IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dhava Gautama
28 July 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API and Web Scrapping
  - Data Wrangling
  - Explanatory Data Analysis with SQL
  - Explanatory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Explanatory Data Analysis Result
  - Interactive analytics screenshot
  - Predictive Analytics result from Machine Learning Lab

Introduction

SpaceX is a company that has revolutionized the space industry by offering Falcon 9 rocket launches for as low as 62 million dollars, while other providers charge upward of 165 million dollars per launch. The company has been able to achieve this by reusing the first stage of the rocket, which has helped reduce the cost of launches. The first stage of the rocket is re-landed and then used on the next mission, which further reduces the cost of launches. As a data scientist working for a startup that is competing with SpaceX, your goal is to create a machine learning pipeline that can predict the landing outcome of the first stage of the rocket in the future. This project is crucial in identifying the right price to bid against SpaceX for a rocket launch.

Key problem included in this project are :

- Identifying factors that influence the landing outcome,
- Find relationship between variable and it effect on landing outcome,
- Find the best condition that needed to increase success landing rate.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from SpaceX API and web scrapping
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
- 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, tune, evaluate classification models

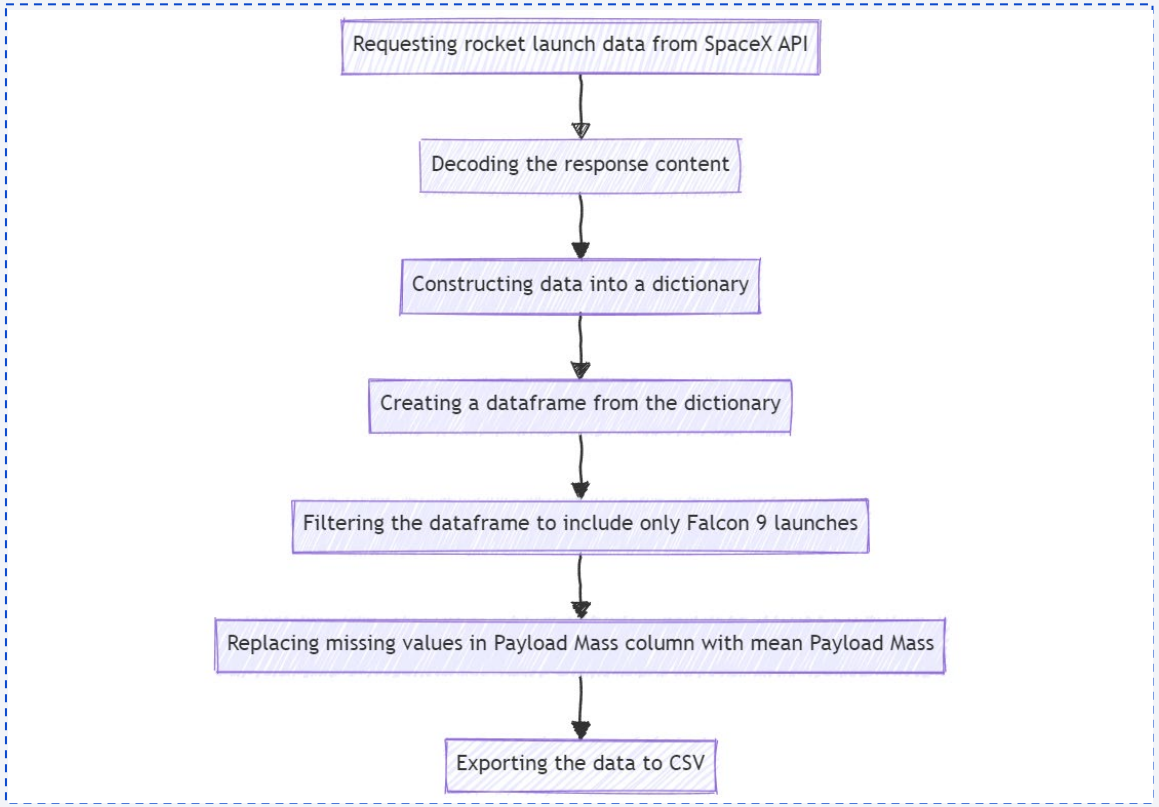
Data Collection

Data collection process combined two key methods: API requests from the SpaceX REST API and web scraping from SpaceX's Wikipedia page. Utilizing both approaches allowed us to gather comprehensive information about SpaceX launches, facilitating a detailed analysis.

This dual-method approach ensured we captured all relevant details for our analysis of SpaceX launches.

Data Collection – SpaceX API

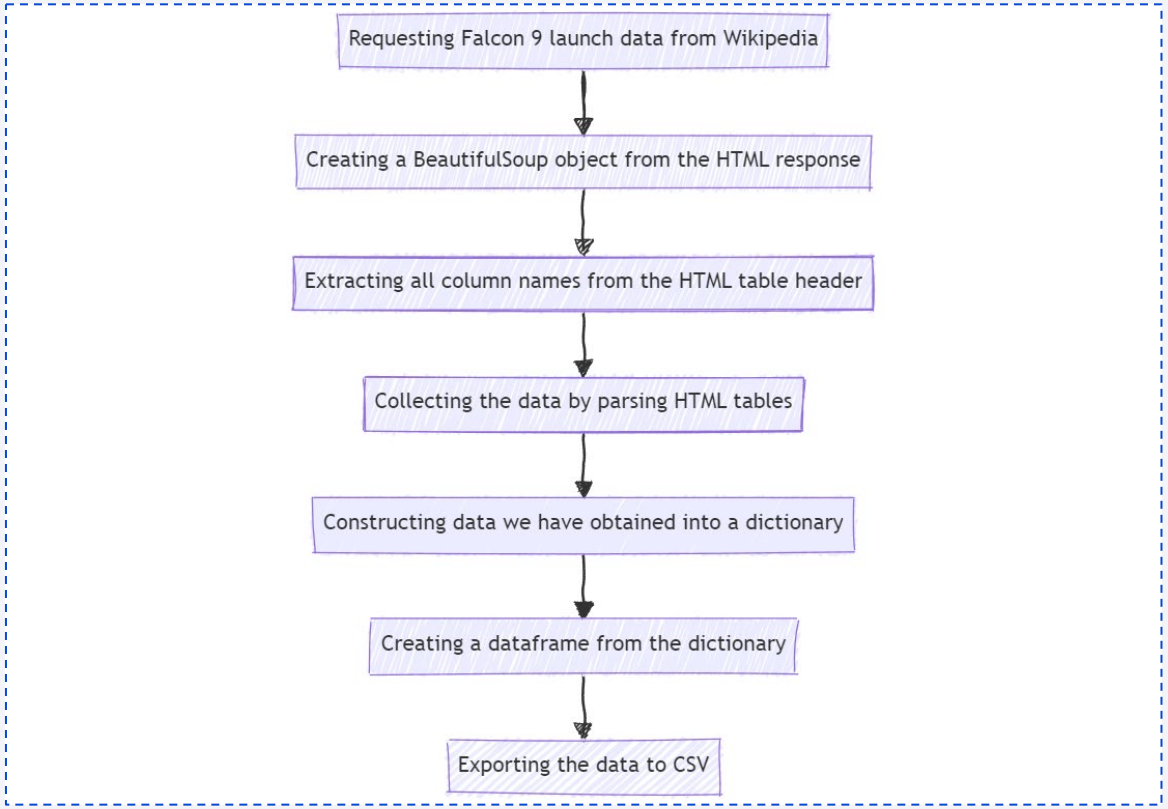
I collect the data from the SPACEX API then filter the data to include only Falcon 9 launches, then the missing values in Payload Mass column is replaced with the respective mean value of the column



