IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Cinthya Cong
May 11, 2025

Outline

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

Executive Summary

This presentation presents a data-driven analysis of SpaceX's Falcon 9 launch records, leveraging data science methodologies to gain insights into launch performance, success rates, and factors influencing mission outcomes. The goal is to support better decision-making regarding mission planning and reusability strategies. To achieve this objective, I will use the following methodologies:

- Data Collection
- Data Wrangling
- Exploratory Data Analysis (EDA)
- Geospatial Visualization
- Visualization Tools
- Machine Learning Preparation

The analysis highlights that Falcon 9 launch success improves with flight experience and varies by launch site, orbit, and payload. High success rates are seen at KSC LC-39A and VAFB SLC-4E after key flight thresholds, and in orbits like ES-L1, GEO, HEO, and SSO. Heavier payloads perform better in Polar, LEO, and ISS orbits. These insights support optimized mission planning and reusability strategies.

Introduction

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.

In this capstone, a series of analytical questions arise that guide strategic decision-making regarding mission planning and booster reusability. By exploring launch outcomes, success rates by site and orbit, and booster performance, I aim to extract insights that support SpaceX in optimizing future missions and making data-driven operational choices.

- What is the success rate of SpaceX launches over time?
- What booster versions have the highest landing success rates?
- Can machine learning help predict the success of future launches?

Section 1

Methodology

Methodology

This project analyzes SpaceX launch data to support better decision-making in mission planning and rocket reusability strategies.

- Data Collection & Processing:
  Launch data was extracted from Wikipedia, cleaned, and structured using pandas, addressing missing or inconsistent entries.
- Exploratory Data Analysis (EDA):
  SQL queries and visualizations were used to explore success rates, launch site performance, and payload characteristics.
- Interactive Visual Analytics:
  Launch locations were mapped using Folium, and dynamic dashboards were built using Plotly Dash for deeper insights.
- Predictive Analysis:
  Classification models, such as logistic regression, were developed and fine-tuned to predict mission success, with performance evaluated using accuracy metrics.

This methodology enables data-driven decisions to optimize launch reliability and booster recovery strategies.

Data Collection

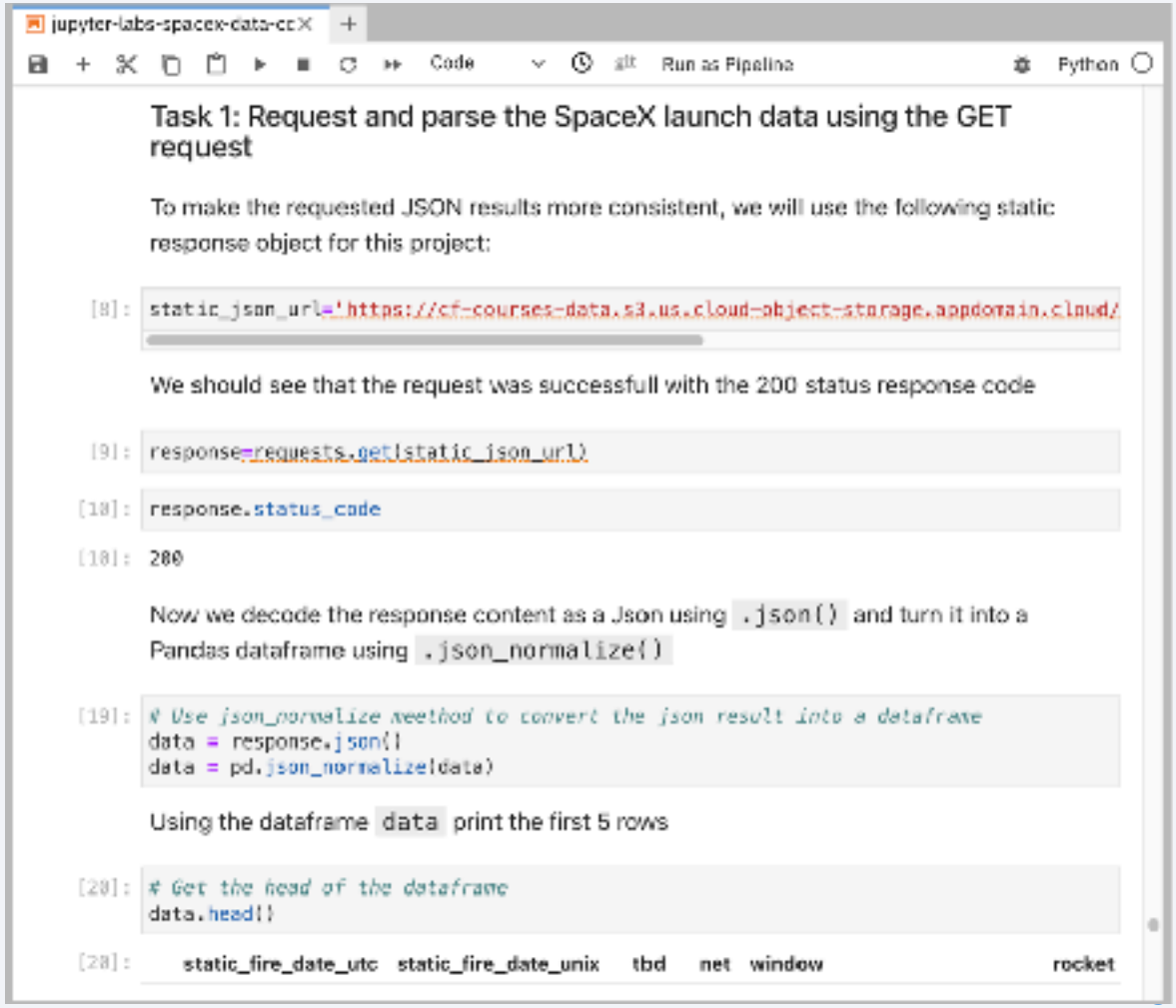
The data was collected using various methods

- Data collection was done by sending a GET request to the Wikipedia page of Falcon 9 launches.
- Next, we parsed the HTML response using BeautifulSoup to locate the launch record tables.
- I then extracted relevant data fields such as Flight Number, Launch Site, Payload, Orbit, etc.
- The extracted HTML table rows were cleaned and converted into a structured pandas DataFrame.
- In addition, I checked and handled missing or inconsistent values during the data wrangling stage.
- The objective was to create a reliable dataset suitable for further analysis and machine learning.

Data Collection – SpaceX API

The data collection process began by sending a GET request to the SpaceX REST API to retrieve launch data in JSON format. This data was then normalized into a tabular structure using pandas.json_normalize. After parsing, the dataset was cleaned and filtered to handle missing or irrelevant values. Finally, it was enriched by merging with additional launch information scraped from Wikipedia, resulting in a clean and comprehensive dataset ready for analysis.

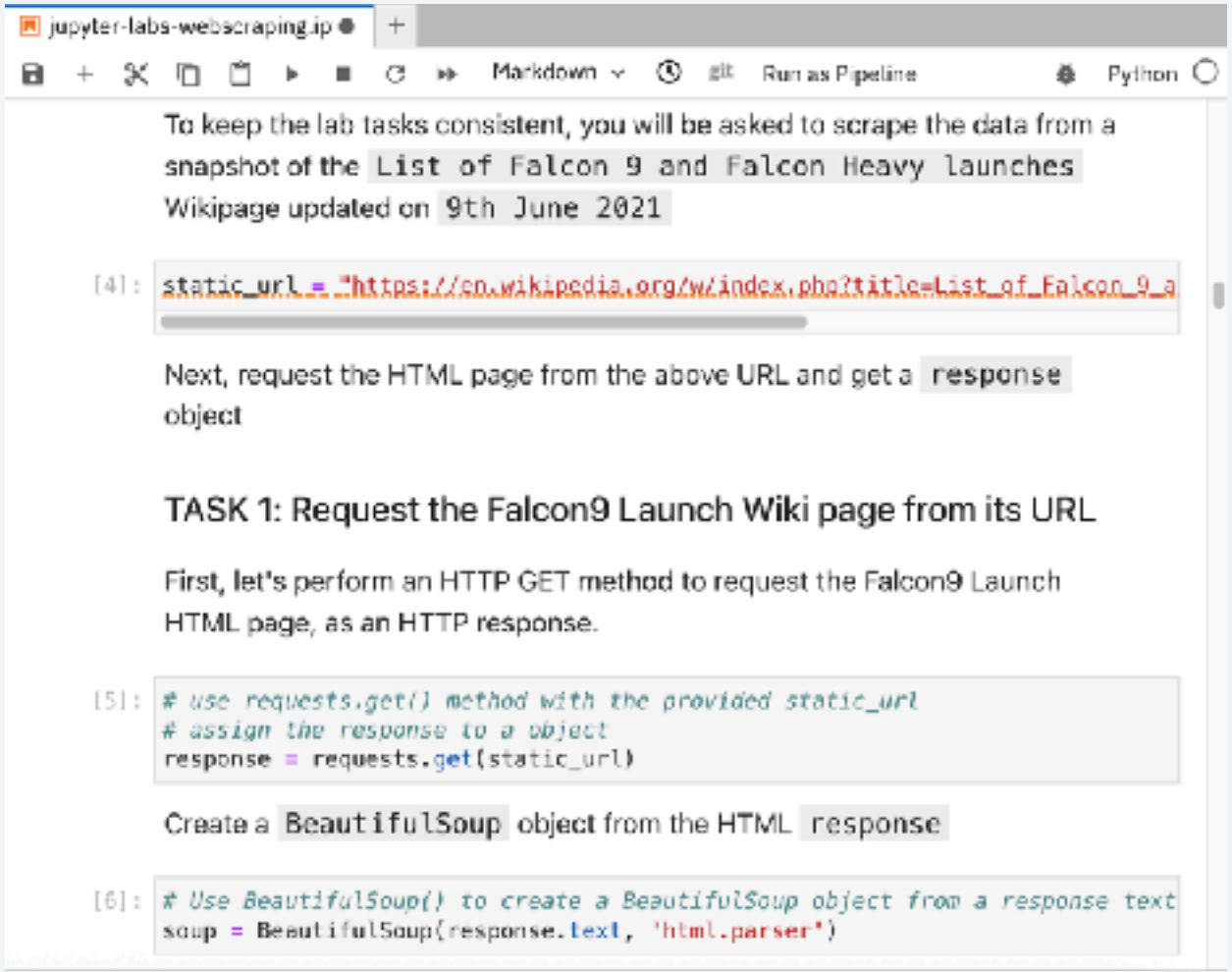
GitHub URL of the completed SpaceX API calls notebook: https://github.com/cinthyacong/Data-Science-Capstone-SpaceX/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

I scraped Falcon 9 launch data from Wikipedia using requests and BeautifulSoup, extracted tables, cleaned and transformed the data (e.g., dates, payload mass), and stored it in a pandas DataFrame for analysis.

