IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Manny Fomundam
August 24, 2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Dats Wrangling
  - Exploratory Data Analysis
    - With Data Visualization
    - With SQL
  - Building a Dashboard with Plotly Dash
  - Predictive Analysis (Classification
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive Analytics Demo (Screenshots)
  - Predictive Analysis Results

Introduction

- Project background and context
  - SpaceX is the SpaceX has revolutionized the space industry by significantly reducing the cost of access to orbit. While a traditional rocket launch can cost over $165 million, SpaceX offers its Falcon 9 launches for $62 million. These dramatic savings are primarily achieved through the reusability of the rocket's first stage.
  - This project directly ties technical success to financial outcome. By predicting whether the first stage of a Falcon 9 rocket will land successfully, we can effectively model the cost of a launch. Using public data and machine learning, we build a classifier to make this critical prediction.
- Problems you want to find answers
  - How do different variables like payload mass, launch site, number of flights, and orbits affect the success of the first landing stage
  - Does the rate of successful landings increase over the years
  - What is the best ML algorithm that can be used for binary classification in this case

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected through SpaceX Open Source Rest API
  - Data was also collected through ‘List of Falcon 9 and Falcon Heavy Launches’ from Wikipedia
- Perform data wrangling
  - Transformed categorical data with One Hot Encoding for ML algorithms and also removed any empty and unnecessary information
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Decision Trees were deployed to determine the best classification methods.

Data Collection

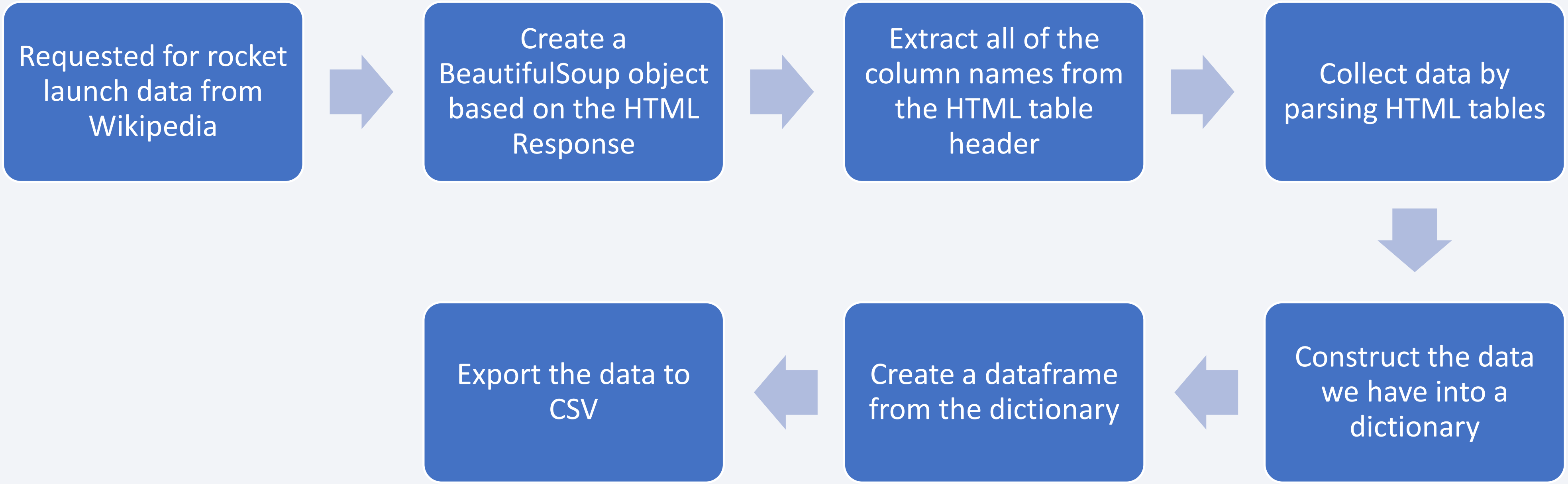
- SpaceX API
  - Collected past launch data from SpaceX from their open-source API
  - Retrieved and processed data with GET Request
  - Made sure the data only included Falcon 9 launches
  - Dealt with missing payload weights from secret missions
- Web Scraping
  - Asked for past Falcon 9 and Falcon Heavy launch data from Wikipedia
  - Accessed the Falcon 9 launch page
  - Retrieved all columns from HTML Table
  - Parsed and transformed the table into a pandas dataframe, which makes it easier to analyze.

Data Collection – SpaceX API

- Request rocket launch data from SpaceX API
- Decode the response content using .json() and turn it into a dataframe with .json.normalize()
- Request needed info about the launches from SpaceX API with applying custom functions
- Build data we have gained into a dictionary
- Create a dataframe
- Filter the dataframe to only have Falcon 9 launches
- Relace missing values of the Payload Mass column with calculated.mean() for the column
- Export the data to a csv file

GitHub Link 1

Data Collection - Scraping



[GitHub Link 2]

Data Wrangling

Perform EDA and determine Training Labels → Calculate the number of launches for each site → Calculate the number and occurrence of each orbit → Calculate the number and occurrence of mission outcomes per orbit type → Create a landing outcome label based on the Outcome column → Export data to CSV

[GitHub Link 3]

EDA with Data Visualization

- Charts that were Plotted
  - Flight Number vs Payload Mass, Flight Number vs Launch Site, Payload Mass vs Launch Site, Orbit Type vs Success Rate, Flight Number vs Orbit Type, Payload Mass vs Orbit Type, and Success Rate Yearly Trend
- Charts Used
  - Scatter Plots
    - Show relationships between variables.
  - Bar Charts
    - Shows comparisons among discrete categories.
  - Line Charts
    - Show trends in data over time.

GitHub Link 4

EDA with SQL

[GitHub Link 5]

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Rank the 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.

Build an Interactive Map with Folium

- Added a marker with Circle, Popup Label with text of NASA Johnson Space Center based on its latitude and longitude coordinates as a starting position
- Did the same with all Launch Sites using latitude and longitude coordinates to show their relationship with their locations and proximity to the Equator and coasts.
- Made colored markers of success and failed (green and red) launches using Marker cluster to show which launch sites had high levels of success.
- Also added colored lines to show distance between launch site KSC LC-39A (example) and its proximities to places like Railway, Highway, Coastline and Closest City.

