IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Neha M Dharwad
26/06/2025

Outline

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

Executive Summary

Summary of methodologies:

1. **Data Collection**: Via SpaceX API and web scraping (Python, BeautifulSoup)
2. **Data Wrangling**: Handled missing values, formatted fields, encoded categories
3. **EDA**: With Data Visualization and SQL
4. Building an Interactive map with **Folium**
5. Building a dashboard with **Plotly**
6. **Predictive Analysis**

Summary of all results

1. Success rates varied by **launch site** and **booster version**
2. **Payload mass** showed a moderate impact on success
3. **Decision Tree and SVM** models performed best, with ~85–90% accuracy
4. SQL revealed top-performing sites and boosters

Introduction

Project background and context:

SpaceX, a leader in commercial spaceflight, has conducted numerous rocket launches with varying degrees of success. Understanding the factors that contribute to a successful launch is critical for improving mission reliability and reducing costs. With rich historical launch data publicly available, this project leverages data science to gain insights and build predictive models for launch outcomes.

Problems you want to find answers

1. Which launch sites have the highest success rates?
2. How does payload mass affect mission success?
3. Do booster types influence launch outcomes?
4. Can we predict the success of a future launch based on available features?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web scraping from Wikipedia
- Perform data wrangling
  - Filter data, handle missing values and use One hot encoding to prepare data
- 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 best results.

Data Collection

The data used for this project was collected from **two primary sources**:

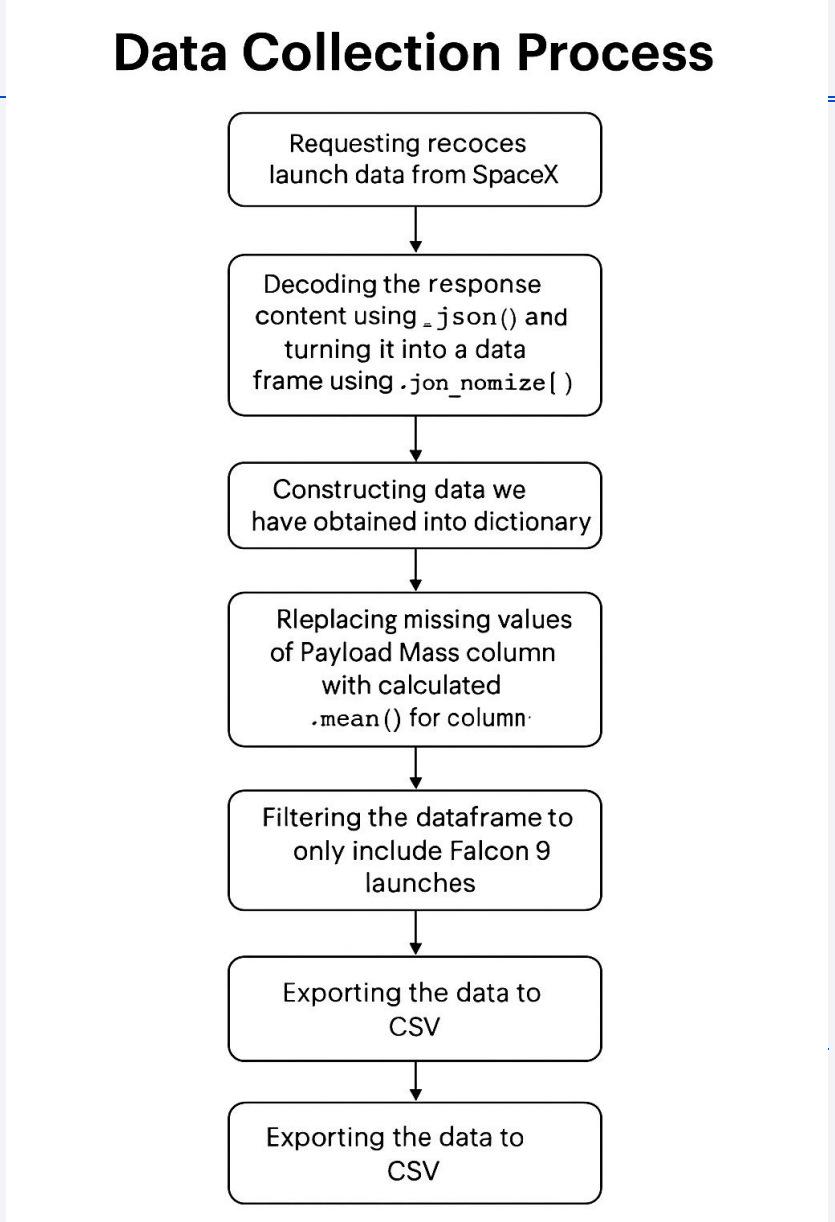
1) **SpaceX REST API**
    - Accessed launch history data including rocket configurations, launch sites, payloads, and mission outcomes.
    - Data was retrieved in JSON format using Python's `requests` library.
    - Nested JSON structures were flattened using `pandas.json_normalize()`.

2) **Web Scraping (Supplementary Data)**
    - Used `BeautifulSoup` to extract additional metadata such as payload details and booster reuse info from structured HTML pages (e.g., Wikipedia).

After collection, the raw data was cleaned and filtered to focus on **Falcon 9** missions only.

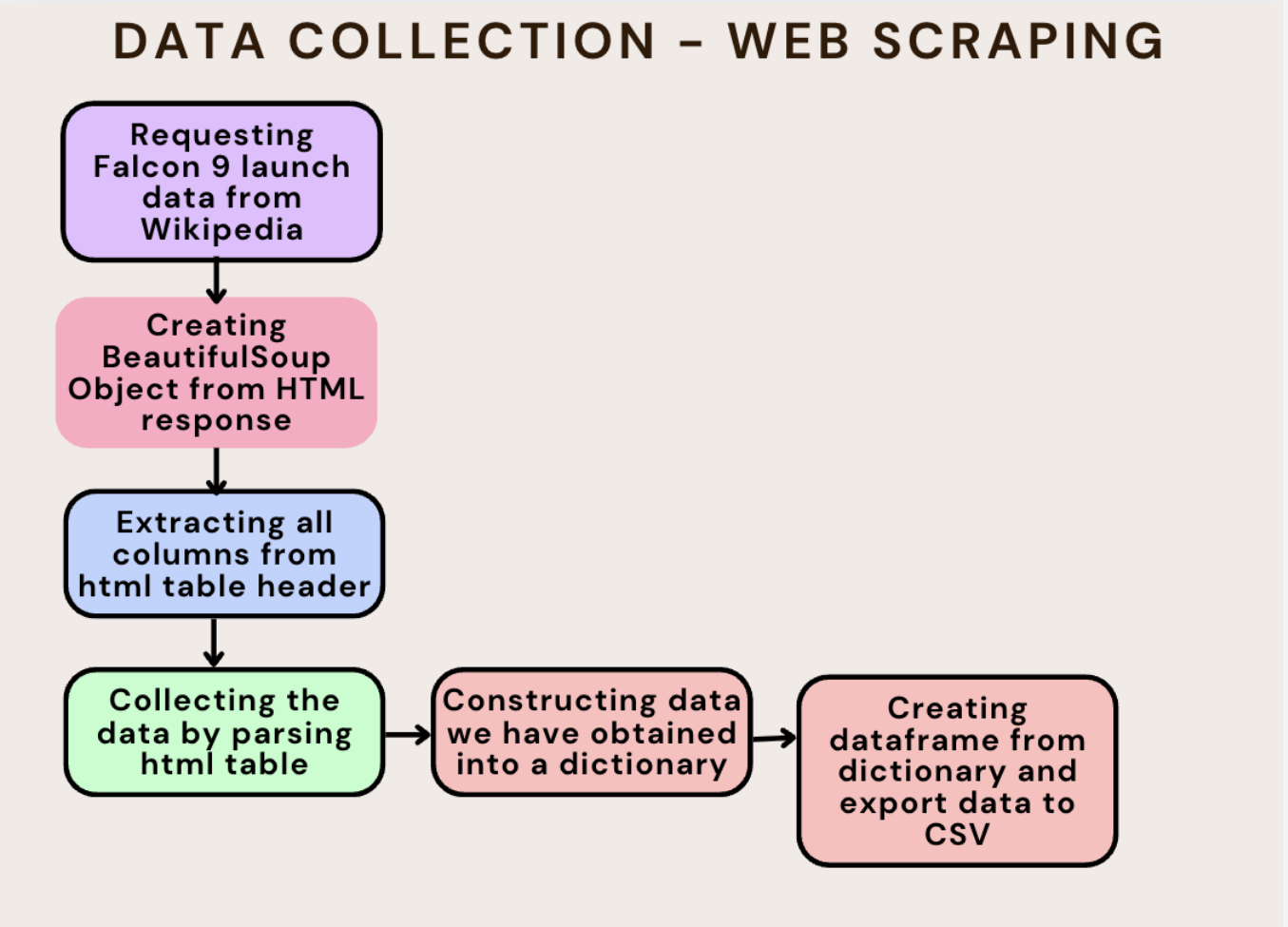
Data Collection – SpaceX API

GitHub Link: [Data Collection - SpaceX API]



