IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Joshua Soriano
17/January/2024

Outline

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

Executive Summary

- Summary of methodologies
  - Gathered data using two different methods from two publically accessible sources, then sorted the data by narrowing the range of data and eliminating null values.
  - Detail-oriented exploratory data analysis using database queries and visualisations.
  - Additional interactive data analysis using dashboard components and maps
- Summary of all results
  - The observations obtained from the Retrieved Launch Data showed a correlation between the Launch Site, Payload Mass, Launch Number, Orbit Type, and Landing Outcome.
  - Additionally, the four machine learning models all performed similarly well in predicting the landing outcome of a launch, despite the fact that Launch Sites share certain geographic and infrastructure features.

Introduction

- SpaceX is leading the commercial sector in the present "Space Race" era thanks to their economical launch strategy, which involves reusing the initial rocket stage.
- According to SpaceX's website, the cost of launching a Falcon 9 rocket is approximately $62 million; the same cost is around $165 million for other providers.
- In an effort to rival SpaceX, our company, Space Y, uses data science methods to forecast whether SpaceX will deploy the first rocket stage. Future launch costs can be calculated by evaluating the likelihood that the first stage will land and hence be reused.

Section 1

Methodology

Methodology - Executive Summary

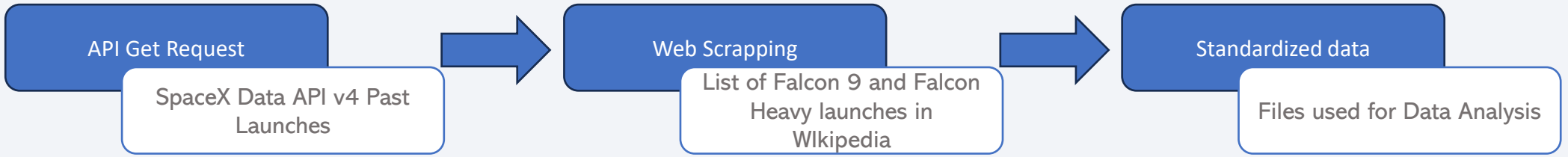
- Data collection methodology:
  - Completed web scraping with Beautiful Soup library to gather history launch records for the Falcon 9. Source: Wikipedia's list of Falcon 9 and Falcon Heavy missions;
  - Sent a get request to the SpaceX API using the requests library. Source: SpaceX Data API v4 Past missions
- Perform data wrangling
  - Only Falcon 9 is being filtered
  - Handle missing value by replacing by the means

Methodology - Executive Summary

- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using standardised data that has been divided into training and test sets
  - Using classification trees, K Nearest Neighbours (KNN), logistic regression, and support vector machines (SVM), I isolated hyperparameters from the training data.
  - Methods were tested with a test set, and accuracy was displayed in confusion matrices.

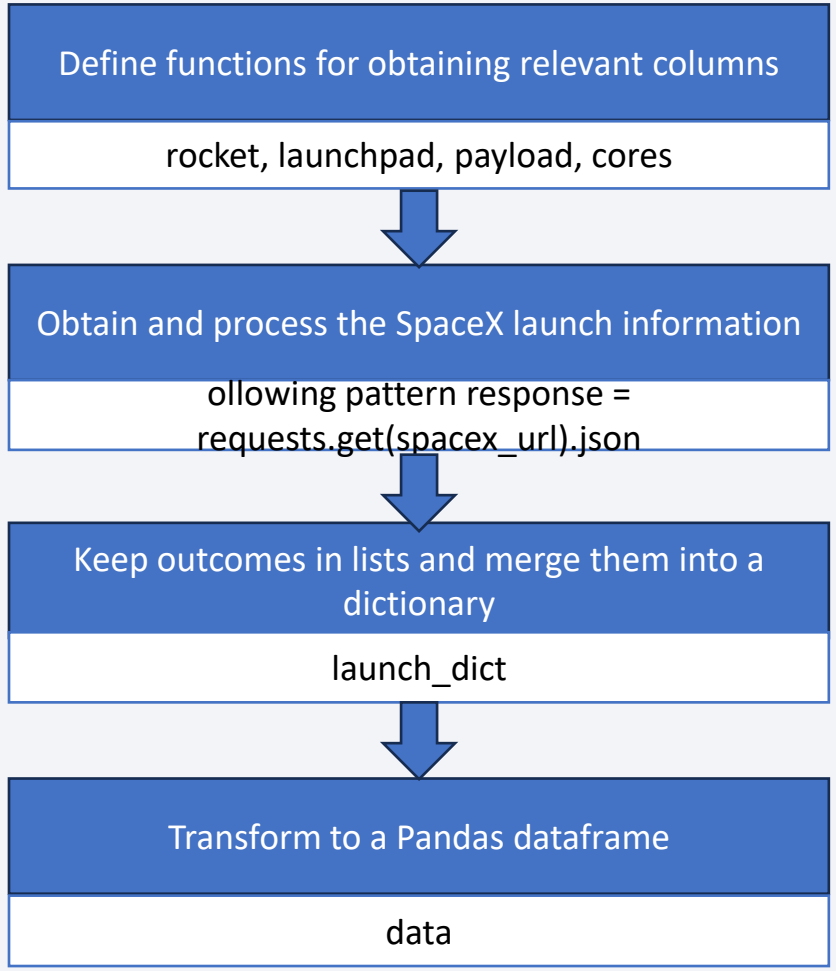
Data Collection

- Individual data sets were gathered from several sources using two separate techniques.
  - SpaceX Data API v4 Past Launches obtained through a get request from SpaceX
  - List of Falcon 9 and Falcon Heavy launches that Beautiful Soup was used to scrape from Wikipedia
- Standardized data sets produced by the leadership team using data that was retrieved



