

# Winning Space Race with Data Science

<Dang Phuoc Hung>  
<February 24,2026>



# Outline

---

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

# Executive Summary

---

- In this capstone project, we aim to predict whether the SpaceX Falcon 9 first stage will successfully land by applying various machine learning classification models.
- The project workflow includes data collection and preprocessing, exploratory data analysis, interactive data visualization, and predictive modeling.
- Through data analysis, we observe that certain launch features, such as payload, orbit, and launch site are associated with mission outcomes(success or failure).
- Among the models evaluated, the Decision Tree algorithm demonstrated the best performance in predicting first stage landing success.

# Introduction

---

- In this capstone project, we aim to predict whether the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 launches at approximately \$62 million per launch, while competitors typically charge more than \$165 million. A significant portion of this cost advantage comes from the ability to reuse the first stage booster. Therefore, accurately predicting landing success helps estimate the overall launch cost and provides valuable insight for companies competing with SpaceX in bidding for launch contracts.
- It is important to note that not all unsuccessfully landings are accidental, in some missions, SpaceX intentionally performs controlled landings in the ocean.
- The primary research question of this project is given a set of launch characteristics, such as payload mass, orbit type, and other features, can we accurately predict whether the Falcon 9 first stage will land successfully.

Section 1

# Methodology

# Methodology

---

## Executive Summary

- **Data collection methodology:**
  - Collected launch data from SpaceX REST API. Supplemented data through web scraping. Combined data sets for unified analysis.
- **Perform data wrangling**
  - Cleaned and processed raw data. Handled missing values. Encoded categorical variables. Created binary target variable for classification.
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
  - Built multiple classification models. Tuned hyperparameters using GridSearchCV. Evaluated models using accuracy and confusion matrix. Identified best-performing model.

# Data Collection

---

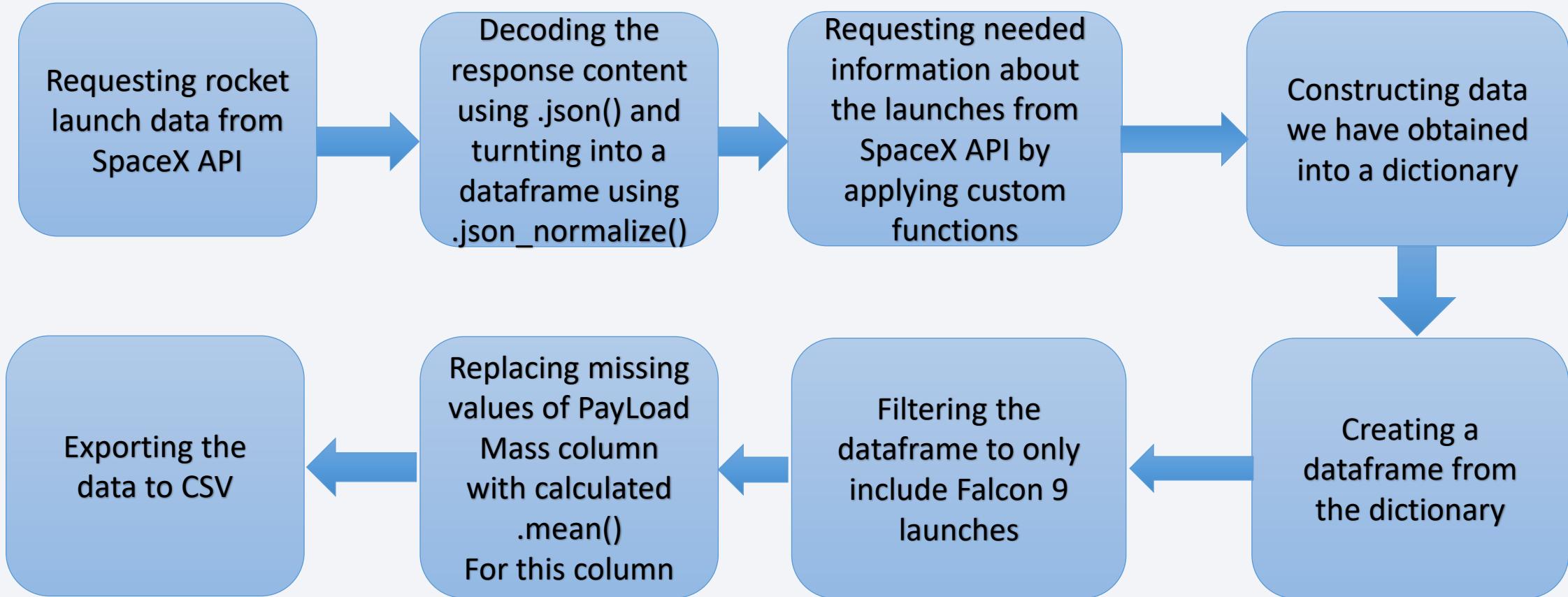
## Data Collection Process

- SpaceX Rest API:
  - Retrieved structured launch data.
  - Filtered for Falcon 9 missions.
  - Extracted payload, orbit, booster, launch site.
- Web Scraping:
  - Scraped Falcon 9 launch records from Wikipedia.
  - Extracted mission outcomes.
  - Validated and supplemented API data.

