



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Quirino Ygot
30th August 2023



Outline

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

Executive Summary

- Summary of methodologies
 - Data Collection through API
 - Data Collection with Web Scraping
 - Data Wrangling
 - Exploratory Data Analysis with SQL
 - Exploratory Data Analysis with Data Visualization
 - Interactive Visual Analytics with Folium
 - Machine Learning Prediction
- Summary of all results
 - Exploratory Data Analysis result
 - Interactive analytics in screenshots
 - Predictive Analytics result

Introduction

- Project background and context

Space X 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 Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- Problems you want to find answers

- What factors determine if the rocket will land successfully?
- The interaction amongst various features that determine the success rate of a successful landing.
- What operating conditions needs to be in place to ensure a successful landing program?

Section 1

Methodology

Methodology

Executive Summary

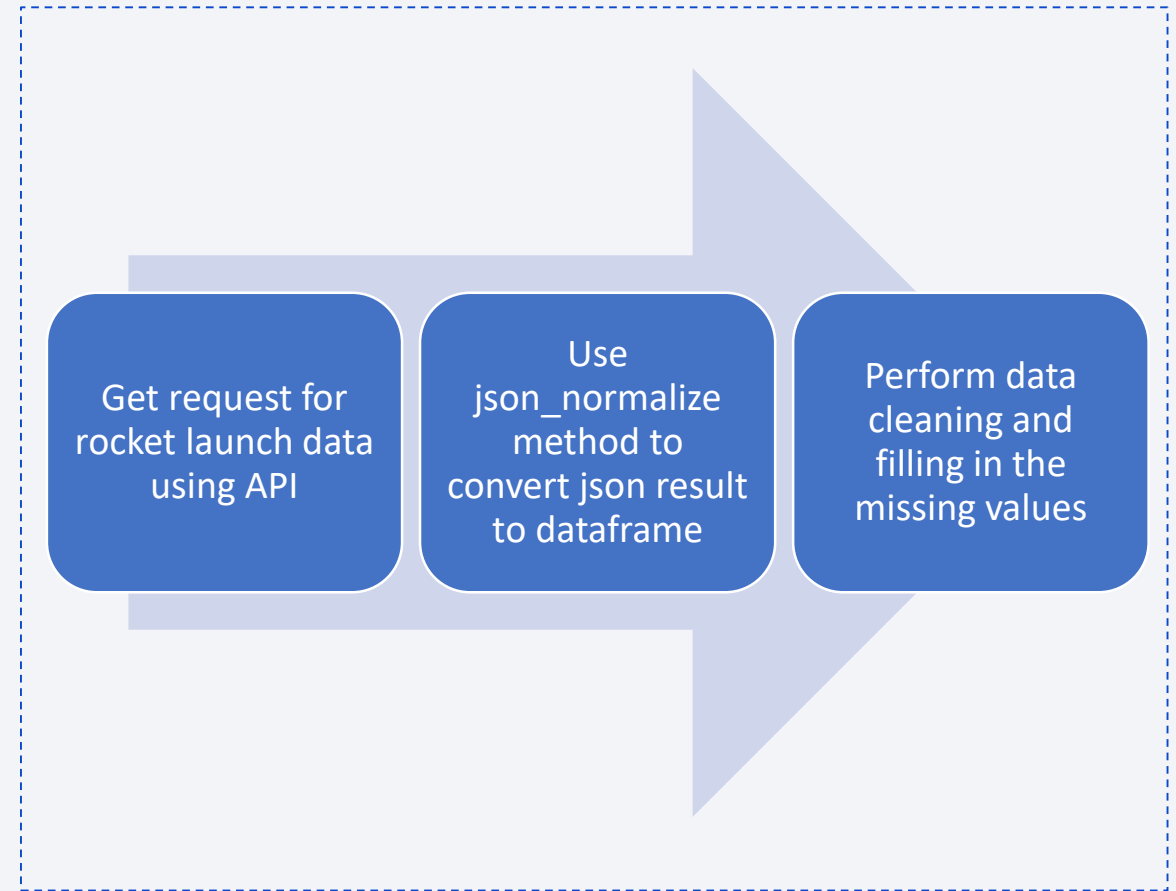
- Data collection methodology:
 - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
 - One-hot encoding was applied to categorical features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
 - How to build, tune, evaluate classification models

Data Collection

- Describe how data sets were collected.
 - Data collection was done using get request to the SpaceX API.
 - Next, I decoded the response content as a Json using `.json()` function call and turn it into a pandas dataframe using `.json_normalize()`.
 - I then cleaned the data, checked for missing values and fill in missing values where necessary.
 - In addition, I performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup.
 - The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for future analysis.

