IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ivan Afanaskin
February 14, 2025

Outline

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

Executive Summary

Summary of Methodologies:

- **Exploratory Data Analysis (EDA):** SQL queries, data visualization (scatter plots, bar charts, line charts).
- **Interactive Analytics:** Folium maps for launch site locations, success rates, and proximities.
- **Predictive Analysis:** Classification models (Logistic Regression, SVM, Decision Tree, KNN) to predict Falcon 9 landings.
-

Summary of Results:

- Success rates improved significantly after 2013, peaking at 80–90%.
- Heavy payloads showed higher success rates for specific orbits (LEO, ISS, Polar).
- Decision Tree achieved the highest accuracy (94.44%), making it the best model for landing prediction.
- CCAFS SLC-40 had the highest launch success ratio among all sites.

Conclusion:

- Data-driven insights and machine learning models **enhance mission planning and optimize cost efficiency for Falcon 9 launches.**

Introduction

Project background and context

- The commercial space industry is evolving, and private companies are making space travel more accessible.
- SpaceX has significantly reduced launch costs by reusing the first stage of the Falcon 9 rocket.
- A new competitor, Space Y, aims to enter the market by predicting whether the first stage can be recovered.

Problems

- Can we predict the successful landing of the Falcon 9 first stage?
- What factors influence the first stage recovery?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

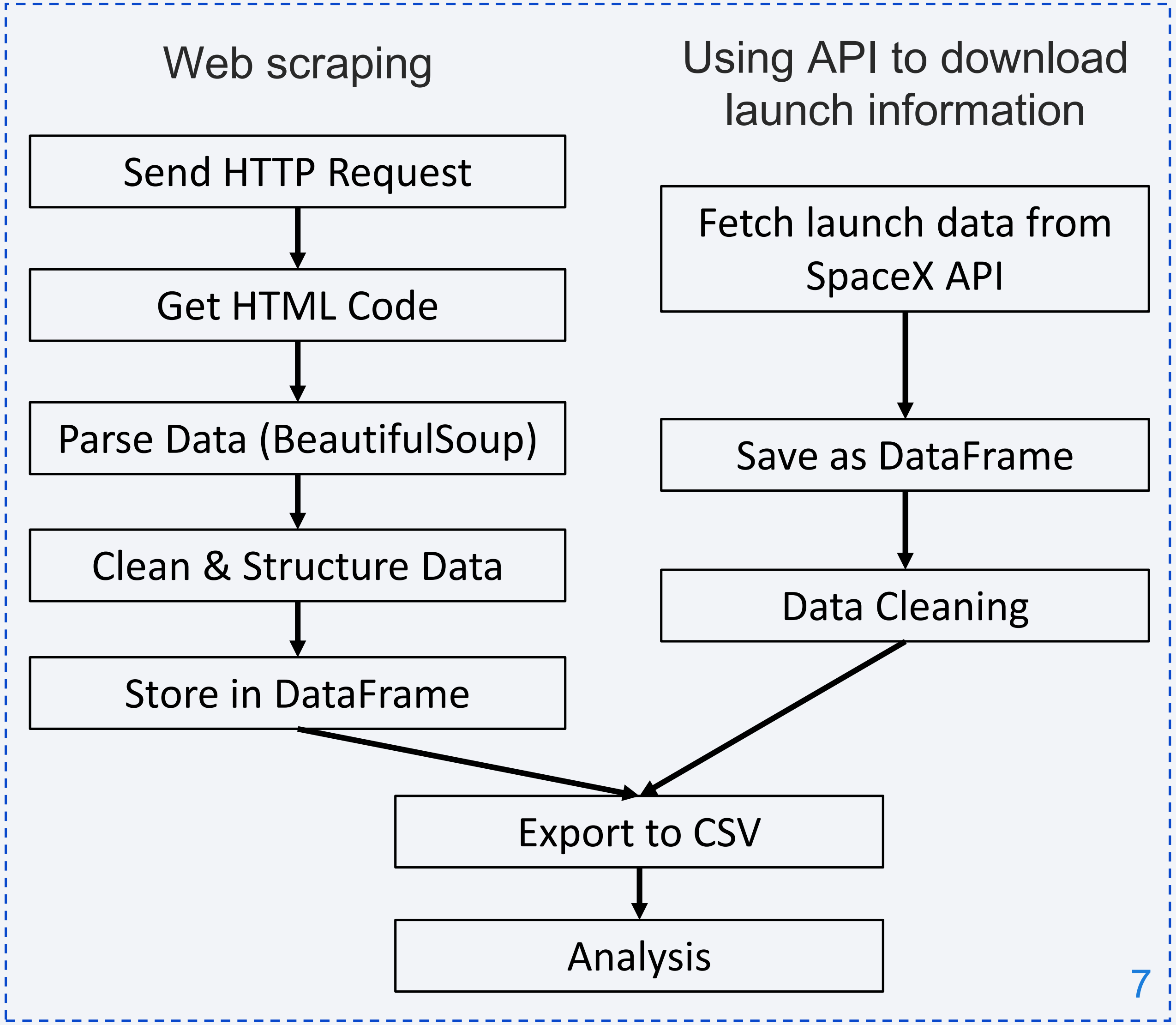
Data Collection

Data sources

1. SpaceX API.
2. Web scraping from SpaceX website.

How data was collected

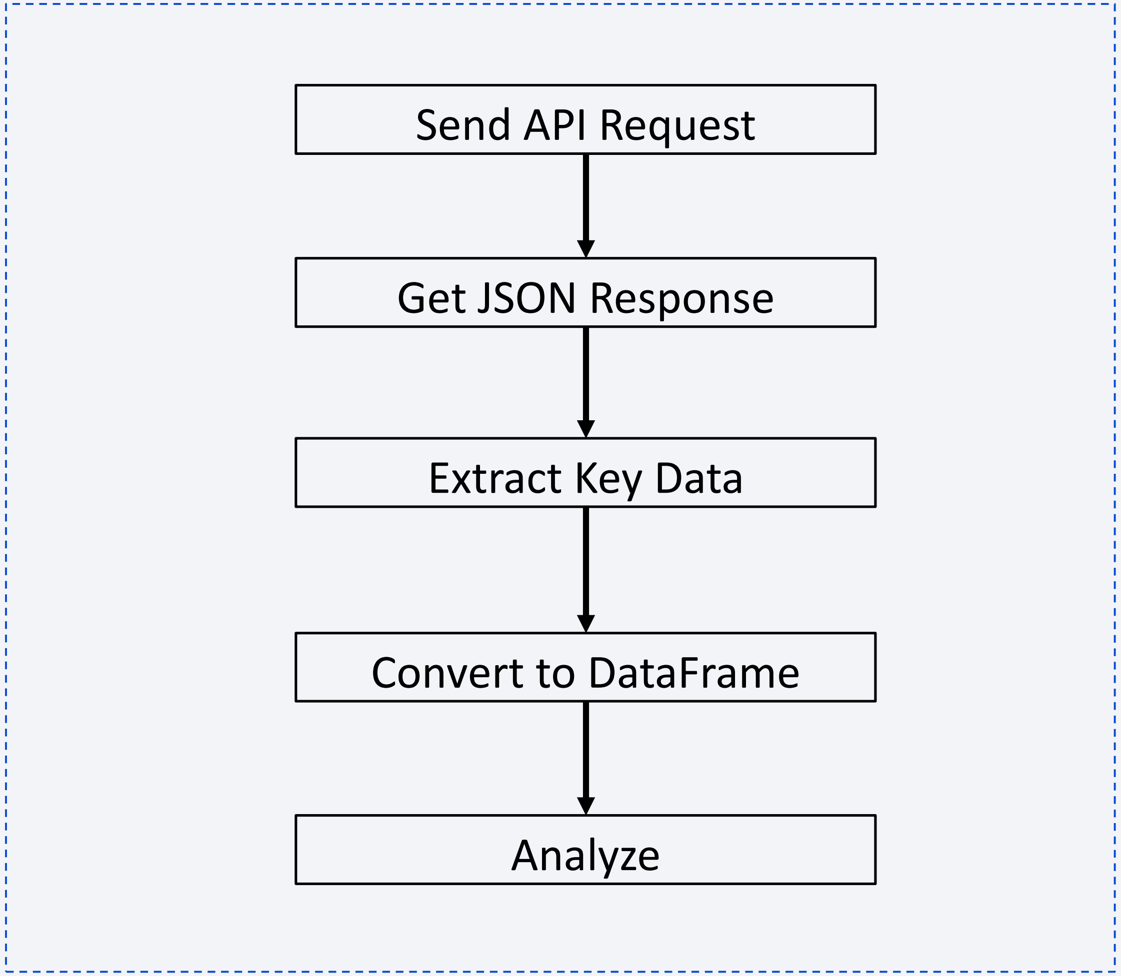
1. Using API to download launch information.
2. Web scraping to get additional data.



Data Collection – SpaceX API

Process Description

- SpaceX launch data was retrieved using REST API.
- The API provided information about launch date, payload, orbit, rocket version, first stage landing status, and other parameters.
- Used Python module requests to send GET request and receive JSON response.
- Data was processed and stored in Pandas DataFrame for further analysis.



External Reference (GitHub Link)

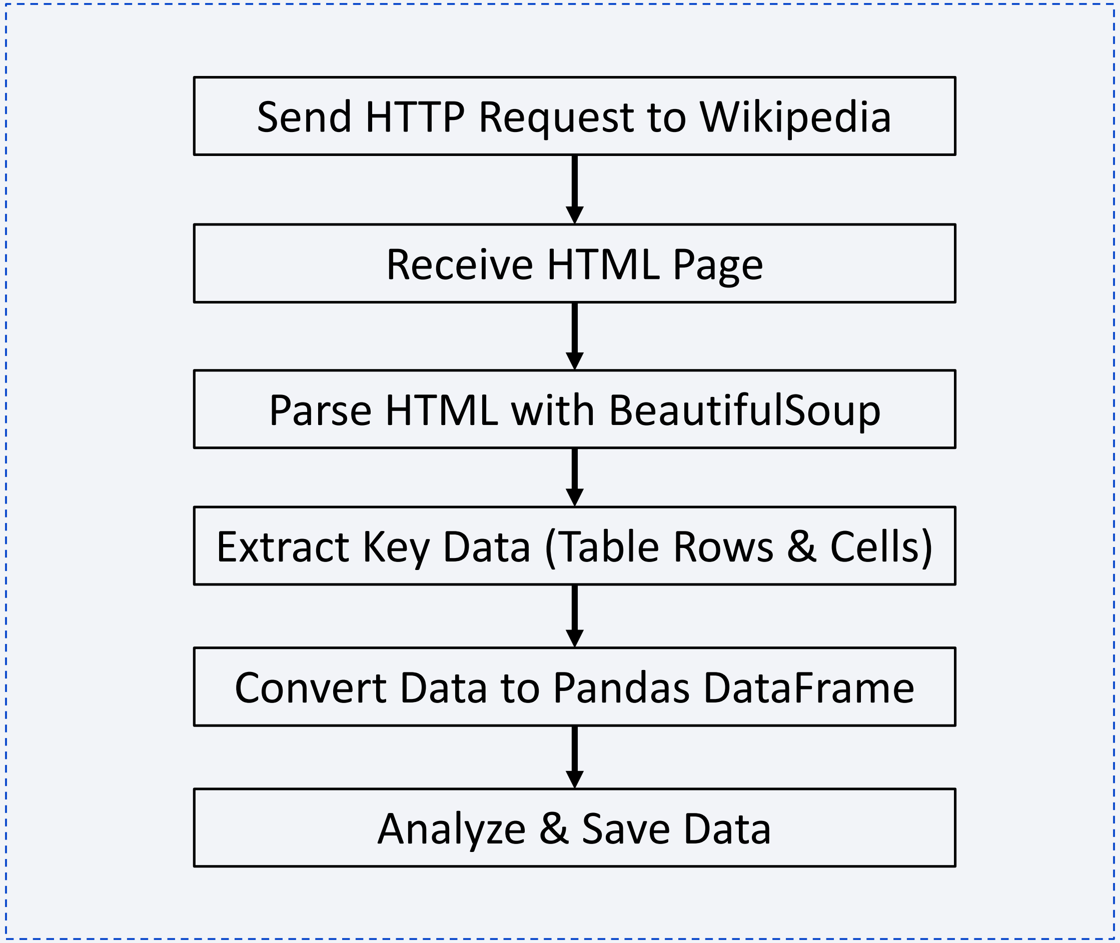
Jupyter Notebook (API Data Collection):

https://github.com/Garu-2000/Applied-Data-Science-Capstone/blob/main/1_SpaceX_Data-collection_API.ipynb

Data Collection - Scraping

Process Description

- Web Scraping method was used to collect data.
- Data source: Wikipedia (page with Falcon 9 launch history).
- Python module BeautifulSoup was used to parse the HTML code of the page.
- Key data was extracted: launch date, payload, mass, orbit, first stage landing status.
- Data was processed, saved in Pandas DataFrame and exported to CSV.



