IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Onur İŞCİL
August 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with SQL and Python
  - Visual analysis with falioum
  - Dashboards with Plotly Dash
  - Predictive Analysis by Machine Learning
- Summary of all results
  - Sufficient information was gathered using the SpaceX API and web scraping techniques.
  - The relationship between successful and unsuccessful launches was revealed through Exploratory Data Analysis (EDA).
  - With the application of machine learning techniques, the probability of a launch's success or failure, and consequently the estimated cost of the launch, was predicted.

Introduction

- Project background and context
  - Space transportation is an extremely expensive and challenging endeavor. Rockets, in particular, account for a significant portion of the cost. While many companies estimate the cost of space transportation to be around 165 million, SpaceX offers this service for 62 million dollars. The reason for this difference is the reusability of their rockets. Therefore, whether the rockets successfully return to Earth or not is the main factor determining the cost
- Problems you want to find answers
  - The main factor in calculating the cost of a launch is predicting the success rate based on the launch site, payload amount, and the target orbit. If we develop an algorithm that calculates this, we can estimate the cost and identify the most successful launch conditions

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data collection method consists of two parts:
    - Data was obtained using the SpaceX API
    - Data was also gathered through web scraping from Wikipedia
- Perform data wrangling
  - Unwanted parts of the data were filtered out, and missing data was completed.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The data, after being split into training and test sets and standard scaled, was used to train various classification algorithms, and the best algorithm was selected.

Data Collection

- During the data collection phase, two main sources and various techniques were primarily focused on.
- These are the SpaceX API and Web Scraping, respectively.
- The data obtained from both approaches were subsequently transferred to dataframes and finally to CSV files.
- The details of the method can be seen in the following two slides.

Data Collection – SpaceX API

- The data retrieval process using the Space X-API, as shown in the flowchart above, involves obtaining the JSON data, filtering it, and then saving it

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longit
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577
5	8	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577
6	10	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577
7	11	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610
8	12	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577

GitHub link to the code

Data Collection - Scraping

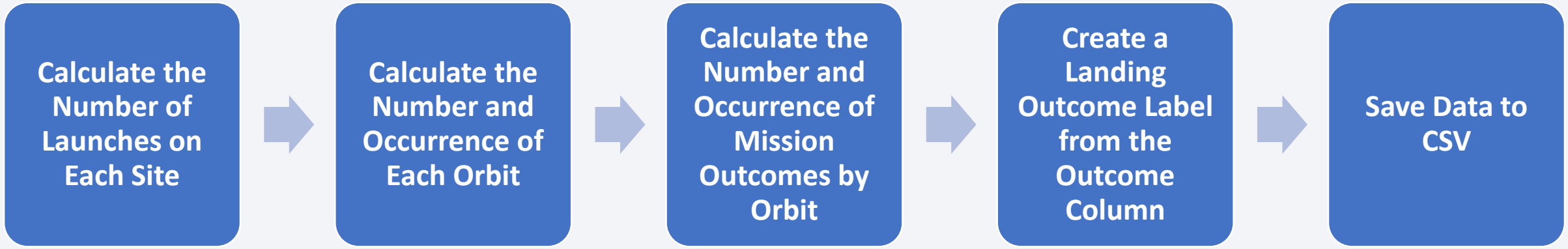
Connect to Website and Fetch Data: → Parse HTML Content: → Identify and Extract Table → Extract Data → Convert to DataFrame and Save to CSV

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.07B0003.18	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.07B0004.18	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.07B0005.18	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.07B0006.18	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.07B0007.18	No attempt\n	1 March 2013	15:10
...	...	...	...	...	...	...	...	...	...	...	...
222	117	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1051.10657	Success	9 May 2021	06:42
223	118	KSC	Starlink	~14,000 kg	LEO	SpaceX	Success\n	F9 B5B1058.8660	Success	15 May 2021	22:56
224	119	CCSFS	Starlink	15,600 kg	LEO	NASA	Success\n	F9 B5B1063.2665	Success	26 May 2021	18:59
225	120	KSC	SpaceX CRS-22	3,328 kg	LEO	Sirius XM	Success\n	F9 B5B1067.1668	Success	3 June 2021	17:29
226	121	CCSFS	SXM-8	7,000 kg	GTO	NaN	NaN	F9 B5	NaN	6 June 2021	04:26

The process of connecting to the web address, locating the data within the address, and subsequently retrieving and saving the data is shown step by step in the above flowchart.

[GitHub link to the code]

Data Wrangling



- The dataset contains various types of information. For example, values of 1 or 0 represent successful or unsuccessful landing attempts, with 'True Ocean' indicating a successful landing in the ocean and 'False Ocean' indicating an unsuccessful landing in the ocean. Additionally, it includes extensive information about launch pads and details about which orbits the launches occurred. At this stage, the data has been processed to handle this information.

GitHub link to the code

EDA with Data Visualization

- Various graphical techniques were used to explore the relationships between variables. For example, scatter plots were used to observe the correlation and relationship between two variables, bar charts to examine the relationship between categorical and numerical data, and line graphs to identify trends

- Scatter Plot Graphs
  - Flight Number X Payload Mass
  - Flight Number X Launch Site
  - Payload vs. Launch Site
  - Orbit X Flight Number
  - Payload vs. Orbit Type
  - Orbit vs. Payload Mass
- Bar Graph
  - Success rate X Orbit
- Line Graph
  - Success rate vs. Year

