IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Thomas Steele
July 4, 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

The following methodologies were used to analyze and draw inferences from the data:

- Data Collection using Web Scraping and extracting information from the SpaceX API
- Exploratory Data Analysis (EDA), including data wrangling, SQL queries and data refinement for analysis, data visualization and interactive visual analytics
- Machine Learning predictive analysis.

To better focus competitive pricing, a prediction model has been developed to determine if the first stage landing will be successful, based on key variables including:

Payload	**Flight Number**	**Legs**
Orbit	**Grid Fins**	**LandingPad**
Launch Site	**Booster Version**	**Reused Count**

Introduction

Project Background and Context:

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. If it can be determined that the first stage will land successfully, the cost of a launch can be competitively priced if there are alternative companies bidding against SpaceX for a given payload launch.

Problems to solve:

- Can the successful first stage return landing be predicted accurately and repeatably?
- What parameters or features are most important in predicting the first stage successful landing?

Section 1

Methodology

Methodology - 1

Summary of Methodologies

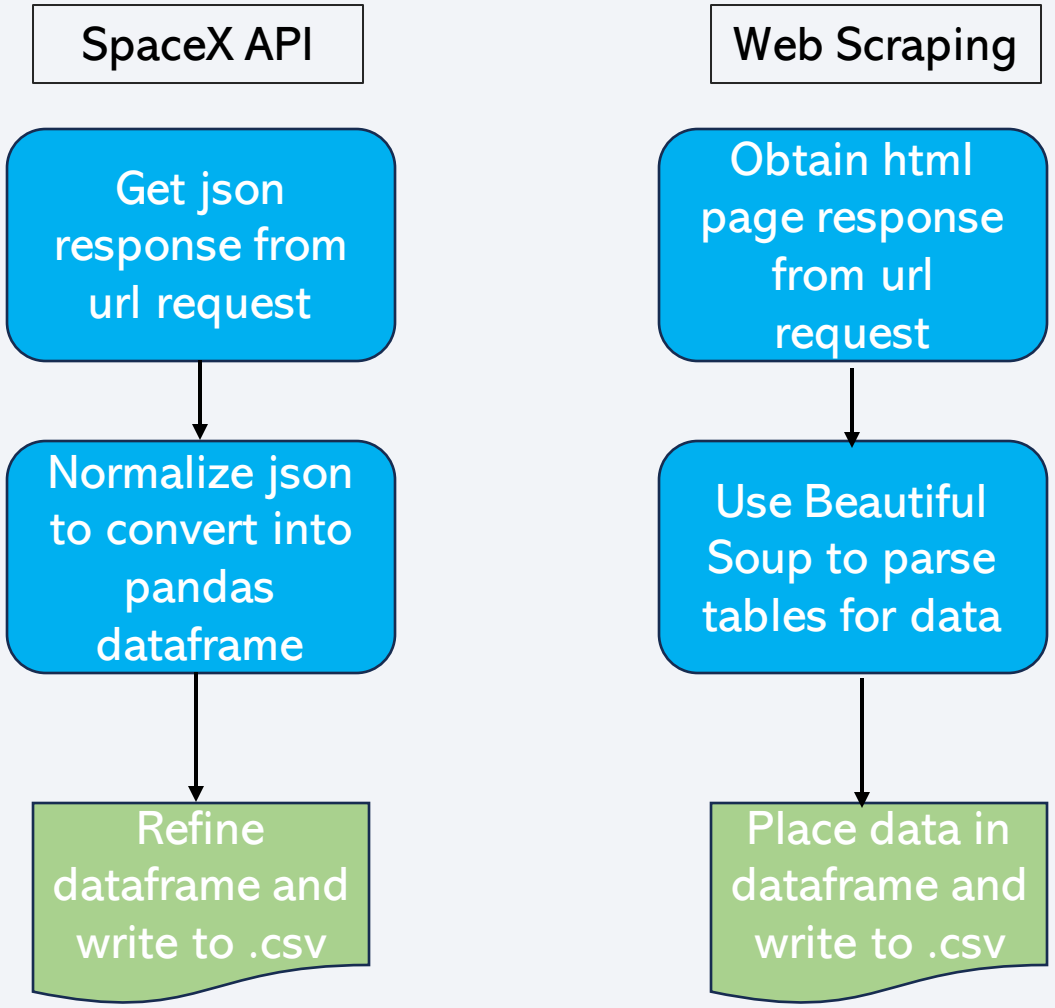
- Data collection:
  - The public SpaceX API was used to gather rocket launch information
  - Web scraping was used to acquire Falcon heavy launch records from Wikipedia
- Data Wrangling:
  - Missing data replaced with mean value imputation
  - Calculations of key metrics such as:
    - number of launches on each site
    - number and occurrences of each orbit
    - number and occurrences of mission outcome per orbit type
  - Simplified landing outcomes with a Class label to identify the dependent target variable

Methodology - 2

- Exploratory Data Analysis (EDA) using visualization and SQL:
  - Visualizing relationships between launch site, payload, flight number, and orbit
  - One Hot encoding for important categorical values needed for prediction
- Interactive visual analytics using Folium and Plotly Dash:
  - Visualizing Success/Failure of Launch Sites and Payloads
- Predictive analysis using classification models:
  - The classifier with the best score is determined
  - Classifiers tested include:
    - Support Vector Machine
    - Linear Regression
    - Decision Tree
    - K Nearest Neighbor
    - Gradient Boosting

Data Collection

- The public SpaceX API was used to gather rocket launch information from: https://api.spacexdata.com/v4/launches/past. An http get request to this url will return a large json response content to be processed.
- Web scraping was used to acquire Falcon heavy launch records from Wikipedia at: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches. An http get request to this url will return raw html page content, where tables need to be parsed and processed.



Data Collection – SpaceX API

- SpaceX API REST calls data collection flowchart

- For details, see the Notebook at GitHub URL:

https://github.com/gi1no/Capstone-Project/blob/Master/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – Web Scraping

- Wikipedia Web Scraping data collection flowchart

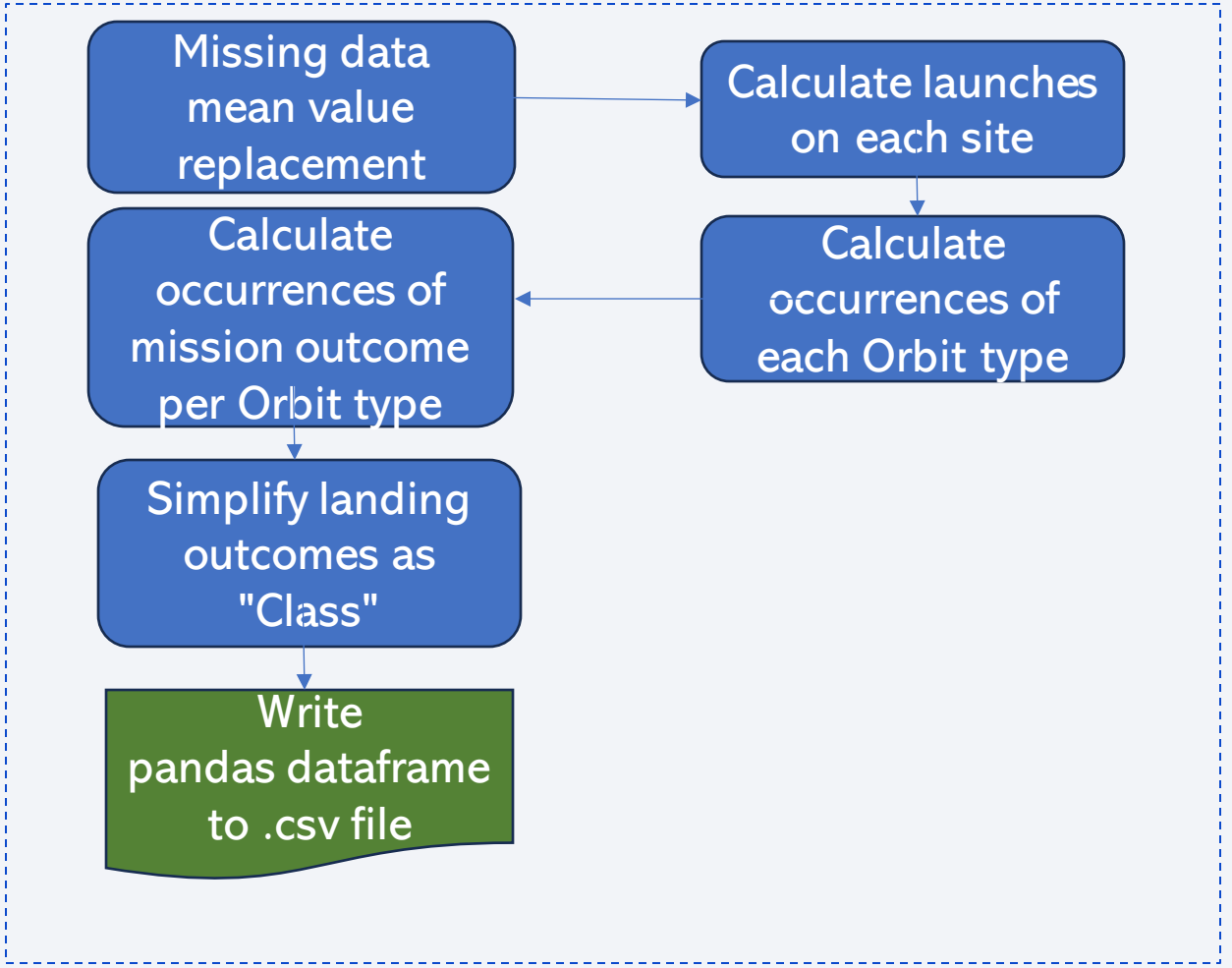
- For details, see the Notebook at GitHub URL:

https://github.com/gi1no/Capstone-Project/blob/Master/jupyter-labs-webscraping.ipynb

Data Wrangling

- Missing data replaced with mean value imputation
- Calculations of key metrics such as:
  - number of launches on each site
  - number and occurrences of each orbit
  - number and occurrences of mission outcome per orbit type
- Simplified landing outcomes with a Class label to identify the dependent target variable
- For details, see the Notebook at GitHub URL:

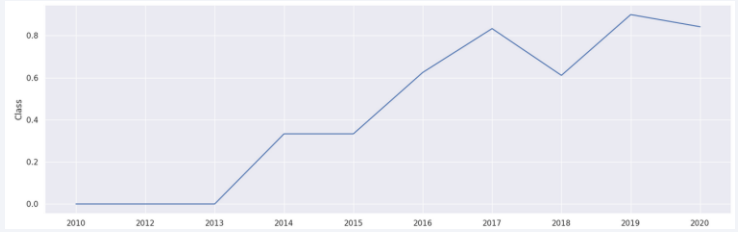
https://github.com/gi1no/Capstone-Project/blob/Master/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

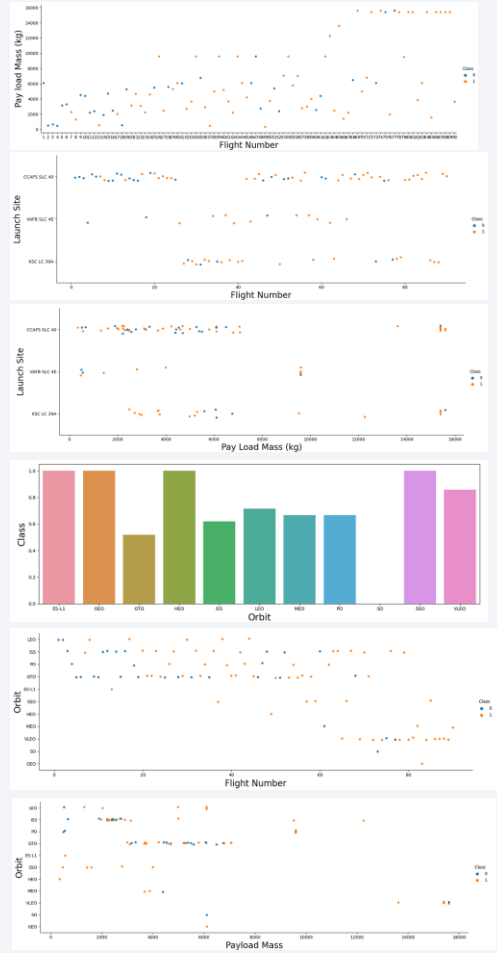


EDA with Data Visualization

Summary of charts and graphs plotted:

- Scatter Plot of Flight Number vs Payload to determine effects of Payload on Launch Outcome (Class).
- Scatter Plot of Flight Number vs Launch Site to determine effects of site on Launch Outcome (Class).
- Scatter Plot of Payload vs Launch Site to determine effects of each on Launch Outcome (Class).
- Bar Plot of Orbit vs Class to determine if Orbit has any effect on Launch Outcome
- Scatter Plot of Flight Number vs Orbit to determine effects of each on Launch Outcome
- Scatter Plot of Payload vs Orbit to determine effects of each on Launch Outcome
- Launch Success per year for reference





GitHub URL: https://github.com/gi1no/Capstone-Project/blob/Master/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Retrieved the four unique names of the launch sites
- Displayed 5 records where launch sites began with the string 'CCA'
- Displayed the total payload mass carried by boosters launched by NASA (CRS)
- Displayed average payload mass carried by booster version F9 v1.1
- Listed the date of the first successful landing outcome in ground pad
- Listed the names of the boosters which had landing outcome "Success in drone ship" and have payload mass greater than 4000 but less than 6000
- Listed the total number of successful and failed mission outcomes
- Listed the names of the booster_versions which have carried the maximum payload mass.
- Listed the records which will display the month names, failed landing_outcomes in drone ship, booster versions, & launch_site for the months in the year 2015
- Ranked the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20, in descending order
- GitHub URL: https://github.com/gi1no/Capstone-Project/blob/Master/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Build an Interactive Map with Folium

- Circles for all four launch sites were added to the map to easily zoom in to each site
- Markers were added to each site to show, at a glance, the total number of launches
- The Marker Clusters were added to each site and color coded to indicate success (green) or failure (red) of launch outcomes. This provided a visual cue as to which sites were more successful
- The distance between a launch site and example surrounding landmarks (highway, city, etc.) were indicated and a line drawn to highlight the location, direction, and distance
- GitHub URL: https://github.com/gi1no/Capstone-Project/blob/Master/lab_jupyter_launch_site_location.jupyterlite.ipynb

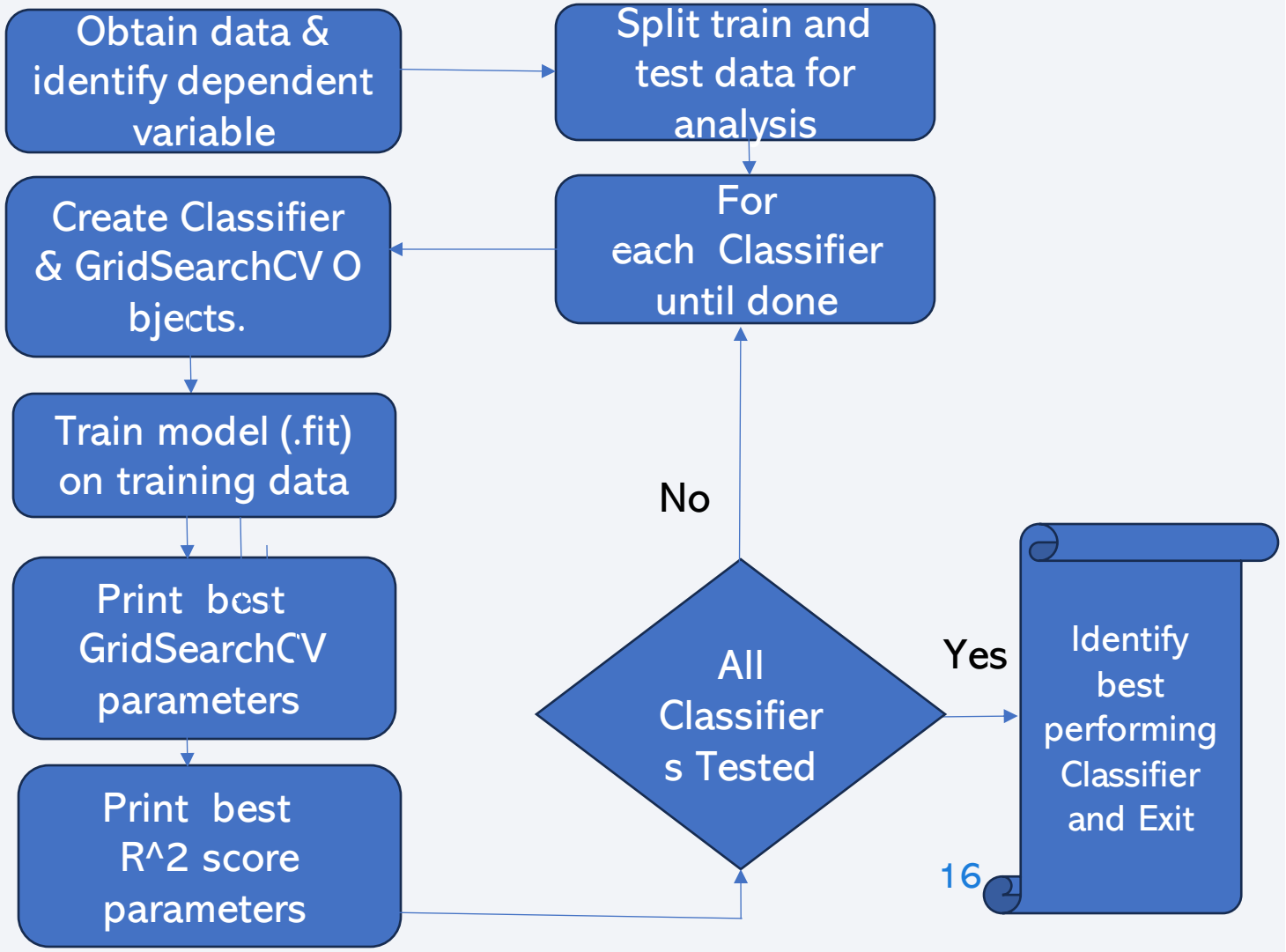
Build a Dashboard with Plotly Dash

- A pie chart showing the successful launches for all sites provides a concise visual comparison of launch performance
- A drill-down to each launch site reveals a pie chart showing the specific success and failure rate for the selected launch site
- An interactive Scatter Plot of Launch Outcome vs Payload for each Booster version allowed for the slider bar selection of payload ranges to reveal relationships among these potential launch success parameters.

- GitHub URL: https://github.com/gi1no/Capstone-Project/blob/Master/spacex_dash_app.py

Predictive Analysis (Classification)

- Split the data into train and test sets
- For each of the following classifiers, use GridSeach to find best performance parameters and report R^2 score:
  - Linear Regression
  - SVC - Support Vector Machine
  - DecisionTreeClassifier
  - KNeighborsClassifier
  - GradientBoostingClassifier
- GitHub URL: https://github.com/gi1no/Capstone-Project/blob/Master/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

