Data Science Capstone Project

Arman Moradi Fard

12/2024

Overview Summary

- Gathered data from the SpaceX API and SpaceX's Wikipedia page. Created a 'class' column to label successful landings. Analyzed the dataset using SQL, data visualizations, folium maps, and interactive dashboards. Selected relevant columns as features and converted categorical variables into binary format using one-hot encoding. Standardized the data and utilized GridSearchCV to optimize hyperparameters for machine learning models. Visualized the accuracy scores of all models.

- Developed four machine learning models: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K-Nearest Neighbors. Each model achieved a similar accuracy of approximately 83.33% but tended to over-predict successful landings. Additional data is required to improve model performance and accuracy.

Introduction

Background

- SpaceX offers the most competitive pricing in the industry ($62 million vs. $165 million USD), primarily because of its capability to recover a portion of its rockets (Stage 1).
- SpaceY aims to rival SpaceX's success.

Problem

- SpaceY has assigned us the task of developing a machine learning model to predict the successful recovery of Stage 1 rockets.

Methodology

Data Collection Methodology:

- Integrated data from the SpaceX public API and Wikipedia page.
- Conducted data wrangling to clean and preprocess the dataset.
- Classified landings as either successful or unsuccessful.
- Performed exploratory data analysis (EDA) using SQL and visualizations.
- Created interactive visualizations with Folium and Plotly Dash for deeper insights.
- Conducted predictive analysis with classification models.
- Optimized model performance using GridSearchCV.

Data Collection Overview

The data collection involved retrieving information through API requests from SpaceX's public API and web scraping a table from SpaceX's Wikipedia page.

The following slides will present a flowchart for the API data collection process, followed by a flowchart illustrating the web scraping methodology.

SpaceX API Data Columns:

- FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Wikipedia Web-Scraped Data Columns:

- Flight No., Launch Site, Payload, PayloadMass, Orbit, Customer, Launch Outcome, Booster Version, Booster Landing, Date, Time

Data Wrangling

Creating Training Labels for Landing Outcomes

- Define a training label where successful landings are assigned a value of 1, and failures are assigned a value of 0.
- The Outcome column consists of two components: Mission Outcome and Landing Location.
- Introduce a new column, class, which is set to 1 if the Mission Outcome is successful and 0 otherwise.

Value Mapping:

- **Successful (1):** True ASDS, True RTLS, True Ocean
- **Failure (0):** None None, False ASDS, None ASDS, False Ocean, False RTLS

Exploratory Data Analysis Through Visualization

EDA was conducted on key variables including Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year.

Plots Used:

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Orbit vs. Success Rate
- Flight Number vs. Orbit
- Payload vs. Orbit
- Yearly Success Trend

Exploratory Data Analysis Using SQL

Data Loading and Querying Process

- Uploaded the dataset into the IBM DB2 database.
- Used SQL Python integration to perform queries.
- Conducted queries to gain deeper insights into the dataset.
- Extracted information on launch site names, mission outcomes, payload sizes for various customers, booster versions, and landing outcomes.

Creating an Interactive Map Using Folium

Folium Map Analysis

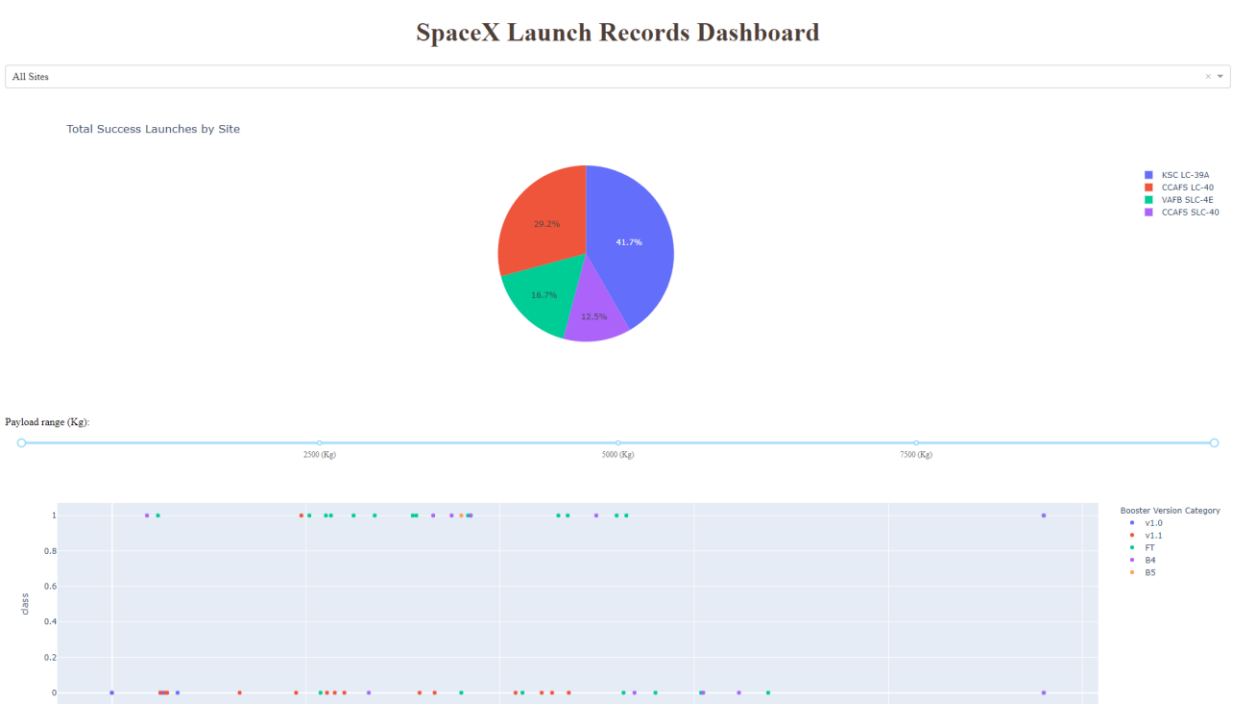
- Created Folium maps to highlight launch sites, as well as successful and unsuccessful landings.
- Included proximity analysis to key features such as railways, highways, coastlines, and cities.
- This visualization helps to understand the strategic placement of launch sites and illustrates the relationship between landing success and location.

