IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nakul Sreekanth
11/25/2024

Outline

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

Executive Summary

Summary of methodologies

- Data Collection from SpaceX API
- Web scraping from SpaceX Wikipedia page
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Summary of all results

- ES-L1, GEO and HEO are among best orbits for mission success
- Increase in success rate from 2013-2020
- KSC LC 39-A is the most successful Launch Site
- The 2000 kg to 4000 kg range for the Payload Mass has the highest success rate
- The FT Booster Version has the highest success rate
- All predictive methods work equally well, with an 83.3% accuracy rate

Introduction

Background and Context

- The commercial space industry is growing, and SpaceX is a current leader in the market, offering affordable options for customers
- SpaceX's success can be attributed to the reusability of the First Stage of their rocket
- SpaceY is a new company that aims to compete with SpaceX
- SpaceY wants to use SpaceX's data to help their own company make decisions

Problems you want to find answers

- Predicting First Stage Reusability
    - Finding different predictors of mission success

Section 1

Methodology

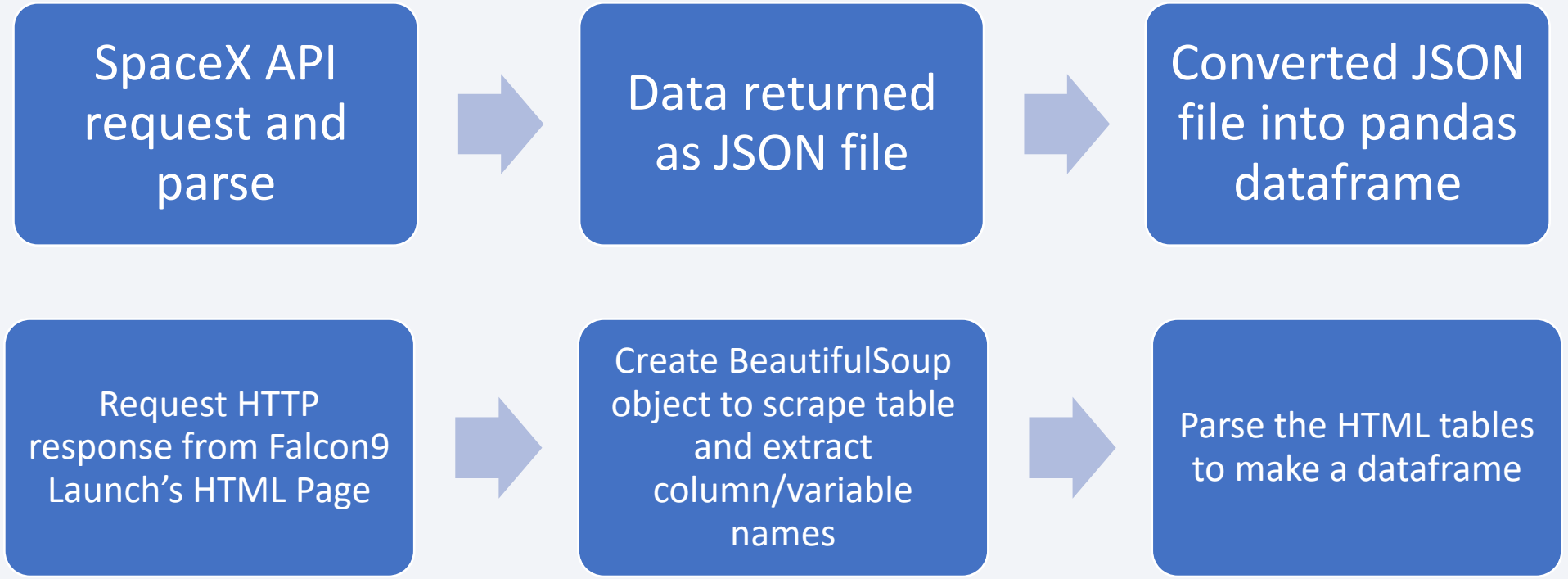
Methodology

Executive Summary

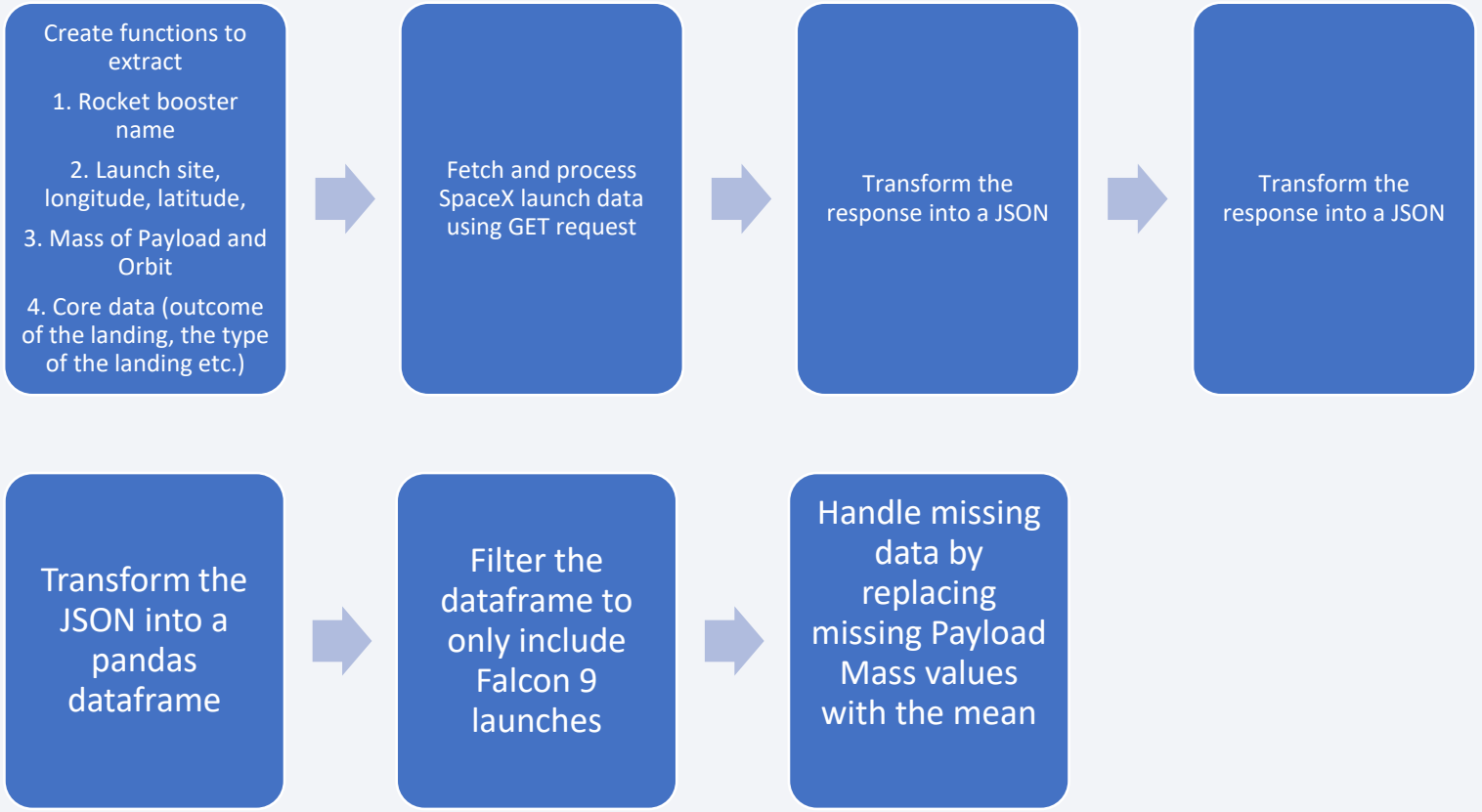
- Data collection methodology:
  - SpaceX Rest API to obtain rocket launch data
  - Wikipedia SpaceX page for Web Scraping
- Perform data wrangling
  - Data content was normalized and transformed into a pandas dataframe
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Used machine learning and Grid Search method to find the best parameters for SVM, Classification Trees and Logistic Regression

