IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Laszlo Pal Danyi
28-JUL-2024

Outline

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

Executive Summary

- **Summary of methodologies**
- Data collection through:
  - API
  - Web scraping
- Data Wrangling
- Exploratory Data Analysis (EDA) with:
  - SQL
  - Data Visualization
- Interactive GEO Data visualization using Folium Map
- Machine Learning (ML) Algorithms for Prediction
- **Summary of all results**
- Exploratory Data Analysis results
- Interactive Chart/Visualization results
- Predictive Analysis results

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
- Determine the factors that most significantly influence:
    - Whether the Falcon 9 first stage will land successfully
    - The various factors of a successful mission.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected through SpaceX API and web scraping from Wikipedia
- Perform data wrangling
  - Data was cleansed, organized and applied one-hot encoding for feature classification.
- 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 Collection Process:**

1. Initial Research:

1. Identified key data sources and gathered preliminary information.
2. Determined the scope of data needed for analysis (e.g., launch dates, outcomes, payload details).

2. API Integration:

1. Connected to the SpaceX API to fetch data on all historical launches.
2. Parsed and stored the data in a structured format (e.g., JSON, CSV).

3. Web Scraping:

1. Used Beautiful Soup to scrape data from Wikipedia about Falcon 9 launches.
2. Extracted relevant details such as launch site information, mission descriptions, and booster versions.
3. Converted the extracted data into a Pandas DataFrame for further processing.

4. Data Cleaning:

1. Cleaned and preprocessed the collected data to ensure consistency.
2. Handled missing values, removed duplicates, and standardized formats.

Data Collection – SpaceX API

- We collected data using SpaceX REST API calls, integrating and processing it into a clean, comprehensive dataset using Pandas. The workflow included API integration, data extraction, cleaning, merging, and validation to ensure data accuracy and completeness.
- https://github.com/Haseulhoff/SpaceX-FinalAssignment/blob/main/jupyter-labs-spacex-data-collection-api%20(SPCXPT1)-LPD.ipynb

Task 1: Request and parse the SpaceX launch data using the GET request

To make the requested JSON results more consistent, we will use the following static response object for this project:

```
[9]: static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

```
[19]: # Use json_normalize meethod to convert the json result into a dataframe
json_data = response.json()
df = pd.json_normalize(json_data)
```

Using the dataframe `data` print the first 5 rows

```
[20]: # Get the head of the dataframe
df.head(5)
```

Data Collection - Scraping

- We employed web scraping using Beautiful Soup to extract Falcon 9 launch data from Wikipedia. The process involved parsing HTML content, extracting relevant information, handling missing values, and formatting the data into a structured Pandas DataFrame.
- https://github.com/Haseulhoff/SpaceX-FinalAssignment/blob/main/jupyter-labs-webscraping%20(spcxpt1.2)-LPD.ipynb

[6]:
```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
```

Next, request the HTML page from the above URL and get a `response` object

TASK 1: Request the Falcon9 Launch Wiki page from its URL

First, let's perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.

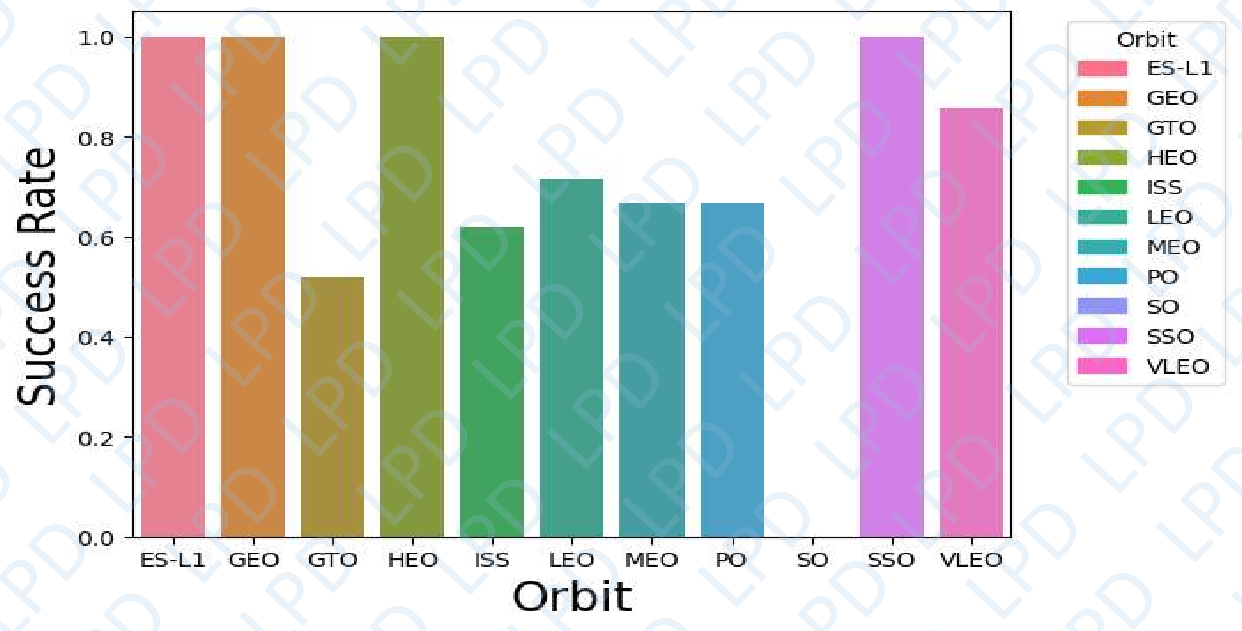
[8]:
```
# use requests.get() method with the provided static_url
# assign the response to a object

# Perform the HTTP GET request and assign the response to an object
response = requests.get(static_url)

# Check the status code to ensure the request was successful
if response.status_code == 200:
    print("Request was successful")
    # Create a BeautifulSoup object from the response text content
    soup = BeautifulSoup(response.text, 'html.parser')
else:
    print(f"Failed to retrieve the page. Status code: {response.status_code}")

