IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Miguel Angel Quiñones Ramirez
4/22/2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using SpaceX API and Web Scrapping
  - Data Wrangling using Pandas
  - Exploratory Data Analysis (EDA) using SQL and with visualization
  - Interactive Visual Analytics with Folium and Plotly Dash
  - Predictive Analysis using Sklearn
- Summary of all results
  - Collection and Wrangling of data from a public source
  - The data was further analyzed with visual representations
  - EDA was used to identify the best features to predict the success of launchings
  - A Machine Learning (ML) model with the best features was developed to predict the success

Introduction

- Project background and context
  - One reason of SpaceX´s success is because of their relatively inexpensive rocket flights that reuse the first stage of the rocket
  - SpaceY would like to compete with SpaceX. For that reason, the data of past launches will be analyzed and used to train a model that predicts whether the first stage will land successfully and will be reused.
- Problems you want to find answers
  - How to predict if the first stage of the rocket will land successfully and can be reused to save costs

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API https://api.spacexdata.com/v4/rockets/
  - Web Scraping from Wikipedia https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling
  - A landing outcome label was added to each launch based on their results
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was normalized, divided in training and test datasets, and evaluated in four different classification models. Later, hyperparameter tuning was performed to find the best

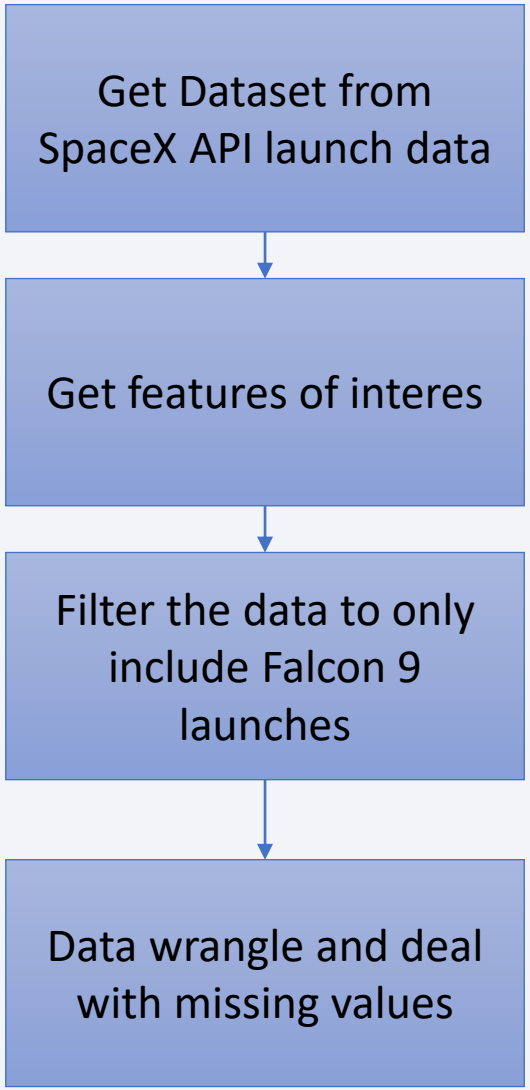
Data Collection

- The first dataset, containing information about the date, site and outcome of the launches before and during 2020, was collected using request calls to the SpaceX API (https://api.spacexdata.com/v4/rockets/) as well as
- Further data was obtained by web scraping the Wikipedia page of Falcon launches in 2020 (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)

Data Collection – SpaceX API

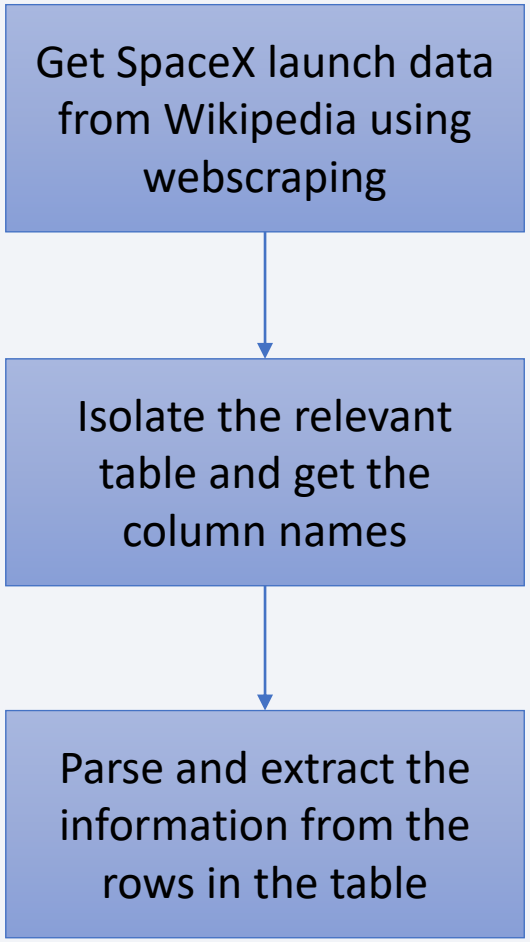
- The dataset was collected using the requests library and stored in a pandas dataframe
- Features of interest were stored in a dictionary
- The data was filtered to include only Falcon 9 launches.
- Missing values from PayloadMass were dealt with

- Source: https://github.com/MiguelQr/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

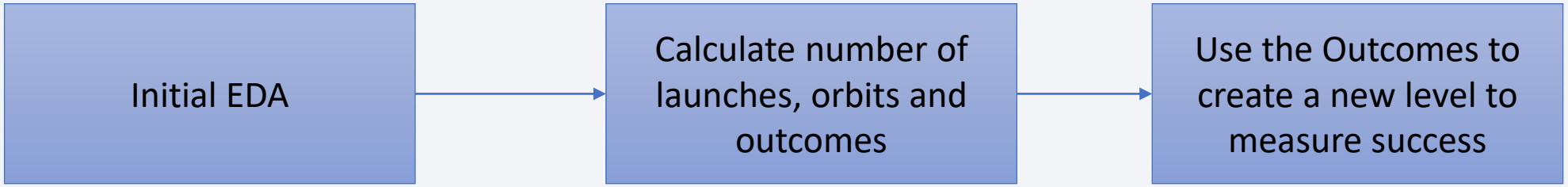
- The Wikipedia webpage html data was collected using the beautifulsoup library and stored in a pandas dataframe
- The table with the dataset is isolated and the columns are extracted
- The rows of the table are parsed and put in a dataframe



- Source: https://github.com/MiguelQr/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

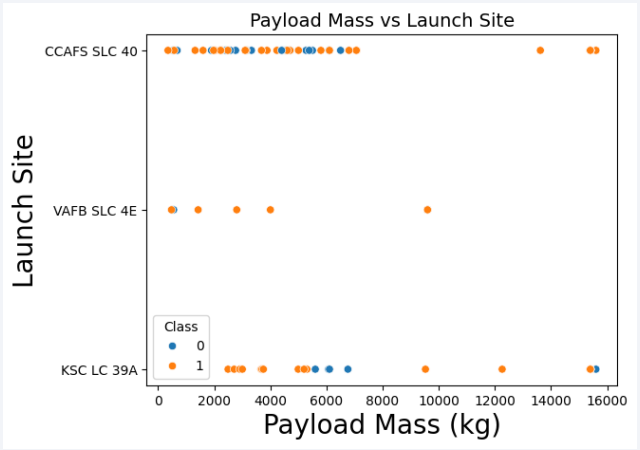
- EDA was performed to know the dataset better
- The number of launches on each site, orbits and landing outcomes in those orbits were calculated
- A landing outcome label corresponding to the Outcome of the launch was added to the dataframe, which can be further used to determine the success rate

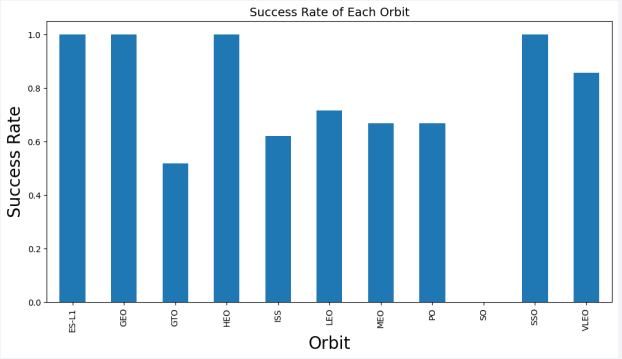


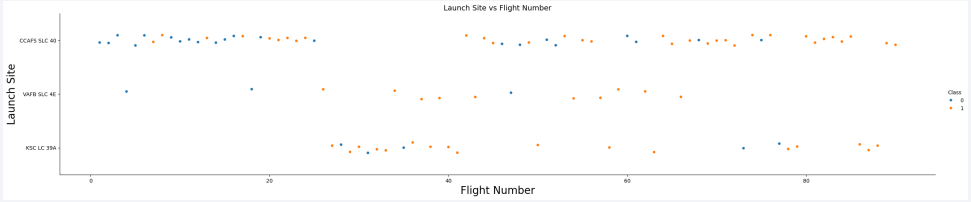
- Source: https://github.com/MiguelQr/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

