Winning Space Race with Data Science

Luis Otero
1/22/22

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Building an Interactive Map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive Analysis (Classification)
- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive Analytics Demo in Screenshots
  - Predictive Analysis Results

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
  - 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 you want to find answers
  - How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
  - Does the rate of successful landings increase over the years?
  - What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Web API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- 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

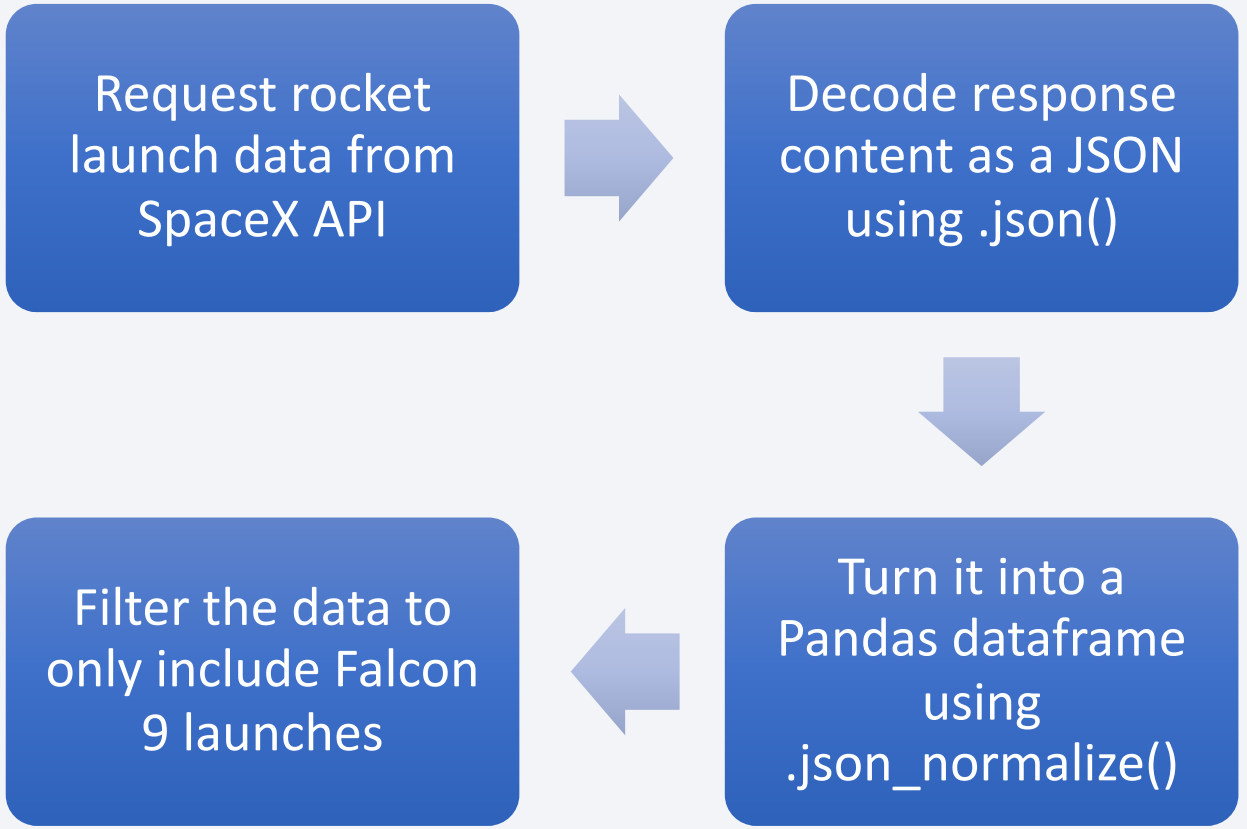
Data Collection

- We obtained the first dataset by requesting rocket launch data from SpaceX API
- We obtained the second dataset by performing web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled “List of Falcon 9 and Falcon Heavy launches”
- We had to use both data collection methods in order to get complete information about the launches for a more detailed analysis.

WIKIPEDIA
The Free Encyclopedia

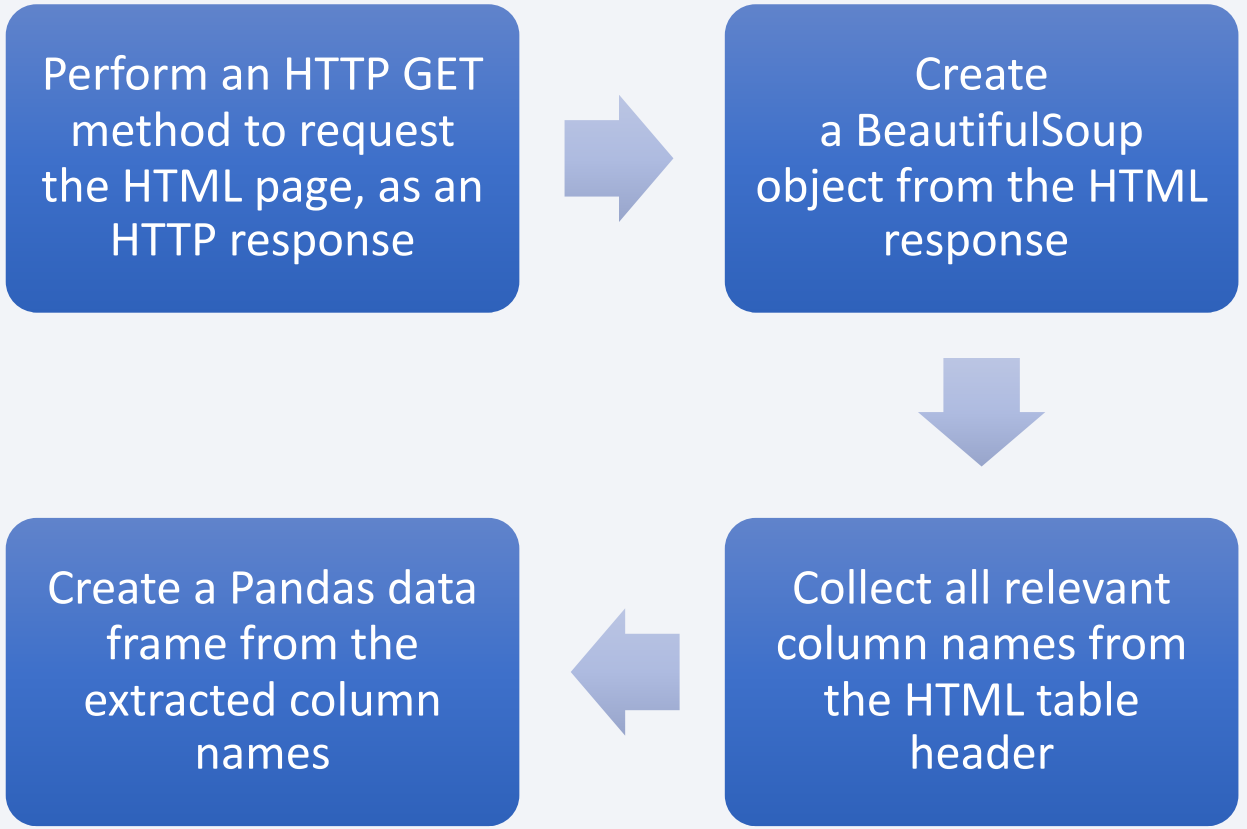
Data Collection – SpaceX API

- We used a get request to the SpaceX API to collect the data.
- Once we got the requested data, we decoded the response content as a JSON and turned it into a Pandas data frame
- We then performed some basic formatting and cleaning.
- Data Collection API notebook