GitHub link to the code

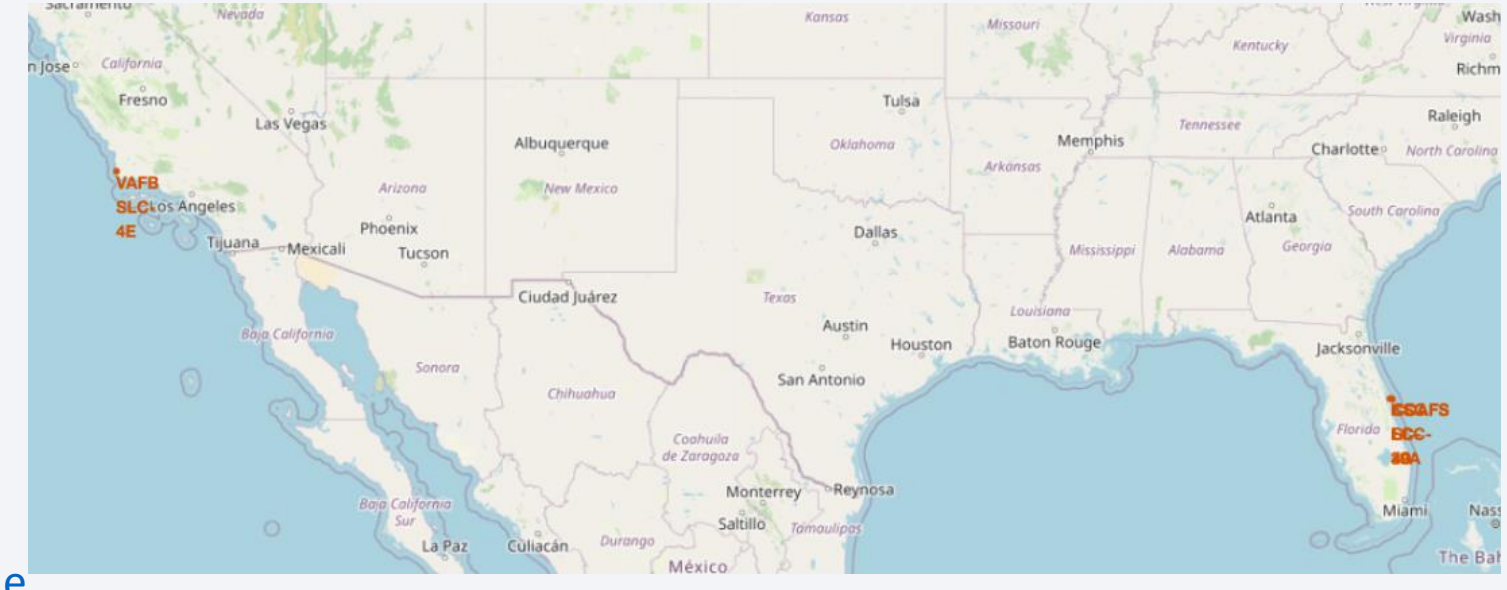
EDA with SQL

- During this stage, Exploratory Data Analysis (EDA) was performed using SQLAlchemy to execute SQL commands within Python. The following questions were addressed:
  - Unwanted parts of the data were filtered out, and missing data was completed.
  - Display the names of the unique launch sites in the space missions.
  - Display the total payload mass carried by boosters launched by NASA (CRS).
  - Display the average payload mass carried by the booster version F9 v1.1.
  - List the date when the first successful landing outcome on a ground pad was achieved.List the total number of successful and failed mission outcomes.
  - List the names of the booster versions that have carried the maximum payload mass using a subquery.
  - Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20, in descending order.

GitHub link to the code

Build an Interactive Map with Folium

- At this stage, a map was created using Folium to visualize the launch sites.
- The launch sites were then marked on the map.
- Information about successful and failed launches was added to these marked sites, and the markers were color-coded based on the number of launches and the success rate



[GitHub link to the code]

Build a Dashboard with Plotly Dash

- At this stage, a dashboard was created using the data we obtained.
- DASH and Plotly libraries were used for this purpose.
- In the dashboard, pie charts and scatter plots were created based on launch sites, payload range, and success rate. The dashboard allows for the interactive selection and filtering of launch sites and payload range information

GitHub link to the code

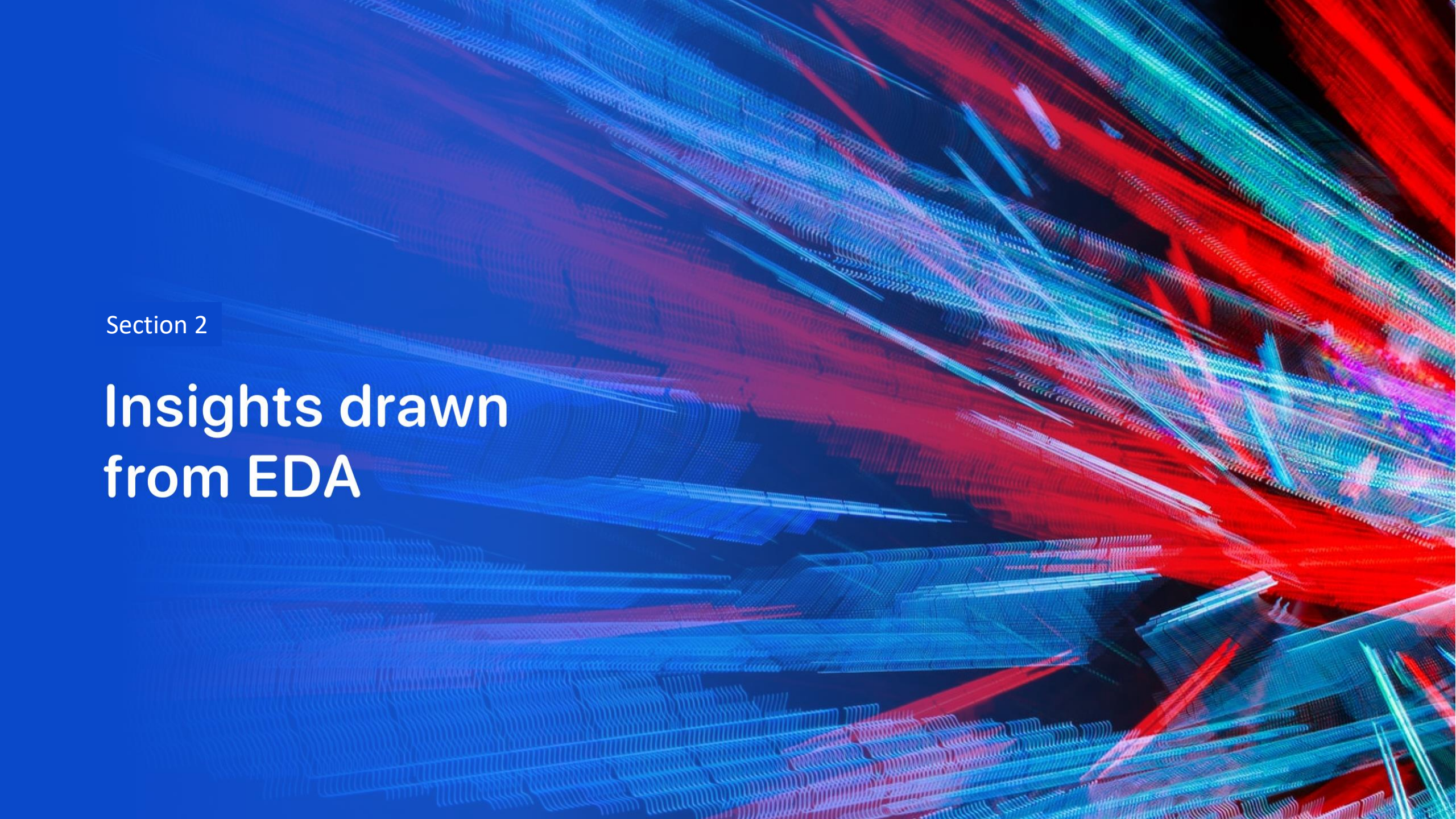
Predictive Analysis (Classification)

- At this stage, the data was subjected to the StandardScaler function, and after splitting it into test and training sets (with test_size=0.2), a technique was developed to predict the system's performance using;
  - logistic regression,
  - support vector machine,
  - decision tree,
  - and k-nearest neighbors methods.
- GridSearchCV was used to optimize different hyperparameters