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts



Data Collection – SpaceX API

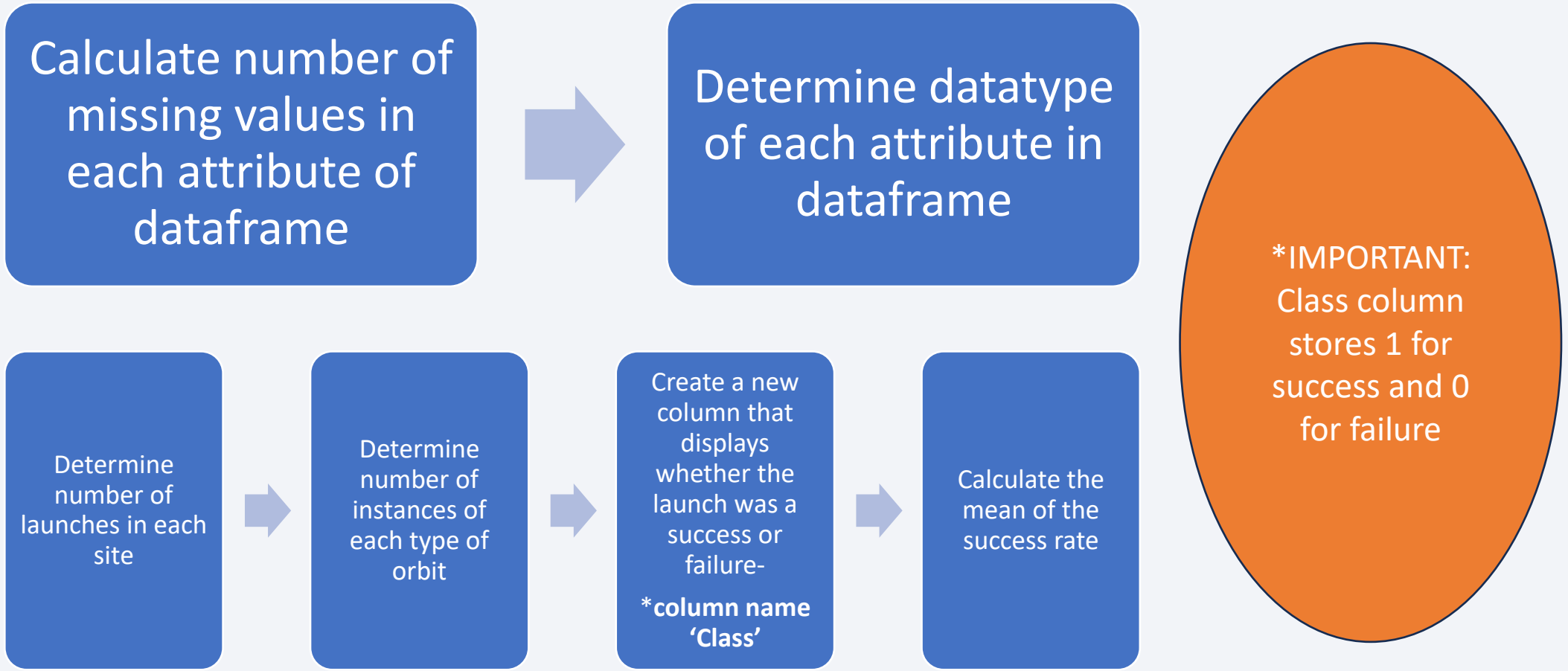


Retrieving SpaceX API

Data Collection - Scraping

[Web scraping Falcon 9 Data from Wikipedia]

Data Wrangling



[Data Wrangling and EDA for SpaceX Falcon 9 First Stage Landing Prediction]

EDA with Data Visualization

Charts plotted to draw potential correlation to mission success

- Catplot of Flight Number vs. Launch Site
- Catplot of Payload Mass vs. Launch Site
- Bar plot of Orbit Type vs. Average Success Rate
- Catplot of Flight Number vs. Orbit Type
- Catplot of Payload Mass vs Orbit Type
- Line plot of Yearly success

Data Visualization

EDA with SQL

Performed Queries to find

- All Launch Site Names
- Records of Launch Site Names beginning with CCA
- Total Payload Mass in NASA Launches
- Average Payload Mass by F9 v1.1
- First Successful Ground Landing Date
- Successful Drone Ship Landing with Payload between 4000 kg and 6000 kg
- Total Number of Successful and Failure Mission Outcomes
- Boosters Carried Maximum Payload
- 2015 Launch Records
- Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

SQL For EDA

Build an Interactive Map with Folium

Created an Interactive map of Folium including

- Markers to show the different launch sites
- Circles to visually represent launch sites on map
- Marker clusters to group multiple launch sites together to avoid cluttering
- Added line to show distance from launch site to nearest coastline

Folium Map Analysis

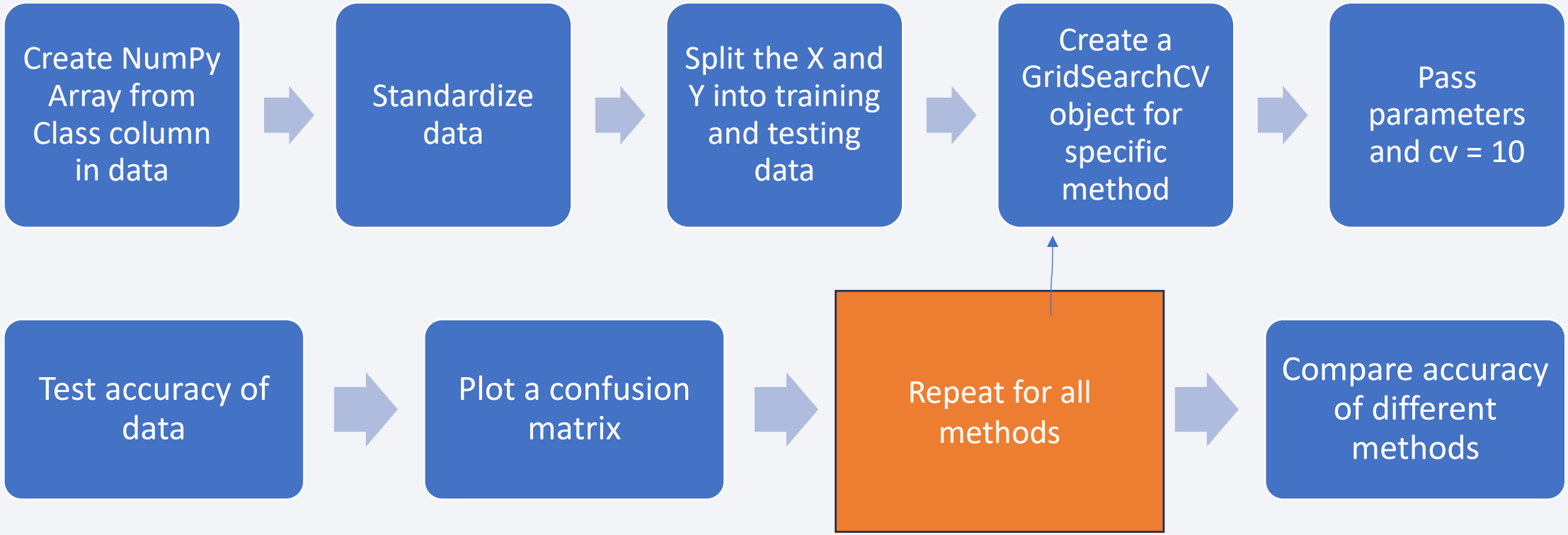
Build a Dashboard with Plotly Dash

Made a dashboard of

- Pie Chart for Total Success Launches for all sites
  - Shows which launch site has the highest success rate
- Pie Chart for Successful Launches for Launch Site with Highest Launch Success Ratio
  - Shows the statistics of the most successful Launch Site
- Catplot for Payload Mass vs. Launch Outcome & Payload
  - Helps find correlation between Payload Mass and mission success for different Launch sites

Plotly Dash Code

Predictive Analysis (Classification)



Machine Learning Predictive Analysis

Results

Exploratory data analysis results

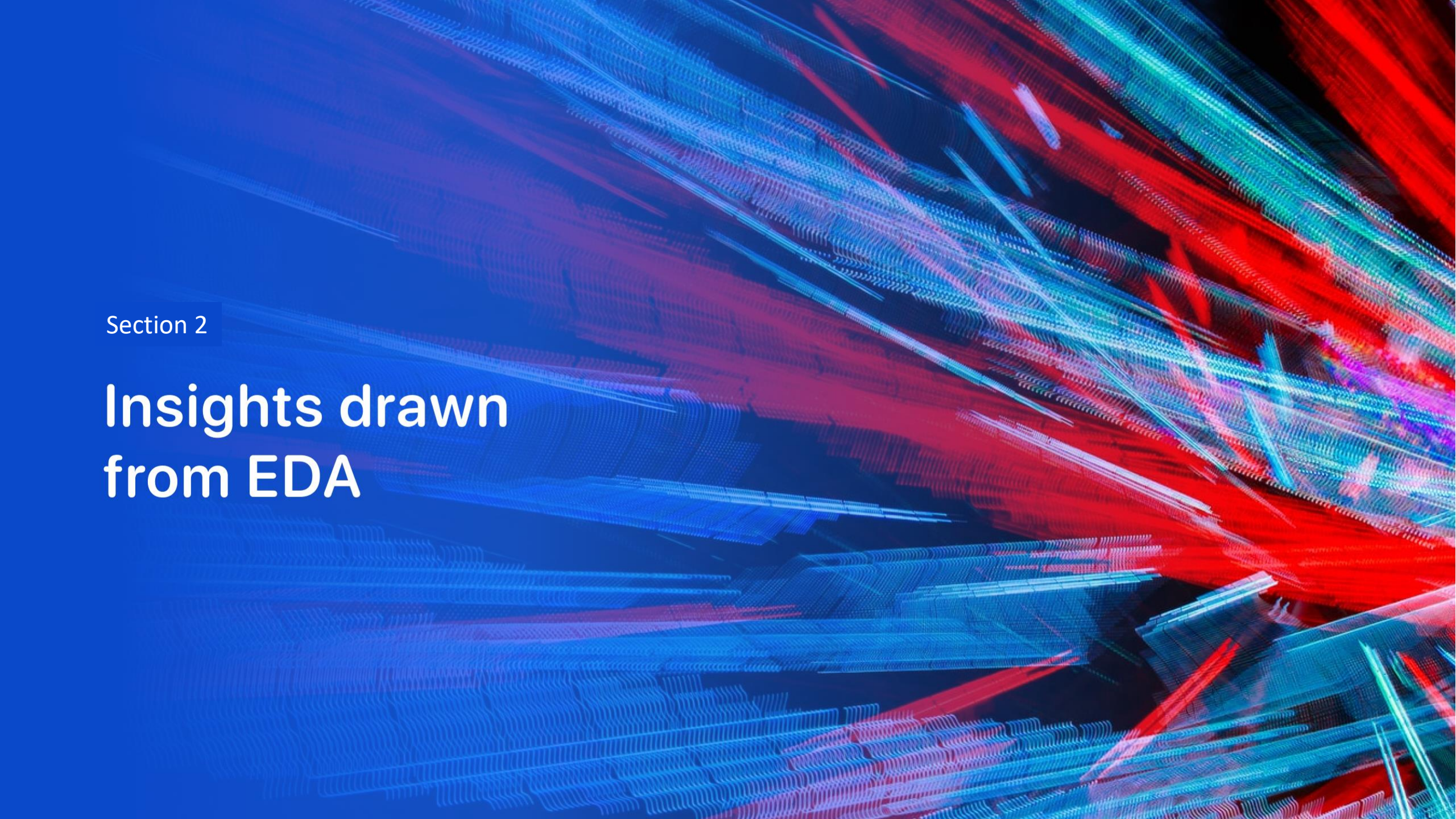
- 3 distinct launch sites
- ES-L1, GEO and HEO are among best orbits for mission success
- Increase in success rate from 2013-2020