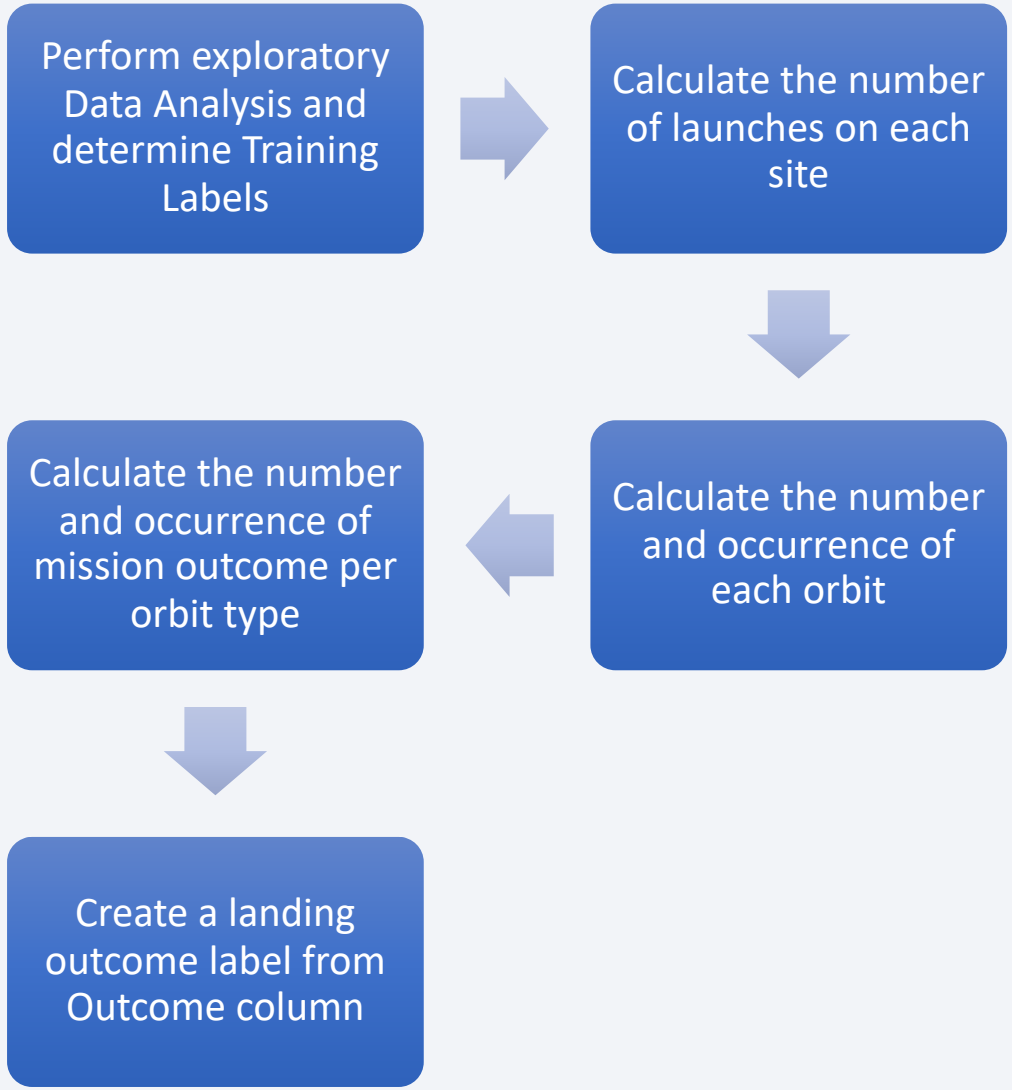
Data Collection - Scraping

- We used web scraping to extract a Falcon 9 launch records HTML table from Wikipedia
- After obtaining the data, we then parsed the table and converted it into a Pandas data frame
- Data Collection with Web Scraping



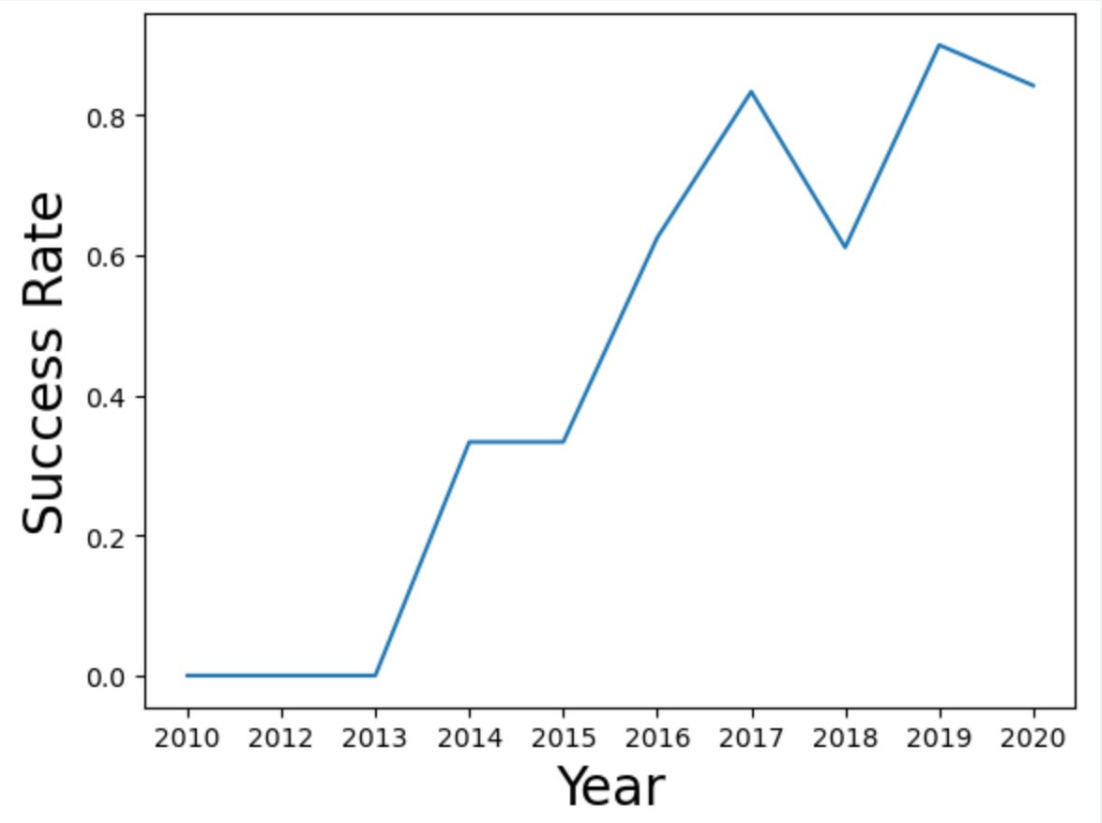
Data Wrangling

- We performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- We converted booster landings into Training Labels with “1” meaning the landing was successful and “0” meaning the landing was unsuccessful.
- Data Wrangling Notebook



EDA with Data Visualization

- We created scatter plots and bar charts to visualize the relationships between various columns such as Flight Number vs. Launch Site, Payload vs. Orbit Type, and Success Rate vs. Orbit Type
- We created a line chart to visualize the launch success yearly trend
- EDA with Data Visualizations Notebook



EDA with SQL

- Within the Jupyter notebook, we established a connection with the SQL database on IBM Cloud and executed SQL queries to perform an exploratory data analysis
- We wrote and executed SQL queries to find out for instance:
  - The names of the unique launch sites in the space mission
  - The total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - The total number of successful and failure mission outcomes
- EDA with SQL Notebook

