IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Szu-Yu Chen

07/19/2024

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

SpaceX Background

SpaceX is the most successful company of the commercial space age, 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. Therefore, if we can determine if the first stage will land, we can 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

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

Section 1

Methodology

Methodology

- Data collection methodology:
  - SpaceX Rest API & Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - 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

Data Collection

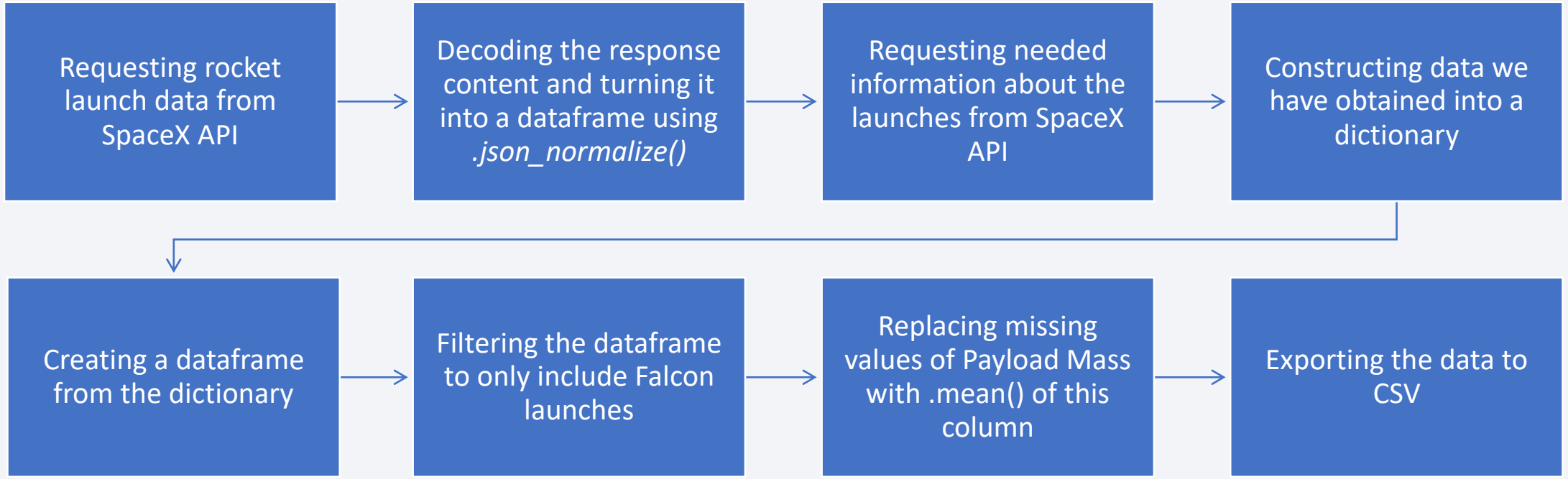
- Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from Wikipedia entry.
- To get complete information about the launches for a more detailed analysis using both of these data collection methods.
  - Columns obtained by SpaceX REST API:

    FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

  - Columns obtained by Web Scraping:

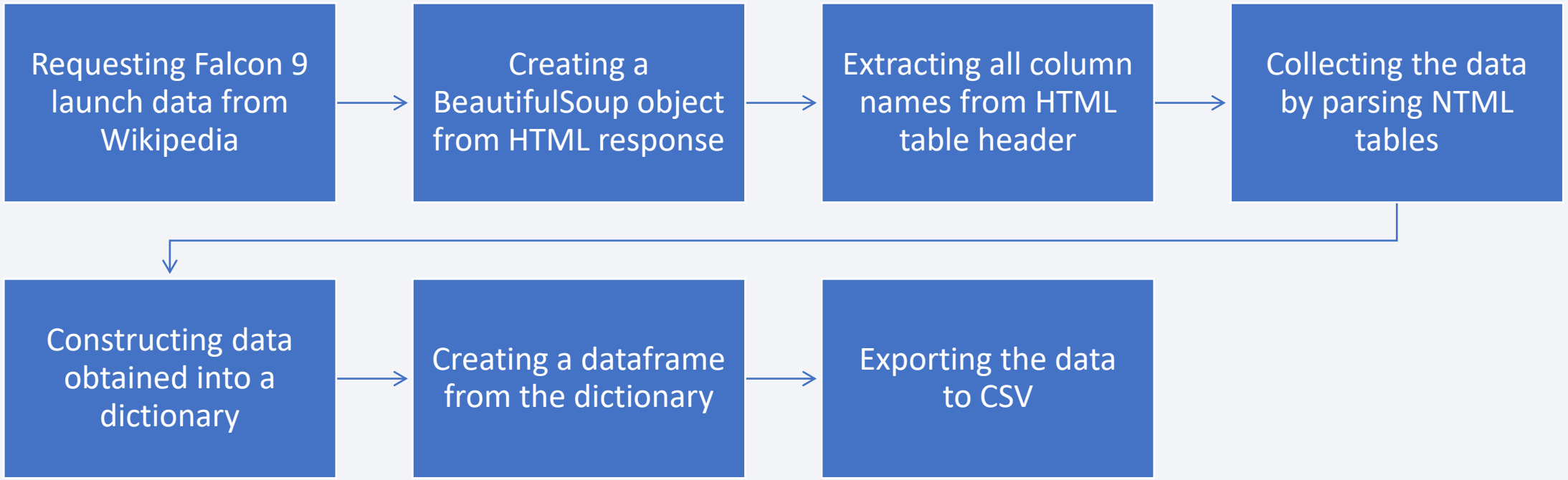
    Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



