IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Fabrice Achu Ngando
October 22, 2025

Outline

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

Executive Summary

Project Goal

Predicting whether the Falcon 9 first stage will successfully land after launch is crucial for understanding mission efficiency and cost optimization. Accurate predictions of landing success enable better estimation of launch expenses, as reusable rockets greatly lower overall costs for SpaceX and future commercial space ventures.

Project Overview

Data Collected:

- SpaceX REST API
- IBM Skills Network datasets
- Performed data wrangling and cleaning to prepare datasets.
- Conducted exploratory data analysis (EDA) using visualizations and SQL queries.
- Built interactive analytics tools using Folium (maps) and Plotly Dash (dashboard).
- Developed and tuned machine learning models (Logistic Regression, SVM, Decision Tree, KNN) using GridSearchCV.

Executive Summary

Key Findings

- Launch site and payload mass are the strongest predictors of landing success.
- Missions to the ISS have the highest success rate.
- Reused boosters show a clear pattern of higher reliability over time.
- Decision Tree Classifier achieved the highest test accuracy (93.33%), outperforming other models.

Conclusion

The analysis confirms that reusability and mission parameters are strong drivers of Falcon 9 success. Accurate landing predictions can help SpaceX and its partners optimize mission planning and reduce launch costs through reliable booster recovery.

Model	Validation Accuracy	Test Accuracy
Logistic Regression	83%	83%
Support Vector Machine	88%	83%
Decision Tree	95%	93%
K-Nearest Neighbors	83%	83%

Introduction

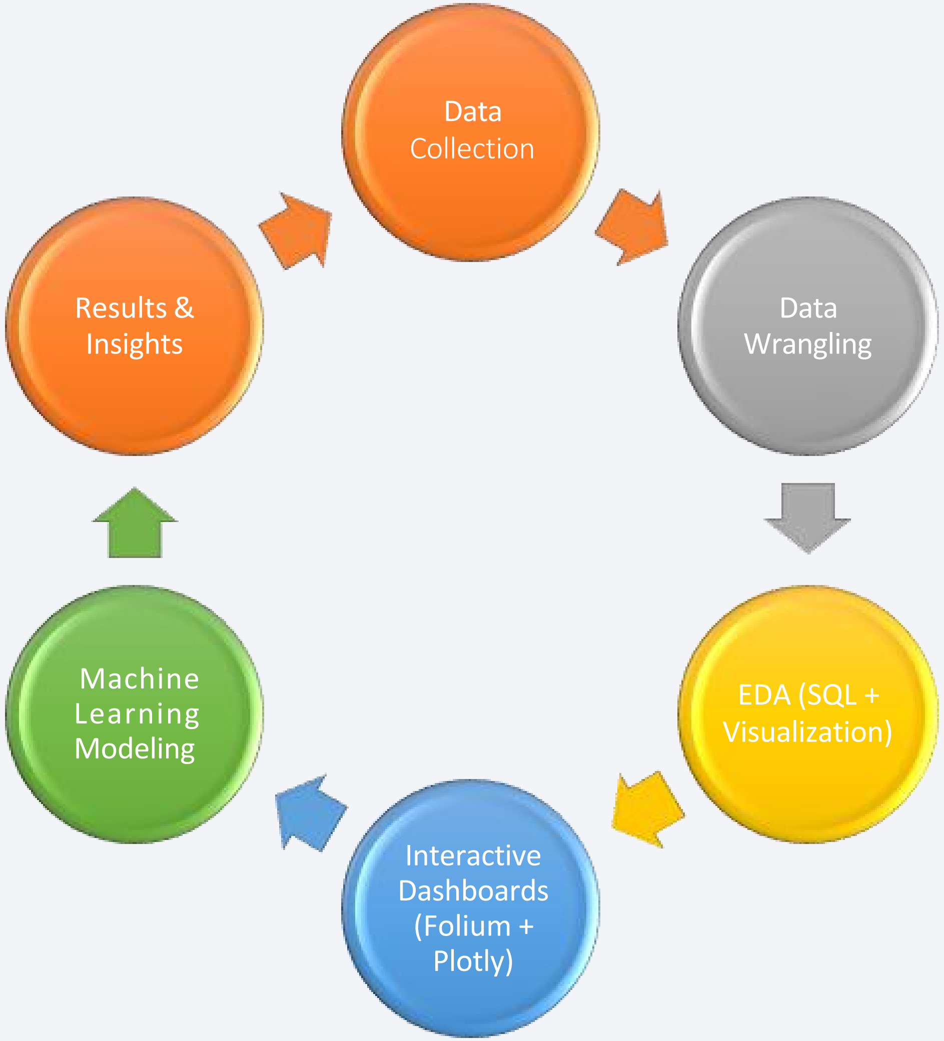
Project Background and Context

SpaceX has revolutionized the commercial space industry by significantly reducing the cost of space travel. While competitors charge over $165 million per launch, a Falcon 9 mission costs roughly $62 million, largely due to SpaceX's groundbreaking ability to reuse the rocket's first stage. This project focuses on predicting the success of Falcon 9 first-stage landings using publicly available data and machine learning techniques. By identifying the key factors that contribute to a successful landing, the project also provides insights into estimating the cost efficiency of future launches.

Key Questions

- How do factors like payload mass, launch site, number of flights, and orbit type impact the success of first-stage landings?
- Has the success rate of first-stage landings improved over the years?
- Which machine learning algorithm performs best for this binary classification problem?

Introduction



Project Flow:

Data Collection →
Data Wrangling →
EDA (SQL + Visualization) →
Interactive Dashboards (Folium + Plotly) →
Machine Learning Modeling →
Results & Insights

Section 1

Methodology

Methodology

Data Collection Methodology

- Collected data from the SpaceX REST API and through web scraping from Wikipedia.
- Conducted data wrangling by:
    - Filtering and cleaning the dataset
    - Handling missing values
    - Applying One-Hot Encoding to prepare the data for binary classification

Exploratory and Visual Analysis

- Performed exploratory data analysis (EDA) using visualization tools and SQL queries.
- Created interactive visualizations with Folium and Plotly Dash to explore spatial and temporal patterns.

Predictive Analysis

- Built and fine-tuned classification models to predict first-stage landing success.
- Evaluated model performance to identify the most accurate and reliable algorithm.

Data Collection

GitHub Link

Data Collection Process

The dataset was collected using two approaches: retrieving data through the SpaceX REST API and web scraping from SpaceX's official Wikipedia page. Combining both methods ensured a complete and reliable dataset, enabling a more comprehensive analysis of Falcon 9 launch records.