GitHub URL: https://github.com/cinthyacong/Data-Science-Capstone-SpaceX/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

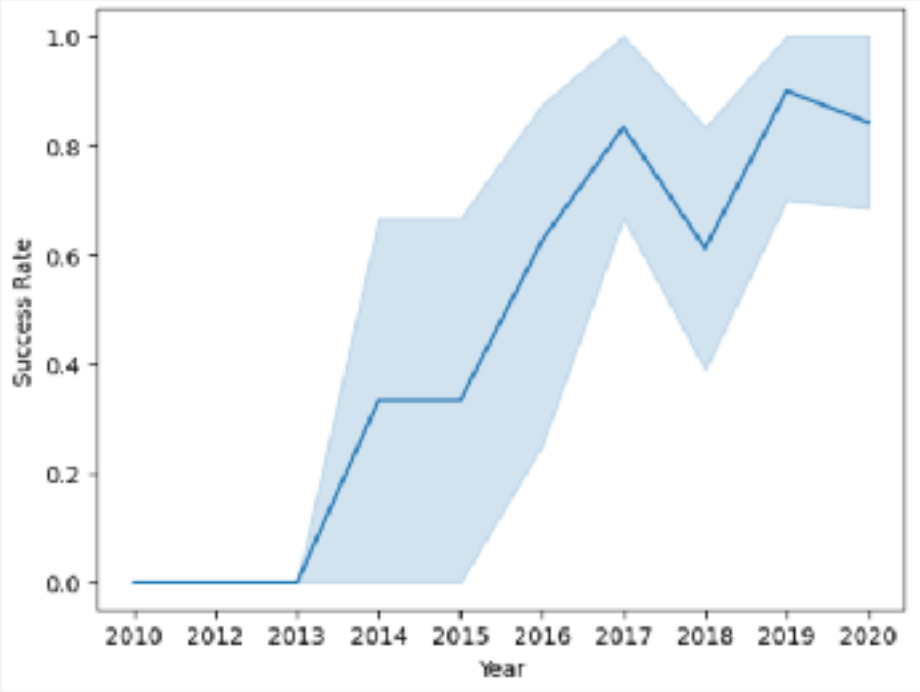
The data wrangling process began by loading SpaceX data into pandas DataFrames. Irrelevant and redundant columns were removed, and column headers were renamed for consistency. I handled missing or inconsistent values. Payload mass units were standardized, and nested or irregular data was flattened to ensure clean, analyzable structure, preparing it for further exploration and modeling.

GitHub URL: https://github.com/cinthyacong/Data-Science-Capstone-SpaceX/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

To gain insights into the SpaceX missions, several charts were plotted for visual analysis:

- Bar Chart: Used to show the average success rate (Class) by orbit type to identify orbits with higher reliability.
- Line Chart: Plotted success rates over the years to analyze performance trends and improvements.
- Scatter Plot: Visualized the relationship between flight number, launch site, and success class to detect spatial and temporal patterns.
- Map (Folium): Plotted launch site locations for geographical context and clustering of launches.



GitHub URL: https://github.com/cinthyacong/Data-Science-Capstone-SpaceX/blob/main/edadataviz.ipynb

EDA with SQL

To better understand the SpaceX dataset, SQL queries were used to explore launch site usage, mission outcomes, payload statistics, and trends over time. Below is a summary of the key queries performed:

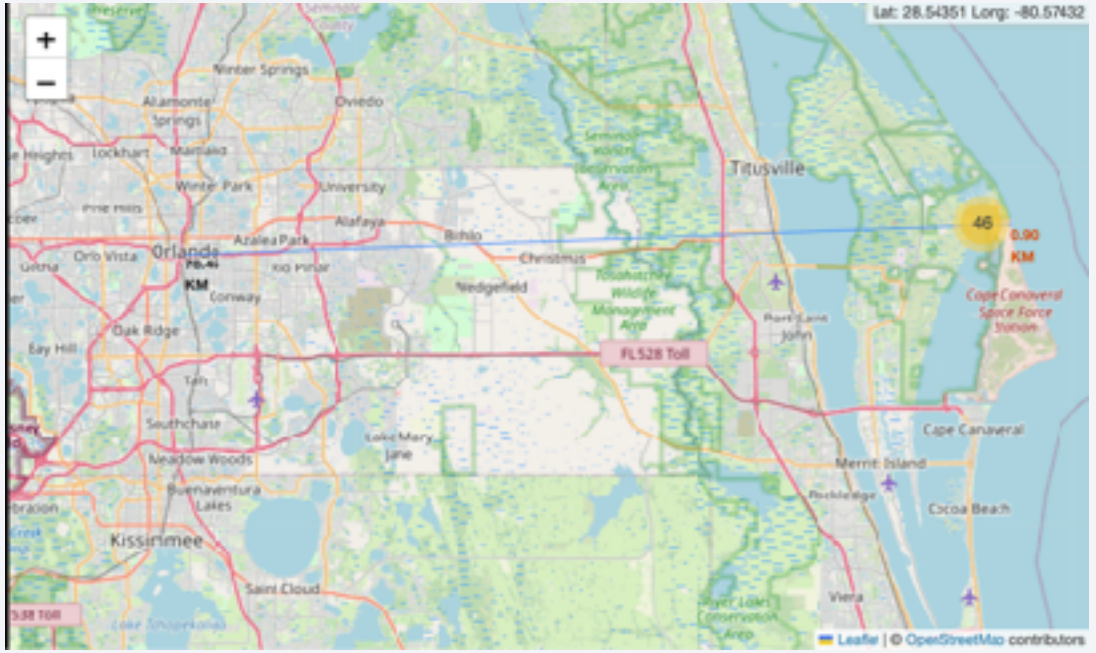
- Selected distinct launch sites to identify all unique SpaceX launch locations.
- Filtered first 5 records of the dataset for initial inspection.
- Queried launches starting with ‘CCA’ to explore site-specific launches.
- Used LIKE and SUBSTR() functions to extract launches from the year 2015 and specific months.
- Summed up payload mass for specific customers (e.g., NASA CRS missions).
- Calculated average success rate by orbit type using GROUP BY and AVG() functions.
- Retrieved monthly breakdowns of launch outcomes using SUBSTR(Date, 6, 2) to extract month.
- Sorted records by date to analyze trends and filter outcomes over time.

GitHub URL: https://github.com/cinthyacong/Data-Science-Capstone-SpaceX/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

To visualize SpaceX launch site data interactively, I used Folium to create a map that includes key spatial elements:

- Markers: Placed on each launch site to identify its geographic location.
- Circle Markers: Added to highlight launch density or significance, with radius and color indicating mission outcomes.
- Popups: Included with markers to display site names and other launch details interactively.
- MarkerCluster: Used to group markers that are geographically close together, enhancing readability when zoomed out.

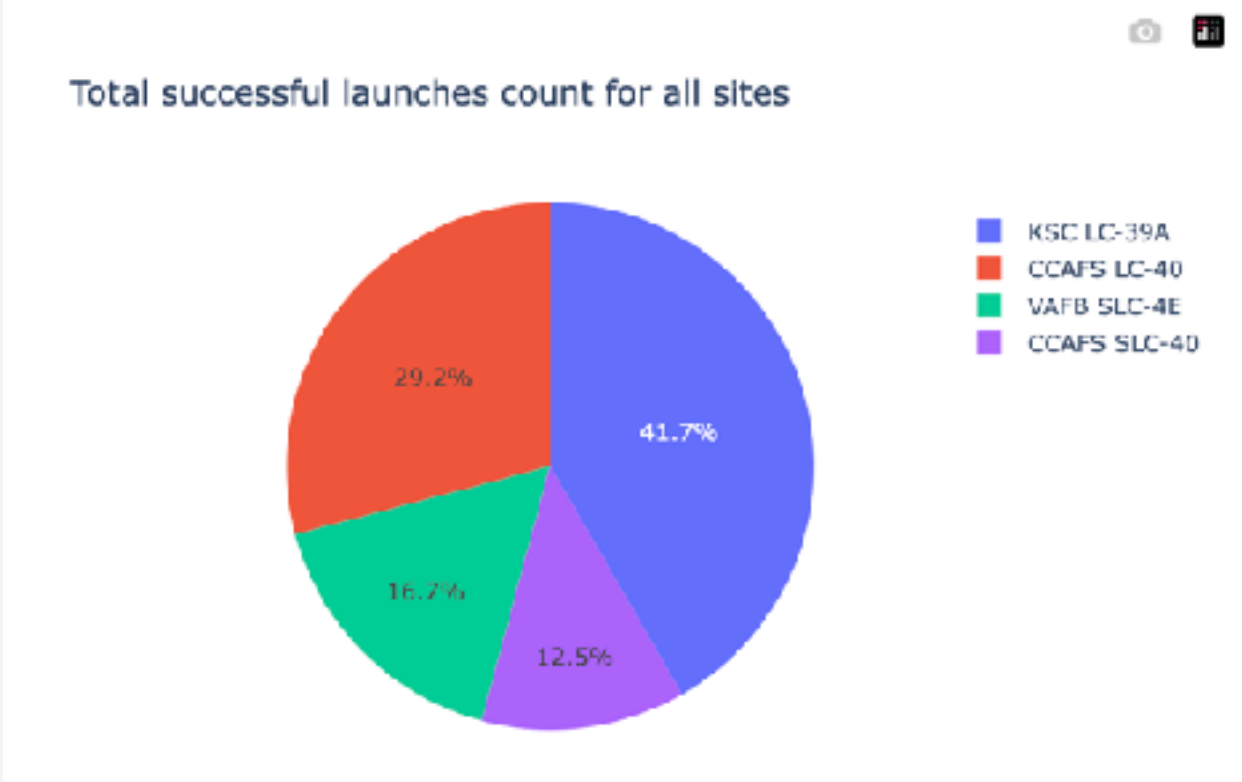


GitHub URL: https://github.com/cinthyacong/Data-Science-Capstone-SpaceX/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

I built a Plotly Dash dashboard with dropdowns to select launch sites and sliders to filter payload range. I added pie charts to display success rates and scatter plots to explore the relationship between payload and mission outcomes. These interactive elements allow me to better understand how different variables influence launch success.

