IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Partho Sarothi Das
25th September 2025

Outline

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

Executive Summary

- **Methodologies**
  - Collected SpaceX data using REST API and web scraping
  - Performed data wrangling to clean and transform datasets
  - Conducted EDA with SQL and visualizations
  - Built interactive analytics with Folium maps and Plotly Dash
  - Developed and evaluated classification models for launch outcome prediction
- **Results**
  - Identified patterns between payload, launch site, orbit type, and mission success
  - Built interactive dashboards for visual exploration of launch performance
  - Found the best-performing classification model with high accuracy in predicting mission success

Introduction

Project Background:

SpaceX aims to reduce the cost of space travel with reusable rockets.

Problem Statement:

Can we predict launch success and understand factors (payload, orbit, site) influencing outcomes?

Objective:

Perform data-driven analysis and predictive modeling on SpaceX launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Extracted launch data from SpaceX API (REST calls), Scraped additional data from SpaceX Wikipedia page, Stored results in CSV for processing

- Data wrangling

  Cleaned missing/invalid values, Created derived columns, Standardized payload mass and booster categories

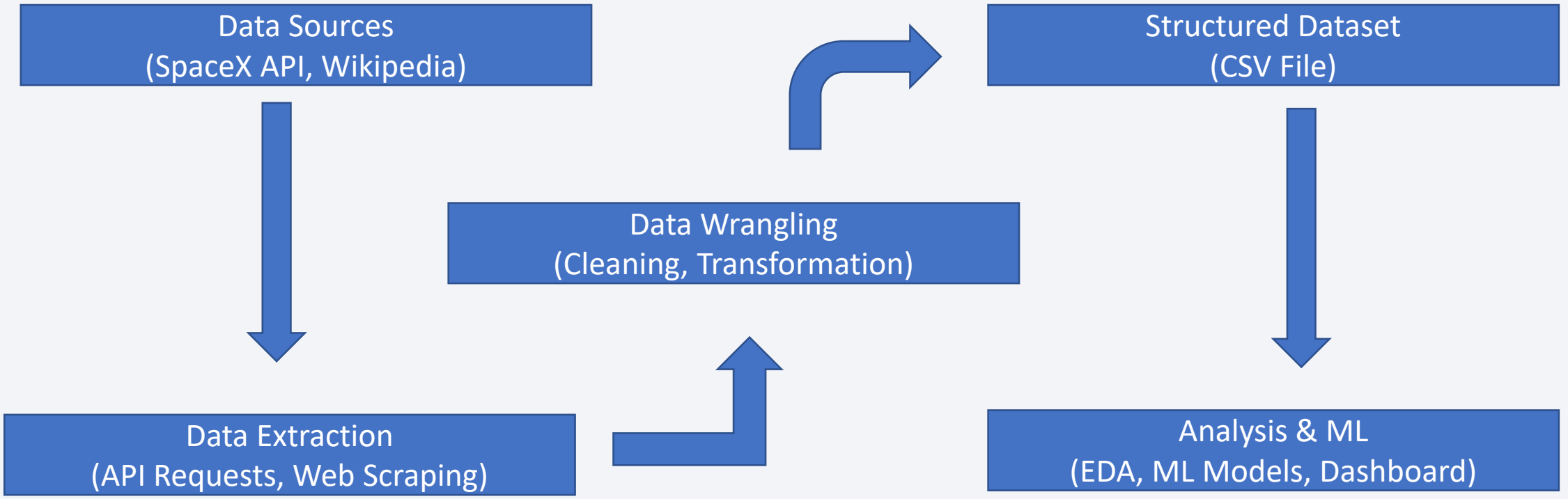
- Perform exploratory data analysis (EDA) using visualization and SQL

- Perform interactive visual analytics using Folium and Plotly Dash

- Perform predictive analysis using classification models

  Built classification models (Logistic Regression, Decision Tree, KNN, SVM), Hyperparameter tuning for best performance, Evaluated models with accuracy scores and confusion matrix.

