



IBM Developer  
SKILLS NETWORK

# Winning Space Race with Data Science

Adán Faramiñán  
05/06/2022



# Outline

---

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

# Executive Summary

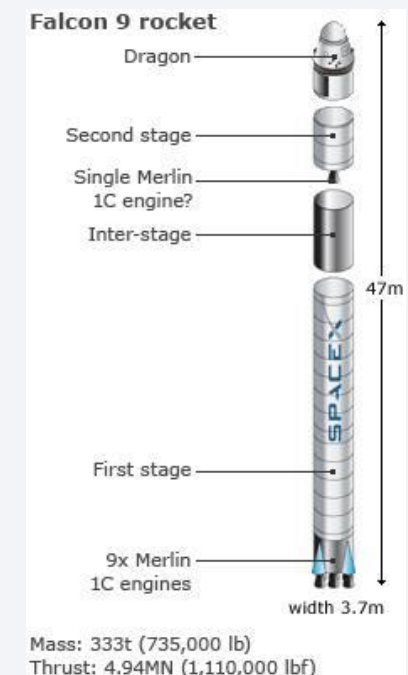
---

- Using public data from SpaceX, an algorithm was developed to predict the probability of reusing stage 1 of a rocket. The results showed that landing success could be predicted with an accuracy of 0.833. The algorithms with the best performance were Support Vector Machine and k Nearest Neighbor.
- This analysis makes it possible to develop a plan to avoid the destruction of part of the fuselage and, consequently, avoid significant economic losses.

# Introduction

---

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. SpaceX's Falcon 9 launch like regular rockets.
- Unlike other rocket providers, SpaceX's Falcon 9 Can recover the first stage. Sometimes the first stage does not land. Sometimes it will crash. Other times, SpaceX will sacrifice the first stage due to the mission parameters like payload, orbit, and customer.
- In this project, information about SpaceX is collected and analyzed. The study's main objective was to determine the probability of refusing stage 1 through artificial intelligence algorithms and public data.





Section 1

# Methodology

# Methodology

---

## Executive Summary

- Data collection methodology:
  - SpaceX Rest API and Web Scrapping (Wikipedia=
- Perform data wrangling
  - The process of data wrangling may include one hot encoding, data visualization and data aggregation.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Linear regression (LR), support vector machine (SVM), k-near nighbour (KNN) and tree decision algorithms were evaluated. Before the data was split to train and test.

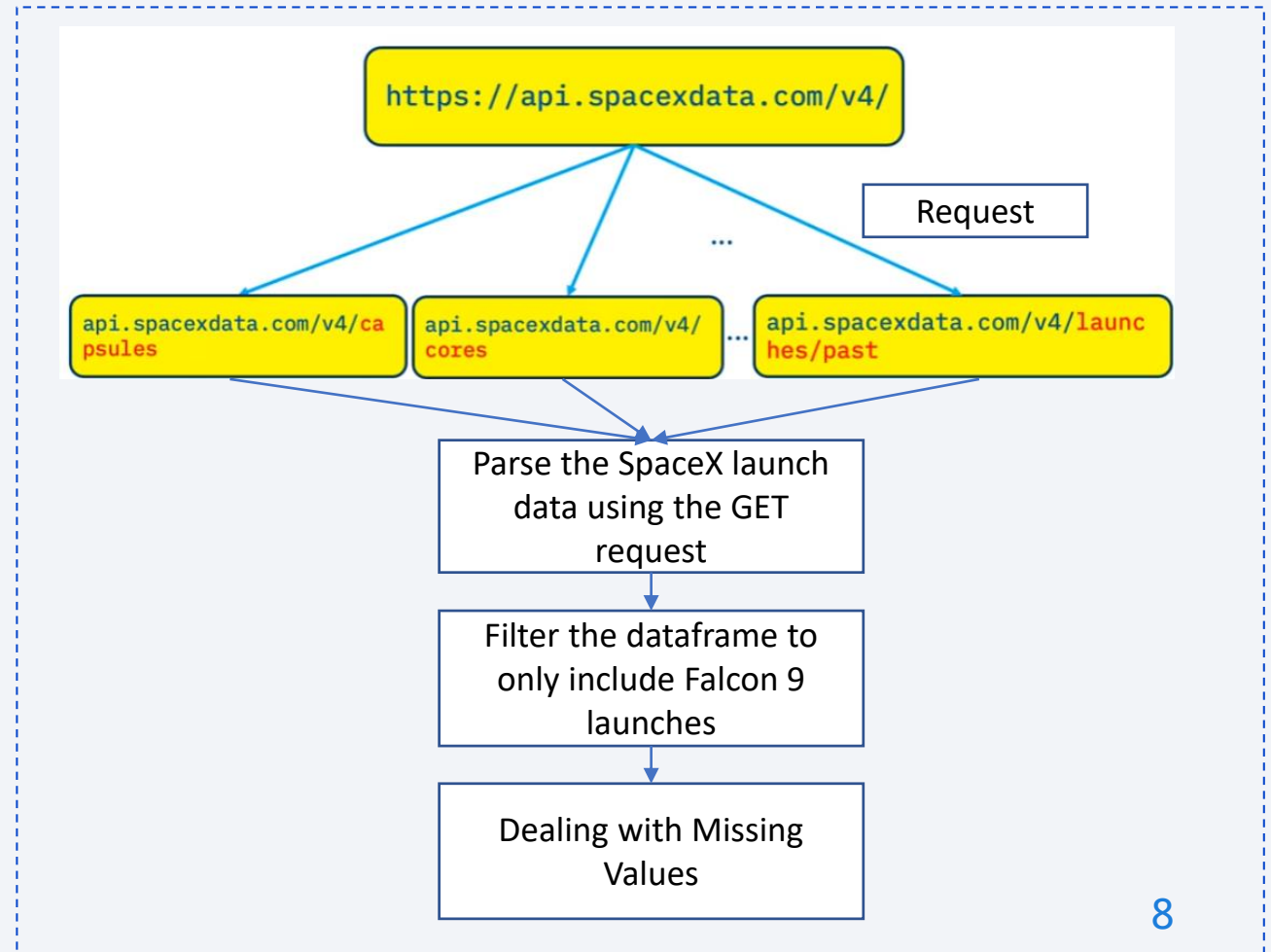
# Data Collection

---

- Two methods were chosen to collect SpaceX public data: web scraping and API requests.
- Web scraping is data scraping used for extracting data from websites. The web scraping software may directly access the World Wide Web using the Hypertext Transfer Protocol or a web browser. It is a form of copying in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis (Wikipedia).
- An application programming interface (API) is a connection between computers or between computer programs. It is a type of software interface, offering a service to other pieces of software. An API request occurs when a developer adds an endpoint to a URL and makes a call to the server (Wikipedia).