# Print the title of the page to verify
print(soup.title.string)
```

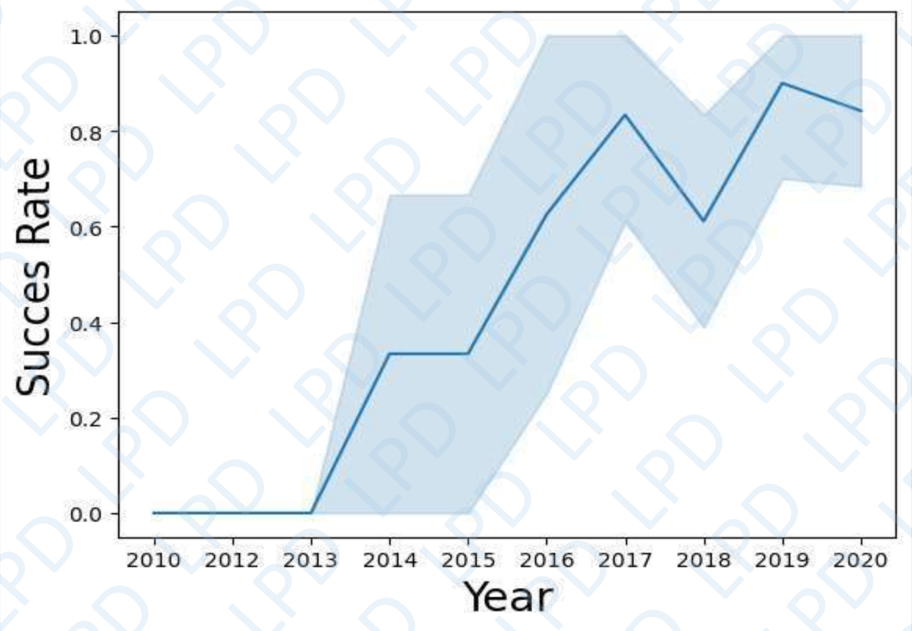
```
Request was successful
List of Falcon 9 and Falcon Heavy launches - Wikipedia
```

Data Wrangling

- Data wrangling involved cleaning and transforming the collected data to ensure it was suitable for analysis.
- This process included handling missing values, standardizing formats, converting data types, and creating new variables.
- Additionally, we computed the number of launches by each launch site and the occurrences of different orbits to gain more insights.
- https://github.com/Haseulhoff/SpaceX-FinalAssignment/blob/main/labs-jupyter-spacex-Data%20wrangling%20(spcxpt2.0)-LPD.ipynb

EDA with Data Visualization





We deep dived into details and analyzed certain features, such as Launch Sites, Orbits, Payload Mass trends of outcomes and for better insights we created charts that allows us to gain knowledge.

https://github.com/Haseulhoff/SpaceX-FinalAssignment/blob/main/edadataviz(spcxpt4.1)-LPD.ipynb

EDA with SQL

Extracted relevant columns from the SpaceX launch dataset.
•Retrieved records for specific launch sites.
•Counted the number of successful and failed launches.
•Summarized the total payload mass for each launch site.
•Extracted records for specific orbit types.
•Retrieved the number of launches for each launch site and orbit combination.
•Sorted the dataset by payload mass in descending order.
•Filtered payload mass records within a specific date range.

- https://github.com/Haseulhoff/SpaceX-FinalAssignment/blob/main/jupyter-labs-eda-sql-coursera_sqllite-(spcxpt3)-LPD.ipynb

Build an Interactive Map with Folium

Map Objects Added to Folium Map:

- **Markers and circles:**
  - **Classification:** Markers and circles were used to indicate launch sites with distinct colors or symbols based on the classification of launch success (0 for unsuccessful, 1 for successful).
  - **Rationale:** To visually distinguish between successful and unsuccessful launches at each site, aiding in the analysis of launch performance.
- **Lines:**
  - **Distance Calculation:** Lines were drawn to show distances between launch sites and various proximities such as highways, railways, and coastlines.
  - **Rationale:** To illustrate the geographical context and proximity of launch sites to critical infrastructure, aiding in the assessment of logistical and environmental factors affecting launch success.

- https://github.com/Haseulhoff/SpaceX-FinalAssignment/blob/main/lab_jupyter_launch_site_location-(spcxpt4.2)LPD.ipynb

Build a Dashboard with Plotly Dash

- **Dropdown List for Launch Site Selection:**
  - **Description:** Allows users to select a specific launch site or view data for all sites.
  - **Rationale:** Enables focused analysis on individual sites or comparisons across all sites.
- **Pie Chart for Launch Success Counts:**
  - **Description:** Displays total launch counts by site or success/failure counts for a selected site.
  - **Rationale:** Provides a clear visual representation of launch distributions and success rates.
- **Range Slider for Payload Selection:**
  - **Description:** Allows users to select a payload range (kg).
