IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

LUCIANO CONDE PERES
FEB 2024

Outline

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

Executive Summary

- The methodology applied in this project consists of using a data-driven approach, utilizing publicly available data gathered via REST APIs, leveraging machine learning techniques, interactive visual analytics with Folium and Plotly Dash in order to perform predictive analysis using classification models.
- Summary of results
  - Falcon 9 obtained a success landing rate of 66.67%
  - ES-L1, GEO, HEO and SSO orbits present the highest success rate
  - Launch sites should be close to the coast as well as Equator line as much as possible
  - Payload mass is an important factor to determine the success of a launch, the range with highest success rate is between 3000 and 4000 kg
  - The 4 evaluated ML prediction methods perform similarly, with same accuracy score of 0.8333333333333334

Introduction

- In the rapidly evolving landscape of commercial space exploration, the emergence of companies like SpaceX, Blue Origin, Virgin Galactic, and Rocket Lab has revolutionized the industry, making space travel more accessible than ever before. One of the key factors contributing to SpaceX's success is its ability to significantly reduce launch costs through the innovative reuse of the first stage of its Falcon 9 rockets.
- The project objective is to delve into the dynamics of rocket launches, pricing strategies, and the critical question of whether SpaceX will reuse the first stage for each mission, offering Space Y a data-driven understanding of SpaceX's practices and, consequently, insights into optimizing their own launch operations for cost efficiency and competitiveness.

Section 1

Methodology

Methodology

Executive Summary

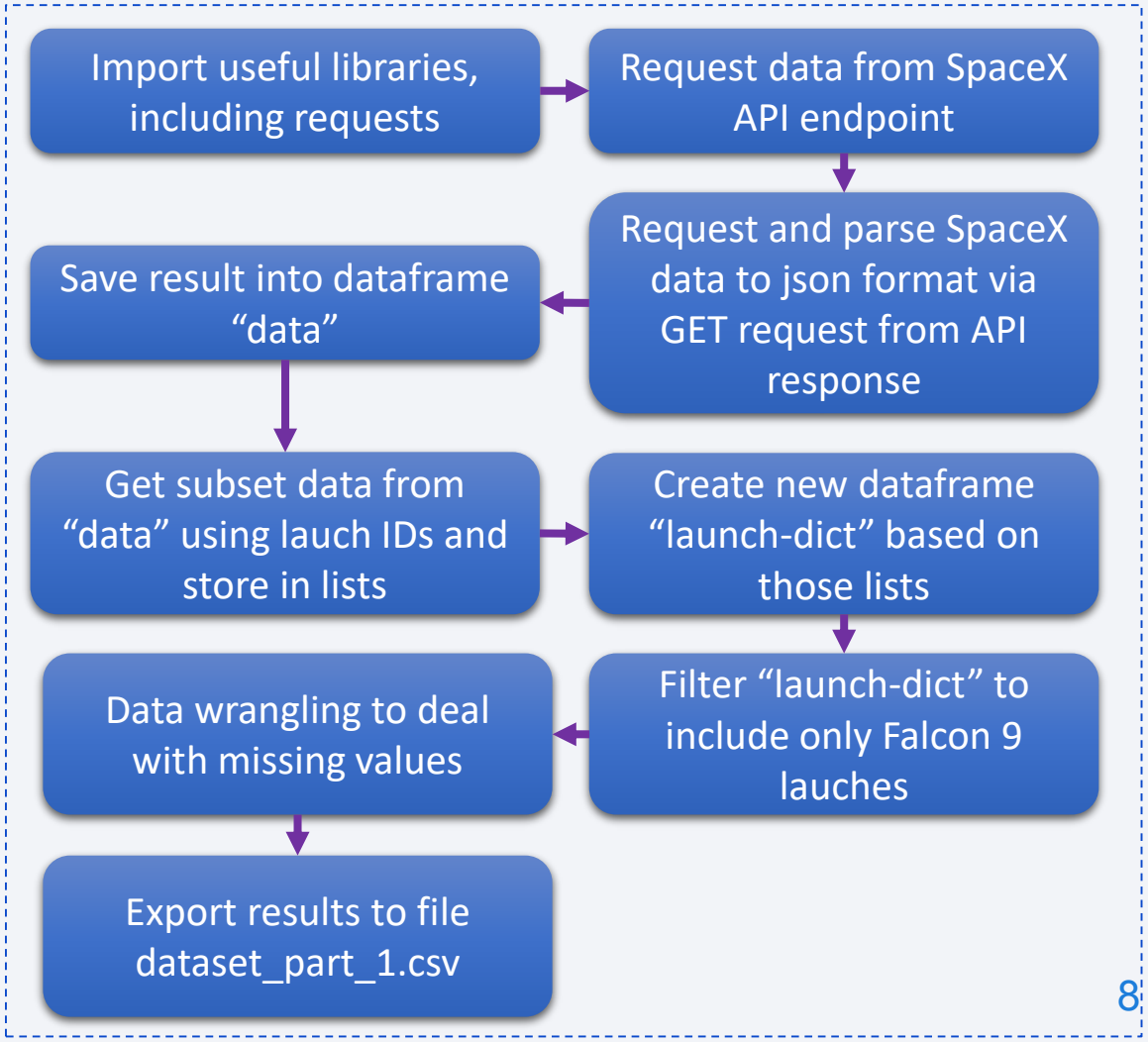
- Data collection methodology
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- Datasets were collected using:
  - **SpaceX REST API** via **Python's requests library**, using the endpoint **api.spacexdata.com/v4/launches/past** – this API gives the information about the rocket used, payload delivered, launch specifications, landing specifications and landing outcome
  - **Web scrapping** using **Python BeautifulSoup** package to obtain some HTML tables from following Wikipedia website: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922. This site contains valuable Falcon 9 launch records.

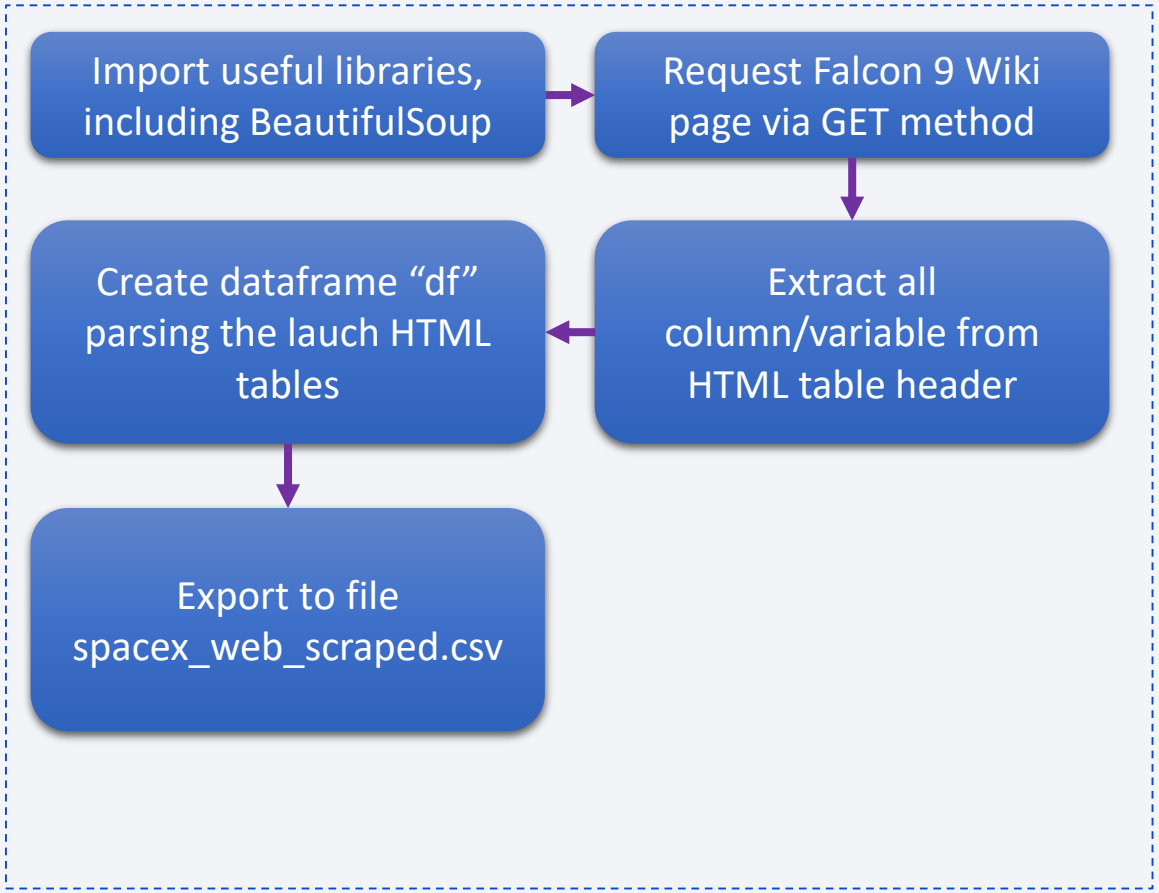
Data Collection – SpaceX API

- SpaceX API calls notebook link
- Dataset generated including only Falcon 9 launches data link



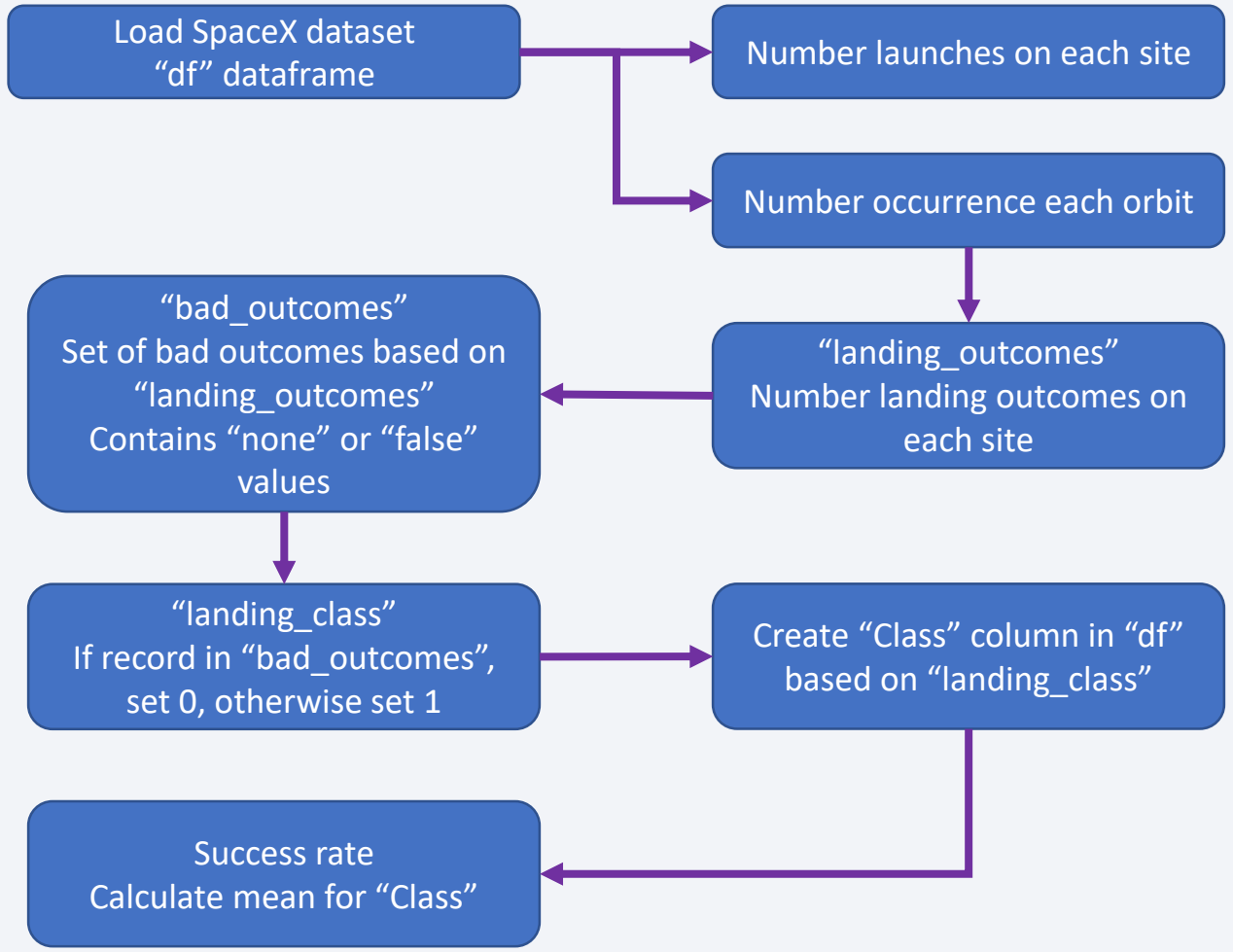
Data Collection - Scraping

- Web scraping notebook reference link



Data Wrangling

- Exploratory Data Analysis of SpaceX dataset related to Falcon 9
- Determine Training Labels for training supervised models
- Convert landing outcomes into Training Labels:
  - 1 -> successful
  - 0 -> unsuccessful
- Landing outcomes success rate calculated:
  - 66.667%
- Data wrangling notebook reference link