Developing a Dashboard with Plotly Dash

Dashboard Features

- Includes a pie chart and a scatter plot for interactive data exploration.
- The **pie chart** displays the distribution of successful landings across all launch sites or can be filtered to show success rates for individual launch sites.
- The **scatter plot** accepts two inputs: selection of all launch sites or a specific site, and a payload mass slider ranging from 0 to 10,000 kg.
- The pie chart provides insights into launch site success rates.
- The scatter plot helps analyze how success rates vary with launch sites, payload mass, and booster version categories.

Results

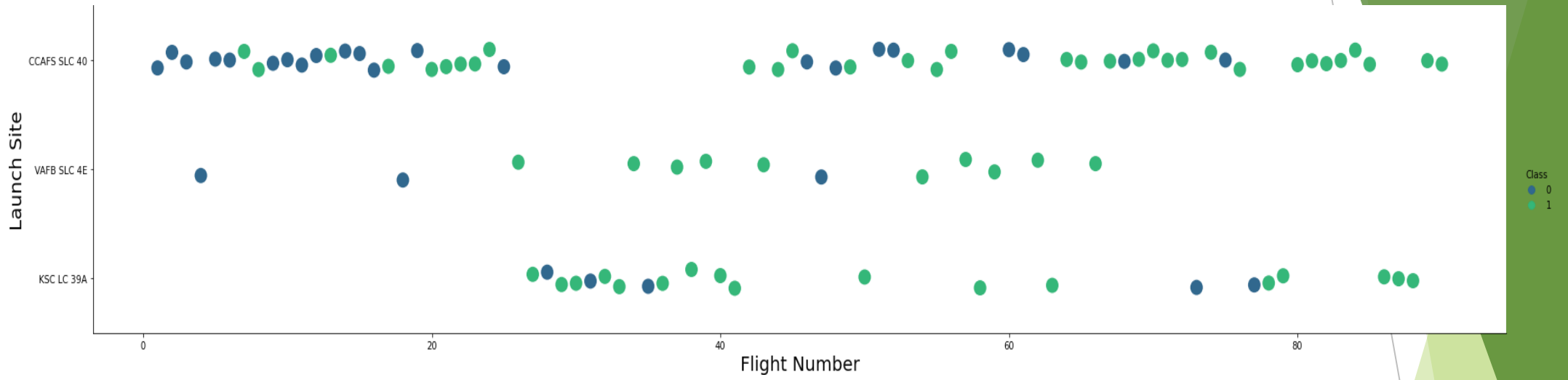


Plotly Dashboard Preview

This slide provides a preview of the Plotly dashboard. The subsequent slides will present:

- Results of Exploratory Data Analysis (EDA) with visualizations
- EDA insights using SQL
- Interactive maps created with Folium
- Final model results, achieving approximately 83% accuracy

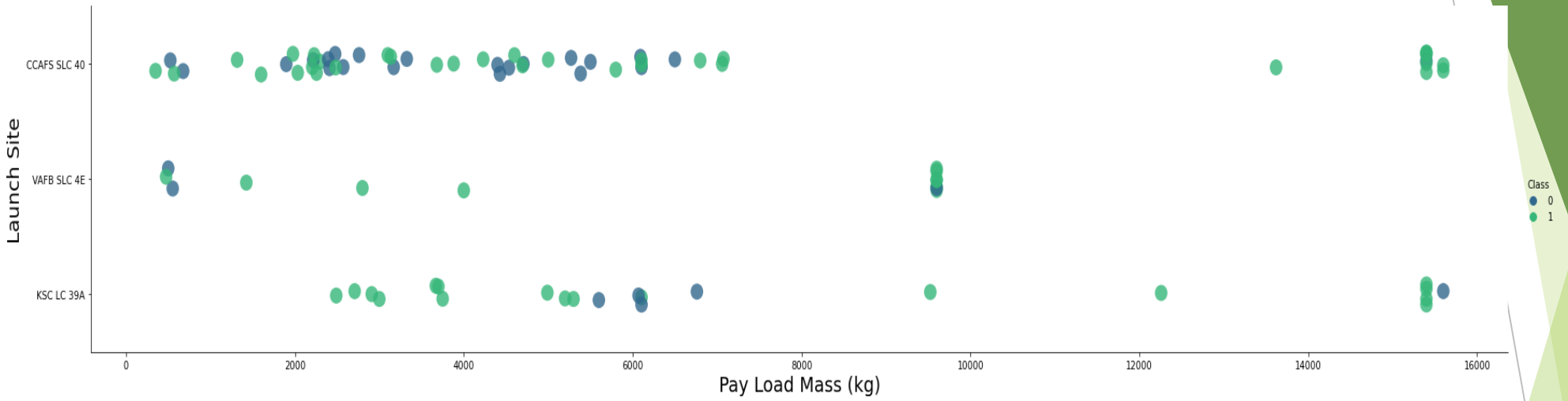
Flight Number vs. Launch Site



Color Key: Green represents successful and Purple represents unsuccessful launches.

- The graphic indicates an increasing success rate over time, as reflected by the Flight Number.
- A notable breakthrough seems to have occurred around Flight 20, leading to a significant improvement in success rates.
- CCAFS emerges as the primary launch site due to its high launch volume.

