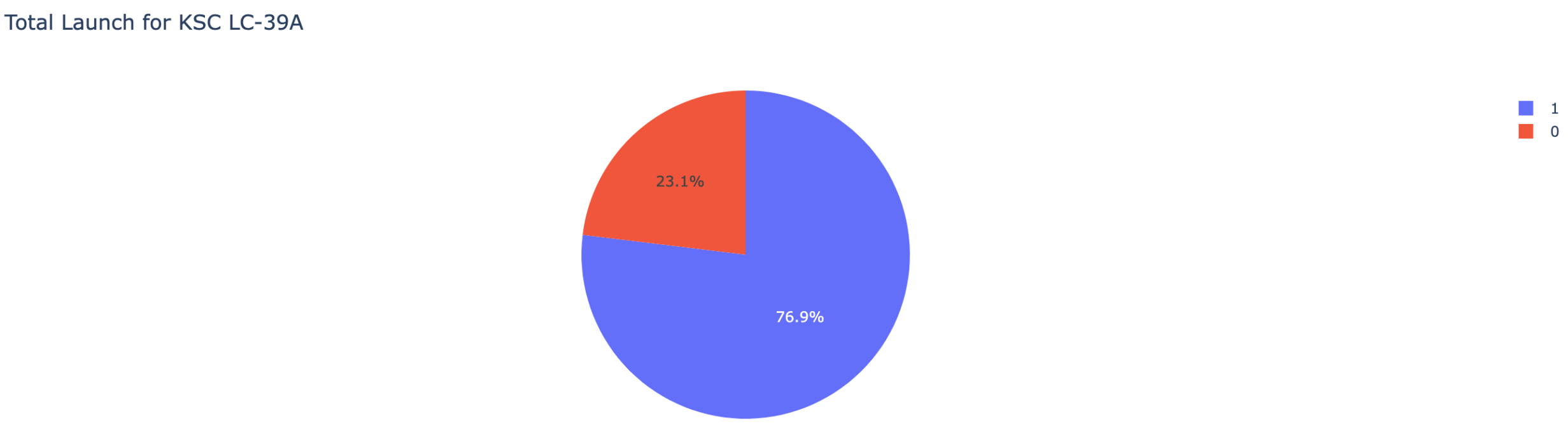

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



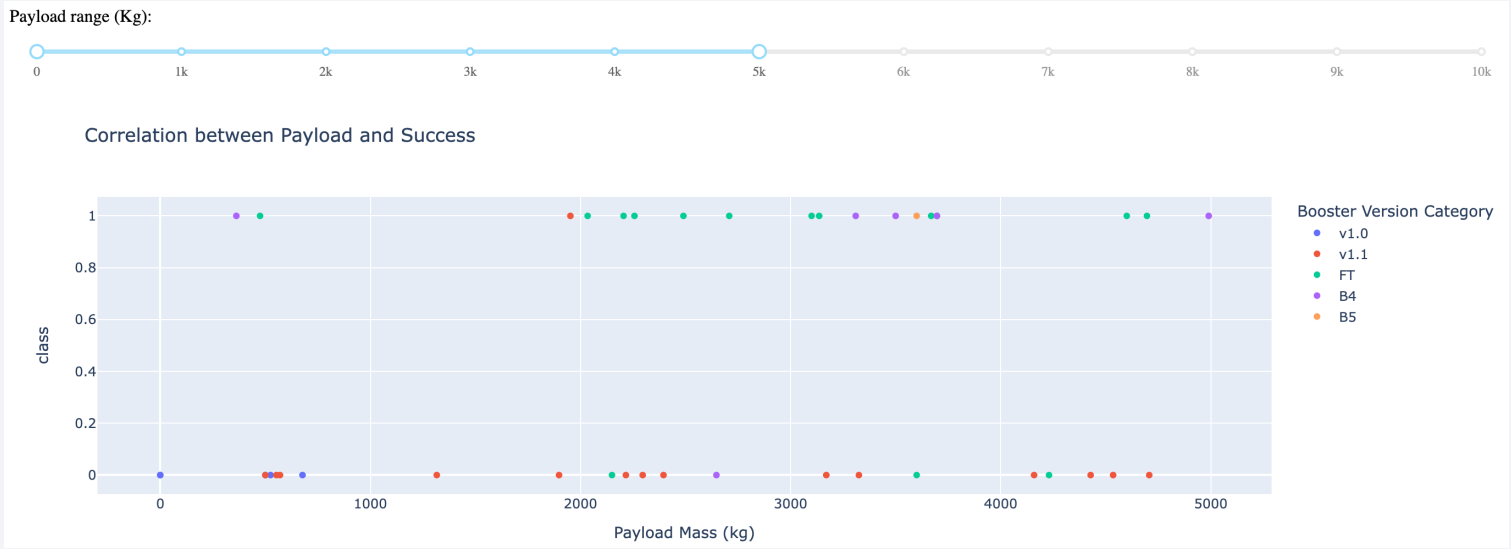
This pie chart best describes launch success count for all sites.
The one with the highest success rate is KSC LC-39A.