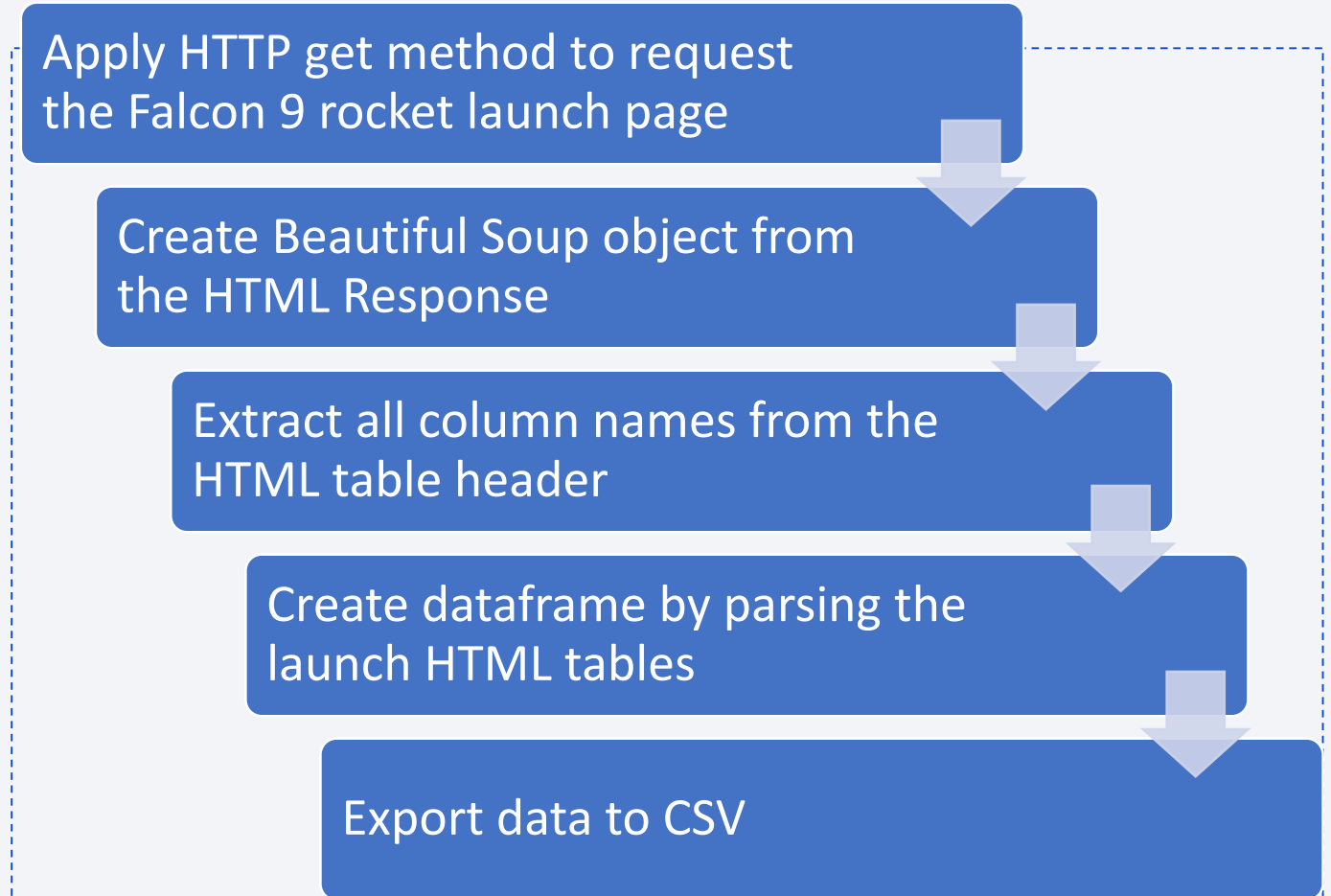
Data Collection - SpaceX API

- I used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.
- Here is the link to the notebook: [IBM-Data-Science-Capstone/Data Collection API.ipynb](https://github.com/kerygot16/IBM-Data-Science-Capstone/blob/main/Collection%20API.ipynb) at main · kerygot16/IBM-Data-Science-Capstone (github.com)



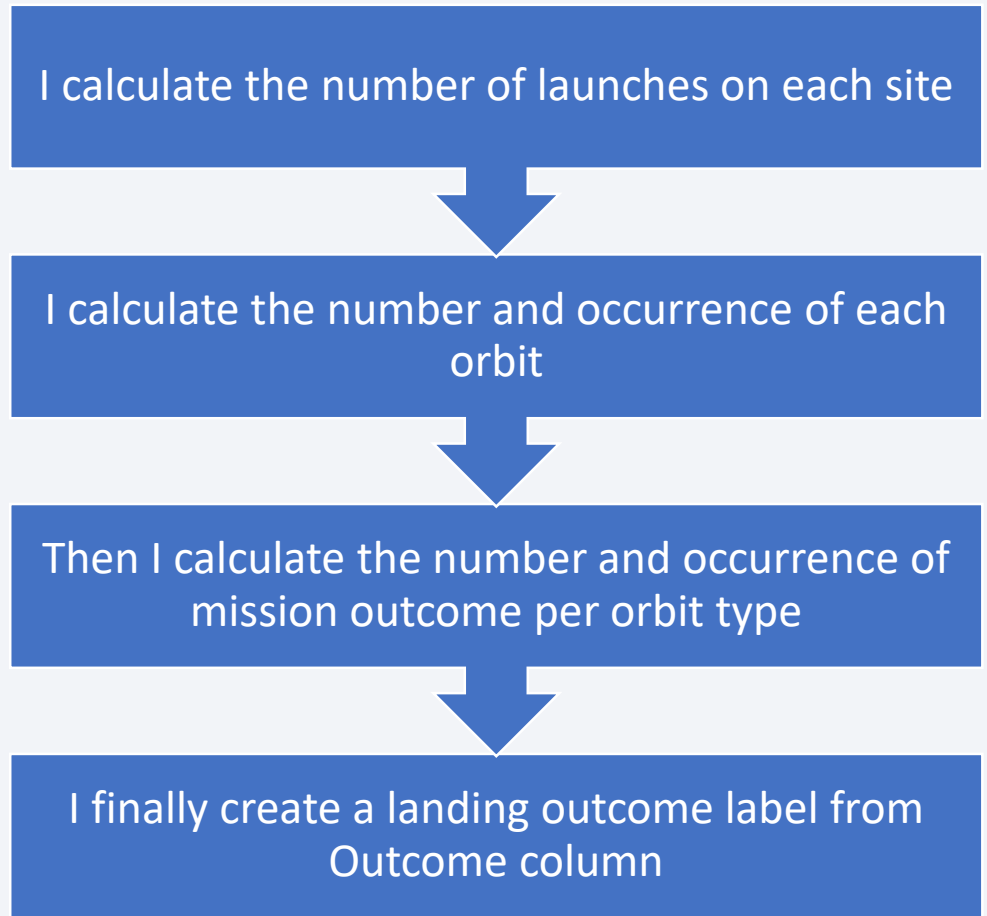
Data Collection - Scraping

- I applied web scraping to webscrap Falcon 9 launch records with BeautifulSoup
- I parsed the table and converted it into a pandas dataframe
- Here is the link to the notebook: [IBM-Data-Science-Capstone/Data Collection With Webscraping.ipynb](https://github.com/kerygot16/IBM-Data-Science-Capstone/blob/main/Data%20Collection%20With%20Webscraping.ipynb) at main · kerygot16/IBM-Data-Science-Capstone (github.com)

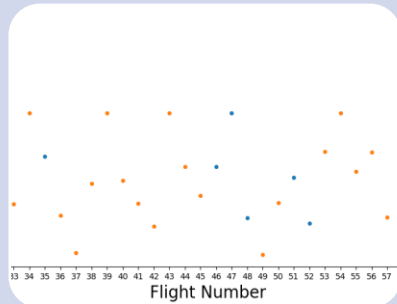


Data Wrangling

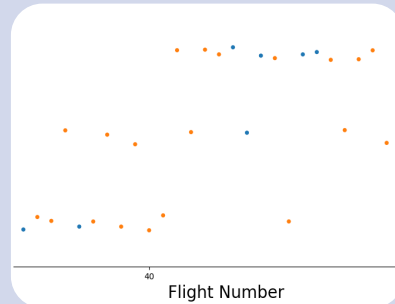
- I performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- I convert the outcomes into Training labels with 1 if the booster successfully landed and 0 if it was unsuccessful.
- Here is the link of the notebook: [IBM-Data-Science-Capstone/Data Wrangling.ipynb](https://github.com/kerygot16/IBM-Data-Science-Capstone/blob/main/Data%20Wrangling.ipynb) at main · kerygot16/IBM-Data-Science-Capstone (github.com)



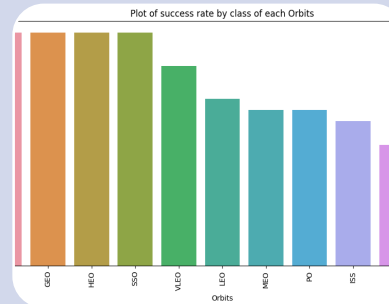
EDA with Data Visualization



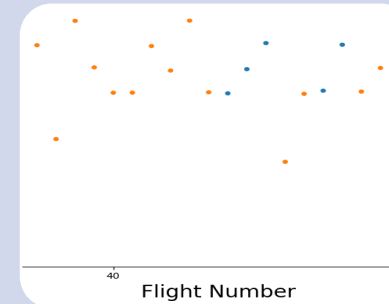
Using scatter plot, I plot out the Flight Number and Payload Mass Variables to see how would they affect the launch outcome



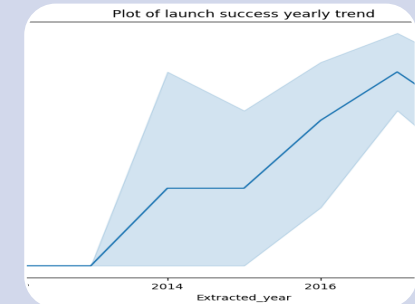
I then used scatter plot to visualize the relationship between Flight Number and Launch Site as well as the relationship between Flight Number and Launch Site.



