Winning Space Race with Data Science

Nelson Espiritu

October 25, 2023

Outline

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

Executive Summary

- These are the following methodologies used::
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Building Interactive Maps with Folium
  - Building a Dashboard using Plotly Dash
  - Predictive Analysis (Classification
- Summary of all results
  - The above mentioned method and its methodologies is an effective procedure to follow to convey useful business decision

Introduction

- By using public information and data science methodologies, it is our aim to determine the cost of a launch and predict whether SpaceX can reuse the first stage of a launch.
- It is also our aim to determine how specific variables affect the success of a first stage landing.
- Does the rate of successful landings increase over time?

Section 1

Methodology

Methodology

- Data collection methodology:
  - The following data were obtained from:
    - Space X REST API
    - Web Scraping from Wikipedia
- Perform data wrangling
  - Data was filtered
  - Dealt with missing or null values
  - Prepared the data for 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
  - Data was divided into combination of training and data sets and evaluated using four classification models

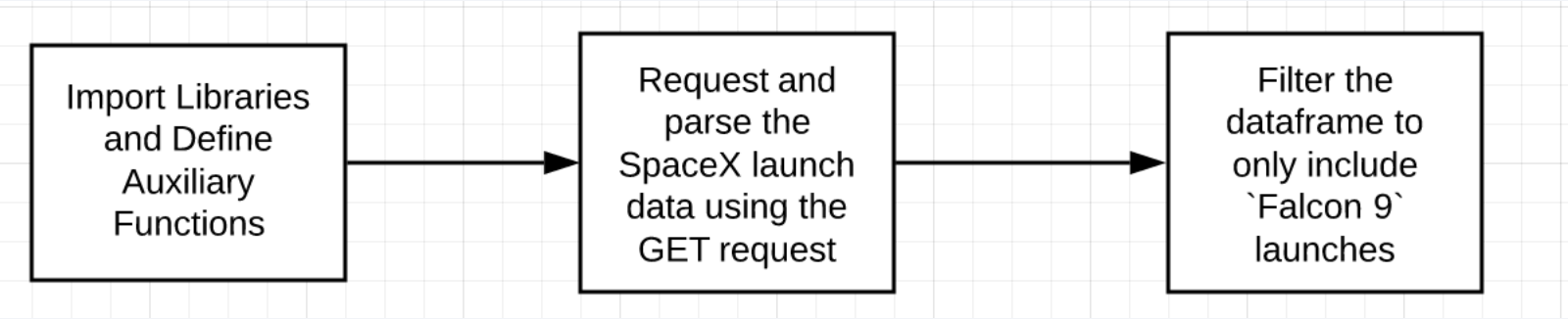
Data Collection

- Data collection involved the process of API requests from SpaceX REST API and Web Scraping from a table in SpaceX's Wikipedia entry.

- The following data and columns were collected both from the two mentioned sources:

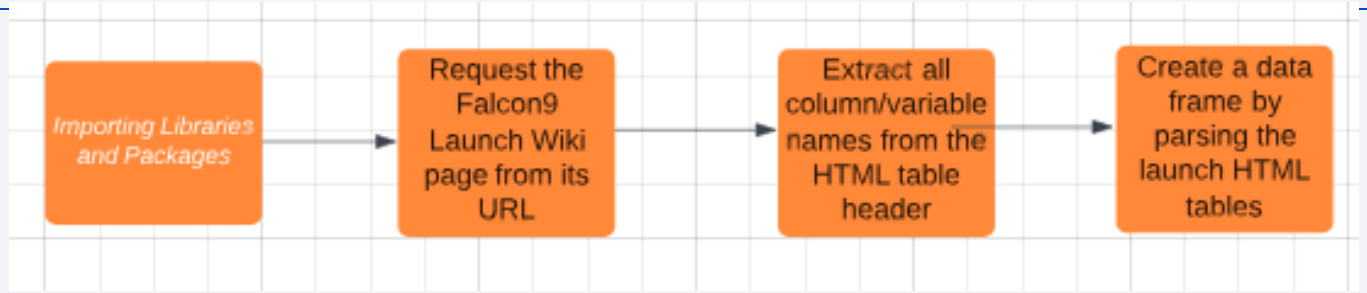
FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, Customer, LaunchSite, Outcome, Flights, Gridfins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude, Time

Data Collection – SpaceX API



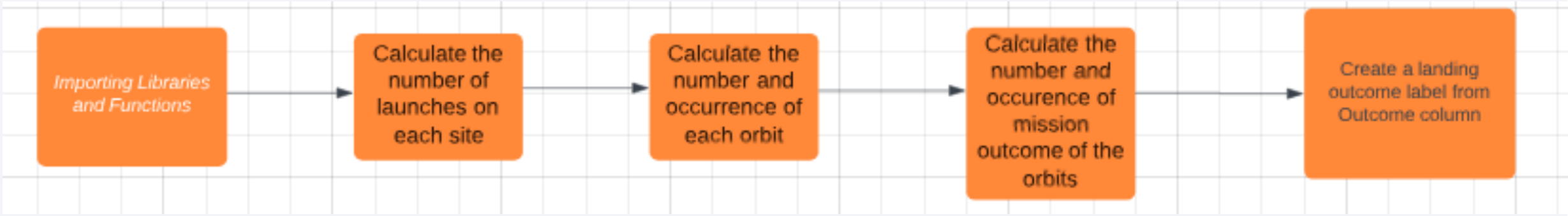
- We always begin by importing the necessary libraries to ensure successful data collection, getting requesting the data from the source and filtering necessary information for our study, in this case, Falcon 9 launches.
- Source: https://github.com/Tisisith/Applied-Data-Science-Capstone-Project/blob/bde8fa8d673dd6c611121aacfea7809d10cbf9be/Data%20Collection.ipynb

Data Collection - Scraping



- By importing packages and libraries for scraping, you can now begin the process of extracting important data from HTML tables and also putting those data into a useful data frame
- Source: https://github.com/Tisisith/Applied-Data-Science-Capstone-Project/blob/bde8fa8d673dd6c611121aacfea7809d10cbf9be/Web scraping.ipynb

Data Wrangling



- First, we import necessary libraries and functions. Then we load the data set and calculate needed values using dataframe functions
- Source: https://github.com/Tisisith/Applied-Data-Science-Capstone-Project/blob/bde8fa8d673dd6c611121aacfea7809d10cbf9be/Data%20Wrangling.ipynb

EDA with Data Visualization

- Th following charts were plotted: Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend
- Scatter plots were used to show relationship between variables.
- Bar charts show comparisons of values among categorical variables.
- Line charts show trends in data over time.
- Source: https://github.com/Tisisith/Applied-Data-Science-Capstone-Project/blob/bde8fa8d673dd6c611121aacfea7809d10cbf9be/EDA%20with%20Visualization.ipynb

EDA with SQL

Using SQL, we queried for the following data:

- Names of the unique launch sites in the space mission
- 5 records where launch sites begin with ’CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass greater than 4,000 but less than 6,000
- Total number of successful and failed mission outcomes
- Names of the booster versions which carried the maximum payload mass
- Failed landing outcomes in drone ship, the booster versions and launch site names for the months in year 2015
- Rank of landing outcomes (failure of drone ship) or Success (ground pad) between the date 2010-06-04 and 2017-03-20 in descending order

Build an Interactive Map with Folium

- Using Folium, we were able to add markers indicating launch sites
- We also added colored markers for successful (green) and failed (red) launches to show which sites have high success rates
- We added colored lines to indicate distance between launch site CCAFS SLS-40 and its proximity to the nearest coastline, railway, highway, and city
- Source: https://github.com/Tisisith/Applied-Data-Science-Capstone-Project/blob/bde8fa8d673dd6c611121aacfea7809d10cbf9be/Folium.ipynb

Build a Dashboard with Plotly Dash

- Using Dashboard with Plotly Dash:
  - We created a dropdown list with launch sites
  - We created a Pie Chart showing successful launches
  - We provided a slider of payload mass range
  - Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

  - Source: https://github.com/Tisisith/Applied-Data-Science-Capstone-Project/blob/bde8fa8d673dd6c611121aacfea7809d10cbf9be/Plotly.py

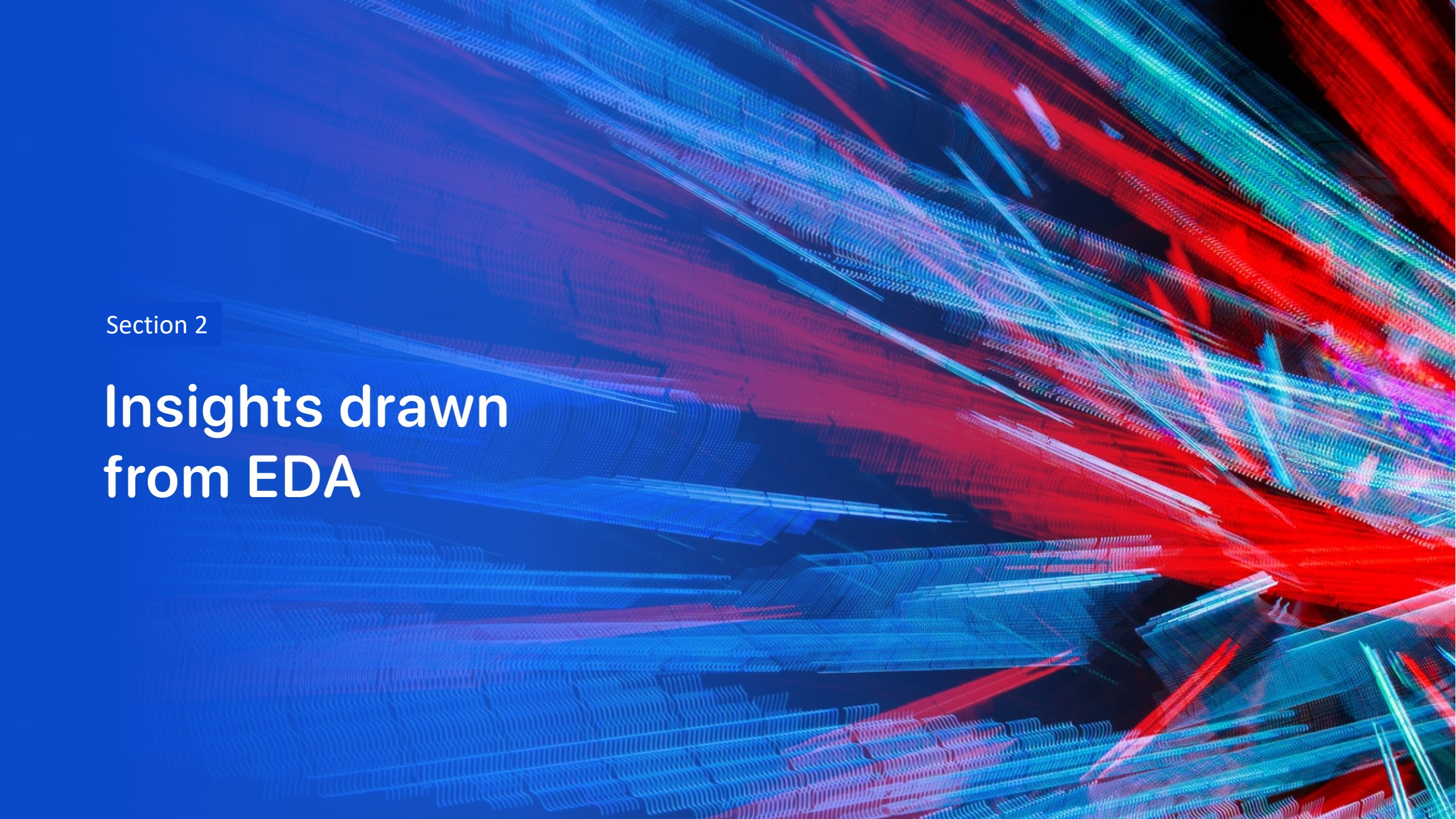
Predictive Analysis (Classification)

- First, we create NumPy array
- Standardize, fit and transform the data
- Split the train and testing data sets
- Calculate accuracy on the test data for all models
- Assess confusion matrix
- Identify the best model for classification

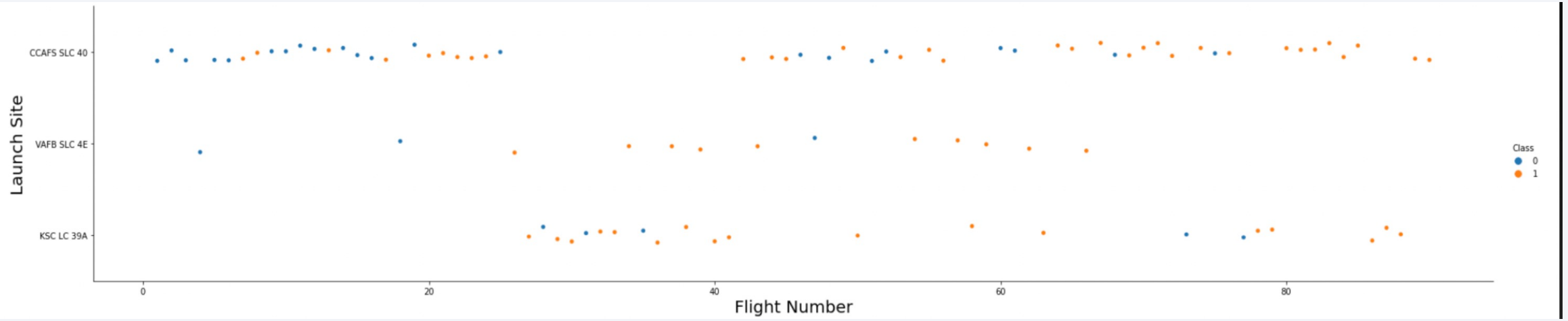
- Source: https://github.com/Tisisith/Applied-Data-Science-Capstone-Project/blob/bde8fa8d673dd6c611121aacfea7809d10cbf9be/Machine%20Learning.ipynb

