IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gbolahan (GB) Oladeji

2023/07/24

Outline

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

Executive Summary

We employ data science methodologies, including Data Collection, Exploratory Analysis, Visualization, and Machine Learning Modeling, to predict the likelihood of successful landings for the SpaceX Falcon9 first stage.

Notably, our findings indicate an increasing trend in successful landings over time due to advancements made during trials throughout the years. Further details to be discussed in subsequent slides.

Introduction

Project background and context

- Space exploration is a costly venture that was previously only dominated by governmental agencies due to enormous funding requirements. However, new private options like SpaceX are disrupting the landscape by competing on technology efficiency, pricing, and faster delivery. Thereby providing more options and increased launch activities. For example, SpaceX as of 2013 was about $15mm cheaper than the nearest competitor[].
- One of the key push by private enterprises as a business case is the ability to reuse the rocket, which if successful, would further lower the cost of rocket launch by an order of magnitude.

Problem To Solve:

- With reuse being a key business strategy, we need to use data science to solve the challenge of achieving success launch rate as well as predict the likelihood of successful landing of the rocket ultimately leading to reuse.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts
- The data were collected through various sources including:
  - (1) SpaceX REST API which is an open-source RESPI for launch, rocket, core, capsule, Starlink, launchpad, and landing pad data.
  - (2) Collection through Web scraping from the Wikipedia page containing the SpaceX launch data.
- These would be discussed in detail in the next slide.

Data Collection – SpaceX API

- Show on the right is the flowchart of the data collection via the SpaceX API
- API:

https://api.spacexdata.com/v4/launches/past

- Data Collection Notebook:

https://github.com/devtage-goladeji/ai-capstone/blob/main/spacex-data.ipynb

Task1: (Collection) Request & Parse the SpaceX launch data using Get Request

↓

Task2: Filter the DataFrame to only include Falcon 9 launches

↓

Task 3: (Wrangling) Deal with Missing Values (– fill via mean value)

↓

Task 4: Save Data for further analysis

Data Collection - Scraping

- With the process, the launch records were scraped from HTML in a Wikipedia:

https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

- Notebook GitHub URL:

https://github.com/devtage-goladeji/ai-capstone/blob/main/jupyter-labs-webscraping.ipynb

CSV produced:

https://github.com/devtage-goladeji/ai-capstone/blob/main/spacex_web_scraped.csv

TASK 1: Request the Falcon9 Launch Wiki page from its URL via HTTP Get Method and BeautifulSoup object

↓

Task2: Extract all column/variable names from the HTML table header using HTML tags with soup object

↓

Task 3: Create a data frame by parsing the launch HTML tables

↓

Task 4: Save Data for further analysis spacex_web_scraped.csv

Data Wrangling

- **Describe how data were processed**

In the data wrangling process, we performed some exploratory data analysis to understand which data elements are labels and which are features. Also, the data were analyzed for quality assurance and issues such as missing values were addressed.

- **Key Tasks:**

1) Load Data dataset_part_1.csv	2) Identify & Calculate % of missing values	3) Identify numerical & categorical columns	4) Calculate # of launches per site
5) Calculate # & occurrence of each orbit	6) Calculate # & occurrence of orbits mission outcomes	7) Create Landing Outcome label	8) Export/save result data

- **Notebook GitHub URL:** https://github.com/devtage-goladeji/ai-capstone/blob/main/labs-jupyter-spacex-Data-wrangling.ipynb
- **Data Generated:** https://github.com/devtage-goladeji/ai-capstone/blob/main/dataset_part_2.csv

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts

Charts	Rationale
Scatter Plot – Flight Number vs Payload Mass	How continuous launch attempt & payload variables affects launch outcome. Appears the more the payload, the less successful the launch
Scatter Plot – flight Number vs Launch Site	Visualize the relationship between Flight Number and Launch Site – and see if the success rate is tied to the launch site. We see that CCCAFS LC-40 has a higher success rate, but equally more launches as well.
Scatter Plot – Payload Mass vs Launch Site	Visualize the relationship between Payload and Launch Site – No rocket was launched for heavy payload at VAFB-SLC launch site. Also, success outcome was more within less heavy payload (less than 7000kg)
Bar Chart – Orbit Type vs Class Mean	Visualize the relationship between the success rate of each orbit type. Here we can see that ES-L1, GEO, HEO, and SSO had higher success rates than other orbit types.
Scatter Plot – flight Number vs Orbit	Here we also see that LEO success appears inversely correlated to the number of flights, however no such relationship with GTO which has success across all flight numbers.
Scatter Plot - Payload Mas vs Orbit	Find the relationship between payload mass and orbit. The relationship is a mixed bag; Heavy payload success is more seen for VLEO, but for GTO, both heavy payload success and failures are recorded
Line graph – Year vs Success rate	Visualize the launch success yearly trend – we see from the trend the increasing success rate as the years progress.

- EDA visualization Notebook GitHub URL: https://github.com/devtage-goladeji/ai-capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

  Data Generated: https://github.com/devtage-goladeji/ai-capstone/blob/main/dataset_part_3.csv

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
    - Select Distinct Launch Site
    - Display 5 records where launch sites begin with '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 the ground pad was achieved
    - List the names of the boosters with success in drone ship and with payload mass > 4000 and < 6000
    - List the total number of successful and failed mission outcomes
    - List names of the booster version which have carried maximum payload mass.
    - List records with drone failure landing outcomes in 2015
    - Rank landing outcomes (failure or success) between 2010/06/04 and 2017/03/20 in descending order
- EDA SQL Notebook GitHub URL:

https://github.com/devtage-goladeji/ai-capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Summary of map objects created & added:

- Folium.Circle object for the launch sites
- Folio.MarkerCluster as a container for the Folio.Marker
- Folium.Icon
- Folium.Marker
- Folium.PolyLine –

Rationale:

- Objects were added to provide visual clarity and interpretation to map related data. For example, the color for the launch outcome helps put into perspective, the success/failure of each launch within each launch site.

Notebook GitHub URL:

https://github.com/devtage-goladeji/ai-capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- As part of the exploratory data visualization effort, the following were added during the exercise:
- Dropdown list for Launch site selection (providing data slices via point and click)
- Pie chart – provide a breakdown of launch sites # of launches and outcomes
- Slider to filter the payload
- Scatter diagram to show a correlation between payload and launch success s/ outcomes
- Callbacks to respond to click events for pie charts and the payload slider/scatter diagram
- Explain why you added those plots and interactions:

- Notebook GitHub URL: https://github.com/devtage-goladeji/ai-capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

The following is the summary of the model development process in arriving at the best classification model: Build the machine learning pipeline:

Data Preprocessing → Train Test Split → Train | GridSearchCV → Scoring / Deployment

Data Preprocessing: Profile data, handle missing values using advanced techniques, apply suitable encoding for categorical data, and consider different feature scaling methods (used standardization classifier here).

Train-Test Split: Divide data into training and testing sets; use cross-validation to improve generalization and avoid overfitting.

Hyperparameter Tuning: Utilize GridSearchCV to find optimal hyperparameters. Used different estimators including LogisticRegression, SVM, DecisionTree, KNN

Model Selection: Evaluate models using various metrics; choose the best-performing model for deployment.

Interpretability: Use interpretable models like linear regression or decision trees.

Model Deployment: Thoroughly test the model with real-world data before deploying it.

Continuous Improvement: Iterate and refine the process to enhance model accuracy and robustness over time.

