Winning Space Race with Data Science

P.Bhat
01/29/2026

Outline

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

Executive Summary

Methodologies Summary

- Data collection with API
- Data collection using Web Scraping
- Data wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Visualization
- Interactive Visual Analytics with Folium
- Interactive Dashboard with Plotly Dash
- Machine Learning Prediction

Results Summary

- Improved success Rate over Time
- KSC LC-39A had highest success Rate by Launch Site
- Highest success Rate by Orbit for SSO, ES-L1, HEO, GEO
- Lightening payloads increases their success rate
- Algorithms can be classified ass highly accurate and not-highly accurate

Introduction

Project background and context

- SpaceX launches Falcon 9 launches at $62 million dollars per mission, as compared to their competitors (around $165 million).
- They have this edge due to huge cost reduction from first stage reuse for their rockets.
- Competitors can also gain this edge, by calculating whether their rocket's first stage will land successfully, which will greatly reduce overall rocket launching costs.

Problems you want to find answers

- Variables influencing a rocket's successful landing?
- How is probability of a safe landing impacted by the interconnectedness of different features?
- Operational conditions needed to be met for consistent successful landings.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was gathered through the SpaceX API, along with additional information obtained by scraping Wikipedia pages.
- Perform data wrangling
  - Categorical variables were transformed using one-hot encoding, relevant entries were selected, and nested fields were flattened while handling missing 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
  - Developed, tune, assessed, and compared multiple classification models for prediction tasks.

Data Collection

- Data collection and transformation methods:

  1. SpaceX web API

     i. REST/SOAP-based API calls were used to fetch .json responses with data.

     ii. ii. Data was cleaned/cleansed/enriched (where possible) and converted to pandas data frame.

  2. Wikipedia:

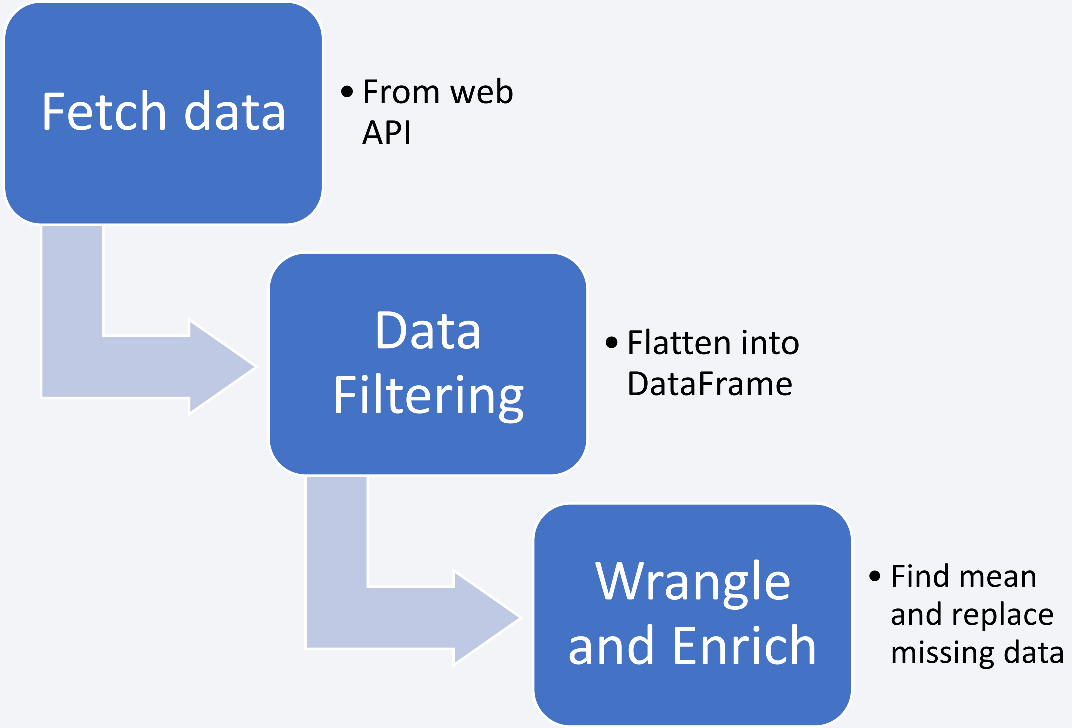
     i. Web scraping data collected from Falcon 9's Wikipedia pages (such as launch records).

     ii. ii. Extracted data was converted to pandas data frame.

Data Collection – SpaceX API

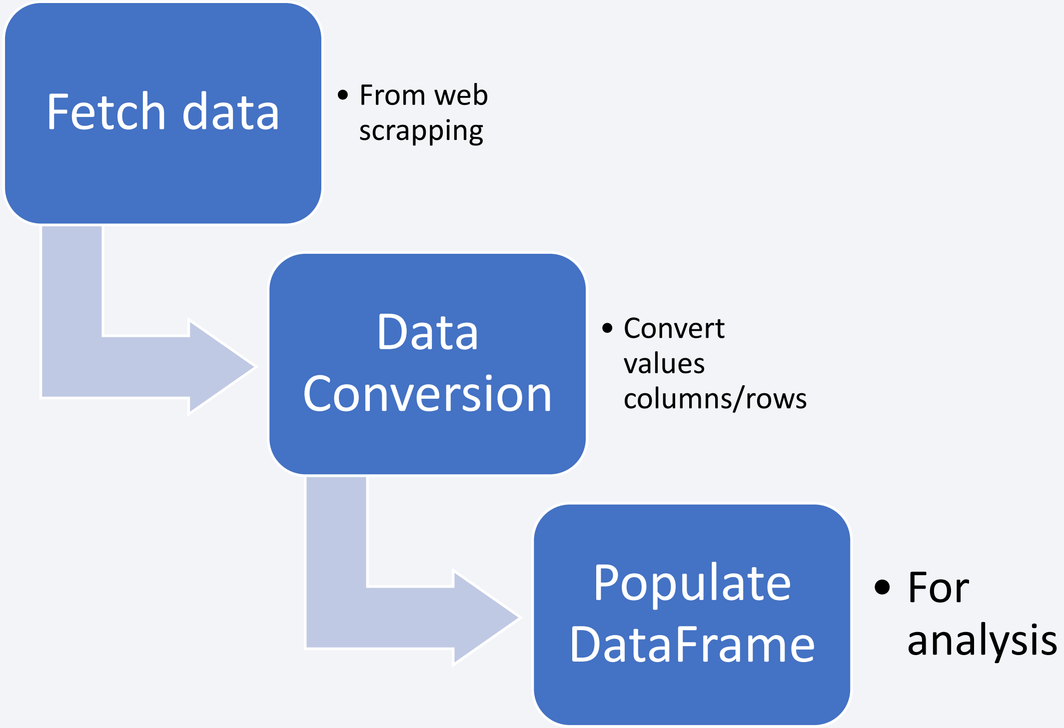
- Data was fetched from SpaceX API
- The initial data was cleaned
- Then it was wrangled and re-formatted

- GitHub URL - [Link]



Data Collection - Scraping

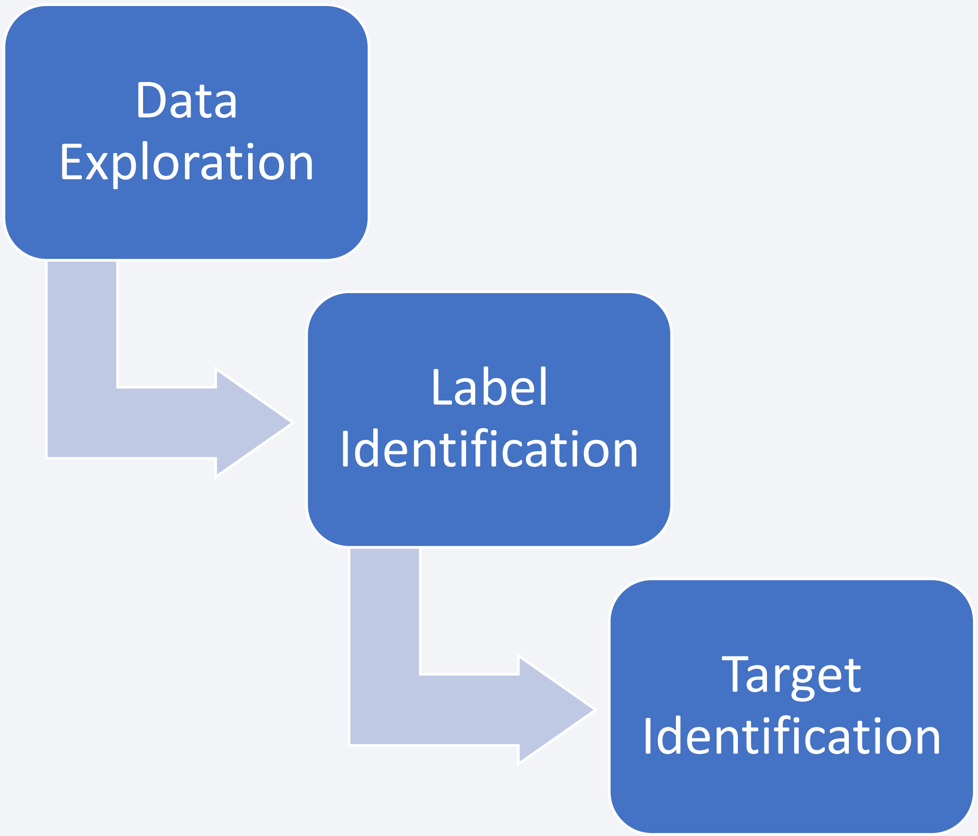
- HTTP 403 error (web scrapping not allowed on Wikipedia)
- Planned to scrape data from web (e.g., Wikipedia)
- Planned to extract using BeautifulSoup and parse into Pandas DataFrame
- GitHub URL - [Link]