GitHub URL: SpaceX API

Data Collection - Scraping



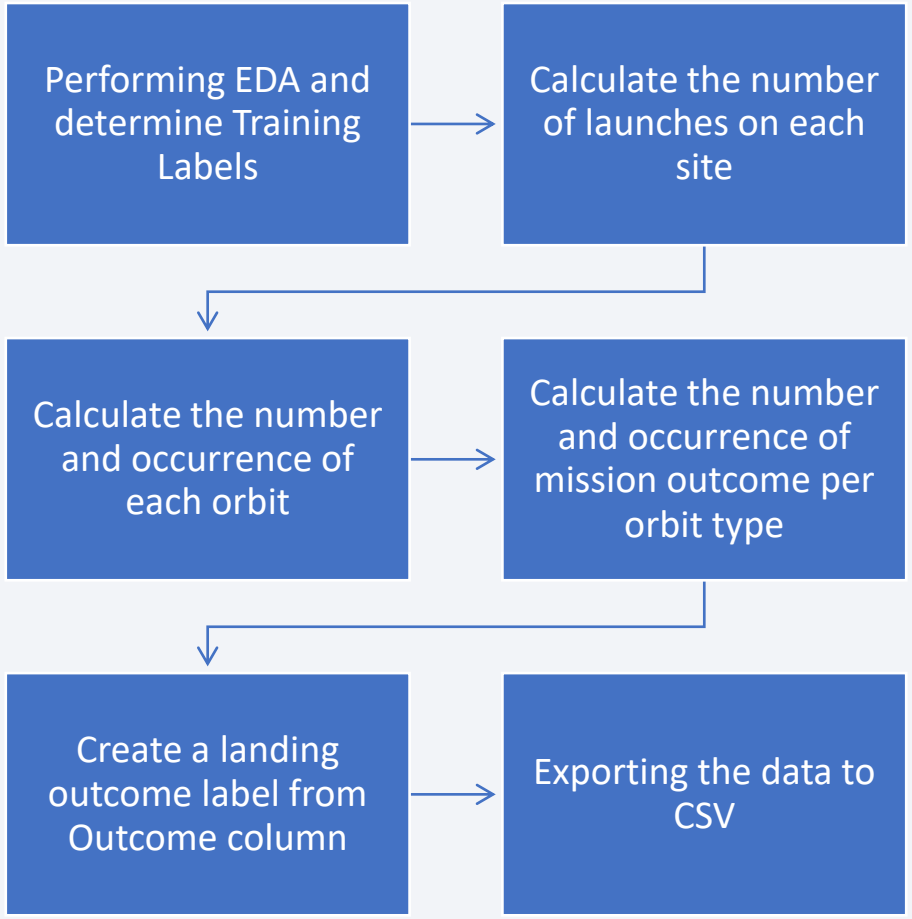
GitHub URL: SpaceX Web Scraping

Data Wrangling

There are several different cases where the booster did not land successfully.

- True Ocean: the mission outcome was successfully landed to a specific region of the ocean
- False Ocean: the mission outcome was unsuccessfully landed to a specific region of the ocean
- The RTLS: the mission outcome was successfully landed to a ground pad
- False RTLS: the mission outcome was unsuccessfully landed to a ground pad
- True ASDS: the mission outcome was successfully landed on a drone ship
- False ASDS: the mission outcome was unsuccessfully landed on a drone ship.

Converting those outcomes into Training Labels with “1” meaning booster successfully landed; “0” meaning unsuccessful.



GitHub URL: Data Wrangling

EDA with Data Visualization

- Scatter plot: showing the relationship between variables.
  - Flight Number vs. Payload Mass, Flight Number vs. Launch Sit, Payload Mass vs. Launch Site, Flight Number vs. Orbit Type, Payload Mass vs. Orbit Type
- Bar chart: showing comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.
  - Orbit Type vs. Success Rate
- Line chart: showing trends in data over time
  - Success Rate by Year

GitHub URL: EDA Data Visualization

EDA with SQL

- 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 the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- 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 coloured 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 coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City

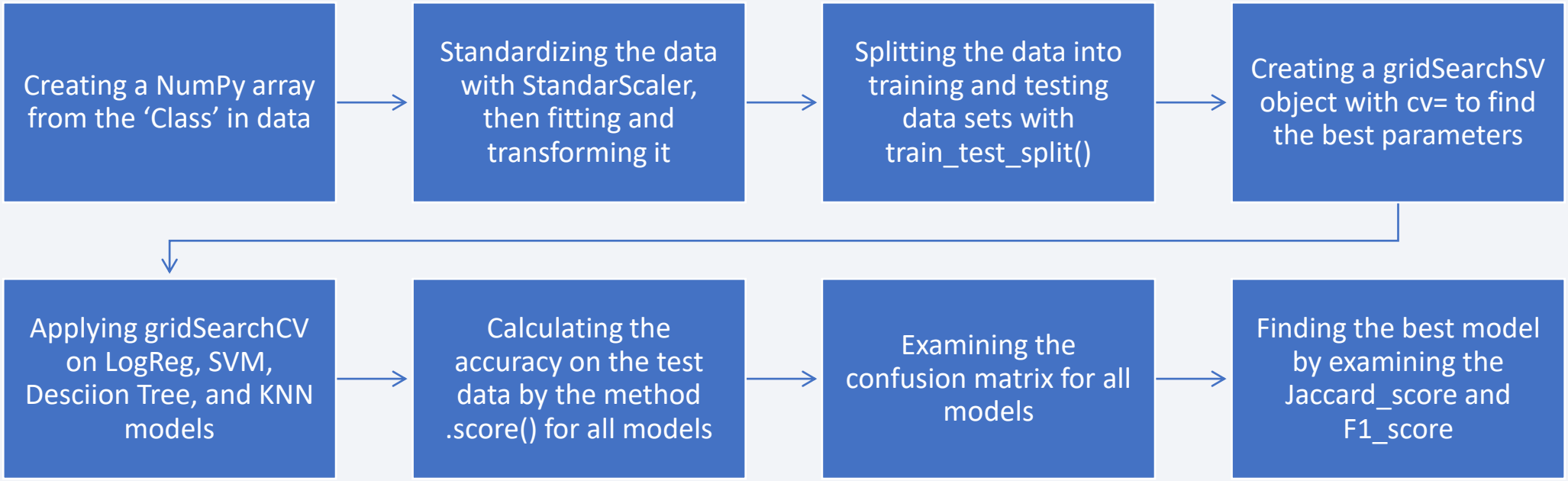
GitHub URL: Interactive Map

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List
  - A dropdown list to enable Launch Site selection
- Pie Chart showing Success Launches(All/Certain Sites)
  - Showing the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected
- Slider of Payload Mass Range
  - To select Payload range
- Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions
  - Showing the correlation between Payload and Launch Success

