IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gareth Last
28th September 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Interactive Visual Analytics with Folium
  - Predictive Analysis with Machine Learning
- Summary of all results
  - Exploratory Data Analysis
  - Interactive Analytics in Screenshots
  - Predictive Analytics 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.

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 SpaceX for a rocket launch.

- Problems to answer:
  - 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 does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API by requesting and parsing the SpaceX data using the GET request. Web scraping from Wikipedia
- Perform data wrangling
  - Data wrangling was performed with some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised model.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Discovering new patterns in the data with visualization techniques such as scatter plots.
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification machine models were built to achieve this goal.

Data Collection

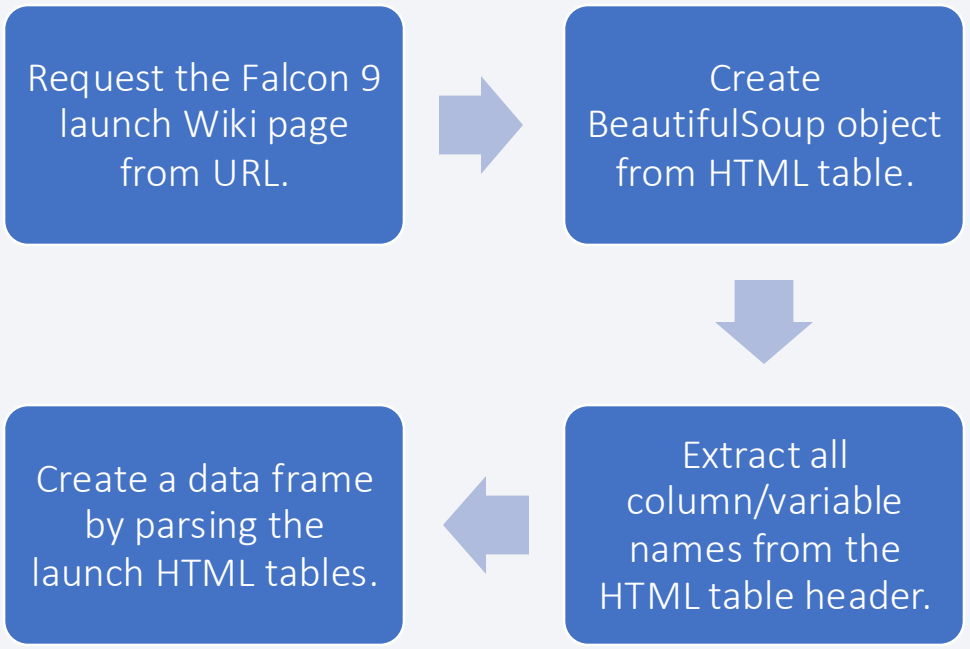
- Data sets were collected and processed through the following steps:
  - Data collection was done using the GET request to the SPaceX API.
  - We decoded the response content as a Json using .json() function call and turned it into a pandas dataframe using the pandas library. Using .json_normalize().
  - The data was cleaned, checked for missing values and these were filled in where necessary.
  - Web scraping from Wikipedia for Falcon 9 launch records was done using BeautifulSoup.

Data Collection – SpaceX API

- We used the GET request to the SPaceX API to collect data, clean the requested data and did some data wrangling and formatting to store the data in a data-frame.
- Link to the notebook is: https://github.com/GarethLast1/CapstoneSpaceX/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

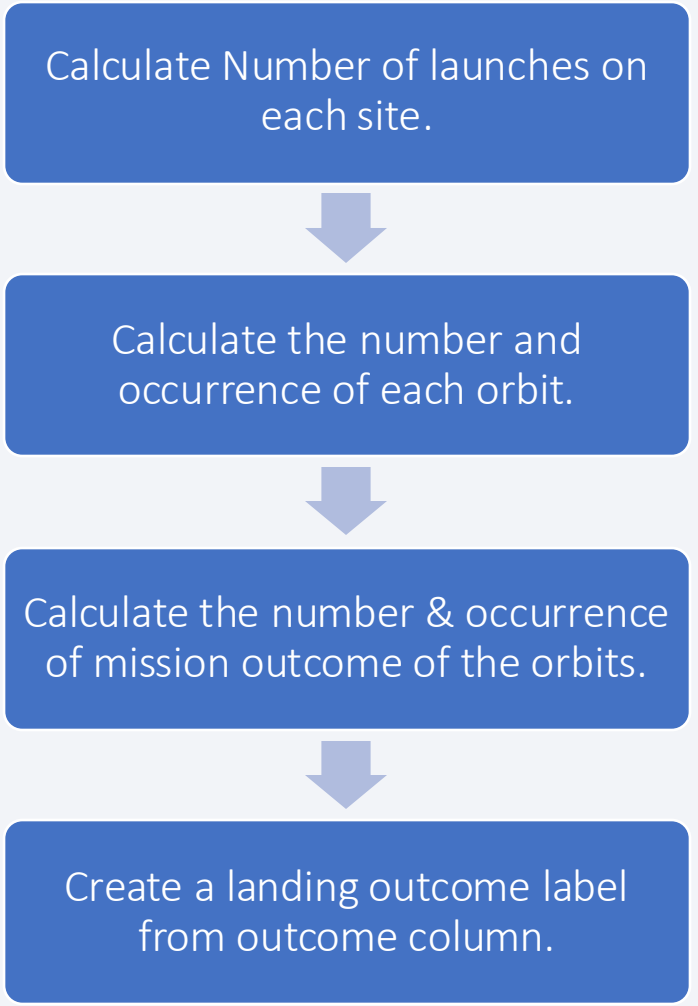
Data Collection - Scraping

- We applied web scraping to web-scrap Falcon 9 launch records with Beautiful Soup.
- We parsed the table and converted it into Pandas data-frame.
- Link to the notebook is: https://github.com/GarethLast1/CapstoneSpaceX/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- We performed exploratory data analysis and determined the training labels.
- We calculated the number of launches at each site, and the number and occurrence of each orbit.
- We created the landing outcome form the outcome column and exported the results to a CSV file.
- The GitHub URL to the notebook is: https://github.com/GarethLast1/CapstoneSpaceX/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Summary
  - Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
  - We visualized the relationship between flight number and launch site.
  - Visualized the relationship between payload mass and launch site.
  - Visualized the relationship between success rate of each orbit type.
  - Visualized the relationship between FlightNumber and orbit type.
  - Visualized the relationship between Payload Mass and orbit type.
  - Visualized the launch success yearly trend.