External Reference (GitHub Link)

Jupyter Notebook (Web Scraping Data Collection):

https://github.com/Garu-2000/Applied-Data-Science-Capstone/blob/main/2_SpaceX_Data-collection_Webscraping.ipynb

Data Wrangling

Process Description

1. Data cleaning: removing duplicates, missing values and correcting data formats.

2. Data merging: integrating information from API and Web Scraping into a single DataFrame.

3. Creating new features: highlighting key parameters (e.g. binary flag of successful landing).

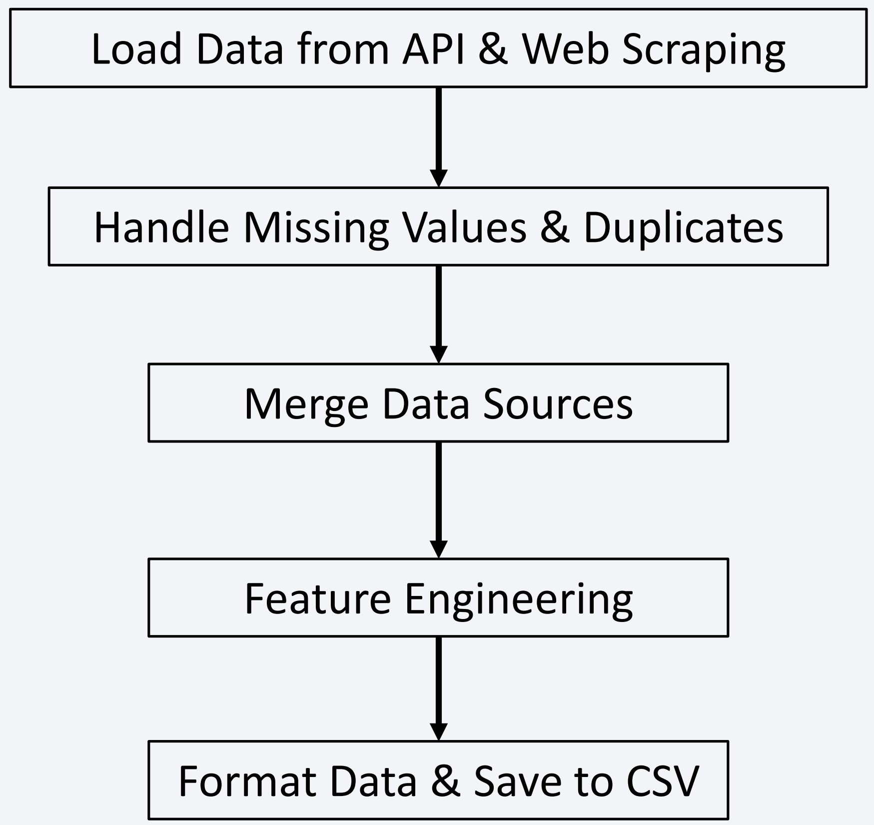
4. Data formatting: bringing data to a form convenient for analysis (data type, columns, indexing).

5. Data export: saving the cleaned set to CSV for further analysis.

External Reference (GitHub Link)

Jupyter Notebook (Web Scraping Data Collection):

https://github.com/Garu-2000/Applied-Data-Science-Capstone/blob/main/3_SpaceX_Data-wrangling.ipynb



EDA with Data Visualization

EDA revealed key patterns that could affect the probability of a successful landing of the Falcon 9 first stage:

- **Flight Number vs. Payload Mass** (Scatter Plot) – showed that the probability of a successful landing increases with an increase in the flight number.
- **Flight Number vs. Launch Site** (Scatter Plot) – compared landing successes at different launch sites.
- **Payload Mass vs. Launch Site** (Scatter Plot) – revealed which sites launched the heaviest payloads.
- **Orbit Type vs. Success Rate** (Bar Chart) – showed that LEO and ISS orbits have a higher landing success rate.
- **Flight Number vs. Orbit Type** (Scatter Plot) – demonstrated that success at LEO is related to launch experience, while there is no connection at GTO.
- **Payload Mass vs. Orbit Type** (Scatter Plot) – analyzed how payload mass affects the probability of landing.
- **Yearly Launch Success Trend** (Line Chart) – showed that launch success has been steadily increasing since 2013.

External Reference (GitHub Link)

Jupyter Notebook (EDA with Data Visualization):

https://github.com/Garu-2000/Applied-Data-Science-Capstone/blob/main/5_SpaceX_EDA_Data-vizualization.ipynb

EDA with SQL

SQL queries used in the analysis

1. List unique launch sites.
2. List top 5 launches from sites starting with "CCA".
3. Total payload mass launched by NASA (CRS).
4. Average payload mass for Falcon 9 v1.1.
5. Date of first successful landing on ground pad.
6. List of boosters that successfully landed on drone ship and had payload mass between 4000 and 6000 kg.
7. Total number of successful and failed missions.
8. Boosters that carried maximum payload.
9. Show drone ship landing failures in 2015.
10. Ranking of landing outcomes in the period from 2010-06-04 to 2017-03-20.

External Reference (GitHub Link)

Jupyter Notebook (EDA with SQL):

https://github.com/Garu-2000/Applied-Data-Science-Capstone/blob/main/4_SpaceX_EDA_SQL.ipynb

Build an Interactive Map with Folium

What map objects were added?

1. Launch Site Markers (Folium.Marker)

- Each launch site is marked on the map with **markers** and labeled with its name.
- **folium.Marker** and **folium.DivIcon** were used.

2. Circles Around Launch Sites (Folium.Circle)

- **Circles (folium.Circle)** with a 1000m radius were added to highlight areas around the launch sites.
- A popup label displays the name of the launch site.

3. Clustered Markers for Successful/Failed Launches (Folium.MarkerCluster)

- Launch markers are colored **green** (successful) and **red** (failed).
- **folium.plugins.MarkerCluster** was used.

4. Distance Measurements (Folium.PolyLine and Folium.Marker)

- Distances from launch sites to: **Coastline, Highway, Railway, Nearest City.**
- **folium.PolyLine** was used to draw distance lines, and **folium.Marker** displayed the calculated distances in km.

Why were these objects added?

- **Marking launch sites** helps visualize their geographic locations.
- **Color-coded launch outcome markers** allow quick success/failure analysis.
- **Measuring distances to nearby locations** helps identify geographical constraints and dependencies.

External Reference (GitHub Link)

Jupyter Notebook (Build an Interactive Map with Folium):

https://github.com/Garu-2000/Applied-Data-Science-Capstone/blob/main/6_SpaceX_Location_Folium.ipynb

Build a Dashboard with Plotly Dash

What plots and interactions were added?

1. Dropdown (Launch Site Selection)

- Allows selecting **a specific launch site** or viewing data for **all sites**.
- Affects the display of both **Pie Chart and Scatter Plot**.

2. Pie Chart (Total Successful Launches)

- Displays the **proportion of successful launches** for each launch site.
- If a specific site is selected, it shows the **success vs. failure ratio** for that site.
- Implemented using **Plotly Express Pie Chart**.

3. Range Slider (Payload Mass Filter)

- Enables **filtering launches** based on **payload mass range**.
- Affects the Scatter Plot visualization.

4. Scatter Chart (Payload vs. Success Correlation)

- Analyzes the **relationship between payload mass and launch success**.
- Allows filtering by **Launch Site and Payload Mass**.
- Implemented using **Plotly Express Scatter Plot**.

Why were these elements added?

- **Dropdown** – Helps analyze launch success by location.
- **Pie Chart** – Highlights which sites have higher success rates.
- **Range Slider** – Allows detailed analysis of payload impact.
- **Scatter Chart** – Helps identify trends between payload mass and success rate.

External Reference (GitHub Link)

Jupyter Notebook (Build a Dashboard with Plotly Dash):
https://github.com/Garu-2000/Applied-Data-Science-Capstone/blob/main/7_SpaceX_Dash_App.py

Predictive Analysis (Classification)

How was the model built and evaluated?

1. Data Preprocessing

- Loaded SpaceX launch data.
- Standardized features using **StandardScaler**.
- Split into **training (80%)** and **test (20%)** sets.

2. Training Multiple Classification Models

- **Logistic Regression**
- **Support Vector Machine (SVM)**
- **Decision Tree**
- **K-Nearest Neighbors (KNN)**

3. Hyperparameter Tuning

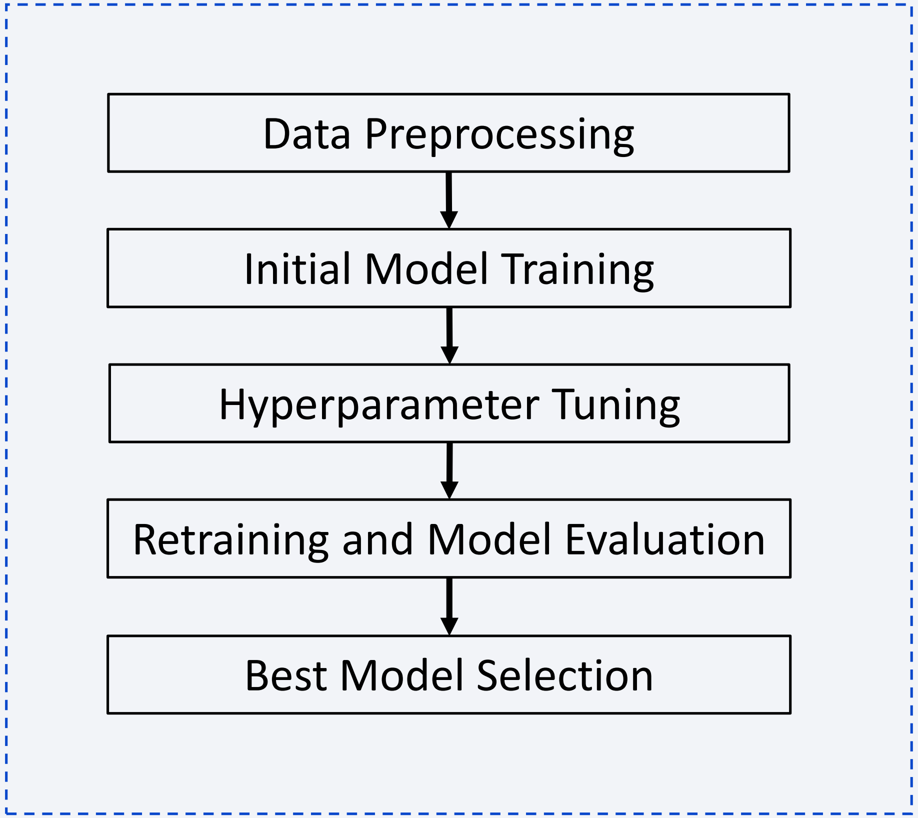
- Used **GridSearchCV (10-fold cross-validation)**.
- Optimized parameters like **C, kernel, max_depth, n_neighbors**.

4. Model Evaluation

- **First evaluation**: Models trained with default parameters.
- **Second evaluation**: Models retrained after tuning hyperparameters.
- Used **Accuracy Score** and **Confusion Matrix**.

5. Best Model Selection

- **Decision Tree (94.4% accuracy)** outperformed all models.
- Other models (Logistic Regression, SVM, KNN) had **83.3% accuracy**.
- Decision Tree was chosen as the best predictor for Falcon 9 landings.