GitHub URL: Dash App

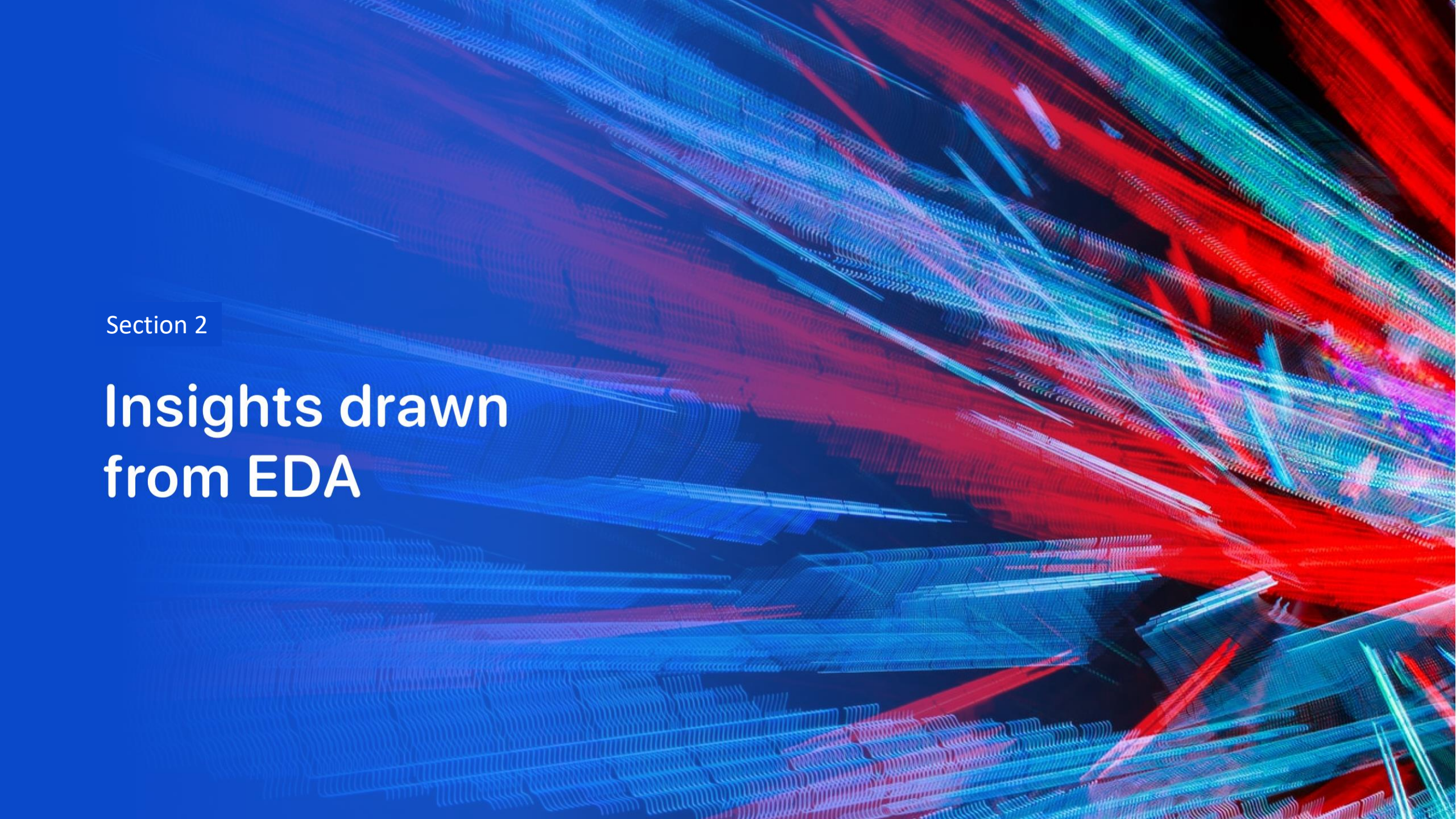
Predictive Analysis (Classification)



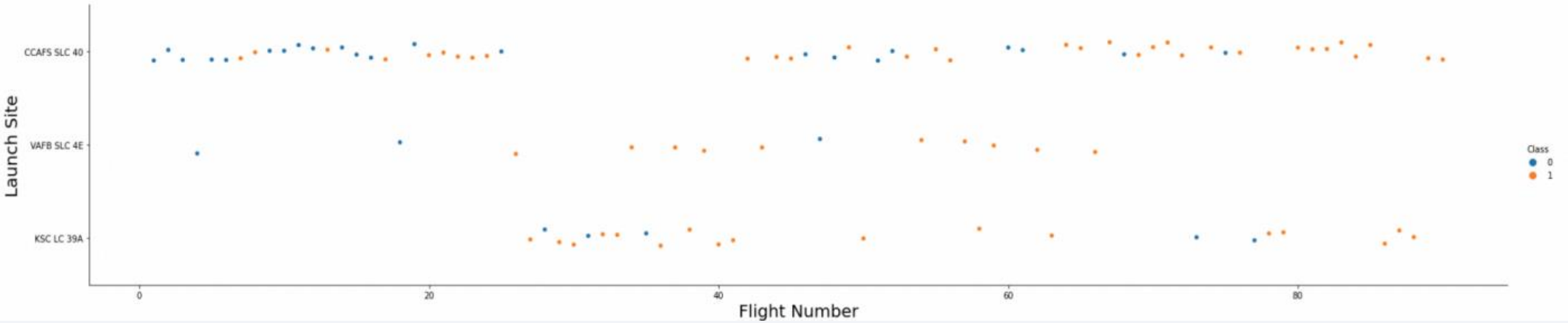
GitHub URL: Predictive Analysis

Results

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

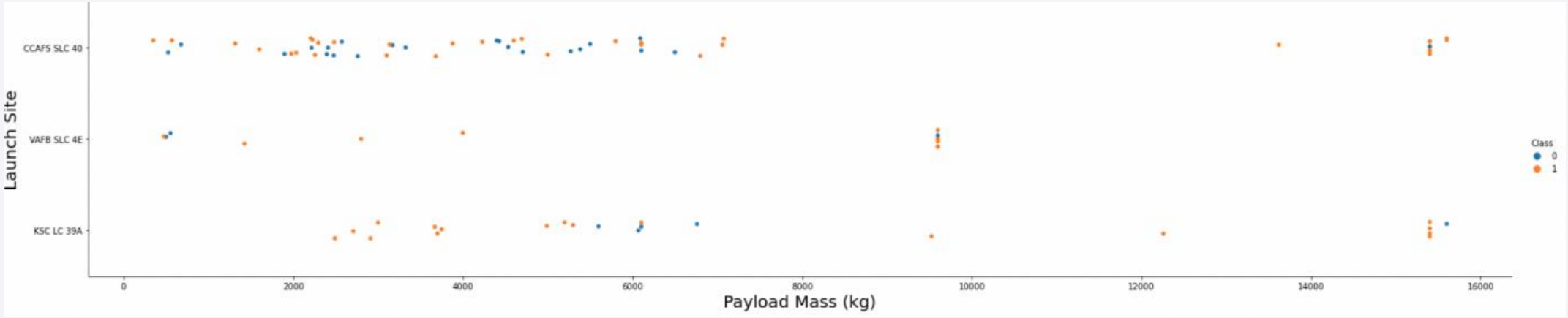

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- The earliest flights all failed while the latest flights all succeeded
- The CCAFS SLC 40 launch site has about a half of all launches
- VAFB SLC 4E and KSC LC 39A have higher success rates
- It can be assumed that each new launch has a higher rate of success

