Winning Space Race with Data Science

Santiago Castrillón Zambrano
09/29/2022

Outline

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

Executive Summary

Methodologies:

- Data Collection
  - Spacex Web API
  - Web Scrapping
- Exploratory Data Analysis
  - Data Cleaning
  - Data Visualization
  - Interactive Visual Analytics
- Prediction Using Machine Learning
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Trees
  - K-Nearest Neighbors (KNN)

Results:

- The Exploratory Data Analysis allowed to identify important information about the success of launches.
- The Machine Learning methods used showed similar accuracy scores on the test set.

Introduction

In this project, we'll make a prediction on how well the Falcon 9 first stage will land. On its website, SpaceX promotes Falcon 9 rocket launches for USD62 million; other suppliers charge upwards of USD165 million for each launch.

A large portion of the savings is due to SpaceX's ability to reuse the first stage. Therefore, if we can figure out if the first stage will land, we can figure out how much a launch will cost. If a different business want to compete with SpaceX for a rocket launch, it may use the information provided here. You will receive an overview of the issue and the resources you need to finish the course in this module.

Section 1

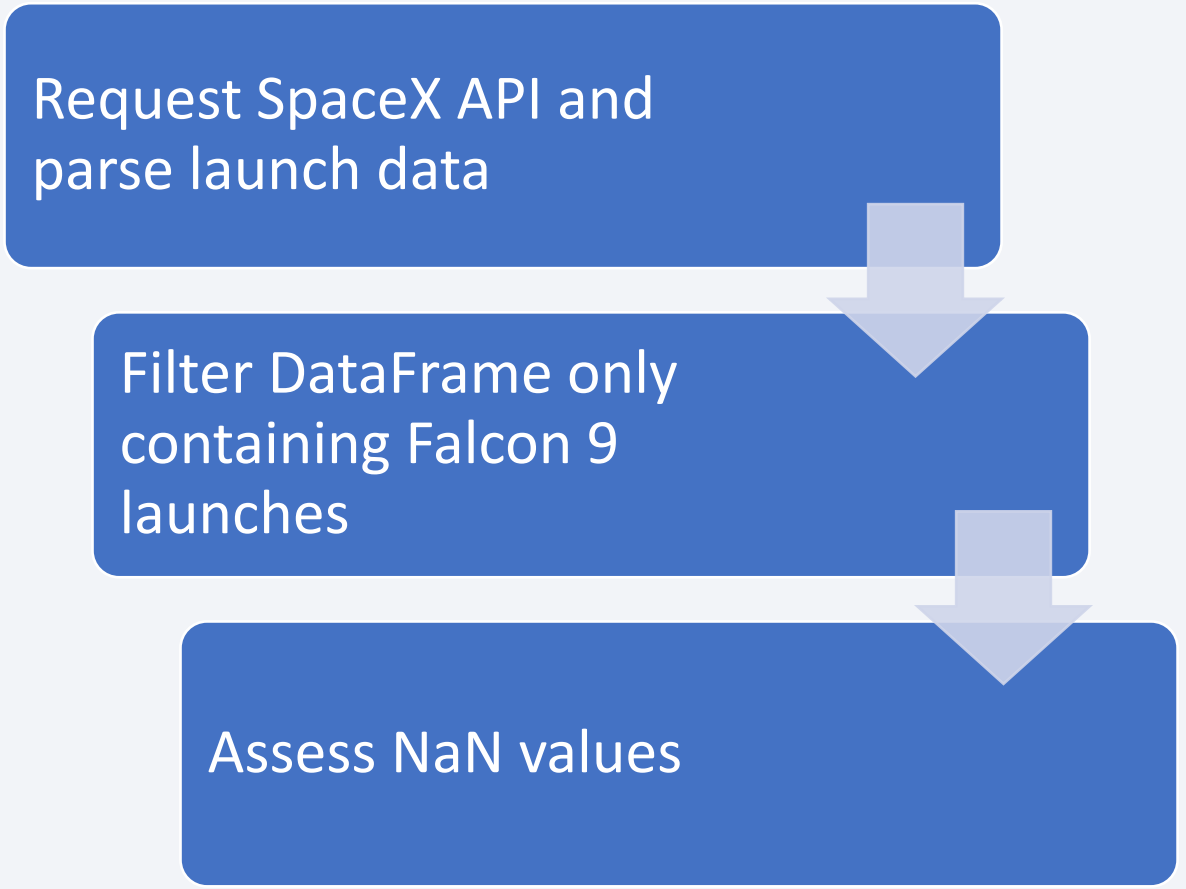
Methodology

Methodology

- **Data collection methodology:**
  - Data from Space X was obtained from 2 sources:
    - Space X API
    - WebScraping
- **Perform data wrangling :**
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
  - The data collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

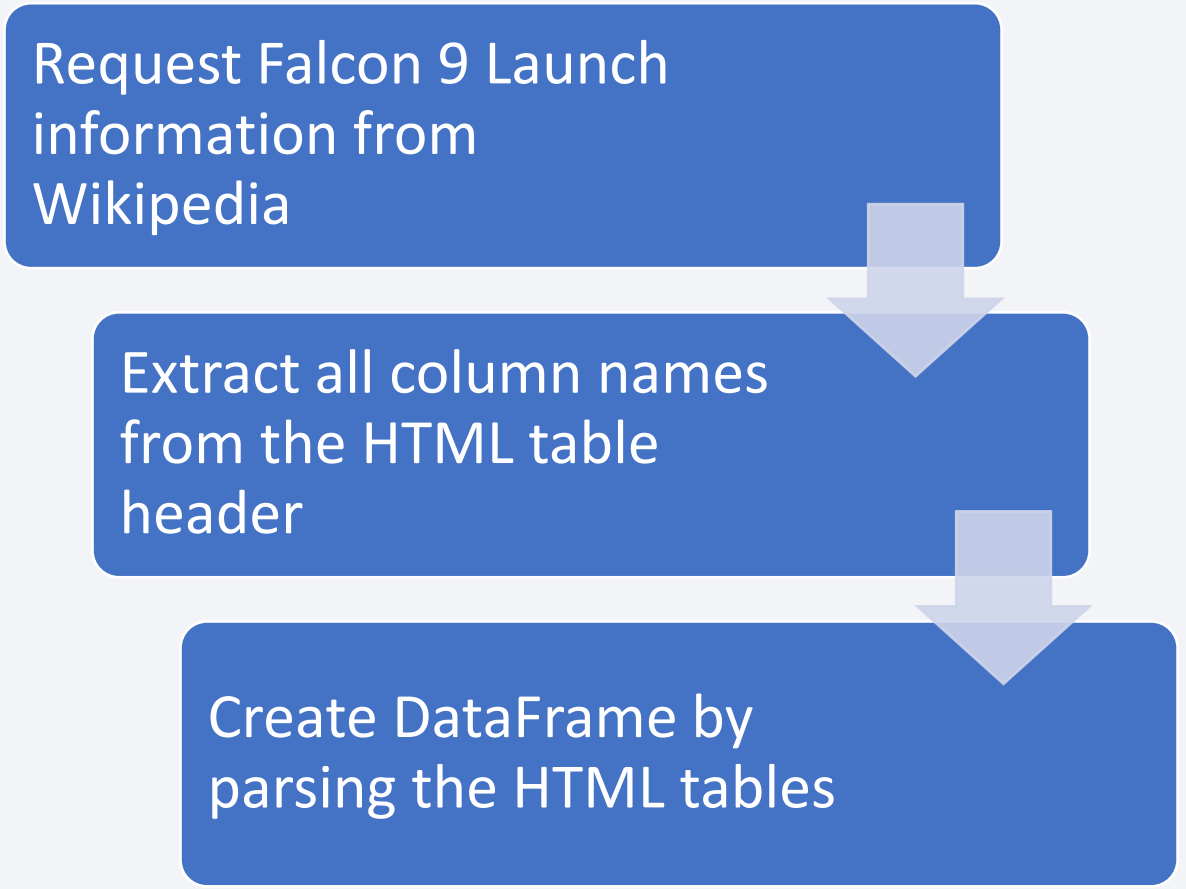
Data Collection – SpaceX API

- Providing a public API, SpaceX is where information can be acquired, then employed.
- The usage of this API resembled the following a flowchart
- GitHub URL of the completed notebook