External Reference (GitHub Link)

Jupyter Notebook (Predictive Analysis: Classification):

https://github.com/Garu-2000/Applied-Data-Science-Capstone/blob/main/8_SpaceX_Machine-Learning-Prediction.ipynb

Results

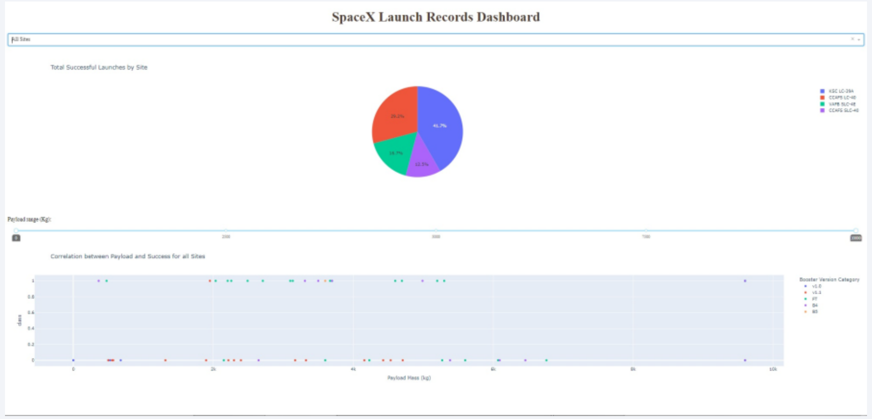
Exploratory Data Analysis (EDA) Results

1. **Flight Number vs. Success Rate**
    - The **higher the flight number**, the **more likely** the first stage will land successfully.
    - Payload mass also plays a role: **even with heavier payloads, successful landings are frequent**.
2. **Launch Site Insights**
    - **VAFB-SLC launch site** has **no heavy payload launches** (>10,000 kg).
3. **Orbit and Success Rate**
    - **Highest success rates: ES-L1, GEO, HEO, SSO**.
    - **Lower success rates: VLEO**.
    - **In LEO**, success correlates with **flight number**.
    - **In GTO**, no clear relationship between **flight number and success**.
4. **Payload Mass Impact**
    - **Higher payloads → better success rates** for **Polar, LEO, ISS orbits**.
    - **In GTO, both successes and failures occur equally**, making it harder to predict.
5. **Yearly Success Rate Trend**
    - **Since 2013, the success rate of Falcon 9 landings steadily increased until 2020**.

Predictive Analysis Results

1. **Best performing model: Decision Tree** (**94.4% accuracy**).
2. Other models (Logistic Regression, SVM, KNN) had **83.3% accuracy**.
3. **Confusion Matrix Analysis:**
    - Decision Tree minimized **false positives** and **false negatives**.
    - Logistic Regression and SVM had slightly more misclassifications.

Interactive analytics demo in screenshots



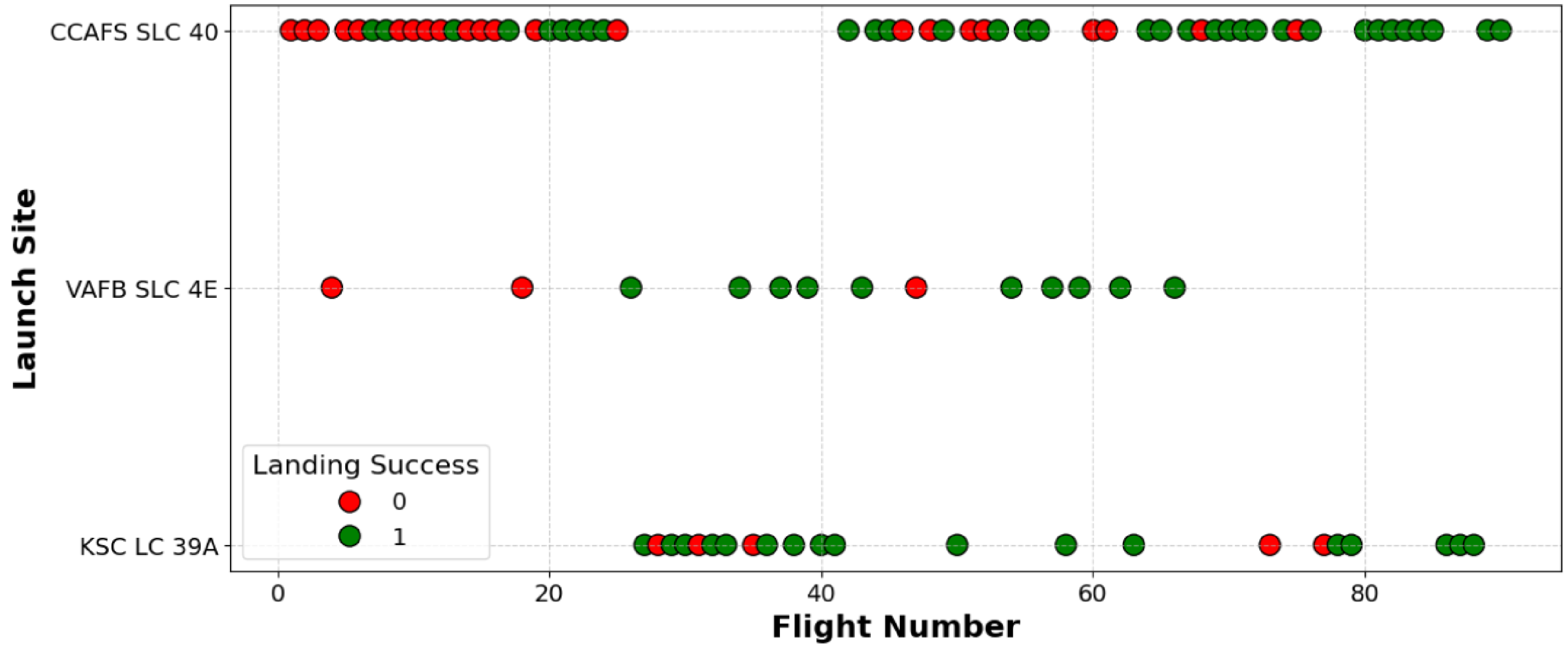
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Interpretation:

- Each **point represents a launch**.
- **Color indicates the success of the first stage landing** (success = 1, failure = 0).
- **VAFB SLC-4E does not handle heavy payloads**, which might explain its differences compared to other sites.



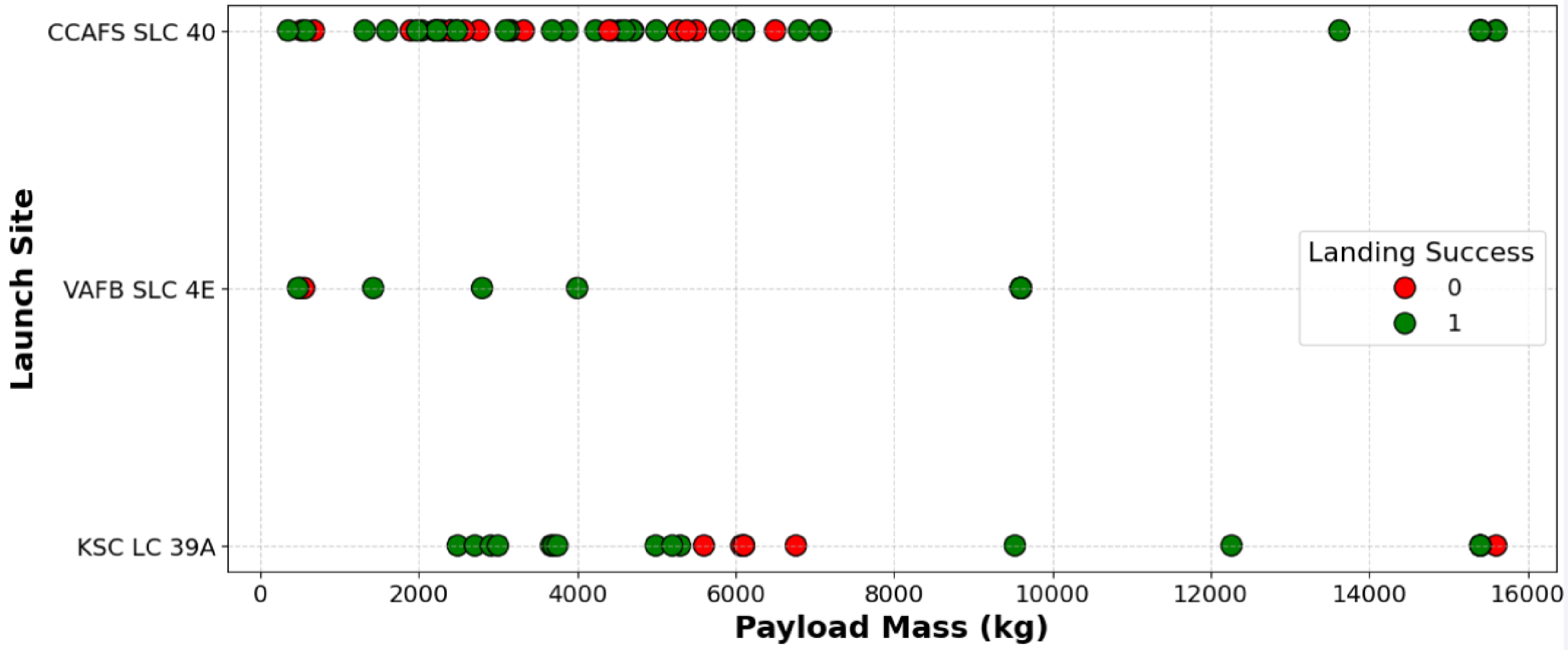
Key takeaways:

- Higher flight numbers correlate with more successful landings.
- CCAFS SLC-40 has the highest number of launches.
- VAFB SLC-4E has fewer successful landings than other sites.

Payload vs. Launch Site

Interpretation:

- Each **point represents a launch**.
- **Color indicates the success of the first stage landing** (success = 1, failure = 0).
- **VAFB SLC-4E does not handle heavy payloads**, which might explain its differences compared to other sites.



Key observations:

1. Launches from VAFB SLC-4E are limited to light and medium payloads (up to 10,000 kg).

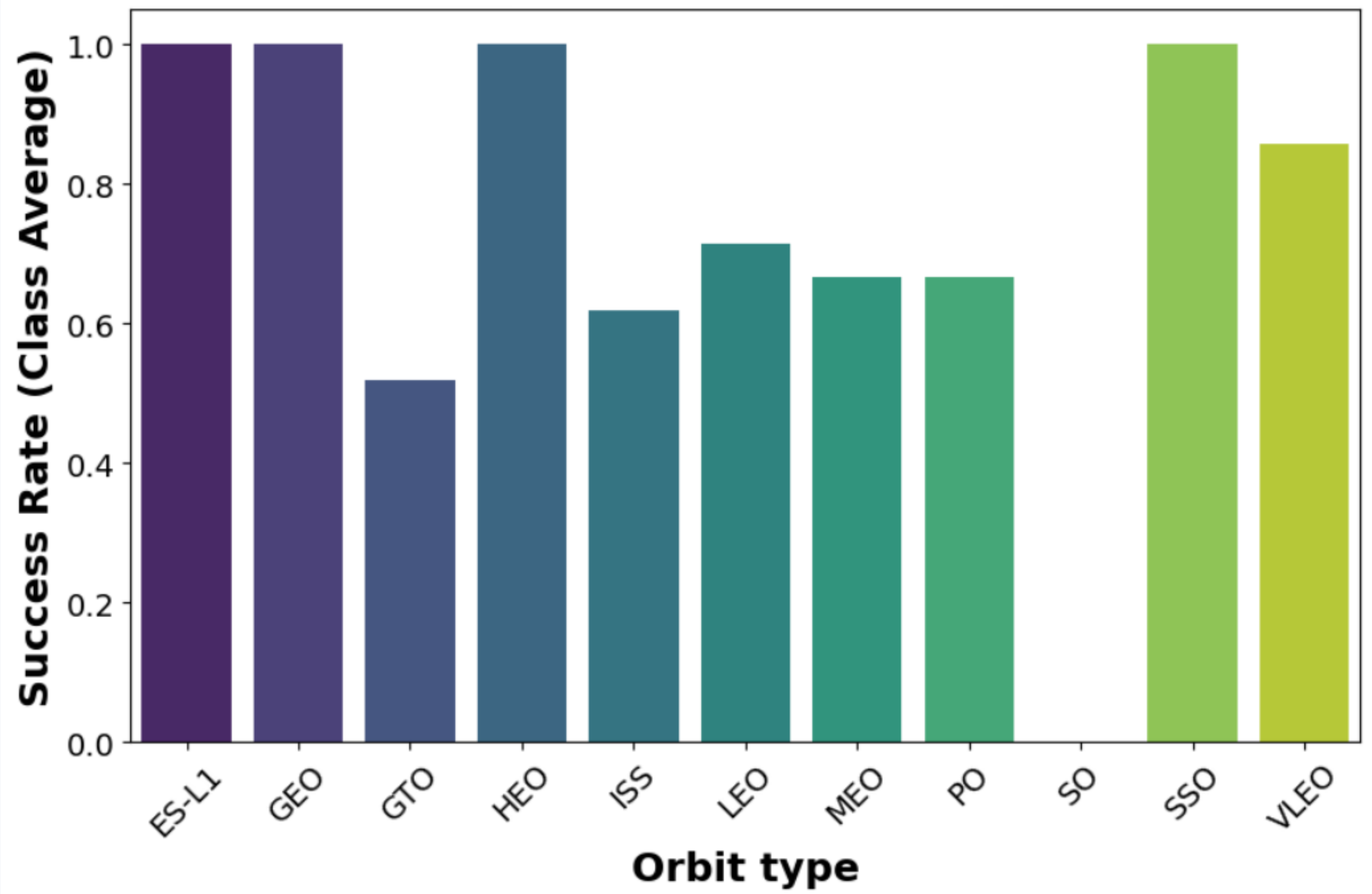
2. CCAFS LC-40 and KSC LC-39A support a wider range of payload masses, including heavy payloads (>10,000 kg).

