IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tanmay Deokule
26th June 2025

Outline

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

Executive Summary

Summary of Methodologies Used:

1. Data Collection
2. Data Wrangling
3. Exploratory Data Analysis with Data Visualization
4. Exploratory Data Analysis with SQL
5. Building an Interactive Map with Folium
6. Building a Dashboard with Plotly Dash
7. Predictive Analytics (Classification Task)

Summary of Results

1. Exploratory Data Analysis results
2. Interactive analytics demo in screenshots
3. Predictive analysis results

Introduction

Objective: To predict whether first stage of Falcon-9 rocket launches will land successfully

Project background and Context

- SpaceX, a successful company of the commercial space, is making space travel affordable.
- The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars.
- Other providers cost upward of 165 million dollars each.
- Much of the savings is because SpaceX can reuse the first stage.
- If one can determine if the first stage will land, one will be able to determine the cost of a launch.
- Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

Questions to be Answered

- How do variables such as payload mass, launch site, number of flights, and target orbit affect the success of the first stage landing ?
- Does the rate of successful landings increase over the years ?
- What is the best algorithm that can be used for binary classification in this case ?

Section 1

Methodology

Methodology
Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia (Library: Beautiful Soup)
- Perform data wrangling
  - Filtering Data to keep only the Relevant one.
  - Dealing with Missing values by deleting corresponding rows or replacing missing values with mean or median values of a particular feature.
  - Using One Hot Encoding to prepare the data to a binary classification
- 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 ensure the best results (Library: SciKit-Learn)

Data Collection - Overview

- Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.
- Combine into data frame to get complete Dataset using Pandas.

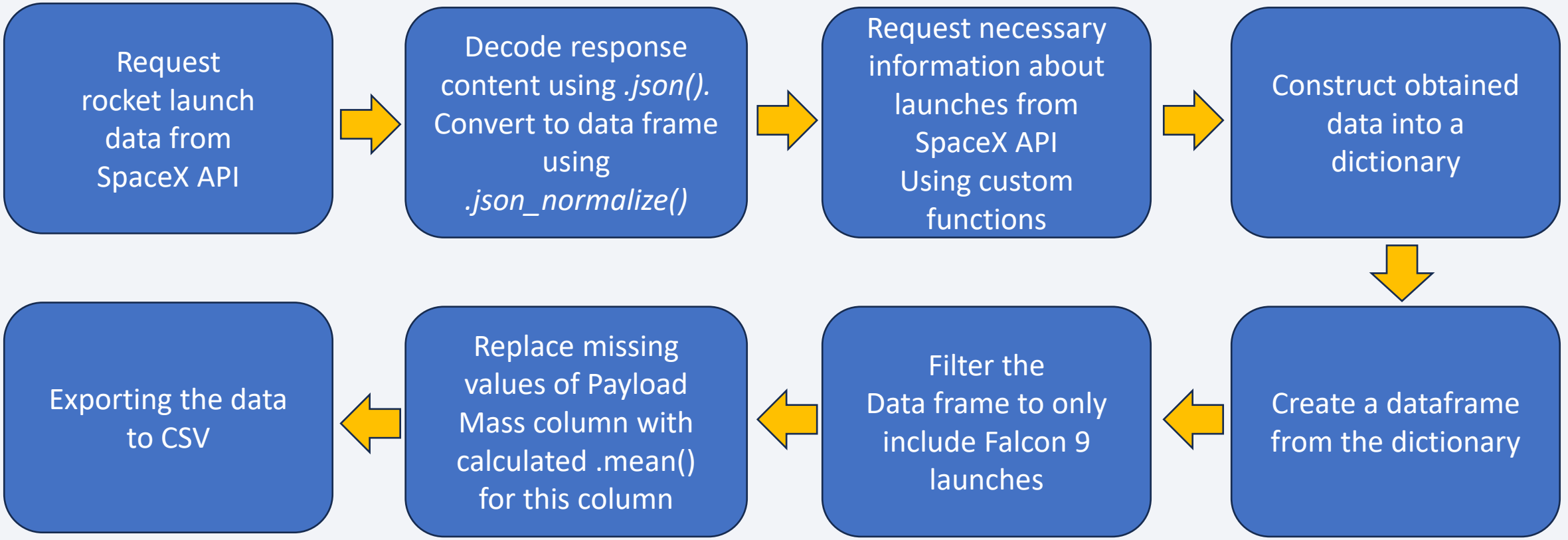
Data Columns are obtained by using SpaceX REST API:

FlightNumber, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, GridFins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude

Data Columns are obtained by using Wikipedia Web Scraping:

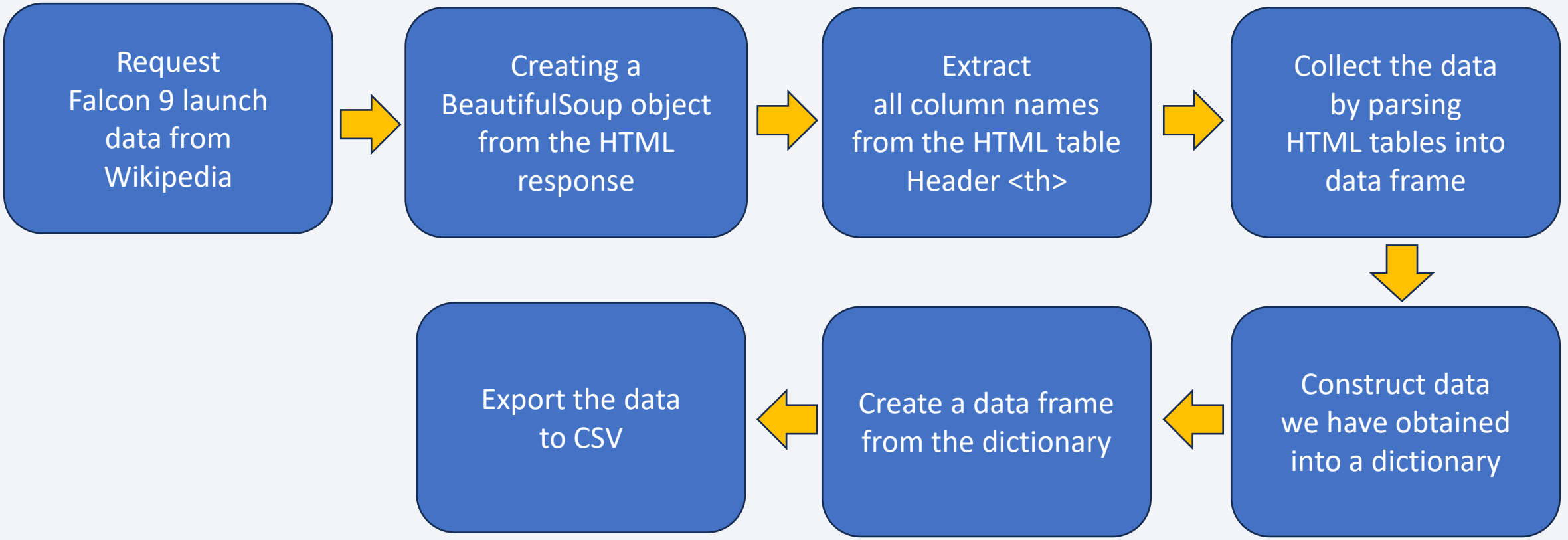
Flight No., Launch site, Payload, Payload Mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



GitHub URL : Data Collection – SpaceX API

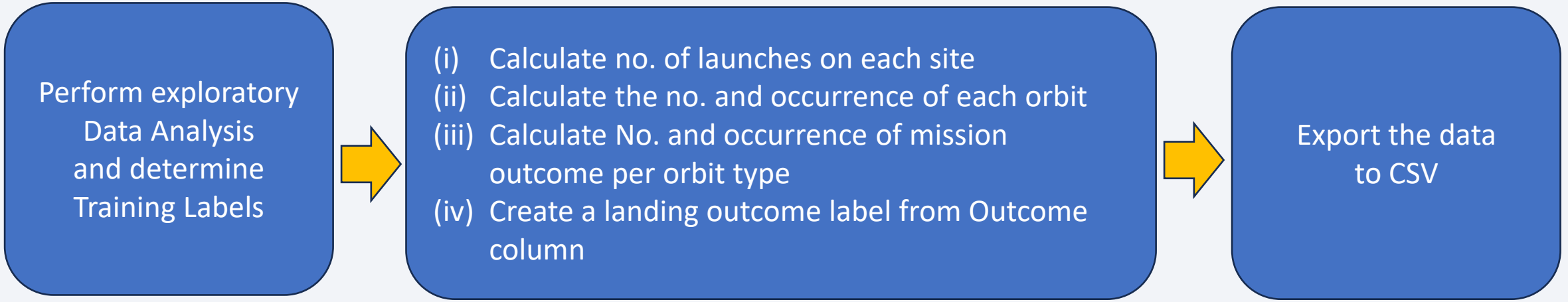
Data Collection - Scraping



GitHub URL : Data Collection - Scraping

Data Wrangling

- Clean dataset using Pandas and Numpy
- Convert complicated / categorical outcomes into Training Labels with “1” meaning booster’s successful landing and “0” meaning landing was unsuccessful.



GitHub URL : Data Wrangling

EDA with Data Visualization

Plotted Charts:

i. Flight Number vs. Payload Mass,
ii. Flight Number vs. Launch Site,
iii. Payload Mass vs. Launch Site,
iv. Orbit Type vs. Success Rate,
v. Flight Number vs. Orbit Type,
vi. Payload Mass vs Orbit Type
vii. Success Rate Yearly Trend

GitHub URL : EDA with Data Visualization

EDA with SQL

SQL Queries Performed:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved.
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing all the booster versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.
- Listing the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

GitHub URL : EDA with SQL

Build an Interactive Map with Folium

Markers of all Launch Sites:

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Colored Markers of the launch outcomes for each Launch Site:

- Added colored Markers of success (Green) and failed (Red) launches using Marker Cluster to
- identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

- Added colored Lines to show distances between the Launch Site KSC LC-39A (for example) and its proximities like Railway, Highway, Coastline, Closest City etc.

