IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

R D
May 2022

Outline

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

Executive Summary

- Using Python, we gathered, cleaned, explored, analyzed, and modeled via machine learning SpaceX launch data to determine if the Falcon 9 first stage will land successfully.
- Success varies by **Launch Site**, **Orbit**, **Payload Mass** and **Booster Version**. A **Decision Tree Classifier** has the best prediction outcomes.

Introduction

- SpaceX space launches save money because SpaceX can re-use the first stage.
- If we can predict the whether the first stage will land or not, we can determine the cost of a launch.

Section 1

Methodology

Methodology

Executive Summary

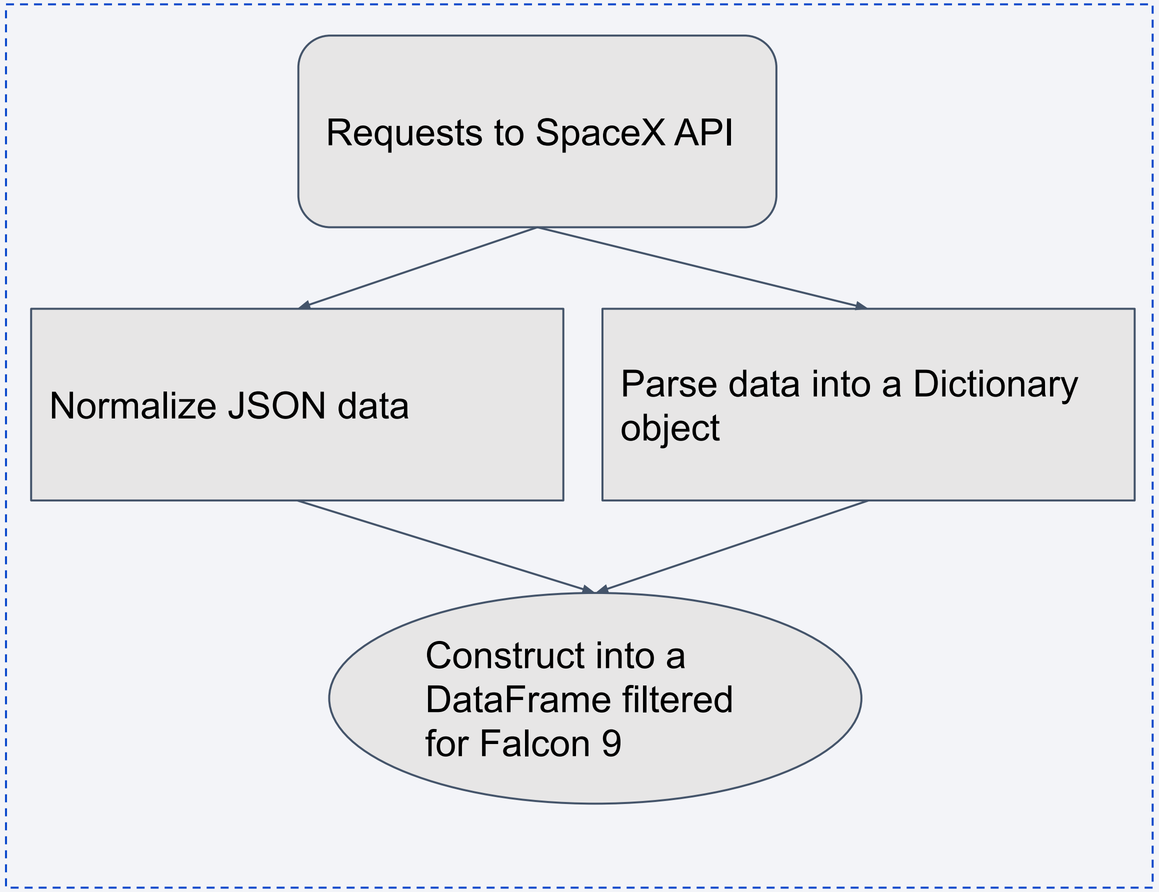
- Data collection methodology:
  - Collected using SpaceX API and web scraping with BeautifulSoup
- Perform data wrangling
  - Cleaned using Python Pandas library
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Datasets were collected making requests to SpaceX API and through web scraping from Wikipedia and the BeautifulSoup library.

Data Collection – SpaceX API

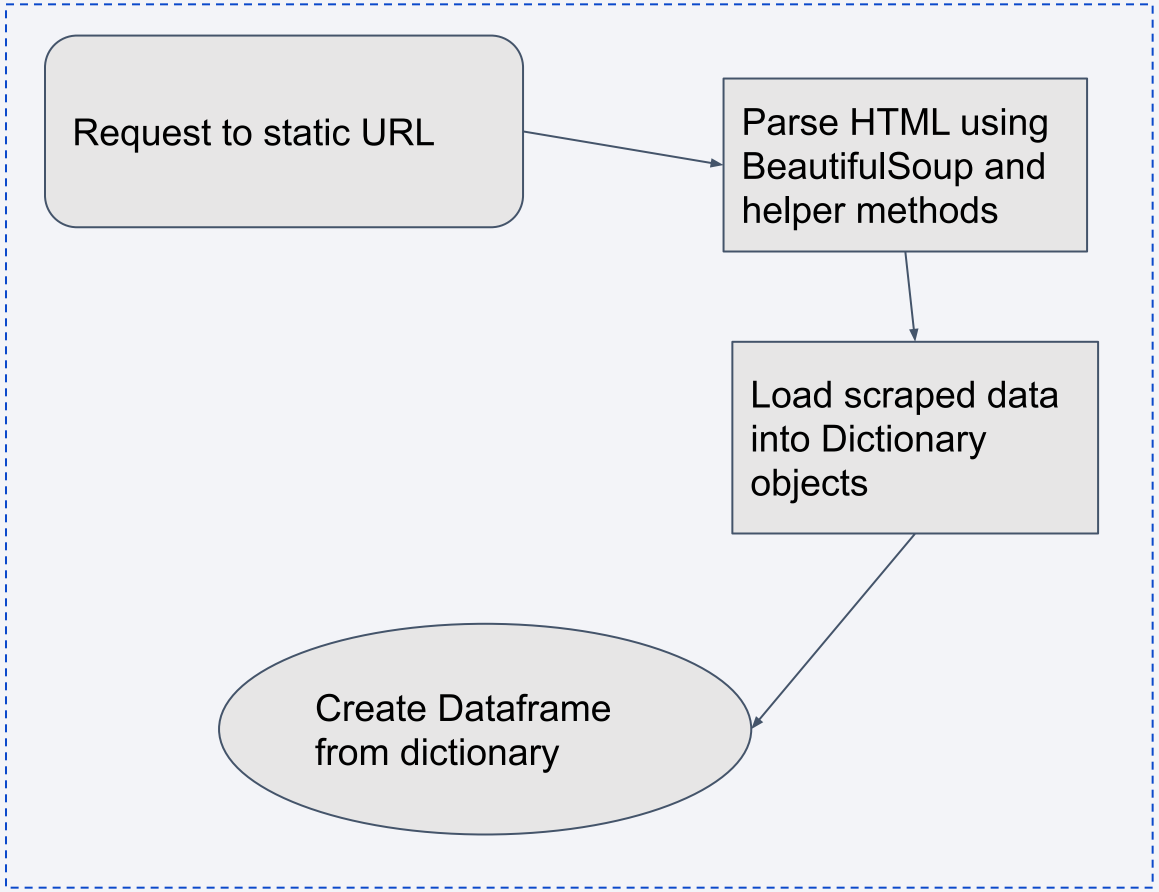
- REST calls to SpaceX API to:
  - get the Booster Version for each launch
  - get the Launch Site
  - get the Payload
  - get Core data (outcome, flight, grid, reused, legs, landing pad)
- Filter, clean up missing data
- Github: [Lab 1: Collecting the data]



Data Collection - Scraping

- Falcon 9 Launch Records scraped from Wikipedia HTML page
- BeautifulSoup and helper functions process scraped data

- Github: [Data Collection with Web Scraping]



Data Wrangling

- Performed Exploratory Data Analysis to find patterns in the data for future prediction models.
- Calculated:
    - the number of launches on each site
    - the number and occurrence of each orbit
    - the number and occurrence of mission outcome per orbit type
    - the landing outcome
- Github: Lab 2: [Data Wrangling]

EDA with Data Visualization

- To explore what makes a successful launch, we visualized the relationships between:
    - Flight Number and Launch Site
    - Payload and Launch Site
    - Success rate and Orbit type
    - Flight Number and Orbit type
    - Payload and Orbit type
- Plotted the yearly success trend
- Github: [Exploring and Preparing Data]

EDA with SQL

- Explored the data using SQL queries to become more familiar with the data
    - Distinct Launch Sites
    - Calculating the Payload Mass by Customer, Booster Version
    - Getting values by date and range
    - Embedding queries and aggregating counts
    - Getting data on successful landing and mission outcomes
- Github: [EDA with SQL]

Note for peer reviewer: Unable to load the data into IBM Cloud DB2, I ran these queries locally using MySQL Workbench.

Build an Interactive Map with Folium

- Marked all Launch Sites to familiarize ourselves with the data
- Added successful/failed launch marker clusters to help understand what makes a successful launch
- Calculated the distance between two coordinates to search for relationships in the data

- Github: [Launch Sites Locations Analysis with Folium]

Note for peer reviewer: While my maps render locally in Jupyter Notebook, they don't render on Github for security reasons. You'll have to download, run on Jupyter notebook, and trust the notebook to view in your browser. (Or just believe my code!)

Build a Dashboard with Plotly Dash

- Pie chart of Successes/Failures by Launch Site
- Scatter plot of success (Class) by Booster Version, Payload Mass

- Can we see what makes for a successful launch?

- Github: SpaceX Dash App (python code)

Predictive Analysis (Classification)

- Through one hot feature encoding of the data, created models to test/train.
- Then, we trained and tested different models and evaluated their scores:
  - Logistic regression
  - Support vector machine
  - K Nearest Neighbors
  - Decision Tree
- Github: Machine Learning Prediction

Results: Exploratory Data Analysis

- How many launches are on each site?

Use the method `value_counts()` on the column `LaunchSite` to determine the number of launches on each site:

In [5]:
```
# Apply value_counts() on column LaunchSite
df['LaunchSite'].value_counts()
```

Out[5]:
```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

Results: Exploratory Data Analysis

- What's the mission outcome per orbit type?

```python
# landing_outcomes = values on Outcome column
landing_outcomes = df['Outcome'].value_counts()
landing_outcomes
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: Outcome, dtype: int64
```

Results: Exploratory Data Analysis

- We create a new column Class to capture the outcome (1 for success, 0 for failure) of each launch.

```
bad_outcomes=set(landing_outcomes.keys()[[1,3,5,6,7]])
bad_outcomes
```

{'False ASDS', 'False Ocean', 'False RTLS', 'None ASDS', 'None None'}

TASK 4: Create a landing outcome label from Outcome column

Using the `Outcome`, create a list where the element is zero if the corresponding row in `Outcome` is in the set `bad_outcome`; otherwise, it's one. Then assign it to the variable `landing_class`:

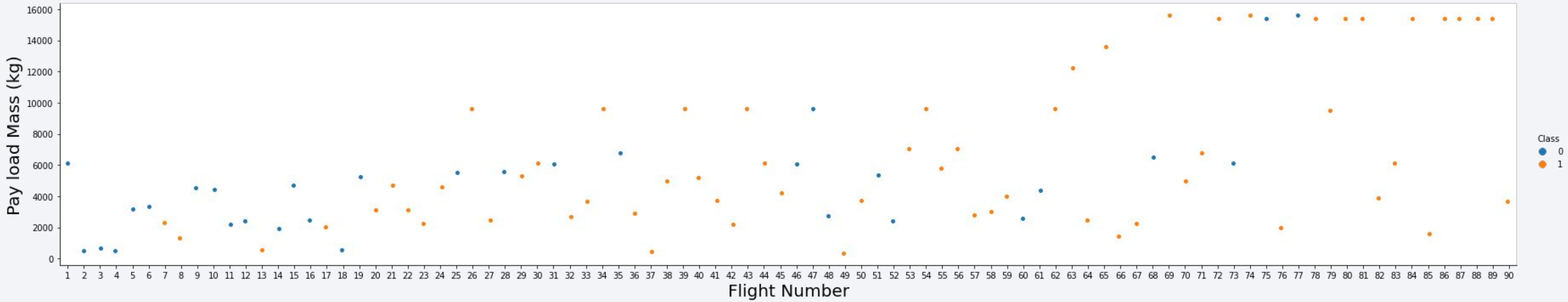
```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise
landing_class = []
for i in df['Outcome']:
    if i in set(bad_outcomes):
        landing_class.append(0)
    else:
        landing_class.append(1)
```

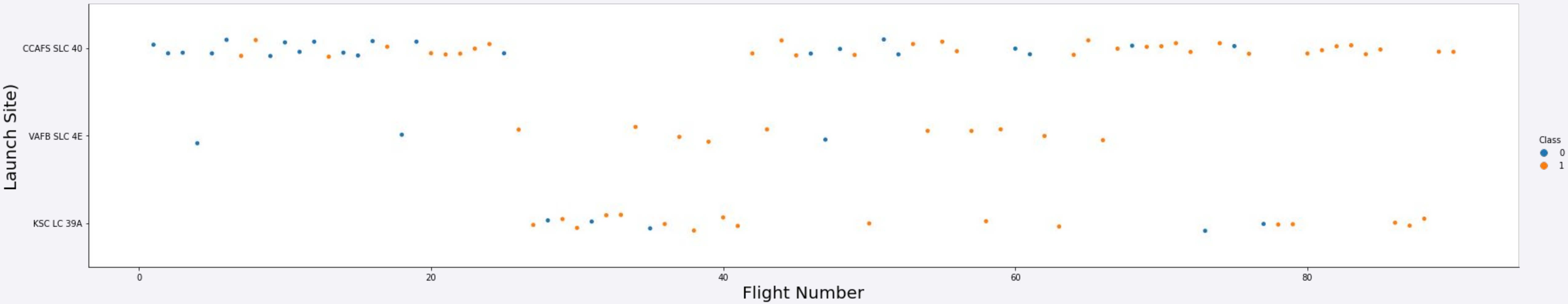
This variable will represent the classification variable that represents the outcome of each launch. If the value is zero, the first stage did not land successfully; one means the first stage landed Successfully