Data Collection - Scraping

GitHub Link: Data collection - Scraping



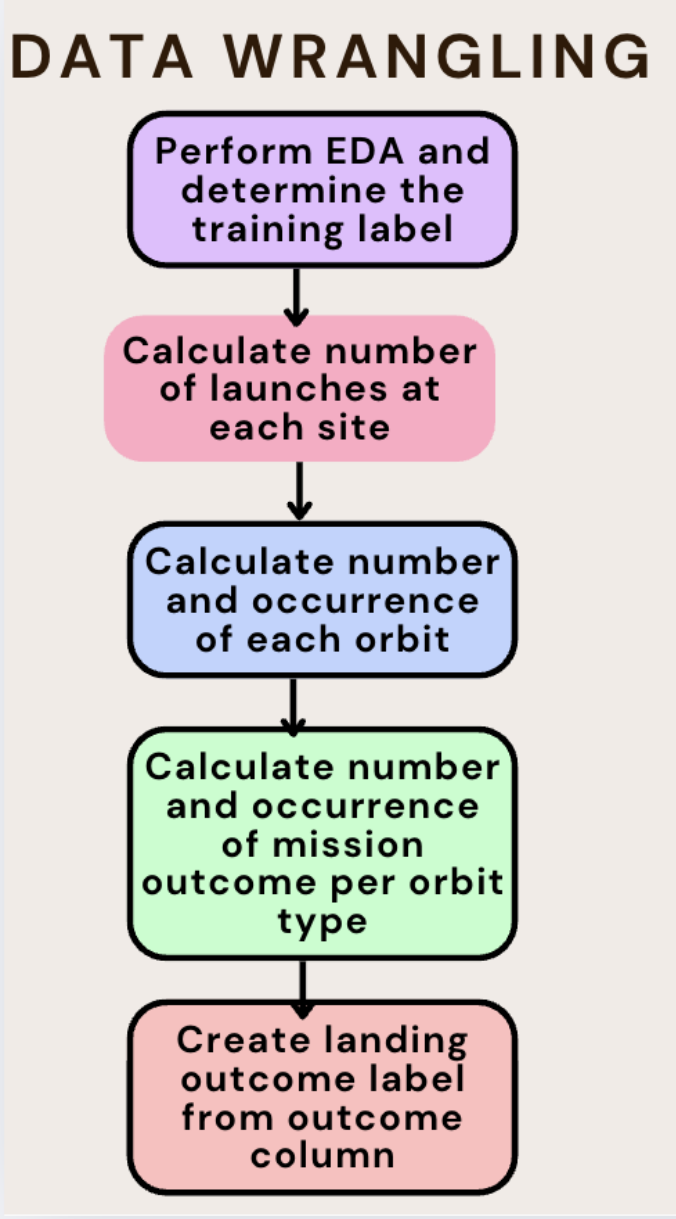
Data Wrangling

GitHub Link: Data Wrangling

The dataset includes multiple scenarios where the Falcon 9 booster did not land successfully. In some cases, a landing attempt was made but failed due to issues such as accidents. For example:

True Ocean indicates a successful landing in a designated ocean zone, whereas
False Ocean refers to an unsuccessful ocean landing attempt.
True RTLS (Return To Launch Site) means the booster successfully landed back on the ground pad,while **False RTLS** indicates a failed landing attempt at the same site.
True ASDS (Autonomous Spaceport Drone Ship) refers to a successful landing on a drone ship,while **False ASDS** signifies an unsuccessful landing attempt on a drone ship.

To simplify analysis, these various outcomes were converted into binary training labels:
"1" indicates a successful booster landing, and **"0"** represents a failure.



EDA with Data Visualization

GitHub Link: EDA with Data Visualization

A variety of charts were plotted to explore relationships in the data, including:

- 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
- Yearly Trend of Success Rate

Chart Interpretations

- Scatter Plots
  Used to reveal potential relationships between numerical variables. If a clear trend is observed, these variables may be useful features in a machine learning model.
- Bar Charts
  Help compare discrete categories. They highlight how different categorical features (like orbit type or launch site) relate to outcomes such as success rates.
- Line Charts
  Visualize trends over time, particularly helpful for analyzing progress or performance year by year (e.g., success rate over time).

EDA with SQL

GitHub Link: EDA with SQL

- Retrieved the names of all unique launch sites used in missions.
- Displayed 5 records where launch site names start with **'CCA'**.
- Calculated the **total payload mass** delivered by boosters under NASA's CRS missions.
- Computed the **average payload mass** carried by the **F9 v1.1** booster version.
- Identified the **date of the first successful landing** on a ground pad.
- Listed boosters that successfully landed on a drone ship and carried payloads **between 4000 and 6000 kg**.
- Counted the **number of successful and failed mission outcomes**.
- Retrieved booster versions that have carried the **maximum payload mass**.
- Extracted data on **failed drone ship landings** in 2015, along with booster versions and launch site names.
- Ranked landing outcomes (e.g., *Failure – drone ship*, *Success – ground pad*) **between 2010-06-04 and 2017-03-20** in descending order of count.

Build an Interactive Map with Folium

Github Link: Interactive Map with Folium

Launch Site Markers:

- Placed a circular marker on the map for the **NASA Johnson Space Center**, using its latitude and longitude as the initial reference point.
- Added similar markers (with circles, popup labels, and text labels) for **all other launch sites**, using their coordinates to illustrate geographical locations and their distance from the equator and nearby coastlines.

Colored Markers Based on Launch Outcomes:

- Used **color-coded markers—green** for successful launches and **red** for failed ones—grouped with Marker Clusters to help visualize the success rate at each launch site.

Distance Visualization to Nearby Infrastructure:

- Drew **colored lines** from the launch site (e.g., **KSC LC-39A**) to nearby points of interest, such as **railways**, **highways**, **coastlines**, and the **nearest city**, to show relative distances and surroundings.

Build a Dashboard with Plotly Dash

GitHub Link: Plotly Dash Dashboard

Launch Site Selection Dropdown: Implemented a dropdown menu that allows users to choose a specific launch site or view data for all sites collectively.

Pie Chart Displaying Launch Successes (All Sites / Selected Site):

Integrated a pie chart to visualize the count of successful launches across all sites.

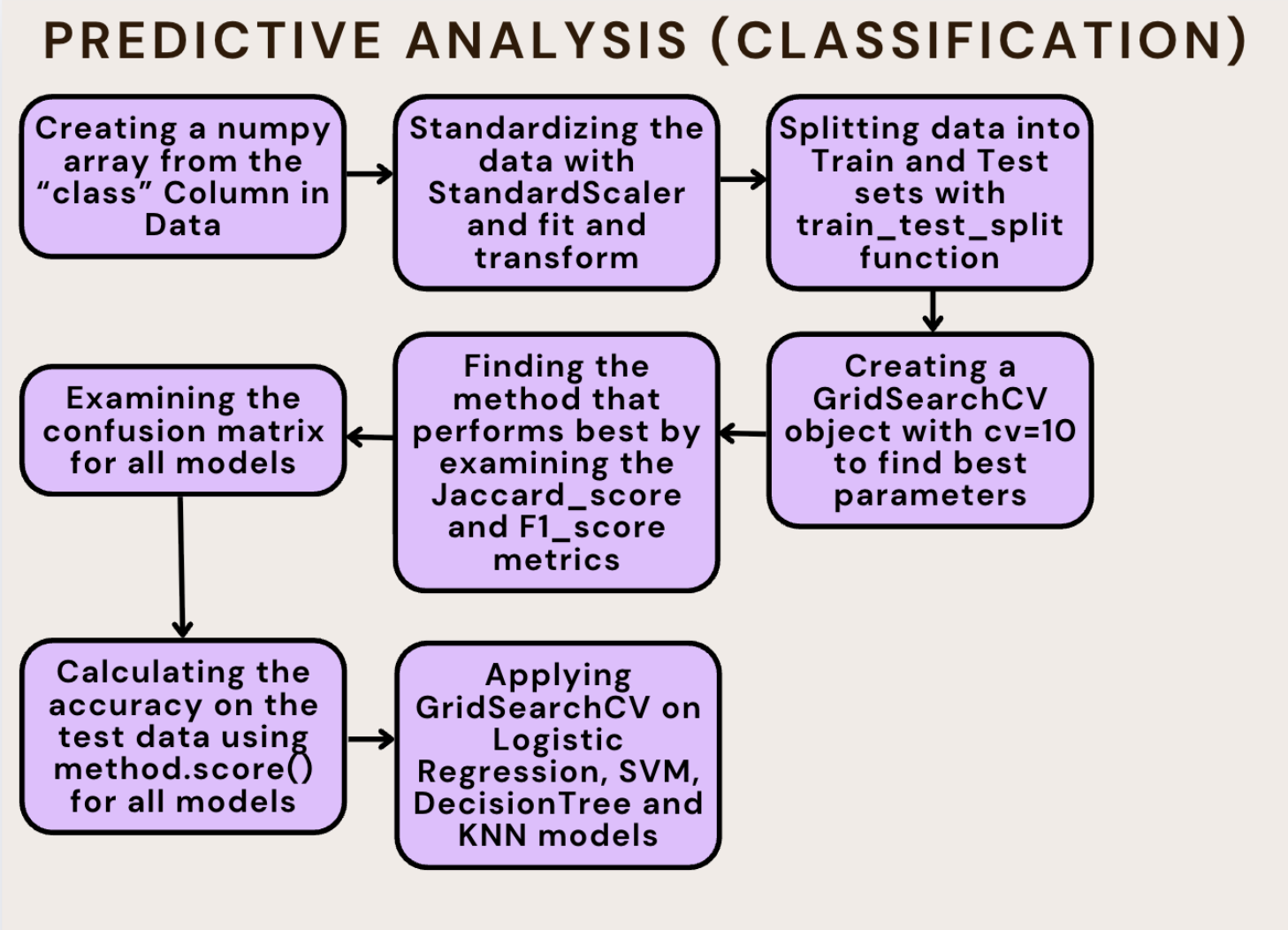
When a specific site is selected, the chart updates to compare successful vs. failed launches at that site.