Data Collection – SpaceX API Github

Data Collection - Scraping

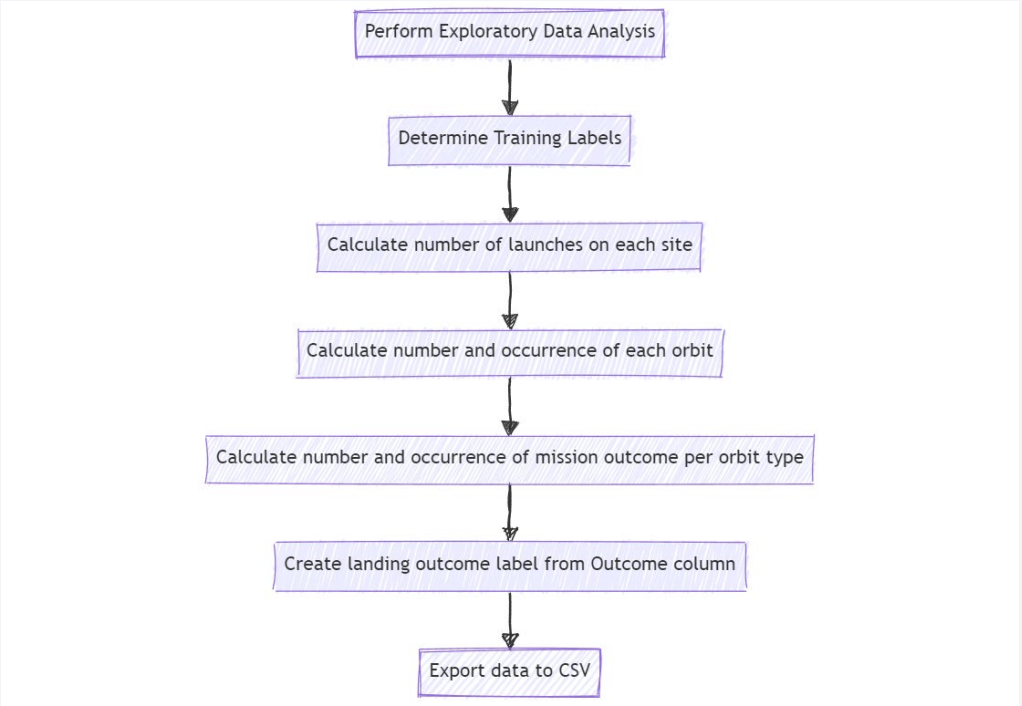
The data scrapped from the Wikipedia page then extracted by using BeautifulSoup, then the extracted data is constructed to be a dataframe.



Data Collection – Scraping Github

Data Wrangling

- Data Exploration: The data wrangling process began with performing exploratory data analysis to gain insights into the data structure and content.
- Label Determination: Next, training labels were determined to prepare the data for modeling.
- Feature Engineering: The data was then processed to extract relevant features
- Data Transformation: A new feature, landing outcome label, was created from the existing Outcome column.
- Data Export: Finally, the processed data was exported to CSV for further analysis or modeling.



Data Wrangling Github

EDA with Data Visualization

Chart Plotted

- Relationship between Flight Number and Launch Site
- Relationship between Payload and Launch Site
- Relationship between success rate of each orbit type
- Relationship between FlightNumber and Orbit type
- Relationship between Payload and Orbit type
- Launch success yearly trend

EDA with Data Visualization Github

EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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

EDA with SQL Github

Build an Interactive Map with Folium

Visualizing Launch Sites and Outcomes

Launch Site Markers:

- A marker with a circle, popup label, and text label was added to denote the NASA Johnson Space Center, utilizing its precise latitude and longitude coordinates as the starting location.
- Markers with circles, popup labels, and text labels were added to identify all launch sites, leveraging their latitude and longitude coordinates to illustrate their geographical locations and proximity to the Equator and coastlines.

Launch Outcome Indicators:

- Colored markers were used to distinguish between successful (green) and failed (red) launches, employing a marker cluster to highlight launch sites with relatively high success rates.

Proximity Analysis:

- Colored lines were added to visualize the distances between a launch site (e.g., KSC LC-39A) and its surrounding features, including railways, highways, coastlines, and the closest city.

Build an Interactive Map with Folium Github

Build a Dashboard with Plotly Dash

Launch Site Selection:

- A dropdown list was integrated to facilitate the selection of a specific launch site, enabling users to focus on individual site performance.

Launch Outcome Analysis:

- A pie chart was added to provide a comprehensive view of launch outcomes, displaying the total number of successful launches across all sites and, when a specific site is selected, the proportion of successful versus failed launches for that site.

Payload Mass Filtering:

- A slider was incorporated to allow users to select a specific payload mass range, enabling the exploration of launch outcomes within defined payload parameters.

Payload Mass vs. Success Rate Correlation:

- A scatter chart was created to illustrate the relationship between payload mass and launch success rate for different booster versions, providing insights into the impact of payload mass on launch outcomes.

Build a Dashboard with Plotly Dash Github

Predictive Analysis (Classification)

Step 1: Data Preparation

- Create a NumPy array from the "Class" column in the data
- Standardize the data using StandardScaler (fit and transform)

Step 2: Data Split

- Split the data into training and testing sets using train_test_split function

Step 3: Model Selection and Hyperparameter Tuning

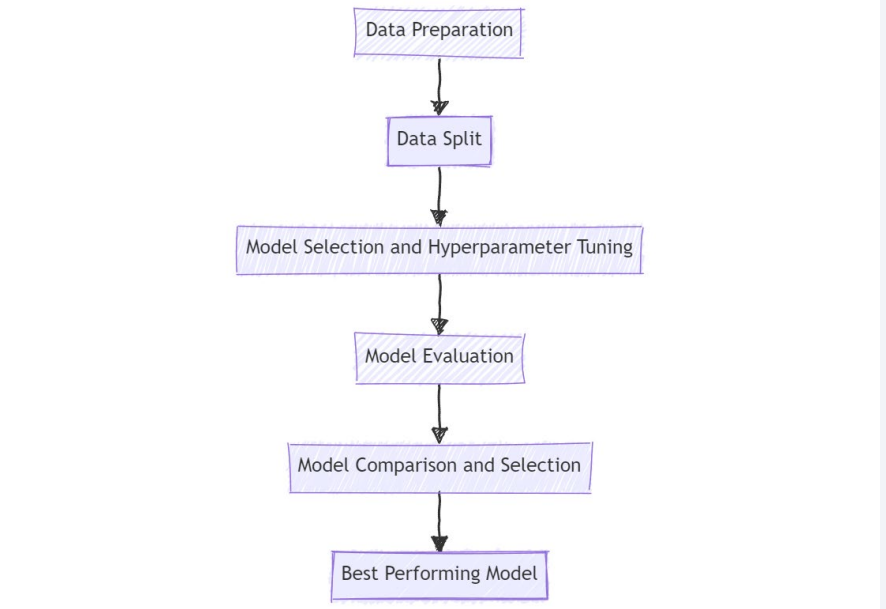
- Step 4: Model EvaluaCreate a GridSearchCV object with cv = 10 to find the best parameters
- Apply GridSearchCV to LogReg, SVM, Decision Tree, and KNN models

tion

- Calculate the accuracy on the test data using the .score() method for all models
- Examine the confusion matrix for all models

