IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Emmylou Bader
05.08.2025

Outline

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

Executive Summary

This project analyzes SpaceX launch data to explore the factors influencing landing success rates. We used exploratory data analysis to gain insights into key factors from the SpaceX data. Using interactive visualizations and geographical maps, we examined launch sites, payload ranges, and booster versions. The Kennedy Space Center showed the highest success rate, while the Vandenberg Space Force Base had the lowest. A machine learning model was built to predict landing outcomes based on payload mass, booster version, and launch site. Among the tested algorithms, Logistic Regression and Support Vector Machines achieved the highest accuracy. The results provide valuable insights into SpaceX's reusable rocket technology and support data-driven decision-making for future launches.

Introduction

- SpaceX has fundamentally changed the aerospace industry by developing reusable first-stage rockets. This innovation significantly reduces launch costs and gives the company a major competitive edge.
- For competitors, it is crucial to understand how reliable these reusable systems truly are. The success rate of launches directly affects pricing models and market access.
- By analyzing historical SpaceX launch data, we aim to assess mission outcomes and identify factors that influence launch success — especially by exploring correlations between key variables such as payload mass, launch site, and booster recovery.
- From a competitor's perspective, these insights could help improve launch reliability, optimize operational strategies, and ultimately increase competitiveness in a market dominated by SpaceX.

Section 1

Methodology

Methodology

Executive Summary

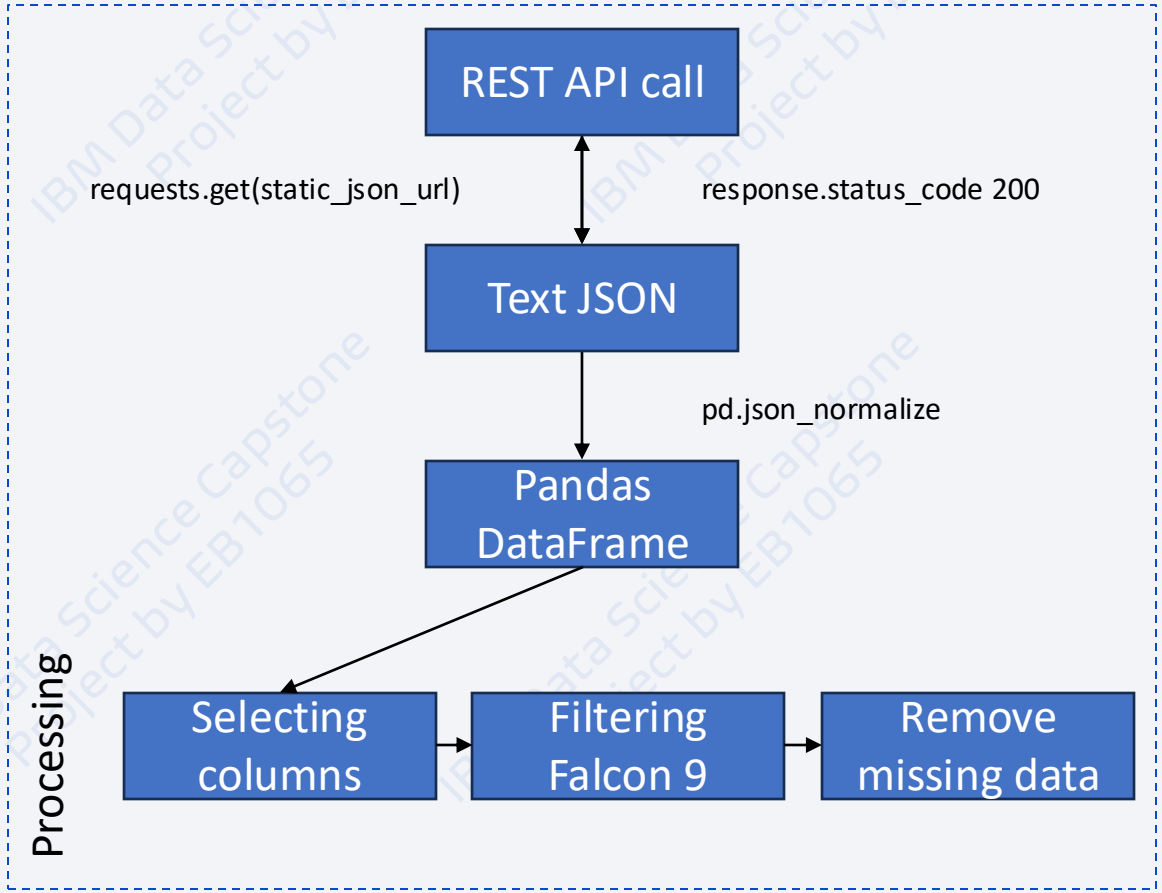
- Data collection methodology:
  - Collecting the data via the SpaceX API (GET request/JSON response) and web scraping from Wikipedia
- Perform data wrangling:
  - Normalizing and cleaning the data (including only Falcon 9 launches), removing missing values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardizing the data, splitting for training and testing, tuning with best fitting hyperparameters, comparing different models for best prediction results

Data Collection

- The first data source is the SpaceX REST API for past launch data with endpoint: https://api.spacexdata.com/v4/launches/past
- The data here was collected via API call (GET request) and returned text in a JSON format.
- The data contains information about booster type, payload mass, success/failure of the launch and flight number.
- The second data source is Wikipedia: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- These tables contain more Falcon 9 specific information and were collected via web scraping with Python BeautifulSoup.
- It contains information about landing outcome, customer and flight number for example.

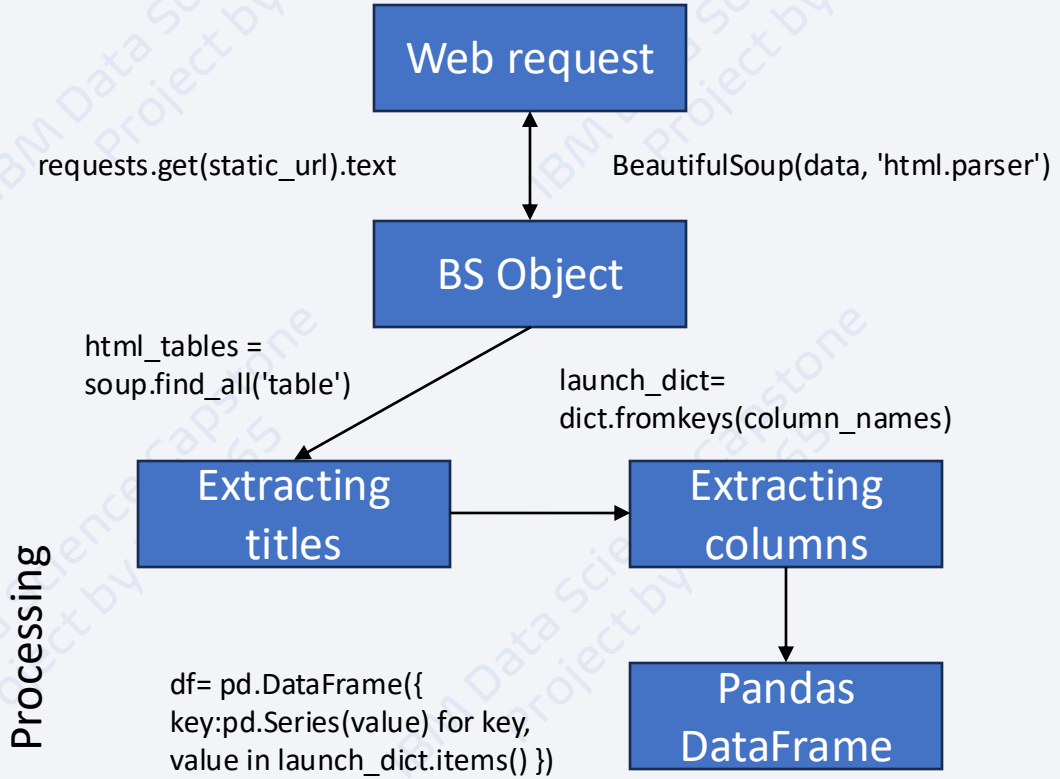
Data Collection – SpaceX API

- For the project, SpaceX launch data was collected programmatically using the official REST API with requests library. The GET request to the endpoint returned launch data in JSON format, which included mission outcomes, payload mass, launch dates, and more. The data was normalized using Pandas to load it into a DataFrame for further processing and analysis (e.g. filtering, dealing with missing data).
- Find the Jupyter notebook here: https://github.com/EB1065/datascience/blob/0c450368080a2a0fdfcb6da15e278a40a2f1c16a/jupyter-labs-spacex-data-collection-api.ipynb

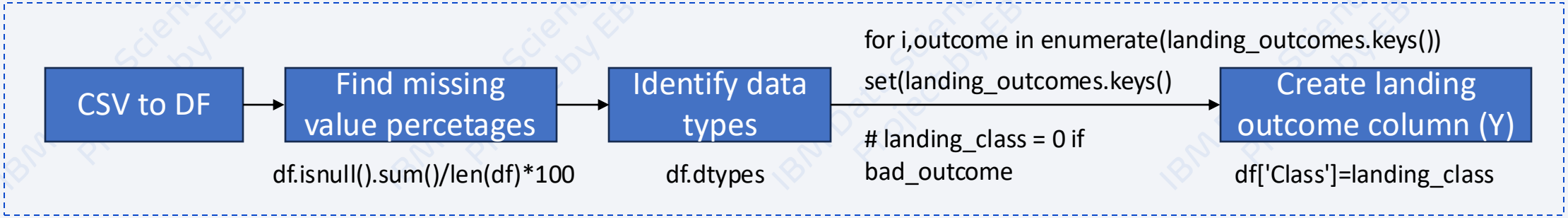


Data Collection - Scraping

- For the project, SpaceX launch data was collected from Wikipedia tables using Web Scraping. BeautifulSoup library was used to parse the text for data. By identifying the table structures in the HTML text the column headings were extracted and used to extract the data in the columns to a dictionary that was then written into a Pandas DataFrame.
- Find the Jupyter notebook here: https://github.com/EB1065/datascience/blob/0c450368080a2a0fdfcb6da15e278a40a2f1c16a/jupyter-labs-webscraping.ipynb



Data Wrangling



- Operational goals: first exploratory data analysis for data insights, determining training labels for machine learning
- The prepared SpaceX launch data was first loaded into a Pandas DataFrame.
  We analyzed the percentage of missing values and determined the data types of each column.
  Categorical fields such as launch sites and orbit types were reviewed to identify these as possible categories. For machine learning preparation, we engineered a new column called Class to represent landing success numerically: A predefined set of landing outcomes was classified as *unsuccessful* and mapped to 0, while all other outcomes were considered *successful* and mapped to 1.
- Find the Jupyter notebook here:
  https://github.com/EB1065/datascience/blob/633ecfafe1f7e1ff38225728cc2f437da58766a4/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