Build an Interactive Map with Folium

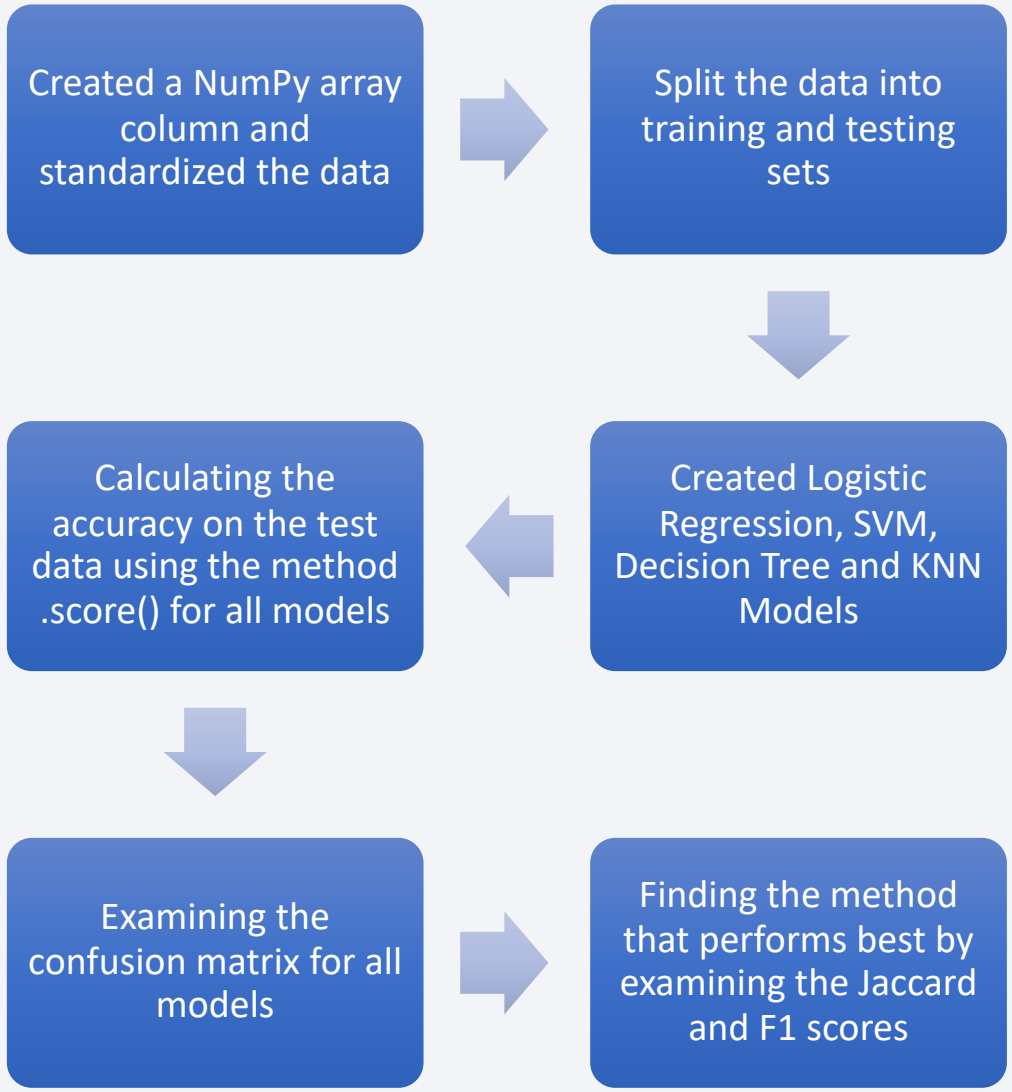
- Marked all Launch Sites on a Map
  - Created a folium Map object, with an initial center location to be NASA Johnson Space Center at Houston, Texas.
  - Created and added folium.Circle and folium.Marker for each launch site on the site map
- Marked the success/failed launches for each site on the map
  - Created markers for all launch records.
  - If a launch was successful (class=1), then we used a green marker and if a launch was failed, we used a red marker (class=0)
- Calculated the distances between a launch site to its proximities
  - Marked down a point on the closest coastline using MousePosition and calculated the distance between the coastline point and the launch site
  - Drew a PolyLine between a launch site to the selected coastline point
- Interactive Visual Analytics with Folium Notebook

Build a Dashboard with Plotly Dash

- Built a Plotly Dash application for users to perform interactive visual analytics on SpaceX launch data in real-time.
- This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.
- Added a callback function is to render a pie chart visualizing launch success counts based on selected site dropdown
- Added a Range Slider to Select Payload and find if variable payload is correlated to mission outcome.
- SpaceX App Python

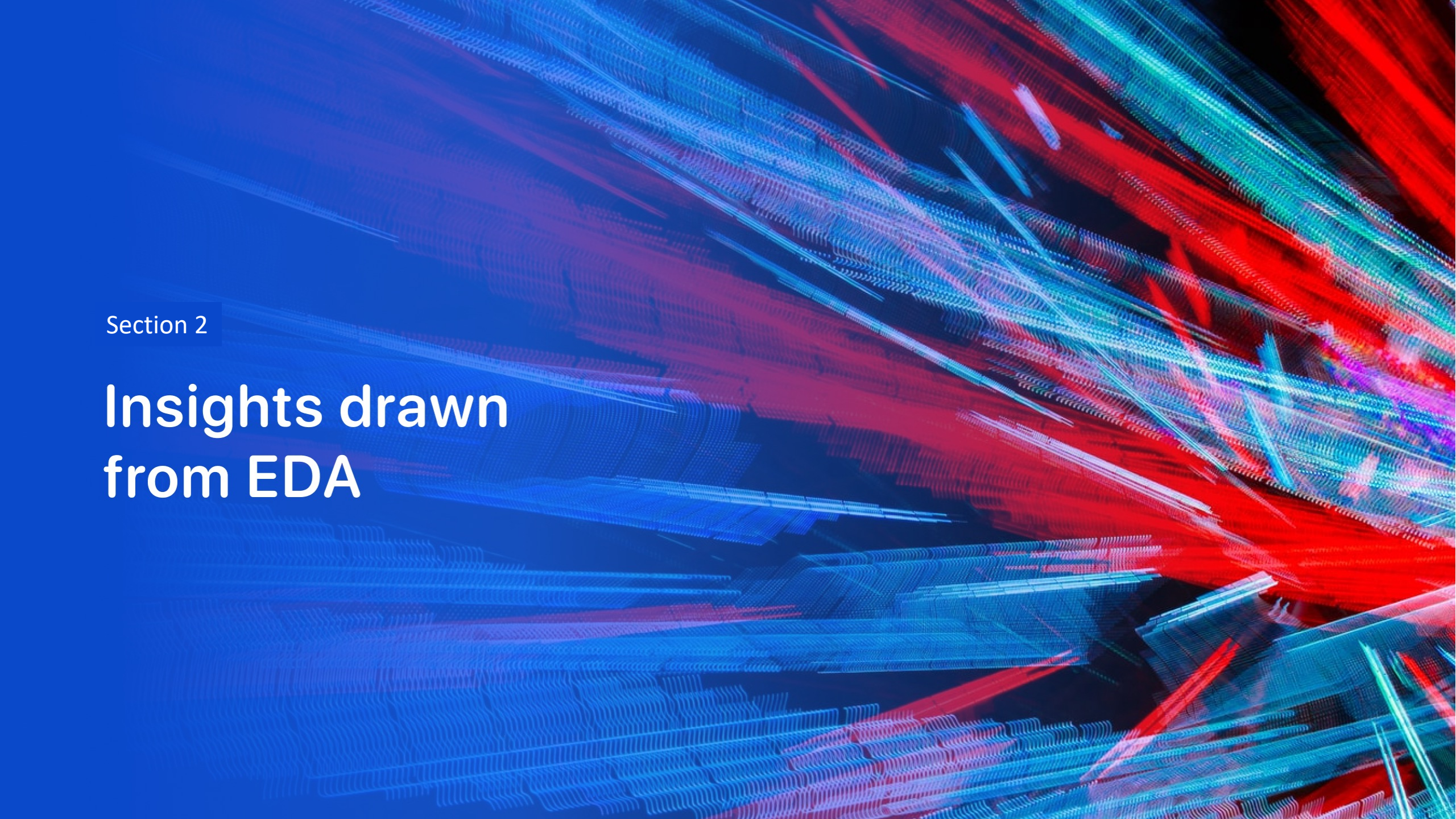
Predictive Analysis (Classification)

- We loaded the data using NumPy and Pandas, transformed the data, and split it into training and testing datasets.
- We built different machine learning models using GridSearchCV:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K-Nearest Neighbors
- Calculated the accuracy of each model and examined each with confusion matrices
- SpaceX Machine Learning Prediction Notebook