Step 5: Model Comparison and Selection

- Find the method that performs best by examining the accuracy, best scores, and F1_score metrics



Predictive Analysis (Classification) Github

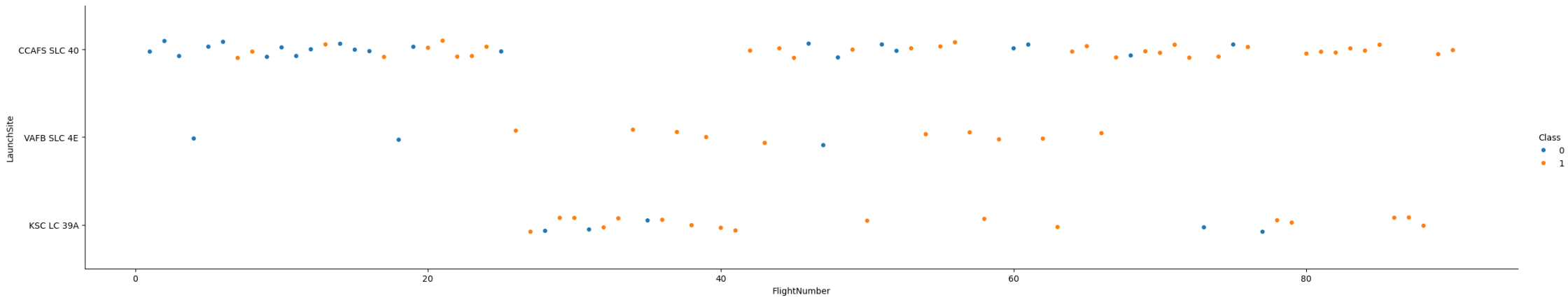
Results

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

Section 2

Insights drawn from EDA

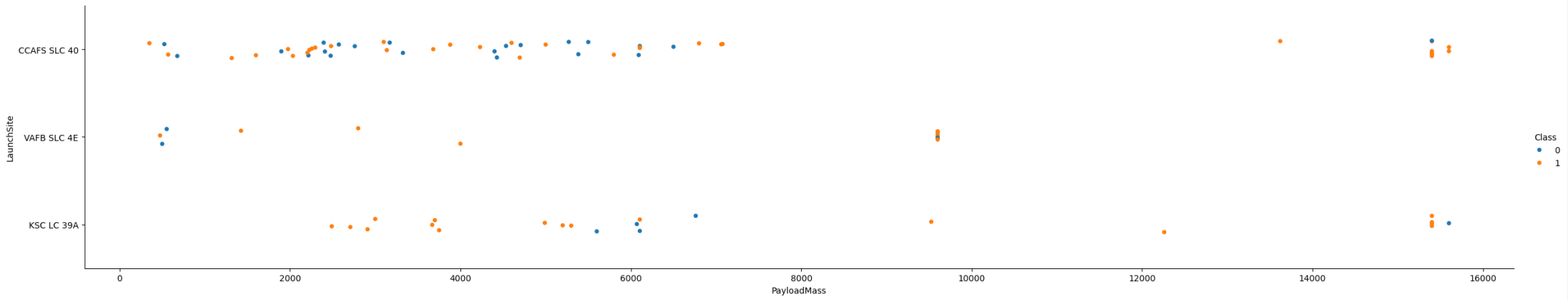
Flight Number vs. Launch Site



Trends:

- The earliest flights all failed, while the latest flights have all succeeded, suggesting an overall improvement in success rates over time.
- It can be inferred that each new launch has a higher rate of success.

Payload vs. Launch Site



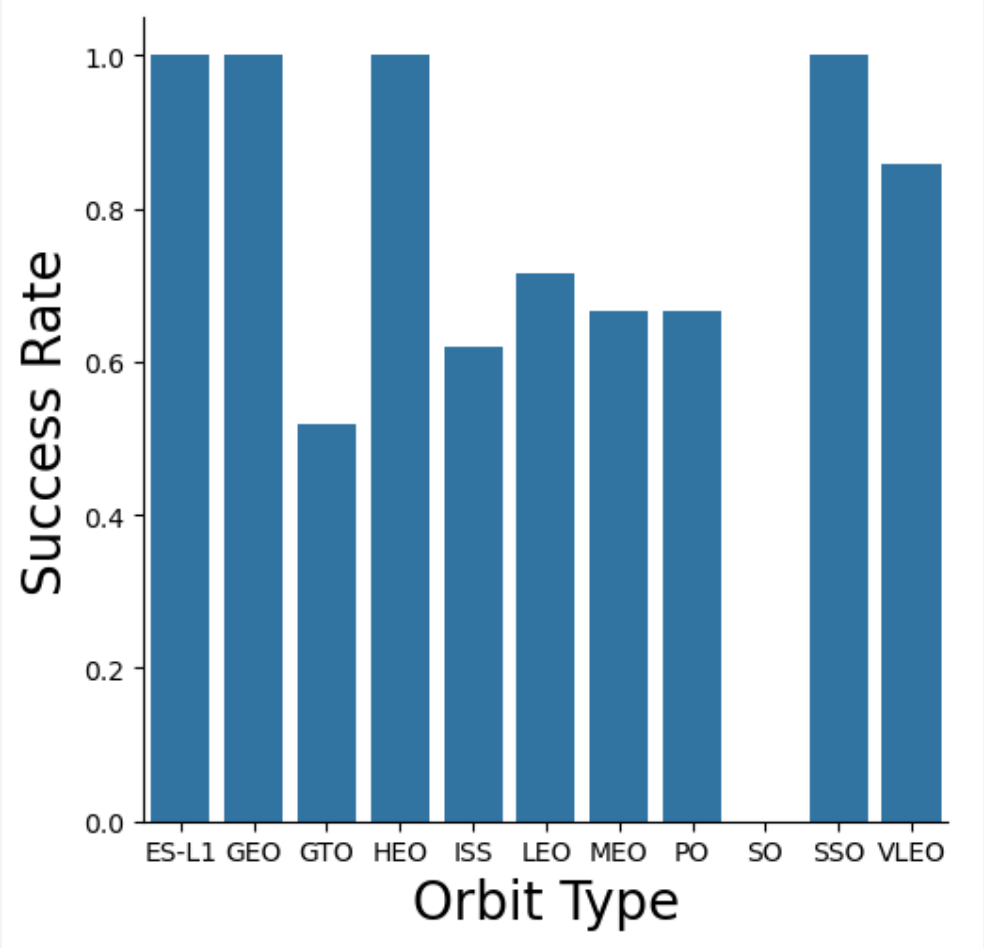
- All VAFB SLC 4E payload mass are below 10000 kg
- For payload under 5500 kg the launch site KSC LC 39A has 100 % success rate

