IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Prakhar Agrawal
22/01/2026

OUTLINE

EXECUTIVE SUMMARY

SUMMARY OF METHODOLOGIES

- **Data Collection:** Leveraged SpaceX API and Web Scraping (BeautifulSoup).
- **Data Wrangling:** Cleaned 26 missing landingPad entries; 80 total features.
- **EDA:** SQL (Db2/SQLite) and Visualization (Matplotlib/Seaborn).
- **Interactive Viz:** Folium (MarkerClusters) and Plotly Dash.
- **Predictive Modeling:** GridSearchCV for LogReg, SVM, Tree, and KNN.

SUMMARY OF ALL RESULTS

- **Site Performance:** CCAFS SLC 40 leading with 55 records.
- **Landing Success:** 66.7% success rate; 31 Drone Ship (ASDS) landings.
- **Model Accuracy:** Consistent test accuracy of 83.33% across models.
- **Optimal Parameters:** SVM 'sigmoid' kernel identified as top performer.
- **Conclusion:** Flight numbers and payload ranges are key success predictors.

INTRODUCTION

PROJECT BACKGROUND & CONTEXT

Industry Revolution: SpaceX disrupted the market by prioritizing rocket reusability to reach multi-planetary goals.

The Falcon 9 Rocket: The workhorse of modern missions, being the first orbital-class rocket capable of re-flight.

Cost Efficiency: Reusing first-stage boosters reduces launch costs from $150M to ~$62M, creating a competitive edge.

PROJECT OBJECTIVES

Landing Prediction: Build a machine learning pipeline to predict successful first-stage landings.

Commercial Intelligence: Help competitors estimate "real" costs to bid effectively in the commercial market.

Data Exploration: Identify how site, orbit, payload, and version impact booster recovery likelihood.

Section 1

Methodology

METHODOLOGY

Data Collection Methodology

Using **SpaceX Rest API** and **Web Scraping** from Wikipedia.

Perform Data Wrangling

Filtering data, handling missing values, and applying **One Hot Encoding**.

Exploratory Data Analysis (EDA)

Utilizing **Visualization** tools and **SQL** queries.

Interactive Visual Analytics

Mapping with **Folium** and building dashboards with **Plotly Dash**.

Predictive Analysis

Building, tuning, and evaluating **Classification Models** for optimal accuracy.

DATA COLLECTION

COLLECTING THE DATASETS

A multi-faceted approach was used to gather the necessary data, combining real-time telemetry from the **SpaceX REST API** and historical mission records via **Web Scraping** from Wikipedia.

The collection process ensures a robust feature set by **cross-referencing API technical specifications** with historical mission descriptions and outcomes.

Post-collection, the data undergoes **Filtering** to isolate Falcon 9 launches and **Wrangling** to prepare for predictive modeling.

DATA COLLECTION – SPACEX API

Using the SpaceX API, we performed **REST API Calls** to the launches endpoint. The complex **JSON** structure was processed into a tabular format for analysis.

Key methodologies included **JSON Normalization** and the use of custom helper functions to extract rocket names, launch site locations, and core telemetry data from their respective API endpoints.

View SpaceX API Collection Notebook

DATA COLLECTION - SCRAPING

For historical launch data, we leveraged **BeautifulSoup** to parse Wikipedia's launch tables, identifying mission-specific details like booster versions and payload customers.

The **Web Scraping** process involved targeted **HTML Table Extraction**, followed by string cleaning to handle units (kg/lbs) and formatting mission outcomes for classification.

View Web Scraping Notebook

DATA WRANGLING

Raw mission outcomes were processed to establish binary classification targets. We analyzed landings to define the **Training Labels** necessary for supervised machine learning.

CATEGORICAL MAPPING:

- **Successful (Label=1):** Includes landings on drone ships (ASDS), ground pads (RTLS), and specific ocean regions.
- **Unsuccessful (Label=0):** Includes crashes, landing failures, or missions where no attempt was made.

INSIGHTS SUMMARY:

Wrangling addressed **26 null values** in the landingPad feature and ensured PayloadMass was complete. One-Hot Encoding then prepared **80 features** for predictive modeling.

GitHub URL: Data Wrangling

EDA WITH DATA VISUALIZATION

Exploratory Data Analysis utilized visual plotting to identify key relationships between mission variables and the likelihood of a successful booster landing.

Scatter Plots: Analyzed **Flight No. vs. Launch Site** and **Payload vs. Launch Site** to detect success clusters.

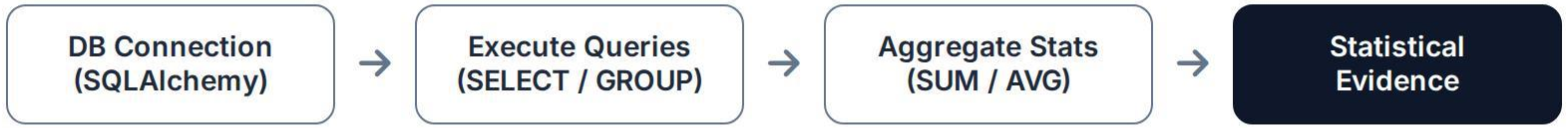
Bar Charts: Evaluated **Success Rates by Orbit** to identify the most efficient landing trajectories (SSO, HEO, etc).

Line Charts: Tracked the **Yearly Success Trend** to demonstrate improved reliability across mission versions.

GitHub URL: EDA with Data Visualization

EDA WITH SQL

Advanced data querying was performed using SQL to extract specific mission metrics and perform deep-dive analysis on landing outcomes and payload distributions.



MISSION METADATA:

- Identified **unique launch sites** (CCAFS, KSC, VAFB).
- Calculated total **NASA CRS Payload** mass carried by boosters.
- Analyzed average payload mass for the **Falcon 9 v1.1** booster version.

OUTCOME ANALYTICS:

- Extracted date of the **first successful ground landing**.
- Ranked boosters by **maximum payload mass** carried.
- Ranked counts of landing outcomes between **2010-06-04 and 2017-03-20**.

GitHub URL: EDA with SQL

BUILD AN INTERACTIVE MAP WITH FOLIUM

Integrated multiple interactive geospatial objects into a single map instance to evaluate logistical proximities and reusability outcomes across all SpaceX launch sites.

MARKERS & OUTCOME MAPPING:

- Added Markers with Circle, Popup, and Text Labels for NASA JSC and all launch sites to show geographical locations and coastal proximity.
- Used coloured Markers to distinguish outcomes: **Green** for success and **Red** for failed launches using Marker Clusters to identify high-success sites.

PROXIMITY & LOGISTICS:

- Added coloured PolyLines to measure distances between the **KSC LC-39A** site and logistical infrastructure.
- Analyzed distances to **Railway, Highway, Coastline, and Closest City** to understand the geographical constraints impacting mission safety and efficiency.

GitHub URL: Interactive Visual Analytics with Folium

BUILD A DASHBOARD WITH PLOTLY DASH

Developed a web-based interface for dynamic exploration of mission datasets, enabling stakeholders to filter performance metrics in real-time.

