IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

François Pape Diouf
13/03/2025

Outline

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

Executive Summary

Objective

The goal of this project is to analyze the cost implications of SpaceX's Falcon 9 rocket launches by developing a predictive model that determines whether the first stage will land successfully. This analysis provides insights into cost-saving opportunities and competitive benchmarking for other aerospace companies.

Results

Launch success improved with experience; the initial years (2010–2013) saw no successful missions, while post-2013 showed consistent progress. **KSC LC-39A** at Kennedy Space Center recorded the highest success rate (**76.9%**) and the most successful launches. **Coastal launch sites** offered strategic advantages in terms of safety and logistics. **Heavy payloads** exceeding 10,000 kg were associated with higher success rates, particularly for **LEO, Polar,** and **ISS** orbits, whereas the **GTO** orbit showed the lowest success rate with no clear link between flight number and success. **ES-L1**, **GEO**, **HEO**, and **SSO** orbits achieved a 100% success rate. On the modeling side, all tested models performed similarly on test data; however, model selection considered **F1-score**, **ROC-AUC**, and standard error, with **K-Nearest Neighbors (KNN)** emerging as the best-performing model.

Methodologies

1. **Data collection:** historical Falcon9 launch records were collected from Wikipedia using web scraping. Information about launches suck as the rocket used, payload delivered, rocket specifications, landing specifications and outcome were collected from SpaceX Rest API

2. **Data cleaning and wrangling:** missing values in the PayloadMass column were imputed using mean. Landing outcomes were converted into a classification variable(1 for successful, 0 for unsuccessful)

3. **Exploratory data analysis (EDA):** data was explored using SQL, Pandas and Matplotlib. Relationships between variables such as Flight Number and Launch Site, Payload Mass and Launch Site, Orbit type and Success rate… were visualized and based on those preliminary insights, features were selected for model building. Categorical features were encoded into dummies variables and the data saved into a csv file model building.

4. **Interactive visual analytics and Dashboard:** launch sites, launch outcomes and the distance between launch sites and its proximities (city, railway, highway) were mapped

5. **Predictive Analysis (Classification):** previously processed data was load and numeric features were standardized. Data was then split into training and validation sets. We trained and tunned several machine learning models such as Logistic regression, support vector machine, decision tree and kneighbors classifier.

Introduction

The advertisement of Falcon 9 rocket at a launch cost of $63 million by SpaceX, has revolutionized the space industry. In fact, other providers launches can exceed $165 million per launch. The main reason of this cost efficiency is the reusability of the Falcon 9 first stage. SpaceX can drastically reduce the overall cost of each launch by successfully landing and reusing the first stage.

This project objective is to develop a predictive model which will determine whether the Falcon 9 first stage will land successfully. This will help assess the cost implications of rocket launches. SpaceX but also other competitors willing to bid against SpaceX for rocket launches can gain valuable insights by understanding the factors that influence landing success.

Throughout this project, we will try to gather insights to understand the cost dynamics of rocket launches and provide strategic advantages in the competitive space industry

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - historical Falcon9 launch records were collected from Wikipedia using web scraping. Information about launches suck as the rocket used, payload delivered, rocket specifications, landing specifications and outcome were collected from SpaceX Rest API
- Perform data wrangling
  - missing values in the PayloadMass column were imputed using mean. Landing outcomes were converted into a classification variable(1 for successful, 0 for unsuccessful)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

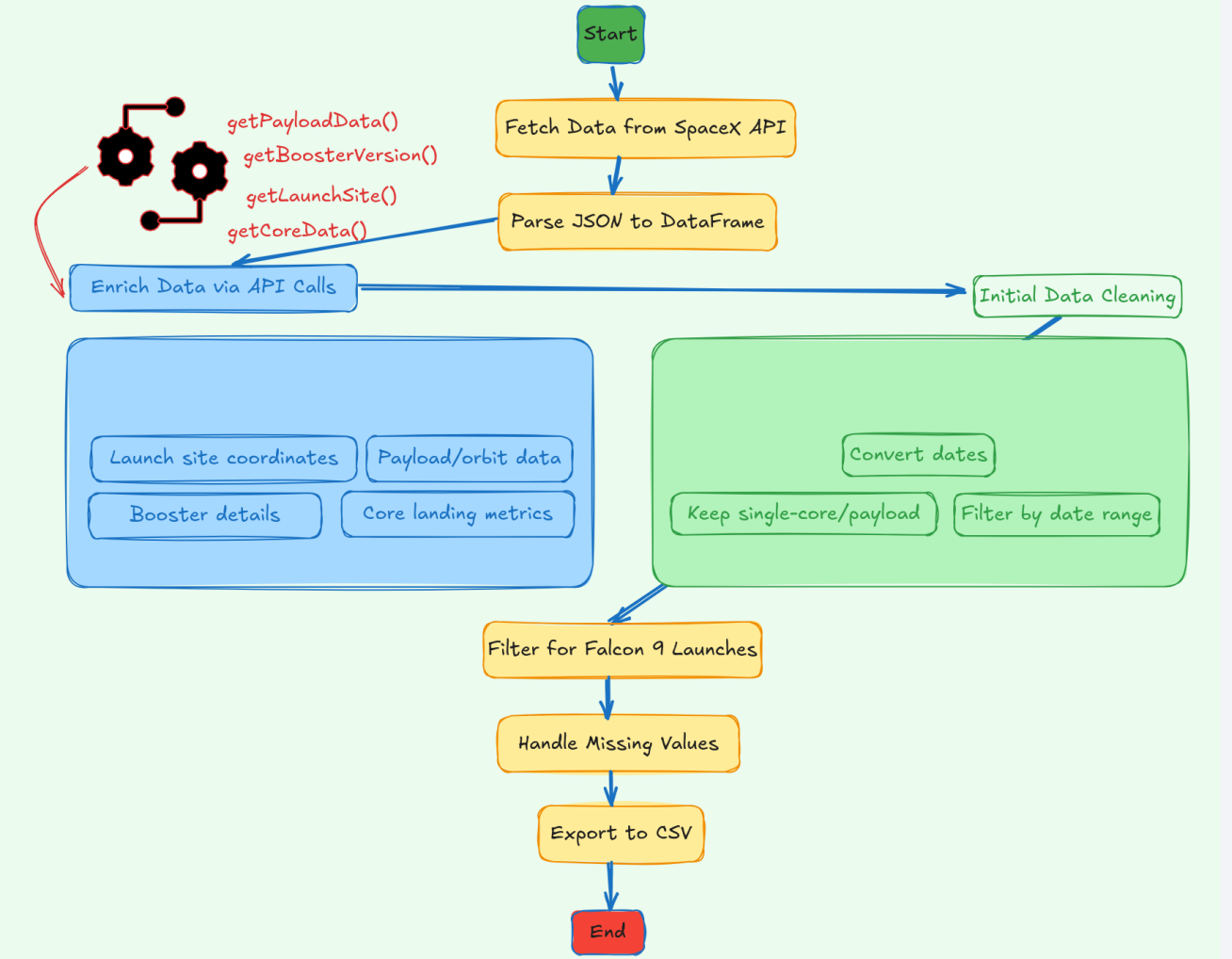
Data Collection

- historical Falcon9 launch records were collected from Wikipedia using web scraping.
- Information about launches suck as the rocket used, payload delivered, rocket specifications, landing specifications and outcome were collected from SpaceX Rest API

Data Collection – SpaceX API

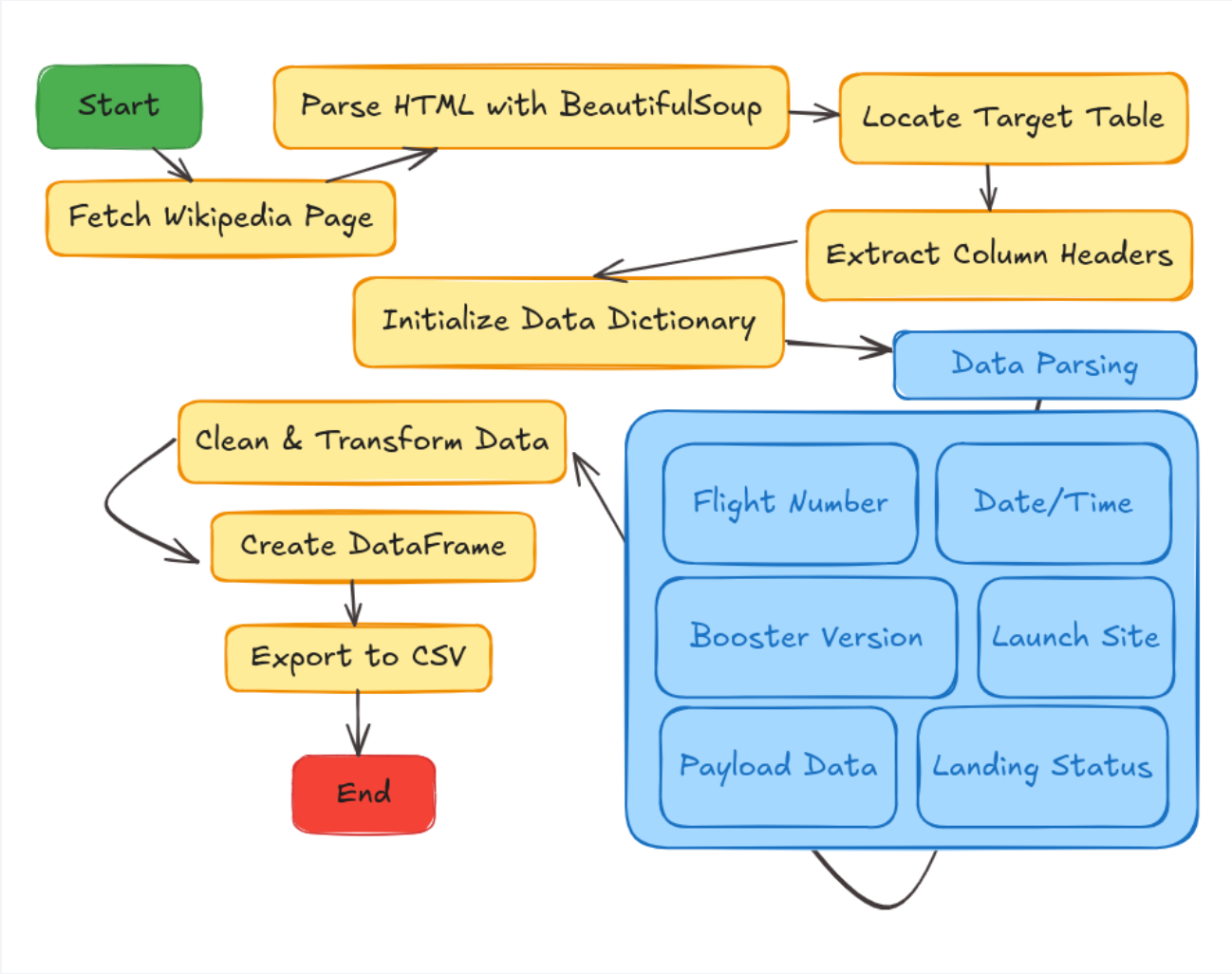
Information about launches suck as the rocket used, payload delivered, rocket specifications, landing specifications and outcome were collected from SpaceX Rest API

https://github.com/dioufra/IBM_data_science_capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

