IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

MG Ishaque

05 May, 2025

Outline

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

Executive Summary

Objective: Predict SpaceX Falcon 9 first-stage landing success using machine learning models

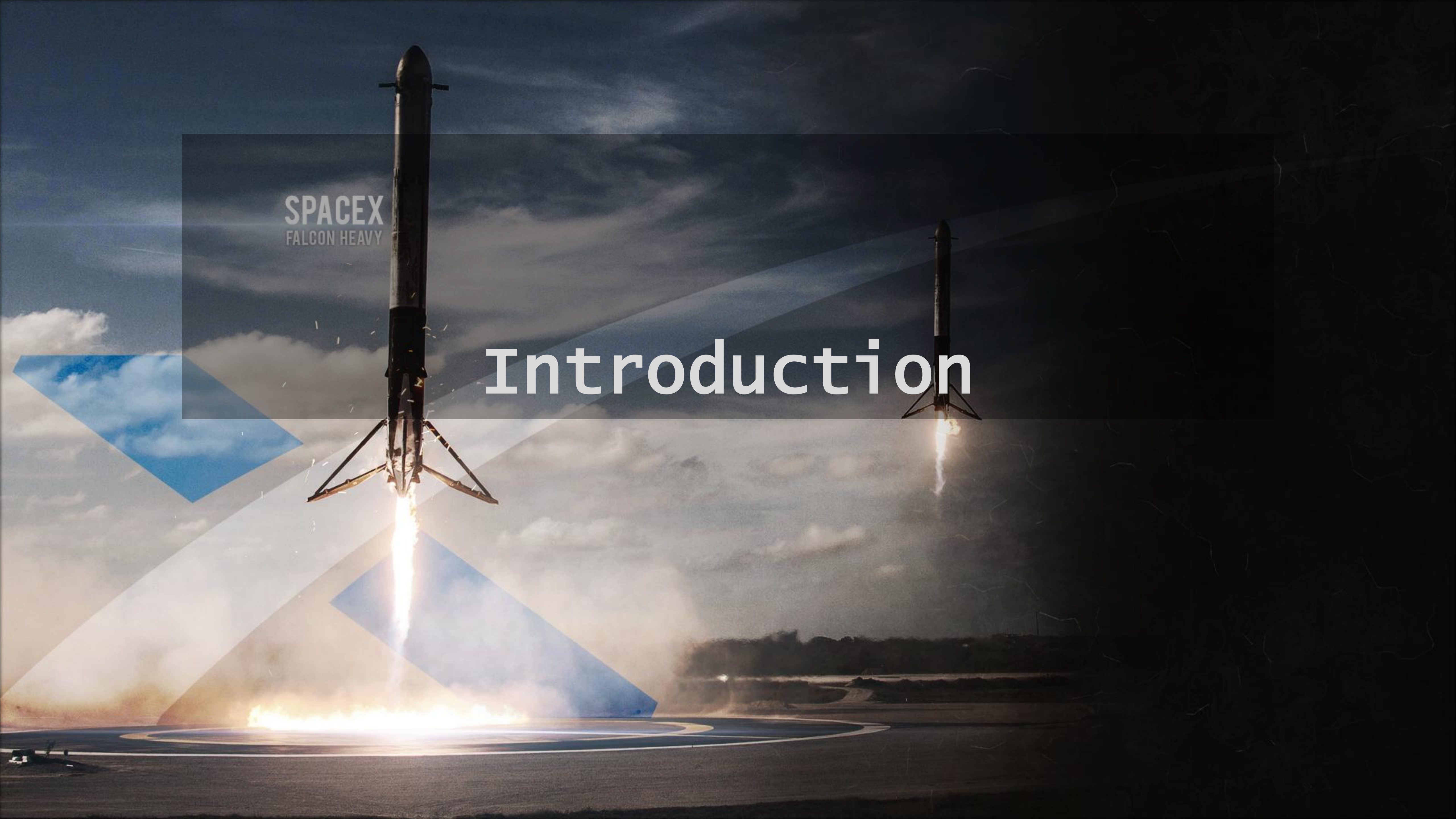
Methodology

- Data collection, wrangling, and EDA
- Interactive visualizations (Plotly Dash, Folium)
- Classification models: Logistic Regression, SVM, KNN, Decision Tree

Key Results:

- Logistic Regression, SVM, and KNN models all achieved an accuracy of 83.3%, indicating strong and consistent performance across these algorithms.
- Critical features: Payload mass, Booster version, Launch site. Certain booster types and launchpads had higher reliability

Impact: Supports mission planning, cost estimation, and booster reuse strategy



Introduction

Why Predicting Falcon 9 First-Stage Landing Success Matters?

🚀 1. Cost Efficiency

- Reusable Falcon 9 boosters significantly reducing the cost per launch.
- First-stage landings are critical for this cost-saving strategy.

📊 2. Mission Planning

- Informed Decisions: Accurate predictions help mission planners assess risk.

📈 3. Commercial Viability:

- Reliable predictions boost investor trust and attract commercial and government clients by proving mission dependability.

🔍 Project Objective

- Build machine learning models to predict Falcon 9 first-stage landing success using SpaceX historical launch data
- Leverage historical SpaceX Falcon 9 launch data for predictive insights
- Analyze data to identify key factors affecting Falcon 9 first-stage landing outcomes

Section 1

Methodology

Methodology Summary

1. Data Collection: Public API (https://api.spacexdata.com/), Wikipedia scraping, course datasets

2. Data Wrangling: JSON normalization, table joins, feature extraction, NULL imputation

3. EDA: Statistical visualizations, SQL queries

4. Interactive Analytics: Folium for maps, Plotly Dash for dashboarding

5. Machine Learning Models
   - Classification algorithms (LogReg, SVM, DT, KNN)
   - Tuning with GridSearchCV
   - Evaluated using performance metrics (accuracy, precision, recall)

Data Collection

🔗 **Sources of Data:**

- **SpaceX API** https://api.spacexdata.com/: provided structured JSON data on launches, rockets, payloads, and cores.
- **Wikipedia**: Web-scraped Falcon 9 mission tables using BeautifulSoup (HTML parsing).
- **IBM Course Datasets**: Additional CSV files containing metadata and launch details.

Data Collection – SpaceX API

Source: https://api.spacexdata.com/v4/launches/query

Process:

- **Initiate API Connection**: Use requests.get on https://api.spacexdata.com/v4/launches/past to retrieve past launch JSON.
- **Normalize JSON**: Apply pd.json_normalize to flatten nested JSON into DataFrame.
- **Supplement with Metadata**: Query /rockets, /payloads, /cores, /launchpads endpoints via API calls and merge by ID.
- GitHub URL: [Data Collection Notebook]

Flowchart of API Data Processing

- Send GET request to API
- Extract nested data
- Convert date format
- Use defined functions to generate specific columns of data
- Combine separate columns into a DataFrame
- Filter out all launches with rockets other than the Falcon 9
- Handle missing values

Data Collection: Wikipedia Web-Scraping

🔗 **Source:**
https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Extracted:
- HTML tables containing launch records
- Flight number, date, mission, landing outcome

Cleaning:
- Dropped unnecessary rows
- Formatted column names
- Converted into Pandas DataFrames

GitHub URL: [Web Scraping Notebook]

- Web Scrape the page to get the entire HTML text
- Create a BeautifulSoup object from the response text content
- Select the tables
- From the launch table, extract the column names from the <th></th> tags
- Create a Pandas DataFrame by parsing the launch tables

Data Wrangling

Merging and Cleaning

- Combined API & web-scraped data
- Filtered for Falcon 9 launches
- Standardized column names & datetime

Feature Engineering

- Created success flag, orbit, payload mass, etc.
- Encoded categorical variables
- Mission outcome: 1 = success, 0 = failure

NULL Handling

- Imputed missing PayloadMass values
- Preserved NULLs for one-hot encoding
- Resolved foreign key references

Final Dataset

- Cleaned, structured DataFrame ready for modeling

GitHub URL: Data Wrangling Notebook

EDA with Data Visualization

Scatter plot Flight Number vs. Launch Site	Scatter Plot Payload Mass vs. Launch Site	Scatter plot Flight Number vs. Orbit Type	Scatter plot Payload Mass vs. Orbit Type	Yearly Success Trend
To show how each site has been utilized over time.	To assess whether payload mass and launch site jointly impact success probability	To understand how launch goals evolved with mission experience.	To reveal mission complexity and its effect on landing outcomes.	Shows progress over time—critical for stakeholders to trust SpaceX reliability.

GitHub URL: [EDA Visualization Notebook]

EDA with SQL

- SQL queries performed in the EDA SQL Notebook:
  - Identified Unique Launch Sites
  - Calculated Total Payload Mass by Customer
  - Filtered Records for Specific Launch Site and Payload Conditions
  - Extracted Year from Launch Date
  - Counted Successful Landings by Orbit Type
  - Computed Average Payload Mass by Orbit
  - Analyzed Yearly Success Trend

GitHub URL: [EDA SQL Notebook]

Build an Interactive Map with Folium

📍 Markers added for :

- All SpaceX launch sites and NASA Johnson Space Center
- Purpose: To visually identify and label key locations on the map for better spatial understanding.

Circles added around each lauch site

- To highlight the spatial coverage of each site and enhance visibility on the map.

📏 Lines (Polylines) Drawn between: CCAFS LC-40 and:

- Nearby coastline
- Rail line
- Perimeter road
- Purpose: To measure distances from the launch site to nearby infrastructure—useful for logistics, safety, and accessibility analysis.

