

# Winning Space Race with Data Science

TOSHIO SEKIGUCHI  
09/16/2022



# Outline

---

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

# Executive Summary

---

## Summary of methodologies

The process of data science is this. First, collected the data from SpaceX site with SpaceX API and from Wikipedia with Web Scraping. Second, edited the data to create a landing outcome label ('Class') for further Data Analysis and Machine Learning. Third, Performed exploratory data analysis (EDA) using visualization and SQL to find out characteristics of data and correlation among columns, such as success ratio vs. orbit type. Forth, performed interactive analytics with Folium and Plotly Dash to understand the key success factor of launch, such as launch success vs. payload. Last, performed predictive analysis using classification models to find out which model is best suited for predicting future launch result.

## Summary of all results

- KSC LC-39A is the highest success ratio site.
- Overall after 20 flights, success ratio seems improved.
- Orbit type ES-L1, GEO, HEO, and SSO, and VLEO shows highly success rate.
- Success ratio of heavy lifting more than 5k kg payload was less than 30%, i.e., 3/11.
- Machine Learning using Tree model shows the highest accuracy score of 0.89; we can predict if a future-launch success with features such as Payload, Booster Version, Launch Site, Orbit Type, etc.

# Introduction

---

## Project background and context

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.

## Problems you want to find answers

- Find some patterns in the data and determine what would be the label for training supervised models.
- Predict if the first stage will land given the data from the history of launch records.

Section 1

# Methodology

# Methodology

---

## Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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

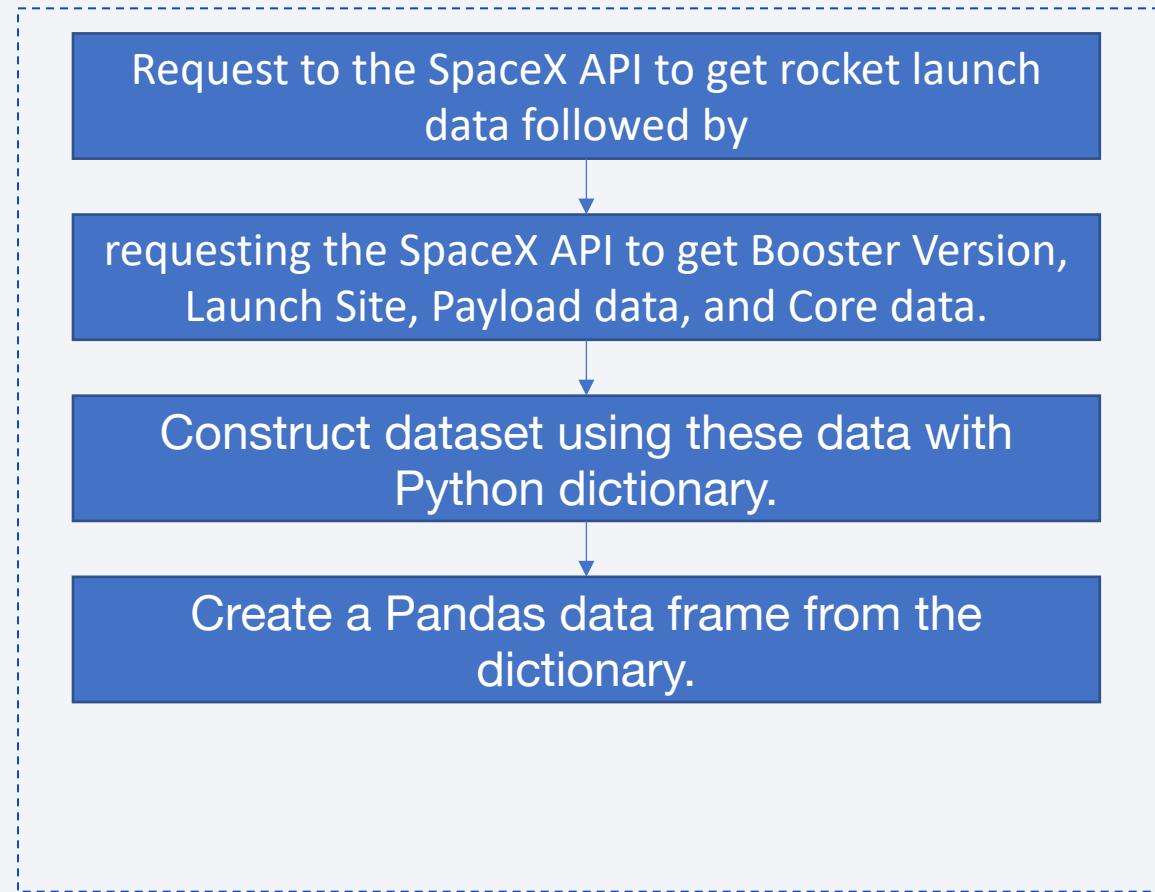
# Data Collection

---

- Data sets were collected in two ways: SpaceX API and Web Scraping.
- The following two slides describe how data sets were collected.

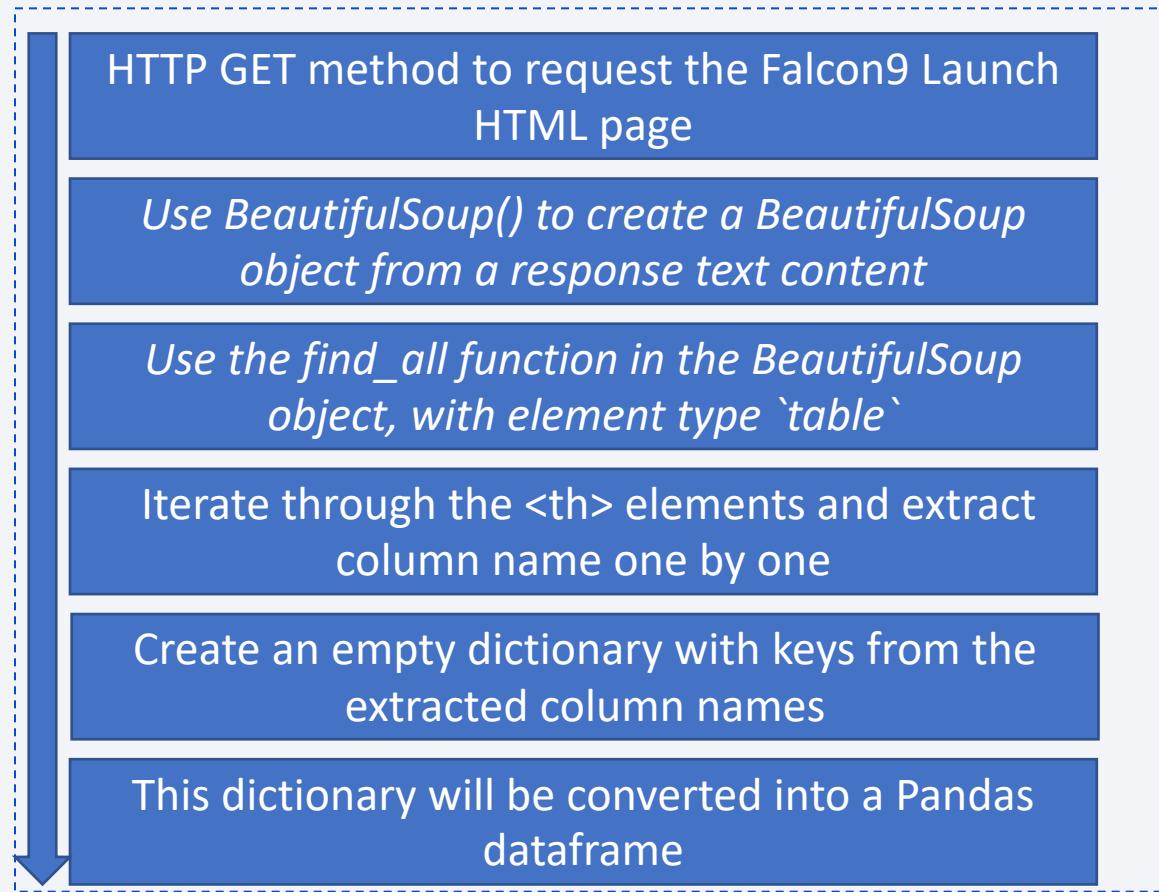
# Data Collection – SpaceX API

- The data collection with SpaceX REST calls is presented in the right flowchart.
- <https://github.com/tosshee/course-a-ds-capstone-project/blob/a6c299ad9560dfa638e861eccafabf6a52ce9c22/jupyter-labs-spacex-data-collection-api.ipynb>



