IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Franklin A. Ubiera
10/2025

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

SpaceY, a new and ambitious rocket launch company, aims to offer more competitive prices in the space industry. To achieve this goal, the company is conducting an in-depth study of its main competitor, **SpaceX**, to understand the key factors that enable **SpaceX** to maintain the most cost-effective launch prices in the market.

SpaceX's data will be used to answer questions through exploration, analysis, and machine learning algorithms. These methods will facilitate the process of obtaining insights from data.

Section 1

Methodology

Methodology

Executive Summary

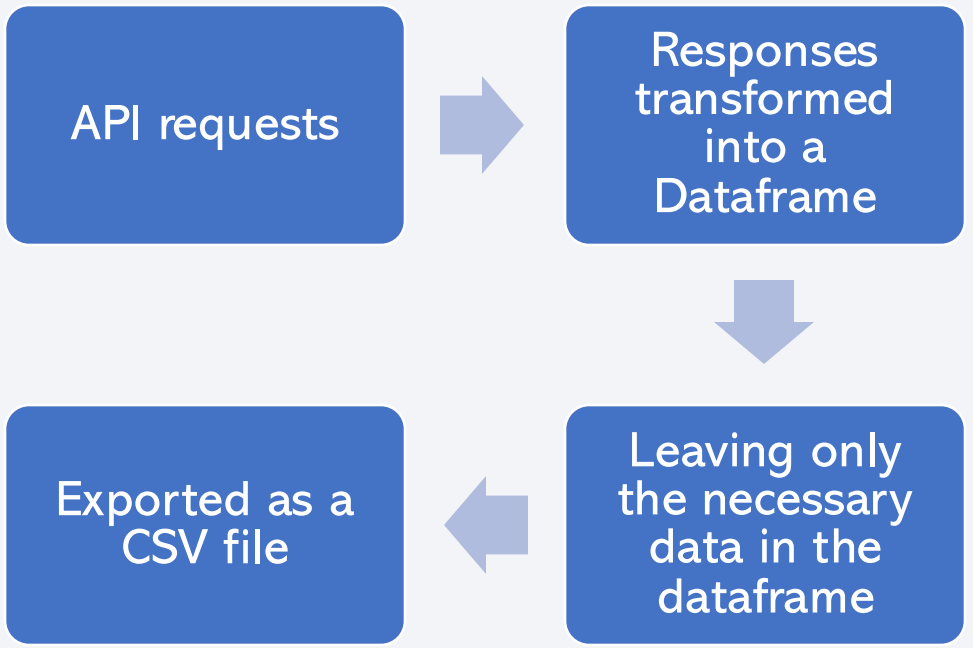
- Data collection methodology:
  - The data was collected by sending requests to the SpaceX API and scraping Wikipedia pages containing public SpaceX records.
- Perform data wrangling
  - An exploratory analysis of the data was performed to identify missing values and data types, as well as useful patterns, in order to create a new output feature that determines whether the outcome of a landing is positive or negative.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection – SpaceX API

1. Launch data was requested from the SpaceX API and converted into a **pandas** DataFrame.
2. Some columns contained **IDs**, which were used to make additional API requests.
3. The retrieved information was stored in **lists**, then converted into **dictionaries**, and finally used to replace the IDs in the DataFrame.

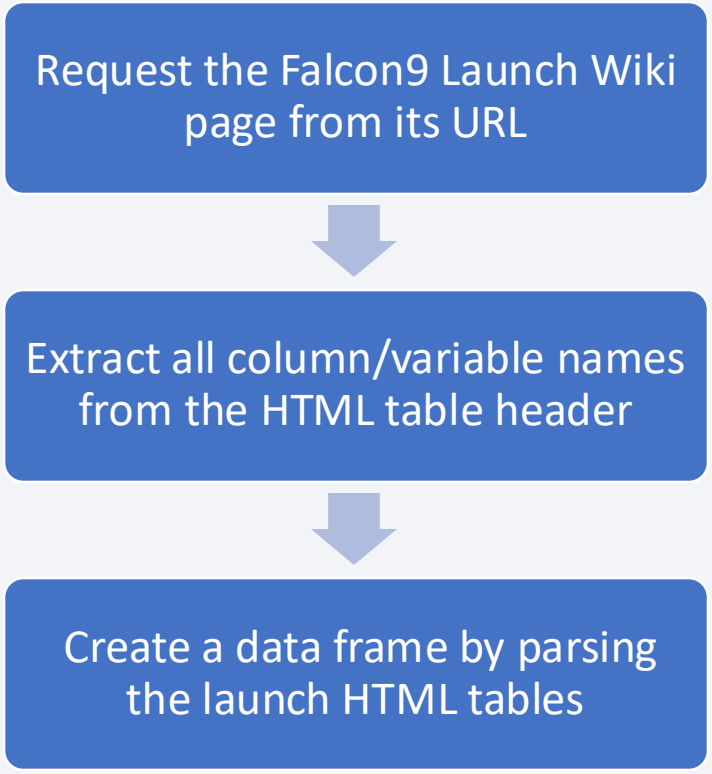
SpaceX API calls notebook:

https://github.com/AbelU1999/Applied_Data_Science_Capstone/blob/0ca2a086eb5f9f17523ee321312447956b213fc0/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- The Falcon 9 launch records were extracted from a Wikipedia table using web scraping.
- This data was parsed to create a DataFrame, and then exported as a CSV file.
- Web scraping notebook: https://github.com/AbelU1999/Applied_Data_Science_Capstone/blob/c732bcb19a06eb4e1f159579777cdd5b9417a8d9/jupyter-labs-webscraping-bak-2025-09-08-15-32-56Z.ipynb



Data Wrangling

Positives and Negatives outcomes were separate, and then created a class column with assigned labels 1 (positive), 0 (negative)

- Once the dataframe has been loaded, it was checked for missing values, columns data types, and landing occurrences per launch site and orbit with their outcomes.
- The outcomes were separated into positive and negative, and then classes (1,0) were created.

- Data wrangling related notebooks: https://github.com/AbelU1999/Applied_Data_Science_Capstone/blob/0ca2a086eb5f9f17523ee321312447956b213fc0/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Catplot: Categorical plots were used to visualize the relationship between different variables, and understand how it affect the outcome.
- Bartplot: With a bar chart we visualize the success rate of landings per orbits.
- Lineplot: Thanks to a line plot we observed how the success rate has increased yearly

EDA with data visualization
notebook: https://github.com/AbelU1999/Applied_Data_Science_Capstone/blob/0ca2a086eb5f9f17523ee321312447956b213fc0/edadataviz.ipynb

EDA with SQL

- Fetching the unique launch sites

```
SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

- Showing 5 records from Cape Canaveral

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

- Total sum of payload mass kg for NASA

```
SELECT sum("PAYLOAD_MASS__KG_") AS "Total_NASA_PLM_KG" FROM SPACEXTABLE WHERE
"Customer" = "NASA (CRS)";
```

- Average payload mass carried by F9 v1.1 booster

```
SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Booster_Version" = "F9 v1.1";
```

EDA with SQL

- Date of the first successful landing in a ground pad

```
select MIN(Date) from SPACEXTABLE where "Landing_Outcome" = "Success (ground pad)";
```

- Names of the boosters which have success in drone ship and have payload mass between 4K and 6K

```
select "Booster_Version" from SPACEXTABLE where "Landing_Outcome" = "Success (drone
ship)" and "PAYLOAD_MASS__KG_" between 4000 and 6000;
```

- Total number of success and failed mission outcomes

```
select (select count("Mission_Outcome") from SPACEXTABLE where "Mission_Outcome" like
"Success%%") as Successful, (select count("Mission_Outcome") from SPACEXTABLE where
"Mission_Outcome" like "Failure%%") as Failure from SPACEXTABLE limit 1;
```

EDA with SQL

- Booster that had carried the maximum payload mass

```
SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT
MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);
```

- Failures landing in drone ship for the months in 2015

```
SELECT substr(Date, 6, 2) AS Months, (SELECT "Landing_Outcome" FROM SPACEXTABLE WHERE
"Landing_Outcome" = "Failure (drone ship)") AS landing_outcomes, "Booster_Version",
"Launch_Site"

    FROM SPACEXTABLE

    WHERE substr(Date, 0, 5) = '2015';
```

EDA with SQL

- Rank of count of landing outcomes between 2010-06-04 and 2017-03-20,in descending order.

