IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Muhammad Daffa Hilmy
11-October-2023
https://github.com/Dahgorago/IBM-Data-Science

Outline

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

Executive Summary

- Data was acquired from the public SpaceX API and the SpaceX Wikipedia page. A 'class' column was added to categorize successful landings. The data was analyzed using SQL, visualization techniques, folium maps, and dashboards. Relevant columns were selected as features, and categorical variables were converted to binary using one hot encoding. The data was standardized, and GridSearchCV was utilized to determine the optimal parameters for machine learning models. The accuracy scores of all models were visualized.

- Four machine learning models were generated: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. These models yielded comparable results, achieving an accuracy rate of approximately 83.33%. However, they tended to over-predict successful landings. Additional data is necessary to enhance model accuracy and make more informed determinations.

Introduction

Project Background

- The era of commercial space exploration has arrived.
- SpaceX offers the most competitive pricing at $62 million compared to $165 million USD.
- This is mainly attributed to their capability to recover a portion of the rocket (Stage 1).
- Space Y aims to rival SpaceX in the space industry.

Problem Statement

- Space Y has assigned us the responsibility of training a machine learning model to anticipate the success of Stage 1 rocket recovery.

Section 1

Methodology

Methodology

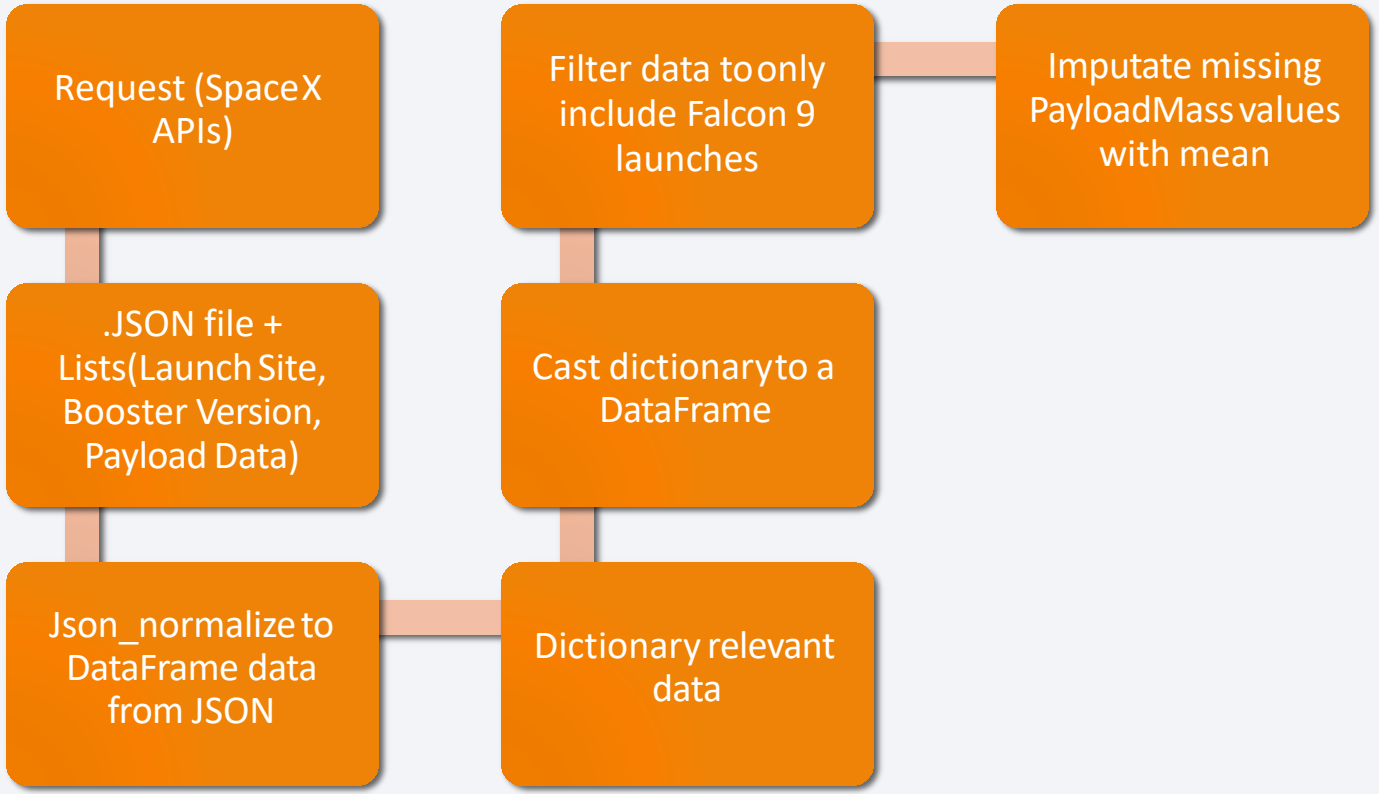
Executive Summary

- Data collection methodology:
  - Data was gathered from SpaceX's public API and web scraping a specific table in SpaceX's Wikipedia page, encompassing critical columns needed to train machine learning models for predicting the success of Stage 1 rocket recovery, aligning with Space Y's goals..
- Perform data wrangling
  - Generating a new training label, 'class', based on successful and failed landing outcomes, where 'class' is set to 1 if 'Mission Outcome' is true and the landing location is ASDS, RTLS, or Ocean, and 0 for other cases.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Predictive analysis utilizing classification techniques, including Logistic Regression, Support Vector Machine, Decision Tree, and K Nearest Neighbors.

Data Collection

- The process of gathering data involved utilizing a combination of API requests from SpaceX's public API and extracting information from a table within SpaceX's Wikipedia entry through web scraping.
- In the upcoming slide, there will be a depiction of the flowchart illustrating the data collection procedure from the API, followed by another slide illustrating the flowchart for data collection through web scraping.
- The data columns retrieved from the SpaceX API encompass FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.
- On the other hand, the data columns acquired through web scraping from Wikipedia consist of Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.

