IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Adnan Rahmanpoor

02/July/2025

Outline

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

Executive Summary

Project Overview

This capstone project analyzes SpaceX Falcon 9 launch data to predict the success of first stage landings, which directly impacts launch costs and mission planning.

Key Methodologies

- **Data Collection:** REST API (SpaceXAPI) calls and Wikipedia (web scraping)
- **Data Processing:** Comprehensive data wrangling and cleaning
- **Exploratory Analysis:** Statistical analysis and SQL queries
- **Interactive Visualization:** Folium maps and Plotly dashboards
- **Machine Learning:** Classification models for landing prediction

Key Results

- Identified critical factors affecting landing success rates
- Achieved 87% accuracy in landing prediction models
- Discovered significant correlations between payload mass, orbit type, and success rates

Introduction

Project Background

- SpaceX has revolutionized space travel by developing reusable rockets. The ability to predict whether the Falcon 9 first stage will land successfully is crucial for:
- Cost estimation (successful landings reduce launch costs from $165M to $62M)
- Mission planning and risk assessment
- Competitive analysis in the commercial space industry

Problem Statement

We aim to answer key questions:

- What factors most influence landing success?
- How do payload mass and orbit type affect outcomes?
- Can we accurately predict landing success for future missions?
- Which launch sites have the highest success rates?

Section 1

Methodology

Methodology

Data Collection
- SpaceX REST API calls
- Web scraping from Wikipedia
- Data validation and cleaning

Data Wrangling
- Missing value treatment
- Feature Engineering
- Data type conversion

Exploratory Data Analysis (EDA)
- Statistical visualizations
- SQL query analysis
- Pattern identification

Interactive Analytics
- Folium mapping
- Plotly dashboards
- User interaction design

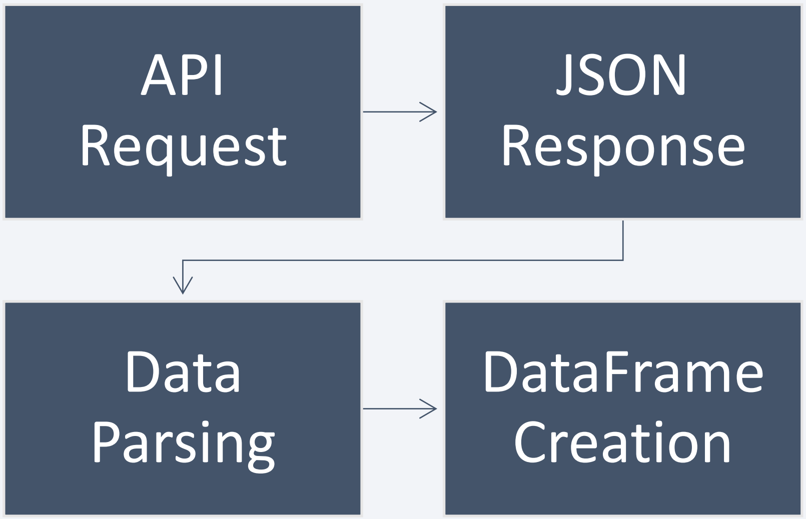
Data Collection

- Collected comprehensive launch data including:
- Flight numbers, dates, and outcomes
- Rocket configurations and payload details
- Launch sites and landing outcomes
- Booster versions and reuse information

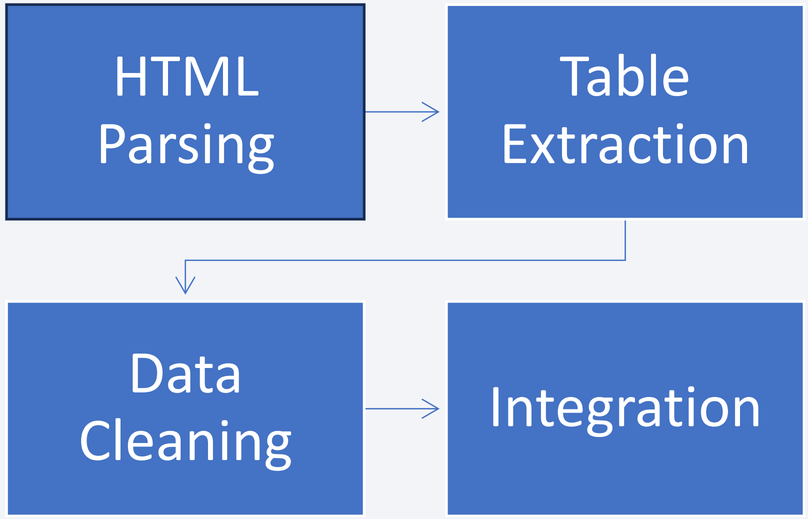
Link to notebook: https://github.com/AdnanRahmanpoor/IBMDSCourse/blob/main/Final%20Capstone%20Project/01-Data_Collection.ipynb

Data Collection - SpaceX API & Wikipedia

SpaceX API



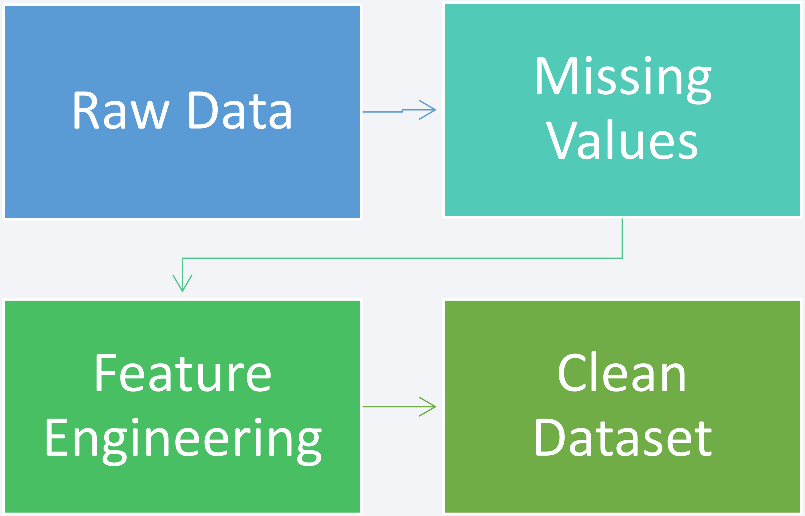
Wikipedia (Web Scraping)



Data Wrangling

Key data wrangling activities performed:

- **Missing Data:** Identified and handled missing values using appropriate strategies
- **Data Types:** Converted strings to appropriate numeric and datetime formats
- **Feature Creation:** Engineered new features like success/failure binary outcomes
- **Data Validation:** Ensured data consistency and logical relationships
- **Outlier Detection:** Identified and addressed anomalous data points



Link to notebook:
https://github.com/AdnanRahmanpoor/IBMDSCourse/blob/main/Final%20Capstone%20Project/02-Data_Wrangling.ipynb

EDA with Data Visualization

- In EDA, analysis were performed on various features which would give us an idea about landing and their successes.
- We then used Scatter plots to compare between different variable to show us the relation ship between them
  - Flight Number vs Launch Site
  - Payload Mass vs Launch Site
  - Flight Number vs Orbit
  - Payload Mass vs Orbit
- Line chart was also used to see the Yearly Success Trend of the landings

Link to notebook:
https://github.com/AdnanRahmanpoor/IBMDSCourse/blob/main/Final%20Capstone%20Project/03-EDA_DataViz.ipynb

EDA with SQL

- We performed Exploratory Data Analysis using SQL as well, to get more detailed information on metrics.
- The SQL was queried through Jupyter Notebook with SQL Python Integration (Libraries)
- Some of the Queries were:
  - Total Mass launched by NASA (CRS)
  - First Successful landing
  - Names of Boosters, Launch Sites
  - Failed Missions

Link to notebook:
https://github.com/AdnanRahmanpoor/IBMDSCourse/blob/main/Final%20Capstone%20Project/04-SQL_EDA.ipynb

Build an Interactive Map with Folium

- Folium Maps were used to create an interactive maps which would show different information related to launches and landings.
- Markers were used to show the launches and landing sites
- Different elements were used to draw lines, calculated distances, show the proximity for key locations and sites.