Launch site with highest launch success ratio



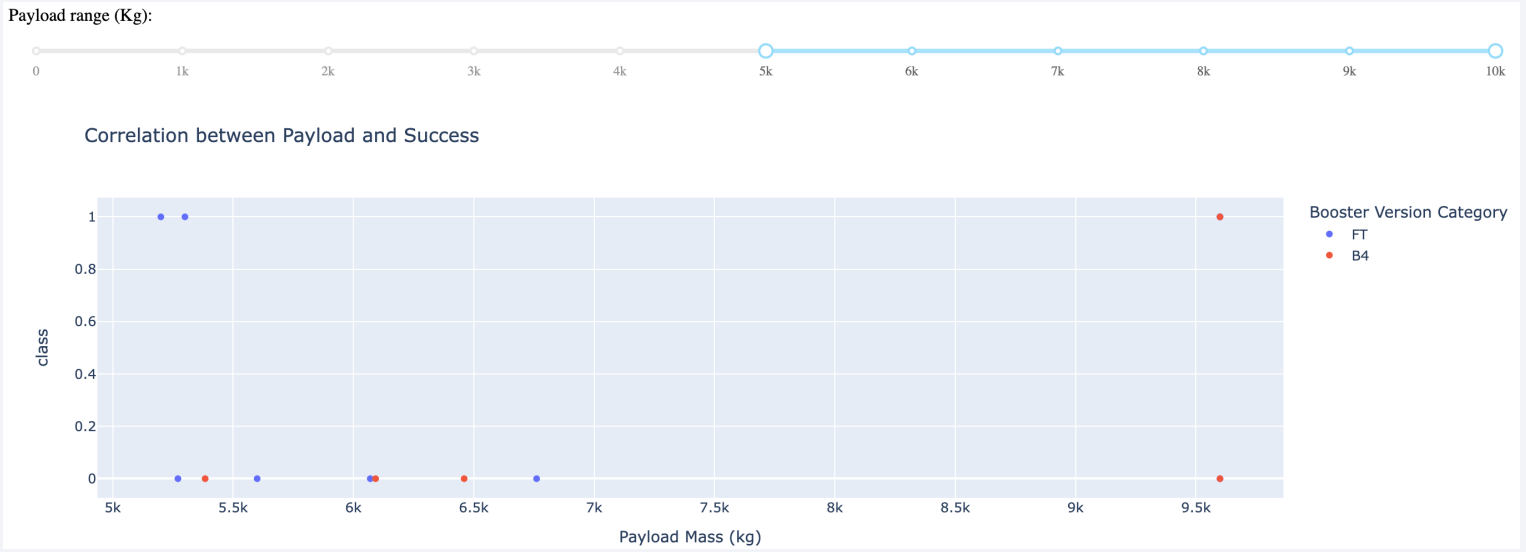
For KSC LC-39A, successful launches represent the 76.9% of all launches in this site, while failed launches only represent a 23.1%

Payload vs. Launch Outcome for all sites



These scatter plot show the relationship between Payload Mass and Launch Outcomes for all sites, categorized by Booster Version.