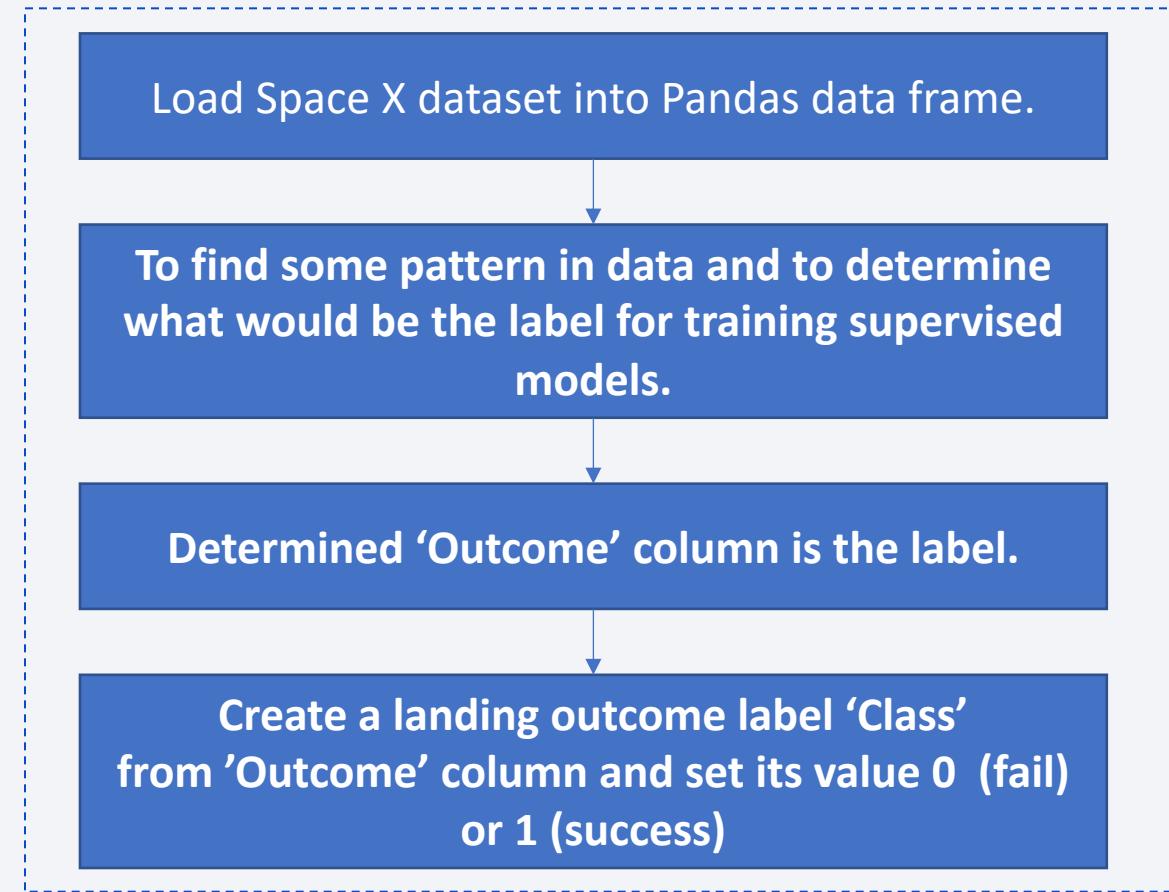
# Data Collection - Scraping

- The web scraping process is presented in the right flowchart.
- <https://github.com/tosshee/coursera-ds-capstone-project/blob/d0b069869f38b7c61d860a0994e5c9e7f96b7af2/jupyter-labs-webscraping.ipynb>



# Data Wrangling

- How data was processed is described in the right flowchart.
- <https://github.com/tosshee/coursera-ds-capstone-project/blob/1e4e8f6463d17f792a5083a839170ad2ca56d8d3/labs-jupyter-spacex-Data%20wrangling.ipynb>



# EDA with SQL

---

- SQL queries you performed on the following:
  - 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.
  - 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. Use a subquery
  - 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.
- [https://github.com/tosshee/coursera-ds-capstone-project/blob/f19437c4b0f9561bffba92cd9b4e092f610a5512/jupyter-labs-eda-sql-coursera\\_sqlite.ipynb](https://github.com/tosshee/coursera-ds-capstone-project/blob/f19437c4b0f9561bffba92cd9b4e092f610a5512/jupyter-labs-eda-sql-coursera_sqlite.ipynb)

# EDA with Data Visualization

---

- Finding out correlation with key factors of success, scatter plots are used:
  - Flight number (attempts) vs. Launch site
  - Payload vs. Launch site
  - Flight number (attempts) vs. Orbit type
  - Payload vs. Orbit type
- Finding out whether Orbit Type is the key factor of success, comparison between Success Ratio and Orbit Type is described by bar chart.
- Looking at the trend of success over years, line chart is used.
- <https://github.com/tosshee/coursera-ds-capstone-project/blob/5ed4a7858791ba91083745c3997f43ad2d17a6e1/jupyter-labs-eda-dataviz.ipynb>

# Build an Interactive Map with Folium

---

- Added each site's location with a circle on a map; added markers to show the success or failed launches for each site; added lines to show the distance from the launch site to its nearest seashore and to the Cape Canaveral Space Force Station.
- [https://github.com/toshee/coursera-ds-capstone-project/blob/ba83892380bf39535b713a10a4ed9a1bcace44d8/lab\\_jupyter\\_launch\\_site\\_location.ipynb](https://github.com/toshee/coursera-ds-capstone-project/blob/ba83892380bf39535b713a10a4ed9a1bcace44d8/lab_jupyter_launch_site_location.ipynb)
- <https://github.com/toshee/coursera-ds-capstone-project.git> [See map\_#.png]

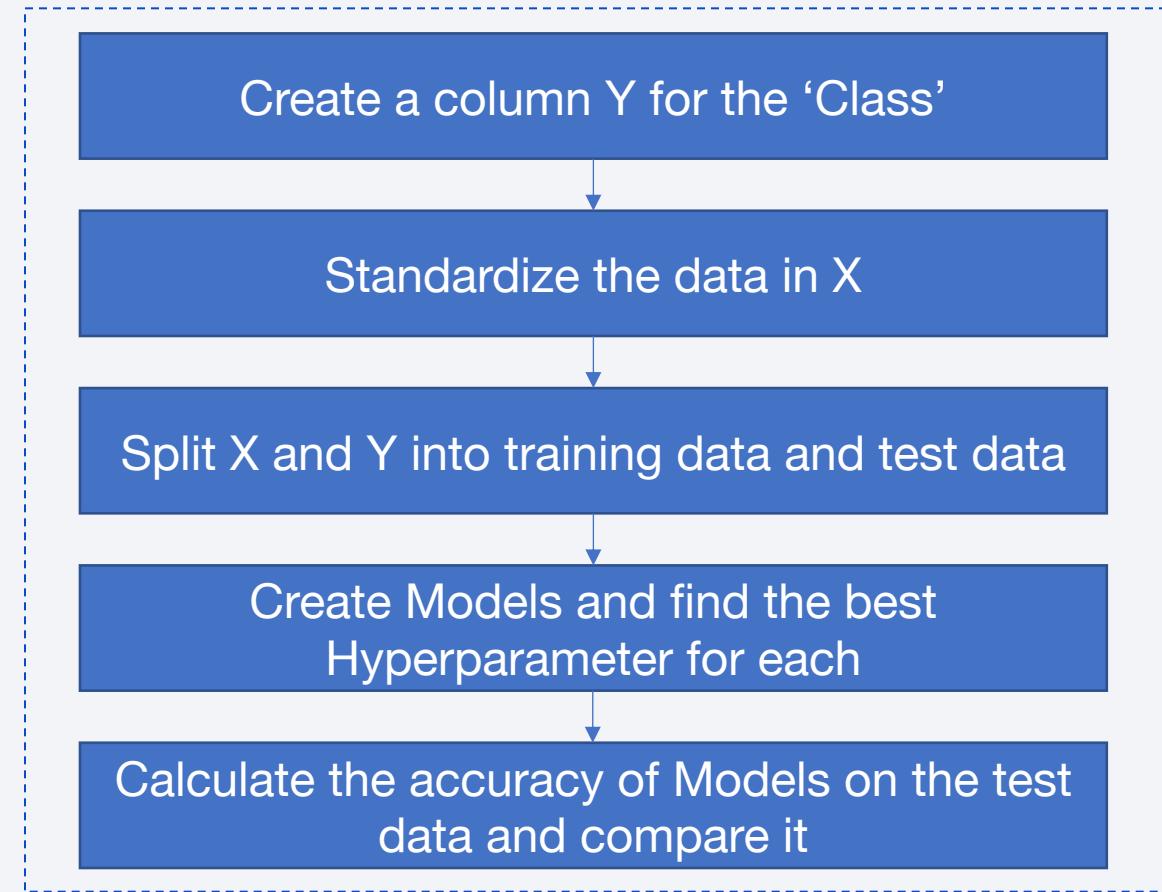
# Build a Dashboard with Plotly Dash

---

- The Pie chart tells us the successful launch by site.
- The Plots tells us the correlation between payload and success, by site and by payload interactively.
- <https://github.com/tosshee/coursera-ds-capstone-project/blob/343de6cbeabd9e7cd190848c1bc3c2ee4995cc5a/spacex-dash-app.py>
- Refer to Appendix - Dash

