IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rodrigo Elesbão
July 23rd, 2023

Outline

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

Executive Summary

Summary of methodologies

- Understanding rocket launching as product, and how/why reusing rocket stages impacts its cost.
- Applying Python programming to collect data via online requests and web scraping.
- Performing exploratory analysis over the collected data converted into data frames to find patterns.
- Using plotting and visual analytics as insight tools.
- Developing a Machine Learning model that could be used to state trends and projections.

Summary of all Results

- The collected data can effectively compound a scenery on the SpaceX Falcon 9 rocket landing outcomes.

Introduction

Project Background and Context:

- SpaceX has gained worldwide attention for a series of historic milestones
- It is the only private company to return a spacecraft from low-earth orbit
- They advertise Falcon 9 rocket launches on its website with a cost of 62 million dollars; much of the savings is because SpaceX can reuse the first stage

Existing issues and Opportunities:

- Despite the price, the launch success rate may depend on factors such as payload mass, orbit type, location, etc
- Therefore, the launch cost can be determined if Falcon 9 rocket's first-stage landing success rate can be determined
- Such information can be used if an alternate company wants to bid against SpaceX for a rocket launch

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- SpaceX API request, Web Scraping from Wikipedia, SpaceX Dataset SQL request

Perform data wrangling

- Handling gathered data into Pandas data frame with Numpy, Matplotlib.pyplot, and Seaborn

Perform exploratory data analysis (EDA) using visualization and SQL

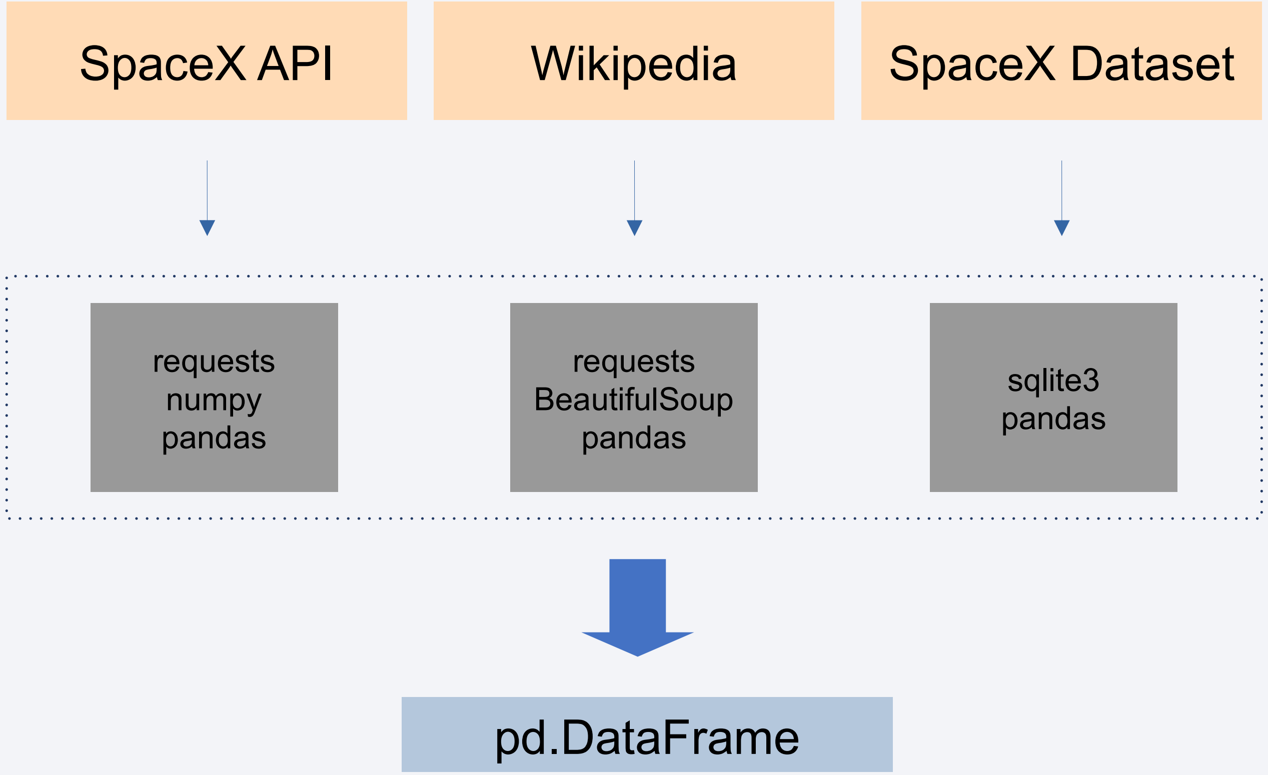
Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- Applying and evaluating machine learning pipelines with Python libraries over data frames

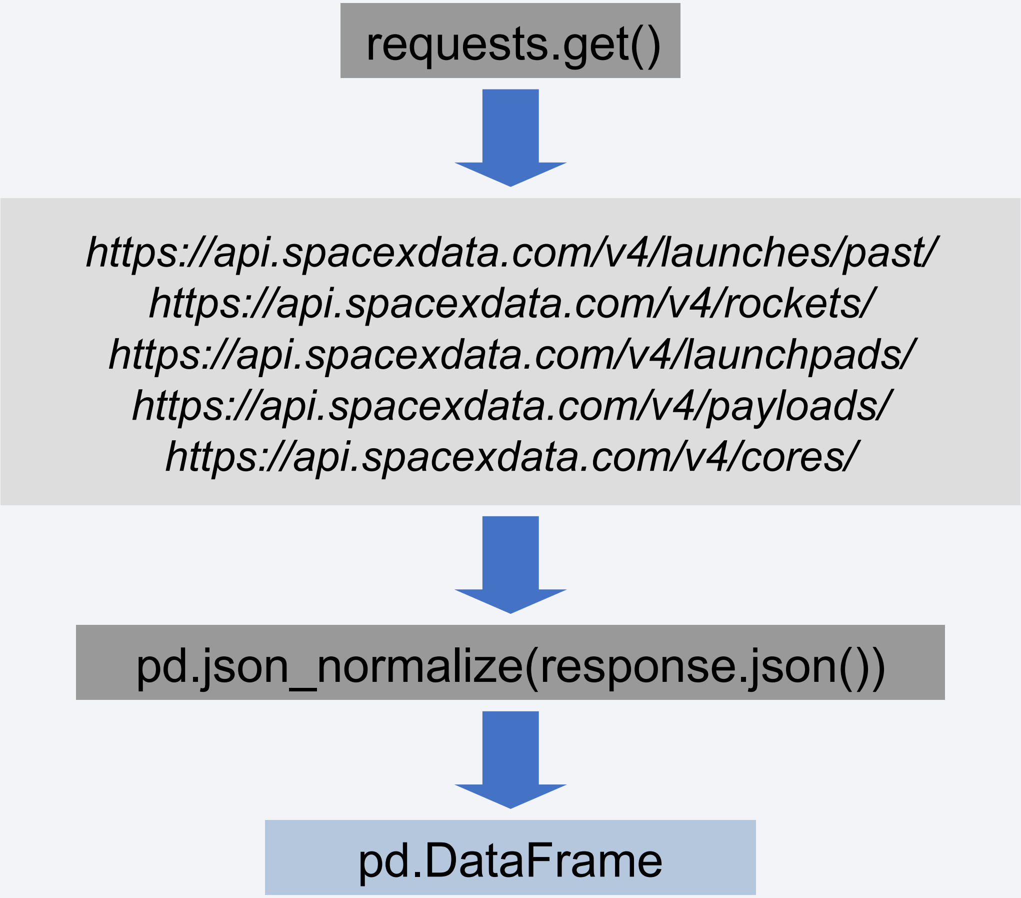
Data Collection

- Requests to SpaceX API
- Web scraping to collect records from a Wikipedia page
- Complementary SQL requests to a SpaceX Dataset
- All data are merged into Pandas data frames



Data Collection – SpaceX API

- Requests to Spacex API compound the main launch data source
- The results of the requests are then normalized and gathered into a Pandas data frame



GitHub: [1_1_jupyter-labs-spacex-data-collection-api.ipynb]

Data Collection – Scraping

- Related data collected from Wikipedia compound complementary data about launch records
- The results of web scrapping are then parsed and gathered into a Pandas Data Frame

requests.get().txt

↓

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

↓

BeautifulSoup(, 'html.parser')

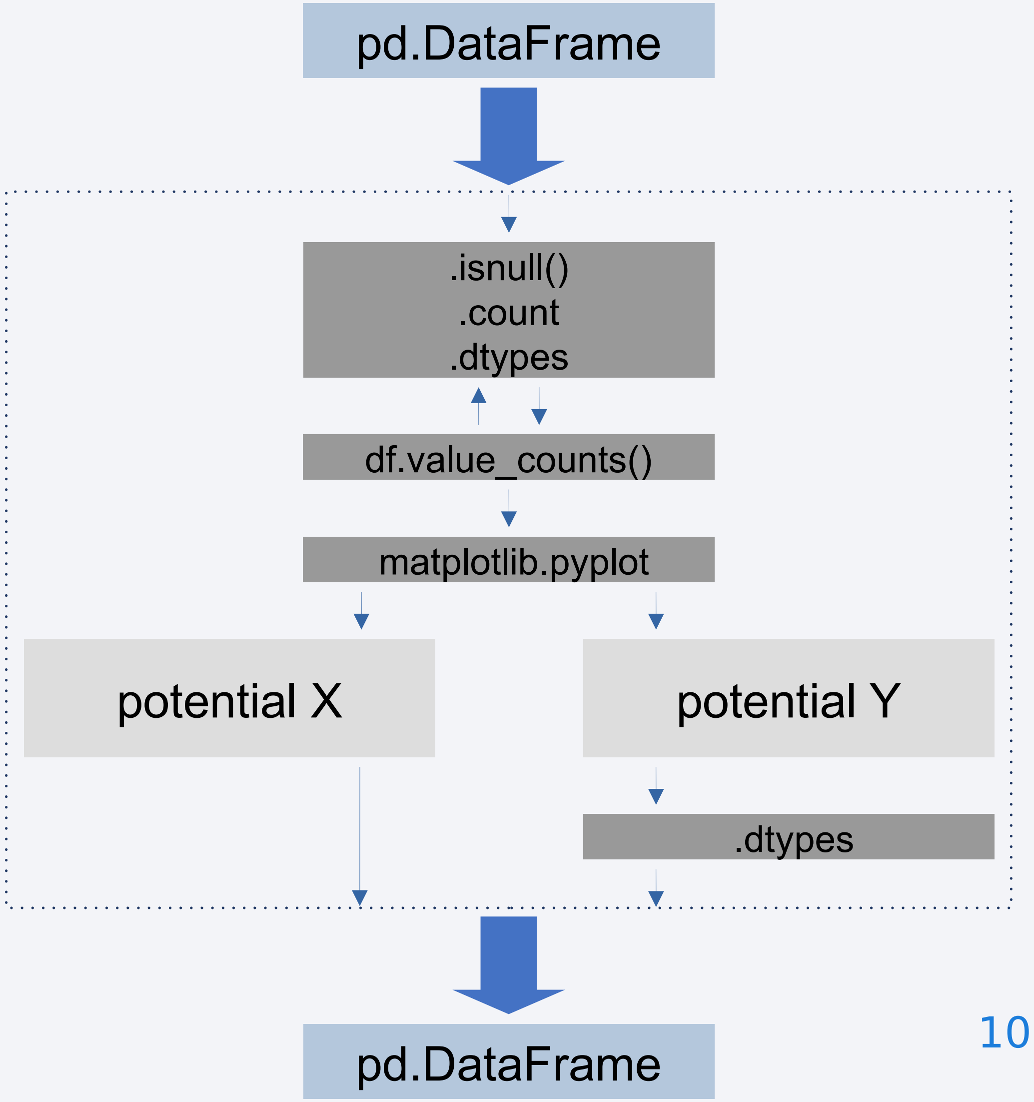
↓

pd.DataFrame

GitHub: 1_2_jupyter-labs-webscraping.ipynb

Data Wrangling

Using data previously collected via SpaceX API into a data frame using Pandas and Numpy



GitHub: 1_3_labs-jupyter-spacex-Data_wrangling.ipynb

EDA with Data Visualization

Plots used for exploratory data visualization:

- flight number vs launch site
- number of launches on each site
- launch site vs payload mass
- orbit vs class
- number and occurrence of orbits
- flight number vs orbit type
- payload vs orbit
- year vs class

GitHub: [2_2_jupyter-labs-eda-dataviz.ipynb]

EDA with SQL

SQL queries to SpaceX Dataset:

- fetch all the unique launch site names
- fetch first 5 occurrences of launch site names begin with ‘CCA
- fetch total payload carried by boosters from NASA
- return average payload mass carried by booster version F9 v1.1
- fetch date of the first successful landing outcome on ground pad
- list names of boosters which have successfully landed on drone ship and had payload mass between 4000 Kg and 6000 Kg
- calculate the total number of successful and failure mission outcomes
- list names of the boosters which have carried the maximum payload mass
- list failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- rank count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

GitHub: [2_1_jupyter-labs-eda-sql-coursera_sqllite.ipynb]

Build an Interactive Map with Folium

Folium Map: launch sites location (over the USA map)

- **Map:** .Map(location=[29.559684888503615, -95.0830971930759], zoom_start=5)

.add_child(.Circle())

.add_child(.Marker())

Folium Map: launch outcomes for each site location

.add_child(.Marker(.Icon))

Folium Map: Nearby Infrastructure

- .add_child(Marker(icon=(DivIcon(html='distance'))))
- .add_child(.Polyline())

GitHub: 3_1_lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

App layout:

```
app.layout = html.Div(children=[
```

Input: select launch site (one or all)

```
    dcc.Dropdown(id='site-dropdown')
```

Output: render a pie chart graph to show success launches

```
    html.Div(dcc.Graph(id='success-pie-chart'))
```

Input: select a payload mass range

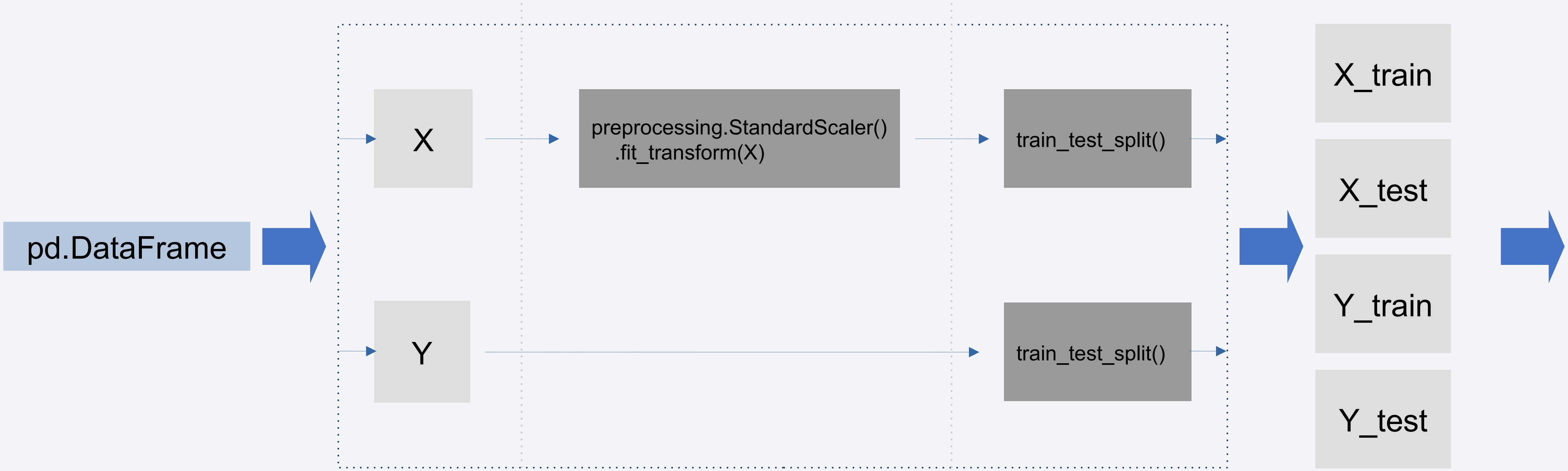
```
    dcc.RangeSlider(id='payload-slider')
```

Output: render a scatter plot to display values for Payload Mass (kg) and class (success rate)

```
    html.Div(dcc.Graph(id='success-payload-scatter-chart')) ])
```


GitHub: [3_2_2_spacex_dash_app.py]

Predictive Analysis (Classification)



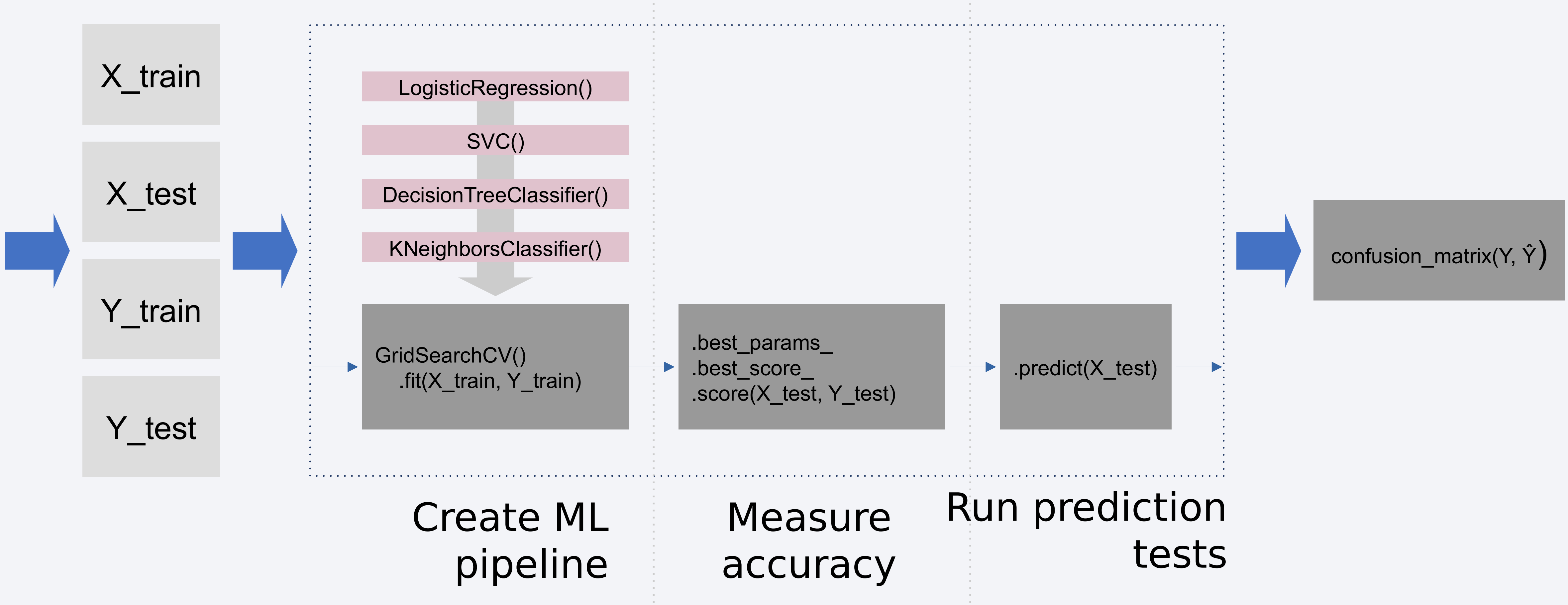
Assign data to X, Y variables

Standardize the data in X

Split X, Y into training/test data

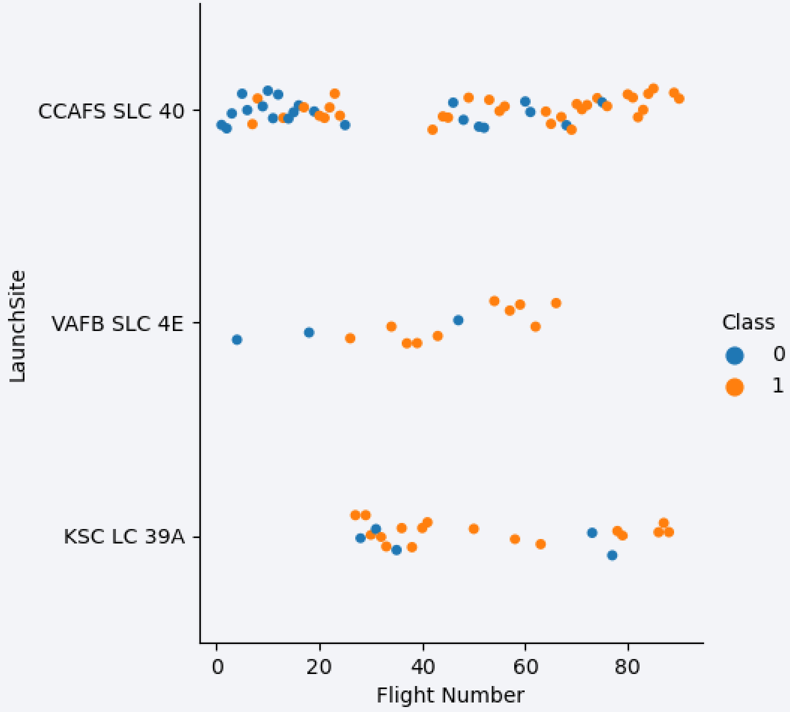
GitHub: [4_1_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb]

Predictive Analysis (Classification)

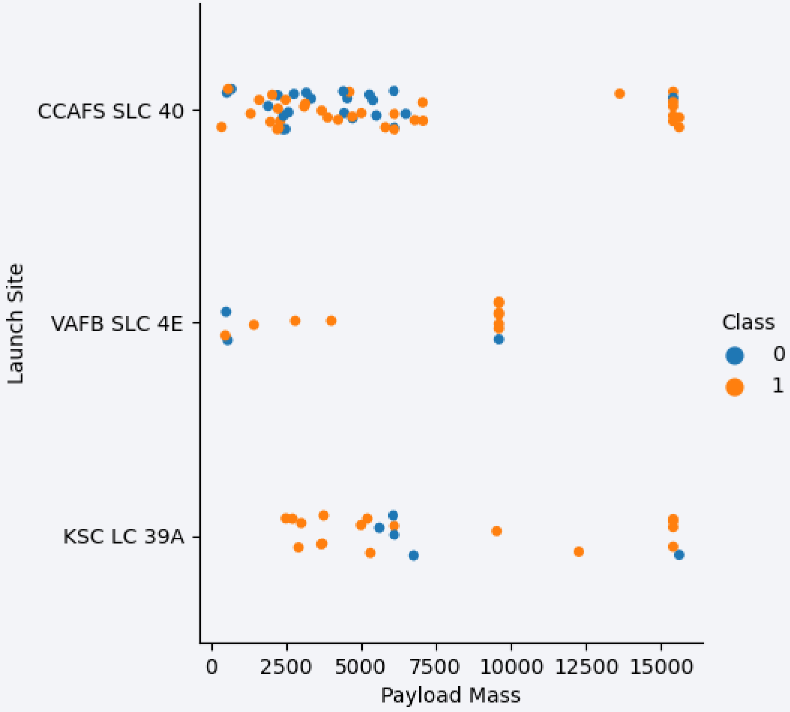


GitHub: [4_1_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb]