Data Collection – SpaceX API

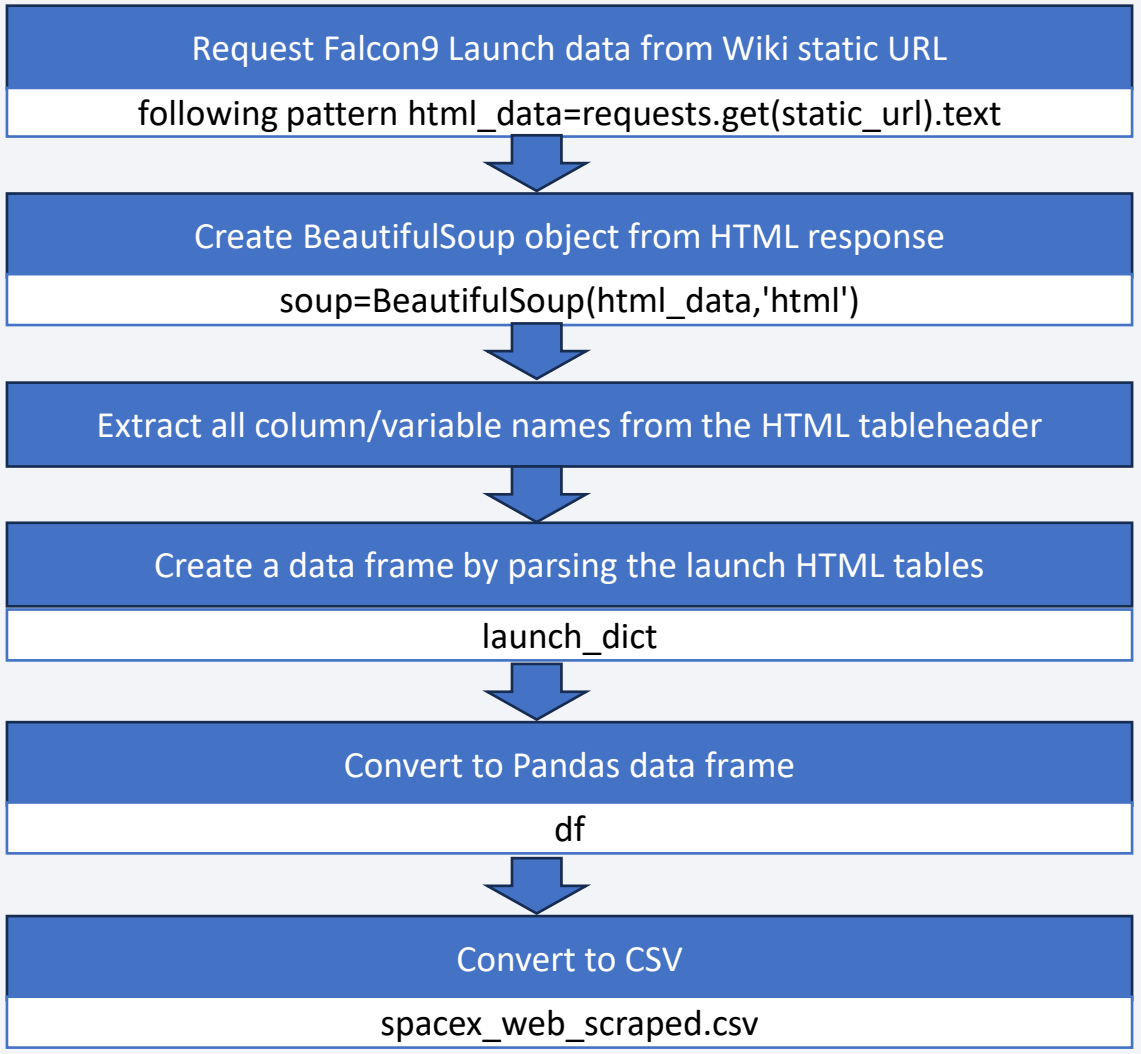
- SpaceX launch data was requested using GET request forms, and helper functions were built to retrieve relevant columns from the SpaceX REST API.
- Lists of the results were kept, compiled into a dictionary, and then converted to a Pandas data frame.



Reference: https://github.com/JFSmetrix/Applied-Data-Science-Capstone/blob/main/Week%201%20-%20Lab%201%20Collecting%20the%20data.ipynb

Data Collection – Web Scraping

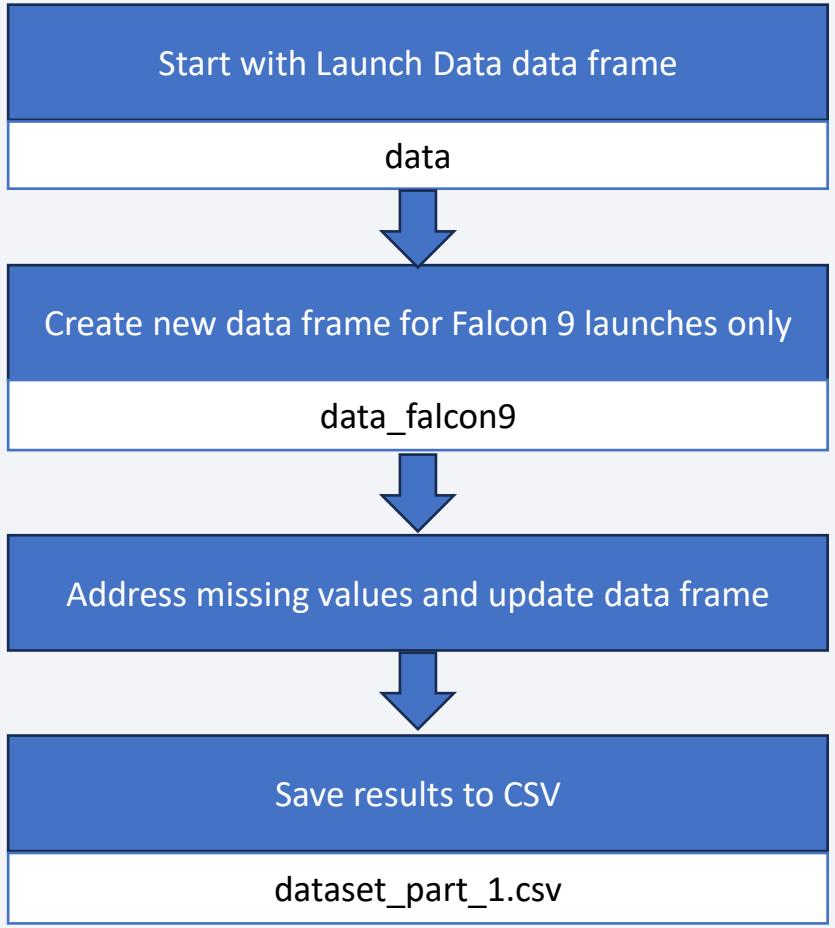
- Using a GET request, Falcon 9 launch data was obtained from Wikipedia and transformed into a BeautfulSoup object.
- The HTML table header included the column names, while the tables itself contained the data.
- The output was first saved as lists, then concatenated into a dictionary, and finally converted into a CSV file and Pandas data frame.



Reference: https://github.com/JFSmetrix/Applied-Data-Science-Capstone/blob/main/Week%201%20-%20Web%20scraping%20Falcon%209%20and%20Falcon%20Heavy%20Launches%20Records%20from%20Wikipedia.ipynb

Data Wrangling

- Started by using the SpaceX API's Launch Data data frame.
- Only the Falcon 9 launches remained after the Falcon 1 launches were eliminated using the BoosterVersion column.
- When landing pads were not in use, LandingPad was permitted to keep its None values.
- The column mean value was utilised to substitute NaN values in the payload mass.



Reference: https://github.com/JFSmetrix/Applied-Data-Science-Capstone/blob/main/Week%202%20-%20Lab%202%20Data%20wrangling.ipynb

EDA with Data Visualization

The initial attempt at exploratory data analysis was done through data visualisation. For this reason, the charts that follow were created:

- Using categorical charts (catplots), ascertain the potential influence of two variables on the result:
  - Flight Number and Payload Mass
  - Launch Site and Flight Number
  - Launch Site and Payload Mass
  - Flight Number and Orbit Type
  - Payload Mass and Orbit Type
- To see the relationship between each orbit type's success rate, use a bar chart (barplot).
- A line chart, or lineplot, that shows the patterns in average launch success over time.

Reference: https://github.com/JFSmetrix/Applied-Data-Science-Capstone/blob/main/Week%202%20-%20Exploring%20and%20Preparing%20Data.ipynb

EDA with SQL

The subsequent attempt at exploratory data analysis was conducted through SQL database queries. For this reason, the following inquiries were made:

- Query Launch Site details
  - The space mission's distinct launch sites' names
  - Documents indicating that CCAFS launch sites
- Query Payload Mass details
  - Total mass of payload carried by NASA-launched boosters
  - The names of the boosters that have been successful in drone ships and whose payload mass is between 4,000 and 6,000 kg, together with their average payload mass, are listed in version F9 v1.1.
  - The names of the launchers with the largest mass payloads
- Query Outcome details
  - Date of the ground pad's first successful landing outcome. The total number of missions, both successful and unsuccessful
  - The records from 2015 that show the drone ship's landing resulted in a failure, together with the booster version, launch site, and launch month.
  - A list of successful landing outcomes arranged in descending order with their corresponding count from 2010-04-06 to 2017-03-20.

