IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kamban Loganathan
02/11/2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using API
  - Data Collection with web scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results

Introduction

- ❖ In this Capstone Project I take the role of Data Scientist
- ❖ To estimate the price of each launch of SpaceX
- ❖ Do this by gathering information about Space X and creating dashboards for the team.
- ❖ And also determine if SpaceX will reuse the first stage.
- ❖ Instead of using rocket science to determine if the first stage will land successfully, I will train a machine learning model and use public information to predict if SpaceX will reuse the first stage

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Get requests to the SpaceX API and web scraping from wikipedia
- Perform data wrangling
  - Clean the data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building an optimal machine learning model

Data Collection

The Data sets are collected by: SpaceX API request. and Web Scraping

Step 1	Step 2	Step 3	Step 4	Step 5
Please enter the URL of the page you'd like to analyze for this project.	Request and parse the SpaceX launch data using the GET request	Convert the response content to JSON format and load it into a Pandas DataFrame.	Now use the API again to get information about the launches using the IDs given for each launch	Filter the data frame to only include Falcon 9 launches and replace null values and get required output

Data Collection – SpaceX API

- How Data Collection has done is given in a form of flow chart for an overview. For completed notebook link is given below

- https://github.com/KambanI/IBM-Capstone-Project--SpaceX/blob/2a9dc8453bf88ad6e27f1ebdf63e6f5dcf7949f1/SpaceX%20Data%20Collection-%20IBM%20Capstone%20Project.ipynb

Make request to SpaceX API

Decode the response content as a JSON

Turn JSON into pandas DataFrame

Use the API again to get information about the launches using the IDs given for each launch

Construct our dataset using the data we have obtained.

Data Collection - Scraping

- A flow chart provides an overview of the data collection process through web scraping. The link to the completed notebook is provided below.

- https://github.com/Kambanl/IBM-Capstone-Project--SpaceX/blob/074ba5904ef313faccd8fd4f3560f3a76d8ae186/SpaceX%20Web%20Scrapping-%20IBM%20Capstone%20Project.ipynb

Falcon 9 Launch Wiki page from its URL

↓

Create a Beautifulsoup object from the HTML

↓

Extract all column names from the HTML table

↓

Create an empty dictionary with keys from the column names

↓

Fill up the dictionary with launch records extracted from table rows.

↓

Convert the dictionary into a CSV dataset

Data Wrangling

- Data Wrangling process is given in a flow chart for a overview. For completed notebook link given below

- https://github.com/Kambanl/IBM-Capstone-Project--SpaceX/blob/0f39b524233f49b73427c06cd399b338ac2c9f58/SpaceX-%20Data%20Wrangling%20-%20Capstone%20Project.ipynb

Data Wrangling process is given in a flow chart:

EDA with Data Visualization

- **Types of Charts Used** :
  - Scatter plot -
    - Flight Number vs Payload Mass
    - Flight Number vs Launch Sites
    - Payload and Launch Sites
    - Flight Number and Orbit Type
    - Payload and Orbit Type
  - Bar chart – Success rate of each orbit
  - Line plot – success rate and Date EDA with Data Visualization complete notebook link is given below

Github link:

https://github.com/Kambanl/IBM-Capstone-Project--SpaceX/blob/7783ec40bd3ab94ed31bd66399af1ffa1289ac53/SpaceX-%20EDA%20with%20Data%20Visualization%20Capstone%20Project.ipynb

EDA with SQL

- Summary of SQL queries that were used: -
    - Display the names of the unique launch sites in the space mission
    - Display 5 records where launch sites begin with the string 'CCA -
    - Display the total payload mass carried by boosters launched by NASA (CRS)
    - Display average payload mass carried by booster version F9 v1.1 -
    - List the date when the first successful landing outcome in ground pad was acheived
    - List the booster names which have success in drone ship & payload mass greater than 4000 but less than 6000
    - List the total number of successful and failure mission outcomes
    - List the names of the booster versions which have carried the maximum payload mass. Use a subquery
    - List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
    - 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
- GitHub Link: https://github.com/Kambanl/IBM-Capstone-Project--SpaceX/blob/53a56fee1a0a5a24d8fa52a8fa9527b61bc6e38c/SpaceX-%20EDA%20with%20SQL%20Capstone%20Project.ipynb

Build an Interactive Map with Folium

- Folium Markers were used to show the SpaceX launch sites and their nearest important landmarks like railways, highways, cities and coastlines.
- Polylines were used to connect the launch sites to their nearest landmarks.
  - Red represents rocket launch failures
  - Green represents the successe

➔ Github Link: https://github.com/Kambanl/IBM-Capstone-Project--SpaceX/blob/15c96ee1eb07cc6f42bcde557a3a6746882426c2/SpaceX-%20Interactive%20Visuals%20with%20Follium.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Build a Dashboard with Plotly Dash

- Pie charts and scatter plots were used to Visualize SpaceX launch records.
- These charts showed the rocket launch success rate for each launch site, providing insights into factors that might influence success rates, such as payload mass and booster version.
- Successful launches were represented by 1 while failures were represented by 0.

❖ Github Link
https://github.com/Kambanl/IBM-Capstone-Project--SpaceX/blob/1b26abeefc78a729748947bc8cdd3167e2fddfe3/SpaceX-%20Dashboard%20Capstone%20Project.py

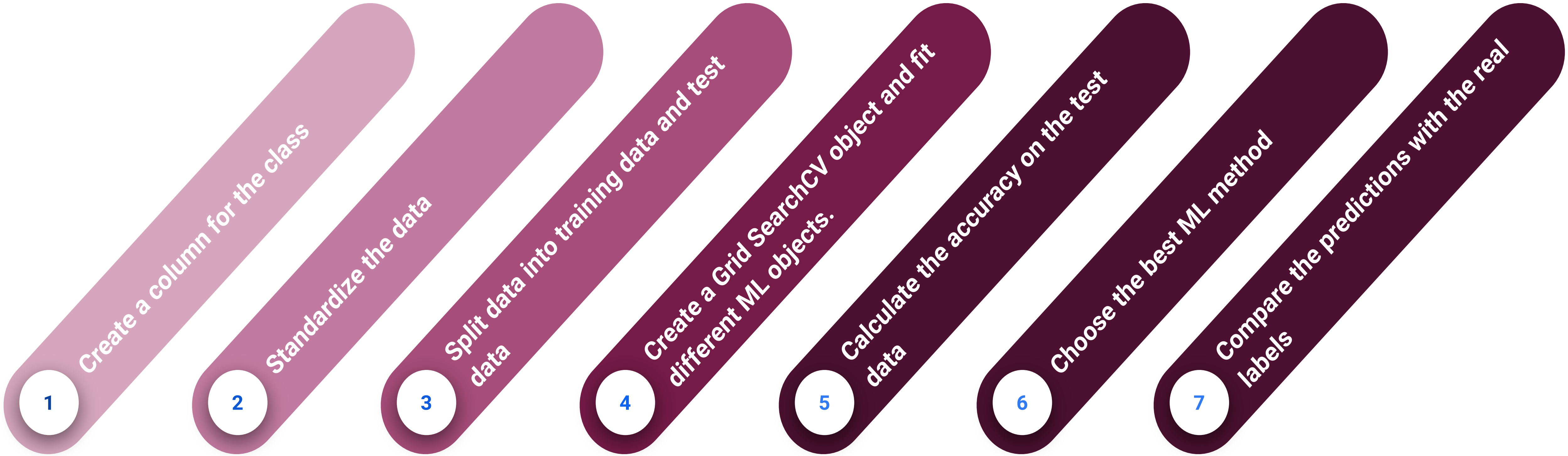
Predictive Analysis (Classification)

Scikit-learn, a machine learning library, was used for predictive analysis. The process involved:

- Building a machine learning pipeline to predict whether the first stage would successfully land based on the given data.

