IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

Project Objective:

- Analyze SpaceX launch data to identify key factors influencing launch success and predict future outcomes.

Overall Methodology:

- Data collection using the SpaceX API and web scraping
- Data cleaning and transformation
- Exploratory Data Analysis (EDA) and visualizations (charts, interactive maps, dashboards)
- Predictive modeling using classification techniques

Key Results:

- Identification of relationships between payload, orbit types, and launch success
- Comparative evaluation of different predictive models

Introduction

Background and Context:

- SpaceX is revolutionizing the aerospace industry with innovative launch technologies.

Problem Statement:

- Determine the key factors driving launch success and use classification models to predict outcomes.

Project Justification:

- Optimize future launches and support decision-making in the aerospace sector.

Section 1

Methodology

Methodology

Data collection methodology:

- Extracted launch data using the SpaceX API (information on launches, sites, boosters, payloads, etc.)
- Supplemented the dataset with additional data gathered via web scraping.

- Perform data wrangling

  Handling missing values, converting categorical variables, and normalizing numerical data.

- 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

- Our datasets were collected using a dual approach:

SpaceX API Calls:

- **Key Phrases:** REST API, JSON extraction, endpoint calls, SpaceX API
- We made RESTful API calls to the SpaceX API endpoints to retrieve comprehensive launch data. This includes information on launch sites, booster versions, payload details, and flight outcomes.

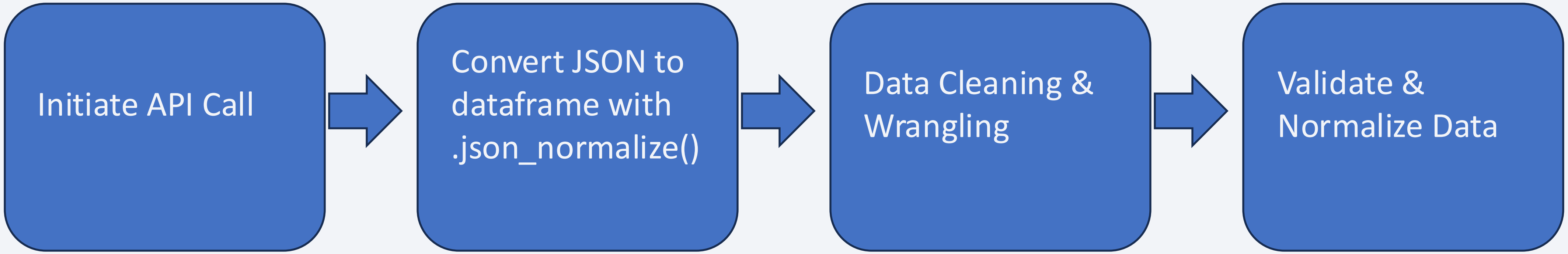
Web Scraping:

- **Key Phrases:** Web scraping, HTML parsing, data extraction, supplemental data
- In addition to the API data, we used web scraping techniques to extract complementary data from public sources. This helped in gathering additional context and filling in any missing information from the API.

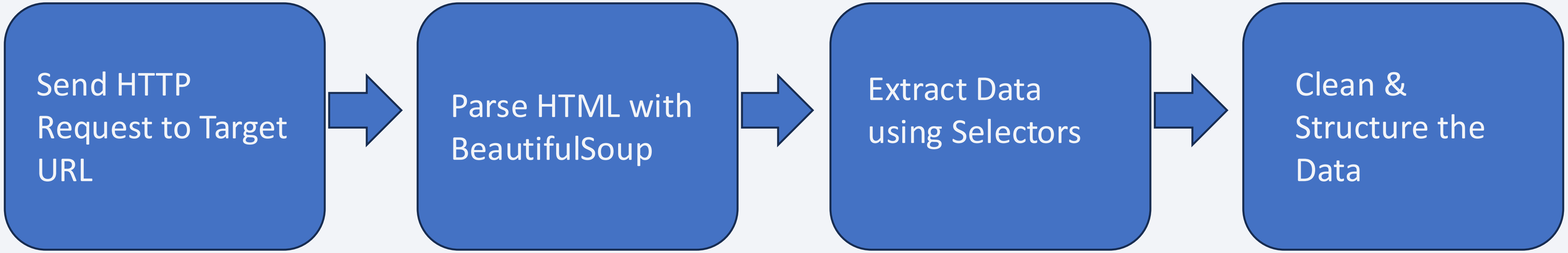
Data Aggregation:

- **Key Phrases:** Data wrangling, merging datasets, data validation
- After collecting data from both sources, we cleaned and transformed the datasets. This process involved handling missing values, converting categorical variables, and normalizing numerical data. The final step was to merge all data sources into a unified dataset for analysis.

Data Collection – SpaceX API



Data Collection - Scraping



Data Wrangling

Key Phrases:

- Data Cleaning
- Data Transformation
- Handling Missing Values
- Data Normalization
- Feature Engineering
- Data Integration

Process Overview:

- **Import Raw Data:**
  - Load raw datasets from the SpaceX API and web scraping sources.
- **Data Cleaning:**
  - Remove duplicates, handle or impute missing values, and correct inconsistencies.
- **Data Transformation:**
  - Convert data types, normalize numerical features, and encode categorical variables.
- **Feature Engineering:**
  - Create new variables that help in analysis, such as aggregated metrics or derived features.
- **Data Integration:**
  - Merge cleaned and transformed datasets into a unified dataset ready for analysis and modeling.

https://github.com/Sebastian2nunez/IBM-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling-v2.ipynb

EDA with Data Visualization

Charts Plotted and Their Purpose

- **Scatter Plots:**
  - **Flight Number vs. Launch Site:**
    - **Purpose:** To visualize the distribution of launches across different sites, helping to identify any patterns or clusters in launch frequency.
  - **Payload vs. Launch Site:**
    - **Purpose:** To examine the relationship between payload mass and launch site, which can indicate how payload size might influence launch outcomes.
- **Bar Chart:**
  - **Success Rate by Orbit Type:**
    - **Purpose:** To compare the success rates of launches across different orbit types. This chart helps highlight which orbits are associated with higher or lower success probabilities.
- **Line Chart:**
  - **Yearly Launch Success Trend:**
    - **Purpose:** To track changes in launch success rates over time, revealing trends or improvements in performance across

https://github.com/Sebastian2nunez/IBM-Capstone/blob/main/jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

https://github.com/Sebastian2nunez/IBM-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

