IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Francis Emmanuel Calingo
March 06, 2025

Outline

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

Executive Summary

- Outlined below is a list of methodologies implemented for this initiative:
  - **Data pre-processing:** data collection via SpaceX API and web scraping, data wrangling
  - **Exploratory Data Analysis:** SQL, Python (Pandas, Matplotlib)
  - **Data visualization:** Mapping with Folium, dashboarding techniques with Plotly Dash
  - **Predictive Analysis:** Classification Modelling
- Summary of all results
  - Sourcing of data successful, with exploratory data analysis allowing us to determine most important features to help us predict the success of landings.
  - Using said features, the best performing classification model was deployed to predict performance of space landings.

Introduction

The Cold War-era Space Race between the United States and the Soviet Union, which captivated much of the world, may have long passed us, but a new space race between our company and our primary competitor, SpaceX, could commence, ***if we execute our plans effectively.*** In order to do that, we first have to understand why SpaceX is such a successful company. That is because unlike other competitors, SpaceX keeps their rocket launches relatively inexpensive by reusing their first stages from their rockets. It then follows that studying factors that maximizes the success of a launch is key to keeping costs down.

By scraping available data to determine and visualize the cost of each launch, then using that information to feed a machine learning model and train it to help us determine of SpaceX will reuse the first stage of a rocket launch. We will also use the available information to determine the best launch sites to optimize the probability of a successful landing.

Introduction (cont'd)

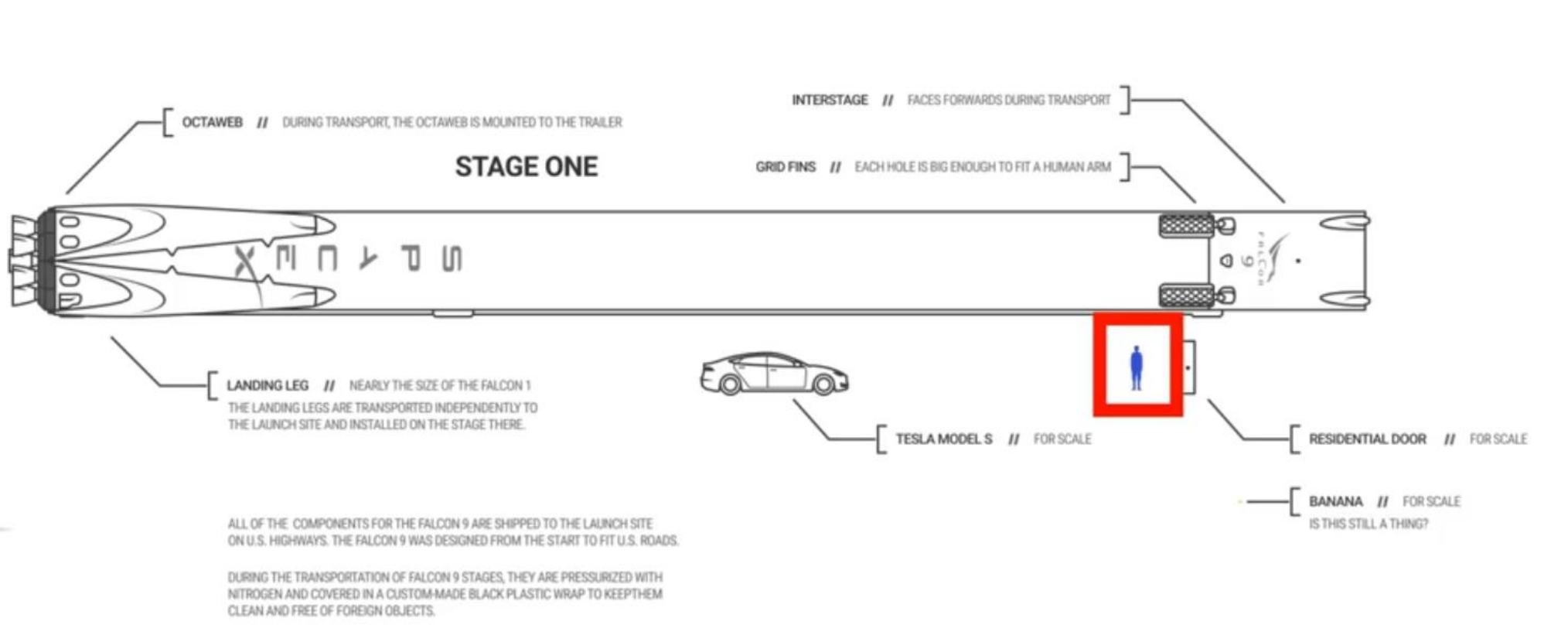


Diagram of a rocket's Stage One.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Scraped using SpaceX API: https://api.spacexdata.com/v4/rockets/
  - Scraped from the Wikipedia article "List of Falcon 9 and Falcon Heavy launches": https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling
  - Exploratory data analysis for ascertaining descriptive statistical patterns, and determining training labels for the machine learning model.

Methodology (cont'd)

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - After data wrangling and exploratory data analysis, data was standardized, assigned into either a training or testing set, and the test set fed into four different machine learning models, where each of their performance were evaluated before deployment.

Data Collection

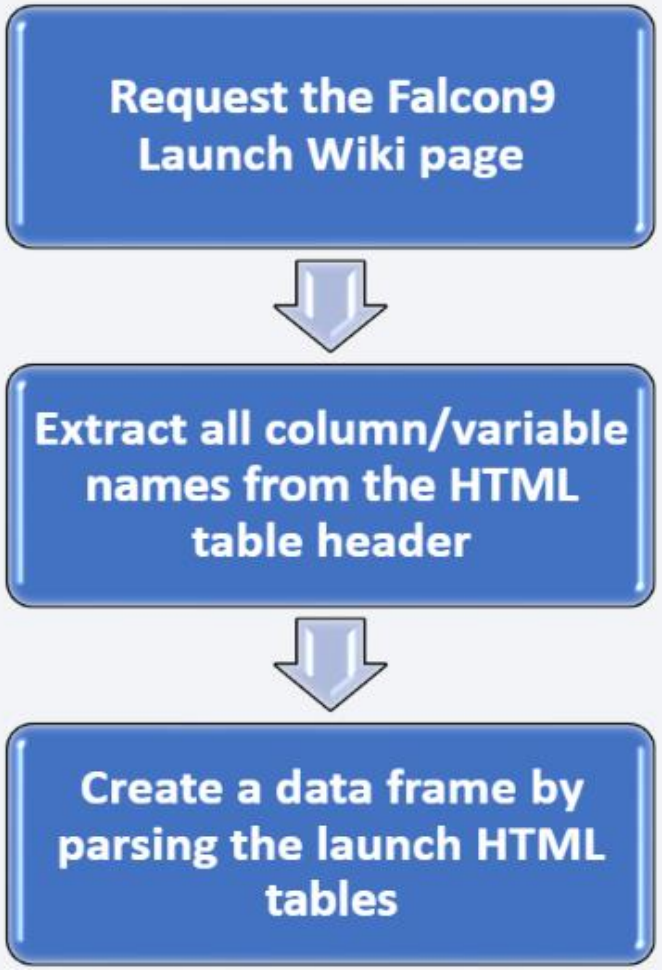
- Data was collected from two sources:
  - SpaceX API: https://api.spacexdata.com/v4/rockets/
  - Wikipedia article "List of Falcon 9 and Falcon Heavy launches": https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- The flowchart on the right represents the workflow implemented to collect data from the SpaceX API.
- An application program interface (API) is a mechanism that allows for two computer applications to communicate and connect with each other. Many websites, from the National Hockey League to YouTube, have their own API, which allows local machines such as your laptop and mobile phone to access them and their data.
- GitHub link to notebook for SpaceX API calls: https://github.com/Francis-Calingo/IBM-Capstone-2/blob/main/SpaceY_DataCollection_API.ipynb