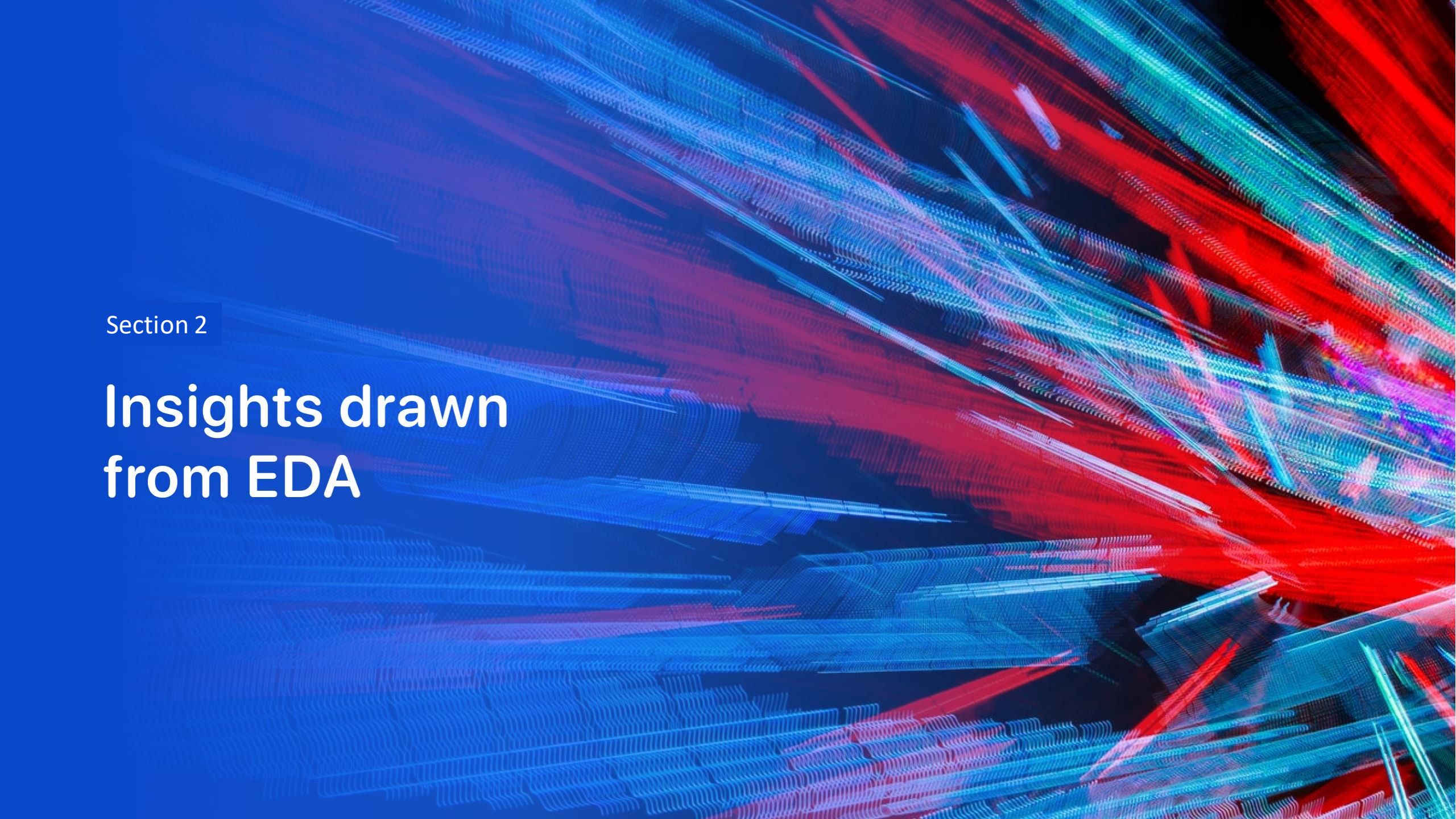


Results

- Exploratory data analysis results:
  - SSO is the only real Orbit with a perfect record on successful launches (5). While Orbits HEO, GEO, and ES-L1 are also perfect, they have only (1) data point
  - Launch site KCS LC 39A has good launch outcomes, except for payloads around 6,000 kg.
  - Since 2013, launch success has generally continued to increase
- Interactive analytics demo in screenshots results:
  - Launch site KCS LC 39A had the best launch outcome
  - For low Payloads, The FT booser version does well (green). But the v1.1 does not (red)
- Predictive analysis results:
  - The Decision Tree classifier was the best performing predictive model. It had an 89% fit to the test/validation data after training on an independent set of data. It can be used to predict a successful launch outcome based on several key parameters

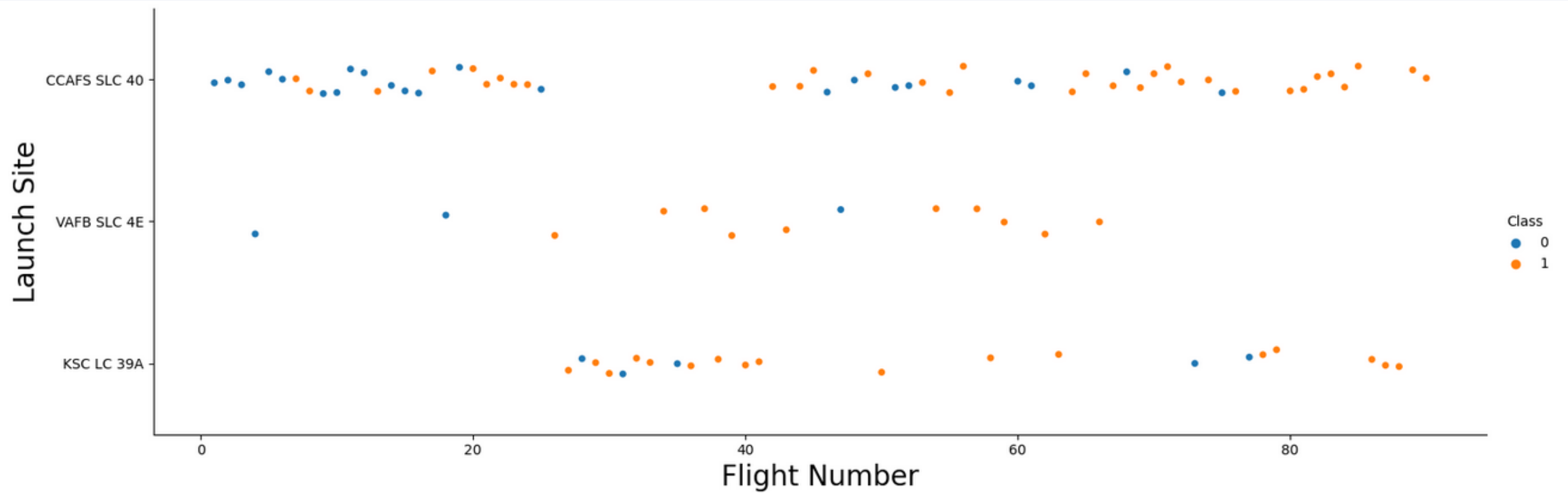
Section 2

Insights drawn from EDA



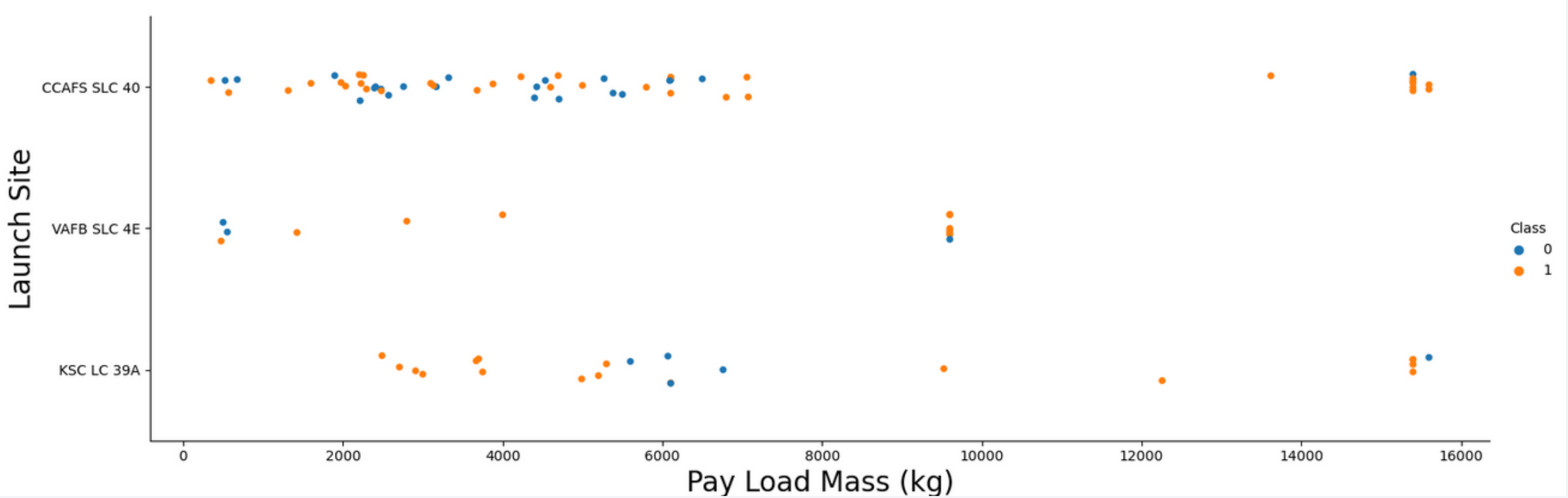
Flight Number vs. Launch Site

- As the number of flights increase the number of successes increases for each launch site.
- CCAFS SLC 40 has the most launches