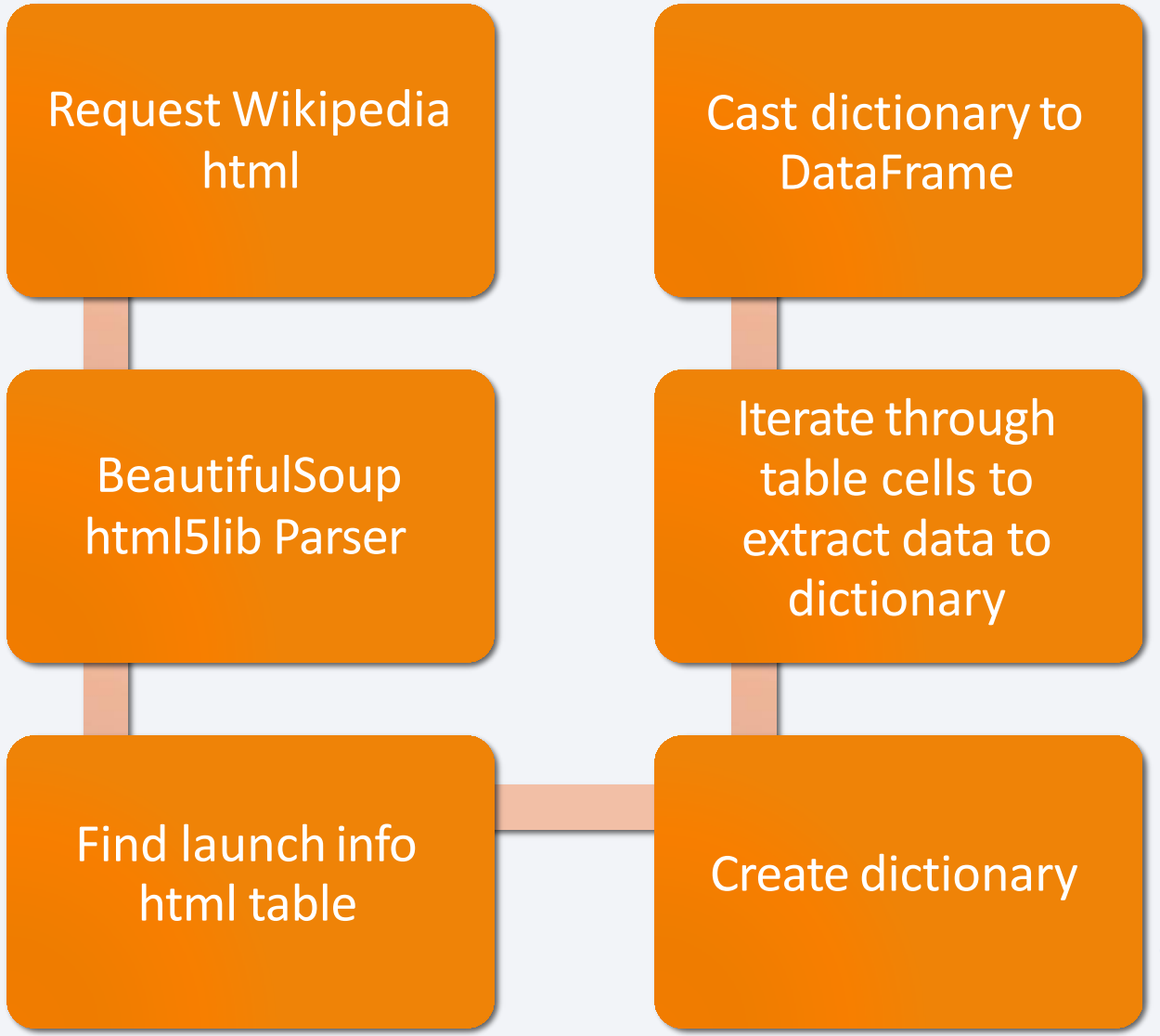
Data Collection – SpaceX API



- Github URL : https://github.com/Dahgorago/IBM-Data-Science/blob/main/Data%20Science%20Capstone%20Project/1.%20Collecting%20The%20Data/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Github URL : https://github.com/Dahgorago/IBM-Data-Science/blob/main/Data%20Science%20Capstone%20Project/1.%20Collecting%20The%20Data/jupyter-labs-webscraping.ipynb



Data Wrangling

Generate a training label that signifies successful landings as 1 and unsuccessful ones as 0. The label comprises two elements: 'Mission Outcome' and 'Landing Location.' Introduce a new 'class' column in the training label, assigning a value of 1 if 'Mission Outcome' is confirmed and 0 otherwise. Assign the following value mappings:

'True' for ASDS, RTLS, and Ocean in 'Mission Outcome' to be set as 1.

'None' for both components or 'False' for ASDS, Ocean, and RTLS in 'Mission Outcome' to be set as 0.

Github URL : https://github.com/Dahgorago/IBM-Data-Science/blob/main/Data%20Science%20Capstone%20Project/2.%20Data%20Wrangling/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

Conducted exploratory analysis on Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year variables.

Utilized various plots for comparison, including Scatter plots, line charts, and bar plots, to assess relationships between Flight Number, Payload Mass, Launch Site, Orbit, and Success Rate, determining their relevance for potential integration into the machine learning model during training.

Github URL : https://github.com/Dahgorago/IBM-Data-Science/blob/main/Data%20Science%20Capstone%20Project/3.%20Exploratory%20Data%20Analysis/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Uploaded the dataset into an IBM DB2 Database.
- Utilized SQL Python integration for querying.
- Executed queries to enhance comprehension of the dataset, focusing on obtaining details like launch site names, mission outcomes, diverse customer payload sizes, and booster versions, along with landing outcomes.

Github URL : https://github.com/Dahgorago/IBM-Data-Science/blob/main/Data%20Science%20Capstone%20Project/3.%20Exploratory%20Data%20Analysis/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Folium maps display Launch Sites, both successful and unsuccessful landings, and offer a proximity illustration to significant areas like Railway, Highway, Coast, and City.
- This visualization aids in comprehending the rationale behind the placement of launch sites and provides a clear view of successful landings concerning their respective locations.

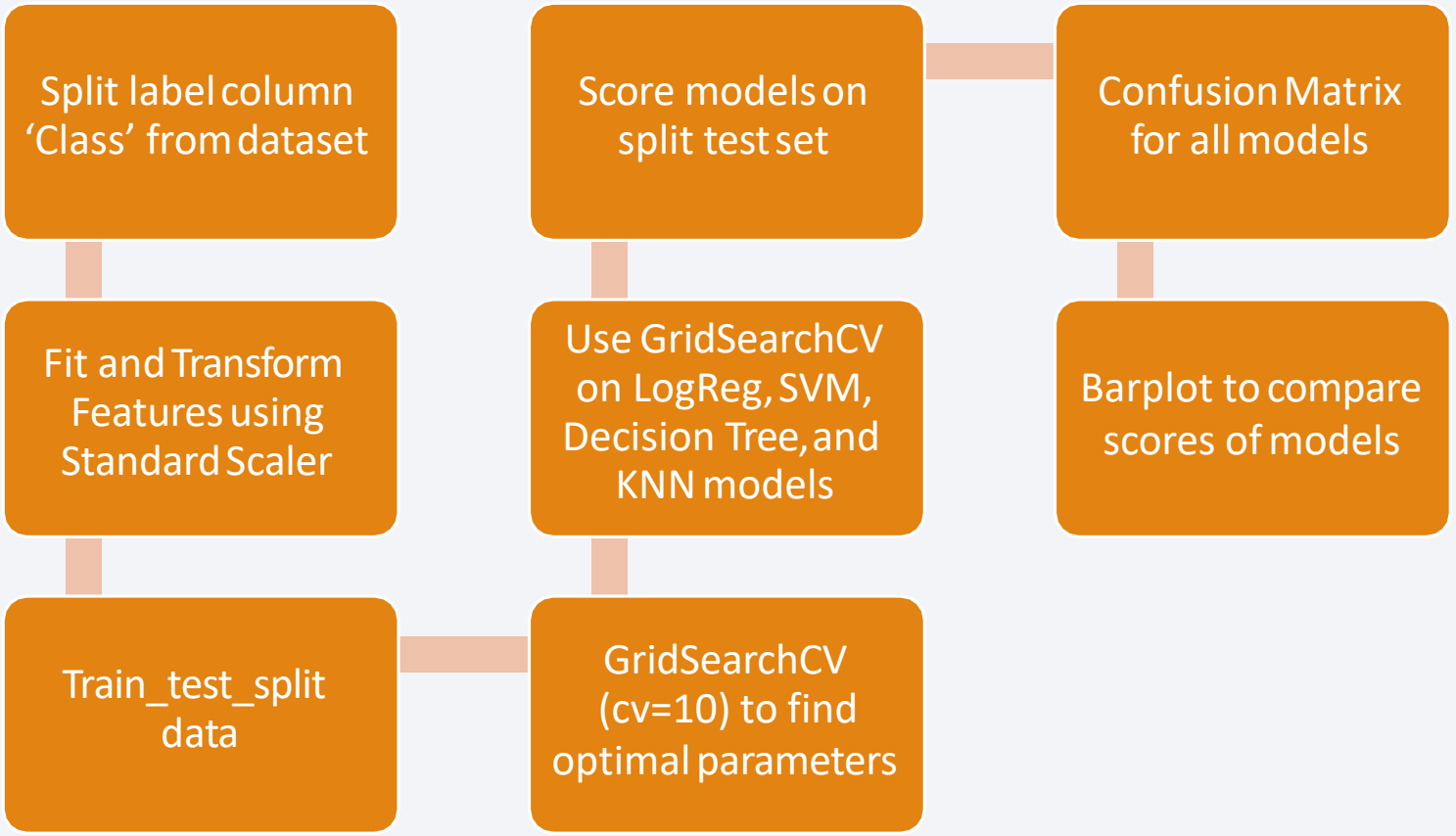
- Github URL : https://github.com/Dahgorago/IBM-Data-Science/blob/main/Data%20Science%20Capstone%20Project/4.%20Interactive%20Visual%20Analytics/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- The dashboard features both a pie chart and a scatter plot.
- The pie chart provides options to display the distribution of successful landings across all launch sites or the success rates of individual launch sites.
- The scatter plot allows users to select either all sites or an individual site and adjust the payload mass using a slider ranging from 0 to 10000 kg.
- The pie chart serves as a visualization tool for presenting launch site success rates.
- The scatter plot is useful in observing the variation in success rates across launch sites, payload masses, and booster version categories.