INTERACTIONS:

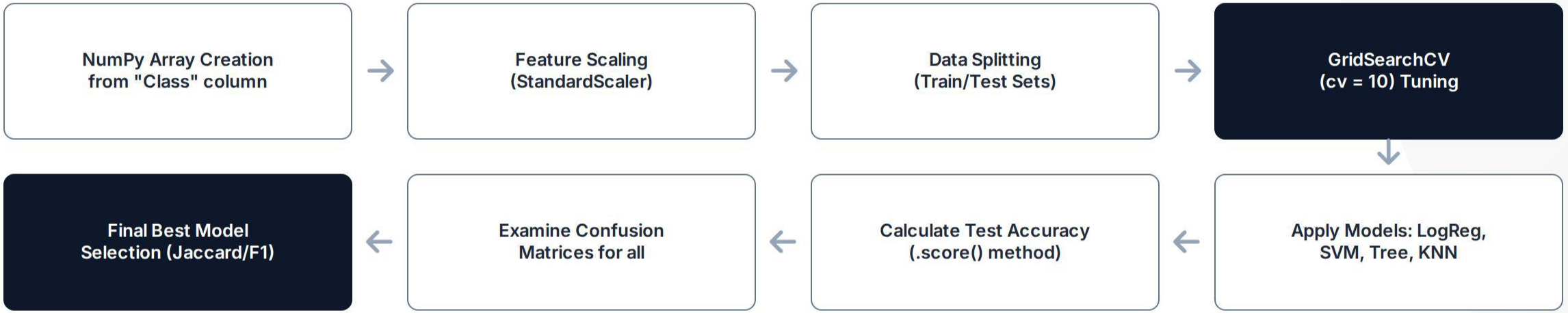
- **Dropdown List:** To enable specific Launch Site selection.
- **Payload Slider:** To filter missions by mass range.

VISUAL ANALYTICS:

- **Pie Charts:** Success vs. Failure counts for site-specific or global views.
- **Scatter Charts:** Payload Mass vs. Success Rate correlation by Booster Version.

GitHub URL: SpaceX Dash App

PREDICTIVE ANALYSIS (CLASSIFICATION)

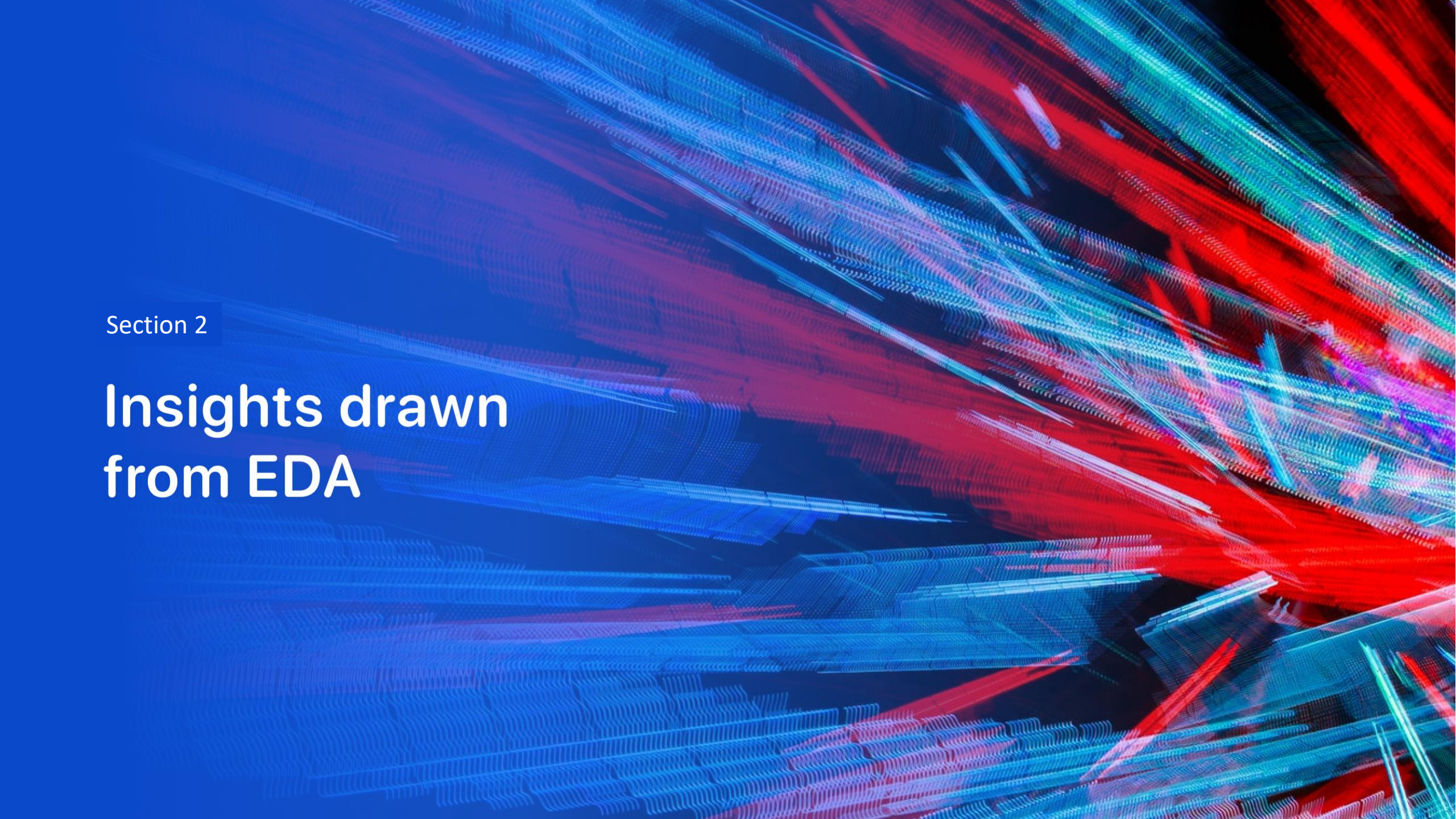


Summary: Built and evaluated four classification models to find the best performing algorithm. Optimized for reusability prediction by tuning hyperparameters to maximize F1 and Jaccard scores.

