



IBM Developer  
SKILLS NETWORK

# Winning Space Race with Data Science

Juan Pablo Cabanillas  
June 17<sup>th</sup> 2025



# Outline

---

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

# Executive Summary

---

- **Summary of methodologies**
  - Data wrangling
  - Data Collection
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis
- **Summary of all results**
  - The exploratory analysis revealed key patterns.
  - I presented insights with interactive visual tools.
  - The classification models provided measurable prediction results.

# Introduction

---

- Project background and context
  - SpaceX has made space missions more affordable by reusing its rocket's first stage. My goal in this project is to predict whether that stage will successfully land, using public data and machine learning models.
- Problems you want to find answers
  - How do features like payload mass, launch site, flight number, and orbit influence landing success?
  - Has the success rate improved over time?
  - Which classification model works best for this prediction?



Section 1

# Methodology

# Methodology

## Executive Summary

---

- Data collection methodology:
  - I used the SpaceX REST API and also scraped Wikipedia to gather complete launch data.
- Perform data wrangling
  - I filtered out unnecessary data.
  - I handled missing values.
  - I applied One-Hot Encoding to make categorical features usable for machine learning.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - I trained and optimized several classifiers to predict landing outcomes.

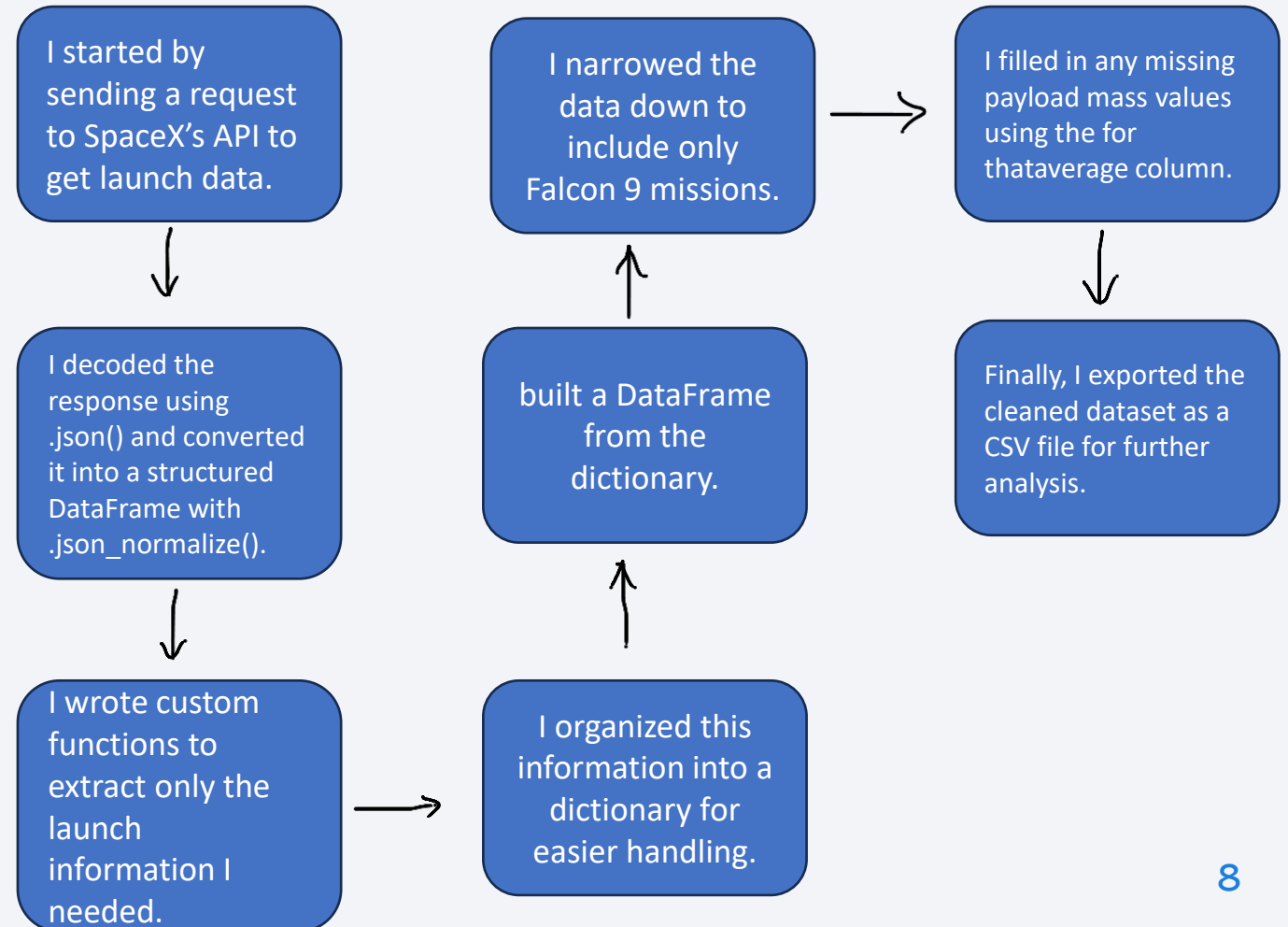
# Data Collection

---

- Describe how data sets were collected.
  - I retrieved launch data through custom functions.
  - I turned API responses into a usable DataFrame.
  - I filtered for Falcon 9 launches only.
  - I filled in missing payload values with their mean.
  - I exported everything to a CSV file for analysis.
- You need to present your data collection process use key phrases and flowcharts
  - I scraped launch tables from Wikipedia using BeautifulSoup.
  - I parsed headers and rows into a dictionary and then into a DataFrame.
  - I exported this data to a CSV as well.

# Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts
- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