GitHub link to the code

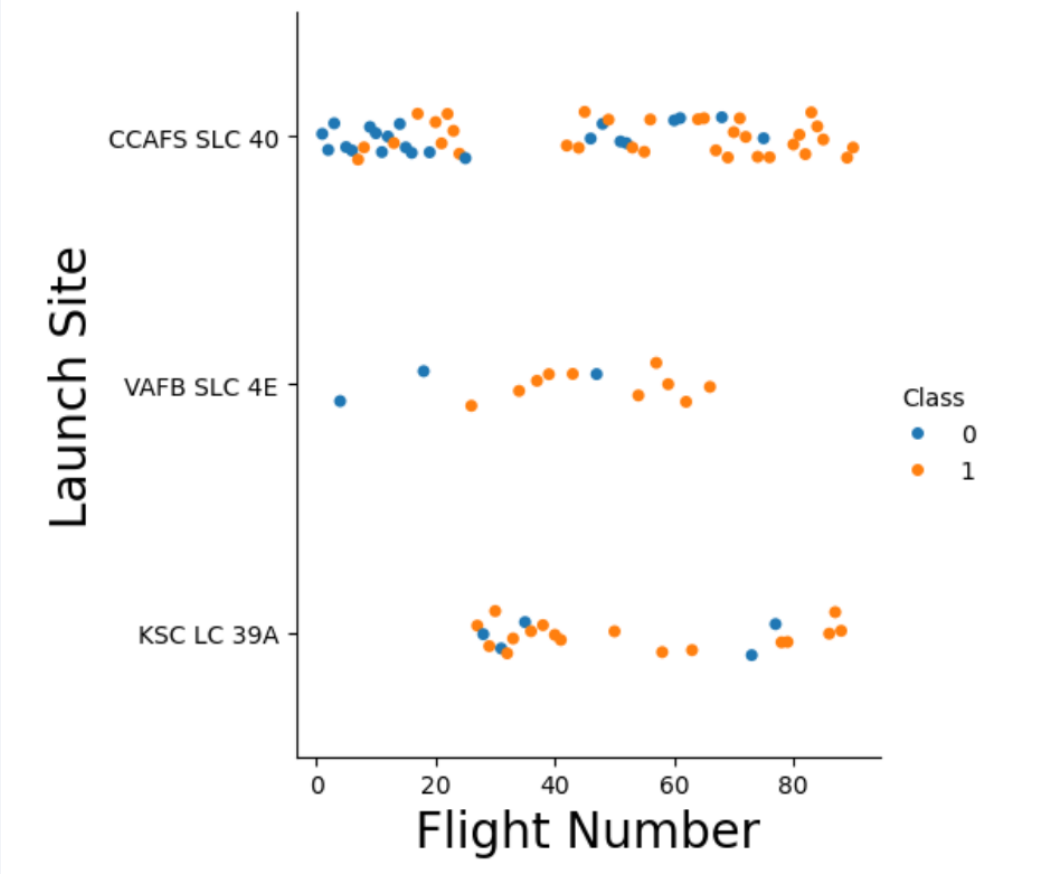
Results

- At the conclusion of all analyses, we encounter three different categories of results:
  - Exploratory data analysis results
  - Interactive analytic results obtained through the dashboard
  - Predictive analysis results

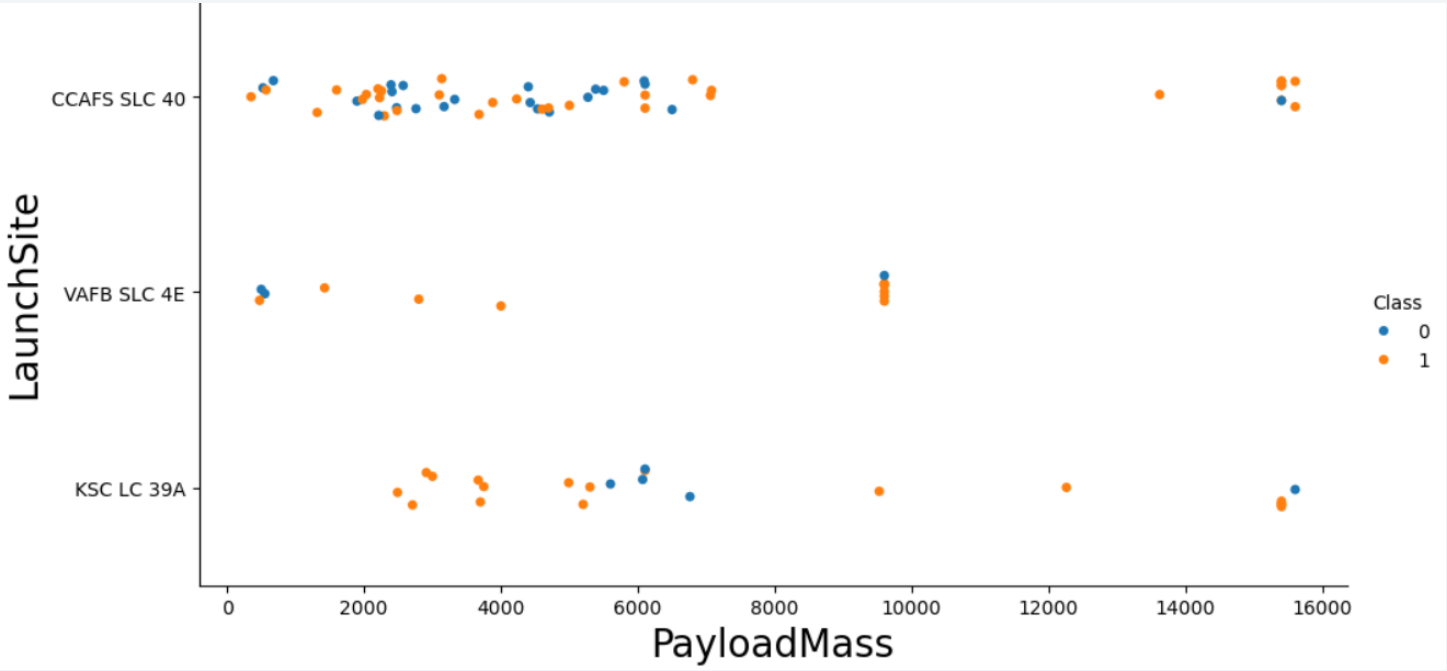

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- Specifically, at the launch sites 'KSC LC 39A' and 'VAFB SLC 4E,' it is observed that rockets with more than 20 flights are typically launched, and the overall success rate is higher compared to another launch site, 'CCAFS SLC 40