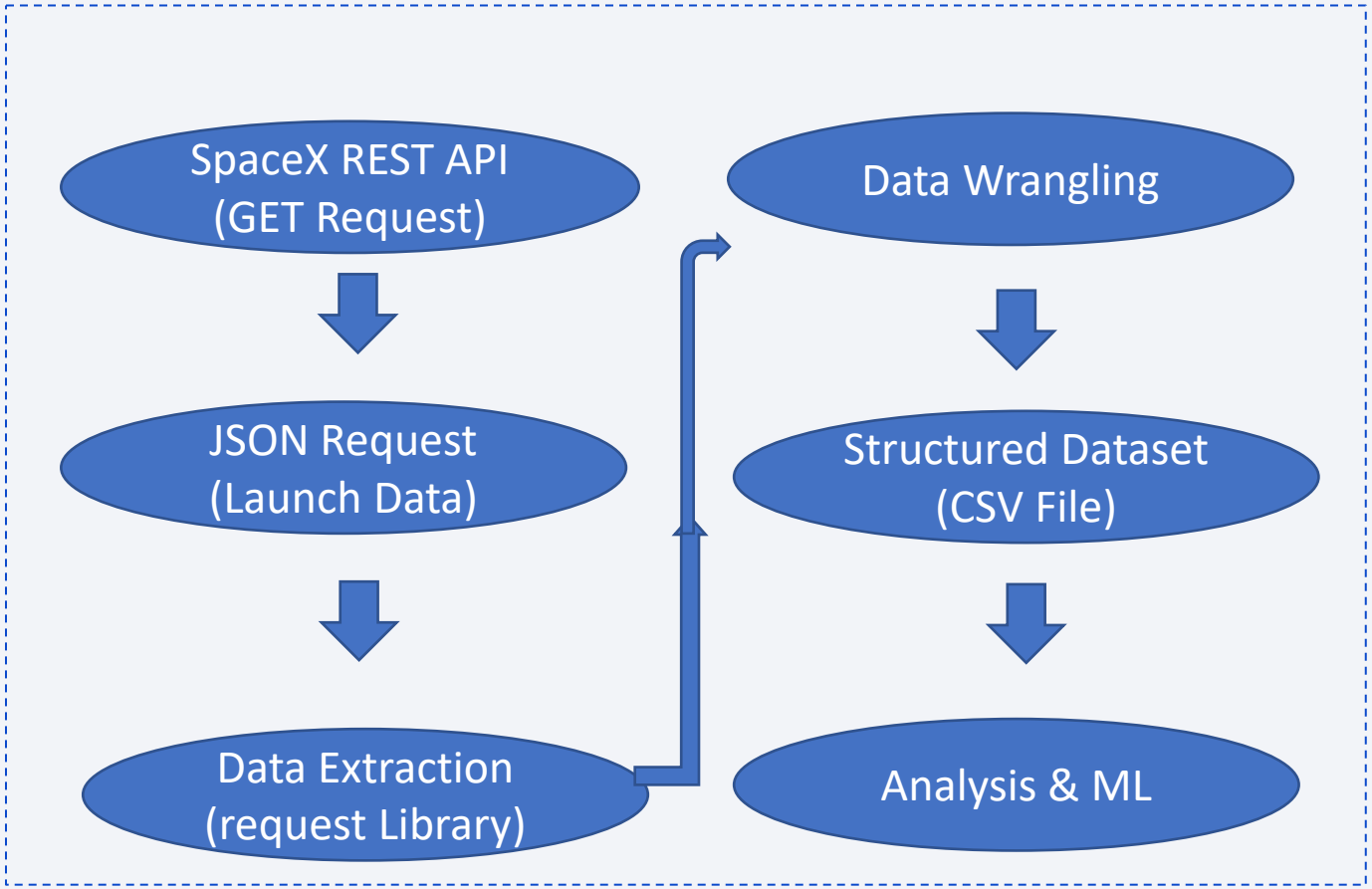
Data Collection



Data Collection – SpaceX API

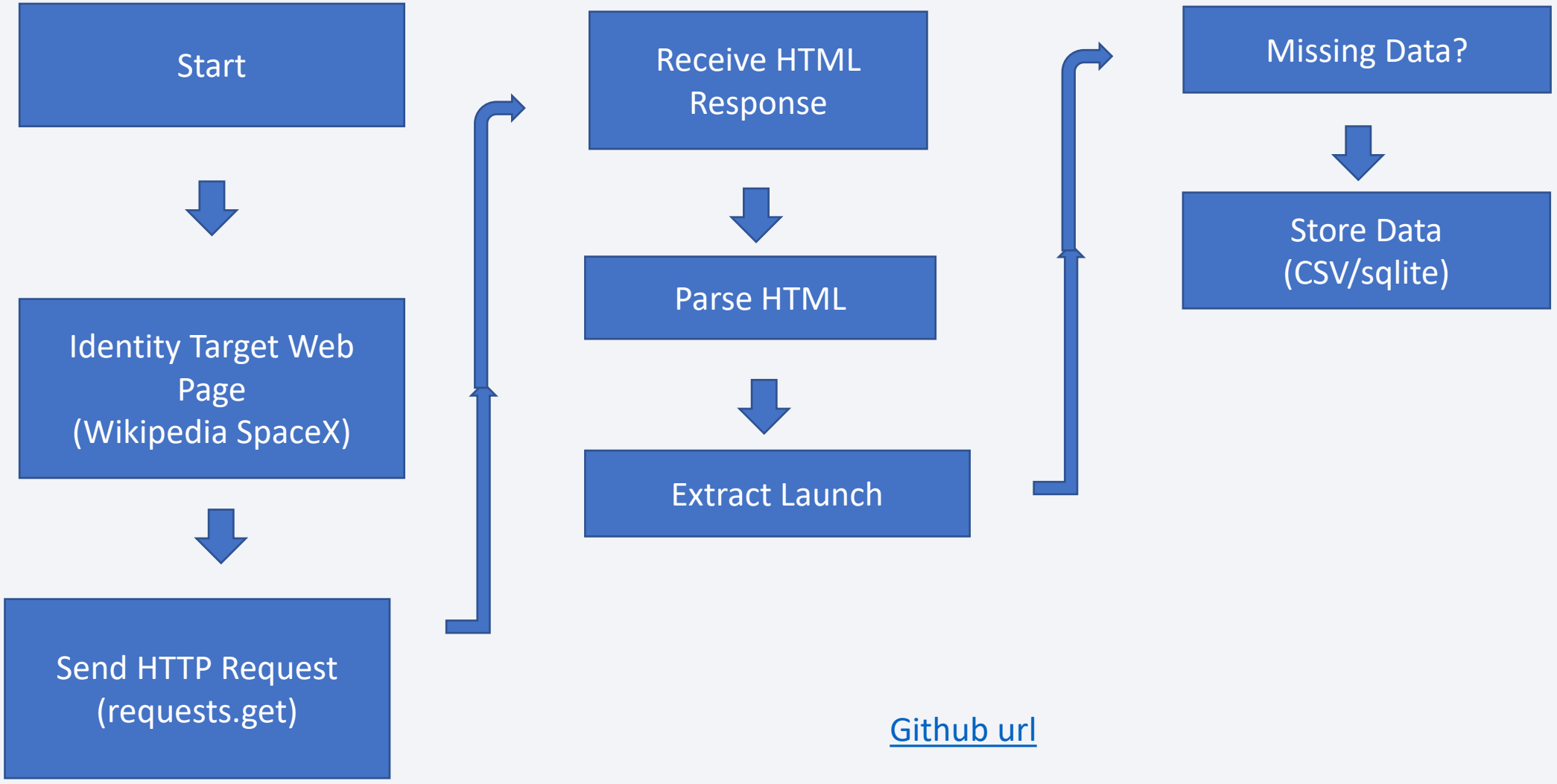
- Key Phrases to

1. Rest API Requests (GET method)
2. JSON Response Parsing
3. Data Wrangling with Pandas
4. Data Cleaning
5. Structured Dataset Creating (CSV)



GitHub URL:

Data Collection - Scraping

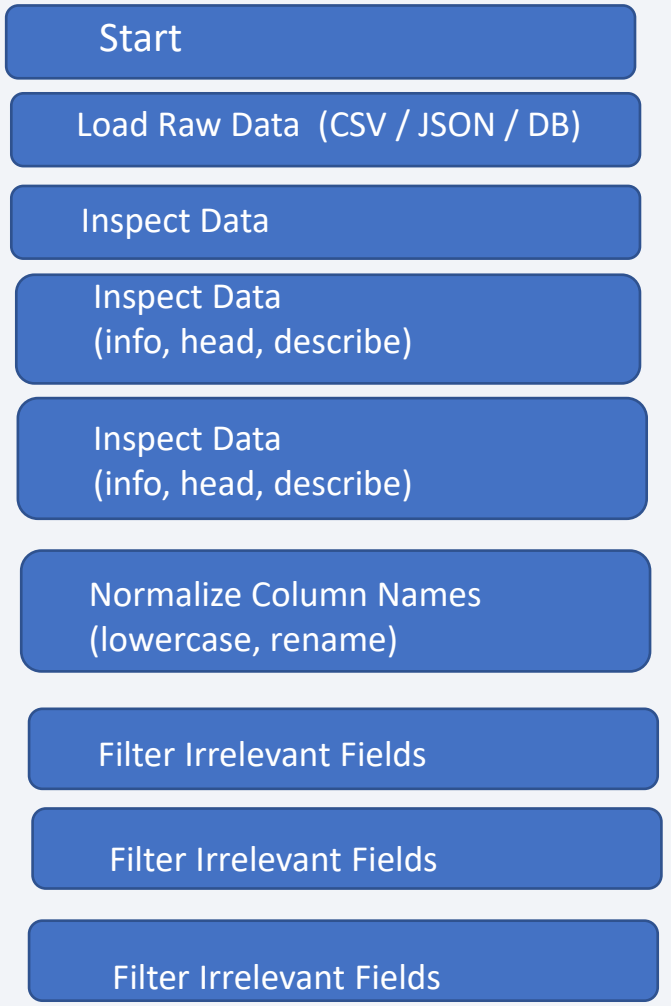


Github url

Data Wrangling

- In the data set, there are several different cases where the
- booster did not land successfully. Sometimes a landing was
- attempted but failed due to an accident; for example, True
- Ocean means the mission outcome was successfully landed
- to a specific region of the ocean while False Ocean means
- the mission outcome was unsuccessfully landed to a specifi c region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means
- the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission
- outcome was unsuccessfully landed on a drone ship.
- We mainly convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful.

GitHub URL



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).

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
- 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

Github: 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.

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.

GitHub URL

Build a Dashboard with Plotly Dash

Dropdown Menu (Launch Site Selection)

Allows users to select:

Pie Chart (Success Rate Distribution)

Shows proportion of successful vs failed launches.

Scatter Plot (Payload vs. Launch Outcome)

X-axis: Payload Mass (kg)

Y-axis: Launch Success (0 = Failure, 1 = Success)

Range Slider (Payload Mass Filter)

Updates scatter plot in real-time.