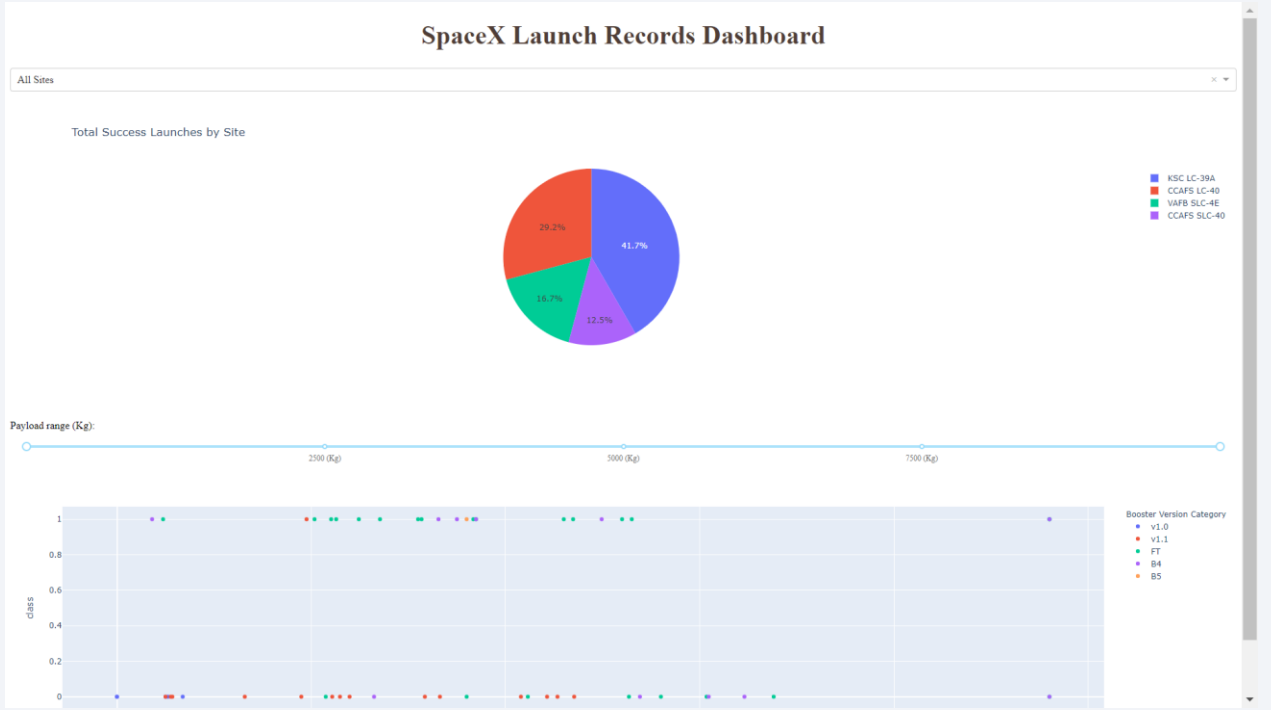
- Github URL : https://github.com/Dahgorago/IBM-Data-Science/blob/main/Data%20Science%20Capstone%20Project/4.%20Interactive%20Visual%20Analytics/spacex_dash_app.py

Predictive Analysis (Classification)

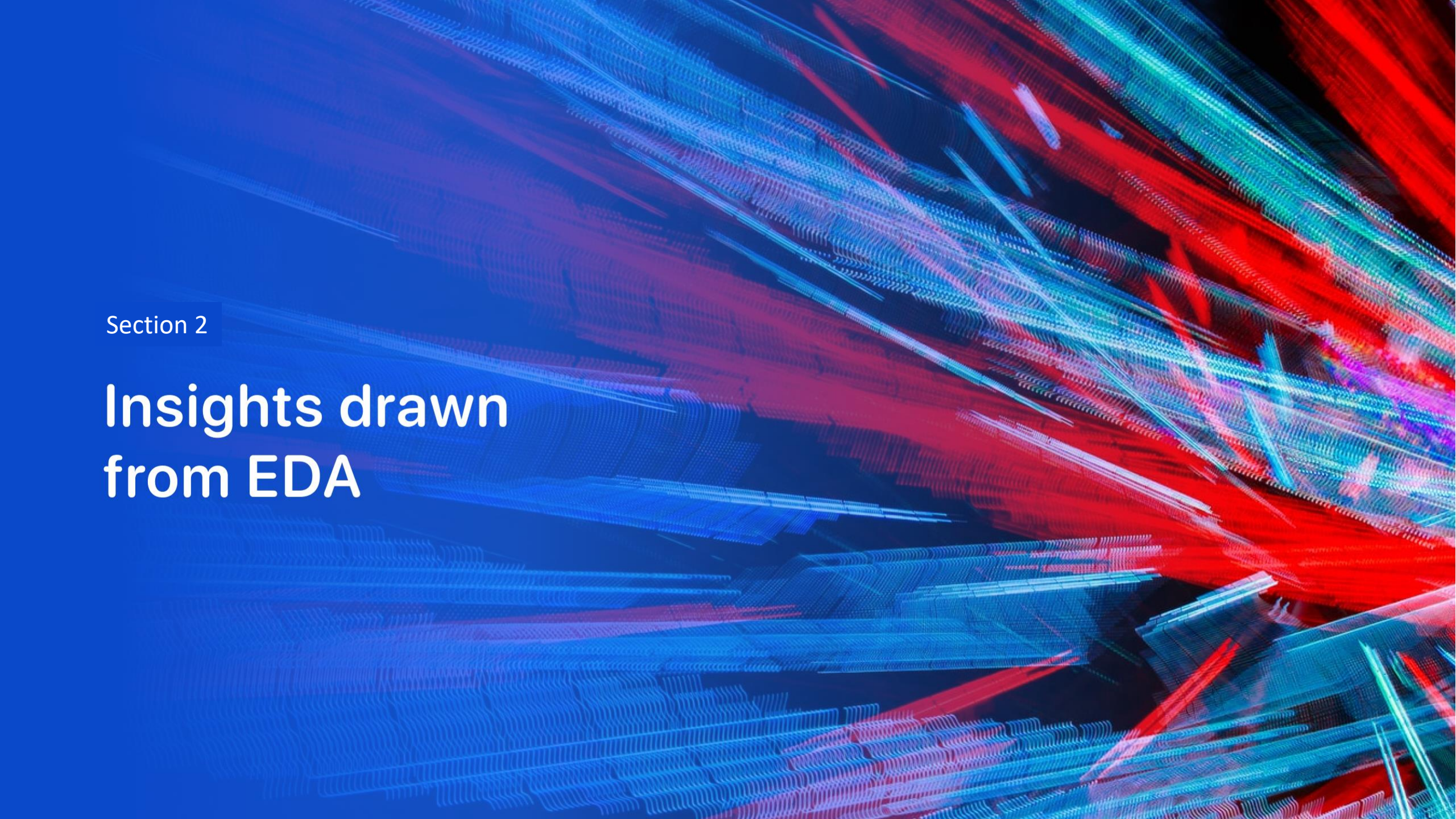


- Github URL : https://github.com/Dahgorago/IBM-Data-Science/blob/main/Data%20Science%20Capstone%20Project/4.%20Interactive%20Visual%20Analytics/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

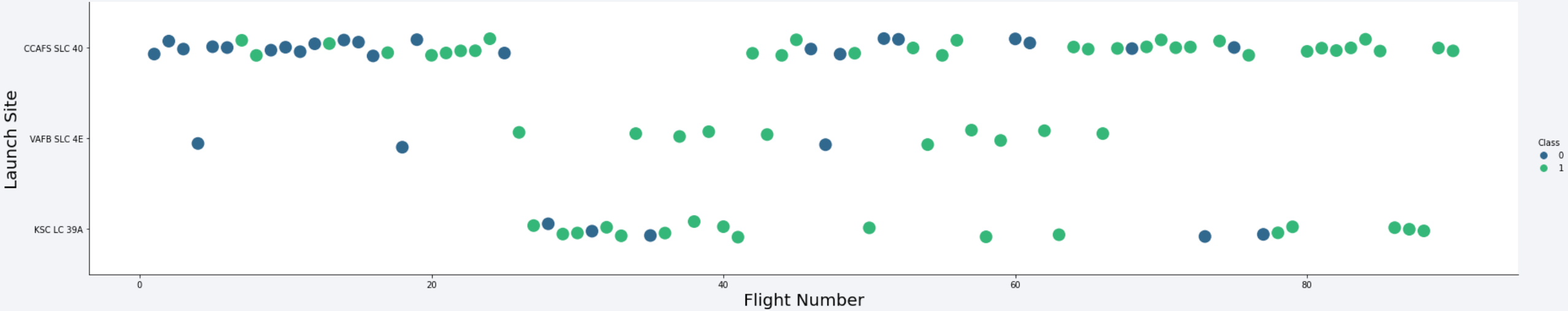
Results



Here's a glimpse of the Plotly dashboard. The subsequent sections will present the outcomes of Exploratory Data Analysis (EDA) using visualization, EDA utilizing SQL, an Interactive Map using Folium, and ultimately, our model results showcasing an accuracy of approximately 83%.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