GitHub URL: Folium Map Notebook

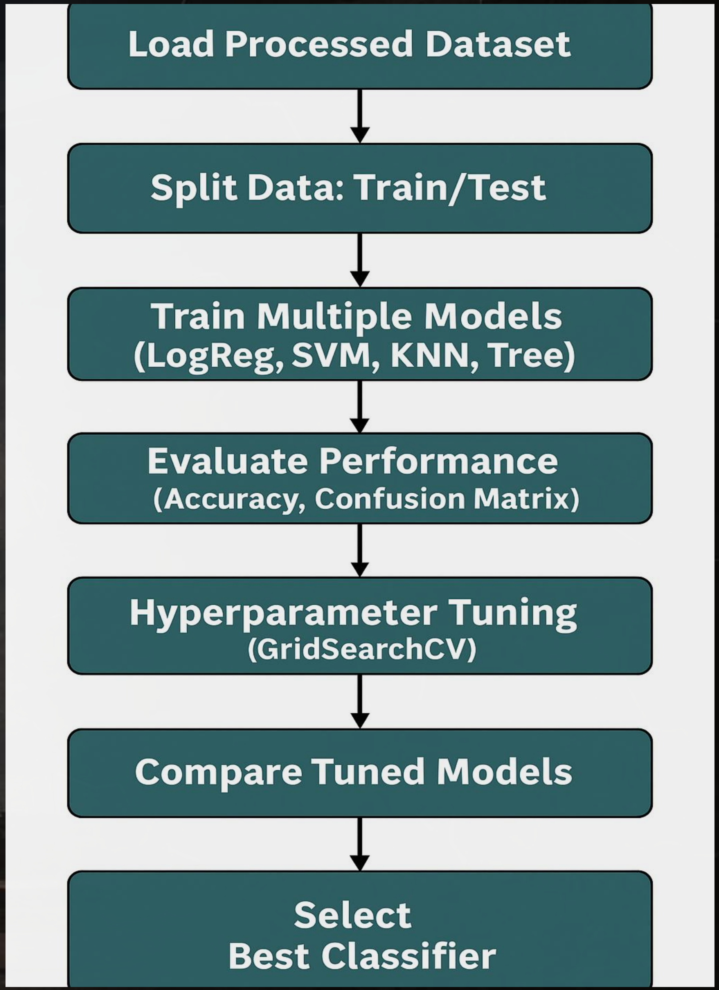
Interactive Plotly Dash Dashboard

Component	Type	Purpose / Insight Provided
Launch Site Dropdown	Dropdown Menu	Enables user-specific analysis of different launch sites or all collectively
Pie Chart	Pie Chart	Displays Pie Chart: a quick summary of success rate either overall or per site.
Payload Mass Slider	Range Slider	Enables users to filter the payload mass range for dynamic graph updates
Scatter Plot	Scatter Plot	Visualizes how payload mass impacts success and highlights booster version performance.

GitHub URL: [Plotly Dash App]

Predictive Analysis (Classification)

📊 Model Development Stage	✍ Description
Data Preparation	Loaded dataset, performed train-test split
Model Building	Logistic Regression, SVM, Decision Tree, KNN
Model Evaluation	Accuracy, confusion matrix, classification report
Hyperparameter Tuning	GridSearchCV on SVM (C, kernel), DT (max_depth), KNN (n_neighbors)
Model Improvement	Used optimal hyperparameters to retrain and improve models
Final Selection	Selected best model based on accuracy and F1-score



GitHub URL: [Machine Learning Notebook]

RESULTS

Results Summary

Category	Insight / Result
EDA: Launch Sites	KSC LC-39A: ~76.9% success rate; not the highest performing site
EDA: Payload Impact	Non-linear relation; mid-range payloads more successful
EDA: Booster Version	Falcon 9 FT associated with higher success
EDA: Launch Year Trend	More recent years show improved success rates
Interactive Charts	Pie charts (success per site), scatter plots (payload vs. success), sliders for filtering payload/site
Dashboard Filters	Interactive Dash dashboard includes dropdowns and sliders for payload mass and site selection
Best ML Model (CV/Test)	Decision Tree – CV: 0.875, Test Acc: 0.889
Best Model Confusion Matrix	Logistic Regression, SVM, KNN – all around 0.833 test accuracy
Confusion Matrix Insight	Decision Tree has best accuracy, but possible higher false negatives; SVM/LogReg more balanced

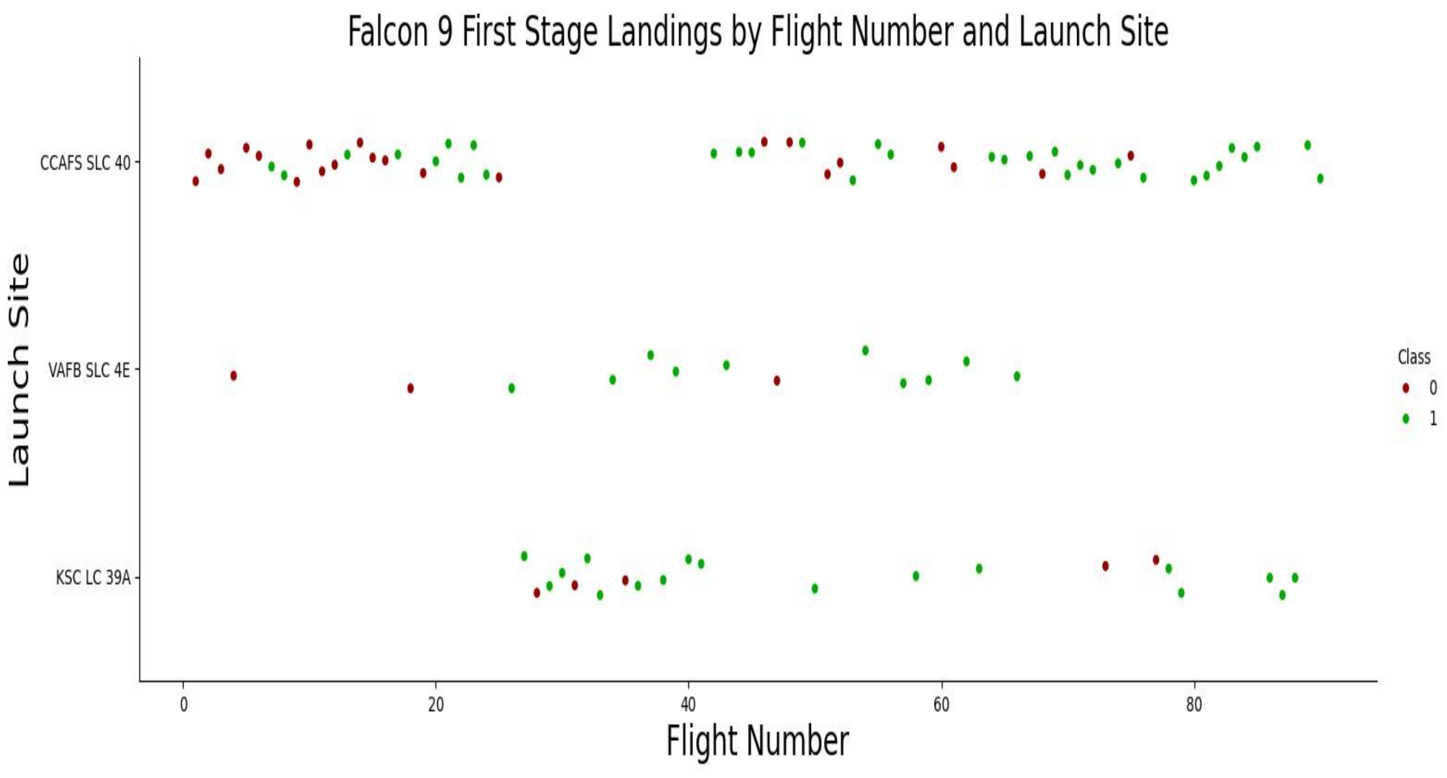
Section 1

Insights drawn from EDA

Flight Number vs. Launch Site

Insight:

- **Operational Focus:** Frequent use of specific sites suggests reliability and preference.
- **Learning Curve:** Early flights at each site may show more failures; later flights trend toward success
- **Site Maturity:** Some sites may have fewer flights but higher success rates.

