Winning Space Race with Data Science

Shedrack Chinonso
17 Aug 2025

Outline

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

Executive Summary

- Collected and pre-processed SpaceX launch data (e.g., payload mass, booster version, landing outcomes).

- Applied **exploratory data analysis (EDA)** with visualizations (bar charts, scatter plots, timelines).

- Used **statistical analysis** and **machine learning** models (Support vector Machine, KNN, logistic regression, decision trees) to predict successful landings.

- Validated models using **cross-validation** and accuracy scores.

Summary of Results

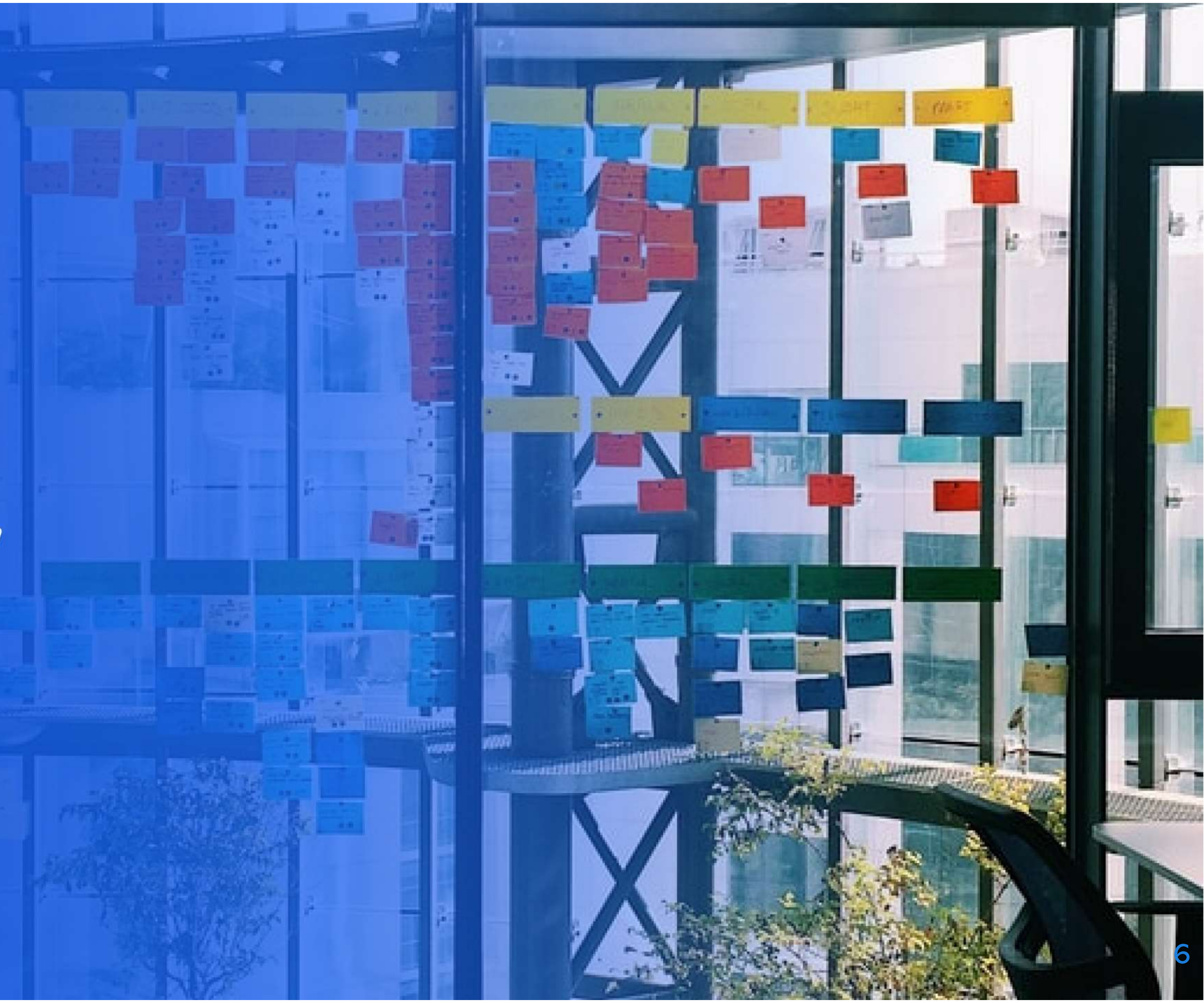
- Identified key factors influencing landing success: payload mass, booster version, and launch site.
- Achieved **83% prediction accuracy** with the best ML model (mention which performed best, e.g., Decision Tree or KNN).
- Found that success rates improved significantly after 2015 with booster upgrades.
- Overall trend: **increasing reliability of SpaceX launches**, supporting future reusability and cost efficiency.

Introduction

- SpaceX's ability to cut launch costs depends on the consistent recovery of rocket boosters, but success is influenced by multiple factors such as payload, booster type, and launch site.
- By analyzing historical SpaceX launch data with data science methods, we can uncover the drivers of landing success and build predictive models to support future missions.

Section 1

Methodology



Methodology

•Data was collected from the SpaceX API and publicly available datasets.

•Data wrangling was performed to clean, standardize, and prepare the records.

•The processed dataset was structured into model-ready tables with relevant features.

•Exploratory Data Analysis (EDA) was carried out using visualizations and SQL queries.

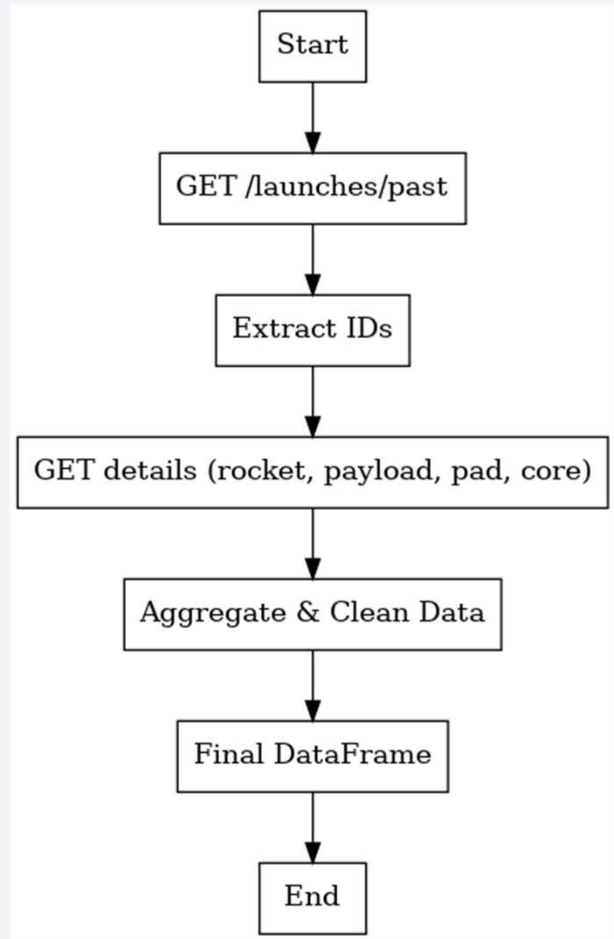
•Interactive visual analytics were developed using Folium maps and Plotly Dash dashboards.

•Predictive analysis was conducted with classification models to forecast landing success.