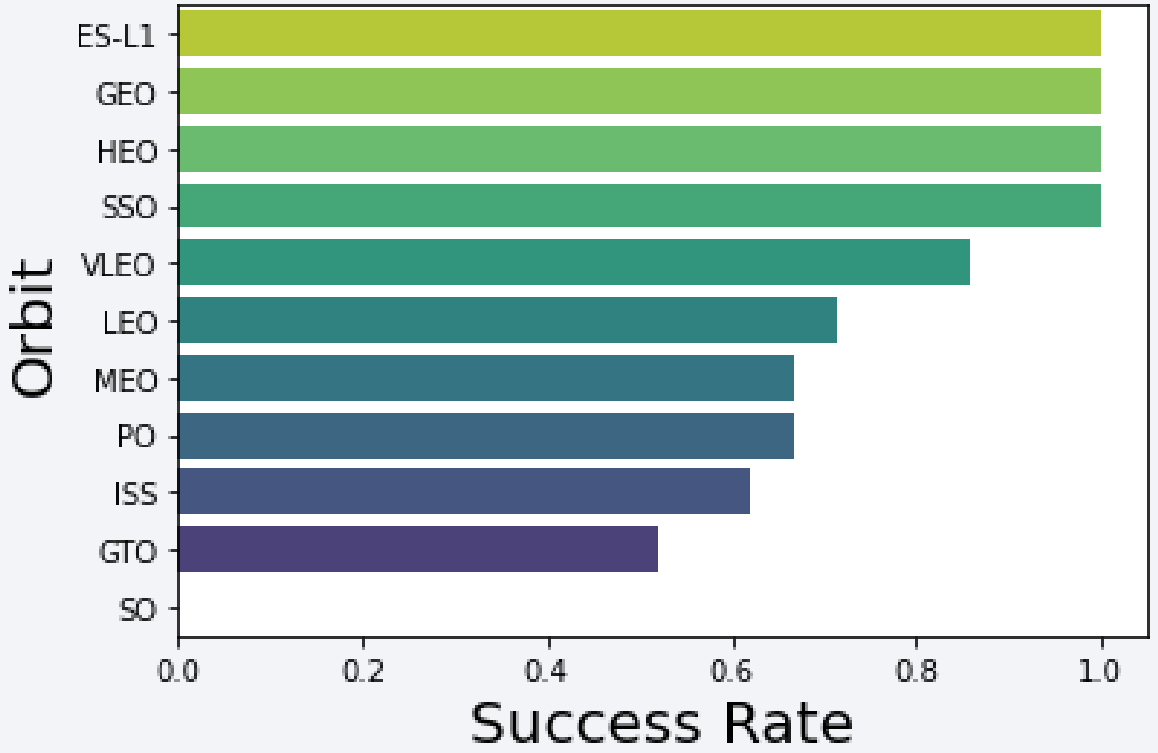
- Green indicates successful launch; Purple indicates unsuccessful launch.
- The visual indicates a rise in the success rate as Flight Number progresses. There seems to be a significant improvement in success rate around flight 20, marking a potential breakthrough. CCAFS seems to dominate as the primary launch site, displaying the highest volume.

Payload vs. Launch Site

- Green indicates successful launch; Purple indicates unsuccessful launch.
- The payload mass predominantly ranges from 0 to 6000 kg. Additionally, distinct launch sites seem to utilize varying payload masses.

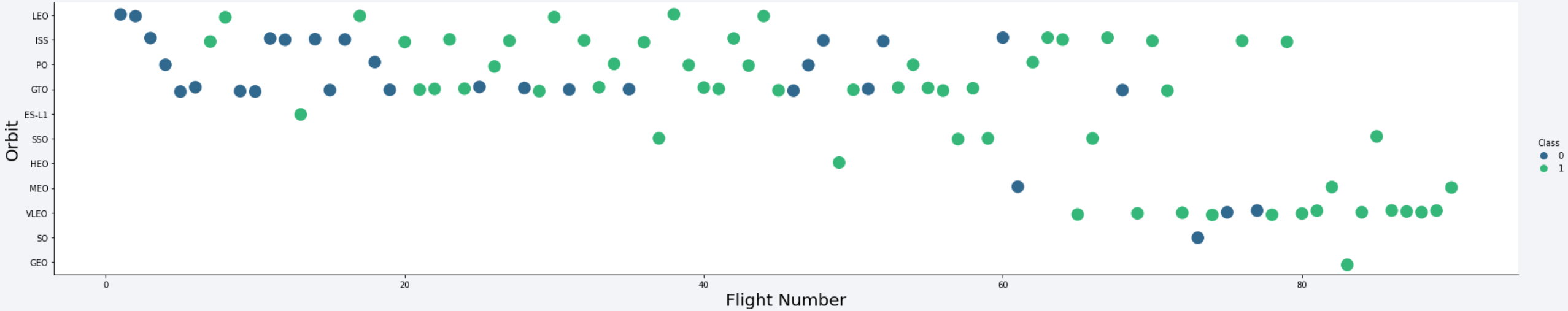
Success Rate vs. Orbit Type

Success Rate Scale with
0 as 0%
0.6 as 60%
1 as 100%



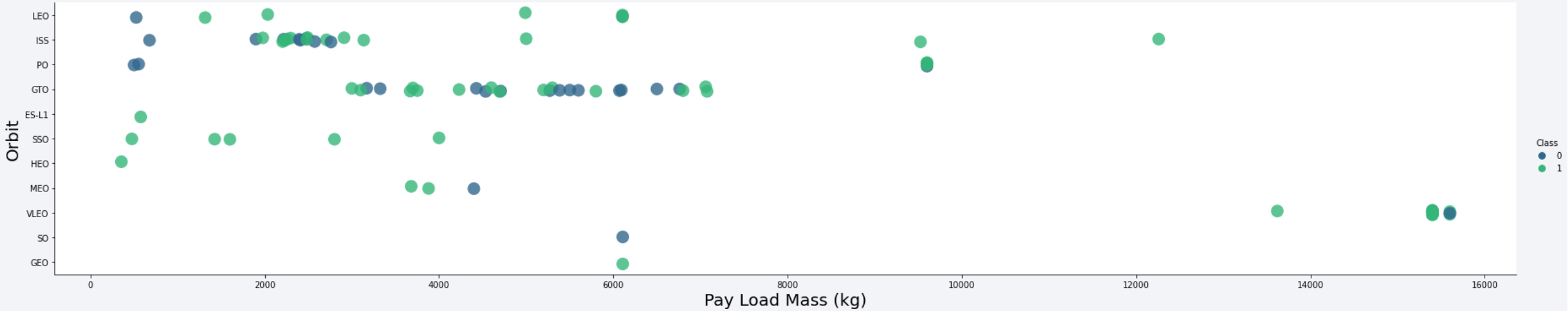
- ES-L1 (1), GEO (1), HEO (1) exhibit a perfect success rate (sample sizes in brackets).
- SSO (5) demonstrates a flawless success rate.
- VLEO (14) showcases a respectable success rate along with a notable number of attempts.
- SO (1) records a 0% success rate.
- GTO (27) reflects an approximate 50% success rate with the largest sample size.

Flight Number vs. Orbit Type



- Green indicates successful launch; Purple indicates unsuccessful launch.
- The preferences for launch orbits shifted with varying flight numbers.
- The launch outcomes appear to align with these changing preferences.
- SpaceX initially opted for LEO orbits, experiencing moderate success, then transitioned to VLEO in recent launches. It seems that SpaceX achieves better performance in lower orbits or Sun-synchronous orbits.