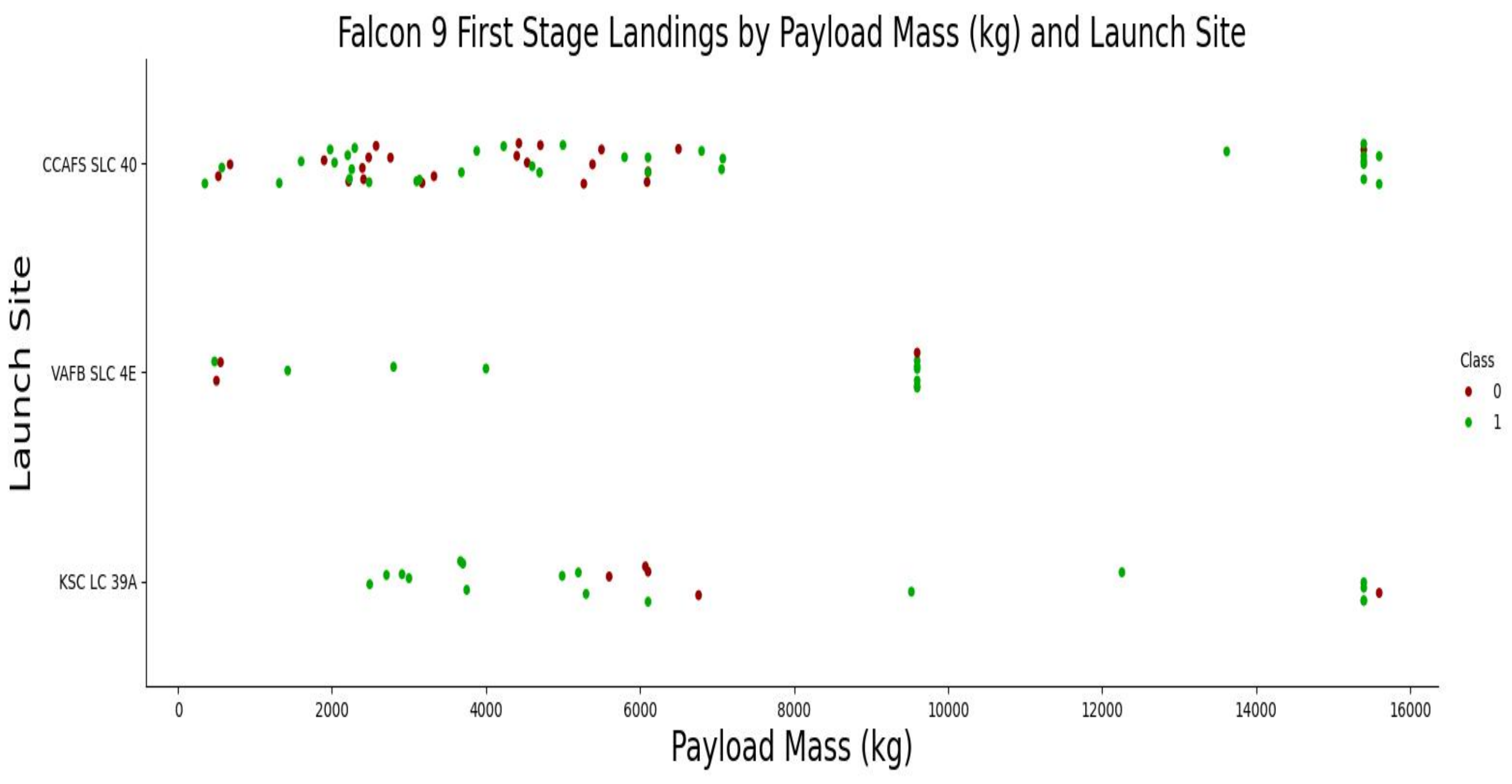


Failed landings are indicated by ● red markers and successful landings by ● green markers

Payload Mass vs. Launch Site

Success Trends by Mass:

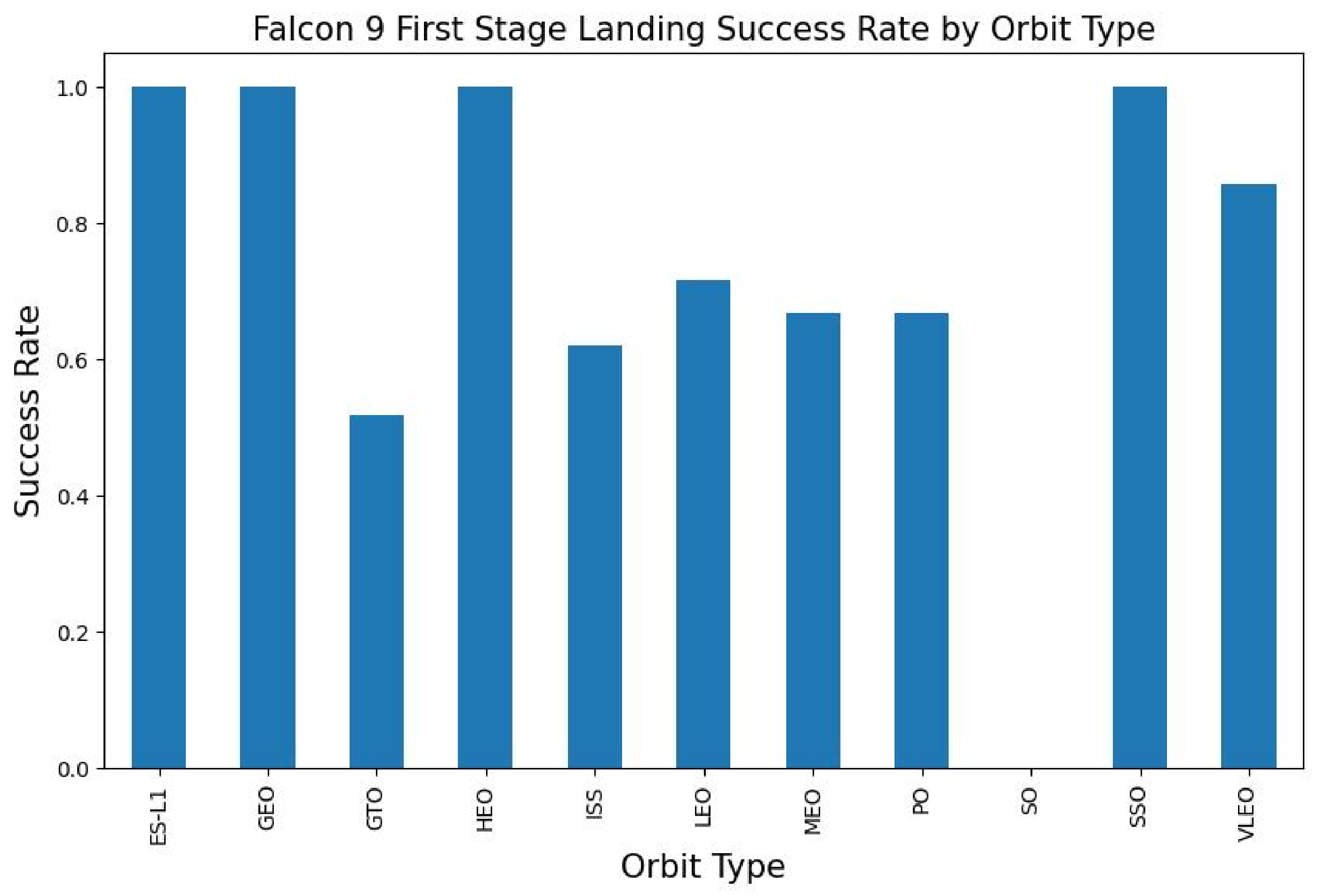
- The larger payloads tend to fail at certain sites.
- VAFB SLC-4E may be used more for polar or LEO missions, often with smaller payloads.
- KSC LC-39A may be use for heavier payloads, for more complex missions like GTO or interplanetary launches
- **Site Optimization:** Guides future launch planning to match payloads with the most reliable launch site



Failed landings are indicated by ● red markers and successful landings by ● green markers

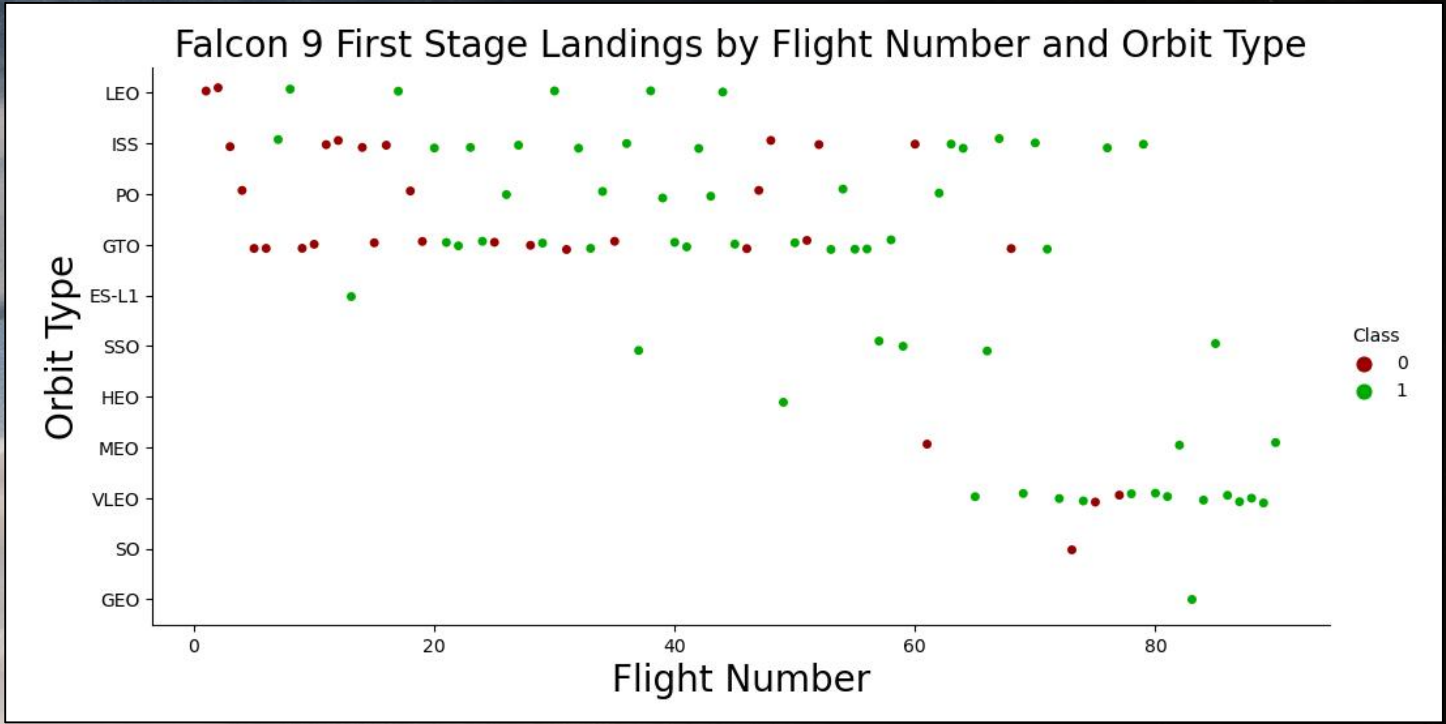
Success Rate vs. Orbit Type

- ES-L1, SSO, HEO and GEO orbits have no failed first stage landings.
- SO orbits have no successful first stage landings.
- ISS, LEO, MEO, PO has similar successful first stage landings.
- GTO has around than 50% successful first stage landings.