Data Wrangling

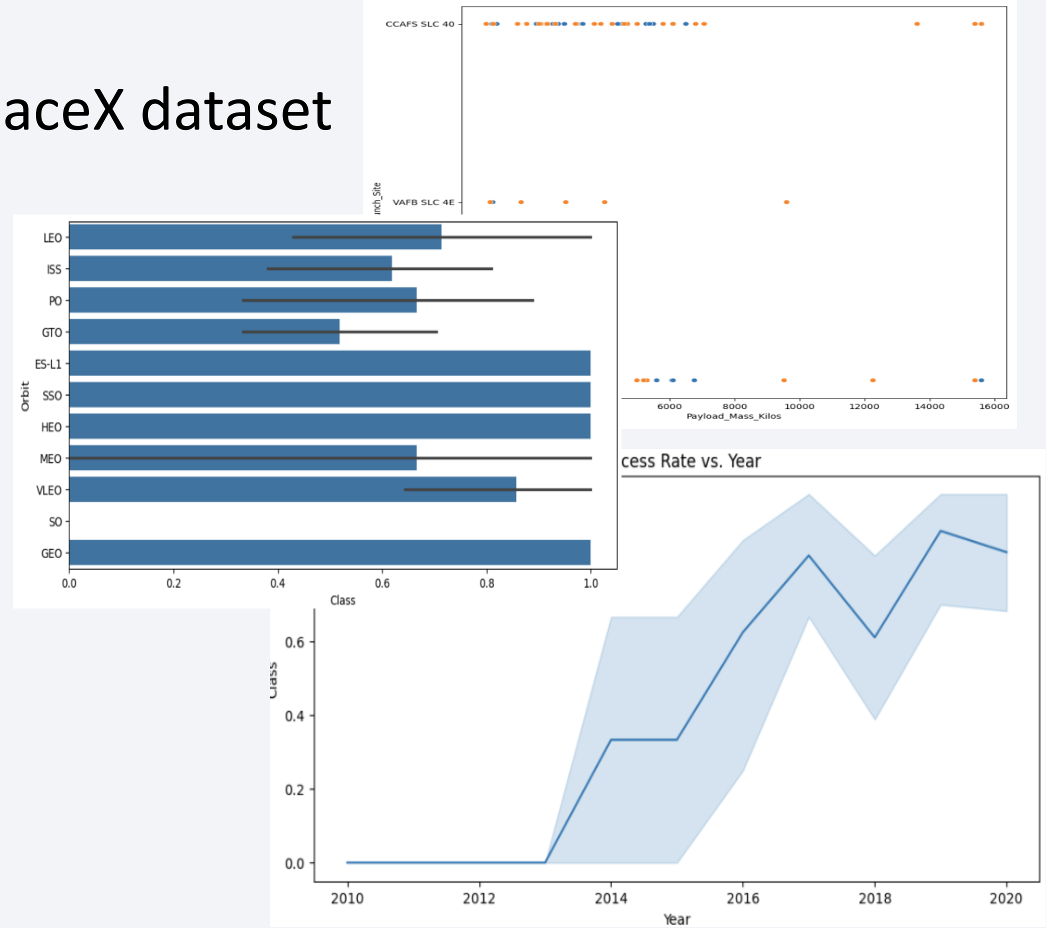
- Identified the labels needed from Exploratory Data Analysis (EDA)
- Perform calculations of launches and orbital occurrences and write outcome data into csv file.

- GitHub URL - Link



EDA with Data Visualization

- Conducted EDA with Data Visualization for SpaceX dataset
- Visualized the relationships between:
  - Flight number and payload to launch Site
  - success rate and flight number to each orbit type
  - Yearly launch success trend.
- Feature engineering
- Conversion to float64
- GitHub URL - [Link]



EDA with SQL

- Conducted EDA with SQL to SpaceX dataset which we injected into a PostgreSQL from Jupyter notebook.
- Wrote multiple SQL queries
- EDA analysis resulted in fetching –
  - names of launch site that are unique
  - payload mass by booster rockets
  - average payload mass vs F9
  - success vs failure counts, etc.
- GitHub URL - [Link]

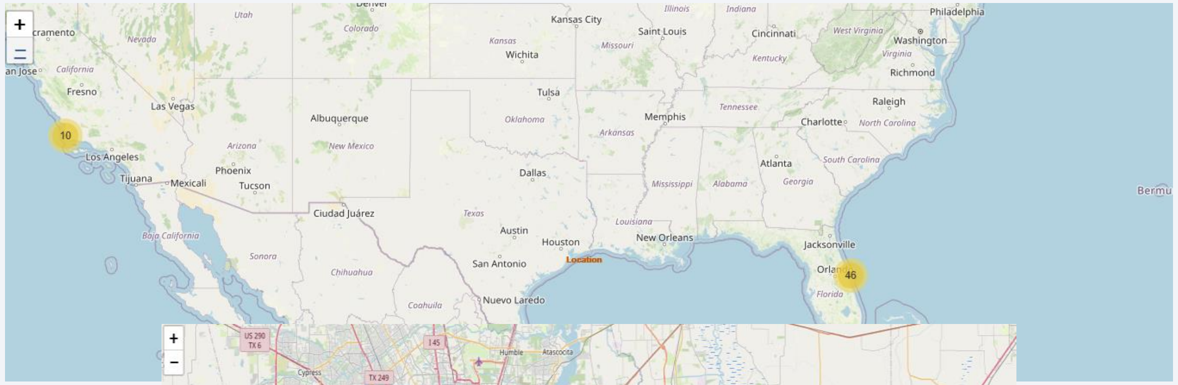
```sql
SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

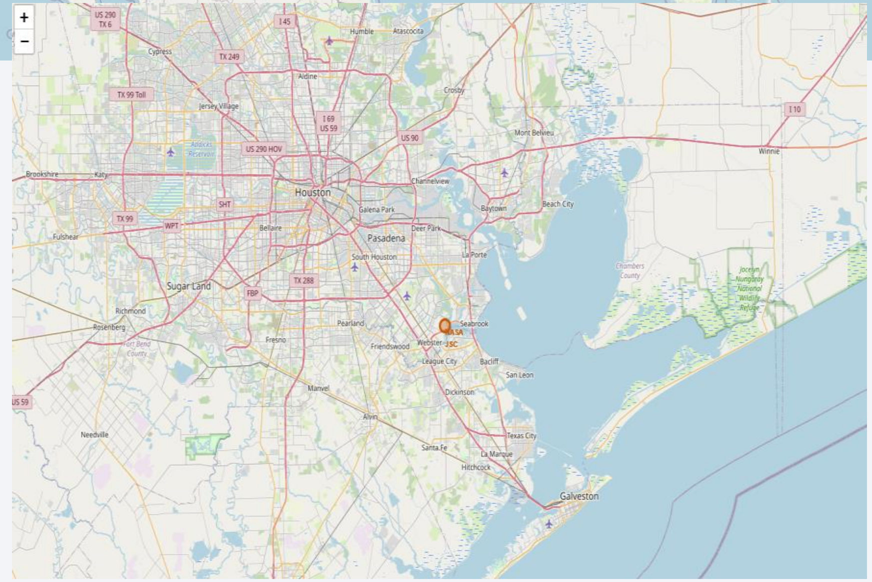
SUCCESSFUL_LAUNCH_DATE
2015-12-22