•Classification models were built, tuned, and evaluated using cross-validation and accuracy metrics

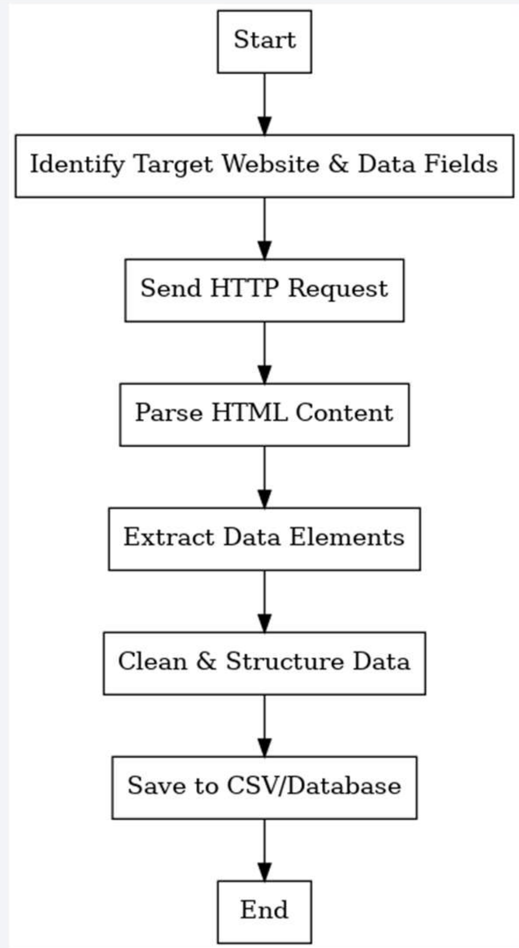
Data Collection – SpaceX API

- Data was collected from the SpaceX REST API using HTTP requests. Responses were retrieved in JSON format, then transformed into pandas Data Frames.
- After performing data cleaning and wrangling, the processed dataset was stored for analysis. The process is illustrated with a flowchart showing data extraction, transformation, and storage.
- The SpaceX API Calls and results in this Github Repository



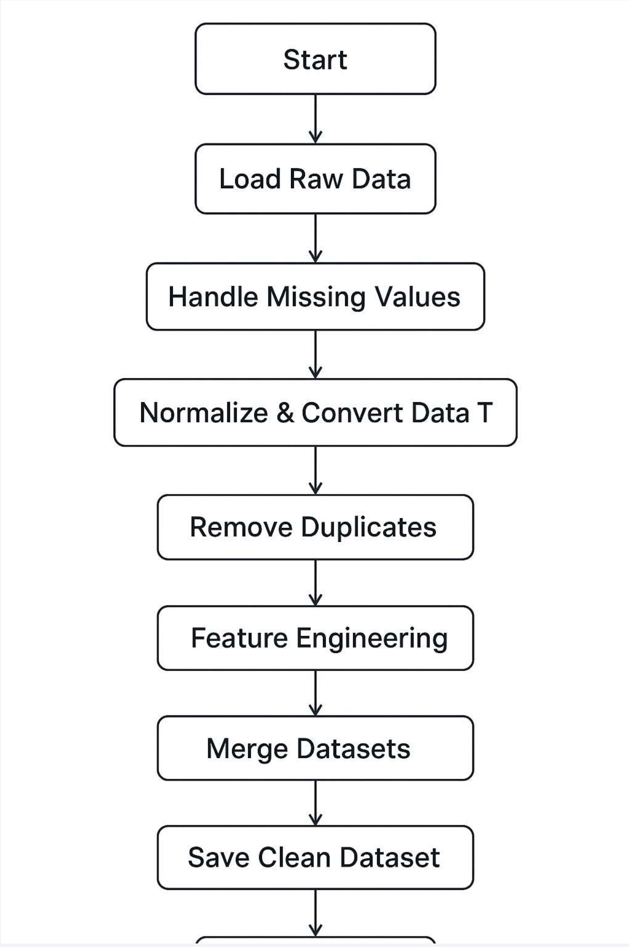
Data Collection - Scraping

- I identified target website and data fields
- Sent HTTP request to web page
- Parsed HTML with BeautifulSoup / lxml
- Extracted relevant elements (e.g., titles, links, tables)
- Cleaned and structured data into a table
- Stored data in CSV/Database for analysis
- The SpaceX Webscraping in this [Github Repository]



Data Wrangling

- Raw data were loaded into a Pandas DataFrame
- Missing values were handled (dropped, filled, or replaced)
- Column formats were normalized and standardized
- Data types were converted (e.g., dates, numbers)
- Duplicates and irrelevant fields were removed
- New features were engineered (e.g., landing success flag, payload bins)
- Data from multiple sources (API endpoints) were merged and joined
- The cleaned dataset was saved for analysis
- GitHub URL of the completed data wrangling related notebooks.



EDA with Data Visualization

• Bar Charts was used to compare categorical data such as launch outcomes across different Launch sites.

• Pie Charts was used to show proportions of successful vs. failed launches.

• Line Charts was used to visualize launch frequency and trends over time.

• Scatter Plots was used to identify relationships between payload mass and Flight Number.

• Maps (Folium) was used to show spatial distribution of launch sites and landing outcomes.

GitHub URL of the completed EDA with data visualization notebook

EDA with SQL

- Retrieved launch records with booster version and payload mass.
- Counted total number of successful vs. failed mission outcomes.
- Found the date of the first successful landing on a ground pad.
- Listed boosters with successful drone ship landings and payload mass between 4000–6000 kg.
- Calculated the average payload mass for a specific booster version (e.g., F9 v1.1).
- Identified booster versions that carried the maximum payload mass using a subquery with aggregation.

- GitHub URL of your completed EDA with SQL notebook

Build an Interactive Map with Folium

- Markers were placed at launch sites to indicate their exact geographic locations.
- Circles were drawn around launch sites to highlight proximity zones and area of interest.
- Lines were used to show distance between launch sites and their nearest cities or coastlines.
- Popups/Labels were attached to markers to display launch site names and key details when clicked.

These objects were added to make the map more interactive and to visually analyze how geography (site location, proximity, and distance) relates to SpaceX launch operations.

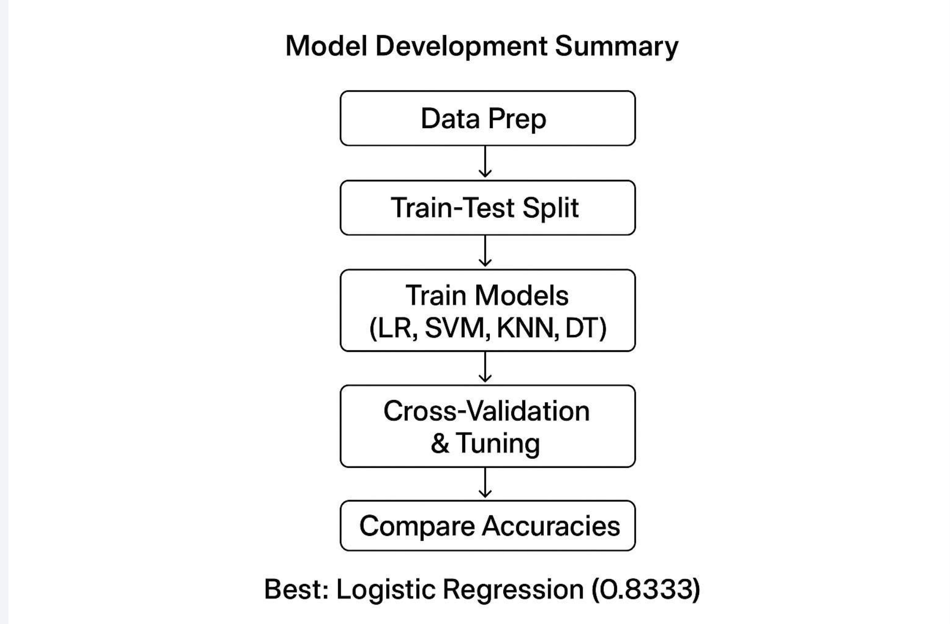
GitHub URL of the completed interactive map with Folium map

Build a Dashboard with Plotly Dash

- The dashboard includes a scatterplot to reveal relationships and trends between numerical variables with interactive filtering and zooming, and a pie chart to summarize categorical distributions like success vs. failure outcomes.
- These plots were chosen to balance detailed analysis with high-level insights, making the data both explorable and easy to interpret.
- GitHub URL of the completed Plotly Dash lab

