IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

La Pyae Phyo
2nd April 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection using SpaceX REST API and web scraping techniques
  - Data wrangling to create success/fail outcome variable
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Interactive Visual Analysis with Folium
  - Interactive Dashboard with Plotly Dash
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis results
  - 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.

- ## Desirable answers

  - The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets.
  - Where is the best place to make launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collect data using SpaceX REST API and web scraping techniques
- Perform data wrangling
  - Wrangle data - by filtering the data, handling missing values and applying one hot encoding - to prepare the data for analysis and modeling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build models to predict landing outcomes using four different classification models. Tune and evaluate models to find best model and parameters.

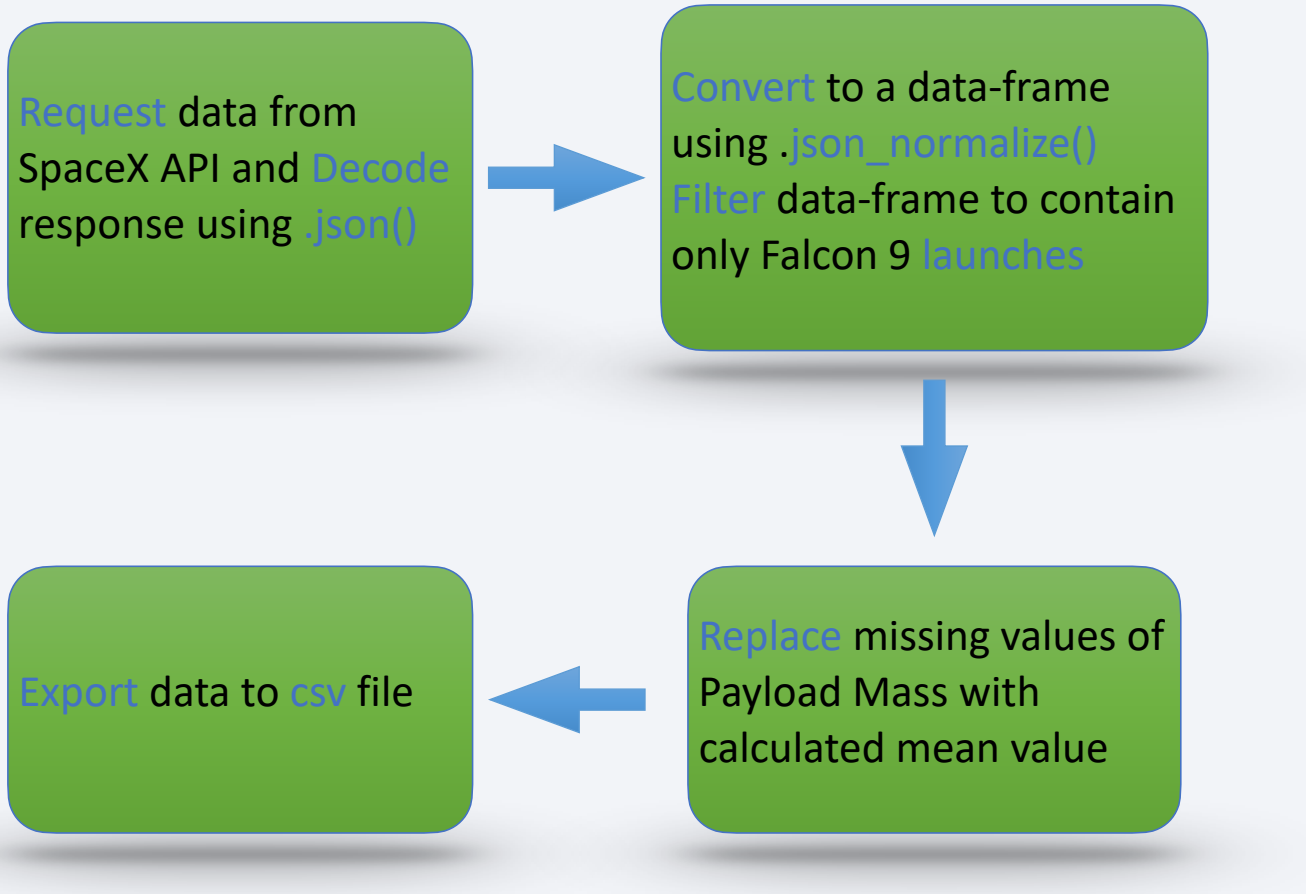
Data Collection

• Data sets were collected from Space X API (https://api.spacexdata.com/v4rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping techniques.

Data Collection – SpaceX API

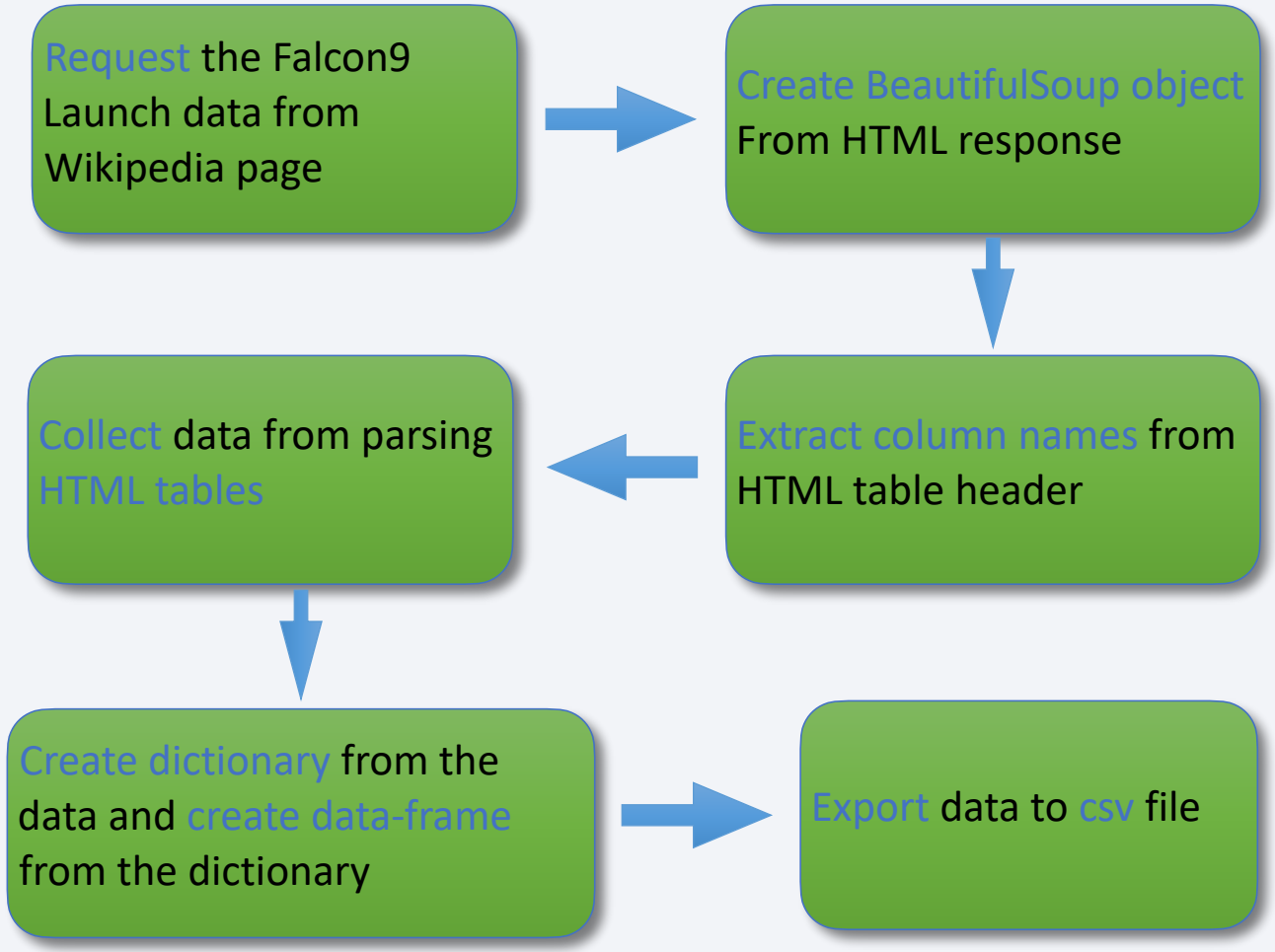
- We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.