  NOTE: Screenshots in second section of presentation (Insights drawn from EDA)
- The GitHub URL of the completed EDA with data visualization notebook: https://github.com/GarethLast1/CapstoneSpaceX/blob/main/edadataviz.ipynb

EDA with SQL

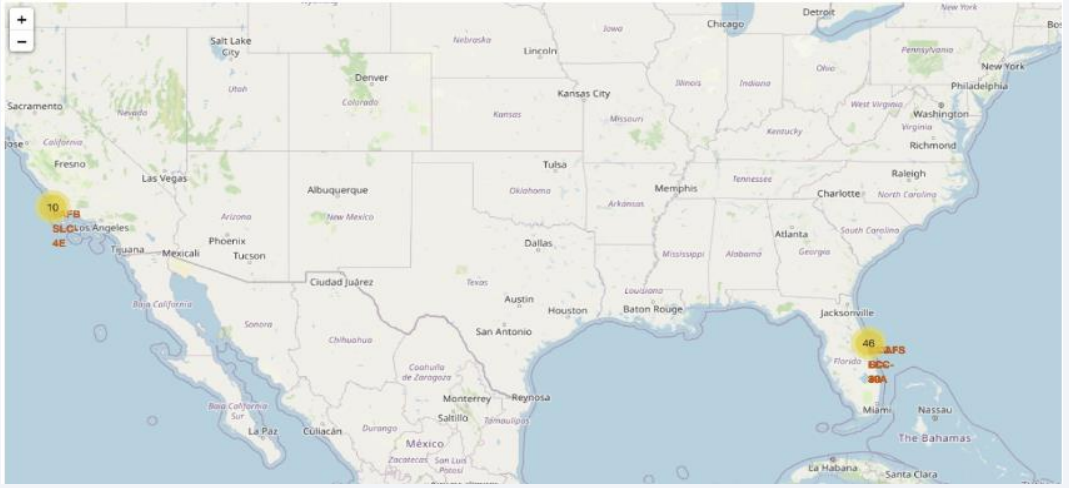
- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name begins with the string 'CCA';
  - Total pay load mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing out comes in drone ship, their booster versions, and launch site names for in year 2015;
  - Rank of the count of landing outcomes (such as Failure (droneship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

  NOTE: Screenshots of SQL queries in next section of presentation (Insights drawn from EDA)

- The GitHub URL of the completed EDA with SQL notebook source code: https://github.com/GarethLast1/CapstoneSpaceX/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

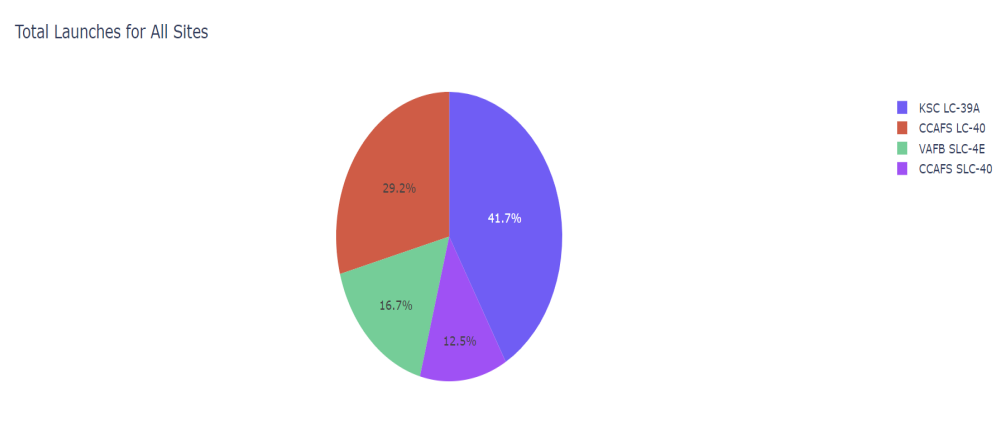
- The map objects such as markers, circles, lines, etc. that I created and added to a folium map
  - Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site;
  - Lines are used to indicate distances between two coordinates.

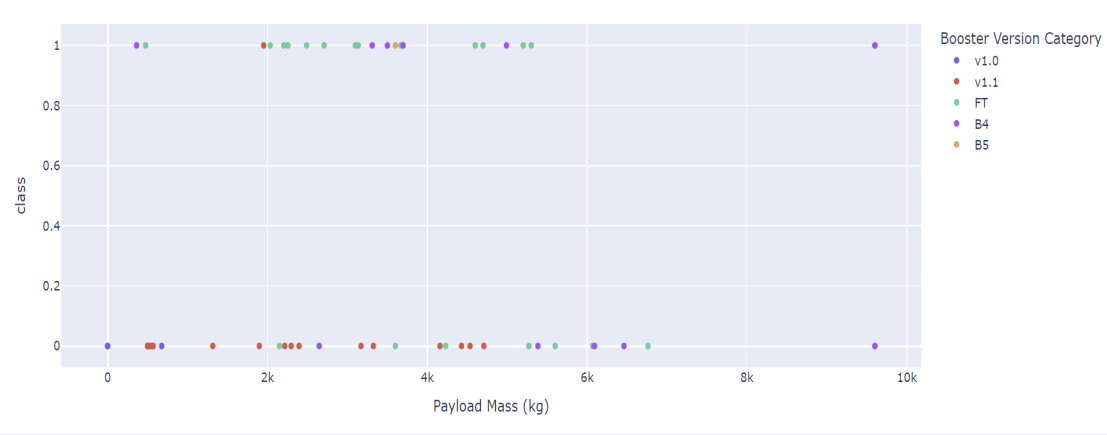


- The GitHub URL of the completed interactive map with Folium map: https://github.com/GarethLast1/CapstoneSpaceX/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Summary of plots/graphs and interactions that were added to a dashboard
  - We built an interactive dashboard with Plotly dash
  - We plotted pie charts showing the total launches by a certain sites
  - We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.

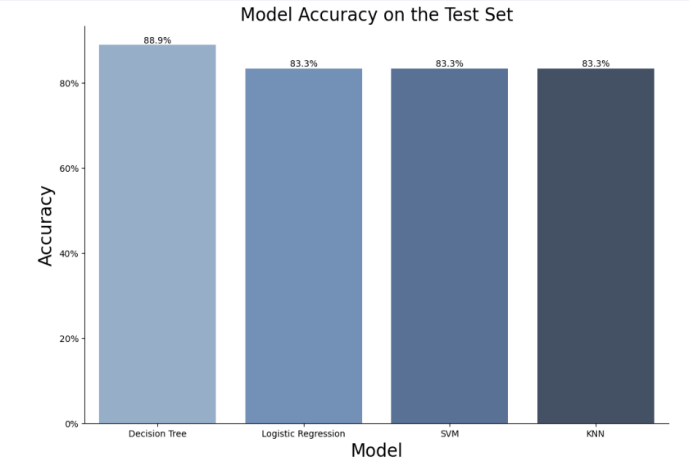


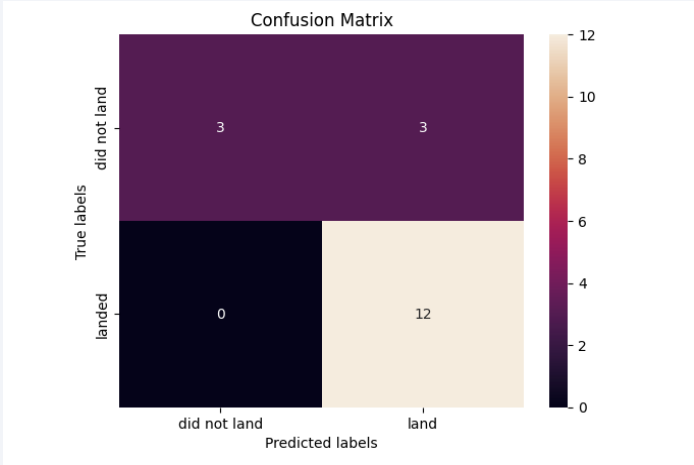


- The GitHub URL of the completed Plotly Dash lab:
  https://github.com/GarethLast1/CapstoneSpaceX/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- How we built, evaluated, improved, and found the best performing classification model
  - We loaded the data using numpy and pandas, transformed the data, split our data into training and testing.
  - We built different machine learning models and tune different hyperparameters using GridSearchCV.
  - We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
  - We found the best performing classification model.