Notebook GitHub URL:

https://github.com/devtage-goladeji/ai-capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

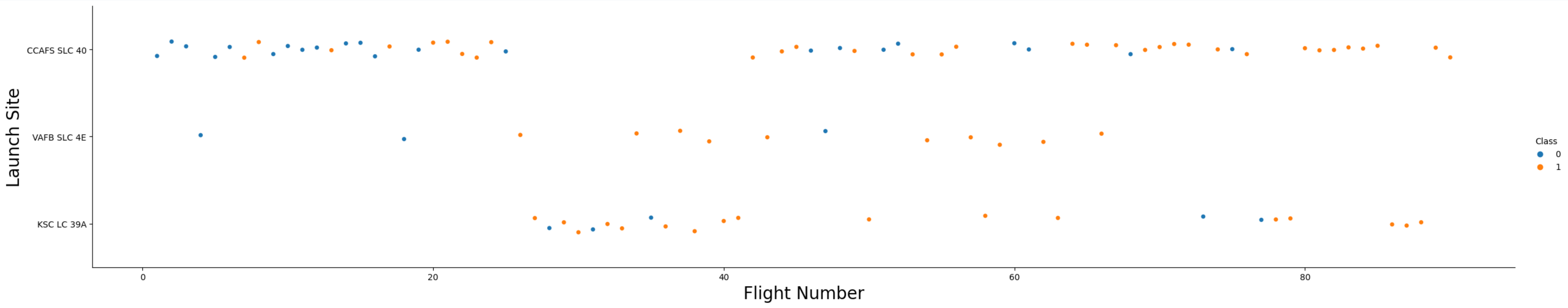
Results

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

Section 2

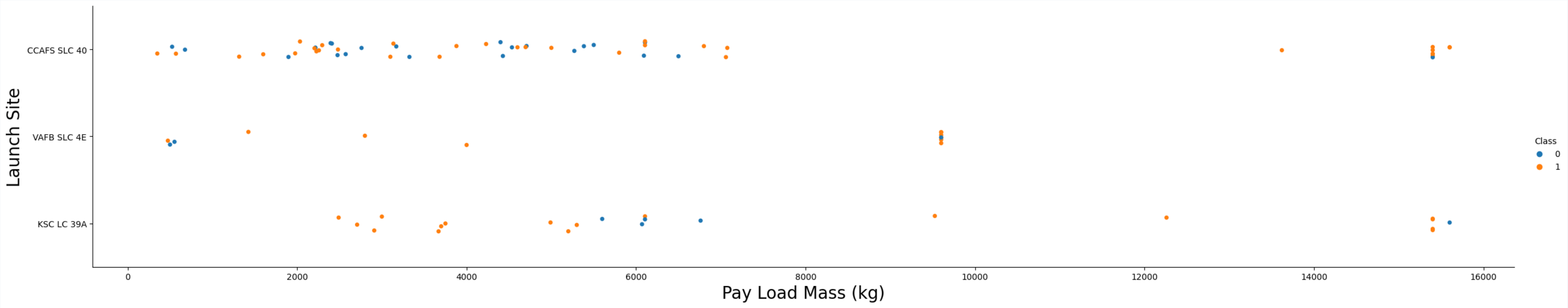
Insights drawn from EDA

Flight Number vs. Launch Site



- VAFB SLC 4E had fewer launches and fewer positive outcomes compared to the other 2 launch sites. Also, it appears further subsequent launches were prioritized for both CCAFS SLC 40 and KSC LC 39A

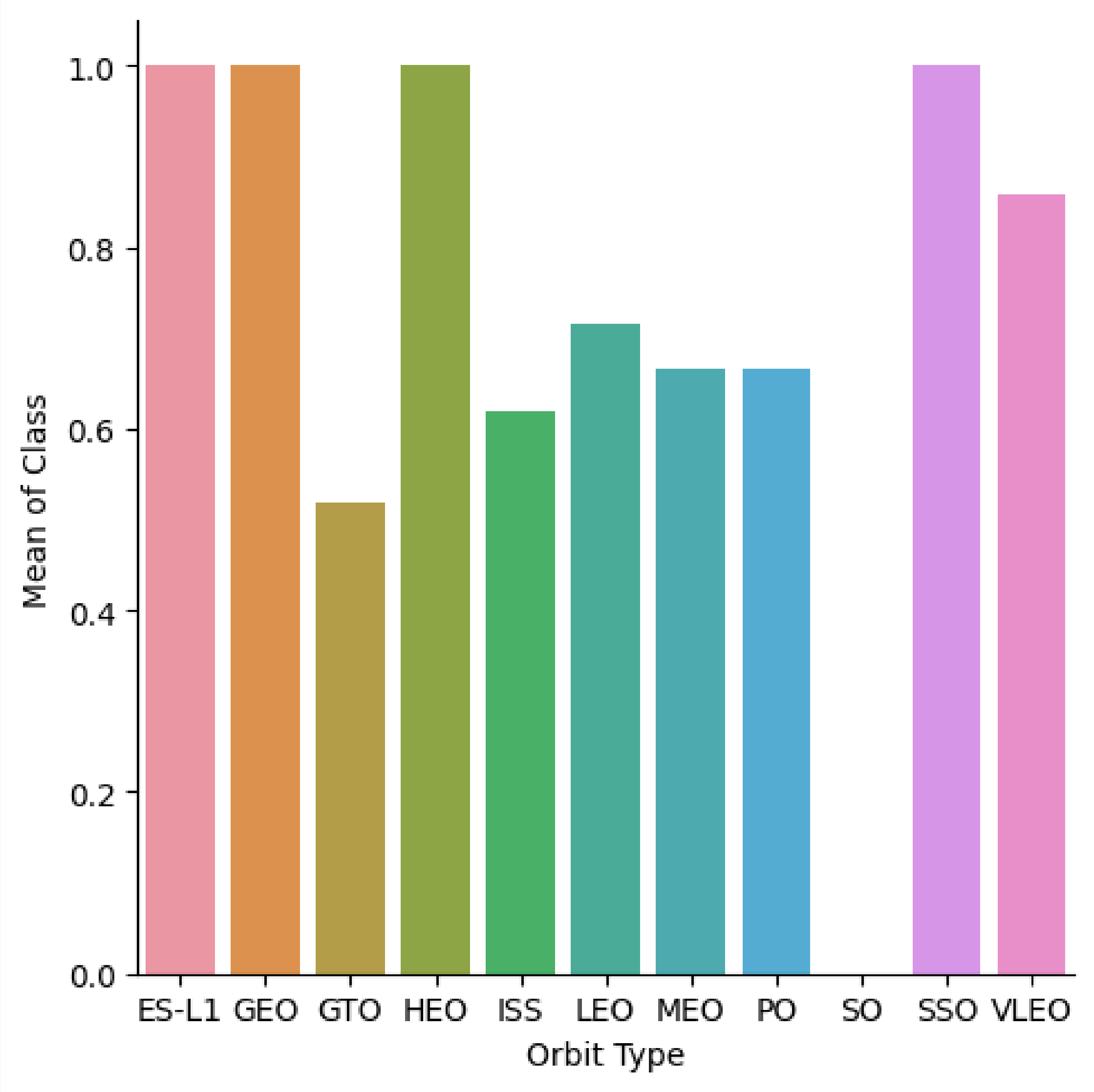
Payload vs. Launch Site



Shows a scatter plot of Payload vs. Launch Site

It can be observed that the Launch sites CCAPFS SLC 40 & KSC LC39A appear to have better success outcomes for heavy payload than the rest.