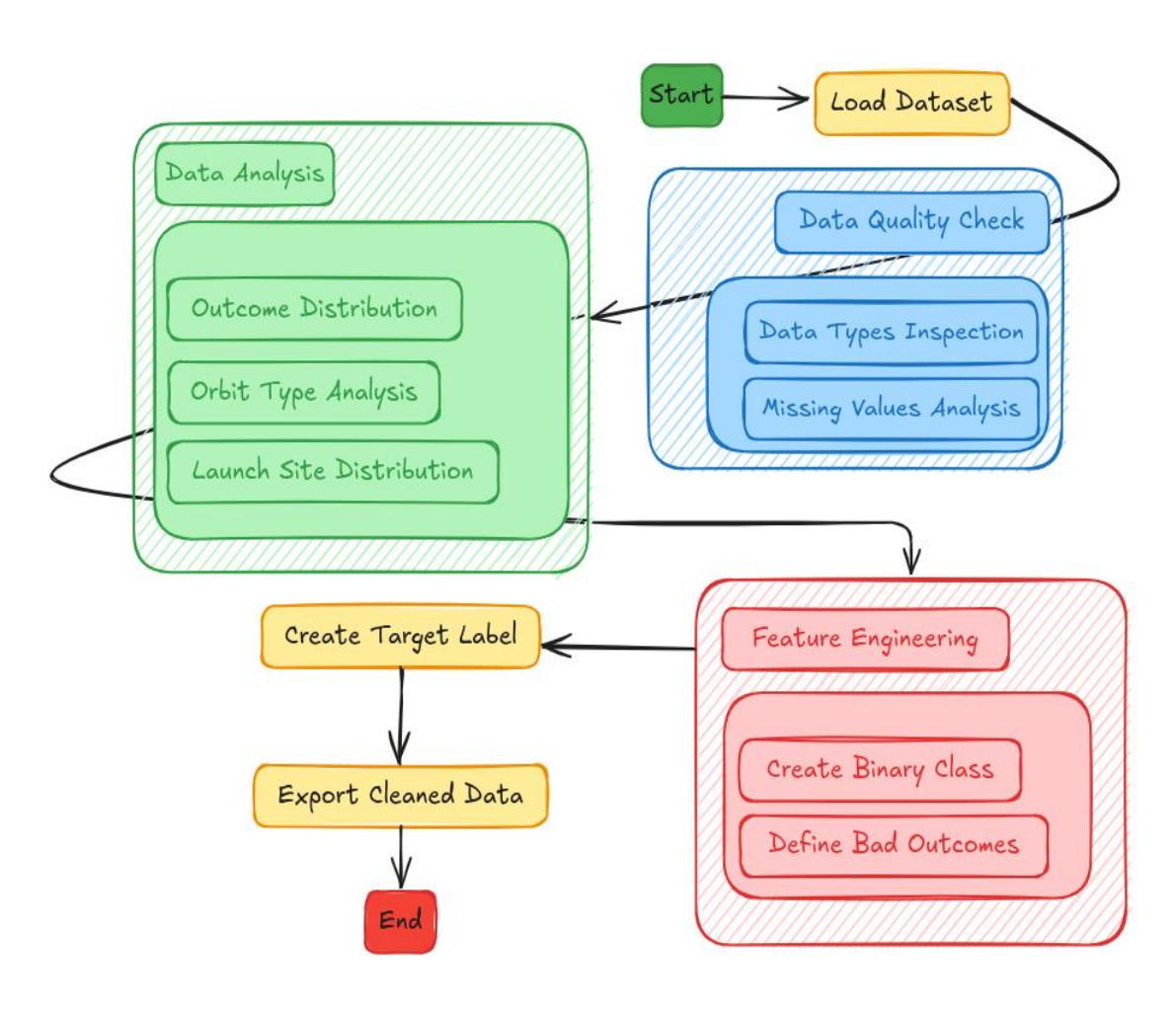
- Historical Falcon9 launch records were collected from Wikipedia using web scraping. The html content of the page was fetched with **requests**. Using **BeautifulSoup**, the content was parsed into a soup object. Details such as the **date**, the **booster version**, the **landing status**, and the **payload mass** were then extracted using helper functions.
- https://github.com/dioufra/IBM_data_science_capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

The workflow begins by loading the dataset and performing initial checks on data quality, types, and missing values. Next, feature engineering transforms outcomes into a binary classification (success/failure) with defined failure criteria, followed by the creation of a clear target variable for modeling. Finally, the cleaned data is exported in a structured format (e.g., CSV/parquet) with documented transformations, ensuring analysis-ready quality

https://github.com/dioufra/IBM_data_science_capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

1. FlightNumber vs. PayloadMass

- Type: Scatter plot (hue=Class)
- Purpose: Reveal how launch attempts and payload weight correlate with success.

2. FlightNumber vs. LaunchSite

- Type: Scatter plot (hue=Class)
- Purpose: Compare success trends across launch sites over time.

3. PayloadMass vs. LaunchSite

- Type: Scatter plot (hue=Class)
- Purpose: Identify payload capacity limits by site.

4. Orbit Success Rates

- Type: Bar chart (mean Class by Orbit)
- Purpose: Compare success rates across orbit types.

5. Yearly Success Trend

- Type: Line chart (Year vs. mean Class)
- Purpose: Track annual improvements in landing success.

6. Feature Engineering

- Actions: One-hot encoded Orbit, LaunchSite, and LandingPad.

https://github.com/dioufra/IBM_data_science_capstone/blob/main/edadataviz.ipynb

EDA with SQL

1. Unique Launch Sites

- Identify all unique launch locations

2. Launch Sites Starting with 'CCA'

- Filter records for Cape Canaveral sites (first 5 entries)

3. Total Payload Mass for NASA (CRS)

- Find average payload capacity for the F9 v1.1 booster.

4. Average Payload for F9 v1.1

- Find average payload capacity for the F9 v1.1 booster.

5. First Successful Ground Landing Date

- Identify the earliest successful ground pad landing.

6. Boosters with Drone Ship Success (Payload 4k–6k kg)

- List boosters that succeeded with mid-range payloads.

7. Mission Outcomes Count

- Aggregate success/failure rates.

8. Boosters with Max Payload Mass

- Identify high-capacity boosters.

9. 2015 Drone Ship Failures by Month

- Analyze temporal trends in failures.

10. Landing Outcomes Ranking (2010–2017)

- Compare landing success rates over a 7-year period.

https://github.com/dioufra/IBM_data_science_capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Launch Site Markers**
  - Added circles (radius: 1000m) and labeled markers to highlight SpaceX launch sites.
- **Success/Failure Visualization**
  - Used MarkerCluster with color-coded markers (green = success, red = failure) to show launch outcomes.
- **Proximity Analysis**
  - Calculated distances to coastline, highways, railways, and cities.
  - Added distance markers and connecting lines for spatial context.
- **Purpose:** To analyze launch site locations, success patterns, and logistical relationships with nearby infrastructure.

https://github.com/dioufra/IBM_data_science_capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- **1. Interactive Components**
  - Dropdown Menu:
    - Allows selection of specific launch sites or "All Sites"
    - filter data dynamically.
  - Payload Range Slider:
    - Adjusts the payload mass range (0–10,000 kg)
    - analyze success rates by payload weight.
- **2. Visualizations**
  - Pie Chart:
    - Shows success vs. failure rates
    - For all launch sites or single selected site
  - Scatter Plot:
    - Displays correlation between payload mass and launch success,

Why These Elements Were Added

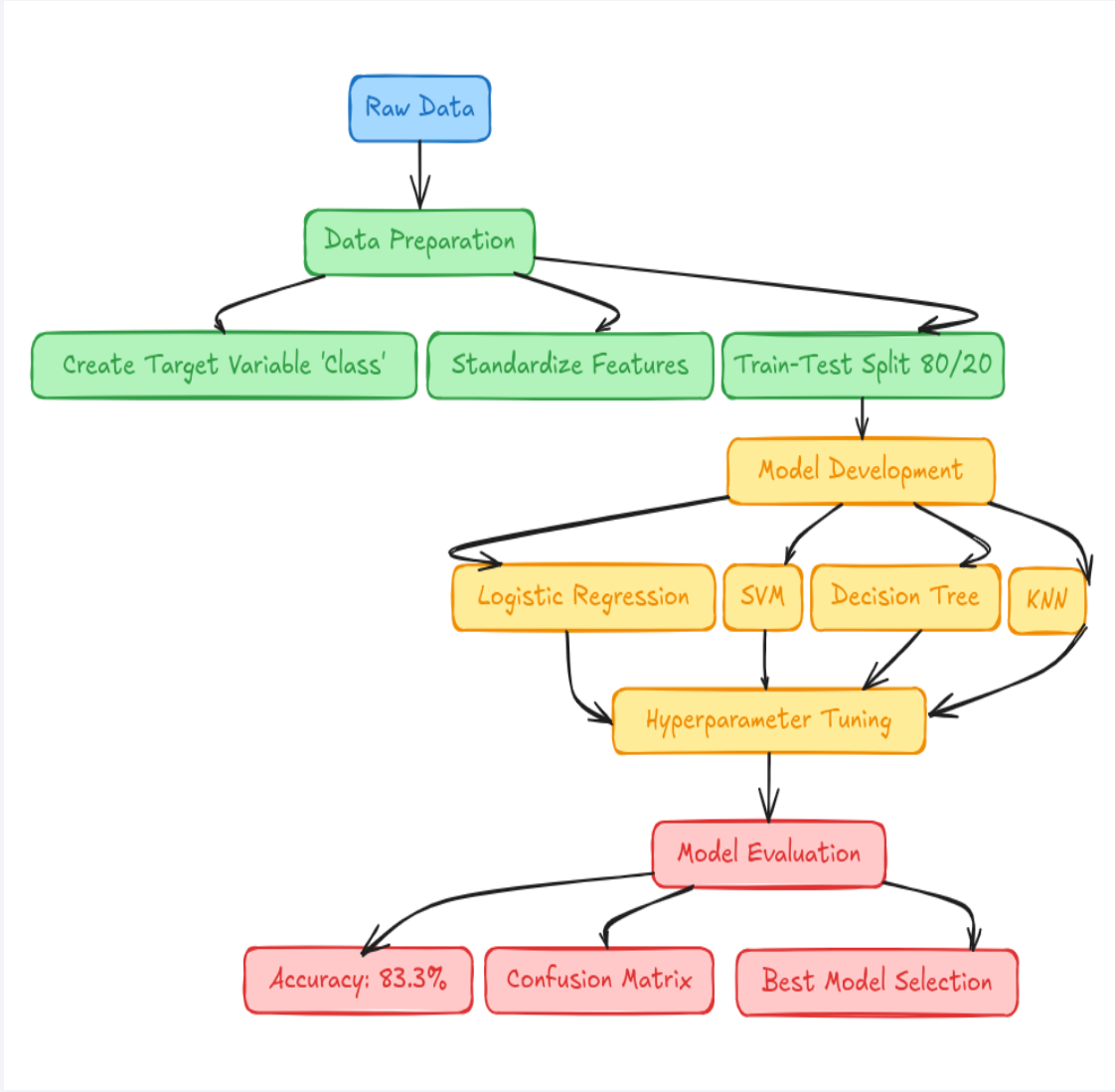
- Dropdown & Pie Chart:
  - Quickly compare success rates across sites
  - Or look into a single site's performance.
- Payload Slider & Scatter Plot:
  - Investigate impact of heavier payloads on success rates
  - And how booster versions perform.