Reference: https://github.com/JFSmetrix/Applied-Data-Science-Capstone/blob/main/Week%202%20-%20Assignment%20SQL%20Notebook%20for%20Peer%20Assignment.ipynb

Build an Interactive Map with Folium

- An interactive map was created using Folium to find geographic patterns related to the launch site.Explain why you added those objects
- For this reason, the following map objects were defined and added to the map:
  - Markers and circles at every launch location
  - A marker indicating the closest shoreline to VAFB SLC-4E, along with a line connecting the two
  - Markers in a marker cluster for each launch outcome
  - VAFB SLC-4E's closest city, Lompoc, California, is marked with a marker, and a line links the two

Reference: https://github.com/JFSmetrix/Applied-Data-Science-Capstone/blob/main/Week%203%20-%20Launch%20Sites%20Locations%20Analysis%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

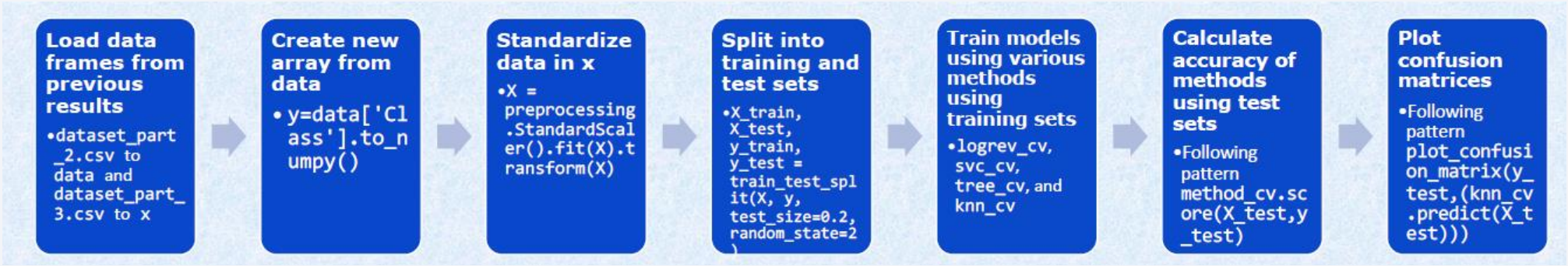
An application for a dashboard designed to provide interactive visual examination of SpaceX launch data.

- Interaction: All launch sites are selected by default from a drop-down list.
  - Visualization: By default, the pie chart shows the total number of successful launches for all sites; when a particular launch site is chosen, however, the outcome counts for that site are displayed.
- Interaction: A payload range selector ranging from 0-10,000 kg in 1,000 kg
  - Visualization: A scatter chart illustrates the relationship between payload and launch success using the chosen payload range.
- Questions this helped to answer:
  - Which site has the largest successful launches and which has the highest launch success rate?
  - Which payload range(s) has the highest and lowest launch success rates?
  - Which F9 Booster version has the highest launch success rate?

Reference: https://github.com/JFSmetrix/Applied-Data-Science-Capstone/blob/main/Week3%20-%20spacex_dash_app.py

Predictive Analysis (Classification)

- Training and test sets were created by loading, standardising, and dividing data frames from earlier outcomes.
- The approaches of Logistic Regression, Support Vector Machine (SVM), Classification Trees, and K Nearest Neighbours (KNN) were utilised to determine the optimal hyperparameter.
- Each of the four above-mentioned approaches' accuracy was computed and plotted.



Reference: https://github.com/JFSmetrix/Applied-Data-Science-Capstone/blob/main/Week%204%20-%20Machine%20Learning%20Prediction.ipynb

Results - EDA

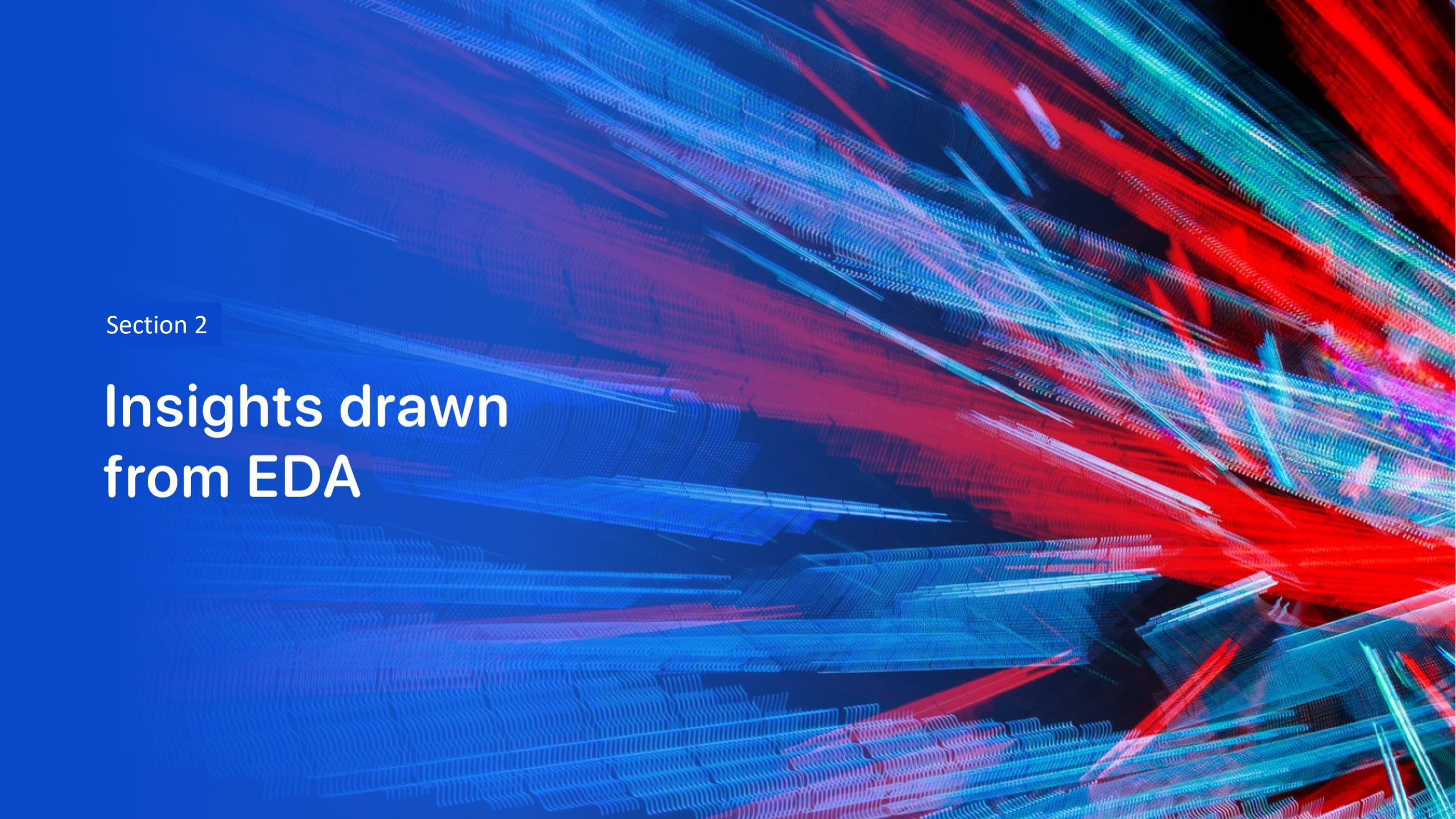
- Launch locations with varying success rates: VAFB SLC 4E with 77%, KSC LC-39A with 60%, and CCAFS LC-40 with 60%.
- The majority of flights take place at CCAFS SLC-40, and higher flight numbers indicate greater success than lower flight numbers.
- The VAFP-SLC launch facility has not conducted any rocket launches weighing more than 10,000 kg in payload.
- The ES-L1, GEO, HEO, and SSO orbit types had the highest success rates.
- There doesn't seem to be any correlation between flight number and success for GTO, however for LEO, success rates are correlated with the number of flights.
- Landings in ISS, LEO, and polar orbits often yield greater success rates and larger cargoes. There are no obvious tendencies for GTO in this regard.
- Growth in Success Rate Trends from 2013 to 2020