EDA with Data Visualization

- Exploratory Data Analysis and Feature Engineering using Pandas, Matplotlib and Seaborn
- Main charts utilized in this step
  - Scatter plot – useful to view relationship between numerical and categorical variables, it was observed in our dataset:
    - Payload Mass vs Success Rate, FlightNumber vs LaunchSite, Launch sites vs. Payload Mass, FlightNumber vs. Orbit type, Payload Mass vs. Orbit type
  - Bar chart – good for representing an aggregate or statistical estimate for a numeric variable with the height/size of each rectangle
    - It was used to visualize the success rate of each orbit type
  - Line chart provides great visualizations for trend analysis
    - It was used to show the average launch success trend along the years
- EDA with data visualization notebook reference link

EDA with SQL (1/2)

- A connection was stablished to the database “my_data1.db”, than a table "SPACEXTBL" was created in order to have SpaceX dataset added from “Spacex.csv” file.
- Using “sqlite3” to perform magic queries, several information was observed as follows:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA‘
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved.

EDA with SQL (2/2)

- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the “booster_versions” which have carried the maximum payload mass
- 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
- 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

- EDA with SQL notebook link

Build an Interactive Map with Folium

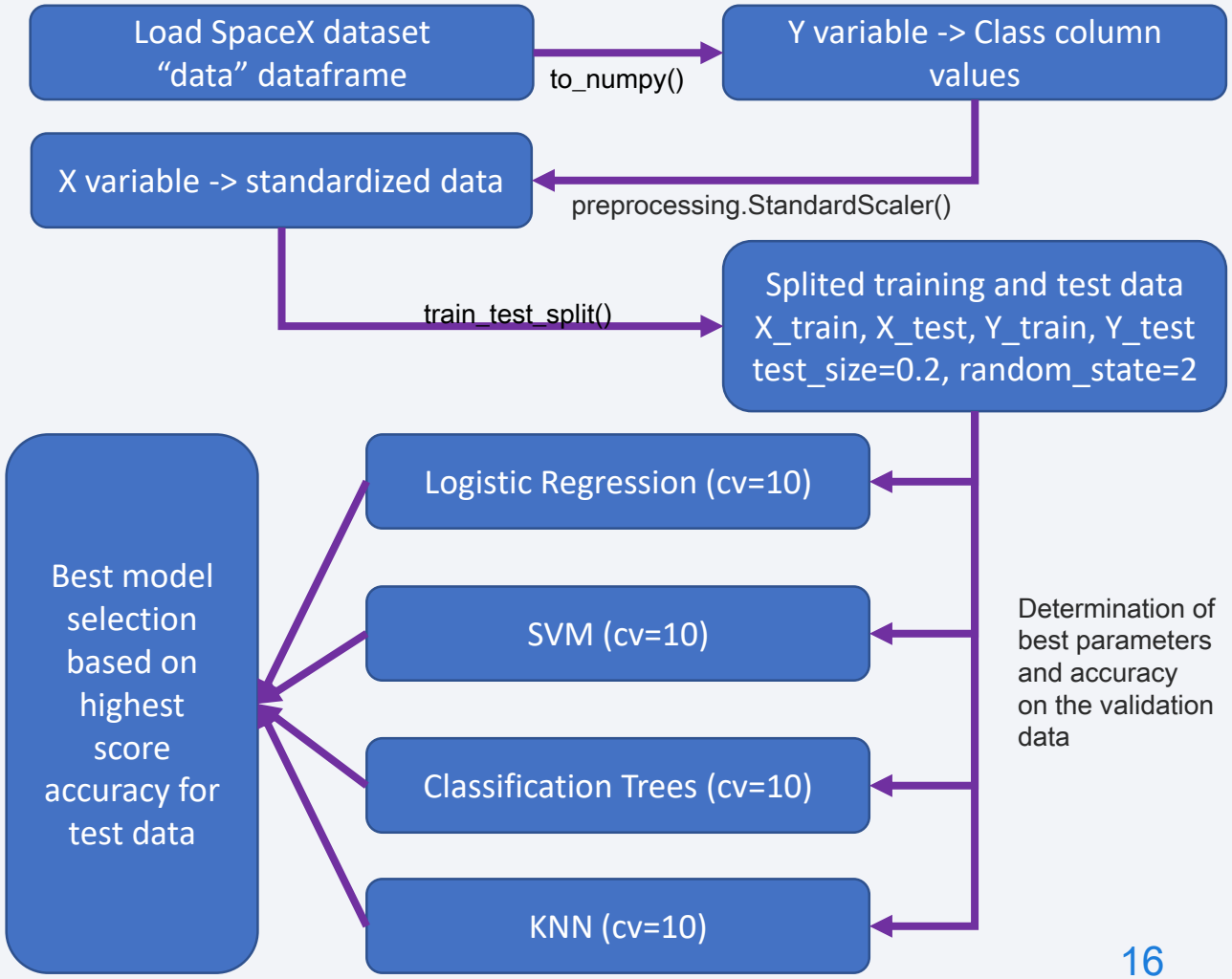
- Objective to explore geographical patterns about launch sites
- Map objects were added to folium map as follows:
  - **Marker**: identifies the launch sites in the map
  - **Circle**: to add a circle area with a text label in order to highlight a launch site based on its specific coordinate
  - **Marker Cluster**: since many launch records have the exact same coordinate, Marker clusters simplify the map containing many markers having the same coordinate.
  - **Polyline**: to illustrate in the map the calculated distances between a launch site to its proximities
- Interactive map with Folium reference notebook link

Build a Dashboard with Plotly Dash

- Objective to obtain insights to answer some questions, such as:
  - Which site has the largest successful launches
  - Which site has the highest launch success rate
  - Etc.
- For this interactive dashboard, it was included following graphs:
  - **Pie chart** - displays which launch site has the largest success count (class=0 vs. class=1). Dropdown menu allows selection of different launch sites.
  - **Scatter plot** – displays payload correlation to mission outcome and identify some visual patterns. A slider allows to easily select different payload ranges.
- Plotly Dash notebook link

Predictive Analysis (Classification)

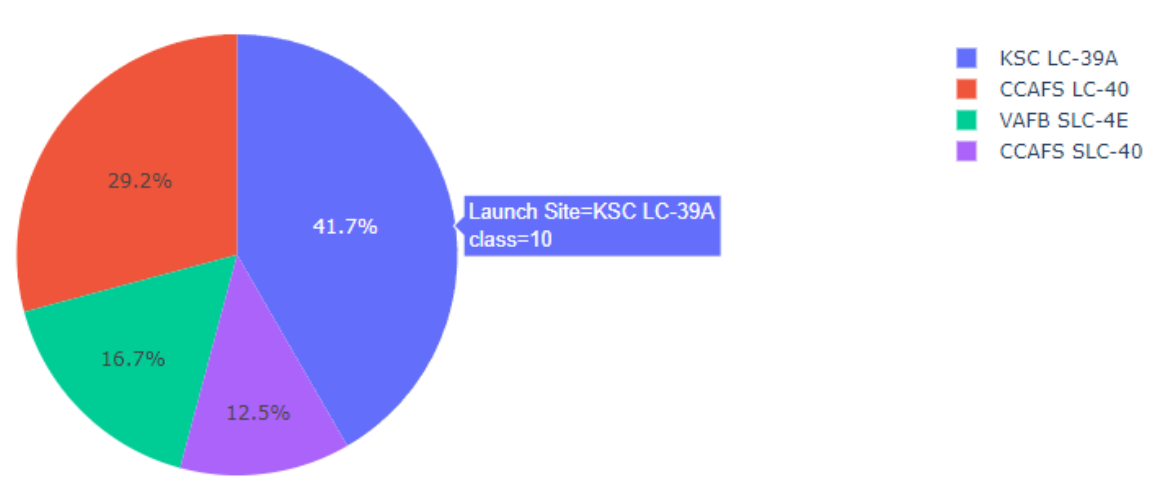
- Machine learning pipeline to predict if the first stage will land given the data from the preceding steps
- ML models evaluated:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Classification Trees
  - K-Nearest Neighbors
- Discover the method which performs best using test data
- Predictive analysis notebook link



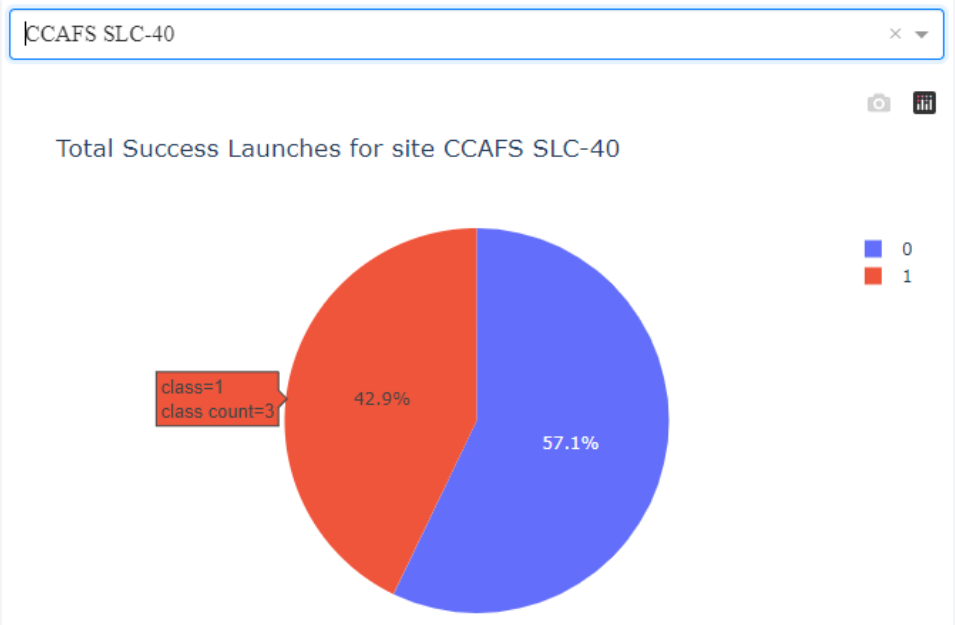
Results – Exploratory Data Analysis

- It was observed from the dataset containing 90 records, a success landing rate of **66.67%** for Falcon 9
- Drop ship ASDS is the booster with highest number of successful landing outcomes: **41**
- The average payload mass carried by booster version F9 v1 is **2,928.4 kg**
- The first successful landing outcome in ground pad was achieved in **Dec 22, 2015**
- 4 launch sites where utilized in the space mission
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