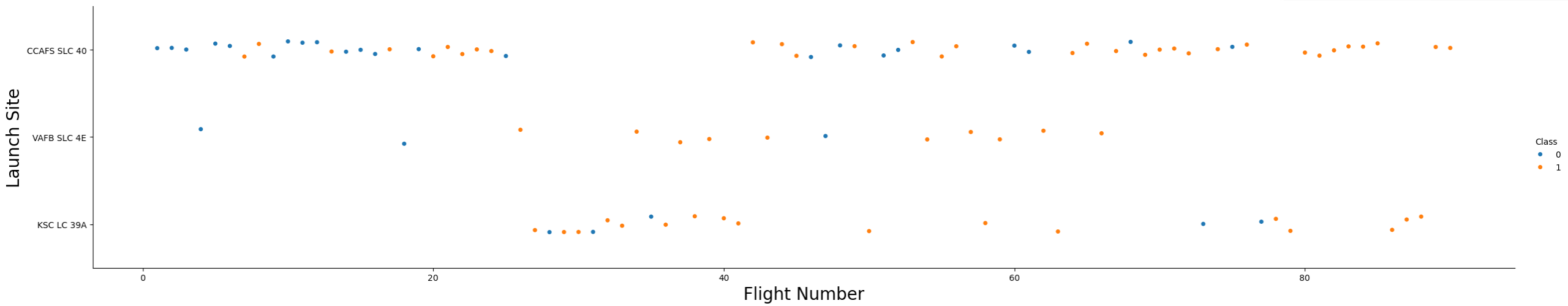
GitHub URL: Machine Learning Prediction

RESULTS

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


Section 2
Insights drawn from EDA

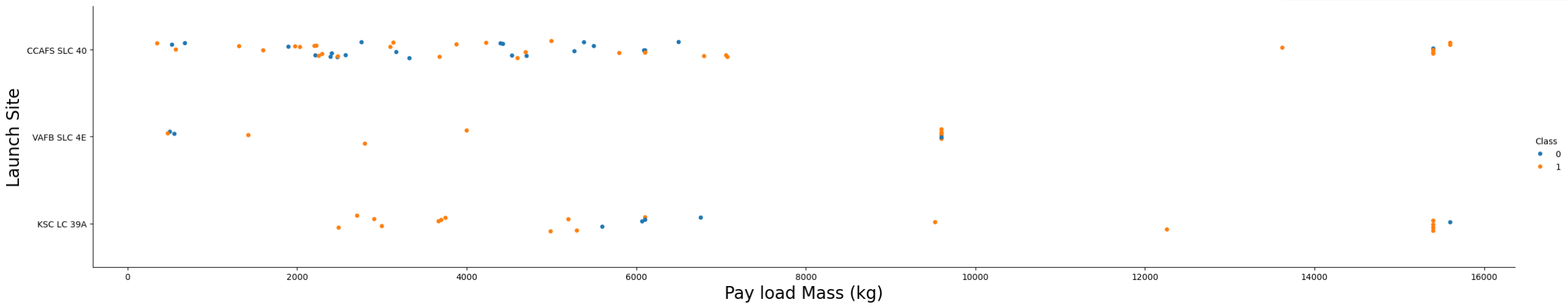
FLIGHT NUMBER VS. LAUNCH SITE



EXPLANATION:

- The earliest flights all failed while the latest flights all succeeded.
- The **CCAFS SLC 40** launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

PAYLOAD VS. LAUNCH SITE



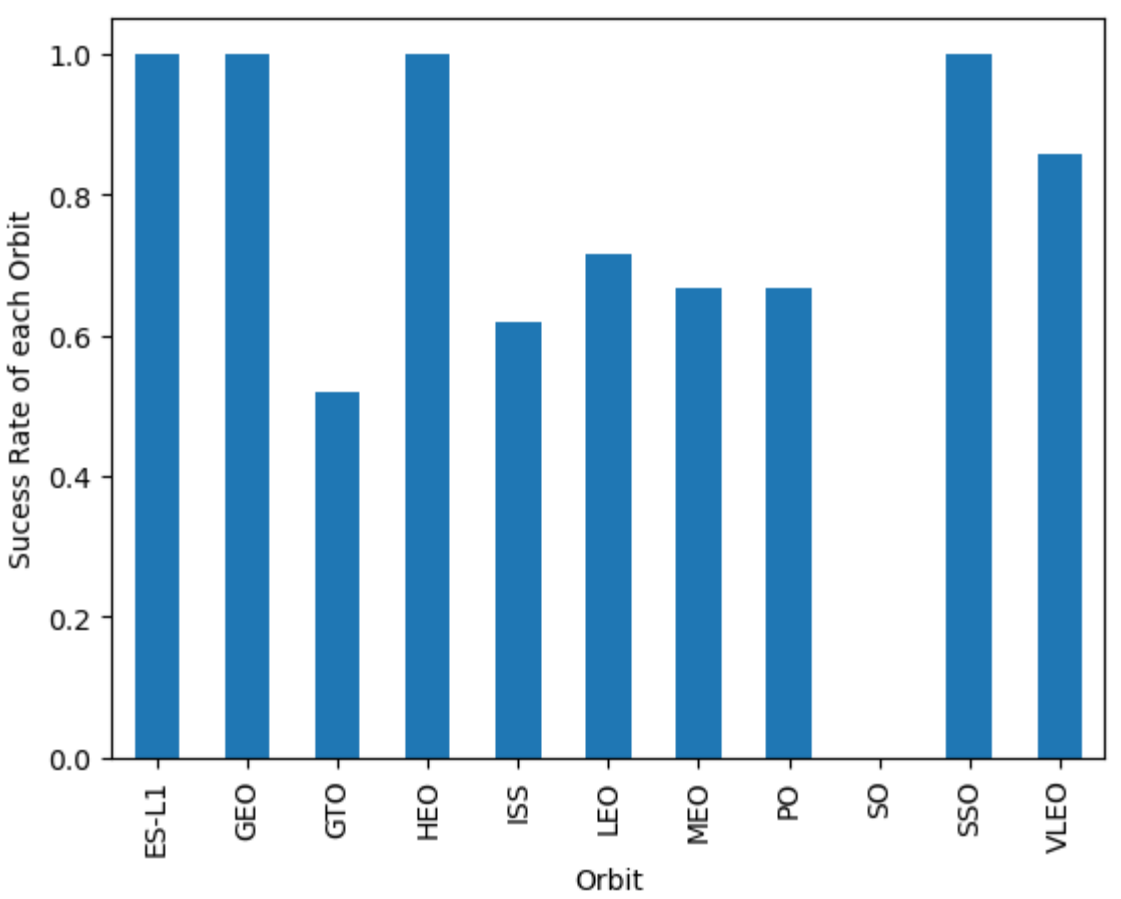
EXPLANATION:

- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass **over 7000 kg** were successful.
- **KSC LC 39A** has a 100% success rate for payload mass under 5500 kg too.
- Payload capacity varies significantly across the three primary launch sites.

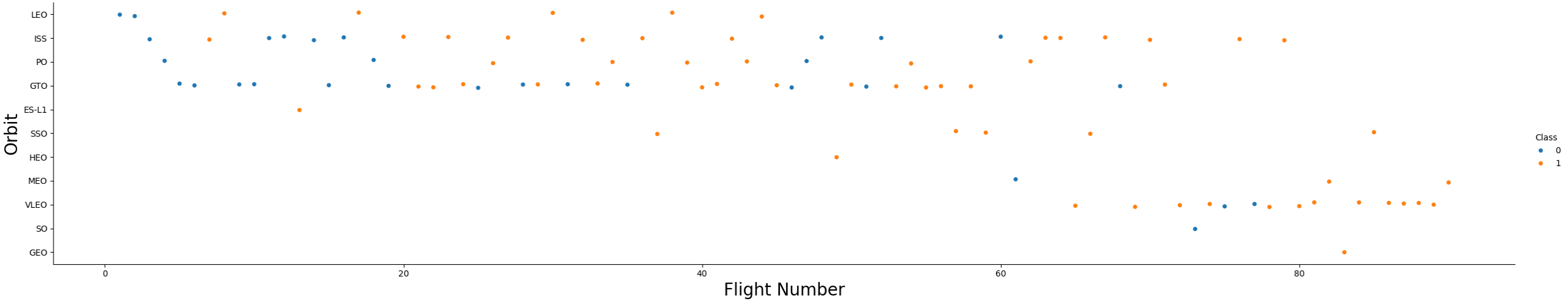
SUCCESS RATE VS. ORBIT TYPE

EXPLANATION:

- **100% Success Rate:**
  - ES-L1, GEO, HEO, SSO
- **0% Success Rate:**
  - SO
- **50% - 85% Success Rate:**
  - GTO, ISS, LEO, MEO, PO