Data Collection - Scraping

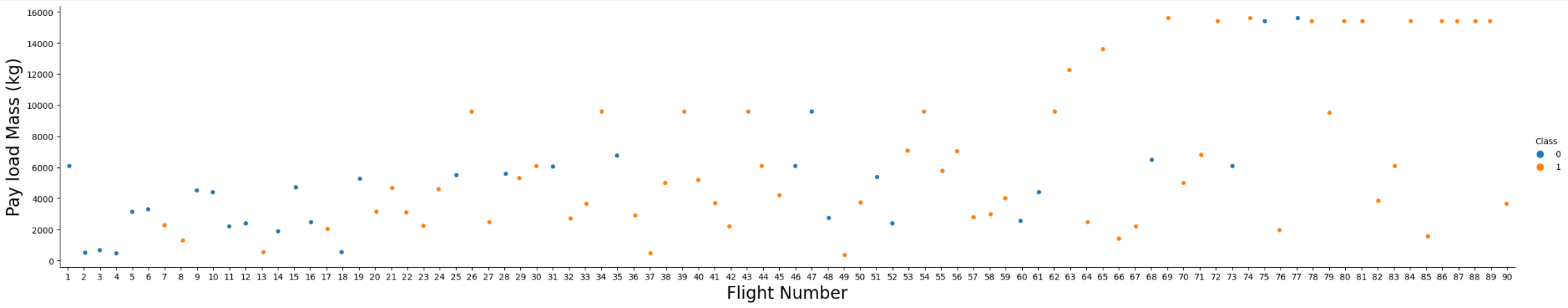
- Data from SpaceX launches is obtained from Wikipedia
- GitHub URL of the completed notebook



Data Wrangling

- Performed Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- Compute launches by site, orbit occurrences, and mission outcomes by orbit type.
- Create landing outcome label from Outcome column
- GitHub URL of the completed notebook

EDA with Data Visualization



We can see that the initial stage has a higher chance of successfully landing as the number of flights rises. It appears that the cargo mass is also significant; the heavier the payload, the less probable it is that the first stage would return.

- GitHub URL of the completed notebook

EDA with SQL

- **SQL queries:**
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name begin with the string 'CCA';
  - Total payload mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.
  - GitHub URL of the completed notebook

Build an Interactive Map with Folium

- Using Folium Maps, Markers, circles, lines and marker clusters were used
  - Markers indicate points like launch sites
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site
  - Lines are used to indicate distances between two coordinates.
  - These maps allow us to better understand the problem and the data. We can see all launch sites, their surroundings and the number of successful and unsuccessful landings

GitHub URL of the completed notebook

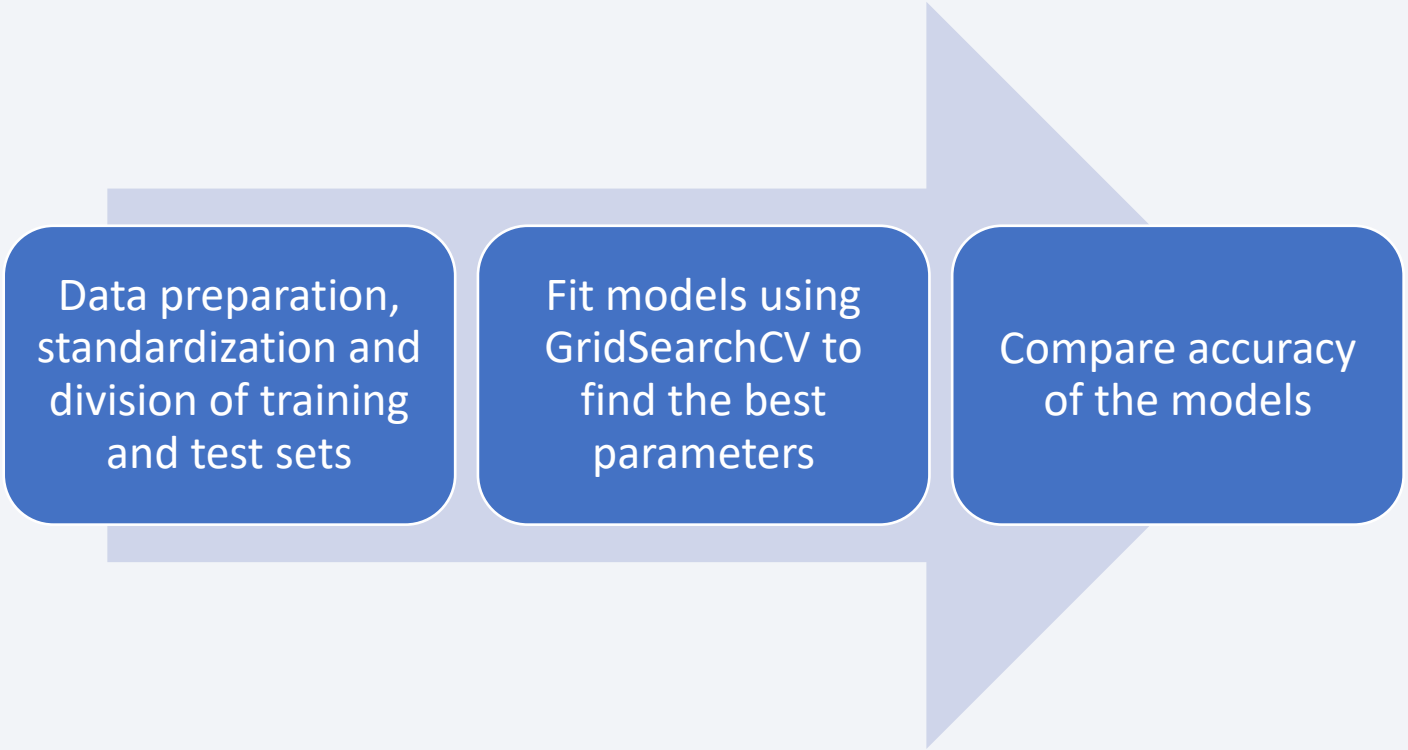
Build a Dashboard with Plotly Dash

- Dashboard has dropdown, pie chart, rangeslider and scatter plot components
  - Dropdown allows user to choose the launch site or all launch sites
  - Pie chart showing the total success comparing all sites or the success and the failure for the launch site chosen with the dropdown.
  - Rangeslider allows a user to select a payload mass
  - Scatter chart showing relationship between Success and Payload Mass
- This combination allowed for a quick analysis of the relationship between payloads and launch sites, assisting in determining the best place to launch based on payloads.

GitHub URL of the completed notebook

Predictive Analysis (Classification)

