



IBM Developer  
SKILLS NETWORK

# Winning Space Race with Data Science

Luis Felipe Klaus  
March 12, 2024



# Outline

---

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

# Executive Summary

---

- **Summary of methodologies**
  - Data Collection using SpaceX API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis (EDA) with Data Visualization
  - Exploratory Data Analysis (EDA) with SQL
  - Interactive Visual Analysis with Folium and Dash (Plotly)
  - Predictive Analysis with Machine Learning
- **Summary of all Results**
  - Exploratory Data Analysis (EDA) results
  - Interactive analytics (screenshots)
  - Predictive analysis results

# Introduction

---

- Project background and context

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

- Problems you want to find answers

The target of this project is to predict if the First Stage of the SpaceX Falcon 9 will land successfully given the Launch Site, payload mass, and several other operating conditions.



Section 1

# Methodology

# Methodology

---

## Executive Summary

- Data Collection Methodology
  - SpaceX Rest API
  - Web Scraping (Wikipedia)
- Perform data wrangling
  - Null and irrelevant data cleaned
  - One-hot encoding applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - LR, SVM, DT and KNN models were trained, tested and evaluated to find best classifier

# Data Collection

---

- Data was collected from multiple sources using different methods:
  - The first batch of data was collected through multiple GET requests to the SpaceX API and casted into a pandas data frame. Data was formatted and filtered to keep only Falcon 9 launches. Lastly, missing values were replaced by the mean value for the feature.
  - Additional data was collected through Beautiful Soup Web Scraping from SpaceX Falcon 9 launch records Wikipedia page. The HTML was parsed, processed and finally casted into a pandas data frame.

# Data Collection – SpaceX API

---

- Data was collected through multiple GET requests to SpaceX API, and casted into a pandas data frame. Then it was formatted and filtered to keep only Falcon 9 launches. Lastly missing values were replaced by the mean value for the feature.
- Source: <https://github.com/duckbox72/bm-applied-data-science-capstone/blob/main/1-spacex-data-collection-api.ipyn>

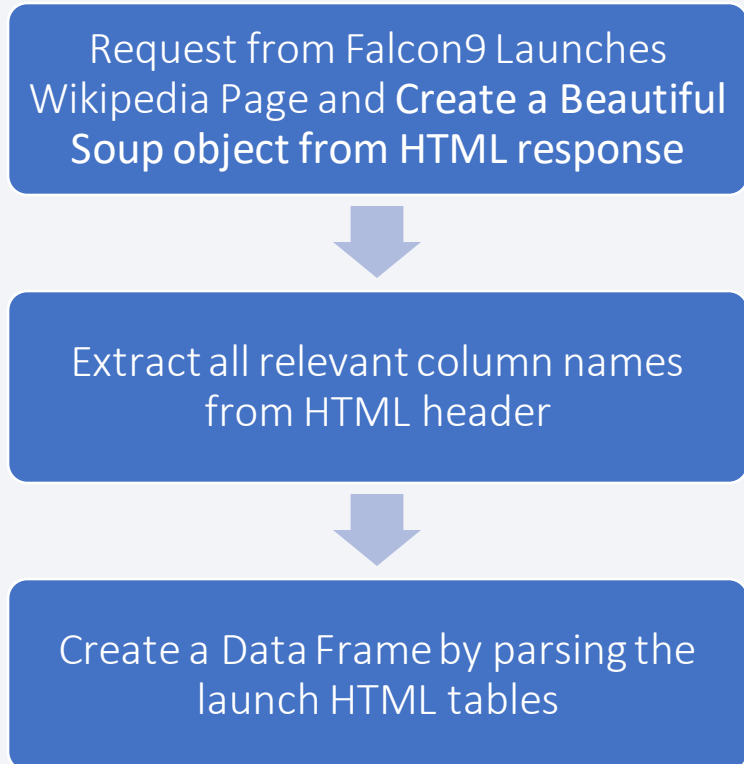




# Data Collection - Scraping

---

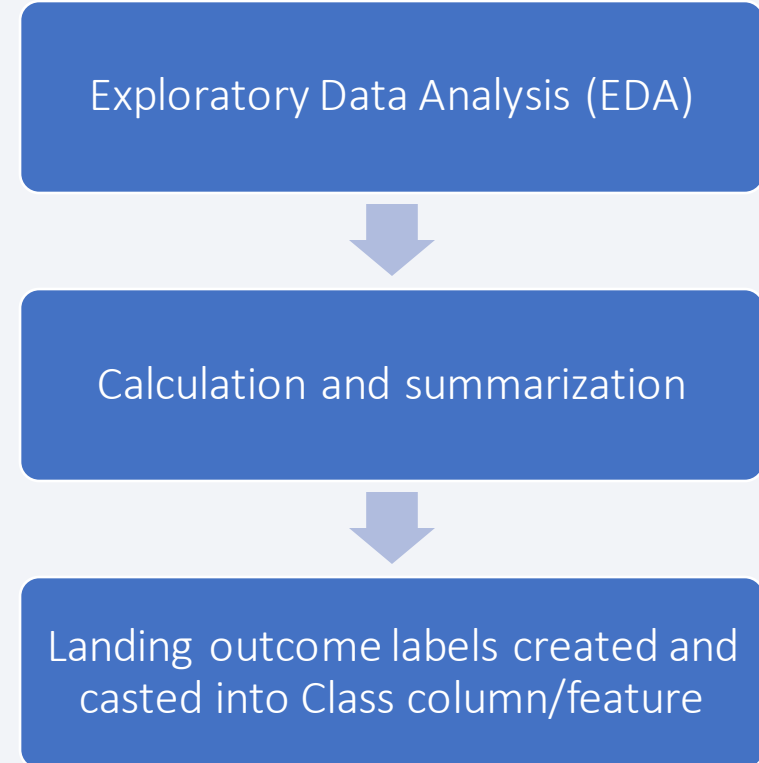
- Data was also acquired from SpaceX Falcon9 Wikipedia page. Beautiful Soup was used for web scraping, to extract relevant column names and values.
- Source: <https://github.com/duckbox72/ibm-applied-data-science-capstone/blob/main/2-spacex-data-collection-webscraping.ipynb>



# Data Wrangling

---

- Some Exploratory Data Analysis (EDA) was performed in the dataset
- The number of launches per site, and the number and occurrence outcomes per orbit were calculated
- A binary landing outcome label was created from the Outcome column, and populated into the Class column
- Source: <https://github.com/duckbox72/ibm-applied-data-science-capstone/blob/main/3-spacex-data%20wrangling.ipynb>



# EDA with Data Visualization

---

- EDA with data visualization was performed to gain insights on the relationship between multiple pairs of features. For that matter a number of charts were plotted:
  - Payload mass vs Flight number (scatter point chart)
  - Flight number vs Launch site (scatter point chart)
  - Payload mass vs Launch site (scatter point chart)
  - Success rate for each Orbit (bar chart)
  - Flight number vs Orbit type (scatter point chart)
  - Launch success trend over the years (line chart)
- Additionally relevant features were filtered and then One-hot encoding applied to categorical columns
- Source: <https://github.com/duckbox72/ibm-applied-data-science-capstone/blob/main/5-spacex-eda-data-vizualization.ipynb>

# EDA with SQL

---

- Data was loaded to a database and further EDA was performed using SQL queries to:
  - List the names of the unique launch sites
  - List five records where launch site name begin with "CCA"
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display the average payload mass carried by booster version F9 v1.1
  - Display the date when the first successful landing outcome in the ground pad was achieved
  - List boosters which successfully landed in a drone ship with a payload between 4000 and 6000 KG
  - List the total number of success and failure mission outcomes
  - List the names of booster versions that have carried the maximum payload mass
  - List failure landing outcome in drone ship, displaying month, booster version and launch site in 2015
  - Rank in descending order the count of all landing outcomes between 6-4-2010 and 3-20-2017
- Source: <https://github.com/duckbox72/ibm-applied-data-science-capstone/blob/main/4-spacex-eda-sql.ipynb>