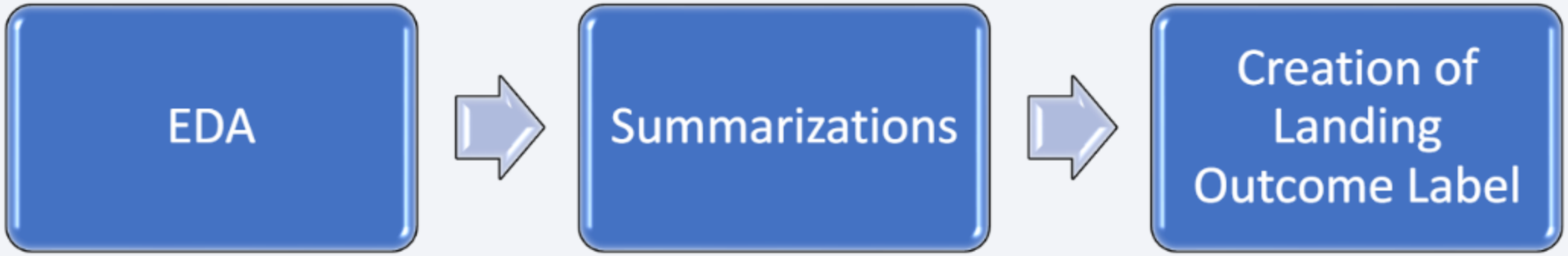
Data Collection - Scraping

- The flowchart on the right represents the workflow implemented to collect data from the Falcon 9 Launch Wikipedia article via web scraping.
- Web scraping is the process of collecting publicly accessible content from a website and saving it in a database or file.
- GitHub link to notebook for web scraping: https://github.com/Francis-Calingo/IBM-Capstone-2/blob/main/SpaceY_DataCollection_Webscraping.ipynb



Data Wrangling

- Below represents the flow chart for the data wrangling process
- Exploratory Data Analysis (EDA) and data cleaning (e.g., checking for null values) were performed on the dataset. The results allowed us to summarize the following: raw launch count by site, number and occurrences of each orbit type and mission outcomes.
- The last step was creating a landing outcome label from the Outcome column of the dataset.
- GitHub link to notebook for data wrangling: https://github.com/Francis-Calingo/IBM-Capstone-2/blob/main/SpaceY_DataWrangling.ipynb



EDA with Data Visualization

- 8 plots were created:
  - Scatter point charts to compare a pair of variables of interest:
    - Flight number vs. launch site
    - Payload mass vs. launch site
    - Flight number vs. orbit type
    - Payload mass vs. orbit type
  - 1 bar chart to compare categorical data against a single metric. In this, the categorical data was orbit type, and the metric was success rate of each type.
  - 1 line chart to plot the progression of a metric over time. In this case, the launch success rate of each launch from 2010-2020.
- GitHub link to EDA-Data Visualization notebook: https://github.com/Francis-Calingo/IBM-Capstone-2/blob/main/SpaceY_EDA_Viz.ipynb

EDA with SQL-Queries

- GitHub link to SQL notebook: https://github.com/Francis-Calingo/IBM-Capstone-2/blob/main/SpaceY_EDA_SQL.ipynb
- SQL Queries used:
  - To display the names of the unique launch sites used for the mission:

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

  - To display launch sites whose names start with 'CCA':

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

EDA with SQL-Queries (cont'd)

- To display the total payload mass carried by boosters launched by NASA (CRS):

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';
```

- To display average payload mass carried by booster version F9 v1.1:

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

- To list the date when the first successful landing outcome in ground pad was achieved:

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

- To 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)';
```

EDA with SQL-Queries (cont'd)

- To list the total number of successful and failure mission outcomes:

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

- To list the names of the BOOSTER_VERSION which have carried the maximum payload mass:

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

EDA with SQL-Queries (cont'd)

- To 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:

```
%sql SELECT substr(Date,6,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing_Outcome] \
FROM SPACEXTBL \
where [Landing_Outcome] = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

- To 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 count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '2010-06-04' AND '2017-03-20' group by [Landing_Outcome] order by count_outcomes DESC;
```

Build an Interactive Map with Folium

- The following map objects were added to the interactive map:
  - **Markers;** Used to mark singular locations in a map. In this case, launch sites.
  - **Circles;** Used for demarcating areas around certain coordinates, such as the area in near proximity to a launch site.
  - **Marker clusters;** Used to group multiple occurrences or locations within a single coordinate. In this case, successful and failed launches associated for each launch site.
  - **Lines;** used for indicating distances between two coordinates. In this case, the distance between launch sites and its proximities.
- GitHub link to Folium Map: https://github.com/Francis-Calingo/IBM-Capstone-2/blob/main/SpaceY_FoliumMap.ipynb

Build a Dashboard with Plotly Dash

- The dashboard contains two types of plots: pie graph and scatter plot.
- For each site, the pie chart is meant to visualize the proportion of successful and unsuccessful launches for the particular site.
- For the aggregated successful launches, the pie chart visualizes the proportion of each sites that contributed to the total successful launches.
- The scatter plot is meant to visualize the success (or lack thereof) of each booster version by payload mass (kg).
- By combining both analyses, we are able to draw conclusions about the best site and booster specifications to optimize launch success.
- GitHub link to Plotly Dash notebook: https://github.com/Francis-Calingo/IBM-Capstone-2/blob/main/SpaceY_Dashboard.ipynb

Predictive Analysis (Classification)

- The model development process, as outlined in the flow chart below, was implemented on four machine learning algorithms: Logistic Regression, Support vector Machine (SVM), Decision Tree Classifier, and K-Nearest Neighbours (KNN) Classifier.

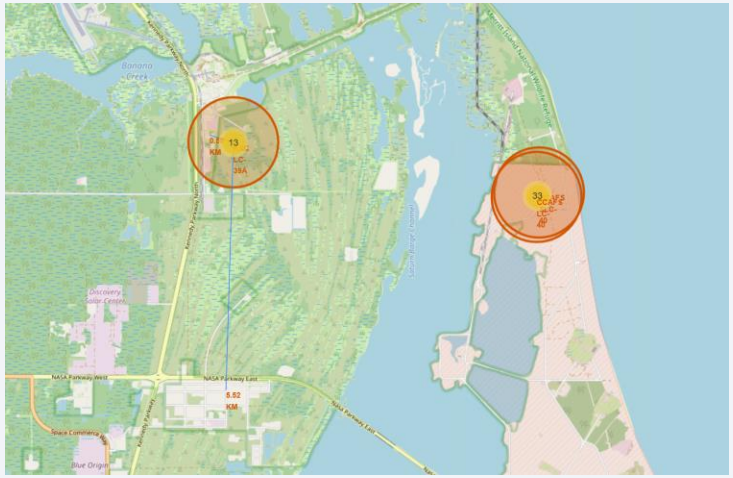
- GitHub link to the Predictive Analysis notebook: https://github.com/Francis-Calingo/IBM-Capstone-2/blob/main/SpaceY_MLPredictiveModel.ipynb

Results-Exploratory Data Analysis

- Space X uses 4 different launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E
- The success rate for each site improved over time.
- Very high success rate for payloads over 8000 kg for launch sites, and 9000 kg for orbits.
- Orbit types ES-L1, GEO, HEO, and SSO are the most successful orbit types.
- Total payload mass for NASA launches: 111,268 kg.

