IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ibrahim Koicha
18th January 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using SpaceX API
  - Data Collection using Web Scrapping
  - Data Wrangling
  - Exploratory Data using SQL
  - Data Visualization using Pandas and Matplotlib
  - Launch Sites Analysis with Folium-Interactive Visual Analytics and Ploty Dash
  - Machine Learning Language Prediction
- Summary of all results
  - EDA Results
  - Interactive Visual Analytics and Dashboard
  - Predictive Analysis

Introduction

- **Project background and context**

SpaceX promotes Falcon 9 rocket launches on its website at a price of 62 million dollars, significantly lower than other providers whose costs exceed 165 million dollars per launch. The key factor contributing to these savings is SpaceX's ability to reuse the first stage. Consequently, if we can predict the successful landing of the first stage, we can estimate the overall cost of a launch. This insight becomes valuable when assessing competitive bids from other companies vying for rocket launch opportunities against SpaceX.

- **Problems you want to find answers**

In this final project, our aim is to forecast the successful landing of the Falcon 9 first stage by analyzing data sourced from the Falcon 9 rocket launches featured on its official website.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

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.

We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

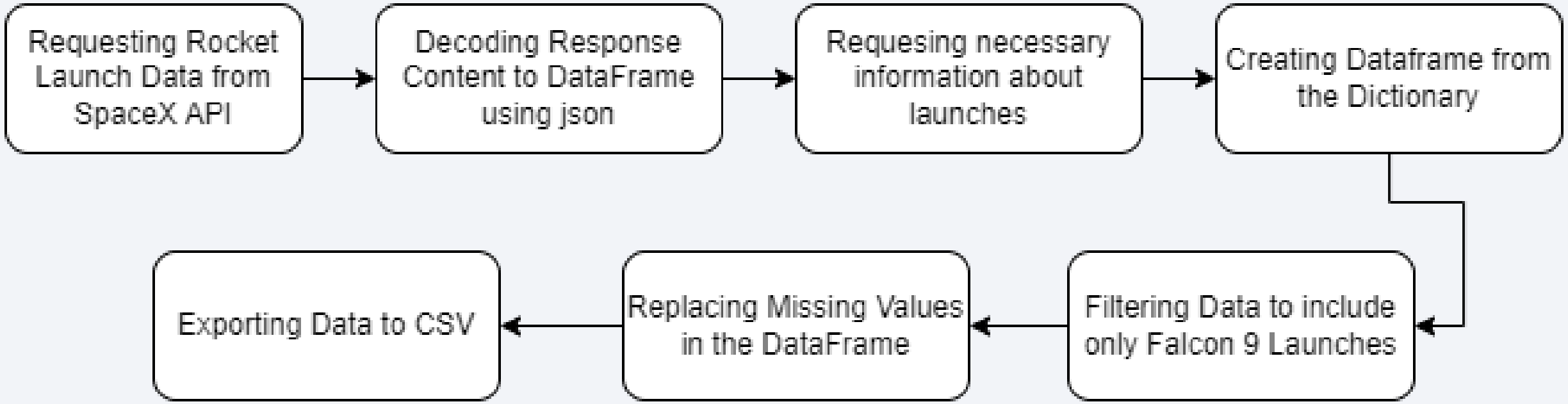
Data Columns are obtained by using **SpaceX REST API**:

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

Data Columns are obtained by using **Wikipedia Web Scraping**:

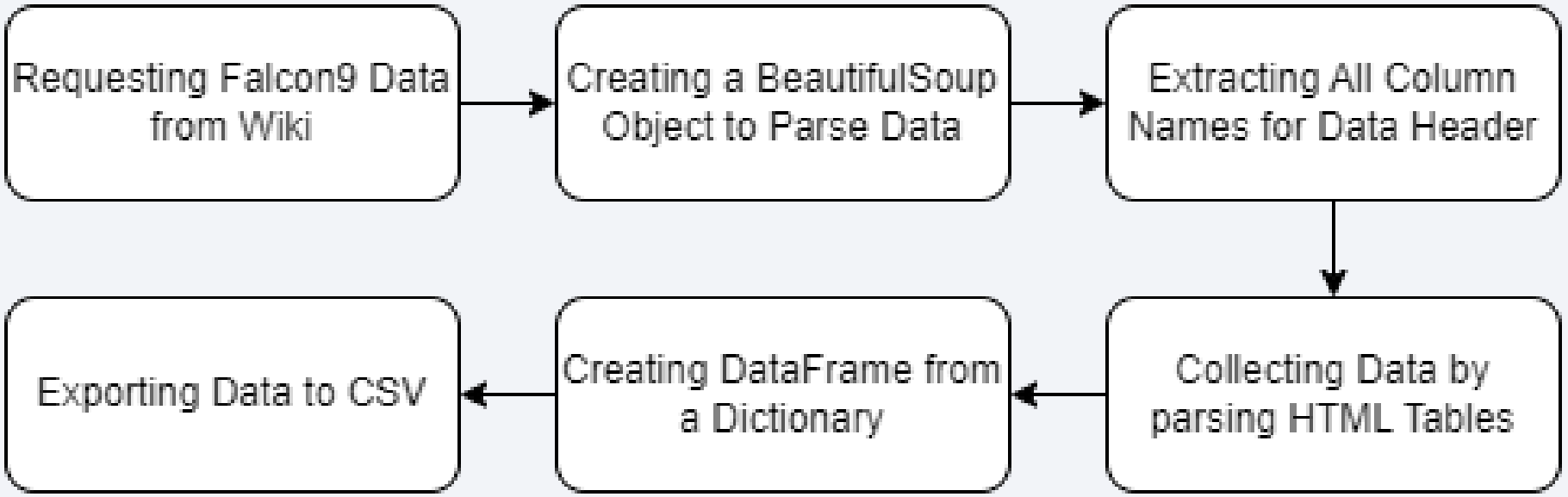
- Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



Data Collection using SpaceX API

Data Collection – SpaceX API

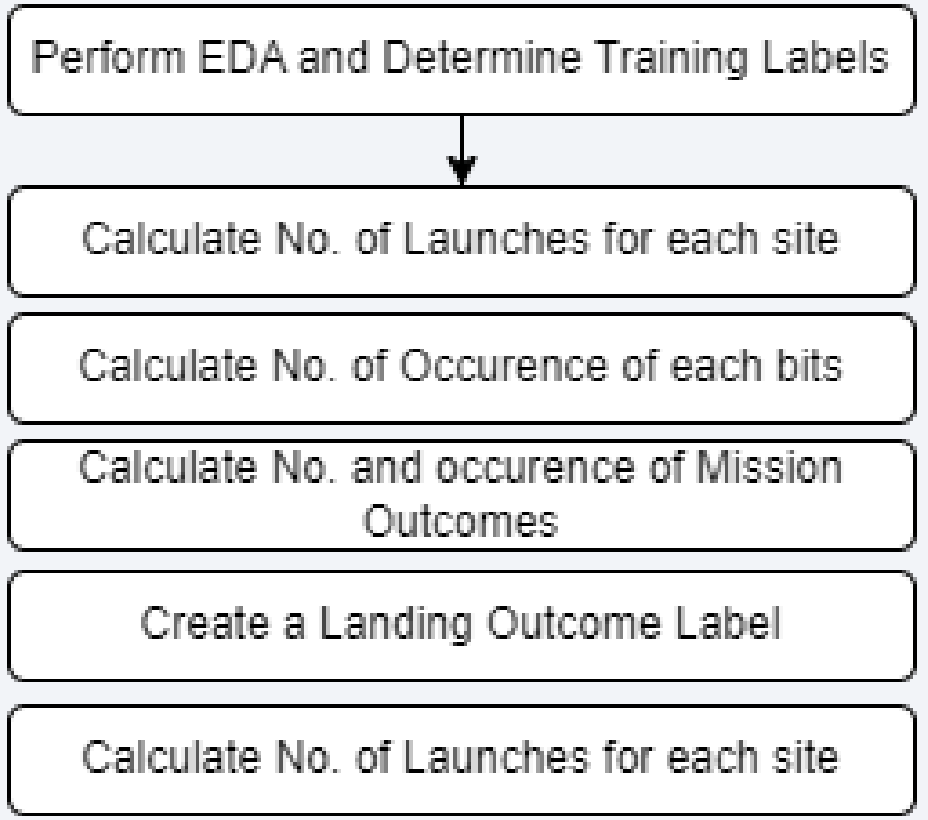


Data Collection using Web Scrapping

Data Wrangling

Within the dataset, various scenarios depict instances where the booster failed to land successfully. This could be attributed to attempted landings resulting in accidents. For instance, a "True Ocean" designation signifies a successful landing in a specific region of the ocean, while "False Ocean" indicates an unsuccessful landing in a designated ocean region. Similarly, "True RTLS" represents a successful ground pad landing, whereas "False RTLS" denotes an unsuccessful attempt at landing on a ground pad. "True ASDS" indicates a successful landing on a drone ship, and "False ASDS" indicates an unsuccessful landing on a drone ship.