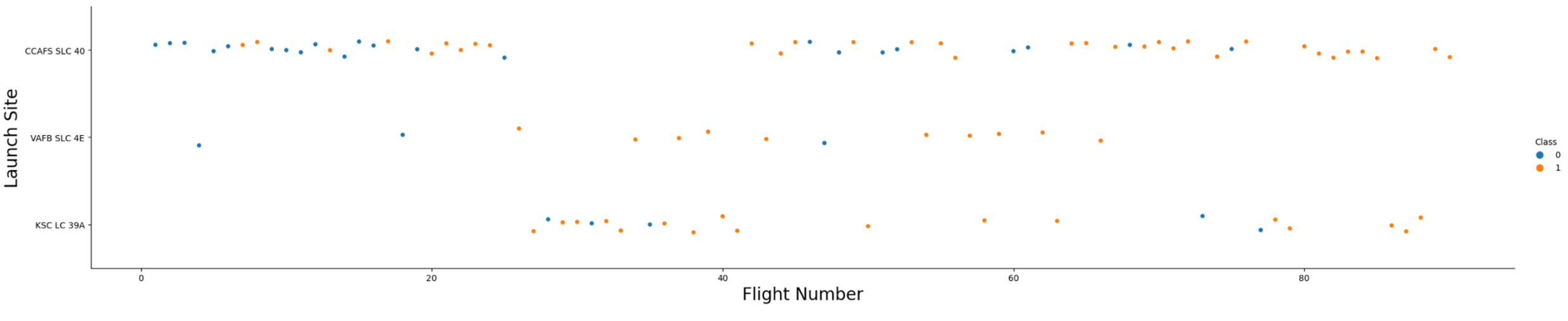


Results

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

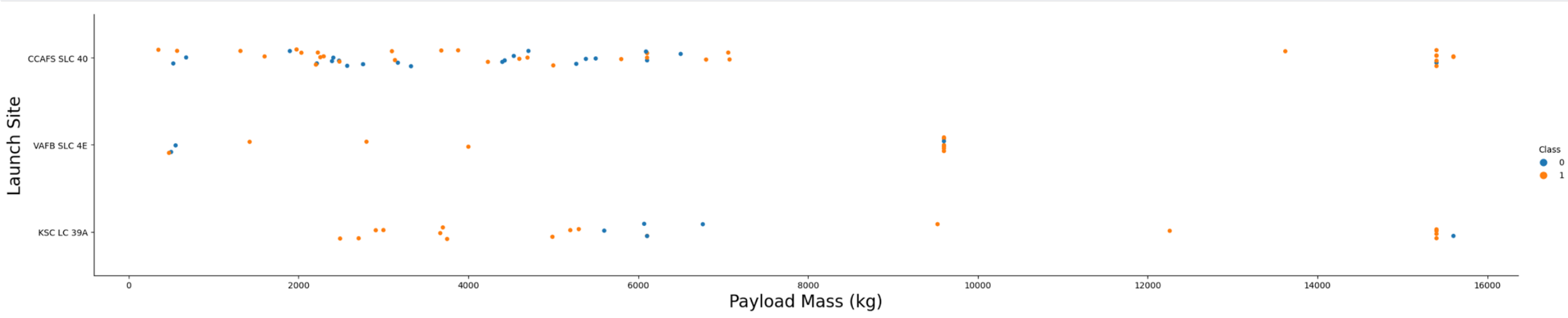

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.

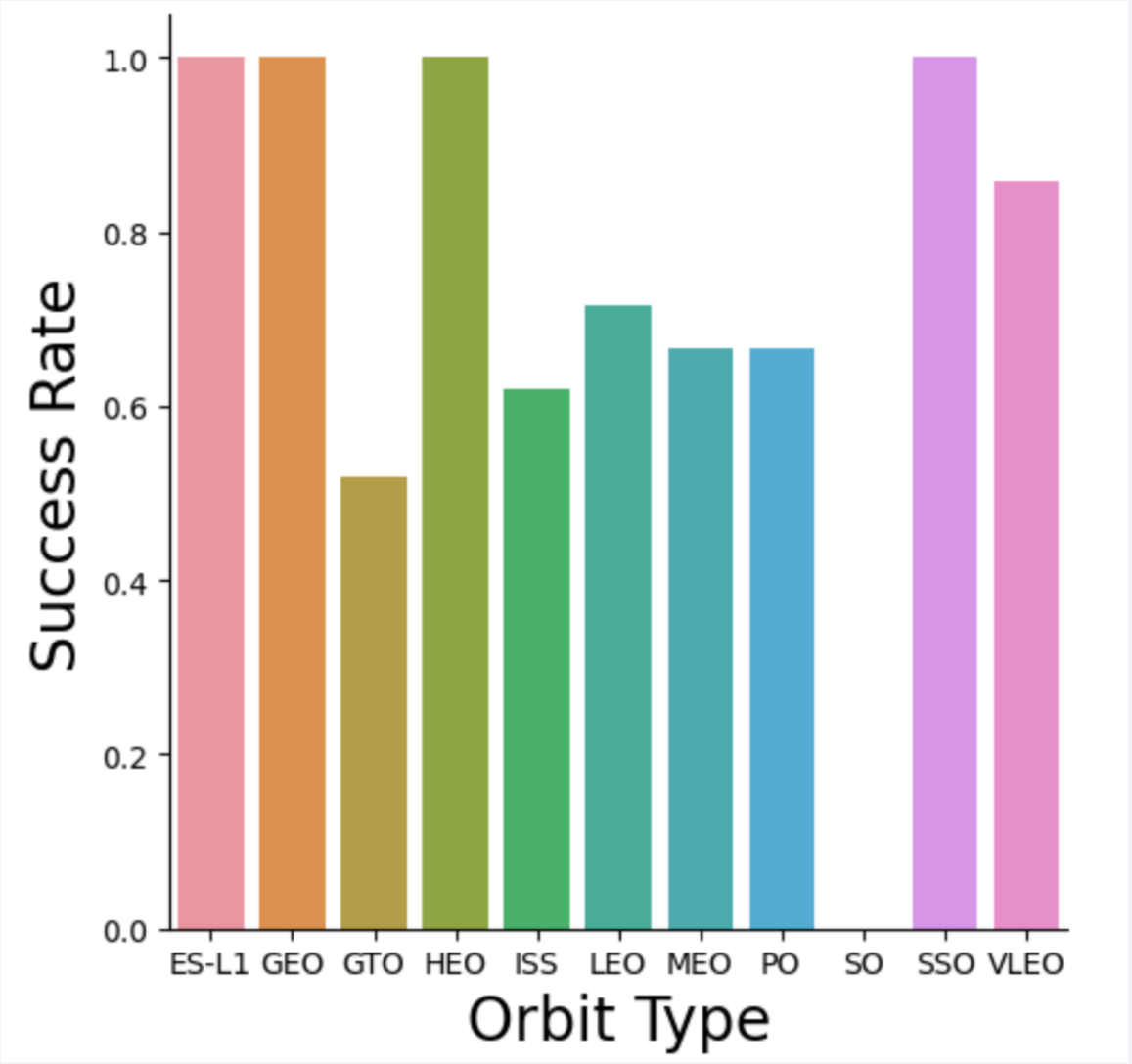
Payload vs. Launch Site



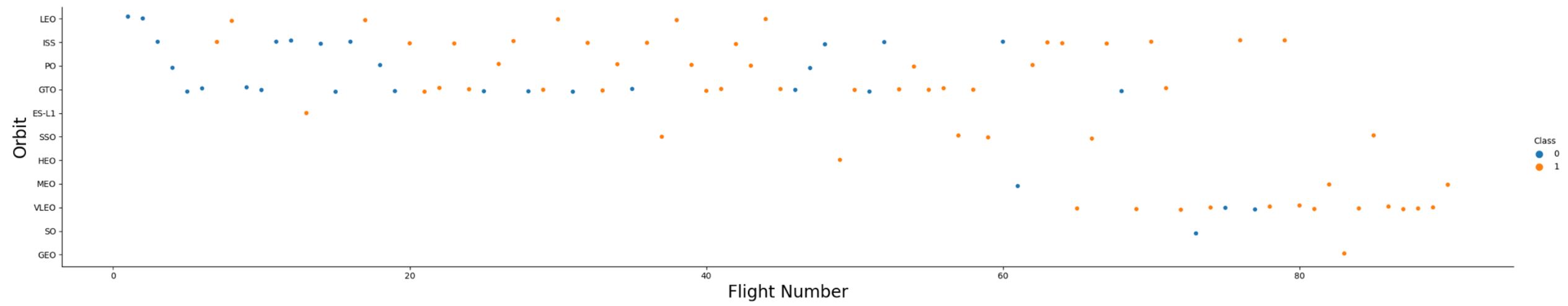
- For every launch site the higher the payload mass, the higher the success rate.
- The scatter plot displays that for the VAFB-SLC launch site, there are no rockets launched for heavy payload mass (greater than 10000).