Flight Number vs. Orbit Type

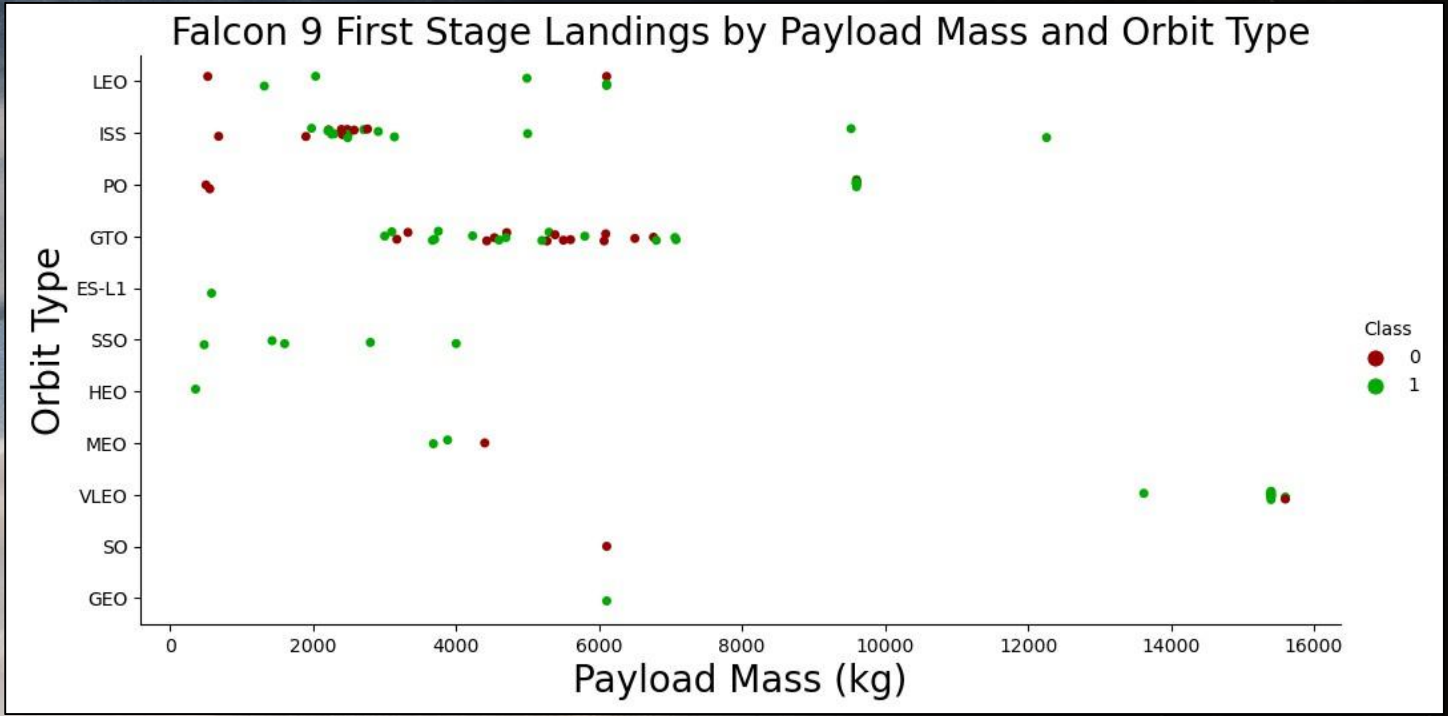
- There is a positive correlation between flight number and success rate.
- Larger flight numbers were associated with higher success rates.
- Falcon 9, are specialized for specific orbits, LEO, ISS, PO, GTO, VLEO



Failed landings are indicated by ● red markers and successful landings by ● green markers

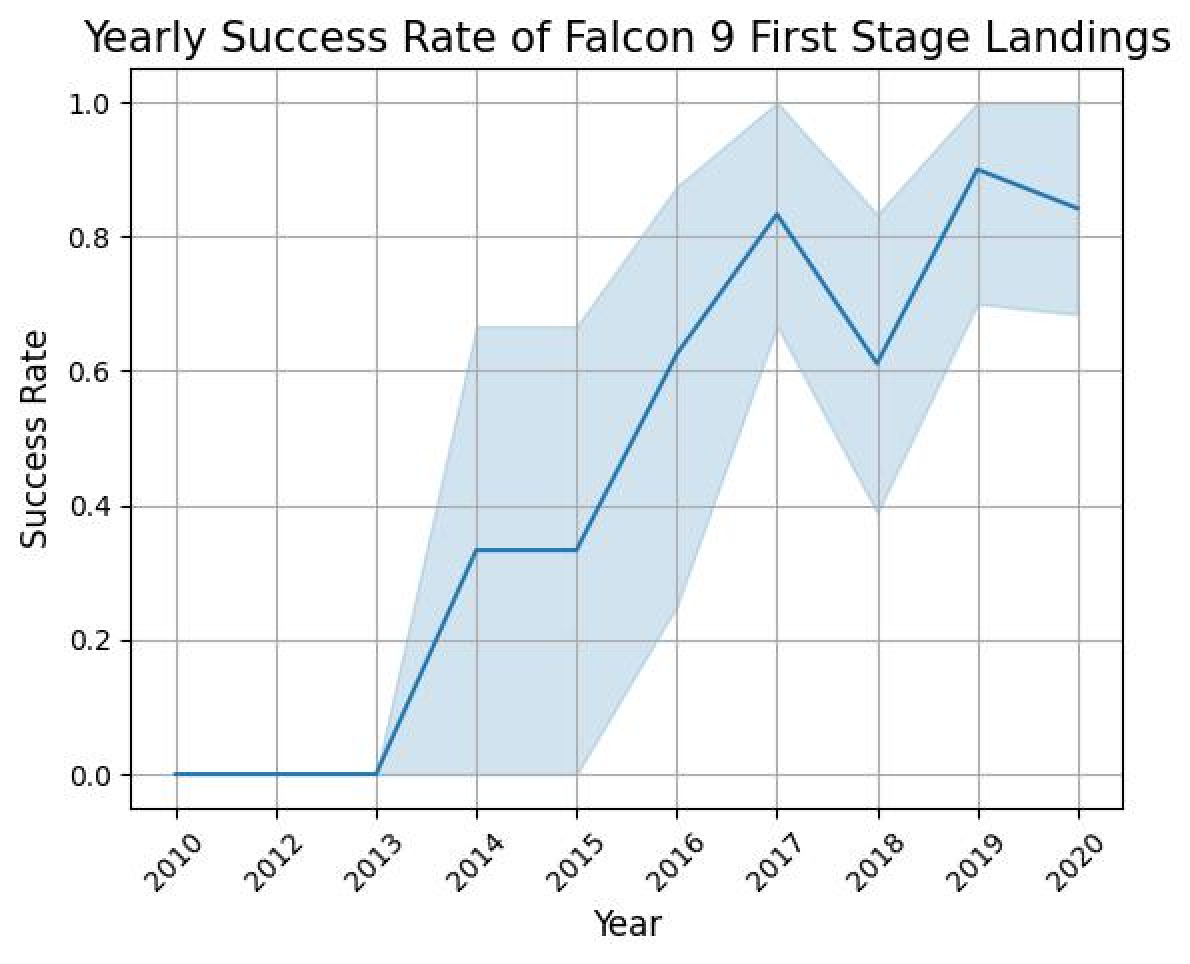
Payload vs. Orbit Type

- Failures are rare and don't cluster around any specific orbit or payload size.
- Success rate appeared to have no obvious correlation with payload mass.



Launch Success Yearly Trend

- **Indicates increasing consistency and fewer failures over time.**
- 2010–2013: No successful landings; experimental phase.
- 2014-2016: Rapid improvement in landing technology; success rate rises above 60%.
- 2017–2020: High reliability achieved; peak near 90% in 2019.



All Launch Site Names

- **Question:** Display the names of the unique launch sites in the space mission
- **query = 'SELECT DISTINCT Launch_Site FROM SPACEXTABLE'**
- **Result:**

	Launch_Site
0	CCAFS LC-40
1	VAFB SLC-4E
2	KSC LC-39A
3	CCAFS SLC-40

- **Explanation:** There are four unique launch sites.