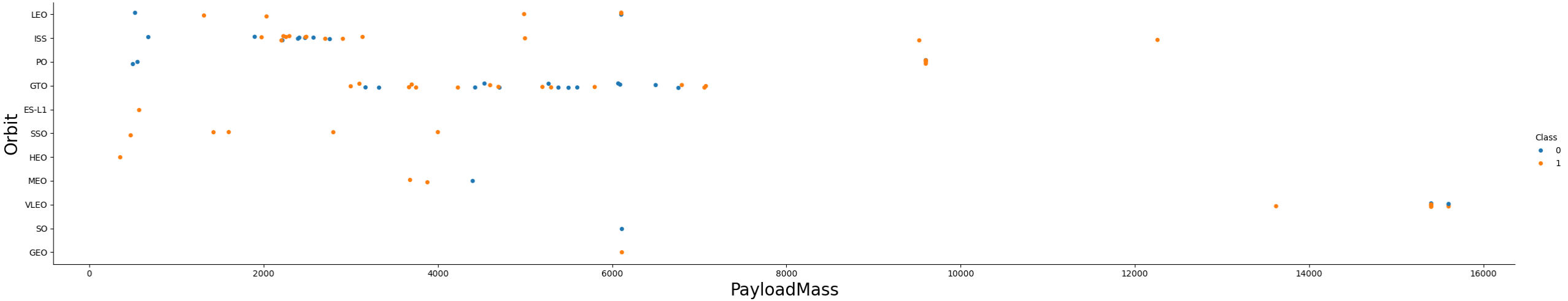
FLIGHT NUMBER VS. ORBIT TYPE



EXPLANATION:

- In the **LEO orbit**, success appears positively related to the number of flights.
- In contrast, there seems to be no relationship between flight number when in **GTO orbit**.

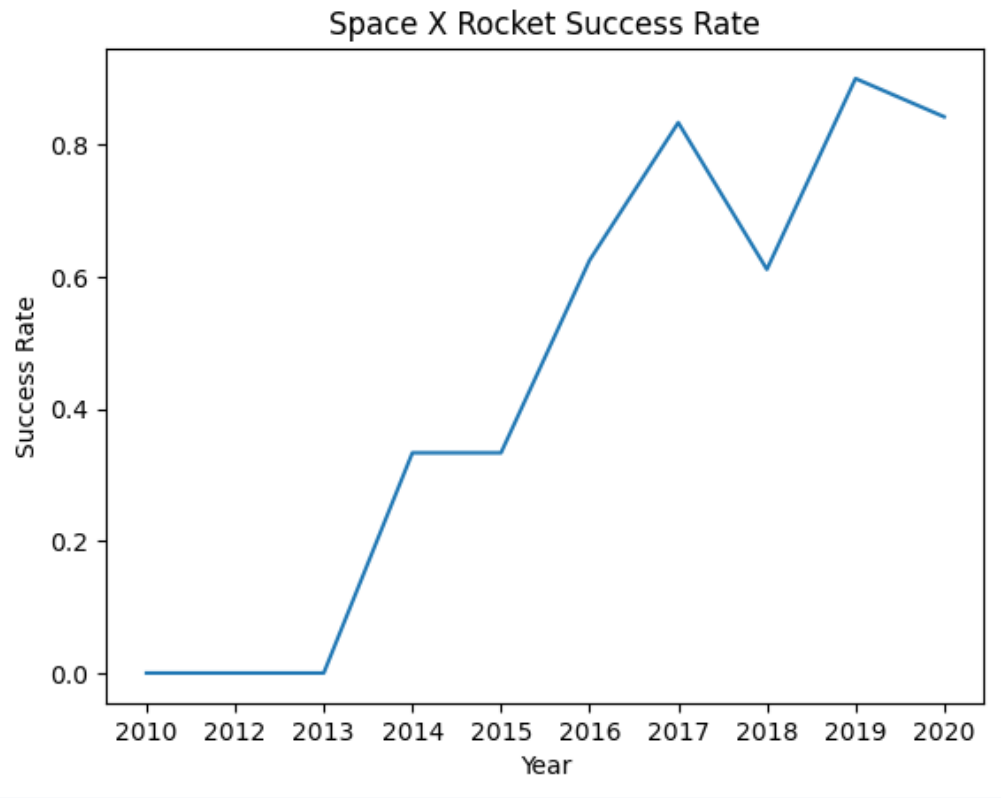
PAYLOAD MASS VS. ORBIT TYPE



EXPLANATION:

- Heavy payloads have a **negative influence** on GTO orbits.
- Heavy payloads have a **positive influence** on GTO and Polar LEO (ISS) orbits.

LAUNCH SUCCESS YEARLY TREND



EXPLANATION:

- Since **2013**, we can see a consistent increase in the SpaceX Rocket success rate.
- The data demonstrates technological maturation and improved booster recovery reliability over time.

ALL LAUNCH SITE NAMES

```
In [10]: %sql SELECT DISTINCT "LAUNCH_SITE" FROM SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

Out[10]:

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

EXPLANATION:

The use of **DISTINCT** in the query allows to remove duplicate LAUNCH_SITE, providing a clean list of unique operational pads.

LAUNCH SITE NAMES BEGIN WITH 'CCA'

In [11]:
```
%sql SELECT * FROM SPACEXTBL WHERE "LAUNCH_SITE" LIKE '%CCA%' LIMIT 5
```

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

Out[11]:

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

EXPLANATION:

The **WHERE** clause followed by **LIKE** clause filters launch sites that contain the substring CCA. LIMIT 5 shows the first 5 records from the filtered results.

TOTAL PAYLOAD MASS

```
In [12]: %sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "CUSTOMER" = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
Out[12]: SUM("PAYLOAD_MASS__KG_")
                            45596
```

EXPLANATION:

This query utilizes the **SUM()** function to return the total aggregate payload mass carried on missions where the customer is identified as **NASA (CRS)**.

AVERAGE PAYLOAD MASS BY F9 V1.1

```
In [13]: %sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "BOOSTER_VERSION" LIKE '%F9 v1.1%'
 * sqlite:///my_data1.db
Done.
Out[13]: AVG("PAYLOAD_MASS__KG_")
         2534.6666666666665
```

EXPLANATION:

This query returns the average of all payload masses where the booster version contains the substring **F9 v1.1**, used to benchmark performance of that specific version.

FIRST SUCCESSFUL GROUND LANDING DATE

In [19]:
```
%sql SELECT MIN("DATE") FROM SPACEXTBL WHERE "Landing_Outcome" LIKE '%Success%'
```

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

Out[19]:

MIN("DATE")
2015-12-22

EXPLANATION:

With this query, we select the oldest successful landing. The WHERE clause filters dataset in order to keep only records where landing was successful. With the **MIN function**, we select the record with the oldest date.

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

```
1  %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.
```

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

EXPLANATION:

This query returns the booster version where landing was successful and payload mass is between **4000 and 6000 kg**. The WHERE and AND clauses filter the dataset.

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

```
%%sql SELECT \
    (SELECT COUNT(MISSION_OUTCOME) \
     FROM SPACEXTBL \
     WHERE MISSION_OUTCOME LIKE '%Success%') AS SUCCESS, \
    (SELECT COUNT(MISSION_OUTCOME) \
     FROM SPACEXTBL \
     WHERE MISSION_OUTCOME LIKE '%Failure%') AS FAILURE

```

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

SUCCESS	FAILURE
100	1

EXPLANATION:

With the first SELECT, we show the subqueries that return results. The first subquery counts the successful mission. The second subquery counts the unsuccessful mission. The WHERE clause followed by **LIKE clause** filters mission outcome. The **COUNT function** counts records filtered.

BOOSTERS CARRIED MAXIMUM PAYLOAD

```
1
2  %sql SELECT DISTINCT "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

EXPLANATION:

We used a subquery to filter data by returning only the heaviest payload mass with **MAX function**. The main query uses subquery results and returns unique booster version (**SELECT DISTINCT**) with the heaviest payload mass.