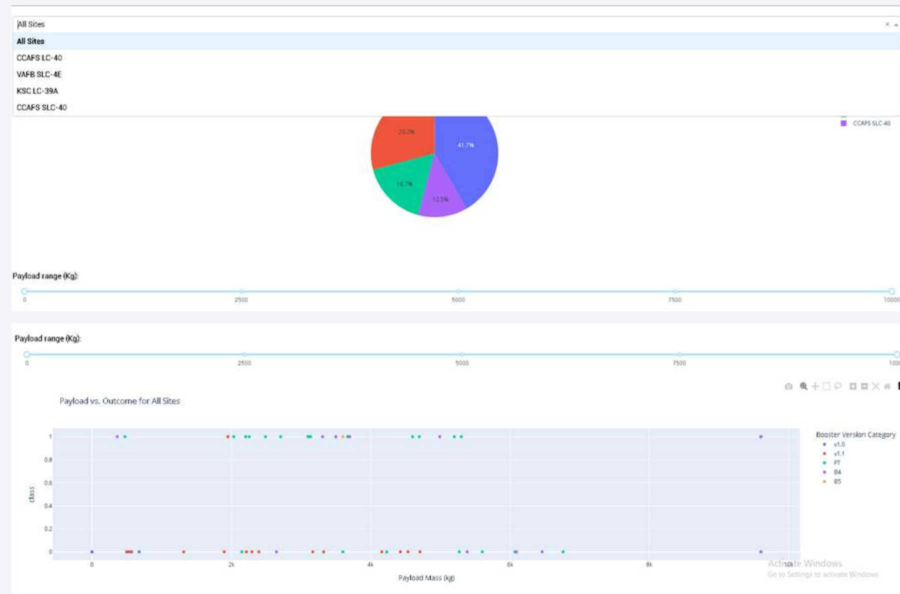
Predictive Analysis (Classification)

- Data Preparation: Cleaned, encoded, normalized, and split data into 80/20 train-test sets.
- Model Building: Trained Logistic Regression, SVM, KNN, and Decision Tree.
- Evaluation: Used 10-fold cross-validation and accuracy as the main metric.
- Improvement: Applied GridSearchCV for hyperparameter tuning.
- Best Model: Logistic Regression performed best with accuracy = 0.8333, slightly above SVM.
- GitHub URL of the completed predictive analysis lab



Results

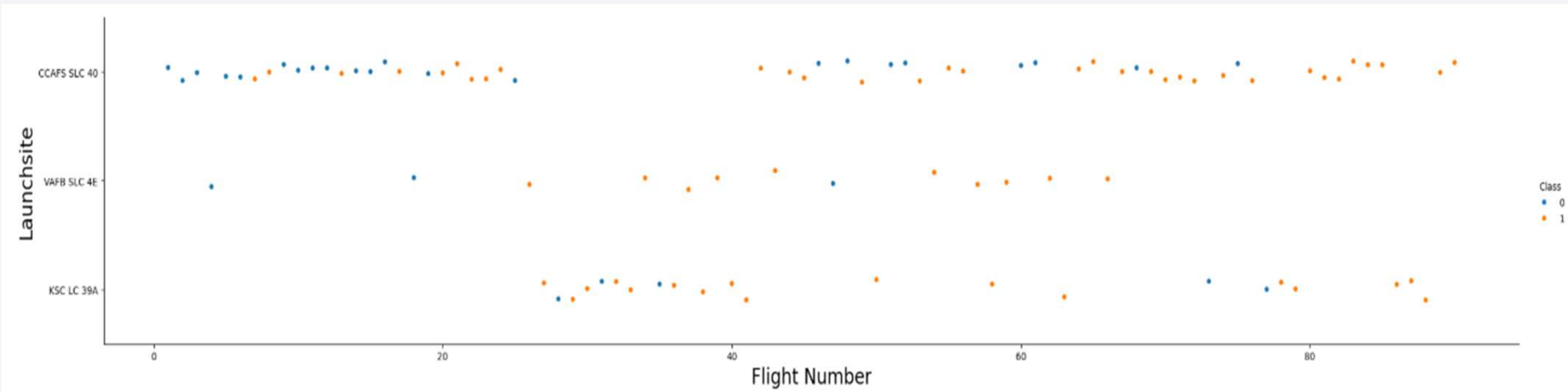
- **Exploratory Data Analysis (EDA) Results**
- Launch outcomes showed that a majority were successful, with a smaller proportion of failures.
- Payload mass distribution revealed most launches carried medium payloads (2,000–6,000 kg).
- Launch site analysis indicated some sites had significantly higher success rates than others.
- Booster version trend showed newer versions had higher success probabilities.
- **Predictive Analysis Results**
- Multiple models tested: Logistic Regression, SVM, KNN, Decision Tree.
- After cross-validation and tuning, Logistic Regression achieved the best accuracy: 0.8333.
- The model can reliably predict launch success probability given payload, site, and booster features.