To visualize the relationship between success rate of each orbit type, I used the bar chart



I also used the scatter point chart to visualize the relationship between Flight Number and Orbit Type and Payload with respect to the Orbit Type.



I then visualize the launch success yearly trend using the line chart.

EDA with SQL

- I loaded the SpaceX dataset into a PostgreSQL database without leaving the jupyter notebook.
- I then applied EDA with SQL to get any insights from the data. I wrote queries to find out relevant information such as:
 - The names of unique launch sites in the space mission.
 - The total payload mass carried by booster version F9 v1.1
 - The total number of successful and failure mission outcomes.
 - The failed landing outcomes in drone ship, their booster version and launch site names.
- Here is the link to the notebook on EDA with SQL: [IBM-Data-Science-Capstone/EDA with SQL.ipynb at main · kerygot16/IBM-Data-Science-Capstone \(github.com\)](https://github.com/kerygot16/IBM-Data-Science-Capstone/blob/main/EDA%20with%20SQL.ipynb)

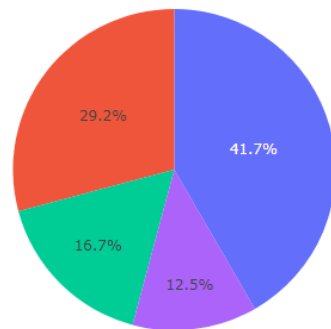
Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- I marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.
- I then assigned the feature launch outcomes (failure or success) to class 0 and 1, such that 0 for failure and 1 for success.
- Using the color-labeled marker clusters, I identified which launch sites have relatively high success rate.
- I calculated the distance between a launch site to its proximities and answered some questions such as:
 - Are launch sites near railways, highways and coastlines?
 - Do launch sites keep certain distance away from cities?
- Please refer to the link for the complete notebook: [IBM-Data-Science-Capstone/Interactive Visual Analytics with Folium.ipynb at main · kerygot16/IBM-Data-Science-Capstone \(github.com\)](https://github.com/kerygot16/IBM-Data-Science-Capstone/blob/main/IBM-Data-Science-Capstone/Interactive%20Visual%20Analytics%20with%20Folium.ipynb)

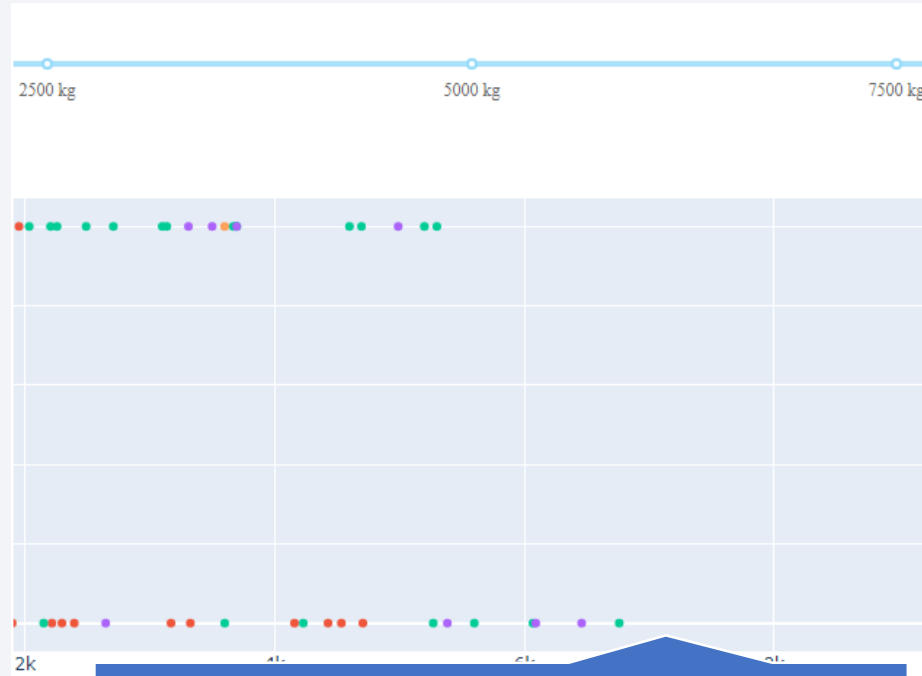
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard

ites



I plotted pie chart in the dashboard showing the total launches by a certain site.




I used scatter plot to visually observe how payload may be correlated with mission outcomes for selected site(s).


Here is the link for the notebook: [IBM-Data-Science-Capstone/spacex_dash_app.py](https://github.com/kerygot16/IBM-Data-Science-Capstone/blob/main/IBM-Data-Science-Capstone/spacex_dash_app.py) at main · kerygot16/IBM-Data-Science-Capstone (github.com)

Predictive Analysis (Classification)


I loaded the data using numpy and pandas, transformed the data, split into training and testing.



I built different machine learning models such as KNeighbors, Decision Tree, Logistic Regression and Support Vector. And then I tune different hyper parameters using GridSearch CV



I used accuracy as the metric for our model, improved the model using feature engineering and algorithm tubing.



And finally, I found the best performing classification model which is the Decision Tree with 87% score.

Here is the link of the notebook: [IBM-Data-Science-Capstone/Machine Learning Prediction.ipynb](https://github.com/kerygot16/IBM-Data-Science-Capstone/blob/main/IBM-Data-Science-Capstone/Machine%20Learning%20Prediction.ipynb) at main · kerygot16/IBM-Data-Science-Capstone (github.com)