## Flowcharts

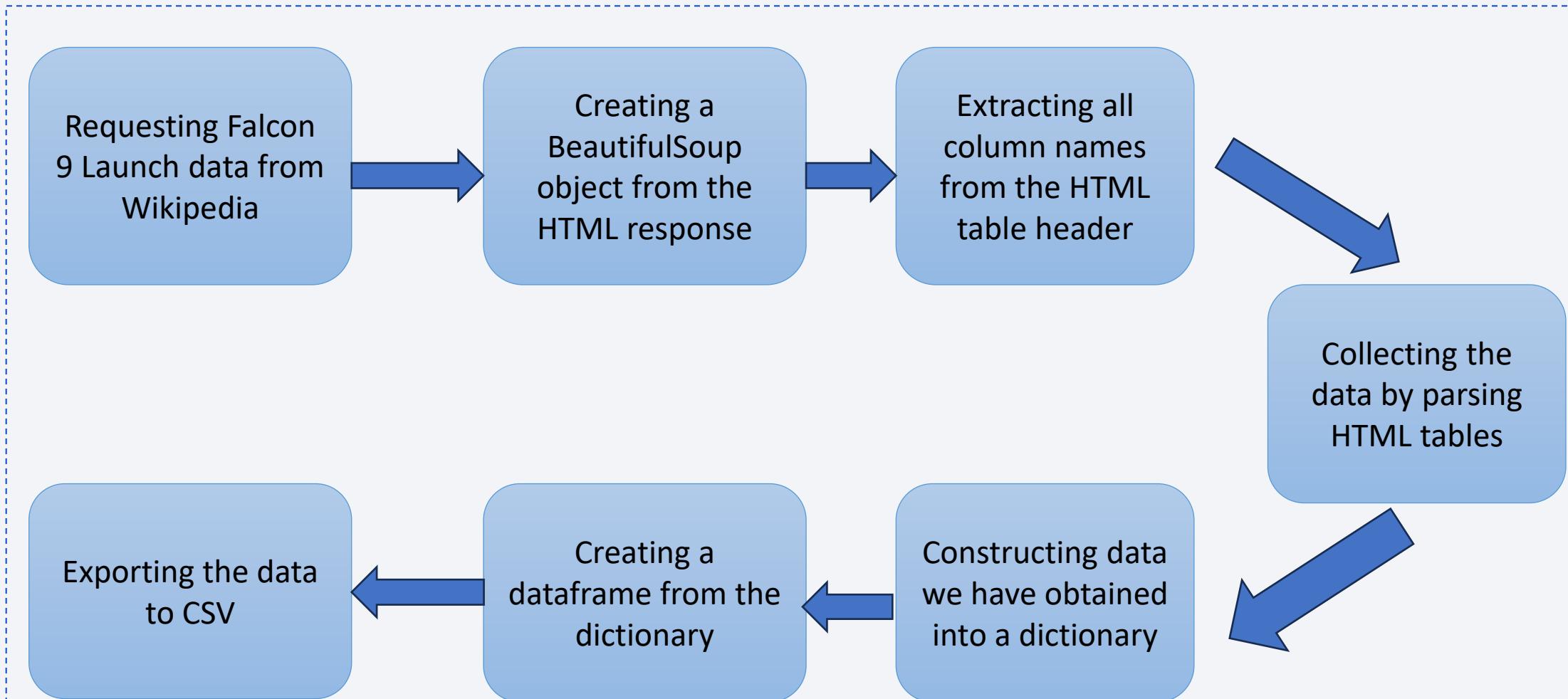
- SpaceX API → JSON Data → DataFrame
- Web Scraping → HTML Tables → DataFrame
- DataFrames → Merge → Final Dataset

# Data Collection – SpaceX API



[GitHub URL: Data Collection API](#)

# Data Collection - Scraping



# Data Wrangling

- The dataset includes multiple landing outcome categories, reflecting different landing attempts and locations. Some missions resulted in successful landings, while others failed due to accidents or controlled outcomes.
- For example, **True Ocean** indicates a successful landing in a designated ocean area, whereas **False Ocean** indicates an unsuccessful ocean landing. Similarly, **True RTLS** and **False RTLS** represent successful and unsuccessful landings on a Ground pad, respectively. **True ASDS** and **False ASDS** indicate successful or failed landings on a drone ship.
- For modeling purposes, all landing outcomes were converted into binary training labels, where “**1**” represents a successful landing and “**0**” represents an unsuccessful landing.

Perform exploratory Data Analysis and determine Training labels

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column

Exporting the data to CSV

# EDA with Data Visualization

---

Several charts were plotted to explore relationships between key variables, including:

- 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
- Yearly Success Rate Trend

Scatter plots were used to examine relationship between continuous variables and to identify potential patterns useful for machine learning models.

Bar charts were used to compare performance across discrete categories, highlighting differences between launch sites and orbit types.

Line charts were used to analyze trends over time, particularly to observe changes in launch success rates across years.

# EDA with SQL

---

The following SQL queries were conducted to explore and analyze the dataset:

- Identified unique launch site names
- Retrieved records where launch sites begin with “CCA”
- Calculated total payload mass carried by NASA(CRS) missions
- Computed average payload mass for booster version F9 v1.1
- Determined the date of the first successful ground pad landing
- Listed boosters that successfully landed on a drone ship with payload between 4000-6000 kg
- Calculated total counts of successful and failed missions
- Retrieved failed drone ship landings in 2015 along with booster and launch site details
- Ranked landing outcomes between 2010-06-04 and 2017-03-20 in descending order

# Build an Interactive Map with Folium

---

## Launch Site Markers

- Added circle markers with popup and text labels for NASA Johnson Space Center as the reference location
- Plotted all launch sites using latitude and longitude coordinates
- Visualized their geographic positions relative to the equator and coastlines

## Color-Coded Launch Outcomes

- Added green markers for successful launches and red markers for failed launches
- Used Marker Cluster to highlight launch sites with higher success rates

## Proximity Distance Analysis

- Drew colored lines to illustrate distances from KSC LC-39A(example site)
- Measured proximity to railway, highway, coastline, and nearest city

# Build a Dashboard with Plotly Dash

---

## Launch Site Selection

- Implemented a dropdown menu to allow users to select a specific launch site or view all sites

## Success Rate Visualization

- Created a pie chart displaying total successful launches for all sites
- When a specific site is selected, the chart shows Success vs Failure distribution