- **Unique Launch Sites Query:**
- Retrieves a list of all distinct launch site names.
- **Launch Site Filtering Query:**
- Retrieves 5 records where the launch site names begin with "CCA" to inspect specific entries.
- **Total Payload by NASA Boosters Query:**
- Calculates the total payload mass carried by boosters associated with NASA.
- **Average Payload Mass Query:**
- Computes the average payload mass for the booster version F9 v1.1.
- **First Successful Ground Landing Date Query:**
- Finds the earliest date of a successful ground pad landing.
- **Successful Drone Ship Landing Query:**
- Lists boosters that have successfully landed on a drone ship with payload mass between 4000 and 6000.
- **Mission Outcomes Summary Query:**
- Calculates the total number of successful and failed mission outcomes.
- **Maximum Payload Boosters Query:**
- Retrieves the names of the boosters that carried the maximum payload mass.
- **Failed Drone Ship Landings (2015) Query:**
- Retrieves details (booster version and launch site) for failed drone ship landing outcomes in the year 2015.
- **Ranked Landing Outcomes Query:**
- Ranks the count of landing outcomes (e.g., failure on drone ship, success on ground pad) between two specified dates in descending order.

Build an Interactive Map with Folium

- **Markers (Launch Sites):**
- Placed markers at each SpaceX launch site to visualize their exact locations.
- **Purpose:** Helps identify the geographical distribution of launch sites and compare their proximities to key infrastructure.
- **Colored Circles (Launch Outcomes):**
- Added color-coded circles representing launch success (green) and failure (red).
- **Purpose:** Provides a quick visual representation of launch outcomes at different sites.
- **Lines (Proximity Analysis):**
- Created lines connecting launch sites to nearby infrastructure, such as railways, highways, and coastlines.
- **Purpose:** Analyzes how close launch sites are to essential transportation networks, which could impact logistics and safety.
- **Popups (Detailed Information):**
- Configured interactive popups displaying launch details when clicking on a marker.
- **Purpose:** Enables users to explore launch-specific details without cluttering the map.

https://github.com/Sebastian2nunez/IBM-Capstone/blob/main/lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

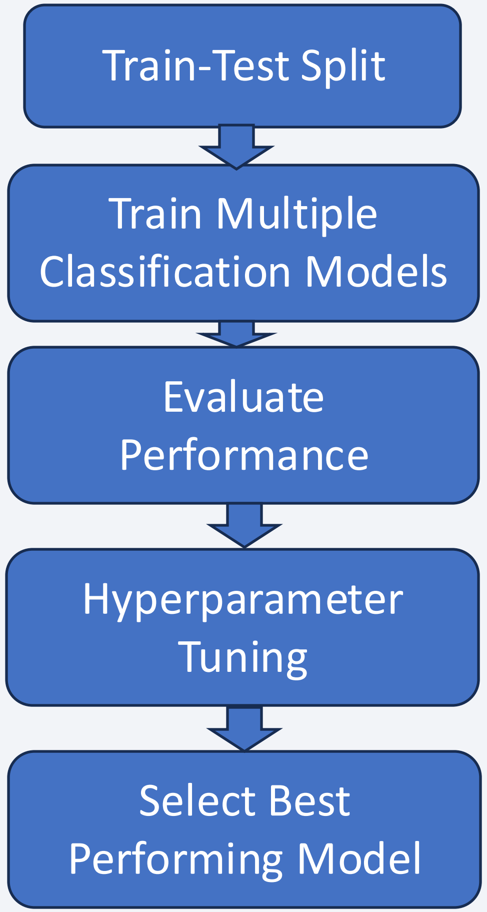
- **Plots and Graphs Added:**
- **Pie Chart – Launch Success Count for All Sites**
    - **Purpose:** Displays the overall distribution of successful and failed launches across all launch sites, providing a high-level success rate overview.
- **Pie Chart – Highest Launch Success Ratio Site**
    - **Purpose:** Highlights the launch site with the highest success ratio, helping identify the most reliable location for launches.
- **Scatter Plot – Payload vs. Launch Outcome**
    - **Purpose:** Shows the relationship between payload mass and launch success, helping analyze how payload weight affects launch success probability.
- **Dropdown Menu – Launch Site Selection**
    - **Purpose:** Allows users to filter data by specific launch sites, making it easier to analyze performance variations between locations.
- **Range Slider – Payload Mass Selection**
    - **Purpose:** Enables users to adjust the payload range dynamically, helping explore how different payload sizes impact launch success.
- **Hover and Click Tooltips**
    - **Purpose:** Provides additional details when hovering over data points, improving the user experience and making the dashboard more informative.

https://github.com/Sebastian2nunez/IBM-Capstone/blob/main/Dash.py

Predictive Analysis (Classification)

- **Data Preprocessing:**
  - Handled missing values, normalized numerical features, and encoded categorical variables.
  - **Key Phrases:** Data normalization, feature scaling, one-hot encoding, train-test split.
- **Model Selection:**
  - Tested different classification models, including **Logistic Regression, Decision Trees, Random Forest, and K-Nearest Neighbors (KNN)**.
  - **Key Phrases:** Supervised learning, binary classification, model training.
- **Model Evaluation:**
  - Evaluated models using **accuracy, precision, recall, and F1-score**.
  - **Generated confusion matrices** to analyze misclassifications.
  - **Key Phrases:** Performance metrics, confusion matrix, classification report, ROC curve.
- **Model Improvement:**
  - **Hyperparameter tuning** was performed using GridSearchCV and RandomizedSearchCV to optimize the models.
  - **Key Phrases:** Hyperparameter optimization, cross-validation, GridSearchCV, feature importance.
- **Best Performing Model:**
  - Based on evaluation metrics, the best-performing model was selected for final predictions.
  - **Key Phrases:** Best accuracy, highest F1-score, model deployment.

https://github.com/Sebastian2nunez/IBM-Capstone/blob/main/Spacex-Machine-Learning-Prediction-Part-5-v1.ipynb