Results – Interactive Dashboard (1/3)

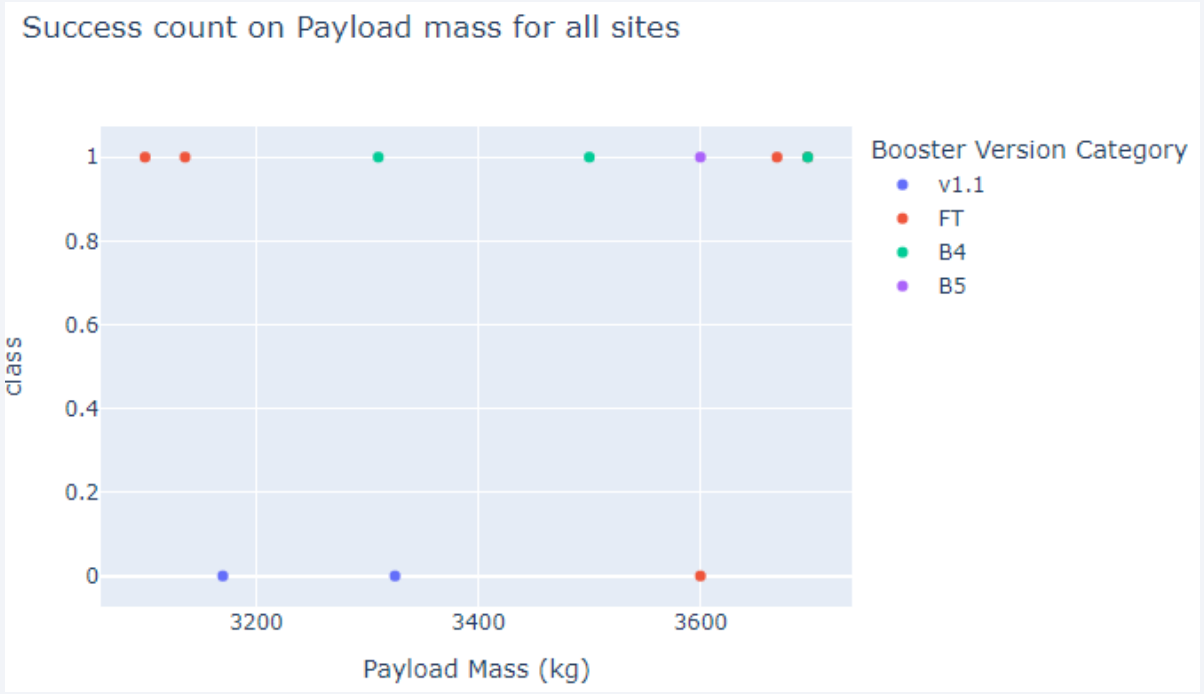


Site with largest number of successful launches:
KSC LC-39A with **10 occurrences**

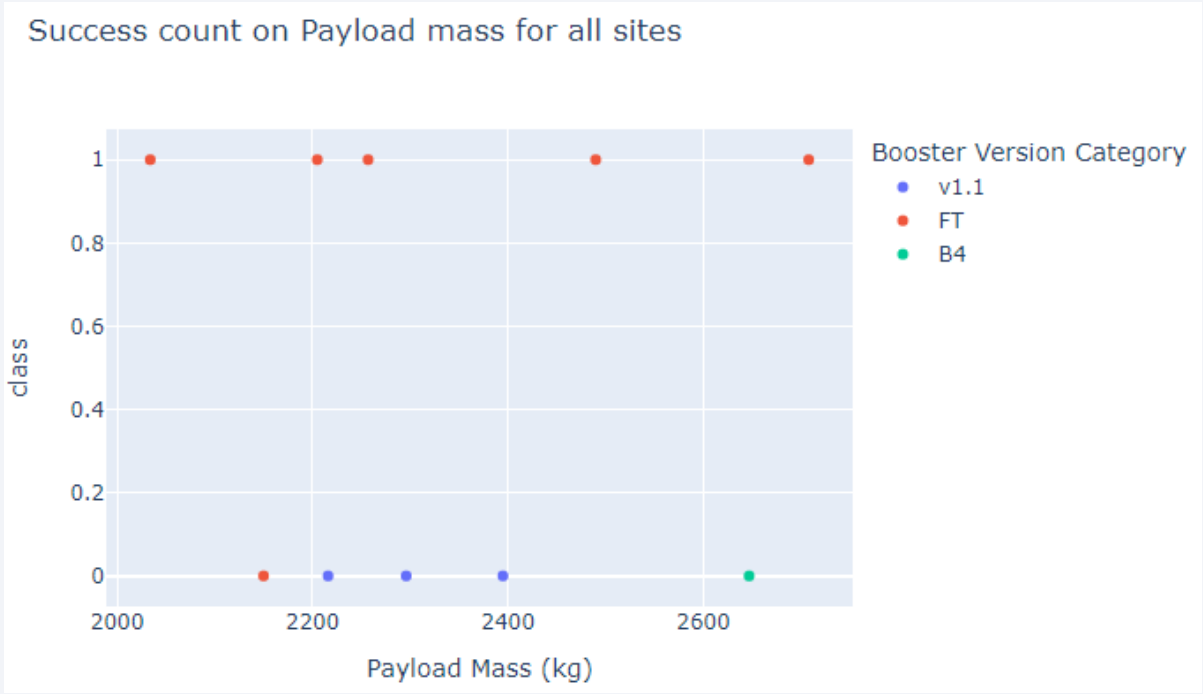


Site with highest launch success rate:
CCFAFS SLC-40 with **42,9%**

Results – Interactive Dashboard (2/3)



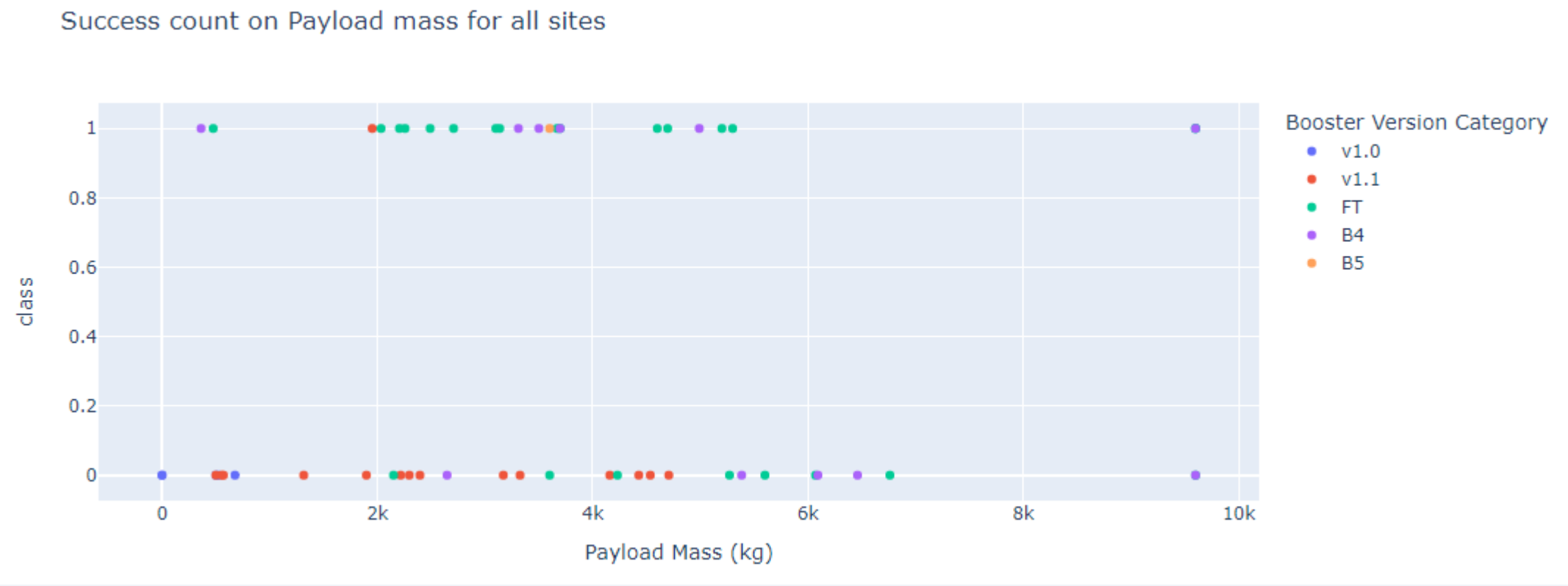
Payload range with highest launch success rate: 3000 – 4000 kg



Payload range with lowest launch success rate: 2000 – 3000 kg / 5000 – 6000kg

Results – Interactive Dashboard (3/3)

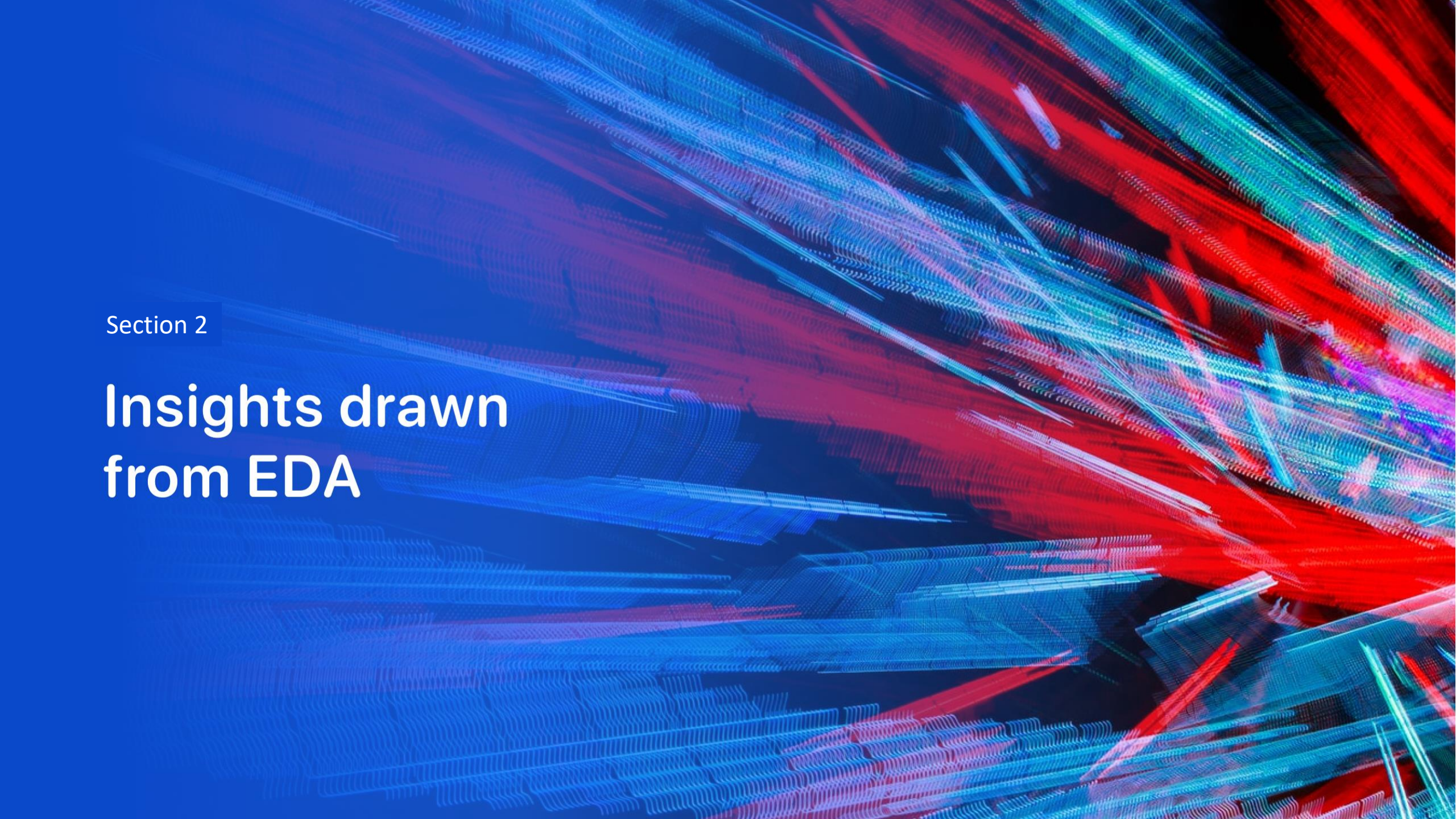
- F9 Booster version with the highest launch success rate: **FT**