GitHub URL : Building Interactive Map with Folium

Build a Dashboard with Plotly Dash

User Input Objects

- **Drop Down List for Launch Site Selection:** Added a dropdown list to enable Launch Site selection. This dropdown lets you select either ALL launch sites or individual launch sites.
- **Slider of Payload Mass Range:** Added a slider to select Payload range

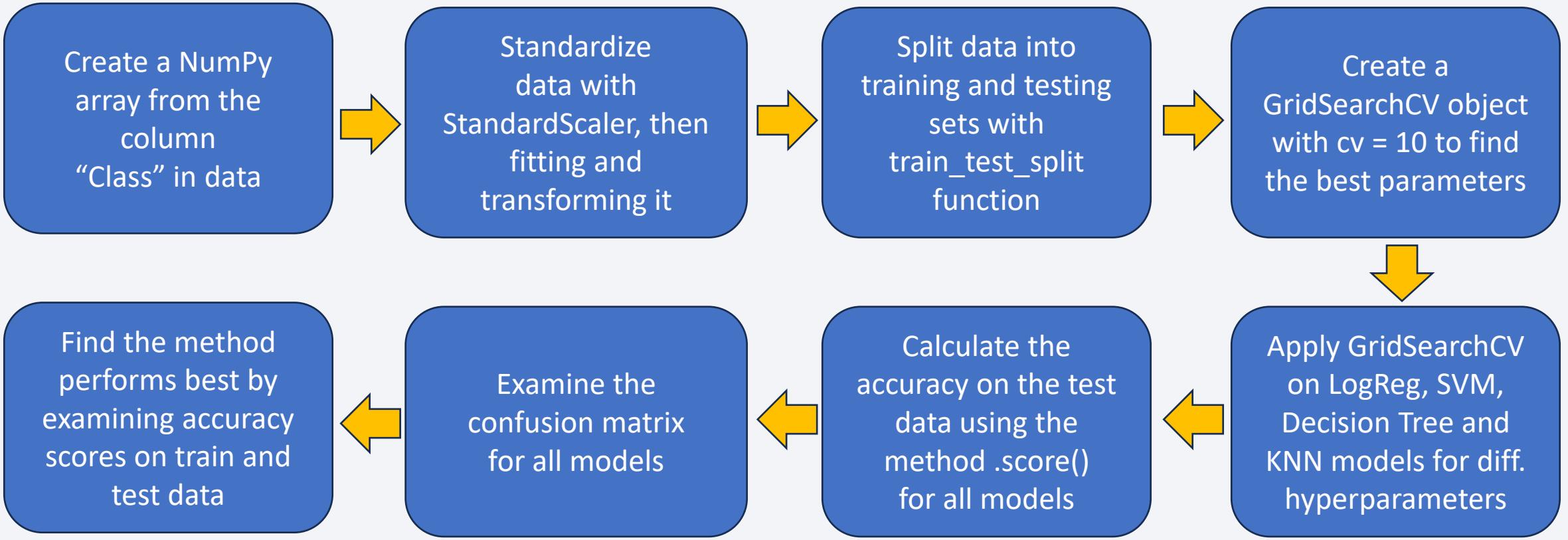
Result Display Plots

- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.
- Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions: Added a scatter chart to show the correlation between Payload and Launch Success based on payload mass range selected on slider.

GitHub URL : Python Code for Dashboarding with Plotly Dash

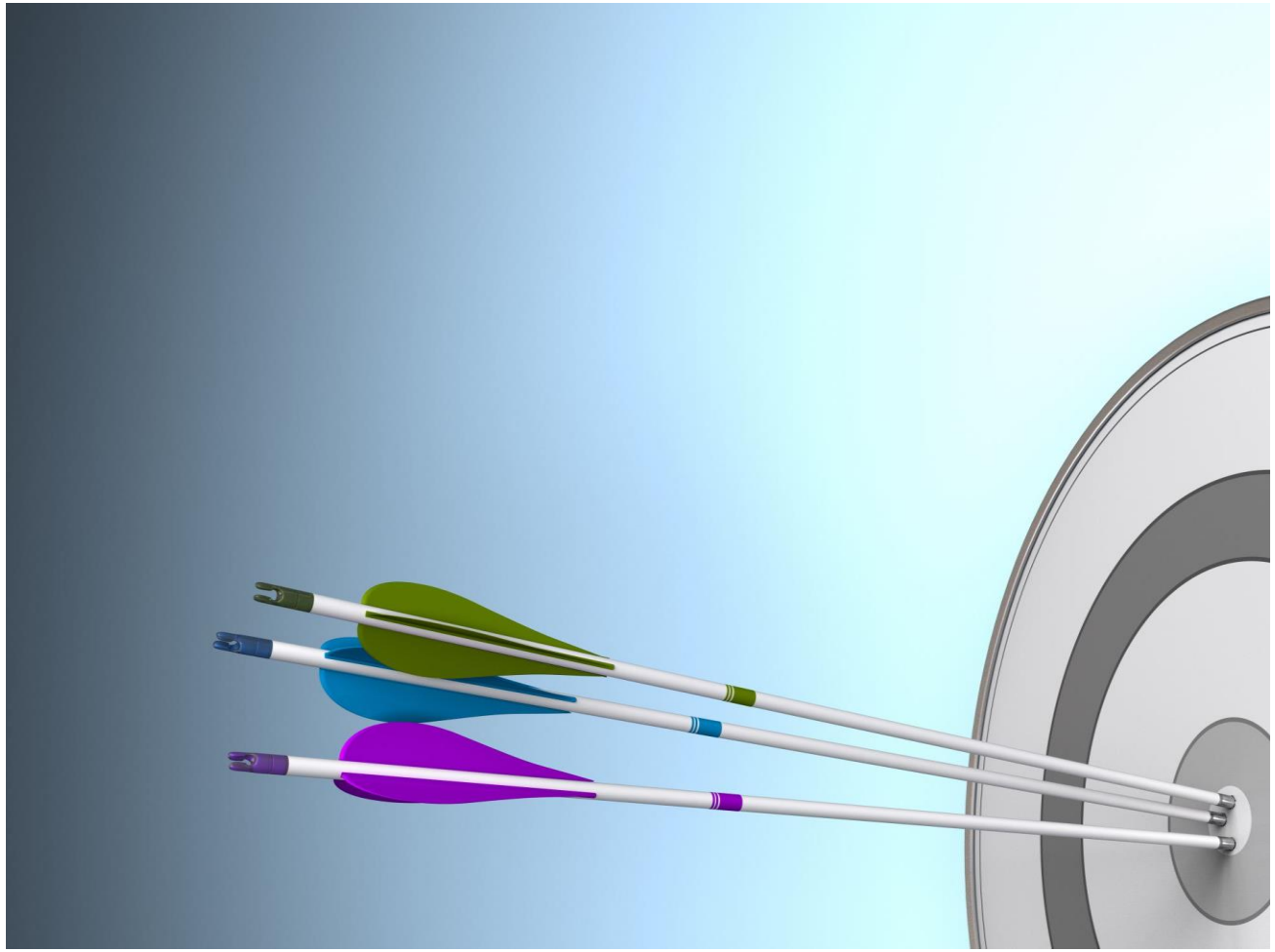
Predictive Analysis (Classification)

Flowchart on how ML Model was built, evaluated, improved and best performing classification model was identified.