2015 LAUNCH RECORDS

```
1  %sql SELECT substr(Date,6,2) AS Month, Booster_Version, Launch_Site FROM SPACEXTBL WHERE Landing_Outcome='Failure (drone ship)' AND substr(Date,1,4)='2015'
2
```

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

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

EXPLANATION:

This query returns month, booster version, launch site where landing was unsuccessful and landing date took place in 2015. Substr function process date in order to take month or year. **Substr(DATE, 4, 2)** shows month. **Substr(DATE, 7, 4)** shows year.

RANK LANDING OUTCOMES BETWEEN 2010-06-04 AND 2017-03-20

```
1  %sql SELECT Landing_Outcome, COUNT(*) AS Outcome_Count FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Outcome_Count DESC;
2
```

```
 * 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

EXPLANATION:

This query returns landing outcomes and their count where mission was successful and date is between 04/06/2010 and 20/03/2017. The GROUP BY clause groups results by landing outcome and ORDER BY COUNT DESC shows results in decreasing order.

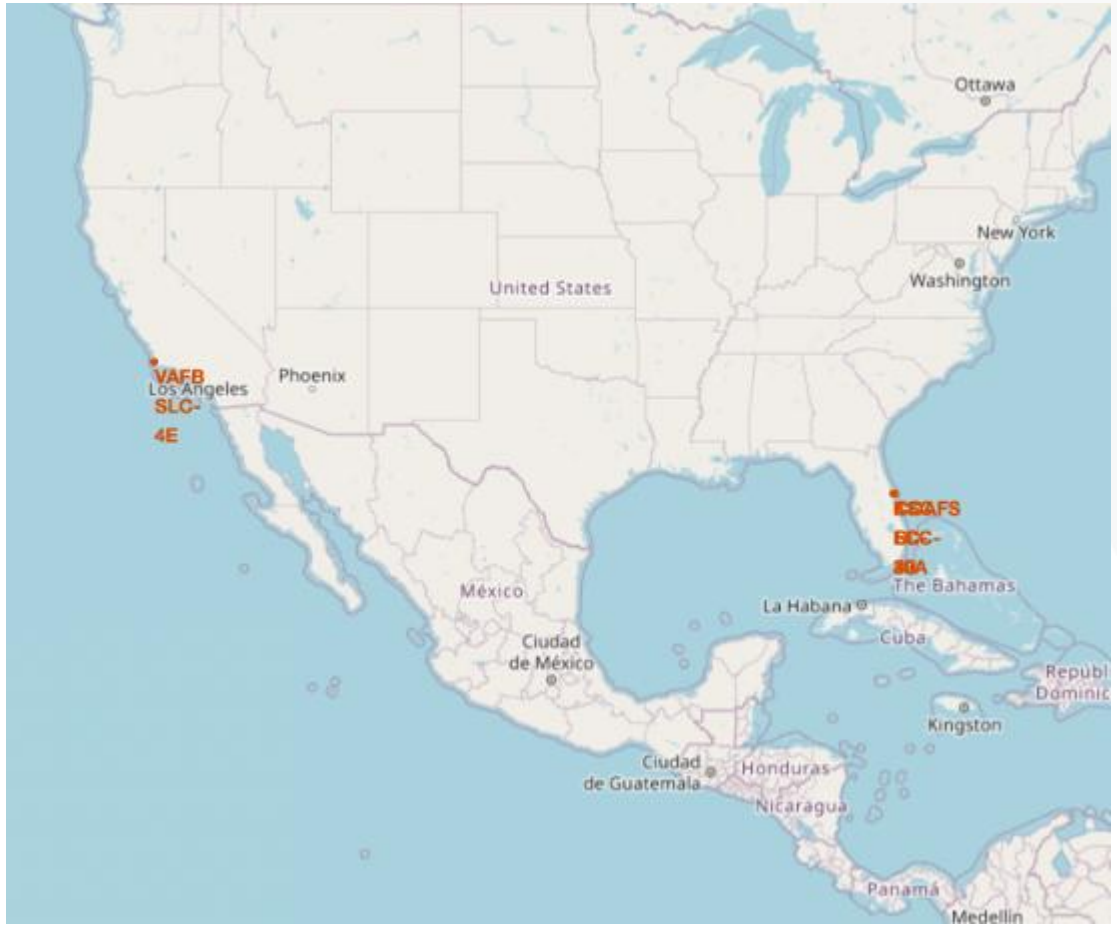
Section 3

Launch Sites Proximities Analysis