Results

- Exploratory data analysis results
  - Launch success has improved over the years
  - Orbits ES-L1, GEO, HEO and SSO have 100% success rates
  - Among landing sites, KSC LC-39A has the highest success rate
- Visual and interactive analytics
  - Launch sites are near the equator, all are close to the coast
  - Launch sites are far enough from cities, highway and railways.
- Predictive analysis results
  - Decision Tree model is the best predictive model among 4 classification models

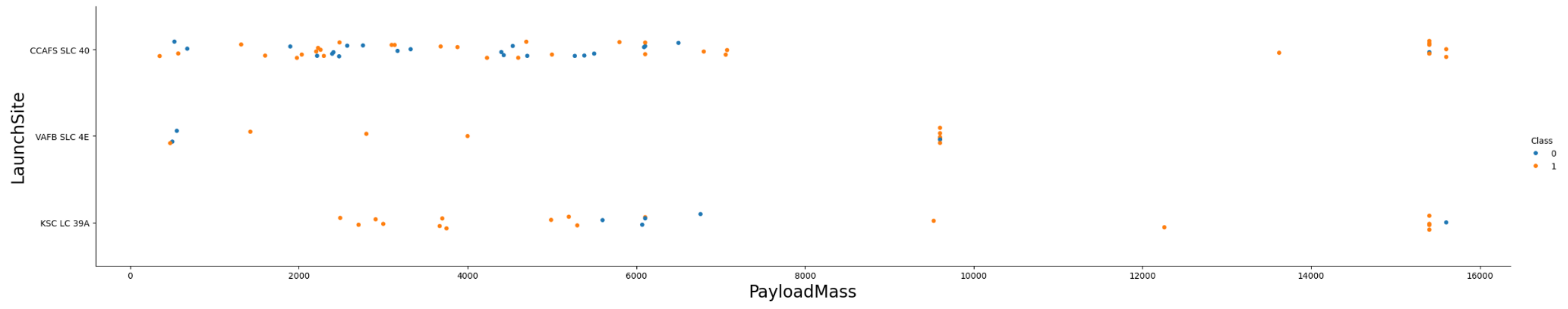

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

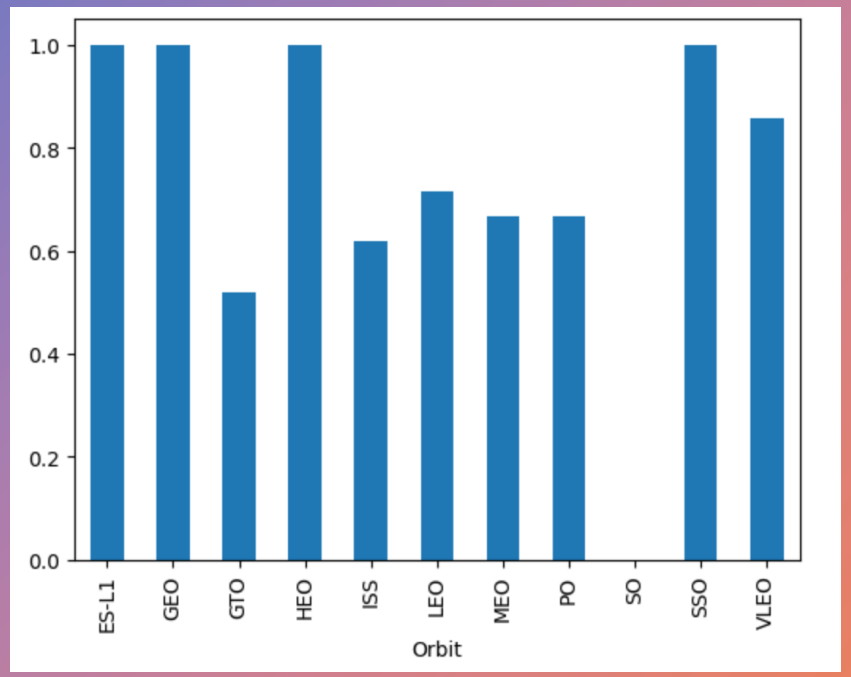


- Data shows that earliest flights all failed and the recent launches succeeded.
- Newer launches has higher success rates

Payload vs. Launch Site



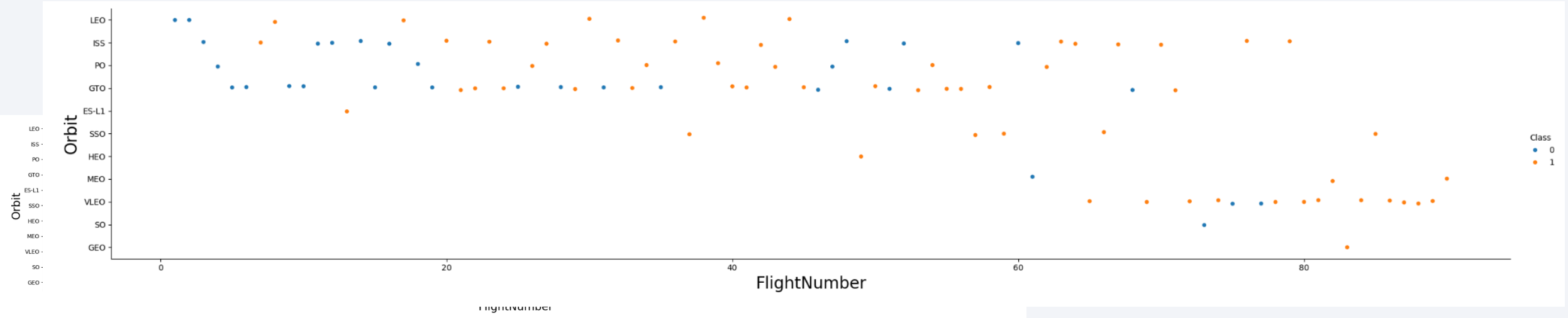
- The higher the payload mass of a launch site, the higher the success rate
- The launches with a payload mass of over 7000kg were most successful.



Success Rate vs. Orbit Type

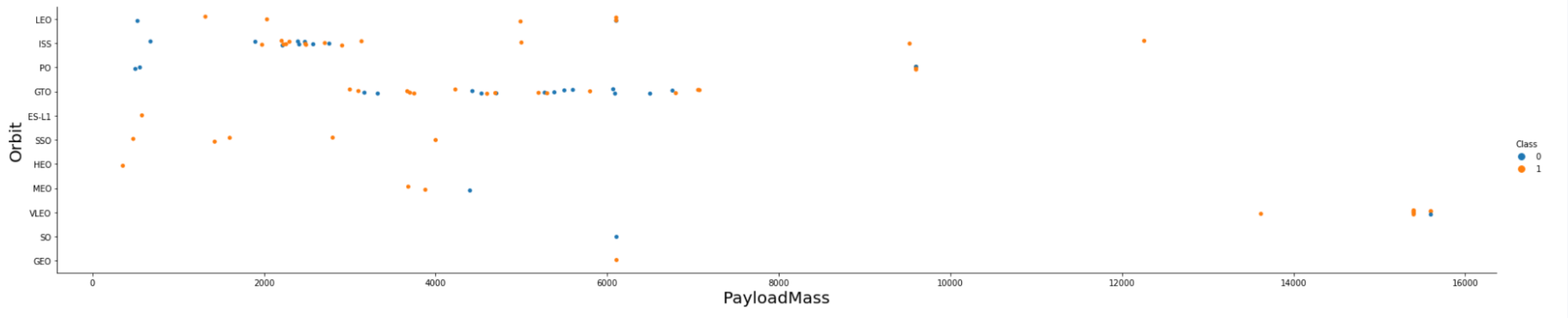
- 4 Orbits have 100% success rates
- 1 has 0% success rate
- 5 orbits have between 50% - 85% success rate

Flight Number vs. Orbit Type



- You should see that 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.