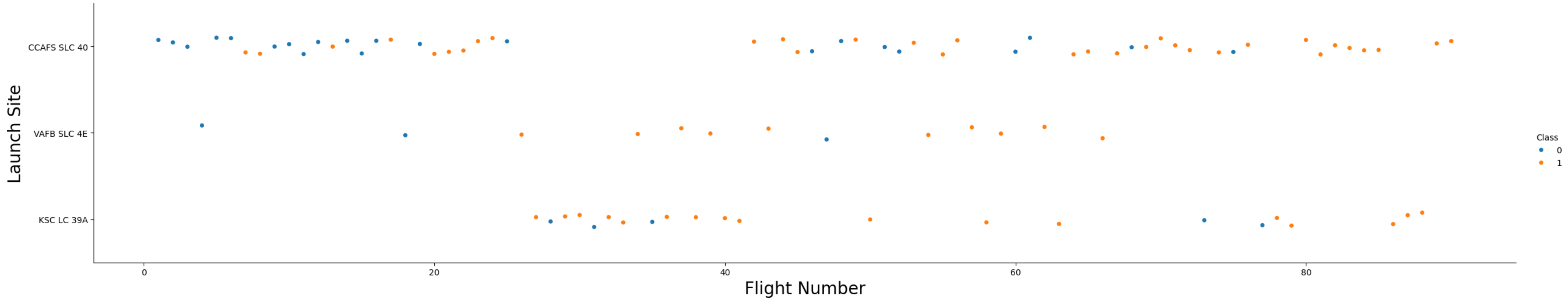


Results

- **Flight Number vs. Launch Site:**
- Showcases the distribution of launches across different sites.
- **Finding:** Some launch sites have significantly more flights than others.
- **Payload vs. Launch Outcome:**
- Displays the impact of payload mass on launch success.
- **Finding:** Heavier payloads have a lower success rate, particularly in certain orbit types.
- **Success Rate by Orbit Type:**
- Compares the success rates of launches across different orbits.
- **Finding:** Some orbit types, like GTO (Geostationary Transfer Orbit), have lower success rates.
- **Yearly Launch Success Trend:**
- Analyzes SpaceX's success rate over time.
- **Finding:** Success rates have improved significantly in recent years due to technological advancements.

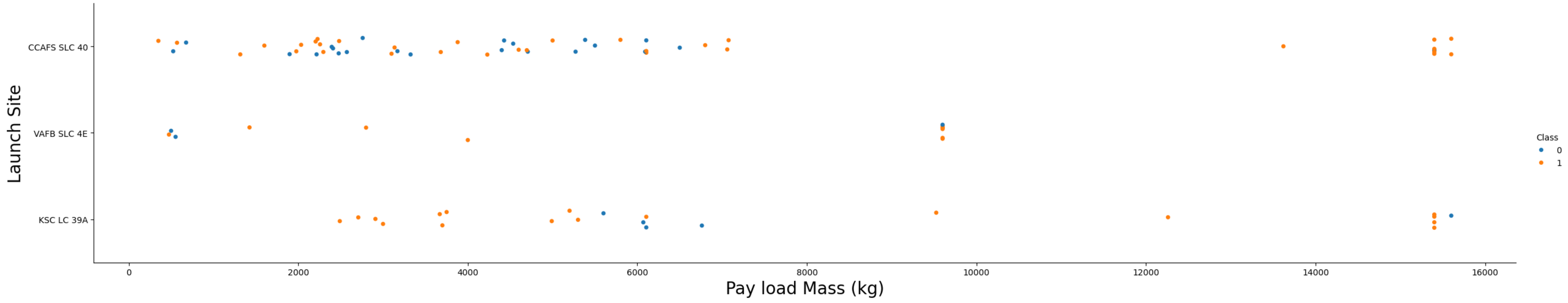
Section 2

Insights drawn from EDA



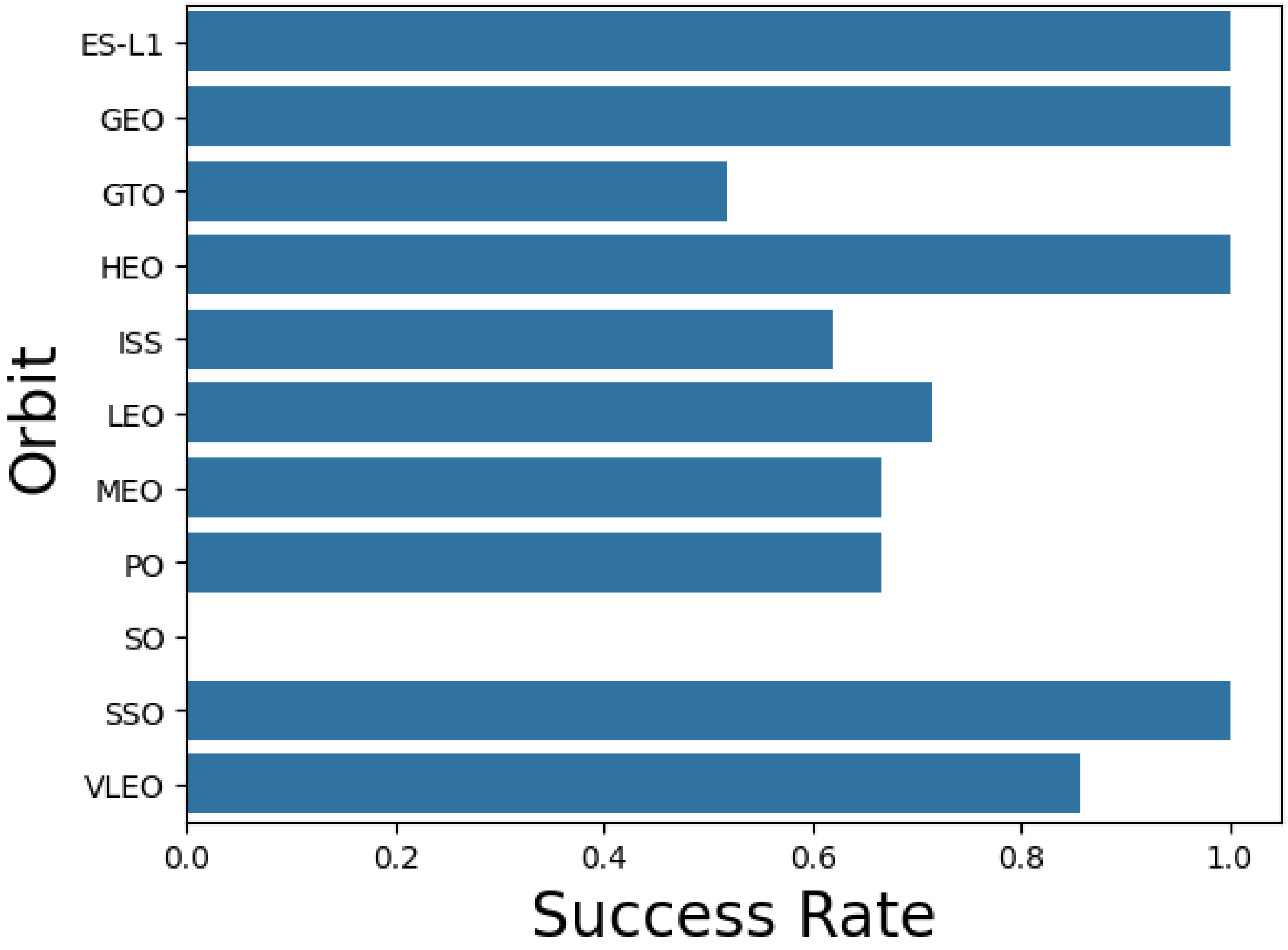
Flight Number vs. Launch Site

- KSC LC 39A and VAFB SLC 4E Has a high rate of success
- The CCAFS SLC 40 rate of success is round of 50%