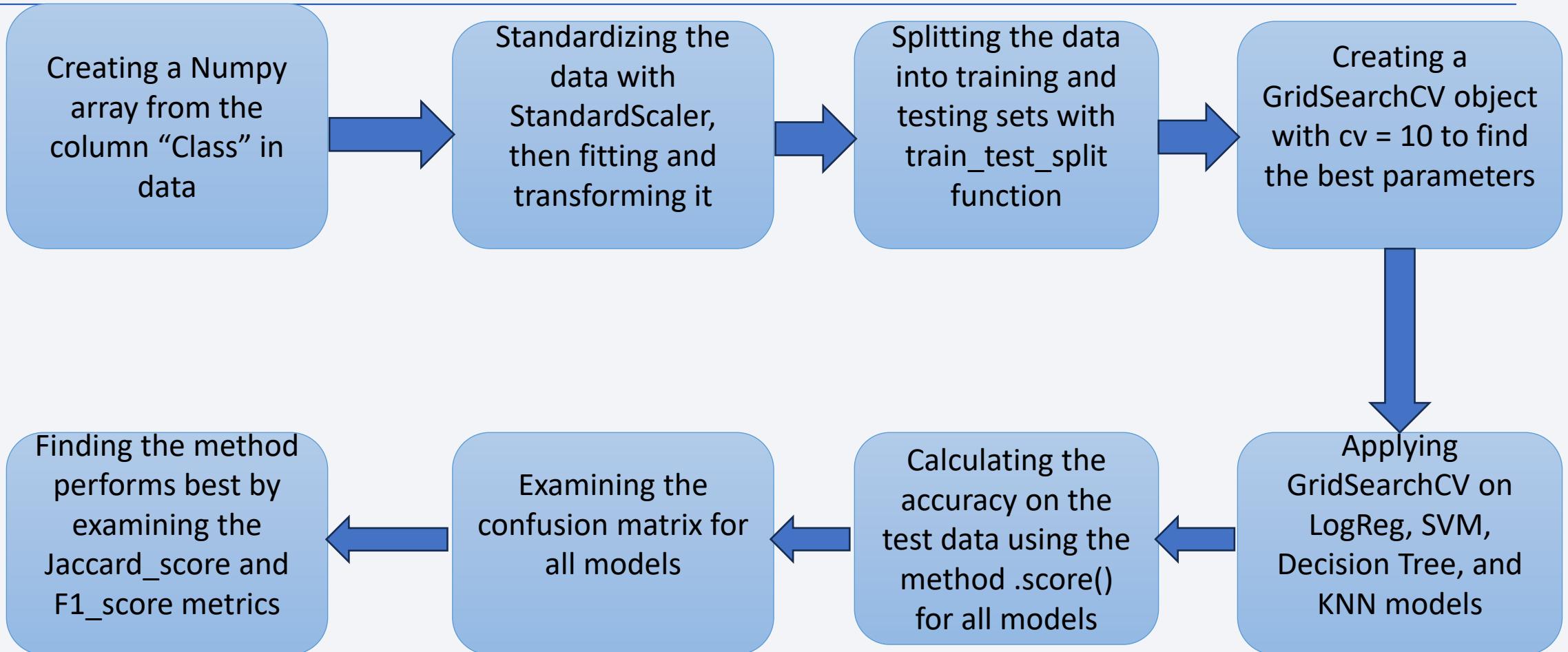
## Payload Range Filter

- Added a range slide to dynamically filter payload mass values

## Payload vs Launch Outcome Analysis

- Developed a scatter plot to illustrate the relationship between payload mass and landing success
- Colored by booster version to highlight performance differences

# Predictive Analysis (Classification)



# Results

---

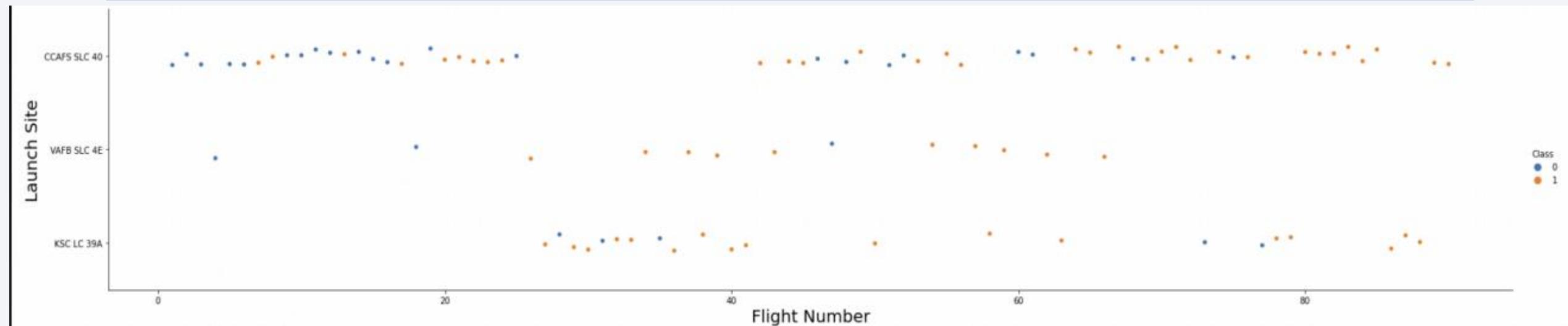
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

The background of the slide features a complex, abstract digital visualization. It consists of numerous thin, glowing lines that create a sense of depth and motion. The lines are primarily blue and red, with some green and purple highlights. They form a grid-like structure that curves and twists across the frame, resembling a three-dimensional space or a network of data points. The overall effect is futuristic and dynamic.

Section 2

## Insights drawn from EDA

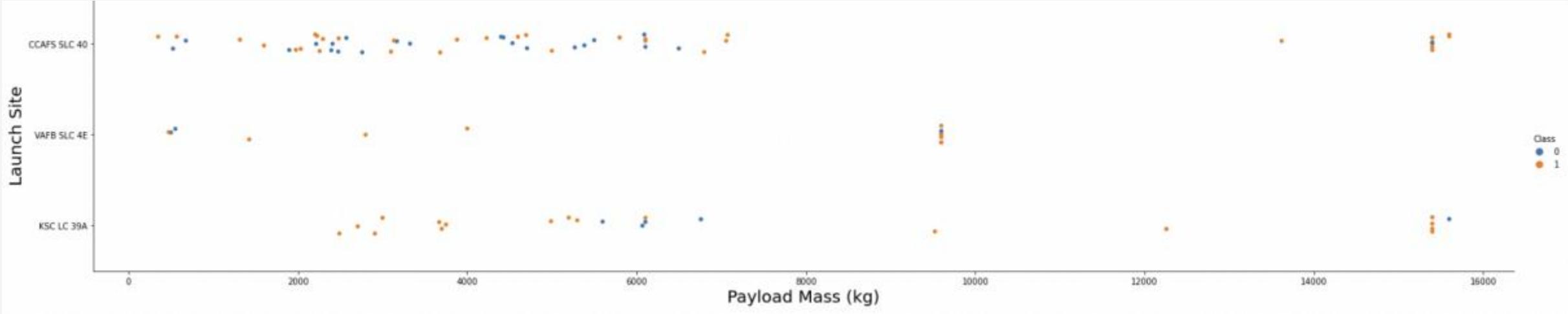
# Flight Number vs. Launch Site



## Key Observations

- Early missions experienced more failures, while more recent launches show consistently higher success rates.
- CCAFS SLC-40 accounts for approximately half of all launches.
- VAFB SLC-40 and KSC LC-39A demonstrate relatively higher landing success rates.
- Overall, the data suggests that launch success probability has improved over time, likely due to technological advancements and operational experience.