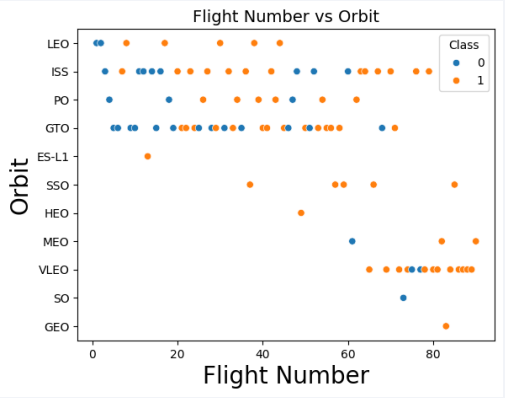
EDA with Data Visualization

- Scatterplots and barplots were used to visualized the relationship between two variables, some of them include:
  - Flight number and launch site
  - Payload and launch site
  - Success rate and orbit
  - Flight number and orbit









- Source: https://github.com/MiguelQr/Applied-Data-Science-Capstone/blob/main/edadataviz.ipynb

EDA with SQL

- The following SQL queries were performed:
  - Unique launch sites
  - Launch sites that begin with CCA
  - Payload mass carried by NASA (CSR) boosters
  - Average payload mass carried by F9 v1.1 boosters
  - Date of the first successful landing
  - Boosters successful in drone ship with payload mass between 4000 and 6000
  - Successful and failure mission outcomes
  - Booster versions which have carried the maximum payload mass
  - Month, booster version and launch site of failure landing outcomes in 2015
  - Count of landing outcomes between 2010-06-04 and 2017-03-20
- Source: https://github.com/MiguelQr/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

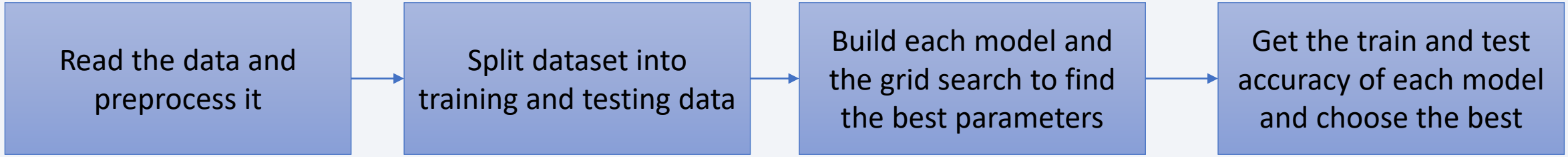
- Different markers, circles and lines were added to a folium map:
  - A marker and a circle for each launch site and their respective area
  - A marker for each success and failure within the areas
  - A marker between a launch site and a coastline
  - A polyline between the selected launch site and coastline
  - A polyline between a different selected launch site and a nearby city
- Source: https://github.com/MiguelQr/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- A Pie Chart was added to show the percentage of success launches by site
  - The site can be selected in a dropdown menu, and selecting the option ‘All’ allows to compare successful launches between sties.
- A Scatter Plot was added to show the correlation between payload and launch success
  - A slider limits the range of the payload that is shown in the plot
- Source: https://github.com/MiguelQr/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

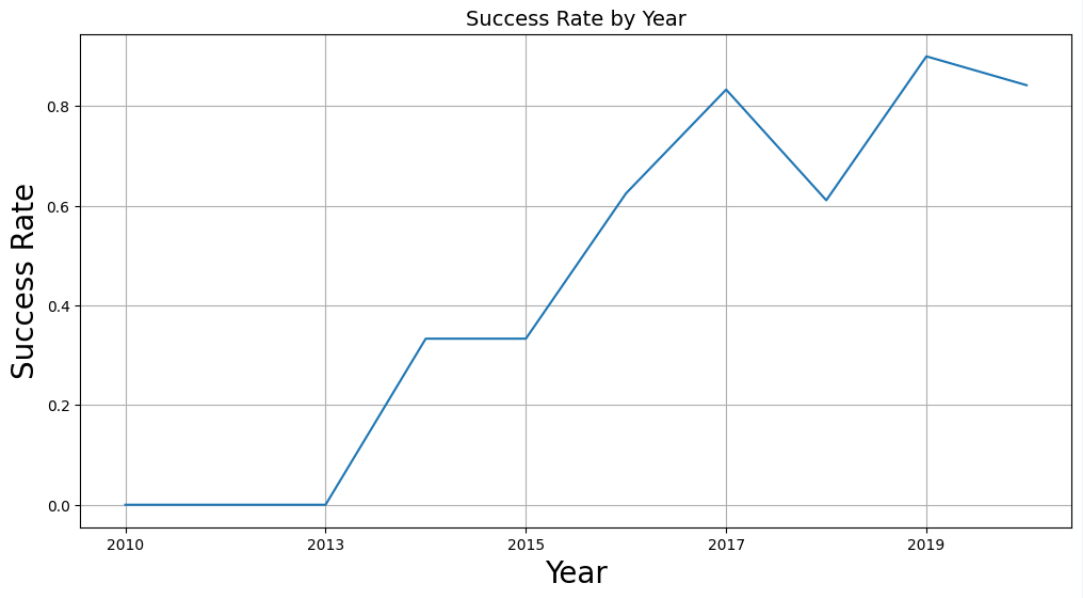
- Four classification models were built and evaluated: Logistic Regression, SVM, Decision Tree and KNN:
- For each one a Grid Search was used for hyperparameter tuning, and the accuracy and test accuracy was measured with the best parameters found.
- The best performing model was the one with the best accuracy.



- Source: https://github.com/MiguelQr/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5_blank.ipynb

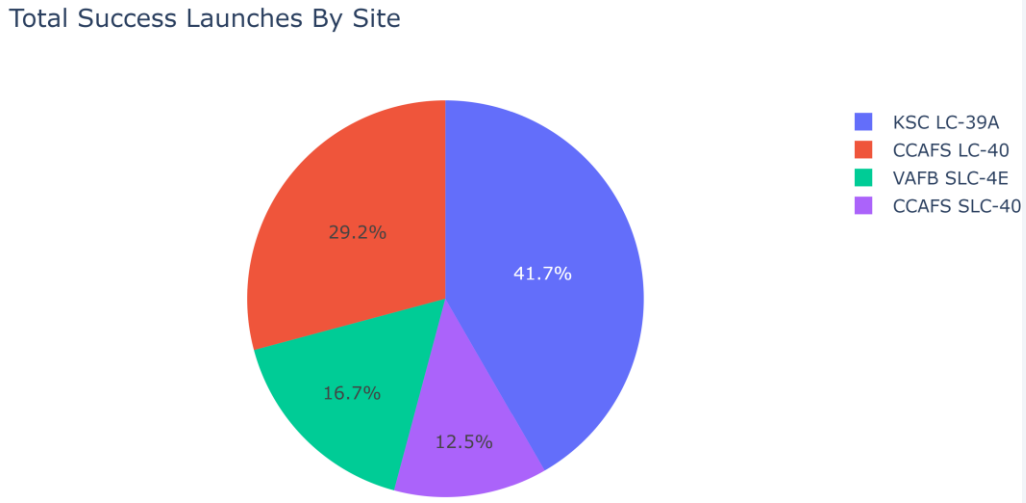
Results

- Exploratory data analysis results
  - The are four different launch sites
  - The first successful landing date was in 2015-12-22
  - 98% of the missions were successful
  - Success rate increased by the year