# Data Collection - SpaceX API

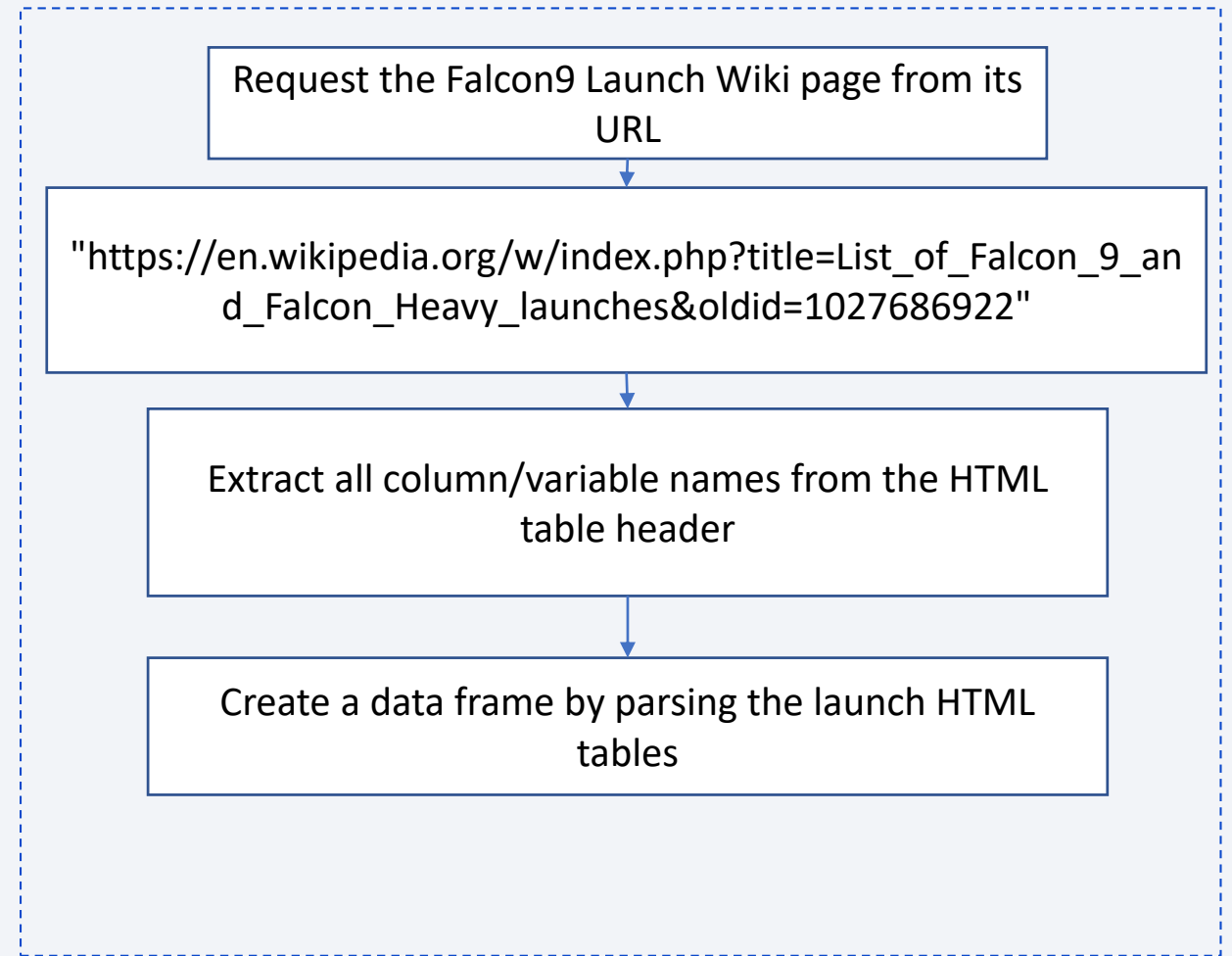
- In this section the data collection with SpaceX REST calls is presented.
- The flowchart is shown in the right figure.
- [GitHub URL](#)





# Data Collection - Scraping

- In this section the data collection with csv file and SQL query.
- The flowchart is shown in the right figure.
- [GitHub URL](#)

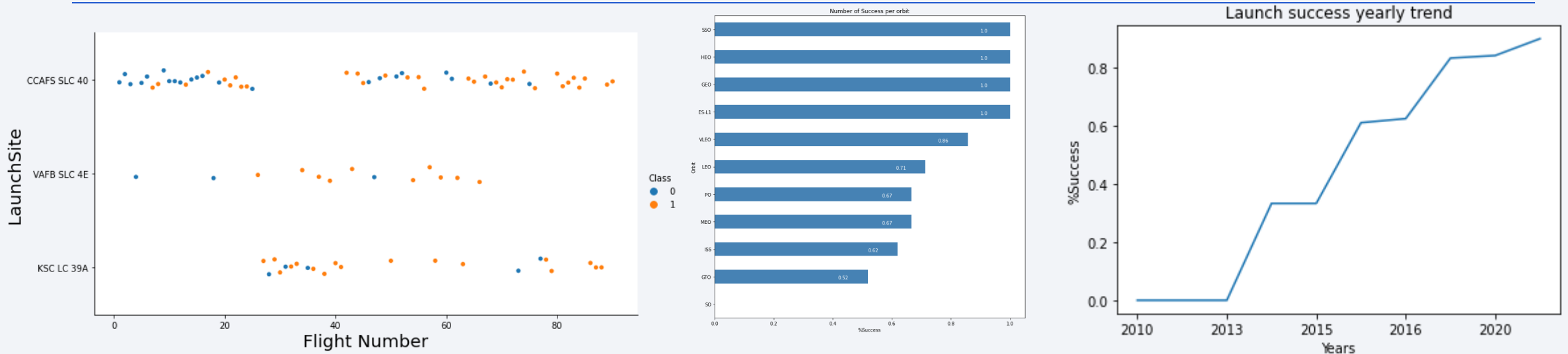


# Data Wrangling

---

- In this step were made following the tasks:
  1. the number of launches on each site was calculated,
  2. the number and occurrence of mission outcome per orbit type was calculated (If the value is zero, the first stage did not land successfully; one means the first stage landed Successfully)
  3. The function `get_dummies` was used to apply OneHotEncoder to the column Orbits, LaunchSite, LandingPad, and Serial.
  4. [GitHub URL step 1 and 2](#)
  5. [GitHub URL step 3 \(Feature Engine\)](#)

# EDA with Data Visualization



A [catplot](#) to see how the FlightNumber (indicating the continuous launch attempts.) and Payload variables would affect the launch outcome (0 fail and 1 success) was done.

Also, two [catplot](#) more Visualize the relationship between Flight Number and Launch Site and the relationship between success rate of each orbit type was done.

I did a [catplot](#) and a [bar chart](#) to visualize the relationship between success rate of each orbit type. The barchart help us to find which orbits have high sucess rate (SSO, GEO, HEO and ES-L1)

[GitHub Link](#)

Finally, I did two [catplot](#) more to visualize the relationship between success rate of each orbit type considering Flight Number and Payload. The last graph was a time series between percentage of success and Years (2015-2019). The 2019 was a year with more success rate.

# EDA with SQL

---

- The data was analyzed to SQL Queries ([GitHub link with results](#)). The tasks were:
  1. Display the names of the unique launch sites in the space mission
  2. Display 5 records where launch sites begin with the string 'CCA'
  3. Display the total payload mass carried by boosters launched by NASA (CRS)
  4. Display average payload mass carried by booster version F9 v1.1
  5. List the date when the first successful landing outcome in ground pad was achieved.
  6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  7. List the total number of successful and failure mission outcomes
  8. List the names of the booster\_versions which have carried the maximum payload mass. Use a subquery
  9. List the failed landing\_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  10. Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

# Build an Interactive Map with Folium

---

- In this step, the following tasks was developed:
  1. Mark all launch sites on a map
  2. Mark the success/failed launches for each site on the map
  3. Calculate the distances between a launch site to its proximities
- [GitHub Link with results](#)



# Build a Dashboard with Plotly Dash

---

- The dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart. The app is composed for
  1. Launch Site Drop-down Input Component
  2. A callback function to render success-pie-chart based on selected site dropdown
  3. A Range Slider to Select Payload
  4. A callback function to render the success-payload-scatter-chart scatter plot
- [GitHub link](#)

# Predictive Analysis (Classification)

---

- In order to develop a model which predicted the success rate, the follows tasks was done:
  1. Create a column for the class
  2. Standardize the data
  3. Split into training data and test data
  4. Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
  5. Find the method performs best using test data
- [GitHub Link](#)

# Results

---

- Through the data analysis, it can be affirmed that the success of maintaining the "First Stage" increased with the growth over the years. The missions destined for the GEO, HEO, ESL1 and SSO orbits had a 100% success rate. On the other hand, launches from KSC LC 39a had the highest success rate.
- Regarding the prediction of success, the Support Vector Machine (SVM), K-Nearest Neighbor (kNN), and Logistic Regression (LR) algorithms were the ones with the best performance considering precision and confusion matrix.



The background of the slide is an abstract composition. It features a dark blue field on the left side, which transitions into a complex pattern of diagonal streaks in shades of blue, red, and cyan on the right. These streaks have a textured, almost woven appearance. Overlaid on this pattern is a faint, light blue grid that recedes into the distance, creating a sense of depth and perspective.