- Classification models: Logistic Regression, SVM, Decision Trees, KNN



GitHub URL of the completed notebook

Results

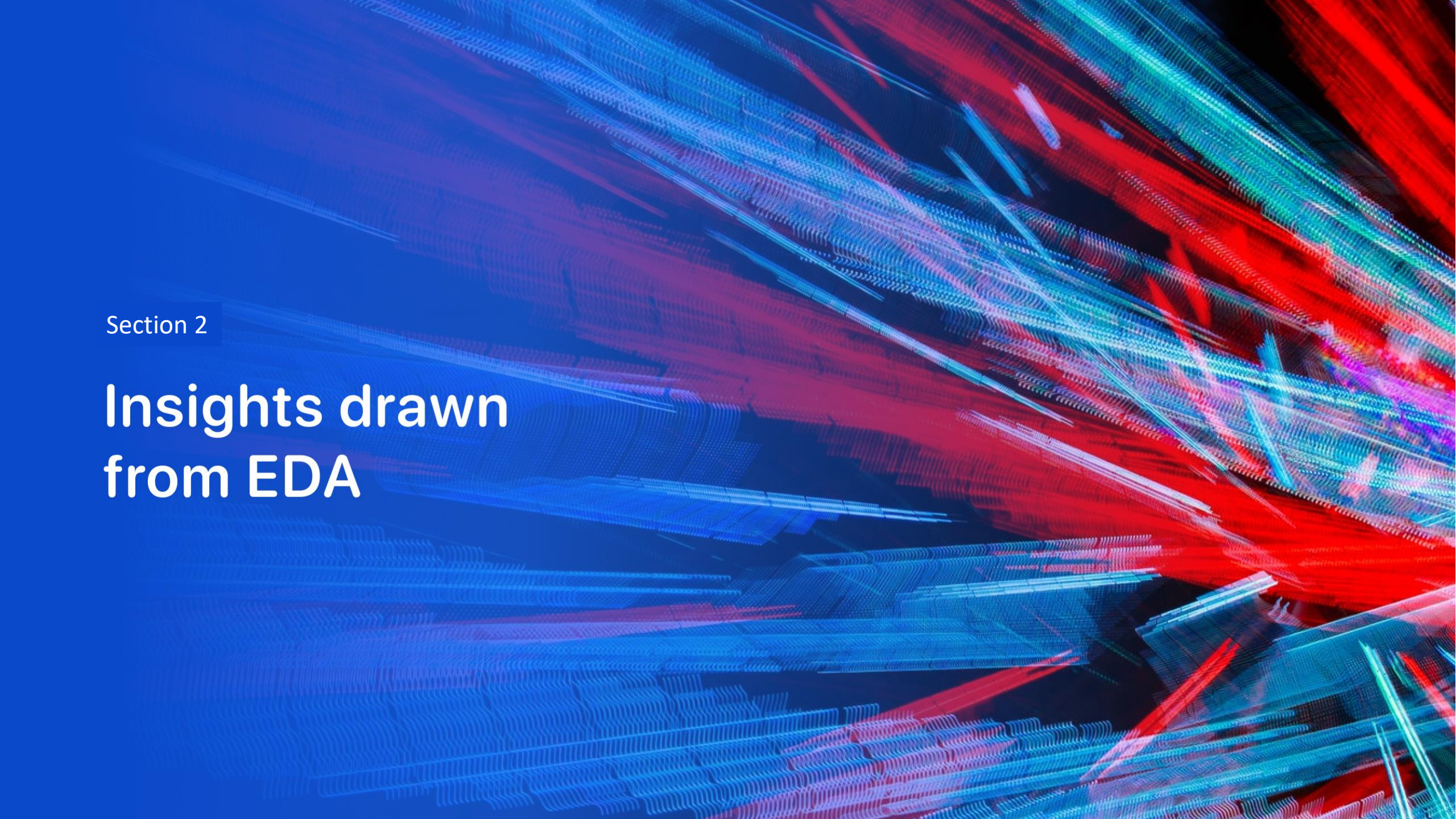
Results of exploratory data analysis:

- Space X launches from four different locations.
- The first launches were carried out by Space X and NASA;
- The F9 v1.1 booster's average payload is 2,928 kg.
- Many Falcon 9 booster versions successfully landed in drone ships with payloads greater than the average
- Almost all mission outcomes were positive
- In 2015, two booster versions failed to land in drone ships: F9 v1.1 B1012 and F9 v1.1 B1015.
- As time passed, the number of successful landings increased.

Results

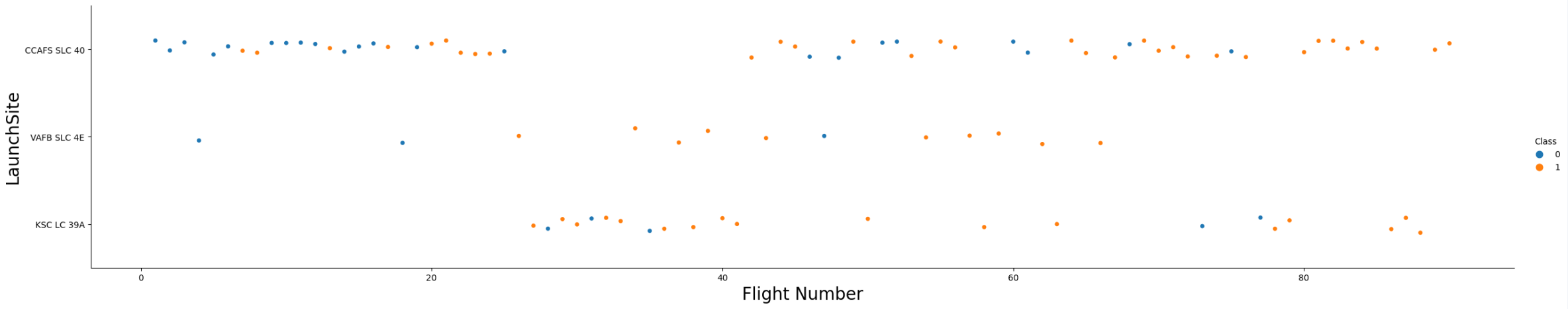
- All the Machine Learning models show the same accuracy in the test data set.
- Given that train and test accuracy is close there are low probabilities of over fitting.
- It could be interesting to increase the sample size to determine the better model.

	Train_Score	Test_Score
KNN	0.847222	0.833333
Decision Tree	0.875000	0.833333
SVM	0.847222	0.833333
LogisticRegression	0.847222	0.833333


Section 2
Insights drawn from EDA

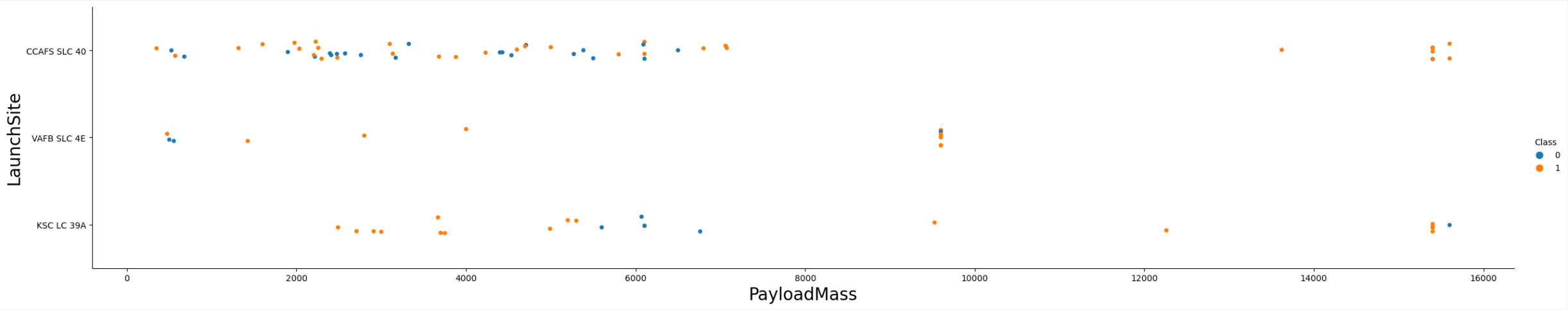
Flight Number vs. Launch Site

- It is possible to confirm that the best launch site nowadays is CCAF5 SLC 40, where most recent launches were successful, followed by VAFB SLC 4E and KSC LC 39A.