GitHub URL: https://github.com/cinthyacong/Data-Science-Capstone-SpaceX/blob/main/spacex-dash-app.py



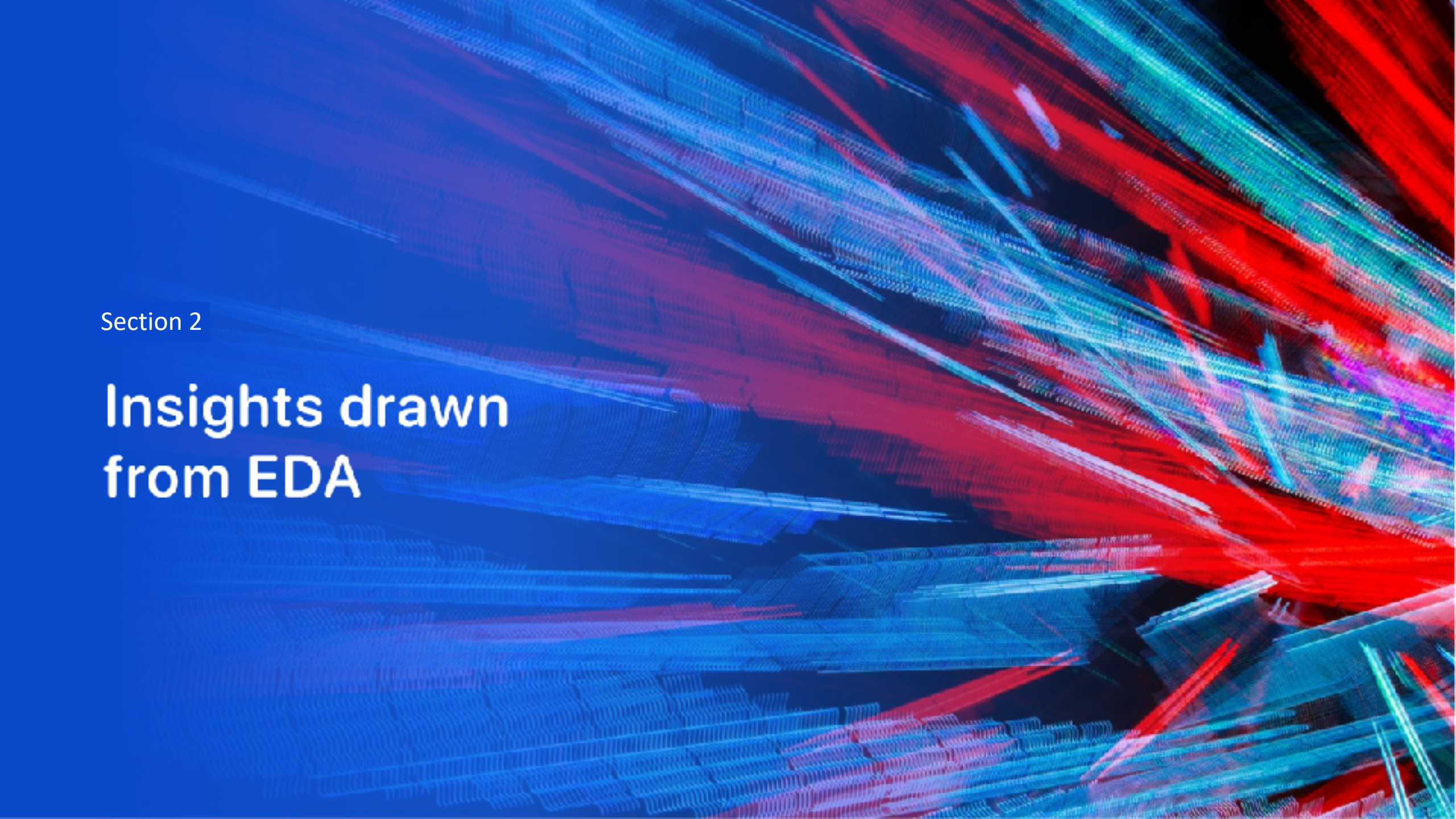
Predictive Analysis (Classification)

I built and evaluated several classification models to predict launch success. First, I prepared and standardized the data, then split it into training and testing sets. I trained models like Logistic Regression, KNN, SVM, and Decision Tree, and evaluated their performance using accuracy scores. To improve results, I applied hyperparameter tuning with GridSearchCV and selected the best-performing model based on overall accuracy.

GitHub URL: https://github.com/cinthyacong/Data-Science-Capstone-SpaceX/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