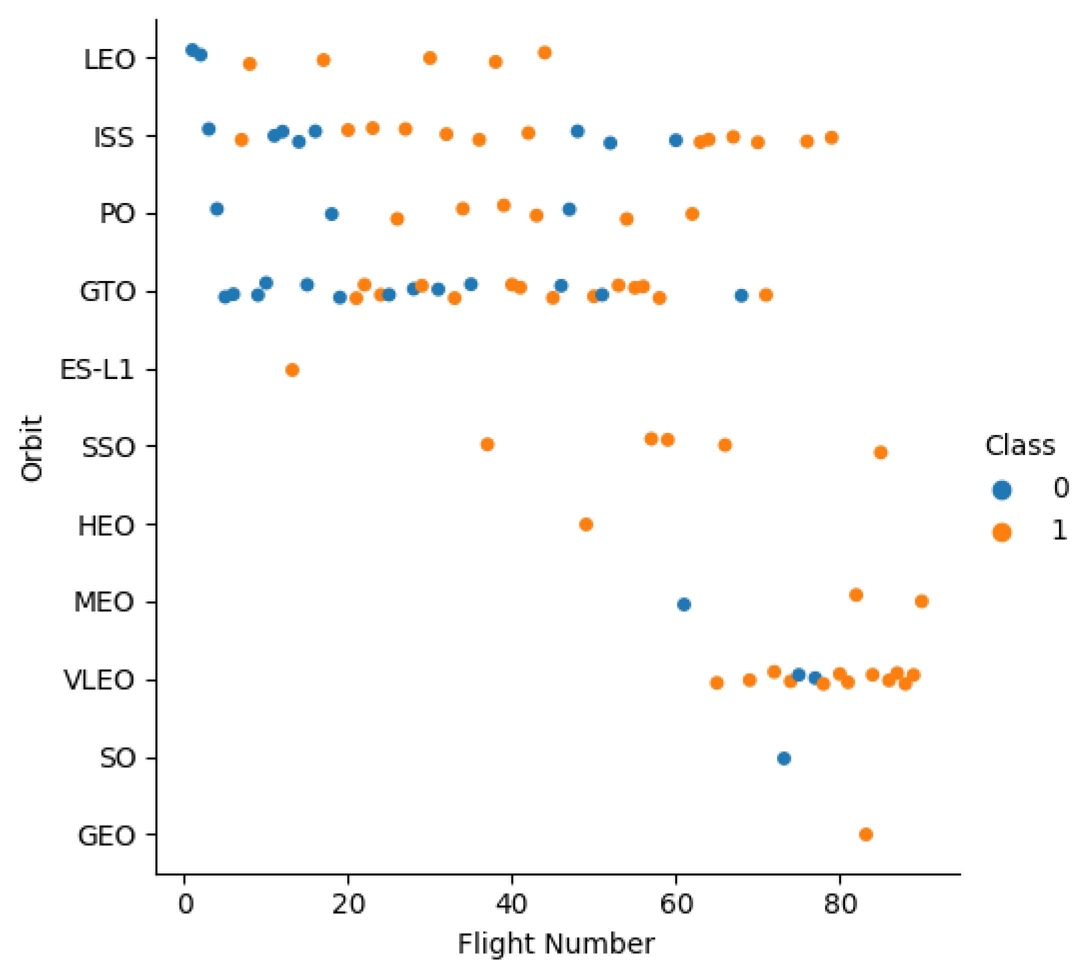
Success Rate vs. Orbit Type



- Shows a bar chart for the success rate of each orbit type

More successful outcome can be observed for E-L1, GEO, HEO, SSO and VLEO

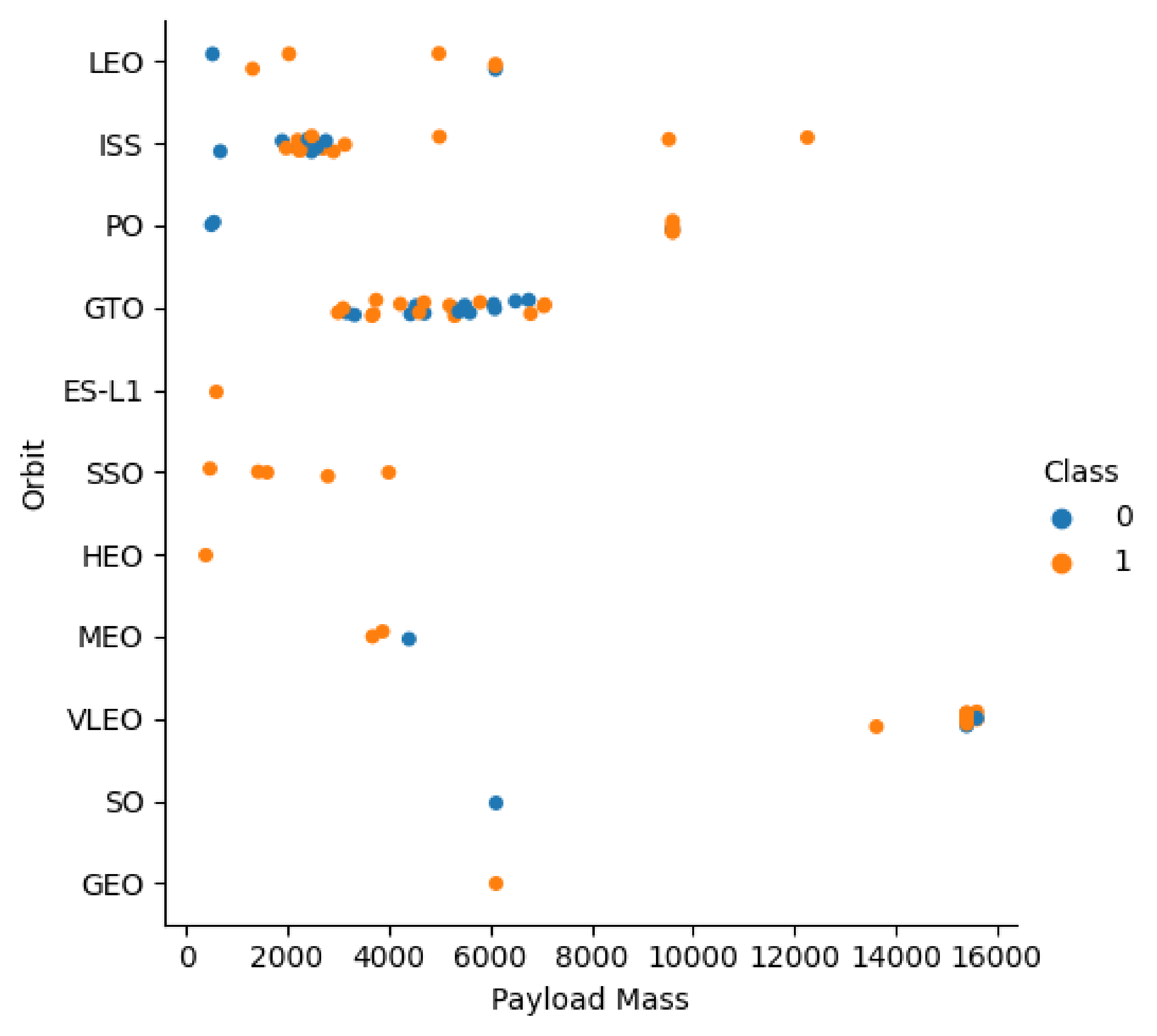
Flight Number vs. Orbit Type



Shows a scatter point of Flight number vs. Orbit type

- From the observation, in the LEO orbit, the Success appears related to the number of (early) flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