```
SELECT Date, "Landing_Outcome", COUNT("Landing_Outcome") AS "COUNT"
 FROM SPACEXTABLE
 WHERE Date > '2010-06-04' AND Date < '2017-03-20'
 GROUP BY "Landing_Outcome" ORDER BY "COUNT" DESC;
```

- EDA with SQL notebook: https://github.com/AbelU1999/Applied_Data_Science_Capstone/blob/938faa47d6d2c6f9fe7a4468e3a505fe97c1677f/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Circles with labels were used to identify each launch site, and a marker cluster was included to indicate success or failure. A green marker indicated a successful launch, while a red marker indicated a failed launch.
- The distance between the base and its surroundings (cities, railways, coastline, etc.) was measured using a line and a label indicating the approximate distance.

Interactive map with Folium map
Notebook: https://github.com/AbelU1999/Applied_Data_Science_Capstone/blob/2c34b919150e8628ce4b2a68a51baadf54d06935/lab_jupyter_launch_site_location.ipynb

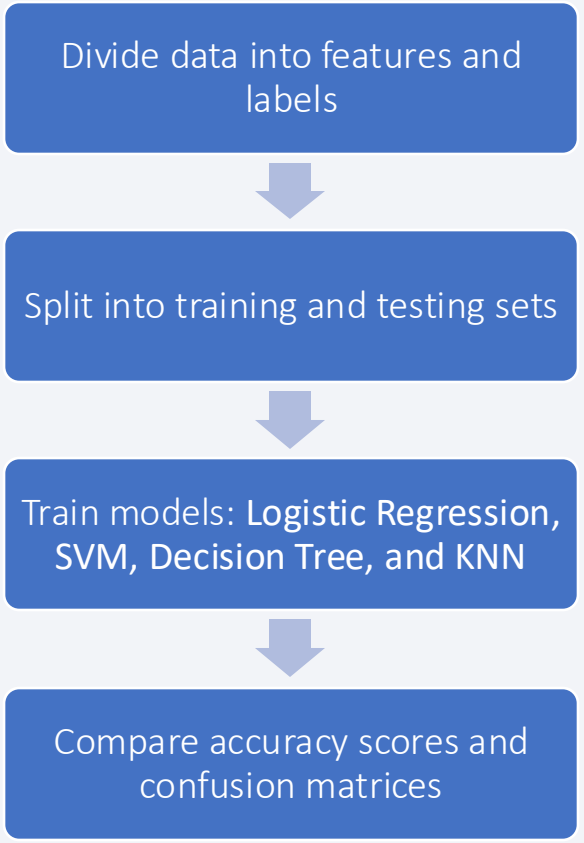
Build a Dashboard with Plotly Dash

- Dash shows two graphs, a scatter chart to show the correlation between payload and launch success, allowing us to select among the launch sites or all of then thanks to a dropdown menu, and a **pie chart** showing the total successful launches count for all sites, or showing the Success vs. Failed counts for a specific selected from a dropdown menu.

Plotly Dash