# Payload vs. Launch Site



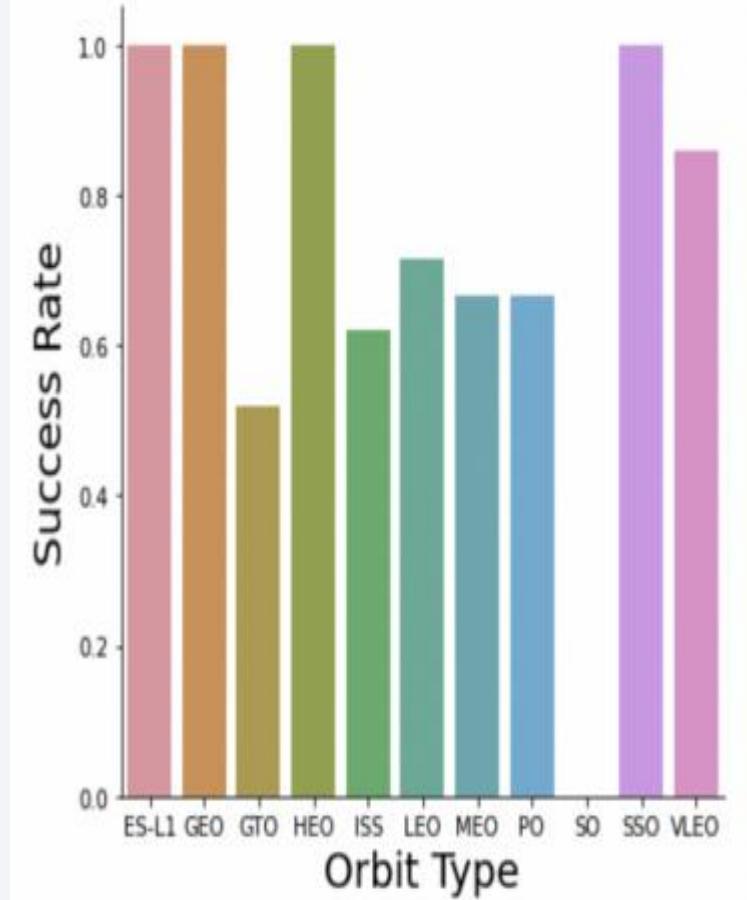
## Key Observations

- Across all launch sites, higher payload mass is generally associated with a higher landing success rate.
- Most missions carrying payloads are above 7000 kg resulted in successful landings.
- KSC LC-39A achieved a 100% success rate for missions with payload mass below 5500 kg.

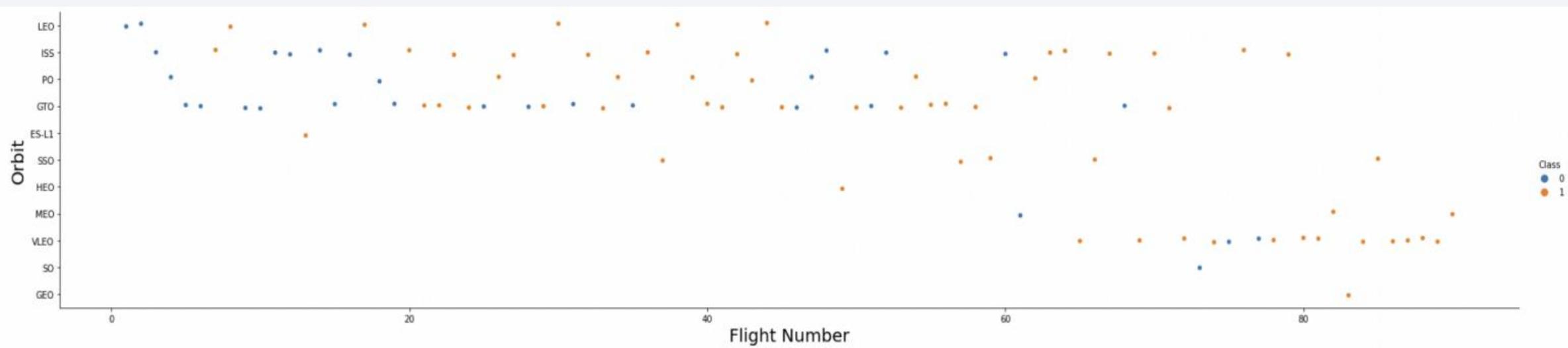
# Success Rate vs. Orbit Type

## Key Findings

- **Orbits with 100% success rate**
  - ES-L1,GEO,HEO, and SSO
- **Orbits with 0% success rate**
  - SO
- **Orbits with moderate success rates(50%-85%)**
  - GTO,ISS,LEO, MEO, and PO



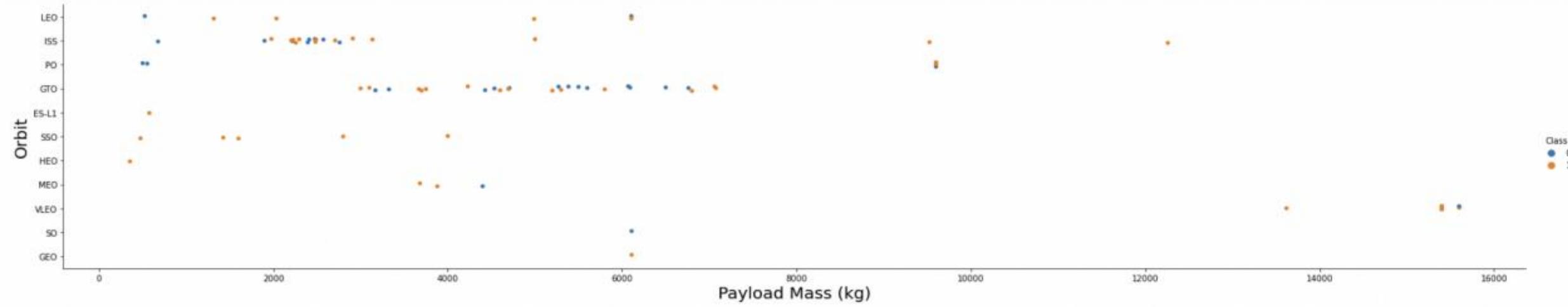
# Flight Number vs. Orbit Type



## Key Observations

- In the LEO orbit, landing success appears to increase with the number of flights, suggesting improvement over time.
- In contrast, for the GTO orbit, there is no clear relationship between flight number and landing success.

