IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Akrem Zaghdoudi>
<11/08/2024>

Outline

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

Executive Summary

- Summary of methodologies :

- Data Collection through API and Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL and Pandas
- Exploratory Data Analysis with Data Visualization (Matplotlib)
- Interactive Visual Analytics with Folium
- Machine Learning Prediction (Classification)

- Summary of all results :

- Exploratory Data Analysis helped to choose the best features affecting the landing success.
- Interactive analytics in screenshots that provided helpful insights.
- Predictive Analytics showed the best model to predict .

Introduction

- **Project background and context**

- SpaceX has revolutionized the space industry by significantly reducing launch costs, primarily due to the reusability of the Falcon 9 rocket's first stage. While a typical Falcon 9 launch costs $62 million, competitors' prices can soar to $165 million per launch. The ability to predict whether the Falcon 9's first stage will land successfully is crucial for estimating the overall cost of a launch. This prediction model can also provide valuable insights for other companies, like a hypothetical competitor Space Y, seeking to bid against SpaceX for rocket launches

- **Problems We want to find answers :**

- How can we accurately estimate the total cost of a launch ?
- Where could be the best location to make those launchs ?
- What are the factors that play a significant role in determining whether the Falcon 9 first stage will land successfully ?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from two sources : Space X API and WebScrapping (Wikipedia)
- Perform data wrangling
  - Collected data was enhanced by creating a landing outcome labelbased on the outcome data, after summarizing and analyzing the relevant features
- Perform exploratory data analysis (EDA) using Data visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Collected data was normalized and split into training and test datasets, then evaluated with four different classification models, with each model's accuracy .

Data Collection

- The data sets were collected from: :

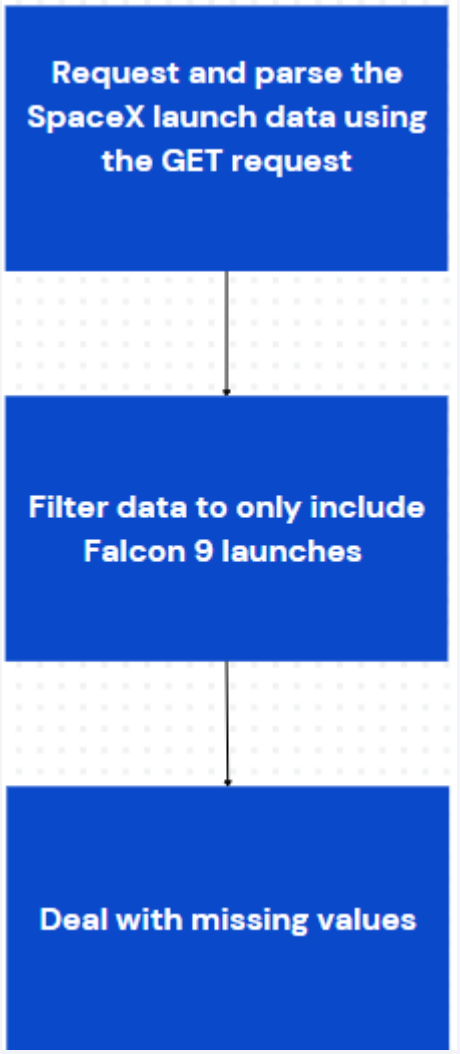
  - The SpaceX API , the data collection was done using get request method .Next, we decoded the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize().

  - A permanently linked Wikipedia page with launch data in HTML tables which we performed web scraping from with BeautifulSoup. We extracted the records then parse the table and convert it to a pandas dataframe for future analysis

- We then cleaned the data, checked for missing values and fill in missing values where necessary.

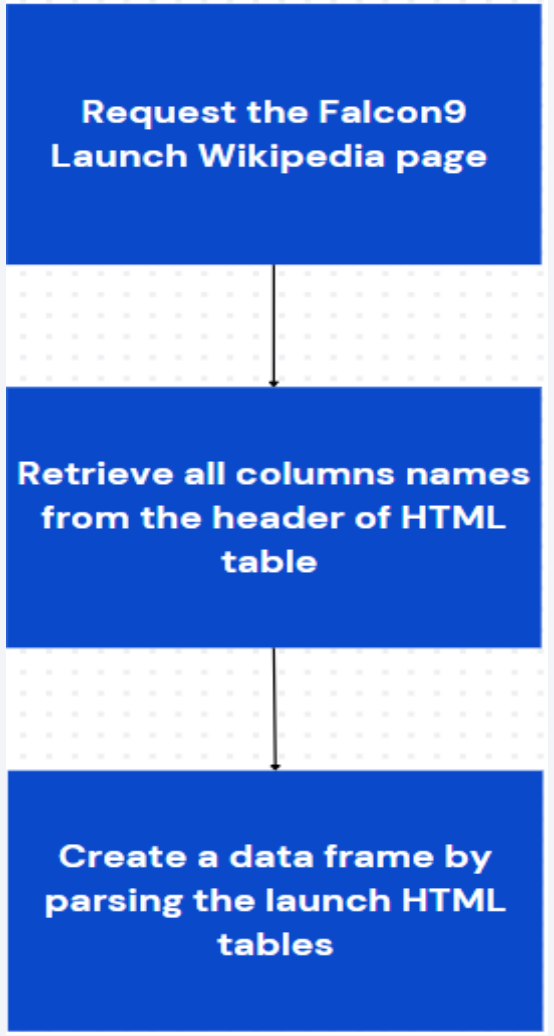
Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used; (« https://api.spacexdata.com/v4/rockets/ »)
- GitHub URL of the completed notebook: https://github.com/zaghdoudiakrem/Data_science_capstone/blob/main/1-Collecting_the_data_from_SpaceX_API.ipynb



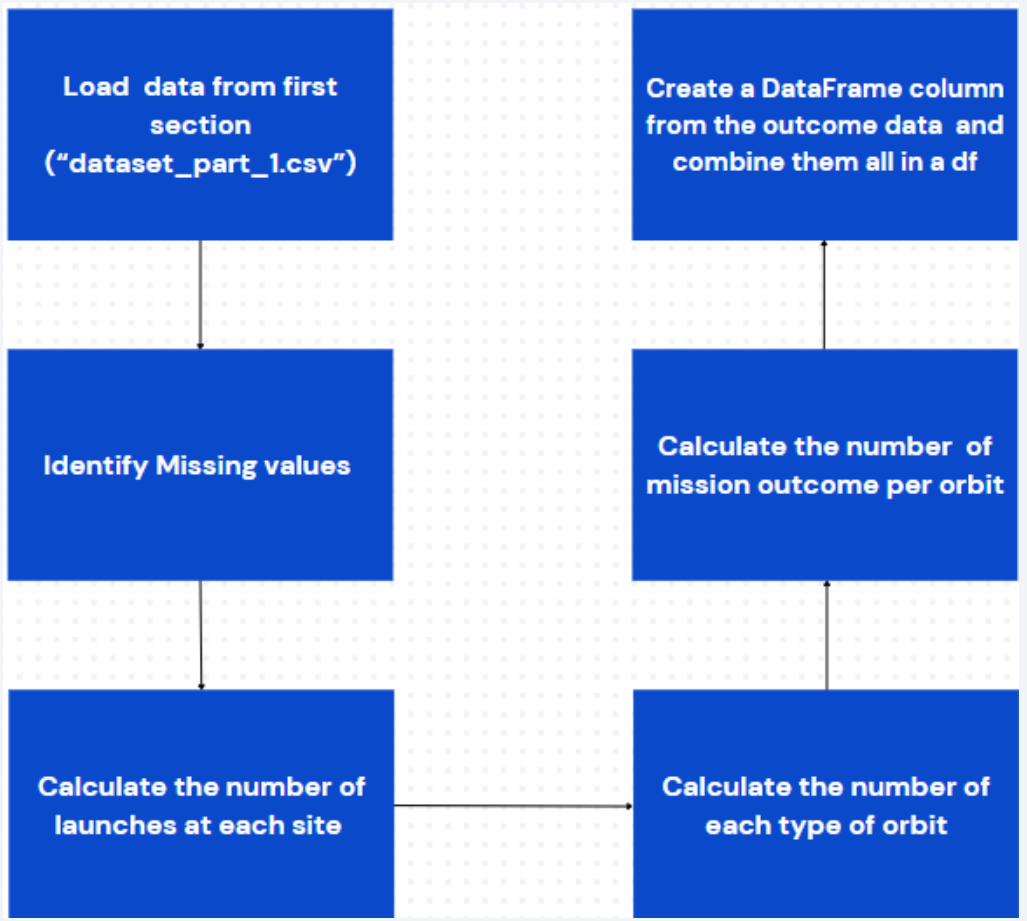
Data Collection - Scraping

- Data was scraped from Wikipedia(«https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922»)
- GitHub URL of the completed notebook: https://github.com/zaghdoudiakrem/Data_science_capstone/blob/main/2-Web_scraping_from_Wiki.ipynb



Data Wrangling

- We performed EDA to identify patterns in the data , analyze the dataset and establish the appropriate training labels
- GitHub URL of the completed notebook:
- https://github.com/zaghdoudiakrem/Data_science_capstone/blob/main/3-Data_wrangling.ipynb



EDA with Data Visualization

- Many charts were plotted we used them to analyze :
- **Launch Site Trends:**
  - Scatterplot illustrating the relationship between mission outcome, Launch Site, and Flight Number.
  - Scatterplot showing the relationship between mission outcome, Launch Site, and Payload.
- **Orbit Type Trends**
  - Bar chart depicting the relationship between mission outcome and Orbit Type.
  - Scatterplot displaying the relationship between mission outcome, Orbit Type, and Flight Number.
  - Scatterplot revealing the relationship between mission outcome, Orbit Type, and Payload.
- **Temporal Trends**
  - Line plot demonstrating the trend of mission outcomes over the years.