Results – Interactive Analysis

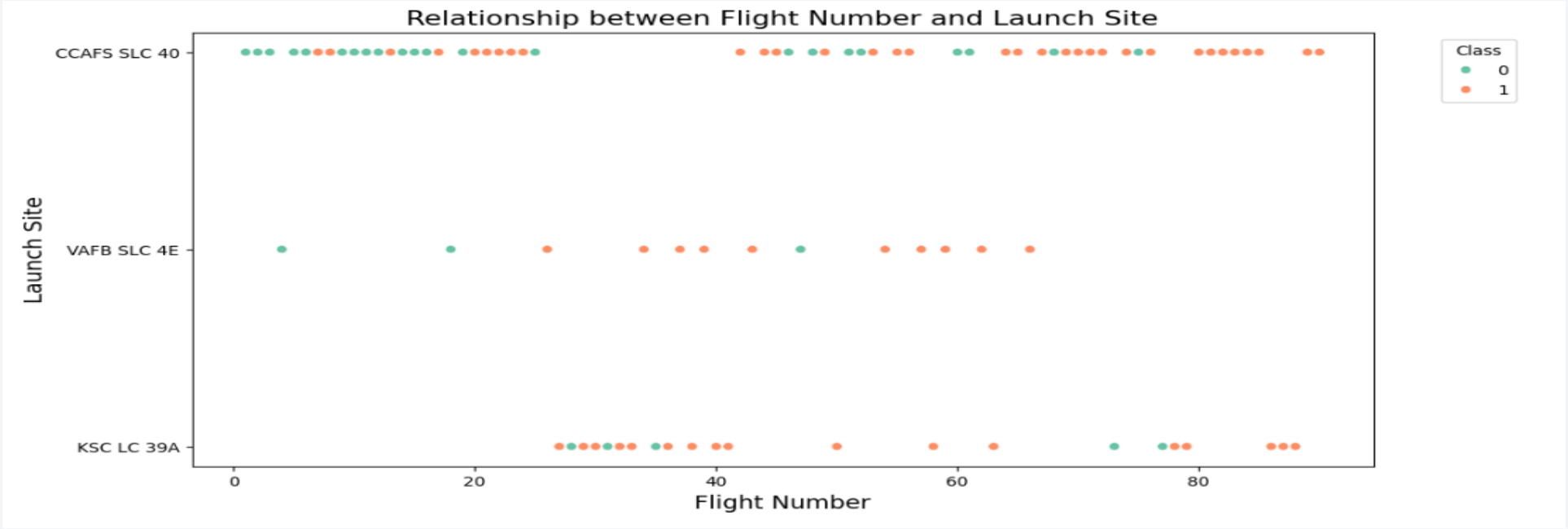
- Map study shows that:
  - Launch sites are farther from cities and highways.
  - Launch sites are near the shore and on a rail line.
- Dashboard study revealed that CCAFS SLC-40 had the highest launch success ratio and KSC LC-39A had the most successful outcomes overall.
- The most successful outcomes were found in the payload mass range of 2,000–5,500 kg, while the least successful range was found in the range of 5,500–10,000 kg.
- The category with the most successful outcomes was the FT booster version category, followed by the B4 booster version category.

Results – Predictive Analysis

- The four methods analysed yielded an accuracy score of 0.8333333333333334.
- The confusion matrices produced identical outcomes for each of the four approaches.