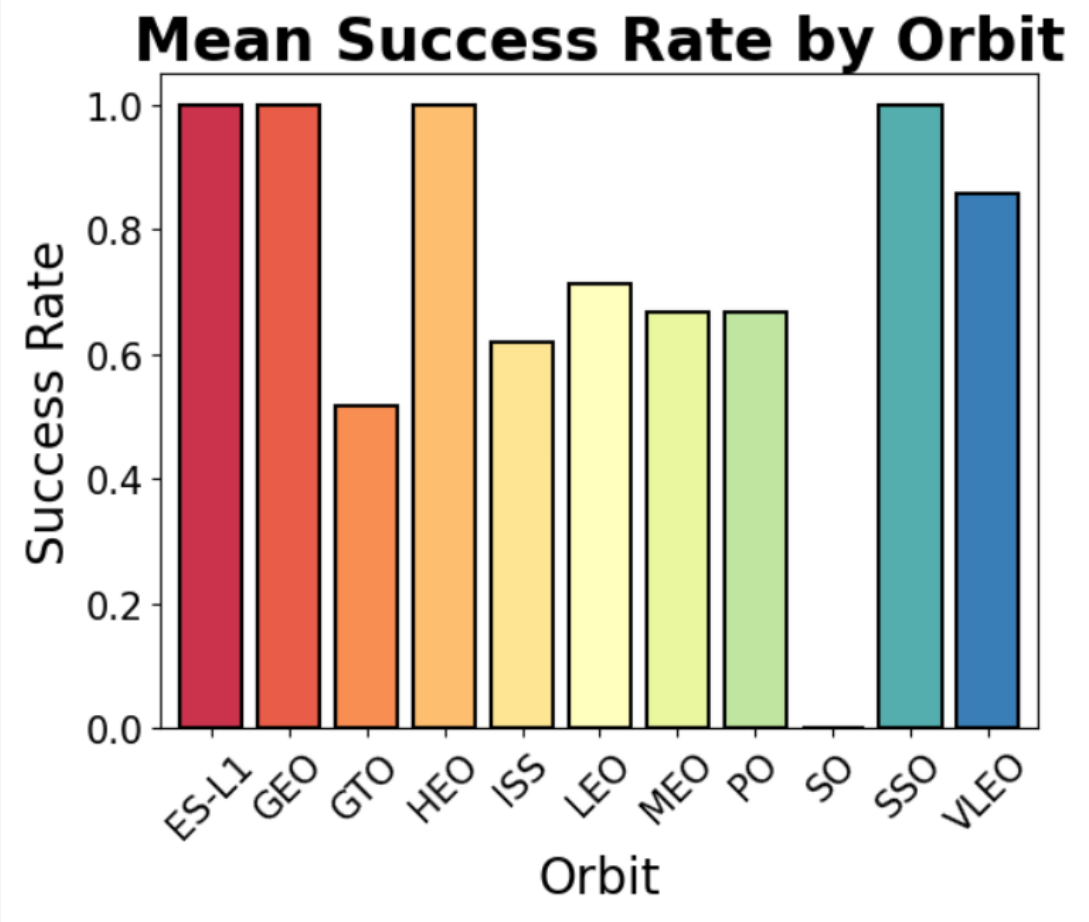
Payload vs. Launch Site



- Another observation is that the VAFB-SLC launch site is generally not used for heavy payloads (greater than 10,000 kg)

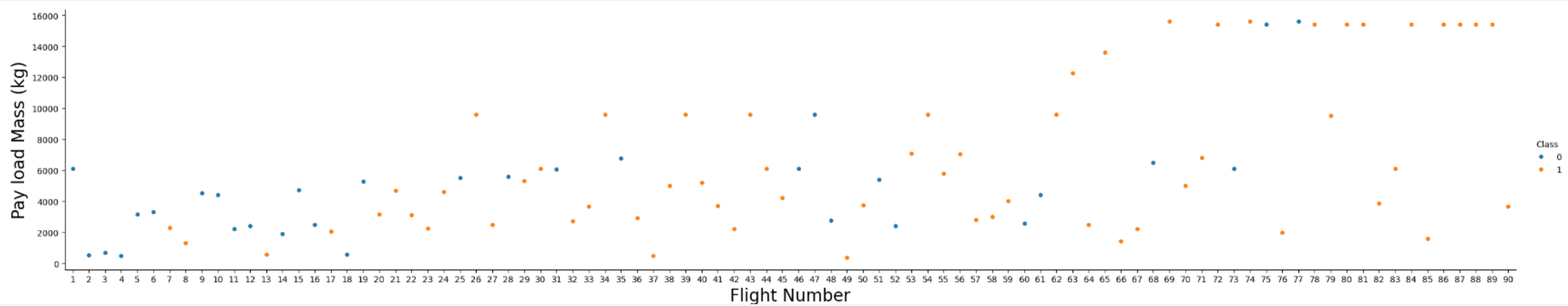
Success Rate vs. Orbit Type

- Various success rates have been observed across different orbits. Specifically, while the success rate in the SO orbit is 0, the success rates in the ES-L1, GEO, HEO, and SSO orbits are at least 1. Additionally, the VLEO orbit has shown a success rate exceeding 0.8

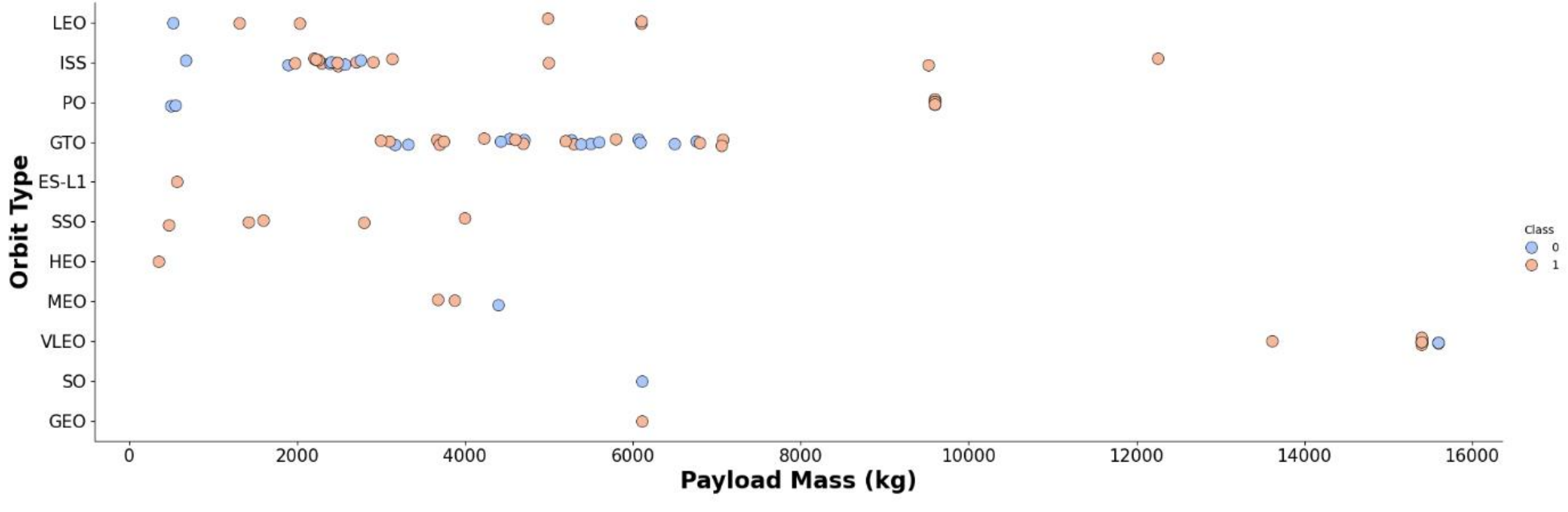


Flight Number vs. Orbit Type

- When examining the Flight Count vs. Payload Mass graph, it is observed that even in launches with higher masses, rockets with a higher flight count have a higher rate of successful landings compared to rockets with a lower flight count