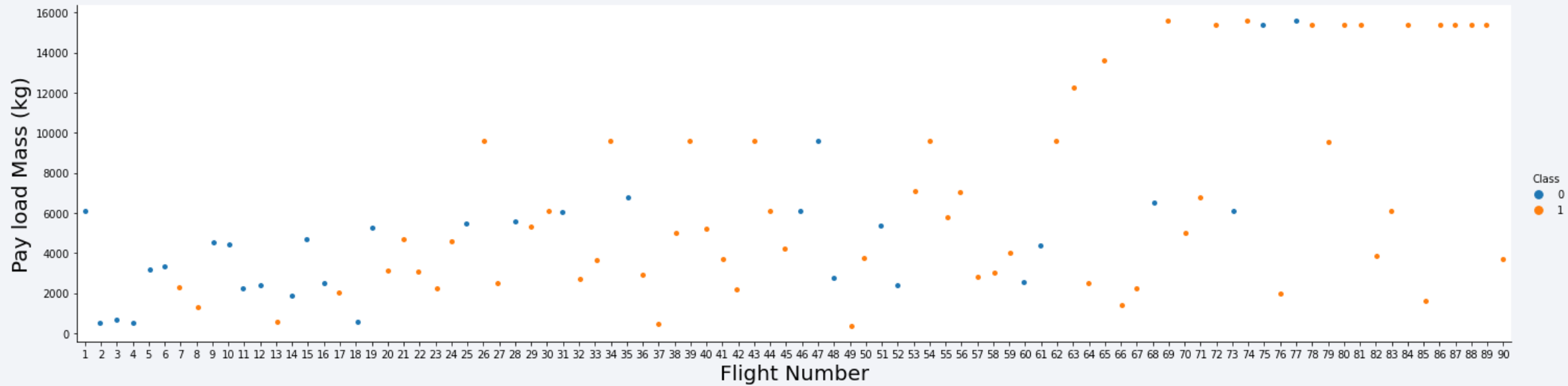
Section 2

# Insights drawn from EDA



# Flight Number vs. Launch Site

---

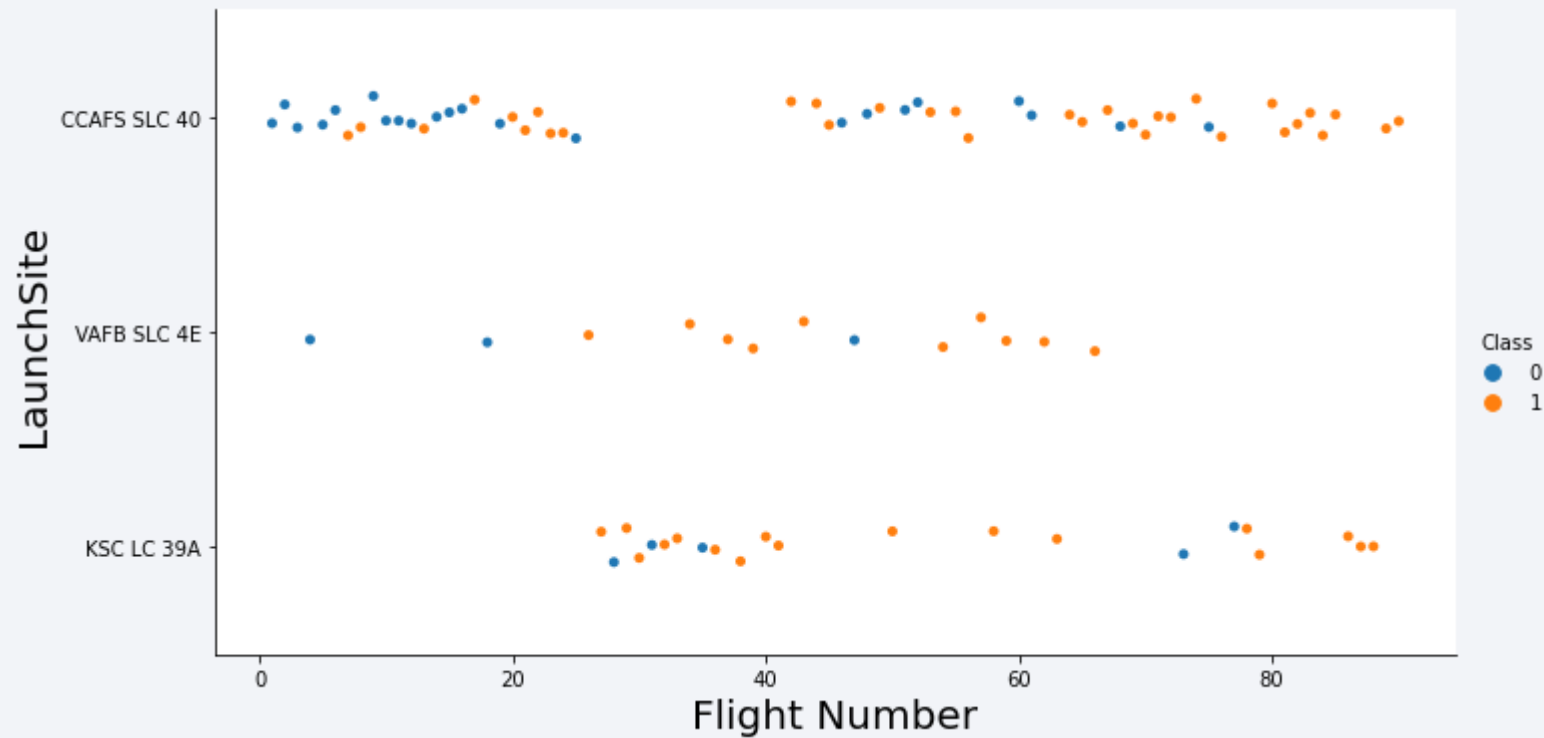


- We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.