```
df['Class']=landing_class
```

Results: Exploratory Data Analysis

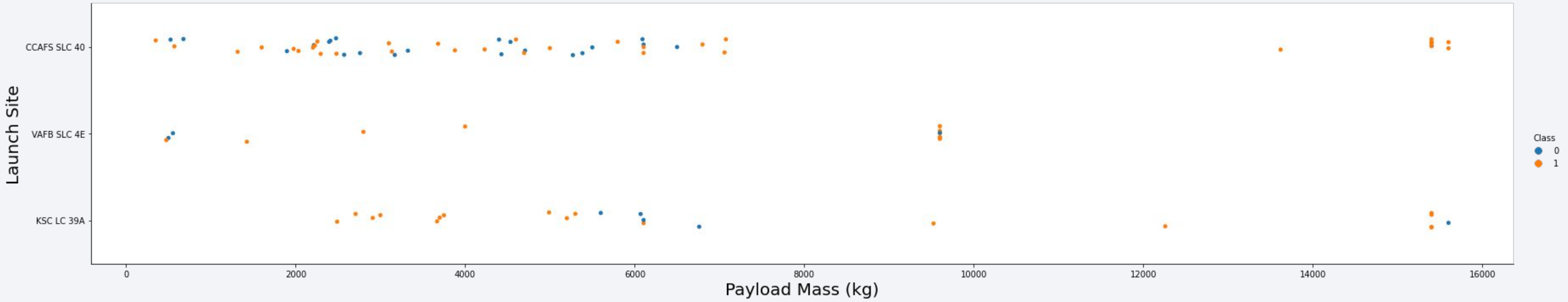
- Now we can visualize the Class against other columns.

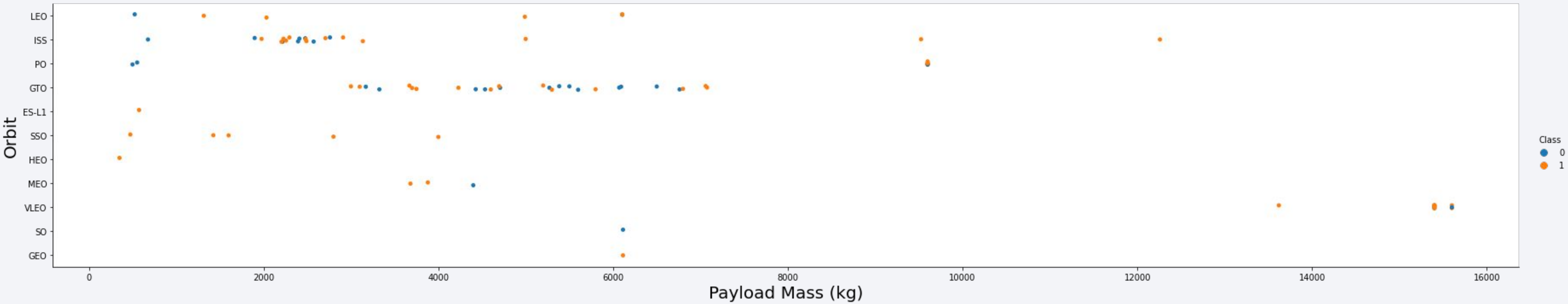




Results: Exploratory Data Analysis

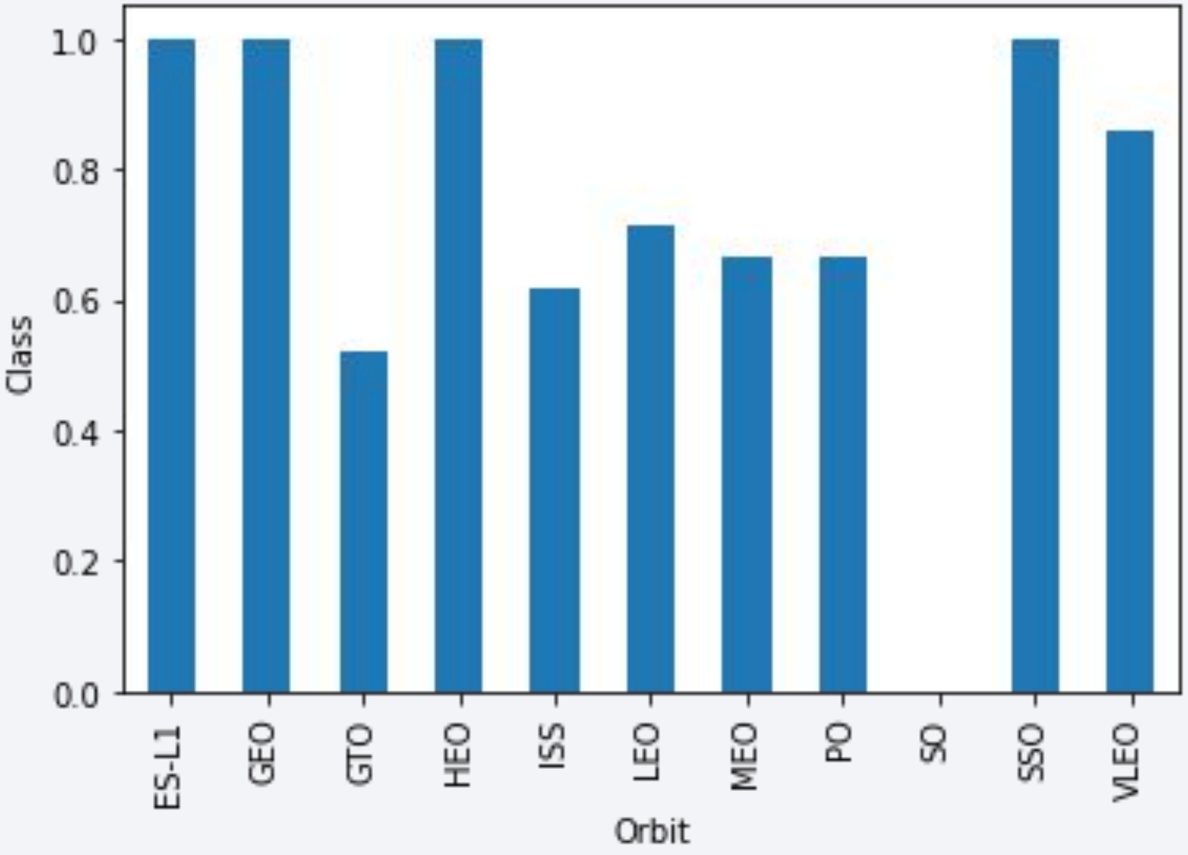
- Now we can visualize the Class against other columns.

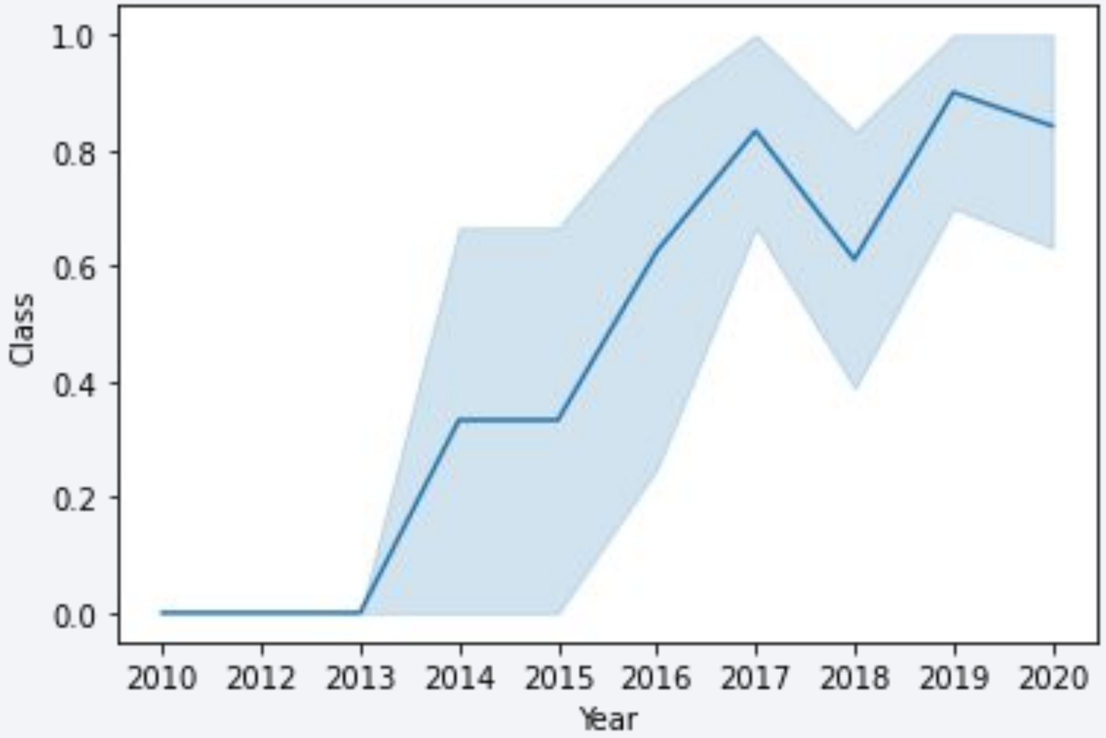




Results: Exploratory Data Analysis

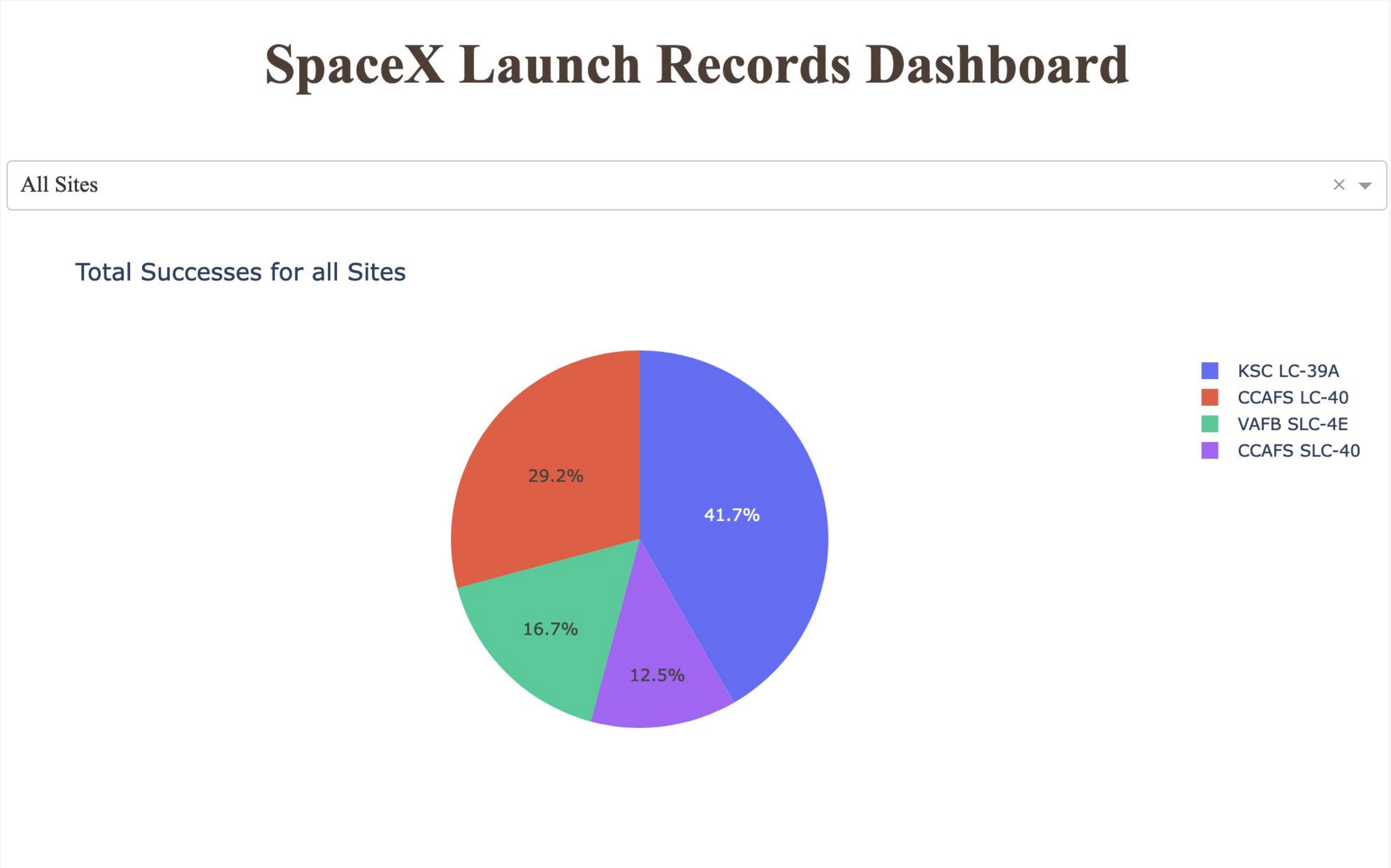
- Now we can visualize the Class against other columns.





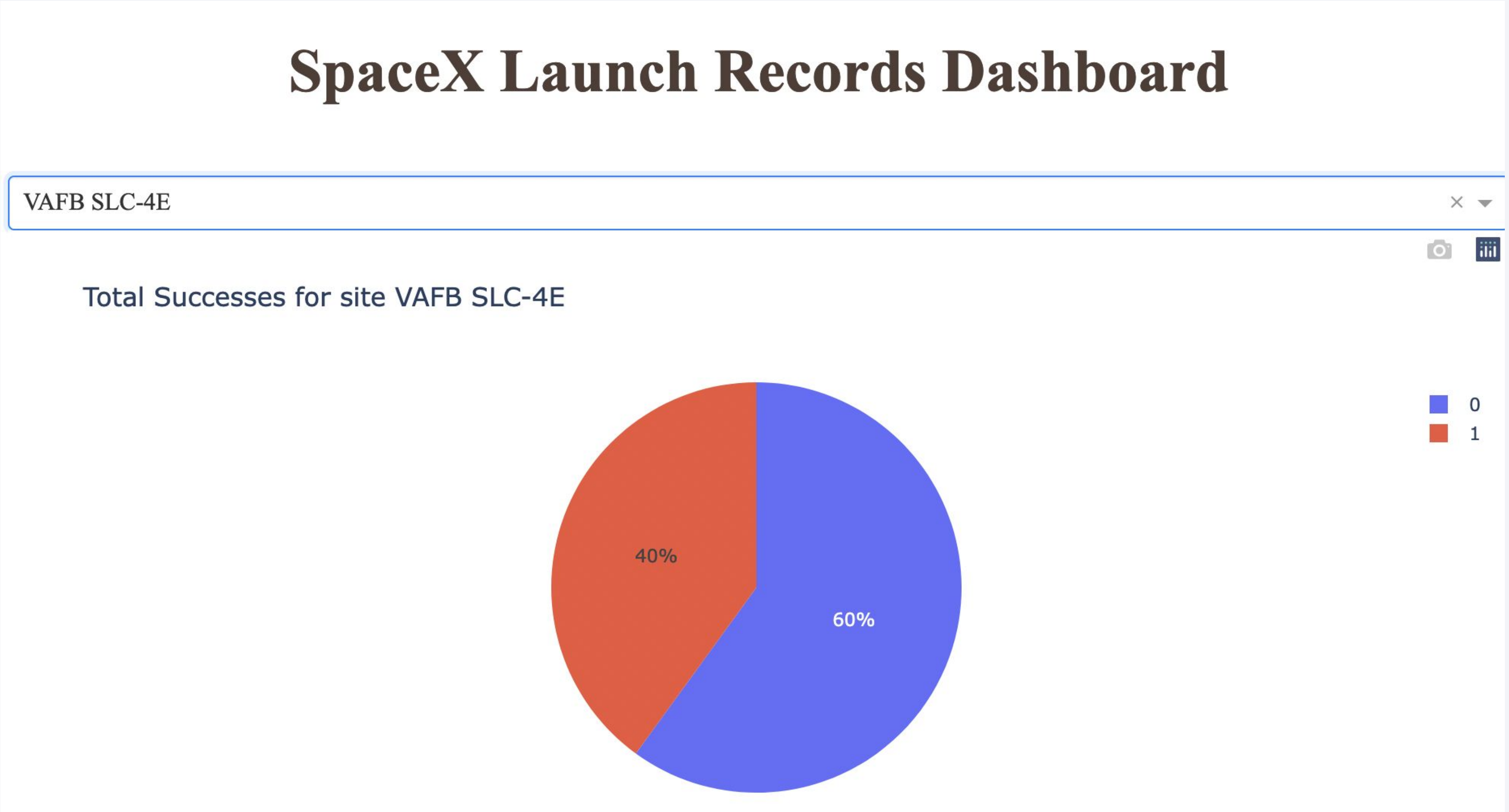
Results: Interactive analytics demo