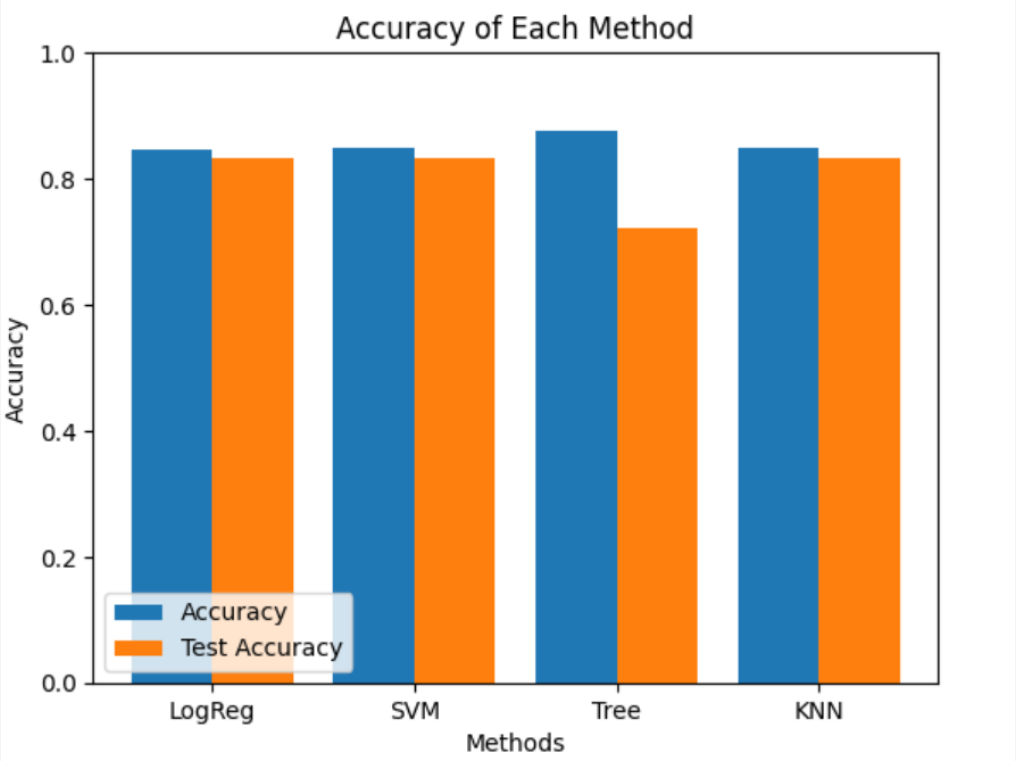
Results-Interactive Geospatial Analysis

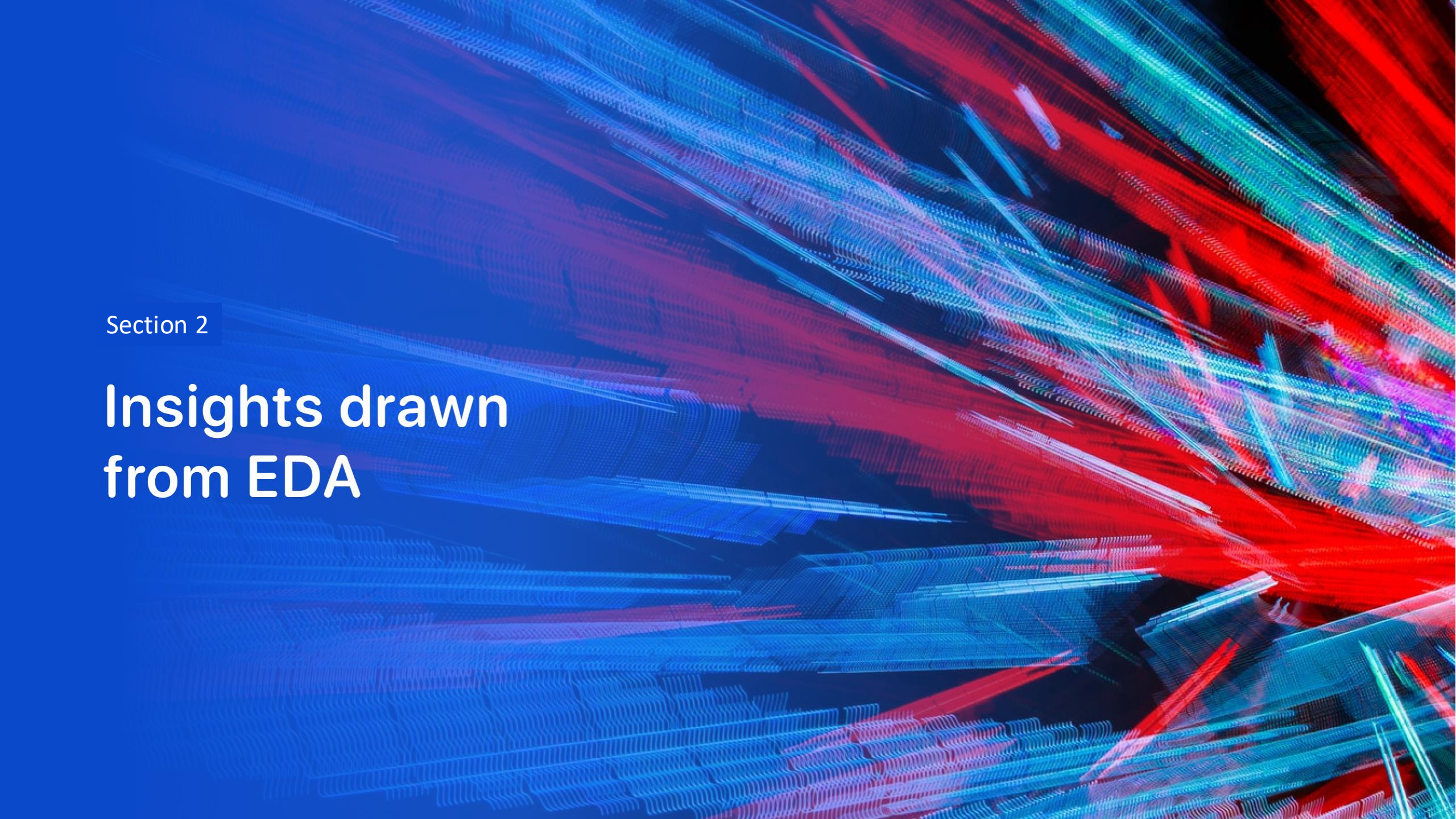
- The sites were concentrated in the coastal part of Southern California and Florida, possibly due to safety considerations in the event of a failed launch, which allows debris to have a better chance of falling into the ocean rather than highly populated centres.
- Despite the relative isolation, there is sufficient infrastructure in the vicinity of the launch sites to help sustain them.