Interactive analytics demo in screenshots

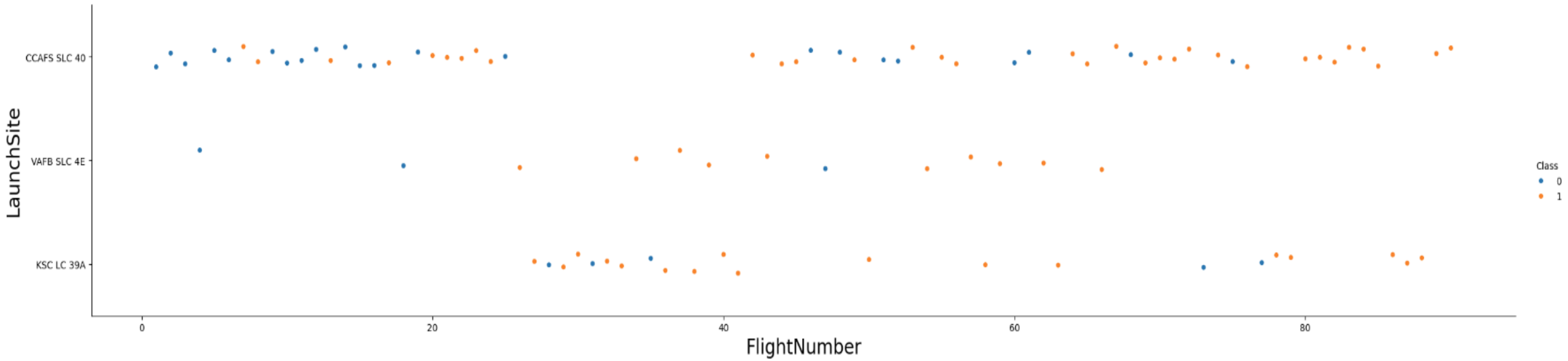
- KSC LC 39-A is the most successful Launch Site
- The 2000 kg to 4000 kg range for the Payload Mass has the highest success rate
- The FT Booster Version has the highest success rate

Predictive analysis results

- All predictive methods work equally well, with an 83.3% accuracy rate

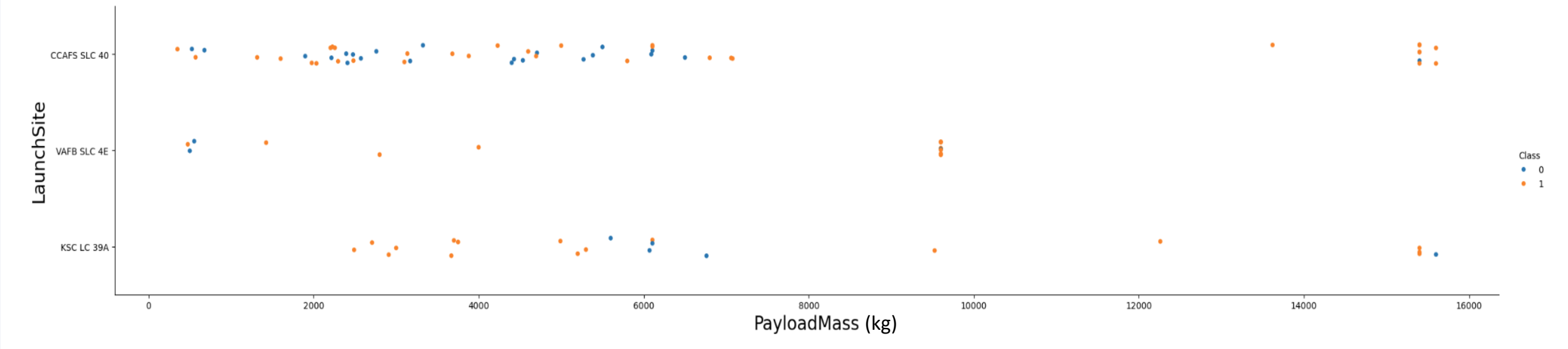

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Plot shows 3 distinct launch sites
- Most launches taken from CCA SLC 40 site
- No apparent correlation between Flight Number and mission success for CCAFS SLC 40 Site
- Fewest launches from VAFB SLC 4E site
- Few launches with a high Flight Number from VAFB SLC 4E site
- Increased success as Flight number increases for from VAFB SLC 4E site
- Few launches with low Flight Number from KSC LC 39A site
- No apparent correlation between mission Flight Number and mission success for KSC LC 29A site

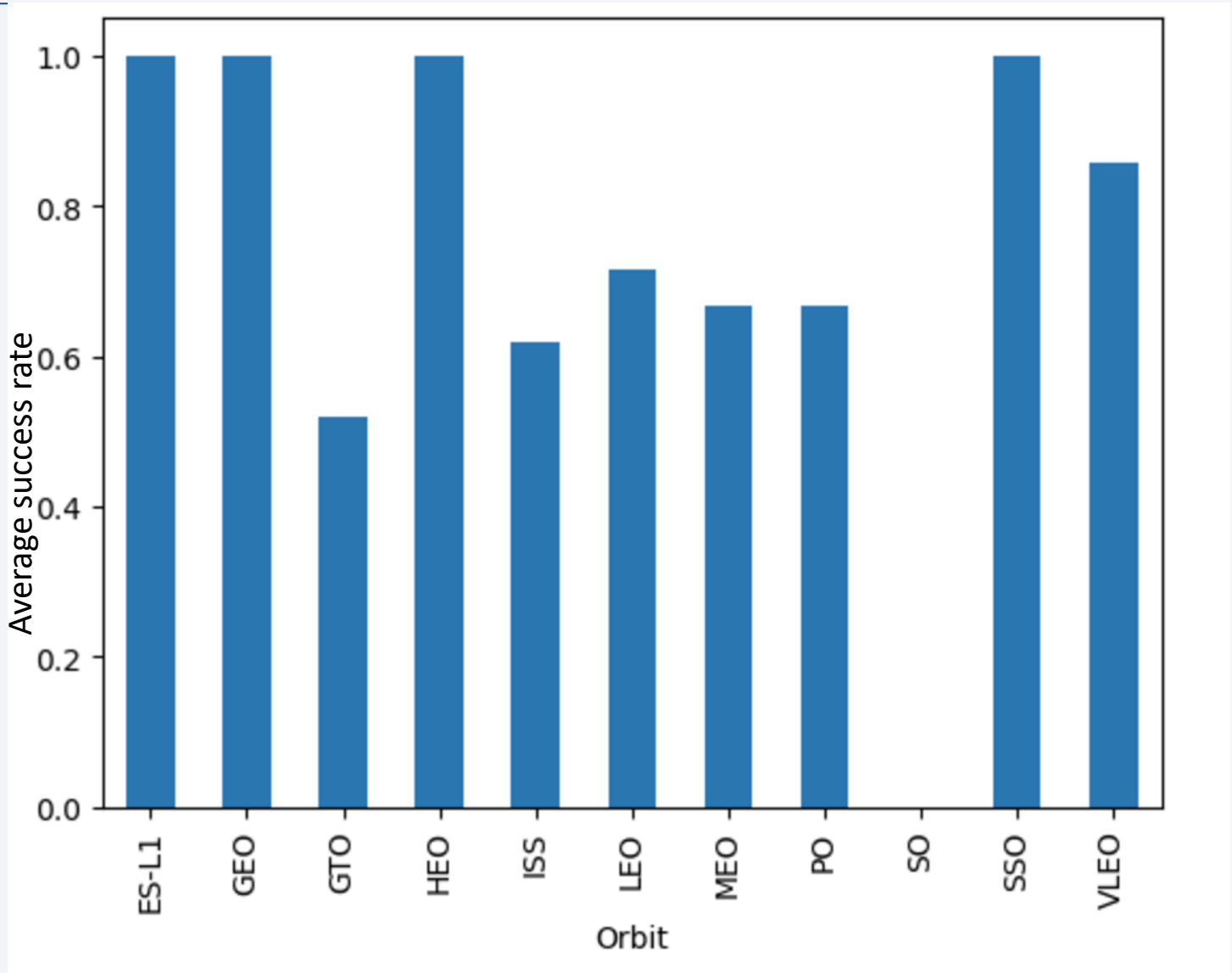
Payload vs. Launch Site



- No launch for Payload Mass from CCAFS SLC 40 site in the range from 8000 kg to 13000 kg
- Higher Payload Mass correlated with higher success rate for the CCAFS SLC 40
- No Payload Mass over 10000 kg launched in the VAFB SLC site
- Higher Payload Mass correlated with higher success rate for the VAFB SLC site
- Lower Payload Mass correlated with higher success rate for KSC LC 39A site