# Payload vs. Orbit Type



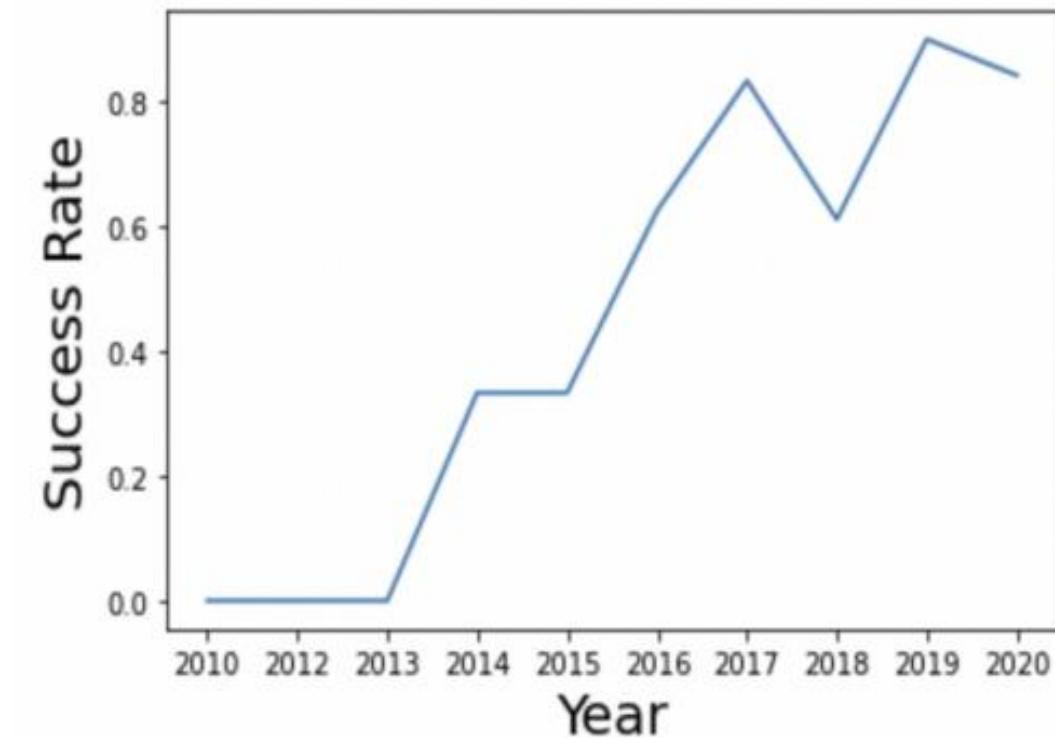
## Key Observations

- Heavy payloads appear to negatively impact landing success in **GTO** missions.
- In contrast, heavier payloads are associated with higher success rates **Polar LEO(ISS)** missions.

# Launch Success Yearly Trend

## Explanation:

- The landing success rate has steadily increased from 2013 to 2020, indicating continuous improvement in SpaceX's landing technology and operational reliability.



# All Launch Site Names

```
%sql select distinct launch_site from SPACEXDATASET;

* ibm_db_sa://wzf08322:****@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io901
08kqb1od81cg.databases.appdomain.cloud:31198/bludb
Done.
Out[4]:
launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E
```

## SQL Result- Unique Launch Sites

- Retrieved and displayed the names of all distinct launch sites involved in the missions.

# Launch Site Names Begin with 'CCA'

%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;										
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od81cg.databases.appdomain.cloud:31198/bludb Done.										
DATE	time_utc	booster_version	launch_site	payload	payload_mass_kg	orbit	customer	mission_outcome	land	reused
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure	Yes
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	Yes
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success		Yes
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success		Yes
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success		Yes

## Explanation

- Retrieved five records where the launch site name starts with “CCA” to analyze missions conducted at Cape Canaveral locations

# Total Payload Mass

```
%sql select sum(payload_mass_kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';

* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bl
edb
Done.

total_payload_mass

45596
```

## Explanation

- Calculated the total payload mass carried by boosters launched under NASA's CRS missions.

# Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass_kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';  
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od81cg.databases.appdomain.cloud:31198/bl  
|db  
)done.  
  
average_payload_mass  
2534
```

## Explanation

- Calculated the average payload mass carried by the F9 v1.1 booster version

# First Successful Ground Landing Date

---

```
%sql select min(date) as first_successful_landing from SPACEXDATASET where landing_outcome = 'Success (ground pad)';
```

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od81cg.databases.appdomain.cloud:31198/bl  
edb  
Done.
```

**first\_successful\_landing**

2015-12-22

## Explanation

- Identified the date of the first successful landing on a ground pad

# Successful Drone Ship Landing with Payload between 4000 and 6000

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 SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ b  
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od81cg.databases.appdomain.cloud:31198/bl  
idb  
done.  
  
booster_version  
F9 FT B1022  
F9 FT B1026  
F9 FT B1021.2  
F9 FT B1031.2
```

## Explanation

- Retrieved the names of boosters that successfully landed on a drone ship with payload masses between 4000 kg and 6000 kg.

# Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bl
  db
  Done.
```

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

## Explanation

- Calculated the total number of successful and failed mission outcomes.

# Boosters Carried Maximum Payload

## Explanation

- Identified the booster versions that carried the maximum payload mass.

### **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

```
%sql select monthname(date) as month, date, booster_version, launch_site, landing_outcome from SPACEXDATASET  
where landing_outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od81cg.databases.appdomain.cloud:31198/bl  
edb  
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)

## Explanation

- Retrieved failed drone ship landing outcomes in 2015, including their booster versions and corresponding launch sites.

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

```
: %%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
  where date between '2010-06-04' and '2017-03-20'
    group by landing__outcome
    order by count_outcomes desc;

* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od81cg.databases.appdomain.cloud:31198/bl
edb
Done.
```

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

## Explanation

- Ranked the counts of landing outcomes (e.g., Failure – drone ship, Success – ground pad) between June 4, 2010 and March 20, 2017 in descending order.

The background of the slide is a photograph taken from space at night. It shows the curvature of the Earth against a dark blue-black void of space. City lights are visible as numerous small white and yellow dots, primarily concentrated in the lower right quadrant where the United States appears. In the upper left quadrant, the green and yellow glow of the aurora borealis is visible in the upper atmosphere.

Section 3

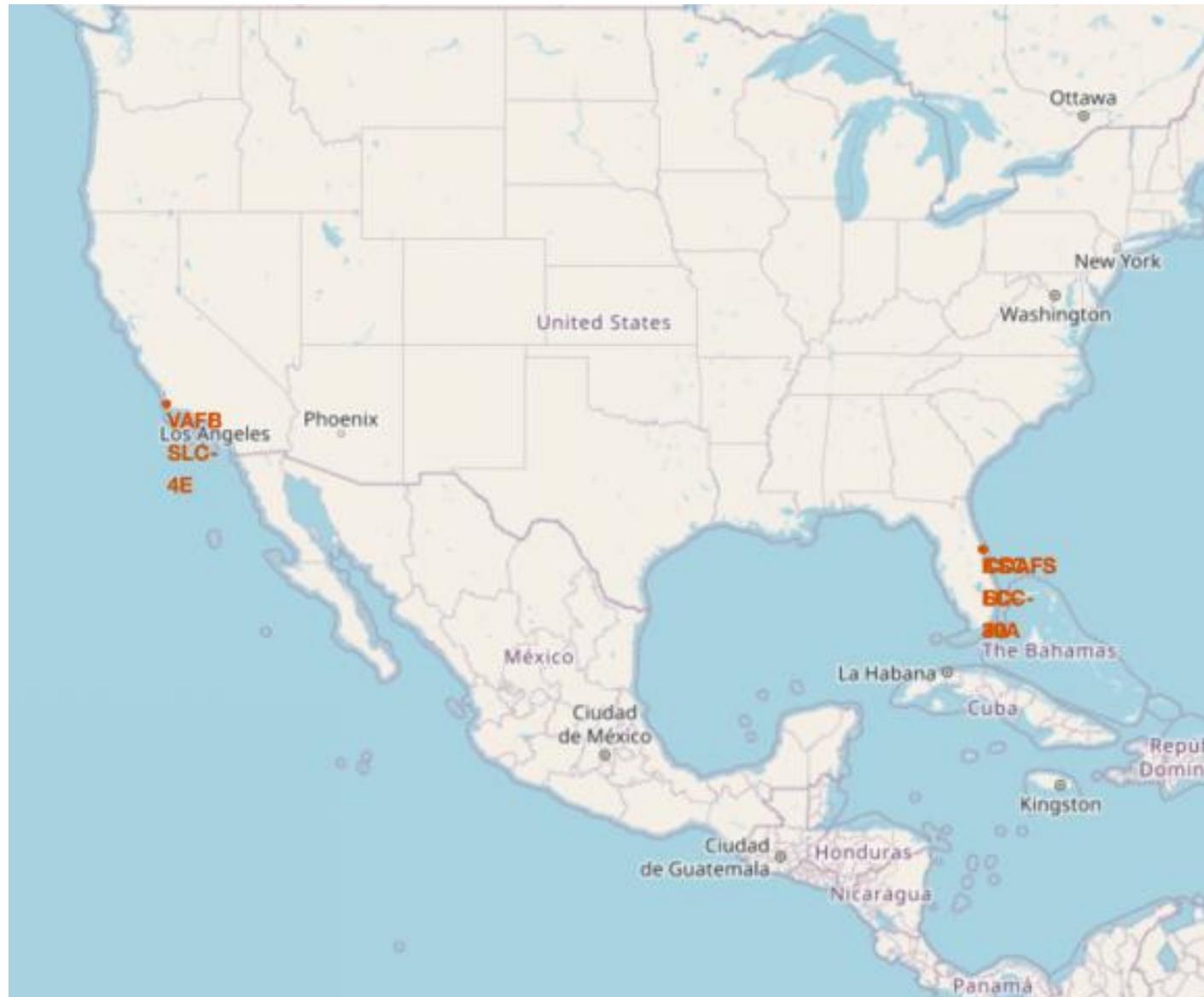
# Launch Sites Proximities Analysis

## All launch sites' location markers on a global map

---

### Geographic Location Insights

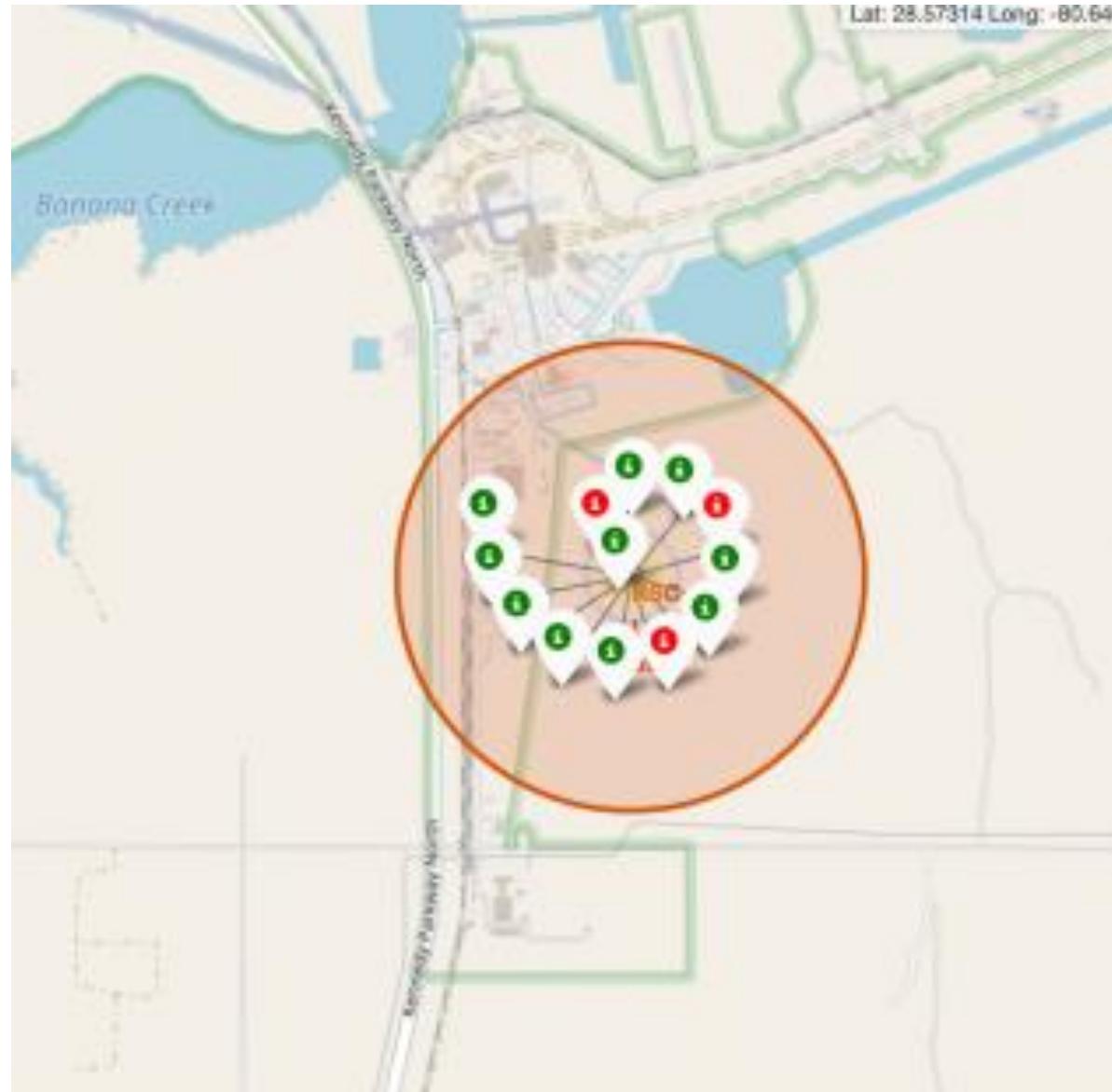
- Most launch sites are located near the **Equator**, where the Earth's rotational speed is highest(approximately 1670 km/h). Launching from near the Equator provides additional velocity due to inertia, helping spacecraft reach and maintain orbit more efficiently.
- All launch sites are positioned close the coastline. Launching rockets over the oceans reduces safety risks by minimizing the chance of debris falling over populated areas.



## Colour-labeled launch records on the map

### Launch Outcome Visualization

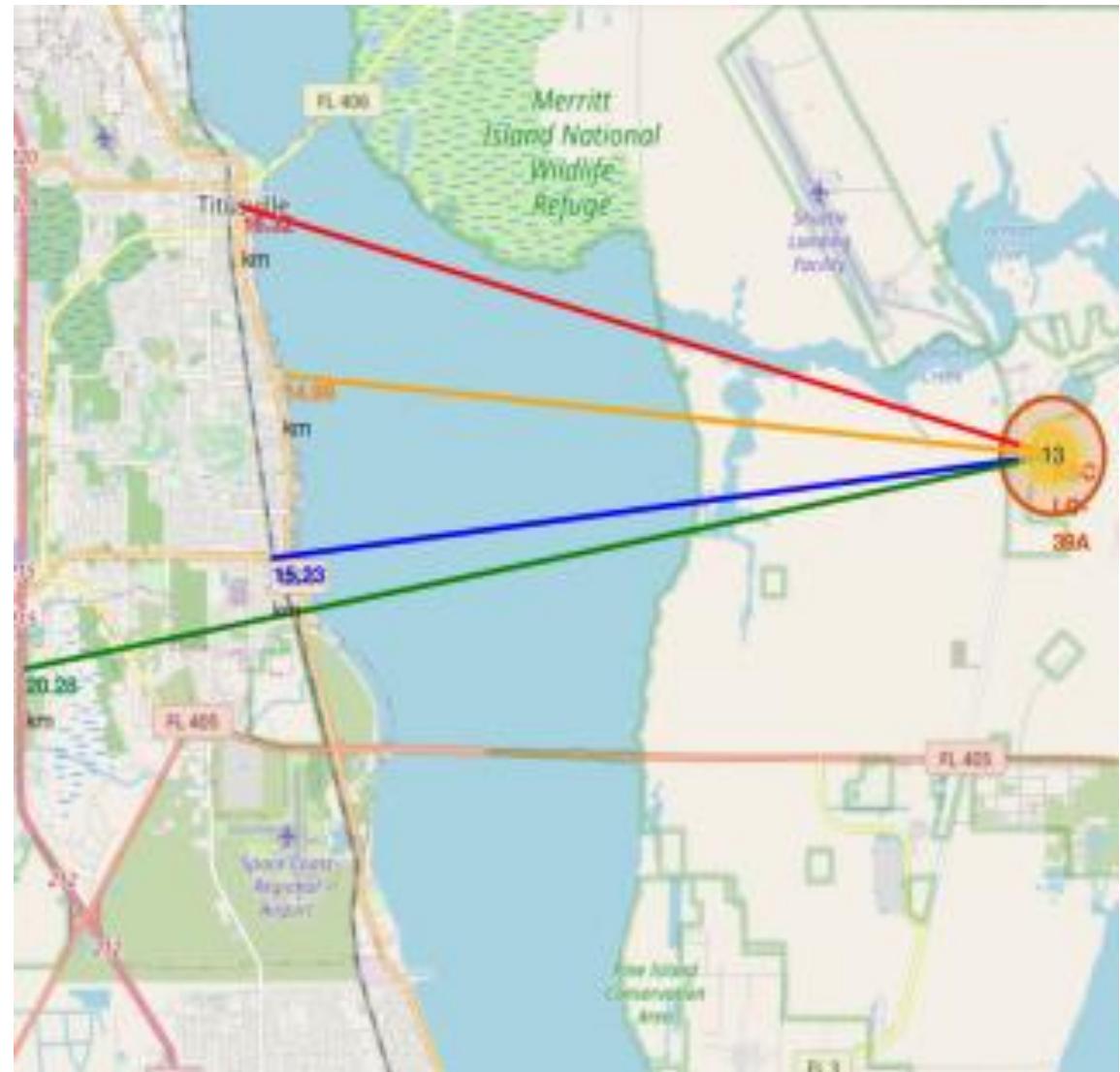
- Color-coded markers help quickly identify launch site performance
  - **Green markers:** represent successful launches