- Success for all Launch Sites



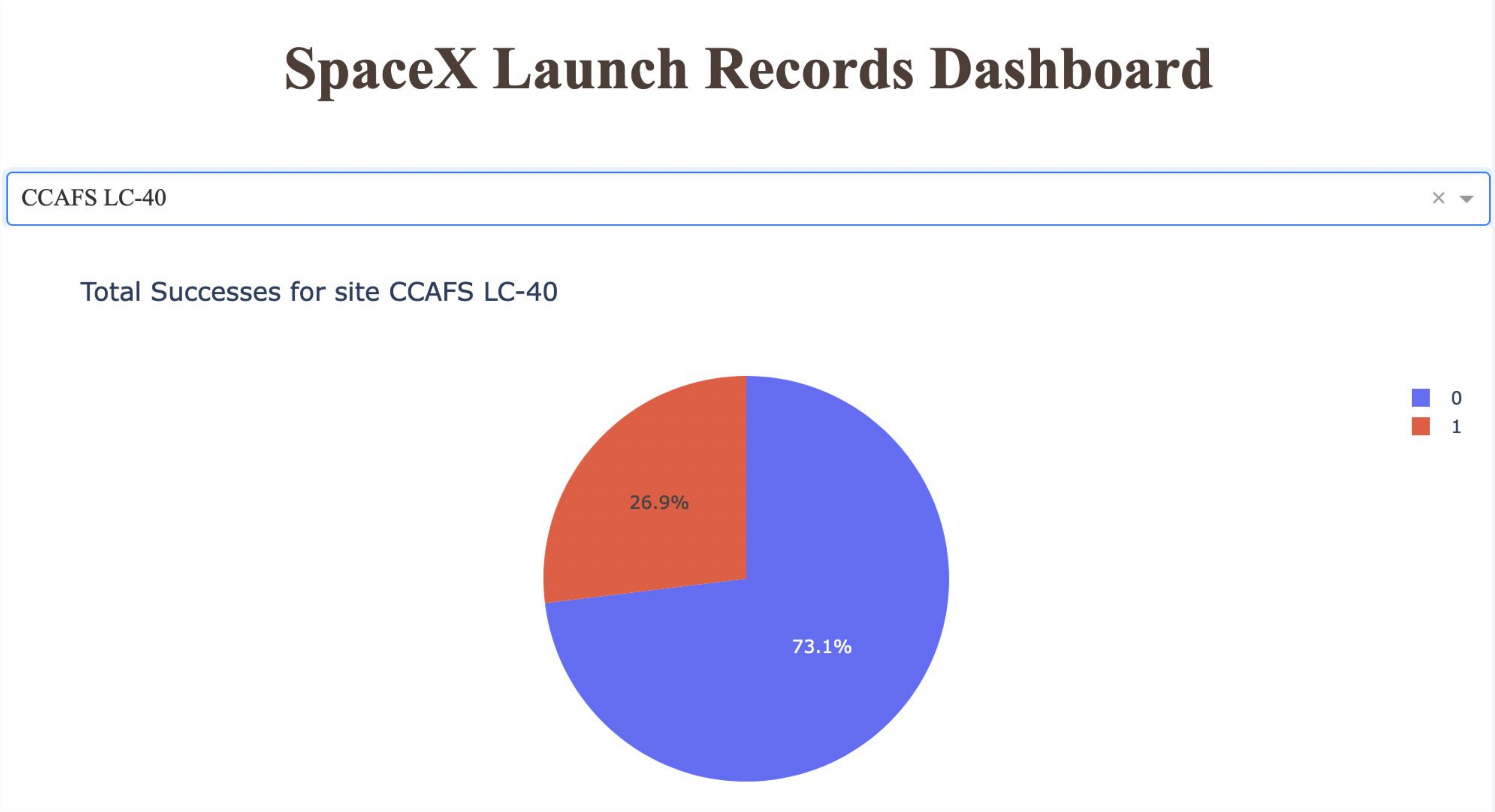
Results: Interactive analytics demo

- Success vs Failure for each Launch Site



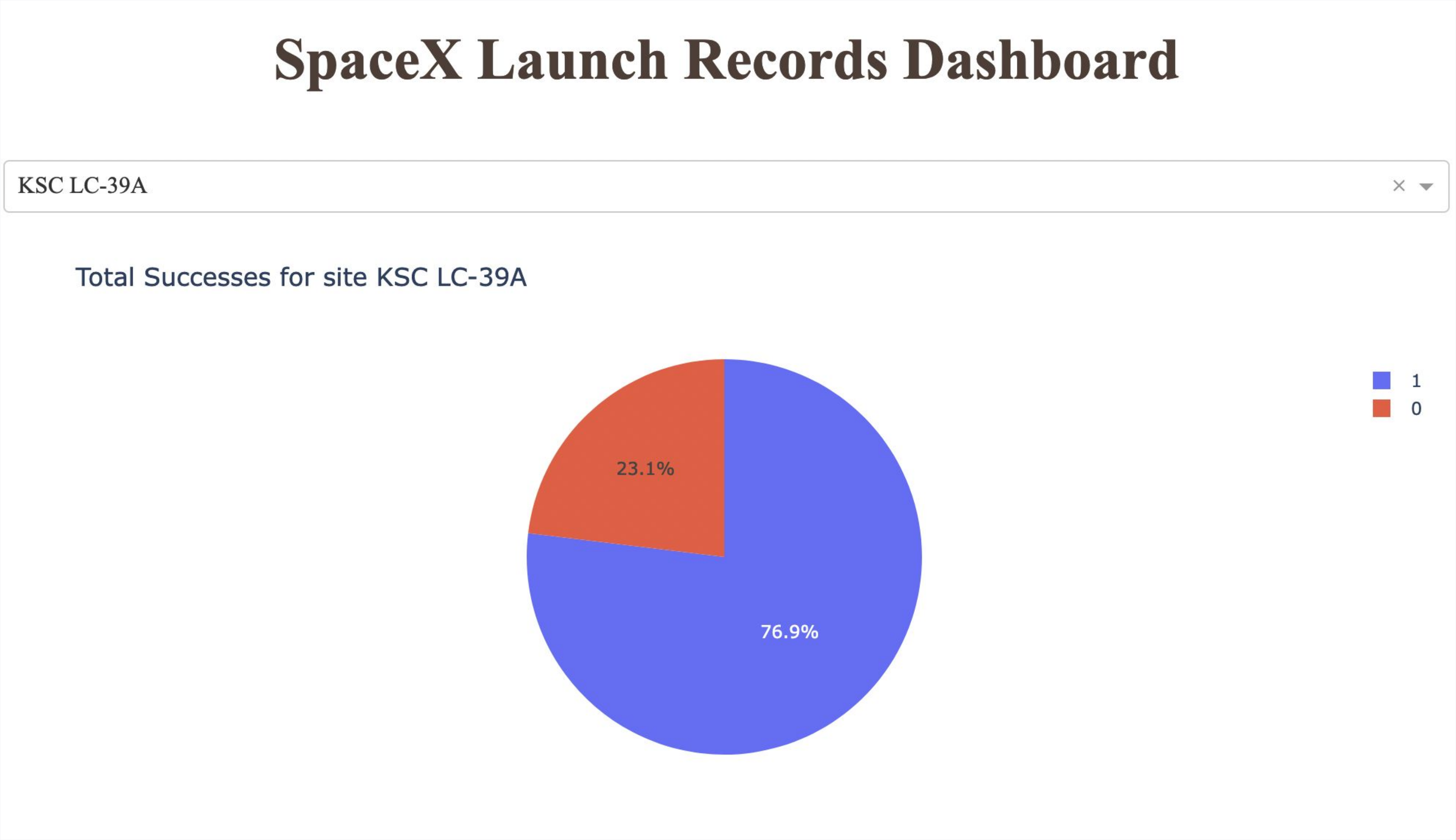
Results: Interactive analytics demo

- Success vs Failure for each Launch Site



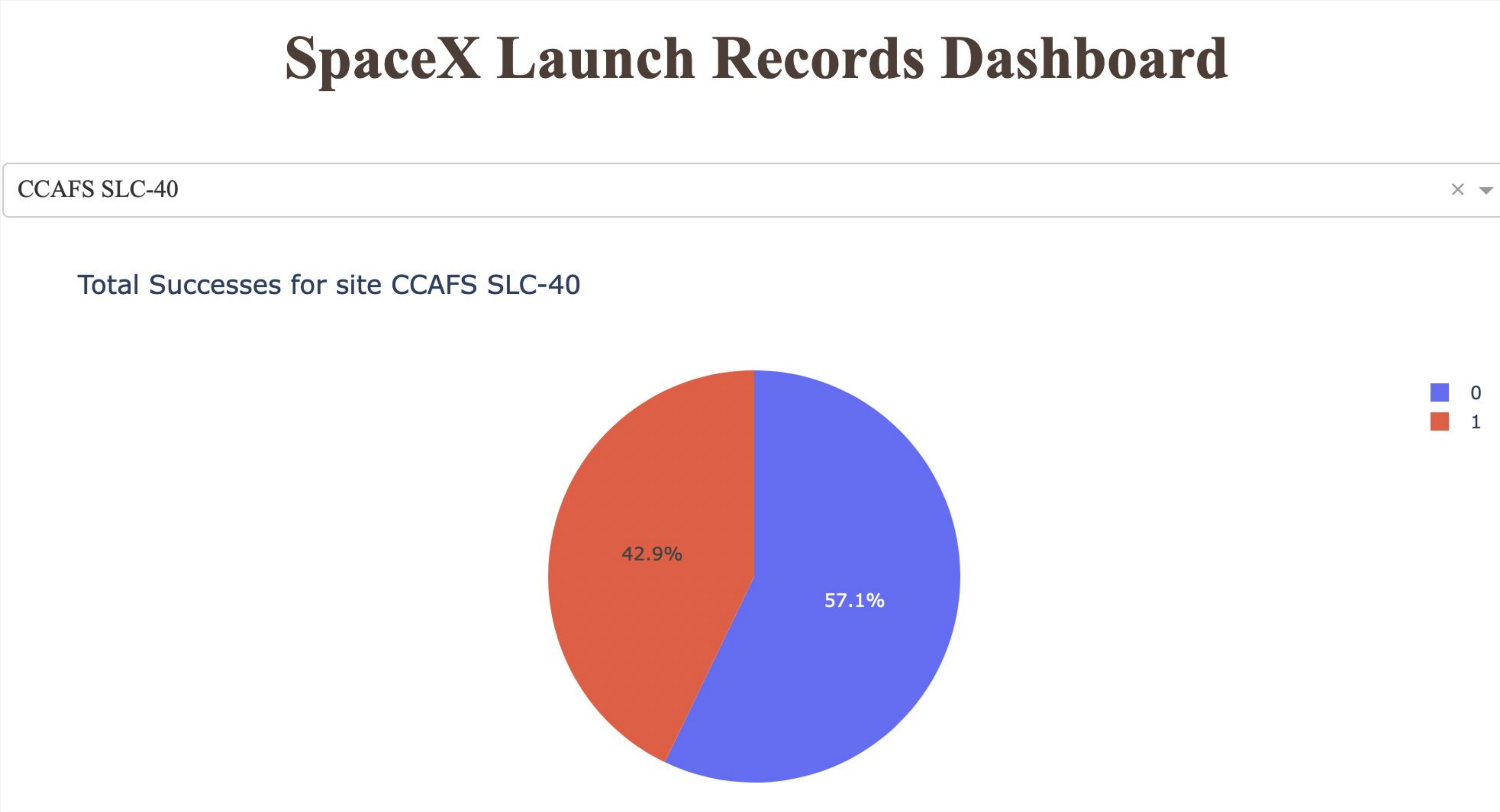
Results: Interactive analytics demo

- Success vs Failure for each Launch Site



Results: Interactive analytics demo

- Success vs Failure for each Launch Site



Results: Interactive analytics demo

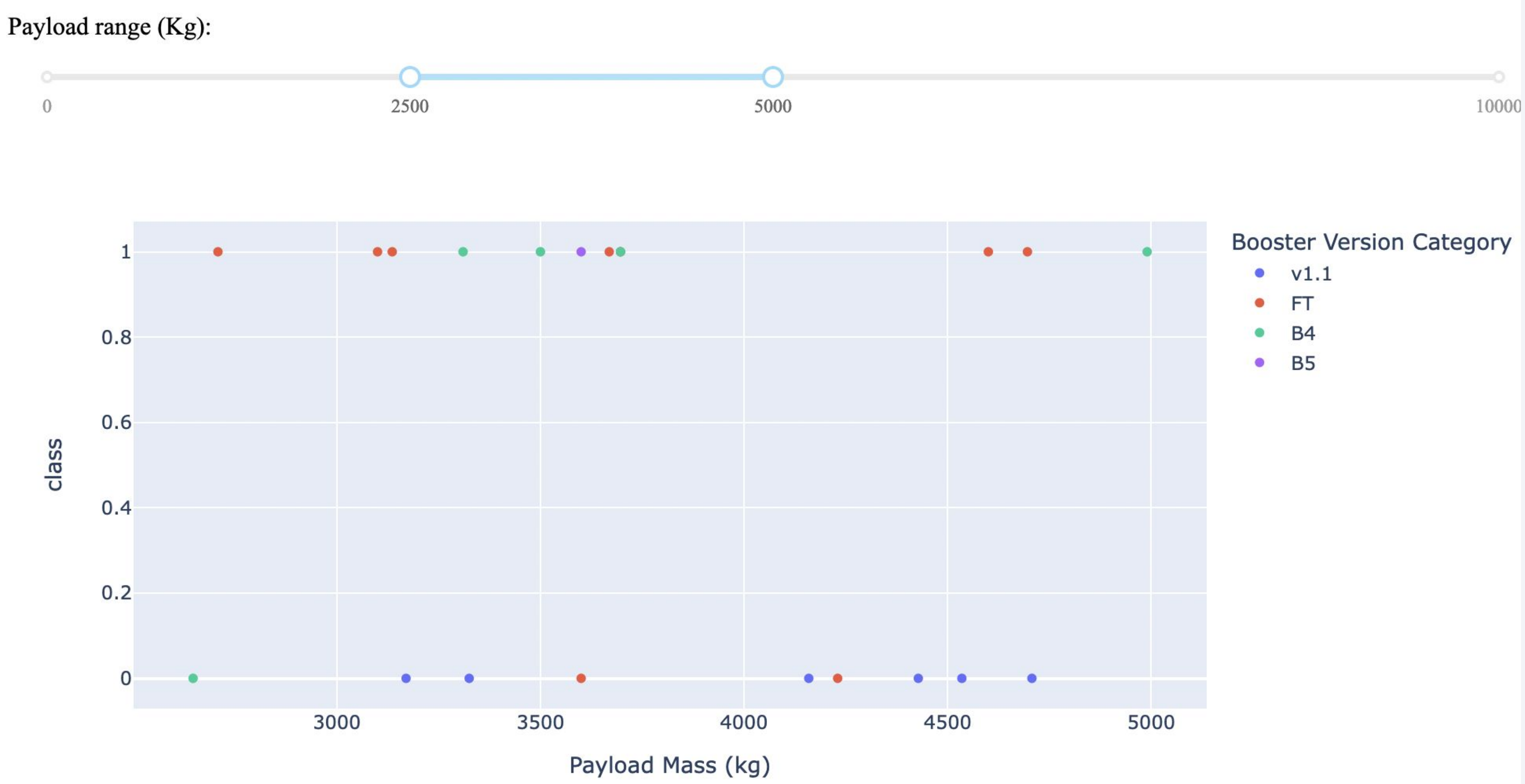
