IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mahmoud Essam
3 Aug. 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API and Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL and Data Visualization “Dash”
  - Interactive Visual Analytics with Folium “Maps”
  - Machine Learning Prediction “Classification of Success or fail”
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

  SpaceX 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 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch

  ## Problems you want to find answers

  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions need to be in place to ensure a successful landing program?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data has been obtained from 2 main resources
    - SpaceX API (https://api.spacexdata.com/v4/launches/past)
    - Web Scraping using BS (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Data wrangling
  - Was processed to check up full data and make sure everything was stable and there were no null, with making target value for classification process

Methodology

Executive Summary

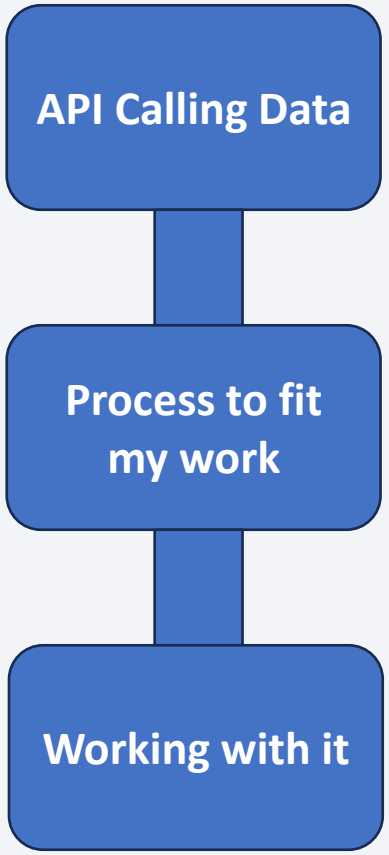
- Perform exploratory data analysis (EDA) using Matplotlib and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - After the process of collecting data, use all needed models to solve a classification problem, “SVM – KNN – Decision T – Logistic Regression” and make comparison between all of them and chose best model with best hyperparameters for the data.

Data Collection

- Describe how data sets were collected.
  - The data has been collected using the Requests of API link https://api.spacexdata.com/v4/launches/past Which helped us to ask the server “having the data” to make a request and retrieve this data to perform our needed EDA and perform needed action
  - The second step was getting the data of Rockets using Web scraping from Wikipedia, and this problem was handled using the Beautiful Soup package from Python and retrieved all needed tables from it.

Data Collection – SpaceX API

- To obtain data from the API, we initiated requests to the server.
- These requests were made to retrieve simple data related to their operations.
- This approach facilitated the acquisition of essential information about their historical launches, including both successful and failed missions.
- We also conducted processes like data type verification and configuration setup to ensure the data's accuracy and readiness for use.



Data Collection - Scraping

- We needed to scrap the rockets and their works and types, so we accessed to Wikipedia Link of rockets specifically Rocket number 9
- After scraping data and making a copy of the HTML code, made some fetching for the needed tables to put it in our side to know why some of the rockets failed or even what's the reason for failure
- Lastly, the copy of this data with us now with all the needed details to work with it in the process of wrangling

Data Wrangling

- The purpose of this process was to make a full checkup on our data and find if there's something interesting in this data like:
  - Knowing the number of launches on each site which helped us to know which launch site has a lot of launches performed at
  - Checking the number and occurrence of each orbit, by knowing how level those rockets have reached
  - Knowing the number and occurrence of mission outcomes of the orbits
  - Lastly, Made simple column for success and failures

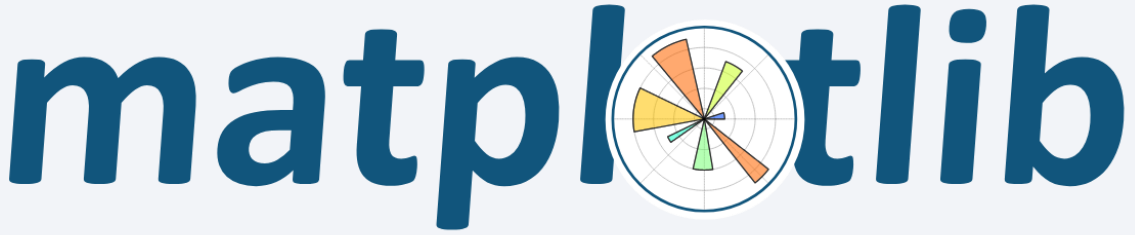
By the way
The mean percent of success in launches was about 66.67%

EDA with SQL

- The most impressive part was Expletory this data using SQL, and the only purpose of this part was to know the reasons for bad landing outcomes
  - Reasons why that landing was bad
  - Number of successes and failed landings
  - How amount of payload mass handled
  - When was the first success, and who was the reason for this success

EDA with Data Visualization

- Just Simple Visualizations have happened to know the percent of which launch site was the best or even to know simple reasons for failures happened
  - Knowing From the patterns and draws.
  - Percent of Success and fails happened along the duration of all launches
  - Checking of all relations between each feature in the data



Build an Interactive Map with Folium

- The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors and hopefully, we could discover some of the factors by analyzing the existing launch site locations.

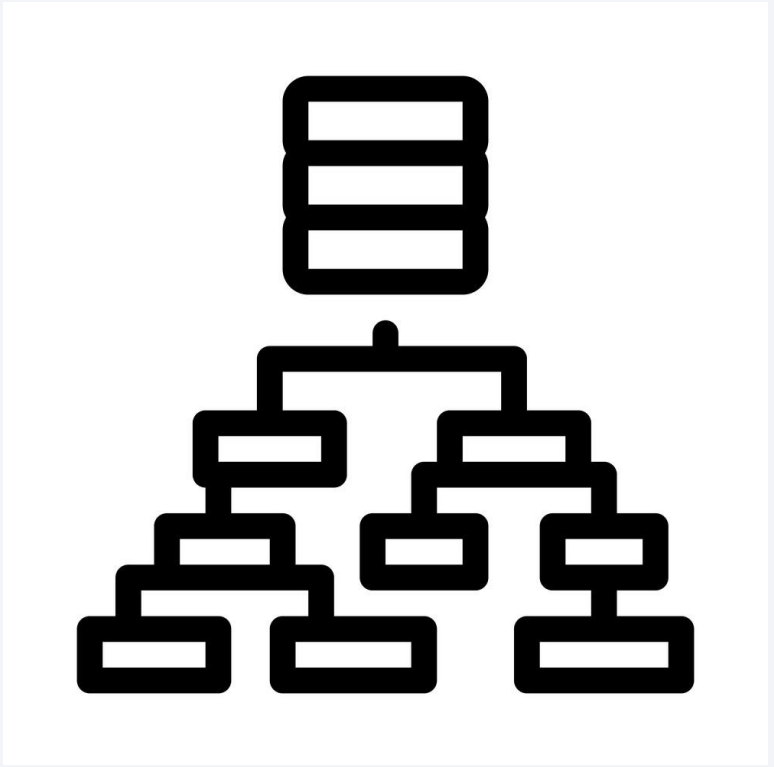
Build a Dashboard with Plotly Dash

- Enhance the interactivity of Plotly to create a dynamic and engaging data exploration experience, resembling an application where users actively interact with the data to find solutions.
  - The goal is to analyze the success and failure rates of various launch sites. Specifically, we want to:
  - Investigate the success and failure rates of each launch site.
  - Examine the payload details for each launch site.
  - Determine the success rate of payloads launched from each site and their corresponding quantities.