- Github link:https://github.com/Lapyae-Phyo/Applied_data_science_capstone/blob/master/data_collection_api.ipynb/



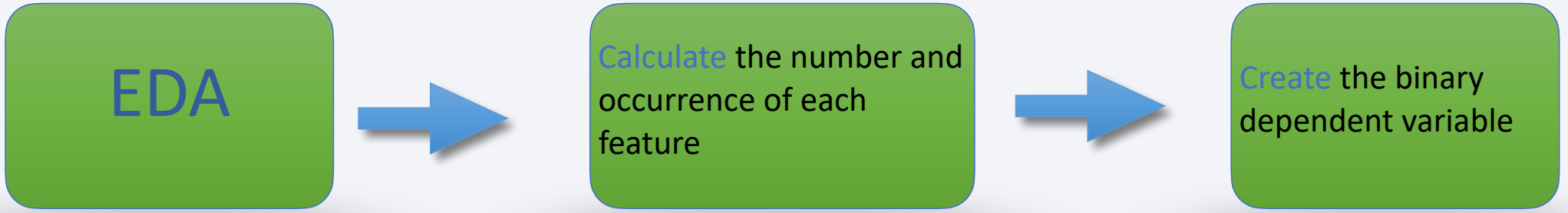
Data Collection - Scraping

- We applied web scrapping to get Falcon 9 launch records with BeautifulSoup
- We parsed the table and converted it into a pandas dataframe.
- Github link:https://github.com/Lapyae-Phyo/Applied_data_science_capstone/blob/master/Data_collection_webscraping.ipynb



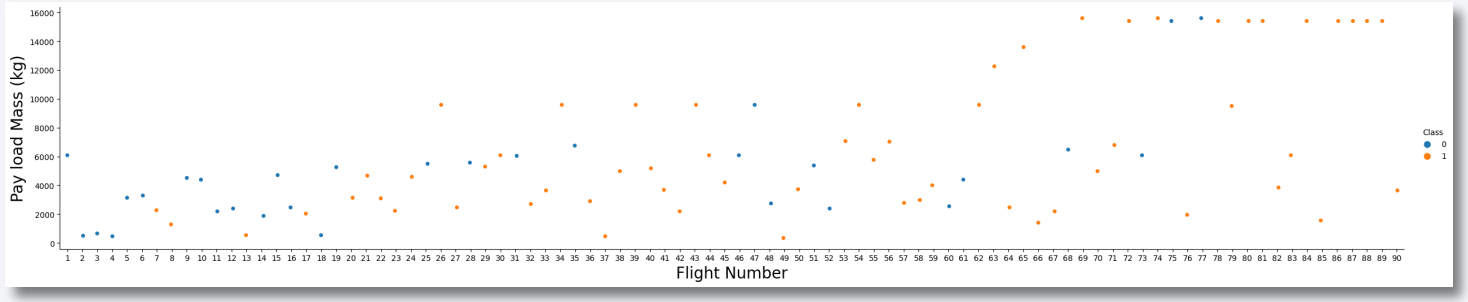
Data Wrangling

- Perform EDA and Determine data labels
- Calculate:
  - Number of launches for each site
  - Number and occurrence of orbit
  - Number and occurrence of mission outcome per orbit type
- Create binary landing outcome column (dependent variable)
- Github link: https://github.com/Lapyae-Phyo/Applied_data_science_capstone/blob/master/Data_wrangling.ipynb

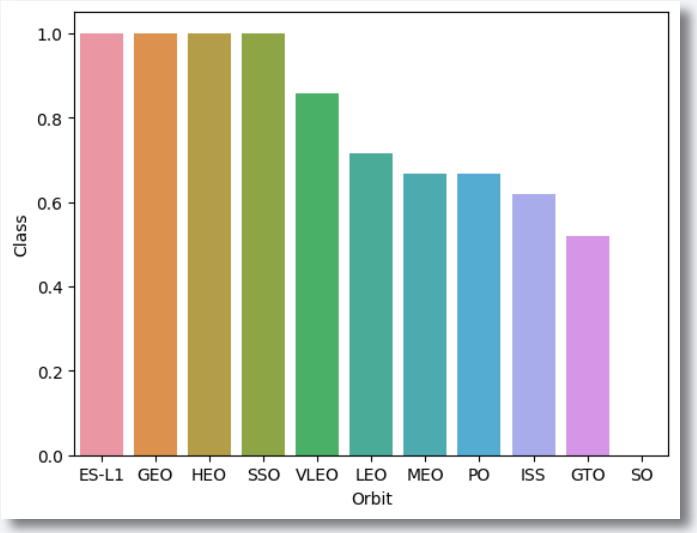


EDA with Data Visualization

- **Scatter plots**
  - Flight Number vs Payload Mass
  - Flight Number vs Launch Site
  - Flight Number vs Orbit Type
  - Payload Mass vs Launch Site
  - Payload Mass vs Orbit Type



  - To view relationship between each feature and can be useful for machine learning if a relationship exists.



- A bar chart displays comparisons of success rates among different orbits.
- Github link: https://github.com/Lapyae-Phyo/Applied_data_science_capstone/blob/master/EDA_with_Data_Visualization.ipynb

EDA with SQL

- We applied EDA with SQL to get insight from the data.The following SQL queries were performed:
  - Names of the unique launch sites in the space mission
  - Top 5 launch sites which name begin with the string 'CCA’
  - Total payload 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 outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.
- Github link: https://github.com/Lapyae-Phyo/Applied_data_science_capstone/blob/master/EDA_with_SQL.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were added to 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

  and
  - Lines are used to indicate distances between two coordinates.
- Github link: https://github.com/Lapyae-Phyo/Applied_data_science_capstone/blob/master/Launch_Site_Location_Analysis.ipynb

Build a Dashboard with Plotly Dash

- We build an interactive dashboard with Plotly Dash
- We plotted pie chart showing total success launches by sites
- We plotted scatter plot showing the relationship between payload and success for all sites by different booster version.
- Github link: https://github.com/Lapyae-Phyo/Applied_data_science_capstone/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- We build four different classification models to predict the first stage will land or not from the following steps:
  - Create NumPy array from the Class column
  - Standardize the data with StandardScaler
  - Split the data using train_test_split
  - Create a GridSearchCV object with cv=10 for parameter optimization
  - Apply GridSearchCV on different algorithms: Logistic Regression, Support Vector Machine, Decision Tree Classifier and K-Nearest Neighbor Classifier.
  - Calculate accuracy on the test data using .score() for all models
  - Assess the confusion matrix for all models
  - Compare the results and identify the best model
- Github link: https://github.com/Lapyae-Phyo/Applied_data_science_capstone/blob/master/SpaceX_Machine_Learning_Prediction.ipynb

