IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

LEANDRO MANASSI PANITZ
29st September 2023

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data Collection through API and Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Interactive Dashboard with Ploty Dash
  - Machine Learning Prediction
- **Summary of all results**
  - Exploratory Data Analysis (EDA) results
  - Interactive Maps and Dashboard
  - Predictive Analytics results

Introduction

- **Project background and context**

SpaceX promotes Falcon 9 rocket launches on its website at a price of $62 million. In contrast, other providers charge over $165 million per launch. A significant portion of these savings comes from SpaceX's ability to reuse the first stage of the rocket. With this in mind, the objective of this project is to analyze data from Falcon 9 launches and develop a predictive machine learning model to determine whether the first stage will land successfully. If we can predict this outcome, we can better estimate the cost of a launch. Such insights would be invaluable for any competing company looking to bid against SpaceX for rocket launch contracts.

- **Problems you want to find answers**
    - The factors that determine whether a rocket will land successfully.
    - The interaction of the various factors that determine the success rate of a successful landing.
    - The ideal operating conditions to achieve the best landing success rate.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collected through the SpaceX REST API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - Data filtering, dealing with missing values, data preparation with creation of new variables
- 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, tuning and evaluation of classification models to guarantee optimal outcomes

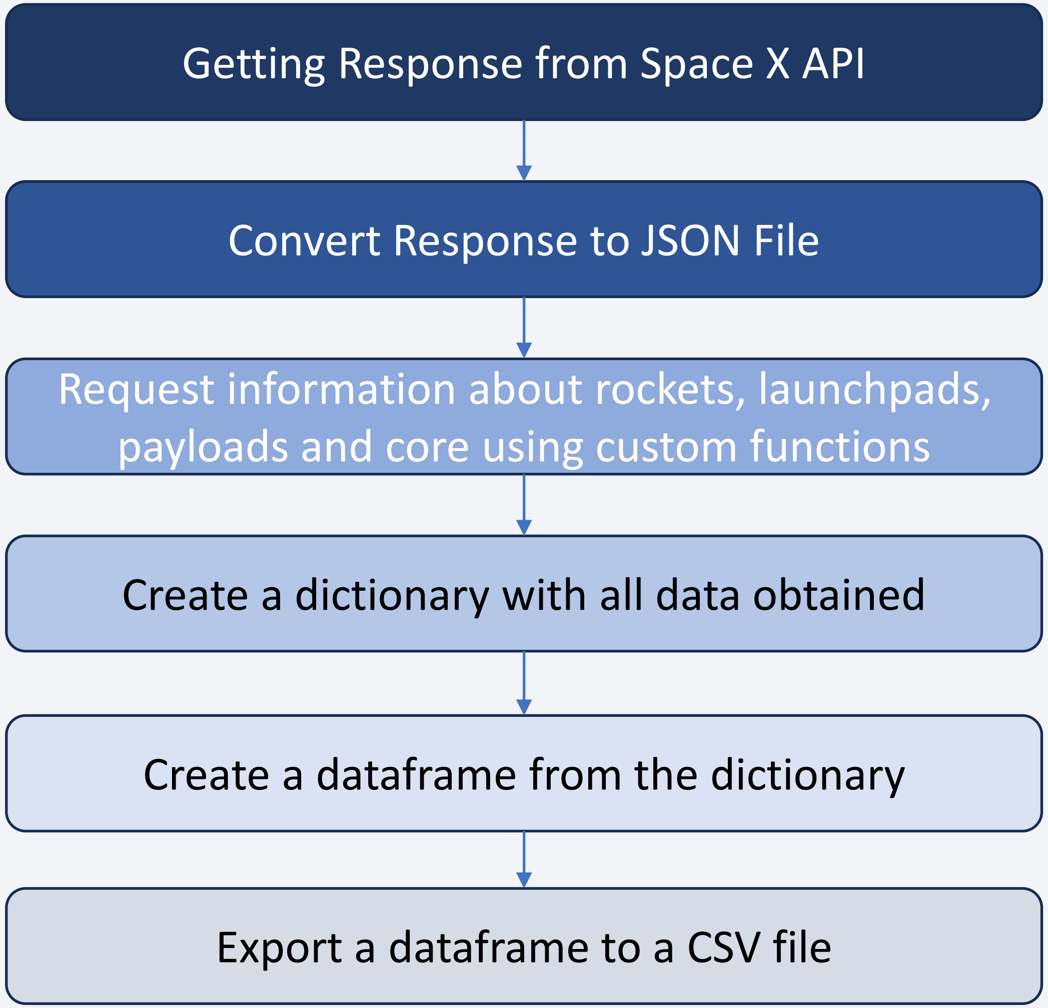
Data Collection

- Data sets collected:
  - **SpaceX API**: Open Source REST API for launch, rocket, core, launchpad, and landing pad data
    - https://api.spacexdata.com/v4/
  - **Wikipedia**: List of Falcon 9 and Falcon Heavy launches
    - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- The data was obtained from the Open SpaceX API following the steps shown in the flowchart presented

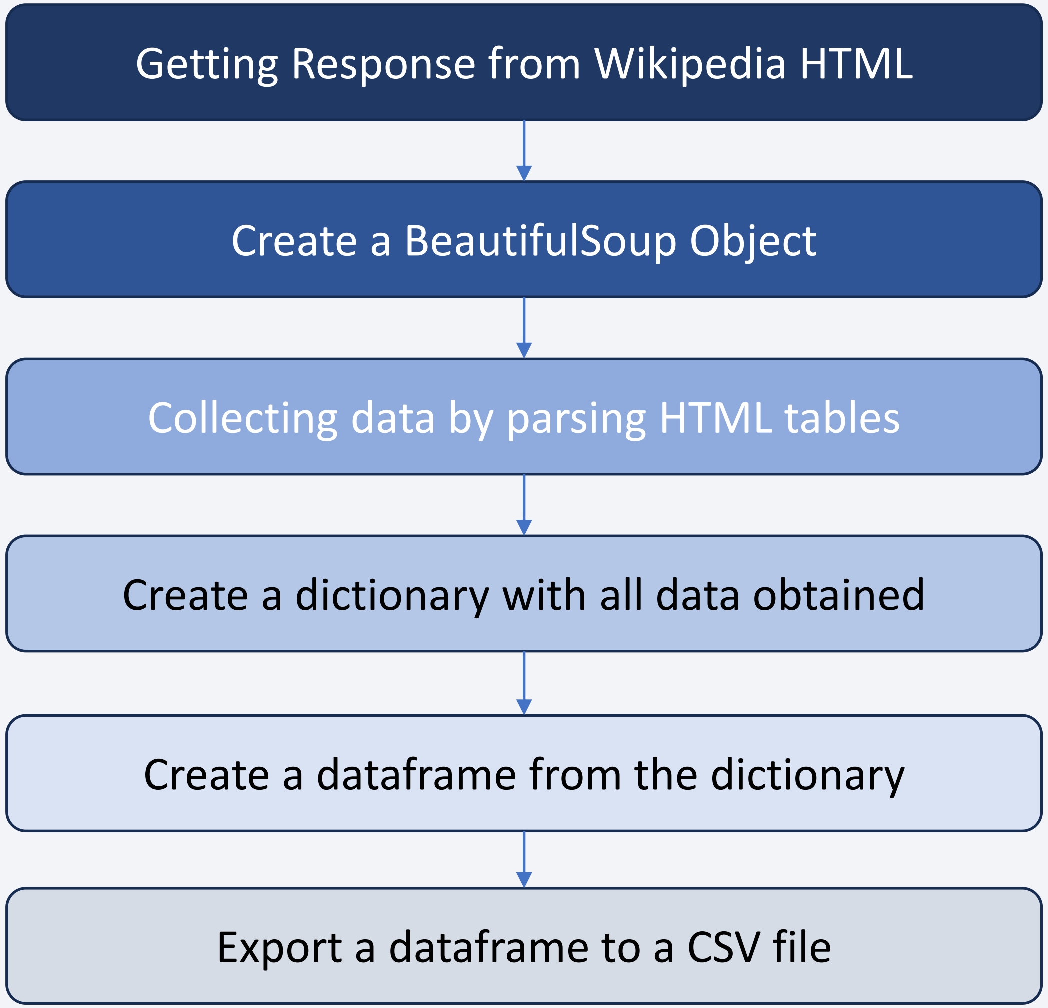
- GitHub URL of the completed SpaceX API calls notebook: https://github.com/lepanitz/IBM---Final-Project/blob/main/01%20jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- The data was obtained from the Wikipedia Webpage following the steps shown in the flowchart presented