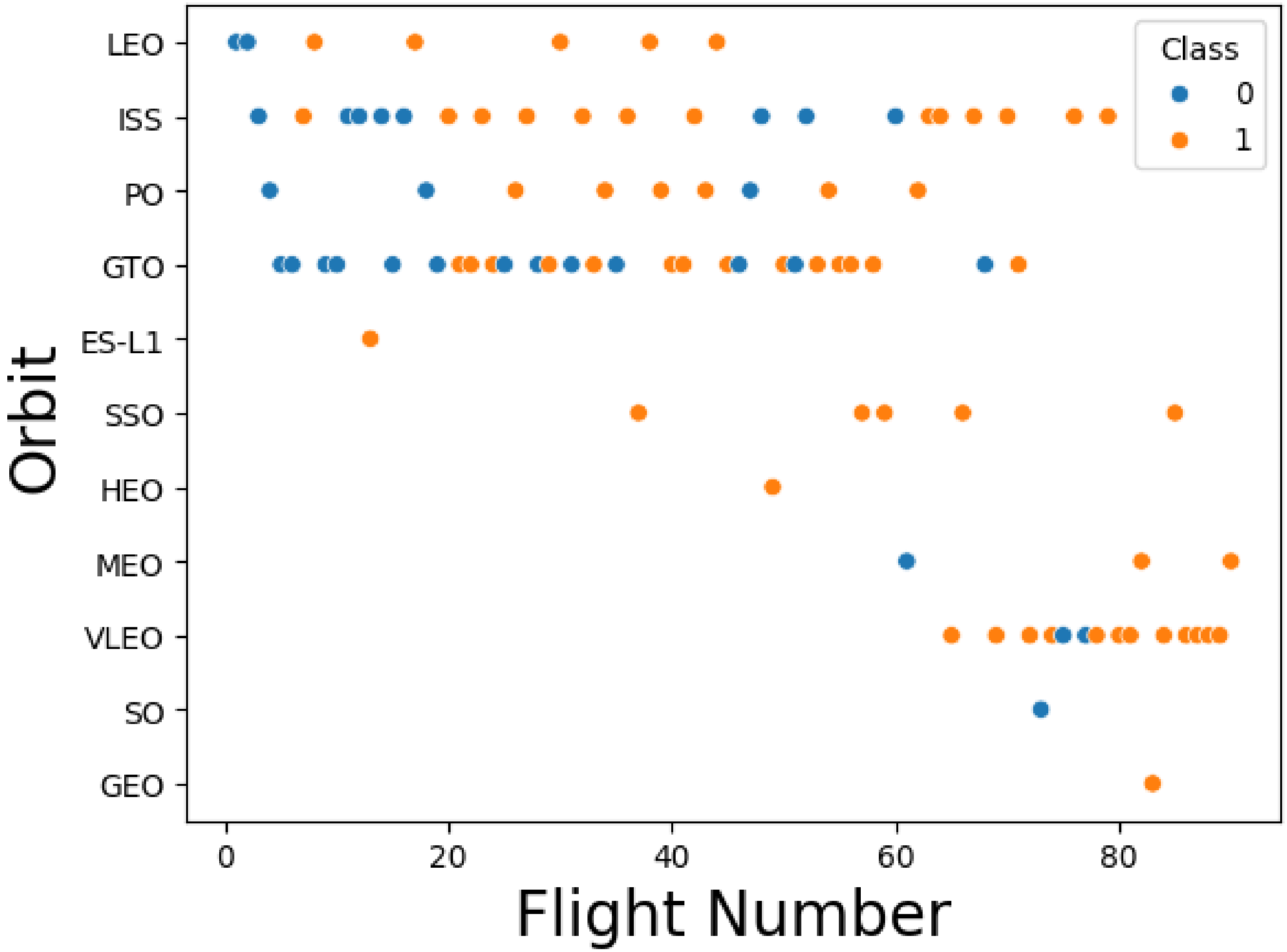
Payload vs. Launch Site

- VAFB SLC 4E has the best rate of success than the others
- With more Mass, the probability of success is improve