To streamline these outcomes, we primarily convert them into training labels, assigning a value of "1" to indicate a successful booster landing and "0" to signify an unsuccessful landing.



Data Wrangling

EDA with Data Visualization

- 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 show the relationship between variables. If a relationship exists, they could be used in machine learning model.
- Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.
- Line charts show trends in data over time (time series).

Data Visualization

EDA with SQL

- Performed SQL queries:
  - 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
  - Listing the failed landing outcomes in drone ship, their booster versions and launch site names 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

EDA Using 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.
- Coloured 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.

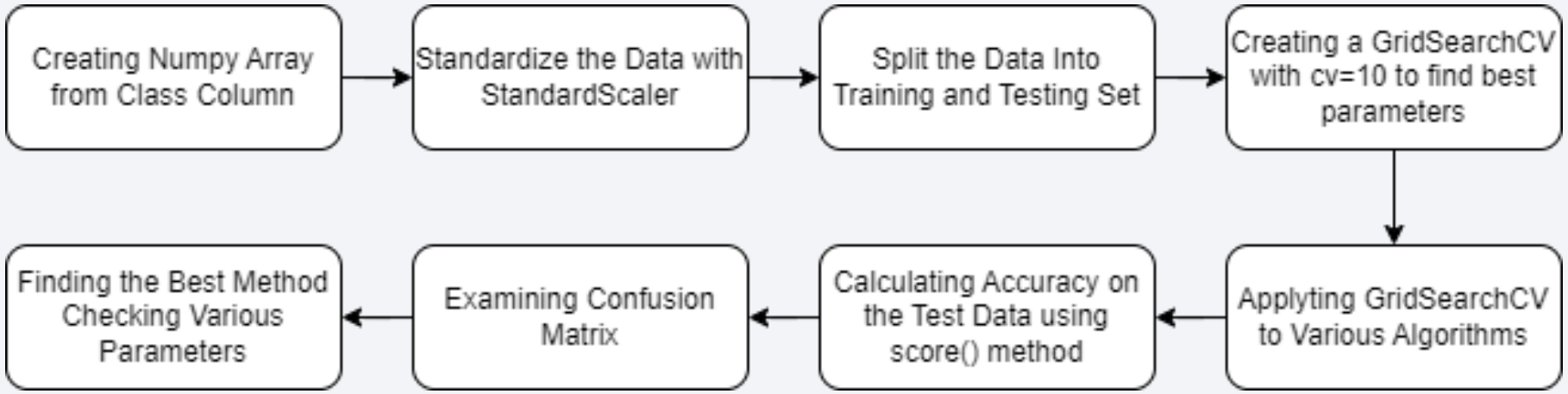
Interactive Map using Folium

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:
  - Added a dropdown list to enable Launch Site selection.
- Pie Chart showing Success Launches (All Sites/Certain Site):
  - 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.
- Slider of Payload Mass Range:
  - Added a slider to select Payload range.
- 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.

Dashboard Using Plotly Dash

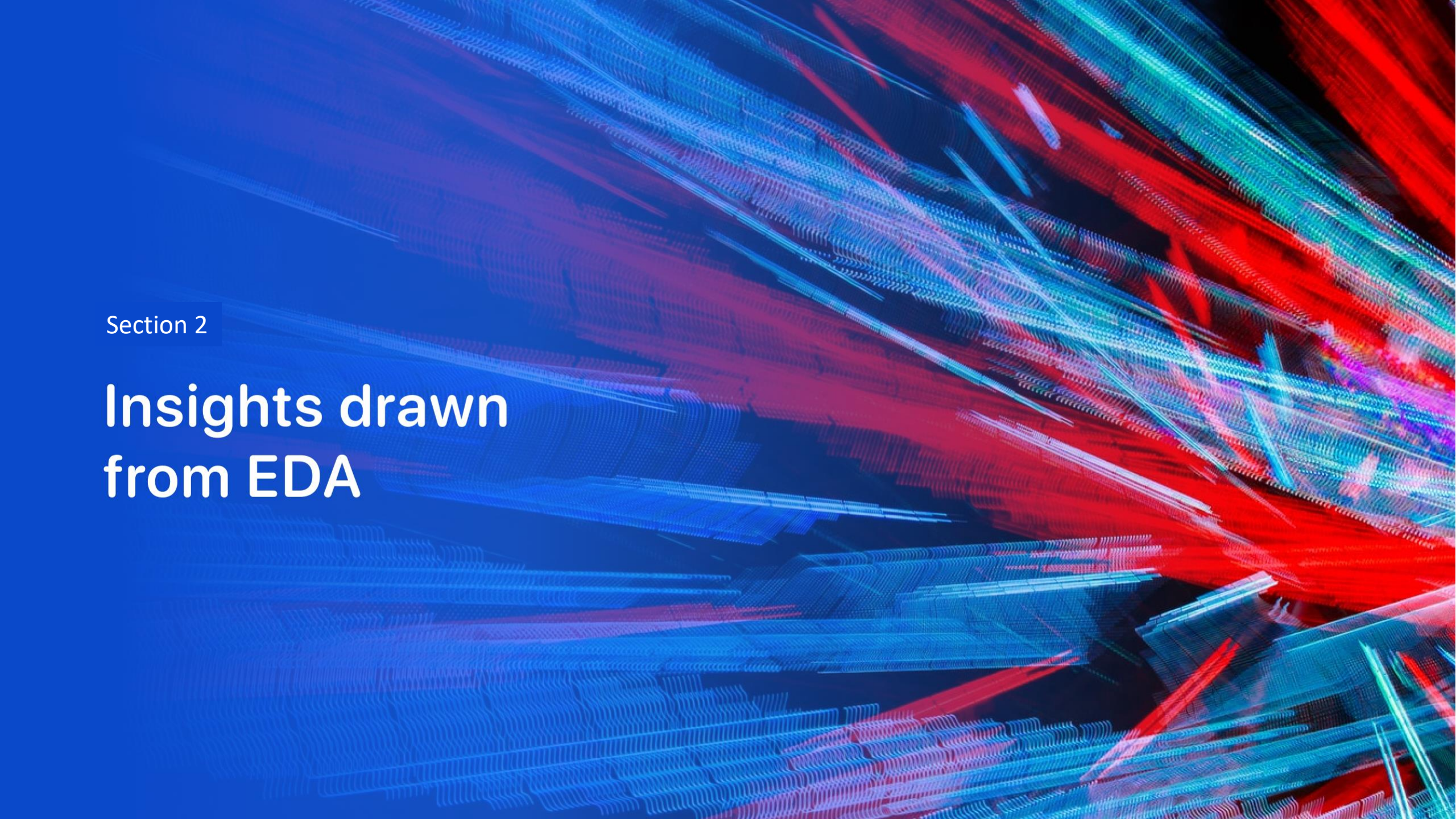
Predictive Analysis (Classification)