GitHub URL of the completed notebook:
https://github.com/zaghdoudiakrem/Data_science_capstone/blob/main/5-Data_Visualization.ipynb

EDA with SQL

- **We loaded the Space X dataset into the corresponding table in a Db2 database**
- **We executed SQL queries to get answers :**
- Identify the unique names of launch sites in the space mission dataset.
- Determine the top 5 launch sites with names beginning with 'CCA'.
- Calculate the total payload mass carried by boosters launched by NASA (CRS).
- Find the average payload mass for the booster version F9 v1.1.
- Retrieve the date of the first successful landing outcome on a ground pad.
- List the boosters that have successfully landed on a drone ship with a payload mass between 4000 and 6000 kg.
- Count the total number of successful and failed mission outcomes.
- Identify the booster versions that have carried the maximum payload mass.
- Analyze failed landing outcomes on a drone ship in 2015, including booster versions and launch site names. Finally, rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20,
- GitHub URL of the completed notebook:
- https://github.com/zaghdoudiakrem/Data_science_capstone/blob/main/4-EDA_with_sql.ipynb

Build an Interactive Map with Folium

- **Summary of Map Objects Added to the Folium Map and their purpose :**
- Markers: Identified each launch site and the NASA Johnson Space Center.
- Purpose: Pinpointed specific locations of launch sites and major centers like NASA Johnson Space Center.
- Circles: Highlighted areas around each launch site.
- Purpose: Emphasized the vicinity of launch sites to provide spatial context.
- Lines: Displayed distances from CCAFS LC-40 to nearby features such as the coastline, rail line, and perimeter road.
- Purpose: Measured and visualized distances to assess logistics and accessibility.
- Marker Clusters: Grouped launch events with color coding to represent success rates.
- Purpose: Facilitated identification of sites with higher success rates and understanding of launch patterns.

- GitHub URL of the completed notebook: https://github.com/zaghdoudiakrem/Data_science_capstone/blob/main/6-Launch_site_location.ipynb

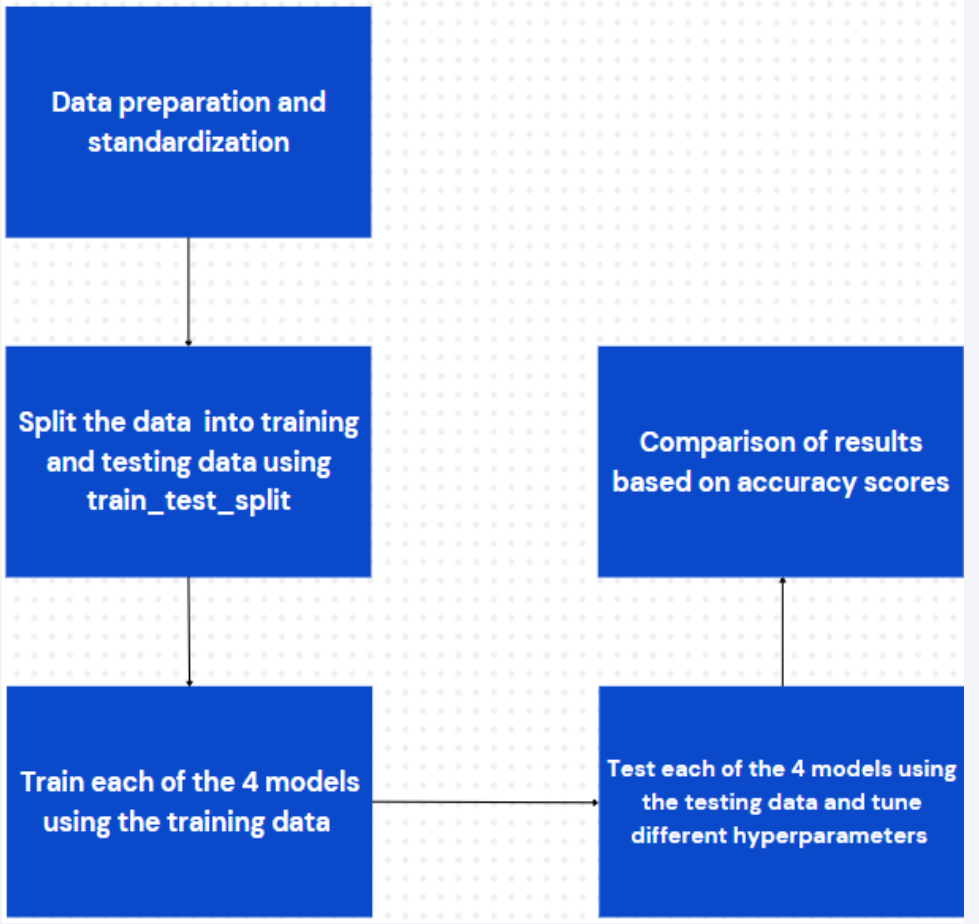
Build a Dashboard with Plotly Dash

- **Summary of Plots/Graphs and Interactions Added to the Dashboard:**
- **Pie Charts:**
  - **Description:** Displayed the distribution of total launches by site and provided a visual breakdown of launches at different sites, allowing users to compare the volume of launches across sites.
- **Scatter Plots:**
  - **Description:** Illustrated the relationship between mission outcomes and payload mass for different booster versions and enabled analysis of how payload mass correlates with launch outcomes across various booster versions.
- **Input Dropdown:**
  - **Description :** Allowed users to select one or all launch sites for the pie chart and scatterplot.
- **Input Slider:**
  - **Description :** Filtered payload masses for the scatterplot and enabled users to focus on specific ranges of payload masses, enhancing the analysis of payload impact on mission outcomes.

GitHub URL of the completed script:
https://github.com/zaghdoudiakrem/Data_science_capstone/blob/main/spacex_dash_app.py

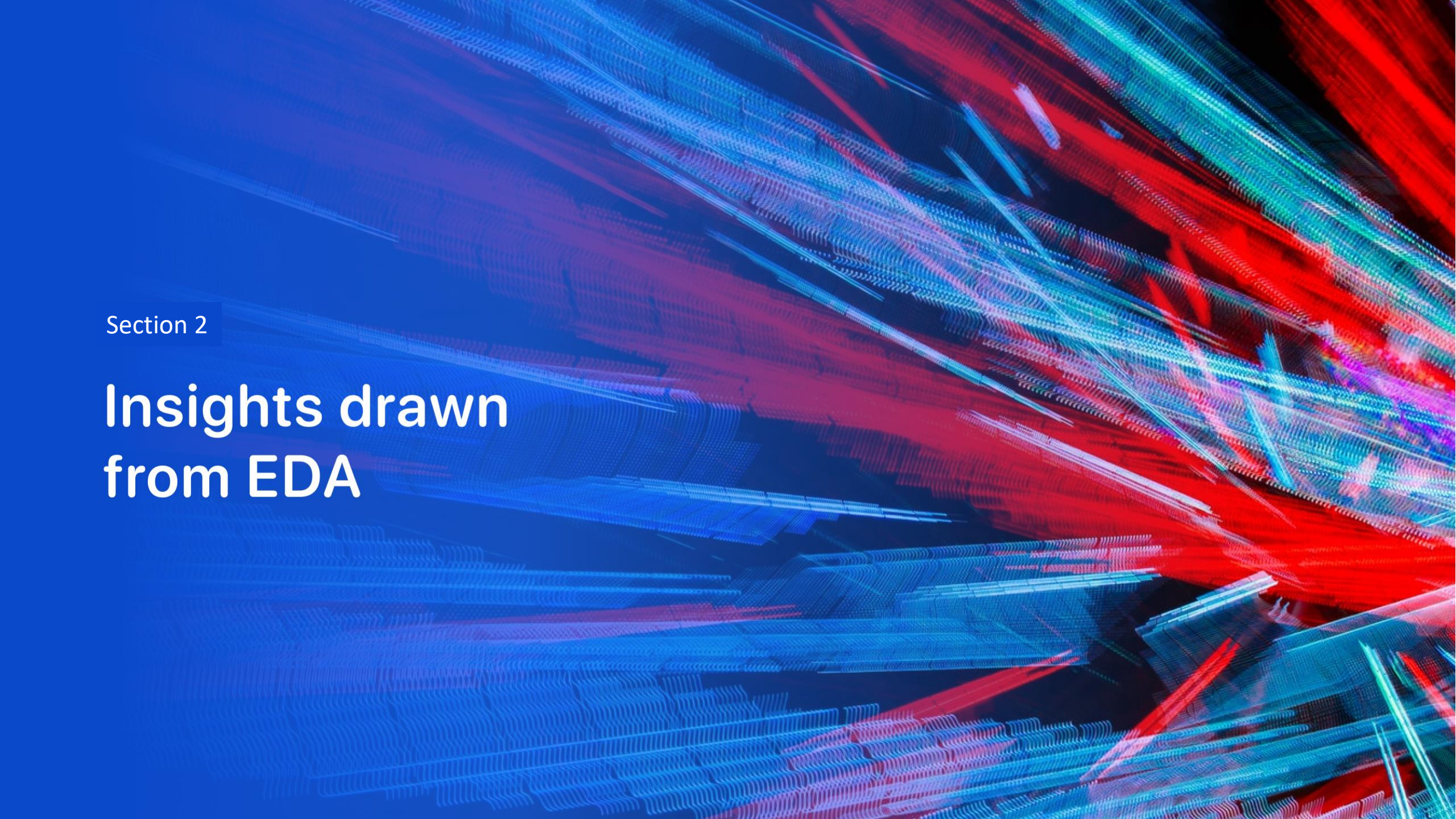
Predictive Analysis (Classification)