https://ds-capstone-dash-app.onrender.com/

https://github.com/dioufra/IBM_data_science_capstone/blob/main/spacex_dash_app.py

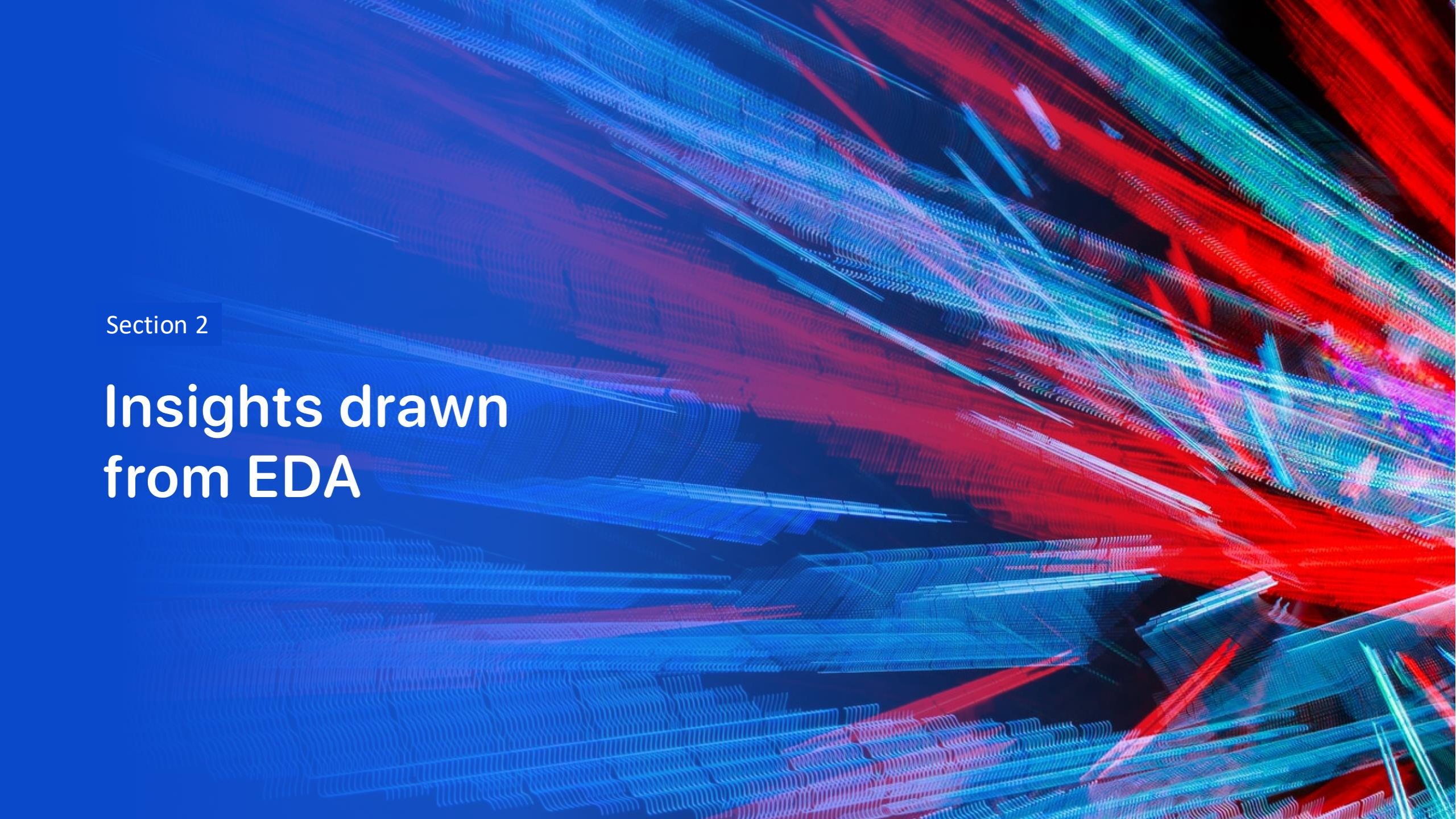
Predictive Analysis (Classification)

- **1. Data Preparation**
  - Engineered binary target: Class (1=landed, 0=failed)
  - Standardized 18 features using **StandardScaler**
  - Split data (80% train / 20% test) with fixed random state
- **2. Model Development**
  - Tested 4 algorithms with **GridSearchCV** (10-fold CV):
    - Logistic Regression
    - SVM (Support Vector Machine)
    - Decision Tree (max_depth=4, gini criterion)
    - KNN (K Nearest Neighbors)

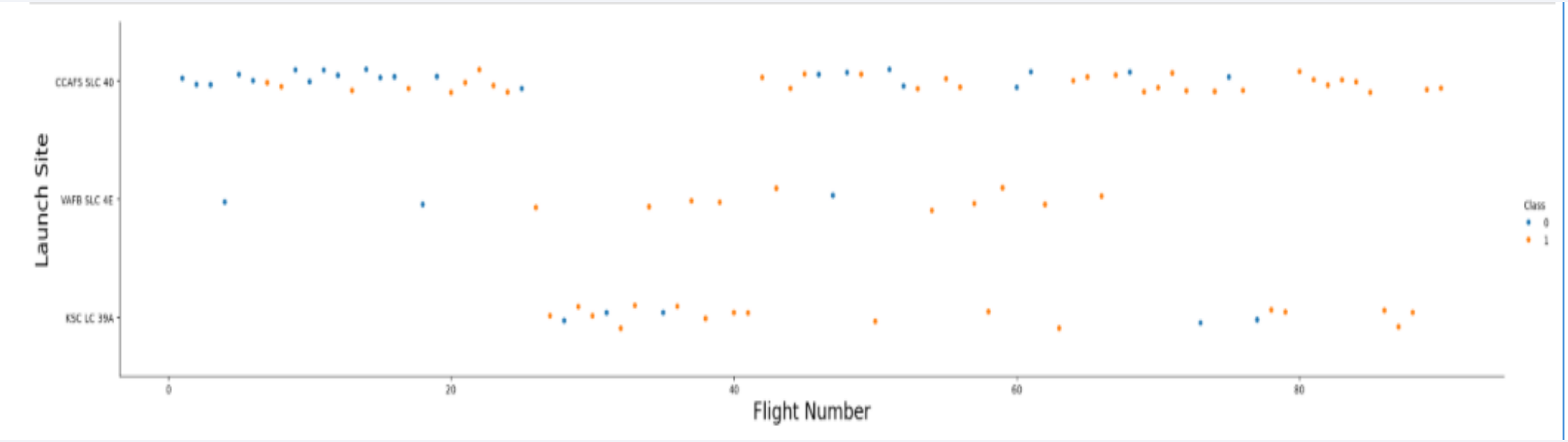


Results

- **Launch Patterns & Success Rates:**
  - Flight Experience Matters: Launch success improved with experience; initial years (2010–2013) had no successes, whereas post-2013 showed improvement.
- **Launch Site Insights:**
  - LC-39A at KSC had the highest success rate (76.9%) and most successful launches.
  - Coastal launch sites provided safety and logistical advantages.
- **Payload & Orbit Effects:**
  - Heavy payloads (>10,000 kg) had high success, especially for LEO, Polar, and ISS orbits.
  - GTO orbit had lowest success rate with no apparent correspondence between flight number and success.
  - ES-L1, GEO, HEO, and SSO orbits experienced a 100% success rate.
- **Modeling & Prediction:**
  - Both models performed alike with test data.
  - The choice of models made use of F1-score, ROC-AUC, and Standard Error.
  - K-Nearest Neighbors (KNN) was ranked the best among performing models.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



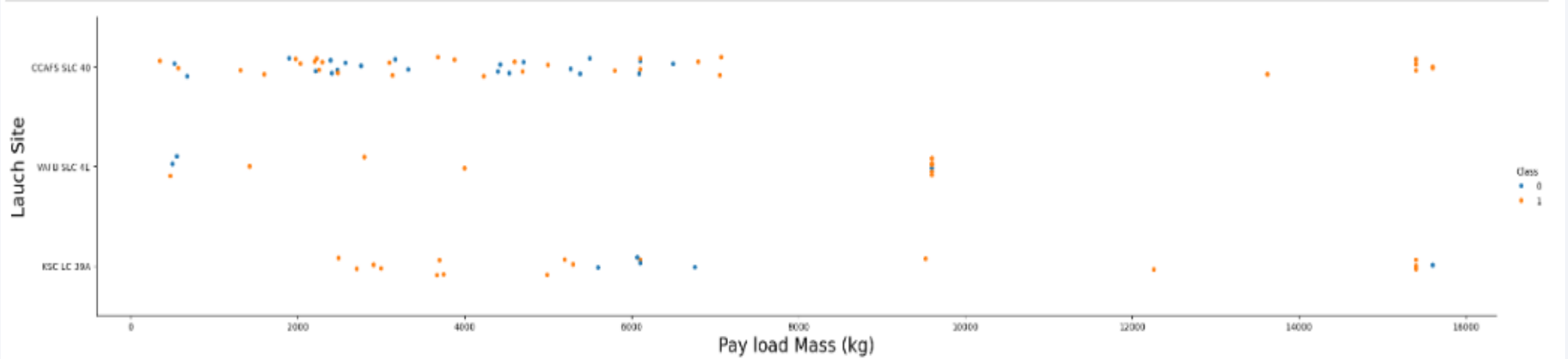
Description:

- Relationship between Flight number and Launch site
- X-axis (Flight Number): order of the flight
- Y-axis (launch site).
- Blue dot represents a failed missions
- Yellow dot represents successful missions

Observations:

- some launch sites were used more frequently than others.
- As the number of flight increases, the mission outcome is likely to be successful.
- It also indicates that SpaceX learned from past failures, leading to better outcomes as flight numbers increased.