# Build an Interactive Map with Folium

---

- Folium was used to analyze the launch sites locations and it's proximities. For that purpose a folium Map object was created and populated with various objects:
  - Circles and markers were used to add a highlighted area with a text label around each launch site
  - On each site, marker clusters were added to accommodate multiple markers, representing each a launch record, and colored green or red according to the outcome, making it ease identify success rates
  - Lines and markers were also added to represent the distance between the launch site and it's surrounding landmarks like cities, shores, roads, railroads, airports
- Source: [https://github.com/duckbox72/ibm-applied-data-science-capstone/blob/main/6-spacex-launch\\_site\\_location-folium.ipynb](https://github.com/duckbox72/ibm-applied-data-science-capstone/blob/main/6-spacex-launch_site_location-folium.ipynb)



# Build a Dashboard with Plotly Dash

---

- A Plotly Dash dashboard application was built to display interactive visual analysis in real-time
  - A pie chart was included to visually represent the success launch ratios, offering the option to view data for all sites collectively or to examine each site's performance individually.
  - A scatter chart was also plotted to represent the relationship between payload mass (kg) and outcome, also offering the option to view data for all sites collectively or each site individually. In addition a color label was used to distinguish among booster versions
- Source: [https://github.com/duckbox72/ibm-applied-data-science-capstone/blob/main/7-spacex\\_dash\\_app.py](https://github.com/duckbox72/ibm-applied-data-science-capstone/blob/main/7-spacex_dash_app.py)

# Predictive Analysis (Classification)

---

- Prior to performing predictive analysis the dataset was divided into features (X) and target (Y). Then the features were standardized using StandardScaler and resultant data was split into training and testing data using the train\_test\_split function. Finally the classifier objects were created and trained through a GridSearchCV so the optimum hiperparameters were selected and implemented in each one of them.
- Each classifier (LR, SVM, DT and KNN) was then submitted to test data in order to obtain the accuracy score and the predicted outcomes. The actual labels were then compared against the predicted labels through a confusion matrix.
- Source: <https://github.com/duckbox72/ibm-applied-data-science-capstone/blob/main/8-spacex-machine-learning-prediction.ipynb>

Dataset divided into features and target.  
Features standardized



Data split into training and testing sets,  
submitted to grid search and trained



Each classifier tested for score accuracy  
and their predictions compared to  
actual values through a confusion matrix

# Results

---

- Exploratory data analysis results
  - SpaceX operations utilized four distinct launch sites
  - The first successful landing outcome in ground pad was achieved in December 2015
  - Payload mass tended to increase through the operation years
  - Following 2013, there was a trend of increasing rates over the subsequent years
- Insights obtained from interactive analytics
  - Launch sites are positioned along the coastline, with the majority of operations taking place at facilities situated in Florida.
  - KSC LC-39A had the best success ratio and boasted an impressive a 76.9% success outcome
- Predictive analysis results
  - Decision Tree classifier had the best performance, achieving a training accuracy of 0.8892 and a testing score of 0.8888.



The background of the slide is an abstract composition. It features a dark blue field on the left side, which transitions into a complex pattern of diagonal streaks in shades of blue, red, and teal on the right. These streaks have a textured, almost woven appearance. Overlaid on this pattern is a faint, light blue grid that recedes into the distance, creating a sense of depth and perspective.