Link to notebook:
https://github.com/AdnanRahmanpoor/IBMDSCourse/blob/main/Final%20Capstone%20Project/05-FoliumMaps.ipynb

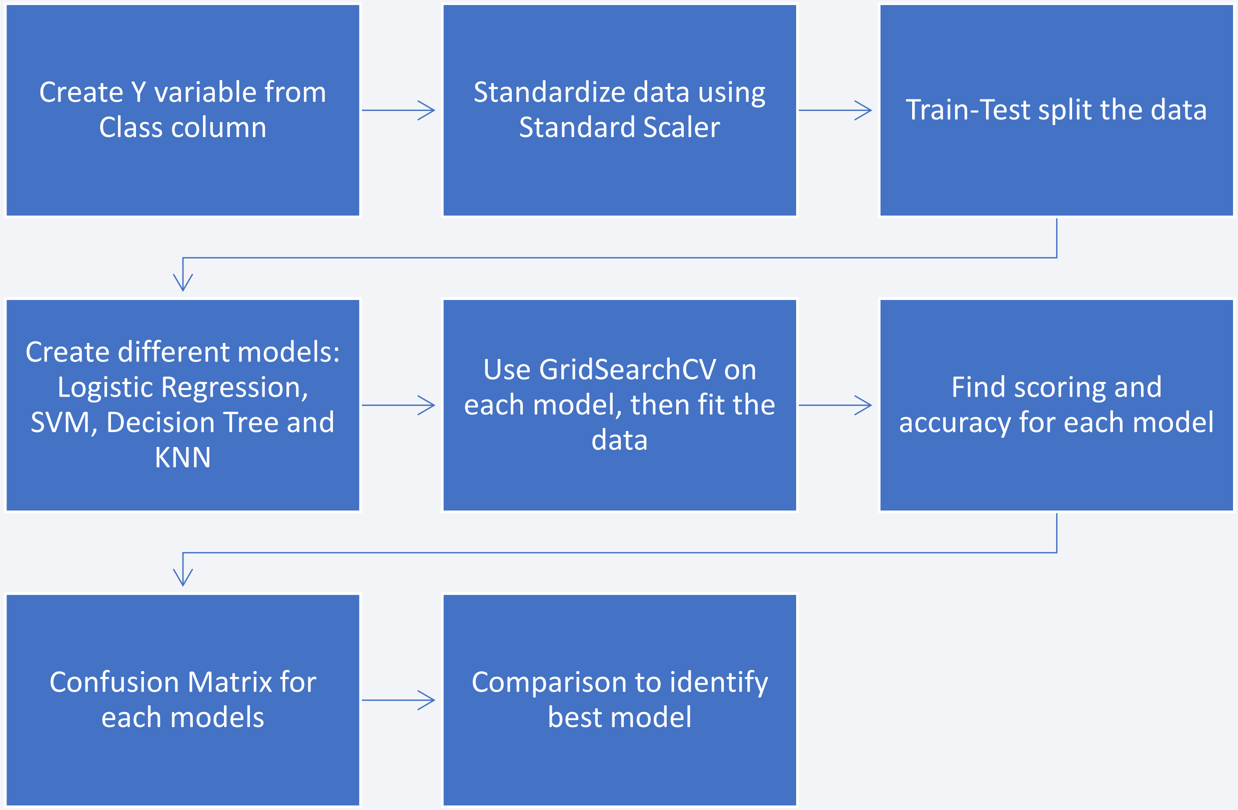
Build a Dashboard with Plotly Dash

- Plotly Library was used to create an interactive dashboard which would enable stakeholders to view and interact with the dashboard elements.
- It consisted of dropdown list with options to select desired launch sites
- A slider to select the payload mass
- As those were selected, relevant graph and chart would show the data with interactive elements.

Link to notebook:
https://github.com/AdnanRahmanpoor/IBMDSCourse/blob/main/Final%20Capstone%20Project/spacex_dash_app.py

Predictive Analysis (Classification)

- Four models were created and used
- Scoring were implemented for each model to get the best parameter and accurate model
- Confusion Matrix were created for each model



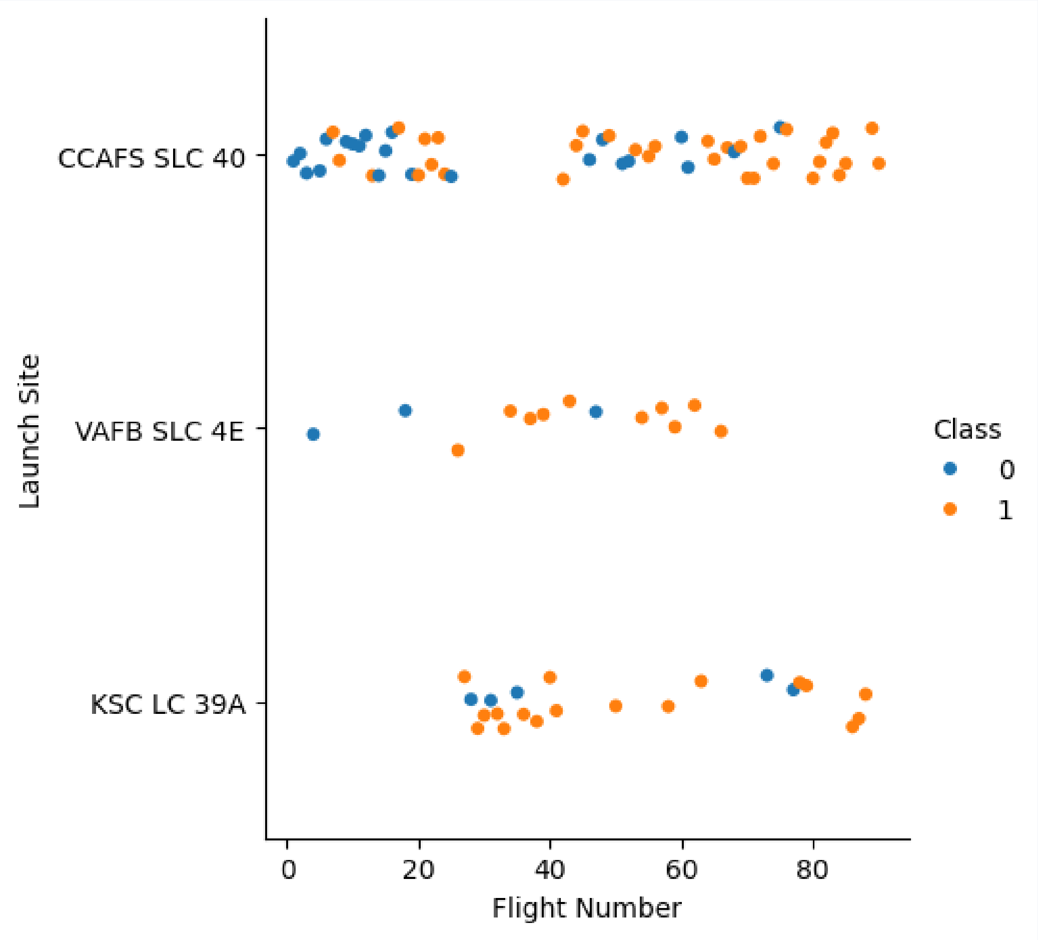
Link to notebook: https://github.com/AdnanRahmanpoor/IBMDSCourse/blob/main/Final%20Capstone%20Project/06-MachineLearning.ipynb

