IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

André Aguiar

07th May 2025

Outline

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

Executive Summary

Summary of methodologies:

- Data Collection through API and Web Scrapping
- Data Wrangling
- Exploratory Data Analysis
- Interactive Maps and Analytics Dashboard
- Machine Learning Models for Predictive Analysis

Summary of all results

- Comprehensive Analysis through Data Visualizations
- Models Evaluation for Prediction Task

Introduction

Project background and context

The aim of this project is to understand and predict the success of Falcon 9 landings by identifying the key factors that contribute to a successful outcome.
According to SpaceX, each Falcon 9 launch costs approximately $62 million. This is significantly lower than the estimated $165 million charged by competitors, positioning SpaceX as a leading candidate for future launches. The primary reason for this cost advantage is SpaceX's ability to reuse the rocket's first stage. Therefore, accurately predicting the success of first stage landings is critical to evaluating the true cost of a launch. For competitors aiming to bid against SpaceX, this insight is essential for SpaceX to remain competitive.

Problems you want to find answers

- What factors influence the likelihood of a successful rocket landing?
- How do different features interact to affect the landing success rate?
- What operating conditions are required to ensure a reliable landing program?

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology

SpaceX REST API

Web Scrapping from Wikipedia

Perform data wrangling

One-hot encoding for categorical features and feature selection

Perform exploratory data analysis (EDA) using visualization and SQL

SQL queries and plotting data to understand data patterns

Perform interactive visual analytics using Folium and Plotly Dash

Visualizing spatial data with Folium Maps and Dashboard creation

Perform predictive analysis using classification models

Built and evaluated multiple models using cross-validation to predict landing outcomes.

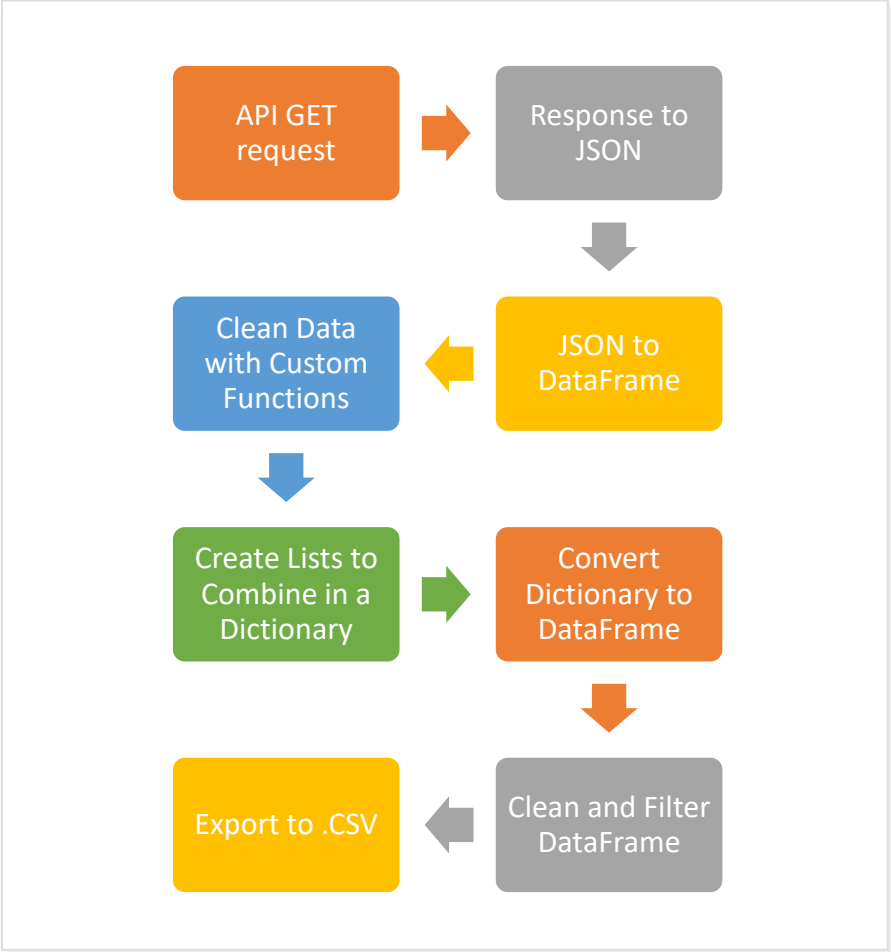
Data Collection

Data Collection process:

- Data was collected through GET requests to the SpaceX API.
- The response content was decoded and converted into a pandas DataFrame.
- Data was cleaned by checking for and handling missing values where necessary and preprocessed for further analysis.
- Additionally, Web Scraping was performed using BeautifulSoup to extract Falcon 9 launch records from Wikipedia.
- The launch records were parsed from an HTML table and transformed into a pandas DataFrame for further analysis.

Data Collection – SpaceX API

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1060
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	13	B1058
91	88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1051
92	89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	12	B1060
93	90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	8	B1062

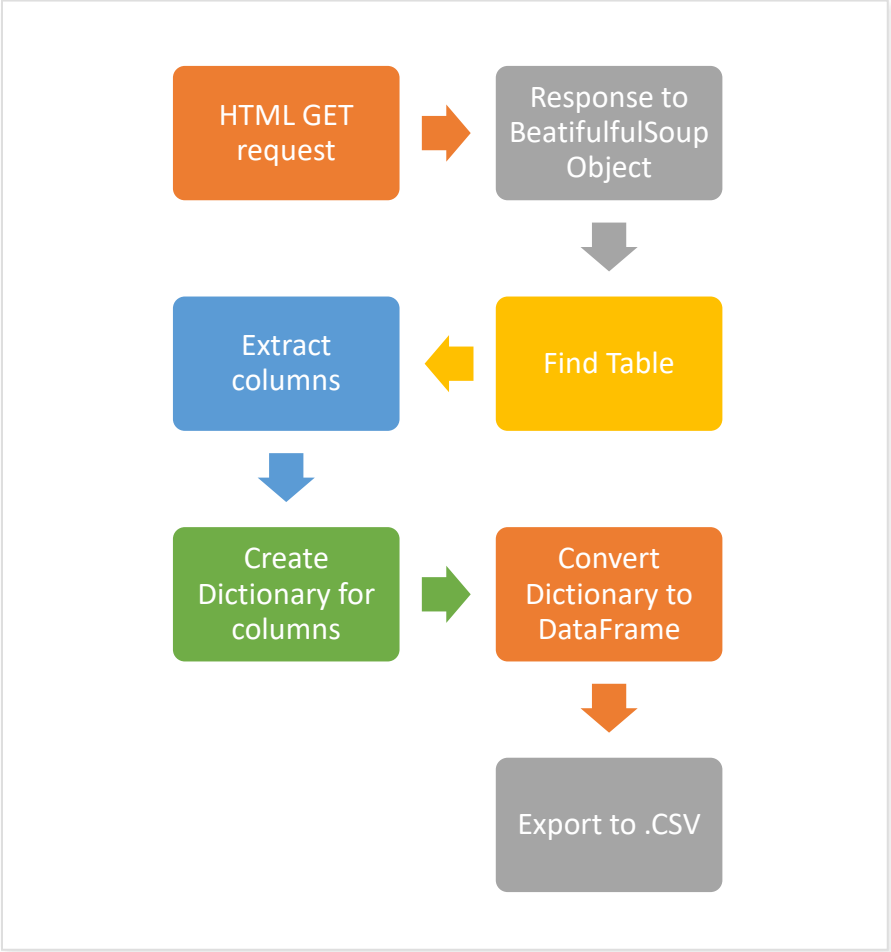


- GitHub URL

Data Collection - Scraping

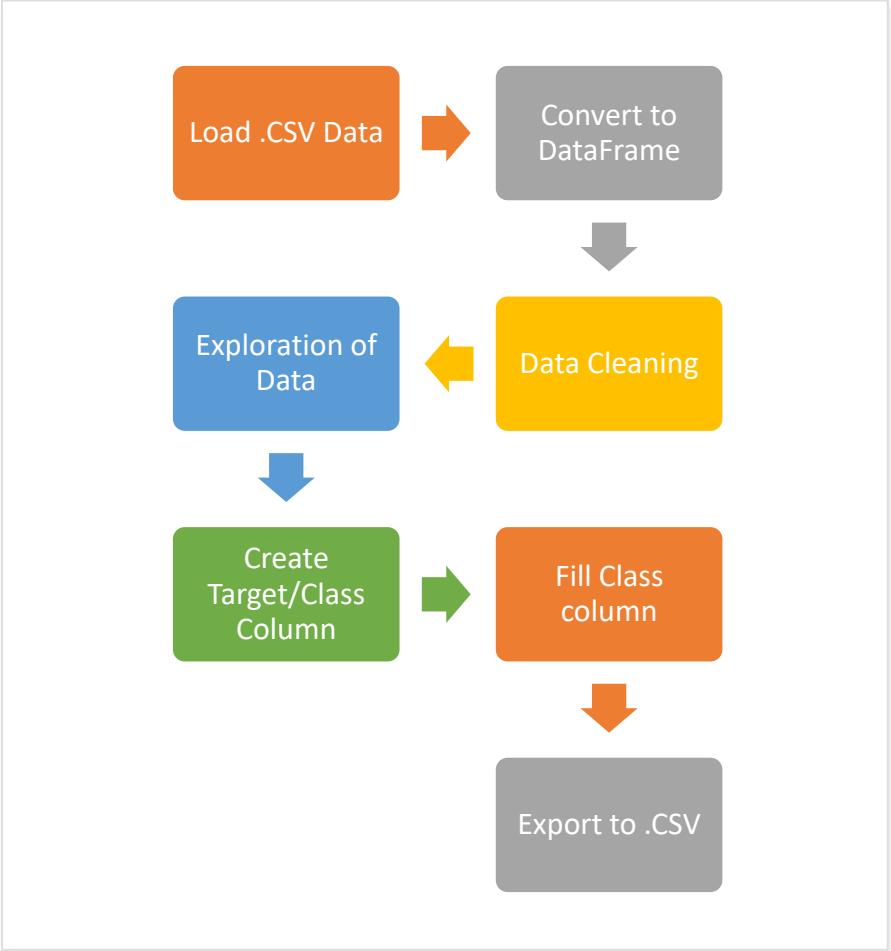
	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.07B0003.18	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.07B0004.18	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.07B0005.18	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.07B0006.18	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.07B0007.18	No attempt\n	1 March 2013	15:10
...	...	...	...	...	...	...	...	...	...	...	...
116	117	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1051.10657	Success	9 May 2021	06:42
117	118	KSC	Starlink	~14,000 kg	LEO	SpaceX	Success\n	F9 B5B1058.8660	Success	15 May 2021	22:56
118	119	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1063.2665	Success	26 May 2021	18:59
119	120	KSC	SpaceX CRS-22	3,328 kg	LEO	NASA	Success\n	F9 B5B1067.1668	Success	3 June 2021	17:29
120	121	CCSFS	SXM-8	7,000 kg	GTO	Sirius XM	Success\n	F9 B5	Success	6 June 2021	04:26