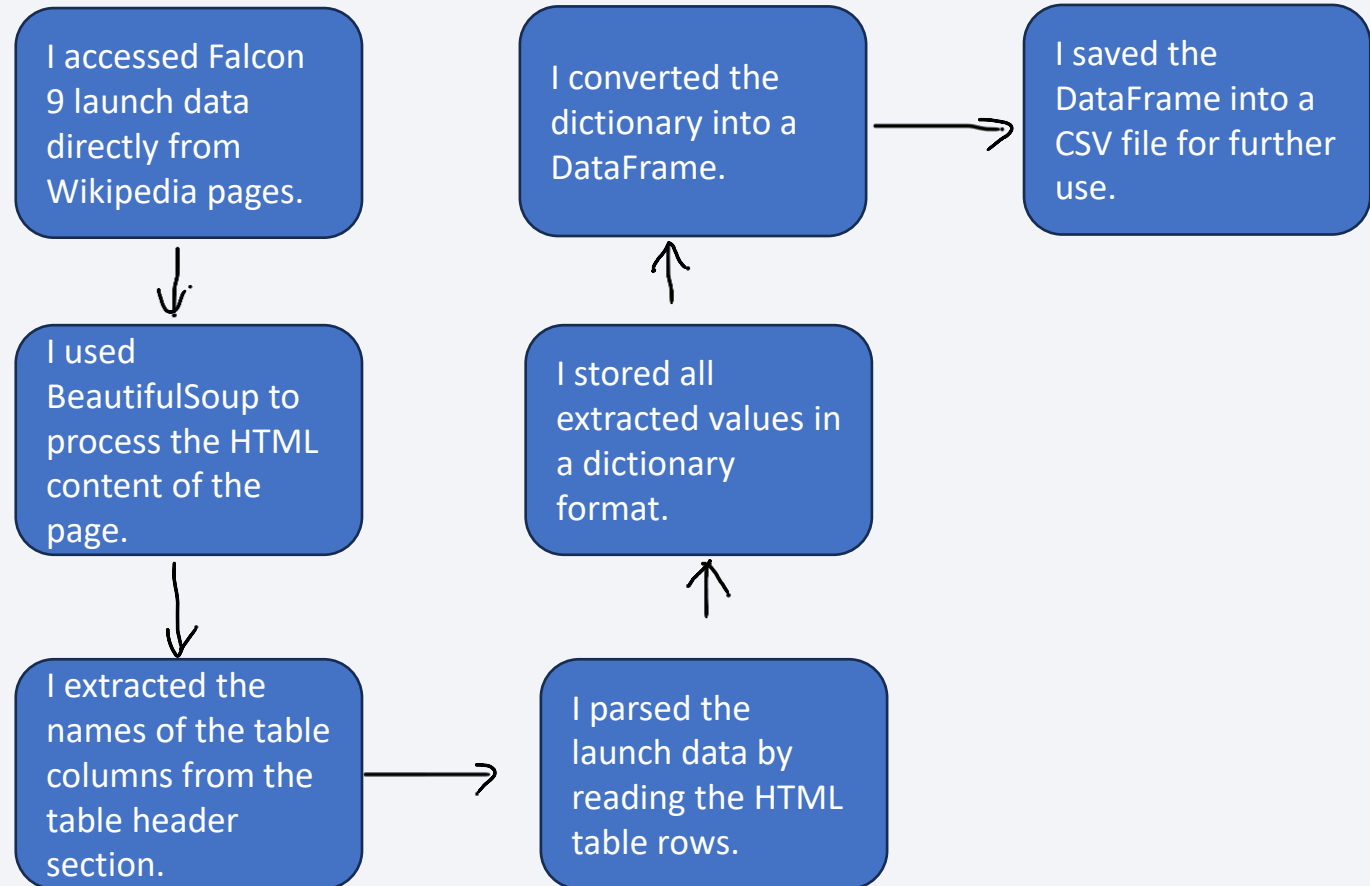




# Data Collection - Scraping

---

- Present your web scraping process using key phrases and flowcharts
- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose



# Data Wrangling

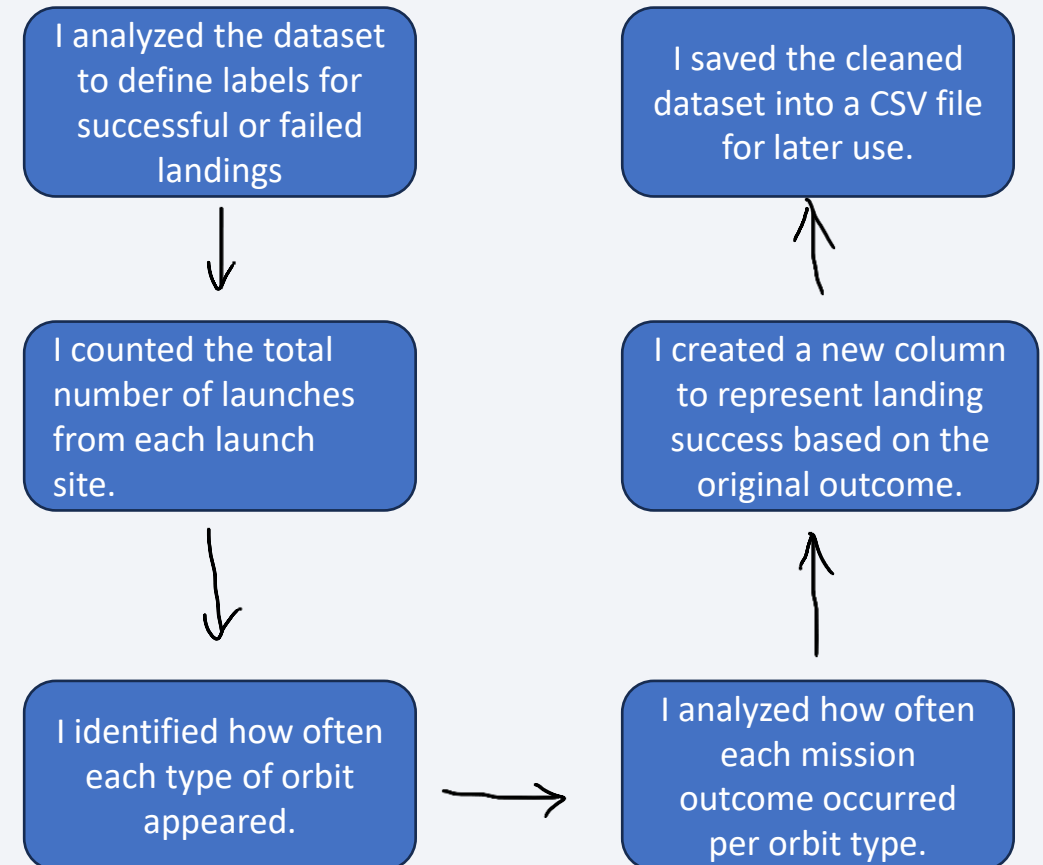
In the dataset, there were multiple variations of failed landings depending on the outcome. For example, a booster could have aimed for the ocean or ground but still failed due to a technical issue. To simplify, I grouped all landing results into two categories:

- “1” means the booster successfully landed.
- “0” means the landing attempt failed.

This involved interpreting values like:

- **True Ocean** → landed successfully in the ocean
- **False Ocean** → failed ocean landing
- **True RTLS** → landed successfully on ground
- **False RTLS** → failed ground landing
- **True ASDS** → landed successfully on drone ship
- **False ASDS** → failed drone ship landing

These were all reclassified into binary labels to train the machine learning model more effectively



# EDA with Data Visualization

---

As part of the exploratory analysis, I created several visualizations to better understand the dataset:

**Charts generated include:**

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Orbit Type vs. Success Rate
- Flight Number vs. Orbit Type
- Payload Mass vs. Orbit Type
- Yearly trend of Success Rates

Each visualization helped identify patterns and potential correlations between different features. Here's what each chart type helped reveal:

- **Scatter plots** were used to detect relationships between variables. If a meaningful connection was found, that insight could inform the machine learning model.
- **Bar charts** were useful for comparing different categories against a measurable value, highlighting variations between groups.
- **Line charts** showed how metrics changed over time, helping to track performance or behavior across years.

# EDA with SQL

---

I performed a series of SQL queries to explore the structured data and uncover important patterns. These queries helped me answer specific questions and validate insights from the visual analysis. Here's what I explored:

- Retrieved the list of all unique launch sites used in the missions.
- Selected 5 launch records where the site name starts with “CCA”.
- Calculated the total payload mass launched by NASA (CRS missions).
- Found the average payload mass for the booster version F9 v1.1.
- Identified the exact date of the first successful ground landing.
- Listed boosters that successfully landed on drone ships and carried payloads between 4000 and 6000 kg.
- Counted how many missions were successful and how many failed.
- Found booster versions that carried the highest payload mass overall.
- Retrieved failed drone ship landings from 2015, including their booster versions and launch site details.
- Ranked the frequency of different landing outcomes (success on ground pad vs failure on drone ship) between June 4, 2010, and March 20, 2017.

