IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

María de los Ángeles Naranjo Muñoz
7 October 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection with API and Web Scrapping
  - Data Wrangling
  - EDA SQL
  - Interactive Maps Launch Sites (Folium)
  - Dashboard (Plotly Dash)
  - Predictive Analysis
- Summary of all results
  - EDA results
  - Interactive Analytics
  - Predictive Analysis

Introduction

This project is part of the IBM Data Science Capstone.

Goal → predict whether the first stage of the Falcon 9 rocket will successfully land. This is crucial because SpaceX can significantly reduce launch costs by reusing the first stage.

To achieve this, we will explore the following questions:

- Can we accurately predict the landing outcome of Falcon 9's first stage?
- What are the most influential factors that impact the success of the landing (e.g., payload mass, orbit type)?
- How can this prediction help other companies compete against SpaceX?

Section 1

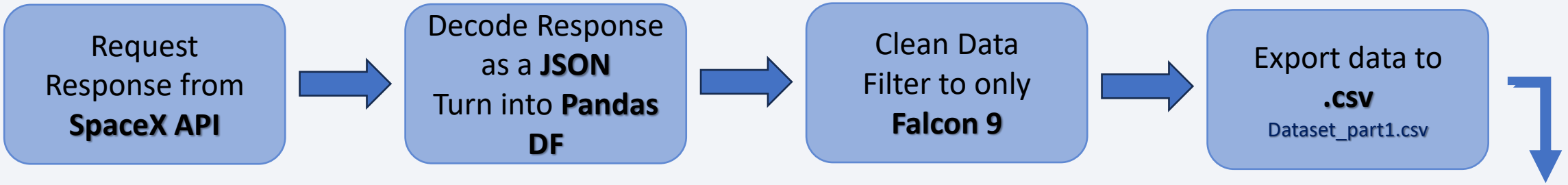
Methodology

Methodology

Executive Summary

- **Data Collection:** Gathered data through the SpaceX API and web scraping techniques.
- **Data Preparation:** Performed data wrangling, including filtering and handling missing values. Applied One Hot Encoding to process categorical data.
- **Exploratory Data Analysis (EDA):**Utilized visualization tools and SQL for in-depth data exploration.
- **Interactive Visual Analytics:** Created interactive visualizations using Folium and Plotly Dash.
- **Predictive Analysis:** Built and evaluated classification models to predict outcomes.
- **Model Optimization:** Tuned and validated models to improve accuracy and performance.

Data Collection : SpaceX API



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	FlightNum	Date	BoosterVe	PayloadMa	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPa	Block	ReusedCo	Serial	Longitude	Latitude
2	1	04/06/2010	Falcon 9	6123.548	LEO	CCSFS SLC	None Non	1	FALSE	FALSE	FALSE		1	0	B0003	-80.5774	28.56186
3	2	22/05/2012	Falcon 9	525	LEO	CCSFS SLC	None Non	1	FALSE	FALSE	FALSE		1	0	B0005	-80.5774	28.56186
4	3	01/03/2013	Falcon 9	677	ISS	CCSFS SLC	None Non	1	FALSE	FALSE	FALSE		1	0	B0007	-80.5774	28.56186
5	4	29/09/2013	Falcon 9	500	PO	VAFB SLC 4	False Ocea	1	FALSE	FALSE	FALSE		1	0	B1003	-120.611	34.63209
6	5	03/12/2013	Falcon 9	3170	GTO	CCSFS SLC	None Non	1	FALSE	FALSE	FALSE		1	0	B1004	-80.5774	28.56186
7	6	06/01/2014	Falcon 9	3325	GTO	CCSFS SLC	None Non	1	FALSE	FALSE	FALSE		1	0	B1005	-80.5774	28.56186
8	7	18/04/2014	Falcon 9	2296	ISS	CCSFS SLC	True Ocea	1	FALSE	FALSE	TRUE		1	0	B1006	-80.5774	28.56186
9	8	14/07/2014	Falcon 9	1316	LEO	CCSFS SLC	True Ocea	1	FALSE	FALSE	TRUE		1	0	B1007	-80.5774	28.56186
10	9	05/08/2014	Falcon 9	4535	GTO	CCSFS SLC	None Non	1	FALSE	FALSE	FALSE		1	0	B1008	-80.5774	28.56186

Click Here Access Data Collection Code

Data Collection : Scraping

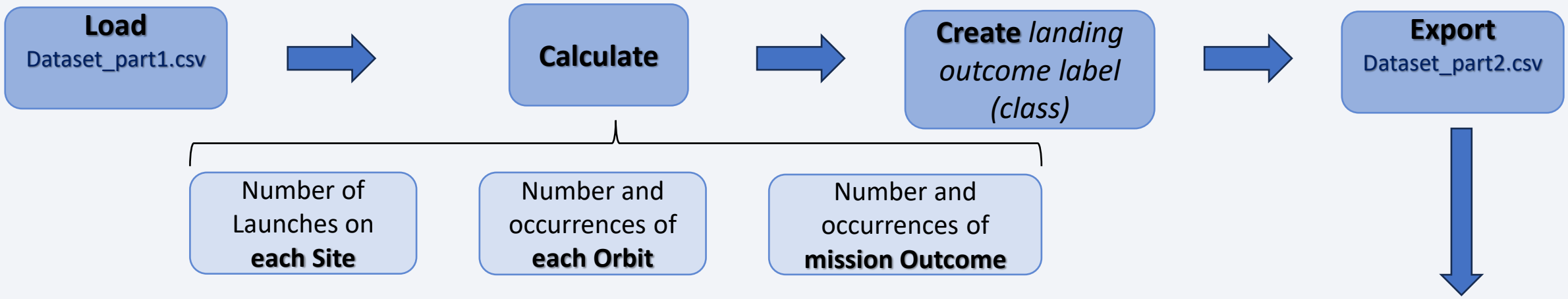
Perform **HTTP GET** to request Falcon 9 Launch → Extract all variables from **HTML Table** → Create DF Export to **.csv**

Spacex_web_scraped.csv

	A	B	C	D	E	F	G	H	I	J	K
1	Flight No.	Launch site	Payload	Payload m	Orbit	Customer	Launch ou	Version Bc	Booster la	Date	Time
2	1	CCAFS	Dragon Sp	0	LEO	SpaceX	Success	F9 v1.07B(	Failure	04-Jun-10	18:45
3	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.07B(	Failure	08-Dec-10	15:43
4	3	CCAFS	Dragon	525 kg	LEO	NASA (COT	Success	F9 v1.07B(	No attemp	22-May-12	07:44
5	4	CCAFS	SpaceX CR	4,700 kg	LEO	NASA (CRS	Success	F9 v1.07B(	No attemp	08-Oct-12	00:35
6	5	CCAFS	SpaceX CR	4,877 kg	LEO	NASA (CRS	Success	F9 v1.07B(	No attemp	01-Mar-13	15:10
7	6	VAFB	CASSIOPE	500 kg	Polar orbit	MDA	Success	F9 v1.17B1	Uncontroll	29-Sep-13	16:00
8	7	CCAFS	SES-8	3,170 kg	GTO	SES	Success	F9 v1.1	No attemp	03-Dec-13	22:41
9	8	CCAFS	Thaicom 6	3,325 kg	GTO	Thaicom	Success	F9 v1.1	No attemp	06-Jan-14	22:06
10	9	Cape Cana	SpaceX CR	2,296 kg	LEO	NASA (CRS	Success	F9 v1.1	Controlled	18-Apr-14	19:25

Click Here Access Data Collection Code(Scraping)

Data Wrangling



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	FlightNum	Date	BoosterVe	PayloadMa	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPa	Block	ReusedCo	Serial	Longitude	Latitude	Class
2	1	04/06/2010	Falcon 9	6104.959	LEO	CCAFS SLC	None Non	1	FALSE	FALSE	FALSE		1	0	B0003	-80.5774	28.56186	0
3	2	22/05/2012	Falcon 9	525	LEO	CCAFS SLC	None Non	1	FALSE	FALSE	FALSE		1	0	B0005	-80.5774	28.56186	0
4	3	01/03/2013	Falcon 9	677	ISS	CCAFS SLC	None Non	1	FALSE	FALSE	FALSE		1	0	B0007	-80.5774	28.56186	0
5	4	29/09/2013	Falcon 9	500	PO	VAFB SLC 4	False Ocea	1	FALSE	FALSE	FALSE		1	0	B1003	-120.611	34.63209	0
6	5	03/12/2013	Falcon 9	3170	GTO	CCAFS SLC	None Non	1	FALSE	FALSE	FALSE		1	0	B1004	-80.5774	28.56186	0
7	6	06/01/2014	Falcon 9	3325	GTO	CCAFS SLC	None Non	1	FALSE	FALSE	FALSE		1	0	B1005	-80.5774	28.56186	0
8	7	18/04/2014	Falcon 9	2296	ISS	CCAFS SLC	True Ocea	1	FALSE	FALSE	TRUE		1	0	B1006	-80.5774	28.56186	1
9	8	14/07/2014	Falcon 9	1316	LEO	CCAFS SLC	True Ocea	1	FALSE	FALSE	TRUE		1	0	B1007	-80.5774	28.56186	1
10	9	05/08/2014	Falcon 9	4535	GTO	CCAFS SLC	None Non	1	FALSE	FALSE	FALSE		1	0	B1008	-80.5774	28.56186	0

Click Here Access Data Wrangling

EDA with Data Visualization

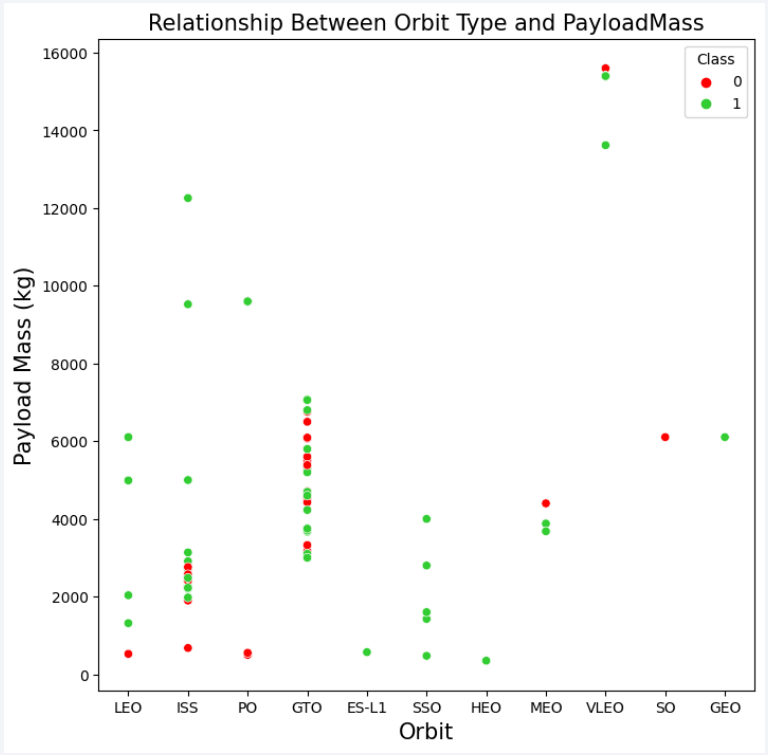
Load
Dataset_part2.csv

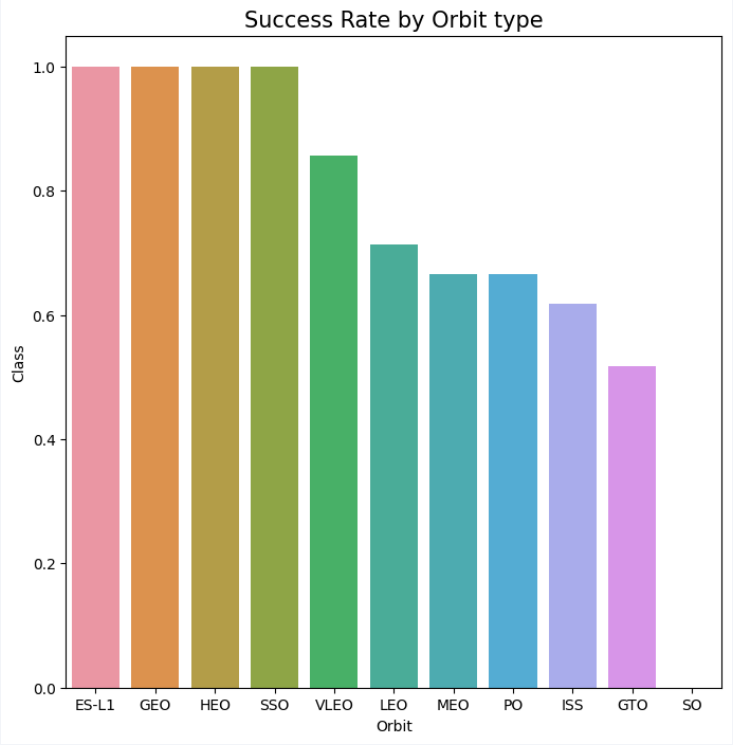
Visualize relationship

- Flight Number vs Launch Site
- Payload vs Launch Site
- Success Rate vs Orbit Type
- Flight Number vs Orbit Type
- Payloadvs Orbit Type