- Class by Launch Site, Payload Mass, Booster Version, Payload range (**all**)

Results: Interactive analytics demo

- Class by Launch Site, Payload Mass, Booster Version, Payload range (**KSC LC-39A Launch Site only**)

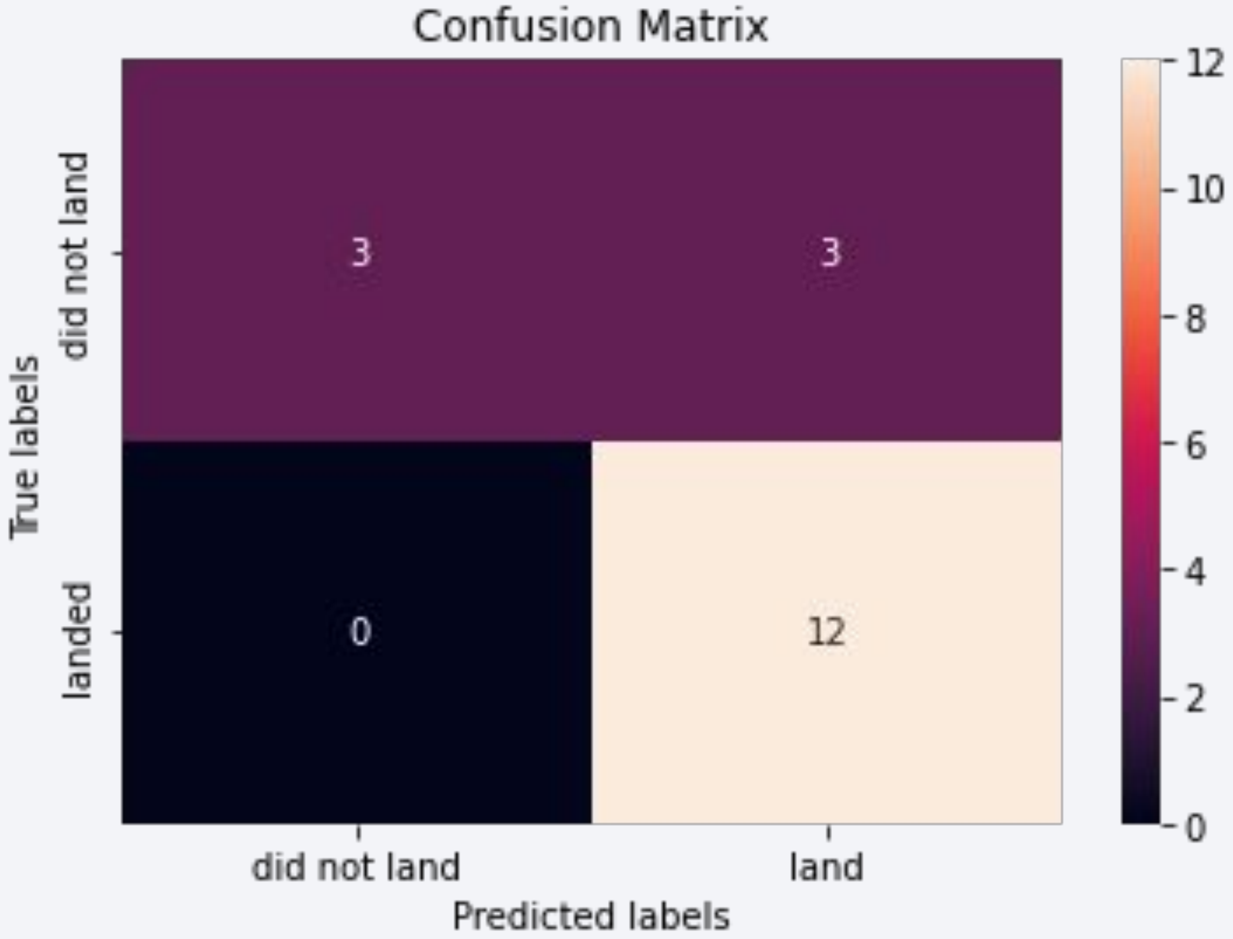
Results: Interactive analytics demo

- Class by Launch Site, Payload Mass, Booster Version, Payload range (**2500-5000 kg**)



Results: Predictive Analytics

- The Confusion Matrix was near-constant across Machine Learning models



Results: Predictive Analytics

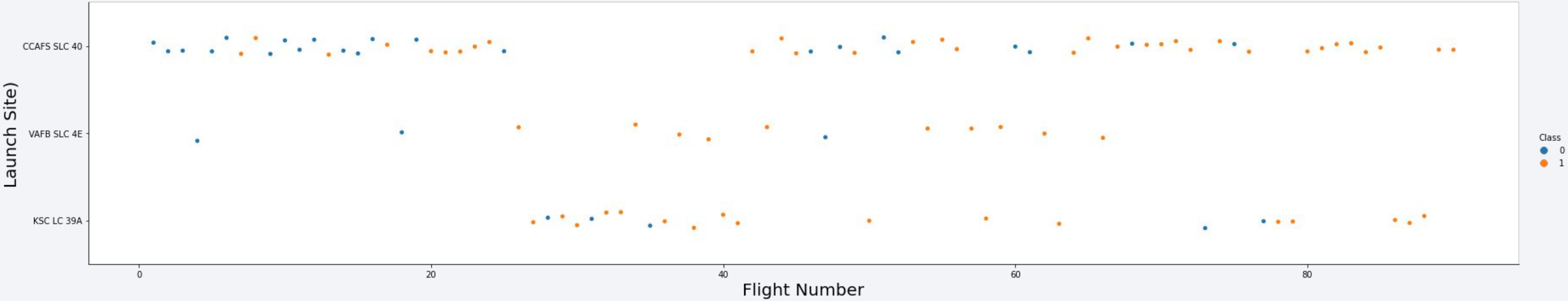
- The **Decision Tree** had the best accuracy score.