# Predictive Analysis (Classification)

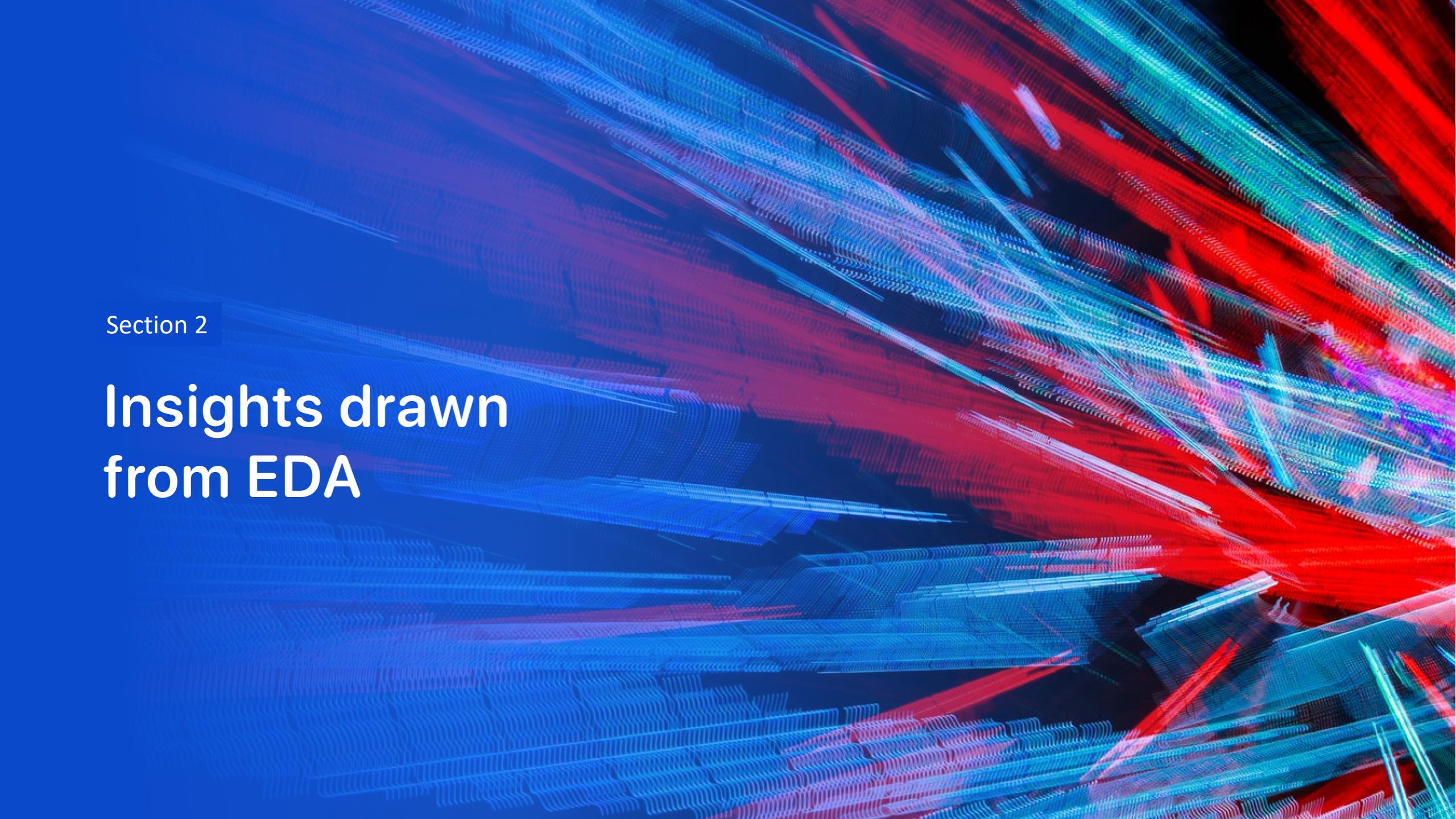
- How to built, evaluated, improved, and found the best performing classification model is shown in the flowchart on the right.
- [https://github.com/tosshee/coursera-ds-capstone-project/blob/592e10a28d7780e84b36bae801aa3513f0b5c043/SpaceX\\_Machine%20Learning%20Prediction\\_Part\\_5.ipynb](https://github.com/tosshee/coursera-ds-capstone-project/blob/592e10a28d7780e84b36bae801aa3513f0b5c043/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb)



# Results

---

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

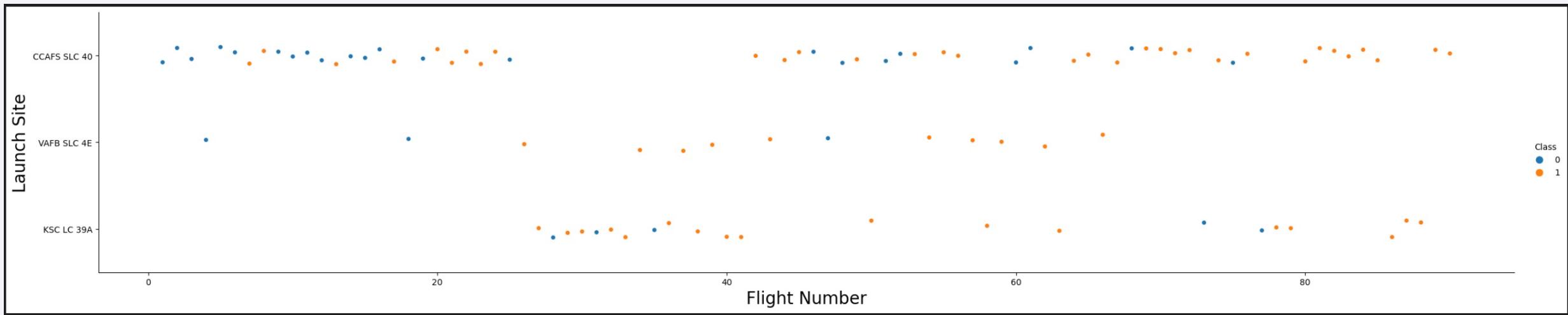
The background of the slide features a complex, abstract digital visualization. It consists of numerous thin, glowing lines that create a sense of depth and motion. The lines are primarily blue and red, with some green and purple highlights. They form a grid-like structure that curves and twists across the frame, resembling a 3D wireframe or a network of data points. The overall effect is futuristic and dynamic, suggesting concepts like data flow, digital communication, or complex systems.

Section 2

## Insights drawn from EDA

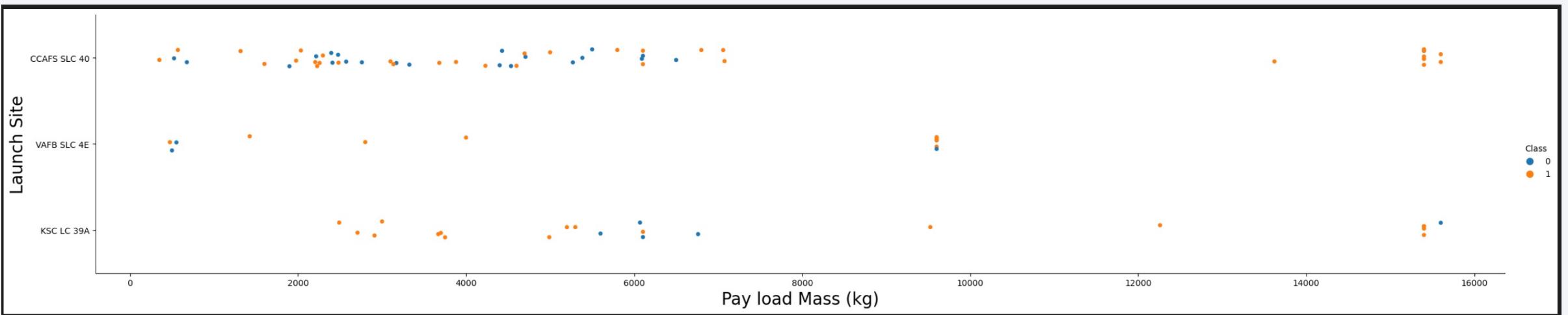
# Flight Number vs. Launch Site

- CCAFS-SLC launch site has the greatest number of launch; it looks improving success ratio after the flight number 70.
- VAFB-SLC launch site and KSC-LC have high success launch ratio.
- Overall after 20 flights, success ratio seems improved.