We see that the decision tree model achieved the highest accuracy.

- **The GitHub URL of the completed predictive analysis lab:**

https://github.com/GarethLast1/CapstoneSpaceX/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Load data using numpy and pandas, transformed the data, split our data into training and testing.

Built different machine learning models and tune different hyperparameters using GridSearchCV.

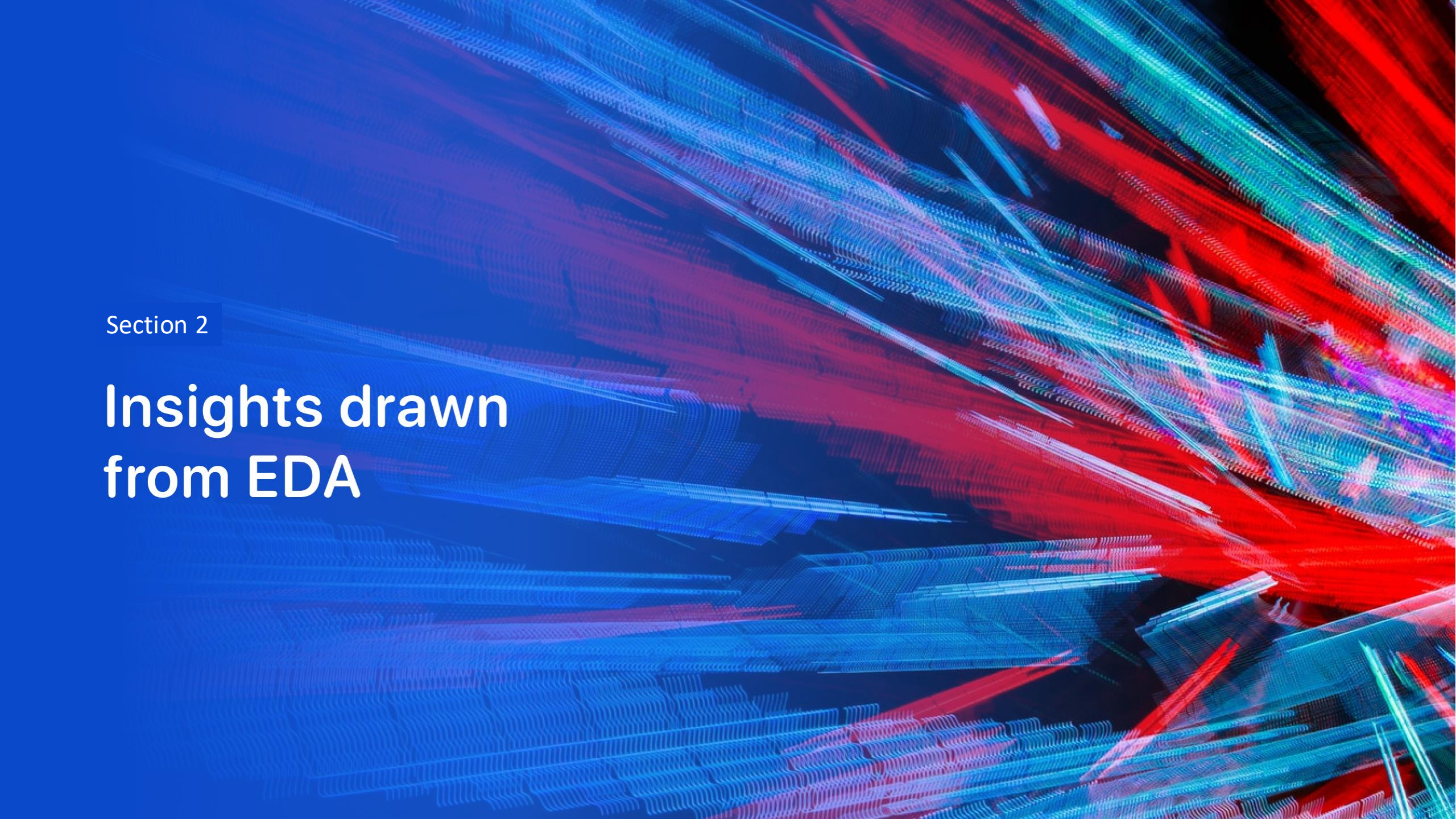
Used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.

Found the best performing classification model.

Results

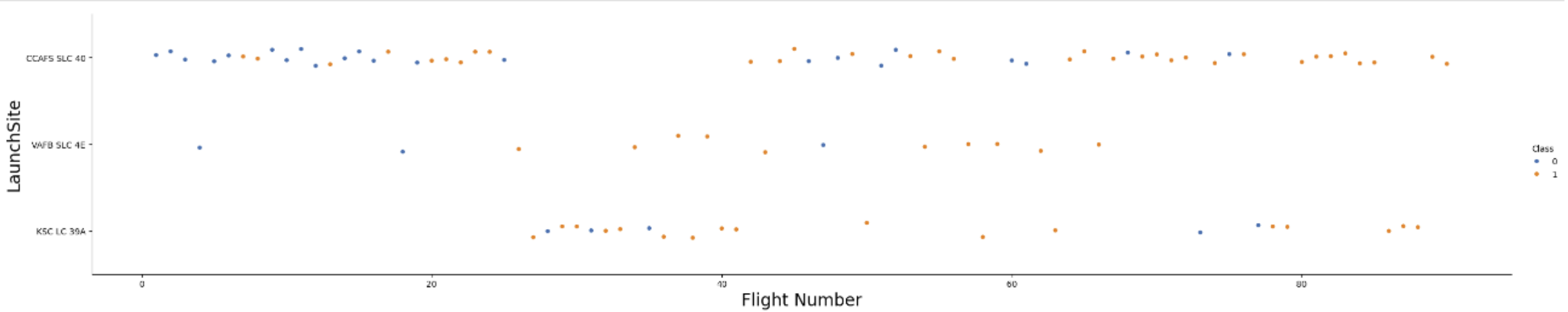
- Exploratory data analysis results
  - SpaceX uses 4 different launch sites.
  - The first launches were done to Space X itself and NASA.
  - The average payload of F9 v1.1 booster is 2,928 kg
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average.
  - Low weighted payloads perform better than heavier payloads.
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015
  - The number of landing outcomes became as better years passed.

NOTE: Screenshots for Interative map, Plotly dashboard and predictive analysis are in there associated sections.


