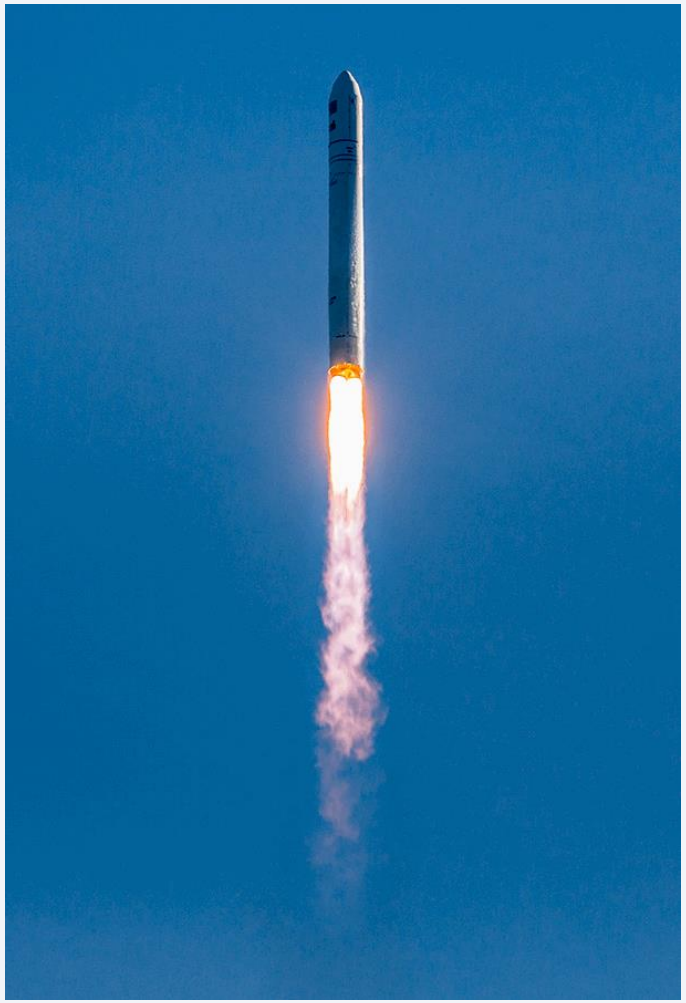
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

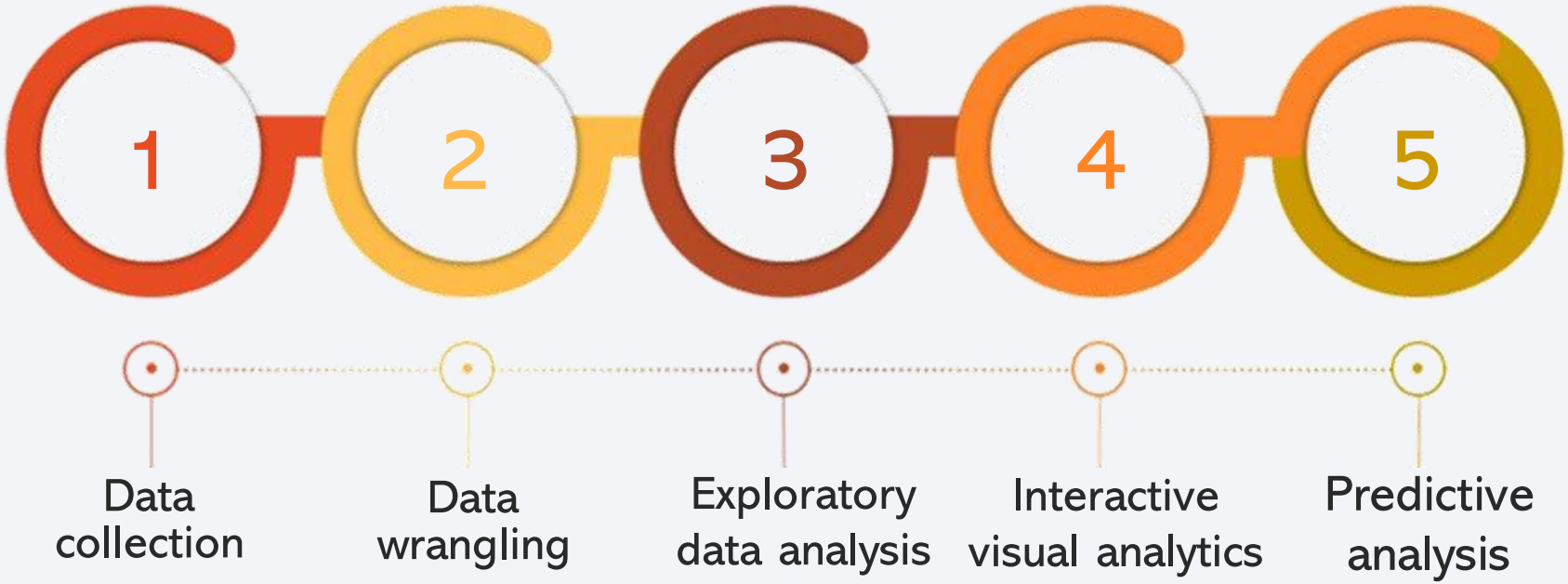
Freya Saima
September 20, 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion



Executive Summary



Abstract

- In this project we present the application of ML algorithms in order to predict the success or no-success of landing of the first stage Space X Falcon 9 rockets.
- The following ML Classification models were used to predict if a the first stage will successfully land: Logistic Regression (LogR), Support Vector Machine (SVM), Decision Tree Classifier (DTree) and K nearest neighbors (KNN)
- As results, LogR, SVM and KNN were able to predict with a test accuracy of 83%, while DTree has an accuracy of 77%

Introduction

Space X 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 Space X can reuse the first stage.

Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

Can we determine if the first stage will land successfully or not?

Section 1

Methodology

Methodology

1. Data collection methodology:
   - SpaceX API: https://api.spacexdata.com/v4/launches/past (json_normalize method)
   - Data: Date, Booster Version, Launch site (Lat, Long), Payload Mass kg, Flights, Reused, Legs, Landing Pad, Block, Reused Count, GridFins, Serial, Orbit, Outcome.
2. Perform data wrangling
   - Missing values: (i) Payload Mass --> mean (ii)
   - Creation of class from Outcomes : Success (1) or No-Success (0)
3. Perform exploratory data analysis (EDA) using visualization and SQL
4. Perform interactive visual analytics using Folium and Plotly Dash
5. Perform predictive analysis using classification models
   - Train data (80%), Test data (20%)
   - Models: Logistic Regression (LogR), Support Vector Machine (SVM), Decision Tree Classifier (DTree) and K nearest neighbors (KNN)
   - Best parameters selection: GridSearchCV

Data Collection

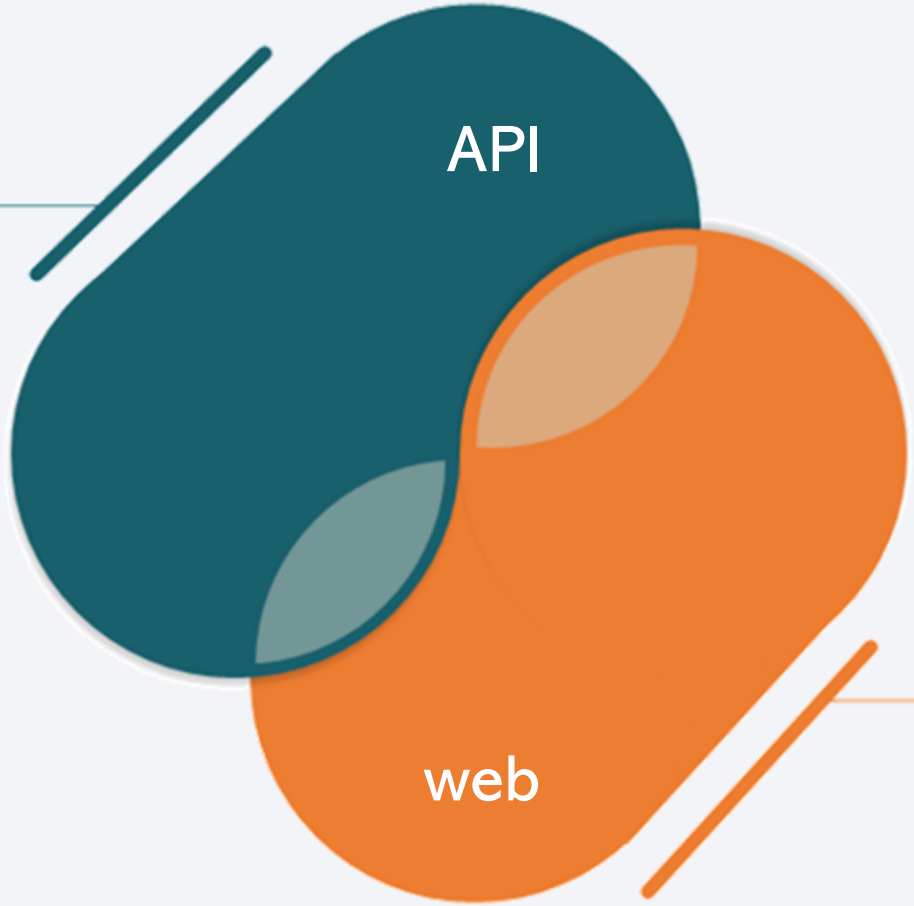
1. API

url:
https://api.spacexdata.com/v4/launches/past

Process:

Acquire data from SpaceX API

Clean data



url:
https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Process:

Extract Table

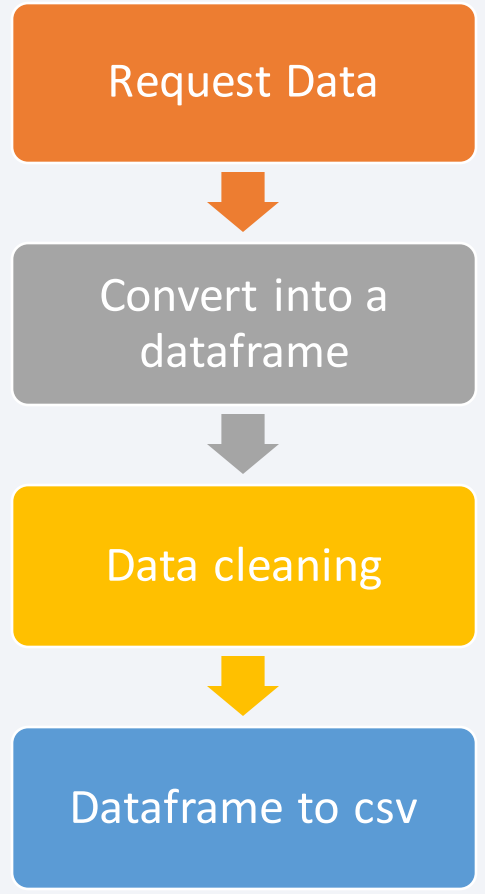
Parse the table

Convert to DataFrame

2. Website

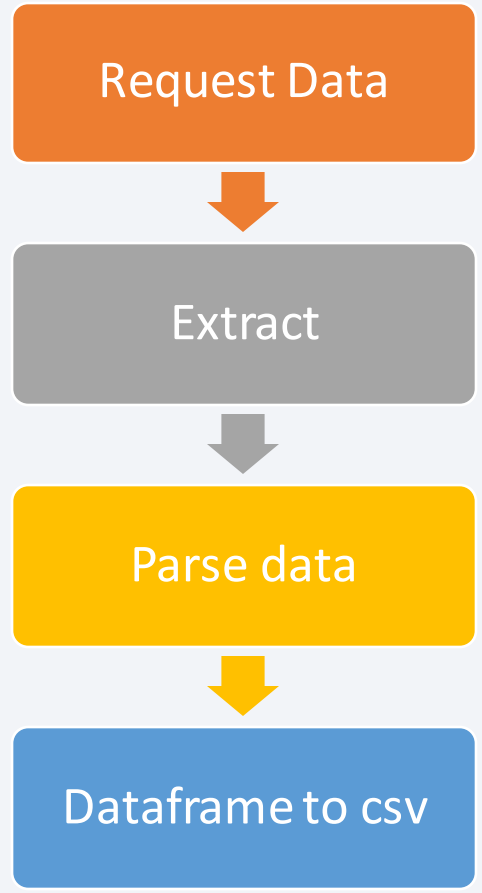
Data Collection – SpaceX API

- Request to the SpaceX API
- Convertion of the json result into a dataframe
- Information request about the launches using the IDs
- Dataframe filtering to only include “Falcon 9” launches
- GitHub URL: https://github.com/FreyaSaima/SpaceX-Launch-Analysis/blob/main/1.%20Spacex-data-collection-api.ipynb



Data Collection - Scraping

- HTTP GET method to request the Falcon9 Launch HTML page
- Create a `BeautifulSoup` object from the HTML `response`
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables
- GitHub URL:https://github.com/FreyaSaima/SpaceX-Launch-Analysis/blob/main/2.%20space_X_webscraping.ipynb



Data Wrangling

- Dealing with Missing Values
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome of the orbits
- Create a landing outcome label from Outcome column

GitHub URL: https://github.com/FreyaSaima/SpaceX-Launch-Analysis/blob/main/3.%20Spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Visualize the relationship between FlightNumber and PayloadMass
Visualize the relationship between Flight Number and Launch Site
Visualize the relationship between Payload and Launch Site
Visualize the relationship between Success rate of each Orbit type
Visualize the relationship between FlightNumber and Orbit type
Visualize the relationship between Payload and and Orbit type
Visualize the launch success yearly trend

Objective

- Best describe the relationship between categorical data
- Compare several categorical data
- Show evolution in the time series data

GitHub URL: https://github.com/FreyaSaima/SpaceX-Launch-Analysis/blob/main/4.%20SpaceX_eda-sql-coursera_sqllite.ipynb

EDA with SQL

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 succesful landing outcome in ground pad was acheived.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster_versions which have carried the maximum payload mass.
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.

SPACE X

Objective

- Find meaningful insights
- Compare several categorical data
- Show key aspects in the dataset

GitHub URL: https://github.com/FreyaSaima/SpaceX-Launch-Analysis/blob/main/5.%20SpaceX_eda-dataviz-v2.ipynb

Build an Interactive Map with Folium

Create and add folium.Circle and folium.Marker for each launch site on the site map to check sites close proximity

Mark the success/failed launches for each site on the map with a MarkerCluster

Calculate the distances between a launch site to its proximities: coast line, railway, city and highway

SPACE X

Objective

- Find meaningful insights in their launch locations
- Find some geographical patterns
- Show key aspects in the outcomes and Launch sites locations

GitHub URL : https://github.com/FreyaSaima/SpaceX-Launch-Analysis/blob/main/6.%20SpaceX_launch-site-location-Dashboard.ipynb

Build a Dashboard with Plotly Dash

- Launch site drop down input component
- A success-pie-chart based on the selected site dropdown
- A ranger slider to select payload
- A success-payload-scatter plot based on the selected site dropdown

Objective

- Inspect the relationship of success rate between launch site and payload

GitHub URL: https://github.com/FreyaSaima/SpaceX-Launch-Analysis/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Data Split: Train data (80%), Test data (20%)

ML Classification Models: Logistic Regression (LogR), Support Vector Machine (SVM), Decision Tree Classifier (DTree) and K nearest neighbors (KNN)

Best parameters selection: GridSearchCV

Model evaluation: Confusion Matrix, Accuracy score

GitHub URL: https://github.com/FreyaSaima/SpaceX-Launch-Analysis/blob/main/7.%20SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

Results

Exploratory data analysis results

There is a correlation between launch site and success rate

There is a correlation between launch site and booster version

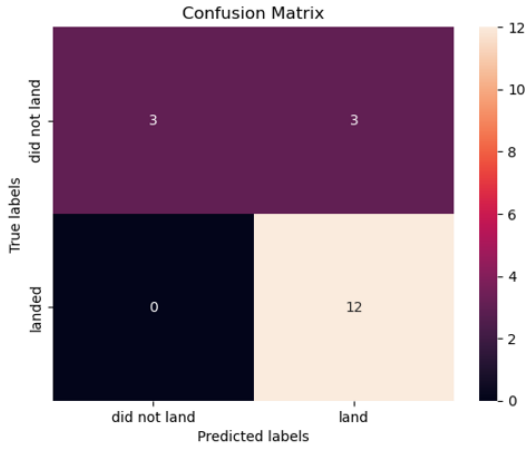
Payload mass is also associated with the success rate: more payload, less likely the first stage will return