```sql
SELECT Landing_Outcome, COUNT(*) as TOTAL_COUNT FROM SPACEXTABLE WHERE Date BETWEEN "2010-06-04" AND "2017-03-20" GROUP BY Landing_Outcome ORDER BY TOTAL_COUNT DESC;
```

Build an Interactive Map with Folium

- All launch sites were displayed on a Folium map using markers, circles, and indicators for launch success or failure.
- Distances between each launch site and its surrounding proximities were computed.
- GitHub URL - [Link]()



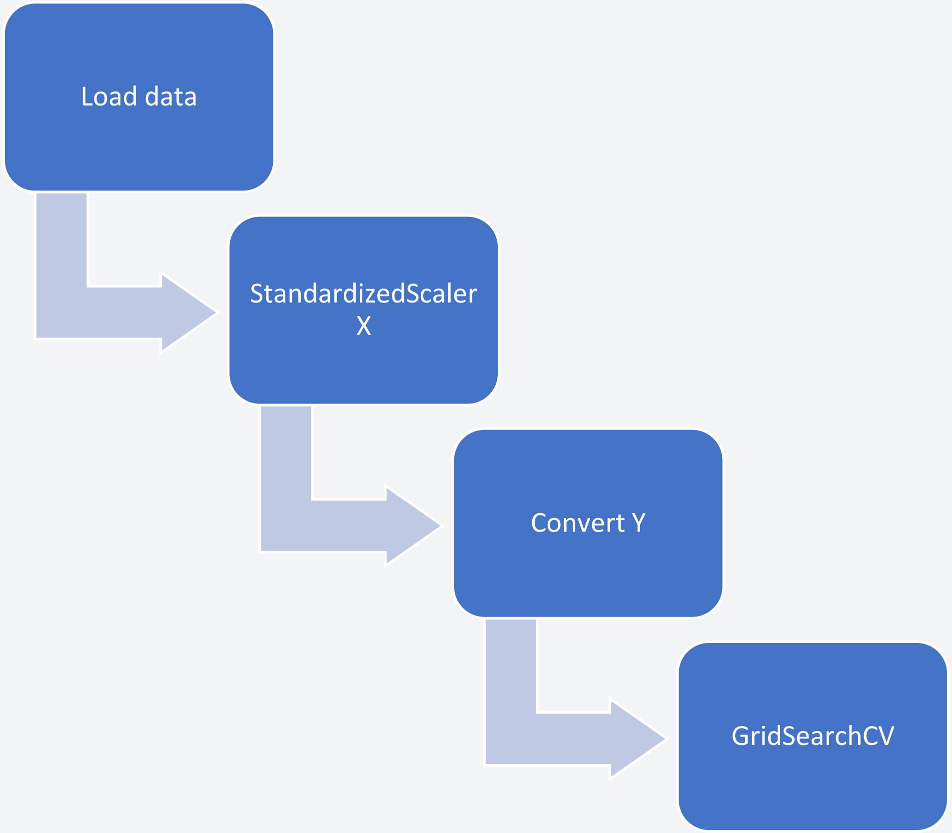


Build a Dashboard with Plotly Dash

- Designed an interactive dashboard with Plotly Dash.
- Visualized total launches per site using pie charts.
- Explored the relationship between launch outcome and payload mass (Kg) across booster versions with a scatter plot.
- GitHub URL - Link

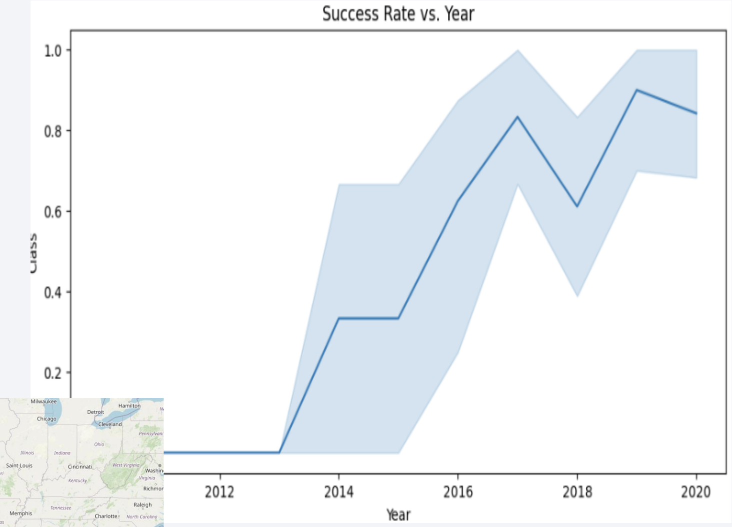
Predictive Analysis (Classification)

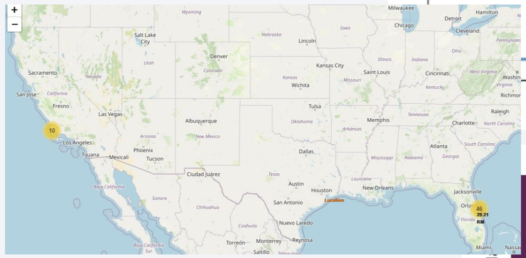
- Preprocessed data with NumPy and Pandas and split into train/test sets.
- Using GridSearchCV, built and tuned machine learning algorithms (Logistic Regression, SVC, Decision Tree Classifier, K Neighbors Classifier)
- Improved performance with feature engineering and algorithm optimization.
- Selected the best-performing classification model.
- GitHub URL - [Link]

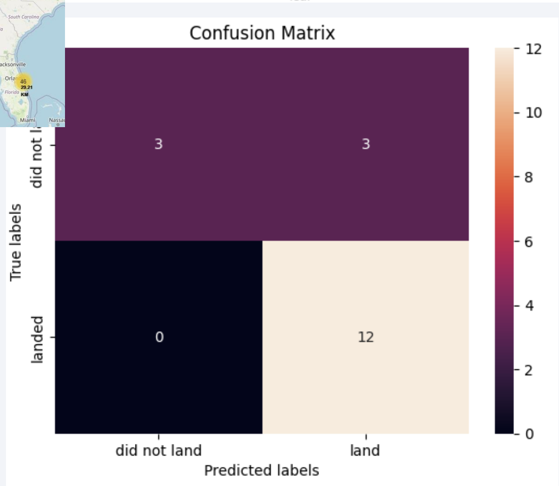


Results

- Exploratory data analysis results