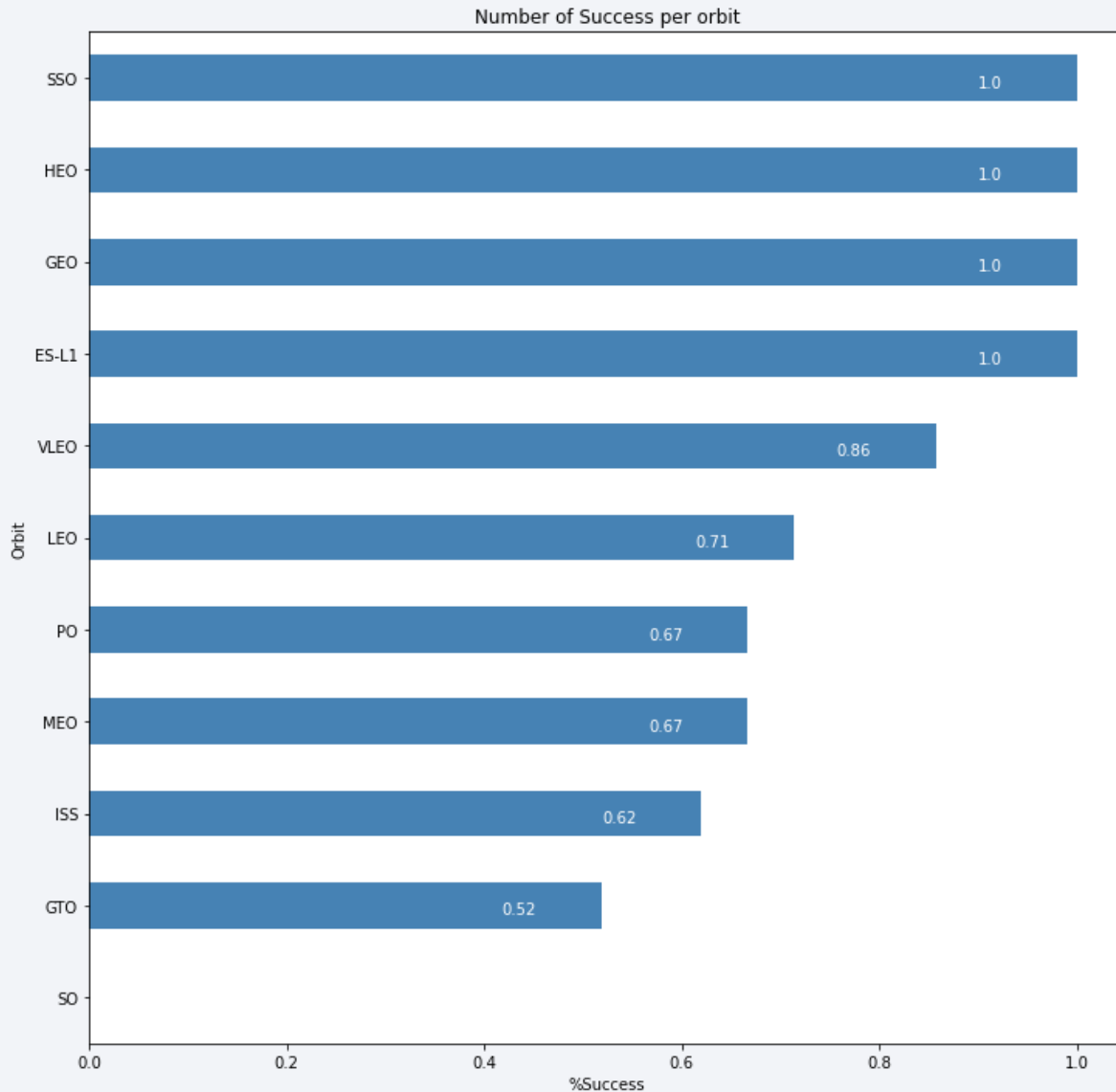
# Payload vs. Launch Site

---

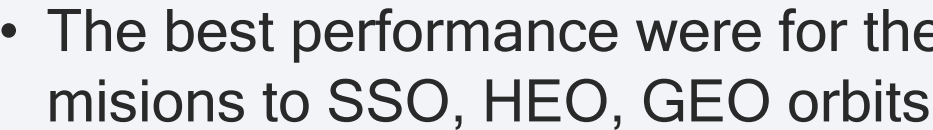


- KSC LC 39 was a Launch Site with the highest success rate

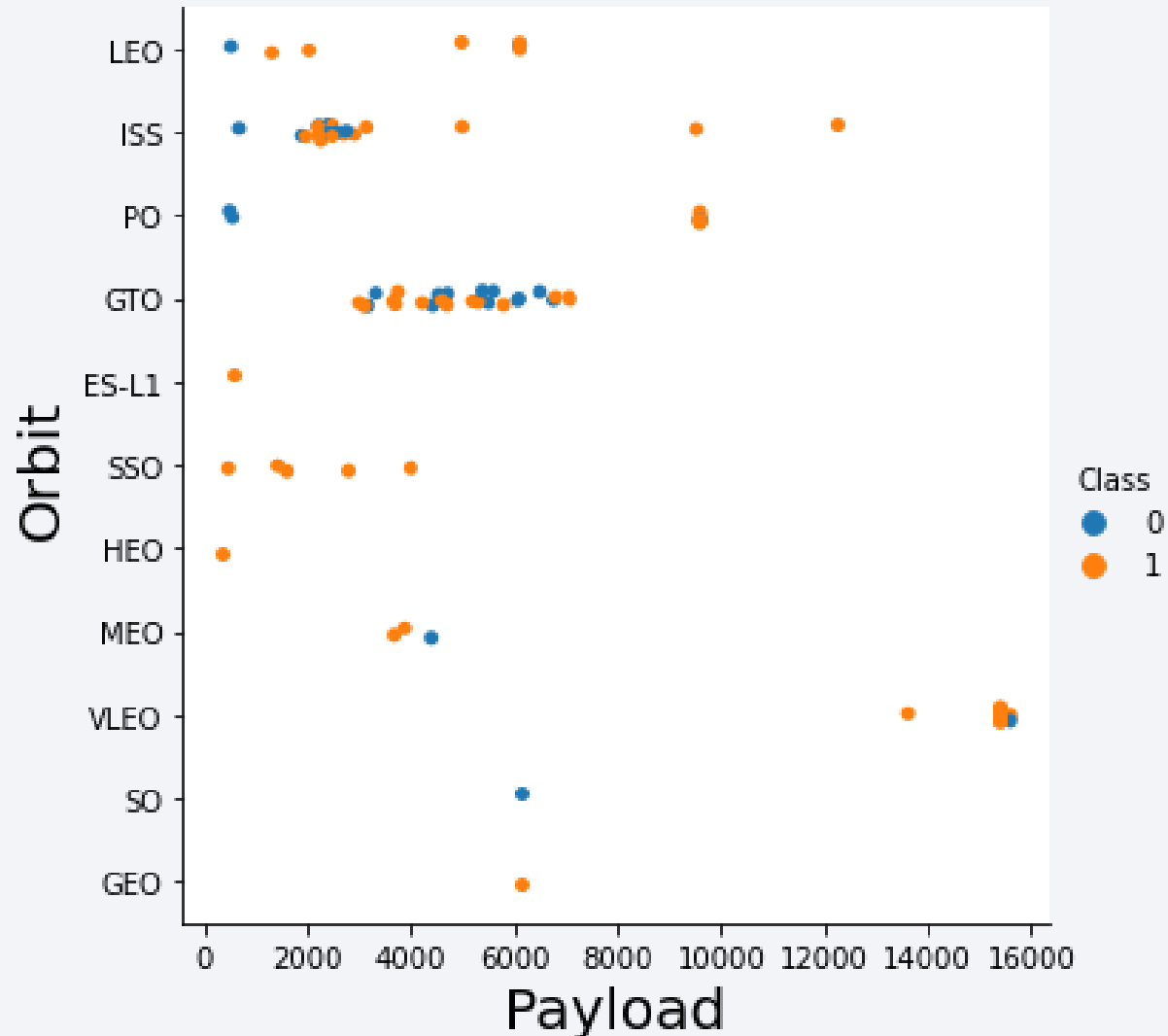
# Success Rate vs. Orbit Type



- The best performance were for the missions to SSO, HEO, GEO orbits



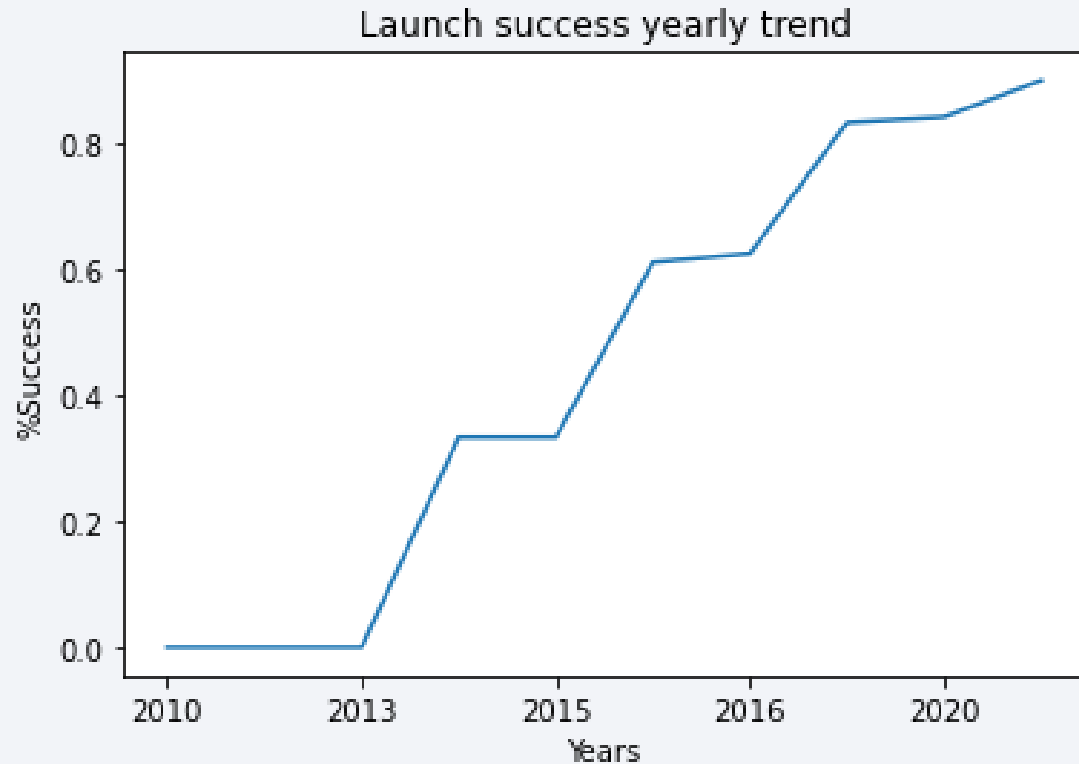
# Payload vs. Orbit Type



- This graph shows that there was only one mission to GEO, ES-L1 and HEO, so statements about these data may be biased. However, it can be said that missions to SSO tend to be successful.