Success Rate vs. Orbit Type

- Orbits with 100% success rate:
  - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate:
  - SO

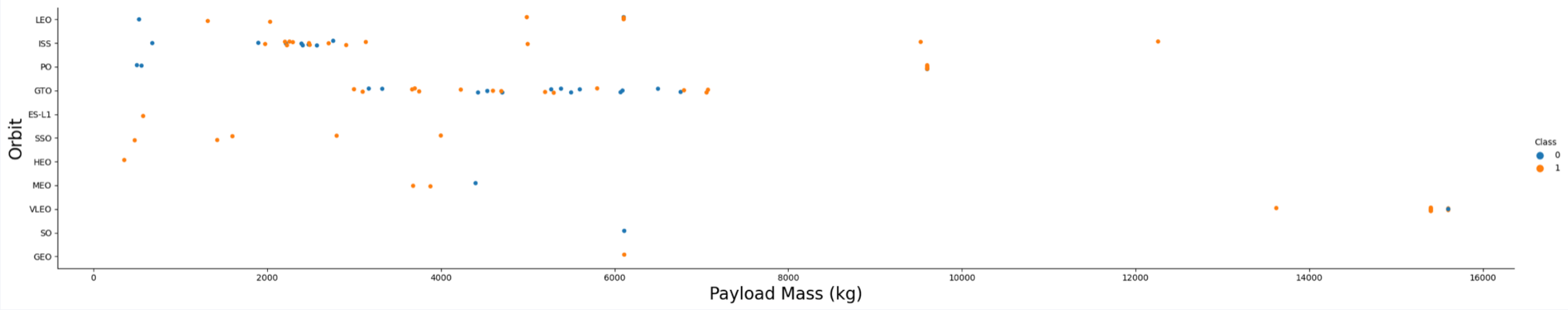


Flight Number vs. Orbit Type



- In the LEO orbit, success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

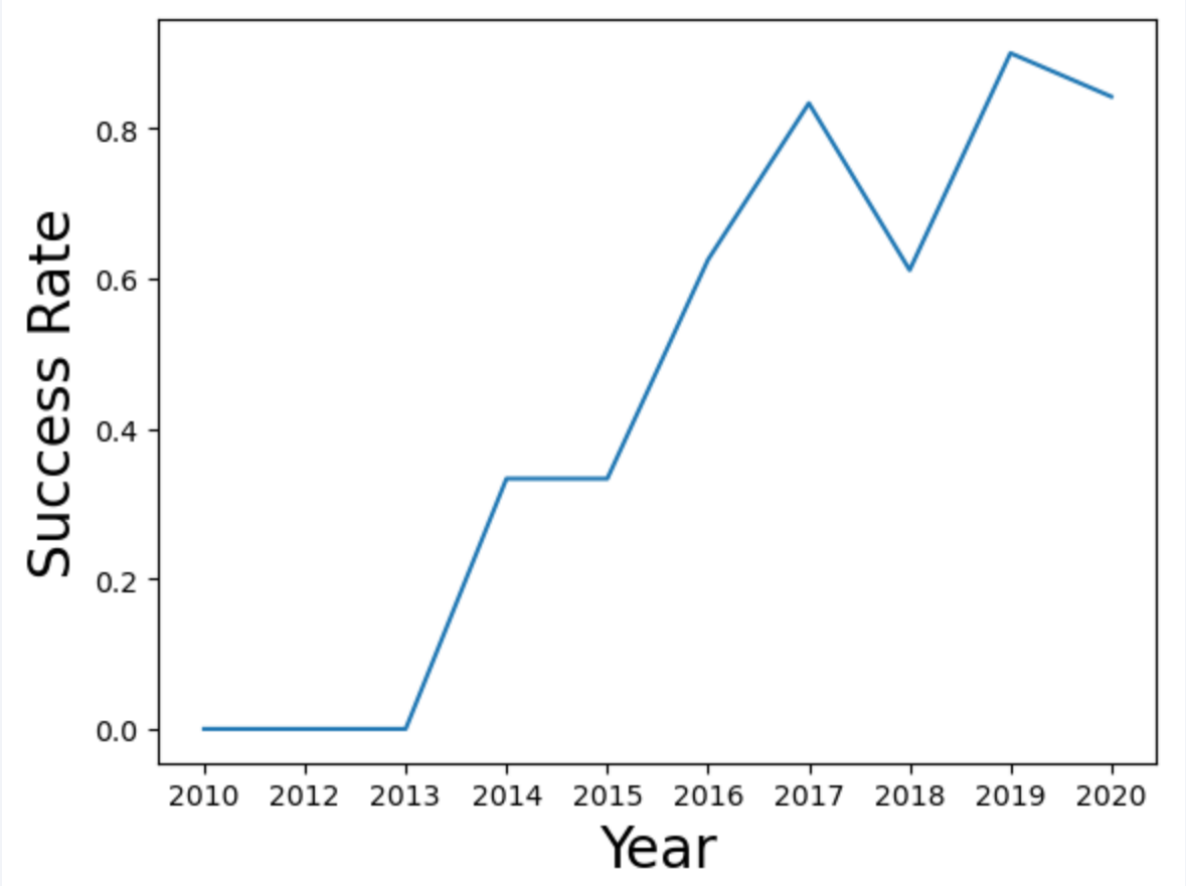
Payload vs. Orbit Type



- With heavy payloads, the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there.

Launch Success Yearly Trend

- Success rate kept increasing until 2020



All Launch Site Names

In [12]:
```
%sql SELECT DISTINCT launch_site FROM SPACEX;
```

```
 * ibm_db_sa://hpl83722:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[12]:

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

- Displays the names of the unique launch sites in the space mission

Launch Site Names Begin with 'CCA'

In [14]:
```
%sql SELECT * FROM SPACEX WHERE launch_site LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://hpl83722:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[14]:

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

- Displays 5 records where launch sites begin with the string 'CCA'

Total Payload Mass

In [16]:
```
%sql SELECT SUM(payload_mass__kg_) AS total_payload_mass FROM SPACEX WHERE customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://hpl83722:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[16]:

total_payload_mass
45596

- The total payload mass carried by boosters launched by NASA (CRS) is 45,596 kg

Average Payload Mass by F9 v1.1

```
In [18]: %sql SELECT AVG(payload_mass__kg_) AS average_payload_mass FROM SPACEX WHERE booster_version like 'F9 v1.1%';
 * ibm_db_sa://hpl83722:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud
Done.
Out[18]: average_payload_mass
                        2534
```

- Average payload mass carried by booster version F9 v1.1 is 2534 kg.

First Successful Ground Landing Date

```
In [32]: %sql SELECT MIN(DATE) AS first_success FROM SPACEX WHERE Landing_Outcome = 'Success (ground pad)';

 * ibm_db_sa://hpl83722:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appd
Done.
Out[32]: first_success
         2015-12-22
```

- The dates of the first successful landing outcome on ground pad is Dec. 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

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

In [34]:
```
%sql SELECT booster_version from SPACEX WHERE Landing_Outcome = 'Success (drone ship)' and payload_mass__kg_ BETWEEN 4000 AND 6000;
```

```
 * ibm_db_sa://hpl83722:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[34]:

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

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

Total Number of Successful and Failure Mission Outcomes

In [30]:
```
%sql SELECT mission_outcome, COUNT(*) as number from SPACEX GROUP BY mission_outcome;
```

```
 * ibm_db_sa://hpl83722:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.c
Done.
```