- [GitHub URL]



Data Wrangling

- The data was loaded from the .CSV into a DataFrame.
- To understand the data, a few calculations and value counts were performed (launches per site, launches per orbit, launch outcomes counts).
- Create column for the Class (target) variable (landing outcome label: 0 – if not successful; 1 – if successful)
- Calculate the success rate
- Export to .CSV



GitHub URL

EDA with Data Visualization

To explore the relationships between variables and uncover patterns in the data, several visualizations were created. These visual tools help identify which features may contribute to the success of rocket landings.

Scatter Plots:

FlightNumber and PayloadMass

FlightNumber and LaunchSite

Payload Mass and Launch Site

FlightNumber and Orbit

Payload Mass and Orbit

Bar Chart:

Success rate by Orbit Type

Line Chart:

Launch Success Rate by Year

GitHub URL

EDA with SQL

- Names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Booster versions that have carried the maximum payload mass
- Launches with month names, failure landing outcomes in drone ship, booster versions and launch site for the months in year 2015
- Ranked count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

GitHub URL

Build an Interactive Map with Folium

Launch sites were marked on a Folium map along with map objects such as:

- Markers
- Circles
- Lines

Launch outcomes were classified as binary values:

- 0 for failure
- 1 for success.

Color-coded marker clusters were used to visually assess which launch sites exhibited relatively higher success rates.

Distances from each launch site to nearby features were calculated to support spatial analysis.

- City
- Highway
- Railway
- Coastline

Key questions addressed included:

- Are launch sites located near railways, highways, or coastlines?
- Do launch sites maintain a certain distance from populated urban areas?

GitHub URL

Build a Dashboard with Plotly Dash

Interactive Dashboard with Plotly and Dash

- Data loaded from .CSV
- Pandas Dataframe

Elements for filtering and data selectors

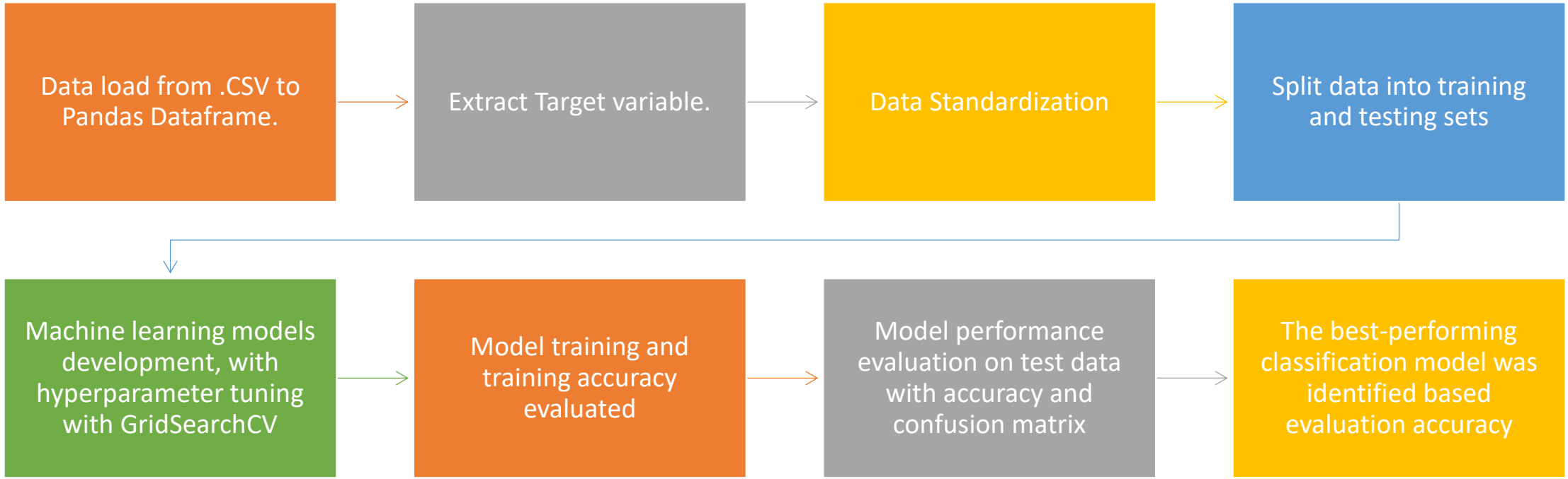
- Dropdown Menu for Launch Site selection (default – All sites included)
- Slider filter for Payload Mass range selection

Data Visualizations

- Pie Chart for successful launches per Site / Successful launches for All Sites
- Scatter plot for correlation between Payload Mass and Launch Success per Booster Version

GitHub URL

Predictive Analysis (Classification)



