IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

K Gagan Deep
08-12-2025

Outline

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

Executive Summary

Summary of Methodologies

- **Data collection**: Gathered data via SpaceX API and web scraping.
- **Data Wrangling**: Cleaned data using Pandas.
- **EDA & SQL**: Visualized trends and queried databases to find patterns.
- **Interactive Analytics**: Built geospatial maps (Folium) and dashboards (Plotly Dash).
- **Predictive Analysis**: Trained classification models (Logistic Regression, SVM, Decision Tree Classification, KNN);

Summary of All Results

- **Trends**: Launch success rates have significantly improved since 2013.
- **Key Insight**: KSC LC-39A is the most successful launch site.
- **Safety**: Launched occur near coastlines for safety reasons.
- **Prediction**: The best classification model achieved an accuracy of ~83.33%(Decision Tree).

Introduction

Project Background and Context

- SpaceX has revolutionized the aerospace industry by making rocket launched more affordable through reusability.
- The key to this cost reduction is the ability to recover and reuse the **Falcon 9 first stage.**
- A standard launch costs roughly **$62 million,** but reusing the first stage can save significantly on manufacturing costs.

Problem Statement

- The main challenge is to determine the likelihood of a successful first-stage landing before the launch occurs.
- We want to answer: **“Will the first stage land successfully?”**
- Accurate predictions allow competitors and stakeholders to estimate the true cost of a launch if a rocket lands, the cost is much lower.

Section 1

Methodology

Methodology

Executive Summary

Data collection: Collected data using SpaceX REST API and Web Scraping from Wikipedia.

Data Wrangling: Cleaned data, handled missing values, and performed One-Hot Encoding using Pandas.

Exploratory Data Analysis (EDA): Visualized launch trends using Matplotlib/Seaborn and queried data with SQL.

Interactive Visual Analytics: Create geospatial maps with Folium and dynamic dashboards with Plotly Dash.

Predictive Analysis: Built and tuned classification models (Logistic Regression, SVM, KNN, Decision Tree).

Data Collection

SpaceX REST API:

- Used ‘request’ library to fetch JSON data from ‘api.spacexdata.com’.
- Extracted core features: Rocket, Launch Pad, and Landing Outcome.

Web Scraping (Wikipedia):

- Targeted historical Falcon 9 launch records using BeautifulSoup.
- Parsed HTML tables to recover missing flight data.

Data Integration:

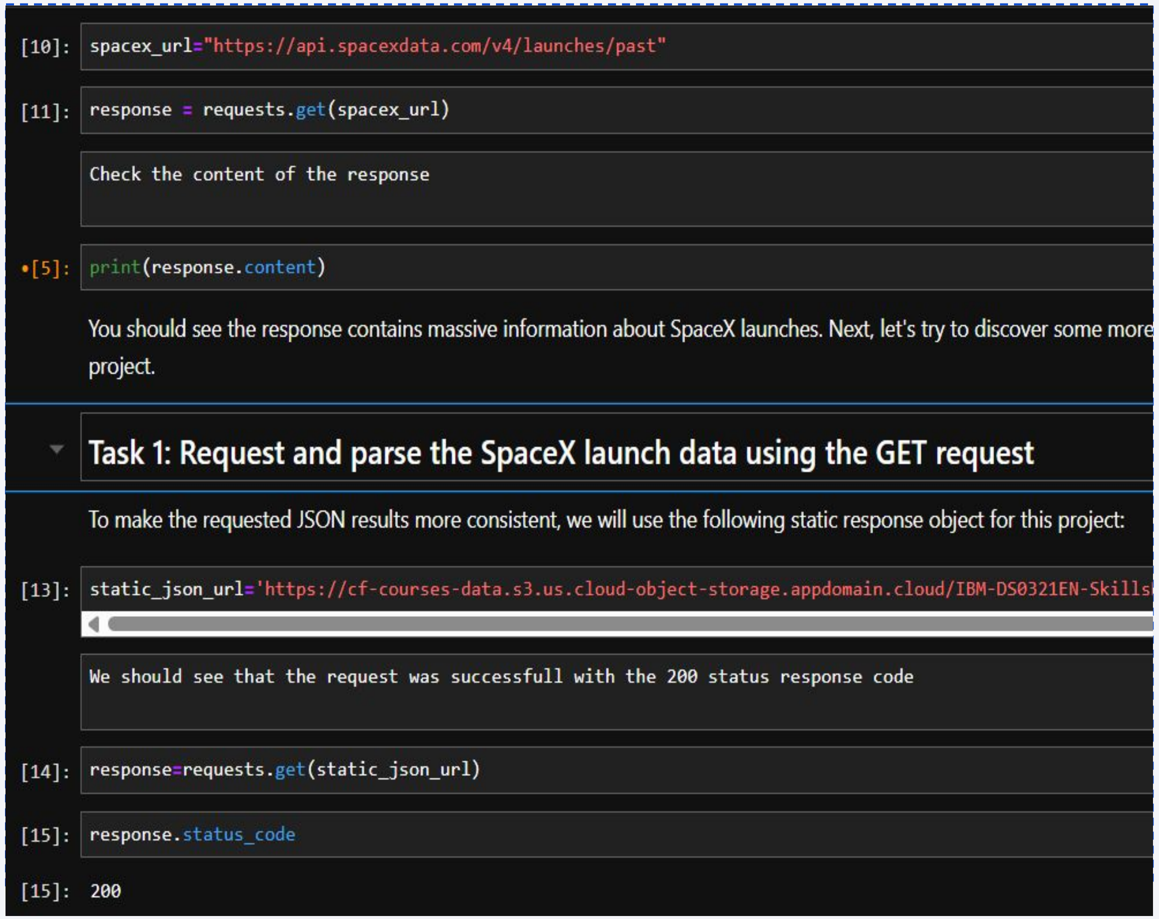
- Merged API and Scraped datasets to ensure a complete launch catalog.

Data Collection – SpaceX API

Api Request Strategy:

- Used 'request' library to fetch JSON data from api.spacex.data.com .
- Filtered specifically for Falcon 9 launch records using Rocket ID.
- Extracted key features: Payload Mass, Orbit, Launch Site, and Landing Outcome.

Github URL: [SpaceX REST API]()



Data Collection - Scraping

Web Scraping Strategy:

- Target the “List of Falcon 9 launched” on Wikipedia using BeautifulSoup.
- Parsed HTML tables to extract launch records missing from the API.
- Collected key details: Flight Number, Date, Time, and Booster Version.

Github URL: [Web Scraping Wiki]

[14]:
```python
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"

headers = {
    "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) "
                  "AppleWebKit/537.36 (KHTML, like Gecko) "
                  "Chrome/91.0.4472.124 Safari/537.36"
}
```

Next, request the HTML page from the above URL and get a `response` object

TASK 1: Request the Falcon9 Launch Wiki page from its URL

First, let's perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.