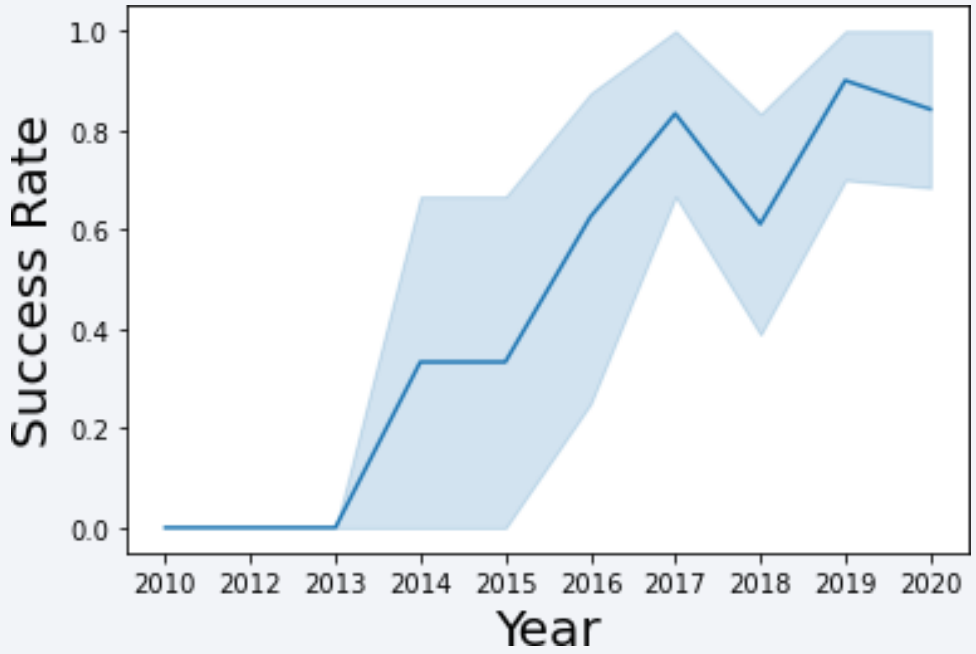
Payload vs. Orbit Type



- Green indicates successful launch; Purple indicates unsuccessful launch.
- There appears to be a relationship between payload mass and the orbit.
- LEO and SSO exhibit lower payload masses.
- On the other hand, VLEO, one of the most successful orbits, predominantly has payload mass values at the higher range.

Launch Success Yearly Trend

95% confidence interval
(light blue shading)



- Overall success rates have shown a consistent upward trend since 2013, with a minor decline in 2018.
- Recent success rates have been approximately 80%.

All Launch Site Names

Retrieve distinct launch site names from the database.

It's probable that CCAFS SLC-40 and CCAFSSLC-40 correspond to the same launch site due to data entry inaccuracies.

The previous name for the site was CCAFS LC-40.

There are likely only three unique values for the launch site: CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E.

```
In [4]: %%sql
        SELECT UNIQUE LAUNCH_SITE
        FROM SPACEXDATASET;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2
Done.
```

Out[4]:

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

Launch Site Names Begin with 'CCA'

In [5]:
```
%%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.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

- The initial five records in the database featuring a Launch Site name starting with "CCA."

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

sum_payload_mass_kg
45596

- The mentioned query calculates the aggregate mass of payload in kilograms for cases where NASA was the customer.
- CRS, an acronym for Commercial Resupply Services, signifies that these payloads were dispatched to the International Space Station (ISS).

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-8
Done.
```

avg_payload_mass_kg
2928

- This query computes the mean payload mass for launches employing the booster variant F9 v1.1.
- The average payload mass for F9 1.1 falls at the lower limit of our payload mass spectrum.

First Successful Ground Landing Date

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

first_success
2015-12-22

- The query provides the date of the initial successful ground pad landing.
- The inaugural successful ground pad landing occurred towards the conclusion of 2015.
- Successful landings in a broader sense commenced in 2014.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

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

- The query lists the four booster variants with successful drone ship landings and a payload mass ranging from 4000 to 6000 (exclusive).

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-
Done.
```

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

- The query provides a tally for each mission outcome.
- SpaceX demonstrates a mission success rate of almost 99%.
- Notably, the majority of unsuccessful landings are deliberate.
- Intriguingly, one launch has an undetermined payload status, and regrettably, one mission ended in a flight failure.

Boosters Carried Maximum Payload

- The query identifies the booster versions that transported the maximum payload mass of 15600 kg.
- These booster versions share a strong resemblance, all belonging to the F9 B5 B10xx.x variety.
- This strongly suggests a correlation between the booster version and the payload mass being transported.

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
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

```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

- The query provides details for 2015 launches where stage 1 failed to land on a drone ship, including the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch Site. There were two instances of such occurrences in 2015.

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

```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

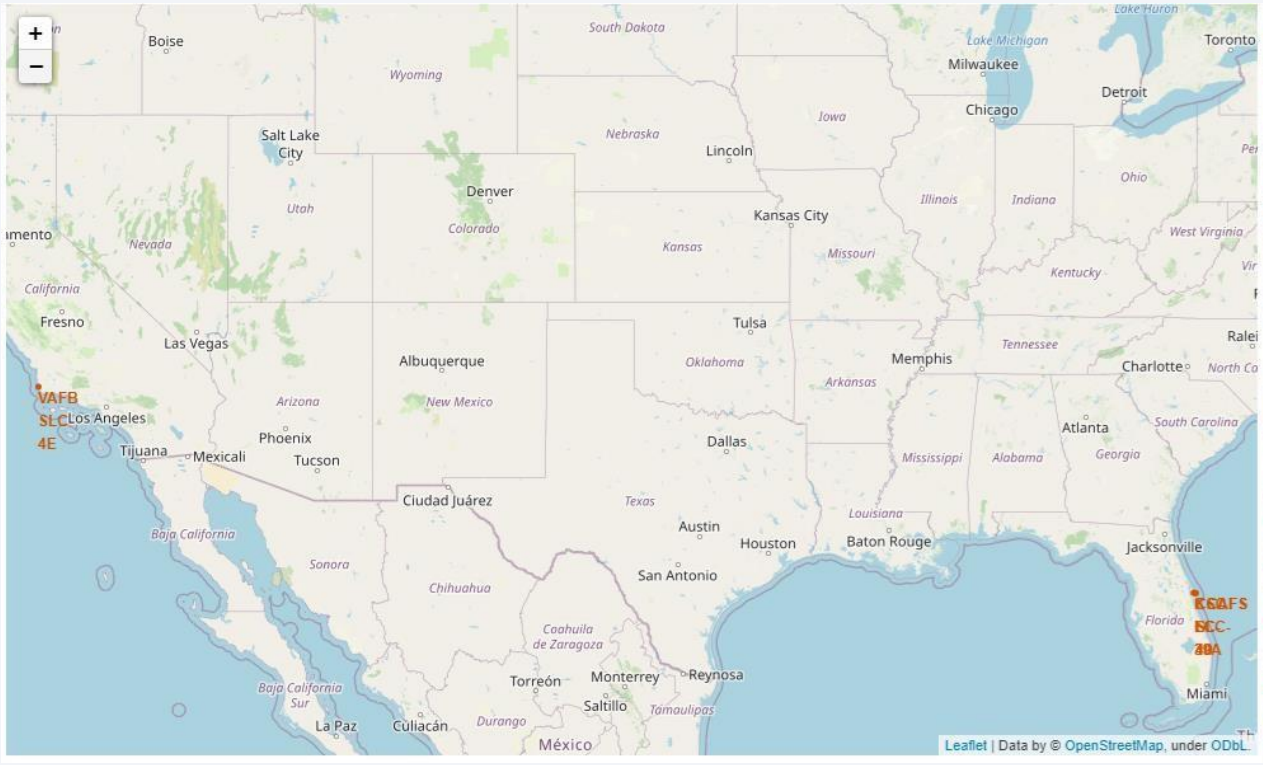
```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