Success Rate vs. Orbit Type

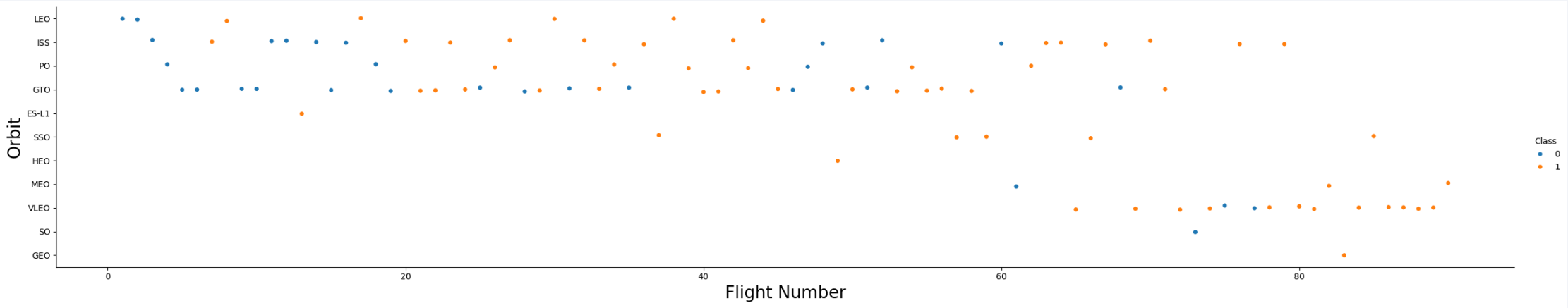
Orbit with 100% Success Rate are ES-L1, GEO, HEO, SSO

Orbit with Success Rate higher than 0% less than 100% are GTO, ISS, LEO, MEO, PO, VLEO

Orbit with 0% Success Rate is SO

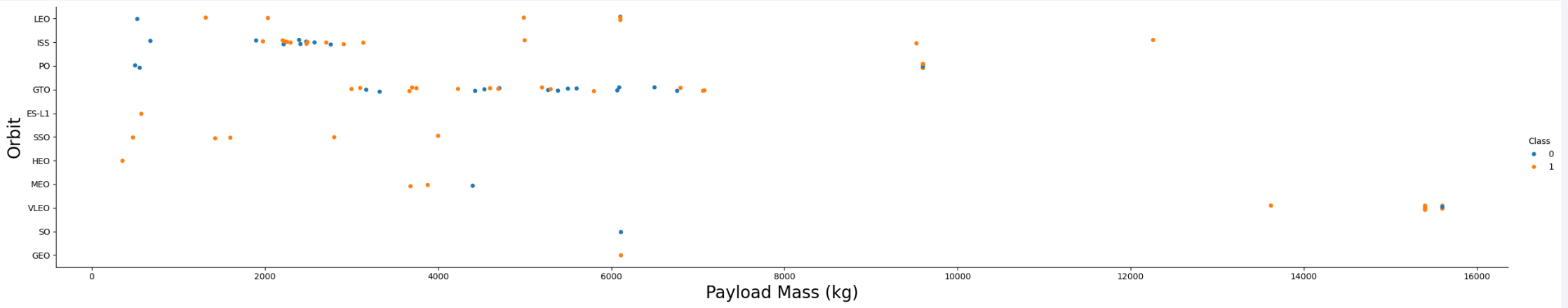


Flight Number vs. Orbit Type



LEO orbit the 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.

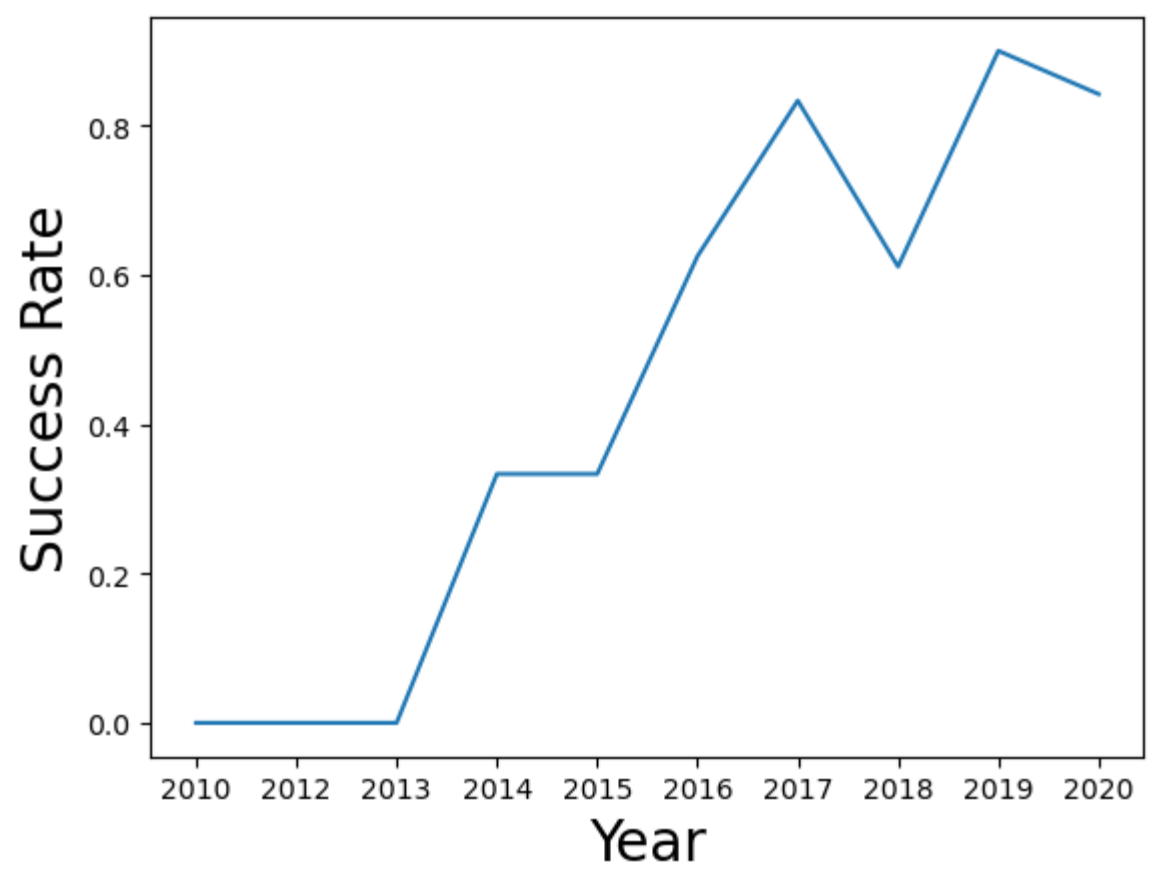
Payload vs. Orbit Type



- 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 here.