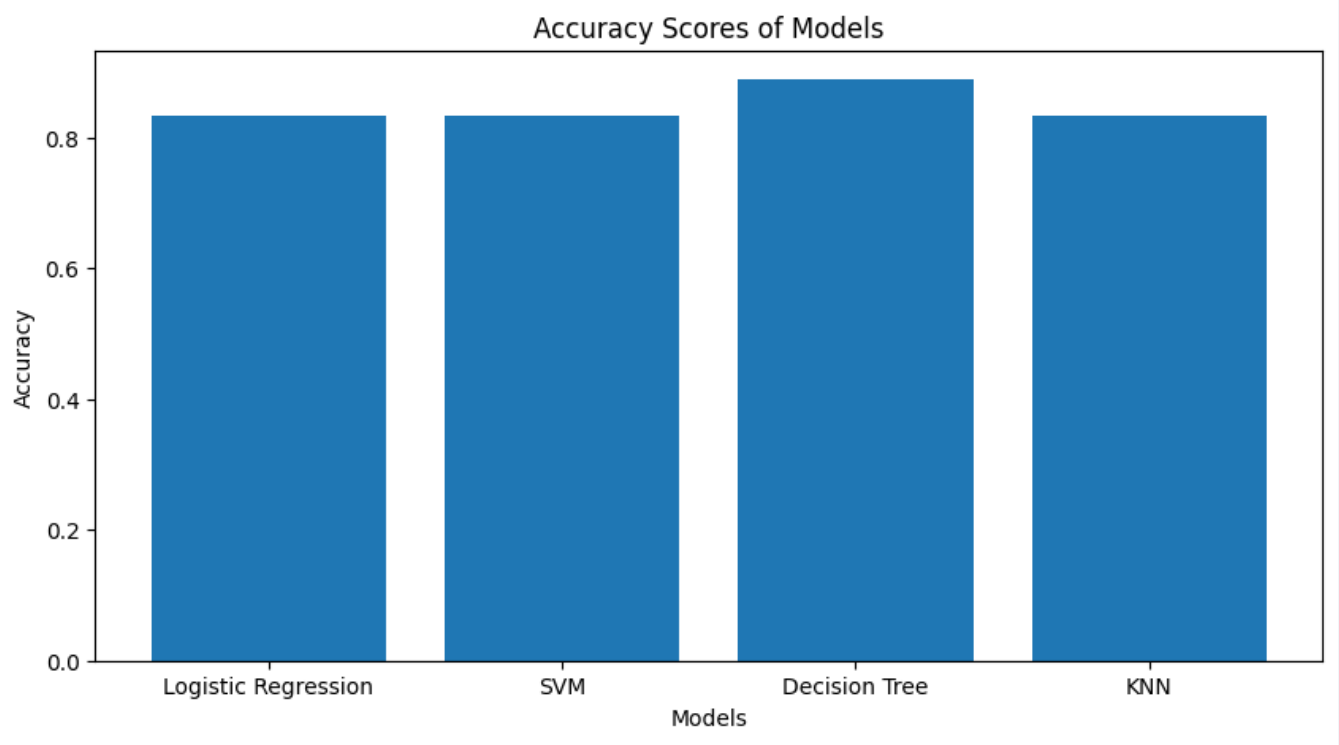
Results

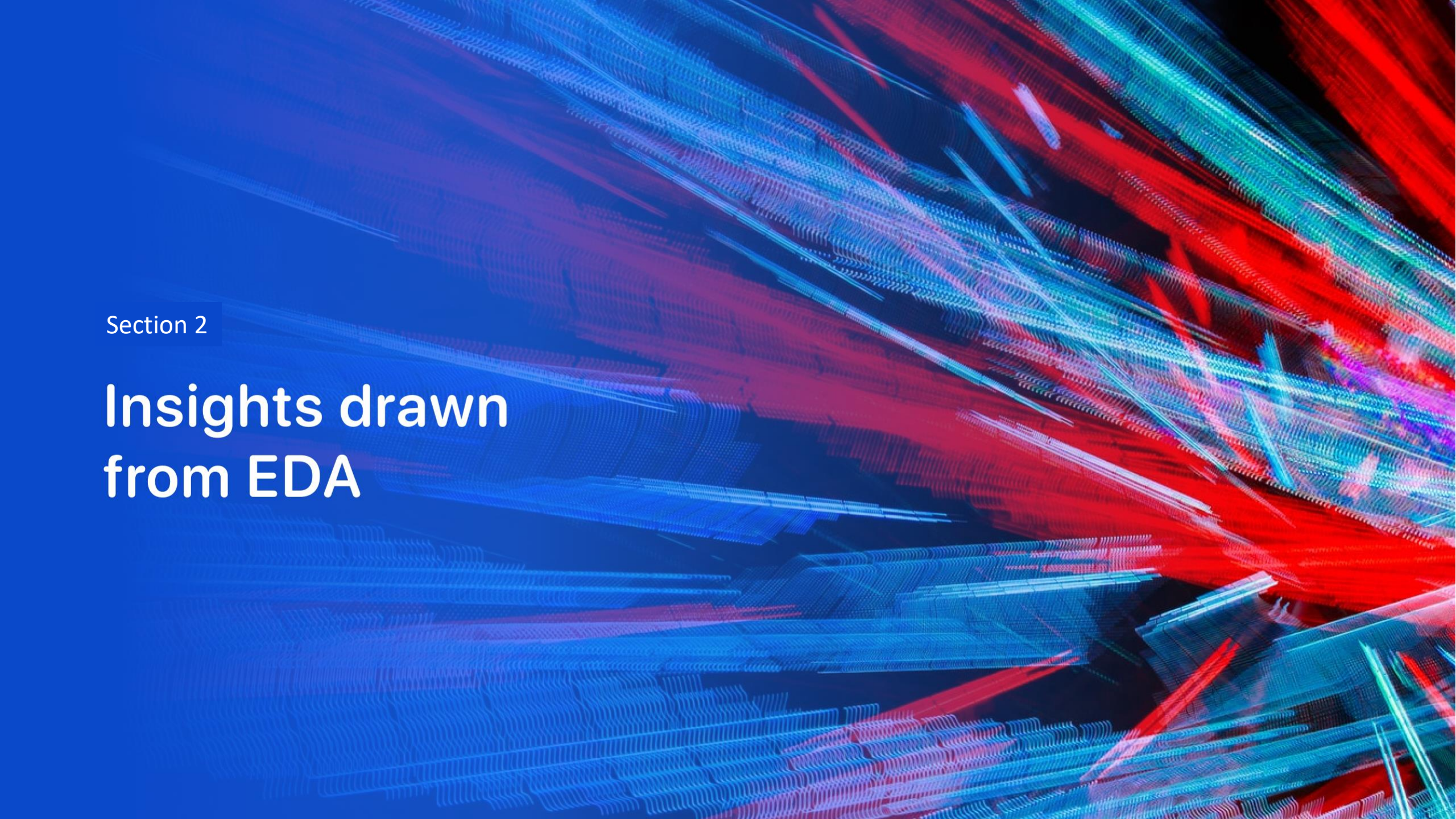
- Interactive visual analysis results
  - Out of the four launch sites, KSC LC-39A had the most successful launches
  - The four launch sites were near the coast



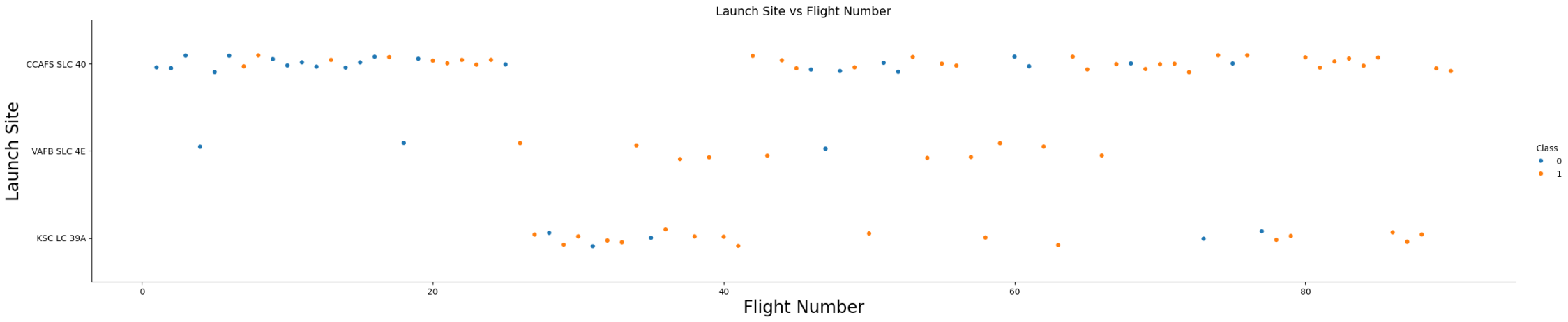
Results

- Predictive analysis results
  - Out of the four models, the Decision Tree, with its best hyperparameters, had the highest accuracy of 88%