[23]:
```python
# use requests.get() method with the provided static_url and headers
# assign the response to a object
response = requests.get(static_url, headers=headers)
```

Create a `BeautifulSoup` object from the HTML `response`

[26]:
```python
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(response.text, 'html.parser')
```

Print the page title to verify if the `BeautifulSoup` object was created properly

[27]:
```python
# Use soup.title attribute
print(soup.title)
```

```
<title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

Data Wrangling

Data Processing Strategy:

- Filtered the dataset to retain only Falcon 9 launch records.
- Replaced missing values in ‘Payload Mass’ with the mean of the columns.
- Created a binary Clasa label (1 for Success, 0 for Failure) for prediction.

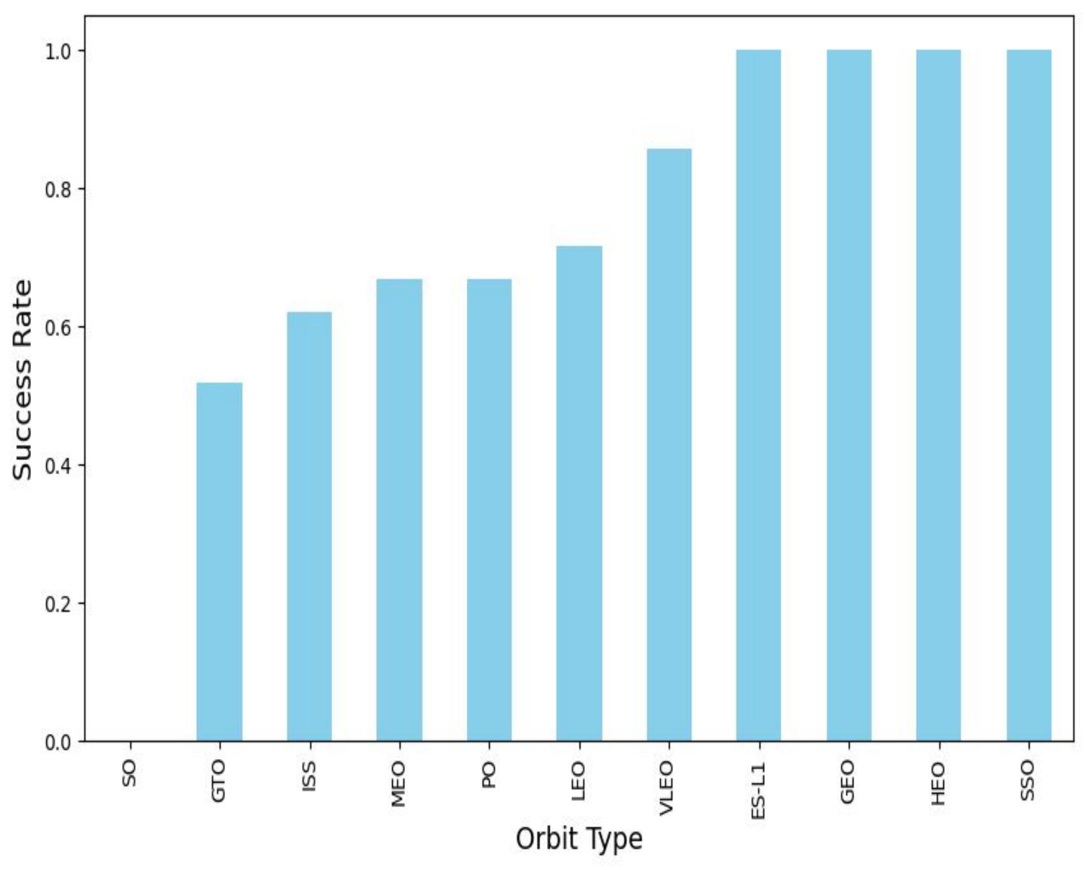
FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude
1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366
2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366
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
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
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

Github URL: [Data Wrangling]

EDA with Data Visualization

Visualisation Strategy:

- Utilized Matplotlib and Seaborn to visualize launch trends and correlations.
- Plotted Scatter Charts to analyze the relationship between Flight Number and Launch Site.
- Created Bar Chart to compare the Success Rate across various Orbit types.



Github URL: [Exploratory Analysis]

EDA with SQL

SQL Query Strategy:

- Executed SQL queries to filter, sort, and aggregate launch data.
- Calculated key metrics such as total payload mass and success counts.
- Ranked landing outcomes to identify performance trends over specific date ranges.

Key Insights Extracted:

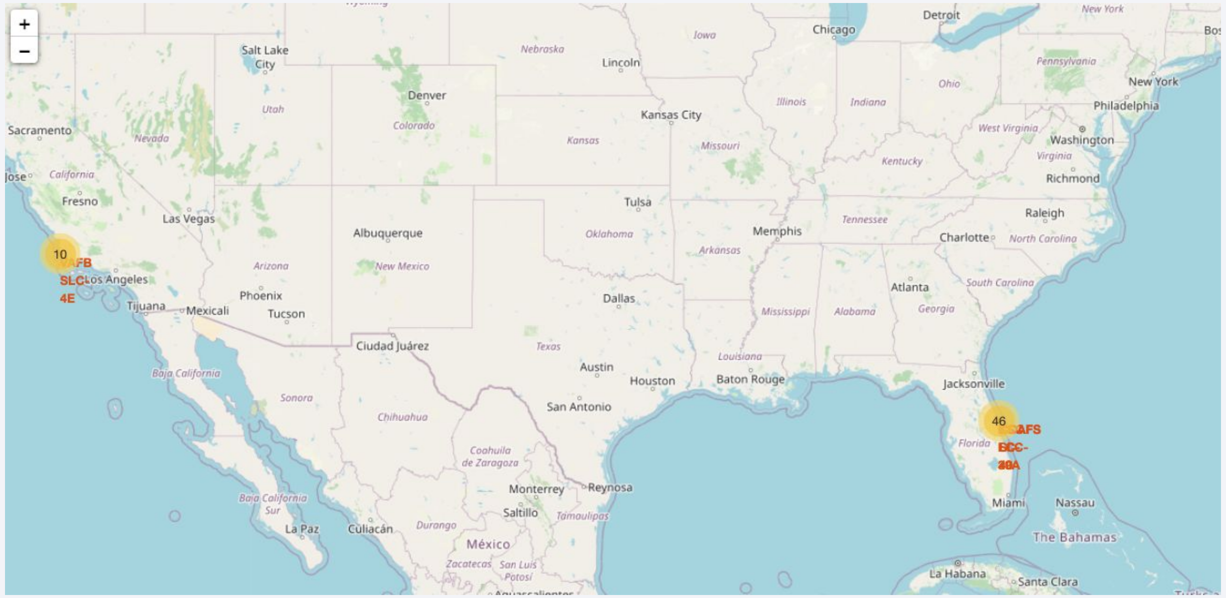
- Identified unique launch sites to understand SpaceX operational bases.
- Determined the specific date of the first successful ground pad landing.
- Ranked landing outcomes between 2010 and 2017 to visualize the trends of increasing reliability.

Github URL: EDA with SQL

Build an Interactive Map with Folium

Geospatial Analysis Strategy:

- Visualize launch site locations on a global map using Folium makers and circles.
- Added color-coded markers to distinguish between successful (green) and failed (red) landings.