- Hyper-parameters were evaluated using GridSearchCV() and the best was selected using '.best_params
- Using the best hyper-parameters, each of the four models were scored on accuracy by using the testing data set.
- GitHub URL of the completed notebook:
- https://github.com/zaghdoudiakrem/Data_science_capstone/blob/main/7-SpaceX_Machine_Learning_Prediction_.ipynb



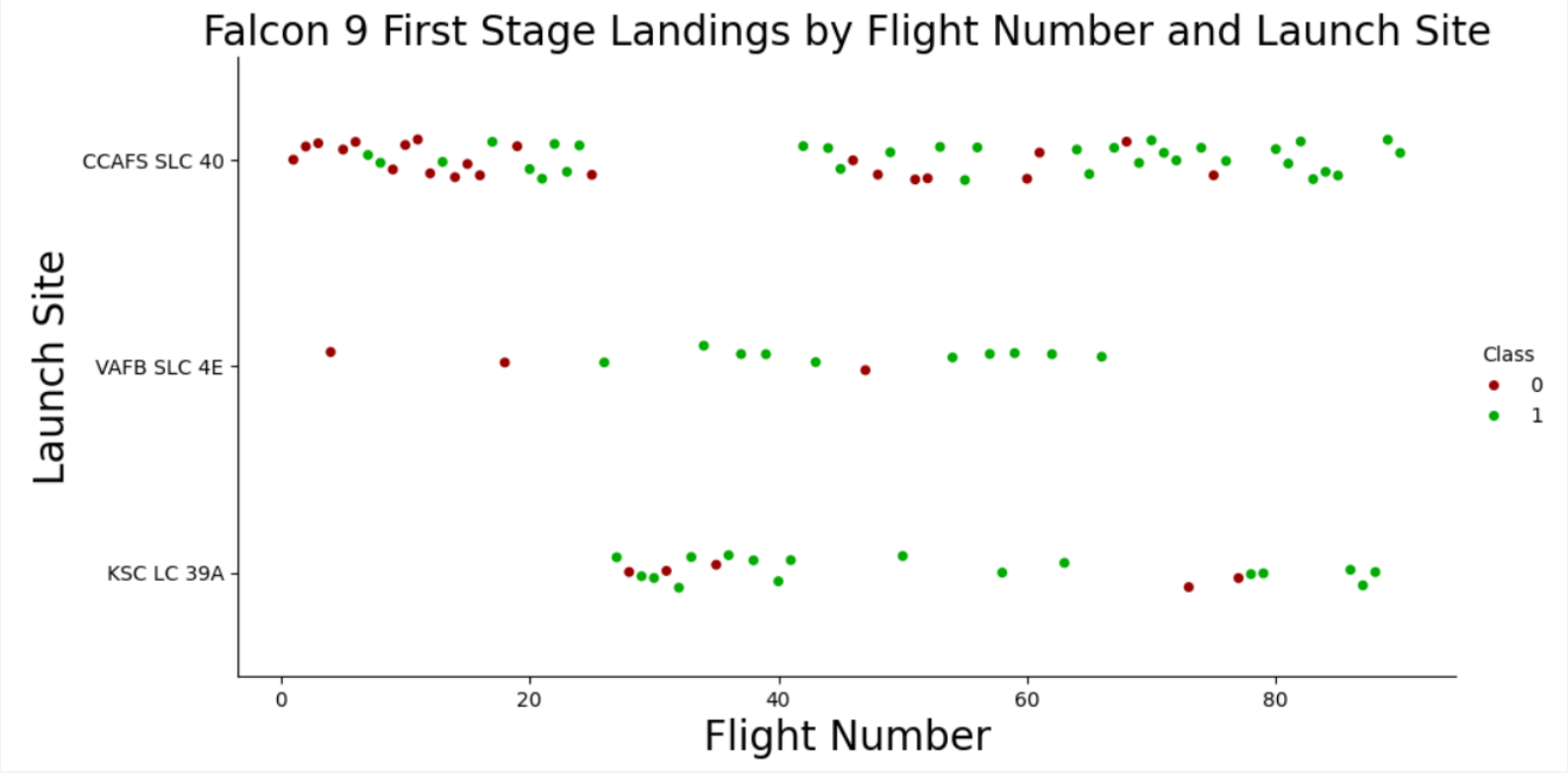
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

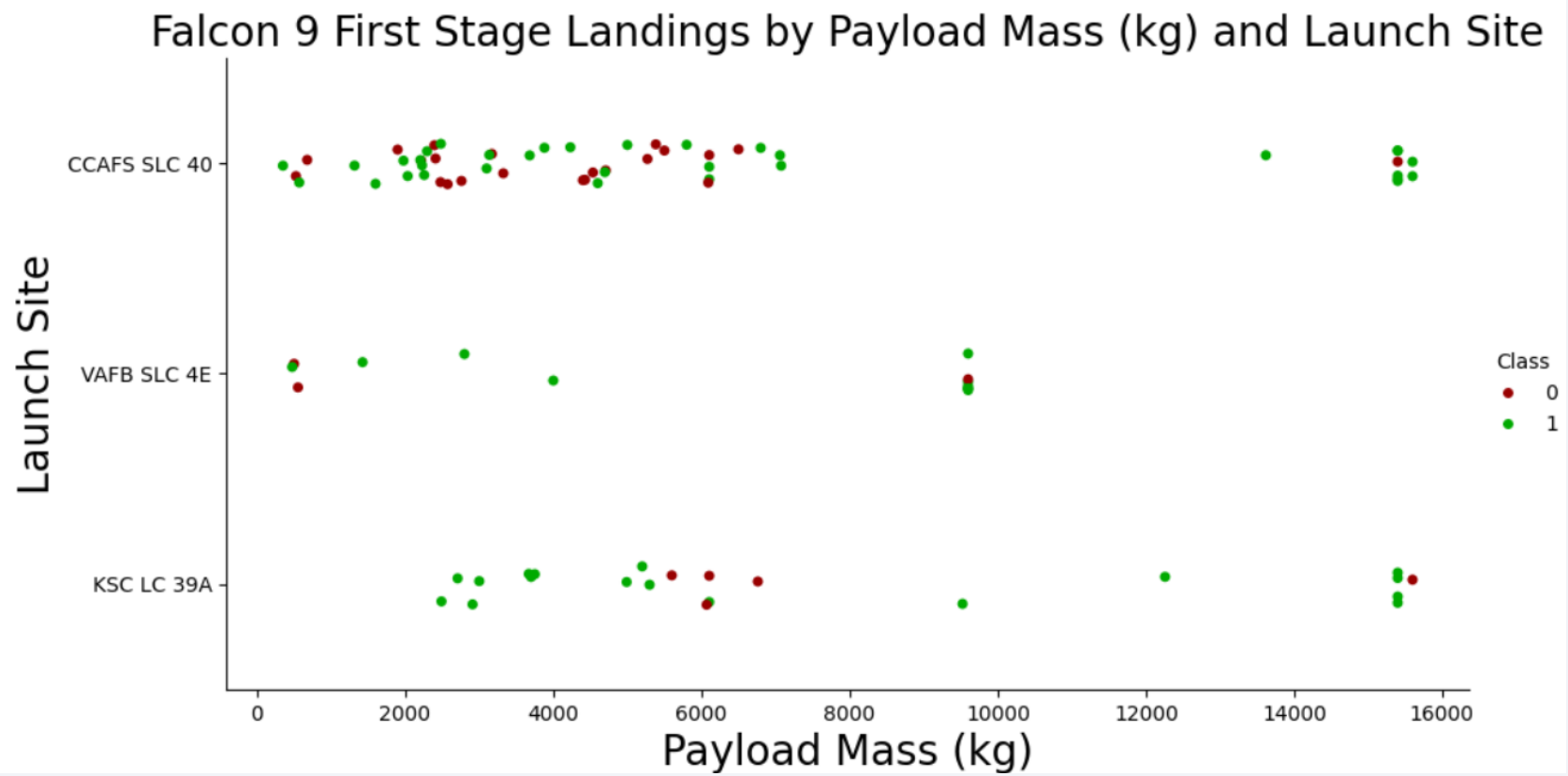
Flight Number vs. Launch Site

- CCAFS SLC 40 is currently the most successful launch site, with VAFB SLC 4E and KSCLC 39A following. The overall success rate has improved over time. While a higher payload mass at CCAFS SLC 40 is generally associated with a better success rate, the visualization does not clearly establish a direct link between launch site success and payload mass.