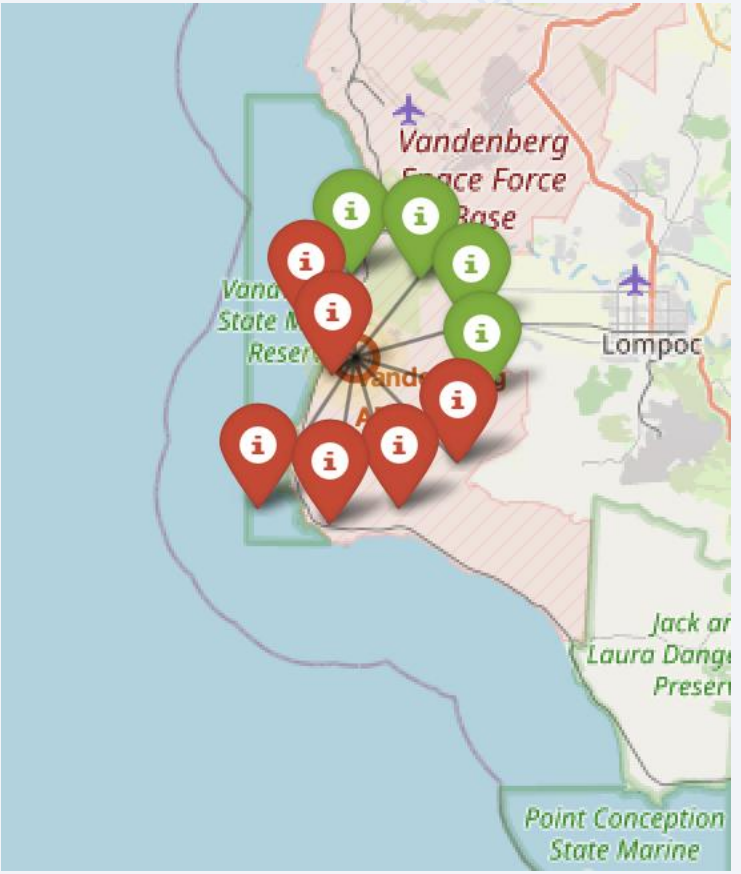
- Operational goal: Exploring correlations within the data to support effective feature selection and preparation.
- For this purpose the following diagrams were plotted:
  - Flight Number vs. Payload Mass (kg) (Scatter plot)
  - Flight Number vs. Launch Site (Scatter plot)
  - Payload Mass (kg) vs. Launch Site (Scatter Plot)
  - Orbit Type vs. Success Rate (Bar plot)
  - Orbit Type vs. Flight Number (Scatter plot)
  - Payload Mass (kg) vs. Orbit Type (Scatter plot)
  - Success Rate by Launch Year (Line plot)
- After selecting the relevant features One-hot Encoding was applied to make all values numerical and they were converted to Float64.
- Find the Jupyter notebook here: https://github.com/EB1065/datascience/blob/7403c9197f716e6d39e105df26b44e58df951918/edadataviz.ipynb

EDA with SQL

- Operational goal: Deriving insights through targeted SQL queries on the SpaceX dataset.
- The following SQL queries were performed on the SpaceX dataset:
  - Displaying the unique names of all launch sites
  - Displaying 5 records where launch sites begin with the string 'CCA'
  - Displaying the total payload mass carried by boosters launched by NASA (CRS) (45596 kg)
  - Displaying average payload mass carried by booster version F9 v1.1 (2534.7 kg)
  - Listing the date when the first successful landing outcome in ground pad was achieved (2015-12-22)
  - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes (61/40)
  - Listing all the Falcon 9 booster versions that have carried the maximum payload mass
  - Listing records with month names, with failure landing outcomes in drone ship, booster versions, launch sites, for all months in the year 2015
  - Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20
- Find the Jupyter notebook here: https://github.com/EB1065/datascience/blob/6f27bf633ec99560f0fb172b5063f9243a637404/jupyter-labs-eda-sql-coursera_sqllite.ipynb

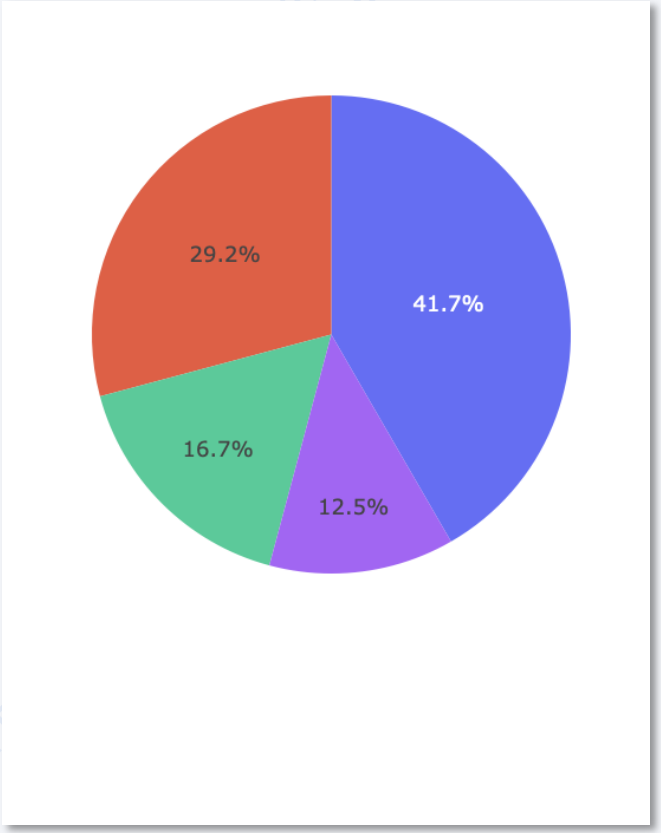
Build an Interactive Map with Folium

- Operational goal: Visualizing the data and geographical distribution and surroundings of the different SpaceX launch sites using an interactive Folium map.
- We added circles for launch sites, color-coded markers for launches and marked points of interest with lines from the launch site.
- This presents the location of the launch sites in global context, the successful and failed launches and the importance of appropriate infrastructure around a successful launch site. This improves cognitive comprehension of spatial and operational relationships.
- Find the Jupyter notebook here: https://github.com/EB1065/datascience/blob/36f3b1d9f368fc0fe15adb9b21a8761685876a3d/lab_jupyter_launch_site_location.ipynb

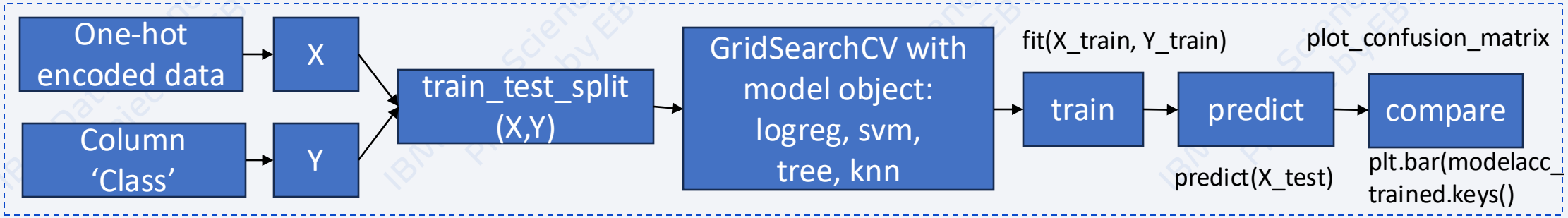


Build a Dashboard with Plotly Dash

- The dashboard provides interactive insights into SpaceX launch performance. Users can select either all launch sites or a specific site via a dropdown menu.
  - A pie chart displays the distribution of successful launches — either across all sites or for the selected launch site only.
  - A scatter plot visualizes the correlation between payload mass and launch outcome, dynamically adapting based on the selected site.
  - Additionally, a slider allows users to filter the data by payload range, refining the view and enabling more targeted analysis.
- This interactive setup allows users to explore launch success patterns and payload impacts in an intuitive and visually accessible way.
- Find the Python file here: https://github.com/EB1065/datascience/blob/82a3ba154a839af4b0bf4ed3667d949b6b7ad1e4/spacex-dash-app.py



Predictive Analysis (Classification) (1)



- Operational goal: Predict the success of future (SpaceX) launches using classification models, enabling data-driven decision-making and risk assessment for upcoming (own) missions.
- Find the Jupyter notebook here: https://github.com/EB1065/datascience/blob/22874e0610ef544f6acecb35b28f36c6279d8925/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb
- Read the complete description of the process on the next slide

Predictive Analysis (Classification) (2)

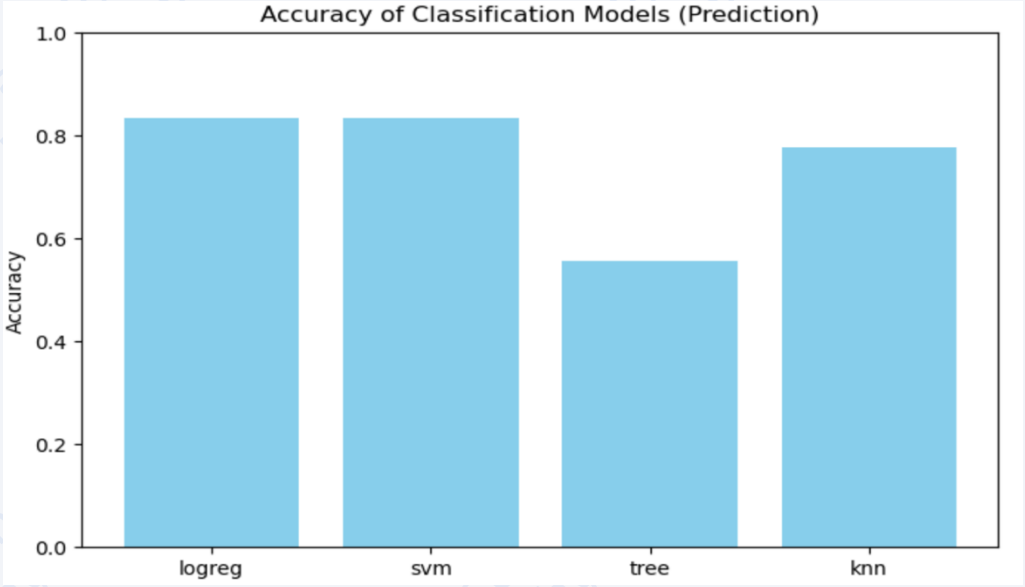
Finding the best performing classification model (1):

- The classification process began with loading the regular dataset and extracting the Class column as the target variable **Y**. The feature set **X** was obtained from the one-hot encoded dataset. Both were stored as DataFrames.
- As part of preprocessing, StandardScaler() was applied to **X** to standardize the feature values. The data was then split into **training and test sets** (80/20 split).
- A **Logistic Regression** model was trained using **GridSearchCV** with 10-fold cross-validation and different hyperparameter values for **C** (0.01, 0.1, 1). Ridge regularization (penalty='l2') was used. Grid search identified the best hyperparameter and the corresponding accuracy score.
- Model performance was then evaluated on the test set using the **accuracy score** and a **confusion matrix** heatmap to analyze the predictions.