Payload vs. Launch Site



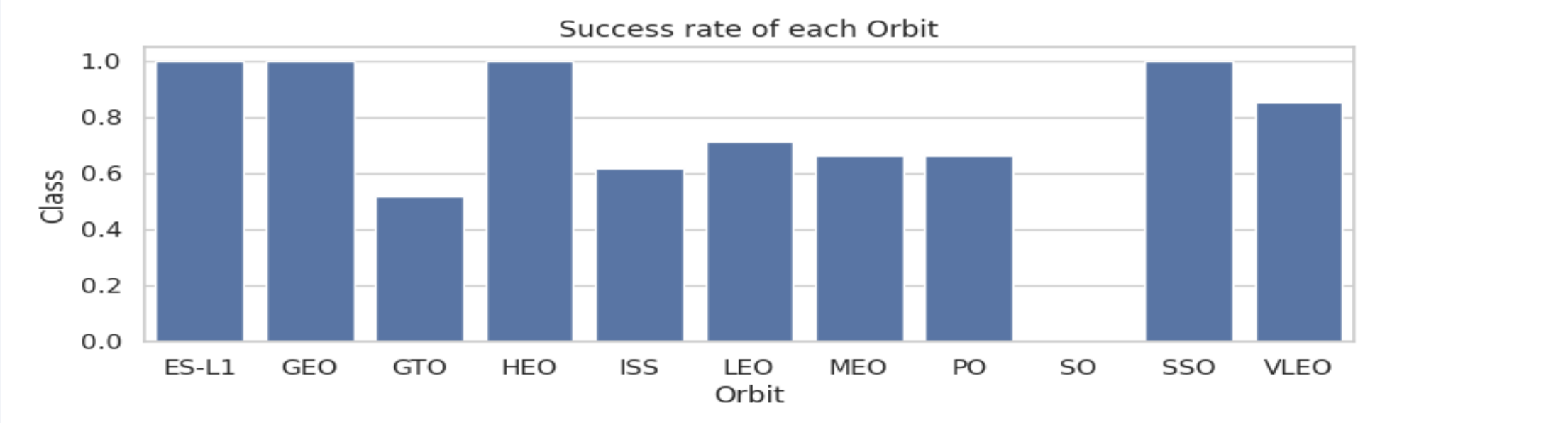
Description:

- Relationship between Payload mass Launch site
- X-axis (Payload mass): mass of the payload
- Y-axis (launch site).
- Blue dot represents a failed missions
- Yellow dot represents successful missions

Observations:

- Almost all rockets launched for heavy payload mass were successful
- there are no rockets launched for heavy payload mass(greater than 10000) for VAFB-SLC

Success Rate vs. Orbit Type



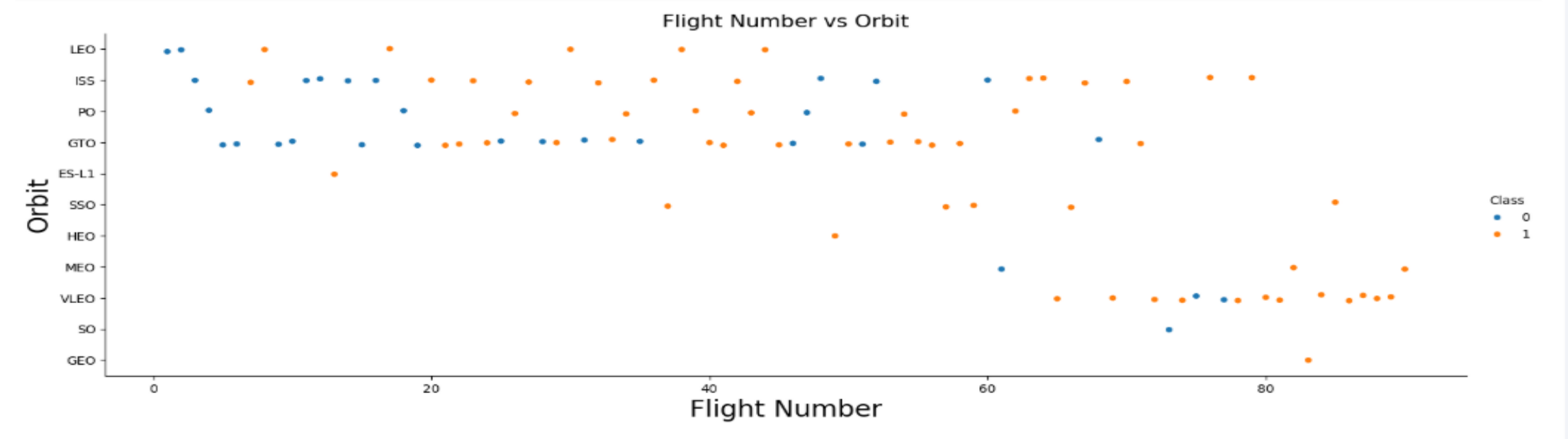
Description:

- Success rate of each obit
- X-axis (Orbit): orbits (LEO, ISS, GTO, ...), destination of each mission.

Observations:

- ES-L1, GEO, HEO and SSO orbits have the highest success rate (1)
- ISS, MEO, LEO, PO show a moderate success rate between 0.6 and 0,7
- GTO has the lowest success

Flight Number vs. Orbit Type



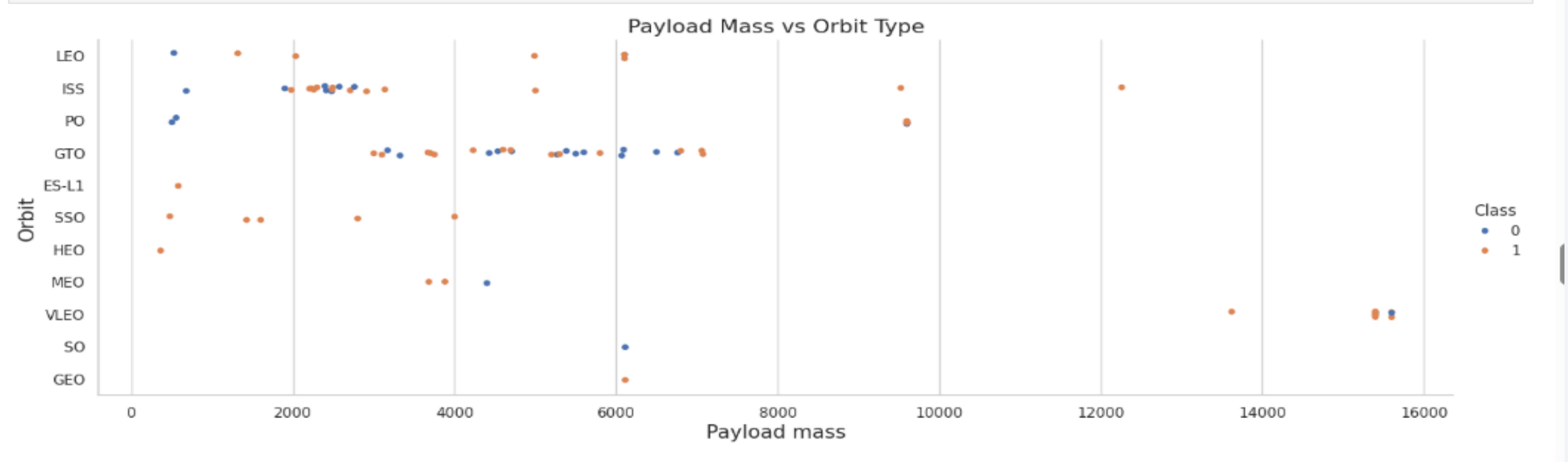
Description:

- Relationship between flight number and Orbit
- X-axis (Flight Number): order in which missions were launched.
- Y-axis (Orbit): orbits (LEO, ISS, GTO, ...), destination of each mission.
- Blue dot represents a failed missions
- Yellow dot represents successful missions

Observations:

- 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



Description:

- Relationship between Payload Mass vs Orbit Type
- X-axis (Payload mass): mass of the payload
- Y-axis (Orbit): orbits (LEO, ISS, GTO, ...), destination of each mission.
- Blue dot represents a failed missions
- Yellow dot represents successful missions

Observations:

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

Launch Success Yearly Trend

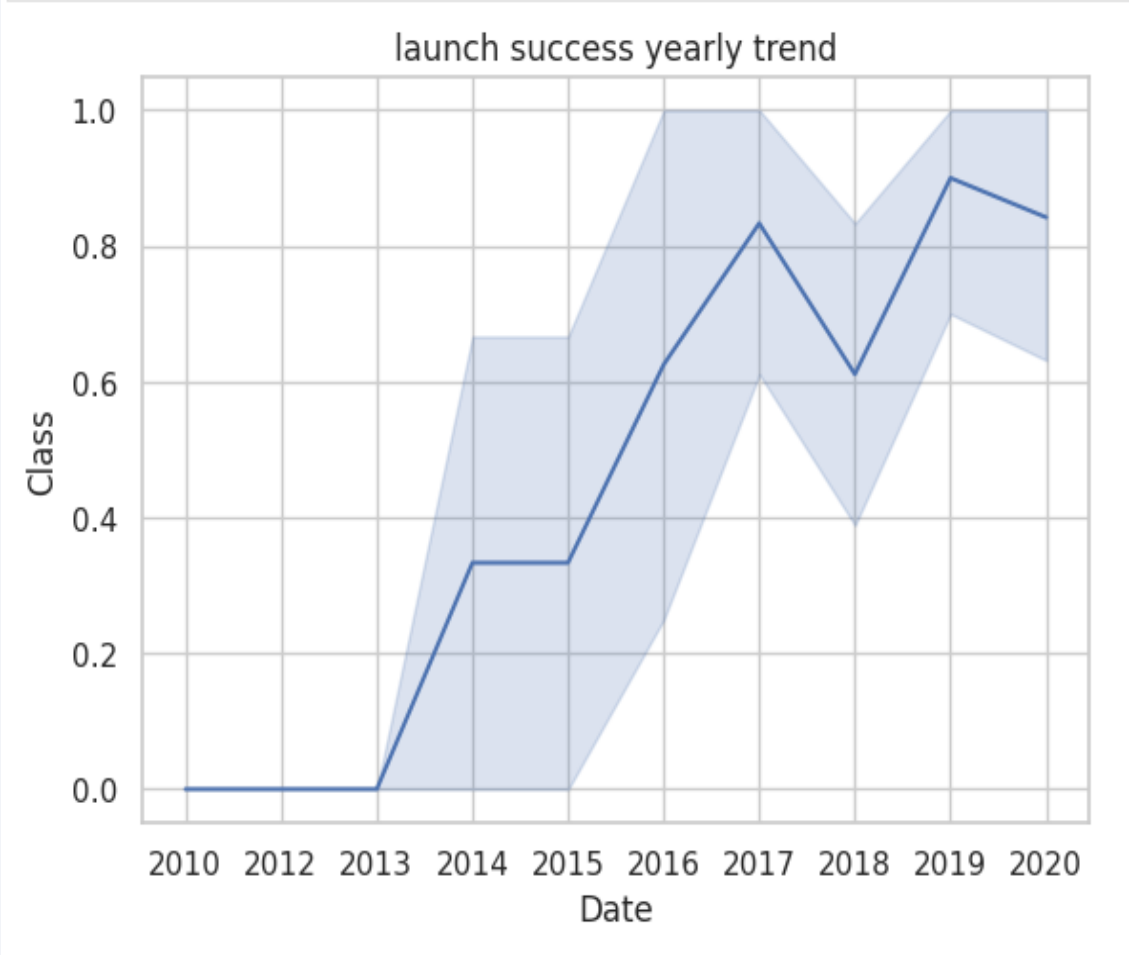
This line plot shows the trend of **launch success rates over time** from **2010 to 2020**, with an associated confidence interval (shaded area).

Axes and Labels:

- **X-axis (Date):** Represents the years from 2010 to 2020.
- **Y-axis (Class):** Represents the success rate of launches, ranging from **0 (no success)** to **1 (100% success rate)**.
- **Shaded Region:** Represents the confidence interval, showing variability in success rates.

Observations:

- **Early Years (2010 - 2013):**
  - No successful launches (success rate = 0).
- **2013 - 2020:**
  - success rate kept increasing



All Launch Site Names

- This SQL query retrieves the unique launch sites used in space missions from the **SPACEXTABLE** database.
- Four unique launch sites are identified:
  - **CCAFS LC-40**
  - **VAFB SLC-4E**
  - **KSC LC-39A**
  - **CCAFS SLC-40**

The objectif of this analysis was to identify where spaceX missions are launched from

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
```

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

Launch Site Names Begin with 'CCA'

- This SQL query retrieves distinct rows from the **SPACEXTABLE** where the **Launch_Site** name starts with `'CCA'`, limited to **5 records**.

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

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

- The SQL query calculates the sum of the PAYLOAD_MASS_KG column for all records where the Customer is '**NASA (CRS)**'.

- The query result shows that the total payload mass carried by SpaceX boosters for **NASA (CRS)** missions is **45,596 kg.**
- This indicates that **SpaceX** has transported a significant amount of cargo (over 45 metric tons) to support **NASA**'s missions

```
%%sql SELECT SUM("PAYLOAD_MASS__KG_")
    AS TOT_PAYLOAD_MASS
    FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```


TOT_PAYLOAD_MASS
45596

Average Payload Mass by F9 v1.1

- The SQL query calculates the average of the PAYLOAD_MASS_KG column for all records where where the Booster_Version is '**F9 v1.1**'.

- This suggests that the **F9 v1.1** had an average payload capacity of approximately **2.93** metric tons per mission.

```
%%sql SELECT AVG("PAYLOAD_MASS__KG_")
    AS AVG_PAYLOAD_MASS
    FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
```


AVG_PAYLOAD_MASS
2928.4

First Successful Ground Landing Date

- The SQL query retrieves the minimum (oldest) date (MIN(Date)) from the SPACEXTABLE where the **Landing_Outcome** was 'Success (ground pad)'.
- The result is labeled as Date and marks a historic milestone in SpaceX's reusable rocket technology
- The query result shows that the earliest successful ground pad landing of a SpaceX booster occurred on **December 22, 2015**.

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


Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The query result lists the booster versions that successfully landed on a drone ship and carried a payload mass between **4,000** kg and **60,000** kg.
- The results show **7 boosters**:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1029.1
  - F9 FT B1021.2
  - F9 FT B1036.1
  - F9 B4 B1041.1
  - F9 FT B1031.2

```
%%sql SELECT Booster_Version
    FROM SPACEXTABLE
    WHERE Landing_Outcome = 'Success (drone ship)'
    AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 60000
```

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1036.1
F9 B4 B1041.1
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The query counts:
  - Successful_Missions: Records where Mission_Outcome starts with "Success".
  - Failed_Missions: Records where Mission_Outcome starts with "Failure".

- Success Rate: **100 out of 101** missions were **successful**, indicating a 99% success rate.
- Failure Rate: Only **1 failure** was recorded, demonstrating SpaceX's high reliability.

```
%%sql SELECT
    COUNT(CASE WHEN Mission_Outcome LIKE 'Success%' THEN 1 END) AS Successful_Missions,
    COUNT(CASE WHEN Mission_Outcome LIKE 'Failure%' THEN 1 END) AS Failed_Missions
FROM SPACEXTABLE;
```


Successful_Missions	Failed_Missions
100	1

Boosters Carried Maximum Payload

- The query uses a subquery to first find the maximum payload mass (MAX(PAYLOAD_MASS_KG)) in the table, then lists all boosters that carried that exact mass. The results highlight SpaceX's ability to repeatedly launch heavy payloads with reused boosters, a cornerstone of their cost-effective model.
- The results can be seen in the picture

```
%%sql SELECT Booster_Version
    FROM SPACEXTABLE
    WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE )
```

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

- The table shows two failed drone ship landing attempts involving early versions of SpaceX's Falcon 9 boosters:
- January Failure:
    - Booster: F9 v1.1 B1012
    - Launch Site: CCAFS LC-40 (Cape Canaveral)
    - Outcome: Failed to land on the drone ship.
- April Failure:
    - Booster: F9 v1.1 B1015
    - Launch Site: CCAFS LC-40
    - Outcome: Another drone ship landing failure.