  - **Rationale:** Facilitates analysis of how different payload ranges affect launch success rates.
- **Scatter Chart for Payload vs. Launch Success:**
  - **Description:** Shows the correlation between payload mass and launch success, with points colored by booster version.
  - **Rationale:** Visualizes the relationship between payload size and launch outcomes.

- https://github.com/Haseulhoff/SpaceX-FinalAssignment/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Model Development Process:

Developed, tested, and optimized models (Logistic Regression, SVM, Decision Tree, KNN) to predict Falcon 9 first stage landing success using GridSearchCV. Split data into training and test sets and calculated accuracy scores.

Standardized features and optimized hyperparameters for best performance.

https://github.com/Haseulhoff/SpaceX-FinalAssignment/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5-LPD.ipynb

Results

•**Exploratory Data Analysis:** Analyzed launch success rates, payload distributions, and launch site frequencies using visualizations.

•**Interactive Analytics Demo:** Created interactive dashboards with pie charts and scatter plots to showcase launch outcomes.

•**Predictive Analysis:** Built and evaluated multiple classification models to identify the best-performing one.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

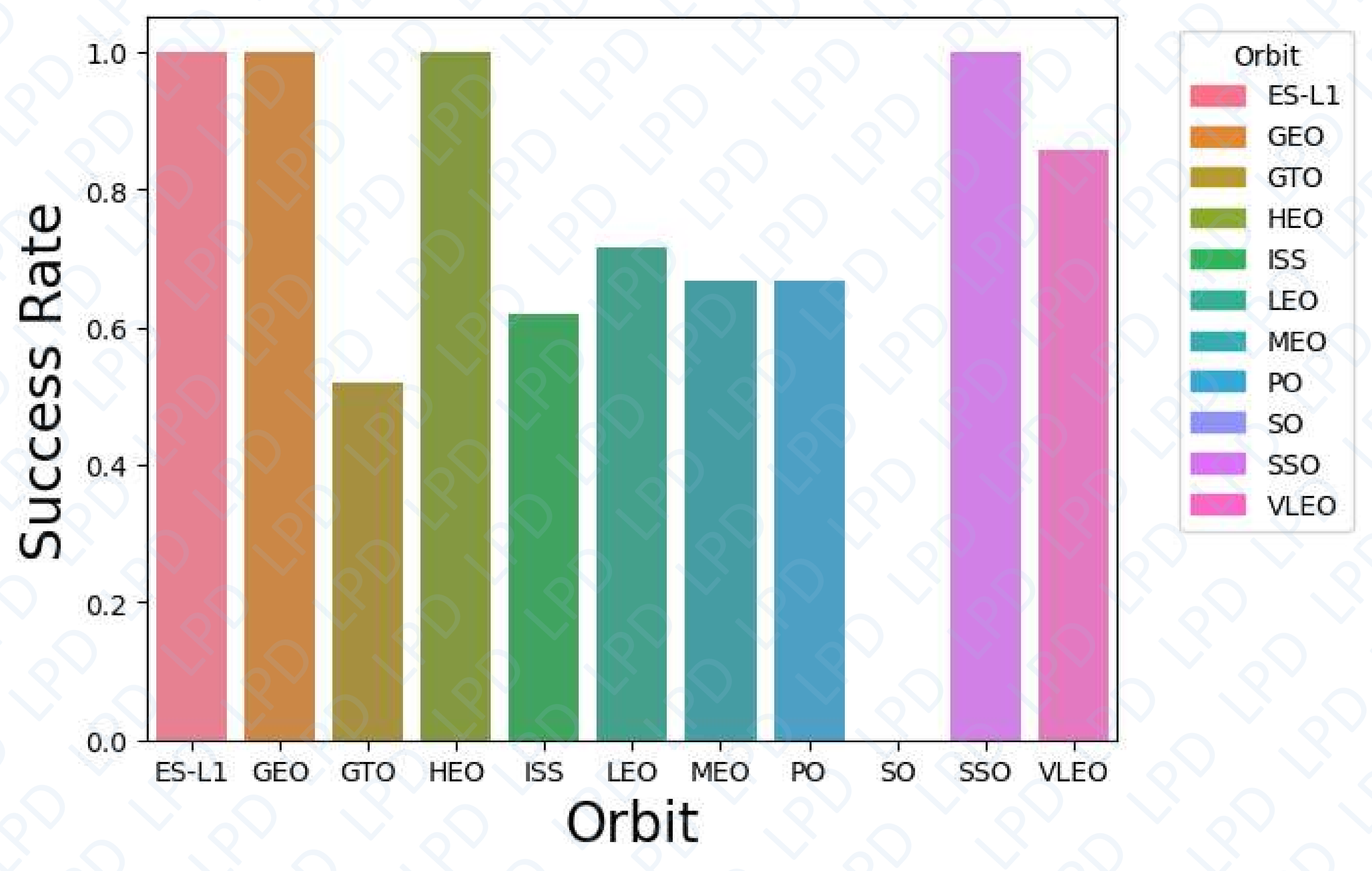
- In this plot graph, we can conclude that the higher the Flight Number at a certain Launch Site, the higher the chance for success.
- We can also see that at Launc Site ‘CCAFS SLC 40’ is where Falcon 9 launches started, in which we see many non-successful outcome, during the beginnings. This will significantly influence the summary of success rate for each Launch Site.
- We also note that ‘KSC LC 39A’ Launch Site is where SpaceX commenced Falcon 9 launches the latest.

[SEE PAGE 39 FOR LAUNCH SITE SUCCES RATE PIE-GRAPH]

Payload vs. Launch Site

- As we know the time line of each Launch site we see that the Low-Middle Weighted Payloads success rate is varies
- However, we see that over ~9500 KG Payload there are clusters of high success rate.

Success Rate vs. Orbit Type

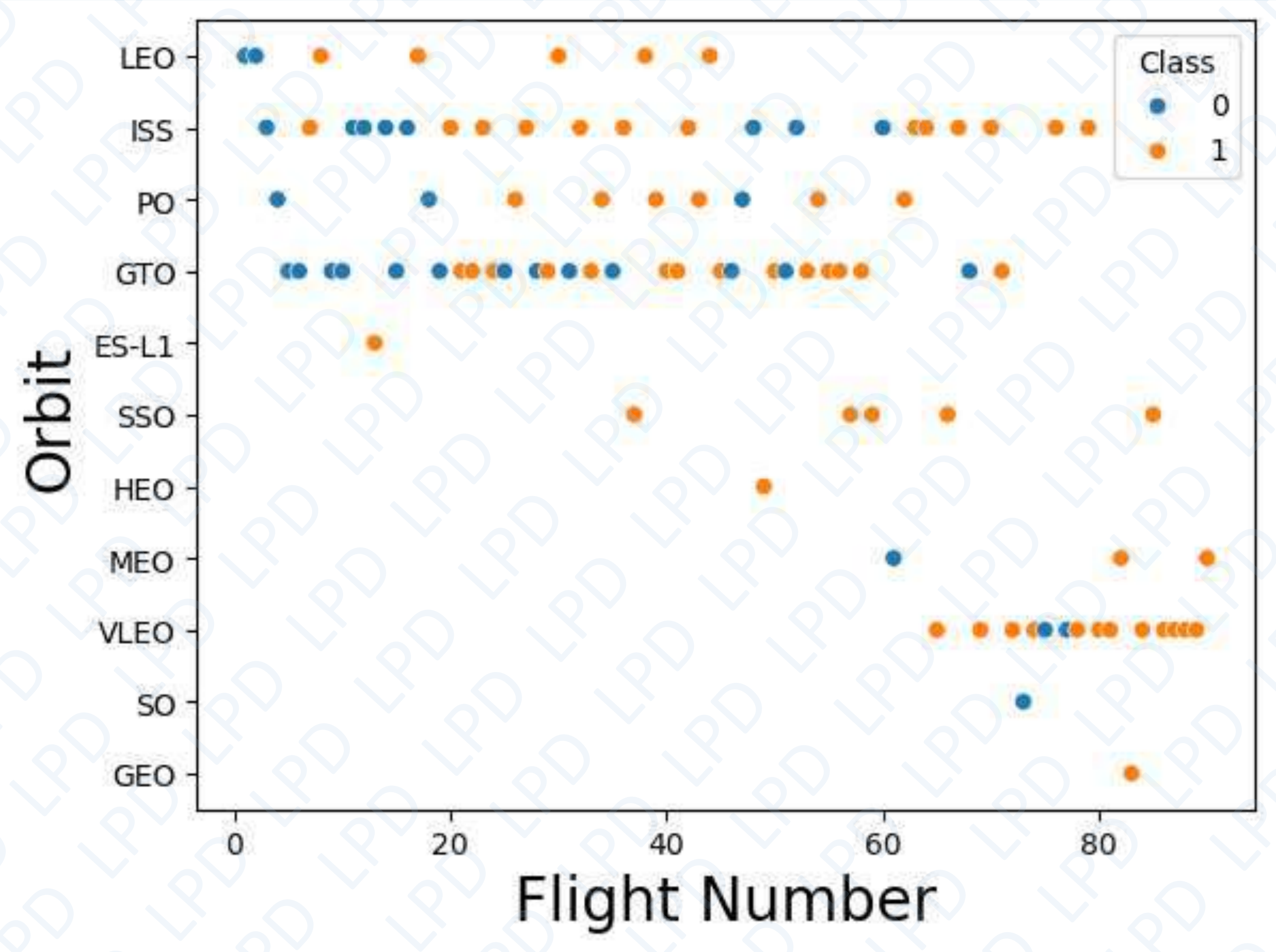


Considering the Orbit Type as a factor of Success we can see that

- ES-L1
- GEO
- HEO
- SSO

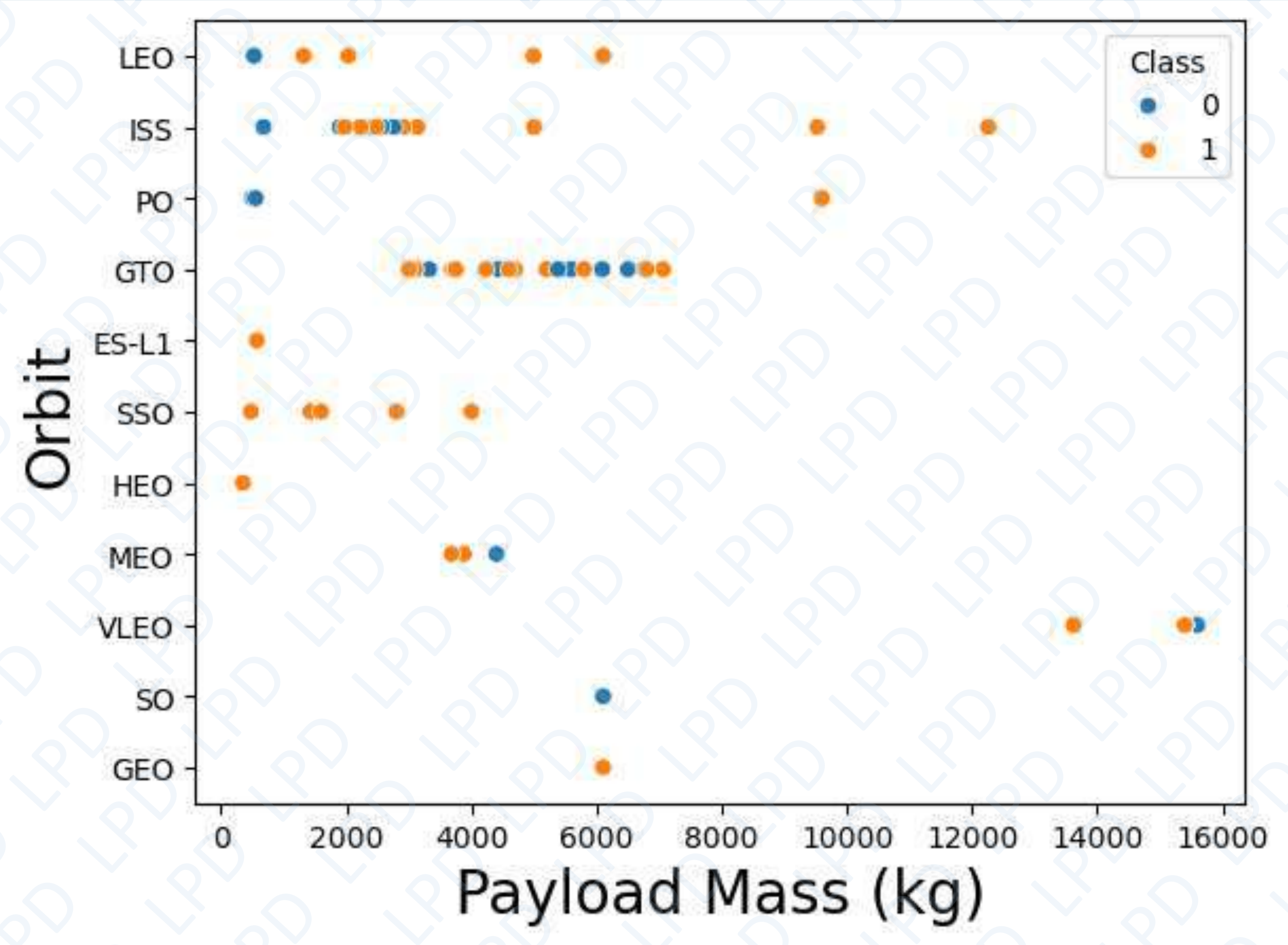
have 100% success rate

Flight Number vs. Orbit Type



- As it shown, **ISS** launches have a clear pattern of gaining experience by going through several failures lead to more successful outcome, while **GTO** Orbit missions have the most failures.

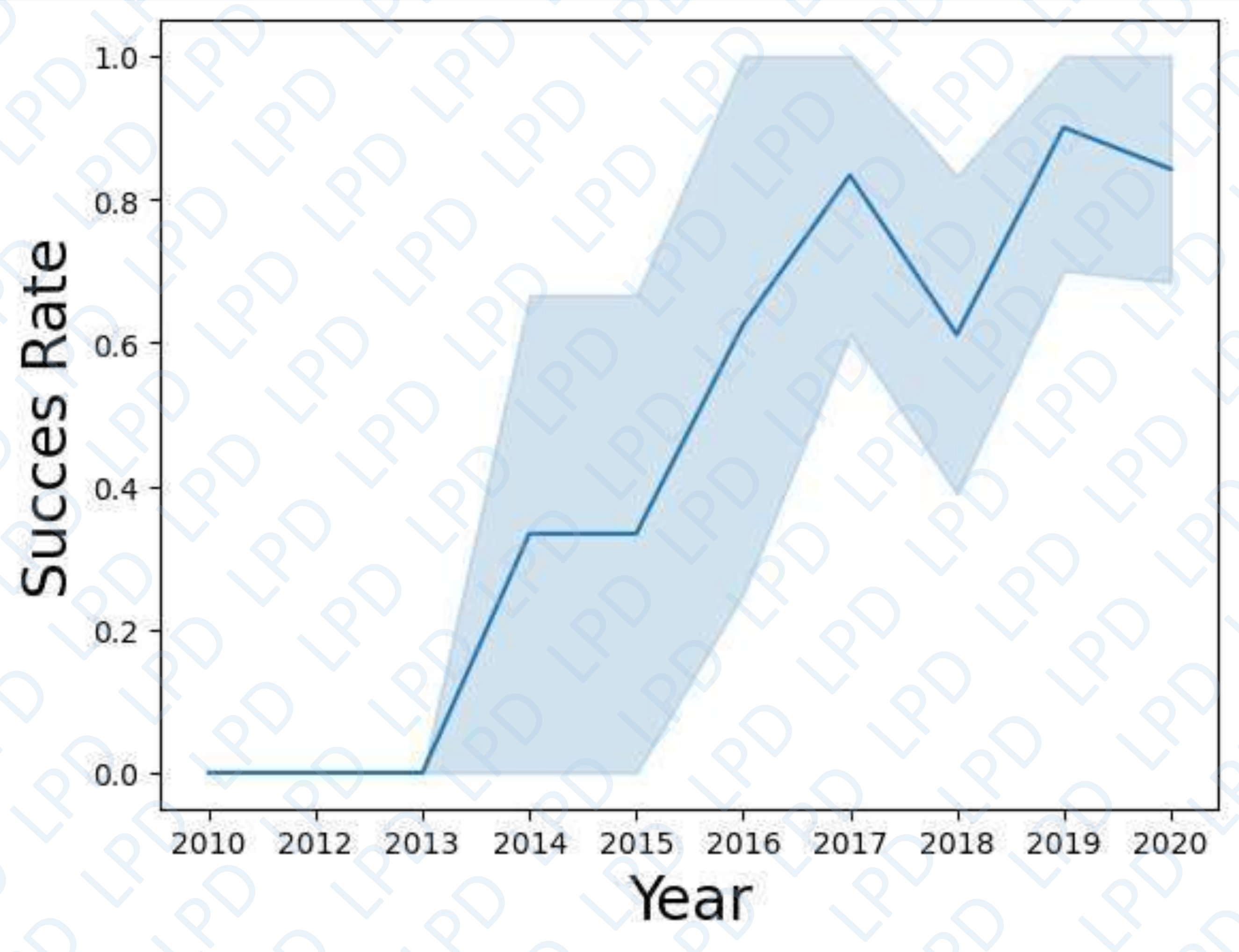
Payload vs. Orbit Type



VLEO orbit missions have the heaviest payloads, While SSO mission have one of the lowest.

GTO Orbit Missions with lower weighted payloads have more success rate than their heavily weighted payloads. We can see a pattern where we can cut the payload ranges visually into two sides.

Launch Success Yearly Trend



- There is a clear trend of improvement with a significant decline in 2018 and a bounce back in 2019.

All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