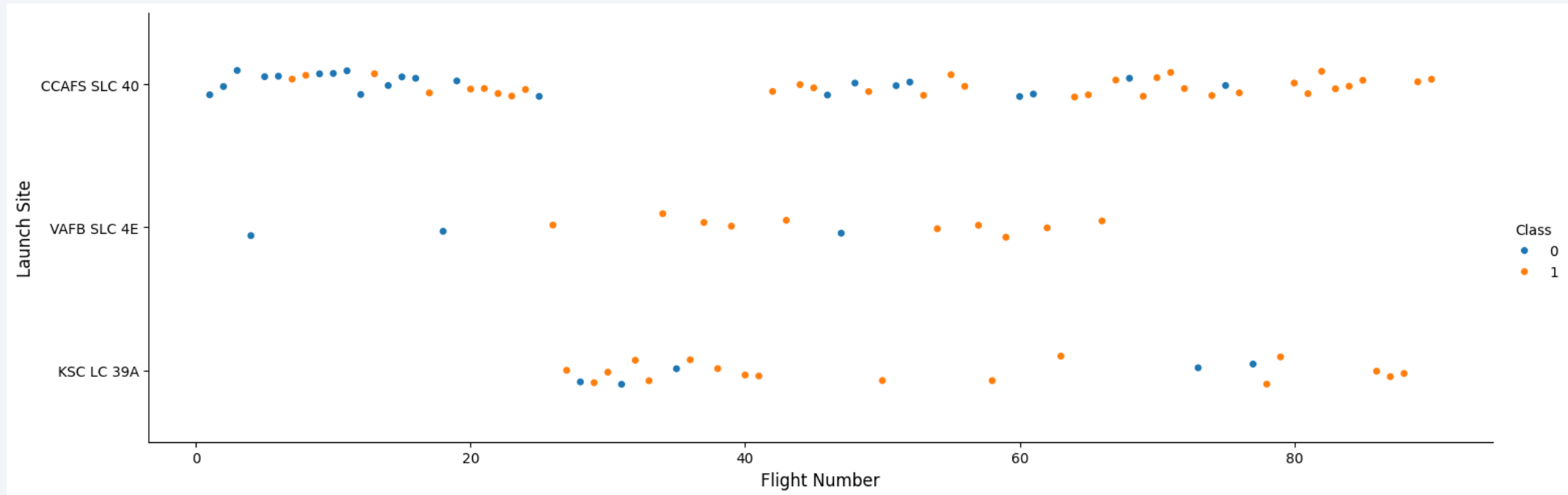
Section 2

# Insights drawn from EDA



# Flight Number vs. Launch Site

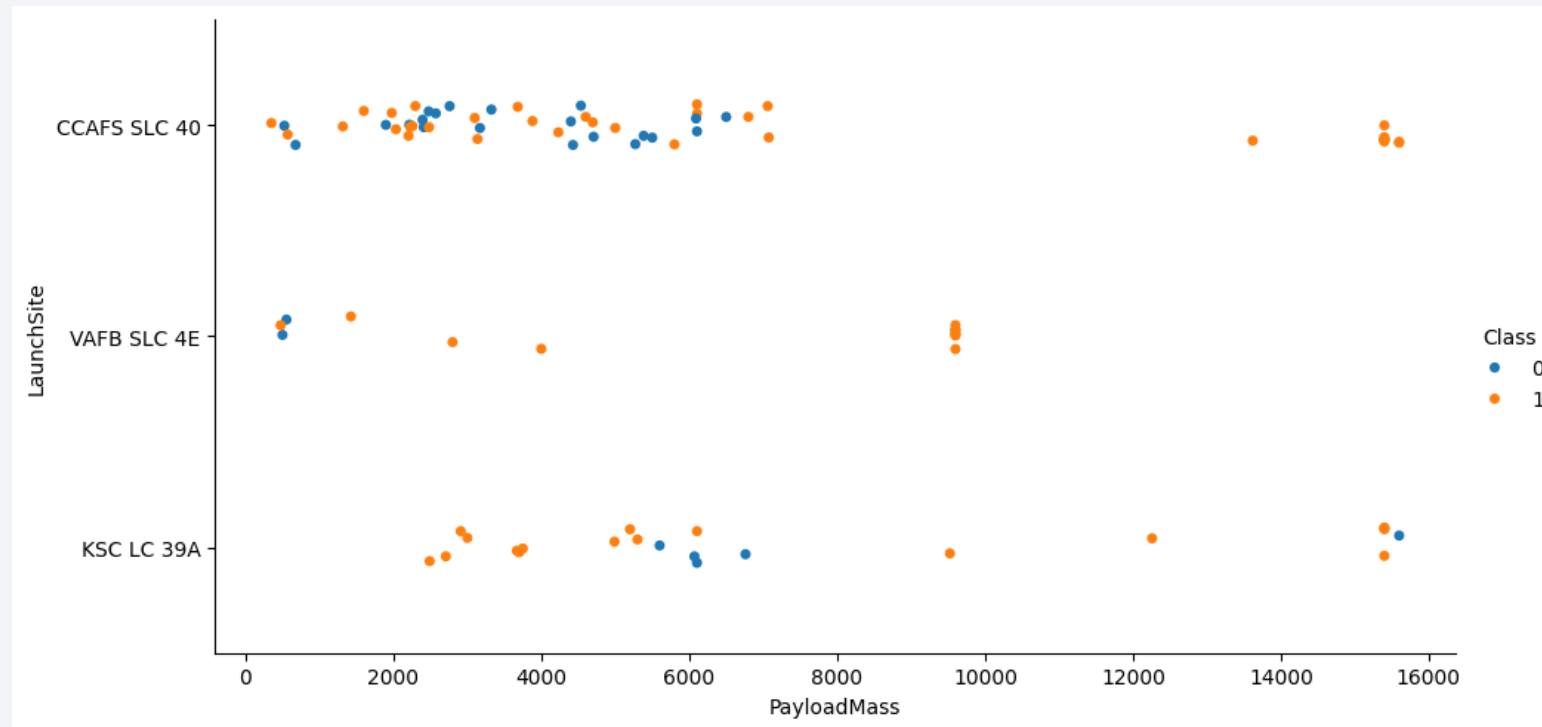
- The graph below illustrates the relation between flight numbers and launch sites. It reveals a notable trend: the occurrence of successful outcomes (class 1) increased significantly in later flights compared to earlier ones.





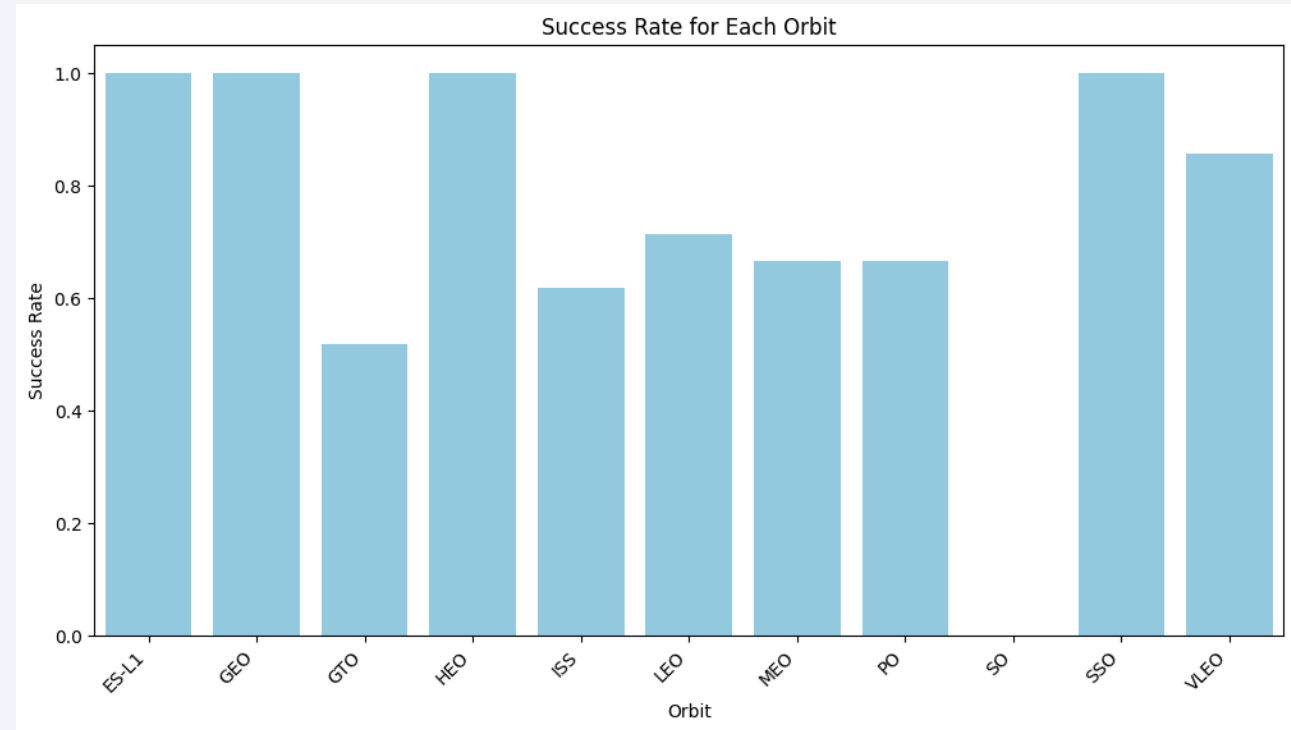
# Payload vs. Launch Site

- The plot below displays the relationship between payload mass and launch sites. It shows that no rockets launched from VAFB-SLC had a payload over 10000 KG. A trend of successful outcomes was also detected for rockets launched from any site carrying a payload mass over 7000 KG



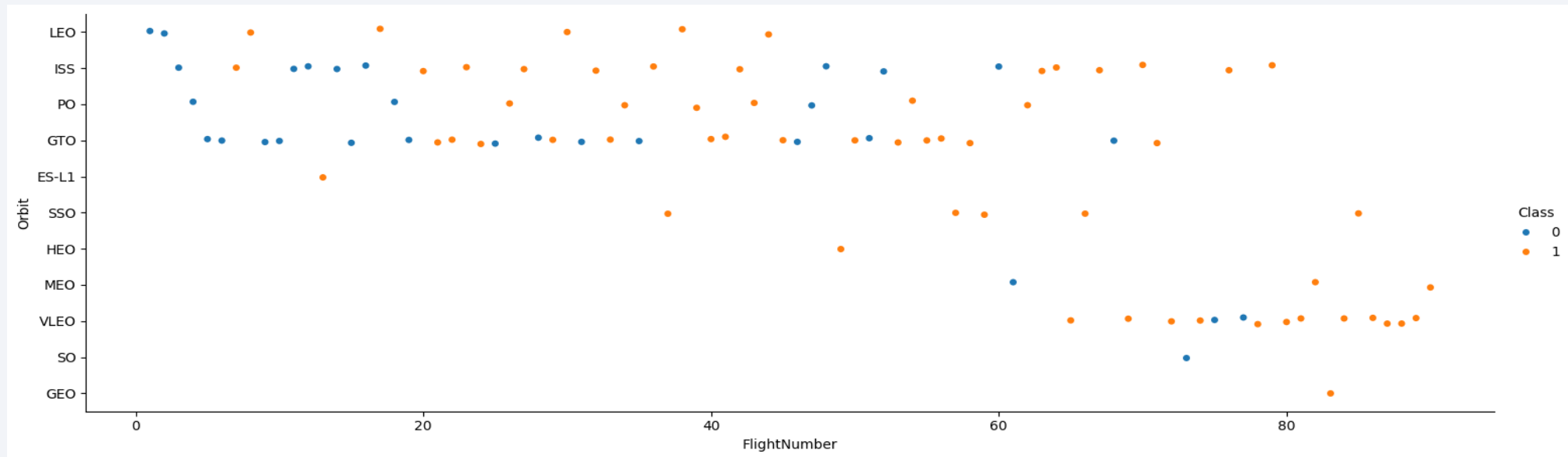
# Success Rate vs. Orbit Type

- The adjacent bar chart shows the relationship between the operation orbits and the success outcome rates. It reveals that orbits ES-L1, GEO, HEO and SSO boosted flawless success rates while orbit SO experienced no success landing outcomes.



