



IBM Developer
SKILLS NETWORK

Winning Space Race With Data Science

Nguyen Thi Tuong Vy
November 9, 2023



Outline

Executive Summary

Introduction

Methodology

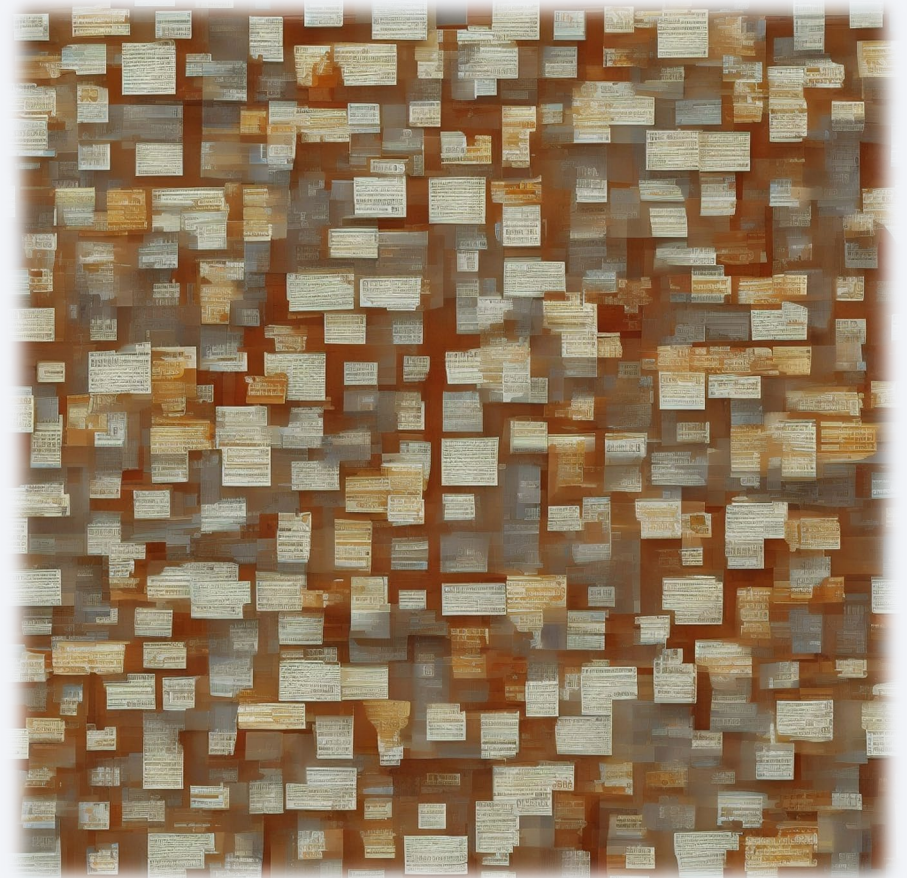
Results

Conclusion



Executive Summary

- Summary of methodologies
 - Data Collection through API
 - Data Collection with Web Scraping
 - Data Wrangling
 - Exploratory Data Analysis with SQL
 - Exploratory Data Analysis with Data Visualization
 - Interactive Visual Analytics with Folium
 - Machine Learning Prediction
- Summary of all results
 - Exploratory Data Analysis result
 - Interactive analytics in screenshots
 - Predictive Analytics result from Machine Learning Lab



Introduction

SpaceX has gained worldwide attention for a series of historic milestones. It is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010. SpaceX advertises **Falcon 9** rocket launches on its website with a cost of **62 million dollars** whereas other providers cost upward of **165 million dollars** each, much of the savings is because Space X can **reuse the first stage**. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. To do this, we can use public data and machine learning models to **predict** whether SpaceX – or a competing company – can reuse the first stage.

The problems included:

- Identifying **factors** that **influence** the **landing outcome**
- The **relationship** between each factor and affection to the outcome
- Best predictive **model for predicting** if the first stage will land

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
 - SpaceX Launch Data was collected using [SpaceX REST API](#) and [web scrapping](#) from Wikipedia
- Perform data wrangling
 - SpaceX Launch Data was processed using [one-hot encoding](#) for categorical features, Wrangle data to create success/fail outcome variable
- Perform exploratory data analysis (EDA) using [visualization](#) and [SQL](#)
- Perform interactive visual analytics using [Folium](#) and [Plotly Dash](#)
- Perform predictive analysis using [classification models](#)
 - How to build, tune, evaluate classification models

Data Collection - Data Sources

REST API

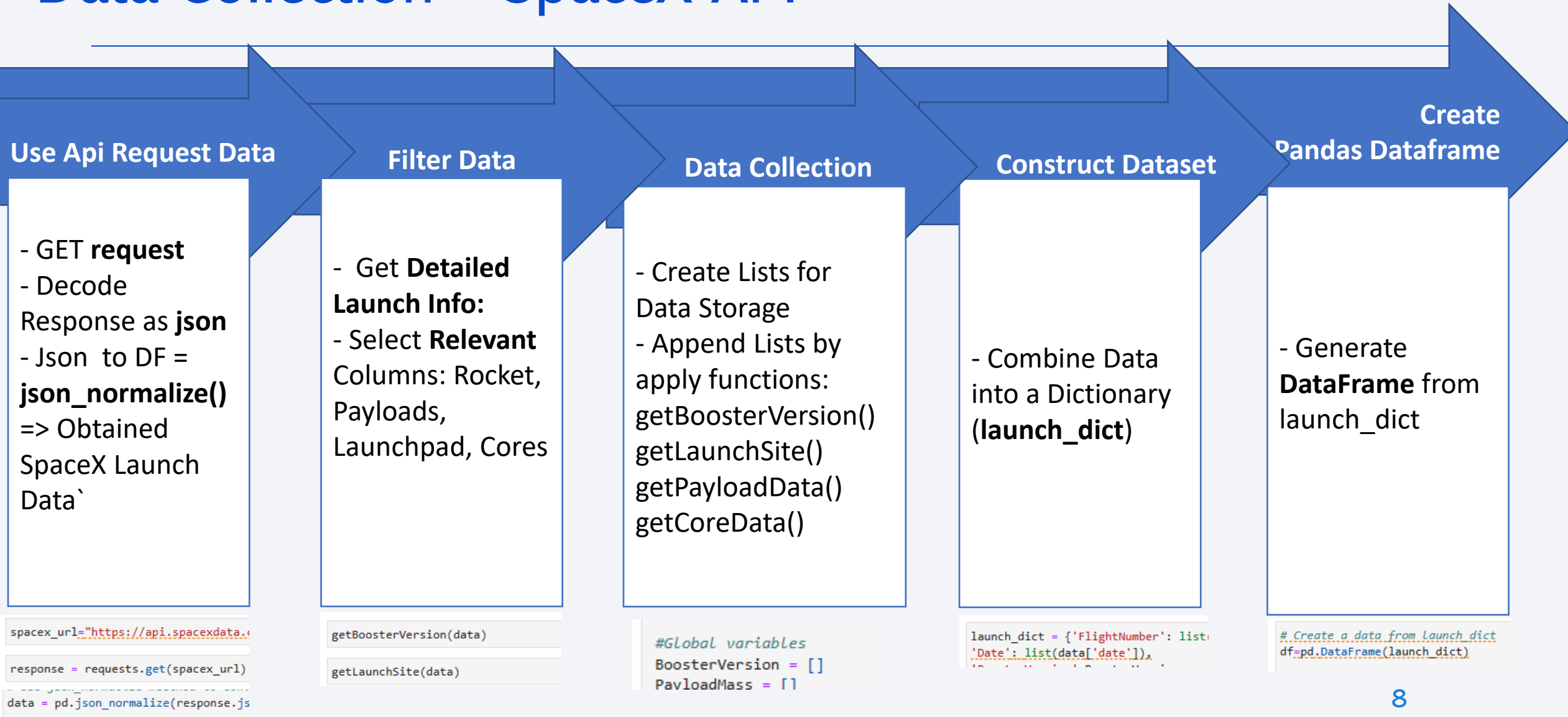
- Request and parse SpaceX launch data.
- Get booster version, launch site, payload data, and core data.
- Filter and format data, including date conversion.
- Create lists for data storage.
- Apply functions to populate lists.
- Construct a dataset by combining data into a dictionary.
- Create a Pandas Dataframe from the dictionary.



Web Scraping

- Request the Falcon9 Launch Wiki page.
- Create a BeautifulSoup object to parse the HTML.
- Extract column/variable names from the HTML table header.
- Create a dictionary for data storage with keys from column names.
- Populate the dictionary with launch records.
- Create a Pandas DataFrame from the dictionary.