- GitHub URL of the completed Wikipedia Web scraping notebook: https://github.com/lepanitz/IBM---Final-Project/blob/main/02%20jupyter-labs-webscraping.ipynb



Data Wrangling

- Through Data Wrangling, the data was cleaned, unified and organized and Exploratory Data Analysis (EDA) was carried out

- GitHub URL of the completed Data Wrangling notebook: https://github.com/lepanitz/IBM---Final-Project/blob/main/03%20labs-jupyter-spacex-Data%20wrangling.ipynb

Imported the dataframe created from the SpaceX API (dataset_part_1.csv)

↓

Calculated the number of launches on each site

↓

Calculated the number and occurrence of each orbit

↓

Calculated the number and occurence of mission outcome of the orbits

↓

Created a landing outcome label from Outcome column

↓

Exported a dataframe to a CSV file (dataset_part_2.csv)

EDA with Data Visualization

- Charts Plotted:
  - Relationship between Flight Number and Launch Site
  - Relationship between Payload and Launch Site
  - Relationship between success rate of each orbit type
  - Relationship between Flight Number and Orbit type
  - Relationship between Payload and Orbit type
  - Launch success yearly trend
- Through the plotted graphs it was possible to visualize the existence or not of relationships between the variables in order to understand how they affect successful landings.
- GitHub URL of the completed EDA with Visualization notebook: https://github.com/lepanitz/IBM---Final-Project/blob/main/05%20jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- SQL queries performed:
  - Names of the unique launch sites in the space mission
  - 5 records where launch sites begin with the string 'CCA’
  - The total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first succesful landing outcome in ground pad was achieved
  - Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Total number of successful and failure mission outcomes
  - Names of the booster_versions which have carried the maximum payload mass. Use a subquery
  - Records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
  - 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

- We chose to use PostgreSQL to persist the data on a local machine.

- GitHub URL of the completed EDA with SQL notebook: https://github.com/lepanitz/IBM---Final-Project/blob/main/04%20jupyter-labs-eda-sql-coursera_sqllite_postgree.ipynb

Build an Interactive Map with Folium

- **Mapping Launch Locations**: All launch sites have been pinpointed on our interactive map.
- **Integration of Map Objects**: To depict the success or failure of each launch, we employed a mix of markers, circles, and lines on the folium map.
- **Categorizing Launch Outcomes**: We designated values to classify the outcomes of the launches: 0 stands for a failure, whereas 1 signifies success.
- **Visual Identification of Performance**: Through color-labeled marker clusters (RED and GREEN), we can clearly discern which launch sites boast a higher success rate.
- **Distance Analysis**: We computed the distances between a launch site and its nearby locations, offering insights about its strategic placement and potential influencing factors on launch success.

- GitHub URL of the completed Analysis with Folium notebook: https://github.com/lepanitz/IBM---Final-Project/blob/main/06%20lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- **Dashboard Creation**: Leveraged Plotly Dash to develop a dynamic and user-friendly interface.
- **Launch Site Analysis**: Introduced pie charts to depict the distribution of total launches across specific sites, allowing for a comparative view.
- **Exploring Outcome vs. Payload Dynamics**: Utilized scatter graphs to visualize the correlation between launch outcomes and payload mass in kilograms, segmented further by various booster versions.
- **Diverse Visualization Tools**: Integrated both categorical and continuous data types for a comprehensive understanding of launch patterns.

- GitHub URL of the completed Dashboard with Plotly Dash notebook: https://github.com/lepanitz/IBM---Final-Project/blob/main/07%20spacex_dash_app.py

Predictive Analysis (Classification)

- **Data Preparation**: Loaded datasets using numpy and pandas, followed by necessary transformations.
- **Data Splitting**: Segmented the data into training and testing sets to validate the model's performance.
- **Model Building**: Developed various machine learning models and optimized hyperparameters using GridSearchCV.
- **Evaluation Metric**: Employed accuracy as the primary metric to assess and refine our model's performance.
- **Model Optimization**: Continual enhancement of the model through feature engineering and algorithm tuning.
- **Highlighting the Best Model**: Identified and spotlighted the top-performing classification model after all adjustments.

- GitHub URL of the completed Predictive Analysis notebook: https://github.com/lepanitz/IBM---Final-Project/blob/main/08%20SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

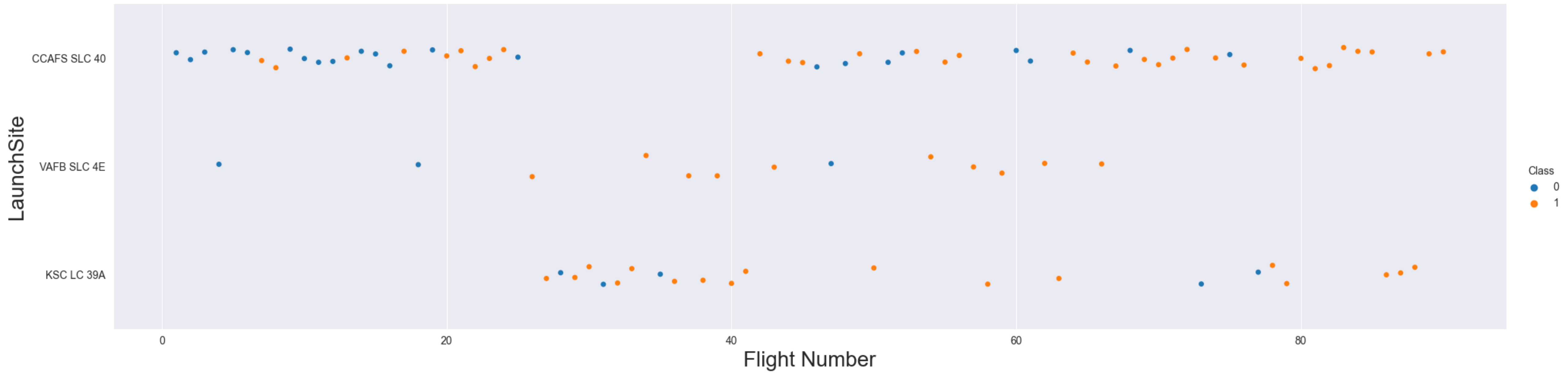
Outcomes are divided into three primary categories:

- **Initial Insights**: Conclusions drawn from exploratory data analysis.
- **Visual Demonstrations**: Screenshots showcasing the interactive analytics experience.
- **Predictive Outcomes**: Results and findings from the predictive analysis phase.