Section 2
Insights drawn from EDA

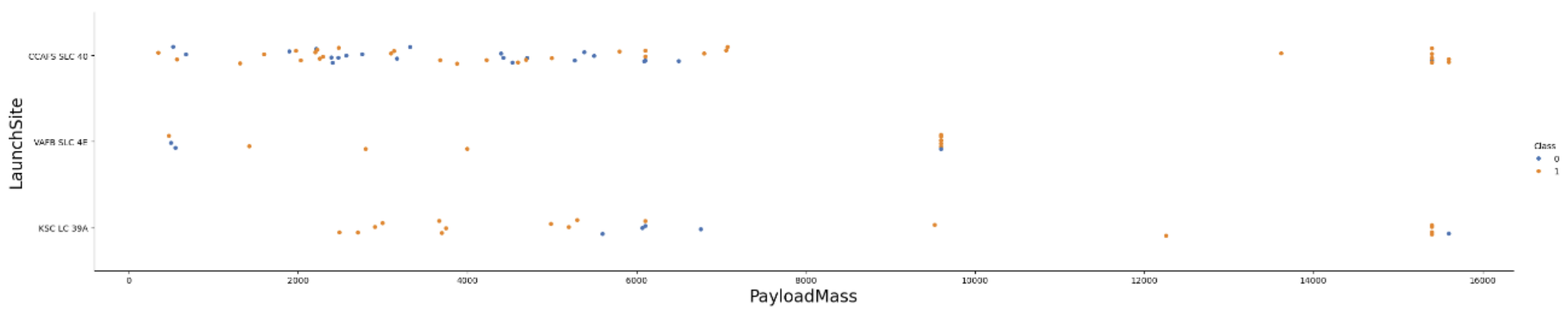
Flight Number vs. Launch Site

From the plot, we see that the larger the flight amount at the launch site, the greater the success rate,



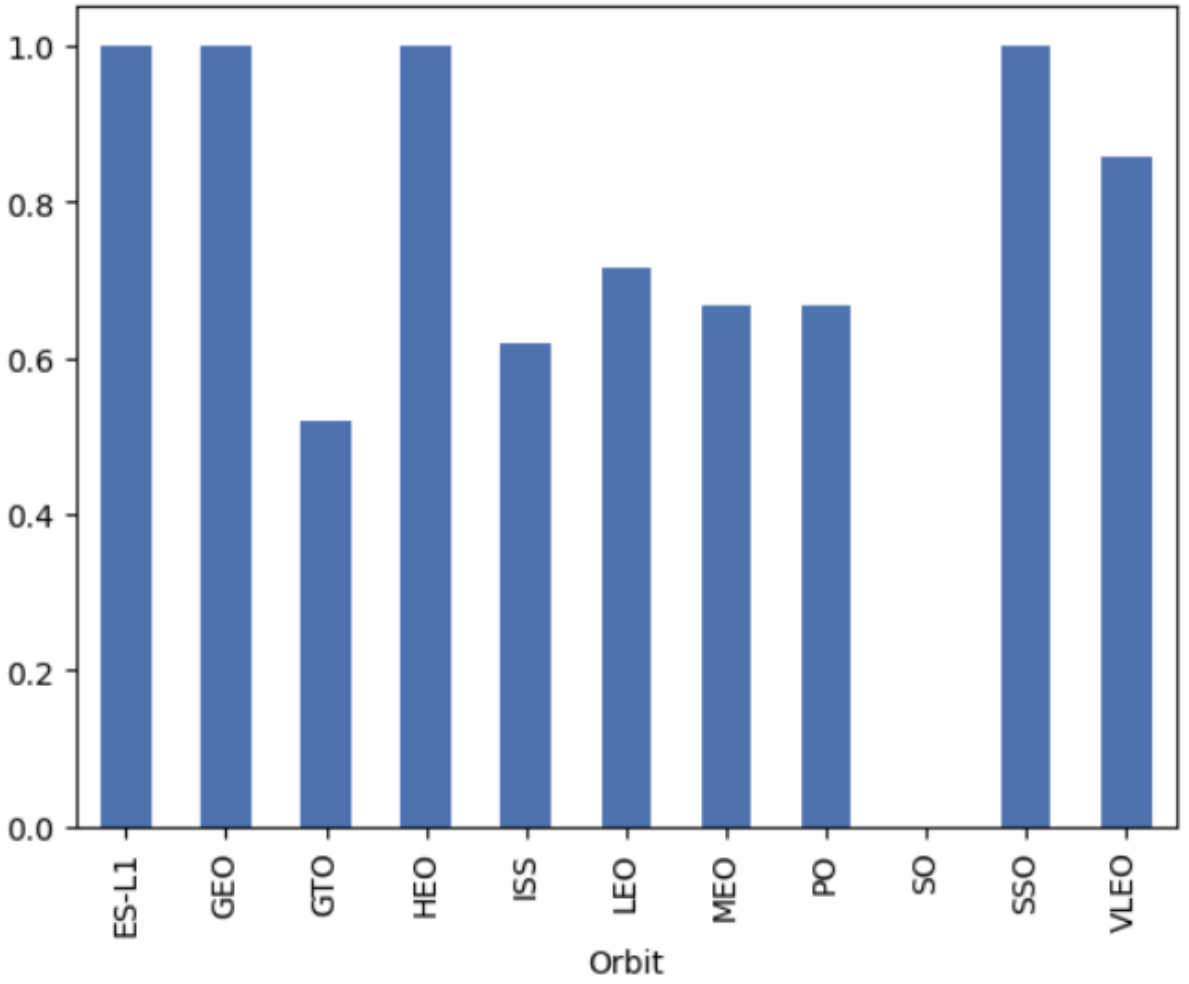
Payload vs. Launch Site

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



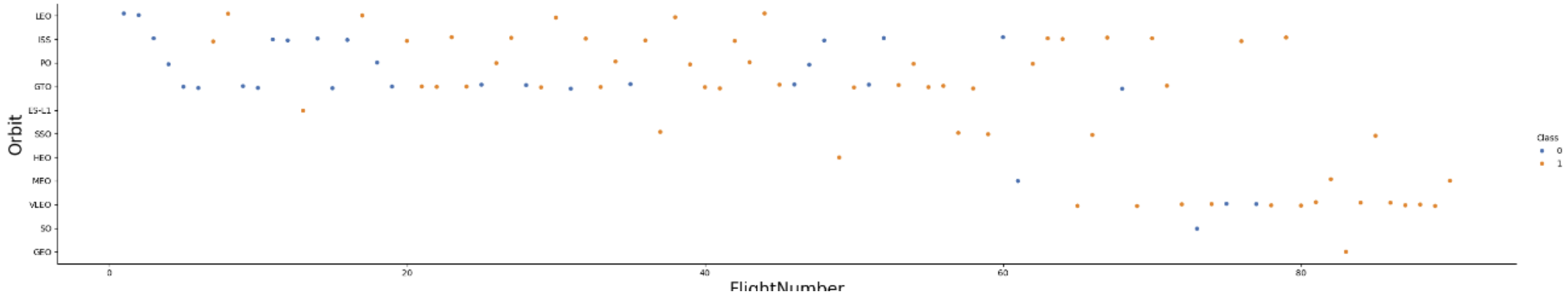
Success Rate vs. Orbit Type

From the plot, we can see that ES-L1, GEO, HEO, SSO had the most success rate.



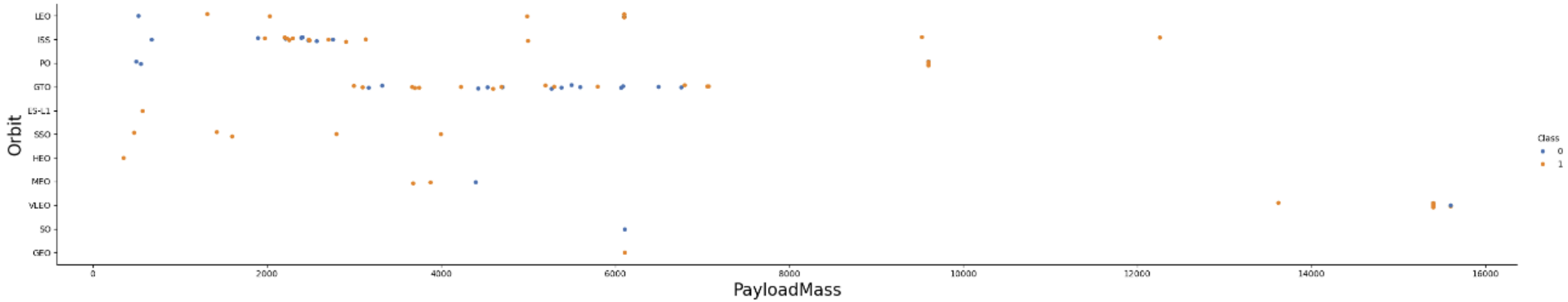
Flight Number vs. Orbit Type

We can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.