# Build an Interactive Map with Folium

---

## **Placing Markers for Launch Sites:**

- I began by marking the NASA Johnson Space Center on the map using its latitude and longitude, adding a popup and label to identify the location.
- Then I added similar markers for all other launch sites to highlight their geographic positions and show how close they are to the Equator and nearby coastlines.

## **Color-coded Launch Results by Site:**

- I used colored markers to represent each launch's outcome: green for successful landings and red for failures.
- To make the map more readable, I used marker clusters so that it's easy to spot which sites have better success rates.

## **Showing Distances from Launch Site to Nearby Features:**

- I drew colored lines from the KSC LC-39A site to surrounding points of interest, such as railroads, highways, the coastline, and the nearest city.
- This helped visualize how strategic the site's location is in terms of safety and accessibility.



# Build a Dashboard with Plotly Dash

---

## **Dropdown Menu for Launch Sites:**

- I added a dropdown component that allows users to choose a specific launch site from the list.

## **Pie Chart to Display Launch Successes (All or Selected Site):**

- I included a pie chart that shows the number of successful launches across all sites.
- If a specific site is selected, the chart updates to compare successes versus failures just for that location.

## **Payload Mass Range Slider:**

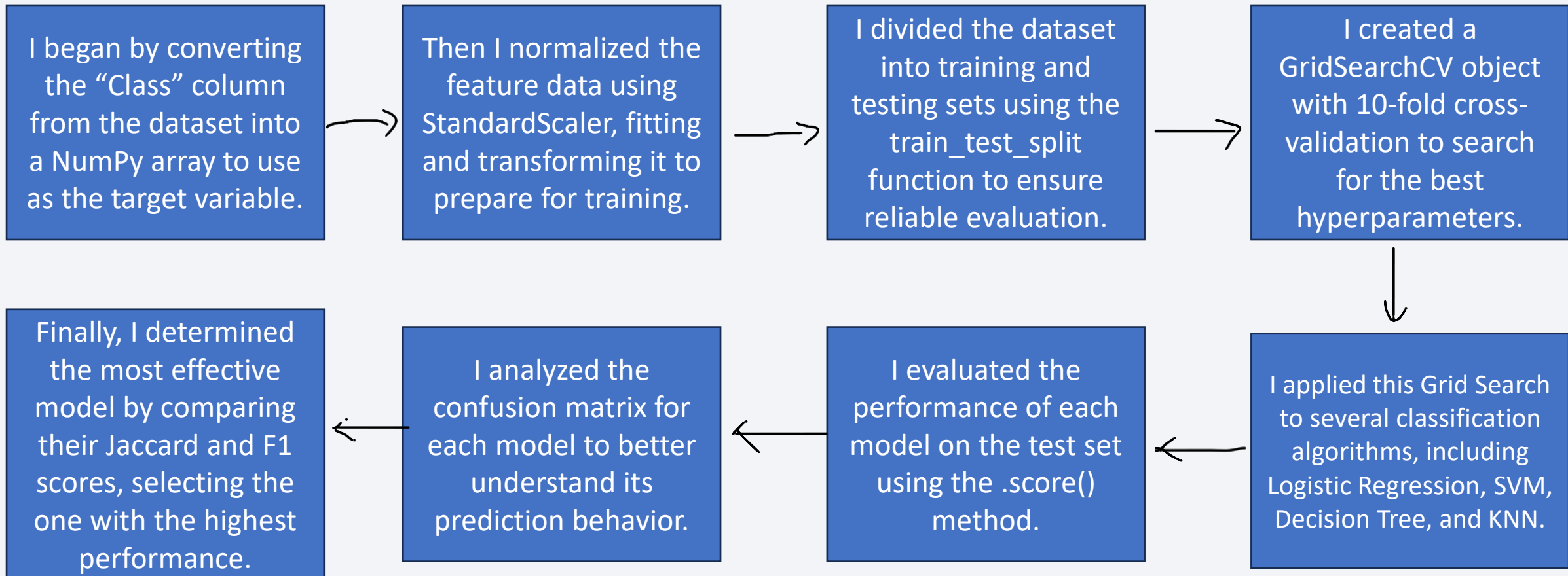
- I implemented a slider to filter the dataset by payload weight, so users can explore results based on different payload sizes.

## **Scatter Plot for Payload vs. Success by Booster Type:**

- I created a scatter plot that visualizes how payload mass relates to the likelihood of a successful launch, split by booster version.

# Predictive Analysis (Classification)

---



# Results

---

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



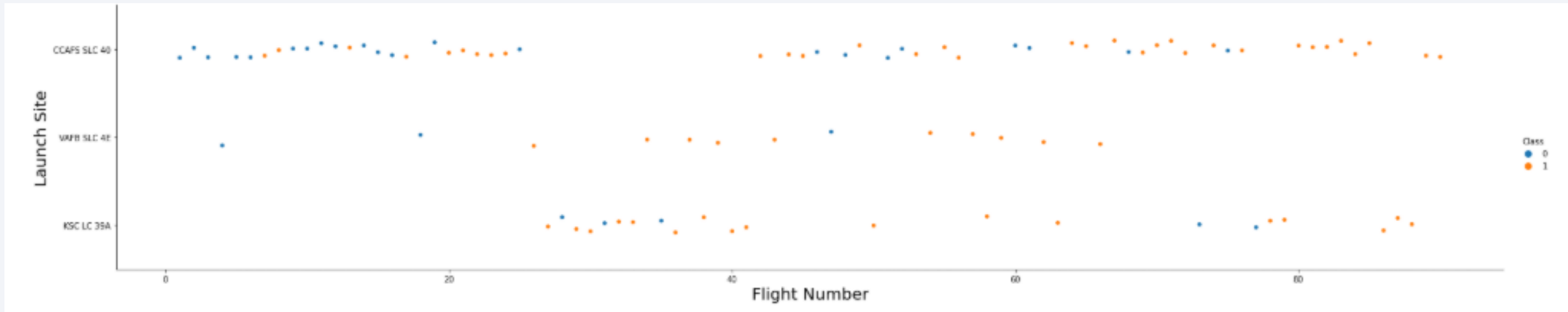


Section 2

# Insights drawn from EDA



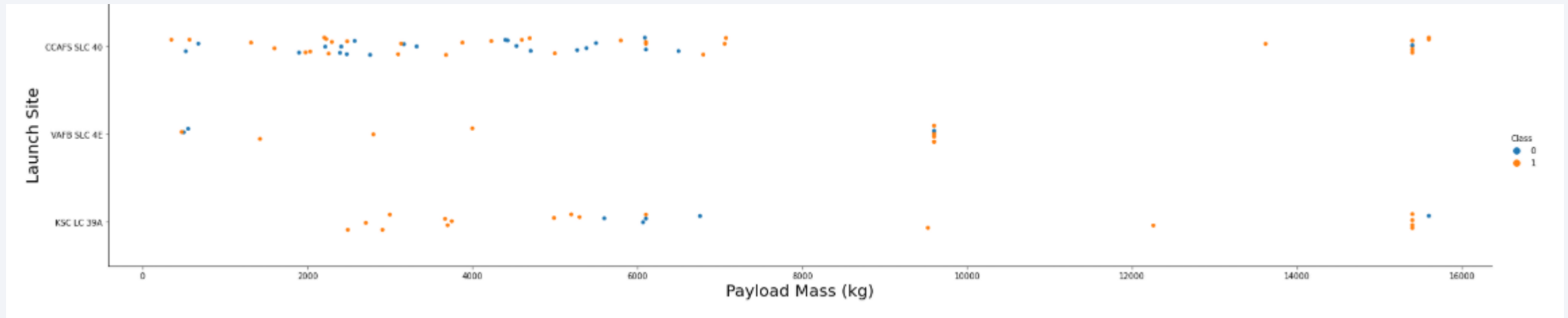
# Flight Number vs. Launch Site



- The first few missions were unsuccessful, but as the flight numbers increased, the success rate improved significantly.
- Nearly 50% of all launches took place at the CCAFS SLC 40 launch site.
- The VAFB SLC 4E and KSC LC 39A locations show notably better performance in terms of successful landings.
- This trend suggests that with each new mission, the chances of success become higher.