Section 2

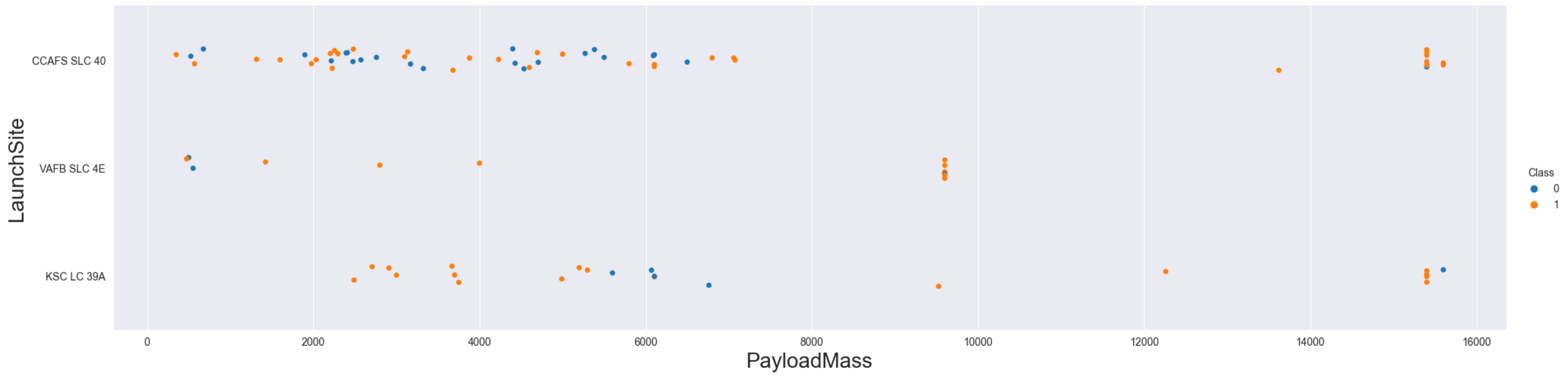
Insights drawn from EDA

Flight Number vs. Launch Site



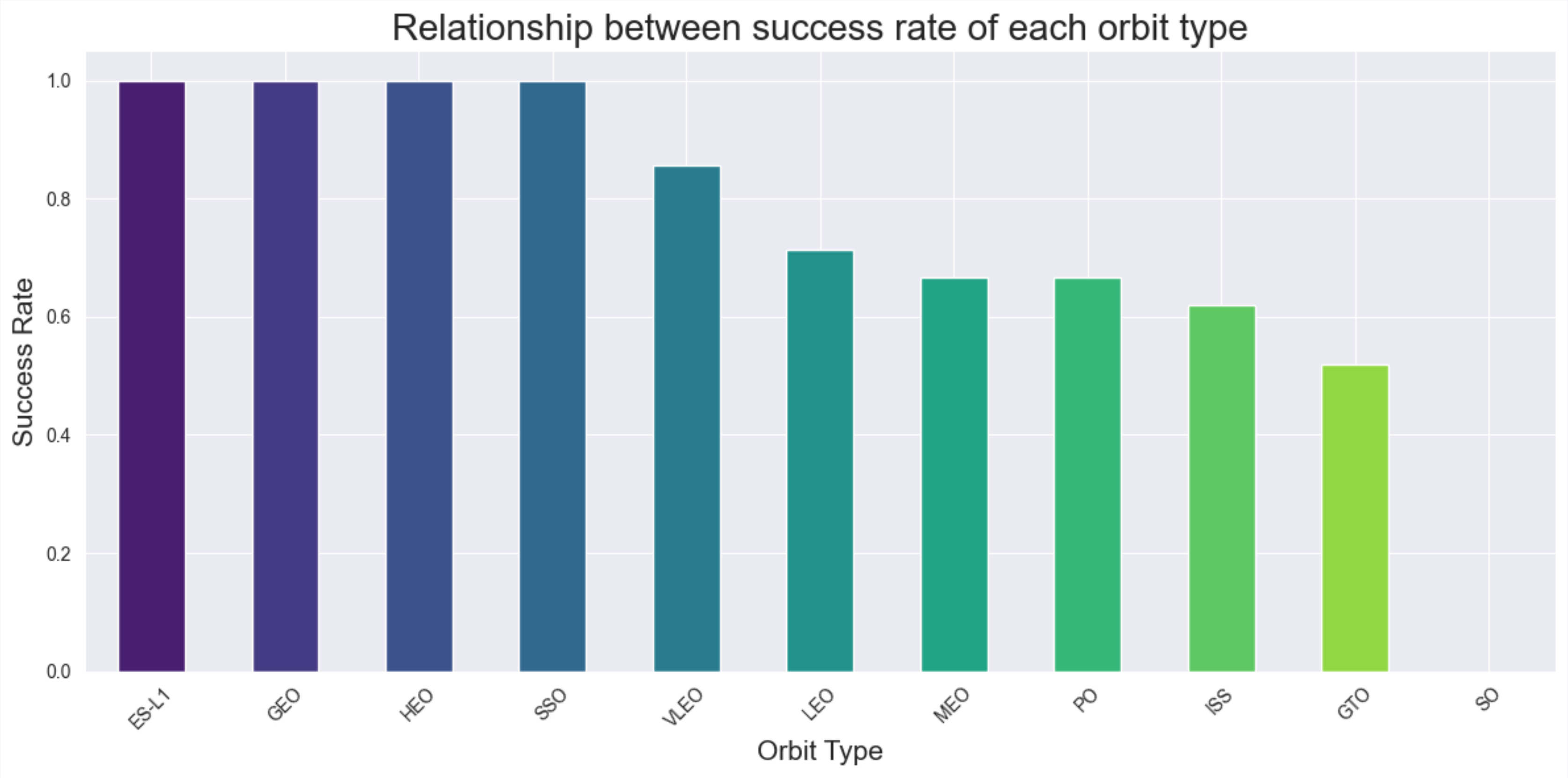
- The success rate increases with the number of flights.
- Launch sites with a higher number of flights tend to have a greater success rate.

Payload vs. Launch Site



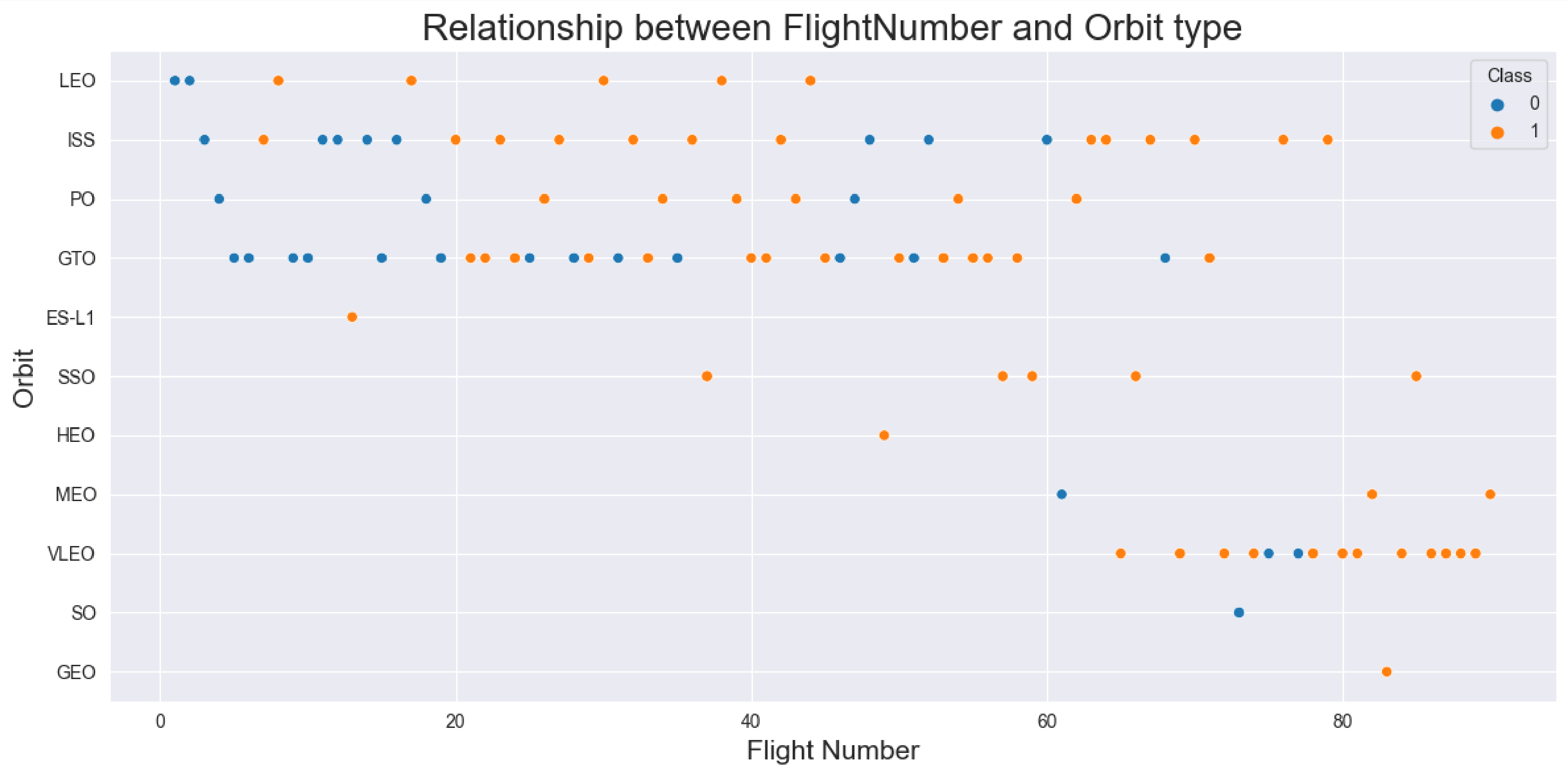
- The KSC LC 39A and CCAFS SLC 40 launch sites are more suitable for payloads greater than 12,000 kg.
- Independente do local de lançamento, voos com carga maior do que 8.000 tem uma taxa de sucesso de lançamento maior.