- Calculated and visualized distances to nearest coastlines, railways, and highways for safety analysis.



Github URL: [Interactive Map Folium]

Build a Dashboard with Plotly Dash

Dashboard Strategy:

- Built an interactive web application using Plotly Dash to enable real-time data exploration.
- Designed callback functions to dynamically update charts based on user-selected filters.

Key Features & Interaction:

- Integrated a Dropdown Menu to toggle launch site selection for granular performance analysis.
- Visualized outcomes using dynamic Pie Charts for success counts and Scatter Plots for payload analysis.

Github URL: SpaceX Dashboard

Predictive Analysis (Classification)

Model Development Strategy:

- Standardize the data using StandardScaler and split it into training and testing sets.
- Trained and evaluated four classification models: Logistic Regression, SVM, Decision Tree, and KNN.
- Tuned hyperparameters using GridSearchCV to find the best performing parameters for each models.
- Calculated accuracy scores on test data to determine the optimal model for deployment.

Github URL: Predictive modeling

Results

Exploratory Data Analysis Results:
- Presented key insights derived from SQL queries and data visualizations.
- Highlighted trends in launch success rates, orbit types, and payload capacities.

Interactive Analytics Demo in Screenshots:
- Showcased geospatial analysis using Folium map screenshots.
- Demonstrated dynamic interactions from Plotly Dash application.

Predictive Analysis Results:
- Summarized classification model performance and accuracy scores.
- Visualized model evaluation using the Confusion matrix.

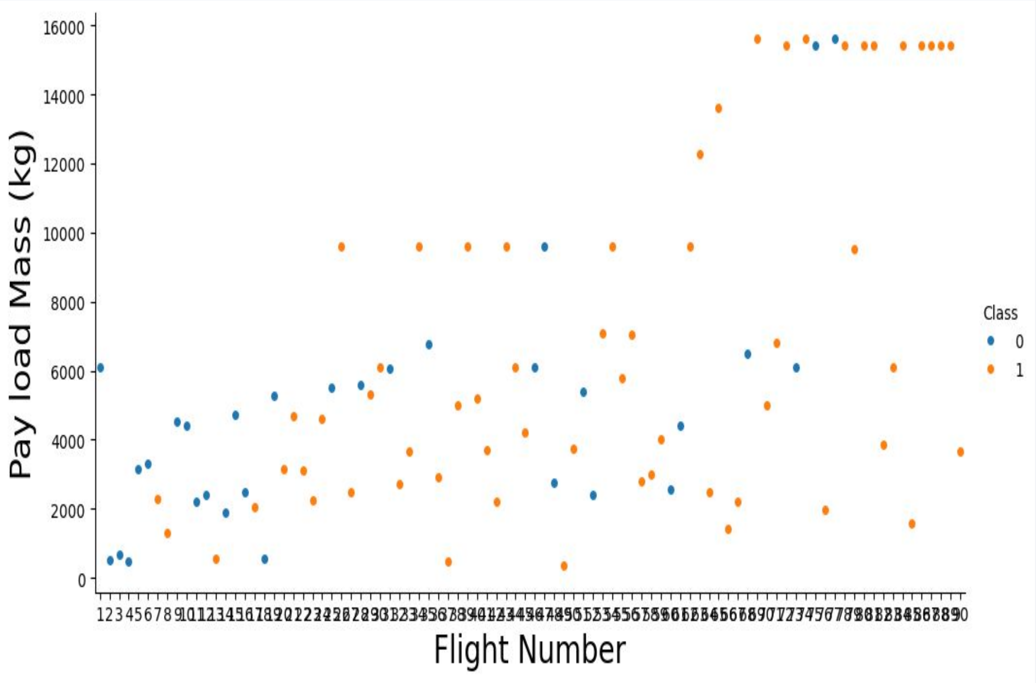
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Scatter Plot Analysis:

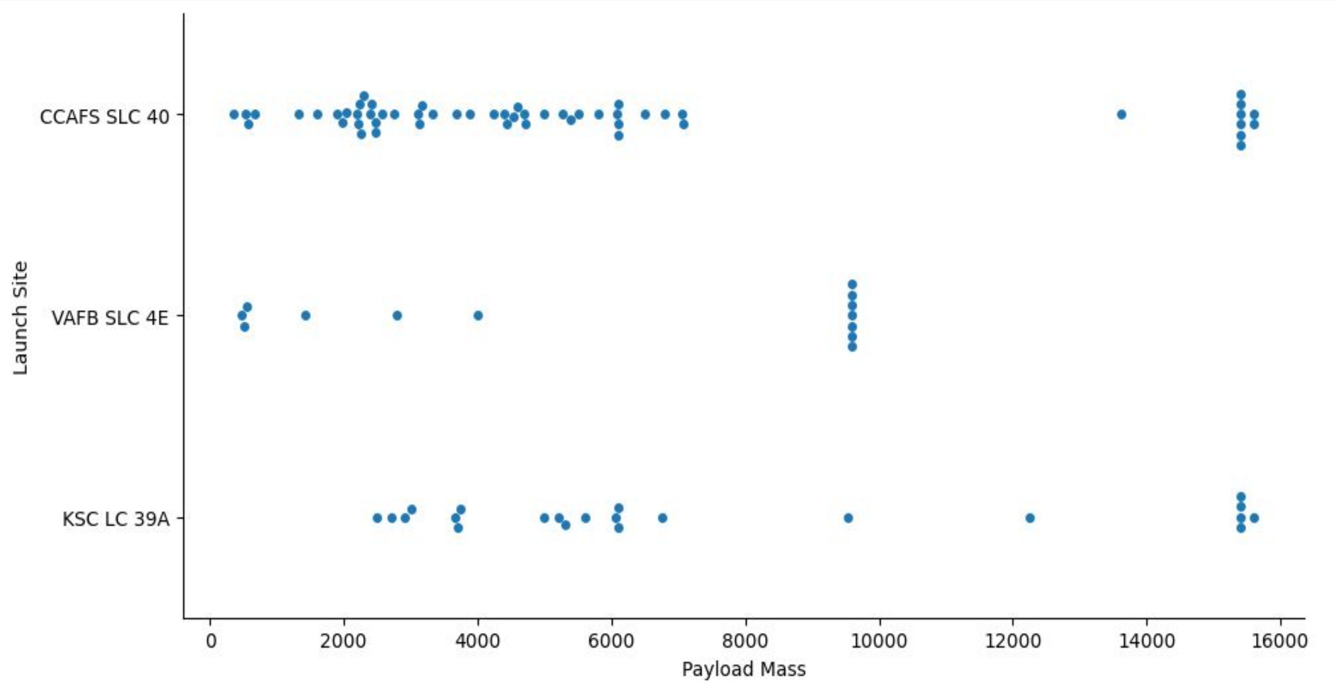
- **Visual Trend:** The plot reveals a strong correlation between Flight Number and Landing Success. As flight increases, the ratio of successful landing (Class 1) improves significantly.
- **Site Performance:**
  CCAFS SLC-40: Shows a mix of early failures and later successes.
  VAFB SLC-4E & KSC LC-39A: Exhibit higher consistency and success rates, particularly in later missions.



Payload vs. Launch Site

Scatter Plot Analysis:

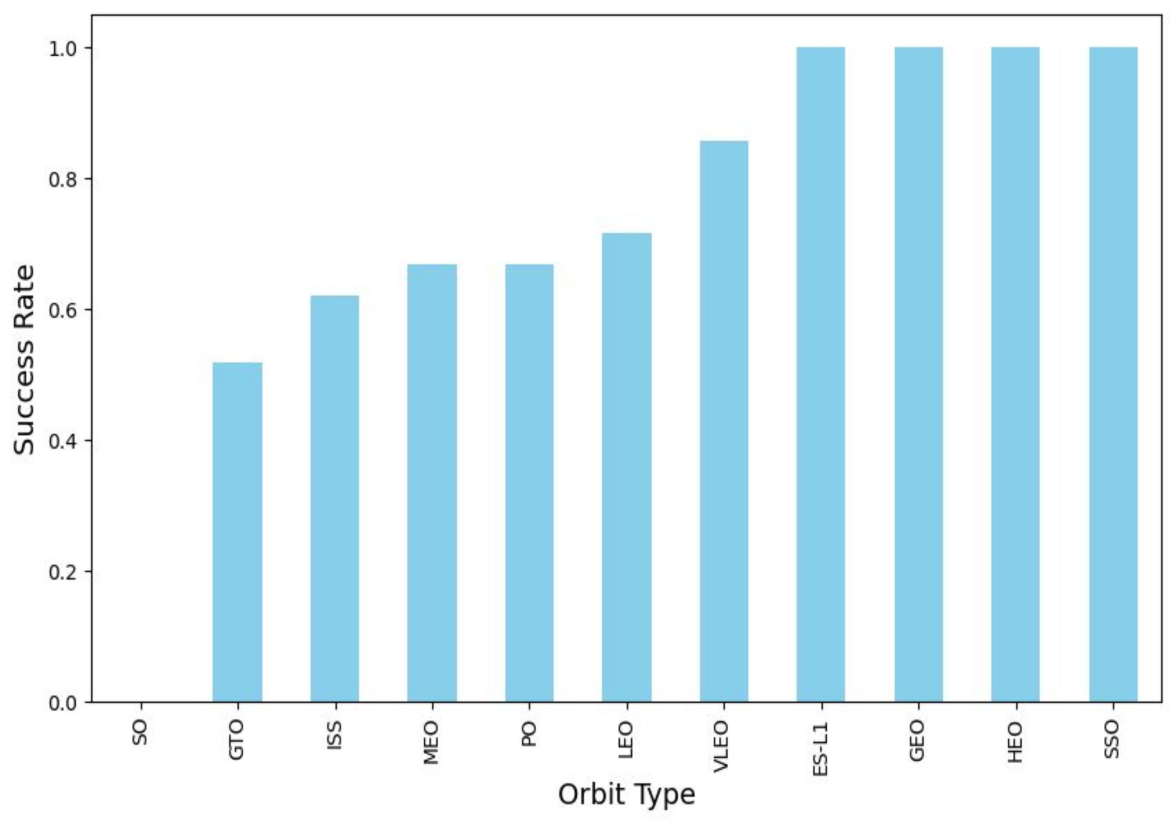
- **VAFB SLC-4E:** Launches only lighter payloads (less than 10,000 kg).
- **Heavy Payloads:** CCAFS SLC-40 AND KSC LC-39A handle the heaviest missions.