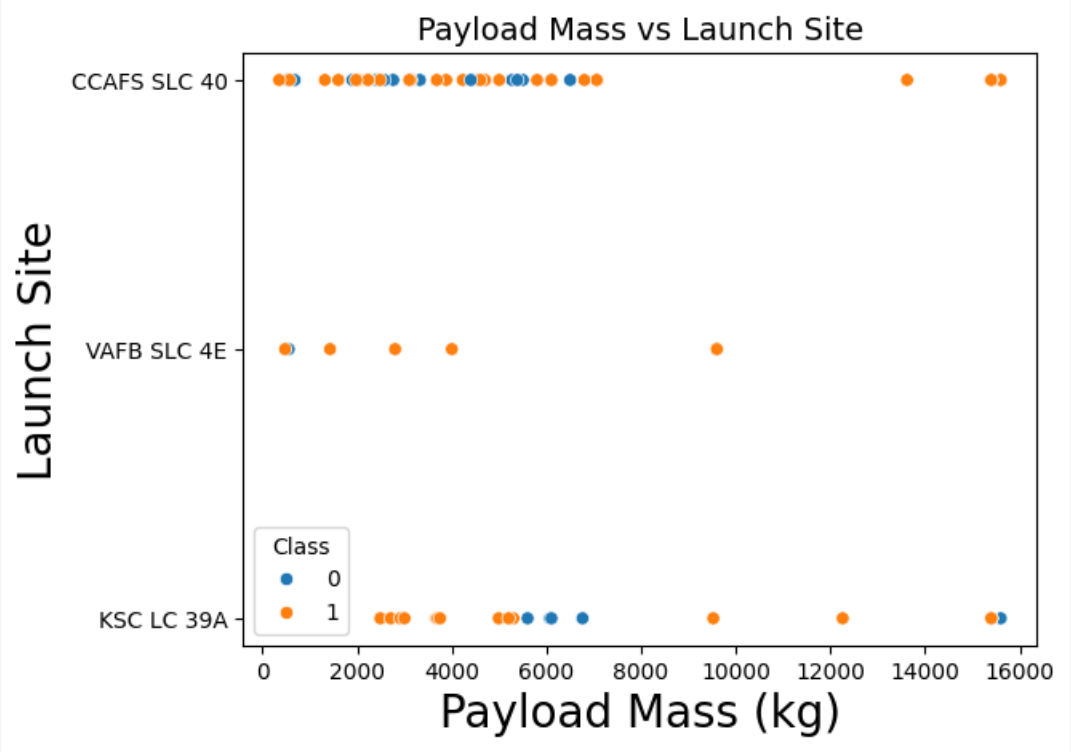

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



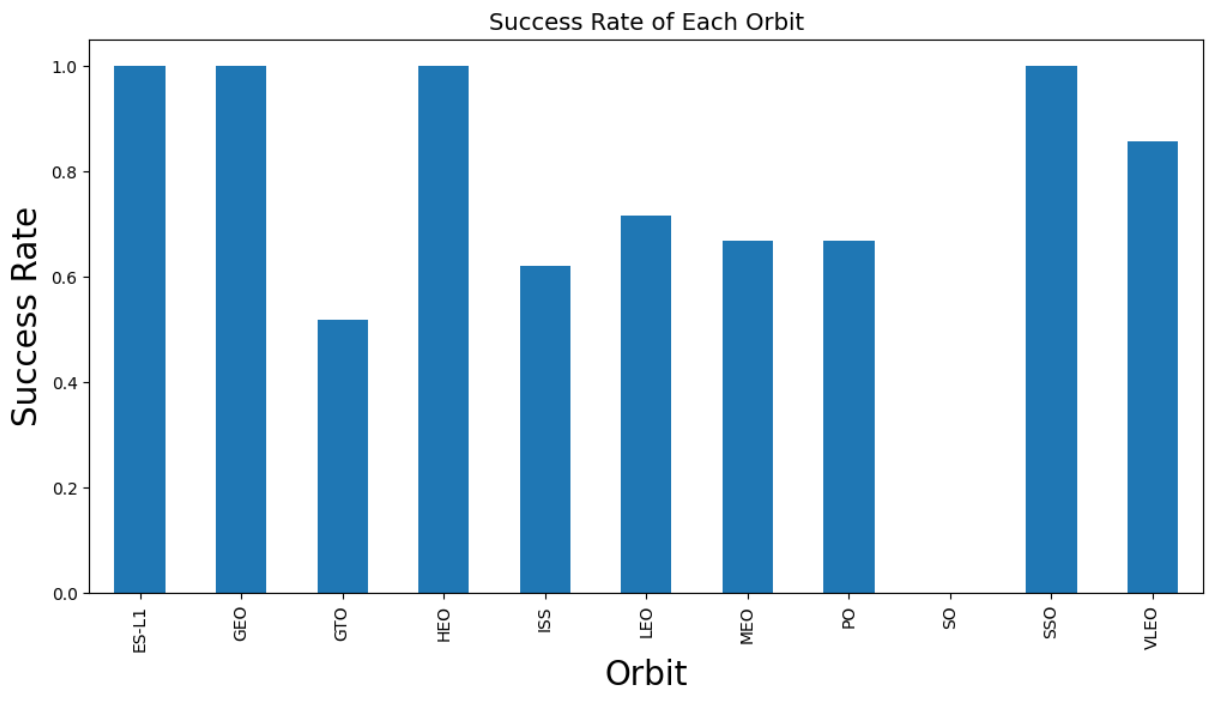
- The best launch site is CCAPS SLC 40, with both the most amount of launches and successes
- In general, the success rate of the launches has improved over time

Payload vs. Launch Site



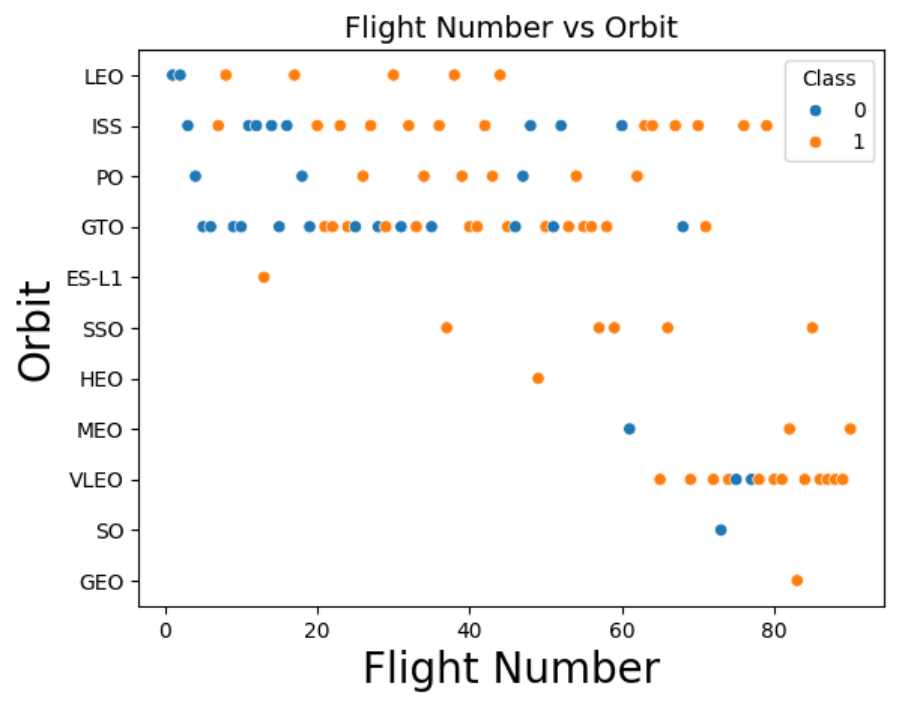
- Only in CCAPS SLC 40 and KSC LC 39A there are payloads above 10,000 kg
- Payloads above 8,000 kg seem to have a better success rate

Success Rate vs. Orbit Type



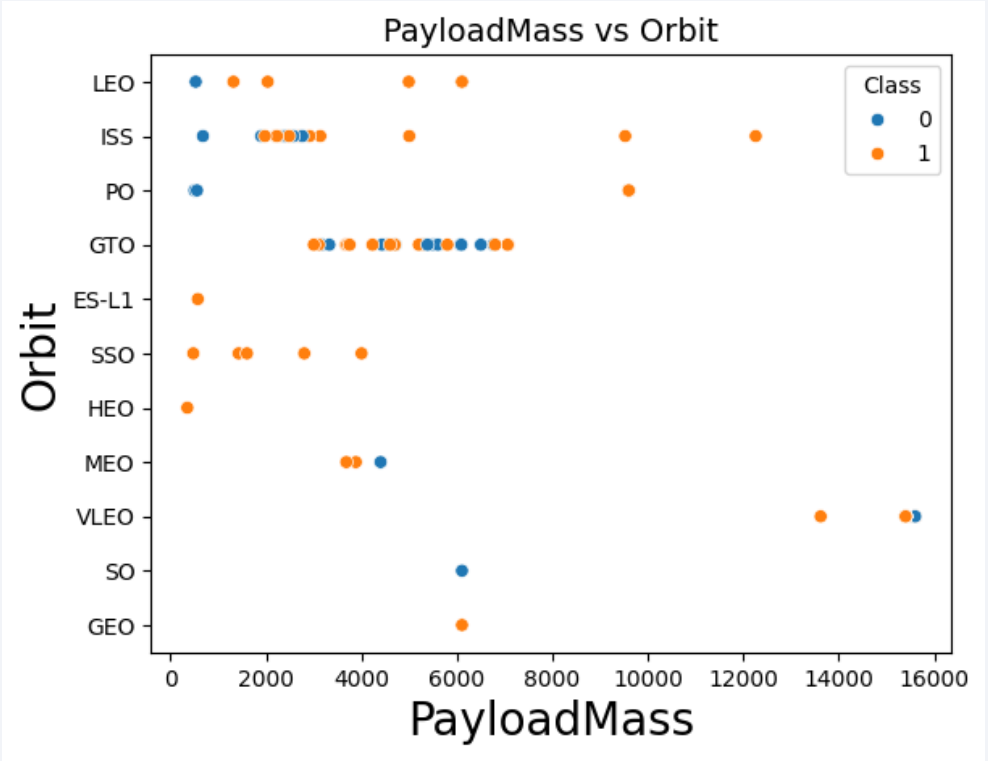
- The orbits with perfect success rate are ES-L1, GEO, HEO and SSO
- All orbits seem to have above 50% success rate