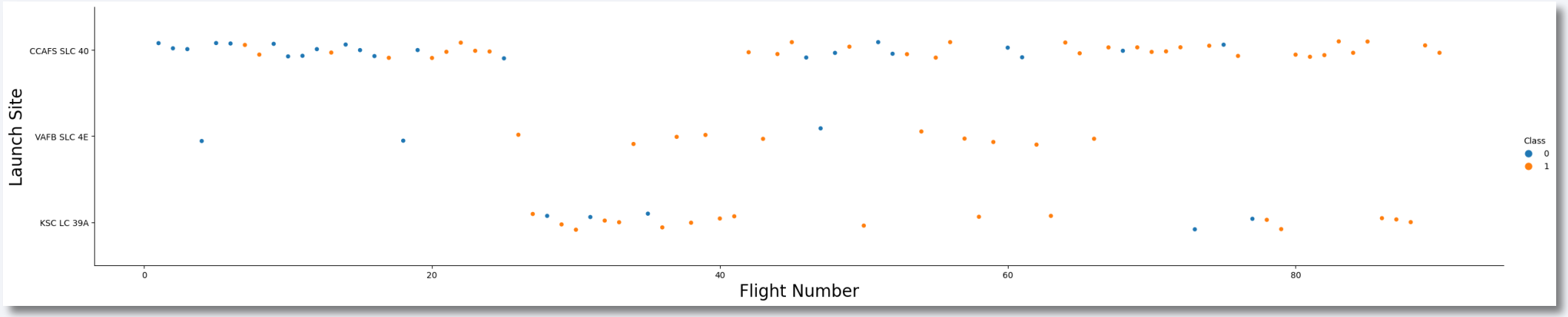
Results Summary

- **Exploratory data analysis results**
  - Launch success has improved over time
  - KSC LC-39A has the highest success rate among landing sites
  - Orbits ES-L1, GEO, HEO and SSO have a 100% success rate
- **Interactive analytics results**
  - Most launch sites are near the equator, and all are close to the coast
  - Launch sites are far enough away from anything a failed launch can damage (city), while still close enough to bring people and material to support launch activities (highway, railway)
- **Predictive analysis results**
  - Decision Tree model is the best predictive model for the dataset

Section 2

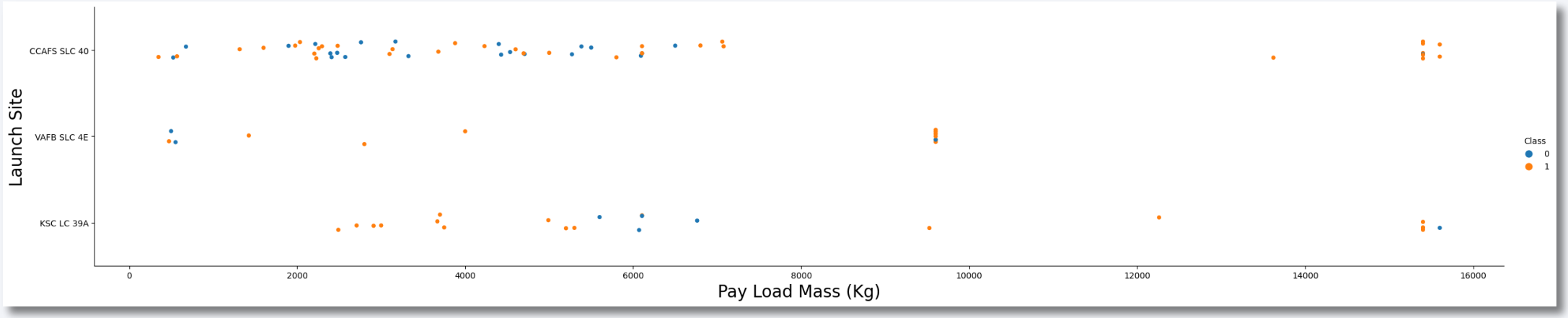
Insights drawn from EDA

Flight Number vs. Launch Site



- Earlier flights had a lower success rate (blue = fail)
- Later flights had a higher success rate (orange = success)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher success rate

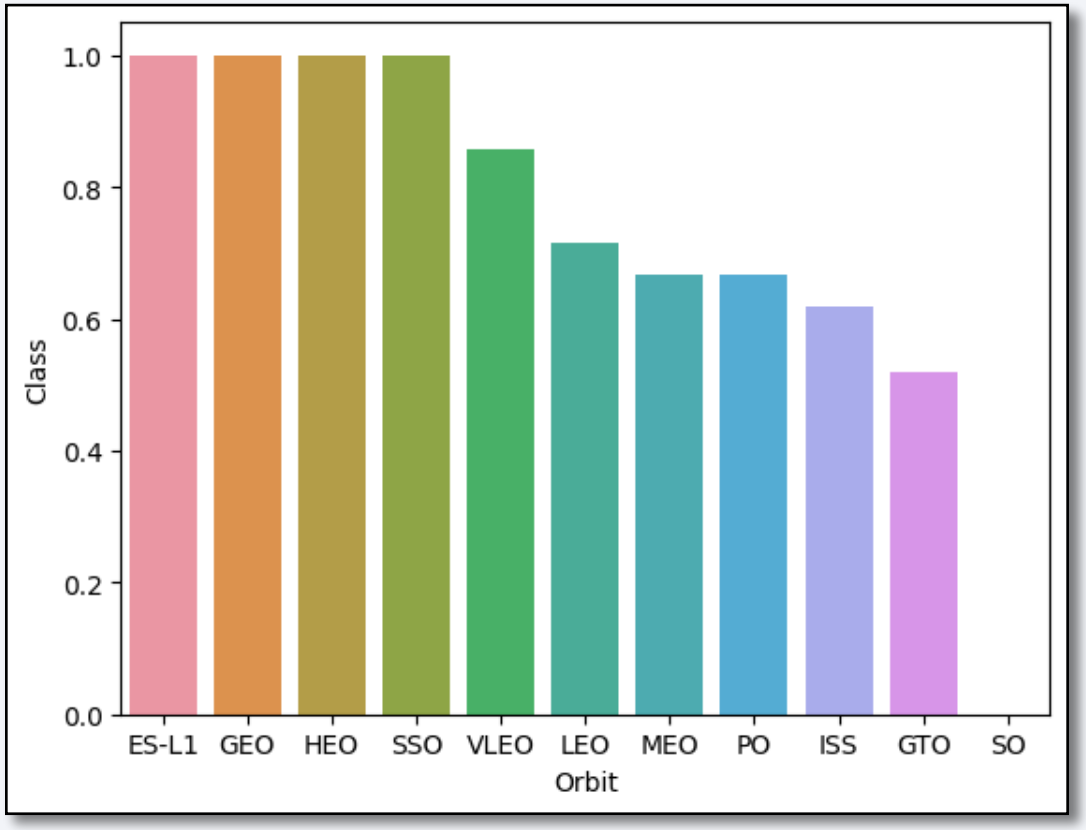
Payload vs. Launch Site



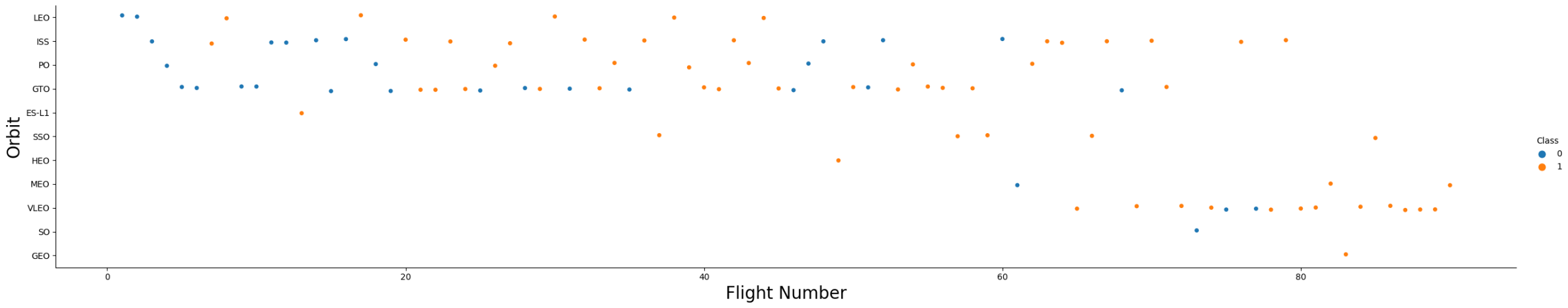
- Typically, the greater the payload mass (kg), the higher the success rate
- Most launches with a payload greater than 7,000kg were successful
- KSC LC39A has a 100% success rate for launches less than 5,500kg