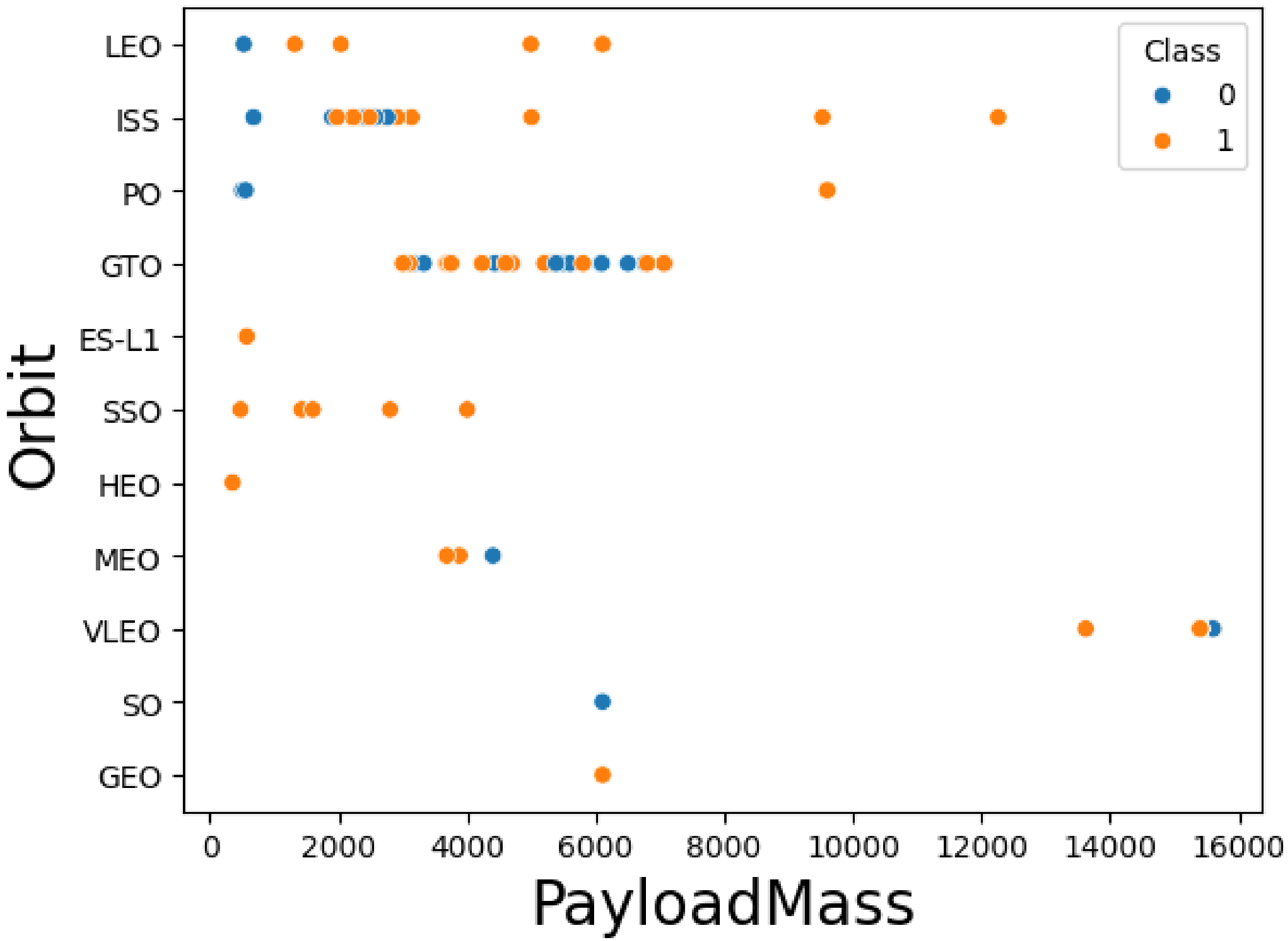
Success Rate vs. Orbit Type

- The best success rate was for ES-L1, GEO ,HEO and SSO with an 100% of success.
- With another orbit the rate of success decrease above all for SO and GTO.



Flight Number vs. Orbit Type

- VLEO has the greater rate of success.
- The worst rate of success was for GTO this was for above 20 flight number.
- Generally the firsts flight number was the worsts

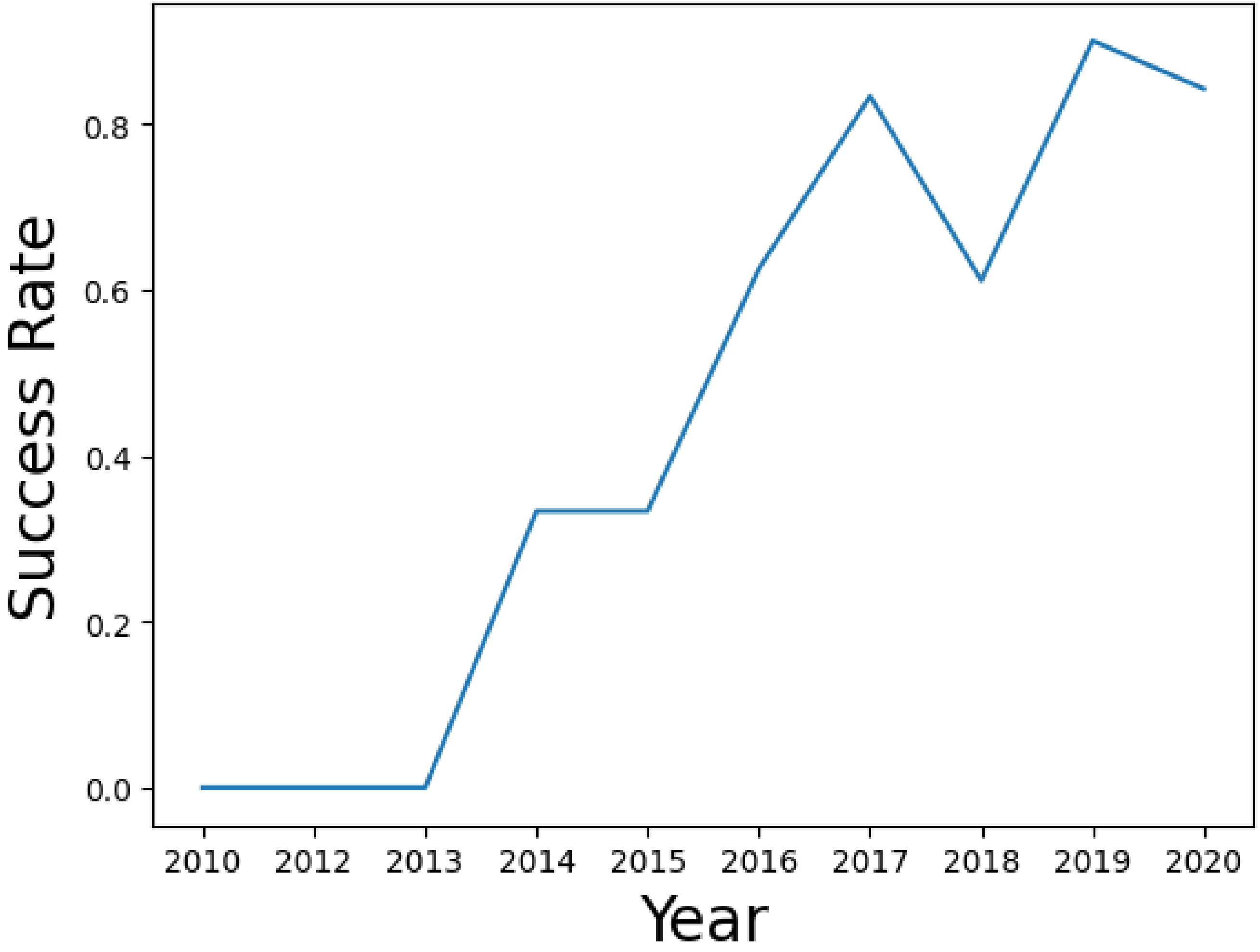


Payload vs. Orbit Type

- For the orbit SSO with a Pay load mass between 0 and 4000 the nicer rate of success.
- For ISS between 0 and 4000 the success decrease and is not very sure was succesfuly.