Month_Name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- This query 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(Landing_Outcome) AS Total
    FROM SPACEXTABLE
    WHERE Date BETWEEN ' 2010-06-04' AND '2017-03-20'
    GROUP BY Landing_Outcome ORDER BY Total DESC
```

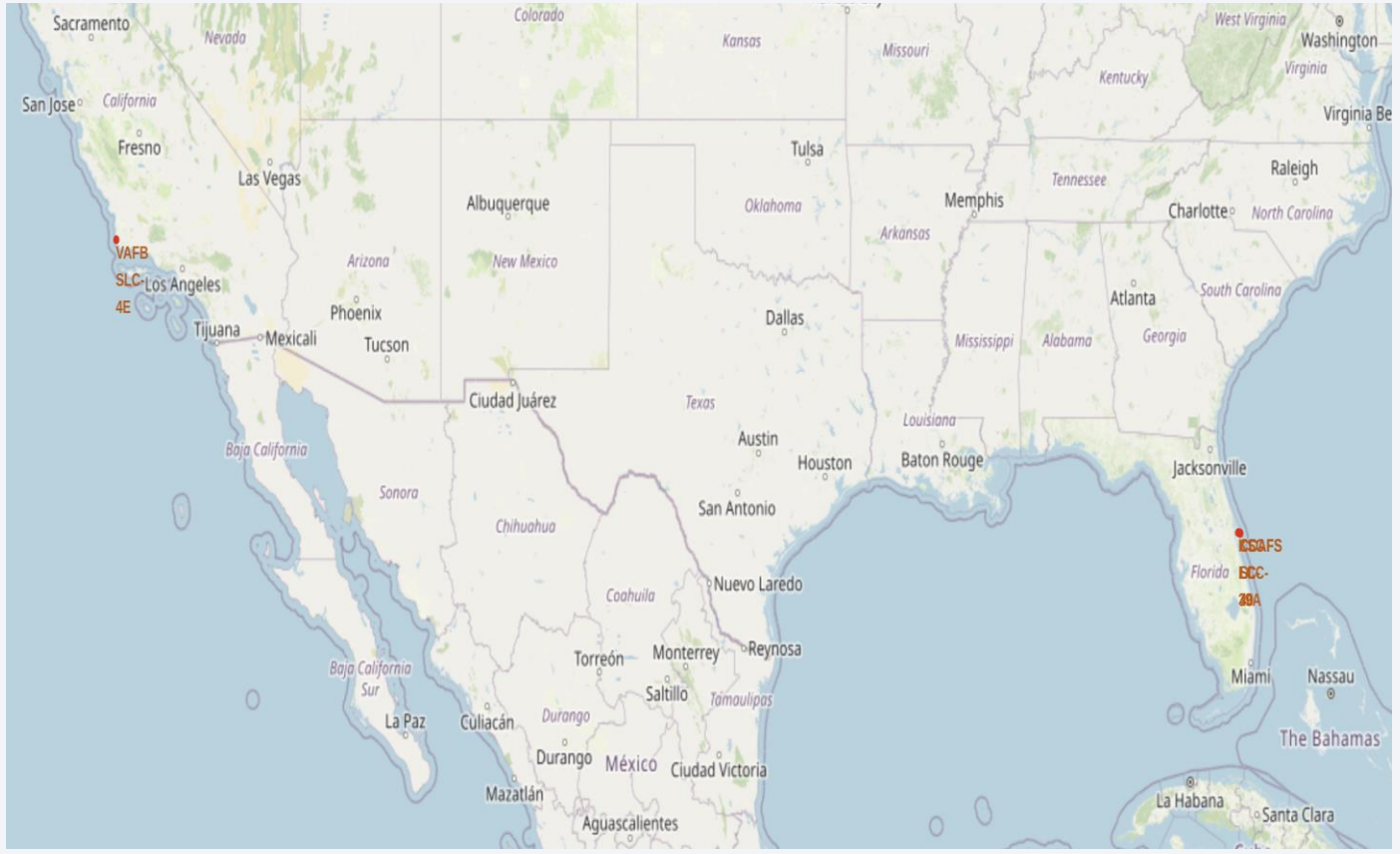
Landing_Outcome	Total
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

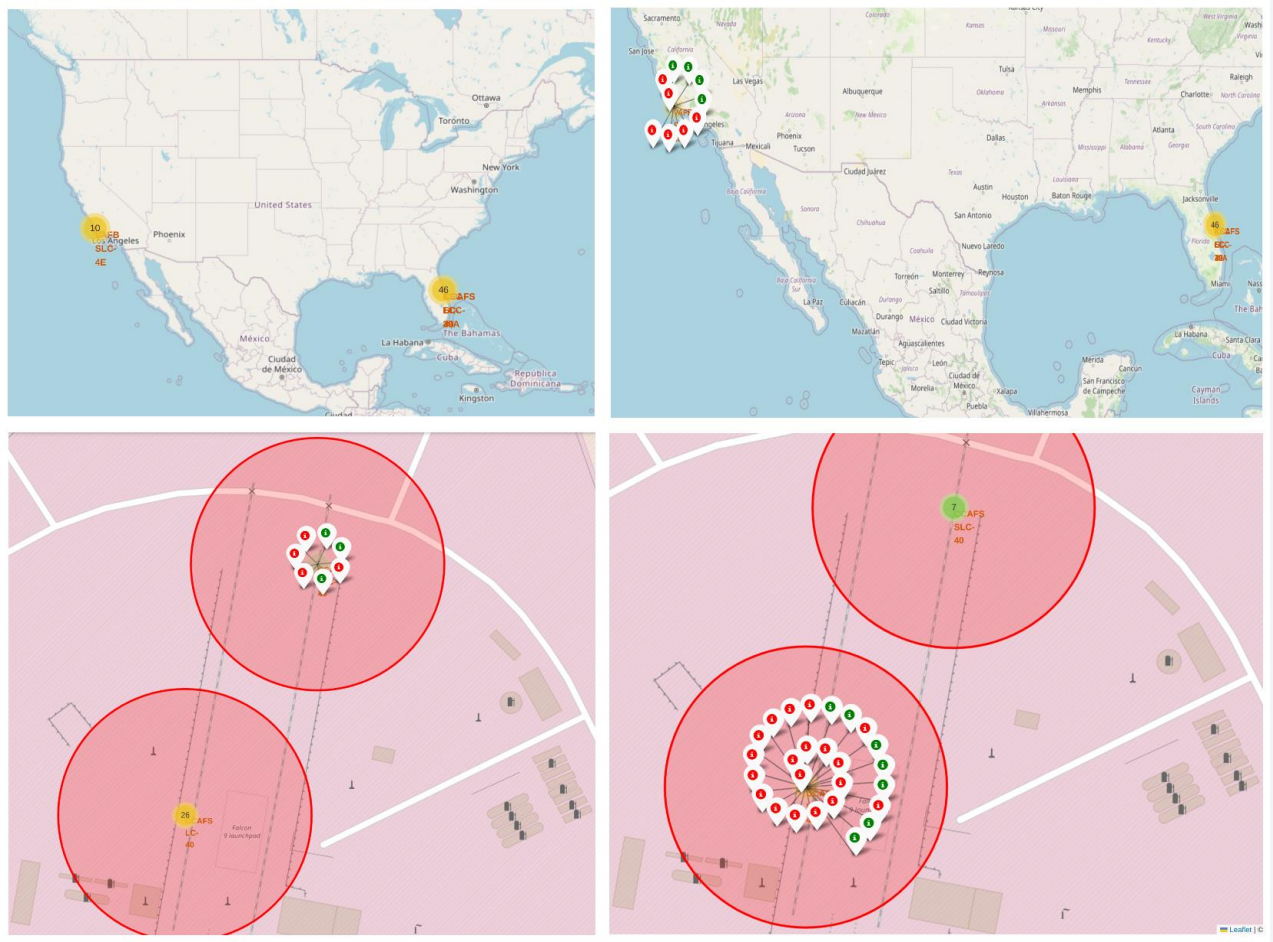
Launch Sites Locations

- SpaceX priorities proximity to coastal zone than equator line
  - Ocean trajectories minimize risk to populated areas.
  - Enables controlled water landings for failed launches.



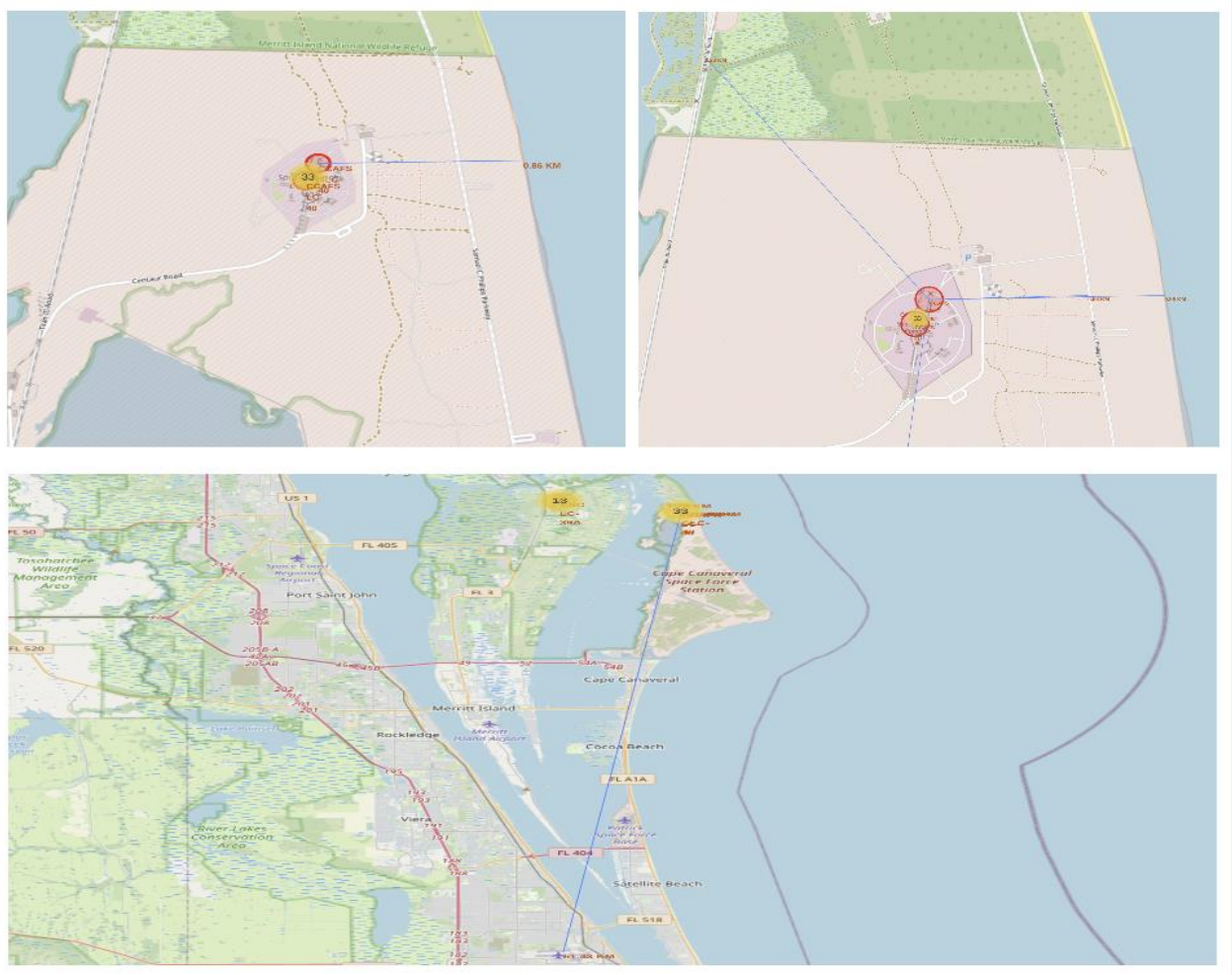
Launch Site Outcomes

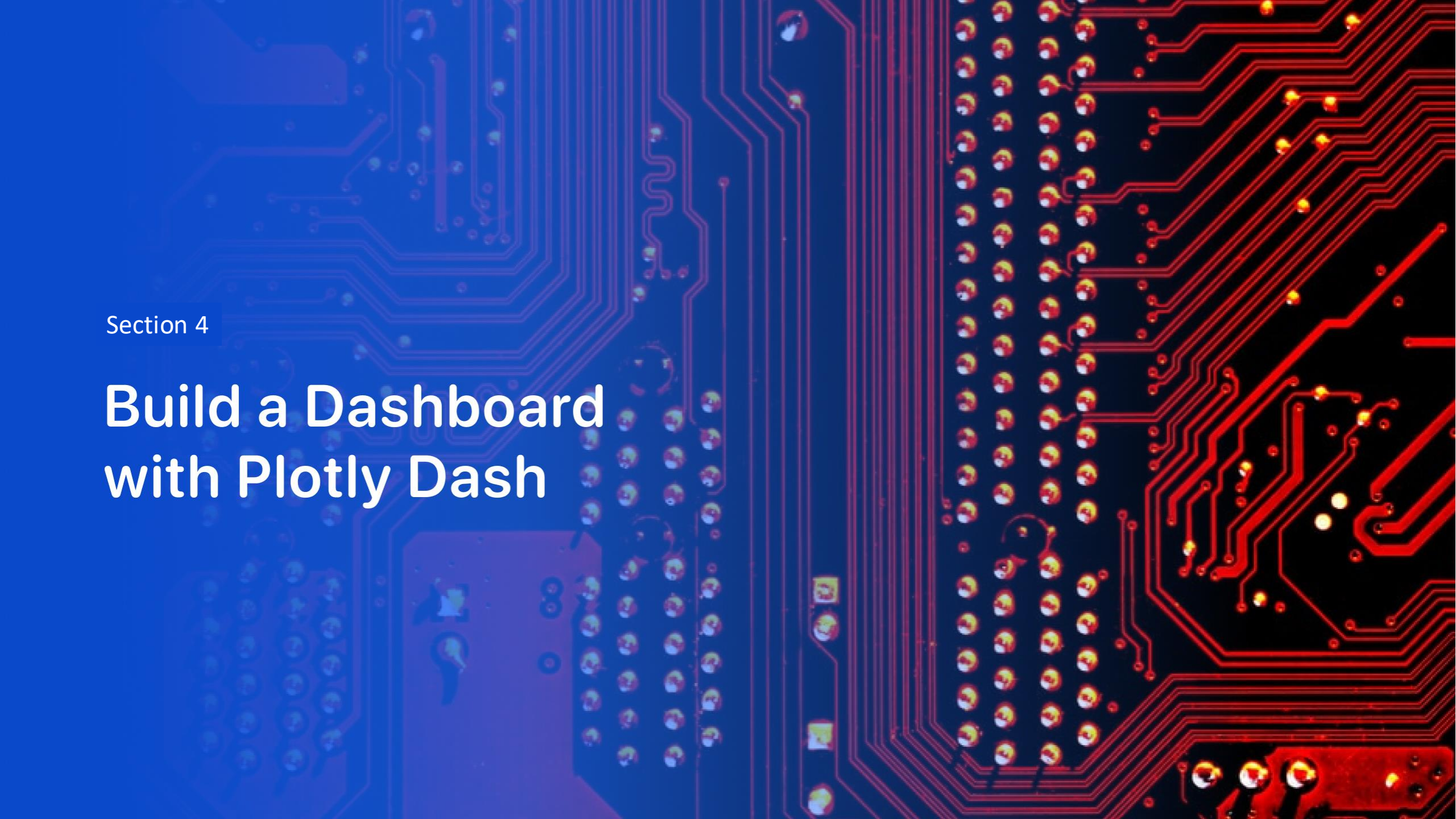
- **Key Findings & Elements:**
  - Success/Failure Distribution – High-success regions (e.g., Cape Canaveral) contrast with sites showing frequent failures, suggesting correlations with infrastructure or operational factors.
  - Geographic Clusters – Launch sites near equatorial zones (e.g., Kourou) exhibit higher success rates, likely due to favorable orbital mechanics.
  - Outlier Patterns – Isolated failure clusters may indicate site-specific risks (weather, technology).
- **Insight**: The map highlights the impact of location on launch outcomes, supporting data-driven site selection for future missions.



Launch Sites Proximities

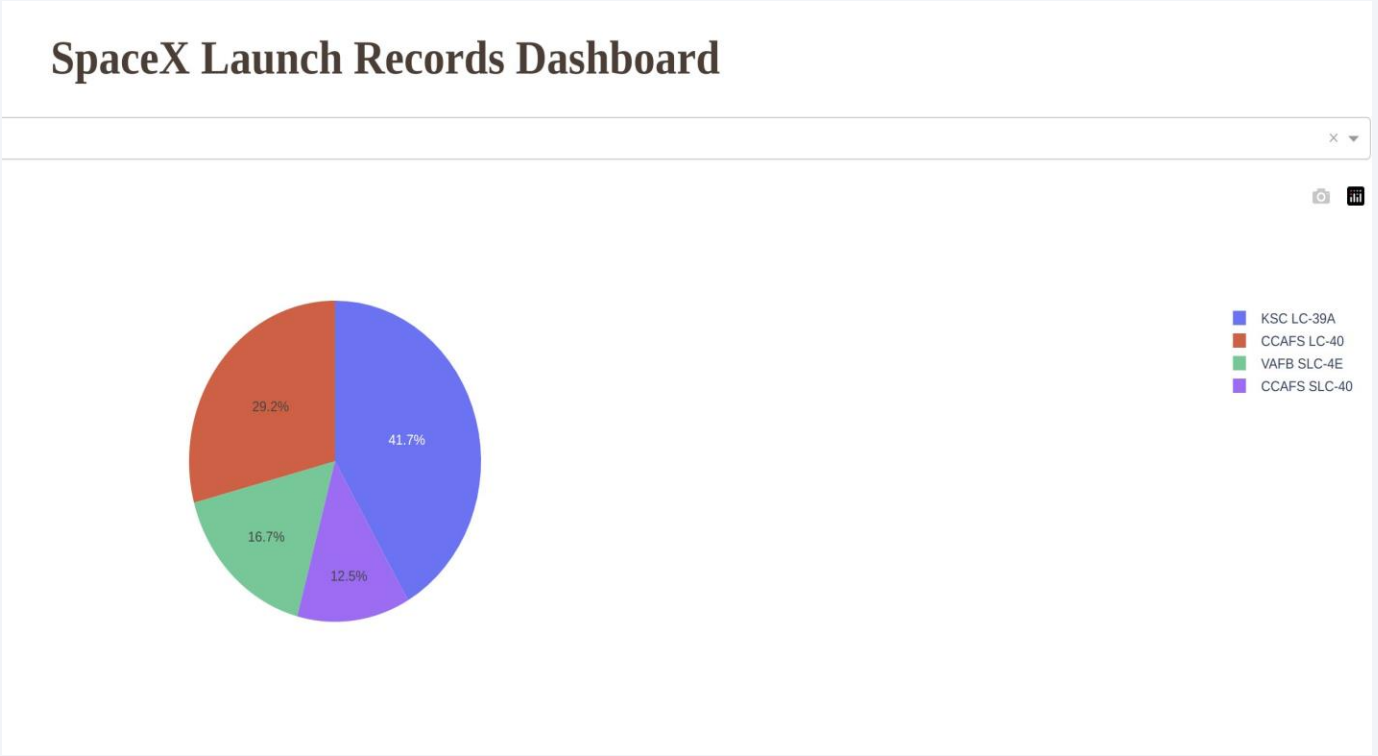
- **Key Criteria for Launch Site Location**
  - **Near Coastline**:
    - Safe launch trajectory over the ocean
    - Enables emergency water landings
    - Limits risk to people and infrastructure
  - **Near Highways**:
    Simplifies movement of personnel and equipment.
  - **Near Railways**:
    Essential for transporting large and heavy components.
  - **Far from Cities**:
    Reduces danger to densely populated areas in case of failure.




Section 4
Build a Dashboard with Plotly Dash

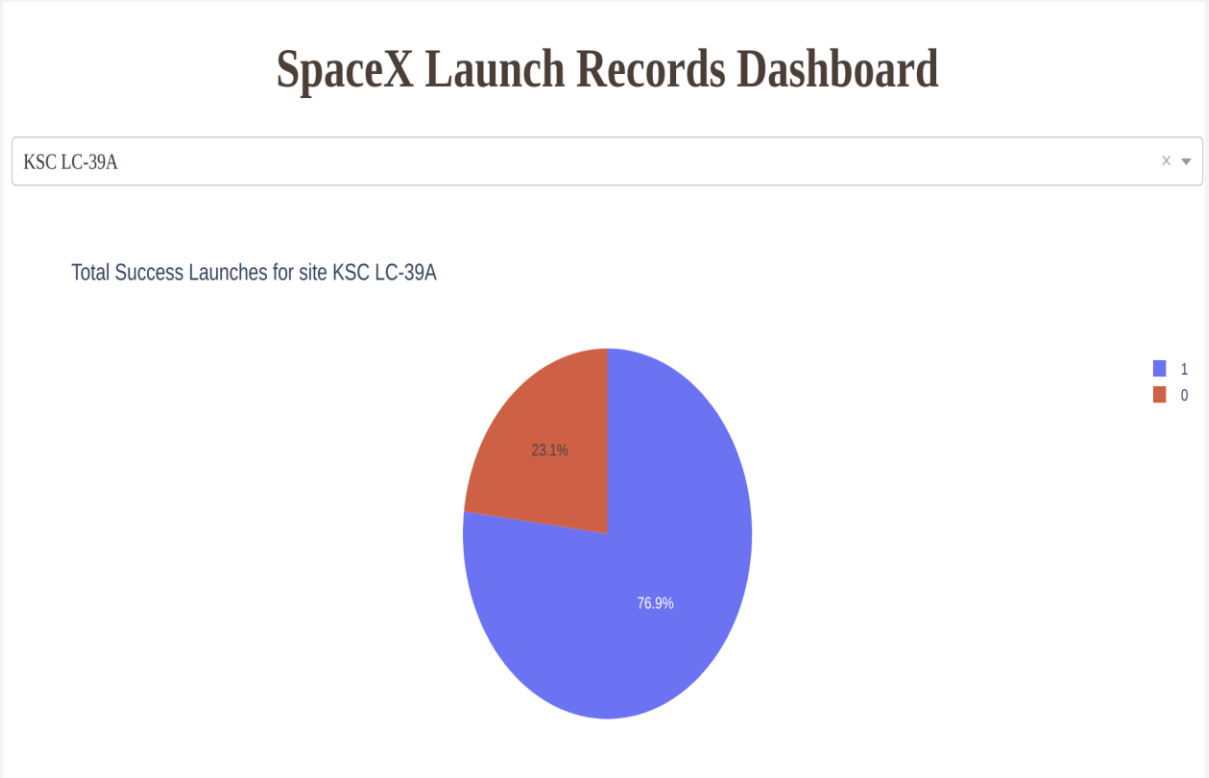
Launch Sites Success Rate

- **KSC LC-39A** launch site has the highest number of successful launches.
- **CCAFS LC-40** followed with a percentage of **29.2 %**
- **VAFB SLC-4E** and **CCAFS SLC-40** account for a smaller portions with **16.7%** and **12.5%** respectively.



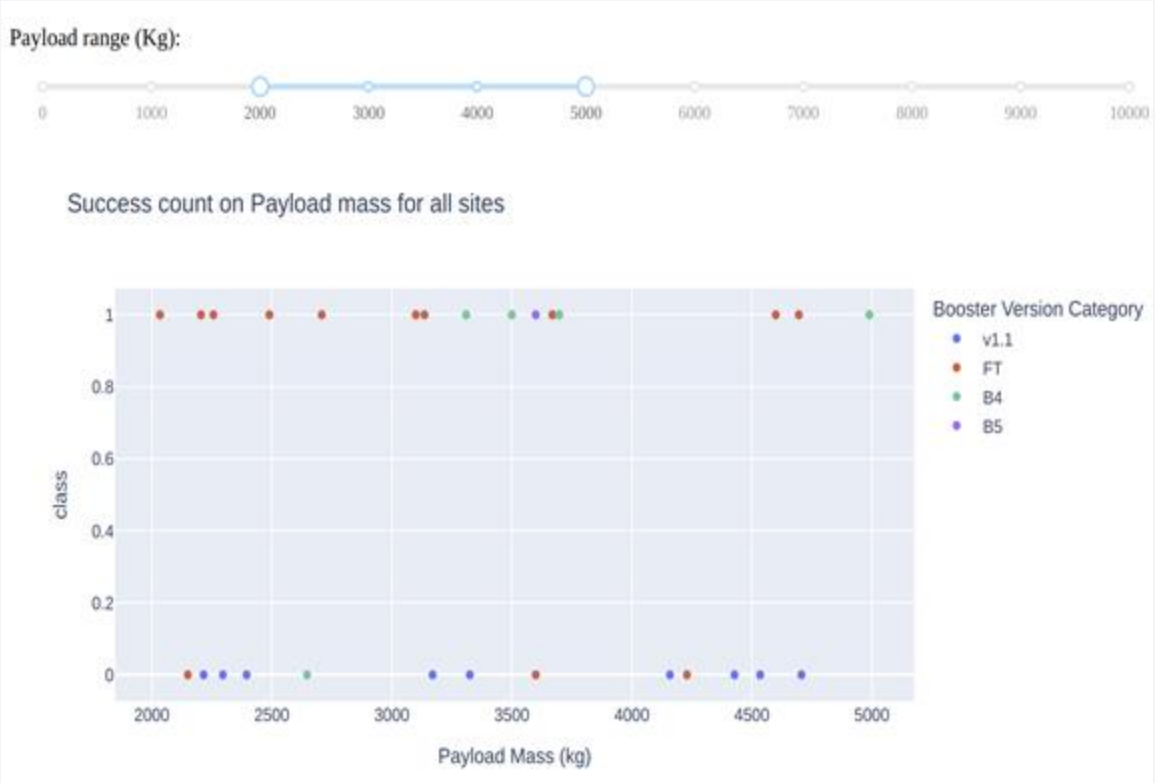
KSC LC-39 Launch Site Success Rate

- The dashboard reports launch success rates at the **KSC LC-39A** facility, a launch complex at NASA's Kennedy Space Center used by SpaceX.