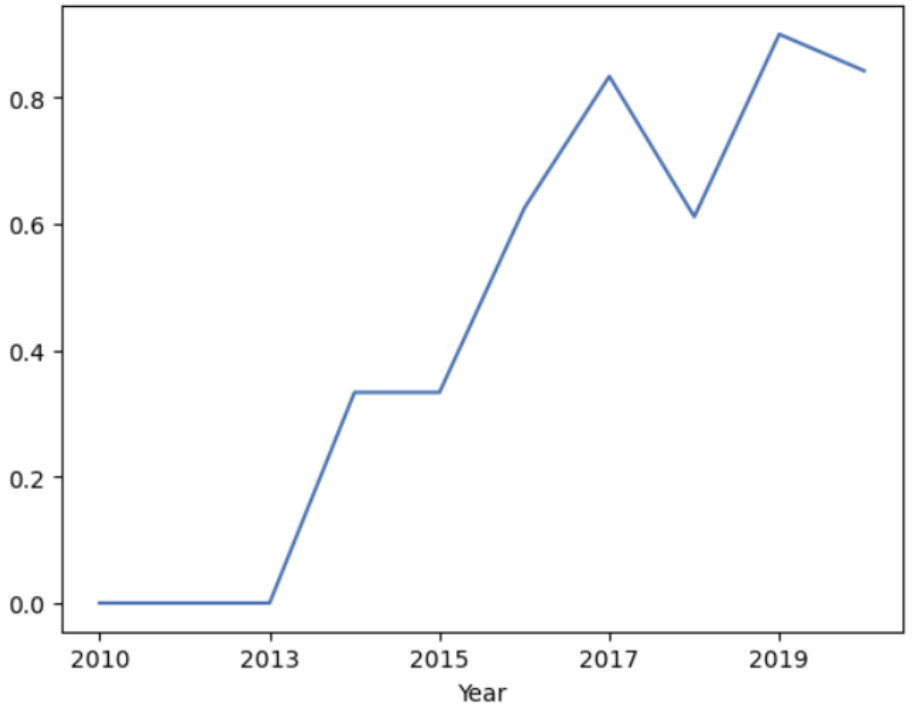
Payload vs. Orbit Type

- 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 success rate since 2013 kept increasing until 2020



All Launch Site Names

We used the key word DISTINCT to show only unique launch sites from the SpaceX data.

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

[18]:
```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTABLE ORDER BY 1;
```

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

[18]:

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

Launch Site Names Begin with 'CCA'

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

[23]:
```
%sql SELECT * FROM SPACEXTABLE WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

[23]:

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

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

```
[1]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTABLE WHERE PAYLOAD  LIKE '%CRS%';
 * sqlite:///my_data1.db
Done.
```

[1]:

TOTAL_PAYLOAD
111268

Average Payload Mass by F9 v1.1

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

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

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

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

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

```
%sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTABLE 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

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 SPACEXTABLE 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

Only 1 Failure in flight

```
%sql SELECT Mission_Outcome, COUNT(*) AS QTY FROM SPACEXTABLE GROUP BY Mission_Outcome ORDER BY Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

The booster which have carried the maximum payload mass

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE) ORDER BY Booster_Ver
```

```
 * 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

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

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

Booster_Version	Launch_Site

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

We 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.

We applied the GROUP BY clause to group the landing outcomes and the ORDER BY clause to order the grouped landing outcome in descending order.