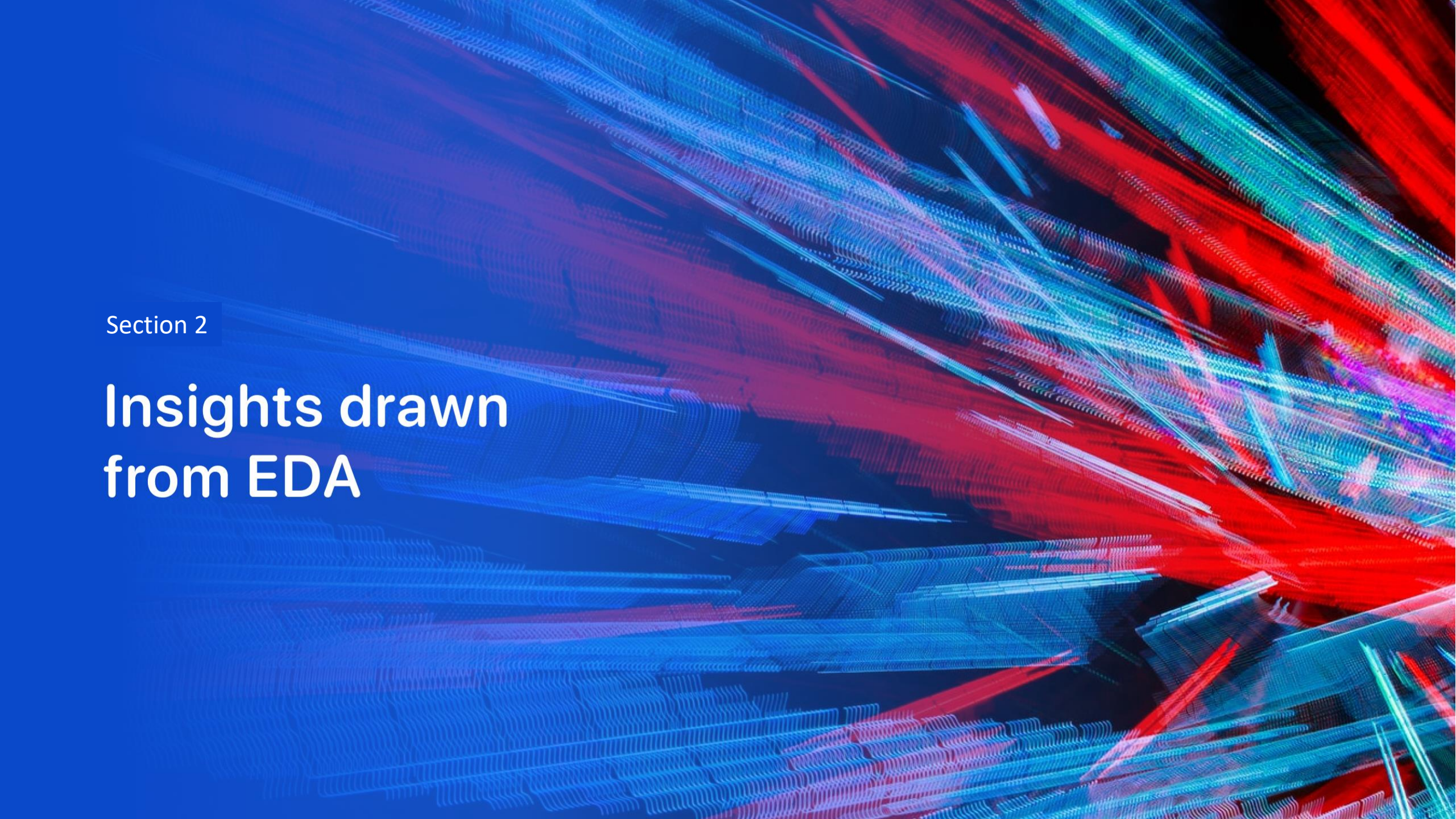


GitHub URL : Predictive Analytics - Classification

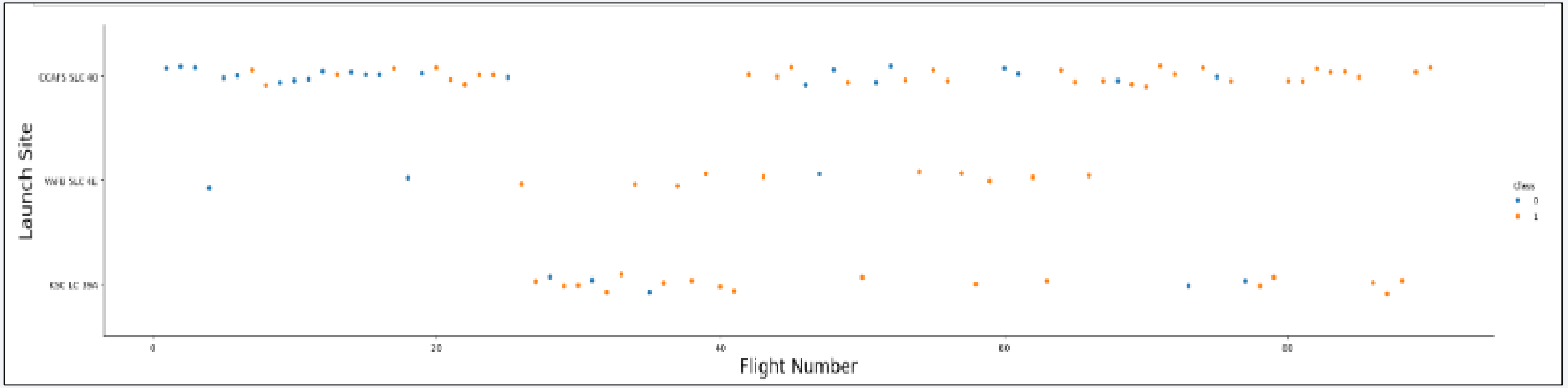
Results



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


Section 2
Insights drawn from EDA

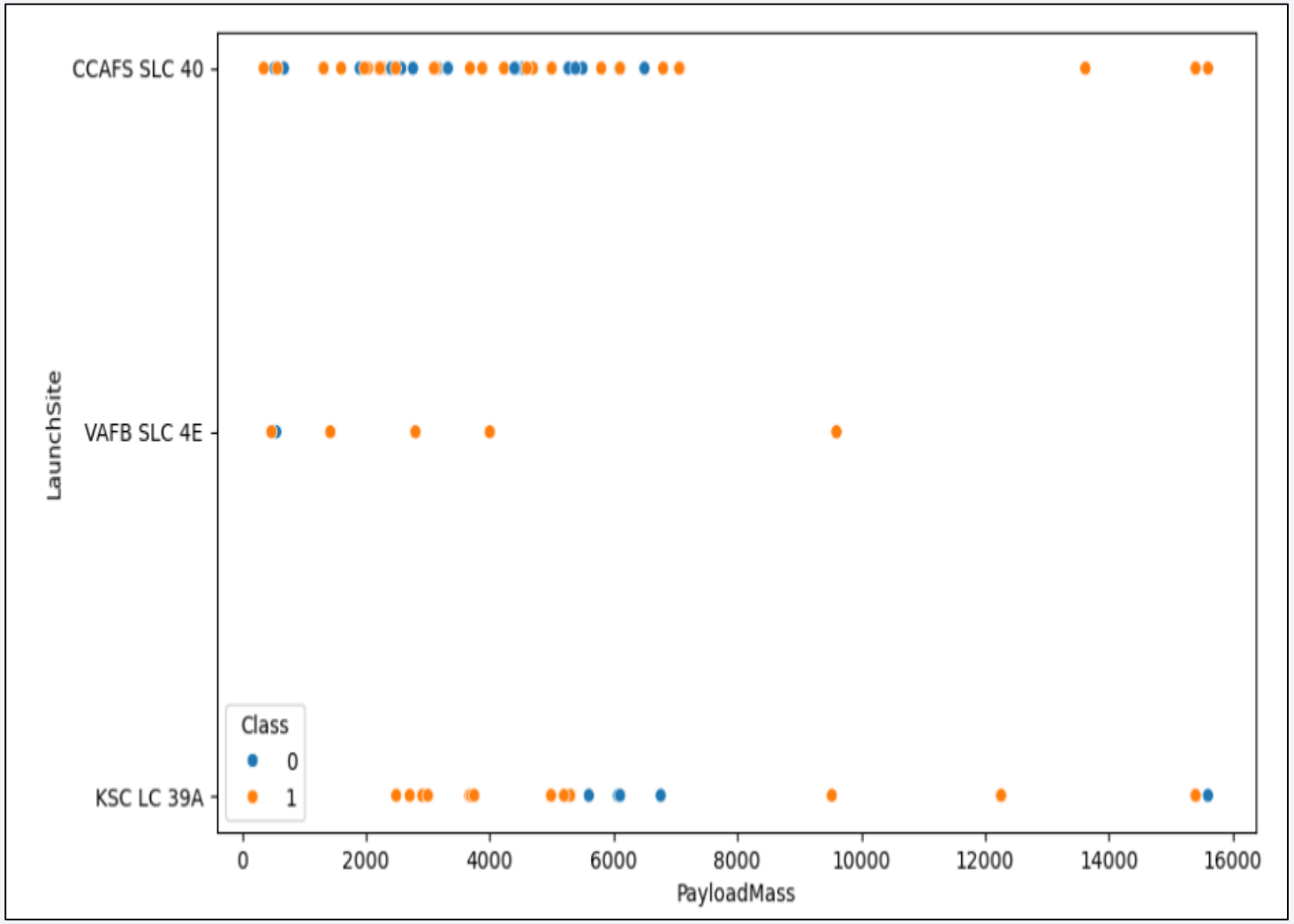
Flight Number vs. Launch Site



Explanation:

- Earliest flights had higher proportion of failed launches while the latest flights have all succeeded.
- The CCAFS SLC 40 launch site has the maximum number of launches.
- VAFB SLC 4E and KSC LC 39A have highest success rates among all launch sites.
- It can be assumed that each new launch has a higher rate of success.

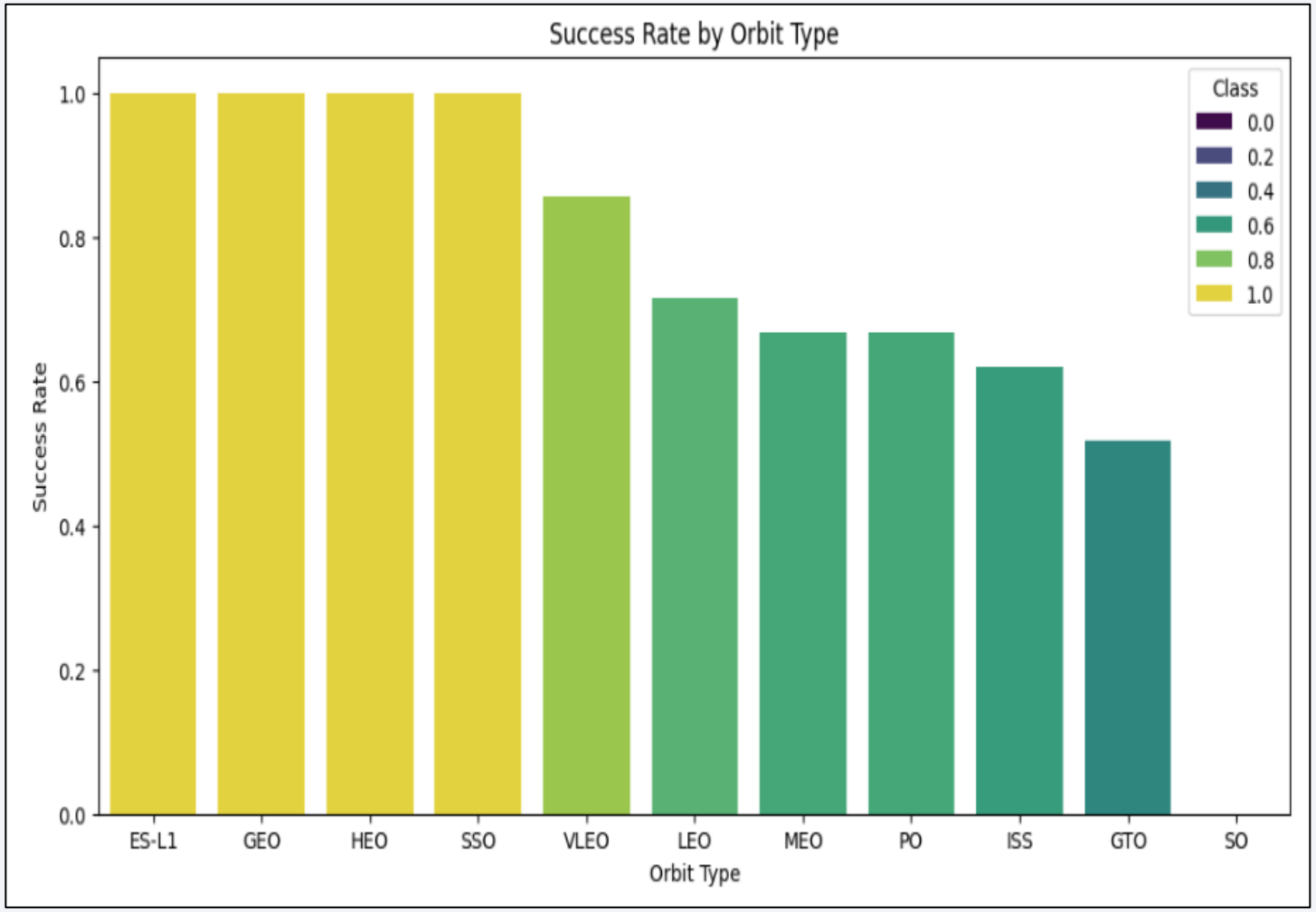
Payload vs. Launch Site



Explanation:

- Higher the payload mass, higher is the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg.
- VAFB SLC 4E has 100% success rate above 1000 kg. payload
- CCAFS SLC 40 is the preferred launch site under 7000 kg. payload.