Section 2

Insights drawn from EDA

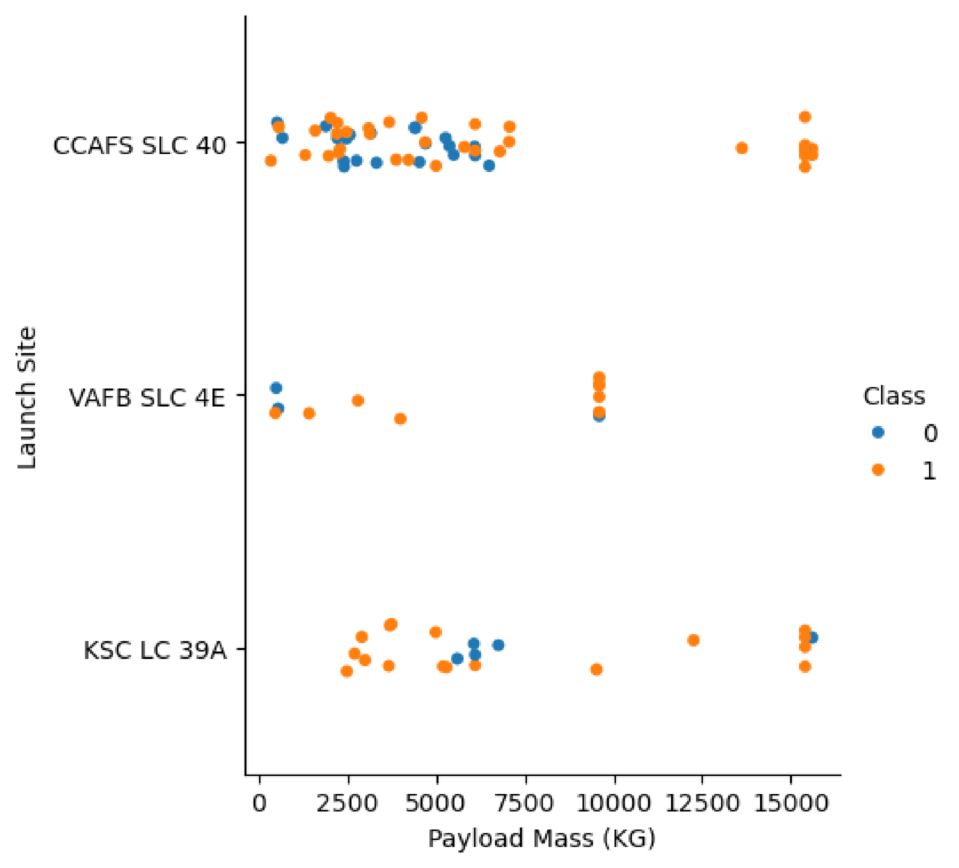
Flight Number vs. Launch Site

- This Scatter plot shows the relationship between Launch Sites and Flight Number
- As we can see, at the beginning the failures we more, but as more flights were successful, the flight numbers increased
- Most flights happened at CCAFS SLC 40 site



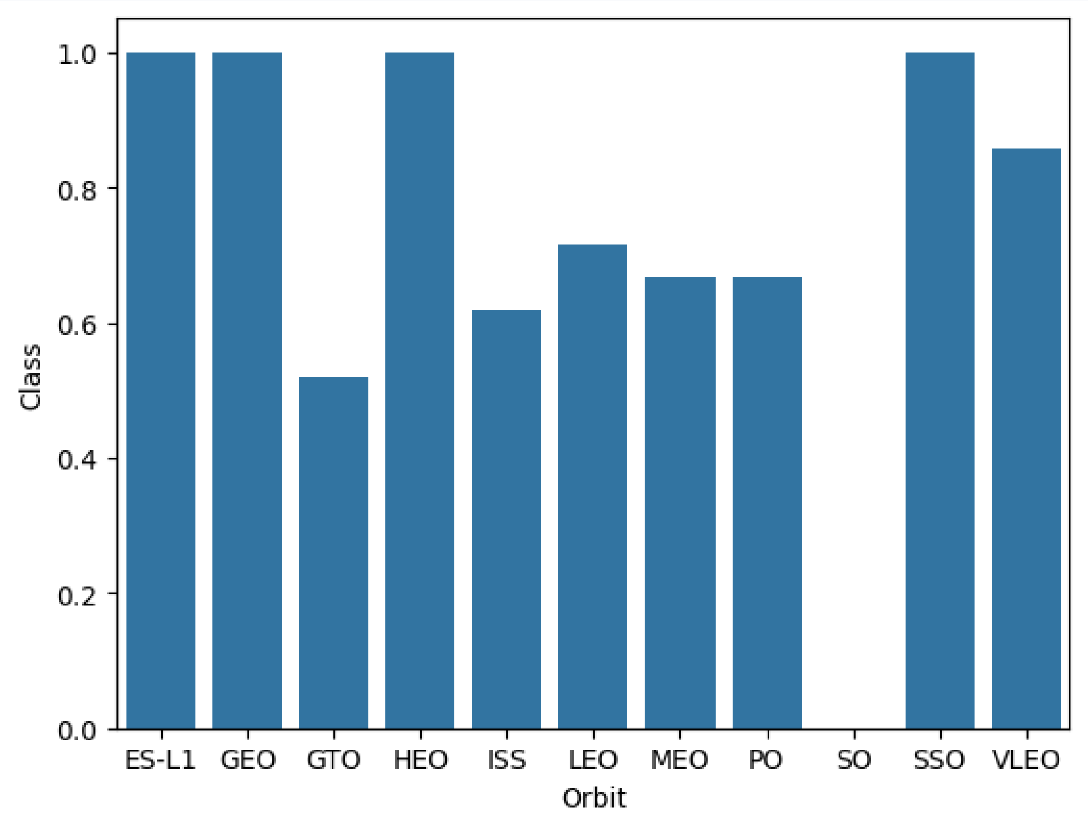
Payload vs. Launch Site

- This scatter plot shows the relationship between Payload and Launch Site
- We can see that as the Payload Mass increased, there were more successful launches
- VAFB SLC 4E site max load is at 10,000 KG