Data Columns from SpaceX REST API:

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

Data Columns from Wikipedia Web Scraping:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

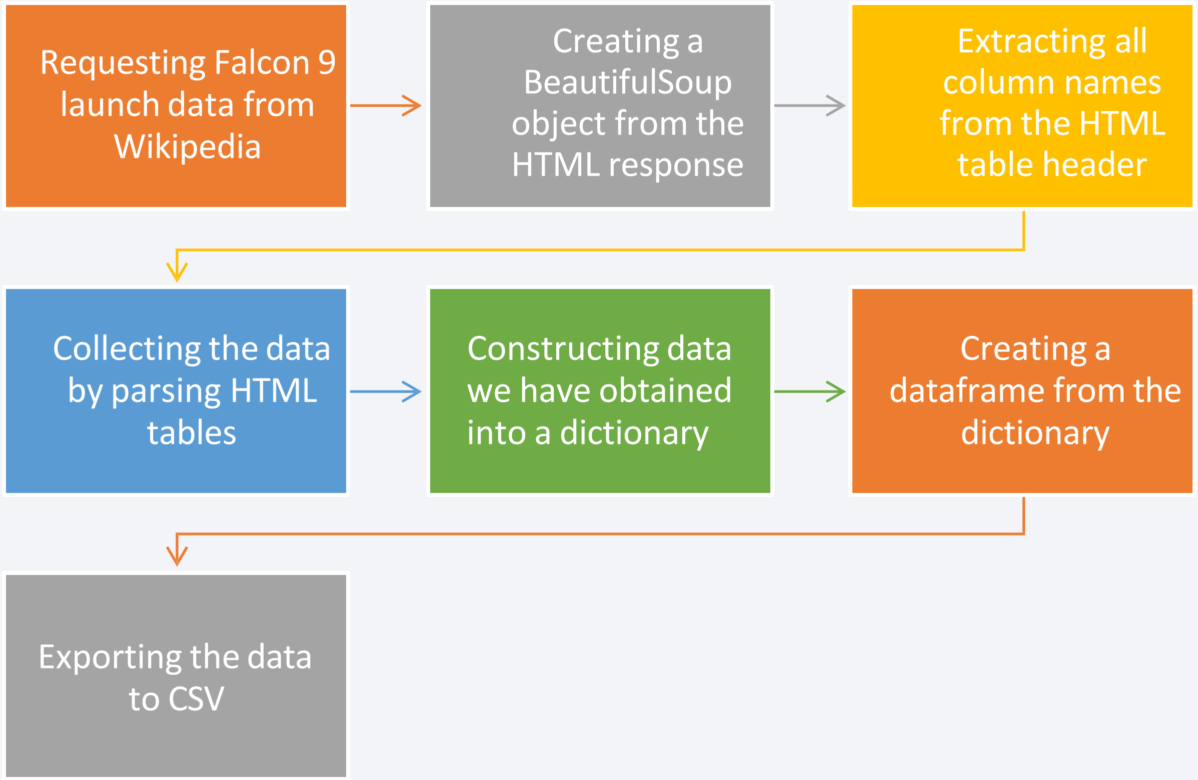
GitHub Link

1. Requesting rocket launch data from SpaceX API
2. Decoding the response content using .json() and turning it into a dataframe using .json_normalize()
3. Requesting needed information about the launches from SpaceX API by applying custom function
4. Constructing data we have obtained into a dictionary
5. Creating a dataframe from the dictionary
6. Filtering the dataframe to only include Falcon 9 launches
7. Replacing missing values of Payload Mass column with calculated .mean() for this column
8. Exporting the data to CSV

Data Collection - Scraping

[GitHub Link](#)

I initiated the data collection via web scraping by requesting the Wikipedia page for SpaceX launches of the Falcon 9. Using BeautifulSoup, the HTML response was parsed and relevant launch-data columns were identified and extracted. These values were then assembled into a Python dictionary, converted to a pandas DataFrame, and finally exported as a CSV file for further analysis.



Data Wrangling

GitHub Link

The dataset contains multiple instances where the Falcon 9 booster did not achieve a successful landing. Some missions attempted landings but ended in failure due to various incidents. The landing outcomes are categorized as follows:

- **True Ocean** – Booster landed successfully in a specific ocean region.
- **False Ocean** – Booster failed to land successfully in a specific ocean region.
- **True RTLS** – Booster landed successfully on a ground pad (Return to Launch Site).
- **False RTLS** – Booster failed to land successfully on a ground pad.
- **True ASDS** – Booster landed successfully on a drone ship (Autonomous Spaceport Drone Ship).
- **False ASDS** – Booster failed to land successfully on a drone ship.

These outcomes were later simplified into binary training labels: 1 for Success and 0 for Failure.

Data Wrangling

GitHub Link

1. Perform exploratory Data Analysis and determine Training Labels
2. Calculate the number of launches on each site
3. Calculate the number and occurrence of each orbit
4. Calculate the number and occurrence of mission outcome per orbit type
5. Create a landing outcome label from Outcome column
6. Exporting the data to CSV

EDA with Data Visualization

GitHub Link

Several visualizations were developed to analyze various aspects of the Falcon 9 launch data. These included comparisons such as **Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs. Orbit Type, and the yearly trend of Success Rate**. Scatter plots were employed to explore relationships between numerical variables, helping to identify potential predictors for machine learning models. Bar charts were used to compare categorical data and show how different factors influenced launch outcomes, while line charts illustrated trends over time, particularly in the success rate and other key variables.

EDA with SQL

GitHub Link

- Retrieved the unique launch site names from the dataset.
- Displayed five records of launch sites beginning with "CCA."
- Calculated the total payload mass for NASA (CRS) missions.
- Determined the average payload mass for booster version **F9 v1.1**.
- Identified the date of the first successful ground pad landing.
- Listed boosters with successful drone ship landings and payloads between **4000–6000 kg.**
- Counted the total number of successful and failed mission outcomes.
- Found the booster versions that carried the maximum payload mass.
- Listed failed drone ship landings in **2015**, including booster versions and launch sites.
- Ranked landing outcomes (success/failure) between **2010-06-04** and **2017-03-20** in descending order.

Build an Interactive Map with Folium

GitHub Link

- **Launch Site Markers:** A starting marker was added at the NASA Johnson Space Center with a circle, popup, and text label based on its latitude and longitude. Additional markers were created for all other launch sites to visualize their geographic positions and proximity to the equator and nearby coastlines.
- **Launch Outcome Visualization:** Colored markers represented launch results — green for successful missions and red for failed ones. A Marker Cluster was used to easily identify which launch sites achieved higher success rates.
- **Distance Mapping:** Colored lines were drawn from the KSC LC-39A launch site to nearby features such as railways, highways, the coastline, and the closest city, allowing clear visualization of spatial relationships and distances.

Build a Dashboard with Plotly Dash

GitHub Link

The interactive dashboard included several dynamic features for better data exploration:

- **Launch Site Dropdown:** A dropdown menu was added to let users select a specific launch site for focused analysis.
- **Success Rate Pie Chart:** A pie chart displayed the total number of successful launches across all sites, or the ratio of successful to failed launches when a specific site was chosen.
- **Payload Mass Slider:** A slider allowed users to filter and view data within a selected payload mass range.
- **Payload vs. Success Scatter Plot:** A scatter chart illustrated the relationship between payload mass and launch success across different booster versions.

Predictive Analysis (Classification)

[GitHub Link](#)

STEP 1
Creating a NumPy array from the column "Class" in data

STEP 2
Standardizing the data with StandardScaler, then fitting and transforming it

STEP 3
Splitting the data into training and testing sets with train_test_split function

STEP 4
Creating a GridSearchCV object with cv = 10 to find the best parameters

STEP 5
Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

STEP 6
Calculating the accuracy on the test data using the method .score() for all models

STEP 7
Examining the confusion matrix for all models

STEP 8
Finding the method performs best by examining the Jaccard_score and F1_score metrics