```
logreg_cv :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
logreg_cv accuracy : 0.8464285714285713

svm_cv :(best parameters)  {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
svm accuracy : 0.8482142857142856

knn_cv :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
knn_cv accuracy : 0.8482142857142858

tree_cv :(best parameters)  {'criterion': 'entropy', 'max_depth': 16, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'splitter': 'random'}
tree_cv accuracy : 0.8875
```

Section 2

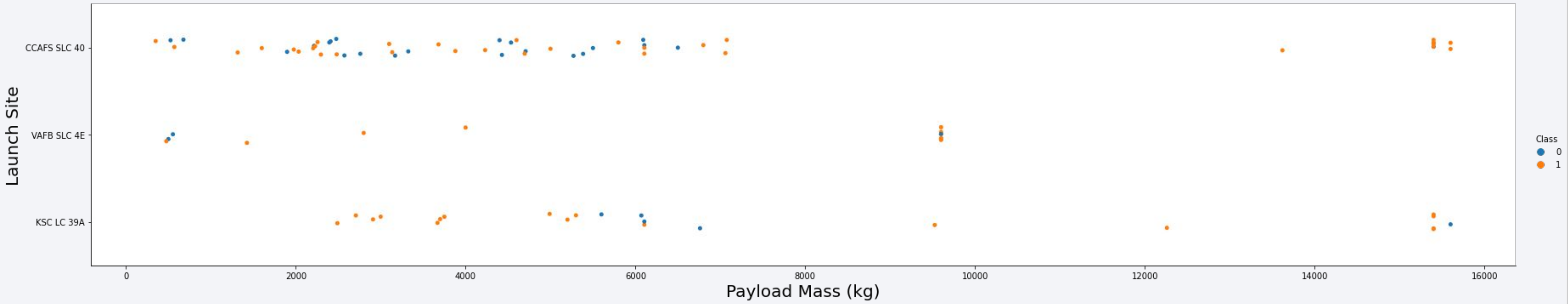
Insights drawn from EDA

Flight Number vs. Launch Site



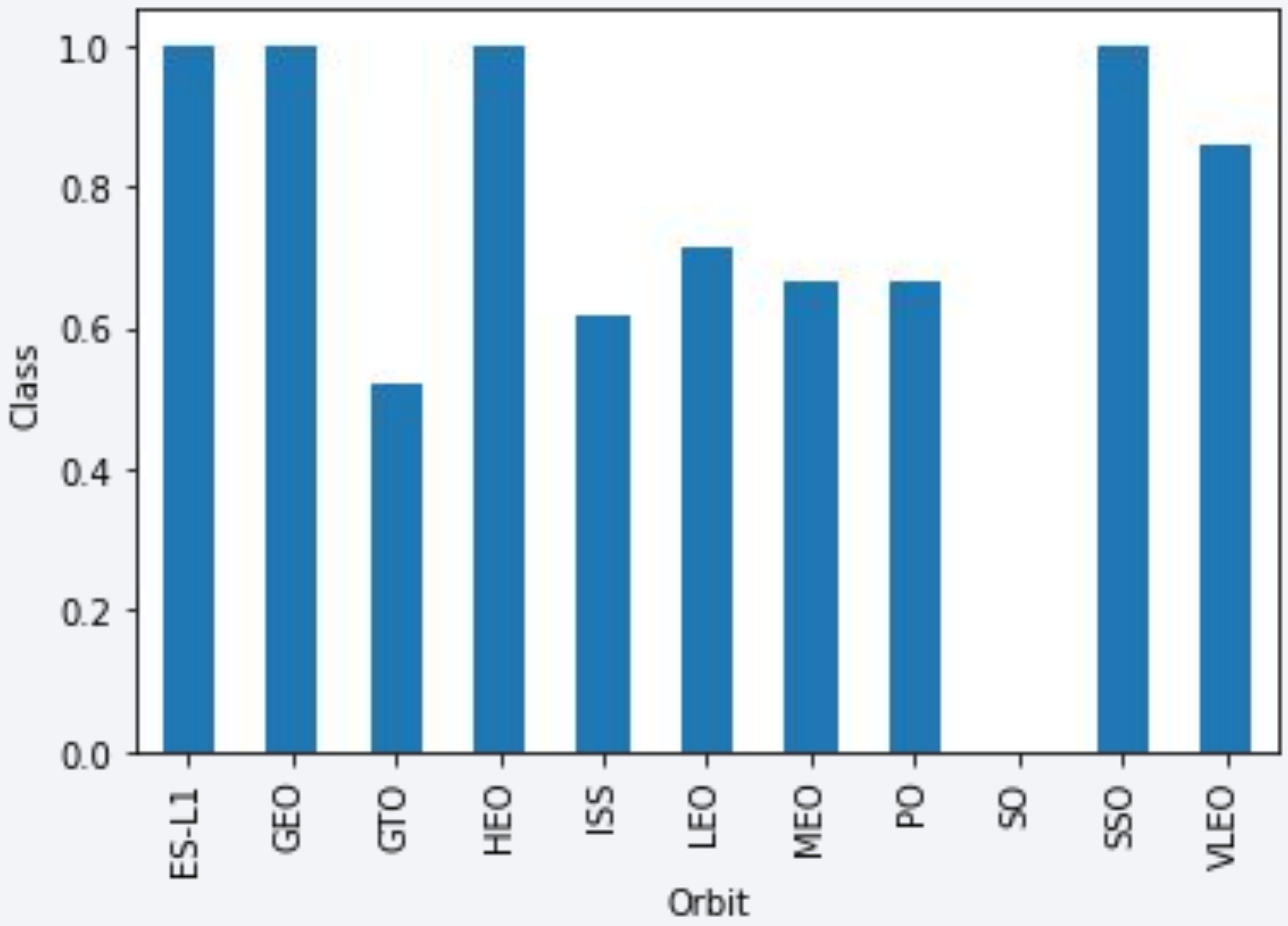
Later Flight Numbers have more success as launches become more successful over time.

Payload vs. Launch Site



The highest Payloads have more successes but fewer launches, though most successes occur between 2000-6000 kgs.

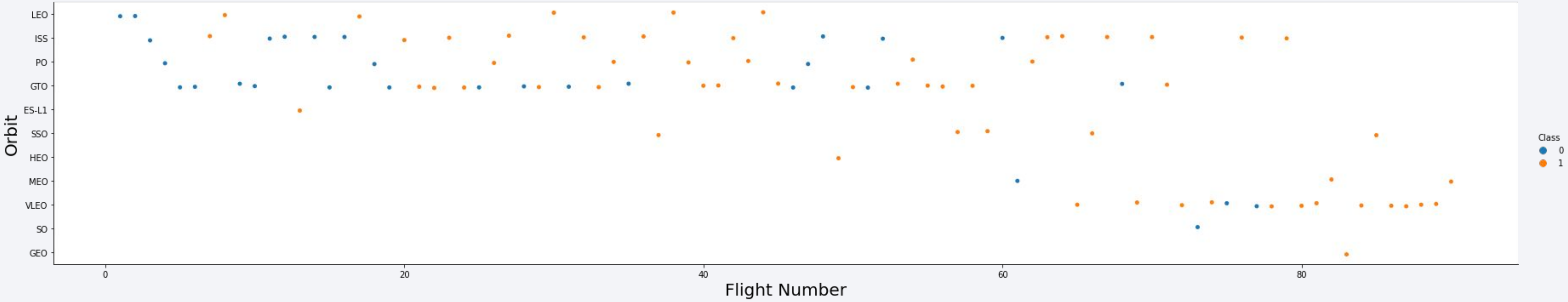
Success Rate vs. Orbit Type



The SO Orbit type has no successes, while the most successful Orbit Types are:

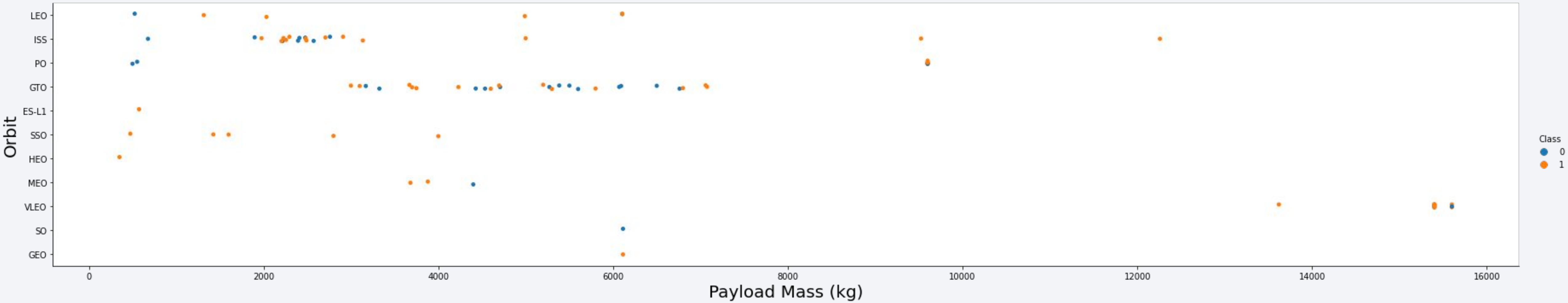
- ES-L1
- GEO
- HEO
- SSO

Flight Number vs. Orbit Type



There are more successes with later Flight Numbers, with new orbit types used as time goes on.

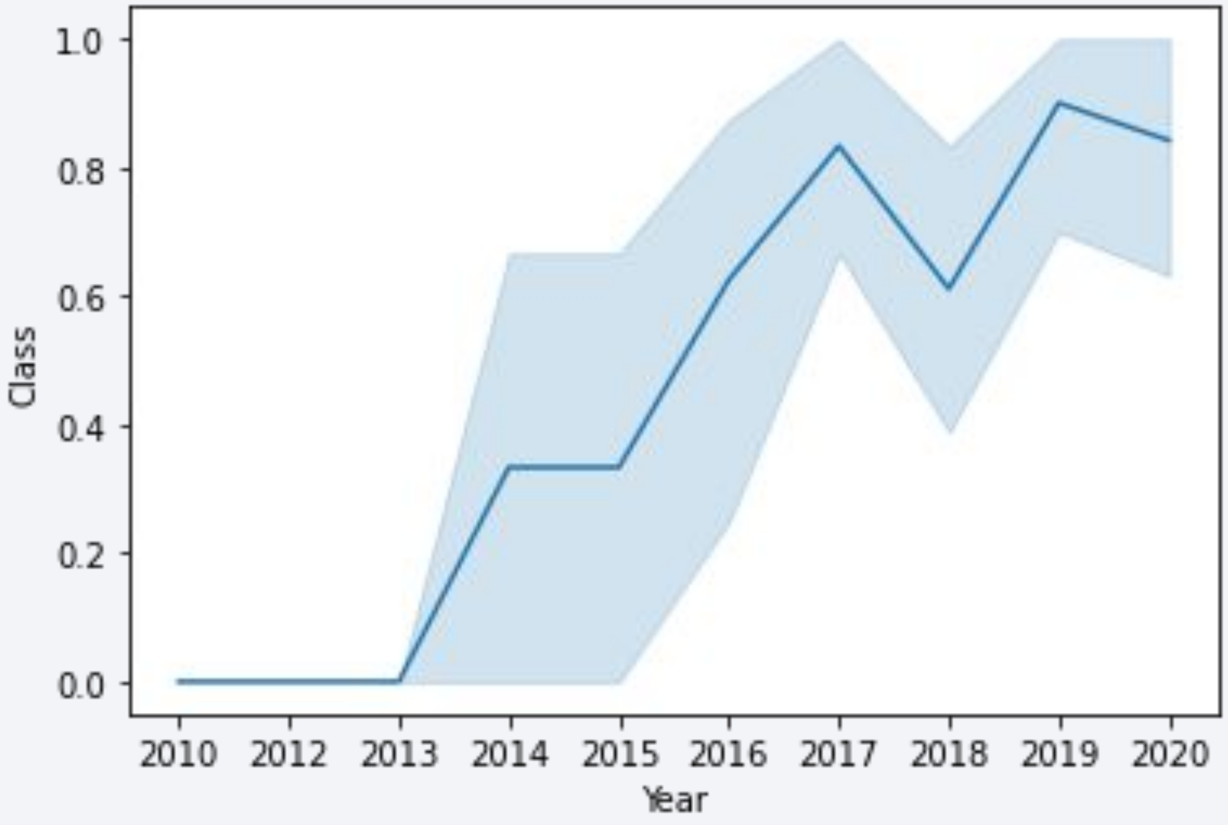
Payload vs. Orbit Type



With heavy payloads, PO, LEO and ISS have more success. ES-L1, SSO and HEO have more success with lighter payloads.

However for GTO we cannot distinguish this well as there's quite a mix.

Launch Success Yearly Trend



With more experience and more data, over time the success rate increases each year.

All Launch Site Names

There are 4 Launch Sites in the Falcon 9 data.

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

Launch Site Names Begin with 'KSC'

While all missions were successful, launch site KSC LC-39A hosts several different customers with varying orbits, payloads, and landing outcomes.

Date	Time_UTC	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing__Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

- Aggregated the SUM of all **Payload Mass (kg)** values for the ***NASA (CRS)*** **Customer**.
- 45,596 kg

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