  - Orbits including ES-L1, Sun-Synchronous Orbit (SSO), Highly Elliptical Orbit (HEO), and Geostationary Orbit (GEO) often achieve impressive success levels.
  - Success rate increased steadily from 2013 to 2020.
- Interactive analytics demo in screenshots
  - CCAFS SLC 40 has conducted the most launches, whereas VAFB SLC 4E demonstrates a greater rate of successful landings.
  - Launch sites are close to the coast and away from cities, highways and railways.
- Predictive analysis results
  - Best Performing Algorithm is Decision Tree.





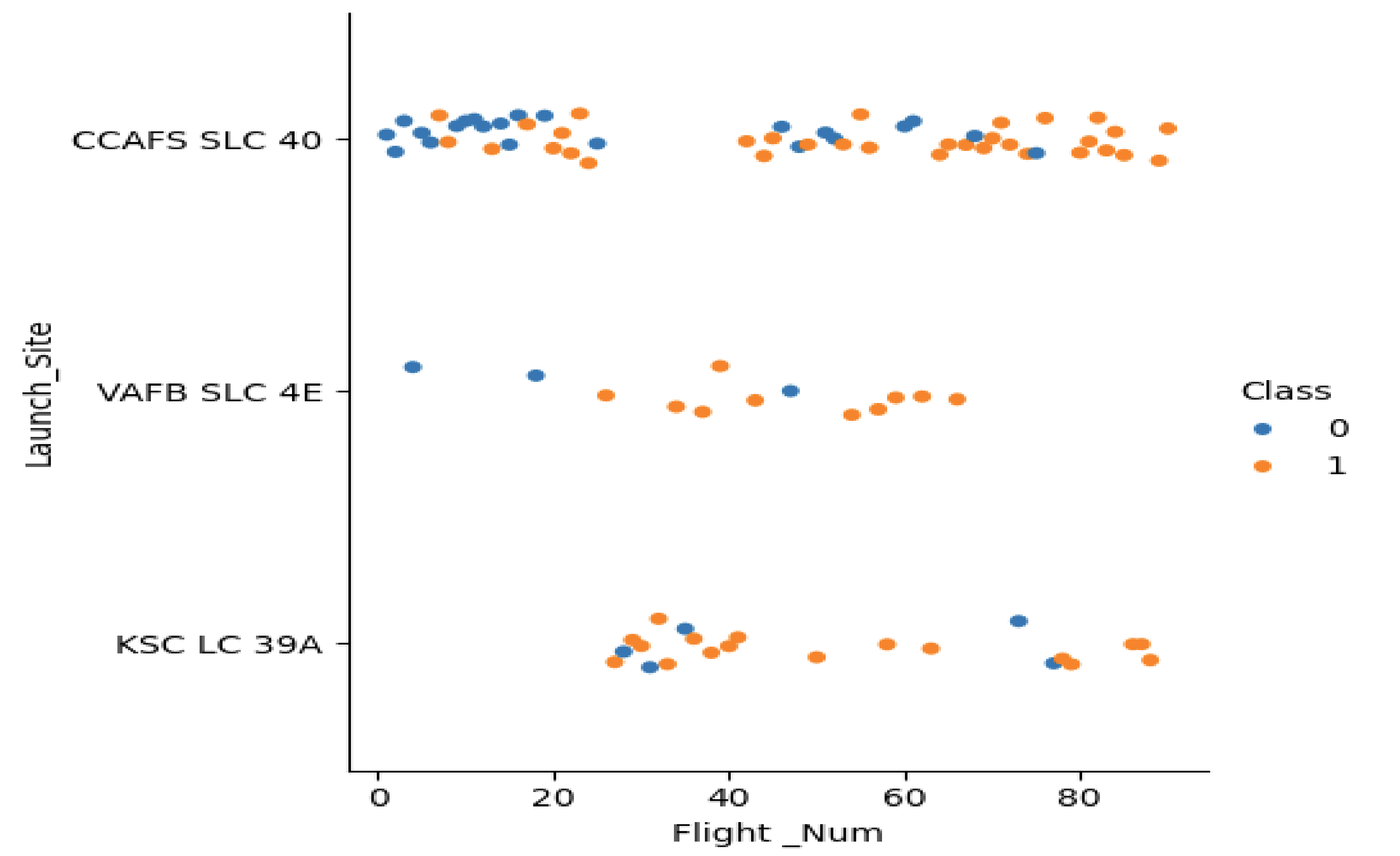


Section 2

Insights drawn from EDA

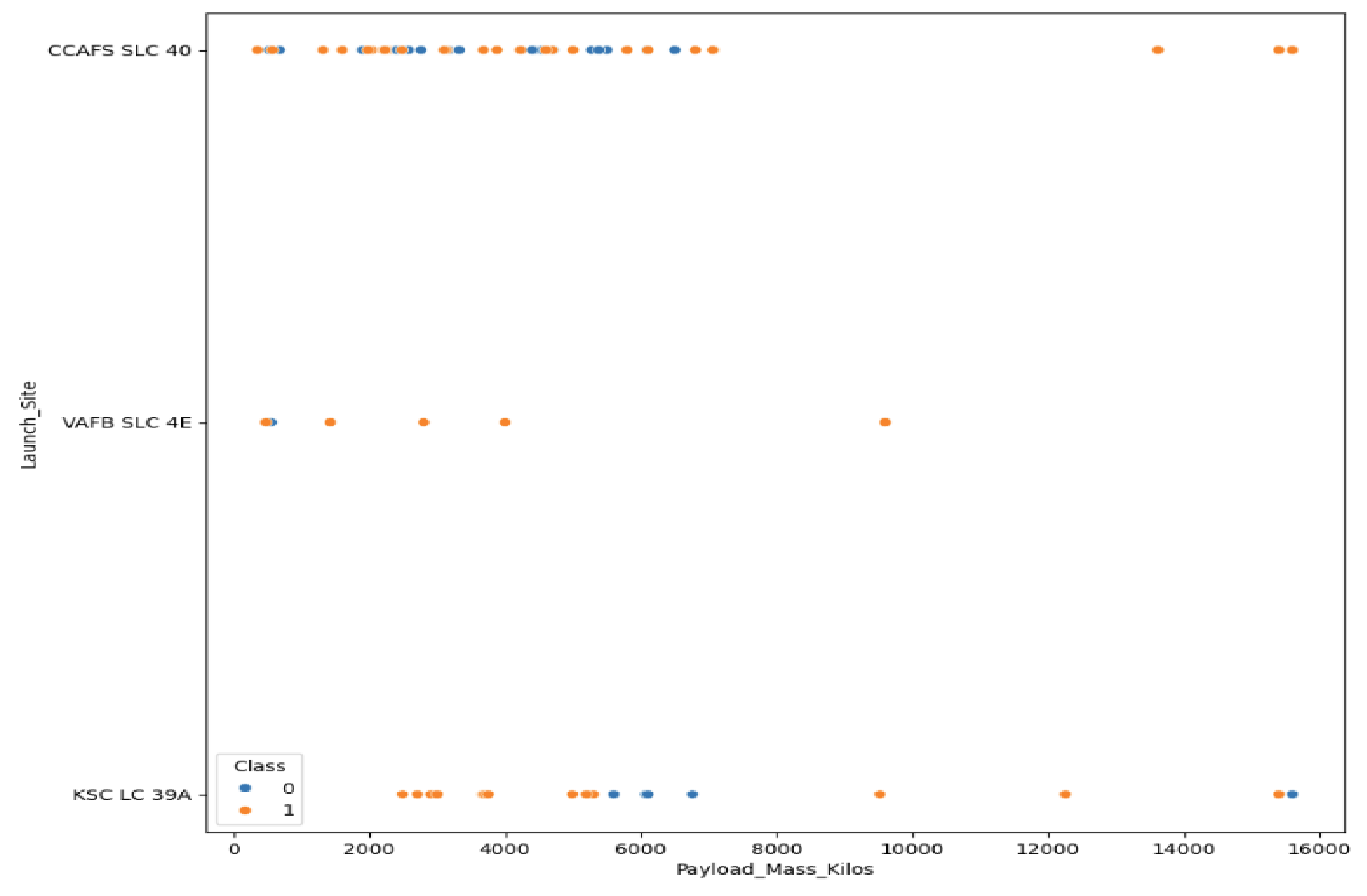
Flight Number vs. Launch Site

- All launch sites experienced both successful and failed landings, with success rates increasing over time.
- Early missions had higher failure rates, reflecting ongoing technological and process improvements.
- CCAFS SLC 40 has the highest number of launches, while VAFB SLC 4E shows a higher proportion of successful landings.



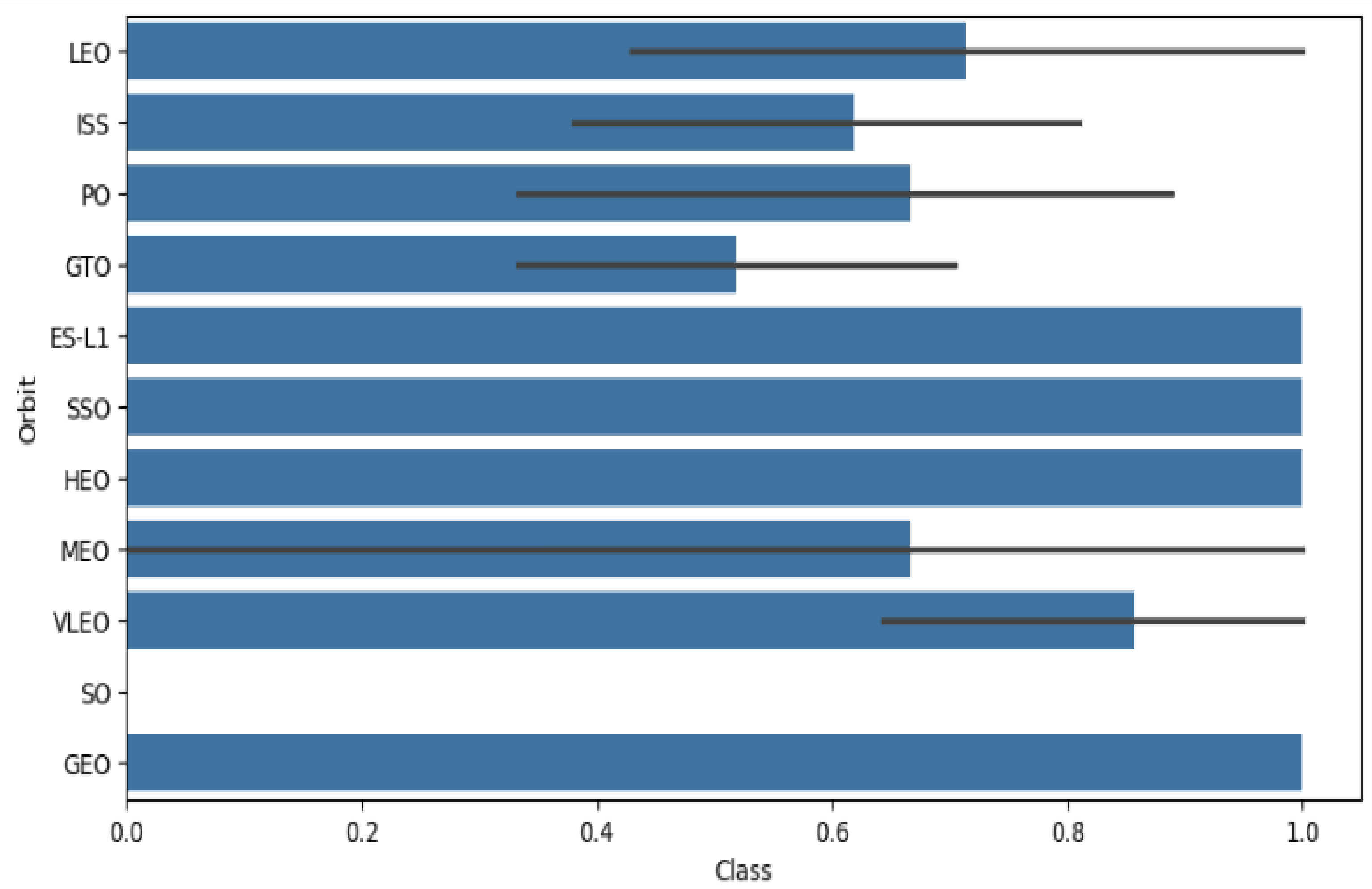
Payload vs. Launch Site

- All launch sites handle a broad range of payload masses, from lighter to heavier payloads.
- Earlier missions generally carried lighter payloads and account for most landing failures.
- This pattern suggests that technological or operational advancements improved success rates for heavier payloads.



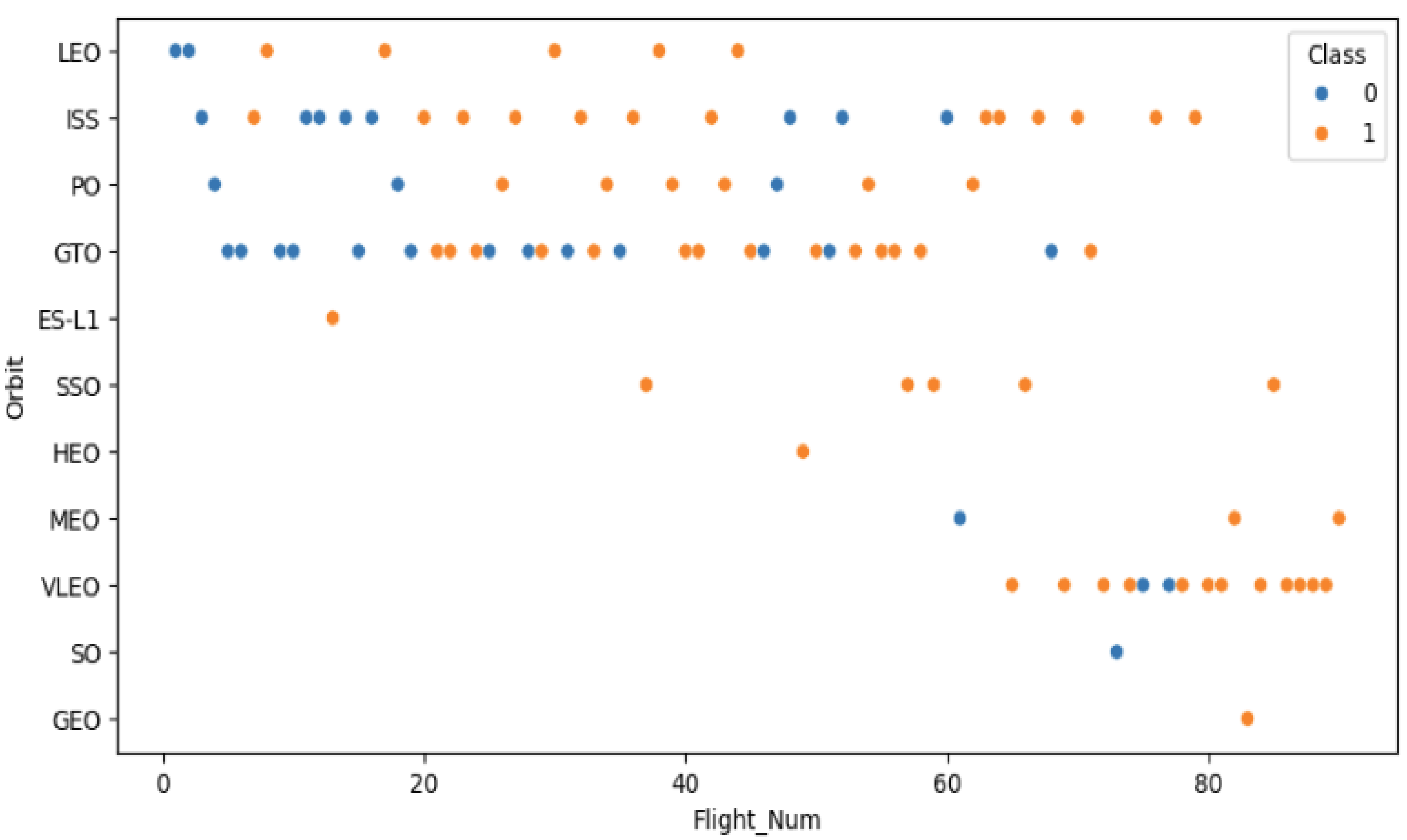