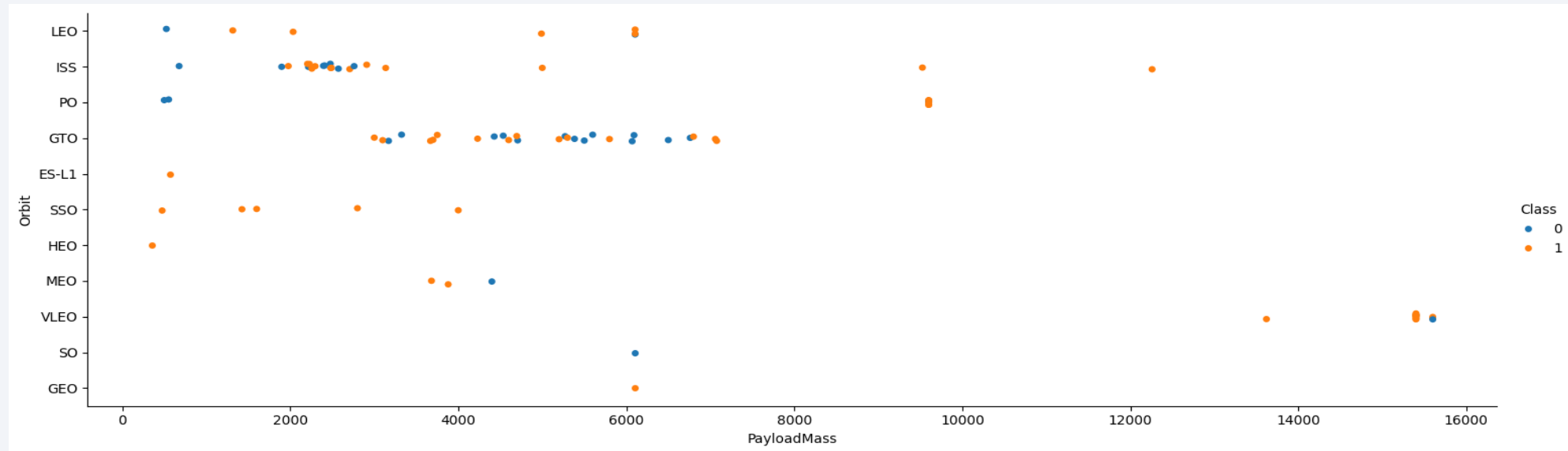
# Flight Number vs. Orbit Type

- The scatterplot below displays the relation between the flight number and various service orbits, with data points color-coded to distinguish between successful and failed landing outcomes.
- Notably, it suggests that the LEO orbit success appears to be related to the number of flights flown. Conversely, it appears to be no correlation between flight number and performance when operating in GTO orbit. Additionally it highlights a high success ratio in VLEO orbit, although operations in this orbit were only attempted after more than 60 flights.



# Payload vs. Orbit Type

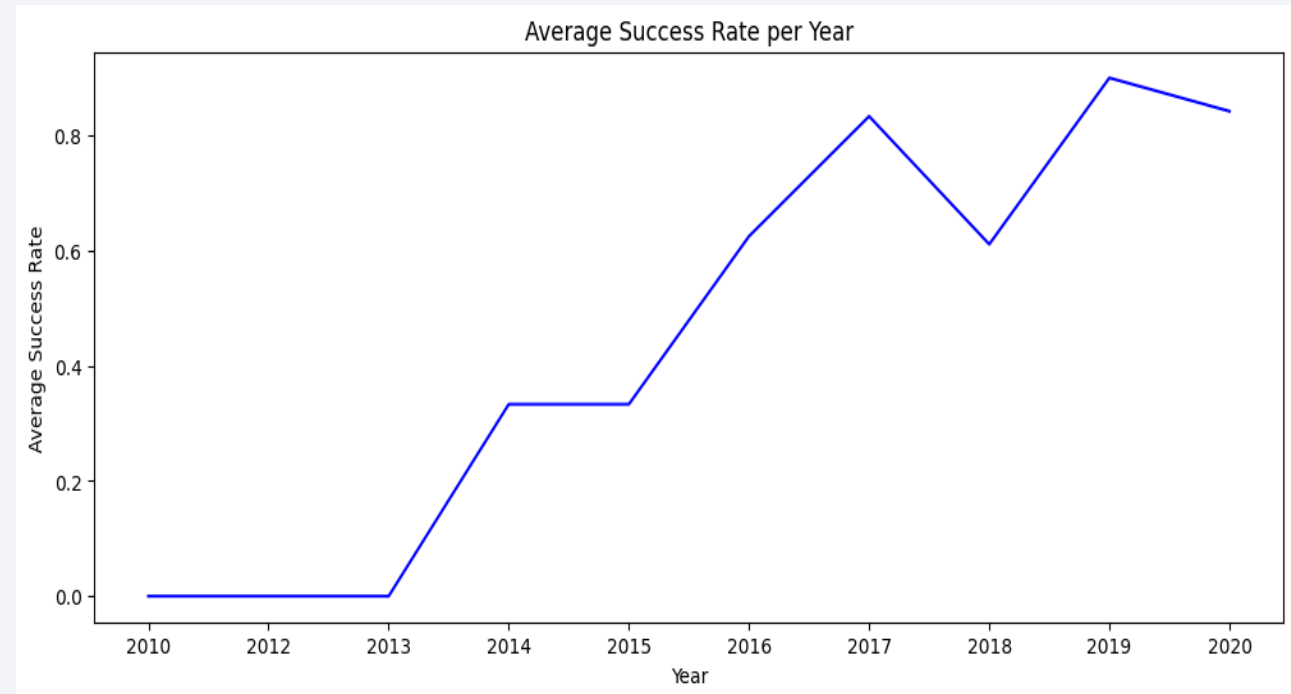
- The scatterplot below displays the relation between the payload mass in kg and various service orbits, with data points color-coded to distinguish between successful and failed landing outcomes.
- It reveals that the orbits LEO, ISS and PO exhibit greater successful landing rates when carrying heavier payloads. However for GTO orbit a trend could not be distinguished as positive and negative outcomes occurred for any payload masses



# Launch Success Yearly Trend

---

- The adjacent line chart shows the average success rate for each year of operation
- From the plot it can be observed that success rates started to improve from 2013 and kept improving over time which suggests a direct correlation between experience and success outcomes.