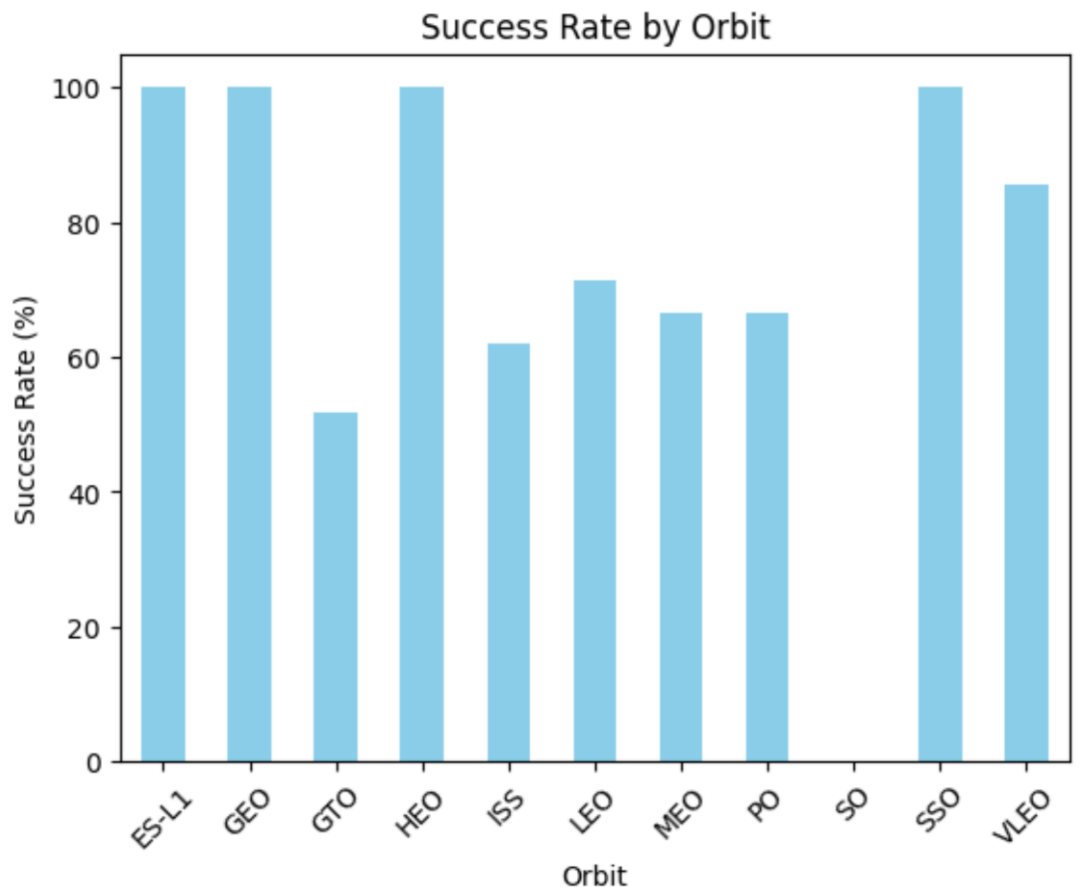
Payload vs. Launch Site



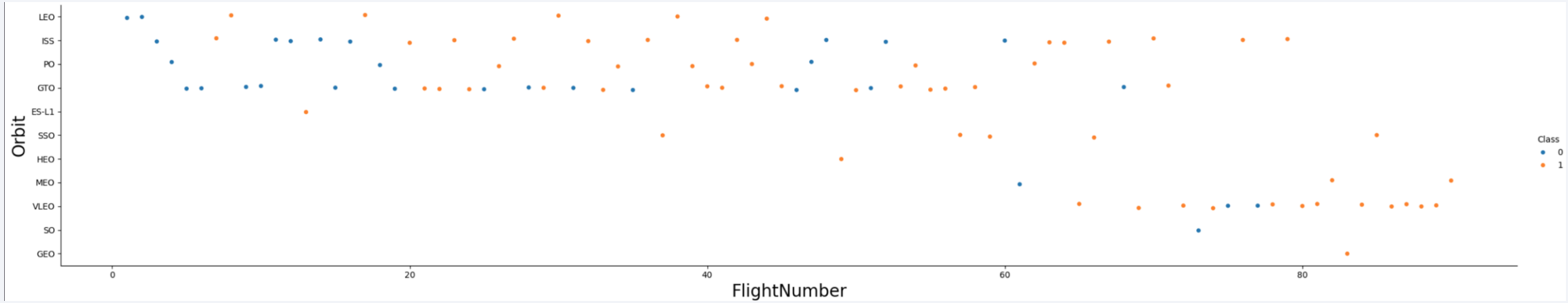
- For every launch site the higher the payload mass, the higher the success rate
- Most of the launches with payload mass over 7000 kg were successful
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too

Success Rate vs. Orbit Type

- Orbits with 100% success rate:
  - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate:
  - SO
- Orbits with success rate between 50% and 85%:
  - GTO, ISS, LEO, MEO, PO