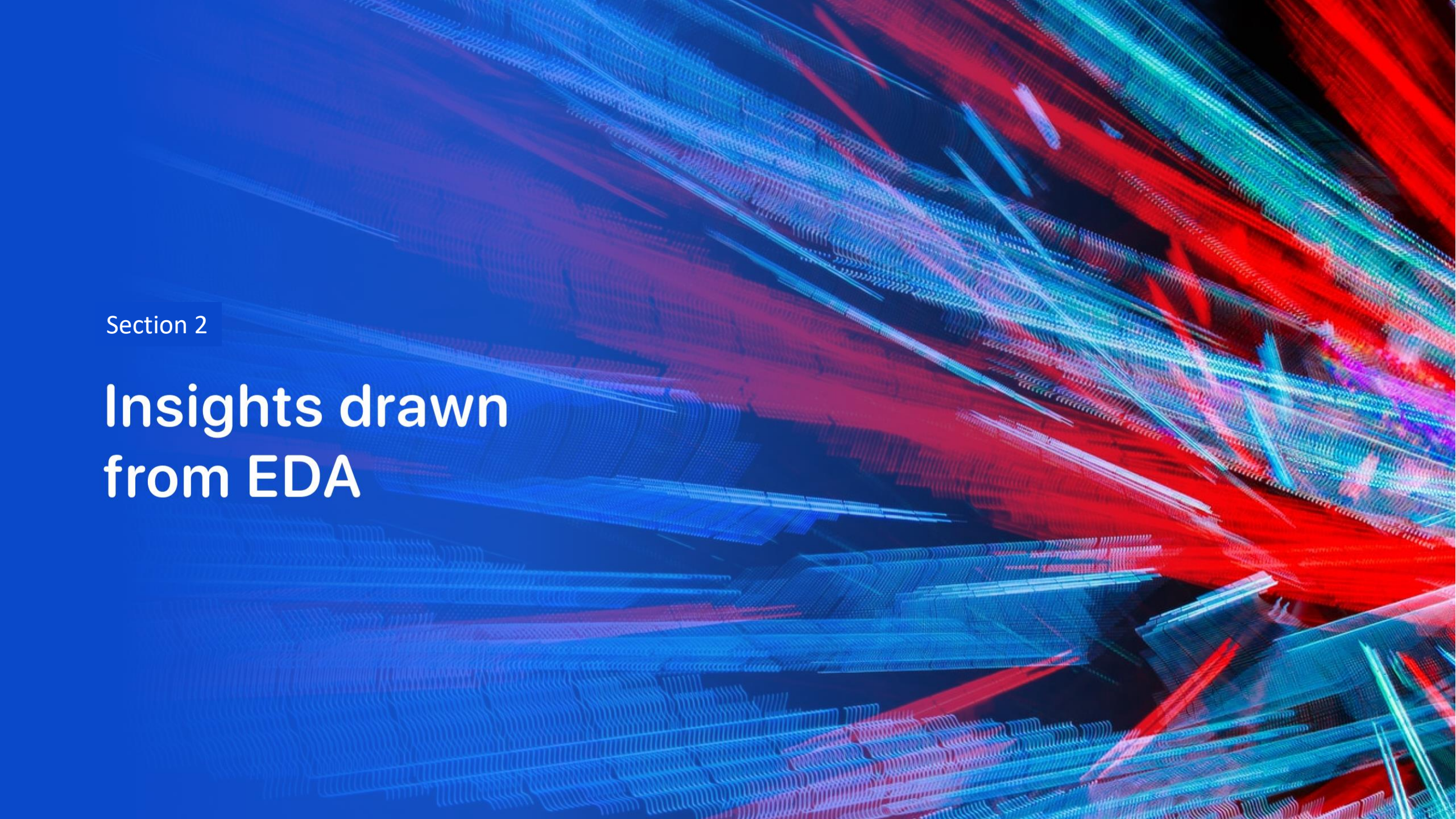
GitHub URL

Results

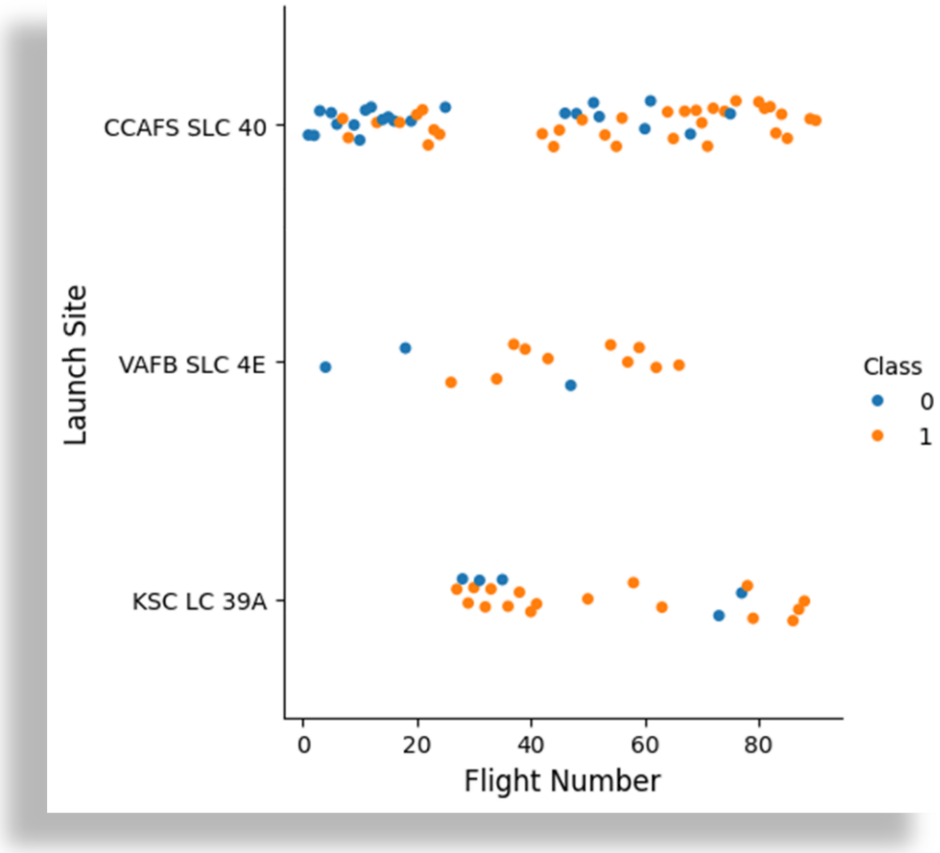
EXPLORATORY DATA ANALYSIS RESULTS

INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS

PREDICTIVE ANALYSIS RESULTS

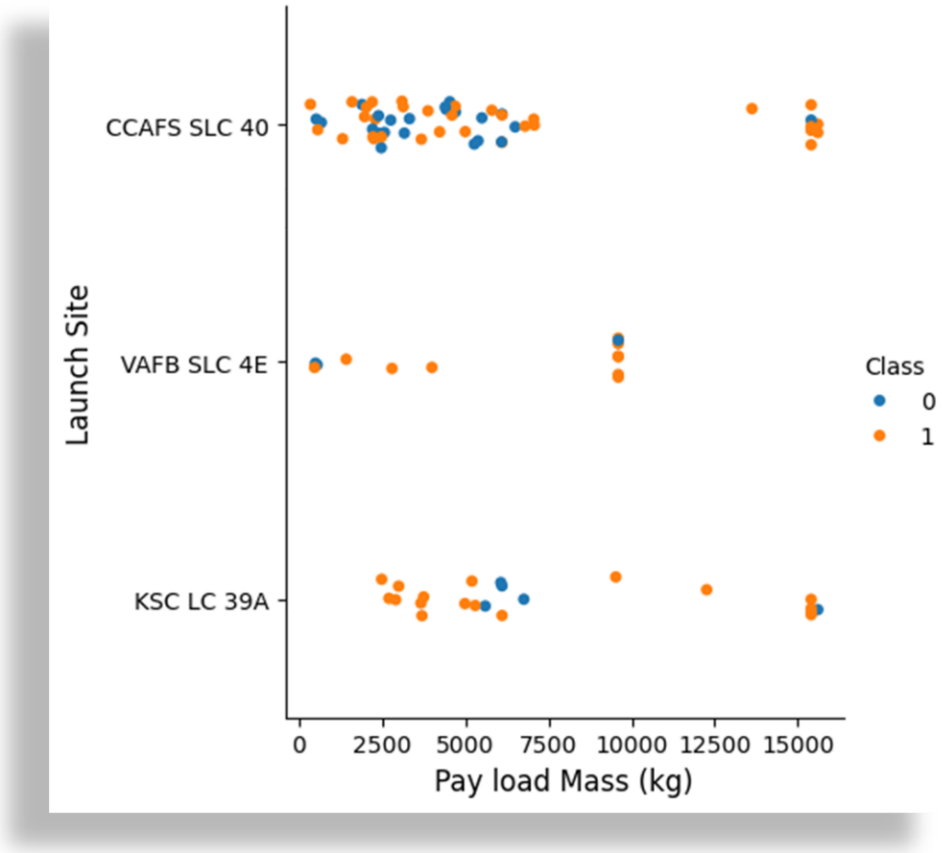

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



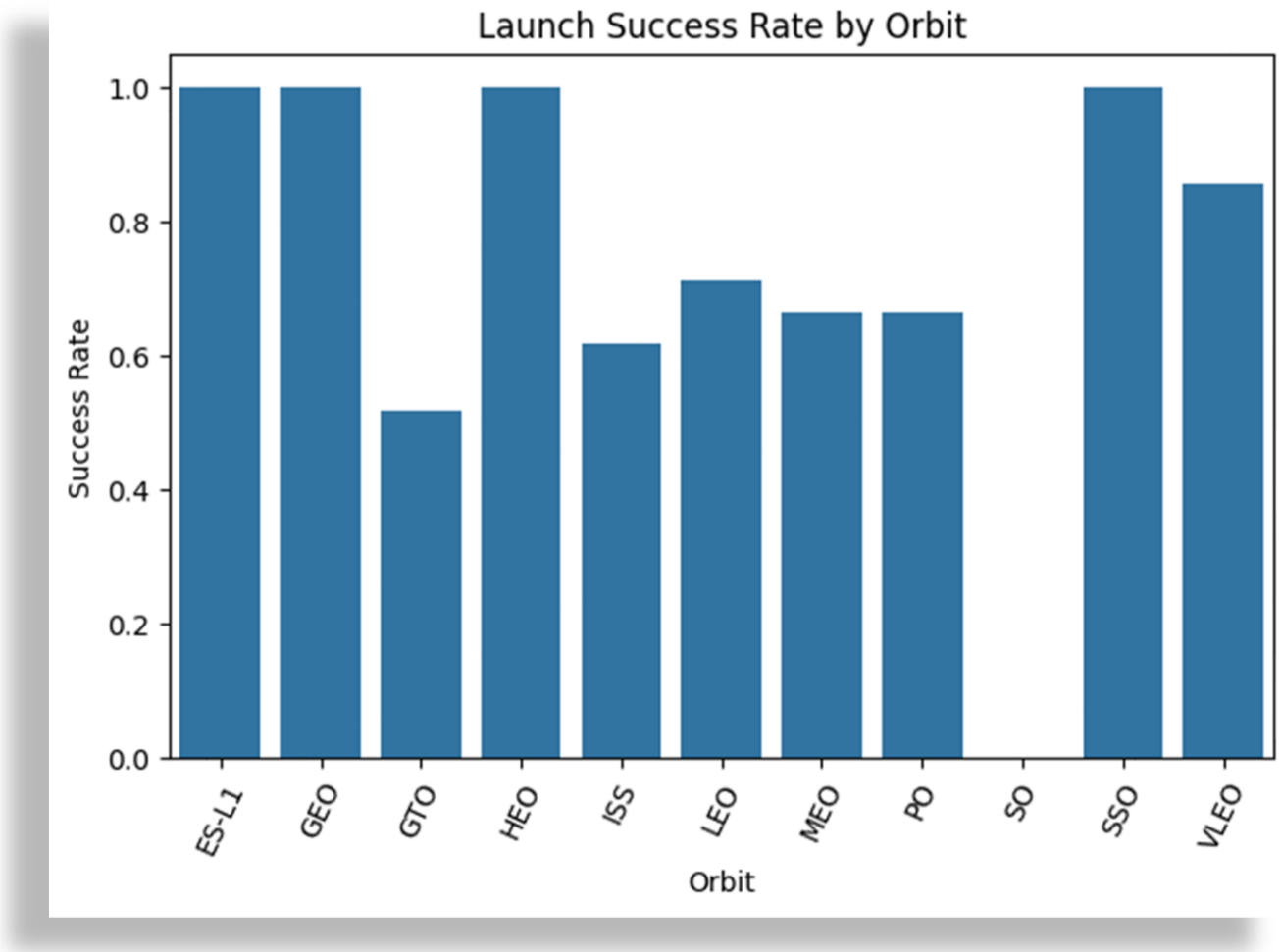
- It is possible to see a clear trend: as more launches flights occur on each site the launch success rate increases.