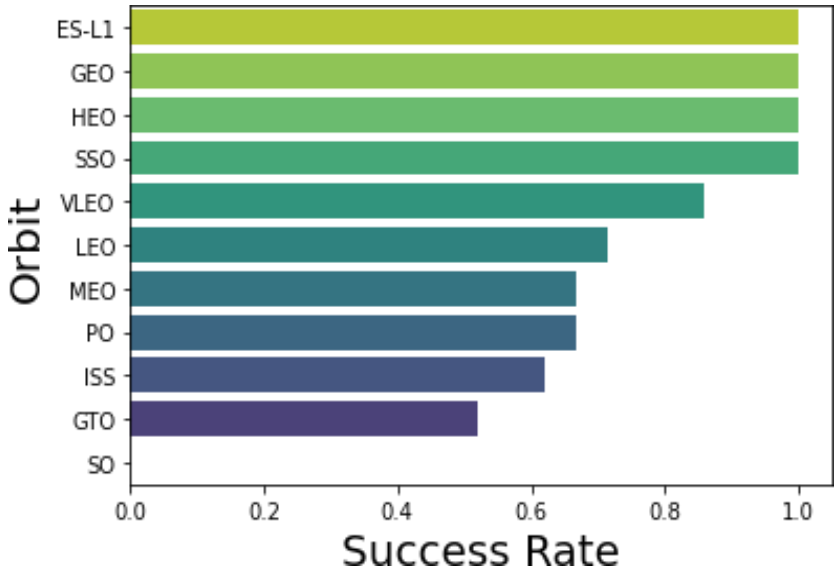
Payload Mass vs. Launch Site



Color Key: Green represents successful and Purple represents unsuccessful launches.

- The majority of payload masses range between 0 and 6,000 kg.
- Different launch sites appear to handle varying payload mass ranges.

Success rate vs. Orbit type

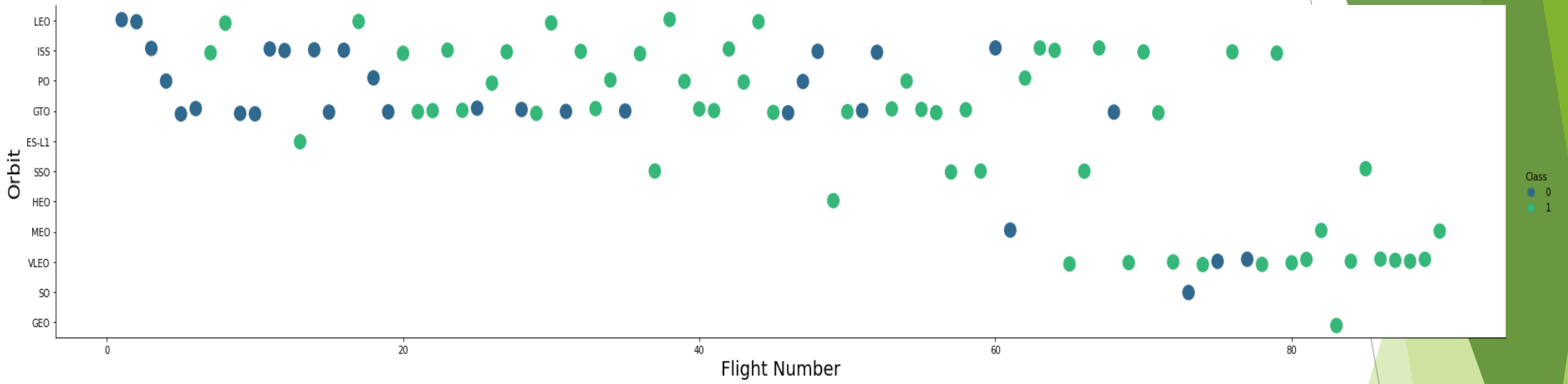


- ES-L1 (1), GEO (1), and HEO (1) each have a 100% success rate (sample sizes in parentheses).
- SSO (5) also achieves a 100% success rate.
- VLEO (14) demonstrates a decent success rate with a notable number of attempts.
- SO (1) has a 0% success rate.
- GTO (27) has the largest sample size but only about a 50% success rate.

Success Rate Scale

- 0 represents 0% success.
- 0.6 represents 60% success.
- 1 represents 100% success.

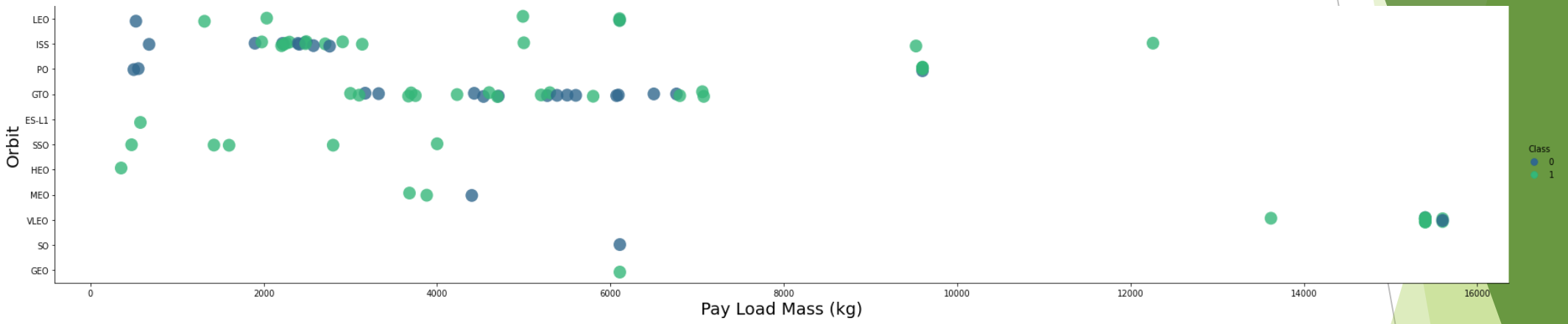
Flight Number vs. Orbit type



Color Key: Green represents successful and Purple represents unsuccessful launches.

- Launch orbit preferences evolved over time, as indicated by Flight Number.
- Launch outcomes appear to correlate with these changing preferences.
- Initially, SpaceX focused on LEO orbits, which achieved moderate success, before shifting back to VLEO in recent launches.
- SpaceX tends to perform better in lower orbits or Sun-synchronous orbits.