Success Rate vs. Orbit Type



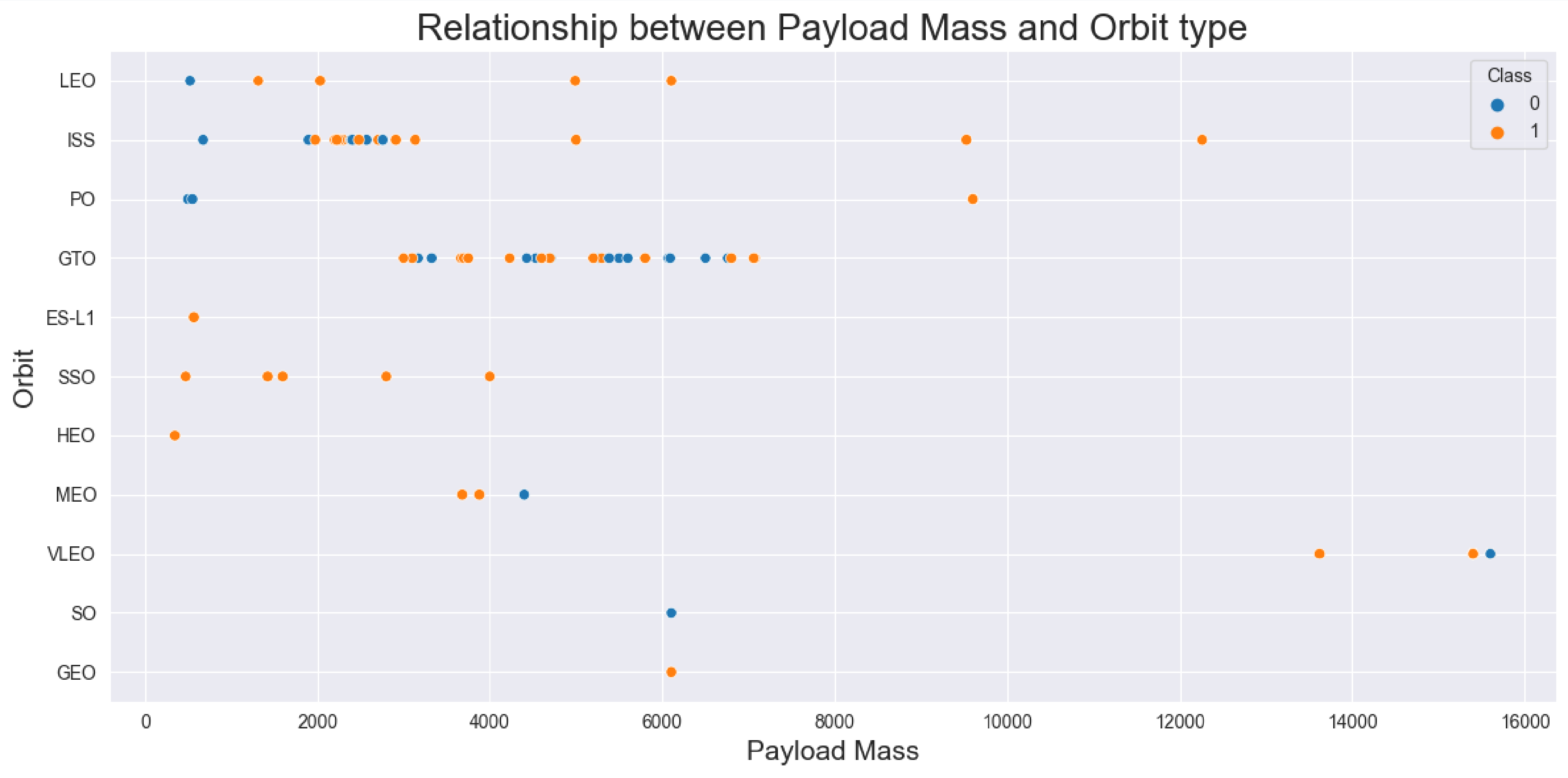
- Das 11 órbitas possíveis, 4 órbitas (ES-L1, GEO, HEO e SSO) apresentam uma de taxa de sucesso de 100%.

Flight Number vs. Orbit Type



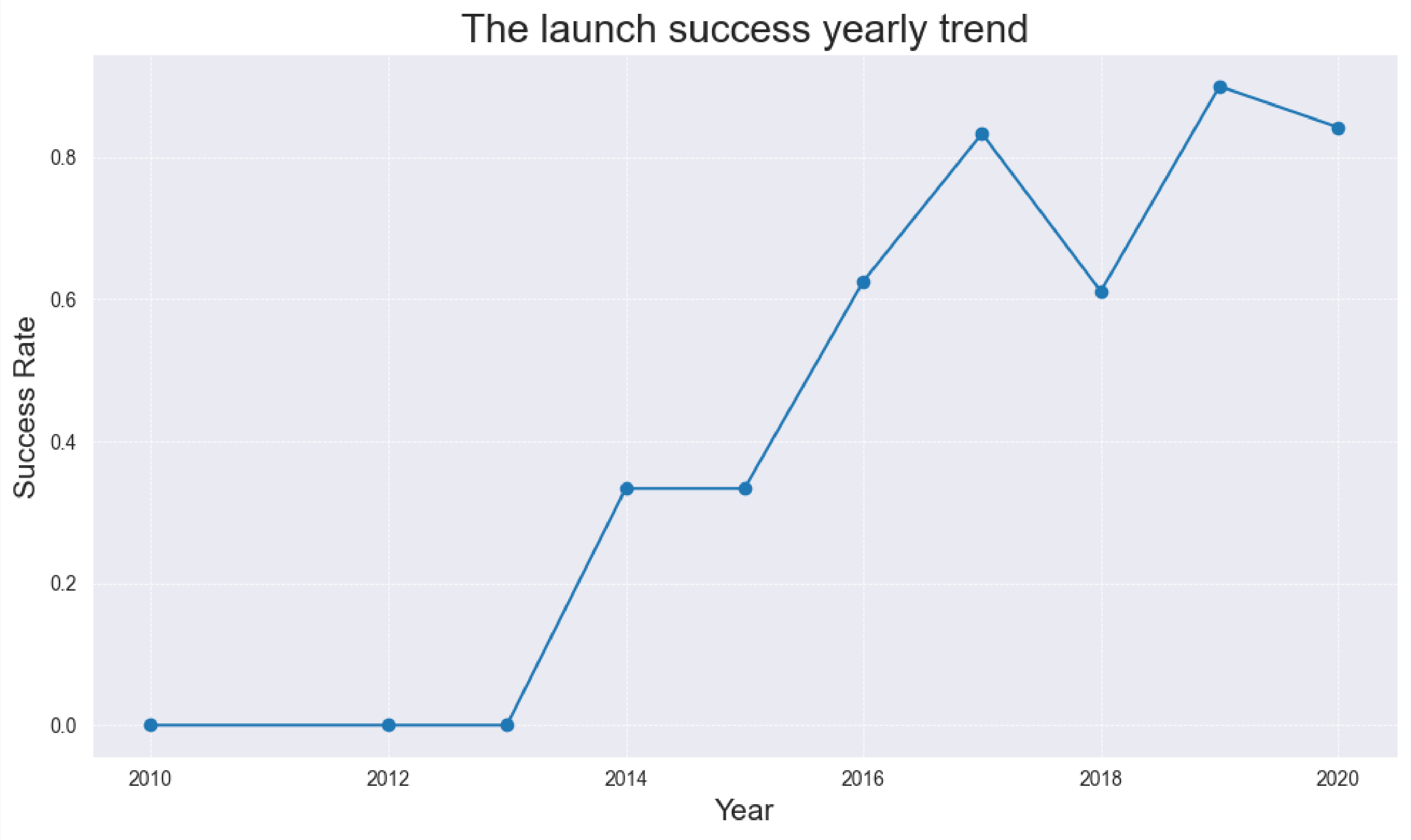
- A taxa de sucesso aumenta em praticamente todas as órbitas.
- As órbitas mais utilizadas são a ISS e a GTO.
- Parece haver uma migração de voos para a órbita VLEO mais recentemente.