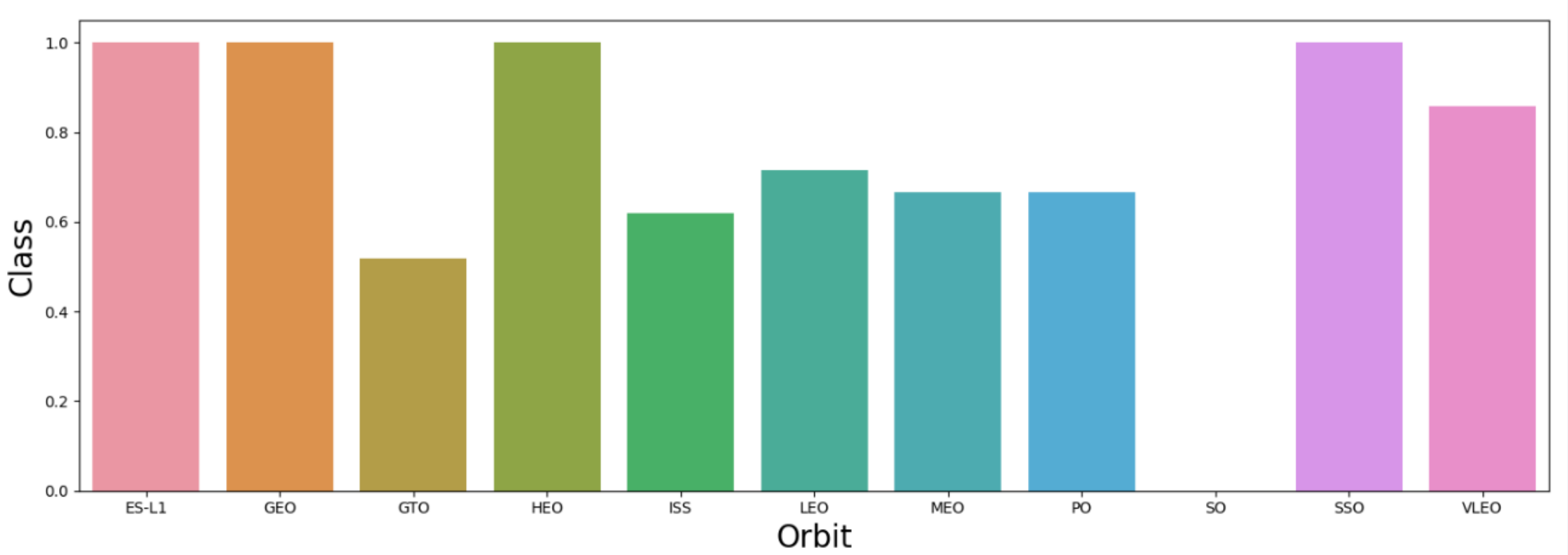
Payload vs. Launch Site

- KSC LC 39A has great success with payloads above and below 6,000 kg
- CCAFSSLC 40 has good success with payloads over 6,000 kg
- VAFB SLC 4E shows promise, more data needed



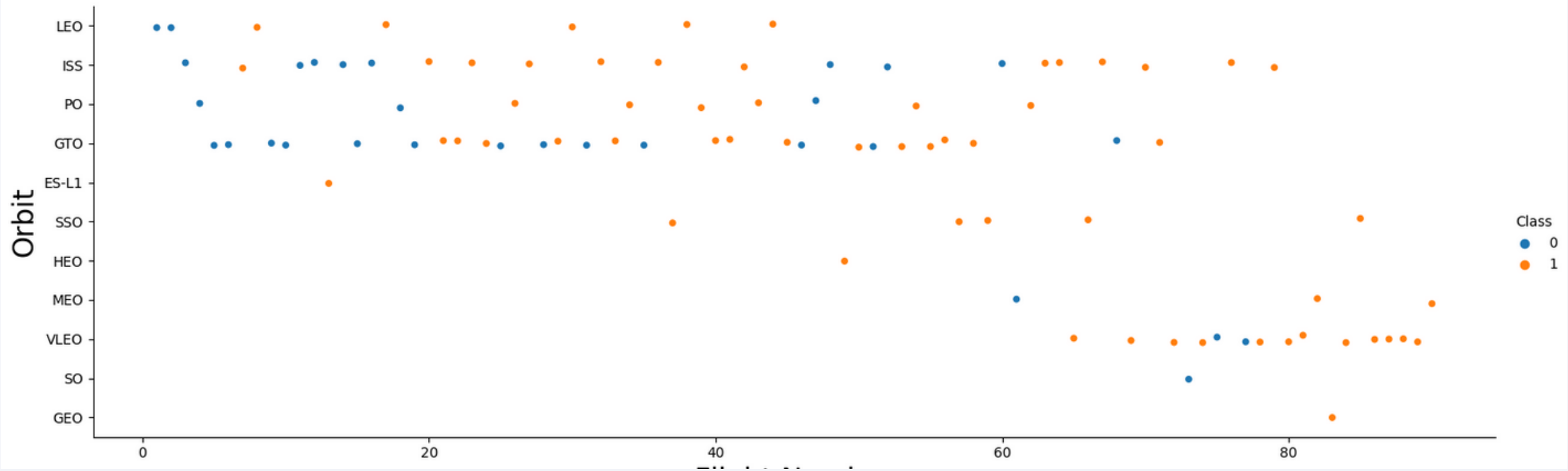
Success Rate vs. Orbit Type

- Orbit types ES-L1, GEO, HEO, and SSO have the highest success rate of stage 1 return
- Consider having customers who want a GTO or ISS orbit, pay a premium?



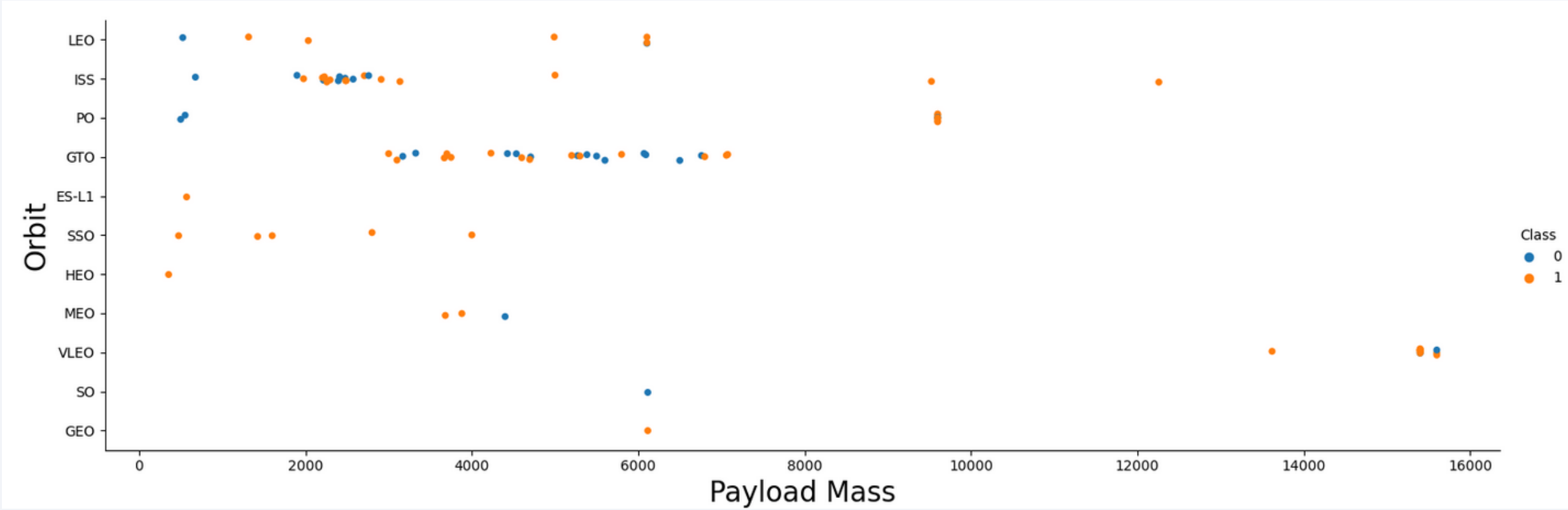
Flight Number vs. Orbit Type

- Although the LEO orbit seems to have great success with increased Flight Numbers, Orbit type seems to have no correlation to Flight Number for the most part.



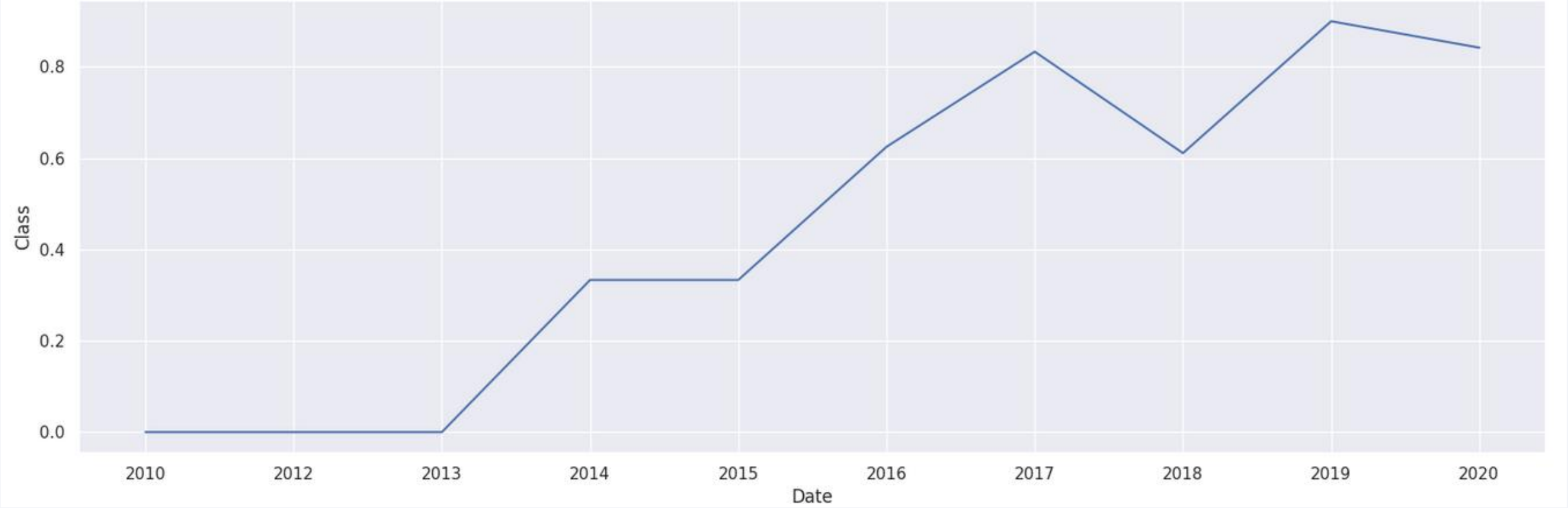
Payload vs. Orbit Type

- LEO, ISS and PO have great success with heavier payloads
- The GTO orbit has inconsistent success in the 2,000 – 8,000 kg. range



Launch Success Yearly Trend

- Although there was a slight dip in 2018, from 2013 through 2020 there has been a steady increase in the stage 1 return success rate.