Github url:

Predictive Analysis (Classification)

Data Collection (API & Web Scraping)

Data Wrangling (Clean, Encode, Feature Scaling)

Train-Test Split (80% 20%)

Build Classifiers (LR, SVM, DT, KNN)

Baseline Evaluation (Accuracy Scores)

Hyperparameter Tuning (GridCV)

Re-Evaluation (Validation Scores)

Final Model (Best Accuracy → SVM RBF Kernel)

Github: ML Prediction

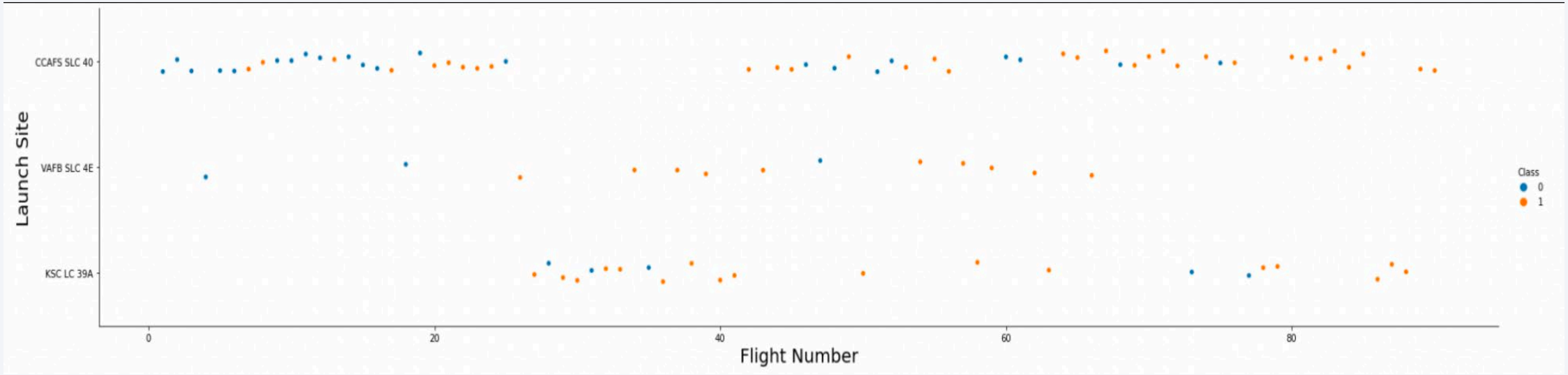
Results

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

Section 2

Insights drawn from EDA

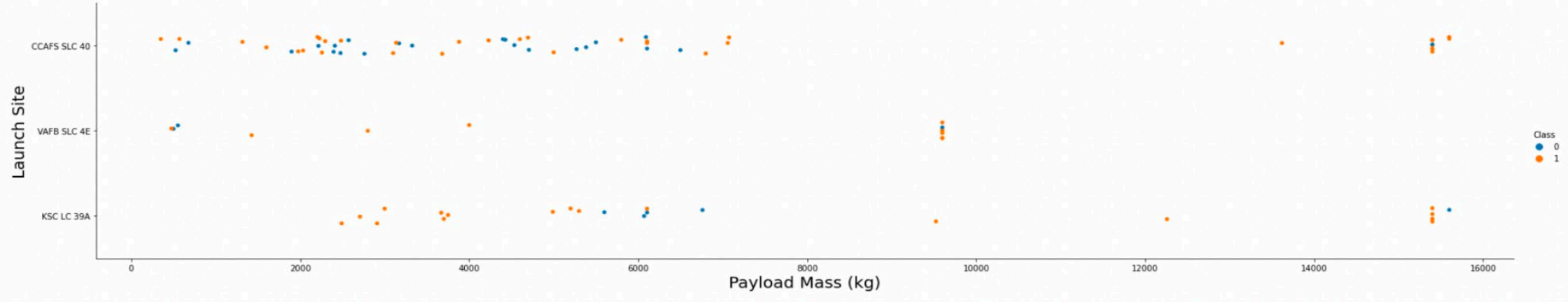
Flight Number vs. Launch Site



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:

X-axis = Payload Mass (kg)

Shows how heavy the payload was for each launch.

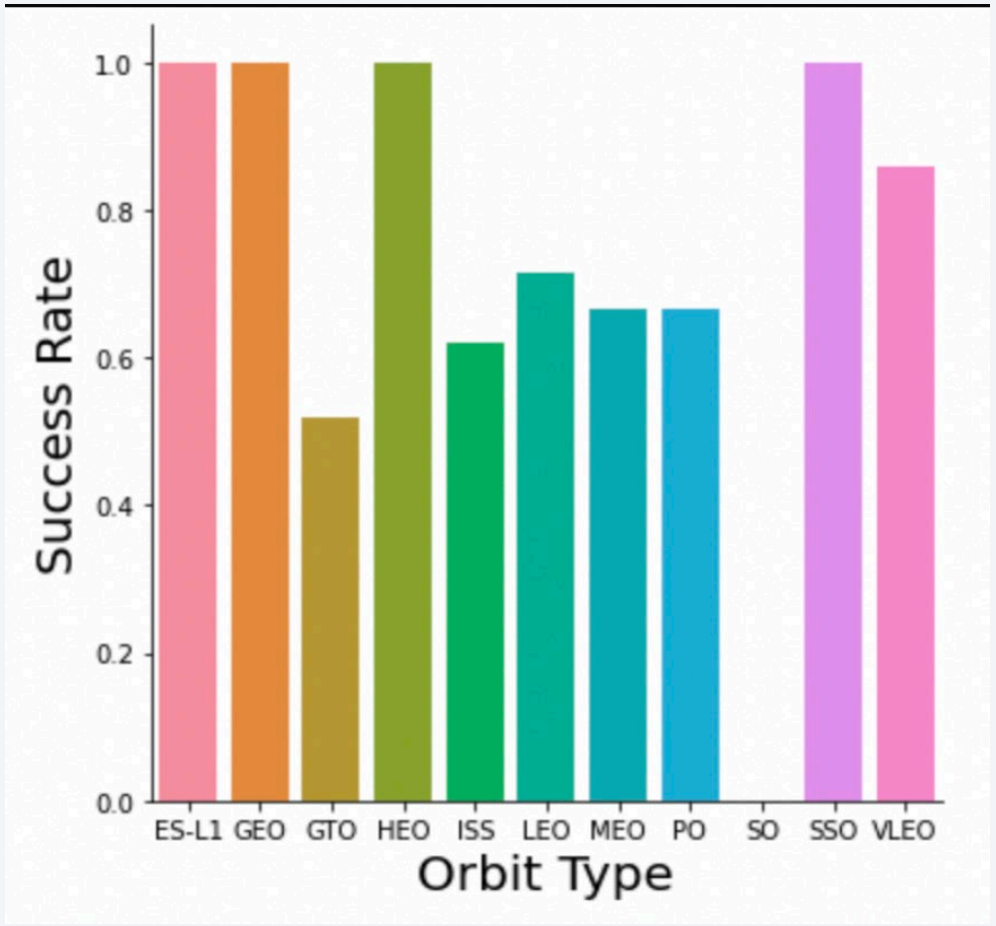
Y-axis = Launch Site

Different SpaceX launch pads (e.g., CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E).