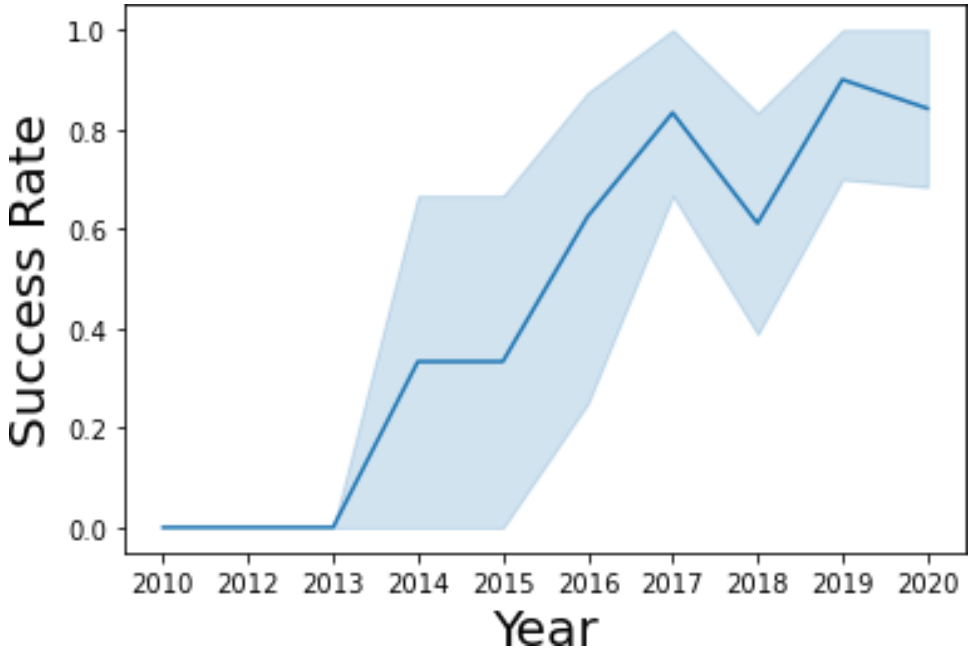
Payload vs. Orbit type



Color Key: Green represents successful and Purple represents unsuccessful launches.

- Payload mass appears to be linked to orbit type.
- LEO and SSO orbits typically have relatively low payload masses.
- The other highly successful orbit, VLEO, tends to feature payload mass values at the higher end of the range.

Yearly Trend of Launch Success



95% confidence interval
(light blue shading)

- Overall success rates have generally increased since 2013, with a minor decline in 2018.
- In recent years, success rates have stabilized at approximately 80%.

All Launch Site Names

```
In [4]: %%sql
        SELECT UNIQUE LAUNCH_SITE
        FROM SPACEXDATASET;

         * ibm_db_sa://ftb12020:***@0c77d6f2
        Done.
```

Out[4]:

launch_site
CCAFS LC-40
CCAFS SLC-40
CCAFSSLC-40
KSC LC-39A
VAFB SLC-4E

- Queried unique launch site names from the database.
- CCAFS SLC-40 and CCAFSSLC-40 likely represent the same launch site due to data entry inconsistencies.
- CCAFS LC-40 appears to be a previous name for the site.
- There are likely only three unique launch site values: CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E.

Launch Site Names Starting with CCA

In [5]:
```
%%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Retrieved the first five entries where the launch site name begins with CCA.

Total Payload Mass from NASA

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

sum_payload_mass_kg
45596

- This query calculates the total payload mass (in kg) for launches where NASA was the customer.
- CRS (Commercial Resupply Services) indicates that these payloads were destined for the International Space Station (ISS).

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-8(
Done.
```

avg_payload_mass_kg
2928

- This query computes the average payload mass for launches utilizing the booster version F9 v1.1.
- The average payload mass for F9 v1.1 falls on the lower end of the overall payload mass range.

Date of First Successful Ground Pad Landing

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

first_success
2015-12-22

- This query retrieves the date of the first successful ground pad landing.
- The first ground pad landing occurred in late 2015.
- Successful landings, in general, began appearing from 2014 onwards.

Successful Drone Ship Landings with Payloads Between 4,000 and 6,000

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

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

- This query identifies the four booster versions that achieved successful drone ship landings with a payload mass between 4,000 and 6,000 kg (exclusive).

Count of Each Mission Outcome

```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-
Done.
```

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

- This query provides a count of each mission outcome.
- SpaceX achieves its mission outcomes approximately 99% of the time, indicating that most landing failures are intentional.
- Notably, one launch has an unclear payload status, and one mission experienced an in-flight failure.

Boosters Carrying the Maximum Payload

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
Done.
```

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- This query retrieves the booster versions that carried the maximum payload mass of 15,600 kg.
- These booster versions are closely related, all belonging to the F9 B5 B10xx.x series.
- This suggests a strong correlation between payload mass and the specific booster version utilized.

Records of Failed Drone Ship Landings in 2015

```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

This query retrieves the month, landing outcome, booster version, payload mass (kg), and launch site for 2015 launches where Stage 1 failed to land on a drone ship.

Ranked Counts of Successful Landings (2010-06-04 to 2017-03-20)