Success Rate vs. Orbit Type

- From the bar chart:
  - ES-L1,GEO,HEO,SSO have 100% success rate
  - GTO, ISS, PO, MEO, LEO and VLEO have 50%-80% success rate
  - SO has 0% success rate

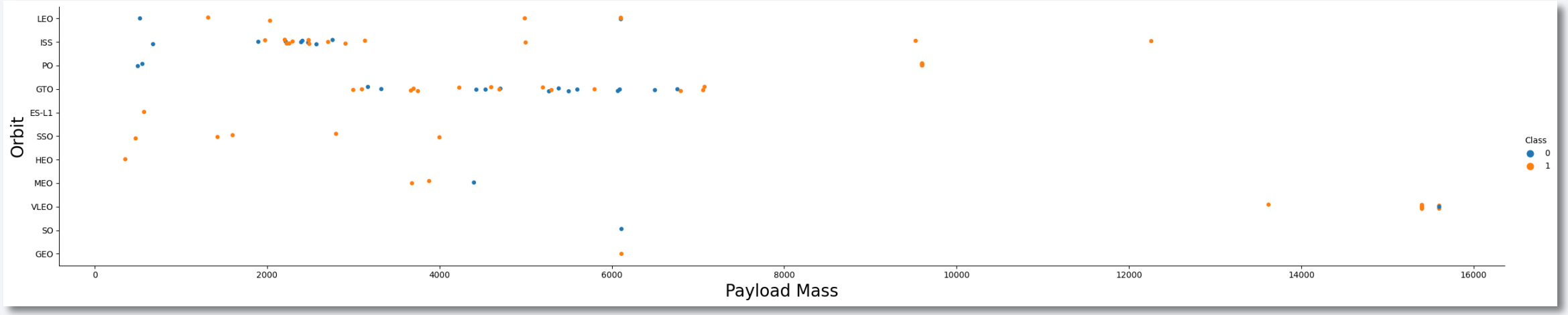


Flight Number vs. Orbit Type



- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit

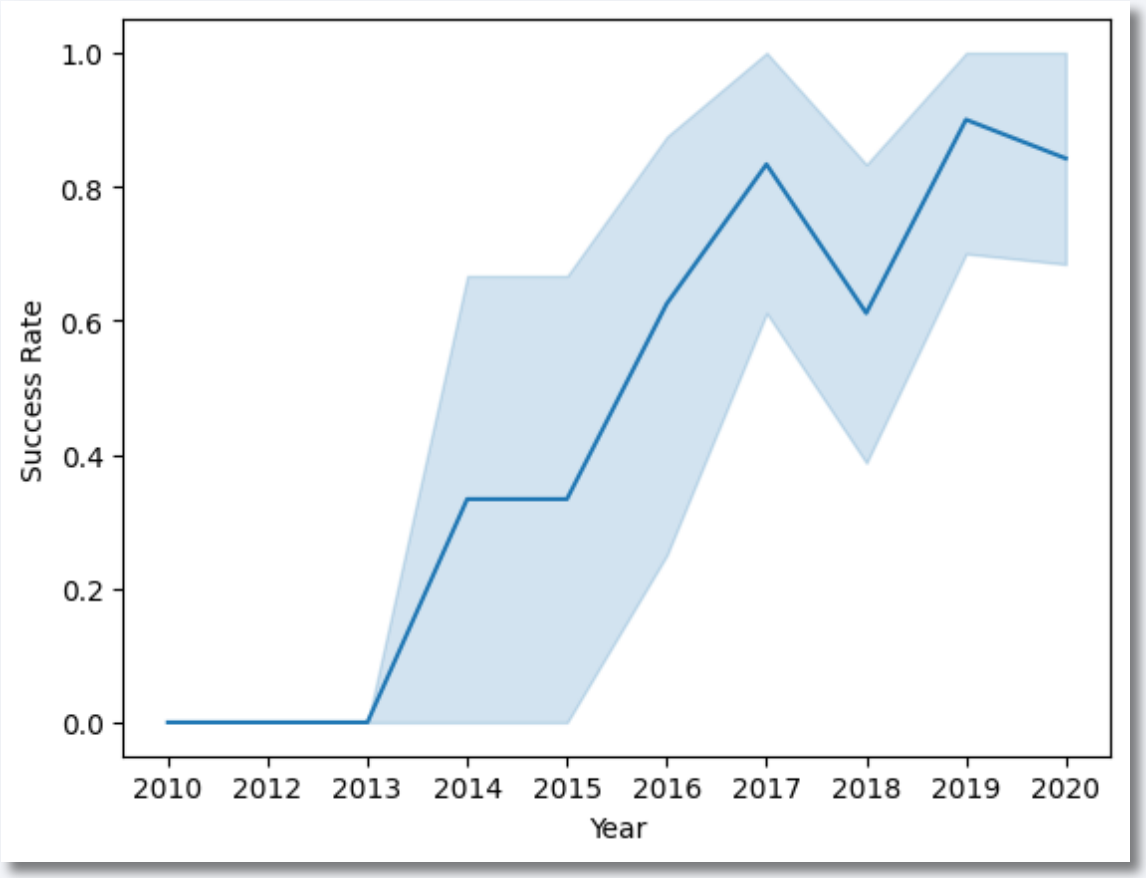
Payload vs. Orbit Type



- Payload mass between 7,000kg and 14,000kg has 100% success rate
- Heavy payloads are better with LEO, ISS and PO orbits

Launch Success Yearly Trend

- The success rate started increasing from 2013 and kept until 2020.
- It seems that the first three years were a period of adjust and improvement of technology.



All Launch Site Names

- We used the key word **DISTINCT** to show only unique launch sites from the SPACEXTABLE

```
%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE
```

executed in 4ms, finished 23:17:27 2024-03-15

```
 * 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'

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

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

executed in 4ms, finished 23:17:27 2024-03-15

```
 * 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

- We used the query above to display 5 records where launch sites begin with `CCA`
- We used the Keyword **WHERE** to filter the results with **LIKE `CCA`** and **LIMIT 5** to display only 5 records.

Total Payload Mass

- We calculated the total payload carried by boosters from NASA as 45596 using the query below.
- Used **SUM** keyword to calculate the total payload

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

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

executed in 4ms, finished 23:17:27 2024-03-15

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

TotaPayloadMass
45596

Average Payload Mass by F9 v1.1

- We calculated the average payload mass carried by booster version F9 v1.1 as 2928.4 from the query below.
- We used **AVG** keyword to average the payload mass

Display average payload mass carried by booster version F9 v1.1

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

executed in 4ms, finished 23:17:27 2024-03-15

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

AveragePayloadMass
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
- We used **MIN** keyword to find the earliest date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome LIKE '%ground pad%'
```

executed in 4ms, finished 23:17:27 2024-03-15

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- We used the **WHERE** clause to filter for boosters which have successfully landed on drone ship and applied the **AND** condition to determine successful landing with payload mass **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

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

executed in 12ms, finished 15:49:03 2024-04-03

```
 * 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

- We Calculated the total number of successful and failure mission outcomes
- We used **GROUP BY** for grouping the mission outcome and **COUNT** for counting the records for each group.

List the total number of successful and failure mission outcomes

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

executed in 3ms, finished 23:17:27 2024-03-15

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

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

Boosters Carried Maximum Payload

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

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

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

executed in 4ms, finished 23:17:27 2024-03-15

```
 * 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

- We used a combinations of the **WHERE** clause, **LIKE** and **AND** to filter the failed landing outcomes in drone ship
- SQLLite does not support monthnames. So we need to use **substr(Date, 6,2)** as month to get the months and **substr(Date,0,5)='2015'** for year.

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, Landing_Outcome, Booster_Version, Launch_Site
FROM SPACEXTABLE
WHERE substr(Date, 0,5)='2015'AND LOWER(Landing_Outcome) LIKE '%failure%' ;
```

executed in 3ms, finished 23:17:27 2024-03-15

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

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

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 2017-03-20.
- 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.