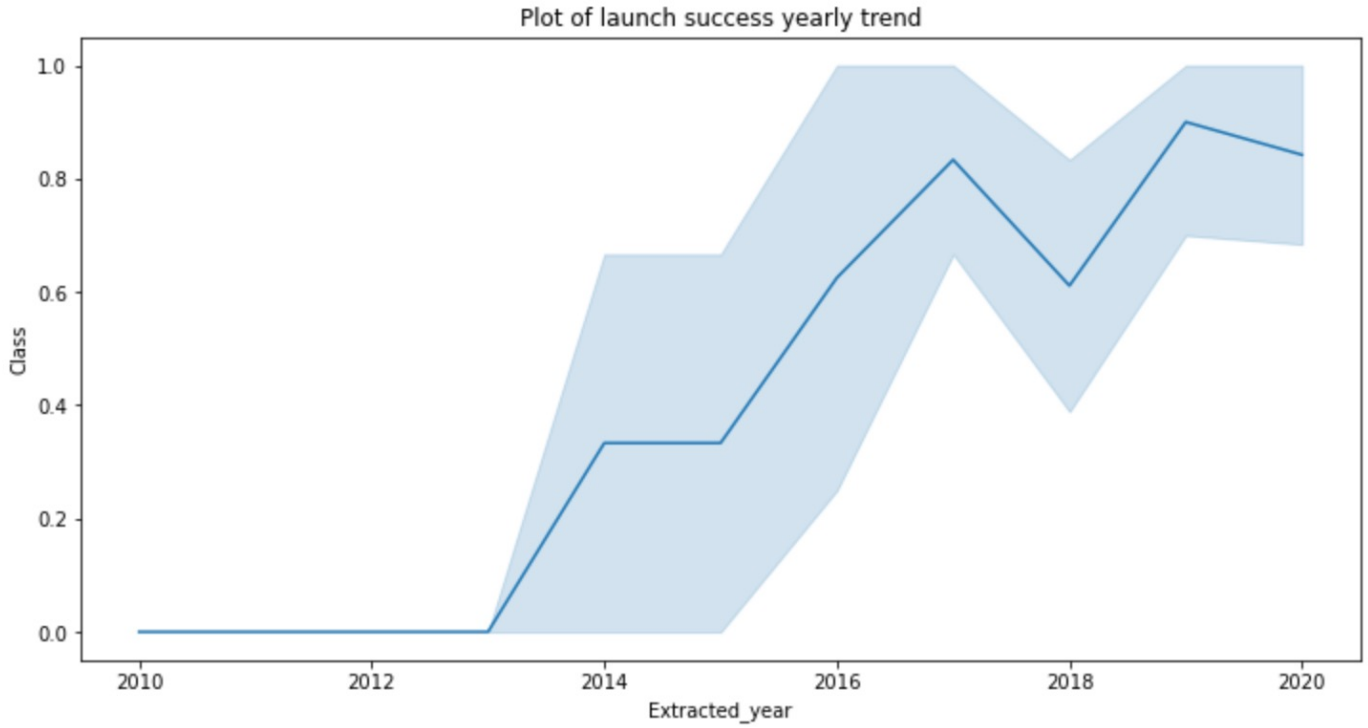
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- We can observe that the sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- By using the the SQL query mentioned, we can generate the names of the launch sites in the space mission

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

```
[8]: sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
 * sqlite:///my_data1.db
Done.
```

[8]:

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

Launch Site Names Begin with 'CCA'

- By using this query, we can generate launches where names begin with the string ’CCA’

```
[9]: sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

[9]:

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

Total Payload Mass

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

```
[10]: sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD LIKE '%CRS%';
```

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

[10]:

TOTAL_PAYLOAD
111268

Average Payload Mass by F9 v1.1

- We can find the average payload mass carried by booster version F9 v1.1

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

```
[12]: sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

[12]:

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

December 22, 2022 was the first successful Ground Landing Date

```
[13]: sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

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

[13]:

FIRST_SUCCESS_GP
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- These are the successful drone ships between 4000 and 6000 that had successful landing

```
[14]: sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND Landing_Outcome = 'Success (drone ship)';
```

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

[14]:

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

Total Number of Successful and Failure Mission Outcomes

- These are the total number of successful and failure mission outcomes

```
[15]: sql SELECT Mission_Outcome, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY Mission_Outcome ORDER BY Mission_Outcome;
```

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

[15]:

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

Boosters Carried Maximum Payload

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

```
[16]: %sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION
```

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

[16]:

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

*** 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.**