- Github Link:
  - https://github.com/Kambanl/IBM-Capstone-Project--SpaceX/blob/8ef1929e7881780183fdaa2f8f4d34234998c327/SpaceX%20-%20Machine%20Learning%20Prediction%20Capstone%20Project.ipynb

Flow Chart for Predictive Analysis:

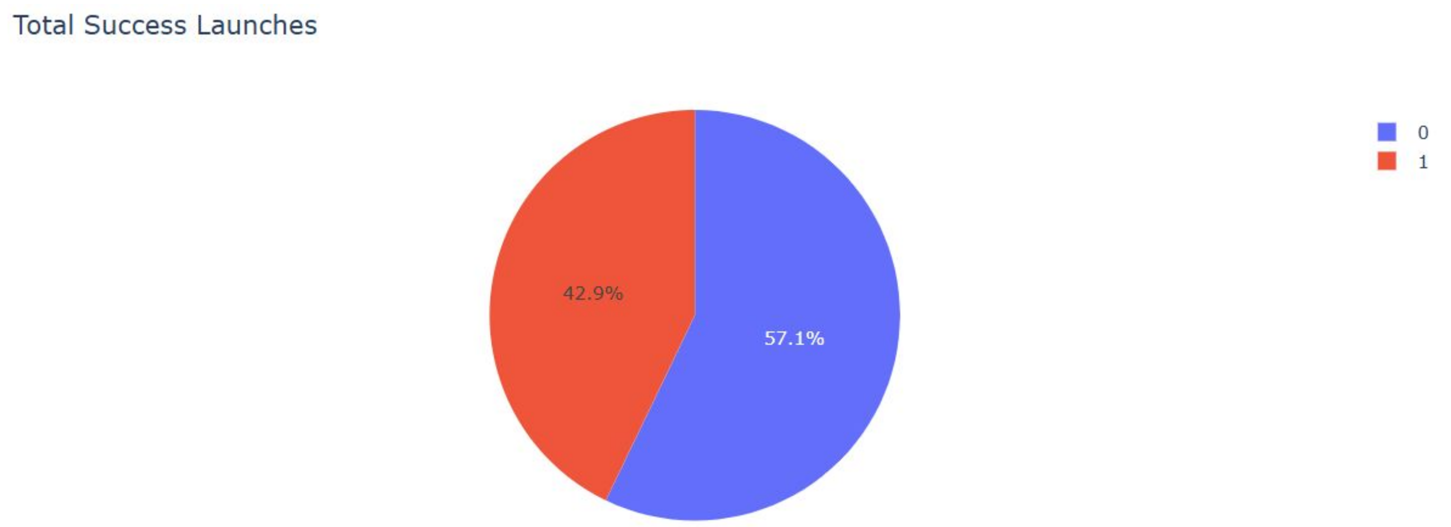


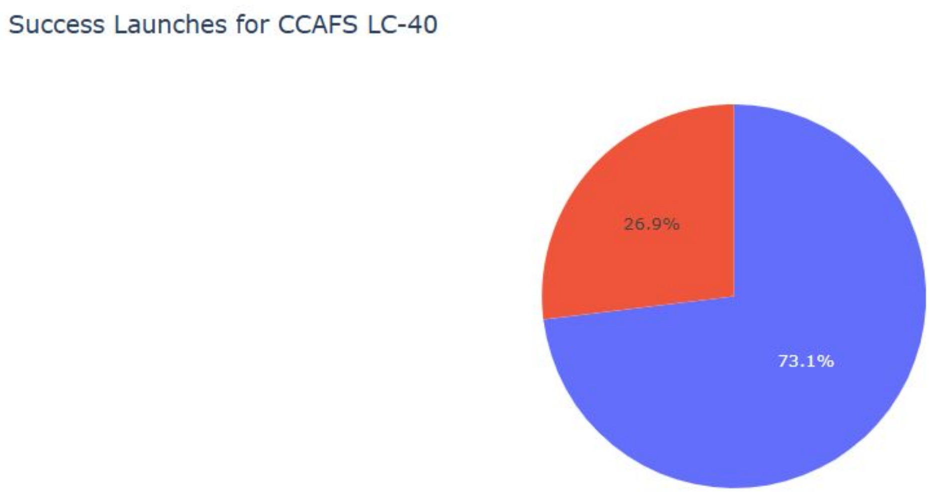
Results

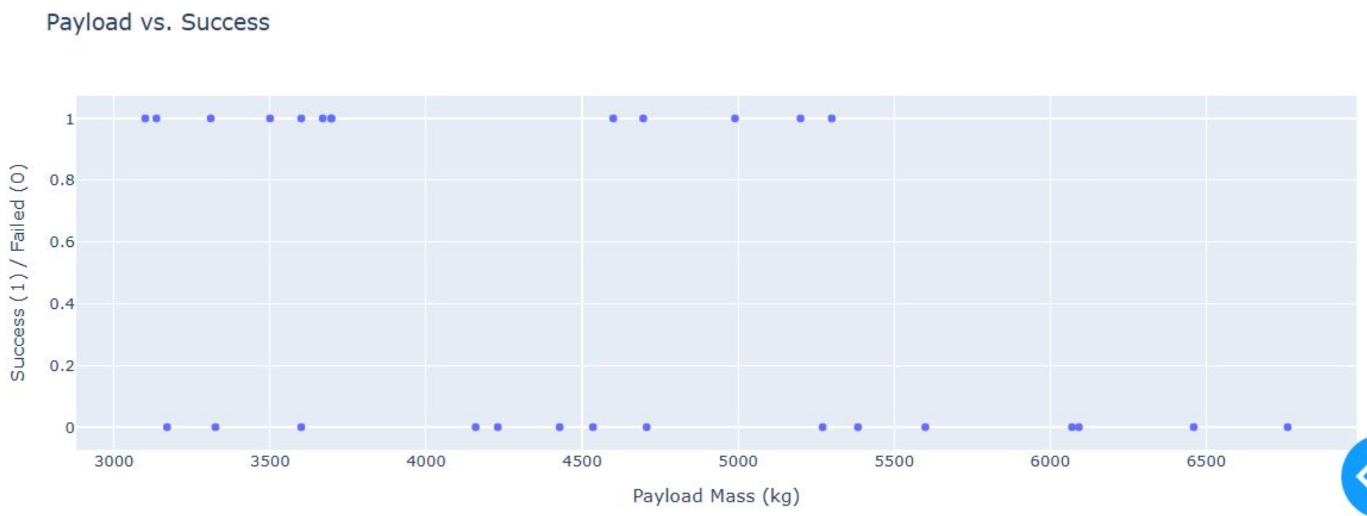
- Exploratory data analysis revealed a moderate correlation between successful landing outcomes and flight number. It also showed a marked increase in successful landings since 2015.
- All launch sites are located near the coast line. Perhaps, this makes it easier to test rocket landings in the water
- Sites are also located near highways and railways. This may facilitate transportation of equipment and research material.

- The machine learning were able to predict the landing success of rockets with an accuracy score of 83.33%.