Section 2

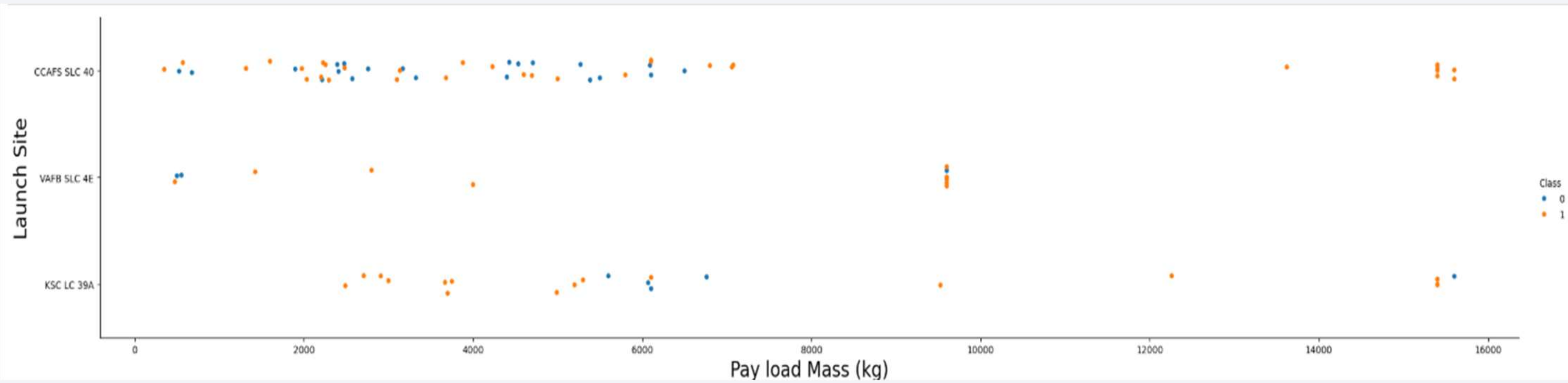
Insights drawn from EDA

Flight Number vs. Launch Site



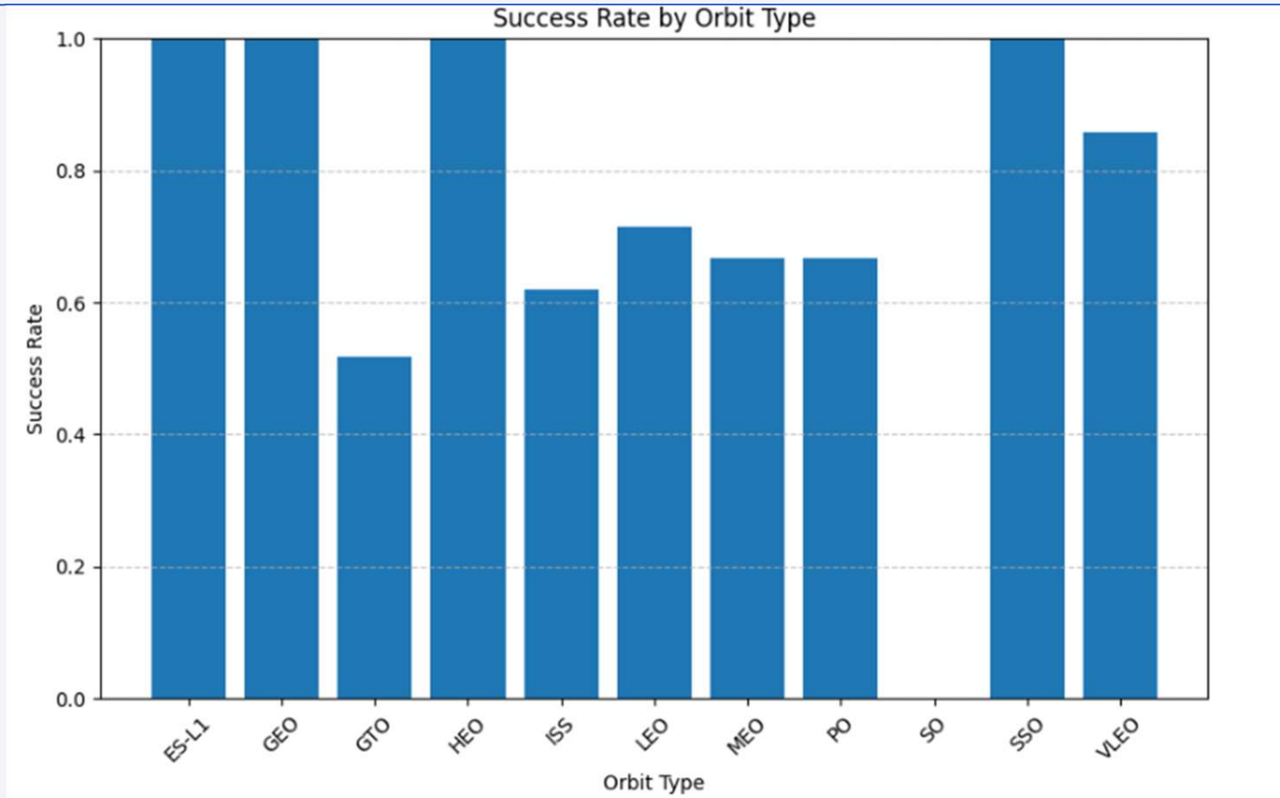
- The scatter plot shows that CCAFS SLC 40 had the most launches across all flight numbers, while KSC LC 39A was mainly used in later flights and VAFB SLC 4E had fewer launches overall. Both successes and failures occurred at all sites, but the success rate improved with higher flight numbers, reflecting growing reliability over time.

Payload vs. Launch Site



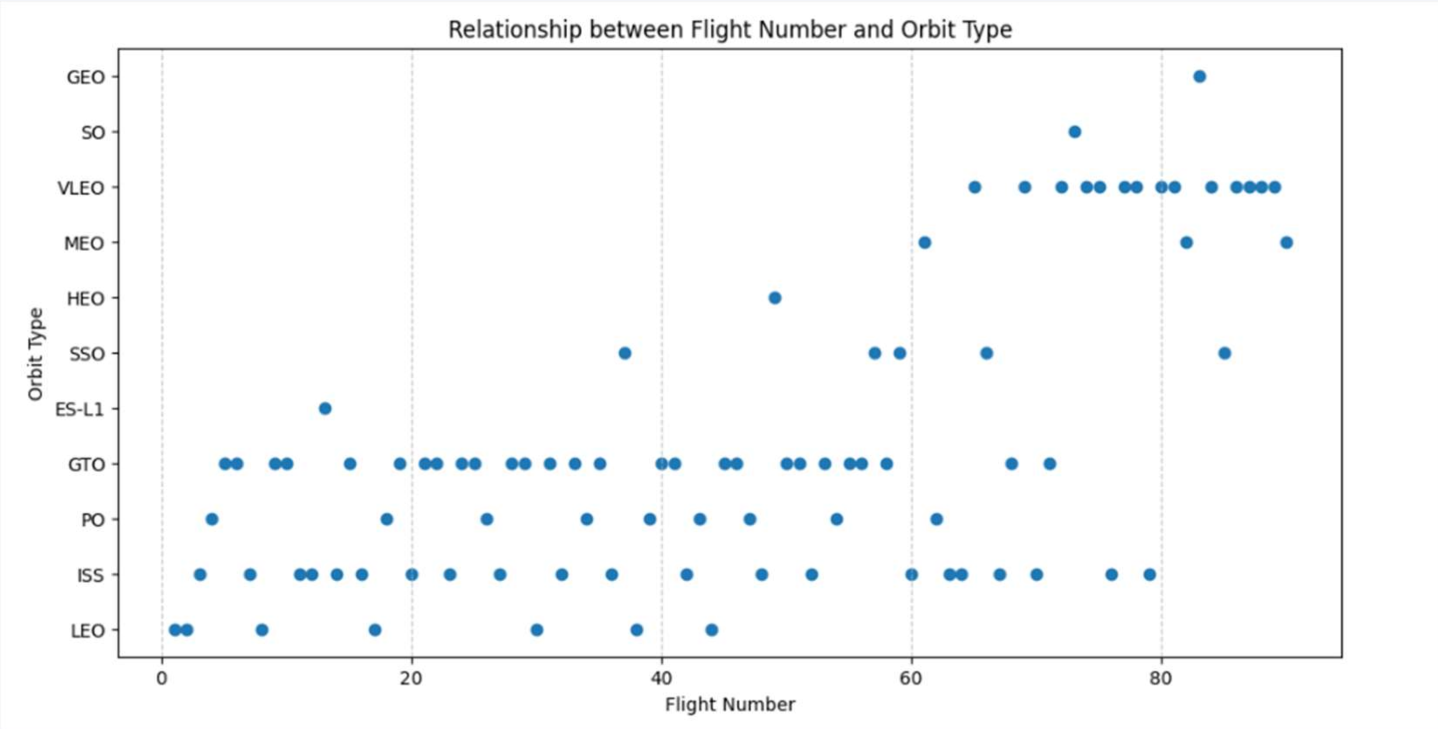
- The plot shows that CCAFS SLC 40 handled the widest range of payloads, while KSC LC 39A managed many of the heaviest missions and VAFB SLC 4E mostly lighter ones. Both successes and failures occurred at all sites, but heavier payload launches generally show more successes, reflecting improved reliability over time.

Success Rate vs. Orbit Type



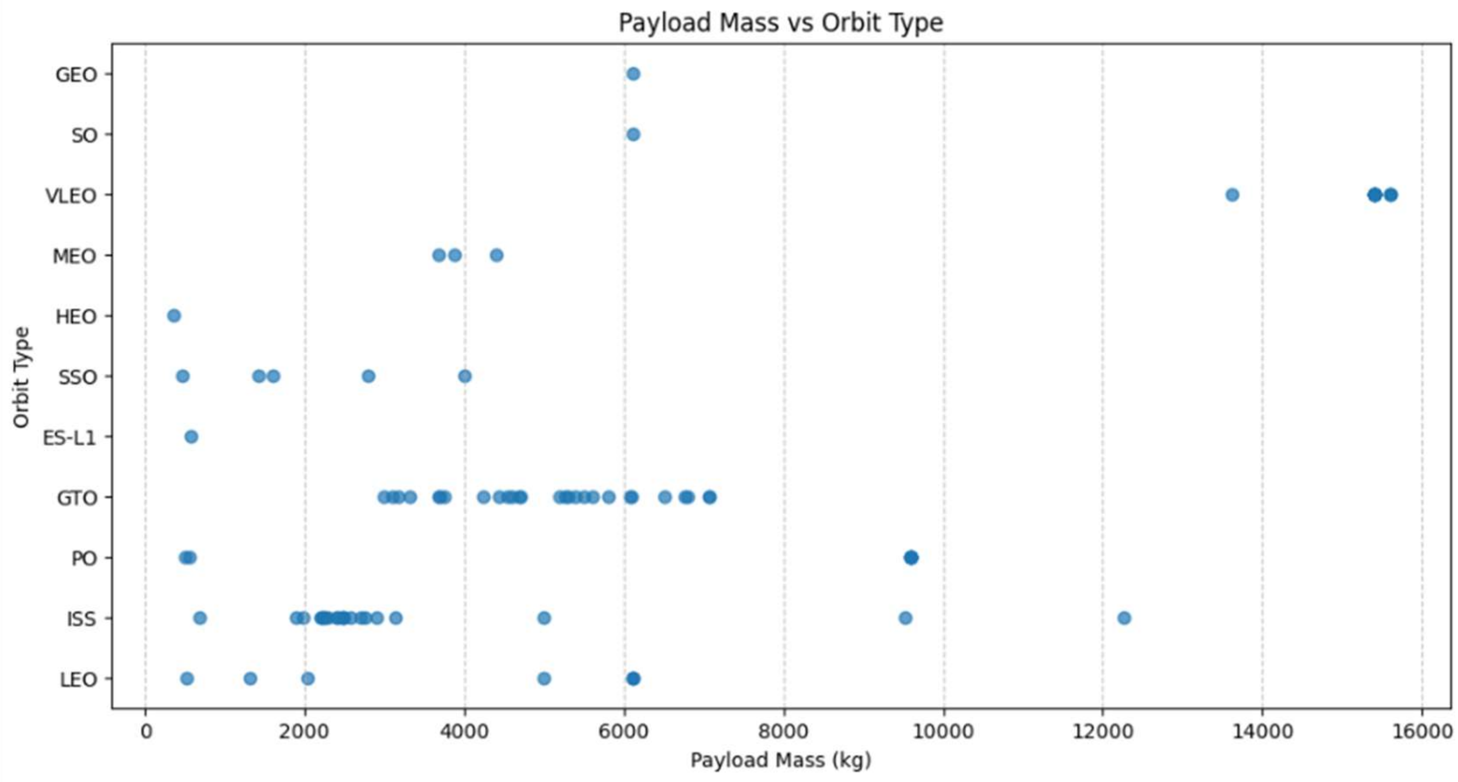
Orbits with 100% success rate: ES-L1, GEO, HEO, and SSO