Launch Site Names That Begin with 'KSC'

- **Task:**Display 5 records where launch sites begin with the string'KSC'
- query = 'SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE "KSC%" LIMIT 5'
- Result:

	Landing_Outcome
0	Success (ground pad)
1	No attempt
2	Success (drone ship)
3	Success (ground pad)
4	No attempt

	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome \
0	2490	LEO (ISS)	NASA (CRS)	Success
1	5600	GTO	EchoStar	Success
2	5300	GTO	SES	Success
3	5300	LEO	NRO	Success
4	6070	GTO	Inmarsat	Success

	Landing_Outcome
0	Success (ground pad)
1	No attempt
2	Success (drone ship)
3	Success (ground pad)
4	No attempt

- **Explanation:** This is a fairly straightforward sampling mechanism used to gain a sense of the data contained in the database table.

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
query = 'SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" LIKE "%NASA (CRS)%"'
```

Result:

```
Total_payload_mass: 48213
```

- **Explanation:** The total payload carried by boosters from NASA is 48,213 kg.

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
query = 'SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Booster_Version" = "F9 v1.1"'
```

Result:

```
Average_payload_mass: 2928.4
```

- **Explanation:** The average payload mass carried by booster version F9 v1.1 is 2,928 kg.

First Successful Ground Landing Date

- **Question:** On which date did the first successful landing outcome on ground pad occur?

```
query = '''
    SELECT MIN(Date) AS "First Successful Landing Outcome Date"
    FROM SPACEXTABLE
    WHERE "Landing_Outcome" LIKE 'Success (ground pad)'
'''
```

Result

First Successful Landing Outcome Date
2015-12-22

- **Explanation:** The first successful landing outcome on ground pad occurred on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- **Question:** What are the names of the boosters which have successfully landed on drone ship and had a payload mass greater than 4000 but less than 6000?

- query = '''

```
    SELECT DISTINCT "Booster_Version"
    FROM SPACEXTABLE
    WHERE "Landing_Outcome" = "Success (drone ship)"
    AND "PAYLOAD_MASS__KG_" > 4000
    AND "PAYLOAD_MASS__KG_" < 6000
'''
```

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

- **Explanation:** The four booster versions that have successfully landed on drone ship with a payload mass greater than 4,000 kg but less than 6,000 kg are listed above.

Total Number of Successful and Failure Mission Outcomes

- **Question:** What was the total number of successful and failed mission outcomes?

- query = '''
```
    SELECT
        (SELECT COUNT(*)
         FROM SPACEXTABLE
         WHERE LOWER("Landing_Outcome") LIKE '%success%') AS "Success",
        COUNT(*) AS "Failure"
    FROM SPACEXTABLE
    WHERE LOWER("Landing_Outcome") NOT LIKE '%success%';
'''
```

- **Result:**

Success	Failure
61	40

- **Explanation:** There were 61 successful and 40 failed mission outcomes.

Boosters Carried Maximum Payload

- **Question:** What were the names of the boosters which have carried the maximum payload mass?

  - Query: SELECT booster_version, payload_mass__kg_ FROM SPACEXDATASET WHERE payload_mass__kg_ = (SELECT max(payload_mass__kg_) FROM SPACEXDATASET);

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

- **Explanation:** The maximum payload mass carried in this dataset is 15,600 kg.
- Twelve (12) separate Falcon 9 boosters carried this amount of payload mass.

2017 Launch Records

- **Task:** List the failed landing_outcomes in drone ship, their booster versions,and launch site names for records in year 2017.

```
query = 'SELECT CASE substr(Date, 6, 2) WHEN "01" THEN "January" WHEN "02" THEN "February"
WHEN "03" THEN "March" WHEN "04" THEN "April" WHEN "05" THEN "May" WHEN "06" THEN "June"
WHEN "07" THEN "July" WHEN "08" THEN "August" WHEN "09" THEN "September" WHEN "10" THEN
"October" WHEN "11" THEN "November" WHEN "12" THEN "December" END AS Month,
"Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE
"Landing_Outcome" = "Failure (drone ship)" AND substr(Date, 0, 5) = "2017"'
```

Result

	Month_Name	Landing_Outcome	Booster_Version	Launch_Site
0	February	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
1	May	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
2	June	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
3	August	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
4	September	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
5	December	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

- **Explanation:** These entries indicate a consistent series of successful landings on ground pads throughout the year 2017.

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

- **Task:** Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.
- **Query:**

```
SELECT `landing__outcome`, count(`landing__outcome`) AS ‘Count’ FROM `SPACEXDATASET` WHERE `DATE`
BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY `landing outcome` ORDER BY count(`landing__outcome`) DESC;
```

- **Result:**

landing__outcome	Count
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

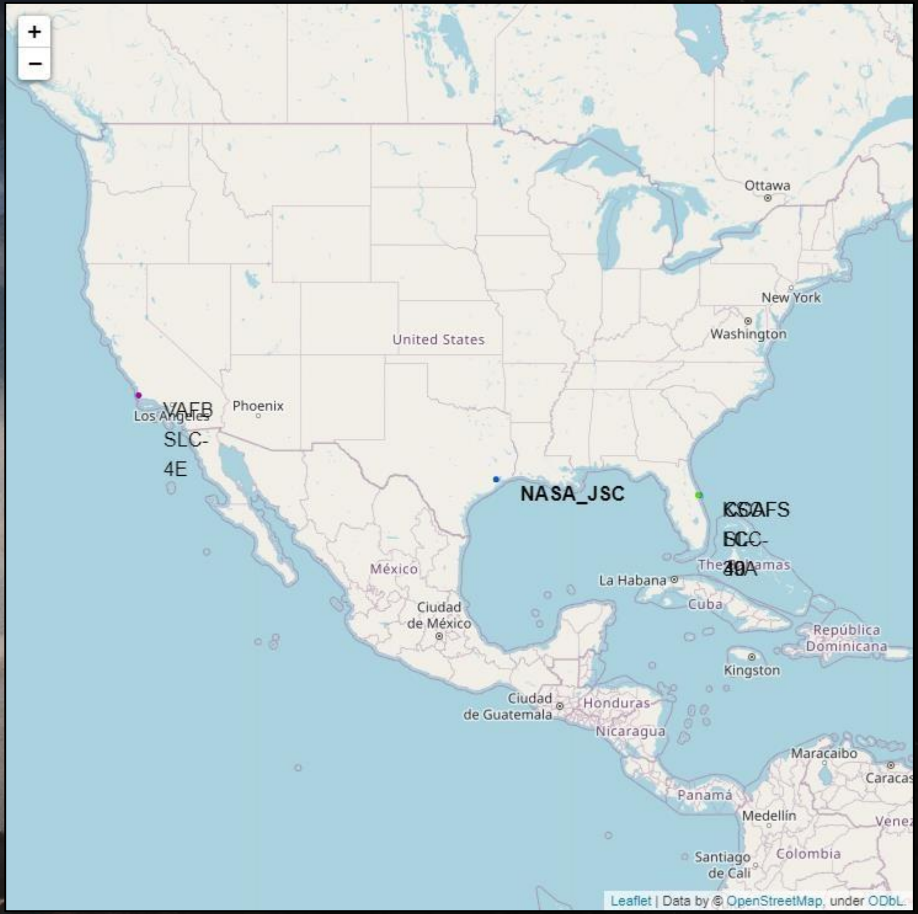
- **Explanation:** The most common landing outcome was ‘No attempt’.

Section 2

Launch Sites Proximities Analysis

Falcon 9 Launch Site Locations

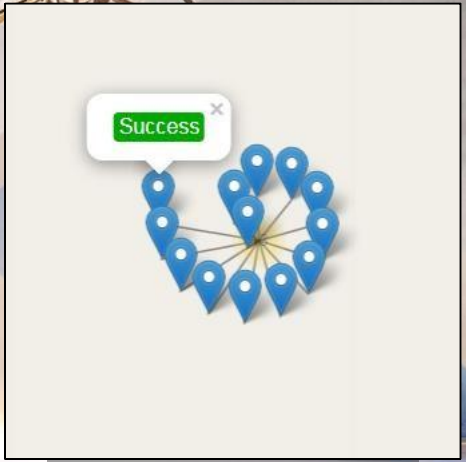
- California, USA
  - VAFB SLC-4E | Vandenberg Air Force Base Space Launch Complex 4E
- Florida, USA
  - KSC LC-39A | Kennedy Space Center Launch Complex 39A
  - CCAFS LC-40 | Cape Canaveral Air Force Station Launch Complex 40
  - CCAFS SLC-40 | Cape Canaveral Air Force Station Space Launch Complex 40
  - *Note: CCAFS LC-40 and CCAFS SLC-40 in the data refer to the same launch site



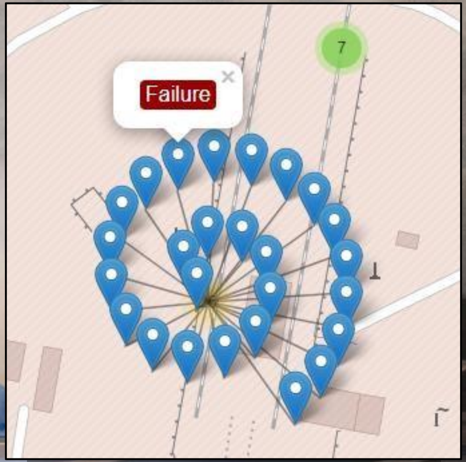
Map Markers of Success/Failed Landings

- The markers display the mission outcomes (Success/Failure) for Falcon 9 first stage landings. They are grouped on the map to be associated with the geographical coordinates for the launch site.
- A sense of a launch site's success rate for Falcon 9 first stage landings can be gleaned from the relative number of green success markers to red failure markers.