[GitHub Link 6]

Build a Dashboard with Plotly Dash

- Added a dropdown list to allow for Launch Site selection
- Added a pie chart to show total success of launches from all sites and the success/fail for each site
- Added a slider for Payload range
- Added a scatter chart to show correlation between Payload and Launch Success

GitHub Link 6

Predictive Analysis (Classification)

1. Create a NumPy array from the “Class” column
2. Standardize the date with StandardScaler, then fitting and transforming it
3. Splitting the data into training and testing sets with train_test_split function
4. Create a GridSearchCV object with the cv = 10 to find the best parameters
5. Apply GridSearchCV on LogReg, SVM, Decision Tree, and KNN Models
6. Calculate the accuracy on the test data using the method.score for all models used
7. Examine the confusion matrix for all models
8. Find the method that that performs the best by examining the Jaccard_score and F1_score metrics

[GitHub Link 7]

Results

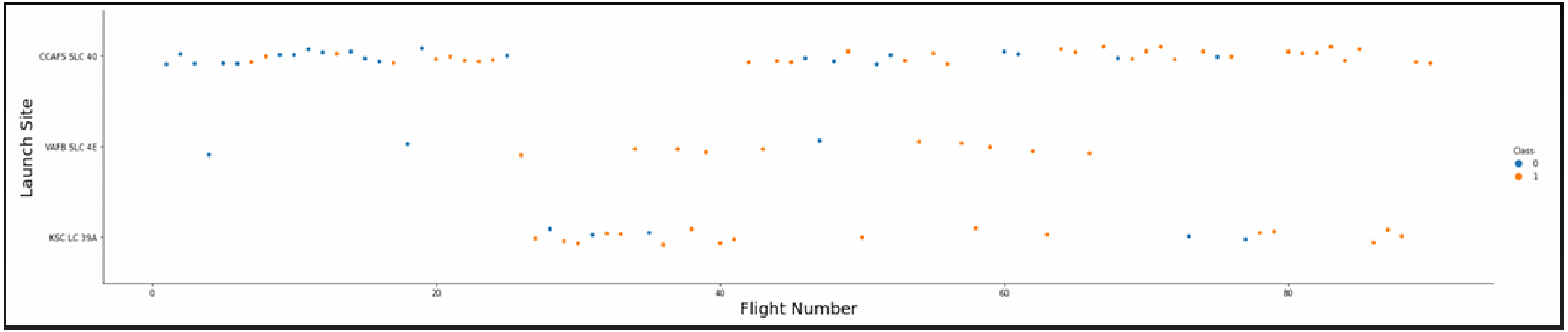
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

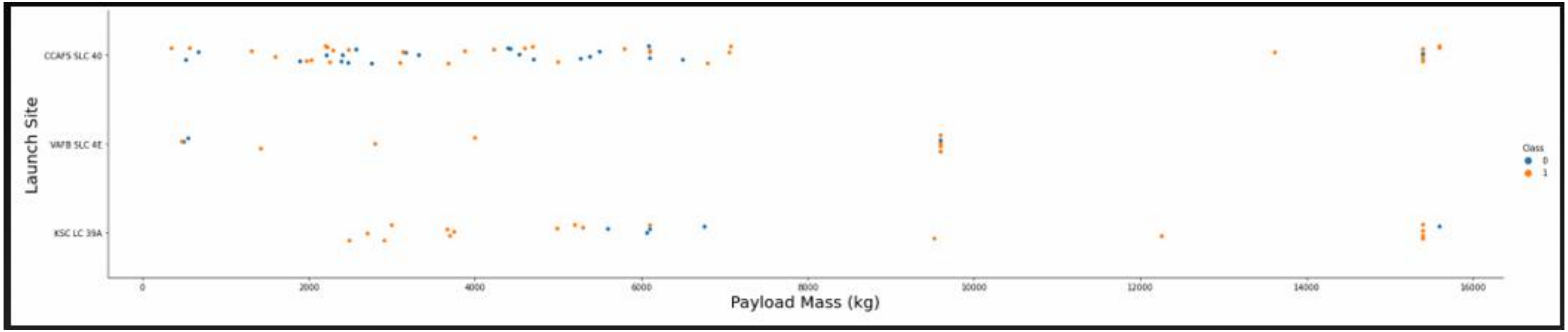
Flight Number vs. Launch Site

The exploratory data analysis uncovered several critical patterns. First, a pronounced temporal trend is evident; initial launch attempts were unsuccessful, whereas all recent missions have succeeded, highlighting SpaceX's rapid technological maturation. Second, the success rate varies considerably by launch site. Although Cape Canaveral's SLC 40 is the most frequently used site, Vandenberg's SLC 4E and Kennedy Space Center's LC 39A exhibit higher success ratios. This correlation suggests that launch location is a non-random factor influencing outcomes. Ultimately, the data supports the conclusion that the probability of a successful landing has increased monotonically over time, a crucial insight for predicting future launch costs and reliability.



Payload vs. Launch Site

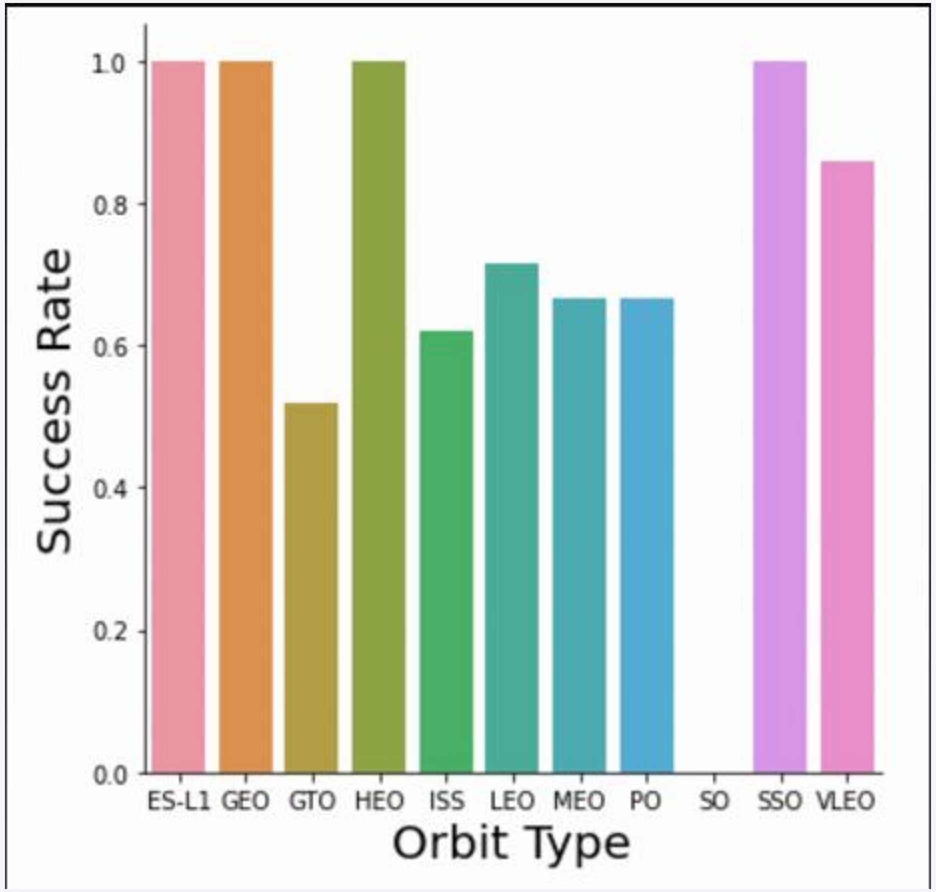
Our analysis reveals a strong positive correlation between payload mass and landing success rate across all launch sites. This trend is particularly evident for heavy payloads; missions with a mass exceeding **7000 kg demonstrate a near-perfect success rate**. Furthermore, **KSC LC-39A** shows exceptional performance, maintaining a **100% success rate even for lighter payloads** under 5500 kg.