There has been an increase in the success rate since 2013

Interactive analytics demo in screenshots

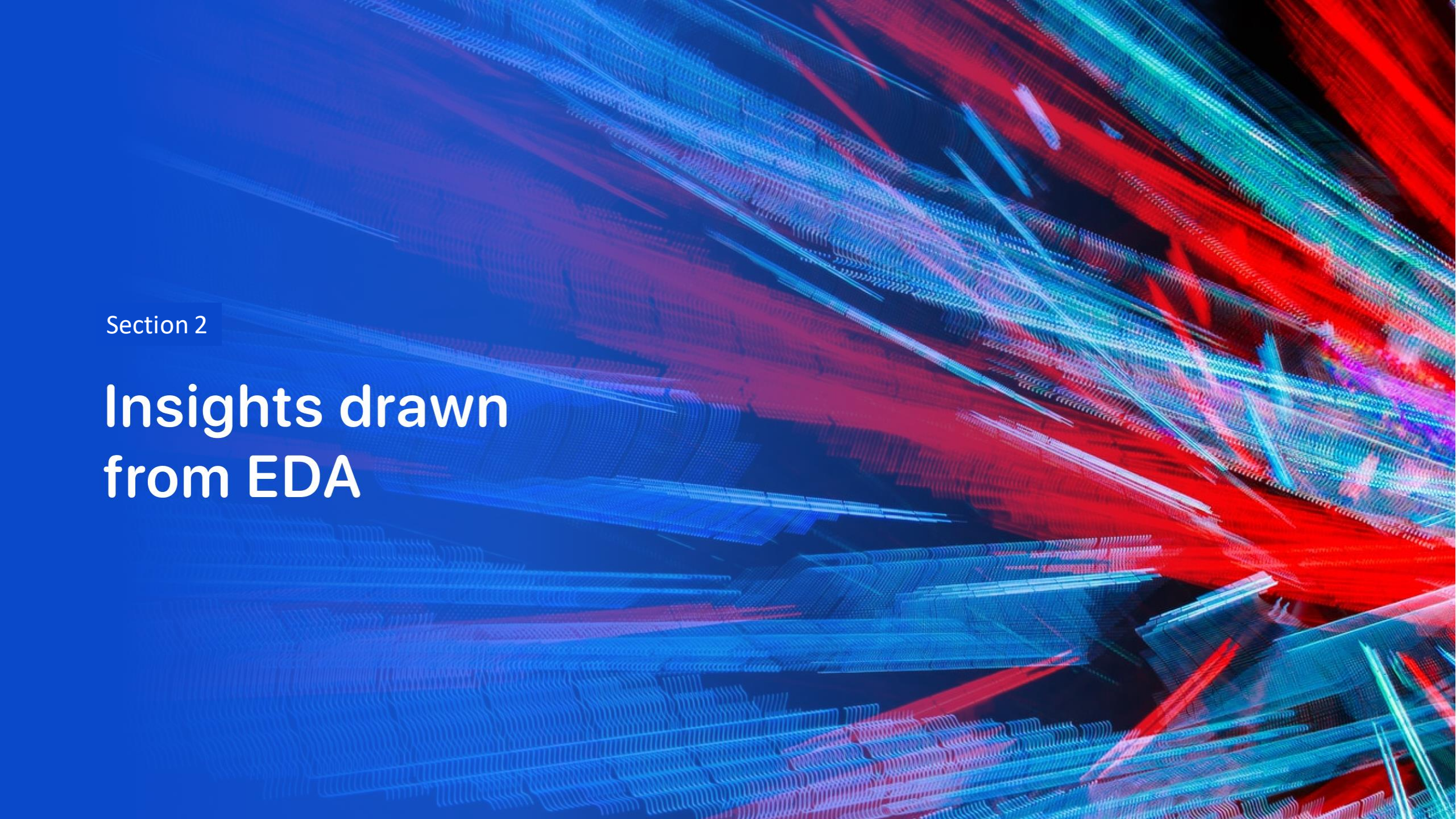
Predictive analysis results

LogR, SVM and KNN =good models

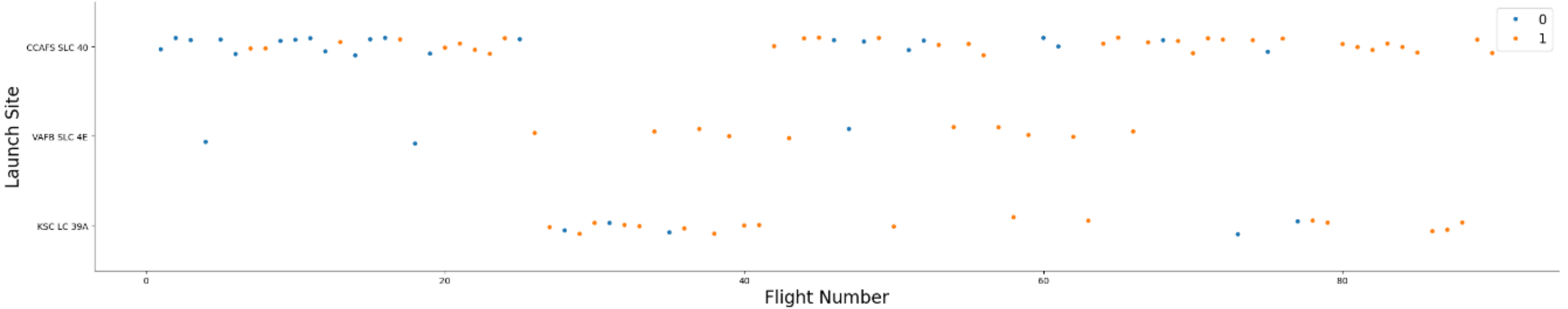


Section 2

Insights drawn from EDA

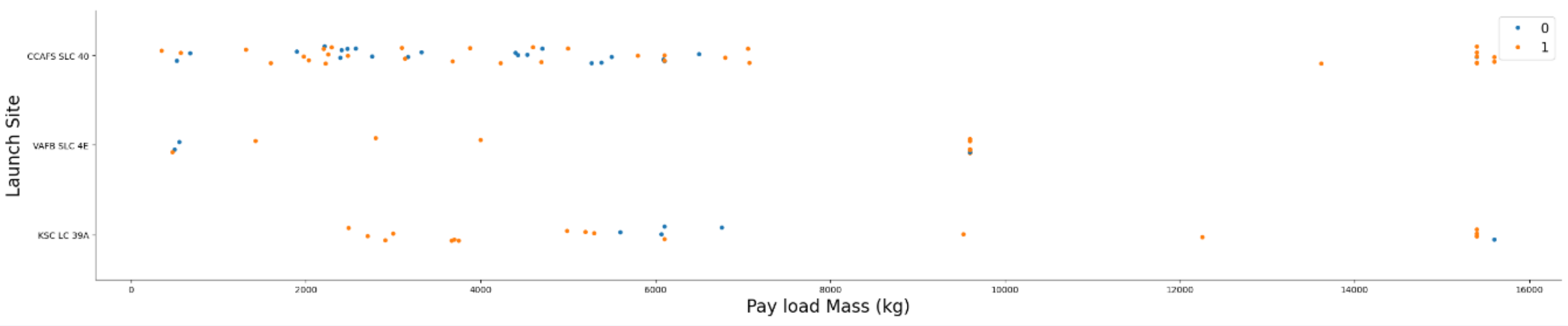


Flight Number vs. Launch Site



Higher success rate in recent Flights

Payload vs. Launch Site

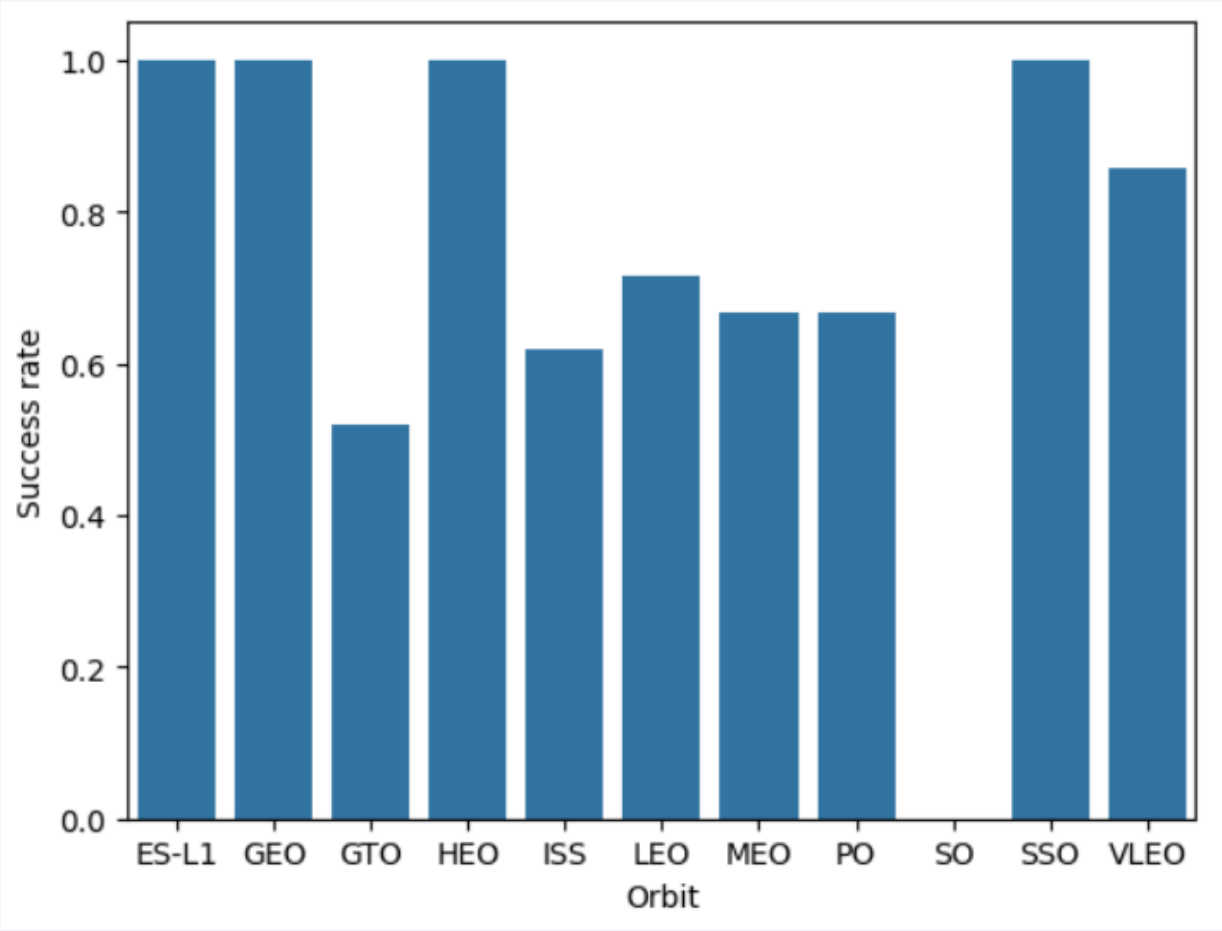


Success more likely at Launch Site CCAFS SLC 40 with high payload

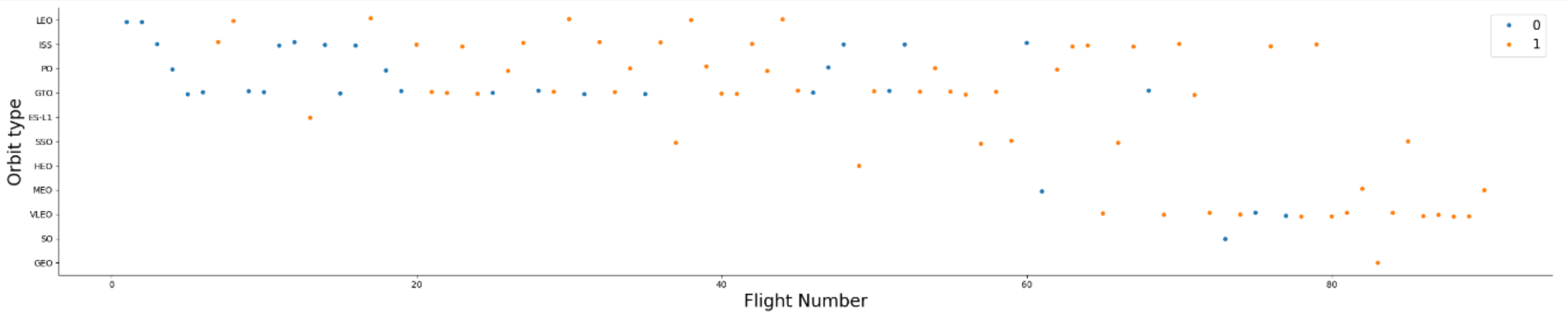
Success more likely at Launch Site KSC LC 39A with moderate payload

Success Rate vs. Orbit Type

Success more likely at orbits ES-L1, GEO, HEO and SSO

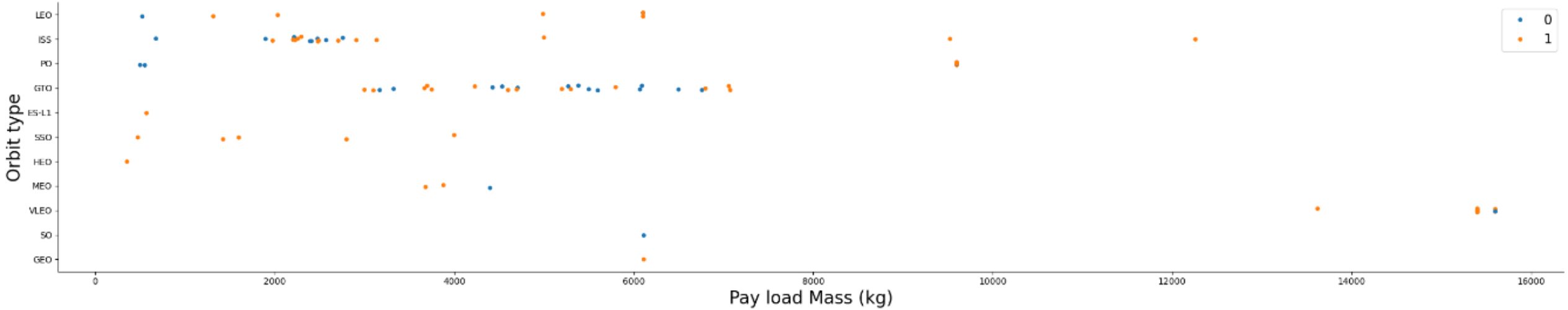


Flight Number vs. Orbit Type



Last Launches have been carried out in HEO, MEO, VLEO, SO and GO orbits, and they have been very successful

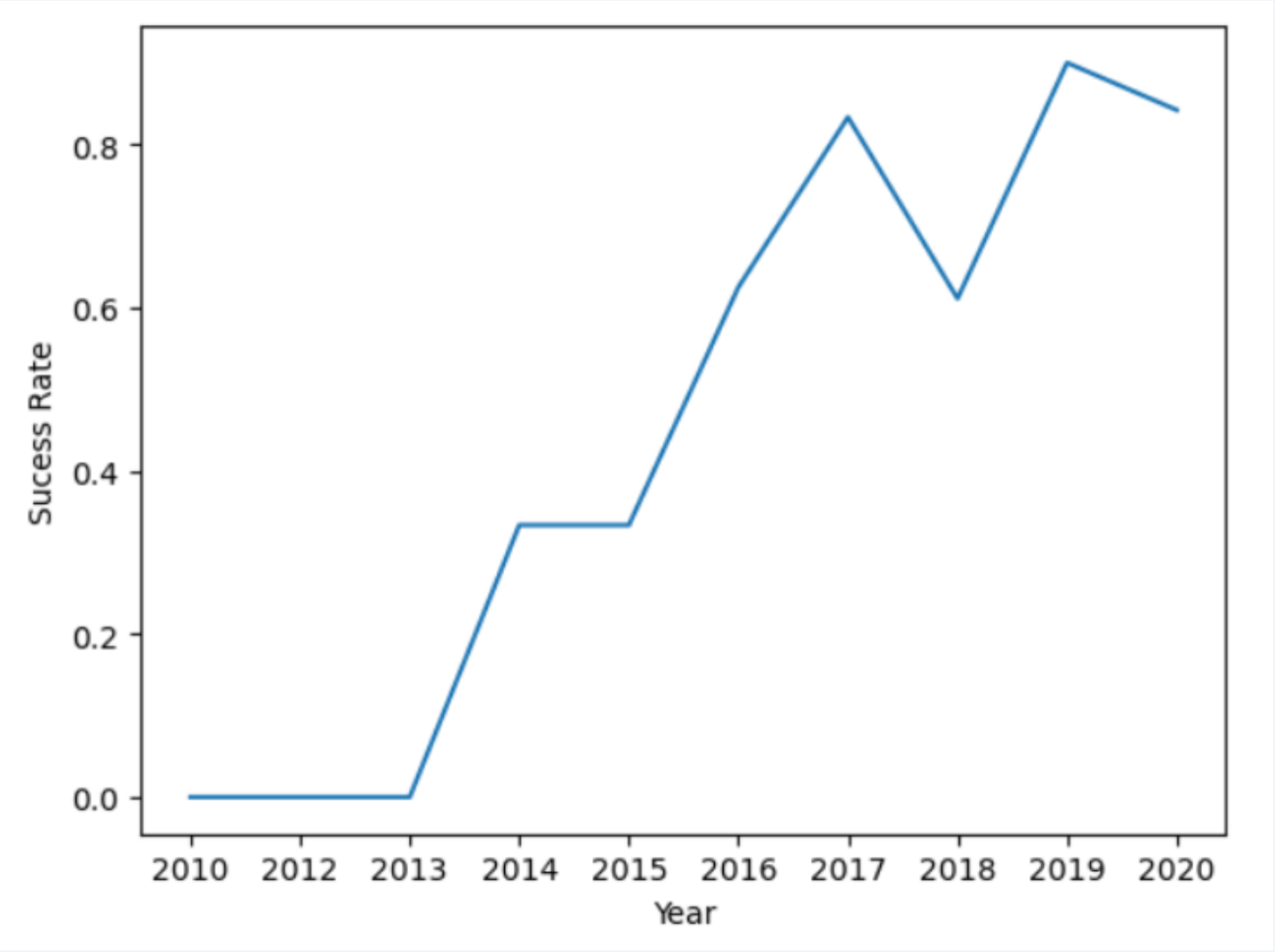
Payload vs. Orbit Type



No clear result. Both successful and failure landings in any orbit

Launch Success Yearly Trend

Success rate has increased dramatically since 2013



All Launch Site Names

4 unique launch sites:

```
%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'

```
%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

All landing outcomes == Failure

Total Payload Mass

```
%sql SELECT PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE CUSTOMER=='NASA (CRS)'
 * sqlite:///my_data1.db
Done.
```

PAYLOAD_MASS__KG_
500
677
2296
2216
2395
1898
1952
3136
2257
2490
2708
3310
2205
2647
2697
2500
2495
2268
1977
2972

The total payload carried by boosters from NASA are very high

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2536.6 kg

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

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

The first successful landing outcome on ground pad was on December, 12, 2015

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

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

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

Done.