Launch Success Yearly Trend

- With the time, we can see an increment of success rate. With a strong positive trend.



```
%sql SELECT Launch_Site, COUNT(Launch_Site) FROM SPACEXTABLE GROUP BY Launch_Site;
```

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

Launch_Site	COUNT(Launch_Site)
CCAFS LC-40	26
CCAFS SLC-40	34
KSC LC-39A	25
VAFB SLC-4E	16

All Launch Site Names

- This query was on sql. This show us all Launch sites and how many launchs each on.

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

Launch Site Names Begin with 'CCA'

- This show 5 launch sites where have the word CCA, and all data with relation of that.

```
%sql SELECT COUNT(Payload) FROM SPACEXTABLE WHERE Payload LIKE '%CRS%';
```

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

COUNT(Payload)
26

Total Payload Mass

- This is a count of payload booster carried by NASA.

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Mean_PAYLOAD_MASS FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9%';
```

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

Mean_PAYLOAD_MASS
6138.287128712871

Average Payload Mass by F9 v1.1

- This query give us the mean of Payload mass for F9

```
%sql SELECT min(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
✓ 0.0s
 * sqlite:///my_data1.db
Done.
```

min(Date)
2015-12-22

First Successful Ground Landing Date

- This query give us the first Successful ground landing of Space X

```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

✓ 0.0s

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

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

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query give us successful drone ship landing with a payload between 4000 and 6000 kg

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

✓ 0.0s

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

Mission_Outcome	COUNT(*)
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Generate + Code + Markdown

Total Number of Successful and Failure Mission Outcomes

- This query classification for success or fail and count that.
- The total of success was 100 and 1 failed in flight.

Boosters Carried Maximum Payload

- This was of the booster version, who has the maximum payload (15,600 kg).

```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
✓ 0.0s

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

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

substr(Date,6,2)	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- This record the failure for 2015 year

2015 Launch Records