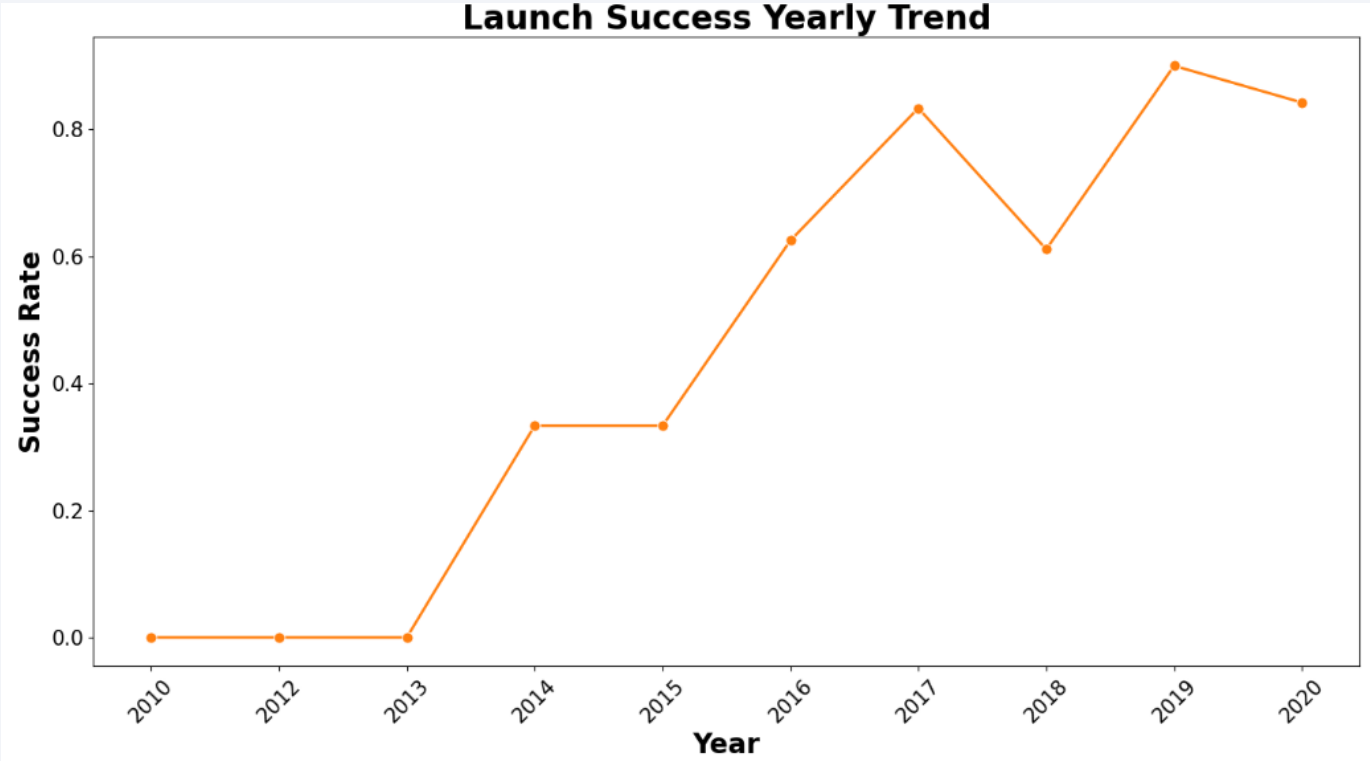


Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



- We can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- In this output, we retrieve information about the launch sites using SQL with the DISTINCT keyword

Task 1

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

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

```
 * 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'

- With this query, we retrieved the first 5 launch pads whose names start with 'CCA

Task 2

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

```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

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

- We calculated the total payload carried by boosters from NASA, and as a result, we obtained a value of 45,596 from this query

Task 3

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

[16]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE customer = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

[16]:

TOTAL_PAYLOAD
45596

Average Payload Mass by F9 v1.1

- In the analysis of the 'booster version F9 v1.1' that we will use in the machine learning section, we found the average payload mass to be 2928.4

Task 4

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

```
[17]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * sqlite:///my_data1.db
Done.
```

[17]:

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

- As the output of the SQL command we used to discover the date of the first successful landing attempt, we obtained the date 2015-12-22

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

FIRST_SUCCESS_GP
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Task 6

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

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND Landing_Outcome = 'Success (drone ship)';
```

```
 * sqlite:///my_data1.db
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Calculation of the Total Number of Successful and Failed Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
%sql SELECT Mission_Outcome, COUNT(*) as Total FROM SPACEXTBL GROUP BY mission_outcome;
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	Total
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- List of the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE, Date FROM SPACEXTBL WHERE DATE LIKE '2015-%' AND Landing_Outcome = 'Failure (drone ship)';
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version	Launch_Site	Date
F9 v1.1 B1012	CCAFS LC-40	2015-01-10
F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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

- 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

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.

```
%sql SELECT Landing_Outcome, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY QTY DESC;
```

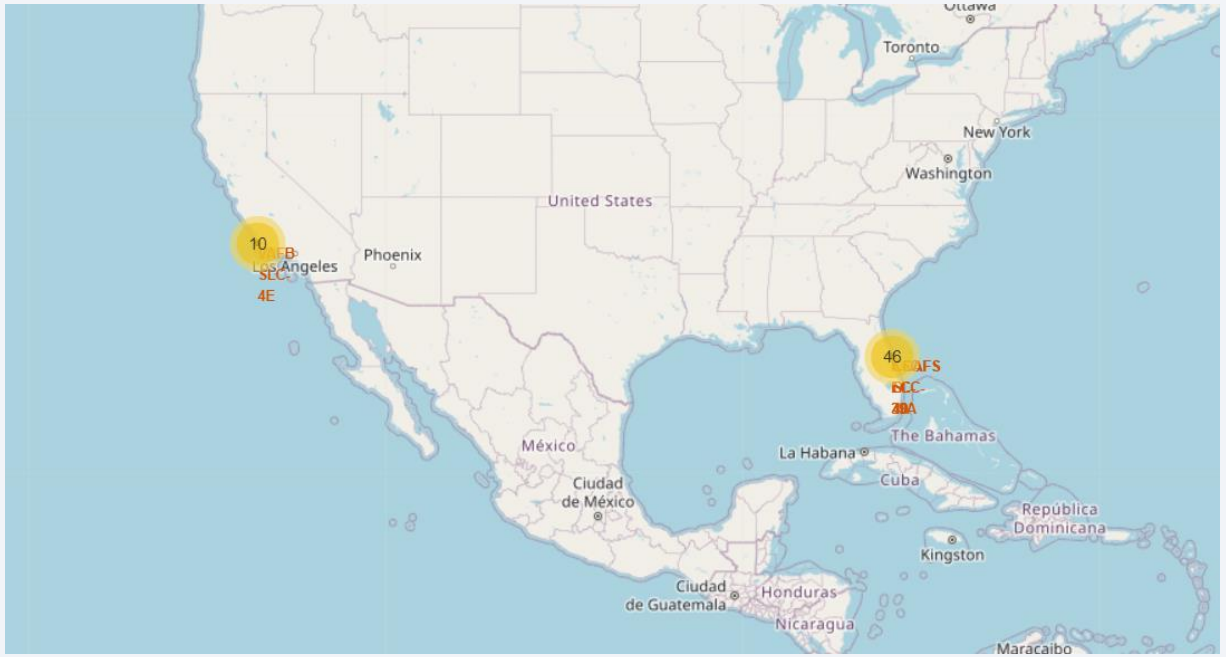
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	QTY
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