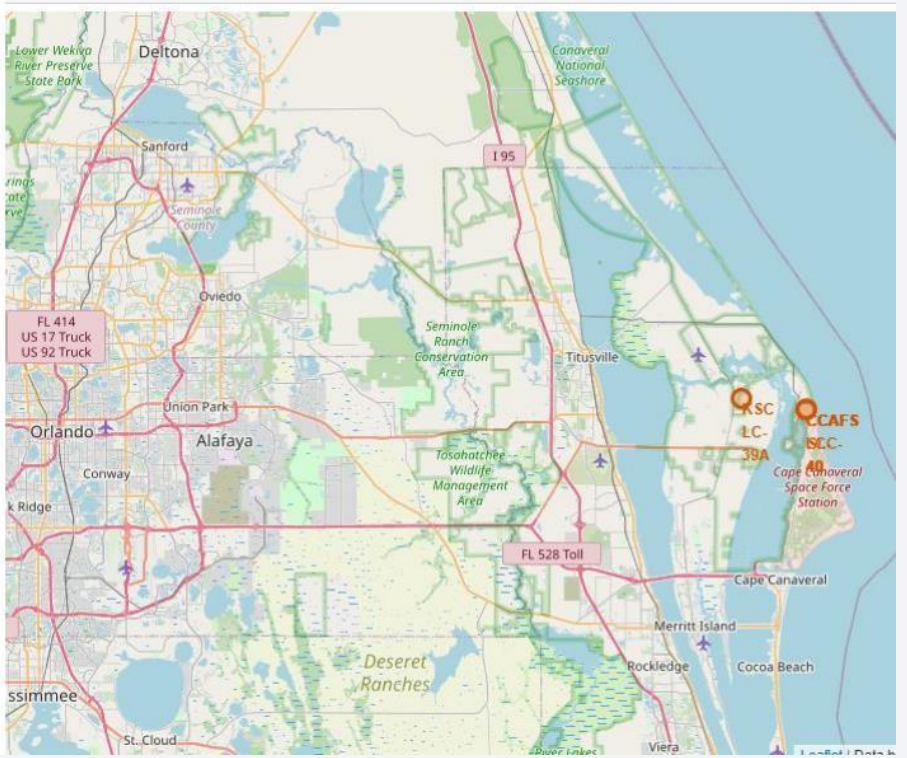
landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

- The query generates a list of successful landings between June 4, 2010, and March 20, 2017, inclusive. There are two types of successful landing outcomes: drone ship and ground pad landings. In total, there were 8 successful landings during this specified time frame.


Section 3
Launch Sites
Proximities Analysis

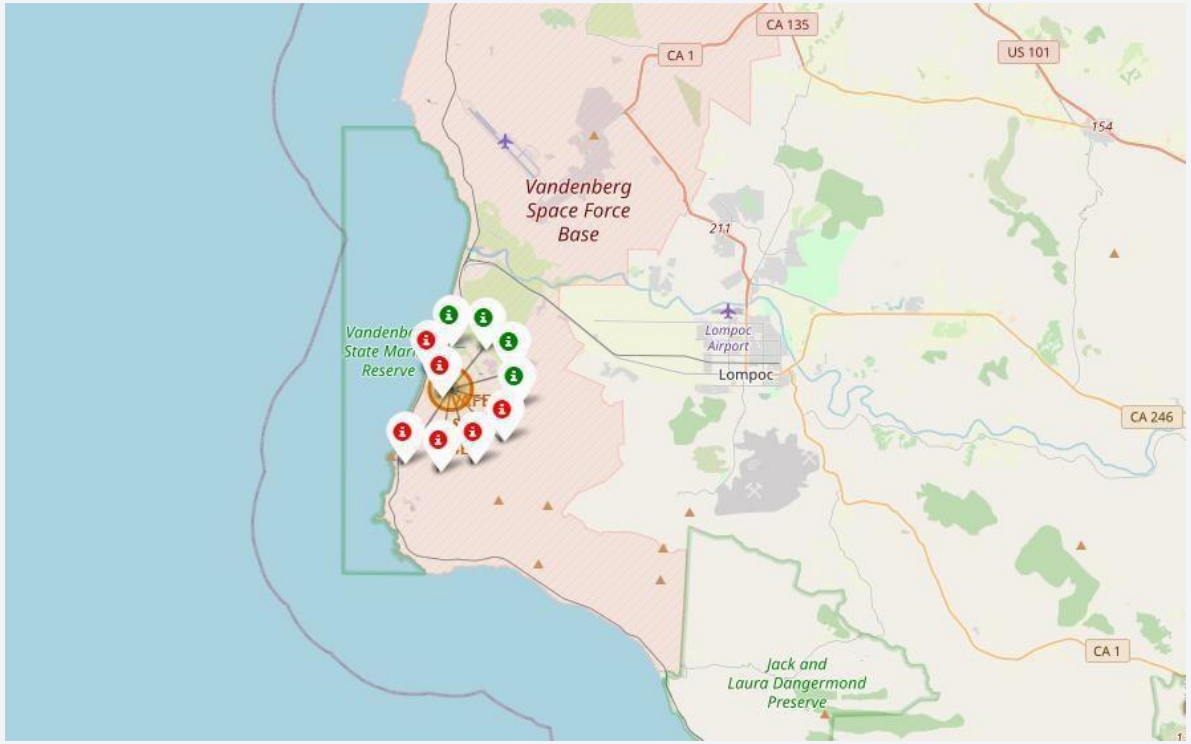
Launch Site Locations





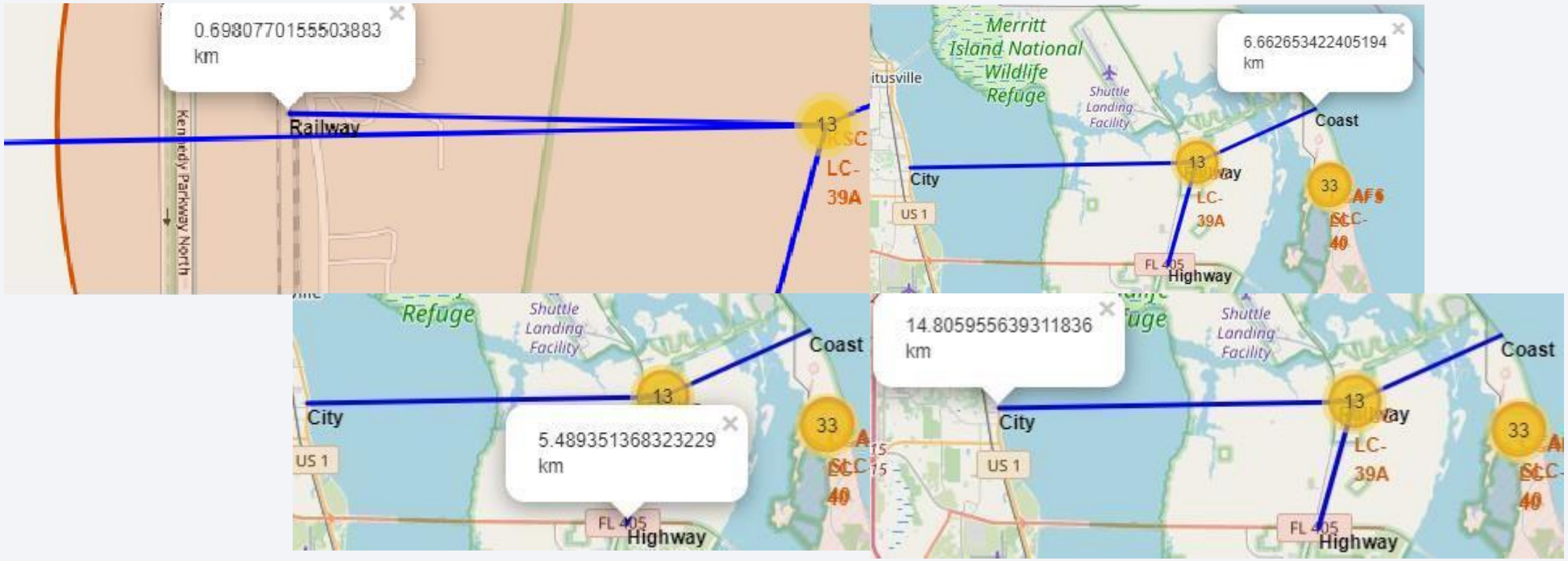
- The left map displays all launch sites in relation to the US map, illustrating their geographical positions.
- The right map specifically highlights the two launch sites in Florida due to their close proximity. Both of these launch sites are located near the ocean.