Predictive Analysis (Classification) (3)

Finding the best performing classification model (2):

- This entire workflow was repeated for **three additional models** under the same conditions
  (80/20 split, 10-fold cross-validation):
  - **Support Vector Machine**,
  - **Decision Tree Classifier**,
  - **K-Nearest Neighbors**.



- The accuracy scores from both the validation and test predictions were stored in two dictionaries. Finally, the results were visualized in **two bar charts**, enabling a direct comparison of model performance and identification of the best-performing classifier.

Results

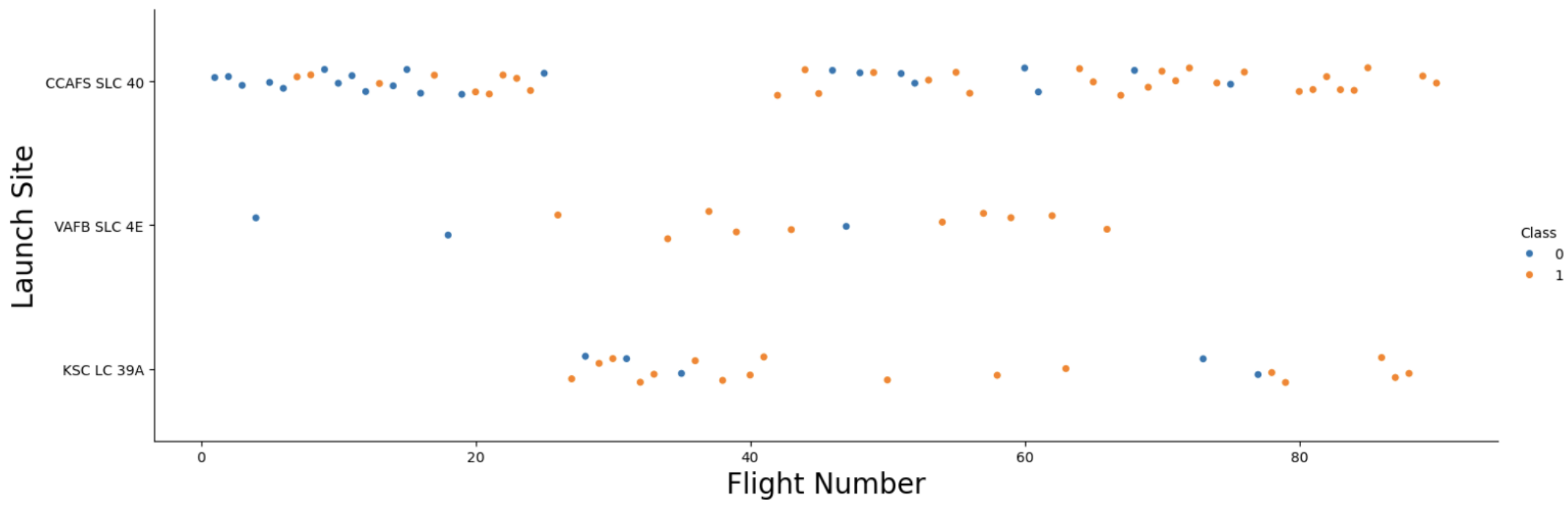
The detailed results are presented in the following chapters:

- Exploratory data analysis results:
  - “Insights drawn from EDA”
- Interactive analytics demo in screenshots:
  - “Launch Sites Proximities Analysis”
  - “Build a Dashboard with Plotly Dash”
- Predictive analysis results:
  - “Predictive Analysis (Classification)”
- Find the global summary at the end of this presentation.

Section 2

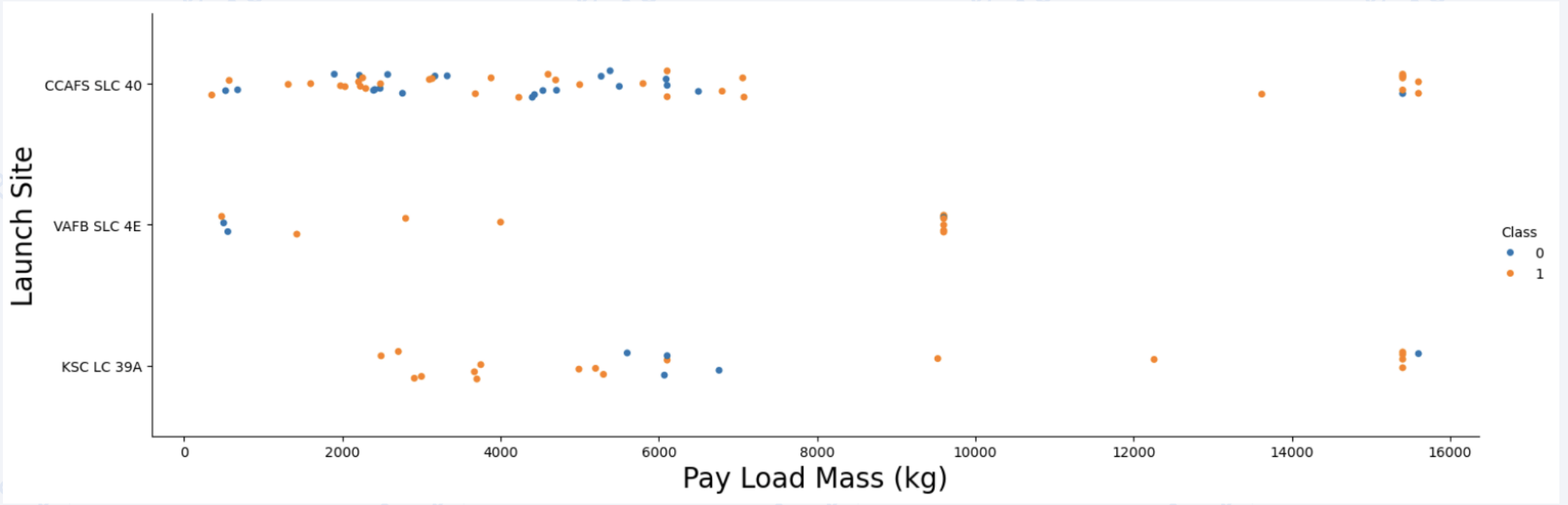
Insights drawn from EDA

Flight Number vs. Launch Site



Explanation: As the number of launches increases, failure rates decrease across all launch sites — suggesting that greater experience leads to higher success rates. While Cape Canaveral recorded the highest number of unsuccessful launches overall, these occurred primarily during early missions. In contrast, within the **last 20 launches**, Cape Canaveral had **only one failure**, making it the **most reliable launch site** based on recent performance.

Payload vs. Launch Site



Explanation: The majority of launches carry a payload of less than 8,000 kg. At Vandenberg, the maximum payload capacity appears to be below 10,000 kg, while heavier payloads are launched from Cape Canaveral or Kennedy Space Center. For these heavy payload missions, **Cape Canaveral shows a slightly higher success rate**. For lighter payloads under 8,000 kg, **Kennedy Space Center performs best**. However, since early testing was predominantly conducted at Cape Canaveral, this comparison may be somewhat misleading.

Success Rate vs. Orbit Type

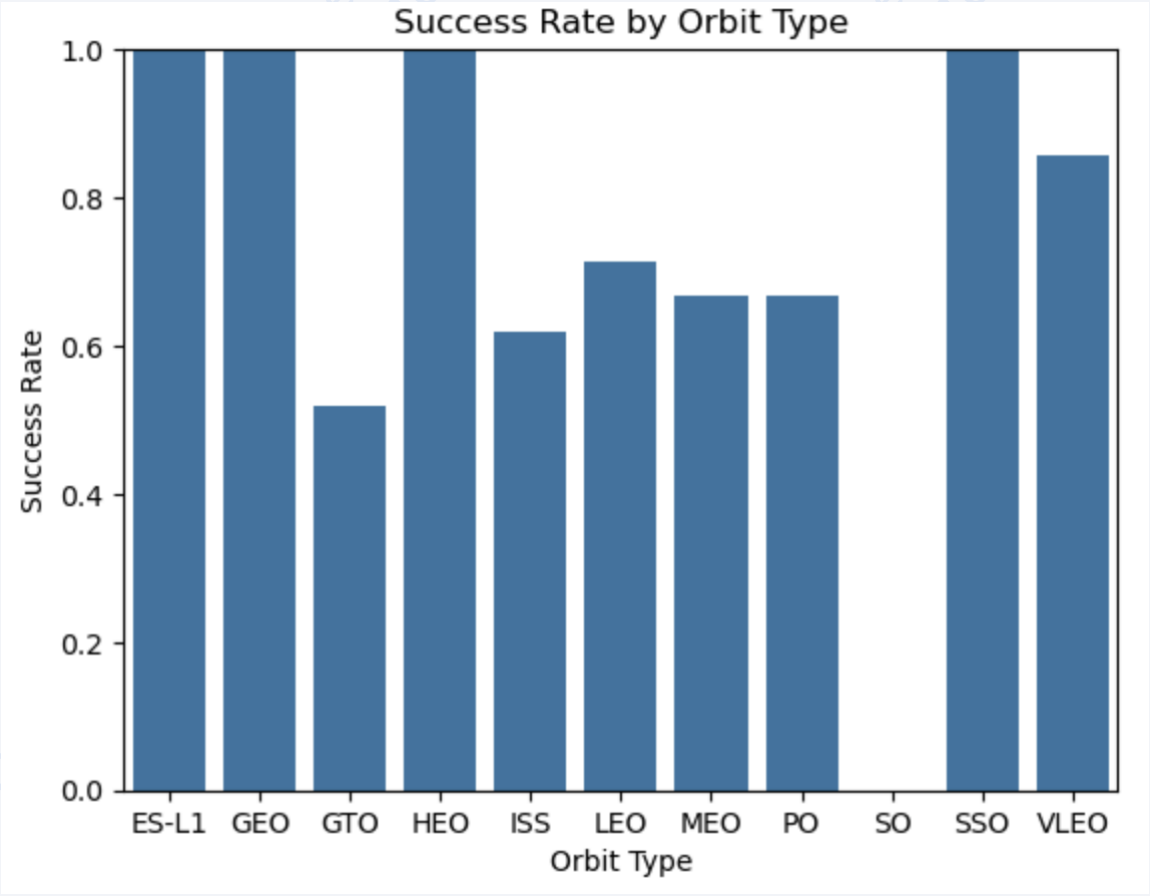
Explanation: The highest success rates (100%) are observed for the following orbits:

- Lagrange Point 1 between Earth and Sun (ES-L1)
- Geosynchronous Orbit (GEO)
- Highly Elliptical Orbit (HEO)

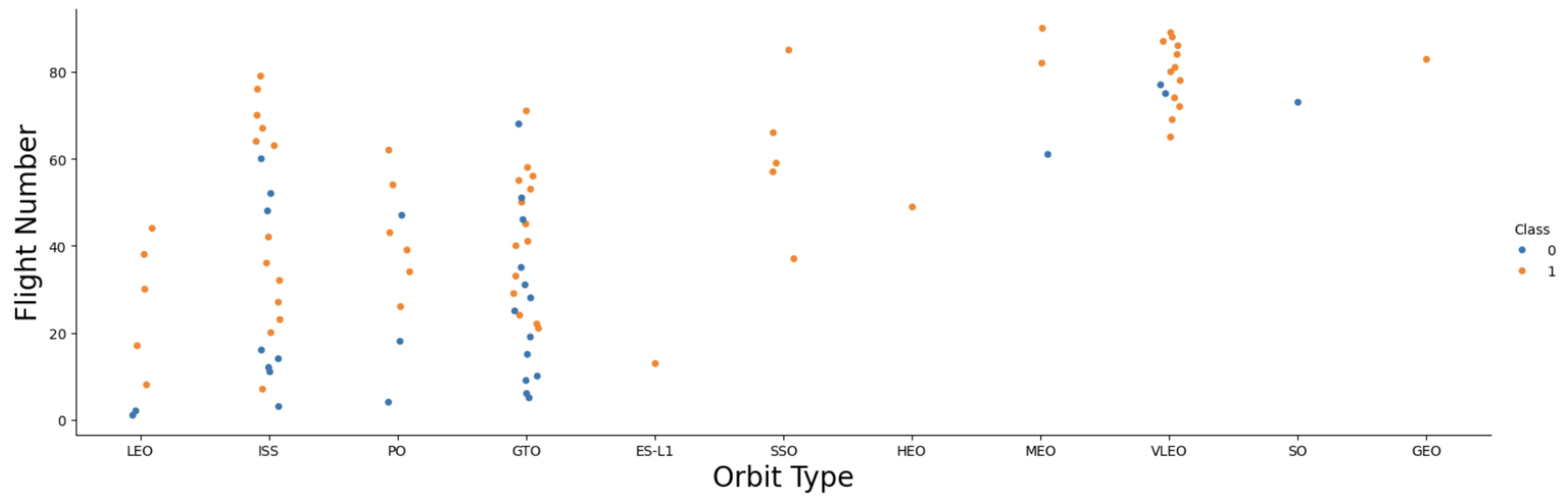
Sun-Synchronous Orbit (SSO) has 100% success rate but as SO is meaning the same kind of orbit, the true success rate is about 84%.

Very Low Earth Orbit (VLEO) follows with a success rate of over 80%.
The lowest success rate is seen for Geostationary Transfer Orbit (GTO), with only about 50% of launches succeeding.

Note: "SO" is equivalent to "SSO" in this dataset but indeed has a zero success rate.

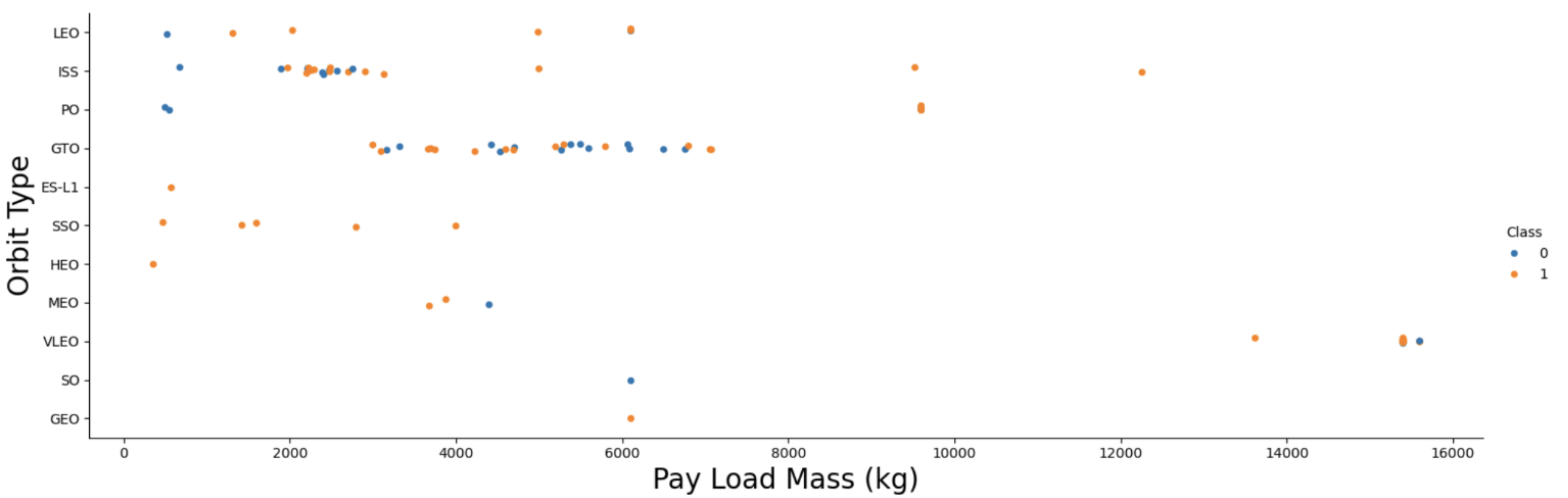


Flight Number vs. Orbit Type



Explanation: Early missions primarily targeted Low Earth Orbit (LEO). As the number of launches increased, there was a shift toward Very Low Earth Orbit (VLEO), which may indicate a higher level of difficulty. A significant number of missions—both early and recent—were also directed to the International Space Station (ISS). Naturally, a higher number of launches increases the likelihood of some failures.
Overall, missions to LEO and VLEO combined show a strong success rate of approximately 77%, despite the high volume of flights.

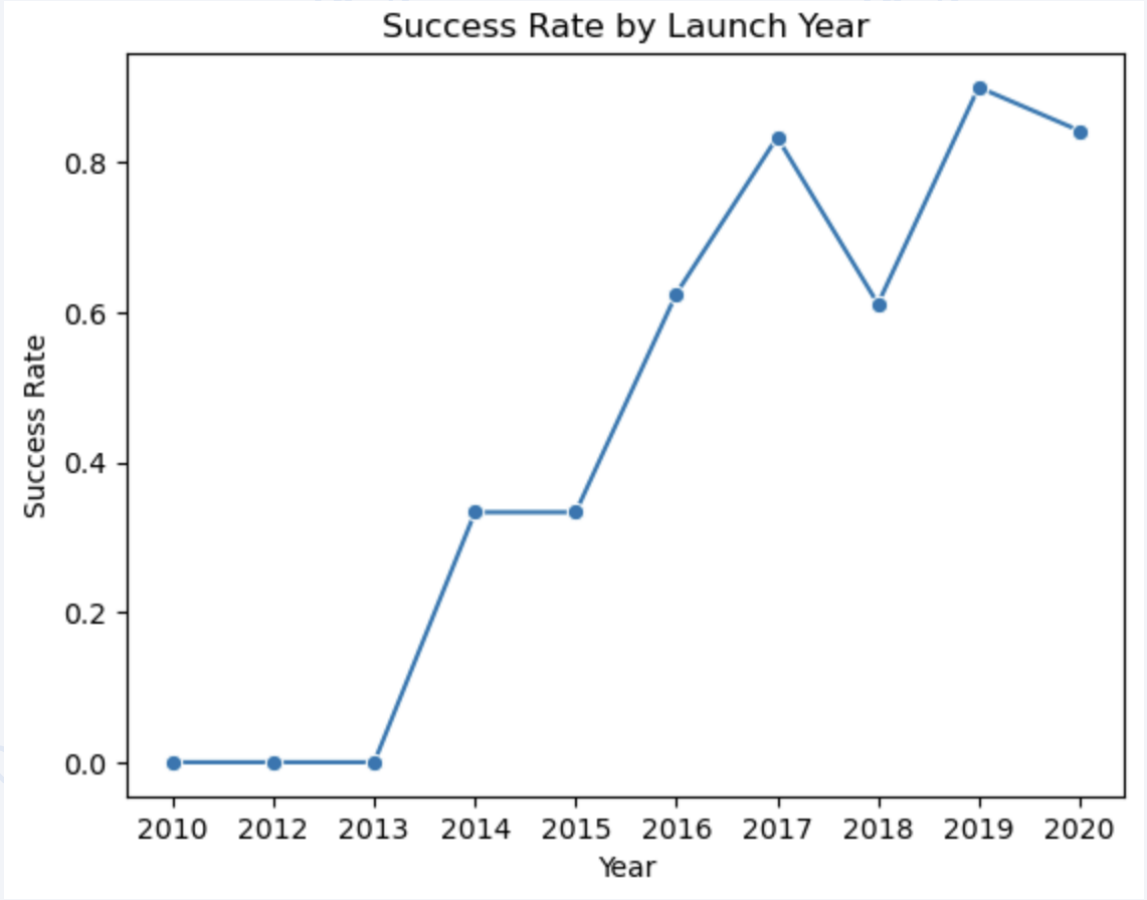
Payload vs. Orbit Type



Explanation: Most missions to all orbits carried payloads under 8000 kg. Very Low Earth Orbit (VLEO) stands out as the only destination with payloads exceeding 13,000 kg—one of which ended in failure. Additionally, two ISS missions carried payloads above 9000 kg, and two heavy-payload launches targeted Polar Orbit (PO).

Launch Success Yearly Trend

Explanation: In the early years, SpaceX had a 0% success rate until 2014. From that point onward, the success rate increased rapidly—reaching over 80% in 2017. A dip followed in 2018, bringing the rate down to around 60%, before climbing to over 90% in 2019. In 2020, another slight decline occurred, settling near 80%.
This fluctuation may be explained in two ways: either a lower overall launch count with the same number of failures, or technical and procedural changes due to experimentation that temporarily reduced the success rate. Overall the success rate is increasing by years.



All Launch Site Names

Query:

```
select distinct Launch_Site
from SPACEXTABLE
```

- Explanation: This query uses DISTINCT to retrieve all unique names of the launch sites that appear in the dataset. It helps identify the different locations from which launches were conducted.