```
%sql SELECT Landing_Outcome, COUNT(*) AS QTY FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY QTY D
```

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

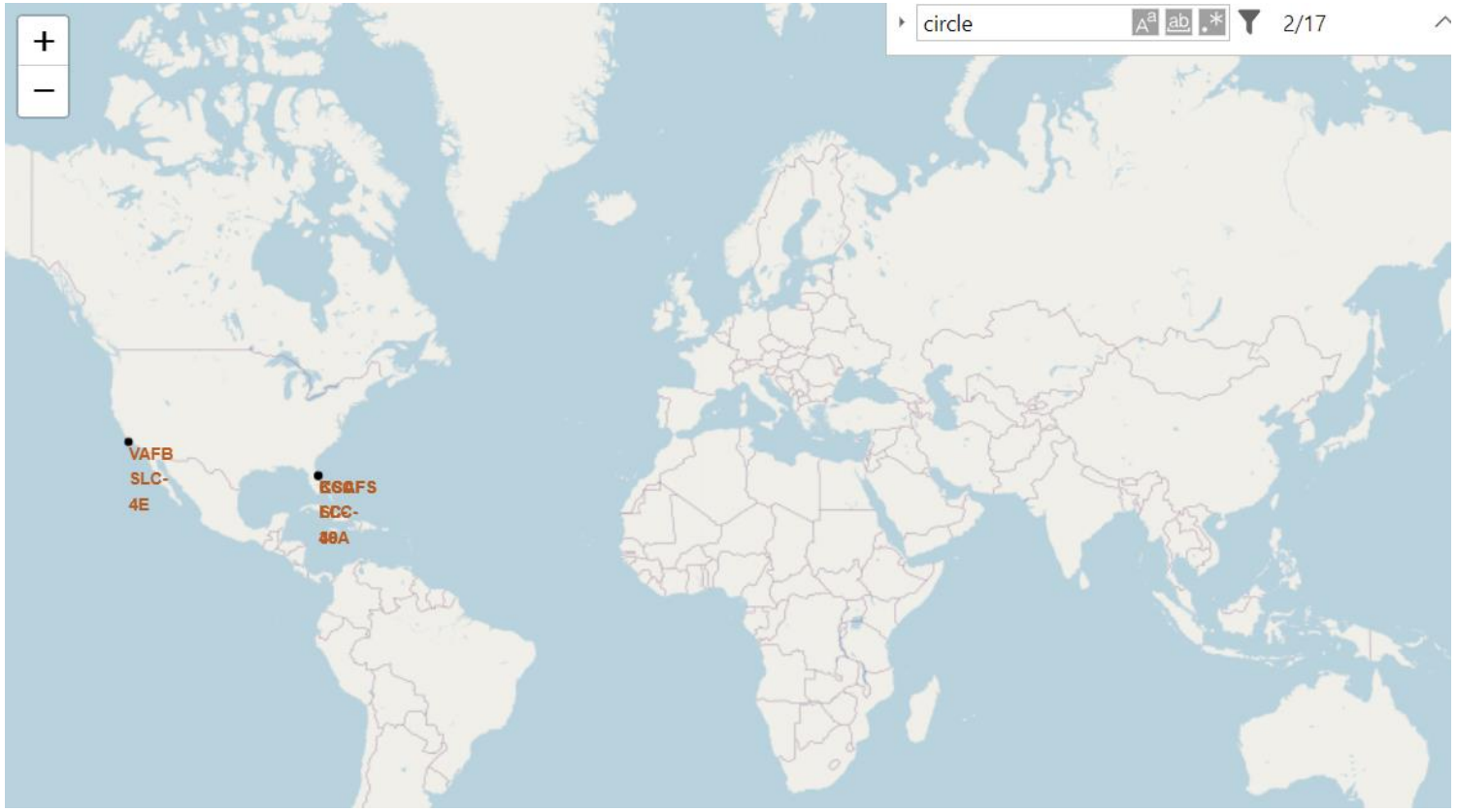
Landing_Outcome	QTY
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

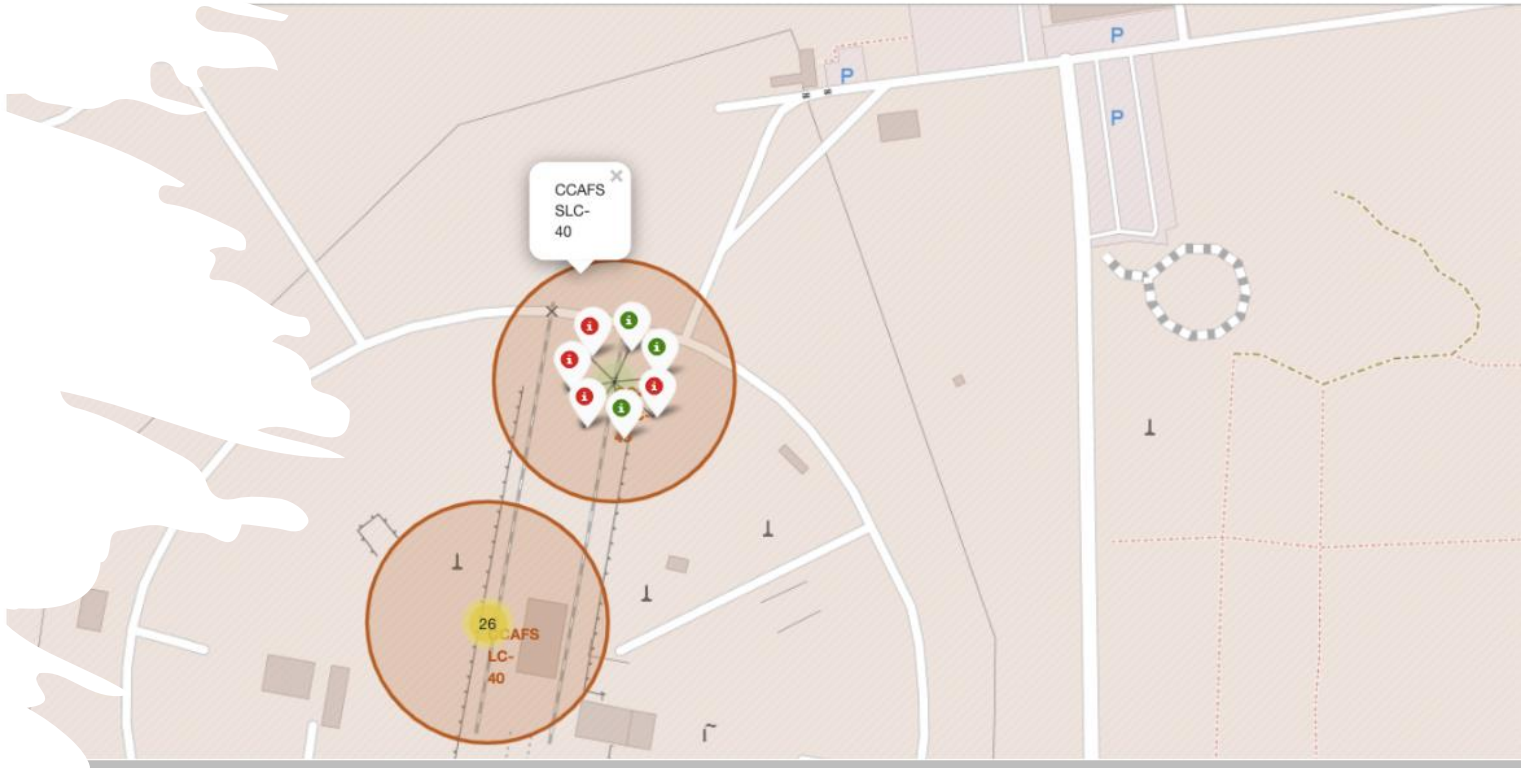
All Launch Sites global map marker

We can see that the SpaceX launch sites are in the United States, East and West coast.



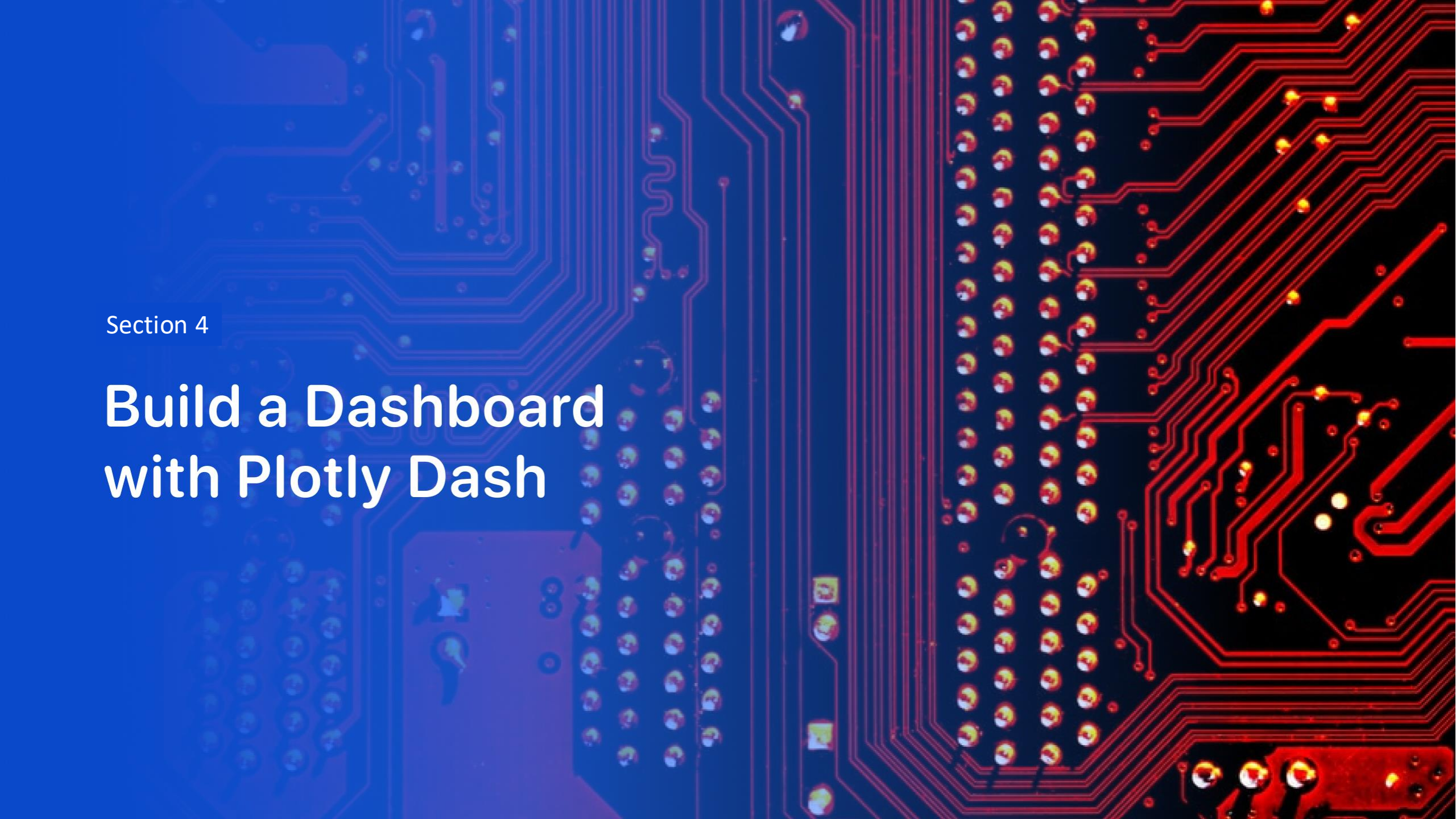
Markers showing launch sites with color labels