Out[30]:

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

- Lists the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

In [37]:
```
%sql SELECT booster_version FROM SPACEX WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEX);
```

```
 * ibm_db_sa://hpl83722:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud
Done.
```

Out[37]:

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

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

2015 Launch Records

In [40]:
```
%sql SELECT Landing_Outcome, booster_version, launch_site FROM SPACEX WHERE Landing_Outcome = 'Failure (drone ship)' AND YEAR(DATE) = 201
```

```
 * ibm_db_sa://hpl83722:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[40]:

landing_outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Lists the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

```
%sql SELECT Landing_Outcome, COUNT(*) AS total FROM SPACEX WHERE date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDEI
```

```
 * ibm_db_sa://hpl83722:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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

- Ranks 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

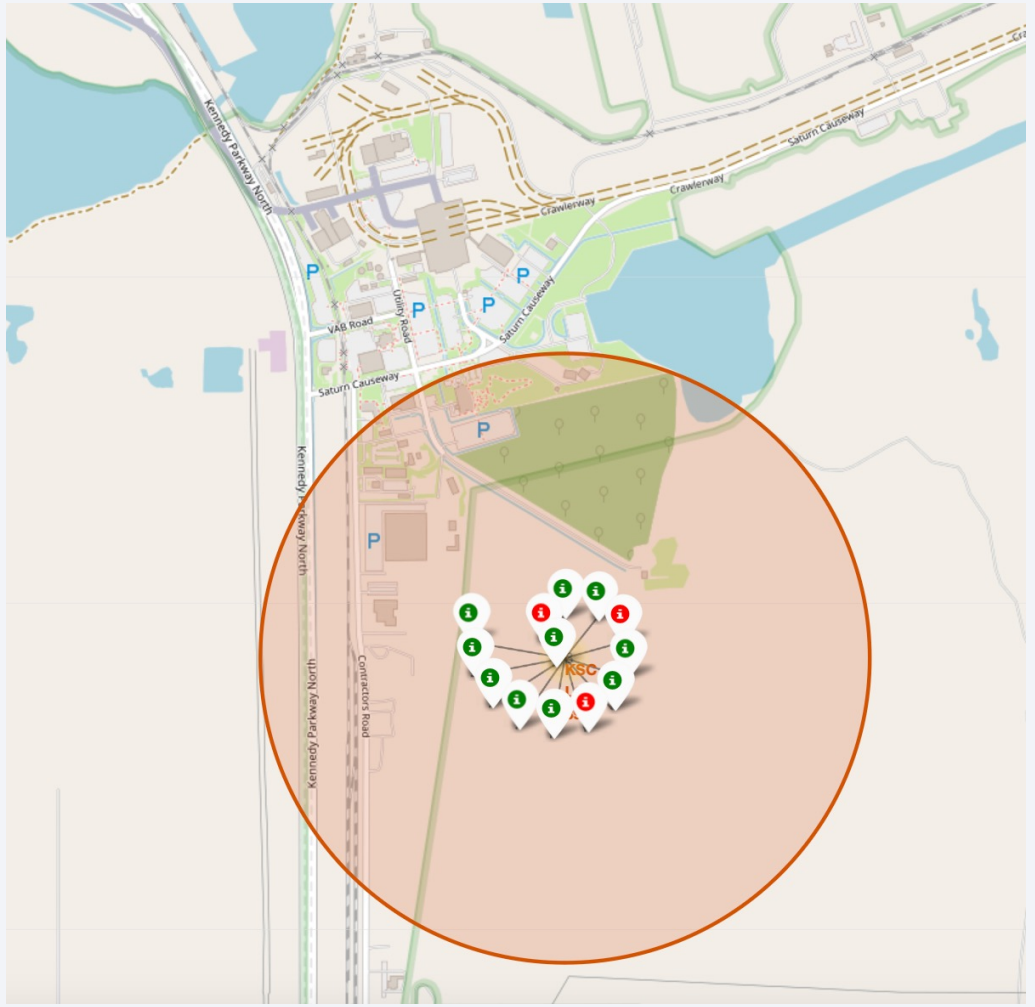

Section 3
Launch Sites
Proximities Analysis

All Launch Sites Global Map Markers

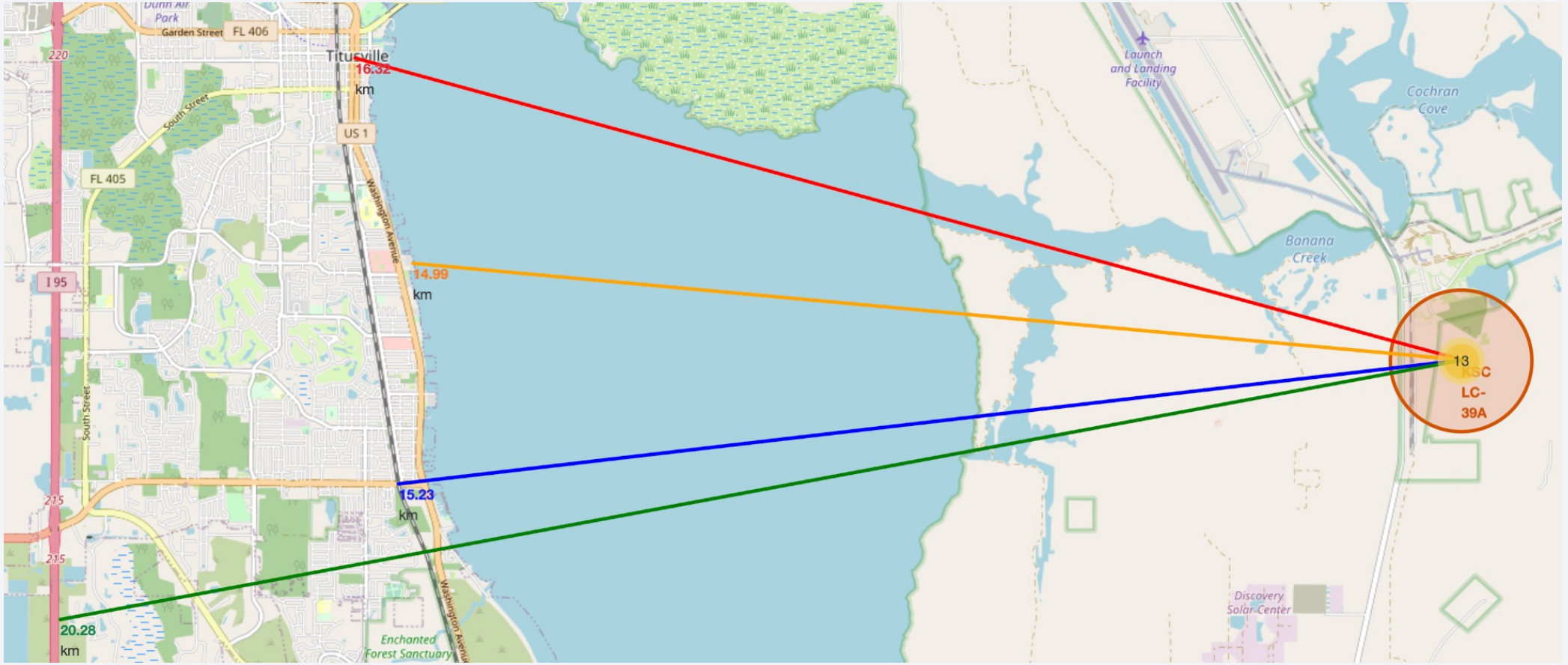
- All launch sites are in proximity to the Equator line
- All launch sites are in very close proximity to the coast