Launch Success Yearly Trend

The success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.



All Launch Site Names

- Find the names of the unique launch sites

Task 1

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

```
[21]: %sql select distinct launch_site from SPACEXTABLE;
```

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

[21]:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

Task 2

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

```
[23]: %sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
 * sqlite:///my_data1.db
Done.
```

[23]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

Task 3

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

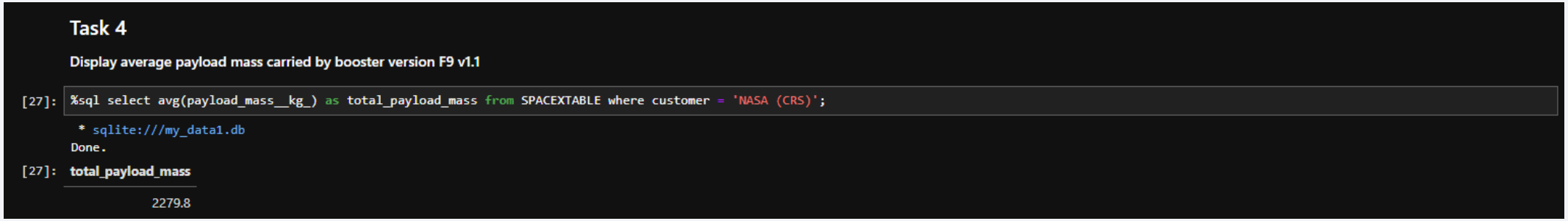
```
[25]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';
```

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

```
[25]: total_payload_mass
      45596
```

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1



Task 4

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

```
[27]: %sql select avg(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';
```

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

[27]:

total_payload_mass
2279.8

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

Task 5

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

Hint:Use min function

```
[31]: %sql select min(date) as first_successful_landing from SPACEXTABLE where landing_outcome = 'Success (ground pad)';
```

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

[31]:

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Task 6

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

```
[35]: %sql select booster_version from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

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

[35]:

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

Task 7

List the total number of successful and failure mission outcomes

```
[37]: %sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

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

[37]:

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

Boosters Carried Maximum Payload

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

Task 8

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

```
[39]: %sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);
```

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

[39]:

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

[49]:
```
%%sql
SELECT strftime('%m', date) AS month, date, booster_version, launch_site, landing_outcome
FROM SPACEXTABLE
WHERE landing_outcome = 'Failure (drone ship)' AND strftime('%Y', date) = '2015';
```

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

[49]:

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)

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

- 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

Task 10

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.

[51]:
```
%%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;
```

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

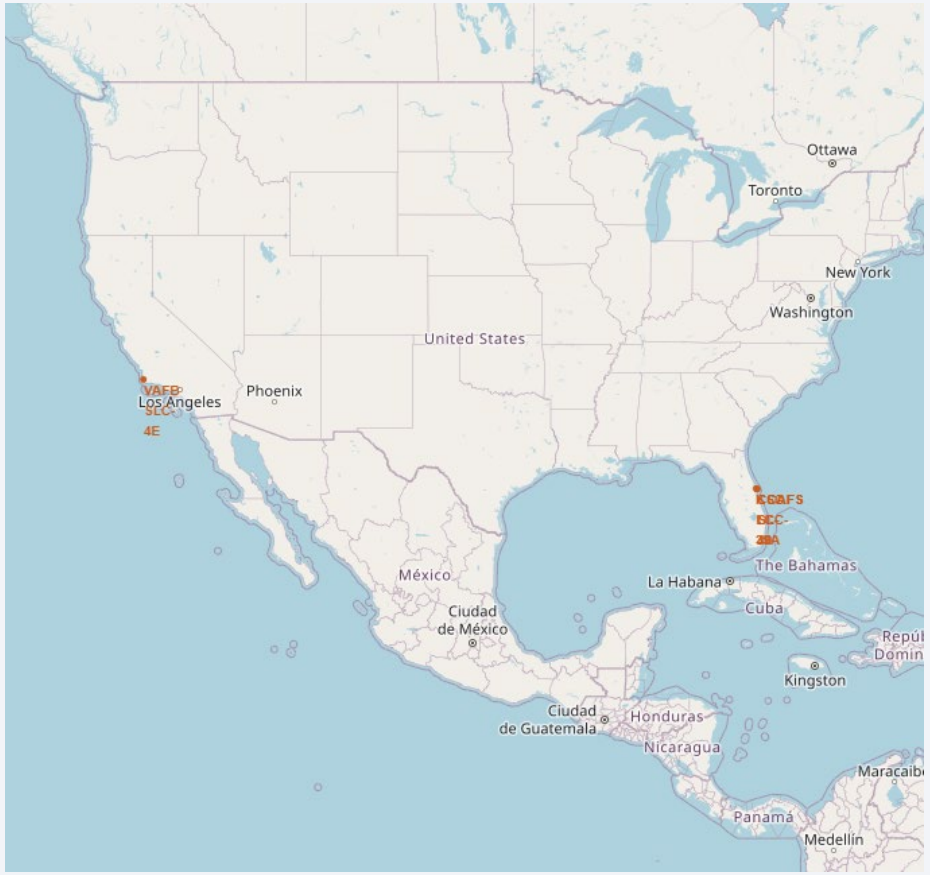
[51]:

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

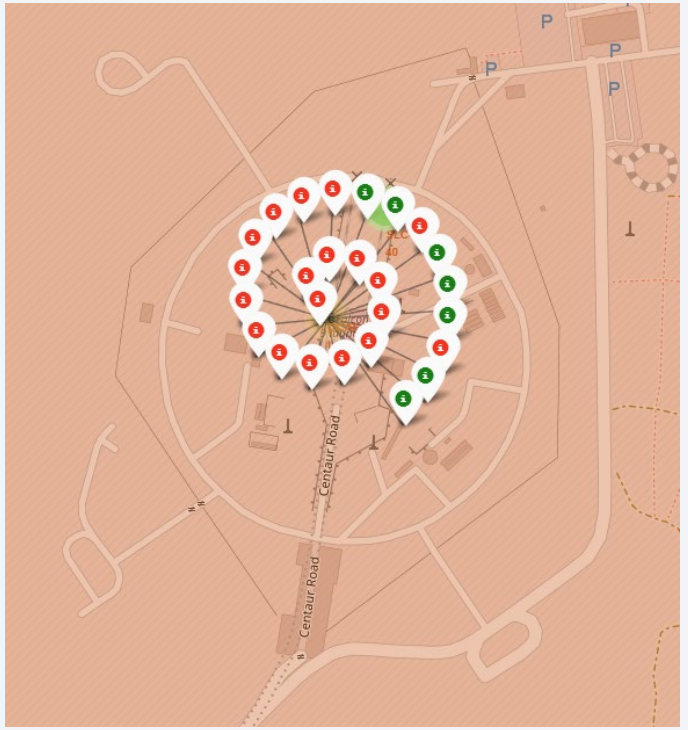
All launch sites' location markers on a global map

Most launch sites are strategically located near the Equator. This proximity provides a significant advantage, as the initial velocity of the launch site is inherited by the spacecraft due to inertia, helping it achieve the necessary speed to stay in orbit. Additionally, all launch sites are situated close to the coast, allowing rockets to be launched over the ocean, thereby minimizing the risk of debris falling or exploding near populated areas.