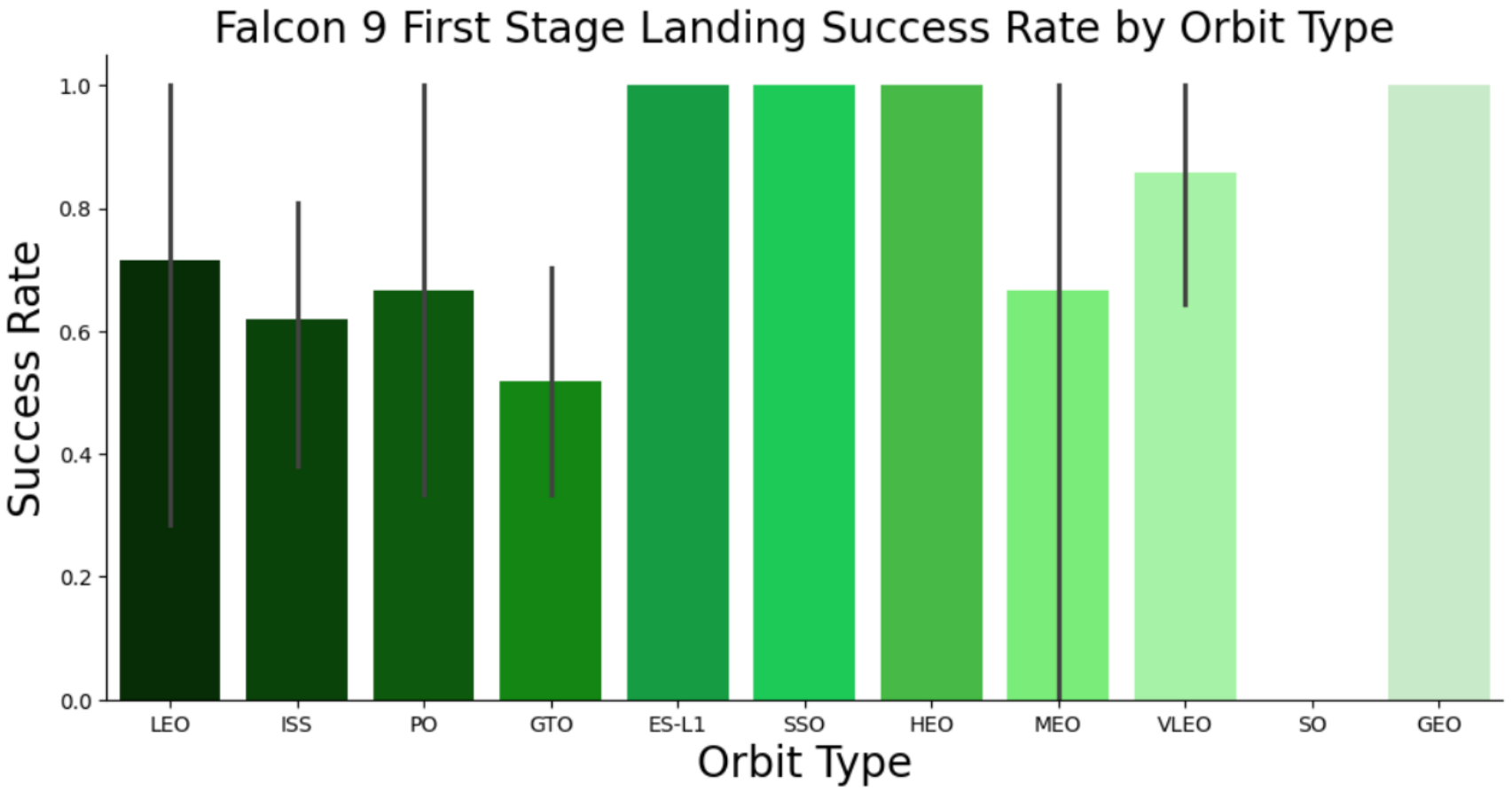
Payload vs. Launch Site

- The greater the payload mass for launch site CCAFS SLC 40 the higher the success rate for the rocket.



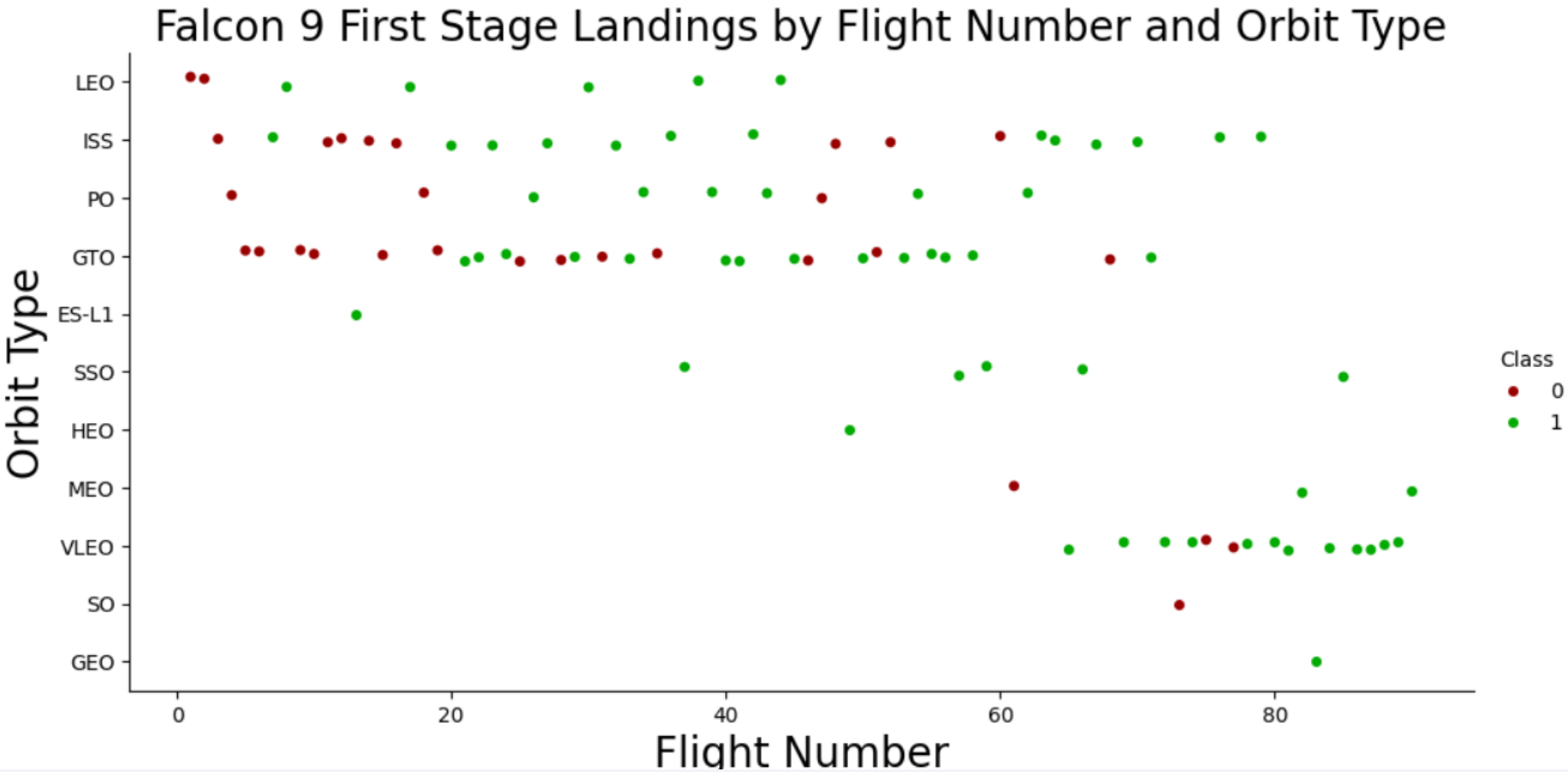
Success Rate vs. Orbit Type

The orbit types ES-L1, GEO, HEO, and SSO have the highest success rates, but it's important to consider the number of launches per orbit type when interpreting these results.