All Launch Site Names

- The unique (i.e., discinct) launch site names were selected from the table holding the dataset and information that was collected

Display the names of the unique launch sites in the space mission

```
%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

Launch Site Names Begin with 'CCA'

- The first 5 launch site names that started with 'CCA' were selected from the table holding the dataset and information that was collected. The operative query phrase "… LIKE 'CCA%' LIMIT 5;" filtered what was required.

Display 5 records where launch sites begin with the string '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
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The sum of all "PAYLOAD_MASS_KG_" entries for customer 'NASA (CRS)' (45,596 kg.) was totaled from the table holding the dataset

Display the total payload mass carried by boosters launched by NASA (CRS)

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

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

SUM("PAYLOAD_MASS__KG_")
45596.0

Average Payload Mass by F9 v1.1

- The average of all "PAYLOAD_MASS_KG_" entries for booster 'F9 v1.1' (2,928.4 kg.) was calculated from the table holding the dataset

Display average payload mass carried by booster version 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

First Successful Ground Landing Date

- While the hint of using min (i.e. "SELECT min("Date") FROM …" was suggested, an alternate method that yielded the same result was used. Date in ascending order limited to one row would show the same date.

```
[11]: %sql SELECT min("Date") FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (ground pad)" Order BY "Date" ASC LIMIT 1;
```

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

[11]:

min("Date")
01/08/2018

```
[12]: %sql SELECT min("Date") FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (ground pad)"
```

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

[12]:

min("Date")
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

- The Booster versions of successful "Landing_Outcome"(s) are listed below
- The WHERE clause filters for 'Success" and payload mass that is between 4,000 and 6,000 kgs.

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql SELECT "Booster_Version" FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (drone ship)" AND "PAYLOAD_MASS__KG_" between 4000 and 6000;
```

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

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

Total Number of Successful and Failure Mission Outcomes

- The successful "Mission_Outcome" count was totaled for all rows LIKE 'Succ%' to get 100
- The failure "Mission_Outcome" count was totaled for all rows LIKE 'Fail%' to get 1

List the total number of successful and failure mission outcomes

```
%sql SELECT COUNT("Mission_Outcome") AS "Successes" FROM SPACEXTBL WHERE "Mission_Outcome" LIKE 'Succ%';
```

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

Successes
100

```
%sql SELECT COUNT("Mission_Outcome") AS "Failures" FROM SPACEXTBL WHERE "Mission_Outcome" LIKE 'Fail%';
```

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

Failures
1

Boosters Carried Maximum Payload

- Distinct Booster Versions are listed for rows in the subselect, where payload mass is the MAX values

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

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

```
 * 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

- The query returns month name as indicated in the Note, along with landing outcome, booster version, and launch site, all WHERE landing outcome is LIKE 'FAIL%' and the year = '2015' as indicated in the Note

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%sql SELECT substr(Date, 4, 2) AS "MONTH", "Landing_Outcome", "Booster_Version", "Launch_Site"  FROM SPACEXTBL \
WHERE "Landing_Outcome" LIKE "Fail%" AND substr(Date,7,4)='2015' ;
```

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

MONTH	Landing_Outcome	Booster_Version	Launch_Site
10	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

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.

```
%sql SELECT COUNT(Landing_Outcome), Landing_Outcome FROM SPACEXTBL \
WHERE DATE(substr(Date,7,4) \
||'-' \
||substr(Date,4,2) \
||'-' \
||substr(Date,1,2)) \
BETWEEN DATE('2010-06-04') AND DATE('2017-03-20') \
GROUP BY Landing_Outcome \
ORDER BY COUNT(Landing_Outcome) DESC;
```

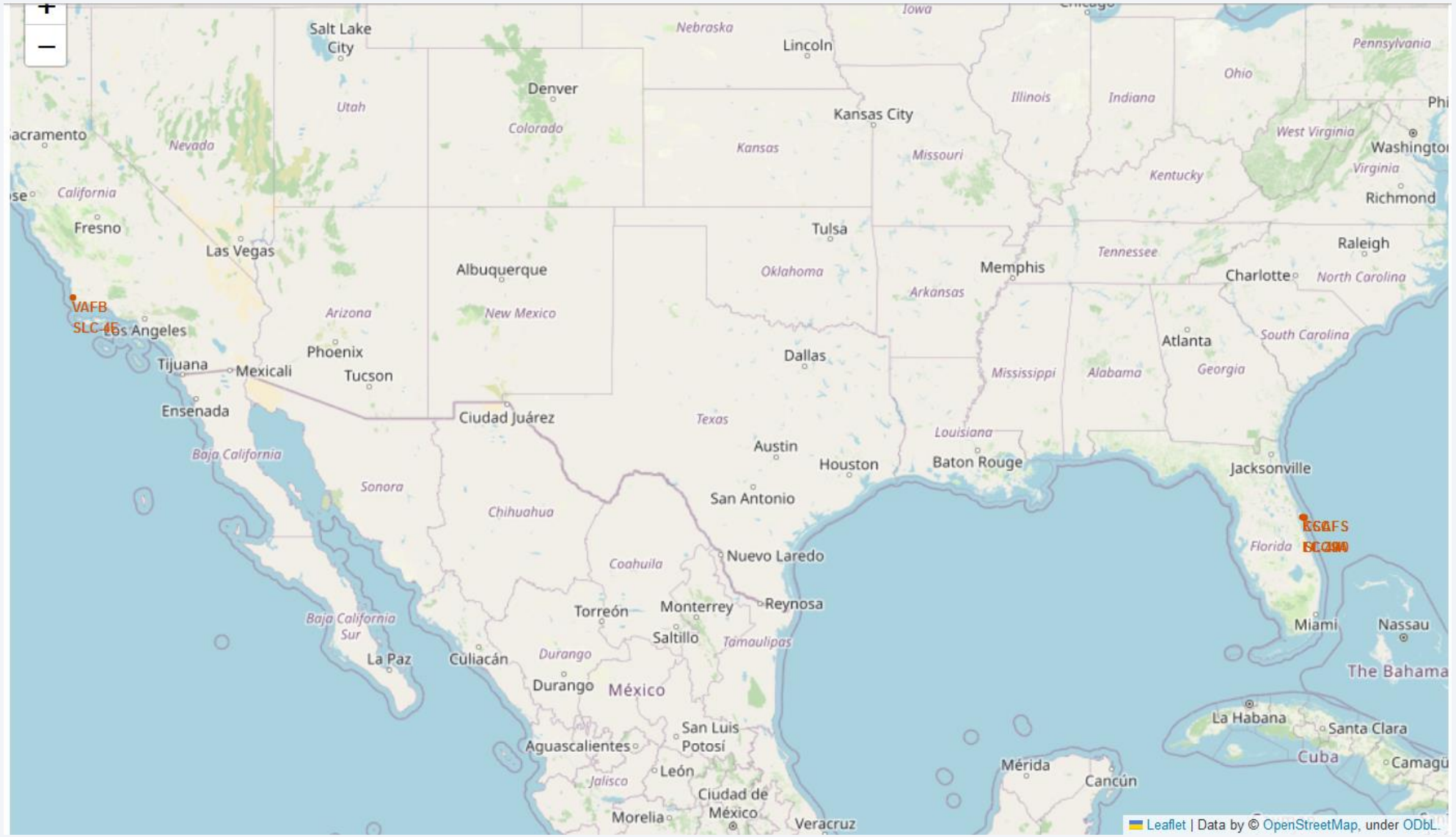
```
 * sqlite:///my_data1.db
Done.
```

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

- There are 8 different landing outcomes listed and their respective counts in descending order. The Date is text in the DB and has to be handled carefully.
- Data is selected between the specified dates. It is Grouped by landing outcome and ordered by that count in descending order to get these results.

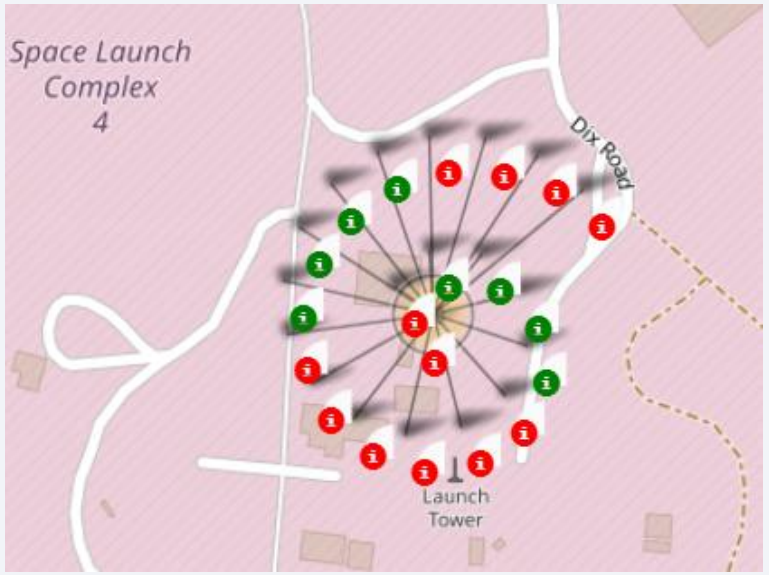

Section 3
Launch Sites
Proximities Analysis

GIS - Launch Sites

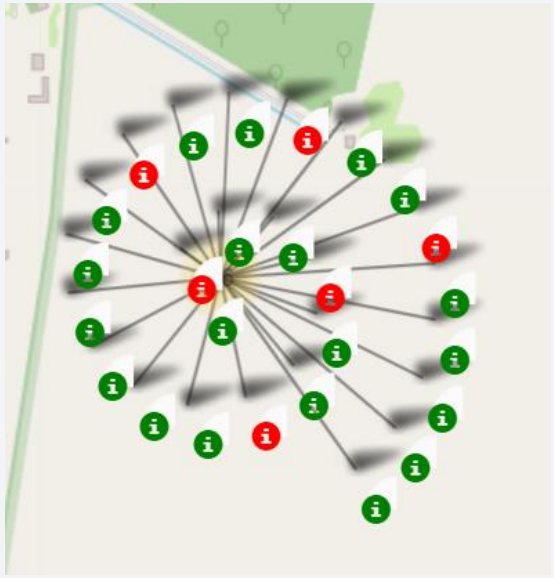


The two launch site locations are VAFB SLC – 4E on the west coast of California (in red). The remaining 3 sites are on the east coast of Florida (in red).