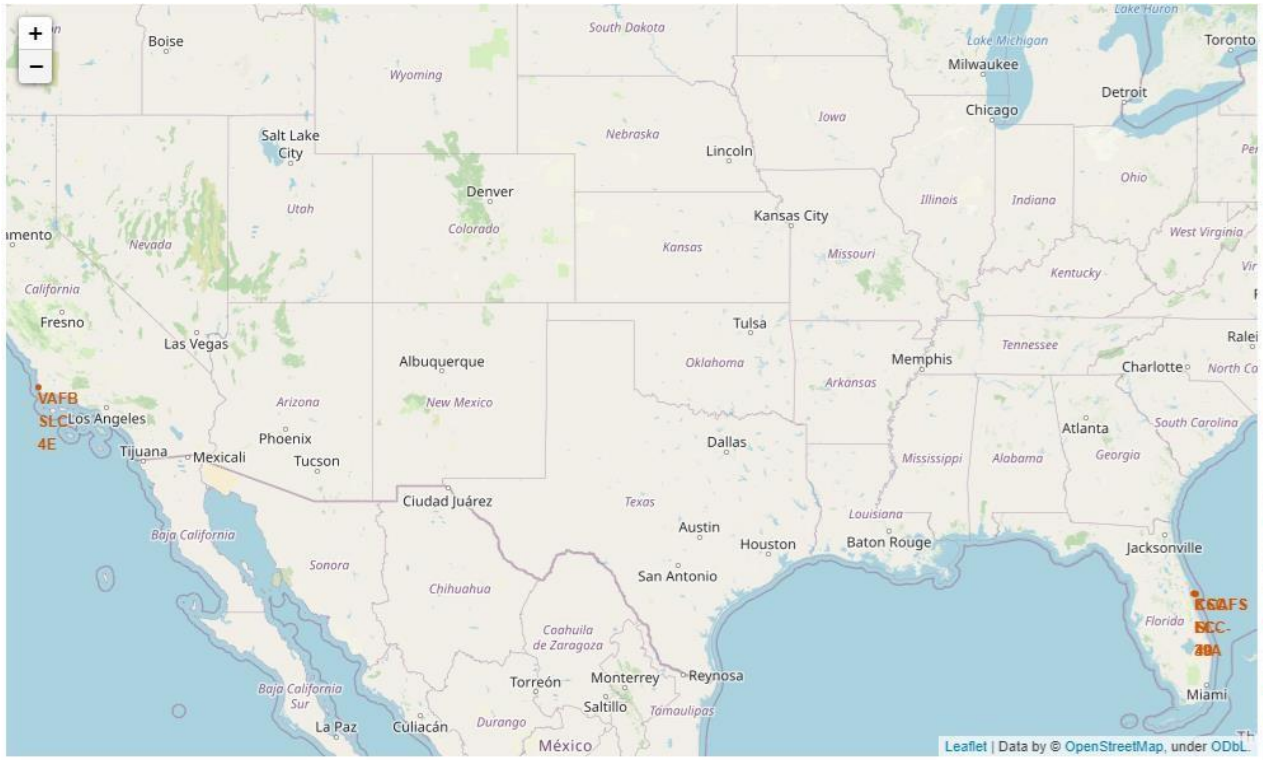
```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

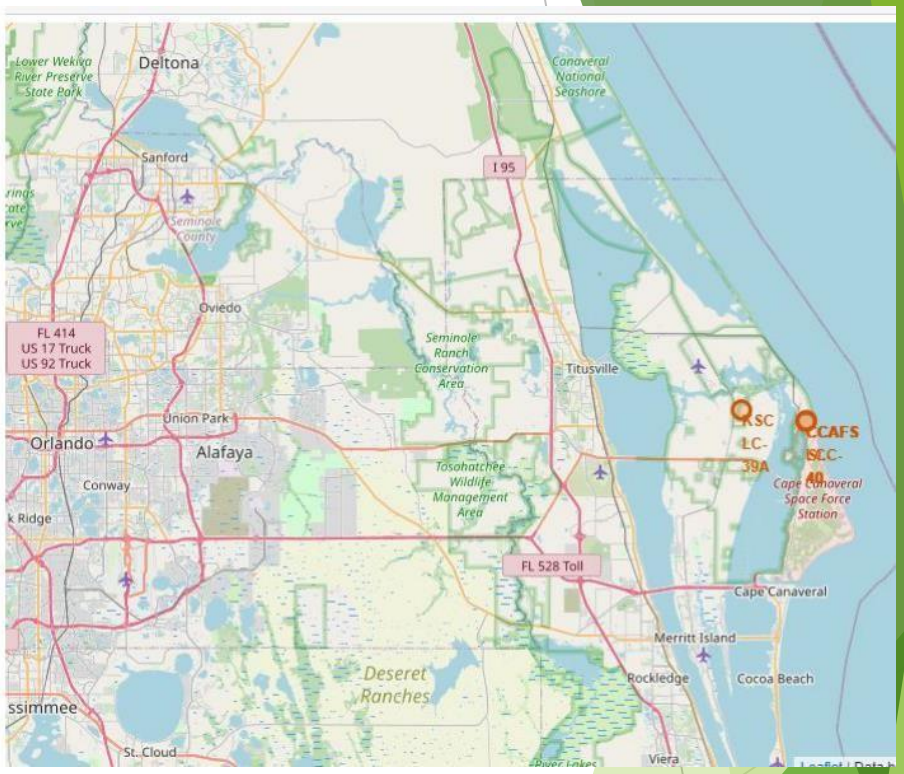
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

- This query retrieves a list of successful landings between 2010-06-04 and 2017-03-20 (inclusive).
- The successful landing outcomes are categorized into two types: drone ship landings and ground pad landings.
- A total of 8 successful landings were recorded during this period.

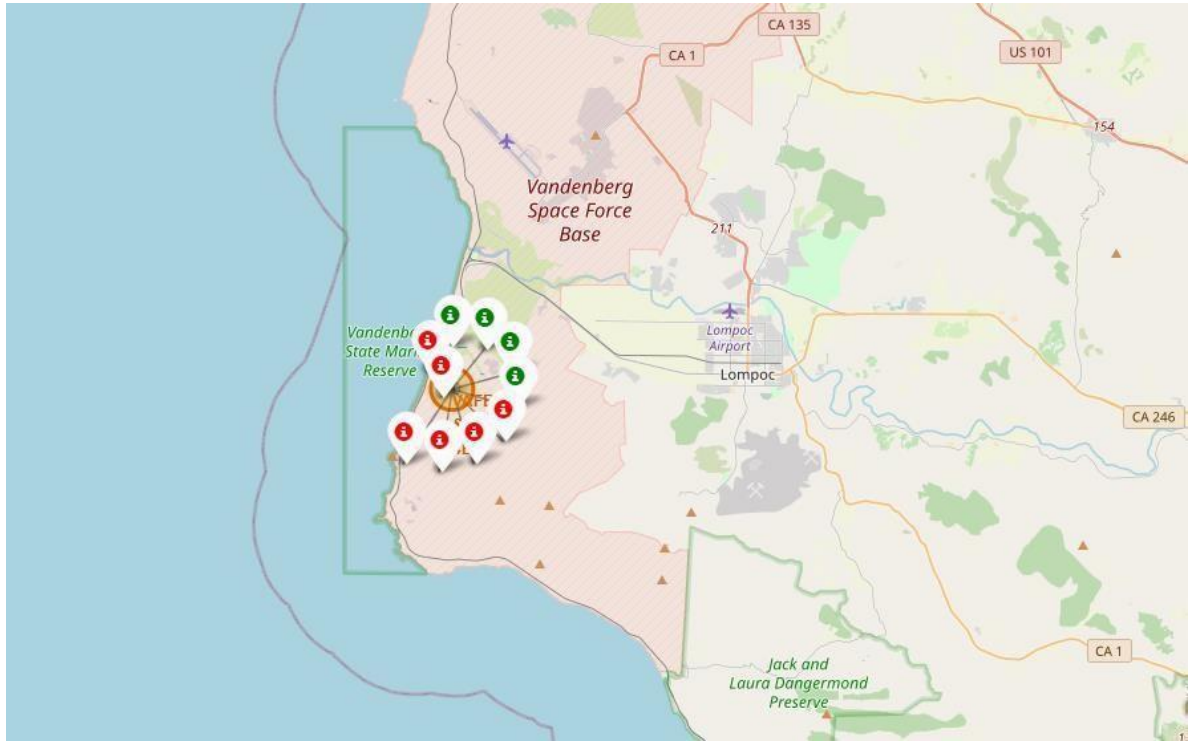
Launch Site Locations





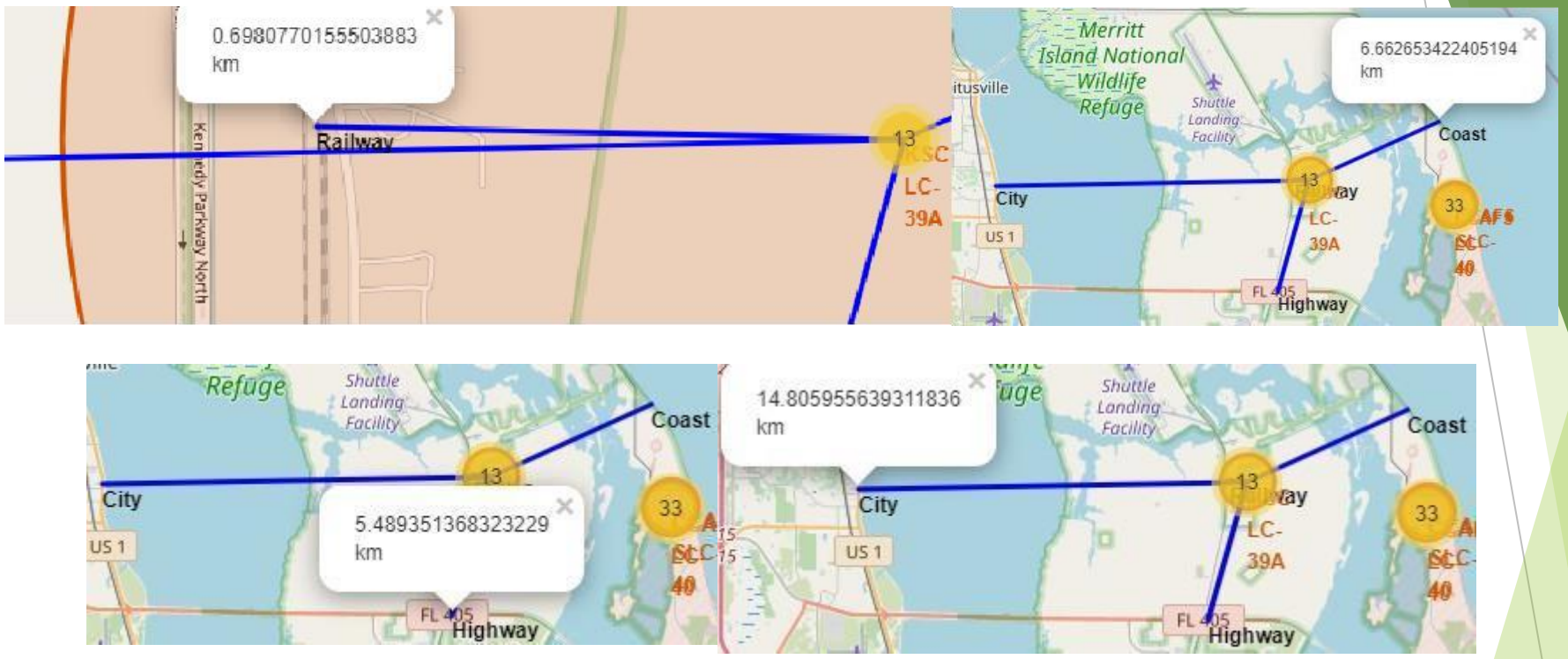
- The map on the left displays all launch sites in relation to the U.S. map.
- The map on the right zooms in on the two Florida launch sites, as they are located close to each other.
- All launch sites are strategically positioned near the ocean.

Color-Coded Launch Markers



- Clusters on the Folium map are interactive and can be clicked to display details of each landing.
- Successful landings are marked with green icons, while failed landings are marked with red icons.
- For example, VAFB SLC-4E shows 4 successful landings and 6 failed landings in this visualization.

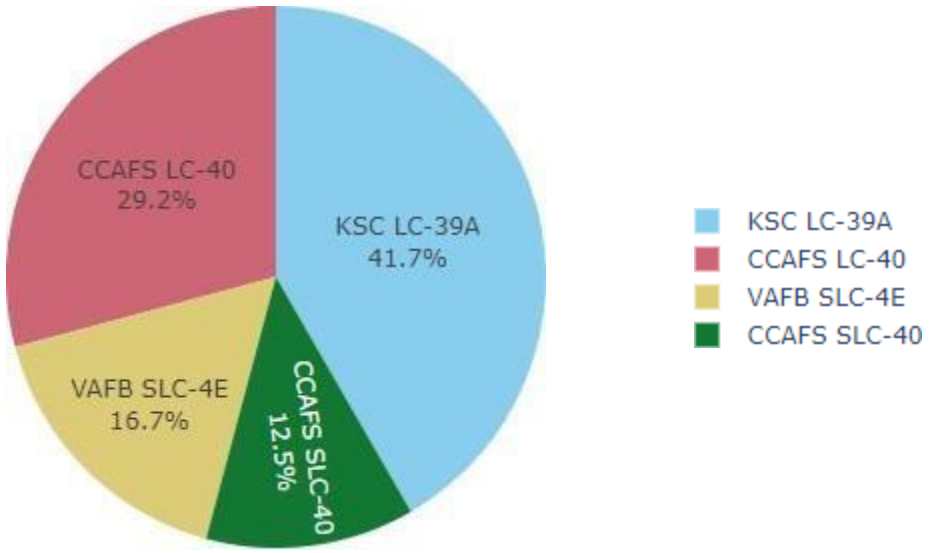