Success Rate vs. Orbit Type

- Orbits such as ES-L1, SSO, HEO, and GEO consistently demonstrate high success rates.
- In contrast, GTO missions show more variable outcomes, indicating potential operational or technological challenges.
- The SO orbit has insufficient data for reliable analysis due to having only one launch.



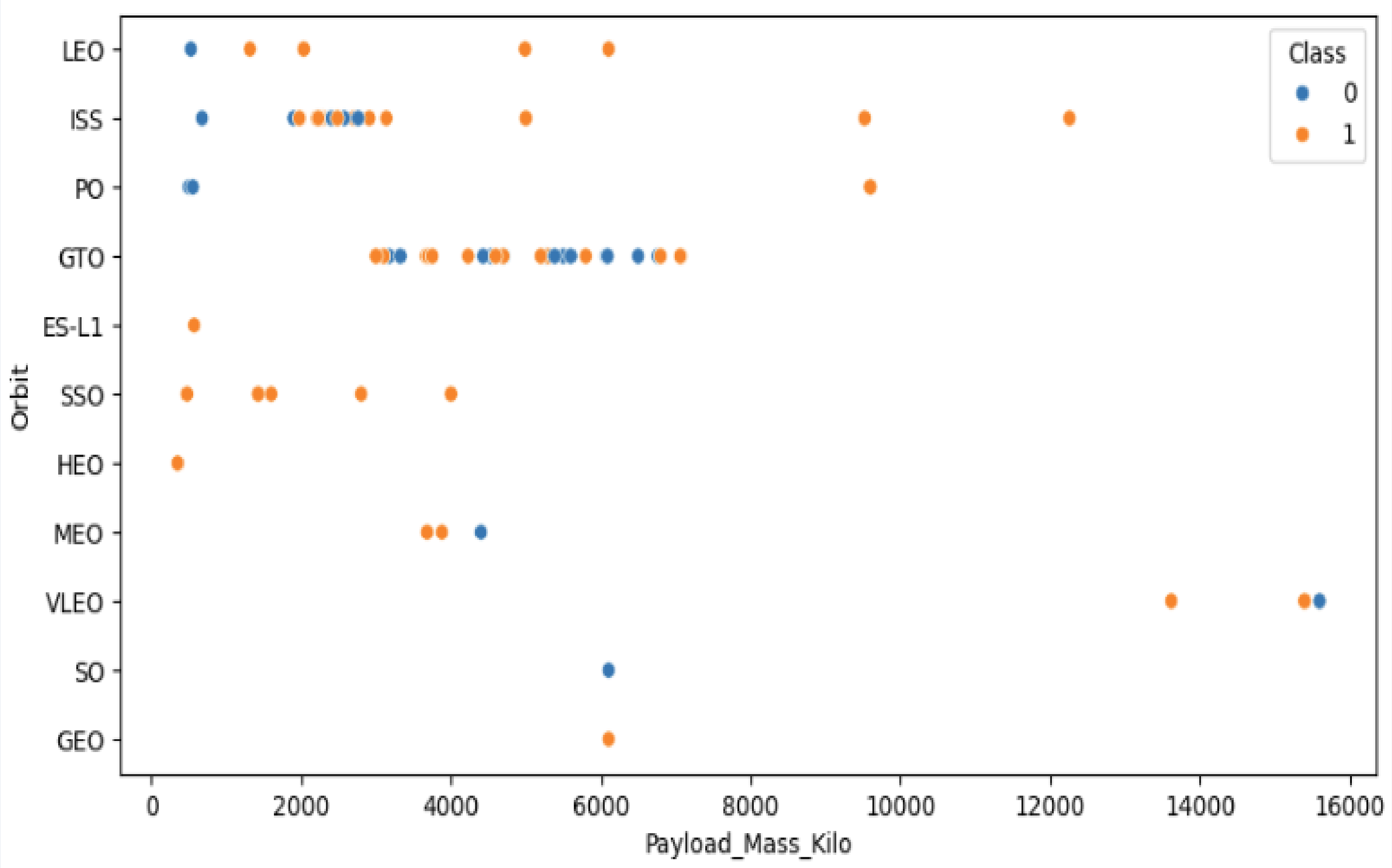
Flight Number vs. Orbit Type

- Multiple orbit types appear across the range of flight numbers, while some are only attempted in later missions.
- Landing success rates improve as flight numbers increase, reflecting accumulated experience and continuous advancements.



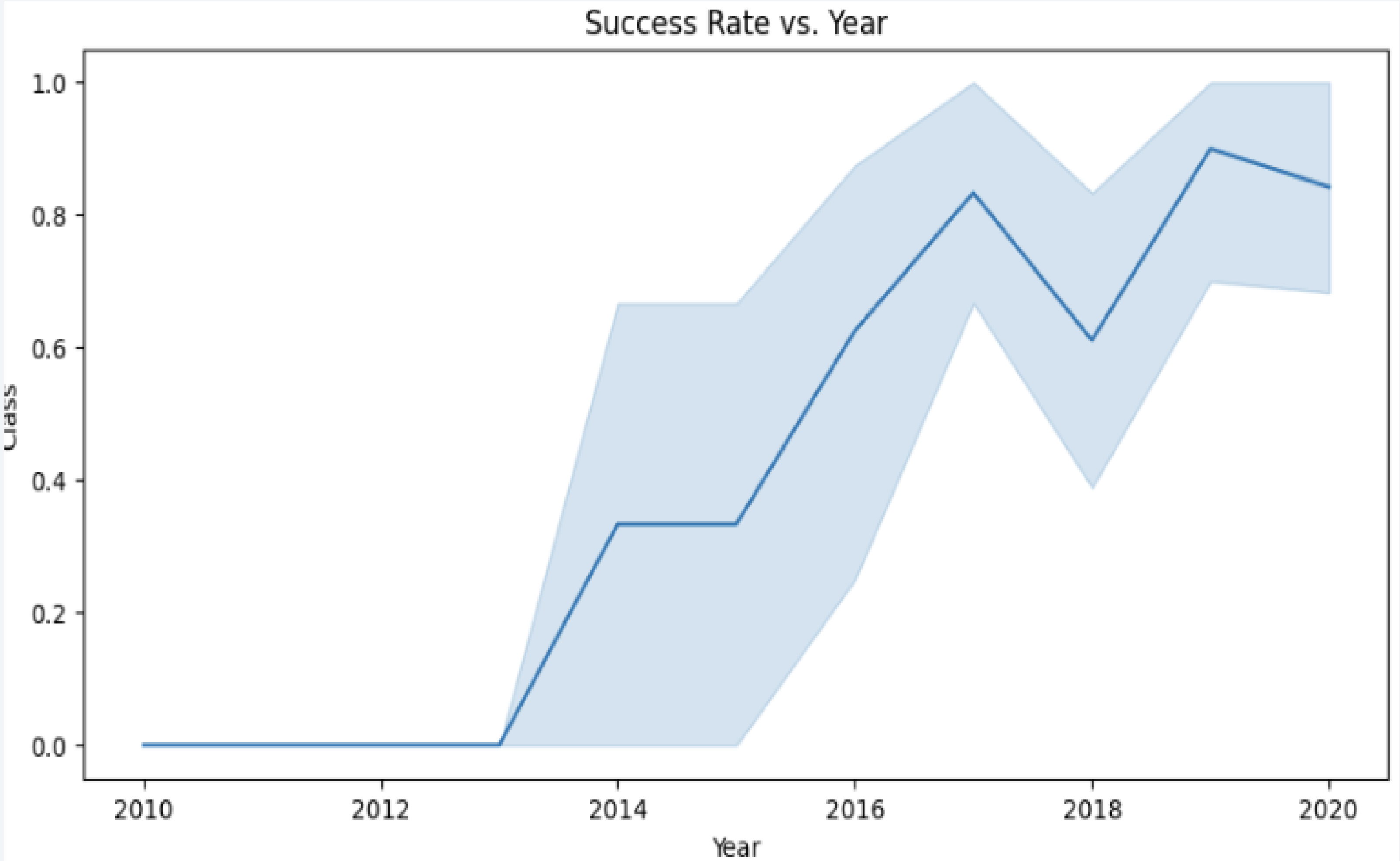
Payload vs. Orbit Type

- Several orbit types span a wide range of payload masses, while SSO, MEO, HEO, and GEO are generally limited to lower payload ranges.
- Orbits with more constrained payload ranges tend to exhibit higher landing success rates.
- Although payload mass alone does not directly determine mission success, its interaction with orbit type shows a meaningful correlation.



Launch Success Yearly Trend

- The yearly trend shows a steady transition from early difficulties to high reliability in first-stage landings.
- Starting in 2016, SpaceX achieved consistent year-over-year improvements in success rates, with a minor decline in 2018.



All Launch Site Names

- There are four unique launch sites