Predictive Analysis (Classification)

- For Each data, suppose there's a model to make it usable for new data, to save a lot of money for the future, we are here trying to test the following Models to retrieve who's the best one
  - Logistic Regression
  - Tree Decision
  - Support Vector Machine
  - K Nearest Neighbor

  Sure to know the best parameters we used Hyperparameter Concept



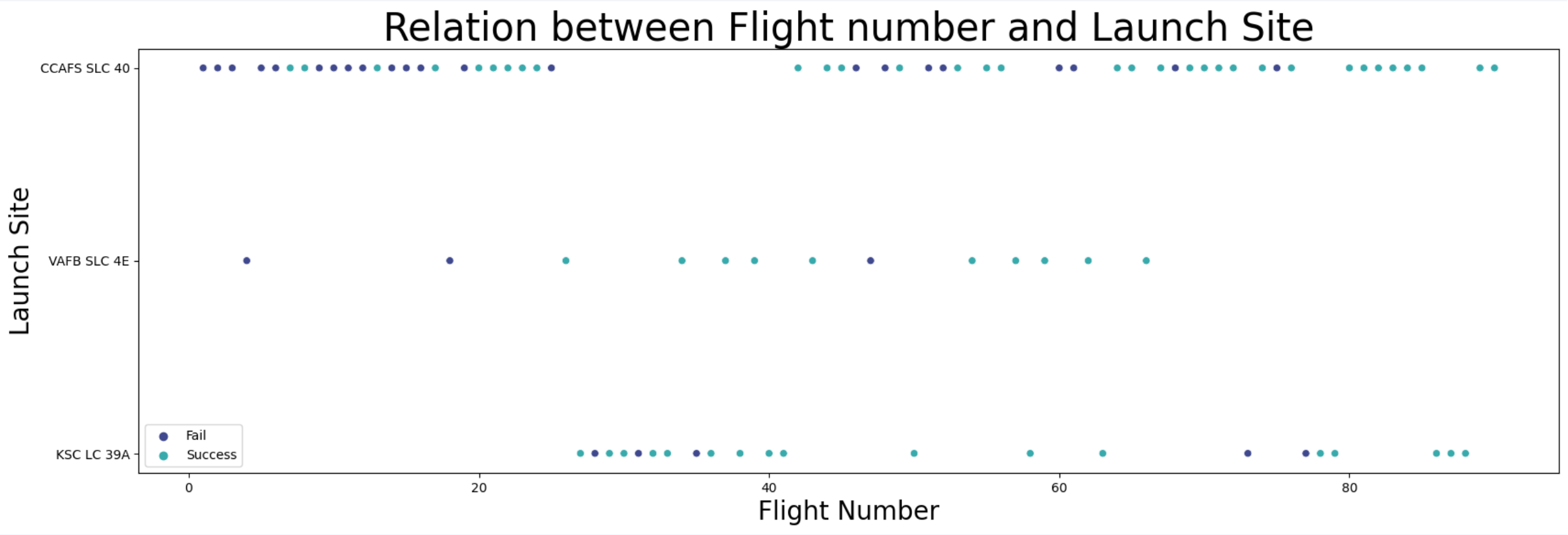
Results

Let’s Start it!

Section 2

Insights drawn from EDA

Viz: Flight Number vs. Launch Site



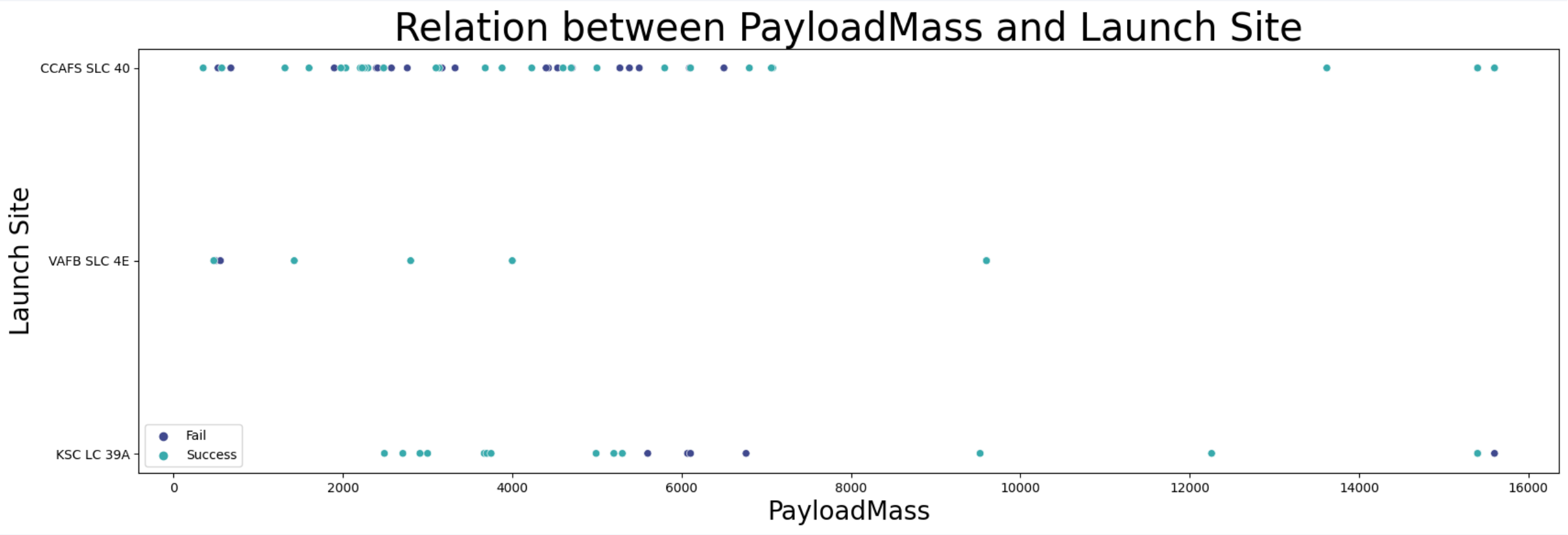
There are Great Notes, Press for the Next Page

Viz: Flight Number vs. Launch Site

1. **CCAFS SLC 40 (Success Rate = 60%)**:
    1. Shows a relatively low success rate of 60%.
    2. More failed launches compared to successful ones.
    3. Has a high number of flight attempts.
2. **KSC LC 39A (Success Rate = 77.72%)**:
    1. Displays a higher success rate of 77.72%.
    2. Majority of points are above the success rate line, indicating more successful launches.
    3. Has a substantial number of flight attempts.
3. **VAFB SLC 4E (Success Rate = 76.9%)**:
    1. Has a success rate close to KSC LC 39A, at 76.9%.
    2. Fewer data points suggest fewer launches.
    3. Fewer failures, possibly indicating a more reliable launch site.