```
[21]: sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Failure (drone ship)' AND DATE_PART('YEAR', DATE) = 2015;
```

Out[18]:

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

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.

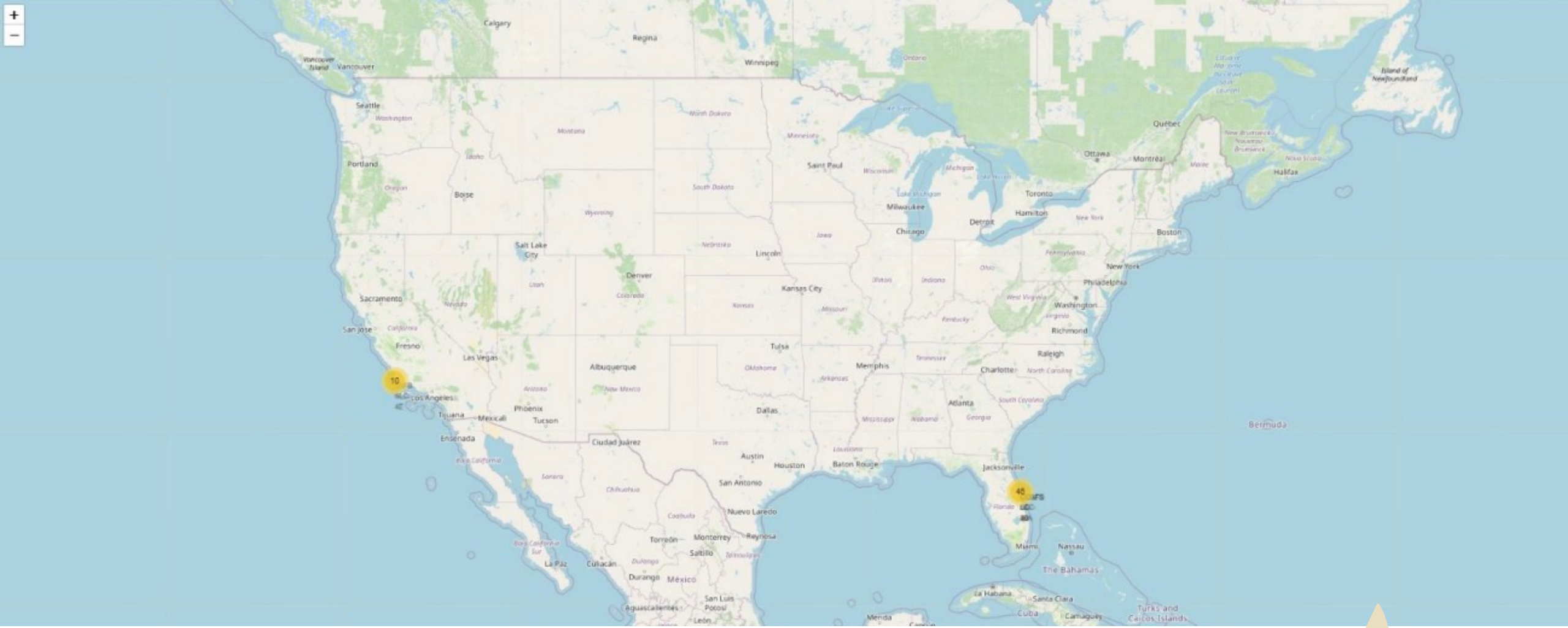
```
[22]: sql SELECT Landing_Outcome, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY L
```

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

[22]:

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

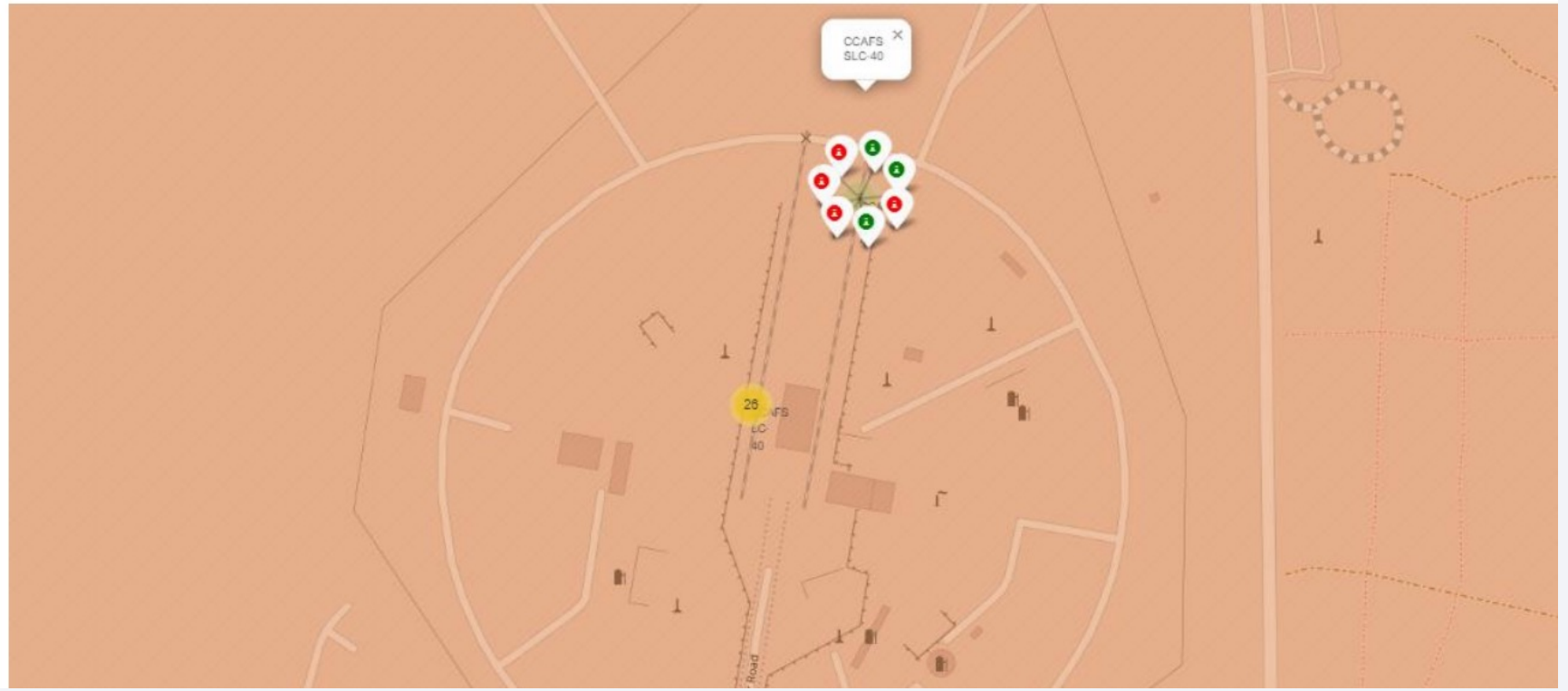

Section 3
Launch Sites
Proximities Analysis



Launch Sites Location

Launch sites are near the equator to take advantage of the earth’s natural rotation.

Launch Outcomes

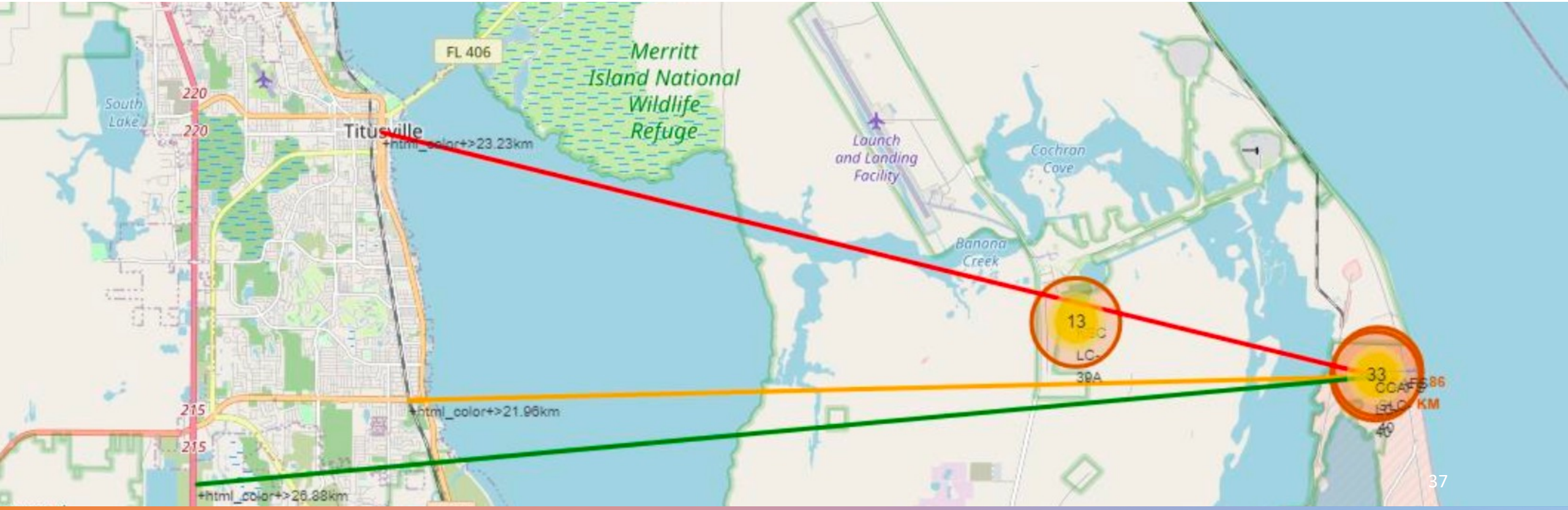


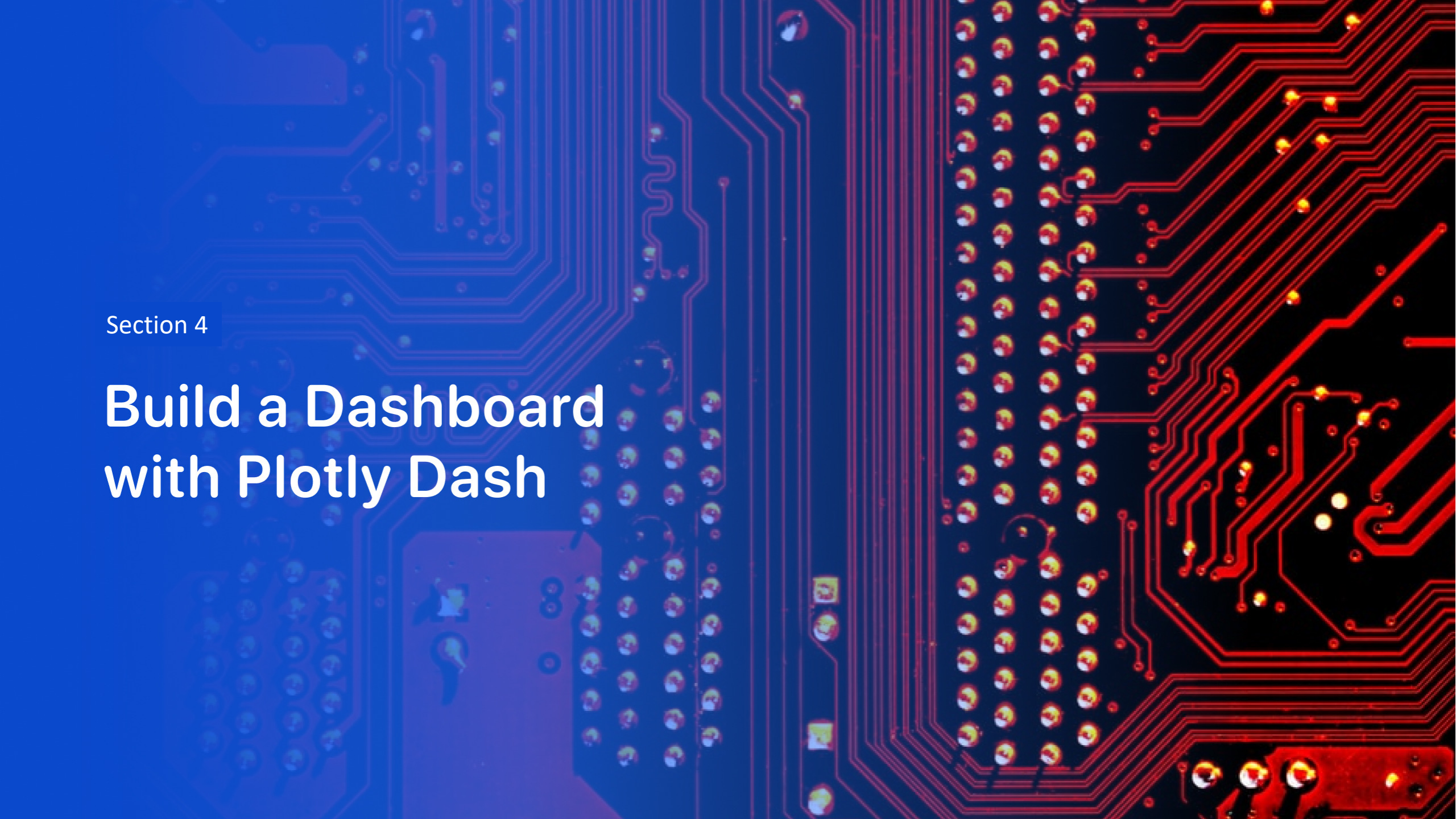