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

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes are:

Successful: 99 Success with 1 payload status unclear

Failure: 1

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

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

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

Boosters Carried Maximum Payload

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

- The names of the booster which have carried the maximum payload mass are F9

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

There were 2 failed landing_outcomes in drone ship for in year 2015

```
%%sql
SELECT Date, SUBSTR(Date, 6,2) AS MONTH, Landing_Outcome, Booster_Version, Launch_Site
FROM (SELECT * FROM SPACEXTBL WHERE Landing_Outcome='Failure (drone ship)') WHERE SUBSTR(Date, 0,5)=='2015'
```

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

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

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

The landing outcomes (Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 are:

Failure : 5

Success: 3

```
%%sql
SELECT Landing_Outcome, COUNT(*) AS TOTAL
FROM (SELECT * FROM SPACEXTBL WHERE Landing_Outcome='Failure (drone ship)' OR Landing_Outcome='Success (ground pad)')
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY TOTAL DESC
```

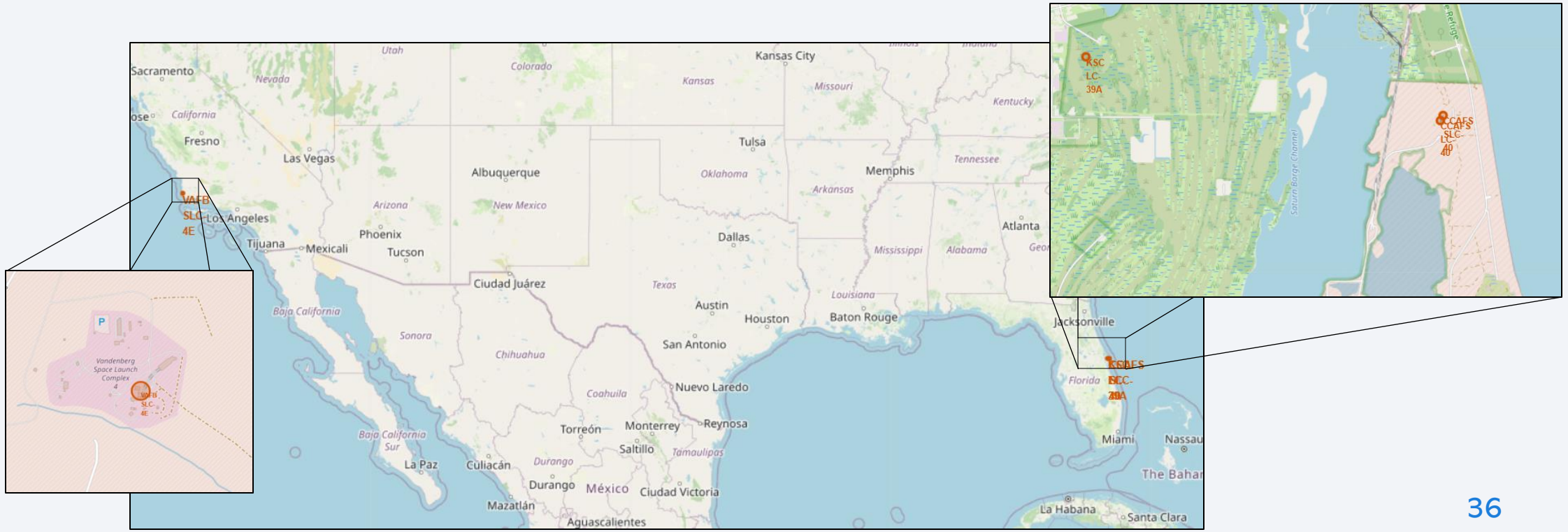
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	TOTAL
Failure (drone ship)	5
Success (ground pad)	3


Section 3
Launch Sites
Proximities Analysis

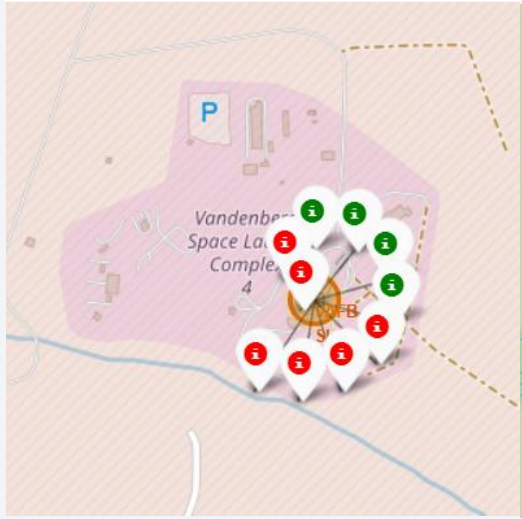
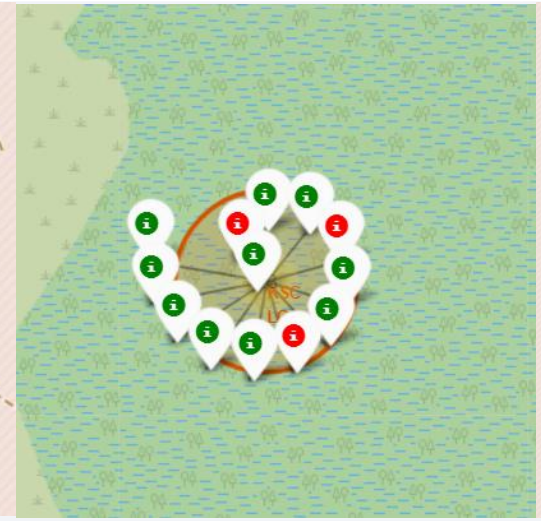
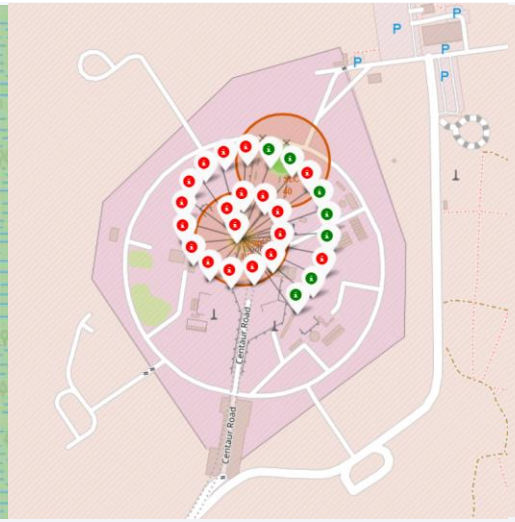
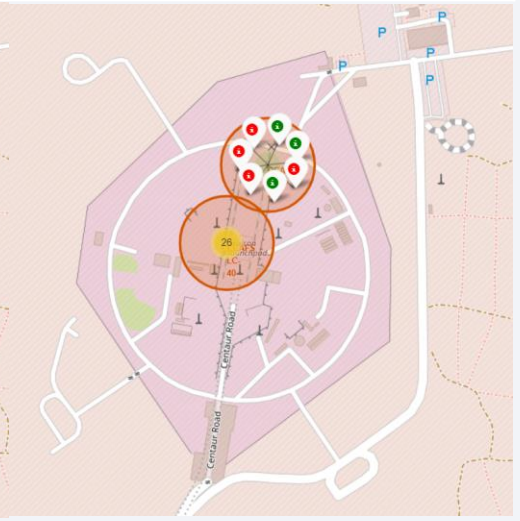
Launch Sites Locations