lab: https://github.com/AbelU1999/Applied_Data_Science_Capstone/blob/3f43498ff7d99e50092baa6b0c8b584747c5c796/spacex-dash-app.py

Predictive Analysis (Classification)

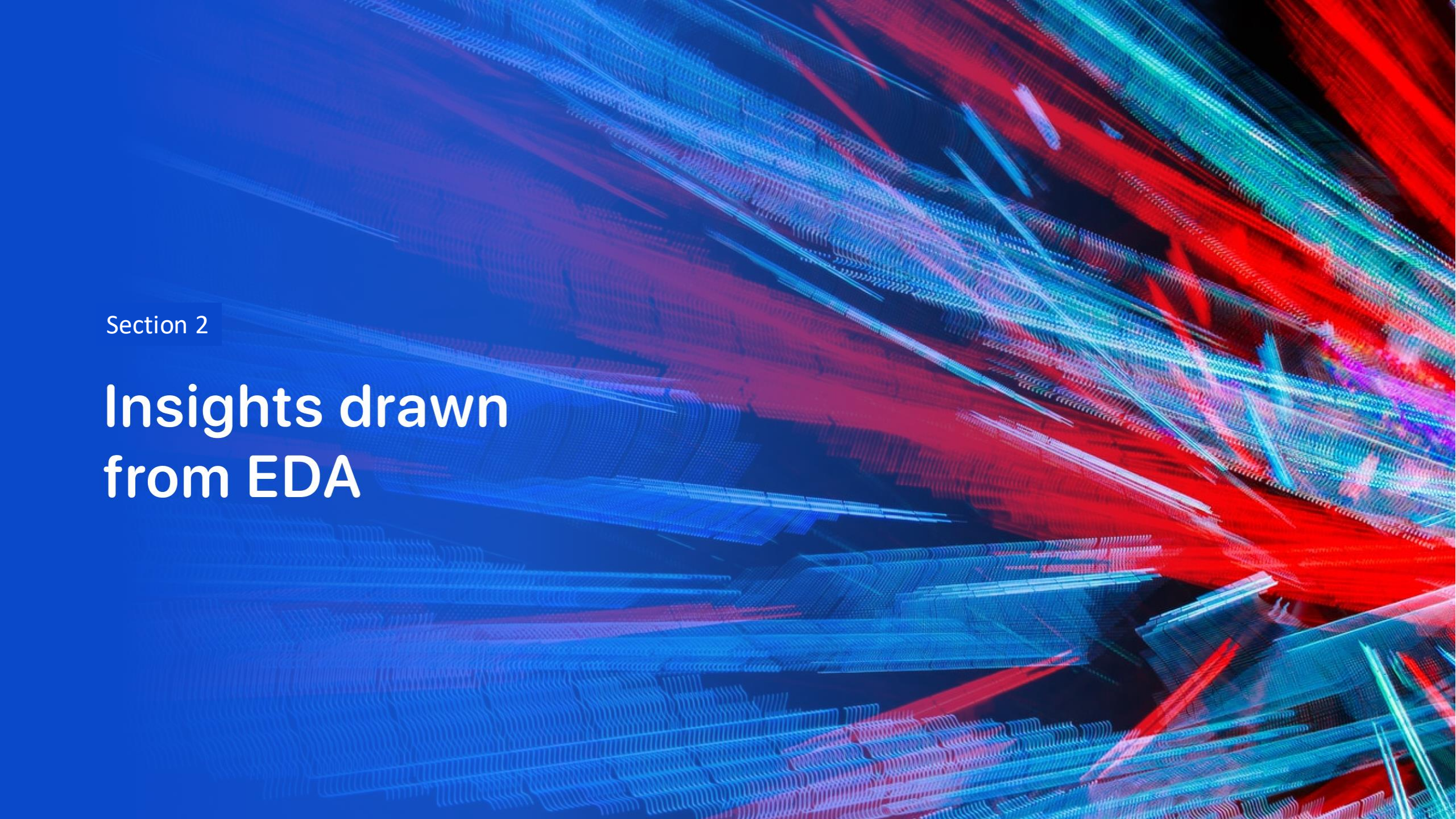
- The dataset was divided into predictive features and class labels.
- Both were split into training and testing sets.
- Four algorithms were evaluated: Logistic Regression, SVM, Decision Tree, and KNN.
- Each model was optimized and tested using the same data.
- Performance was compared using accuracy scores and confusion matrices to find the best model.
- Notebook: https://github.com/AbelU1999/Applied_Data_Science_Capstone/blob/8809a93cc34e9c421aebb0c287e188f1d2f191e4/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Results

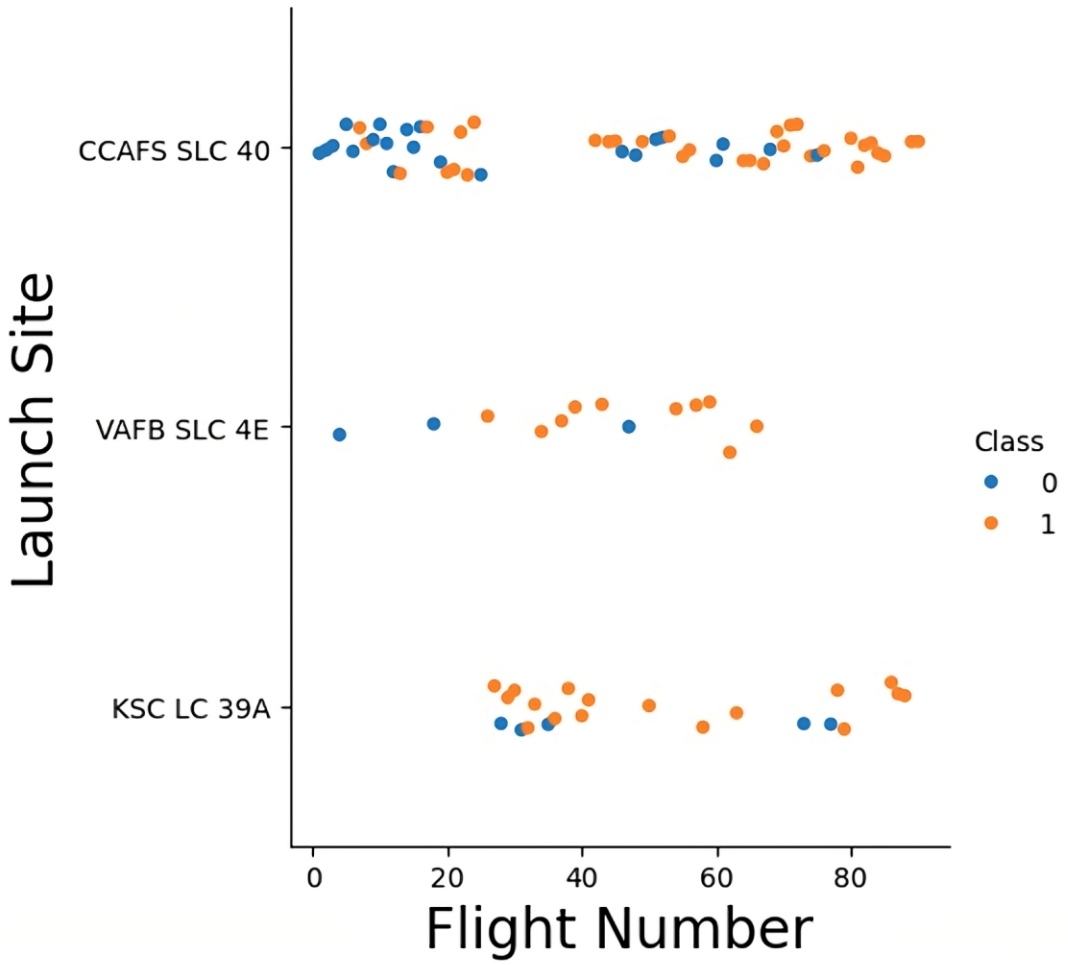
Exploratory Data Analysis (EDA), Interactive Analytics, and Predictive Analysis Results

- The exploratory data analysis revealed clear relationships between **payload mass, orbit type, and landing success rates**.
- **Launch success** increased consistently from **2013 to 2020**, showing improvements in reliability.
- The **interactive maps** demonstrated that all launch sites are **located near coastlines**, and Florida hosts the majority of launches with the highest success rates.
  The **dashboard analysis** showed that **KSC LC-39A** led in successful launches, followed by **CCAFS LC-40**.