# Launch Success Yearly Trend

---



- The success rate since 2013 kept increasing till 2020



# All Launch Site Names

---

```
%sql SELECT unique(launch_site) \
from SPACEXDATASET
```

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

- There are 4 launch site between 2010 and 2021

# Launch Site Names Begin with 'CCA'

---

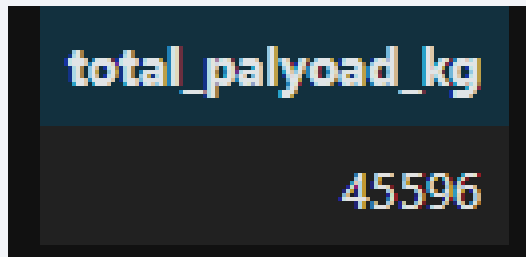
```
%sql SELECT *\nFROM SPACEXDATASET\nWHERE launch_site LIKE 'CCA%'\nLIMIT 5
```

DATE	time_utc	booster_version	launch_site	payload
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2

# Total Payload Mass

---

```
%sql SELECT sum(payload_mass__kg_) as Total_palyoad_kg \  
FROM SPACEXDATASET \  
WHERE customer = 'NASA (CRS)'
```

A terminal window with a dark background. The first line shows the column name 'total\_palyoad\_kg' in a light blue font. The second line shows the result '45596' in a light blue font.

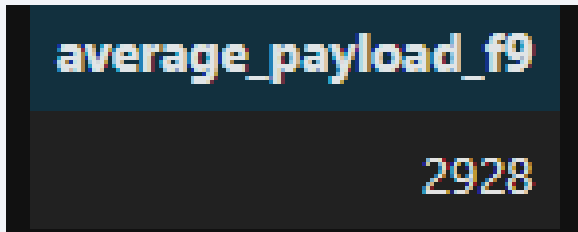
total_palyoad_kg
45596

- The total payload carried by boosters from NASA is 45596

# Average Payload Mass by F9 v1.1

---

```
%sql SELECT avg(payload_mass__kg_) as average_payload_F9\  
FROM SPACEXDATASET\  
WHERE booster_version LIKE 'F9 v1.1'
```



A terminal window with a dark background. The first line shows the column name 'average\_payload\_f9' in a light blue font. The second line shows the result '2928' in a light blue font.

average_payload_f9
2928

- The average payload mass carried by booster version F9 v1.1 is 2928

# First Successful Ground Landing Date

---

```
%sql SELECT avg(payload_mass__kg_) as average_payload_F9\  
FROM SPACEXDATASET\  
WHERE booster_version LIKE 'F9 v1.1'
```

**DATE**  
**2015-12-22**

- The date of the first successful landing outcome on ground pad was December 12<sup>th</sup> 2015

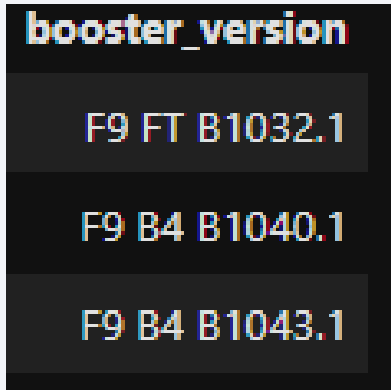
DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)



# Successful Drone Ship Landing with Payload between 4000 and 6000

---

```
%sql SELECT *\nFROM SPACEXDATASET\nWHERE DATE in ( SELECT min(DATE) as DATE\n                FROM SPACEXDATASET\n                WHERE landing__outcome = 'Success (ground pad)' )
```



booster_version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

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

# Total Number of Successful and Failure Mission Outcomes

---

```
%sql SELECT (SELECT count(mission_outcome) as success FROM SPACEXDATASET WHERE  
mission_outcome LIKE 'Success%'),\  
          (SELECT count(mission_outcome) as failure FROM SPACEXDATASET WHERE mission_outcome LIKE  
'Failure%')\  
FROM SPACEXDATASET\  
LIMIT 1
```

success	failure
100	1

- The total number of successful was 100 and there was 1 failure mission

# Boosters Carried Maximum Payload

```
%sql SELECT booster_version\  
FROM SPACEXDATASET\  
WHERE payload_mass__kg_ in ( SELECT max(payload_mass__kg_) as  
payload_mass__kg_\  
FROM SPACEXDATASET )
```

- List of the names of the booster which have carried the maximum payload mass

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

# 2015 Launch Records

---

```
%sql SELECT DATE,landing__outcome, booster_version, launch_site\  
FROM SPACEXDATASET\  
WHERE EXTRACT(YEAR FROM DATE) = 2015 AND landing__outcome = 'Failure (drone ship)'
```

DATE	landing__outcome	booster_version	launch_site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

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

```
%sql SELECT landing__outcome,count(landing__outcome) AS  
TOTAL\  
FROM SPACEXDATASET\  
WHERE DATE BETWEEN '2010-06-04' and '2017-03-20\  
GROUP BY landing__outcome\  
ORDER BY total DESC
```

- 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, in descending order

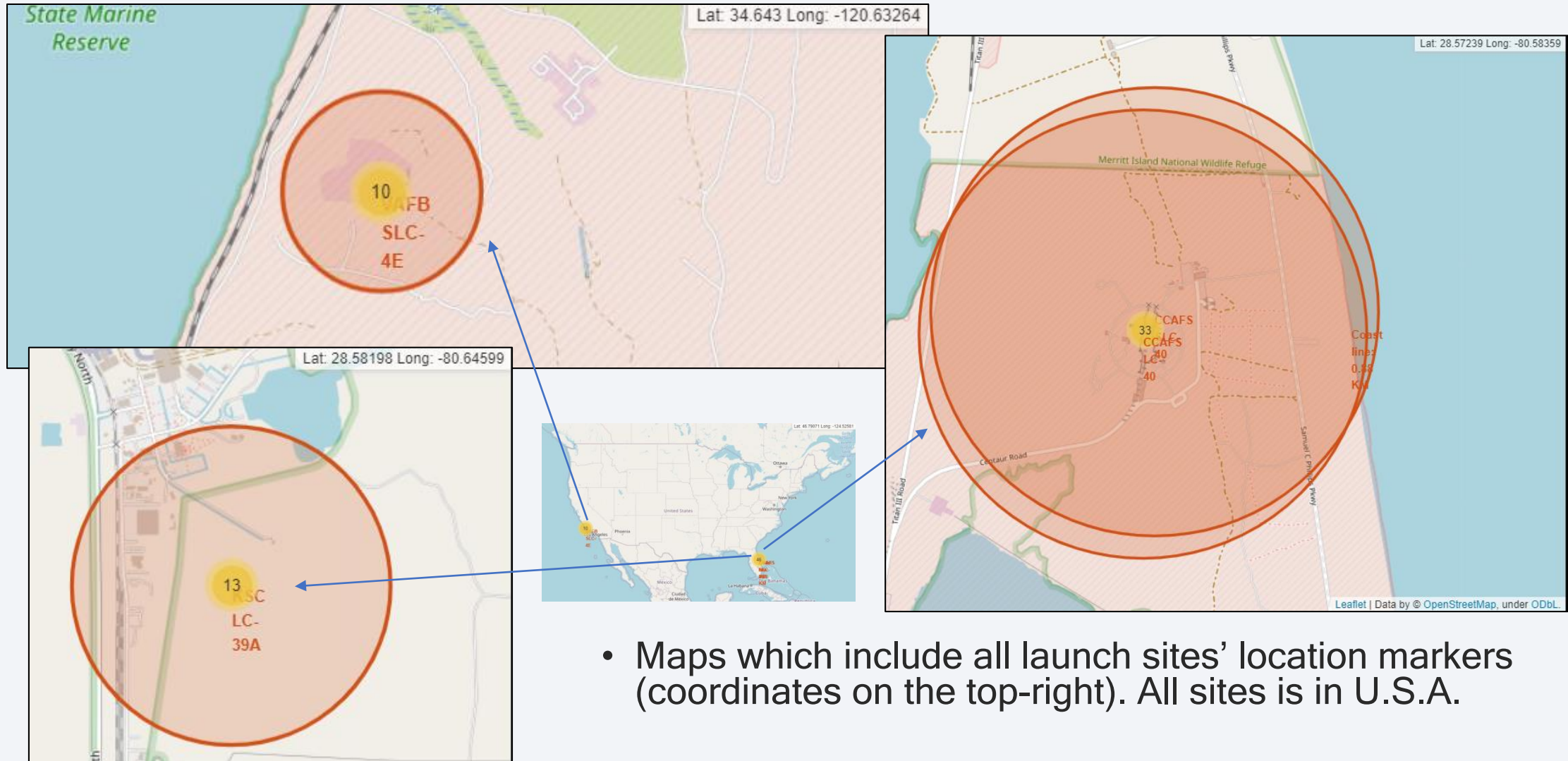
landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