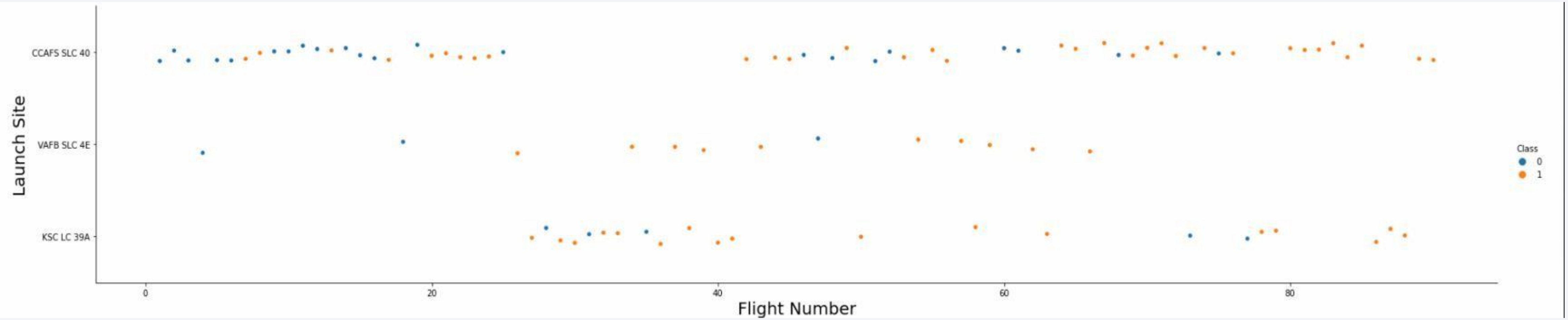
Results

1. **Launch Success Trends:** The success rate of Falcon 9 landings has improved steadily over the years, showing significant progress in booster recovery technology.
2. **Impact of Payload Mass:** Heavier payloads generally had a lower success rate, suggesting that payload mass influences the difficulty of achieving a safe landing.
3. **Launch Site Performance:** Certain launch sites, such as **KSC LC-39A**, showed higher success rates compared to others, indicating better operational efficiency or environmental advantages.
4. **Orbit Type Correlation:** Launches targeting **LEO (Low Earth Orbit)** had a higher probability of successful landings than those aimed at higher or more complex orbits.
5. **Model Accuracy:** Machine learning classification models accurately predicted landing outcomes, confirming that variables like payload mass, orbit type, and flight number are strong predictors of landing success.

Section 2

Insights drawn from EDA

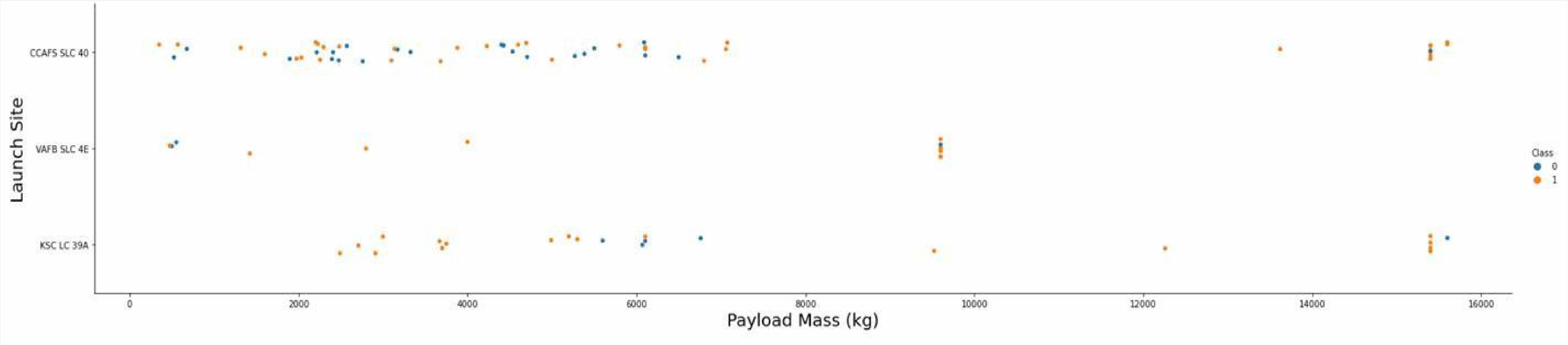
Flight Number vs. Launch Site



Explanation:

- The initial Falcon 9 flights were unsuccessful, while more recent missions have all achieved successful landings.
- The **CCAFS SLC-40** site handled roughly half of all recorded launches.
- Launch sites **VAFB SLC-4E** and **KSC LC-39A** demonstrated higher overall success rates.
- Overall, the data suggests that with each new launch, the probability of success has increased over time.

Payload vs. Launch Site



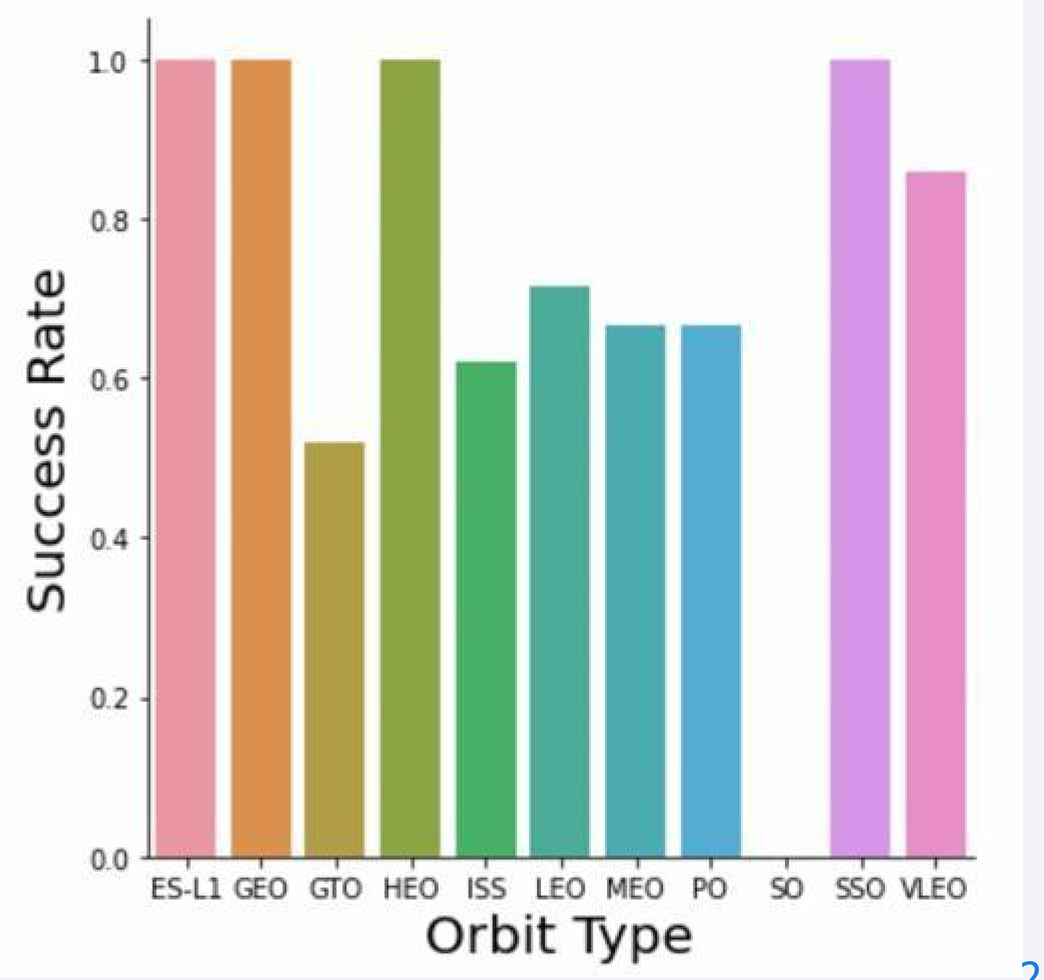
Explanation:

- Across all launch sites, a higher payload mass generally corresponds to a higher success rate.
- Most launches carrying payloads above **7000 kg** were successful.
- The **KSC LC-39A** site achieved a **100% success rate** for launches with payload masses below **5500 kg**.