Data Collection – SpaceX API



Data Collection – Scraping - Falcon9 Launch

Request Wiki Page

- Use **requests.get()** to fetch the Falcon9 Launch HTML page

```
# use requests.get() method  
# assign the response to  
data = requests.get('http://www.spacex.com/launches')
```

Parsed the data

- Use BeautifulSoup to parse the HTML response content

```
# Use BeautifulSoup() to  
soup = BeautifulSoup(data.text, 'html.parser')
```

Extract data

- collect all relevant column names from the HTML table header
- extract column name one by one

```
# Use the find_all function to  
# Assign the result to a list  
html_tables = soup.find_all('table')  
  
for row in first_launch_tables:  
    name = extract_column_names(row)  
    if (name != None and  
        column_names.append(name))
```

Construct Dataset

- Create an empty dictionary with keys extracted name
- Populate the dictionary

```
launch_dict = dict.fromkeys(column_names)  
extracted_row = 0  
# Extract each table  
for table_number, table in enumerate(html_tables):  
    # get table row  
    for rows in table.find('tr'):  
        # check to see if  
        if rows.th:
```

Create Pandas Dataframe

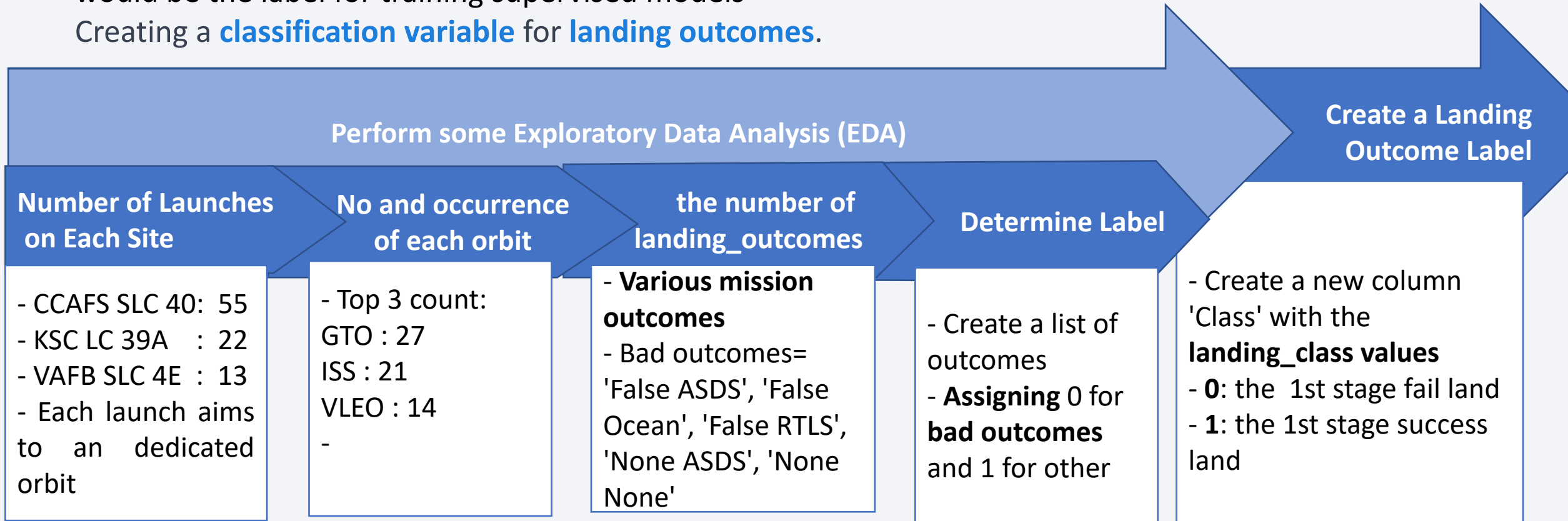
- Create a Pandas DataFrame from the dictionary

```
df = pd.DataFrame(launch_dict)
```

Data Wrangling

Perform some **Exploratory Data Analysis** (EDA) to find some patterns in the data determine what would be the label for training supervised models

Creating a **classification variable** for **landing outcomes**.



EDA with Data Visualization

Explore and visualize the relationships between various factors and their influence on the success or failure of SpaceX launches. The insights gained can inform future analyses and predictions.

Charts

- Scatterplot Flight Number vs. Payload Mass (kg) with Launch Outcome Overlay
- Scatterplot Flight Number vs. Launch Site with Launch Outcome Overlay
- Scatterplot Payload Mass (kg) vs. Launch Site with Launch Outcome Overlay
- Bar Chart Success Rate by Orbit Type
- Scatterplot Flight Number vs. Orbit Type with Launch Outcome Overlay
- Scatterplot Payload Mass (kg) vs. Orbit Type with Launch Outcome Overlay
- Line Chart Launch Success Yearly Trend



EDA with SQL

Summarized SQL queries performed

- Select distinct launch sites from the table SPACEXTBL
- Select launch sites begin with 'CCA', limit to 5 records.
- Calculate total payload mass carried by boosters launched by NASA (CRS)
- Calculate average payload mass carried by booster version F9 v1.1
- Find the date of the first successful landing outcome on a ground pad
- List booster names with successful drone ship landings and payload mass 4000-6000 kg
- Count the total number of successful and failure mission outcomes
- Find the names of booster versions with the maximum payload
- List records with month, failure landing outcomes on a drone ship, booster versions, and launch sites for the months in the year 2015
- Rank the count of landing outcomes between specific dates in descending order

Build an Interactive Map with Folium

The following map objects are created and added to a Folium map:

- **Markers:** are created for launch sites, each launch record, and various points of interest (e.g., [coastline](#), [city](#), [railway](#), [highway](#)). Markers for launch sites indicate the specific locations from which rockets are launched. Markers for launch records show the [success](#) or [failure](#) of each launch.
- **Circles:** circles around launch sites illustrate their general area of influence. Circles help highlight and visualize the proximity of launch sites to significant locations such as coastlines.
- **Lines (Polylines):** are used to connect launch sites to their proximities, such as coastlines, cities, railways, and highways. These lines visually display the [distances](#) between these locations, providing insights into the site's geographic relationships.
- **Marker Clusters:** help organize and group markers that share the [same coordinates](#), making it easier to visualize and interpret the data.

=> provide visual representation and insights into SpaceX launch data and the [geographic characteristics of launch sites](#).

Build a Dashboard with Plotly Dash

SpaceX dashboard app was build to provide insights and enable users to explore SpaceX launch records

Dropdown List

Launch Site Selection

- Allow users to **select** a specific launch site or view data for all sites. Users can **focus** on launches from specific sites or analyze launches collectively for all site

Pie Chart

Total success launch by Site

- Visually represents the total successful launch count. When a specific launch site is selected, it displays the **success-to-failure ratio** for that site, allows users to quickly understand the success rates and make **comparisons** between different launch sites.

Slider

Payload Range Slider

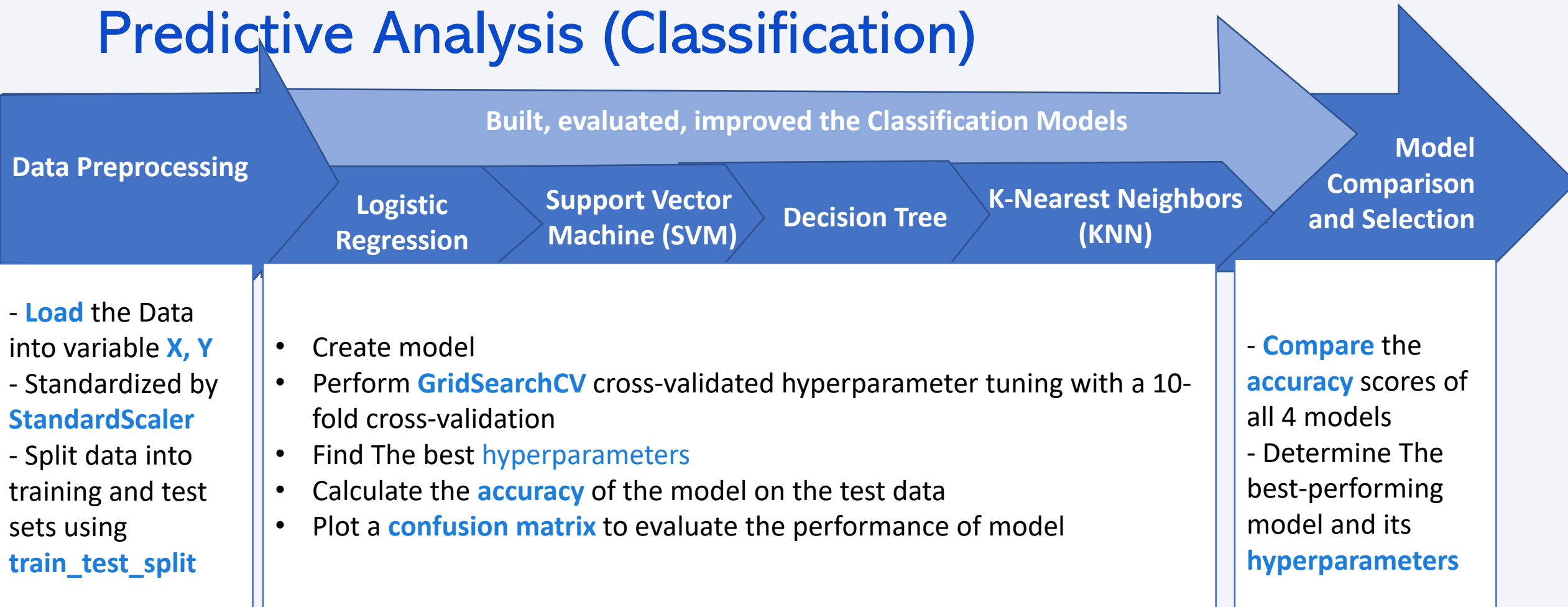
- Enables users to specify a payload range in kg, allows users to **filter the data** on payload mass

Scatter Chart

Correlation between payload mass & launch success

- It uses the payload range selected via the slider and displays data points for each launch. Users can explore how **payload mass relates to launch success** and identify **potential trends**.