Result:

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

Launch Site Names Begin with 'CCA'

Query:

```
select *
from SPACEXTABLE
where Launch_Site like 'CCA%'
limit 5;
```

- Explanation: This query filters the dataset using a WHERE condition to return only records related to launches from the Cape Canaveral launch site.

Result:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2

Total Payload Mass

Query:

```
select sum(cast(PAYLOAD_MASS__KG_ as INTEGER)) as Sum
from SPACEXTABLE
where Customer like '%CRS)';
```

- Explanation: The SUM function is used to calculate the total payload mass. A WHERE condition filters the dataset to include only records related to the customer *NASA CRS*. This allows the calculation of the total payload mass transported for this customer, which amounts to **45,496 kg**.

Result:

Sum
45596

Average Payload Mass by F9 v1.1

Query:

```
select avg(cast(PAYLOAD_MASS__KG_ as INTEGER)) as Average_Payload_mass
from SPACEXTABLE
where Booster_Version like '%F9 v1.1%'
```

- Explanation: This query uses the AVG function to calculate the average payload mass. The WHERE clause limits the calculation to all boosters of the *Falcon 9 v1.1* version, providing insight into the typical payload capacity for this specific rocket configuration.

Result:

Average_Payload_mass
2534.6666666666665

First Successful Ground Landing Date

Query:

```
select min(DATE(Date)) as First_Launch
from SPACEXTABLE
where Landing_Outcome is 'Success (ground pad)'
```

- Explanation: This query identifies the **first successful ground pad landing** by selecting the **minimum date** from the dataset. The WHERE clause filters the data to include only rows where the landing outcome is 'Success (ground pad)', ensuring that only relevant landings are considered.

Result:

First_Launch
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Query:

```
select distinct(Booster_Version) as Success_Boosters
from SPACEXTABLE
where Landing_Outcome is 'Success (drone ship)'
and cast(PAYLOAD_MASS__KG_ as INTEGER) between 4000 and 6000
```

- Explanation: This query retrieves all **distinct booster versions** that had **successful drone ship landings** with a **payload mass between 4000 and 6000 kg**. The WHERE clause filters the data by both the **landing outcome** ('Success (drone ship)') and the **payload mass range**, ensuring only relevant launches are included.

Result:

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

Total Number of Successful and Failure Mission Outcomes

Query:

```
select
case
when Landing_Outcome like 'Success%' then 'Success'
else 'Failure'
end as Outcome,
count(*) as Count
from SPACEXTABLE
group by Outcome;
```

Result:

Outcome	Count
Failure	40
Success	61

- Explanation: This query calculates the **total number of successful and failed landings**. Using a SELECT statement with a CASE expression, all landing outcomes containing the substring 'Success' are categorized as '**Success**', while all others are marked as '**Failure**'. The results are then **grouped and counted** to show the total number of occurrences in each category.

Boosters Carried Maximum Payload

Query:

```
select distinct(Booster_Version)
from SPACEXTABLE
where cast(PAYLOAD_MASS__KG_ as INTEGER) =
    (
        select max(cast(PAYLOAD_MASS__KG_ as
            INTEGER))
        from SPACEXTABLE
    );
```

Result:

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: To list all boosters that carried the **maximum payload**, the **inner query** first filters the dataset using WHERE to find the **maximum payload mass**. The **outer query** then retrieves all **distinct booster versions** (DISTINCT) that match this maximum payload using a second WHERE condition on payload_mass.

2015 Launch Records

- Explanation: This query lists all **failed drone ship landings** from the year **2015**, including **month**, **year**, **booster version**, **launch site**, and **landing outcome**. The **month** and **year** values are extracted from the date column using SUBSTRING. The WHERE clause filters for records containing the year *2015* and a landing outcome of 'Failure (drone ship)'.

Query:

```
select substr(Date,6,2) as Month,
substr(Date,0,5) as Year,
Booster_Version,
Launch_Site,
Landing_Outcome
from SPACEXTABLE
where substr(Date,0,5) = '2015'
and Landing_Outcome = 'Failure (drone ship)'
```

Result:

Month	Year	Booster_Version	Launch_Site	Landing_Outcome
01	2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

Query:

```
select
Landing_Outcome as Outcome,
COUNT(*) as Count
from SPACEXTABLE
where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by Count desc;
```

- Explanation: This query generates a **ranking of landing outcomes** in descending order, limited to the date range between **June 4, 2010, and March 20, 2017**.
  Using WHERE BETWEEN, the dataset is filtered by date. The landing_outcome values are **grouped and counted**, and the results are **sorted by frequency** using ORDER BY COUNT DESC.
  The result reveals that **in most cases, no landing attempt was made** during this period.

Result:

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

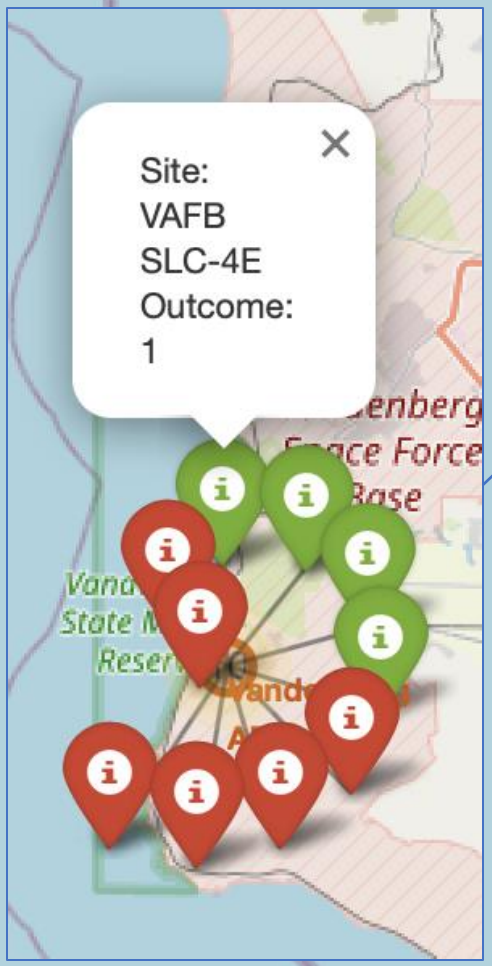
Section 3

Launch Sites Proximities Analysis

Location of the 3 SpaceX Launch Sites

- The **Folium map** displays the **three launch sites** used by SpaceX for its launches.
- These include the **Vandenberg Space Force Base** on the **Pacific Coast**, and **Cape Canaveral** along with the **Kennedy Space Center**, both located on the **Atlantic Coast**.
- The global view highlights that **all launch sites are situated within the United States**. Each site has **access to the ocean**, which is beneficial for testing **landings on drone ships**.

Launch Outcomes on Vandenberg AFB



- The **Folium map** displays **all launches from Vandenberg Space Force Base** (former Air Force Base). The **color of the markers** indicates the **landing outcome**:
  - **Green** represents **successful landings**,
  - **Red** indicates **failed landings**. Clicking on a marker reveals the **launch site** and the **outcome** as text in a popup.
- Out of all launches from this site, **4 landings were successful**, while **6 landings failed**.

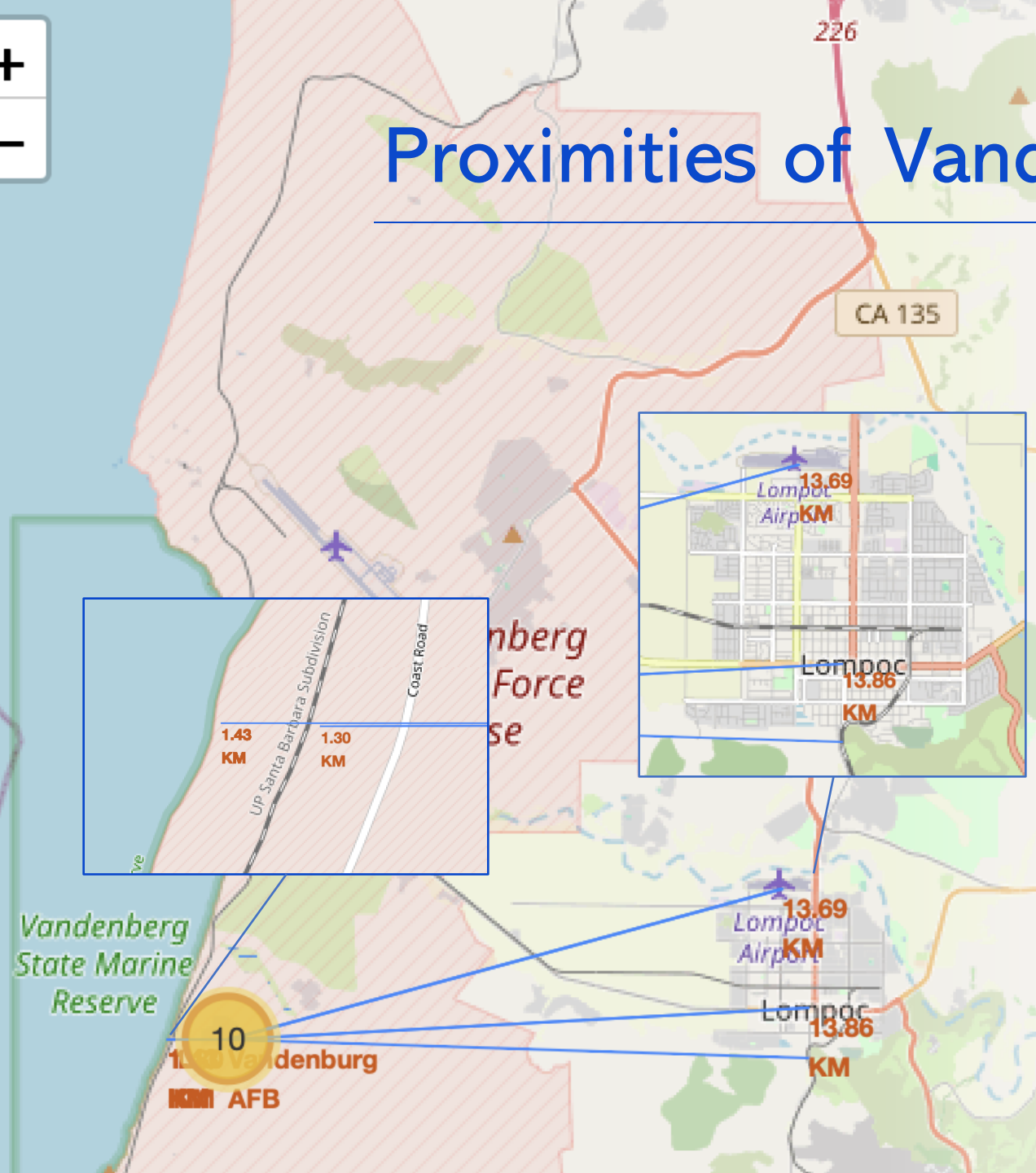
Launch Outcomes at Kennedy Space Center

- The **Folium map** displays **all launches from Kennedy Space Center**.
  The **color of the markers** indicates the **landing outcome**:
  - **Green** represents **successful landings**,
  - **Red** indicates **failed landings**.
    Clicking on a marker reveals the **launch site** and the **outcome** as text in a popup.
- Out of all launches from this site, **10 landings were successful**, while **3 landings failed**.