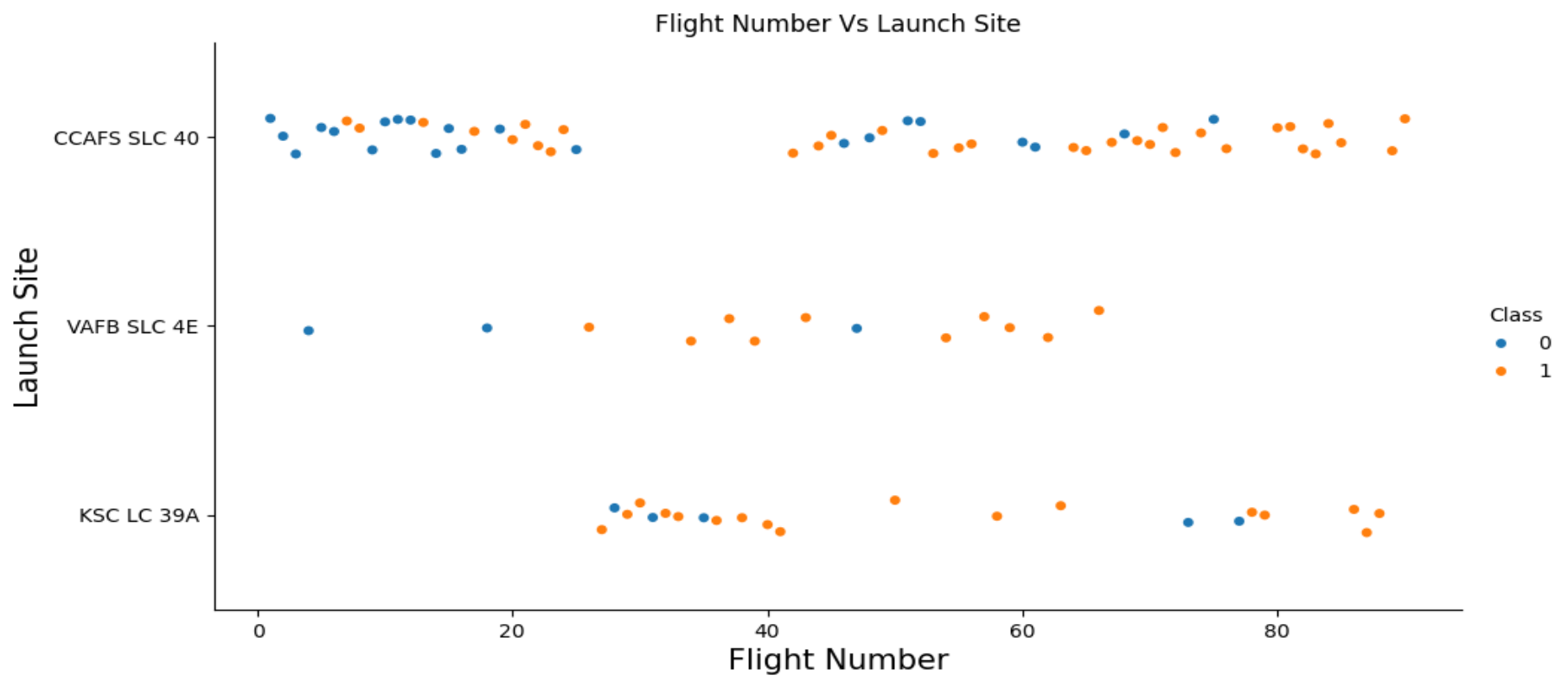
[Predictive Analysis Notebook]