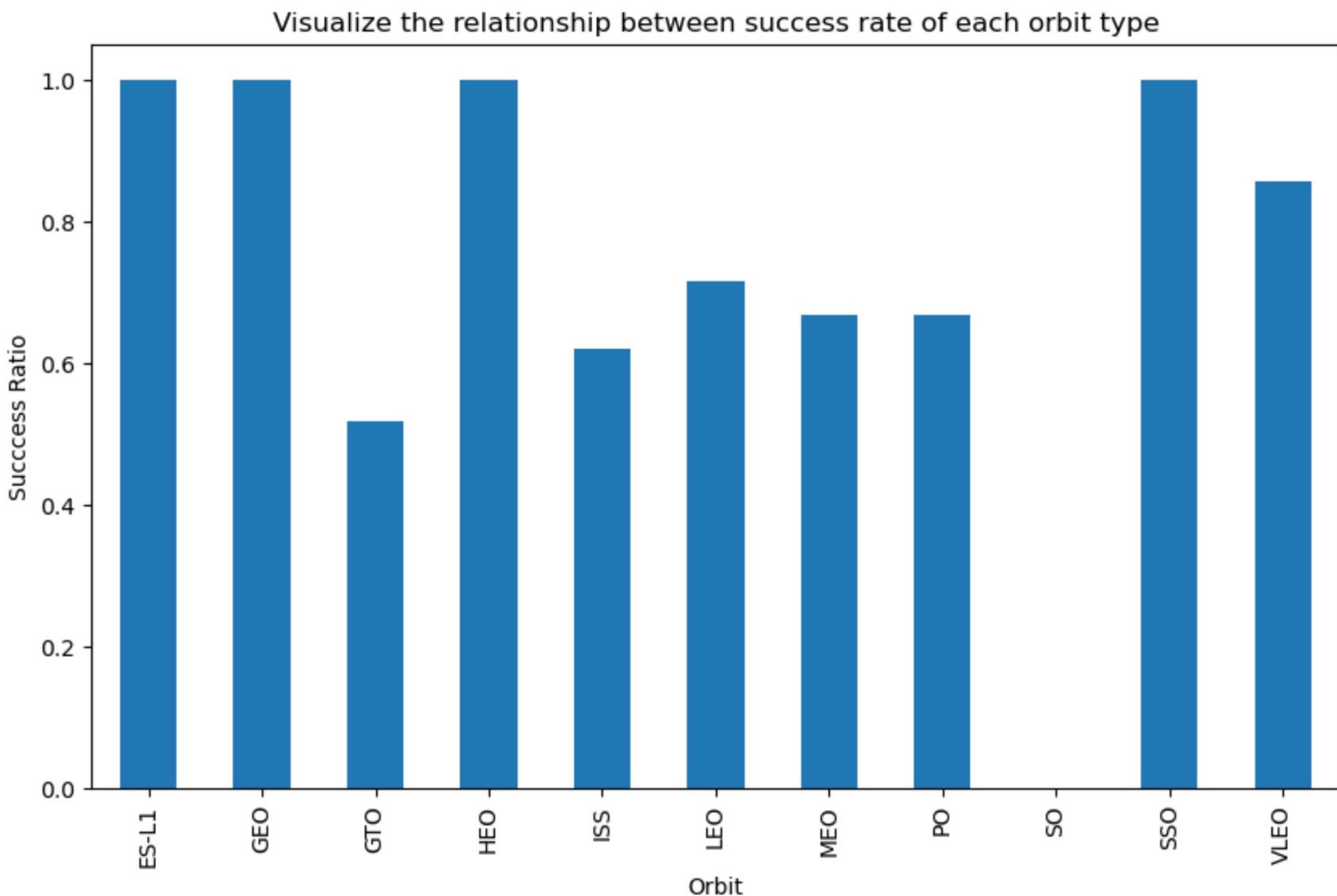
# Payload vs. Launch Site

- VAFB-SLC launch site, there are no rockets launched for heavy payload mass (greater than 10000).
- CCAFS-SLC and KSC-LC launch sites are good at almost all launch with heavy payload mass, greater than 6000 and 8000 respectively.



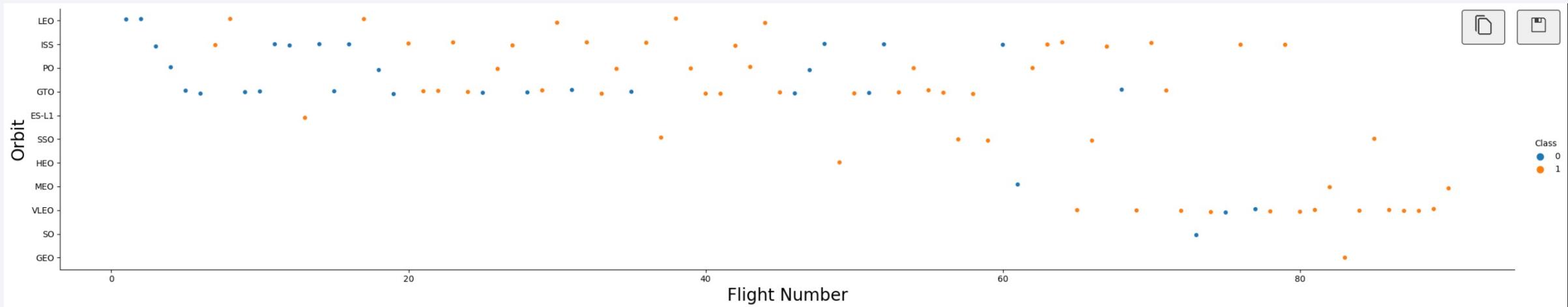
# Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO, and VLEO show highly success rate.



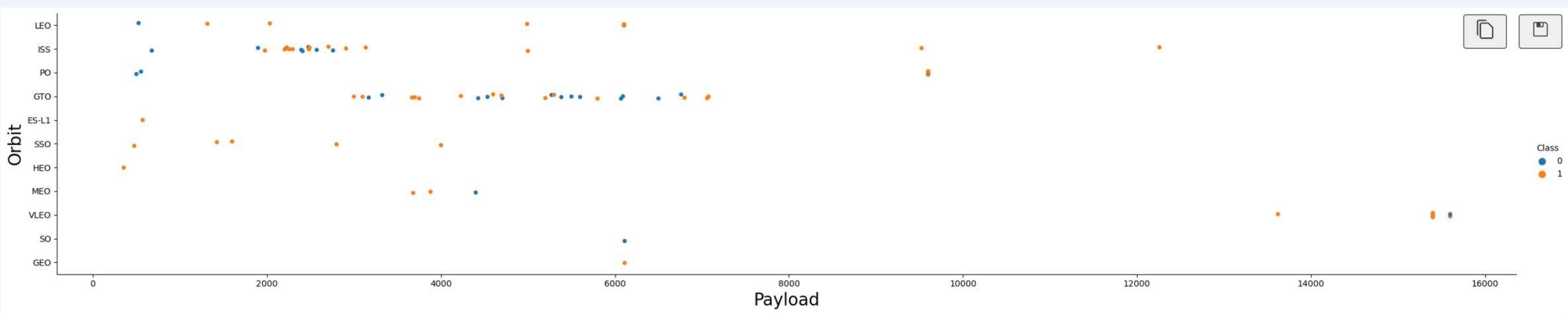
# Flight Number vs. Orbit Type

- Looking at the LEO orbit, the success appears related to the number of flights; on the other hand, there seems to be no correlation with flight attempts when in GTO orbit.
- Looking at the VLEO orbit, the success ratio seems constantly high.



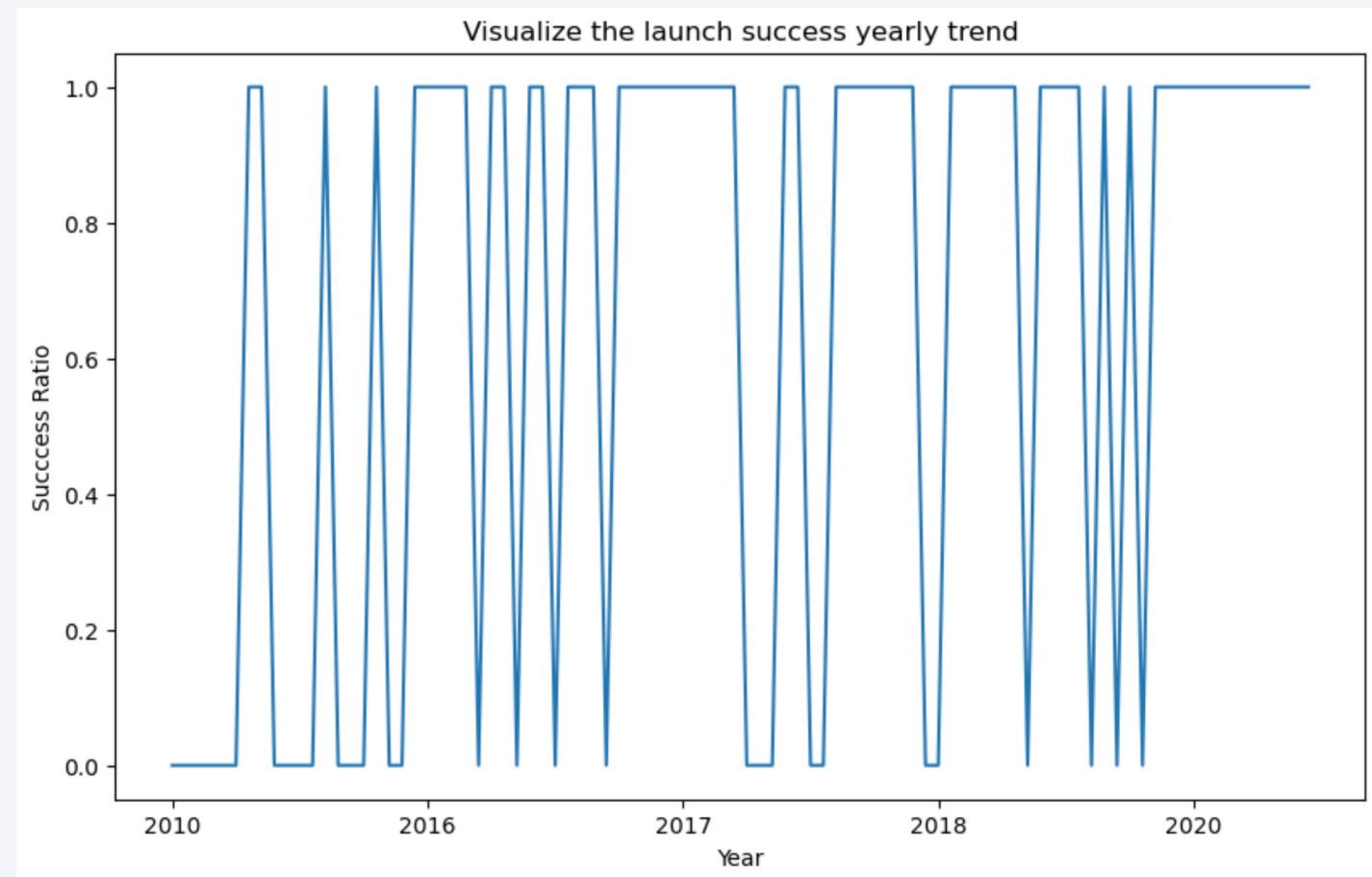
# Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here.