Flight Number vs. Orbit Type



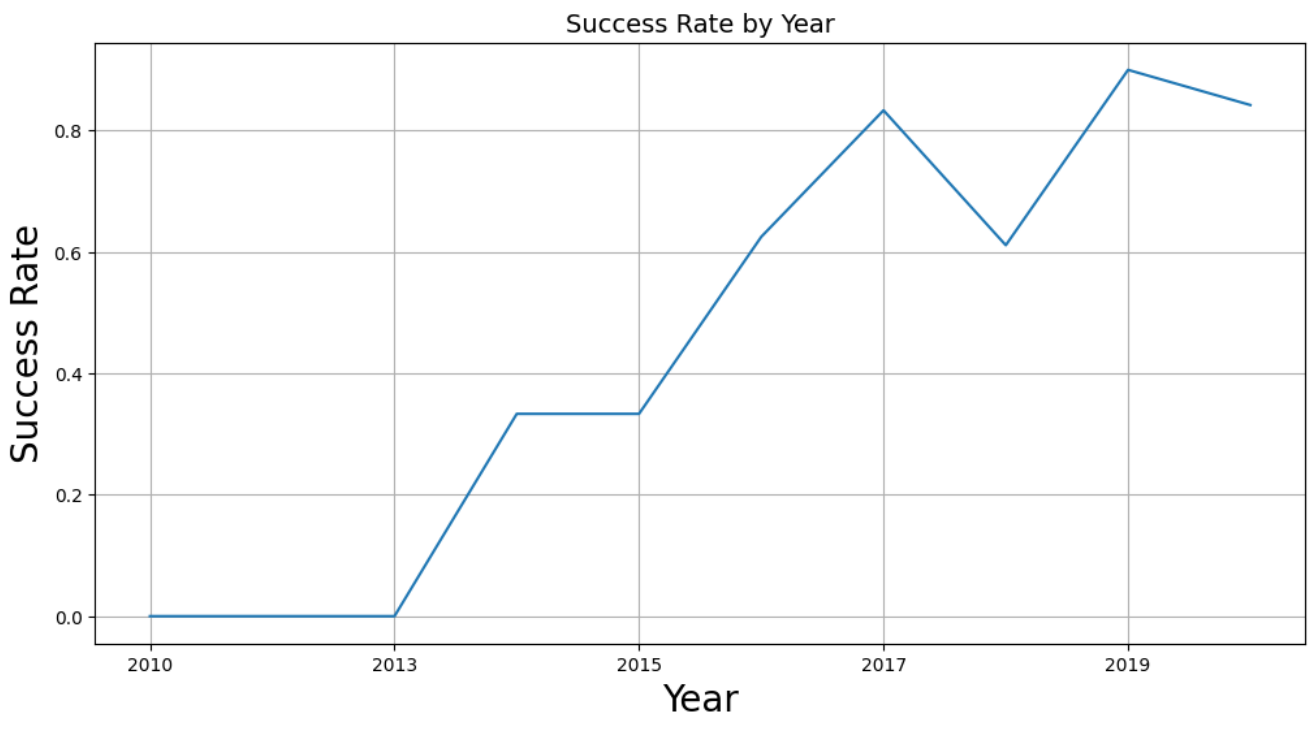
- The VLEO orbit is one of the most recent ones to start, and has a very high success rate
- SSO and HEO have perfect success rates, but very few flights
- Overall, the success rate of all orbits has improved over time

Payload vs. Orbit Type



- ISS has a wide range of payload, while also having a good success rate
- SSO has a perfect success rate, but with small payloads
- Overall, there does not seem to be much correlation

Launch Success Yearly Trend



- Between 2010 and 2013 there was a period with no success rate increase, probably because of development and adjustments
- Starting on 2013, the success rate has constantly increased until 2020. With a small exception in 2018.

All Launch Site Names

- Four different launch site names can be found with a query

```
%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'

- Five records of Cape Carneval launches can be found with a query

```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

Python

```
 * 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

- The total payload mass from boosters launched from NASA (CRS) is 48,213

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

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

Total_Payload_Mass
48213

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2,948.4

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

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

Average_Payload_Mass
2928.4

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad is 2015-12-22

```
%sql SELECT MIN(`Date`) AS First_Successful_Landing_Date FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
```

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

First_Successful_Landing_Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

```
ion 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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes are
  - Failure (in flight) 1
  - Success 98
  - Success 1
  - Success (payload status unclear) 1

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

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

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

Boosters Carried Maximum Payload

- The boosters which have carried the maximum payload mass are the following

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

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

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

2015 Launch Records

- The failed landing_outcomes in drone ship for the year 2015, alongside their booster versions, and launch site names, are:

```
%%sql

SELECT substr(Date, 6, 2) AS Month,
       Mission_Outcome,
       Booster_Version,
       Launch_Site
FROM SPACEXTABLE
WHERE substr(Date, 0, 5) = '2015'
  AND Landing_Outcome LIKE '%Failure (drone ship)%';
```

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

Month	Mission_Outcome	Booster_Version	Launch_Site
01	Success	F9 v1.1 B1012	CCAFS LC-40
04	Success	F9 v1.1 B1015	CCAFS LC-40

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

- The ranking of the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order, is as follows:

```
%%sql

SELECT Landing_Outcome,
       COUNT(*) AS Outcome_Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Outcome_Count DESC;
```

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

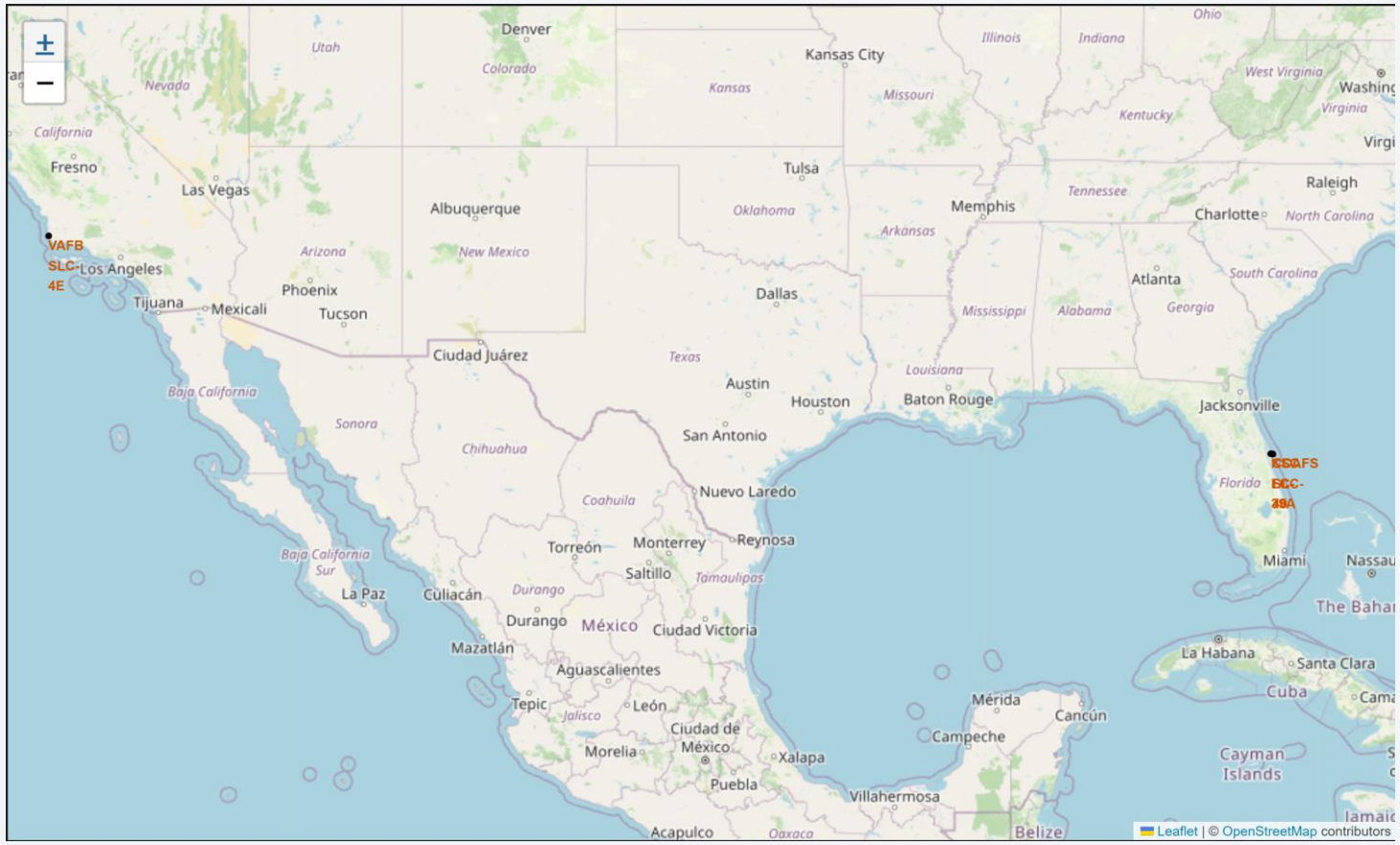
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