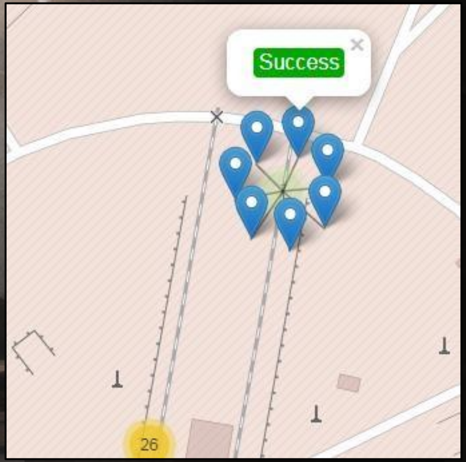
VAFB SLC-4E



KSC LC-39A



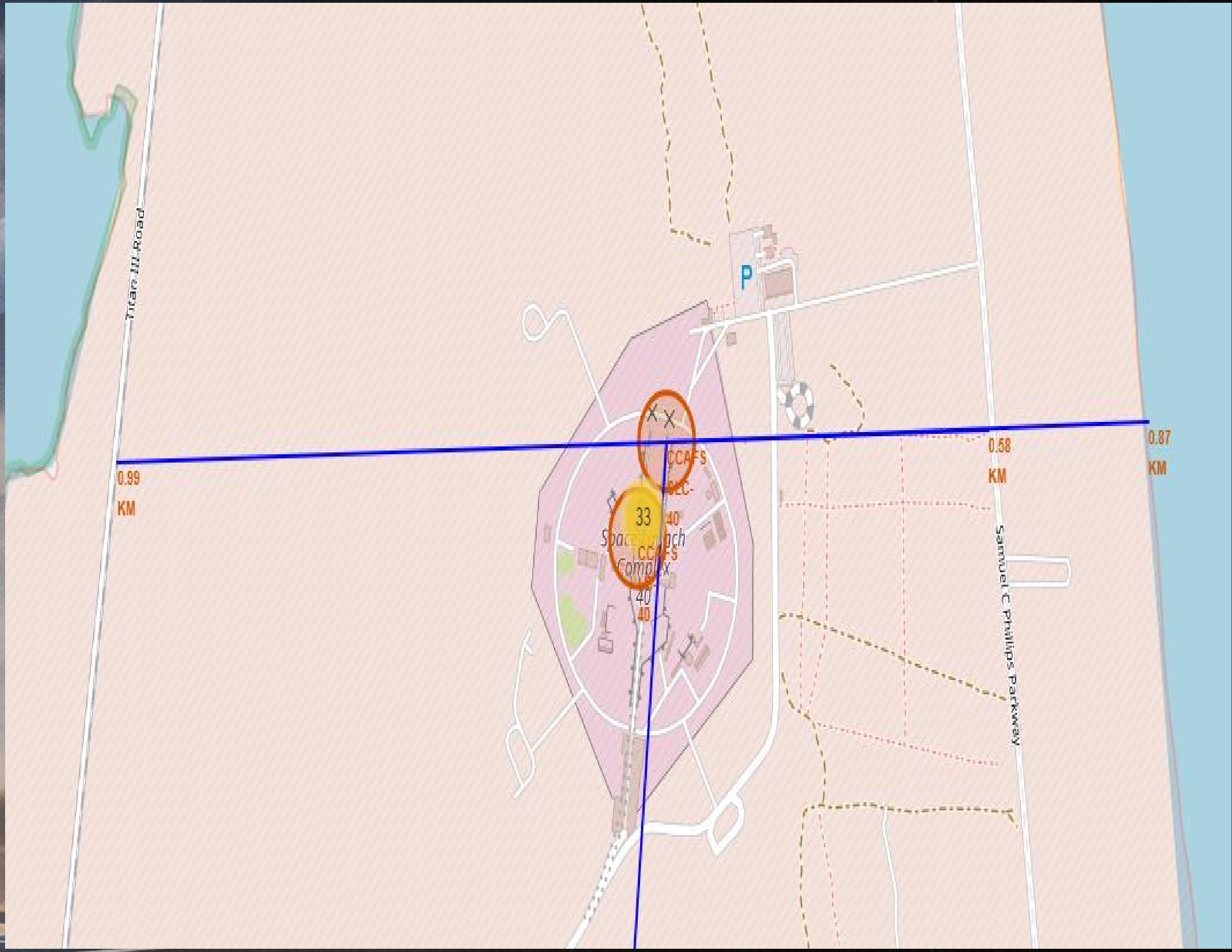
CCAFS LC-40



CCAFS SLC-40

Distance from Launch Site to Proximities

- The CCAFS LC-40 and CCAFS SLC-40 launch sites have coordinates that are close to being.
- Cocoa Beach: Located just south of Cape Canaveral, Cape Canaveral itself encompasses the launch complex.
- The Samuel C. Phillips Parkway is one of the most traveled and important roads on CCSFS
- The coastline is 0.87 km away from CCAFS LC-40.
- The rail line is around 1km away from CCAFS LC-40.