Interactive analytics demo in screenshots:



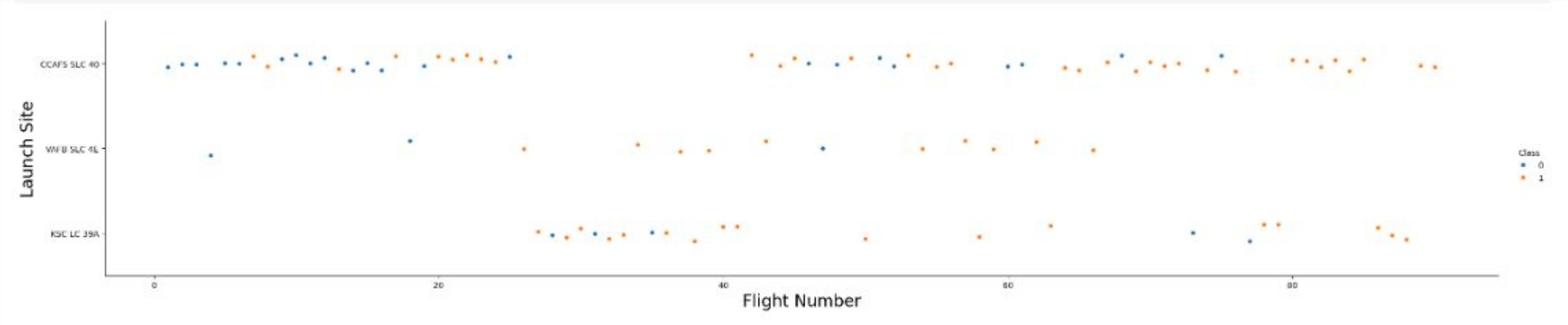




Section 2

Insights drawn from EDA

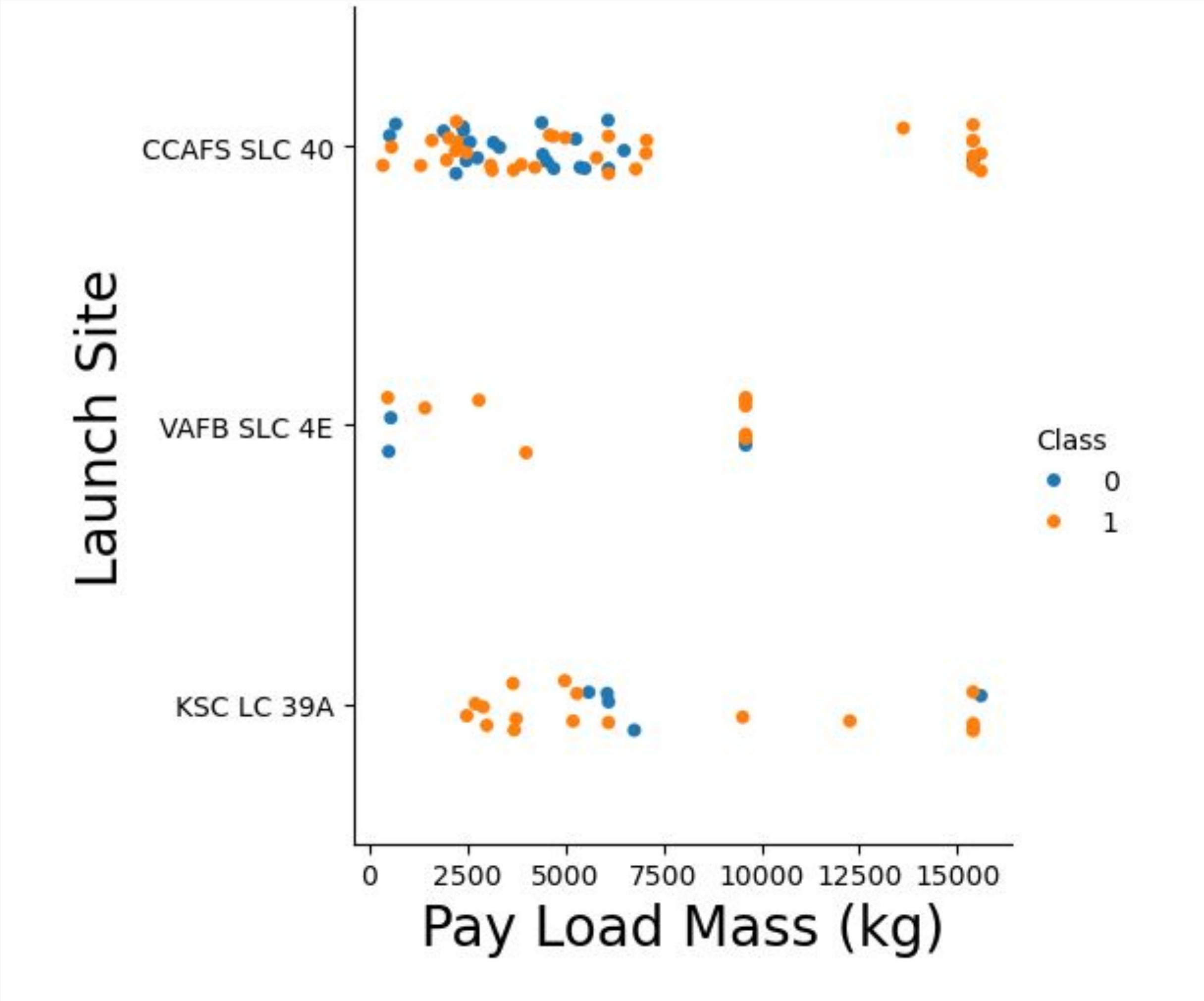
Flight Number vs. Launch Site



➔ It seems that successful landings became more frequent with higher flight numbers, with the CCAFS SLC 40 launch site recording the highest number of landings.

Payload vs. Launch Site

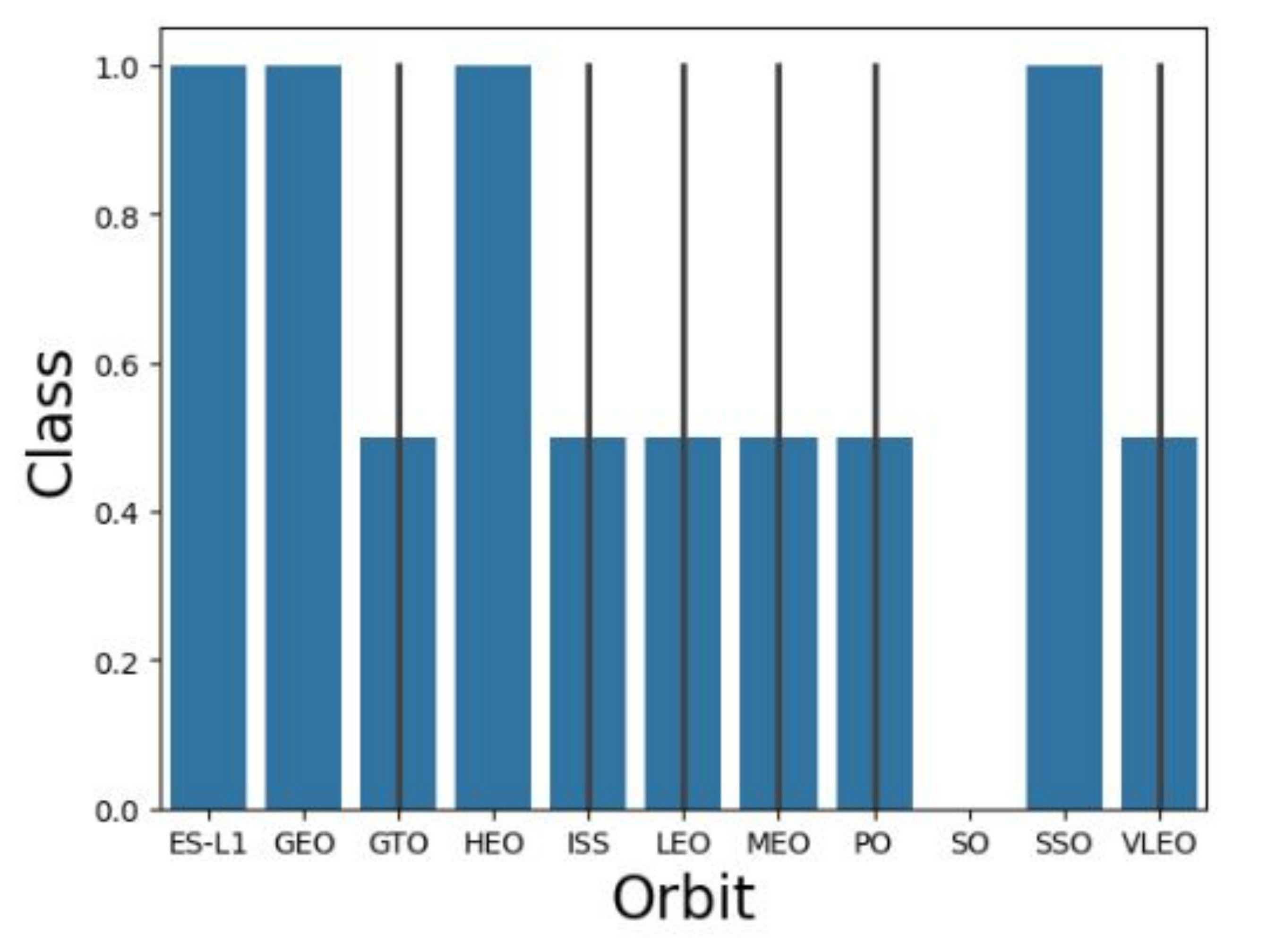
➔ If you look at the scatter plot, you'll notice that no rockets were launched with a heavy payload mass (over 10,000) at the VAFB-SLC launch site.