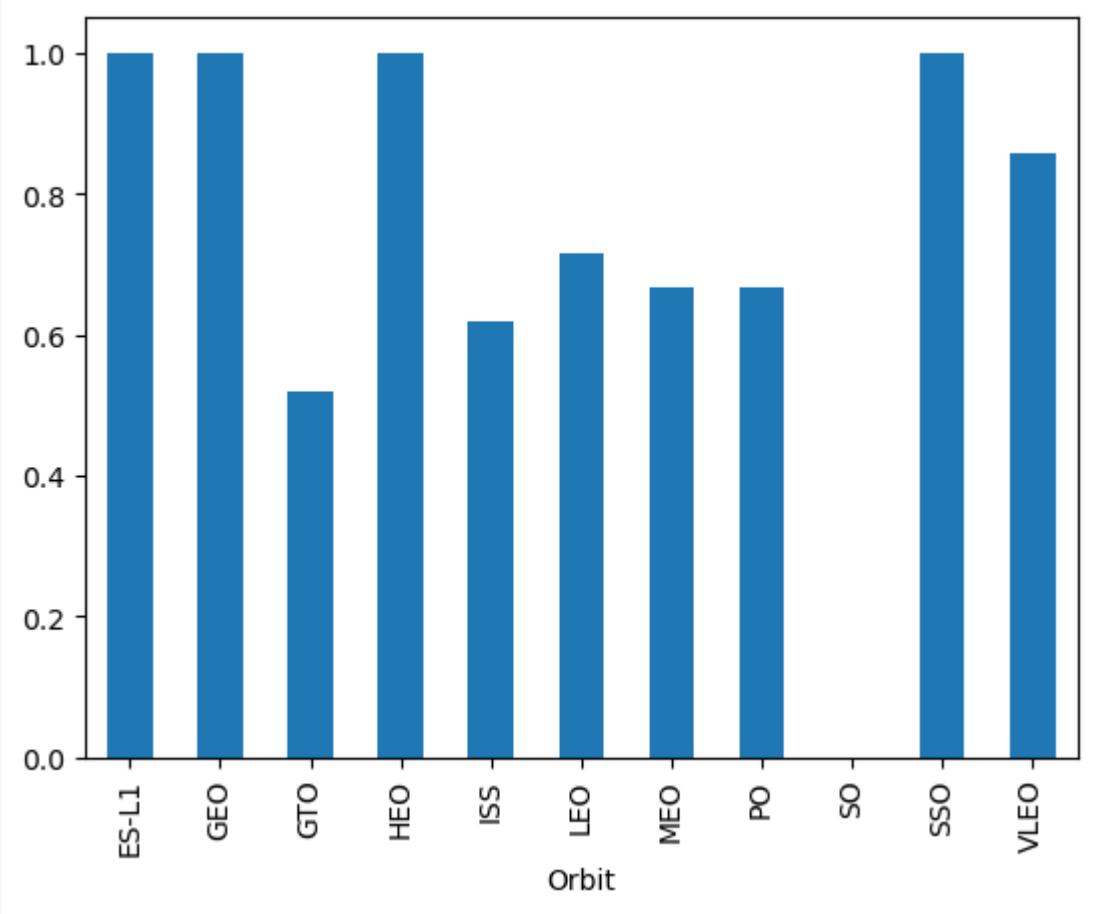
Payload vs. Launch Site

- Payloads weighing more than 9,000kg show a high success rate.
- Payloads weighing more than 12,000kg appear to be limited to the CCAFS SLC 40 and KSC LC 39A launch sites.
- There are no rockets launched for heavy payload mass at the VAFB-SLC launch site (greater than 10000Kg).



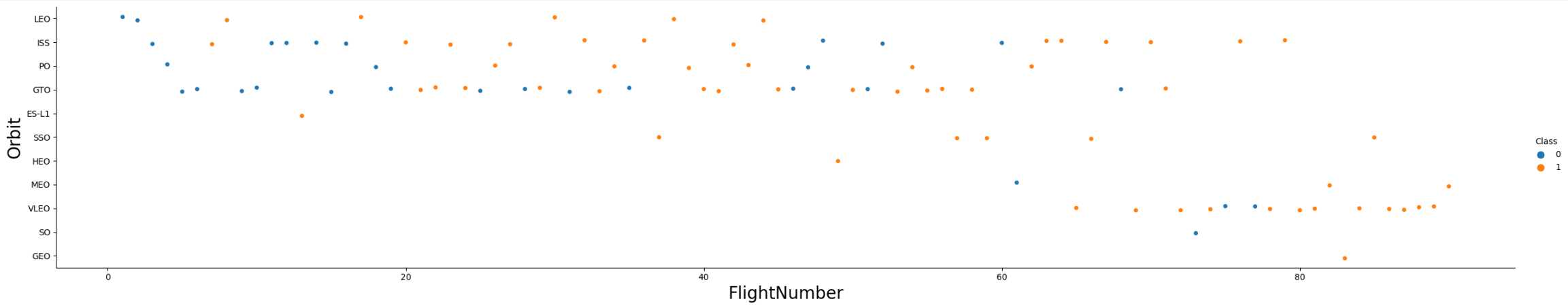
Success Rate vs. Orbit Type

- The highest success rates happen in orbits:
  - **ES-L1**
  - **GEO**
  - **HEO**
  - **SSO.**



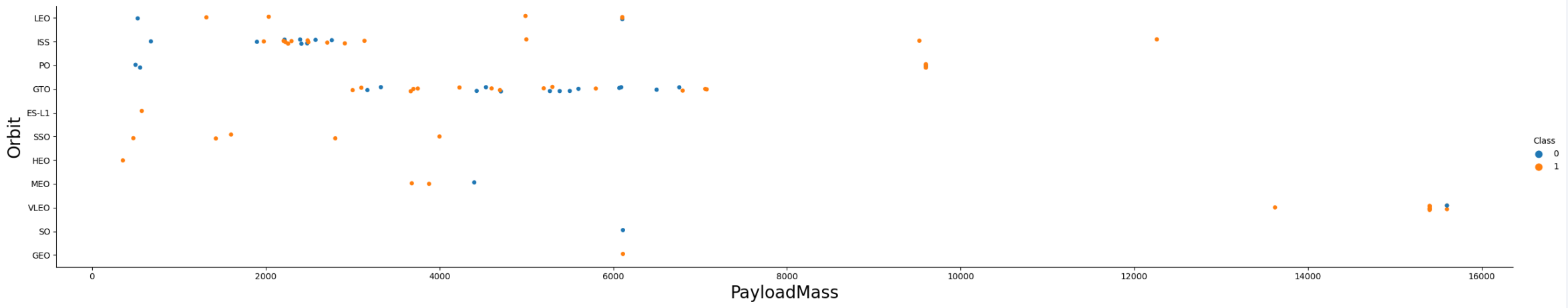
Flight Number vs. Orbit Type

- In LEO orbit, success appears to be related to the number of flights; however, in GTO orbit, there appears to be no relationship between flight number.



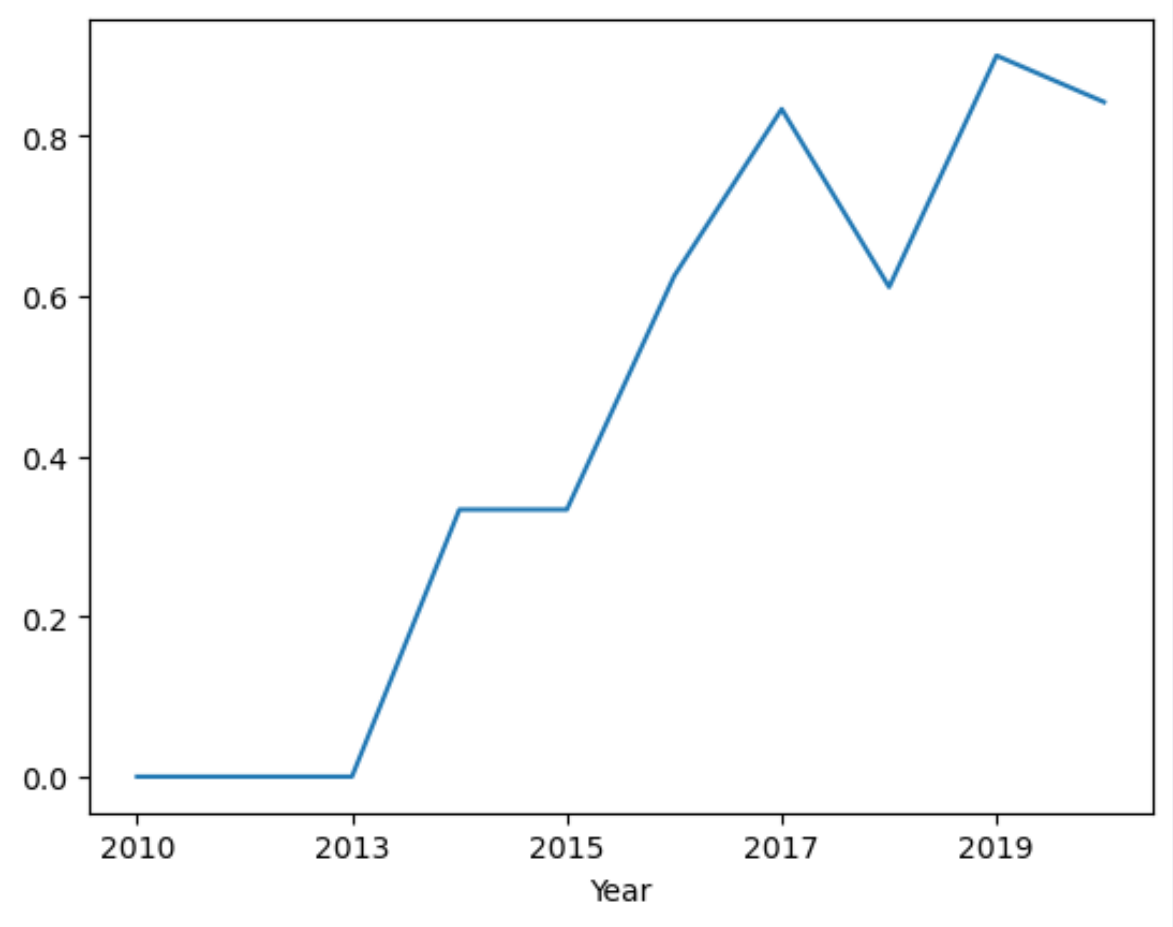
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- For GTO, we can't tell the difference because both positive and negative landing rates (missed missions) are present.