Results

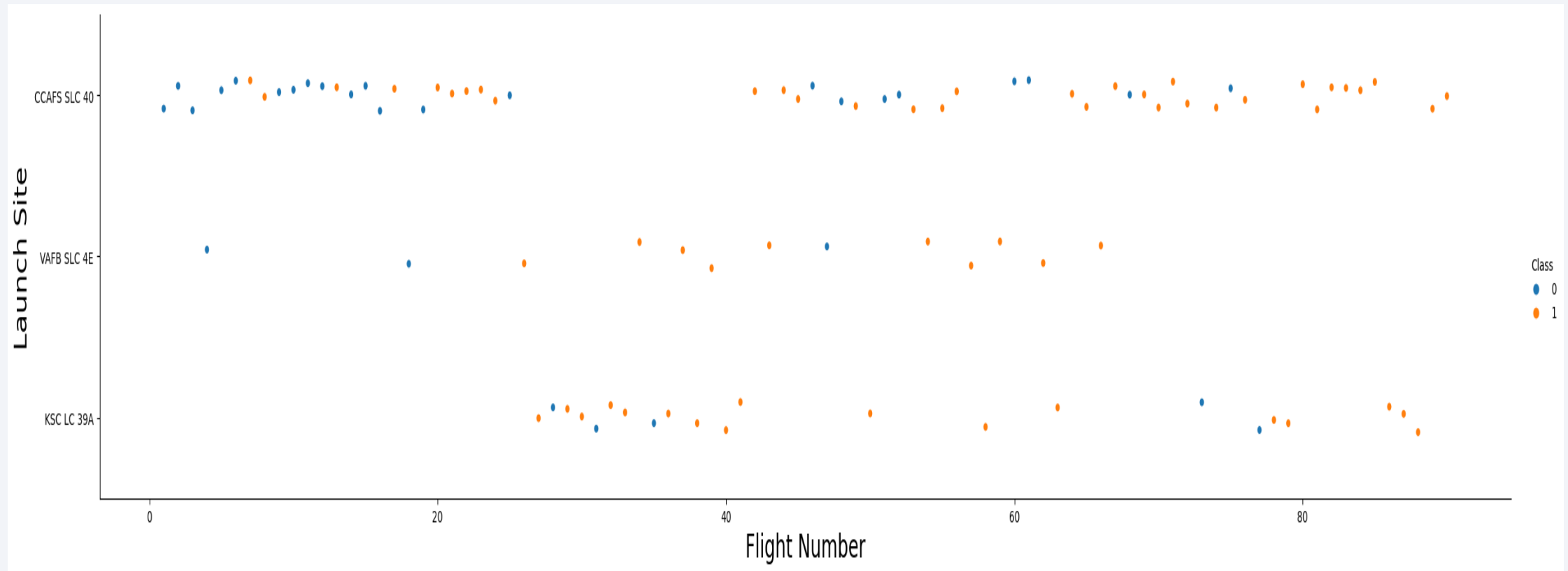
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower half of the image. The overall effect is dynamic and technological.

Section 2

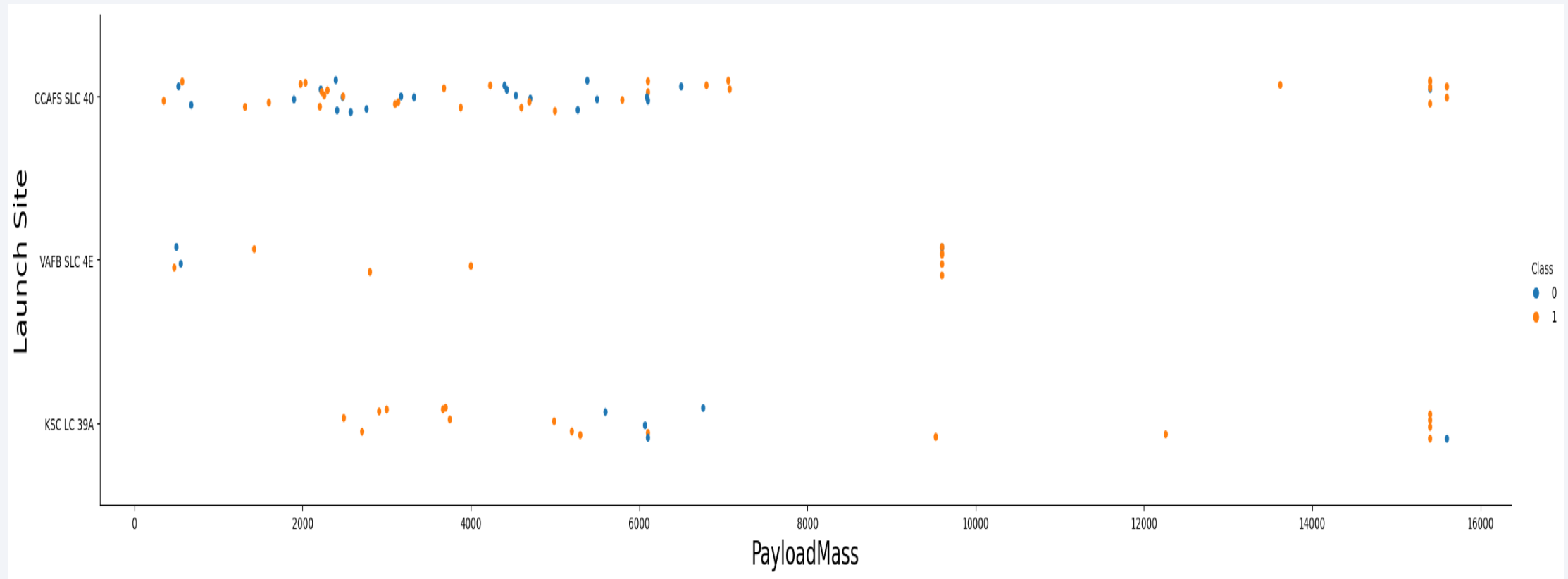
Insights drawn from EDA

Flight Number vs. Launch Site



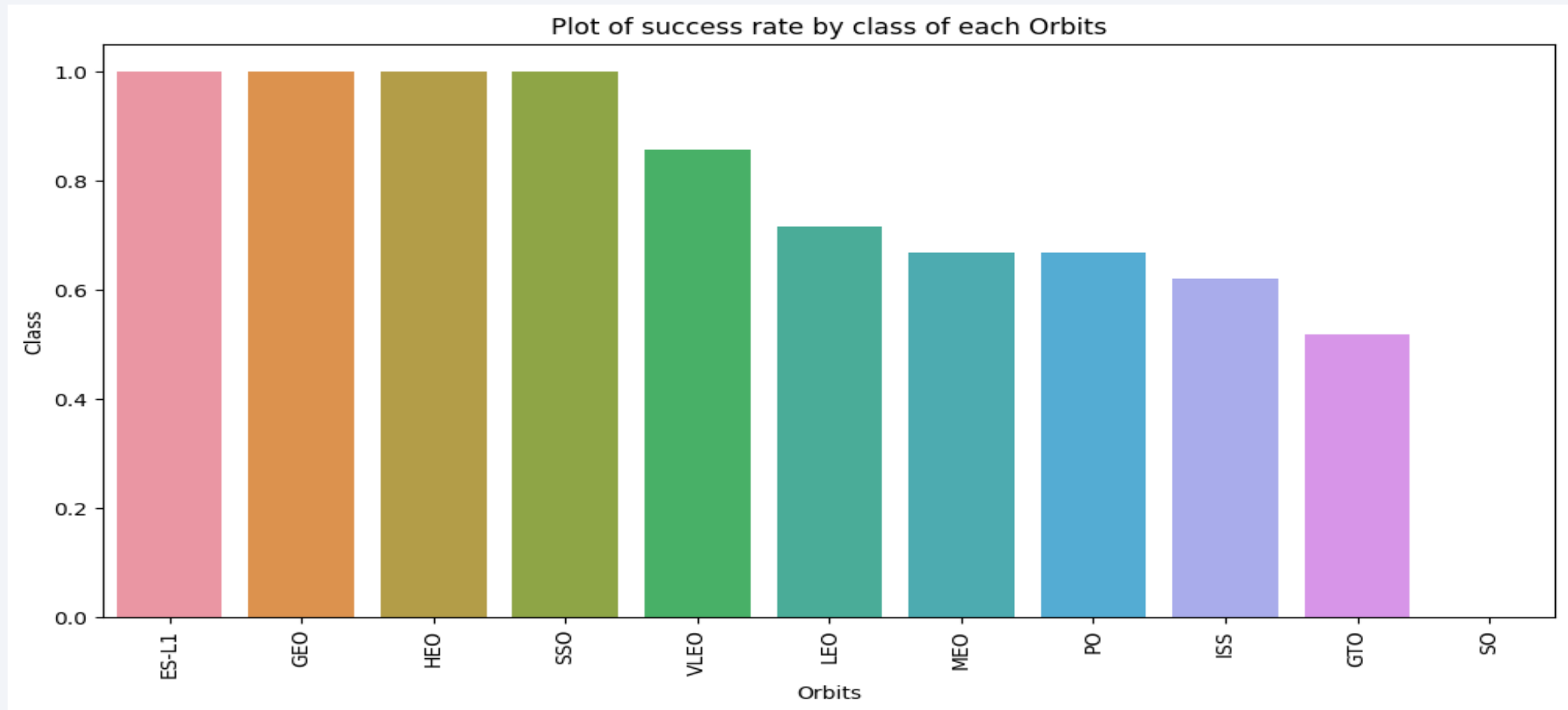
From the plot above, we can infer that the larger the flight amount at a launch site, the greater the success rate at a launch site.