Feature Engineer
Categorical var. to dummy
Cast numerical var. to float64

Convert to .csv
Dataset_part3.csv





	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	FlightNum	PayloadMa	Flights	GridFins	Reused	Legs	Block	ReusedCo	Orbit_ES-L	Orbit_GEO	Orbit_GTO	Orbit_HEO	Orbit_ISS	Orbit_LEO	Orbit_MEO	Orbit_PO	Orbit_SO	Orbit_SSO	Orbit_VLE	LaunchSite	LaunchSite	LaunchSite	LandingPa
2	1	6104.959	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
3	2	525	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
4	3	677	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
5	4	500	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
6	5	3170	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
7	6	3325	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
8	7	2296	1	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
9	8	1316	1	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
10	9	4535	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0

Click Here Access EDA Visualization

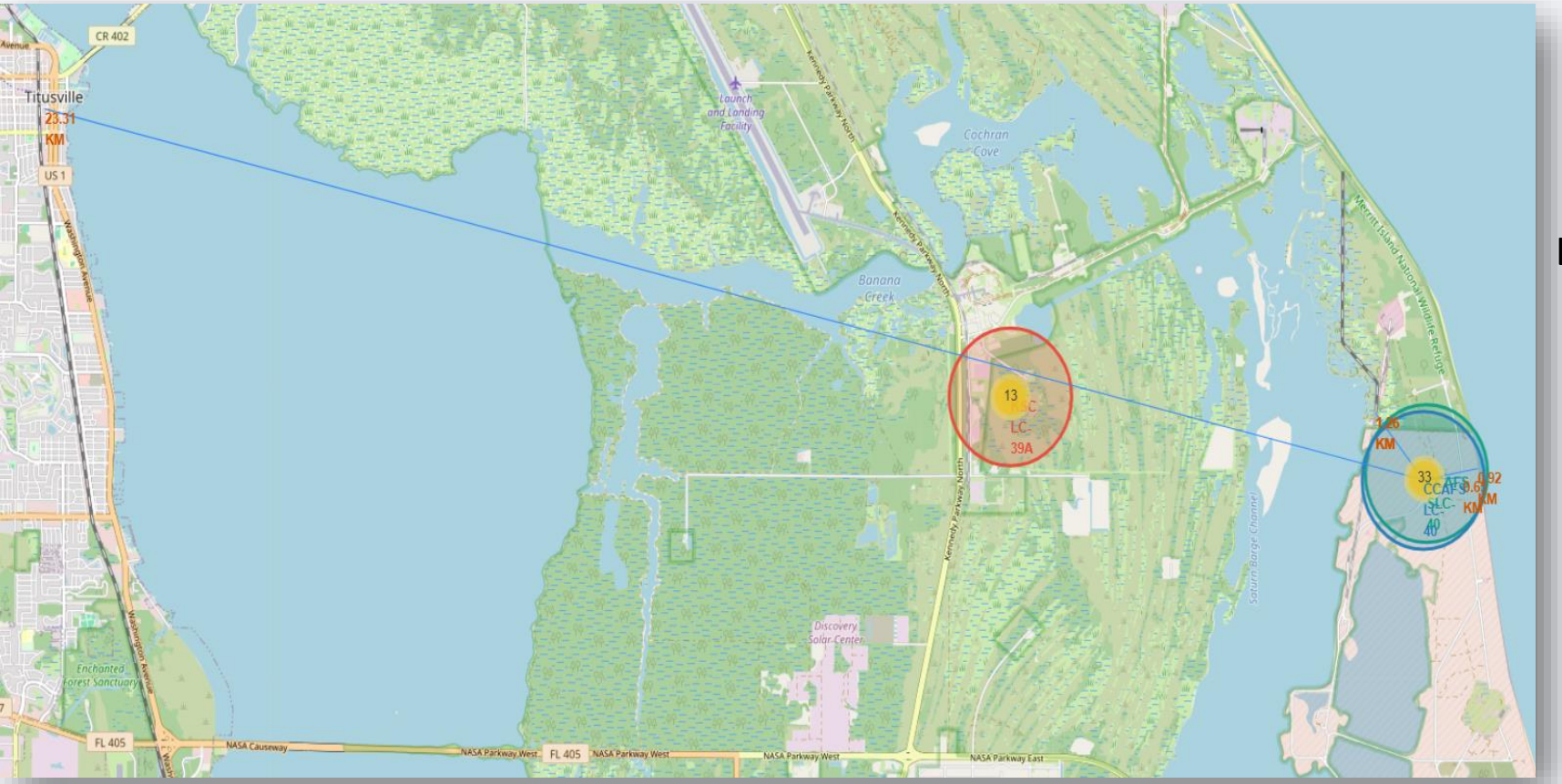
EDA with SQL

1. Display **unique names** of **Launch Sites**
2. Display records where **Launch Site** start '**CCA**'
3. **Total Payload Mass** carried by booster launched by **NASA (CRS)**
4. **AVG Payload Mass** by **F9 v1.1**
5. **First Successful landing** (*ground pad*)
6. Boosters with **success** in **Drone Ship,** Payload between **4000 – 6000 Kg**
7. Total number of **successful** and **failure** **mission outcomes**
8. **Booster_Version** with **Max. Payload Mass (Kg)**
9. **Month names** of failure Landing Outcomes in **Drone Ship** year **2015**
10. **Landing Outcome** between **2010-06-04** and **2017-03-20**


Click Here Access EDA SQL

Build an Interactive Map with Folium

Mark all **Launch Sites** on Map → Mark **success / failed launches** for each sites → Calculate distances **Launch Site vs proximity**

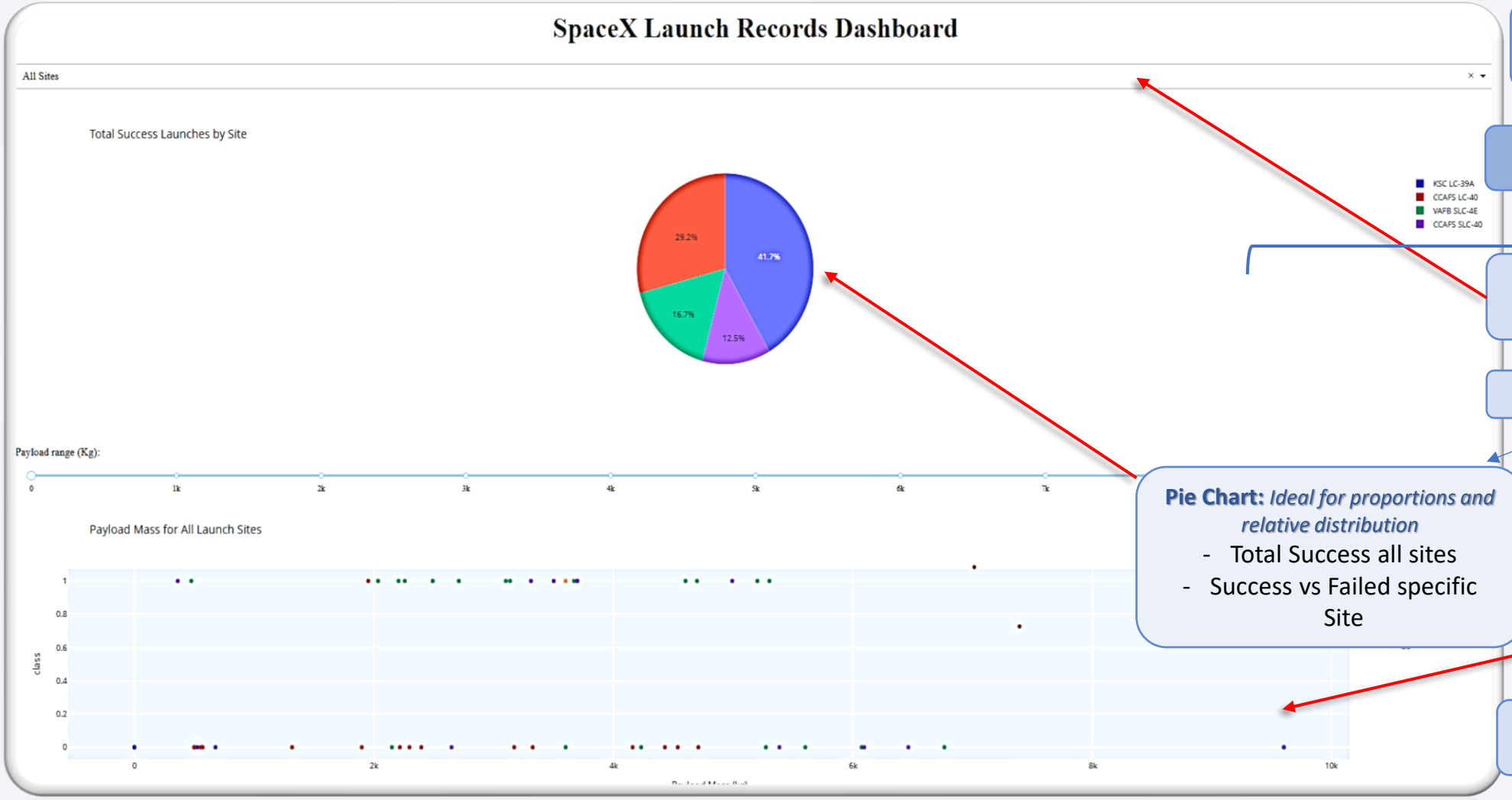


Launch Site CCAFS LC-40:

- Proximity Railway (1.26 km)
- Proximity Highway (0.65 km)
- Proximity Coastline (0.92 km)
- Proximity City (Titusville – 23.31 km)

Click Here Access Interactive Map

Build a Dashboard with Plotly Dash



Create **Dash APP**

App Layout

Dropdown: Enable Launch Site Selection

Graphs:

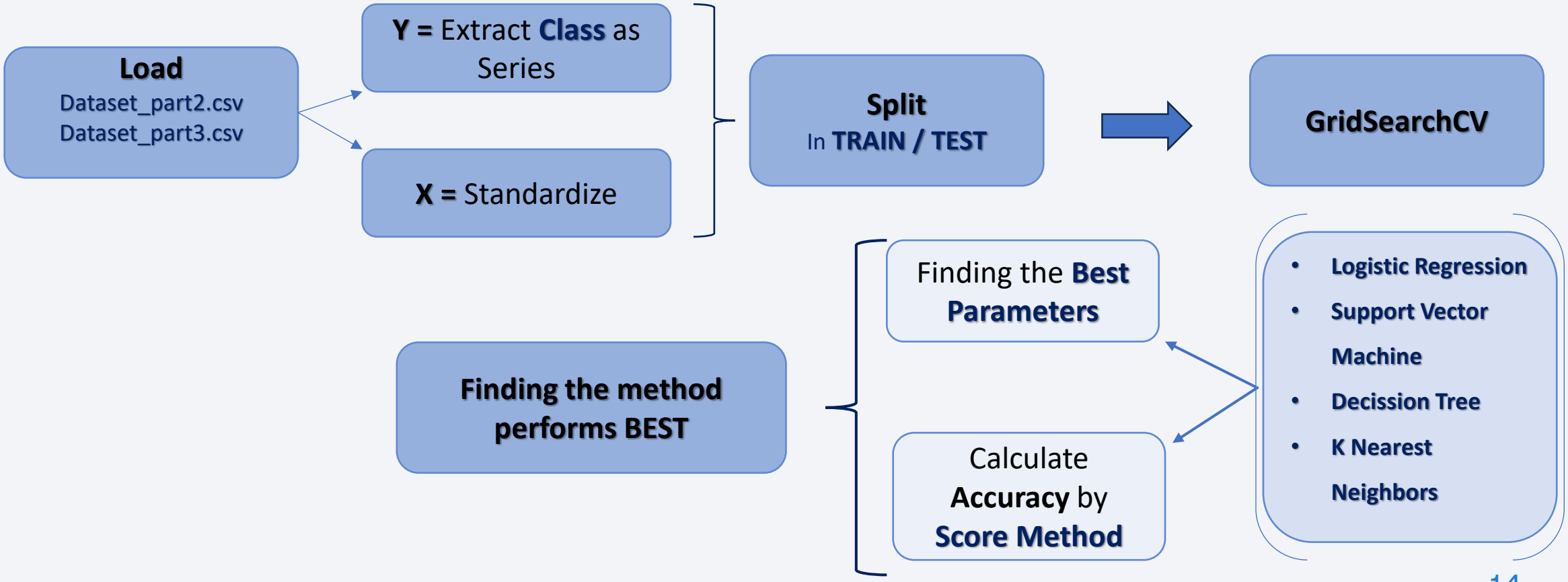
Pie Chart: *Ideal for proportions and relative distribution*
- Total Success all sites
- Success vs Failed specific Site

Scatter Chart: *best method for non linear pattern*
- Correlation Payload vs Launch Sites

RangeSlider: Select Payload Range

Click Here Access Dashboard Code

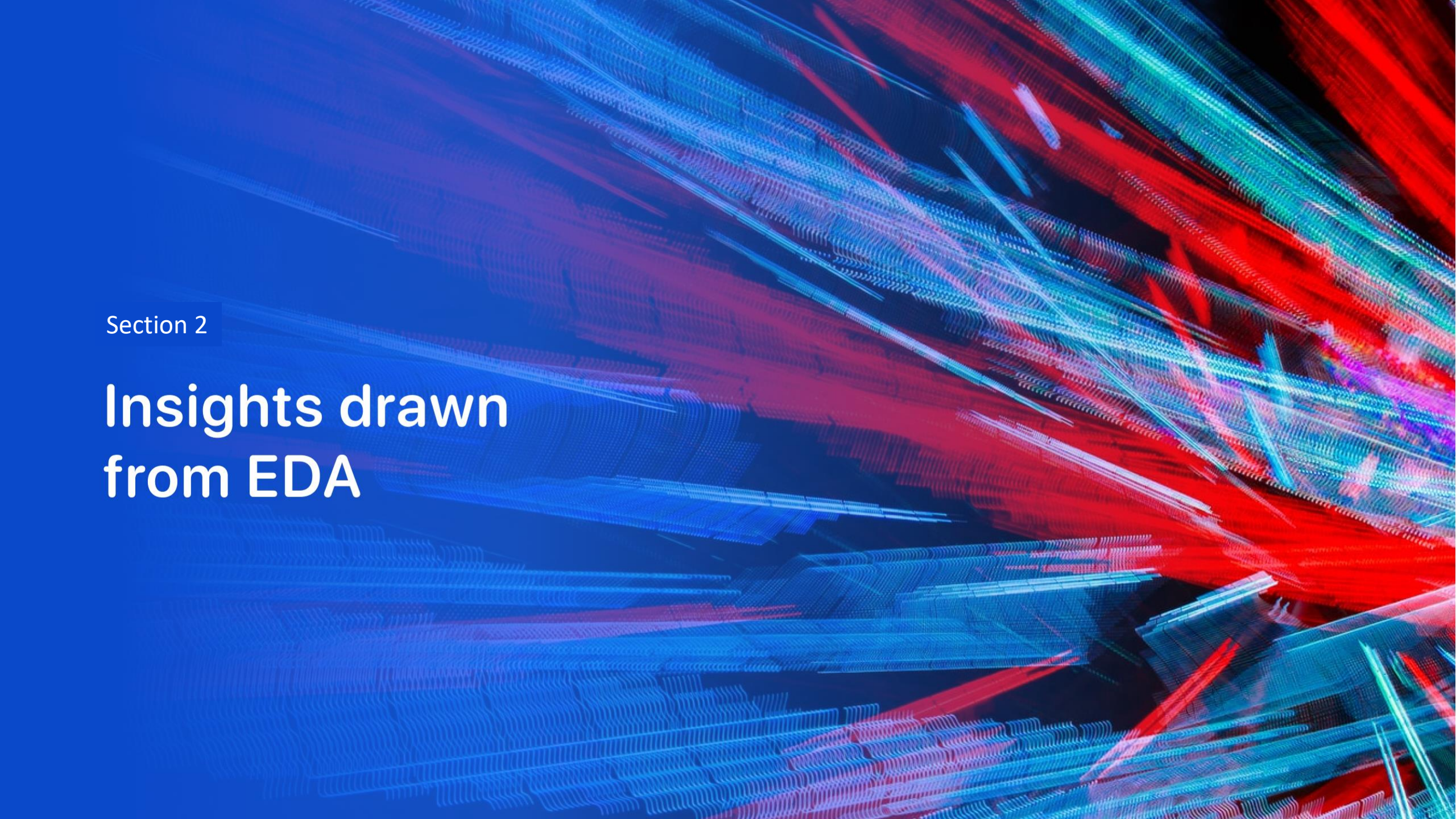
Predictive Analysis (Classification)



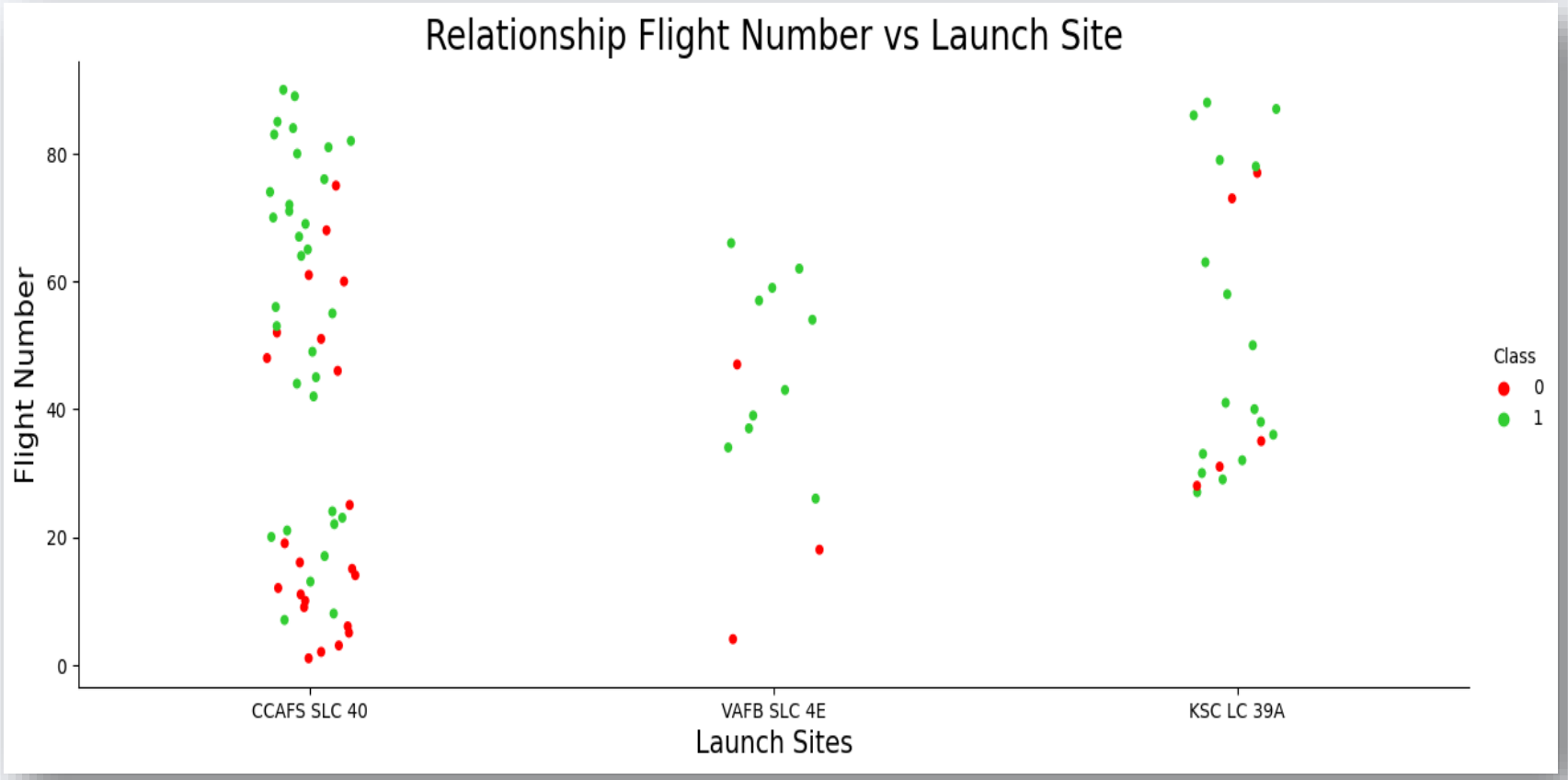
Click Here Access Prediction

Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



CCAFS SLC 40: Consistent launches with a mix of successes and failures, but a trend toward better success rates as the flight numbers increase.

VAFB SLC 4E: Fewer launches with a higher proportion of failures, which may indicate specific challenges at this site.

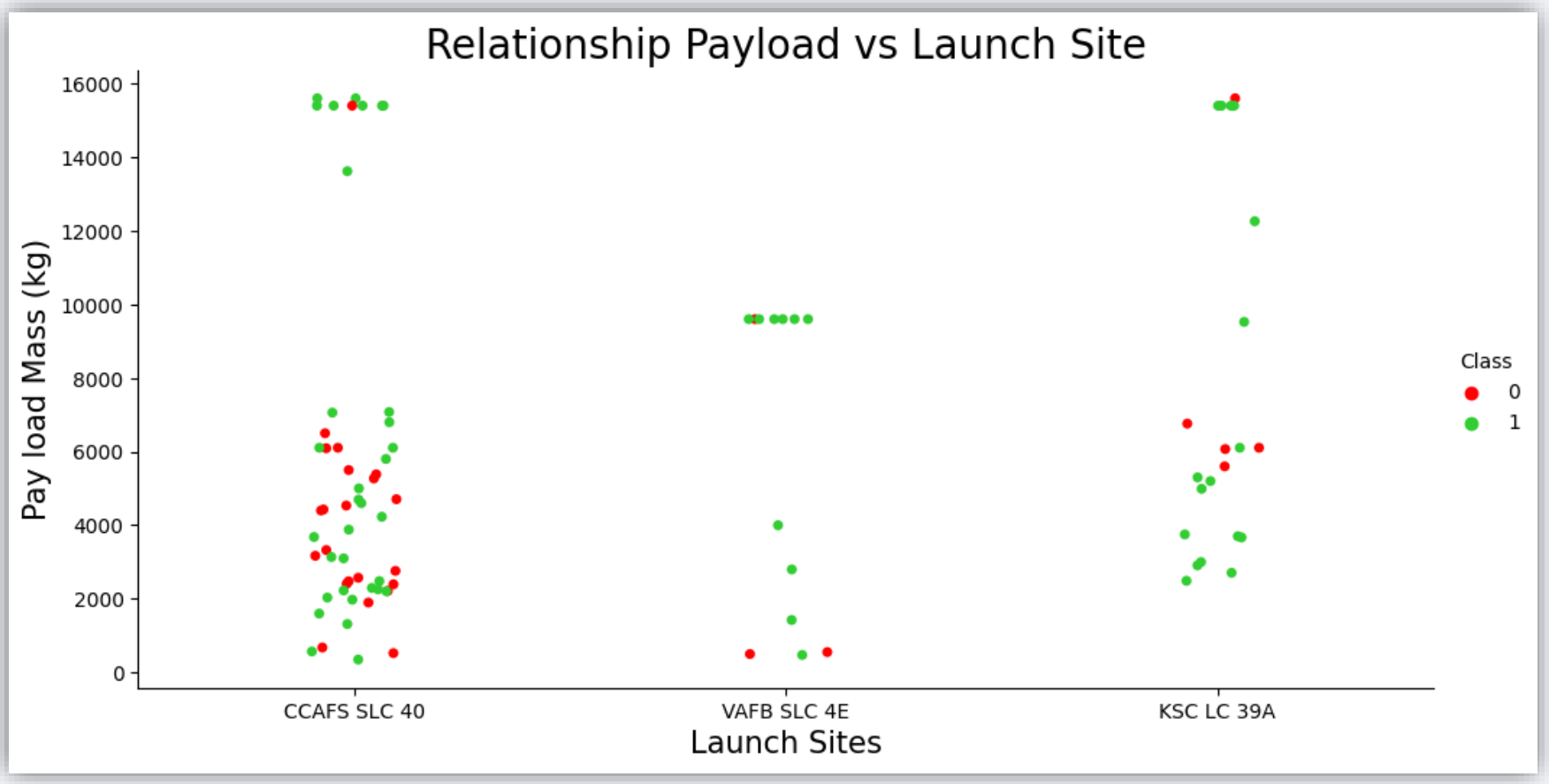
KSC LC 39A: Follows a pattern similar to CCAFS, showing improved success with higher flight numbers.