Payload vs. Orbit Type



- As órbitas ISS, PO e VLEO são preferenciais para maiores cargas.
- As órbitas ES-L1, SSO, HEO, LEO, ISS e MEO são preferenciais para menores cargas.

Launch Success Yearly Trend



- Desde 2013 as taxas de sucesso dos foguetes da Space X vem crescendo.

All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

Launch Site Names Begin with 'CCA'

Query

```
%%sql
select * from "spacextbl"
where "Launch_Site" like 'CCA%'
limit 5;
```

Python

Result

```
 * postgresql+psycopg2://postgres:***@localhost:5432/SIH
5 rows affected.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Explanation

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

Total Payload Mass

Query

```
%%sql
select "Customer", sum ("PAYLOAD_MASS__KG_") from "spacextbl"
where "Customer" = 'NASA (CRS)'
group by "Customer";
```

Result

```
 * postgresql+psycopg2://postgres:***@localhost:5432/SIH
1 rows affected.
```

Customer	sum
NASA (CRS)	45596

Explanation

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

Average Payload Mass by F9 v1.1

Query

```
%%sql
select "Booster_Version", avg ("PAYLOAD_MASS__KG_") from "spacextbl"
where "Booster_Version" like 'F9 v1.1'
group by "Booster_Version";
```

```
 * postgresql+psycopg2://postgres:***@localhost:5432/SIH
1 rows affected.
```

Result

Booster_Version	avg
F9 v1.1	2928.4000000000000000

Explanation

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

First Successful Ground Landing Date

Query

```
%%sql
select min ("Date") from "spacextbl"
where "Landing_Outcome" = 'Success (ground pad)';
```

```
 * postgresql+psycopg2://postgres:***@localhost:5432/SIH
1 rows affected.
```

Result

min
2015-12-22

Explanation

Date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

Query

```
%%sql
select "Booster_Version" from "spacextbl"
where "Landing_Outcome" = 'Success (drone ship)' and "PAYLOAD_MASS__KG_" between 4000 and 6000
group by "Booster_Version"
```

Result

```
* postgresql+psycopg2://postgres:***@localhost:5432/SIH
4 rows affected.
```

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

Explanation

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

Total Number of Successful and Failure Mission Outcomes

Query

```
%%sql
select "Mission_Outcome", count ("Mission_Outcome") from "spacextbl"
group by "Mission_Outcome"
```

Result

```
 * postgresql+psycopg2://postgres:***@localhost:5432/SIH
4 rows affected.
```

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

Explanation

Total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

Query

```
%%sql
SELECT "Booster_Version", "PAYLOAD_MASS__KG_"
FROM "spacextbl"
WHERE "PAYLOAD_MASS__KG_" = (
    SELECT MAX("PAYLOAD_MASS__KG_") FROM "spacextbl"
);
```

Python

```
 * postgresql+psycopg2://postgres:***@localhost:5432/SIH
12 rows affected.
```

Result

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

Explanation

Names of the booster versions which have carried the maximum payload mass

2015 Launch Records

Query

```
%%sql
SELECT
    EXTRACT(MONTH FROM "Date"::date) as month,
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM "spacextbl"
WHERE EXTRACT(YEAR FROM "Date"::date) = 2015
    AND "Landing_Outcome" LIKE 'Failure (drone ship)'
ORDER BY month;
```

Result

```
 * postgresql+psycopg2://postgres:***@localhost:5432/SIH
2 rows affected.
```

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

Explanation

Records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.

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

Query

```
%%sql
SELECT
    "Landing_Outcome",
    COUNT(*) as outcome_count
FROM
    "spacextbl"
WHERE
    "Date"::date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY
    "Landing_Outcome"
ORDER BY
    outcome_count DESC;
```

Python

Result