Red indicates failed launches

Green indicate successful launches

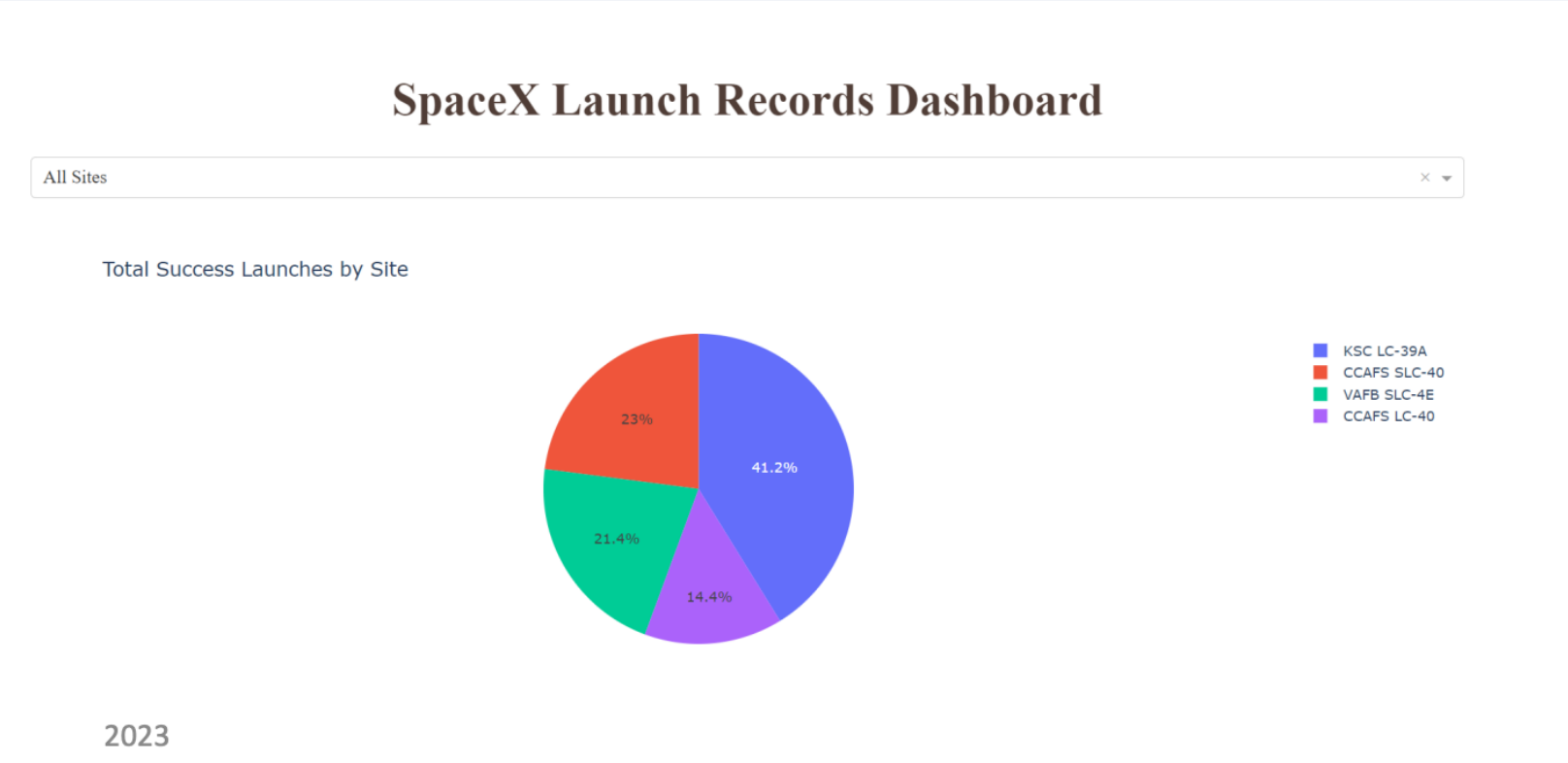
Distance to Nearest Coastline, Railway, City and Highway

- Less than a km from the nearest coastline
- 21.96 km from nearest railway
- 23.23 km from nearest city
- 26.88 km from nearest highway




Section 4
Build a Dashboard with Plotly Dash

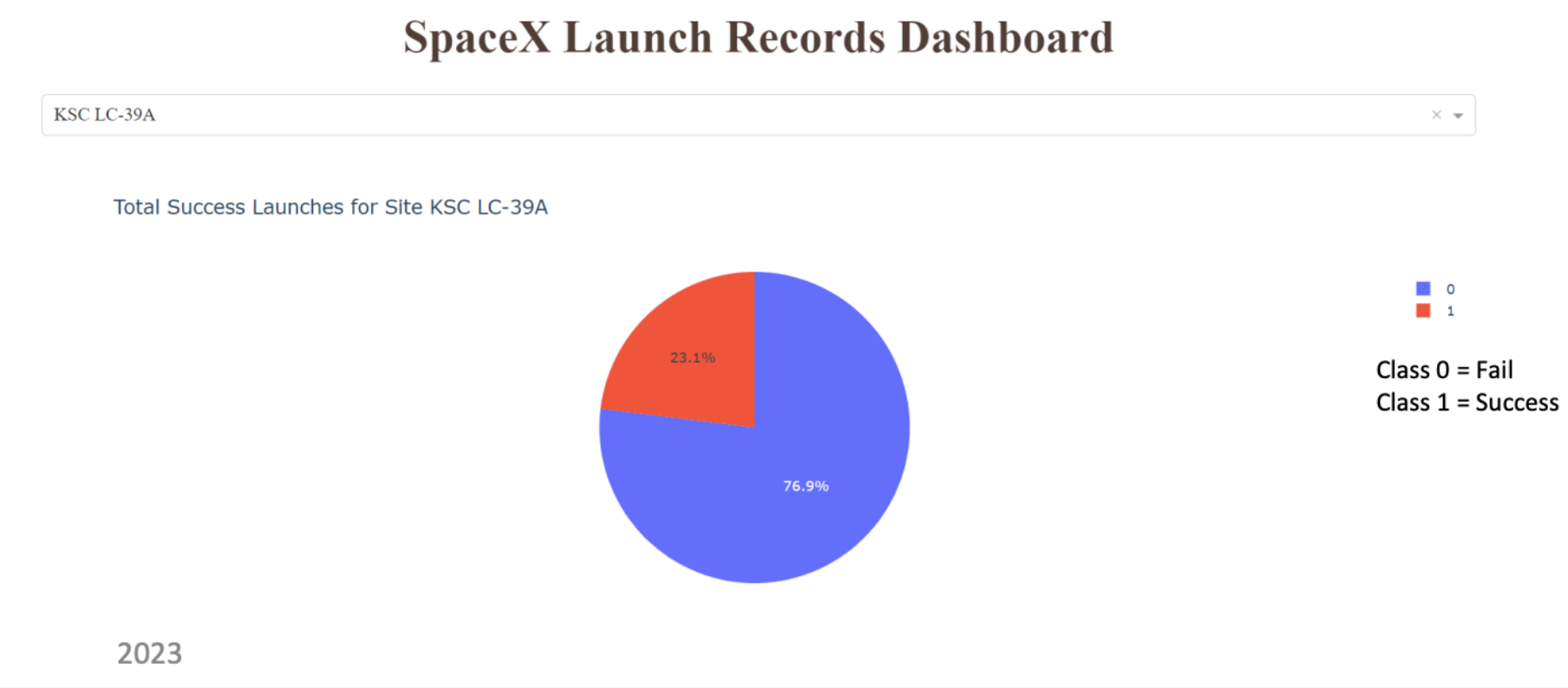
Total Success Launches by Site



- KSC LC-39A has the most successful launch sites (41.2%)

Launch Success of KSC LC-29A

- KSC has 13 total launches (10 successful and 3 failed, 76.9%)



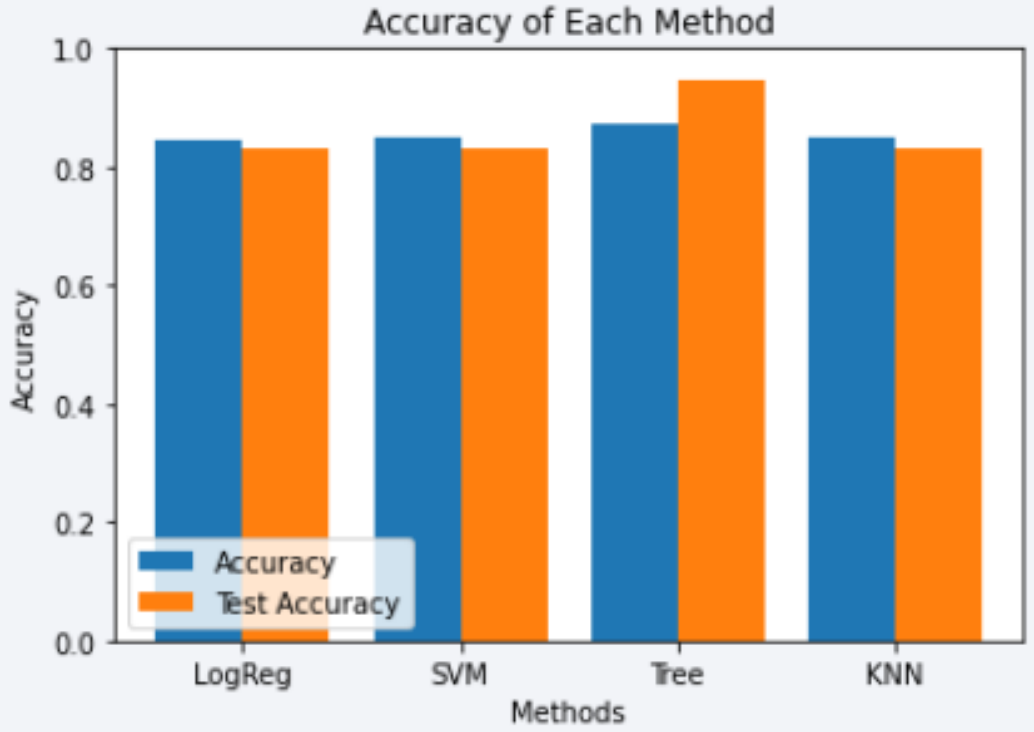
Payload Mass and Success by Booster Version

- The payload between 2,000 and 5,000 have the highest success rate.

Section 5

Predictive Analysis (Classification)

