IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rafael Coba
May 31 3025

Outline

EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

RESULTS

CONCLUSION

APPENDIX

Executive Summary

Summary of Methodologies:

- Data Collection: Data on SpaceX launches was obtained using the SpaceX API. Web scraping was also performed.
- Data Cleaning: Data was processed to ensure proper formatting, including the creation of a “Class” column to indicate landing success or failure.
- Exploratory Data Analysis (EDA): Visualizations (scatter plots, bar charts, line charts) and SQL queries were used to explore patterns in the data, such as success rates by launch site and orbit.
- Interactive Analysis: Interactive maps were created with Folium to visualize launch locations, and a dashboard with Plotly Dash was built to show success rates and correlations with payload mass.
- Predictive Analysis: Classification models (Logistic Regression, SVM, Decision Trees, KNN) were constructed and evaluated to predict Falcon 9 landing success, with hyperparameter tuning using GridSearchCV.

Summary of Results:

- EDA: Patterns were found, such as higher success rates at certain launch sites (e.g., CCAFS) and orbits (e.g., GTO). The success rate has improved over time.
- Interactive Visualizations: The dashboard showed that sites like CCAFS have the highest success rates, and that heavier payloads tend to have different success rates depending on the rocket version.
- Predictive Models: The classification models achieved accuracies around 83-87% on test data. The Decision Trees model performed best, with an accuracy of 87.5% on validation data.

Introduction

- **Context and Background:**

SpaceX has revolutionized the aerospace industry by significantly reducing launch costs through the reuse of the Falcon 9's first stage. Predicting whether the Falcon 9's first stage will successfully land is crucial for estimating launch costs and enabling other companies to compete with SpaceX in bids.

- **Problems to Solve:**

Landing Prediction: Develop a predictive model to determine if the Falcon 9's first stage will successfully land based on historical launch data.

Factor Analysis: Identify which factors (launch site, orbit, payload mass, rocket version) most influence landing success.

Data Visualization: Create interactive visualizations (maps, dashboards) to communicate patterns in launch data, such as success rates by site or correlations with payload mass.

Cost Optimization: Provide useful insights for SpaceY to estimate launch costs and compete with SpaceX.

Section 1

Methodology

Methodology

- Executive Summary
- **Data Collection Methodology:**

SpaceX API: GET requests were used to retrieve launch data from the SpaceX API, including details such as flight number, launch site, orbit, payload mass, and landing outcome (success or failure).

Web Scraping: Additional data was gathered through web scraping, likely from sites such as Wikipedia or SpaceX-related pages, to supplement the API data.

- **Data Processing (Data Wrangling):**

The data from the API and scraping was cleaned to ensure it was in the correct format.

A **"Class"** column was created in the dataset to indicate whether the landing was successful (1) or failed (0).

Numeric variables were standardized using **StandardScaler** to prepare them for modeling.

- **Exploratory Data Analysis (EDA):**

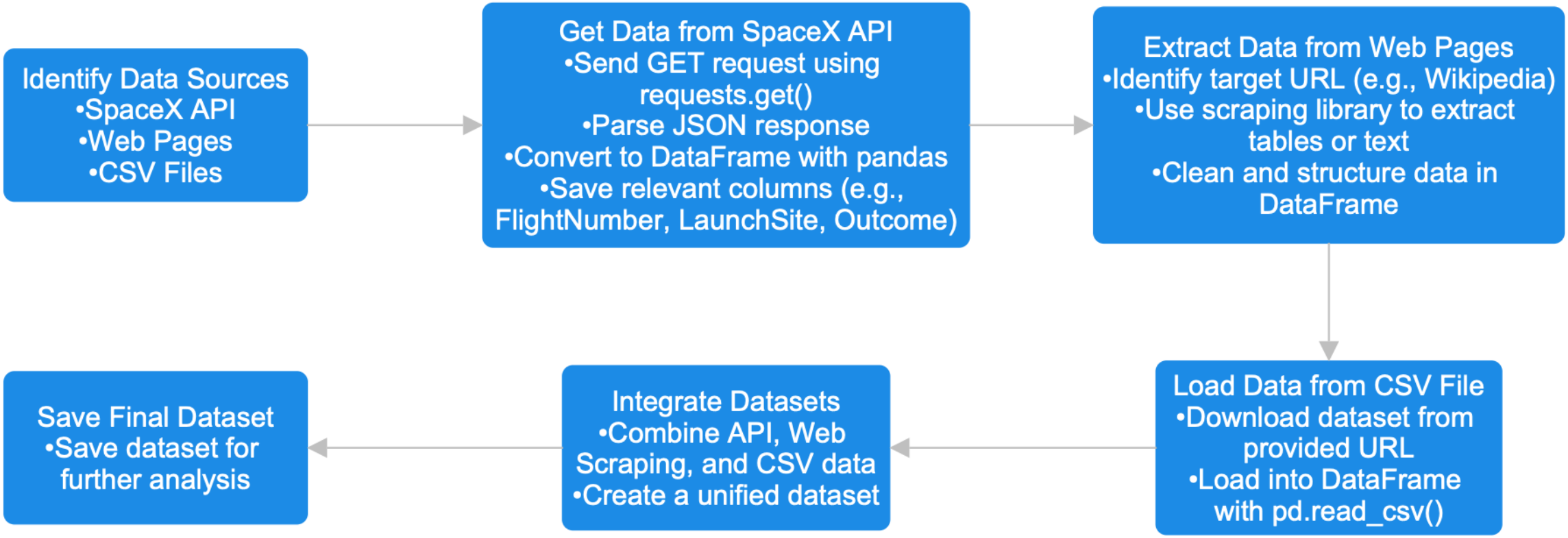
Visualization: Scatter plots (Flight Number vs. Launch Site, Payload vs. Orbit Type), bar charts (Success Rate vs. Orbit Type), and line charts (annual success trend) were generated to identify patterns.

SQL: SQL queries were used to answer specific questions, such as names of launch sites, total payload mass, and dates of successful landings.

Methodology

- **Interactive Visual Analysis:**
- **Folium:** Interactive maps were created with markers for launch sites, landing outcomes (success/failure), and distances to nearby landmarks like roads or coastlines.
- **Plotly Dash:** A dashboard was developed with a dropdown to select launch sites, a pie chart to show success rates, and a scatter plot to illustrate correlations between payload mass and success, with a slider to filter payload ranges.

- **Predictive Analysis (Classification):**
- **Models Built:** Four classification models were trained—Logistic Regression, SVM, Decision Trees, and KNN—to predict the "Class" column (success or failure of landing).
- **Hyperparameter Tuning: GridSearchCV** with cross-validation (cv=10) was used to find the best hyperparameters for each model.
- **Evaluation:** Models were evaluated using accuracy on test data, with the following results:
- Logistic Regression: **83.33%**
- SVM: **83.33%**
- Decision Trees: **72.22%**
- KNN: **83.33%**
- **Best Model:** Although Decision Trees achieved the highest validation accuracy (**87.5%**), its test data performance was lower (**72.22%**). Logistic Regression, SVM, and KNN shared the same test accuracy (**83.33%**), but Logistic Regression is more interpretable and efficient.