Key Location Proximities



Using KSC LC-39A as an example:

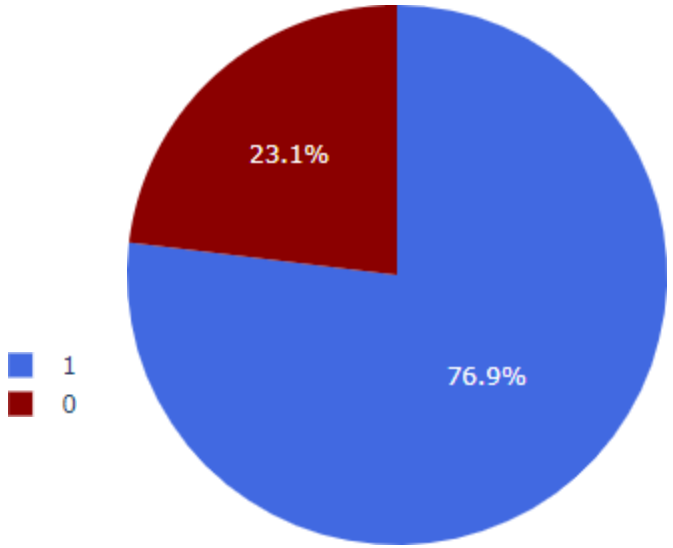
- Launch sites are strategically located near railways to facilitate the transportation of large parts and supplies.
- They are also close to highways to enable efficient transport of personnel and additional supplies.
- Proximity to the coast ensures that in the event of a launch failure, rockets can safely land in the sea, minimizing risk to densely populated areas.

Distribution of Successful Launches Across Launch Sites



- This visualization shows the distribution of successful landings across all launch sites.
- CCAFS LC-40 is the former name of CCAFS SLC-40, meaning CCAFS and KSC have an equal number of successful landings. However, most of the successful landings occurred before the name change.
- VAFB accounts for the smallest share of successful landings, likely due to a smaller sample size and the added challenges of launching from the West Coast.

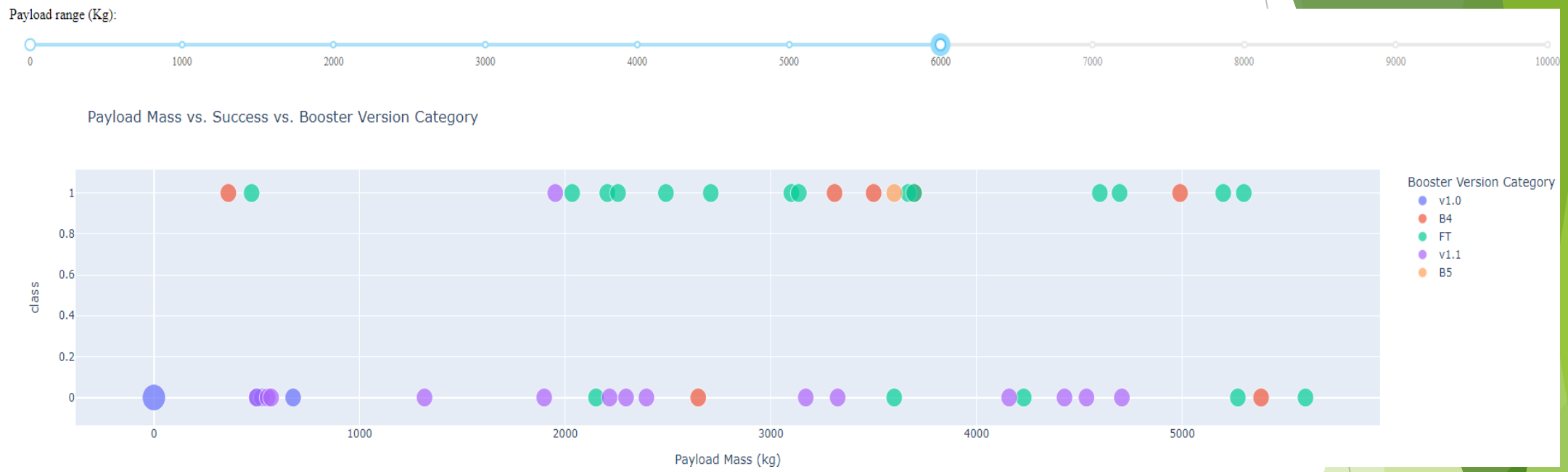
Launch Site with the Highest Success Rate



The success rate for KSC LC-39A is depicted, with blue representing successful landings.

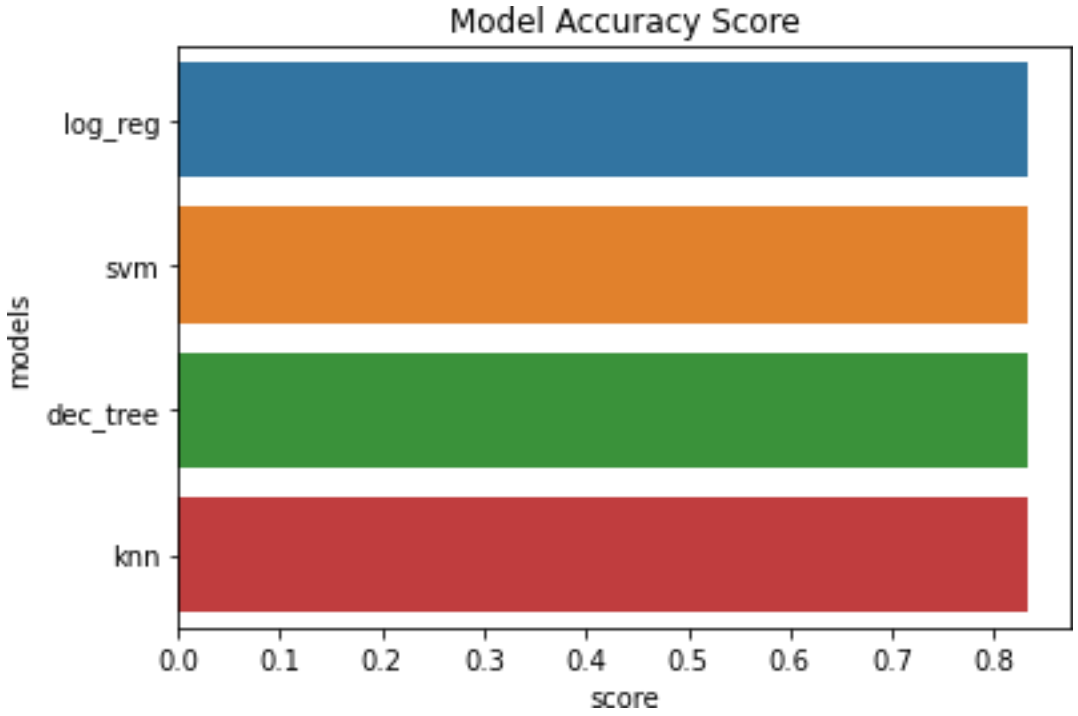