3 out of 4 launch locations are in the east coast



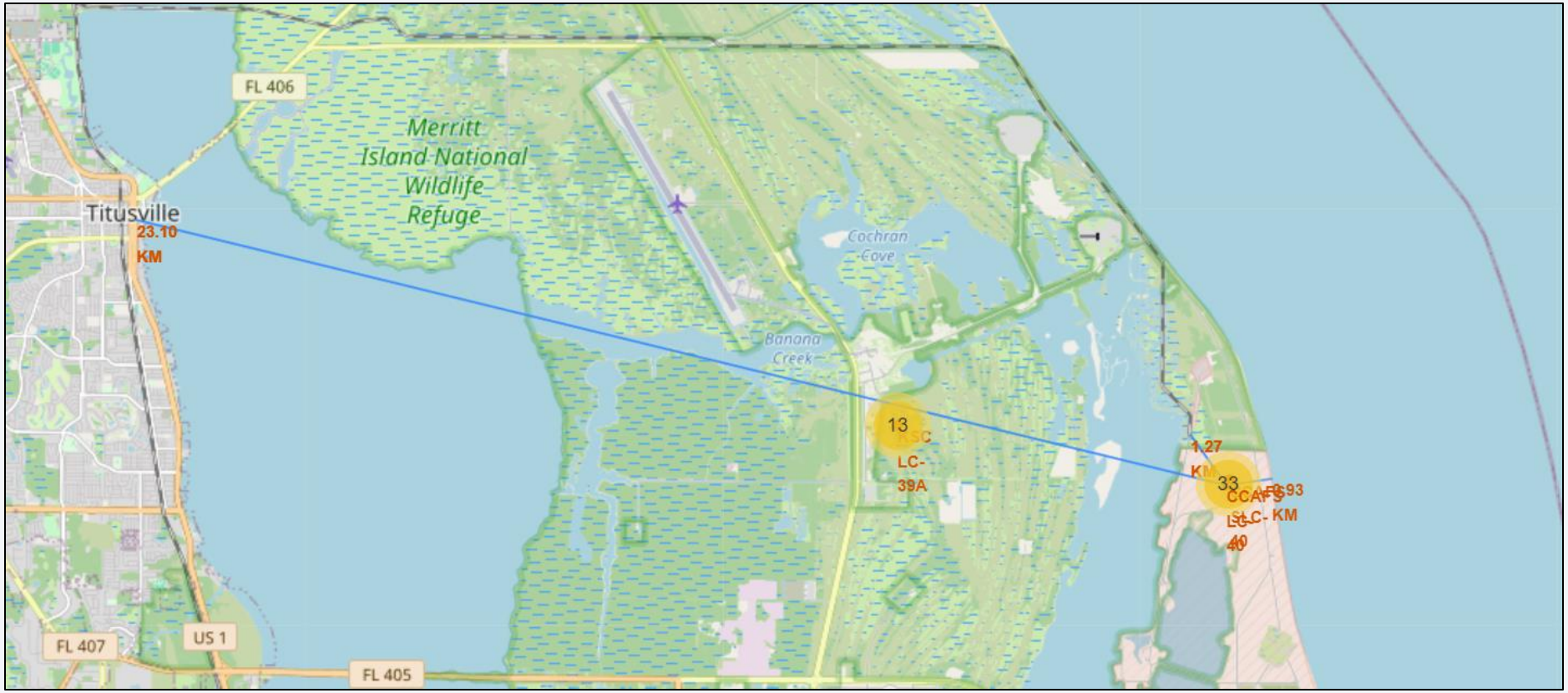
Launch Outcomes

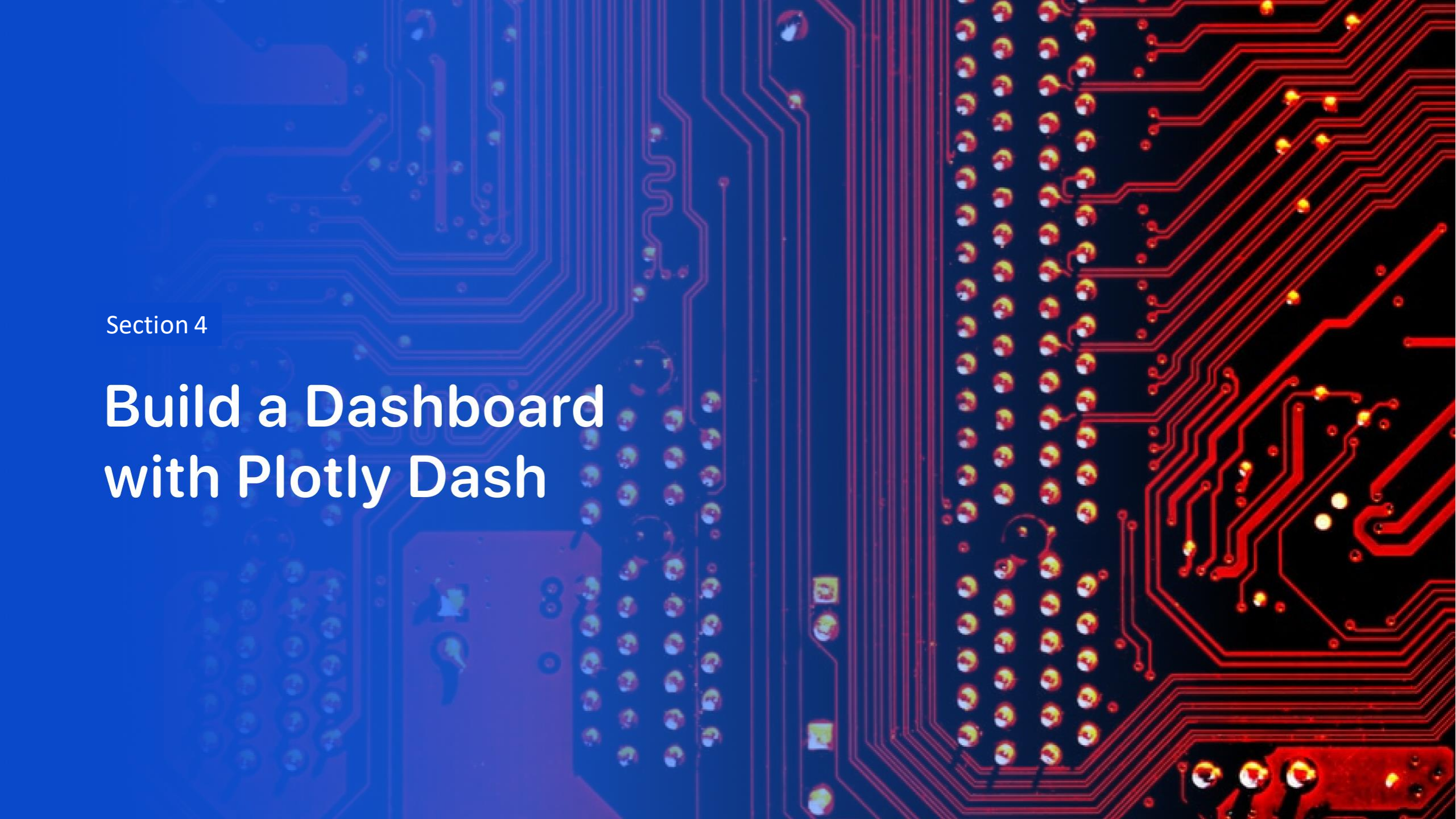
- VAFB SLC-4E has very few launches, most of them failure
- KSC LC 39A has the highest number of successful outcomes
- CCAFS LC-40 has most of launches, most of them failure
- CCAFS SLC-40 has very few launches, 50/50 outcome