Predictive Analysis (Classification)



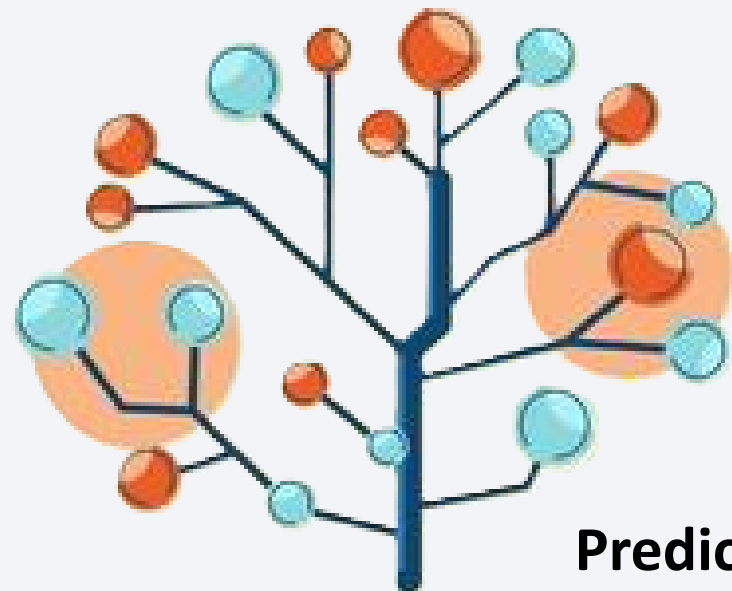
Best model is Decision Tree with a score of 0.875

Results

Exploratory Data Analysis

- Payloads: Launches with **low-weight** perform better than heavier
- Launch success has **improved over time**
- **KSC LC-39A** has the highest success
- The **SSO** orbit exhibits exceptional success

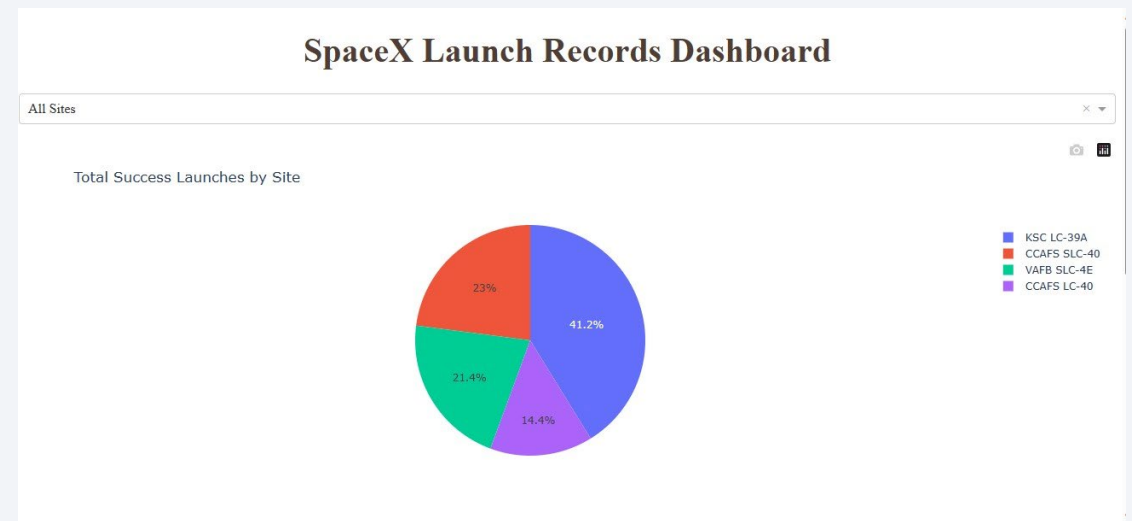
Interactive analytics demo



Predictive analysis results

DECISION TREE

- **Decision Tree** model is the best predictive model for the dataset

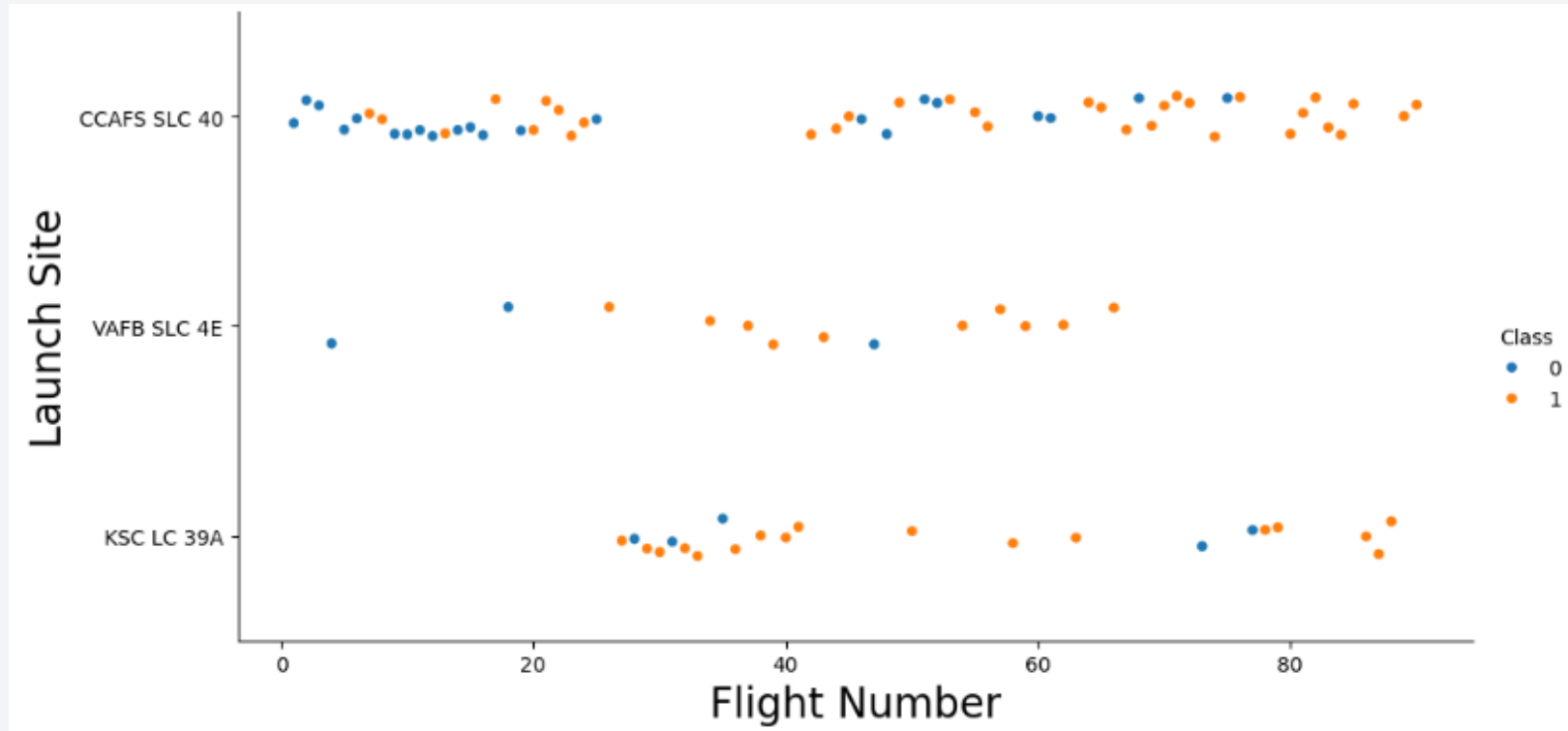


The background of the slide is an abstract composition. It features a solid blue area on the left side, which transitions into a dynamic pattern of diagonal streaks in shades of blue and red on the right. Overlaid on these streaks is a fine, light-colored grid or mesh pattern, giving the impression of a digital or data-driven environment.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



blue = fail

orange = success

Trend: As we move towards higher Flight Numbers, success rate become higher (the color shifts to orange)

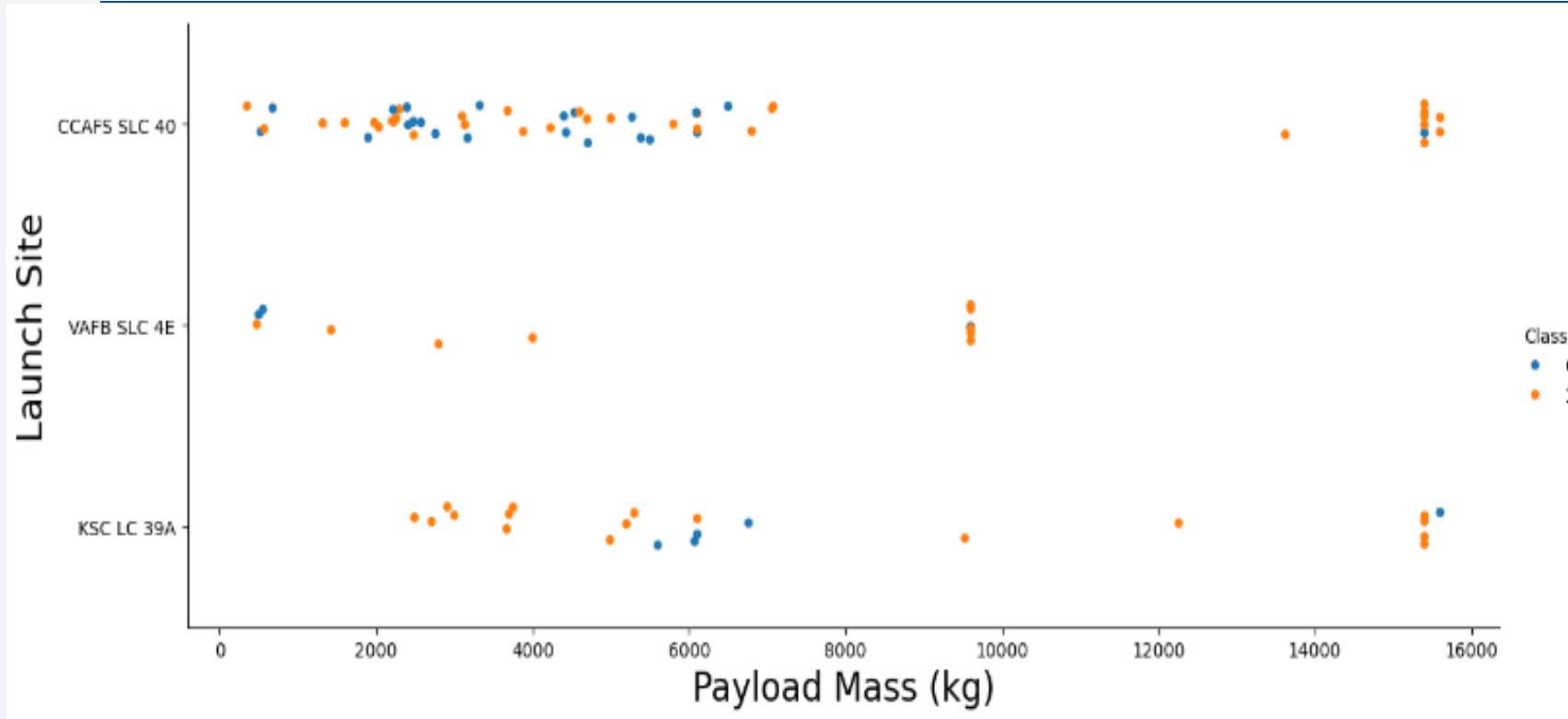
Distribution: ~1/2 launches are at the CCAFS SLC 40 launch site

Launch site success rates: VAFB SLC 4E and KSC LC 39A exhibit a notably higher success rate

SpaceX's more recent launches tend to have a higher success rate.