Results: EDA

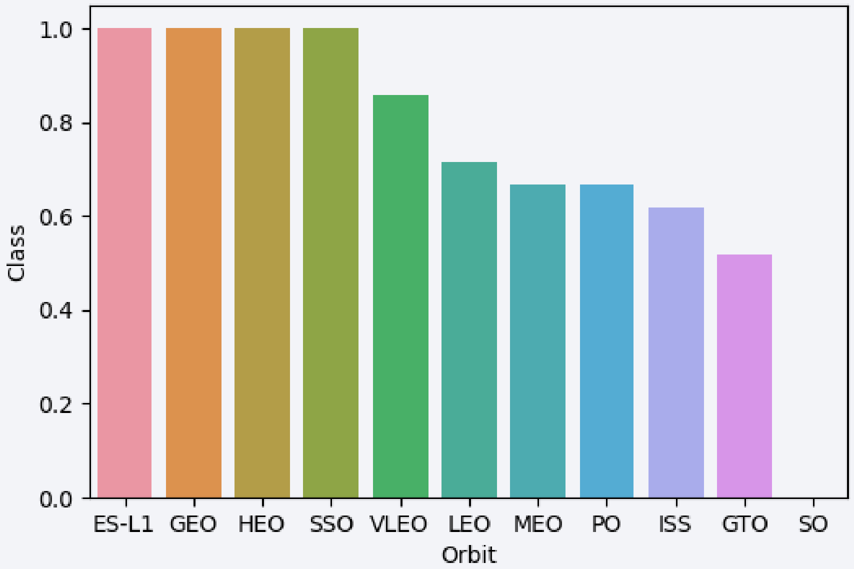


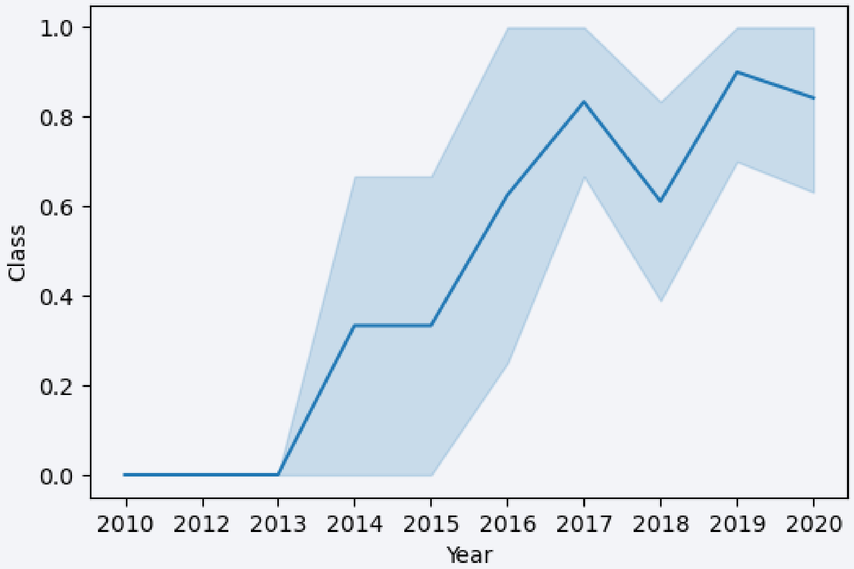
Plots on the evolution of launch and outcome rates per launch site



Plots on the distribution of payload mass among launch sites

Results: EDA



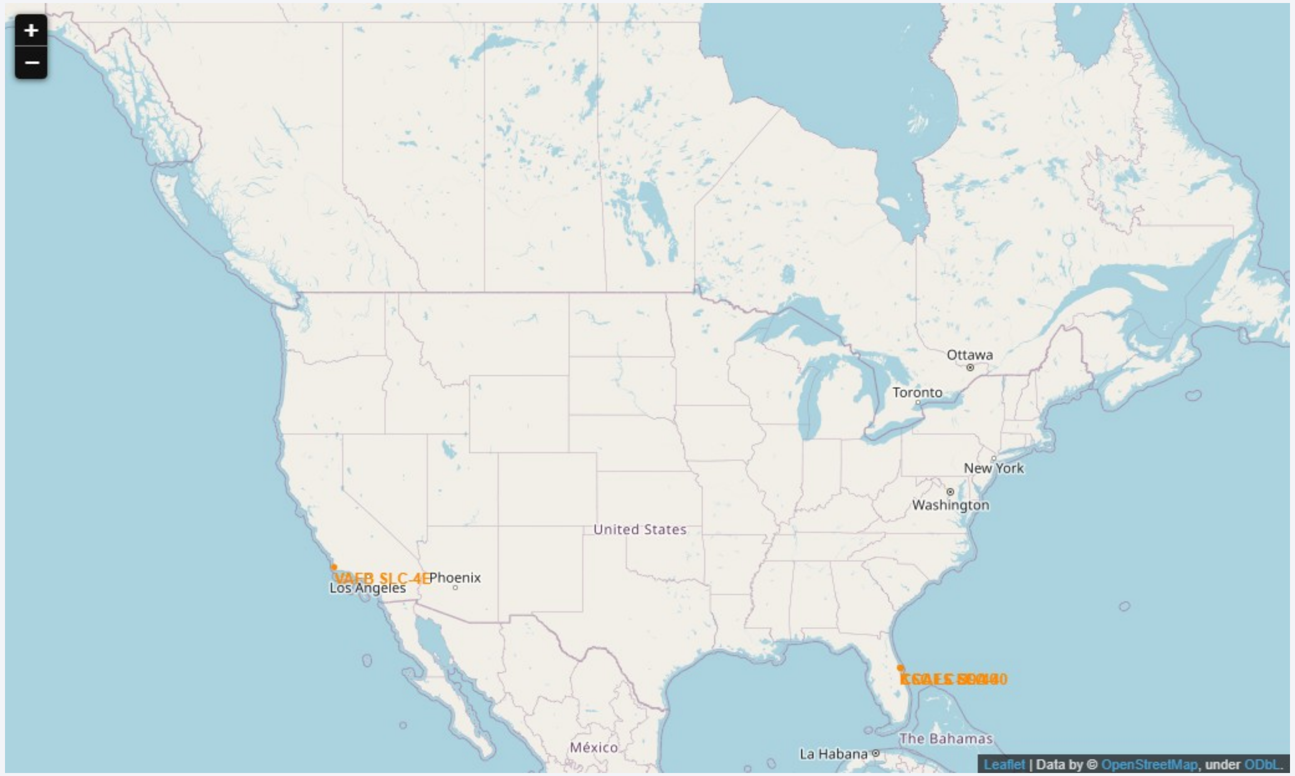


Plots on the distribution, frequency, payload, and success rates by type of orbit

Plot on the overall outcome through years

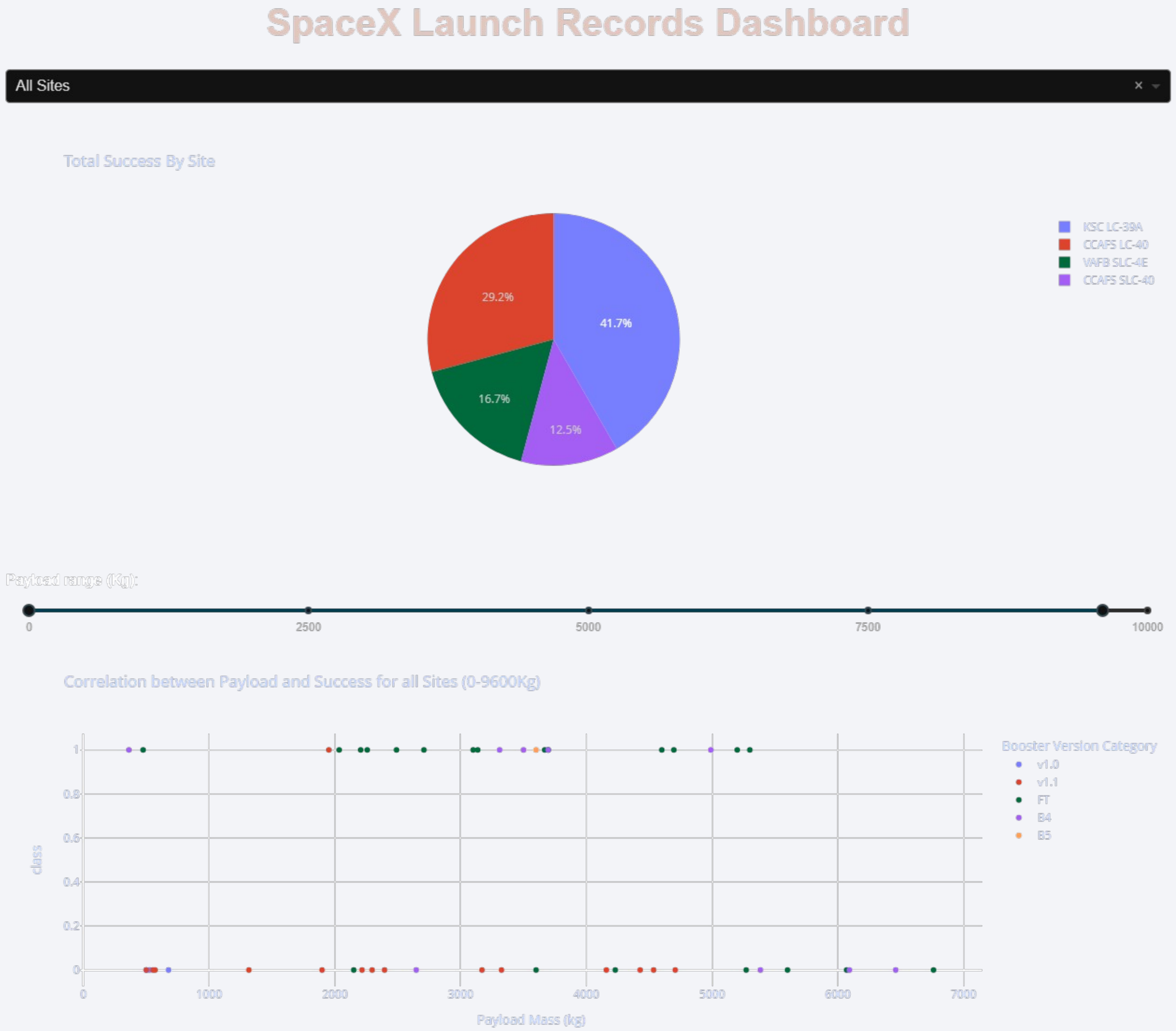
Results: Interactive Analytics

Launch site locations on an interactive map with Folium



Results: Interactive Analytics

Dashboard with Plotly Dash relating launch sites, payload mass, booster category, and outcome

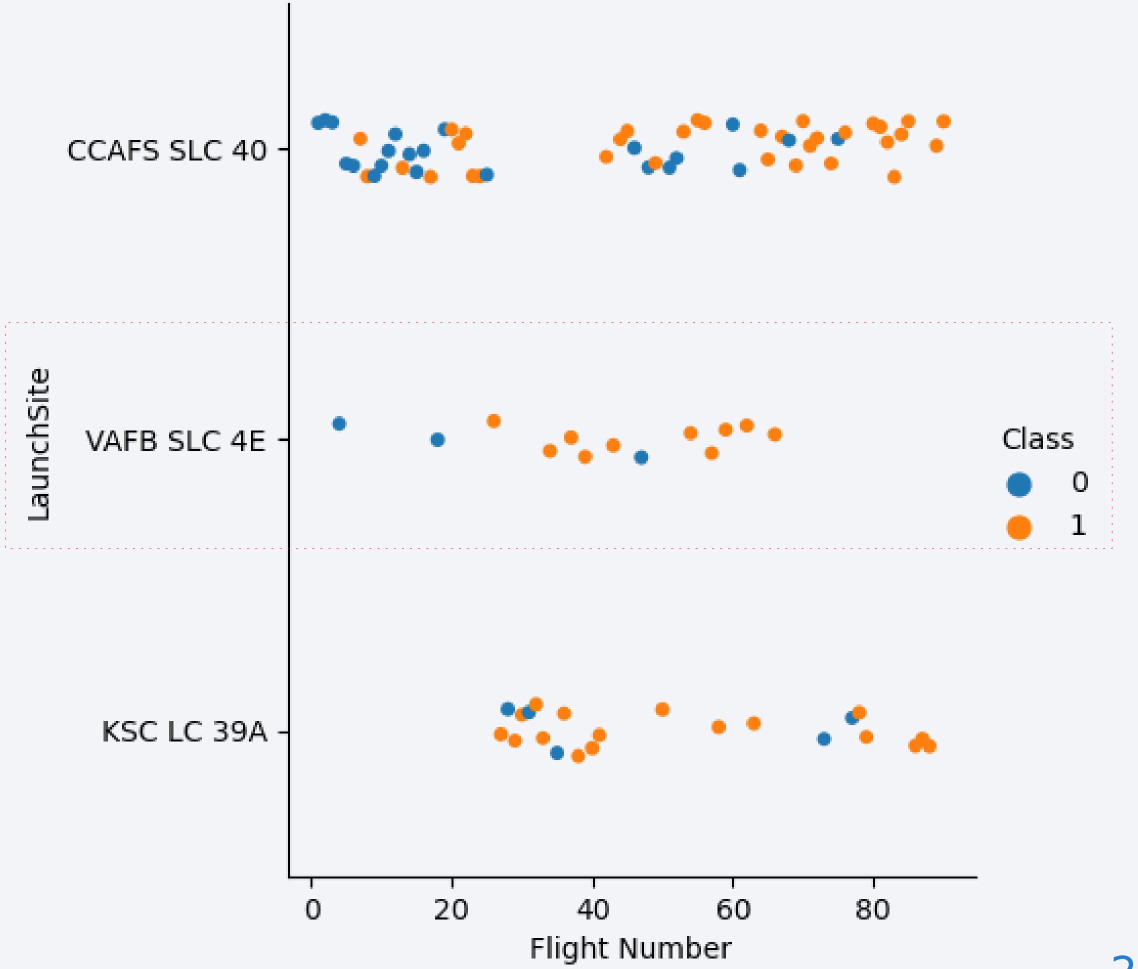


Section 2

Insights drawn from EDA

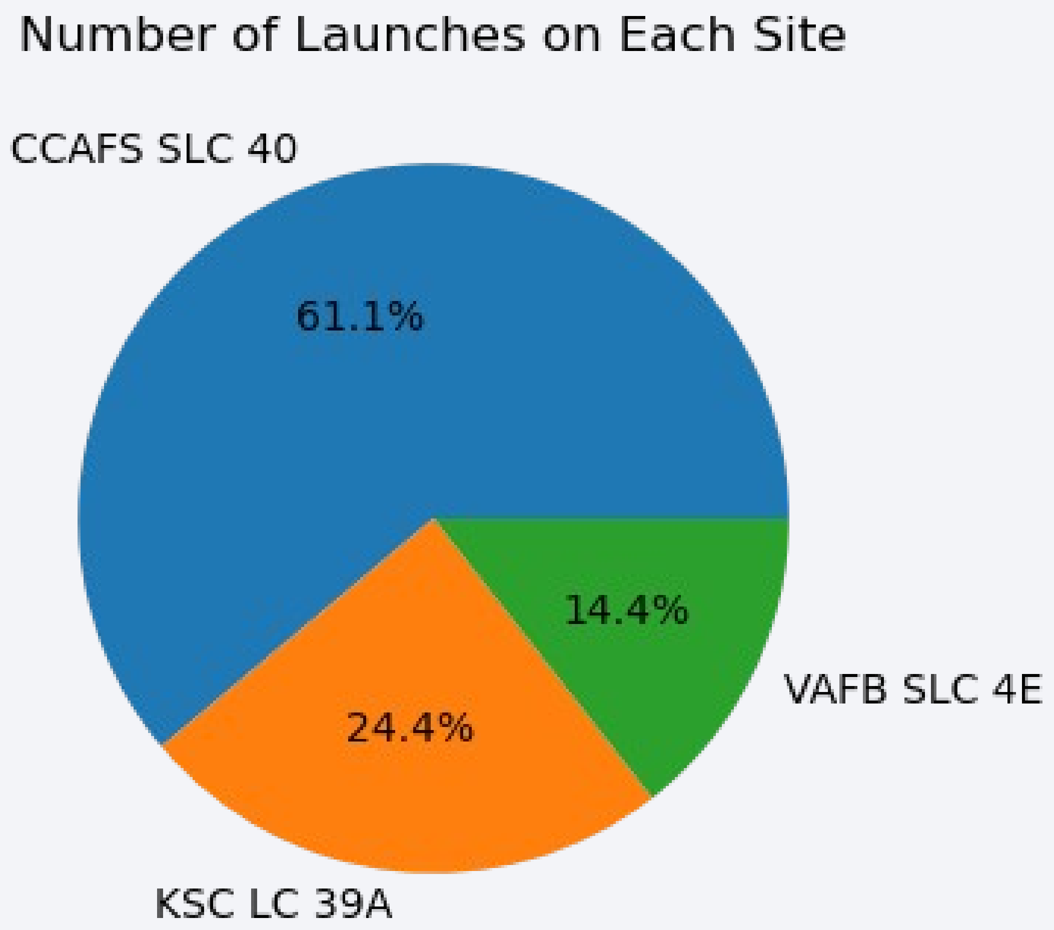
Flight Number vs. Launch Site

- CCAFS SLC 40 has most of the flights
- It's also where the evolution of outcome rates is more outstanding - from the high number of failed landings at the beginning to the rise of the successful landings as of the end of the period



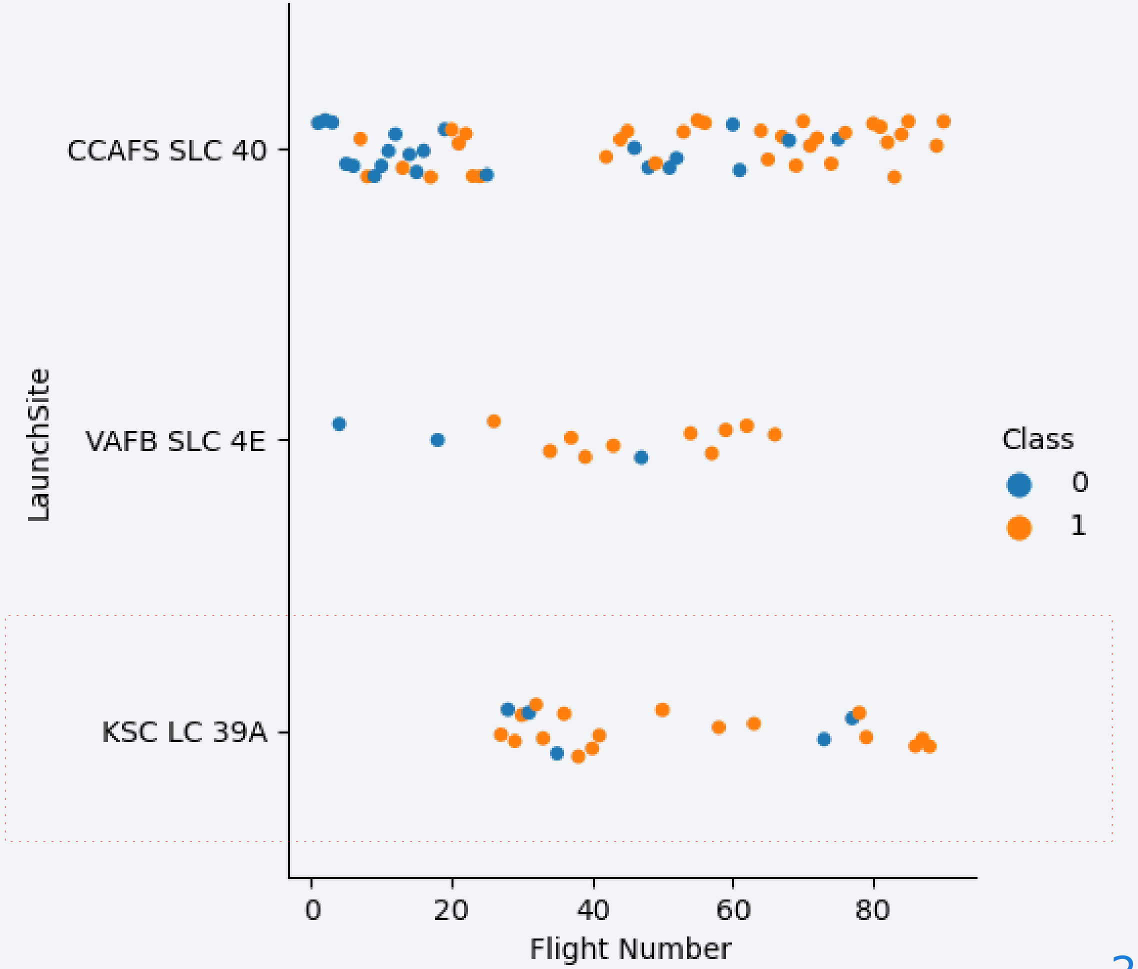
Flight Number vs. Launch Site

- CCAFS SLC 40 has most of the flights
- It's also where the evolution of outcome rates is more outstanding - from the high number of failed landings at the beginning to the rise of the successful landings as of the end of the period



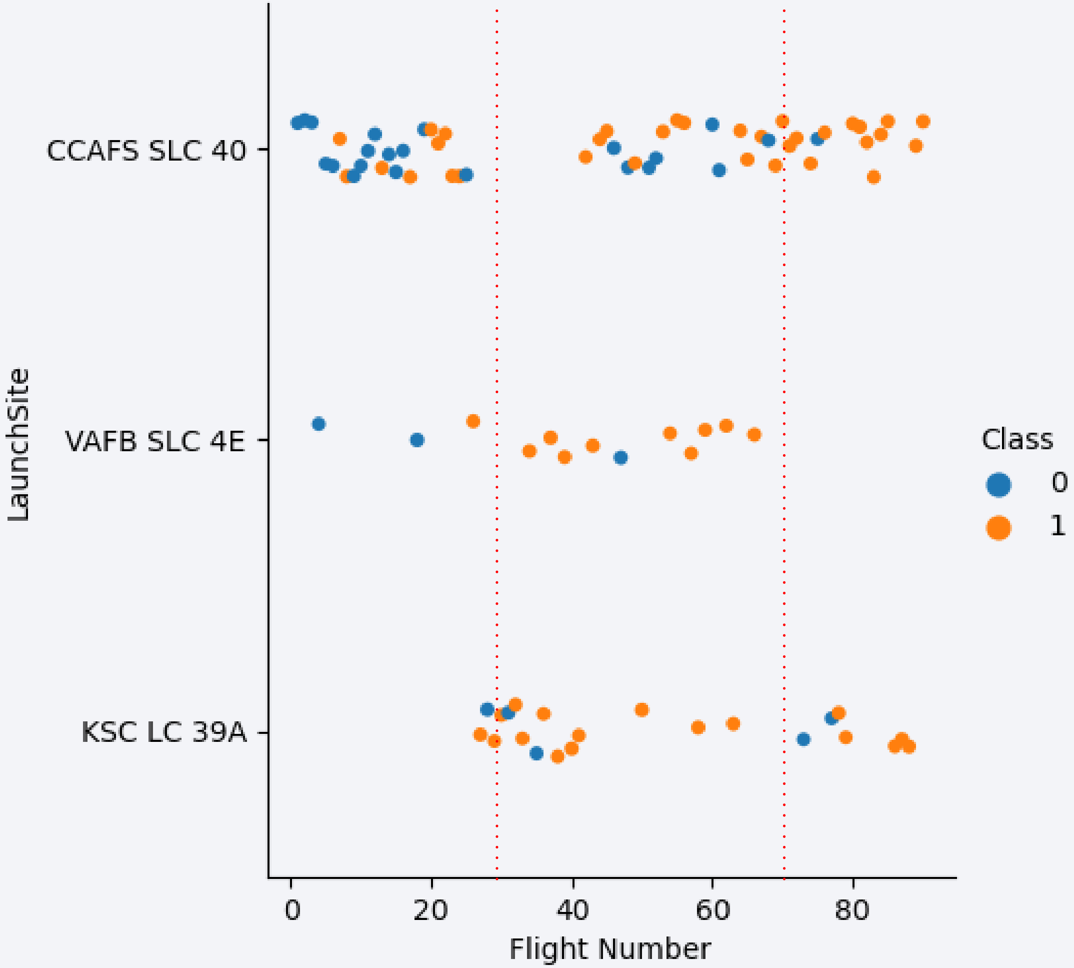
Flight Number vs. Launch Site

- Launch tests started later in KSC LC 39A than the other two
- It has a remarkable success rate