Results - Predictive analysis

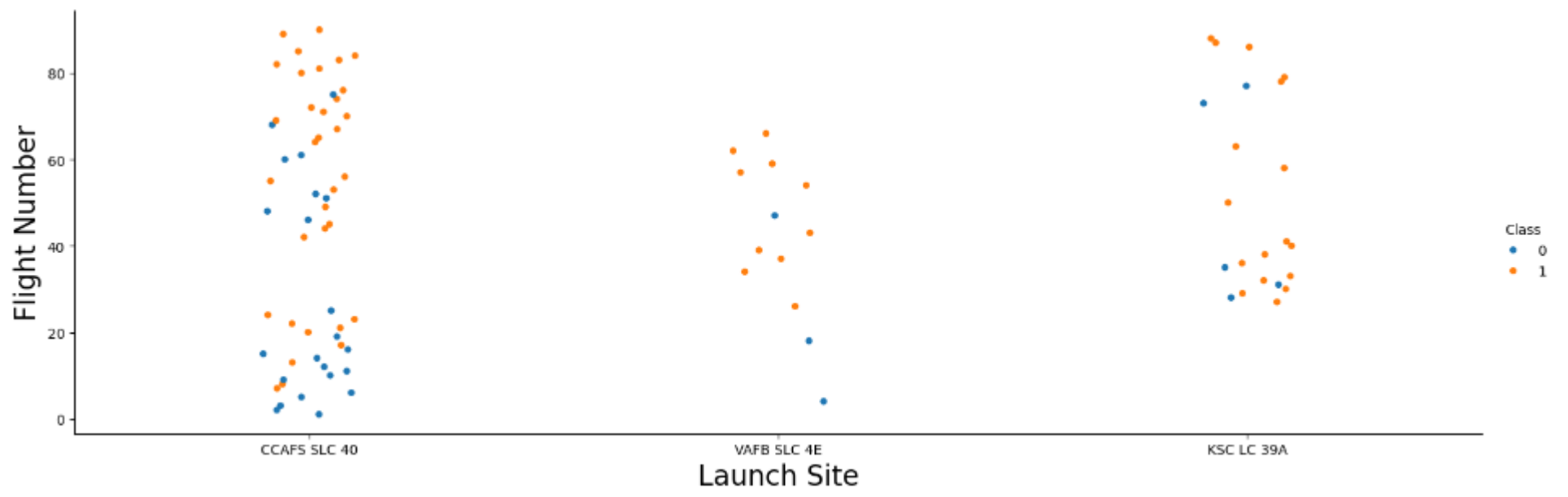
- Sample assessed contained 90 records, 83 columns (parameters)
- All evaluated methods gave exact same accuracy score performance as follows:

Method	Best Hyperparameters	Accuracy (score method)
Logistic Regression	'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'	0.8333333333333334
SVM	'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'	0.8333333333333334
Classification Tree	'criterion': 'entropy', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 5, 'splitter': 'random'	0.8333333333333334
KNN	'algorithm': 'auto', 'n_neighbors': 10, 'p': 1	0.8333333333333334


Section 2
Insights drawn from EDA

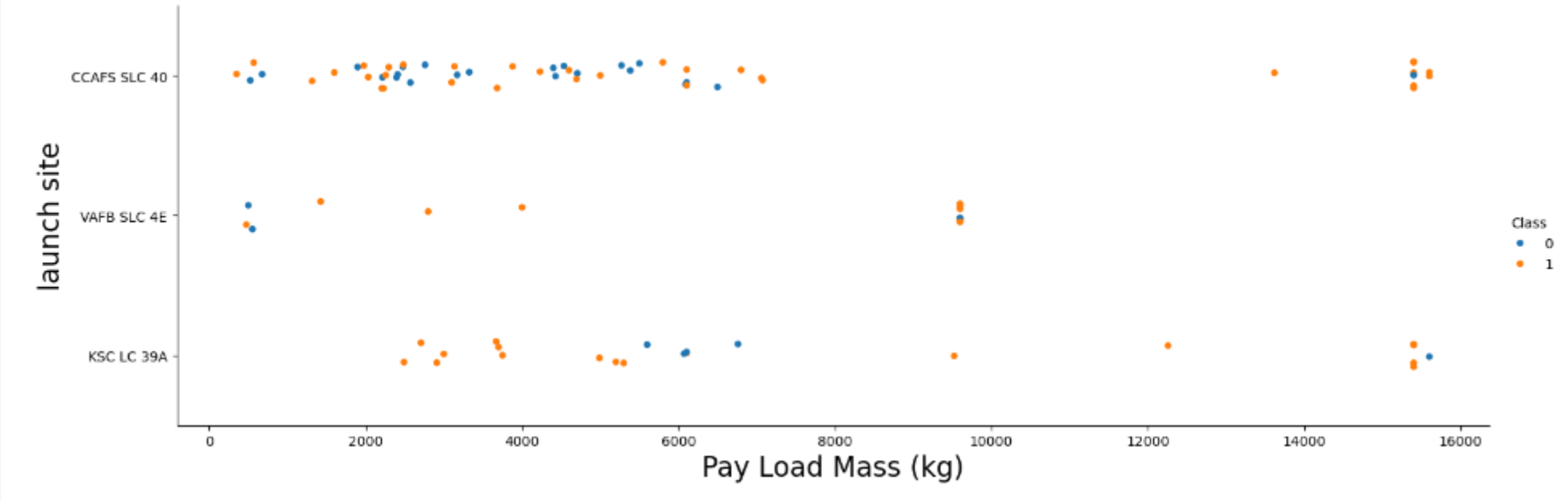
Flight Number vs. Launch Site

- CCAFS SLC-40 has the largest number of launches
- No correlation observed with Flight Number



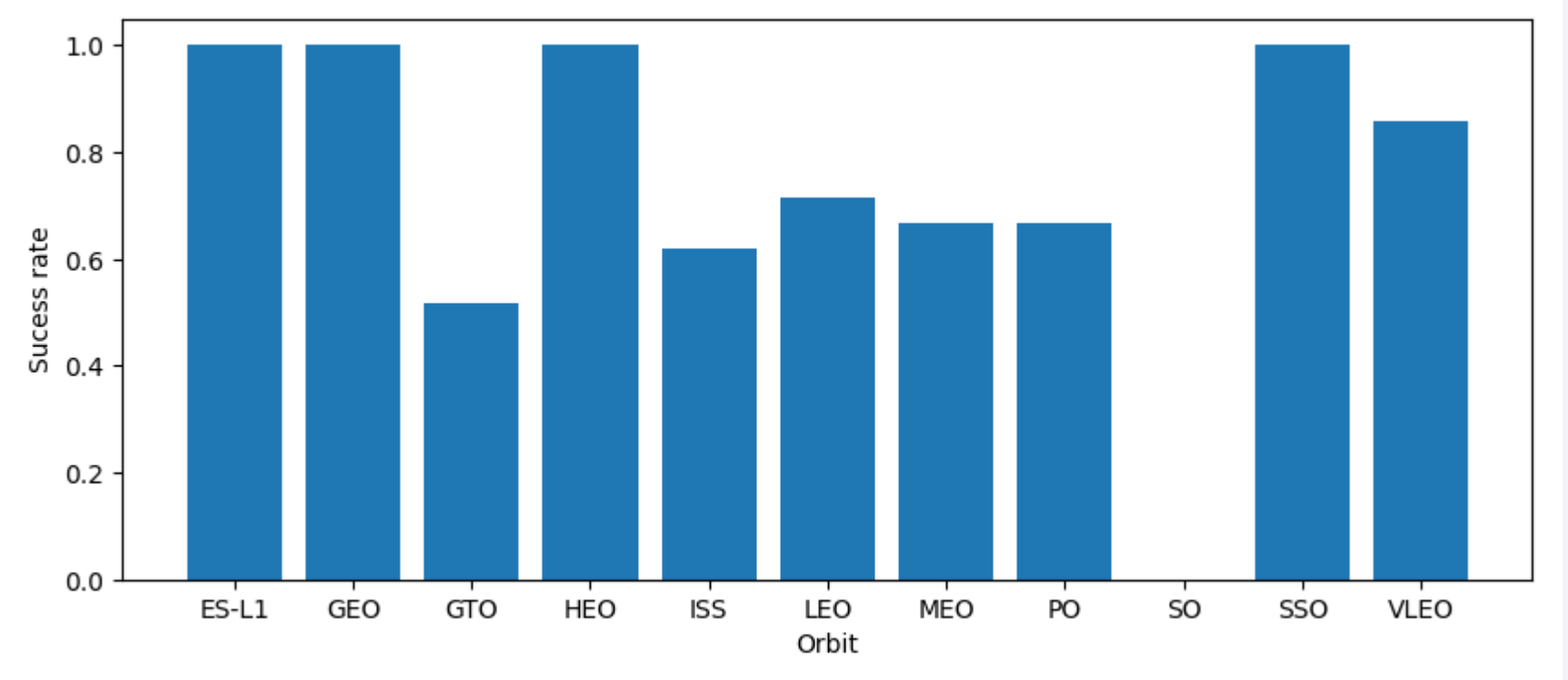
Payload vs. Launch Site

- This chart reveals that lower payload mass has better success rate
- There are no launches from VAFB-SLC heavy payload mass(>10,000).