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 FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Count DESC
```

executed in 4ms, finished 23:17:27 2024-03-15

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

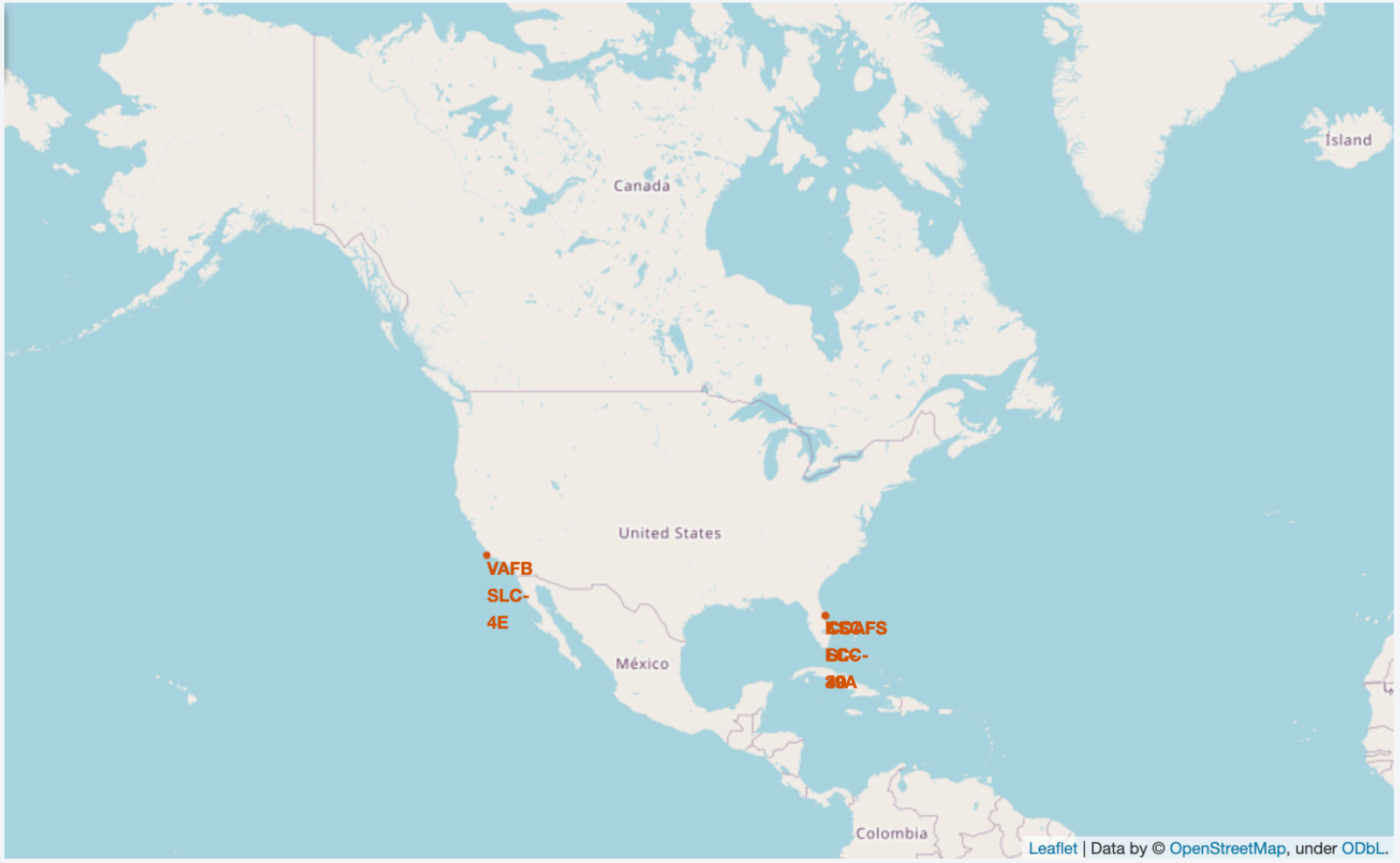
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

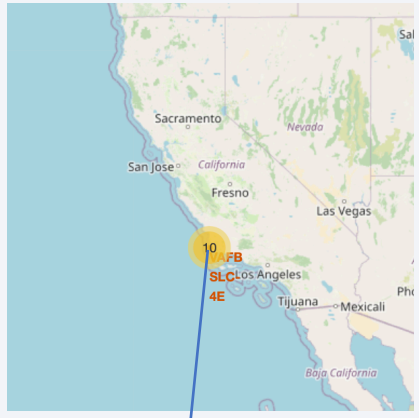
All Launch Sites

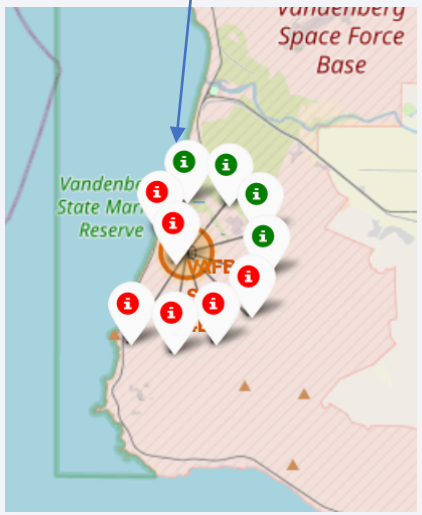
- All launch sites are in the United States of America coasts and near to equator.



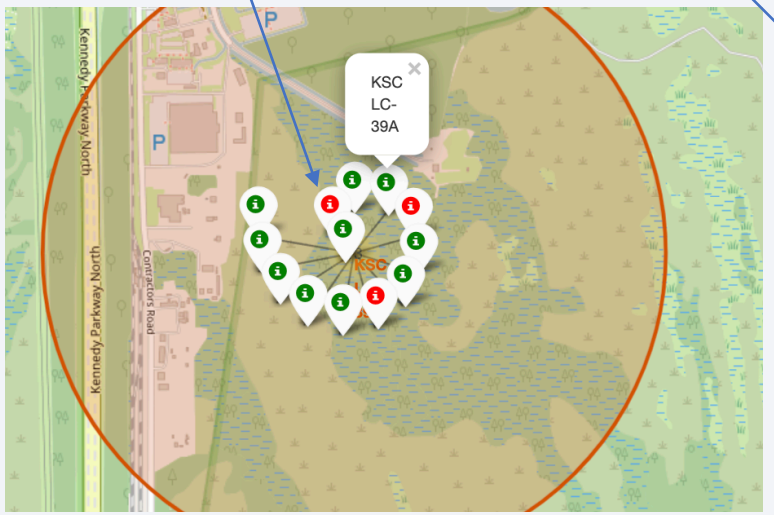
Launch Outcome By Site

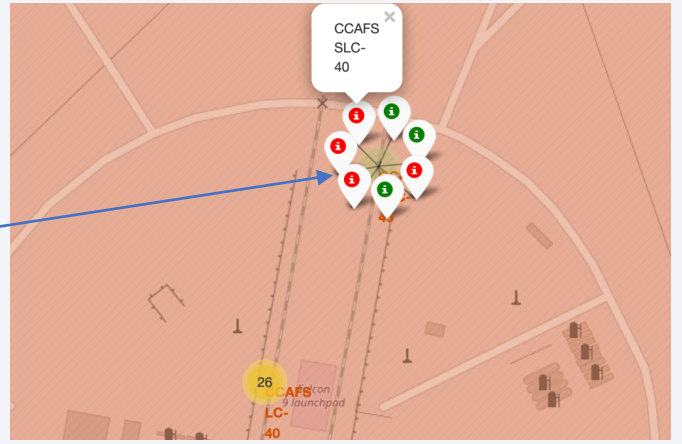
- Green markers indicate successful and red ones indicate failure.

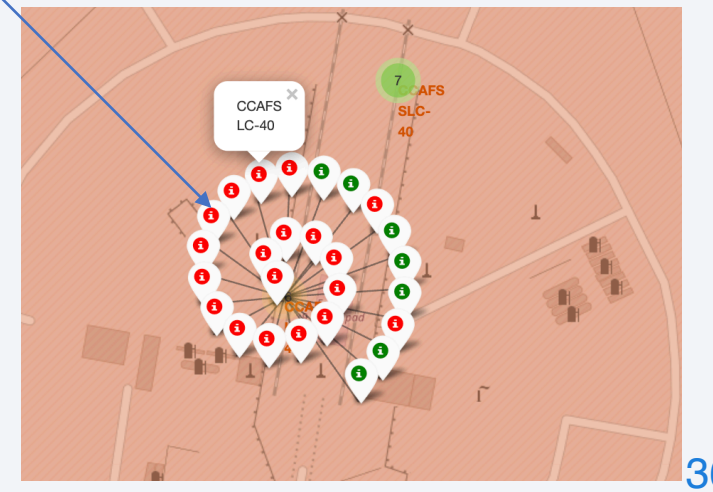