Launch Success Yearly Trend

- Success rate started increasing in 2013 andkept until 2020



All Launch Site Names

In [66]:

```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL ORDER BY 1;
```

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

Out[66]:

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

Launch Site Names Begin with 'CCA'

In [67]:

```
%%sql
SELECT * FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

Out[67]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
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

Total Payload Mass

In [68]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL
WHERE PAYLOAD  LIKE '%CRS%';
```

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

Out[68]:

TOTAL_PAYLOAD
111268

Total payload calculated above, by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

Average Payload Mass by F9 v1.1

In [69]:
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

Out[69]:

AVG_PAYLOAD
2928.4

Filtering data by the booster version above and calculating the average payload mass.

First Successful Ground Landing Date

In [71]:

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL
WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

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

Out[71]:

FIRST_SUCCESS_GP
01-05-2017

By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence.

Successful Drone Ship Landing with Payload between 4000 and 6000

In [72]:

```
%%sql
SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING_OUTCOME = 'Success (drone ship)';
```

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

Out[72]:

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

The **WHERE** and **AND** clauses filter the dataset to return the booster version where landing was successful and payload mass is between 4000 and 6000 kg

Total Number of Successful and Failure Mission Outcomes

In [73]:

```
%%sql
SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL
GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

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

Out[73]:

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

Grouping mission outcomes and counting records for each group led us to get the number of successful and failed mission outcomes

Boosters Carried Maximum Payload

In [75]:
```
%%sql
SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;
```

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

Out[75]:

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

With the **MAX** function, we used a subquery to filter data by returning only the heaviest payload mass. The main query uses subquery results to return the most powerful booster version with the heaviest payload mass.

2015 Launch Records

In [98]:

```
%%sql
SELECT MISSION_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE, substr(Date, 4, 2) as Month  FROM SPACEXTBL
WHERE substr(Date,7,4)='2015';
```

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

Out[98]:

Mission_Outcome	Booster_Version	Launch_Site	Month
Success	F9 v1.1 B1012	CCAFS LC-40	01
Success	F9 v1.1 B1013	CCAFS LC-40	02
Success	F9 v1.1 B1014	CCAFS LC-40	03
Success	F9 v1.1 B1015	CCAFS LC-40	04
Success	F9 v1.1 B1016	CCAFS LC-40	04
Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40	06
Success	F9 FT B1019	CCAFS LC-40	12

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

In [91]:
```
%%sql
SELECT LANDING_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL
WHERE DATE BETWEEN '04-06-2010' AND '20-03-20'
GROUP BY LANDING_OUTCOME ORDER BY QTY DESC;
```

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

Out[91]:

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

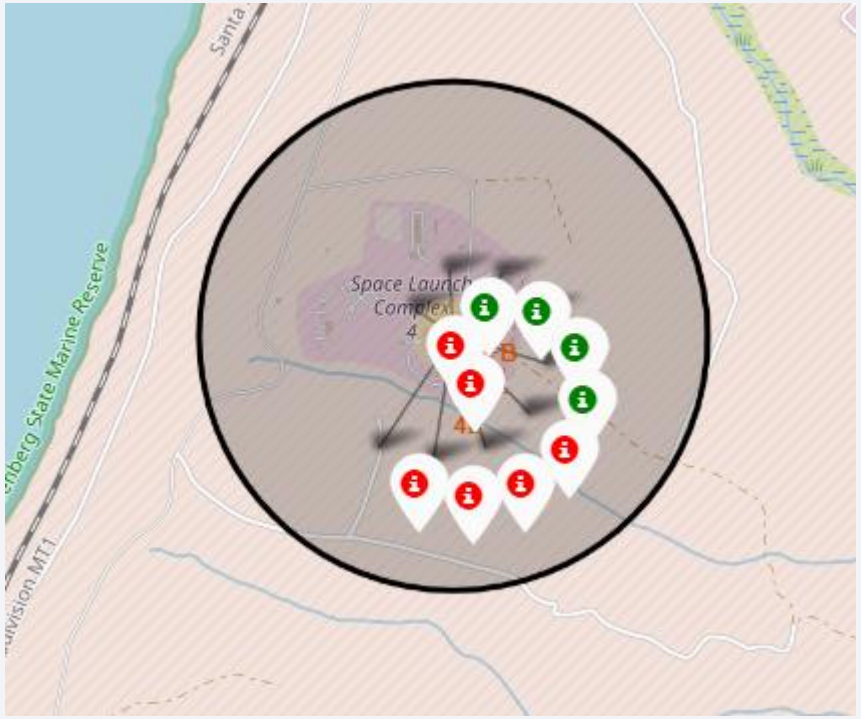
Section 3

Launch Sites Proximities Analysis

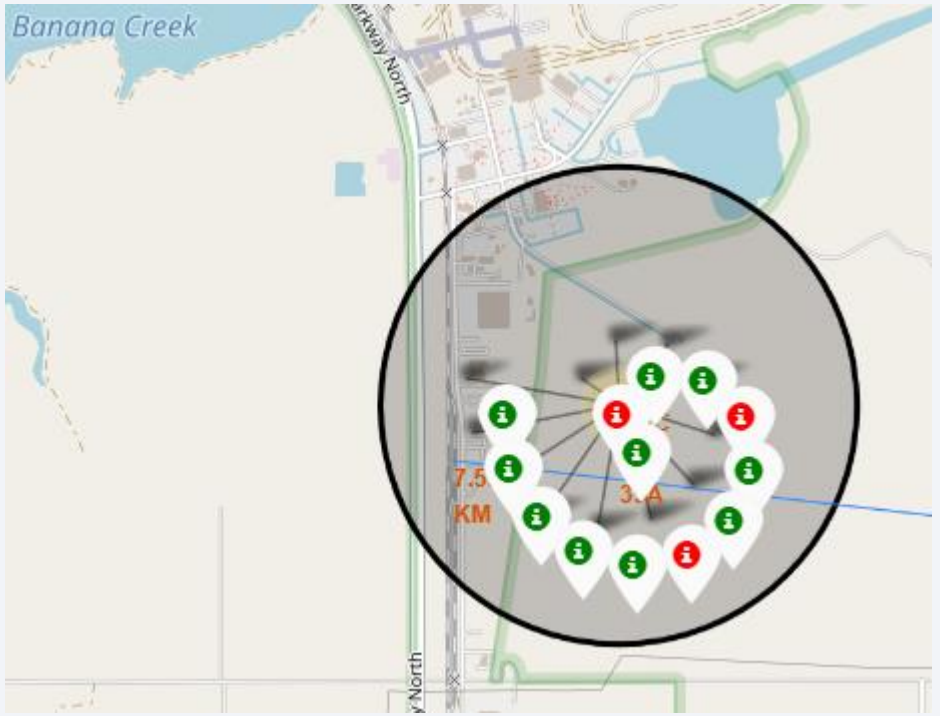
Launch Sites