Flight Number vs. Orbit Type



You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

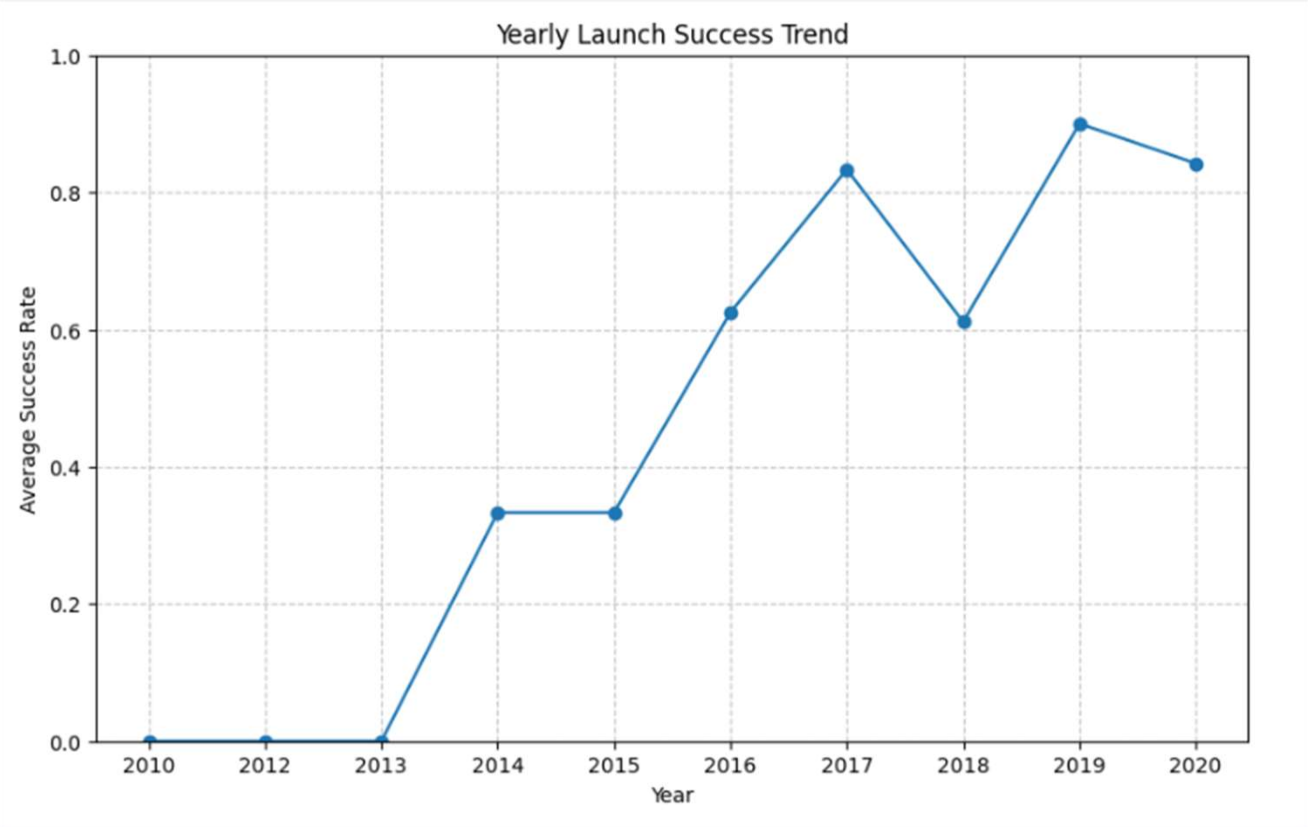
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



you can observe that the success rate since 2013 kept increasing till 2020

All Launch Site Names

- The unique launch sites in the dataset are:
- **CCAFS LC-40**
- **CCAFS SLC-40**
- **KSC LC-39A**
- **VAFB SLC-4E**

- These are the four different locations SpaceX has used for launches. "CCAFS" refers to Cape Canaveral Air Force Station in Florida, "KSC" is the Kennedy Space Center in Florida, and "VAFB" is Vandenberg Air Force Base in California.

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

The query filters launch sites that start with "CCA" (Cape Canaveral Air Force Station). These include **CCAFS LC-40** and **CCAFS SLC-40**. The first 5 rows returned show missions launched from these Cape Canaveral sites.

Total Payload Mass

Total_Payload_Mass
45596

This calculation shows how much payload mass (in kilograms) was launched on missions involving NASA boosters. It filters the dataset for boosters labeled with "NASA" and aggregates their payloads.

Average Payload Mass by F9 v1.1

Average_Payload_Mass
2928.4

This gives the typical payload size carried by **Falcon 9 v1.1** rockets. It helps compare the performance of this booster version to others like **F9 v1.0** or **F9 FT**.

First Successful Ground Landing Date

1]:

First_Successful_Ground_Pad_Landing
2015-12-22

This query found the earliest recorded date where a SpaceX booster successfully landed on a ground pad instead of a drone ship. That milestone was a turning point in reusable rocket technology.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

This query shows the booster versions that successfully landed on a drone ship during missions with **medium payloads (between 4000–6000 kg)**. These results highlight the reusable boosters that proved reliable under such mission conditions.

Total Number of Successful and Failure Mission Outcomes

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

The query counts how many missions ended in **Success** vs **Failure** This helps evaluate SpaceX's overall mission reliability. Historically, most missions were **successful**, with very few failures.

Boosters Carried Maximum Payload

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

The query identifies the booster version(s) that carried the **heaviest payload mass** in the dataset. This highlights SpaceX's most powerful rockets, showing which boosters achieved maximum lift capability.

2015 Launch Records

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

The result shows all missions in 2015 where SpaceX attempted a drone ship landing but failed. It lists the booster version used and the launch site for each failed mission, highlighting the learning curve before achieving consistent drone ship recoveries.