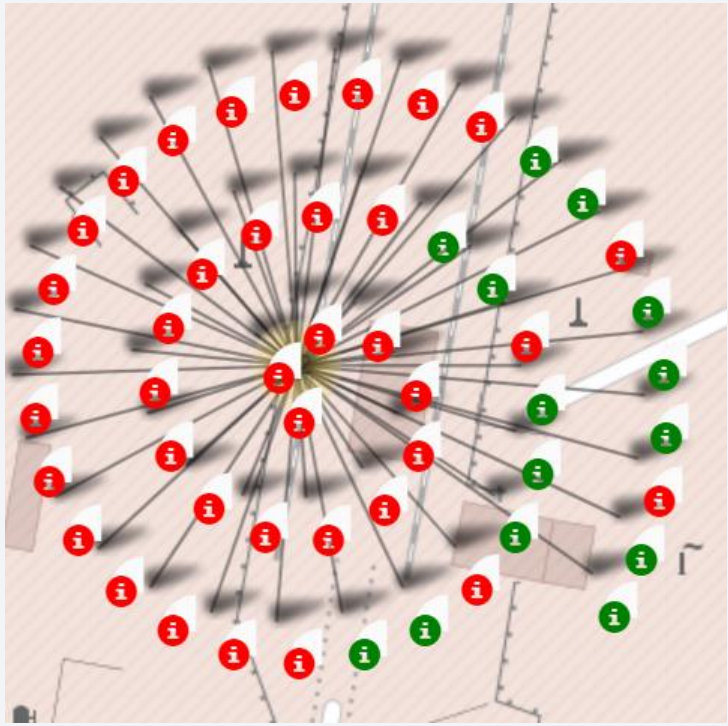
Launch Site Success/Failure Details



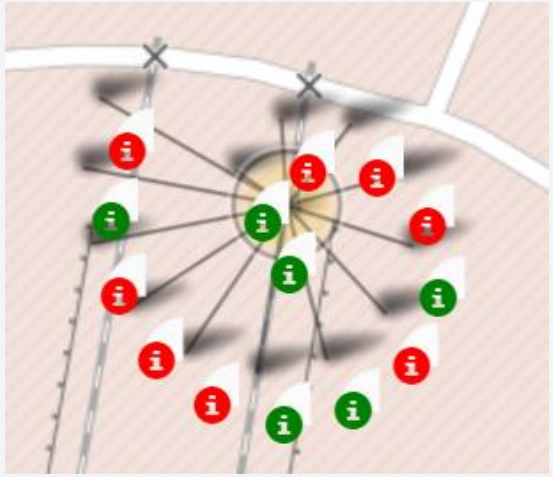
VAFB SLC-4E



KSC LC-39A



CCAFS LC-40

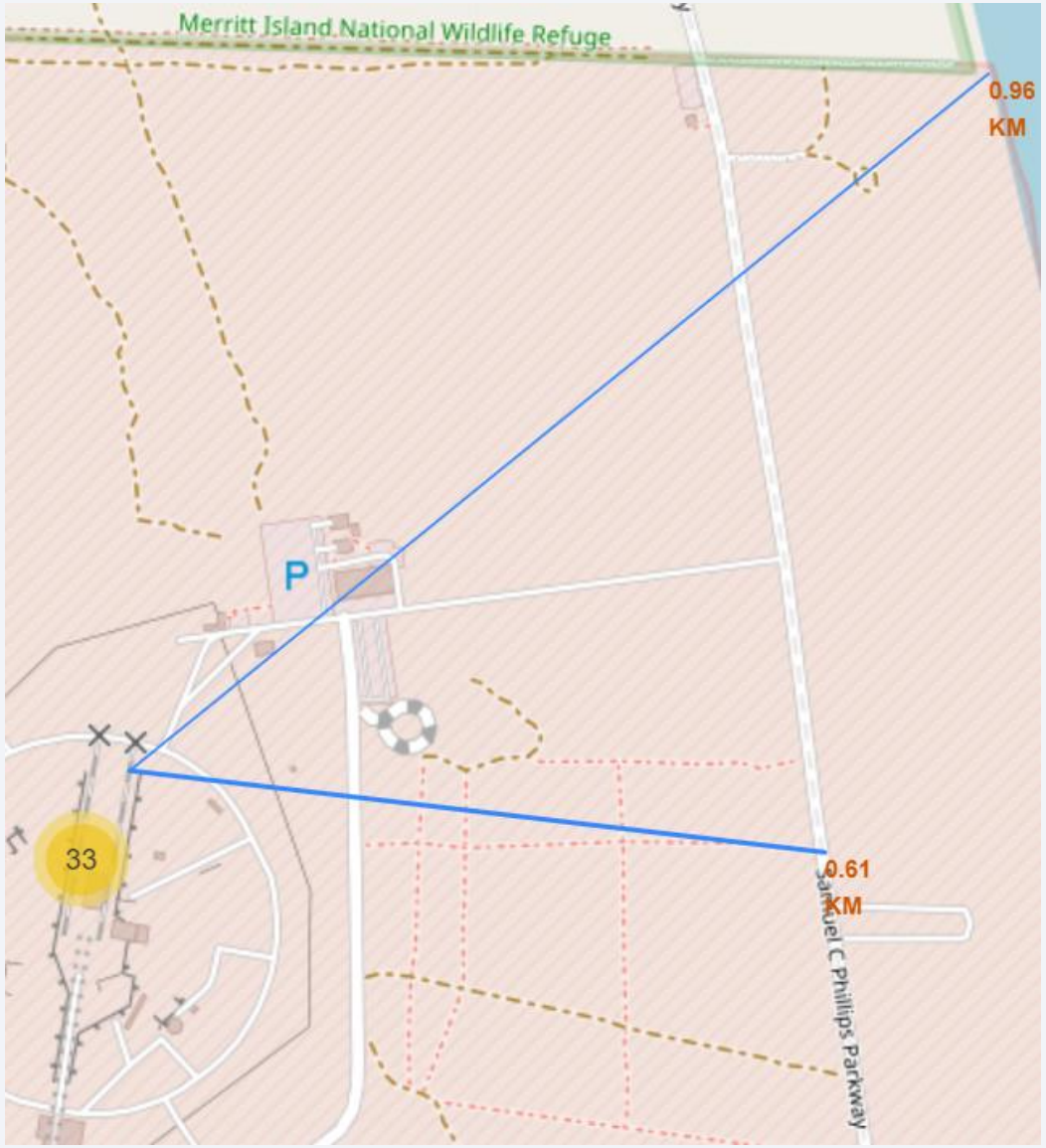


CCAFS SLC-40

Red markers indicate failures.

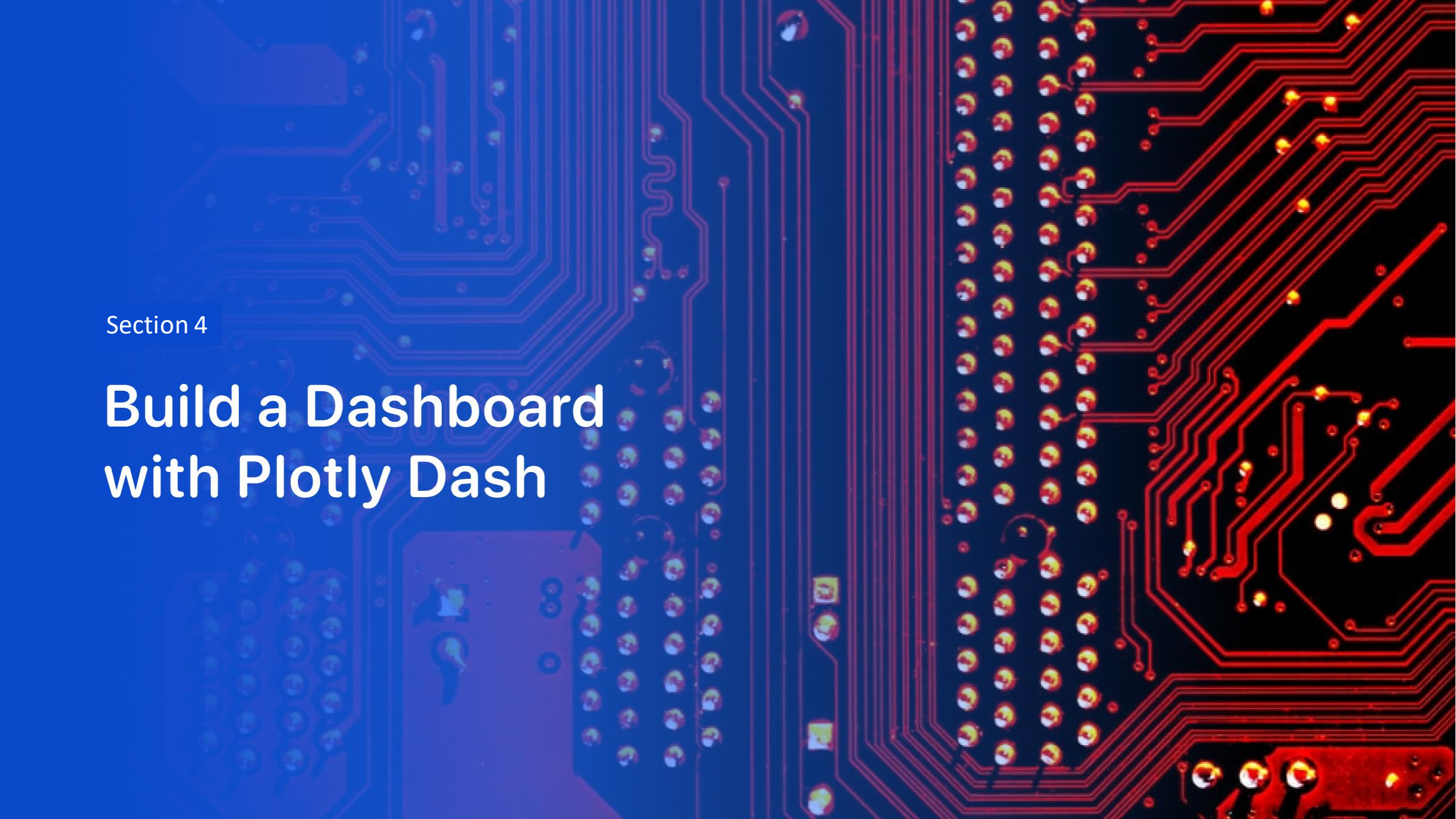
Green markers indicate successes.