This suggests that as SpaceX launches more rockets, particularly from key locations, their ability to land the first stage improves

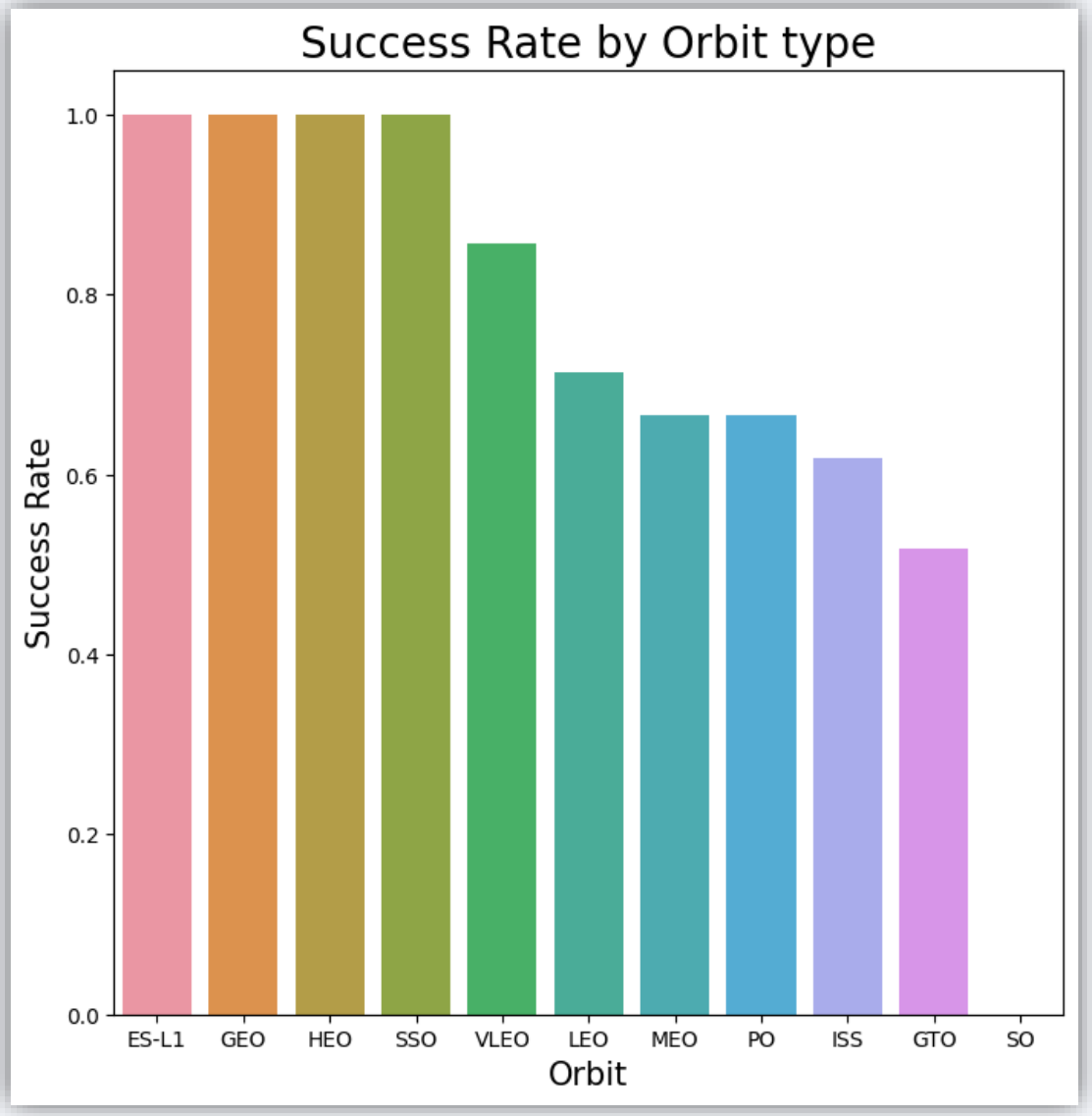
Payload vs. Launch Site

CCAFS SLC 40 | KSC LC 39A :
Handle heavier payloads (exceeding 10,000 kg),having a mix of landing outcomes, indicating that heavier payloads don't necessarily lead to landing success.

VAFB SLC 4E:
Deals with lighter payloads, not exceed 10000 kg



Success Rate vs. Orbit Type

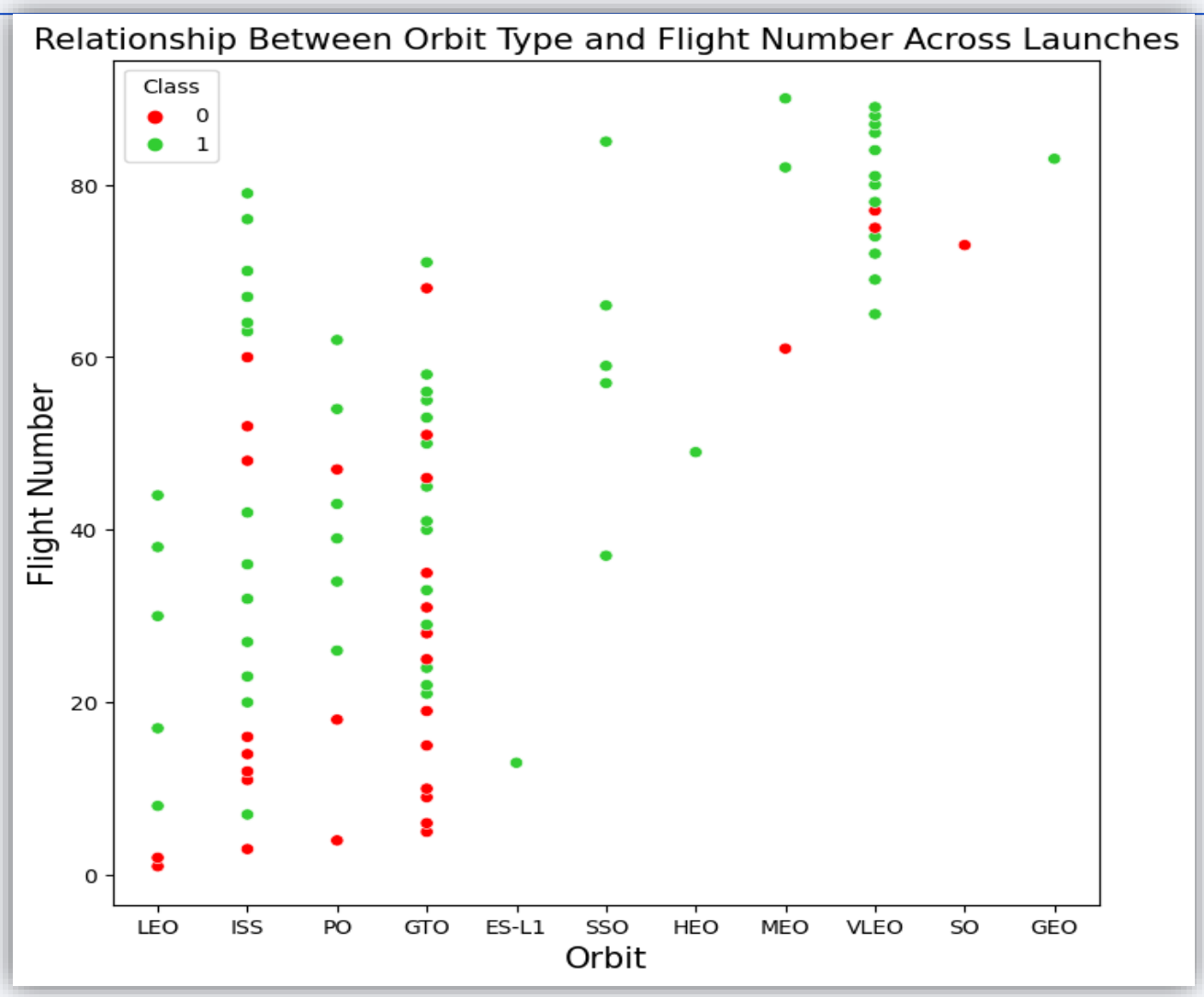


ES-L1, GEO, HEO, SSO:
Perfect success rate (1.00)
Indicates → **high reliability** for launches

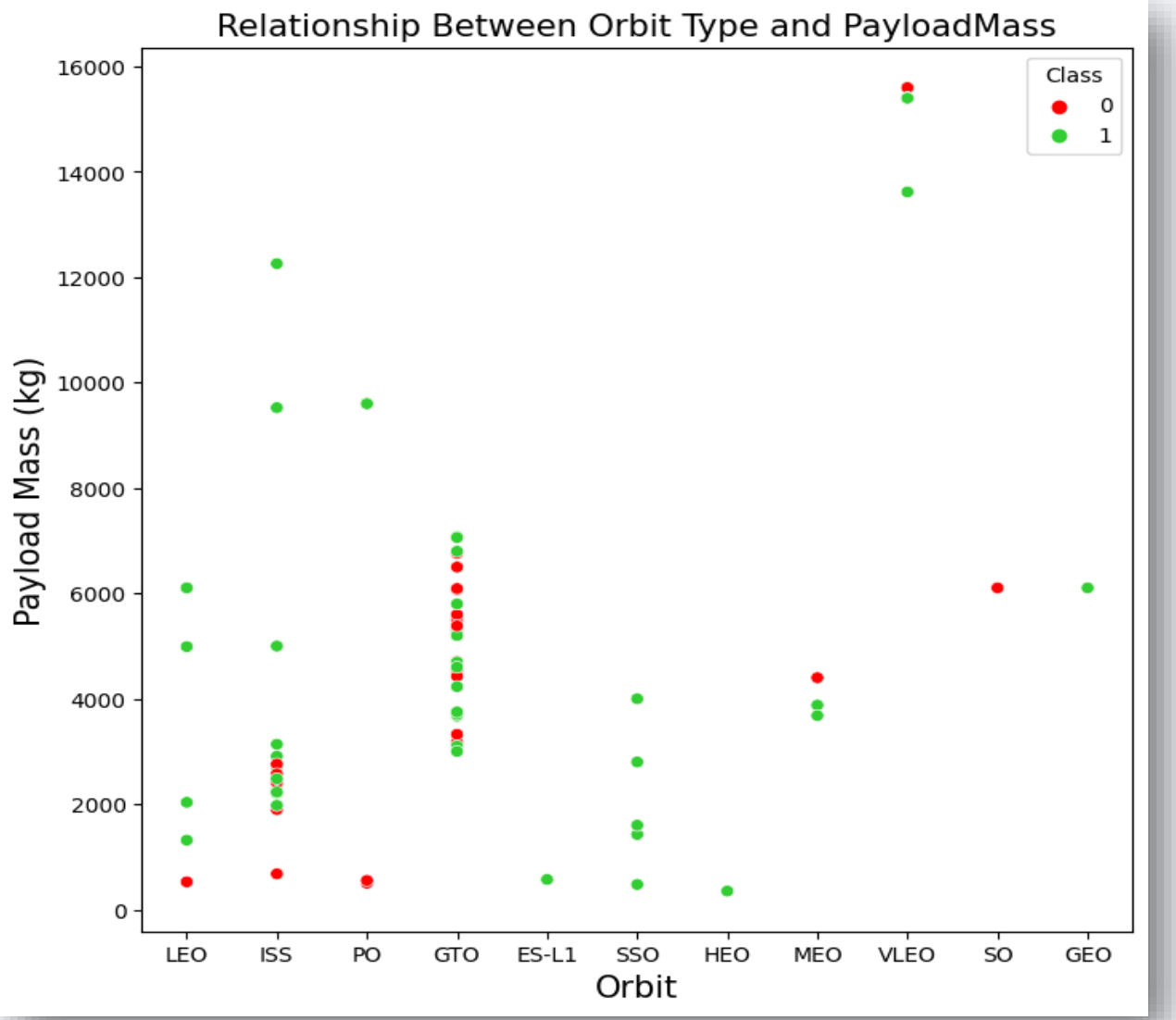
SO:
No recorded success rate (0.00)
Indicates --> potential lack of data or no successful launches for this orbit.