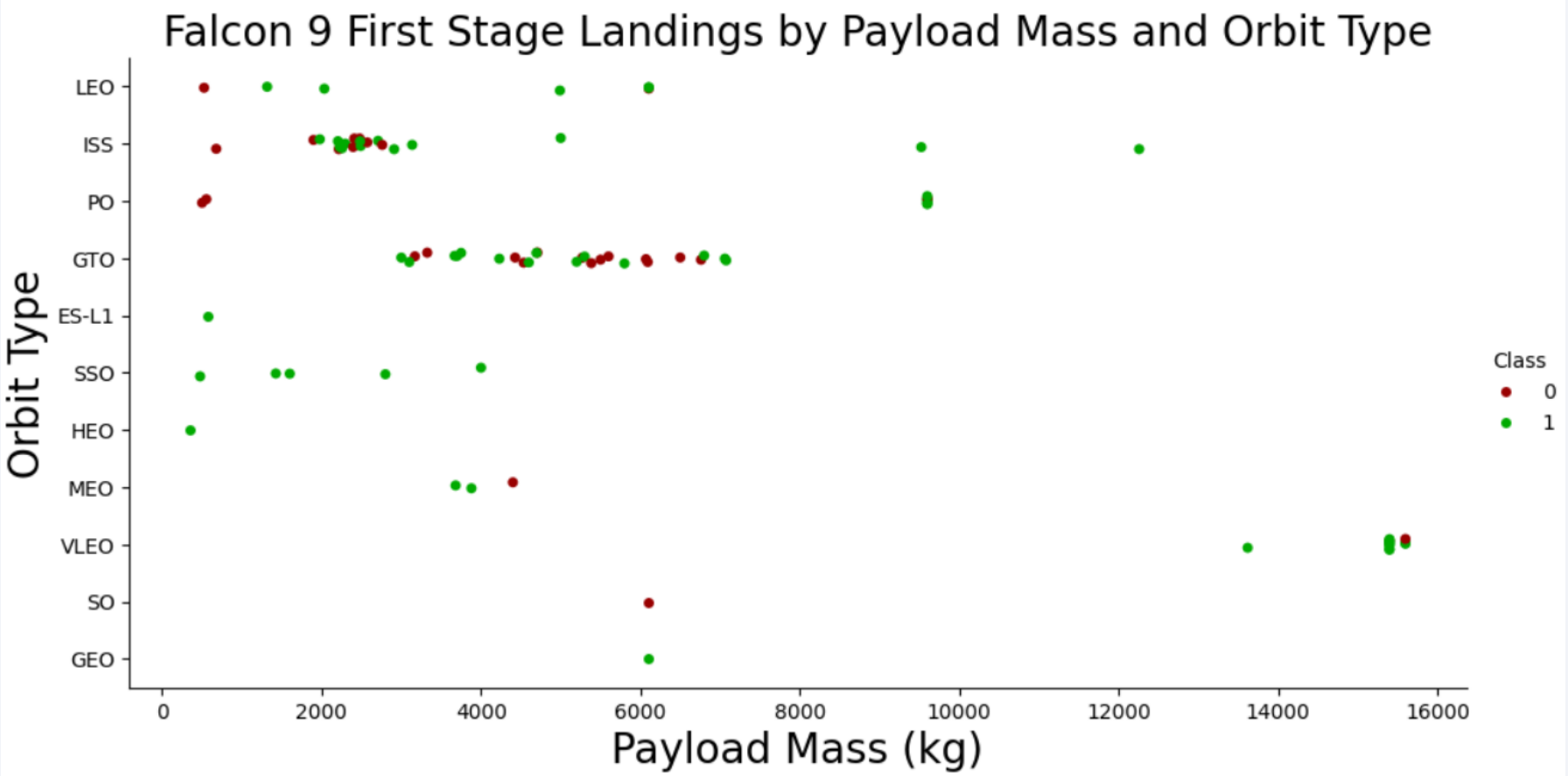
Flight Number vs. Orbit Type

In LEO orbits, success is linked to the number of flights, while in GTO orbits, there is no clear relationship between flight number and success.



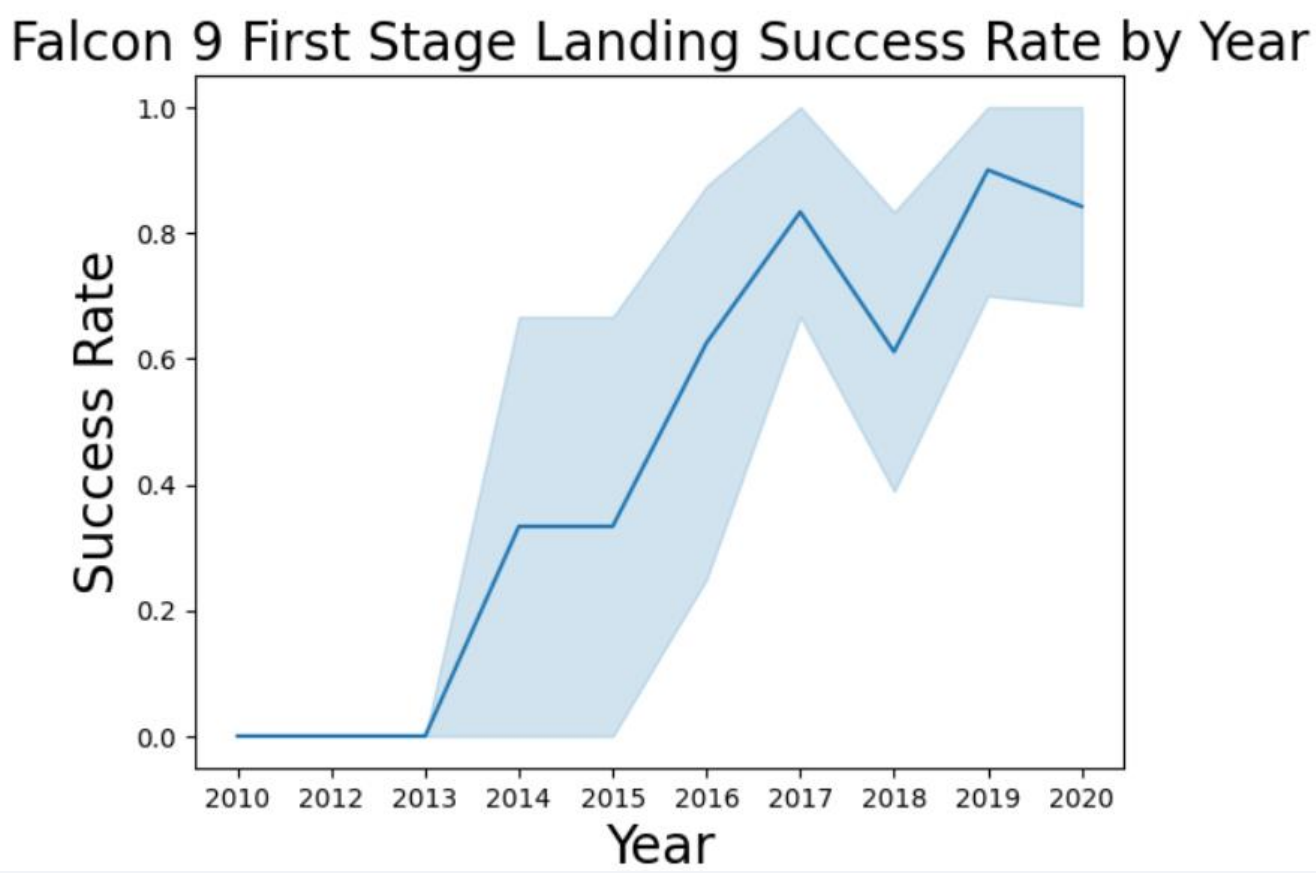
Payload vs. Orbit Type

• Some orbit types have better success rates than others . Success rate appears to have no obvious correlation with payload mass.



Launch Success Yearly Trend

Based on the plot , we can observe that success rate has increased since 2013 and kept on increasing till 2020



All Launch Site Names

- **Query:** SELECT DISTINCT LAUNCH_SITE FROM SPACEXTABLE;
- **Result:**

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

- **Explanation :** As its showen in the result there are 4 unique launch sites

Launch Site Names Begin with 'CCA'

- **Query :** SELECT * FROM SPACEXTABLE WHERE launch_site LIKE 'CCA%' LIMIT 5;
- **Result :**

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- **Explanation :** This is a simple method used to get an idea of the data stored in the database table.