```
%sql SELECT Landing_Outcome, Date FROM SPACEXTABLE WHERE (Landing_Outcome= 'Failure (drone ship)' or Landing_Outcome= 'Success (ground pad)') and Date BETWEEN  '2010-06-04' AND '2017-03-20' ORDER BY Land
```

✓ 0.0s Python

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

Landing_Outcome	Date
Success (ground pad)	2015-12-22
Success (ground pad)	2016-07-18
Success (ground pad)	2017-02-19
Failure (drone ship)	2015-01-10
Failure (drone ship)	2015-04-14
Failure (drone ship)	2016-01-17
Failure (drone ship)	2016-03-04
Failure (drone ship)	2016-06-15

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

- This rank the landing outcome beetween 2010-06-04 to 2017-03-20 in descendent order.

Section 3

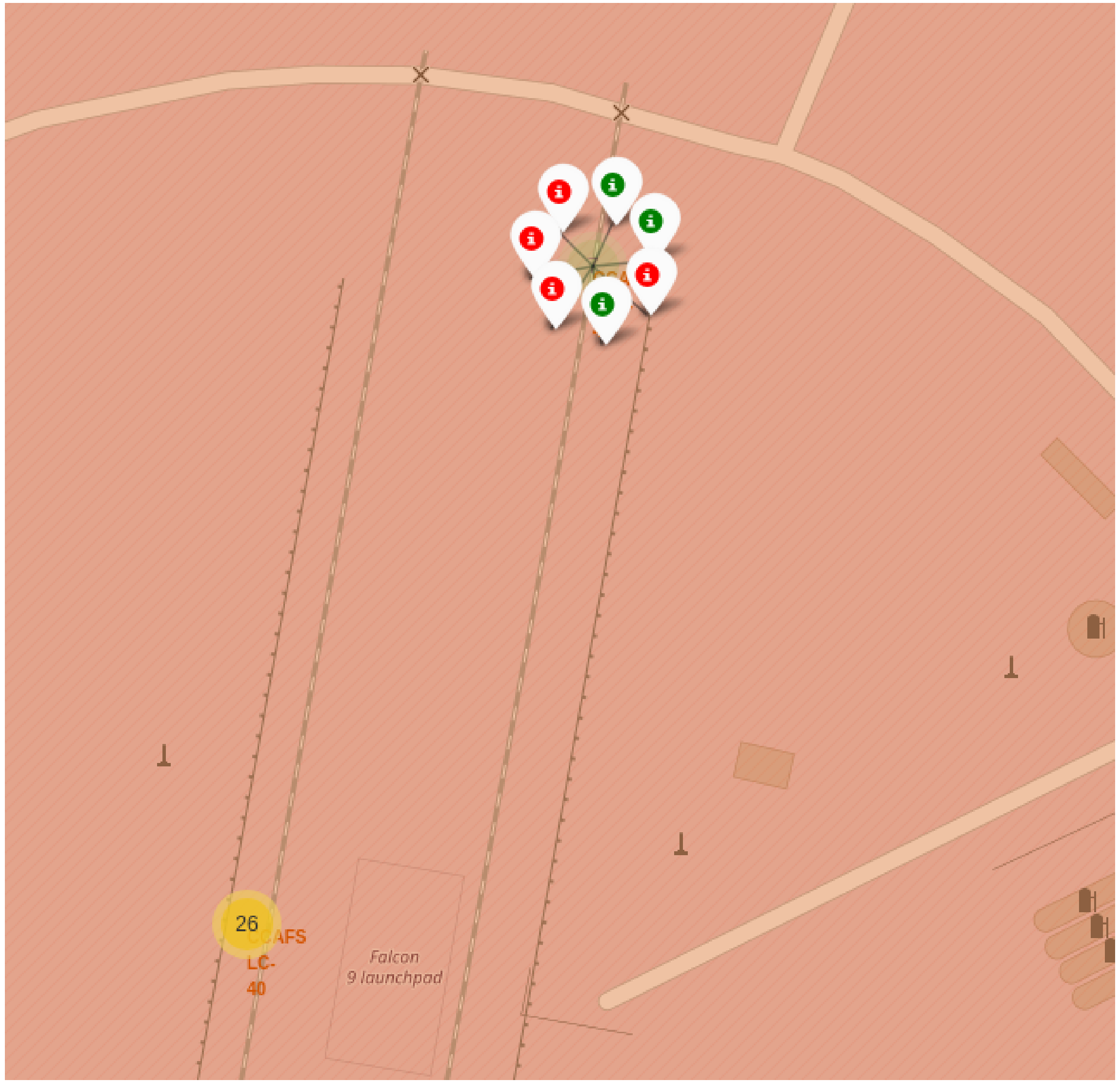
Launch Sites Proximities Analysis

Launch sites location

- The location of launch is very distant with both near of the sea. Travel that distance take a lot of time depending of velocity an object.

<Folium Map Screenshot 2>

- The green marker is for successful launch
- The red marker is for failed launch

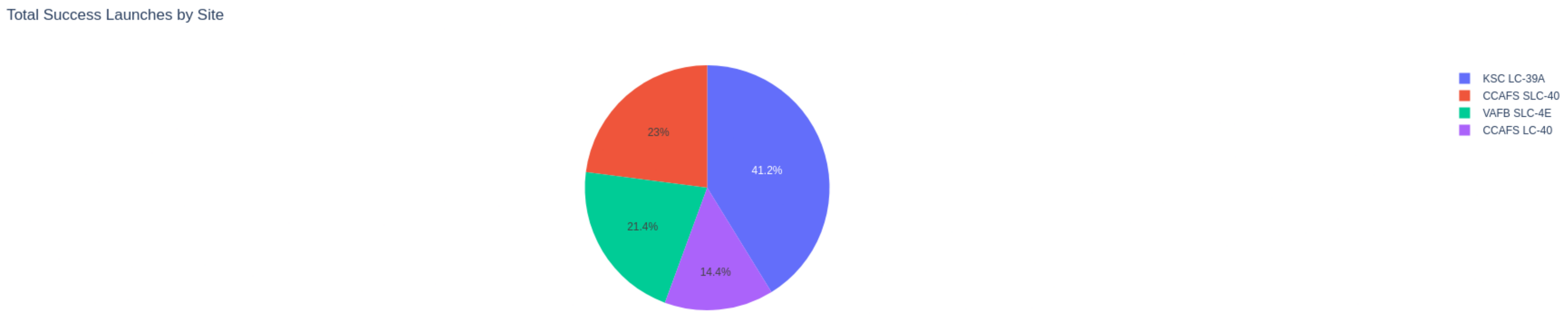


<Folium Map Screenshot 3>

- For middle prong has 13.47 km of distance.

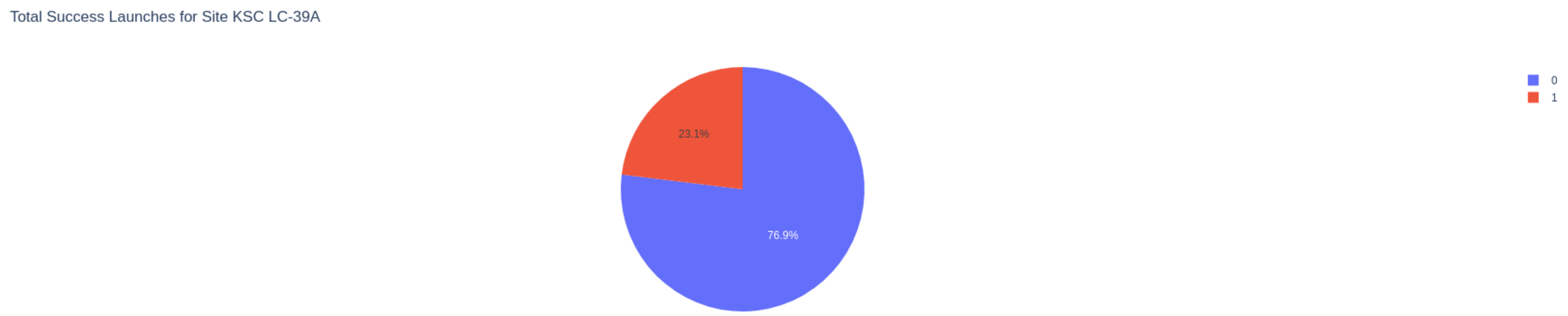
Section 4

Build a Dashboard with Plotly Dash



Pie chart Total success launches by site

- This pie chart classificate the launch sites and put Total success launche by site.

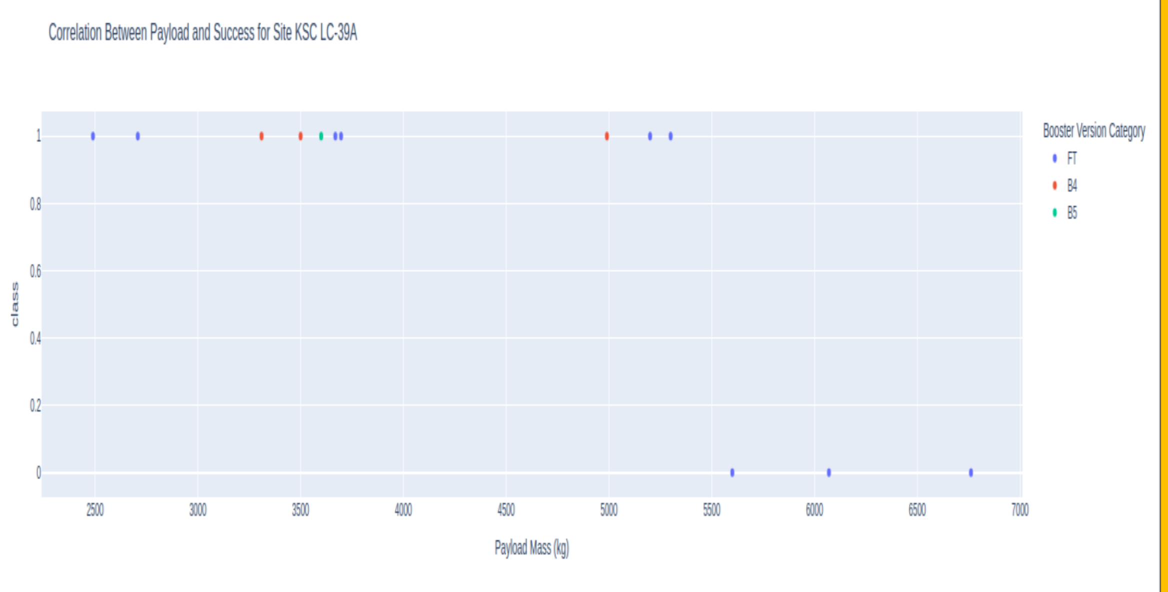


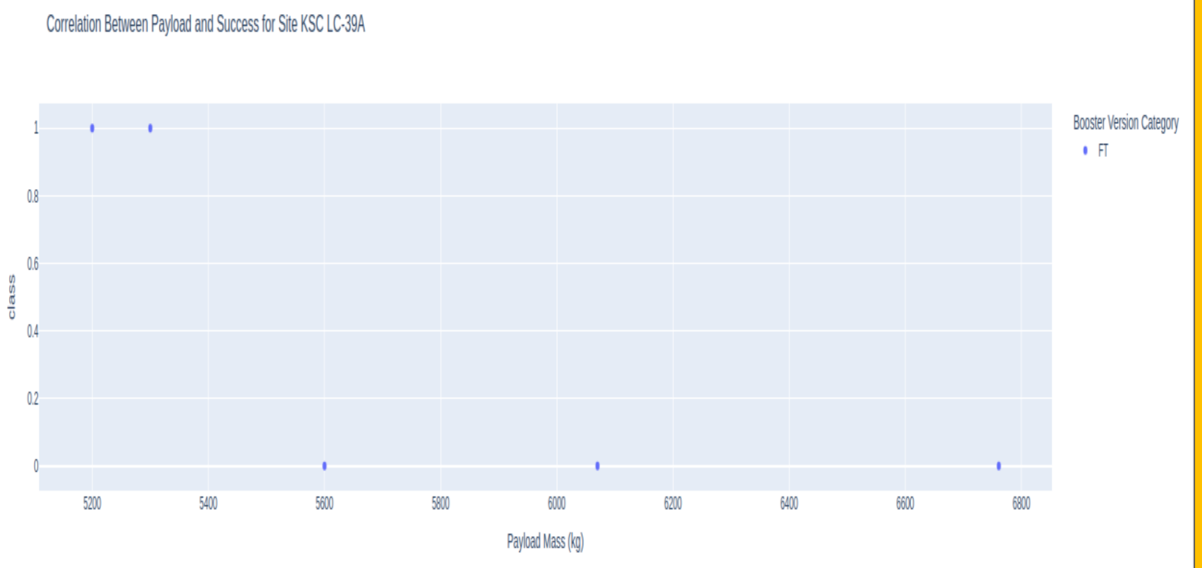
- This show us the greater success ratio. This was for KSC LC-39A

Highest launch success ratio