- The pie chart shows the percentage of **successful (1)** and **failed (0)** launches at the facility.
- **76.9%** of the launches succeeded, with 23.1% failing, representing a reasonably high success rate but one which could be improved.



Payload vs. Launch Outcome

- Successes (y = 1) are dispersed across all payload ranges, but failures (y = 0) are more frequent with v1.1 and FT versions, especially in the 2000–4500 kg range.
- Subsequent booster versions like B4 and B5 have better success rates, even at heavier payloads, indicating increased reliability over time.



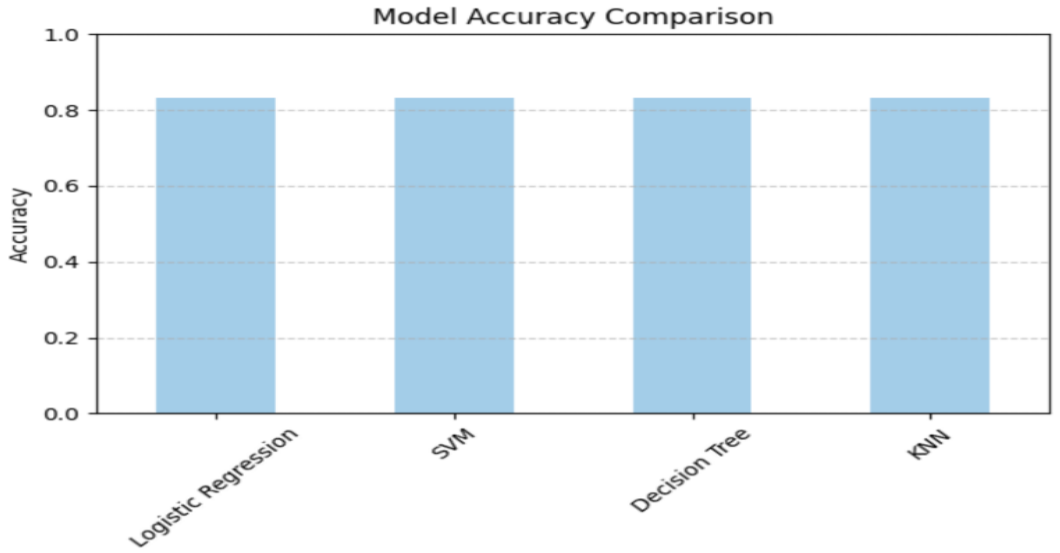
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- All the modest have the same accuracy of the test data
- For model selection different other metrics were used such as f1-score roc-auc. The standard error was also assessed.
- Based on the results KNN was selected as a best performing model

Model	F1-score	ROC AUC	CV Mean Accuracy	CV Std
Logistic Regression	0.889	0.889	0.8222	0.1133
SVM	0.889	N/A	0.8222	0.1018
Decision Tree	0.880	0.791	0.7556	0.1778
KNN	0.889	0.8958	0.8333	0.1139

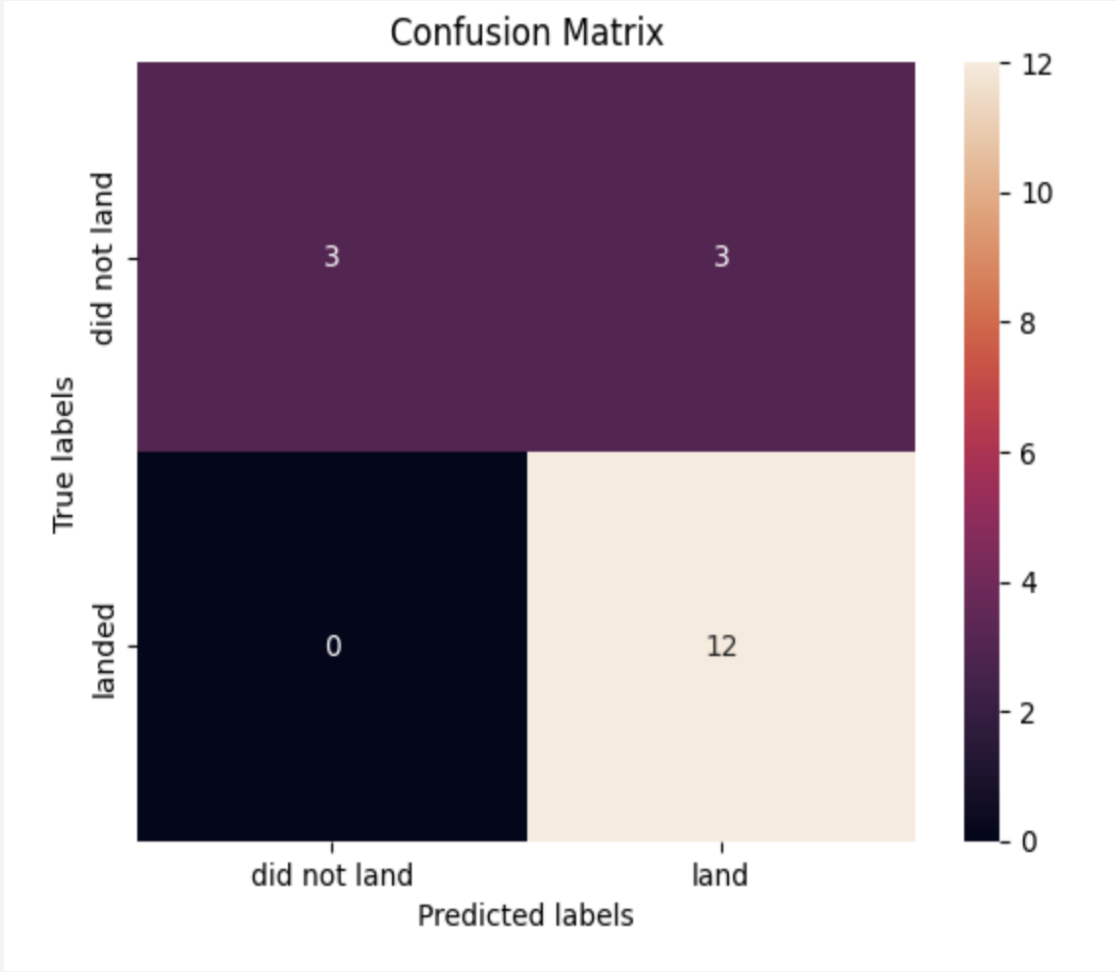


Confusion Matrix

The confusion matrix shows how well the classifier performed in predicting rocket landings. Out of 18 samples:

- **12** successful landings were correctly predicted.
- **3** failures were also correctly predicted.
- **3** failures were mistakenly classified as successful landings.

- No successful landings were misclassified as failures
- This suggests the model is reliable for detecting successful landings



Conclusions

- **Launch Patterns & Success Rates:**
  - Flight Experience Matters: Launch success improved with experience; initial years (2010–2013) had no successes, whereas post-2013 showed improvement.
- **Launch Site Insights:**
  - LC-39A at KSC had the highest success rate (76.9%) and most successful launches.
  - Coastal launch sites provided safety and logistical advantages.
- **Payload & Orbit Effects:**
  - Heavy payloads (>10,000 kg) had high success, especially for LEO, Polar, and ISS orbits.
  - GTO orbit had lowest success rate with no apparent correspondence between flight number and success.
  - ES-L1, GEO, HEO, and SSO orbits experienced a 100% success rate.
- **Modeling & Prediction:**
  - Both models performed alike with test data.
  - The choice of models made use of F1-score, ROC-AUC, and Standard Error.
  - K-Nearest Neighbors (KNN) was ranked the best among performing models.

Appendix

- Code snippets used to select best performing model