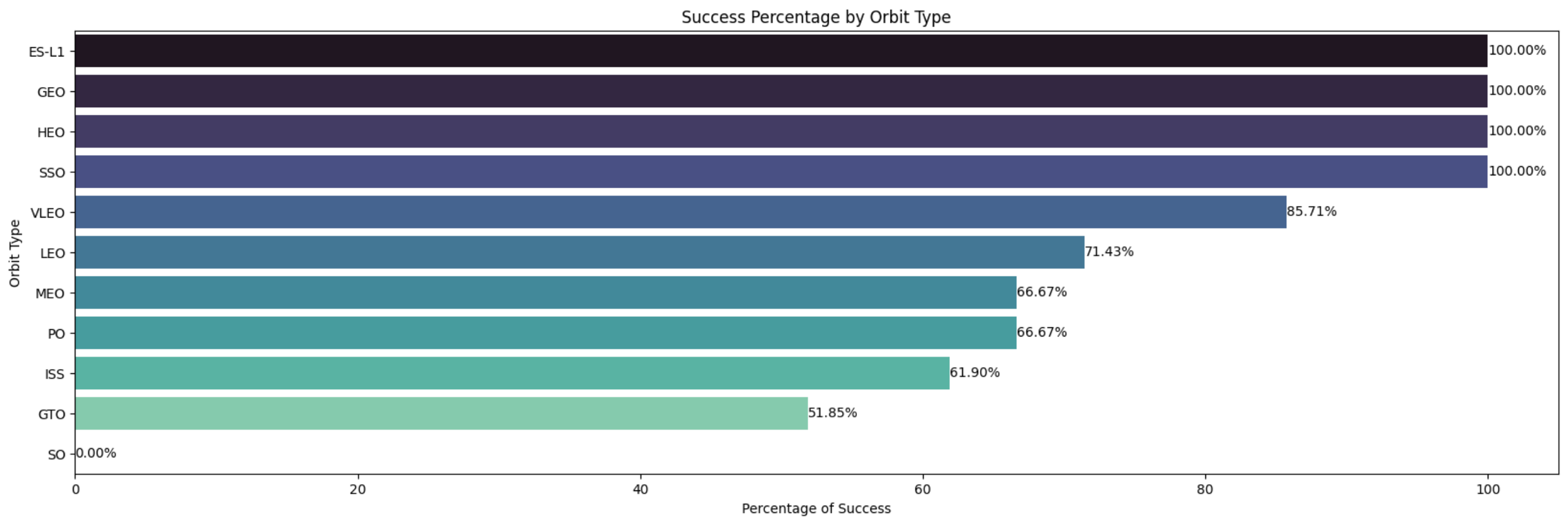
-VAFB SLC 4E's success with fewer launches might be due to favorable conditions and quality control.
-Lessons from CCAFS SLC 40's more launches may improve VAFB SLC 4E's success.
-Some failures at CCAFS SLC 40 succeeded elsewhere, highlighting site-specific factors.

Viz: Payload vs. Launch Site



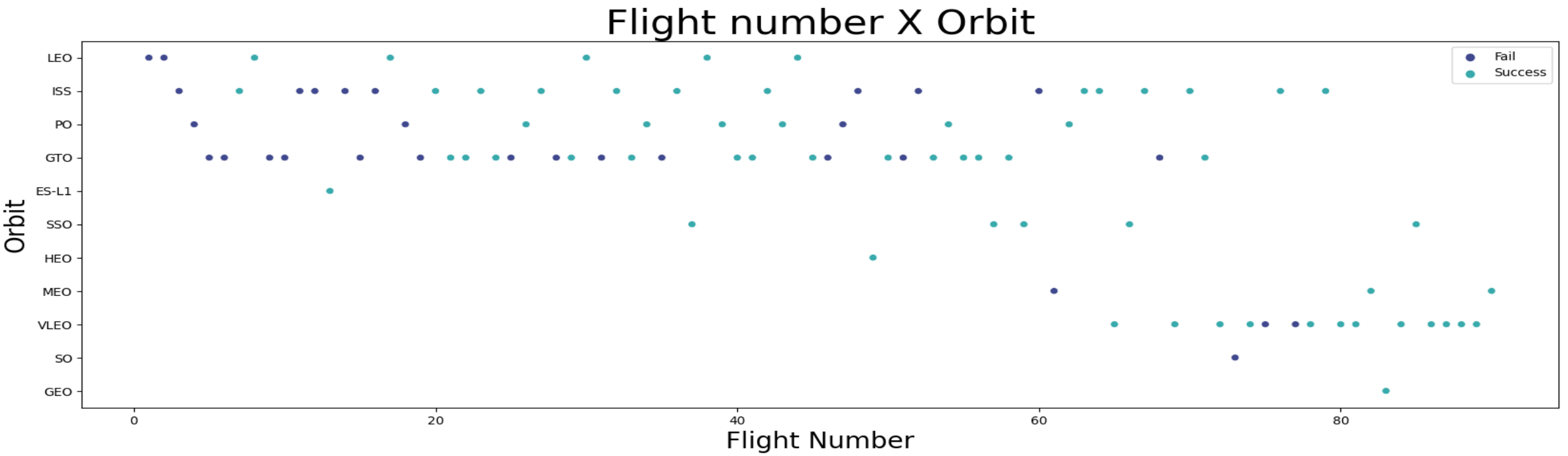
KSC LC 39A wasn't utilized until CCAFS served as the initial testing ground, and VAFB solely paved the path to success. Hence, we can consider CCAFS as the testing site for addressing any challenges until they achieve higher success rates and distinguish this success from other launch facilities.

Viz: Success Rate vs. Orbit Type



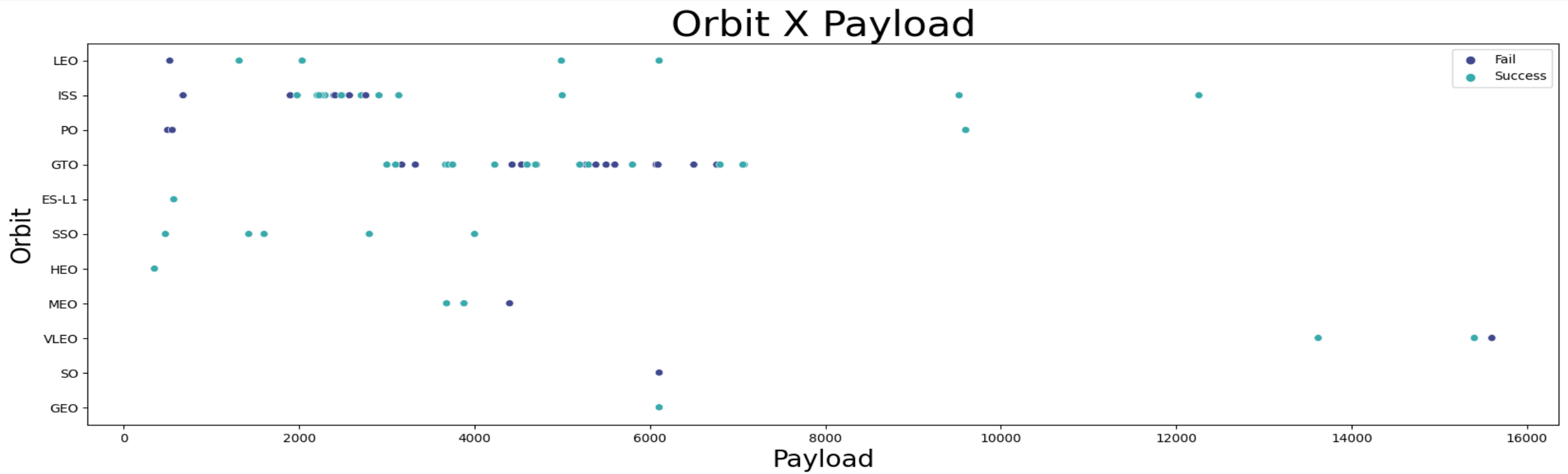
The key factor behind the 100% success rate of GEO, HEO, and ES-L1 missions, despite their distance from Earth, is that they underwent only one attempt. In contrast, GTO and ISS missions were subjected to numerous trials, which resulted in lower success percentages. These latter missions served as testing grounds to eventually achieve a 100% success rate for more distant locations.