# Launch Success Yearly Trend

- Since 2016, the launch success ratio has been increasing.



# All Launch Site Names

- The unique launch sites are CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.

## Task 1

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

```
[11]: %sql select distinct "Launch_Site" from spacextbl;
```

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

Done.

```
[11]: Launch_Site
```

```
-----  
CCAFS LC-40
```

```
VAFB SLC-4E
```

```
KSC LC-39A
```

```
CCAFS SLC-40
```

# Launch Site Names Begin with 'CCA'

- Select five records where launch sites begin with the string “CCA”

▼ Task 2 ↶ ↷ ↴ ↵ ↶ ↸

Display 5 records where launch sites begin with the string 'CCA'

```
[8]: %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     |
|------------|------------|-----------------|-------------|---|------------------|-----------|-----------------|-----------------|---------------------|
| 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

- The total payload mass carried by boosters launched by NASA (CRS) is 45596 KG.

## Task 3

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

```
[9]: %sql select customer, sum("PAYLOAD_MASS_KG_") from spacextbl where Customer = "NASA (CRS);  
* sqlite:///my_data1.db  
Done.  
[9]:   Customer  sum("PAYLOAD_MASS_KG_")  
-----  
NASA (CRS)          45596
```

# Average Payload Mass by F9 v1.1

---

- The average payload mass carried by booster version F9 v1.1 is 2928.4 KG.

## Task 4

Display average payload mass carried by booster version F9 v1.1

```
[10]: %sql select "Booster_Version", avg("PAYLOAD_MASS_KG_") from spacextbl where "Booster_Version" = "F9 v1.1";  
* sqlite:///my_data1.db  
Done.  
[10]: 

| Booster_Version | avg("PAYLOAD_MASS_KG_") |
|-----------------|-------------------------|
| F9 v1.1         | 2928.4                  |


```

# First Successful Ground Landing Date

---

- Date of the first successful landing outcome on ground pad is 22 DEC 2015.

```
[13]: %sql select Date, "Landing _Outcome" from spacextbl where "Landing _Outcome" = "Success (ground pad)" \
order by substr(Date,7,4)||substr(Date, 4, 2)||substr(Date, 1, 2) limit 1;
```

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

```
Done.
```

```
[13]:
```

| Date       | Landing _Outcome     |
|------------|----------------------|
| 22-12-2015 | Success (ground pad) |

## Successful Drone Ship Landing with Payload between 4000 and 6000

- The query shows the boosters which have success in drone ship and have payload mass between 4000 and 6000 Kg.

### Task 6

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

```
[15]: %sql select "Booster_Version", "PAYLOAD_MASS_KG_", "Landing _Outcome" from spacextbl\\
where "Landing _Outcome" = "Success (drone ship)" and ("PAYLOAD_MASS_KG_" between 4000 and 6000);
```

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

Done.

| Booster_Version | PAYLOAD_MASS_KG_ | Landing _Outcome     |
|-----------------|------------------|----------------------|
| F9 FT B1022     | 4696             | Success (drone ship) |
| F9 FT B1026     | 4600             | Success (drone ship) |
| F9 FT B1021.2   | 5300             | Success (drone ship) |
| F9 FT B1031.2   | 5200             | Success (drone ship) |

# Total Number of Successful and Failure Mission Outcomes

- The overall total number of success is 100, whereas the failure is 1.

## Task 7

List the total number of successful and failure mission outcomes

```
[16]: %sql select Customer, "Mission_Outcome", count("Mission_Outcome") as TOTAL \
from spacextbl group by "Mission_Outcome";  
* sqlite:///my_data1.db  
Done.
```

| Customer         | Mission_Outcome                  | TOTAL |
|------------------|----------------------------------|-------|
| NASA (CRS)       | Failure (in flight)              | 1     |
| SpaceX           | Success                          | 98    |
| USAF             | Success                          | 1     |
| Northrop Grumman | Success (payload status unclear) | 1     |

# Boosters Carried Maximum Payload

---

- The booster 'F9 B5 B1048.4' which have carried the maximum payload mass is 15,600 Kg.

## ▼ Task 8

List the names of the booster\_versions which have carried the maximum payload mass. Use a subquery

```
[18]: %sql select "Booster_Version" , (select MAX("PAYLOAD_MASS__KG_")) as MAX_PAYLOAD from spacextbl;  
* sqlite:///my_data1.db
```

Done.

| Booster_Version | MAX_PAYLOAD |
|-----------------|-------------|
| F9 B5 B1048.4   | 15600       |

# 2015 Launch Records

- List the failed landing\_outcomes in drone ship, their booster versions, and launch site names for in year 2015

## Task 9

List the records which will display the month names, failure landing\_outcomes in drone ship ,booster versions, launch\_s

**Note: SQLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get t**

```
[17]: %sql select substr(Date, 4, 2) as month, "Booster_Version", "Landing _Outcome" from spacextbl\
where substr(Date,7,4)='2015' and "Landing _Outcome" = "Failure (drone ship)";
```

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

Done.

| month | Booster_Version | Landing _Outcome     |
|-------|-----------------|----------------------|
| 01    | F9 v1.1 B1012   | Failure (drone ship) |
| 04    | F9 v1.1 B1015   | Failure (drone ship) |

# 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