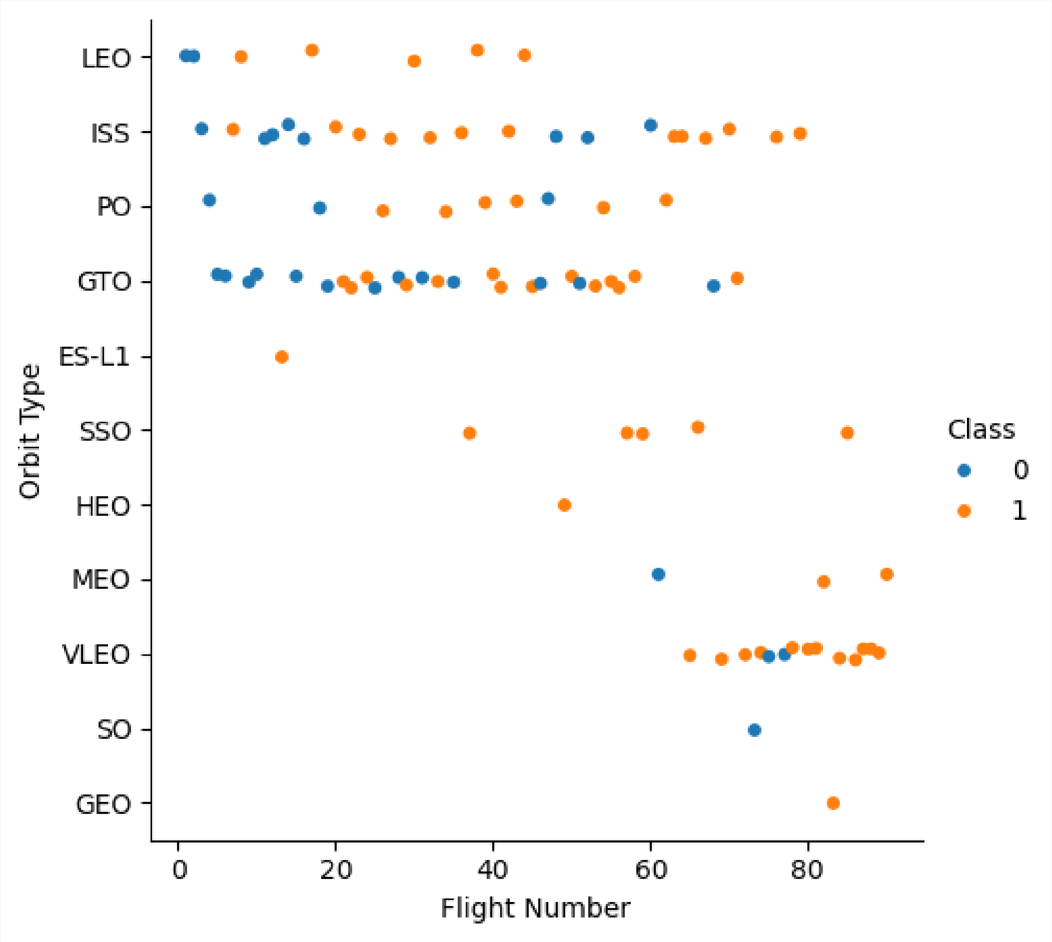
Success Rate vs. Orbit Type

- This bar chart shows us a comparison between Orbit Type and Success rates of flights
- We can see that ES-L1, GEO, HEO & SSO have 100% success rates
- Lowest success rate orbit is SO



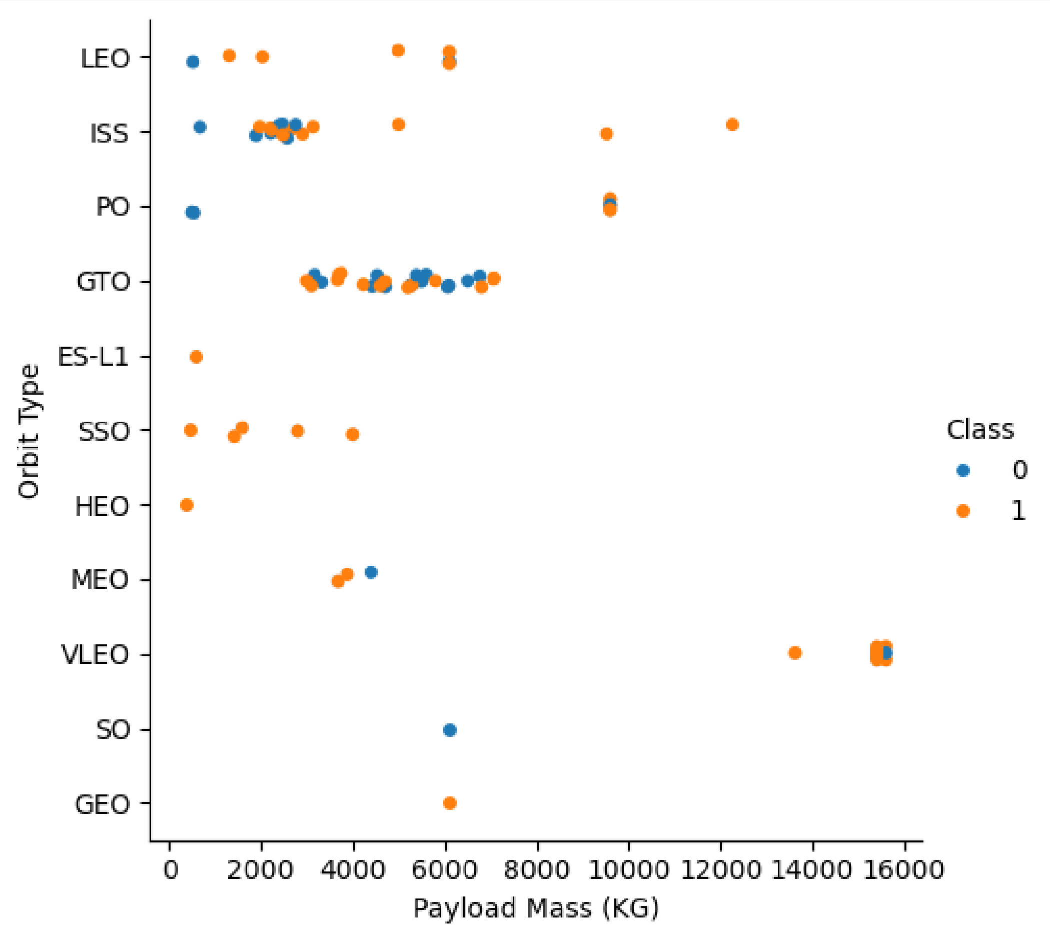
Flight Number vs. Orbit Type

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



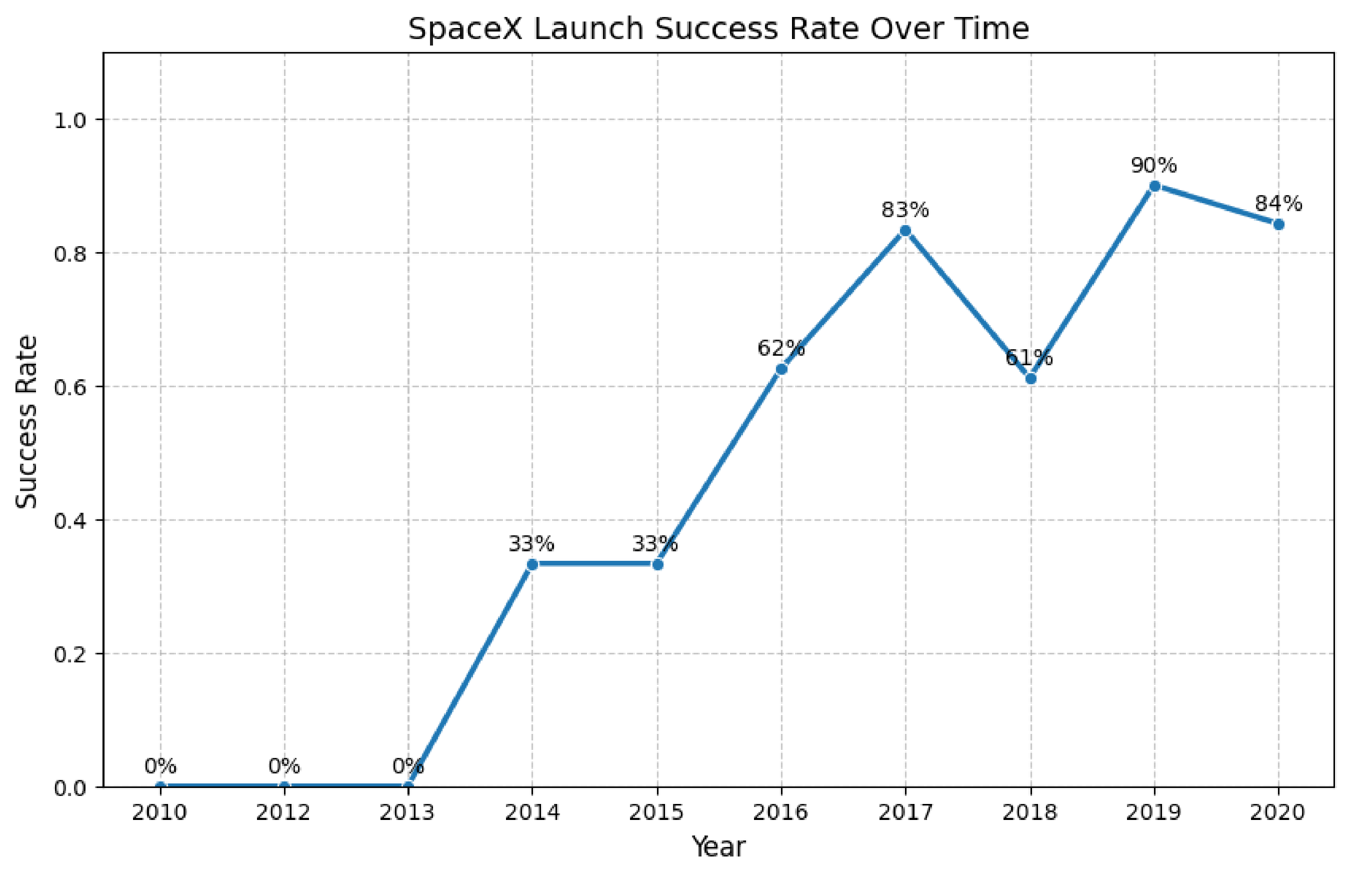
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- For GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch Success Yearly Trend