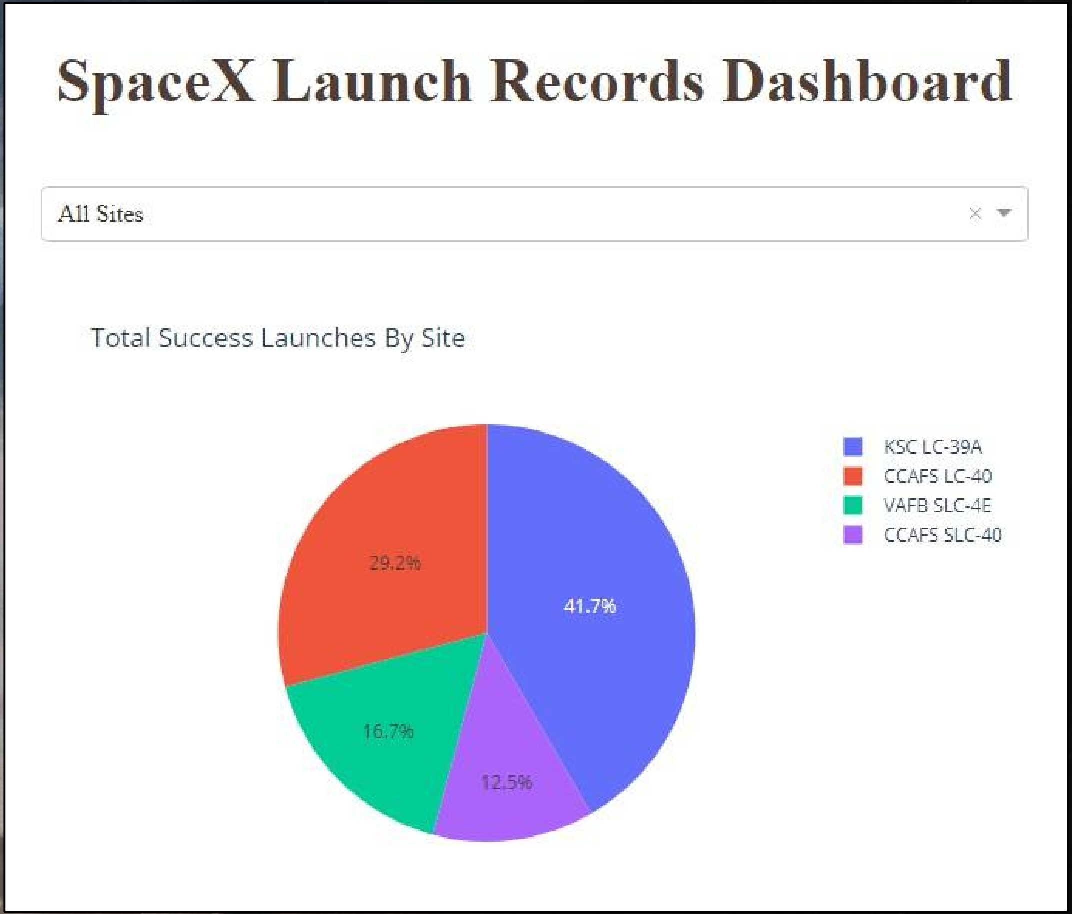
Section 3

Build a Dashboard with Plotly Dash

Launch Success Count for All Sites

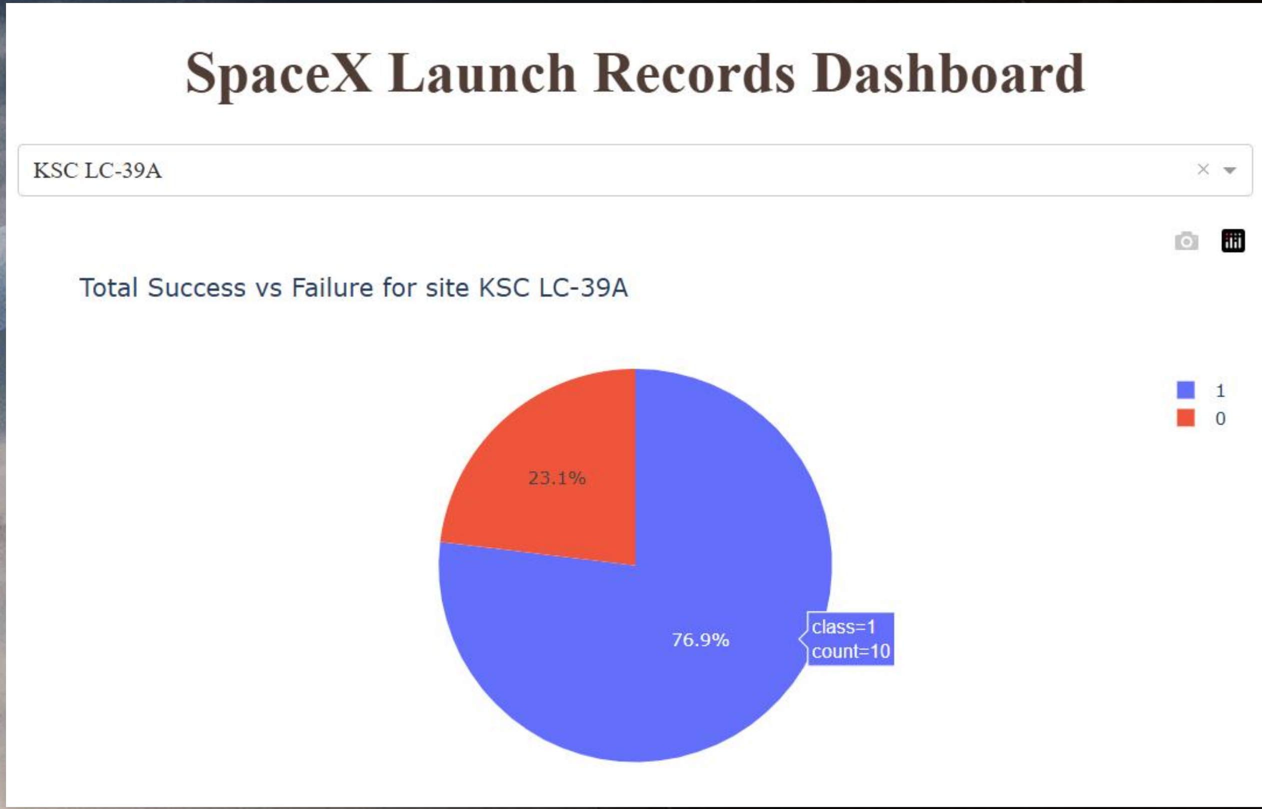
Insights

- KSC LC-39A leads with the highest success rate of launches (nearly 42%).
- CCAFS LC-40 also plays a major role, accounting for about 29%.
- The VAFB SLC-4E and CCAFS SLC-40 sites contribute smaller but still notable portions.



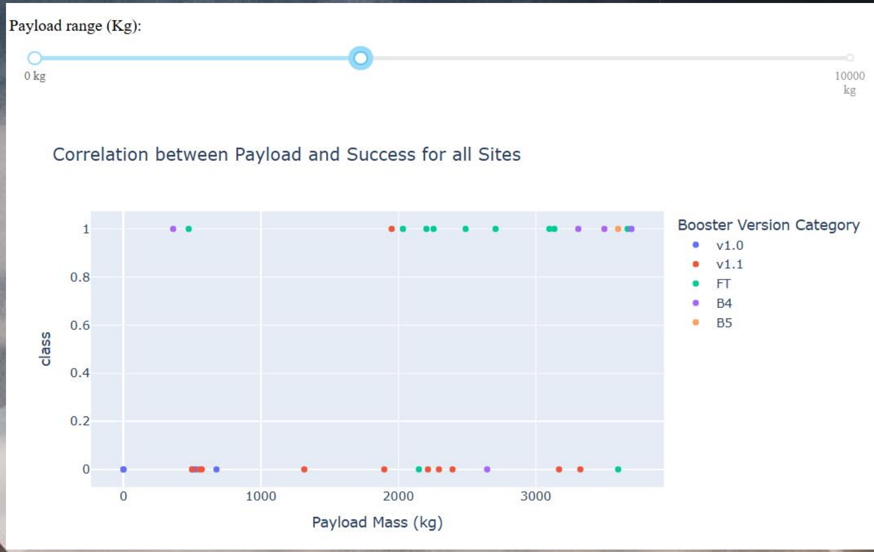
Launch Site with Highest Launch Success Ratio

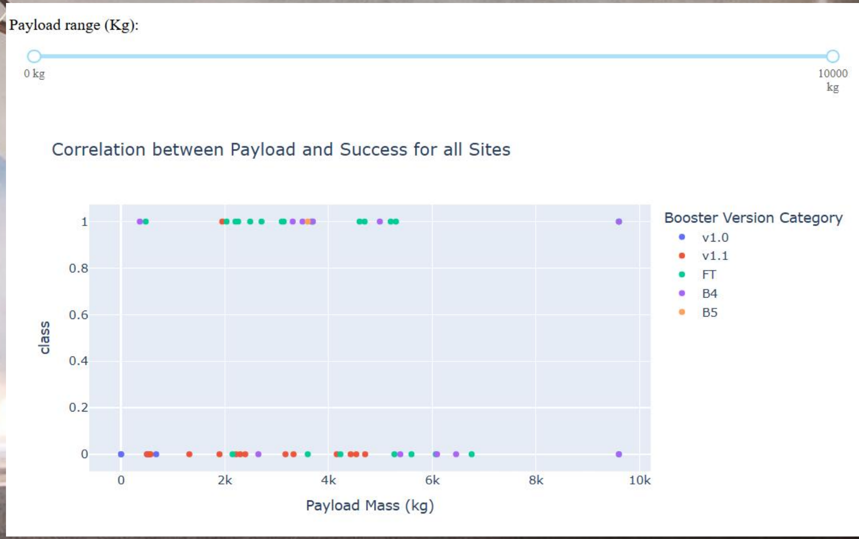
- KSC LC-39A has a much higher success rate (76.9%).
- This suggests LC-39A is a more reliable launch site.
- Indicates greater reliability and possibly newer infrastructure or more critical missions.
- KSC LC-39A may be the preferred site for high-stakes or advanced missions.



Payload vs. Launch Outcome

- These screenshots are of the Payload vs. Launch Outcome scatter plots for all sites, with different payload selected in the range slider.
- The payload range from about 2,000 kg to 5,000 kg has the largest success rate.
- The 'FT' booster version category has the largest success rate.





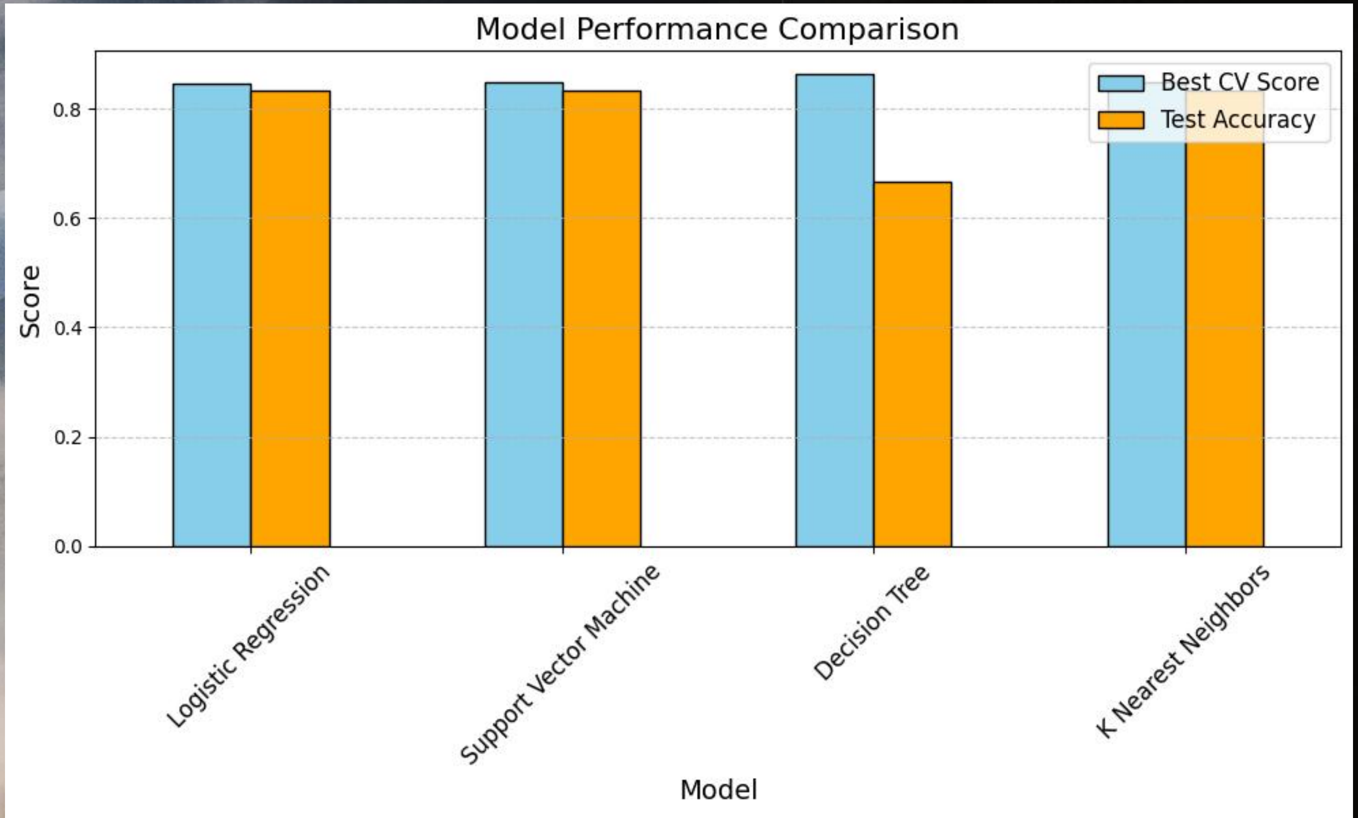
Section 4

Predictive Analysis (Classification)

Classification Accuracy

- ✅ Best Performing Model (Based on Confusion Matrix) is shared by Logistic Regression, SVM, and KNN.