```
SELECT DISTINCT Launch_Site from
SPACEXTABLE;
```

Out[16]:

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

Launch Site Names Begin with 'CCA'

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

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

Total Payload Mass

```
SELECT SUM(PAYLOAD_MASS__KG_) AS
PAYLOAD_MASS_SUM FROM SPACEXTABLE WHERE
Customer='NASA (CRS)';
```

PAYLOAD_MASS_SUM
45596

Average Payload Mass by F9 v1.1

```
SELECT AVG(PAYLOAD_MASS__KG_) AS
PAYLOAD_MASS_AVERAGE FROM SPACEXTABLE WHERE
Booster_Version LIKE 'F9 v1.1%';
```

PAYLOAD_MASS_AVERAGE
2534.6666666666665

First Successful Ground Landing Date

```
SELECT MIN(Date) as SUCCESSFUL_LAUNCH_DATE
FROM SPACEXTABLE WHERE Landing_Outcome =
'Success (ground pad)';
```

SUCCESSFUL_LAUNCH_DATE
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

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

Total Number of Successful and Failure Mission Outcomes

```
SELECT CASE    WHEN Mission_Outcome LIKE
'Success%' THEN 'Success'   WHEN
Mission_Outcome LIKE 'Failure%' THEN
'Failure' END as Mission_Result,COUNT(*)
FROM SPACEXTABLE GROUP BY Mission_Result
```

Mission_Result	COUNT(*)
Failure	1
Success	100

Boosters Carried Maximum Payload

```
SELECT DISTINCT Booster_Version,
PAYLOAD_MASS__KG_ FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_=(SELECT
MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
ORDER BY Booster_Version;
```

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

```
SELECT      CASE strftime("%m",Date)
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,      Landing_Outcome,
Booster_Version, Launch_Site, Date FROM
SPACEXTABLE WHERE
strftime("%Y",Date)="2015" AND
Landing_Outcome="Failure (drone ship)";
```

MONTH_NAME	Landing_Outcome	Booster_Version	Launch_Site	Date
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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