Total Payload Mass

- **Query :** SELECT sum(payload_mass__kg_) AS "Total Payload Mass (kg)" FROM SPACEXTABLE WHERE customer LIKE '%NASA (CRS)%';
- **Result :**

Total Payload Mass (kg)
48213

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

Average Payload Mass by F9 v1.1

- **Query :** SELECT sum(payload_mass__kg_) / count(payload_mass__kg_) AS “Average Payload Mass (kg)" FROM SPACEXTABLE WHERE booster_version LIKE 'F9 v1.1';
- **Result :**

Average Payload Mass (kg)
2928

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

First Successful Ground Landing Date

- **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 ground pad landing occurred on December 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- **Query :** SELECT DISTINCT booster_version FROM SPACEXTABLE WHERE landing__outcome = 'Success (drone ship)' and payload_mass__kg_ BETWEEN 4000 and 6000;
- **Result :**

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

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

Total Number of Successful and Failure 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

- **Query :** SELECT booster_version, payload_mass__kg_ FROM SPACEXTABLE WHERE payload_mass__kg_ = (SELECT max(payload_mass__kg_) FROM SPACEXTABLE);
- **Result :**

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 dataset shows a maximum payload mass of 15,600 kg, which was carried by twelve separate Falcon 9 boosters.

2015 Launch Records

- **Query :** SELECT MONTHNAME(DATE) AS “Month”, landing__outcome, booster_version, launch_site FROM SPACEXTABLE WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;

- **Result :**

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

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

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

- **Query :** SELECT landing__outcome, count(landing__outcome) AS "Count" FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing__outcome ORDER BY count(landing__outcome) DESC;
- **Result :**

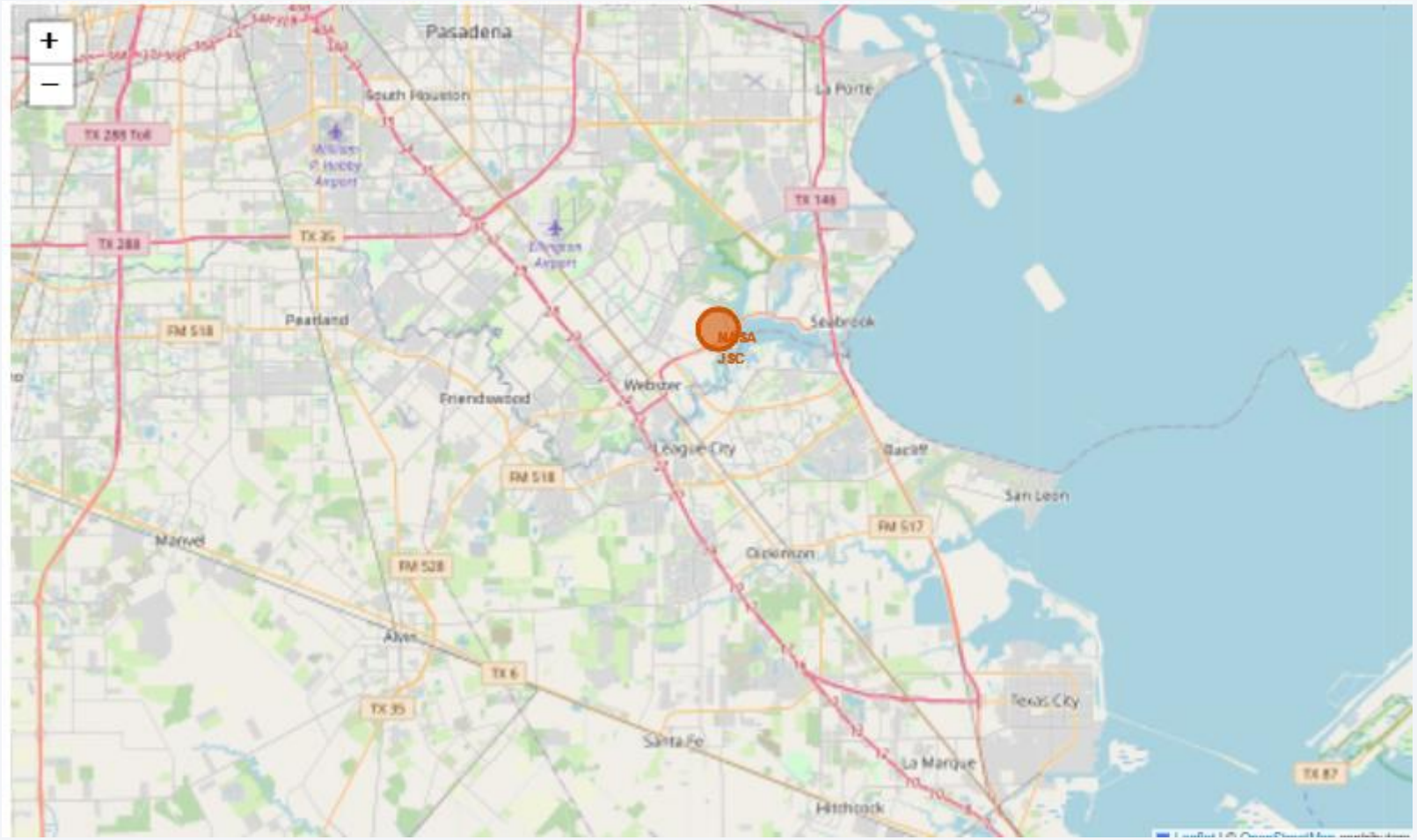
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- **Explanation :** The most common landing outcome was ‘not attempted’ (10).