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

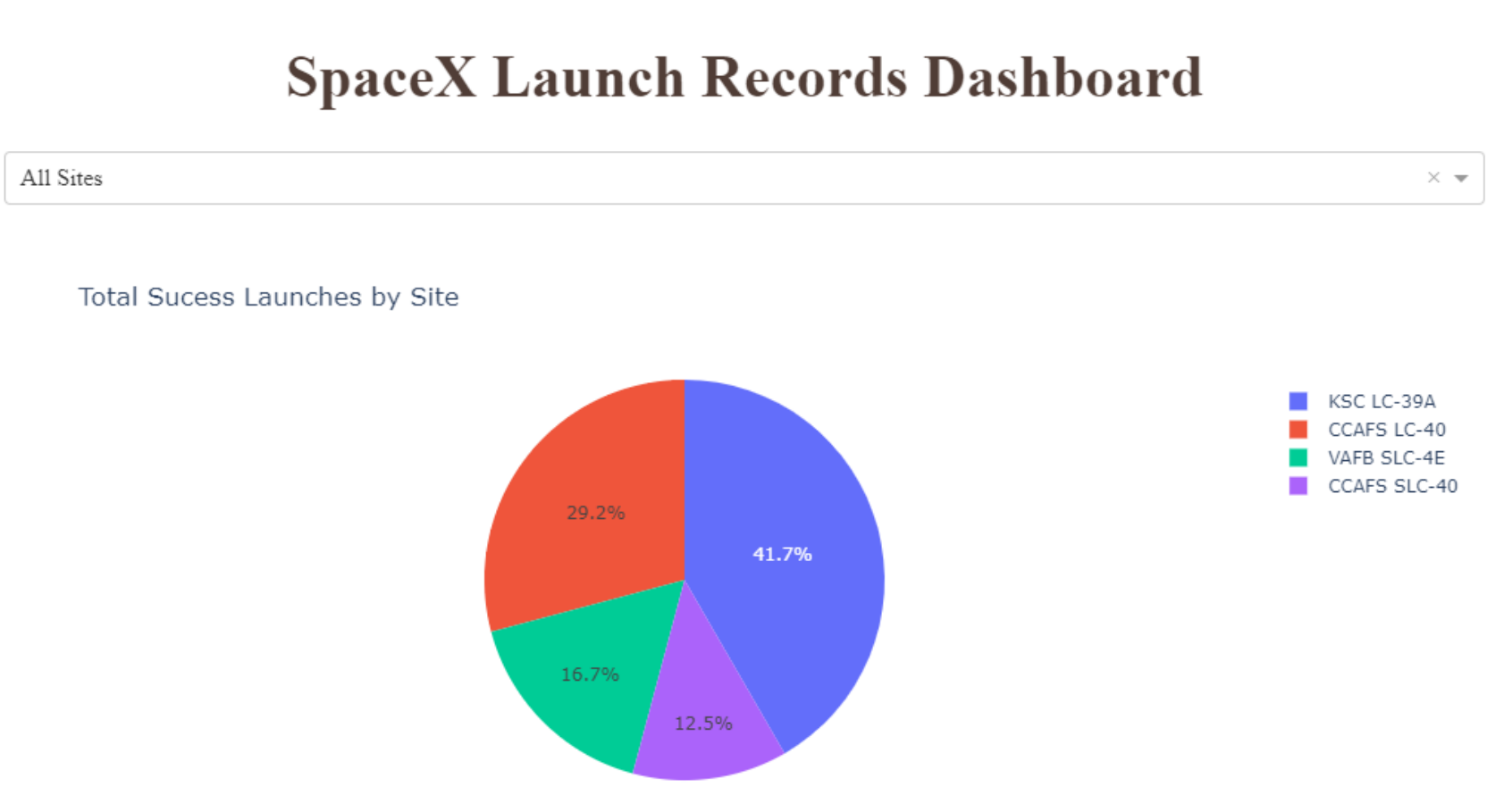
Distance to proximities

The launch site its very close to a railway(1.27 km), a highway, a coastline (0.9 km) and a city (23.1km)




Section 4
Build a Dashboard with Plotly Dash

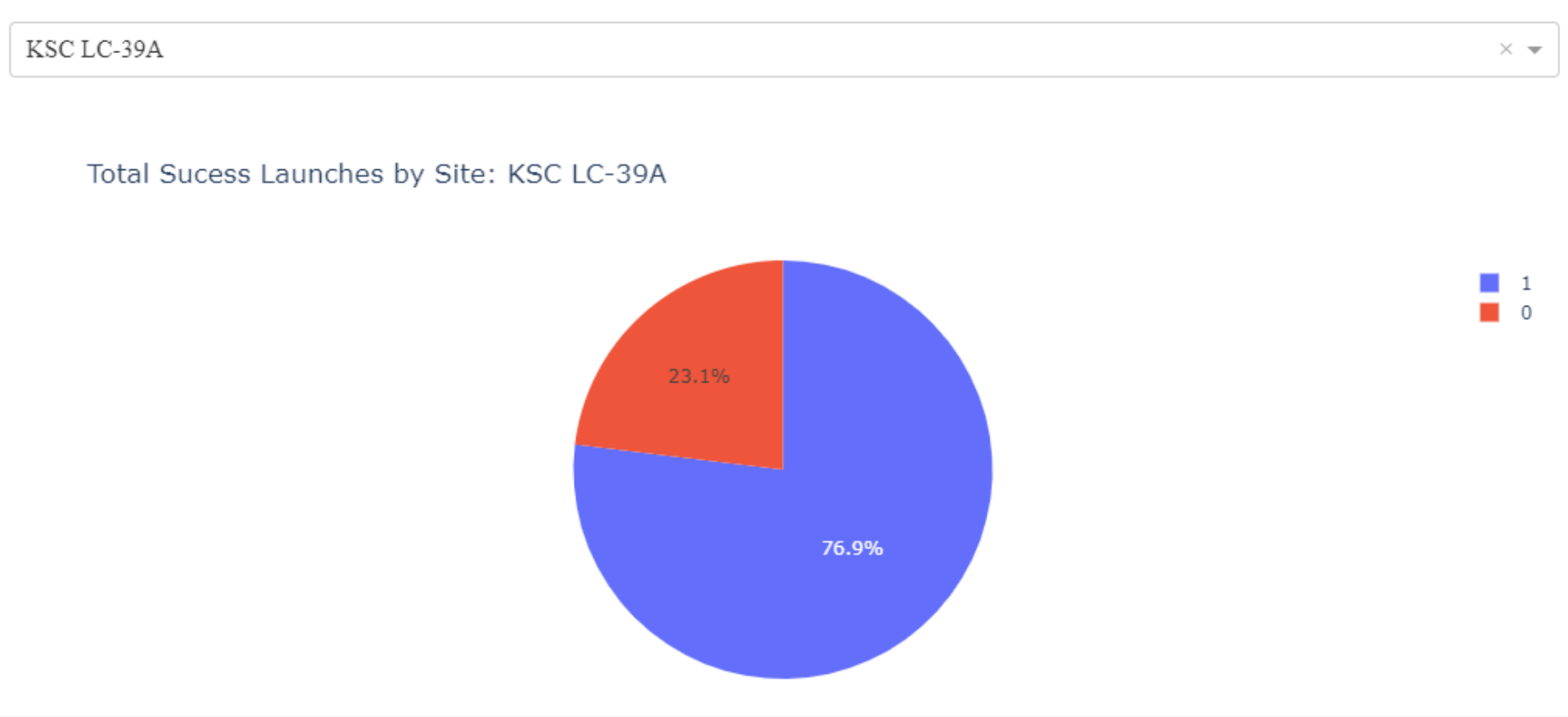
Dashboard: Launch success count for all sites



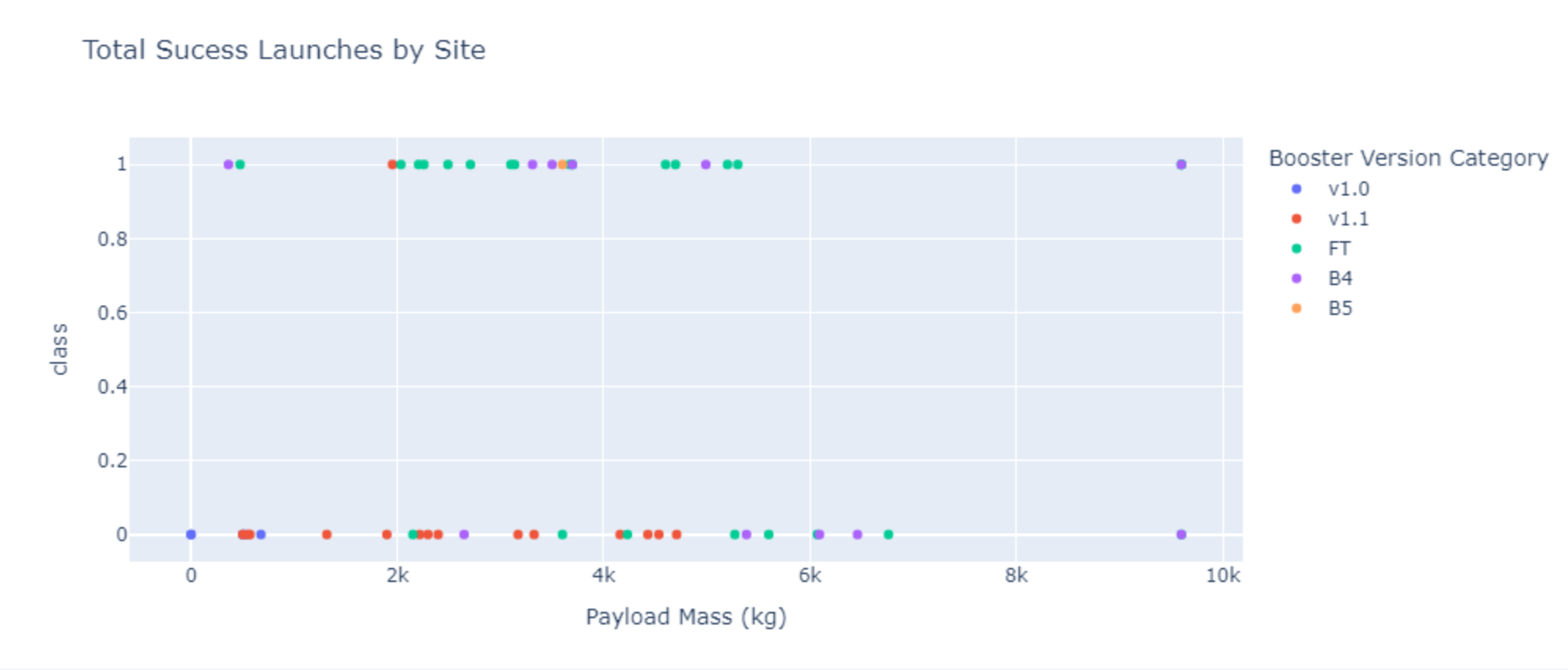
KSC LC-39A Site has the highest success count