- ✅ Best Performing Model Based on CV Score is:
- 🎯 Decision Tree with a CV Score of 0.875000



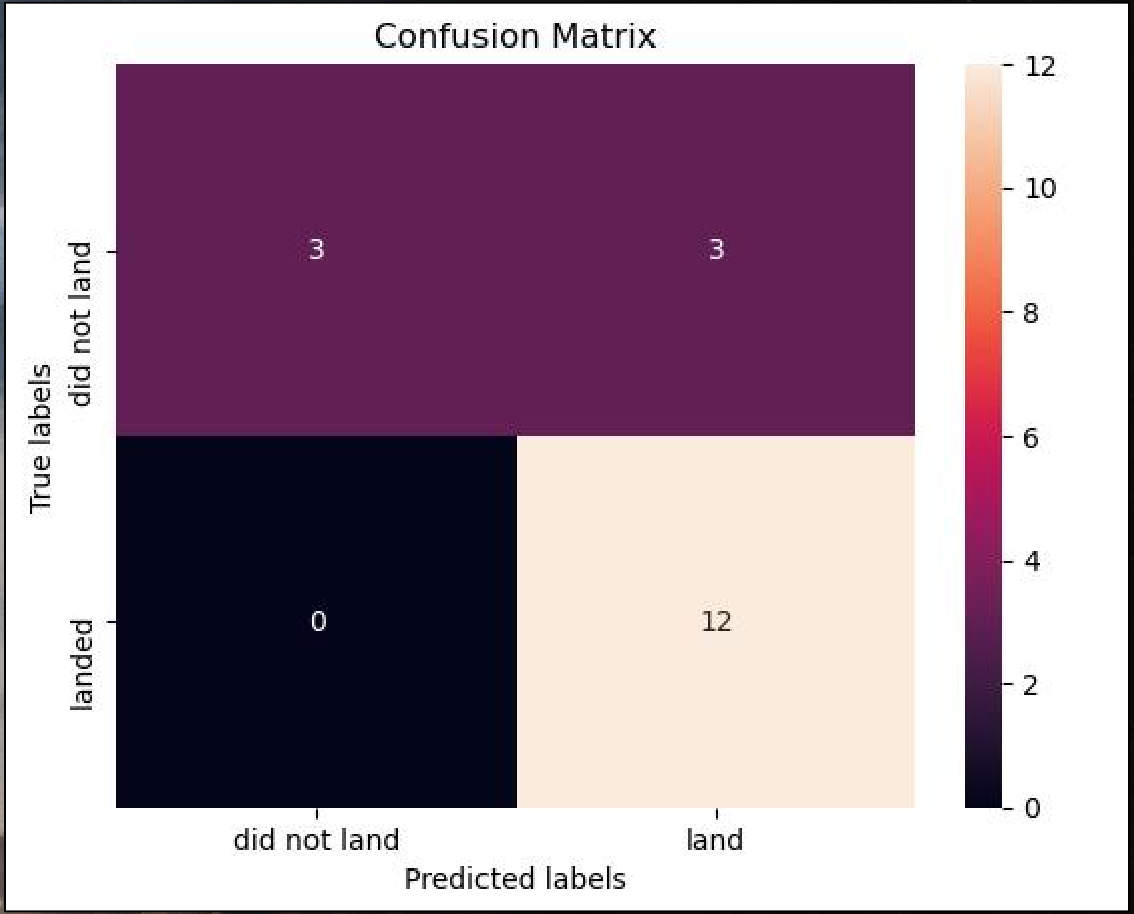
Confusion Matrix

From the confusion matrices the best confusion matrix is shared by:

✓ **Logistic Regression, SVM, and KNN.**

- High precision (80%)
- F1 Score: 88.9%
- Test Accuracy: 83.33%

⊖ Decision Tree, despite a better test accuracy (88.89%), had lower F1 score and more misclassifications in the matrix:

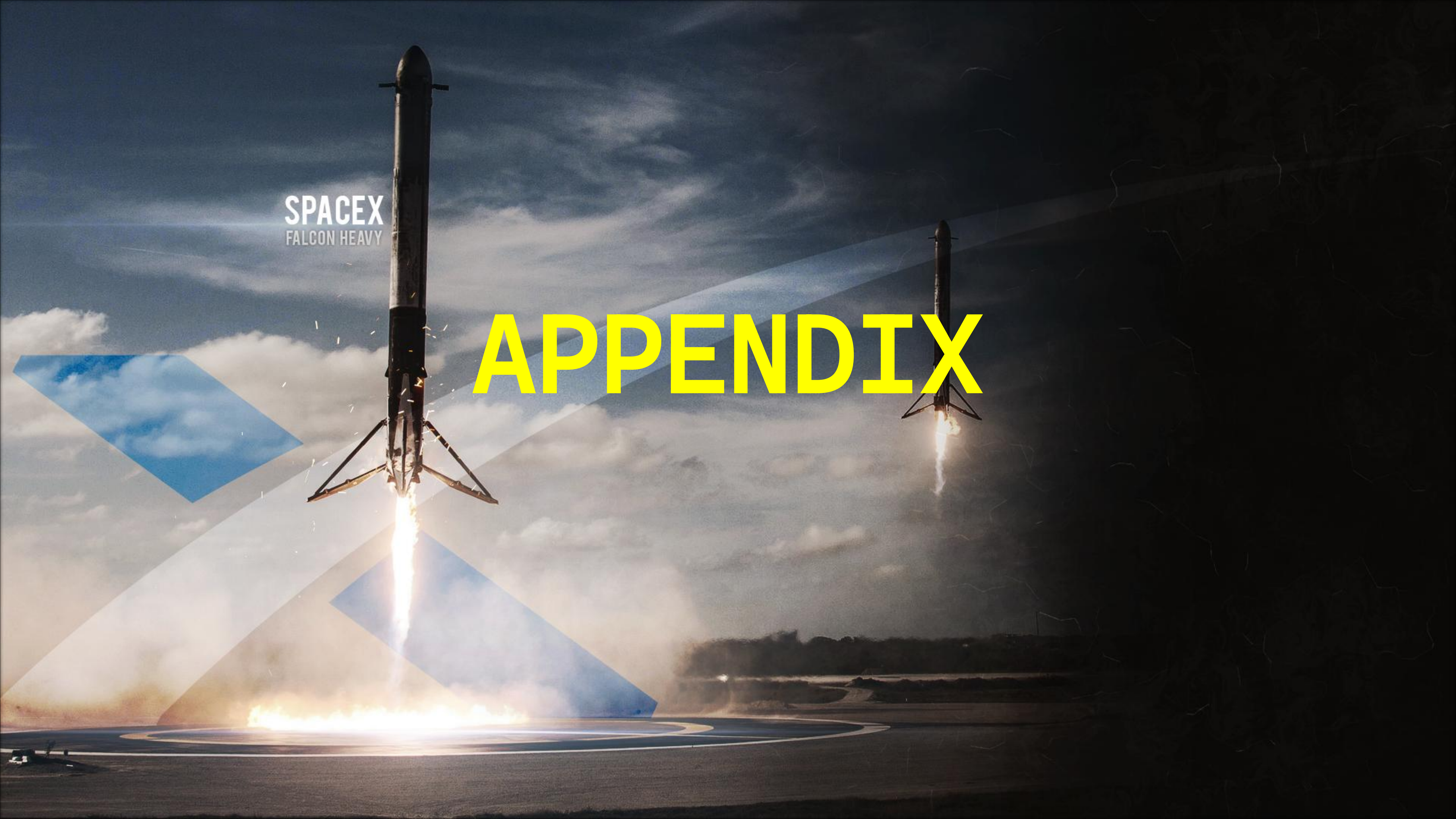


- Prediction Breakdown:
- 12 True Positives and 3 True Negatives
- 3 False Positives and 0 False Negatives

CONCLUSION

- Logistic Regression, SVM, and KNN models all achieved an accuracy of 83.3%, indicating strong and consistent performance across these algorithms.
- Payload mass, booster version, and launch site strongly influenced landing success.
- Certain booster versions and pads yielded higher reliability
- Certain orbits and booster versions correlate with better outcomes
- Launch site selection plays a crucial role in mission outcomes. The site KSC LC-39A had a 76.9% success rate and 23.1% failure rate.

APPENDIX



Initial Data Sources

- **SpaceX API (JSON):** https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json

- **Wikipedia (Webpage):** https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

- **SpaceX (CSV):** https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/labs/module_2/data/Spacex.csv?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-_Channel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2022-01-01

- **Launch Geo (CSV):** https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_geo.csv

- **Launch Dash (CSV):** https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_dash.csv

Processed Data Sets (CSV files)

- GitHub URL Dataset 1 : dataset_part_1
- GitHub URL (Web Scraped): spacex_web_scraped.csv
- GitHub URL (CSV 2): dataset_part_2.csv
- GitHub URL (CSV 3): dataset_part_3.csv
- GitHub URL (Launch Geo): spacex_launch_geo.csv
- GitHub URL (Launch Dash): spacex_launch_dash.csv

Jupyter Notebooks and Plotly Dashboard File

- Data Collection Notebook
- Web Scraping Notebook
- Data Wrangling Notebook
- EDA SQL Notebook
- EDA Visualization Notebook
- Folium Map Notebook
- Plotly Dash App
- Machine Learning Notebook


Thank You