Success Rate vs. Orbit Type

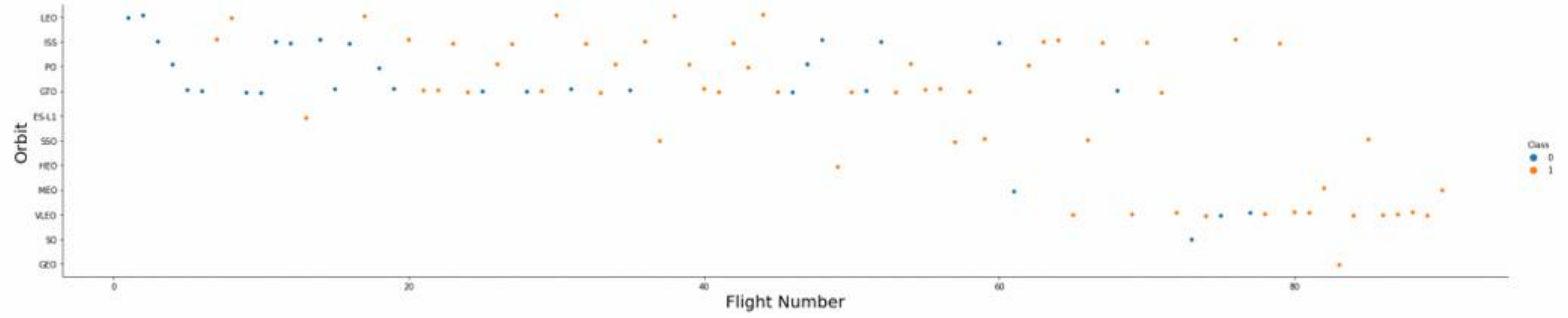
Our analysis of historical launches identifies a strong correlation between the target orbit and the success rate of the first-stage landing. Orbits can be grouped into three distinct tiers:

- **High-Success Orbits (100%):** Missions to **ES-L1, GEO, HEO, and SSO** have consistently resulted in successful landings, making them low-risk profiles for reusability.
- **Variable-Success Orbits (50-85%):** Launches to **GTO, ISS, LEO, MEO, and PO** show moderate success, indicating that other factors like payload mass and mission parameters play a significant role.
- **High-Risk Orbit (0%):** The **SO (Suborbital)** orbit shows a 0% success rate, representing a significant challenge for recovery efforts and a key area for investigation.



Flight Number vs. Orbit Type

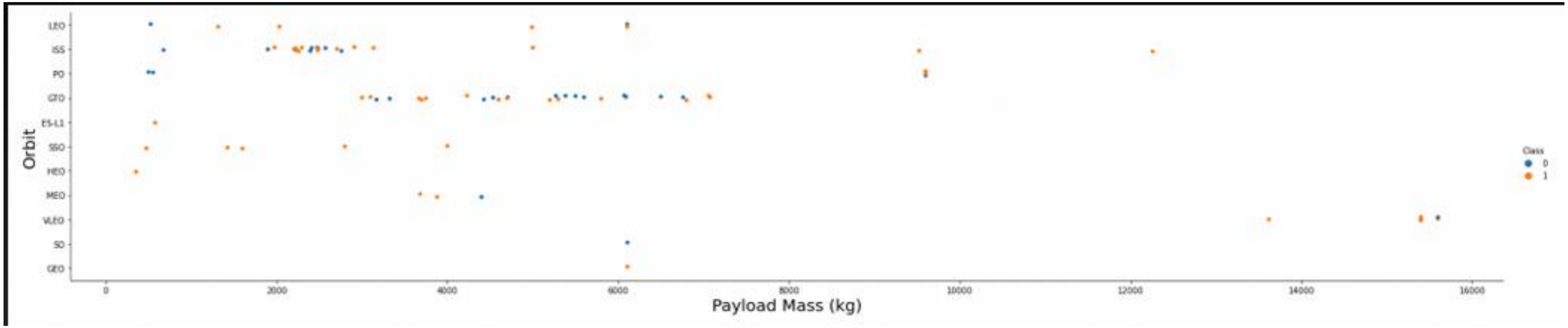
- The success rate for missions in **Low Earth Orbit (LEO)** shows a positive correlation with the number of flights, suggesting improved performance through experience. In contrast, success rates for **Geostationary Transfer Orbit (GTO)** missions remain consistently high regardless of flight number, indicating mastery of this more challenging profile.



Payload vs. Orbit Type

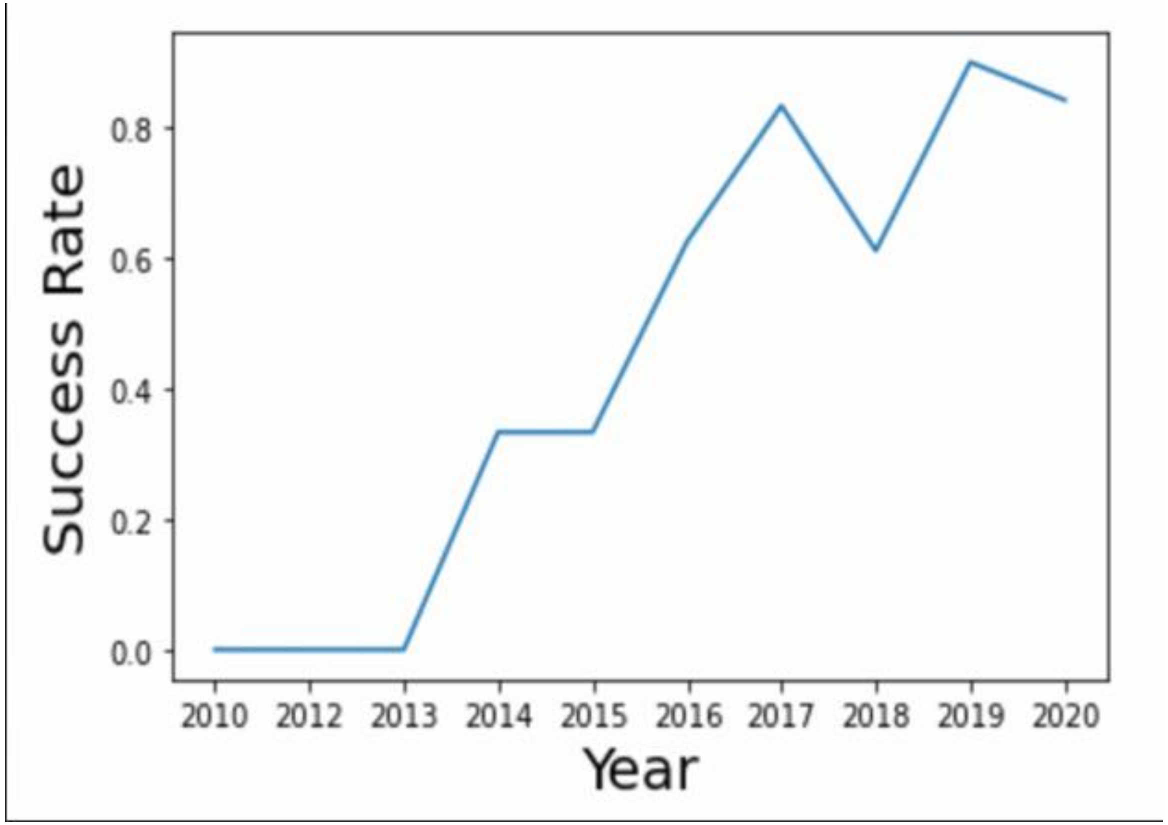
The mass of the payload significantly influences the likelihood of a successful landing, but this effect varies drastically by orbit type:

- **GTO (Geostationary Transfer Orbit):** Heavy payloads have a strong **negative** effect on landing success. Achieving this high-altitude orbit requires the rocket to expend almost all of its fuel, leaving minimal margin for the propellant needed to land the first stage.
- **LEO (Low Earth Orbit) such as ISS or Polar orbits:** Heavy payloads have a **neutral or slightly positive** correlation with landing success. These missions require less energy, leaving sufficient fuel for the rocket to execute its landing maneuvers. The presence of a heavy payload may also indicate a mission profile that was designed for recovery from the outset.



Launch Success Yearly Trend

- Analysis of launch history shows a marked improvement in landing success rates over time. Starting from initial attempts in 2013, SpaceX demonstrated a consistent year-over-year increase in reliability, culminating in a highly successful period leading up to 2020.



All Launch Site Names

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

Task 1

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

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

✓ 0.0s Python

```
 * 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'

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

Task 2

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

[24]
```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```
Python

* sqlite:///my_data1.db
Done.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landi
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failur
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	Failur
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass

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

Task 3

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

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") AS Total_Payload_Mass FROM SPACEXTBL WHERE "Customer" = 'NASA (CRS)';
```

Python

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

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

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

Task 4

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

```
[39]  %sql SELECT AVG("PAYLOAD_MASS__KG_") AS Average_Payload_Mass FROM SPACEXTBL WHERE "Booster_Version" LIKE '%F9 v1.
```

Python

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

Average_Payload_Mass
2534.6666666666665

First Successful Ground Landing Date

Listing the date when the first successful landing outcome in ground pad was achieved

Task 5

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

Hint:Use min function

```
%sql SELECT MIN("Date") AS First_Successful_Landing FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (ground pad
```

[40] Python

* sqlite:///my_data1.db
Done.

First_Successful_Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 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 SPACEXTBL WHERE "Landing_Outcome" = 'Success (drone ship)' AND PAYLOAD_MASS__K
```

Python

```
 * 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

Listing the 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_Number FROM SPACEXTBL GROUP BY Mission_Outcome;
```

[42]

Python

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

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

Boosters Carried Maximum Pay load

Listing the names of the booster versions which have carried the maximum payload mass.

Task 8

List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

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

Python

```
 * 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

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql SELECT CASE WHEN substr("Date", 6, 2) = '01' THEN 'January' WHEN substr("Date", 6, 2) = '02' THEN 'February'
```

Python

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

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

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

- Ranking the 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.

Task 10

Rank the 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.

```
%sql SELECT "Landing_Outcome", COUNT(*) as Count_Outcomes from SPACEXTBL GROUP BY "Landing_Outcome" ORDER BY Cou
```

Python

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

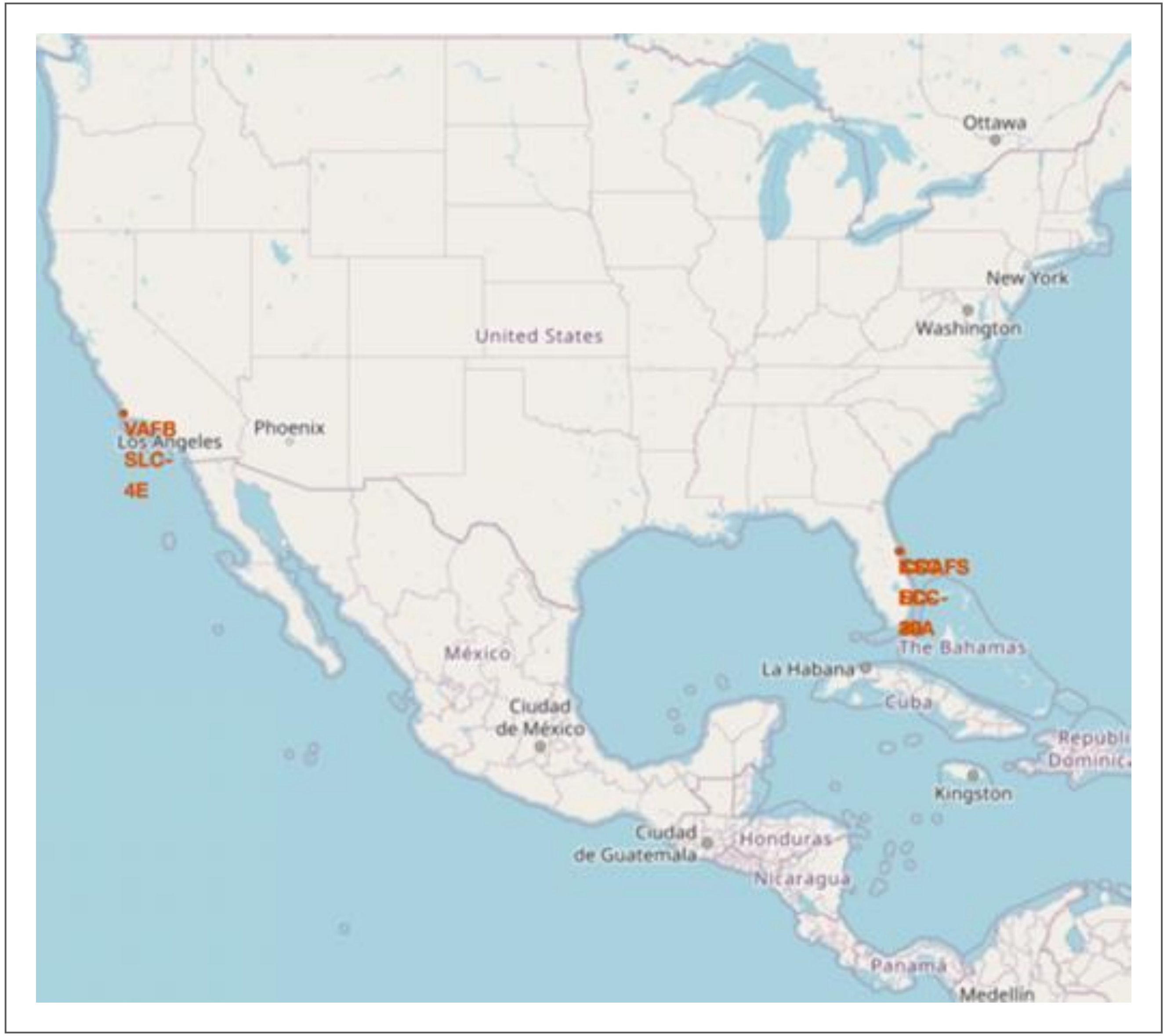
Landing_Outcome	Count_Outcomes
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Failure	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1