Payload vs. Launch Site



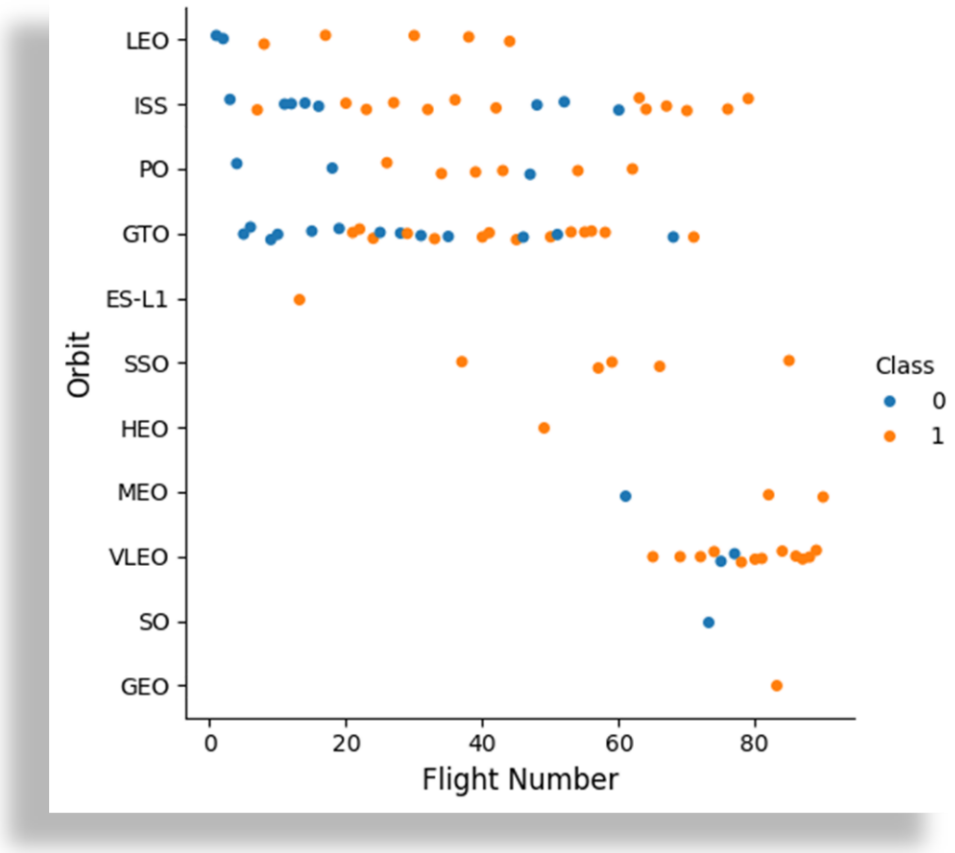
- There’s not a clear trend, although we can see that for CCAFS SLC 40 there’s a considerable success rate for flights with heavy pay load mass.
- For the other two launch sites, there’s seems to a be a considerable success rate for lighter pay load.

Success Rate vs. Orbit Type



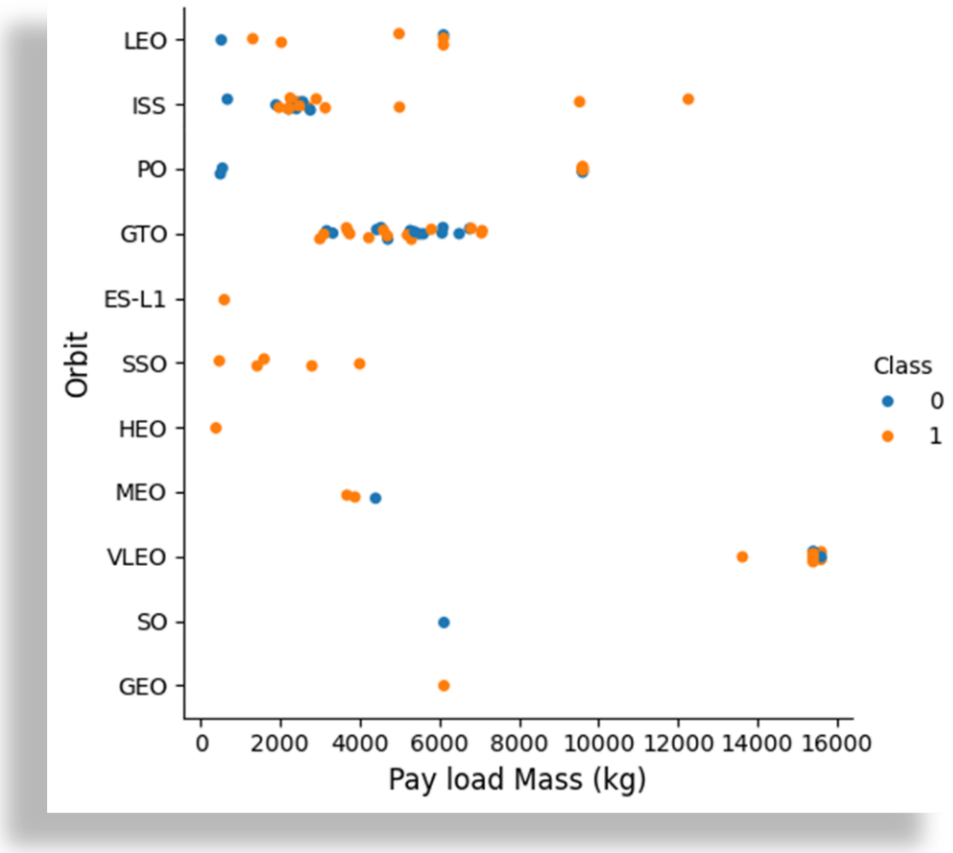
- The orbits ES-L1, GEO, HEO, SSO, VLEO have higher success rates.

Flight Number vs. Orbit Type



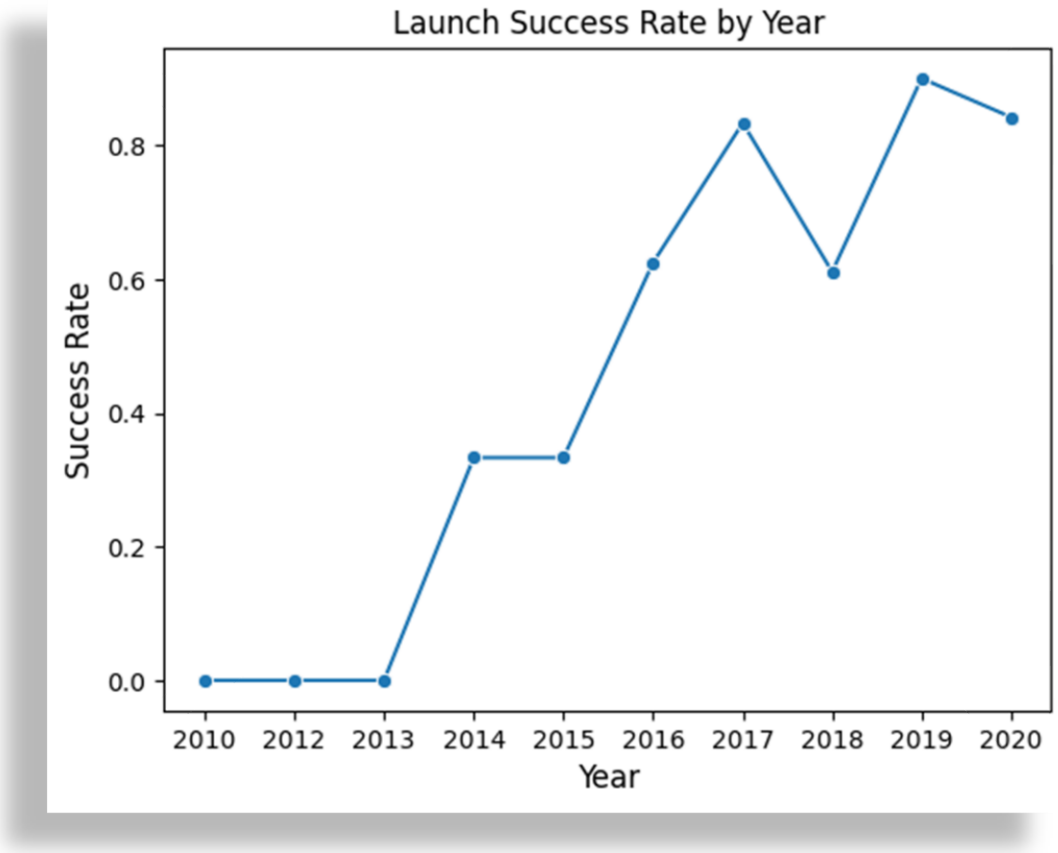
- LEO orbit success rate increases with the number of flights. PO and ISS arguably can be seen as having a similar trend.
- There’s no clear trend on GTO orbit.