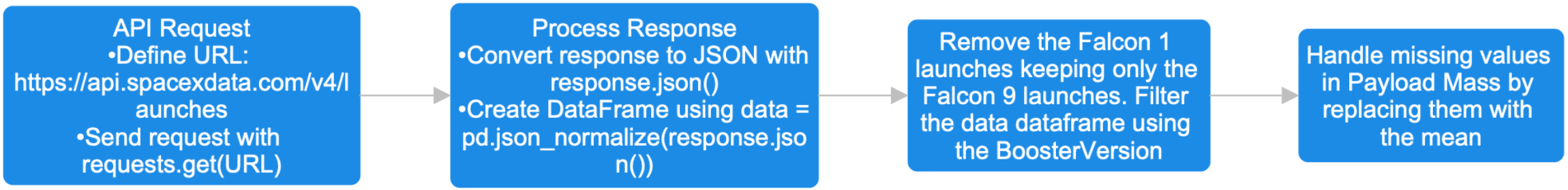


Data Collection

Process:
HTTP GET requests were sent to the SpaceX API endpoints (e.g., **/launches**) using the **requests** library. The JSON response was parsed into a pandas DataFrame, extracting relevant columns such as **flight_number**, **date_utc**, **launchpad**, **payloads**, **cores** (which contains landing information), and **success**.

Key Phrases:

- GET request to /launches to retrieve all launches.
- Parse JSON with pandas to create a DataFrame.
- Select relevant columns: flight_number, launch_site, landing_success.
- Handle missing or inconsistent data.

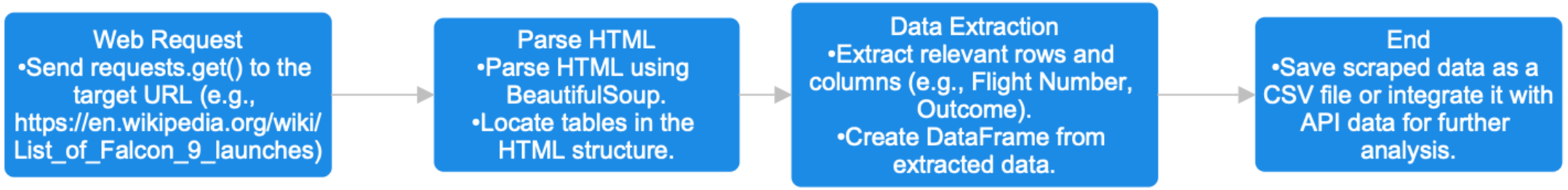


Data Collection – SpaceX API

https://github.com/Rcob24/ADSSpec/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Key Phrases:

- Identify URL with SpaceX data (e.g., Wikipedia).
- Use BeautifulSoup to parse HTML and extract tables.
- Convert scraped data into a pandas DataFrame.
- Clean data to align it with the API format.



Data Collection – Scraping

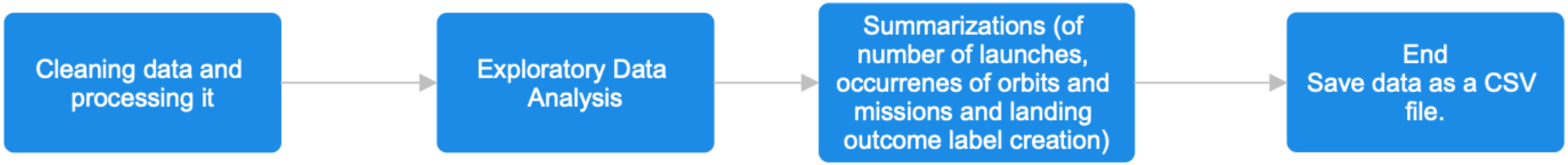
https://github.com/Rcob24/ADSSpec/blob/main/jupyter-labs-webscraping.ipynb

Cleaning: Null values were removed, inconsistent formats were corrected, and columns were renamed for consistency (e.g., converting **cores.landing_success** to **Class** with values **0/1** for failure/success).

Variable Creation: A **Class** column was added to indicate landing outcome (**1** for success, **0** for failure), based on data from the API or scraping.

Standardization: Numeric variables (e.g., **PayloadMass**, **Flights**) were standardized using **StandardScaler** to prepare the data for modeling.

EDA: Computation of number of launches, success rate, landing regions, orbit types and occurrences



Data Wrangling

https://github.com/Rcob24/ADSSpec/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Charts Created and Reasons:

- **Flight Number vs. Launch Site (Scatter Plot):** Shows the distribution of launches by site over time (flight number). Used to identify patterns in the frequency of each site's usage.
- **Payload vs. Launch Site (Scatter Plot):** Analyzes how payload mass varies by launch site. Used to detect if certain sites handle heavier payloads.
- **Success Rate vs. Orbit Type (Bar Chart):** Compares the landing success rate by orbit type (e.g., GTO, ISS). Used to identify which orbits have higher success rates.
- **Flight Number vs. Orbit Type (Scatter Plot):** Explores the relationship between flight number and orbit type. Used to see if orbit types change over time.
- **Payload vs. Orbit Type (Scatter Plot):** Examines the relationship between payload mass and orbit type. Used to detect correlations between payload and orbit.
- **Launch Success Yearly Trend (Line Chart):** Shows the annual trend of landing success rates. Used to evaluate SpaceX's performance improvements over time .

General Reasons:

- **Scatter plots** were used to explore bivariate relationships and detect visual patterns.
- **Bar charts** were used to compare categorical metrics (like success rates).
- **Line charts** were used to visualize temporal trends.
- These visualizations helped identify key factors (site, orbit, payload) that influence landing success.

https://github.com/Rcob24/ADSSpec/blob/main/edadataviz.ipynb

EDA with SQL

Performed SQL Queries:

- **Unique Launch Site Names**
- **Records of Sites Starting with 'CCA'**
- **Total Payload Mass for NASA**
- **Average Payload Mass for F9 v1.1**
- **Date of First Successful Ground Landing**
- **Boosters with Successful Drone Ship Landing and Payload Between 4000-6000 kg**
- **Total Number of Successful and Failed Missions**
- **Boosters with Maximum Payload Mass**
- **Records from 2015 with Failed Drone Ship Landings**
- **Ranking of Landing Outcomes (2010-06-04 to 2017-03-20)**

https://github.com/Rcob24/ADSSpec/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Map Objects Added:
- Circles: Represent locations of launch sites and highlight the approximate area around each launch site.
- Markers with labels: Are used to show site names on the map.
- Markers: Used to show the center of launch site.
- Popups: To provide interactive details when clicking on circles
- Lines: To measure distances between points

https://github.com/Rcob24/ADSSpec/blob/main/lab_jupyter_launch_site_location.ipynb