=> expectation of continuous improvement and learning from earlier launch experiences

Payload vs. Launch Site



Successful High-Payload Launches:

~Most launches with a payload >7,000 kg outcome success

KSC LC 39A : stands out with 100% success rate for launches with payload <5,500 kg

VAFB SLC 4E Payload Limit: there are no rockets launched for heavypayload mass(>10.000kg)

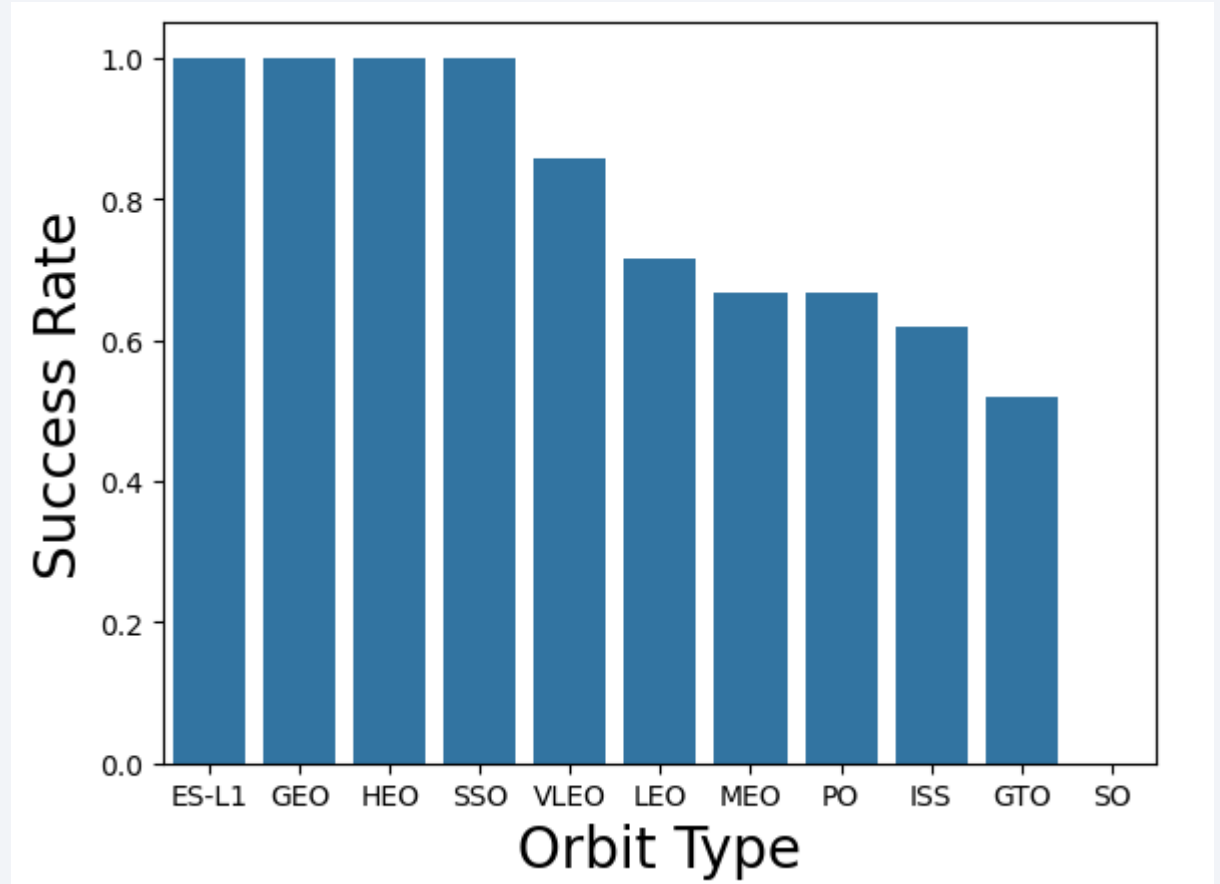
The data suggests a **positive correlation** between the payload mass (measured in kilograms) and the launch success rate.

Success Rate vs. Orbit Type

The relationship between success rate of each orbit type:

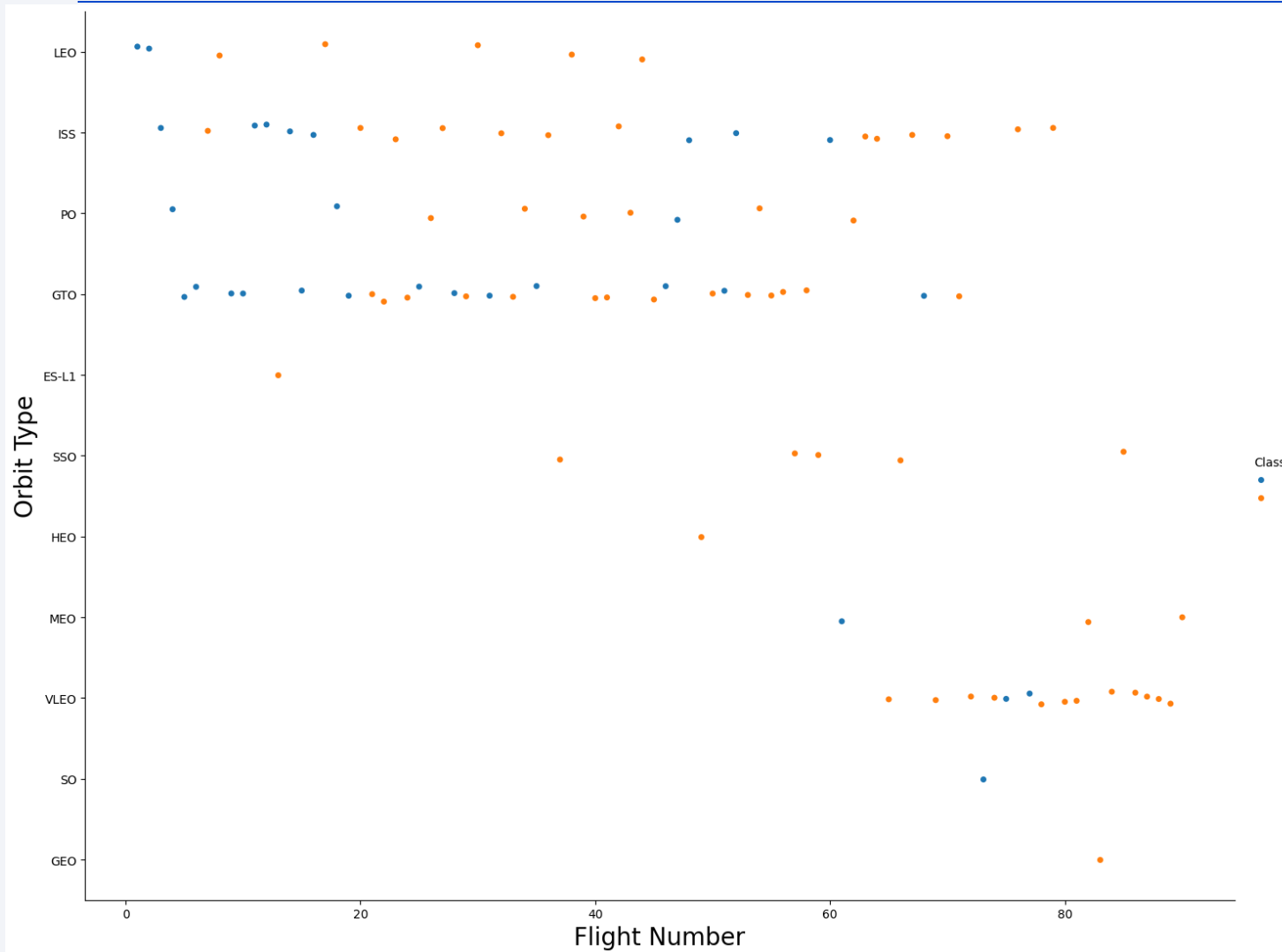
- **100% Success Rate:** ES-L1, GEO, HEO and SSO
- **50%-90% Success Rate:** VLEO, LEO, MEO, PO, ISS, GTO
- **0% Success Rate:** SO

The choice of orbit plays a pivotal role in mission success, and SpaceX's [strategic selection of orbits](#) has led to impressive results in certain categories.



Insights from this data can guide future mission planning and emphasize the importance of orbit-specific considerations.