California Launch Sites



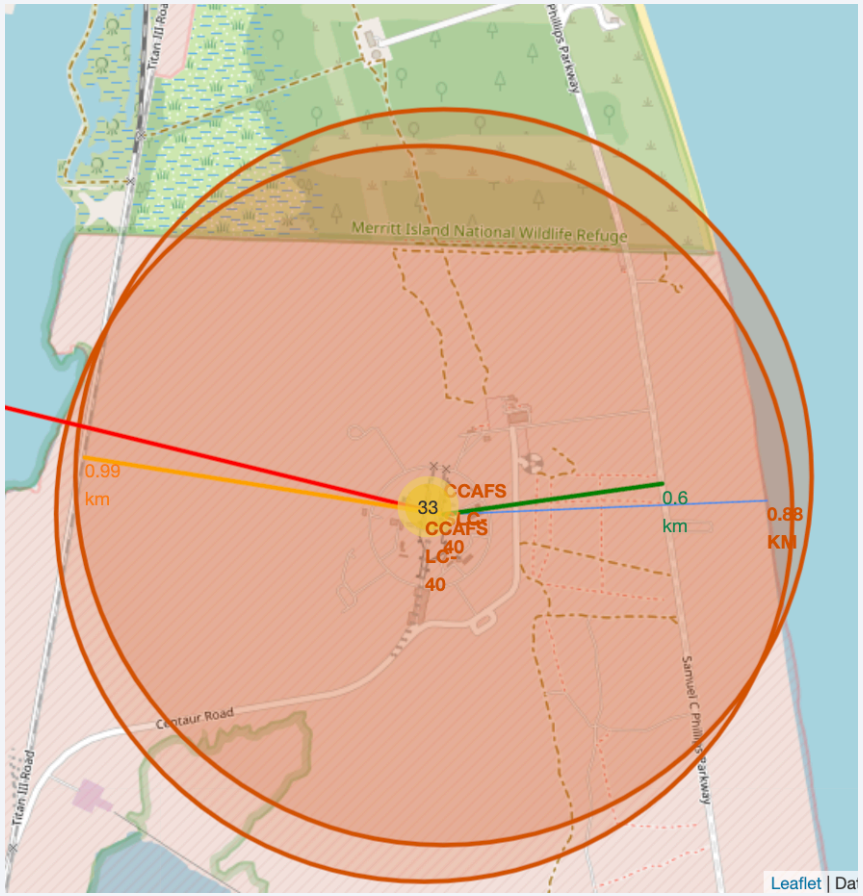


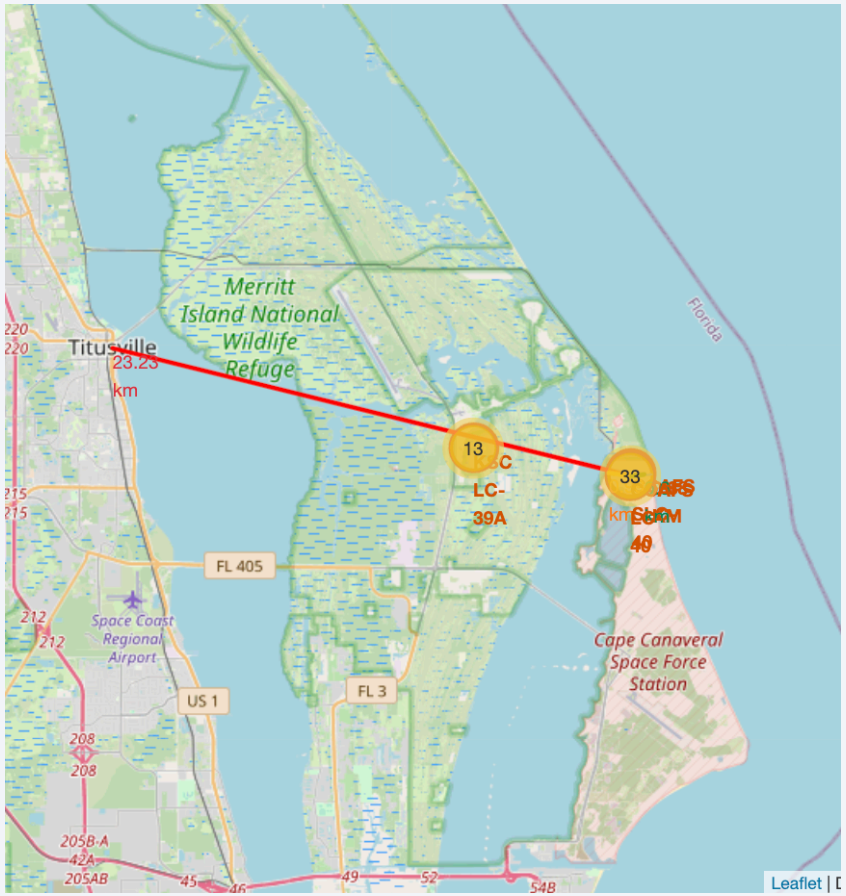


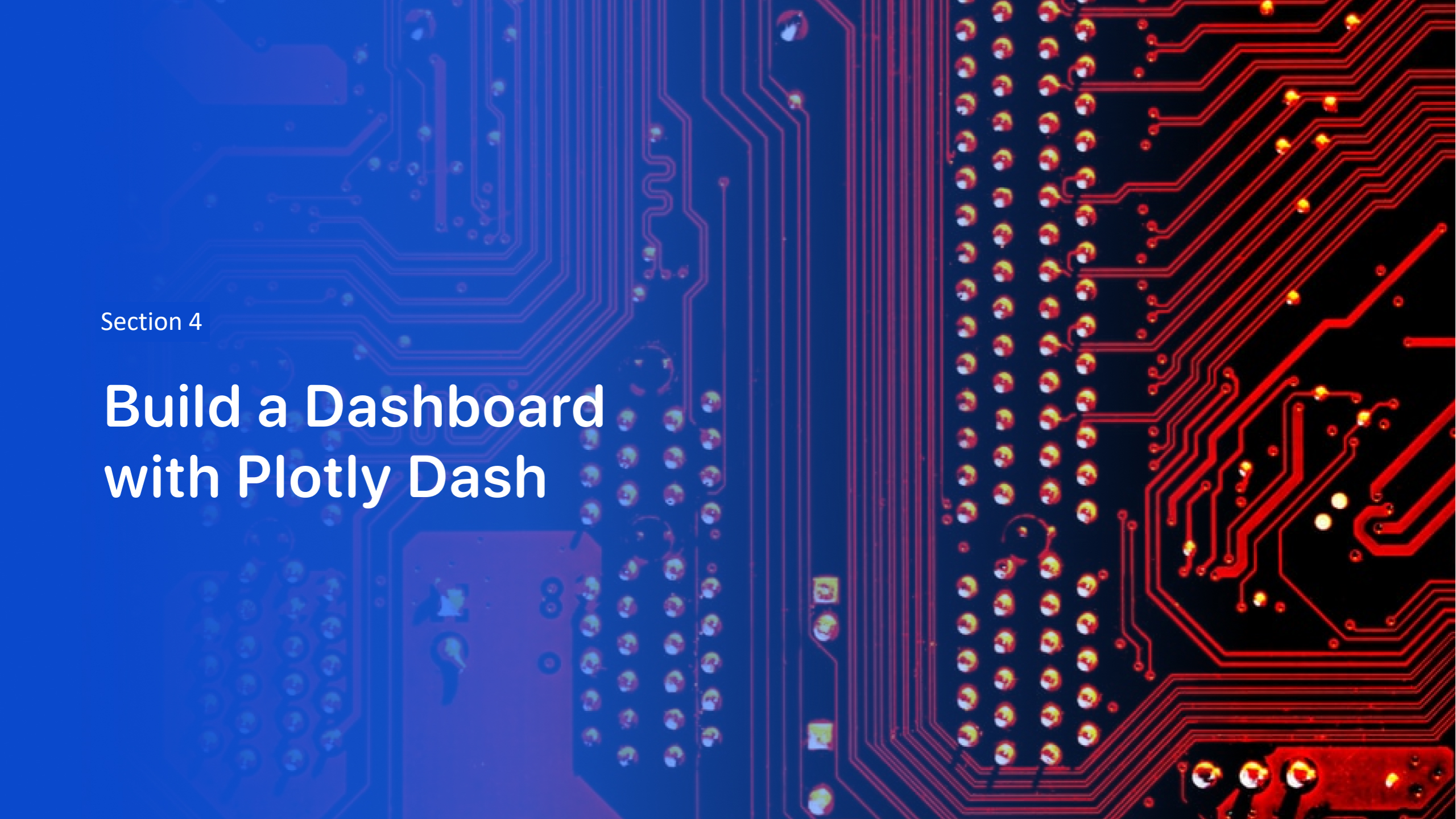
Florida Launch Sites

Launch Site Distance to Landmarks

- Launch site CCAFS SLC40 has good logistics aspects, being near coast, railroad and highway and relatively far from inhabited areas.

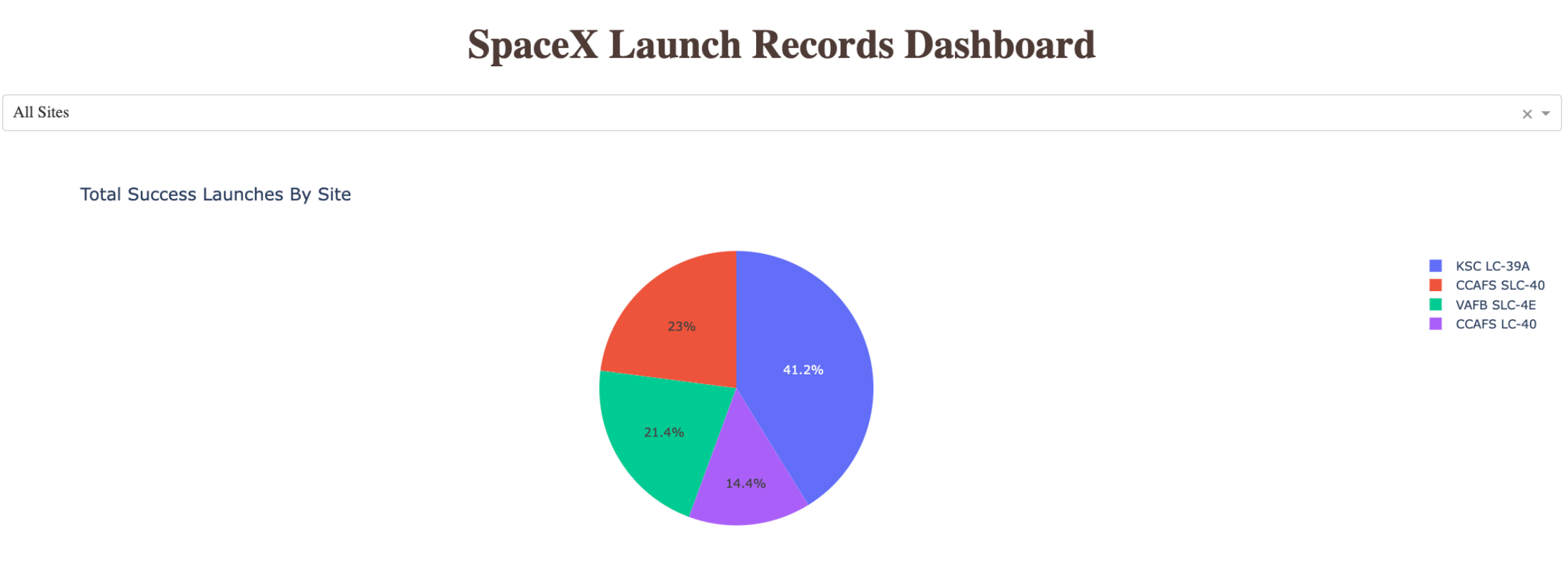





Section 4
Build a Dashboard with Plotly Dash

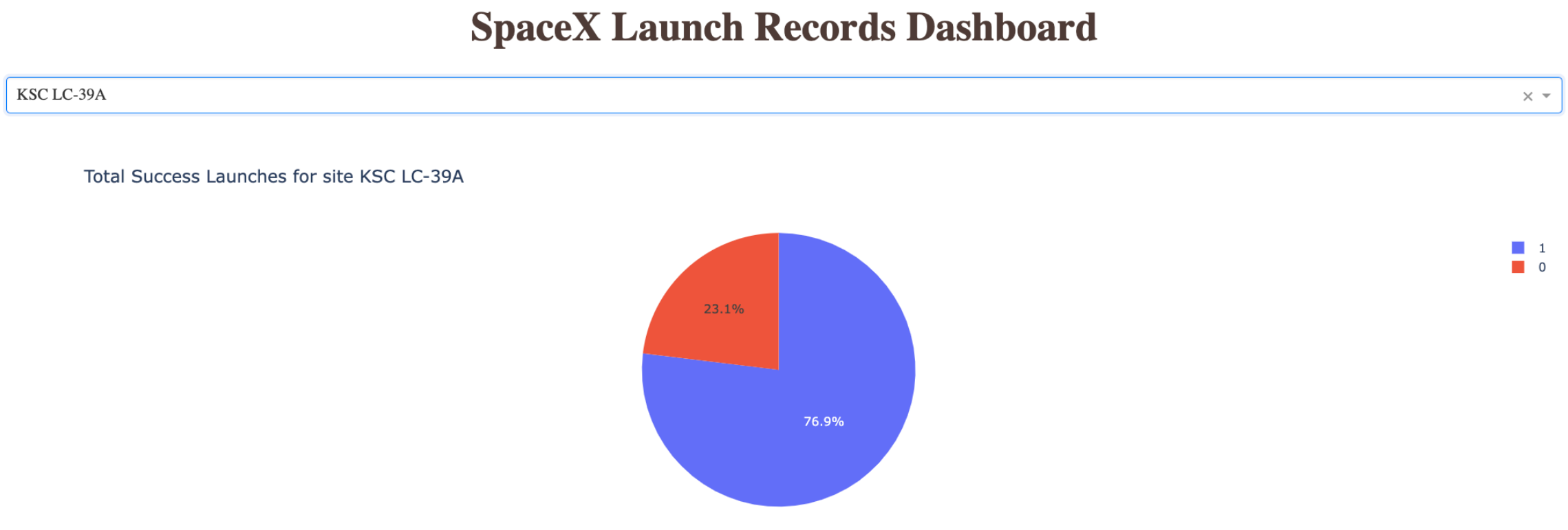
Successful Launches by Site

- KSC LC-39A has the most successful launches amongst launch sites (41.2%)



Launch Success Ratio for KSCLC-39A

- KSC LC-39A has the highest success rate amongst launch sites (76.9%) while getting a 23.1% failure rate.



Payload vs. Launch Outcome