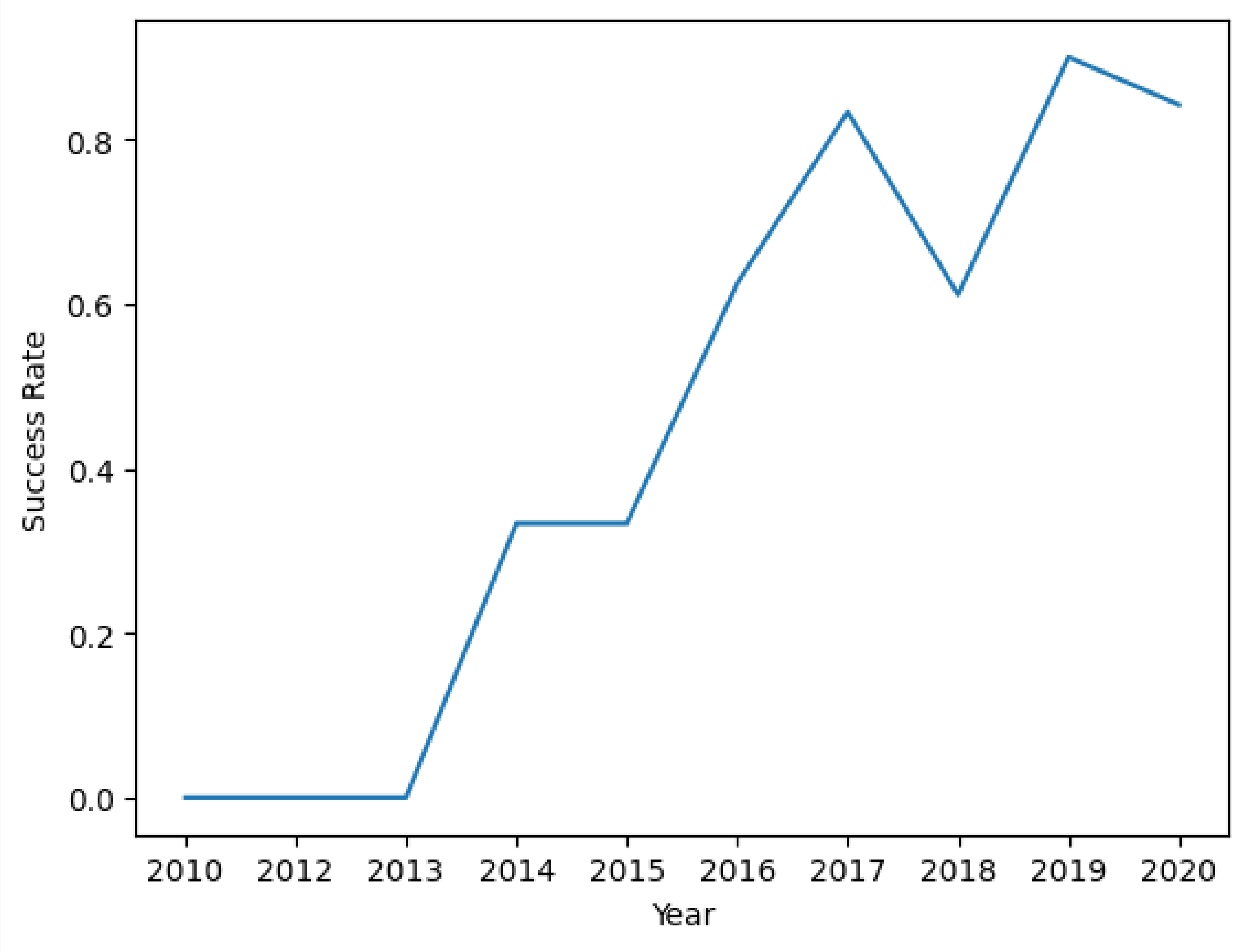
Payload vs. Orbit Type



Shows a scatter point of payload vs. orbit type

- LEO had success on the opposite end of the loads
- ISS had many successful outcomes in the 500 to 2500 payload range.
- PO only had success with lower payloads
- VLEO seems to be the only orbit with higher heavy-load (16,000kg) success
- However, for GTO we cannot distinguish this well as both a positive landing rate and a negative landing(unsuccessful mission).

Launch Success Yearly Trend



- Shows a line chart of the yearly average success rate
- It can be observed that there is an increasing trend of successful outcome apar from the dip in 2018

All Launch Site Names

```
%%sql
select distinct Launch_Site from SPACEXTBL
```

✓ 0.0s

* sqlite:///my_data1.db
Done.

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

- Finds the names of the unique launch sites

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL where Launch_site like 'CCA%' Limit 5
```

[9] ✓ 0.0s Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

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

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_)  from SPACEXTBL where Customer = 'NASA (CRS)'
```

✓ 0.0s

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

sum(PAYLOAD_MASS__KG_)
45596.0

- Calculates the total payload carried by boosters from NASA

Average Payload Mass by F9 v1.1

```
%sql select sum(PAYLOAD_MASS__KG_)  from SPACEXTBL where Booster_Version = 'F9 v1.1'
✓ 0.0s

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

sum(PAYLOAD_MASS__KG_)
14642.0

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

First Successful Ground Landing Date

Hint:Use min function

```
%sql Select MIN(Date) from SPACEXTBL where Landing_Outcome='Success (ground pad)'
```

✓ 0.0s

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

MIN(Date)
01/08/2018

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

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql Select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ between 4000 and 6000 and Landing_Outcome='Success (drone ship)'
```

✓ 0.0s

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

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

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

Total Number of Successful and Failure Mission Outcomes

```
%sql Select trim(Mission_Outcome) as 'Mission Outcomes', Count(*) as 'Count' from SPACEXTBL group by  trim(Mission_Outcome)
```

[23] ✓ 0.0s

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

Mission Outcomes	Count
None	898
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Calculates the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
%sql Select Booster_Version, PAYLOAD_MASS__KG_ \
from SPACEXTBL where  PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_)  from SPACEXTBL)
```

✓ 0.0s

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

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

2015 Launch Records

```
%sql select substr(Date,7,4) "Year", substr(Date,4,2) "Month", Landing_Outcome, Booster_Version, Launch_Site  \
from SPACEXTBL where substr(Date,7,4) = '2015' and Landing_Outcome='Failure (drone ship)'
```

✓ 0.0s

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

Year	Month	Landing_Outcome	Booster_Version	Launch_Site
2015	10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Lists 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'
from SPACEXTBL  where (substr(Date,7,4)||substr(Date,4,2) ||
substr(Date,1,2))   between '20100604' and '20170320'
group by Landing_Outcome  order by 2 desc
```

✓ 0.0s

* sqlite:///my_data1.db
Done.

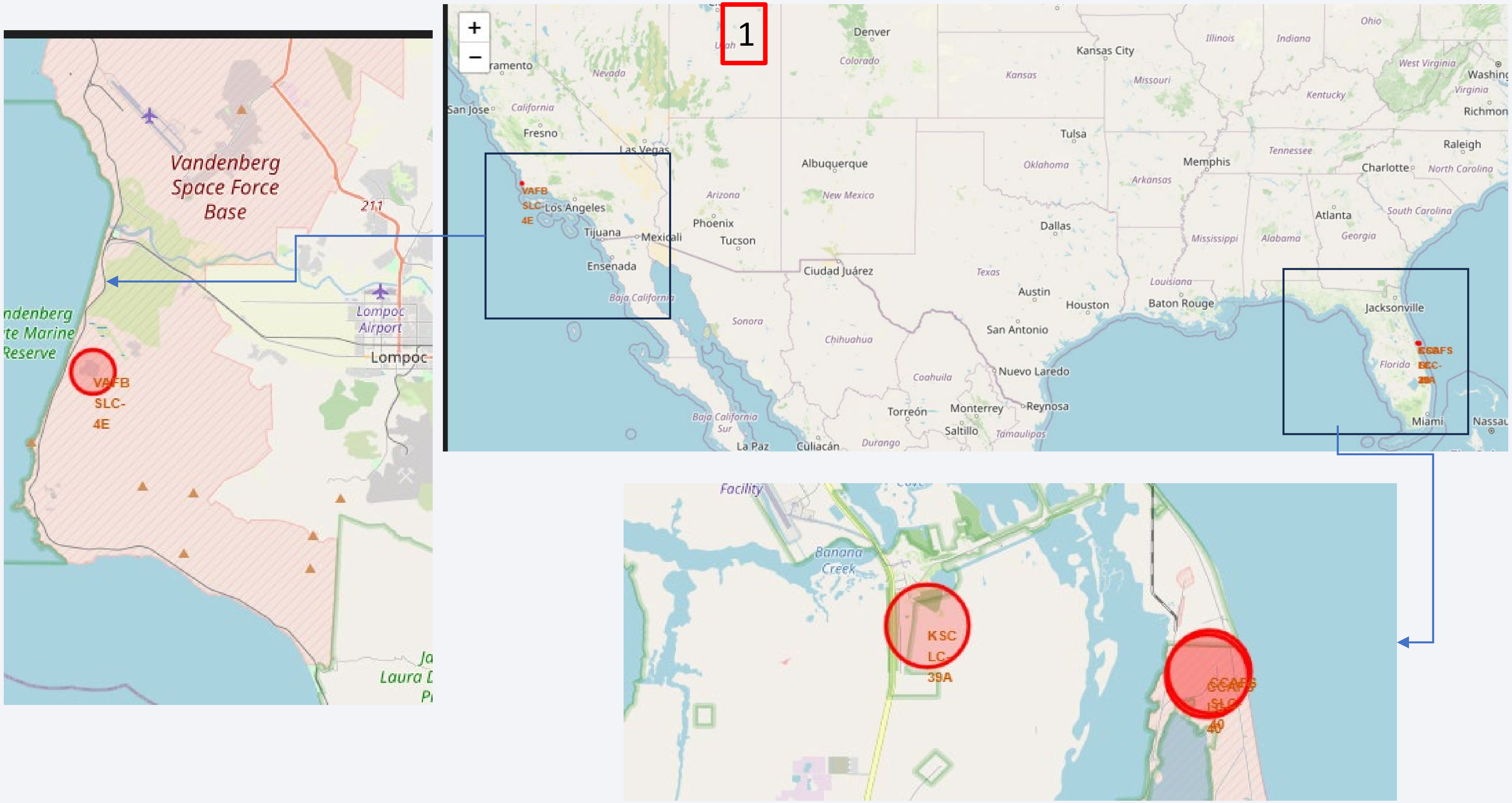
Landing_Outcome	Count
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