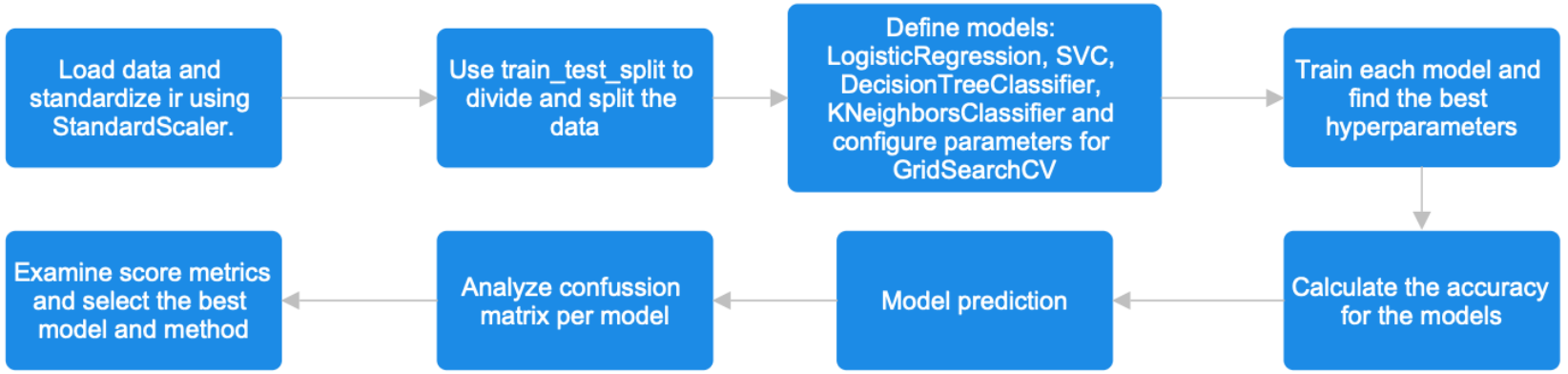
Build a Dashboard with Plotly Dash

- **Charts and Interactive Elements Added:**
- **Dropdown (Dropdown Menu):** A menu to select a specific launch site.
- **Pie Chart:** Shows the distribution of successful launches by site (for all sites) or the proportion of success/failure for a specific site.
- **Range Slider:** Allows filtering of data by payload mass range.
- **Scatter Plot:** Visualizes the relationship between payload mass and landing outcome (**class**, 0 for failure, 1 for success)
- **Reasons for Adding These Elements:**
- **Dropdown:** Allows users to analyze success rates for specific launch sites or at a global level, facilitating comparisons between sites.
- **Pie Chart:** Provides a quick view of the proportion of successful launches, useful for identifying which sites have better performance.
- **Range Slider:** Enables exploration of how payload mass affects landing success, filtering data for specific ranges (e.g., heavy vs. light payloads).
- **Scatter Plot:** Reveals correlations between payload mass, landing success, and rocket version, helping to identify patterns (e.g., certain versions achieve higher success with heavy payloads).

https://github.com/Rcob24/ADSSpec/blob/main/Dashboard.py

Key Phrases:

- Extract **Class** as **Y** and standardize features **X**.
- Split data into training (80%) and testing (20%).
- Optimize hyperparameters with **GridSearchCV** (**cv=10**).
- Evaluate models on test data using **score**.
- Select the model with the best balance of accuracy and simplicity.



Predictive Analysis (Classification)

https://github.com/Rcob24/ADSSpec/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

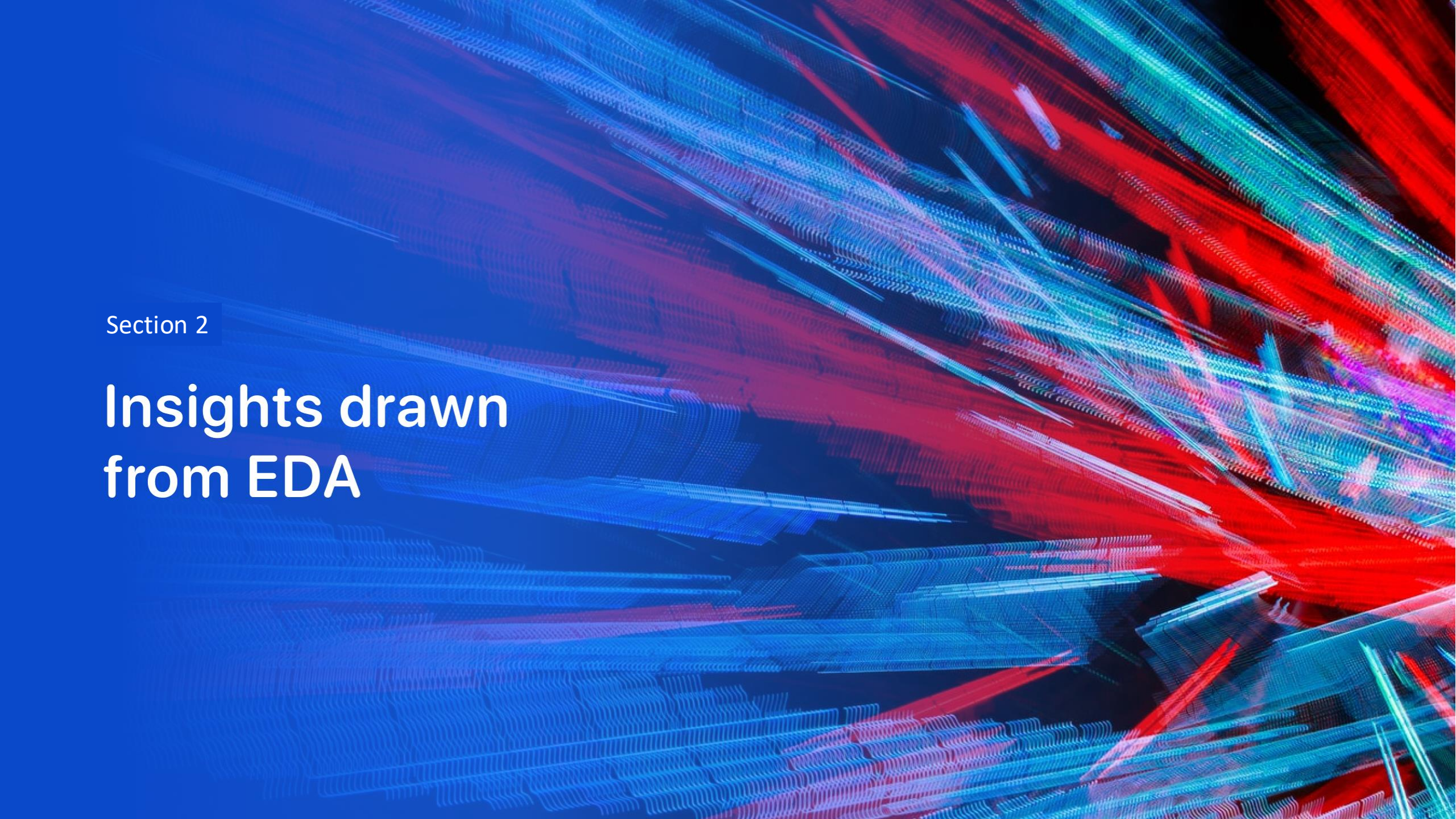
Results

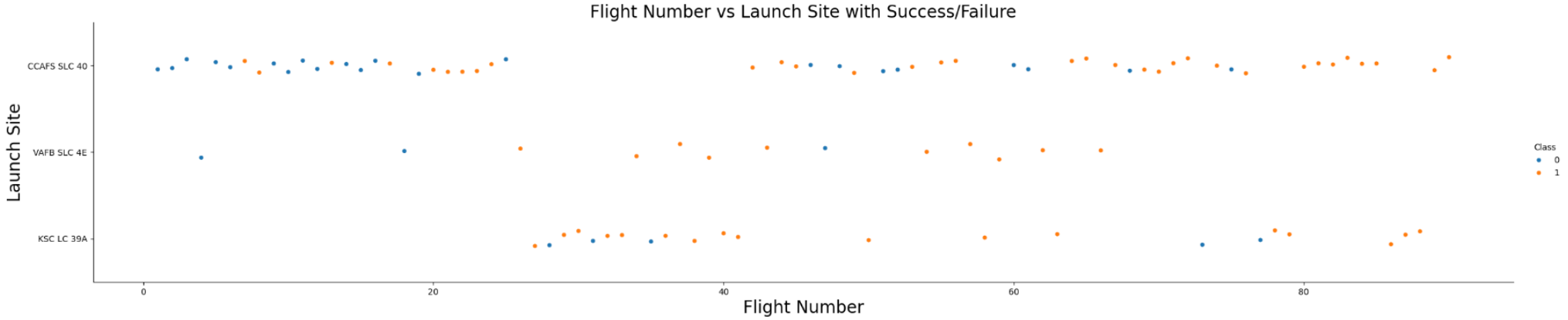
- Greatest success rate was achieved in 2019.
- Orbits with higher success rate are ES-L1, GEO, HEO and SSO (100%). SO has a 0 % success rate. GTO a 50%, ISS a 60%, MEO and PO approximately 65% and LEO a 70% success rate.
- Success rate is different on every launch site.
- Average payload mass carried by booster version F9 v1.1 was 2928.4 kg
- Success rate kept increasing every year (except for 2018) till 2020.