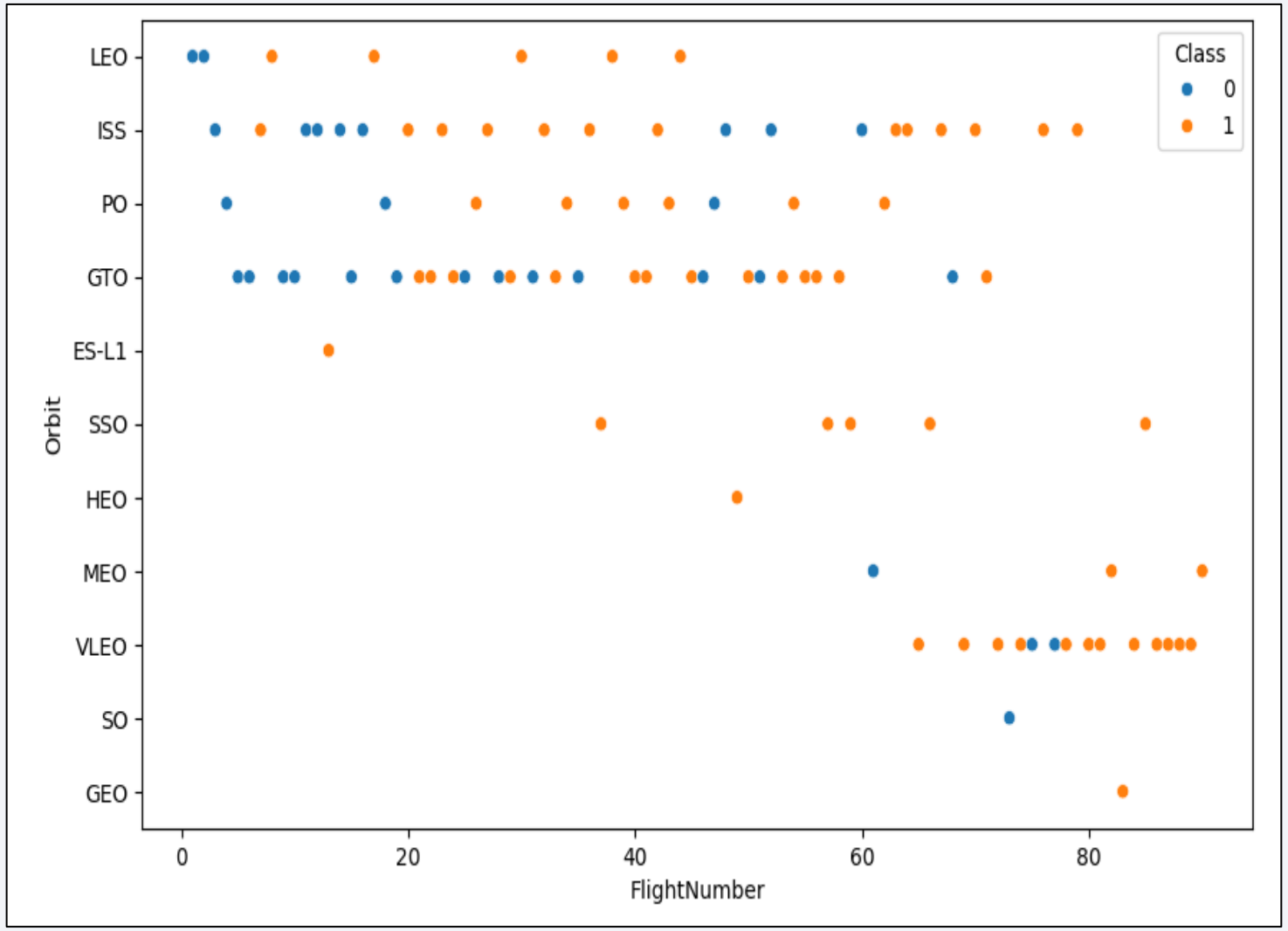
Success Rate vs. Orbit Type



Explanation:

- Orbits with 100% success rate are ES-L1, GEO, HEO, SSO.
- Orbits with 0% success rate are SO.
- Orbits with success rate between 50% and 85% are GTO, ISS, LEO, MEO, PO
- There are no orbit types with success rate ranging between 1% and 50%.

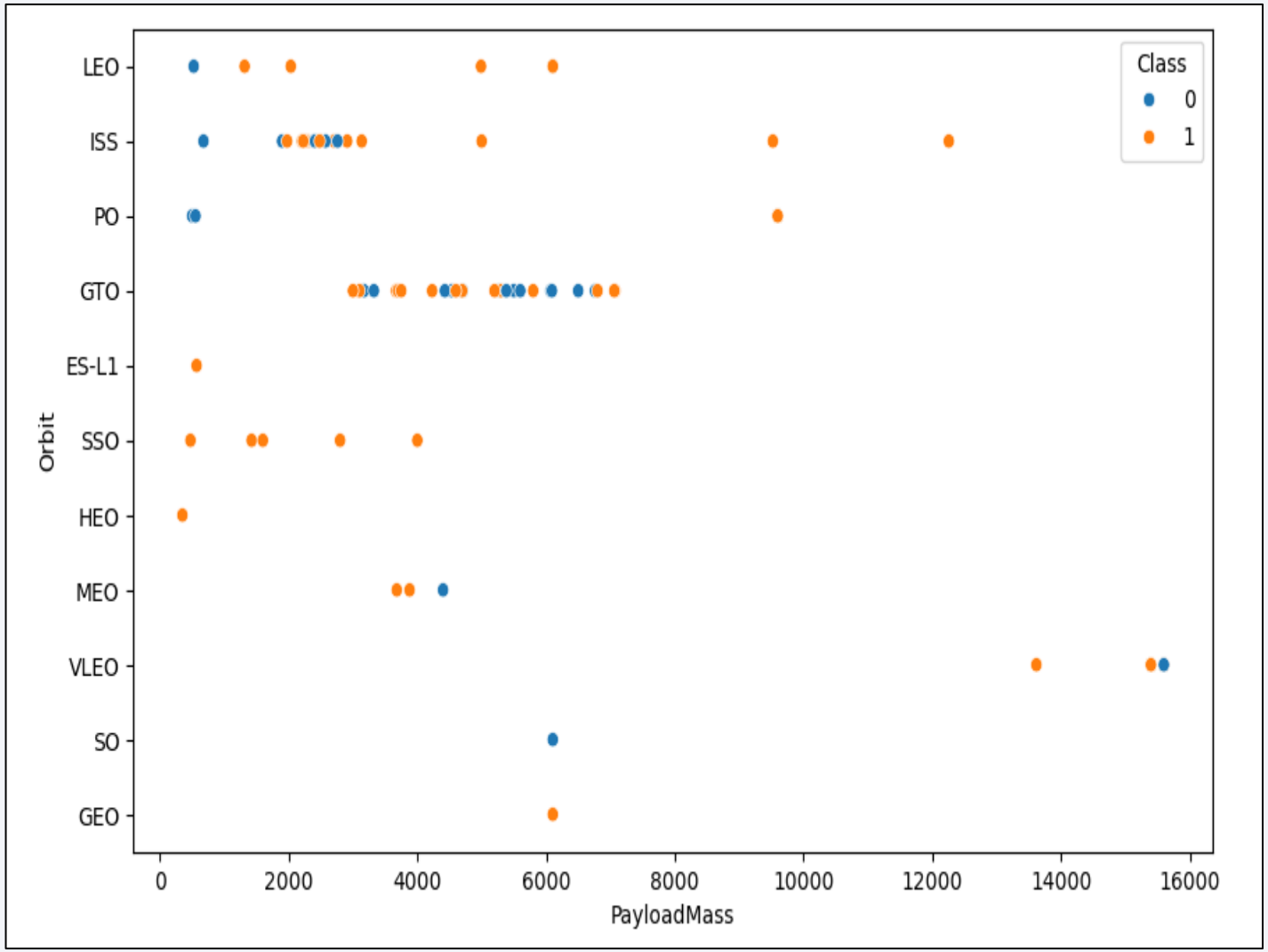
Flight Number vs. Orbit Type



Explanation:

- In the LEO orbit the Success appears related to the number of flights.
- There seems to be no relationship between flight number when in GTO orbit.
- Majority of later launches have been to VLEO orbit.
- No launch to SO Orbit has been successful irrespective of Flight Number
- Maximum Flights have been to ISS and GTO orbits

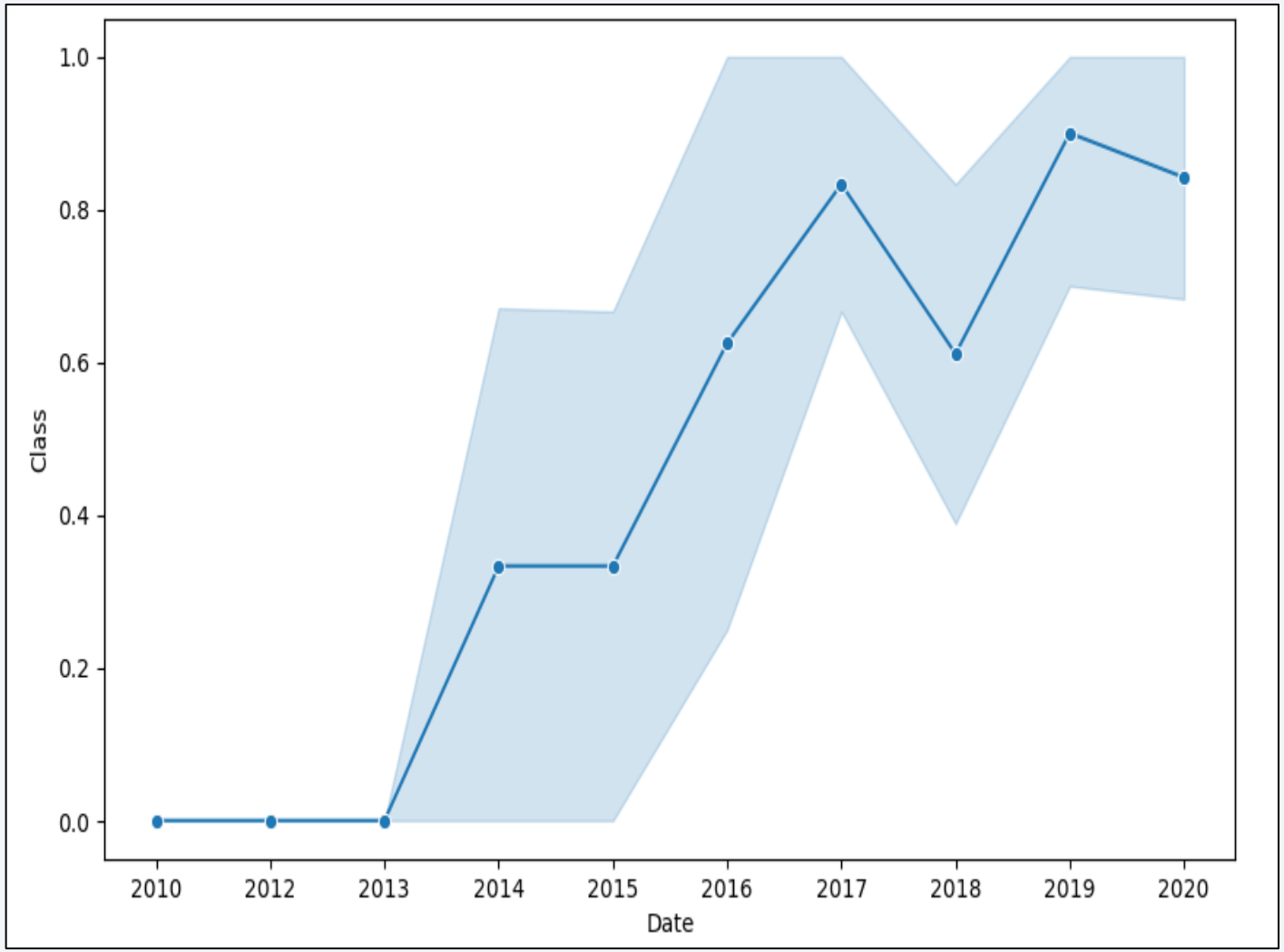
Payload vs. Orbit Type



Explanation:

- Payloads have no influence on success and failure of missions to GTO orbits and positive on GTO and Polar LEO & ISS orbits.
- Highest payloads have been to VLEO orbit
- Success Rate for SSO orbit is highest at 100% irrespective of payload mass.