- 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

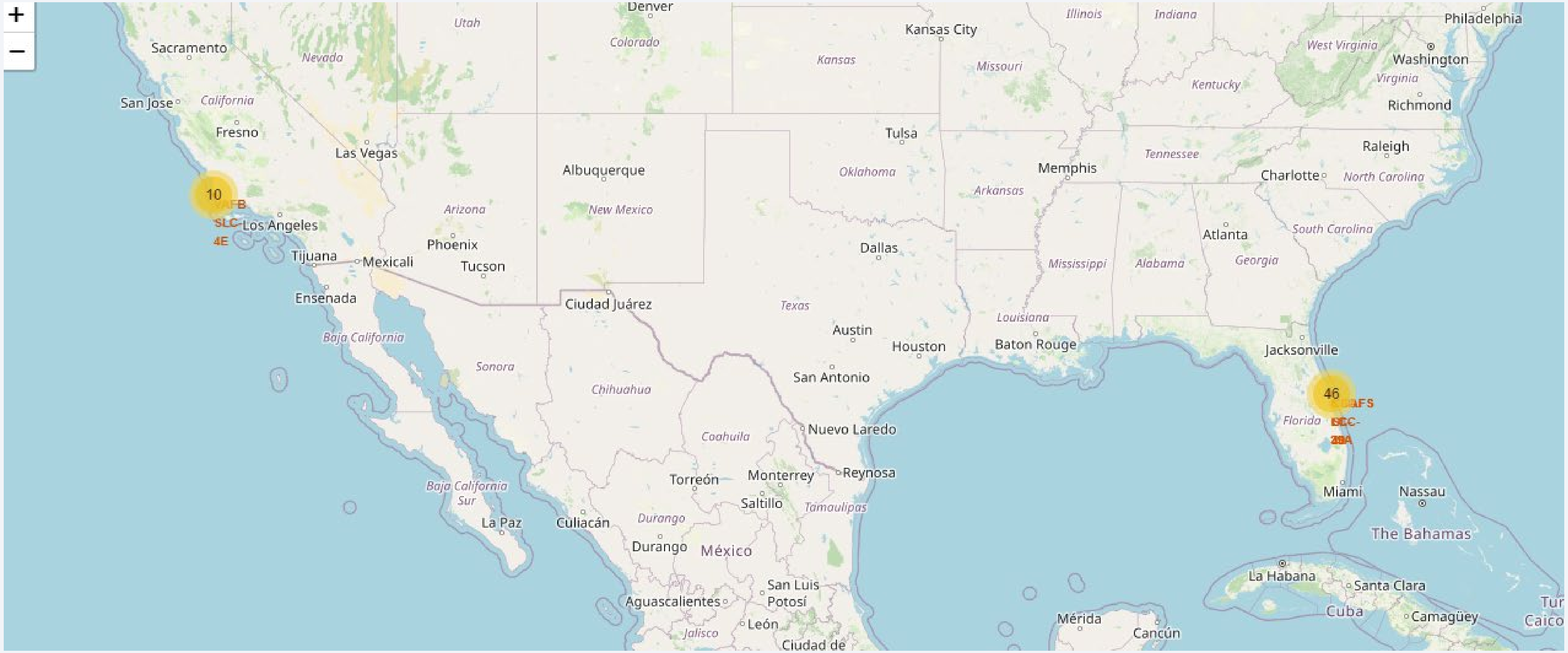
Launch Sites Proximities Analysis

All Launch Sites Map



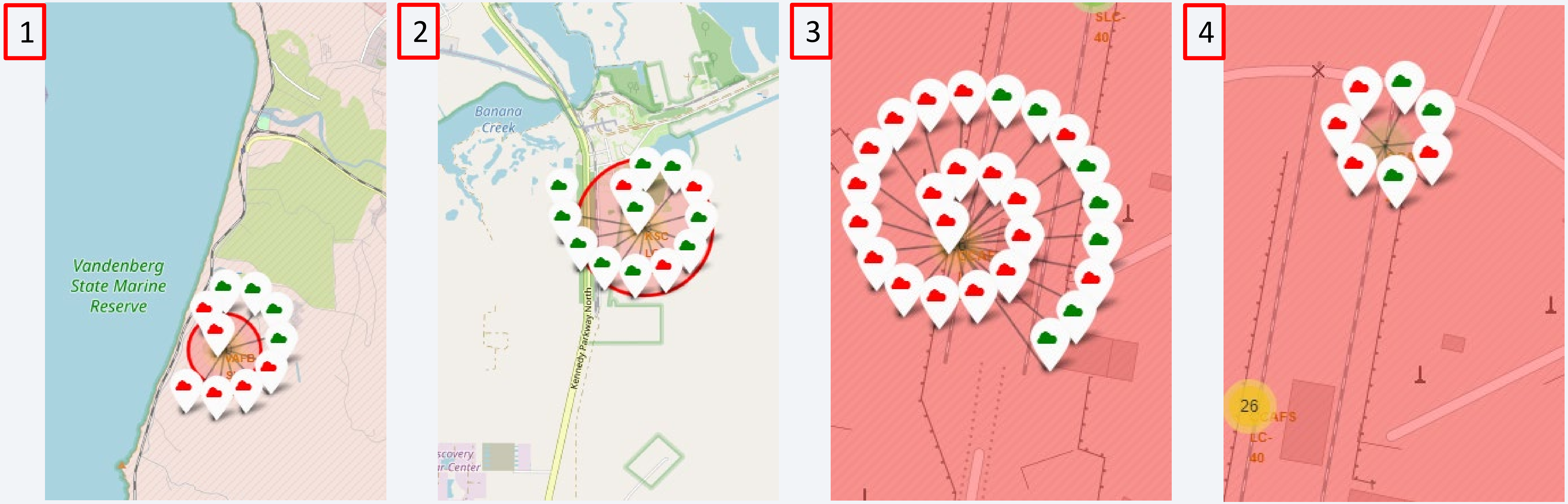
The map [1] shows the 4 launch sites. It is observed that the launch sites are located close to the coastal waters. One of the sites (VAFB SLC-4E) is on the west coast (Los Angeles) and the remaining 3 (KSC LC-39A, CCAFS LC-40, CCAFS SLC-40) are on the east coast (Florida) and in proximity.

Launch Sites Outcomes Map / 2



The map shows launch sites and the number of launches per each site: 10 launches in the Los Angeles area and 46 launches in the Florida area. A drill-down of the outcomes will be presented in the next slide….

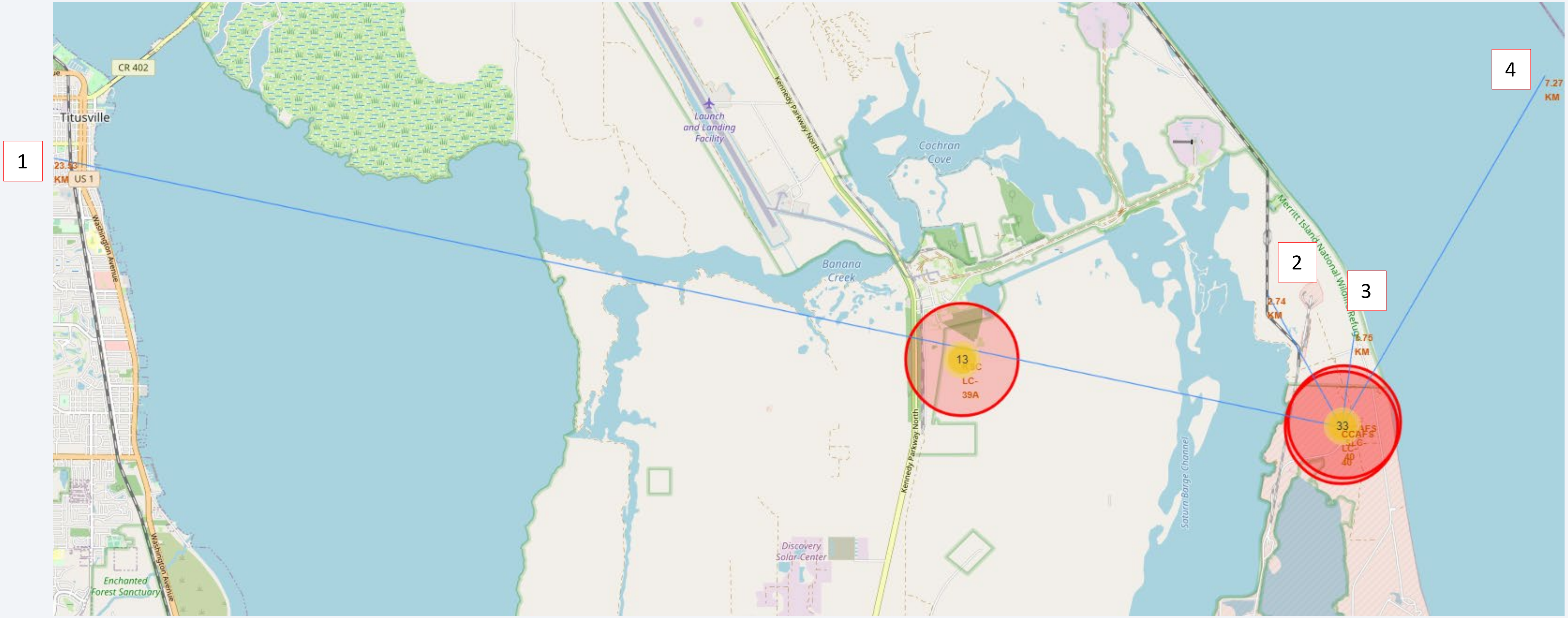
Launch Sites Outcomes Map



This map shows the launch outcomes icons (success = green | failure = red)