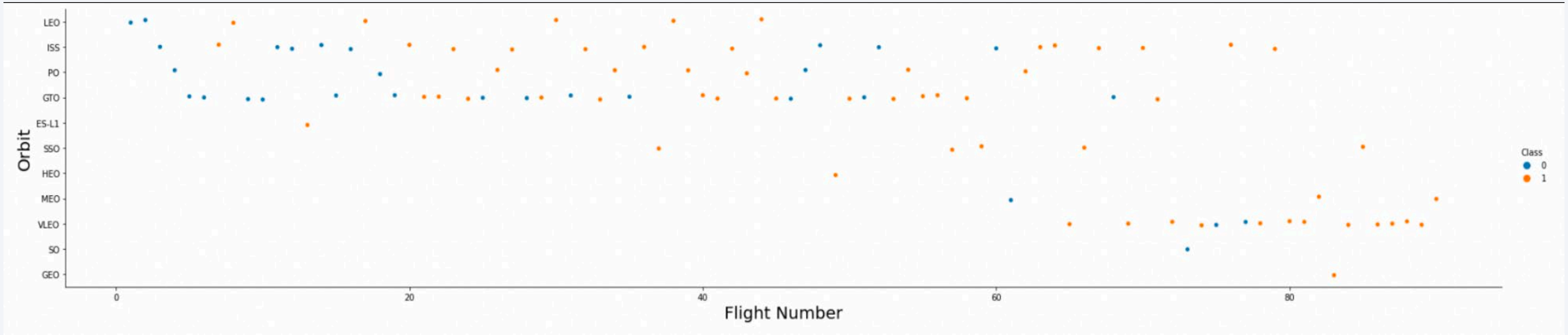
Success Rate vs. Orbit Type

Explanation:

- X-axis = Orbit Type
  - Different orbit destinations (e.g., LEO, GEO, ISS, SSO).
- Y-axis = Success Rate (0–1)
  - Shows the proportion of successful launches for that orbit.
- Bar Heights = Performance
  - Taller bars = higher success rate.
  - Shorter bars = lower reliability for that orbit.



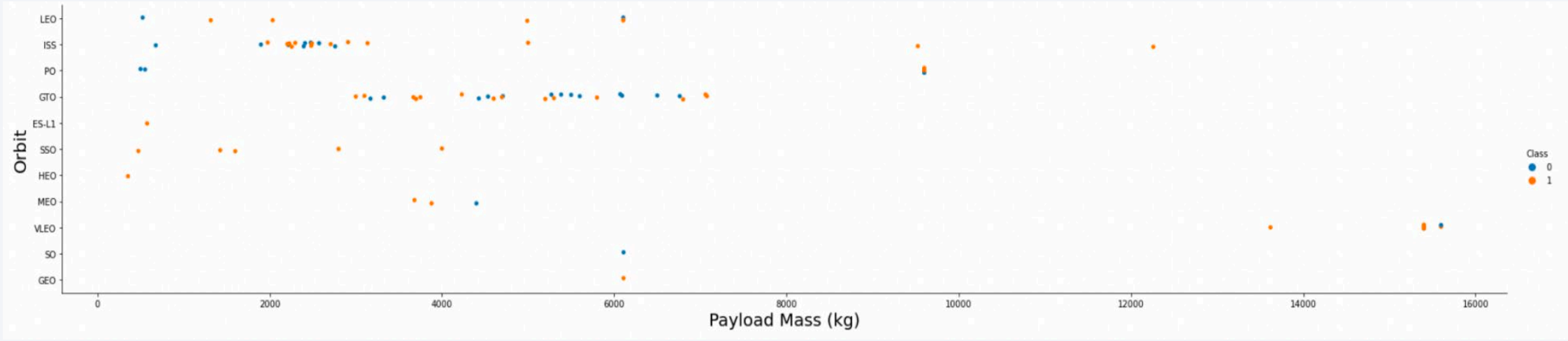
Flight Number vs. Orbit Type



Explanation:

- X-axis = Flight Number
  - Each launch in chronological order (earlier flights on the left, newer ones on the right).
- Y-axis = Orbit Type
  - Each point represents a launch's target orbit (LEO, GTO, ISS, SSO, etc.).

Payload vs. Orbit Type

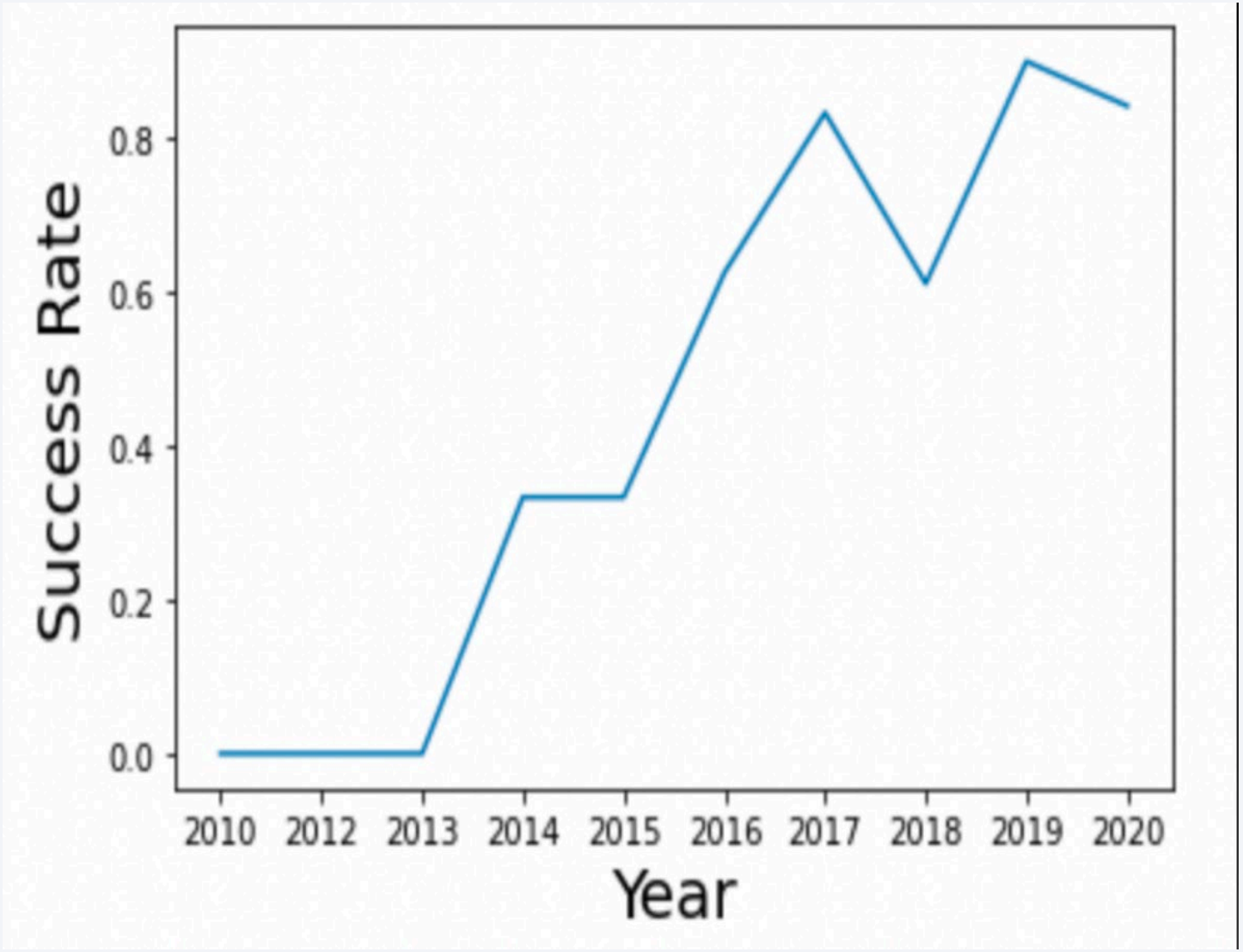


Explanation:

- 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

Explanation:

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

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

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

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

[17]:

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

Launch Site Names Begin with 'CCA'

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

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

[19]:

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 sites begin with the string 'CCA'.

Total Payload Mass

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

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

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

[21]:

Total_Payload_Mass
45596

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

```
[23]: %sql SELECT AVG("PAYLOAD_MASS__KG_") AS Average FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 v1.1%'
```

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

[23]:

Average
2534.6666666666665

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.