Payload vs. Launch Site



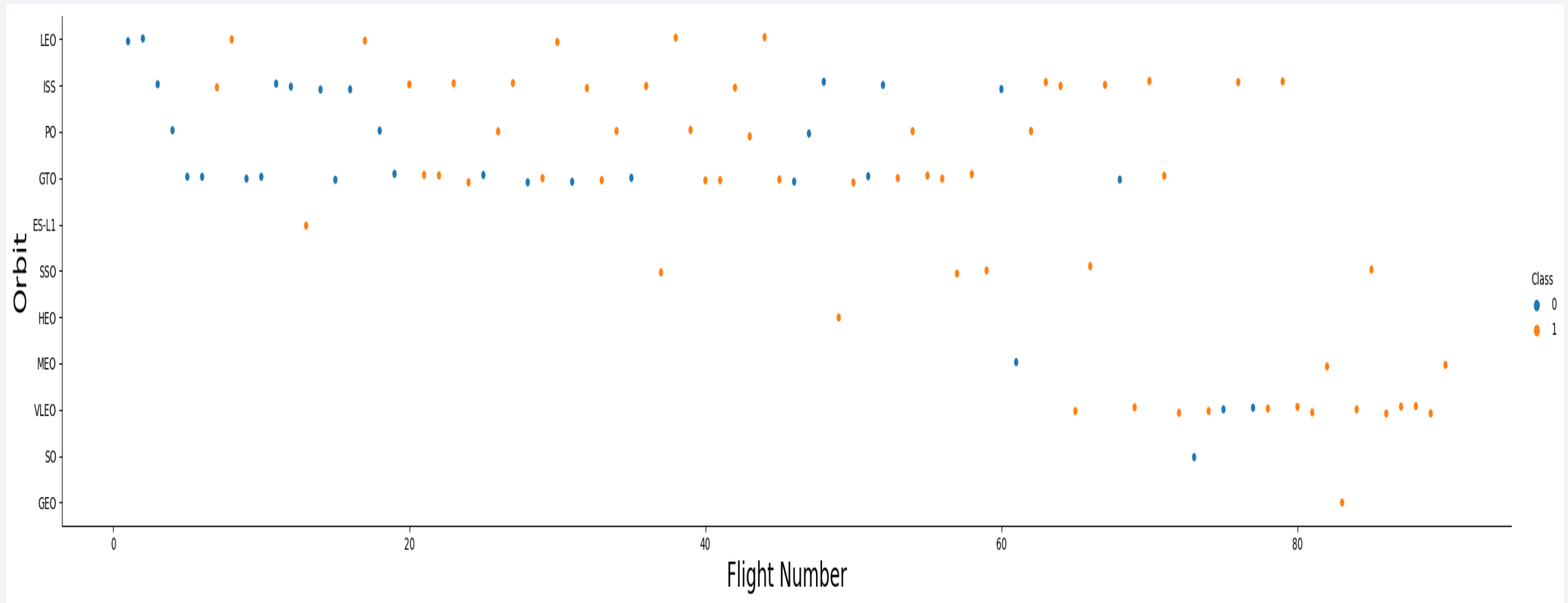
The greater the payload mass for launch site CCAFS SLC 40 the higher the success rate for the rocket.

Success Rate vs. Orbit Type



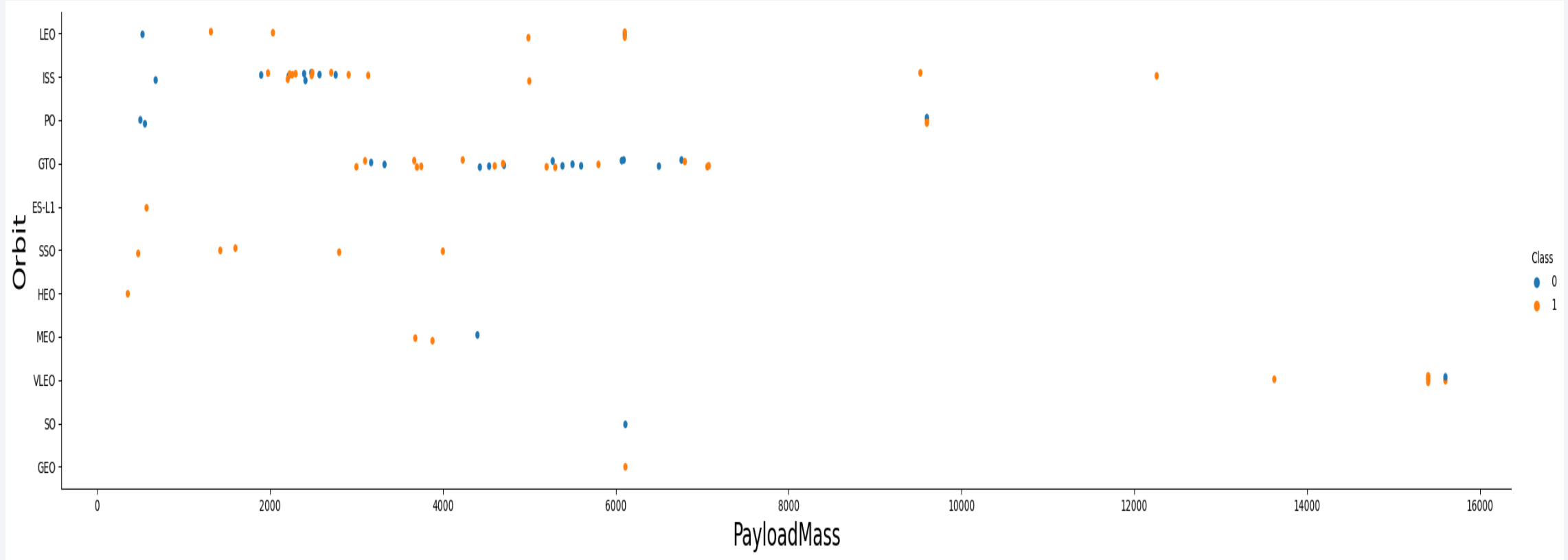
From the plot, we can see that ES-L 1, GEO, HEO, SSO and VLEO had the most success rate.