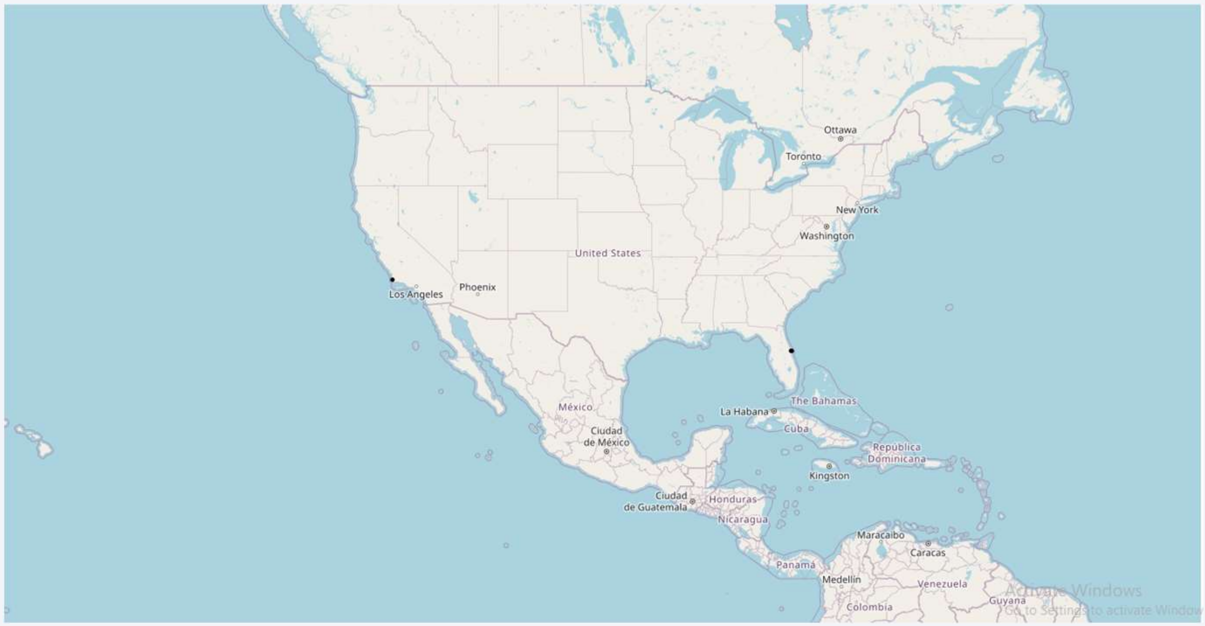
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Landing_Outcome	OutcomeCount
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

This ranking shows that in the early SpaceX missions (2010–2017), most launches did not attempt landings (since reusability trials only began in 2015). As landing technology matured, drone ship successes increased, though failures were common early on.


Section 3
Launch Sites
Proximities Analysis

All launch sites

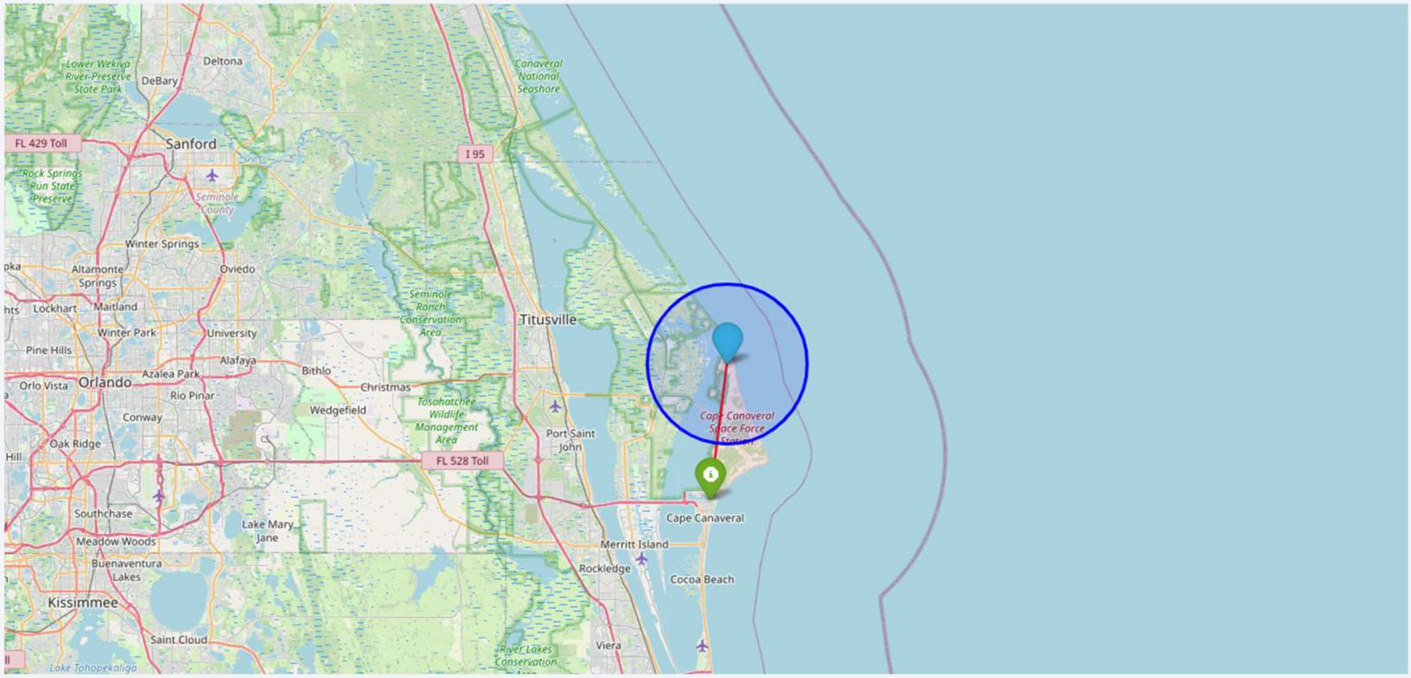


The Folium map shows all SpaceX launch sites marked with circles. indicates that none of the sites are near the Equator
they are mostly in mid-latitudes. However, all launch sites are located close to the coast, allowing rockets to safely launch over water. This placement reflects SpaceX’s priorities for safety and optimal launch trajectories.

Launch Outcomes

The marker cluster map shows all SpaceX launches, with marker colors indicating success or failure. Launch sites with more green markers have higher success rates, while sites with more red markers experienced more failures. Clustering makes it easy to visualize and compare launch site performance across regions.

Selected launch site proximities

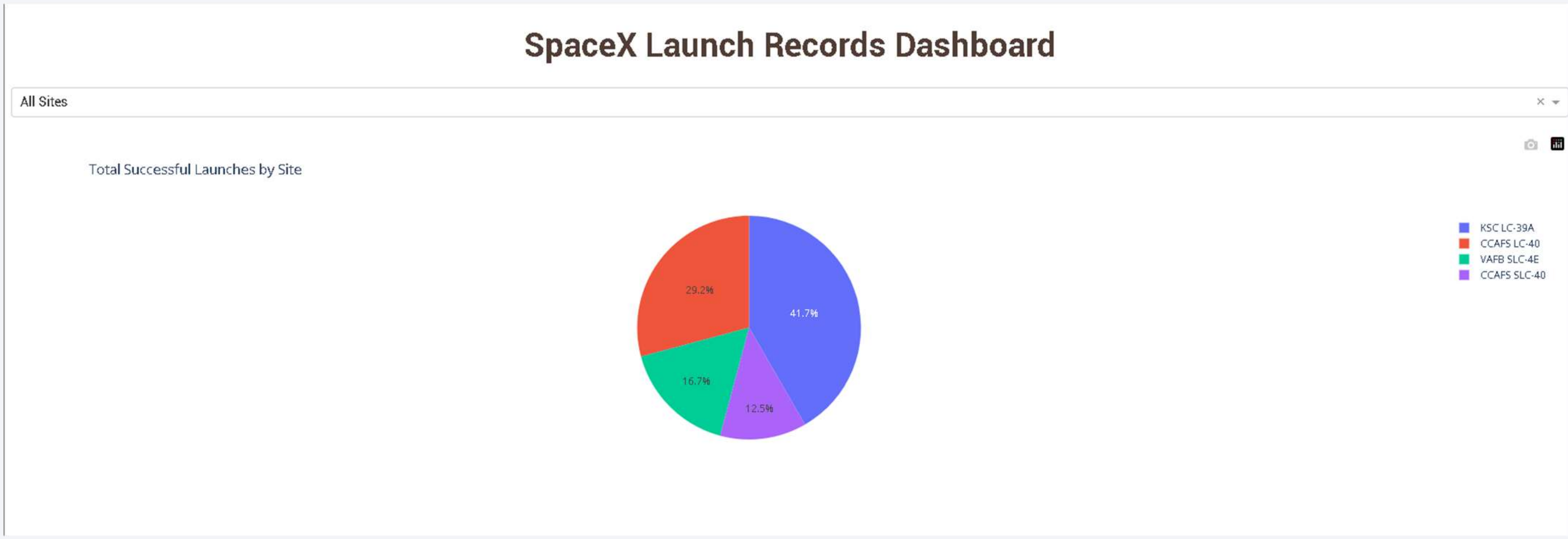


SpaceX launch sites are strategically located close to the coast for safe rocket trajectories while maintaining a safe distance from cities, highways, and railways. The map visualization with markers, lines, and circles clearly illustrates these proximity relationships.