Payload Mass Range Slider: Added an interactive slider to let users filter data based on payload mass range (in kilograms).

Scatter Plot of Payload vs. Success Rate by Booster Version: Created a scatter chart to show how payload mass relates to launch outcomes, categorized by different booster versions.

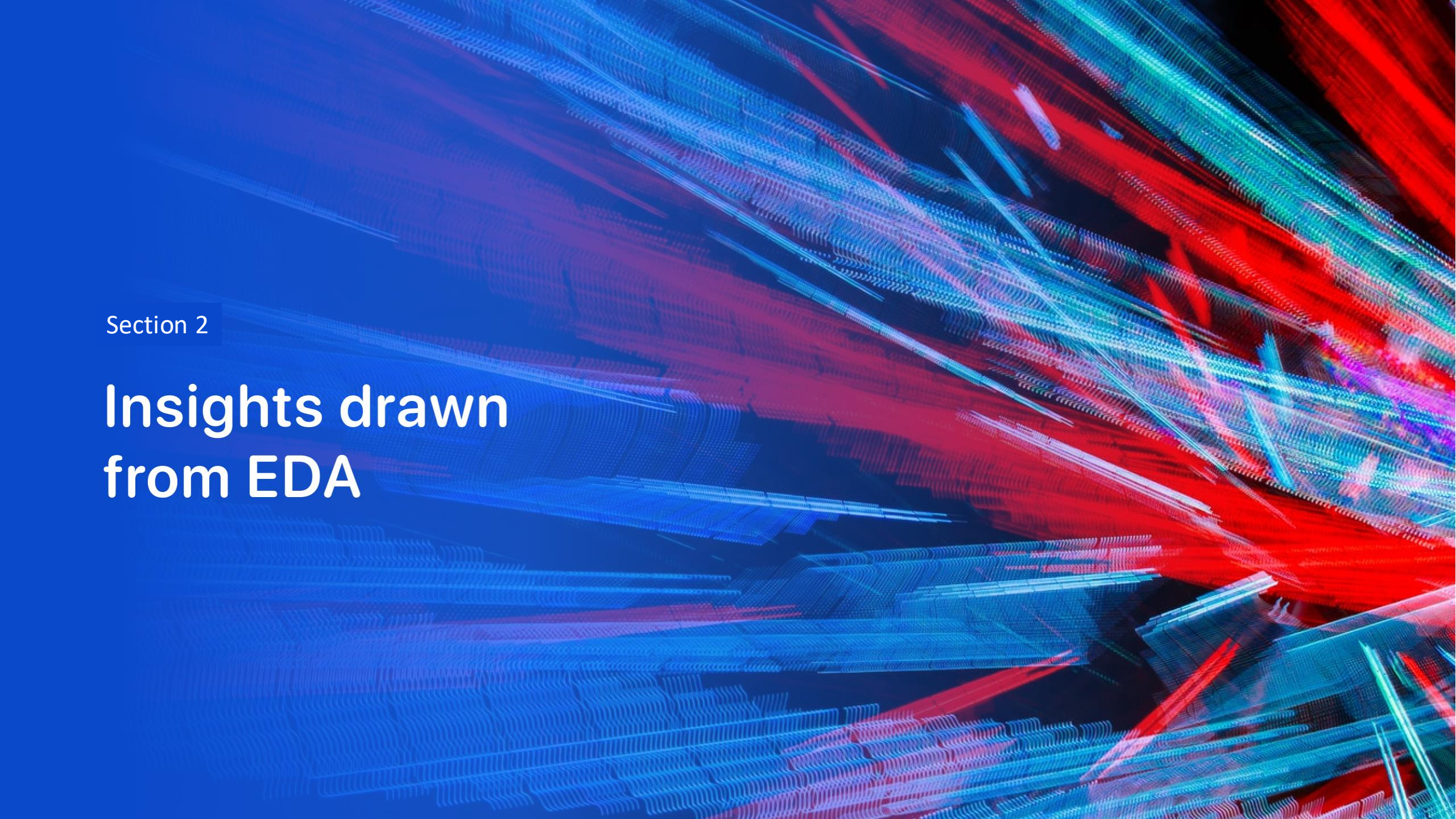
Predictive Analysis (Classification)

GitHub Link: [Machine Learning Prediction]