- We can see that launches are in an upward trend for past 12 years, from 2013
- There have been a dip in 2018 but quickly recovered in 2019.
- Highest Success rate occurred in 2019 at 90% success rate



All Launch Site Names

- We queried the database in SQL to get the name of all Launch Site names
- Launch Sites:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL;
```

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

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

Launch Site Names Begin with 'CCA'

- We queried the database with SQL with the query seen in the screenshot to get the names of Launch site names that began with “CCA”

```
%sql SELECT * FROM SPACEXTBL 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

Total Payload Mass

- We queried the database using SQL to get the total payload mass for the missions launched by NASA (CRS)
- Total Payload Mass: 45,594 KG

```sql
%%sql
SELECT
    SUM("PAYLOAD_MASS__KG_") AS "Total Payload Mass by NASA (CRS)"
FROM
    SPACEXTBL
WHERE
    Customer = "NASA (CRS)";
```

✓ 0.0s

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

Total Payload Mass by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

- The Average Payload Mass by F9 v1.1: 2,928.4 KG

```
%%sql
SELECT AVG("PAYLOAD_MASS__KG_") AS "Average Payload Mass by Booster F9 v1.1 (KG)"
FROM SPACEXTBL
WHERE "Booster_Version" = "F9 v1.1";
```

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

Average Payload Mass by Booster F9 v1.1 (KG)
2928.4

First Successful Ground Landing Date

- The first successful landing that happened on a Ground pas was on 22/12/2015

```
%%sql
SELECT MIN(Date) AS "First Successful Landing"
FROM SPACEXTBL
WHERE "Landing_Outcome" = "Success (ground pad)";
```

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

First Successful Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

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

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

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission
  - 101 missions

```
%sql SELECT COUNT(*) FROM SPACEXTBL;
```

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

COUNT(*)
101

Boosters Carried Maximum Payload

- The booster that carried maximum payload were queried from the database using SQL

```
%%sql
SELECT "Booster_Version"
FROM SPACEXTBL
WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL);
```

```
 * 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

2015 Launch Records

- We queried the 2015 Launch Records
- We first created a Date columns which contained the month and the year
- And then we were able to extract the needed data, i.e. 2015 Launch Records

```
%%sql
SELECT
    CASE substr(Date, 6, 2)
        WHEN '01' THEN 'January'
        WHEN '02' THEN 'February'
        WHEN '03' THEN 'March'
        WHEN '04' THEN 'April'
        WHEN '05' THEN 'May'
        WHEN '06' THEN 'June'
        WHEN '07' THEN 'July'
        WHEN '08' THEN 'August'
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
    END AS month_name,
    "Booster_Version",
    "Launch_Site"
FROM
    SPACEXTBL
WHERE
    substr(Date, 1, 4) = '2015'
```

✓ 0.0s

* sqlite:///my_data1.db
Done.

month_name	Booster_Version	Launch_Site
January	F9 v1.1 B1012	CCAFS LC-40
February	F9 v1.1 B1013	CCAFS LC-40
March	F9 v1.1 B1014	CCAFS LC-40
April	F9 v1.1 B1015	CCAFS LC-40
April	F9 v1.1 B1016	CCAFS LC-40
June	F9 v1.1 B1018	CCAFS LC-40
December	F9 FT B1019	CCAFS LC-40

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

- Here we have the rank by 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
- No attempt at 10 count
- Success (drone ship) 5 counts
- Failure (drone ship) 5 counts