Results

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


Section 2
Insights drawn from EDA

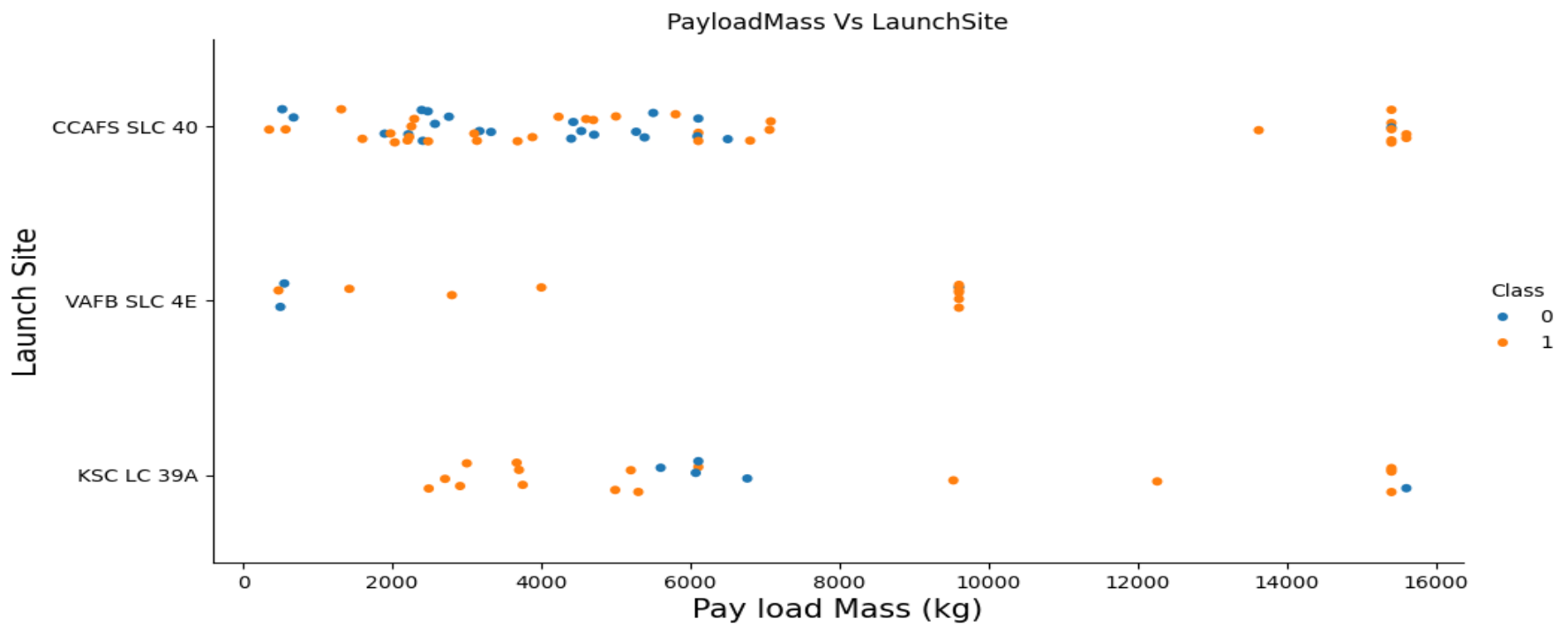
Flight Number vs. Launch Site



Explanation:-

- 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



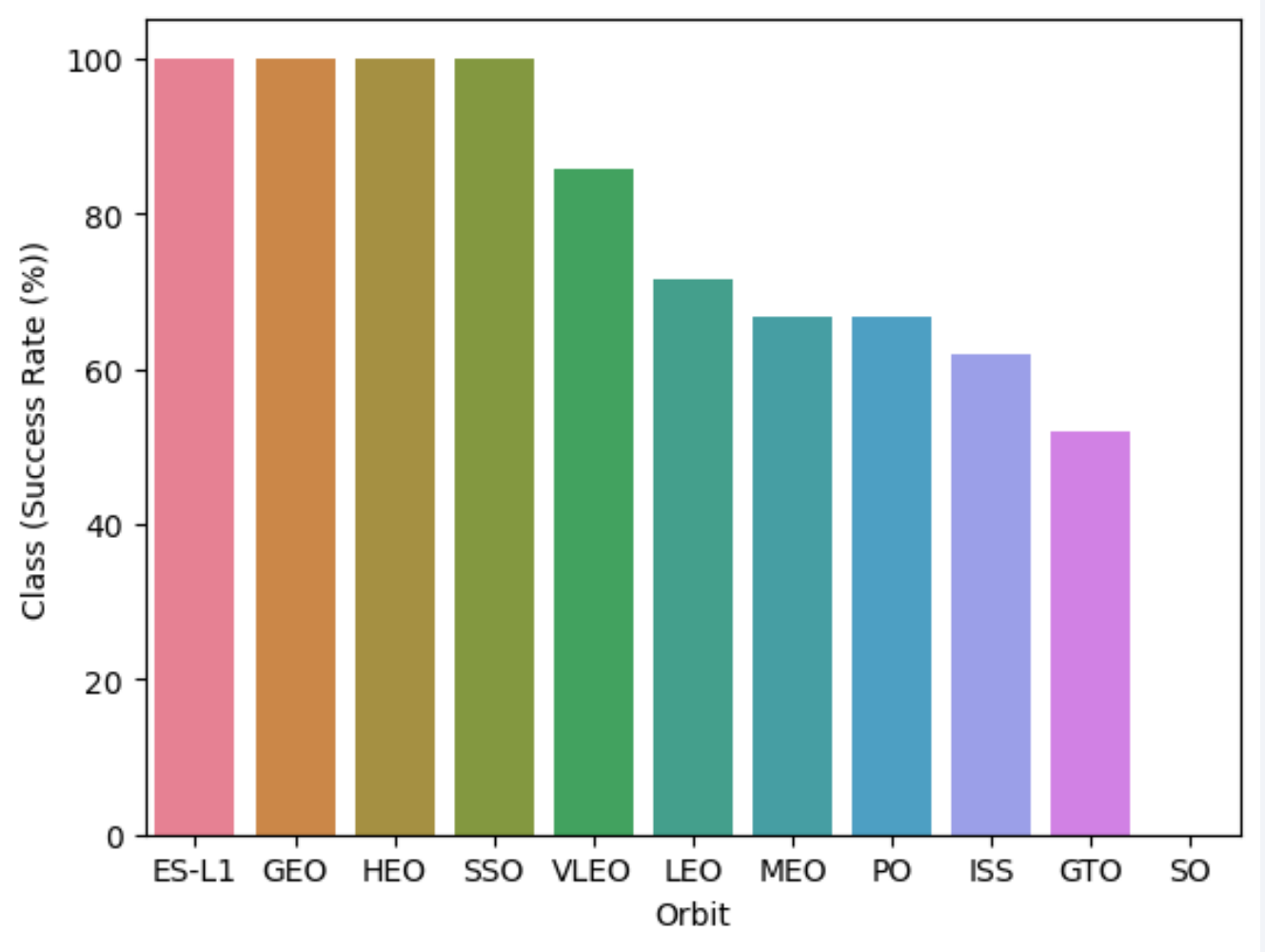
Explanation

- 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

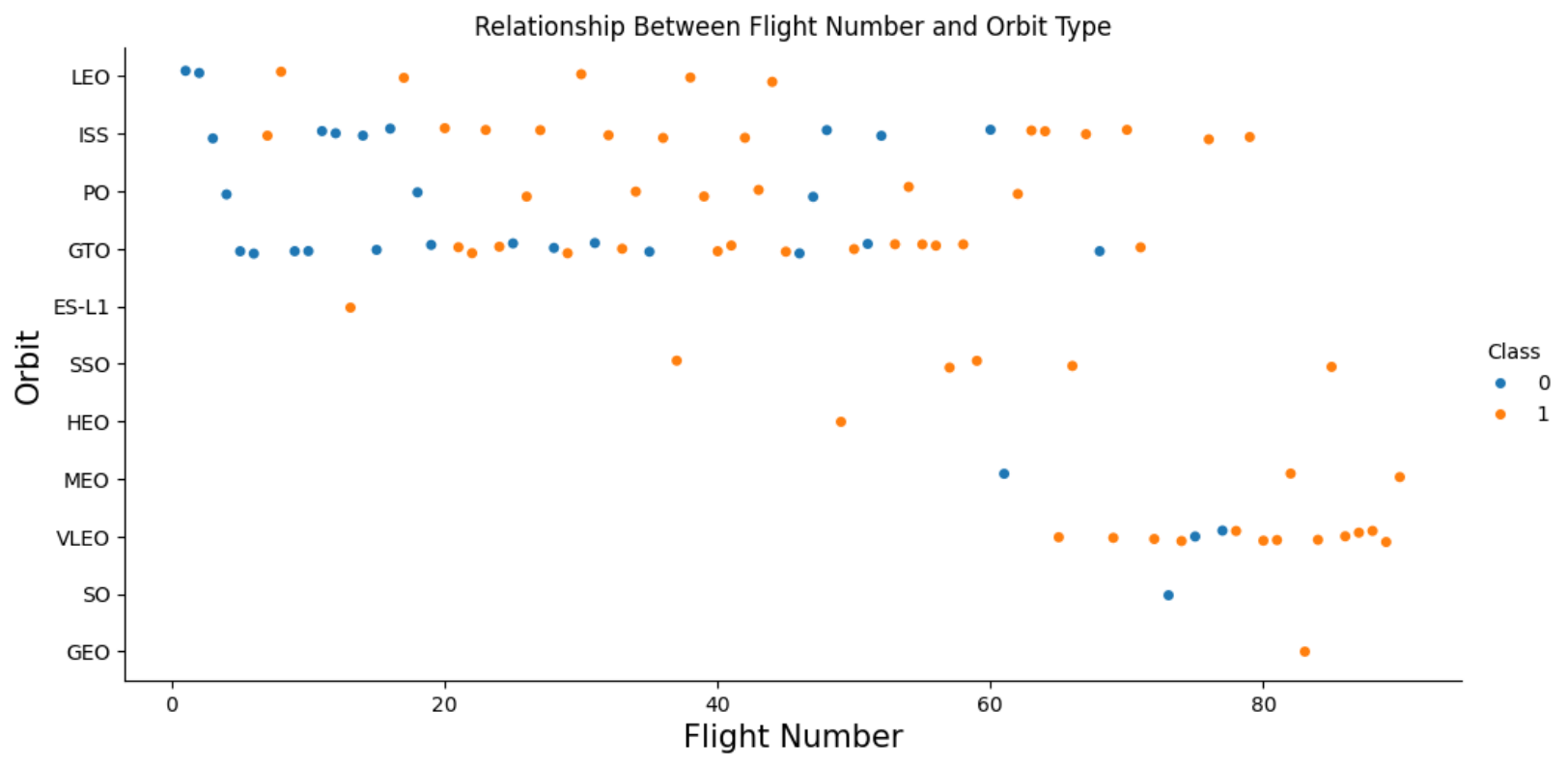
Success Rate vs. Orbit Type

Explanation:

- 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



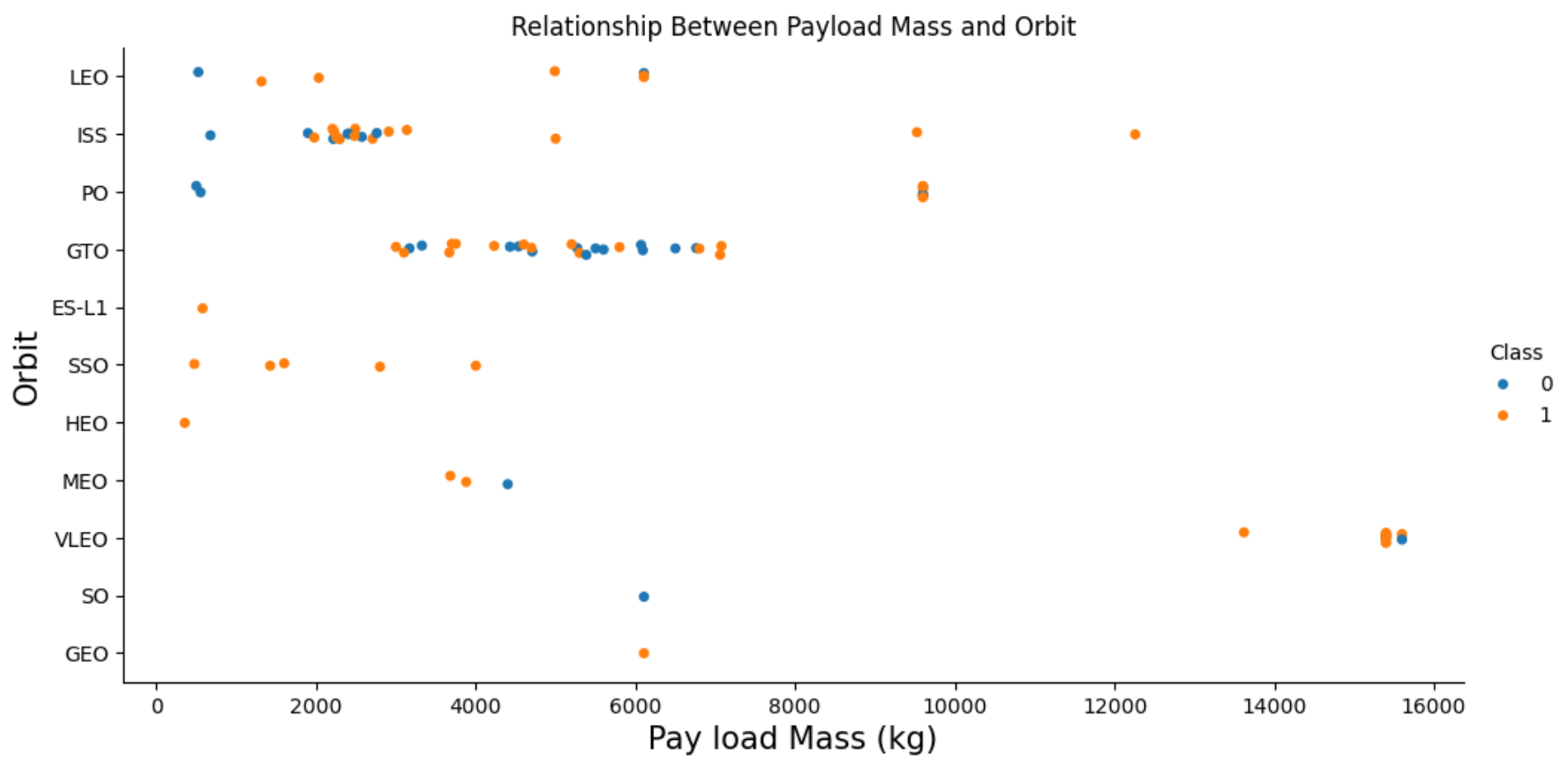
Flight Number vs. Orbit Type



Explanation

- 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 21 in GTO orbit.