Accuracy on test data (methods/classifiers)

- Logistic Regression: 83.33%
- SVM: 83.33%
- Decision Tree: 72.22%
- KNN: 83.33%

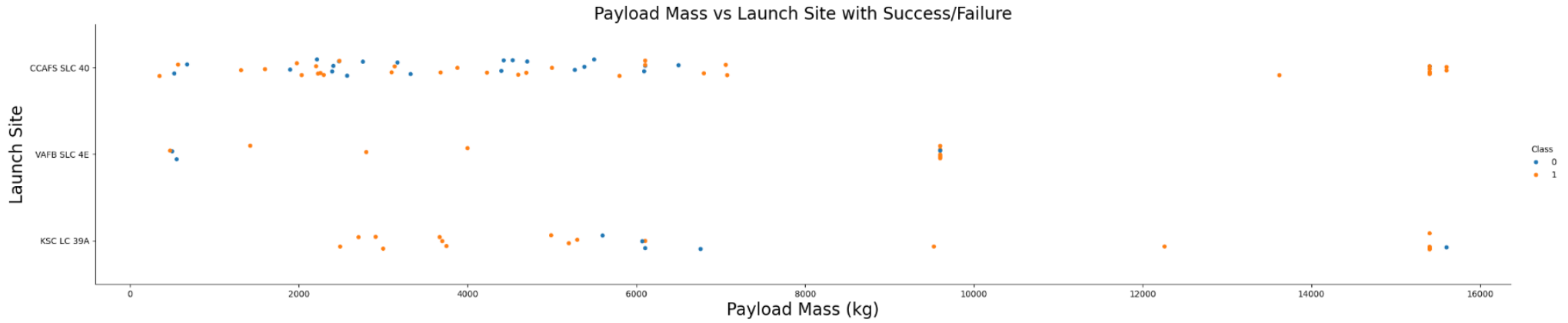
Landing_Outcome	outcome_count
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2


Section 2
Insights drawn from EDA



Flight Number vs. Launch Site

- Most of launches are made in CCAFS SLC 40 Launch site.
- 13 last launches has succeeded, while first 6 failed.
- While VAFB SLC 4E and KSC LC 39A have higher success rates, CCAFS SLC 40 is where most of recent launches were successful.

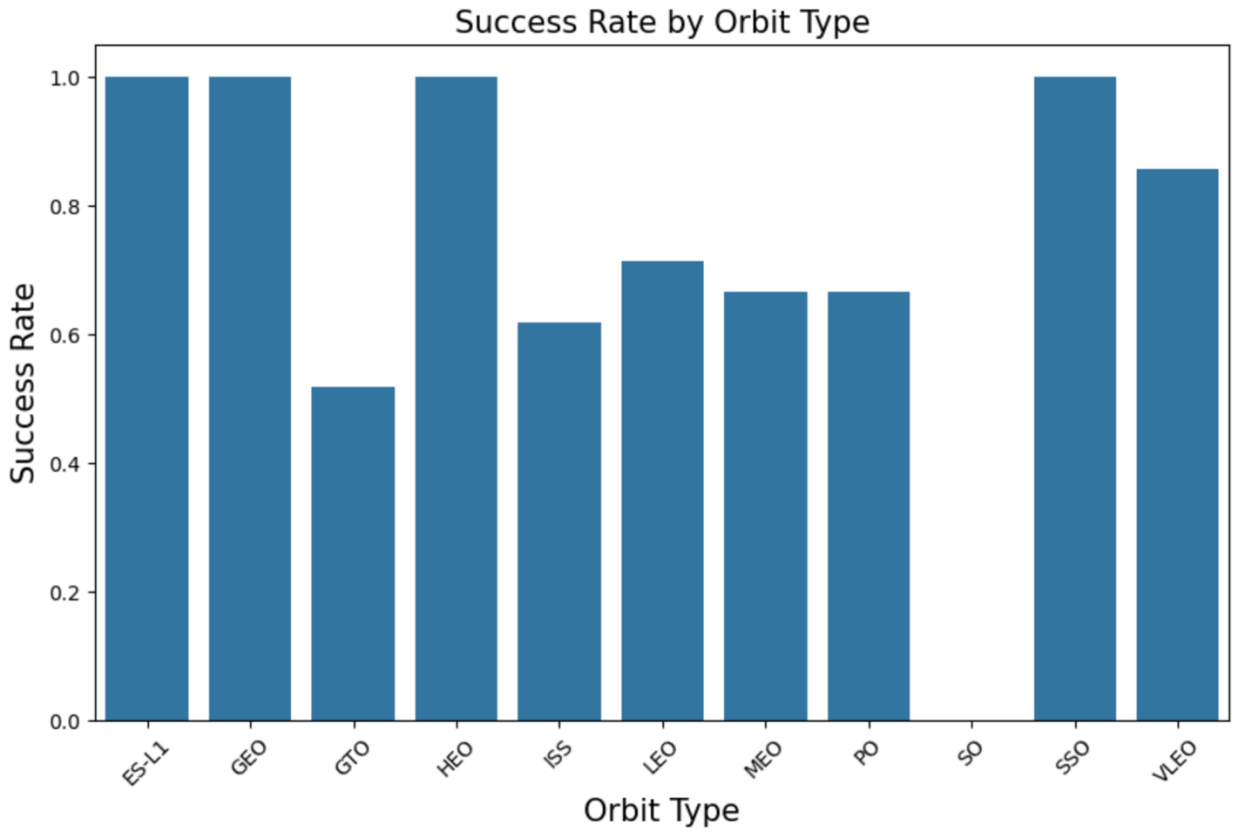


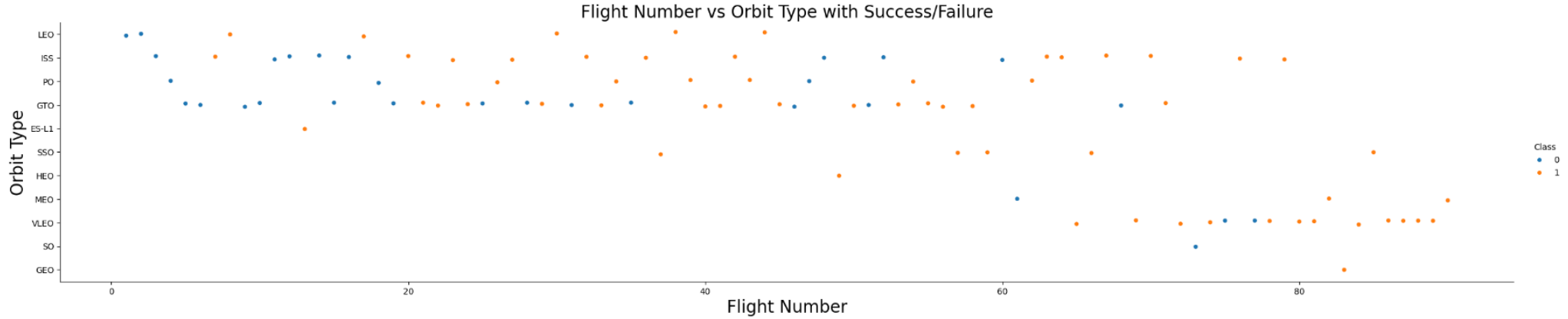
Payload vs. Launch Site

- Over 8000kg of payload mass, success rate is giant.
- There are no launches greater than 10000kg in mass on VAFB SLC 4E
- There is a 100% success rate with a pyload mass of 7000 kg or more in CCAFS SLC 40