18-04-2014	19:25:00	F9 v1.1	CCAFS LC-40	SpaceX CRS-3	2296	LEO (ISS)	NASA (CRS)	Success	Controlled (ocean)
21-09-2014	05:52:00	F9 v1.1 B1010	CCAFS LC-40	SpaceX CRS-4	2216	LEO (ISS)	NASA (CRS)	Success	Uncontrolled (ocean)
10-01-2015	09:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)

Average Payload Mass by F9 v1.1

- **Booster Version *F9 v1.1*** did not carry the highest payload masses.
- Average: 2928.4 kg

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
03-12-2013	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt
06-01-2014	22:06:00	F9 v1.1	CCAFS LC-40	Thaicom 6	3325	GTO	Thaicom	Success	No attempt
18-04-2014	19:25:00	F9 v1.1	CCAFS LC-40	SpaceX CRS-3	2296	LEO (ISS)	NASA (CRS)	Success	Controlled (ocean)
14-07-2014	15:15:00	F9 v1.1	CCAFS LC-40	OG2 Mission 1 6 Orbcomm-OG2 satellites	1316	LEO	Orbcomm	Success	Controlled (ocean)
05-08-2014	08:00:00	F9 v1.1	CCAFS LC-40	AsiaSat 8	4535	GTO	AsiaSat	Success	No attempt

First Successful Ground Landing Date

- The first successful groundpad landing date was **December 22, 2015**.
- In **2017** there were several successful groundpad landings.

Month_2017	Landing__Outcome	Booster_Version	Launch_site
February	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
May	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
June	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
August	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
September	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
December	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

Successful Drone Ship Landing with Payload between 4000 and 6000

- Only 4 different **Booster Versions** have successfully landed a drone ship with a **Payload Mass (kg)** between 4000 and 6000 kg.

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

Total Number of Successful and Failure Mission Outcomes

- While most missions were successful, **Landing Outcomes** were mixed.

Landing _Outcome	Mission_Outcome	Count
Success	Success	38
No attempt	Success	20
Success (drone ship)	Success	14
Success (ground pad)	Success	8
Controlled (ocean)	Success	5
Failure (drone ship)	Success	5
Failure	Success	3
Failure (parachute)	Success	2
Uncontrolled (ocean)	Success	2
No attempt	Success	1
No attempt	Success	1
Precluded (drone ship)	Failure (in flight)	1