Launch Success Yearly Trend



Explanation:

- There is an overall trend of improving launch success rates since 2013 to 2020.
- Success Rate saw a dip in 2018 and 2020.

All Launch Site Names

Task 1

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

```
In [13]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
```

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

Out[13]:

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

Explanation:

- To identify unique launch site names in the space mission, we make use of DISTINCT keyword to display the result, as shown.

Launch Site Names Begin with 'CCA'

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

```
In [15]: %sql SELECT * FROM SPACEXTABLE where Launch_Site like '%CCA%'
 * sqlite:///my_data1.db
Done.
```

Out[15]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcor
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachu
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 (parachu
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No atten
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No atten
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No atten

Explanation:

- The query lists all launch sites which begin with string '%CCA%'. Here we make use of Wild Card '%' to account for different string literals after CCA.
- If exactly 5 records are to be listed, we can do so by including "limit 5" in query. However, the query does not specify exactly 5. hence, all such launch sites are listed out of which any 5 can be chosen.

Total Payload Mass

Task 3

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

In [23]:
```
%sql  select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer like "%NASA (CRS)%"
```

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

Out[23]:

sum(PAYLOAD_MASS__KG_)
48213

Explanation:

- Displaying the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

Task 4

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

In [20]:
```
%sql  select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like "%F9 v1.1%"
```

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

Out[20]:

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

Explanation:

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

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

```
In [21]: %sql select min(Date) from SPACEXTABLE where Landing_Outcome = "Success (ground pad)"
 * sqlite:///my_data1.db
Done.
```

Out[21]:

min(Date)
2015-12-22

Explanation:

- Above query provides the date for the first successful landing outcome on ground pad.
- We check for earliest date using min function on date column provided success condition is met.

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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

In [26]:
```
%sql select DISTINCT Booster_Version from SPACEXTABLE where (Landing_Outcome = "Success (drone ship)") And (PAYLOAD_MASS__KG
```

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

Out[26]:

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

Explanation:

- Above query provides a list of the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000. We use between function to check for mass between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [27]:
```
%sql select Mission_Outcome, COUNT(Mission_Outcome) from SPACEXTABLE GROUP BY Mission_Outcome
```

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

Out[27]:

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

Explanation:

- Listing the total number of successful and failure mission outcomes.
- Result illustrates ‘Mission Outcome’ column requires some cleaning.

Boosters Carried Maximum Payload

Task 8

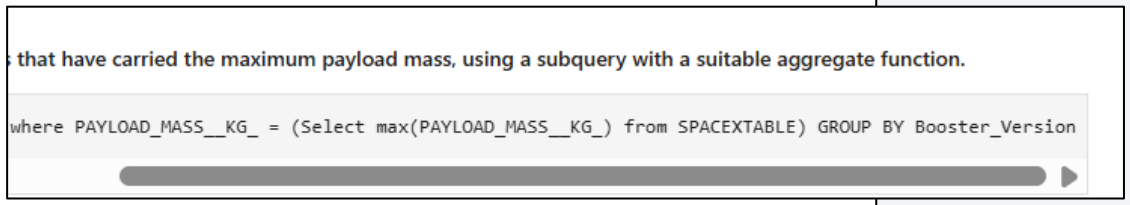
List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.

In [33]:
```
%sql SELECT Booster_version, MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE where PAYLOAD_MASS__KG_ = (Select max(PAYLOAD_MASS__KG_
```

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

Out[33]:

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



Explanation:

- Listing the names of the booster versions which have carried the maximum payload mass.
- Here, we are using a sub query in WHERE Clause of query to identify maximum payload mass.

2015 Launch Records

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

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

In [39]:
```
%sql SELECT substr(Date, 6,2) AS MONTH, substr(Date,0,5) as YEAR, Landing_Outcome, Booster_Version, Launch_Site from SPACEXT
```

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

Out[39]:

MONTH	YEAR	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

```
Booster_Version, Launch_Site from SPACEXTABLE where Landing_Outcome Like "%Failure%" AND substr(Date,0,5) = "2015"
```

```
//my_data1.db
```

Explanation:

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.
- Failure outcome is searched using search string '%Failure%' which uses wildcards.
- Date value is parsed to look for Month Number and Year using 'substr' function.

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

Task 10

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

```
In [47]: %sql SELECT DISTINCT Landing_Outcome, COUNT(Landing_Outcome) AS Count_Of_Outcome from SPACEXTABLE where Date between '2010-6
```

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

Out[47]:

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

Explanation:

- 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.
- We use DESC Clause to sort the counts in descending order.

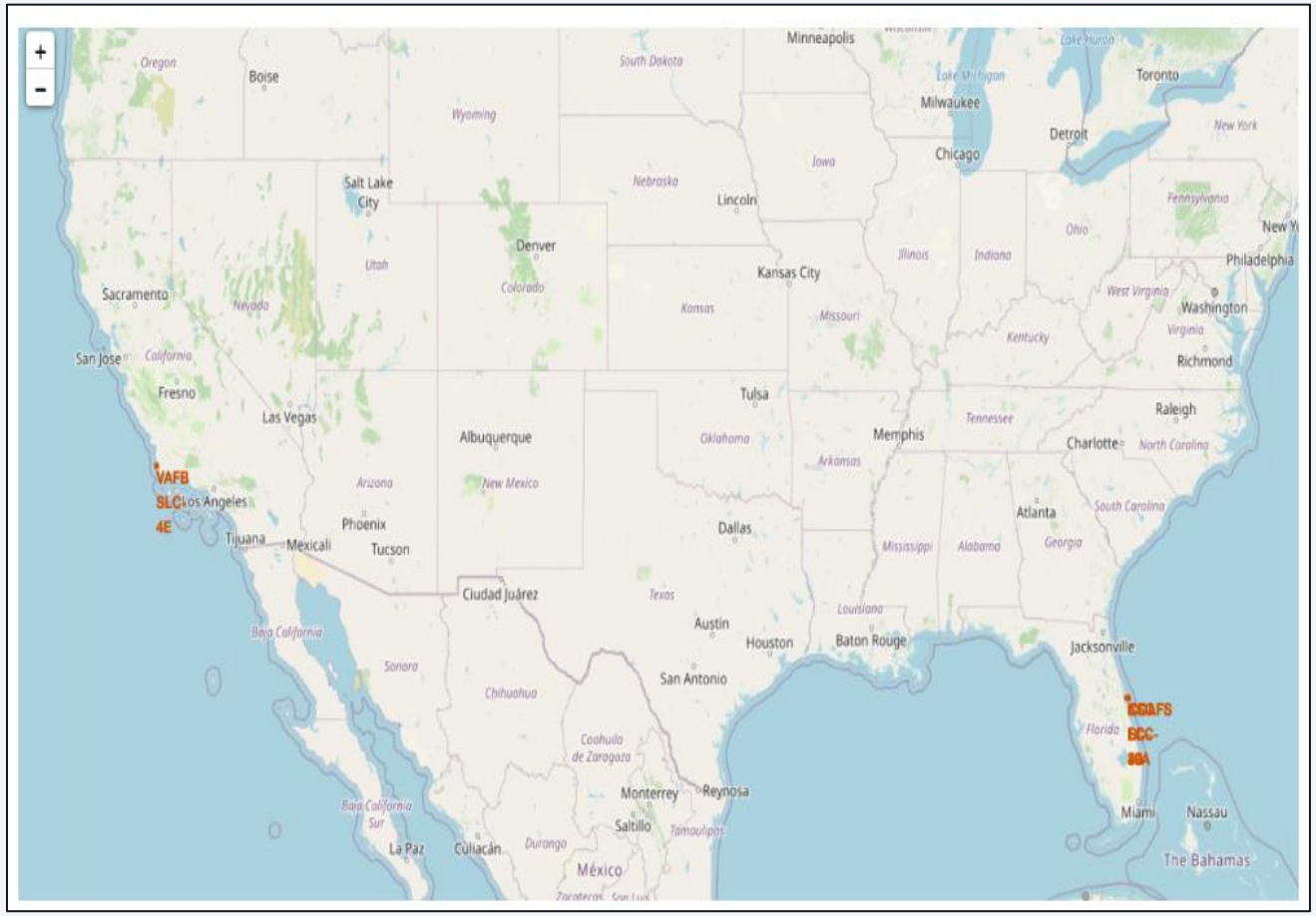
Section 3

Launch Sites Proximities Analysis

Launch Site Locations

Explanation:

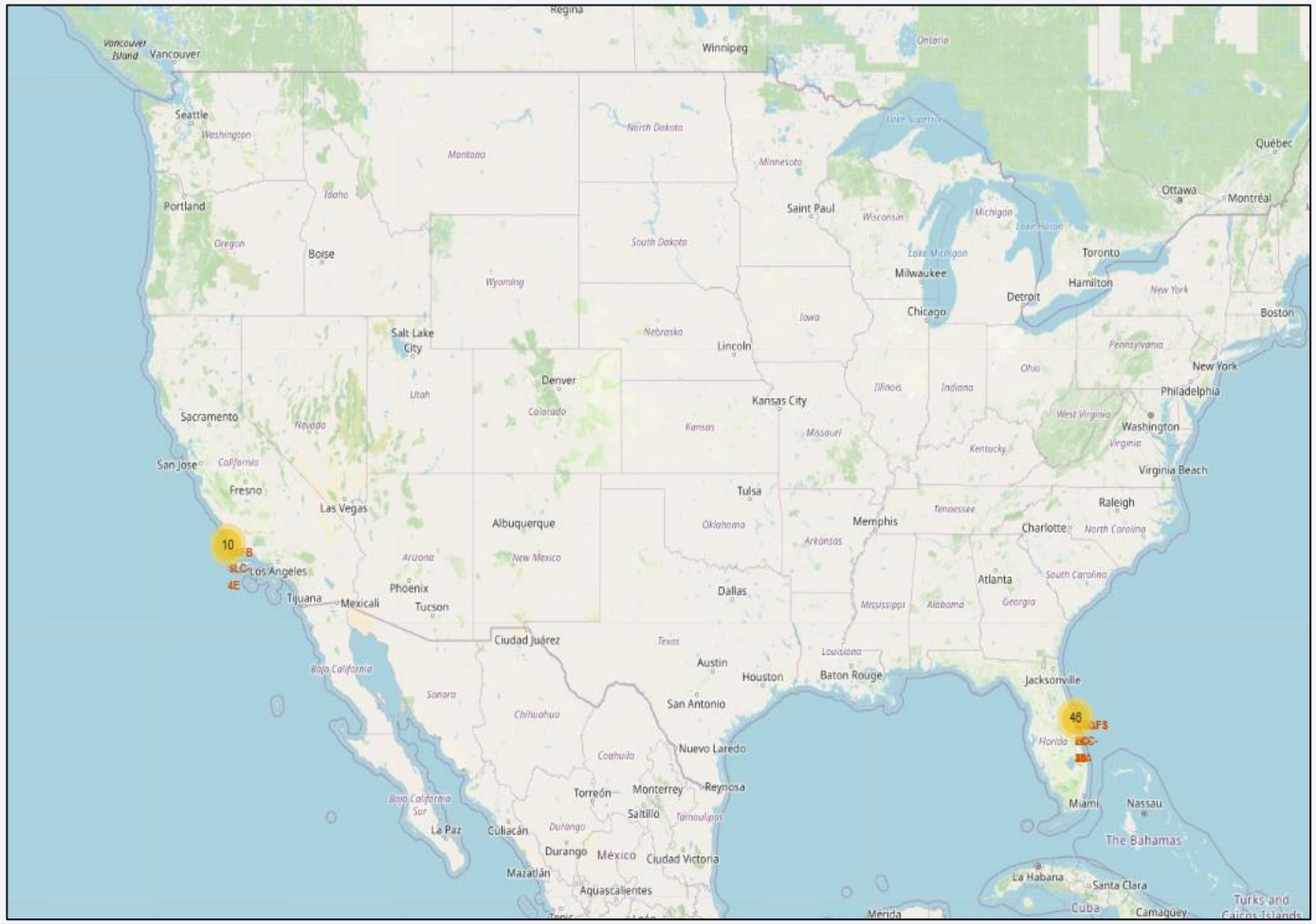
- Most of Launch sites are in proximity to Equator line.
- All launch sites are in very close proximity to the coast.
- There is one launch site on West Coast while remaining on east coast. Locations for the same are marked on the map.



Count of Launches from Launch Sites

Explanation:

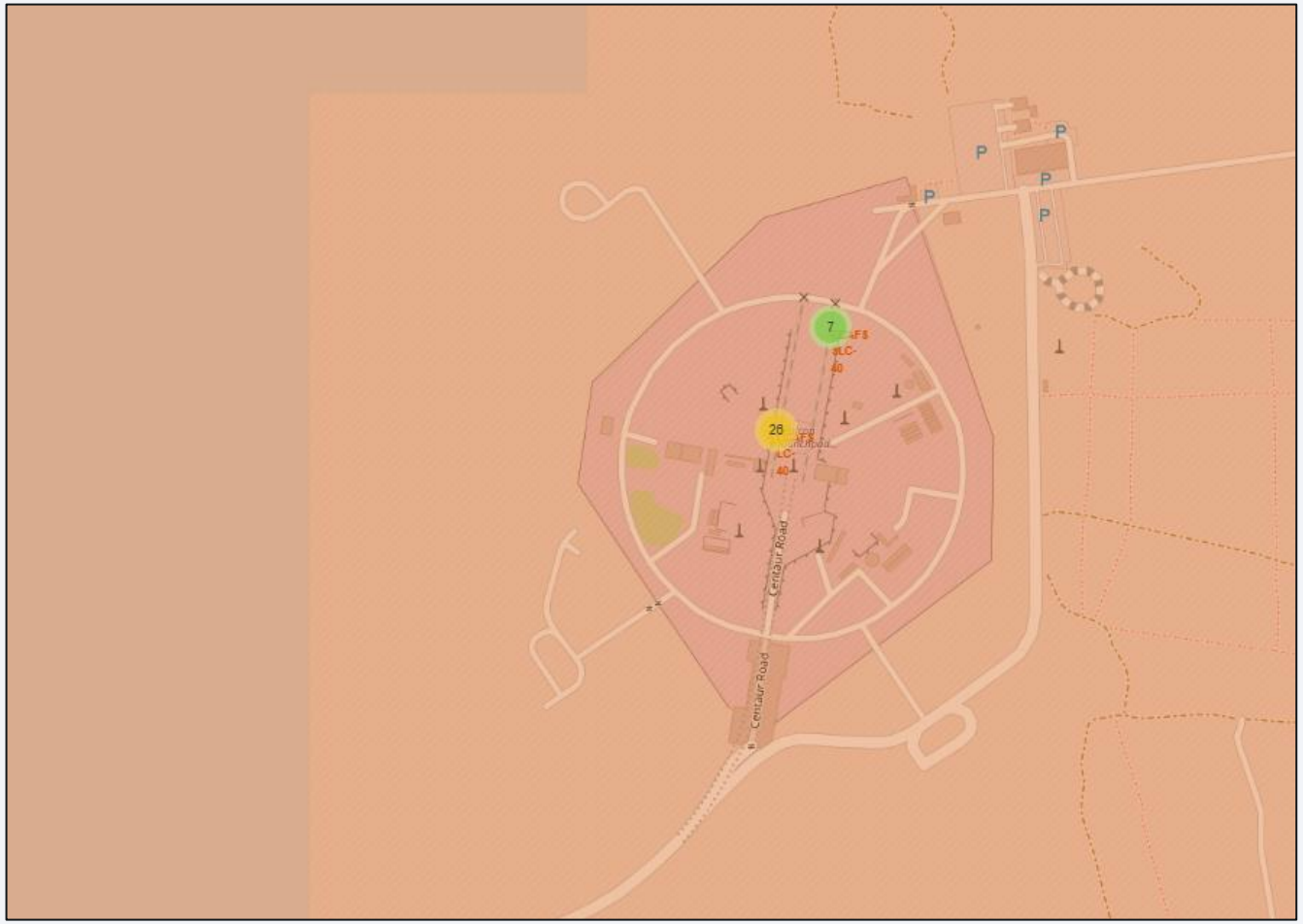
- Count of Launch sites from each Launch Site is plotted.
- In all, there were 10 launches from launch site on West Coast while remaining 46 were from launch sites on east coast. Locationsof launch sites and markers with count of launches are marked on the map.



Count of Successful and Failed Launches from Launch Sites

Explanation:

- Count of Launch sites from each Launch Site is plotted.
- Along with count, count of successful and failure launches is also plotted. They are shown with different colors on the map.