By the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.



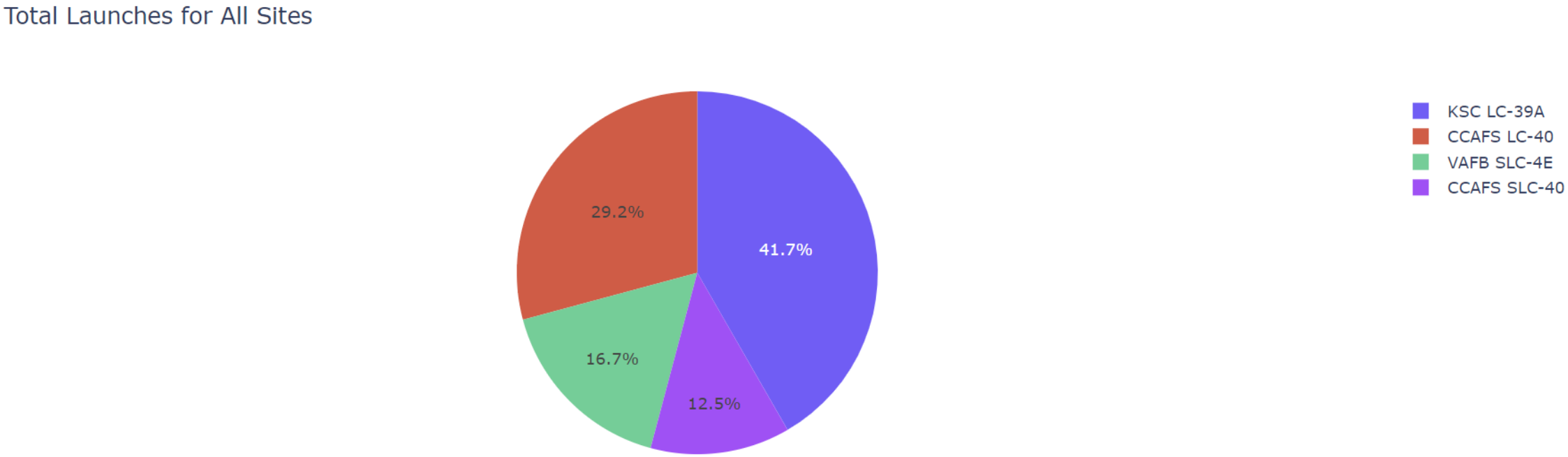
Launchsite distance to landmarks

- 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

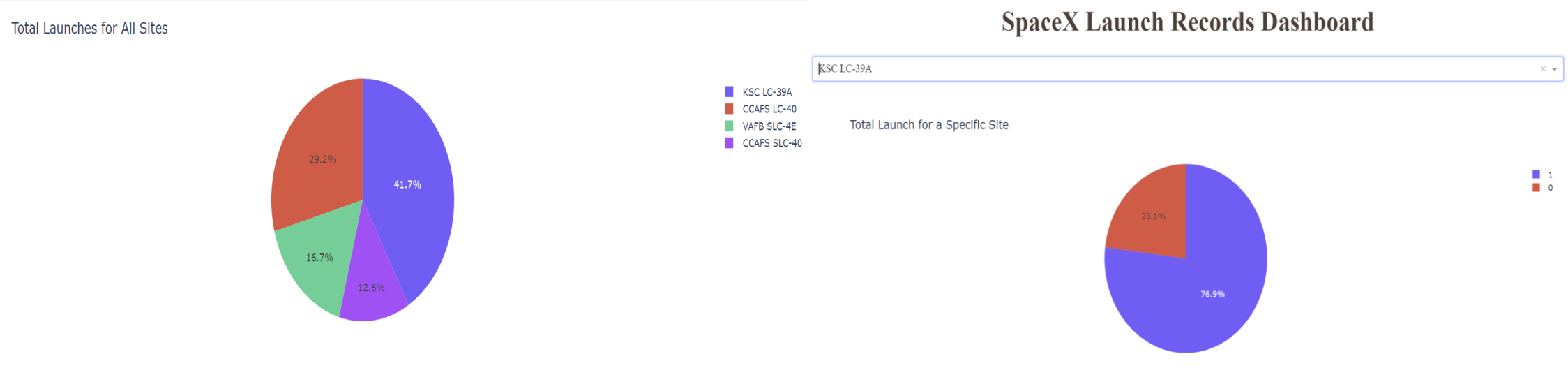
Pie chart showing the success percentage achieved by each launch site.

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



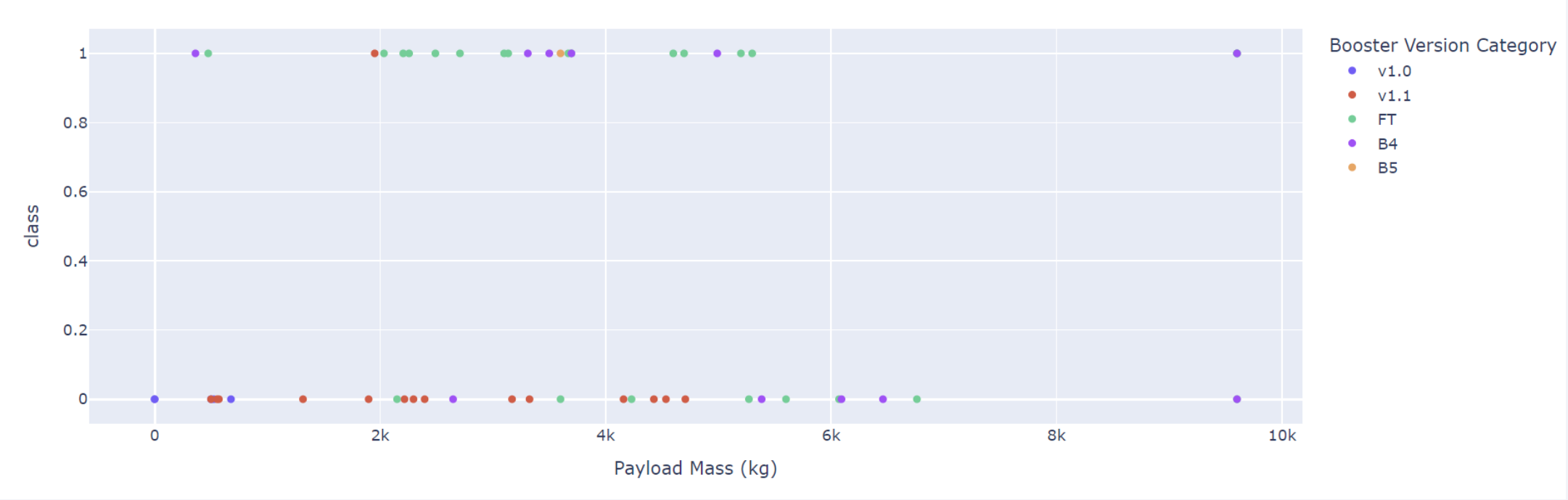
Pie chart showing the launch site with the highest launch success ratio

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



Payload vs. Launch Outcome scatter plot for all sites