```
[32]: %%sql
select "Landing _Outcome",count("Landing _Outcome") as occurence from spacextbl
where ((substr(Date,7,4)||substr(Date, 4, 2)||substr(Date, 1, 2))
      between '20100604' and '20170320') group by "Landing _Outcome"
      order by occurence desc
```

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

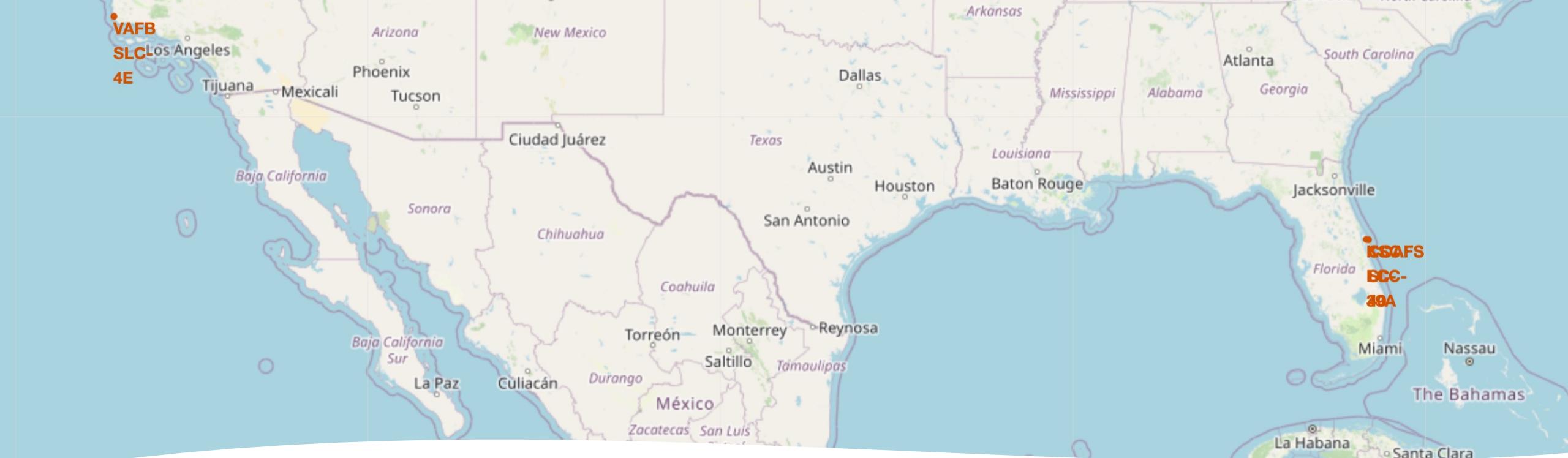
```
Done.
```

| Landing _Outcome       | occurence |
|------------------------|-----------|
| No attempt             | 10        |
| Success (drone ship)   | 5         |
| Failure (drone ship)   | 5         |
| Success (ground pad)   | 3         |
| Controlled (ocean)     | 3         |
| Uncontrolled (ocean)   | 2         |
| Failure (parachute)    | 2         |
| Precluded (drone ship) | 1         |

The background of the slide is a photograph taken from space at night. It shows the curvature of the Earth against a dark blue and black void of space. City lights are visible as small white dots and larger clusters of light, primarily concentrated in the lower right quadrant where the United States appears. In the upper right, there are bright green and yellow bands of the Aurora Borealis (Northern Lights) dancing across the sky.

Section 3