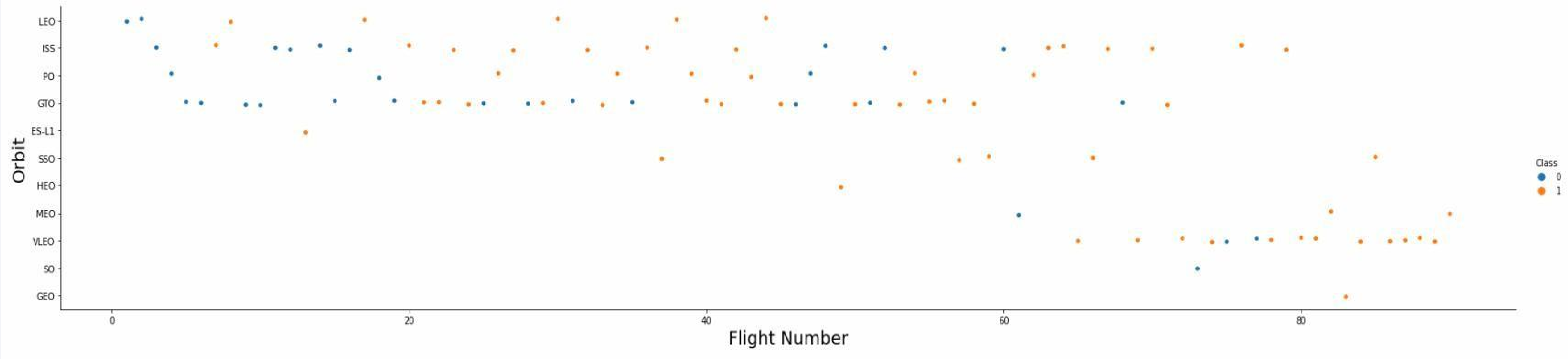
Success Rate vs. Orbit Type

Explanation:

- Orbits with a **100% success rate** include **ES-L1, GEO, HEO,** and **SSO**.
- The **SO** orbit recorded a **0% success rate**.
- Orbits such as **GTO, ISS, LEO, MEO,** and **PO** had **moderate success rates**, ranging between **50% and 85%**.



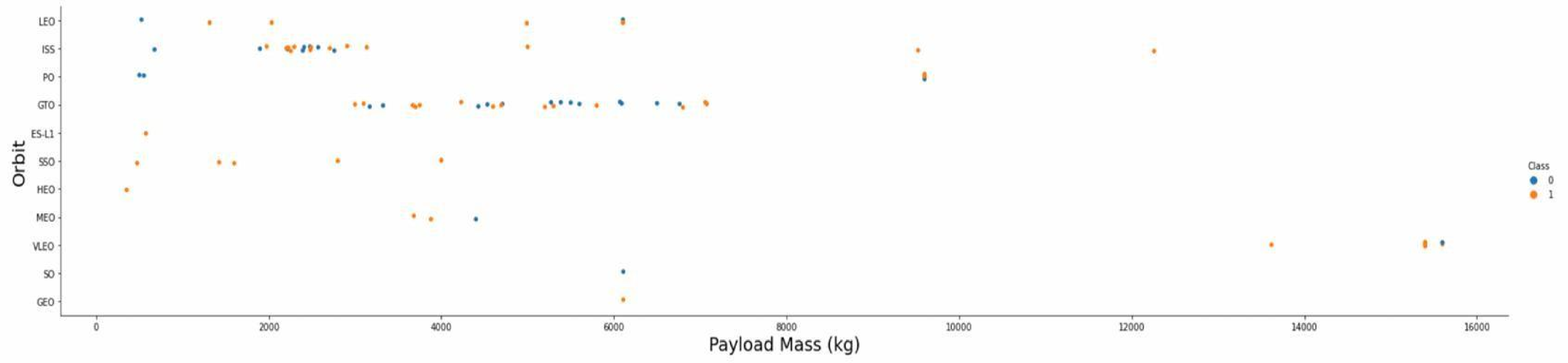
Flight Number vs. Orbit Type



Explanation:

- In the **LEO orbit**, success tends to increase with the number of flights, suggesting that experience and repeated missions improve outcomes.
- In contrast, for the **GTO orbit**, there is no clear relationship between the flight number and landing success.

Payload vs. Orbit Type



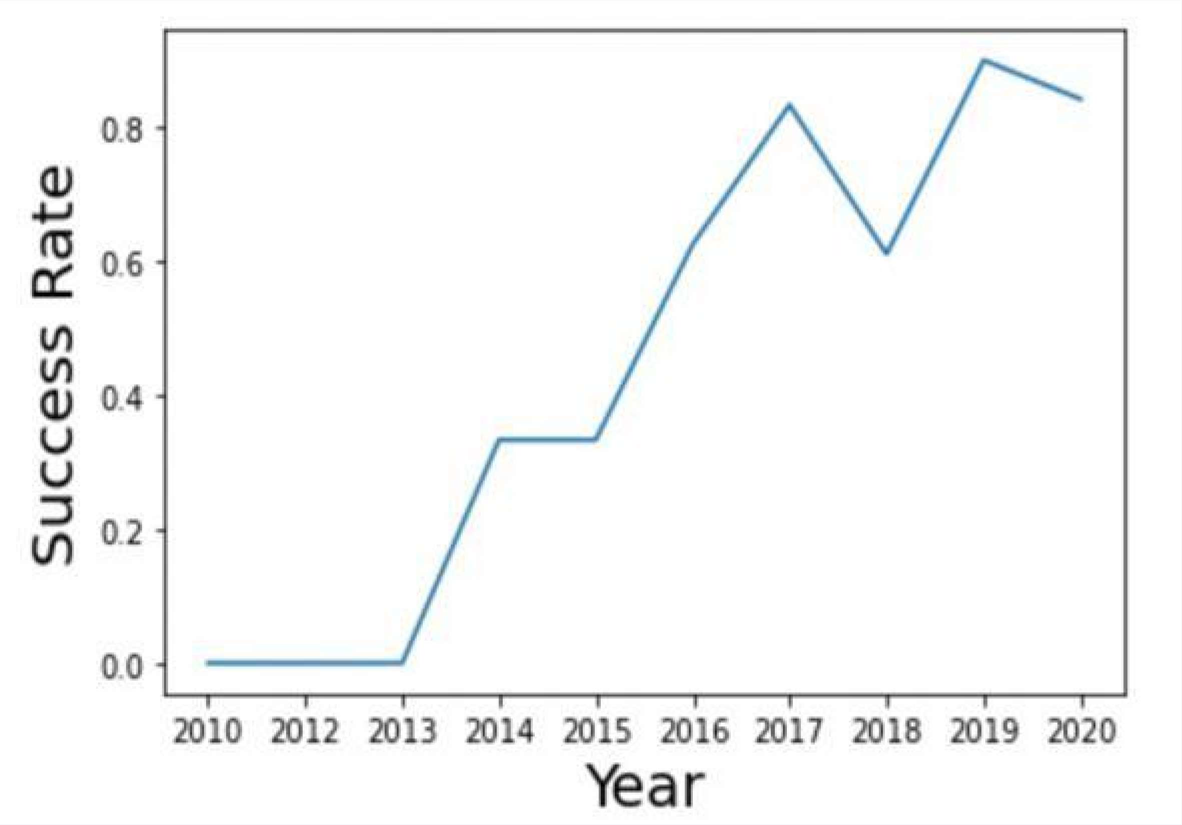
Explanation:

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

Launch Success Yearly Trend

Explanation:

- The success rate since 2013 kept increasing till 2020.