- Predictive models (Logistic Regression, SVM, Decision Tree, and KNN) all achieved an **accuracy of 83%**, confirming that classification algorithms can **reliably predict landing success** using the available features.


Section 2
Insights drawn from EDA

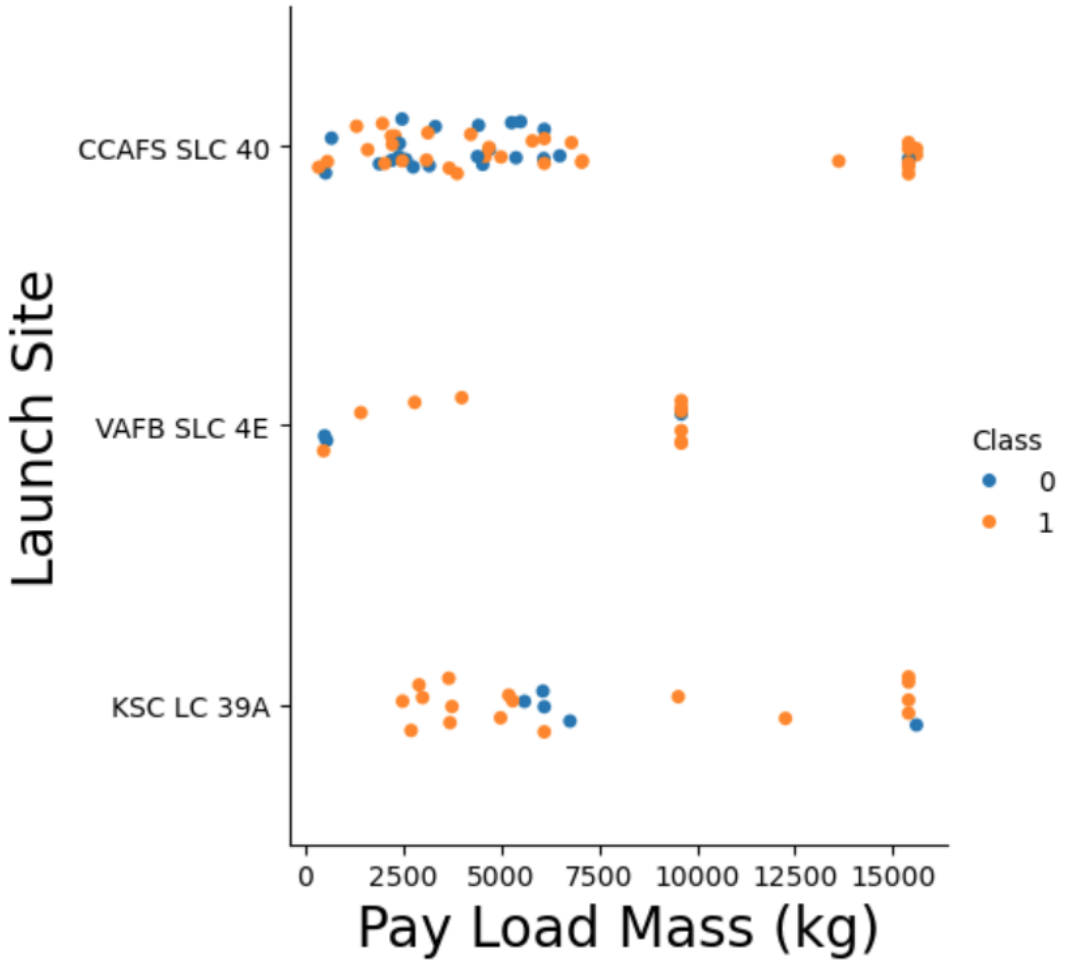
Flight Number vs. Launch Site

- Launches at KSC LC 39A have started after launch number 20.
- Launches at CCAFS SLC 40 have remained constant, pausing when the first cluster of launches began at KSC LC 39A.
- There is no apparent relationship between flight number, launch site, and success.



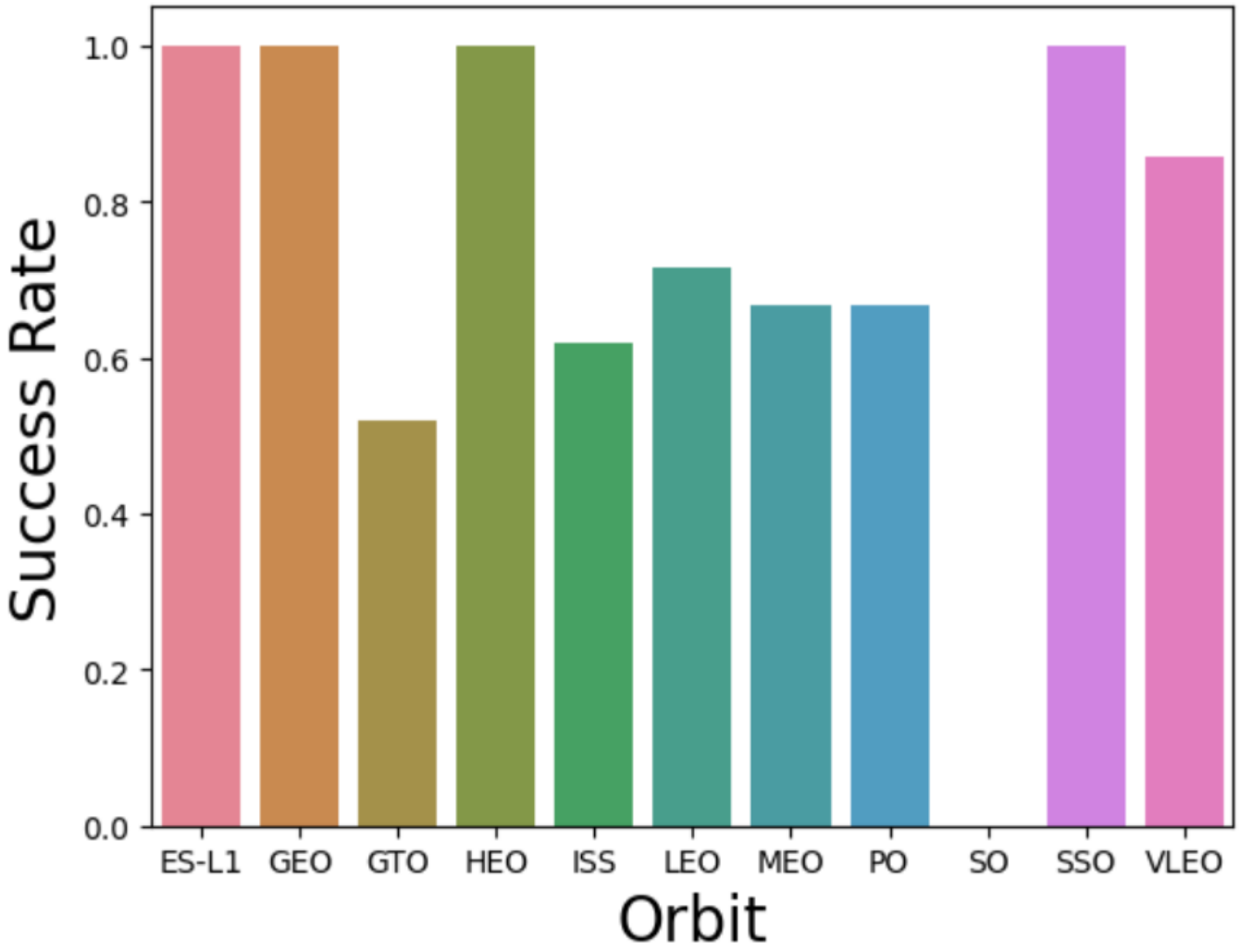
Payload vs. Launch Site

- Most of the launches has a payload mass no greater than 7500 kg.
- At VAFB SLC 4E, there were no launches with a payload exceeding 10,000 kg.
- There is a lower percentage of failed outcomes with higher payloads



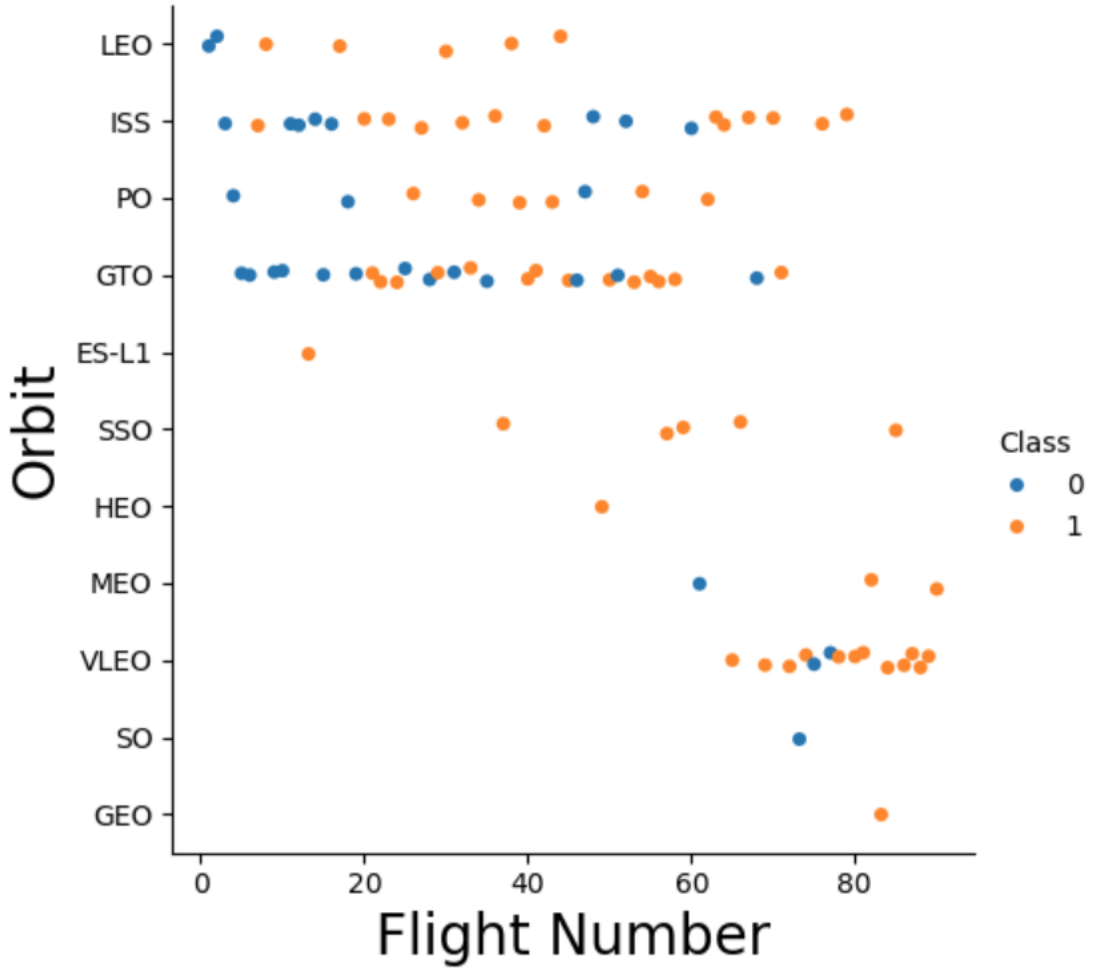
Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO, and SSO apparently has the highest success rate, cause that most of them just have a launch or a few, with a positive outcome.



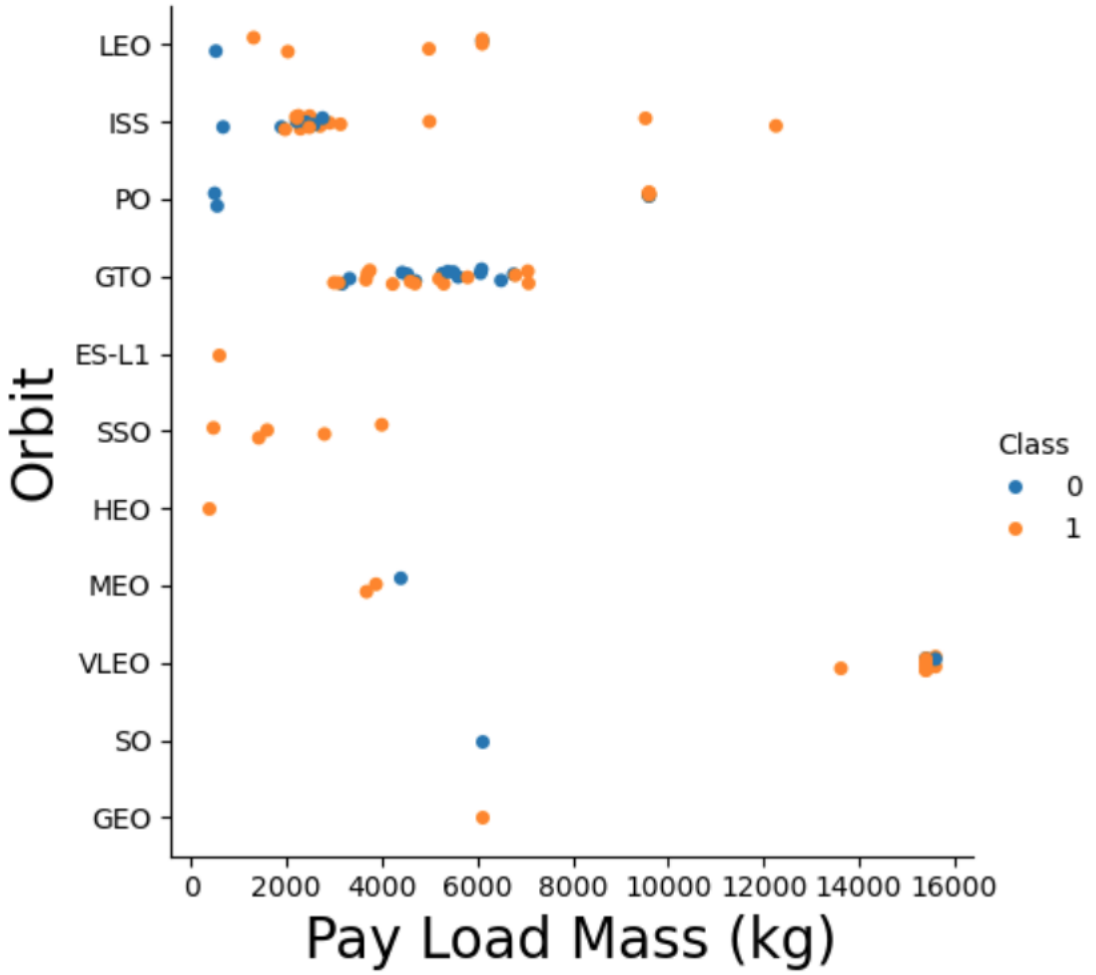
Flight Number vs. Orbit Type

- Launches to orbits as MEO,VLEO, SO, and GEO, begin after flight number 60.
- There appears to be no relationship between flight number and success.



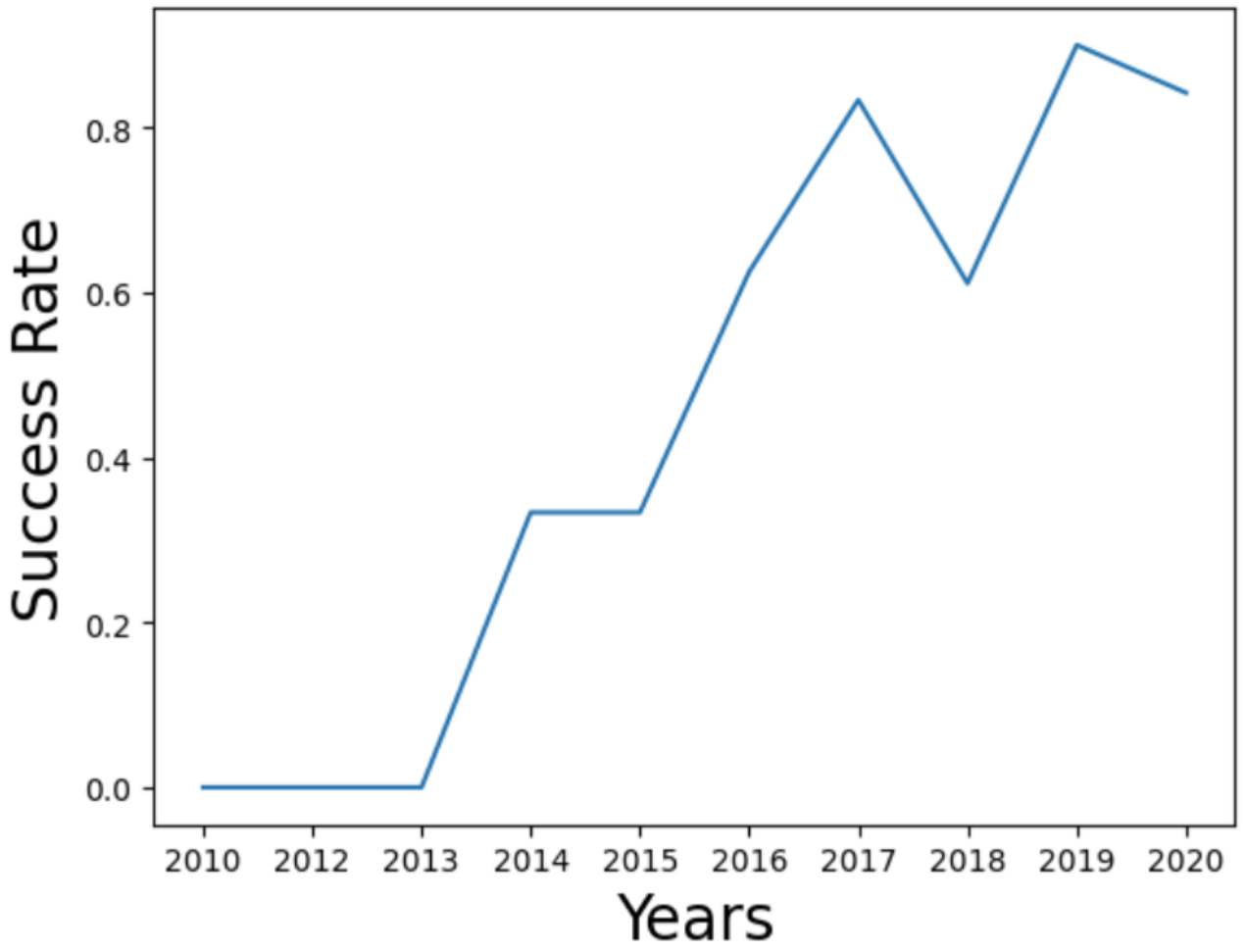
Payload vs. Orbit Type