Dashboard: Launch site with highest launch success ratio

KSC LC-39A Site has 76.9% success



Dashboard: Payload vs. Launch Outcome scatter plot for all sites



Booster version **FT** sems to have the largest success rate

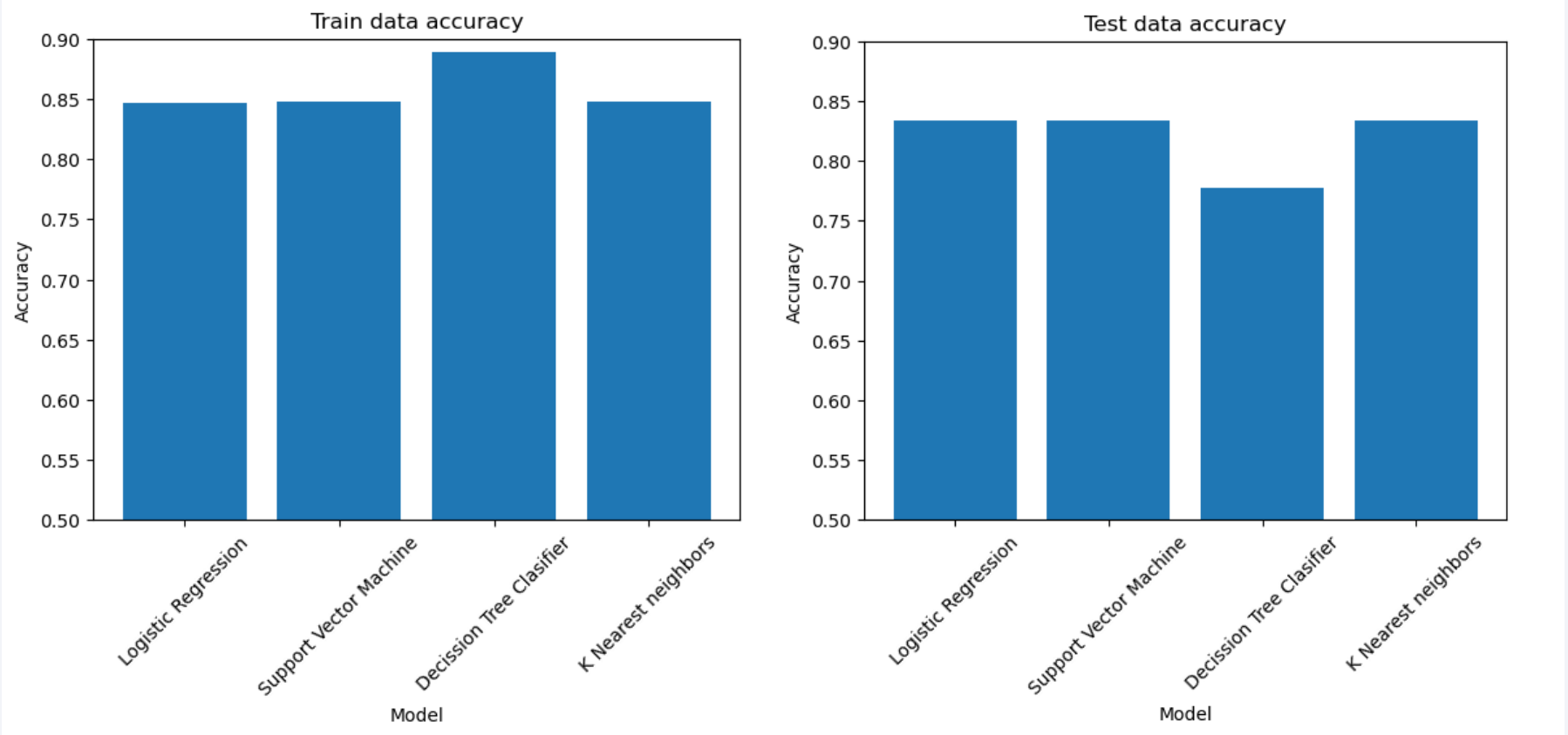
Booster version **v1.1** sems to have the largest failure rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy

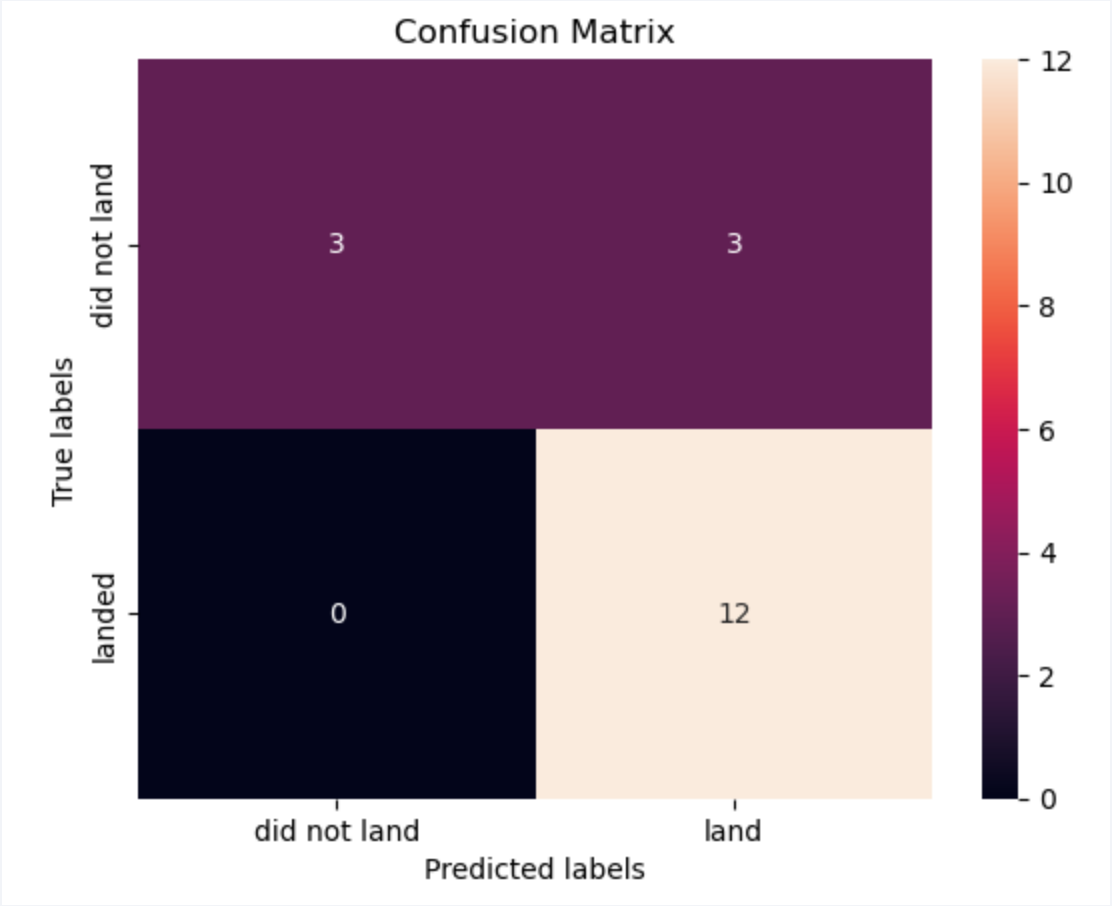
- To train the model, the Decision Tree classifier showed the highest classification accuracy, nevertheless, it has the lowest accuracy for the test data.
- The rest of models have very similar results and consistency within train-test data.



Confusion Matrix

The confusion matrix presented here corresponds to the best performing model, that is, Logistic Regression, Support Vector Machine and K nearest neighbors.

The three models showed the same performance and consistency in accuracy in the train-test data



Conclusions

- There is a correlation between launch site and success rate
- There is a correlation between launch site and booster version
- Payload mass is also associated with the success rate: more payload, less likely the first stage will return
- Most site launches are located in the east coast.
- There has been an increase in the success rate since 2013
- Logistic Regression, Support Vector Machine and K nearest neighbors models are good for predicting the outcome .

Thank you!