3. Some launches from KSC LC-39A reach payload masses of ~15,000 kg, making it the most utilized site for heavy payloads.

4. There is no clear correlation between payload mass and landing success (Class), but on some sites, heavier payloads tend to land successfully more often.

5. CCAFS SLC-40 exhibits the broadest range of launch payloads.

Success Rate vs. Orbit Type



Key observations:

1. The ES-L1, GEO, HEO, and SSO orbits have the highest probability of successful landings.

2. The VLEO orbit shows a slightly lower success rate.

3. The GTO orbit has an almost equal distribution of successful and unsuccessful landings.

4. The ISS and LEO orbits have a moderate success rate.

5. The Polar orbit also demonstrates a high probability of successful landings.

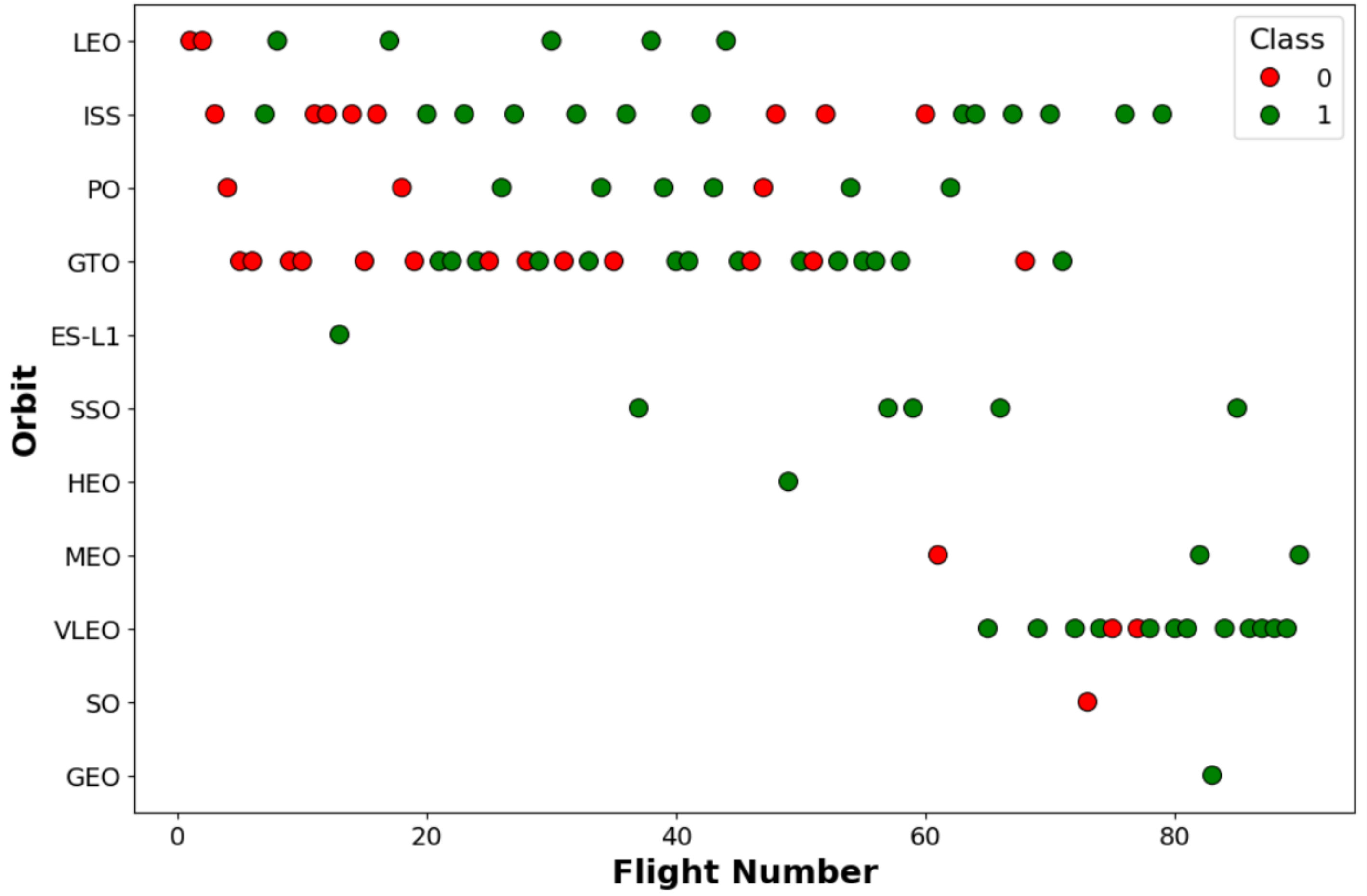
Flight Number vs. Orbit Type

Key observations:

1. For LEO (Low Earth Orbit), the probability of success increases as the number of flights increases.

2. For GTO (Geostationary Transfer Orbit), the number of flights does not significantly affect landing success—both successes and failures appear randomly distributed.

3. Launches to ES-L1, GEO, HEO, and SSO are less frequent but have a high success rate.



4. Most launches occur to LEO and GTO, confirming their popularity for missions.

5. Flights to Polar orbit tend to have a high success rate, especially in later launches.

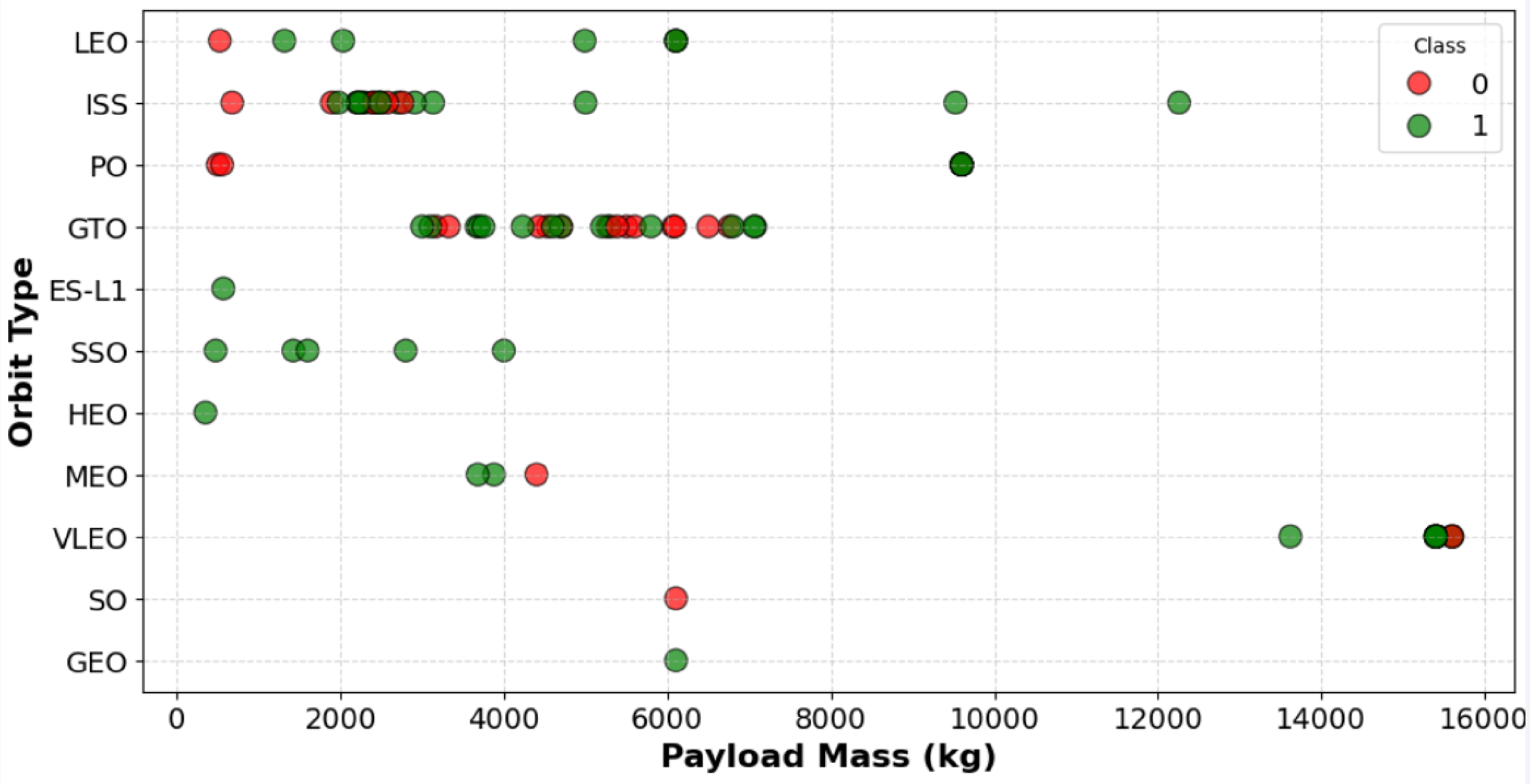
Payload vs. Orbit Type

Key observations:

1. For LEO and ISS orbits, successful landings occur with both light and heavy payloads.

2. On GTO (Geostationary Transfer Orbit), landing outcomes are highly variable, with both successful and failed landings appearing across different payload masses.