  - **Red markers:** represent failed launches
- Based on the visualization, **KSC LC-39A** demonstrates a notably high landing success rate compared to other launch sites.



## Distance from the launch site KSC LC-39A to its proximities

### Proximity Analysis- KSC LC-39A

- The launch site **KSC LC-39A** is located relatively close to key infrastructures
  - Railway: 15.23 km
  - Highway: 20.28 km
  - Coastline: 14.99 km
  - Nearest city(Titusville): 16.32 km
- Given the high velocity of rockets, a failed launch could travel 15-20 km within seconds, posing potential risks to nearly populated or infrastructural areas.



Section 4

# Build a Dashboard with Plotly Dash



# Launch success count for all sites

Total Success Launches by Site



## Dashboard Insight:

- The chart clearly indicates that **KSC LC-39A** has the highest number of successful launches among all launch sites.

# Launch site with highest launch success ratio

Total Success Launches for Site KSC LC-39A



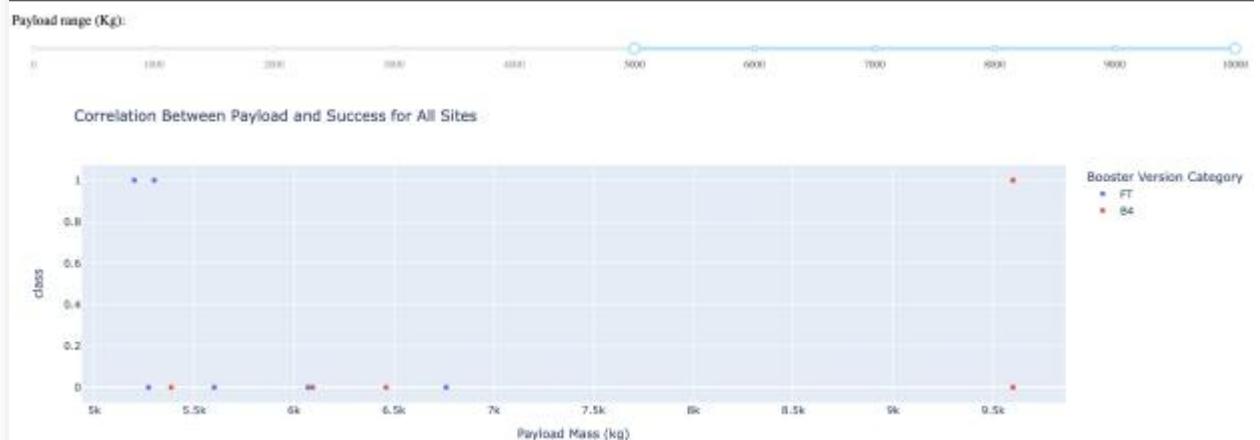
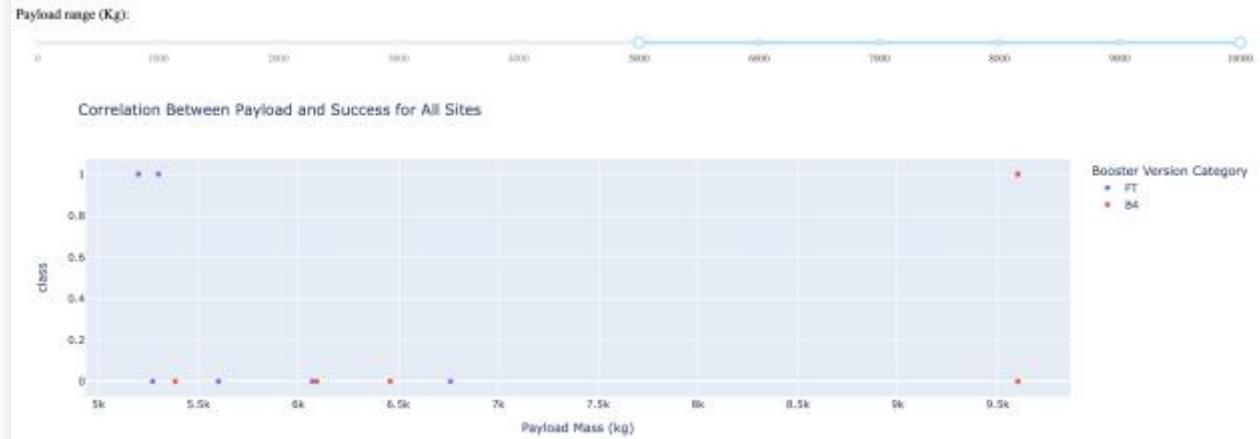
## Launch Success Rate –KSC LC-39A

- **KSC LC-39A** achieved the highest launch success rate at 76.9%, with 10 successful landings and only 3 failures.

# Payload Mass vs Launch Outcome for all sites

## Payload Mass vs Success Rate

- The charts indicate that missions carrying payloads between 2000 kg and 5500 kg have the highest landing success rate.



Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

---

## Model Performance Evaluation

- Based on the test set results, it was not possible to clearly determine the best-performing model, as all methods achieved similar accuracy scores.