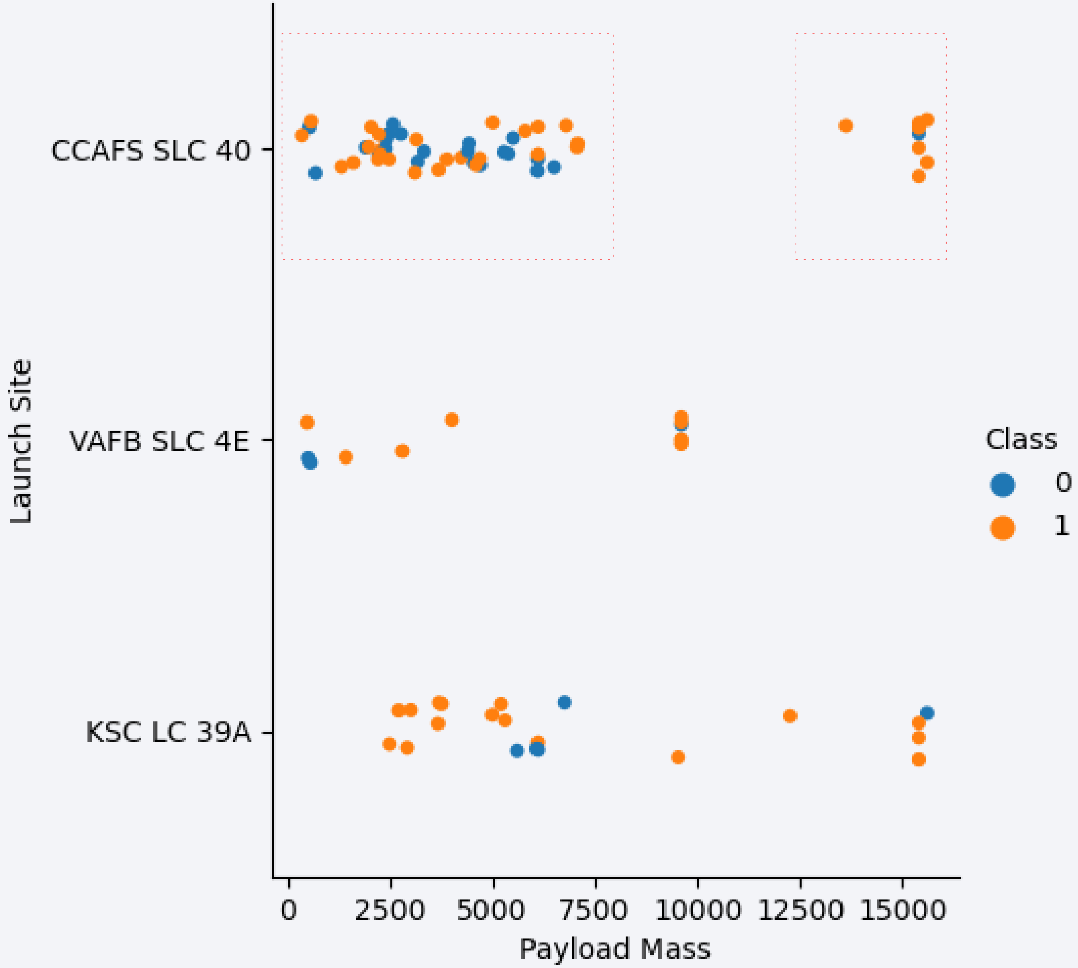
Flight Number vs. Launch Site

In every launch site, successful landing rates are prone to increase as the flight number increases



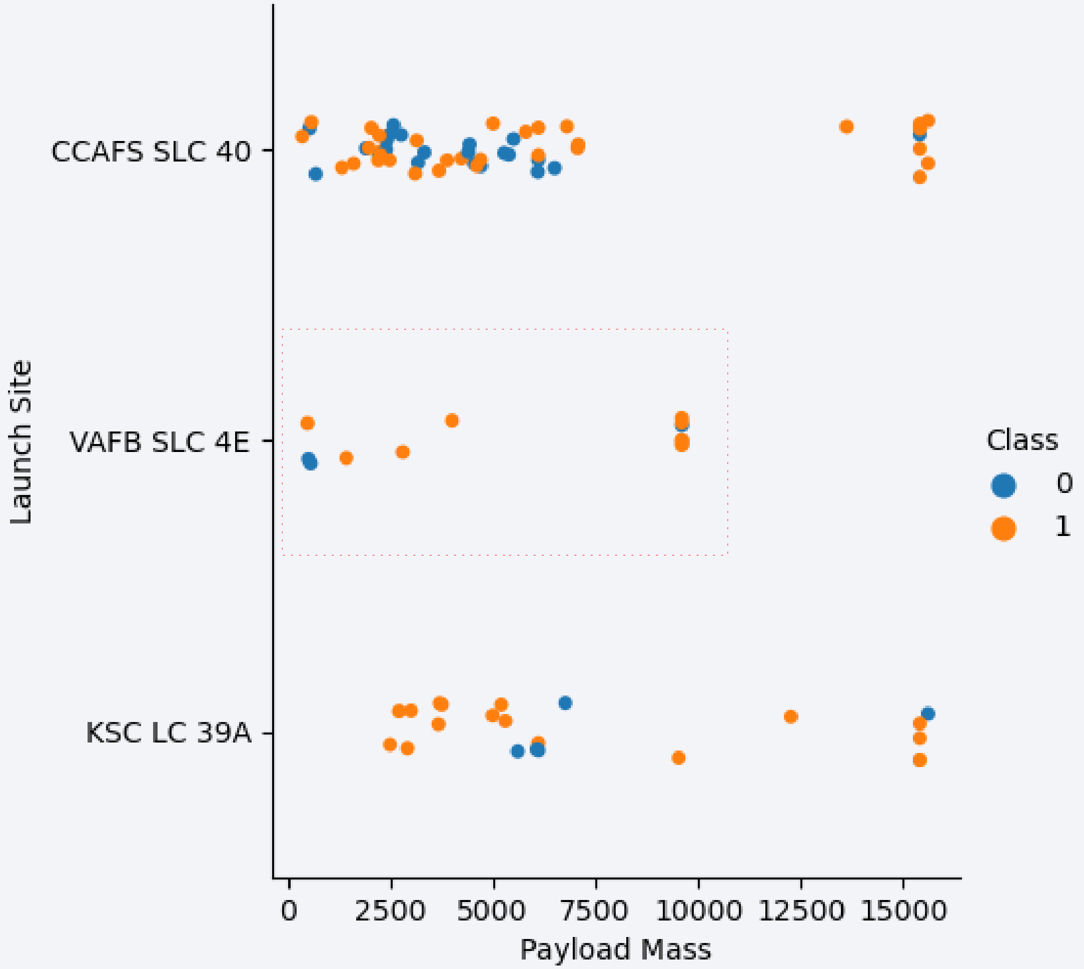
Payload vs. Launch Site

- CCAFS SLC 40 is where most tests with lighter and heavier payloads occurred
- There were more successful high payload launches on this site than any other



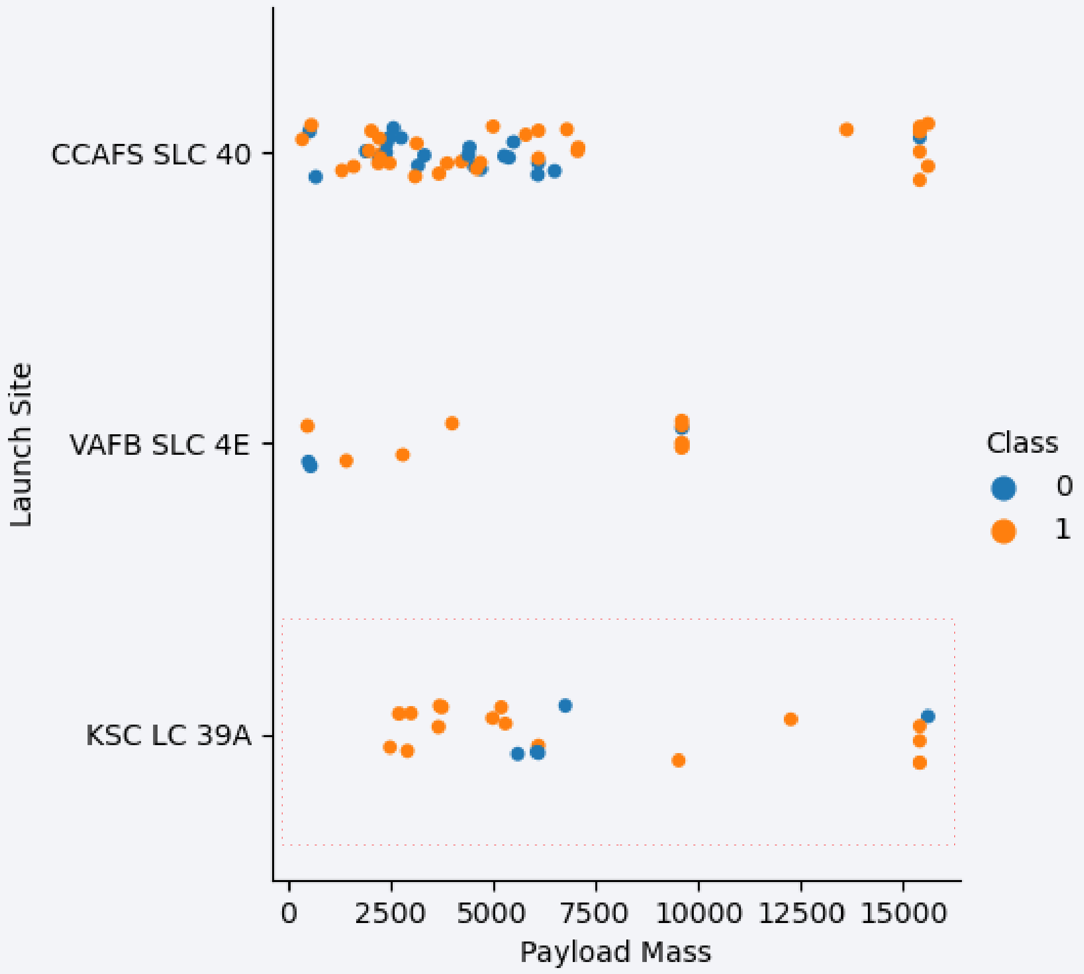
Payload vs. Launch Site

- VAFB SLC 4E has no launches with heavy payloads
- It is also the least used of the launch sites
- Several midweight payload launches took place on this site, if compared to the other sites



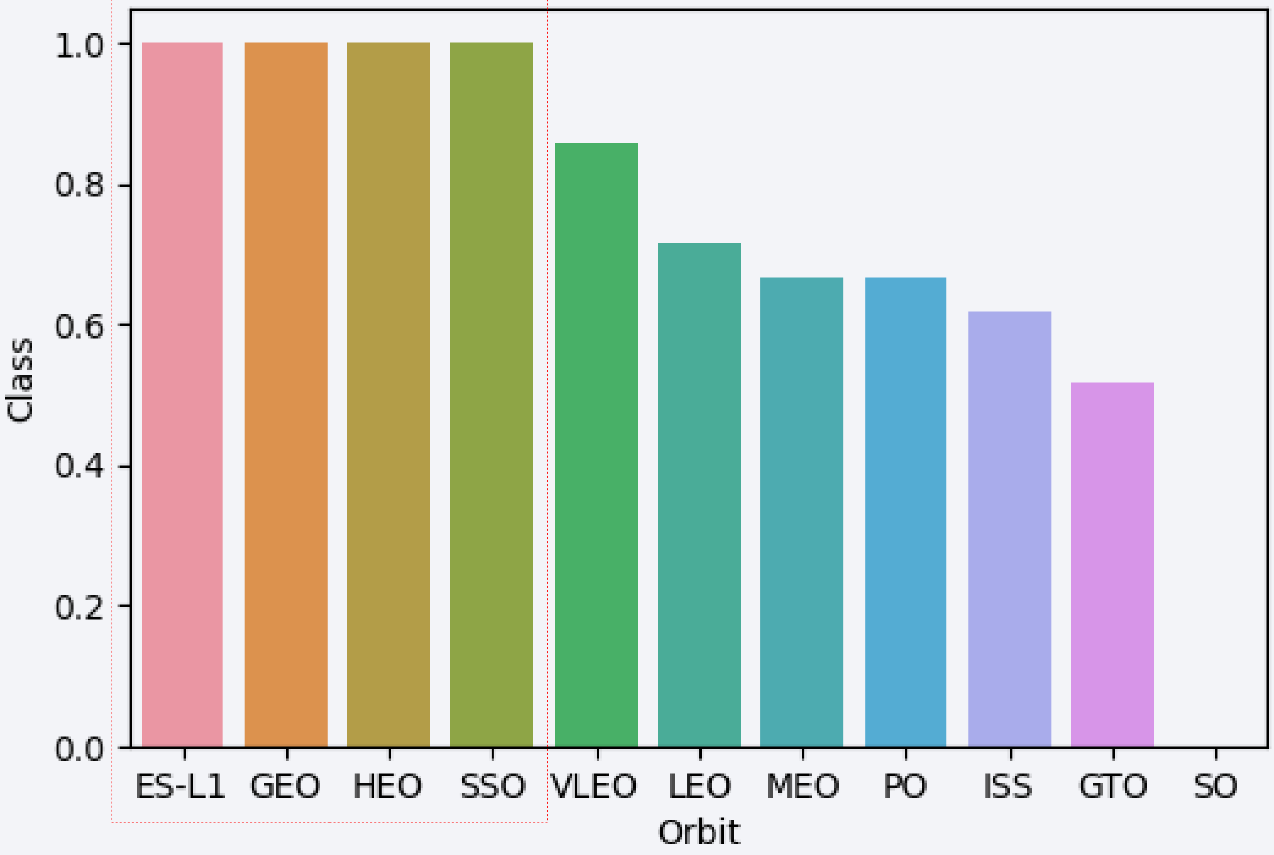
Payload vs. Launch Site

In KSC LC 39, tests involving lighter and heavier payload masses have obtained high success rates



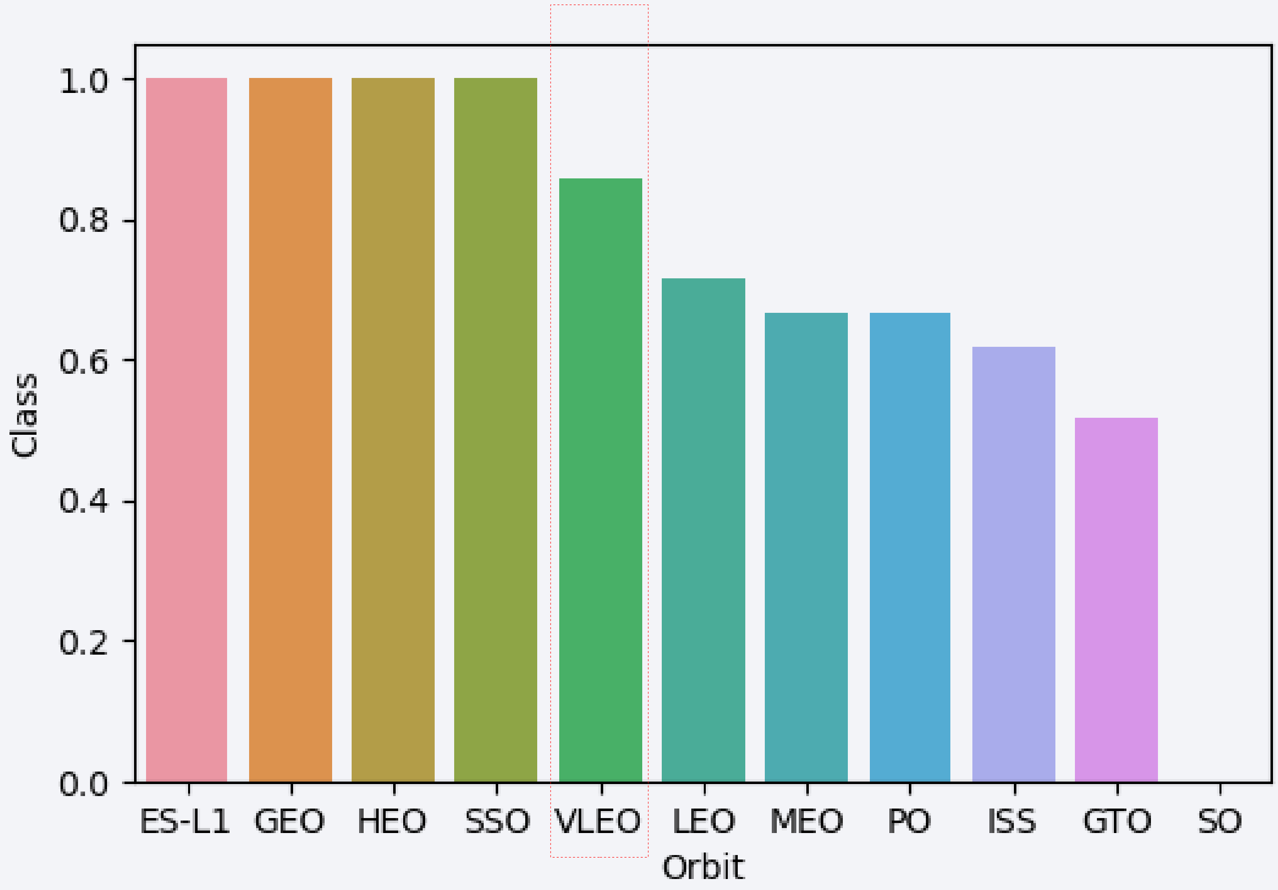
Success Rate vs. Orbit Type

- The highest success rates are related to orbits with low flight attempts
- ES-L1, GEO, and HEO are all medium to high-altitude orbits