Launch Outcomes at Cape Canaveral

- The **Folium map** displays **all launches from Cape Canaveral**.
- For **Cape Canaveral**, the dataset lists **two different launch site names**: **SC-40** and **SLC-40**.
  These entries have **slightly different coordinates**, which causes them to appear as **two separate launch sites** on the map—In reality, this refers to the same physical site—**Space Launch Complex 40**. The difference is likely due to inconsistent naming in the data.
- From all sites most launches were perfomed from Cape Canaveral.
- Out of all launches from this site, **10 landings were successful**, while **23 landings failed**.

Proximities of Vandenburg Space Force Base



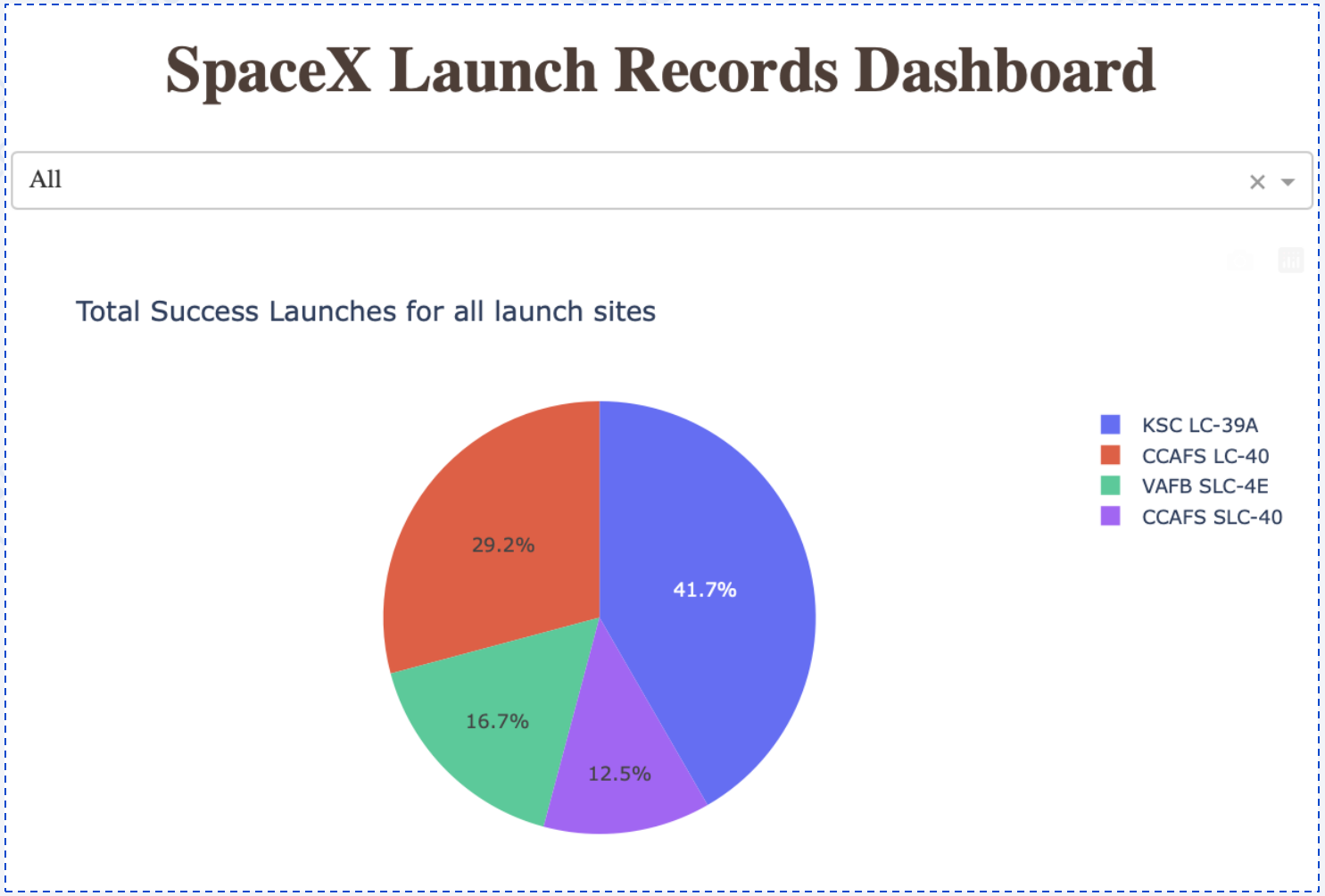
- **Notable proximities of the Vandenberg Space Force Base include:**
  - The coastline lies approximately **1.43 km** from the launch site, while the nearest railway line is just **1.30 km** away.
  - Another railway line can be found in the nearby city of **Lompoc**, located **13.86 km** from the base.
  - Lompoc also hosts a **small airport** (**13.69 km** away) and provides access to the **nearest major road connection**, making it an important transportation hub in the region.
- **Compared to the launch sites in Florida**, these surroundings reflect the **typical proximities SpaceX launch sites are equipped with** — coastal access for drone ship landings, nearby rail and road infrastructure, and proximity to local airports for logistical support.

Section 4

Build a Dashboard with Plotly Dash

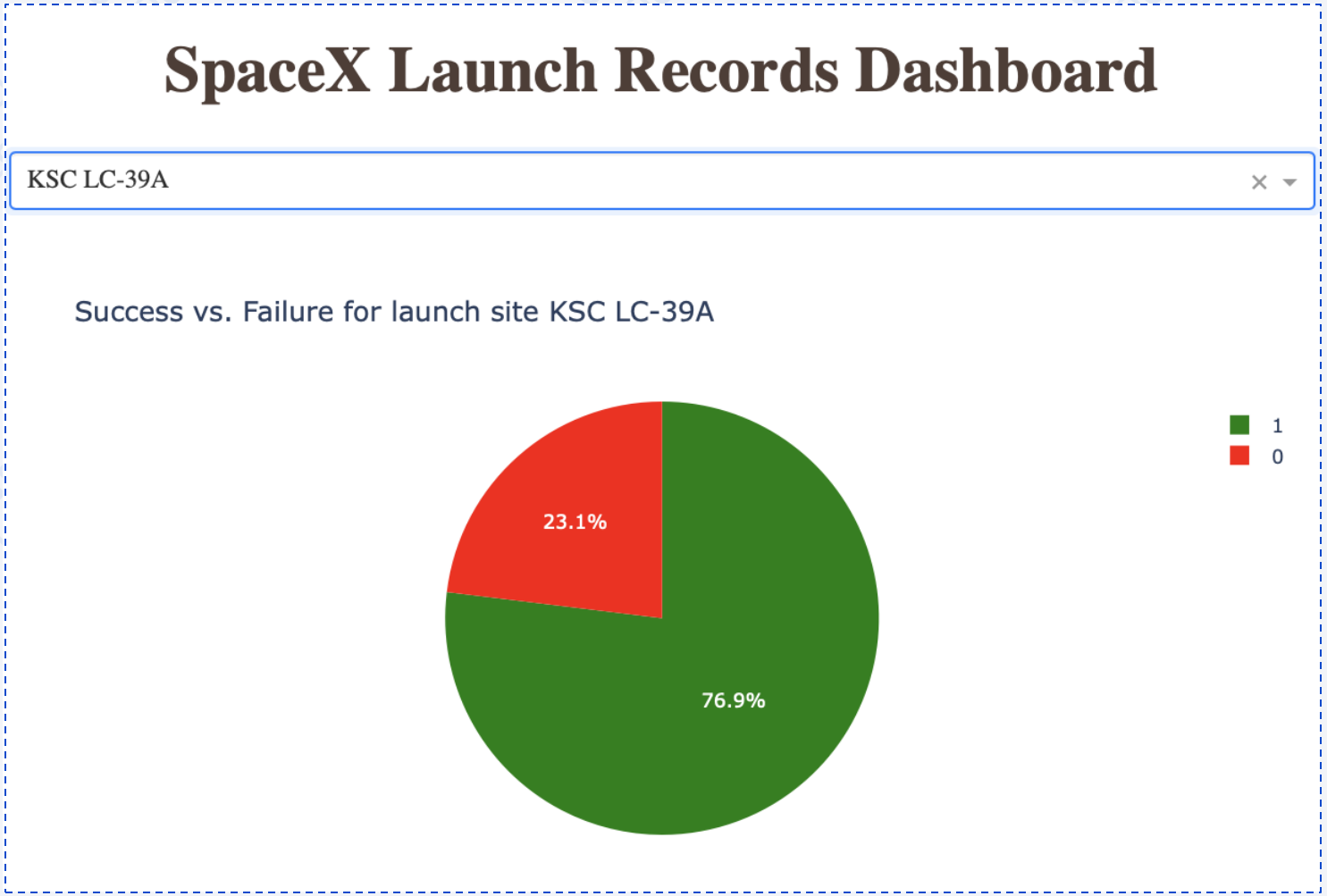
Total Successful Landings for all Launch Sites

- **The initial view of the dashboard displays** either the **total number of successful landings per launch** (ALL is selected), or, if a specific **launch site is selected**, the number of **successful and failed landings** for that site.
- An overview of **all successful landings across all launch sites** reveals that the **Kennedy Space Center ranks second** in terms of successful landings (41,7%).
- If both **Cape Canaveral launch sites** (SC-40 and SLC-40) are combined, **Cape Canaveral accounts for the majority** of successful landings (42,7%).
- **Vandenberg Space Force Base** ranks **last** among the three, with the **fewest successful landings** (16,7%).