[1] VAFB SLC-4E - 10 launches (6 failures, 4 successes)
[2] KSC LC-39A - 13 launches (3 failures, 8 successes)
[3] CCAFS LC-40 - 26 launches (19 failures, 7 successes)
[4] CCAFS SLC-40 – 7 launches (4 failures, 3 successes)

CCAFS LC-40 Launch Site & Distance To Specific Landmarks

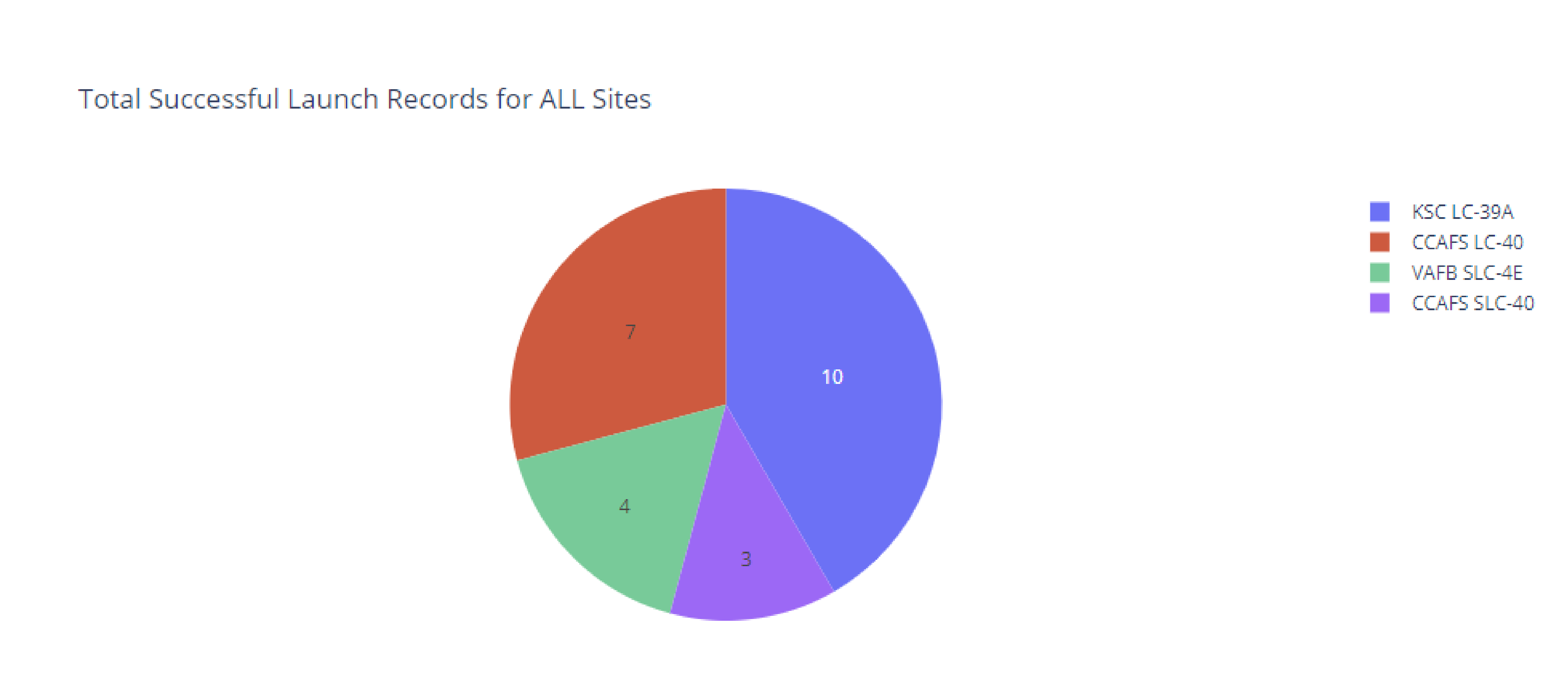


[1] Titusville City (23.53km) [2] Railway line (2.74km) [3] Samuel C Philips Parkway (1.75KM) [4] Florida Coastline (7.27KM)
One of the key observations is the proximity of the site to infrastructures & water, but a safe distance from settlements

Section 4

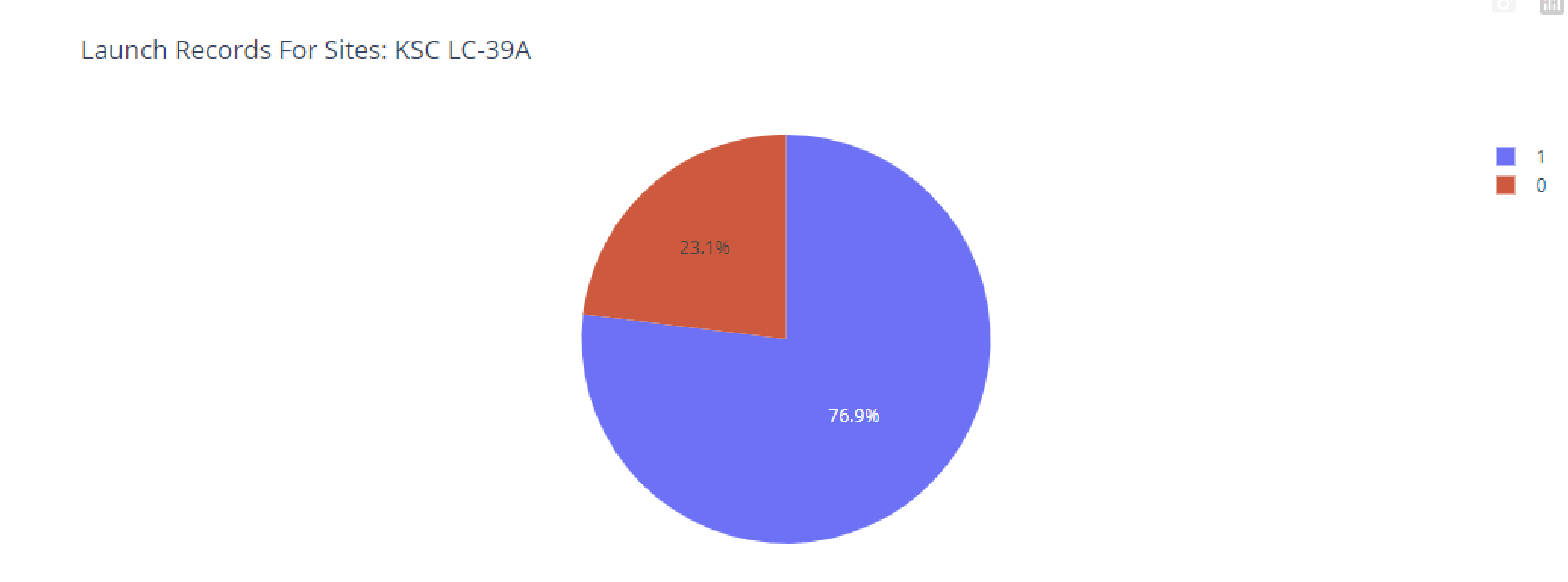
Build a Dashboard with Plotly Dash

Launch Sites Successful Launch Counts



KSC LC-39A and CCAFS LC-40 launch sites have more successful launch counts (10 and 7 respectively) than the remaining 2.