Results-Predictive Analysis

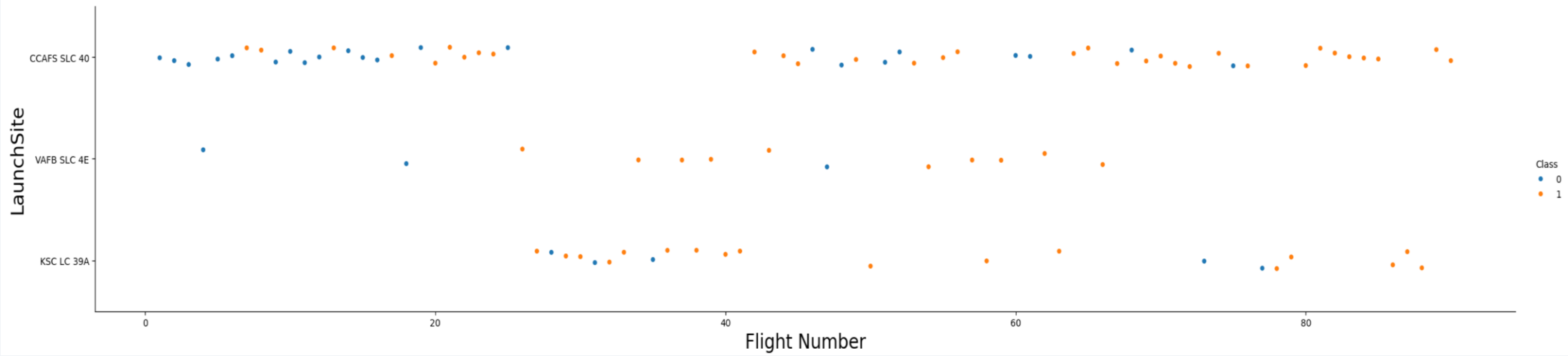
Decision Tree Classifier, despite having the lowest test accuracy, had by far the highest accuracy overall, suggesting that it is the machine learning algorithm that SpaceY should deploy for higher accuracy in predicting successful landings and launches.




Section 2
Insights drawn from EDA

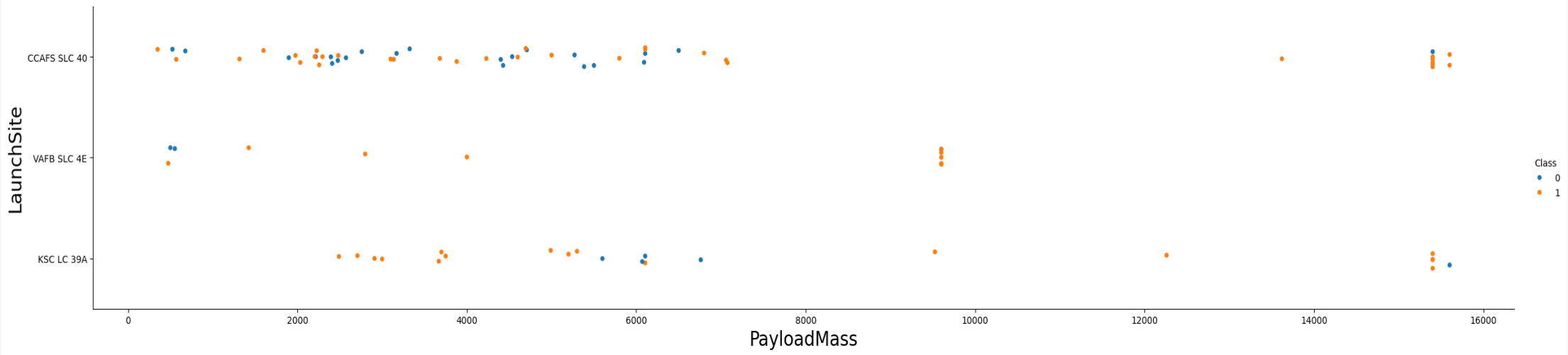
Flight Number vs. Launch Site

From the scatter plot, we can see that high flight numbers tend to be successful (i.e., more often than not, belonging in Class 1) while the reverse is true for lower flight numbers. This makes sense, as we would expect the frequency of successful launches to increase as more flights are implemented and allow us to make improvements over time. In terms of cumulative success rate, KSC LC-39A. However, we see that CCAFS SLC-40 has been pretty successful recently.

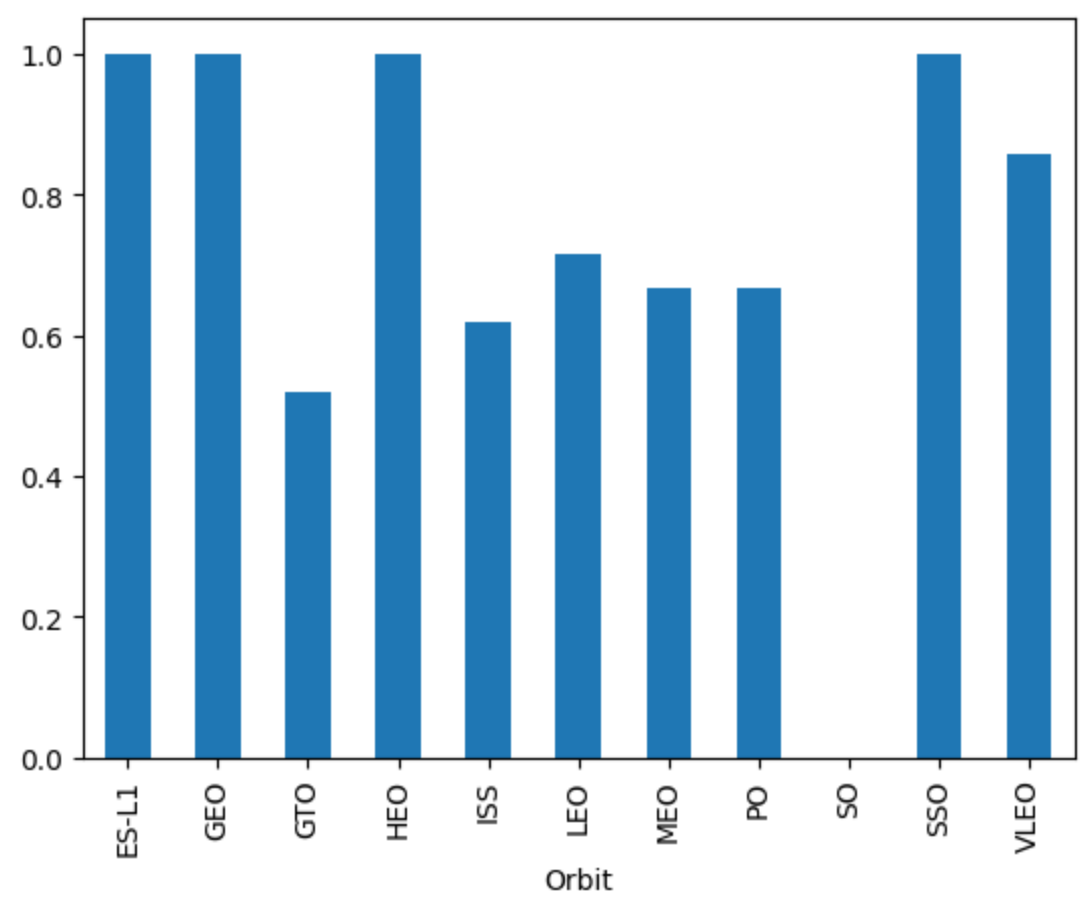


Payload vs. Launch Site

Payloads over 8,000 kg collectively have a very high success rate. Payloads over 12,000 kg seems to be possible only on CCAFS SLC-40 and KSC LC-39A launch sites, while the limit for the VAFB SLC-4E site appears to be 10,000 kg.