- **Outcome:** Higher payloads show varied success rates compared to lighter ones.



Success Rate vs. Orbit Type

Chart Analysis:

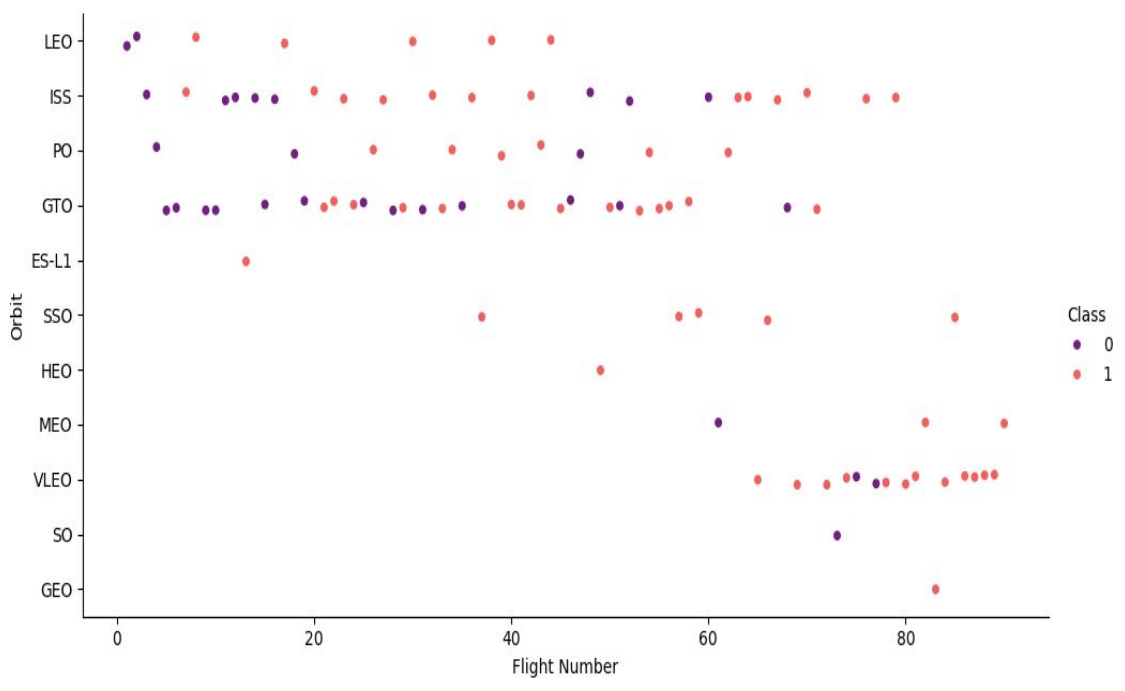
- **High Success:** Orbit like ES-L1, GEO, HEO, and SSO achieved the highest success rates (near 100%).
- **Low Success:** The SO (Sun-Synchronous Orbit) recorded the lowest success rate in the dataset.
- **Trend:** Success variability indicates that orbit trajectory significantly impacts landing difficulty.



Flight Number vs. Orbit Type

Scatter Plot Analysis:

- **LEO Constancy:** Low Earth Orbit (LEO) launches occur consistently across the entire range of flight numbers.
- **Geo Trend:** Geostationary Transfer Orbit (GTO) launches show a distinct increase in frequency as flight numbers get higher.
- **Insight:** Validates the SpaceX expanded its capability to support more complex-high-orbit mission over time.



Payload vs. Orbit Type

Heavy Payloads in LEO:

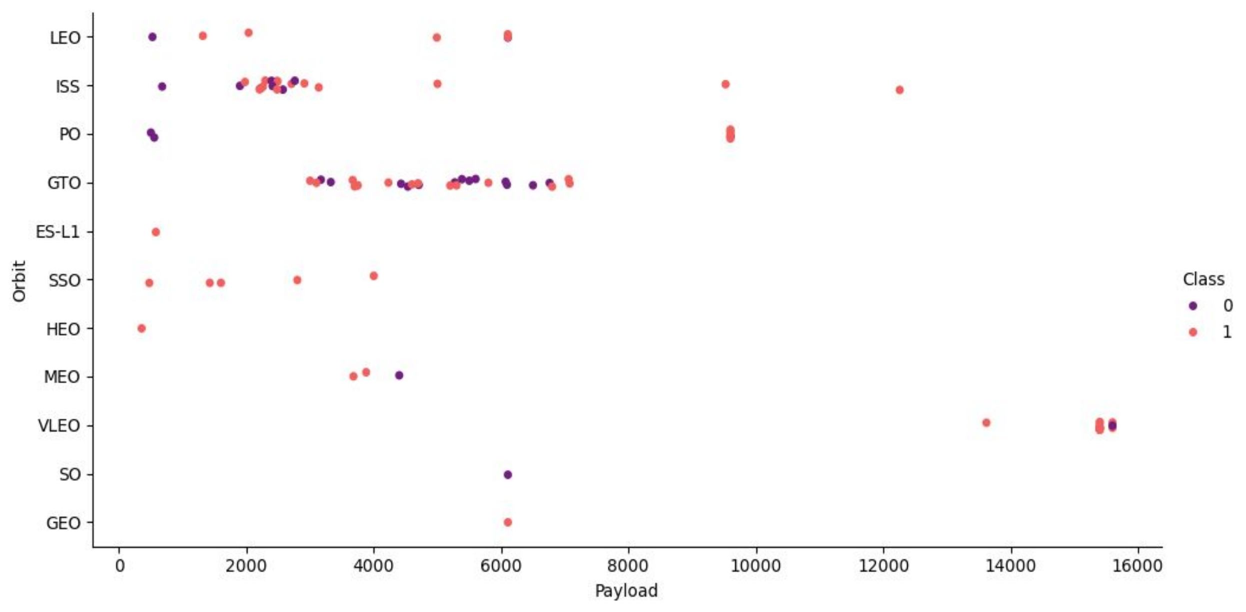
- Highest mass satellites are concentrated in Low Earth Orbit.
- Ex: Starlink Constellations.

Lighter Payloads in High Orbits:

- Payload mass significantly decreases for GTO and GEO.
- Reaching higher altitudes requires more fuel, limiting weight.

Key Insight:

- Inverse relationship: As altitude increases, maximum payloads capacity decreases.



Launch Success Yearly Trend

Positive Trends:

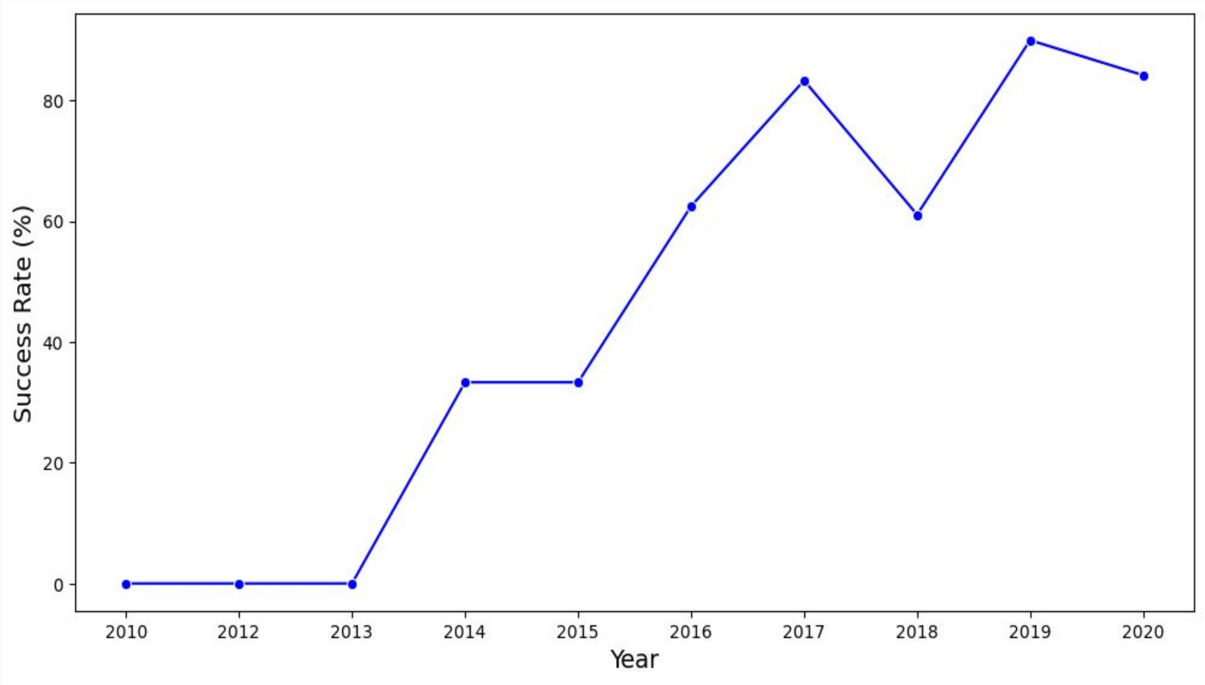
- Success rate has steadily increased since 2013.
- Early failures (2010-2012) were stabilized by 2014.

High Reliability:

- Success rates stabilized near 100% in recent years (2017-2020).
- Demonstrates the maturity of the Falcon 9 platform.

Conclusion:

- Consistent improvements have made SpaceX a reliable launch provider.



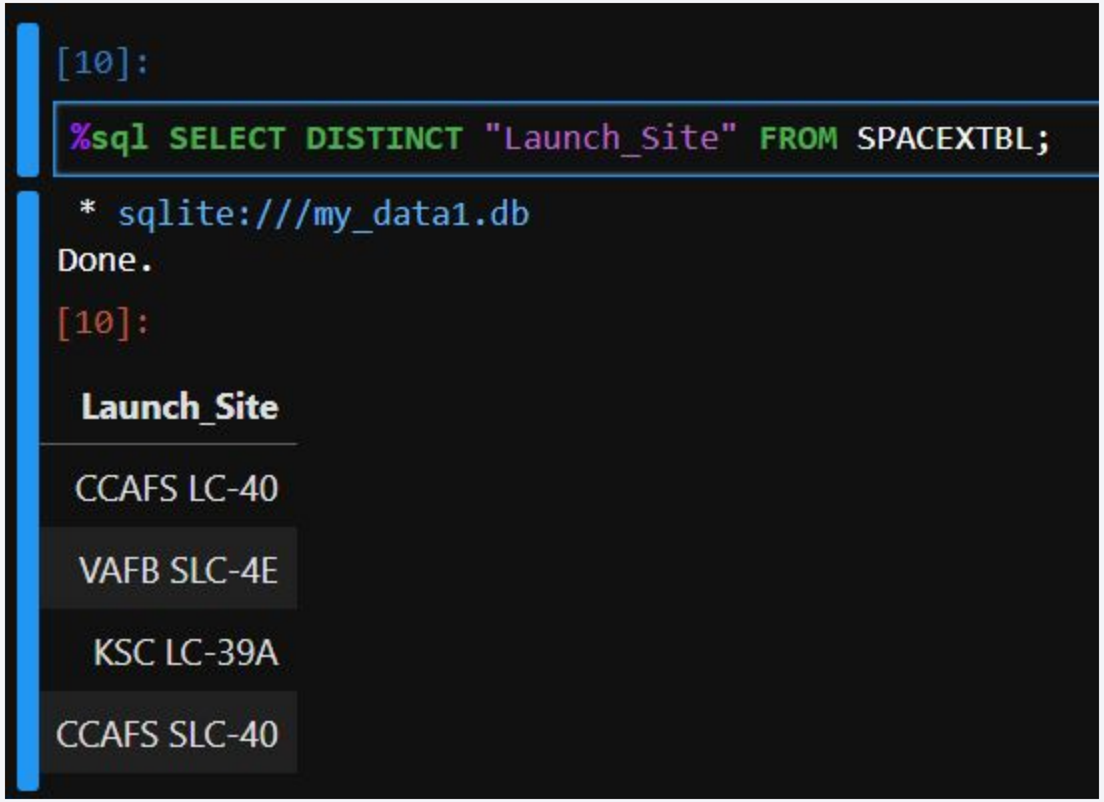
All Launch Site Names

Query Objective:

- Retrieve list of all Unique launch site names.

Result:

- The query returned 4 distinct launch sites:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E



[10]:

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

[10]:

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

Launch Site Names Begin with 'CCA'

Objective

- Filter and display the first 5 records where the launch site starts with “CCA”.

Result:

- The query successfully filtered for Cape Canaveral sites.
- Top 5 records retrieved include launched from CCAFS LC-40.