- The identical test scores may be due to the small test sample size (18 samples). Therefore, all models were further evaluated using the full dataset.
- The evaluation on the complete dataset indicates that the Decision Tree model performs best, achieving the highest overall score and accuracy among all models.

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

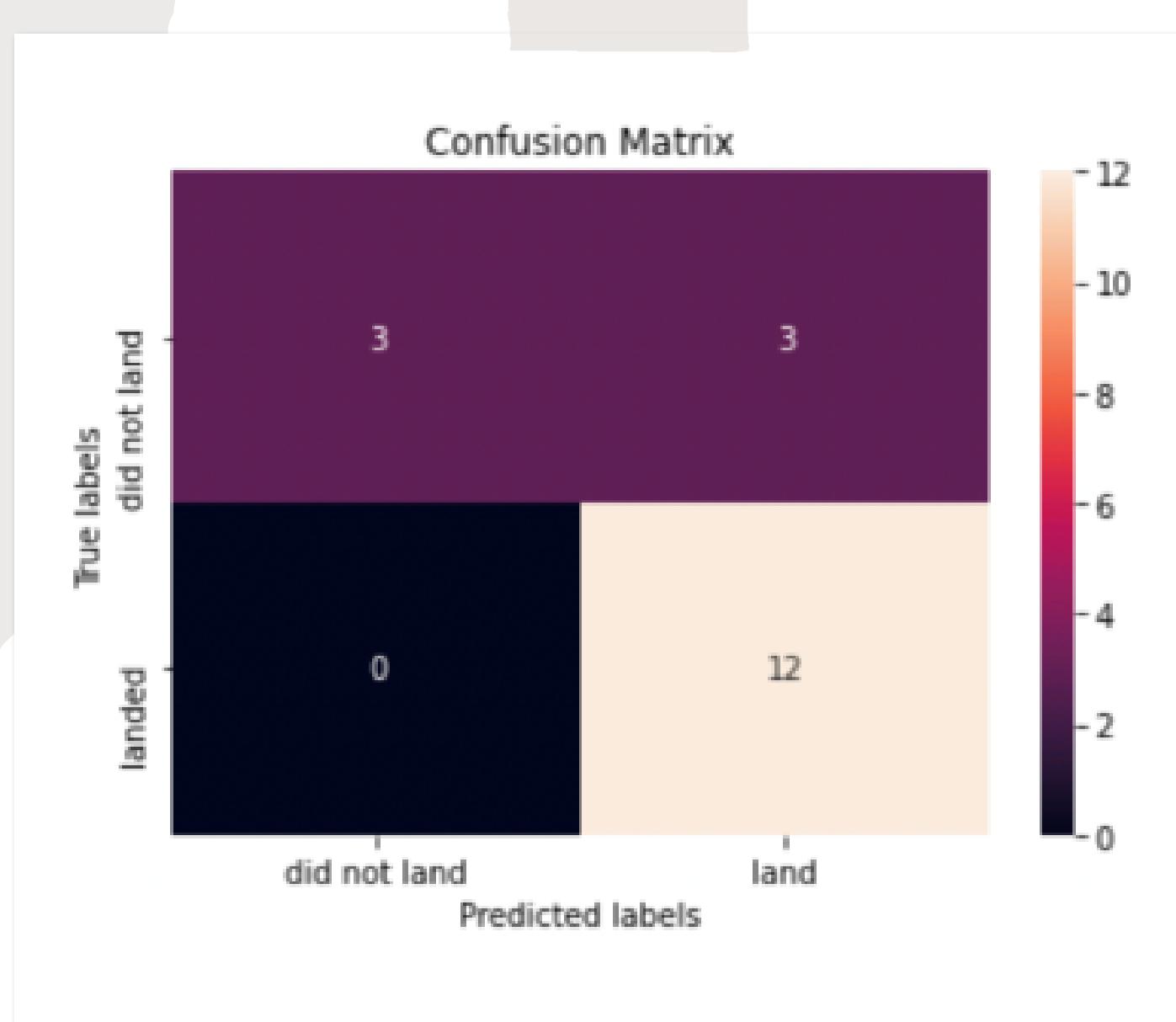
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

# Confusion Matrix

## Confusion Matrix Analysis – Logistic Regression

- The confusion matrix shows that Logistic Regression is able to differentiate between successful and failed landings.
- However, the primary issue observed is a relatively high number of false positives, where unsuccessful landings are incorrectly predicted as successful.

		Predicted Values	
		Negative	Positive
Actual Values	Negative	TN	FP
	Positive	FN	TP



# Conclusions

---

- The Decision Tree model achieved the best overall performance for this dataset.
- Launches with lower payload mass generally demonstrate higher landing success rates compared to heavier payloads.
- Most launch sites are located near the Equator and close to the coastline, optimizing launch efficiency and safety.
- The overall landing success rate has improved over time, indicating technological and operational advancements.
- **KSC LC-39A** has the highest launch success rate among all sites.
- Orbit types **ES-L1, GEO, HEO, and SSO** show a 100% landing success rate.

# Appendix

---

## Special Thanks to

Instructors

Coursera

IBM

Thank you!