- Some orbits as LEO, ISS, and PO has more successful landings with heavy payloads.
- VLEO just has landings with heavy payloads.
- GTO don’t show a correlation between success and payload mass.



Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020.



All Launch Site Names

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

The results of the query show the names of the unique launch sites.

Launch Site Names Begin with 'CCA'

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

- The query returns records for launch sites beginning with "CCA," such as CCAFS LC-40 and CCAFS SLC-40.

Total Payload Mass

Total_NASA_PLM_KG
45596

- We obtain the total sum of payload mass carried by boosters launched by NASA (CRS).
- Being the total 45,596 kg.

Average Payload Mass by F9 v1.1

AVG("PAYLOAD_MASS__KG_")
2928.4

- Average payload mass carried by booster F9 v1.1
- Result: 2,928.4 kg

First Successful Ground Landing Date

MIN(Date)
2015-12-22

- The first successful ground landing occurred on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- With the query we filtered the names of boosters that land successfully in a drone ship, with a payload mass between 4,000 and 6,000 kg.

Total Number of Successful and Failure Mission Outcomes

Successful	Failure
100	1

- The query returns two new columns, one showing total successful mission outcomes and the other showing total failed mission outcomes.

Boosters Carried Maximum Payload

- Query returns booster versions that had carried the maximum payload mass. A total of 12 boosters.

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

2015 Launch Records

- Query results show failed landing outcomes on drone ships for year 2015, all of them launched from CCAFS LC-40 being a total of 7 for that year.

Months	landing_outcomes	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
02	Failure (drone ship)	F9 v1.1 B1013	CCAFS LC-40
03	Failure (drone ship)	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1016	CCAFS LC-40
06	Failure (drone ship)	F9 v1.1 B1018	CCAFS LC-40
12	Failure (drone ship)	F9 FT B1019	CCAFS LC-40

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

Date	Landing_Outcome	COUNT
2012-05-22	No attempt	10
2016-04-08	Success (drone ship)	5
2015-01-10	Failure (drone ship)	5
2015-12-22	Success (ground pad)	3
2014-04-18	Controlled (ocean)	3
2013-09-29	Uncontrolled (ocean)	2
2015-06-28	Precluded (drone ship)	1
2010-12-08	Failure (parachute)	1

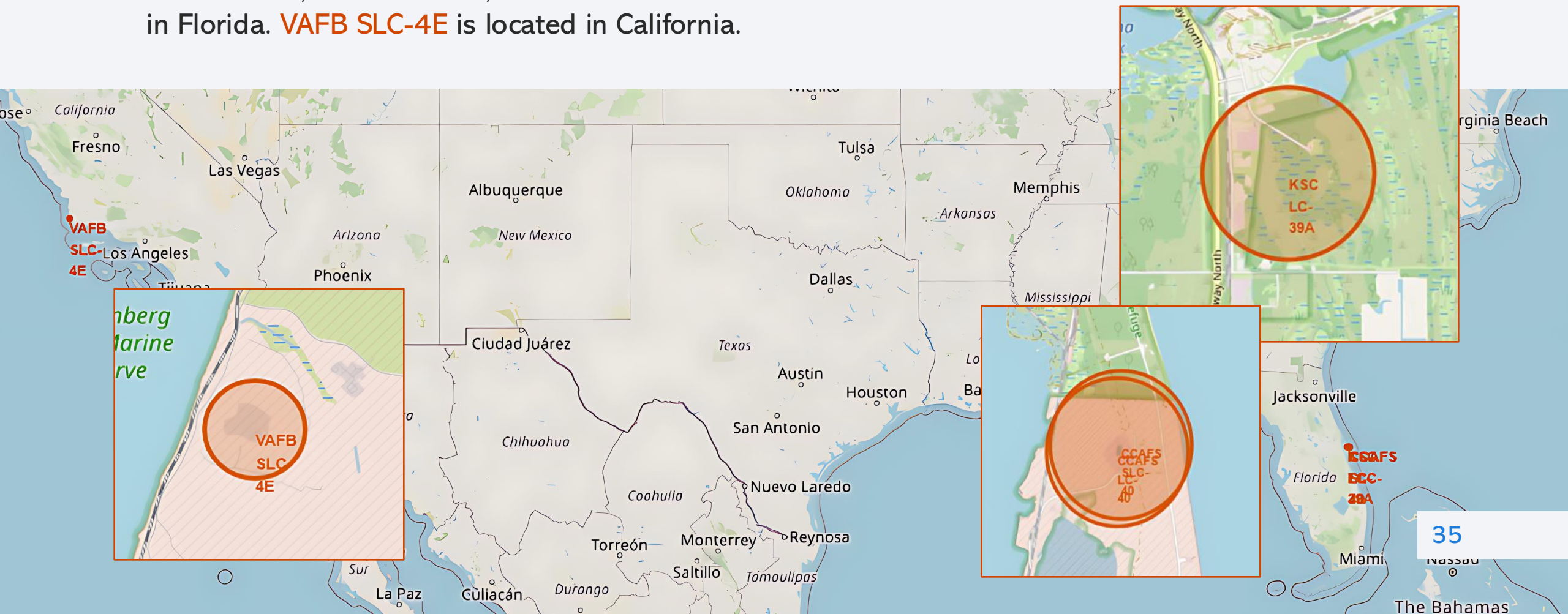
- Results shows year 2016 with most successful landing outcomes (count 5) on drone ship, for year 2015 get a total of 5 failure landings on drone ships. In 2012 no landings were attempted.

Section 3

Launch Sites Proximities Analysis

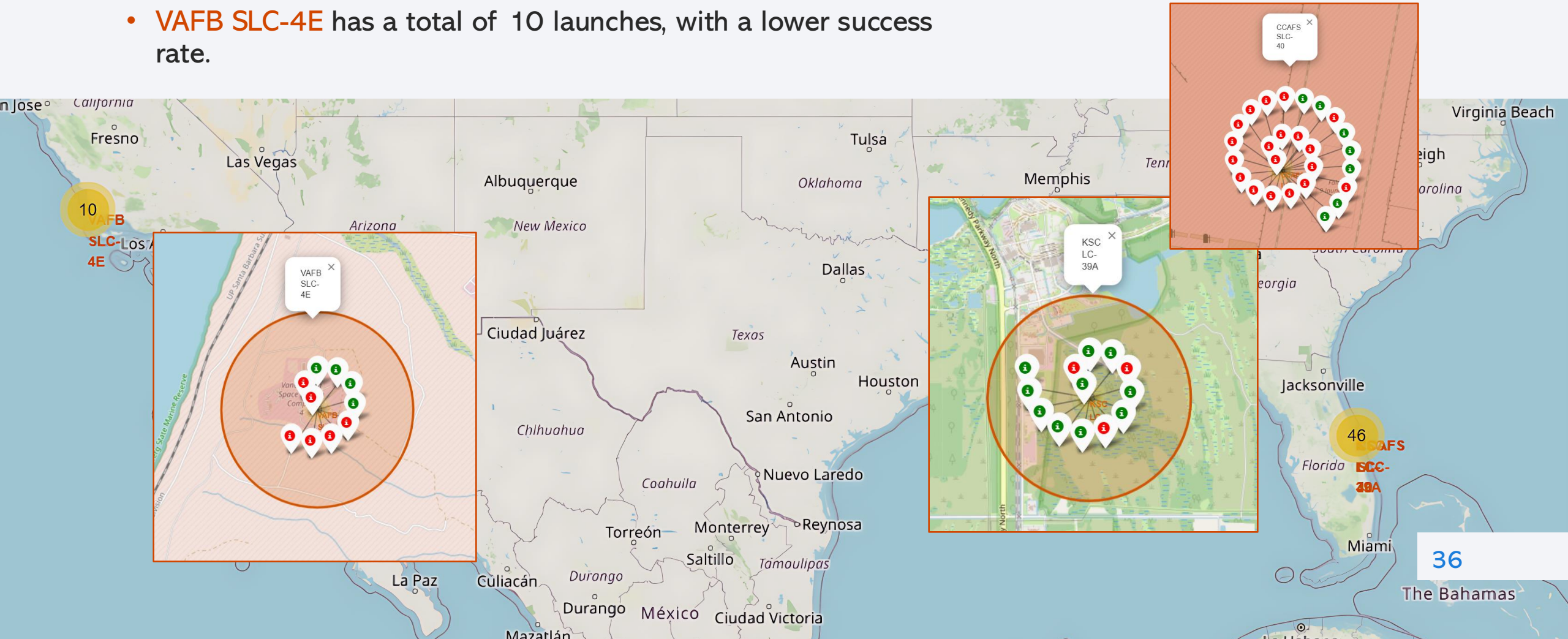
Global Map

- All of the launch sites are located near the coast.