# All Launch Site Names

---

- The following query was performed to identify the names of the unique launch sites

```
%sql SELECT DISTINCT("Launch_Site") FROM SPACEXTABLE;
```

[6]

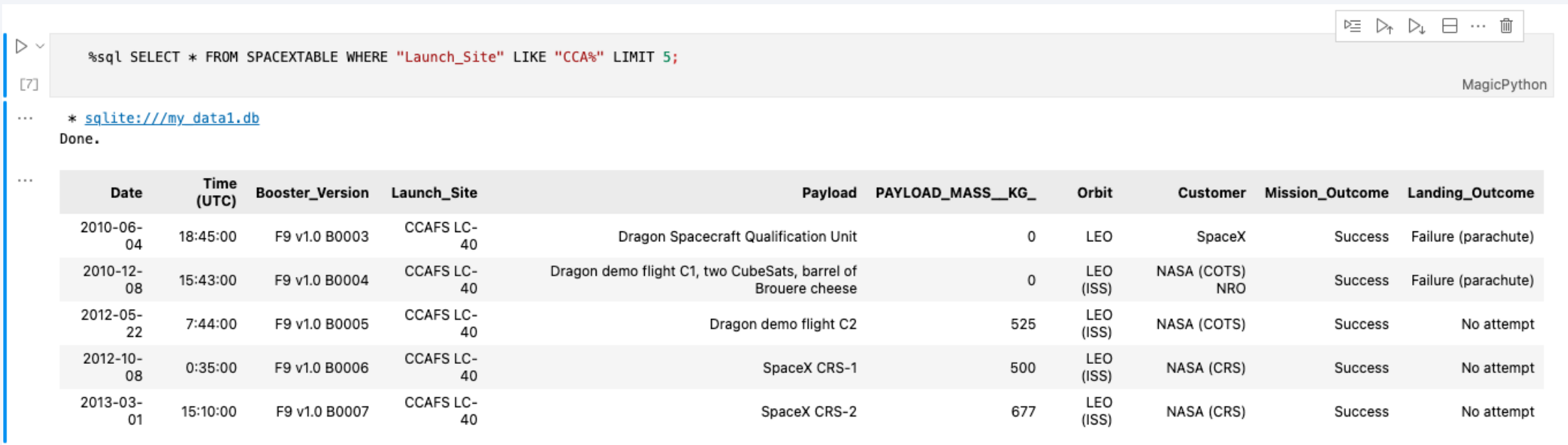
... \* [sqlite:///my\\_data1.db](#)

Done.

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

# Launch Site Names Begin with 'CCA'

- The following query was executed to find 5 records where launch sites begin with 'CCA'



The screenshot shows a Jupyter Notebook interface. At the top, a code cell contains a SQL query: `%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE "CCA%" LIMIT 5;`. Below the code cell, the output shows the connection string `* sqlite:///my_data1.db` and the status `Done.`. Below the output, a table displays the results of the query. The table has 11 columns: Date, Time (UTC), Booster\_Version, Launch\_Site, Payload, PAYLOAD\_MASS\_KG\_, Orbit, Customer, Mission\_Outcome, and Landing\_Outcome. The table contains 5 rows of data, all with Launch\_Site values starting with 'CCA'.

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

# Total Payload Mass

---

- The following query was performed to calculate the total payload carried by boosters from NASA

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" LIKE "NASA (CRS)";
```

[8]

... \* [sqlite:///my\\_data1.db](#)

Done.

... SUM("PAYLOAD\_MASS\_\_KG\_")

45596

# Average Payload Mass by F9 v1.1

---

- The following query was executed to calculate the average payload mass carried by booster version F9 v1.1

```
▷ %sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Booster_Version" LIKE "F9 v1.1%";|
[9]
... * sqlite:///my\_data1.db
Done.
...
  AVG("PAYLOAD_MASS__KG_")
2534.6666666666665
```

# First Successful Ground Landing Date

---

- The following query was performed to find the date of the first successful landing outcome on ground pad using the MIN function. An alternate commented version is also displayed

```
▷ [10] # %sql SELECT "Date" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE "Success (ground pad)" LIMIT 1;

      %sql SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE "Success (ground pad)";

... * sqlite:///my\_data1.db
Done.

... MIN("Date")
    2015-12-22
```



## Successful Drone Ship Landing with Payload between 4000 and 6000

- The query below was executed to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 KG

```
%%sql SELECT "Booster_Version", "PAYLOAD_MASS_KG_" FROM SPACEXTABLE  
WHERE "Landing_Outcome" LIKE "Success (drone ship)"  
AND "PAYLOAD_MASS_KG_" > 4000 AND "PAYLOAD_MASS_KG_" < 6000;
```

[11]

... \* [sqlite:///my\\_data1.db](#)

Done.

...

Booster_Version	PAYLOAD_MASS_KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

# Total Number of Successful and Failure Mission Outcomes

---

- The query below was performed to display the total number of success and failure mission outcomes

```
[13] %%sql SELECT "Mission_Outcome", count("Mission_Outcome") FROM SPACEXTABLE
      GROUP BY ("Mission_Outcome" LIKE "Success%");

... * sqlite:///my\_data1.db
Done.

... 

| Mission_Outcome     | count("Mission_Outcome") |
|---------------------|--------------------------|
| Failure (in flight) | 1                        |
| Success             | 100                      |


```

# Boosters Carried Maximum Payload

- Below query was performed to list the names of the booster versions which have carried the maximum payload mass among all missions

```
[21] %%sql SELECT "Booster_Version", "PAYLOAD_MASS_KG_" FROM SPACEXTABLE
      WHERE "PAYLOAD_MASS_KG_" = (SELECT MAX("PAYLOAD_MASS_KG_") FROM SPACEXTABLE)

... * sqlite:///my\_data1.db
Done.

... 

| 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            |


```

# 2015 Launch Records

---

- The query below listed the month names of failed landing outcomes in drone ship in year 2015, along with their booster versions, and launch site names

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

[32]

... \* [sqlite:///my\\_data1.db](#)

Done.

...

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

- The following query was executed to 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