3. Polar orbits demonstrate a high probability of successful landings, especially for heavy payloads.



4. For SSO, ES-L1, and GEO, payload mass rarely exceeds 6000 kg.

5. HEO and VLEO have fewer launches, but successful landings are more common with lower payload masses.

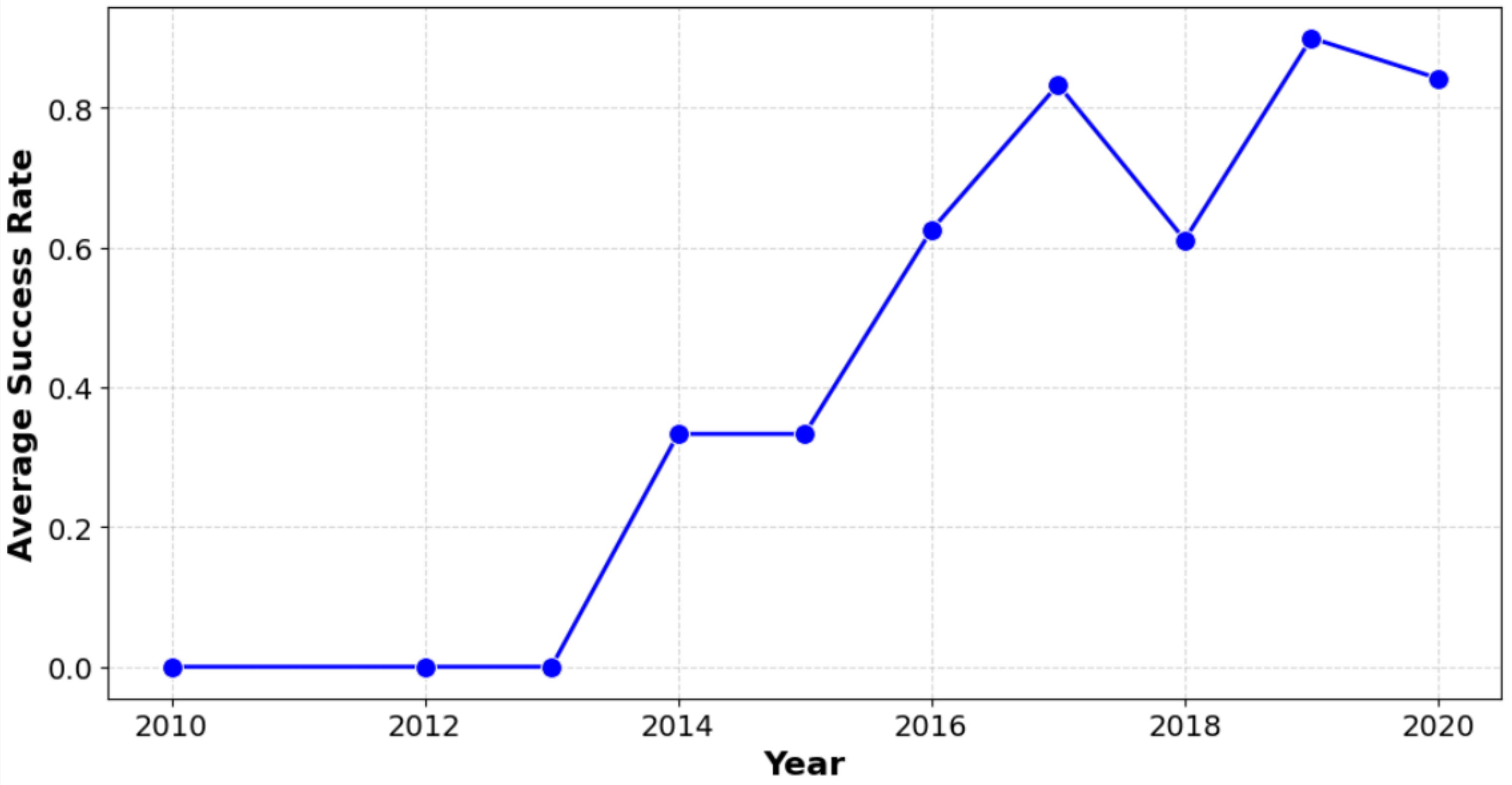
Launch Success Yearly Trend

Key Observations:

1. Before 2013, successful landings were extremely rare or nonexistent.

2. Since 2013, there has been a steady upward trend in first-stage landing success.

3. Starting in 2016, the success rate of landings significantly increased, likely due to improvements in reusability technology.



4. The drop in success rate in 2018 was related to Falcon Heavy test flights, experiments with new technologies, and several failed landings.

5. After 2018, the success rate stabilized, approaching 80–90%.

6. The increase in success is largely attributed to the implementation of landing systems on both floating platforms and ground pads.

All Launch Site Names

SpaceX uses four main launch sites for Falcon 9 missions:

1. CCAFS LC-40 (Cape Canaveral Air Force Station, Launch Complex 40) – one of the most frequently used launch complexes.

2. VAFB SLC-4E (Vandenberg Air Force Base, Space Launch Complex 4E) – used for polar orbit launches.

3. KSC LC-39A (Kennedy Space Center, Launch Complex 39A) – a historic NASA launch site, used for Apollo missions.

4. CCAFS SLC-40 (Cape Canaveral Space Launch Complex 40) – a site for both commercial and government launches.

```
Unique launch sites:
+--------------+
| Launch_Site  |
+--------------+
| CCAFS LC-40  |
| VAFB SLC-4E  |
|  KSC LC-39A  |
| CCAFS SLC-40 |
+--------------+
```

These launch sites are located in Florida and California, allowing SpaceX to conduct launches into different orbital planes.

Launch Site Names Begin with 'CCA'

Key Observations:

Launches conducted at sites beginning with "CCA" (Cape Canaveral Air Force Station):

1. All five launches were conducted from the CCAFS LC-40 launch site.

2. Most missions were part of NASA's COTS and CRS programs, aiming for the ISS.

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

3. The first two launches had unsuccessful landings due to parachute system failures.

4. Later, SpaceX abandoned parachute landing in favor of first-stage landings on platforms.

5. Launches from CCAFS LC-40 continued to improve, leading to an increase in successful landings.

Total Payload Mass

Key Observations:

Total payload mass carried by NASA (CRS) missions:

1. NASA (CRS) missions carried a total of 45,596 kg of payload using SpaceX boosters.

2. These missions were primarily aimed at supplying the International Space Station (ISS).

3. The payloads included cargo, scientific equipment, and supplies for astronauts.

4. Over time, SpaceX improved its payload capacity, allowing for heavier cargo deliveries.

5. NASA's reliance on SpaceX for ISS resupply missions highlights the success of their partnership.

	Total_Payload_Mass
0	45596

Average Payload Mass by F9 v1.1

Key Observations:

Average payload mass carried by Falcon 9 v1.1:

1. The average payload mass carried by the Falcon 9 v1.1 booster was approximately 2,928.4 kg.

2. This version of the Falcon 9 was used from 2013 to 2016 and featured significant improvements over its predecessor, including increased thrust and a redesigned stage separation system.

3. Compared to later Falcon 9 versions, v1.1 had a lower payload capacity, but it was still a major step forward in SpaceX's reusability efforts.

4. Falcon 9 v1.1 supported various missions, including commercial satellite launches and NASA's CRS (Commercial Resupply Services) missions to the ISS.

5. The transition to Falcon 9 Full Thrust (FT) and later versions enabled greater payload capacity and further advancements in first-stage recovery.

	Average_Payload_Mass
0	2928.4

First Successful Ground Landing Date

Key Observations:

First successful ground landing of a Falcon 9 first stage:

1. The first recorded successful ground landing of a Falcon 9 booster occurred on July 22, 2018.

2. This marked a major milestone for SpaceX, proving the feasibility of reusability for orbital-class rockets.

3. Before this, most successful landings were performed on autonomous drone ships in the ocean.

4. Ground landings provide a more cost-effective recovery option, as they reduce the risk of booster damage and simplify logistics.

5. Since this success, SpaceX has continued refining its landing techniques, leading to a high success rate in booster recovery and reuse.

	First_Successful_Landing_Date
0	2018-07-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Key Observations:

Falcon 9 boosters successfully landing on drone ships with payloads between 4000 and 6000 kg:

1. Five Falcon 9 boosters successfully landed on drone ships while carrying payloads in this weight range.

2. The "FT" (Full Thrust) version of Falcon 9 is designed for improved reusability and landing success.

3. Drone ship landings are often used when there is insufficient fuel for a return-to-launch-site (RTLS) landing.

4. This data confirms that SpaceX has achieved multiple successful recoveries of boosters in this payload category.

5. These landings play a crucial role in reducing costs and enabling rapid booster reuse for future missions.

	Booster_Version
0	F9 FT B1020
1	F9 FT B1022
2	F9 FT B1026
3	F9 FT B1021.2
4	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

	Mission_Outcome	Count
0	Failure (in flight)	1
1	Success	98
2	Success (payload status unclear)	1

Key Observations:

Mission Success Rate:

1. The vast majority of missions (98 out of 100) were classified as successful.

2. There was one mission failure during flight, which means the rocket was lost before reaching orbit.

3. One mission had an unclear payload status, indicating uncertainty about whether the payload reached its intended operational state.

Failure Analysis:

- The failure in flight suggests that SpaceX has achieved an extremely high mission success rate over time.
- SpaceX has continuously improved launch reliability, reducing the number of failures over the years.

Conclusion:

The high success rate highlights the efficiency, reliability, and innovation of SpaceX's Falcon 9 program. The company has significantly minimized mission failures, contributing to its industry leadership in commercial spaceflight.

Boosters Carried Maximum Payload

Key Observations:

Maximum Payload Mass:

1. The Falcon 9 Block 5 boosters **B1048, B1049, B1051, B1056, B1058, and B1060** successfully carried the **heaviest payloads** in SpaceX's launch history.

2. These boosters belong to the **Block 5 version** of Falcon 9, which is known for its improved performance, higher thrust, and reusability.

Technical Insights:

- The **Block 5** variant is optimized for **multiple reuses**, making it more cost-efficient for launching heavier payloads.
- These boosters were used in missions involving **high-energy orbits**, including **geostationary and interplanetary missions**.

Conclusion:

The Falcon 9 Block 5 series has demonstrated **exceptional capability** in launching heavy payloads while maintaining reusability. These boosters are crucial for **commercial, governmental, and interplanetary missions**, solidifying SpaceX's dominance in the space industry.

	Booster_Version
0	F9 B5 B1048.4
1	F9 B5 B1049.4
2	F9 B5 B1051.3
3	F9 B5 B1056.4
4	F9 B5 B1048.5
5	F9 B5 B1051.4
6	F9 B5 B1049.5
7	F9 B5 B1060.2
8	F9 B5 B1058.3
9	F9 B5 B1051.6
10	F9 B5 B1060.3
11	F9 B5 B1049.7

2015 Launch Records

	Month	Landing_Outcome	Booster_Version	Launch_Site
0	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
1	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Key Observations:

Failed Landings on Drone Ship in 2015:

1. Two **failed landing attempts** occurred in **January and April 2015**.
2. Both failures happened on **Falcon 9 v1.1** boosters **B1012 and B1015**.
3. The **launch site** for both missions was **CCAFS LC-40** in Florida.

Technical Insights:

- These missions were **early attempts** at landing boosters on drone ships, and SpaceX was still refining its landing technology.
- The Falcon 9 v1.1 version **did not have key improvements** like **stronger landing legs, grid fins, and refined thrust control**, which were later introduced in Falcon 9 Block 5.

Conclusion:

The **2015 failures** were a **critical learning phase** for SpaceX's reusability program. By 2016, **significant improvements** in landing techniques led to **higher success rates**, laying the foundation for the **cost-efficient reusable rocket technology** we see today.

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

Key Observations:

Most Common Landing Outcomes (2010-2017):

1. **"No attempt" (10 launches)** – Many missions in the early years did not include landing attempts.

2. **Success and failure on drone ships (5 each)** – Demonstrates SpaceX's trial-and-error process in offshore landings.

3. **Ground pad landings (3 successful attempts)** – This number increased in later years as landing technology improved.

4. **Controlled ocean landings (3 attempts)** – Indicates some boosters were deliberately ditched for safety reasons.

	Landing_Outcome	Count
0	No attempt	10
1	Success (drone ship)	5
2	Failure (drone ship)	5
3	Success (ground pad)	3
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Failure (parachute)	2
7	Precluded (drone ship)	1

Insights on Failures:

- **Parachute failures (2 times)** – These were early-stage recovery attempts that did not work as planned.
- **Uncontrolled ocean landings (2 times)** – Unplanned water crashes where the booster was lost.
- **Precluded landing (1 case)** – A mission where a landing attempt was originally planned but later canceled due to mission constraints.