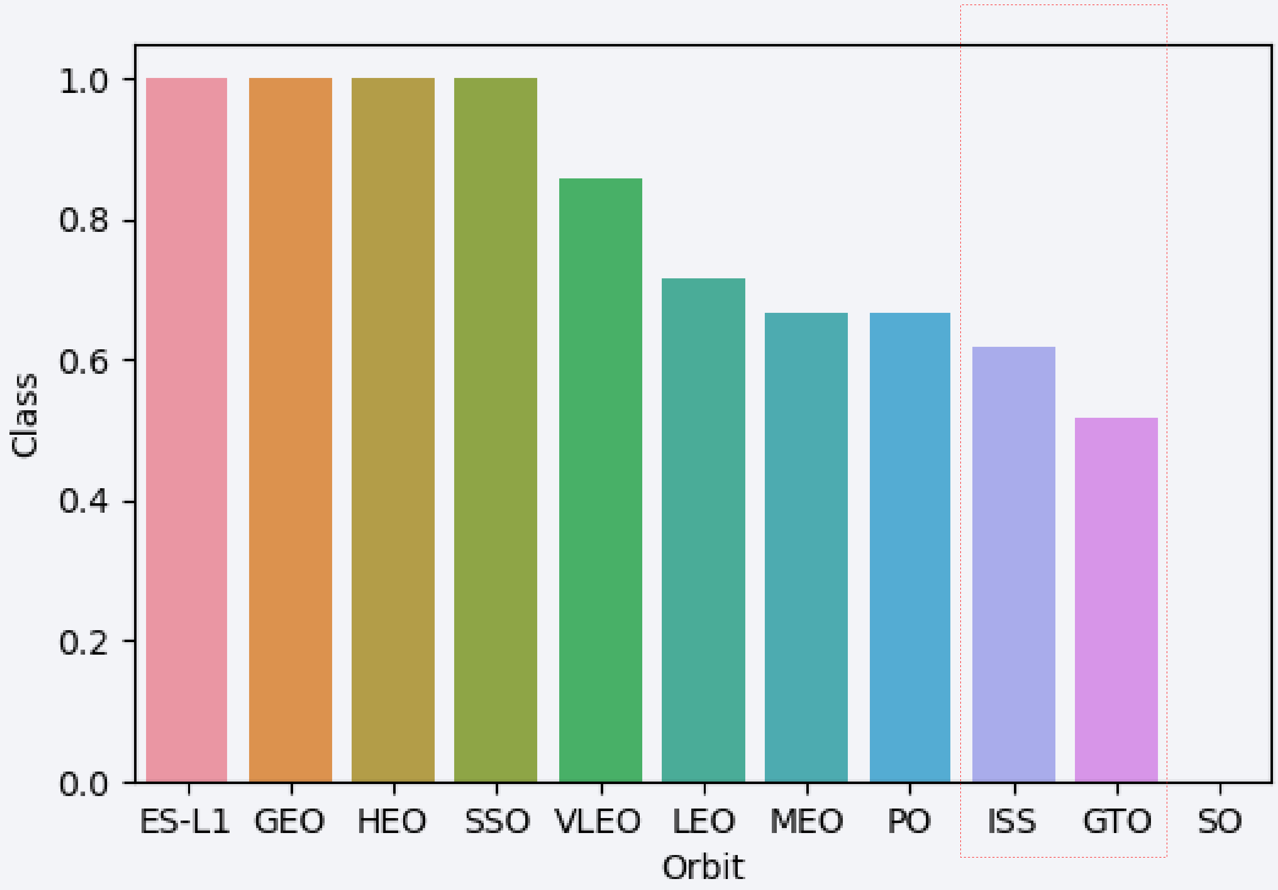
Success Rate vs. Orbit Type

VLEO have an average ratio between success rate and flight attempts



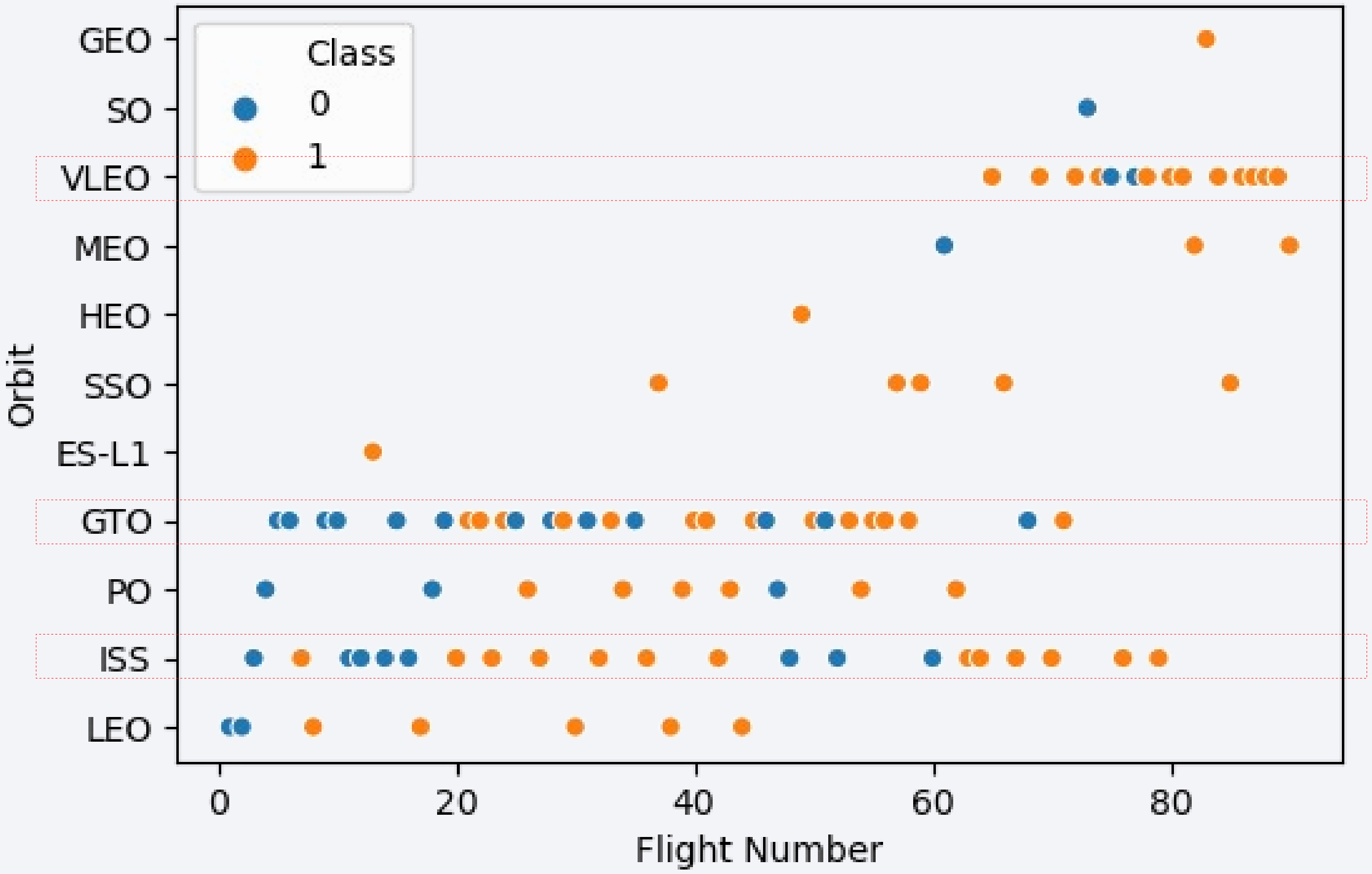
Success Rate vs. Orbit Type

GTO and ISS orbits have low success rate in more flight attempts



Flight Number vs. Orbit Type

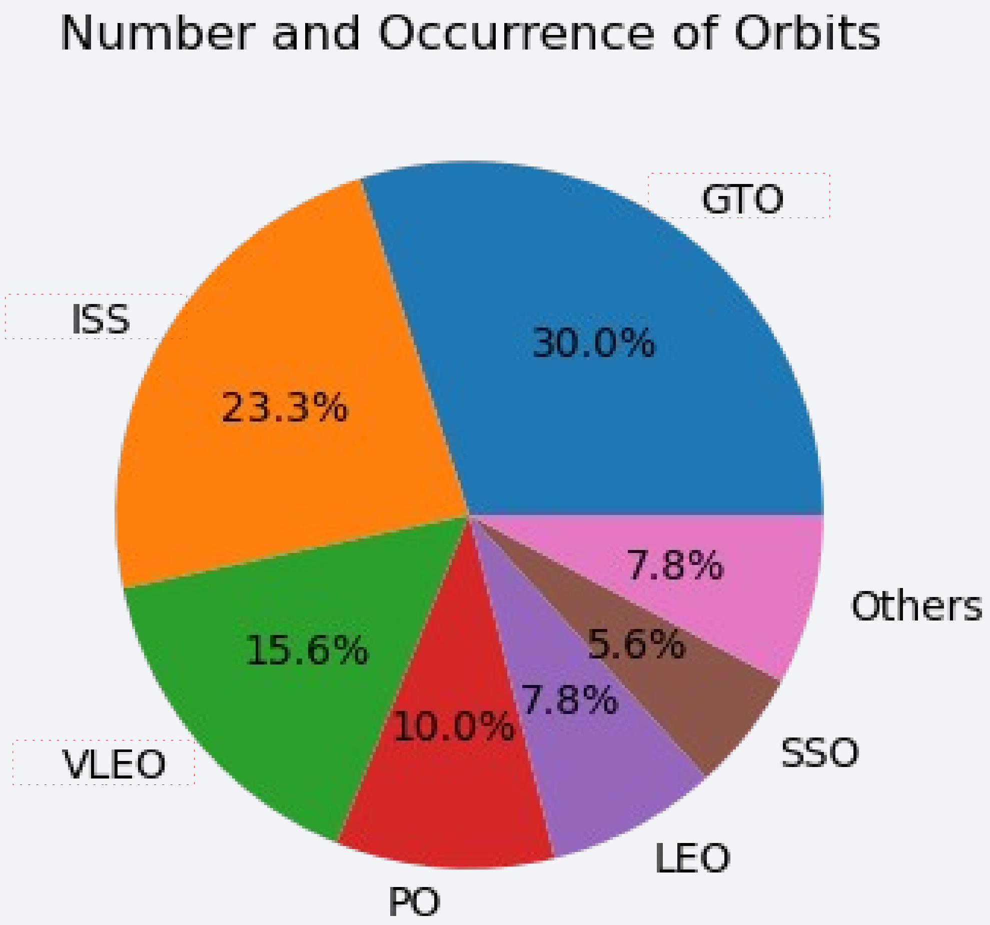
- GTO, ISS, and VLEO are the types of orbits with the most flight attempts
- They also appear to have a shorter gap between subsequential launches



Success Rate vs. Orbit Type

- GTO, ISS, and VLEO are the types of orbits with the most flight attempts

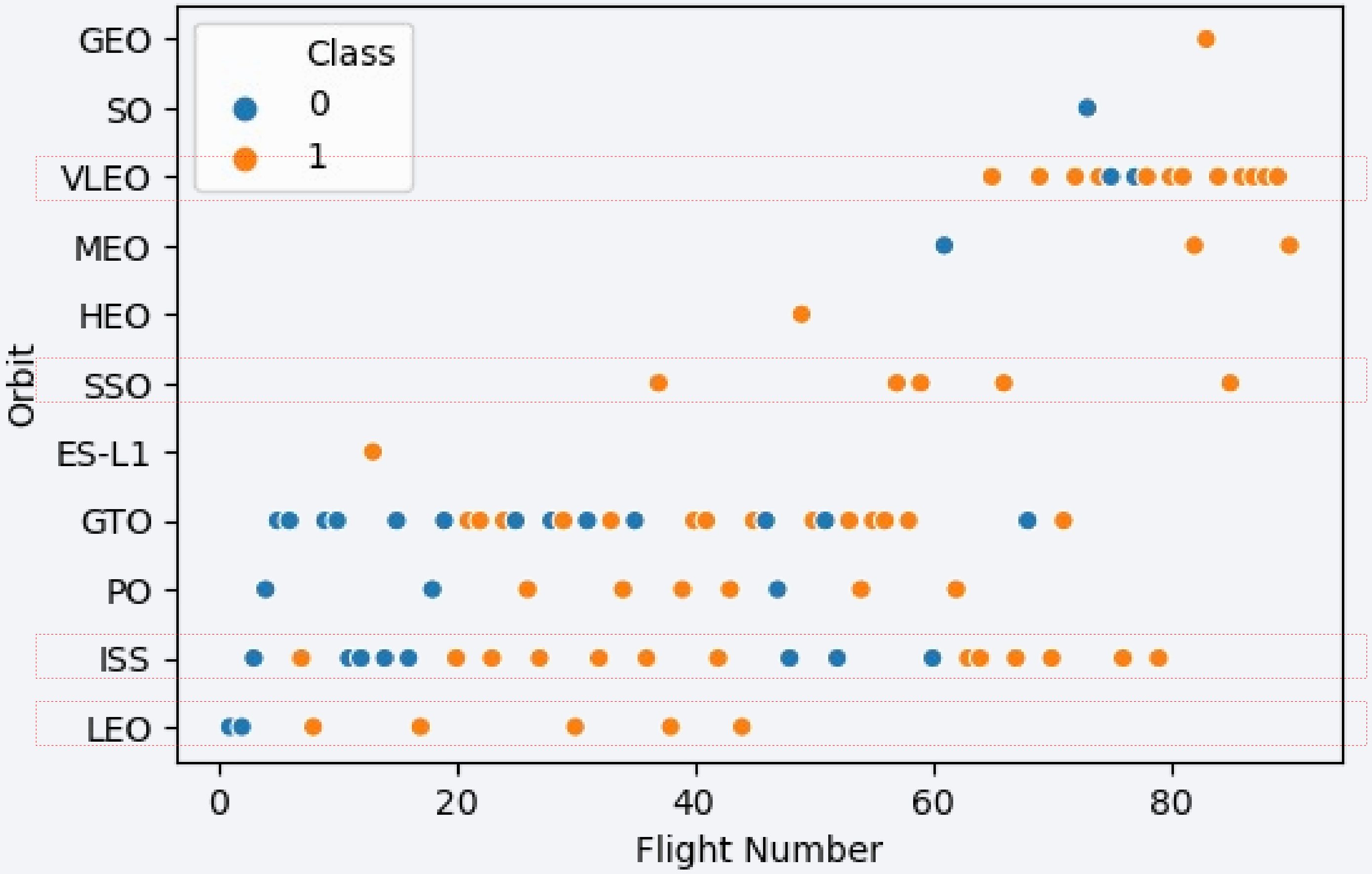
- They also appear to have a shorter gap between subsequential launches



Flight Number vs. Orbit Type

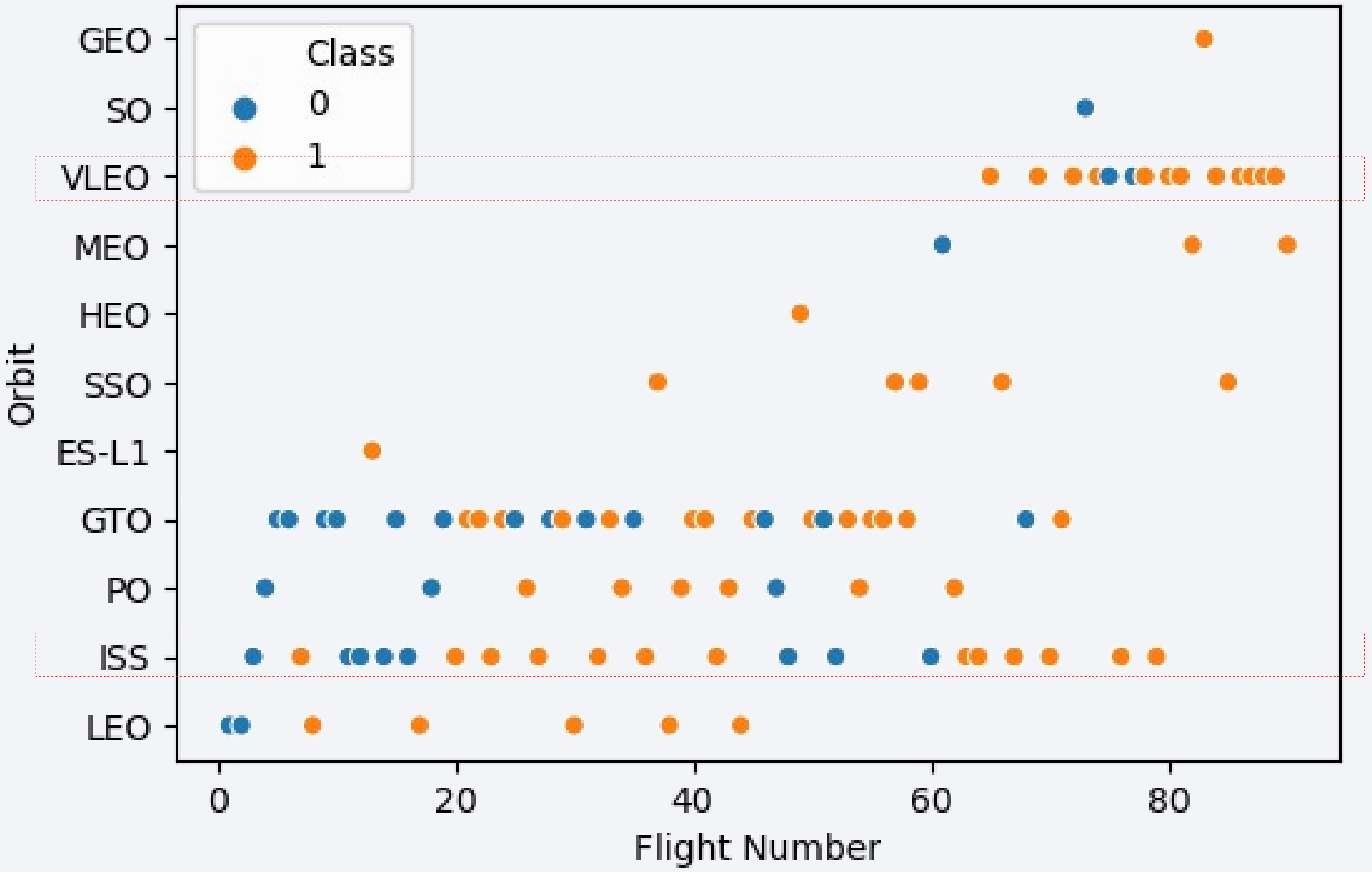
- VLEO, SSO, ISS, and LEO have high successful rates

- SSO and LEO have a small number of launches when compared to the other two



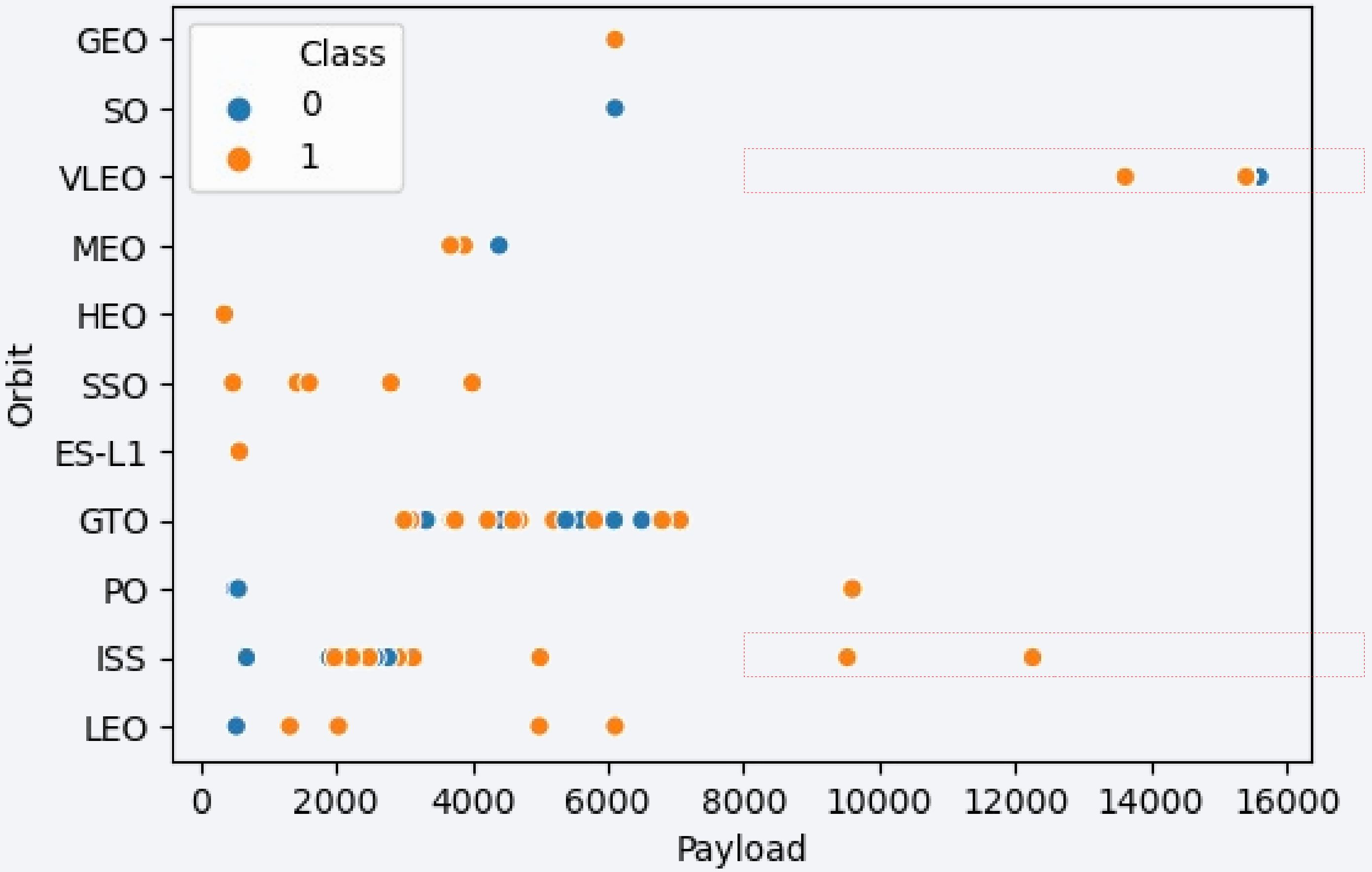
Flight Number vs. Orbit Type

- VLEO and ISS are orbits with similar altitudes
- Together, they may be used to demonstrate that tests in low orbits are very frequent



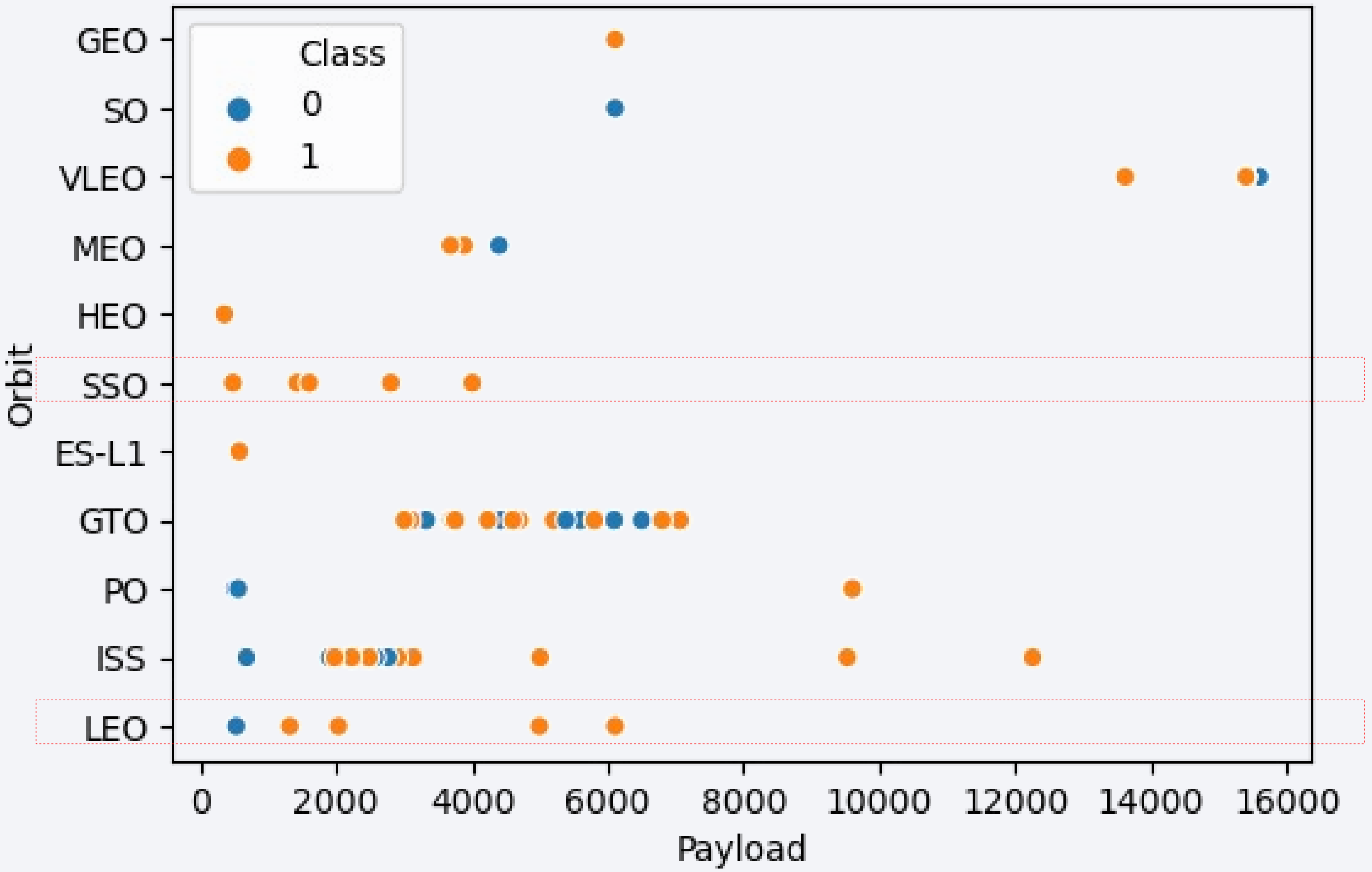
Payload vs. Orbit Type

- VLEO and ISS are orbits where heavy payload flights were attempted



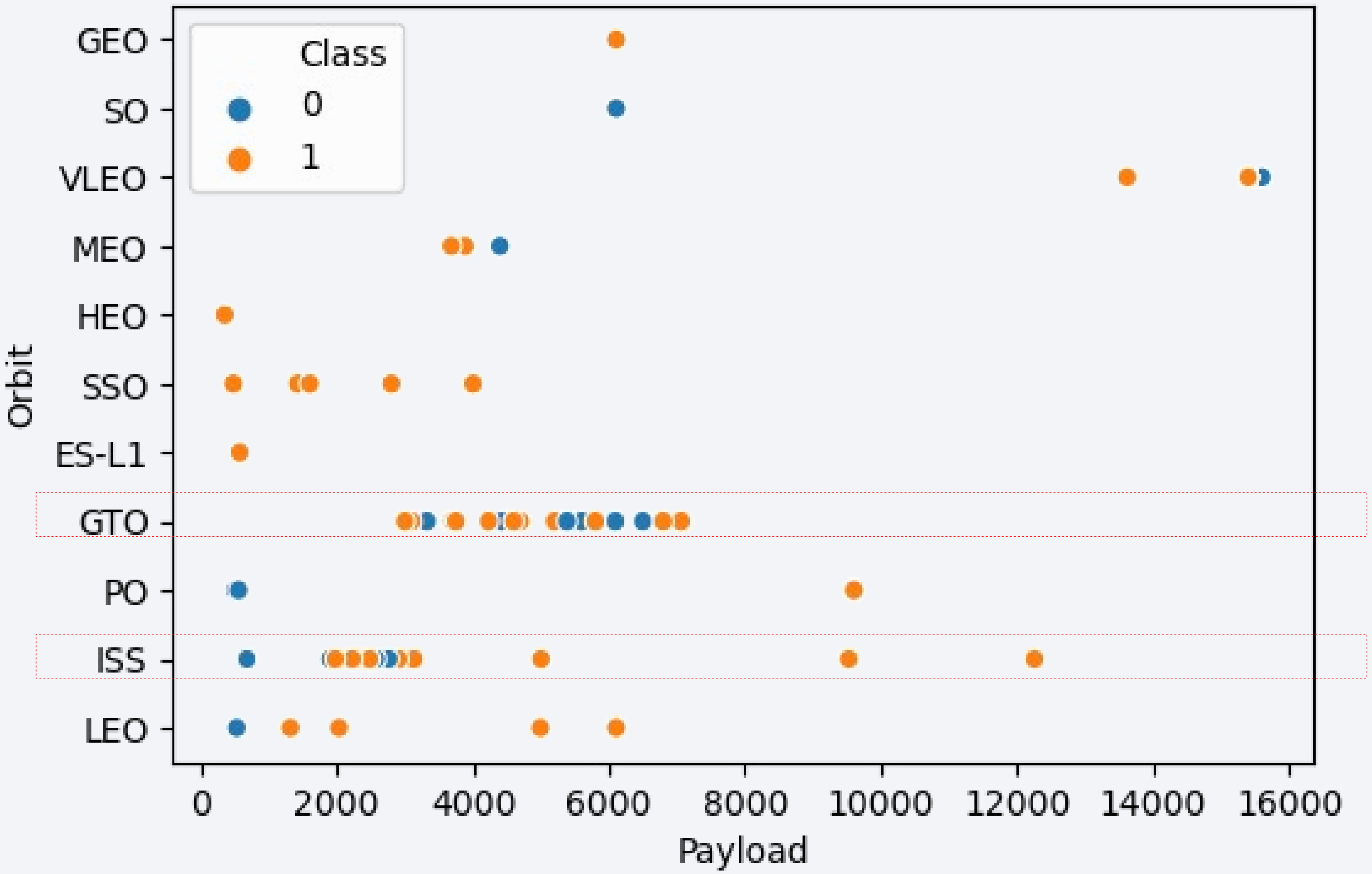
Payload vs. Orbit Type

- SSO and LEO (also low altitude orbits) have high successful rates on medium to low payloads with few launch attempts