Success Rate vs. Orbit Type

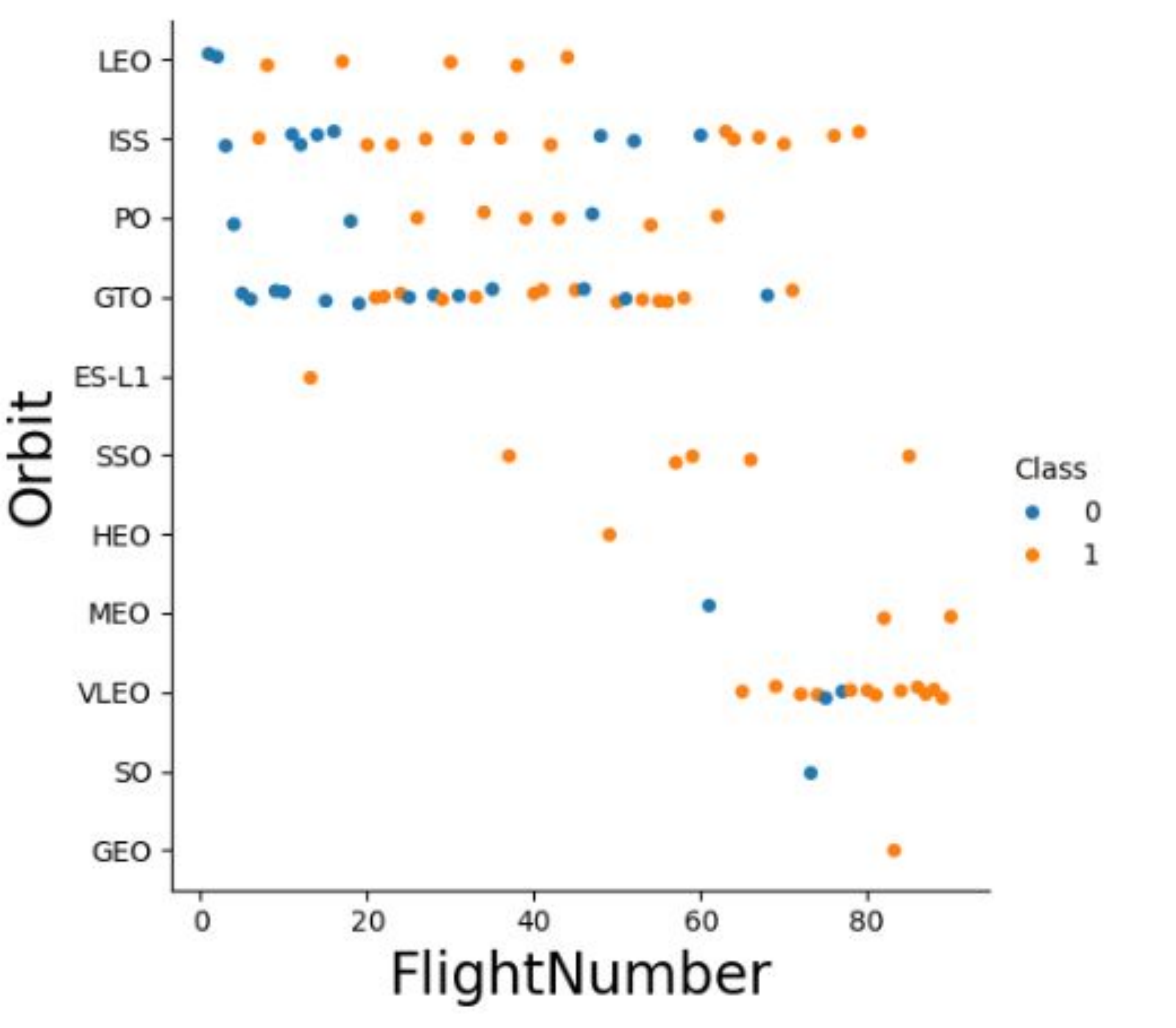
➔ The highest success rate ORBITS are

- **ES-L1**
- **GEO**
- **SSO**
- **HEO**



Flight Number vs. Orbit Type

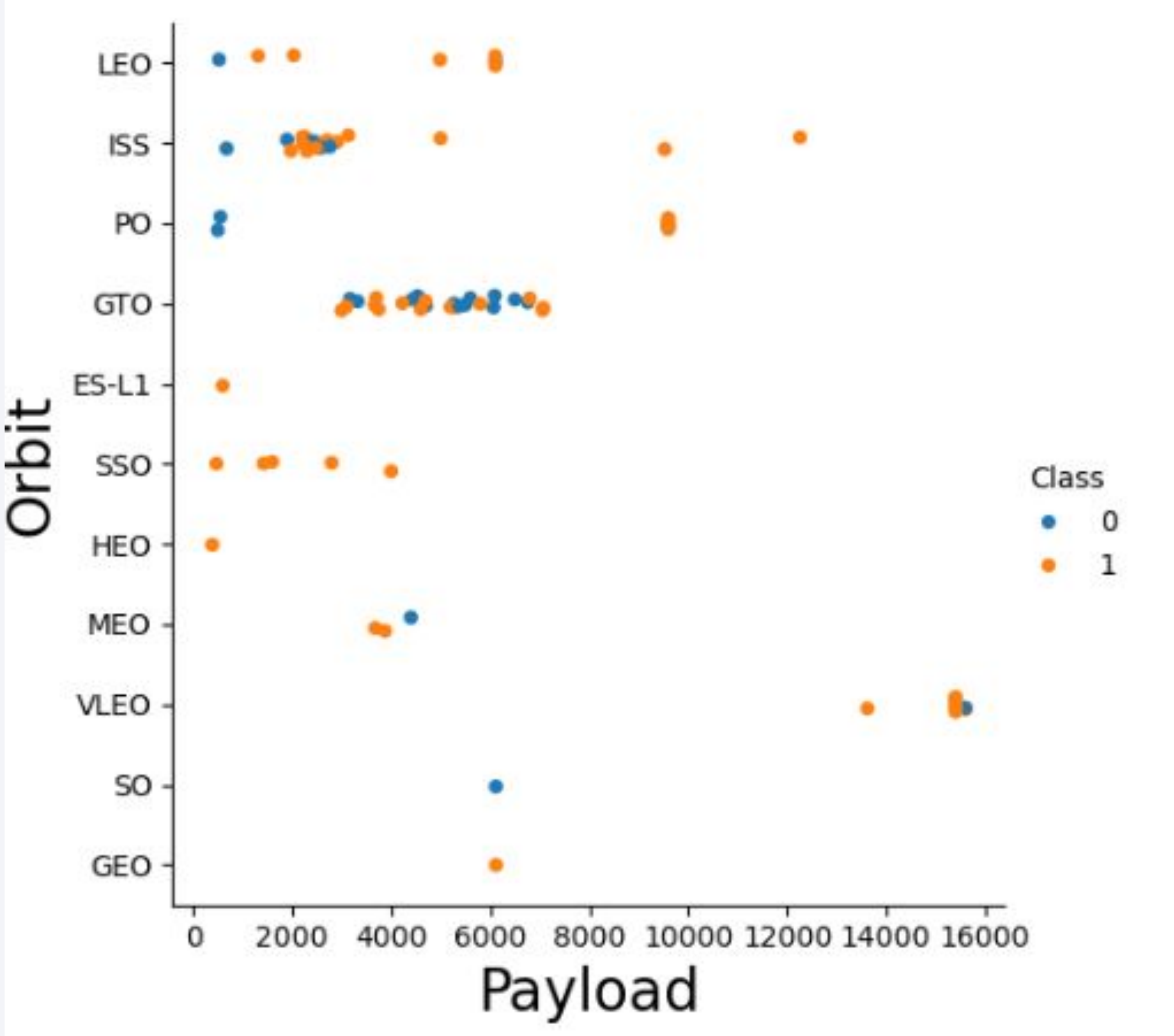
➔ In the LEO orbit, success appears to be linked to the number of flights. However, in the GTO orbit, there seems to be no connection between flight count and success.