KSC LC-39A boasts the highest success rate, with 10 successful landings and 3 failed landings.

Payload Mass vs. Success vs. Booster Version Category



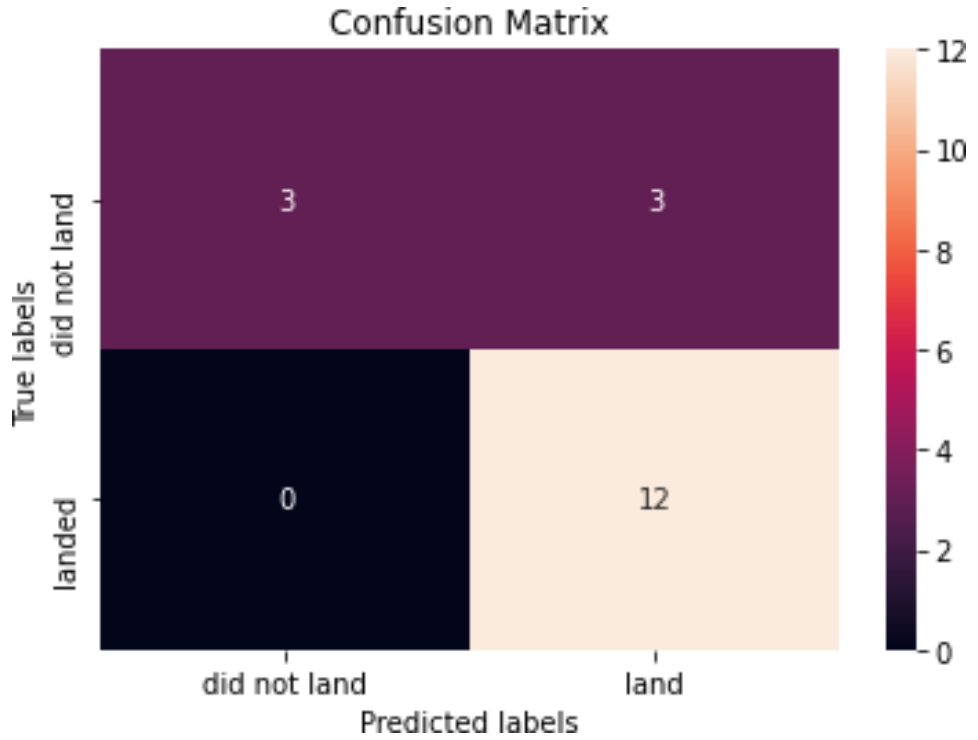
- The dashboard includes a payload range selector, configured from 0 to 10,000 kg instead of the maximum payload of 15,600 kg.
- The Class field indicates successful landings (1) and failures (0).
- The scatter plot visualizes booster version categories using color and represents the number of launches through point size.
- In the 0–6,000 kg range, it is notable that there are two failed landings with payloads of zero kilograms.

Classification Accuracy



- All models achieved a similar accuracy of 83.33% on the test set.
- However, the test set is relatively small, with a sample size of only 18, which may lead to significant variance in accuracy results.
- This variability is particularly evident in the Decision Tree Classifier during repeated runs.
- Additional data is likely needed to accurately determine the best-performing model.

Confusion Matrix



Correct predictions are represented along the diagonal from the top-left to the bottom-right.

- As all models performed equally on the test set, their confusion matrices are identical.
- The models correctly predicted 12 successful landings when the true label was a successful landing.
- They also correctly predicted 3 unsuccessful landings when the true label was unsuccessful.
- However, the models incorrectly predicted 3 successful landings when the true label was unsuccessful (false positives).
- This indicates that the models tend to over-predict successful landings.

CONCLUSION

- Task: Develop a machine learning model for SpaceY, aiming to compete with SpaceX.
- Objective: Predict successful Stage 1 landings to save approximately $100 million per launch.
- Data Source: Combined data from SpaceX's public API and web scraping SpaceX's Wikipedia page.
- Process:
  - Created data labels and stored the dataset in a DB2 SQL database.
  - Built a dashboard for data visualization.
  - Developed a machine learning model achieving 83% accuracy.
- Application: SpaceY’s Allon Mask can leverage this model to predict the likelihood of a successful Stage 1 landing before launch, helping to decide whether the launch should proceed.
- Recommendation: Collect more data to improve model accuracy and refine the selection of the best machine learning model.