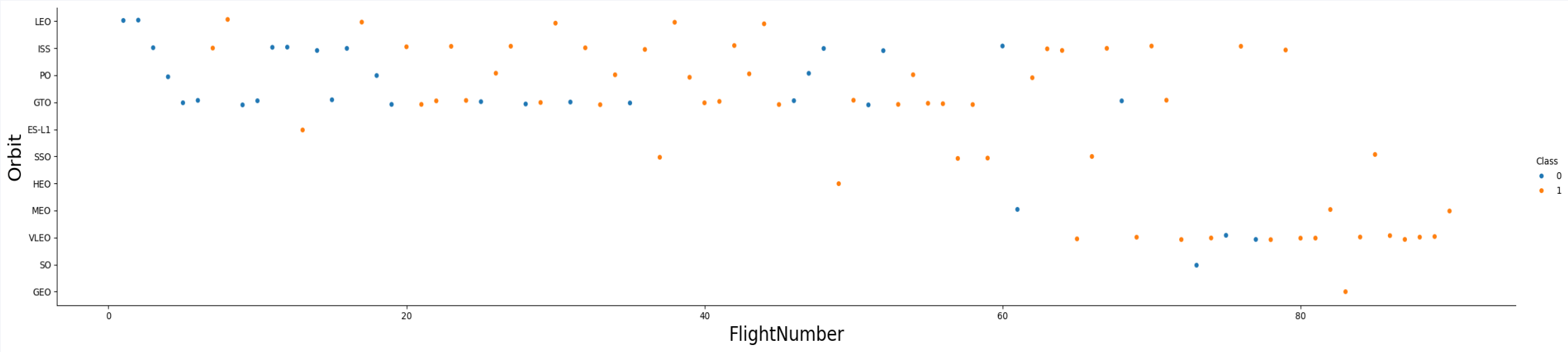
Success Rate vs. Orbit Type



This is a bar plot visualizing the success rate of each orbit type. Orbit types ES-L1, GEO, HEO, and SSO are the most successful orbit types.

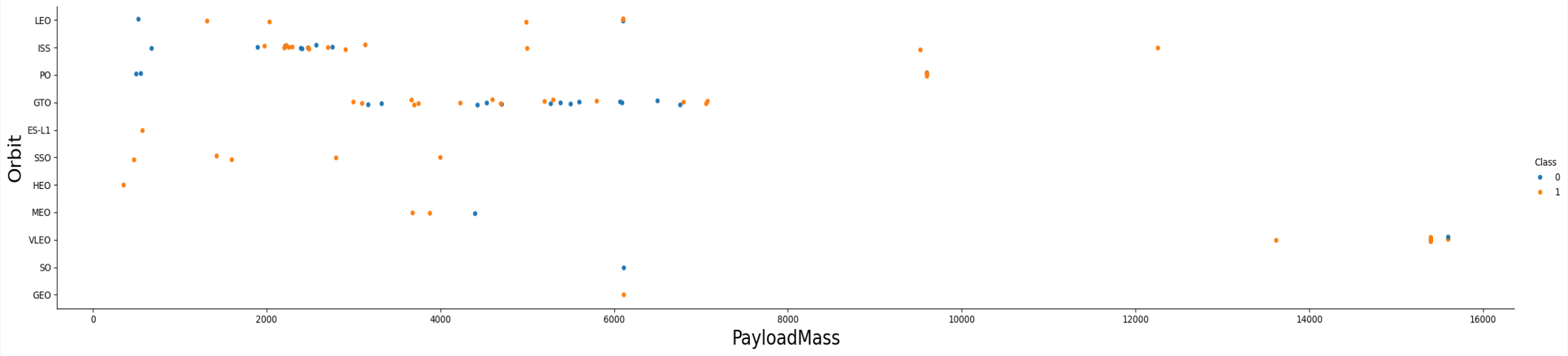
Flight Number vs. Orbit Type

The plot suggests that each orbit type's success rate improved over time. The low frequency of SO and GEO orbits could suggest that they are expensive, despite GEO's success. Conversely, the failure of SO could have been so catastrophic and costly that it was never attempted again.



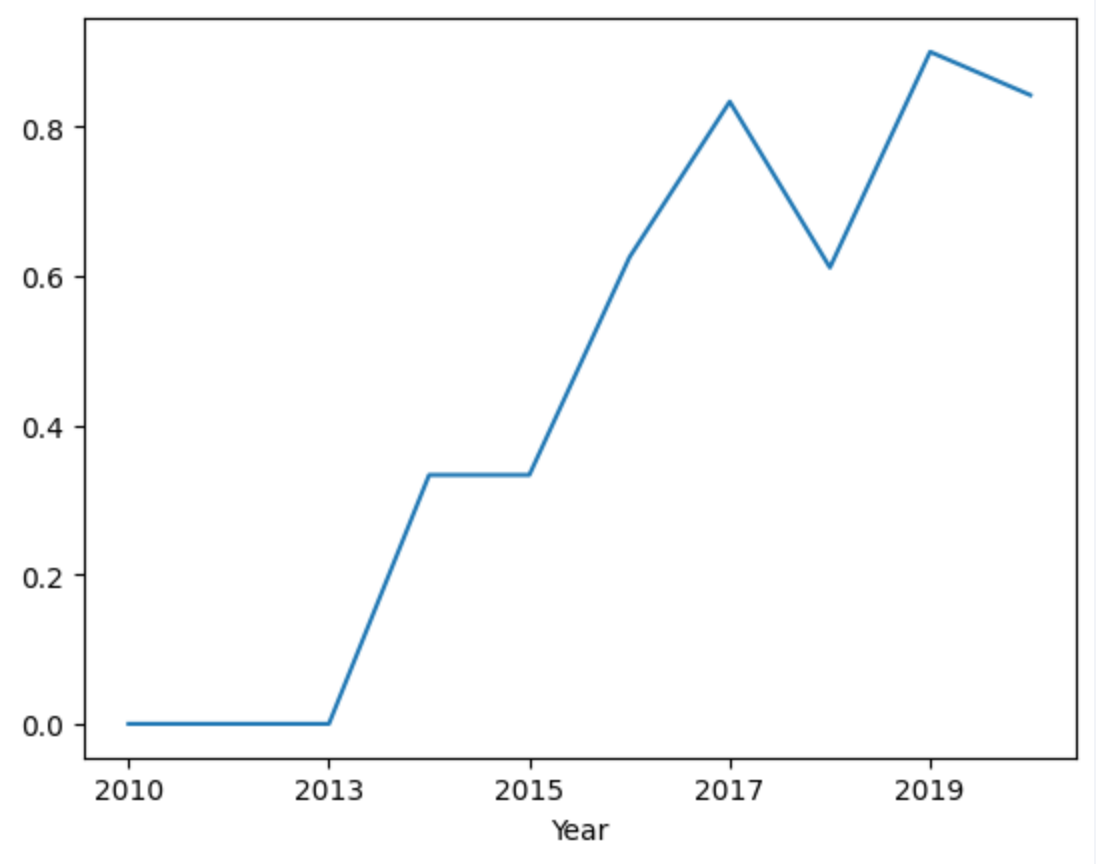
Payload vs. Orbit Type

Despite being comparatively few, orbits with payload mass of over 9000 kg appear to have a very high success rate.