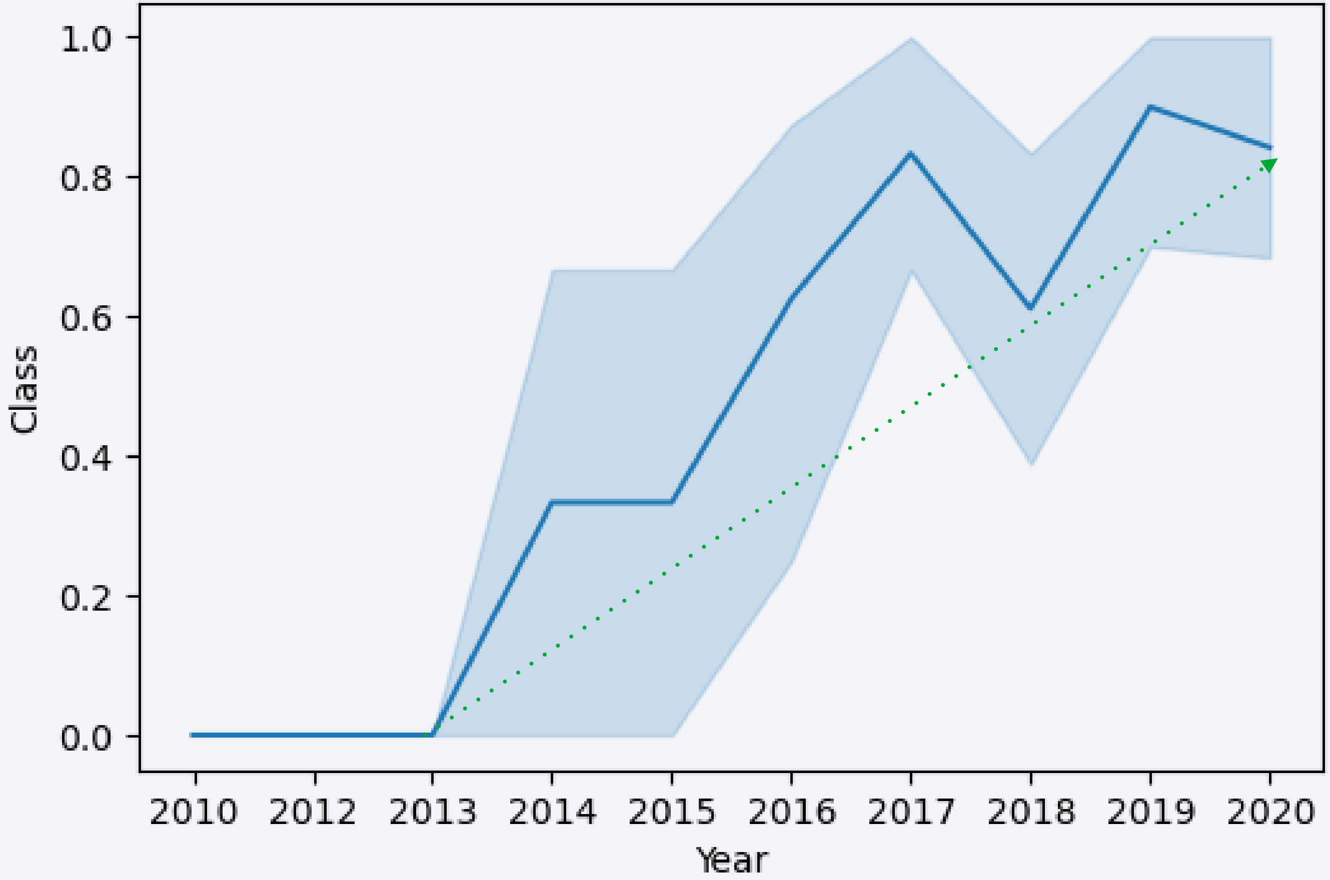
Payload vs. Orbit Type

GTO (high-altitude) and ISS (low-altitude) have a considerable number of launches within very concise payload ranges if compared to the other orbits



Launch Success Yearly Trend

The overall success rate has been increasing since 2013



All Launch Site Names

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

[8]:
```
%%sql
    SELECT DISTINCT Launch_Site
    FROM SPACEXTBL
    WHERE Launch_Site != 'None'
```

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

[8]:

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

SQL query fetching all the unique launch site names (DISTINCT)

Launch Site Names Begin with 'CCA'

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

```
[10]: %%sql
    SELECT Launch_Site
    FROM SPACEXTBL
    WHERE Launch_Site LIKE 'CCA%'
    LIMIT 5
```

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

[10]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

SQL query fetching the first 5 occurrences of launch site names begin with ‘CCA’

Total Payload Mass

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

```
[12]: %%sql
    SELECT SUM(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL
    WHERE Customer like 'NASA (CRS)%'

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

[12]:

SUM(PAYLOAD_MASS__KG_)
48213.0

SQL query fetching the total payload carried by boosters from NASA (SUM(PAYLOAD_MASS_KG_))

Average Payload Mass by F9 v1.1

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

```
[14]: %%sql
    SELECT AVG(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL
    WHERE Booster_Version like 'F9 v1.1%'

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

SQL query returning the average payload mass carried by booster version F9 v1.1

(AVG(PAYLOAD_MASS_KG_))

First Successful Ground Landing Date

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

```
[16]: %%sql
    SELECT MIN(Date)
    FROM SPACEXTBL
    WHERE Landing_Outcome == 'Success (ground pad)'

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

[16]:

MIN(Date)
01/08/2018

SQL query fetching the dates of the first successful landing outcome on ground pad

(MIN(Date))

Successful Drone Ship Landing with Payload 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

[18]:
```
%%sql
    SELECT Booster_Version
    FROM SPACEXTBL
    WHERE Landing_Outcome == 'Success (drone ship)'
    AND PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

[18]:

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

SQL query listing the names of boosters which 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

List the total number of successful and failure mission outcomes

```
[20]: %%sql
    SELECT Mission_Outcome, COUNT(Mission_Outcome) AS count
    FROM SPACEXTBL
    WHERE Mission_Outcome != 'None'
    GROUP BY (Mission_Outcome)
```

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

[20]:

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

SQL query calculating the total number of successful and failure mission outcomes

(COUNT(Mission_Outcome))

Boosters Carried Maximum Payload

SQL query listing the names of the boosters which have carried the maximum payload mass

List the names of the booster_versions which have carried the maximum payload mass

[22]:
```
%%sql
    SELECT DISTINCT(Booster_Version)
    FROM SPACEXTBL
    WHERE PAYLOAD_MASS__KG_ = (
                                SELECT MAX(PAYLOAD_MASS__KG_)
                                FROM SPACEXTBL
                                )
    GROUP BY Booster_Version
```

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

[22]:

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

2015 Launch Records

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

[24]:
```
%%sql
    SELECT SUBSTR(Date, 4, 2) AS Month, Landing_Outcome, Booster_Version, Launch_Site
    FROM SPACEXTBL
    WHERE Landing_Outcome == 'Failure (drone ship)'
    AND SUBSTR(Date,7,4)='2015'
```

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

[24]:

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

SQL query listing 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

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
[26]: %%sql
    SELECT Date, Landing_Outcome, COUNT(Landing_Outcome) AS count
    FROM SPACEXTBL
    WHERE Landing_Outcome like '%Success%'
    AND (SUBSTR(Date,7,4) || SUBSTR(Date,4,2) || SUBSTR(Date,1,2)) BETWEEN '20100604' AND '20170320'
    GROUP BY Landing_Outcome
    ORDER BY count desc
```

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

[26]:

Date	Landing_Outcome	count
22/12/2015	Success (ground pad)	5
04/08/2016	Success (drone ship)	5

SQL query 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

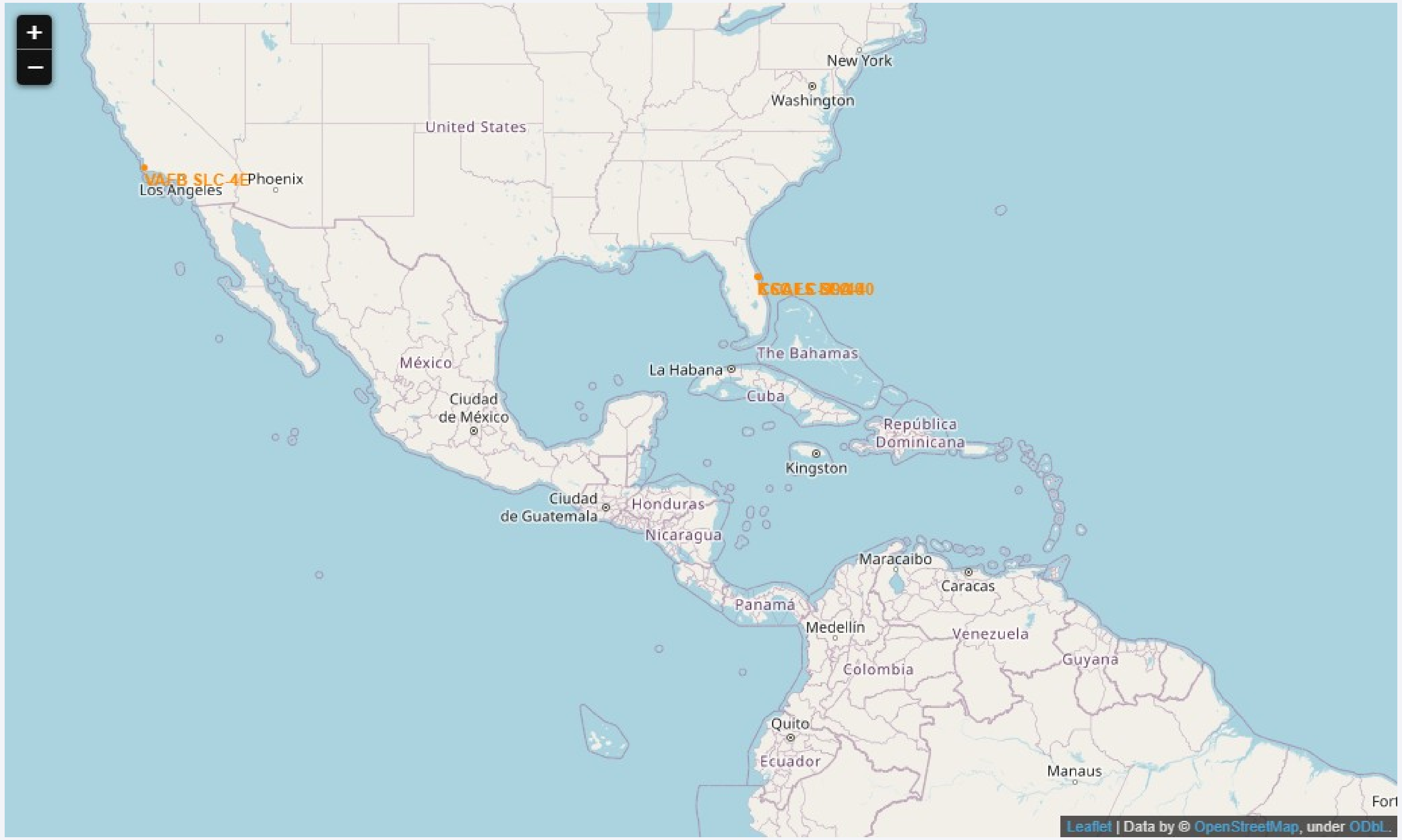
Section 3

Launch Sites Proximities Analysis

Folium Map: Launch Sites Location

Points in common:

- All launch sites are located as Southern as possible
- Proximity to highways, railways and airports
- Proximity to coastline
- Not so close to dense populated areas

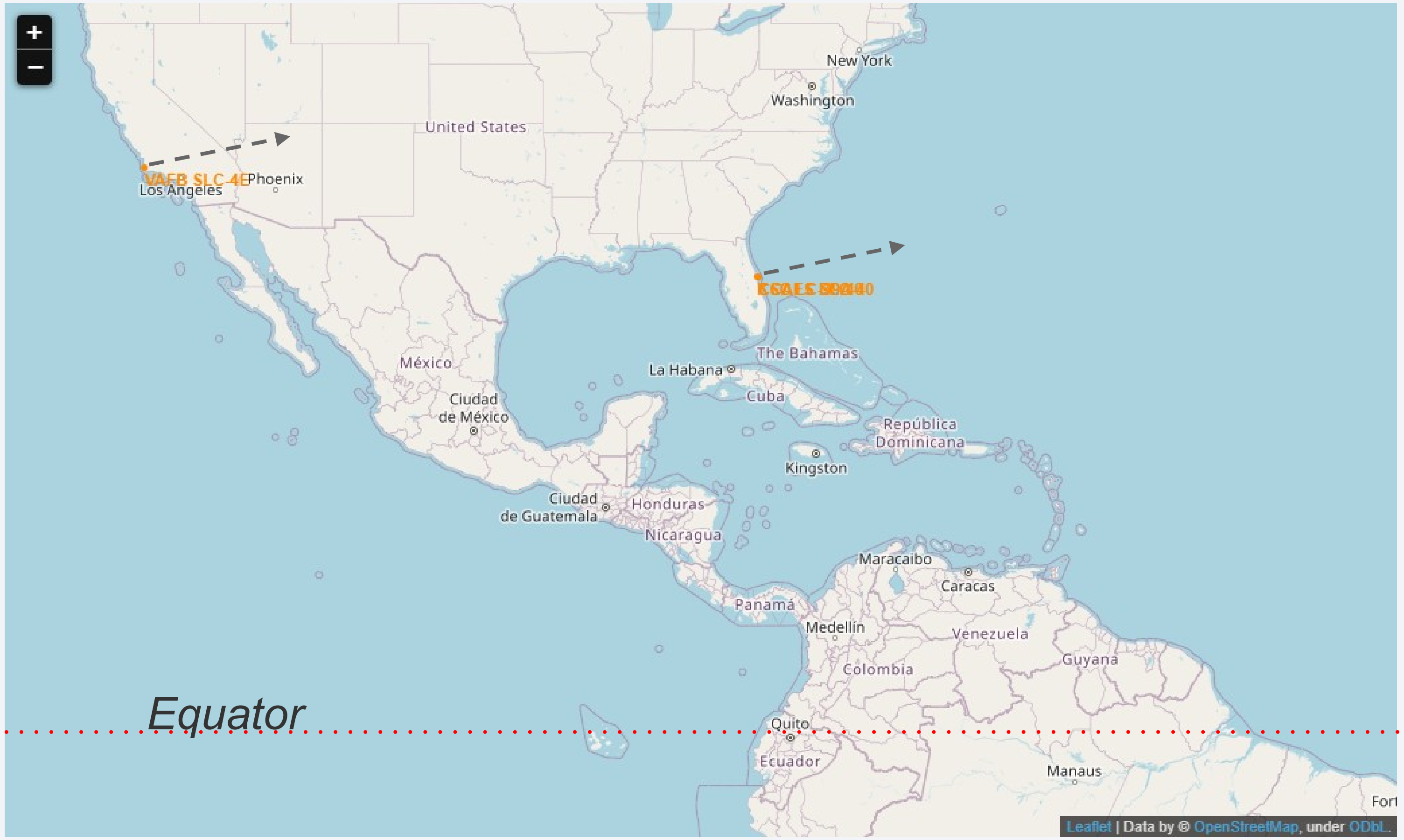


Folium Map: Launch Sites Location

Reason 1:

Rockets can most easily reach satellite orbits if launched near the equator in an easterly direction, as this maximizes use of the Earth's rotational speed (465 m/s at the equator)¹

https://en.wikipedia.org/wiki/Spaceport

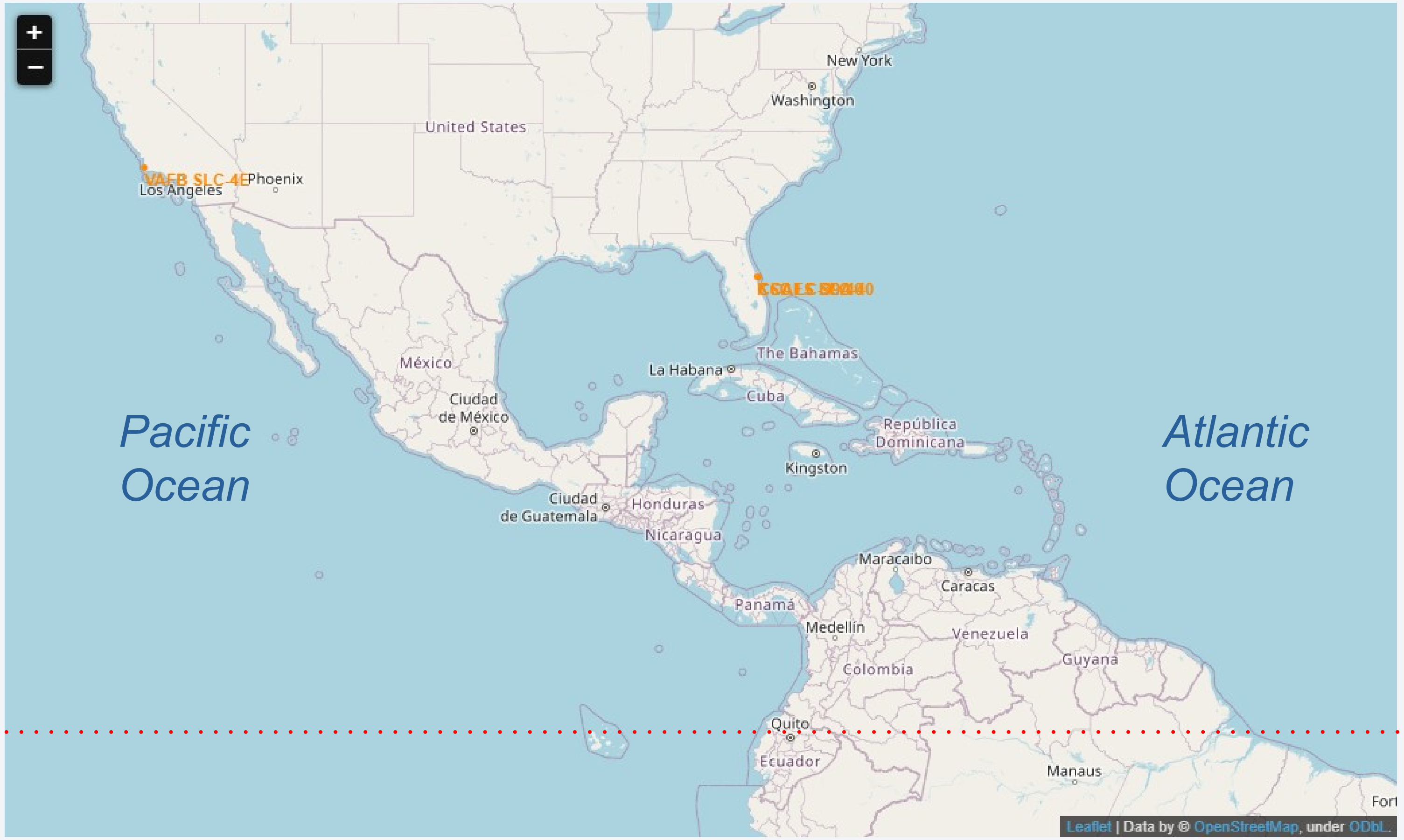


Folium Map: Launch Sites Location

Reason 2:

Launch sites are usually built close to major bodies of water, as far as possible away from major population centers, to ensure that no components are shed over populated areas[2]

https://en.wikipedia.org/wiki/Spaceport

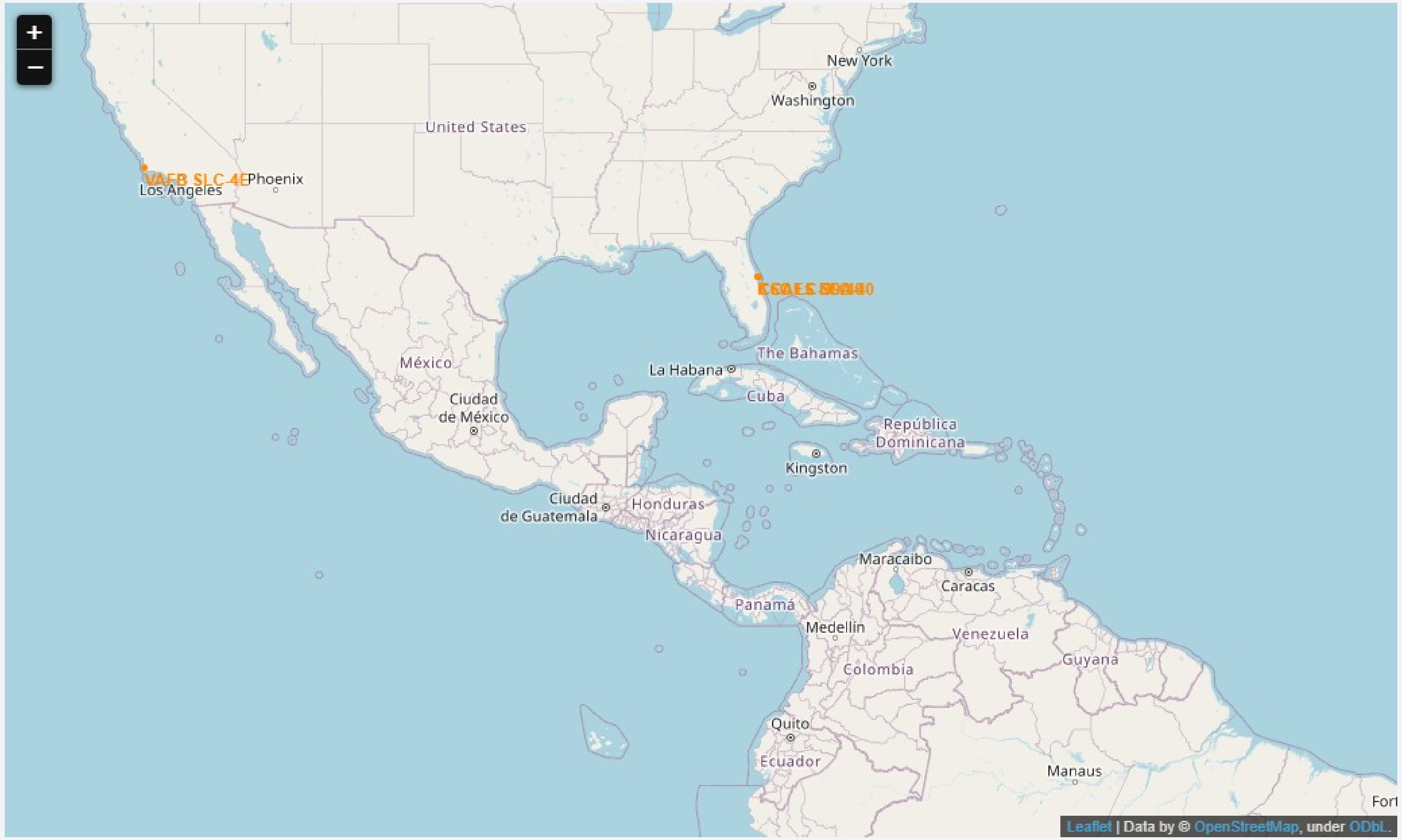


Folium Map: Launch Sites Location

Reason 3:

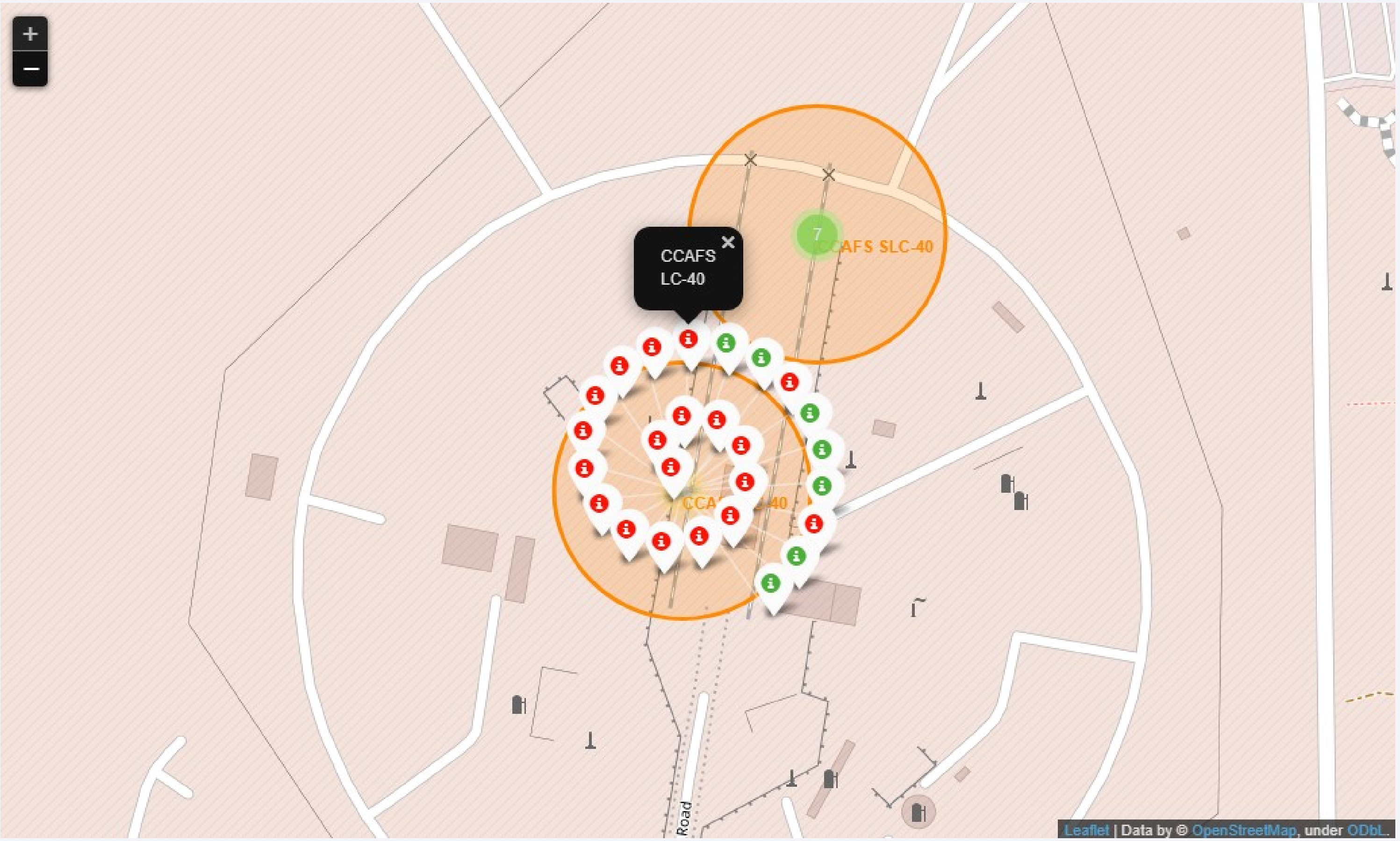
Planned sites of spaceports for sub-orbital tourist spaceflight often make use of existing ground infrastructure, including runways[3]

https://en.wikipedia.org/wiki/Spaceport



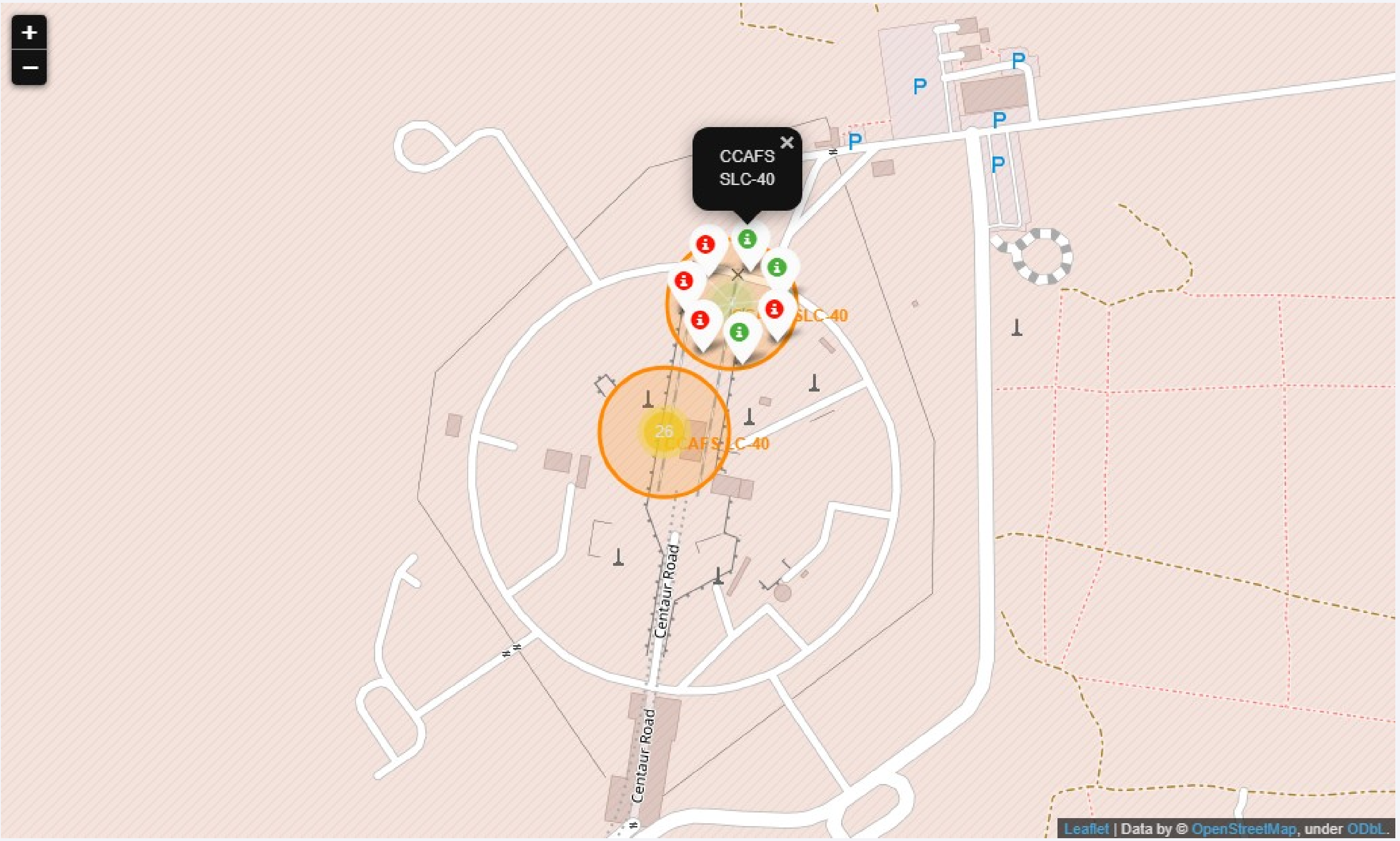
Folium Map: Launch Outcomes for CCAFS LC-40

- CCAFS LC-40 and CCAFS SLC-40 have most of the launches
- It's also where the evolution of outcome rates is more outstanding



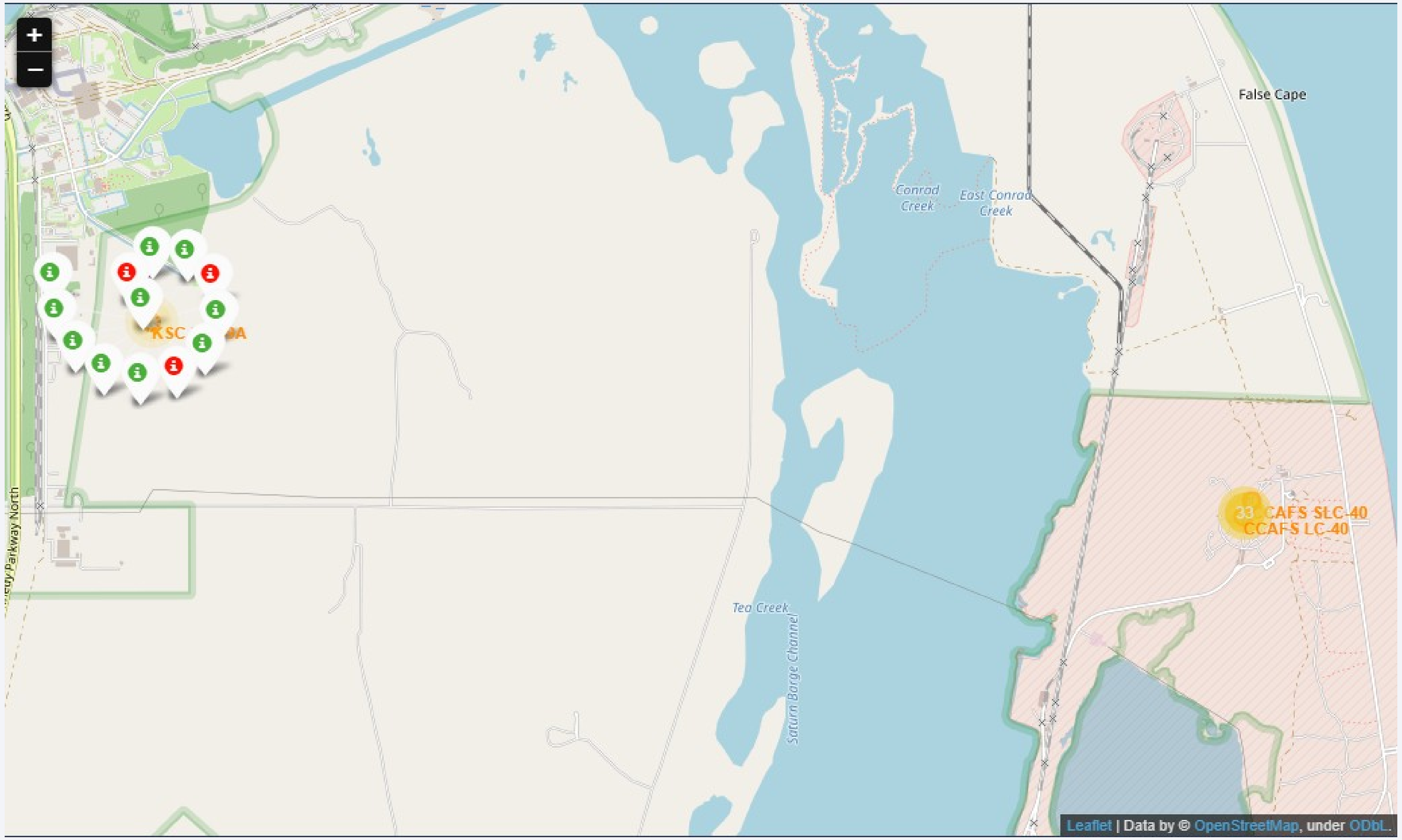
Folium Map: Launch Outcomes for CCAFS SLC-40

- CCAFS LC-40 and CCAFS SLC-40 have most of the launches
- It's also where the evolution of outcome rates is more outstanding



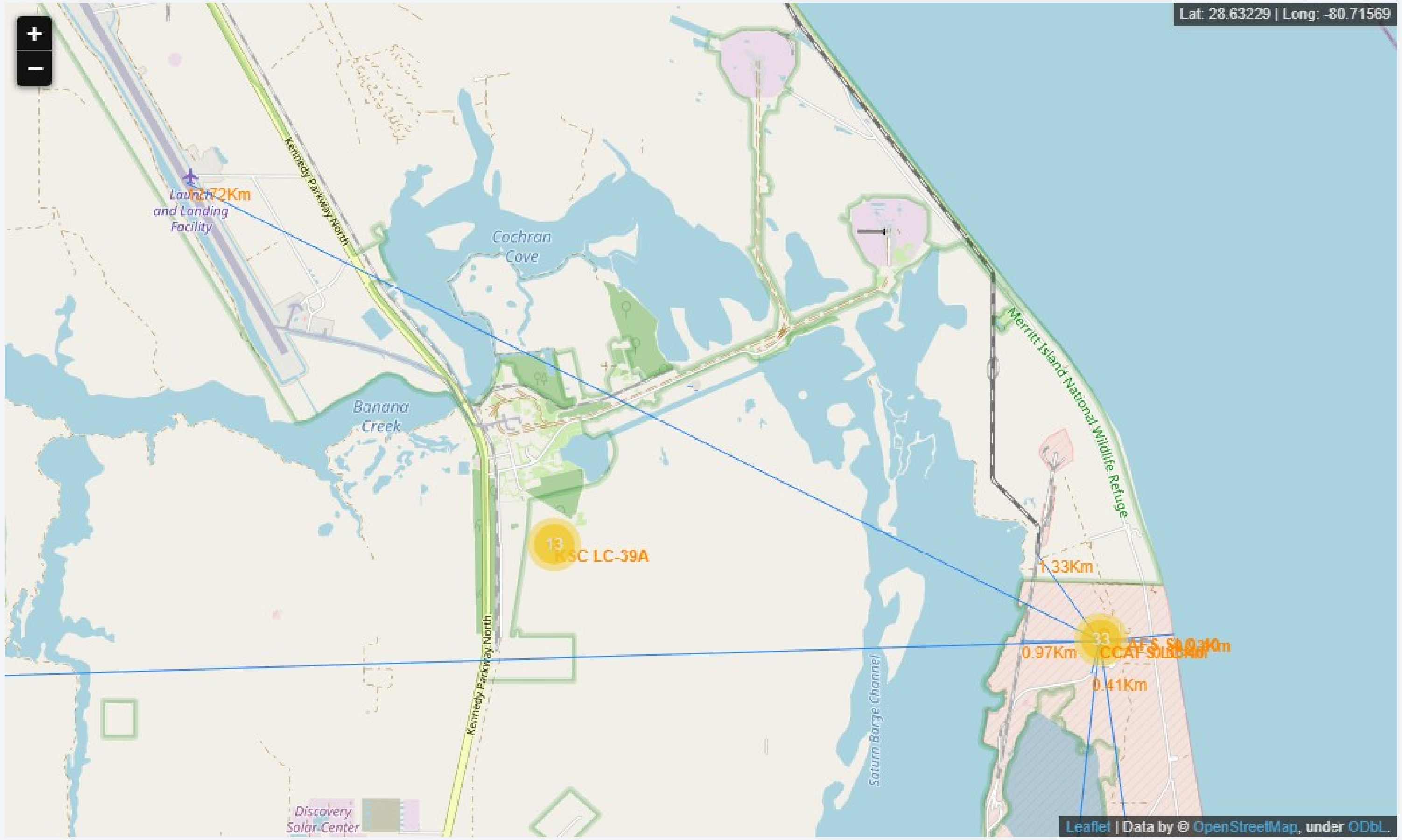
Folium Map: Launch Outcomes for KSC LC-39A

- KSC LC-39A is very close to CCAFS LC-40 and CCAFS SLC-40 and practically shares the same infrastructure
- It is the launch site with the higher success rate among the three



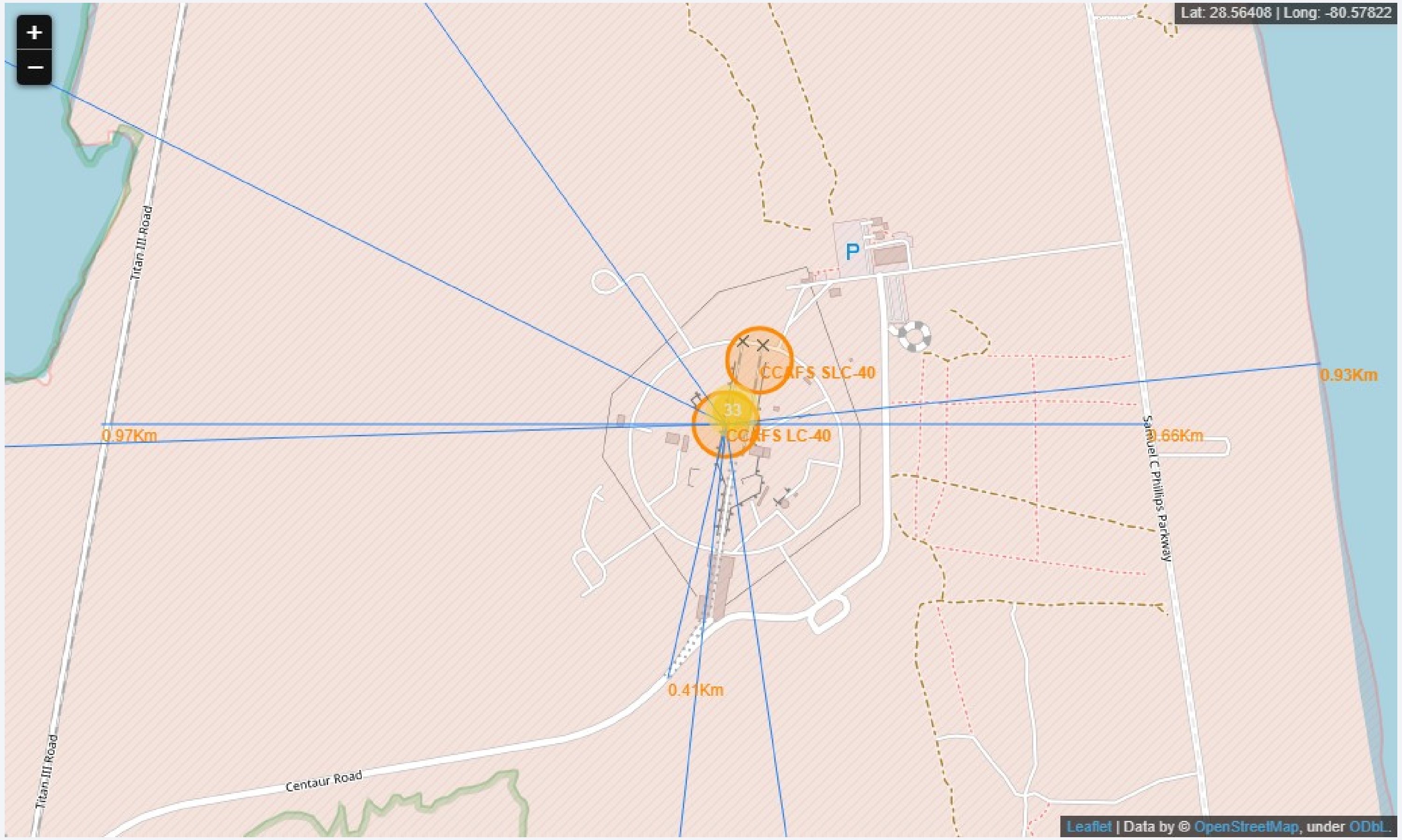