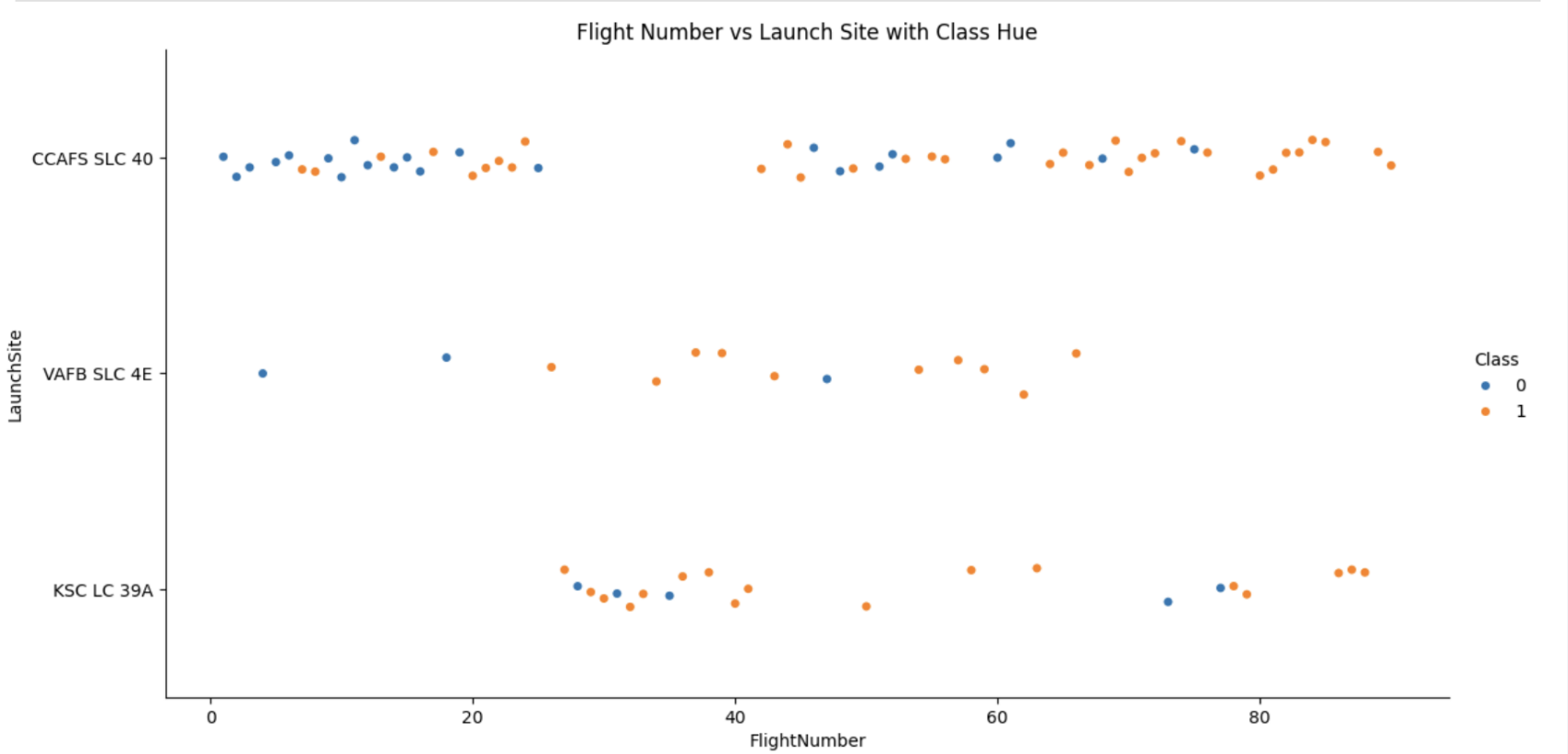


Results

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


Section 2
Insights drawn from EDA

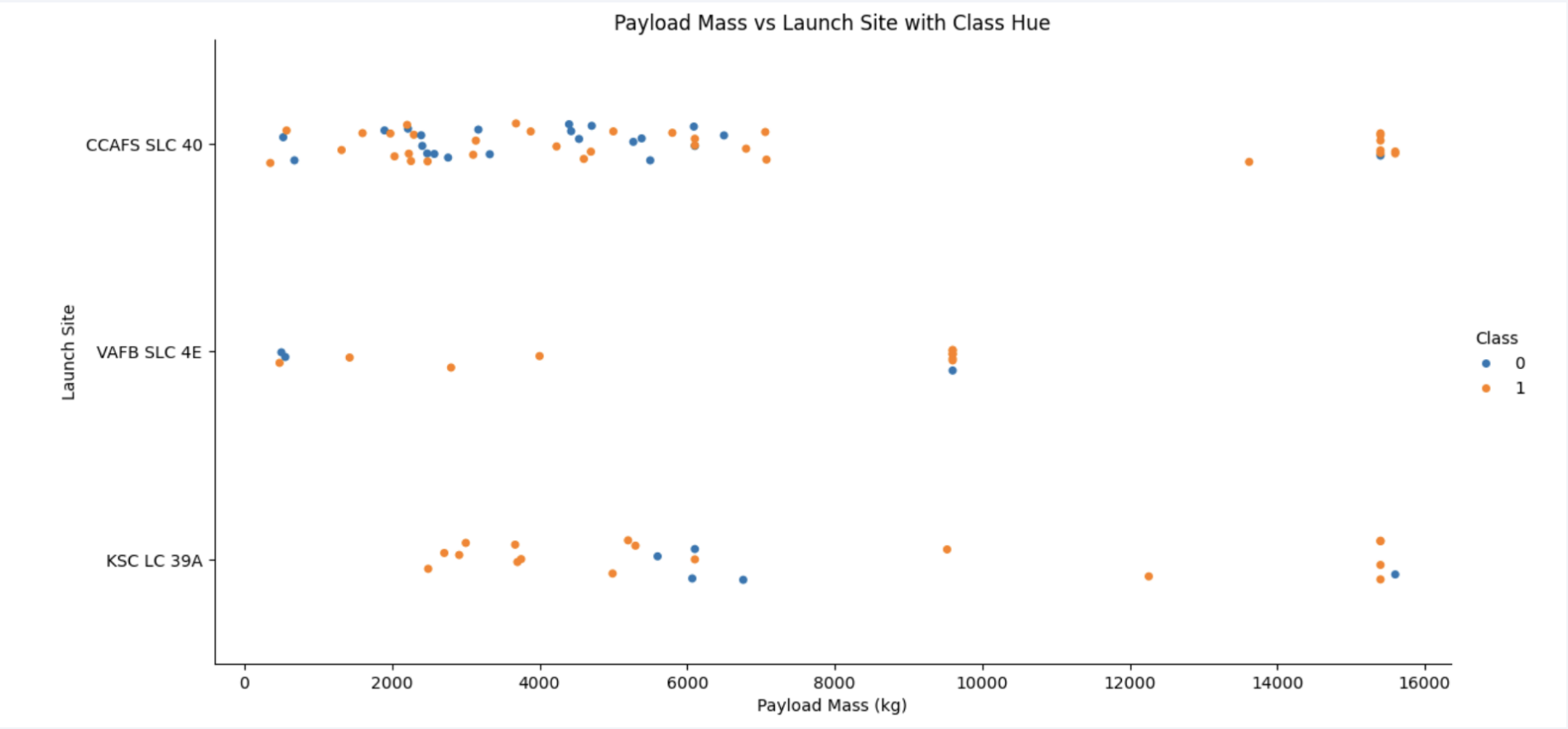
Flight Number vs. Launch Site



Explanation:

- The 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

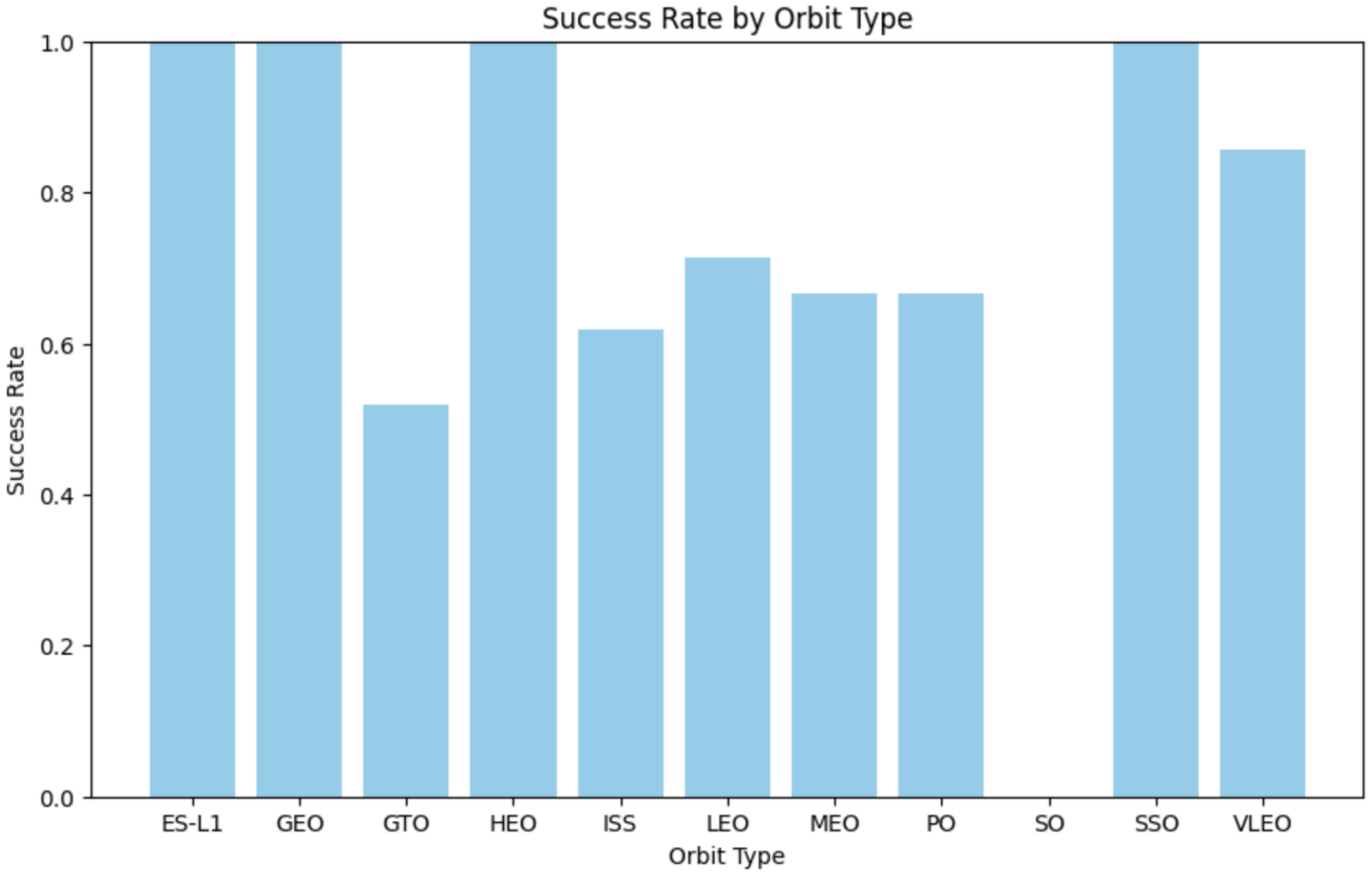
Payload vs. Launch Site



Explanation:

- For every launch site the higher the payload mass, the higher the success
- 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

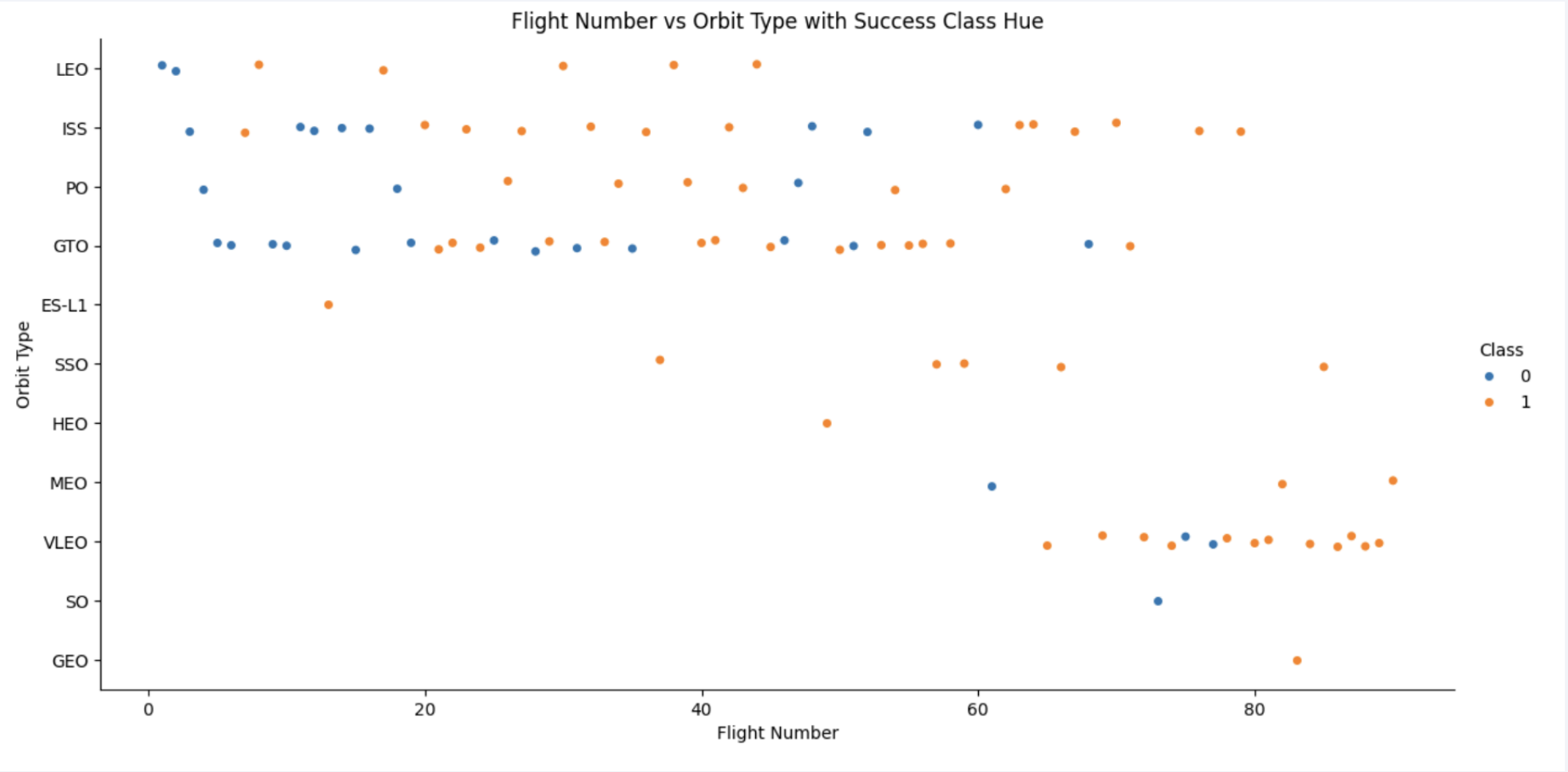
• Orbits with 0% success rate:

- SO

• Orbits with success rate between 50% and 85%:

- GTO, ISS, LEO, MEO, PO

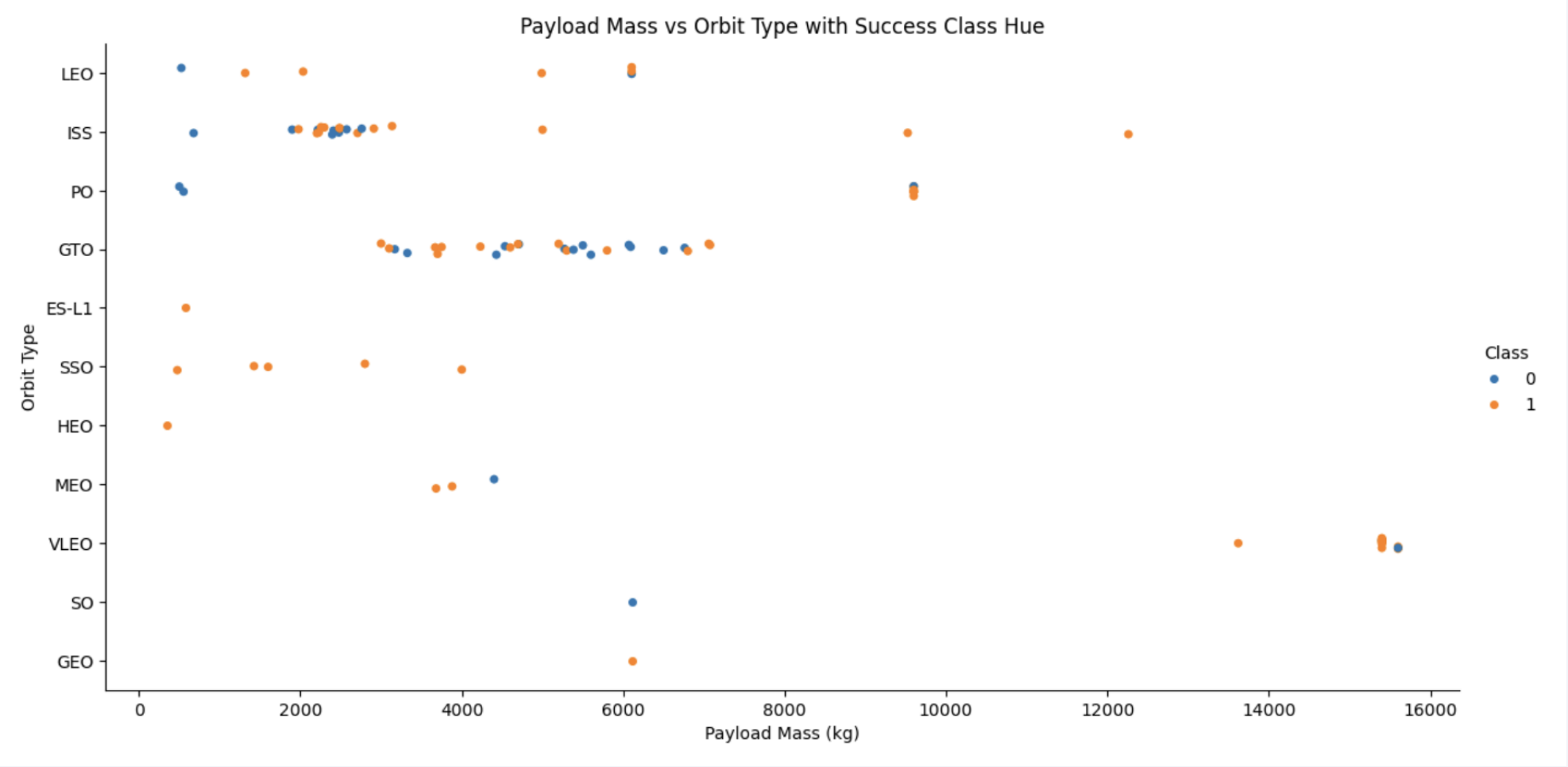
Flight Number vs. Orbit Type



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 in GTO orbit.