Launch Success Yearly Trend

This is a line ploy visualizing the yearly average success rate of rocket launches from the 2010s. We would expect the success rate to increase with time, as shown by the plot, as better innovations, alongside improvements made as a result of failed launches, would increase the success rate.



All Launch Site Names

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

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

The query selects unique occurrences of “LAUNCH_SITE” values from the dataset.

Launch Site Names Begin with '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

The query selects a sample of 5 launches from the list of launch sites that starts with 'CCA'.

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';
```

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

TOTAL_PAYLOAD
111268

The query sums the payload from entries coded 'CRS' (corresponding to NASA) and names the total "TOTAL_PAYLOAD".

Average Payload Mass by F9 v1.1

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

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

AVG_PAYLOAD
2928.4

The query takes all the entries where the booster version is F9 v1.1, and takes the average.

First Successful Ground Landing Date

```
%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

The query uses the MIN function to find the very first date, with the constraint that the landing outcome was "Success (ground pad)".

Successful Drone Ship Landing with Payload between 4000 and 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

The query gives us the names of boosters using the following constraints:

- Successfully landed on drone ship
- Payload mass greater than 4000 kg but less than 6000 kg

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL 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

The query counts all of the mission outcomes, then groups them by missions outcome type.

Boosters Carried Maximum Payload

```
%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

The query takes the maximum payload for each type of booster, then orders the results.

2015 Launch Records

```
%sql SELECT substr(Date,6,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing_Outcome] \
FROM SPACEXTBL \
where [Landing_Outcome] = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