```
%%sql SELECT "Landing_Outcome", COUNT("Landing_Outcome"), "Date" FROM SPACEXTABLE  
WHERE DATE >= '2010-06-04' AND DATE <= '2017-03-20'  
GROUP BY "Landing_Outcome"  
ORDER BY COUNT("Landing_Outcome") DESC;
```

[40]

... \* [sqlite:///my\\_data1.db](#)  
Done.

...

Landing_Outcome	COUNT("Landing_Outcome")	Date
No attempt	10	2012-05-22
Success (drone ship)	5	2016-04-08
Failure (drone ship)	5	2015-01-10
Success (ground pad)	3	2015-12-22
Controlled (ocean)	3	2014-04-18
Uncontrolled (ocean)	2	2013-09-29
Failure (parachute)	2	2010-06-04
Precluded (drone ship)	1	2015-06-28

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 solid blue background on the left and a satellite photograph of Earth on the right. The Earth's surface is dark blue, with numerous bright yellow and orange lights representing cities and urban areas. The horizon of the Earth is visible as a curved line separating the dark surface from the blackness of space.

Section 3

# Launch Sites Proximities Analysis

# Launch Sites Geographic Locations

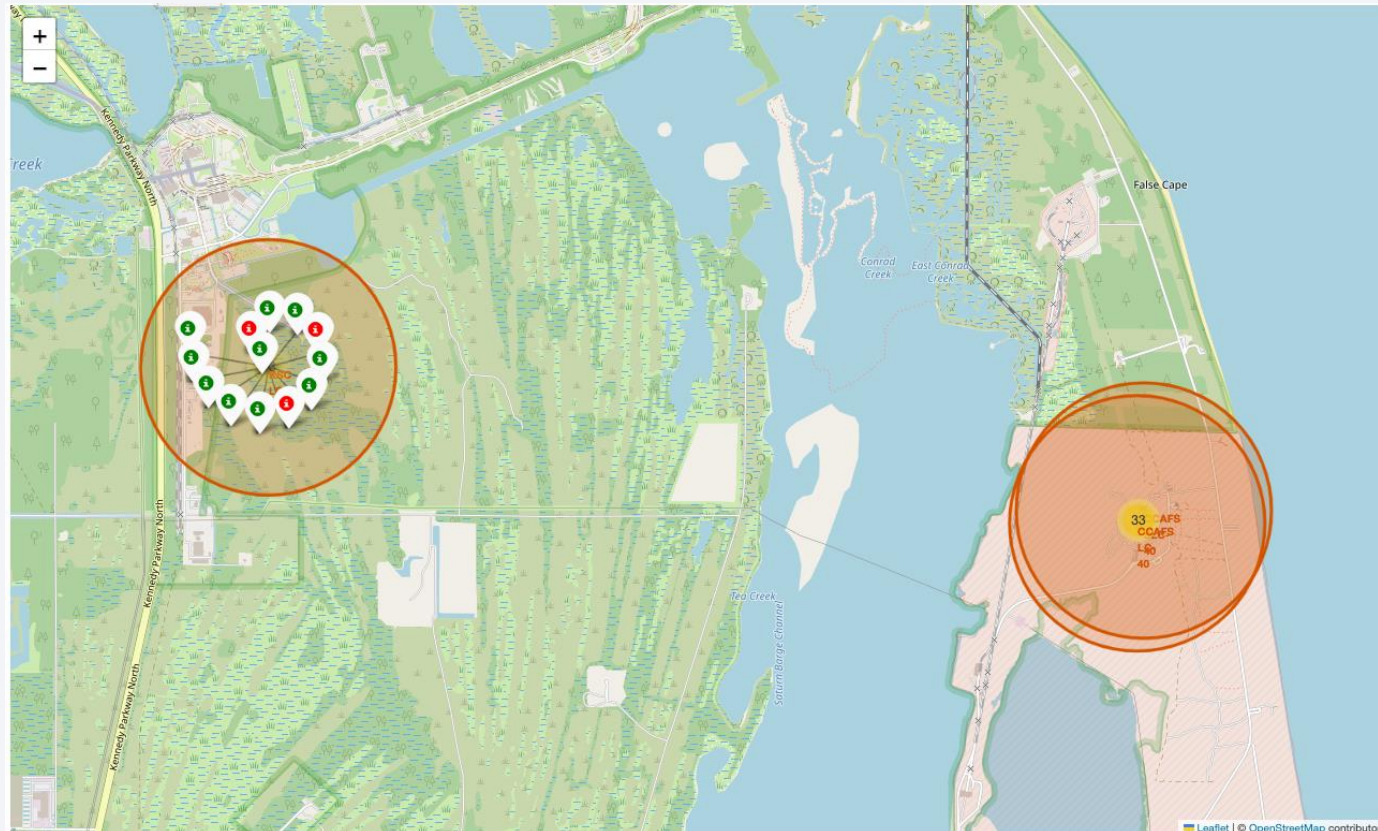
- Below a screenshot of a folium map showing all SpaceX launch site locations. Circle and marker objects were added around each site location to highlight its area and display a text label





# Success vs Failed Launches Cluster Markers

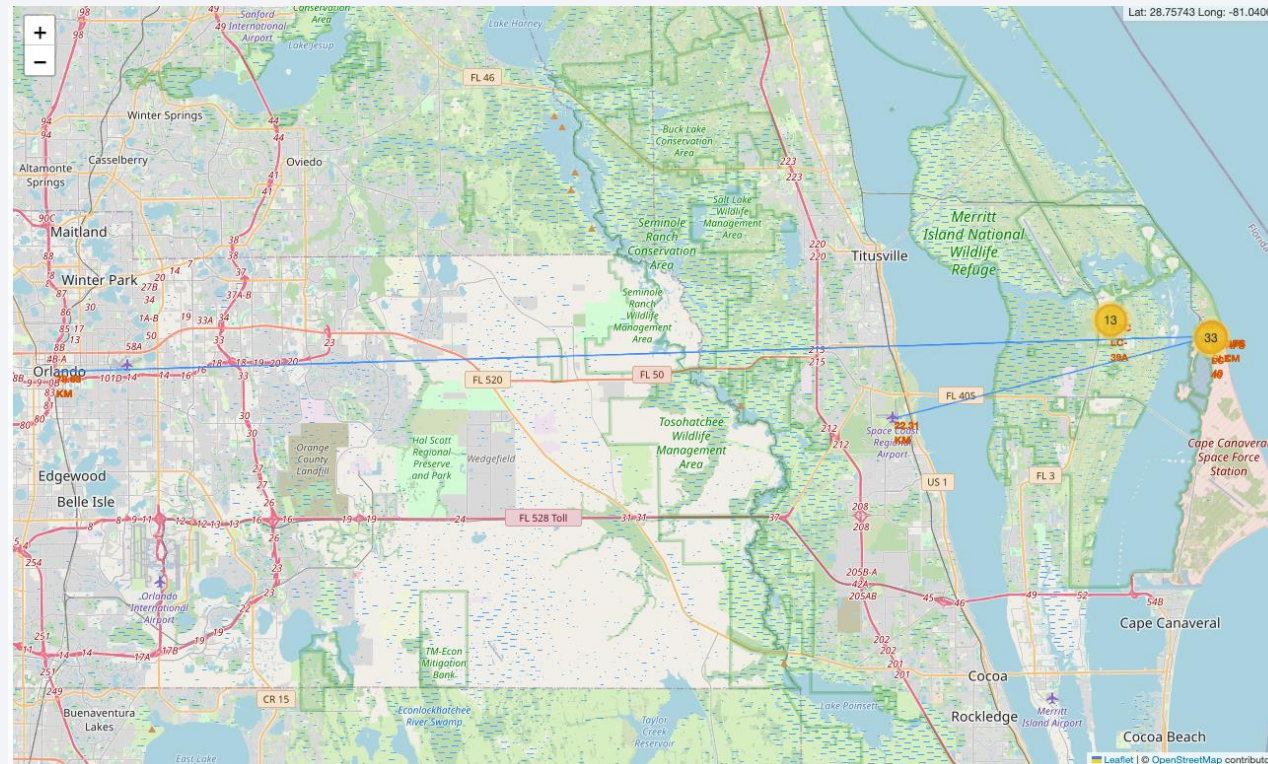