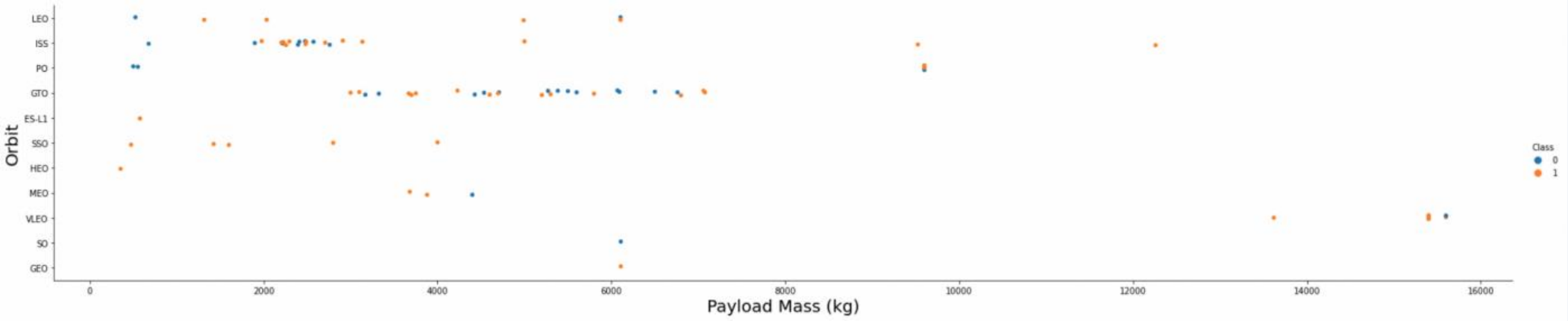


Flight Number vs. Orbit Type



In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

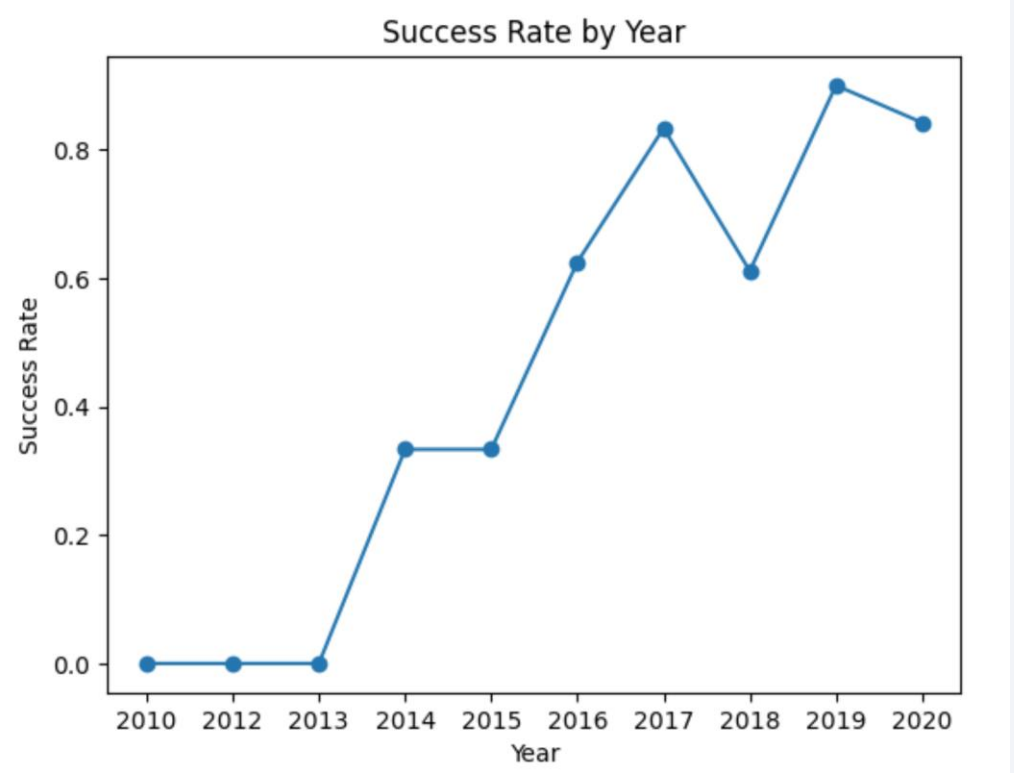
Payload vs. Orbit Type



Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits

Launch Success Yearly Trend

The success rate since 2013 kept increasing till 2020



All Launch Site Names

```
In [9]: %sql SELECT DISTINCT Launch_Site from SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

Out[9]:

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

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

Launch Site Names Begin with 'CCA'

```
%sql SELECT * from SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

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

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS `Total payload mass` FROM SPACEXTABLE WHERE Customer LIKE 'NASA (CRS)'
```

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

Total payload mass
45596

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

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS `Average payload mass` FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%'
```

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

Average payload mass
2534.6666666666665

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

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success'
```

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

MIN(Date)
2018-07-22

Listing the date when the first successful landing outcome in ground pad was achieved

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE '%Success (drone ship)%' AND PAYLOAD_MASS__KG_ BE
```

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

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

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

```
%%sql
SELECT
    CASE
        WHEN "Mission_Outcome" LIKE 'Success%' THEN 'Success'
        WHEN "Mission_Outcome" LIKE 'Failure%' THEN 'Failure'
        ELSE "Mission_Outcome"
    END AS standardized_outcome,
    COUNT(*) AS count
FROM SPACEXTABLE
GROUP BY standardized_outcome;
```

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

standardized_outcome	count
Failure	1
Success	100

Listing the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

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

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

Listing the names of the booster which have carried the maximum payload mass

2015 Launch Records

```
%%sql
SELECT
    substr(Date, 6, 2) AS month,
    Booster_Version,
    Launch_Site,
    Landing_Outcome
FROM
    SPACEXTABLE
WHERE
    Landing_Outcome LIKE '%Failure (drone ship)%'
    AND substr(Date, 0, 5) = '2015';
```

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

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

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

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

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

Landing_Outcome	outcome_count
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

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

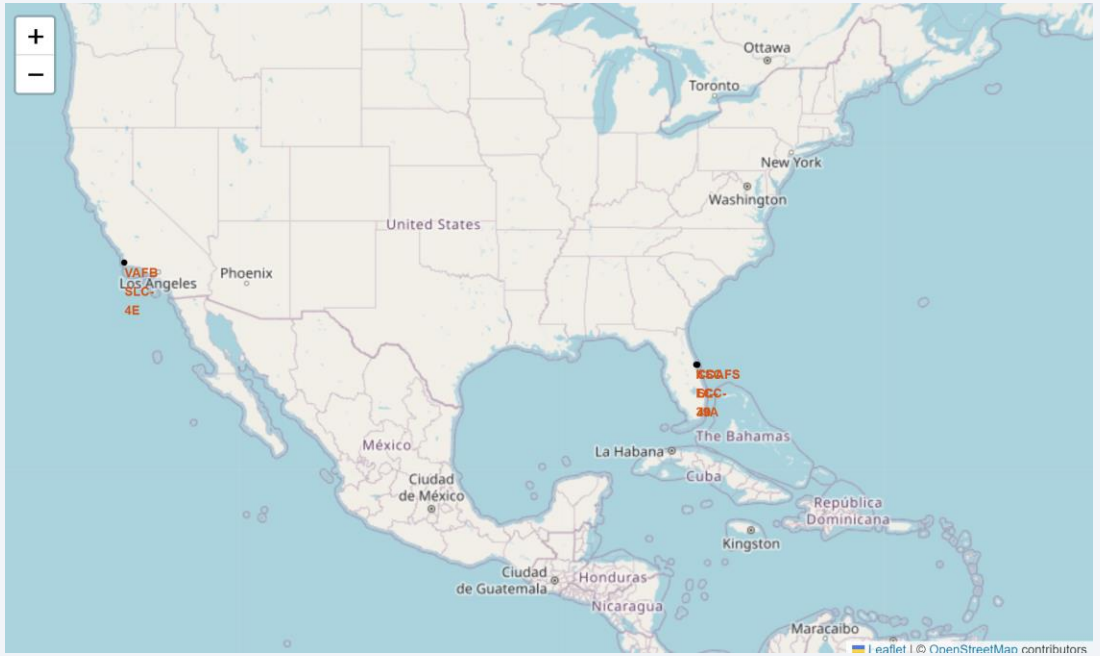
Section 3

Launch Sites Proximities Analysis

Launch Site Location Markers

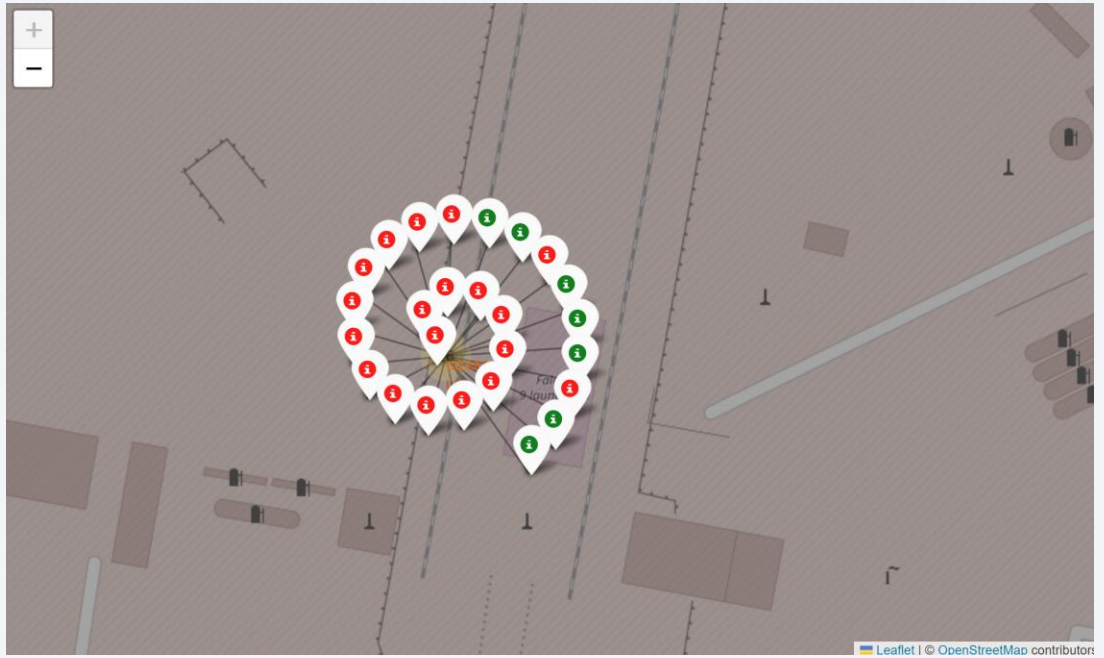
- Most of the Launch sites are in proximity to the Equator line.