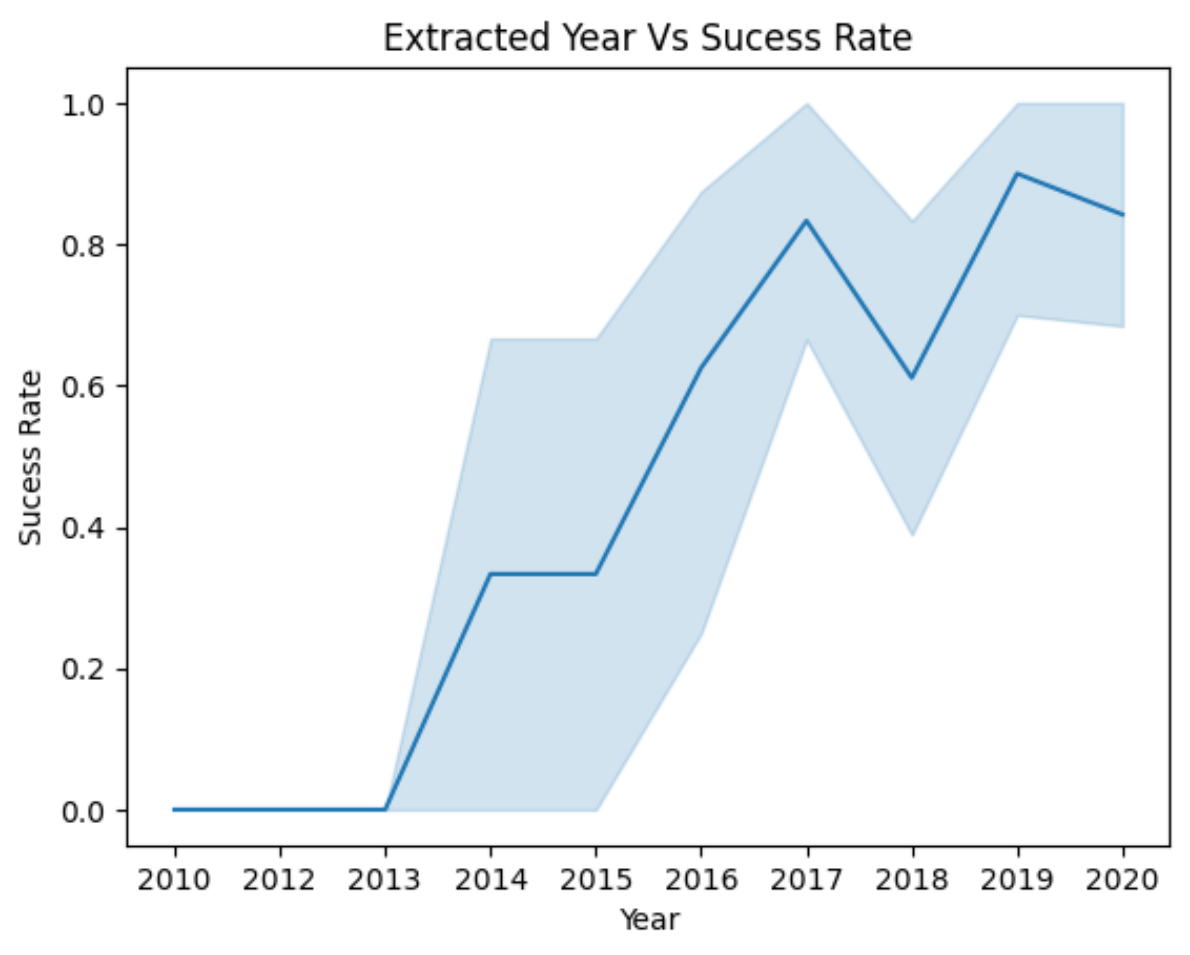
Payload vs. Orbit Type



Explanation:

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

Launch Success Yearly Trend



Explanation:-

- The success rate since 2013 kept increasing till 2020

All Launch Site Names

- Displaying the names of the unique launch sites in the space mission.
- Used 'SELECT DISTINCT' statement to return only the unique launch sites from the 'LAUNCH_SITE' column of the SPACEXTBL table

```
[8]  %sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL
```

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

</>

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

Launch Site Names Begin with 'CCA'

- Displaying 5 records where launch sites begin with the string 'CCA'.
- Used 'LIKE' command with '%' wildcard in 'WHERE' clause to select and display a table of all records where launch sites begin with the string 'CCA'

```
%%sql
SELECT *
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

[9] Python

```
 * 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

Total Payload Mass

- Displaying the total payload mass carried by boosters launched by NASA (CRS).
- Used the 'SUM()' function to return and dispaly the total sum of 'PAYLOAD_MASS_KG' column for Customer 'NASA(CRS)'

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) as "Total_Payload_Mass", Customer
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

[10]

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

Total_Payload_Mass	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

- Displaying average payload mass carried by booster version F9 v1.1
- Used the 'AVG()' function to return and dispaly the average payload mass carried by booster version F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS "Average_Payload_Mass", Customer, Booster_Version
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.1%';
```

[11]

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

Average_Payload_Mass	Customer	Booster_Version
2534.6666666666665	MDA	F9 v1.1 B1003

First Successful Ground Landing Date

- Listing the date when the first successful landing outcome in ground pad was achieved.
- Used the 'MIN()' function to return and display the first (oldest) date when first successful landing outcome on ground pad 'Success (ground pad)'happened.

```
%%sql
SELECT MIN(DATE) AS "First_Successful_Landing"
FROM SPACEXTBL
WHERE Landing_Outcome = "Success (ground pad)";
```

[12]

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

First_Successful_Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Used 'Select Distinct' statement to return and list the 'unique' names of boosters with 4000 and 6000 Limits with Landing Outcome of "Success (Drone ship)"

[29]

```
%%sql
SELECT DISTINCT Booster_Version, Payload
FROM SPACEXTBL
WHERE Landing_Outcome = "Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

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

Booster_Version	Payload
F9 FT B1022	JCSAT-14
F9 FT B1026	JCSAT-16
F9 FT B1021.2	SES-10
F9 FT B1031.2	SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- Listing the total number of successful and failure mission outcomes
- Used the 'COUNT()' together with the 'GROUP BY' statement to return total number of missions outcomes

[31]

```
%%sql
SELECT MISSION_OUTCOME, COUNT(*) AS "MISSION OUTCOMES DETAILS"
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

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

Mission_Outcome	MISSION OUTCOMES DETAILS
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Listing the names of the booster versions which have carried the maximum payload mass
- Using a Subquerry to return and pass the Max payload and used it list all the boosters that have carried the Max payload of 15600kgs