Section 3

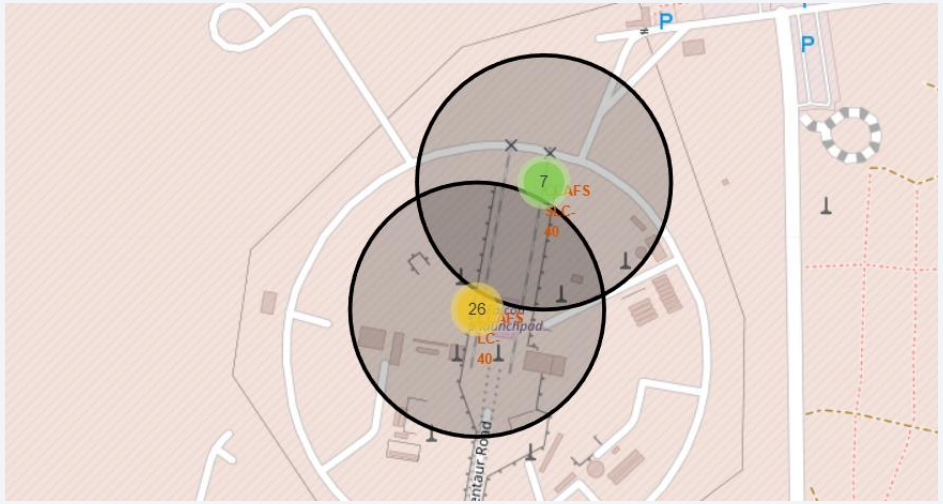
Launch Sites Proximities Analysis

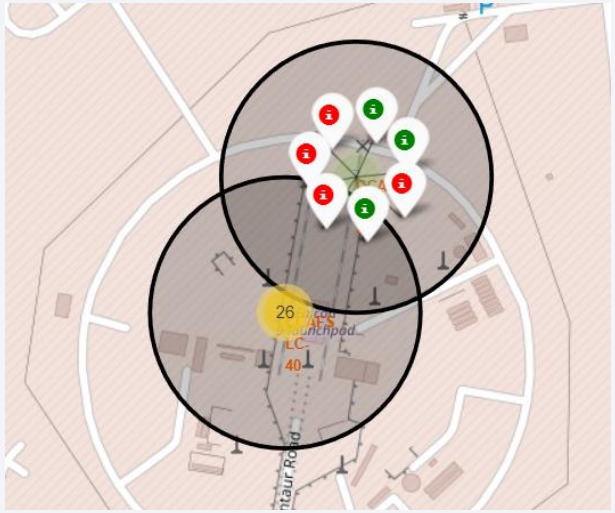
Folium Map

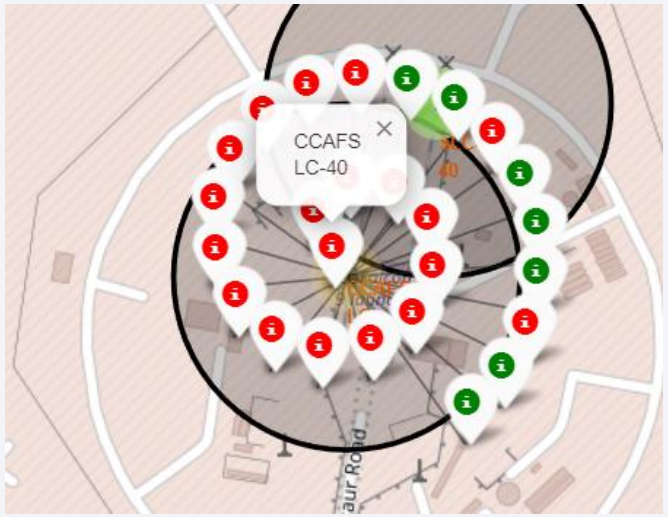


- In the generated Folium map, users can interactively navigate, and the launch sites along with the number of launches at each site can be viewed