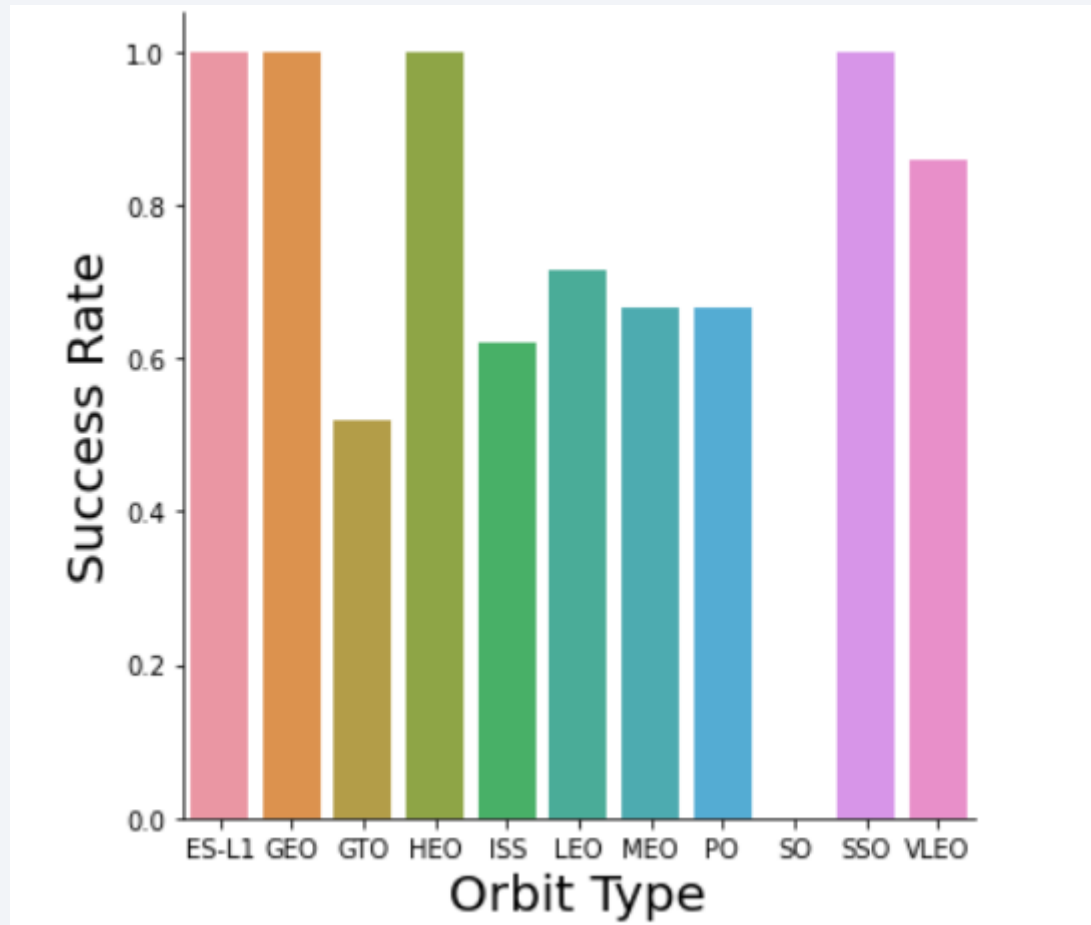


# Payload vs. Launch Site



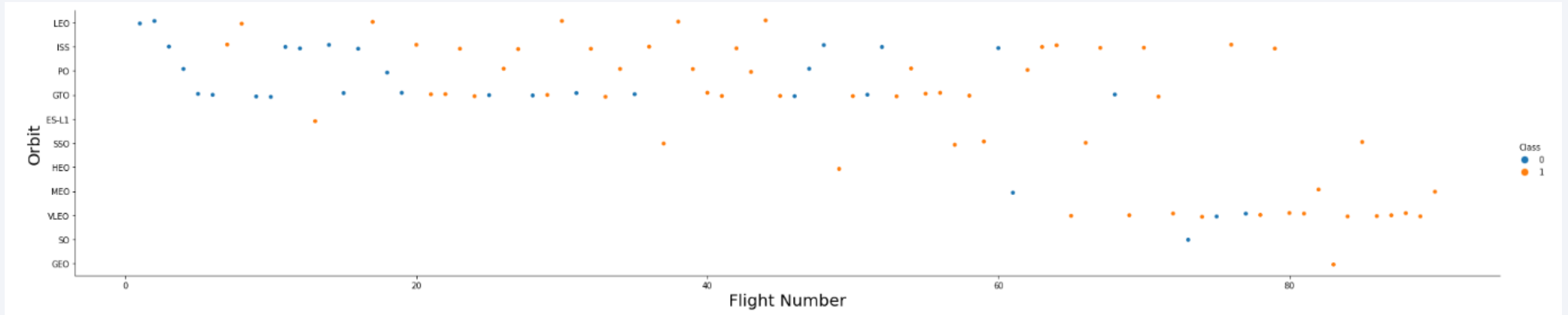
- Across all launch sites, an increase in payload mass tends to be associated with a higher probability of mission success.
- Most launches that carried more than 7000 kg ended up being successful.
- At the KSC LC 39A site, all missions with payloads under 5500 kg were successful, showing a perfect track record in that range.

# Success Rate vs. Orbit Type



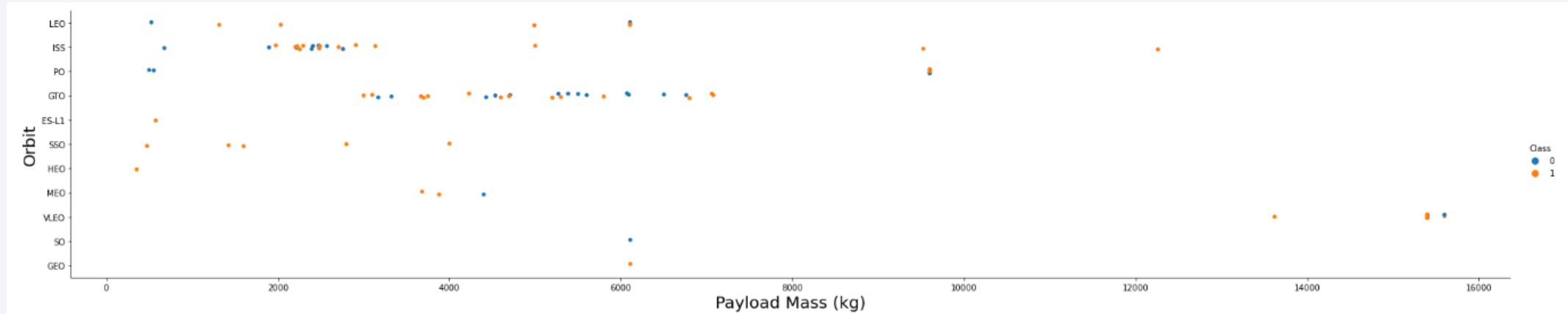
- Some orbit types achieved a perfect track record, showing a 100% success rate. These include: ES-L1, GEO, HEO, and SSO.
- The SO orbit had no successful landings, with a 0% success rate.
- Several orbits had moderate success rates, ranging from 50% to 85%. These include: GTO, ISS, LEO, MEO, and PO.