Section 3

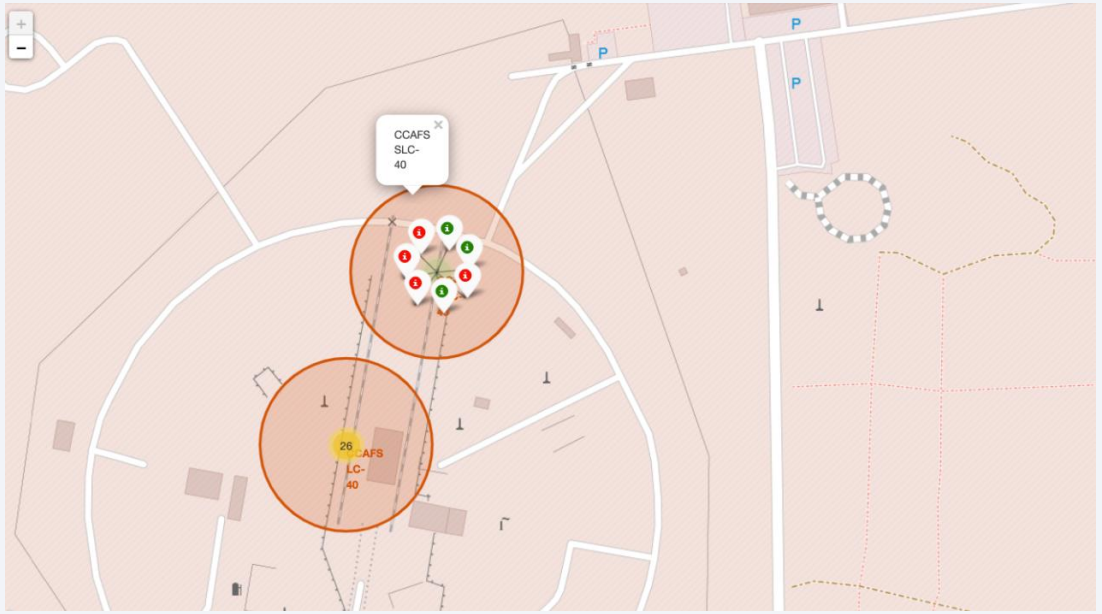
Launch Sites Proximities Analysis

<Falcon 9 Launch Site Locations>



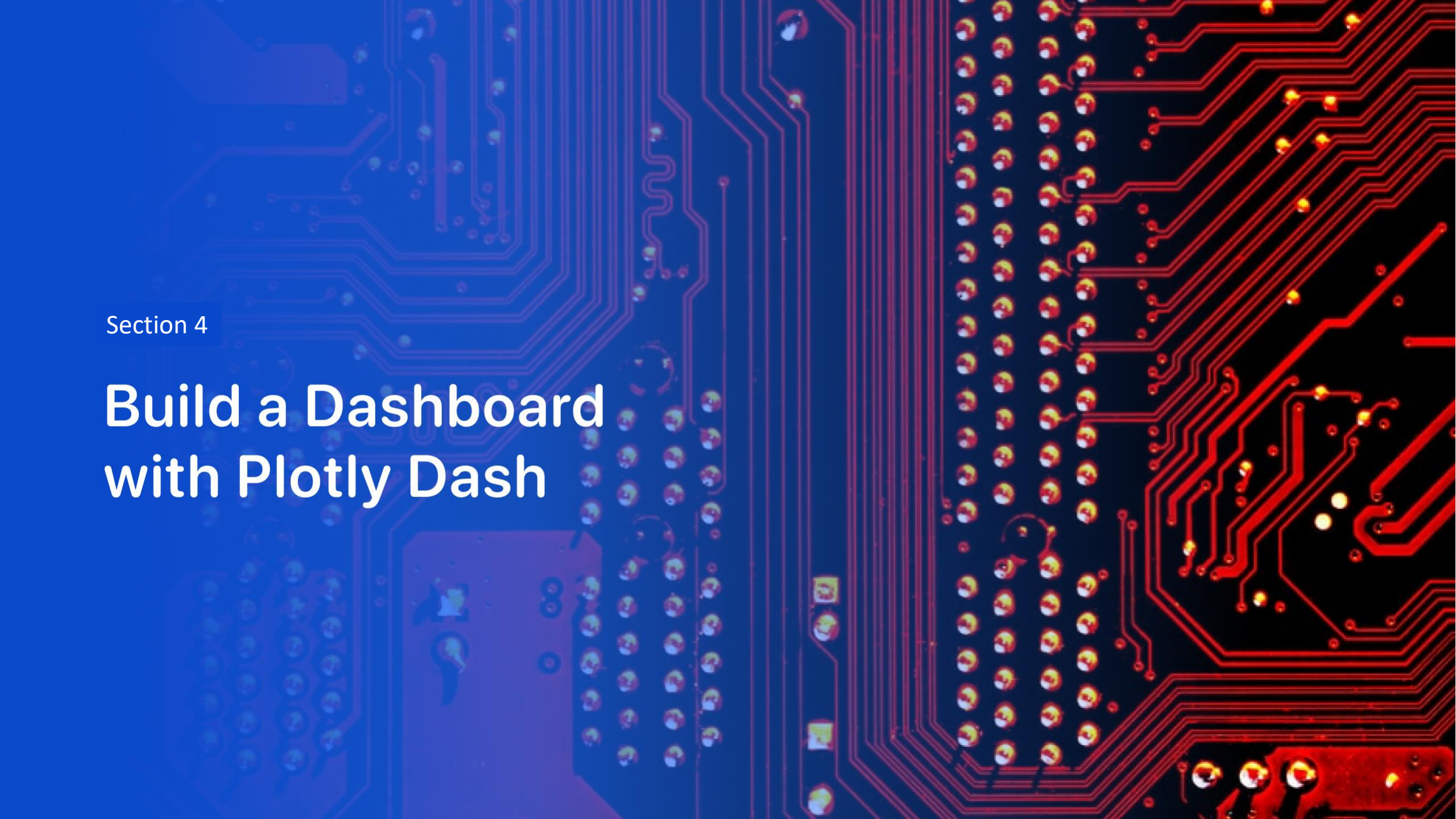
<Map Markers of Success/Failed Landings >

- The markers represent the mission outcomes (Success/Failure) for Falcon 9 first stage landings, organized on the map to correspond with the geographical coordinates of the launch sites.



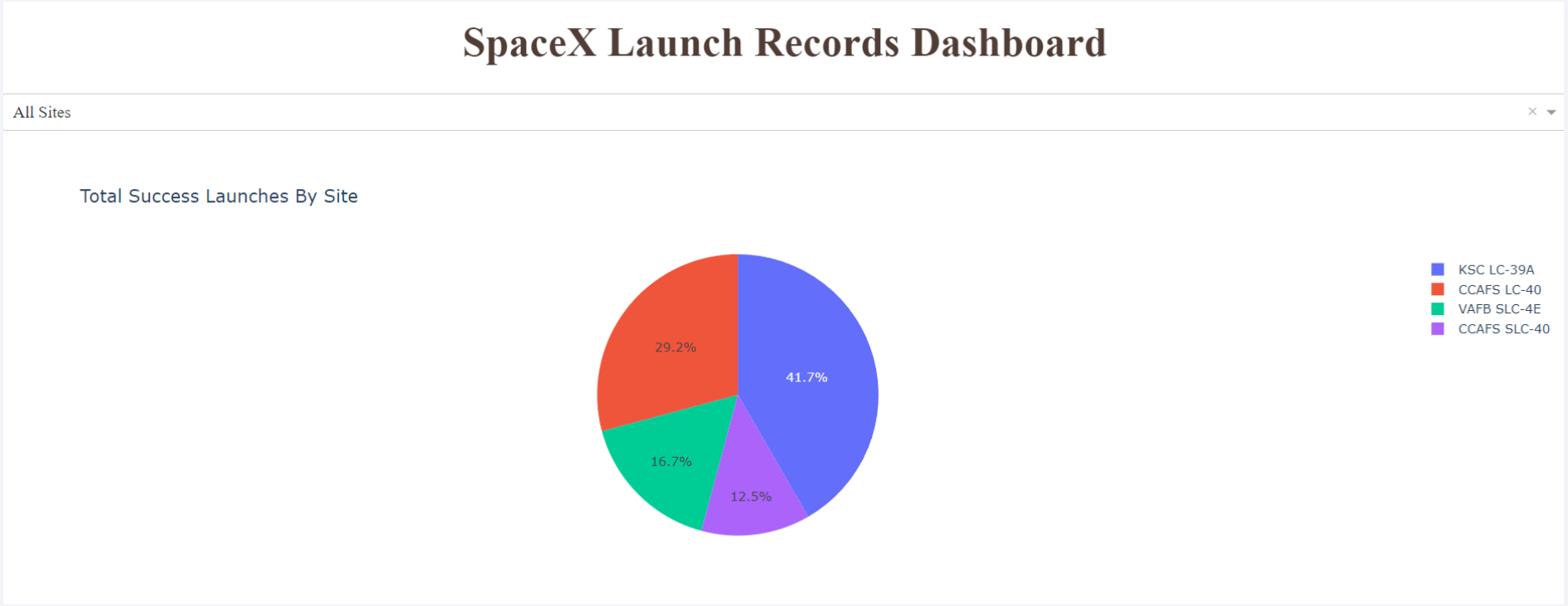
<Distance from Launch Site to Proximities >

- • The perimeter road around CCAFS LC-40 is 0.19 km away from the launch site coordinates.
- • The coastline is 0.92 km away from CCAFS LC-40.
- • The rail line is 1.33 km away from CCAFS LC-40.


Section 4
Build a Dashboard with Plotly Dash

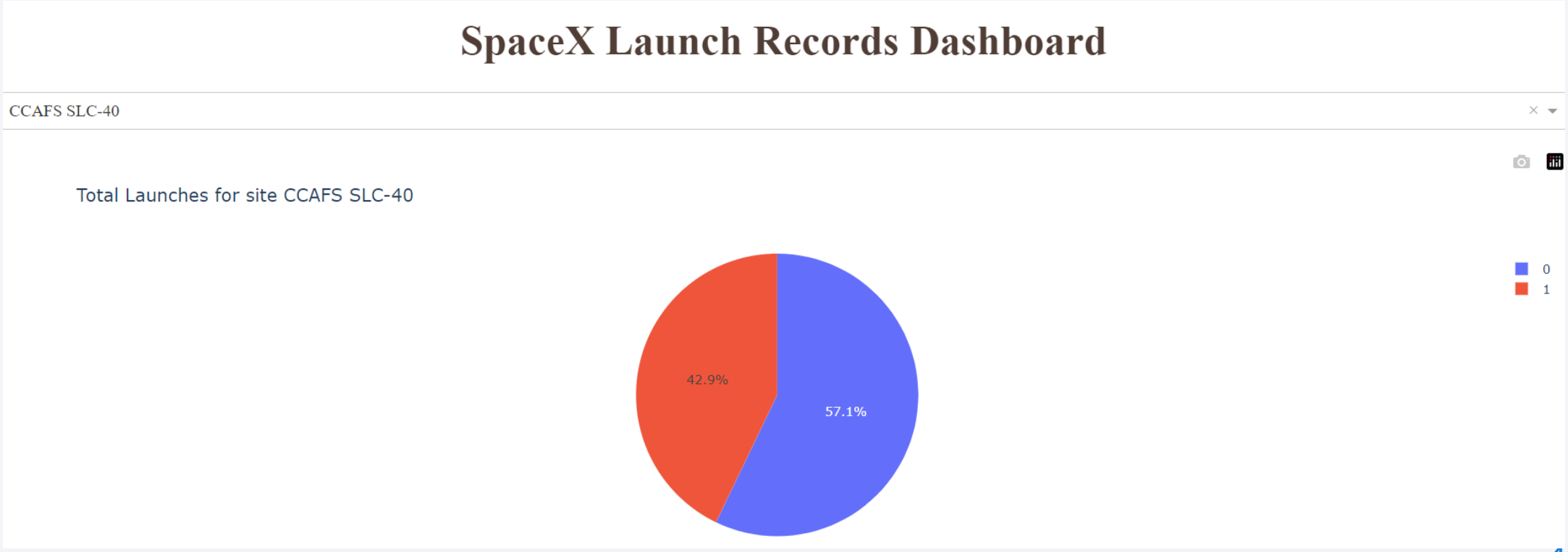
Total success launches by all sites

- KSC LC-39A recorded the most successful launches among all sites.