Color-Labeled Launch Records

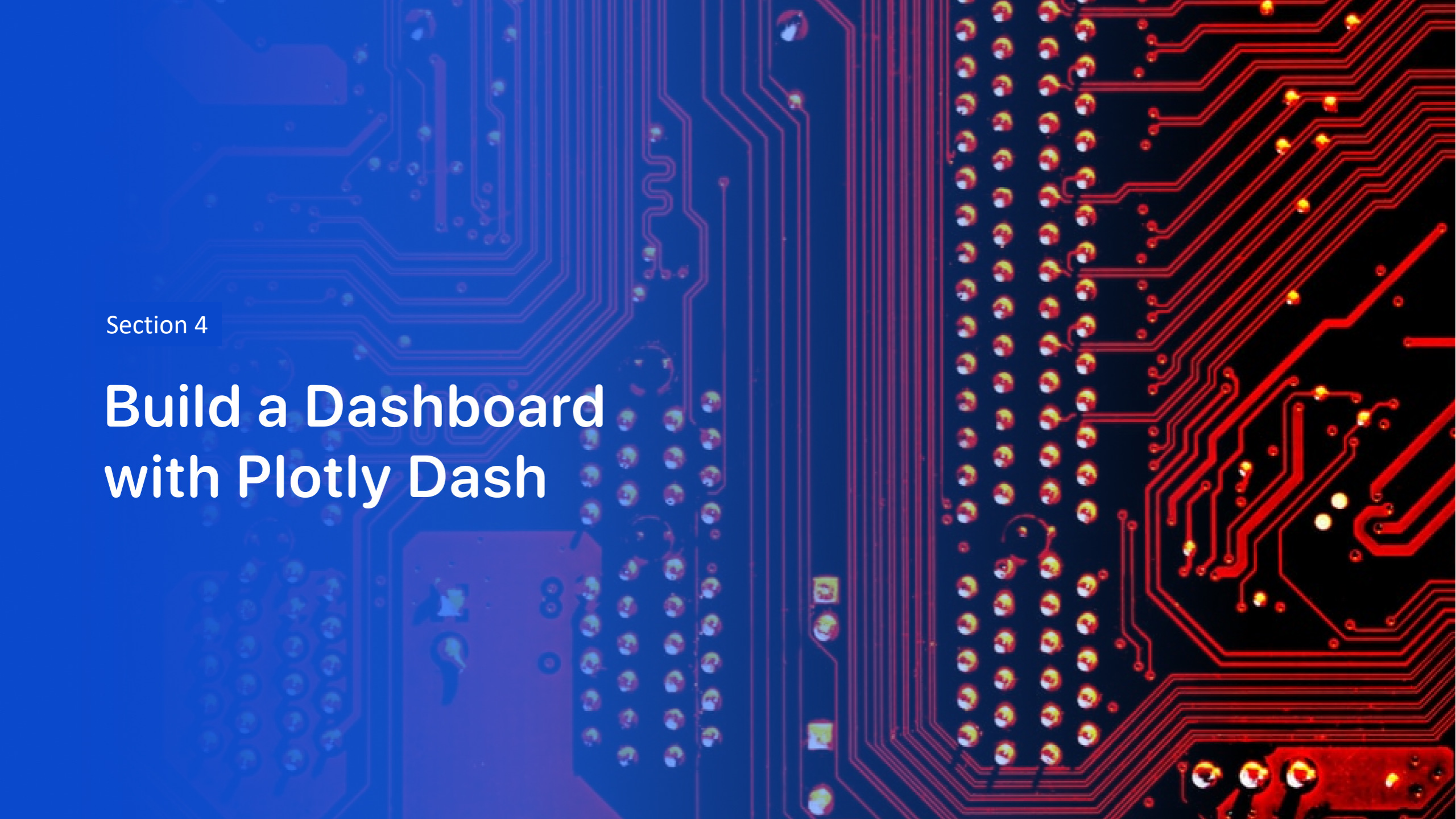
- From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate.



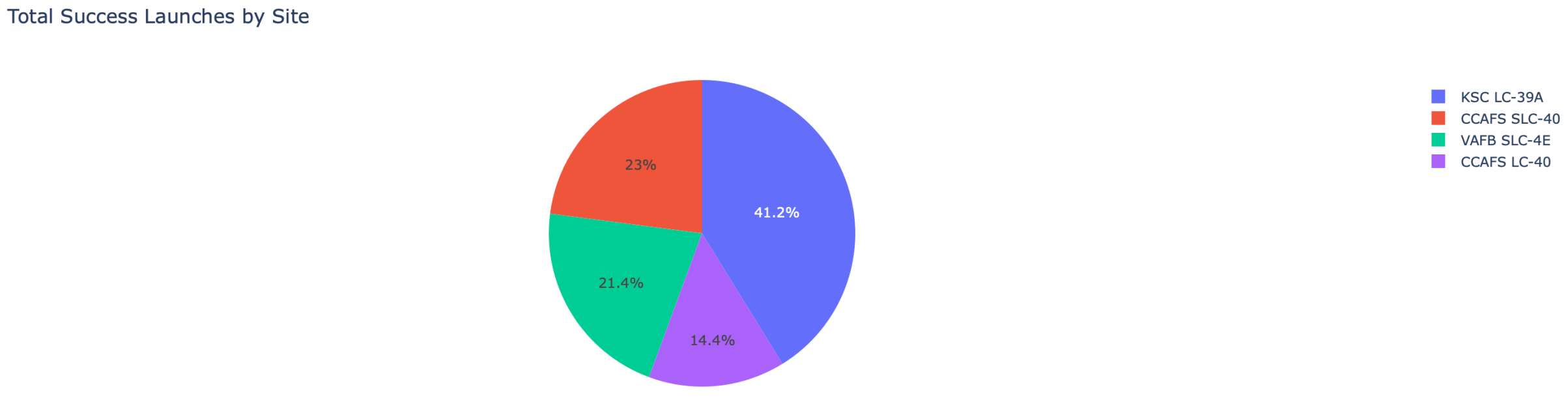
Distance Between Launch Site KSC LC-39A to its Proximities



- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
  - Relatively close to railway (15.23 km)
  - Relatively close to highway (20.28 km)
  - Relatively close to coastline (14.99 km)