It is shown that the success rate is higher for lower payloads



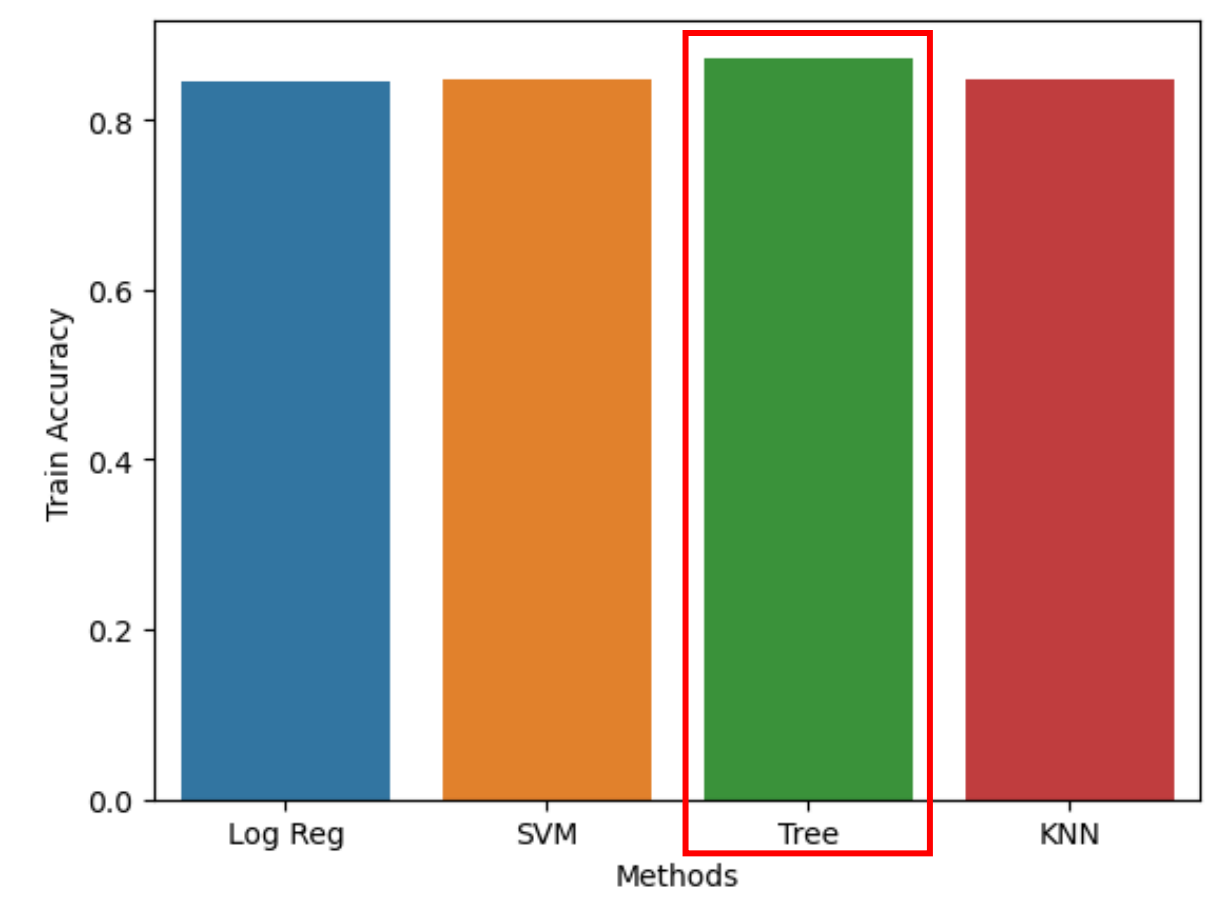
Section 5

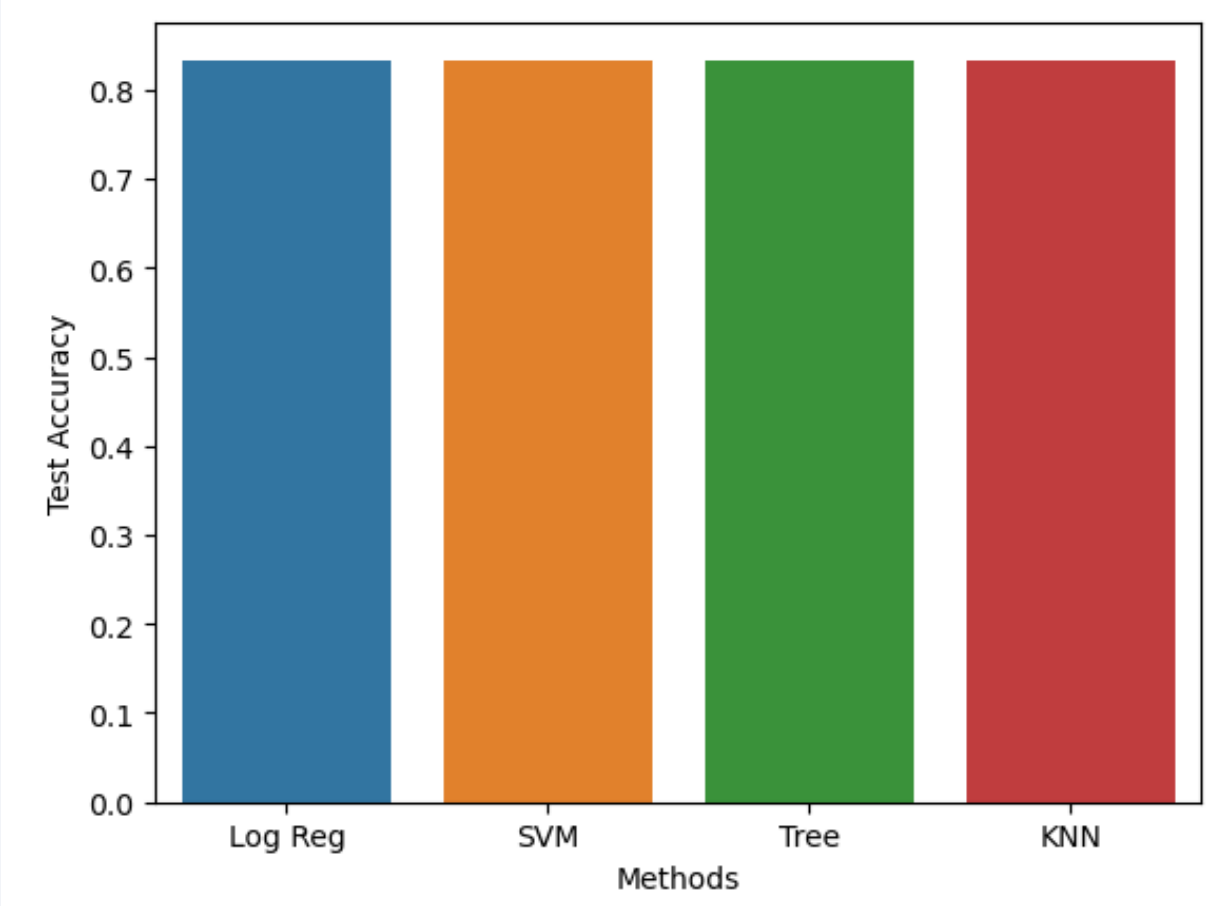
Predictive Analysis (Classification)

Classification Accuracy

	Methods	Train Accuracy	Test Accuracy
0	Log Reg	0.846429	0.833333
1	SVM	0.848214	0.833333
2	Tree	0.873214	0.833333
3	KNN	0.848214	0.833333

For test data, all methods have similar performance. For train data, Decision Tree Classification is superior to the other three