Section 3

Launch Sites Proximities Analysis

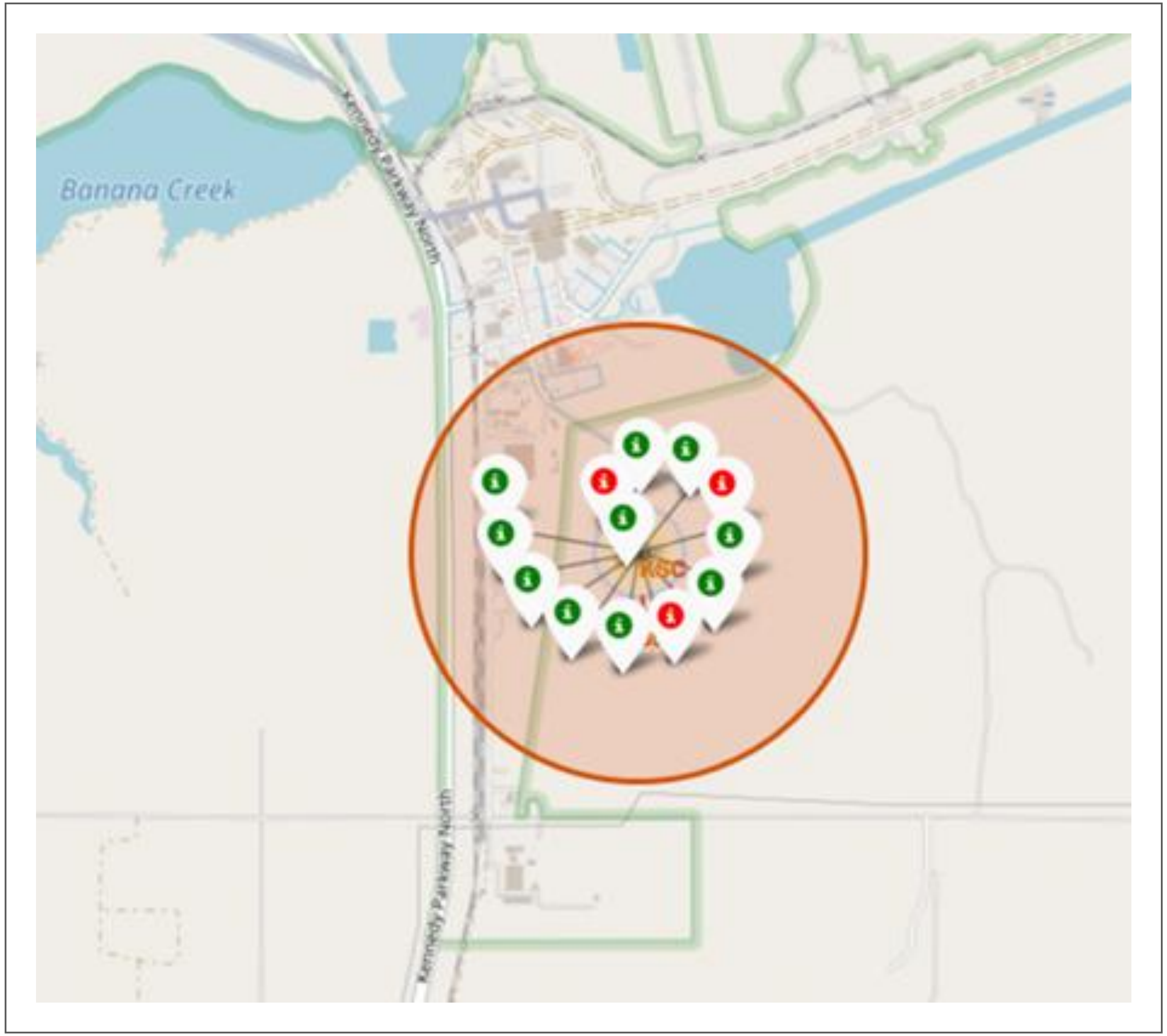
<Folium Map Screenshot 1>

- The geographic location of launch sites is strategically chosen for both **physics** and **safety** reasons.
- **Proximity to the Equator (The Physics Advantage):** Launch sites near the equator capitalize on the Earth's maximum rotational velocity (approximately 1670 km/h). A rocket launched eastward from here already possesses this significant horizontal speed due to rotational inertia. This "free" velocity boost reduces the fuel required to achieve orbital speed, making launches more efficient and cost-effective.
- **Proximity to the Coastline (The Safety Advantage):** Positioning launch pads near the coast allows for launches to be directed over open ocean. This provides a large, unpopulated area for the controlled disposal of spent rocket stages and reduces the risk to human life and infrastructure in the event of a launch failure or if debris falls back to Earth.