Most launches take place at the east coast of the USA.

Outcome Examples



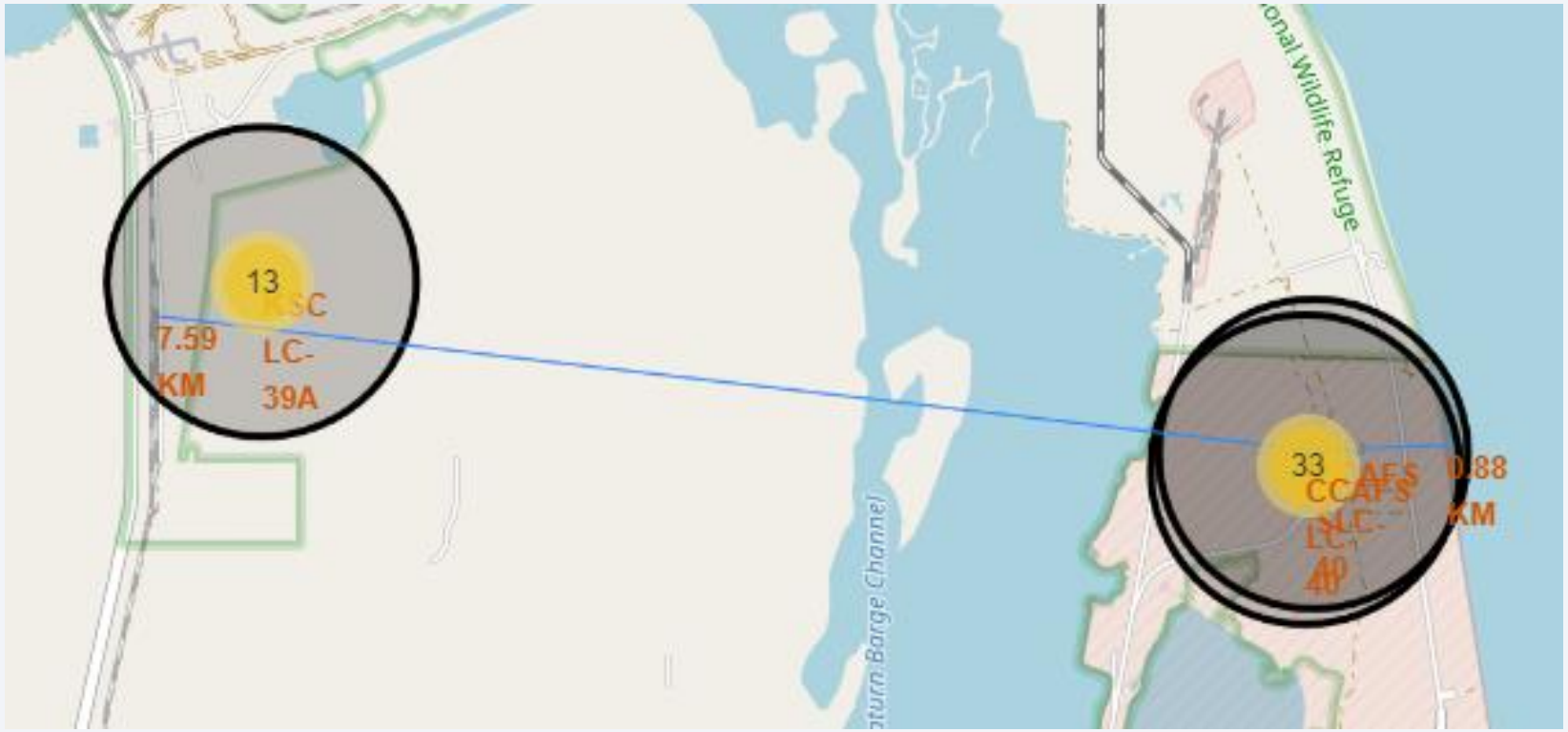
VAFB SLC-4E outcomes.



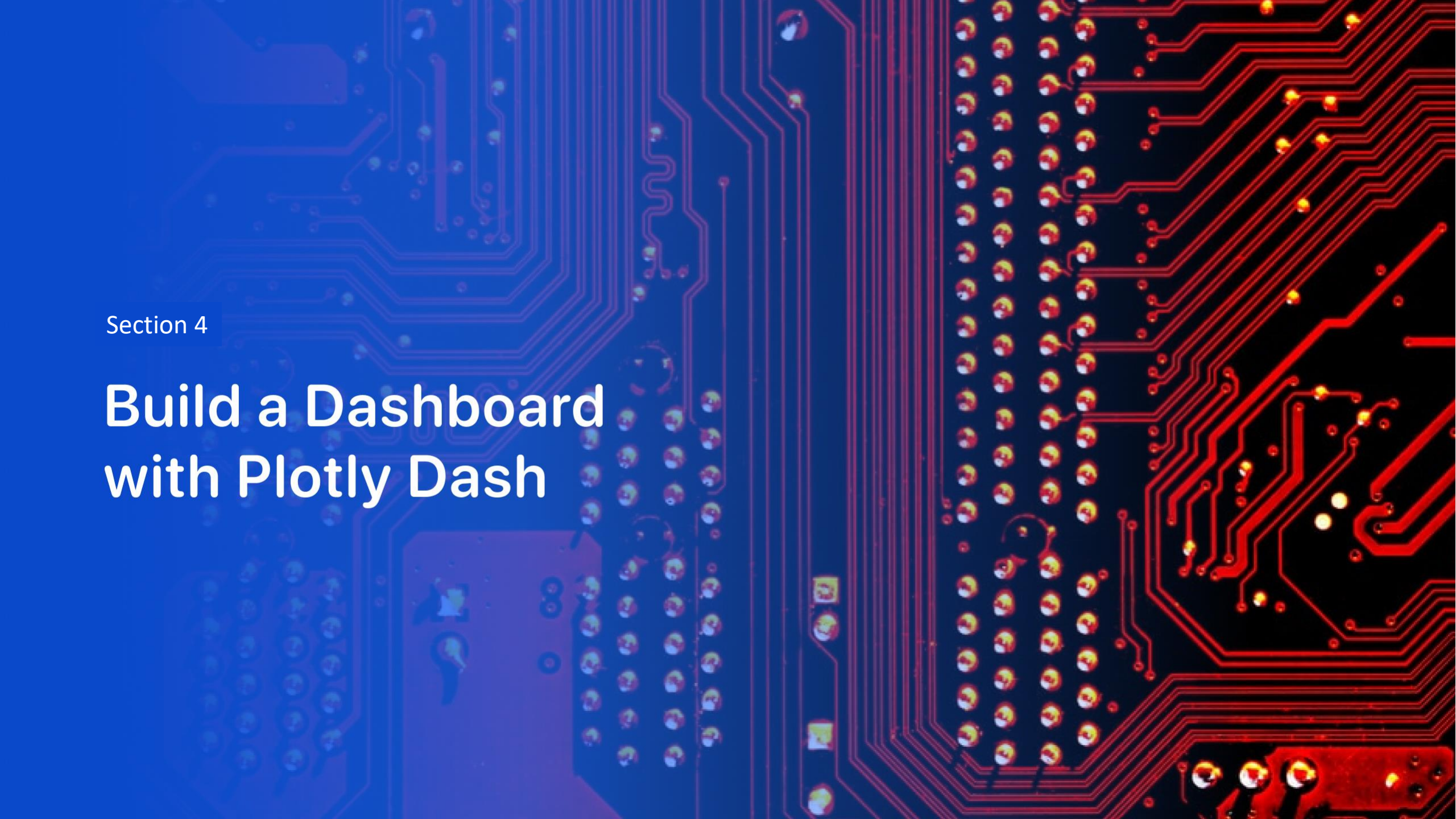
VAFB SLC-4E outcomes.

Green marker represents successful launches. Red marker represents unsuccessful launches.

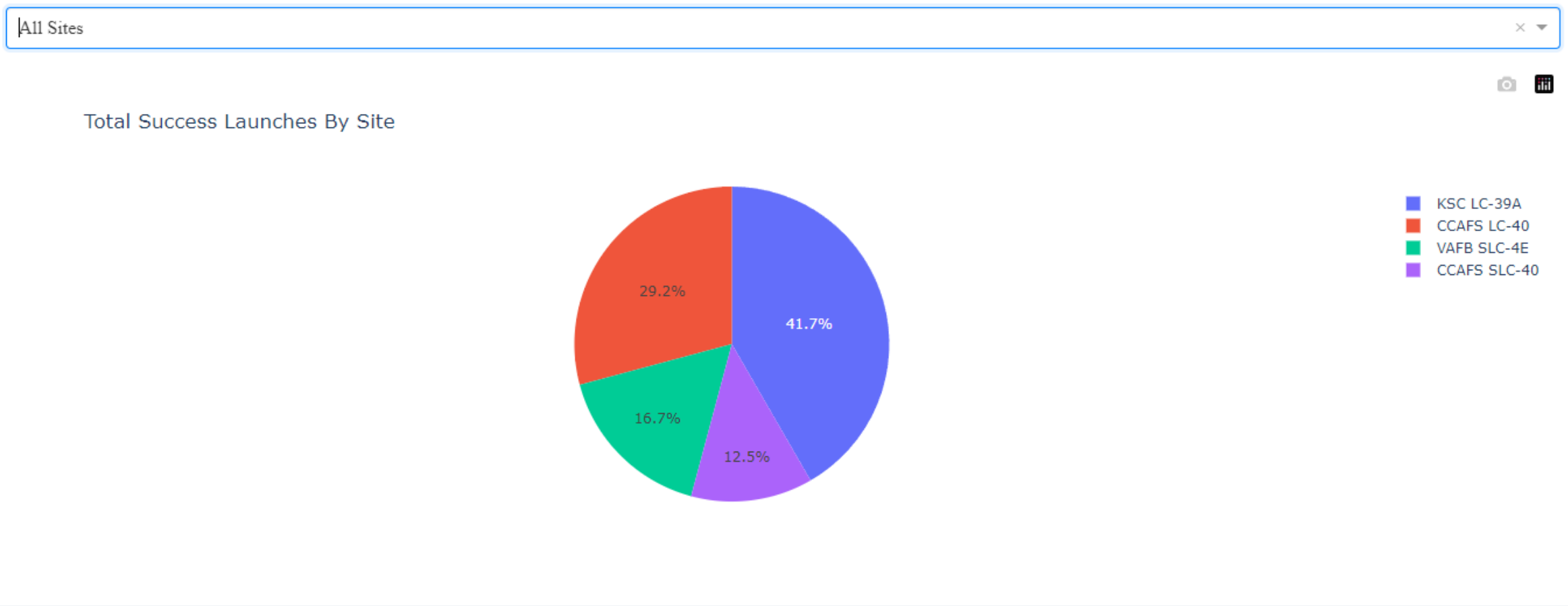
CCAFS SLC-40 is very close to the sea,



Using interactive analytics, it was possible to determine that launch sites are in safe locations, such as near the sea, and had a good logistic infrastructure surrounding them, as it can be shown in the map, they are close to the sea and railways.


Section 4
Build a Dashboard with Plotly Dash

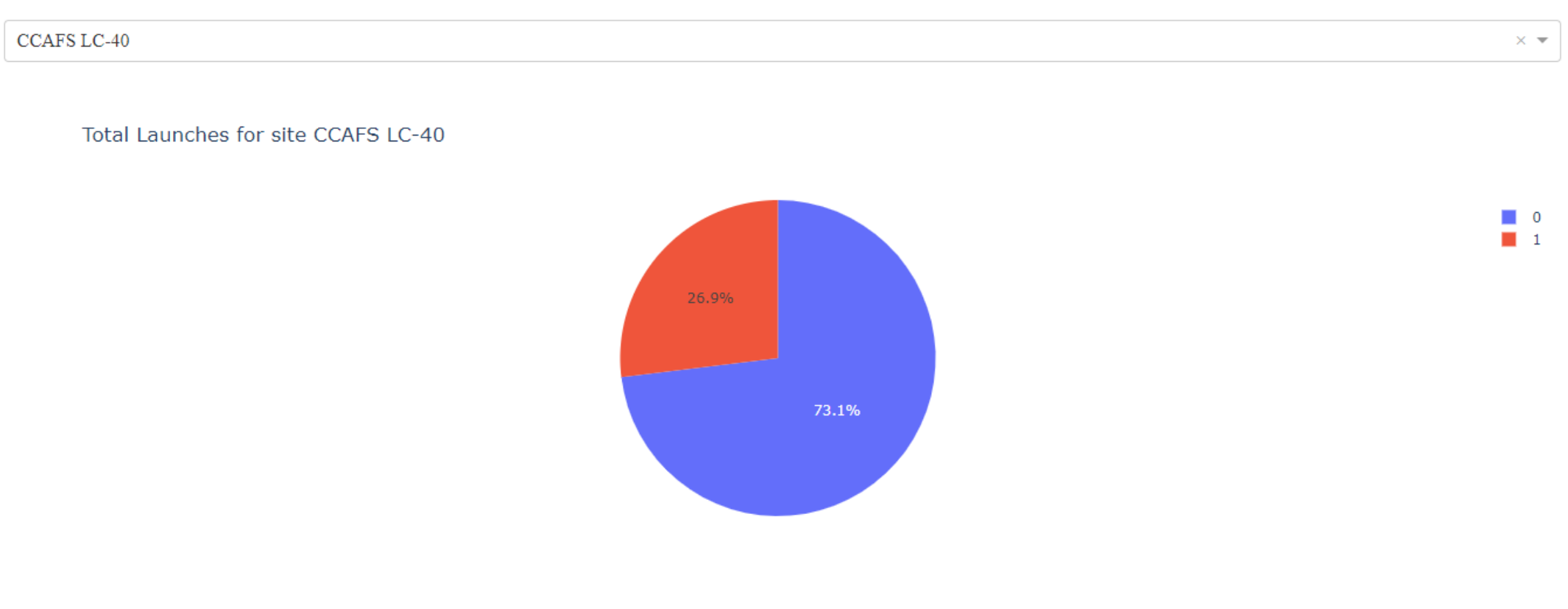
Successful launches per sites



The launch site shows to be an important factor

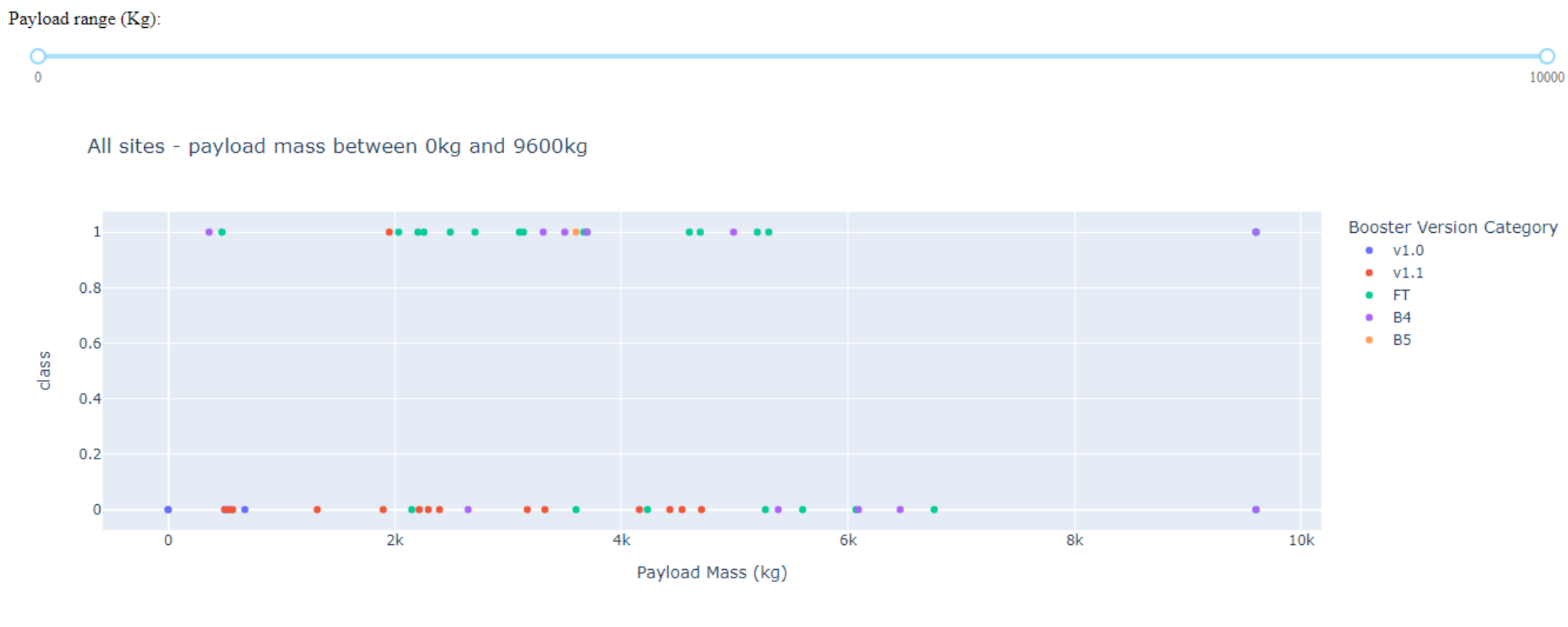
Launch Success in CCAFS LC-40

- Although CCAFS LC-40 has the second highest success rate, most of its launches have ended in failures



Correlation Payload Mass vs Class per Version Category

- The most successful payloads are FT boosters under 6,000kg