All Launch Site Names

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[4]:
```

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

Explanation:

- This query lists all the **unique launch site names** used in the SpaceX missions, helping identify the different locations from which Falcon 9 rockets were launched.

Launch Site Names Begin with 'CCA'

```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;

 * ibm_db_sa://wzf08322:***@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

Explanation:

- This query retrieves **five records** of launch sites whose names start with **“CCA”**, allowing us to view sample entries related to the **Cape Canaveral** launch locations.

Total Payload Mass

```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[6]:
```

total_payload_mass
45596

Explanation:

- This query calculates the **total payload mass** of all missions launched by **NASA (CRS)**, providing insight into the overall cargo capacity handled by SpaceX for NASA's Commercial Resupply Services.

Average Payload Mass by F9 v1.1

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[7]:

average_payload_mass
2534

Explanation:

- This query computes the **average payload mass** for missions using the **F9 v1.1 booster version**, helping to understand the typical payload capacity associated with that specific booster model.

First Successful Ground Landing Date

```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[8]:

first_successful_landing
2015-12-22

Explanation:

- This query identifies the **date of the first successful ground pad landing**, marking a key milestone in SpaceX's progress toward reusable rocket technology.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
000 and 6000;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[9]:

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

Explanation:

- This query lists the **boosters that successfully landed on a drone ship** and carried payloads **between 4000 and 6000 kg**, helping analyze performance under specific payload conditions.

Total Number of Successful and Failure Mission Outcomes

In [10]:
```
%sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[10]:

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

Explanation:

- This query counts the **total number of successful and failed missions**, providing an overview of SpaceX's overall mission success rate and reliability.

Boosters Carried Maximum Payload

```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[11]:
```

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:

- Finds **booster versions** that carried the **maximum payload mass**.

2015 Launch Records

```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
               where landing__outcome = 'Failure (drone ship)' and year(date)=2015;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[12]:
```

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Explanation:

- Lists **failed drone ship landings** in **2015**, along with their **booster versions** and **launch site names**.