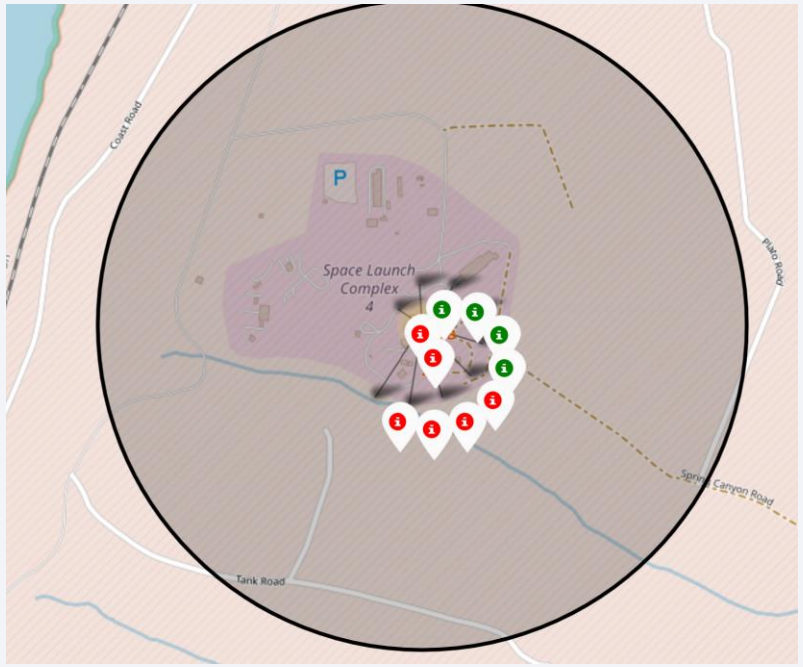
All Launch Sites Location

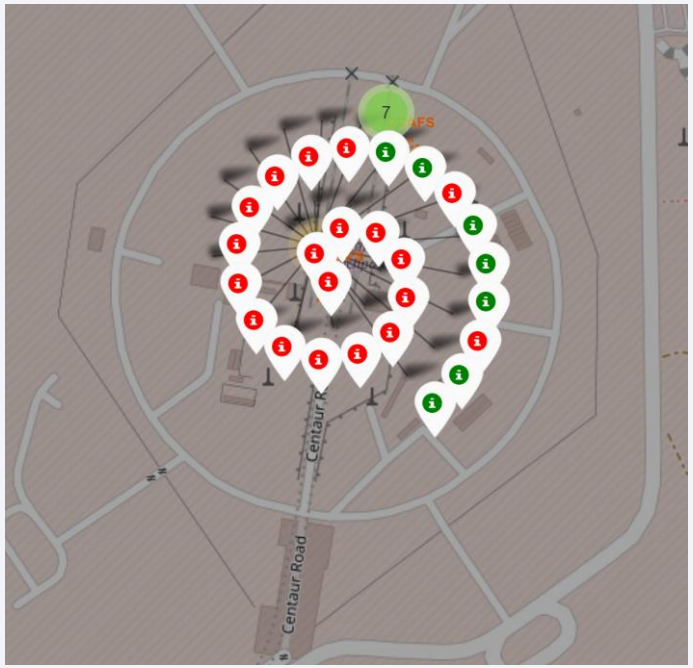
- All launch sites are near the coast for safety measures



Launch Outcomes

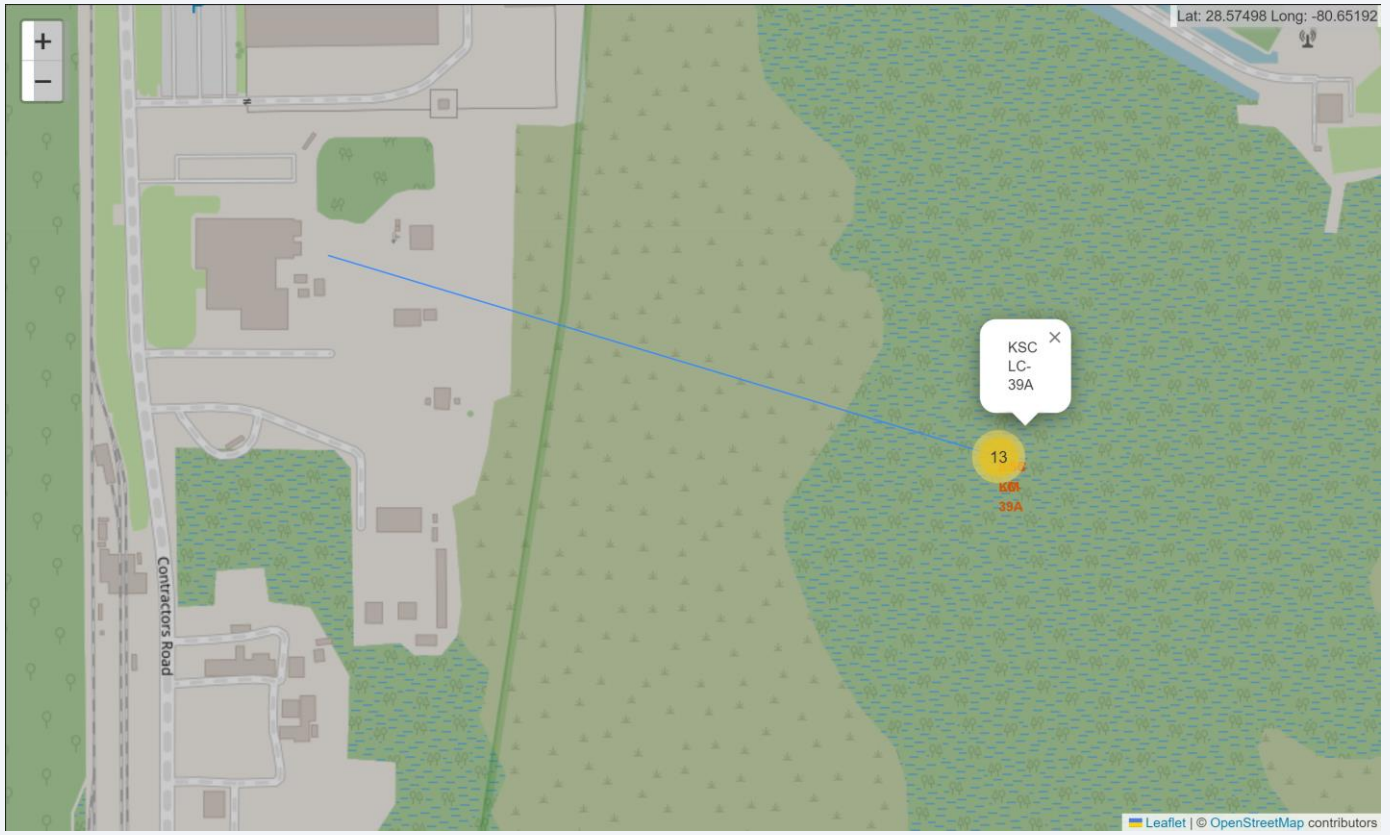
- Green markers indicate successful launches, while red ones indicate failure