Viz: Flight Number vs. Orbit Type



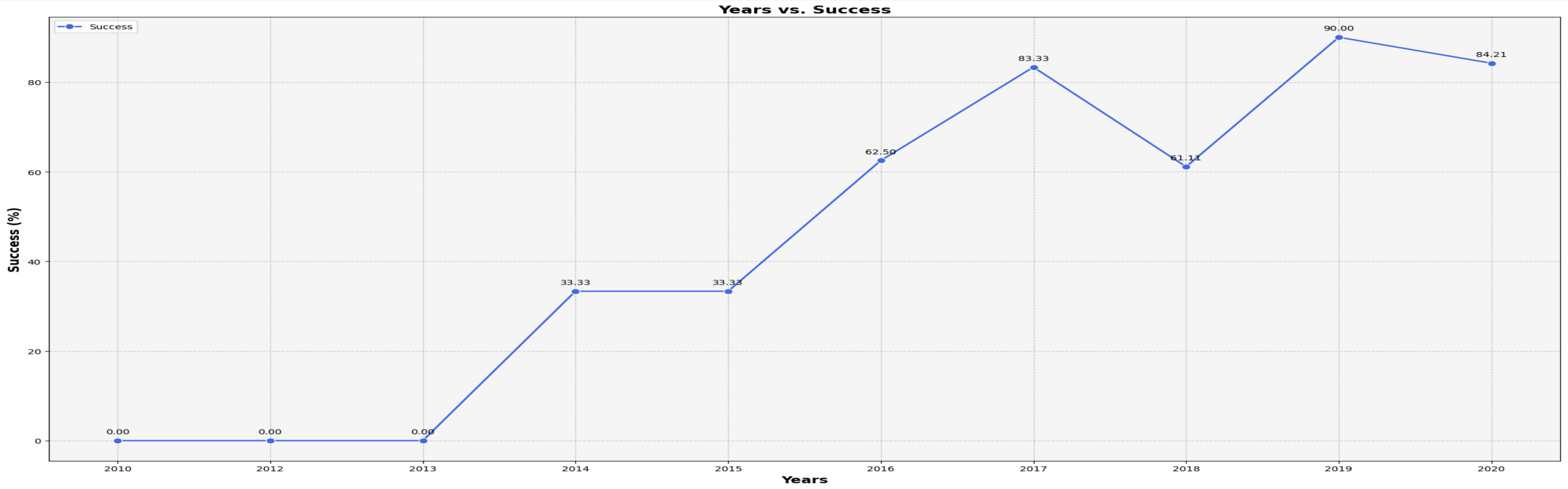
LEO was the initial orbit where testing began, but it faced two initial failures. However, through extensive trials on LEO, ISS, and GTO orbits, these efforts eventually led to success in the later attempts, achieving a 100% success rate for missions like SSO after approximately five launch attempts. Launching to GEO proved to be a significant milestone, as it marked the farthest orbit compared to all previous rocket missions, ultimately setting the stage for subsequent missions to achieve a higher success rate.

Viz: Payload vs. Orbit Type



When dealing with heavy payloads, we tend to observe a higher rate of successful or positive landings for missions targeting Polar, LEO (Low Earth Orbit), and ISS (International Space Station) orbits. However, in the case of GTO (Geostationary Transfer Orbit), it's challenging to make a clear distinction between successful and unsuccessful missions because both positive landing rates and negative landing rates (indicating unsuccessful missions) coexist in this orbit.

Viz: Launch Success Yearly Trend



Starting from 2010, there was a four-year period with no successful missions until 2014. During 2015, the success rate saw an improvement, reaching 33.3%. However, it experienced a surprising increase until 2018, reaching a remarkable 61%. It's worth noting that this surge may have been influenced by the historic Falcon 9 water landing on December 5, 2018. Fortunately, the success rate rebounded and reached its highest point in 2019.

SQL: All Launch Site Names

```
1  %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
```

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

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

The intention is to provide insight into which sites have data readily available.

SQL: Launch Site Names Begin with 'CCA'

```
%%sql
SELECT *
FROM SPACEXTABLE
WHERE Launch_Site like "CCA%"
LIMIT 5
```

Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Let's give a look for CCA sites, which was the most site has launched about a lot of rockets from

SQL: Total Payload Mass

```
1  %%sql
2  SELECT sum(PAYLOAD_MASS__KG_)
3  FROM SPACEXTABLE
4  WHERE Customer = "NASA (CRS)"
```

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

sum(PAYLOAD_MASS__KG_)
45596

The total payload mass carried by boosters launched by NASA (CRS) Was about 45596 KG to space

SQL: Average Payload Mass

```
1  %%sql
2  SELECT AVG(PAYLOAD_MASS__KG_)
3  FROM SPACEXTABLE
4  WHERE Booster_Version like "F9 v1.1%"
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

The Average payload mass carried by Booster Version F9 v1.1 Rocket Was about 2534.6 KG to space

SQL: First Successful Trip!

```
1  %%sql
2  SELECT min(Date) as "first success landing in ground pad"
3  FROM SPACEXTABLE
4  WHERE Landing_Outcome = "Success (ground pad)"
5
```

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

first success landing in ground pad
2015-12-22

The successful landing occurred on December 22, 2015, marking the conclusion of the year with a success rate of 33.3%.

SQL: Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000 and 6000

```

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

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

F9 Boosters of version B's was so incredible

SQL: Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT Mission_Outcome,Count(Mission_Outcome) as "Totals"
FROM SPACEXTABLE
GROUP BY (Mission_Outcome)
```

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

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

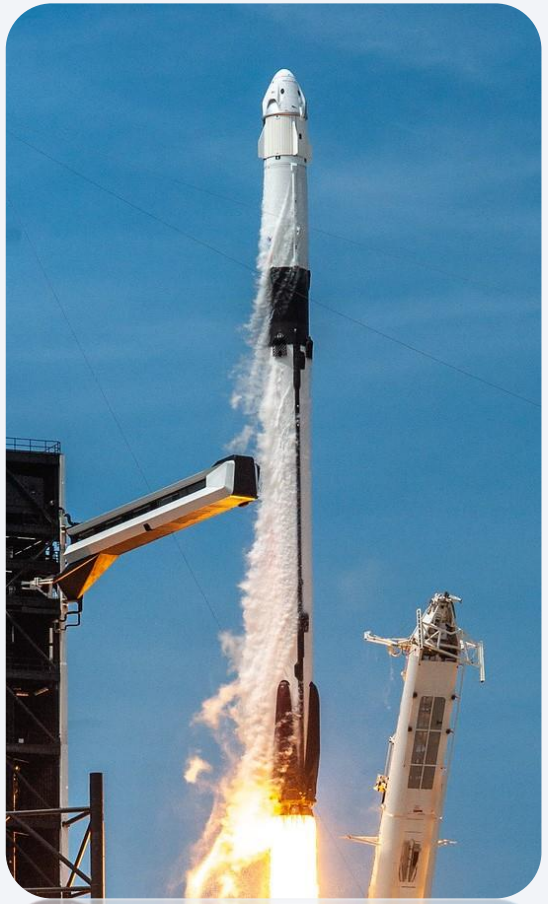
You can say that all the launching trips were so Successful!

SQL: Boosters Carried Maximum Payload