- KSC LC-39A, CCAFS LC-40, and CCAFS SLC-40 are located in Florida. VAFB SLC-4E is located in California.



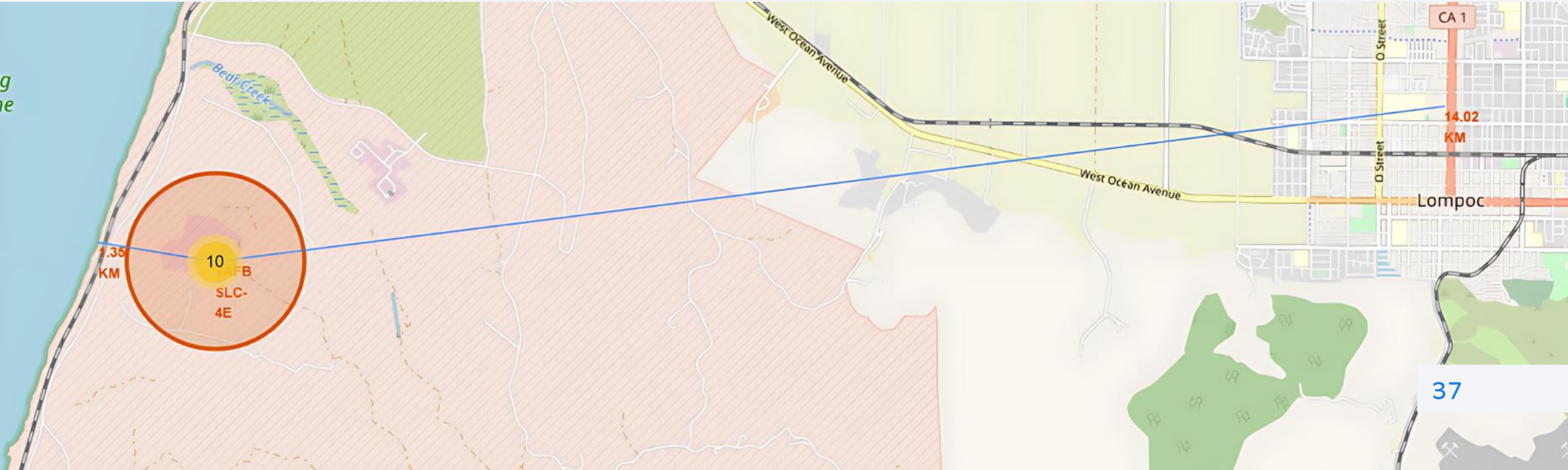
Launch Outcome

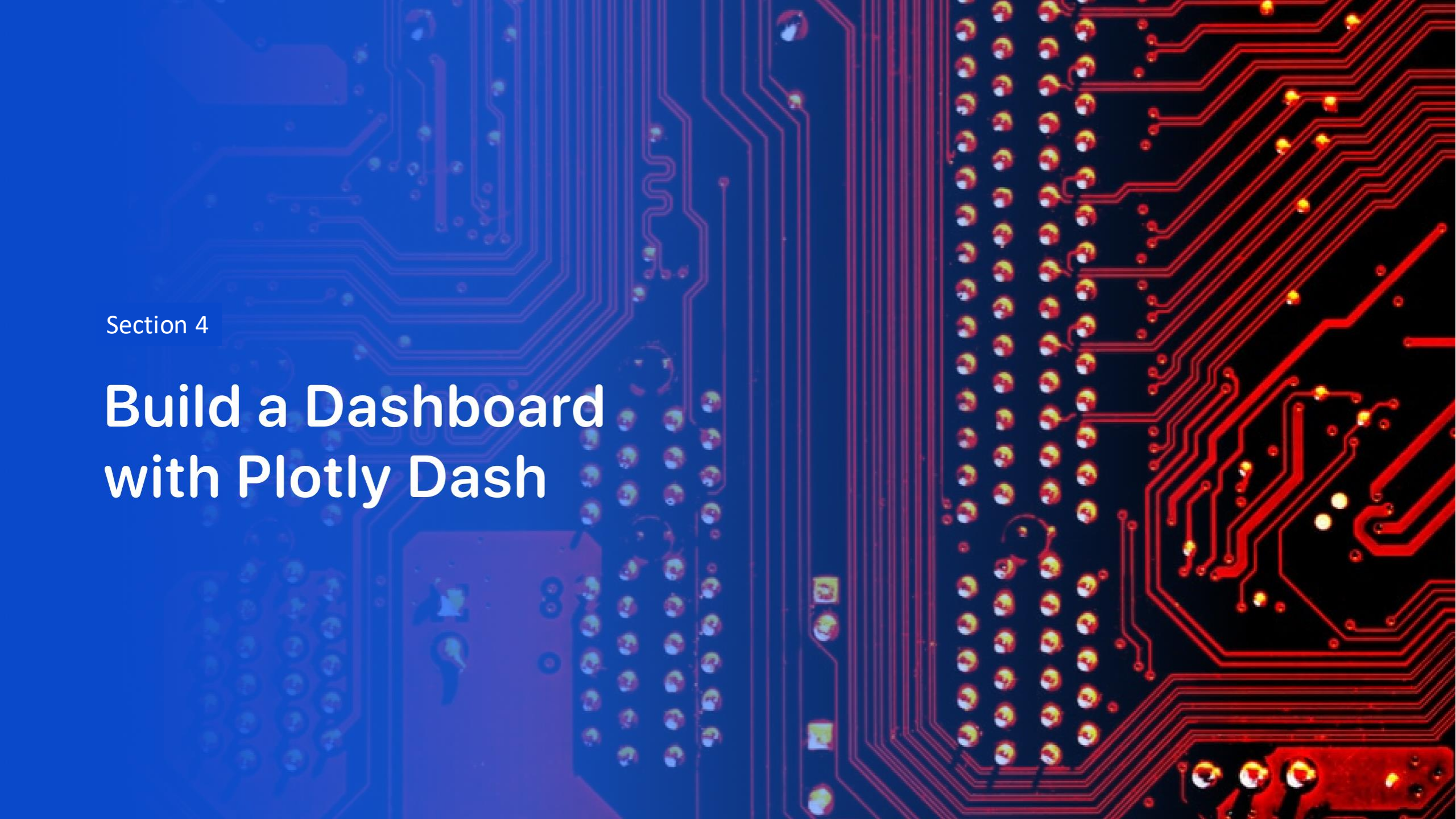
- The majority of launches take place in Florida, with a total of 46. KSC LC-30A has a success rate of approximately 77%.
- VAFB SLC-4E has a total of 10 launches, with a lower success rate.



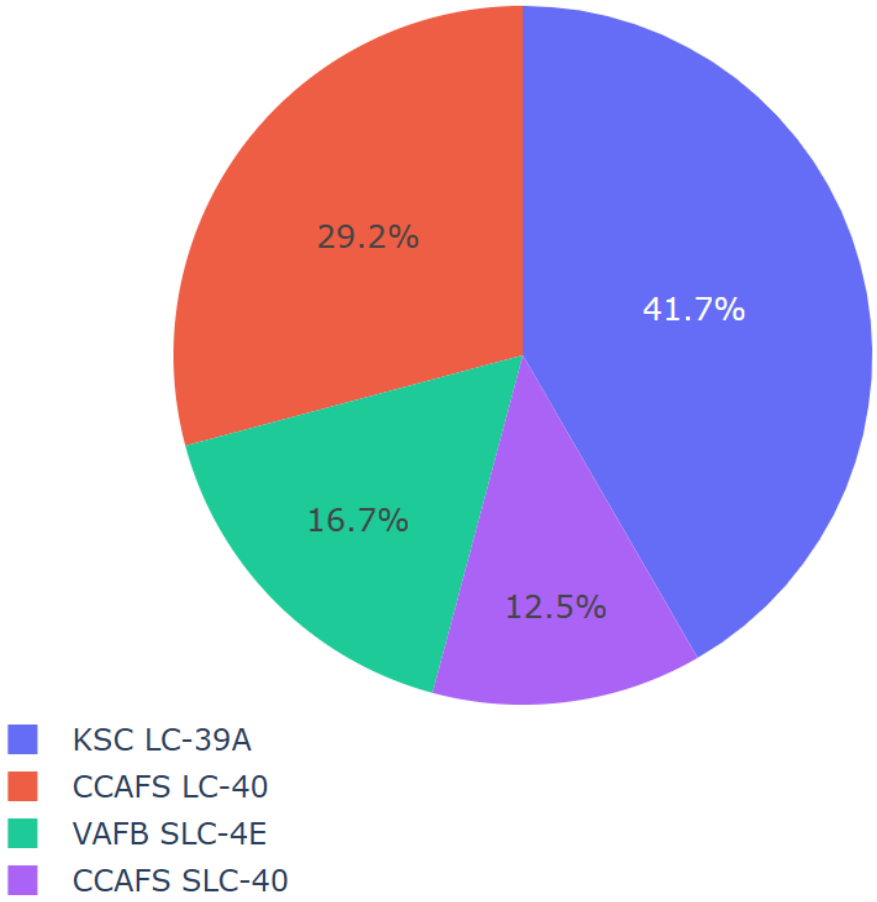
Launch Sites proximities

- The launch sites are close to the coast, and some are near railroad tracks. These sites are farther away from cities and populated areas.




Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by site

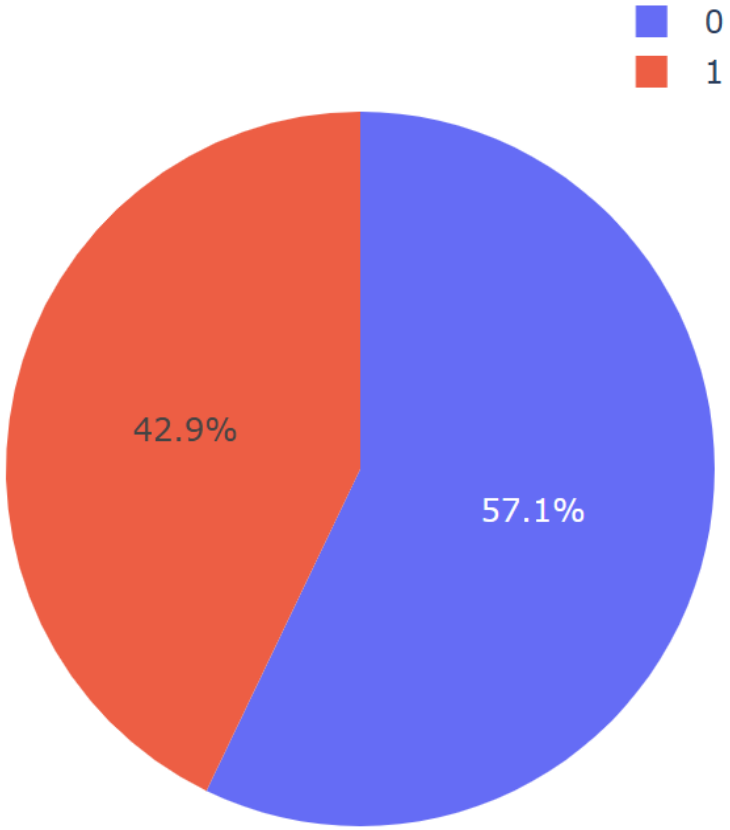


Pie chart. Launch success rate for all sites

The **Pie chart** shows KSC LC-39A leading with a success rate of 41.7%, follow by CCAFS LC-40 with a 29.2%, being CCAFS SLC-40 the minor with a 12.5% success rate.

Site with highest success launches

Despite having only a 12.5% success rate in launches compared to other sites, **CCAFS SLC-40** has the highest percentage of positive launches compared to negative ones.



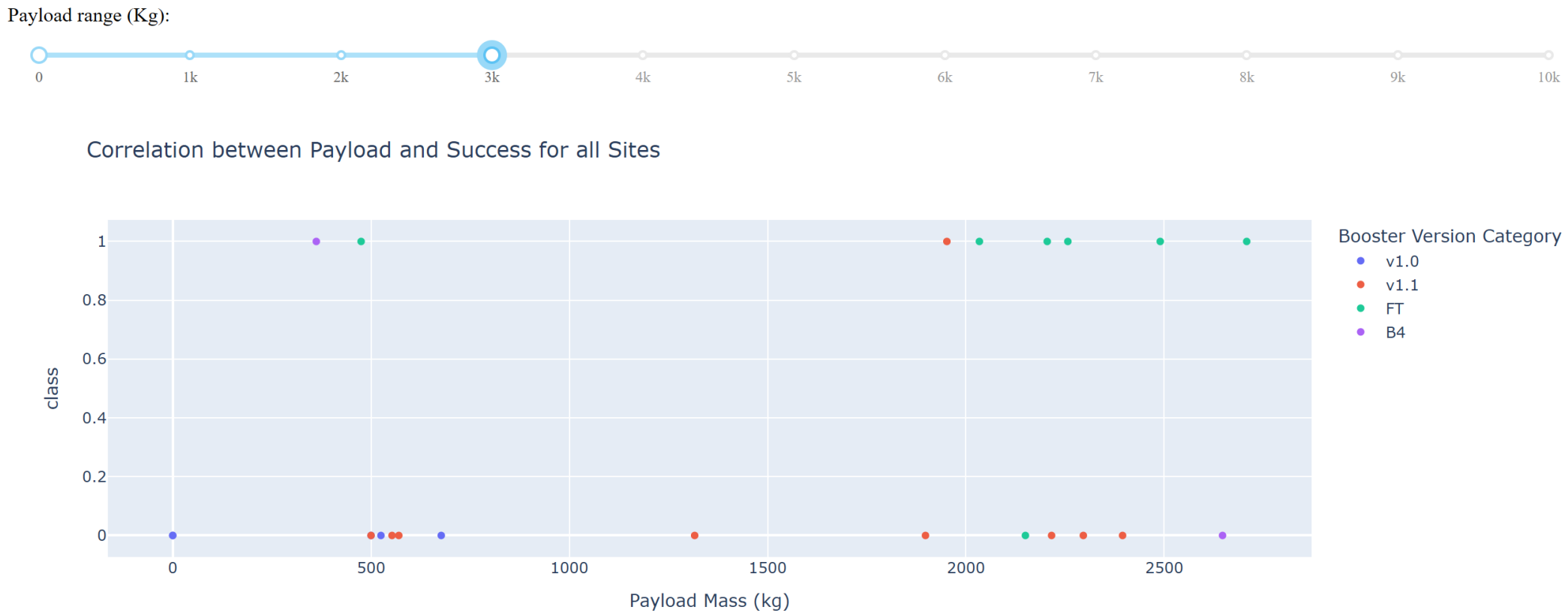
Pie chart. Launch success rate for CCAFS SLC-40

Correlation between payload and success

Booster FT appear to have the largest success rate with payloads going from light to more heavy

Correlation between payload and success