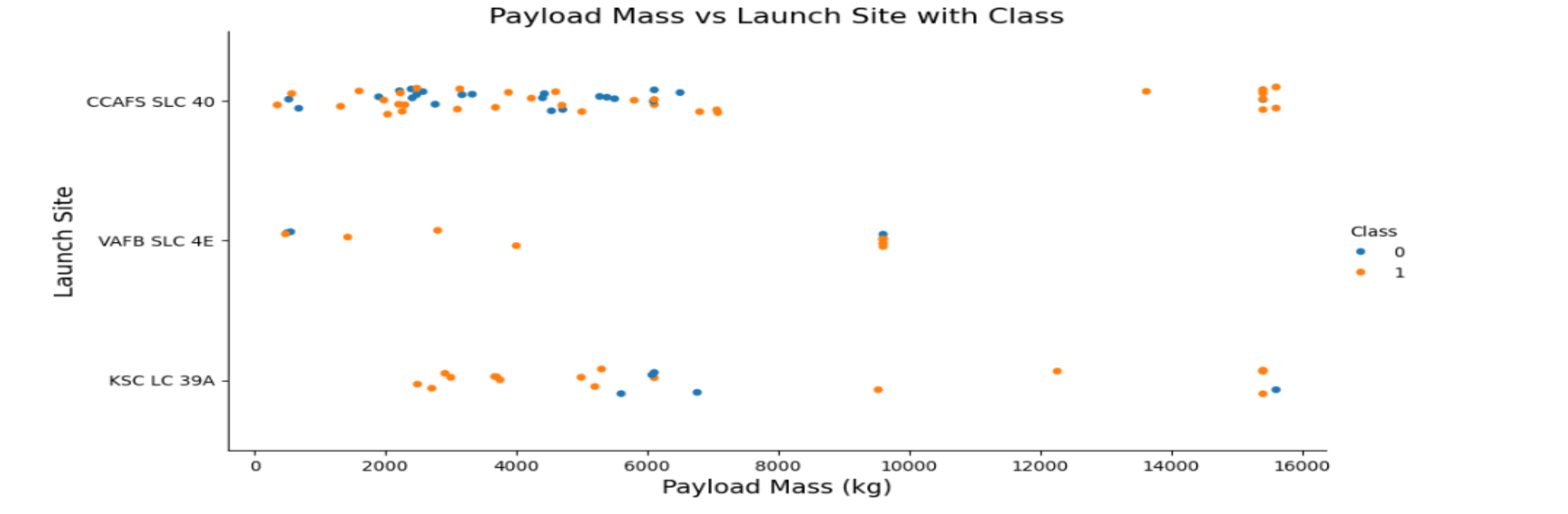

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



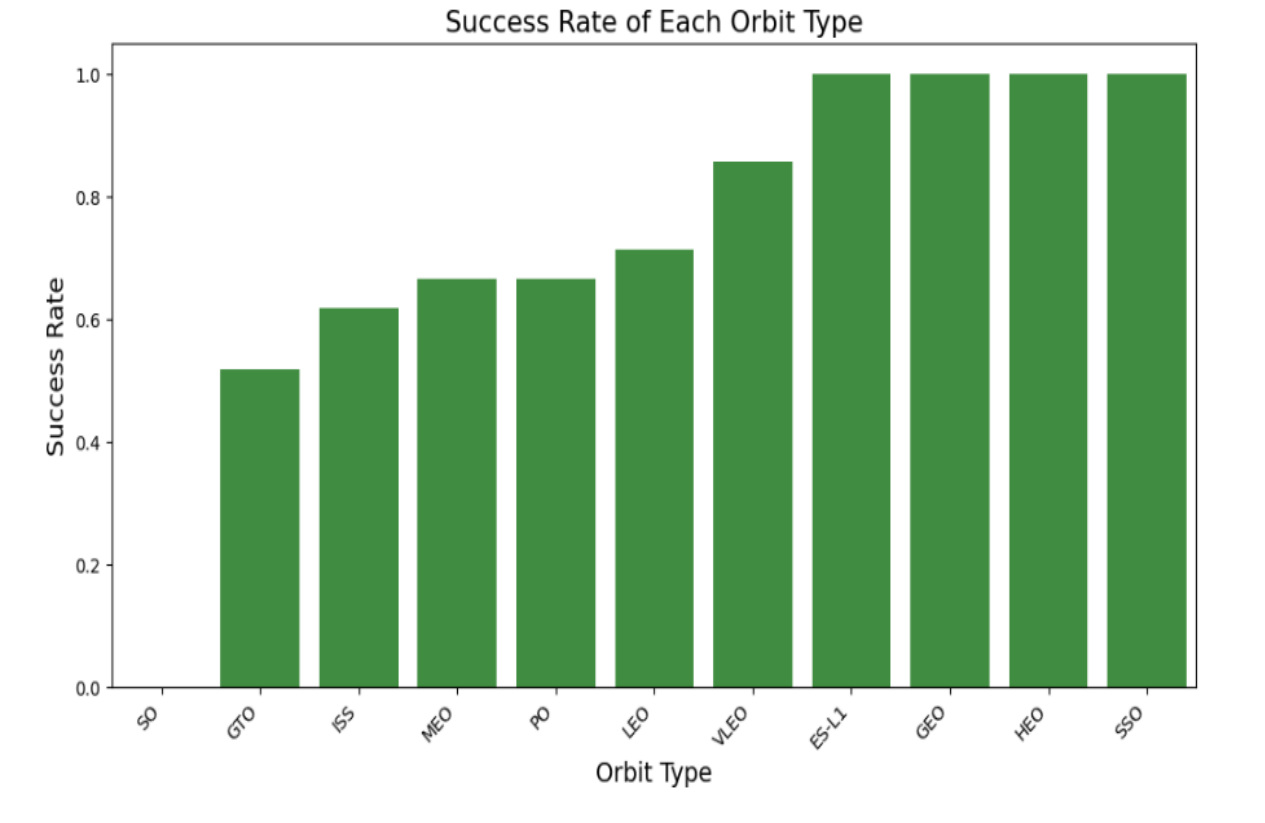
- Launch Site and Flight Number are compared using a scatter plot, where markers are color-coded according to a launch outcome class.
- Class 1, orange, corresponds with a SUCCESS, while Class 0, green, corresponds with an FAILURE.
- The majority of flights take place at CCAFS SLC-40, and higher flight numbers indicate greater success than lower flight numbers.

Payload vs. Launch Site



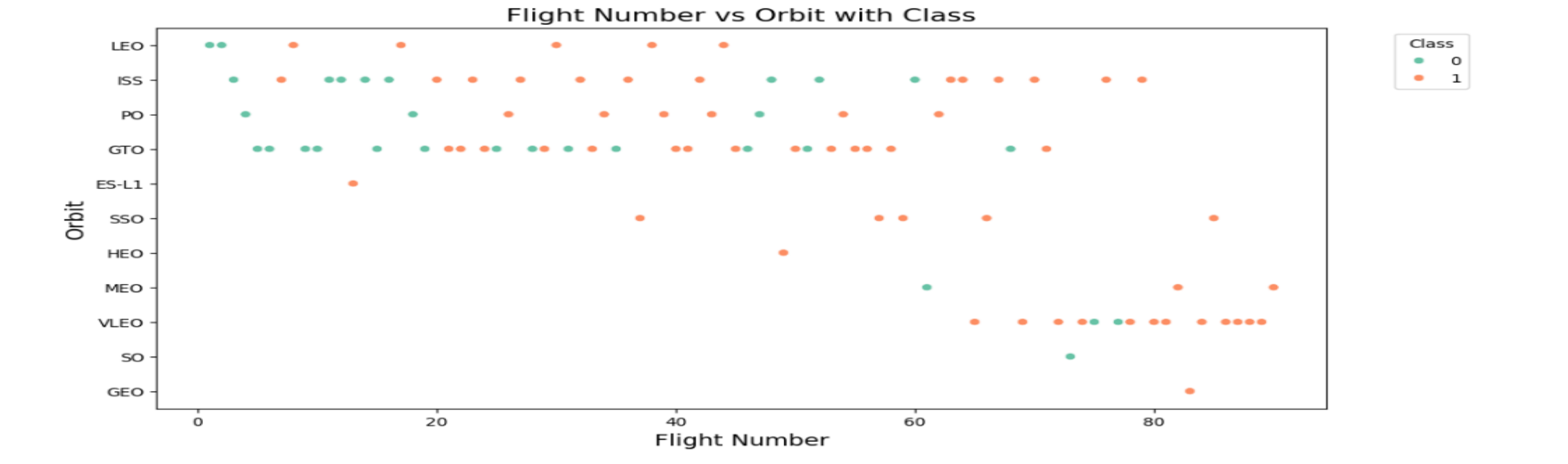
- Launch Site and Payload Mass are compared in a scatter plot, where markers are color-coded according to a launch outcome class.
- Class 1, orange, corresponds with an SUCCESS, while Class 0, blue, corresponds with an FAILURE.
- The majority of missions carried payloads weighing less than 8,000 kg, although those that did carry payloads weighing more than 8,000 kg were nearly always successful.

Success Rate vs. Orbit Type



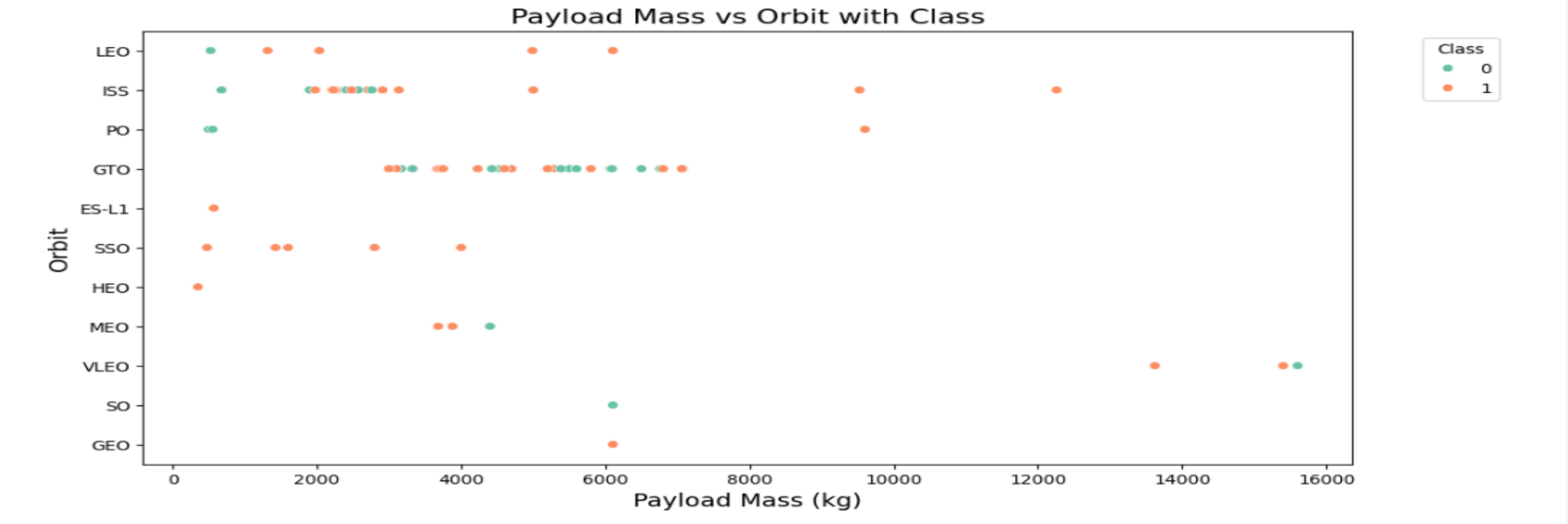
- The orbit types with the best success rates were ESL1, GEO, HEO, and SSO.
- The SO and GTO orbit types had the lowest success percentages.
- The success percentages for LEO, MEO, PO, and ISS orbit types were mediocre.