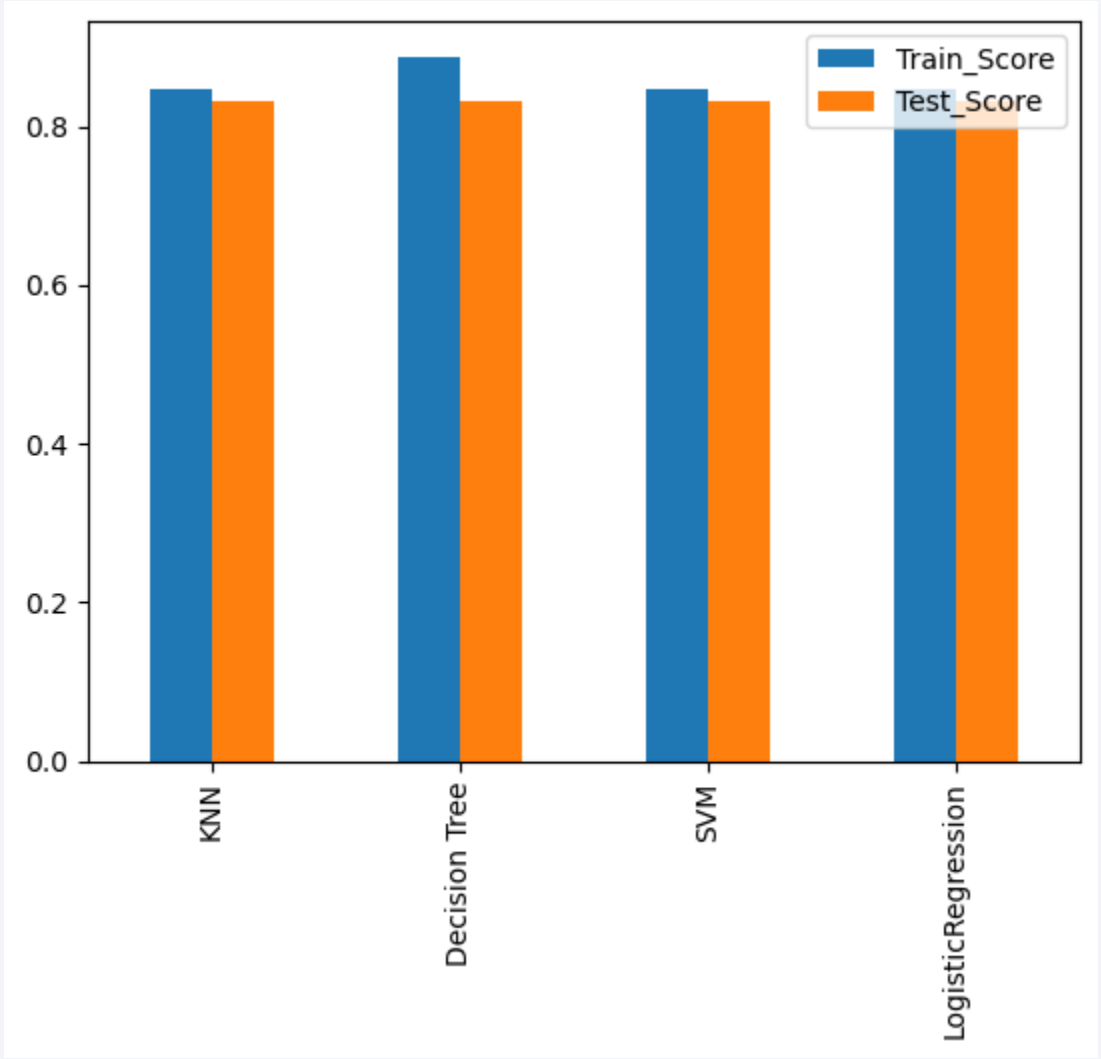


Section 5

Predictive Analysis (Classification)

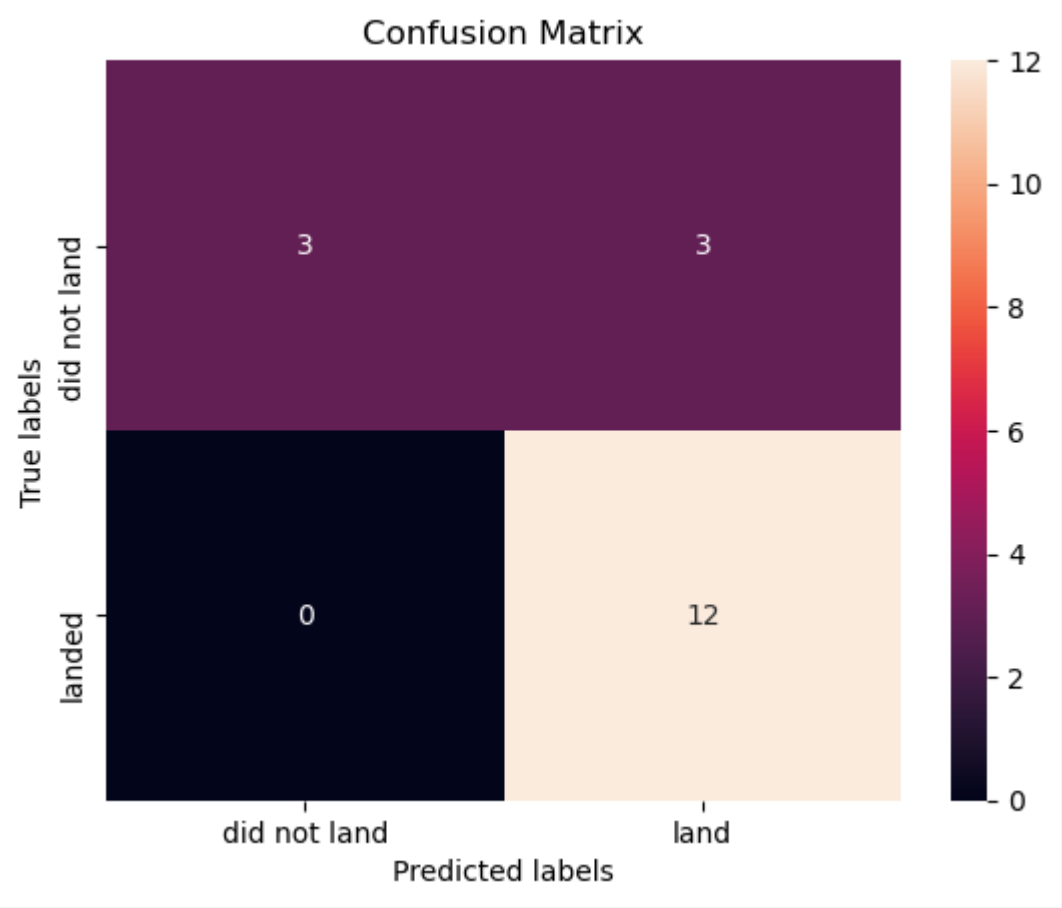
Classification Accuracy

- The Decision Tree model shows the highest accuracy in the train set
- All models show the same accuracy in the test set



Confusion Matrix

- All models show the same confusion Matrix.
- The problem of the models' accuracy comes from false positives



Conclusions

- A mission's success can be explained by several factors, including the launch site, orbit, and, most importantly, the number of previous launches.
- Depending on the orbits, payload mass can be a criterion to consider for mission success. Some orbits necessitate either a light or heavy payload mass. However, low weighted payloads outperform heavy weighted payloads in gene
- Although most mission outcomes are successful, successful landing outcomes appear to improve over time as technology improves.
- All methods showed the same accuracy on the test data although Decision Trees had better performance on the train set.

Thank you!