Flight Number vs. Orbit Type



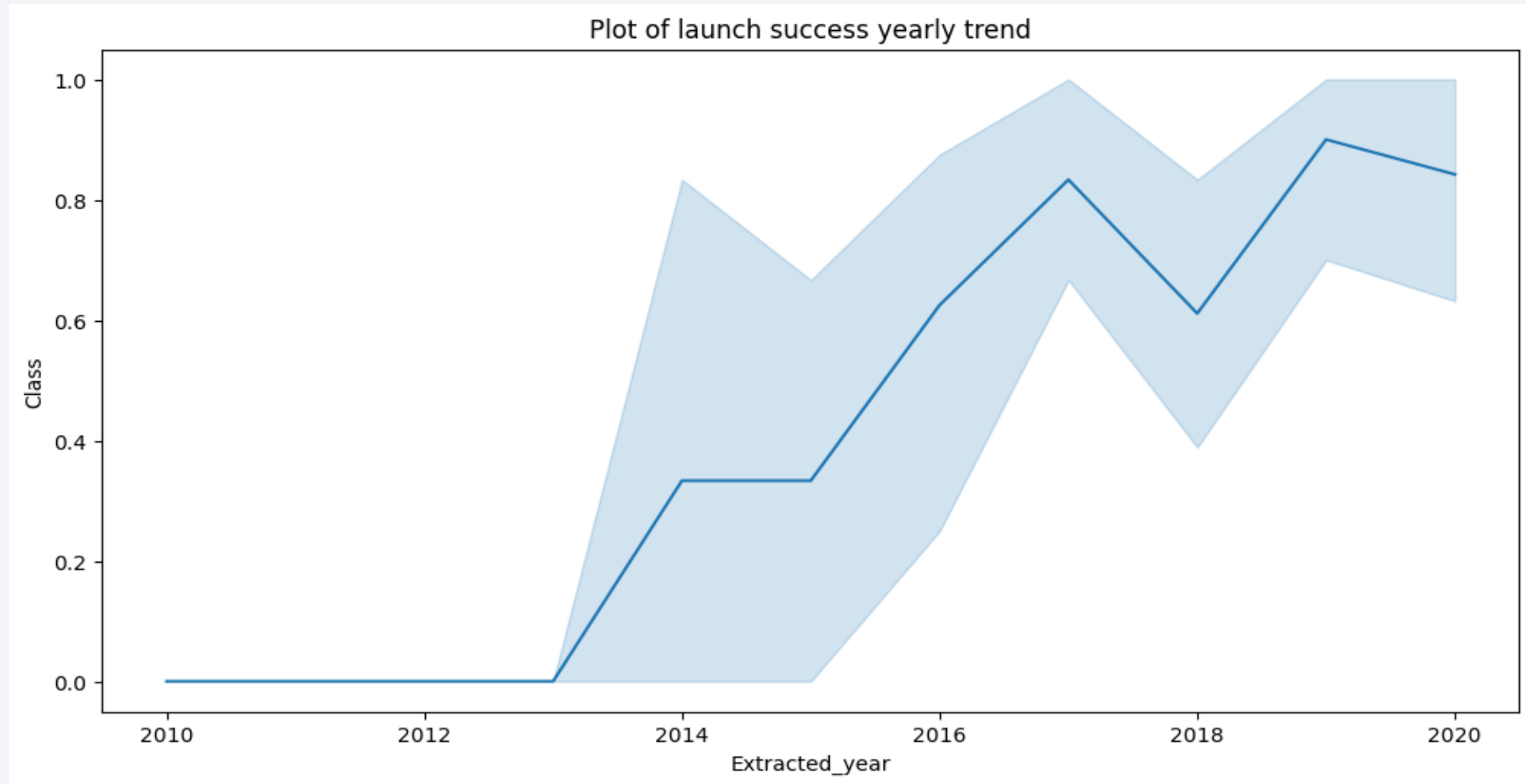
The plot above shows the Flight Number vs. Orbit type. We observe that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship ²¹ between flight number and the orbit.

Payload vs. Orbit Type



We can observe that with heavy payloads, the successful landing are more for PO, LEO and ISS orbits.

Launch Success Yearly Trend



From the plot, we can observe that success rate since 2013 kept on increasing till 2020.

All Launch Site Names

To launch all site names, I used the key word **DISTINCT** to show only unique launch sites from the SpaceX data.



```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

```
* 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'

```
%sql SELECT * FROM SPACEXTABLE 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
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

I used the query above to display 5 records where launch sites begin with `CCA`.

Total Payload Mass

I calculated the total payload carried by boosters from NASA as 45596 using the query below:

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

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

```
Done.
```

```
Total_PayloadMass
```

```
45596
```

Average Payload Mass by F9 v1.1

I calculated the average payload mass carried by booster version F9 v1.1 as 2928.4

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Avg_PayloadMass FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';

* sqlite:///my_data1.db
Done.
Avg_PayloadMass
2928.4
```

First Successful Ground Landing Date

```
%sql SELECT MIN(DATE) AS FirstSuccessful_landing_date FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success (ground pad)';
```

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

```
Done.
```

```
FirstSuccessful_landing_date
```

```
2015-12-22
```

I observed that the dates of the first successful landing outcome on ground pad was 22nd December 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

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

```
Done.
```

```
Booster_Version
```

```
F9 FT B1022
```

```
F9 FT B1026
```

```
F9 FT B1021.2
```

```
F9 FT B1031.2
```

I used the **WHERE** clause to filter for boosters which have successfully landed on drone ship and applied the **AND** condition to determine successful landing with payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
success = %sql SELECT COUNT(Mission_Outcome) AS SuccessOutcome FROM SPACEXTABLE WHERE Mission_Outcome LIKE 'Success%';
failure = %sql SELECT COUNT(Mission_Outcome) AS FailureOutcome FROM SPACEXTABLE WHERE Mission_Outcome LIKE 'Failure%';

print('The total number of successful mission outcome is:', success)
print('The total number of failed mission outcome is:', failure)
```

```
* sqlite:///my_data1.db
Done.
* sqlite:///my_data1.db
Done.
The total number of successful mission outcome is: +-----+
| SuccessOutcome |
+-----+
|      100      |
+-----+
The total number of failed mission outcome is: +-----+
| FailureOutcome |
+-----+
|       1       |
+-----+
```

I used wildcard like '%' to filter for WHERE Mission Outcome was a success or a failure.

Boosters Carried Maximum Payload

I determined the booster that have carried the maximum payload using a subquery in the **WHERE** clause and the **MAX()** function.