# Flight Number vs. Orbit Type



- For missions targeting LEO (Low Earth Orbit), there seems to be a positive link between the number of flights and the likelihood of success — the more missions flown, the better the outcome.
- However, this pattern doesn't hold for GTO (Geostationary Transfer Orbit), where success rates don't appear to be influenced by how many times a mission has flown.

# Payload vs. Orbit Type

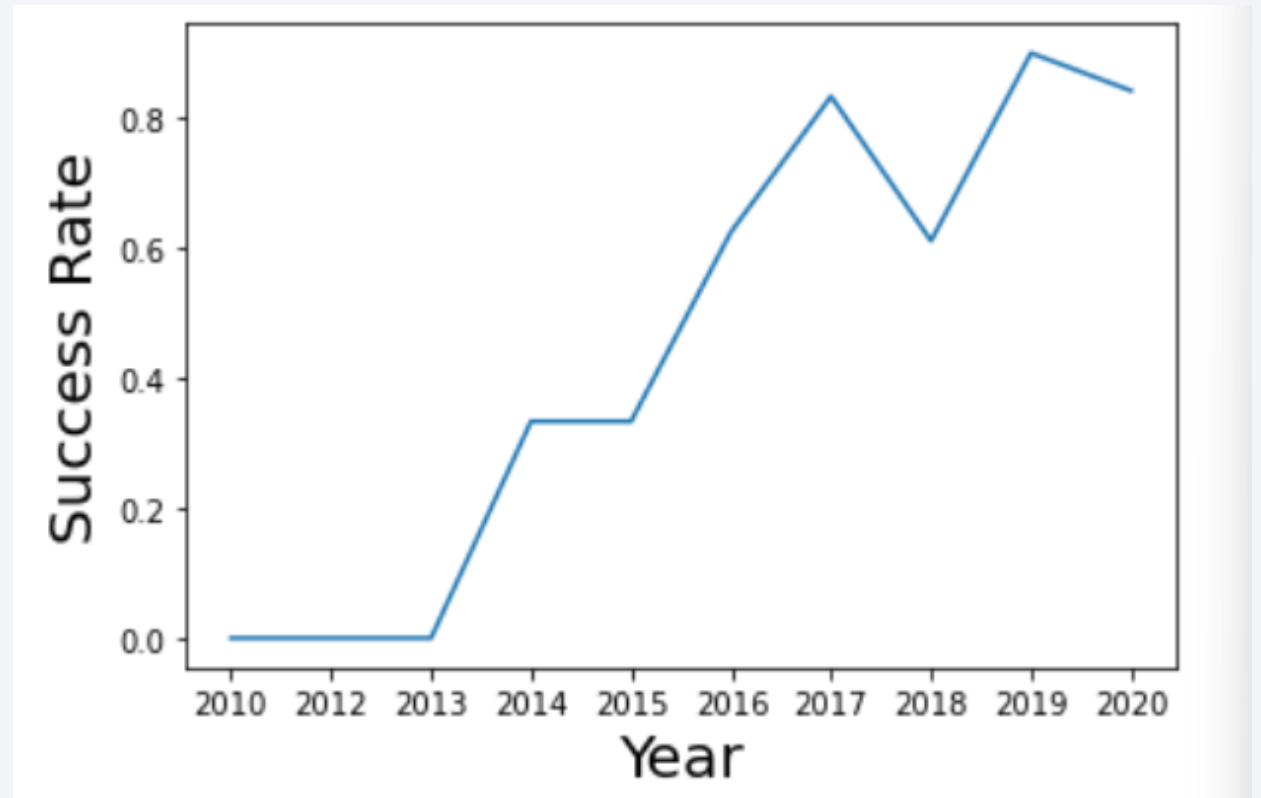


- Heavier payloads tend to lower the success rate when targeting GTO orbits.
- However, these same heavy payloads seem to perform better in missions to Polar LEO or ISS-related orbits, where success is more likely despite the mass.

# Launch Success Yearly Trend

---

- Starting from 2013, the launch success rate showed a steady upward trend and continued to improve consistently through 2020.





# All Launch Site Names

---

- Listing all distinct launch site names that were used during the space missions.

Display the names of the unique launch sites in the space mission dataset.

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

\* ibm\_db\_sa://wzf08322:\*\*\*@0c77d6f2-5da9-48a9-81f8-86b52  
Done.

Out[4]:

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

# Launch Site Names Begin with 'CCA'

- Retrieving 5 entries where the launch site name starts with the prefix “CCA”.

## Task 2

Display 5 records where launch sites begin with the string 'CCA'

In [5]:

```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
```

\* ibm\_db\_sa://wzf08322:\*\*\*@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8l1cg.databases.appdomain.cloud:31198/bludb Done.

Out[5]:

DATE	time_utc	booster_version	launch_site	payload	payload_mass_kg	orbit	customer	mission_outcome	landing_outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

# Total Payload Mass

---

## Task 3

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

```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';  
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8l1cg.databases.appdomain.clo  
Done.
```

```
Out[6]: total_payload_mass
```

```
45596
```

- Calculating the combined payload mass delivered by all booster launches commissioned by NASA under CRS missions.

# Average Payload Mass by F9 v1.1

---

## Task 4

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

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

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

```
Out[7]: average_payload_mass
```

```
2534
```

- Displaying average payload mass carried by booster version F9 v1.1.

# First Successful Ground Landing Date

---

```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8l1cg.databases.appdomain.cloud:31198/bl
Done.
Out[8]: first_successful_landing
2015-12-22
```

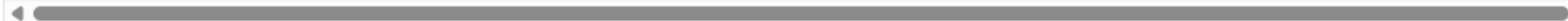
- Listing the date when the first successful landing outcome in ground pad was achieved



# Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ betw
```



```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[9]: **booster\_version**

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

- Extracting the booster versions that successfully landed on a drone ship and carried payloads within the 4000 to 6000 kg range.