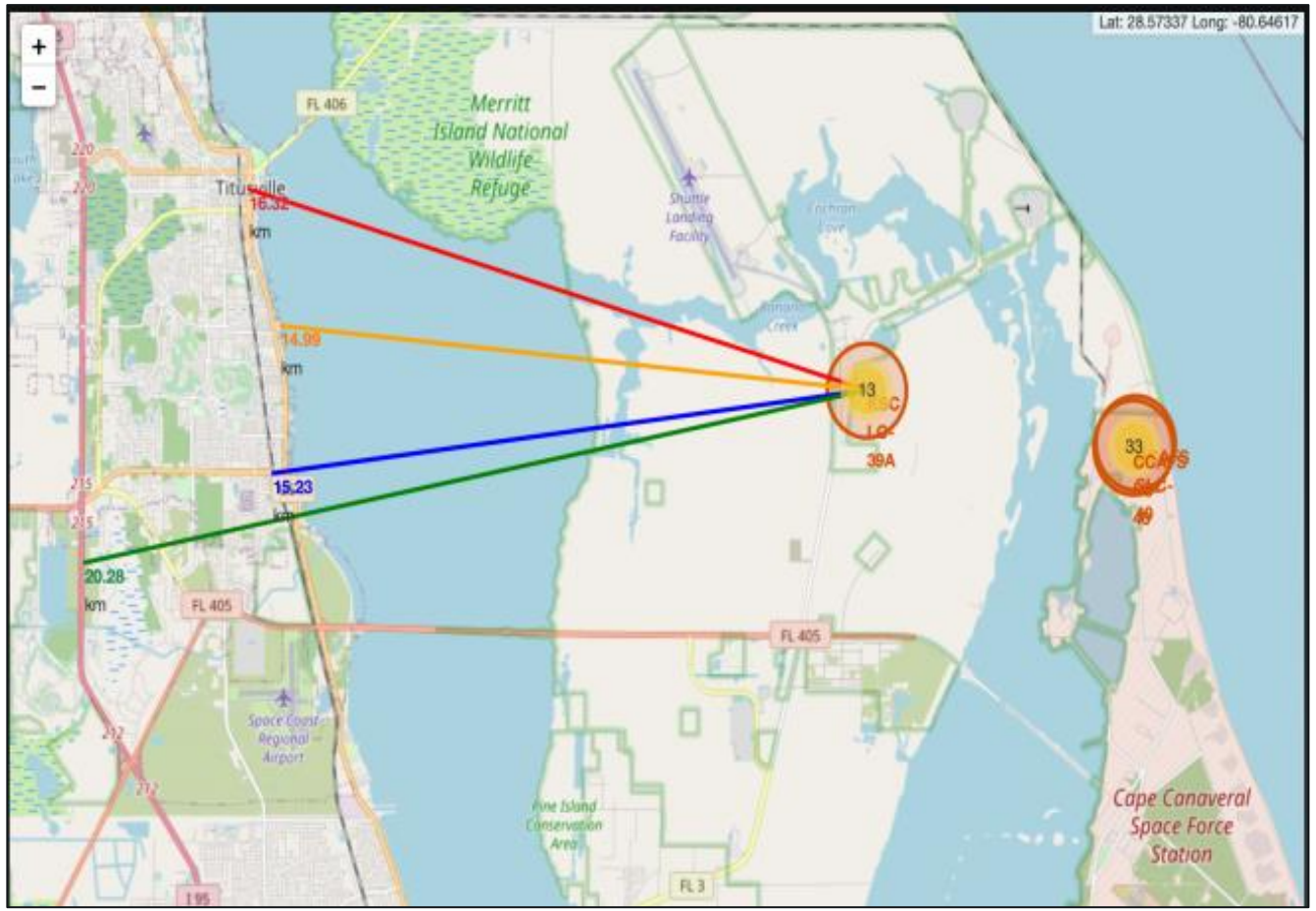


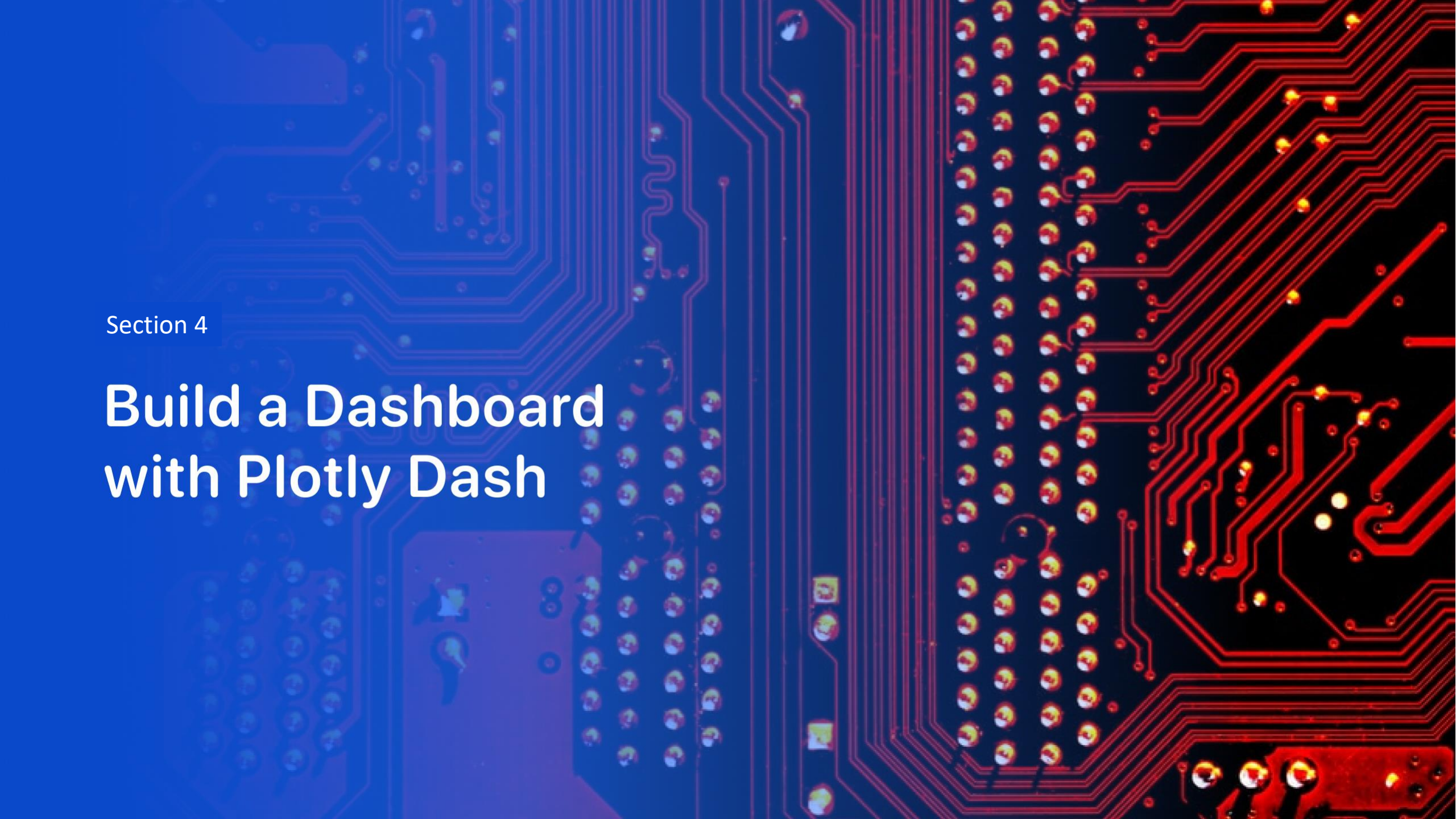
Visual Analysis of KSC LC-39A Launch Site

Explanation:
From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:

- relatively close to railway (15.23 km)
- relatively close to highway (20.28 km)
- relatively close to coastline (14.99 km)

Also, the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).

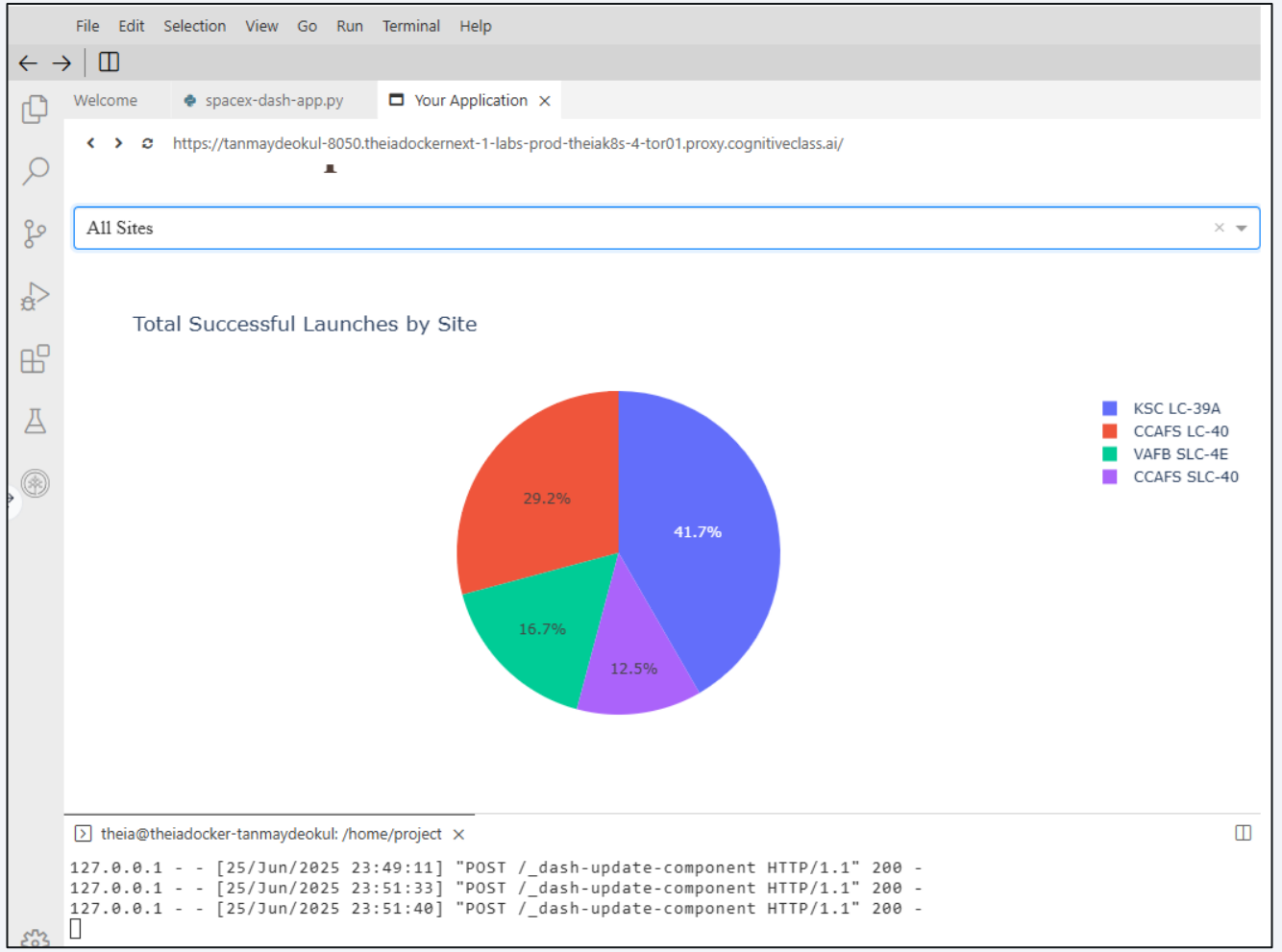



Section 4
Build a Dashboard with Plotly Dash

Comparison of Launch Success across all sites

Explanation:

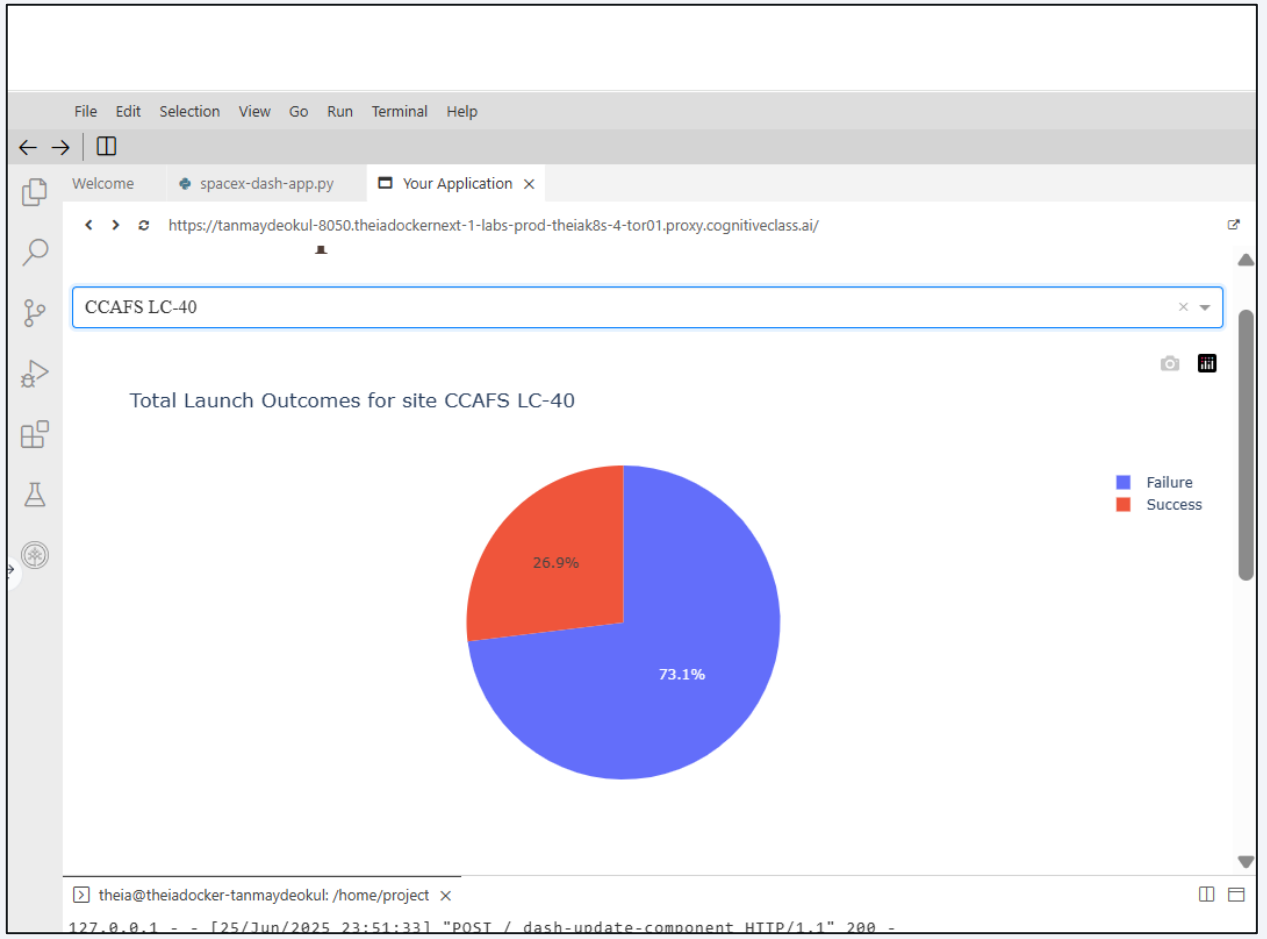
- Figure displays a pie-chart highlighting the 4 launch sites.
- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.