Colour-labeled launch records on the map

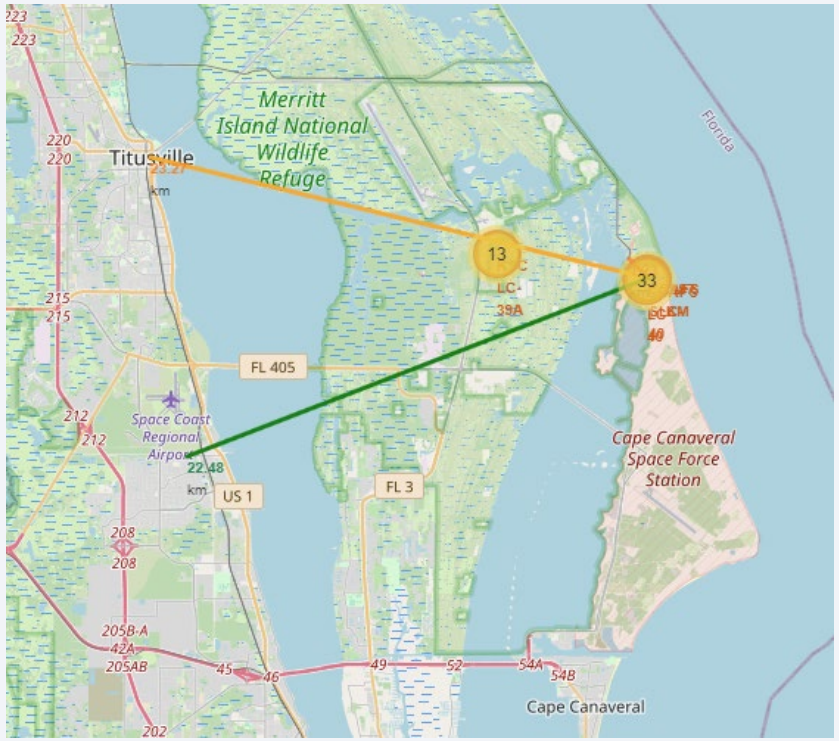
By referencing the color-coded markers, we can quickly identify the launch sites with high success rates. The green markers indicate successful launches, while the red markers denote failed launches, providing a visual representation of each site's performance.