```
%%sql
SELECT DISTINCT Booster_Version, Payload, PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL
    );
```

[33]

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

Booster_Version	Payload	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Starlink 1 v1.0, SpaceX CRS-19	15600
F9 B5 B1049.4	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600
F9 B5 B1051.3	Starlink 3 v1.0, Starlink 4 v1.0	15600
F9 B5 B1056.4	Starlink 4 v1.0, SpaceX CRS-20	15600
F9 B5 B1048.5	Starlink 5 v1.0, Starlink 6 v1.0	15600
F9 B5 B1051.4	Starlink 6 v1.0, Crew Dragon Demo-2	15600
F9 B5 B1049.5	Starlink 7 v1.0, Starlink 8 v1.0	15600
F9 B5 B1060.2	Starlink 11 v1.0, Starlink 12 v1.0	15600
F9 B5 B1058.3	Starlink 12 v1.0, Starlink 13 v1.0	15600
F9 B5 B1051.6	Starlink 13 v1.0, Starlink 14 v1.0	15600
F9 B5 B1060.3	Starlink 14 v1.0, GPS III-04	15600
F9 B5 B1049.7	Starlink 15 v1.0, SpaceX CRS-21	15600

2015 Launch Records

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.
- Used the 'substr()' in the select statement to get the month and year from the date column where substr(Date,7,4)='2015' for year and Landing outcome was 'Failure (drone ship') and return the records matching the filter. Check substr(Date,7,4)='2015' for year and Landing outcome was 'Failure (drone ship') and return the records matching the filter.

[43]

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

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

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

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

- 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.

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

[49]

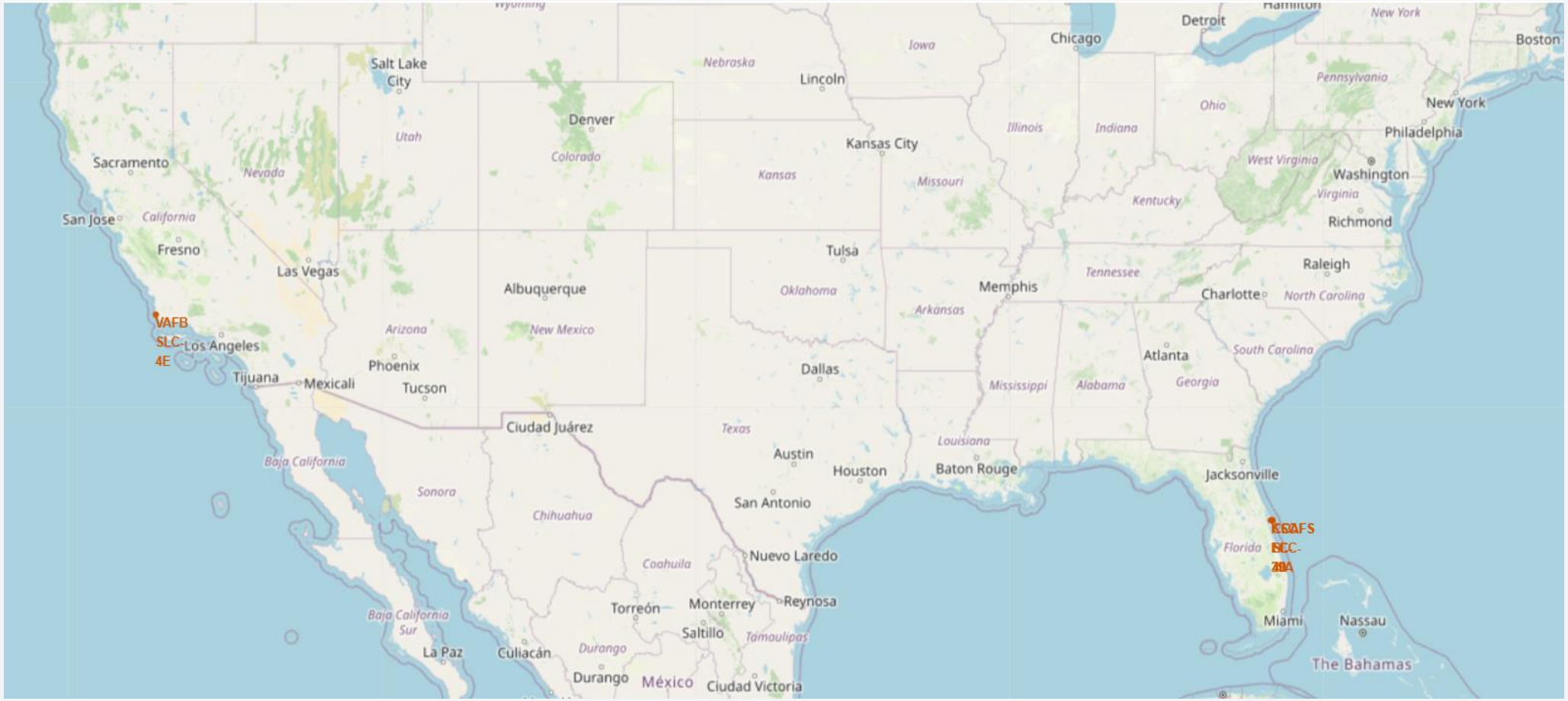