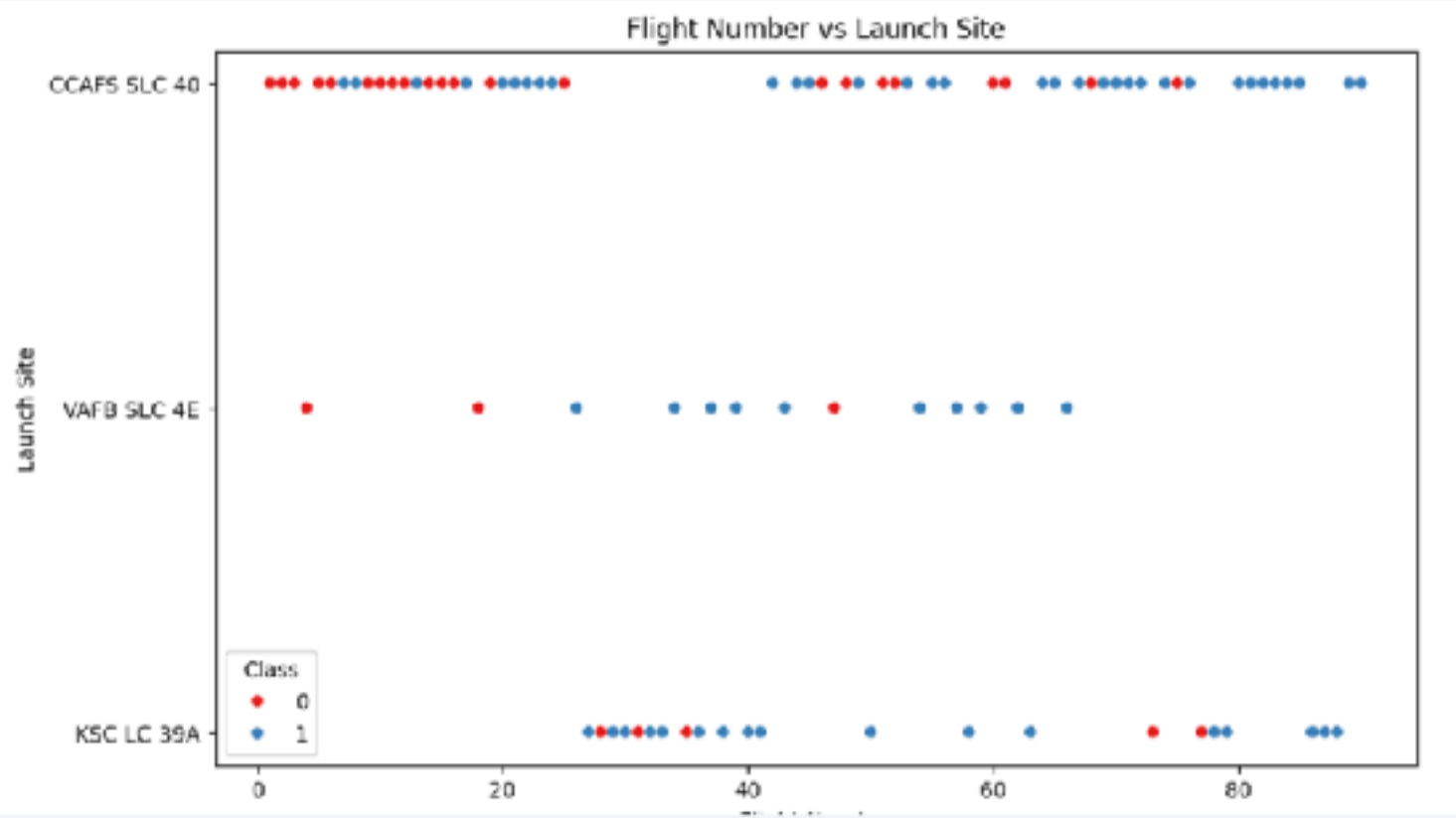
Results

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


Section 2
Insights drawn
from EDA

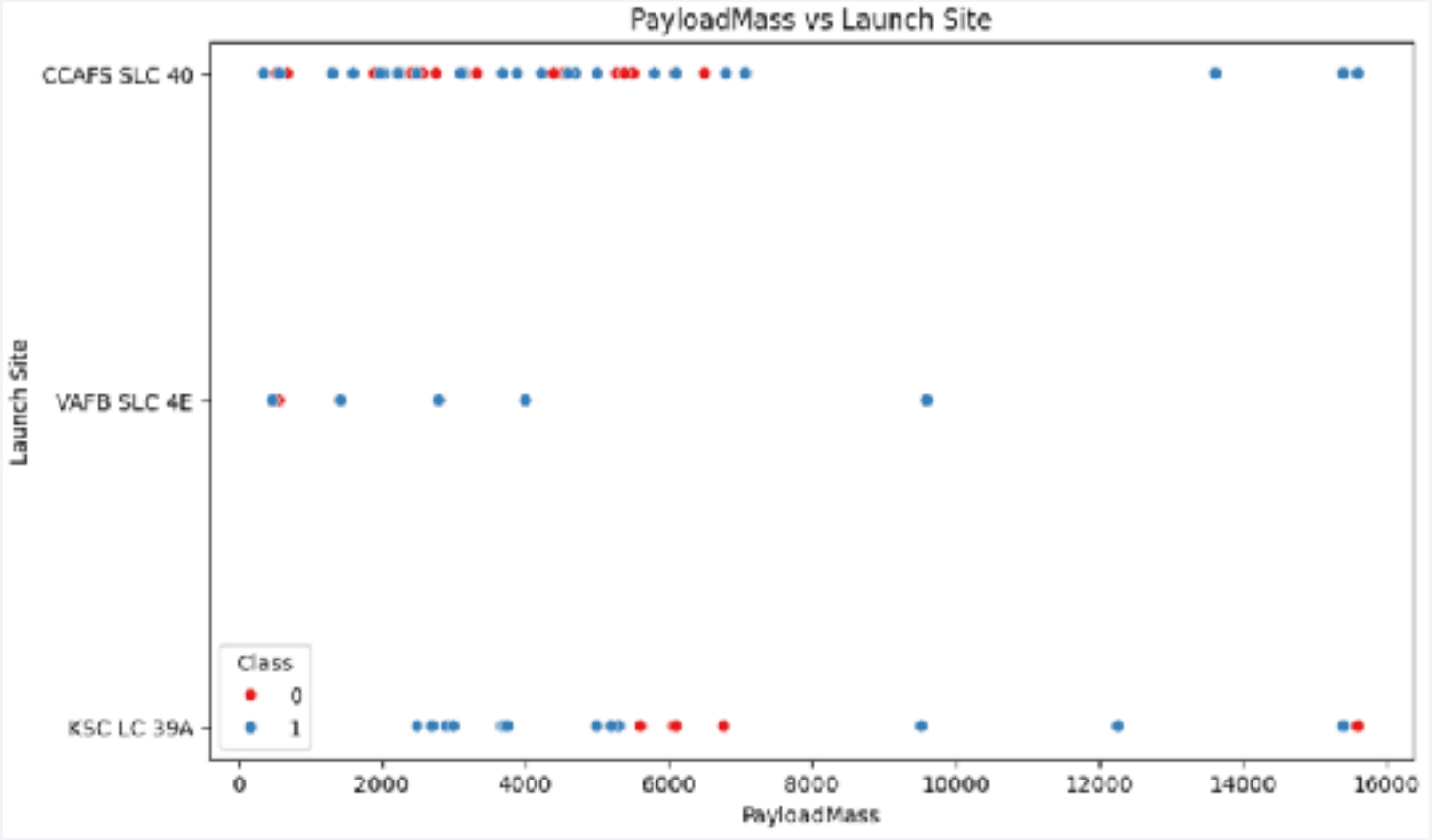
Flight Number vs. Launch Site

As the number of flights increases, the launch success rate also improves across all sites. For example, at VAFB SLC 4E, nearly all launches after the 20th flight were successful, with only one exception. Meanwhile, for the other two launch sites, every launch after the 80th flight was successful. This suggests that experience and continued operations contributed to higher success rates over time.



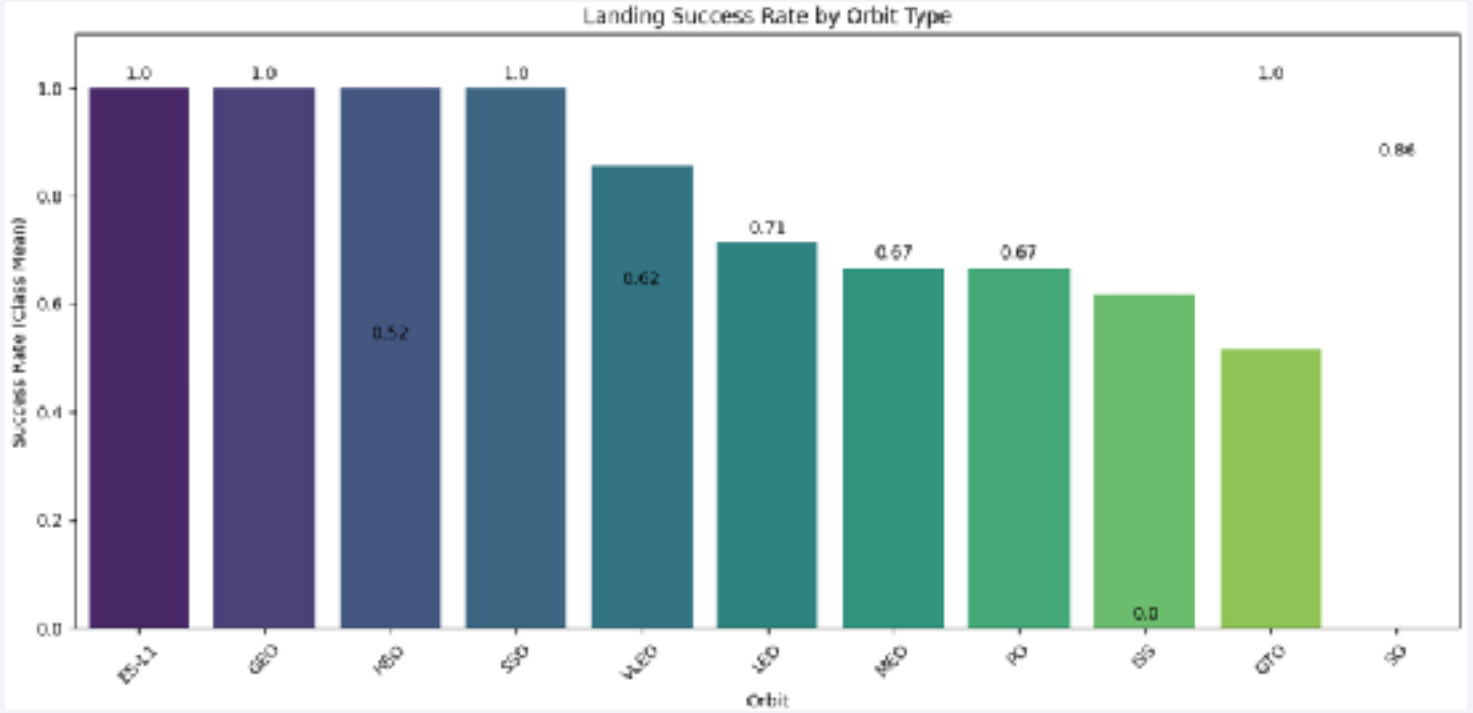
Payload vs. Launch Site

VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000.



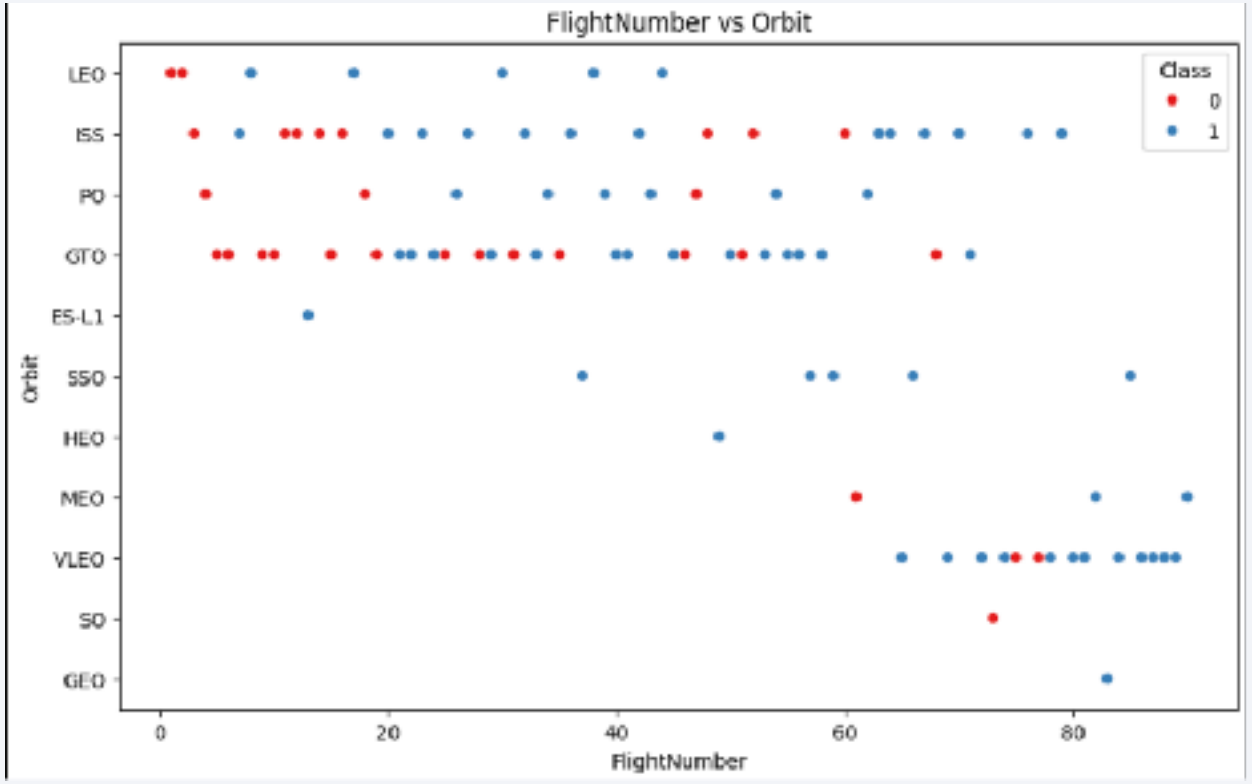
Success Rate vs. Orbit Type

The bar chart shows that some orbits, like ES-L1, GEO, HEO, and SSO, had a perfect 100% success rate. In contrast, the SO orbit had no successful launches. Other orbits, from VLEO to GTO, show a gradual decline in success rates, indicating that certain orbit types are more reliable than others for successful missions.