Success Rate vs. Orbit Type

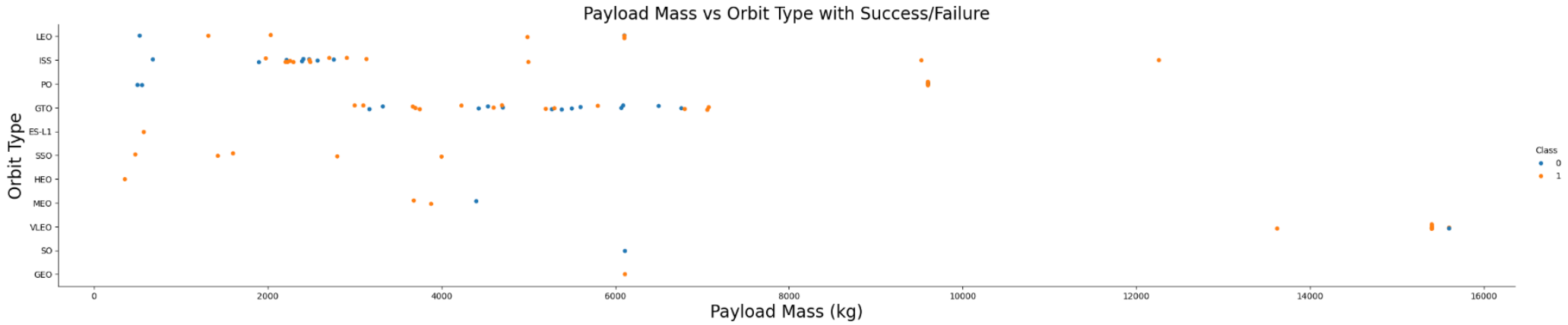
- Orbits with higher success rate are ES-L1, GEO, HEO and SSO (100%). SO has a 0 % success rate. GTO a 50%, ISS a 60%, MEO and PO approximately 65%, LEO a 70% success rate and VLO approximately 85%





Flight Number vs. Orbit Type

- Success rate improved over time
- Some of the orbits are no longer being in function or function sporadically
- The more flight number increases, there is a better success rate on LEO orbit.

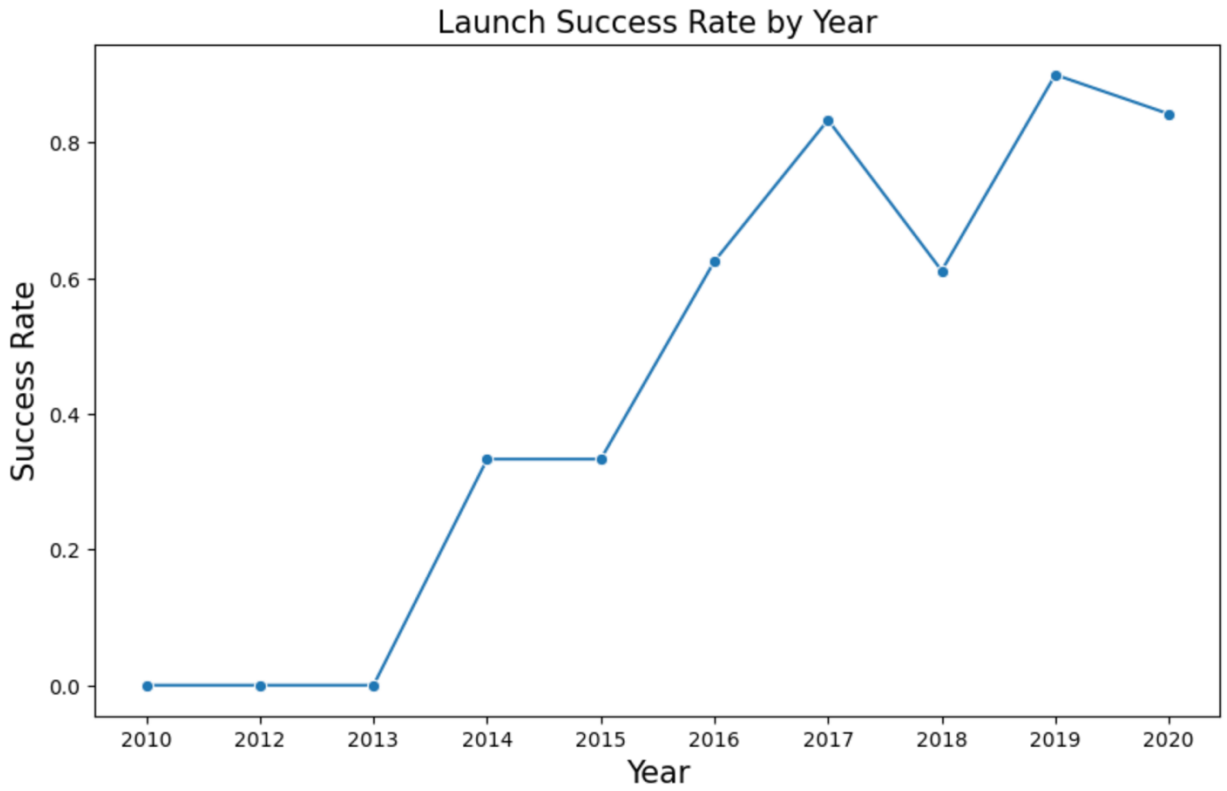


Payload vs. Orbit Type

- There are few launches to SO, MEO and GEO.
- Heavy payloads have a better success rate on LEO and Polar orbits. Light payloads on MEO, SSO and ES-L 1.

Launch Success Yearly Trend

- Success rate increased every year, except for 2018, until 2020.
- Maximum Launch Success rate was in 2019.
- Between 2010 and 2013 there was a success rate of 0%



All Launch Site Names

- The query shows all names of distinct launch sites inside launch_site from the table

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

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
```

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

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

Launch Site Names Begin with 'CCA'

- There are 5 records with launch site names that begin with ”CCA”

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

```
%sql SELECT SUM("Payload_Mass__kg_") AS total_mass FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)'
```

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

total_mass
45596

Total Payload Mass

- Query sums all payloads which code contains ’NASA (CRS)’, giving a total payload mass of 45596kg

```
%sql SELECT AVG("Payload_Mass__kg_") AS avg_mass FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1'
```

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

avg_mass
2928.4

Average Payload Mass by F9 v1.1

- Average payload mass by F9 v1.1 is 2928.4kg

```
%sql SELECT MIN("Date") AS first_success FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)'
```

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

first_success
2015-12-22

First Successful Ground Landing Date

- First successful ground landing was on December 12, 2022.
- Query filters the date to the first one that was a success on ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query filters unique values from the table that meet these two conditions

- Query is: **%sql** SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "Payload_Mass__kg_" > 4000 AND "Payload_Mass__kg_" < 6000

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