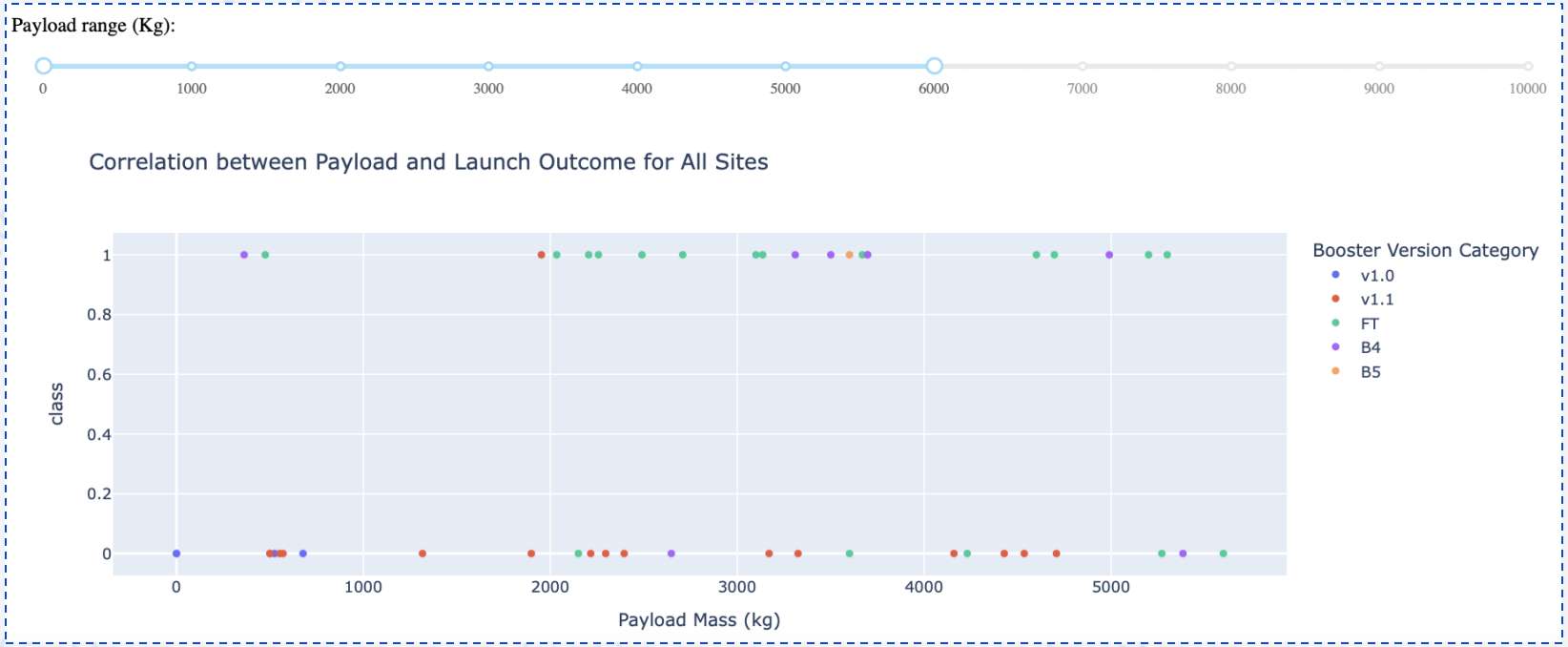
Launch Site with Highest Success Rate

- **When examining individual launch sites**, it becomes clear that the **Kennedy Space Center has the highest success rate**, as **significantly more launches resulted in successful landings than in failures**.



The Impact of Payload on Landing Outcomes (1)

- On the interactive dashboard, a slider allows us to select the payload weight. This filters the scatter plot below accordingly. On this slide, all payloads up to 6000 kg are selected.



- Across all launch sites, it is noticeable that many launches with payloads up to 6000 kg were conducted, but a significant number of these did not result in successful landings. Booster version 1.1 was particularly less successful in this range. In contrast, the FT and B4 booster versions show a significantly higher landing success rate, although they were used in fewer launches.

The Impact of Payload on Landing Outcomes (2)

- On this slide a payload heavier than 6000 kg is selected.

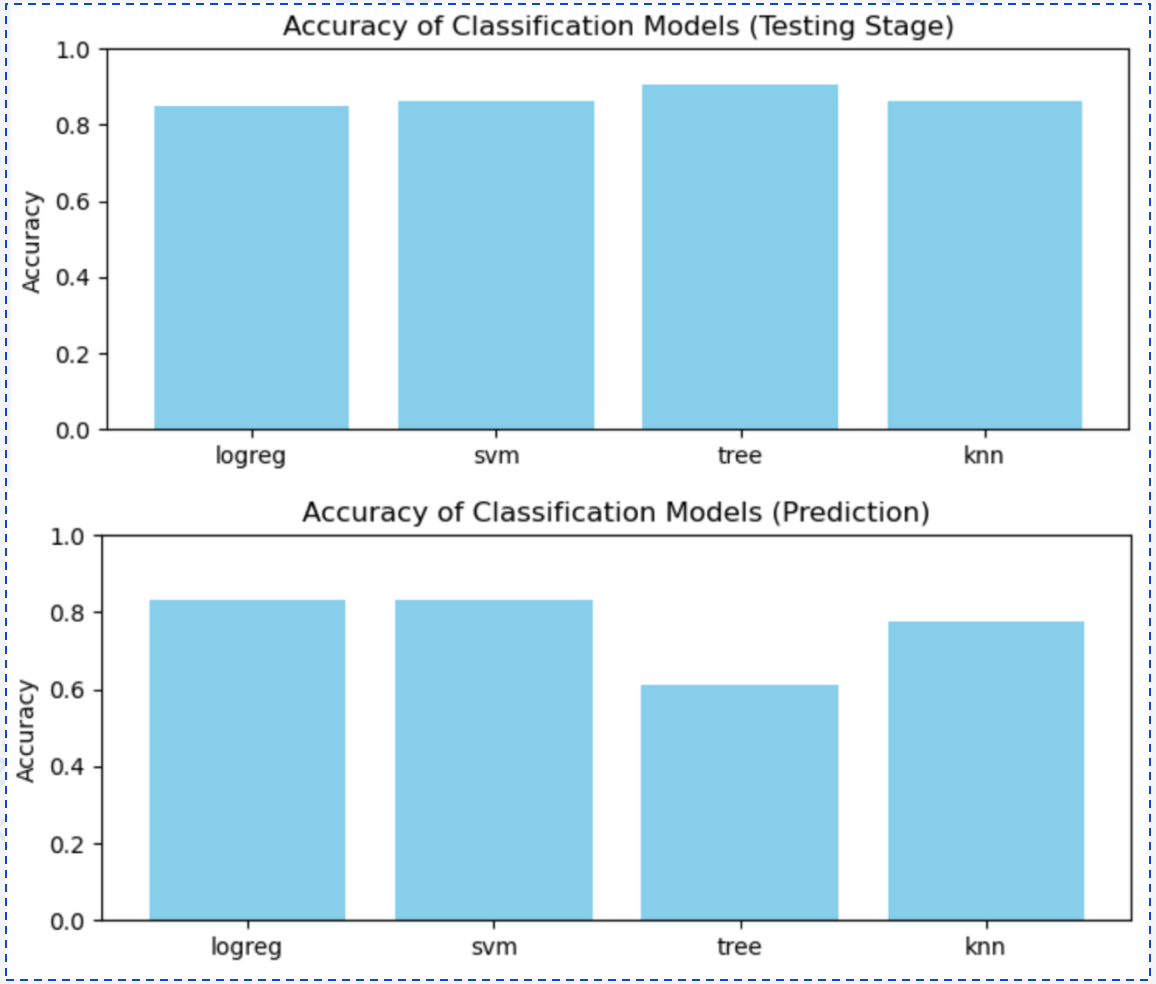
- Across all launch sites, it is now noticeable that only a few launches with payloads more than 6000 kg were conducted, and except one they did not result in successful landings. The only booster version that successfully landed was, again, booster version B4 that did also well with lighter payloads. Thus it can be stated that booster B4 is quite successful amongst all boosters.

Section 5

Predictive Analysis (Classification)

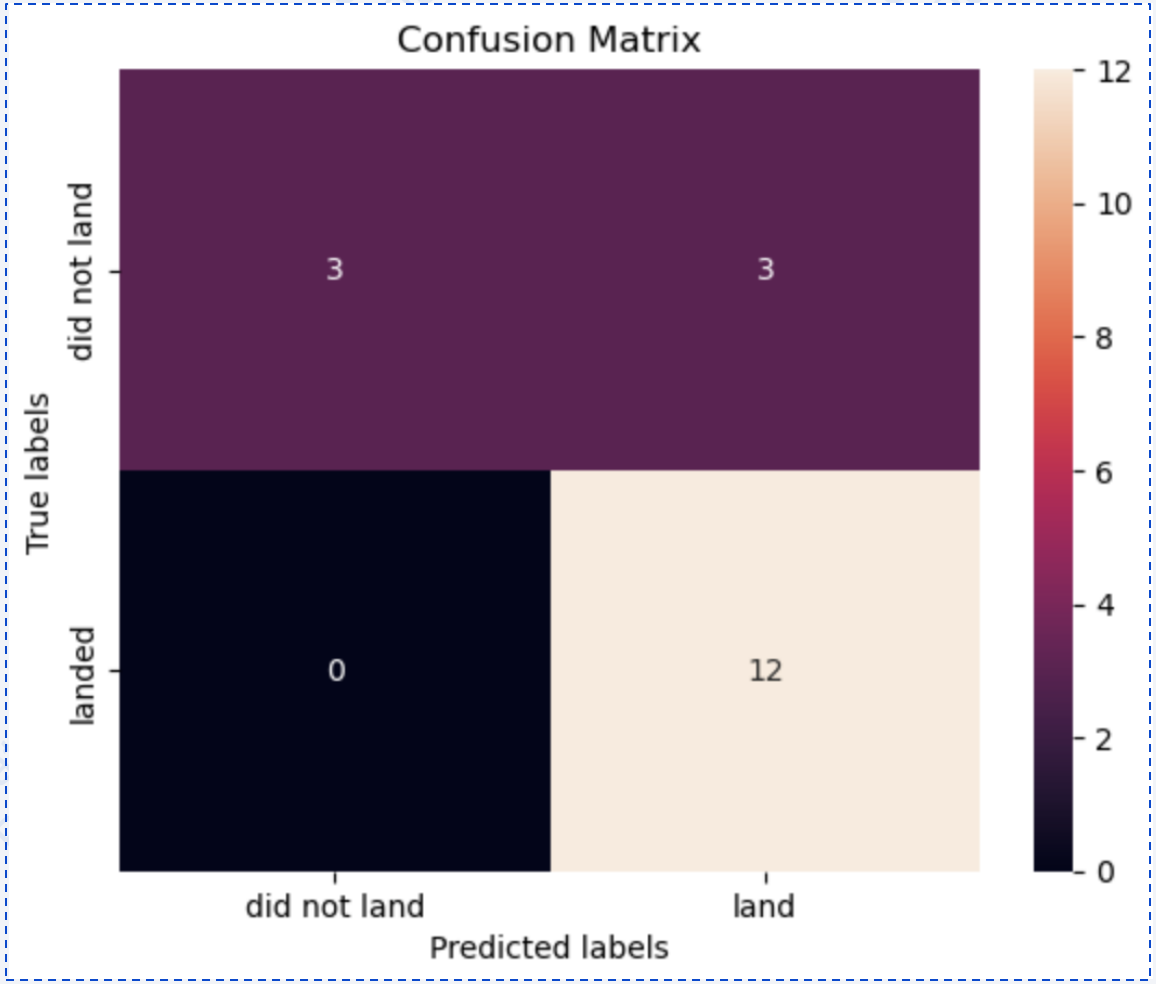
Classification Accuracy

- To determine the best classification algorithm, the following models were tested: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).
- For all models, accuracy was measured using the best_score_ after training and the accuracy_score for prediction performance.
- The results were visualized in a bar chart to enable a direct comparison.
- The best performing models on unseen data are SVM with 86% precision after training and 83% on unseen data and Logistic Regression with 85% precision after training and 83% prediction accuracy on unseen data. Noticeable is also the decision tree model, as it has the best performance after training but performs poorly on unseen data, which is a sign of overfitting. Nonetheless, this model might perform quite well with a different kind of data preparation.
- Also, it must be noted, that the test data consisted of only 18 samples and the whole set of 89, what is too few samples for detailed analysis. Reworking the predictive analysis should include data preparation that cares for missing values using a different method. Also, one of the one-hot encoded parameters should be dropped to prevent overfitting.