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

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

[25]:

MIN(Date)
2015-12-22

Explanation:

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

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

[27]:
```
%sql SELECT DISTINCT("Booster_Version") FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MAS
```

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

[27]:

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

```
[29]: %sql SELECT "Mission_Outcome", COUNT(*) AS Total_Count FROM SPACEXTABLE GROUP BY "Mission_Outcome"
```

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

[29]:

Mission_Outcome	Total_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.

```
[31]: %sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);
 * sqlite:///my_data1.db
Done.
```

[31]:

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

2015 Launch Records

[33]:
```
%%sql
SELECT
  substr(Date, 6, 2) AS Month,
  "Landing_Outcome",
  "Booster_Version",
  "Launch_Site"
FROM SPACEXTABLE
WHERE substr(Date, 0, 5) = '2015'
  AND "Landing_Outcome" LIKE '%Failure%'
  AND "Landing_Outcome" LIKE '%drone ship%';
```

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

[33]:

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 failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

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

[35]:
```
%%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.
```

[35]:

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.

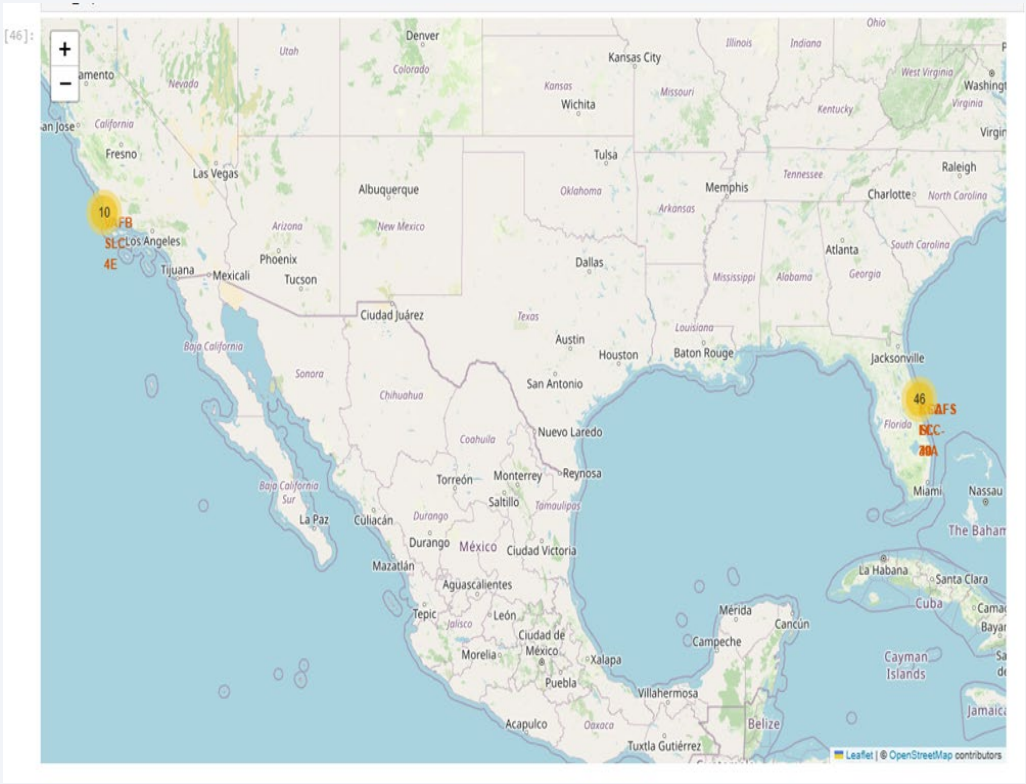
Section 3

Launch Sites Proximities Analysis

All Launch Sites

Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.

• All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.



Color-labeled Launch Outcomes

- From the colour-labeled markers

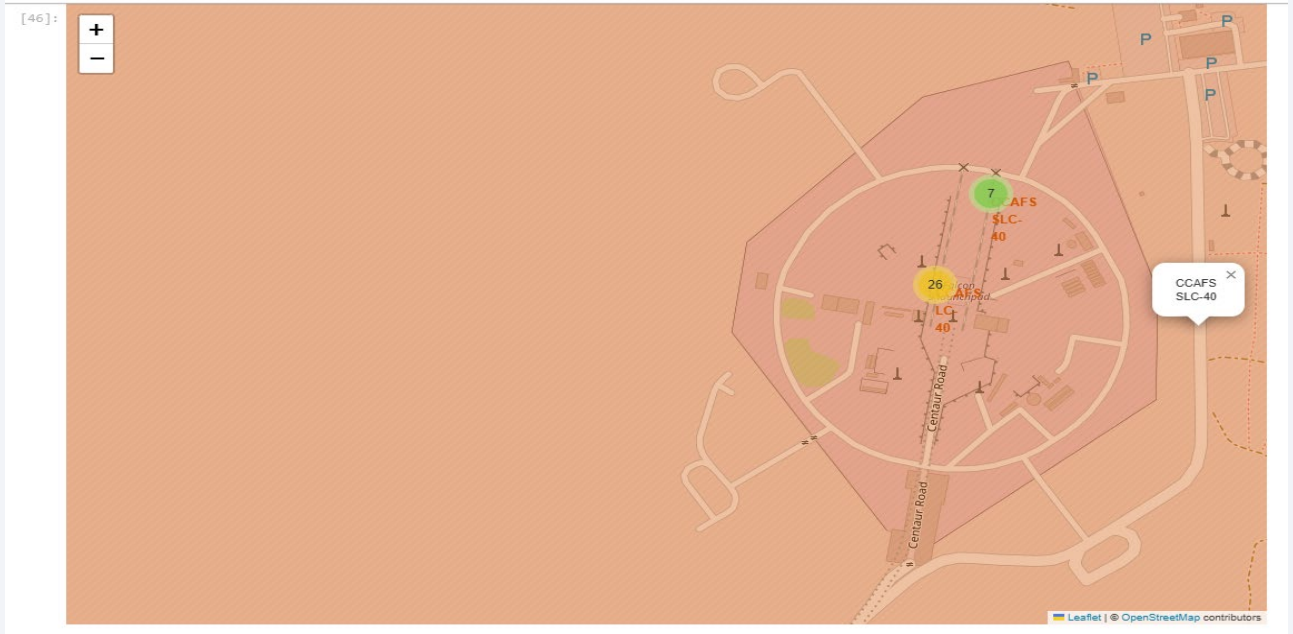
we should be able to easily identify which launch sites have relatively high success rates.