```
1  %%sql
2  SELECT Booster_Version
3  FROM SPACEXTABLE
4  WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

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

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



SQL: 2015 Launch Records

```
%%sql
SELECT substr(Date,6,2) as "Month",substr(Date,0,5) as year , Landing_Outcome, Booster_Version, Launch_Site
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Failure (drone ship)' and year == "2015"

```

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

Month	year	Landing_Outcome	Booster_Version	Launch_Site
10	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

+ Code + Markdown

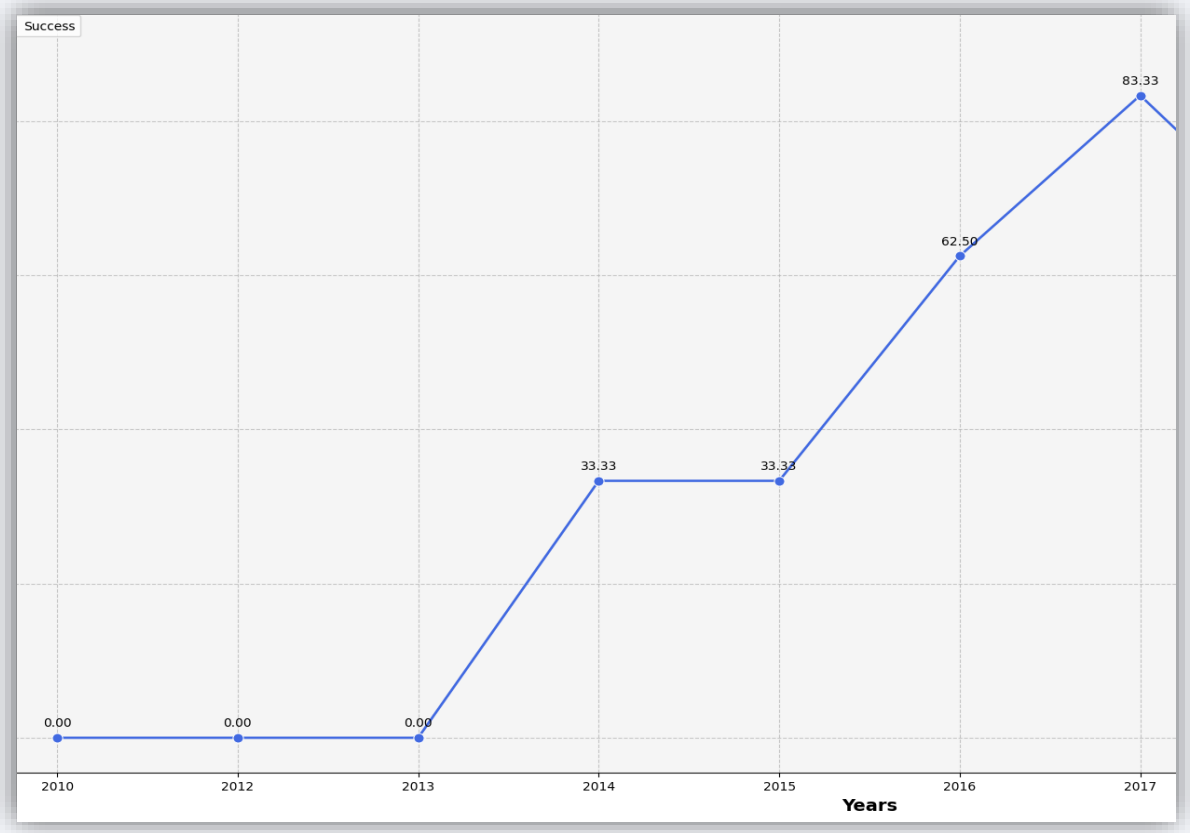
I think that's enough reason to know why the rate was 33%

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

```
%%sql
SELECT Landing_Outcome, count(*) as Quantity
FROM SPACEXTABLE
WHERE DATE between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by Quantity DESC
```

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

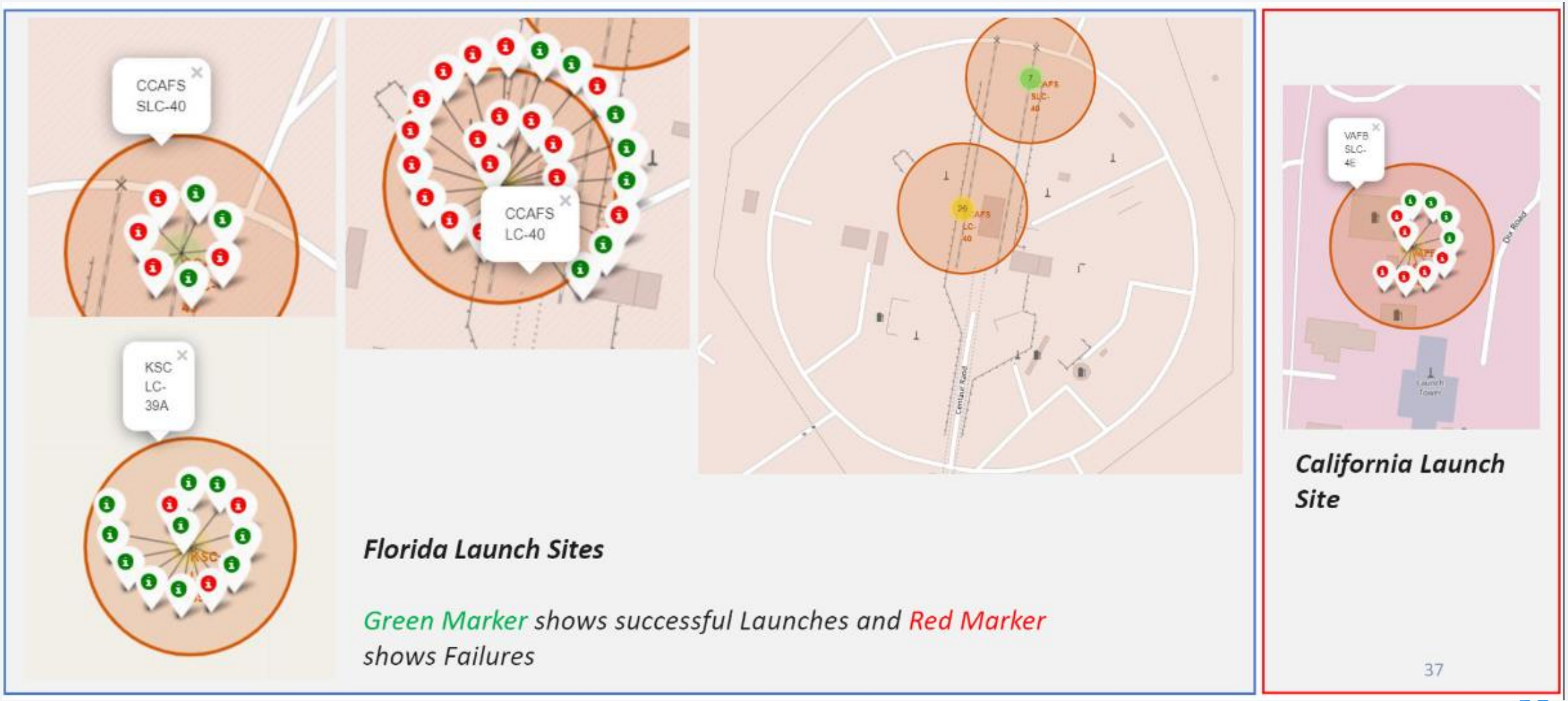
Landing_Outcome	Quantity
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1