Flight Number vs. Orbit Type

Noticeable Trend:
higher flight number correlates with an increased success rate.
(*LEO, ISS, VLEO)*



Payload vs. Orbit Type

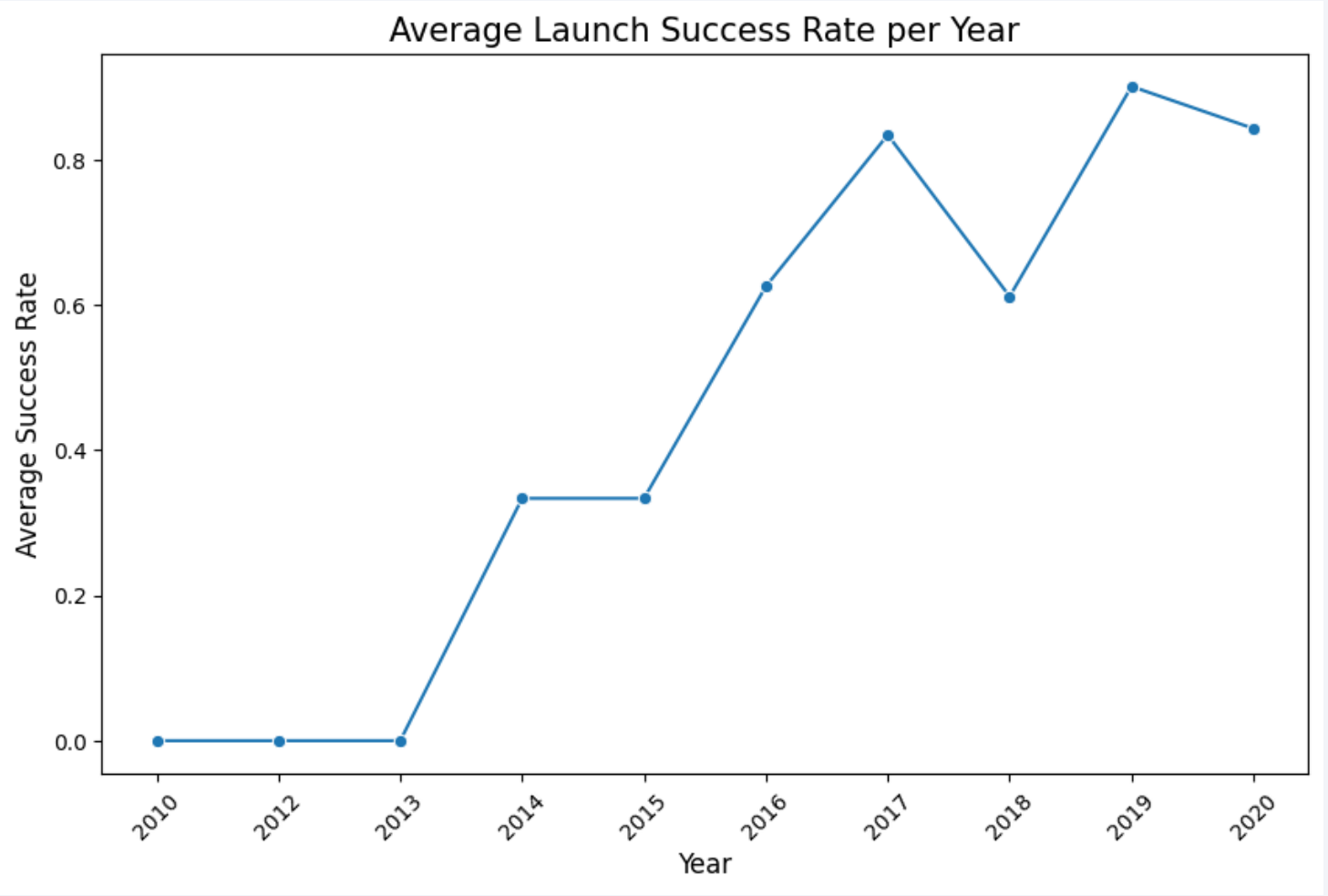


Noticeable Trend:
higher success rate with heavy payloads. (*LEO,PO, ISS)*

GTO displays a mixed successes and failures, making it harder the predictability of missions in this orbit.

Launch Success Yearly Trend

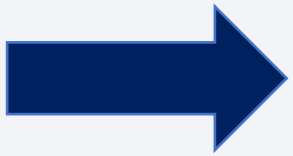
The success rate since 2013 kept **increasing** till 2017 (stable in 2014) and after 2015 it started increasing.



All Launch Site Names

SQL QUERY

```
SELECT DISTINCT Launch_Site
FROM SPACEXTABLE
```



RESULT

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

EXPLANATION:

Using **DISTINCT** will only show **UNIQUE** values in *Launch_Site* column from table SPACEXTABLE

Launch Site Names Begin with 'CCA'

SQL QUERY

```
SELECT *
FROM SPACEXTABLE
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5
```

RESULT

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

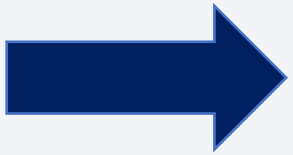
EXPLANATION:

Using **WHERE** Launch_Site **LIKE** **'CCA%'** Filters results to *only include* launch sites that begin with "**CCA**".
LIMIT 5 Restricts the output to the first 5 records found

Total Payload Mass

SQL QUERY

```
SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload, Customer
 FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)'
```



RESULT

Total_Payload	Customer
45596	NASA (CRS)

EXPLANATION:

WHERE Customer = **'NASA (CRS)'** Filters results to *only include* Customer = **'NASA (CRS)'**

SUM() sums the total in the column *PAYLOAD_MASS_KG_* for the filtered records.

Average Payload Mass by F9 v1.1

SQL QUERY

```
SELECT AVG(PAYLOAD_MASS__KG_) AS 'AVG Payload V1.1'
FROM SPACEXTABLE
WHERE Booster_Version = 'F9 v1.1'
```

RESULT

AVG Payload V1.1
2928.4

EXPLANATION:

WHERE Booster_Version = **'F9 v1.1'** Filters results to *only include* that Booster Version **(F9 v1.1)**

AVG() calculates the average *PAYLOAD_MASS_KG_* for the filtered records

First Successful Ground Landing Date

SQL QUERY

```
SELECT MIN(DATE) AS First_Success_Date, Landing_Outcome
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (ground pad)'
```

RESULT

First_Success_Date	Landing_Outcome
2015-12-22	Success (ground pad)

EXPLANATION:

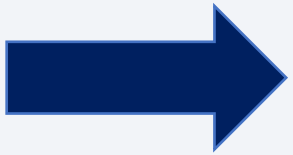
WHERE Landing_Outcome = *'Success (ground pad)'* Filters results to ***only include*** that Landing_Outcome

MIN() show the minimum ***Date*** for the filtered records

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL QUERY

```
SELECT Booster_Version, PAYLOAD_MASS_KG_
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)'
AND PAYLOAD_MASS_KG_ BETWEEN 4000 AND 6000
ORDER BY PAYLOAD_MASS_KG_
```



RESULT

Booster_Version	PAYLOAD_MASS_KG_
F9 FT B1026	4600
F9 FT B1022	4696
F9 FT B1031.2	5200
F9 FT B1021.2	5300

EXPLANATION:

WHERE Landing_Outcome = *'Success (drone ship)'* Filters results to ***only include*** that Landing_Outcome
AND PAYLOAD_MASS_KG_ **BETWEEN** 4000 **AND** 6000 Filters the results to ***show only*** payloads with a mass between 4000 kg and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

SQL QUERY

```
SELECT
    COUNT(CASE WHEN Mission_Outcome LIKE '%Success%' THEN 1 END) AS Count_Succesfull_Mission
    COUNT(CASE WHEN Mission_Outcome LIKE '%Failure%' THEN 1 END) AS Count_Failure_Mission
FROM SPACEXTABLE
```

RESULT

Count_Succesfull_Mission	Count_Failure_Mission
100	1

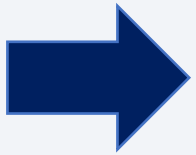
EXPLANATION:

COUNT(CASE WHEN... THEN 1 END). Count the rows where the Mission_Outcome column contains "Success" and "Failure"

Boosters Carried Maximum Payload

SQL QUERY

```
SELECT DISTINCT(Booster_Version), PAYLOAD_MASS_KG_
FROM SPACEXTABLE
WHERE PAYLOAD_MASS_KG_ = (
                        SELECT MAX(PAYLOAD_MASS_KG_)
                        FROM SPACEXTABLE
                                )
```



RESULT

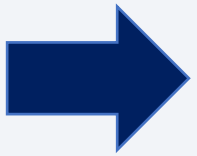
Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

EXPLANATION:

WHERE PAYLOAD_MASS_KG_ = **(SELECT MAX)** Filters results to *only show* the **maximum payload mass**

2015 Launch Records

```
SELECT
   CASE SUBSTR(DATE, 6, 2)
      WHEN '01' THEN 'January'
      WHEN '02' THEN 'February'
      WHEN '03' THEN 'March'
      WHEN '04' THEN 'April'
      WHEN '05' THEN 'May'
      WHEN '06' THEN 'June'
      WHEN '07' THEN 'July'
      WHEN '08' THEN 'August'
      WHEN '09' THEN 'September'
      WHEN '10' THEN 'October'
      WHEN '11' THEN 'November'
      WHEN '12' THEN 'December'
   END AS Month,
   Booster_Version,
   Launch_Site
FROM SPACEXTABLE
WHERE SUBSTR(DATE,0,5)  = '2015'
AND Landing_Outcome = 'Failure (drone ship)'
```



Month	Booster_Version	Launch_Site
January	F9 v1.1 B1012	CCAFS LC-40
April	F9 v1.1 B1015	CCAFS LC-40

EXPLANATION:

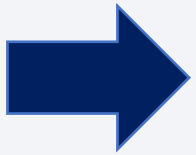
CASE SUBSTR(…) WHEN … THEN … checks the *month number* extracted from the **DATE** and returns the corresponding *month name*.

WHERE SUBSTR(DATE,0,5) = '2015' filters the results to include only records from the year 2015 **WHERE** there was a failure landing on a drone ship.