Payload vs. Orbit Type

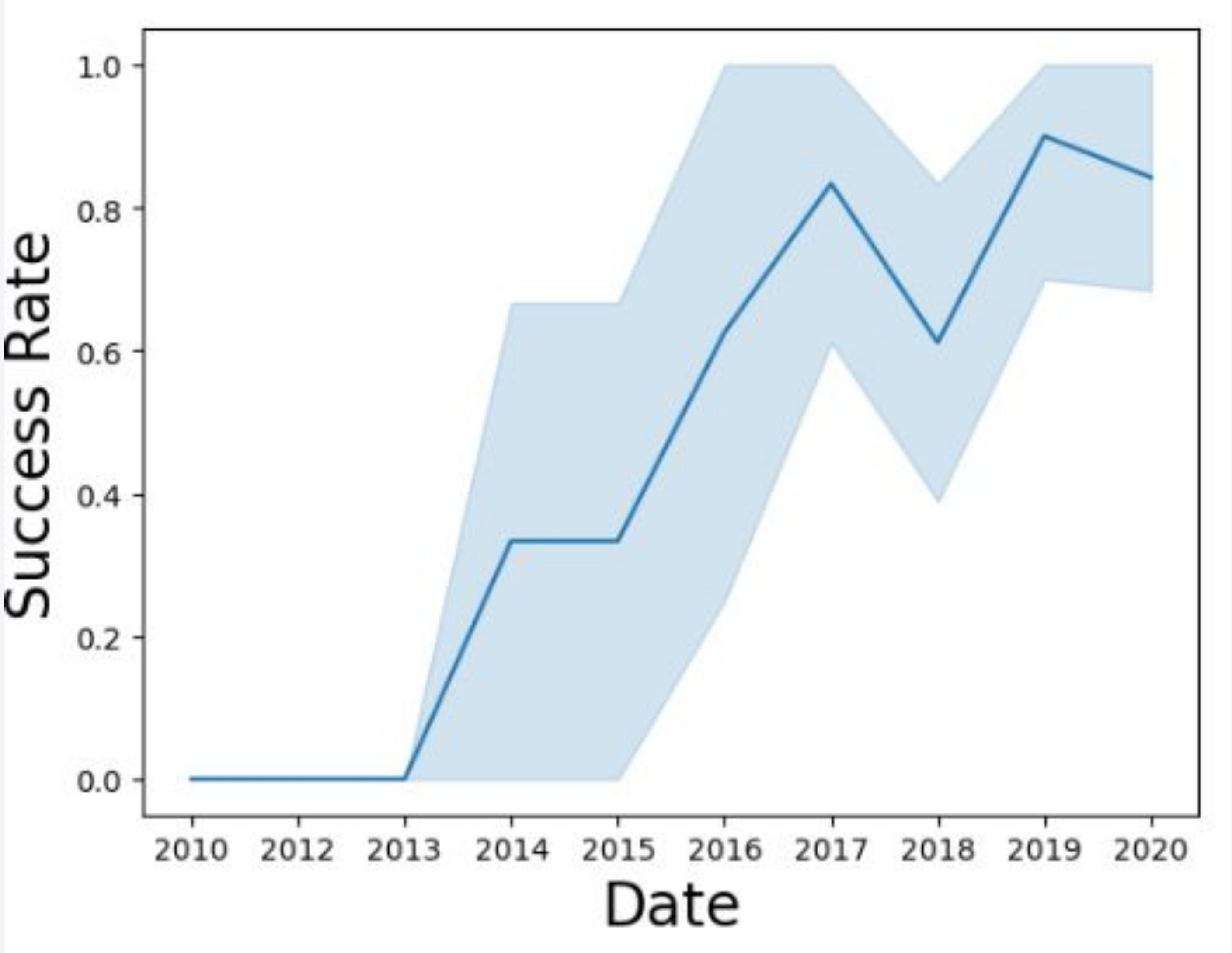
- ➔ For heavy payloads, the landing success rate is higher in Polar, LEO, and ISS orbits.

- ➔ However, in the GTO orbit, successful and unsuccessful landings occur in similar numbers, making it harder to differentiate between the outcomes.



Launch Success Yearly Trend

- It is apparent that the success rate has significantly increased from 2013 to 2020



All Launch Site Names

- Based on the data, here are the names of the launch sites where various rocket landings were attempted:
    - CCAFS LC-40
    - CCAFS SLC-40
    - KSC LC-39A
    - VAFB SLC-4E
- The Average payload mass carried by booster version F9 v1.1: **2928.4 Kg**

Launch Site Names Begin with 'CCA'

- Here are five records of launch sites that start with the letters 'CCA.' Notably, there are other organizations apart from SpaceX conducting rocket tests.

```
%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

Total Payload Mass

```
%%sql
SELECT sum(payload_mass__kg_) AS "Total payload mass (NASA (CRS))" FROM SPACEXTBL WHERE customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://qvc29638:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb?authSource=admin&replicaSet=replset
Done.
```

Total payload mass (NASA (CRS))
45596

- The information in the picture displays the total payload mass carried by boosters launched by NASA

Average Payload Mass by F9 v1.1

```
]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'

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

AVG(PAYLOAD_MASS__KG_)
2928.4

The average payload mass carried by F9 v1.1 was **<u>2928.4 kg</u>**

First Successful Ground Landing Date

```
%%sql
SELECT min(DATE) AS "First successful landing outcome in ground pad" FROM SPACEXTBL WHERE landing__outcome = 'Success (ground pa
d)';
```

```
 * ibm_db_sa://qvc29638:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb?aut
hSource=admin&replicaSet=replset
Done.
```

First successful landing outcome in ground pad
2015-12-22

From the picture given above you can see that the first successful ground pad was in 22 December 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT booster_version FROM SPACEXTBL WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 AND 600
0;
```

```
 * ibm_db_sa://qvc29638:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb?aut
hSource=admin&replicaSet=replset
Done.
```

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

It appears that there only 4 Boosters with a payload mass between 4000 and 6000 they are

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

```
%sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in fligh
```

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

count(MISSION_OUTCOME)
99

The total number of successful and failure mission outcomes are shown above in the picture.

Boosters Carried Maximum Payload

- We found that only 12 boosters have carried the maximum payload mass of 15600 kg

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT max(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * 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

```
%%sql

SELECT booster_version, launch_site,landing__outcome
FROM SPACEXTBL
WHERE landing__outcome = 'Failure (drone ship)'
  AND EXTRACT(YEAR FROM DATE) = 2015
```

```
 * ibm_db_sa://qvc29638:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb?authSource=admin&replicaSet=replset
Done.
```

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

- Only 2 boosters F9 v1.1B1012_CCAFS LC-40 and F9v1.1B1015 CCAFS LC-40 failed to land in 2015

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

- The number of successful landings have increased since 2015.