The query extracts the list of 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 count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '2010-06-04' AND '2017-03-20' group by [Landing_Outcome] order by count_outcomes DESC;
```

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

Landing_Outcome	count_outcomes
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

The query 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

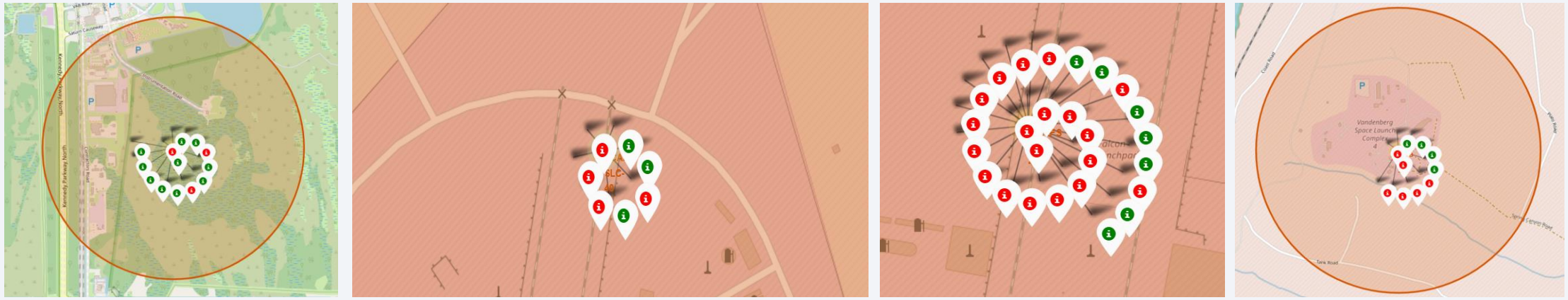
Section 3

Launch Sites Proximities Analysis

Map of All Launch Sites

As noted in an earlier slide, all the launch sites are concentrated in the coastal part of California and Florida, possibly to prevent any potential debris from failed launches from falling into land and populated areas.

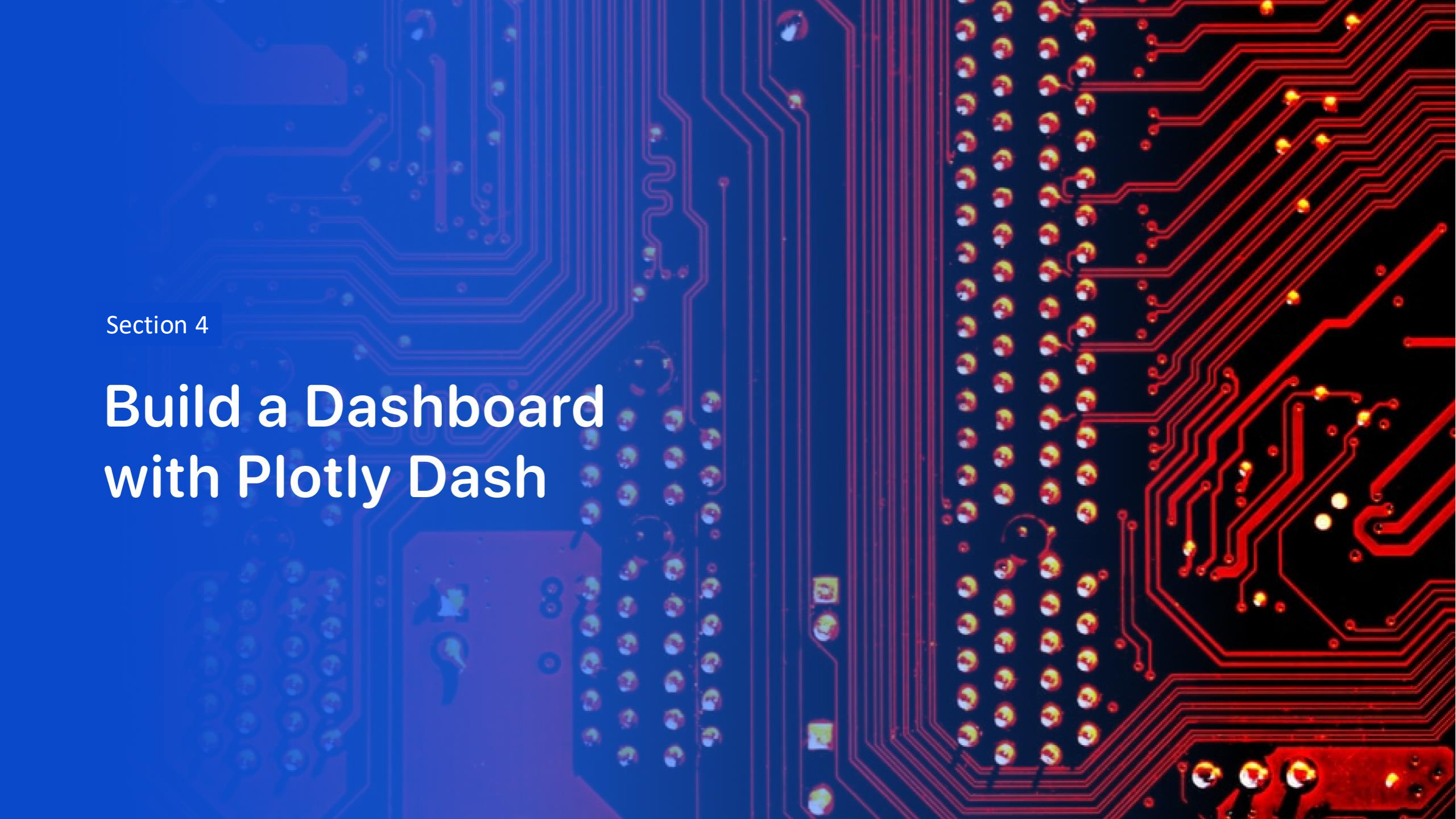
Launch Outcomes of Launch Sites



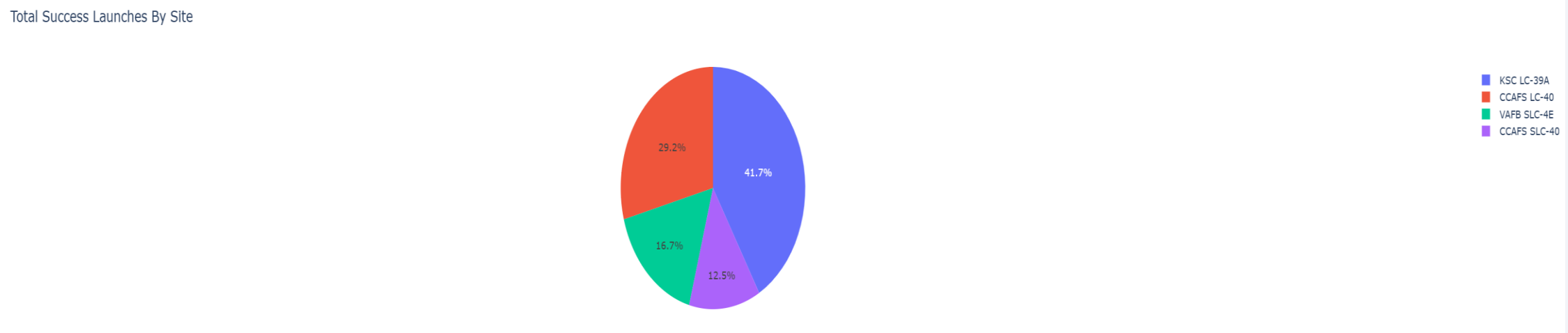
- Green=successful launch, red=unsuccessful launch
- First image (far left): KSC LC-39A
- Second image: CCAFS SLC-40
- Third image: CCAFS LC-40
- Fourth image (far right): VAFB SLC-4E
- Based on the Folium map, KSC LC-39A site is the most successful site.

Proximities of Launch Sites

- Here is the aforementioned KSC LC-39A and its distance to the nearest parking lots.
- A few highways and roads are within the vicinity of the site, suggesting that despite its relative isolation, it is still served well by good road infrastructure.

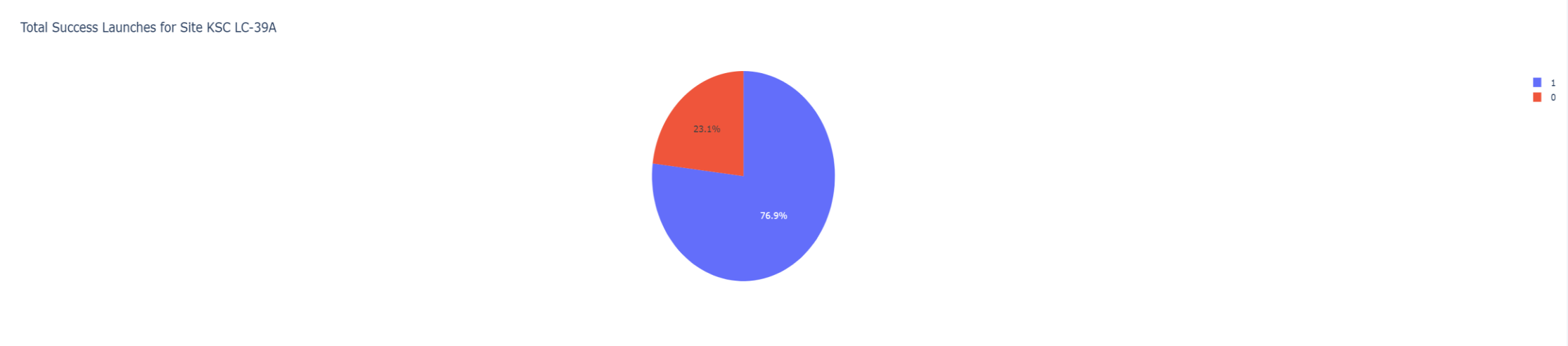

Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Site



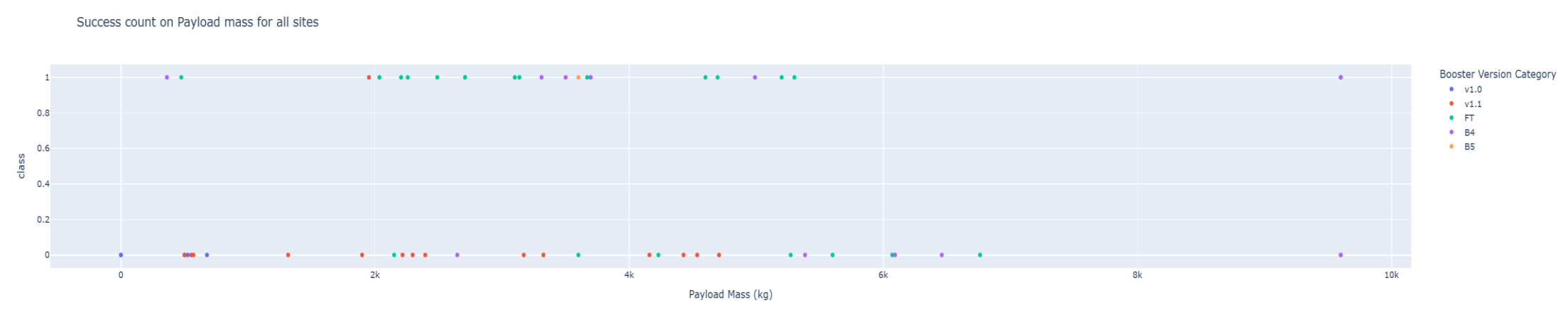
Making up 41.7% of the total number of successful launches, KSC LC-39A is the most successful site when using that metric, followed by CCAFS LC-40 with 29.2%.

Launch Success Rate of KSC LC-39A



Having the highest launch success rate amongst the 4 sites (with a success rate of 76.9%) allowed KSC LC-39A to be the most successful site, as per the chart in the last slide.

Outcomes of Payload Mass (kg) by Site



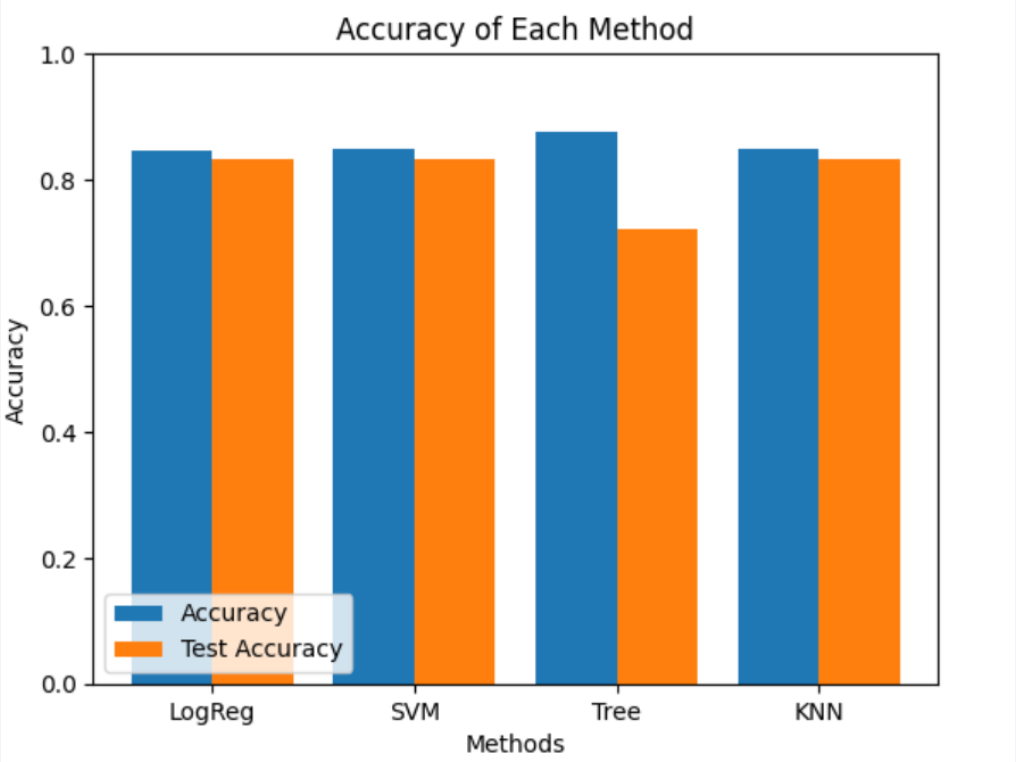
Most launches have a payload mass below 6000 kg. The number of launches with a payload mass greater than 6000 kg is too little to draw concrete conclusions about their successes, but it appears that the outcome is better for launches with payload mass below 6000 kg. FT appears to be the most successful booster based on the proportion of Class=1 points.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