```
 * sqlite:///my_data1.db
Done.
```

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

Section 3

Launch Sites Proximities Analysis

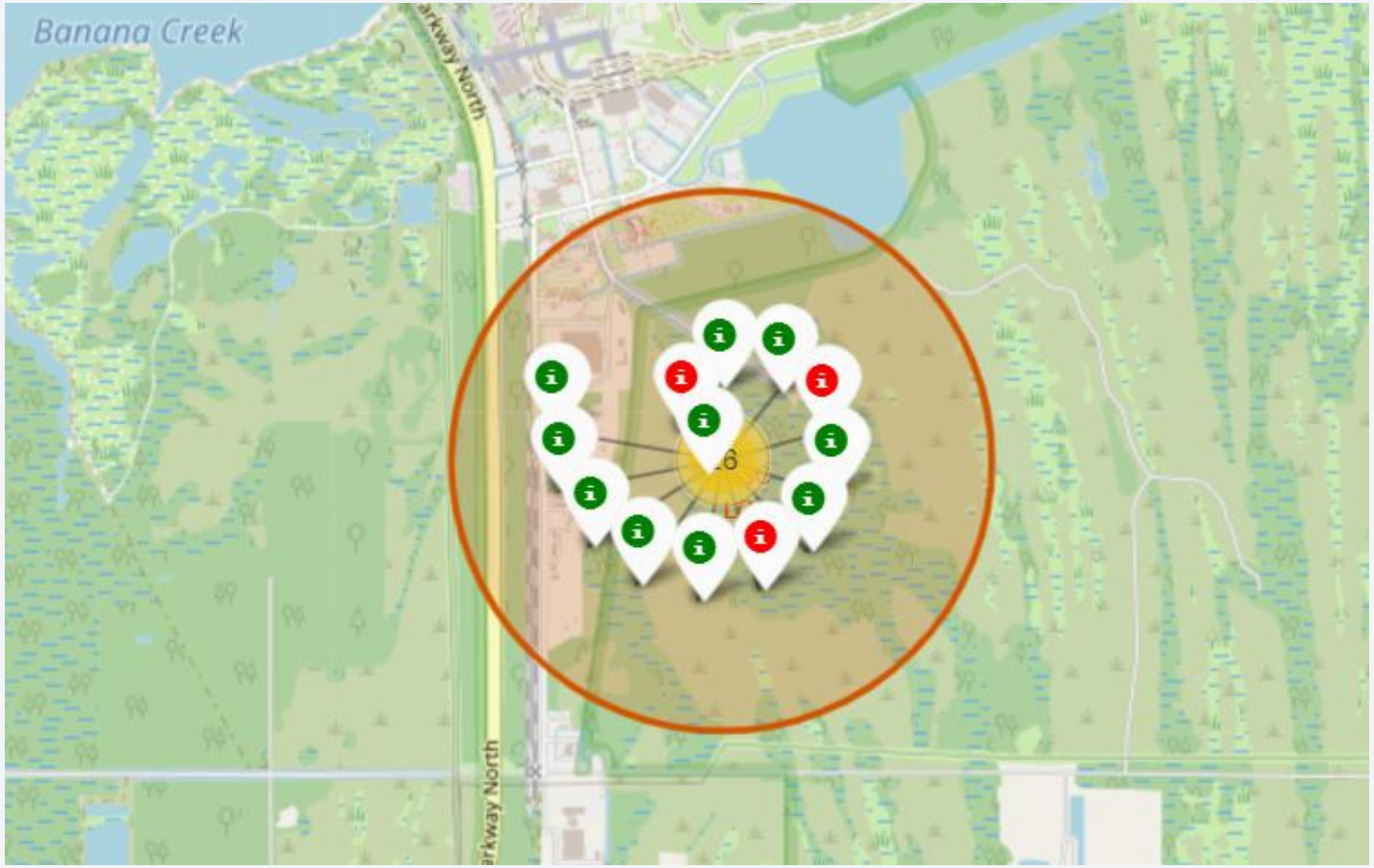
All launch sites' location markers on a global map



All launch sites' location markers on a global map

- Launch sites are strategically located near the Equator for optimal efficiency. Objects at the Earth's Equator are already traveling at a speed of 1670 km/hour due to the planet's rotation. When a spacecraft is launched from the equator, it ascends into space while retaining the Earth's rotational speed through inertia. This velocity aids the spacecraft in maintaining the necessary speed for orbital stability.
- Additionally, all launch sites are situated near coastlines. Launching rockets towards the ocean serves to minimize the potential risk of debris falling or exploding in areas inhabited by people.

Launch outcomes for each site on the map With Color Markers

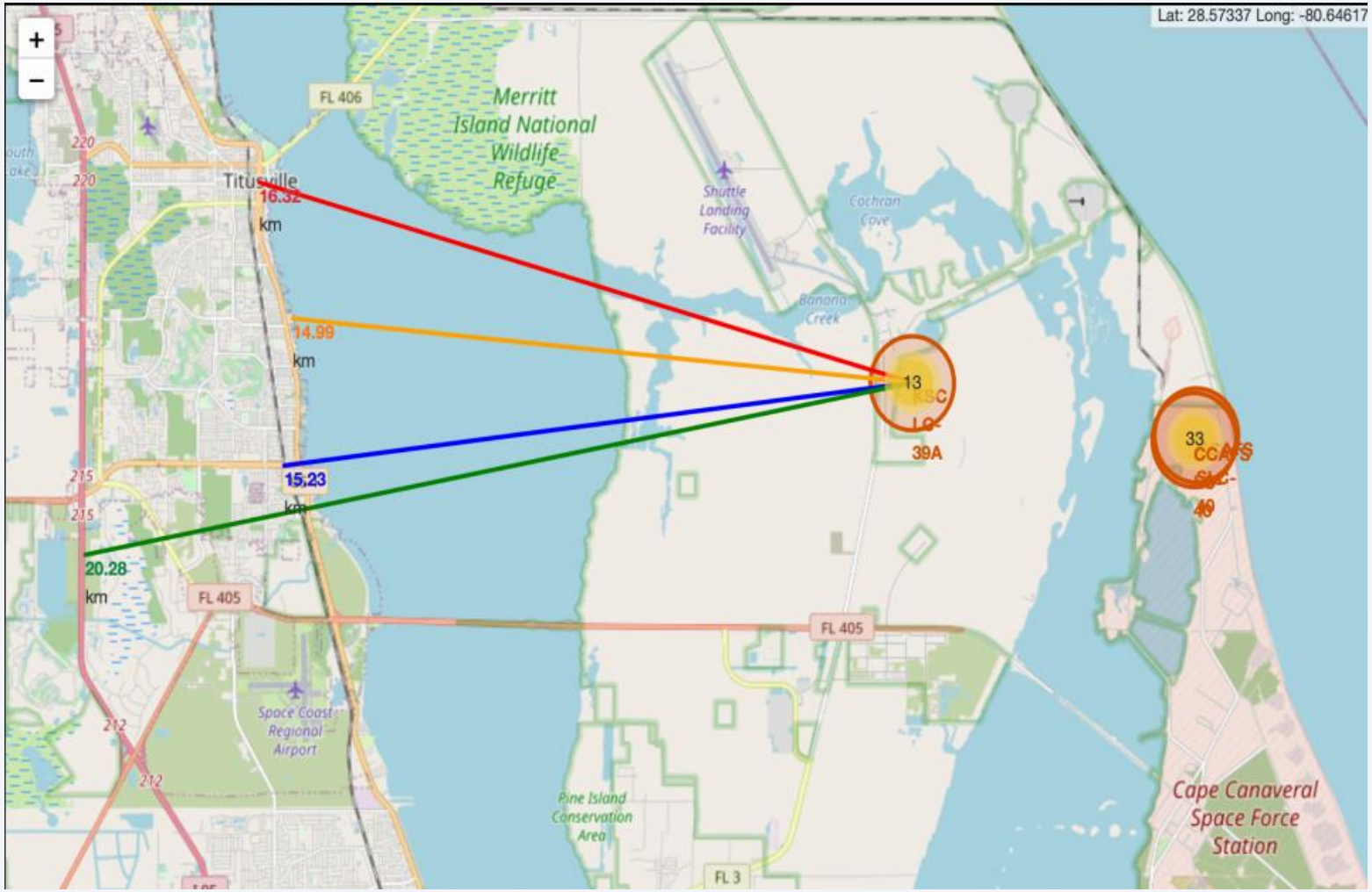


Green Marker = Successful Launch

Red Marker = Failed Launch

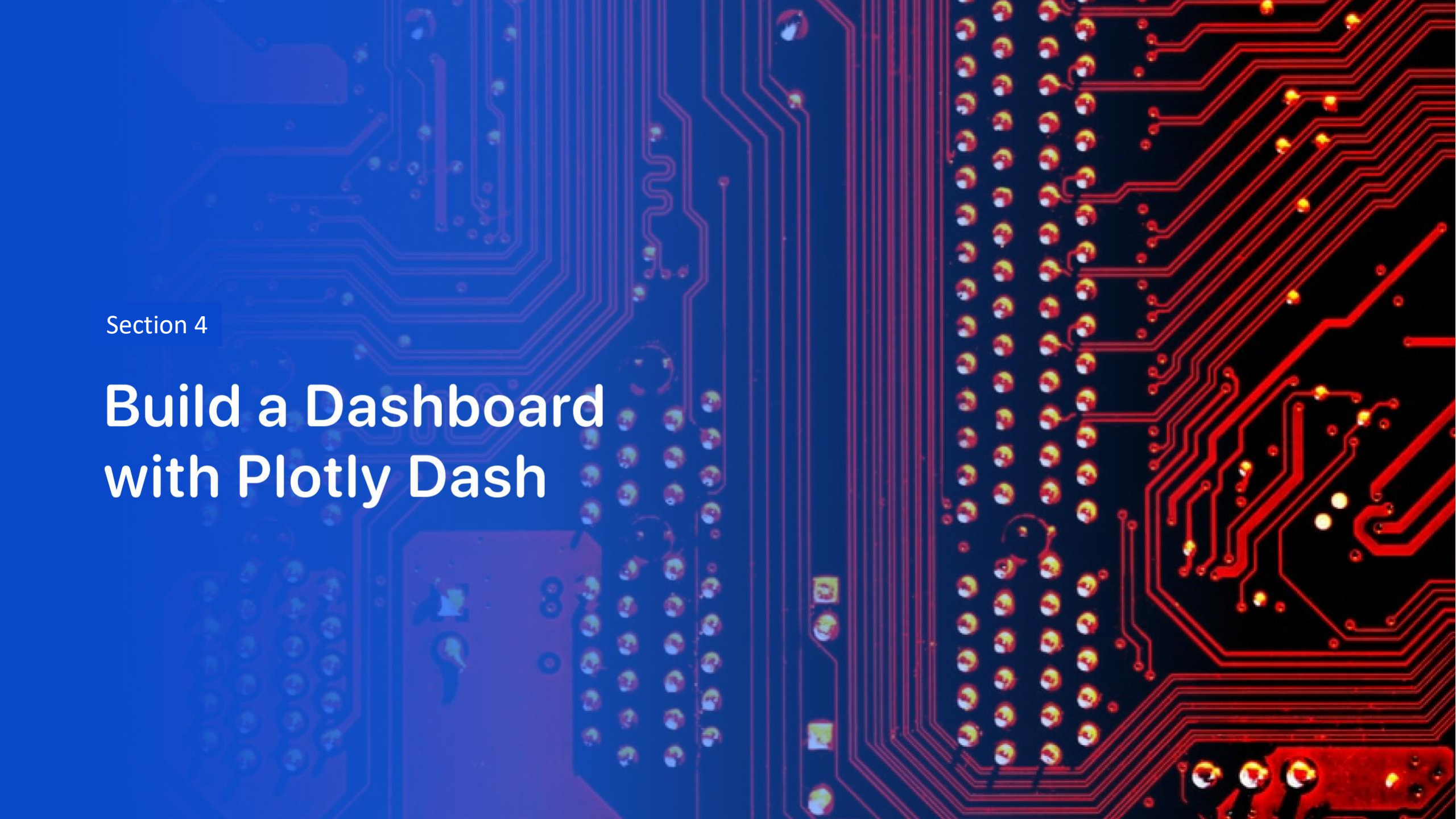
- Launch Site KSC LC-39A has a very high Success Rate.

Distance from the launch site KSC LC-39A to its proximities

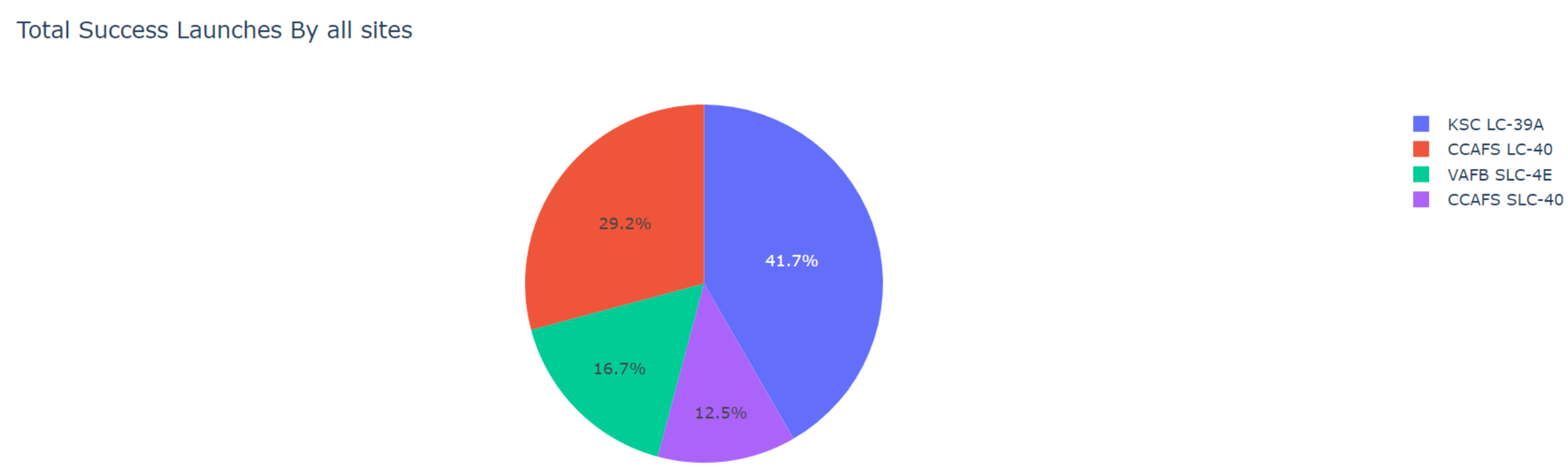


Distance from the launch site KSC LC-39A to its proximities

- From the visual analysis of the launch site KSC LC-39A we can clearly see that the location 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 relatively close to its closest city Titusville (16.32 km).
- Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.

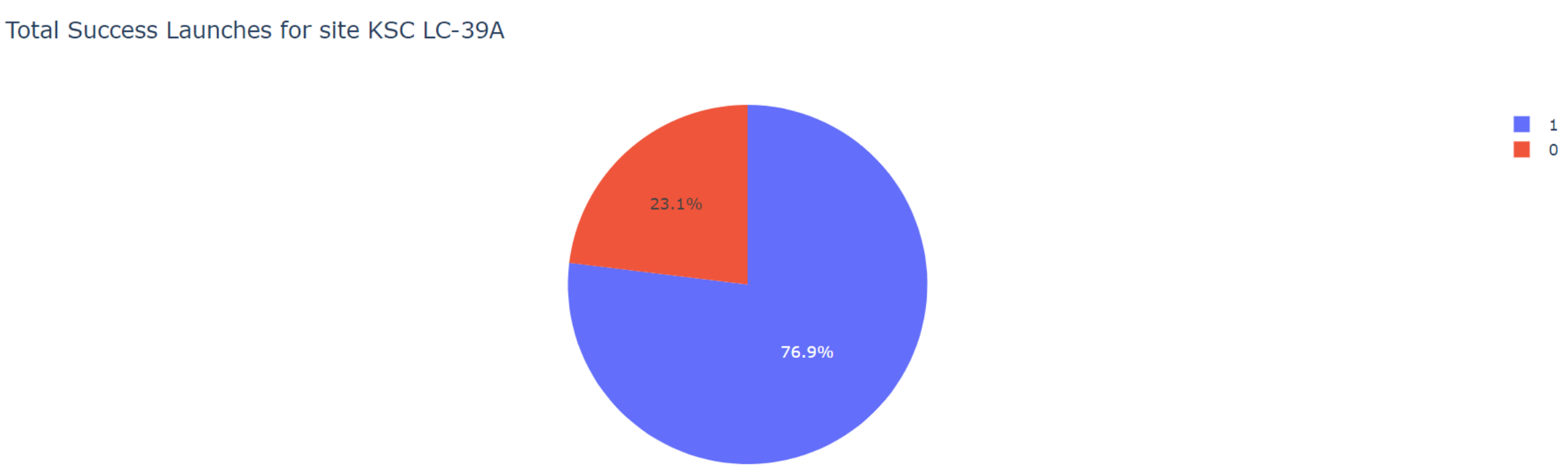

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