```
 * postgresql+psycopg2://postgres:***@localhost:5432/SIH
8 rows affected.
```

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

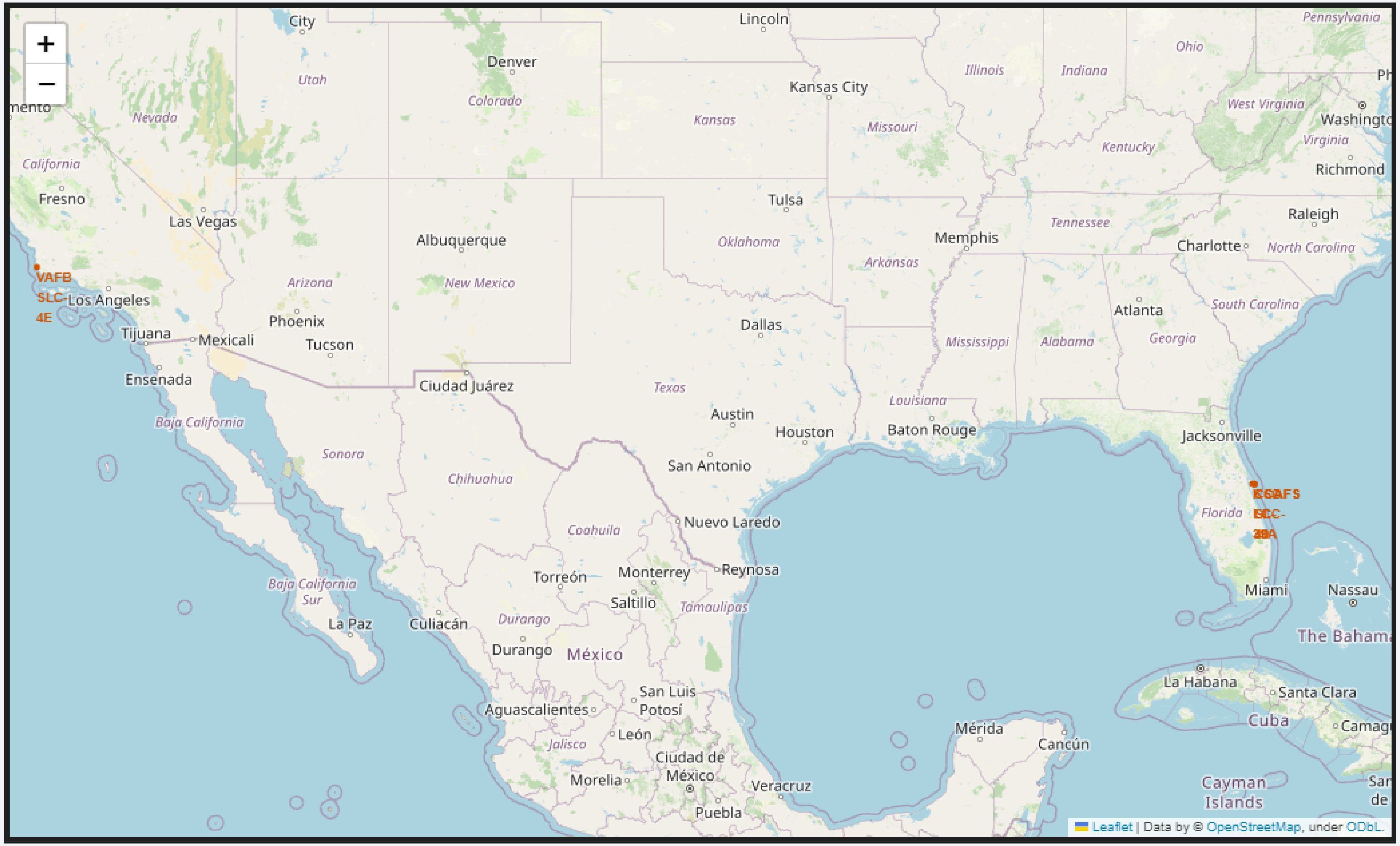
Explanation

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 markers

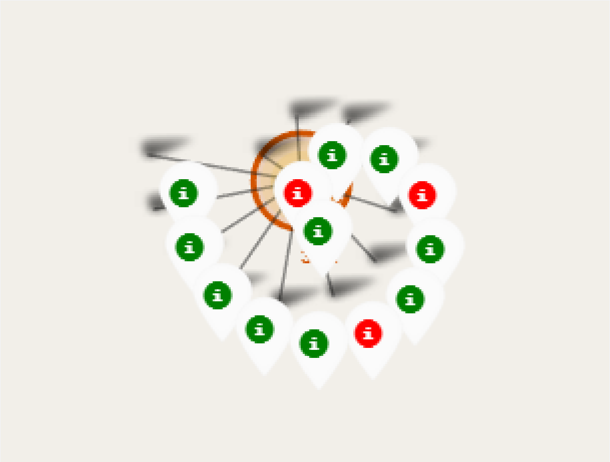


The map displays the location of the four launch sites, which are on the east and west coast of the US.

Markers indicate successful or failed launches for each site on the map.

VAFB SLC-4E

KSC LC-39A



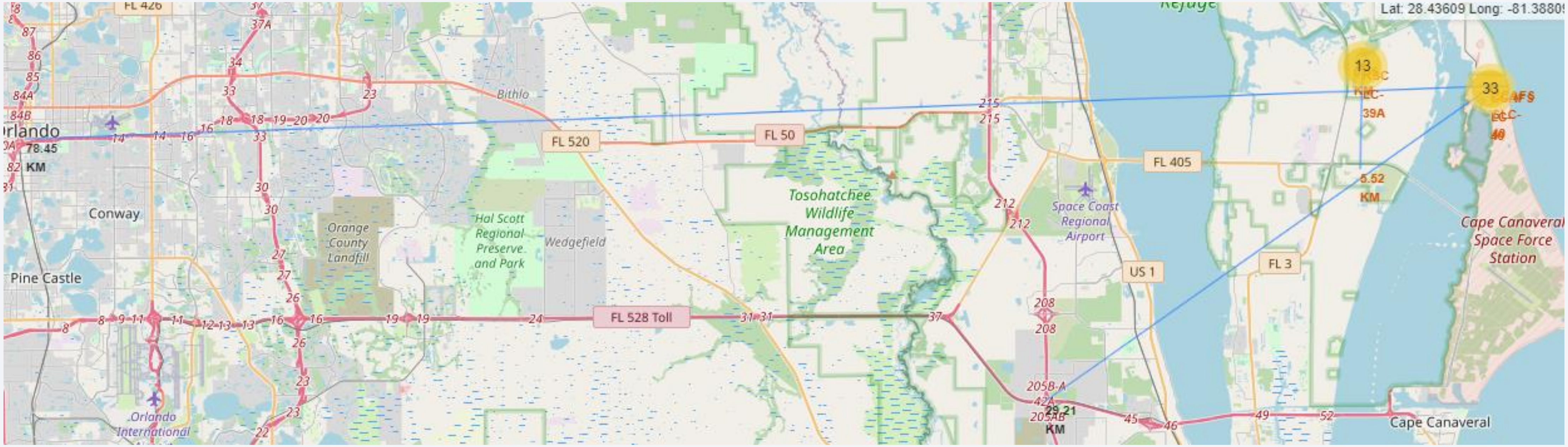
CCAFS LC-40

CCAFS SLC-40

Success

Failure

Distances between a launch site to its proximities

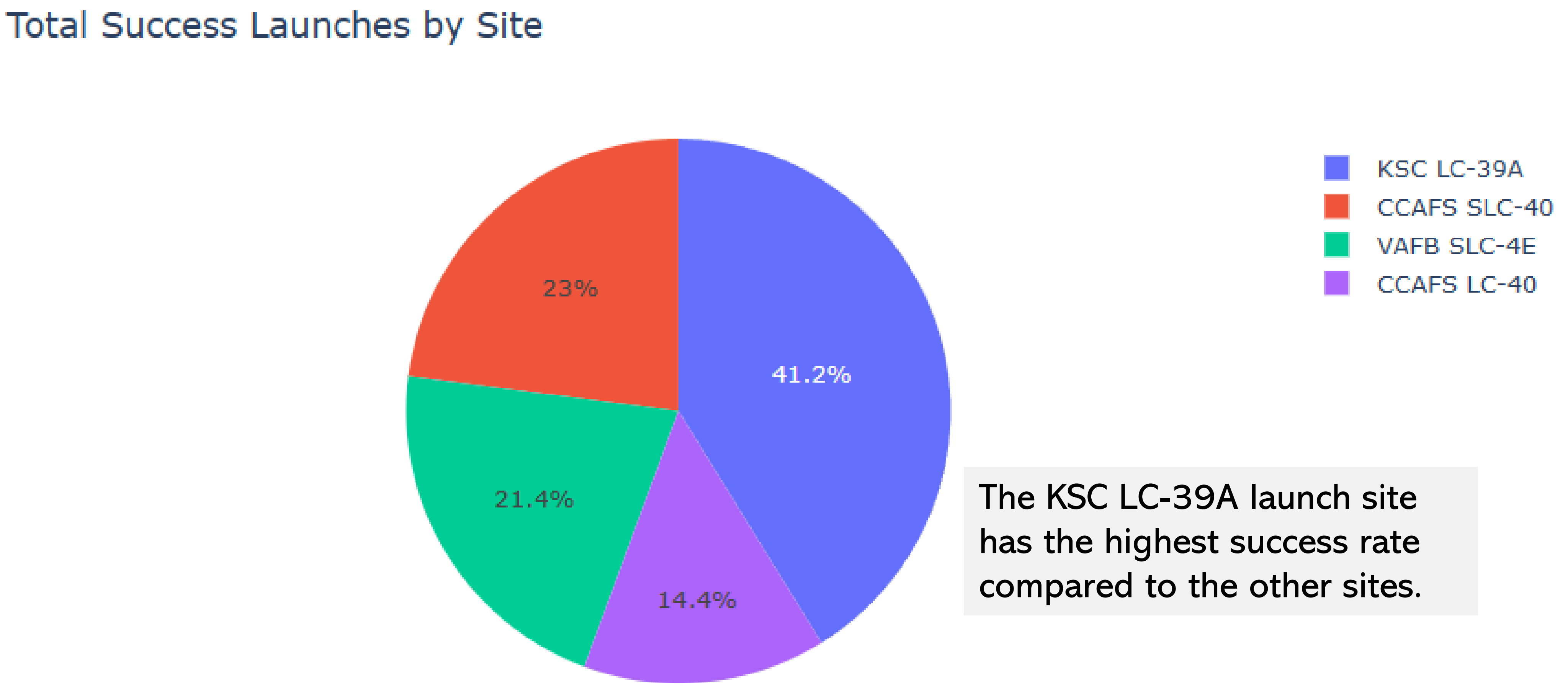


- Launch sites are not in close proximity to railways.
- Launch sites are not adjacent to highways.