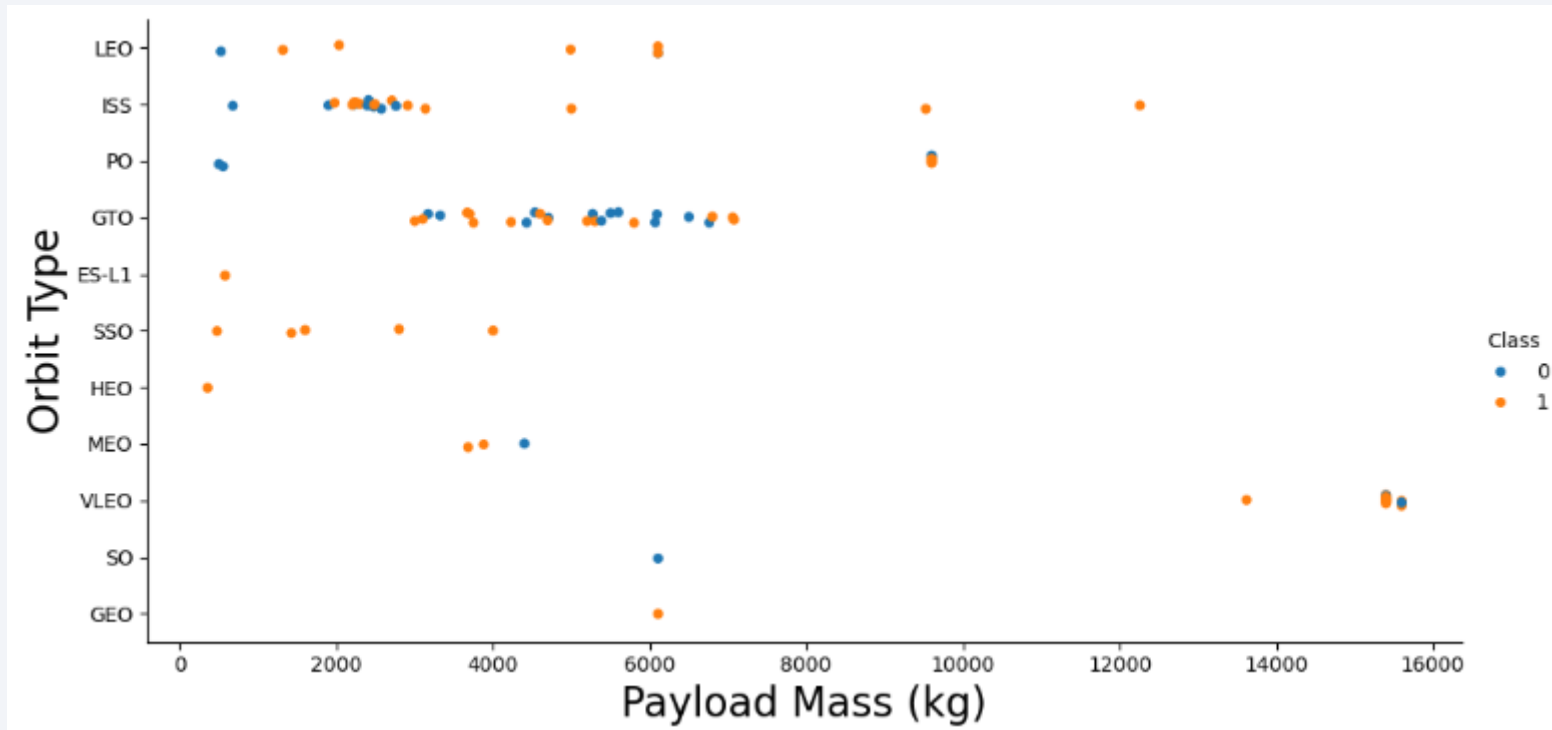
Flight Number vs. Orbit Type



- **General Trend:** Across various orbits, the **success rate** tends to **increase** with the **number of flights**.
- **Highlighting LEO Orbit:** This relationship is highly apparent for the LEO orbit, where **more flights** are associated with a **higher success rate**.
- **GTO Orbit Discrepancy:** In contrast, the GTO orbit deviates from this pattern, showing **no strong correlation** between both attributes.

These insights provide valuable guidance for mission **planning** and emphasize the importance of understanding the dynamics of **different orbits**

Payload vs. Orbit Type



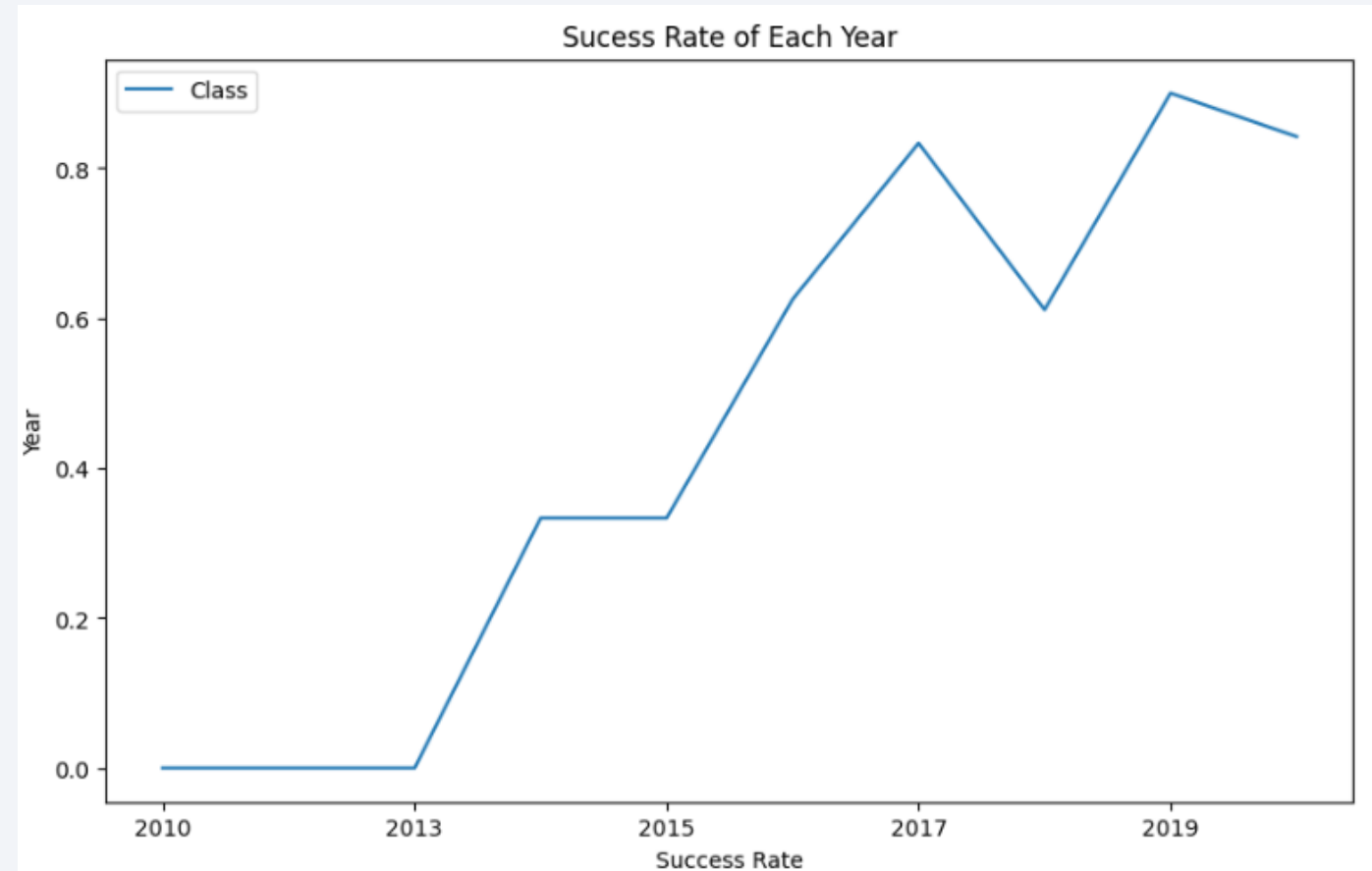
Payloads > 4,000 kg tend to have higher success rates in in **LEO**, **SSO** orbits

Heavier payloads: positive landing rate are more for **ISS** orbits.

Payload selection should align with the target orbit for **optimal mission** success.
Further analysis of GTO missions with heavy payloads is needed to improve success rates.

Launch Success Yearly Trend

- The success rate exhibited increased in 2013-2017.
- However, there was a decline in success in 2017-2018 and 2019-2020.
- **Overall Outlook:** Despite fluctuations, the overall success rate has shown improvement since 2013, indicating positive advancements in SpaceX missions.