Flight Number vs. Orbit Type



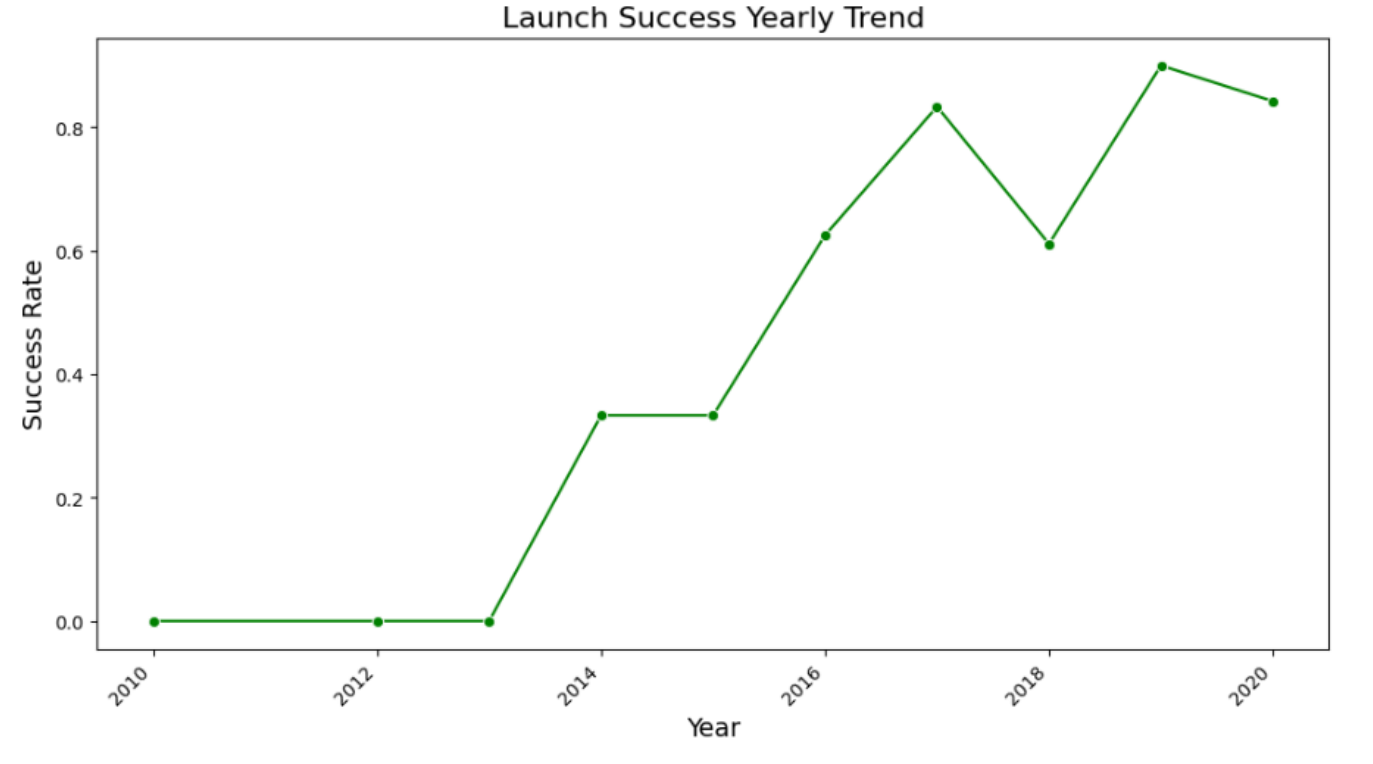
- A scatter plot that color-codes markers according to a launch outcome class compares the Flight Number to the Orbit Type.
- Class 1, orange, corresponds with an SUCCESS, while Class 0, green, corresponds with an FAILURE.

Payload vs. Orbit Type



- A scatter map that color-codes markers according to a launch outcome class compares payload mass to orbit type.
- Class 1, orange, corresponds with an SUCCESS, while Class 0, green, corresponds with an FAILURE.

Launch Success Yearly Trend



- The success rate from 2010 to 2020 rose between 2013 and 2017, following which it experienced a few peaks and valleys until 2020.

All Launch Site Names

- **Task:** Find the names of the unique launch sites
- **Query:** %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
- **Result:**

Task 1

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

```
[10]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
```

[10]:

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

- **Explanation:** There are four unique names for Launch Sites.

Launch Site Names Begin with 'CCA'

- **Task:** Find 5 records where launch sites begin with `CCA`
- **Query:** %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
- **Result:**

Task 2

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:** There were five entries with Launch Sites beginning in "CCA," along with pertinent information.

Total Payload Mass

- **Task:** Calculate the total payload carried by boosters from NASA
- **Query:** %sql SELECT SUM(PAYLOAD_MASS__KG_) as TotalPayloadMass FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
- **Result:**

Task 3

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

```
[12]: %sql SELECT SUM(PAYLOAD_MASS__KG_) as TotalPayloadMass FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
```

[12]:

TotalPayloadMass
45596

- **Explanation:** NASA's boosters carried a total of 45,596 kg of payload.

Average Payload Mass by F9 v1.1

- **Task:** Calculate the average payload mass carried by booster version F9 v1.1
- **Query:** %sql SELECT AVG(PAYLOAD_MASS__KG_) as AveragePayloadMass FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
- **Result:**

Task 4

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

```
[13]: %sql SELECT AVG(PAYLOAD_MASS__KG_) as AveragePayloadMass FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
```

[13]:

AveragePayloadMass
2928.4

- **Explanation:** NASA's boosters carried an average of 2,928.4 kg of payload.

First Successful Ground Landing Date

- **Task:** Find the dates of the first successful landing outcome
- **Query:** %sql SELECT MIN(Date) as FirstSuccessfulLandingDate FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success (ground pad)';
- **Result:**

Task 5

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

Hint:Use min function

```
[14]: %sql SELECT MIN(Date) as FirstSuccessfulLandingDate FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
```

[14]:

FirstSuccessfulLandingDate
2015-12-22

- **Explanation:** On December 12, 2012, was the first successful landing on a ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

- **Task:** List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- **Query:** %sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
- **Result:**

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
[15]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
 * sqlite:///my_data1.db
Done.
```

[15]:

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

Total Number of Successful and Failure Mission Outcomes

- **Task:** Calculate the total number of successful and failure mission outcomes
- **Query:** %sql SELECT Mission_Outcome, COUNT(*) as TotalCount FROM SPACEXTABLE GROUP BY Mission_Outcome;
- **Result:**

Task 7

List the total number of successful and failure mission outcomes

```
[16]: %sql SELECT Mission_Outcome, COUNT(*) as TotalCount FROM SPACEXTABLE GROUP BY Mission_Outcome;
 * sqlite:///my_data1.db
Done.
```

[16]:

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

Boosters Carried Maximum Payload

- **Task:** List the names of the booster which have carried the maximum payload mass
- **Query:** %sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
- **Result:**

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

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

[17]:

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

- **Task:** List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- **Query:** %sql SELECT substr(Date, 6, 2) as Month, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTABLE WHERE substr(Date, 0, 5) = '2015' AND Landing_Outcome LIKE 'Failure (drone ship)';
- **Result:**

Task 9

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.