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

SQL QUERY

```
SELECT Landing_Outcome,
    COUNT(Landing_Outcome) AS Outcome_Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Outcome_Count DESC
```



RESULT

Landing_Outcome	Outcome_Count
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

EXPLANATION:

COUNT(Landing_Outcome**)** Counts the frequency of each landing outcome
WHERE Date within the range (**BETWEEN** '2010-06-04' **AND** '2017-03-20')
GROUP BY group the result by each unique Landing Outcome to calculate the count for each

Section 3

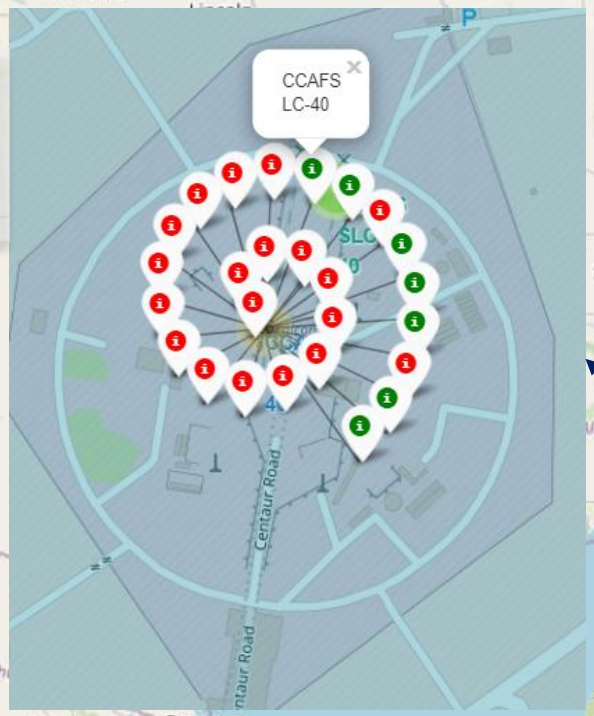
Launch Sites Proximities Analysis

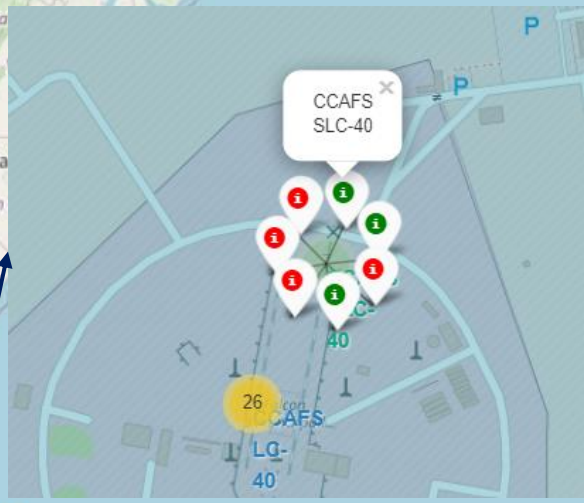
All Launch Sites

All of the observed launch sites are indeed in very close **proximity to the coast**, that **minimizes the risk** to human populations and structures in the event of a failure during launch.

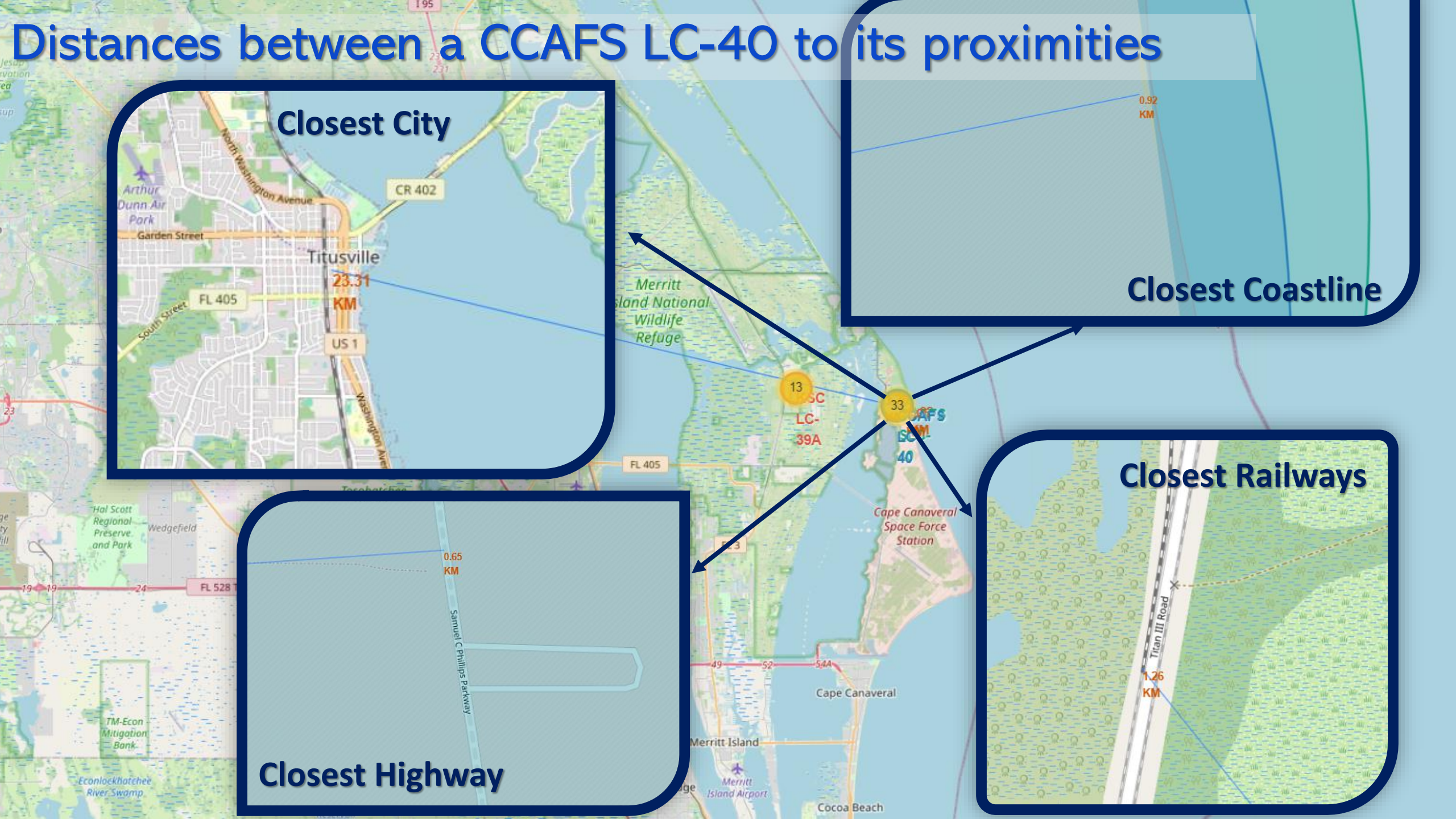
Success / Failed Launches

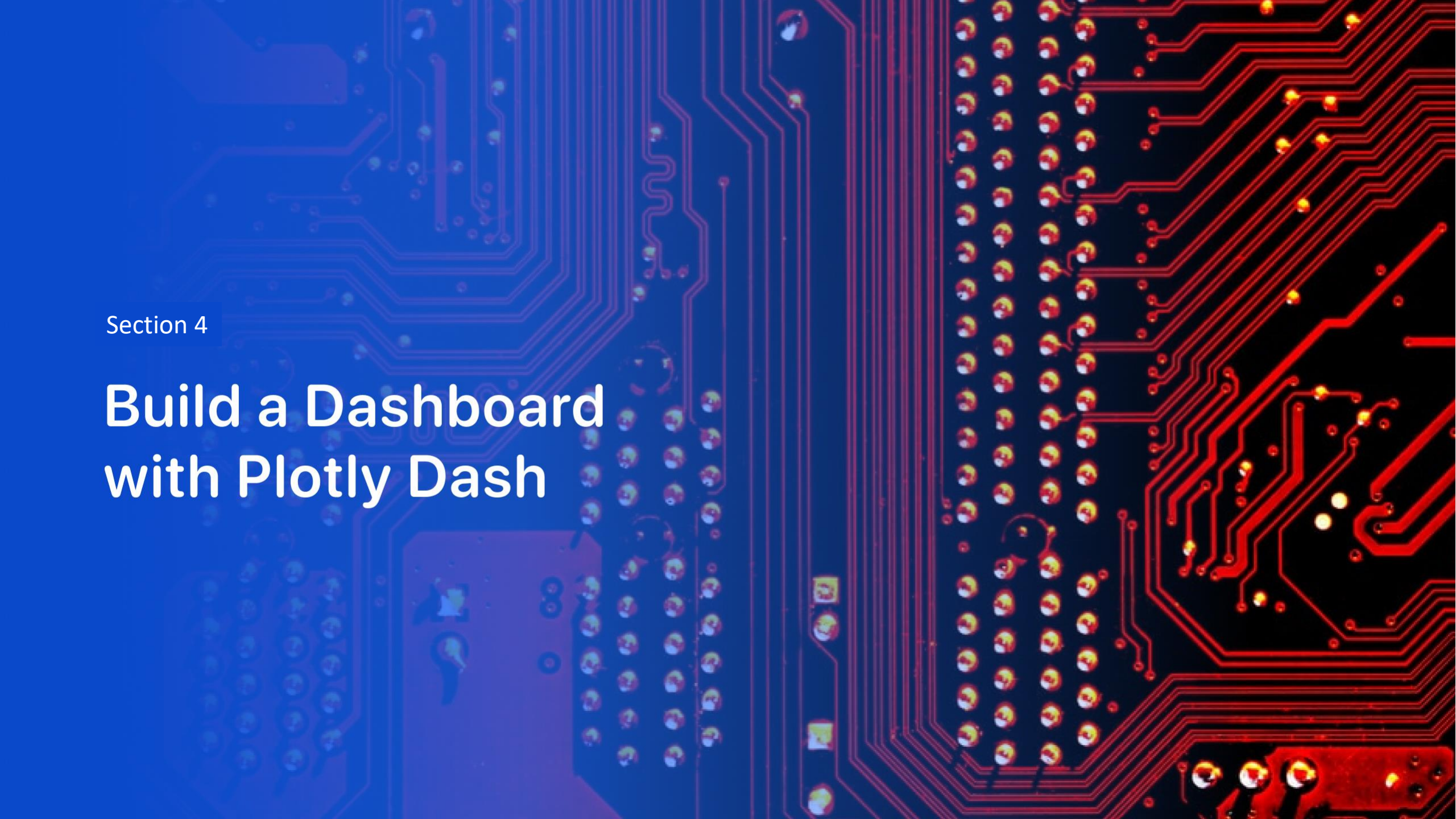
Green Markers → Successful Launches
Red Markers → Failures Launches