Section 3

# Launch Sites Proximities Analysis

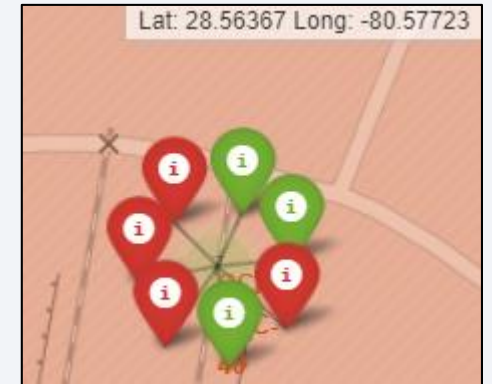
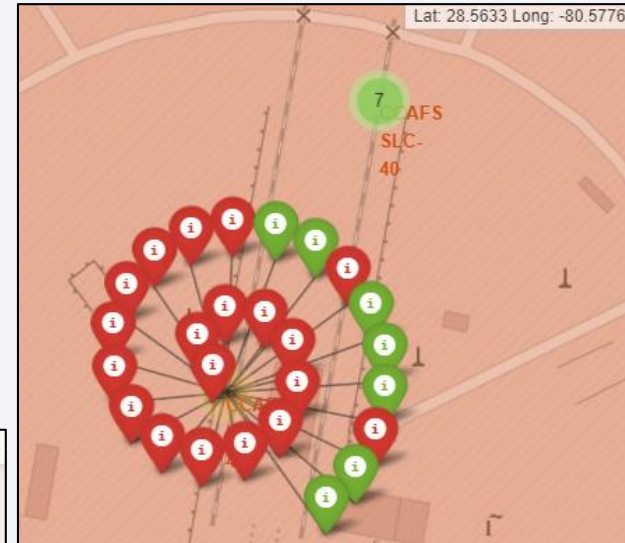
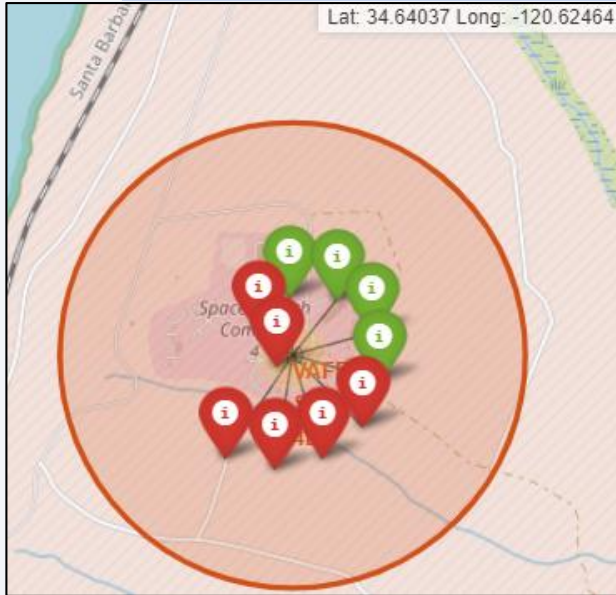
# Launch Sites Locations



- Maps which include all launch sites' location markers (coordinates on the top-right). All sites is in U.S.A.



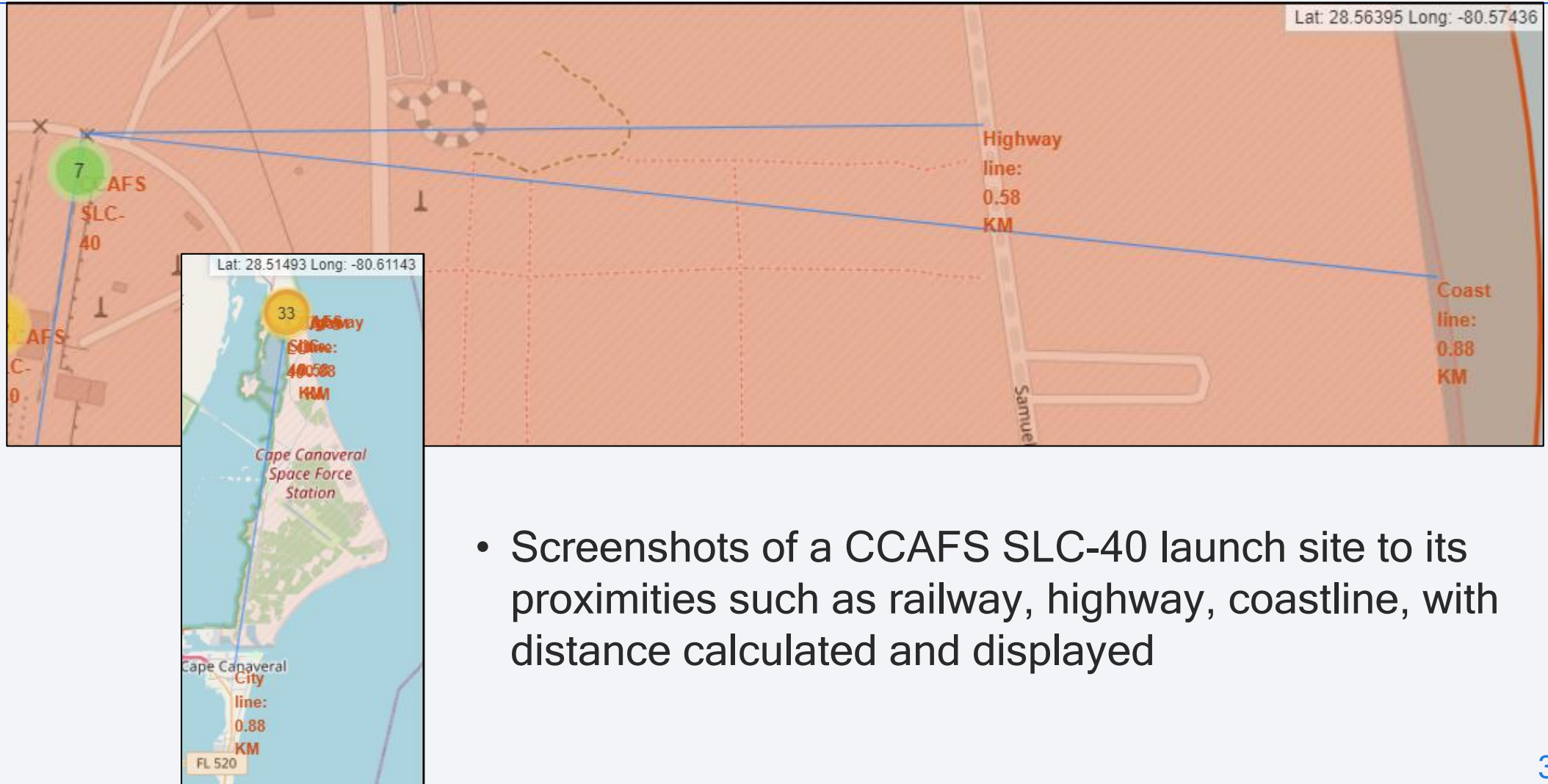
# Launch sites classified by success and failure