```
%sql SELECT "Mission_Outcome", COUNT(*) AS outcome_count FROM SPACEXTABLE GROUP BY "Mission_Outcome"
```

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

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

Total Number of Successful and Failure Mission Outcomes

- As for mission outcome, there were 99 successes, 1 success with payload status unclear and 1 failure.

Boosters Carried Maximum Payload

- Names of booster versions which carried maximum payload mass
- Query is **%sql** SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Payload_Mass__kg_" = (SELECT MAX("Payload_Mass__kg_") FROM SPACEXTABLE)

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

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

2015 Launch Records

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 and the respective month
- Query is %sql SELECT substr("Date", 6, 2) AS month, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Failure (drone ship)' AND substr("Date", 0, 5) = '2015

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

- Ranking of 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

- Query is: %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

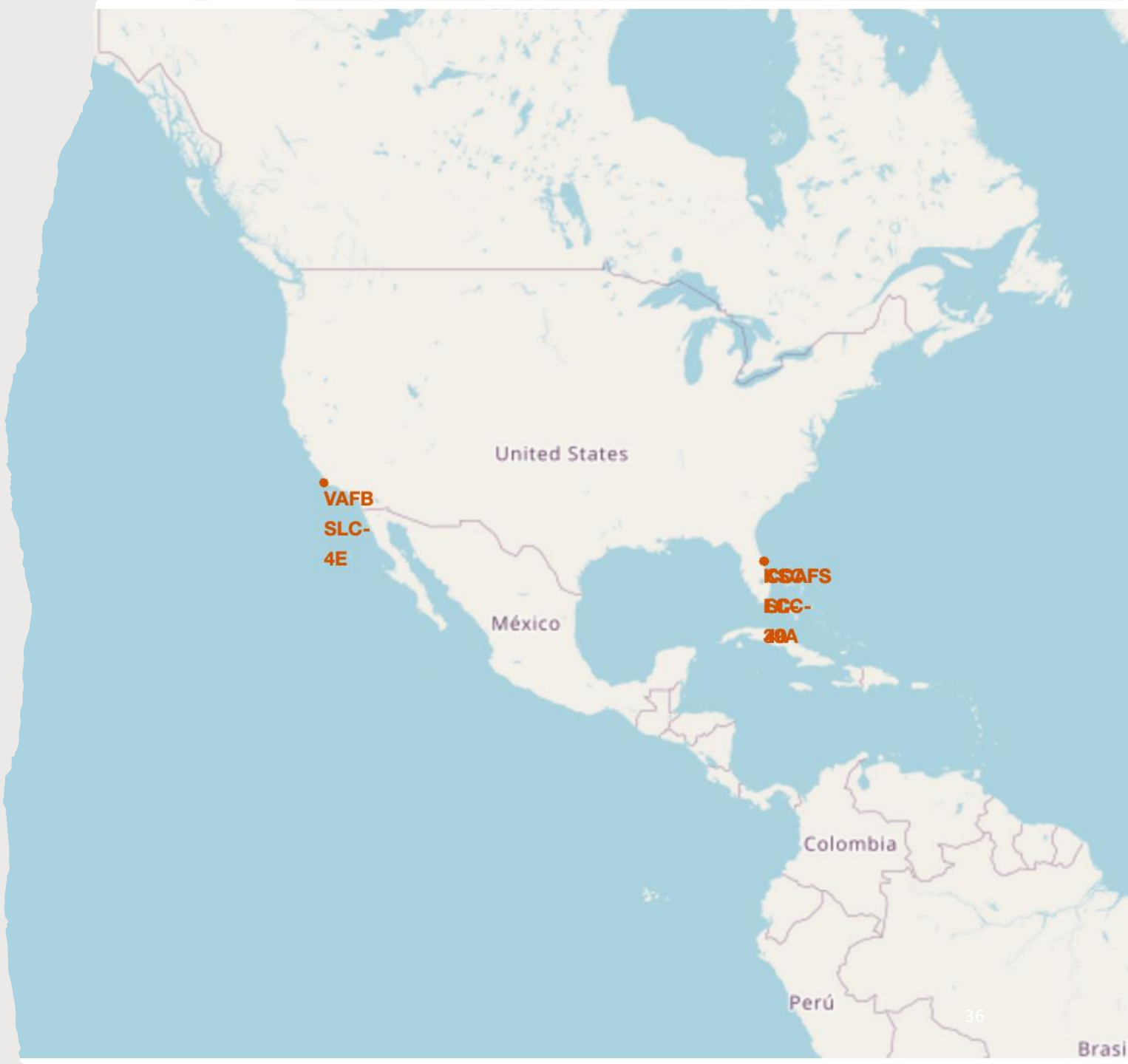
Landing_Outcome	outcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