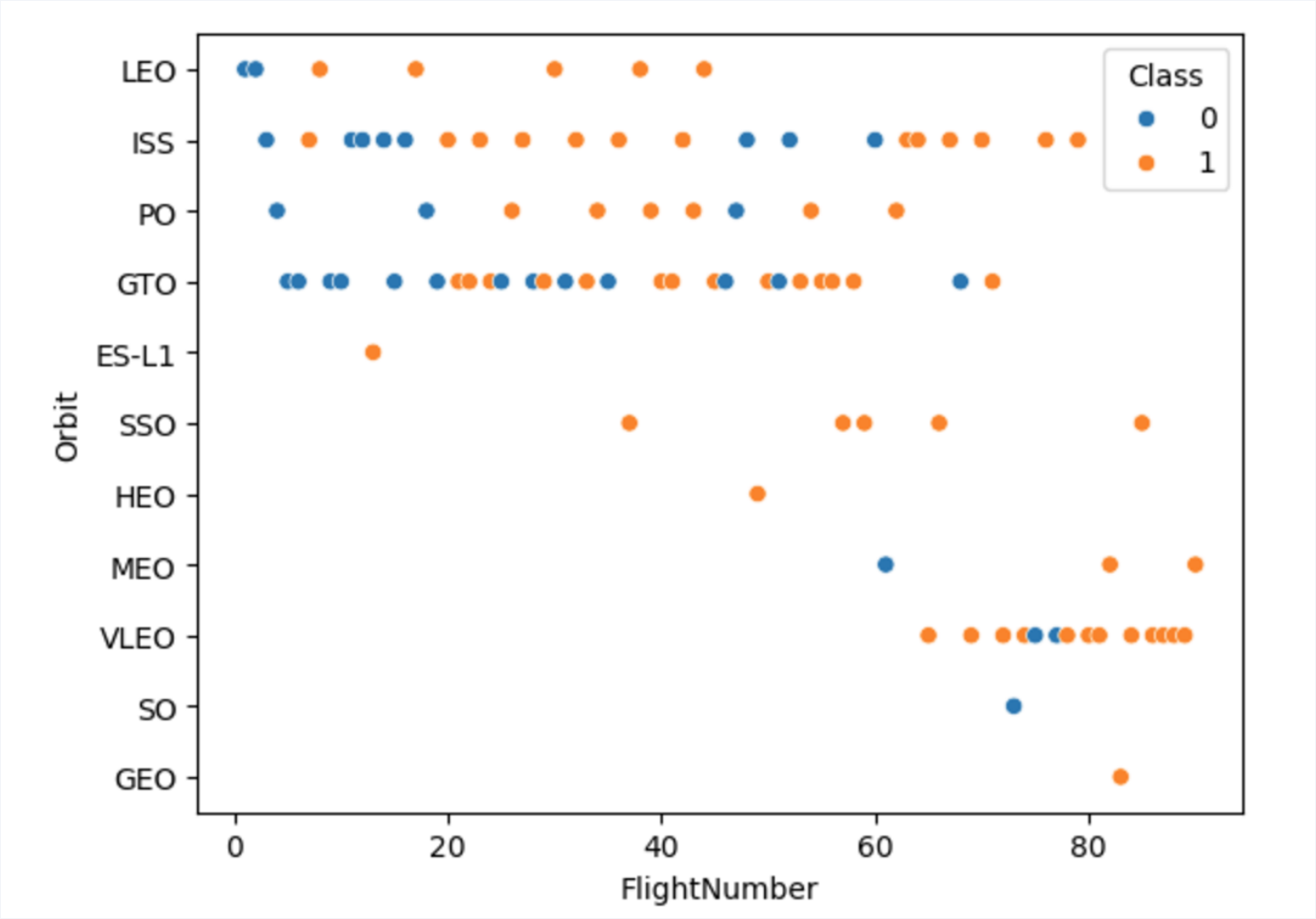
Success Rate vs. Orbit Type

- Orbits with **highest** average success rate are
  - ES-L1
  - GEO
  - HEO
  - SSO
- Orbits with **lowest** average success rate are
  - SO
  - GTO
  - ISS



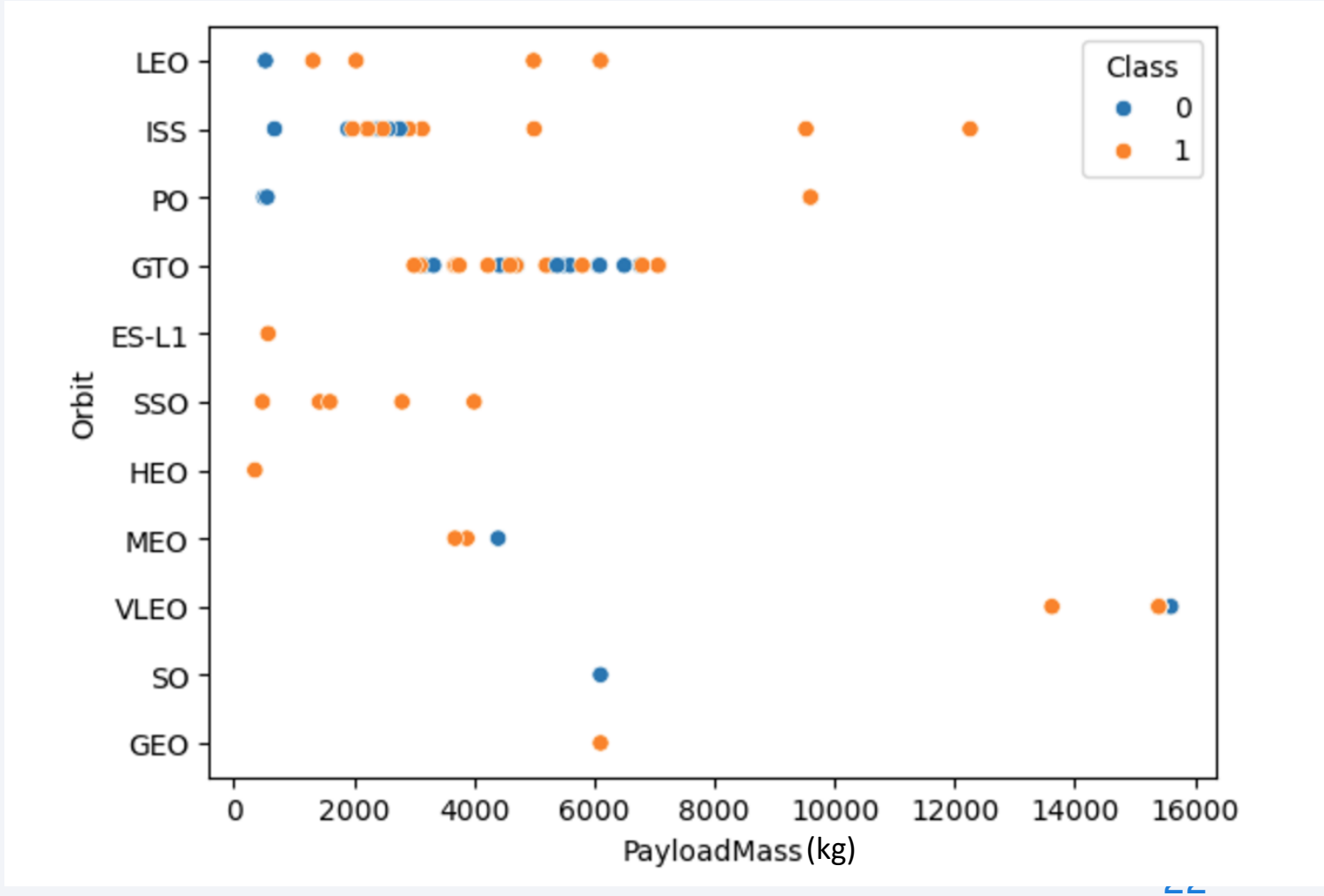
Flight Number vs. Orbit Type

- Higher Flight Number correlated with higher success in the LEO orbit
- SSO has a 100% success rate
- No correlation between Flight Number and success in GTO orbit
- Concentration of high Flight Numbers in VLEO orbit