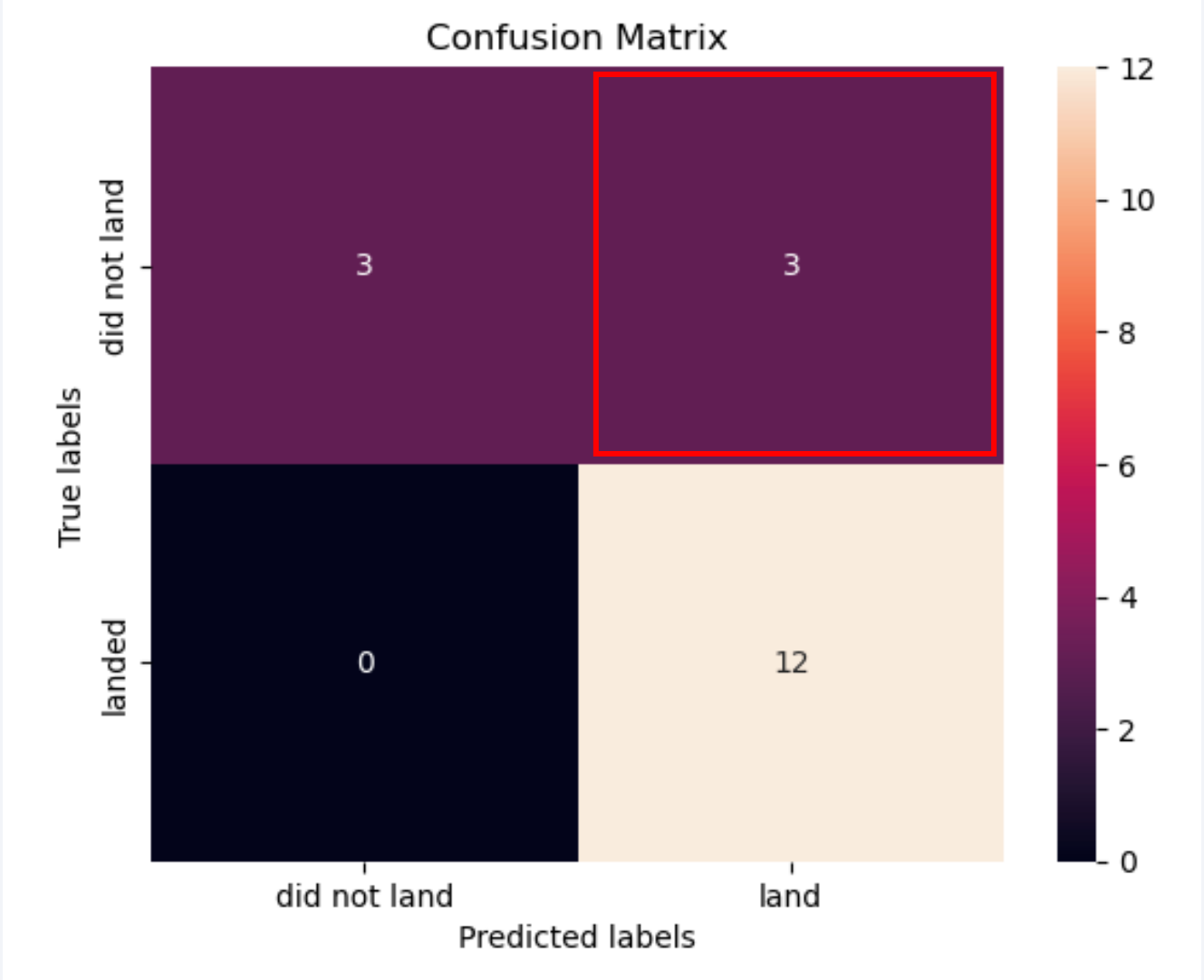




Confusion Matrix of the Decision Tree Classification

Even though the Decision Tree Classification method was superior to the other three methods, there are still false positives in the confusion matrix that represent 17% of the test data

The best way to decide would be to incorporate more data to the models, so that the scores and confusion matrices can be improved



Conclusions

- Specifically, from the EDA, we can say:
  - If the flight number is higher, the more likely is the first stage to land successfully, and this is true for all launch sites.
  - The higher the payload, the more likely the first stage is to return. However, more data is required from certain launch sites, such as VAFB-SLC 4E that does not have payload mass greater than 10000 kg.
  - Plotting success rate vs. orbit type as a bar graph, revealed that ES-L1, GEO, HEO and SSO orbits have higher success rate than the rest.
  - A scatter plot revealed that for most orbit types, as the flight number increases, so does the success rate. However, this relationship proved to be false for the GTO orbit.
  - Plotting payload vs. orbit type as a scatter plot displayed that for LEO, IS and PO, the success rate increases when the payload gets heavier. GTO orbit, on the other hand, does not follow this relationship.
  - Yearly trend of launch success shows that since 2013 success rate increased until 2020, even though there was a slight decrease in 2018.
- EDA seems to indicate that certain variables can be related to a higher success rate, and that certain relationships between variables are not quite as clear as they should be to make a good assessment.

Conclusions

- SQL queries were an important tool to extract information from the data set, such as total and average payload mass, first successful ground landing date, the type of boosters that carried the maximum payload, and more.
- Overall, higher success rate is shown with lower payloads in the scatter plot displayed in the dashboard.
- KSC LC-39 A is the launch site with the highest success rate, according to the information provided in the pie chart illustrated in the dashboard and on the Folium map. Moreover, this success rate does not seem to be explained by the surroundings of KSC LC-39 A.
- Predictive analysis based on logistic regression, SVM, Decision Tree, and KNN, revealed that all methods performed similarly when it came to accuracy and confusion matrices. However, the Decision Tree method was slightly superior in accuracy with respect to the rest.
- More data would be needed to lower false positives from decision matrices and to improve accuracy. Another alternative would be to try another method.

Thank you!