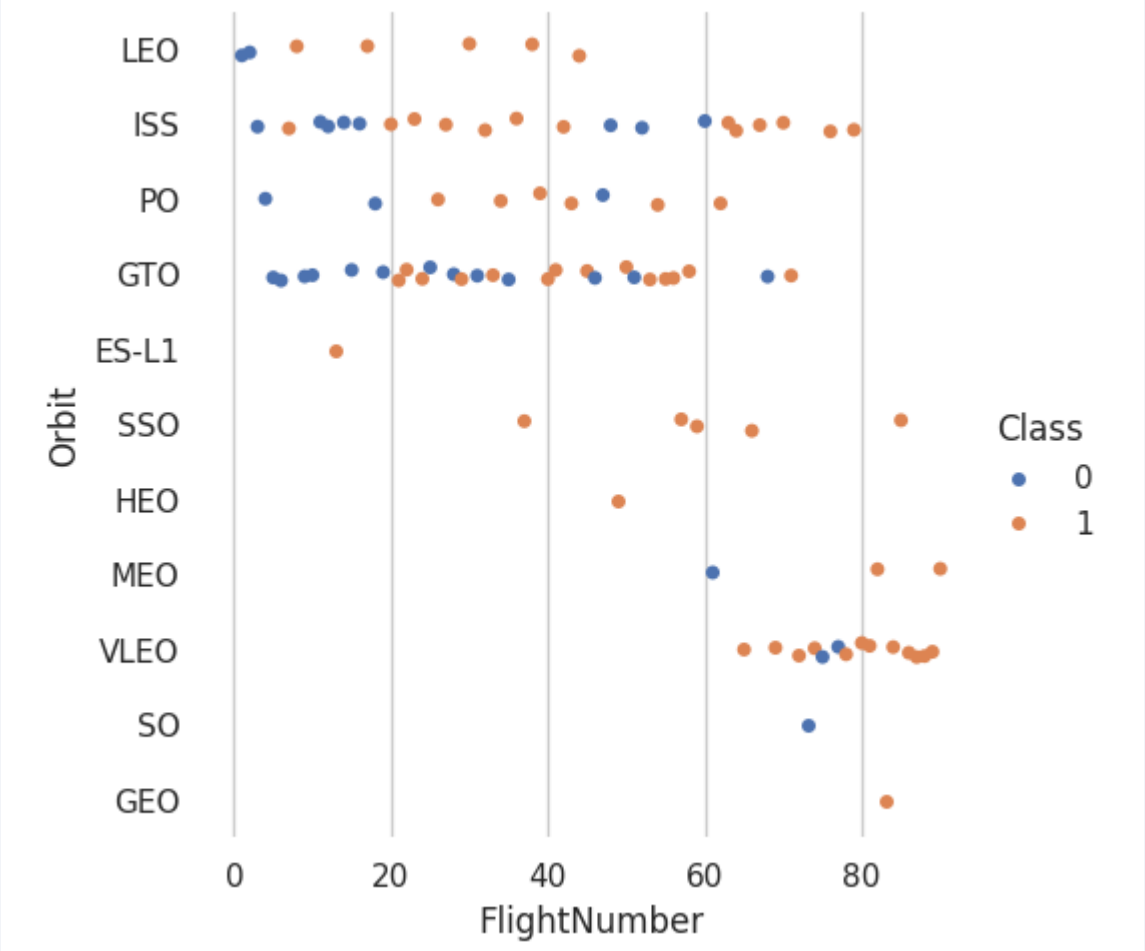
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO present the highest success rate



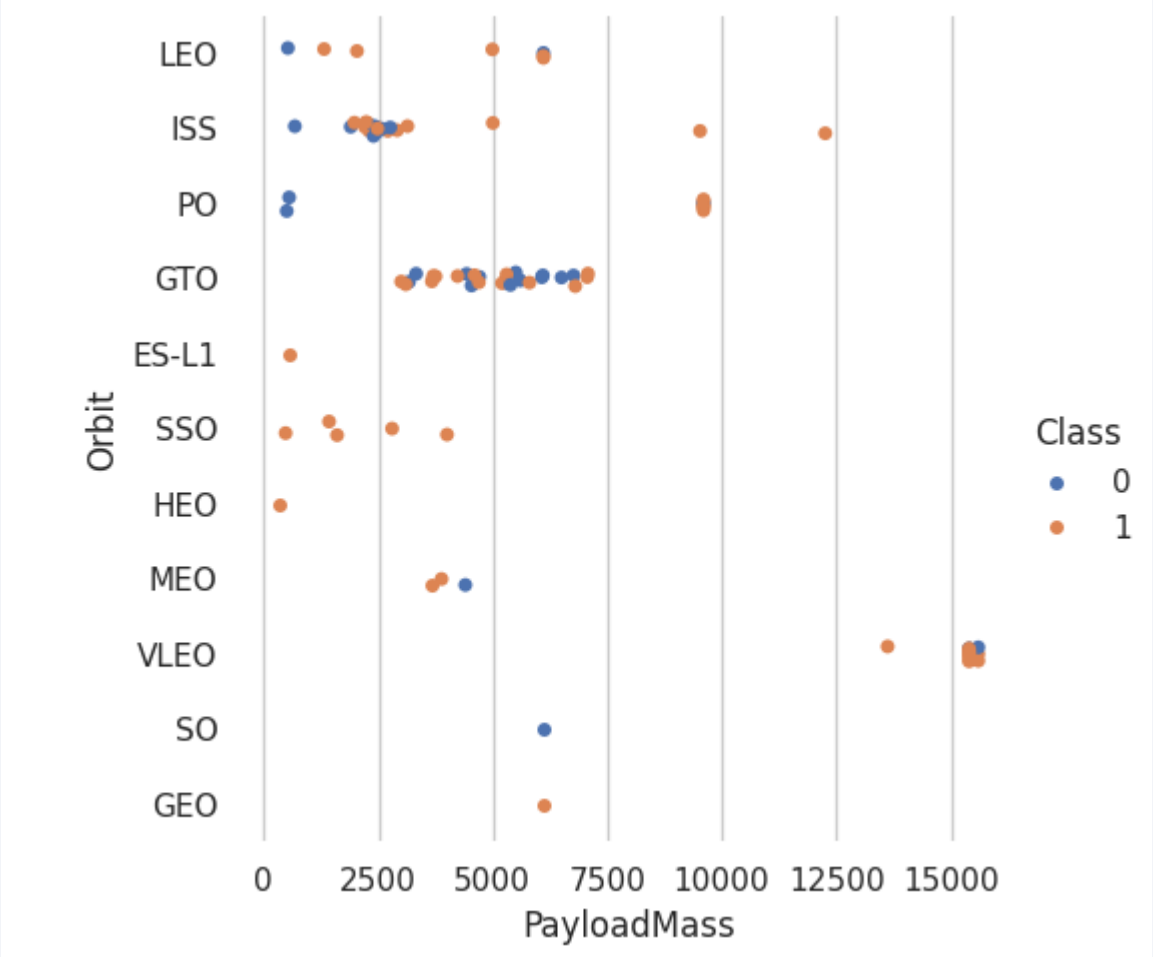
Flight Number vs. Orbit Type

- Success appears related to the number of flights for LEO orbit
- There seems to be no relationship between flight number when in GTO orbit



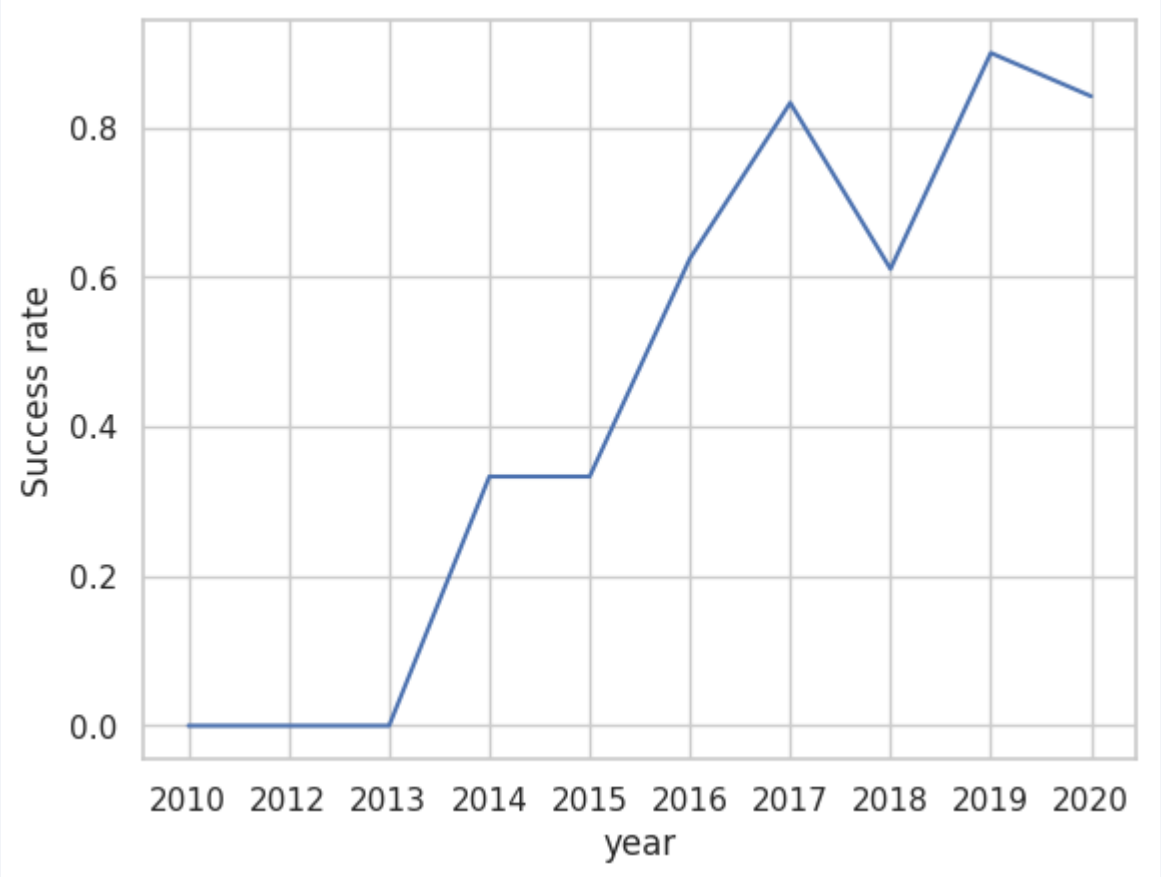
Payload vs. Orbit Type

- Lower payload mass usually has better success rates



Launch Success Yearly Trend

- This chart reveals that success rate since 2013 kept increasing till 2020
- Highest success rate achieved in 2019



All Launch Site Names

- Names of the unique launch sites obtained via SQL query:

  select distinct Launch_Site from SPACEXTBL

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA` using SQL query:

  *select * from SPACEXTBL 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

Total Payload Mass

- Total payload carried by boosters from NASA obtained via SQL query:

```
select Customer, Booster_Version,
sum(PAYLOAD_MASS__KG_) as 'total payload
mass'

from SPACEXTBL

where Customer = 'NASA (CRS)'

group by Customer, Booster_Version
```

Customer	Booster_Version	total payload mass
NASA (CRS)	F9 B4 B1039.2	2647
NASA (CRS)	F9 B4 B1039.1	3310
NASA (CRS)	F9 B4 B1045.2	2697
NASA (CRS)	F9 B5 B1056.2	2268
NASA (CRS)	F9 B5 B1058.4	2972
NASA (CRS)	F9 B5 B1059.2	1977
NASA (CRS)	F9 B5B1050	2500
NASA (CRS)	F9 B5B1056.1	2495
NASA (CRS)	F9 FT B1035.2	2205
NASA (CRS)	F9 FT B1021.1	3136
NASA (CRS)	F9 FT B1025.1	2257
NASA (CRS)	F9 FT B1031.1	2490
NASA (CRS)	F9 FT B1035.1	2708
NASA (CRS)	F9 v1.0 B0006	500
NASA (CRS)	F9 v1.0 B0007	677
NASA (CRS)	F9 v1.1	2296
NASA (CRS)	F9 v1.1 B1010	2216
NASA (CRS)	F9 v1.1 B1012	2395
NASA (CRS)	F9 v1.1 B1015	1898
NASA (CRS)	F9 v1.1 B1018	1952