```
%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

With using SELECT DISTINCT, we retrieve all the Unique values/names of Launch_Site column.

Launch Site Names Begin with 'CCA'

Task 2

Display 5 records where launch sites begin with the string 'CCA'

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

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

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

To find all Launch Sites that begin with **'CCA'**, use the SQL LIKE function with the **%** wildcard.

The **%** symbol represents any sequence of characters, so **'CCA%'** matches any string starting with **'CCA'**

Total Payload Mass

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%%sql
SELECT Customer, SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)'
```

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

Customer	SUM(PAYLOAD_MASS__KG_)
NASA (CRS)	45596

We use the **SUM** function to calculate the total **Payload Mass**, and the WHERE clause to specify the criteria for **NASA** on **Customer** column.

Average Payload Mass by F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

```
%%sql
SELECT Booster_Version, AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTABLE
WHERE Booster_Version LIKE 'F9 v1.1%'
```

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

Booster_Version	AVG(PAYLOAD_MASS__KG_)
F9 v1.1 B1003	2534.6666666666665

- **SELECT Booster_Version, AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass**: Selects the Booster_Version and calculates the average of PAYLOAD_MASS__KG_. The result is aliased as Average_Payload_Mass.
- **FROM SPACEXTABLE**: Specifies the table from which to retrieve the data (SPACEXTABLE in this case).
- **WHERE Booster_Version LIKE 'F9 v1.1%'**: Filters the rows to include only those where the Booster_Version starts with 'F9 v1.1'.
- **GROUP BY Booster_Version**: Groups the results by Booster_Version.

First Successful Ground Landing Date

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%%sql
SELECT MIN(DATE) FROM SPACEXTABLE
WHERE landing_outcome = 'Success (ground pad)'
```

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

MIN(DATE)
2015-12-22

By using **MIN(DATE)** and the **WHERE Clause** on landing_outcome, we retrieve the first Successful landing outcome on ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome LIKE '%Success%'
  AND Landing_Outcome LIKE '%drone ship%'
  AND PAYLOAD_MASS__KG_ > 4000
  AND PAYLOAD_MASS__KG_ < 6000;
```

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

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

- Using **% Wildcard** to retrieve any Success drone ship outcome
- adding AND Conditions WHERE values are < less than or > greater than certain numbers will retrieve the names of Booster versions where Payload mass is > greater than 4k and < less than 6k

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
%%sql
SELECT
    Mission_Outcome,
    COUNT(*) AS Total_Count
FROM
    SPACEXTABLE
GROUP BY
    Mission_Outcome;
```

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

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

- **SELECT Mission_Outcome, COUNT(*) AS Total_Count**: This line selects the Mission_Outcome column and counts the number of occurrences for each unique mission outcome. The result is aliased as Total_Count.
- **FROM launches**: Specifies the table from which to retrieve the data (launches in this case).
- **GROUP BY Mission_Outcome**: Groups the results by each unique value in the Mission_Outcome column.

Boosters Carried Maximum Payload

```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
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

•This query retrieves the **Booster_Version** values for records where the **PAYLOAD_MASS__KG_** is equal to the maximum payload mass.

•The **MAX** function is used in a subquery to find the highest value of PAYLOAD_MASS__KG_ in the SPACEXTABLE.

2015 Launch Records

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

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

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

- This query converts the numerical month from the **Date** column to the corresponding month name using a CASE statement.
- It selects the **Booster_Version, Launch_Site**, and **Landing_Outcome** columns.
- The WHERE clause filters the records to only include those from the year 2015 and with a landing outcome of 'Failure (drone ship)'.

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