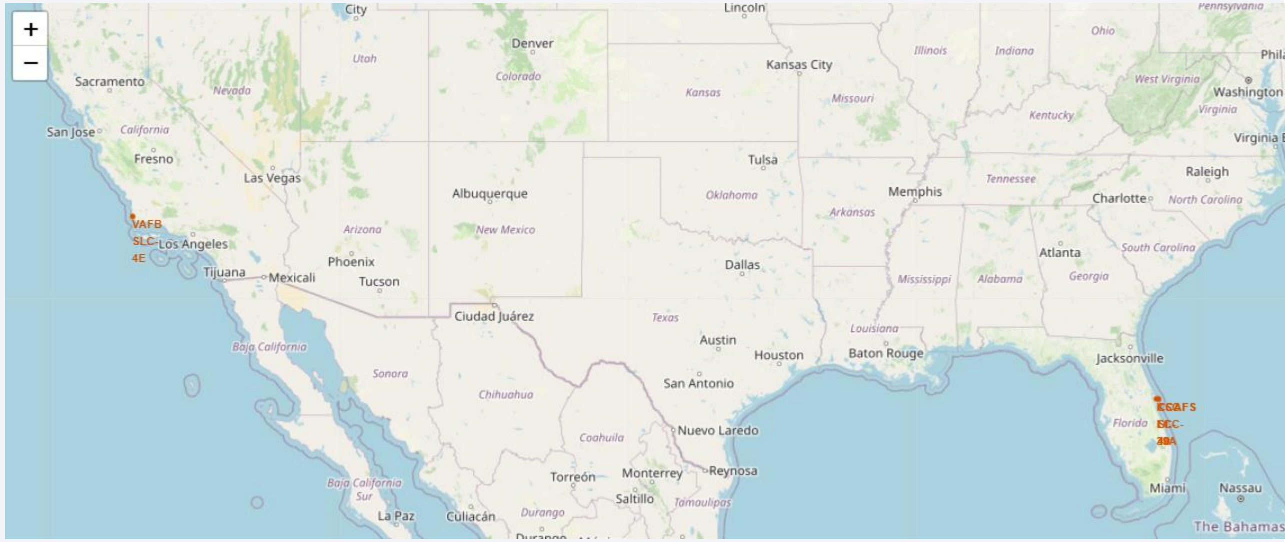
Conclusion:

Between 2010 and 2017, **SpaceX gradually improved its landing success rate**. While **early years saw failures**, **by 2017 drone ship and ground landings became more reliable**, **paving the way for fully reusable rockets**.

Section 3

Launch Sites Proximities Analysis

Global View of SpaceX Launch Sites



Key Insights from the Map:

Proximity to Water Bodies:

All SpaceX launch sites are located **near coastlines**. This minimizes **risk to populated areas** and allows for **safe rocket landings and recovery in the ocean**.

Strategic Geographic Placement:

- **Florida-based sites** (CCAFS & KSC) are ideal for **low-inclination orbits (LEO, GTO, ISS)**, leveraging Earth's rotation for **energy-efficient launches**.
- **Vandenberg SLC-4E (California)** is used for **polar orbits**, launching satellites into sun-synchronous and polar trajectories.

Reusability & Landing Infrastructure:

Each site is **equipped with recovery zones**, including:

- **Ground landing pads** for Falcon 9 & Falcon Heavy.
- **Drone ships positioned offshore** for booster recovery.

Conclusion:

SpaceX **strategically placed its launch sites** to support different mission requirements, **maximizing efficiency and reusability in rocket launches**.

Analysis of SpaceX Successful and Failed Landings

Launch Site Markers:

1. CCAFS LC-40 (Florida)
2. CCAFS SLC-40 (Florida)
3. KSC LC-39A (Florida)
4. VAFB SLC-4E (California)

Color-Coded Landing Outcomes:

- Green markers → Successful first-stage landings.
- Red markers → Failed landings or no landing attempts.

<u>Key Findings from the Map:</u>

- Successful landings are more frequent in Florida:
- Most green markers (successful landings) are concentrated around CCAFS SLC-40 and KSC LC-39A.
- This is due to the development of reusable first-stage technology, which began around 2015.

Failed landings are more common in California and at sea:

- The VAFB SLC-4E site in California has more red markers, likely due to polar orbit launches, which make first-stage recovery more difficult.
- Many failed landings occurred in the ocean, confirming that SpaceX deliberately sacrificed first stages for challenging missions.

Proximity Analysis of SpaceX Launch Sites

Screenshot of the Interactive Folium Map:

The Folium map with marked infrastructure objects shows the distances from a selected SpaceX launch site to: **Railway, Highway, Coastline, Nearest city.**

1. **Object Marking:**

- **Circles and lines on the map** indicate the location of infrastructure objects and connect them to the launch site.
- **Distance labels** in kilometers help evaluate proximity.

2. **Color Coding:**

- **Blue** → Lines between the launch site and infrastructure objects.
- **Orange** → Markers with distance labels to the objects.

<u>**Key Observations:**</u>

- **Railways are at a moderate distance from the launch sites.**
- SpaceX sites are chosen to be away from major railways to **minimize the risk of infrastructure damage** in case of a failed launch.
- **Highways provide convenient access to launch sites.**
- The proximity of highways to launch complexes **simplifies the transportation of rocket components** and equipment.
- **Coastal proximity facilitates launch operations.**
- Most launch sites are **located near the coast**, allowing rockets to **be directed safely over the ocean**, reducing the risk of falling debris over populated areas.
- **Launch sites are far from densely populated areas.**
- SpaceX **strategically locates launch complexes away from cities** to minimize noise and vibration impacts on residents.

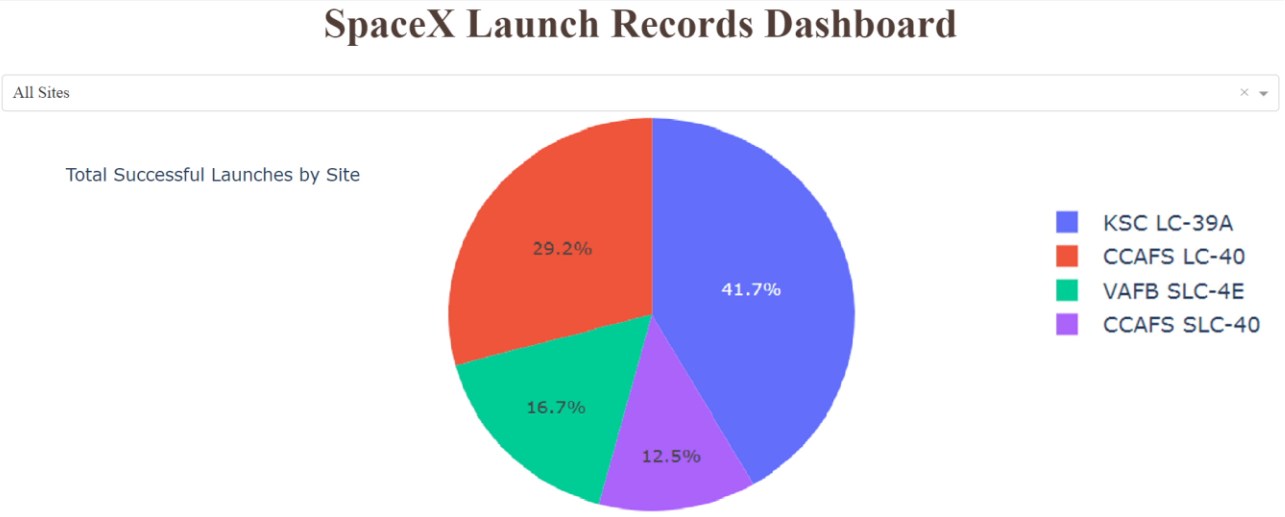
Section 4

Build a Dashboard with Plotly Dash

Dashboard: Total Successful Launches by Site

Content:

- This pie chart represents the total number of successful launches across all SpaceX launch sites.
- Users can select a specific launch site from the dropdown menu to see the success and failure distribution for that site.

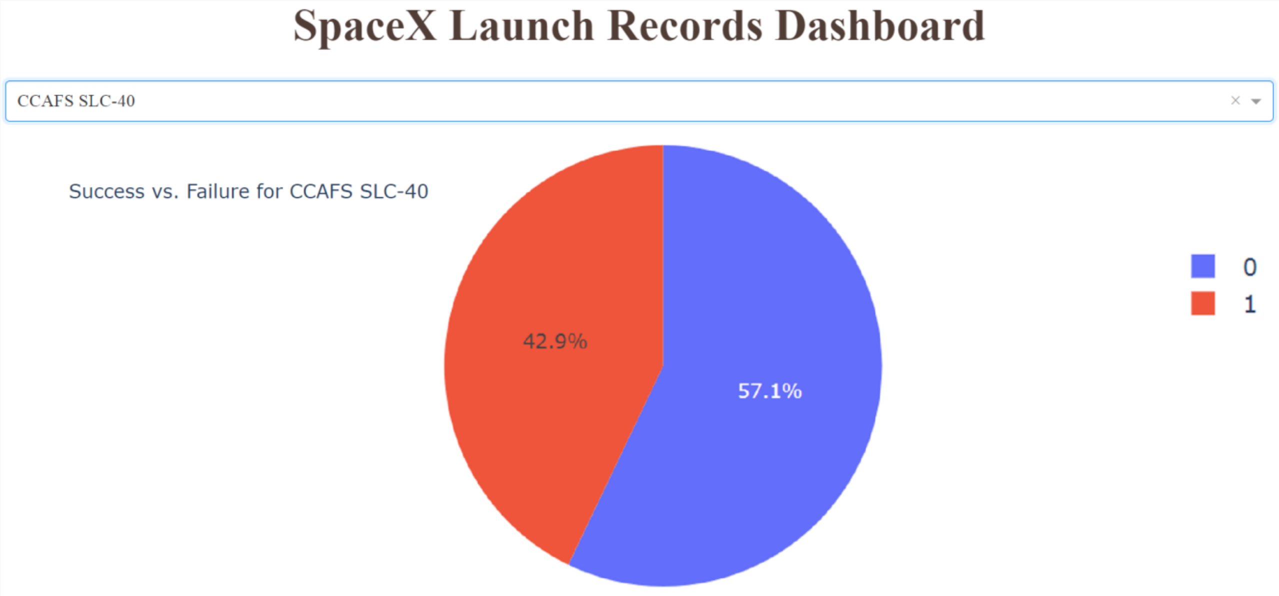


- If "All Sites" is selected, the chart shows a comparison of successful launches between all available launch sites.

Findings:

- The launch site with the highest number of successful missions can be identified.
- Users can compare the relative success rates of different sites.

Dashboard: Highest Launch Success Ratio – CCAFS SLC-40



Content:

- The pie chart displays the **highest launch success ratio**, which belongs to the **CCAFS SLC-40** launch site.
- Users can select this specific launch site from the dropdown menu in the dashboard to visualize its success ratio.
- The chart represents the percentage of successful launches compared to failed ones.

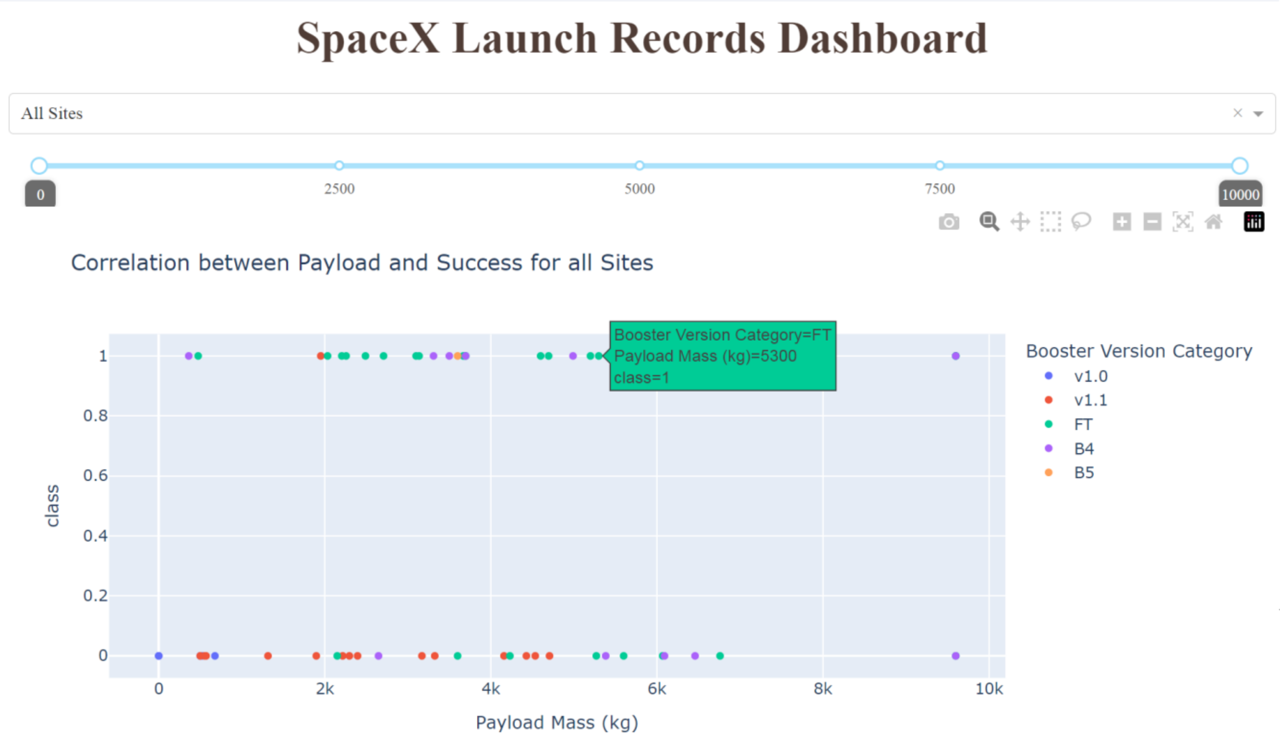
Findings:

- The **CCAFS SLC-40** launch site has the **highest success rate of 42.9%** among all SpaceX launch sites.
- Nearly half of the launches from this site resulted in successful landings.
- This data highlights **CCAFS SLC-40 as the most reliable launch site**, making it a strategic choice for future missions.