Section 4

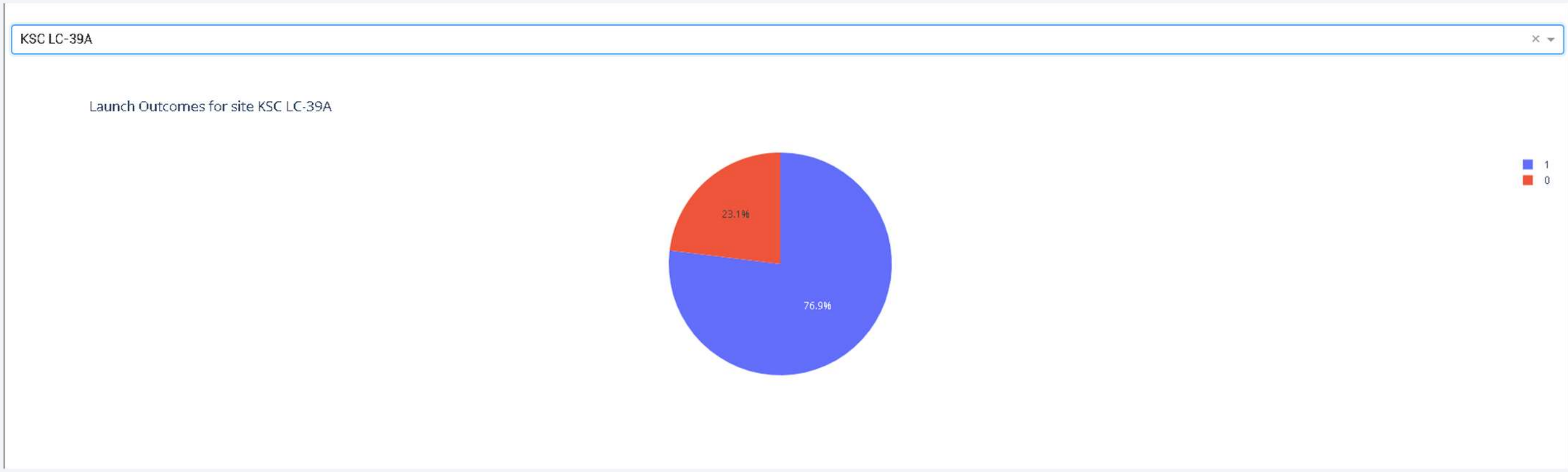
Build a Dashboard with Plotly Dash

SpaceX launch in All Site



The dashboard shows the share of SpaceX's successful launches by site using a pie chart. KSC LC-39A leads with the highest share (41.7%), followed by CCAFS LC-40 (29.2%), VAFB SLC-4E (16.7%), and CCAFS SLC-40 (12.5%). This highlights that Florida sites dominate SpaceX's successful launches.

Launch Site KSC LC-39A



The Launch Site KSC LC-39A has significant amount of successful launch of about 76.9% and unsuccessful launch of 23.1%

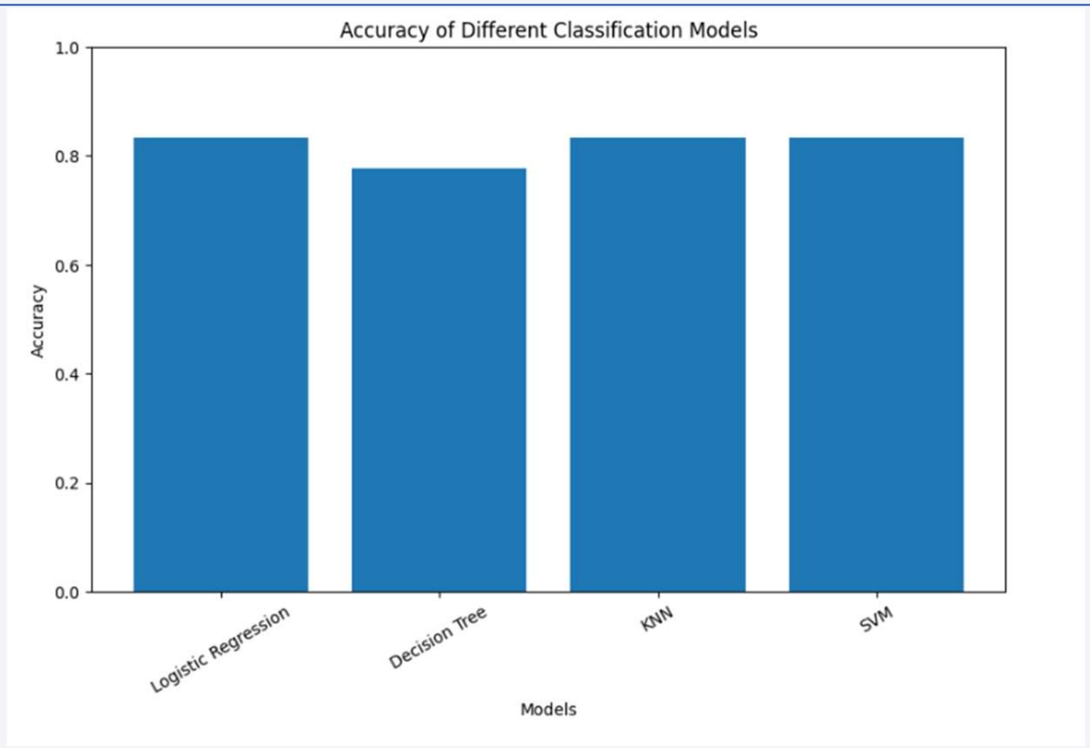
Outcome scatter plot for all sites

From the plot above, FT booster version recorded higher success rate up to 5000KG payload Mass

Section 5

Predictive Analysis (Classification)

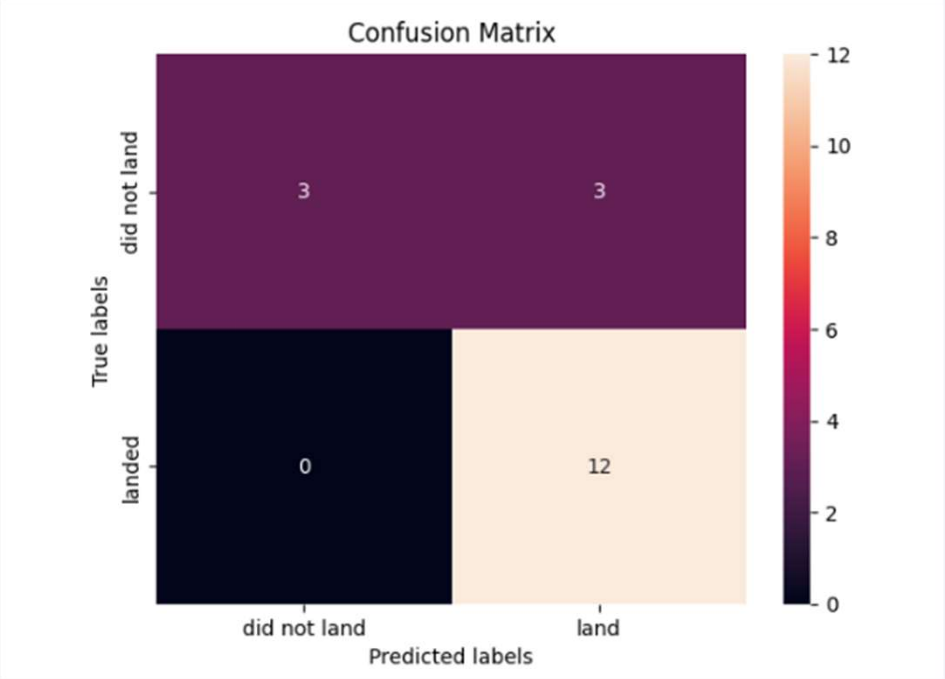
Classification Accuracy



Best performing method: Logistic Regression with accuracy: 0.8333333333333334

Confusion Matrix

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the problem is false positives.
- Overview:
- True Postive - 12 (True label is landed, Predicted label is also landed)
- False Postive - 3 (True label is not landed, Predicted label is landed)