# Total Number of Successful and Failure Mission Outcomes

---

## Task 7

List the total number of successful and failure mission outcomes

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

* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8l1cg.databases.appdc
Done.
```

```
Out[10]:
```

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

- Counting the total number of missions that resulted in either success or failure.

# Boosters Carried Maximum Payload

## Task 8

List the names of the booster\_versions which have carried the maximum payload mass. Use a subquery

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

\* ibm\_db\_sa://wzf08322:\*\*\*@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od81cg.databases.appdomain.cloud:31198/bludb  
Done.

Out[11]: **booster\_version**

F9 B5 B1048.4

F9 B5 B1049.4

F9 B5 B1051.3

F9 B5 B1056.4

F9 B5 B1048.5

F9 B5 B1051.4

F9 B5 B1049.5

F9 B5 B1060.2

F9 B5 B1058.3

F9 B5 B1051.6

F9 B5 B1060.3

F9 B5 B1049.7

- Retrieving the booster versions that carried the heaviest payloads ever recorded in the dataset.

# 2015 Launch Records

---

## Task 9

List the failed landing\_outcomes in drone ship, their booster versions, and launch site names for the in year 2015

In [12]:

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

```
* ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:3:
Done.
```

Out[12]:

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

- Identifying which booster versions were responsible for launching the heaviest payloads recorded in the entire dataset.

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

## Task 10

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

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

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