```
report = pd.DataFrame(data=[logreg_cv_score, svm_cv_score, tree_cv_score, knn_cv_score],
                      index=['Logistic Regression', 'SVM', 'Decision Tree', 'KNN'],
                      columns=['Accuracy'])
report
```

```
report.plot(kind='bar', legend=False, color='skyblue')
plt.title('Model Accuracy Comparison')
plt.ylabel('Accuracy')
plt.ylim(0, 1)  # Assuming accuracy is between 0 and 1
plt.xticks(rotation=45)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```

```
models = {
    "Logistic Regression": logreg_cv.best_estimator_,
    "SVM": svm_cv.best_estimator_,
    "Decision Tree":tree_cv.best_estimator_,
    "KNN": knn_cv.best_estimator_
}
```

```
from sklearn.metrics import precision_score, recall_score, f1_score, roc_auc_score

def evaluate_model(model, X_test, Y_test, name="model"):
    y_pred = model.predict(X_test)
    y_proba = model.predict_proba(X_test)[:,1] if hasattr(model, "predict_proba") else None

    precision = precision_score(Y_test, y_pred)
    recall = recall_score(Y_test, y_pred)
    f1 = f1_score(Y_test, y_pred)
    auc = roc_auc_score(Y_test, y_proba) if y_proba is not None else "N/A"

    print(f"{name}:")
    print(f"  Precision: {precision:.3f}")
    print(f"  Recall:    {recall:.3f}")
    print(f"  F1-score:  {f1:.3f}")
    print(f"  ROC AUC:   {auc}")
    print("-" * 30)


for name, model in models.items():
    evaluate_model(model, X_test, Y_test, name)
```

```
from sklearn.model_selection import cross_val_score

for name, model in models.items():
    scores = cross_val_score(model, X, Y, cv=10)
    print(f"{name}: mean={scores.mean():.4f}, std={scores.std():.4f}")
```

Thank you!