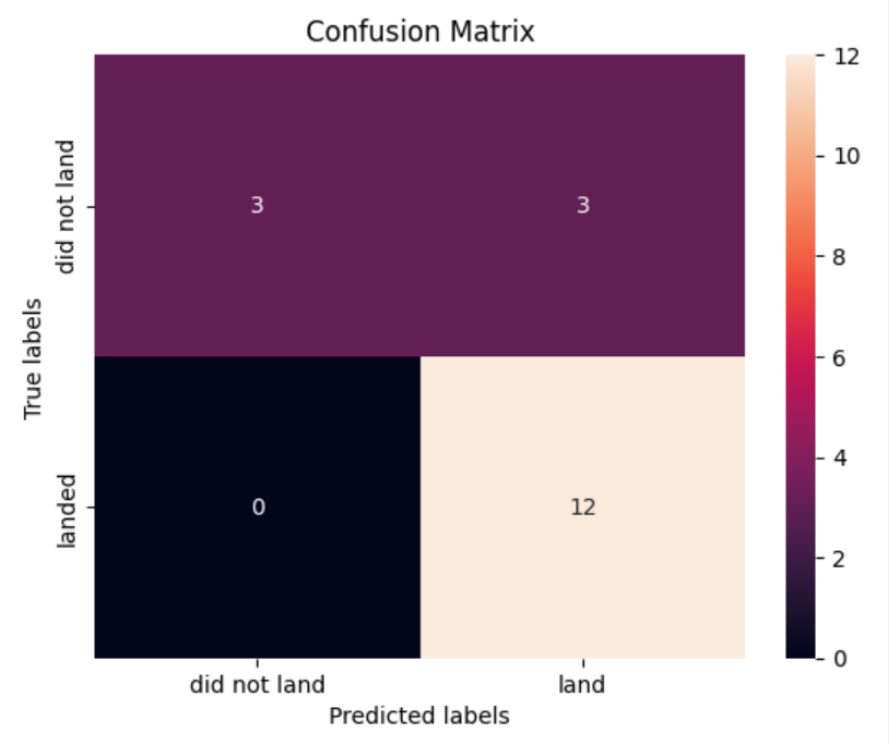
Despite having the lowest test accuracy score, Decision Tree Classifier was nevertheless the method with the highest classification accuracy.



Confusion Matrix

This is the confusion matrix for the aforementioned Decision Tree Classifier. The matrix helped visualized the success by plotting the number of true and false positives, as well as the true and false positives, from the 18 test samples.

- 12 true positives (the classification model correctly predicted that the first stage successfully landed)
- 3 true negatives (the model correctly predicted that the first stage did not land)
- 3 false positives (the model incorrectly predicted that the first stage did successfully land)
- 0 false negatives (the model incorrectly predicted that the first stage did not land)



Conclusions

- The best launch site is KSC LC-39A.
- Launches above 6,000 kg remain comparatively untested compared to launches below 6,000kg, but appear more risky, as none have been successful to date.
- Launches appear to be more successful over the passing years due to continued innovations, highlighting the continued need for further-refined research.
- Decision Tree Classifiers can be deployed to predict successful landings and save money.

Appendix

List of notebook outputs, and their URL links, for this research:

- Data collection with API: https://github.com/Francis-Calingo/IBM-Capstone-2/blob/main/SpaceY_DataCollection_API.ipynb
- Data collection with web scraping: https://github.com/Francis-Calingo/IBM-Capstone-2/blob/main/SpaceY_DataCollection_Webscraping.ipynb
- Data wrangling: https://github.com/Francis-Calingo/IBM-Capstone-2/blob/main/SpaceY_DataWrangling.ipynb
- Exploratory data analysis (EDA) with SQLlite: https://github.com/Francis-Calingo/IBM-Capstone-2/blob/main/SpaceY_EDA_SQL.ipynb

Appendix (cont'd)

- Exploratory data analysis (EDA) with pandas and matplotlib: https://github.com/Francis-Calingo/IBM-Capstone-2/blob/main/SpaceY_EDA_Viz.ipynb
- Geospatial analysis with Folium: https://github.com/Francis-Calingo/IBM-Capstone-2/blob/main/SpaceY_FoliumMap.ipynb
- Dashboarding with Plotly Dash: https://github.com/Francis-Calingo/IBM-Capstone-2/blob/main/SpaceY_Dashboard.ipynb
- Machine Learning Techniques for Predictive Modelling: https://github.com/Francis-Calingo/IBM-Capstone-2/blob/main/SpaceY_MLPredictiveModel.ipynb

Thank you!