```
%%sql
SELECT
    "Landing_Outcome",
    COUNT(*) AS "Outcome_Count"
FROM
    SPACEXTBL
WHERE
    Date BETWEEN '2010-06-04' AND '2017-03-20'
    AND "Landing_Outcome" IS NOT NULL
GROUP BY
    "Landing_Outcome"
ORDER BY
    "Outcome_Count" DESC;
```

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

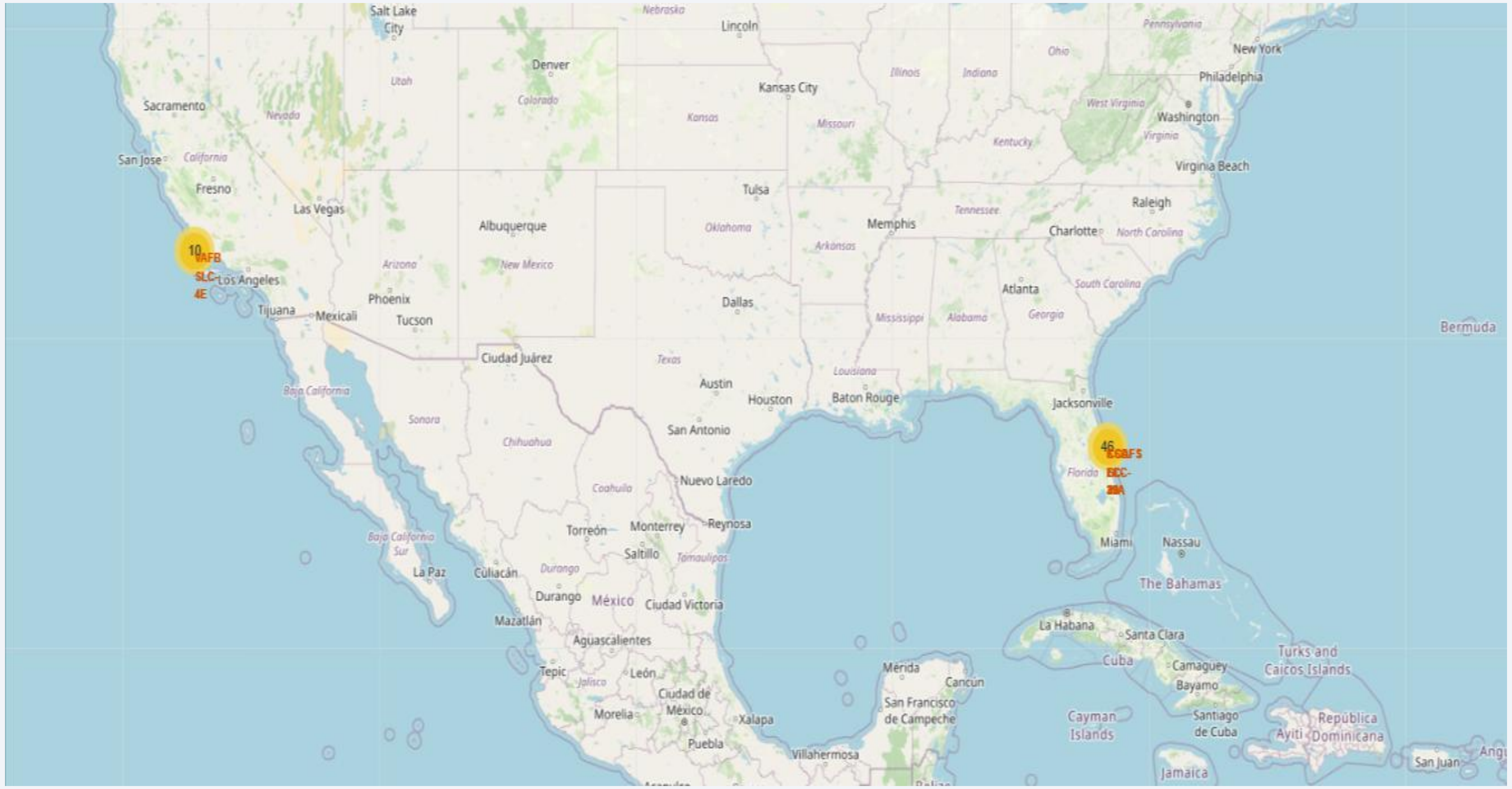
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

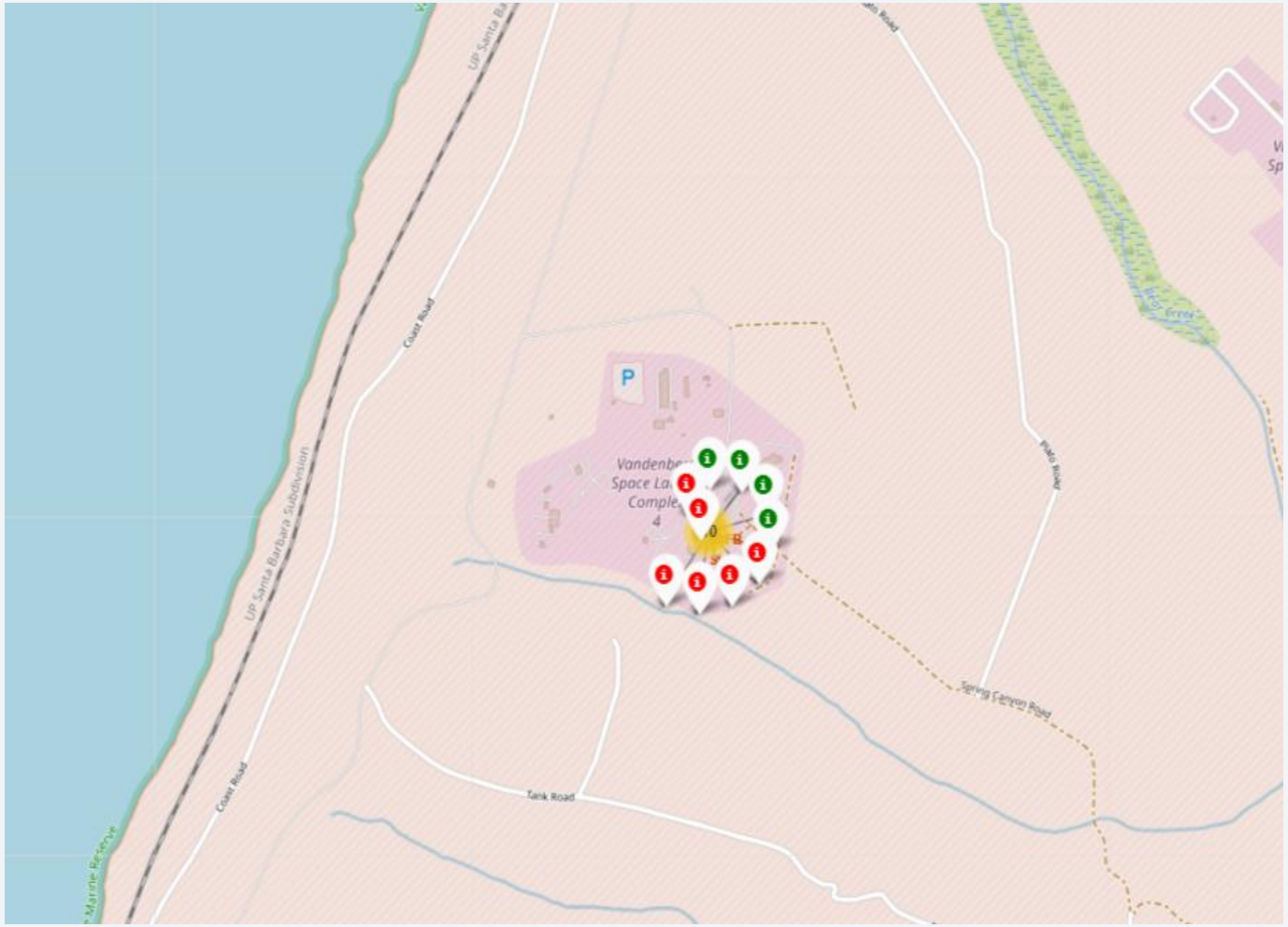
Launch Sites Proximities Analysis

Launch Sites

In this folium map, we can see the launch sites indicated through yellow circle on the map



Launch Markers

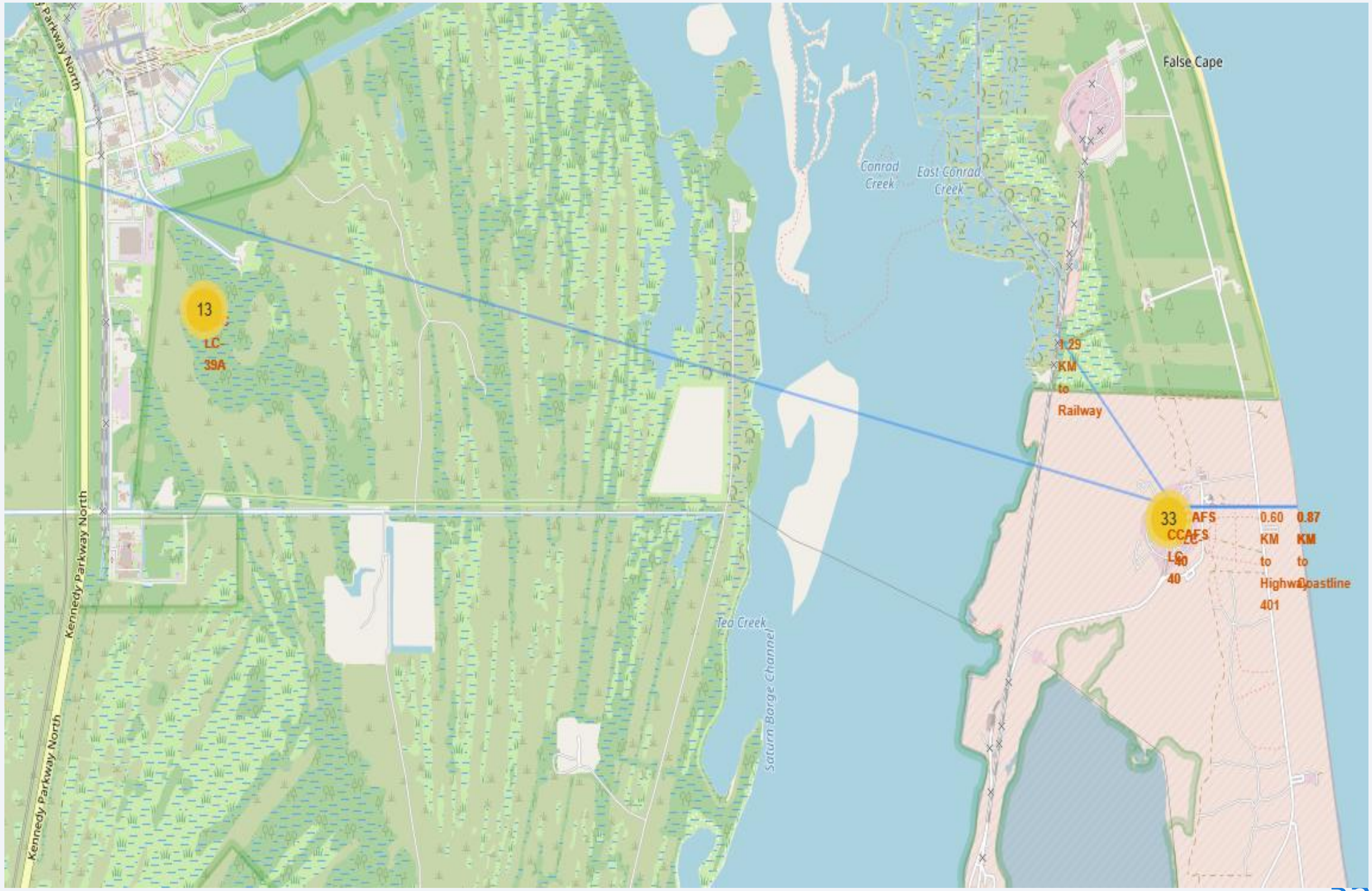


In this Folium Map we have color coded launch markers that were launched from this area

- Green markers are successful launches
- Red markers are failure launches

Proximity

In this Folium Map we have the proximity lines with distance information

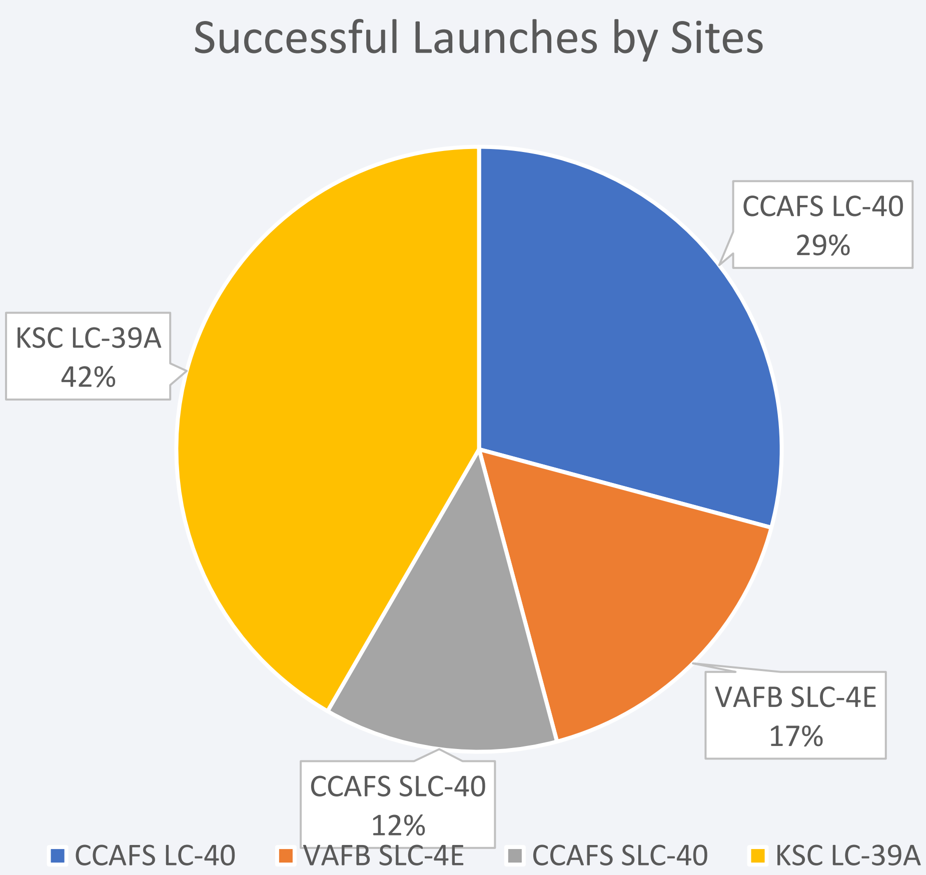


Section 4

Build a Dashboard with Plotly Dash

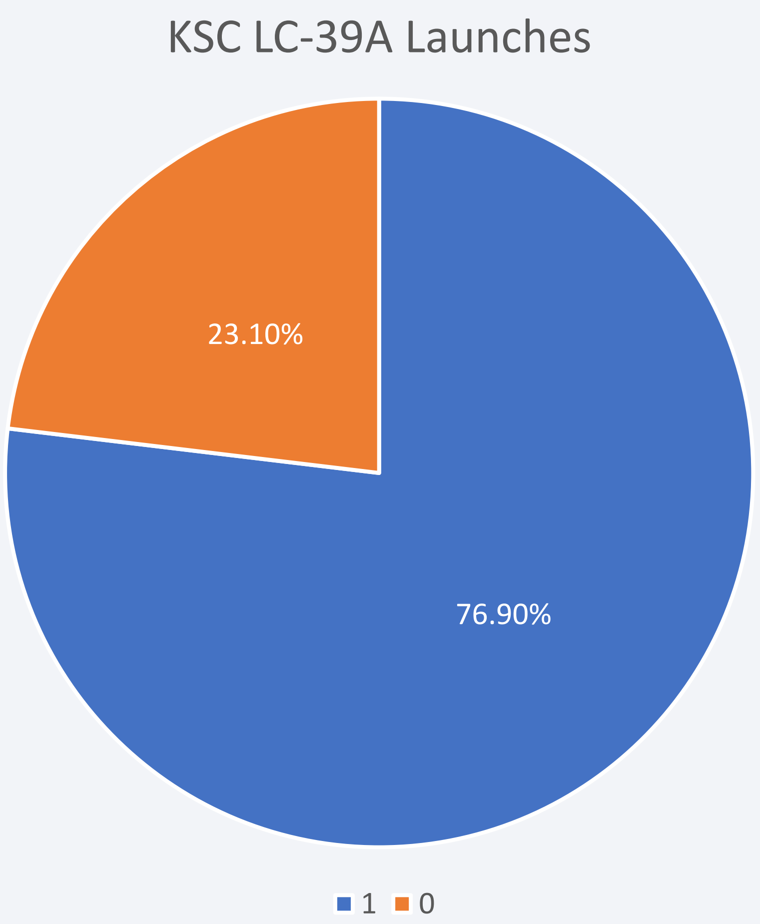
Success Launches – Launch Sites

- This pie chart shows us the distribution of successful launches across the launch sites
- Most of successful launches, launched from KSC LC-39A at 42%