Payload vs. Orbit Type



- Heavy payloads have best success rates for ISS and LEO orbits
- On the other hand, heavy payloads have less success rates for GTO and VLEO orbits
- SSO orbit with lighter payloads have good success rates

Launch Success Yearly Trend



- A steady increase in success rate can be observed up to 2017.
- After 2017 there’s more fluctuation with decreases and increases in success rate

All Launch Site Names

Out[45]:

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

In [45]:

```
%sql select distinct Launch_Site from SPACEXTBL;
```

- Using DISTINCT key word on the select query, it returns the unique Launch Site names
- There are 4 different Launch Sites names

Launch Site Names Begin with 'CCA'

Out[46]:

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

In [46]:

```
%sql select * from SPACEXTBL where Launch_Site like "CCA%" limit 5;
```

- Using LIMIT key word followed by 5 on the select query, it returns max 5 rows.
- The launch site name for all 5 rows displayed begin with CCA

Total Payload Mass

Out[47]:

Total NASA (CRS) payload mass (kg)
45596

```
In [47]: %sql select SUM(PAYLOAD_MASS__KG_) as "Total NASA (CRS) payload mass (kg)" from SPACEXTBL where Customer like "NASA (CRS)";
```

- Using SUM function and filtering with a condition for Customer, the result would be the total payload mass for that specific customer
- The value displayed is already in the desired unit

Average Payload Mass by F9 v1.1

Out[48]:

Average F9 v1.1 Payload Mass (kg)
2534.6666666666665

In [48]:

```
%sql select AVG(PAYLOAD_MASS__KG_) as "Average F9 v1.1 Payload Mass (kg)" from SPACEXTBL where Booster_Version like "%F9 v1.
```

- Using AVG function and filtering with a condition for Booster Version, the result would be the average payload mass for that specific booster version
- The value displayed is already in the desired unit

First Successful Ground Landing Date

Out[49]:

First Success Date
2015-12-22

In [49]:

```
%%sql
select
    Date as "First Success Date"
from
    SPACEXTBL
where
    Date = (select
                MIN(Date)
            from
                SPACEXTBL
            where
                Landing_Outcome like "Success (ground pad)"
            and Mission_Outcome like "Success");
```

- Using subquery to find the min value of Date with the specified conditions for Landing Outcome and Mission Outcome

Successful Drone Ship Landing with Payload between 4000 and 6000

Out[50]:

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

In [50]:

```
%%sql
select
    distinct(Booster_Version)
from
    (select
        Booster_Version
    from
        SPACEXTBL
    where
        Landing_Outcome like "Success (drone ship)"
        and Mission_Outcome like "Success"
        and PAYLOAD_MASS__KG_ > 4000
        and PAYLOAD_MASS__KG_ < 6000);
```

- Using subquery to filter rows between the payload mass values and with the specified conditions for Landing Outcome and Mission Outcome
- Then applying the DISTINCT key word on the subquery result, returning unique booster versions

Total Number of Successful and Failure Mission Outcomes

Out[51]:

trim(Mission_Outcome)	Total Missions
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

```
In [51]: %sql select trim(Mission_Outcome), count(Mission_Outcome) as "Total Missions" from SPACEXTBL group by trim(Mission_Outcome);
```

- Using COUNT function for each value of Mission Outcome returns the number of records for each
- Trim function was applied to unify values that may have spaces
- Note that there is two values with “Success”

Boosters Carried Maximum Payload

Out[52]:

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

In [52]:

```
%%sql
select
    distinct(Booster_Version)
from
    SPACEXTBL
where
    PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

- Subquery to get the value of Max Payload mass
- Then using the max value to filter the rows that have payload mass equals to the max value
- DISTINCT key word to get unique booster version names

2015 Launch Records

Out[53]:

MonthName	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

In [53]:

```
%%sql
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 MonthName, Landing_Outcome, Booster_Version, Launch_Site
from SPACEXTBL
where
    substr(Date,0,5) = '2015'
    and Landing_Outcome like '%Failure (drone ship)%';
```

- Use CASE expression inside the SELECT statement to assign the month name
- SELECT specific columns
- Filter for year 2025 and for specific condition of Landing Outcome

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

Out[54]:

Landing_Outcome	Outcome_Count
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

In [54]:

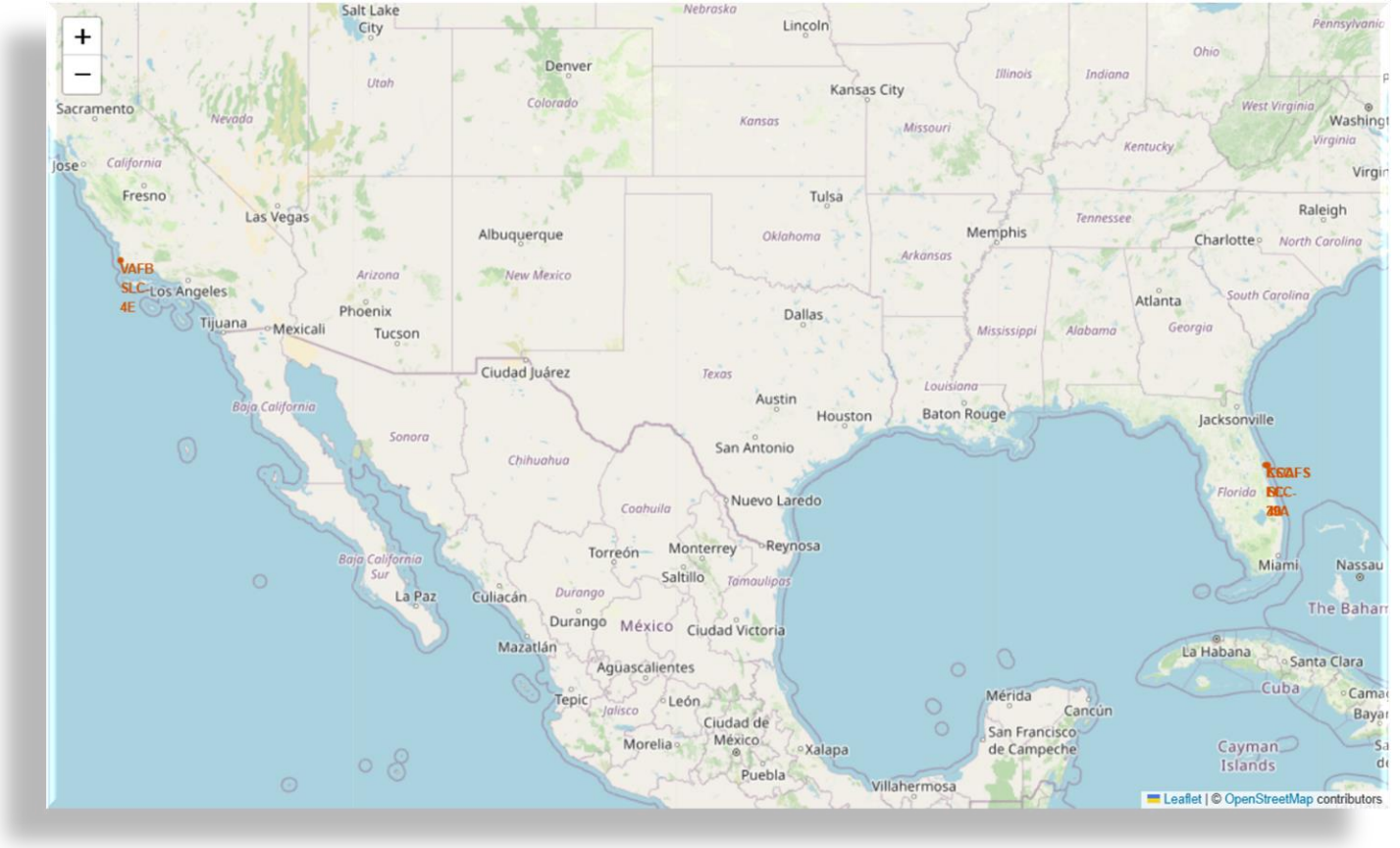
```
%%sql
select
    Landing_Outcome,
    count(*) as Outcome_Count
from
    SPACEXTBL
where
    Date >= '2010-06-04'
    and Date <= '2017-03-20'
group by Landing_Outcome
order by Outcome_Count desc;
```

- Use COUNT(*) on SELECT statement to get number of rows for each group between the specified Date interval
- GROUP BY to group records for each different Landing Outcome
- ORDER BY and DESC on the count itself

Section 3

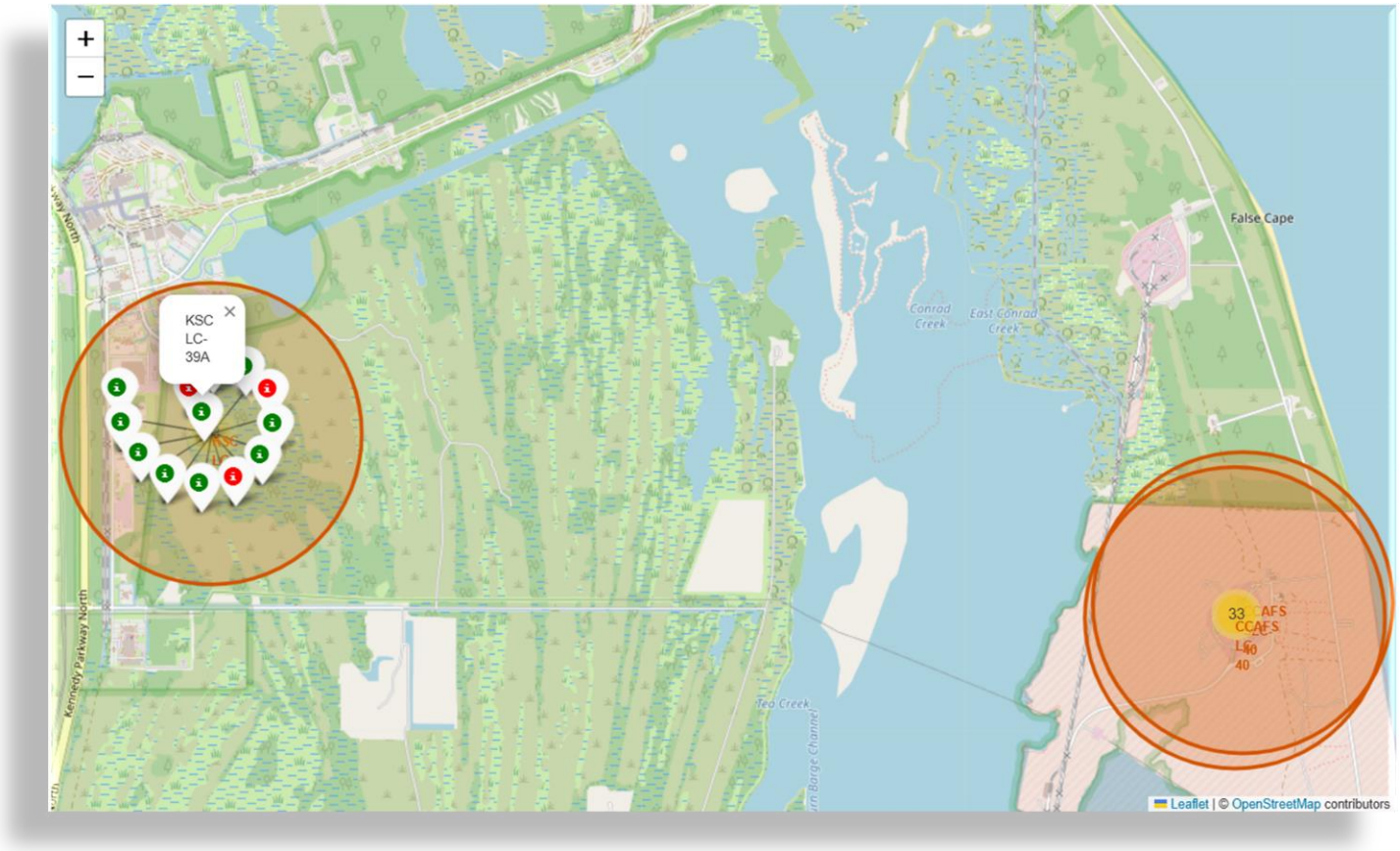
Launch Sites Proximities Analysis

Launch Sites Locations



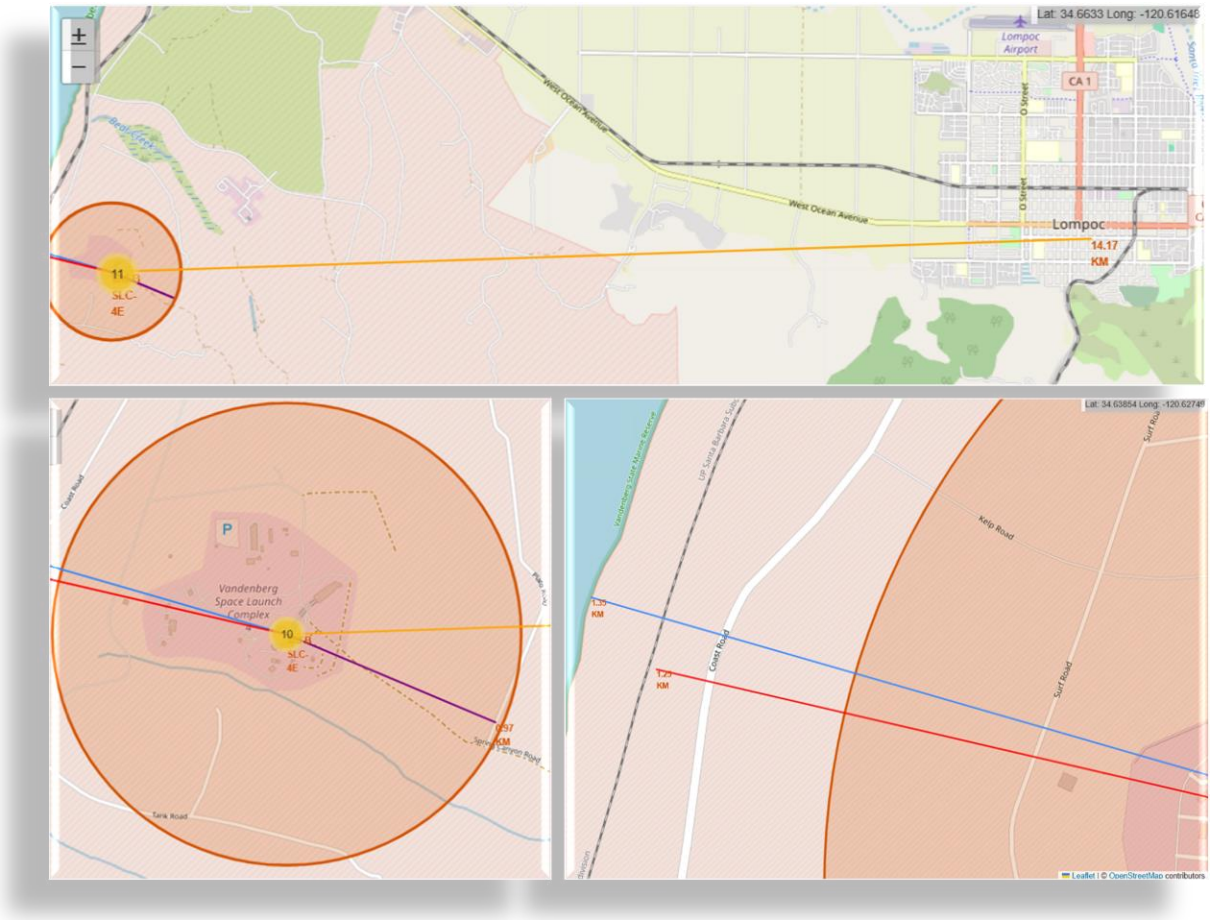
- From the resulting map we can determine that the launch sites are located near the coast of United States of America, one in California coast and the other 3 on the Florida coast

Launch Outcome by Location