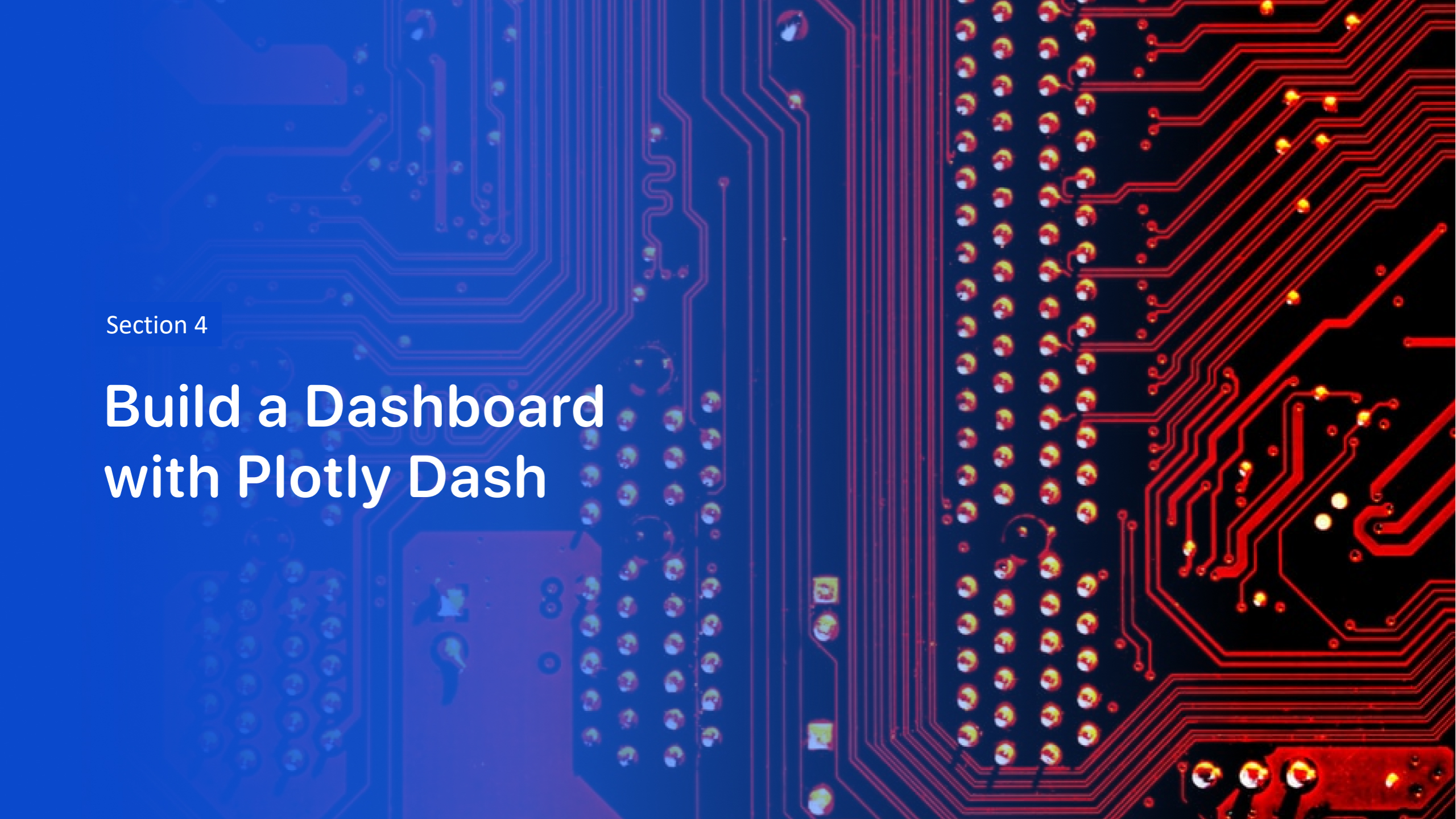


Distance from the launch site CCAFS LC-40 to its proximities

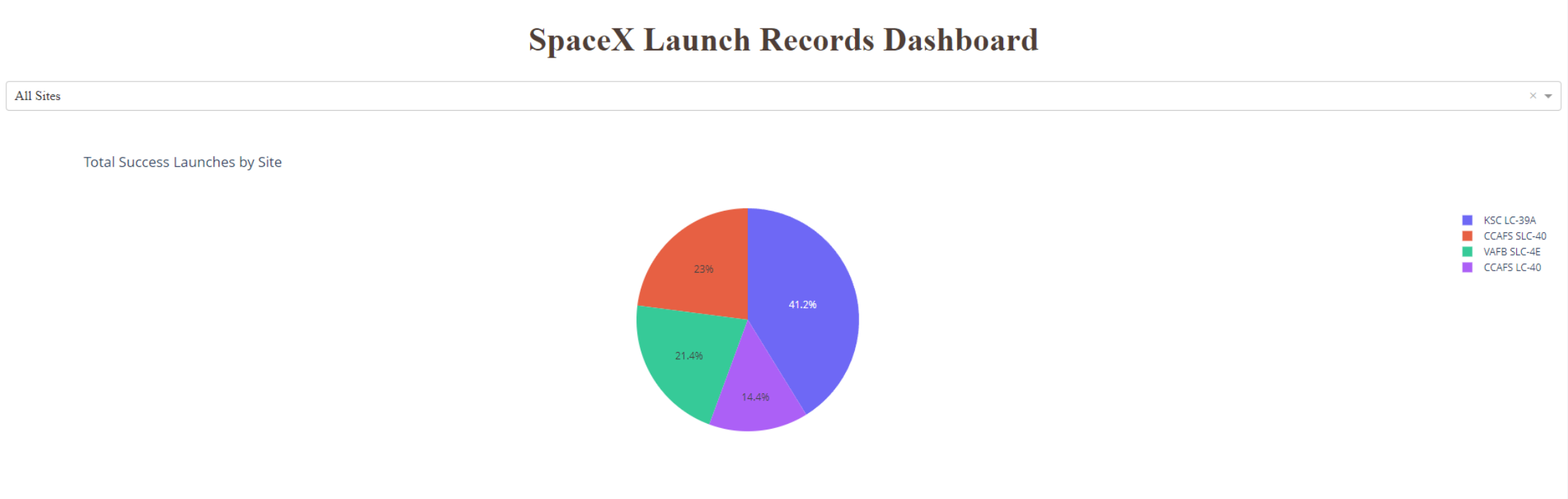
From the visual analysis of the launch site CCAFS LC-40 we can clearly see that it is:

- relative close to railway (1.23 km)
- relative close to highway (22.48 km)
- relative close to coastline (0.87 km)
- relative close to city Titusville (23.27 km).



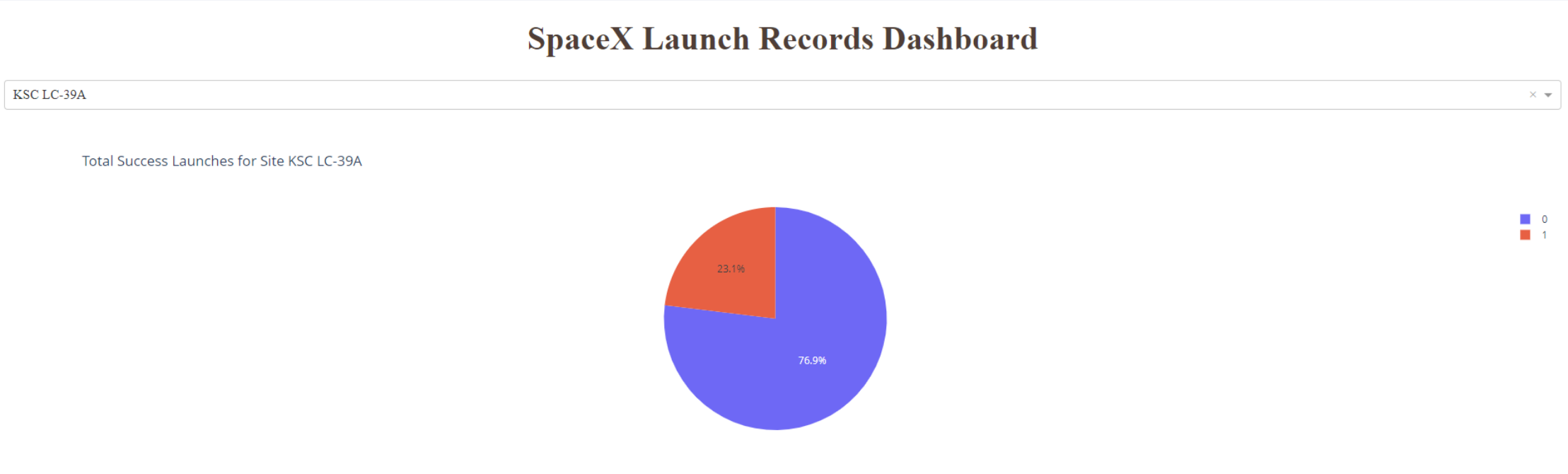

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



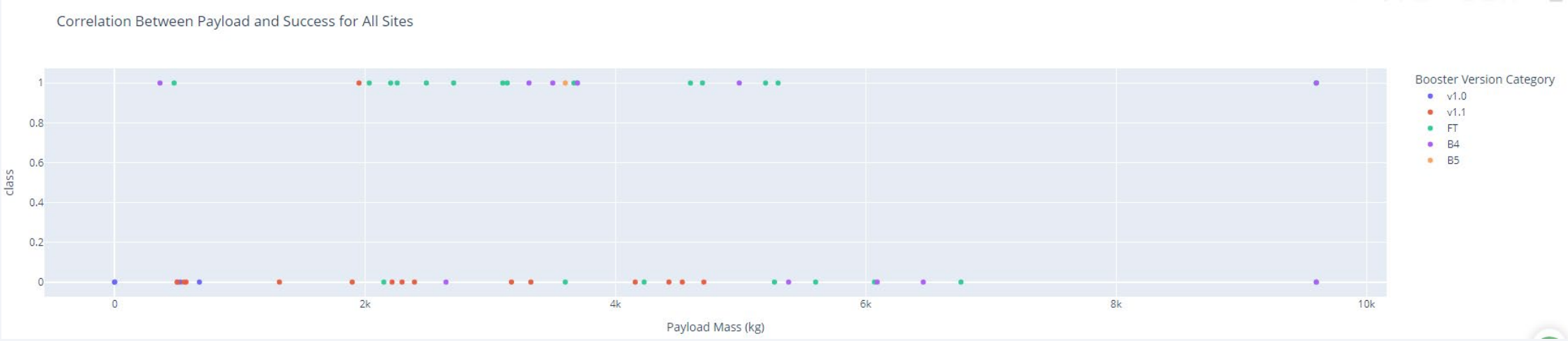
KSC LC-39A have the highest total success launches followed by CCAFS SLC-40 then VAFB SLC-4E, and CCAFS LC-40

Launch site with highest launch success ratio



KSC LC-39A has the highest launch success rate (76.9%)