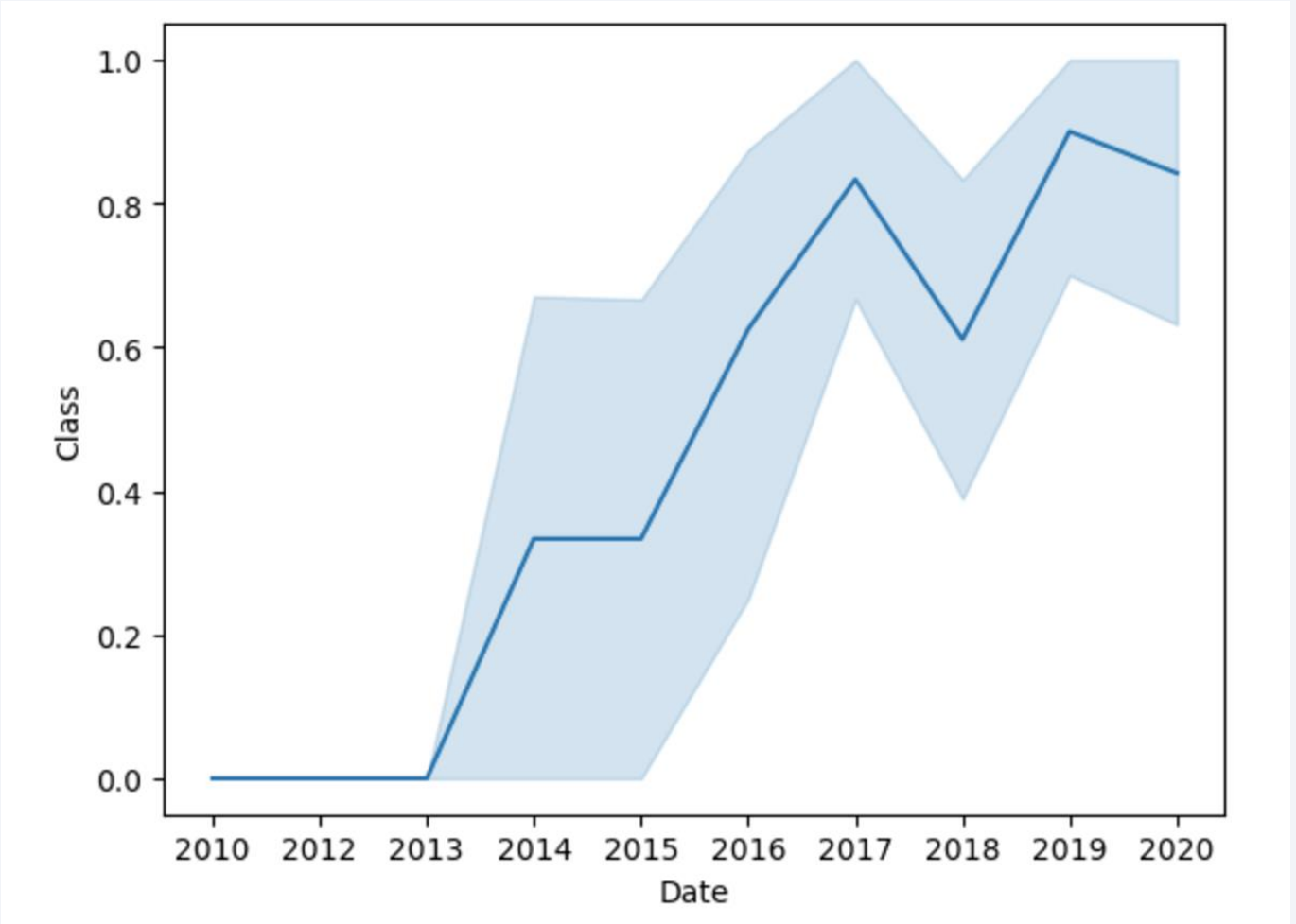
Payload vs. Orbit Type

- Higher Payload Mass correlated with higher success rate for PO, LEO and ISS orbits
- No apparent correlation between Payload Mass and success rate for GTO orbit



Launch Success Yearly Trend

We can observe that the success rate increased from 2013 until 2020



All Launch Site Names

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL
```

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

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

Query takes distinct values from the Launch_Site column, which gives the names of four unique Launch Sites

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL 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

Query finds five records with Launch Site names that begin with SQL using the LIKE operator, and the wildcard % in 'CCA%'. The output is Limit to 5

Total Payload Mass in NASA Launches

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass_KG FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'
```

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

Total_Payload_Mass_KG
45596

Query outputs the sum of the PAYLOAD_MASS_KG column where the boosters are launched by NASA (filtered using the WHERE operator)

The total mass is 45596 kg

Average Payload Mass by F9 v1.1

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

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

AVG(PAYLOAD_MASS__KG_)
2928.4

Query outputs the average of the PAYLOAD_MASS_KG column where the Booster Version is filtered by the WHERE operator to only include the F9 v1.1.

The average mass is 2928.4 kg

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Success (ground pad)'
```

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

MIN(Date)
2015-12-22

Query uses the MIN operator to find the first successful ground landing date

Uses the LIKE operator to find when mission was successful on a ground pad

The first date was 2015-12-22

Successful Drone Ship Landing with Payload between 4000 kg and 6000 kg

```
%sql SELECT DISTINCT Booster_Version, PAYLOAD_MASS__KG_, Landing_Outcome \
FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000 AND Landing_Outcome == 'Success (drone ship)'
```

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

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Query shows unique Booster Versions that have a Payload Mass between 4000 kg and 6000 kg filtered using the WHERE operator, and Landing Outcomes is success in a drone ship

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT COUNT(Mission_Outcome), Mission_Outcome FROM SPACEXTBL GROUP BY Mission_Outcome
```

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

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

Query uses the COUNT operator to count the number of mission outcomes that were successes and failures, using the GROUP BY operator to group by the mission outcome

Boosters Carried Maximum Payload

Query shows the different Booster Versions that carried the maximum payload.

The maximum payload was calculated using a subquery with the MAX operator

```
%sql SELECT Booster_Version,PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

```
%sql SELECT Date, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTBL \
    WHERE Landing_Outcome LIKE 'Failure%' AND substr(Date, 0, 5) = '2015'
```

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

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

Query shows the Date, Booster Version, Launch Site and Landing Outcome for all failures in 2015, using the LIKE operator with the magic % in 'Failure%' to find all the failures, and uses the substring substr(Date, 0, 5) to find the date 2015

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

- Use the COUNT function to count the instances of landing outcomes
- Use the BETWEEN operator to find dates between 2010-06-04 and 2017-03-20
- Group by Landing Outcome using the GROUP BY operator
- Order by descending using the ORDER BY … DESC operator, ordering from highest count to lowest