# Launch Sites Proximities Analysis

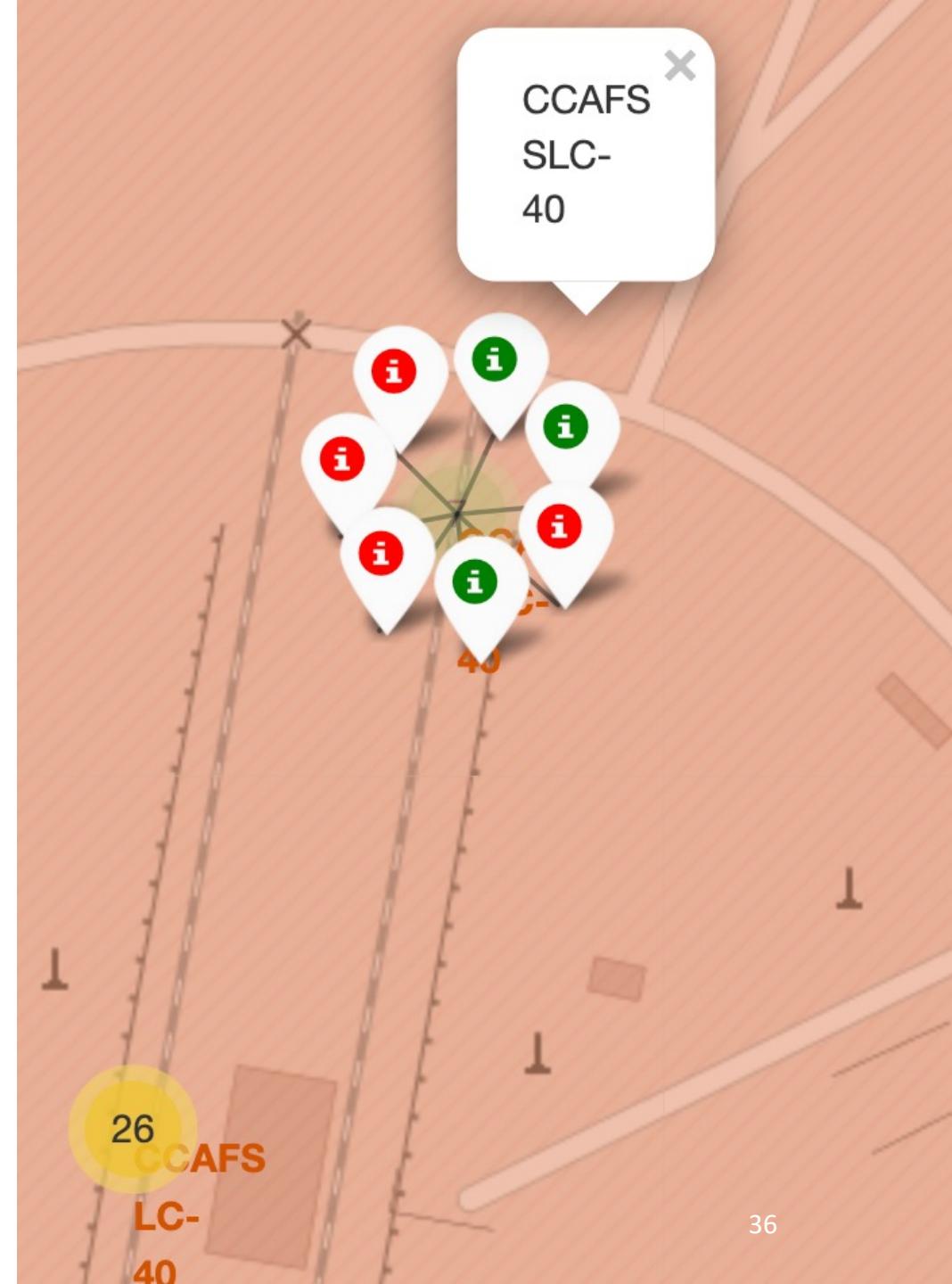


Mark all launch  
sites on a map

- All launch sites are near the coast.
- CCAFS launch sites are near the Equator line.

# Mark the success/failed launch at CCAFS SLC-40

- Here two close places launched SpaceX, 26 and 7 times.
- Green color marks show successful launch, whereas red ones show failure one.

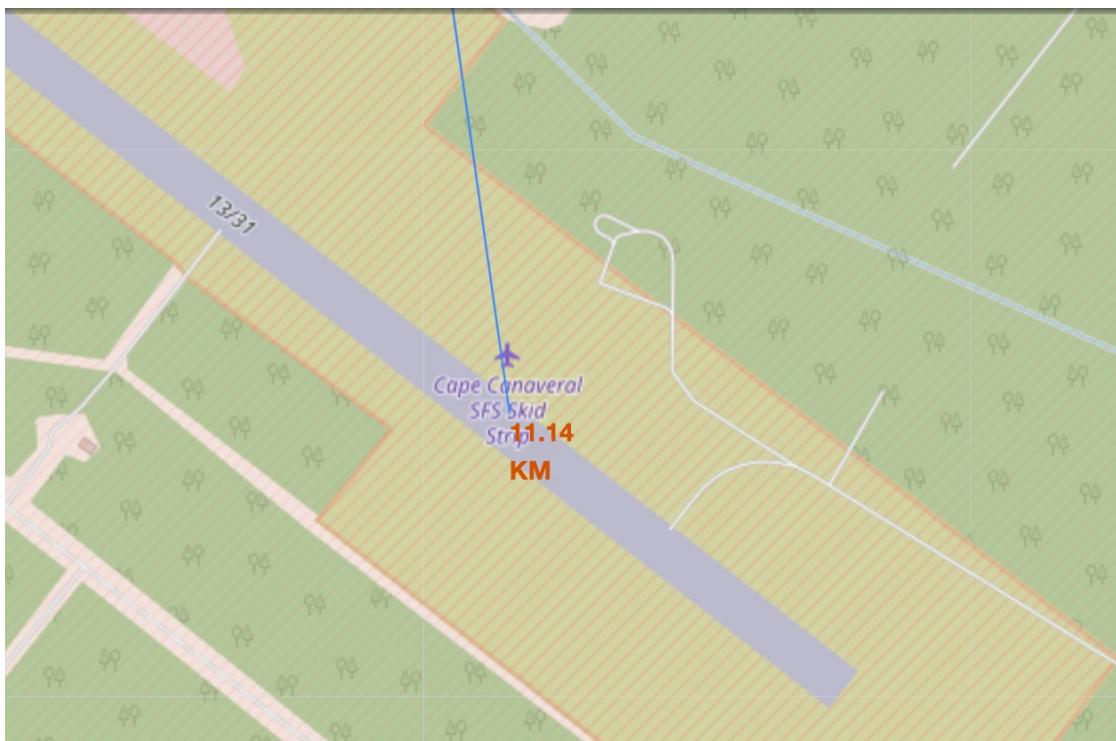


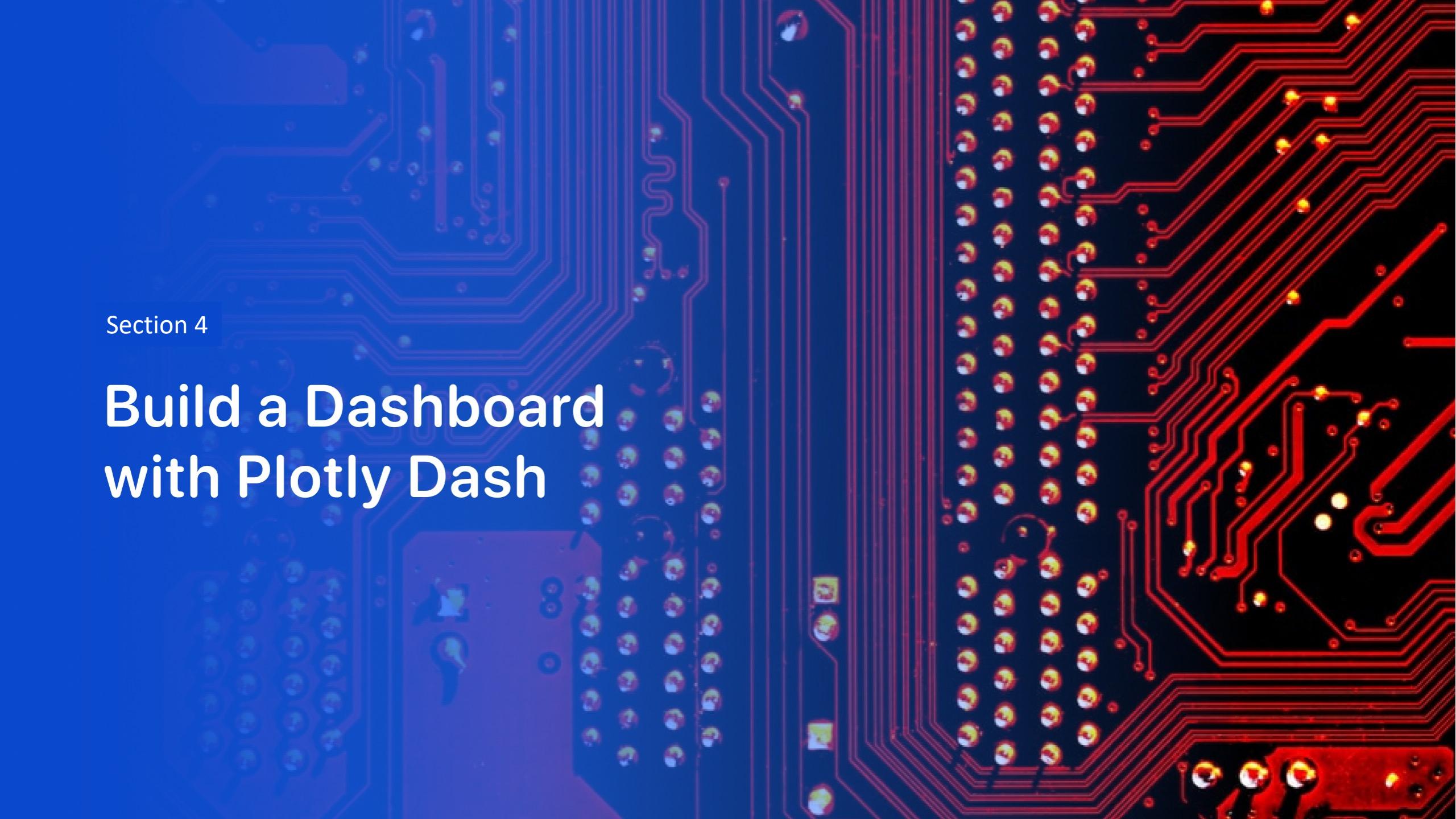


# Distance from CCAFS SLC-40 to CCSFS

---

The distance to the landmark is about 11 KM.





Section 4

# Build a Dashboard with Plotly Dash

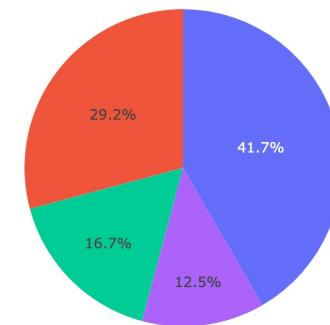
# Success launch for all sites

- 71% of success launch was made by KSC LC-39A and CCAFS LC-40.

## SpaceX Launch Records Dashboard

All Sites

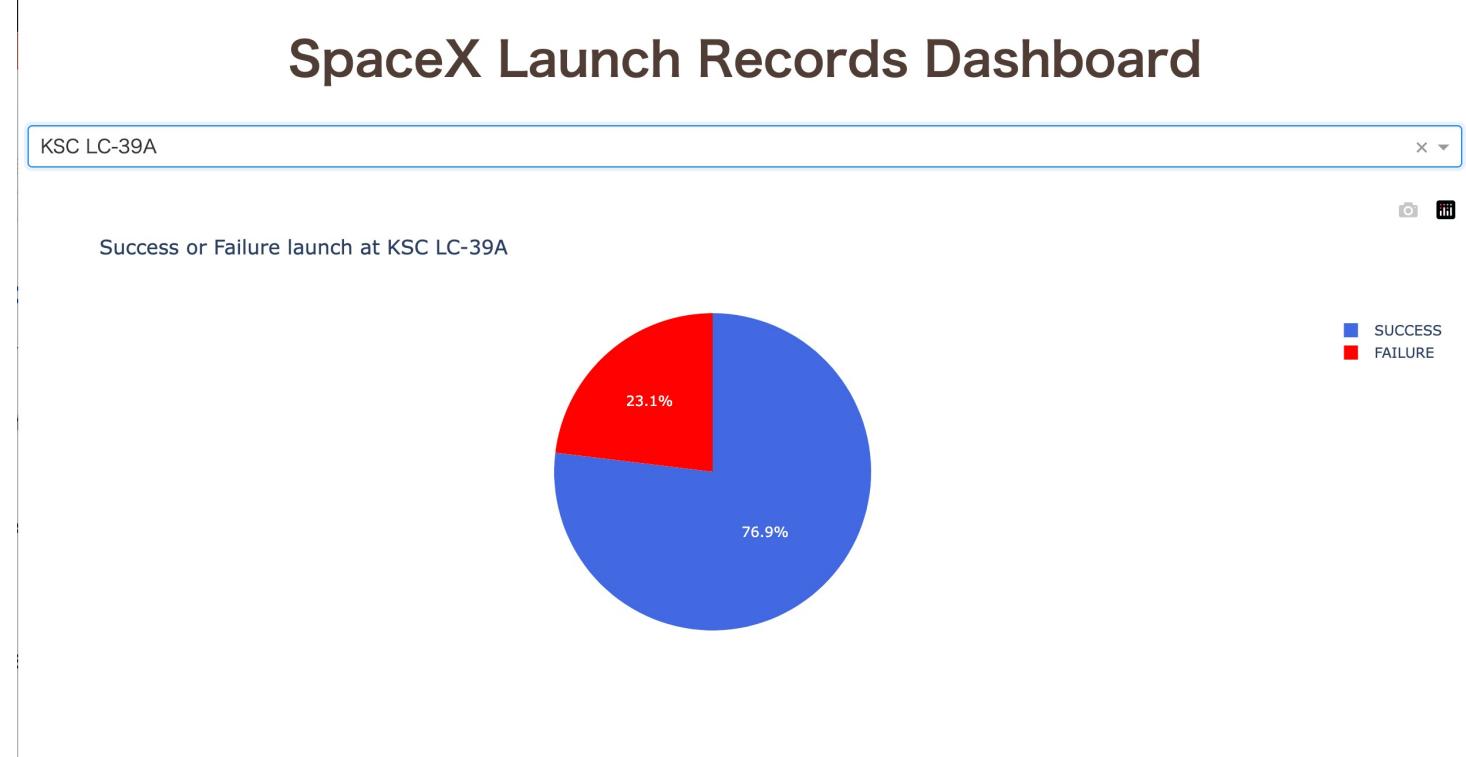
Success Launch By Site



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

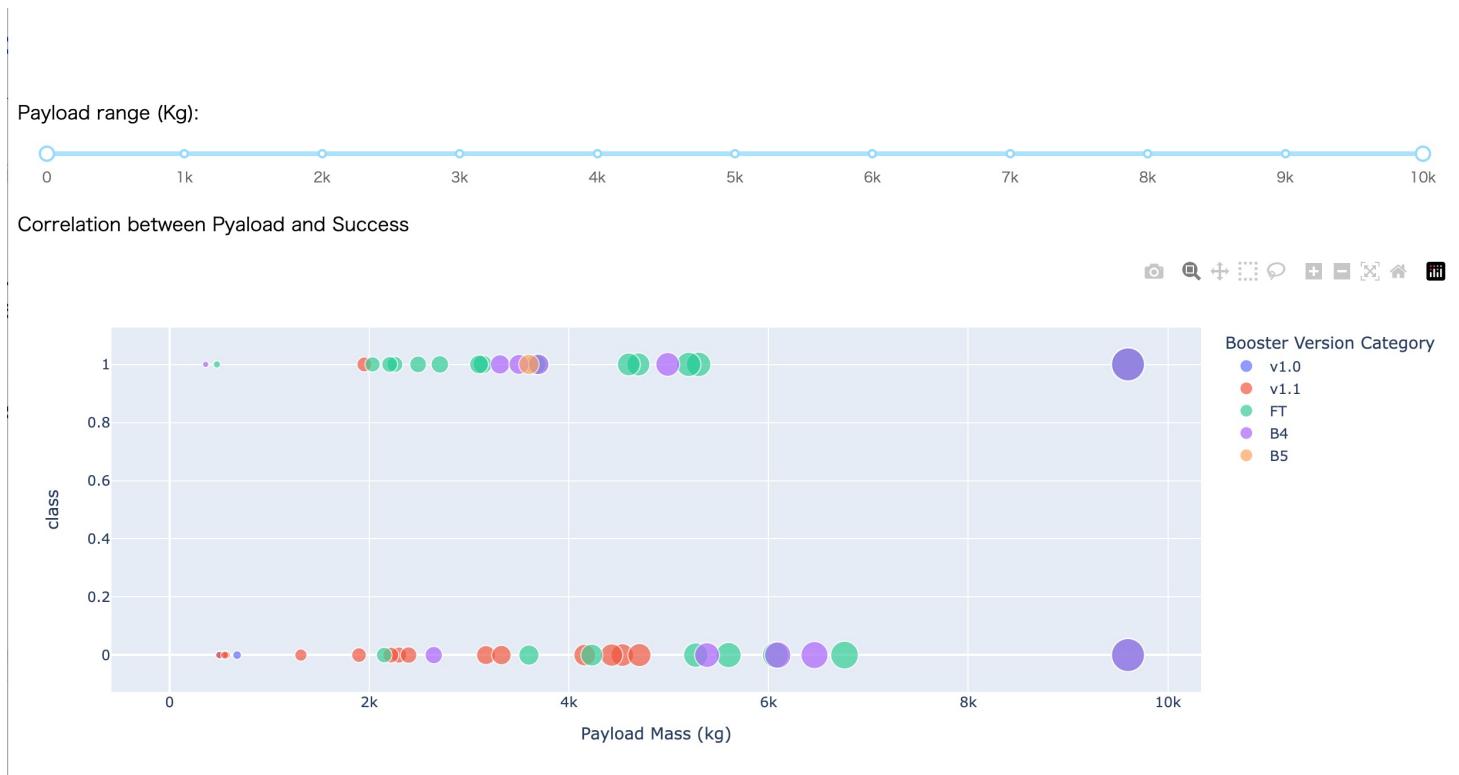
# Highest launch success ratio site

- KSC LC-39A is the highest success ratio site.
- 76.9% was successful launch.



# Payload vs. Launch Outcome

- Booster version v1.0 and v1.1 were mostly failure even with lower payload less than 5k kg.