LAUNCH SITES MARKERS ON A GLOBAL MAP

EXPLANATION:

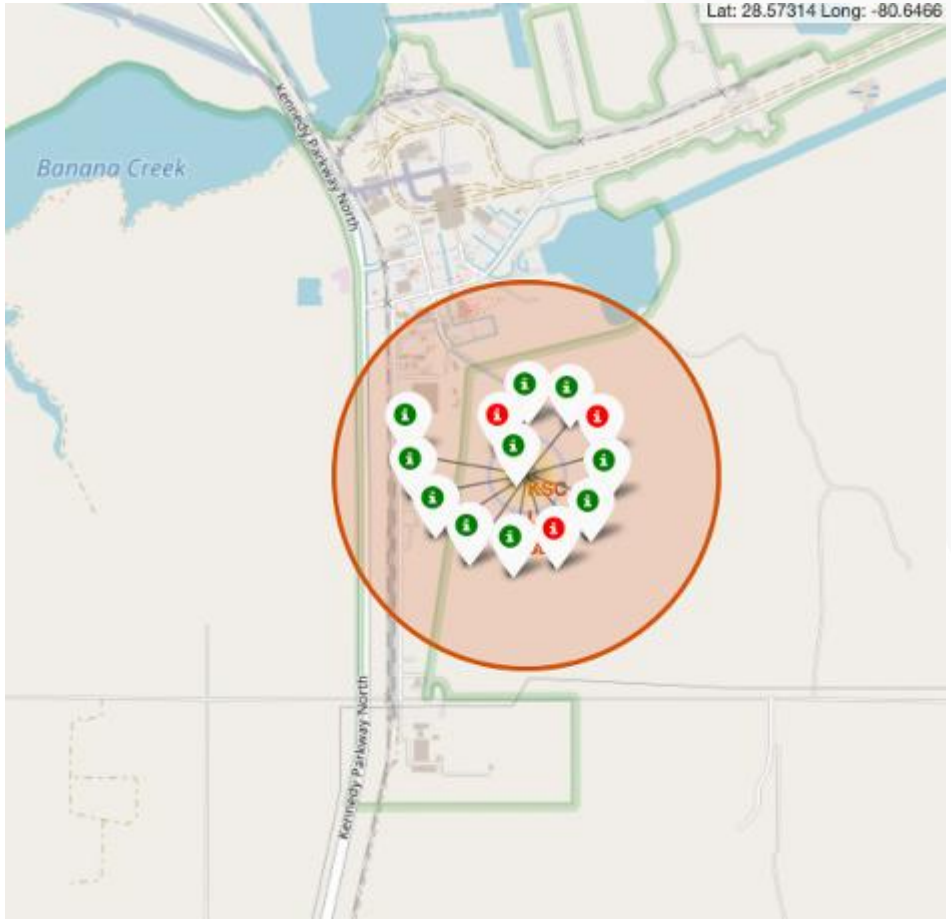
- Launch sites are positioned near the **Equator** (moving at 1670 km/hour).
- Inertia assists the spacecraft in reaching orbital velocity more efficiently.
- Coastal proximity minimizes risk by ensuring launch trajectories and potential debris fall over the ocean rather than populated land.



COLOUR-LABELED LAUNCH RECORDS ON THE MAP

EXPLANATION:

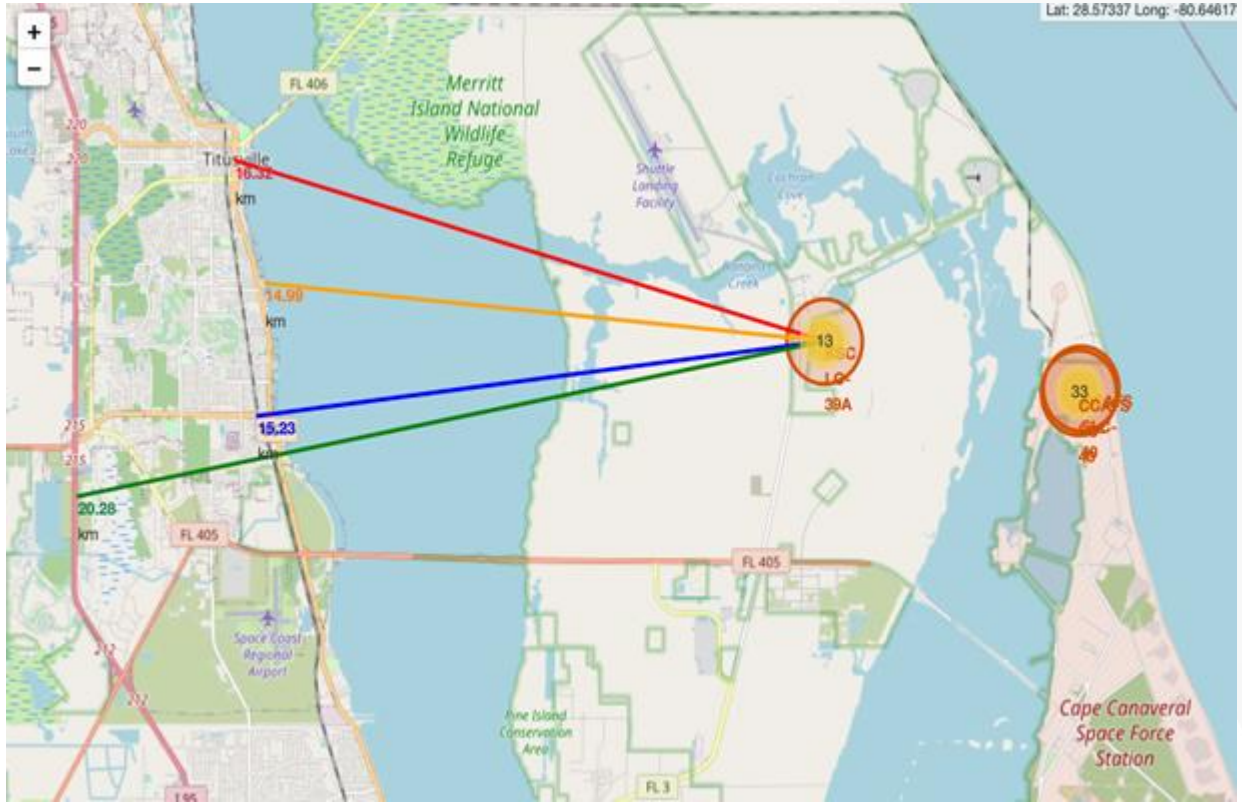
- Colour-labeled markers help easily identify launch site performance:
  - **Green Marker** = Successful Launch
  - **Red Marker** = Failed Launch
- Geospatial clustering indicates that launch site **KSC LC-39A** has a very high success rate.

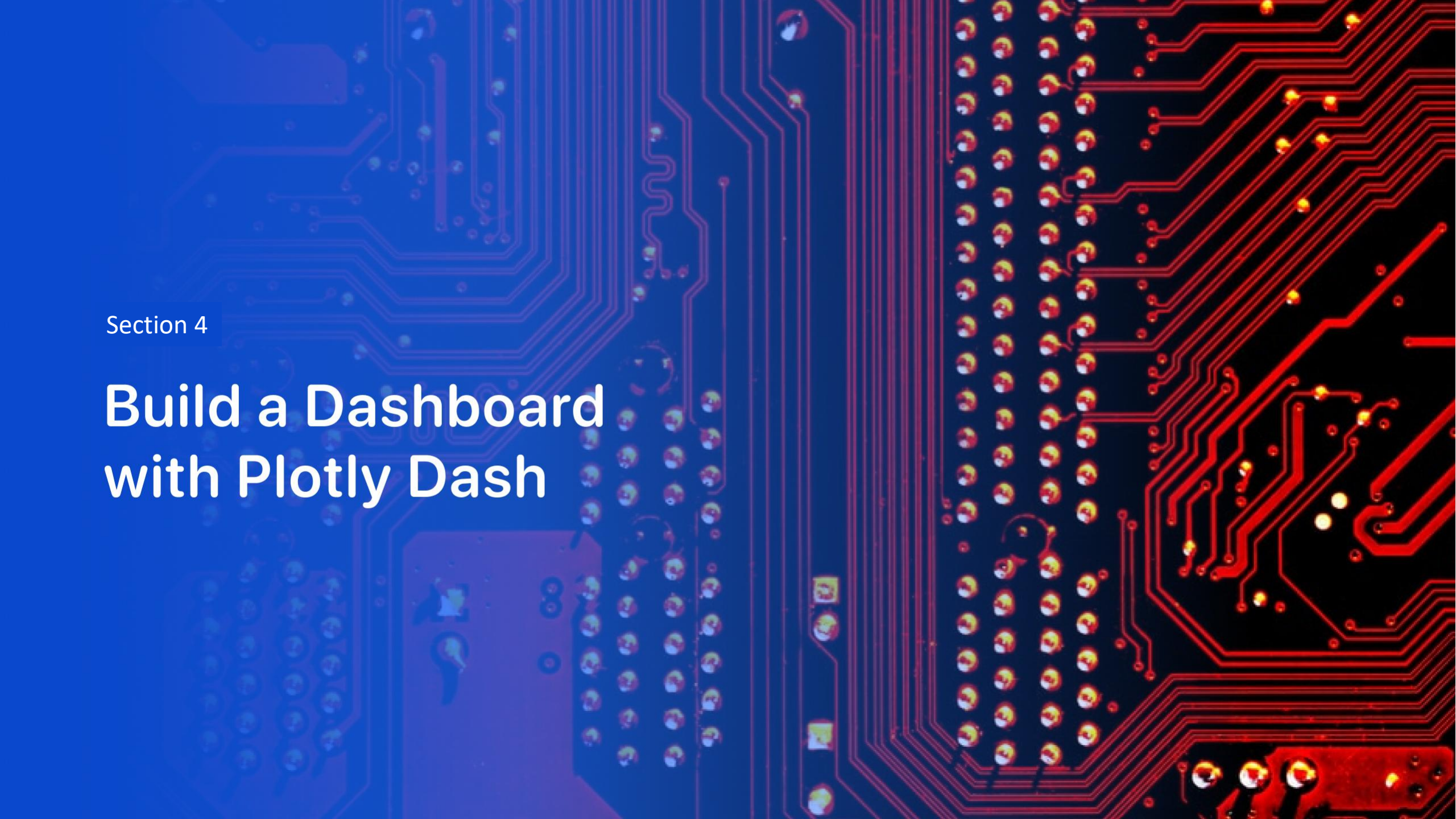


DISTANCE FROM KSC LC-39A TO PROXIMITIES

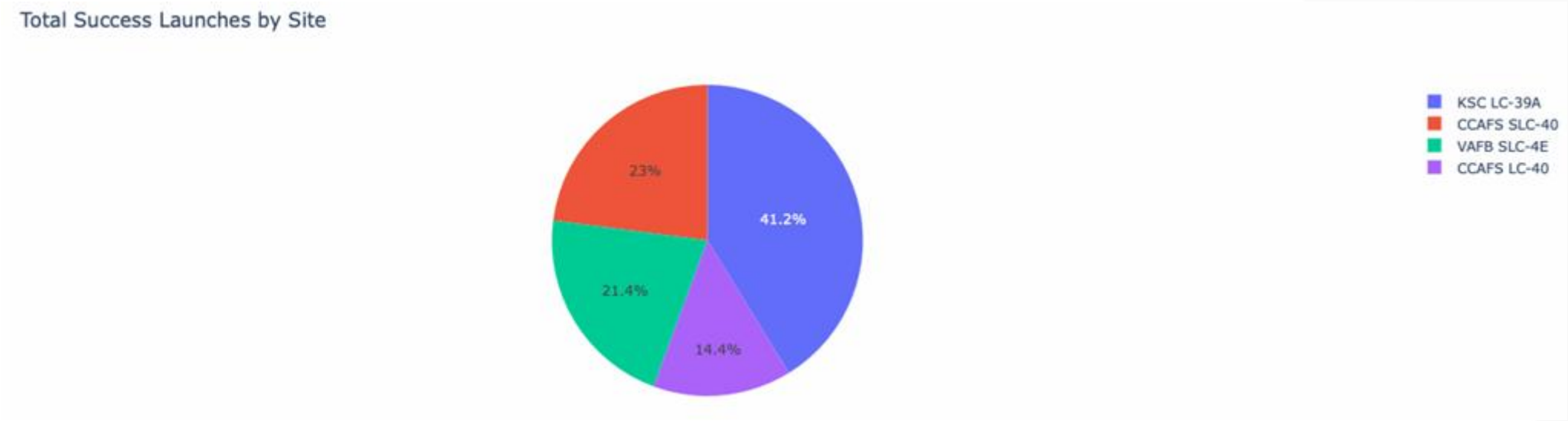
EXPLANATION:

- Visual analysis of **KSC LC-39A** shows it is:
  - Relative close to railway (15.23 km)
  - Relative close to highway (20.28 km)
  - Relative close to coastline (14.99 km)
- The site is also close to **Titusville (16.32 km)**.
- High-speed failed rockets could reach populated areas in seconds, presenting potential safety risks.