- Launch sites are located near coastlines.
- Launch sites maintain a certain distance from cities.

Section 4

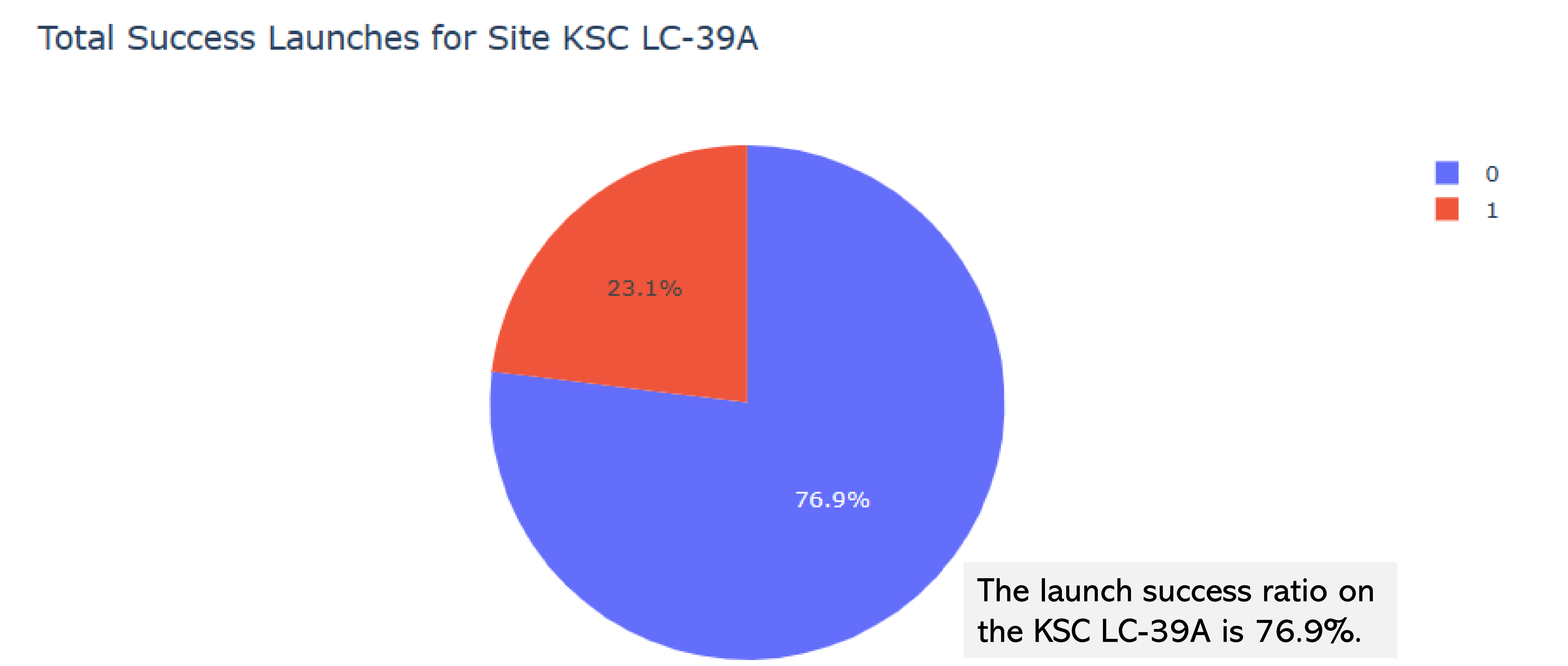
Build a Dashboard with Plotly Dash

Success percentage achieved by each launch site



The KSC LC-39A launch site has the highest success rate compared to the other sites.

Launch site with the highest launch success ratio

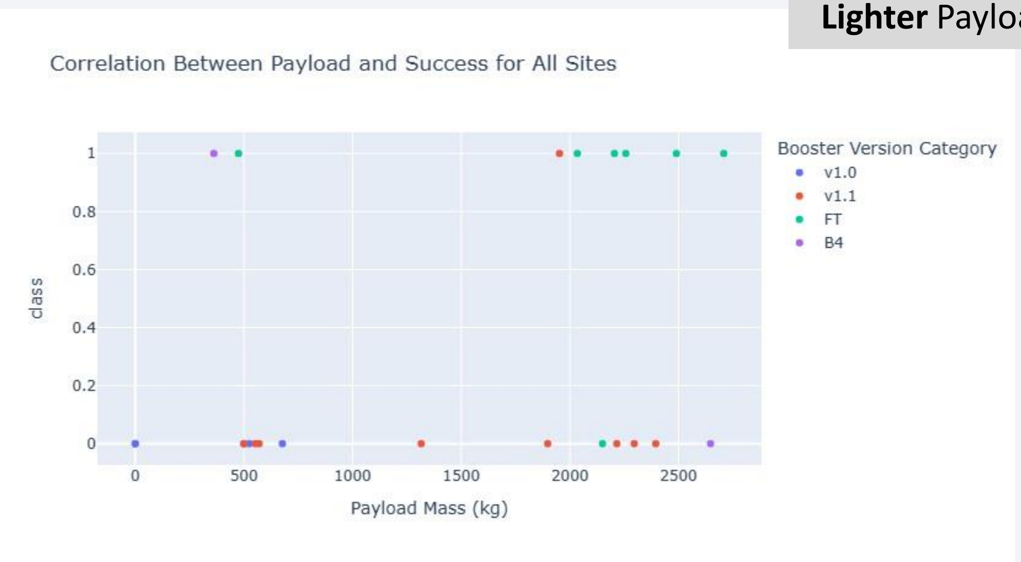


The launch success ratio on the KSC LC-39A is 76.9%.

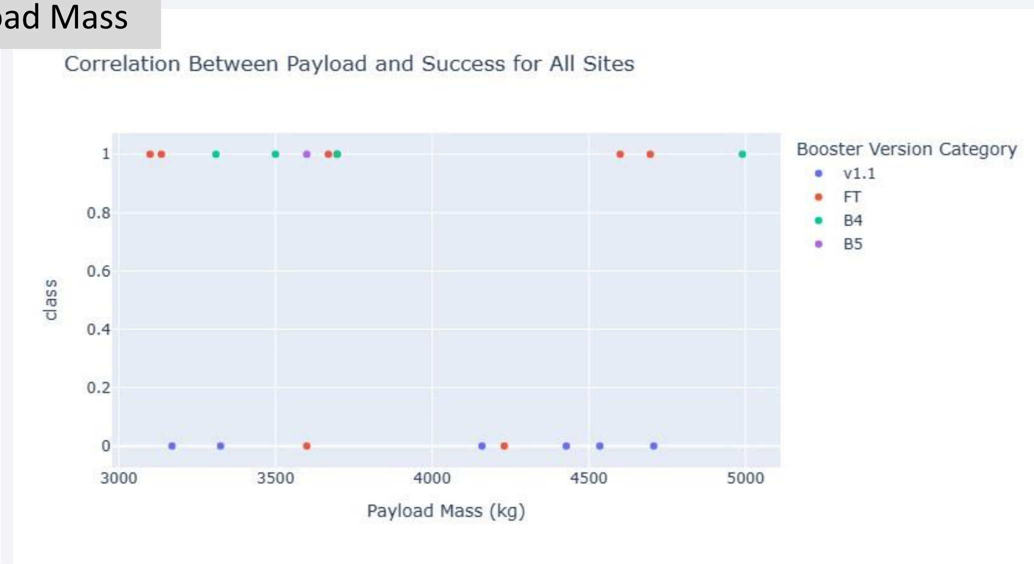
Payload vs Launch Outcome for all sites with different payload Mass

Lighter Payload Mass

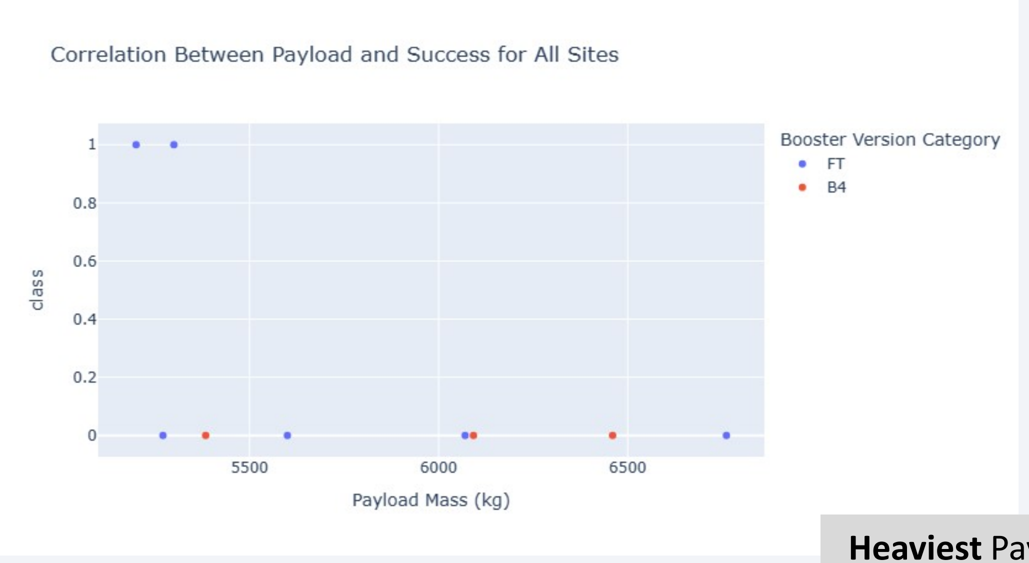
Payload Mass
0 – 3,000



Payload Mass
3,000 – 5,000



Payload Mass
5,000 – 8,000



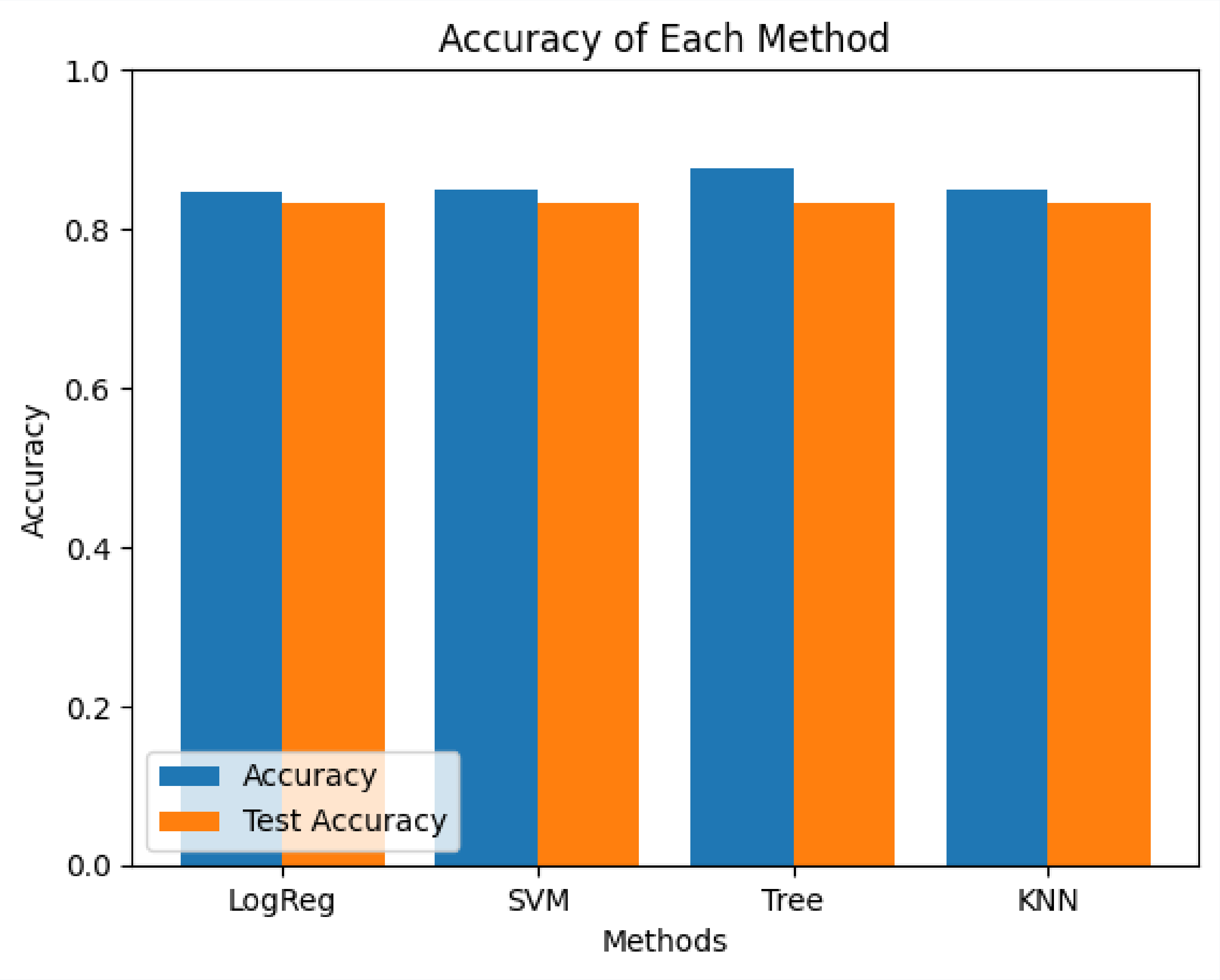