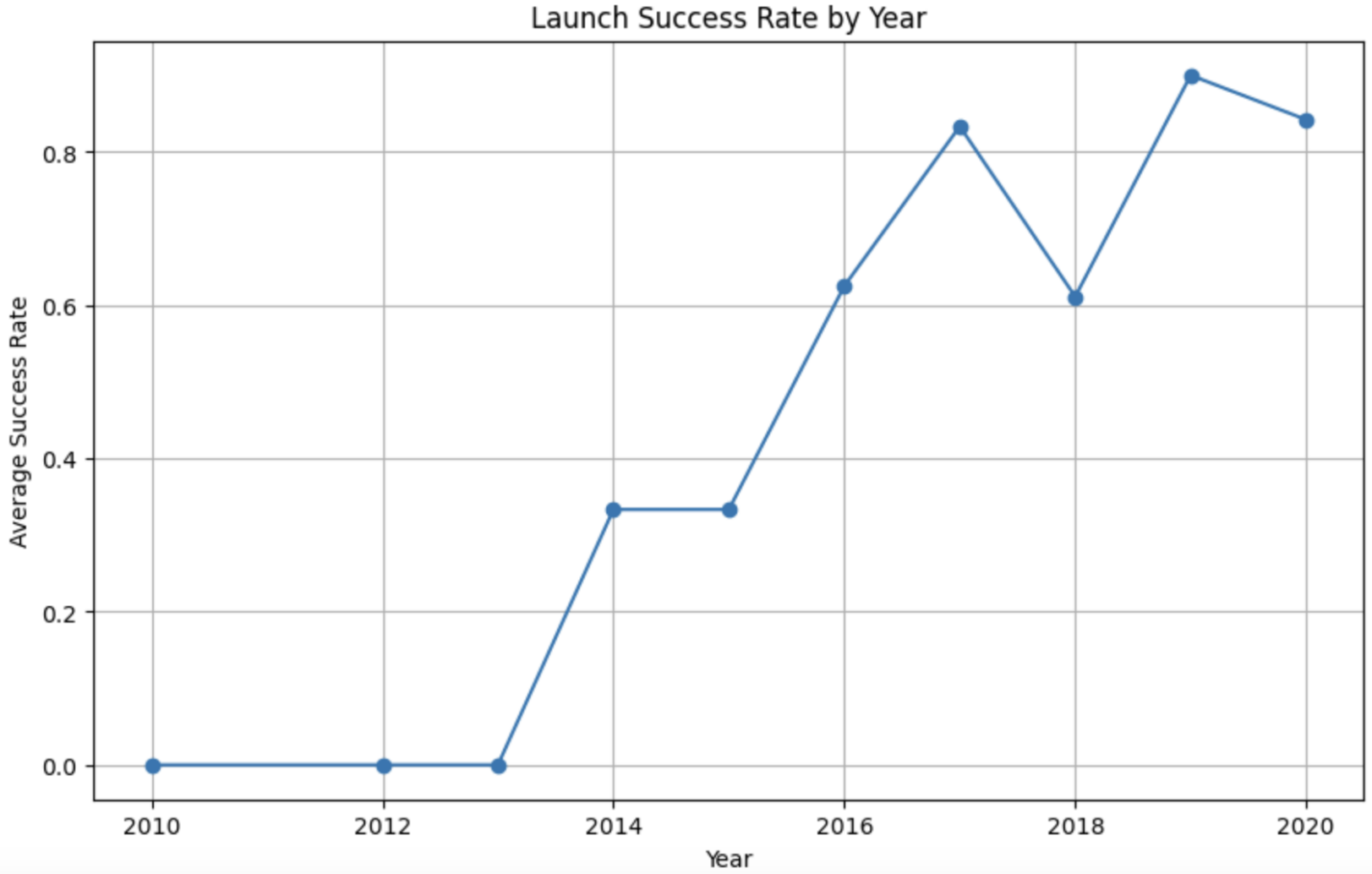
Payload vs. Orbit Type



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

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

In [10]:
```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

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

Out[10]:

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

Explanation:

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

Launch Site Names Begin with 'CCA'

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

[11]:
```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

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

[11]:

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

Explanation:

Displaying 5 records where launch site names begin with the string 'CCA'.

Total Payload Mass

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

[14]:
```
%sql SELECT SUM("Payload_Mass__kg_") AS Total_Payload_Mass FROM SPACEXTABLE WHERE "Customer" LIKE '%NASA (CRS)%';
```

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

[14]:

Total_Payload_Mass
48213

Explanation:

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

Average Payload Mass by F9 v1.1

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

[15]:
```
%sql SELECT AVG("Payload_Mass__kg_") AS Average_Payload_Mass FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';
```

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

[15]:

Average_Payload_Mass
2928.4

Explanation:

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

First Successful Ground Landing Date

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

Hint:Use min function

[16]:
```
%sql SELECT MIN("Date") AS First_Successful_Ground_Landing FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

[16]:

First_Successful_Ground_Landing
2015-12-22

Explanation:

Listing the date when the first successful landing outcome in ground pad was achieved which is **2015-12-22**.

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

[17]:
```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "Payload_Mass__kg_" > 4000 AND "Payload_Mass__kg_" < 6000;
```

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

[17]:

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

Explanation:

Listing the names of the boosters which have success in drone ship and have 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

[18]:
```
%sql SELECT "Mission_Outcome", COUNT(*) AS Outcome_Count FROM SPACEXTABLE GROUP BY "Mission_Outcome";
```

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

[18]:

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

Explanation:

Listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

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

[20]:
```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Payload_Mass__kg_" = (SELECT MAX("Payload_Mass__kg_") FROM SPACEXTABLE);
```

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

[20]:

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

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

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.

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.

[21]:
```
%sql SELECT substr("Date", 6, 2) AS Month,"Landing_Outcome","Booster_Version","Launch_Site" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE '%Failure (drone ship)%'AND substr("Date",
```

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

[21]:

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

Explanation:

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.

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 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.
```

[22]:

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

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.

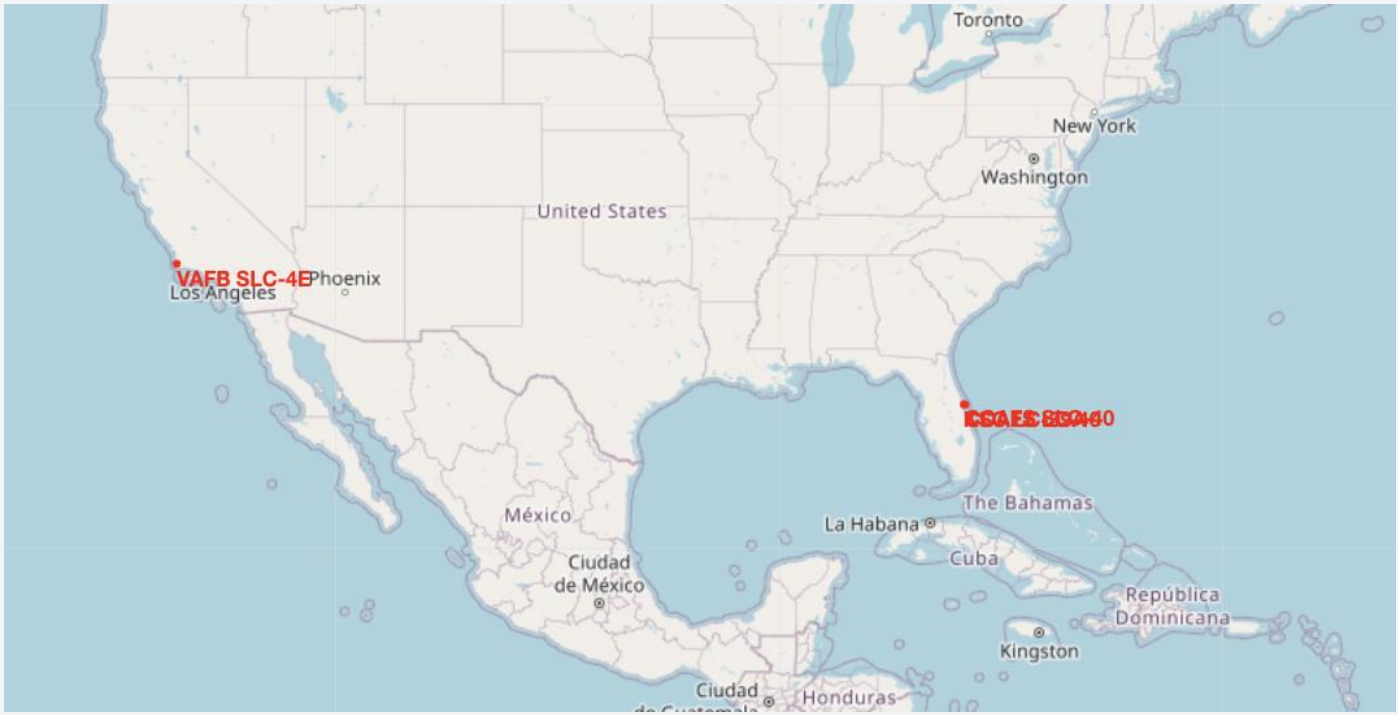
Section 3

Launch Sites Proximities Analysis

All Launch sites location markers on a global map

Explanation:
Most launch sites are located near the equator. This is because the Earth's surface rotates fastest at the equator—about 1670 km/h. Launching from this region allows spacecraft to take advantage of this rotational speed due to inertia, helping them achieve the necessary velocity to enter and stay in orbit more efficiently.