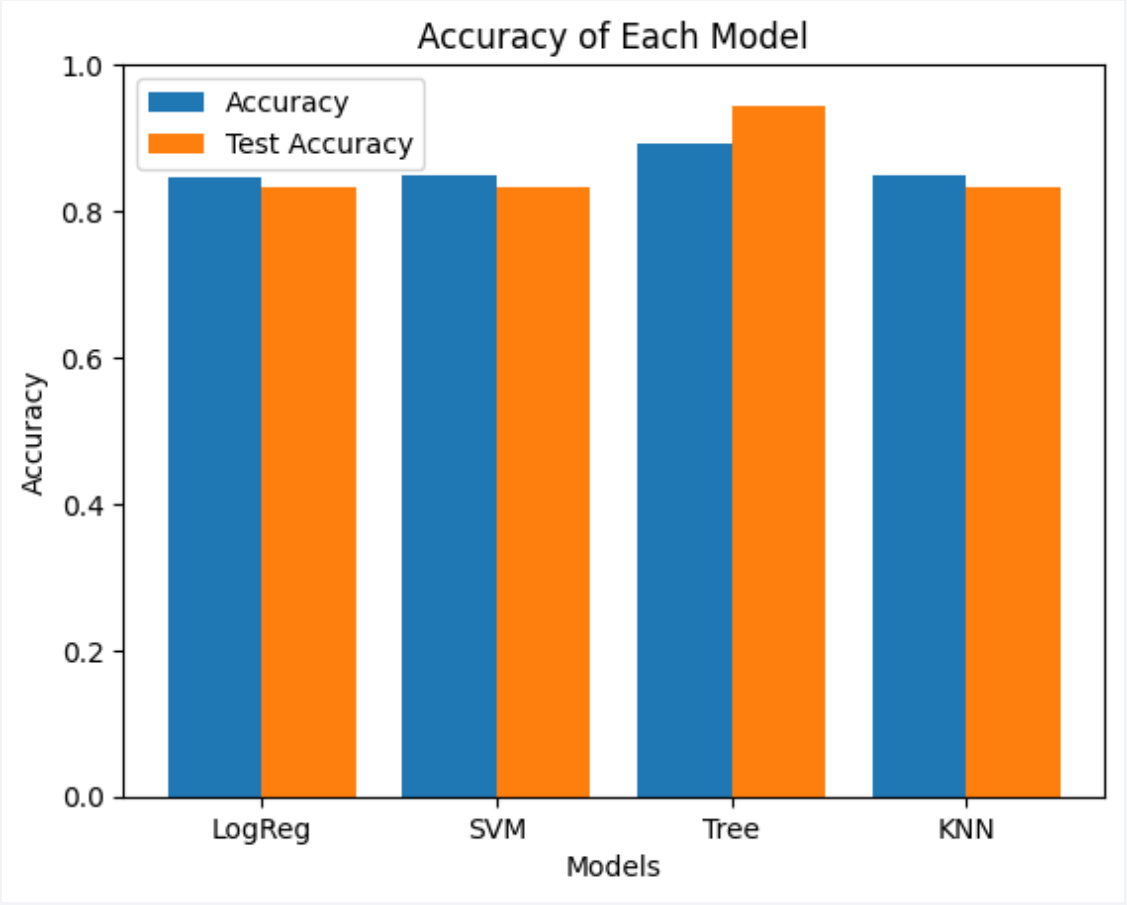
- Payloads between 2,000 kg and 5,000 kg have the highest success rate
- 1 indicating successful outcome and 0 indicating an unsuccessful outcome

Section 5

Predictive Analysis (Classification)

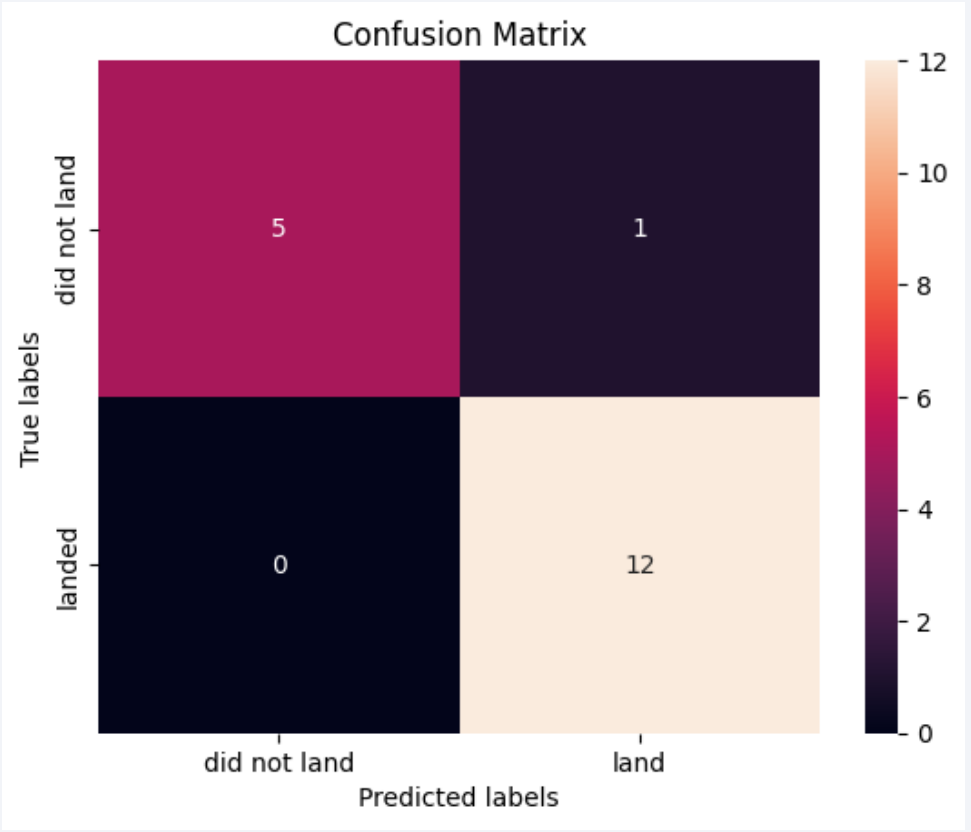
Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over 87%.



Confusion Matrix

- A confusion matrix summarizes the performance of a classification algorithm
- All the confusion matrices were identical
- The fact that there are false positives (Type 1 error) is not good
- Confusion Matrix Outputs:
  - 12 True positive
  - 5 True negative
  - 1 False positive
  - 0 False Negative
- Precision = TP / (TP + FP)
  - 12 / 15 = .80
- Recall = TP / (TP + FN)
  - 12 / 12 = 1
- F1 Score = 2 * (Precision * Recall) / (Precision + Recall)
  - 2 * (.8 * 1) / (.8 + 1) = .89
- Accuracy = (TP + TN) / (TP + TN + FP + FN) = .833



Conclusions

- **Model Performance**: The models performed similarly on the test set with the decision tree model slightly outperforming
- **Equator**: Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters
- **Coast**: All the launch sites are close to the coast
- **Launch Success**: Increases over time
- **KSC LC-39A**: Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- **Orbits**: ES-L1, GEO, HEO, and SSO have a 100% success rate
- **Payload Mass**: Across all launch sites, the higher the payload mass (kg), the higher the success rate

Thank you!