Color-Coded Launch Markers

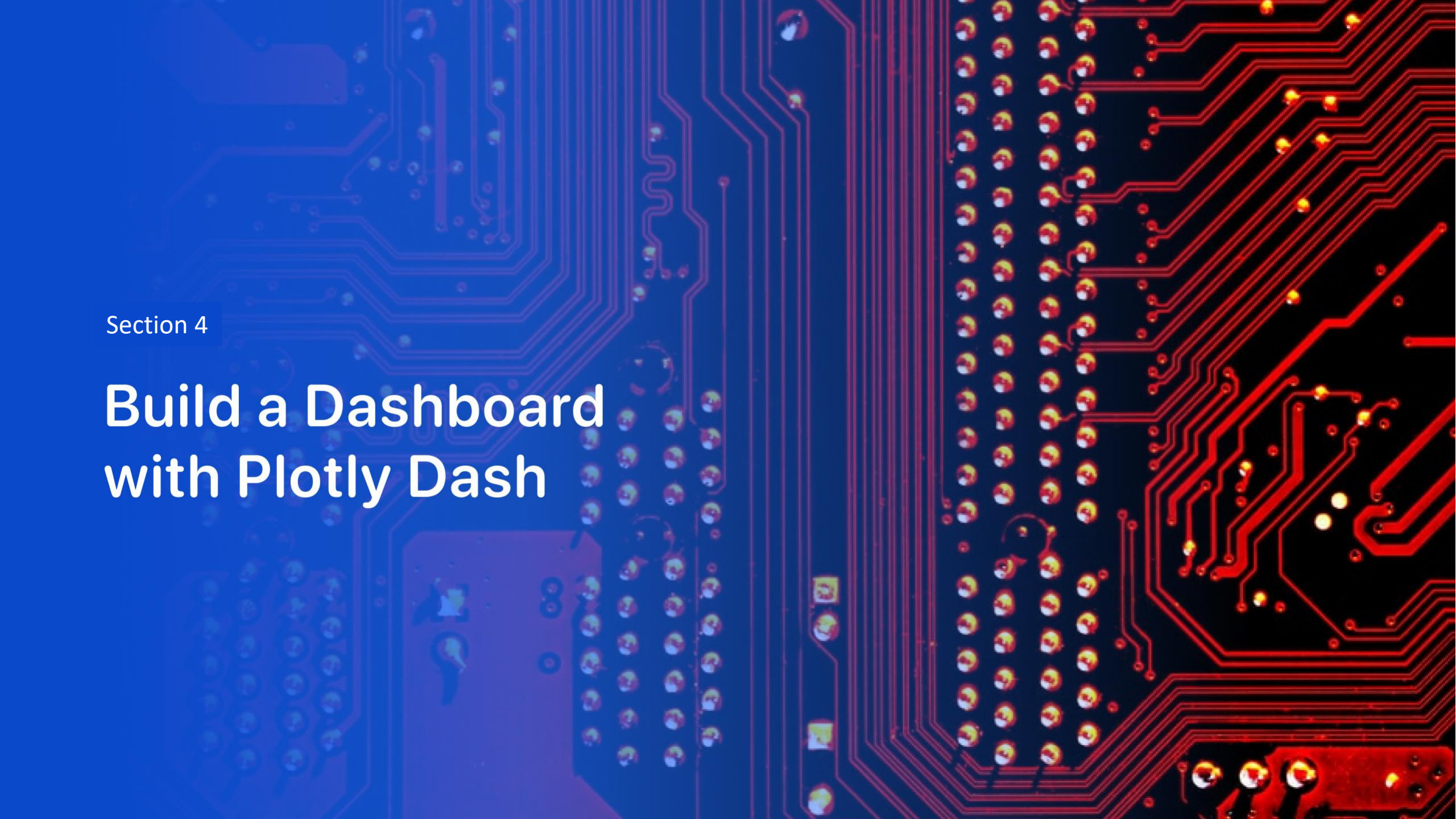


- The Folium map features clickable clusters, each representing successful landings (green icon) and failed landings (red icon). For instance, at VAFB SLC-4E, there have been 4 successful landings and 6 failed landings, as indicated by the respective icons.

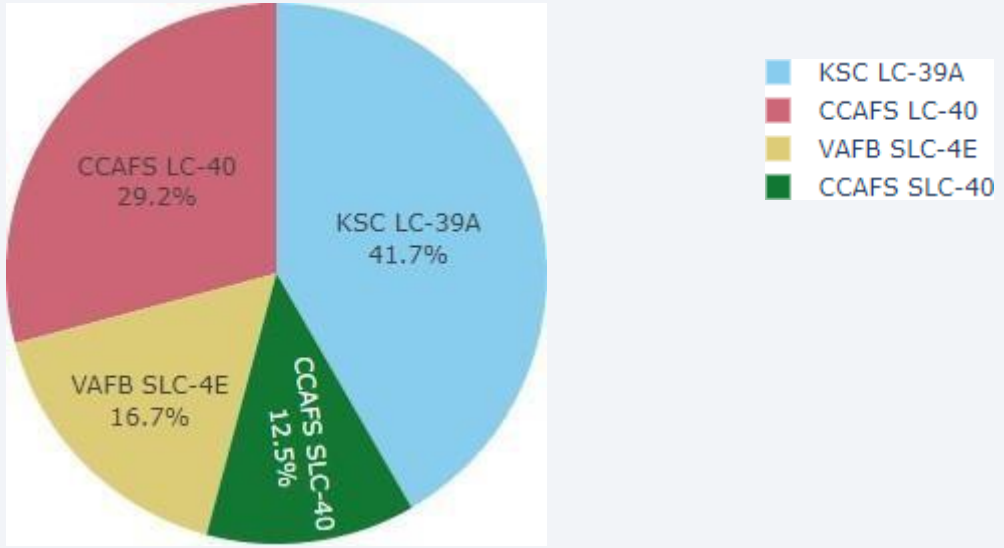
Key Location Proximities



- Taking KSC LC-39A as an illustration, launch sites are strategically positioned in close proximity to railways for efficient transportation of supplies. Additionally, they are situated near highways to facilitate both human transportation and the transport of necessary supplies. Moreover, launch sites are strategically located near coastlines, minimizing risks associated with launch failures by allowing rockets to land in the sea rather than densely populated areas. Furthermore, these sites are generally situated at a distance from heavily populated cities to enhance safety measures.

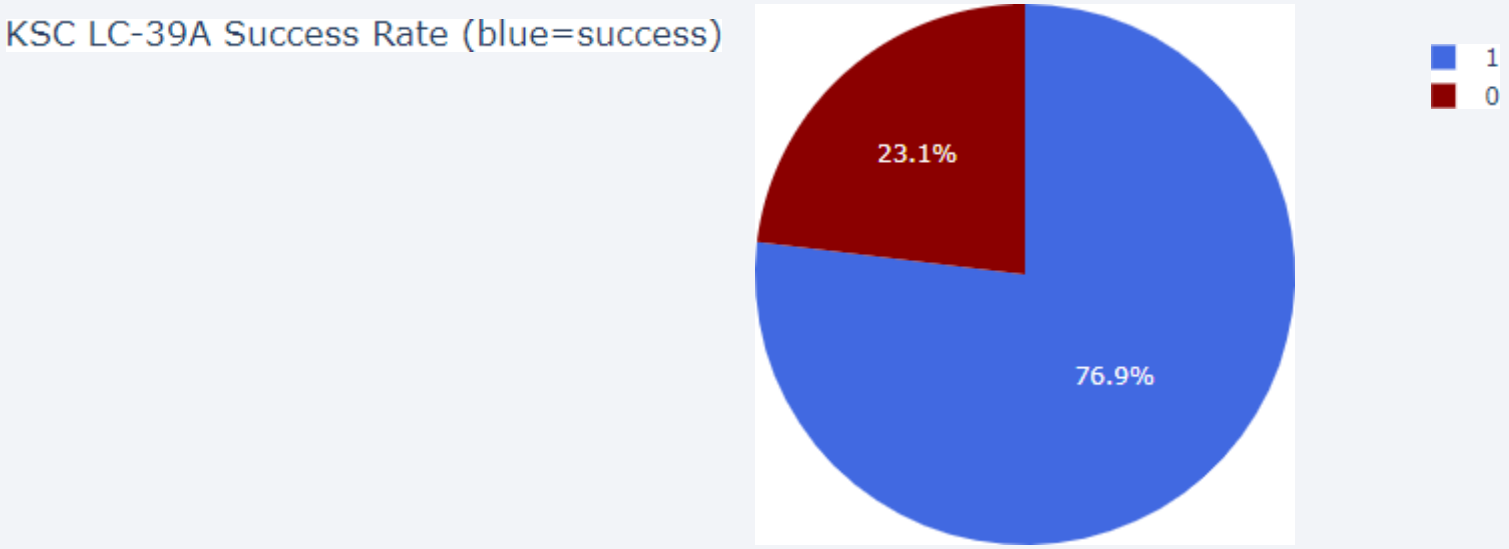

Section 4
Build a Dashboard with Plotly Dash

Successful Launches Across Launch Sites



- The successful landings are distributed across various launch sites, with CCAFS (Cape Canaveral Air Force Station) and KSC (Kennedy Space Center) having an equal number of successful landings. It's noteworthy that a significant portion of these successful landings at CCAFS occurred before the name change from CCAFS LC-40 to CCAFS SLC-40. On the other hand, VAFB (Vandenberg Air Force Base) has the smallest share of successful landings, likely attributed to a smaller sample size and increased launch difficulty associated with launching from the west coast.