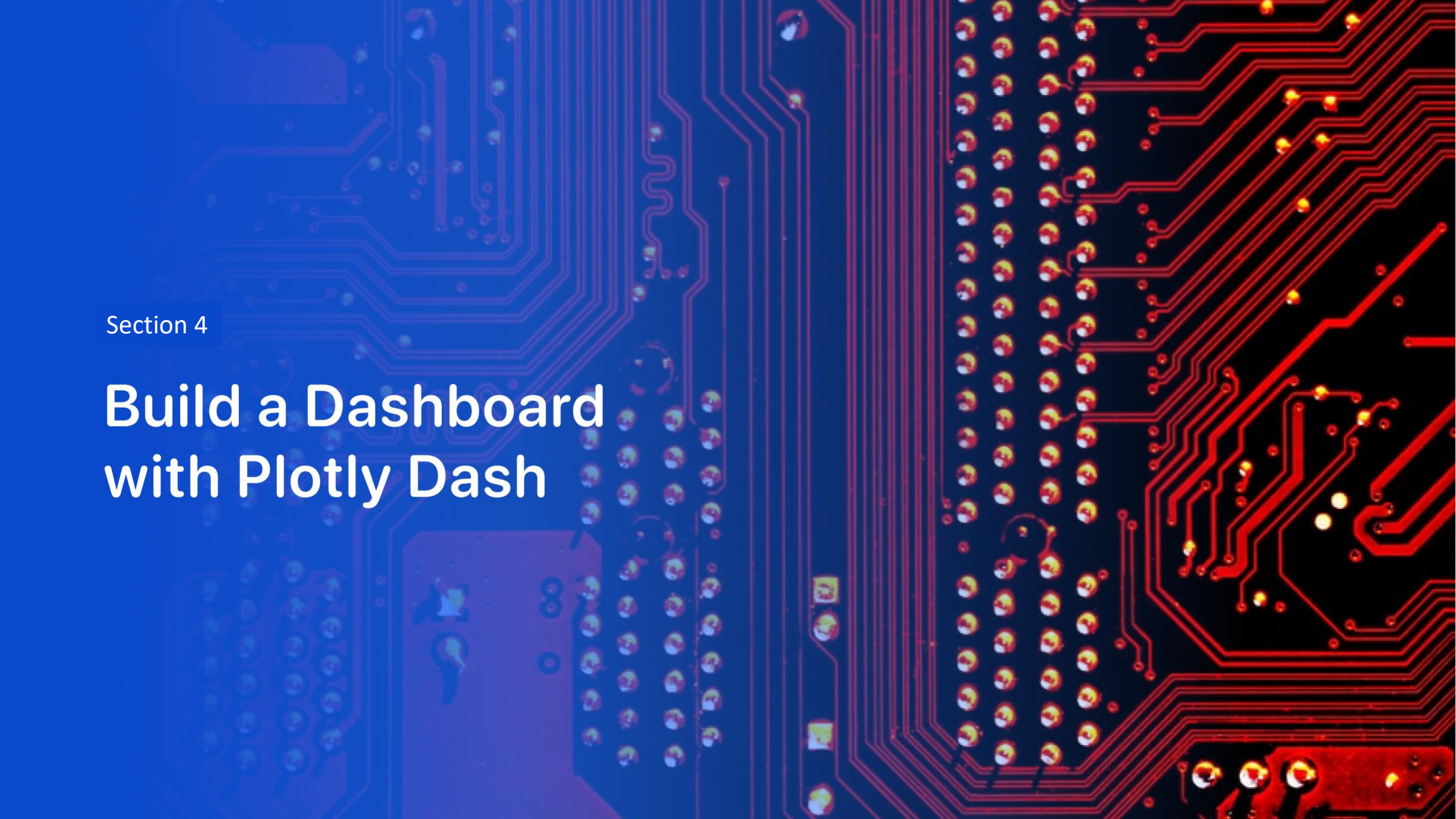


- Green markers show successful launches while red markers show failed launches. This provides a visual insight on the success rate for each site.
- In the image above an example is provided for the launch site with the best launch success rate (KSC LC-39A on the Florida coast).

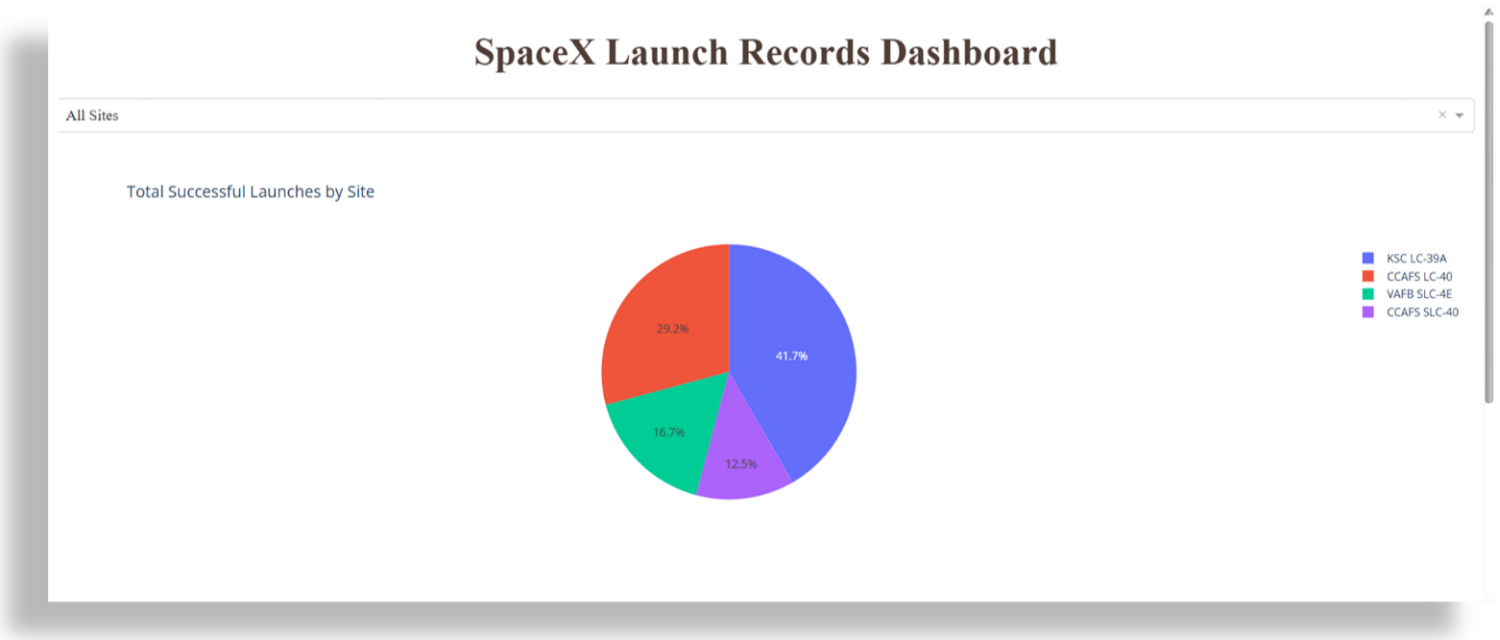
Launch Site Proximities



- By marking certain places or points of interest, and by drawing a polyline with a calculated distance mark we can easily access the proximities to launch sites.
- Considering all launch sites proximities, we can determine that those are usually relatively close to the coastline, to railways and highways while being somewhat far from cities and populated areas.

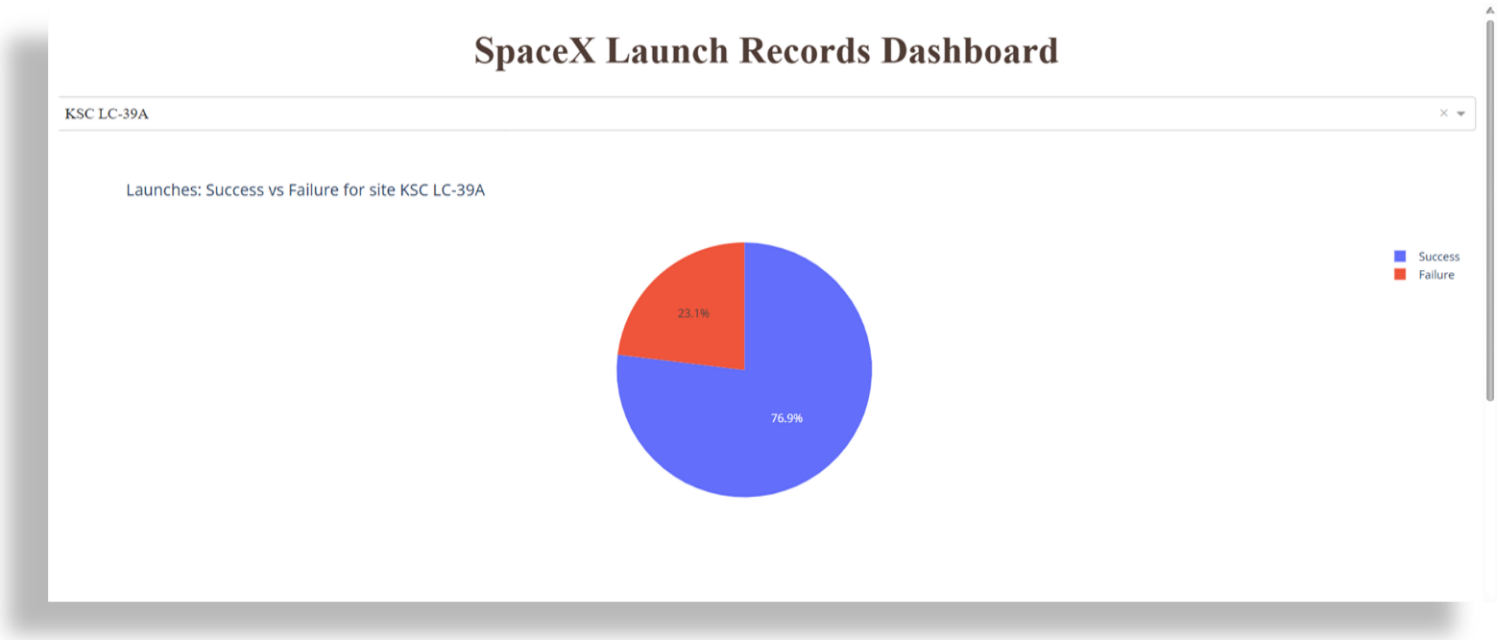

Section 4
Build a Dashboard with Plotly Dash

Launch Success by Launch Site



- The launch site with the most successful launches is KSC LC-39A
- Launch sites KSC LC-39A and CCAFS LC-40 account for more than 70% of the total successful launches

Launch Site with Highest Success Ratio



- The success rate for the the Launch Site KSC LC-39A is 76,9%
- A small fraction of the launches have a failure outcome (23,1%)

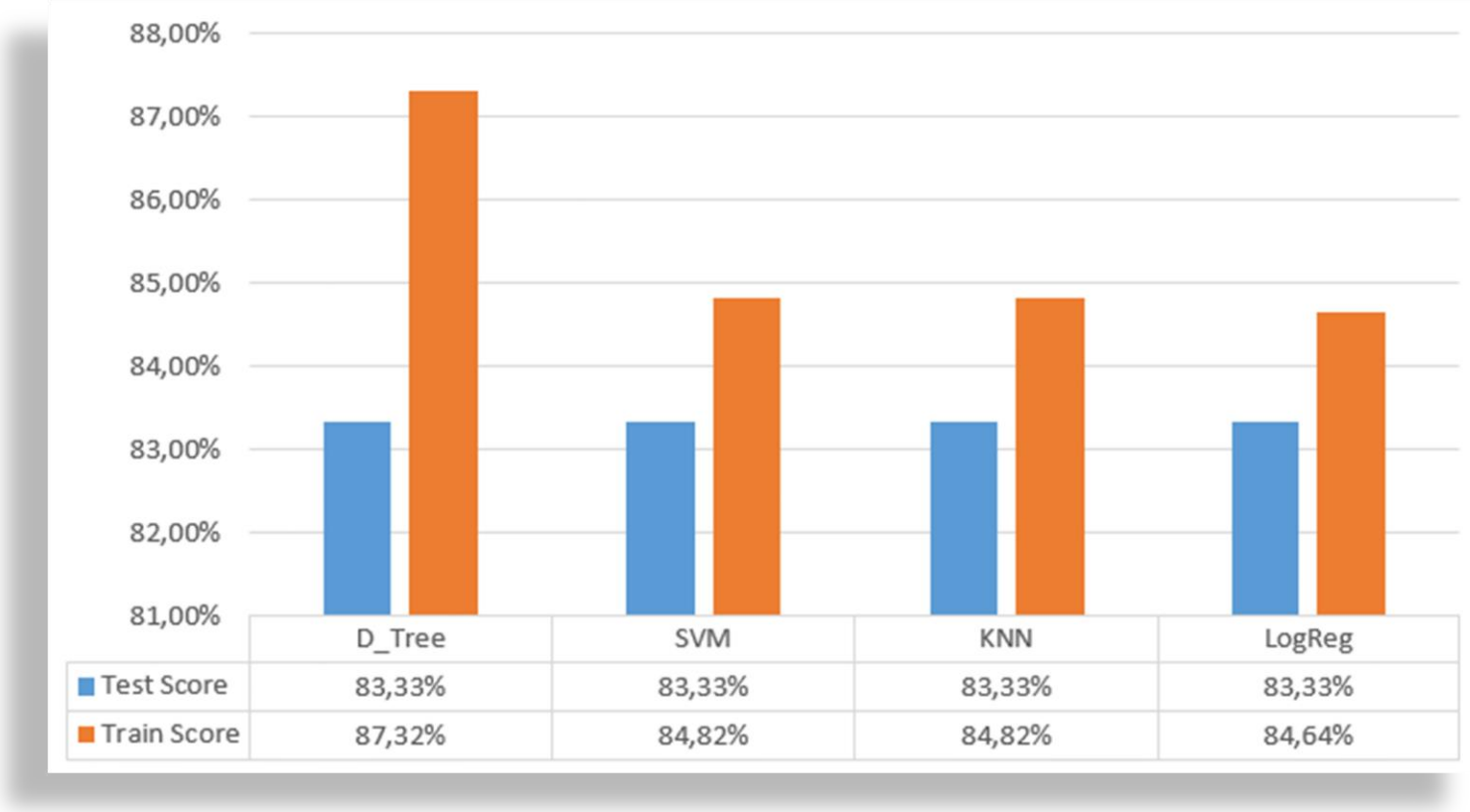
Payload Mass and Success Outcome

- The success rate for lower payload mass is greater than for higher payload mass.
- This trend can be more clearly seen on FT and B4 booster version categories