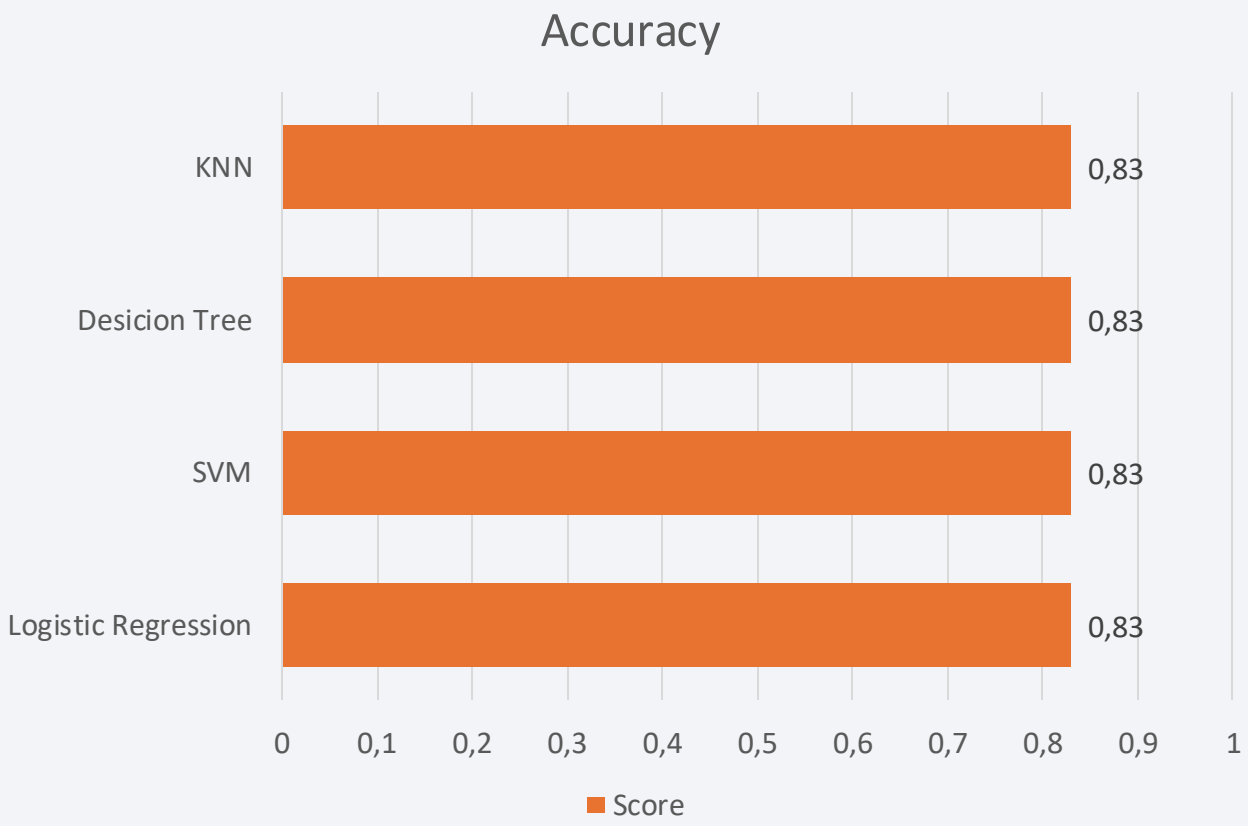
- With lighter loads, the FT booster has a higher positive launch rate.



Section 5

Predictive Analysis (Classification)

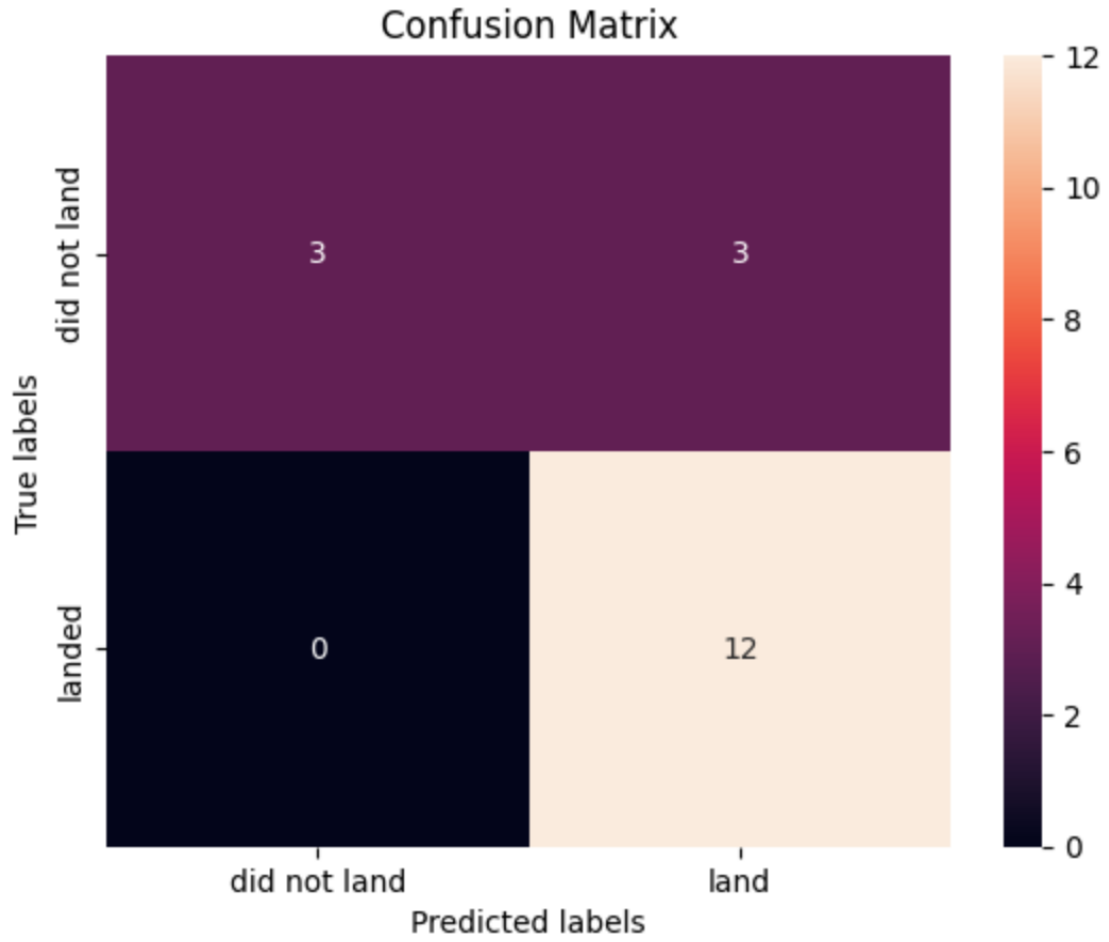
Classification Accuracy



- All models apparently have the same accuracy of 83% with the test data.
- Possibly, if a bigger data set was available, models will learn better the
data, having better results.

Confusion Matrix

- The confusion matrix shows that the model predicted correctly 12 landings and present 3 false positives as landed when not.



Conclusions

- The integration of **data from SpaceX's API and Wikipedia** enabled a comprehensive analysis of launch performance and success factors.
- **Exploratory and SQL-based analyses** provided valuable insights into payload distribution, orbit success rates, and temporal trends in launch outcomes.
- **Interactive visualization tools** such as Folium and Plotly Dash enhanced data interpretation and revealed geographic and operational patterns.
- **Machine learning models** achieved a consistent 83% accuracy, demonstrating that launch success can be predicted effectively from payload and site data.
- Overall, this study shows that **data-driven insights** can guide **SpaceY** in optimizing its operations, improving reliability, and achieving **competitive pricing strategies** in the rocket launch industry.

Thank you!