Flight Number vs. Orbit Type

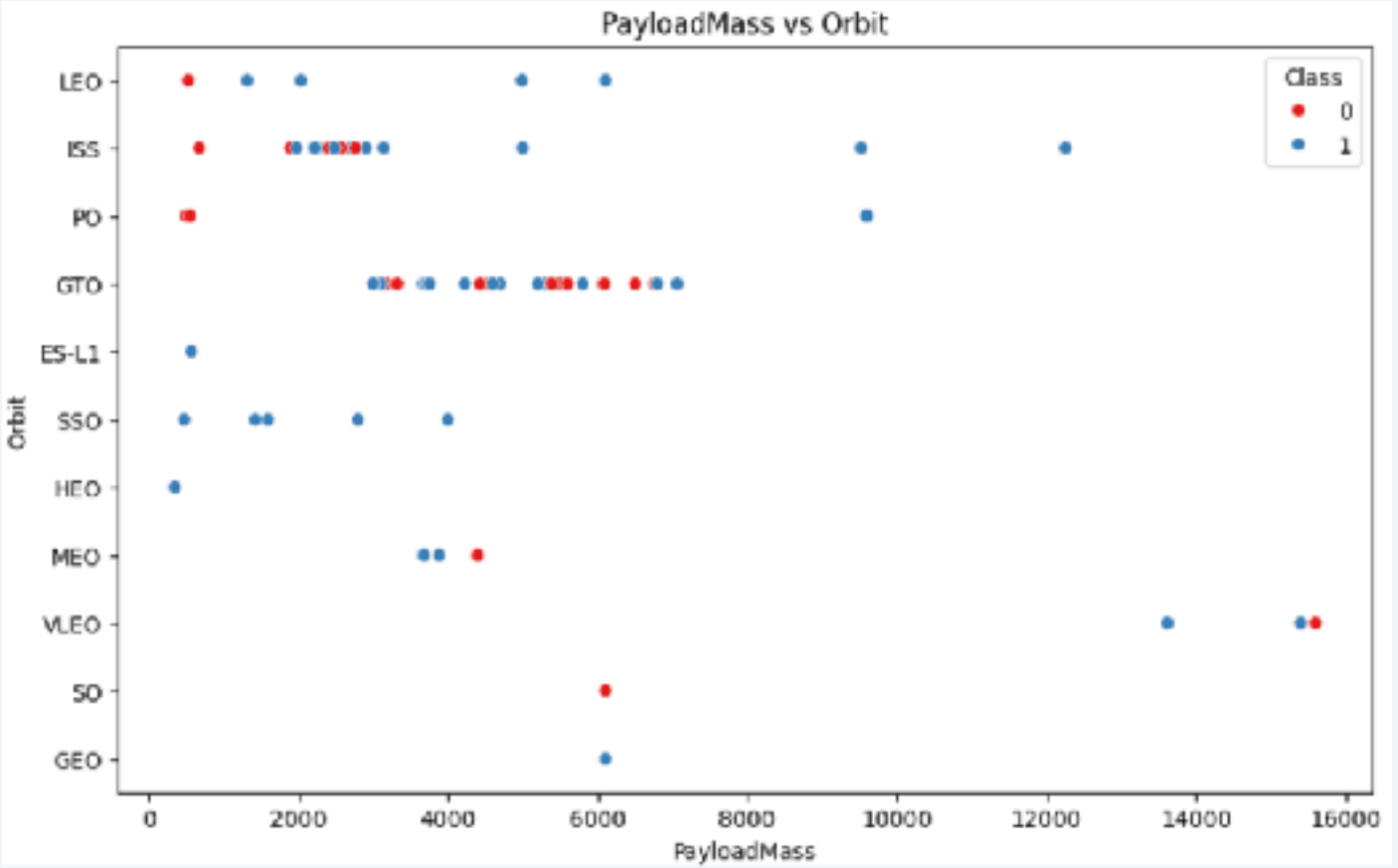
You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.



Payload vs. Orbit Type

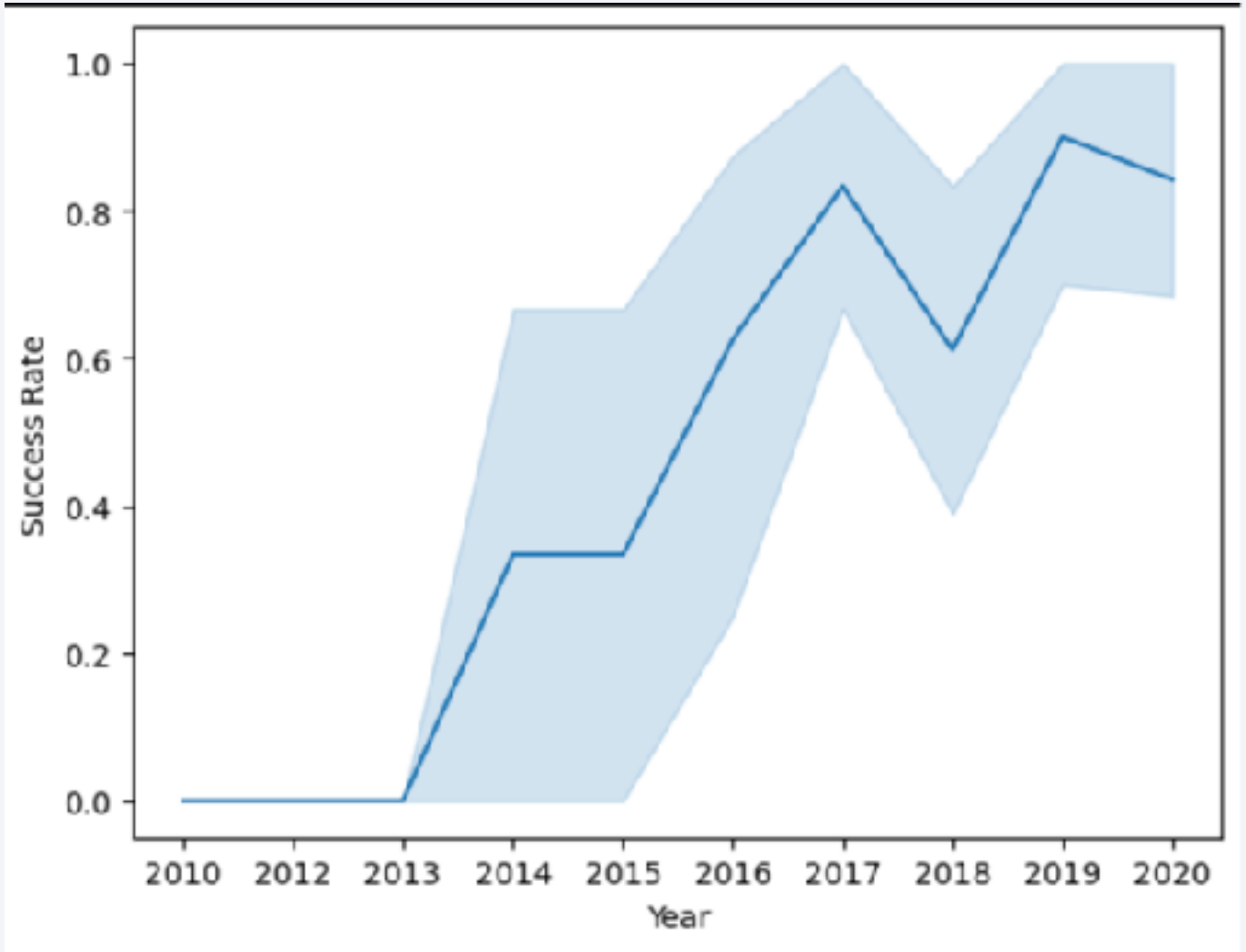
With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch Success Yearly Trend

You can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

The query returns a list of all unique launch sites from the SPACEXTBL table. This helps us identify the different locations SpaceX has used for launching their missions, which is useful for analyzing site-specific performance or trends.

```
In [15]: %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

Out[15]:

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

Launch Site Names Begin with 'CCA'

The query retrieves the first 5 records from the SPACEXTBL table where the LAUNCH_SITE begins with 'CCA'. This filter helps focus on launches that occurred at the Cape Canaveral area, allowing for a targeted analysis of launch activity at that site.

```
In [24]: %sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5
 * sqlite:///my_data1.db
Done.
```

Out[24]:

_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
In [27]: %sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
Out[27]: SUM(PAYLOAD_MASS__KG_)
                          45596
```

This SQL command calculates the total payload mass for all missions where the customer was NASA (CRS). The function SUM() adds up all the values in the PAYLOAD_MASS__KG_ column that meet the condition. This gives insight into how much cargo NASA has transported via SpaceX for CRS missions.

Average Payload Mass by F9 v1.1

```
In [28]: %sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like 'F9 v1.1%'
 * sqlite:///my_data1.db
Done.
Out[28]: AVG(PAYLOAD_MASS__KG_)
         2534.6666666666665
```

This command calculates the average payload mass for all launches that used the F9 v1.1 booster version. The LIKE 'F9 v1.1%' filter ensures that only those records starting with "F9 v1.1" in the Booster_Version column are included. This helps analyze the payload capacity performance of the F9 v1.1 booster variant.

First Successful Ground Landing Date

```
In [29]: %sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
 * sqlite:///my_data1.db
Done.
Out[29]: min(DATE)
         2015-12-22
```

This SQL command retrieves the earliest launch date (MIN(DATE)) where the booster successfully landed on a ground pad. It's useful for identifying the first successful ground landing achieved by SpaceX, helping track milestones in their reusability progress.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
[32]: %sql select Booster_Version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)'
      and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

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

[32]:

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

This SQL query retrieves booster versions that successfully landed on drone ships with payloads between 4000 and 6000 kg, highlighting which boosters perform well under mid-range payload conditions. It provides insights into the capabilities of different booster versions concerning payload weight and landing success on drone ships.

Total Number of Successful and Failure Mission Outcomes

```
[33]: %sql select count(Mission_Outcome) from SPACEXTBL where Mission_Outcome = 'Success'
          or Mission_Outcome = 'Failure (in flight)'

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

[33]:

count(Mission_Outcome)
99

This SQL query counts the number of missions that either succeeded or failed during flight, helping to evaluate overall mission outcomes.

Boosters Carried Maximum Payload

This SQL query returns the booster version used for the mission with the maximum payload mass. It helps identify which booster handled the heaviest payload in the dataset.