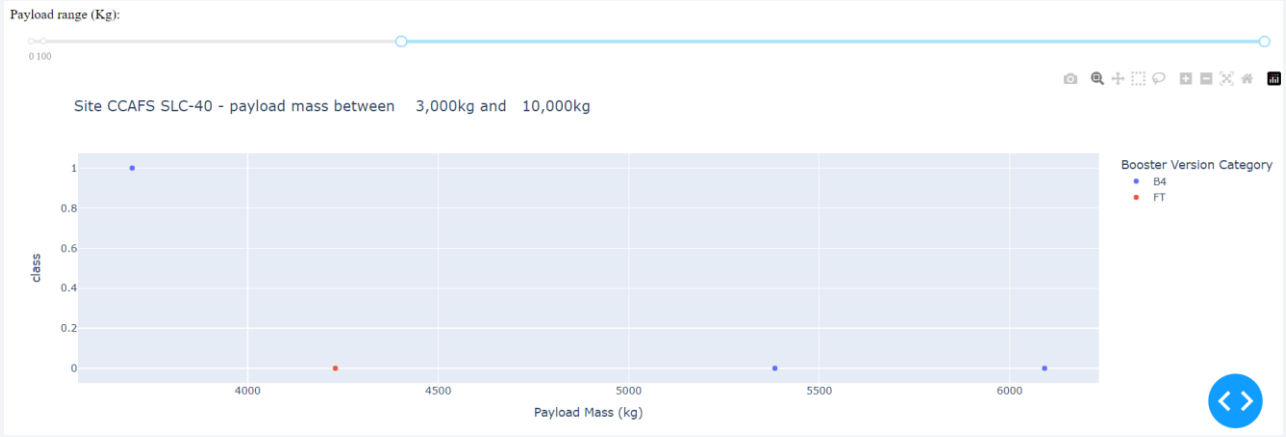
Success rate by site

- KSC LC-39A had a 42.9% success rate (indicated by the ‘0’ Class) and a 57.1% failure rate (indicated by the ‘1’ Class).



Payload vs. Launch Outcome

- Success rates are higher for low-weight payloads compared to heavy ones.

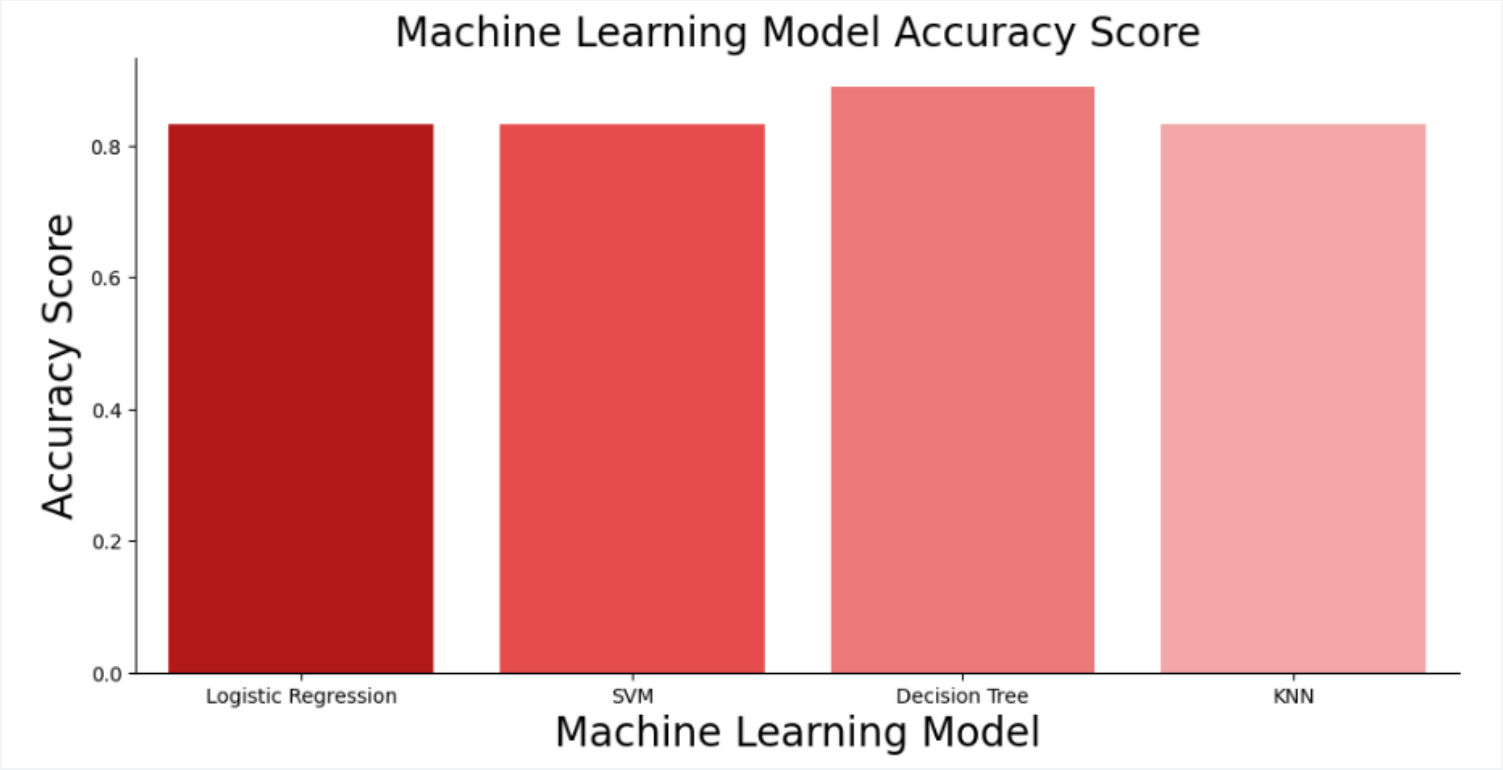


Section 5

Predictive Analysis (Classification)

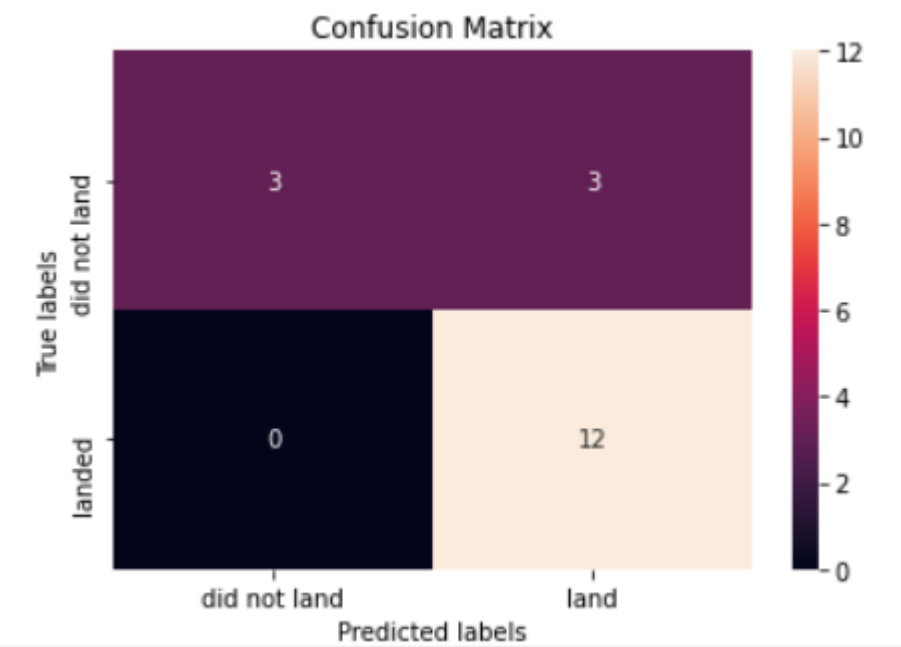
Classification Accuracy

- All models performed comparably, except for the Decision Tree model, which lagged behind the others in performance.



Confusion Matrix

- The confusion matrix for the Decision Tree Classifier demonstrates its accuracy, with a high number of true positives and true negatives indicating effective classification. However, the classifier's primary challenge is the occurrence of false positives, where unsuccessful landings are incorrectly identified as successful



Prediction Breakdown:

- 12 True Positives and 3 True Negatives
- 3 False Positives and 0 False Negatives

Conclusions

- Success Trends: SpaceX's Falcon 9 first stage landing outcomes have improved over time, with success rates increasing notably from 2013 to 2020. KSC LC-39A emerged as the most successful launch site, and orbits such as ES-L1, GEO, HEO, SSO, and VLEO showed high success rates.
- Payload Insights: Launches with payloads over 7,000 kg tend to be less risky, indicating that heavier payloads are associated with better outcomes.
- Predictive Modeling: The Decision Tree Classifier is effective for predicting successful landings, offering a valuable tool for optimizing future missions and enhancing profitability.

Appendix

- For notebook and datasets , check this Github repository link :
- https://github.com/zaghdoudiakrem/Data_science_capstone/tree/main

Thank you!