```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
[11]:
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677

Total Payload Mass

Objective:

- Calculate the total payload mass carried by boosters for NASA mission.

Result:

- The query successfully summed the payload mass for all NASA-affiliated launches.
- Total Mass: 45596 kg

```
%%sql
    SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL
    WHERE "Customer" == 'NASA (CRS)';

 * sqlite:///my_data1.db
Done.
[20]:
SUM("PAYLOAD_MASS__KG_")
                   45596
```

Average Payload Mass by F9 v1.1

Objective:

- Calculate the average payload mass specifically for the F9 v1.1 boosted version.

Result:

- The query calculated the mean mass across all F9 v1.1 missions.
- Average Mass: kg

```
%%sql
    SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL
    WHERE "Booster_Version" == 'F9 v1.1';
```

```
 * sqlite:///my_data1.db
Done.
[13]:
```

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

Objective:

- Identify the specific date of the very first successful landing on a ground pad.

Result:

- The query filtered for successful ground pad outcomes and sorted by date .
- First Success Date: 2015-12-22.

```
%%sql SELECT MIN("Date") FROM SPACEXTBL
    WHERE "Landing_Outcome" == 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
[14]:
```

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Objective:

- List booster names that landed successfully on a drone ship with a payload between 4,100 and 6,000 kg.