```
%sql SELECT COUNT(Landing_Outcome) AS Counts, Landing_Outcome FROM SPACEXTBL \
    WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Counts DESC
```

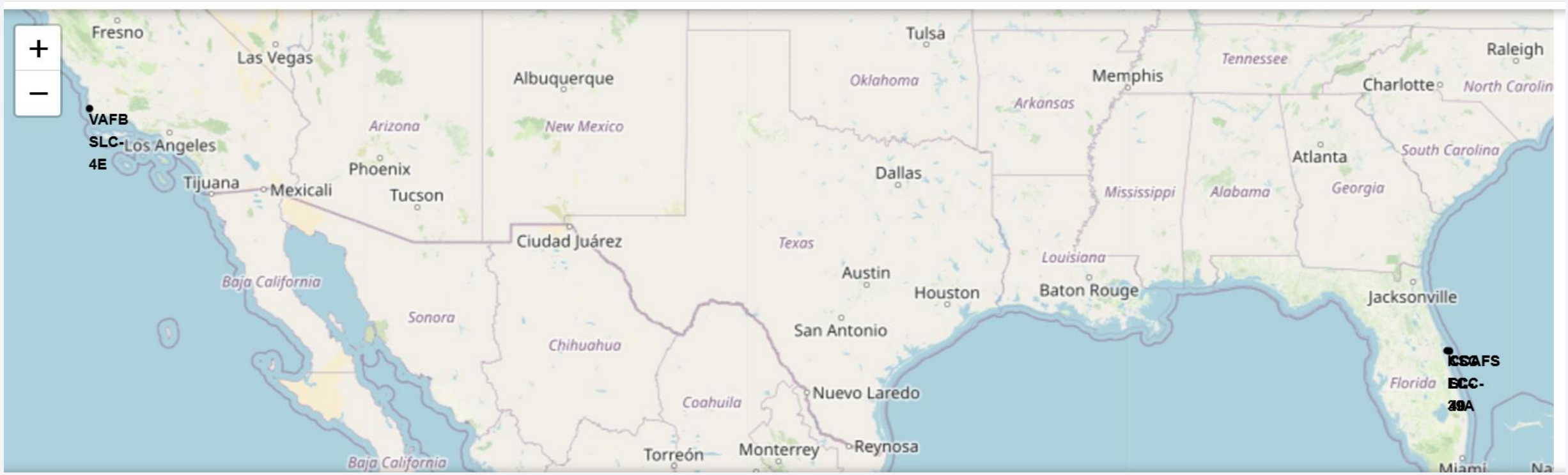
```
 * sqlite:///my_data1.db
Done.
```

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

Section 3

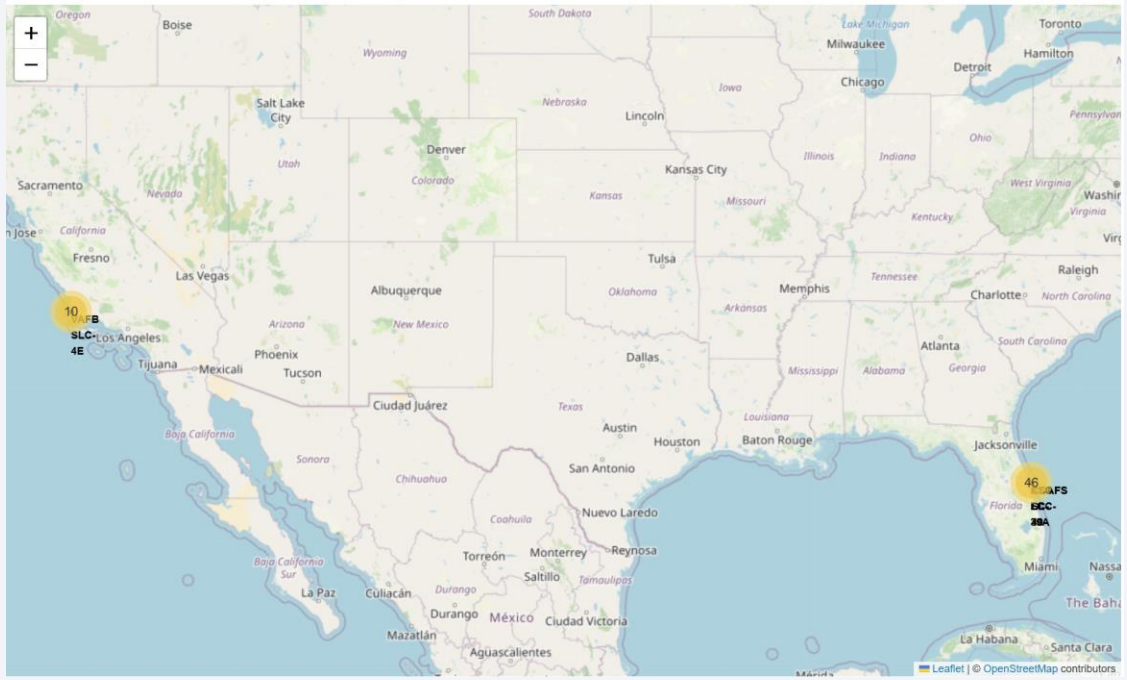
Launch Sites Proximities Analysis

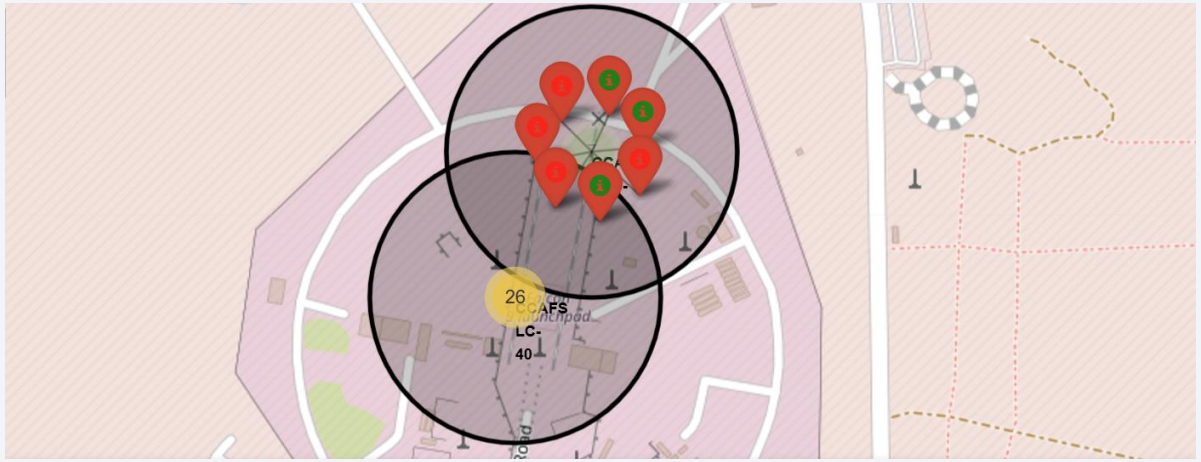
Launch Sites



- All the launch sites are close to a coastline
- The launch sites are in the southern part of the United States, which is the closest part of the country to the equator

Color-labeled Launch Outcomes



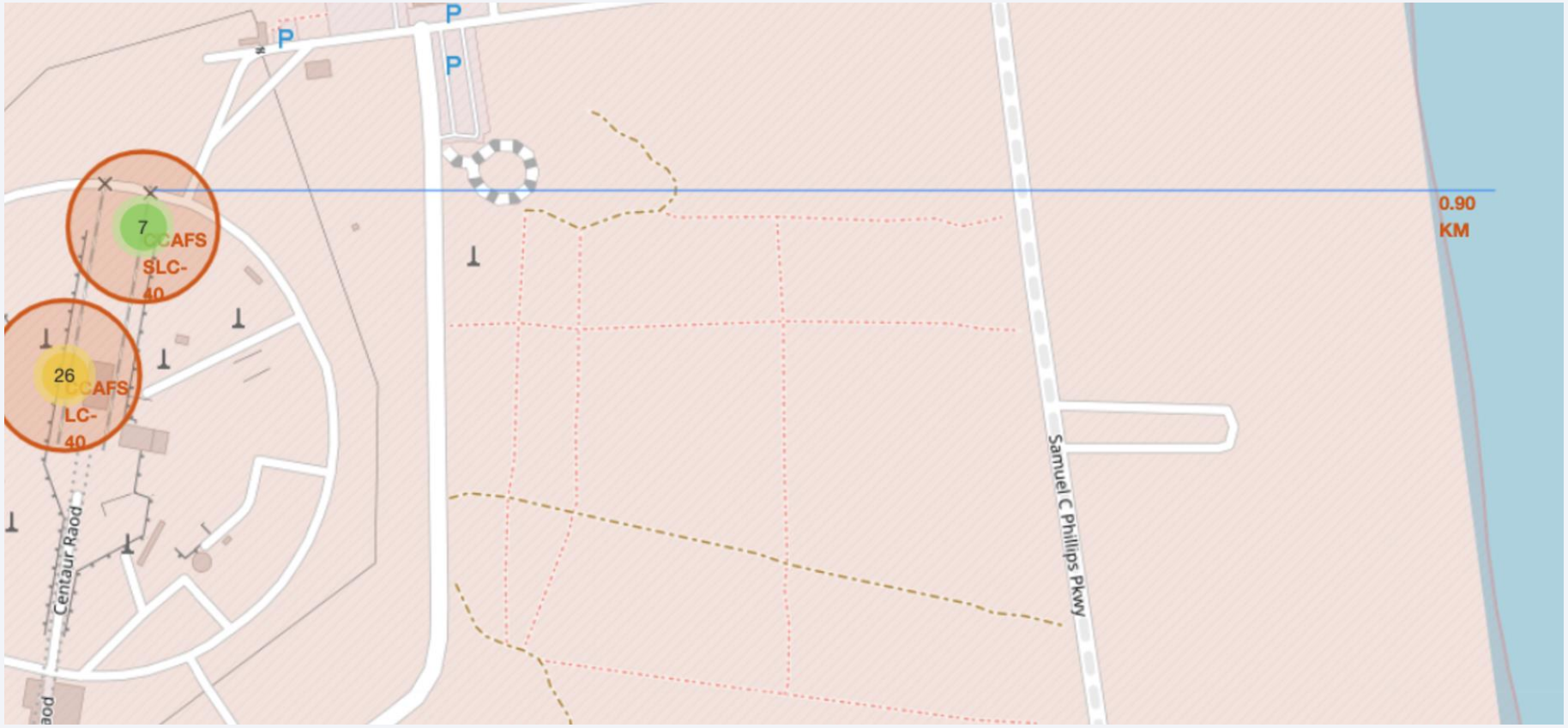


Visual of success at different launch sites is shown

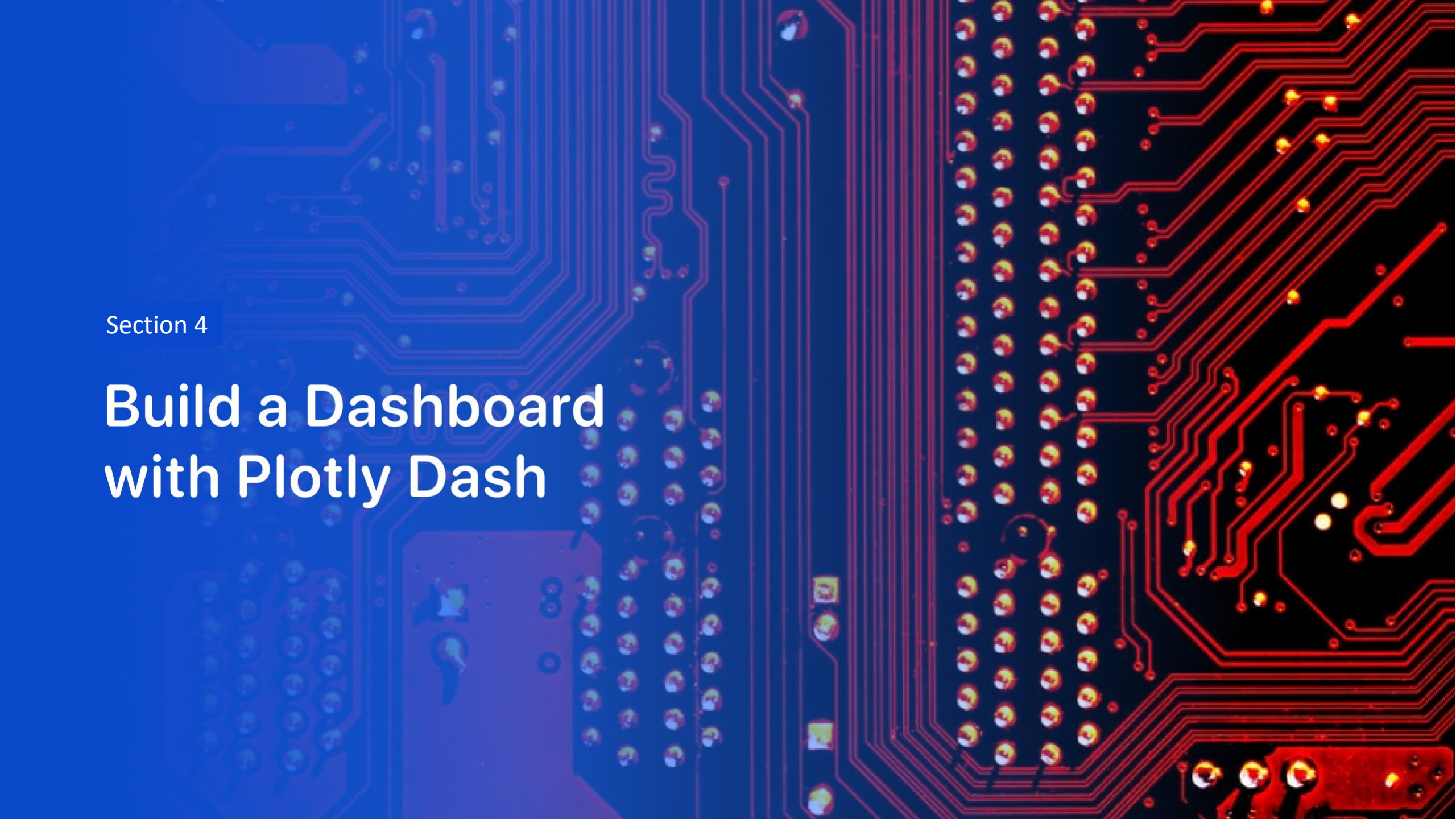
- 10 launches in the left-side of the country, and 46 launches on the right-side of the country

- Zooming into an area of the map, successes and failures at the location can be seen

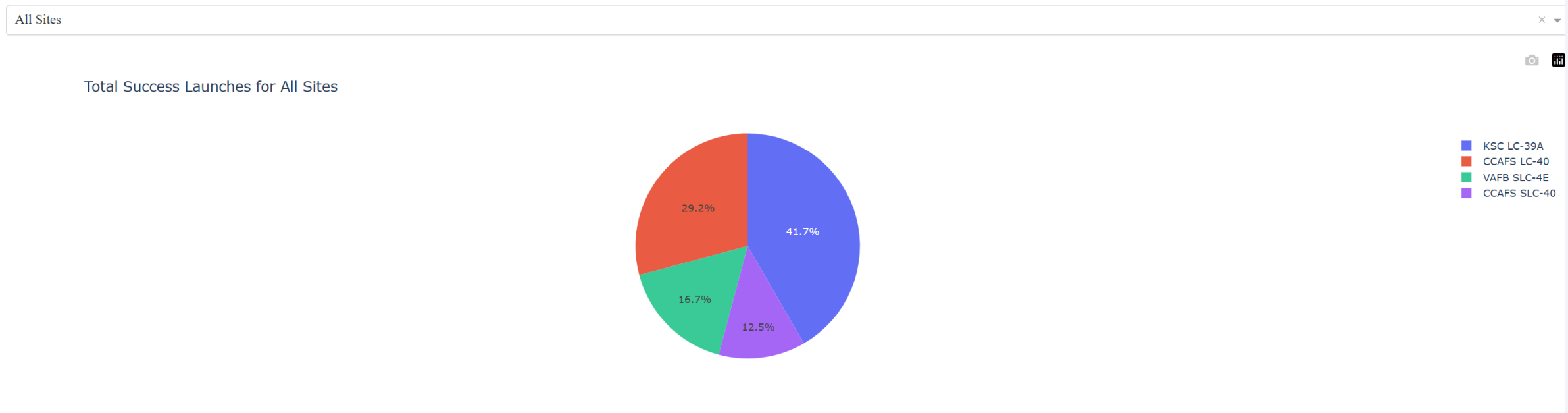