Payload vs. Launch Outcome scatter plot for all sites



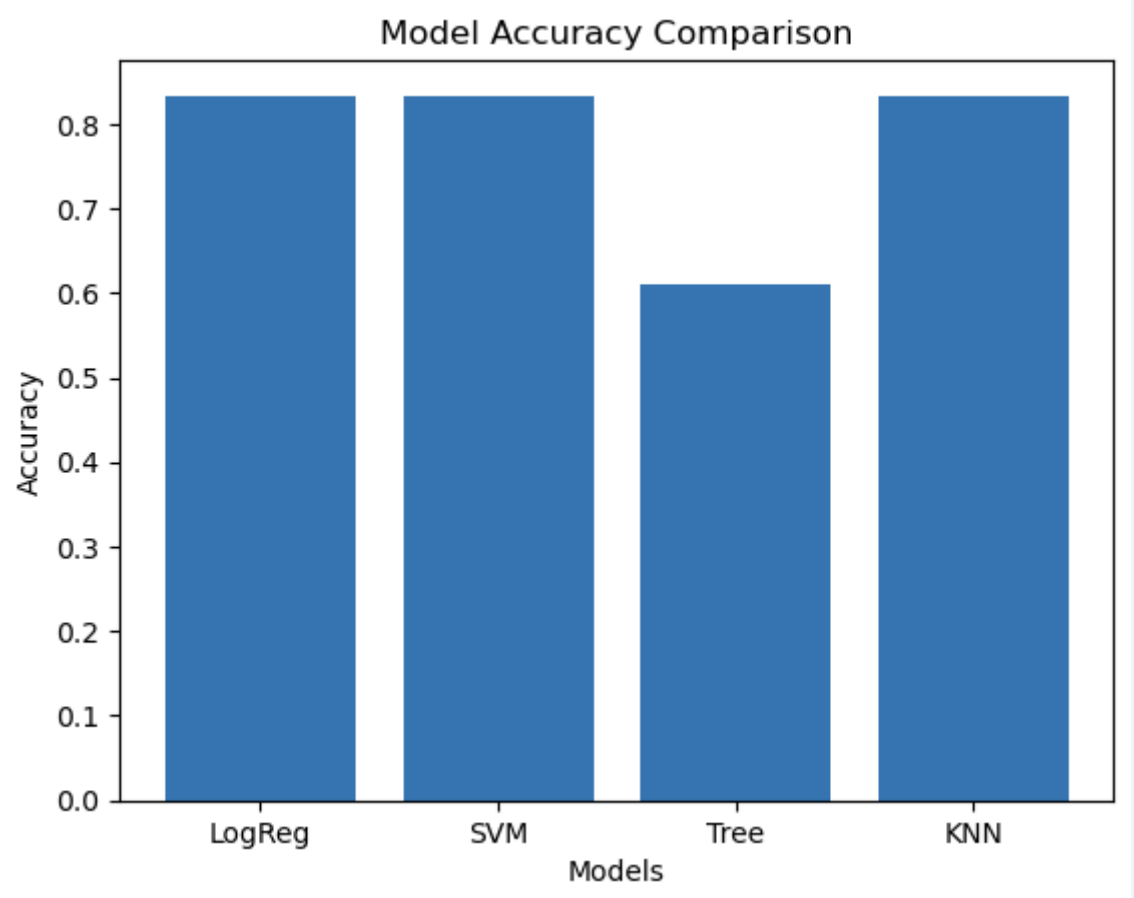
The charts show that payloads between 2000 and 5500 kg have the highest success rate

Section 5

Predictive Analysis (Classification)

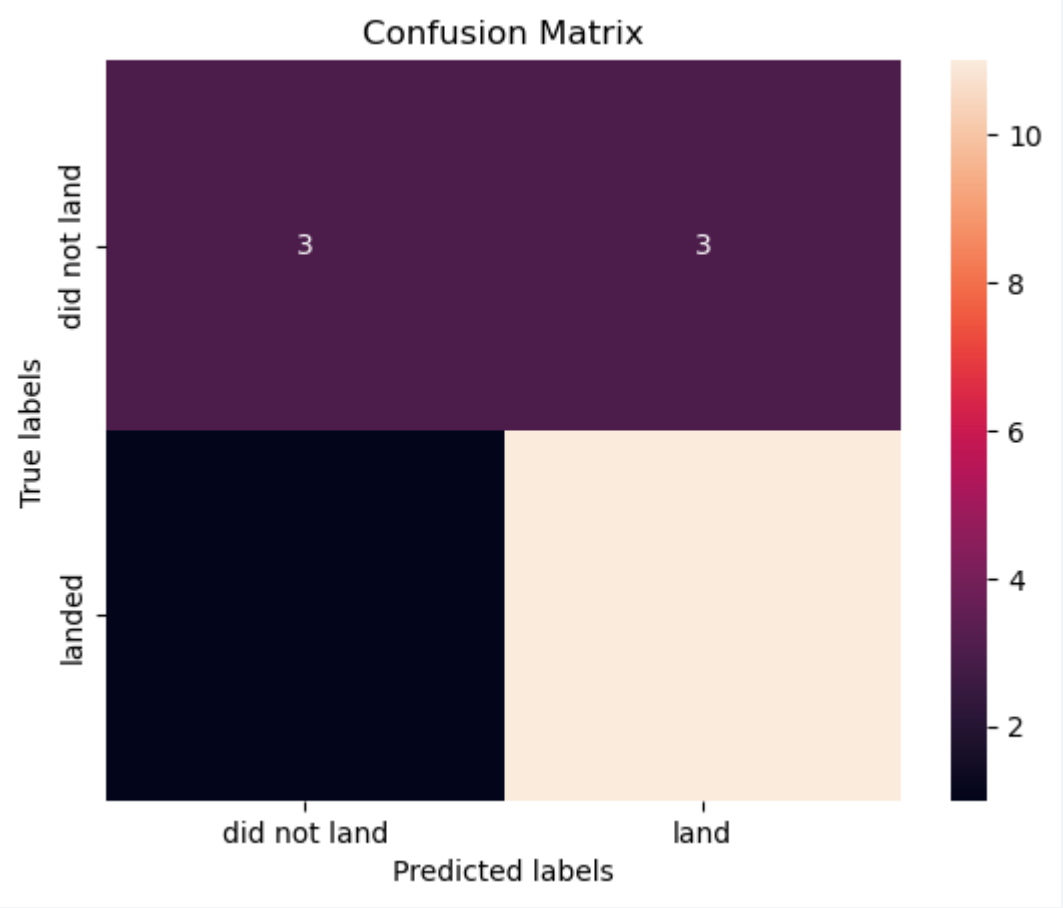
Classification Accuracy

- LogReg, SVM, and KNN have the same accuracy
- The Tree accuracy is lower than the others
- The identical scores on the Same Test Set may be attributed to the limited test sample size of only 18 samples, which could be insufficient to capture significant differences.



Confusion Matrix

Upon examining the confusion matrix, we observe that logistic regression is capable of distinguishing between the different classes. However, a closer inspection reveals that the primary challenge lies in the occurrence of false positives, which is a key area for improvement.



Conclusions

In this project, we aimed to develop a predictive model that determines the likelihood of a Falcon 9 launch's first stage landing, which is a crucial factor in determining the cost of a launch. Our key findings and approach can be summarized as follows:

- We explored the relationships between various features of a Falcon 9 launch, such as payload mass and orbit type, and their impact on the mission outcome.
- We employed multiple machine learning algorithms on historical Falcon 9 launch data to identify patterns and develop predictive models.
- Our models can be used to accurately forecast the outcome of future launches, providing valuable insights for launch cost estimation.

Thank you!