```
Out[13]:
```

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

- Sorting and displaying the different types of landing outcomes based on how frequently they occurred between June 4, 2010, and March 20, 2017 — from the most common to the least.

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The image is a composite of a dark blue sky and a view of the Earth's surface, which is covered in a dense network of city lights and clouds. The lights are concentrated in the lower right portion of the image, while the upper left shows a clear blue sky.

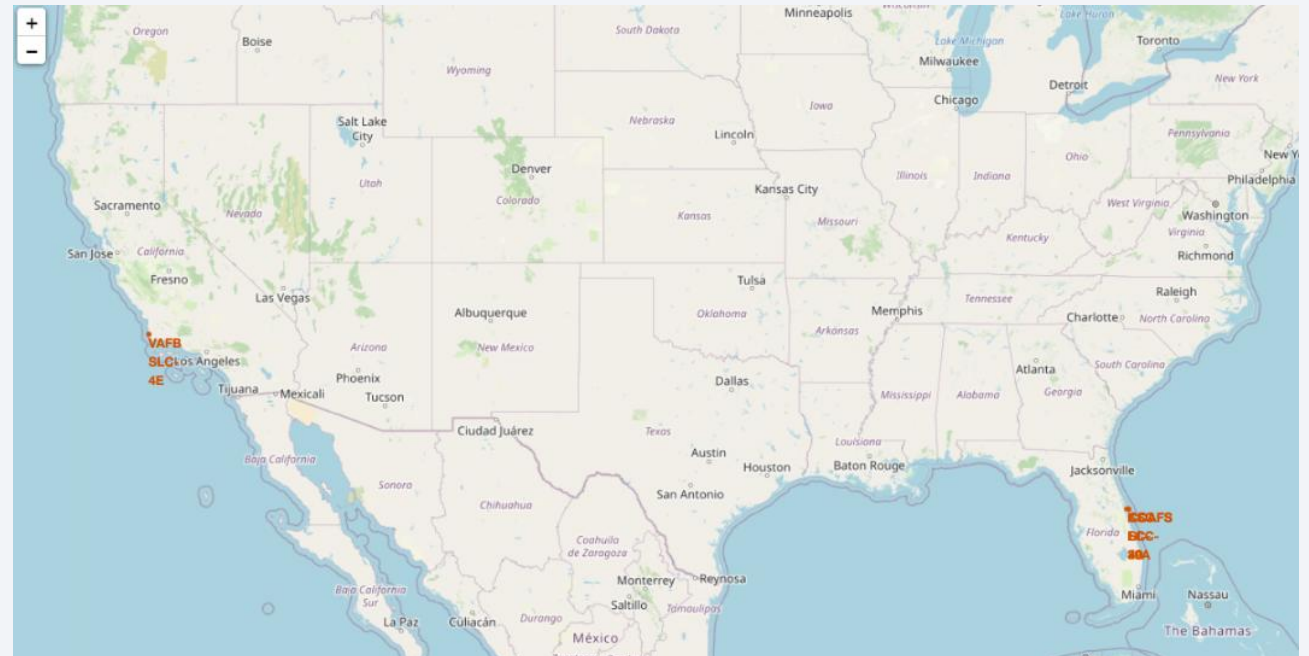
Section 3

# Launch Sites Proximities Analysis

# All launch sites' location markers on a global map

---

- Most launch sites are strategically located near the equator, where the Earth's rotation is fastest—about 1670 km/h. This added speed gives rockets an extra push when launching into orbit, thanks to inertia.
- Additionally, almost all sites are situated close to coastlines. This placement helps ensure safety by directing launches over the ocean, reducing the risk of debris falling near populated areas in the event of failure.

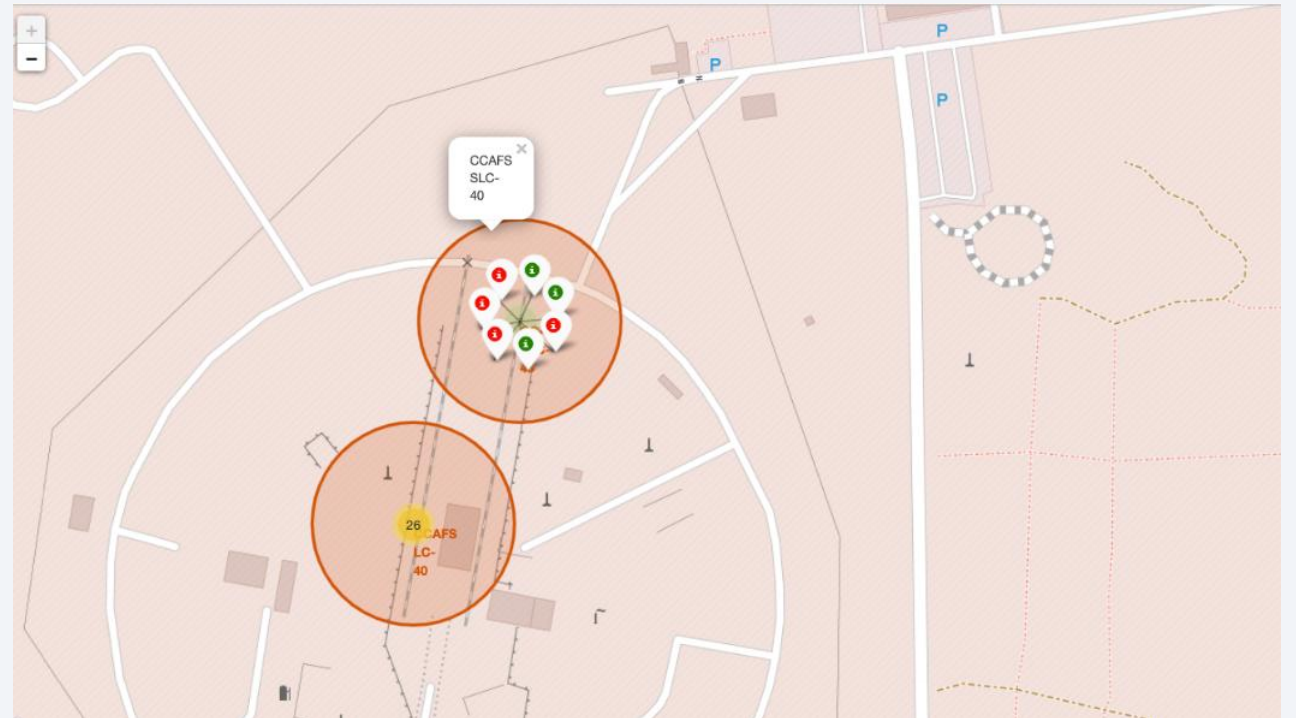




# Colour-labeled launch records on the map

---

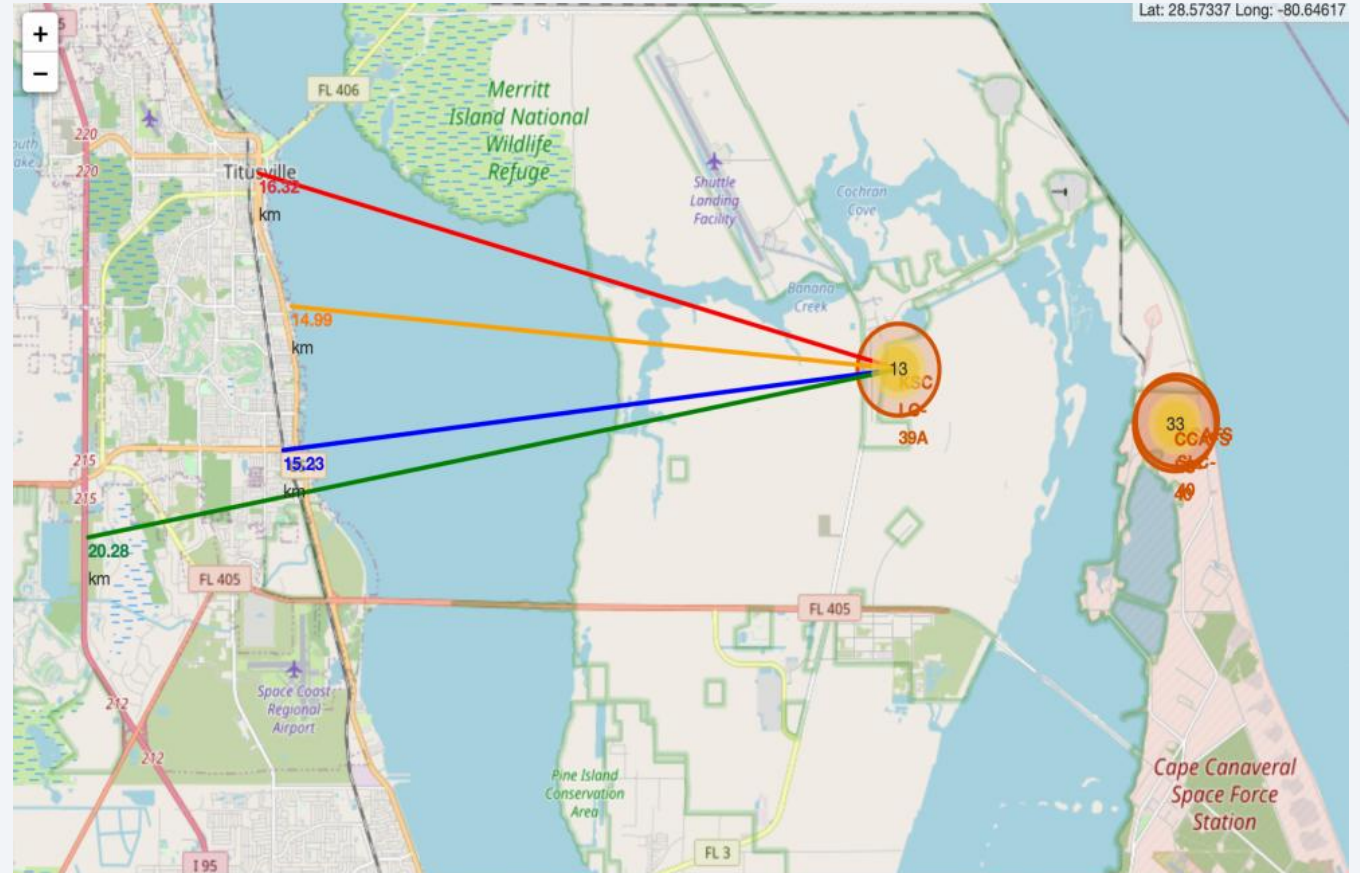
- The colored markers on the map make it easy to spot which launch sites tend to succeed more often.
  - Green markers indicate missions that landed successfully.
  - Red markers represent failed landings.
- The KSC LC-39A site stands out for having an especially high rate of successful launches.





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

- Based on the visual map, the KSC LC-39A launch site is situated relatively close to key infrastructures:
  - Around 15.23 km from the nearest railway
  - About 20.28 km from a major highway
  - Only 14.99 km away from the coastline
- It's also just 16.32 km from the nearest city, Titusville.
- Given the high velocity of rockets, a failed launch could travel 15–20 km in just a few seconds, posing a serious risk to surrounding communities.



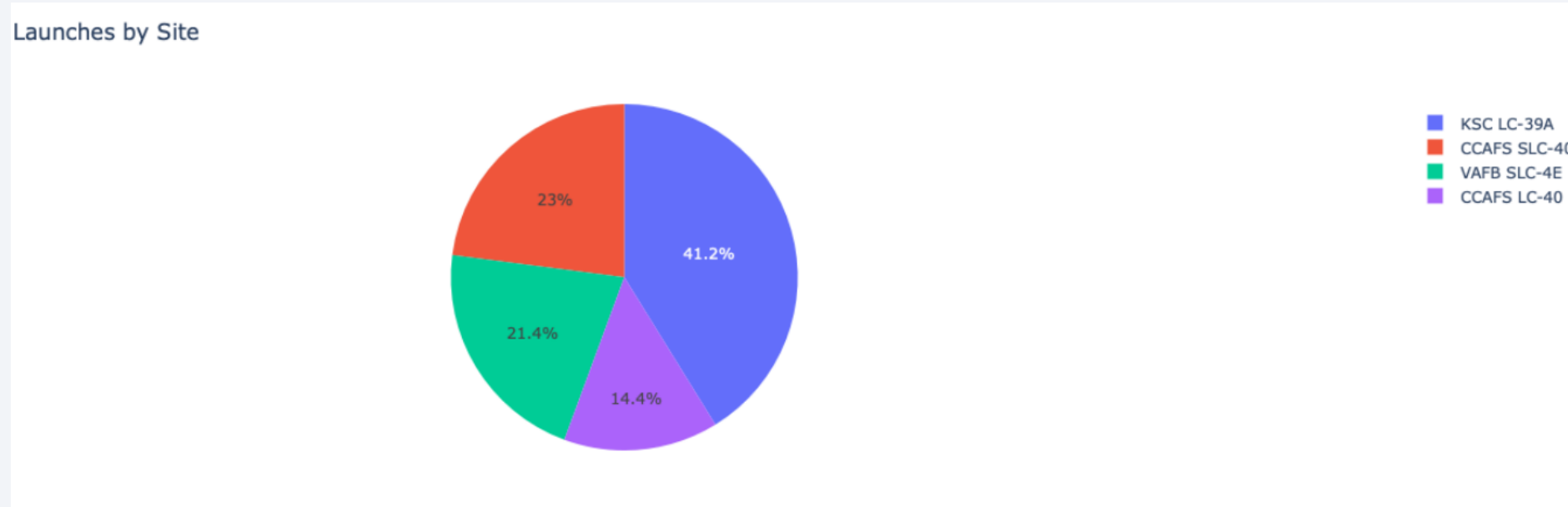


Section 4

# Build a Dashboard with Plotly Dash

# Launch success count for all sites

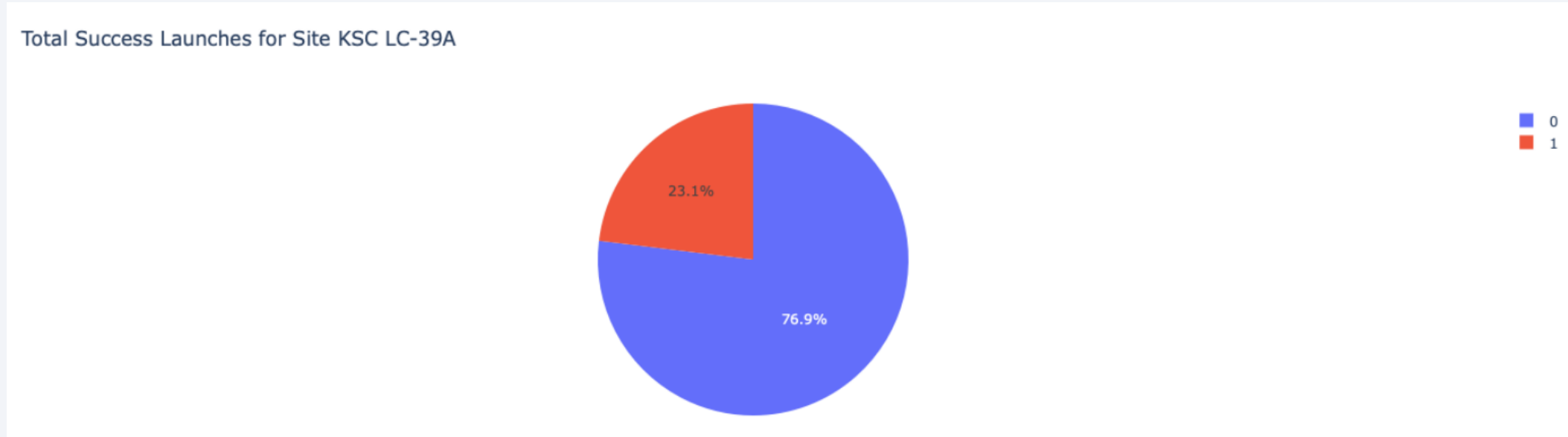
---