```
%%sql

SELECT landing__outcome, COUNT(*) AS "Count"
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' and '2017-03-20'
GROUP BY landing__outcome
ORDER BY Count DESC
;
```

```
 * ibm_db_sa://qvc29638:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb?authSource=admin&replicaSet=replset
Done.
```

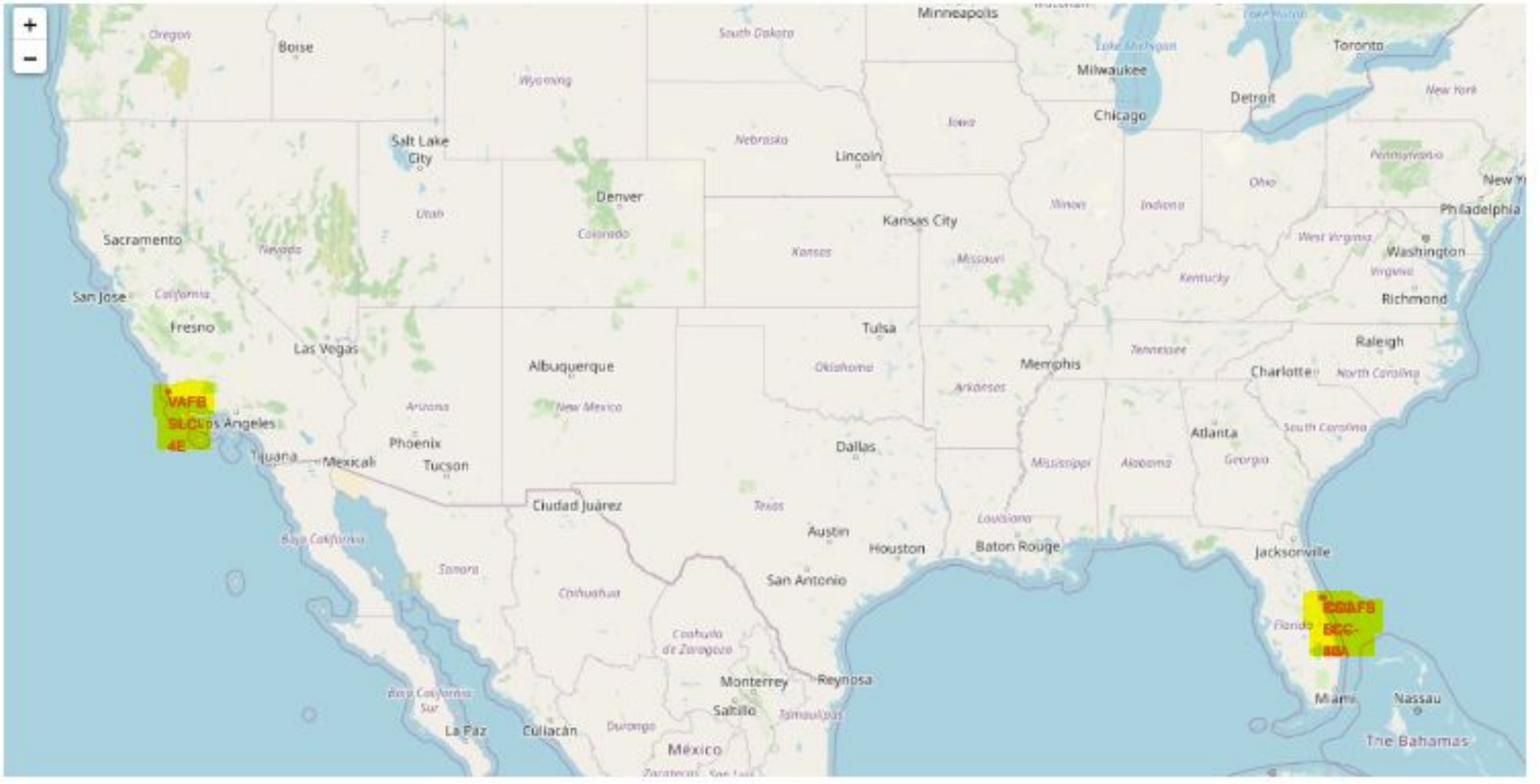
landing__outcome	Count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Launch Site Locations:

- All launch sites are situated very close to the coast and are located a few thousand kilometers away from the equator. Due to the following reasons:
  - Safety Considerations
  - Fuel Efficiency and Orbital Velocity
  - Global reach for satellite Deployment

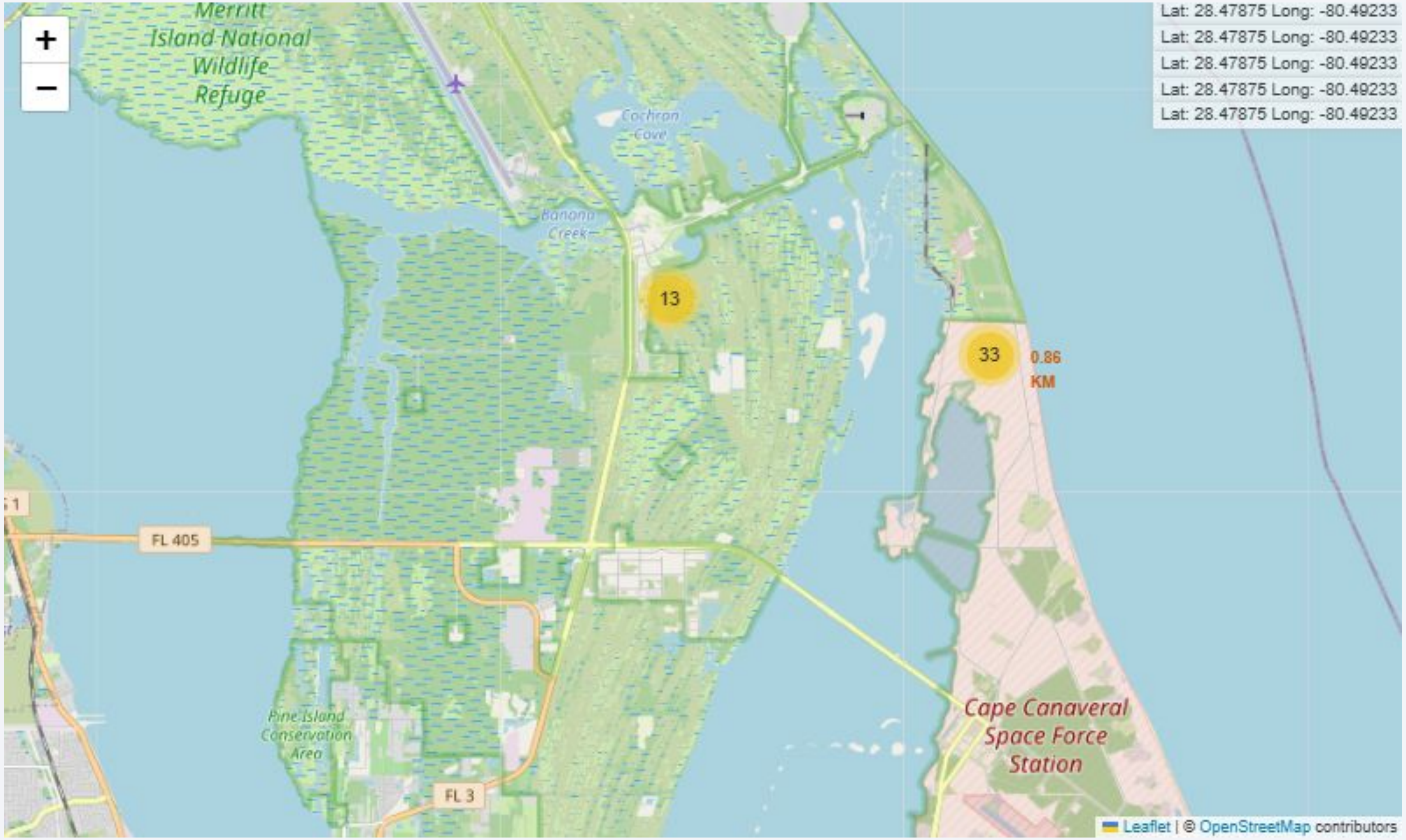


PolyLine between a launch site to the selected coastline point:

- A City Map: 

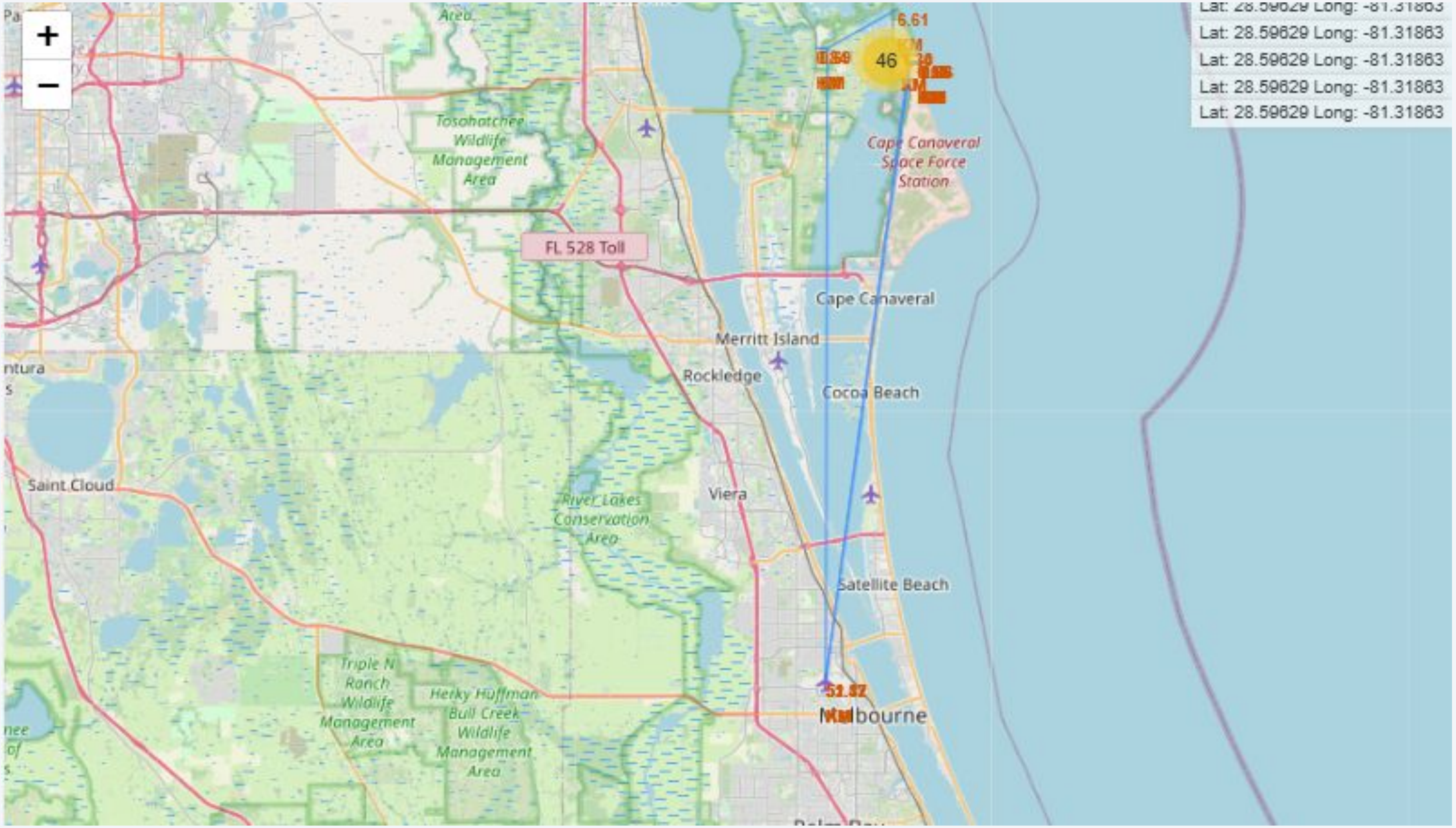
- A railway line: 

- A Highway 



Surrounding Landmarks:

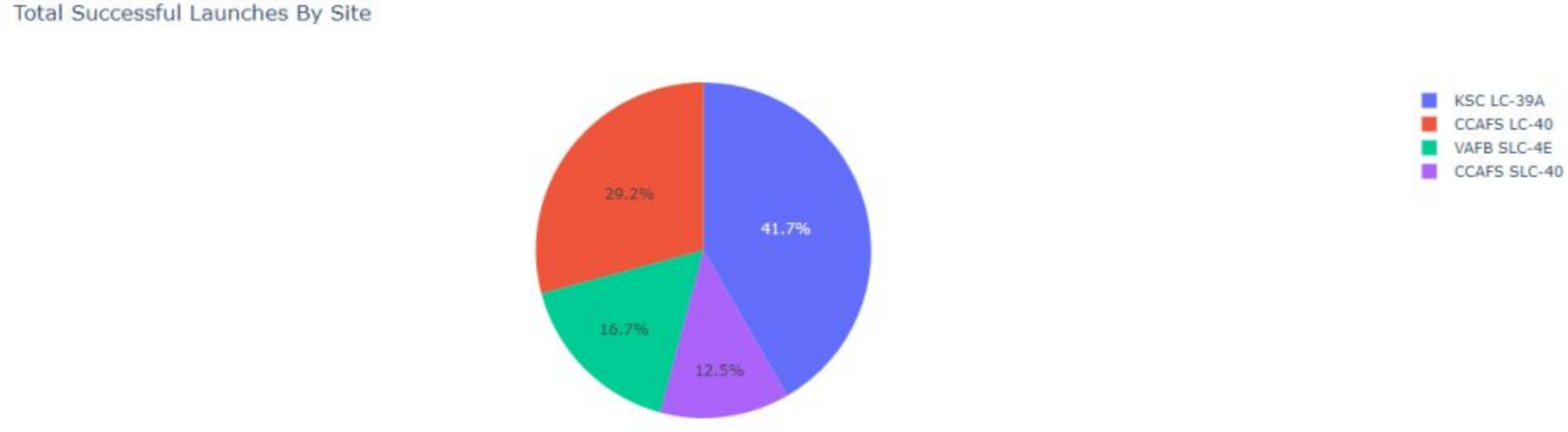
- Launch sites are typically located at least 18 km from cities, likely to reduce the risk of accidents near populated areas.
- Additionally, these sites are often close to railways and highways, possibly to facilitate the transport of rocket components.
- Many launch sites are also near coastlines, as shown by frequent rocket landing tests conducted over bodies of water like the ocean.



Section 4

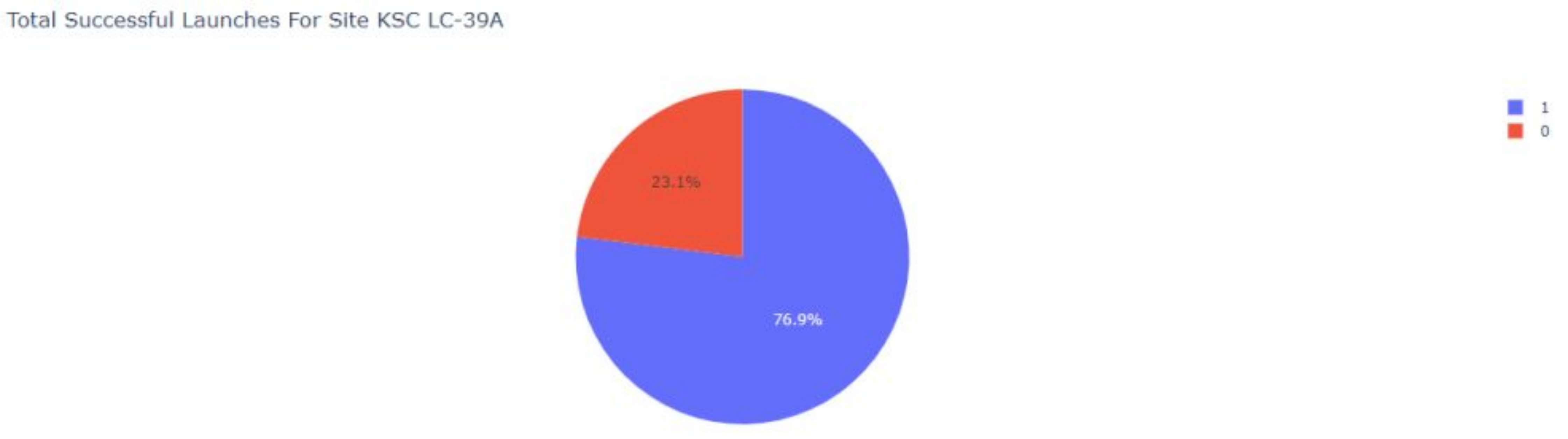
Build a Dashboard with Plotly Dash

Successful launches sitewise:



In summary, KSC LC-39A and CCAFS LC-40 contribute the majority of successful launches, while VAFB SLC-4E and CCAFS SLC-40 contribute fewer successful launches. This distribution may reflect differences in launch site capabilities, mission types, or frequency of use.

Maximum Successful Launches for Site KSC LC-39A:



- ➔ **Success Rate**: The large blue slice suggests that the success rate for launches at KSC LC-39A is high.
- ➔ **Reliable Launch Site:** The high success rate implies that this site is a reliable and well-functioning launch pad.
- ➔ **Possible Reasons for Success**: Factors contributing to the success rate could include advanced technology, experienced teams, and rigorous safety protocols.

Payload Mass vs. Launch Success for All Sites:

Inferences from the Plot:

1. Payload Mass and Launch Success:

- There appears to be no clear correlation between payload mass and launch success.
- Launches with a wide range of payload masses (between 2000 kg and 3600kg) have both successful and unsuccessful outcomes.

2. Booster Version Category:

- There are four distinct booster versions (V1.1, FT, B4, and B5).
- Each version has a mix of successful and unsuccessful launches, indicating that booster version alone does not guarantee success.

3. Potential Factors Influencing Success:

- Other factors beyond payload mass and booster version likely influence launch success. These could include:
  - Weather conditions
  - Equipment reliability
  - Mission complexity
  - Crew/payload health

Section 5

Predictive Analysis (Classification)

Classification Accuracy