Highest Success Rate Launch Site



- KSC LC-39A boasts the highest success rate among the launch sites, achieving 10 successful landings and experiencing 3 failed landings.

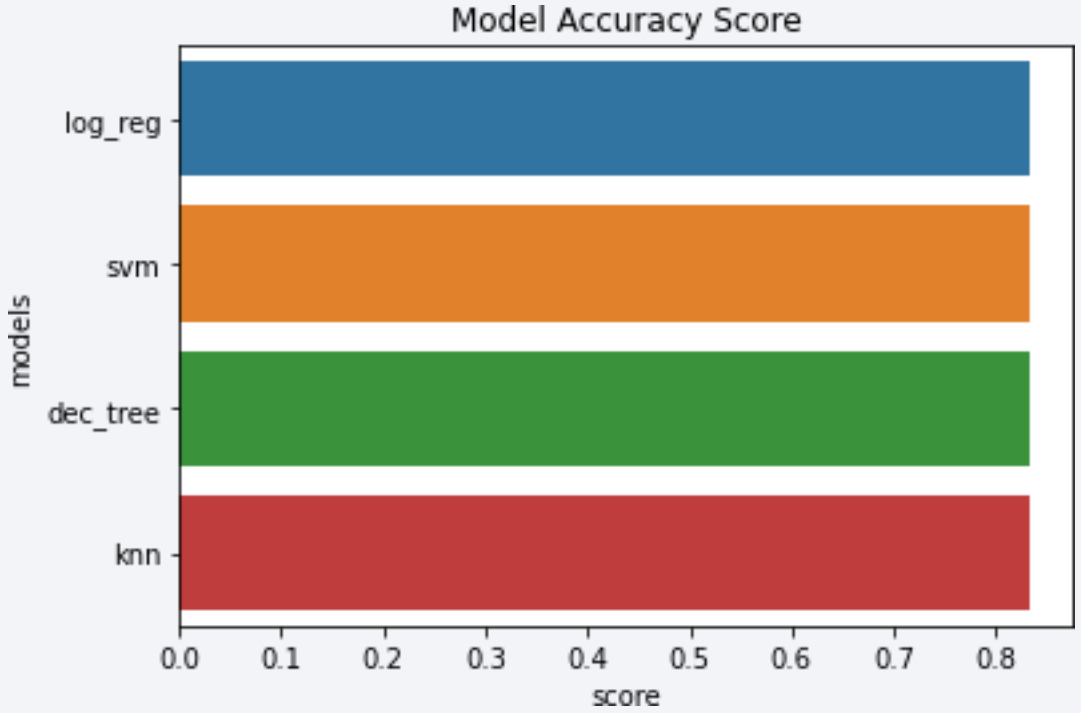
Payload Mass vs Success vs Booster Version Category

- - In the Plotly dashboard, there is a Payload range selector set from 0-10000 instead of the maximum payload of 15600 kg. The "Class" variable is used to indicate 1 for successful landings and 0 for failures. The scatter plot is designed to account for the booster version category using color and represent the number of launches with varying point sizes.
- Interestingly, within the specified payload range of 0-6000 kg, there are two instances of failed landings with payloads recorded as zero kilograms.

Section 5

Predictive Analysis (Classification)

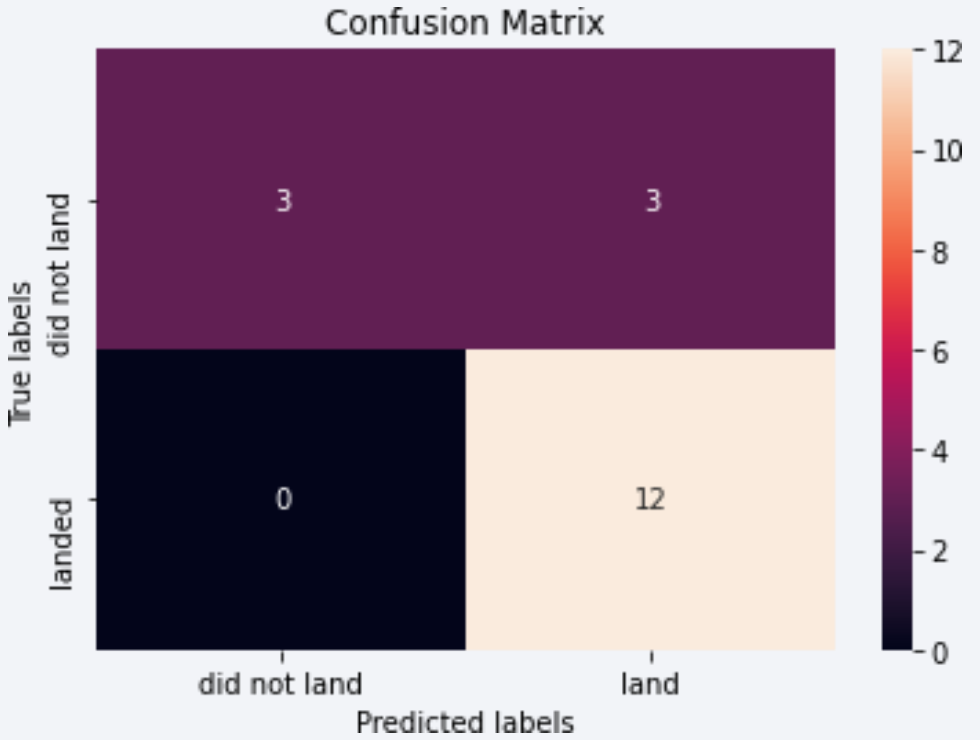
Classification Accuracy



- All models demonstrated nearly identical accuracy on the test set, achieving an accuracy of 83.33%. It's important to highlight that the test size was relatively small, comprising only 18 samples. This limited test size can lead to substantial variance in accuracy results, especially evident in models like the Decision Tree Classifier across repeated runs. To confidently determine the best model, gathering a larger and more diverse dataset is likely necessary.

Confusion Matrix

- Correct predictions align along a diagonal line from the top left to the bottom right in the context of a matrix or plot.
- Given that all models performed similarly on the test set, the confusion matrix remains consistent across all models. Specifically, the models accurately predicted 12 instances of successful landings when the true label was indeed a successful landing. Additionally, they correctly predicted 3 instances of unsuccessful landings when the true label was a failure.
- However, it's important to note that the models made a total of 3 false positive predictions, indicating successful landings when the true label was actually unsuccessful landings. This suggests a tendency of the models to overpredict successful landings.



Conclusions

- Our task was to develop a machine learning model for Space Y, enabling them to bid against SpaceX by predicting successful Stage 1 landings and potentially saving around $100 million USD per launch.
- The data utilized for this model was sourced from a public SpaceX API and web scraping SpaceX's Wikipedia page. We organized the data, created appropriate labels, and stored it in a DB2 SQL database.
- To enhance visualization and understanding, we built a dashboard.
- The resulting machine learning model demonstrated an accuracy of 83%, providing a reliable means for predicting successful Stage 1 landings. This accuracy allows Allon Mask of SpaceY to make informed decisions regarding whether a launch should proceed or not.
- However, to further optimize the model and improve accuracy, we recommend collecting more data. Additional data will facilitate a more comprehensive analysis to determine the best machine learning model for this specific predictive task.

Appendix

- relevant assets like Python code, SQL queries, charts, Notebook outputs, or data sets that may have been created during this project can be found on this github repository : https://github.com/Dahgorago/IBM-Data-Science/tree/main/Data%20Science%20Capstone%20Project

Thank you!