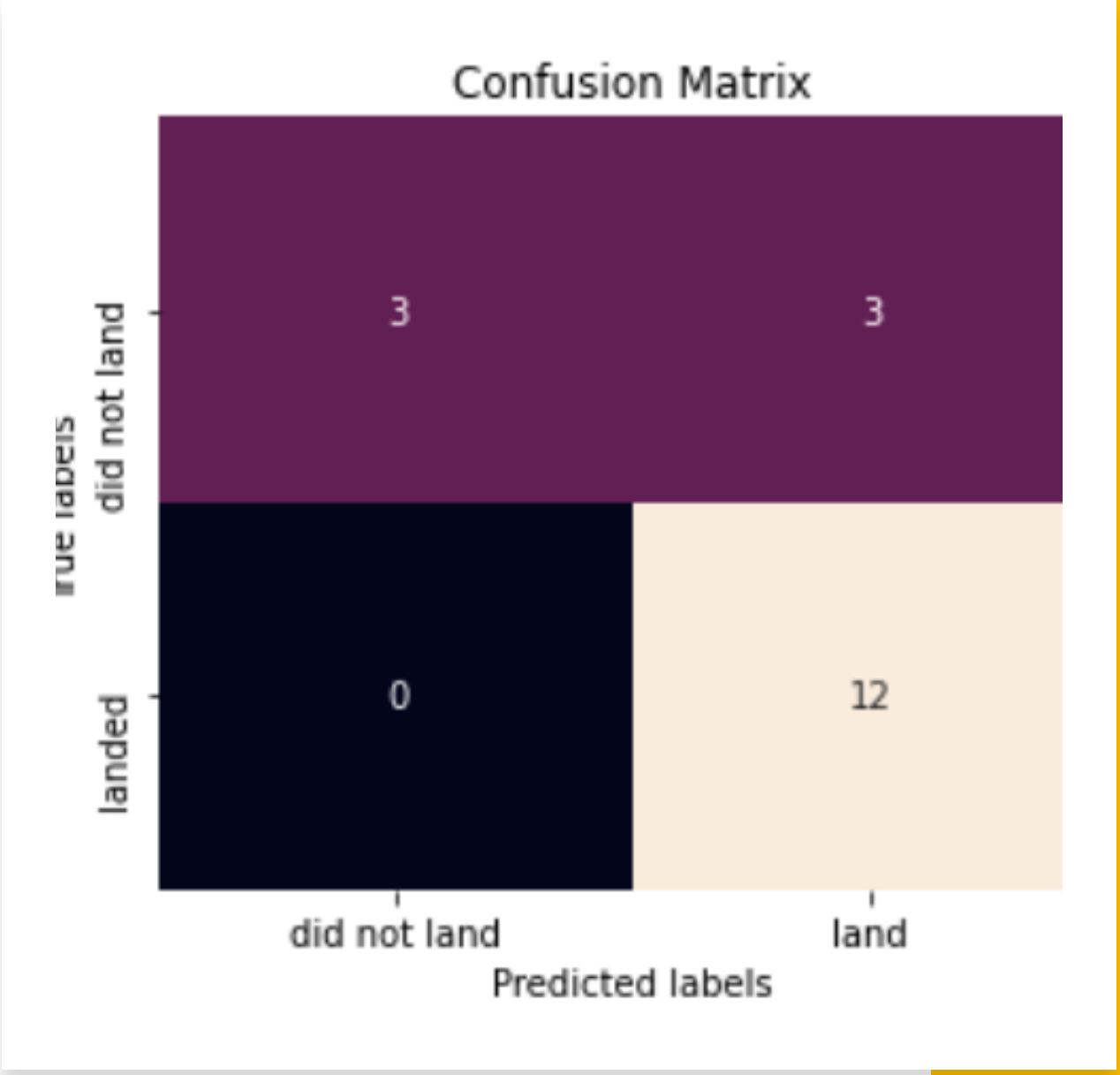
Classification Accuracy



- Out of four classification models, the one that yielded the highest accuracy is Decision Tree, with over 87%

Confusion Matrix

Logistic regression can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- Decision Tree algorithm is the most accurate to use for this dataset.
- Over the years, the success rate of launches increase.
- KSC LC-39A has the highest success rate from all the sites.
- Launches with a low payload mass show better outcomes than launches with a a larger payload mass
- Orbits that has 100% success rates are ES-L1, GEO, HEO and SSO.

Appendix

- https://github.com/Tisisith/Applied-Data-Science-Capstone-Project/blob/bde8fa8d673dd6c611121aacfea7809d10cbf9be/Data%20Collection.ipynb
- https://github.com/Tisisith/Applied-Data-Science-Capstone-Project/blob/bde8fa8d673dd6c611121aacfea7809d10cbf9be/Webscraping.ipynb
- https://github.com/Tisisith/Applied-Data-Science-Capstone-Project/blob/bde8fa8d673dd6c611121aacfea7809d10cbf9be/Webscraping.ipynb
- https://github.com/Tisisith/Applied-Data-Science-Capstone-Project/blob/bde8fa8d673dd6c611121aacfea7809d10cbf9be/EDA%20with%20Visualization.ipynb
- https://github.com/Tisisith/Applied-Data-Science-Capstone-Project/blob/bde8fa8d673dd6c611121aacfea7809d10cbf9be/SQL.ipynb
- https://github.com/Tisisith/Applied-Data-Science-Capstone-Project/blob/bde8fa8d673dd6c611121aacfea7809d10cbf9be/Folium.ipynb
- https://github.com/Tisisith/Applied-Data-Science-Capstone-Project/blob/bde8fa8d673dd6c611121aacfea7809d10cbf9be/Machine%20Learning.ipynb

Thank you!