Result:

- The query filtered the data for specific mass and landing outcomes.
- Identified boosters that met the strict performance and recovery criteria.

```
%%sql SELECT "Booster_Version" FROM SPACEXTBL
    WHERE "Landing_Outcome" == 'Success (drone ship)'
    AND "PAYLOAD_MASS__KG_" > 4000
    AND "PAYLOAD_MASS__KG_" < 6000;
```

```
 * sqlite:///my_data1.db
Done.
[15]:
```

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

Total Number of Successful and Failure Mission Outcomes

Objective:

- Count and categorize all mission outcomes into “Success” or “Failure”.

Result:

- The query grouped the date bt outcome status.
- Demonstrates the overall reliability ratio of the SpaceX program.

```
%%sql SELECT Mission_Outcome, COUNT(*) FROM SPACEXTBL
   GROUP BY Mission_Outcome;

 * sqlite:///my_data1.db
Done.
[16]:
```

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

Boosters Carried Maximum Payload

Objective:

- Identify the specific booster capable of lifting the maximum recorded payload mass.

Result:

- The query returned the names of boosters that achieved the highest mass capacity in the dataset.
- confirms the peak heavy-lift capabilities of the Falcon 9 platform.

```
%%sql SELECT "Booster_Version" FROM SPACEXTBL WHERE
    "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_")
    FROM SPACEXTBL);
```

```
 * sqlite:///my_data1.db
Done.
[17]:
```

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

2015 Launch Records

Objective:

- List failed drone ship landing in 2015, include booster versions and launch sites.

Result:

- Identified specific failures during early experimental landings.
- The query retrieved the associated Booster Version and Launch Site for each failure event.

```
%%sql
    SELECT substr(Date, 6, 2)
    AS Month, Landing_Outcome, Booster_Version, Launch_Site
    FROM SPACEXTBL
    WHERE substr(Date, 0, 5)='2015'
    AND "Landing_Outcome" == "Failure (drone ship)";
```

```
 * sqlite:///my_data1.db
Done.
[18]:
```

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

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

Objective:

- Rank landing outcomes by count between 2010-06-04 and 2017-03-20.

Result:

- The query counts each outcome type (Success vs. Failure) and sorts them in descending order.
- Highlights the most frequent landing results during this specific period.

```
%%sql SELECT "Landing_Outcome", COUNT(*) AS Total FROM SPACEXTBL
    WHERE "Date" BETWEEN '2010-06-04'
    AND '2017-03-20' GROUP BY "Landing_Outcome"
    ORDER BY Total DESC;
```

```
 * sqlite:///my_data1.db
Done.
```

[19]:

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

Section 3

Launch Sites Proximities Analysis

Launch Site Location (Global Map)

Global Overview:

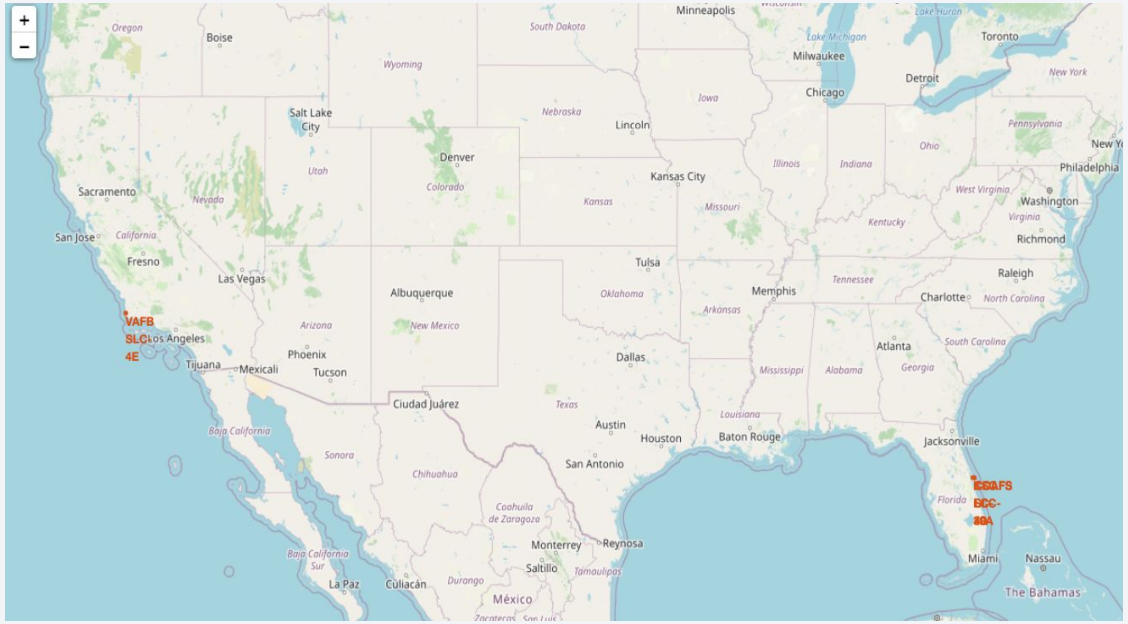
- Map visualizes all SpaceX launch sites marked on the globe.

Strategic Placement:

- All sites are located near coastlines.
- Ensure flight paths remain over water for safety during ascent.

Key Location:

- East Coast (USA): CCAFS SLC-40 & KSC LC-39A.
- West Coast (USA): VAFB SLC-4E.



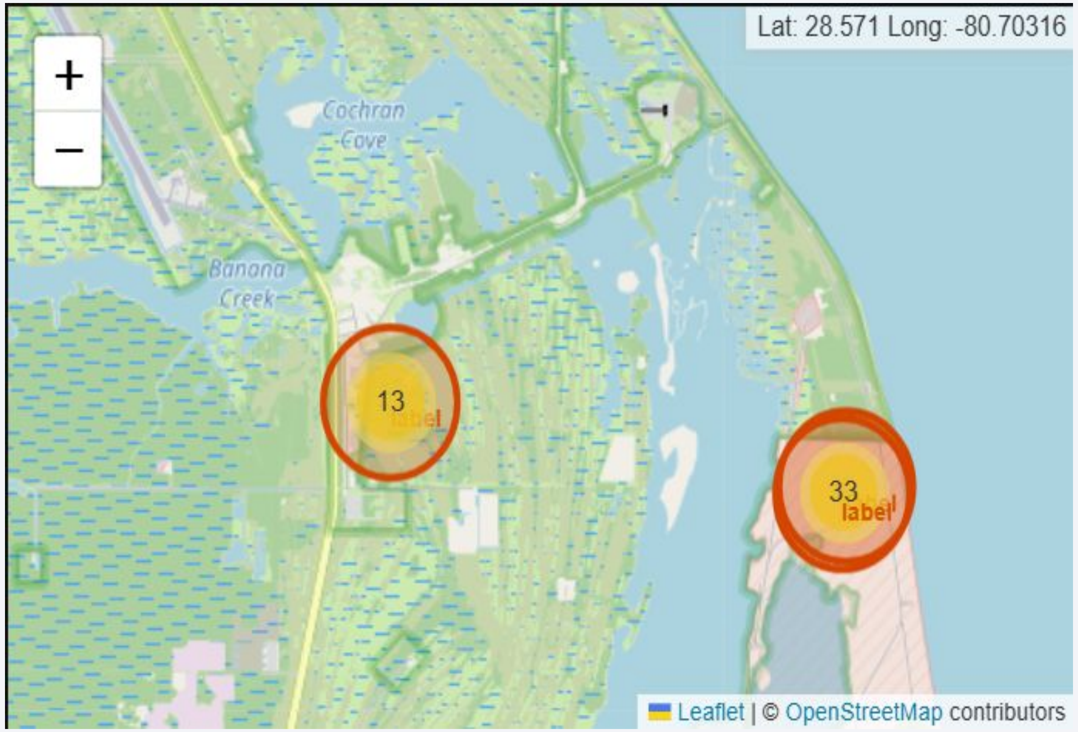
Launch Outcomes Analysis

Color Coding:
- Green Markers: Indicate successful launches.
- Red Markers: Indicates failed launches.