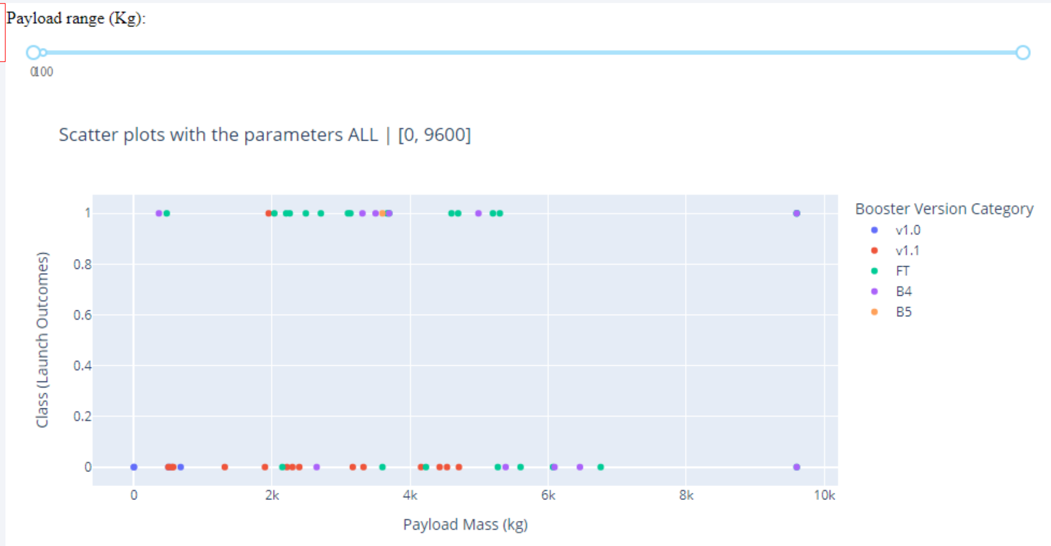
KSC LC-39A Launch Sites Outcome Counts



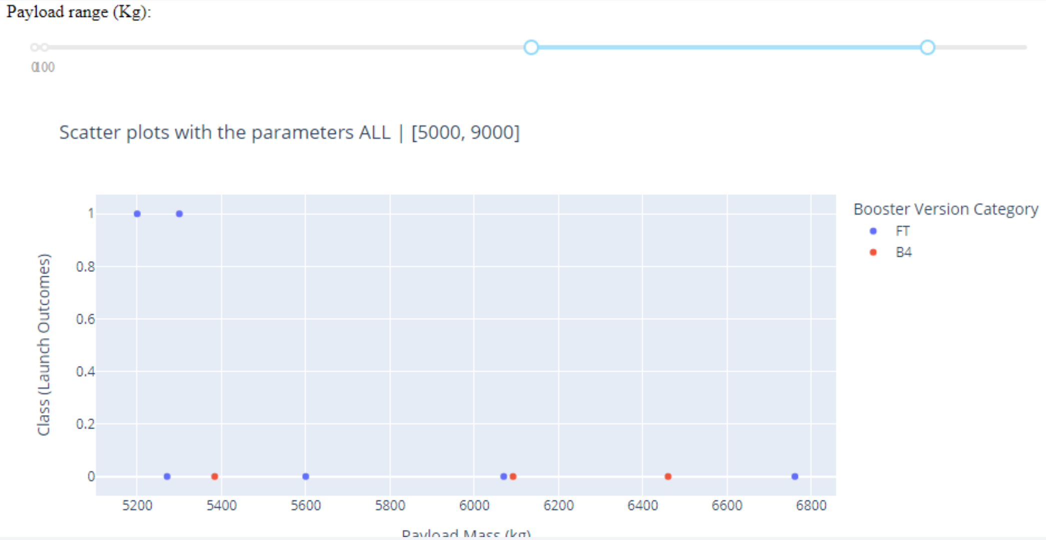
- KSC LC-39A has the highest success ratio based on the % a well as the count

Payload vs. Launch Outcome With Payload Ranges

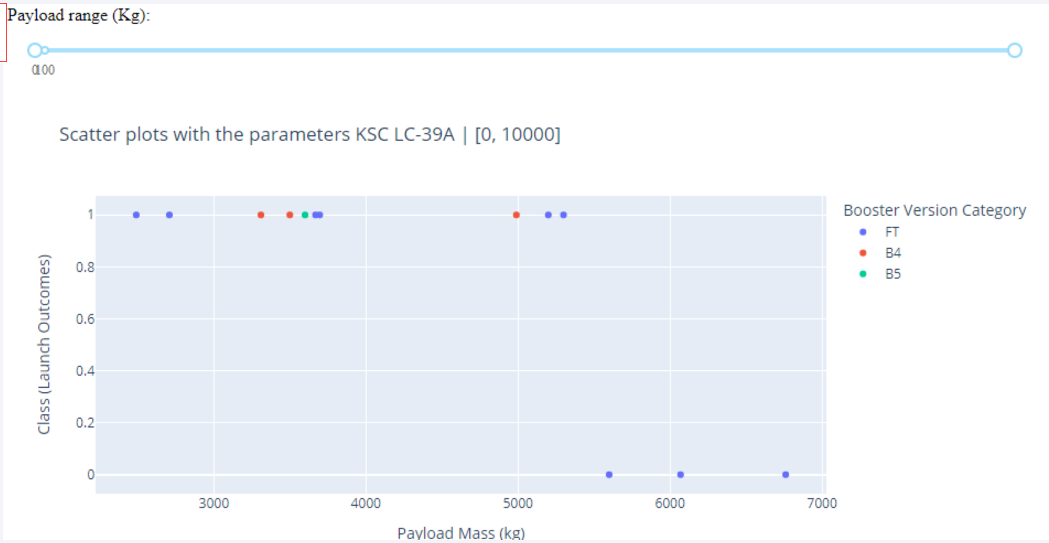
1



2



3



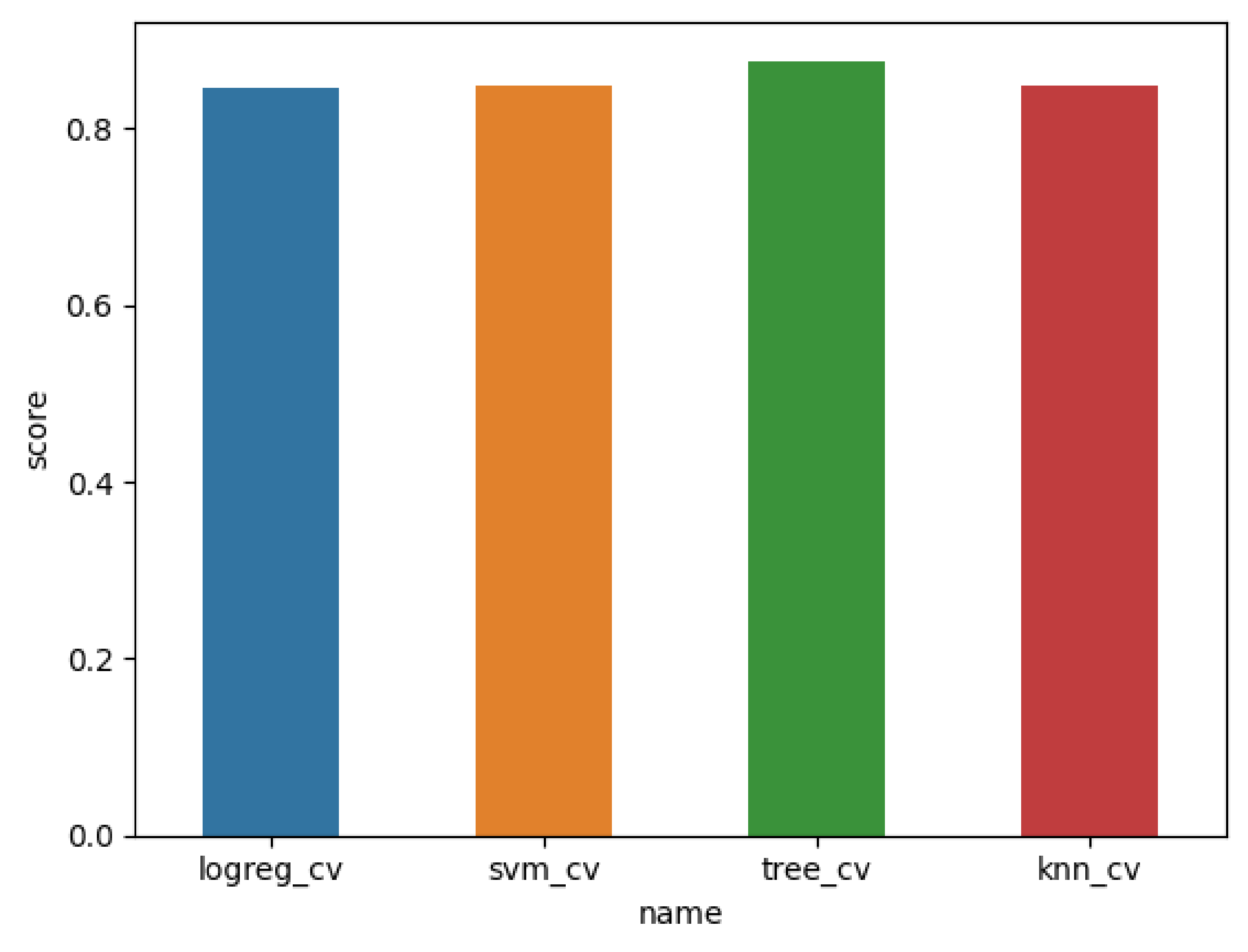
Notes:

[1] All Sites and payload range
[2] – Only FT & B4 booster version categories are within the 5000 to 9000 payloads
[3] For the site with the highest success rate, it is observed that successful launches happen at payloads lower than 5500kg. Payloads above 5000kg all failed.

Section 5

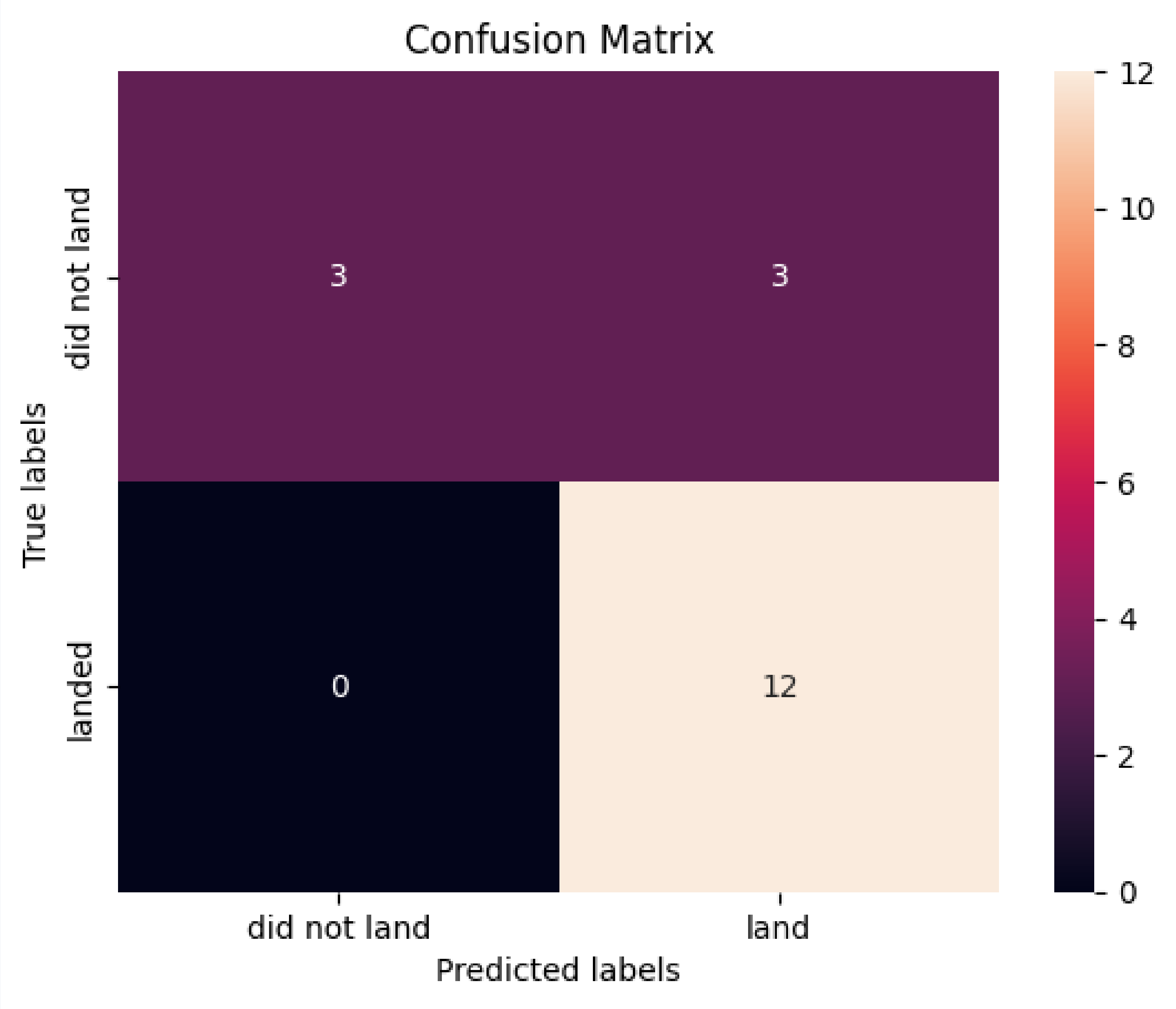
Predictive Analysis (Classification)

Classification Accuracy



- From the diagram, tree_cv is the model with the highest classification accuracy.

Confusion Matrix – tree_cv



The tree_cv confusion matrix shows 100% accuracy in predicting success outcomes. However, for the failure outcome, the accuracy was 50%. Which means we have a 50% false negative.

Conclusions

- **Increasing Reusability**: Over the years, we have observed a consistent upward trend in the rate of reusability, indicating progress in our efforts to reuse space launch vehicles effectively.
- **Challenges in Payload Landing**: Despite advancements, successfully landing heavier payloads remains a significant challenge. Further research and development are needed to address this issue.
- **Data Science for Improved Landing Success**: Leveraging data science, particularly the **tree_cv** technique, we can significantly enhance the likelihood of successful landings. Our predictive model demonstrates an 83% accuracy rate, showing promising results for future missions.
- **Continuous Improvement and Assessment**: While our current model shows promise, it is crucial to continuously gather more data and monitor the model's performance over time. Regular assessments will help identify potential degradation and ensure that the model maintains the highest level of accuracy as our dataset grows....

Appendix

devtage-goladeji . 48b7c11 1 hour ago 23 commits

Submission Overview and Instructions	.	1 hour ago
reference	adding new files	2 days ago
5_Peer_Graded_Assignment_Questio...	Add files via upload	last week
Basics of PowerPoint.pdf	adding new files	2 days ago
Basics of PowerPoint.pptx	adding new files	2 days ago
CapStoneProject-01.ipynb	Master (#1)	2 days ago
Getting_Started_With_Data_science-R...	adding new files	2 days ago
IBM-DS0321EN-SkillsNetwork_labs_...	.	yesterday
IBM-DS0321EN-SkillsNetwork_labs_...	.	7 hours ago
IBM-DS0321EN-SkillsNetwork_labs_...	.	1 hour ago
README.md	Update README.md	last week
Spacex.csv	adding new files	2 days ago
dataset_part_1.csv	changed file	2 days ago
dataset_part_2.csv	new file: dataset_part_2.csv	2 days ago
dataset_part_3.csv	modified: IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-ed...	2 days ago
file001.py	Create file001.py	last week
jupyter-labs-eda-sql-coursera_sqllite....	.	yesterday
jupyter-labs-webscraping.ipynb	modified: jupyter-labs-webscraping.ipynb	2 days ago
labs-jupyter-spacex-Data-wrangling.i...	new file: dataset_part_2.csv	2 days ago
my_data1.db	.	yesterday
spacex-data.ipynb	modified: Submission Overview and Instructions/ds-capstone-template-...	2 days ago
spacex_dash_app.py	.	4 hours ago
spacex_launch_dash.csv	.	yesterday
spacex_web_scraped.csv	web scrapping process run and file created	2 days ago

References

- [1] - SpaceX Challenge Has Arianespace Rethinking Pricing Policies – SpaceNews
- [2] - Space launch market competition – Wikipedia
- [3] – https://en.wikipedia.org/wiki/Space_launch_market_competition
- [4] -

Thank you!