<Dashboard Screenshot 3>

- The charts show payload in different range of payload mass. First for 0 to 7000, and the other between 5000 and 8000.



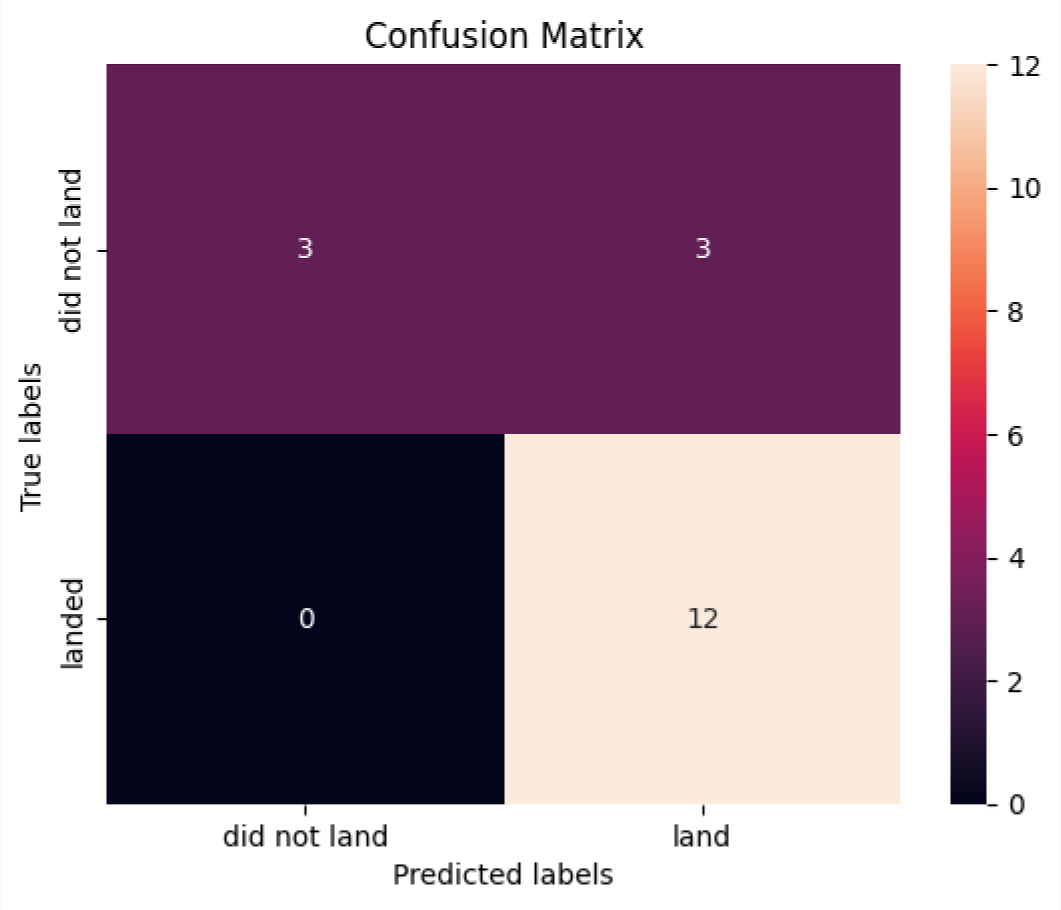


Section 5

Predictive Analysis (Classification)

Confusion Matrix

- This have false positive see in confusion matrix.



Conclusions

- **Factors Influencing Launch Success:**
  - Payload mass, launch site, booster version, and orbit type significantly impact launch success.
  - Lighter payloads tend to have a higher success rate.
  - Some orbit types, such as Geostationary Transfer Orbit (GTO), have lower success probabilities.
- **Trends in SpaceX Launches:**
  - Over time, SpaceX has **improved its launch success rate**, with significant advancements in booster recovery technology.
  - Certain launch sites (e.g., Kennedy Space Center) have a **higher frequency of successful launches** compared to others.
- **Predictive Model Insights:**
  - The best-performing classification model was identified using **accuracy, F1-score, and precision-recall analysis**.
  - Hyperparameter tuning significantly improved model performance, making it more reliable for predicting future launch outcomes.

Implications and Future Considerations:

- The insights gained can help **optimize future launches**, improve mission planning, and reduce failure risks.
- Further improvements could include **deep learning models** or **real-time data streaming analysis** for better predictions.
- Expanding the dataset with **weather conditions** and **real-time telemetry data** could enhance prediction accuracy.

Thank you!