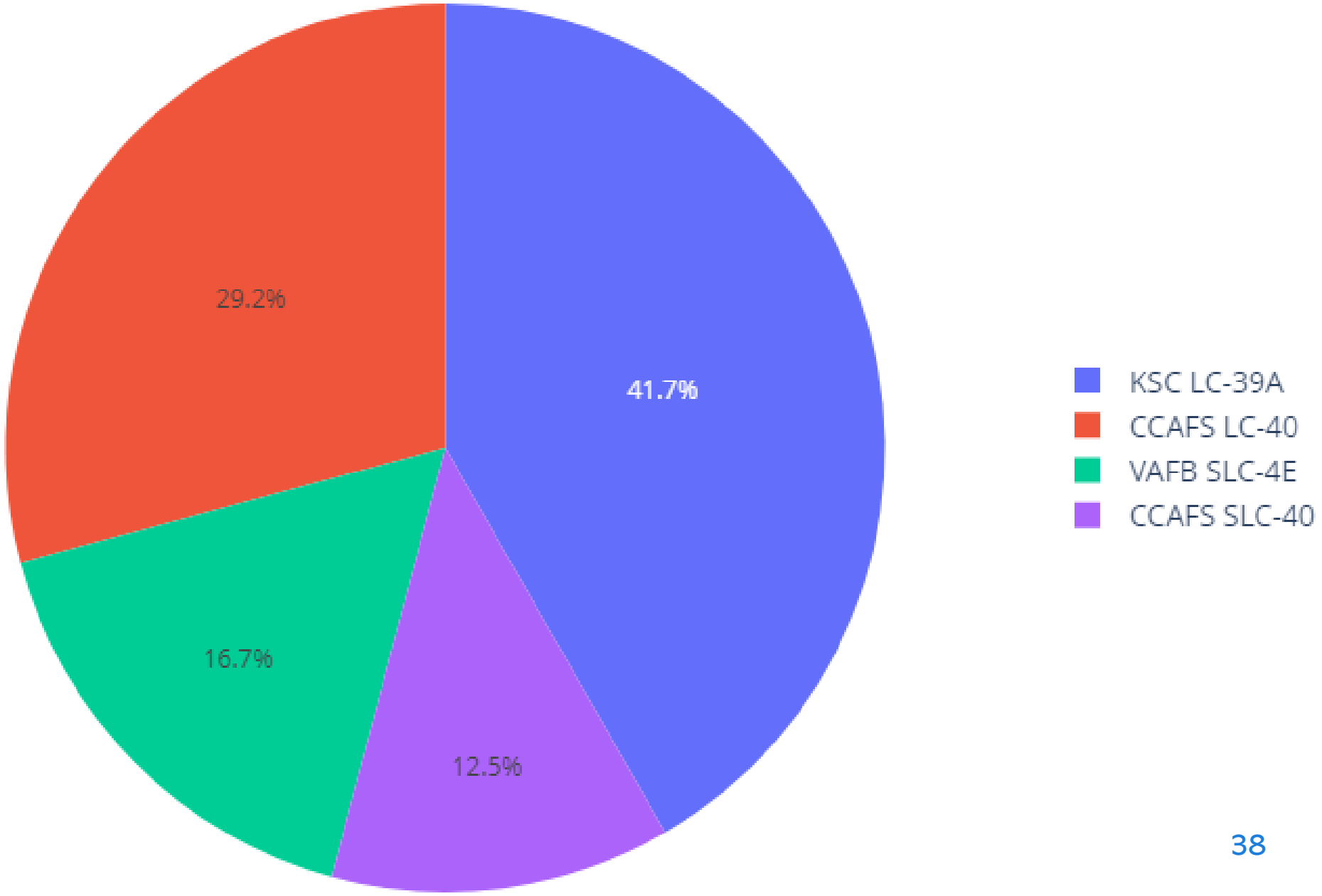
Distances between a CCAFS LC-40 to its proximities




Section 4
Build a Dashboard with Plotly Dash

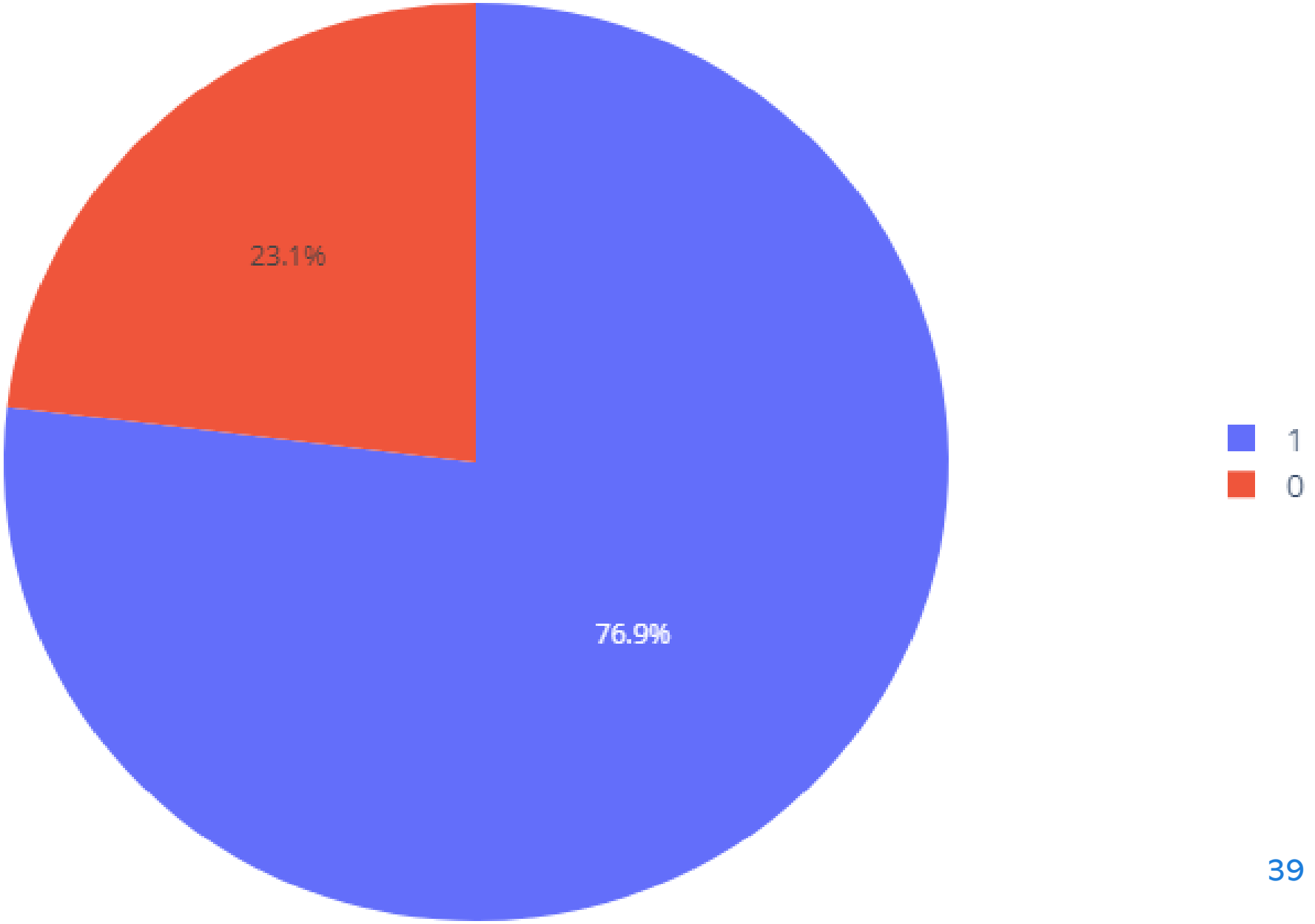
Total Successful Launches

KSC LC-39A recorded the highest number of successful launches whereas **CCAFS SLC-40** had the lowest



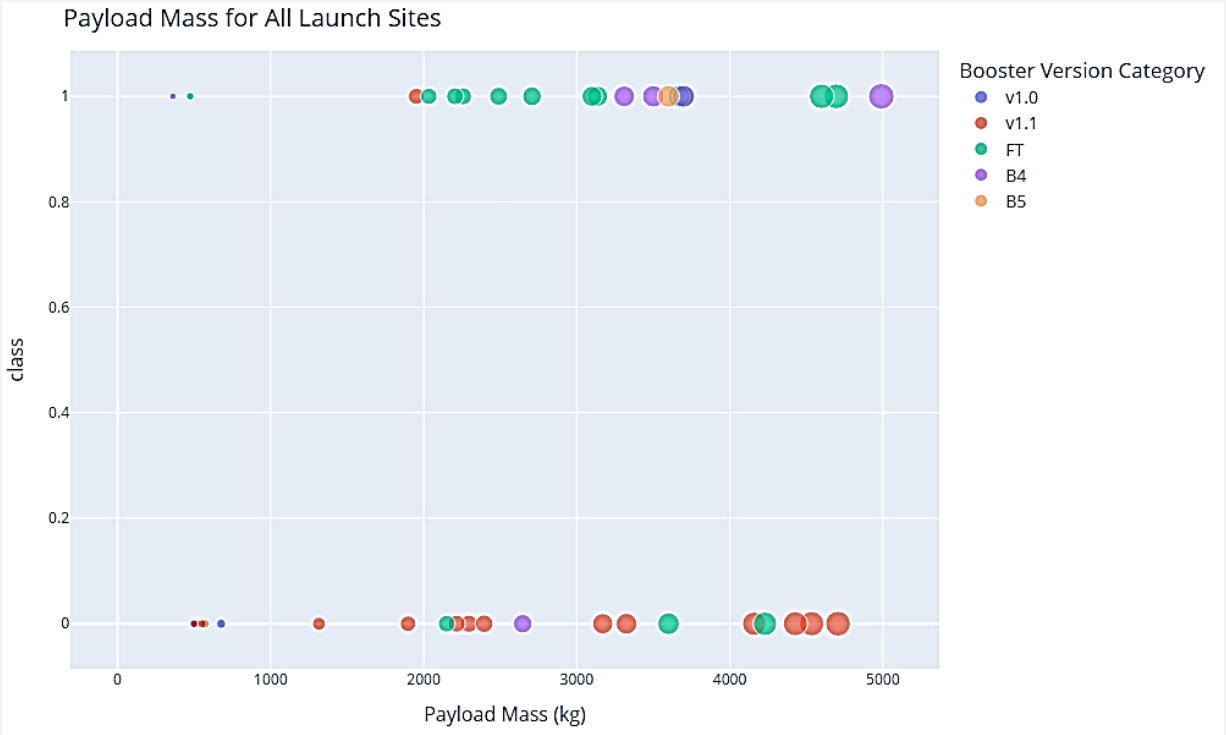
Total Successful Launches for KSC LC-39A

At **KSC LC-39A**, the *success rate* stood at **76.9%**, while the *failure rate* was recorded at **23.1%**.

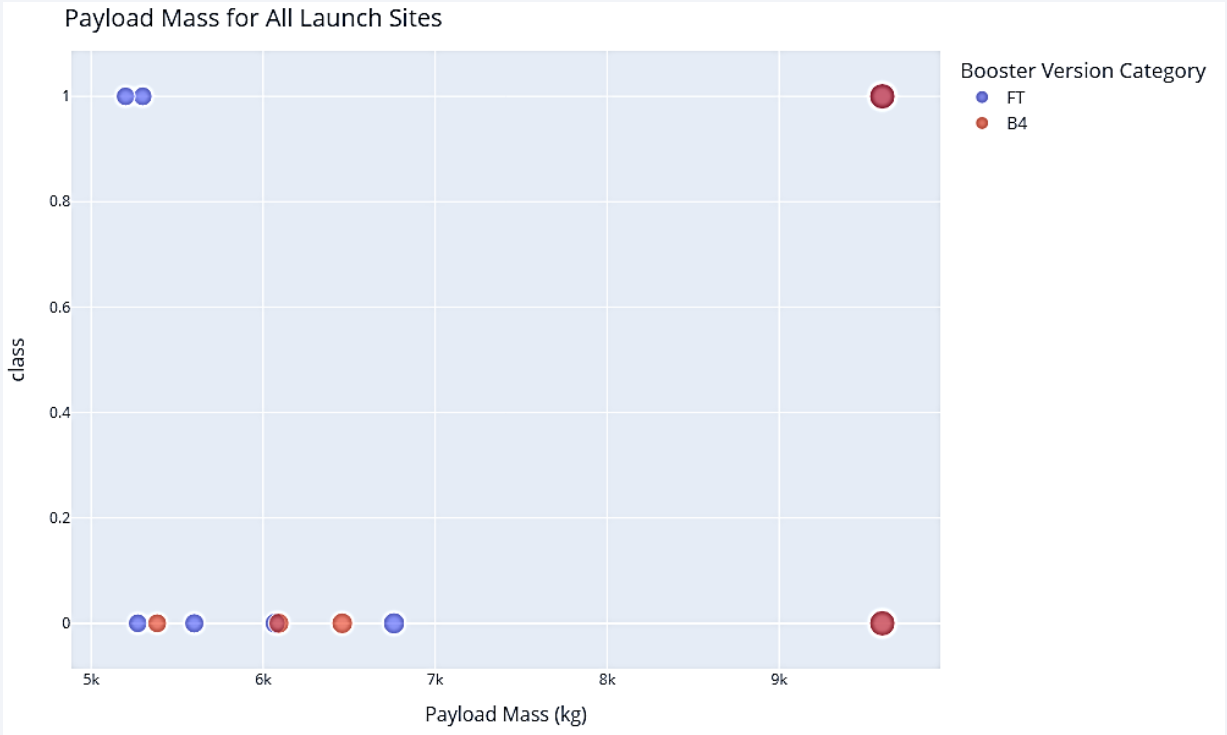


Payload Mass and Success

Lower Payload 0 Kg – 5000 Kg



Higher Payload 5000 Kg – 10000 Kg



- **Highest success rate** (***Class = 1***) is concentrate around payloads between 2000 Kg – 4000 Kg
- For Higher Payload, there are fewer launches in this payload range and the failures outnumber successes.
- The **FT booster** consistently performs well in both low and high payloads, with higher success rates.