Cluster Analysis:
- KSC LC-39A shows the highest density of green markers (Success).
- CCAFS SLC-40 display a mix of outcomes, reflecting early testing phases.

Conclusion:
- Visualizing outcomes reveals reliability trends specific to each launch pad.

Launch Site Proximities



Coastline Proximity (Safety):

- Launch sites are located close to the coast (<1km).
- Ensures failed rockets crash into the ocean, not land.

Logistical Access (Transport):

- Close proximity to railways and highways.
- Essential for transporting heavy boosters and fuel.

Population Safety:

- Sites are positioned far from city centers to minimize risk to the public.

Section 4

Build a Dashboard with Plotly Dash

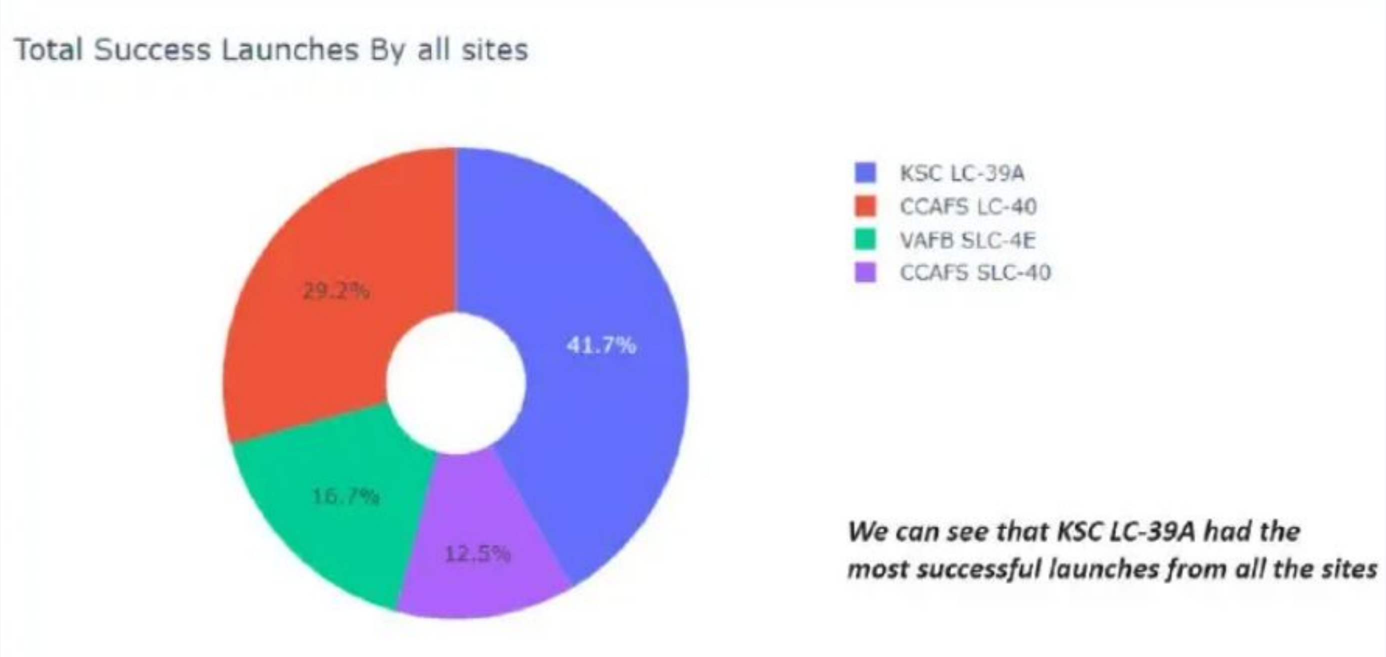
Total Launch Success by Site

Objective:

- Visualize the proportion of successful launched contributed by each site.

Key Findings:

- Leading Site: The chart identified the site with the largest “slice” (highest success count).
- Distribution: Shows how mission success is distributed across the four major launch pads



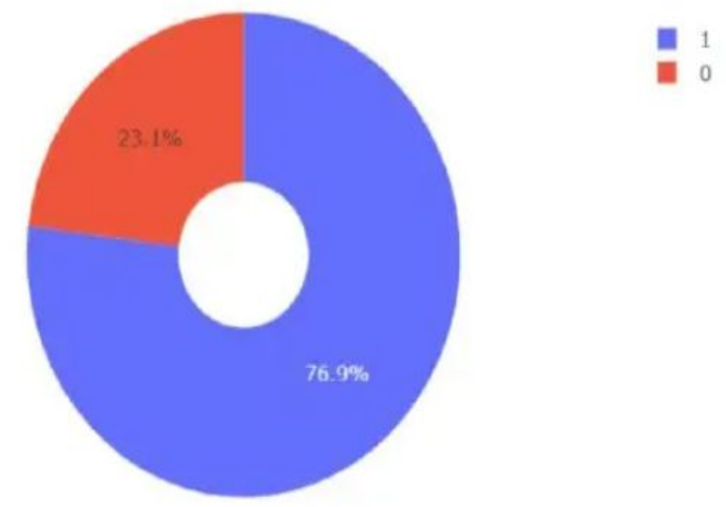
Highest Success Ratio Site

Site Performance:

- Chart isolates the launch site with the highest success-to-failure ratio.
- Displays the breakdown of successful landings (1) vs. failures (0).

Key Finding:

- The large “Success” slice confirms the site as the most reliable in the network.
- High success rate makes it the primary choice for critical payloads.



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Payload vs. Launch Outcome

Payload Range Analysis:

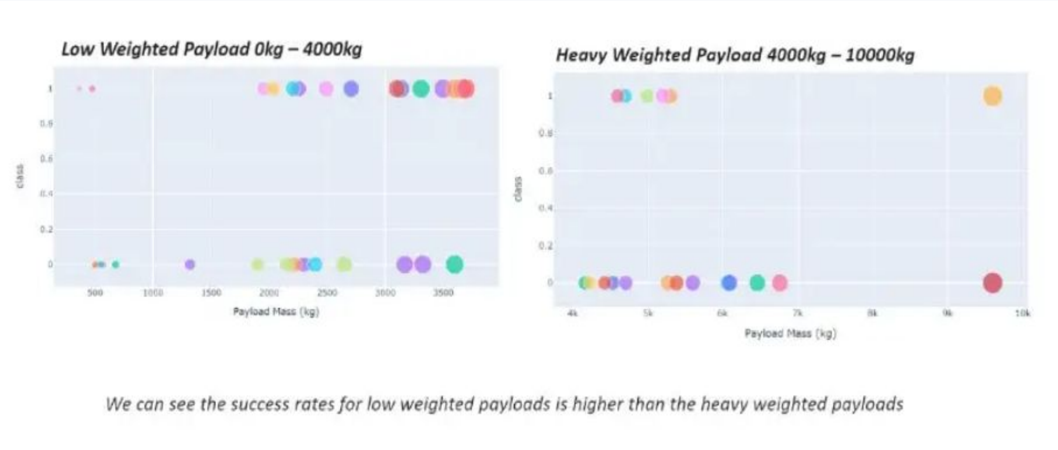
- Visualizes correlation between mass and launch success.
- Finding: Payload between 2,000 kg and 4,000 kg show the highest success rate.