Payload vs. Launch Outcome scatter plot for all sites.



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

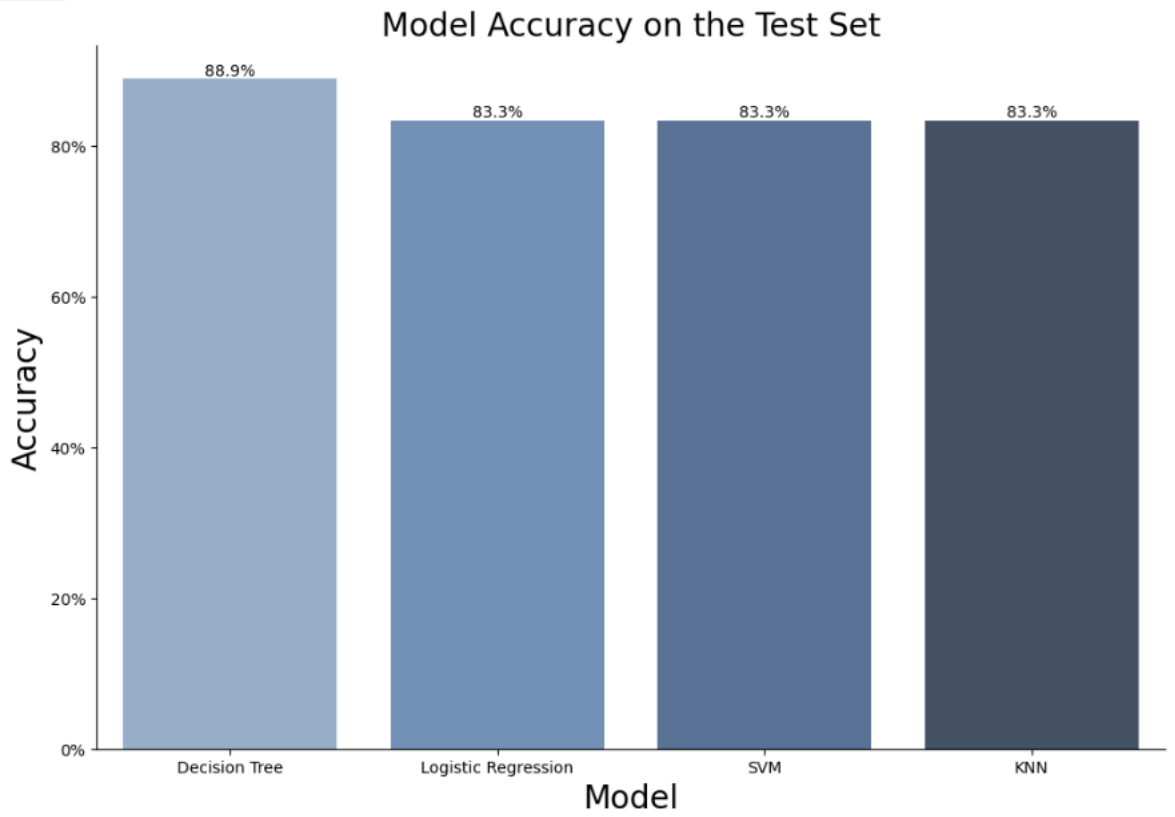
Section 5

Predictive Analysis (Classification)

Classification Accuracy

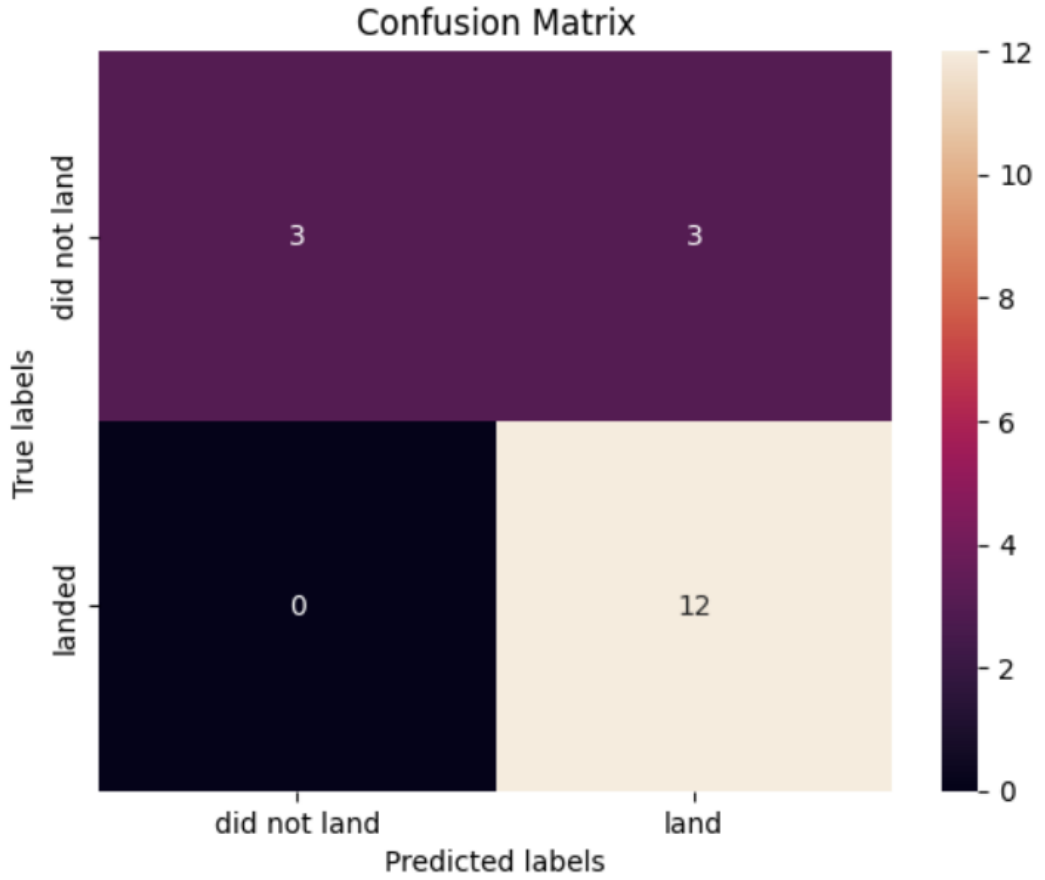
We see that the decision tree model achieved the highest accuracy.

	F1-Score	Precision	Recall	Accuracy
Decision Tree	0.923	0.857	1.0	0.889
Logistic Regression	0.889	0.800	1.0	0.833
SVM	0.889	0.800	1.0	0.833
KNN	0.889	0.800	1.0	0.833



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.

Thank you!