Proximity to Coastline



- Shows that the CCAFS SIC-40 is 0.90 km from the coastline


Section 4
Build a Dashboard with Plotly Dash

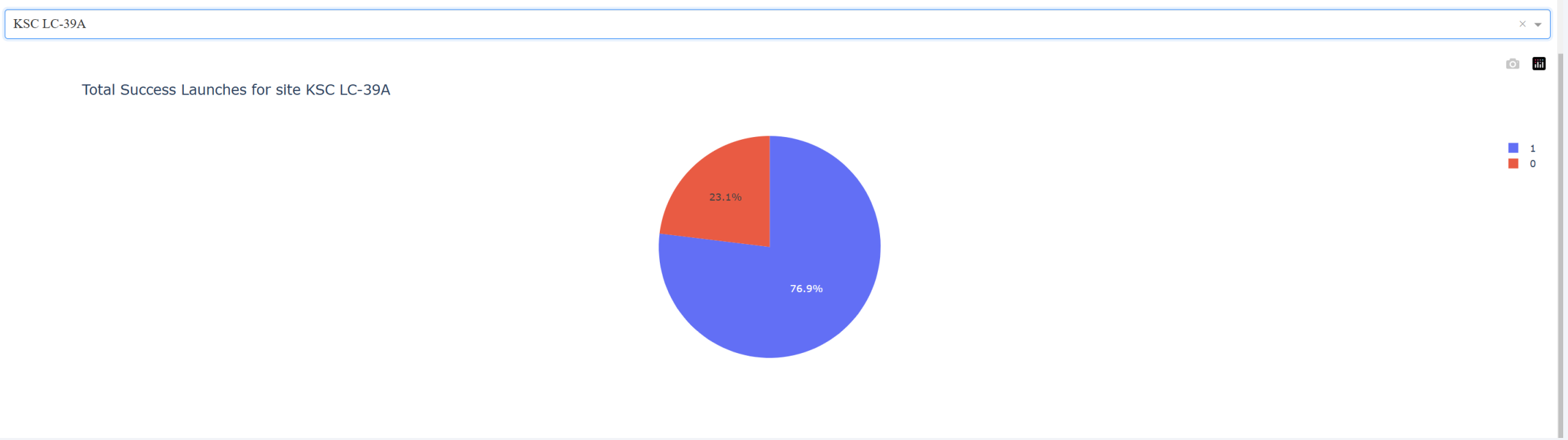
Pie Chart for Total Success Launches for all sites



Pie chart shows the percentage of successful launches by site

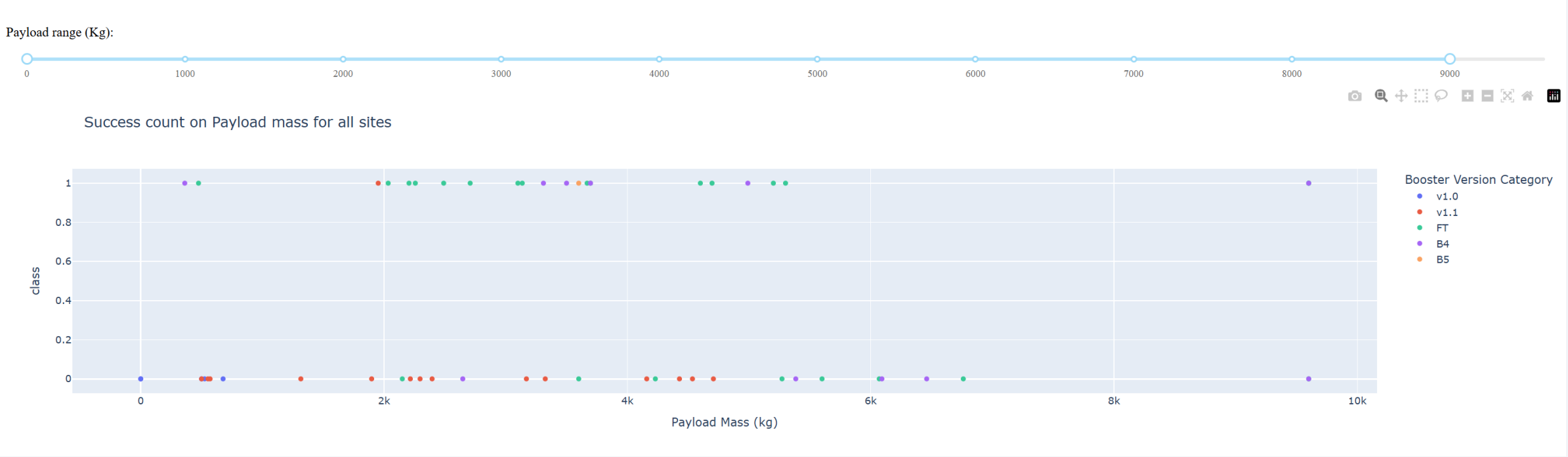
- KSC LC-39-A has the **highest** percentage of total successful launches at 41.7%
- VAFB SLC-4E has the **lowest** percentage of total successful launches at 16.7%

Pie Chart for Successful Launches for Launch Site with Highest Launch Success Ratio



- The Launch Site with the highest Launch Success Ratio was KSC LC -39A
- It had a success ratio of 76.9%

Catplot for Payload Mass vs. Launch Outcome & Payload



Scatter plot shows the range of Payload Masses vs. Success for all sites

- The 2000 kg to 4000 kg range for the Payload Mass has the highest success rate
- The FT Booster Version has the highest success rate

Section 5

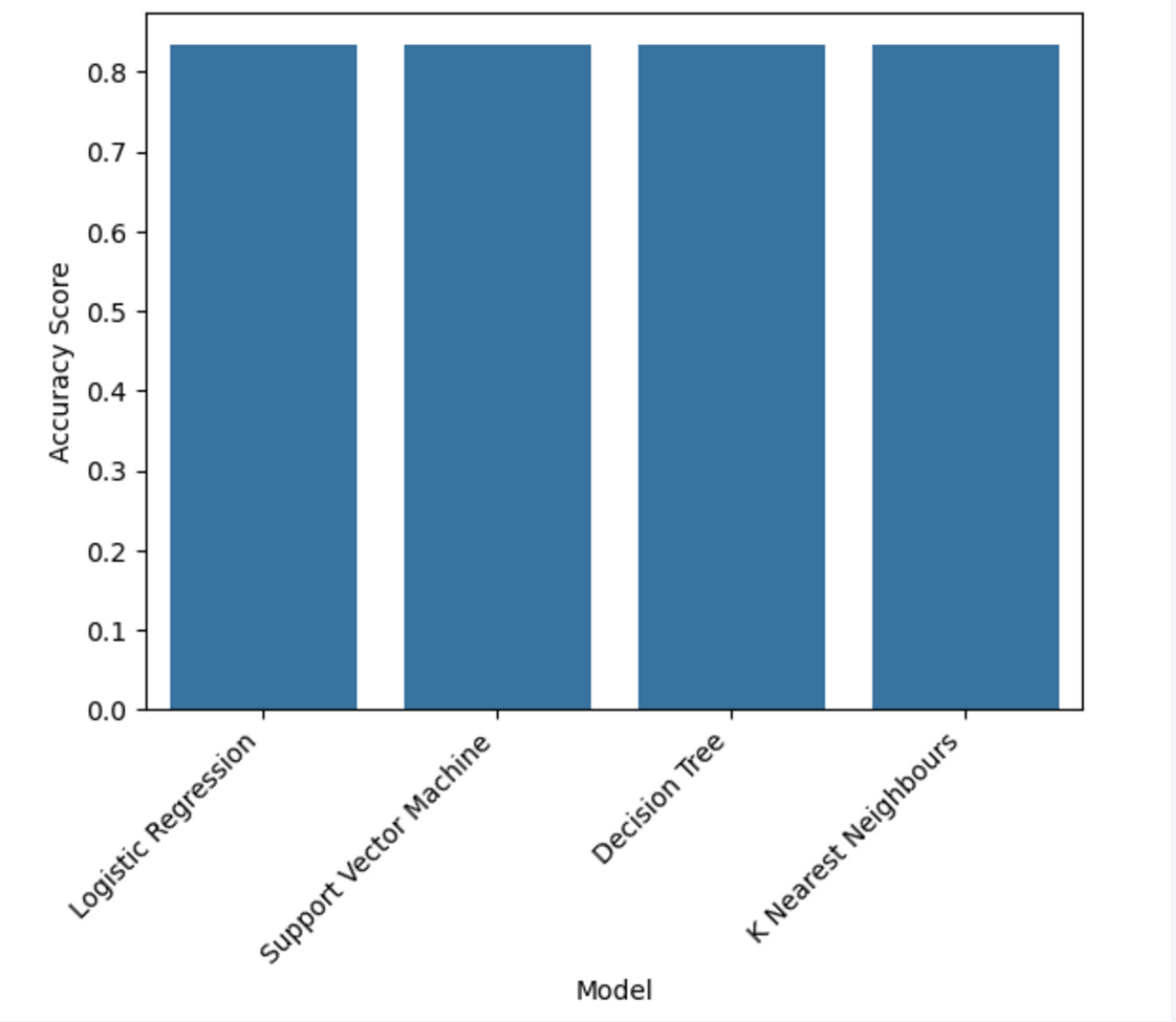