Launch Site Proximity Examples

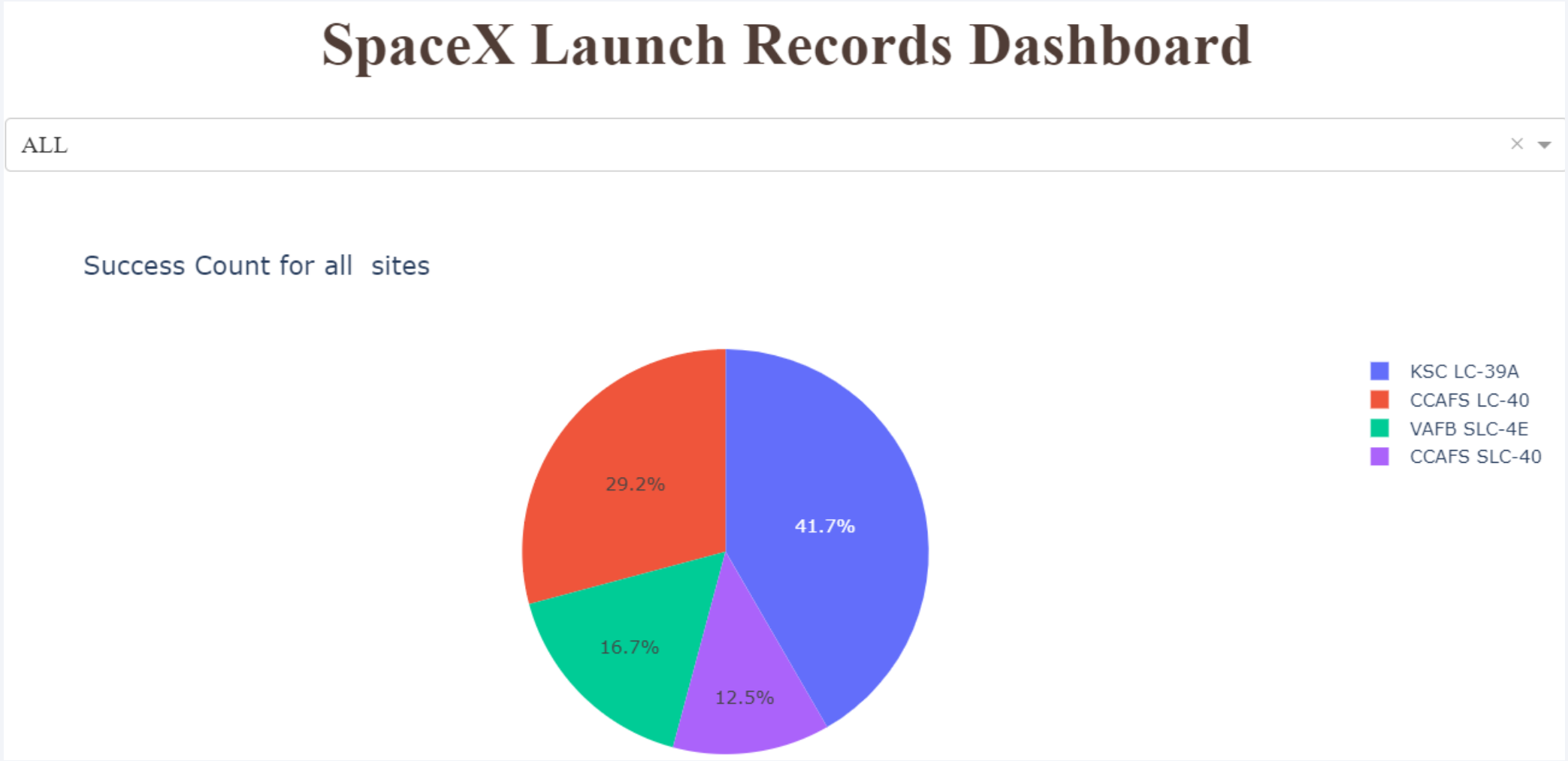


CCAFS SLC-40 is 0.96 km from where Merrit Island National Wildlife Refuge meets the coast (as shown in red, with the distance marker, on the map).

It is also approximately 0.61 km from the Samuel C Phillips Parkway.

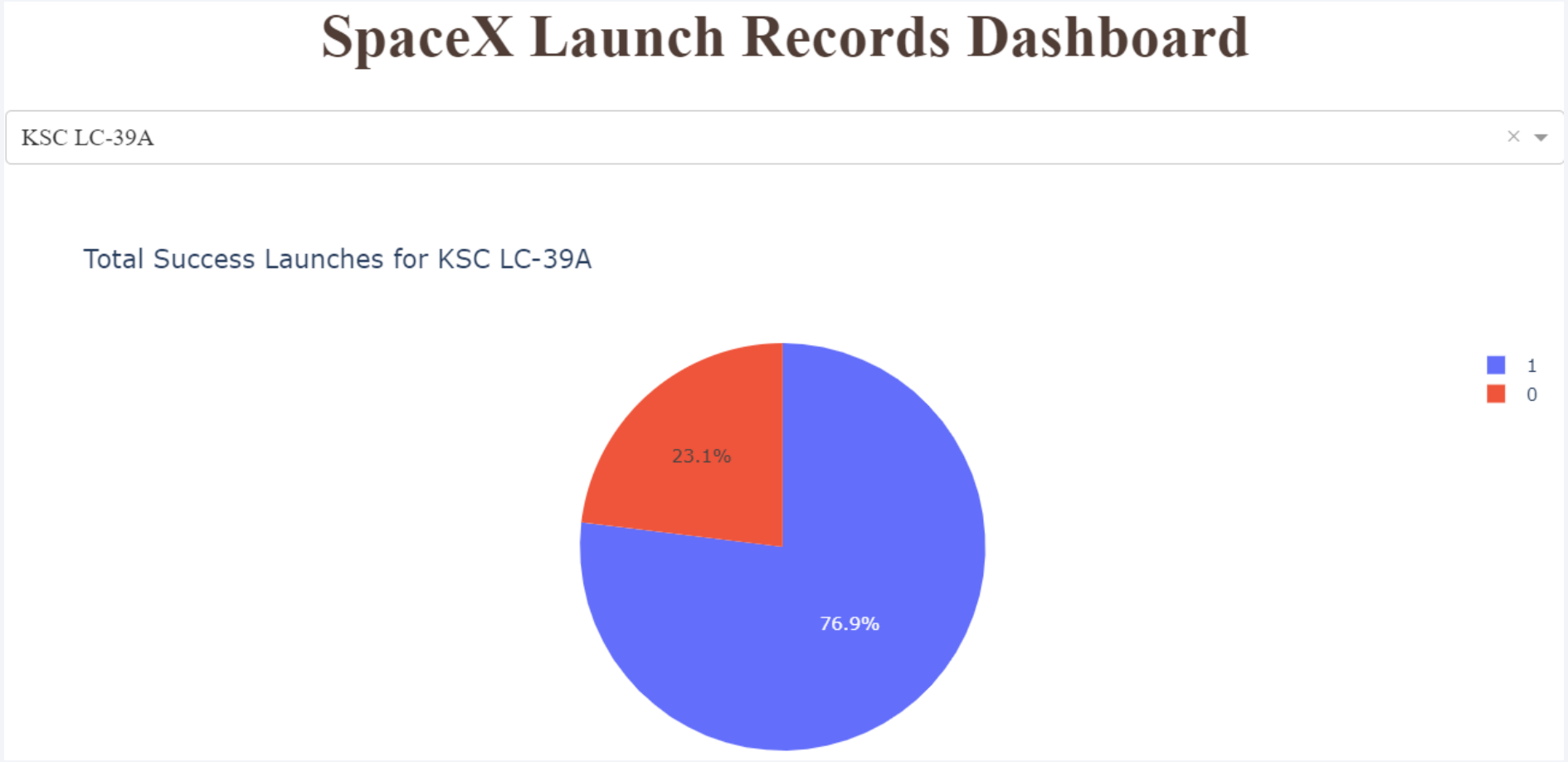

Section 4
Build a Dashboard with Plotly Dash

Launch Success for All Sites



- Site KSC LC-39A has the most successes.

KSC LC-39A Site Launch Details



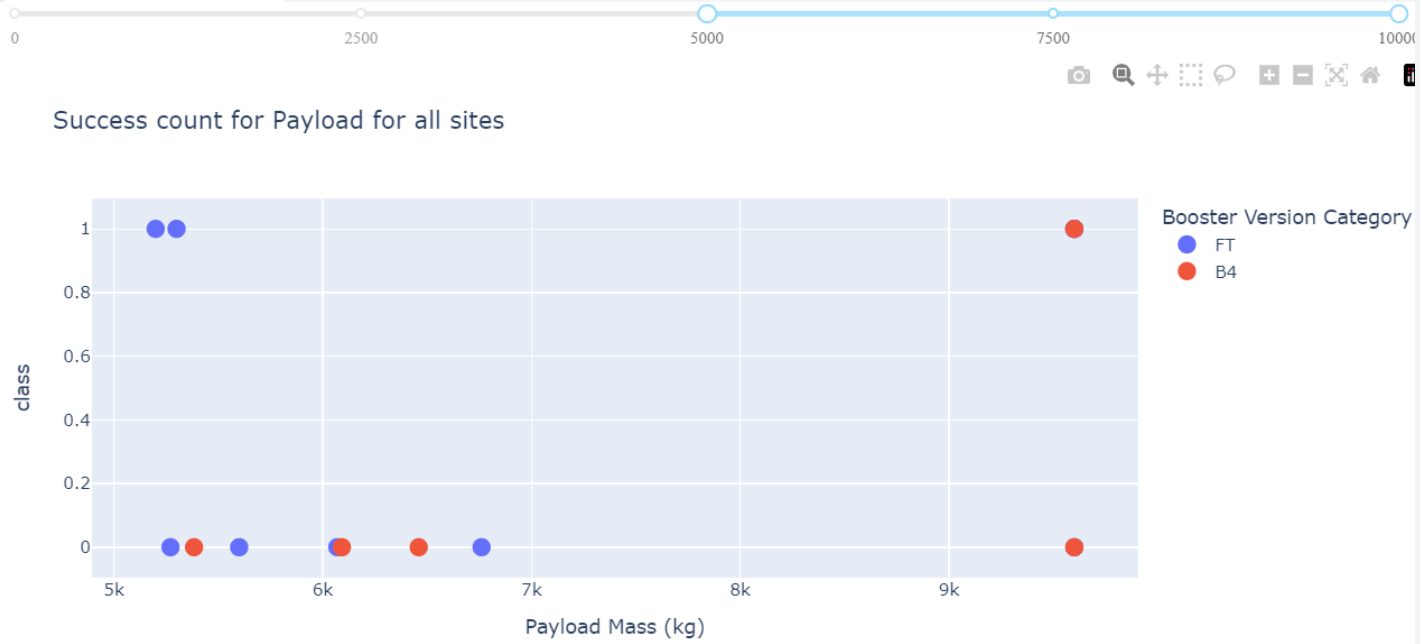
- This site had about a 77% success rate and a 23% failure rate.