  - Green Marker = Successful Launch

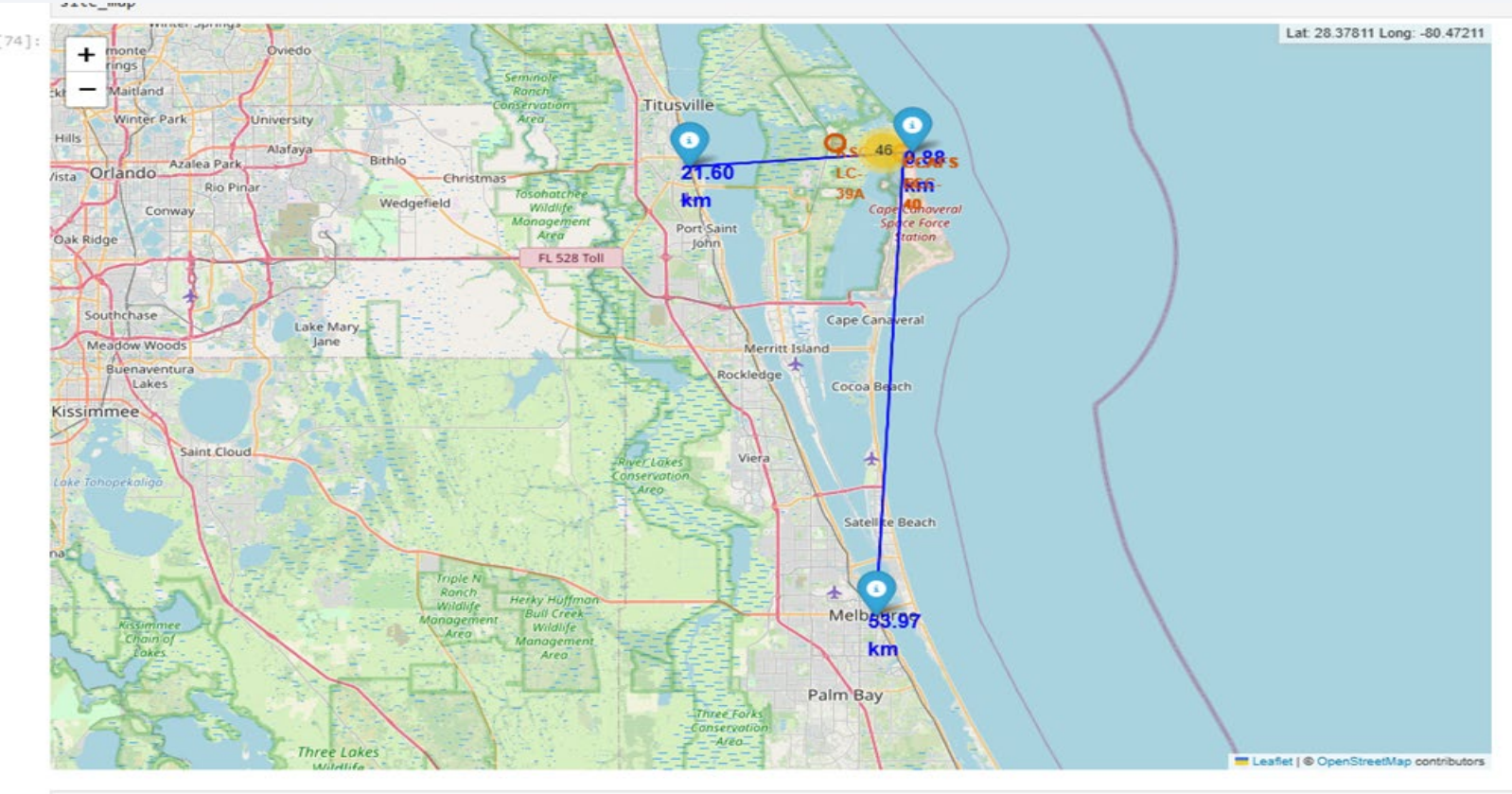
  - Red Marker = Failed Launch

- Launch Site KSC LC-39A has a

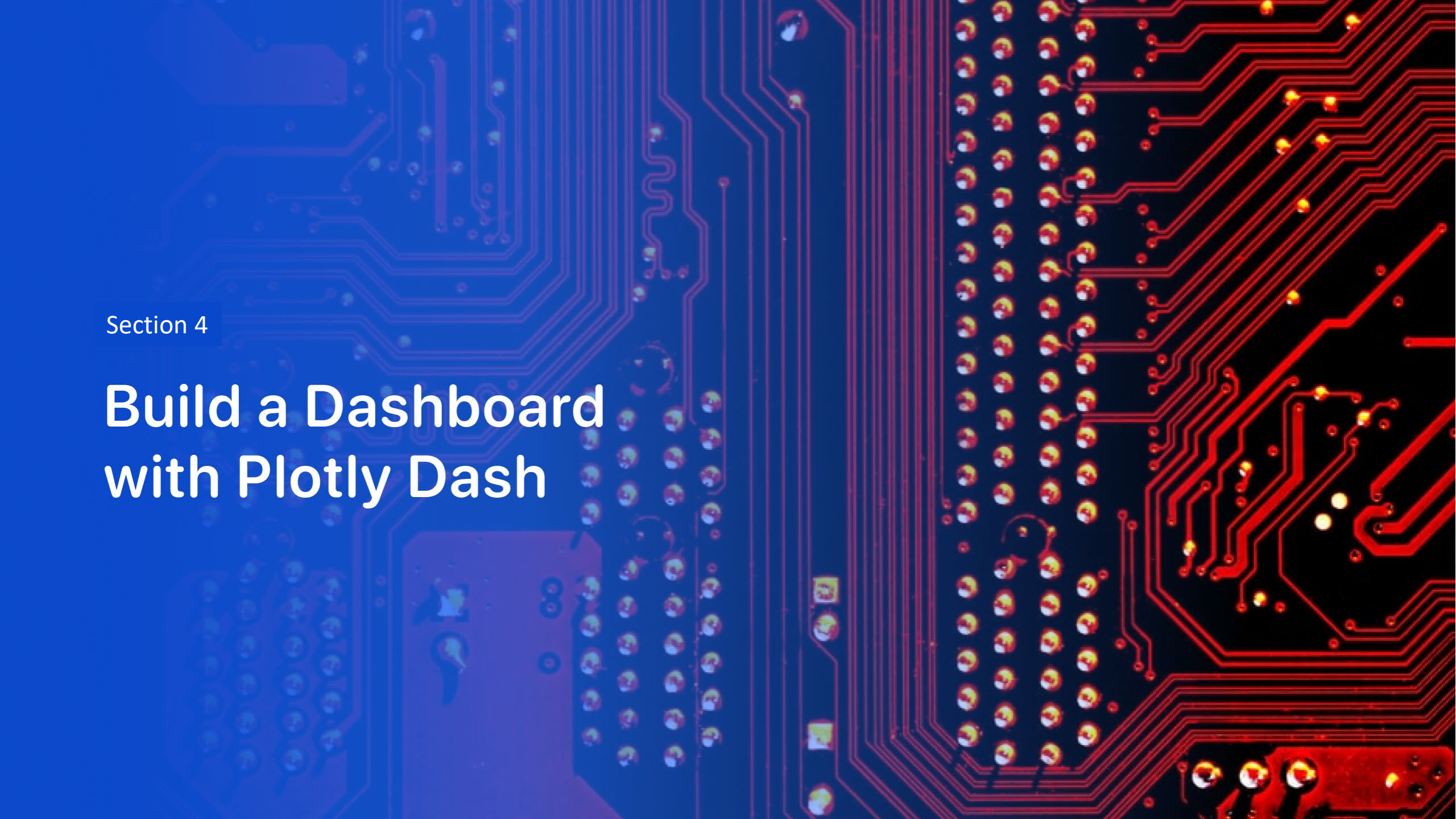
very high Success Rate.