Section 5

Predictive Analysis (Classification)

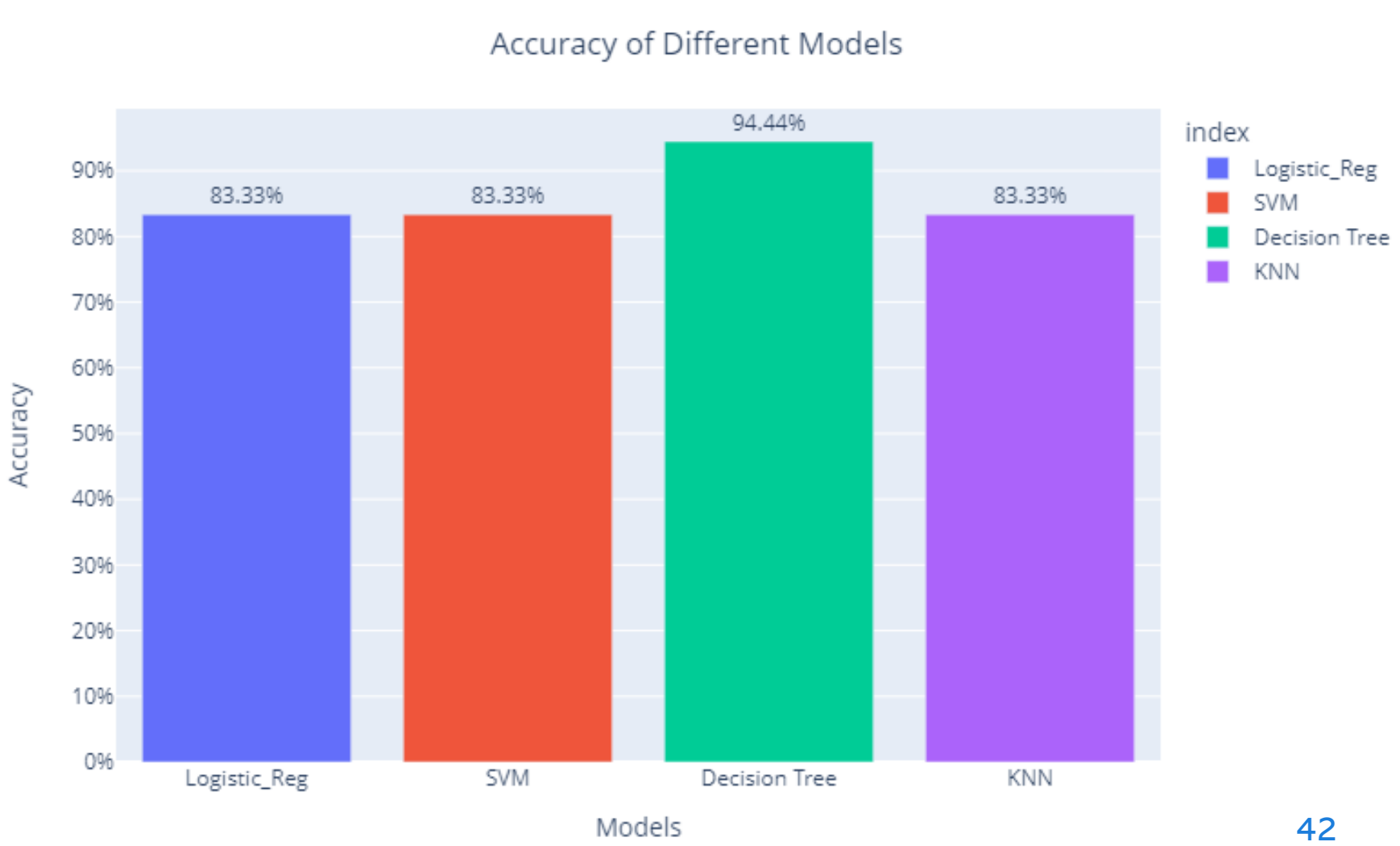
Classification Accuracy

The **Decision Tree** Model has the highest accuracy

The **best parameters**:

{'criterion': **'gini'**,
'max_depth': **2**,
'max_features': **'sqrt'**,
'min_samples_leaf': **4**,
'min_samples_split': **10**,
'splitter': **'best'**}

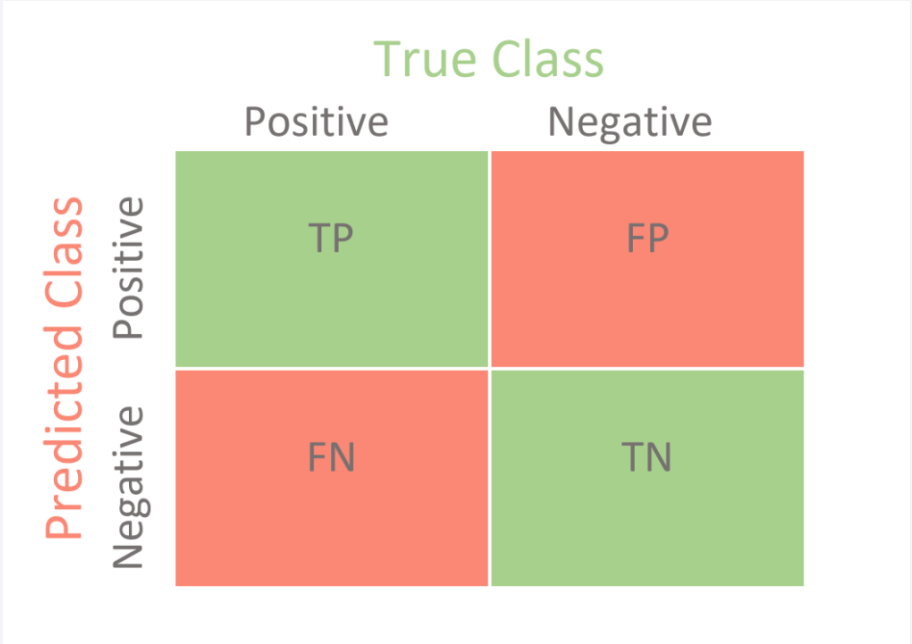
Achieved an **accuracy** of **94.44%** on the test set using X_test and Y_test.

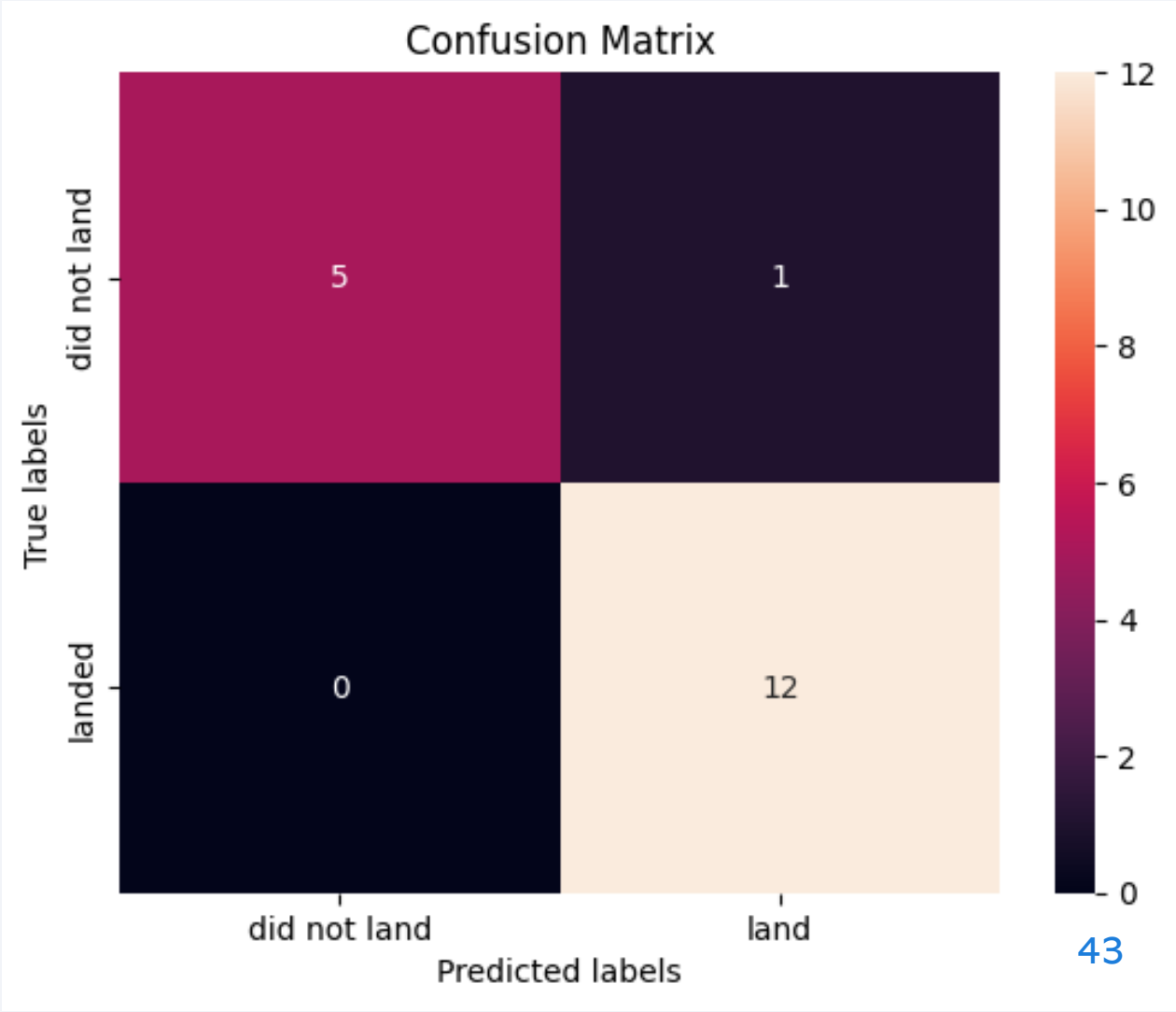


Confusion Matrix Decision Tree

Confusion matrix of decision tree (*model with highest accuracy*)

The decision tree showed **high accuracy**, correctly classifying 17 of 18 cases.
The presence of **false positives** (*Type I errors*) indicates room for improvement, as they reduce the reliability of the model.





Conclusions

- **Model Performance** → The Decision Tree Model is the best performing, achieving the highest accuracy of **94.4%**
- **Proximity Launches** → All launch sites are strategically located **near the coast** to minimize risks to both human populations and infrastructure in case of launch failures, highlighting the importance of safety in mission planning.
- **KSC LC-39A** → Has the highest success rate, making it the **most reliable** launch site
- **Orbits** → ES-L1, GEO, HEO, SSO have a perfect success rate **| High reliability**
- **Success rates** → Trend of **increasing** success rates over time, due to continuous improvements in technology, mission planning, and operational efficiency, indicates growing reliability as the company gains more experience.
- **Booster Version** → The **Falcon 9 FT** booster version consistently performs well across different payload categories

Future analyses could take a closer look at *the Falcon 9 FT booster* version, which has shown high reliability across different payload ranges, to further understand its performance in varying mission profiles.

Appendix

All code can be found on GitHub and accessed by clicking **HERE**.

Thank you!