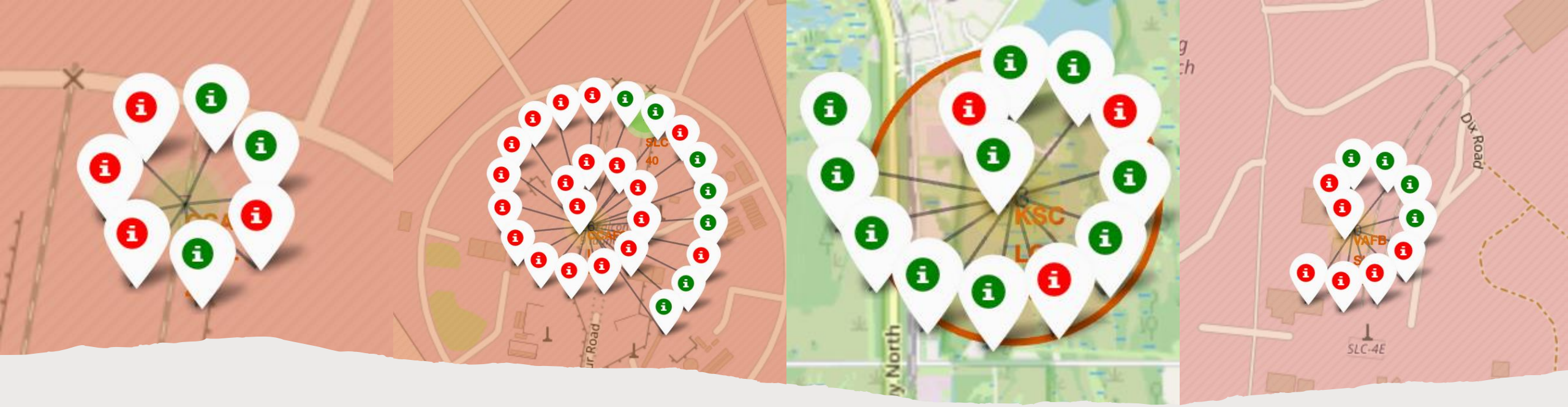
Section 3

Launch Sites Proximities Analysis

Launch Sites

- All launch sites are near oceans/coasts but not far away from roads.
- Florida and California are the states where there are aggrupation of launch sites.



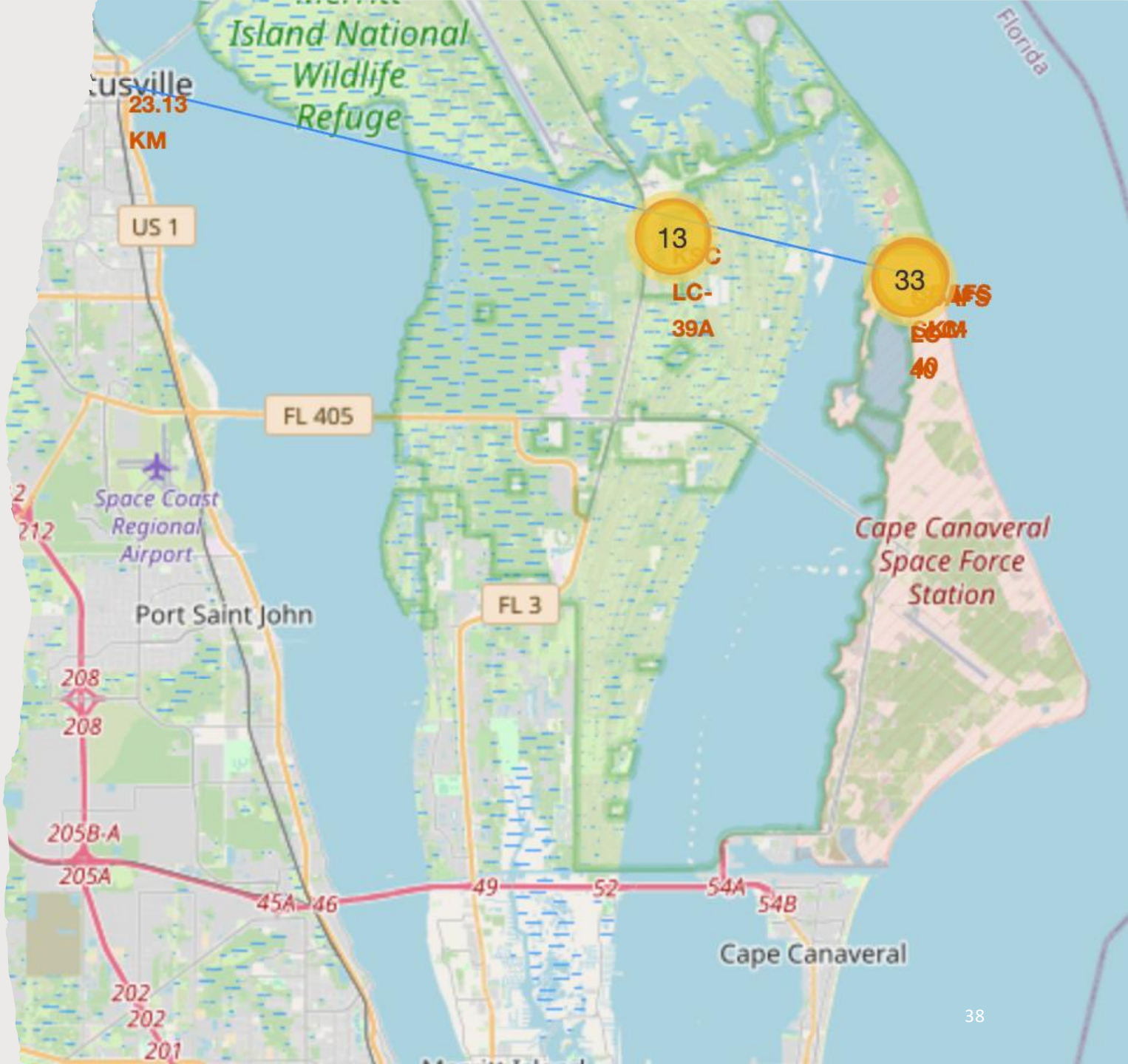


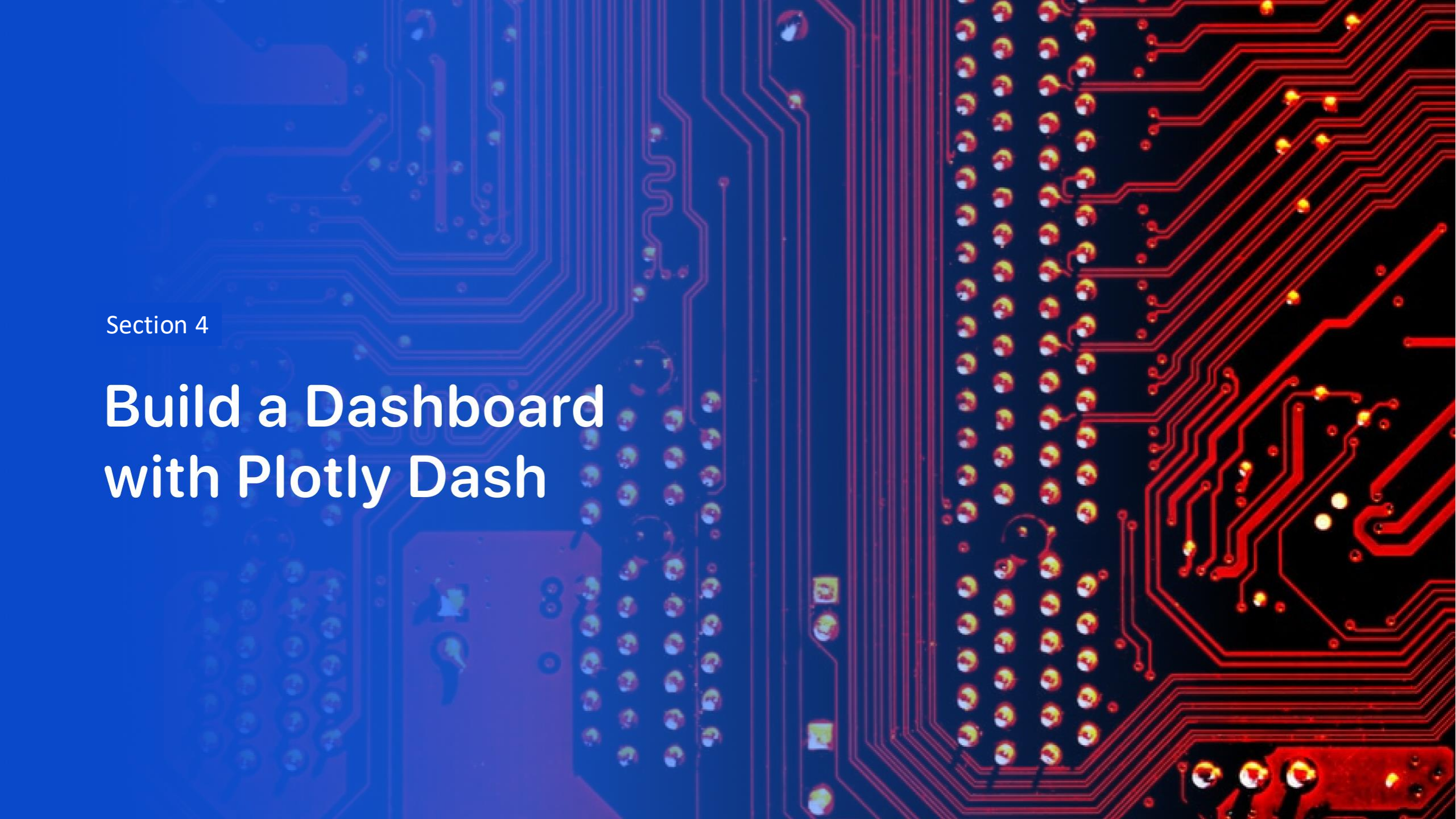
Launch Outcomes Record

- Green marker indicates successful launch, whereas red marker indicates a failed launch.
- KSC LC-39 A has a high success rate.

Safety and Distances

- CCAFS SLC-40 is 23.13 km away from a city like Titusville, but KSC LC-39A is only 16.32km away from the city. So, it could be dangerous for its residents if something goes wrong.
- CCAFS SLC-40 is also near to the coast.




Section 4
Build a Dashboard with Plotly Dash

All Sites

Total Success Launches by Site

Payload ran

Launch success count for all sites

- KSC LC-39A has the most successful launches.
- Success launch variates from the place the launch is done.
- Second site with most success launches is CCAFS LC-40

KSC LC-39A

Total Success Launches for site KSC LC-39A

Payload range (Kg):

Site with highest launch success ratio

- KSC LC-39A has a 76.9% success rate and a 23.1% fail rate.
- KSC LC-39A is the site with highest launch success ratio.

Payload vs Launch Outcome