- This Folium screenshot displays a set of marker clusters for every launch site. Each cluster object shows the total number of flights operates and each individual marker is color coded so the success ratio on each launch location can be easily visualized





# Landmarks at Launch Site Proximities

- The screenshot exhibits lines connecting launch sites to nearby landmarks, with markers indicating the distances of these landmarks to the launch locations. It reveals that all operating sites are situated in close proximity to coastlines, with nearby railroads and highways facilitating logistical operations







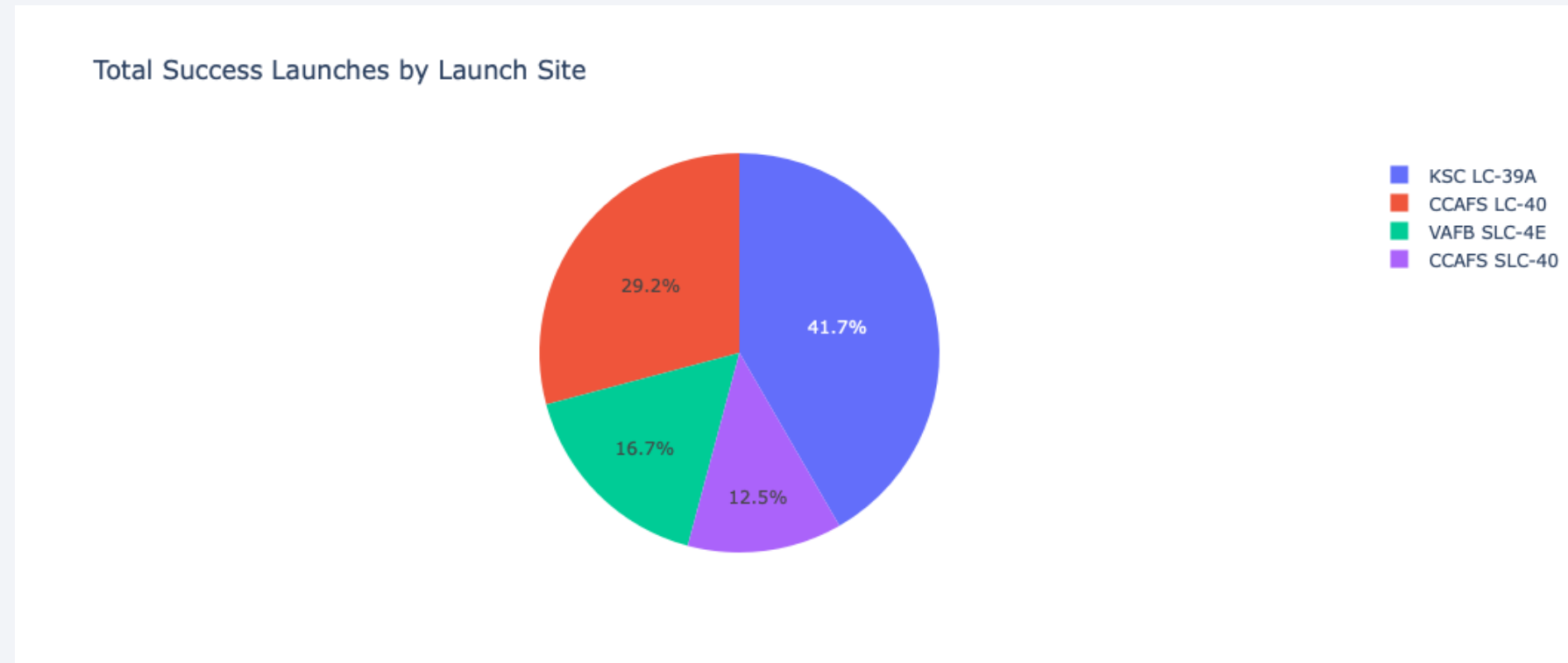
Section 4

# Build a Dashboard with Plotly Dash

# Launch Success Ratio for All Sites

---

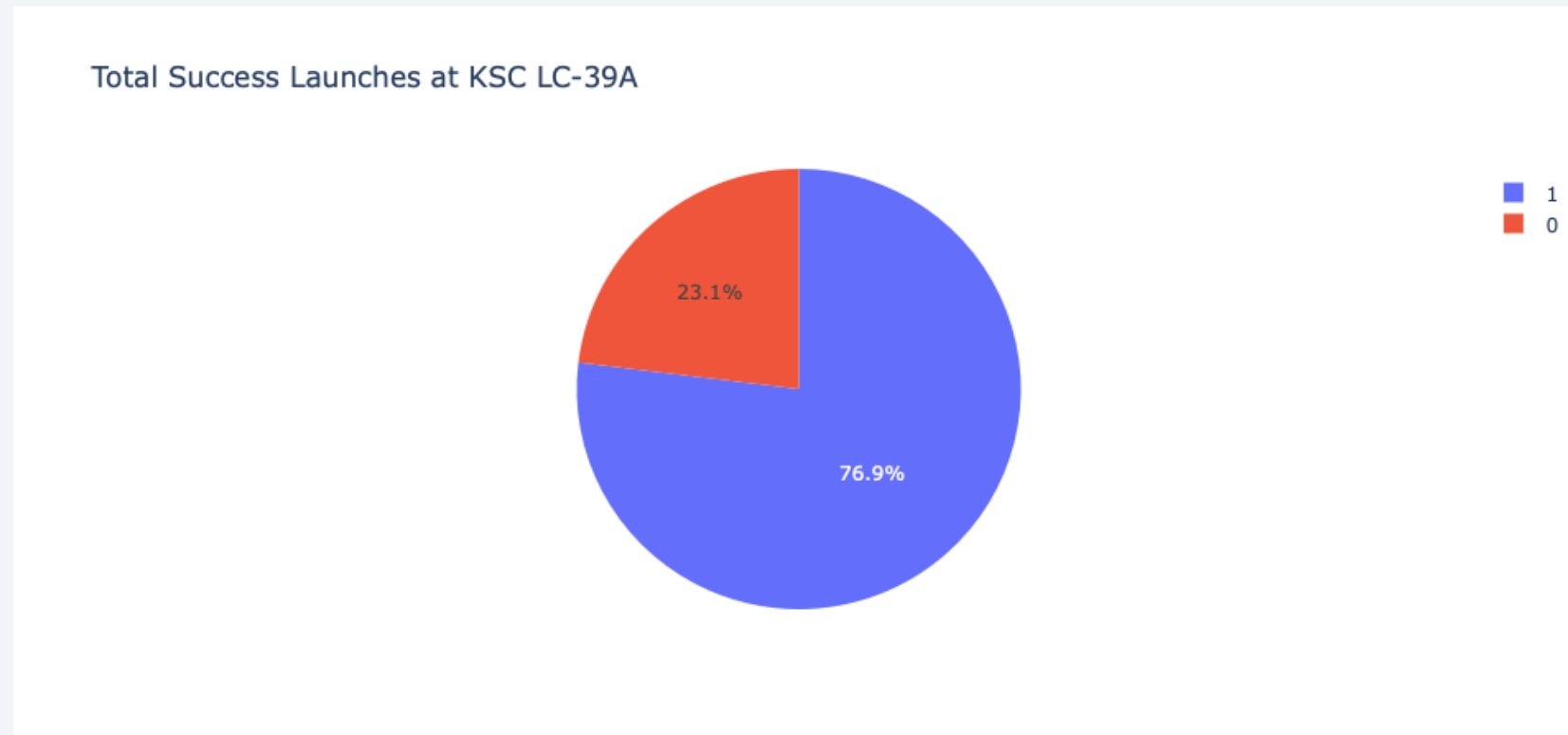
- The image below extracted from the Plotly Dash dashboard and presents a pie chart showcasing the success launch ratio for all launching sites. Notably KSC LC-39A had the most successful launch ratio while CCAFS SLC-40 had the least



# Best Success Ration Launch Site

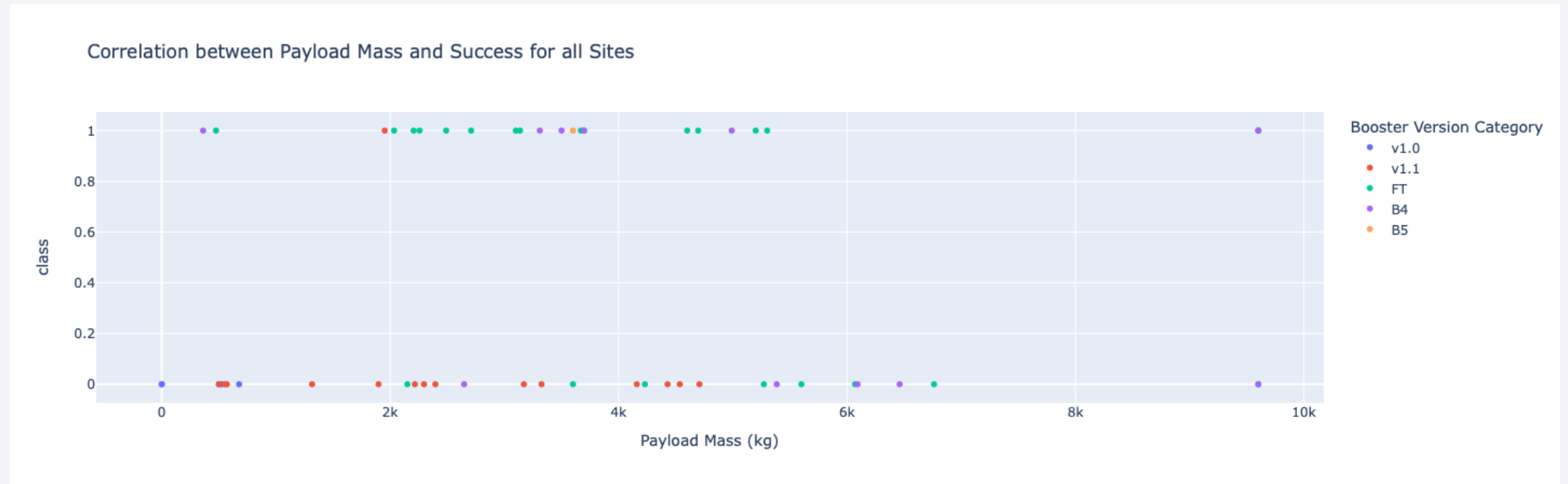
---

- The image below extracted from the Plotly Dash dashboard and presents a pie chart highlighting the highest success launch ratio for site KSC LC-39A, boasting an impressive a 76.9% success rate