All Launch Site Names

Execute SQL queries to get name of the unique launch sites:

```
SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

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

- **CCAFS LC-40** (Cape Canaveral Launch Complex 40)
- **VAFB SLC-4E** (Vandenberg Air Force Base Space Launch Complex 4E)
- **KSC LC-39A** (Kennedy Space Center Launch Complex 39A)
- **CCAFS SLC-40** (Cape Canaveral Space Launch Complex 40)

Launch Site Names Begin with 'CCA'

Get 5 records where launch sites begin with `CCA` use SQL query:

```
SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
6/4/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
12/8/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
10/8/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
3/1/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Calculate the total payload carried by boosters from NASA use SQL query:

```
SELECT Customer, SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

Customer	SUM(PAYLOAD_MASS__KG_)
NASA (CRS)	45596

This query provides insights into the [cumulative payload mass transported by boosters](#) in missions conducted for NASA (CRS). Understanding the total payload is crucial for [mission planning](#) and [resource allocation](#).

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1 use SQL query:

```
SELECT Booster_Version, AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.0%';
```

Booster_Version	AVG(PAYLOAD_MASS__KG_)
F9 v1.0 B0003	340.4

The query provides insights into the average payload mass carried by boosters of the specified version. This information is valuable for assessing the [performance and capabilities](#) of [booster versions](#) in [handling payloads](#).

First Successful Ground Landing Date

The first successful landing outcome on ground pad use SQL query:

```
SELECT Landing_Outcome, date
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (ground pad)'
AND date2 = (SELECT MIN(date2) FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (ground pad)');
```

Landing_Outcome	Date
Success (ground pad)	22/12/2015

The query's results indicate the [historical milestone](#) of the first successful landing of a rocket on a ground pad, providing valuable insights into [SpaceX's achievements](#) in [reusable rocket technology](#).

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_  
FROM SPACEXTBL  
WHERE LANDING_OUTCOME = 'Success (drone ship)'  
AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

The query's results provide information about boosters that successfully landed on drone ships with [specific payload mass ranges](#). This data is valuable for [analyzing the performance of boosters under these conditions](#).

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1021.1	3136
F9 FT B1022	4696
F9 FT B1023.1	3100
F9 FT B1026	4600
F9 FT B1029.1	9600
F9 FT B1021.2	5300
F9 FT B1029.2	3669
F9 FT B1036.1	9600
F9 FT B1038.1	475
F9 B4 B1041.1	9600
F9 FT B1031.2	5200
F9 B4 B1042.1	3500
F9 B4 B1045.1	362
F9 B5 B1046.1	3600

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes use SQL query:

```
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER  
FROM SPACEXTBL  
GROUP BY MISSION_OUTCOME;
```

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

The query identified the total number of mission outcomes, including successes and failures. It provides insights into the distribution of different mission outcomes. The [majority](#) of missions resulted in [success](#), with only a few instances of failures.

Boosters Carried Maximum Payload

List the names of the booster which have carried the maximum payload mass use SQL query:

```
SELECT DISTINCT BOOSTER_VERSION as 'booster_versions carried the max payload mass'
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL);
```

The query reveals **multiple booster versions** that have achieved the **maximum payload capacity**.

Various booster versions have demonstrated the ability to carry the maximum payload mass, **highlighting their significance in space missions**.

booster_versions carried the max payload mass
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

List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 use SQL query:

```
SELECT substr(Date2,6,2) as Month, substr(Date2,0,5) as Year,  
Landing_Outcome, BOOSTER_VERSION, LAUNCH_SITE  
FROM SPACEXTBL  
where Landing_Outcome = 'Failure (drone ship)' and Year='2015';
```

Month	Year	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

In 2015, there were 2 recorded cases of failed drone ship landings, providing valuable data for further analysis. This information can be used to study the causes and improvements needed for successful drone ship landings, ensuring the safety and success of space missions.

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

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order use SQL query:

```
SELECT LANDING_OUTCOME, COUNT(*) AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE DATE2 BETWEEN '2010-06-04' and '2017-03-20'
GROUP BY LANDING_OUTCOME
ORDER BY TOTAL_NUMBER DESC;
```

The query provides a ranked list of landing outcomes within the specified date range.

This ranking can help identify [trends in landing outcomes over time](#) and [guide improvements](#) in space missions for safer and more successful landings.

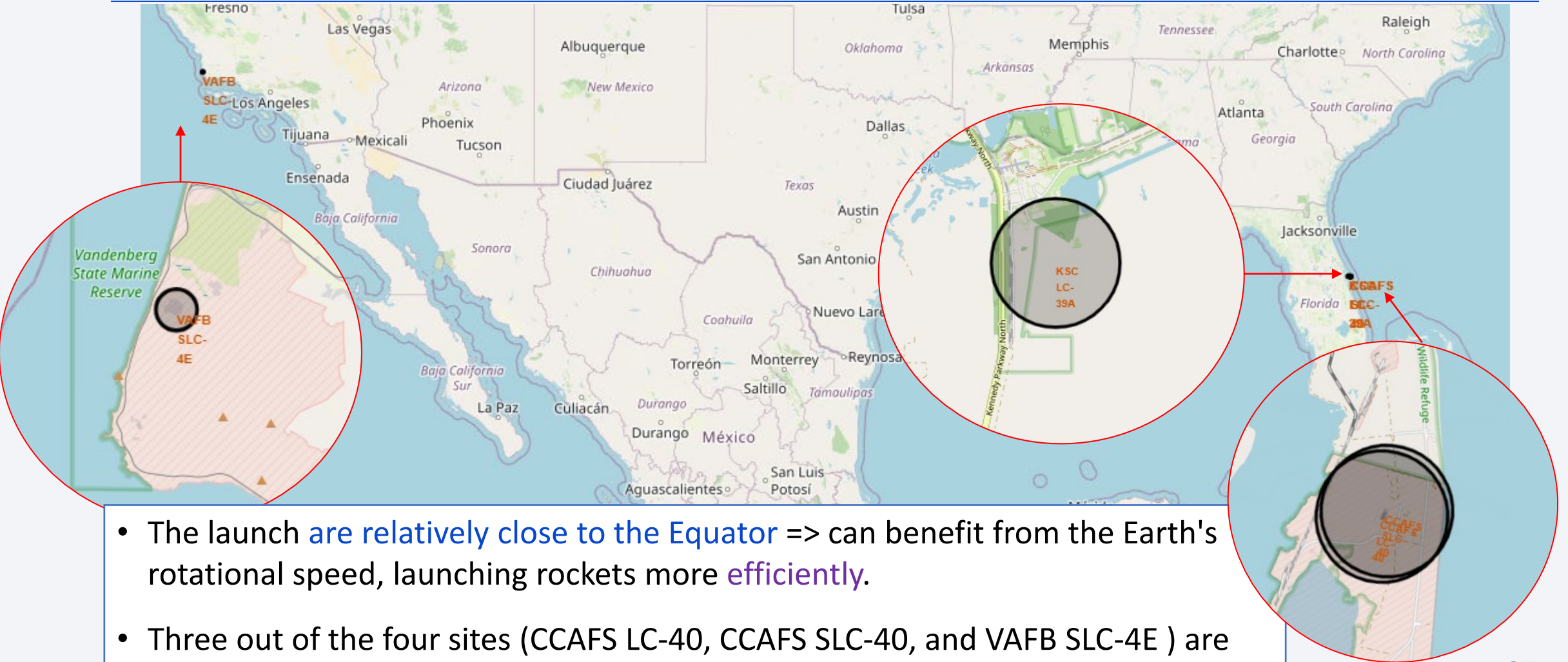
Landing_Outcome	TOTAL_NUMBER
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

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

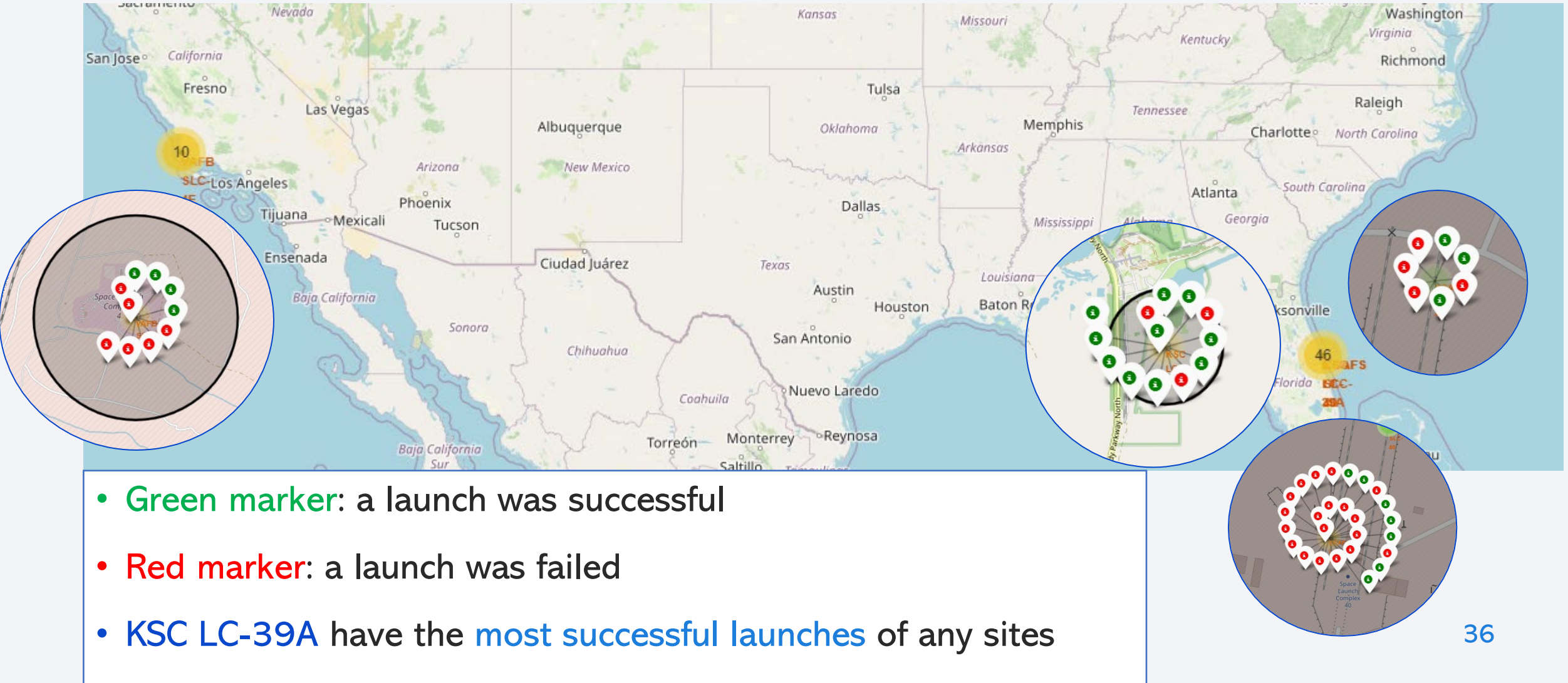
Launch Sites Proximities Analysis

All launch sites' location



- The launch **are relatively close to the Equator** => can benefit from the Earth's rotational speed, launching rockets more **efficiently**.
- Three out of the four sites (CCAFS LC-40, CCAFS SLC-40, and VAFB SLC-4E) are **close to the coast**. KSC LC-39A is located inland, but it's still relatively **close to the coast** => provide a **safe** area for rocket stages to fall back into the ocean.

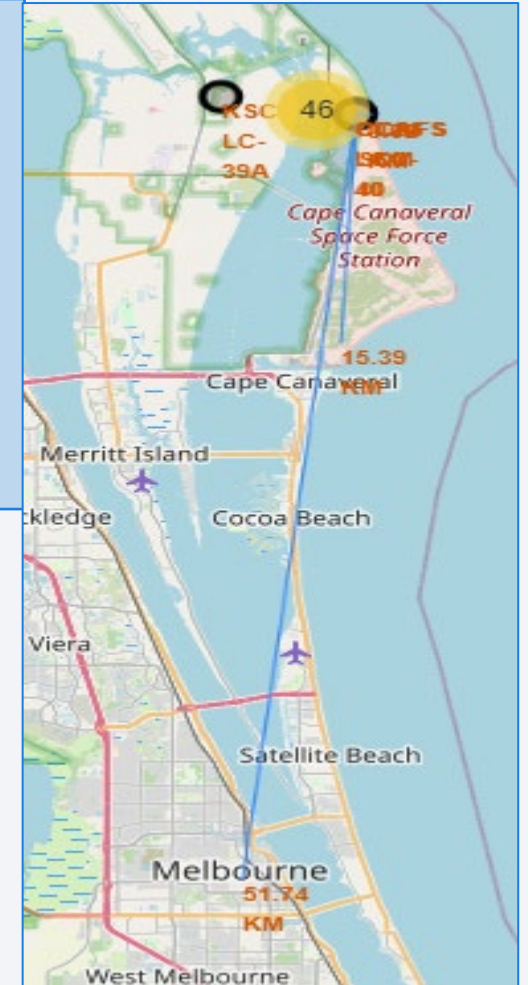
Launch outcomes for each site



The distance from launch site to Proximities



- **Railways:** 15.39 km
- **Highways:** 0.62 km
- **Coastline:** 0.86 km
- **Cities:** 51.74 km



- Launch sites may be close to railways and highways, potentially in proximity to the coastline.
- However, they consistently maintain a considerable distance from cities, aligning with safety and risk mitigation practices.



Section 4

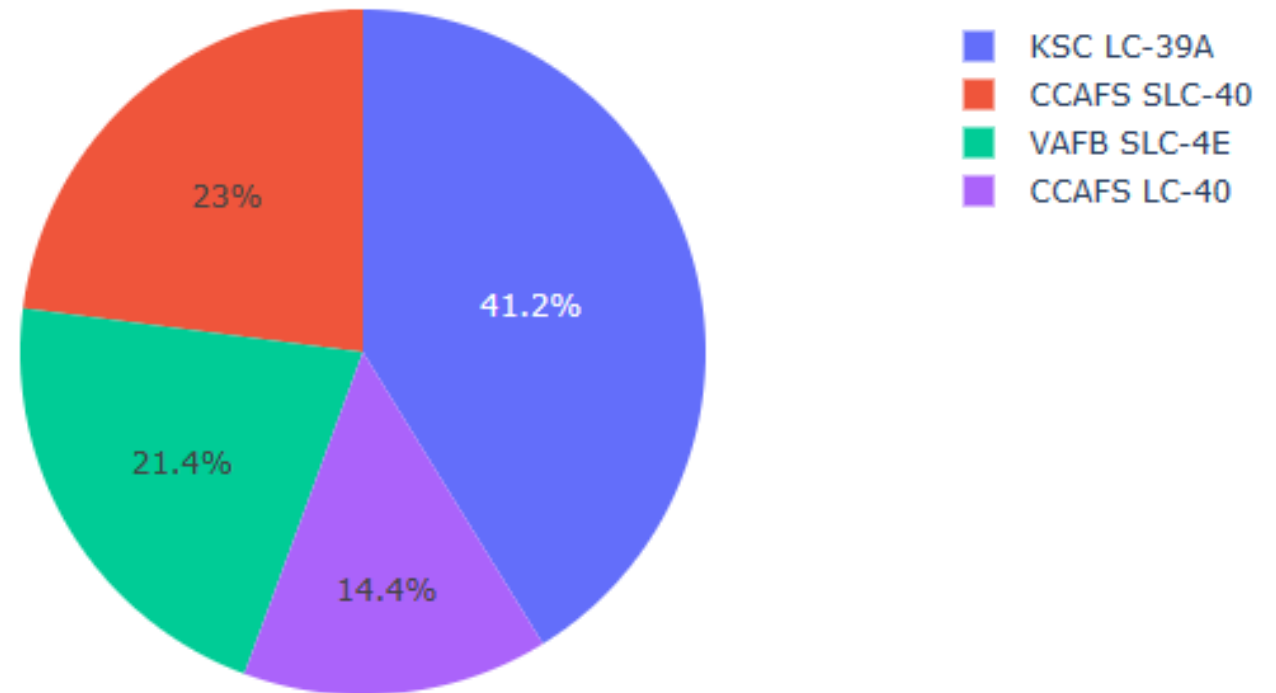
Build a Dashboard with Plotly Dash

Launch Success by Site

Success Rate

- KSC LC-39A: 41.2%
- CCAFS SLC-40: 23%
- VAFB SLC-4E: 21.4%
- CCAFS LC-40: 14.4%

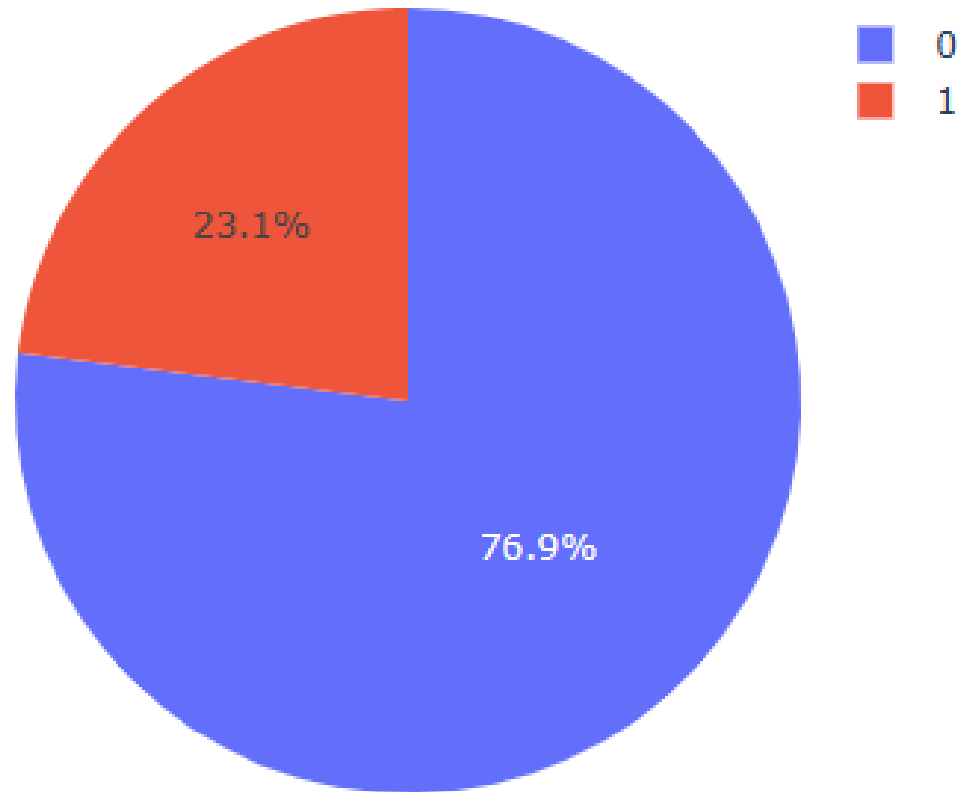
Total Success Launches by Site



KSC LC-39A has the most successful launches amongst all launch sites

KSC LC-39A - highest launch success ratio

Total Success Launches for Site KSC LC-39A



KSC LC-39A:

- Success Rate: 76.9%
- Failure Rate: 23.1%

KSC LC-39A has a relatively **high success rate**(3.33:1), indicating its **reliability** for successful launches.