```
%%sql
SELECT
    Landing_Outcome,
    COUNT(*) as Outcome_Count,
    RANK() OVER (ORDER BY COUNT(*) DESC) as Outcome_Rank
FROM
    SPACEXTABLE
WHERE
    Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY
    Landing_Outcome
ORDER BY
    Outcome_Count DESC;
```

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

Landing_Outcome	Outcome_Count	Outcome_Rank
No attempt	10	1
Success (drone ship)	5	2
Failure (drone ship)	5	2
Success (ground pad)	3	4
Controlled (ocean)	3	4
Uncontrolled (ocean)	2	6
Failure (parachute)	2	6
Precluded (drone ship)	1	8

- SELECT Landing_Outcome, COUNT(*) as Outcome_Count: Selects the Landing_Outcome and counts the number of occurrences for each landing outcome.
- RANK() OVER (ORDER BY COUNT(*) DESC) as **Outcome_Rank**: Assigns a rank to each landing outcome based on the count, in descending order.
- FROM launches: Specifies the table from which to select the data.
- WHERE **Date** BETWEEN '2010-06-04' AND '2017-03-20': Filters the rows to include only those between the specified dates.
- GROUP BY Landing_Outcome: Groups the results by Landing_Outcome to count each unique outcome.
- ORDER BY Outcome_Count DESC: Orders the results by the count in descending order.

Section 3

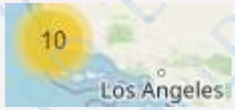
Launch Sites Proximities Analysis

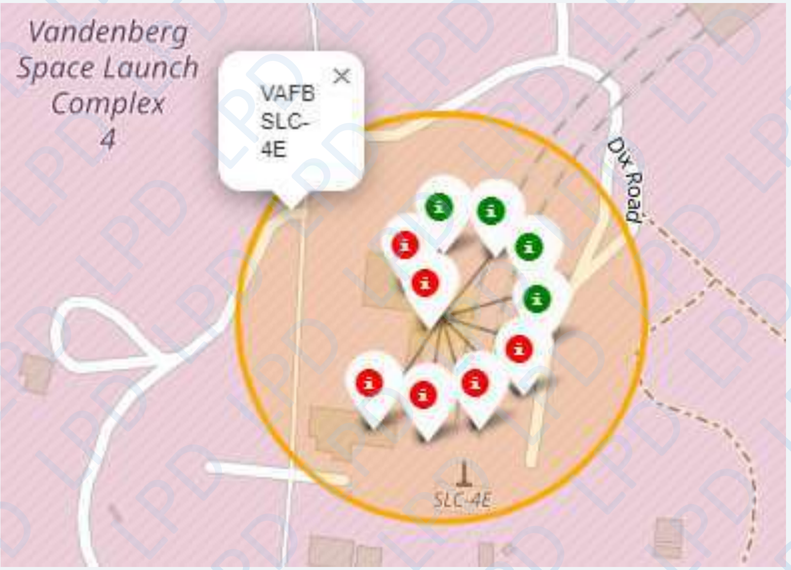
Global Map marked with All Launch sites.

Falcon 9 Launch sites are located in the United States, such as in Florida (FL) and in California (CA).