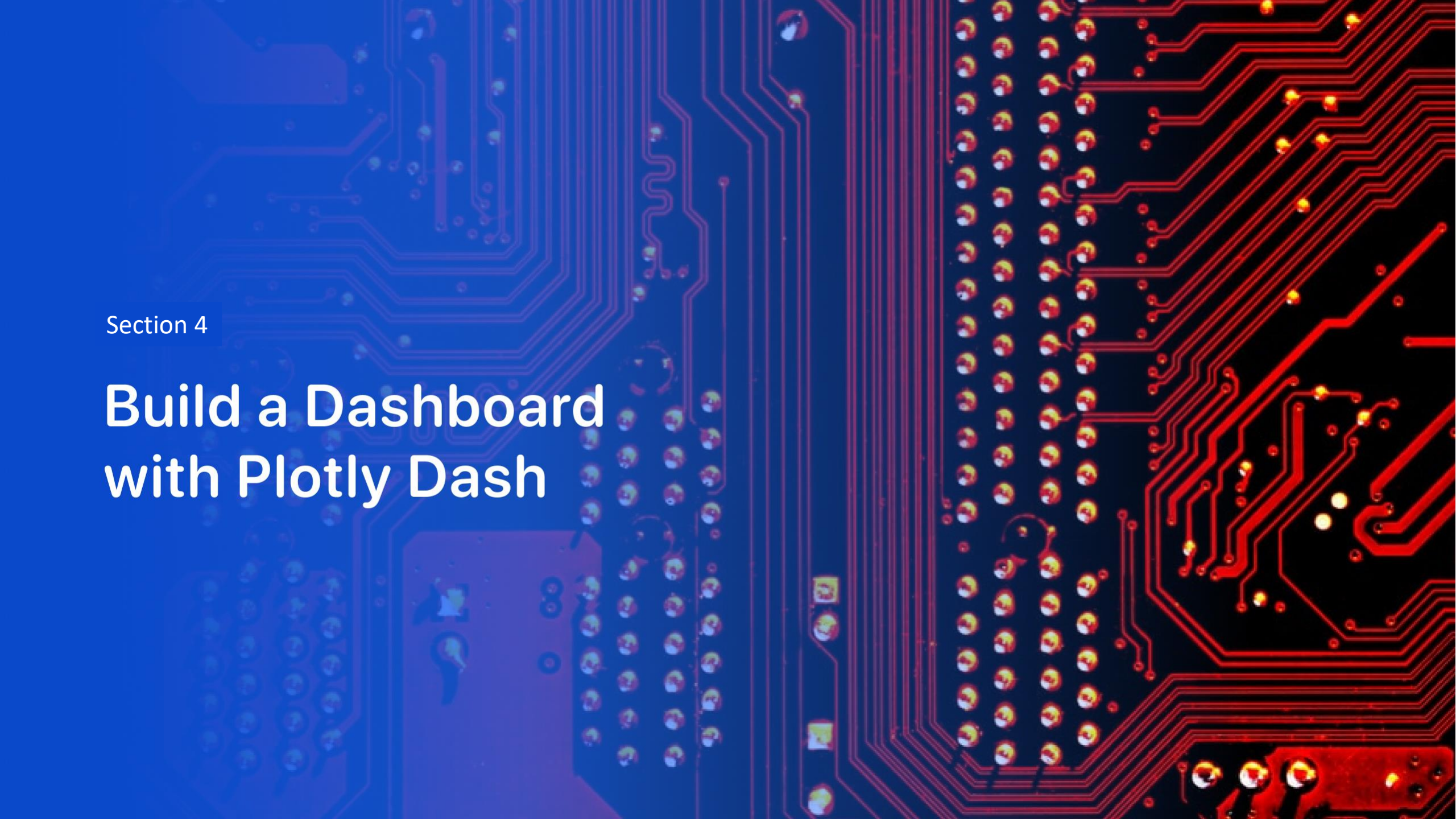
Folium Launch Sites Map



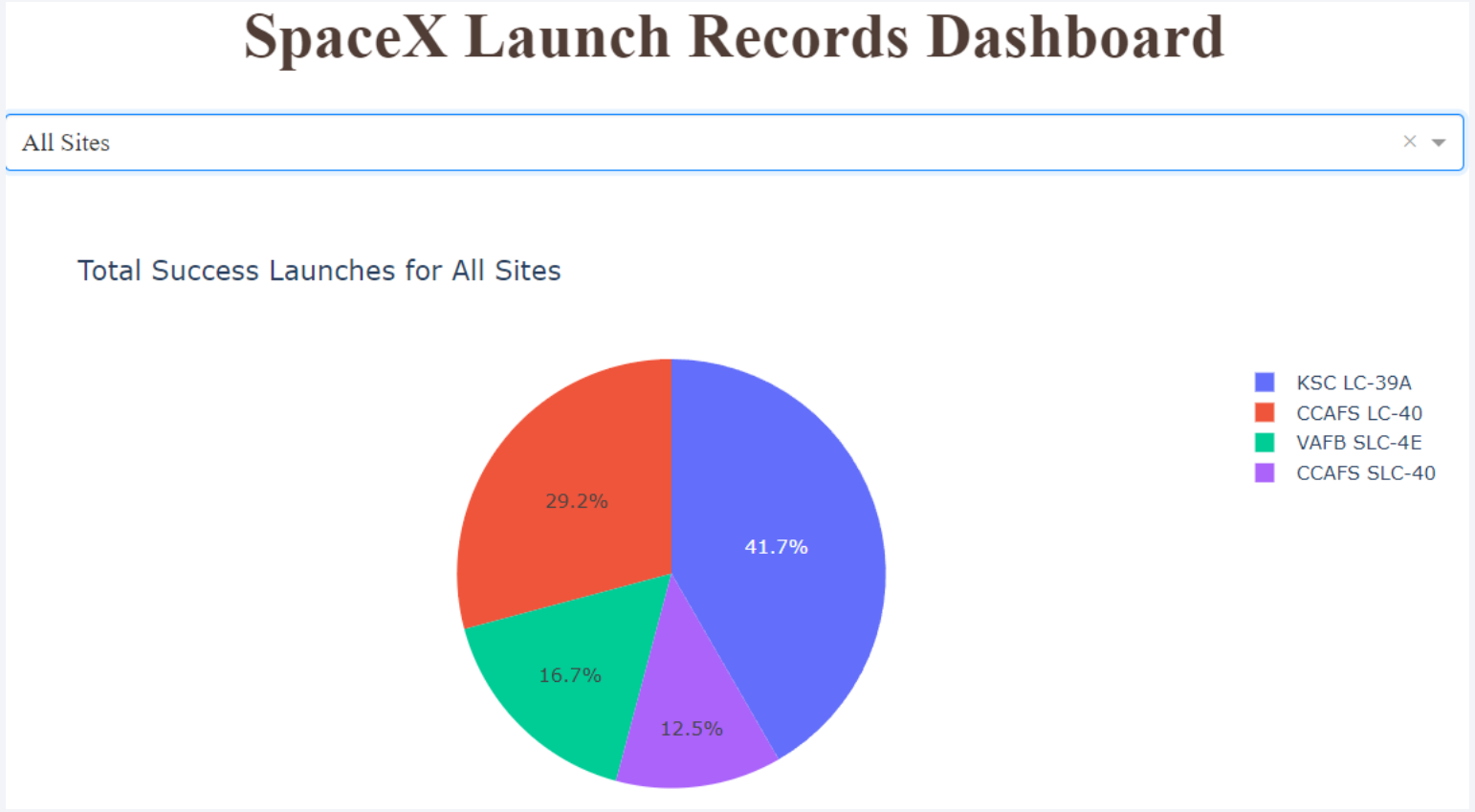




- The colors of the launch sites vary based on the success rate, and when clicked, you can see which launches were successful and which were unsuccessful