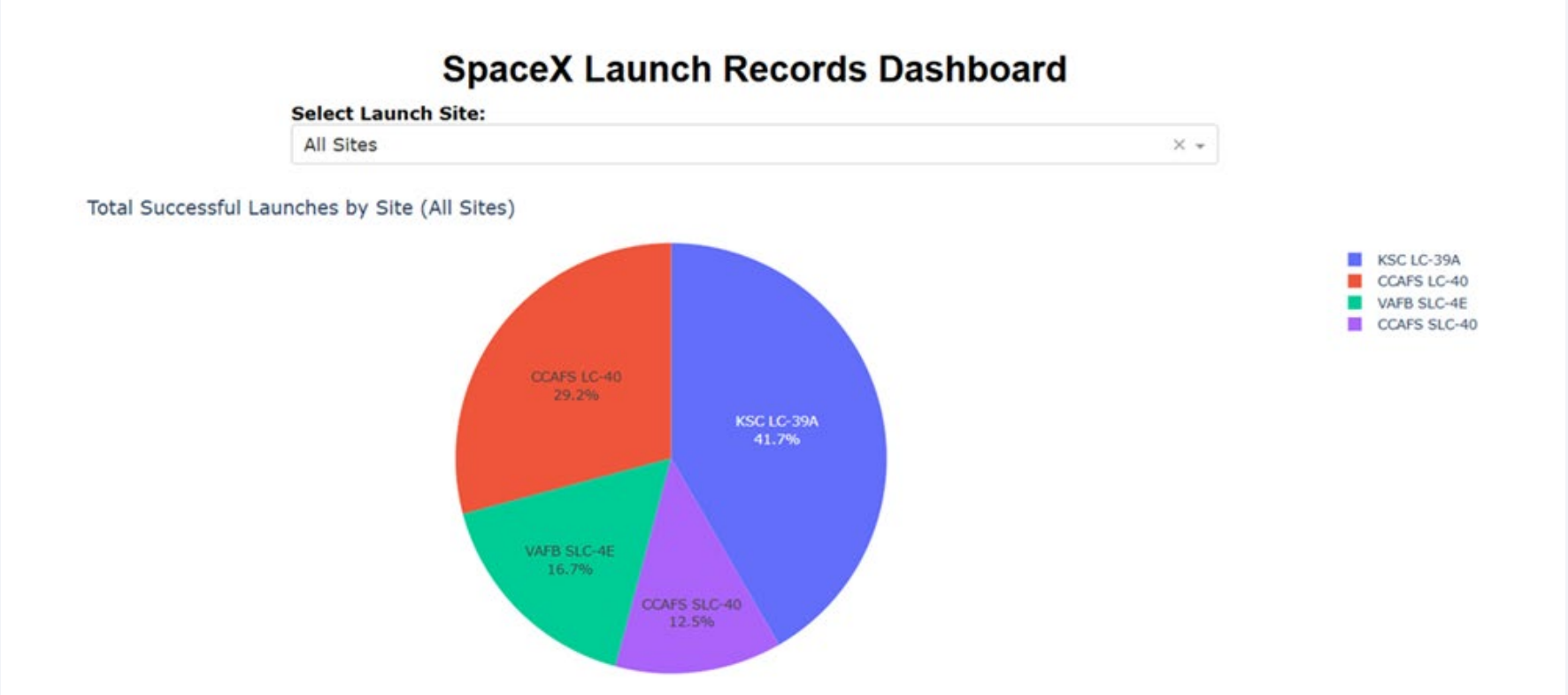
Proximities



- Explain the important elements and findings on the screenshot


Section 4
Build a Dashboard with Plotly Dash

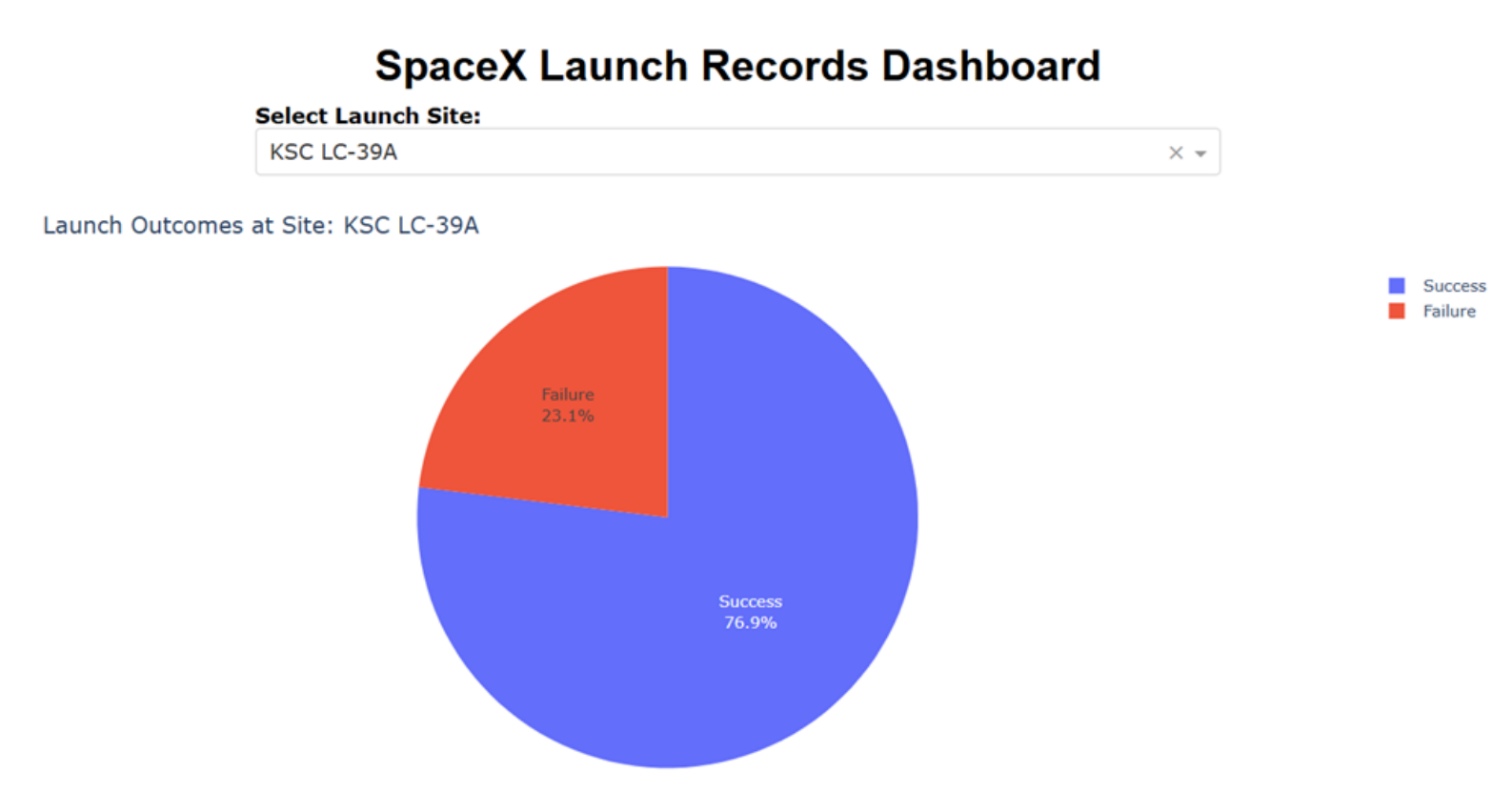
Total Successful Launches Pie Chart



Explanation:

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

Highest Launch Ratio

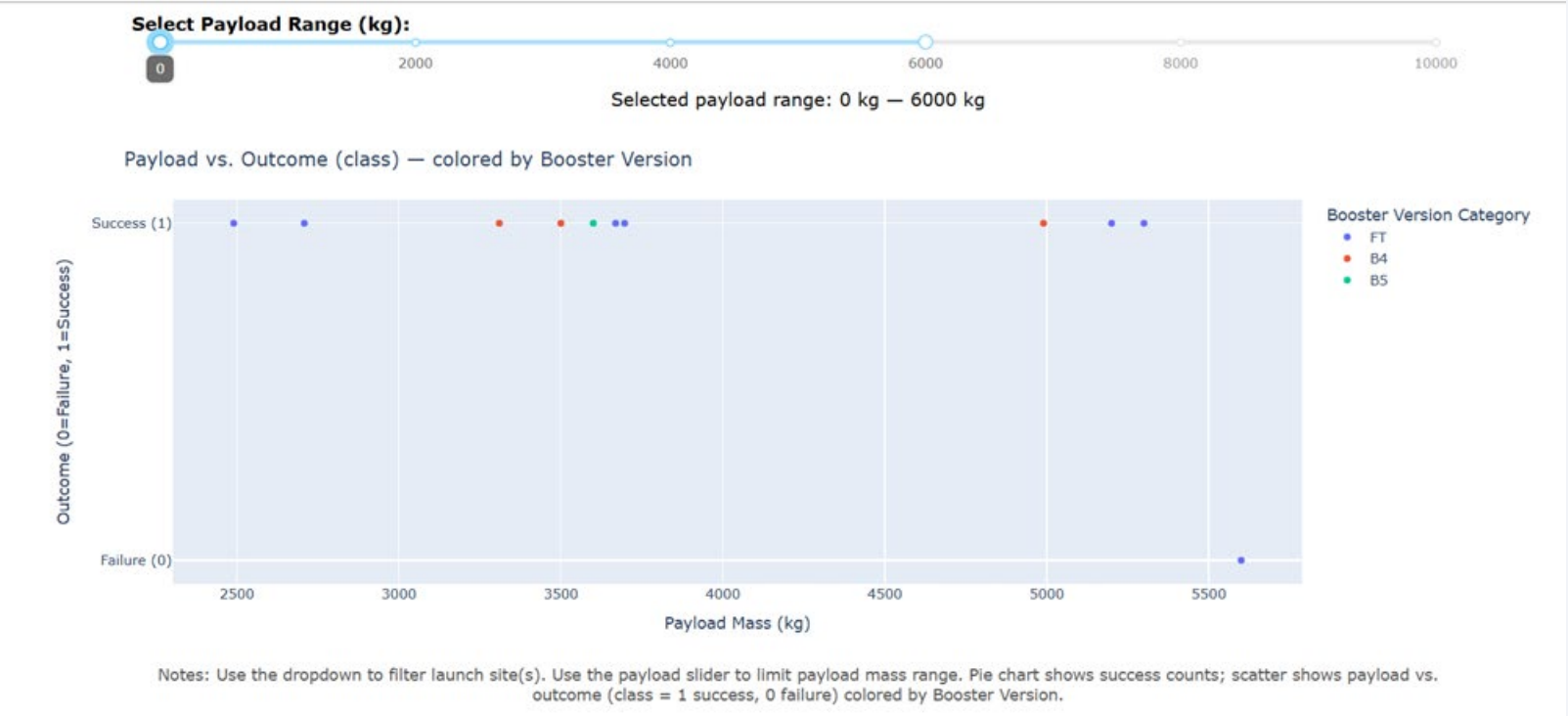


Explanation:

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

Payload Vs Outcome

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



Section 5

Predictive Analysis (Classification)

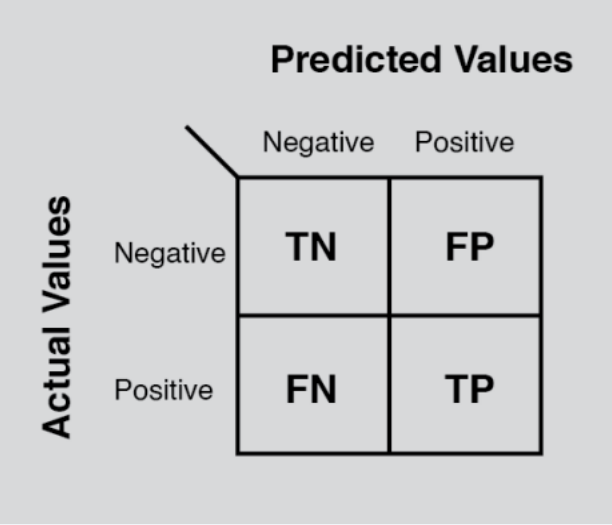
Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart

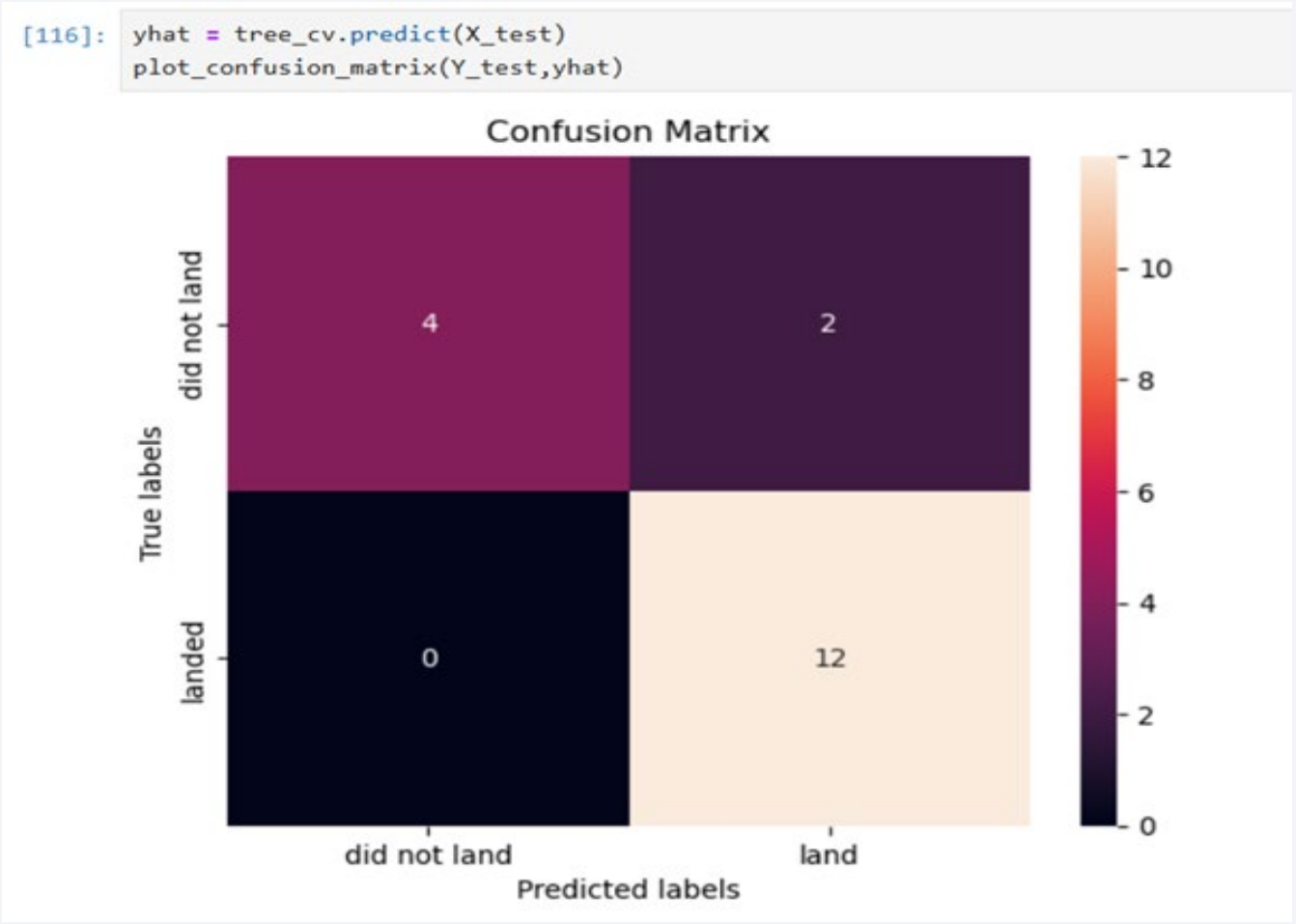
- Find which model has the highest classification accuracy

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.




```
[116]: yhat = tree_cv.predict(X_test)
       plot_confusion_matrix(Y_test,yhat)
```



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

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!