Dashboard: Payload vs. Launch Outcome

Content:

- This dashboard screenshot presents the **scatter plot of Payload Mass vs. Launch Outcome** for all launch sites.
- The payload range can be adjusted using the **range slider**, allowing users to observe variations in success rates across different payload capacities.



Findings:

- **Lighter payloads (below 5000 kg)** tend to have a **higher success rate**, indicating that lower payloads may be easier to manage for successful landings.
- **Heavy payloads (above 8000 kg)** show a **lower success rate**, suggesting that landing challenges increase with payload mass.
- Some **booster versions demonstrate consistently high success rates**, making them more reliable for future missions.
- The **scatter distribution varies across sites**, emphasizing that launch location, payload, and booster type all impact mission success.

Section 5

Predictive Analysis (Classification)

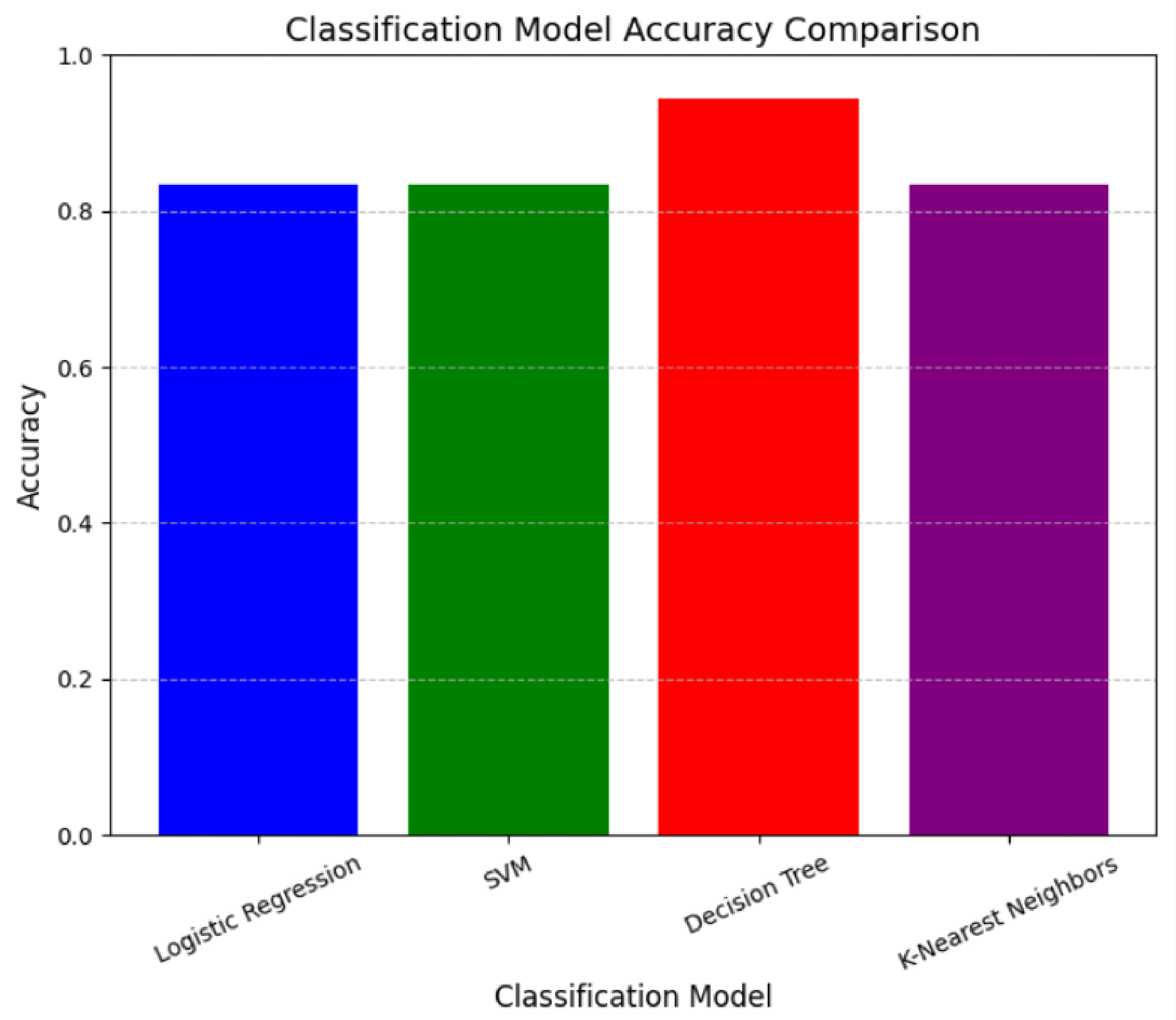
Classification Accuracy

Visualization:

- The bar chart illustrates the **accuracy comparison** among four classification models.
- The **y-axis ranges from 0 to 1**, making it easier to interpret model performance relative to a perfect prediction score.
- The Decision Tree model stands out with a noticeably higher accuracy than the other three models.

Key Findings:

- The **Decision Tree model** demonstrated the **highest classification accuracy** at **94.44%**, outperforming other models.
- **Logistic Regression, SVM, and K-Nearest Neighbors** models achieved identical accuracy levels of **83.33%**.
- The significant performance boost of the **Decision Tree** is likely due to its ability to handle complex decision boundaries and interactions between features.



Confusion Matrix

Confusion Matrix Overview:

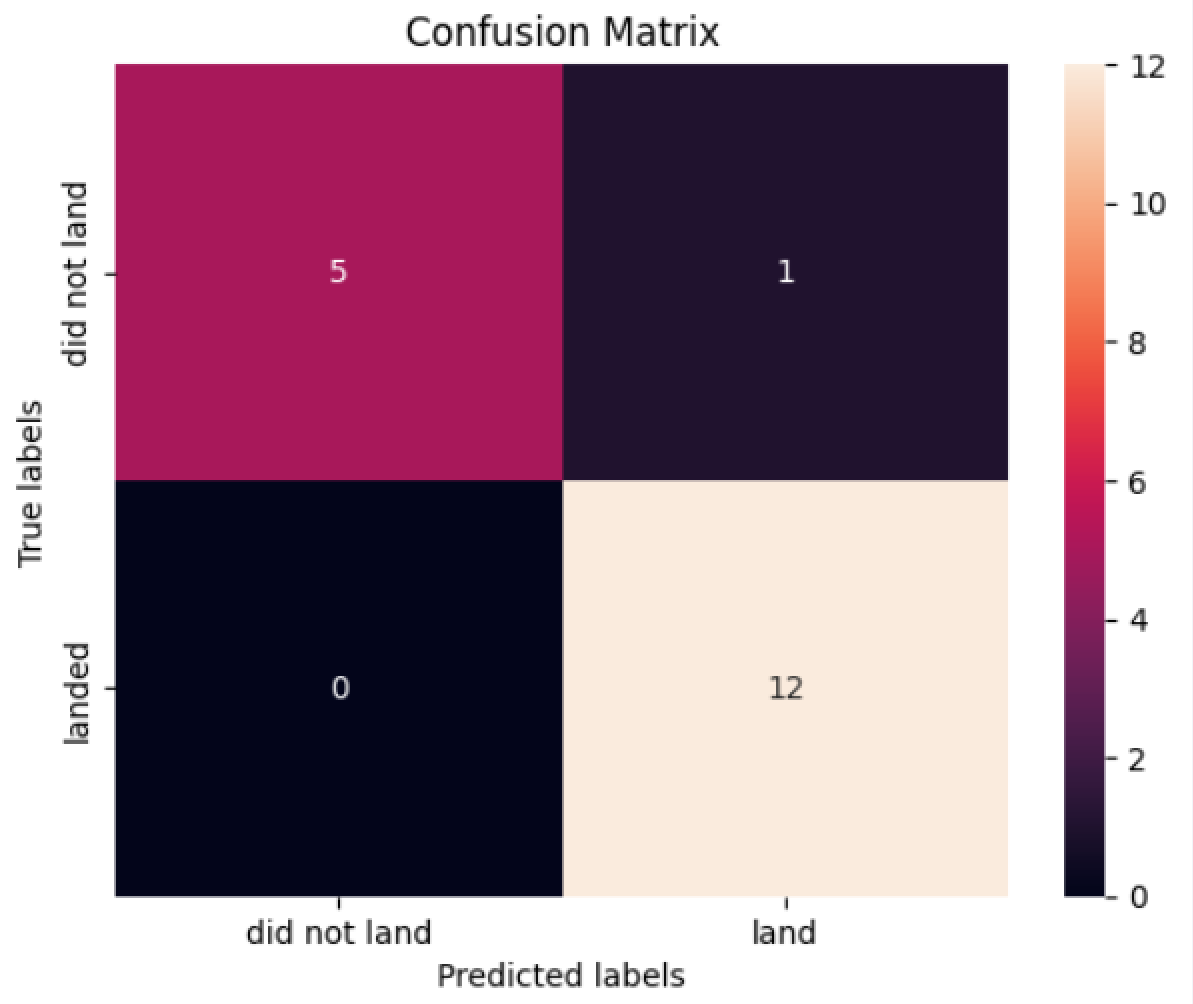
- The confusion matrix for the **Decision Tree model**, which achieved the highest accuracy (**94.44%**), provides insights into how well the model distinguishes between successful and unsuccessful Falcon 9 first stage landings.

<u>Key Observations:</u>

- **Perfect True Positive Rate (TPR): All** successful landings were correctly classified (**12/12**).
- **Zero False Negatives (FN):** The model did not misclassify any successful landing as a failure.
- **One False Positive (FP):** One failed landing was incorrectly predicted as successful.
- **Overall Accuracy:** The model correctly classified **17 out of 18 cases**, demonstrating **high reliability**.

Why is this Important?

- **High precision and recall** make this model ideal for predicting Falcon 9 first stage landings.
- The **low error rate** ensures better decision-making for mission planning and cost optimization.



Conclusions

1. **Exploratory Data Analysis (EDA) Insights:**

- Successful landings increased over time, especially after 2013.
- Heavy payloads tend to have higher success rates for certain orbits (LEO, ISS, Polar).
- The launch site **CCAFS SLC-40** had the highest launch success ratio.

2. **Interactive Analytics & Visualization Findings:**

- Folium maps showed launch site locations and proximity to infrastructure (highways, coastlines, railways).
- Success rates varied significantly by **orbit type**, with **LEO and ISS having the most consistent success**.
- Payload vs. Launch Outcome scatter plots revealed a strong correlation between payload mass and landing success.

3. **Predictive Model Performance:**

- **Decision Tree** achieved the highest accuracy (**94.44%**), outperforming Logistic Regression, SVM, and KNN.
- The **confusion matrix** showed that the Decision Tree model correctly classified **17 out of 18 cases**, with only **one false positive**.
- The model demonstrated **high precision and recall**, making it a reliable tool for predicting Falcon 9 first stage landings.

4. **Business & Technical Implications:**

- Predicting successful landings helps **optimize costs** by maximizing first-stage reuse.
- Identifying key success factors (orbit type, payload mass, launch site) **enhances mission planning**.
- Further improvements in the model could help **reduce false positives** and increase predictive accuracy.