Booster Reliability:

- Plot reveals which booster versions handle specific weight classes best.
- Heavier payloads (e.g., Starlink) also demonstrate high reliability.

Interactive Insight:

- Range slider confirms consistent performance across diverse payload categories.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

Models Evaluated:

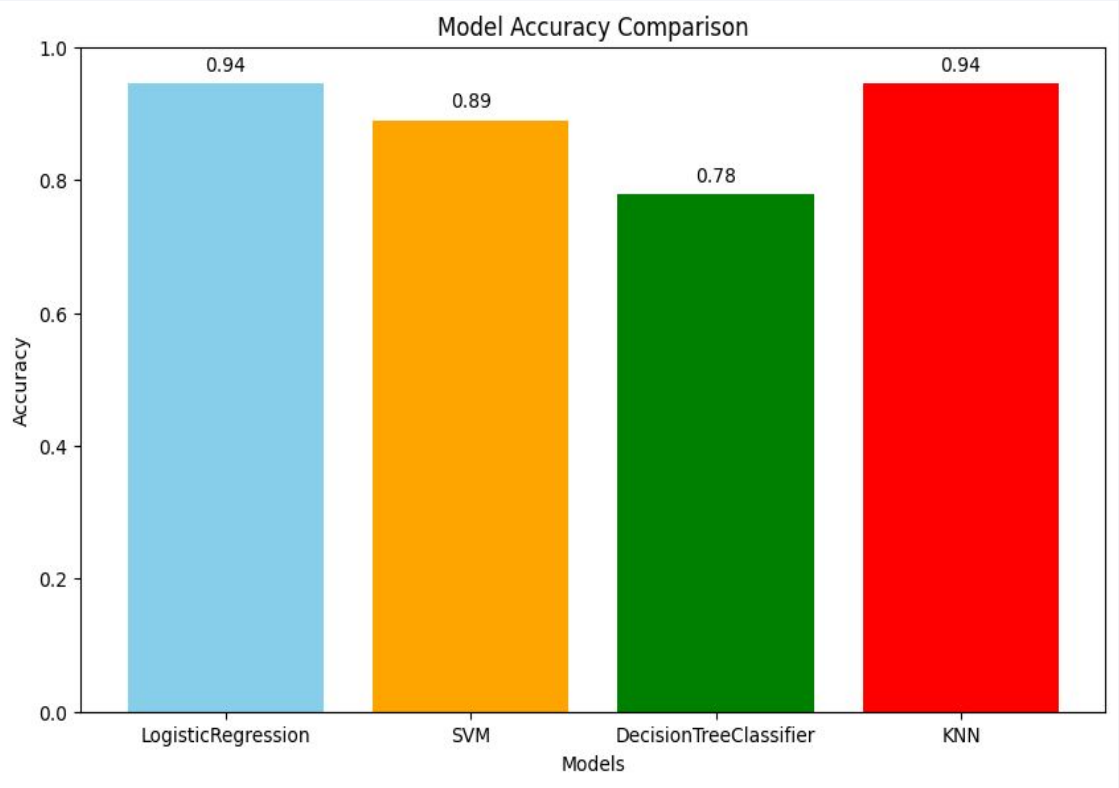
- Logistic Regression, SVM, Decision Tree, and KNN.

Best Performer:

- KNN, Logistic Regression achieved the highest accuracy.
- Outperformed other models in testing.

Conclusion:

- Selected this model for final predictive analysis.



Confusion Matrix

Visualizing Performance:

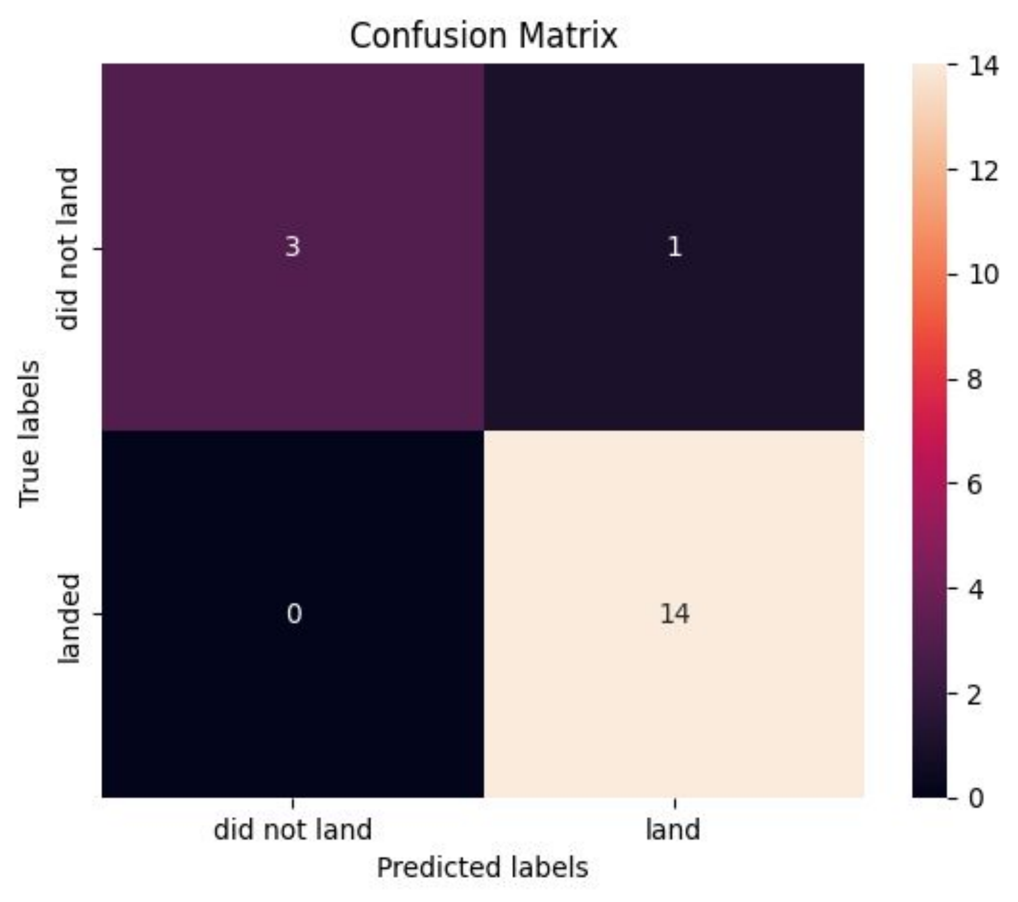
- Grid shows correct predictions vs. errors.
- Diagonal: High values here indicate accurate classification (True Positives/Negatives).

Error Analysis:

- Off-Diagonal: Low values here mean few mistakes (False Positives/Negatives).

Validation:

- Confirms the model reliably distinguishes between successful and failed landings.



Conclusions

Model Performance:

- Machine Learning models successfully predicted landing outcomes with high accuracy.
- Classifiers validated the patterns found during data analysis.

Launch Trends:

- SpaceX success rates have consistently improved over time.
- Technical reliability has stabilized in recent years.

Site Reliability:

- KSC LC-39A proved to be the most reliable launch site.
- Launch site location and proximity to coast are critical safety factors.

Payload Insight:

- Heavy payloads (LEO) achieved high success rates, proving the Falcon 9's heavy-lift capability.

Appendix

Data Collection Source
- SpaceX API Collection Notebook.
- Web Scraping Methodology.

Analysis & Visualization
- EDA and Data Wrangling Code.
- SQL Query Scripts.
- Folium & Plotly Dash Application.

Machine Learning
- Predictive Analysis & Classification Models.
- Model Evaluation Metrics.

Thank you!