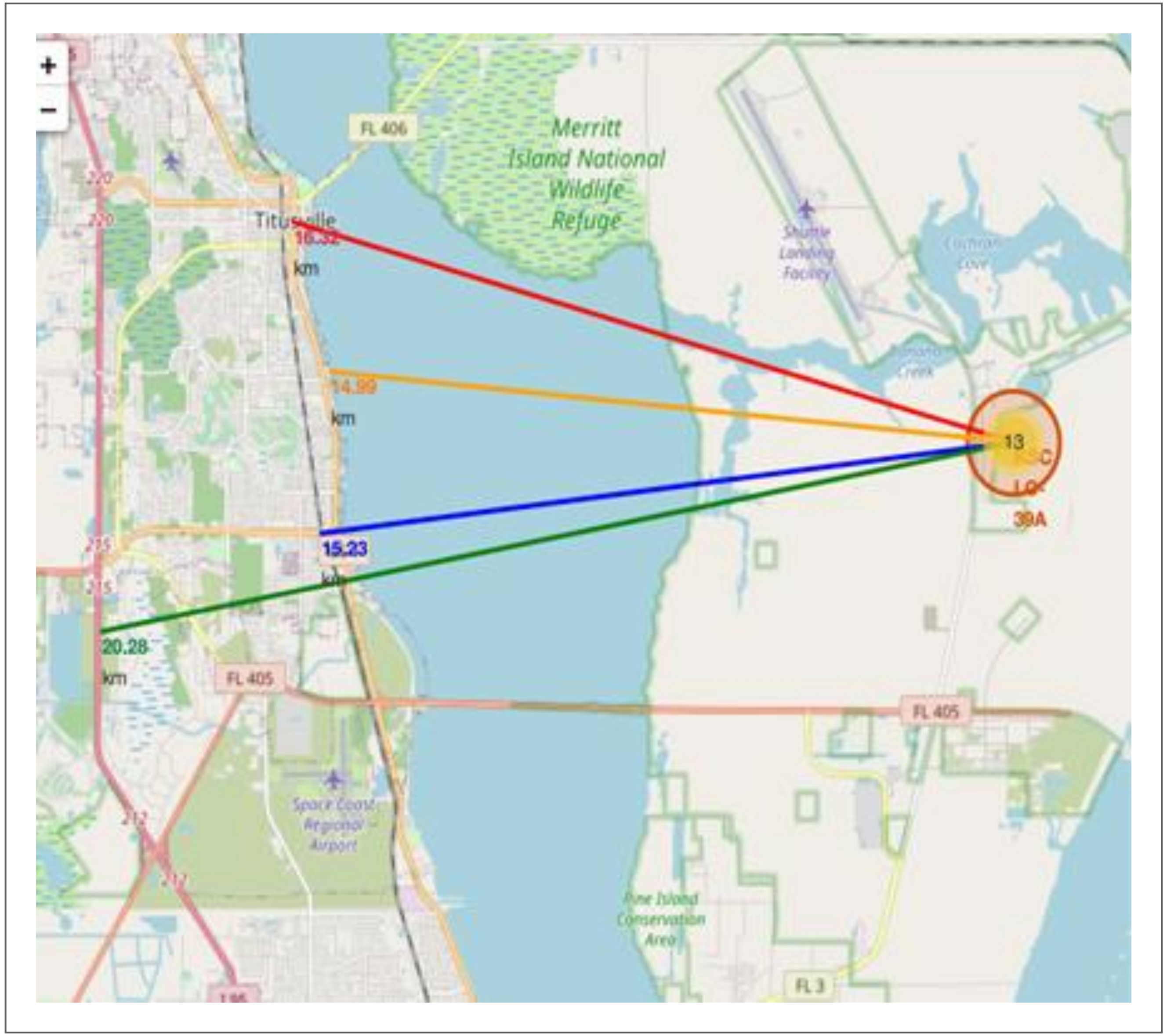
<Folium Map Screenshot 2>

- The success rate of launch sites is effectively communicated through a color-coded marker system on an interactive map. A preliminary visual assessment reveals a clear correlation between launch site and success rate.
- **Successes (Green)** and **Failures (Red)** are plotted geographically.
- **Key Finding:** The launch site **KSC LC-39A** demonstrates an exceptionally high density of successful missions (green markers), establishing it as a primary site for successful Falcon 9 operations.



<Folium Map Screenshot 3>

- **Safety Analysis of Launch Site KSC LC-39A**
- The visual analysis of Launch Complex 39A (KSC) reveals its proximity to key infrastructure:
  - **Railway:** 15.23 km
  - **Highway:** 20.28 km
  - **Coastline:** 14.99 km
  - **Nearest City (Titusville):** 16.32 km
- This proximity presents a significant safety consideration. A rocket failure during launch could propel debris at high velocity, covering distances of 15-20 km in mere seconds and posing a potential risk to populated areas and critical infrastructure.