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

In [13]:
```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

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

Explanation:

- Ranks **landing outcomes** (success or failure) between **2010-06-04** and **2017-03-20** in **descending order** by count.

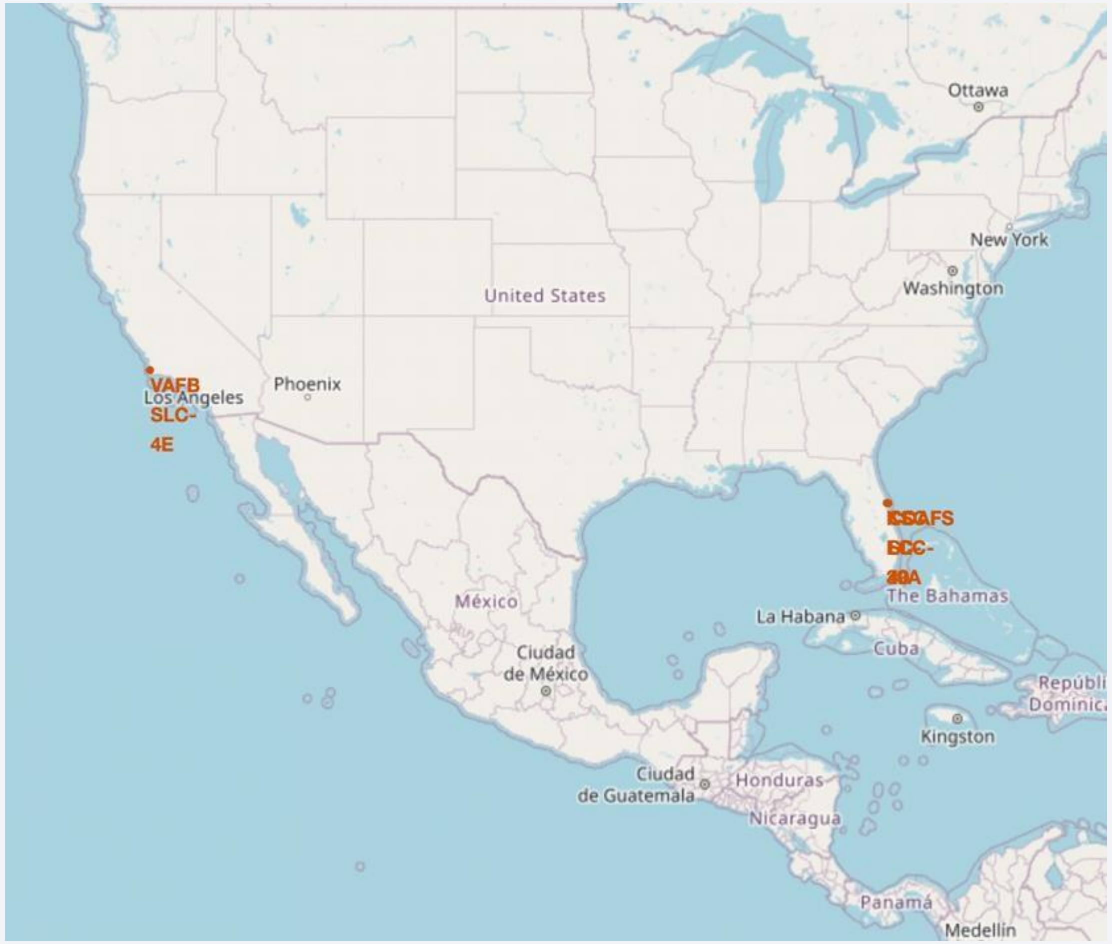
Section 3

Launch Sites Proximities Analysis

All launch sites' location markers on a global map

Explanation:

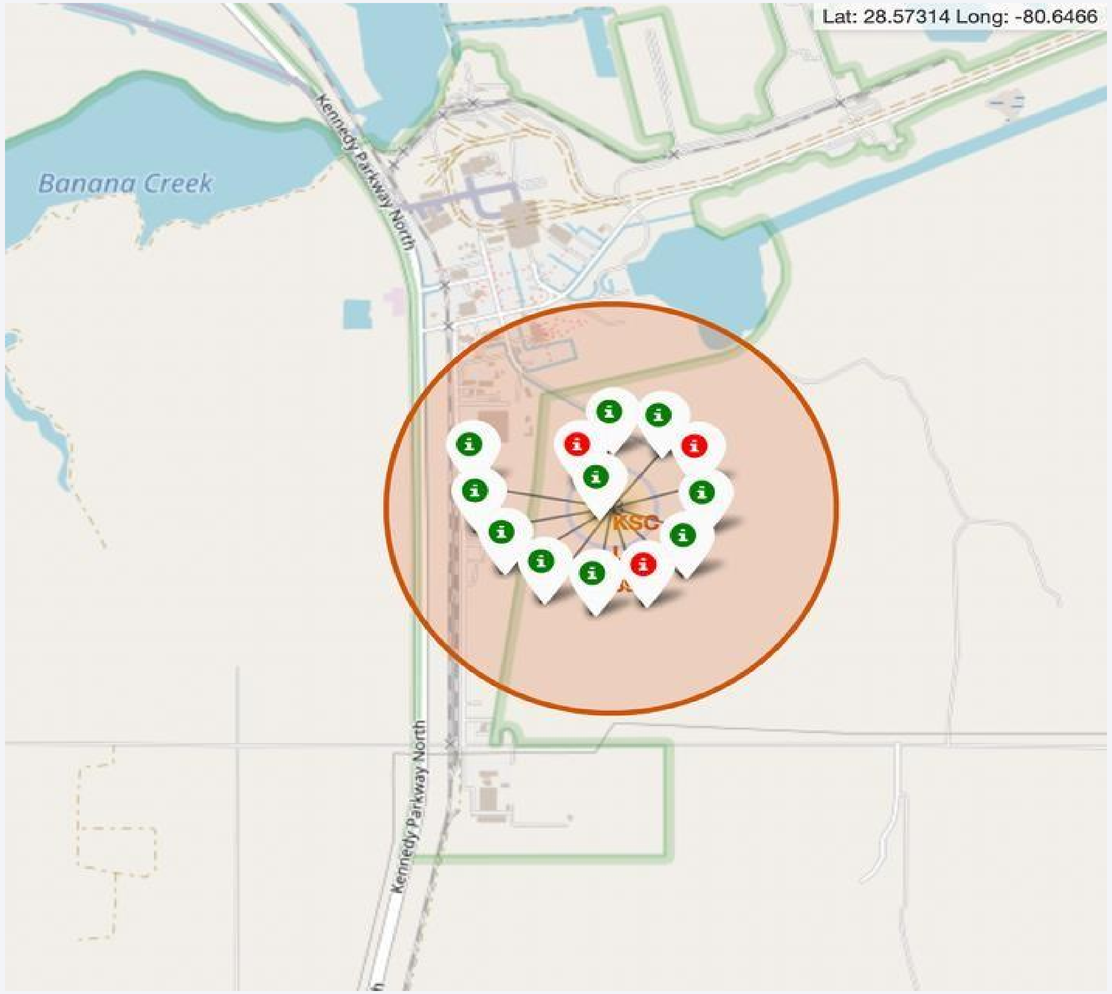
- Most launch sites are located **near the equator**, where the Earth's rotation speed is highest (about **1670 km/h**). Launching from this region provides an **inertial boost**, helping rockets achieve the speed needed to reach orbit more efficiently.
- All launch sites are also positioned **close to the coast**, allowing rockets to be launched over the ocean, which **reduces risk** from falling debris or failed launches near populated areas.



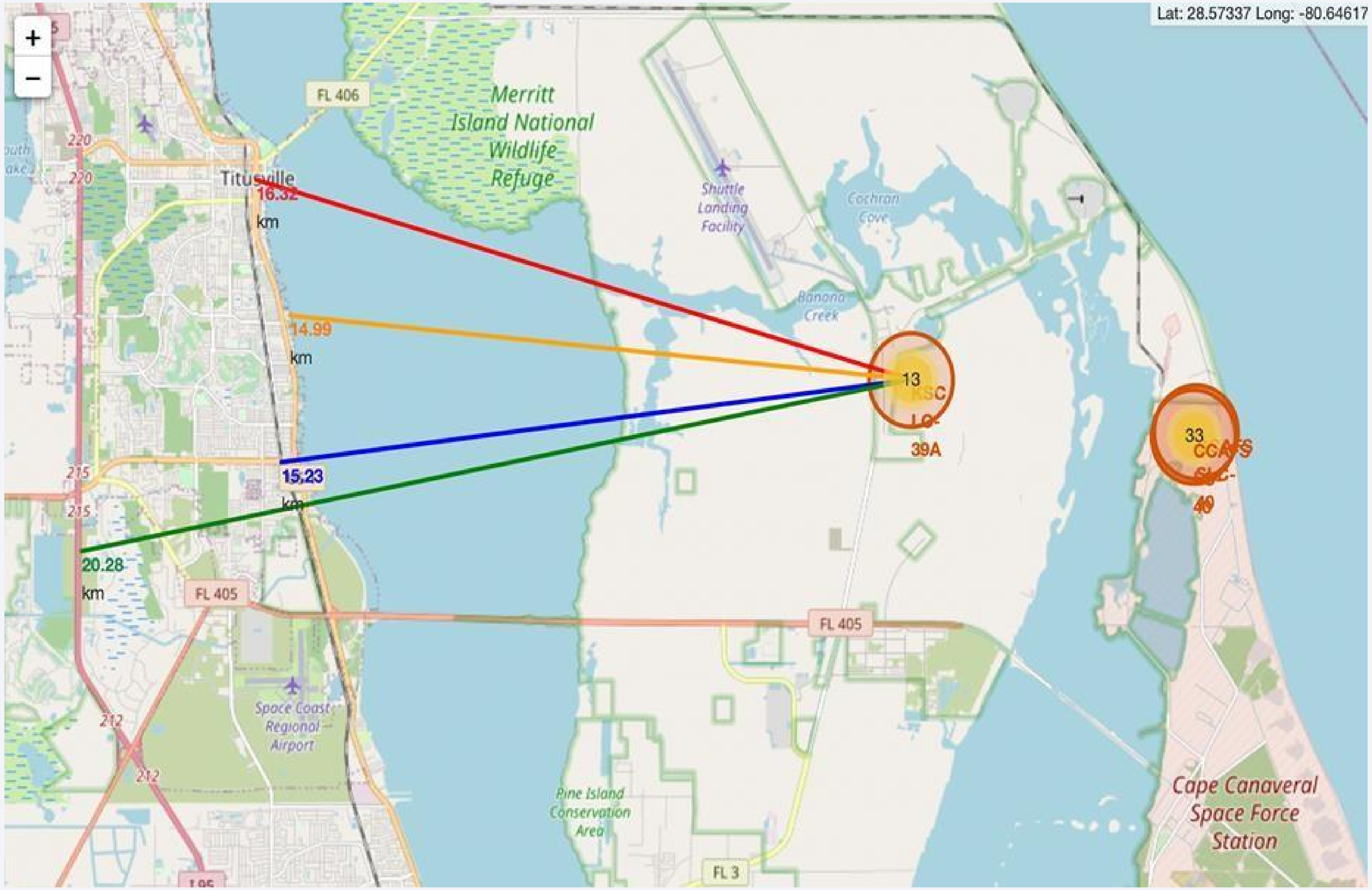
Color-Labeled launch Records on Map

Explanation:

- The **color-coded markers** make it easy to see which launch sites have higher success rates — **green** represents successful launches, while **red** indicates failures.
- The **KSC LC-39A** launch site stands out with a **very high success rate**.



Distance from the launch site KSC LC-39A to its proximities



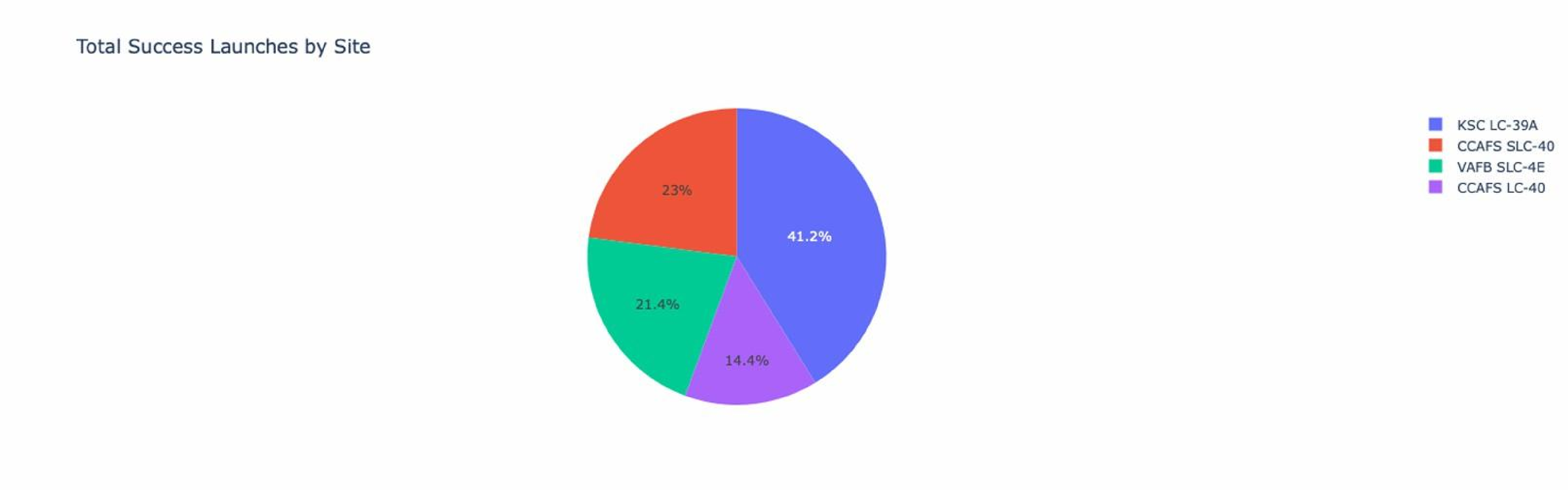
Explanation:

- The visual analysis shows that **KSC LC-39A** is located close to key features: about **15.23 km** from a railway, **20.28 km** from a highway, and **14.99 km** from the coastline.
- The site is also near the city of **Titusville**, roughly **16.32 km away**.
- Since a failed rocket can travel **15–20 km in just a few seconds**, proximity to populated areas poses potential safety risks.

Section 4

Build a Dashboard with Plotly Dash

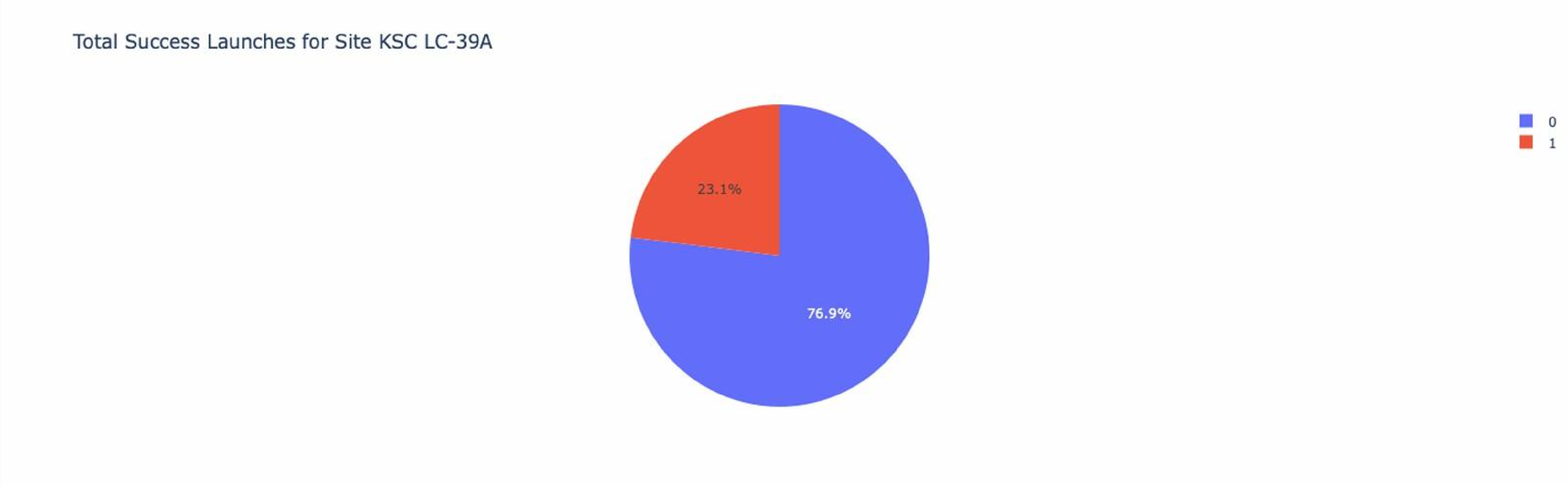
Launch Success Count for all Sites



Explanation:

- The chart indicates that **KSC LC-39A** has recorded the **highest number of successful launches** among all sites.

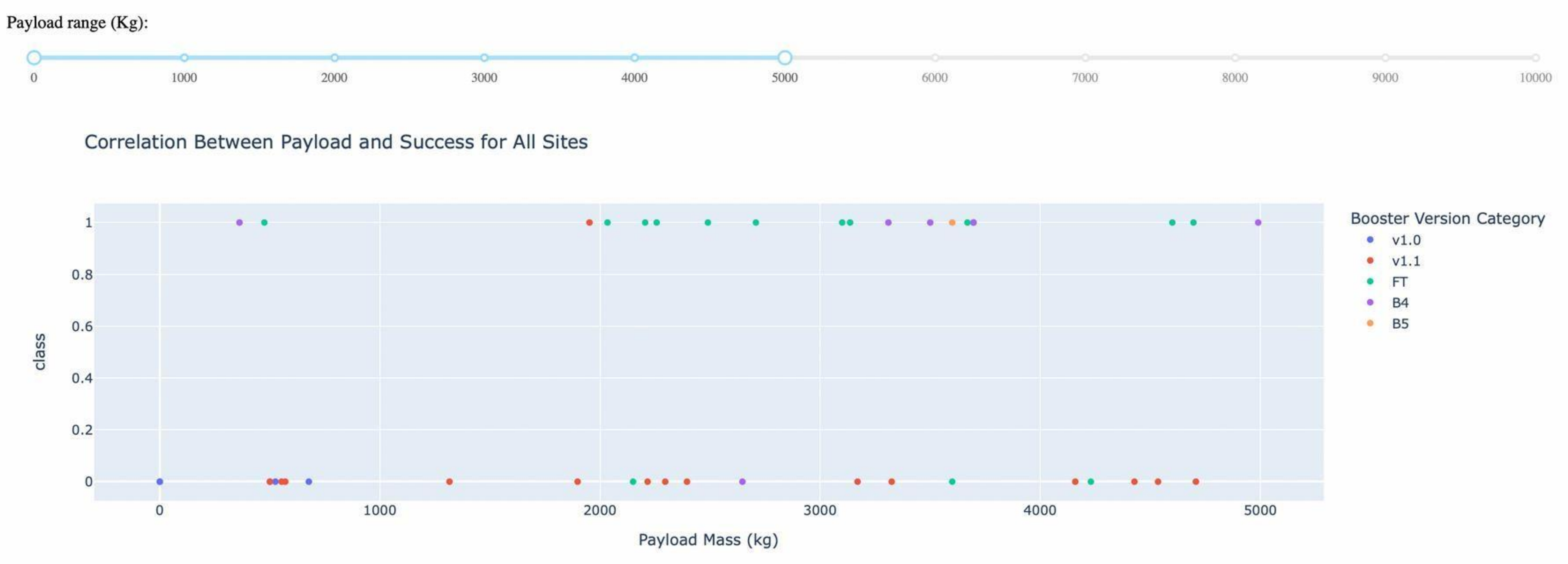
Launch Sites with highest Launch Success Ratio



Explanation:

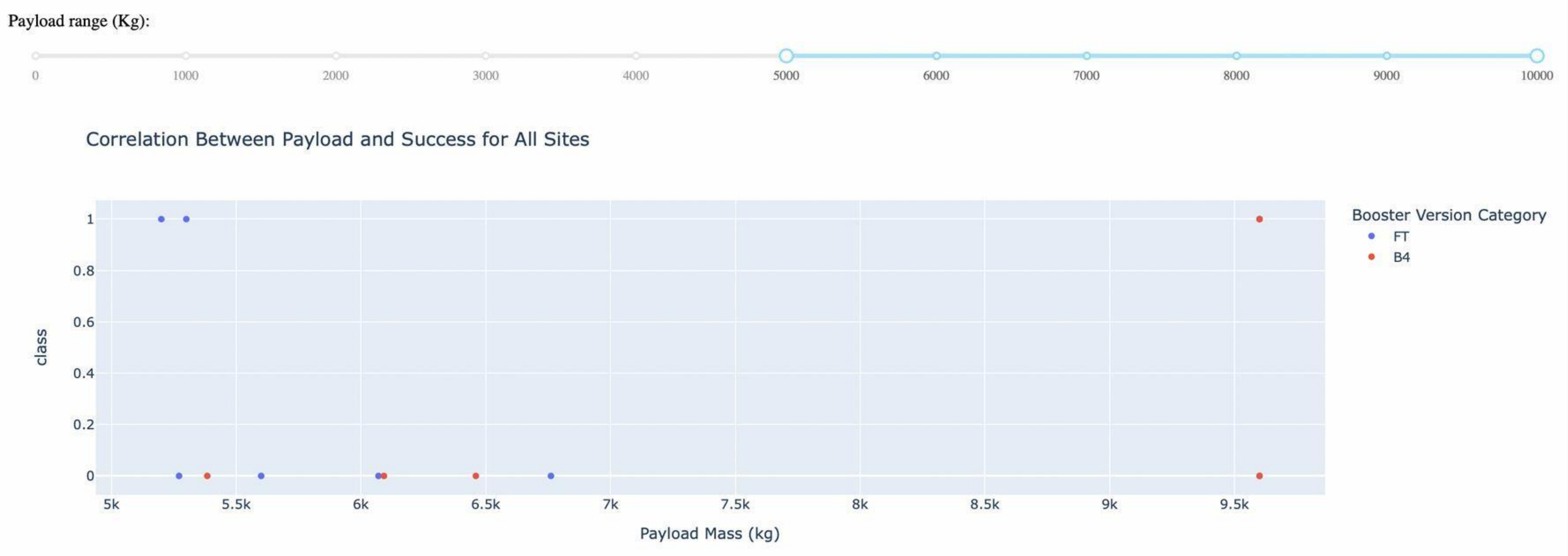
- **KSC LC-39A** achieved the **highest success rate of 76.9%**, with **10 successful** landings and only **3 failures**.

Payload Mass vs. Launch Outcome for all sites



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

Payload Mass vs. Launch Outcome for all sites



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

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Scores and Accuracy of the test set

- The **test set results** did not clearly identify the best-performing model.
- This uncertainty was likely due to the **small test sample size** of only 18 records.

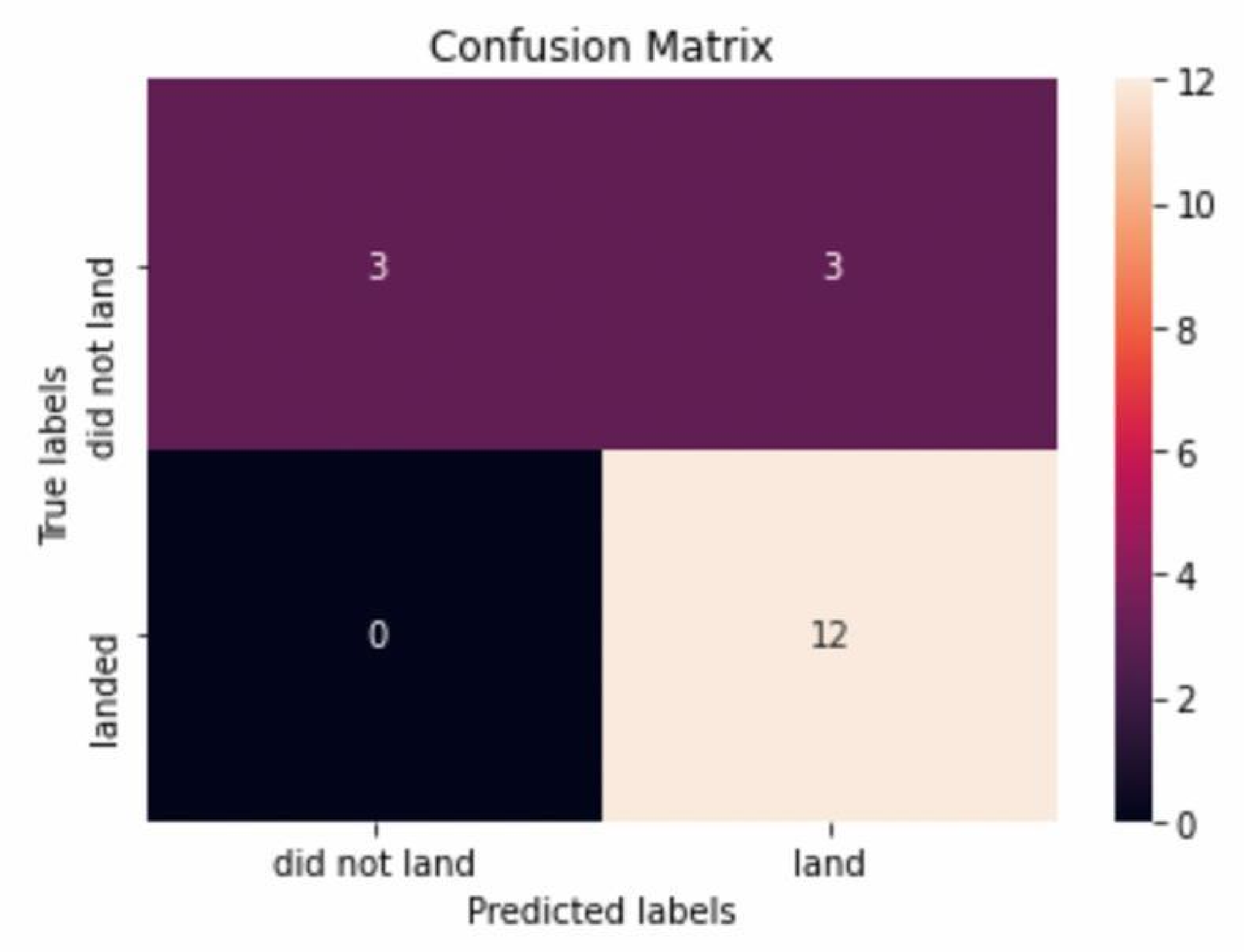
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Scores and Accuracy of the Entire data