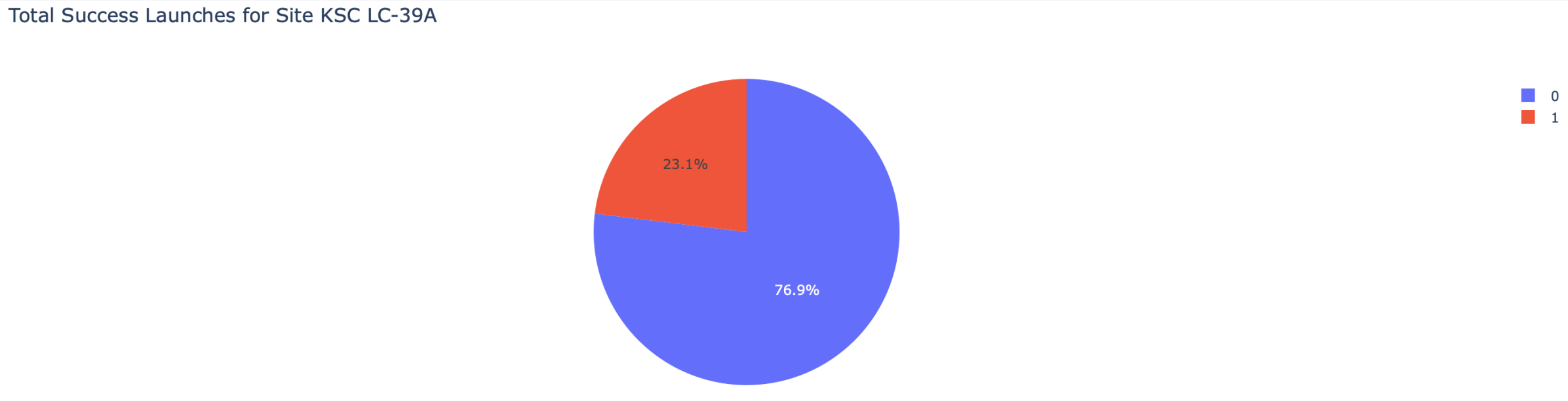

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by Site



- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

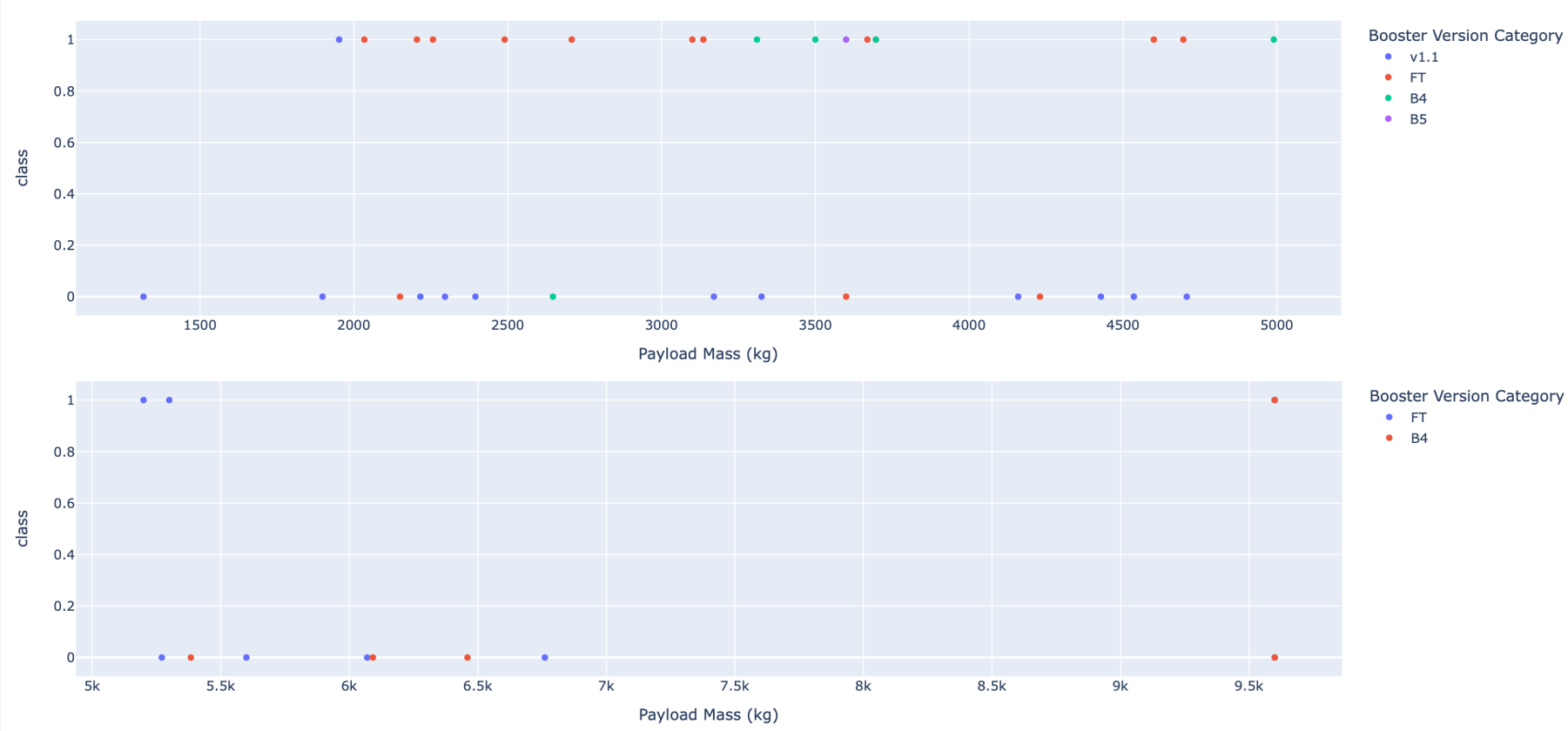
Launch Site with Highest Success Ratio



- Launch Site KSC LC-39A has the highest success ratio with 76.9% of their launches being successes

Payload vs Launch Outcome

- Payloads between 2000 and 5500 kg have the highest success rate



Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Based on the scores of the Test Set, we can not confirm which method performs best.
- The scores of the whole Dataset confirm that the best model is the Support Vector Machine. This model has not only higher scores, but also the highest accuracy.

Test Dataset:

Out[35]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.888889	0.833333

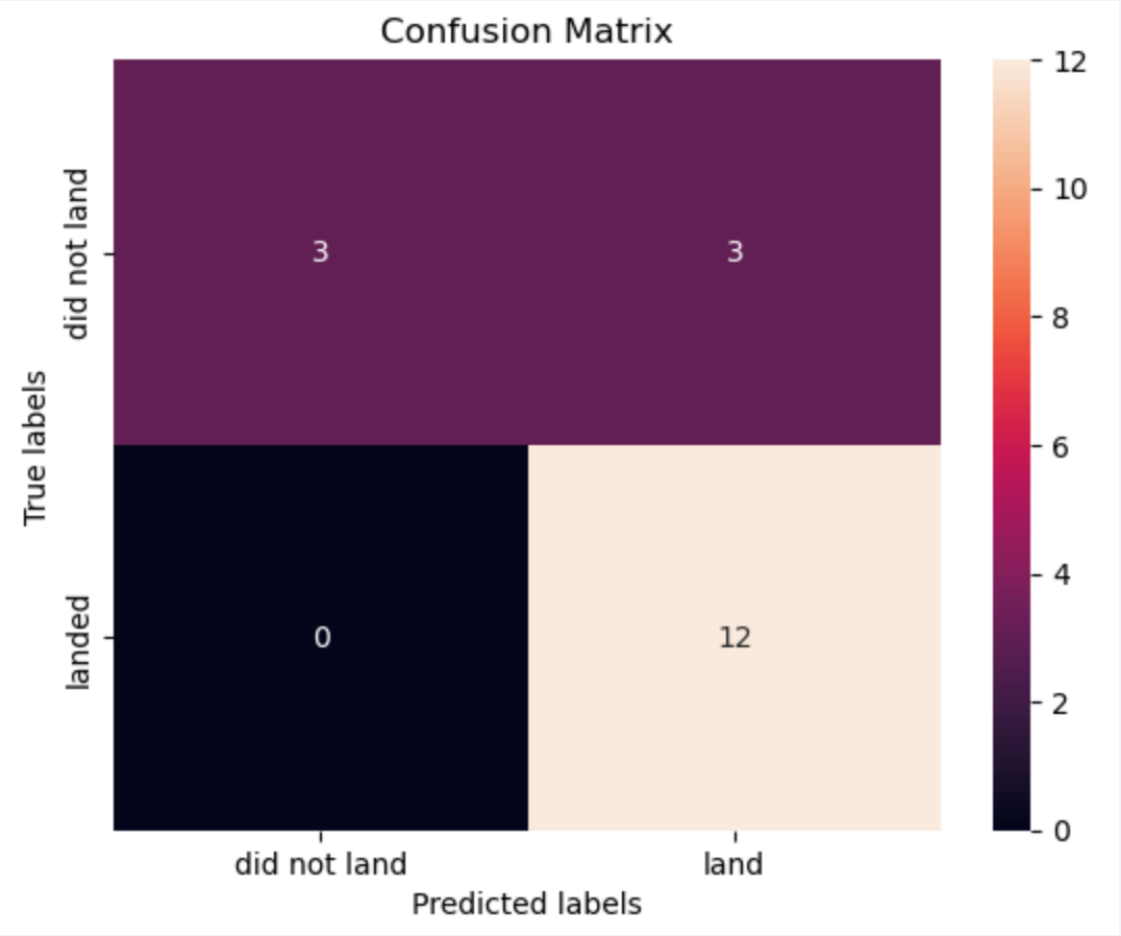
Whole Dataset:

Out[36]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.764706	0.819444
F1_Score	0.909091	0.916031	0.866667	0.900763
Accuracy	0.866667	0.877778	0.822222	0.855556

Confusion Matrix

- Examining the confusion matrix, we see that SVM can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- For every launch site the higher the payload mass, the higher the success rate. However, there is a specific recommended range for more successful launches (2000-5500 kg).
- The success rate of launches has increased over the years.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate
- Out of all the launch sites, KSC LC-39A had the most successful launches of any sites.
- SVM is the best machine learning algorithm for this dataset.

Thank you!