```
SELECT Landing_Outcome, COUNT(*) as
TOTAL_COUNT  FROM SPACEXTABLE
WHERE Date BETWEEN "2010-06-04" AND "2017-
03-20"   GROUP BY Landing_Outcome
ORDER BY TOTAL_COUNT DESC;
```

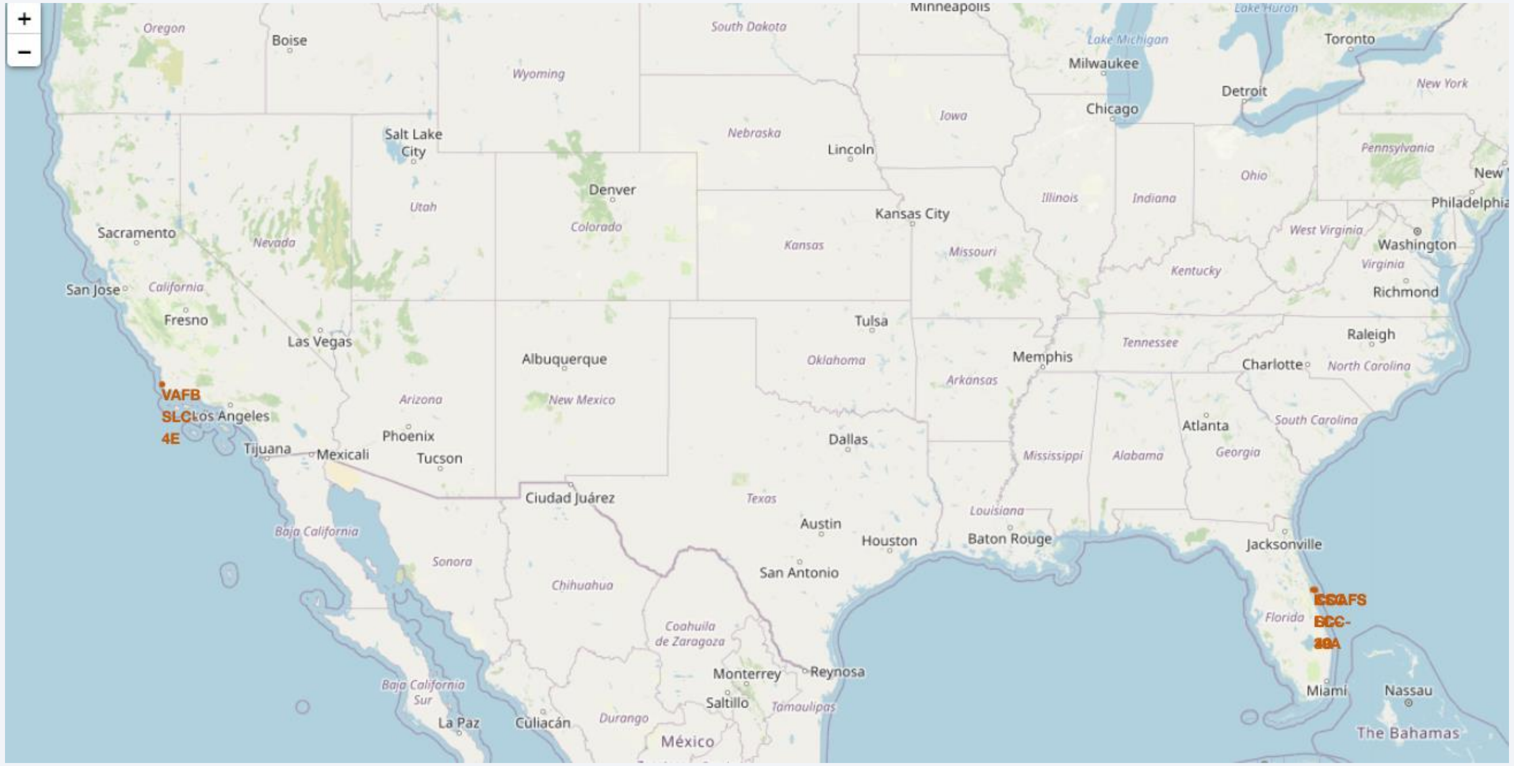
Landing_Outcome	TOTAL_COUNT
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

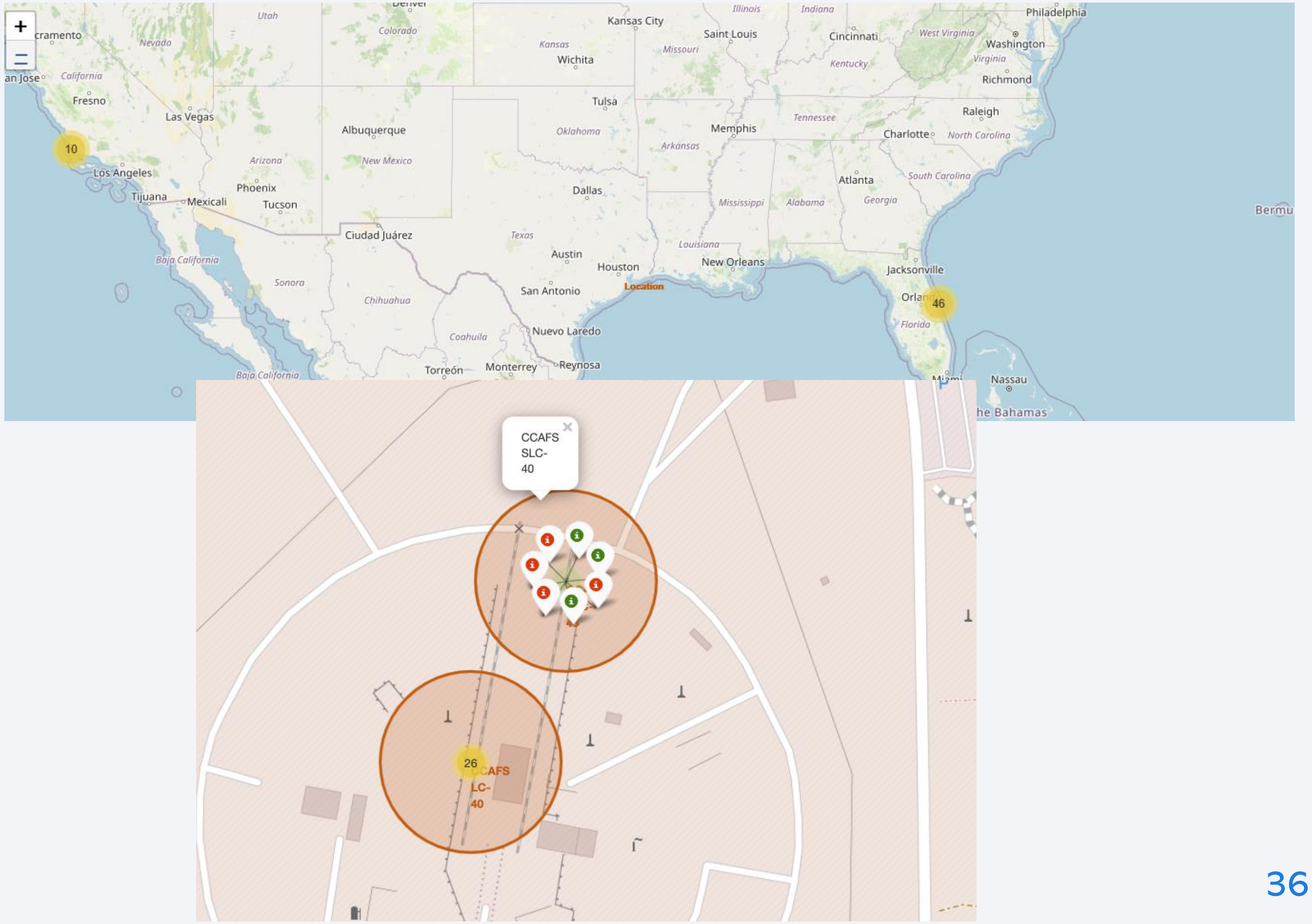
Launch Site Locations

- Launch sites are located near coastal areas in general.
- This is possibly due to p public safety reasons.

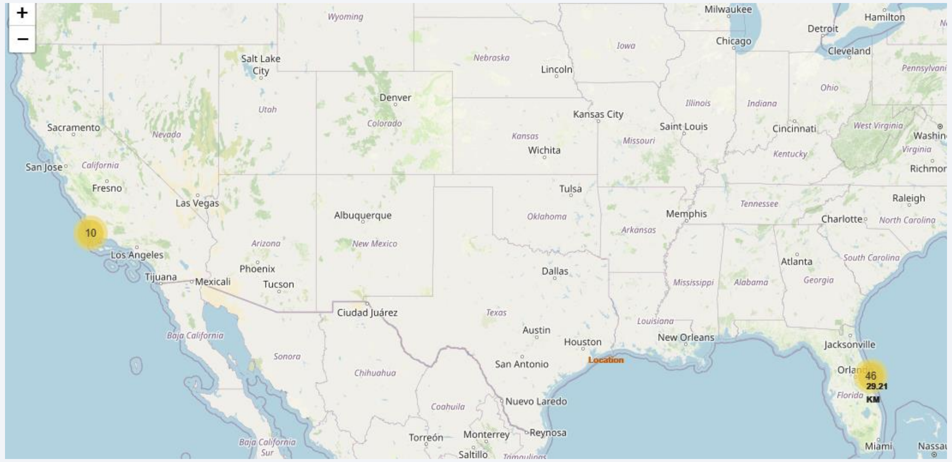


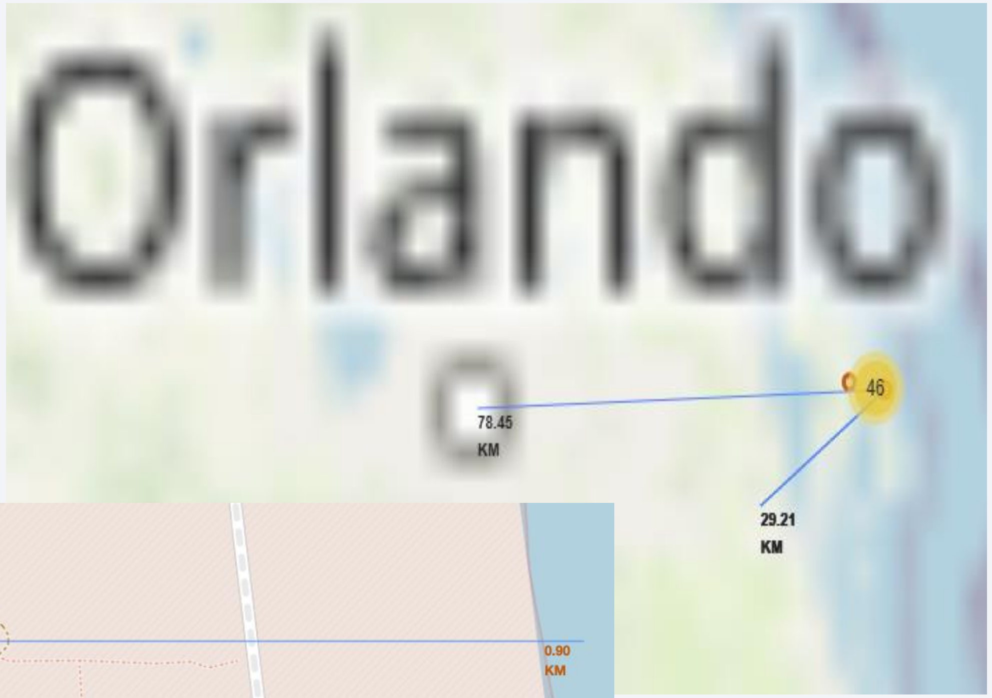
Launch Outcomes

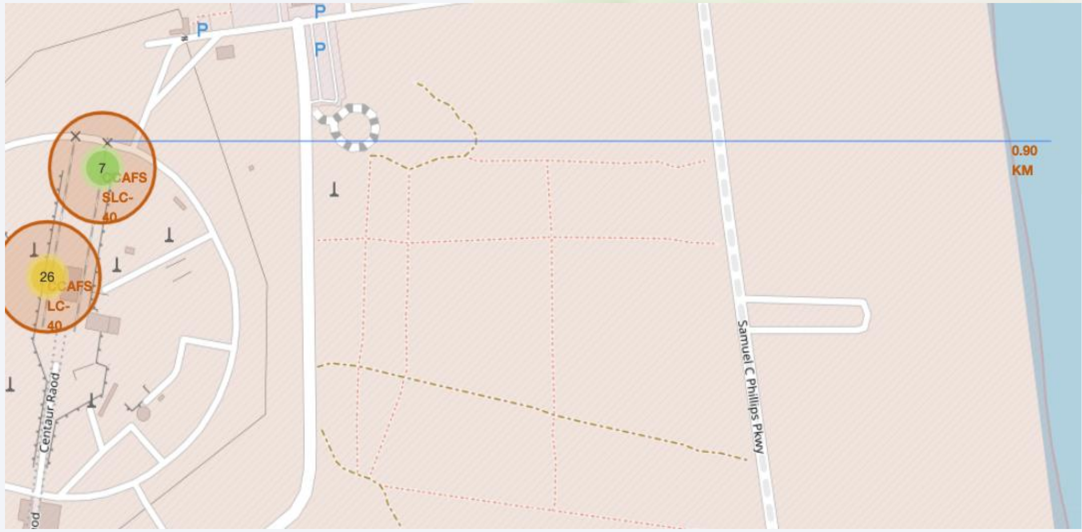
- It is observed that Florida launches were split unevenly between successful (green) while other were not (red).



Launch Location Proximities







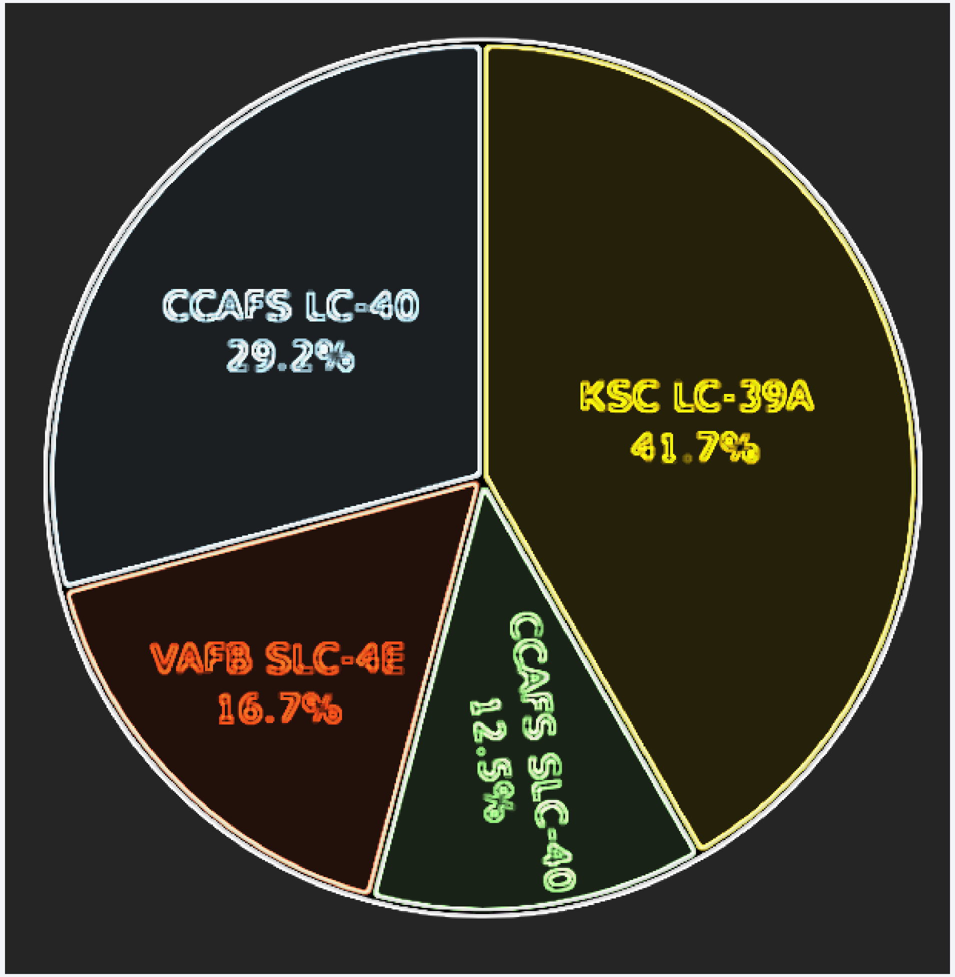
- Launch sites are close to the coast
- Launch sites are away from cities, highways and railways.

Section 4

Build a Dashboard with Plotly Dash

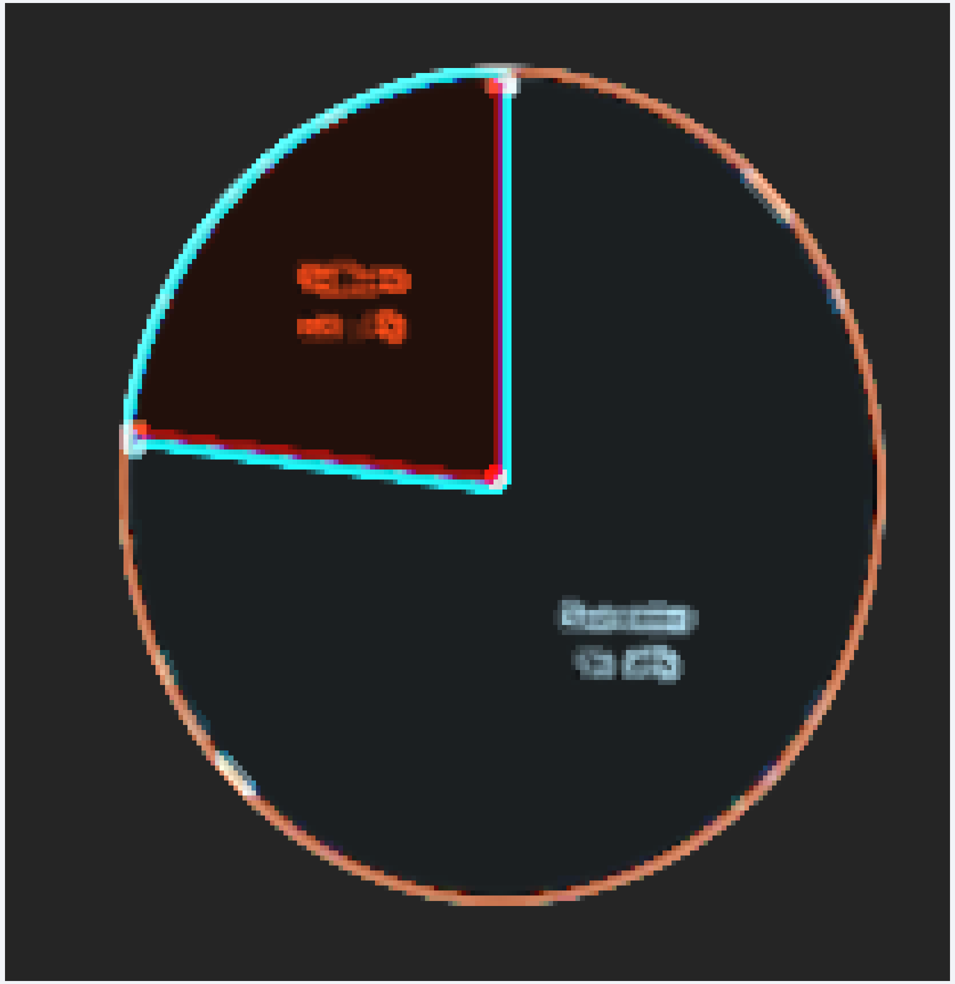
Success Rate by Launch Site

- Most successful launches in order:
  - KSC LC-39A
  - CCAFS LC-40
  - VAFB SLC-4E
  - CCAFS SLC-40



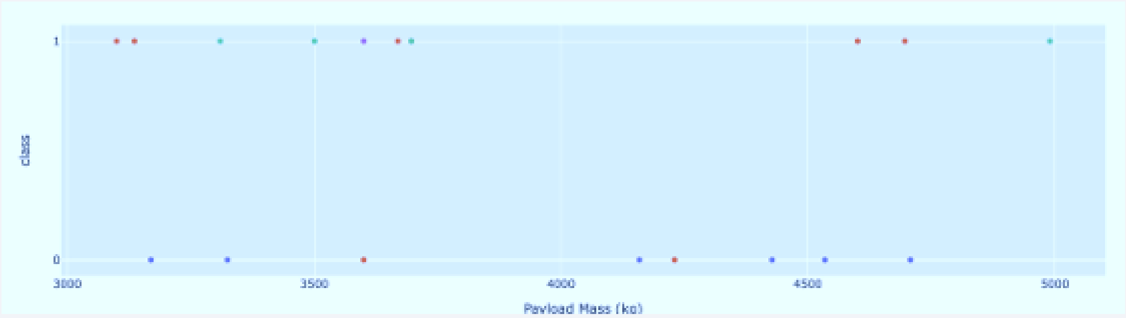
Top Launch Site Success to Failure Rate

- KSC LC-39A launch site stats:
  - Success – 76.9%
  - Failure – 23.1%