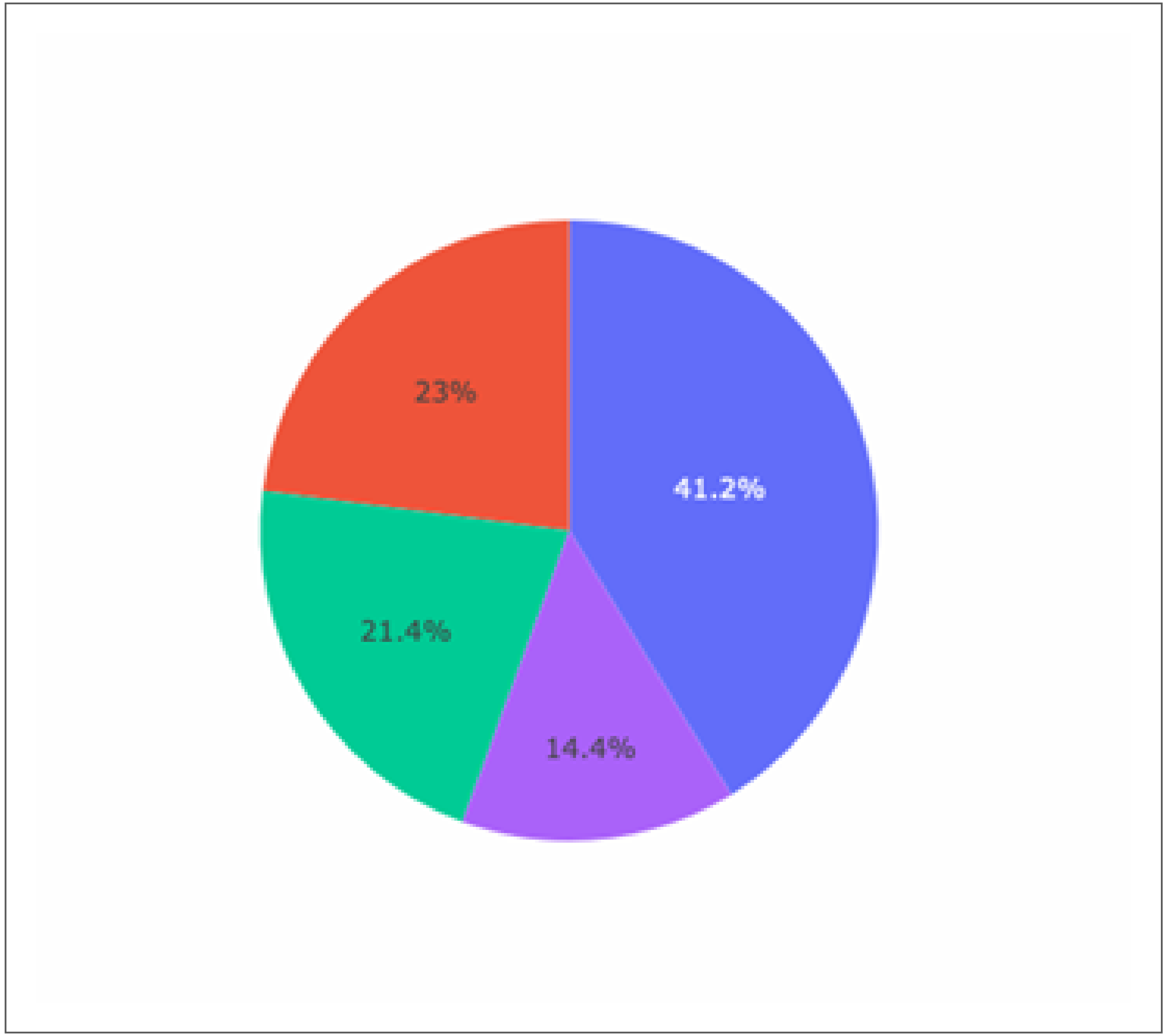


Section 4

Build a Dashboard with Plotly Dash

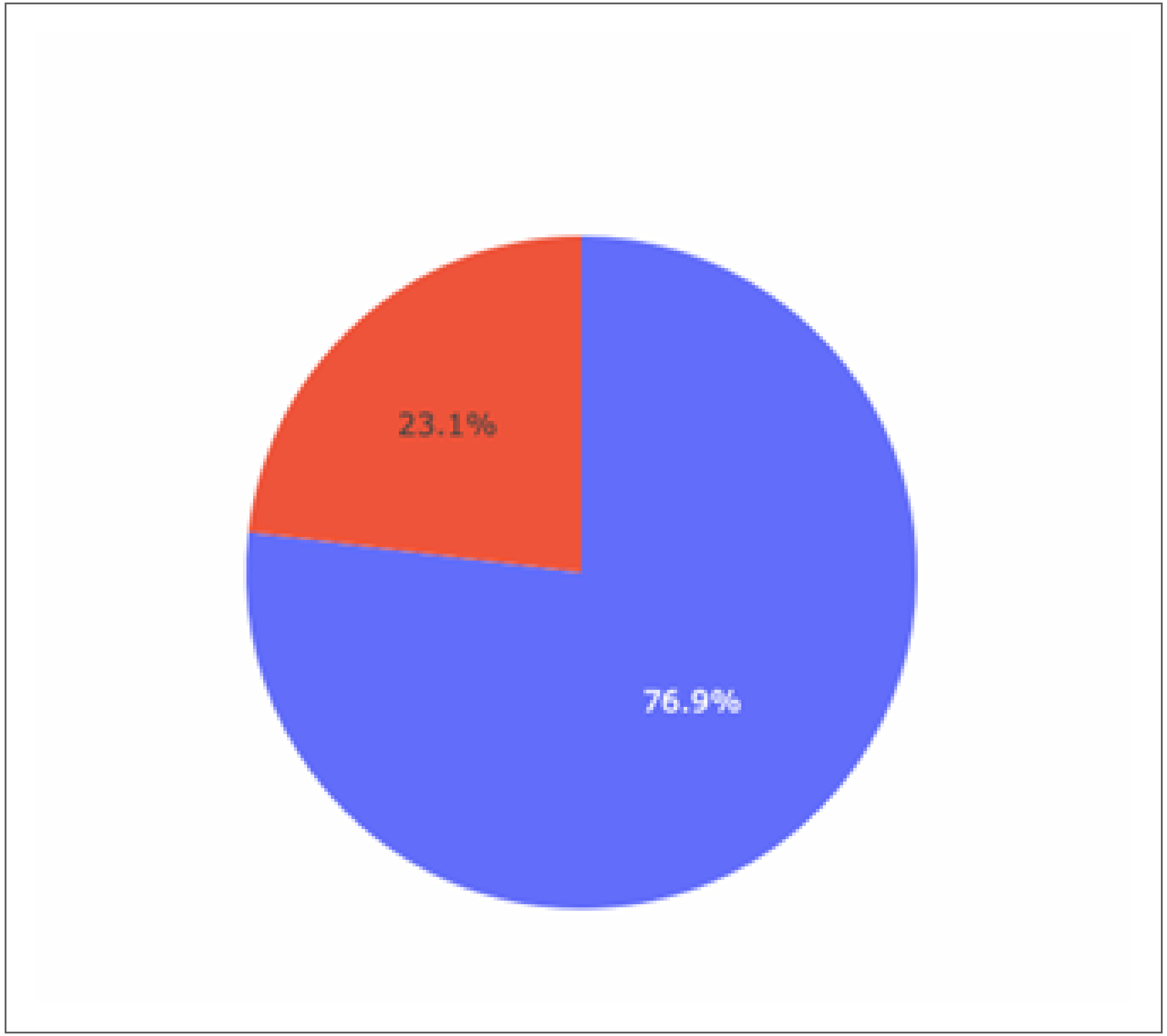
<Dashboard Screenshot 1>

- The data indicates that **KSC LC-39A is the highest-performing launch site**, boasting the greatest number of successful landings.