```
[34]: %sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_=
          (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
 * sqlite:///my_data1.db
Done.
```

[34]:

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

2015 Launch Records

```
[38]: %sql SELECT strftime('%m', "Date") AS "Month", "Landing_Outcome",
          "BoosterVersion", "LaunchSite" FROM SPACEXTBL WHERE "Landing_Outcome"
          LIKE 'Failure%' AND substr("Date", 1, 4) = '2015' ORDER BY "Date" DESC;
```

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

[38]:

Month	Landing_Outcome	"BoosterVersion"	"LaunchSite"
04	Failure (drone ship)	BoosterVersion	LaunchSite
01	Failure (drone ship)	BoosterVersion	LaunchSite

This SQL query retrieves records from the SPACEXTBL table where the landing outcome was a failure and the launch took place in 2015. It extracts the month from the date, along with the landing outcome, booster version, and launch site, and orders the results by date in descending order to help identify when and where failures occurred and which boosters were involved.

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

The SQL query retrieves all records from the SPACEXTBL table where the landing outcome was either 'Failure (drone ship)' or 'Success (ground pad)' and the launch date was between '2010-06-04' and '2017-03-20'. It orders the results by date in descending order to prioritize the most recent relevant missions.

```
[40]: %sql select * from SPACEXTBL where Landing_Outcome = 'Failure (drone ship)'
          or Landing_Outcome = 'Success (ground pad)' and (DATE between '2010-06-04'
      and '2017-03-20') order by date desc
```

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

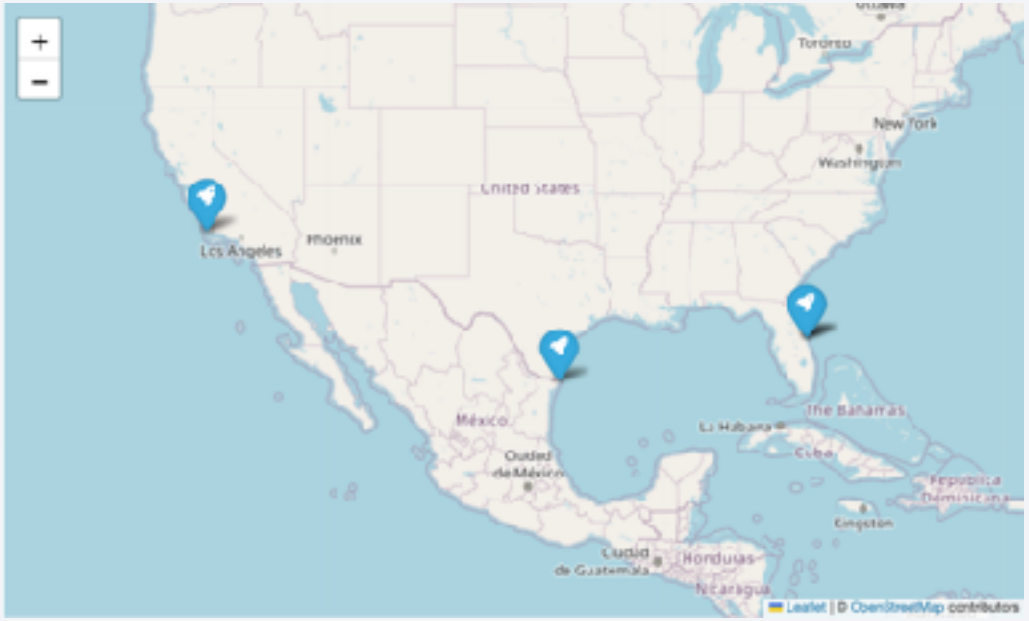
[40]:

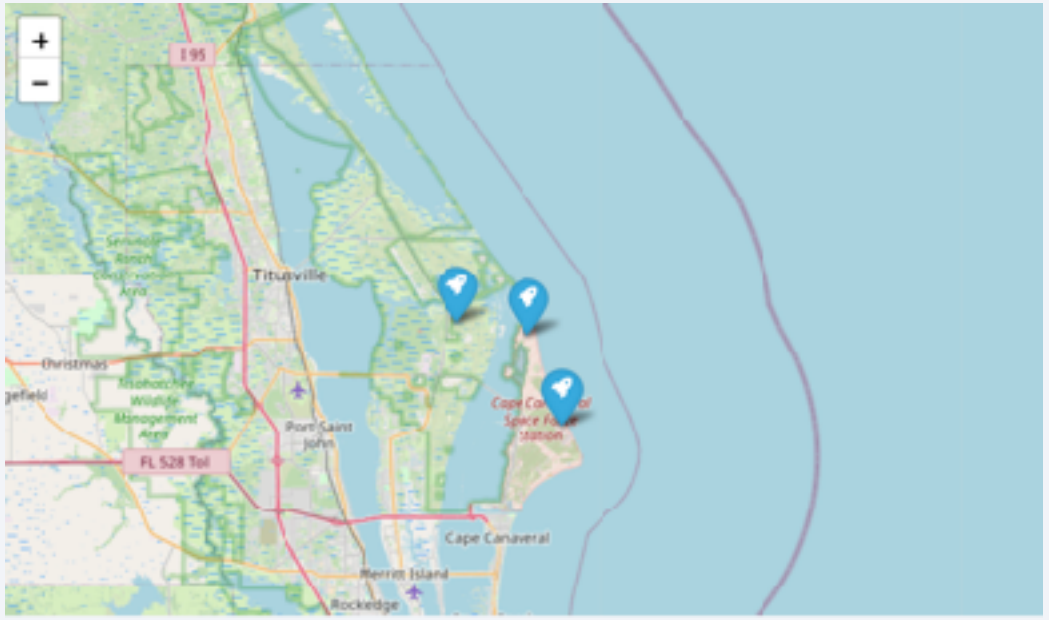
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outco