- All launch sites are very close to the coast because it can minimize the risk of any debris dropping ore exploding near people



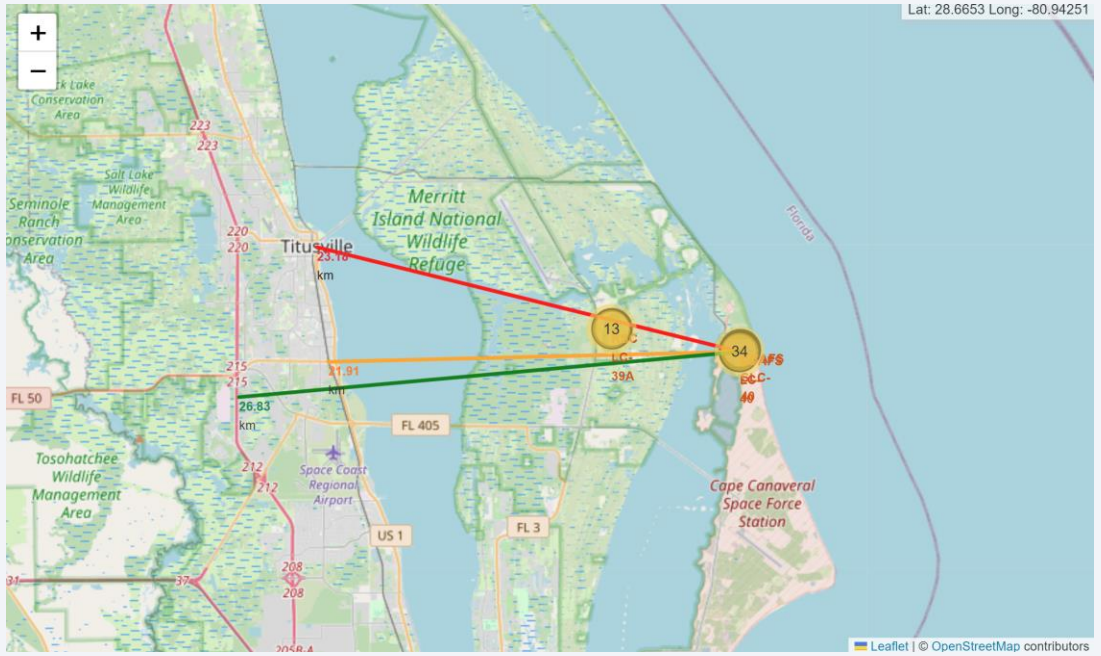
Color-labeled Launch Records

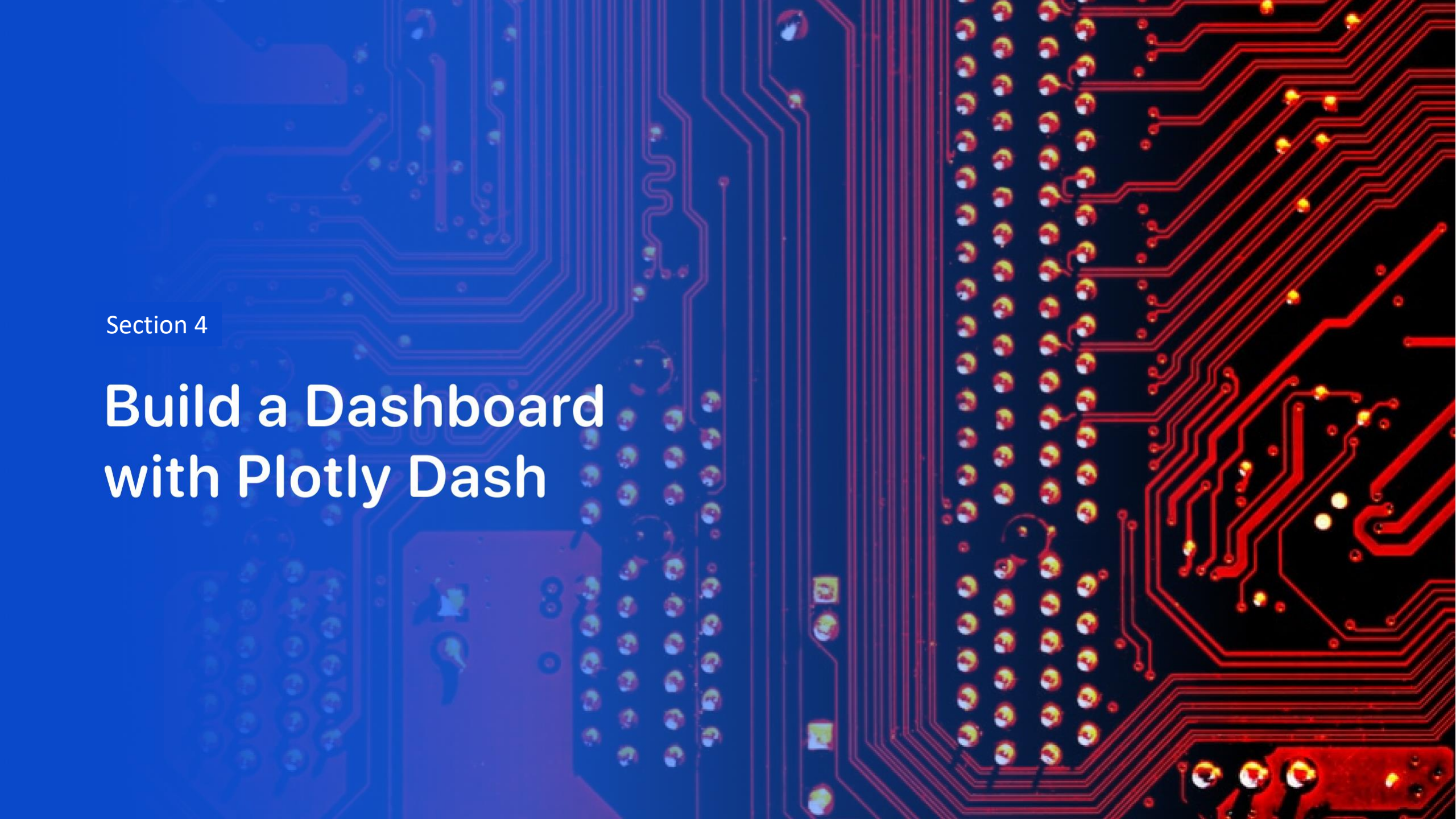
- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates:
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch



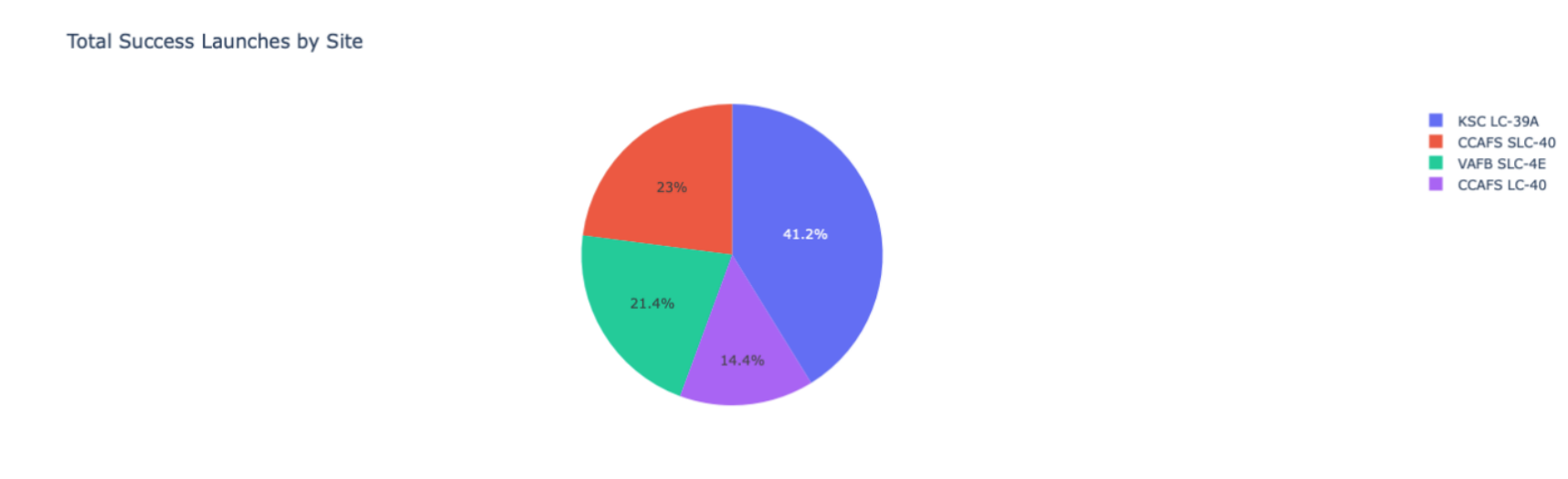
Distance from the CCAFS LC-40

- From the visual analysis of the launch site CCAFS LC-40 we can nearly see that
  - Relative close to railway(21.91 km)
  - Relative close to highway(26.83 km)
  - Relative close to city(23.16 km)
  - Just near the coast
- It could be potentially dangerous to populated areas due to the close distance from the launch site



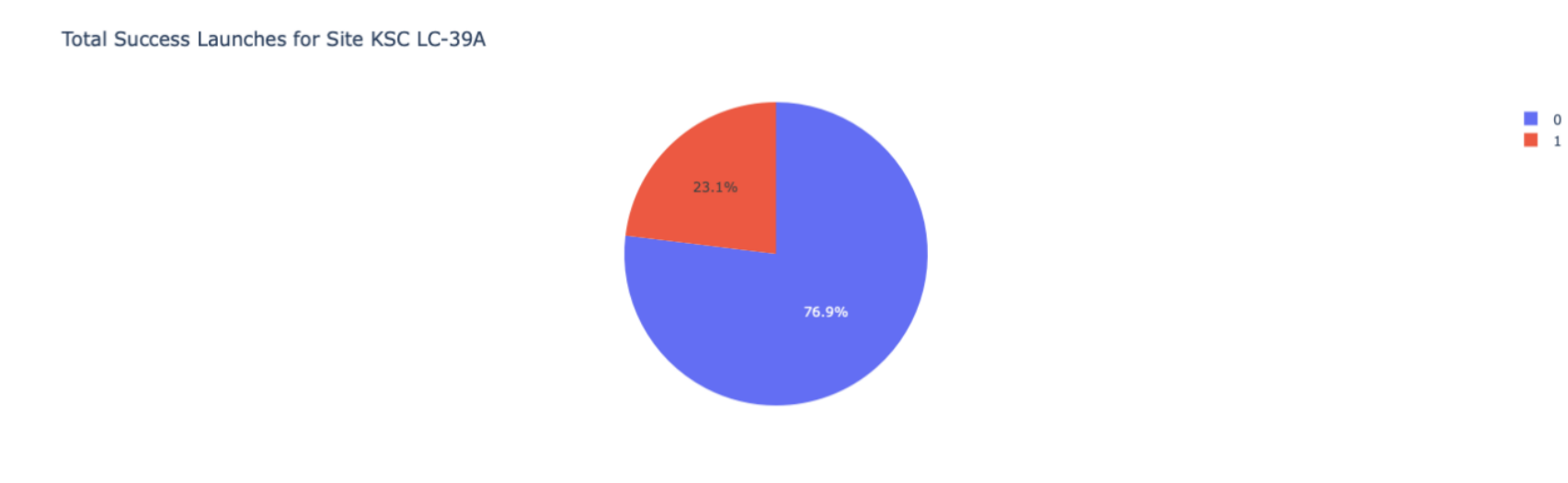

Section 4
Build a Dashboard with Plotly Dash

Successful launch count for all sites



The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches

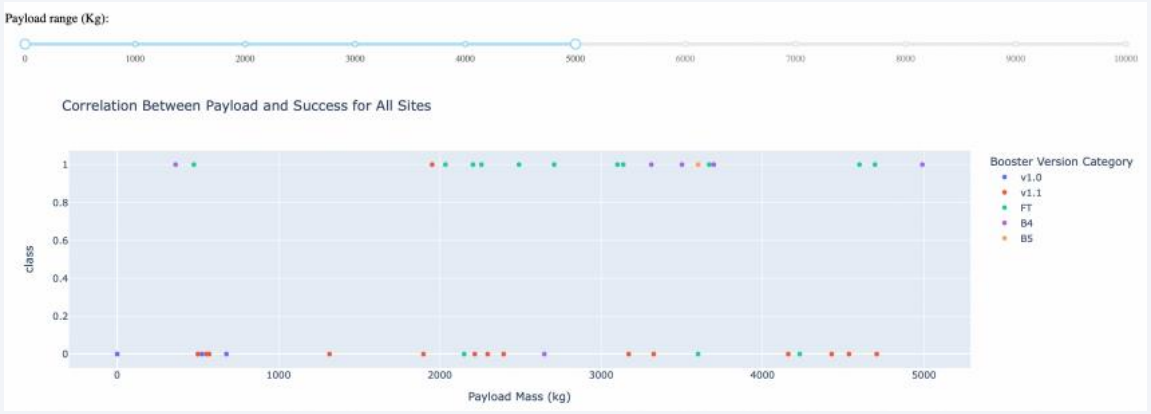
The Highest launch success ratio launch site



KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings

Payload Mass vs, Launch Outcome

Charts show that payloads between 2000 and 5500 kg have the highest success rate

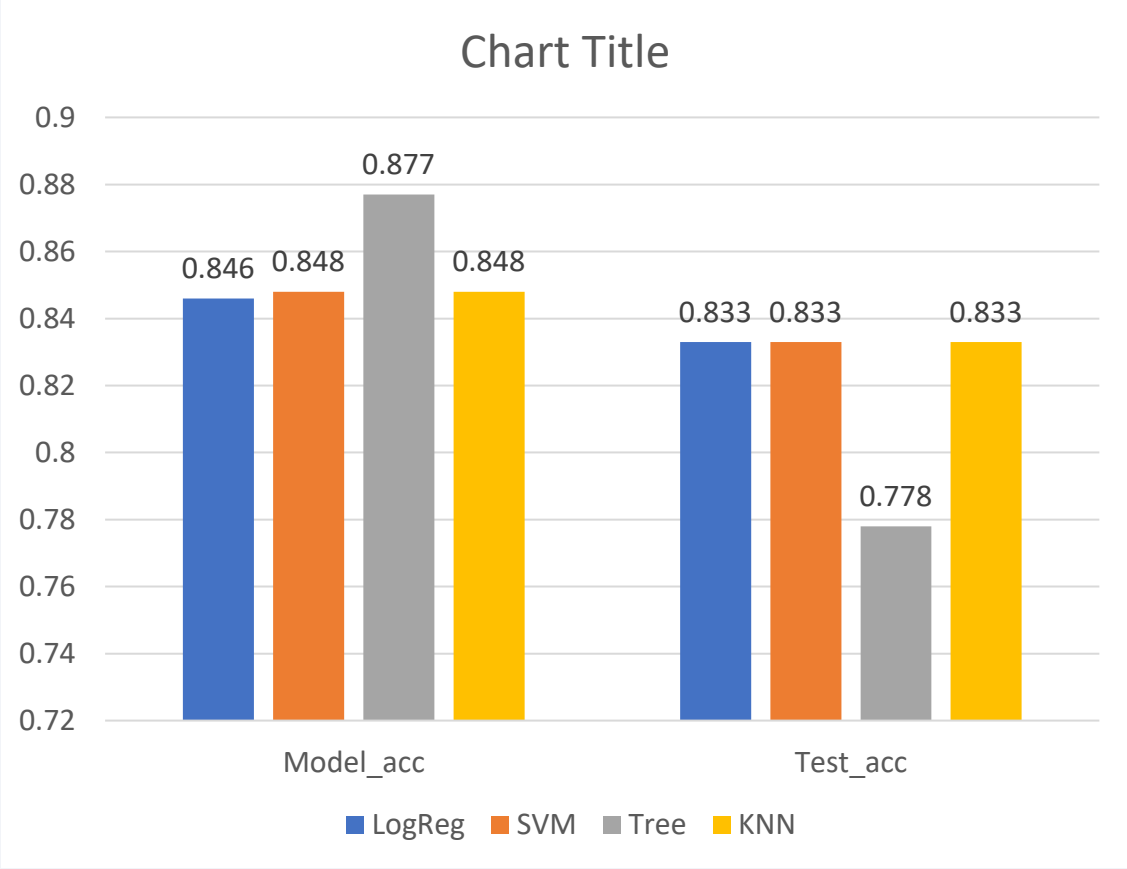


Section 5

Predictive Analysis (Classification)

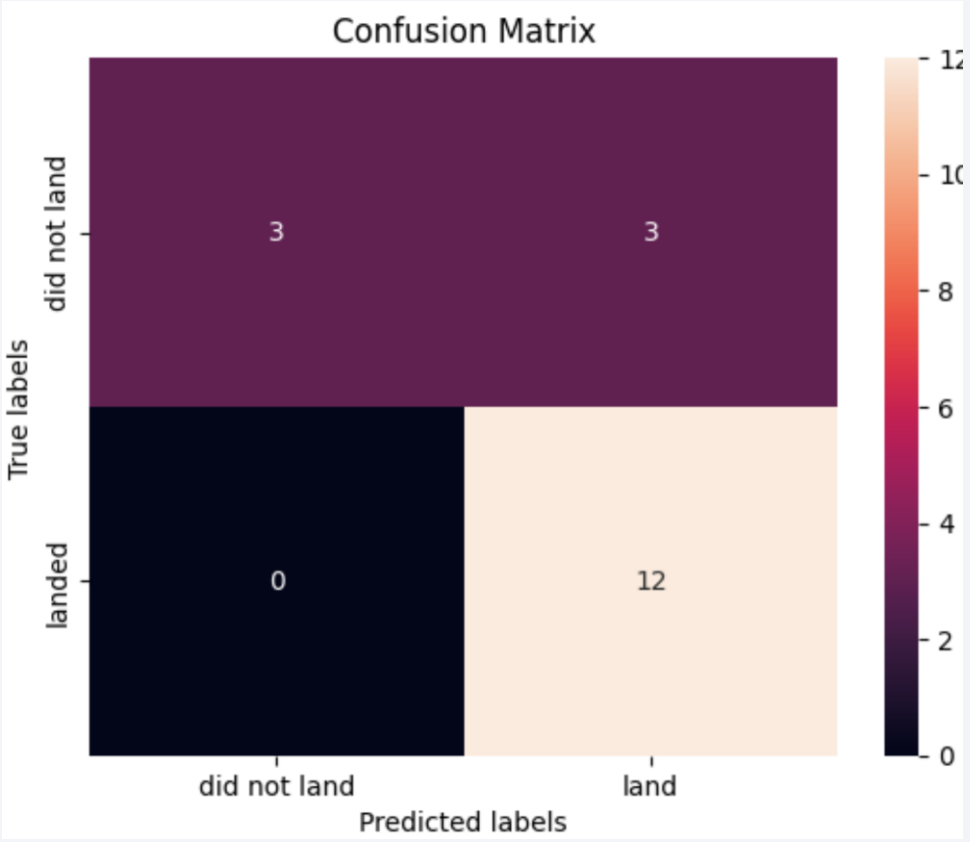
Classification Accuracy

- For the accuracy of the test dataset, the LogReg, SVM, and KNN have the same value
- For the training model accuracy, the SVM and KNN have the same best value
- Based on the above, SVM and KNN are the best classification models



Confusion Matrix

- SVM and KNN have the same confusion matrix
- In the confusion matrix, we can distinguish different classes



Conclusions

- SVM and KNN are 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 near the Equator line and all the sites are close to the coast.
- KSC LC-39A has the highest success rate of the launches form all sites

Thank you!