Confusion Matrix

- This confusion matrix shows the prediction capabilities of the linear regression model (that performed like SVM on unseen data).
- All unsuccessful landings were predicted correctly. Regarding to the successful landings 12 out of 15 were predicted correctly, a 4:1 ratio.
- The only problem are the false positives on successful landings, what is, compared to the other models' predictions, quite good either.
- See appendix for the other three heatmaps

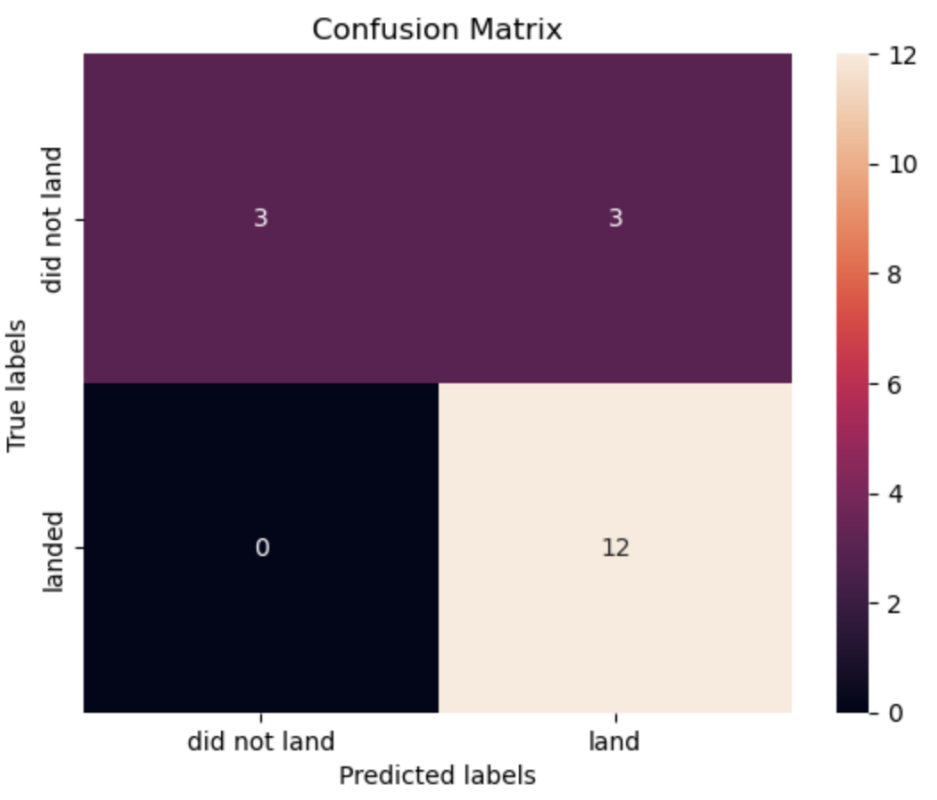


General conclusions drawn from all discussed aspects

- SpaceX experienced high failure rates during its early launch phase, with the first successful ground pad landing achieved only after two years. Our analysis highlights patterns that may help improve success rates from the outset.
- It is advisable to begin with lighter payloads, as they are associated with higher landing success and greater flexibility in selecting launch sites. As operational experience grows, transitioning to heavier payloads becomes more feasible.
- Targeting lower Earth orbits in early missions further increases the chance of success.
- Launch sites with direct access to the ocean are beneficial for preparing drone ship landings. Additionally, proximity to robust transport infrastructure—such as rail lines, highways, and airfields—is critical for logistical efficiency.
- Among all launch facilities, the Kennedy Space Center stands out with the highest landing success rate. In absolute numbers, Cape Canaveral leads with the most successful landings—especially in early-stage missions—underscoring its value as a proven starting point.
- We recommend selecting a booster similar to SpaceX's B4 model, which demonstrated versatility and reliability across both light and heavier payload classes.

Appendix

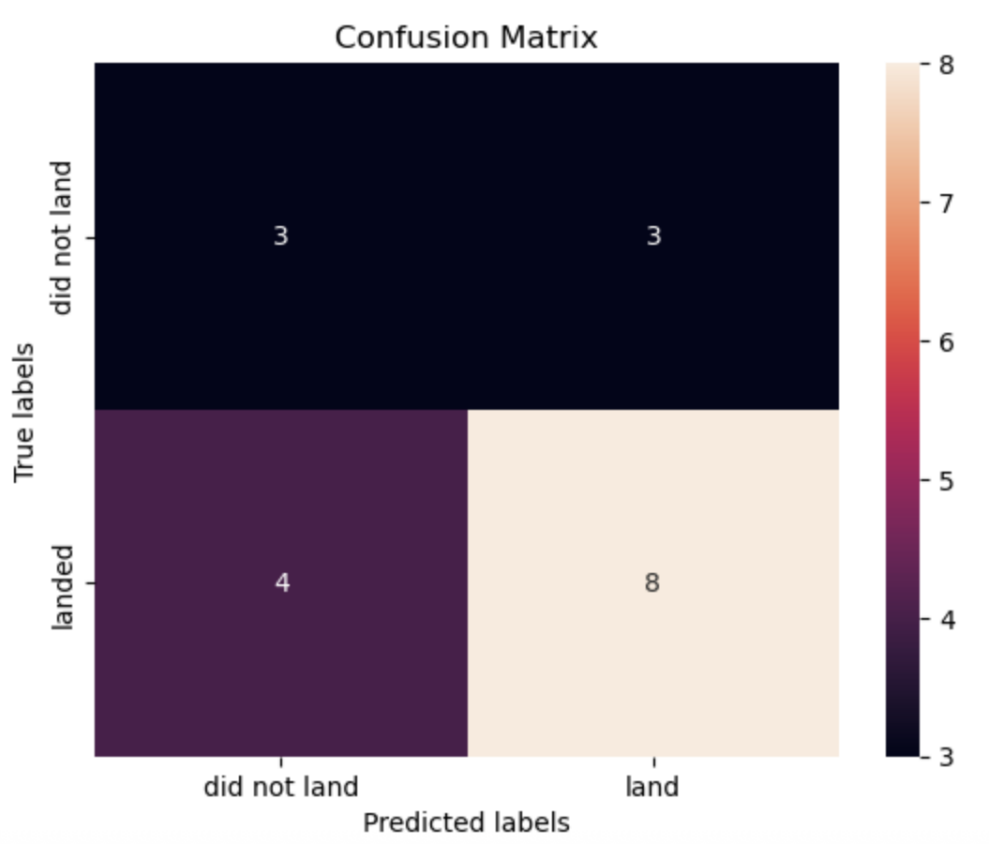
- Confusion Matrix for Support Vector Machine (SVM):




```
yhat_svm=svm_cv.predict(X_test)
modelacc_pred['svm'] =
accuracy_score(Y_test, yhat_svm)
plot_confusion_matrix(Y_test,yhat_svm)
```

Appendix

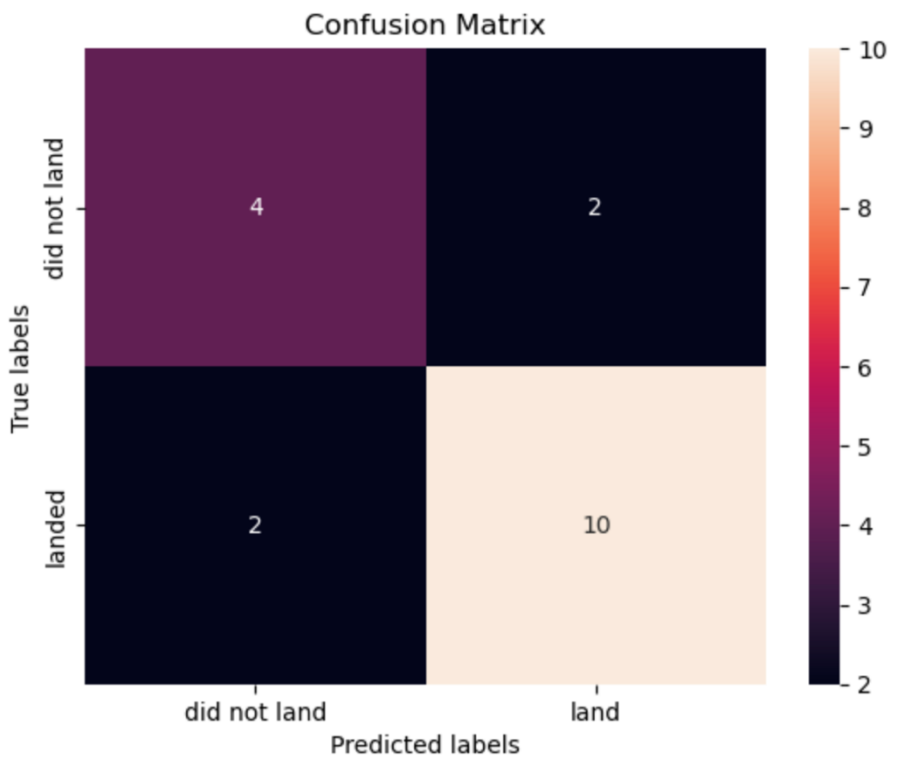
- Confusion Matrix for Decision Tree:




```
yhat_tree = tree_cv.predict(X_test)
modelacc_pred['tree'] =
accuracy_score(Y_test, yhat_tree)
plot_confusion_matrix(Y_test,yhat_tree)
```

Appendix

- Confusion Matrix for K Nearest Neighbors:




```
yhat_knn = knn_cv.predict(X_test)
modelacc_pred['knn'] =
accuracy_score(Y_test,yhat_knn)
plot_confusion_matrix(Y_test,yhat_knn)
```

Appendix

- Find all Jupyter Notebooks here:
  - https://github.com/EB1065/datascience

Thank you!