Launch site with highest launch success ratio

Explanation:

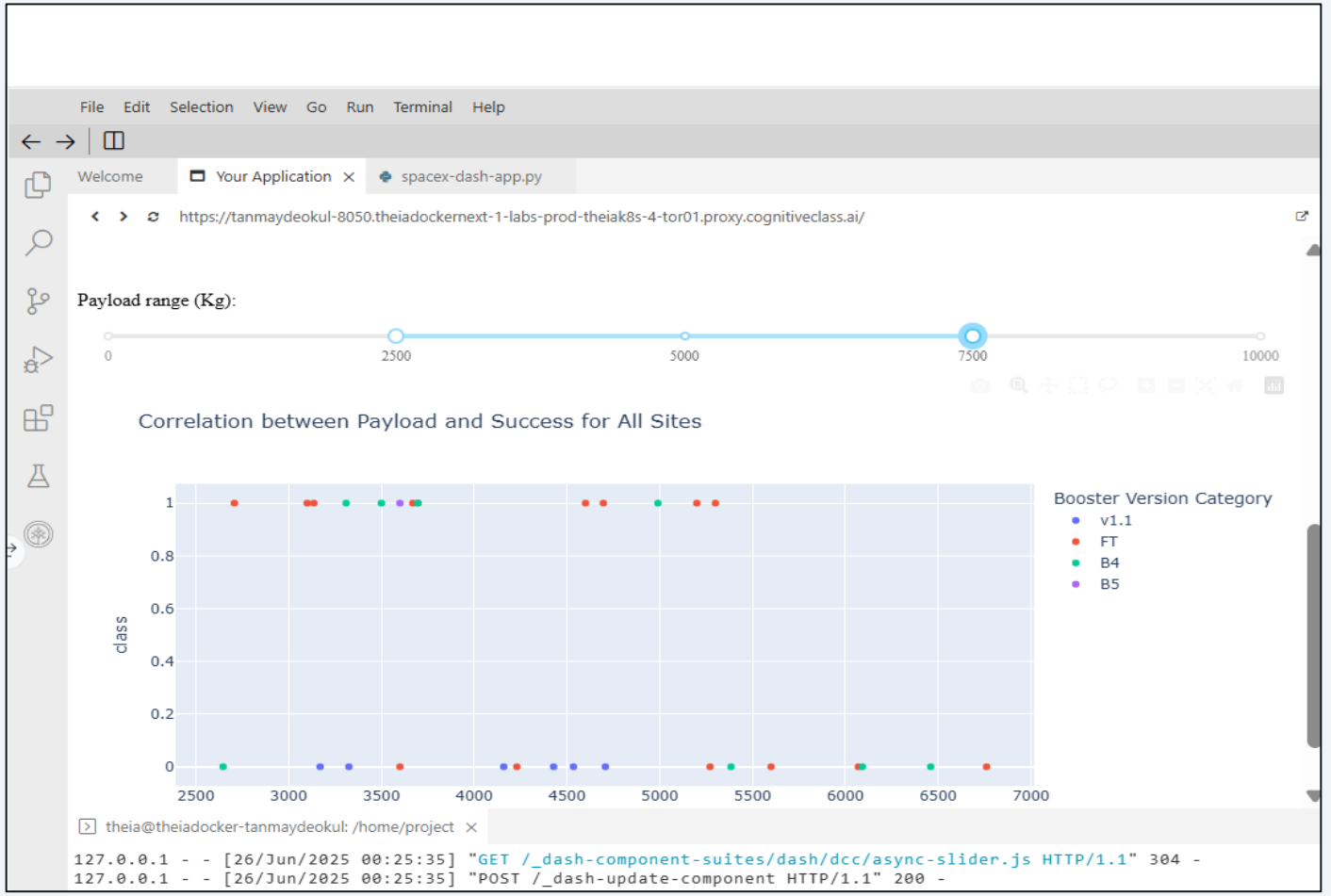
- Figure displays a pie-chart highlighting proportion of successful and failed launches at CCAFS LC-40.
- On comparison, clearly CCAFS LC-40 has the highest launch success ration.
- The chart clearly shows that roughly 70% launches have been successful.



Payload Mass vs. Launch Outcome for Booster Version Categories

Explanation:

- The charts show that payloads between 2000 and 5500 kg have the highest success rate.
- Slider can be used to select the range for payload and accordingly the scatter plot updates itself.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

```
Logistic Regression - Accuracy on Training Data: 0.8464285714285713
Support Vector Machines - Accuracy on Training Data: 0.8482142857142856
Decision Trees - Accuracy on Training Data: 0.8892857142857142
k-Nearest Neighbors - Accuracy on Training Data: 0.8482142857142858
Logistic Regression - Accuracy on Test Data: 0.8333333333333334
Support Vector Machines - Accuracy on Test Data: 0.8333333333333334
Decision Trees - Accuracy on Test Data: 0.8333333333333334
k-Nearest Neighbors - Accuracy on Test Data: 0.8333333333333334
Best Method is: Decision Trees
```

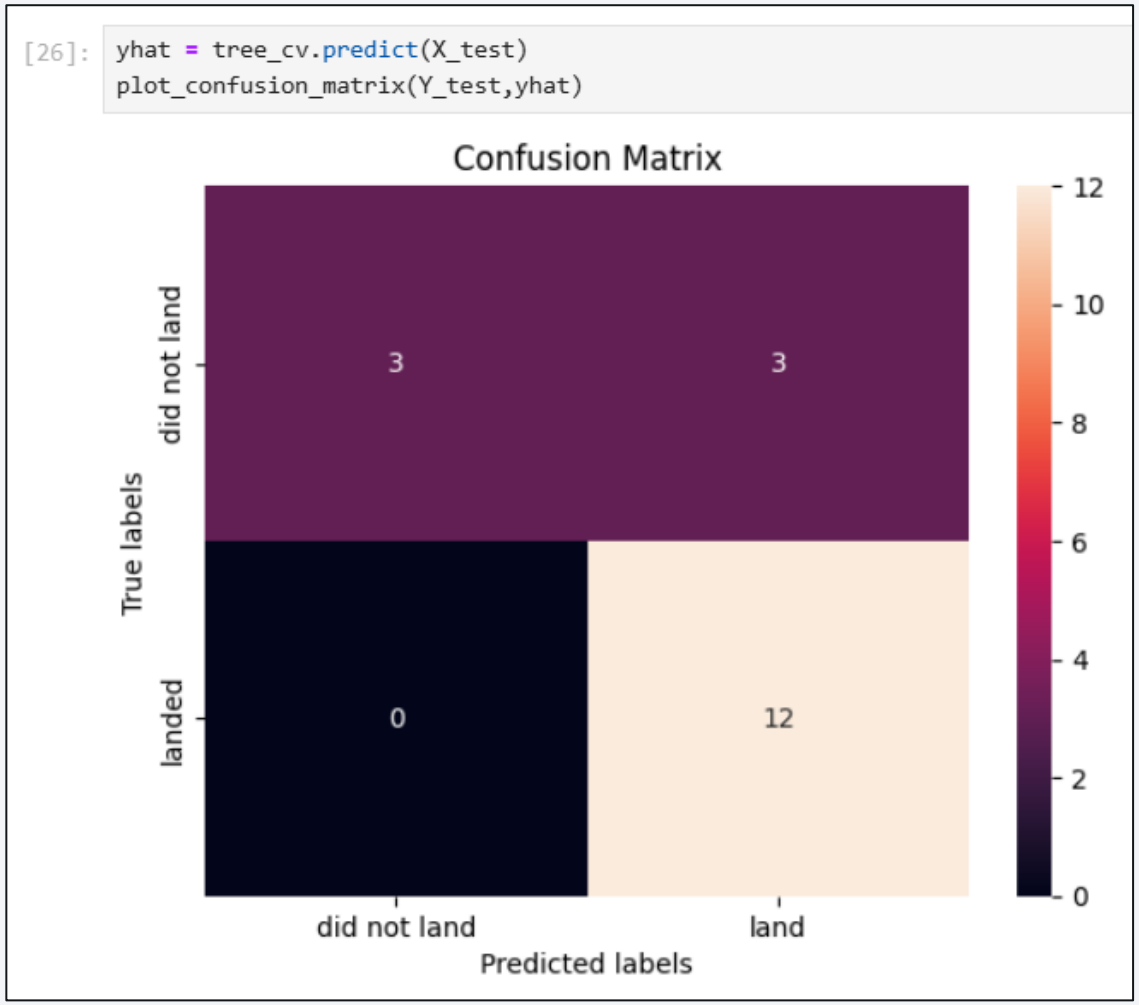
Explanation:

- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
- The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

Confusion Matrix

Explanation:

- Examining the confusion matrix, we see that decision tree can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- **Decision Tree Model is the best algorithm for this dataset.**
- **Launches with a low payload mass show better results than launches with a larger payload mass.**
- **Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.**
- **The success rate of launches increases over the years.**
- **KSC LC-39A has the highest success rate of the launches from all the sites.**
- **Orbits ES-L1, GEO, HEO and SSO have 100% success rate.**

Appendix

- GitHub Repository

Thank you!