2017-12-15	15:36:00	F9 FT B1035.2	CCAFS SLC-40	SpaceX CRS-13	2205	LEO (ISS)	NASA (CRS)	Succ
2017-09-07	14:00:00	F9 B4 B1040.1	KSC LC-39A	Boeing X-37B OTV-5	4990	LEO	U.S. Air Force	Succ
2017-08-14	16:31:00	F9 B4 B1039.1	KSC LC-39A	SpaceX CRS-12	3310	LEO (ISS)	NASA (CRS)	Succ
2017-06-03	21:07:00	F9 FT B1035.1	KSC LC-39A	SpaceX CRS-11	2708	LEO (ISS)	NASA (CRS)	Succ
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Succ
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Succ
2016-07-18	4:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Succ
2016-06-15	14:29:00	F9 FT B1024	CCAFS LC-40	ABS-2A Eutelsat 117 West B	3600	GTO	ABS Eutelsat	Succ
2016-			CCAFS LC-					

Section 3

Launch Sites Proximities Analysis

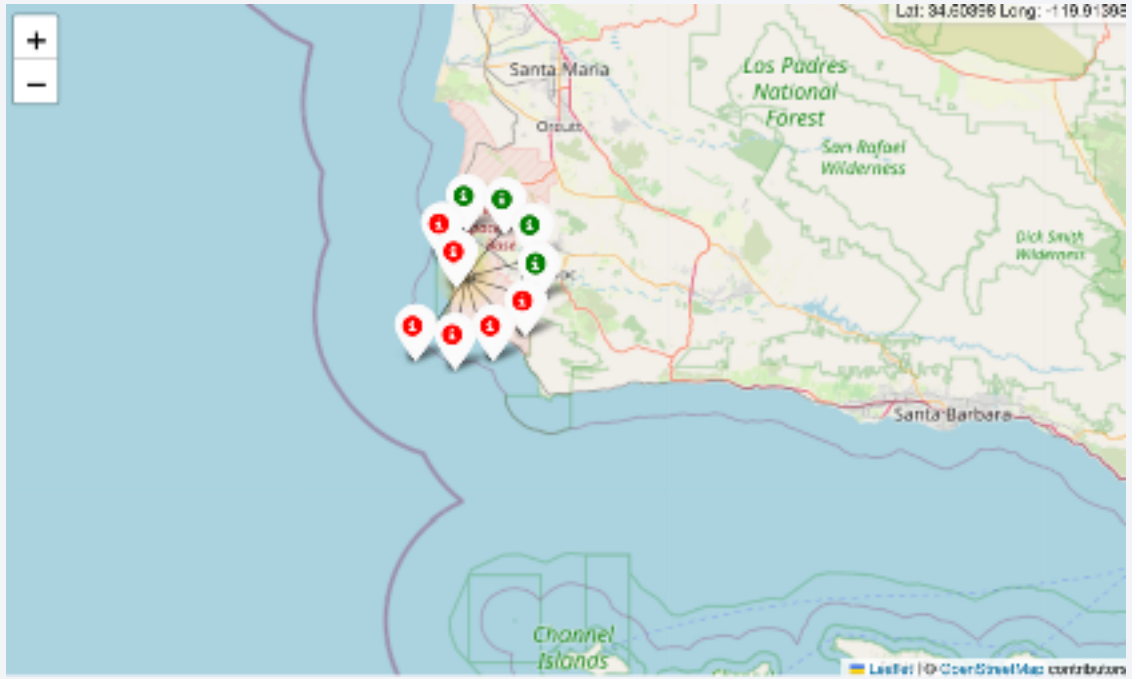
All launch sites on a map

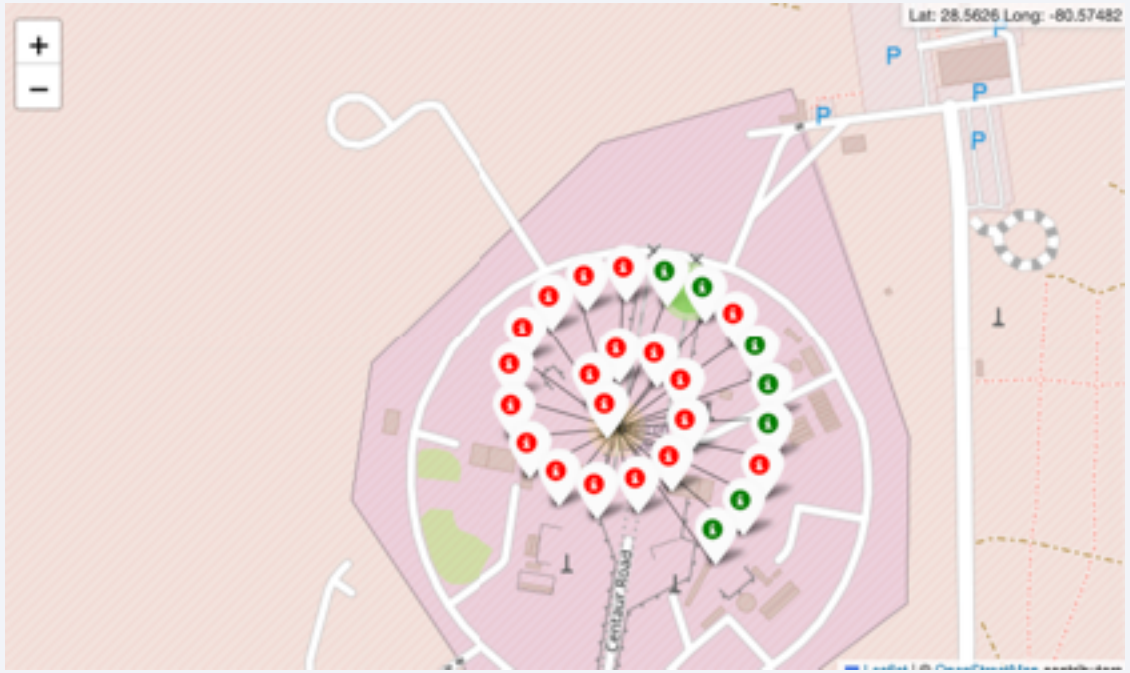




The image shows an interactive map with SpaceX launch sites marked in California, Texas, and Florida. The markers highlight the geographic distribution of launch operations.

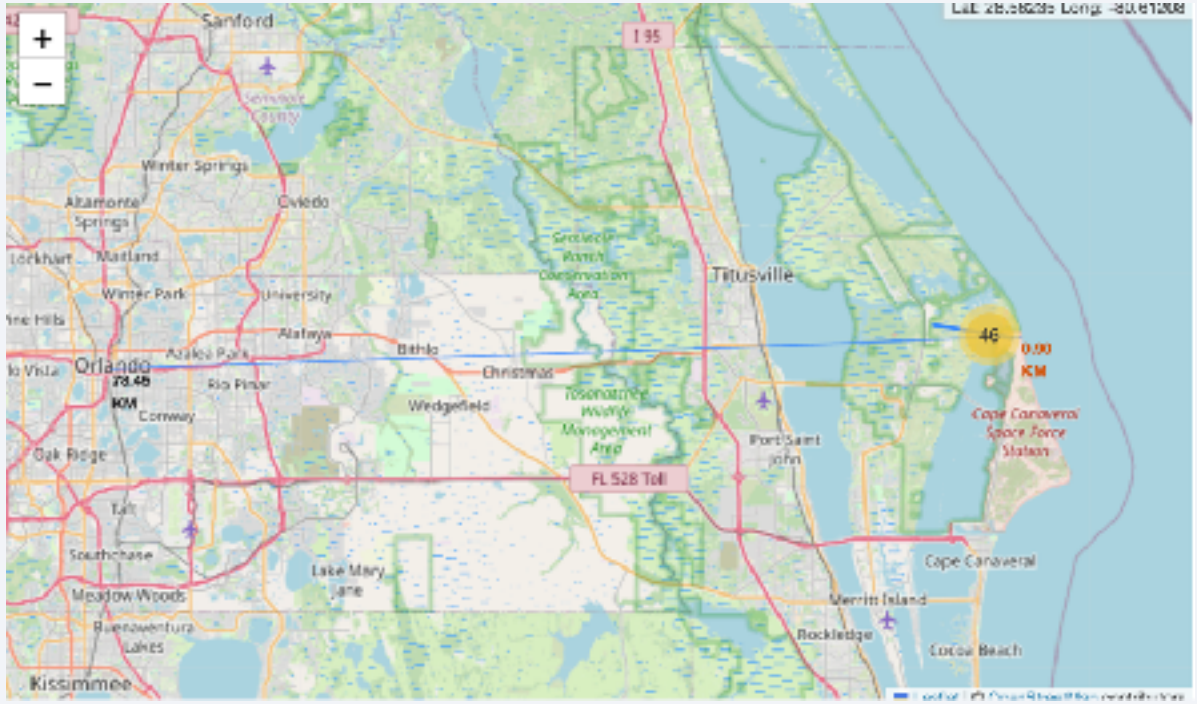
Success/failed launches for each site on the map





It provides a quick and intuitive way to distinguish between successful and failed missions. This visual differentiation helps identify patterns, such as which launch sites or regions tend to have higher success rates, and supports data-driven decision-making in mission planning and resource allocation.

Distances between a launch site to its proximities

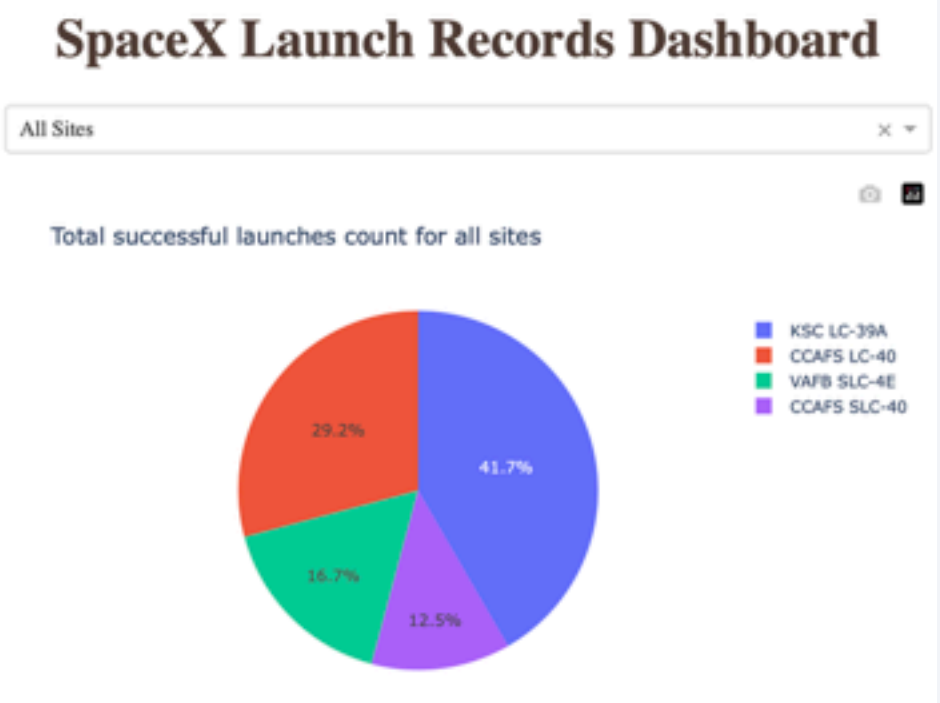


It helps assess logistical and safety considerations, this includes evaluating ease of transport for heavy equipment, accessibility for personnel, and risk management in case of launch failures. It supports strategic planning for infrastructure and launch site optimization.

Section 4

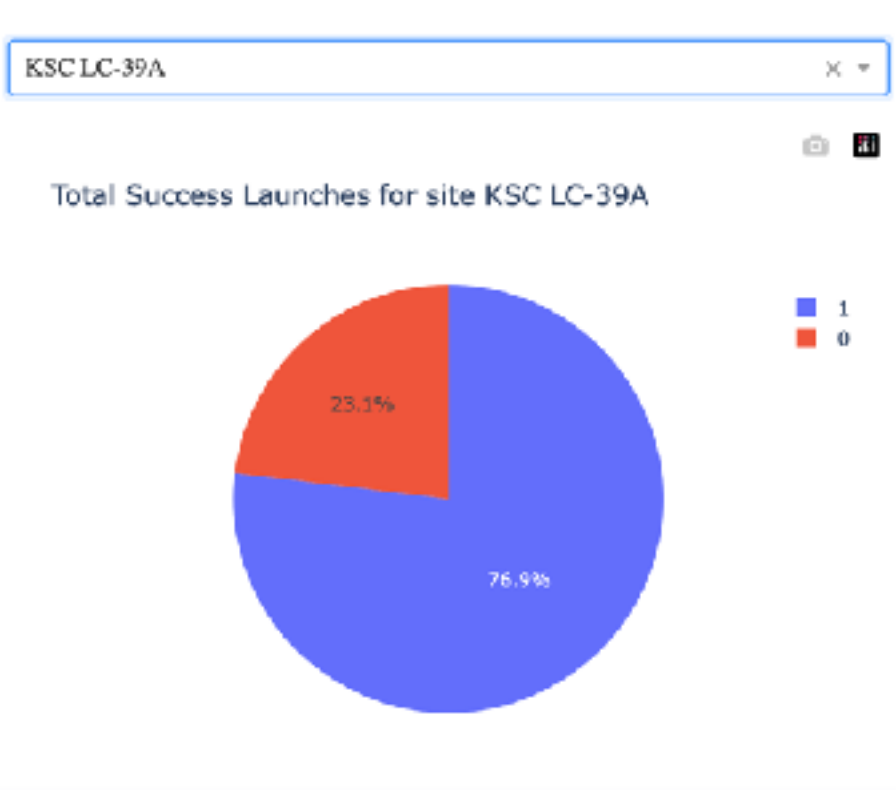
Build a Dashboard with Plotly Dash

Success vs. Failed counts for the site