Average Payload Mass by F9 v1.1

- Using below SQL query, it was obtained the **average payload mass** carried by booster version F9 v1.1 of **2,928,4 kg**

```
select Booster_Version, avg(PAYLOAD_MASS__KG_) as ‘avg payload mass'
from SPACEXTBL
where Booster_Version = 'F9 v1.1'
group by Booster_Version
```

First Successful Ground Landing Date

- **Dec 22nd, 2015** was the date of the first successful landing outcome on ground pad was obtained via SQL query:

```
select *

from SPACEXTBL

where Landing_Outcome = 'Success (ground pad)'

order by Date asc limit 1
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 were obtained via SQL query:

  select Booster_Version, Payload, PAYLOAD_MASS__KG_ from SPACEXTBL where (Landing_Outcome = 'Success (drone ship)') and (PAYLOAD_MASS__KG_ between 4000 and 6000)

Booster_Version	Payload	PAYLOAD_MASS__KG_
F9 FT B1022	JCSAT-14	4696
F9 FT B1026	JCSAT-16	4600
F9 FT B1021.2	SES-10	5300
F9 FT B1031.2	SES-11 / EchoStar 105	5200

Total Number of Successful and Failure Mission Outcomes

- We observed a total number of **100 successful** and only **1 failure mission outcomes**, using below SQL query:

```
SELECT Mission_Outcome, count(Mission_Outcome) as Number_Mission_Outcomes
from SPACEXTBL
group by Mission_Outcome
```

Boosters Carried Maximum Payload

- List of boosters which have carried the maximum payload mass was obtained via SQL, using subqueries as follows:

```
SELECT DISTINCT BOOSTER_VERSION

FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ =
(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)

ORDER BY BOOSTER_VERSION
```

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

2015 Launch Records

- List of failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015. Since SQLLite does not support month names, it was used a formula for month and year as follows:
  - substr(Date, 6,2) as month to get the months
  - substr(Date,0,5)='2015' for year.
- *SQL query:*

```
select substr(Date, 6,2) as Mth, *

from SPACEXTBL

where Landing_Outcome = 'Failure (drone ship)'  and substr(Date,0,5)='2015'
```

Mth	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
01	2015-01-10	9:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
04	2015-04-14	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)

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

- Rank of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order, using below SQL query:

```
SELECT Landing_Outcome, count(*) as Qty

FROM SPACEXTBL

WHERE Date between '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"