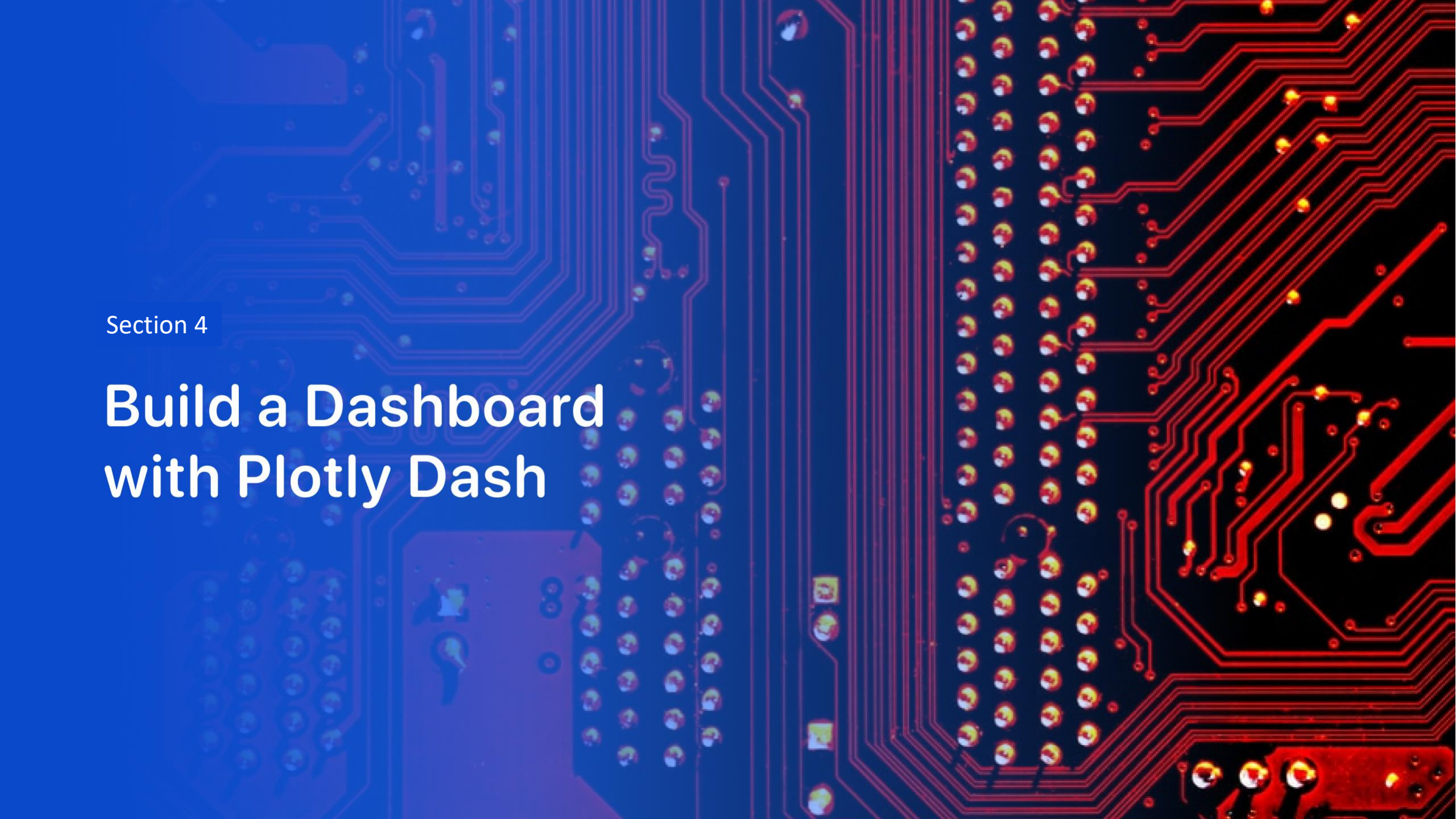




Proximities of launch site KSC LC-39A

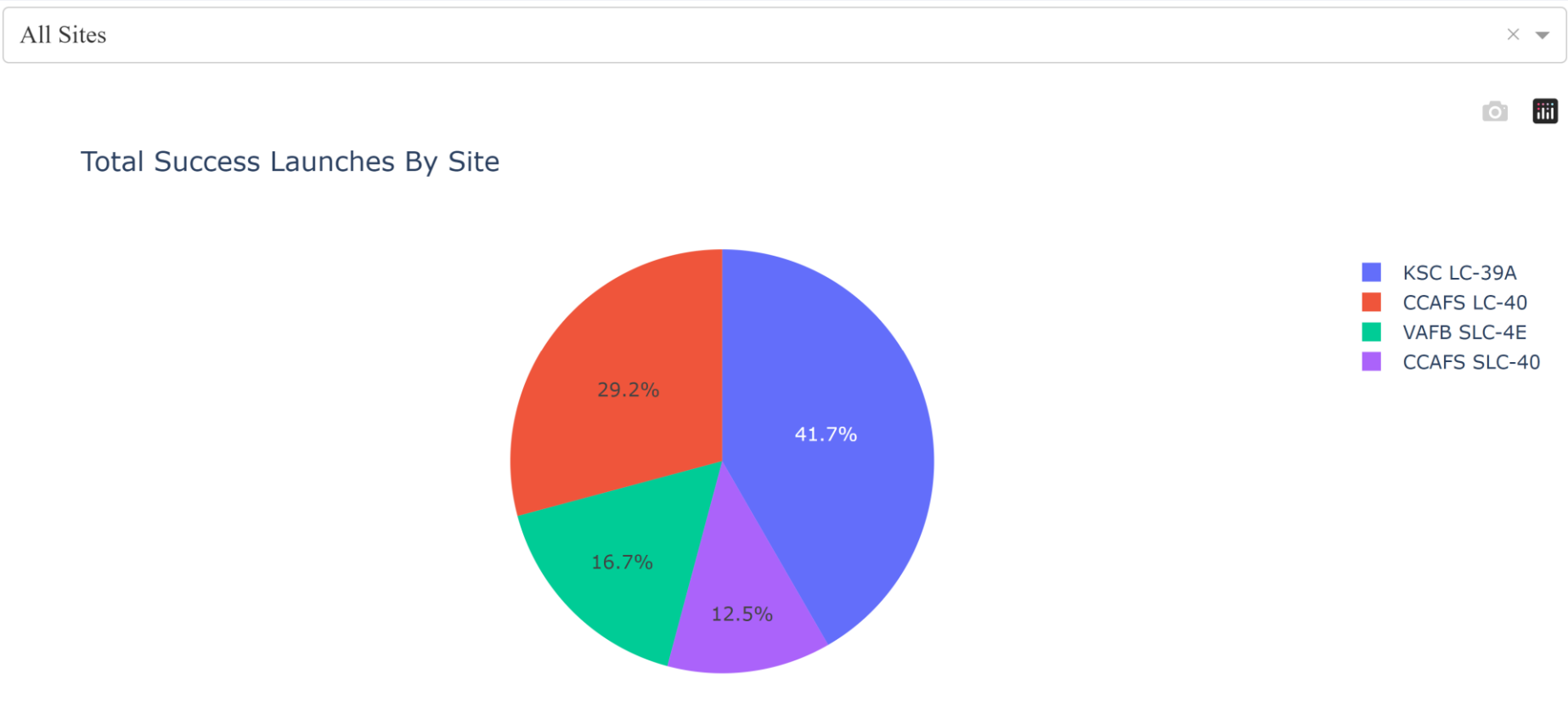
- The launch site KSC LC-39A is far away from the nearest inhabited area




Section 4
Build a Dashboard with Plotly Dash

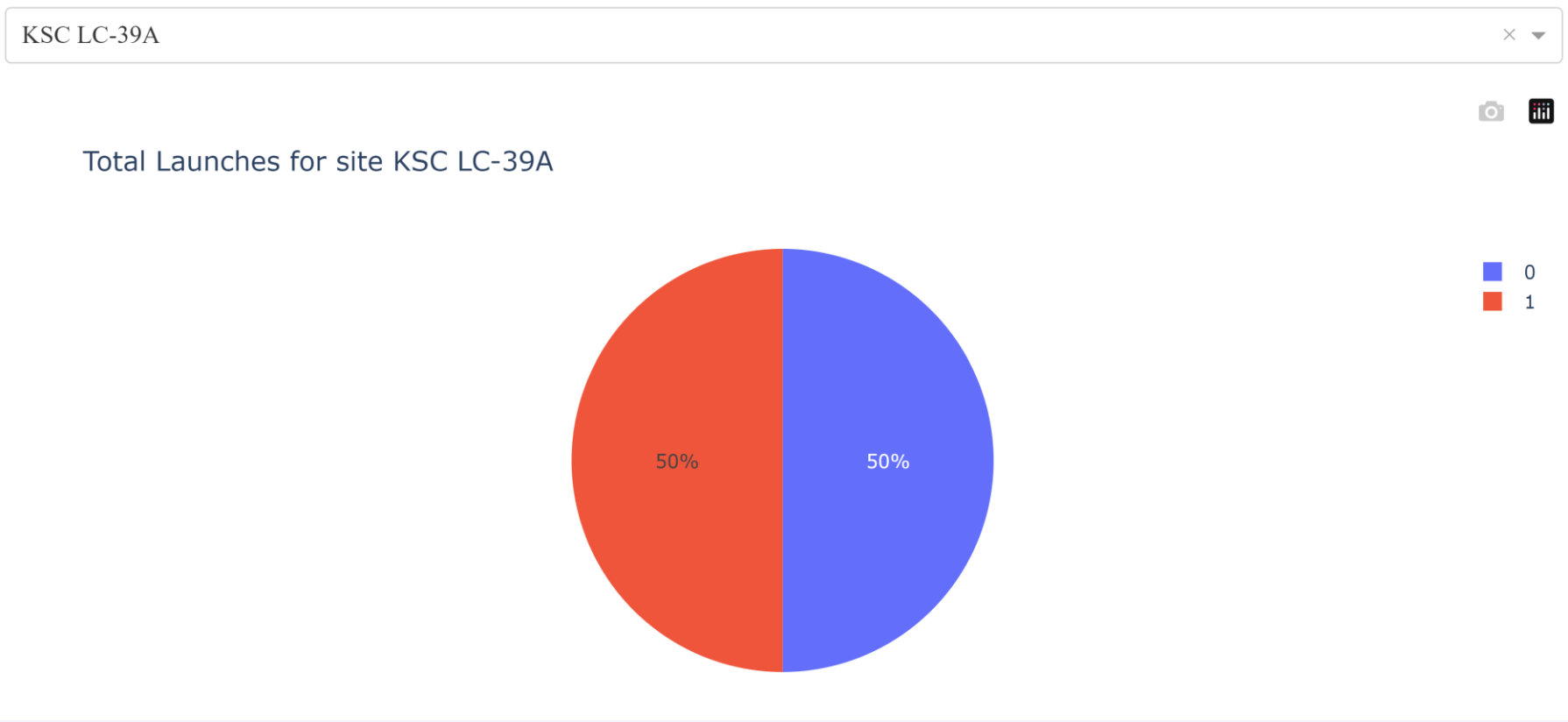
Success Count for Launch Sites

- The launch site KSC LC-39A has the most success launches, probably because it has the highest amount of launches overall