- To ensure reliability, all models were re-evaluated using the **entire dataset**.
- The **Decision Tree Model** emerged as the best performer, with the **highest accuracy and overall scores**.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix



Explanation:

- The confusion matrix shows that **logistic regression** is able to differentiate between classes.
- However, the main issue is the presence of a **high number of false positives**, which affects its accuracy.

Conclusions

- Decision Tree Model Performance: The Decision Tree algorithm demonstrated the highest accuracy and effectiveness for predicting Falcon 9 first-stage landing outcomes.
- Payload Impact: Launches with lighter payload masses were more likely to result in successful landings compared to heavier payloads.
- Launch Site Location: Most launch sites are strategically positioned near the equator and coastlines, optimizing launch efficiency and safety.
- Improving Success Rates: The Falcon 9 success rate has steadily increased over time, reflecting advancements in SpaceX's technology and mission consistency.
- Top Performing Site: KSC LC-39A recorded the highest success rate among all launch sites.
- Orbit Success Rates: Launches targeting ES-L1, GEO, HEO, and SSO orbits achieved a 100% success rate.

Appendix

- Course: IBM Data Science Professional Certificate | Coursera
- Coursera: Coursera | Degrees, Certificates, & Free Online Courses
- All resources are available on: My GitHub Account

Thank you!