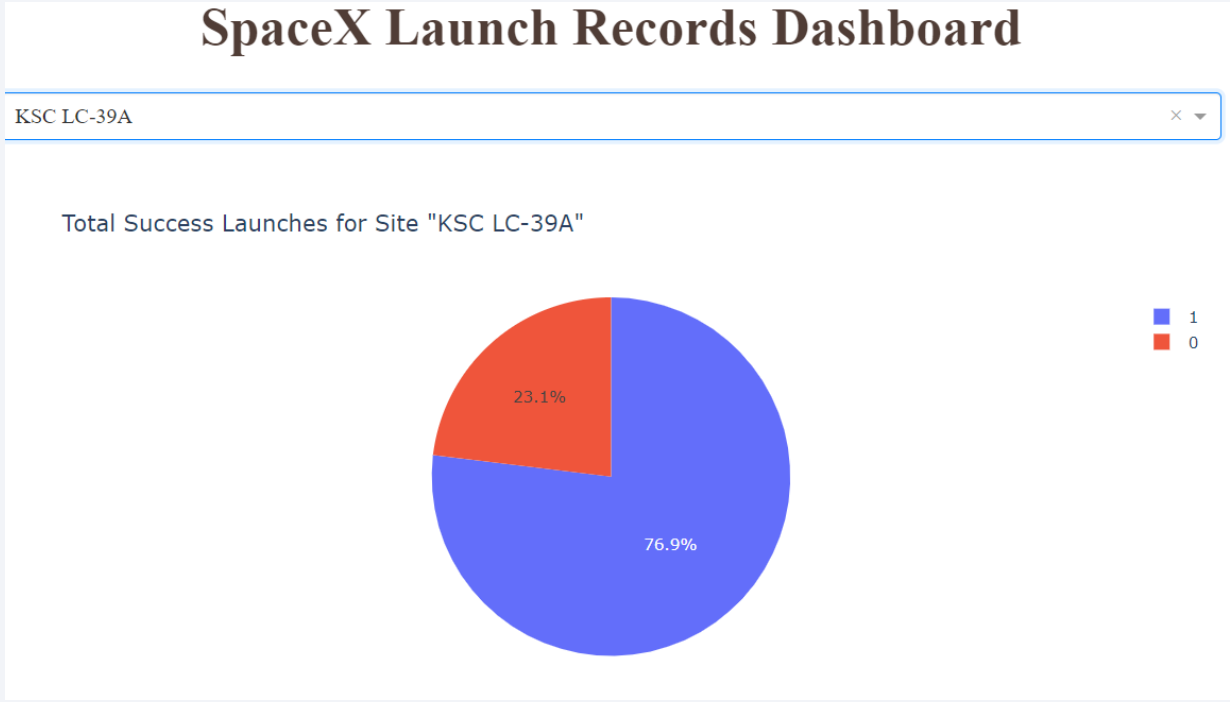

Section 4
Build a Dashboard with Plotly Dash

Dashboard – Pie Chart of All launch Sites



- Through the dashboard, the overall success data by launch site revealed that 41.7% of successful launches occurred at the KSC LC launch pad

Dashboard - Pie Chart of KSC LC launch Site



- For a more detailed analysis, filtering based on the most successful launch pad shown in the previous slide reveals that the 'KSC LC' launch pad has a success rate of 76.9%.

Dashboard – Heavy weighted Launches

- It has been observed that the success rate is low in heavy-weighted launches

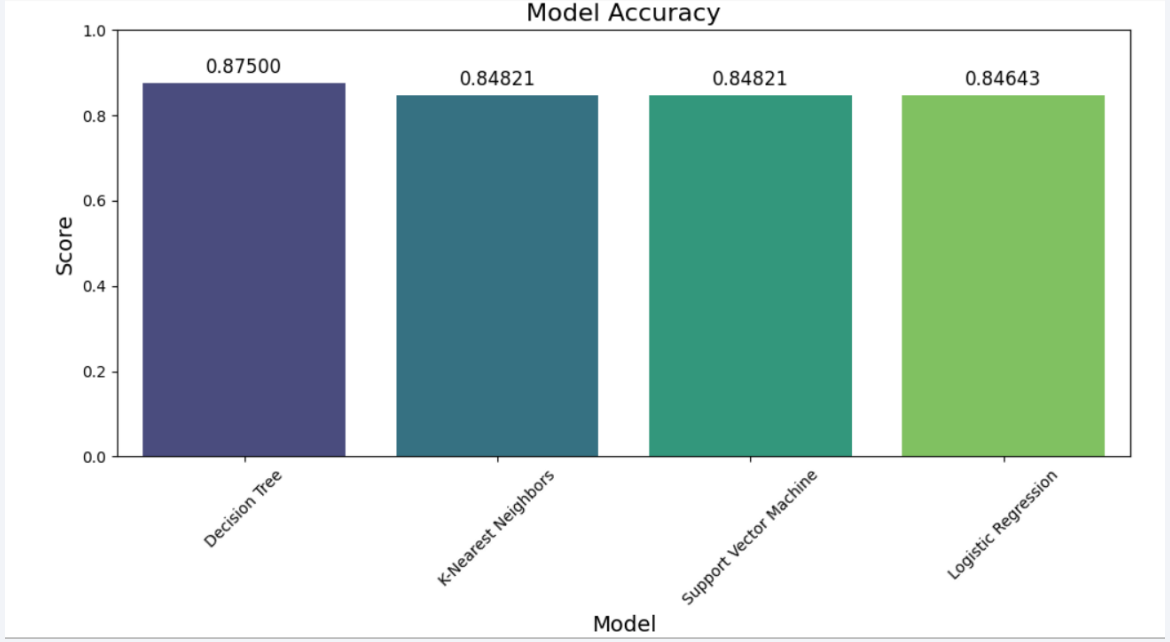
Dashboard - Light weighted Launches

- It has been observed that the success rate is high in low-weighted launches

Section 5

Predictive Analysis (Classification)

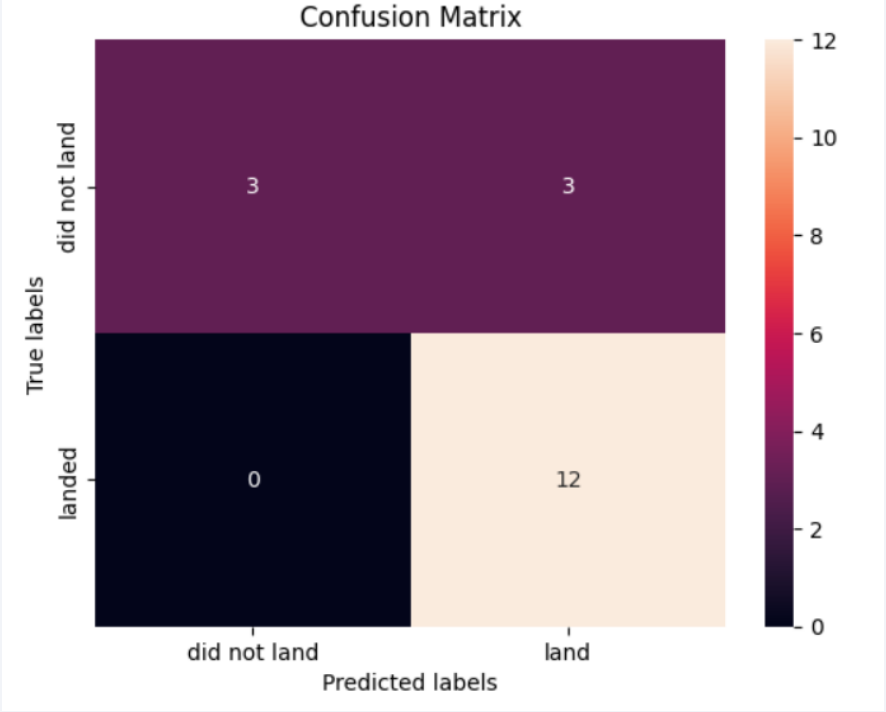
Classification Accuracy



- When evaluating model accuracy, it is observed that the most successful model in predicting launch success is the one created using the Decision Tree approach, while the least successful model is the one created using Logistic Regression. However, even the least successful model has an accuracy score of 0.84643, indicating a good performance

Confusion Matrix

- Examining the performance of the model trained with the Decision Tree technique using the confusion matrix, the visual on the right appears. In the model, only 3 predictions show Type 1 errors, while the remaining 15 predictions are all correct



Conclusions

- Launches with low payloads (5000 kg and below) are more successful
- The KSC LC-39A launch site shows better performance compared to other launch sites.
- Launches to ES-L1, GEO, HEO, SSO, and VLEO orbits are nearly 100% successful.
- Rockets with higher numbers of launches have a higher success rate compared to other rockets.
- The Decision Tree algorithm is the most successful approach in predicting launch success.
- **A prediction made using the Decision Tree method can provide necessary information on the cost of a launch with an accuracy of 0.87500.**

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!