```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_)
FROM SPACEXTABLE) ORDER BY Booster_Version;
```

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

```
Done.
```

```
Booster_Version PAYLOAD_MASS__KG_
```

F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

I used a combinations of the WHERE clause, LIKE, AND, and BETWEEN conditions to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015.

```
%sql SELECT Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTABLE  
WHERE Landing_Outcome LIKE 'Failure (drone ship)' AND Date BETWEEN '2015-01-01' AND '2015-12-31';
```

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

```
Done.
```

Booster_Version	Launch_Site	Landing_Outcome
-----------------	-------------	-----------------

F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
---------------	-------------	----------------------

F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)
---------------	-------------	----------------------

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

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

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

```
Done.
```

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

- I selected Landing outcomes and the COUNT of landing outcomes from the data and used the WHERE clause to filter for landing outcomes BETWEEN 2010-06-04 to 2010-03-20.
- I applied the GROUP BY clause to group the landing outcomes and the ORDER BY clause to order the grouped landing outcome in descending order.

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

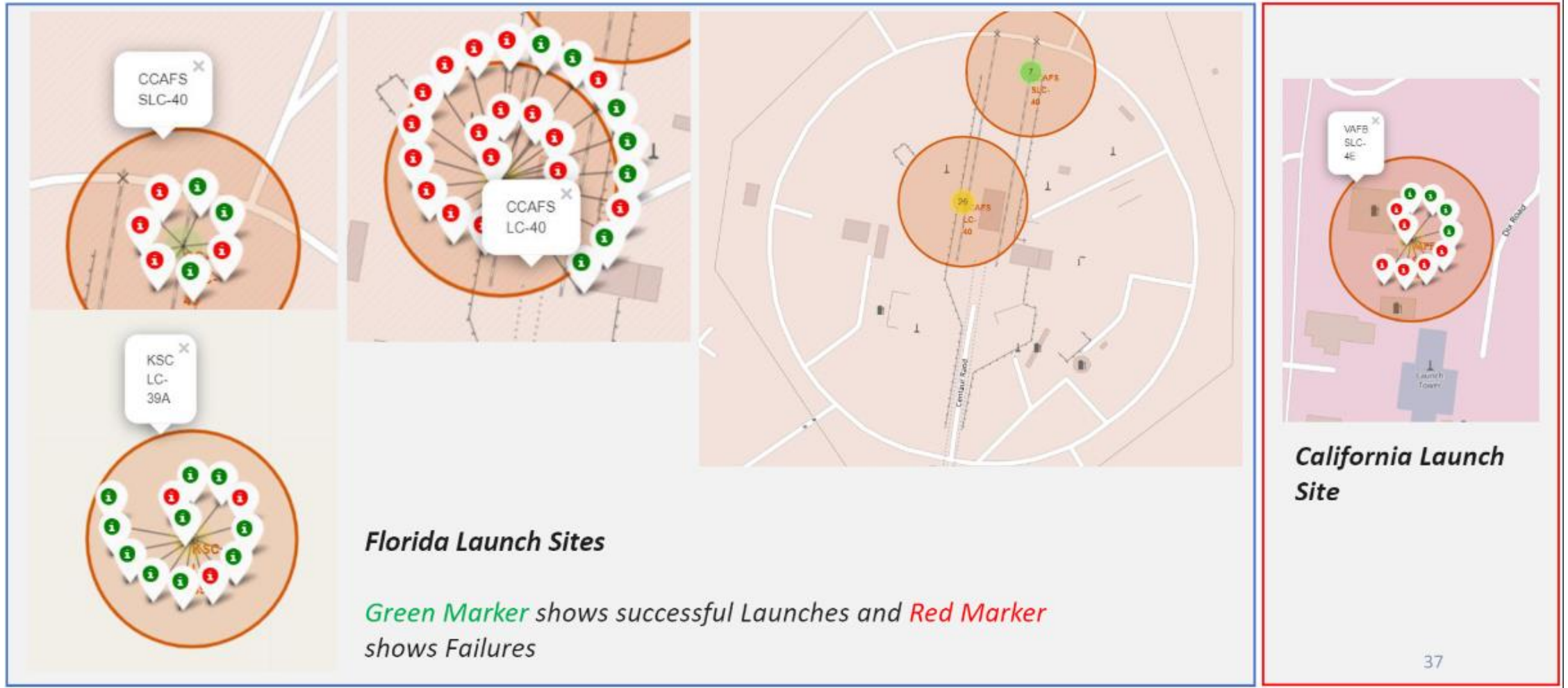
Section 3

Launch Sites Proximities Analysis

All launch sites global map markers



Markers showing launch sites with color labels



Launch Site distance to landmarks



- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes



Section 4

Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site

Total Success Launches By all sites



We can see that KSC LC-39A had the most successful launches from all the sites

Pie chart showing the Launch site with the highest launch success ratio



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider

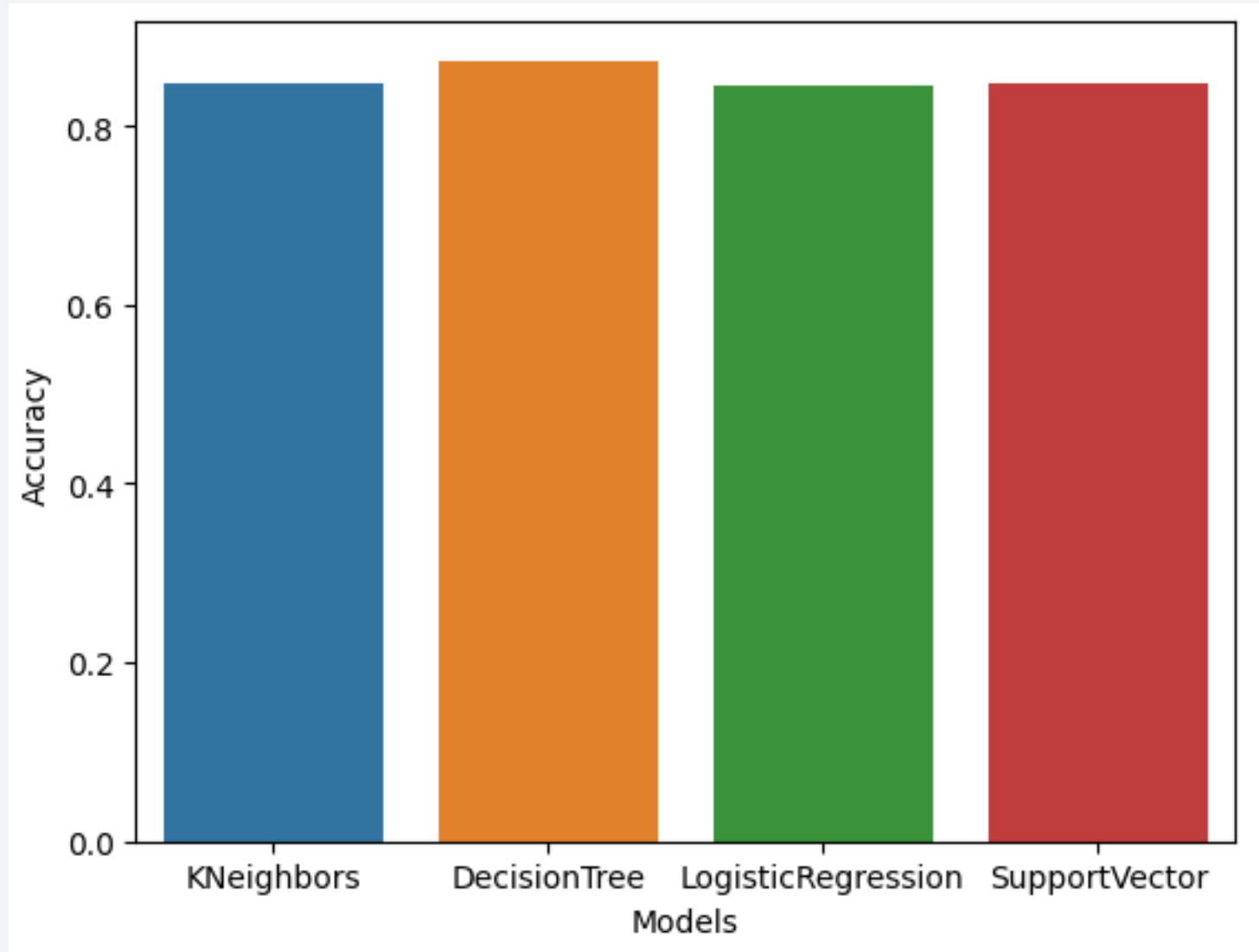


We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

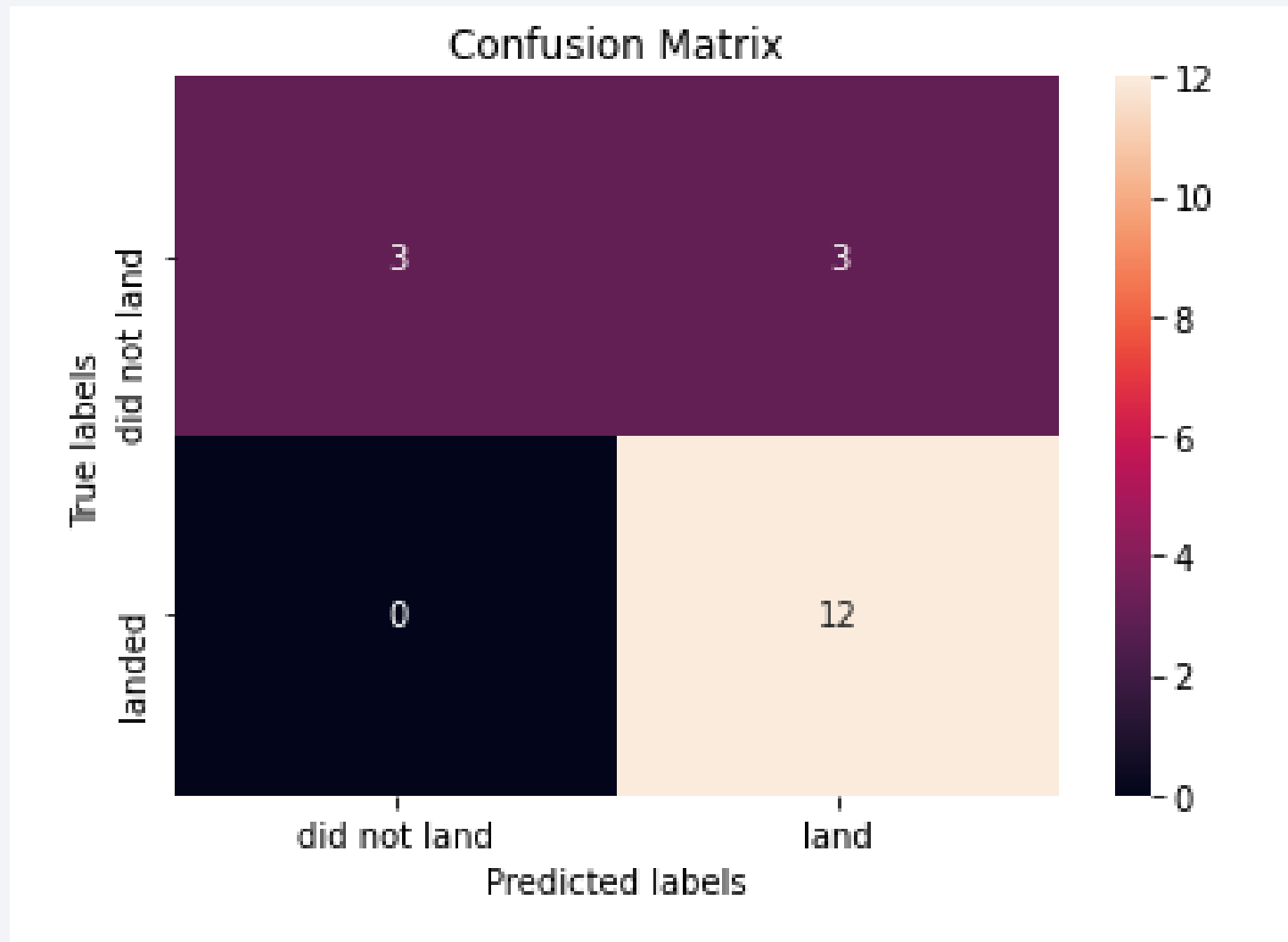
Predictive Analysis (Classification)

Classification Accuracy



Based on the bar chart on the left, the decision tree classifier is the model with the highest classification accuracy

Confusion Matrix



The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.

Conclusions

We can conclude that:

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task.

Appendix

- [IBM-Data-Science-Capstone/Data Collection API.ipynb at main · kerygot16/IBM-Data-Science-Capstone \(github.com\)](#)
- [IBM-Data-Science-Capstone/Data Collection With Webscraping.ipynb at main · kerygot16/IBM-Data-Science-Capstone \(github.com\)](#)
- [IBM-Data-Science-Capstone/Data Wrangling.ipynb at main · kerygot16/IBM-Data-Science-Capstone \(github.com\)](#)
- [IBM-Data-Science-Capstone/EDA with Data Visualization.ipynb at main · kerygot16/IBM-Data-Science-Capstone \(github.com\)](#)
- [IBM-Data-Science-Capstone/EDA with SQL.ipynb at main · kerygot16/IBM-Data-Science-Capstone \(github.com\)](#)
- [IBM-Data-Science-Capstone/Interactive Visual Analytics with Folium.ipynb at main · kerygot16/IBM-Data-Science-Capstone \(github.com\)](#)
- [IBM-Data-Science-Capstone/spacex_dash_app.py at main · kerygot16/IBM-Data-Science-Capstone \(github.com\)](#)
- [IBM-Data-Science-Capstone/Machine Learning Prediction.ipynb at main · kerygot16/IBM-Data-Science-Capstone \(github.com\)](#)

Thank you!