- Booster version categories v1.0 and v1.1 have low success rates

Section 5

Predictive Analysis (Classification)

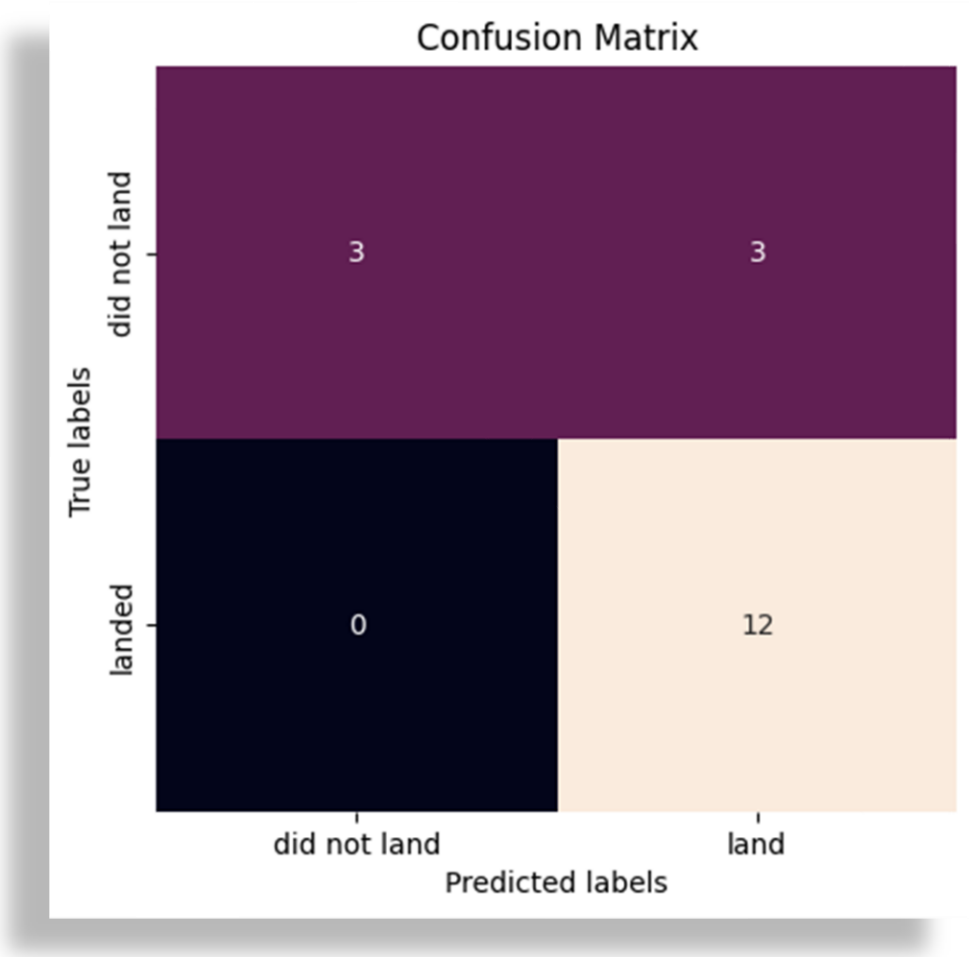
Classification Accuracy



	D_Tree	SVM	KNN	LogReg
Test Score	83,33%	83,33%	83,33%	83,33%
Train Score	87,32%	84,82%	84,82%	84,64%

- By evaluating the accuracy on all models, we can determine that all four models have the same performance (83% accuracy) on test data
- The accuracy on training data is also very similar between the four models, although Decision Tree Classifier comes out as the winner with the highest accuracy (87,32%)

Classification Accuracy



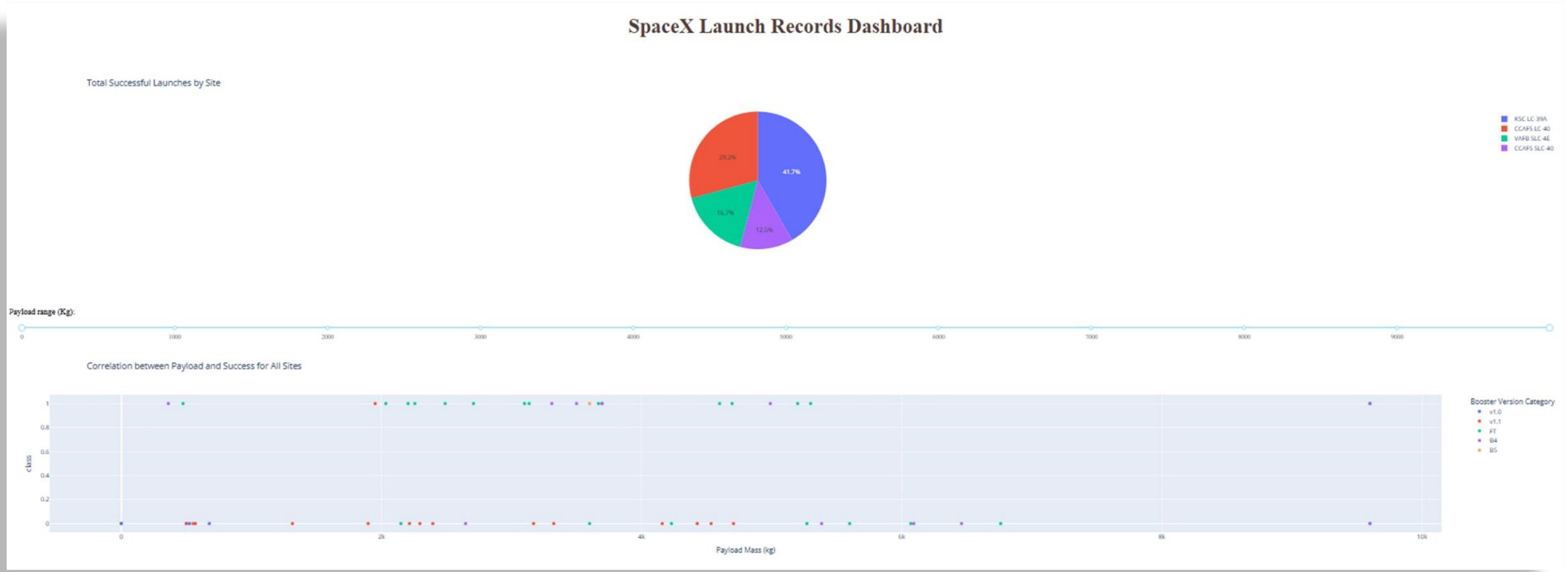
- All four models share the same confusion matrix with 0 False Negative (FN) predictions and 3 False Positive (FP) predictions
- All models performed quite well, despite the considerable FP rate

Conclusions

- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate
- The success rate increased with time (up to 2017 with a steady increase) and number of flights
- Launch Site KSC LC-39A has the higher success rate of all sites
- Payload mass may have a negative impact on success
- Decision Tree Classifier performed better than other classifiers on training/validation data, so we can access it would be the best machine learning model for the task

Appendix

Full preview of Plotly Dash dashboard



Thank you!