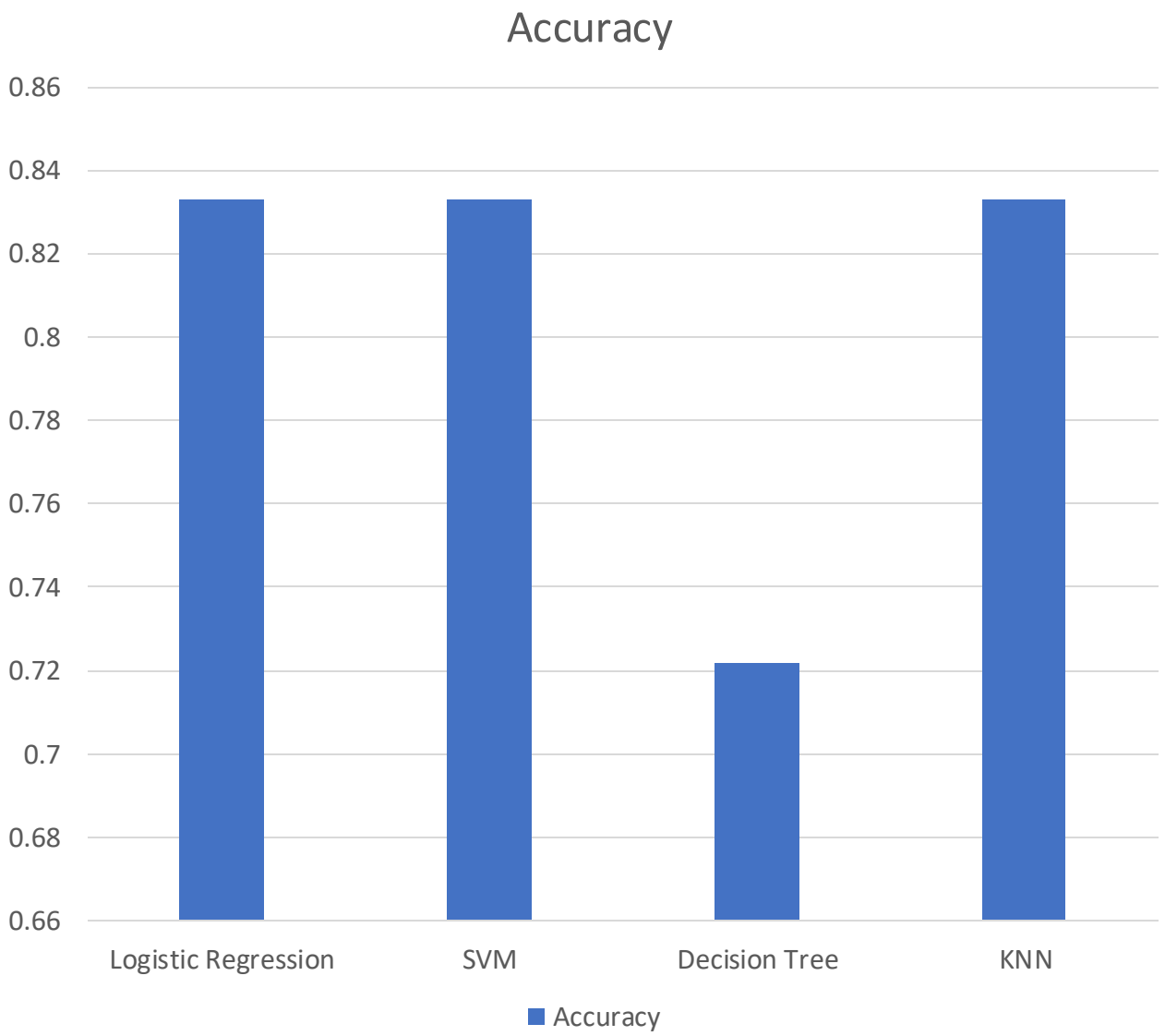
- This plot shows the range between 2000kg and 1000kg of payload mass. It shows that the ones that are successful tend to be below 5300kg and use the FT booster

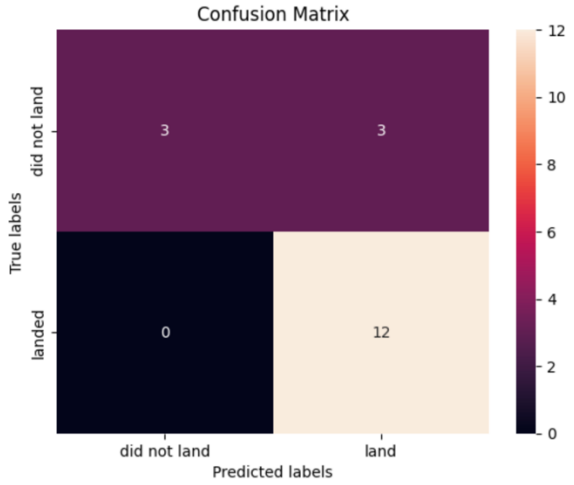
Section 5

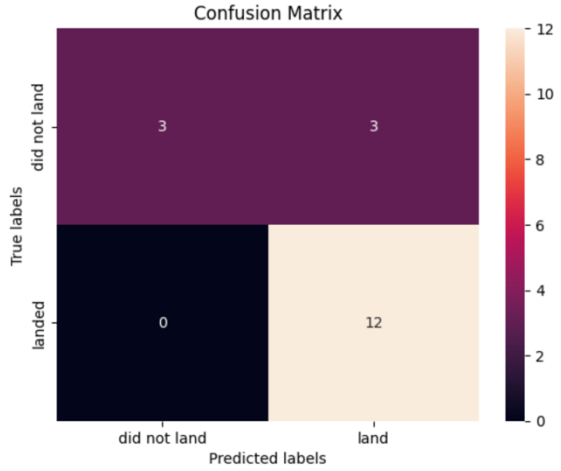
Predictive Analysis (Classification)

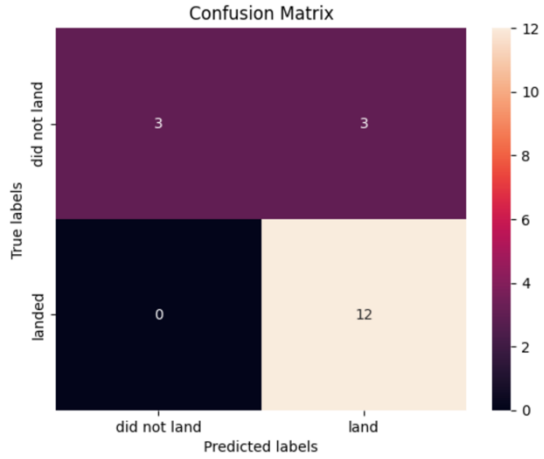
Classification Accuracy

- In this case, the model with the least accuracy was decision tree. As for the one with the most accuracy, Logistic Regression, SVM and KNN had the same accuracy (0.833)









Confusion Matrix

It is possible to see that the confusion matrices for KNN, SVM and Logistic Regression are essentially the same

Conclusions

Key Factors Influencing Landing Success:

- Simple orbits and moderate payloads (2000–5300kg) have higher success rates, especially with modern boosters like the ones from the F series.

Predictive Models are Effective:

- Logistic Regression predicts landings with 83.33% accuracy, offering a simple and reliable approach to estimating launch costs.

Interactive Visualizations Facilitate Analysis:

- Folium maps and the Plotly Dash dashboard reveal geographic and operational patterns, such as the dominance of KSC LC-39A and proximity to coastlines.

- **Launch success tends to increase every year.**

Thank you!