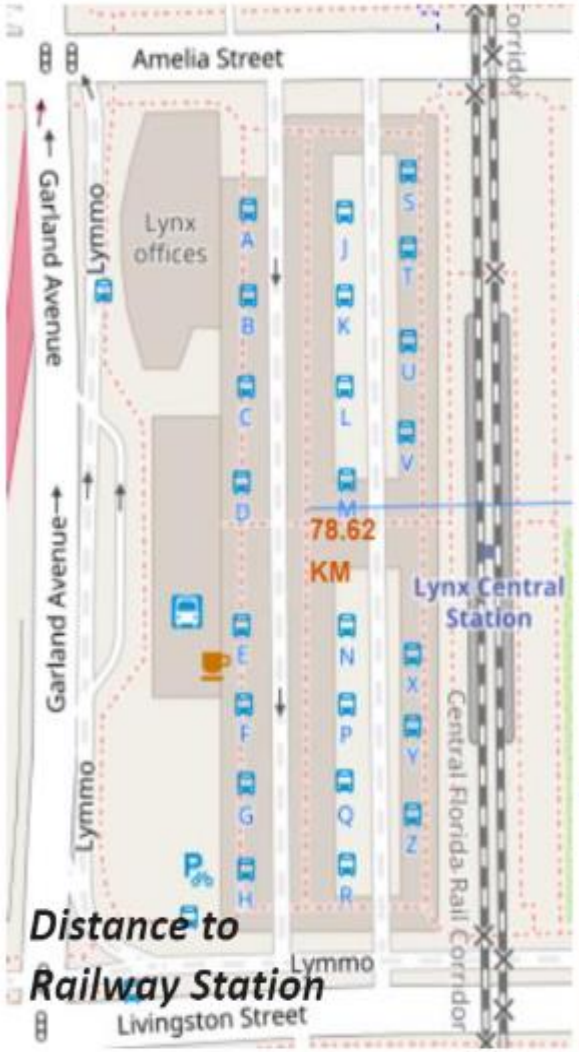

Section 3
Launch Sites
Proximities Analysis

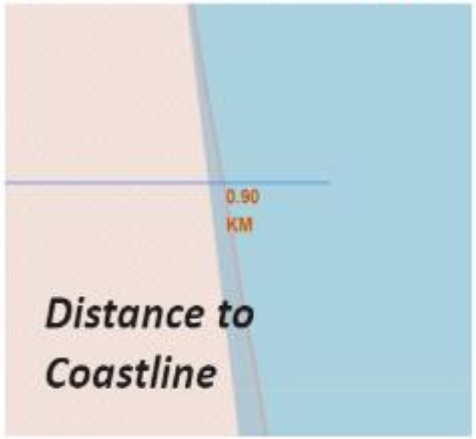
All launch sites global map markers

Markers showing launch sites with color labels

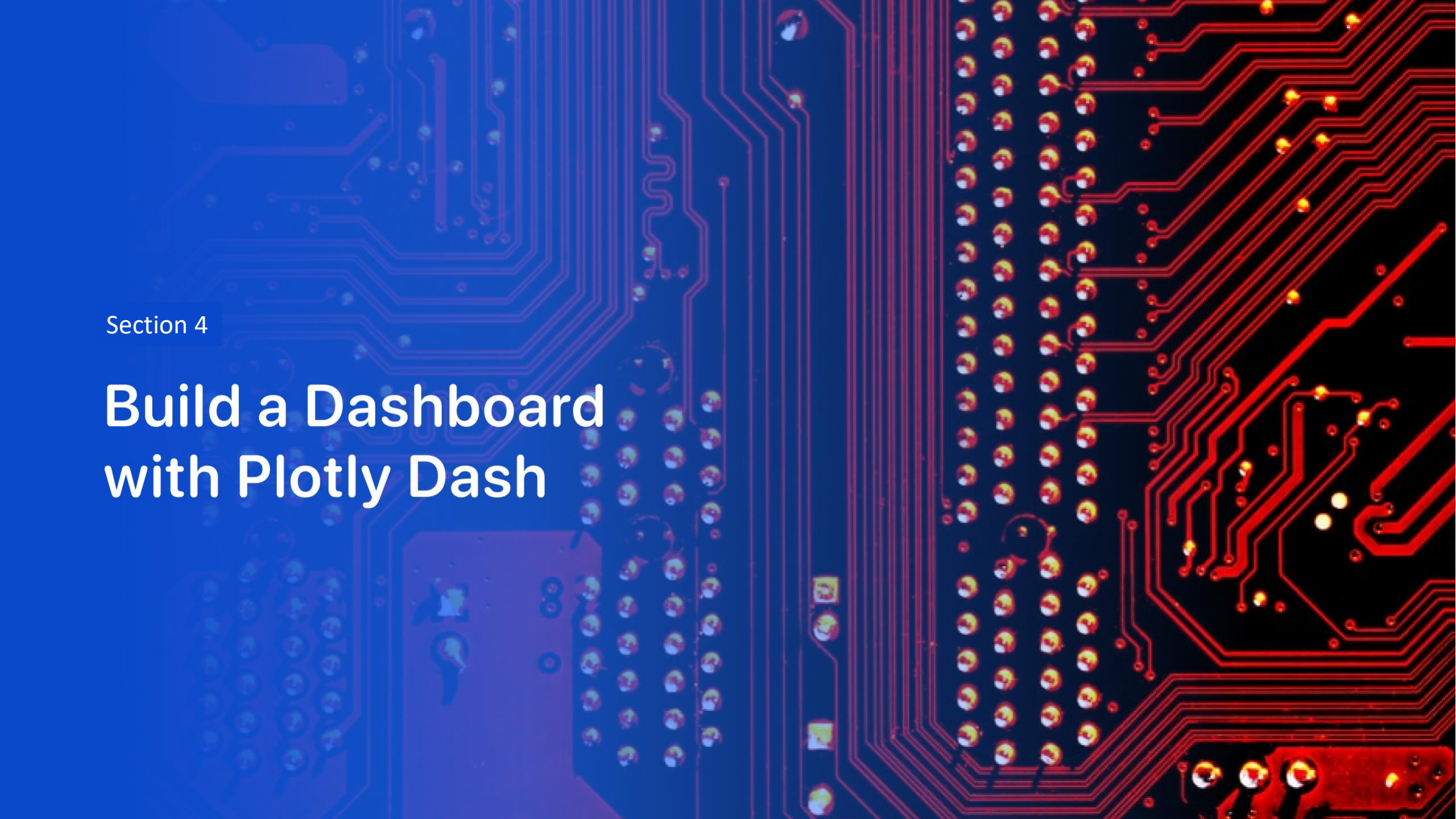


Launch Site distance to landmarks

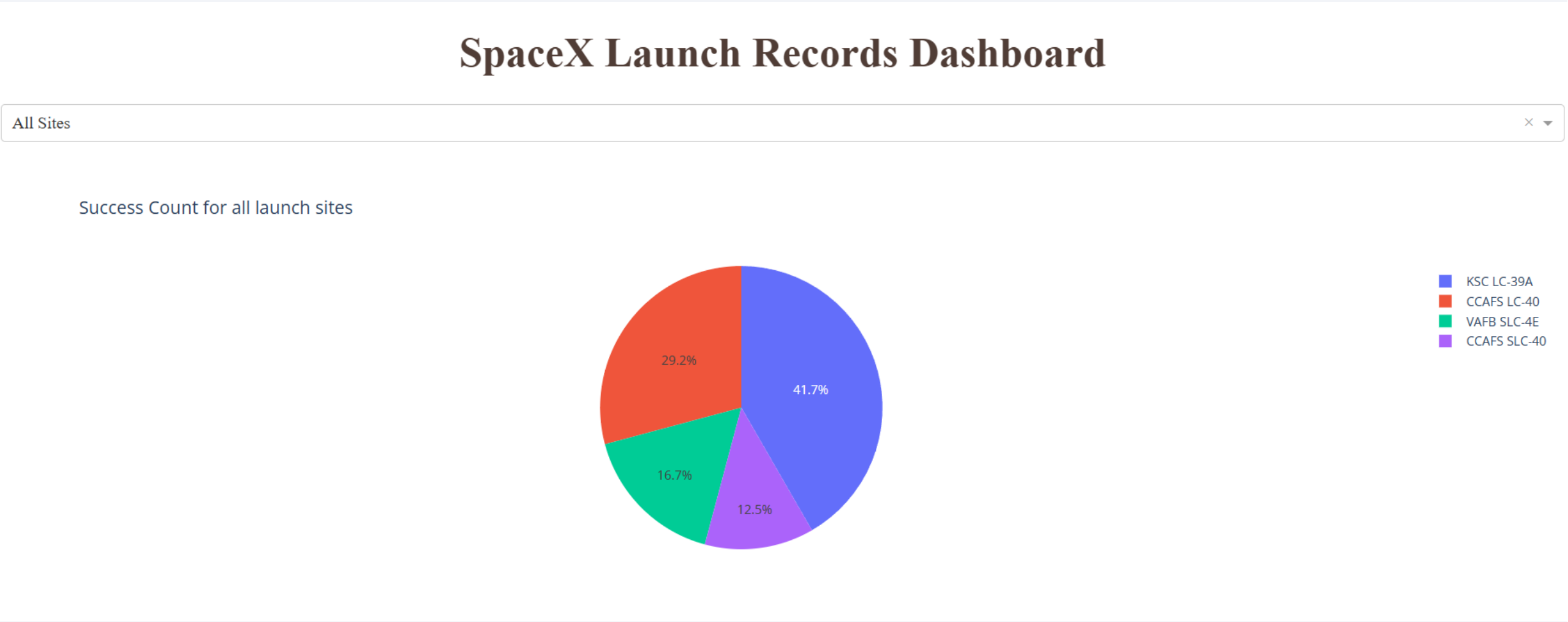




- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

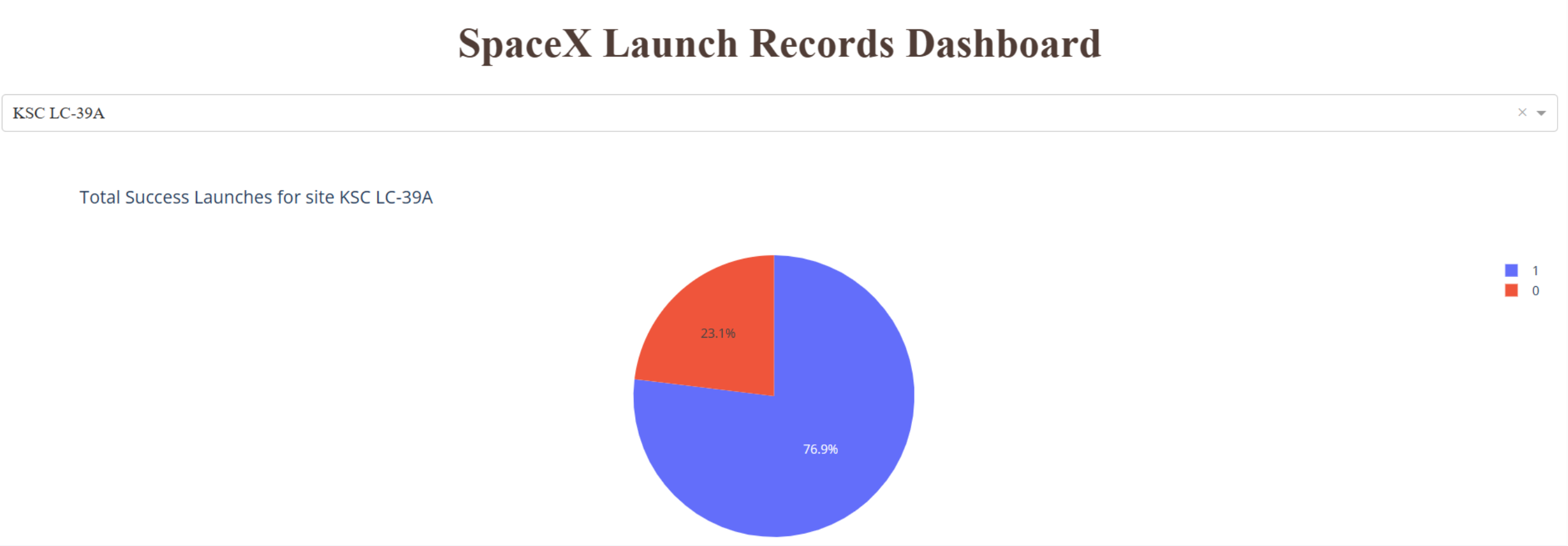

Section 4
Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site



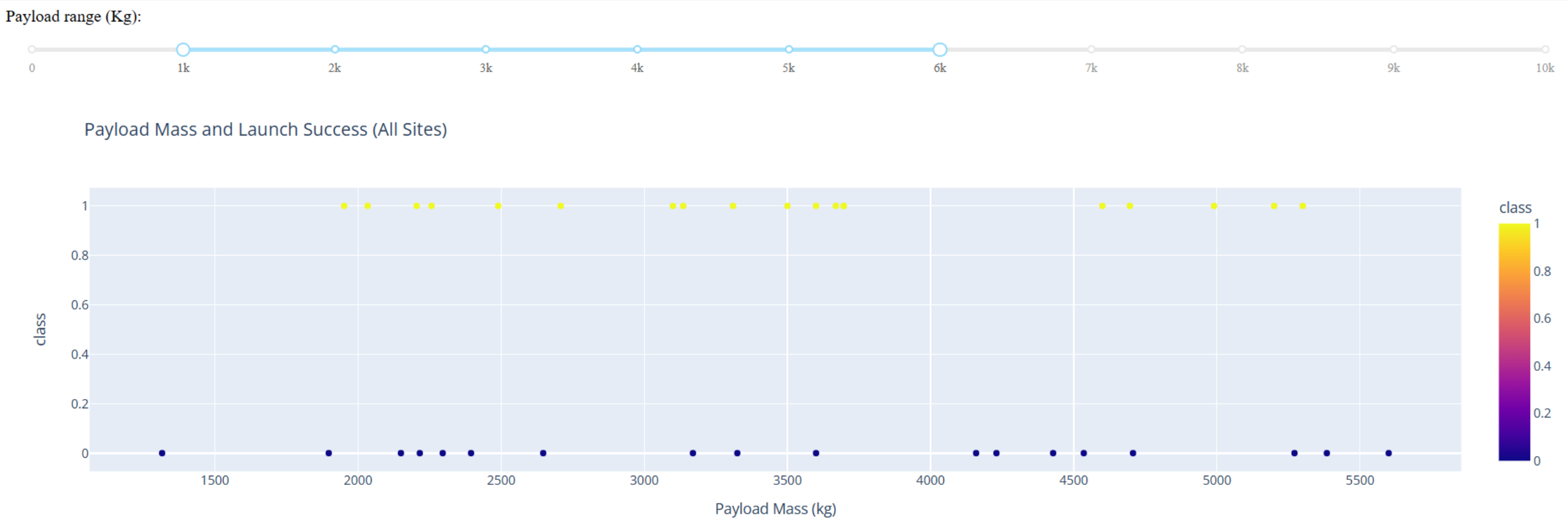
Here will show all success rate of all sites

Pie chart showing the Launch site with the highest launch success ratio



As shown, here KSC Launch Site was the most site with success rate

Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider



Here Payload shown as slider, and send the range to X axis, so it will show the range and figure based of values of slider

Section 5

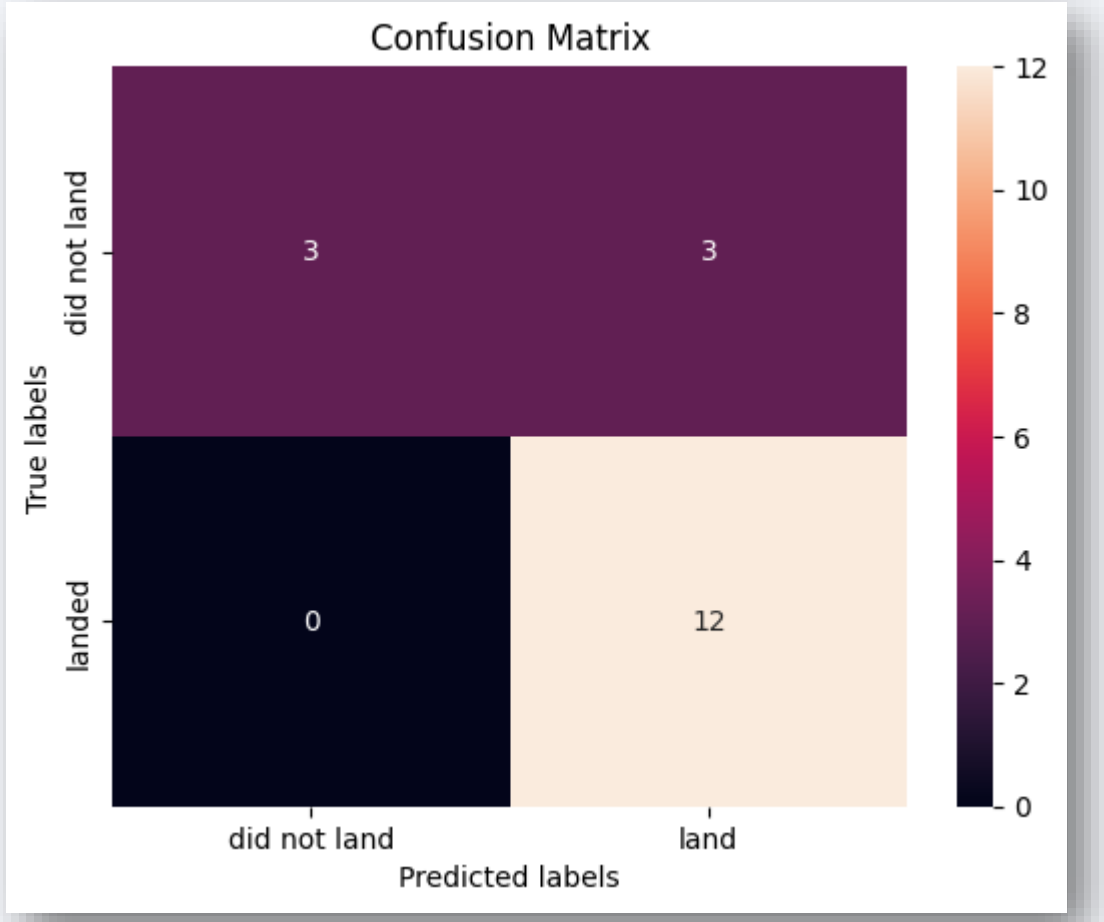
Predictive Analysis (Classification)

Classification Accuracy

Types/Values	Logistic	SVM	KNN	DT
Best Parameters	regularization: 0.1 Penalty: Ridge	Kernal: Sigmoid regularization: 1	Neighbors = 10 Using Manhattan Distance	Depth = 12 Min leafs = 4 Splitting random
Accuracy	84.64%	84.82%	84.82%	87.32%
Test Score	83.34%	83.34%	83.34%	83.34%
Good?	Great	Better than Logistic in accuracy score training	Still better than Logistic, but as same as SVM	Is the best model of all in Accuracy score by training

Confusion Matrix

The degree of Error 1 isn't particularly concerning. The model, by committing Error 1, has successfully reduced the occurrence of false positives, where the rocket hasn't landed but the model incorrectly indicates that it has, albeit to a relatively small extent.



Conclusions

- We can conclude that:
  A higher number of flights at a launch site is associated with a higher success rate at that site, and this trend can potentially positively influence the success rates of other launch sites.

1. The launch success rate demonstrated an upward trend from 2013 to 2020.
2. Orbits such as ES-L1, GEO, HEO, SSO, and VLEO exhibited the highest success rates.
3. KSC LC-39A boasted the highest number of successful launches among all the launch sites.
4. For this specific task, the Decision tree classifier stands out as the most suitable machine learning algorithm over all models.

Thank you!