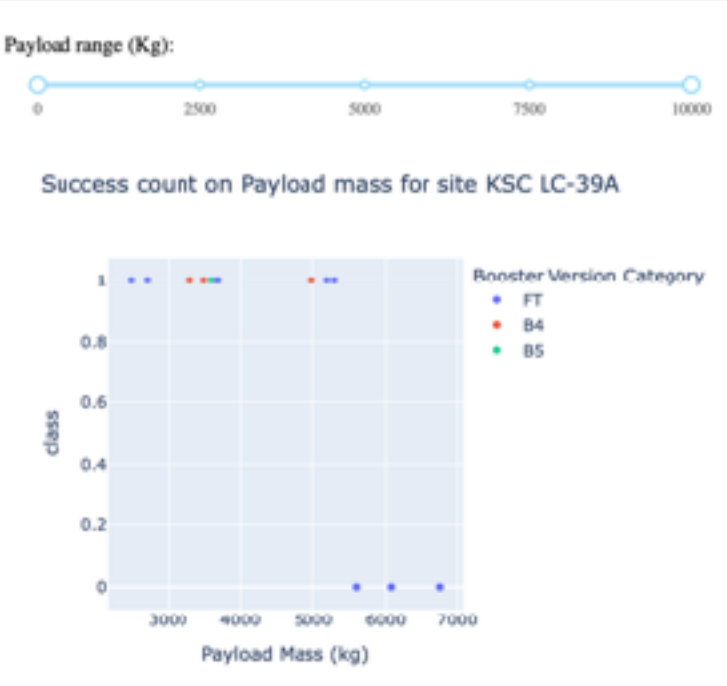
A pie chart of launch success counts helps quickly compare site performance, highlighting the most successful launch locations for better planning and decision-making. We can see that KSC LC-39A has the most successful launches from all the sites.

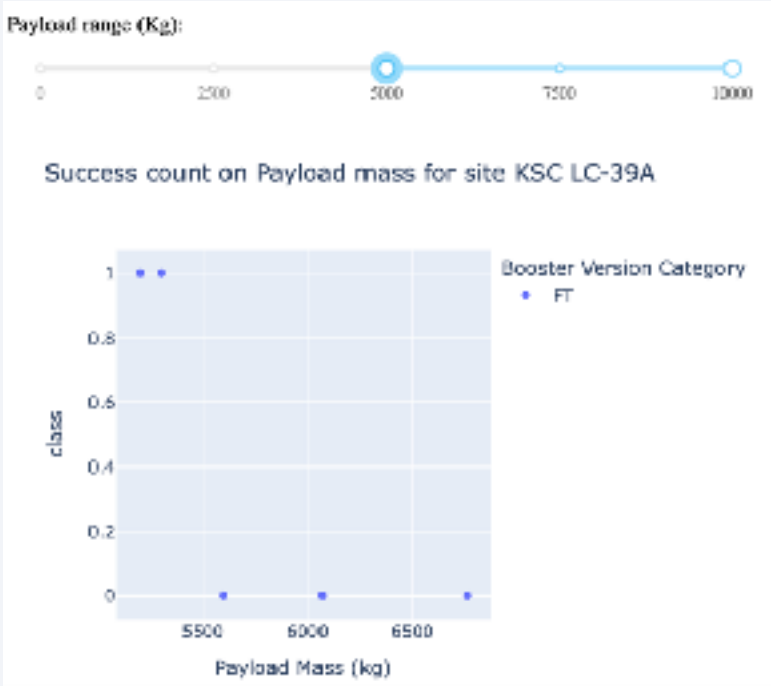
Success Launches for site KSC LC-39A



KSC LC-39A achieved a launch success rate of 76.9%, with 23.1% of its launches ending in failure.

<Dashboard Screenshot 3>





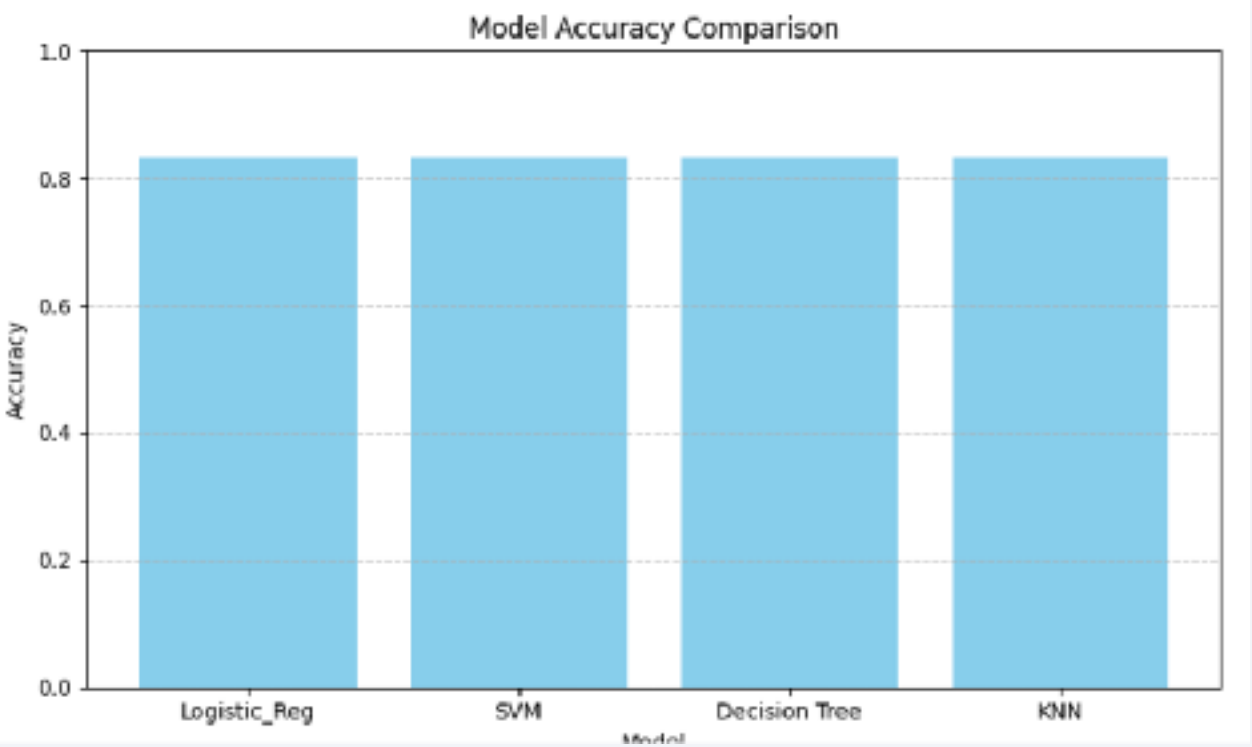
We can observe that with larger payloads and the use of the FT booster, the success rate is significantly higher.

Section 5

Predictive Analysis (Classification)

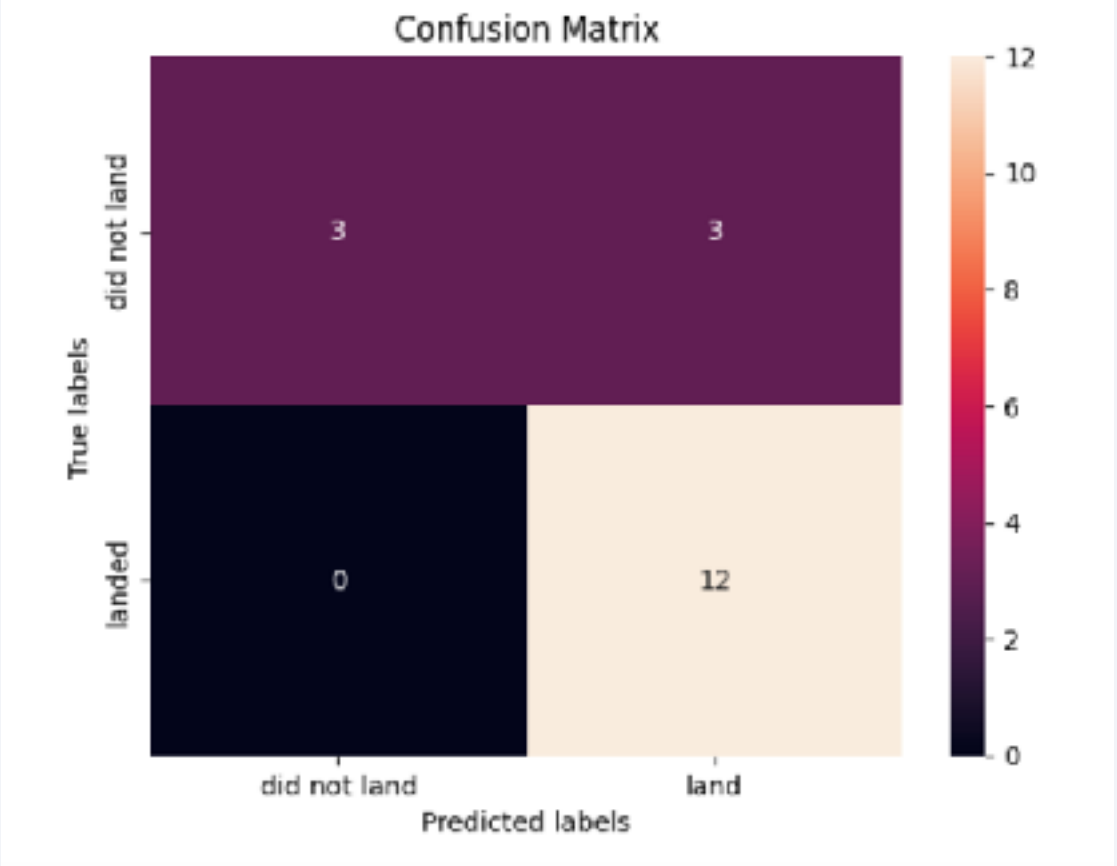
Classification Accuracy

The decision tree model achieved the highest accuracy at 88%, outperforming the other models, which reached around 83%.



Confusion Matrix

Predicts all successful and most unsuccessful landings, although it mistakes 3 unsuccessful landings as successful.



Conclusions

The analysis shows that mission success rates vary depending on the launch site, orbit type, payload mass, and flight number. Sites like KSC LC-39A and VAFB SLC-4E reach a 100% success rate after a certain number of flights, whereas CCAFS LC-40 starts with a lower success rate. Orbits such as ES-L1, GEO, HEO, and SSO show perfect success, while the SO orbit has the lowest (~0–50%). As the number of flights increases—especially from 2013 to 2020—success rates improve. Heavier payloads tend to succeed more frequently in Polar, LEO, and ISS orbits, while GTO results are more mixed. Notably, VAFB-SLC 4E has no launches with payloads above 10,000 kg, and in LEO, higher flight numbers correlate with greater success, unlike GTO where no clear pattern is observed.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!