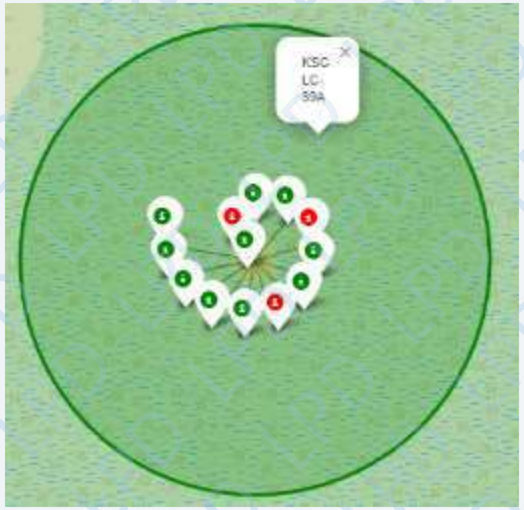
Map of Launch sites and their success rate markers

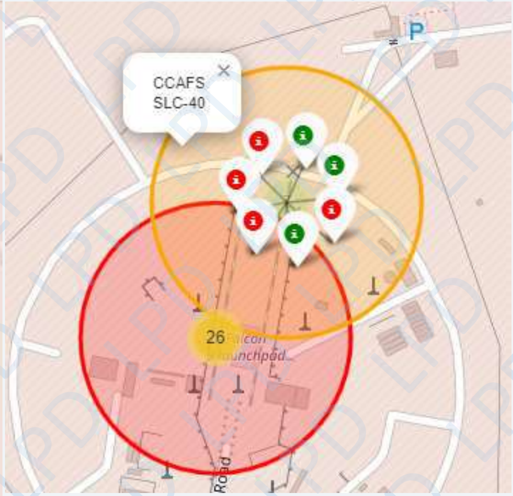
CA. Launch Site





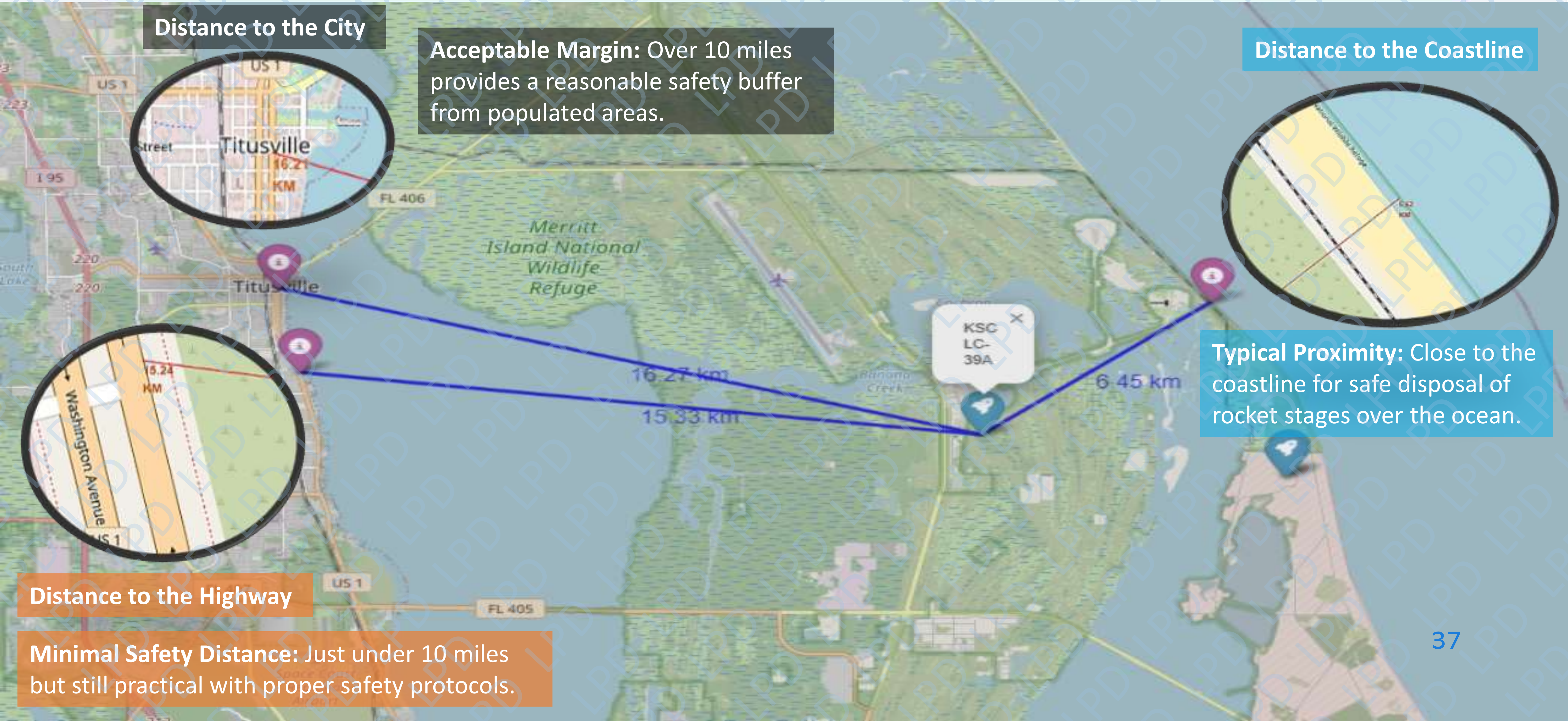
FL. Launch Site





- There are 10 launches count in California and 46 Launches count in Florida.
- Circles are representing the location of **Launch sites** colored by Average Success rate.
- Low Success rate with RED, Moderate Success rate with ORANGE and High Success Rate with GREEN
- Markers and their colors representing each Launch indicating either Successful or Failed outcomes.

Distance from Launch Sites (KSC-LC39A)



Distance to the City

Acceptable Margin: Over 10 miles provides a reasonable safety buffer from populated areas.

Distance to the Coastline

Typical Proximity: Close to the coastline for safe disposal of rocket stages over the ocean.

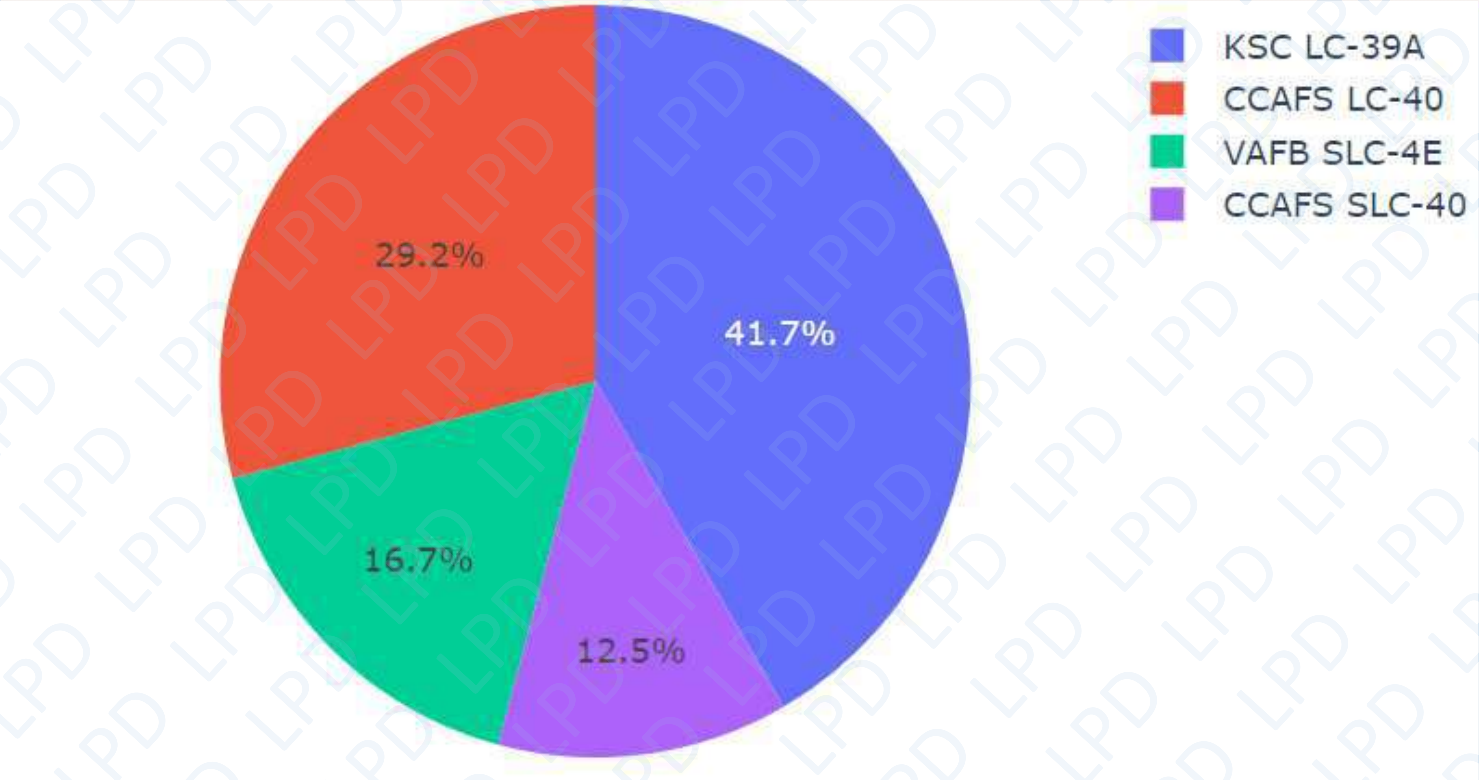
Distance to the Highway

Minimal Safety Distance: Just under 10 miles but still practical with proper safety protocols.

Section 4

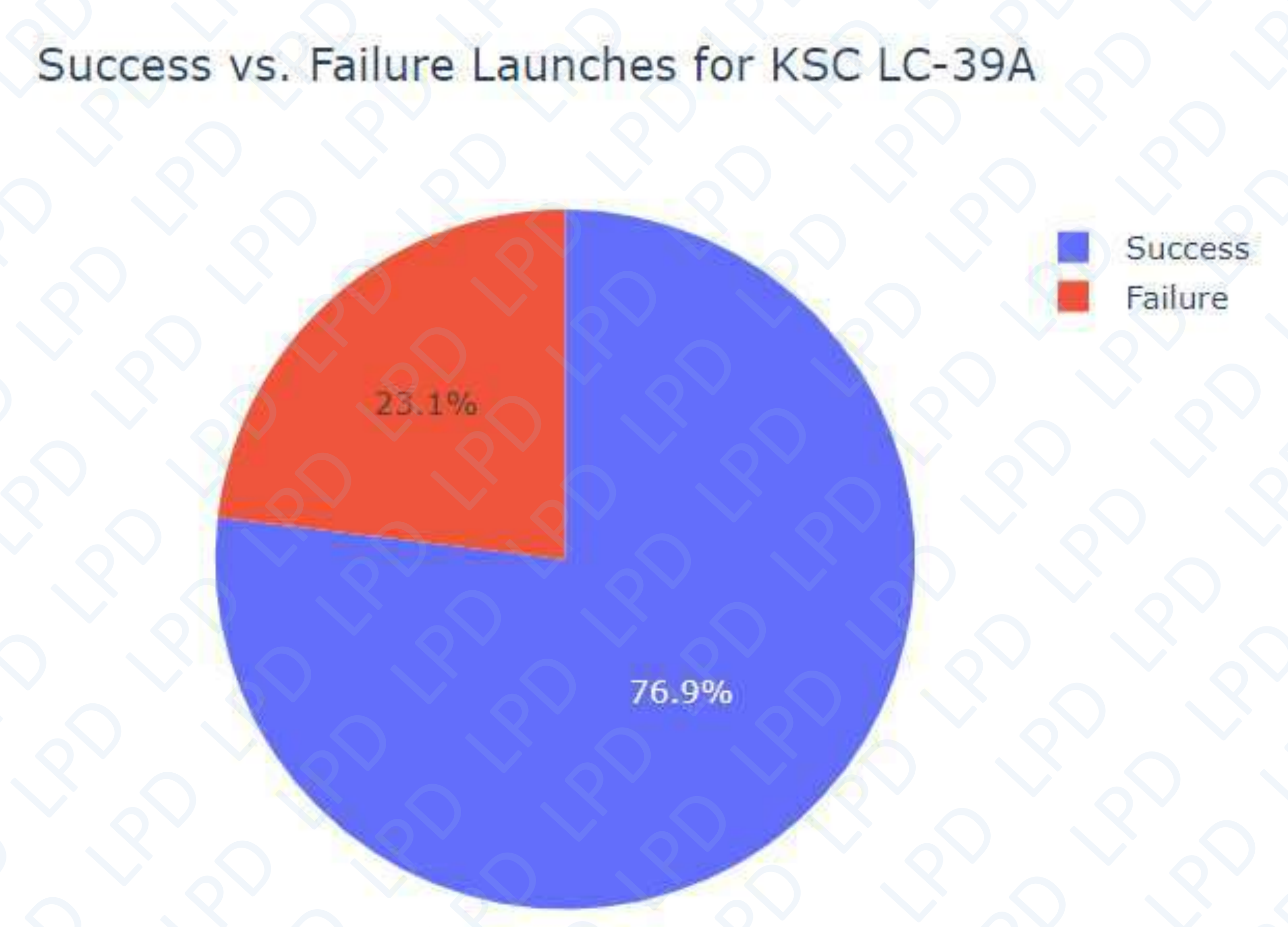
Build a Dashboard with Plotly Dash

Total count of successful launches by Launch Site



- The most successful launches occurred from **KSC LC39A Launch Site** Occupying the whole chart with 41.7%.
- Followed by **CCAFS LC-40** (29.2%) and **VAFB SLC-4E** (16.7%)
- The least numbers of successful launches occurred from **CCAFS SLC-40** occupying only 12.5% of the whole chart.

Launch Site with the most success rate

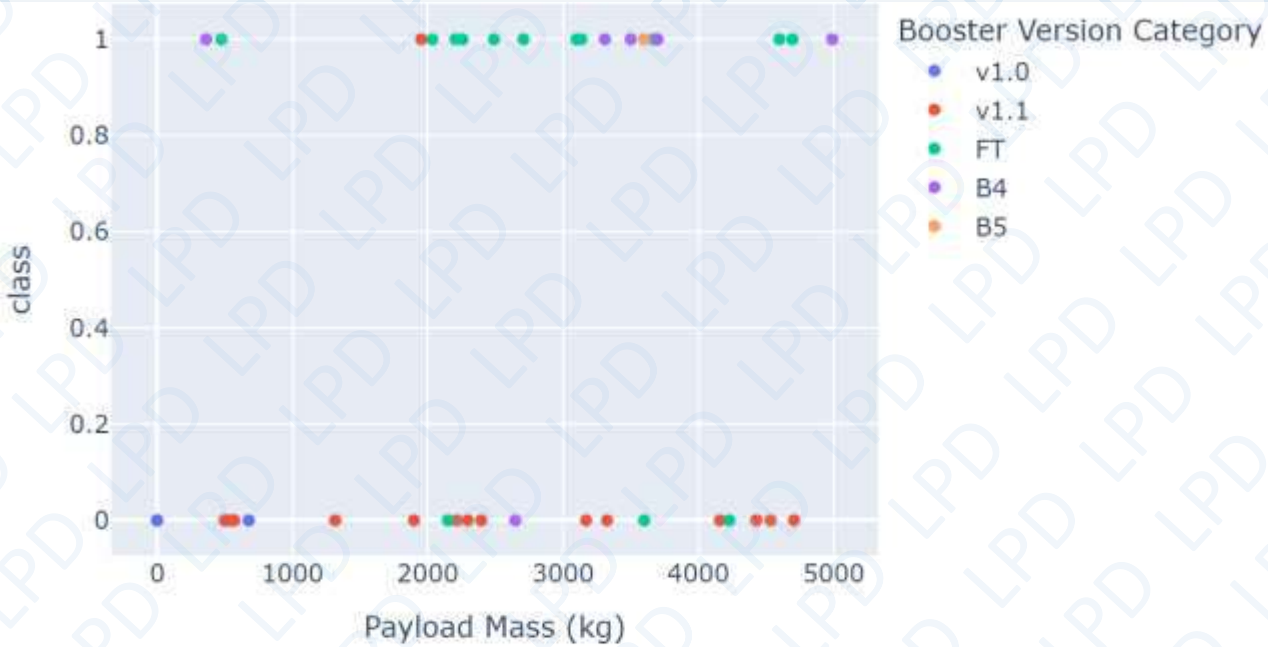