Additionally, all launch sites are situated close to coastlines. Launching rockets over the ocean reduces the risk to human populations by ensuring that any debris or potential explosions occur away from inhabited areas.



Colour-labeled launch records on the map

Explanation:

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

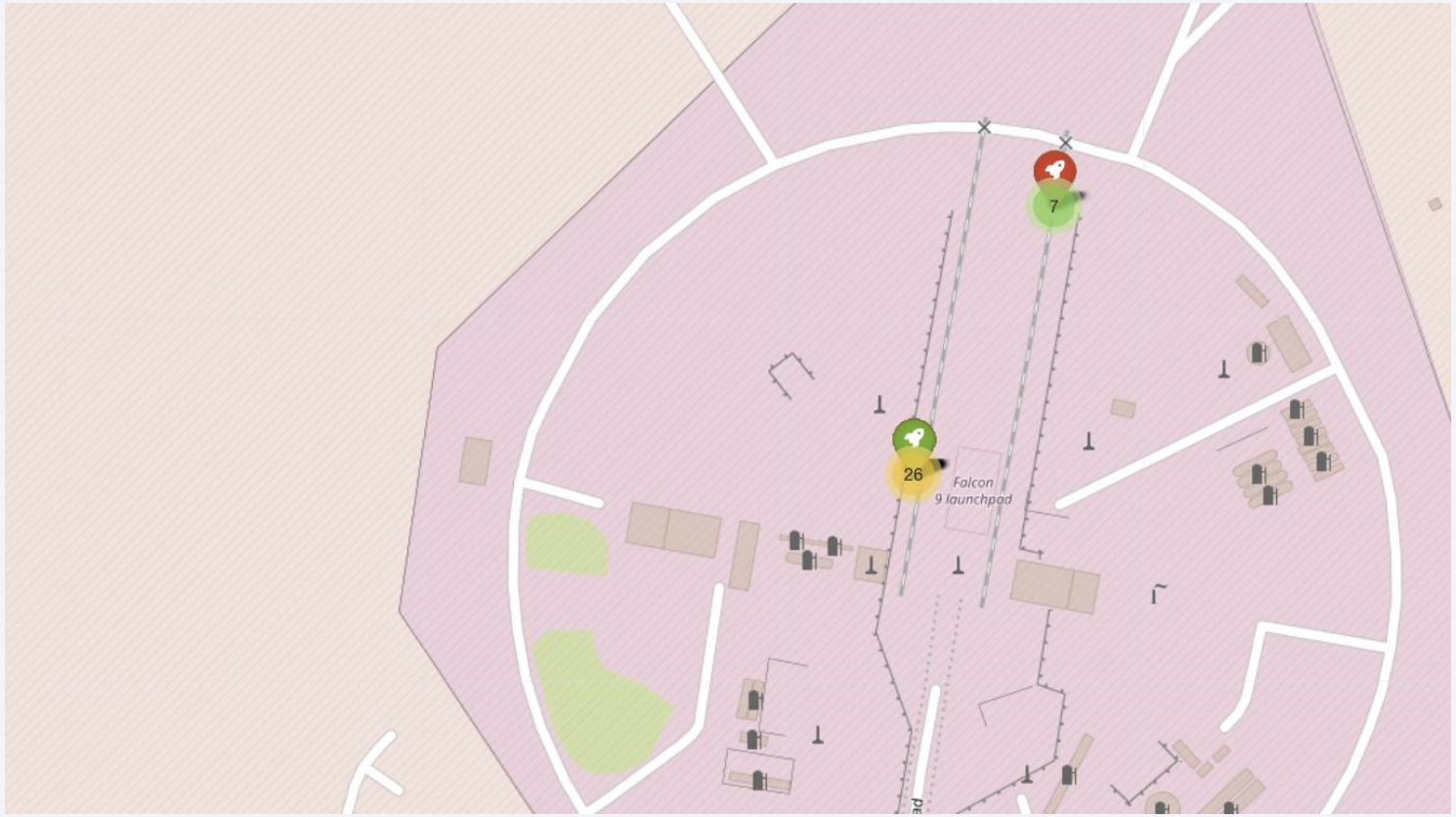
Launch Site KSC LC-39A has a very high Success Rate.

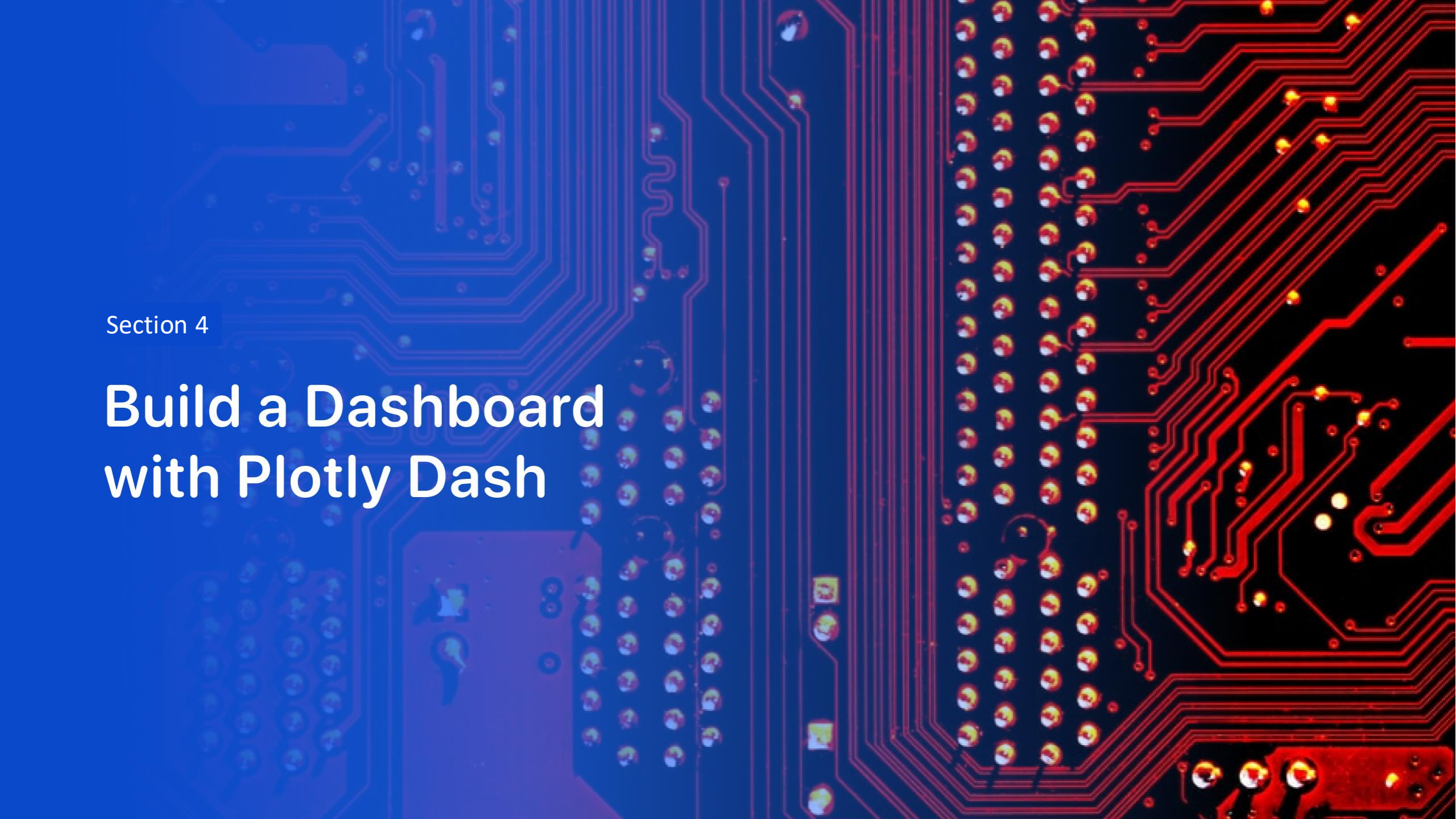
Distance from the launch site KSC LC-39A to its proximities