Folium Map: Nearby Infrastructure

- CCAFS SLC-40, CCAFS LC-40 and KSC LC-39A sites are where 85.5% of the launches have taken place
- Very close to coastline



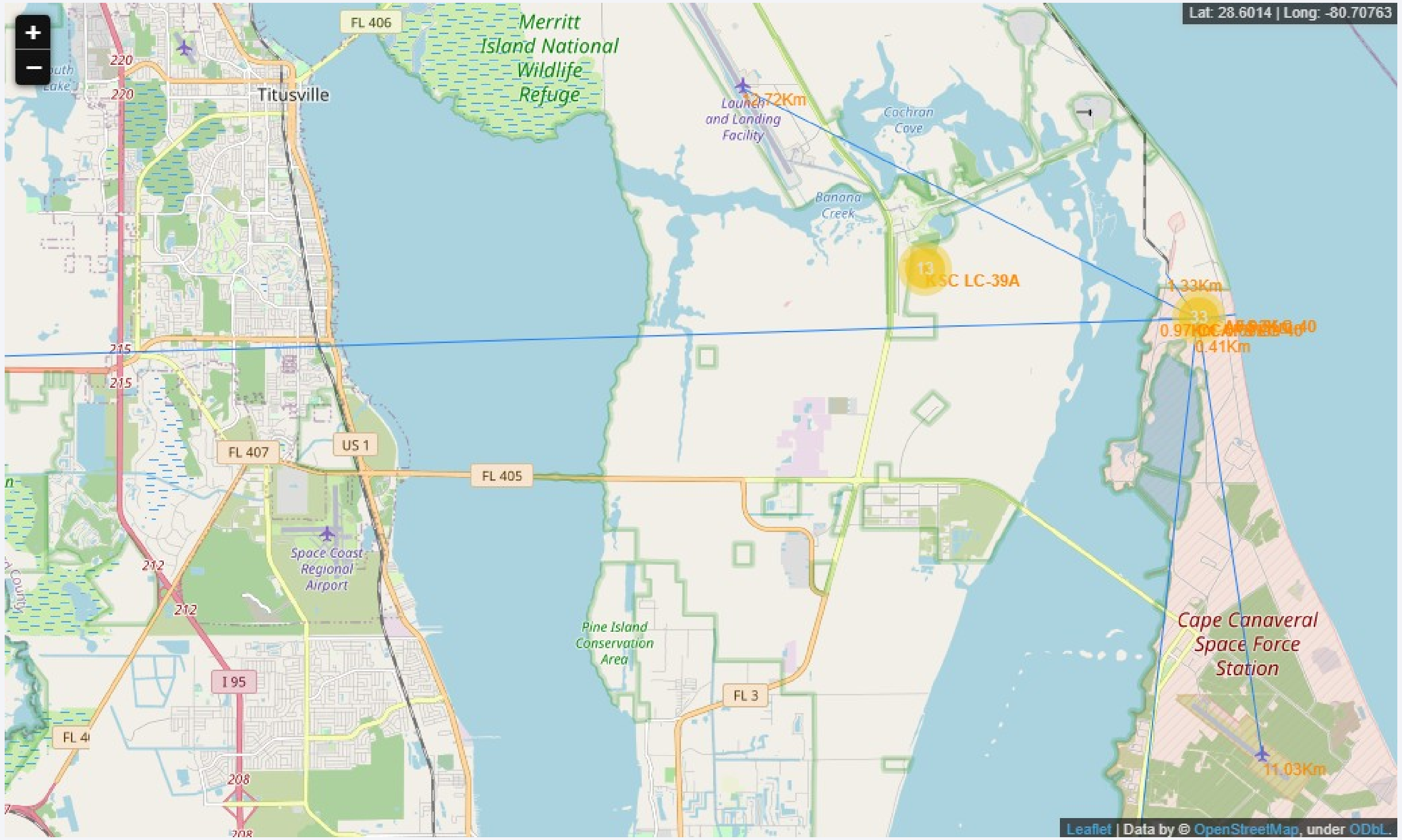
Folium Map: Nearby Infrastructure

Dedicated facilities and direct runways



Folium Map: Nearby Infrastructure

As far away as possible from major population centers

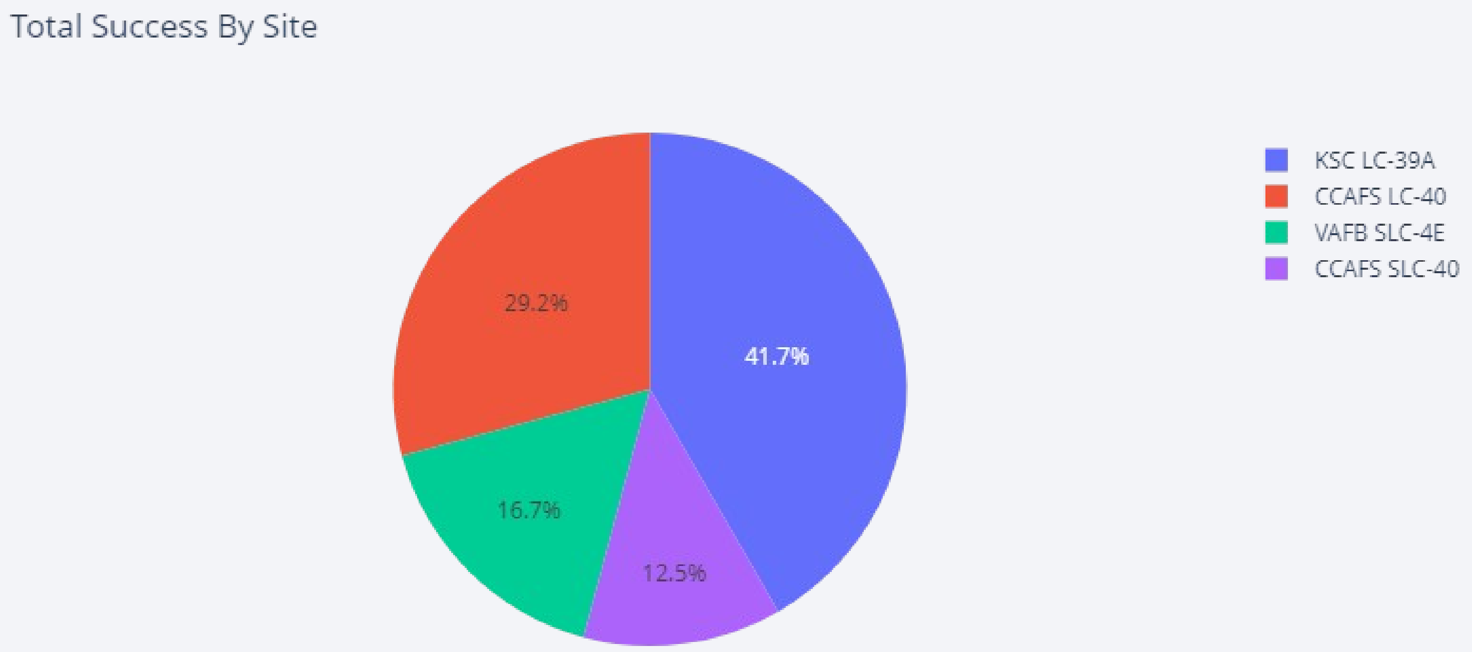


Section 4

Build a Dashboard with Plotly Dash

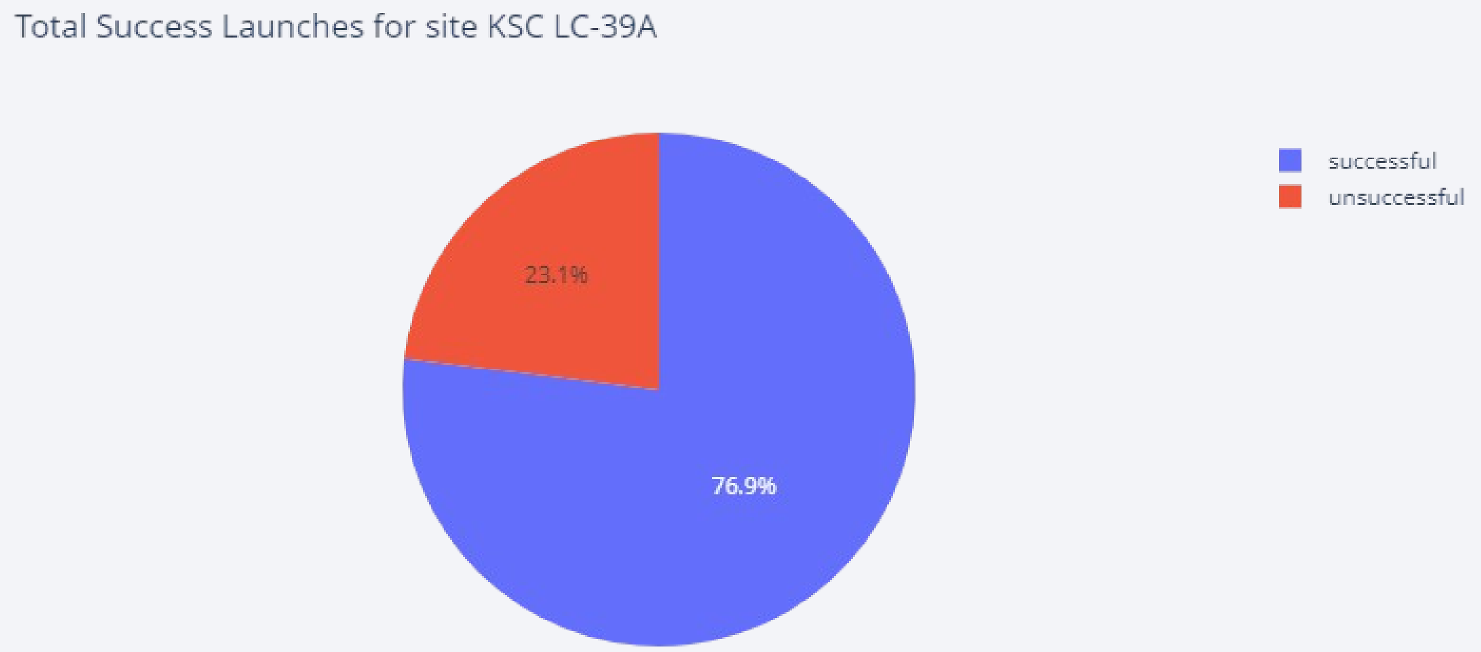
Dashboard: Successful Launch Rates Comparison

- KSC LC-39A has the largest successful launches
- CCAFS LC-40 and CCAFS SLC-40 have, together, a similar success rate to KSC LC-39A



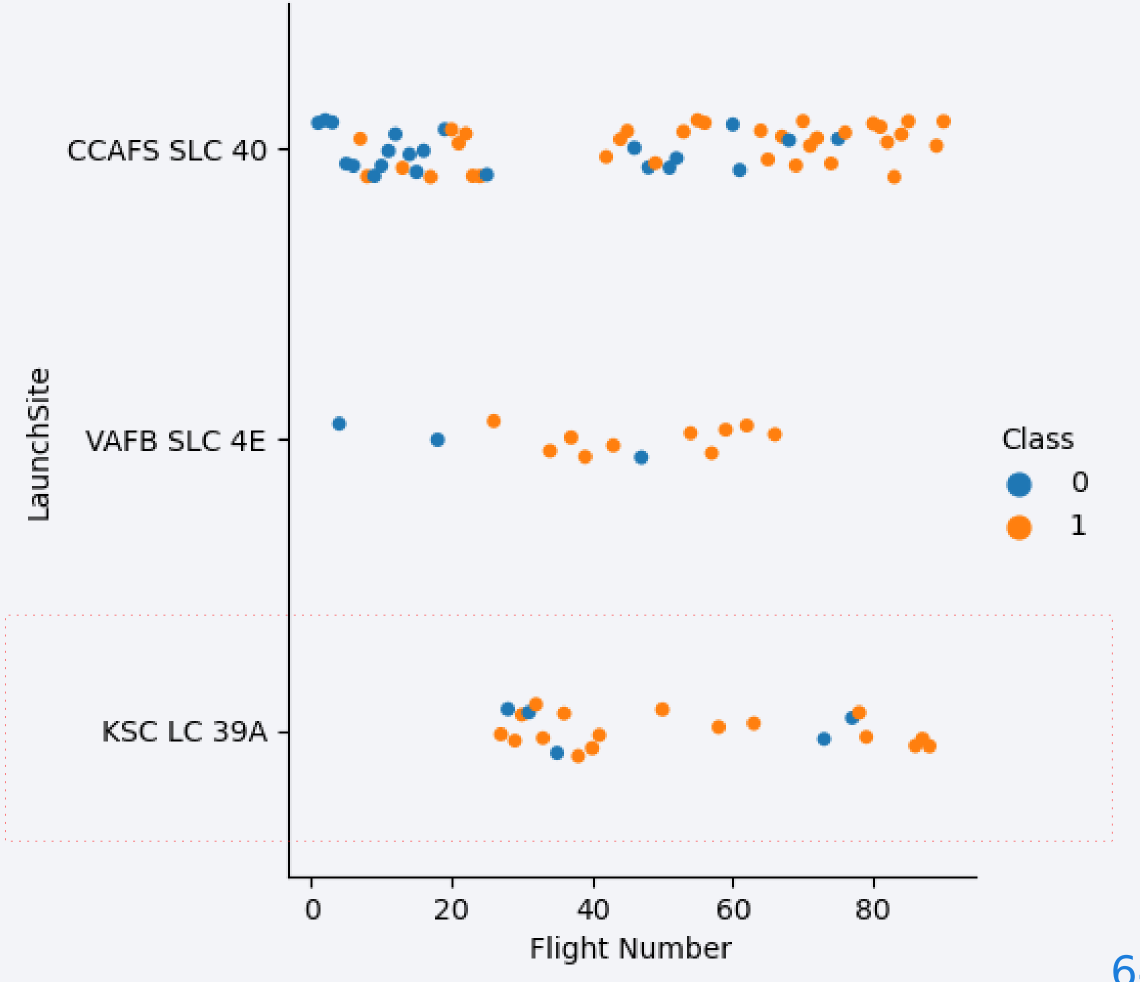
Dashboard: The Most Successful Launch Site

KSC LC-39A has the highest successful launch rate: 76.9%



Dashboard: The Most Successful Launch Site

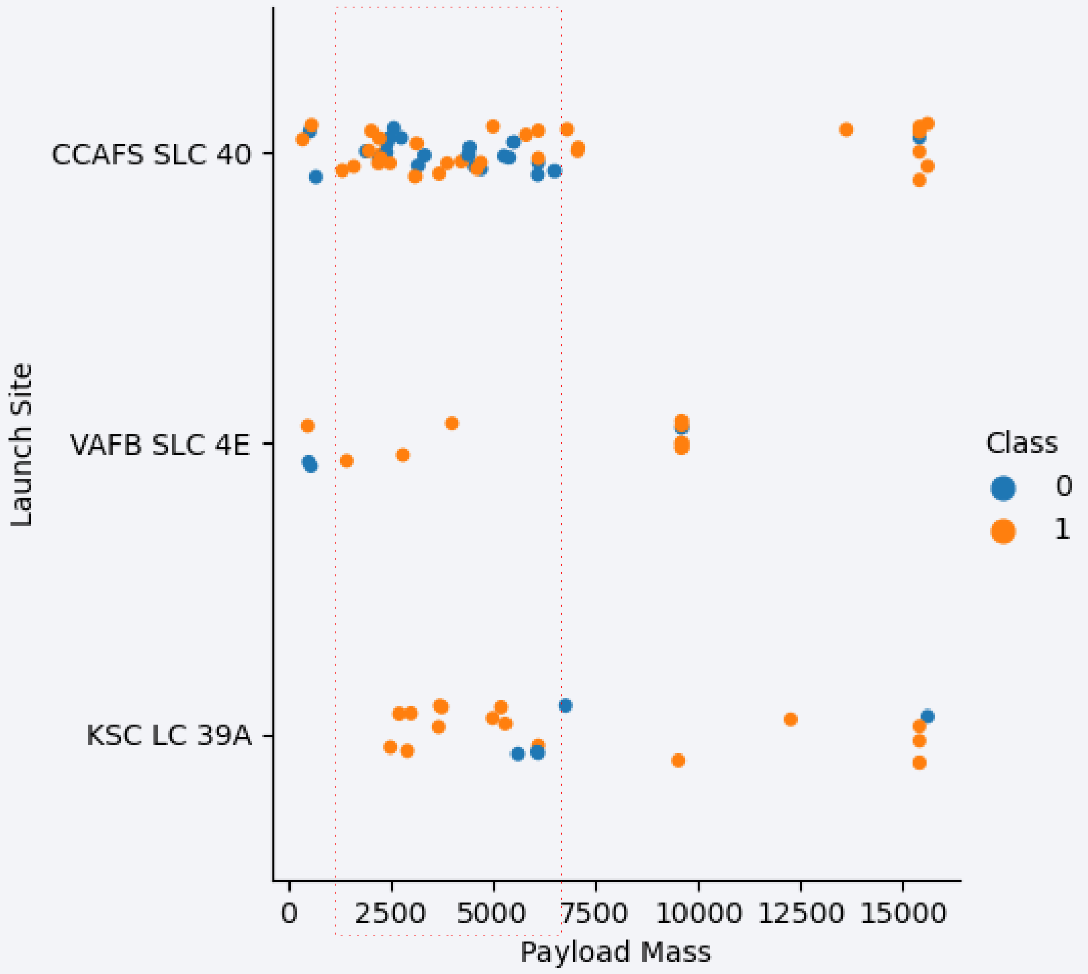
- Launch tests started later in KSC LC 39A than the other two
- Tests involved lighter and heavier payload masses
- It has a remarkable success rate
- Such results are prone to increase as the flight number increases



Dashboard: Correlation Payload vs Launch Success

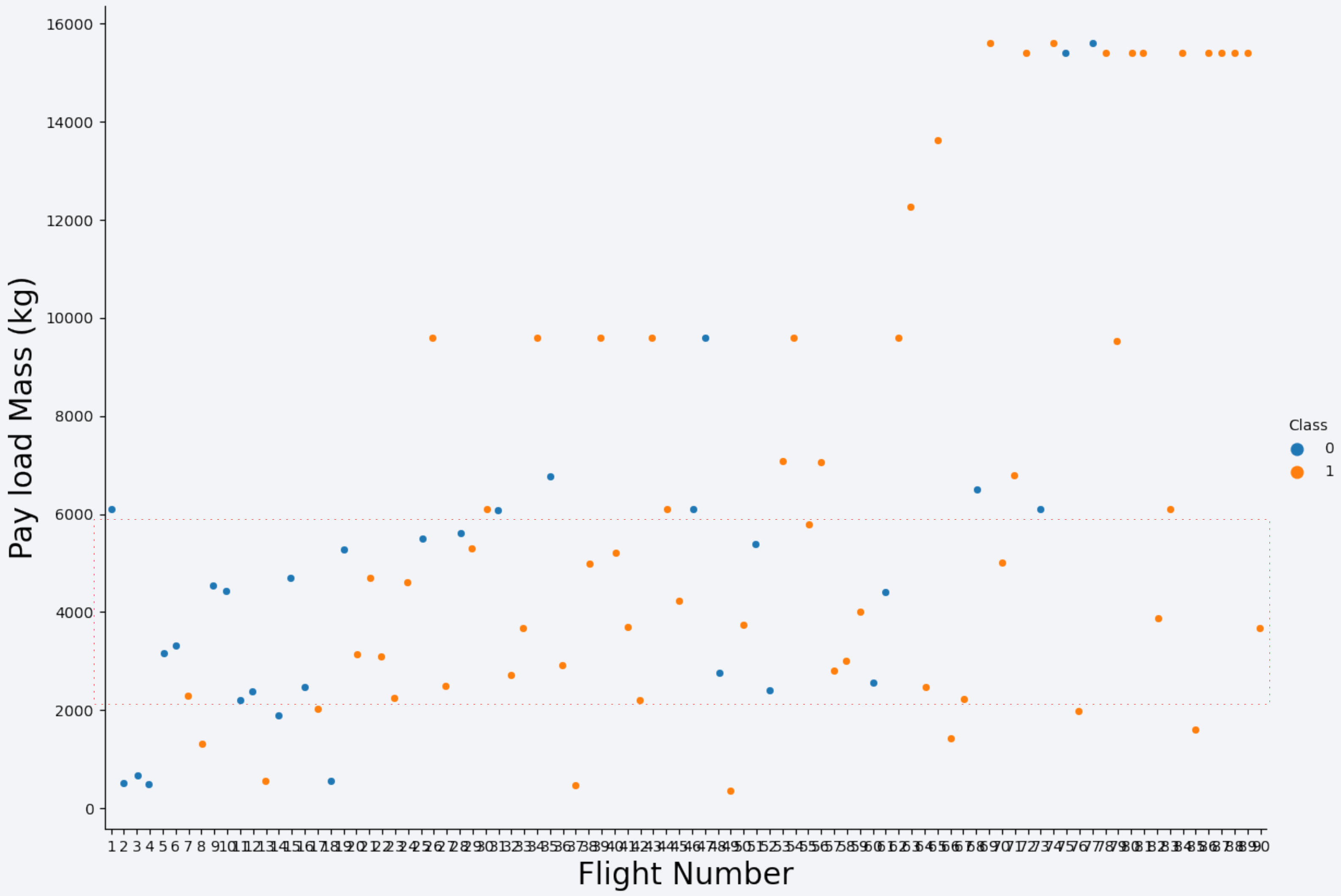
- Most successful launches carried a payload weighing between 2000 Kg and 6000 Kg
- Within this group, most flights carried between 2000 Kg and 4000 Kg

Dashboard: Correlation Payload vs Launch Success



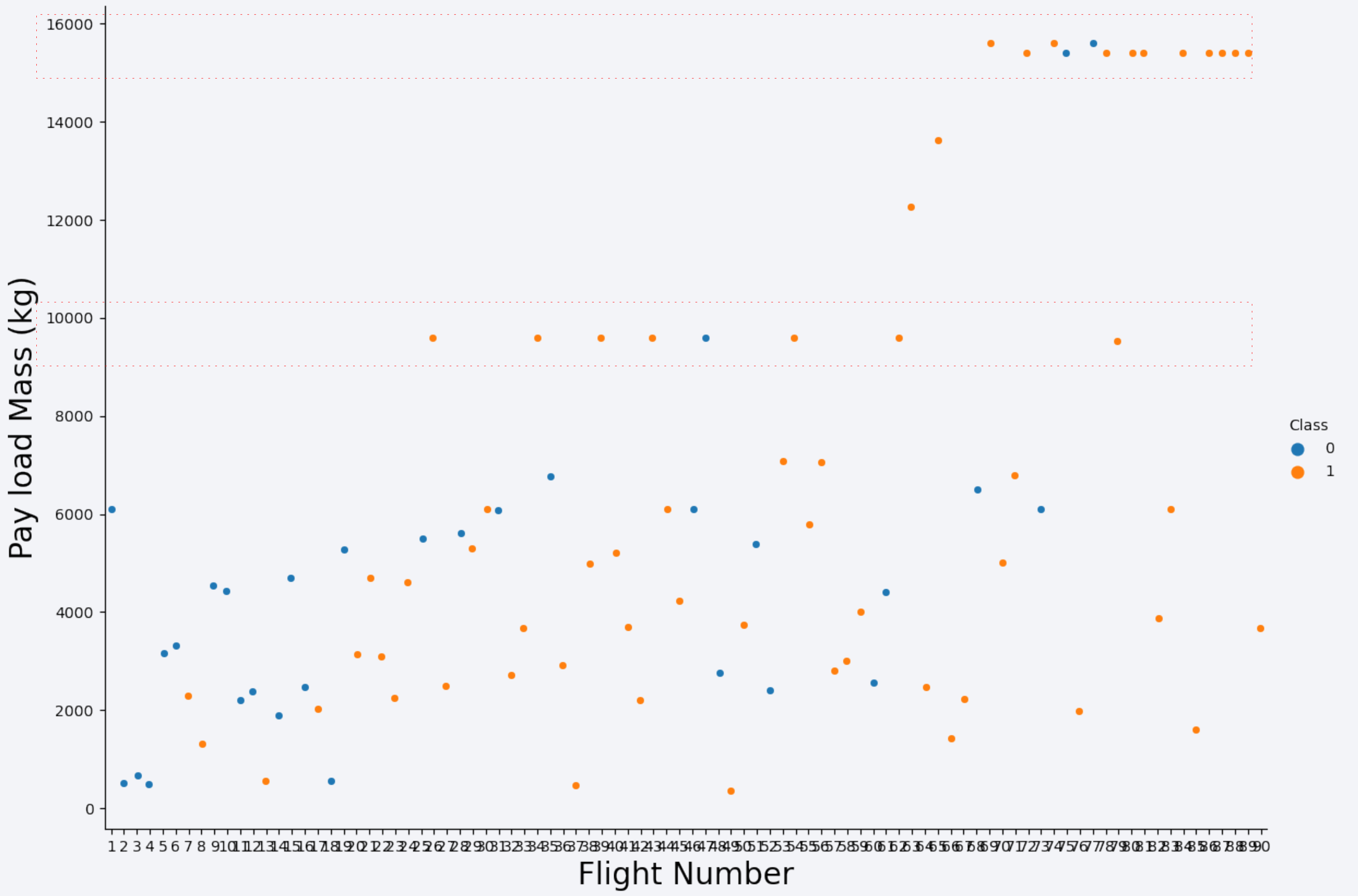
2000 Kg to 6000 Kg is also the most tested payload range

Dashboard: Correlation Payload vs Launch Success



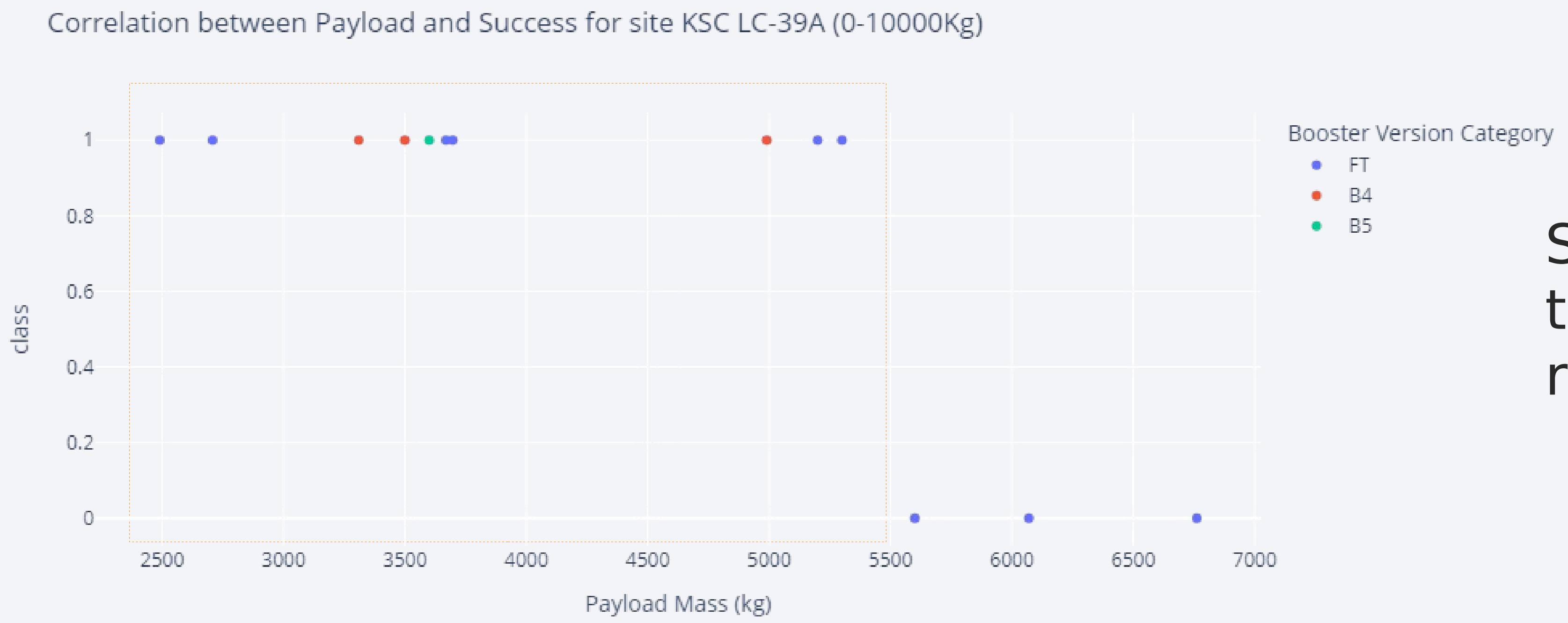
2000 Kg to 6000 Kg is also the most tested payload range

Dashboard: Correlation Payload vs Launch Success



The heaviest payloads (8000 Kg – 16000 Kg) have had high success rates with fewer attempts

Dashboard: Correlation Payload vs Launch Success

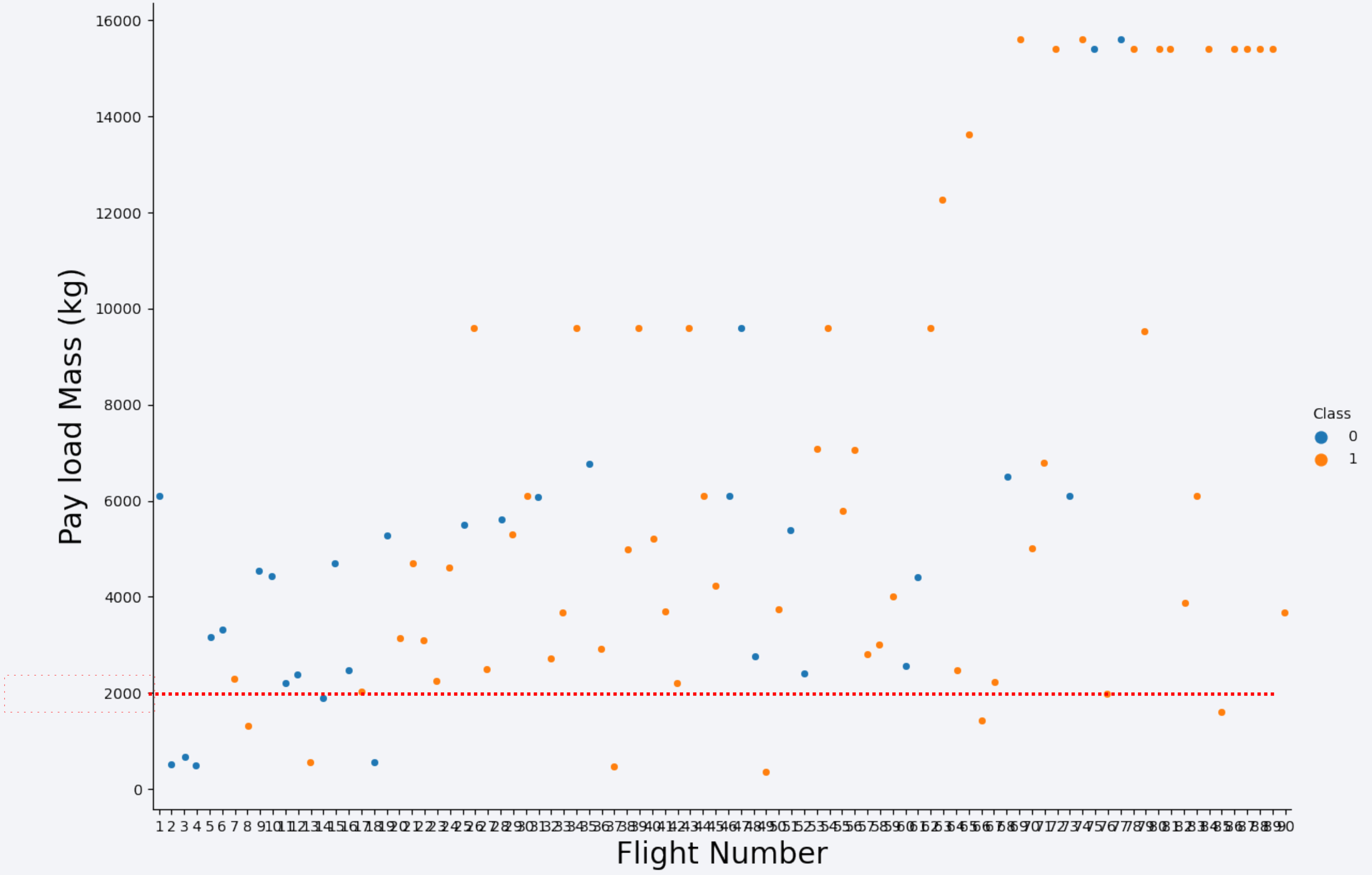


Site KSC LC-3A has the highest success rate

Dashboard: Correlation Payload vs Launch Success

Payloads below 2000 Kg have a very low success rate

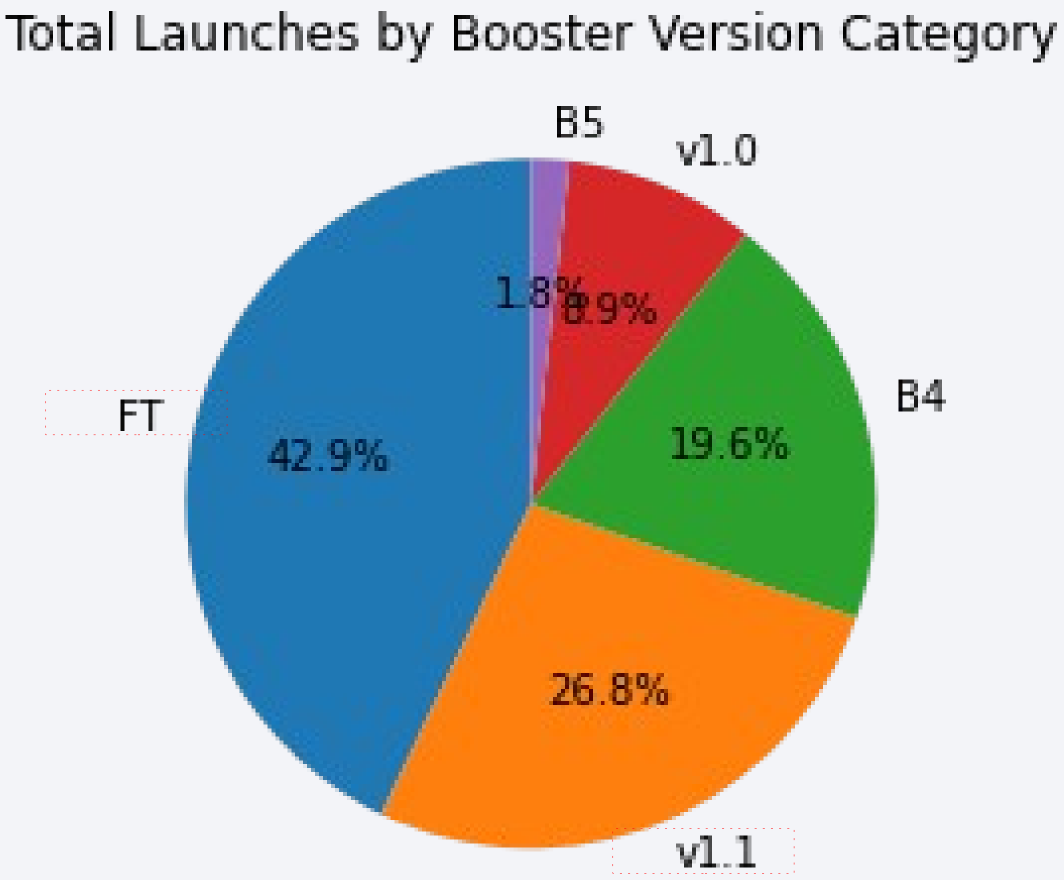
Dashboard: Correlation Payload vs Launch Success



One probable reason is that there were few flights with such low payloads

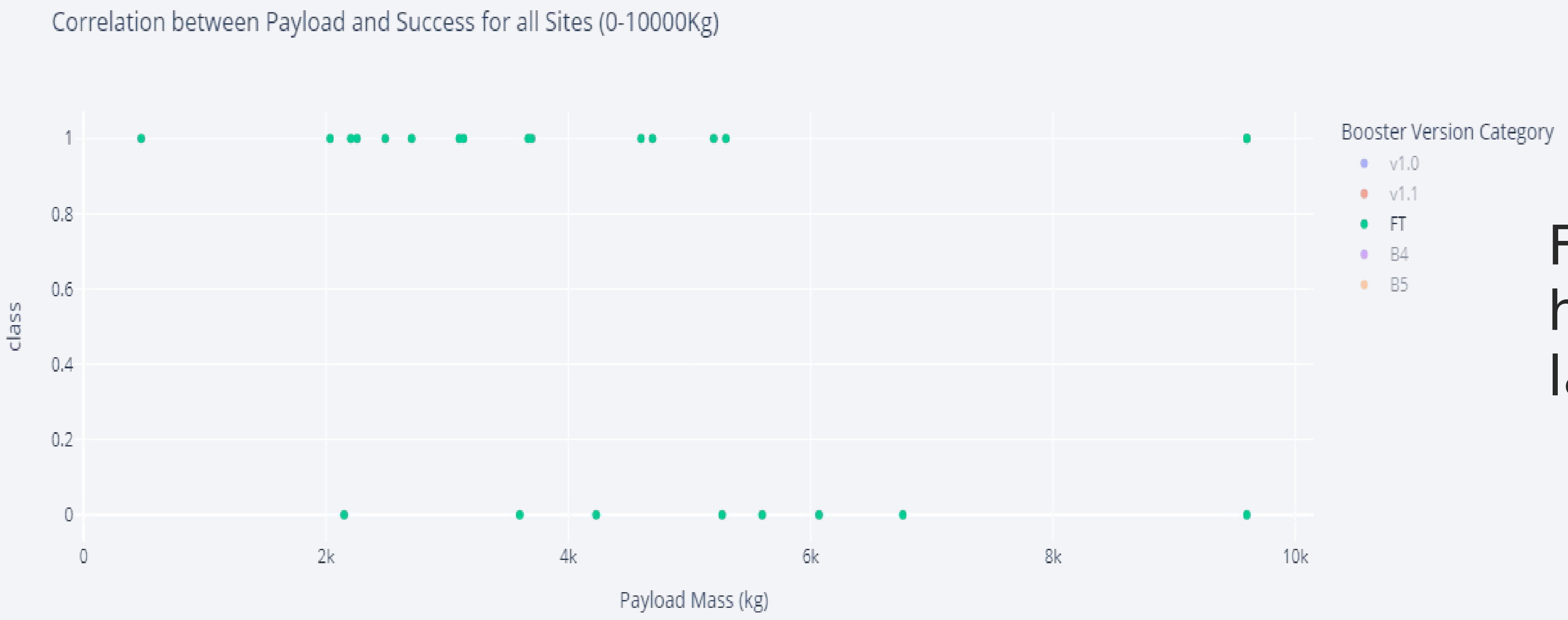
(lack of interest?)

Dashboard: Launch Success by Booster Version



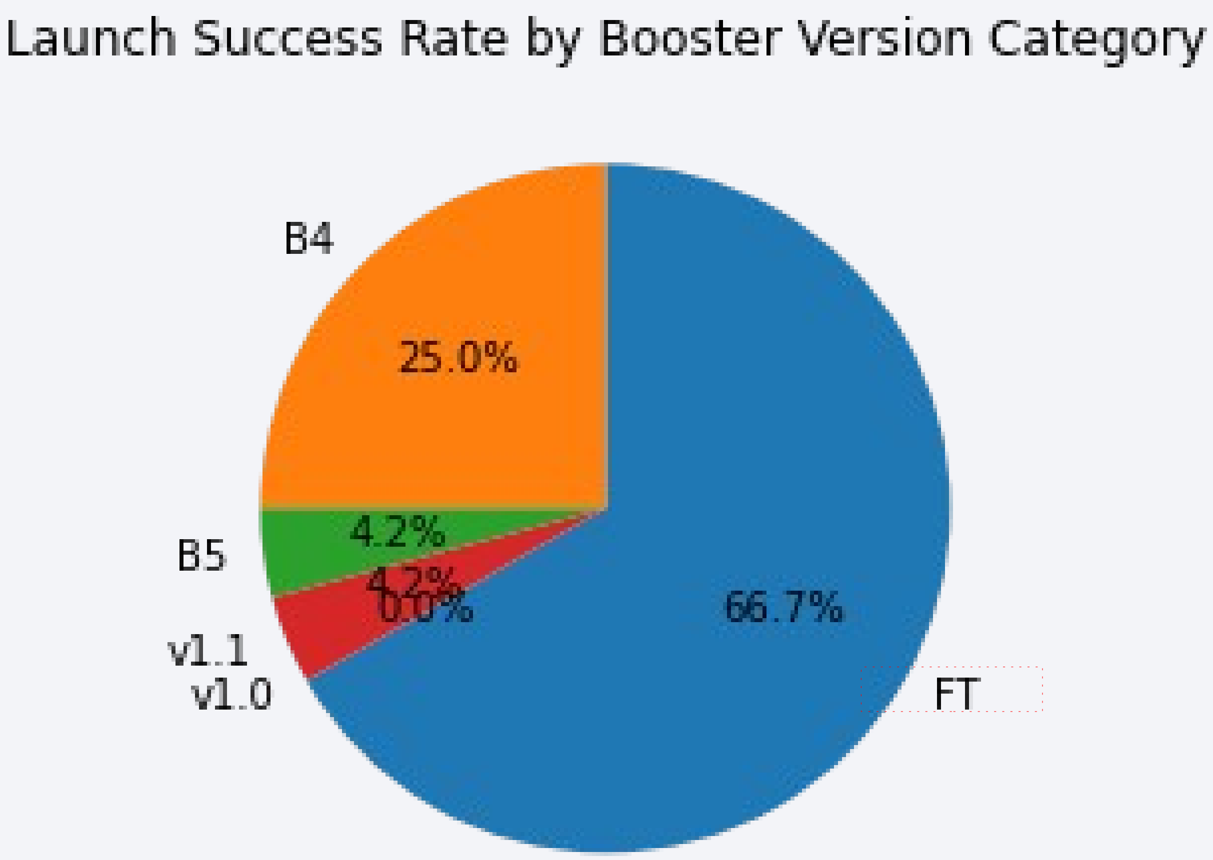
FT and V1.1 were used in 69.7% of launches

Dashboard: Launch Success by Booster Version



FT booster version has the highest launch success rate

Dashboard: Launch Success by Booster Version



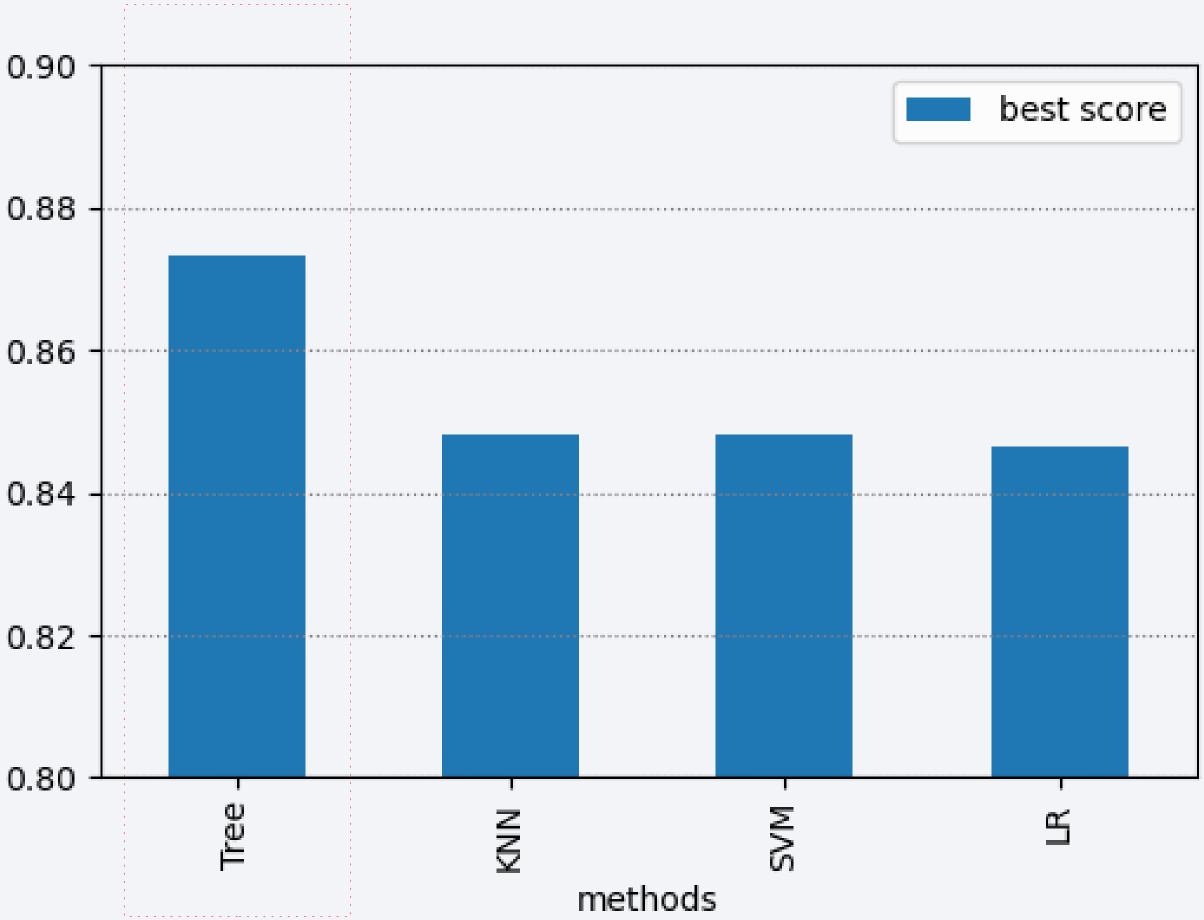