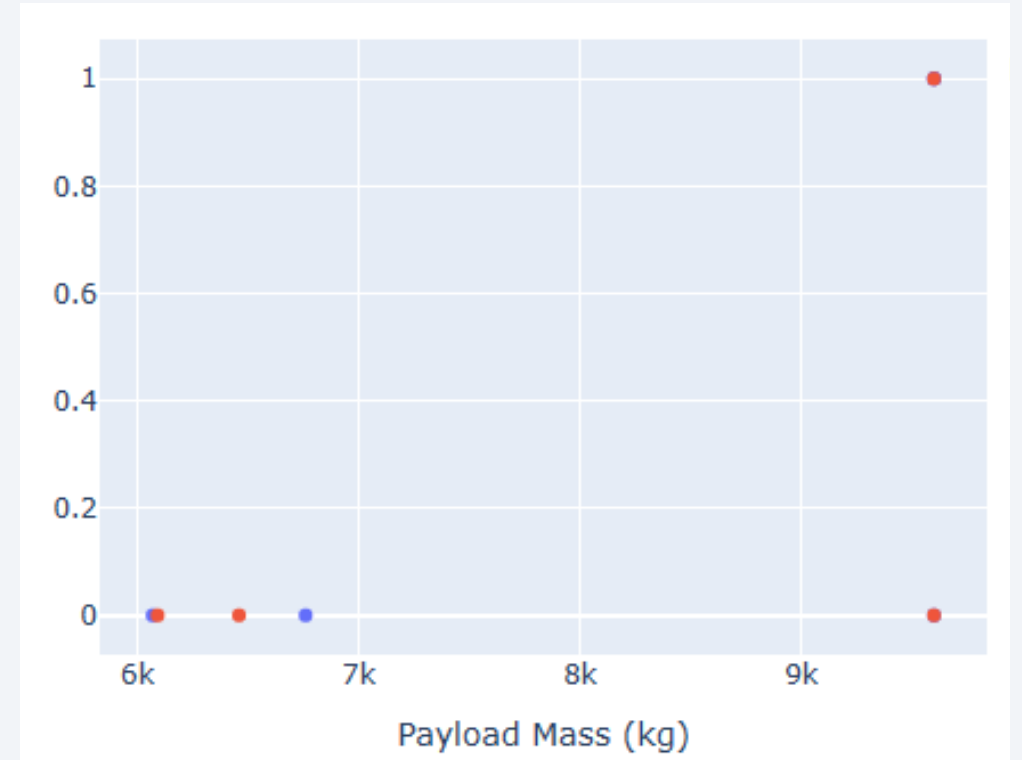
Analyzing the causes of failures can help further improve success rates and mission outcomes.

Payload vs Launch Outcome

Under 6.000kg



6.000-10.000kg



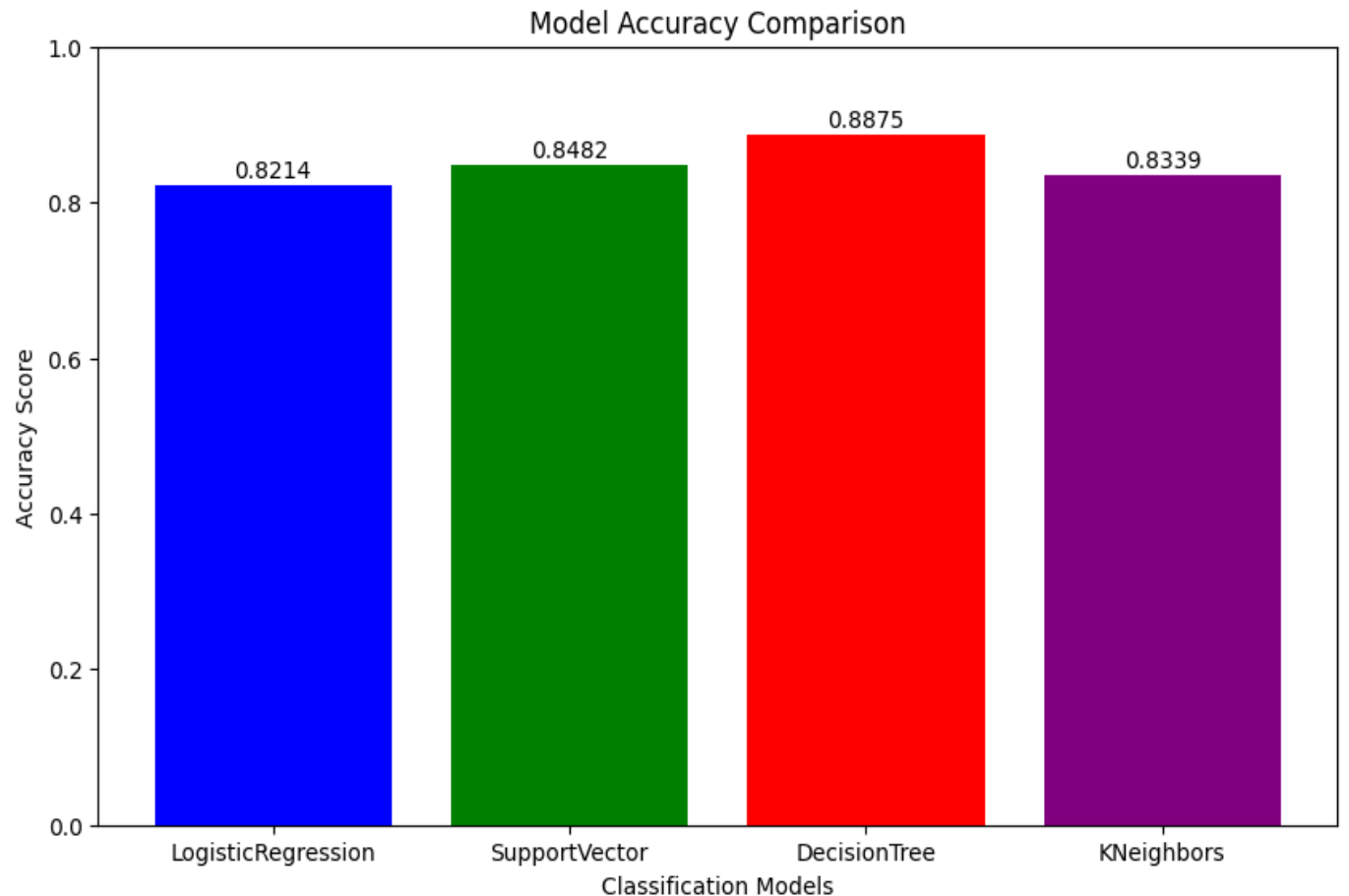
Section 5

Predictive Analysis (Classification)

Classification Accuracy

Best params is :

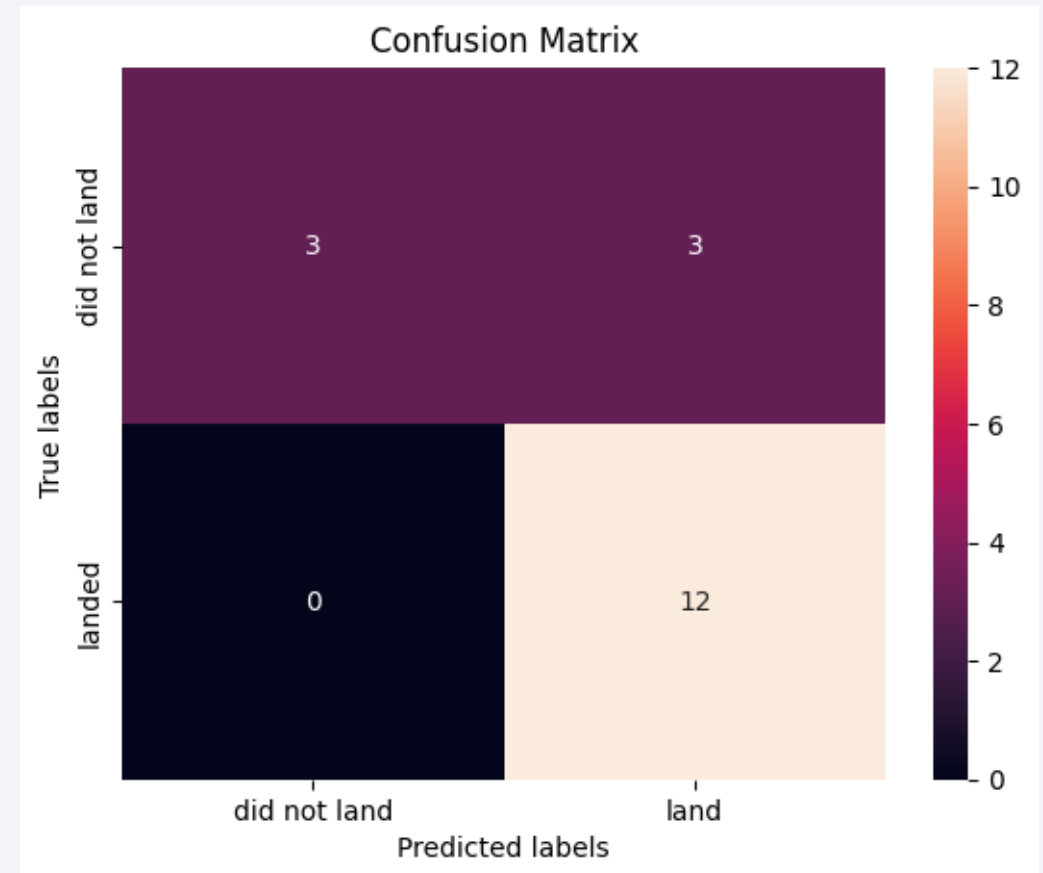
```
{'criterion': 'gini',  
'max_depth': 18,  
'max_features': 'sqrt',  
'min_samples_leaf': 2,  
'min_samples_split': 10,  
'splitter': 'best'}
```



Decision Tree model has the highest classification accuracy (~0.89)

Confusion Matrix

- 3 instances were correctly predicted as "Not Landed" (True Negative).
- 12 instances were correctly predicted as "Landed" (True Positive).
- 3 instances were falsely predicted as "Landed" when they were actually "Not Landed" (False Positive).
- 0 instances were falsely predicted as "Not Landed" when they were actually "Landed" (False Negative).



The **decision tree classifier** model seems to perform well, especially in terms of recall (ability to capture "Landed" instances). However, the specificity is lower, indicating that the model is less accurate in identifying "Not Landed" instances.

Conclusions

Optimal Algorithm: **The Decision Tree Classifier** Algorithm emerges as the most effective machine learning approach for this dataset.

Payload Weight Impact: **Launches with low-weight payloads** (defined as 4000kg and below) demonstrated superior performance compared to heavier payloads.

Success Rate Trend: From 2013 onward, SpaceX's launch success rate has shown a trend increase, suggesting a positive trajectory over the years. This trend indicates a potential for further **improvements in future launches**.

Top Launch Site: **KSC LC-39A** stands out with the highest success rate among all launch sites, boasting an impressive 76.9% success rate.

Orbit Excellence: **The SSO orbit** exhibits exceptional success, achieving a perfect 100% success rate with multiple occurrences, showcasing its reliability as a launch destination.

Thank you!