Success Rate – KSC LC-39A

- This Pie chart shows us the distribution of launches between Successful and failed launches at KSC LC-39A
- 76.90% launches were successful
- 23.1% of the launches failed at KCS LC-39A



Success Launches - Payload Mass by Different Boosters

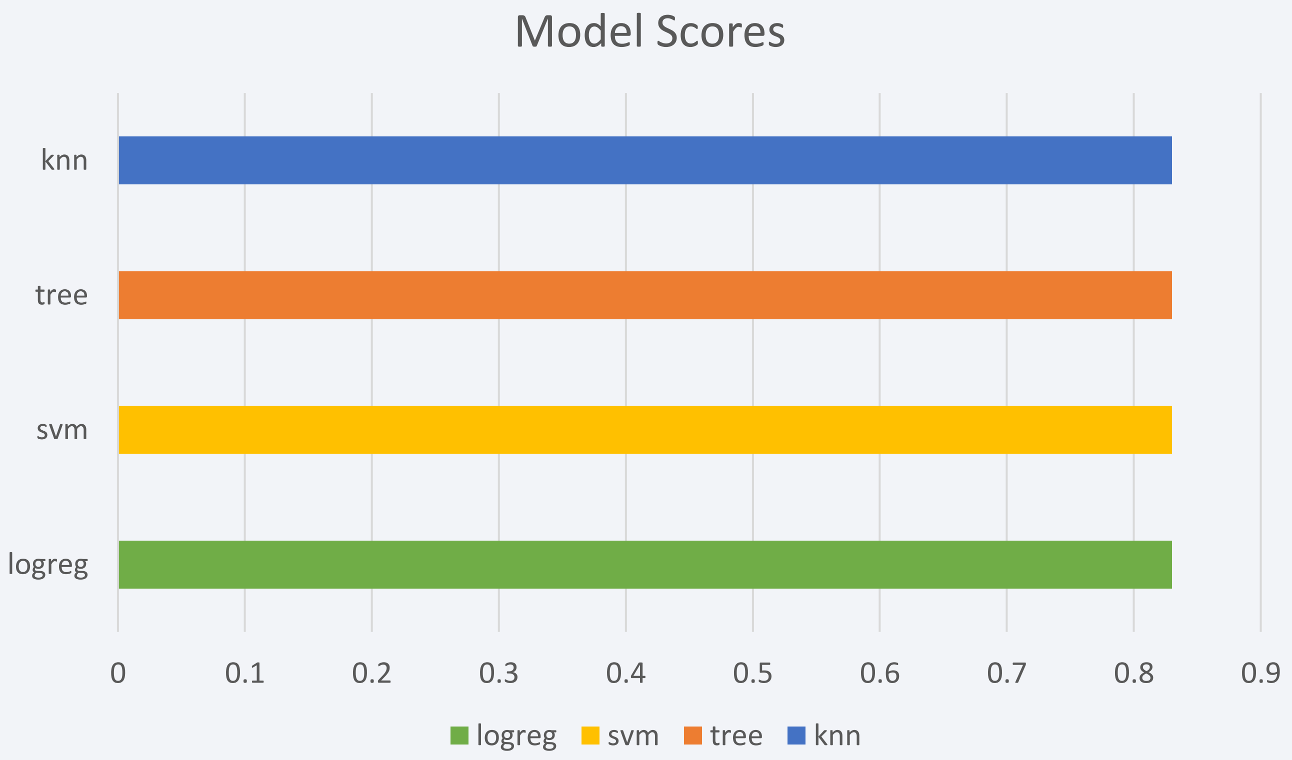
Here we can see that success launches rate by taking into account the relationship between Payload Mass, Successful Launches and Boosters

Section 5

Predictive Analysis (Classification)

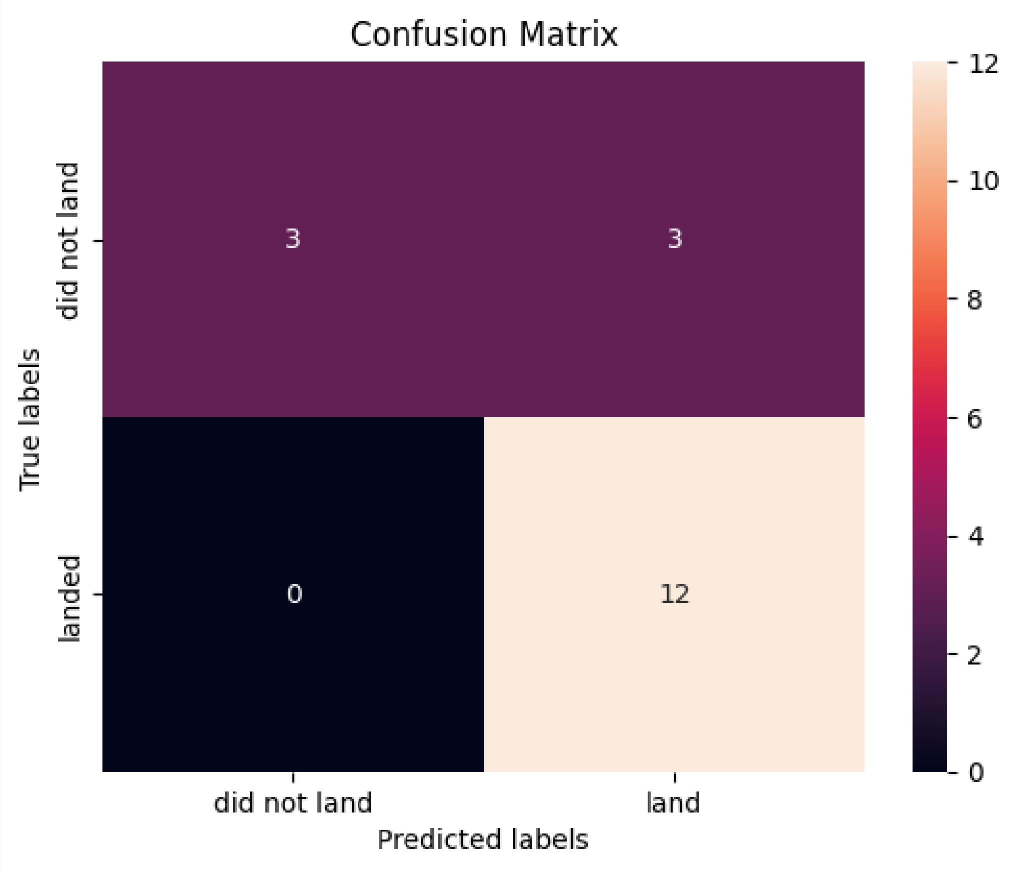
Classification Accuracy

- We built 4 models here
  - Logistic Regression
  - SVM
  - Decision Tree
  - K-Nearest Neighbors
- All models had an accuracy of 83.33%



Confusion Matrix

- **High Recall (100%)**
  - The model **never misses** a successful landing (no false negatives), making it ideal for ensuring recovery operations are always prepared.
- **Moderate Precision (80%)**
  - When predicting a landing, it's correct **80% of the time**, but has **3 false alarms** (preparing for landings that fail).
- **Low Specificity (50%)**
  - Only **half of failed landings** are correctly identified, suggesting room for improvement in reducing wasted resources.
- Great for **mission assurance** (never missing landings), but could optimize cost-efficiency by reducing false positives.



Conclusions

Key Findings

- **Success Rate Improvement:** SpaceX landing success has dramatically improved over time, reaching [XX]% in recent missions
- **Payload Sweet Spot:** Missions with payload mass between 2,000-6,000 kg show optimal success rates
- **Site Performance:** KSC LC-39A consistently outperforms other launch sites
- **Predictive Accuracy:** Machine learning models achieved 84% accuracy in predicting landing outcomes
- **Cost Implications:** Successful predictions can save millions in mission planning and insurance costs

Business Impact

- Enable more accurate cost estimation for future missions
- Support strategic decision-making for launch site selection
- Provide competitive intelligence for space industry analysis
- Facilitate risk assessment and insurance pricing

Future Recommendations

- Incorporate weather data for enhanced predictions
- Analyze booster reuse patterns and their impact on success
- Expand analysis to include Falcon Heavy and Starship data

Appendix

- Github Repo: https://github.com/AdnanRahmanpoor/IBMDSCourse/blob/main/Final%20Capstone%20Project/

Thank you!