Section 4
Build a Dashboard with Plotly Dash

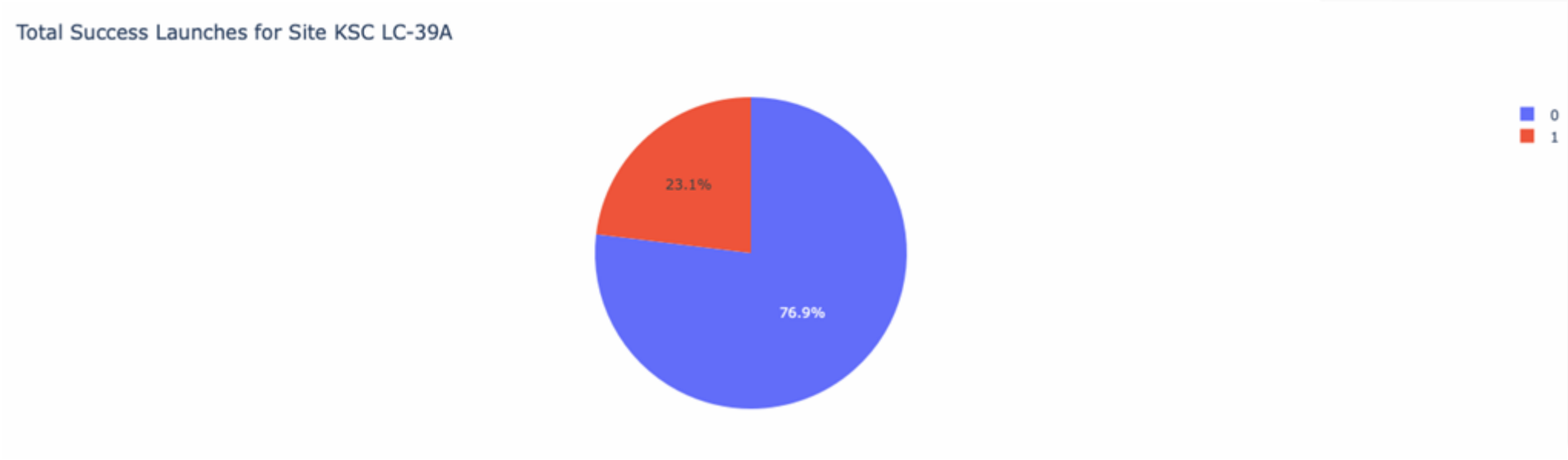
LAUNCH SUCCESS COUNT FOR ALL SITES



EXPLANATION:

- The chart clearly shows that from all the sites, **KSC LC-39A** has the most successful launches.

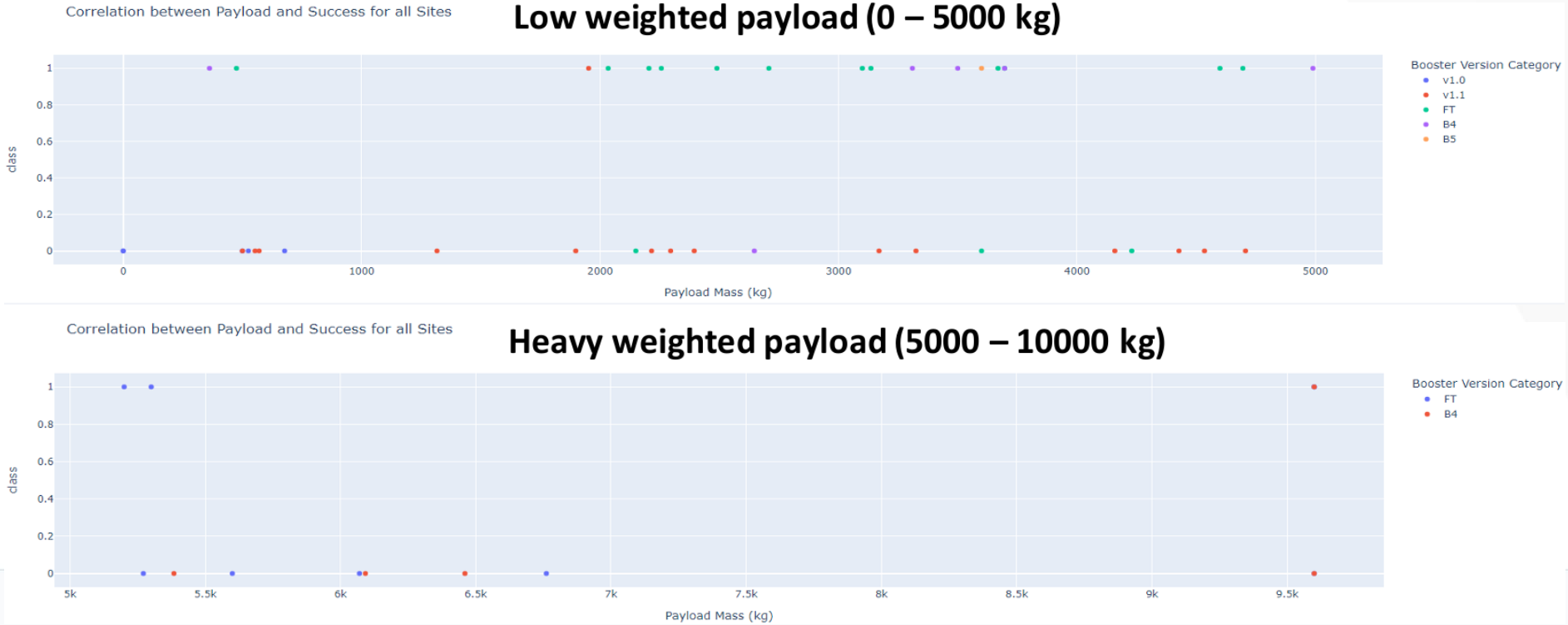
LAUNCH SITE WITH HIGHEST LAUNCH SUCCESS RATIO



EXPLANATION:

- **KSC LC-39A** has the highest launch success rate (**76.9%**) with 10 successful and only 3 failed landings.

PAYLOAD MASS VS. LAUNCH OUTCOME FOR ALL SITES



EXPLANATION:

- The charts show that payloads between **2000 and 5500 kg** have the highest success rate.

Section 5

Predictive Analysis (Classification)

CLASSIFICATION ACCURACY

Explanation:

- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
- The scores of the whole Dataset confirm that the best model is the **SVM Model**. This model has not only higher scores, but also the highest accuracy.

Scores and Accuracy of the Test Set

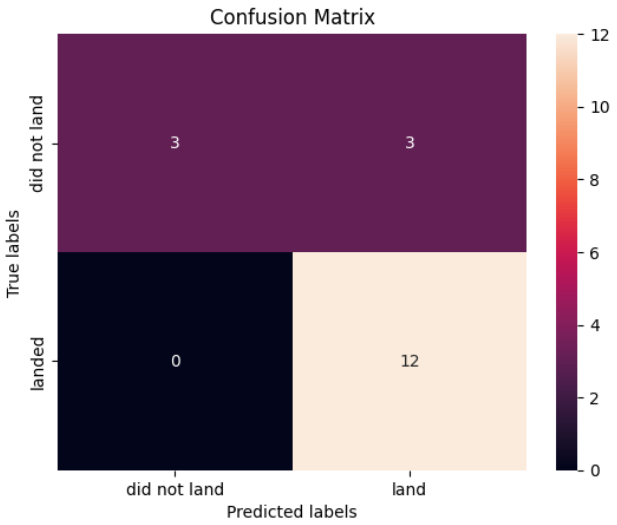
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.733333	0.800000
F1_Score	0.888889	0.888889	0.846154	0.888889
Accuracy	0.833333	0.833333	0.777778	0.833333