ORDER BY "Qty"
```

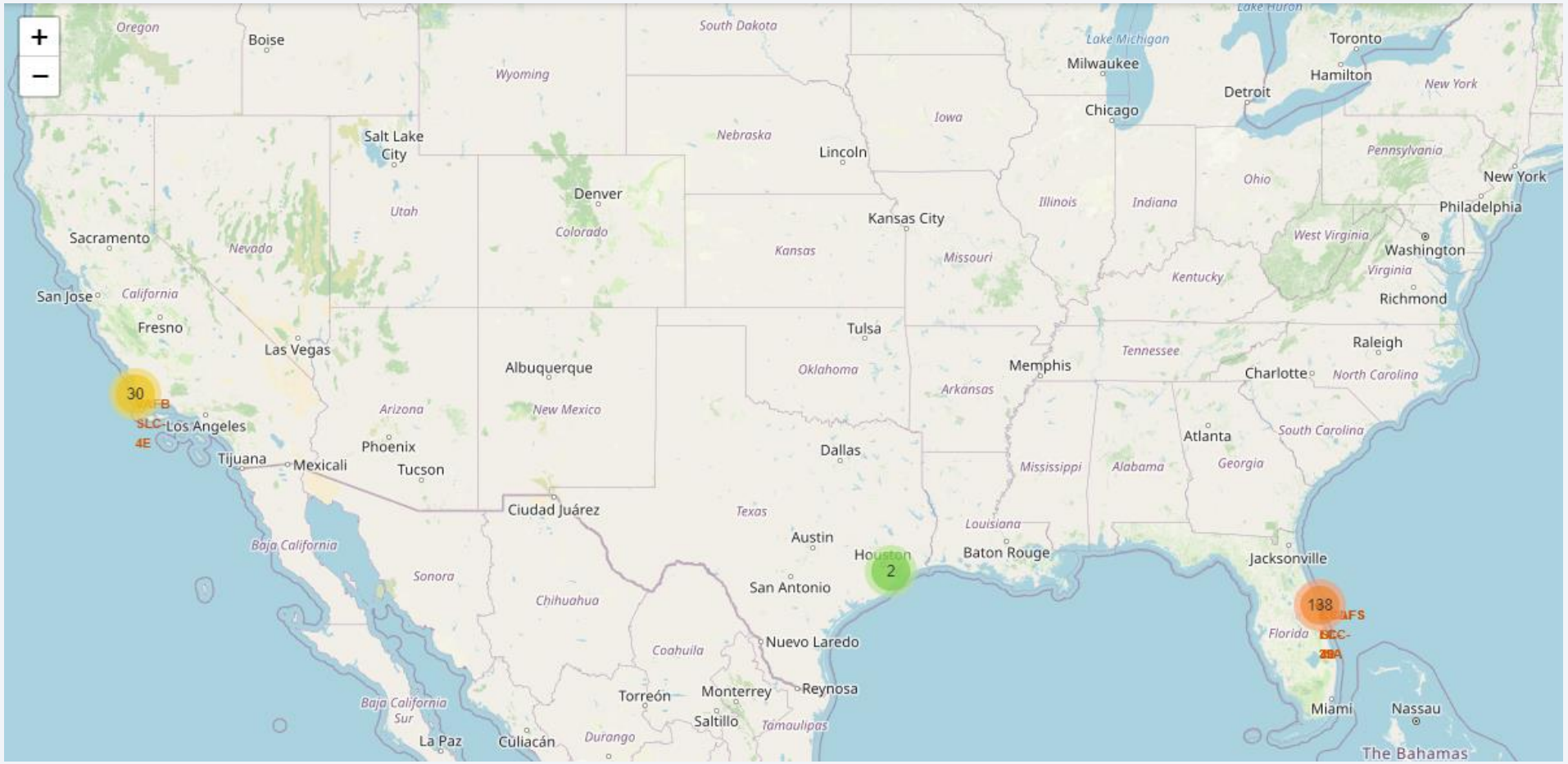
Landing_Outcome	Qty
Precluded (drone ship)	1
Failure (parachute)	2
Uncontrolled (ocean)	2
Controlled (ocean)	3
Success (ground pad)	3
Failure (drone ship)	5
Success (drone ship)	5
No attempt	10

Section 3

Launch Sites Proximities Analysis

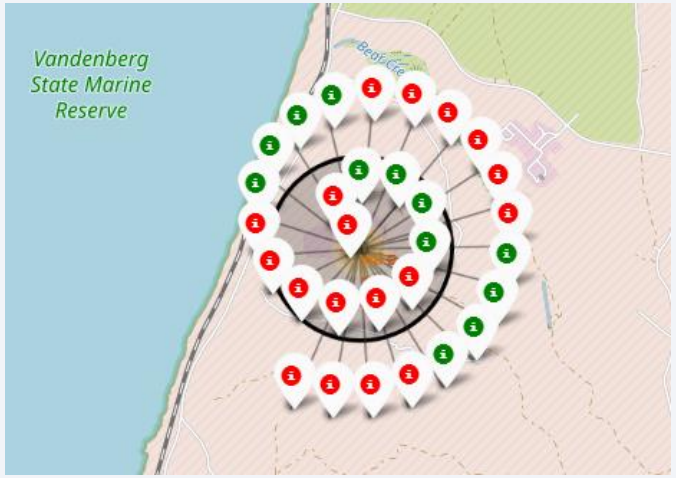
Launch Sites Visualization with Interactive Map

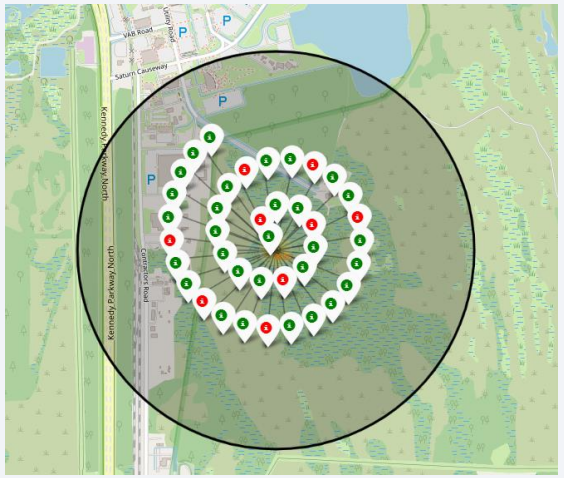
- It observed that launch sites (marked in yellow and orange circles) are close to the coast as well as Equator line

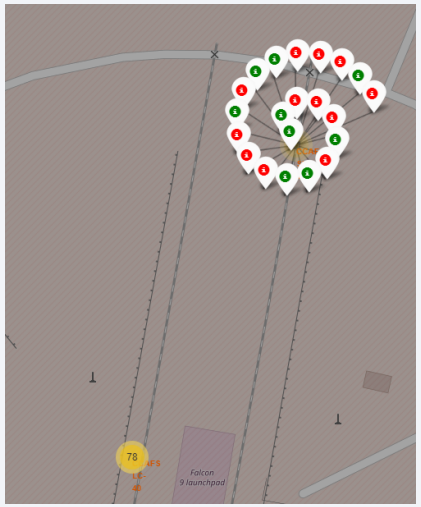


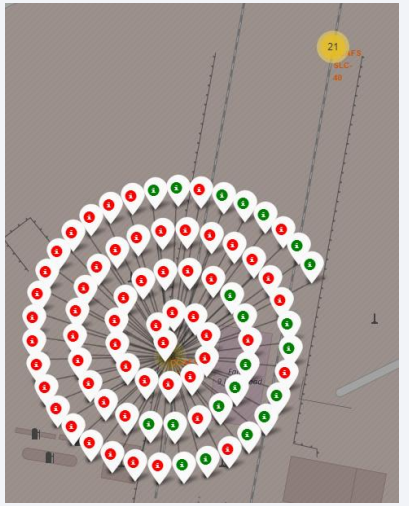
Successful & Failures Launches View with Interactive Map

- Successful launches (class=1) are represented with a green marker and if failed ones, with a red marker (class=0)



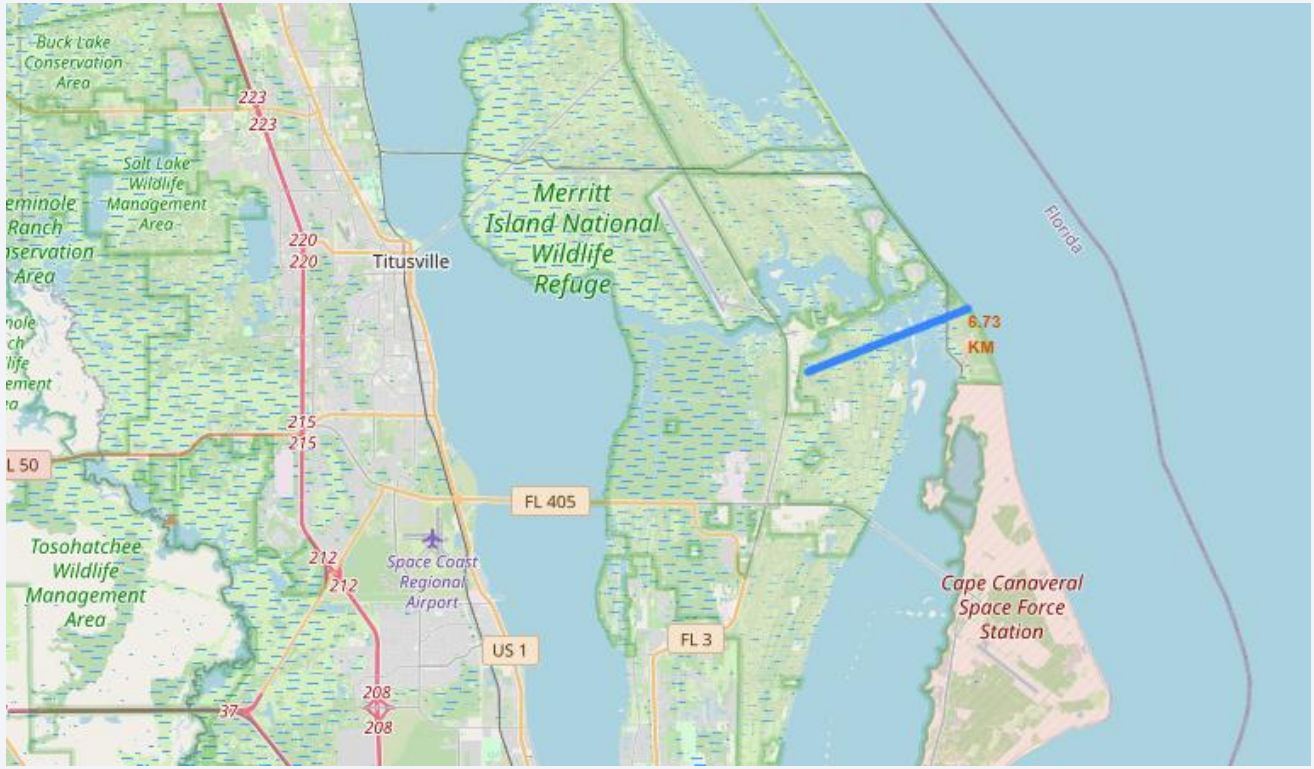


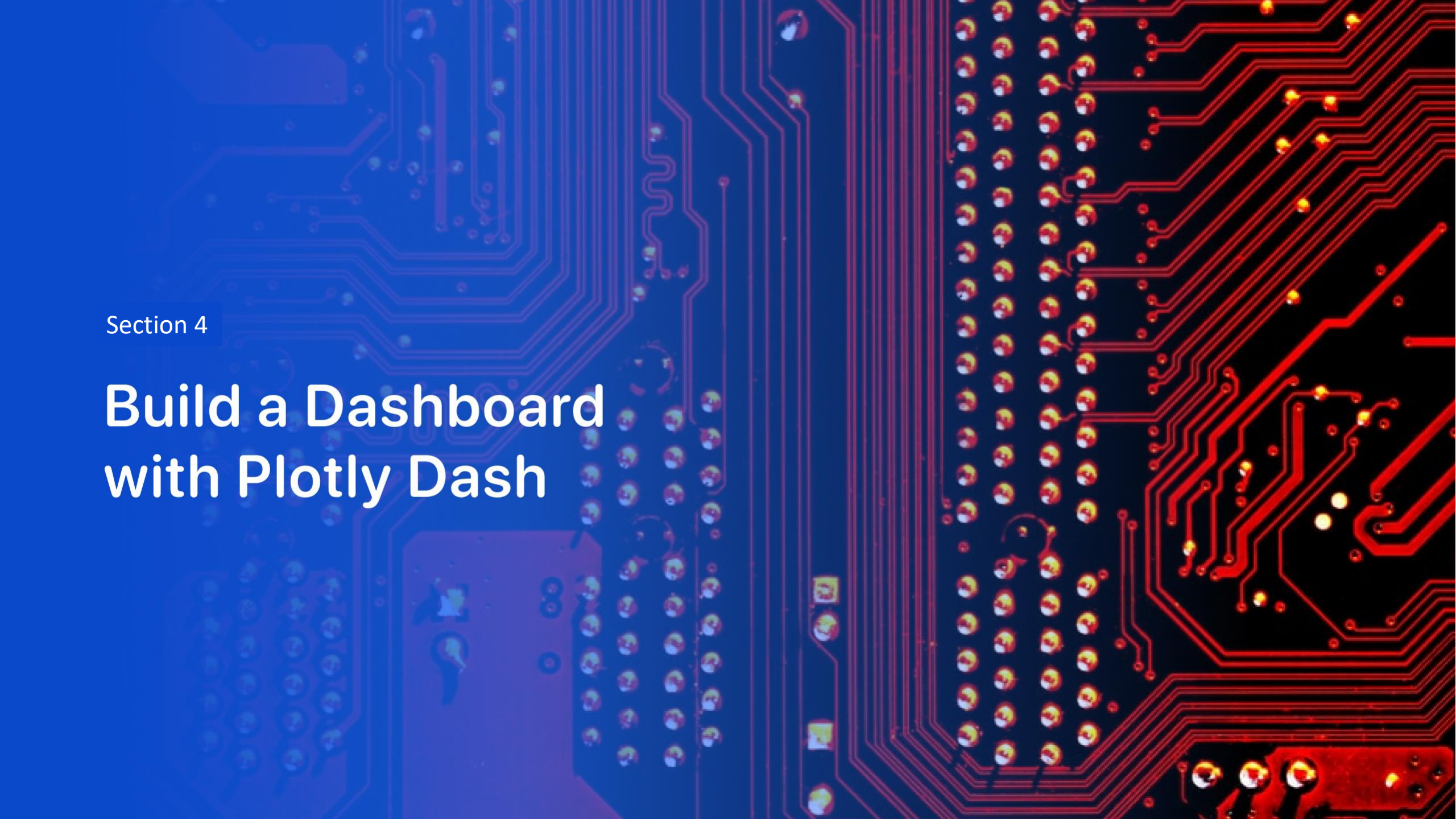




Distance Views with Interactive Map

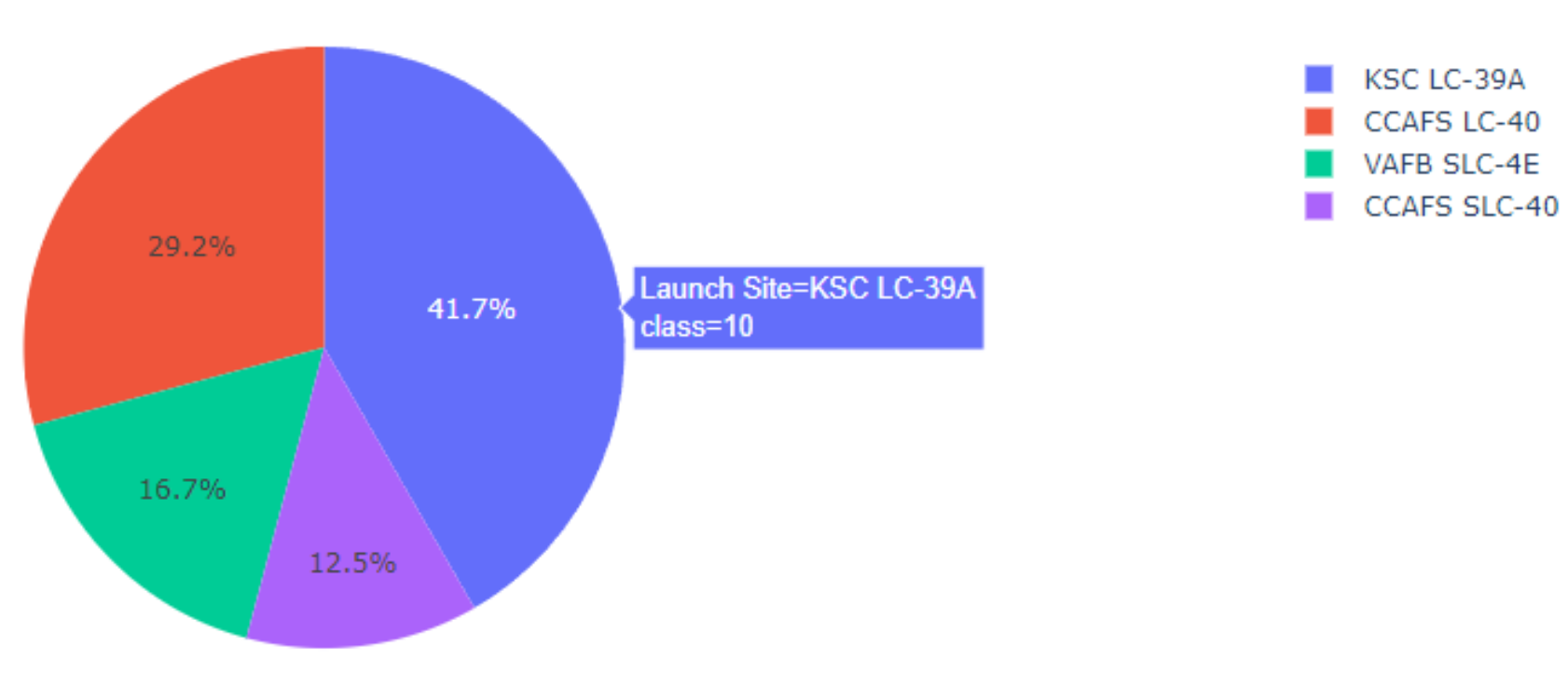
- Distances with points of interests can be calculated and easily visualized in the map
- Below we see an example, the blue line shows the distance between the coast and launch site KSC LC-39A (6.73km)




Section 4
Build a Dashboard with Plotly Dash

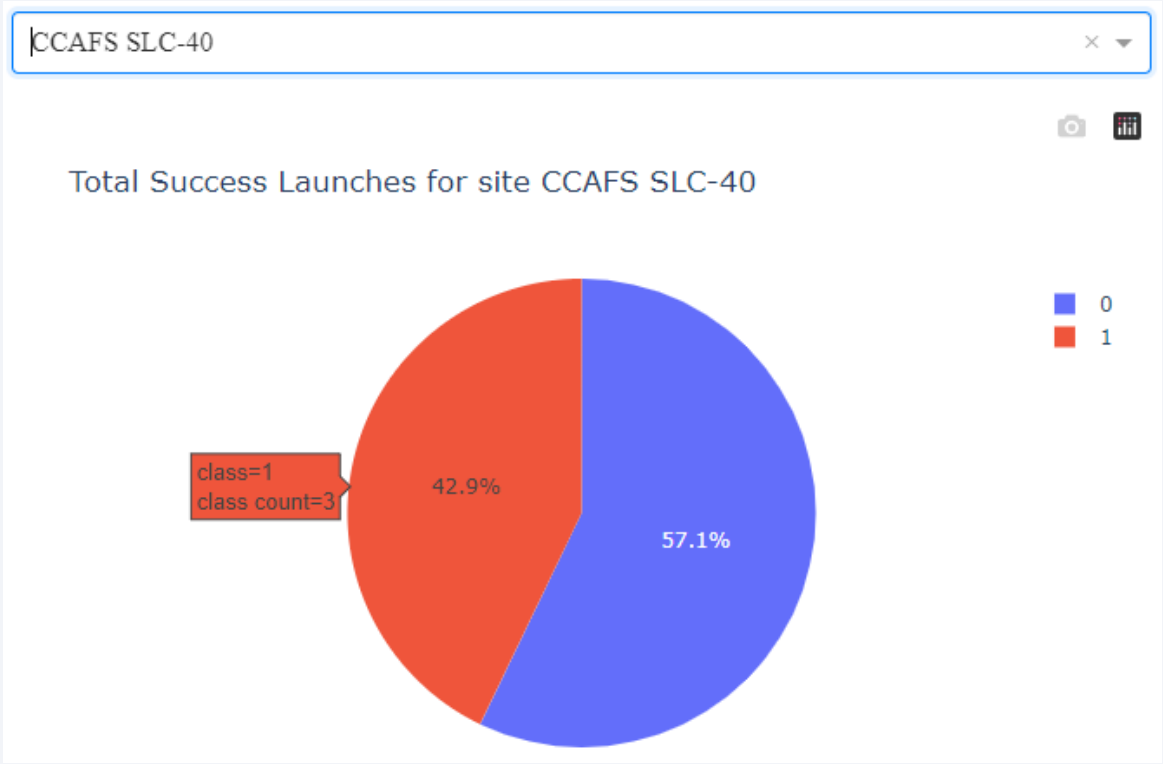
Number of Successful Launches

- **KSC LC-39A** has the largest number of successful launches among the 4 sites



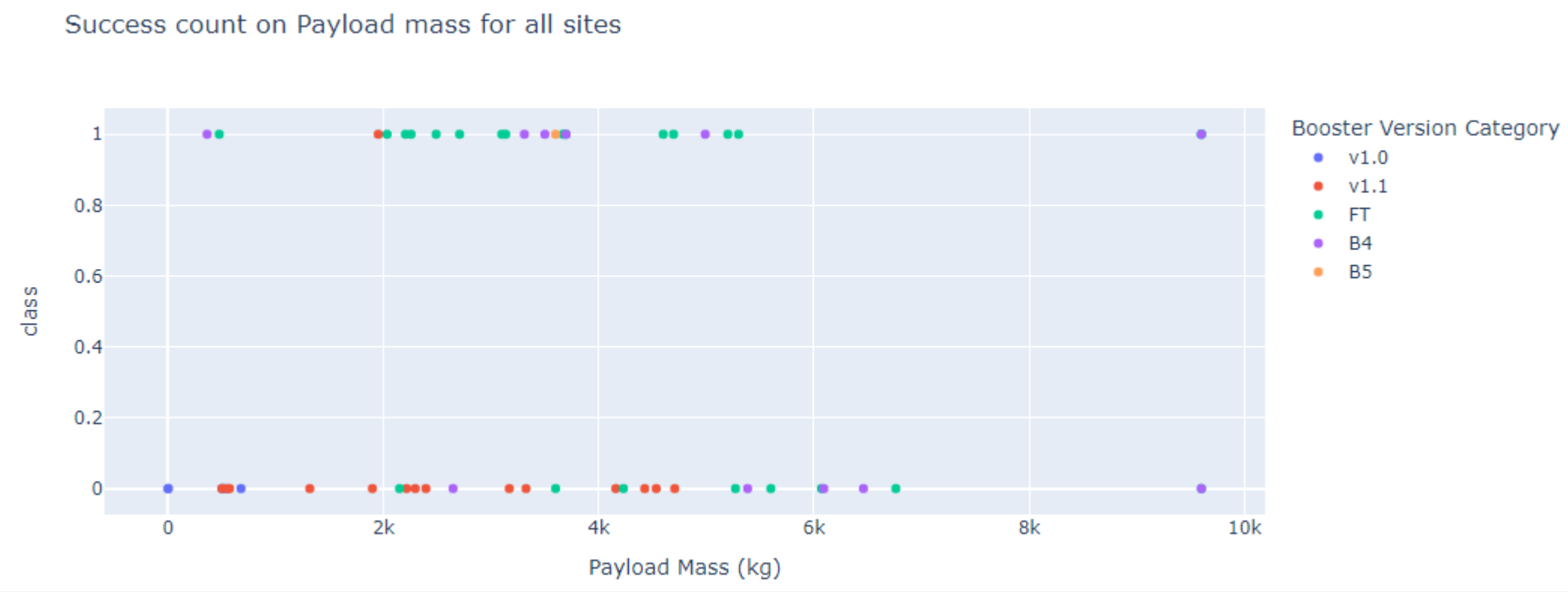
Launch Success Ratio

- **CCAFS SLC-40** is the site with highest success ratio, although it has only 3 successful launches, it is the site with lowest number of launches among the 4 ones