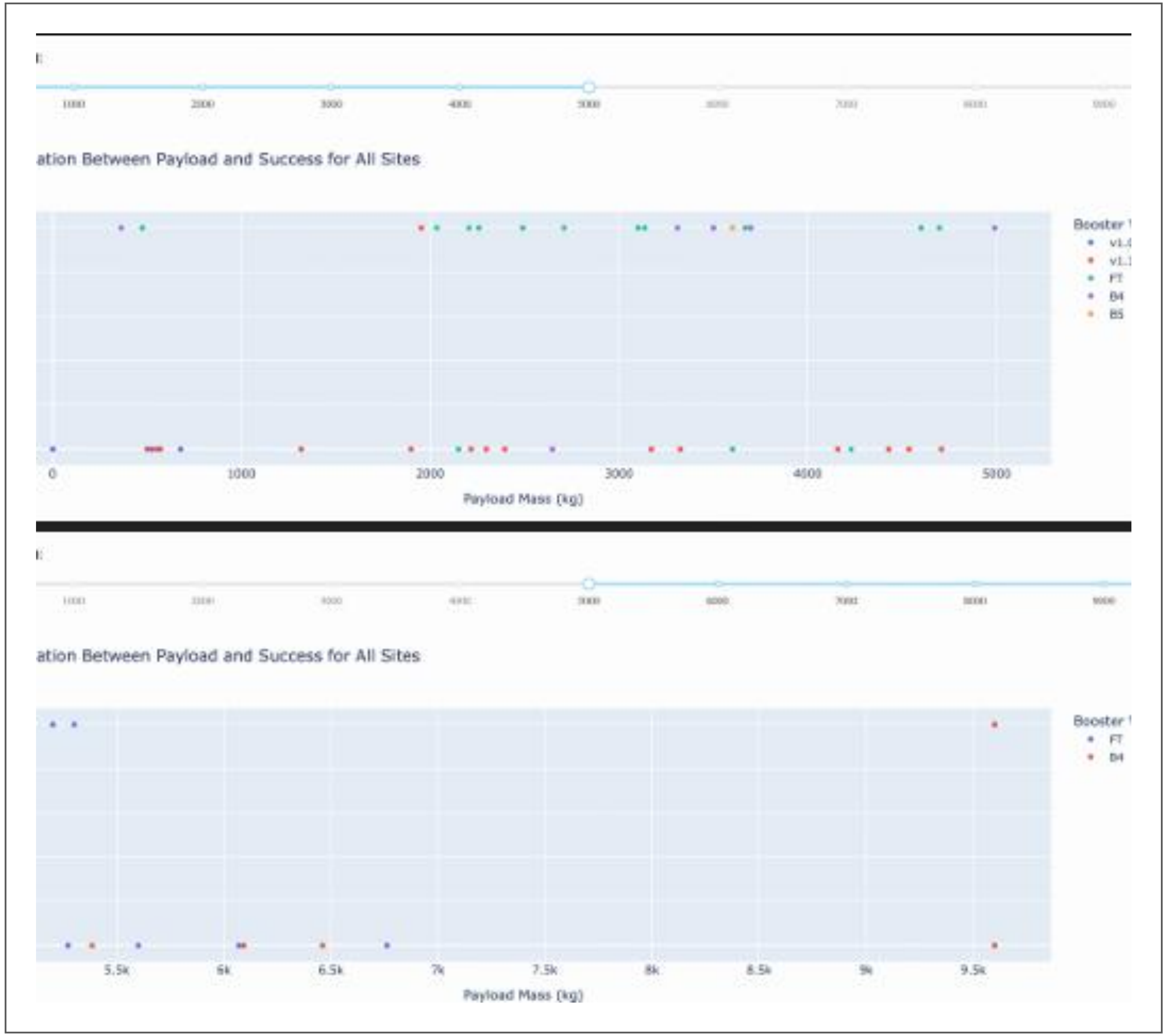
<Dashboard Screenshot 2>

- Among all launch sites, **KSC LC-39A boasts the highest success rate (76.9%)**, with 10 successful landings against only 3 failures.



<Dashboard Screenshot 3>

- **Payload mass is a critical success factor.** Missions with payloads between 2,000 kg and 5,500 kg demonstrate the highest first-stage landing success rate.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

Initial evaluation on the test set (n=18) yielded similar scores across all models, making it difficult to identify a superior performer due to the limited sample size. Consequently, models were evaluated on the entire dataset to leverage more data. This comprehensive analysis confirmed that the **Decision Tree Classifier** was the optimal model, demonstrating the highest accuracy and overall scores.

Scores and Accuracy of the Test Set

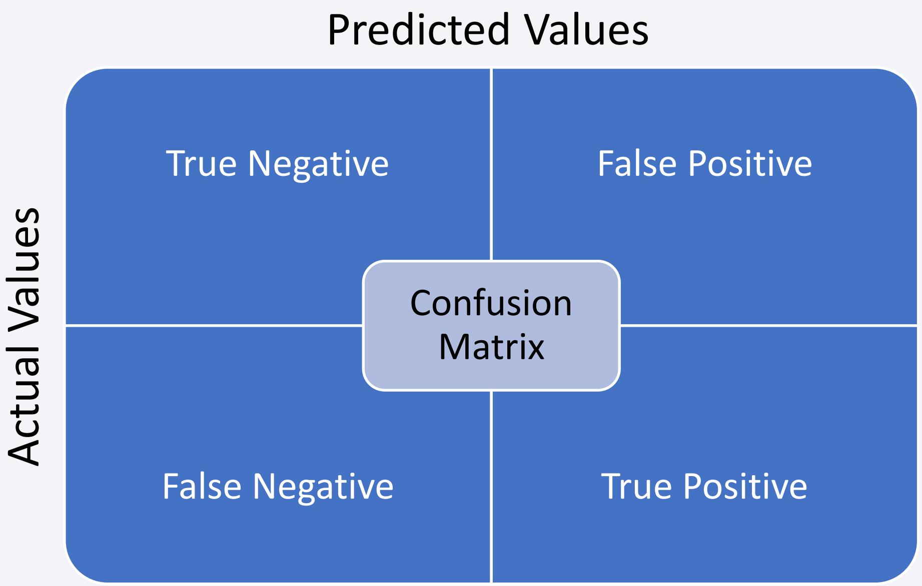
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

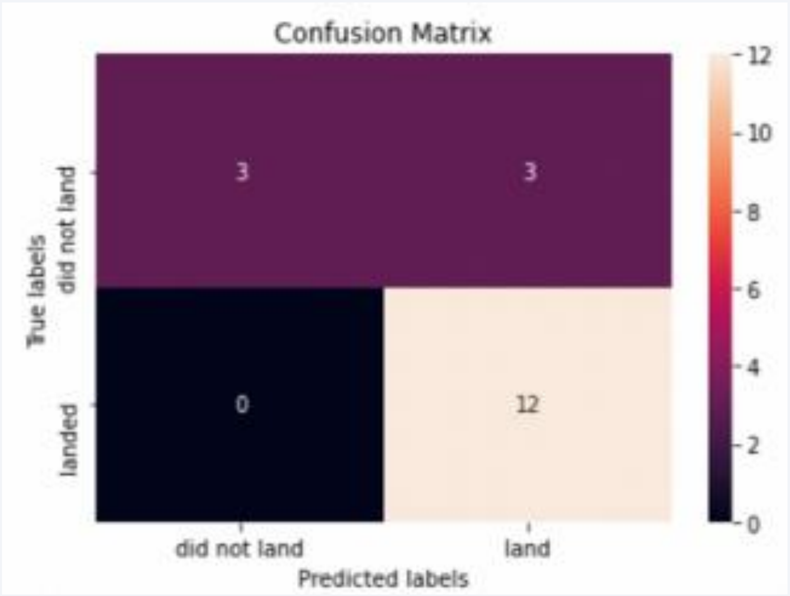
Scores and Accuracy of the Entire Data Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

The confusion matrix indicates that the logistic regression model effectively distinguishes between classes. However, its primary weakness is a higher incidence of false positives (Type I errors), where failed landings are incorrectly predicted as successful.





Conclusions

- Decision Trees is the best ML algorithm
- Launches that have a low payload mass have better results than launches with a larger payload
- Most launch sites are close to the Equator and all sites are close to the coasts
- The success rate has increased over the years
- Highest success rate = KSC LC-39A
- Orbits ES-L1, GEO, HEO and SSO => 100% success rate

Appendix

Thanks to

IBM
Coursera

Thank you!