- Launch site KSC LC-39A has the highest launch success rate at 42% followed by CCAFS LC-40 at 29%, VAFB SLC-4E at 17% and lastly launch site CCAFS SLC-40 with a success rate of 13%

Launch site with highest Launch success ratio



- KSC LC-39A has the highest launch success rate (76.9%).

Payload Mass vs. Launch Outcome for all sites

- The charts show that payloads between 2000 and 5500 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

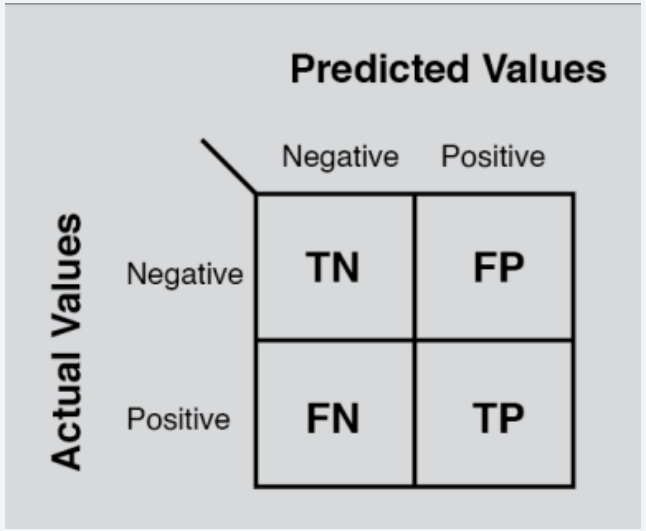
Classification Accuracy

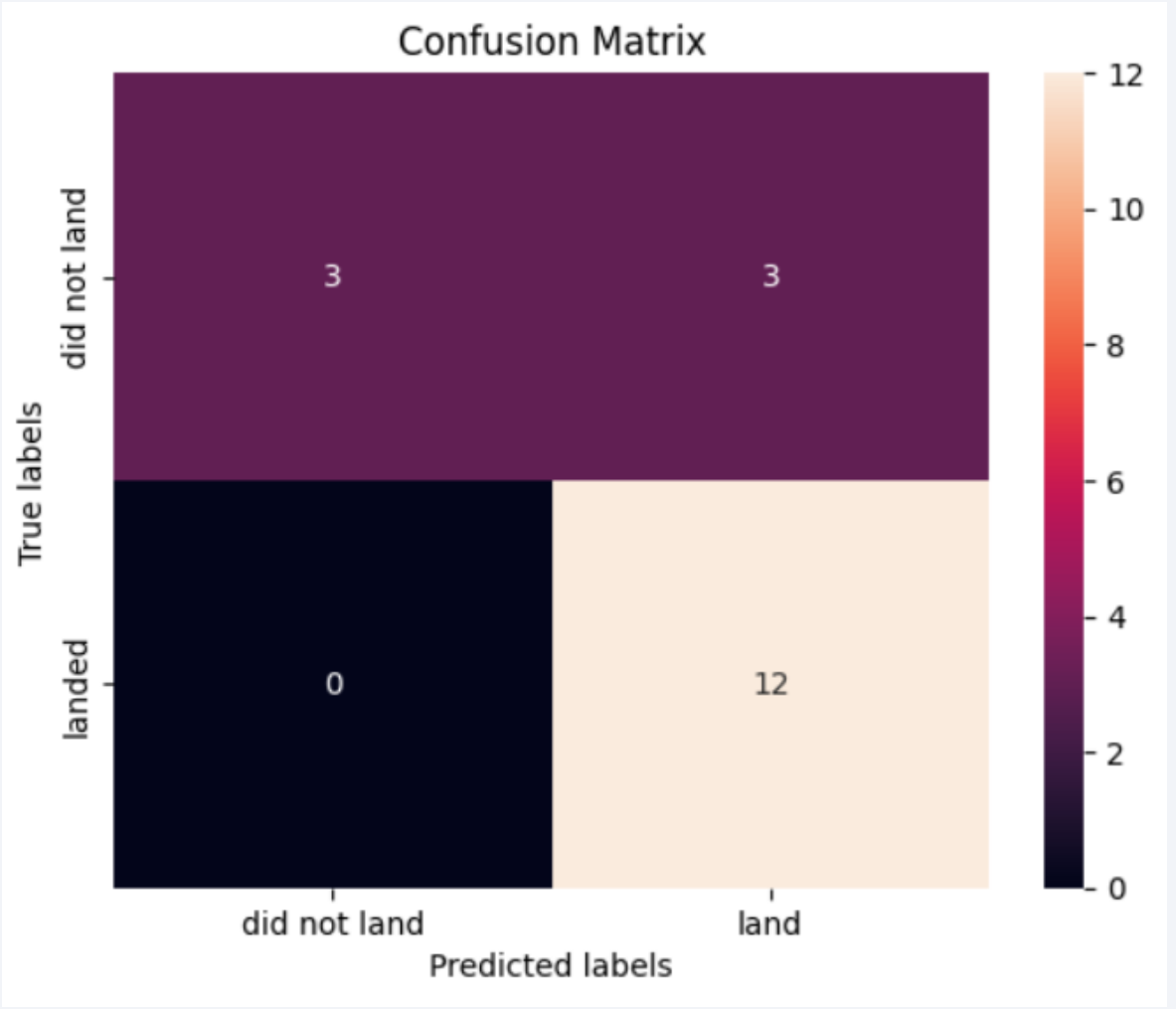
- All Models perform Equally on the Test Data i.e. All have the same Accuracy

Method	Logistic_Reg	SVM	Decision Tree	KNN
Test Data Accuracy	0.833333	0.833333	0.833333	0.833333

Confusion Matrix

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





Conclusions

- The Decision Tree Model is identified as the most suitable algorithm for this dataset.
- Launches with lower payload masses demonstrate superior results compared to those with larger payload masses.
- Most launch sites are positioned near the Equator line, and all sites are in close proximity to coastlines.
- There is an observed upward trend in the success rate of launches over the years.
- KSC LC-39A is distinguished by having the highest success rate among all launch sites.
- Orbits ES-L1, GEO, HEO, and SSO exhibit a perfect 100% success rate.
- In the case of GTO, distinguishing between positive and negative landing rates (successful and unsuccessful missions) is challenging, as both outcomes are present.
- Notably, the success rate has shown a consistent increase from 2013 until 2020

Thank you!

Appendix

- Link to the Datasets can be found in this Link