Payload vs. Launch Outcome

- Payload mass is an important factor to determine the success of a launch
- Payload range with highest launch success rate is 3000 – 4000 kg

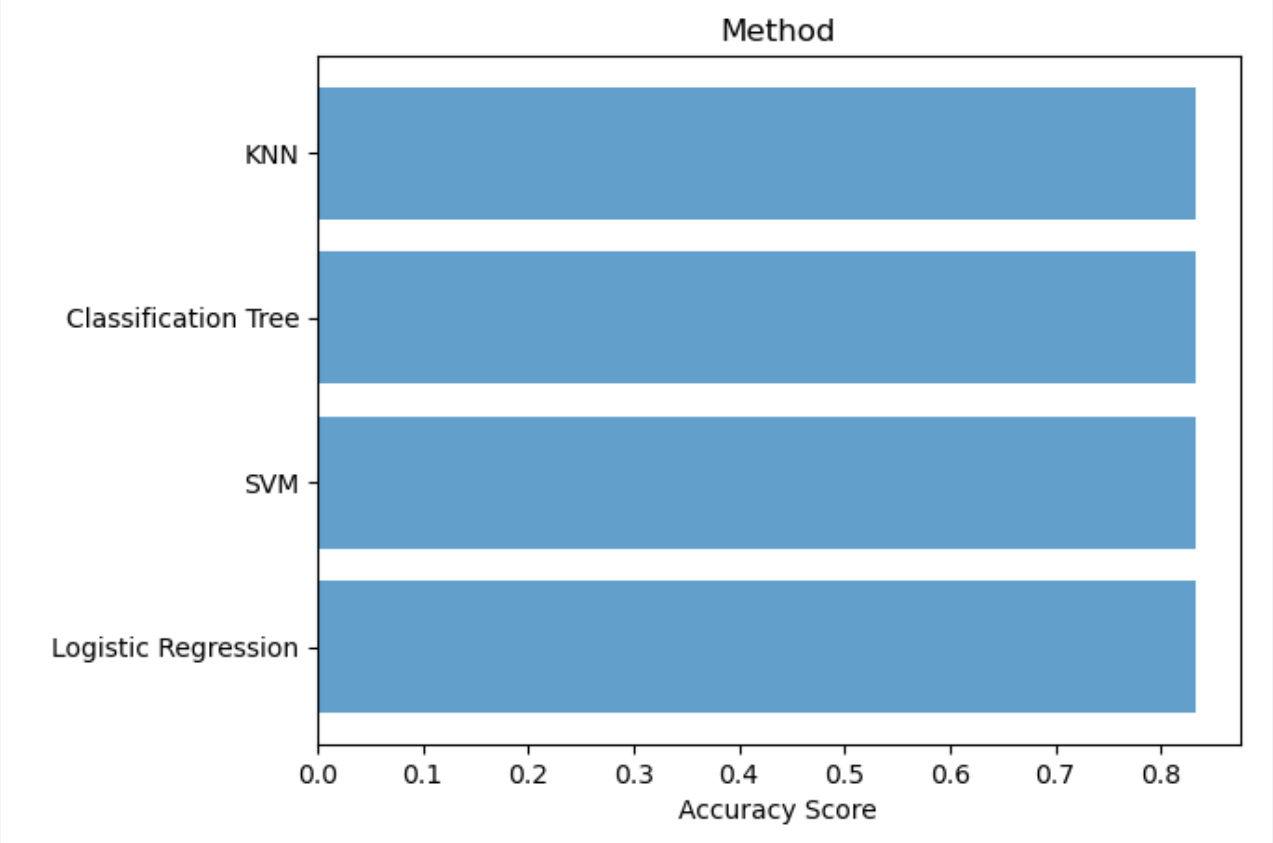


Section 5

Predictive Analysis (Classification)

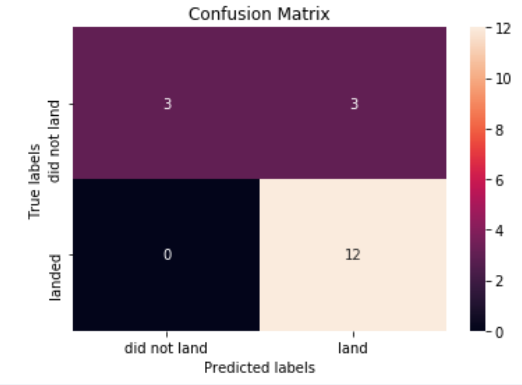
Classification Accuracy

- All models have the same classification accuracy
- Score of **0.8333333333333334**

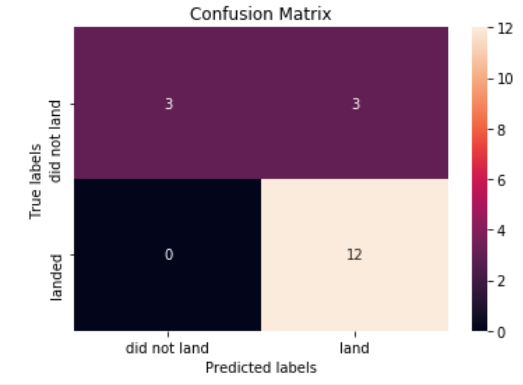


Confusion Matrix

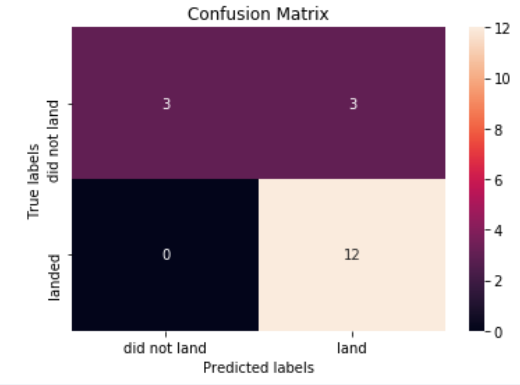
- Confusion matrixes are identical for all 4 evaluated methods since they perform similarly, with same accuracy
- It was observed they can distinguish between the different classes
- The major problem is related to false positives



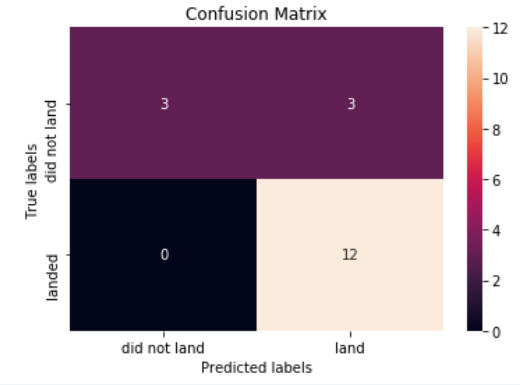
Logistic Regression



SVM



Classification Tree



KNN

Conclusions

- Falcon 9 obtained a success landing rate of 66.67%
- ES-L1, GEO, HEO and SSO orbits present the highest success rate
- Drop ship ASDS is the booster with highest number of successful landing outcomes: **41**
- Launch sites should be close to the coast as well as Equator line as much as possible
- Payload mass is an important factor to determine the success of a launch, the range with highest success rate is between 3000 and 4000 kg
- CCAFS SLC-40 is the site with highest success ratio (57.1%), although KSC LC-39A has the largest number of successful launches (41.7%)
- The 4 evaluated ML prediction methods perform similarly, with same accuracy score of 0.8333333333333334
- The major problem for all the ML methods is related to false positives

Thank you!