Explanation:

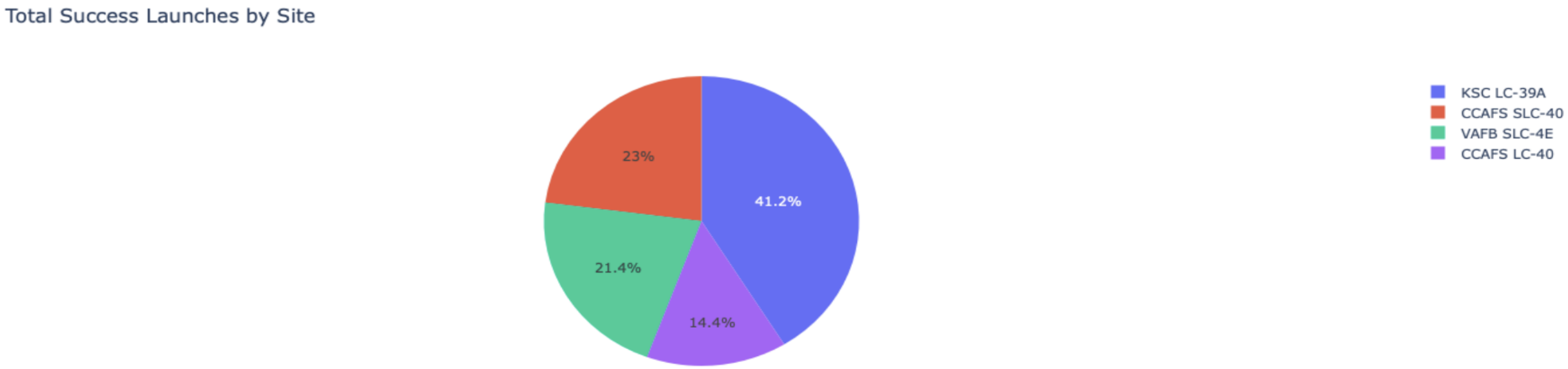
From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
- relative close to railway (15.23 km)
- relative close to highway (20.28 km)
- relative close to coastline (14.99 km)

Also the launch site KSC LC-39A is relative 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

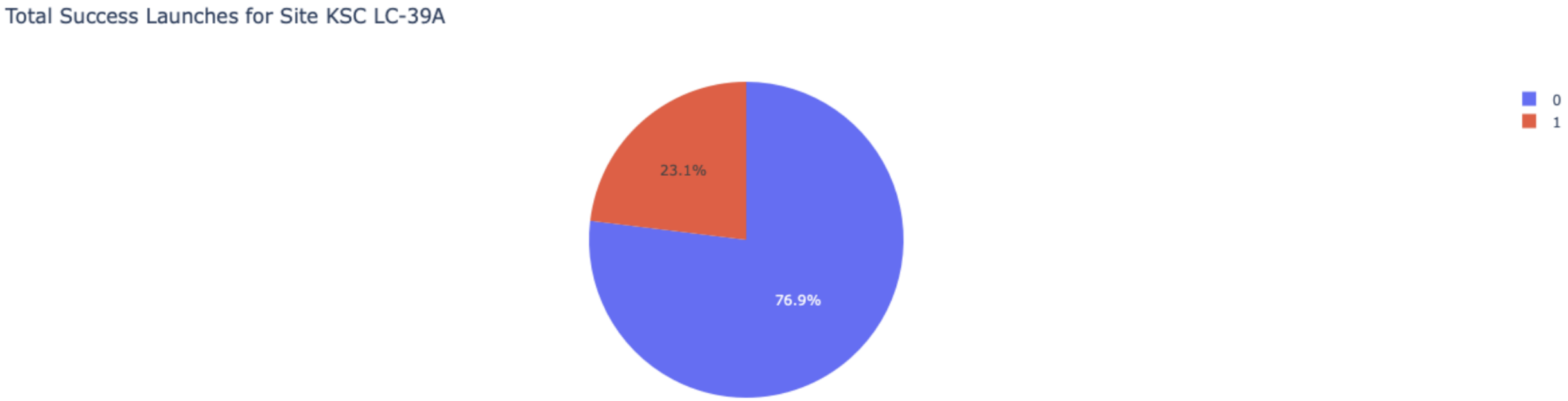
Total Success Launches by Site – Pie Chart



Explanation:

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:

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

Payload Mass vs. Launch Outcome for all sites

Explanation:

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

Section 5

Predictive Analysis (Classification)

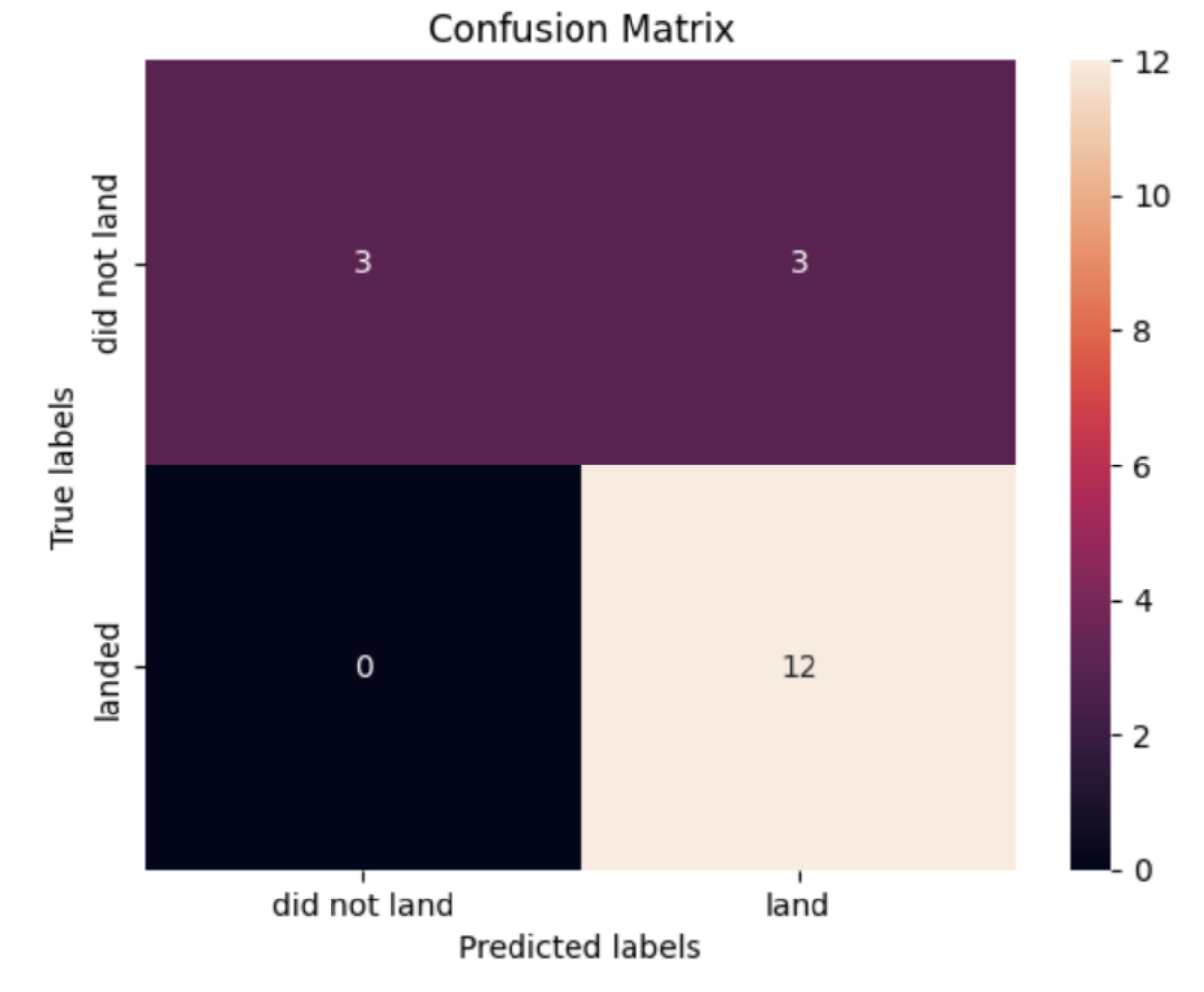
Classification Accuracy

[39]:

	LogReg	SVM	Tree	KNN
F1_Score	0.909091	0.916031	0.900763	0.900763
Accuracy	0.866667	0.877778	0.855556	0.855556

Explanation:
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 logistic regression 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

Thanks to Coursera, IBM for this course!

Thank you!