- KSC LC-39A performed the most success rate of 76.9% of successful launches while getting 23.1% of failure rate.

Payload Mass vs. Success

- We can clearly see that Payloads above 5k have less success ratio rate and only **‘FT’** and **‘B4’** boosters are representing Heavily Weighted Payloads.
- On the Low Middle Weighted Payloads we can see more successful rate and more versions of boosters.
- The most successful Launches are clustering between 2k and 4k Payload Mass mostly represented by the Booster Version of **‘F4’.**

Low-Middle Weighted payloads



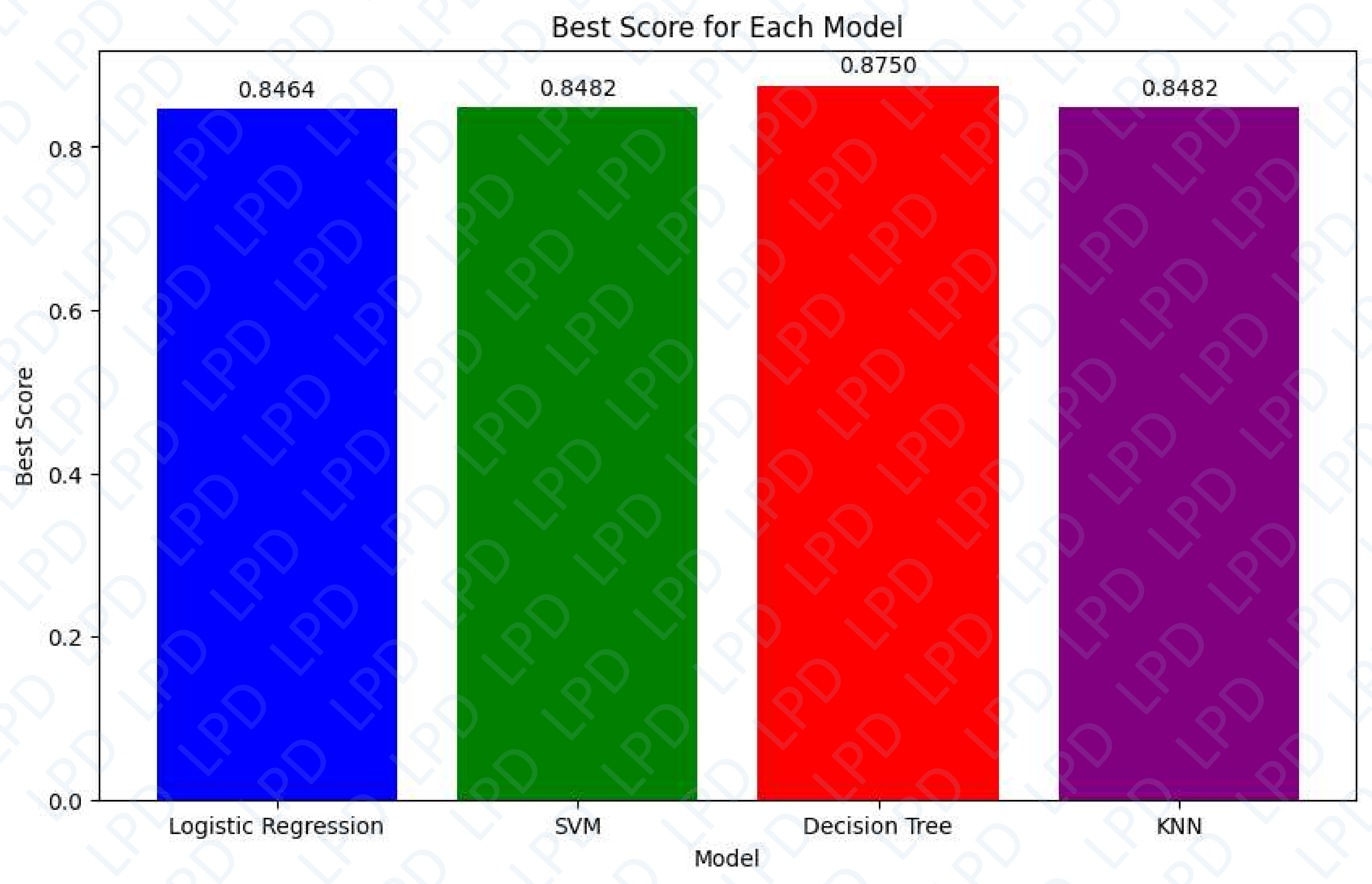
Middle-Heavily Weighted payloads

Section 5

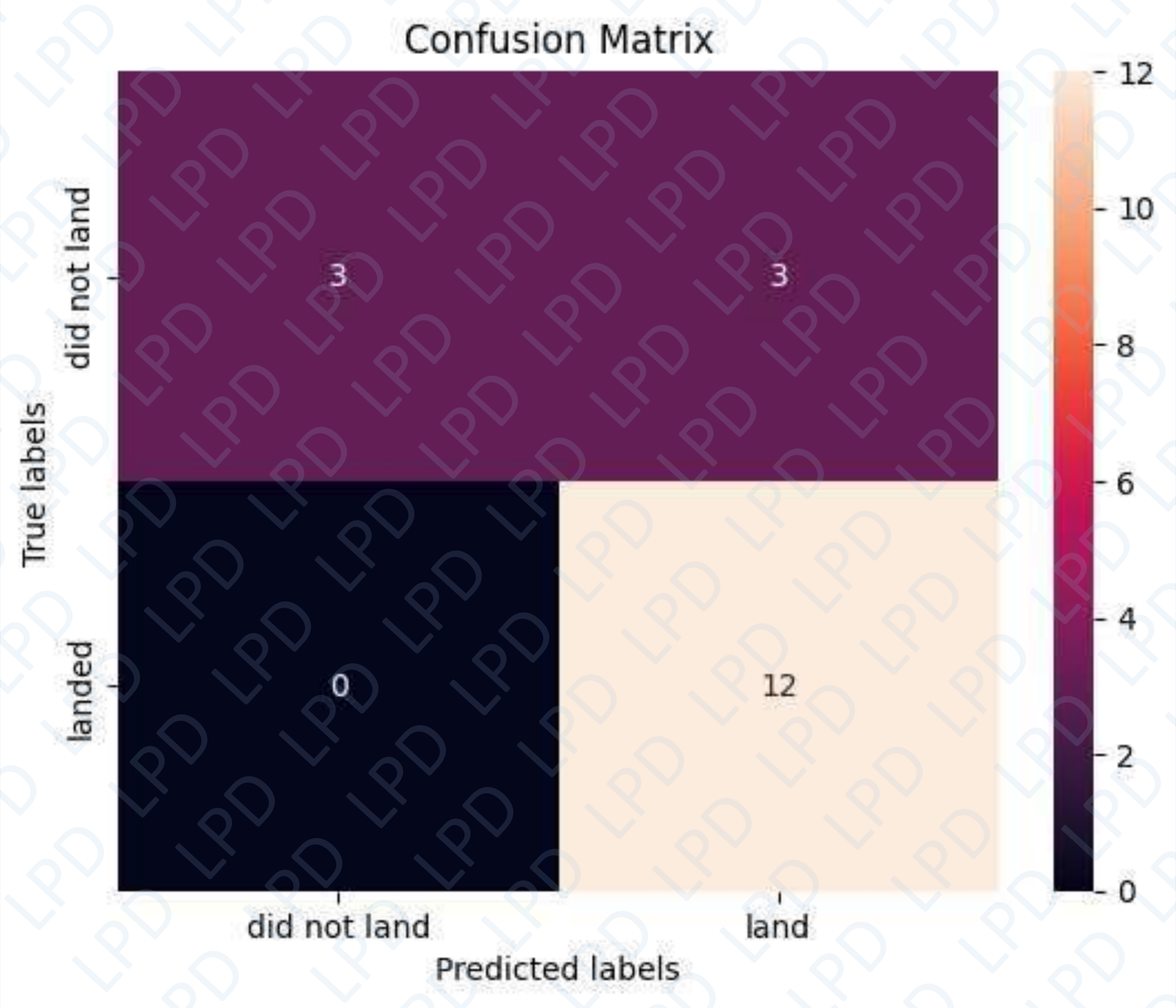
Predictive Analysis (Classification)

Classification Accuracy

Decision Tree has the best score of 0.875



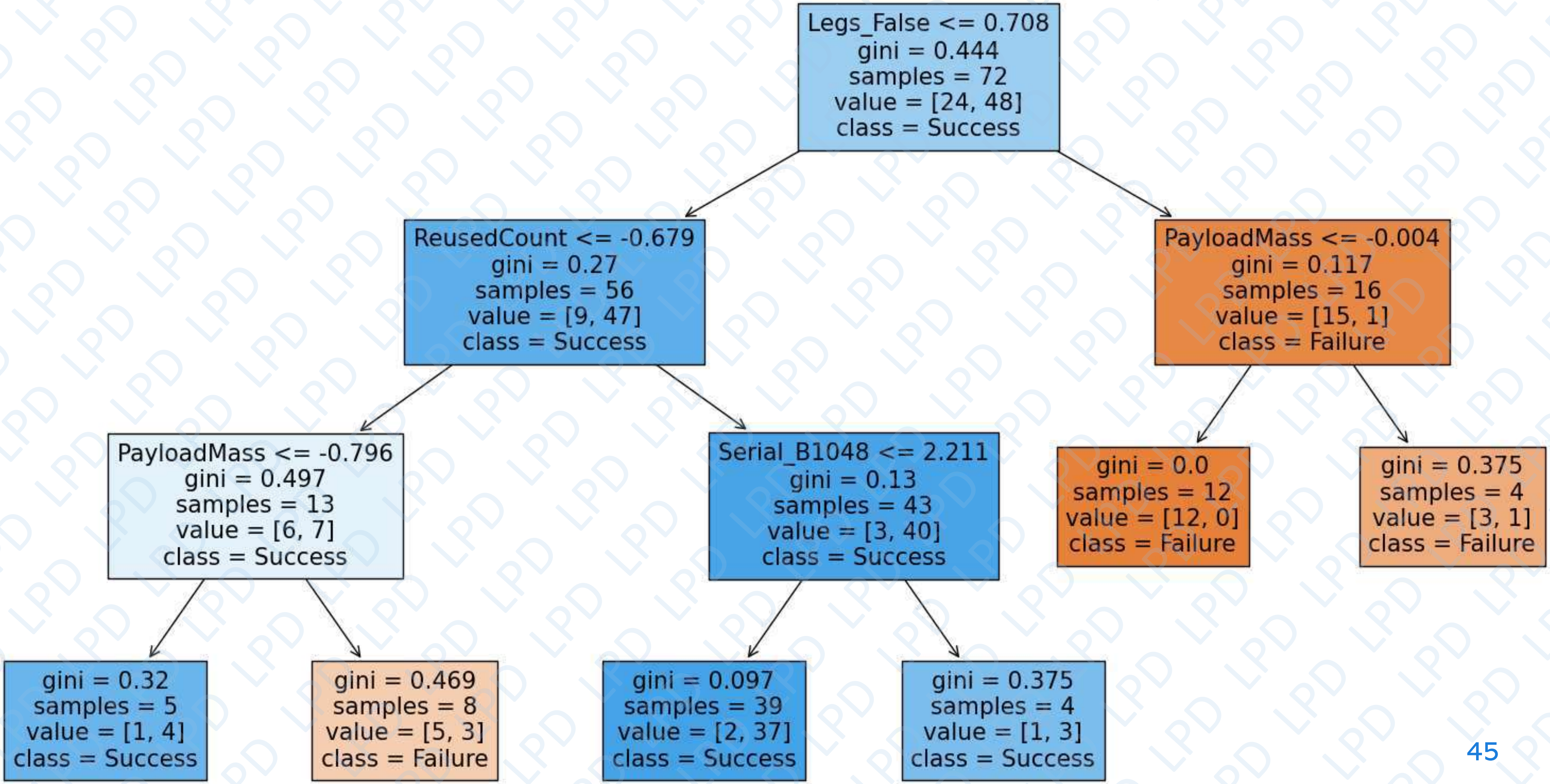
Confusion Matrix



- The confusion matrix shows that when it predicts ‘land’ it **100%** lands.
- However, if it predicts “did not land” it performs **50%**

Did not land: **3/6 (50%)**
Landed: **12/12(100%)**
Overall accuracy: **15/18(83.333%)**

Visualizing Decision Tree results



Interpretation of Decision Tree results

1. Landing Legs Configuration
Critical Factor: Presence of landing legs significantly impacts launch success.

2. Rocket Reuse Count
Critical Factor: The number of times a rocket is reused is inversely related to its success rate.

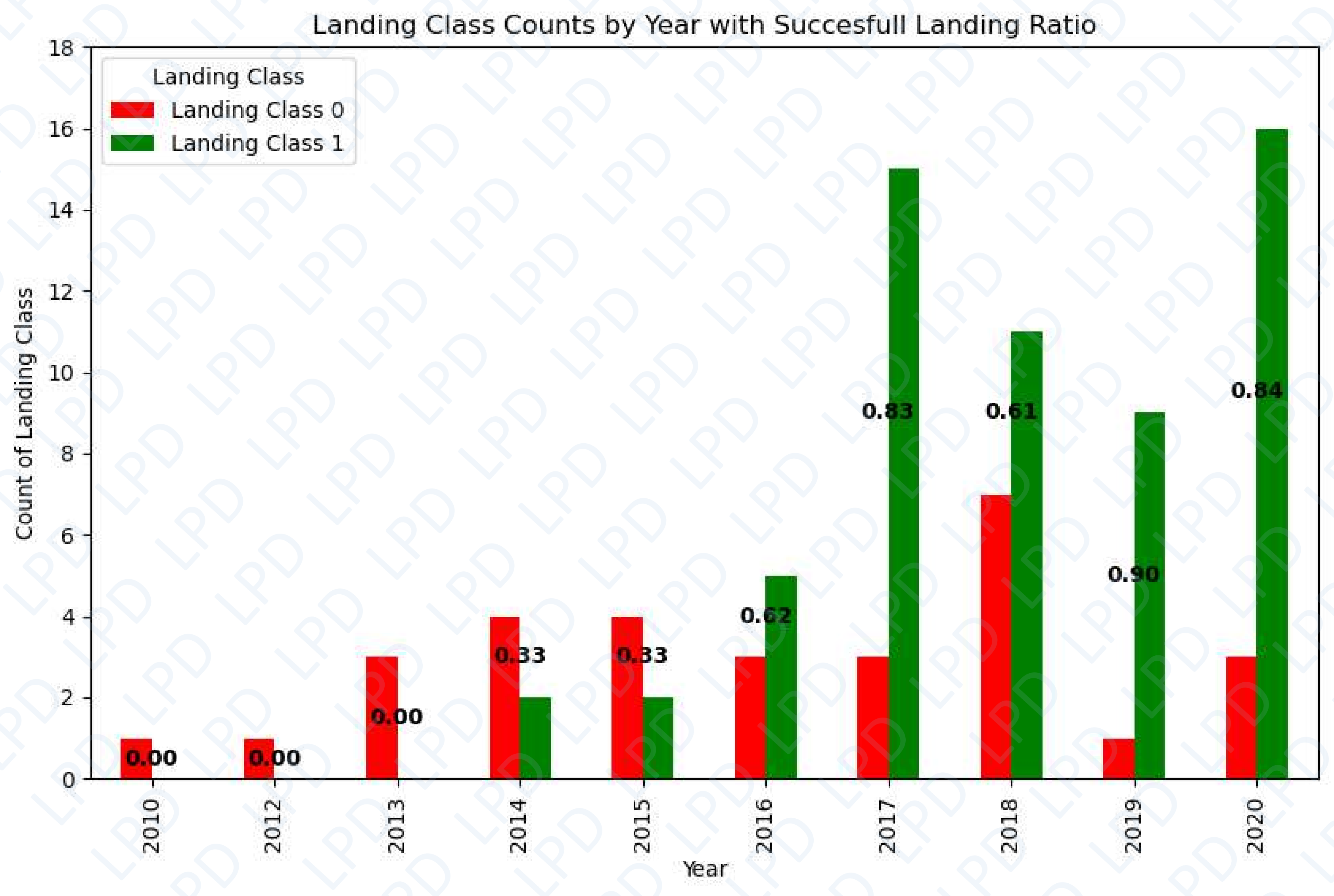
3. Payload Mass Management