Scores and Accuracy of the Whole Set

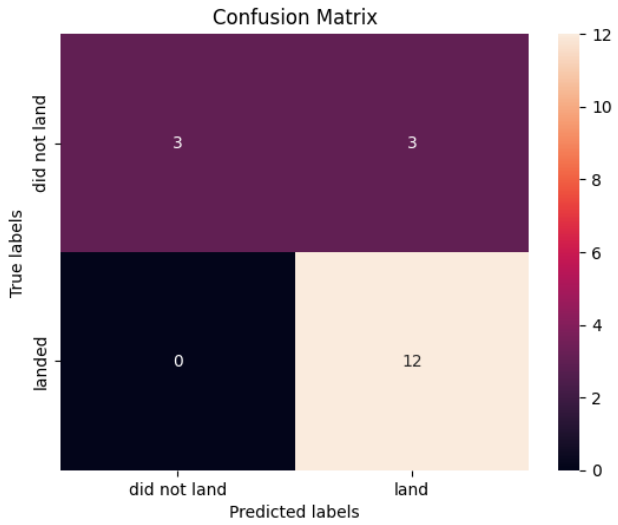
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.826087	0.819444
F1_Score	0.909091	0.916031	0.904762	0.900763
Accuracy	0.866667	0.877778	0.866667	0.855556

CONFUSION MATRIX

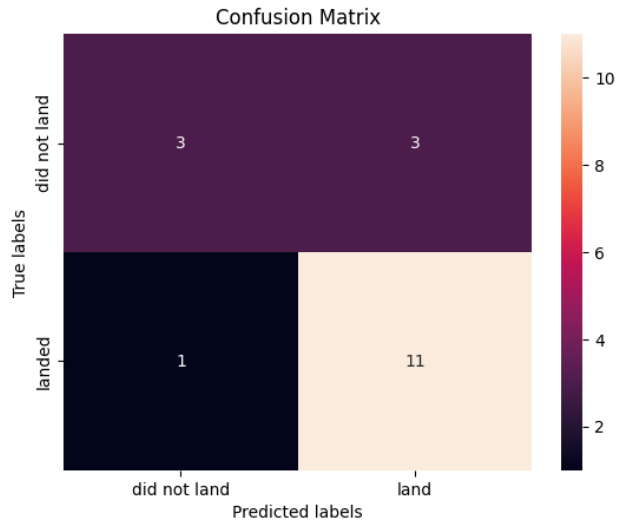
Logistic Regression



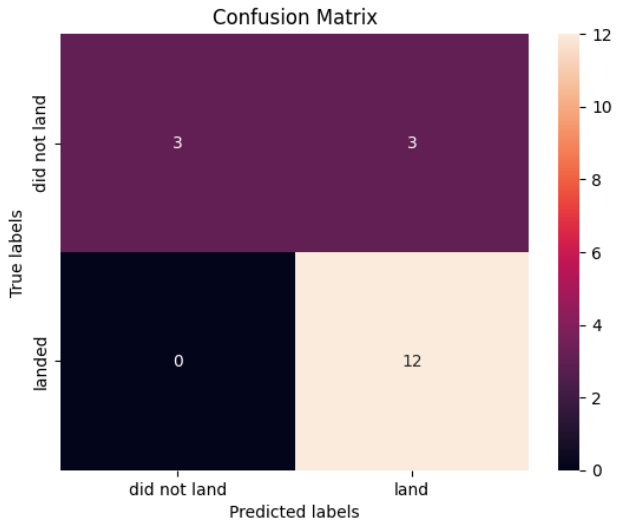
SVM



Decision Tree



KNN



EXPLANATION:

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is **false positives**.

CONCLUSION

KEY TAKEAWAYS

- Mission success is highly correlated with **technological maturity** (Flight Number) and **Payload Mass**.
- Significant reusability milestones post-2013 have led to massive launch cost reductions.
- Machine learning provides a reliable framework for estimating landing success with **83.33% accuracy**.

BUSINESS IMPACT

Accurate landing prediction allows competitors to calculate the **"real" costs** of a SpaceX launch, enabling more effective market competition and strategic bidding.

Expansion of the dataset to include **Starship telemetry** could further enhance predictive precision for next-generation heavy-lift missions.

Thank you!