- The maps show the color-labeled launch outcomes on the map



# Proximity to roads and the coast

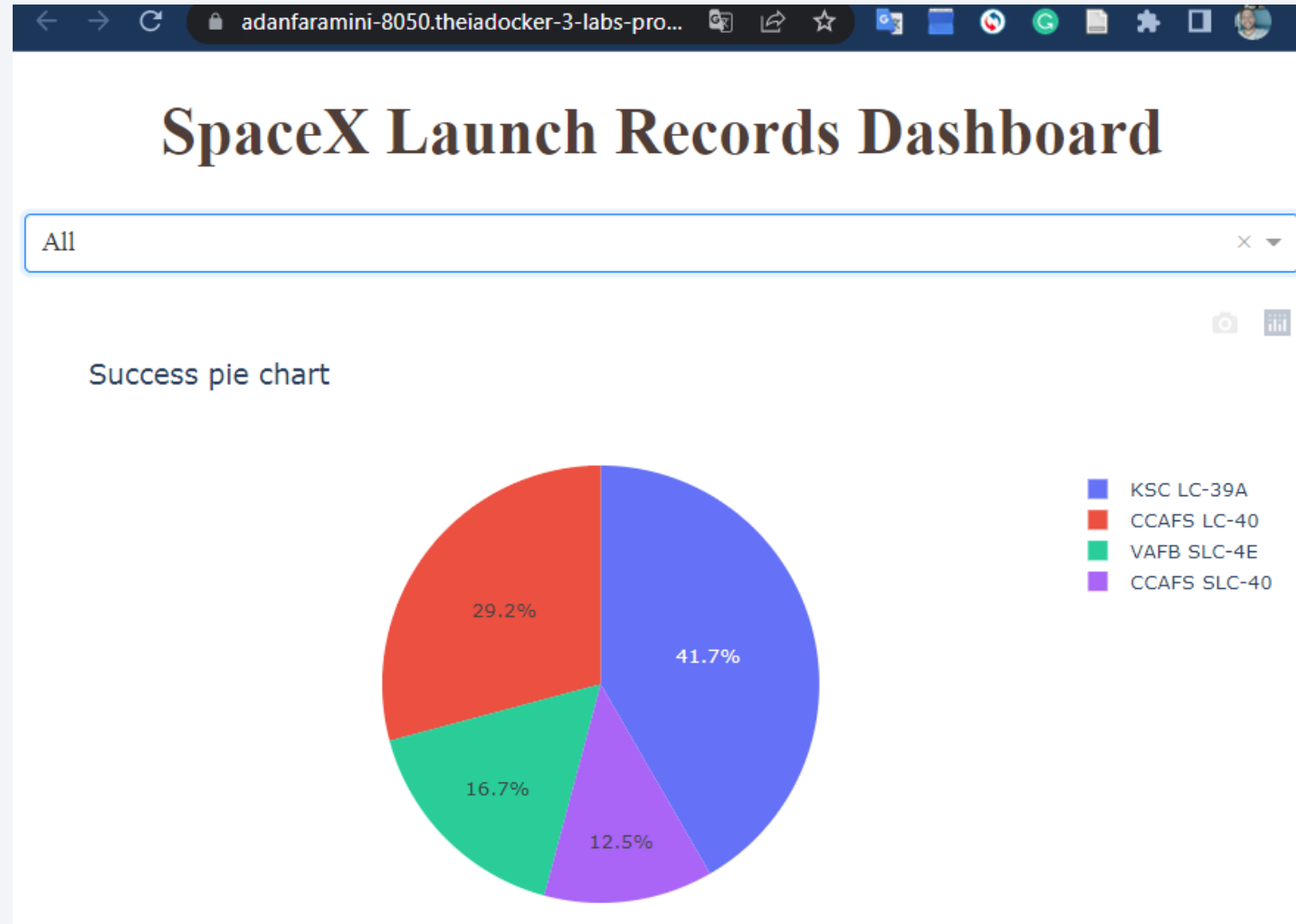




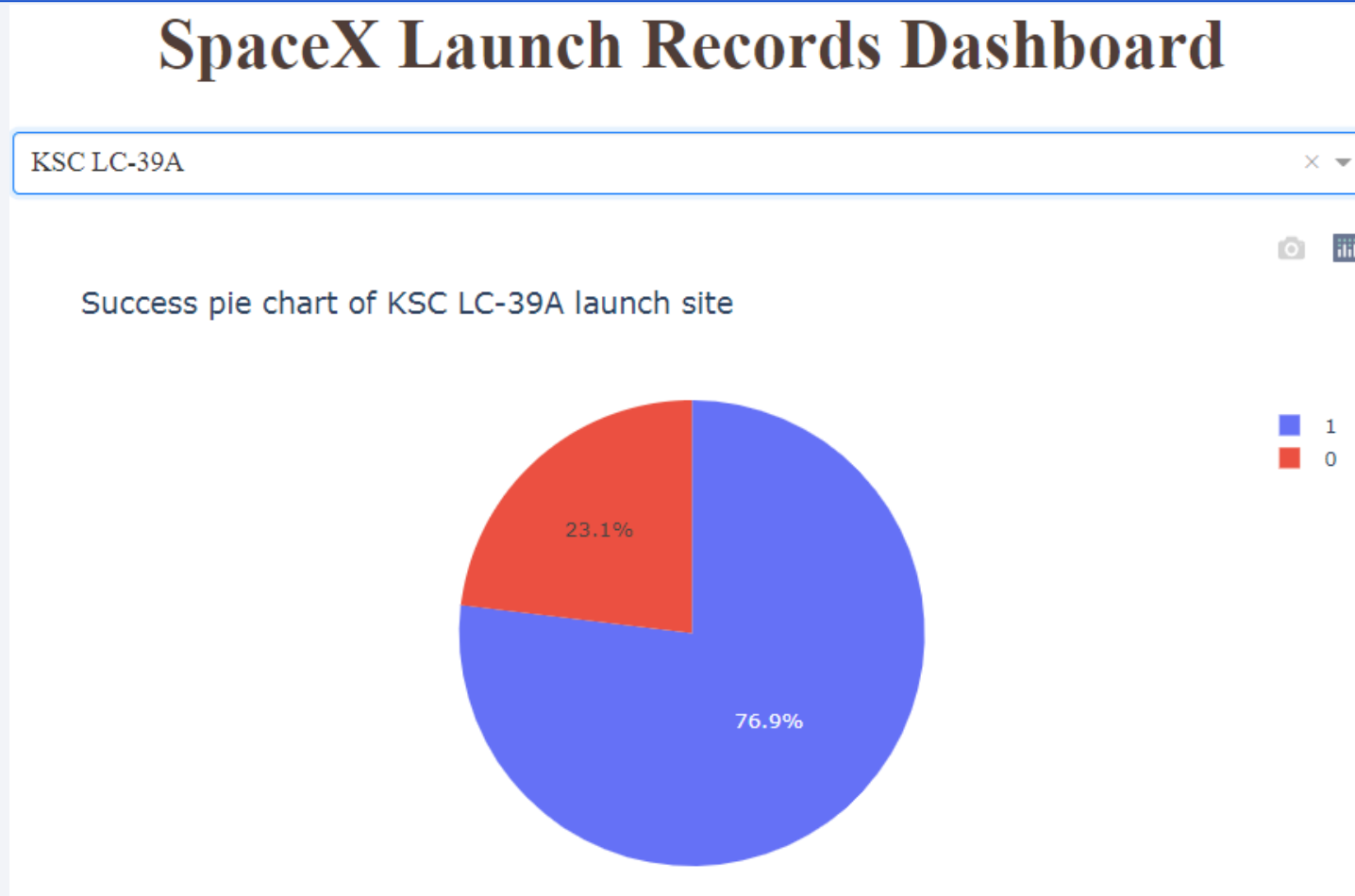
Section 4

# Build a Dashboard with Plotly Dash

# Piechart of launch success count for all sites

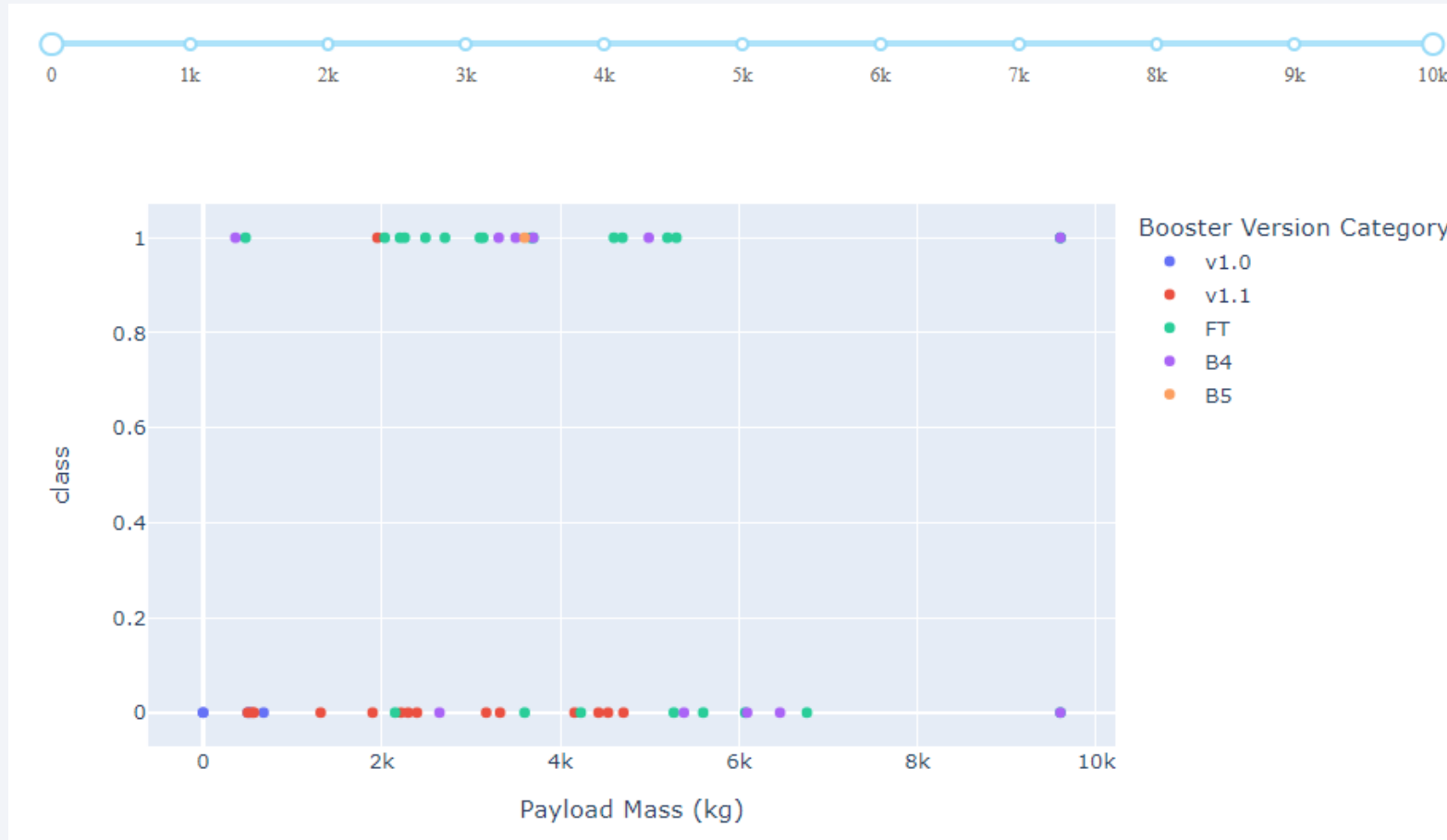


# Piechart of launch success count for all sites



- The launch site with the high ratio success was KSC LC-39A (remember that 1 is to success) with a 76.9%.

# Payload vs. Launch Outcome scatter plot for all sites



- The FT version of Falcon 9 was the most successful in retaining the fuselage, followed by the B4 version.



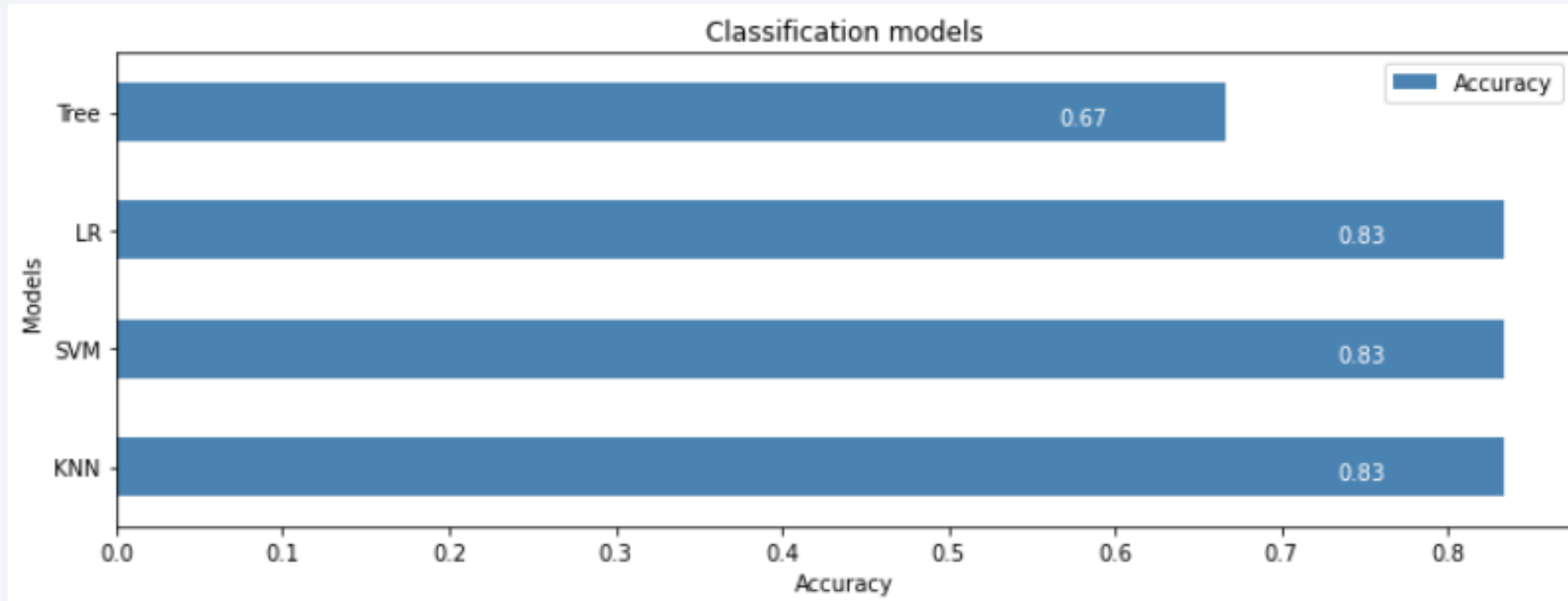


Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

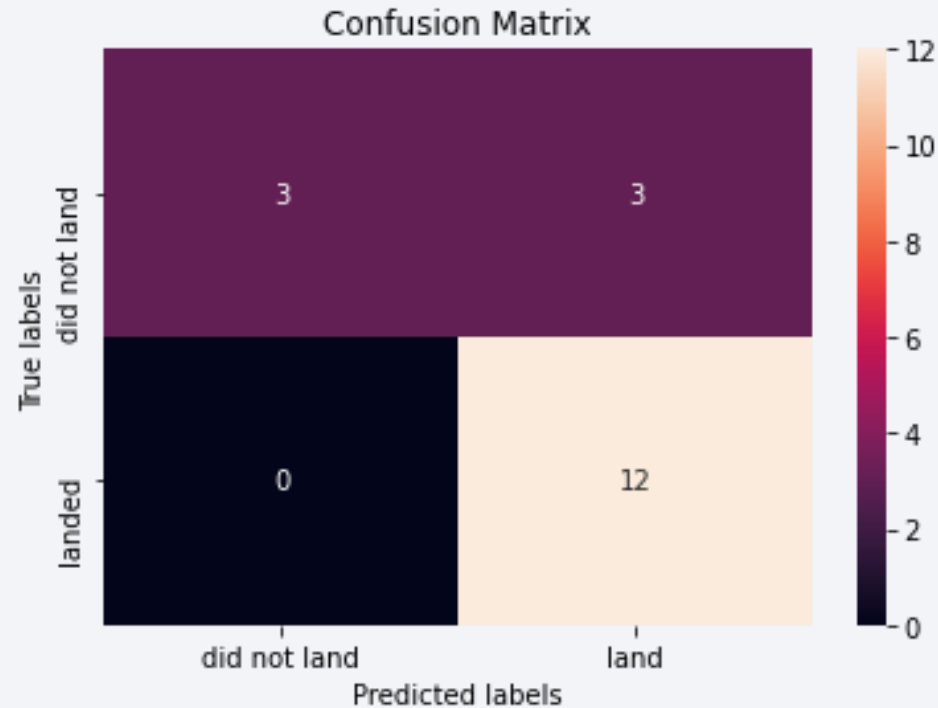
---



- The best models are LR, SVM and kNN

# Confusion Matrix

---



- Confusion matrix of the best performing (SVM) model. Only 3 values were a bad prediction. These are classified as False negative.



# Conclusions

---

- This presentation showed the results obtained by applying data science. For this, public data of the launches of the Falcon rocket of the Tesla company were downloaded. The objective of this work was to analyze through data the possibility of reusing stage 1 of a rocket.
- The study was conducted only for versions of Falcon 9.
- It was observed that the Tesla company managed to increase the percentage of success in reusing stage 1 in an ascending manner between 2013 and 2020. The data showed that payload increases the probability of success. On the other hand, a relationship with success was observed when the rockets were directed to predetermined orbits.
- Classification models were developed in order to estimate the success of the missions. An accuracy close to 0.84 was obtained with the SVM, LR and kNN algorithms.

Thank you!