Critical Factor: The mass of the payload affects the likelihood of success.

4. Serial Number Performance
Critical Factor: Specific rocket serial numbers are linked to better performance..

Conclusions

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- KSC LC-39A is holding the highest success rate of launches.
- The significant rise of Launch success rate commenced in 2013.
- Orbits: ES-L1, GEO, HEO, SSO are holding a 100% success rate each.
- Another mentionable Orbit: There are significantly more Launces targeting **VLEO -** than previously mentioned Orbits - is holding a very high success rate
- The Decision tree is the best performing machine learning algorithm predicting the outcome amongst classifiers.
- SpaceX learns.

Number of Launches and improvement over time.



•**Improvement Over Time**: SpaceX showed a clear improvement in performance year by year, with a notable increase in successful landings as the number of launches grew.
•**Initial Decline**: Following a period of rapid progress, there was a significant decline in success rates and a reduction in the number of launches.
•**High Success Rate Achievement**: After the downturn, SpaceX achieved a remarkable 90% success rate with a modest number of launches.
•**Record Launches and Success Ratio**: The subsequent year saw the highest number of launches and the second-highest ratio of successful landings.
•**Ongoing Commitment**: This trajectory underscores SpaceX's ongoing dedication to enhancing reliability and expanding launch capabilities

Thank you!