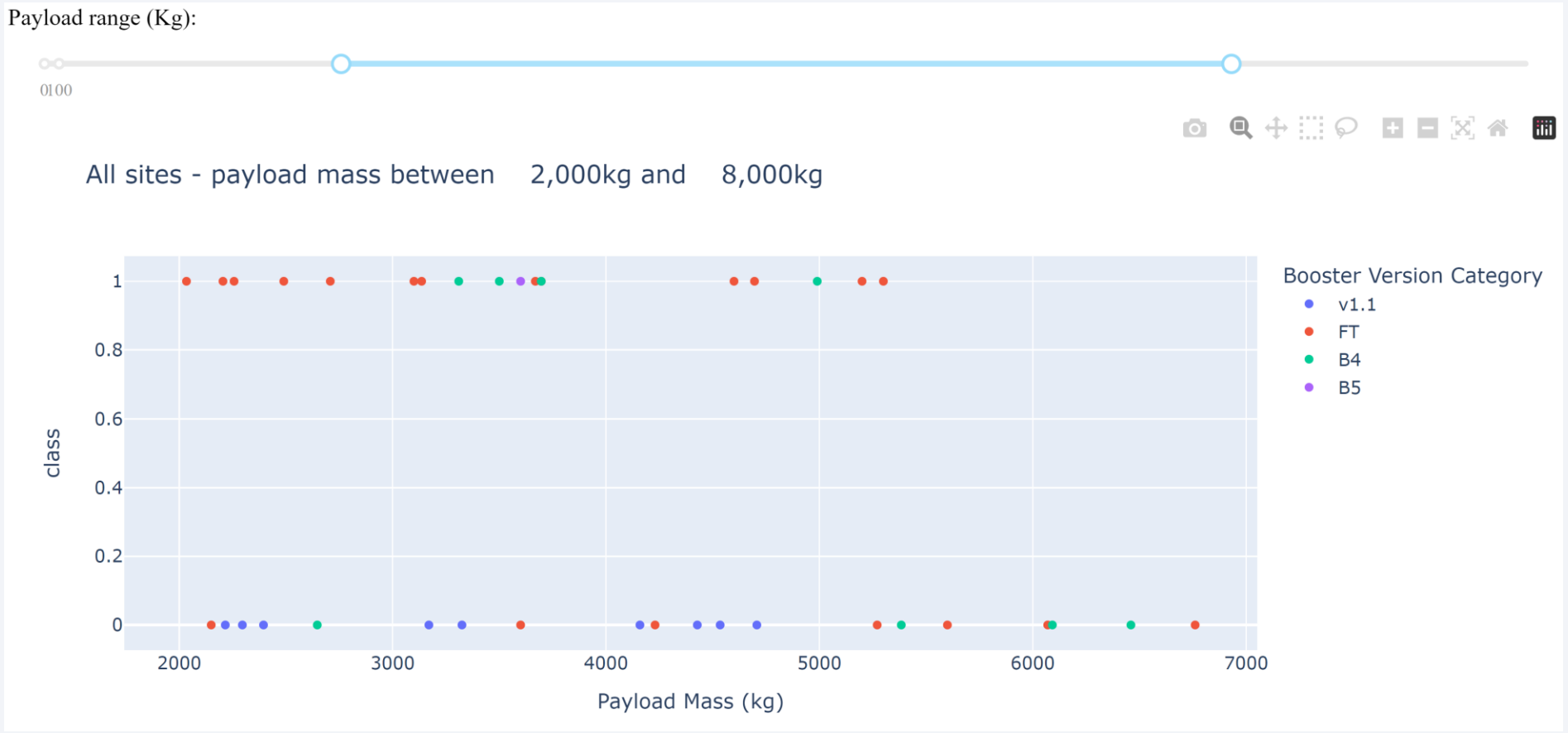
Successful Launches of Site KSC LC-39A

- Although the launch site KSC LC-39A has the most success launches, only 50% of its launches were successful overall



Payload vs. Launch Outcome for All Sites

- In the most common payload range, between 2,000 and 8,000 kg, the booster FT carried the widest range of payload and had the most success

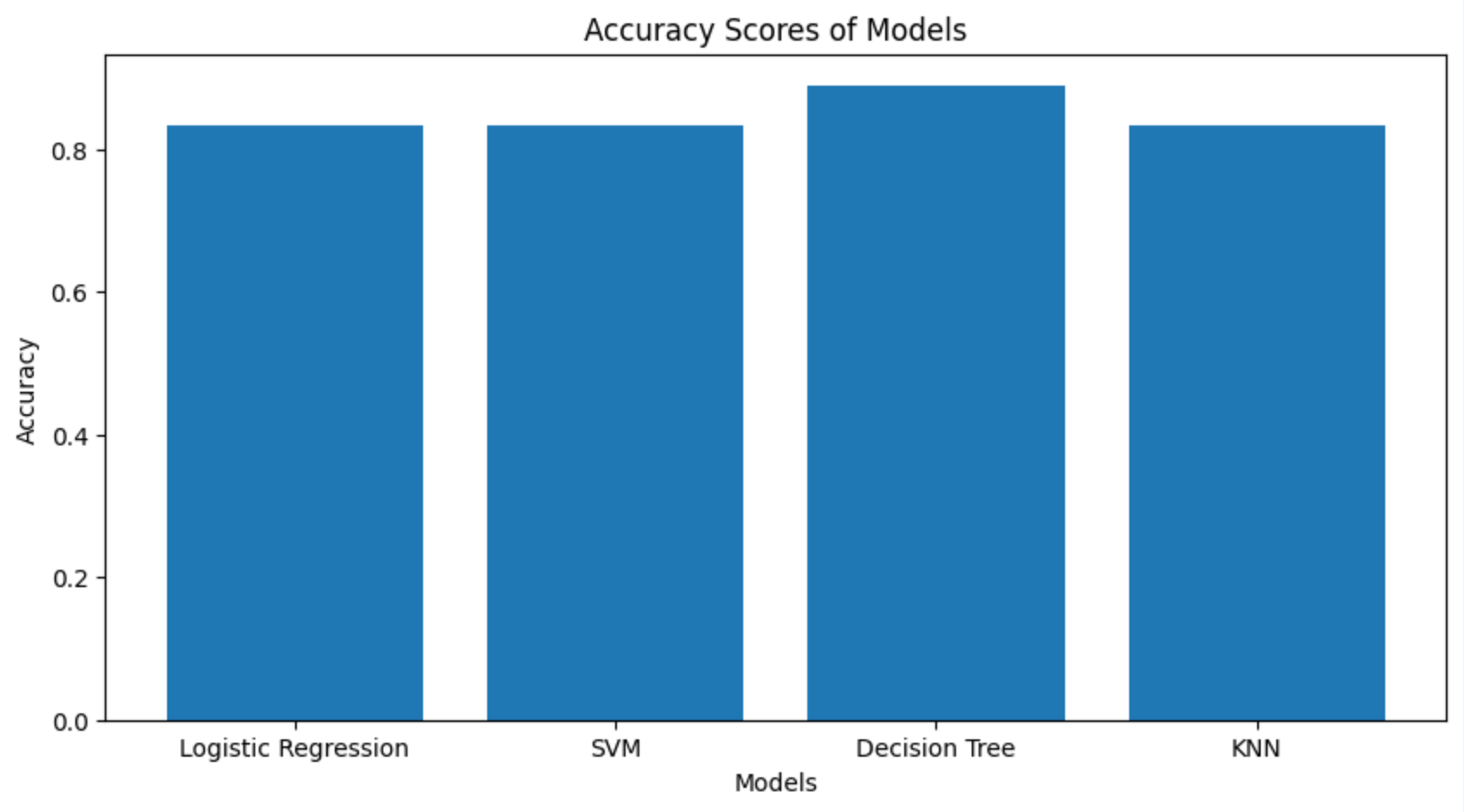


Section 5

Predictive Analysis (Classification)

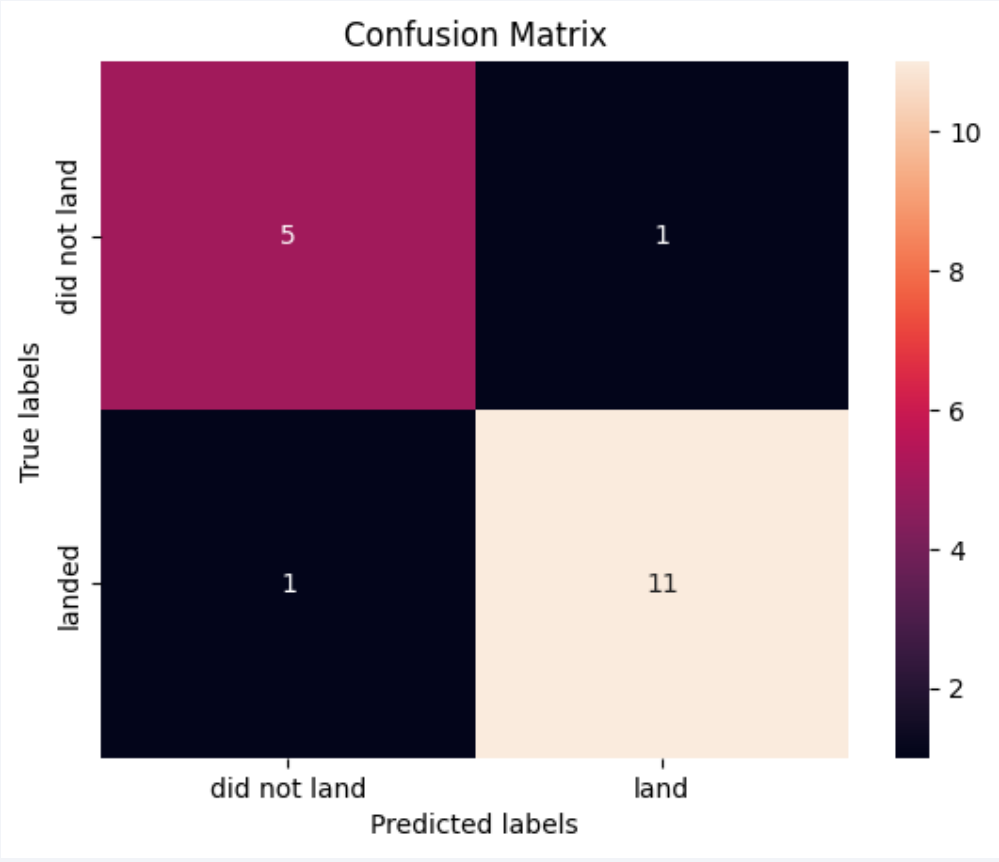
Classification Accuracy

- The model with the highest accuracy is the Decision Tree, with 88%



Confusion Matrix

- The confusion matrix of the Decision Tree model show the highest amount of both True Positives and True Negatives



Conclusions

- Data was analyzed from different sources, and an overall improvement in the success rate of the launches can be seen over time.
- The launch site with the most launches and successes is KSC LC-39A
- All launch sites are strategically located near the coast and distanced from inhabited areas
- Launches with payloads above 8,000kg are more successful
- The Decision Tree Classifier model is the best one that can be used by SpaceY to predict successful landings and increase its profits

Appendix

- For running the notebooks locally, some modifications had to be made to the original notebooks. Such as using the ‘requests’ library in python, instead of using javascript to load the datasets.

Thank you!