- Booster version B4 was 80% of success launch with lower payload less than 5k kg.
- Success ratio of heavy lifting more than 5k kg payload was less than 30%, ie. 3/11.



Section 5

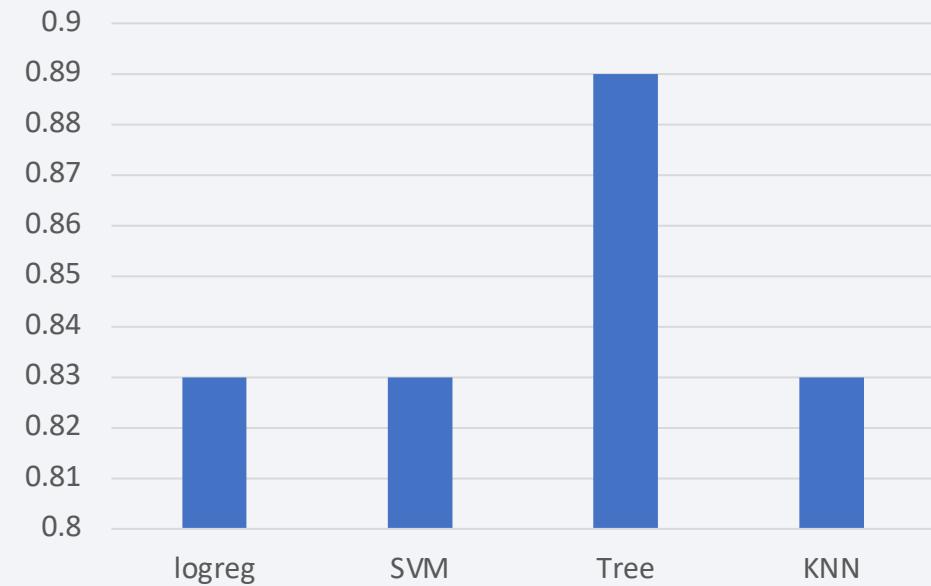
# Predictive Analysis (Classification)

# Classification Accuracy

---

- Visualize the built model accuracy for all built classification models, in a bar chart
- Tree model has the highest classification accuracy

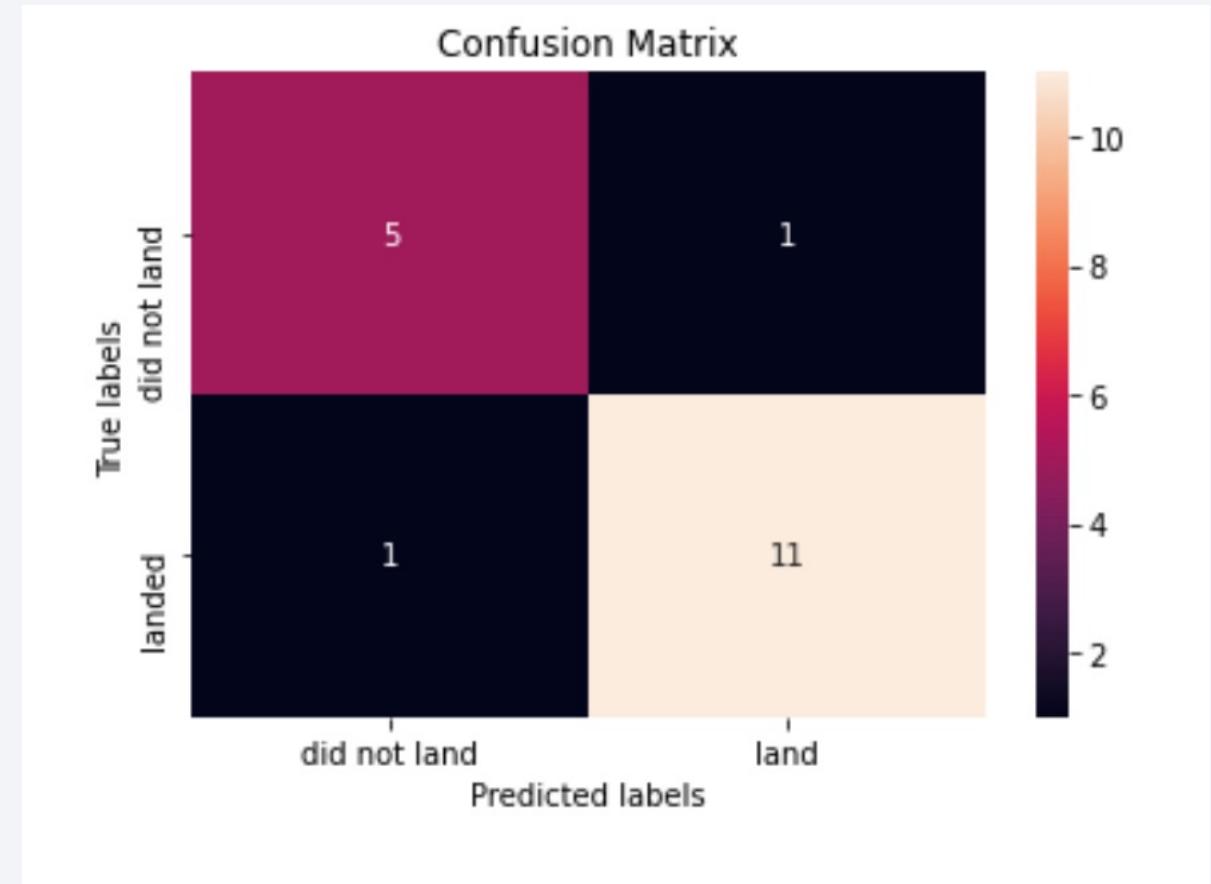
Accuracy on Test data  
(Real world data)



# Confusion Matrix

---

- In terms of accuracy on the test data (real world data), Tree model has the least fault positive and its accuracy score is 0.89.

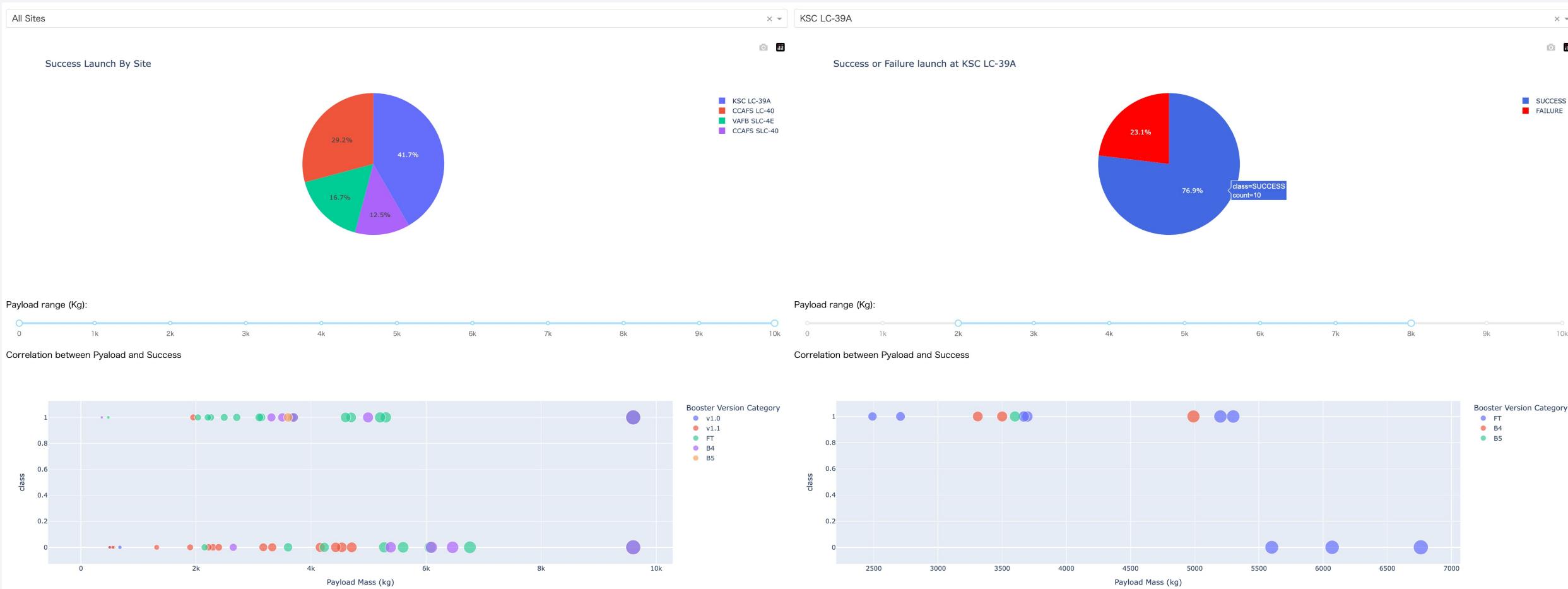


# Conclusions

---

- KSC LC-39A is the highest success ratio site.
- Overall after 20 flights, success ratio seems improved.
- Orbit type ES-L1, GEO, HEO, and SSO, and VLEO shows highly success rate.
- Success ratio of heavy lifting more than 5k kg payload was less than 30%, ie. 3/11.
- Machine Learning using Tree model shows the highest accuracy score of 0.89; we can predict if a future-launch success with features such as Payload, Booster Version, Launch Site, Orbit Type, etc.

# Appendix - Dash



Thank you!