- The visualization clearly highlights that KSC LC-39A leads in terms of total number of successful launches compared to all other launch sites.

## <Dashboard Screenshot 2>

---



- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

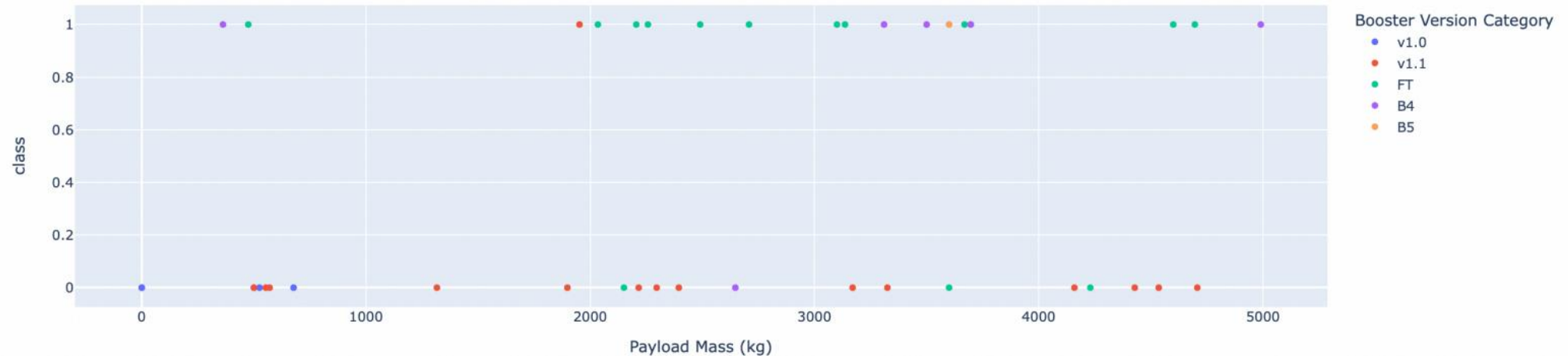


# Payload Mass vs. Launch Outcome for all sites

Payload range (Kg):



Correlation Between Payload and Success for All Sites



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



Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

---

- When evaluating just the test set results, all models appeared to perform equally well, so no single method stood out.
- This may be due to the limited size of the test set, which included only 18 samples. To get a clearer picture, I decided to evaluate all models on the full dataset.
- The results from the entire dataset revealed that the Decision Tree algorithm delivered the strongest performance. It not only had the highest accuracy, but also led in Jaccard and F1 score metrics.

**Scores and Accuracy of the Test Set**

	LogReg	SVM	Tree	KNN
<b>Jaccard_Score</b>	0.800000	0.800000	0.800000	0.800000
<b>F1_Score</b>	0.888889	0.888889	0.888889	0.888889
<b>Accuracy</b>	0.833333	0.833333	0.833333	0.833333

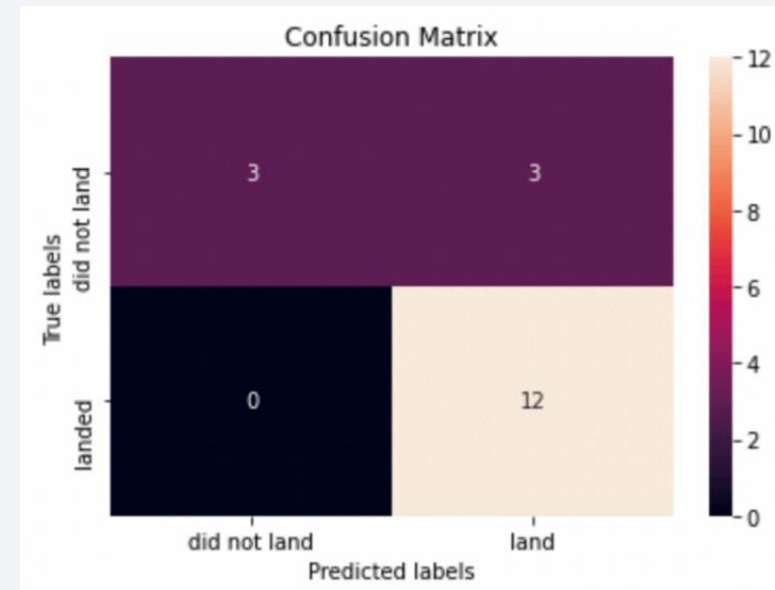
**Scores and Accuracy of the Entire Data Set**

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

# Confusion Matrix

---

- After reviewing the confusion matrix, it's clear that logistic regression is capable of differentiating between the classes.
- However, the main issue lies in the number of false positives—it misclassifies some failed landings as successful.





# Conclusions

---

- Among all tested models, the Decision Tree turned out to be the most effective for this dataset.
- Launches carrying lighter payloads tend to result in more successful landings compared to those with heavier payloads.
- Most launch sites are strategically positioned near the Equator and close to the coast, which helps with both orbital velocity and safety.
- The launch success rate has steadily improved year after year.
- The KSC LC-39A site stands out as the one with the highest number of successful missions.
- Orbits such as ES-L1, GEO, HEO, and SSO showed a 100% success rate.

# Appendix

---

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!