Predictive Analysis (Classification)

Classification Accuracy

Tested the accuracy of four different models

- Logistic Regression
- Support Vector Machine
- Decision Tree
- K Nearest Neighbors

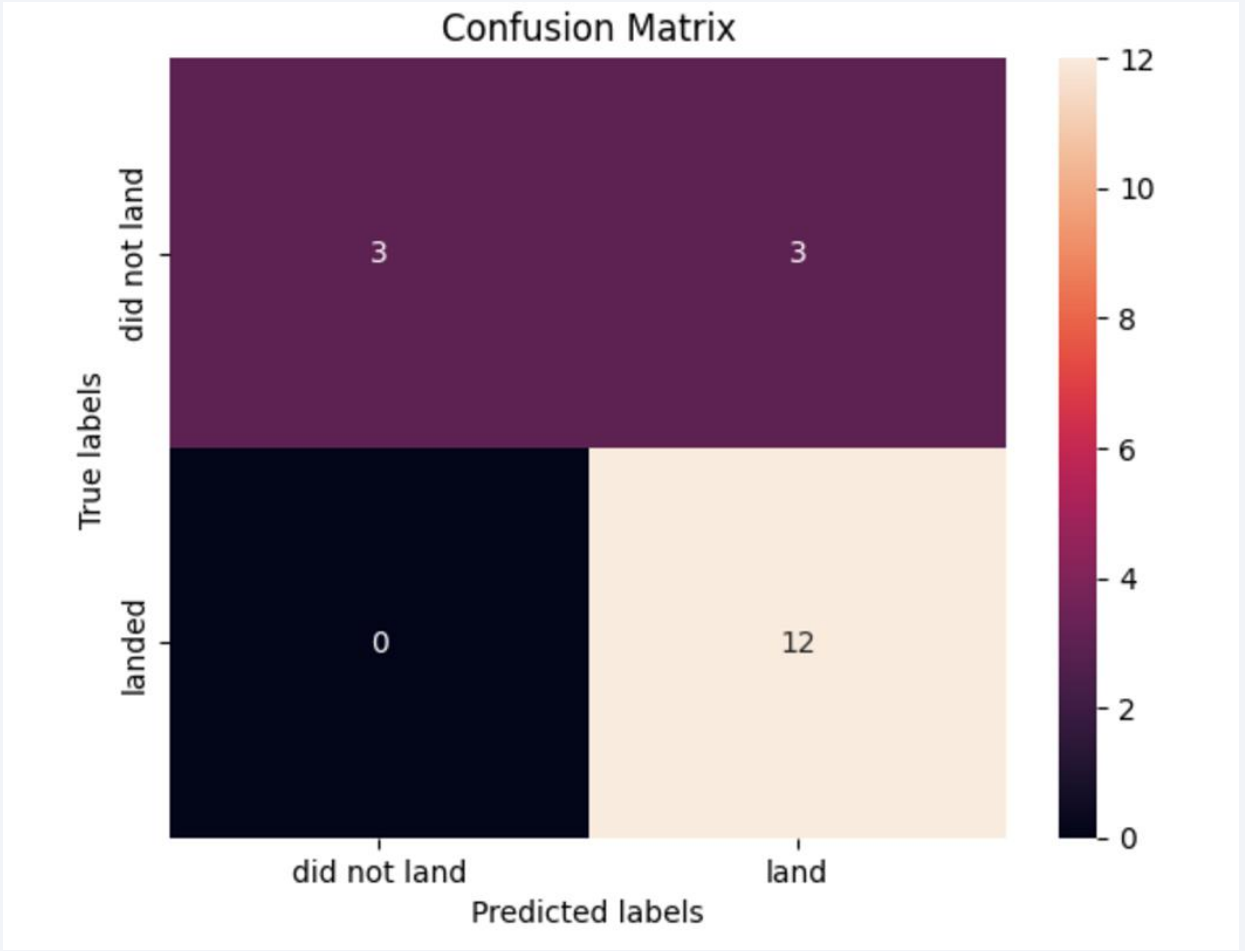
They all have the same accuracy of 0.833%



Confusion Matrix

This is the Confusion Matrix for all the models

- 12 True Positives
- 3 False Positives
- 3 False Negatives
- 0 True Negatives



Conclusions

- The success of SpaceX missions has increased, with significant growth from 2013-2020
- Factors that contribute to a successful mission include
  - Orbit type (ES-L1, GEO, HEO)
  - Payload Mass range (2000 kg- 4000 kg)
  - Booster Version (FT)
  - Launch Site (KSC LC 39-A)
- Most launch sites are found near coastlines, and placed in the southern regions of the United States, which are close to the equator than other regions in the country
- All predictive modelling methods have same accuracy of 83.3%

Appendix

Collection of Notebooks used for this Project

Thank you!