FT booster version has the highest launch success rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Decision Tree is the prediction model with the highest classification accuracy



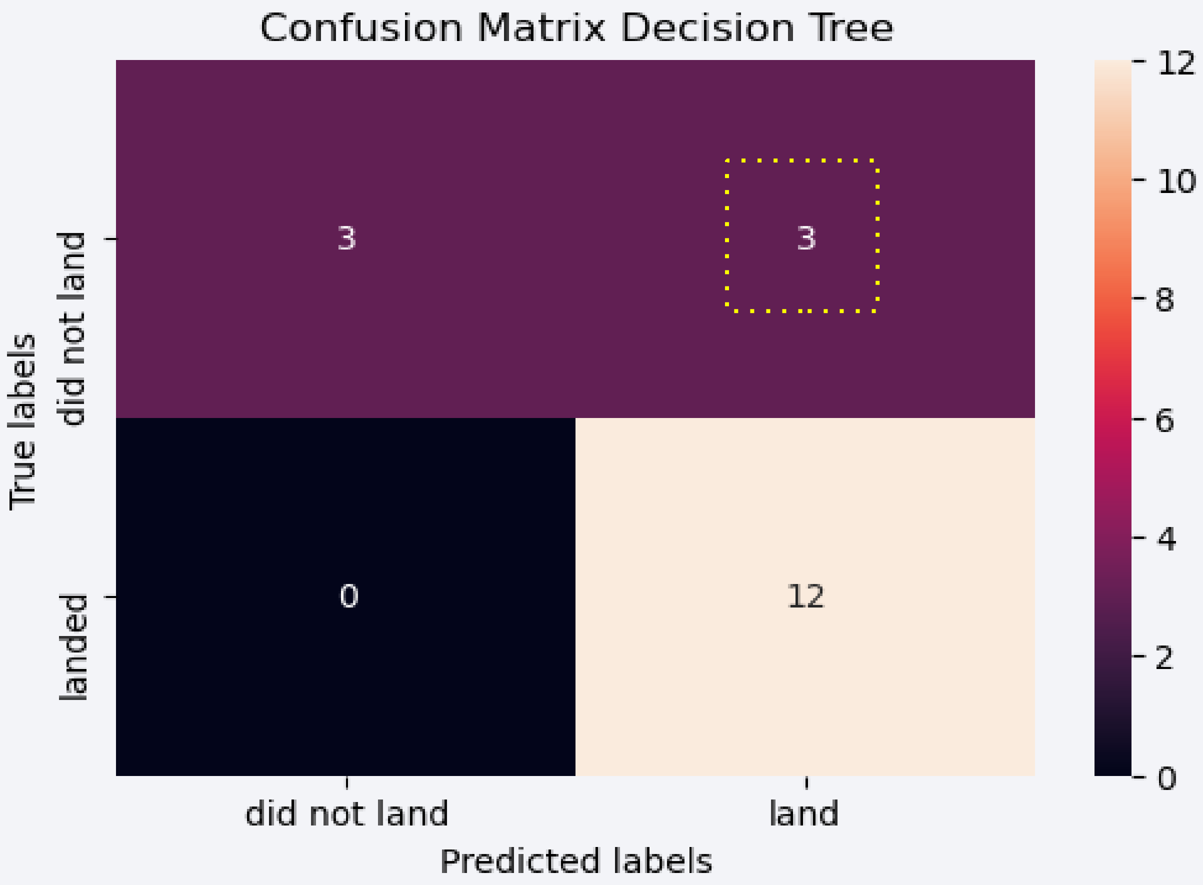
Classification Accuracy

Decision Tree is the prediction model with the highest classification accuracy

	methods	accuracy	best score
2	Tree	0.902778	0.873214
3	KNN	0.861111	0.848214
1	SVM	0.888889	0.848214
0	LR	0.833333	0.846429

Confusion Matrix

- Its performance, viewed through a confusion matrix, is quite similar to the other models
- The 3 in 15 occurrence of false positives has to be considered



Conclusions

The Launch Sites

- KSC LC-39A, CCAFS LC-40, and CCAFS SLC-40 practically share a very well-established infrastructure.
- 85.5% of the launches have taken place there, with a remarkable success rate in tests involving midweight and heavy payload masses.

Conclusions

The Orbits

- VLEO, ISS, and GTO, are the most visited orbits overall.
- They appear to have a shorter gap between subsequential launches within a concise payload range if compared to the other orbits.
- This induces us to see low-orbit and geostationary-orbit launches taken as market niches and cost prediction being more effective over time.

Conclusions

The Payload

- Most flights carried between 2000 kg and 4000 kg.
- Most successful launches carried a payload weighing between 2000 kg and 6000 kg.
- Some distinct heavy payload ranges (above 8000 kg to 16000 kg) have had high success rates with fewer flight attempts.

Conclusions

The Booster

- FT, the most recent booster version, has the far highest success rate and already responds to 42,9% of launches.
- It’s also very well succeeded in loading the most common payload weights when compared to previous versions.

Conclusions

Therefore, in every launch site, successful landing rates have reached a remarkable level of reliability under certain pre-established circumstances (in terms of the type of orbit and payload ranges) and are predictably prone to increase as the flight number increases.

Appendix

Wikipedia pages on landing sites, orbits and Falcon 9 stage boosters:

- https://en.wikipedia.org/wiki/Spaceport
- https://en.wikipedia.org/wiki/List_of_orbits
- https://en.wikipedia.org/wiki/List_of_Falcon_9_first-stage_boosters

Article on the reason behind launch sites being built near Equator and coastlines:

- https://www.scienceabc.com/eyeopeners/why-are-rockets-launched-from-areas-near-the-equator.html

Custom notebook with the Plotly Dash dashboard App using the JupyterDash library:

- https://github.com/RodrigoElesbao/Coursera_IBM_Data_Science_Professional_Certificate/tree/main/10._applied_data_science_capstone/5._present_your_data-driven_insights

Thank you!