Payload vs. Launch Outcome Scatter Plots

- For Low Payloads, The FT booser version does well (green). But the v1.1 does not (red)
- Heavy Payloads have lower success rates per launch

Low Payloads

0 – 5,000 kg.



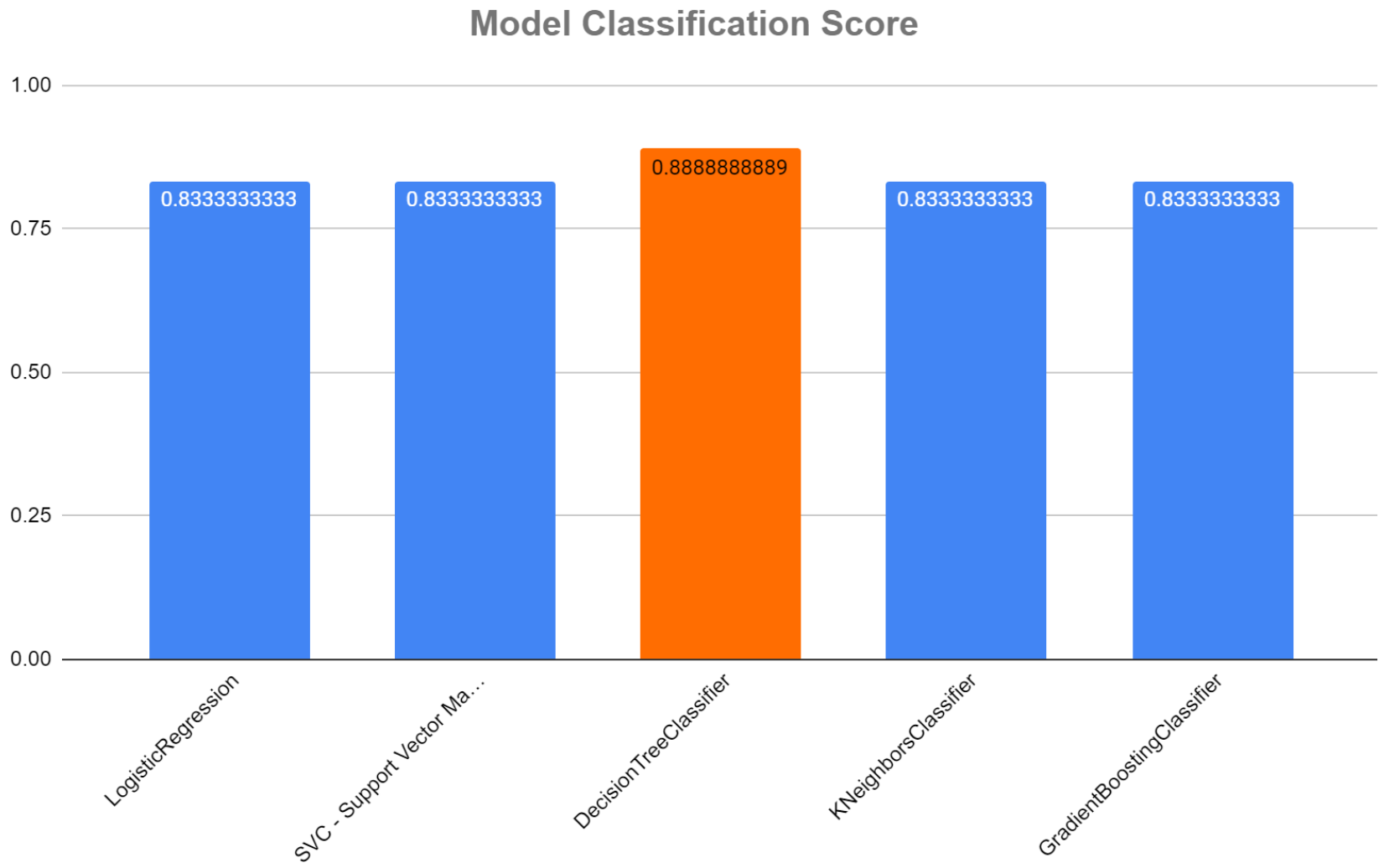
Heavy Payloads

5,000 – 10,000 kg.

Section 5

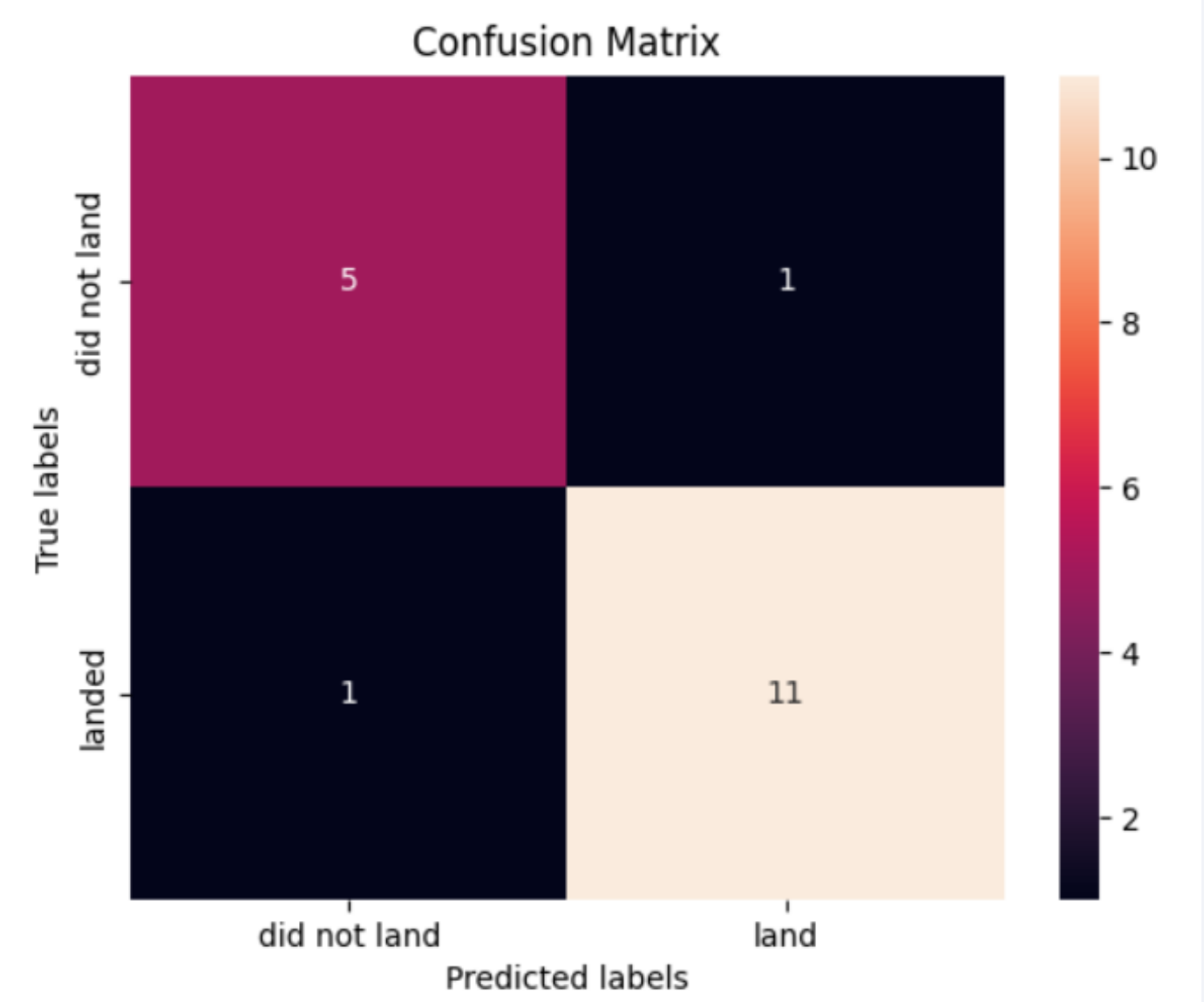
Predictive Analysis (Classification)

Classification Accuracy



- The Decision Tree classifier had the highest R^2 score of all the classifiers.

Confusion Matrix – Decision Tree Classifier



- The Decision Tree Classifier can distinguish between the different classes, better than the other classifiers
- While all of the classifiers have problems identifying increased false positives, the Decision Tree Classifier has less of an issue in this regard than the other classifiers

Conclusions

- The Decision Tree classifier was the best performing predictive model. It had an 89% fit to the test/validation data after training on an independent set of data. It can be used to predict a successful launch outcome based on several key parameters
- Since 2013, launch success has generally continued to increase through the year 2020
- Launch Site KDC LC-39 A has a much better launch outcome success rate compared to the other launch sites. More analysis on why, could benefit other site launch success.
- For lower weight payloads (< 5,000 kg.), the FT booster version does extremely well. Heavy payload (> 5,000 kg.) launch outcome success rate is poor and requires more analysis.
- Launch site CCAFSSLC 40 has good success with payloads over 6,000 kg

Thank you!