Payload Mass
8,000 – 10,000

Heaviest Payload Mass

Section 5

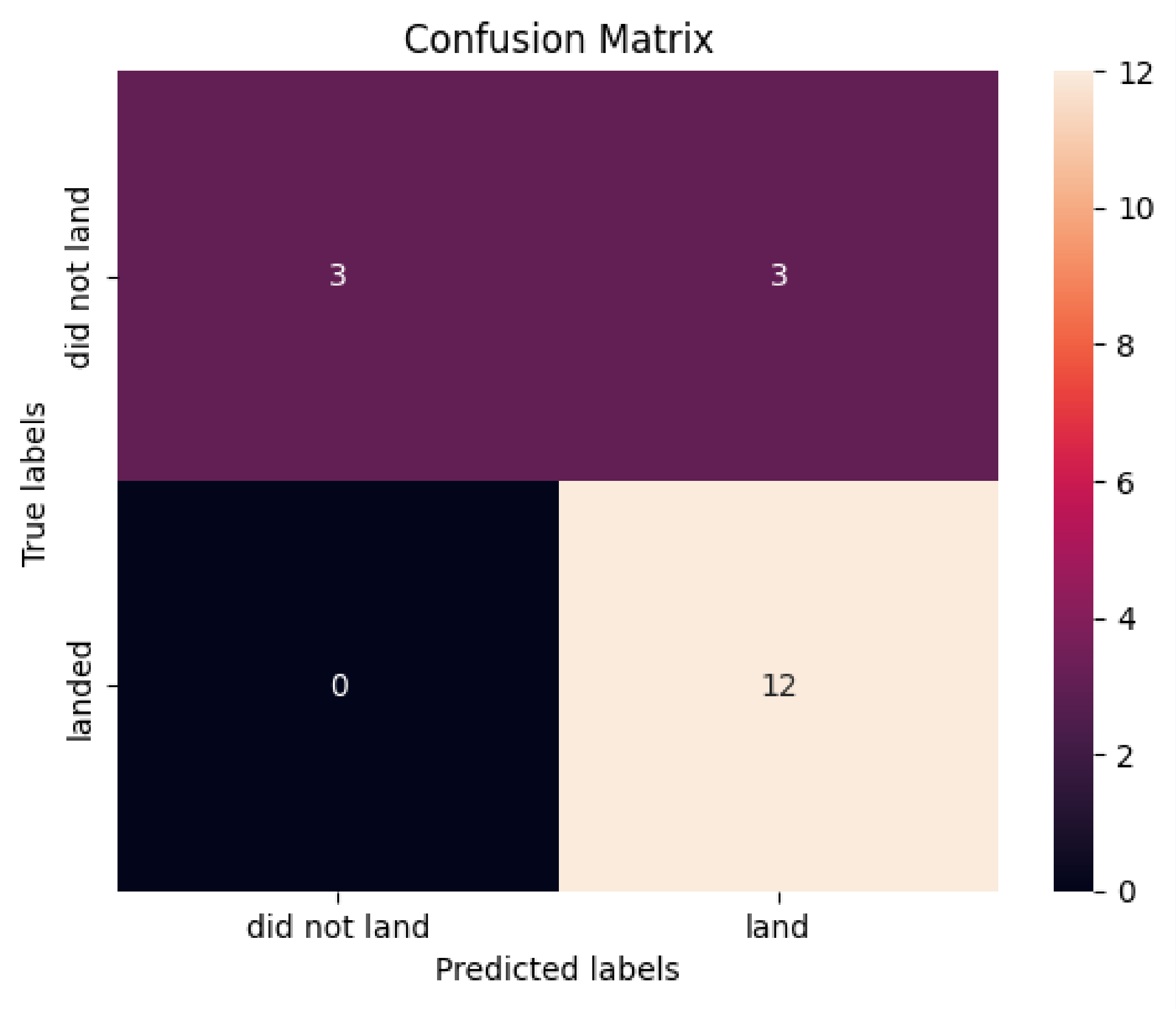
Predictive Analysis (Classification)

Classification Accuracy



The model with the highest classification accuracy is Decision Trees.

Confusion Matrix



Decision Tree Confusion Matrix

Conclusions

- There is a direct correlation between the number of flights from a launch site and its success rate.
- We observed a steady increase in launch success rates from 2013 through 2020.
- The orbits ES-L1, GEO, HEO, SSO, and VLEO have the highest success rates.
- Among all the launch sites, KSC LC-39A recorded the most successful launches.
- The Decision Tree classifier proved to be the most effective machine learning algorithm for this analysis."

Thank you!