Conclusions

- **Point 1:** The analysis of SpaceX launch data revealed clear patterns in launch success rates across different sites and payload ranges, highlighting which launch conditions are most favorable.
- **Point 2:** Classification models were successfully built to predict launch outcomes, with the best-performing model achieving high accuracy, demonstrating the power of data-driven approaches in aerospace operations.
- **Point 3:** Visualization tools, such as scatter plots, bar charts, piechart and line graphs, effectively illustrated relationships between payload mass, launch sites, and success outcomes, aiding in intuitive understanding of the dataset.
- **Point 4:** The project underscores the importance of predictive analytics in space missions, allowing stakeholders to identify risk factors, optimize launch planning, and improve mission success probability.
- **Point 5:** Future work could incorporate additional variables, such as weather conditions, booster reuse, and real-time telemetry, to enhance predictive accuracy and provide more comprehensive insights for SpaceX operations.

Appendix

Load the dataframe

Load the data

[51]:
```
from js import fetch
import io

URL1 = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_2.csv"
resp1 = await fetch(URL1)
text1 = io.BytesIO((await resp1.arrayBuffer()).to_py())
data = pd.read_csv(text1)
```

[52]:
```
data.head()
```

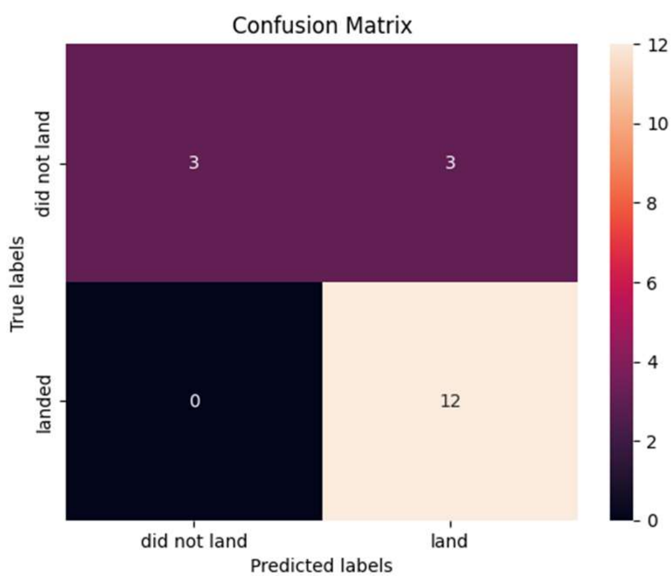
Appendix

```
# Accuracy on the test data
test_accuracy = logreg_cv.score(X_test, Y_test)
print("Test set accuracy:", test_accuracy)
```

```
Test set accuracy: 0.8333333333333334
```

Lets look at the confusion matrix:

```
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Appendix

SQL Queries

- **Title:** Example SQL Queries
- **Content:**
- Count launches by site:

SELECT Launch_Site, COUNT(*) AS Launch_Count

FROM spacex_launches

GROUP BY Launch_Site

ORDER BY Launch_Count DESC;

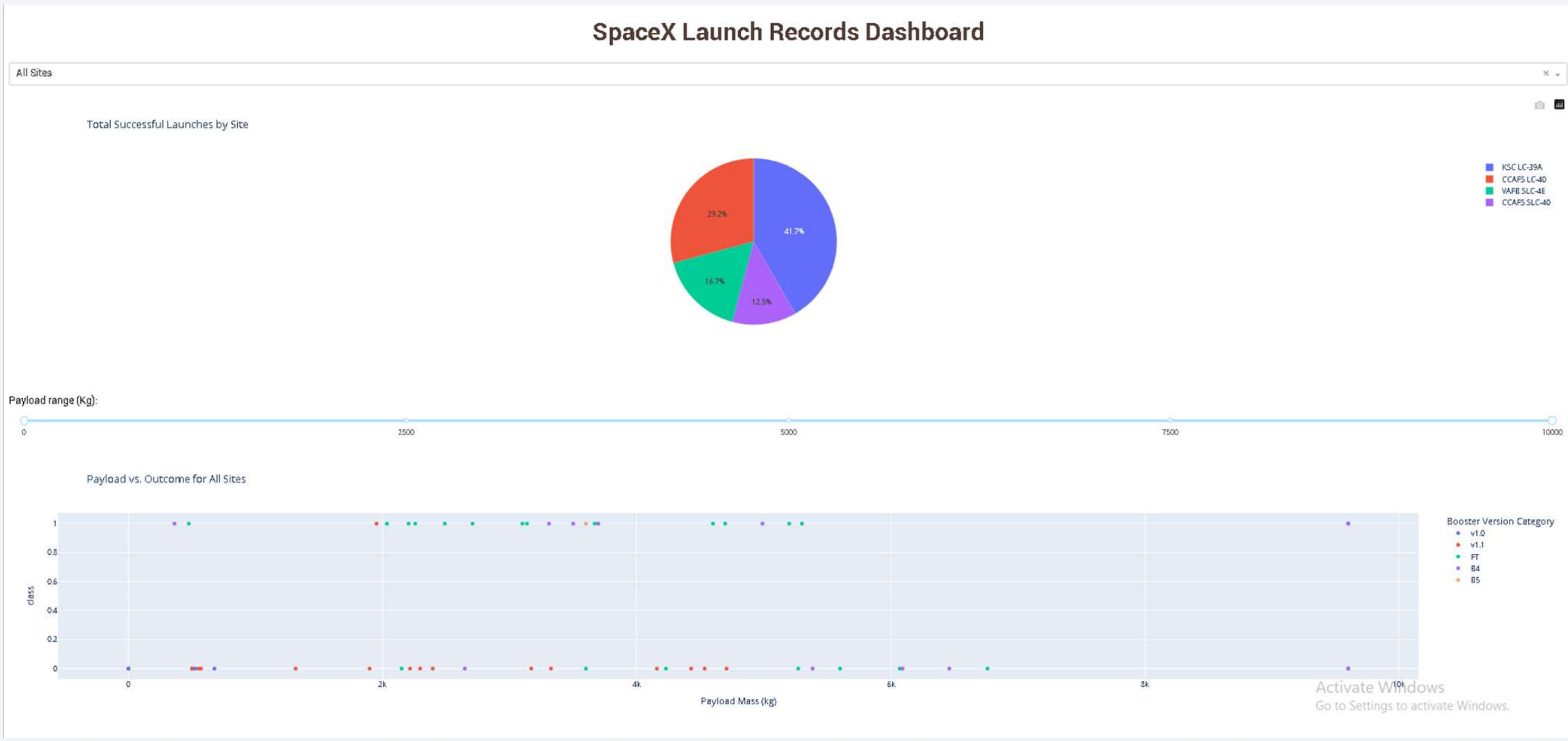
- Count successes and failures per site:
- SELECT Launch_Site, SUM(CASE WHEN Launch_Outcome = 'Success' THEN 1 ELSE 0 END) AS Success_Count,

SUM(CASE WHEN Launch_Outcome = 'Failure' THEN 1 ELSE 0 END) AS Failure_Count

FROM spacex_launches

GROUP BY Launch_Site;

Appendix



SpaceX Launch Records Dashboard
All Sites
Total Successful Launches by Site
41.7%
29.2%
16.7%
12.5%
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
Payload range (Kg):
0
2500
5000
7500
10000
Payload vs. Outcome for All Sites
class
1
0.8
0.6
0.4
0.2
0
0
2k
4k
6k
8k
10k
Payload Mass (kg)
Booster Version Category
v1.0
v1.1
FT
B4
B5
Activate Windows
Go to Settings to activate Windows.


Thank you!