# Payload vs Launch Outcome for All Sites

- The scatterplot below shows the correlation between payload mass and success outcome. The datapoints are color coded so different booster version can be easily identified. It shows that success rate is higher for smaller payloads (up to 4000KG). For heavier payloads the success rate seems to decrease significantly



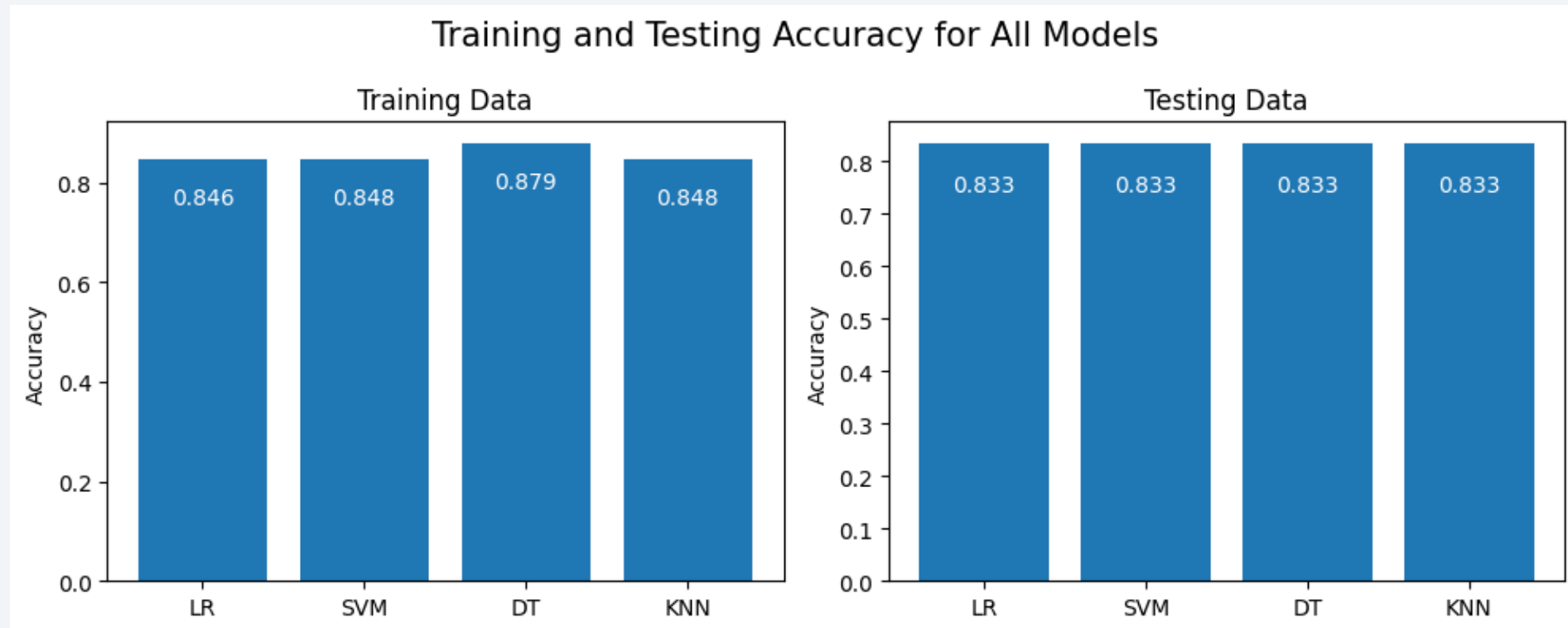
Section 5

# Predictive Analysis (Classification)



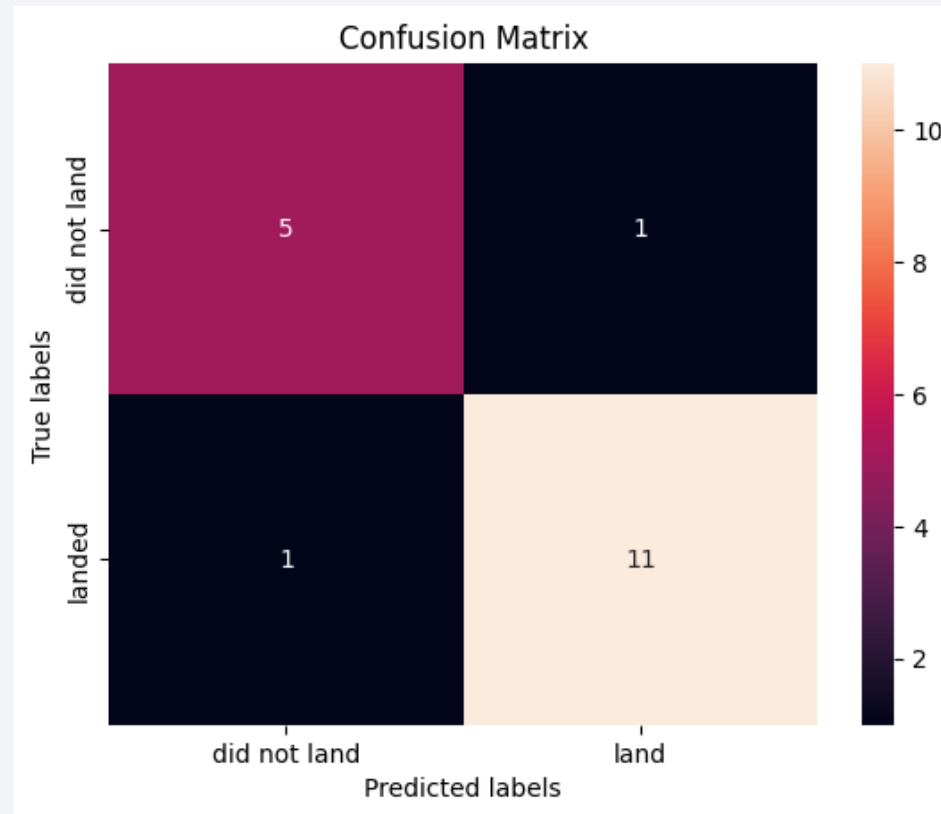
# Classification Accuracy

- As illustrated in the bar charts below, the Decision Tree model showcased the best training accuracy among all models. Interestingly, when subjected to test data, all models demonstrated identical performance outcomes.



# Confusion Matrix

- The Decision Tree model stood out as the best performer, achieving a training accuracy of 0.8892 and a testing score of 0.8888. The confusion matrix below shows that the model experienced only one false positive and one false negative outcome





# Conclusions

---

- Success rates have consistently risen over the years, supported by accumulated experience
- KSC LC-39A had the best success ratio and boasted an impressive a 76.9% success outcome
- Orbits ES-L1, GEO, HEO and SSO had flawless success rates
- Success rates were higher for missions with a payload mass below 4000 KG
- The Decision Tree stood out as the best classifier algorithm for training a model for this dataset

Thank you!