Payload Size to Lauch Success Rate

- The success rate increases for lower weight payloads.
- For higher weight payloads, the success rate decreases.

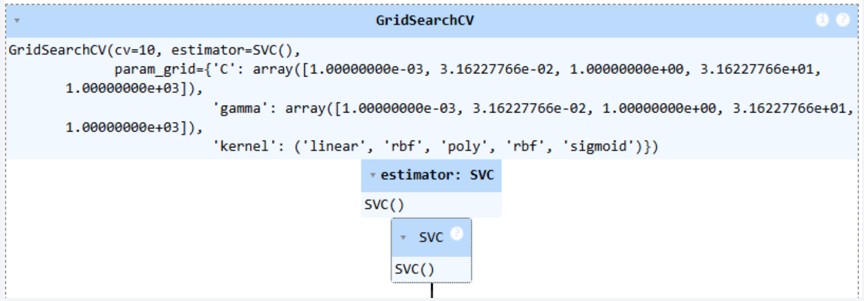


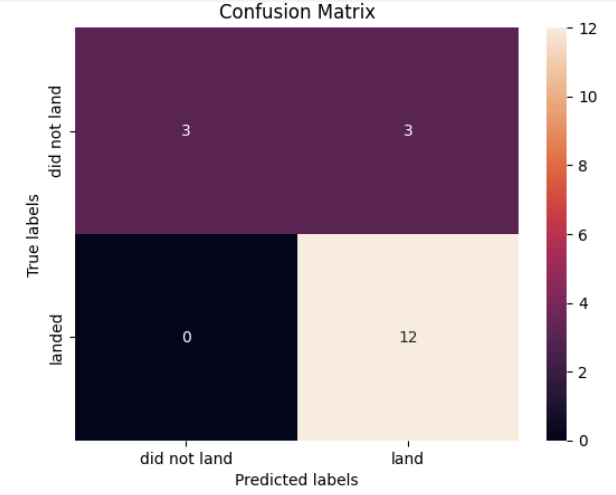
Section 5

Predictive Analysis (Classification)

Classification Accuracy

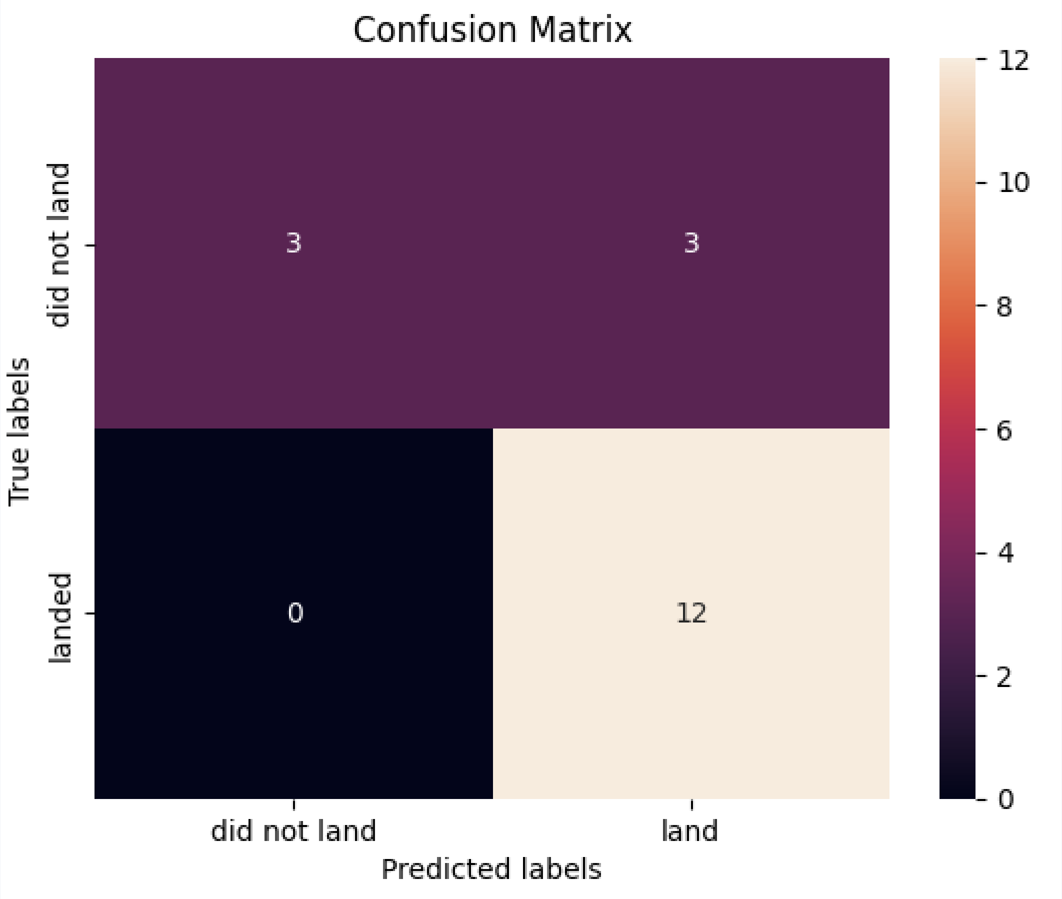
- DecisionTree classifier had the higher accuracy of all the different models tested.





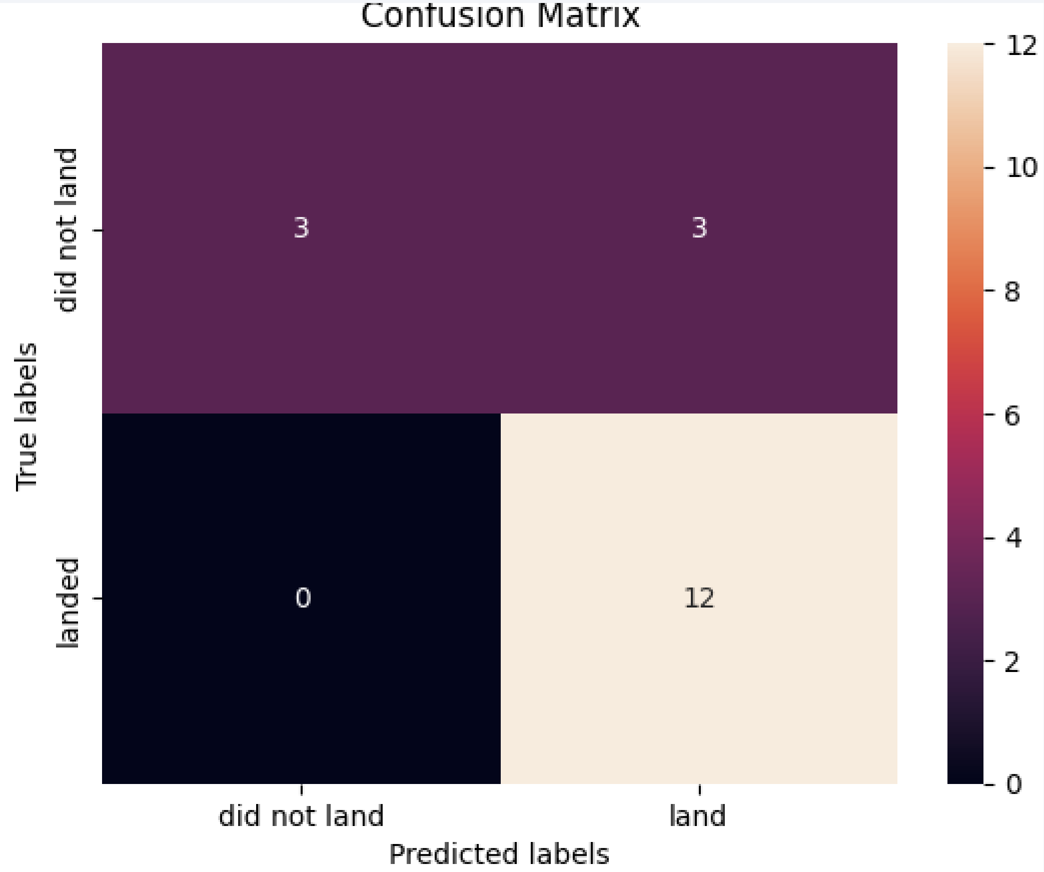
Confusion Matrix

- 12 successful landings were predicted
- 3 unsuccessful landings were predicted
- 3 unsuccessful landings were predicted
- 0 unsuccessful landings were predicted



Conclusions

- Orbits including ES-L1, Sun-Synchronous Orbit (SSO), Highly Elliptical Orbit (HEO), and Geostationary Orbit (GEO) often achieve impressive success levels.
- Success rate increased steadily from 2013 to 2020.
- CCAFS SLC 40 has conducted the most launches, whereas VAFB SLC 4E demonstrates a greater rate of successful landings.
- Launch sites are close to the coast and away from cities, highways and railways.
- Best Performing Algorithm is Tree.



Appendix

- SpaceX API
- List of Falcon 9 and Falcon Heavy launches (Wikipedia)
- Dataset_part_1.csv
- Dataset_part_2.csv
- Spacex_launch_geo.csv
- Spacex_launch_dash.csv

Thank you!