```
accuracy = [svm_cv_score, logreg_score, knn_cv_score, tree_cv_score]
accuracy = [i * 100 for i in accuracy]

method = ['Support Vector Machine', 'Logistic Regression', 'K Nearest Neighbour', 'Decision Tree']
models = {'ML Method':method, 'Accuracy Score (%)':accuracy}

ML_df = pd.DataFrame(models)
ML_df
```

	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	83.333333
3	Decision Tree	83.333333

- The resulting Data Frame provide information about the accuracy scores of four different machine learning models: Support Vector Machine, Logistic Regression, K Nearest Neighbor, and Decision Tree.
- All four models have an accuracy score of 83.3333%.

Confusion Matrix:

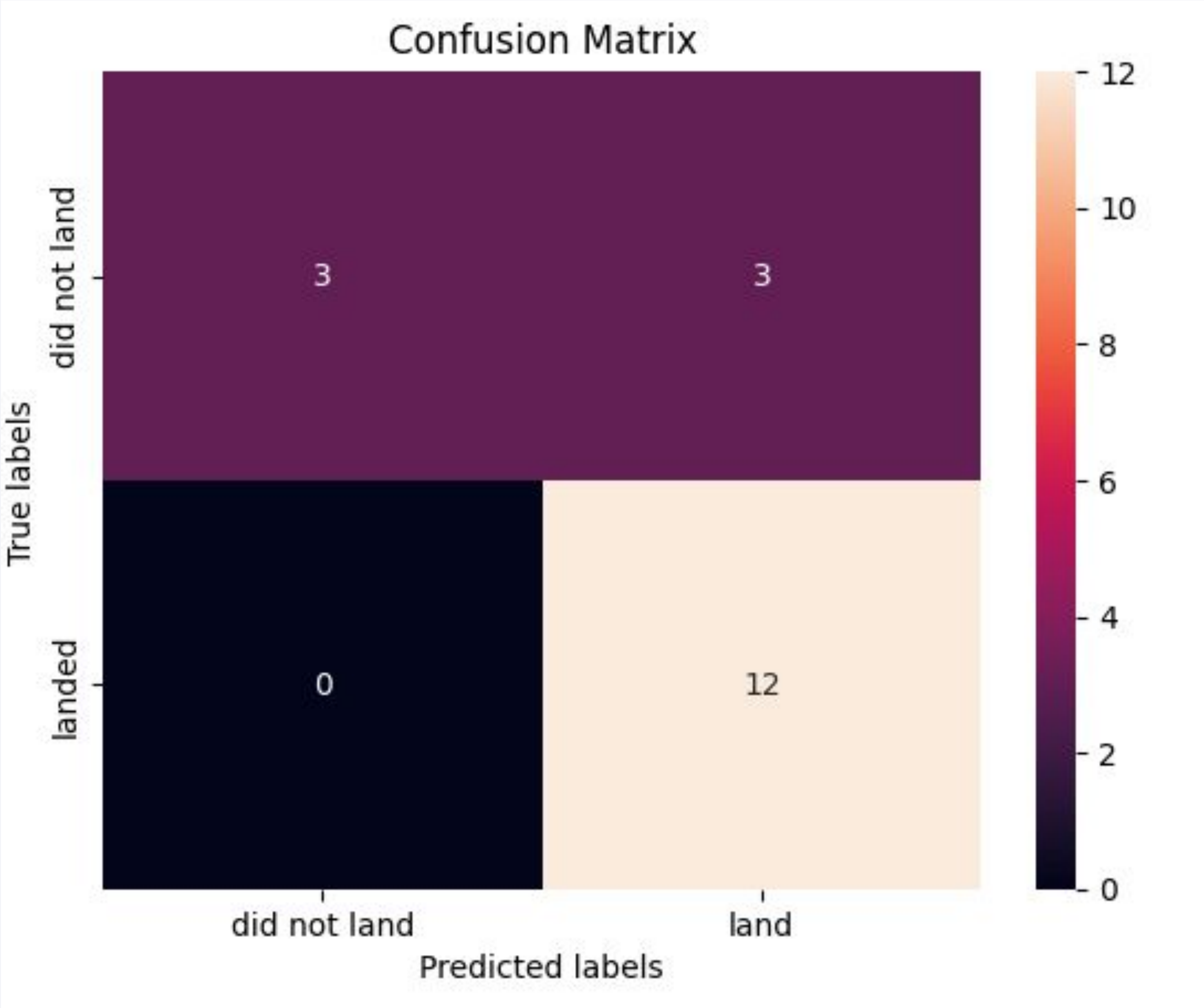
- **Overall Accuracy:** The model's overall accuracy is calculated as (TP+TN)/(TP+TN+FP+FN). In this case, it's (12+0)/(12+0+3+3) = 0.67 or 67%. Moderately accurate.

- **Class-Specific Accuracy:** The model is better at predicting when a rocket lands (12/15 = 80%) compared to when it does not land (0/3 = 0%).

- **Precision:** Precision for the "landed" class is 12/(12+3) = 0.80 or 80% i.e., model predicts a landing, it is correct 80%

- **Recall:** Recall for the "landed" class is 12/(12+3) = 0.80 or 80%. This means the model correctly identifies 80% of the actual landings.

- **F1-Score:** The "landed" class, the F1-score is 2*(0.80*0.80)/(0.80+0.80) = 0.80.



Conclusions

Key Factors Driving SpaceX's Success:

1. **Strategic Site Selection:**

   - Coastal Launch Sites: Minimizes risks to populated areas during landing attempts.
   - KSC LC-39A: Consistently high success rates.
2. **Continuous Improvement:**

   - Increased Landing Success Over Time: Significant improvements observed since 2015.
   - Correlation Between Flight Number and Success: Experience and data-driven refinements lead to better outcomes.

3. All this data was used to train a machine learning model that is able to predict the landing outcome of rocket launches with 83.33% accuracy.

Thank you!