5. **Future improvements:**

- Feature selection, ensemble methods, boosting techniques.

<u>**Final Thought:**</u> The combination of **data analysis, interactive visualization, and machine learning** provides **valuable insights** for optimizing **SpaceX Falcon 9 launches** and ensuring more **cost-efficient and reliable missions**.

Appendix 1: Falcon 9 (WikipediA)

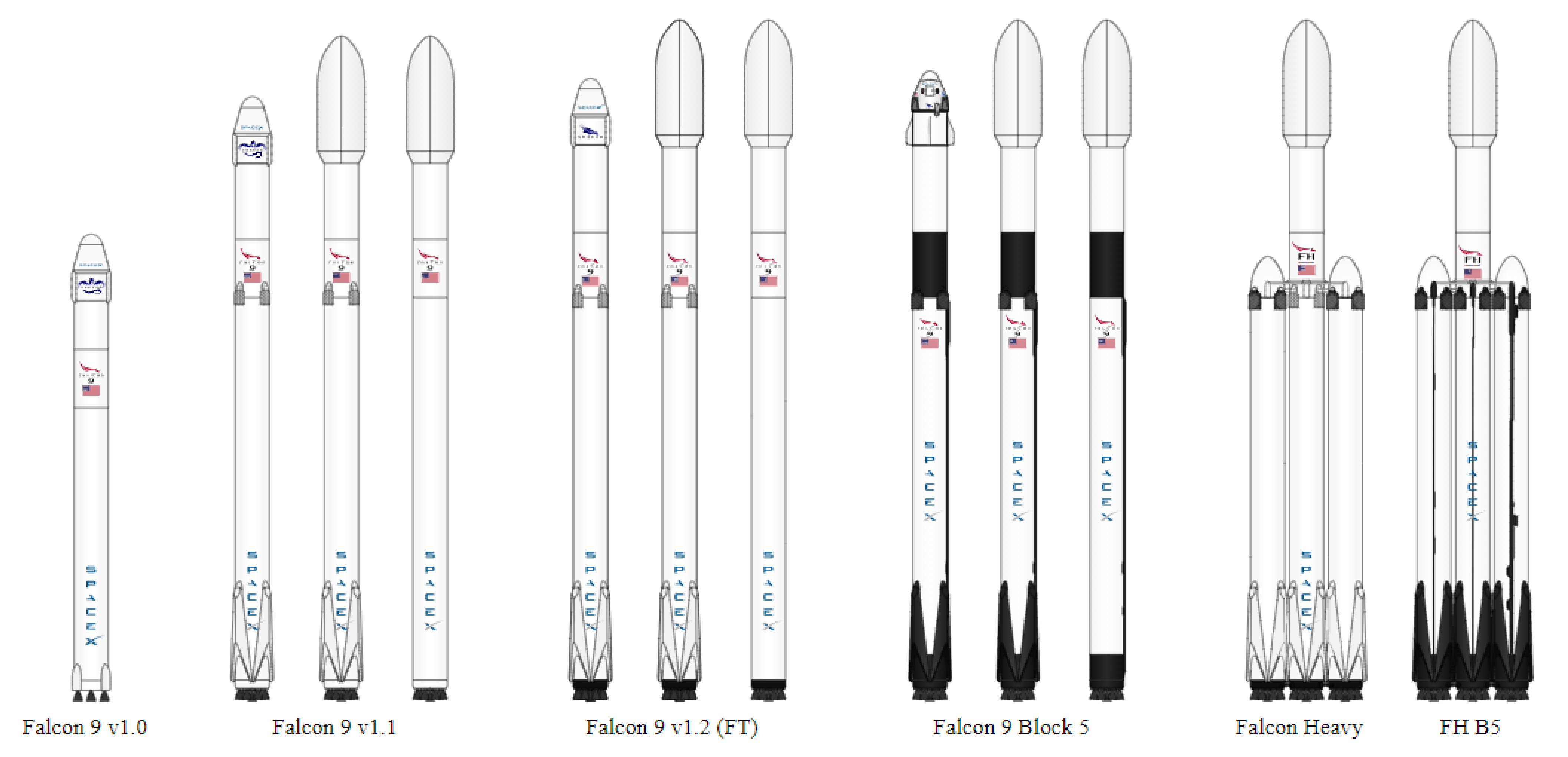


Fig. from: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Appendix 2: Data Collection – SpaceX API – Data Table

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857
5	6	2014-01-06	Falcon 9	3325.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1005	-80.577366	28.561857
6	7	2014-04-18	Falcon 9	2296.000000	ISS	CCAFS SLC 40	True Ocean	1	False	False	True	NaN	1.0	0	B1006	-80.577366	28.561857
7	8	2014-07-14	Falcon 9	1316.000000	LEO	CCAFS SLC 40	True Ocean	1	False	False	True	NaN	1.0	0	B1007	-80.577366	28.561857
8	9	2014-08-05	Falcon 9	4535.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1008	-80.577366	28.561857
9	10	2014-09-07	Falcon 9	4428.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1011	-80.577366	28.561857

The number of launches on each site

```
LaunchSite
CCAFS SLC 40     55
KSC LC 39A       22
VAFB SLC 4E      13
```

The number and occurrence of each orbit

```
Orbit
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
HEO       1
ES-L1     1
SO        1
GEO       1
```

Appendix 3: Confusion Matrix – Classification Errors

What is a Confusion Matrix?
A Confusion Matrix is a **classification error table** that shows how well the model predicted the classes (in our case, whether the Falcon 9 first stage successfully landed or not).

How to Interpret It?

	Predicted: Landed (1)	Predicted: Not Landed (0)
Actual: Landed (1)	✅ **True Positive (TP)** – The model correctly predicted a landing	❌ **False Negative (FN)** – The model incorrectly predicted no landing
Actual: Not Landed (0)	❌ **False Positive (FP)** – The model incorrectly predicted a successful landing	✅ **True Negative (TN)** – The model correctly predicted no landing

Key Metric:
Accuracy = (TP + TN) / (TP + TN + FP + FN)
Shows how **accurate the model is overall**.

Confusion Matrix Insights from Our Project:
✓ **Best performing model: Decision Tree** (**94.4% accuracy**)
✓ **Logistic Regression, SVM, and KNN** achieved **83.3% accuracy**
✓ **Main issue: False Positives (FP)** – The model sometimes **incorrectly predicts a successful landing** when it did not actually occur.

Key Takeaway:
The Confusion Matrix helps us understand where the model makes errors and how we can improve it.

Appendix 4: Project classification models

The project uses **four** classification models:

1. **Logistic Regression**
    - Works well for **linearly separable data**.
    - Outputs the **probability** of the first stage landing.
    - Achieved **83.3% accuracy** in the project.
2. **Support Vector Machine (SVM)**
    - Finds the **optimal hyperplane** to separate classes.
    - Uses different **kernels** (linear, RBF, poly, sigmoid).
    - Achieved **83.3% accuracy** in the project.
3. **Decision Tree**
    - Splits data based on **gini or entropy criteria**.
    - **Easy to interpret** and captures **non-linear relationships** well.
    - **Best-performing model** in the project with **94.4% accuracy**.
4. **K-Nearest Neighbors (KNN)**
    - Classifies a point based on **nearest neighbors**.
    - Works well if **data is well distributed**.
    - Achieved **83.3% accuracy** in the project.

Conclusion:
The best-performing model was Decision Tree (94.4% accuracy). Other models performed **moderately well (83.3%).**

Appendix 5: Classification - Logistic Regression

Key Features

- A **probabilistic** classification model that predicts the probability of a class.
- Uses the **sigmoid function** to transform outputs into probabilities.
- Works well for **linearly separable** problems.

Advantages:

- Easy to interpret results.
- Computationally efficient.
- Provides probability estimates for predictions.

Disadvantages:

- Assumes **linear decision boundary**, which may not always be accurate.
- Sensitive to **outliers**.
- Doesn't work well with **nonlinear relationships**.

Performance in Our Project:

- Accuracy: **83.3%**
- Key issue: Struggles with **complex patterns** in the data.

Appendix 6: Classification - Support Vector Machine (SVM)

Key Features
- Finds the **optimal hyperplane** to separate classes.
- Can use **different kernels** (linear, RBF, polynomial, sigmoid) to map data into a higher-dimensional space.
- Effective for **high-dimensional** data.

Advantages:
- Works well with **small datasets**.
- Handles **high-dimensional data** effectively.
- Robust to **outliers** due to margin maximization.

Disadvantages:
- Computationally expensive for **large datasets**.
- Requires careful selection of **kernel function** and hyperparameters.
- Harder to interpret compared to simpler models.

Performance in Our Project:
- Accuracy: **83.3%**
- Best hyperparameters: **C=1.0, gamma=0.0316, kernel=sigmoid**.
- Key issue: **Tuning kernels** was necessary for improving performance.

Appendix 7: Classification - Decision Tree (Best Model)

Key Features

- Splits data into branches based on **decision rules**.
- Works well with **non-linear relationships**.
- Uses **entropy** or **gini index** to determine splits.

Advantages:

- **Highly interpretable** (resembles human decision-making).
- Handles **non-linear relationships** well.
- need for feature scaling or transformation.

Disadvantages:

- **Prone to overfitting**, especially with deep trees.
- Sensitive to **imbalanced data**.
- Requires **pruning** to avoid unnecessary complexity.

Performance in Our Project:

- Accuracy: **94.4%** (**Best-performing model!**)
- Key hyperparameters:
    - **Criterion: entropy**,
    - **Max depth: 6**,
    - **Splitter: best**,
    - **Min samples split: 5**.

Appendix 8: Classification - K-Nearest Neighbors (KNN)

Key Features

- A **non-parametric** method that classifies data based on the **k-nearest neighbors**.
- Works by measuring the **distance** (e.g., Euclidean) between points.
- The class of a new point is determined by **majority voting** among its neighbors.

Advantages:

- Simple and **easy to implement**.
- Works well with **multi-class problems**.
- No assumption about **data distribution**.

Disadvantages:

- **Computationally expensive** for large datasets.
- Performance depends on **choosing the right k-value**.
- Sensitive to **irrelevant features** and feature scaling.

Performance in Our Project:

- Accuracy: **83.3%**
- Best hyperparameters: **k=10, algorithm=auto, p=1**.
- Key issue: **Performance drops when k is too low or too high**.

Appendix 9: Why did Decision Tree perform the best in the project?

1. Handling Non-Linear Relationships in Data

- Decision Trees **divide the feature space** into regions, allowing them to **capture complex dependencies** between variables.
- Unlike **logistic regression**, which uses a **linear decision boundary**, **decision trees adapt better to intricate patterns**.

Example from the project:
There are **complex dependencies between orbit, payload, and landing success**. Logistic regression failed to capture them effectively, but the decision tree handled them well.

2. Automatic Feature Selection

- Decision Trees **automatically identify the most important features**, filtering out irrelevant ones.
- Unlike **KNN and SVM**, where all features contribute equally, **decision trees focus on the key variables**.

Example from the project:
For classification, the most **influential factors** were: **Payload Mass (kg), Orbit Type, Launch Site**
Other methods, like **KNN**, could have considered less relevant variables, reducing accuracy.

3. Flexibility and Model Power

- Unlike **linear models**, Decision Trees **split data at multiple levels**, making them a powerful tool for complex tasks.
- They allow us to **capture intricate interactions between variables**.

Example from the project:
Decision Trees detected relationships such as:

- Launches from **VAFB SLC-4E** have a **lower success rate** than those from **CCAFS SLC-40**.
- **GTO orbits** are more likely to **fail** compared to **LEO or Polar orbits**.

4. Optimized Hyperparameters

We **fine-tuned the model's parameters**, improving its performance:

- **Criterion: entropy** (considers information gain)
- **Max depth: 6** (prevents overfitting but maintains sufficient depth)
- **Min samples split: 5** (controls data splits)

Without tuning, the tree could have overfitted, but **proper optimization made it powerful and accurate**.

Final Verdict: If the data were **more linear**, logistic regression might have been a better choice. If we had **more data**, SVM could have performed better. However, for **our project, Decision Tree was the optimal solution!**

Thank you!