Success (ground pad)	Success (payload status unclear)	1

Total Number of Successful and Failure Mission Outcomes

- While most missions were successful, **Landing Outcomes** were mixed.

Mission_Outcome	Count
Success	99
Failure (in flight)	1
Success (payload status unclear)	1

Landing _Outcome	Count
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Controlled (ocean)	5
Failure (drone ship)	5
Failure	3
Failure (parachute)	2
Uncontrolled (ocean)	2
No attempt	1
Precluded (drone ship)	1

Boosters Carried Maximum Payload

- Only X **Booster Versions** have carried the maximum payload of ***9,600 kg***.

Booster_Version
F9 FT B1029.1
F9 FT B1036.1
F9 B4 B1041.1
F9 FT B1036.2
F9 B4 B1041.2
F9 B5B1048.1
F9 B5 B1049.2

2017 Launch Records

- As we've already reviewed, 2017 had several successful **Landing Outcomes** by the ***grounding pad***.

Month_2017	Landing__Outcome	Booster_Version	Launch_site
February	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
May	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
June	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
August	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
September	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
December	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

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

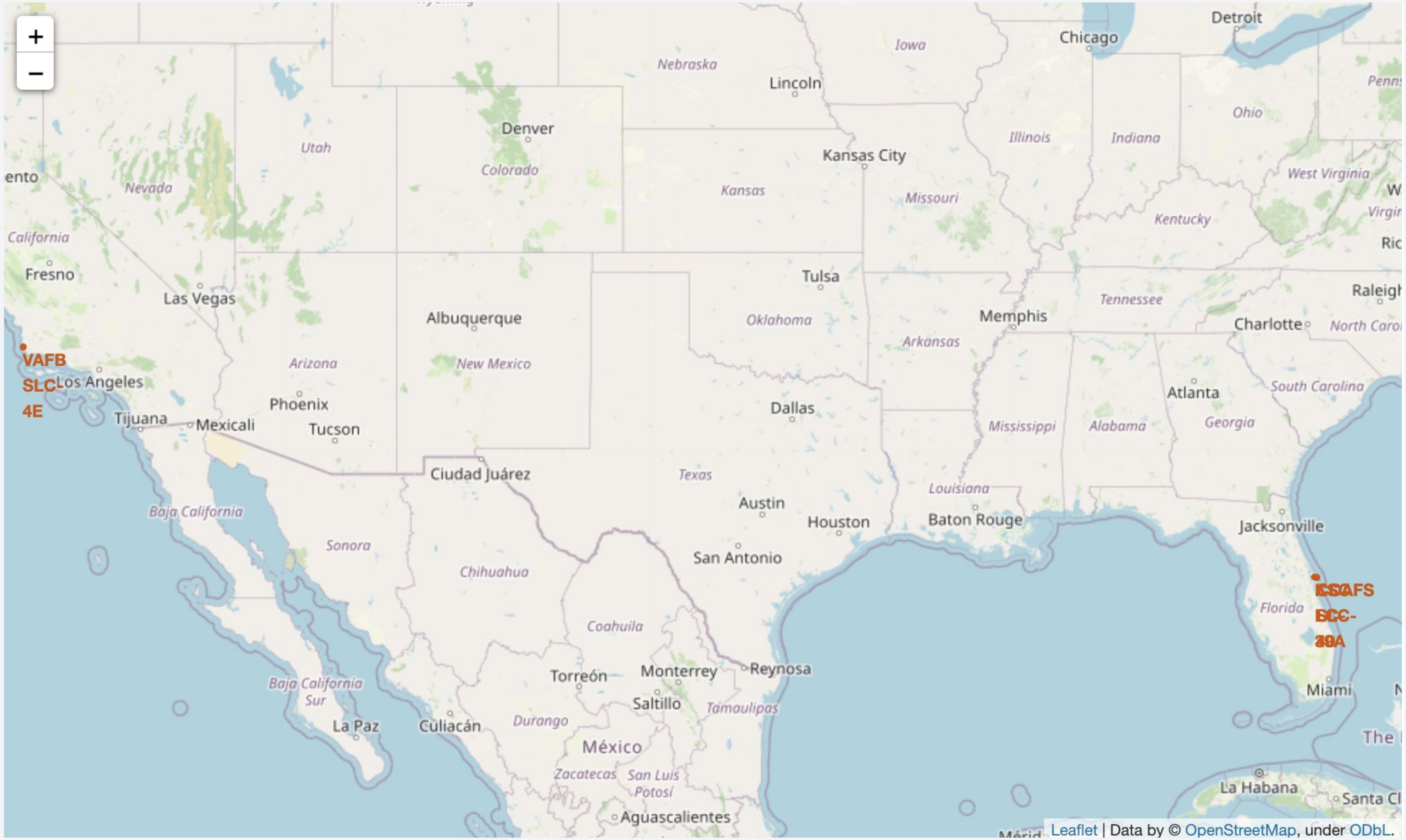
- Between June 4 2010 and March 20 2017, most successful **Mission Outcomes** had ***no landing attempt*** or were done by ***drone***.

# Landing__Outcome	Count of Successes
No attempt	11
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Uncontrolled (ocean)	2

Section 3

Launch Sites Proximities Analysis

All Falcon 9 Launch Sites



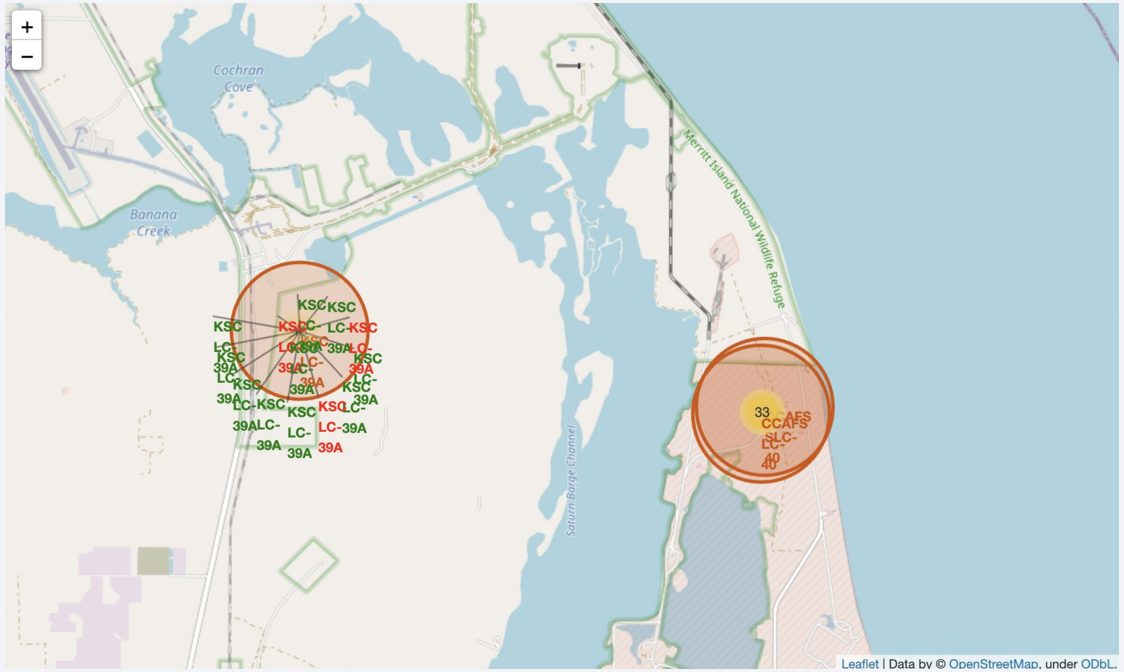
As we’ve seen on an earlier slide, there are 4 **Launch Sites** in our data set. We’ve marked and labelled each of them on a folium map.

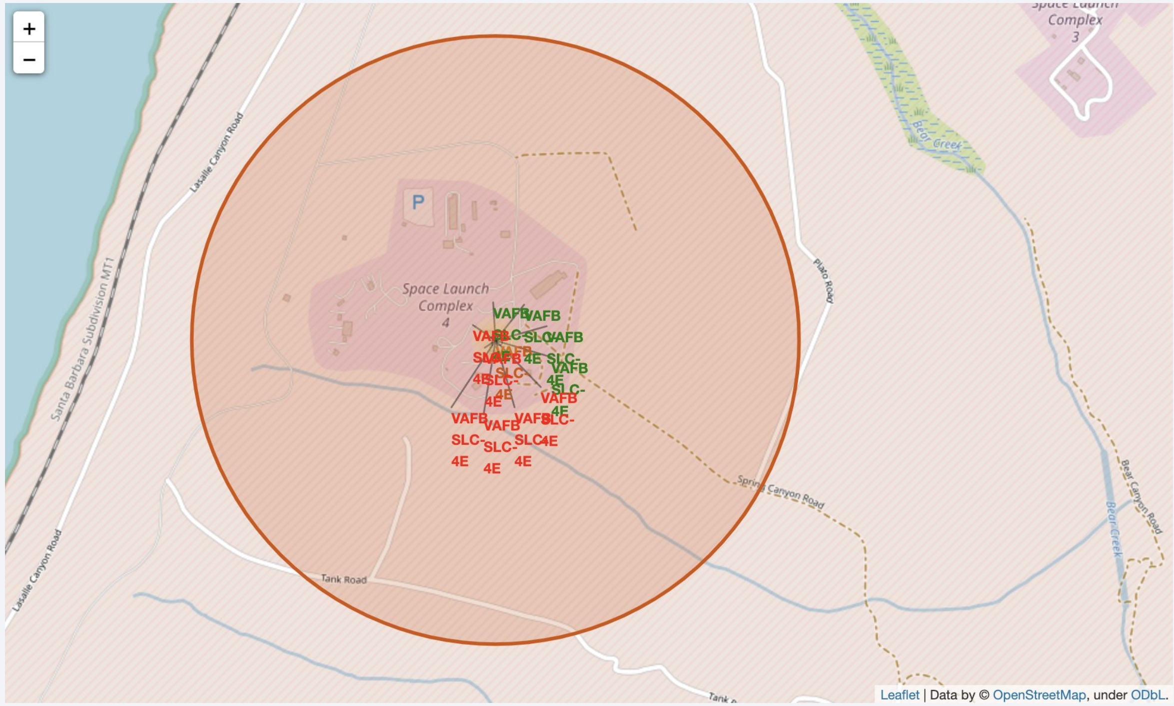
Color Coded Outcome Clusters

We've added each Launch to the map as a cluster marker on its **Launch Site**, color coded by the **Class** (***Success*** in green or ***Failure*** in red).

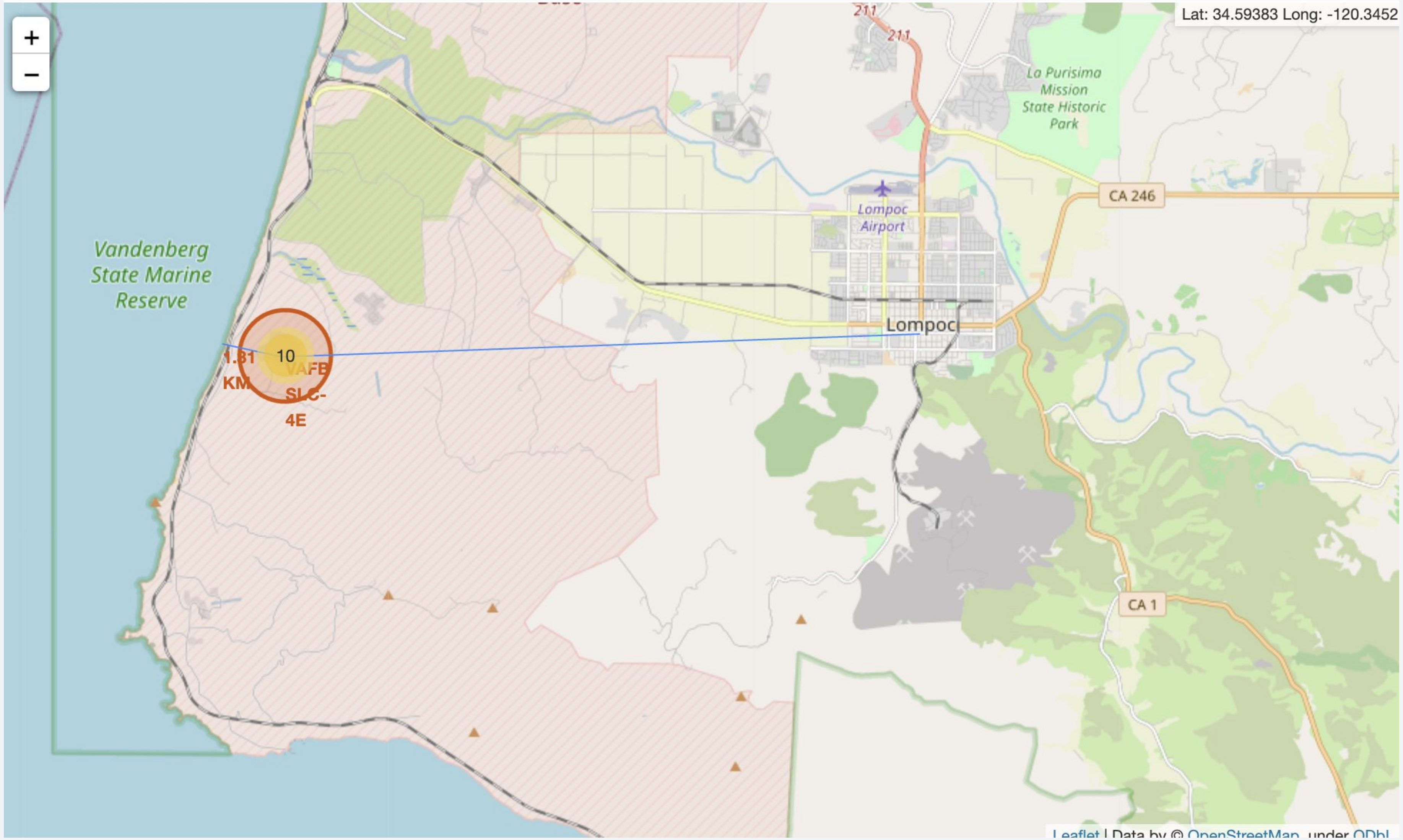
Color Coded Outcome Clusters

If we zoom in, we can visualize the launch outcomes at each site.





Calculating and Marking Distance

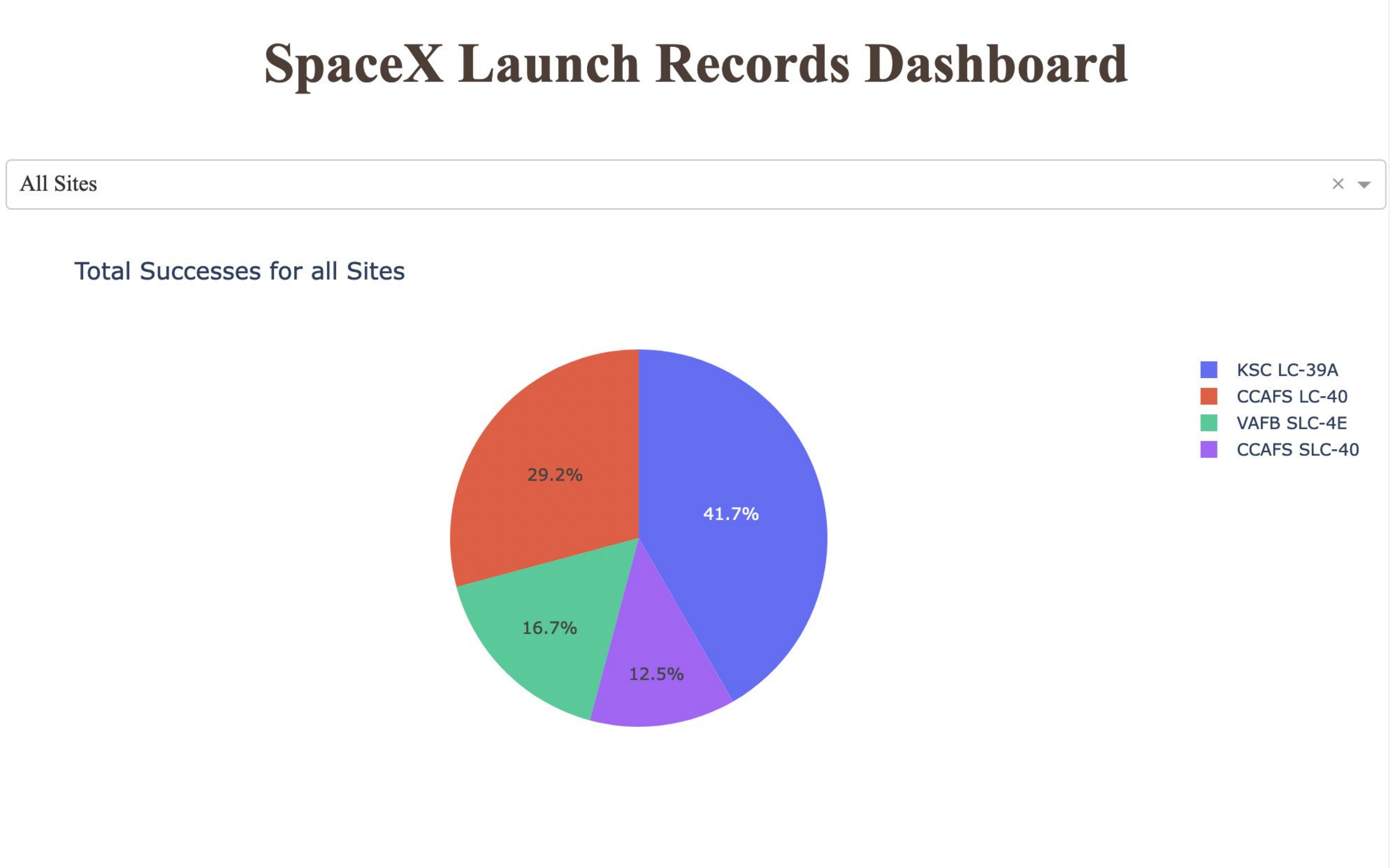


From the **Launch Site *VAFB SLC-4E***, we calculated the distance to the coastline (1.31km) and drew a line to the closest city, ***Lompoci***.

Section 4

Build a Dashboard with Plotly Dash

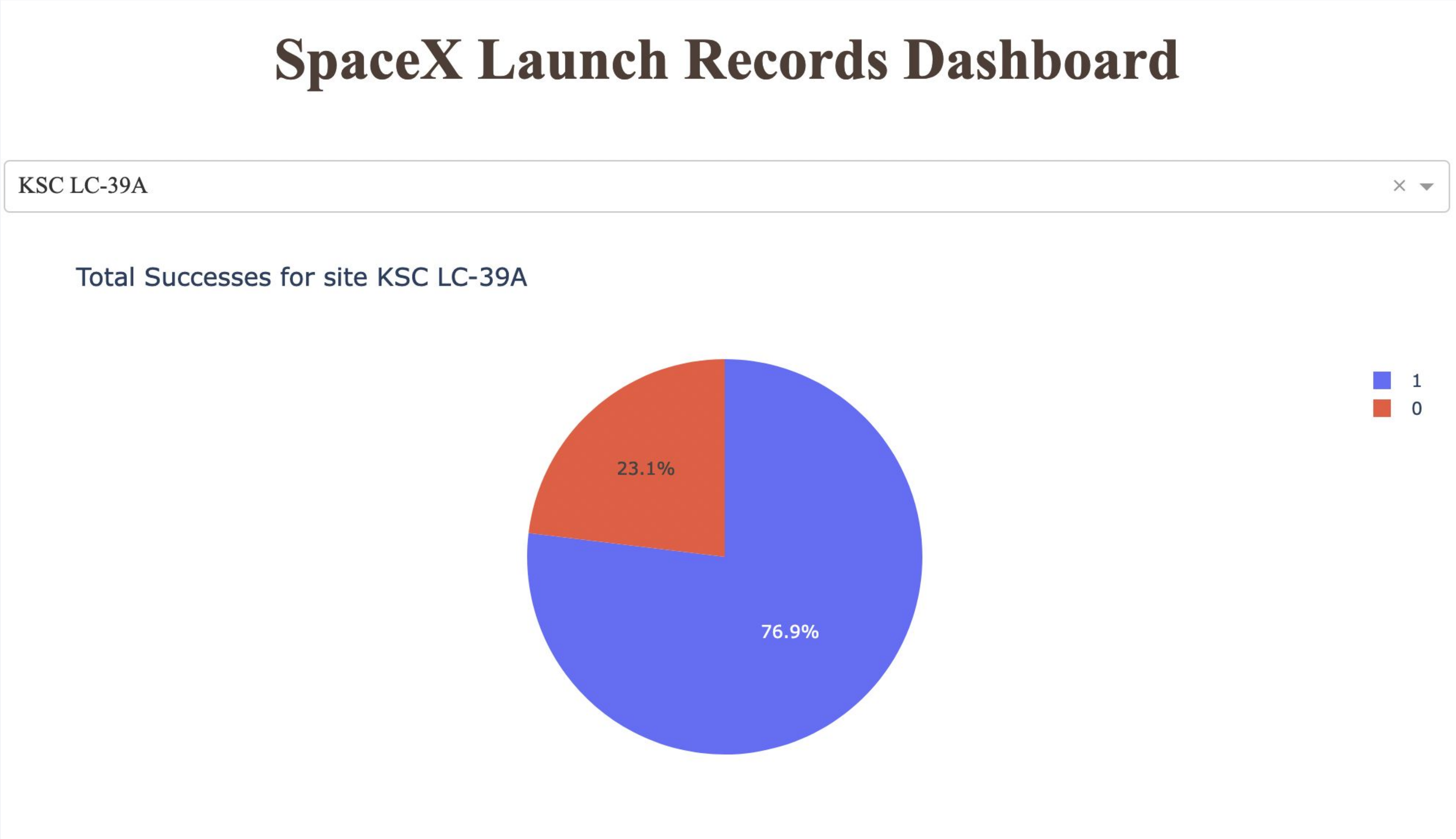
Plotly Dashboard: Launch Site Success



The most successful **Launch Site** is ***KSC LC-39A***, although all **Launch Sites** had successes.

KSC LC-39A: A Successful Launch Site!

- ***KSC LC-39A*** has the most successful launches of the 4 sites, and the highest launch success ratio.



Playing with the elements

The Plotly dashboard helps us interact with the data, seeing what changes as we control for different features.

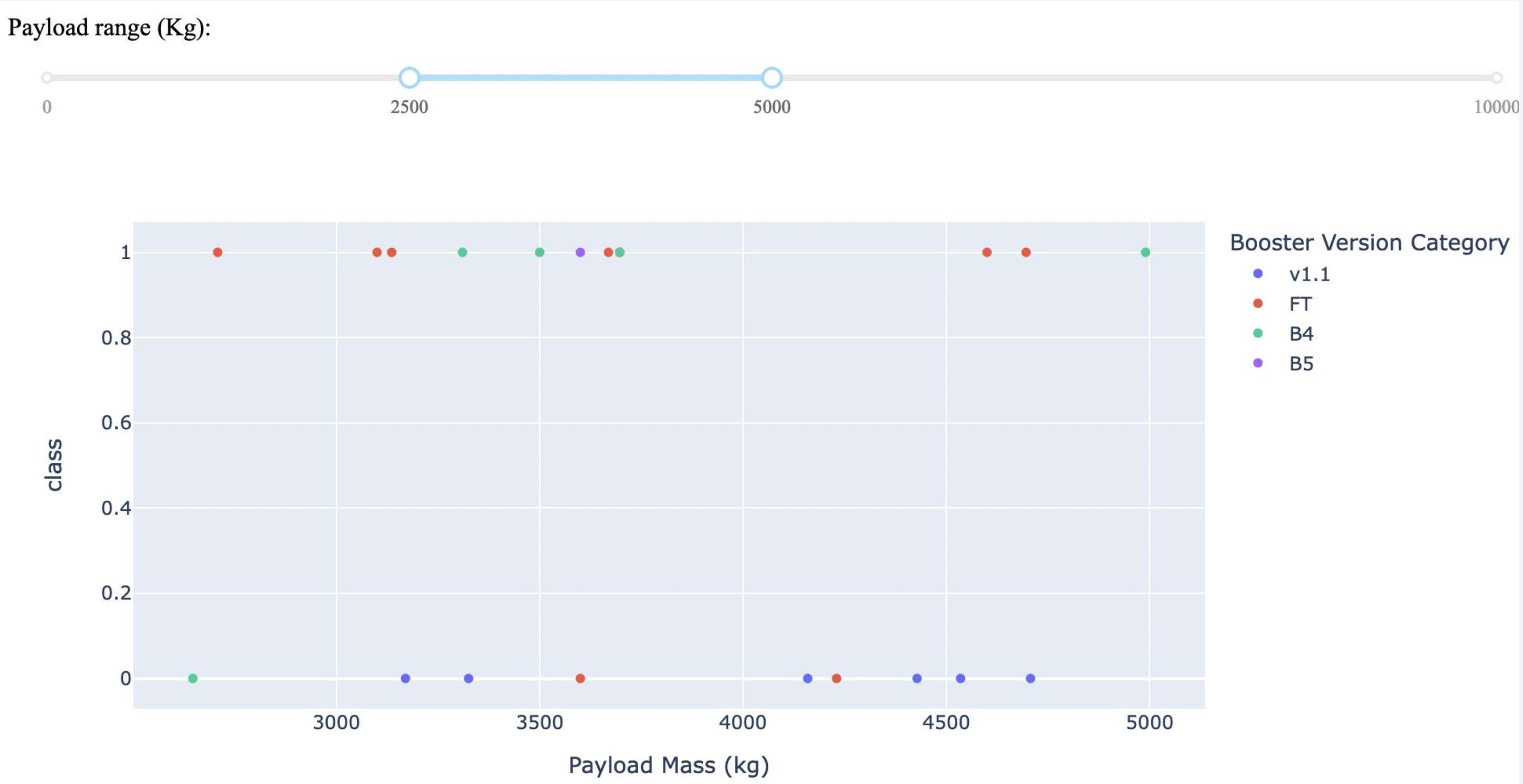
We notice that **Booster Version *v1.1*** is not so successful, while ***FT*** is relatively more successful.

Most successes are concentrated at lower **Payload Mass (kg)** ranges.

Playing with the elements

Limiting the **Payload range** eliminates some **Booster Versions** like ***v1.0***.

We can also see more detail on the **Payload Mass (kg)** of each launch.

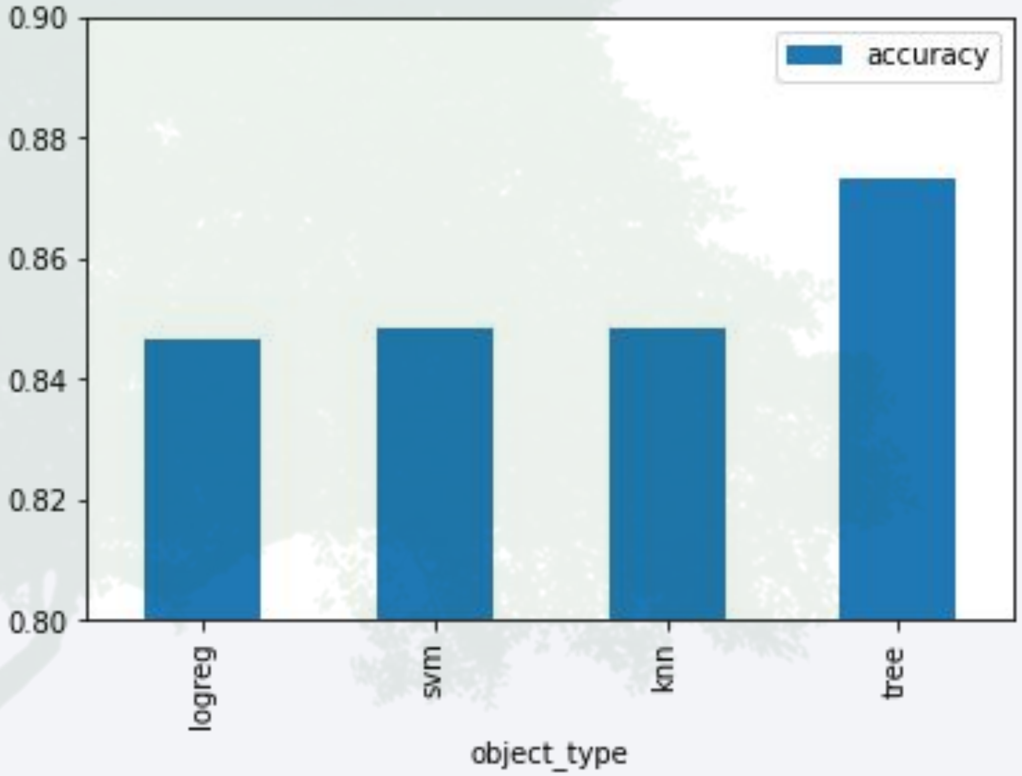


Section 5

Predictive Analysis (Classification)

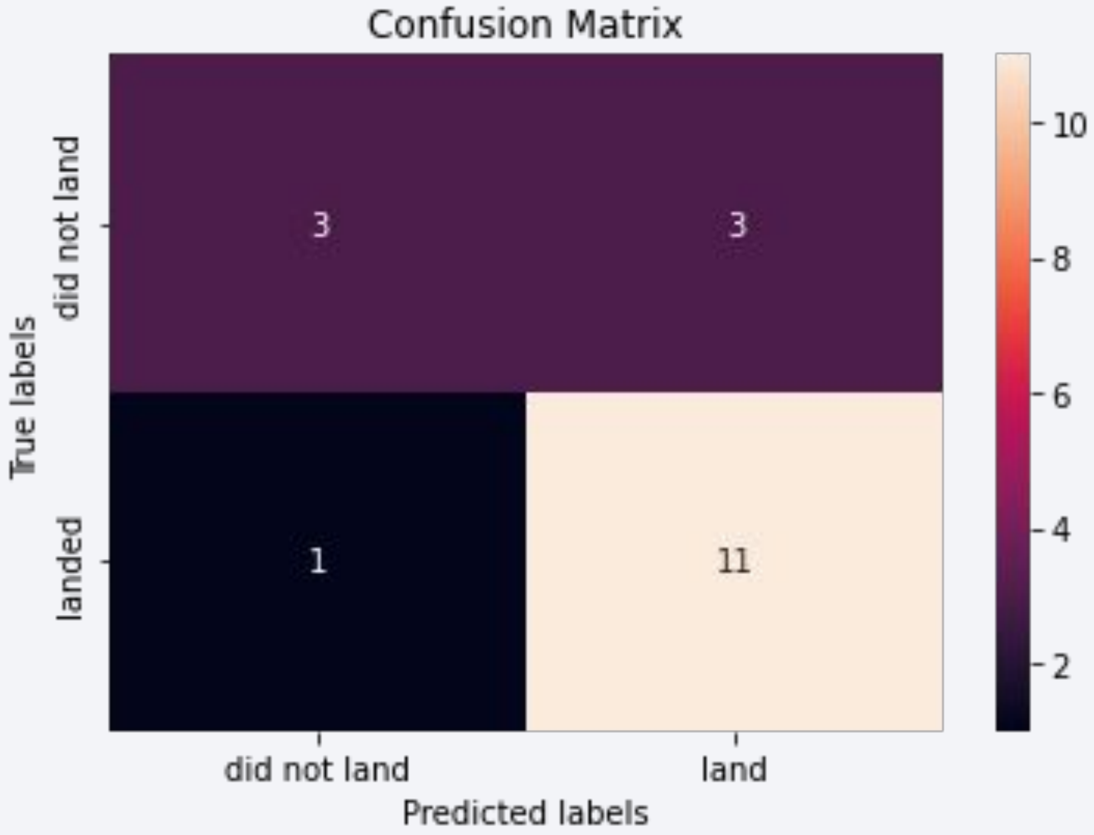
Classification Accuracy

- The **Decision Tree** object has the highest classification accuracy at about 87%.



Confusion Matrix

- All models had a similar confusion matrix, with a tendency towards false positives. The **Decision tree** object had the fewest.



Conclusions

- Launches become more successful (and thus less expensive) over time.
- The most successful **Orbit** types are ***ES-L1***, ***GEO***, ***HEO***, ***SSO***.
- The most successful **Launch Site** is ***KSC LC-39A.***
- The most successful **Booster Version** is ***FT***.
- The **Tree Decision** model has the best accuracy at classifying success.
- **Future Considerations**:
    - Do the most successful **Launch Sites** have a correlation which specific **Orbits** or **Booster Versions**?
    - Are some **Launch Sites** used more often or more recently than others?

Appendix

- Github repo: Data Science and Machine Learning Capstone Project May 2022

Thank you!