```
[18]: %sql SELECT substr(Date, 6, 2) as Month, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTABLE WHERE substr(Date, 0, 5) = '2015' AND Landing_Outcome LIKE 'Failure (drone ship)';
 * sqlite:///my_data1.db
Done.
```

[18]:

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

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

- **Task:** 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
- **Query:** %sql SELECT Landing_Outcome, COUNT(*) as OutcomeCount FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY OutcomeCount DESC;
- **Result:**

Task 10

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.

```
[19]: %sql SELECT Landing_Outcome, COUNT(*) as OutcomeCount FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY OutcomeCount DESC;
 * sqlite:///my_data1.db
Done.
```

[19]:

Landing_Outcome	OutcomeCount
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

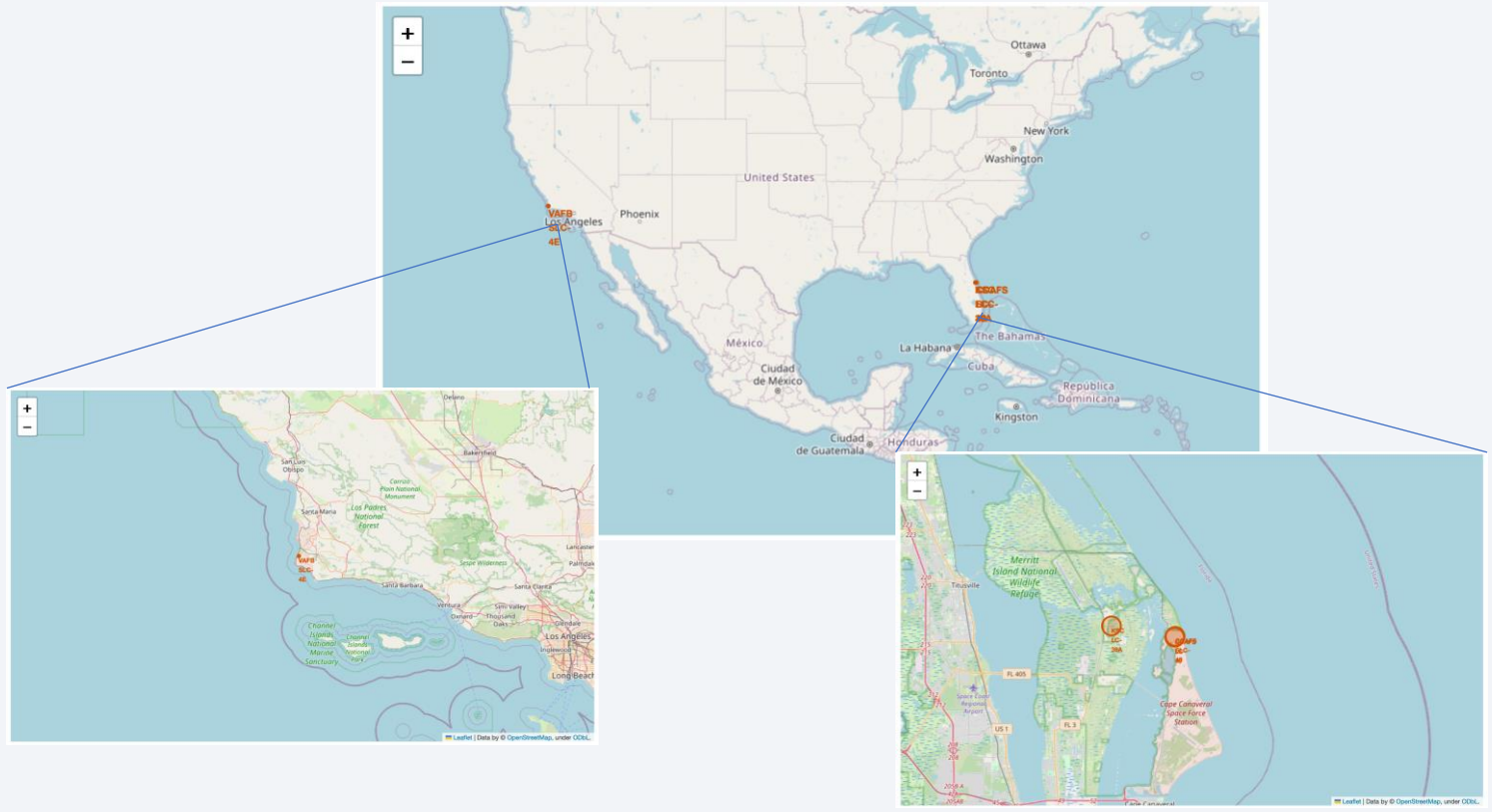
Section 3

Launch Sites Proximities Analysis

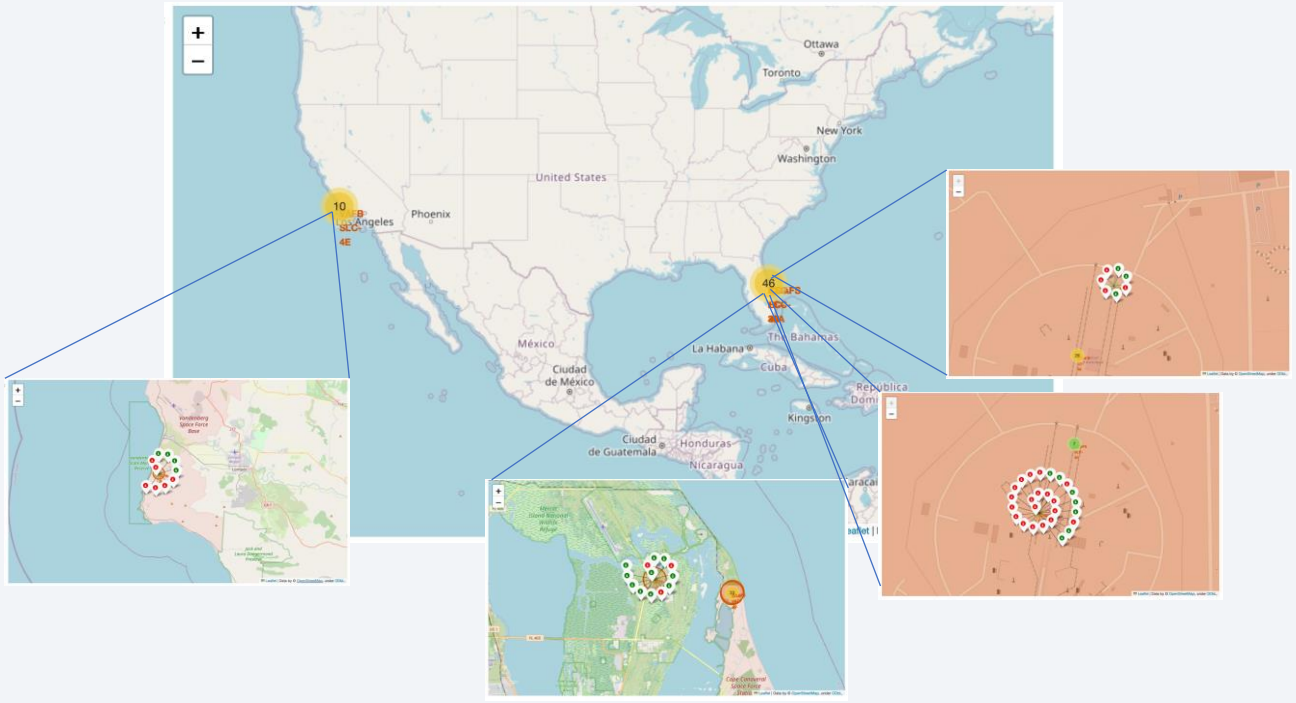
Map of All Launch Site Location

The map was created with circles and burnt orange markings for each of the four launch sites:

- Three in Florida (CCAFS LC-40, CCAFS SLC-40, and KSC LC-39A)
- One in California (VAFB SLC-4E)



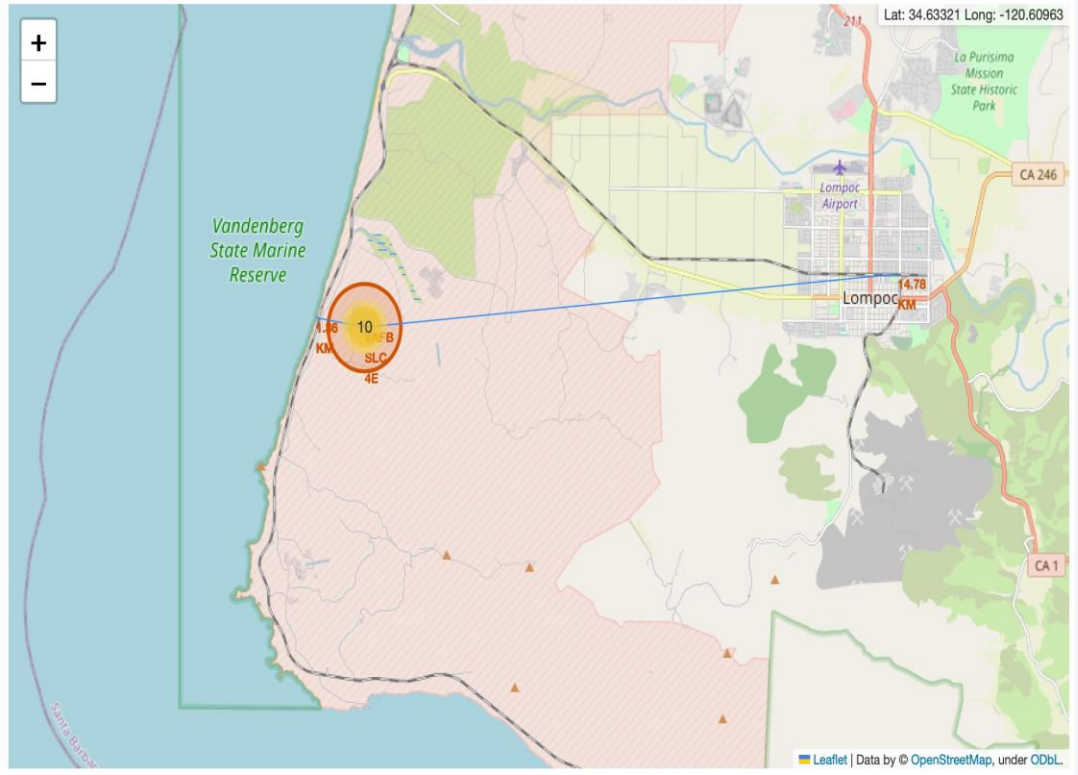
Map of Launch Outcomes

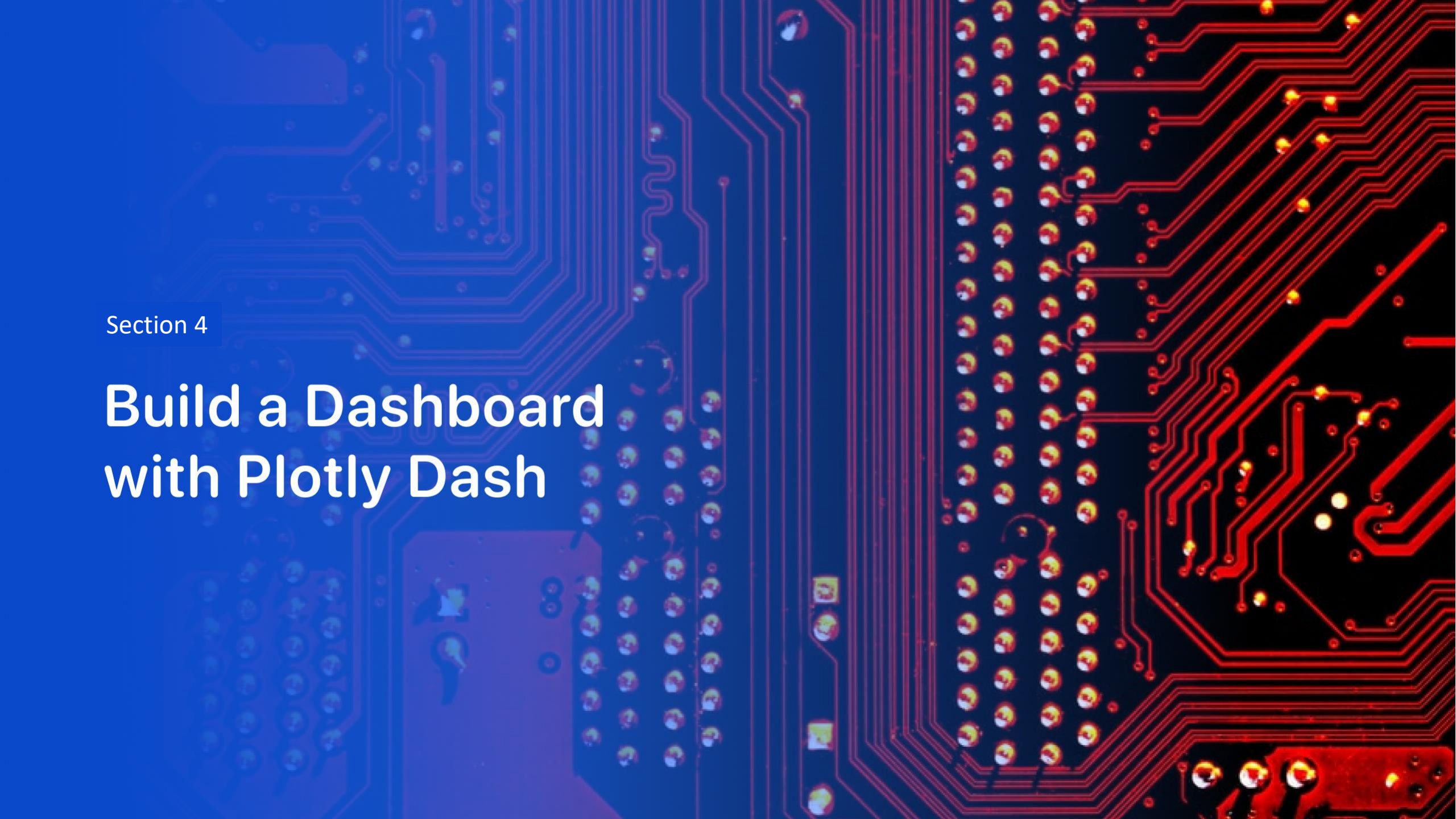


- Clusters, or yellow circles, represent the launch count at each point on a broader scale.
- Upon choosing a site, the results are displayed for every launch.
    - Green denotes accomplishment.
    - Failure is shown with red.

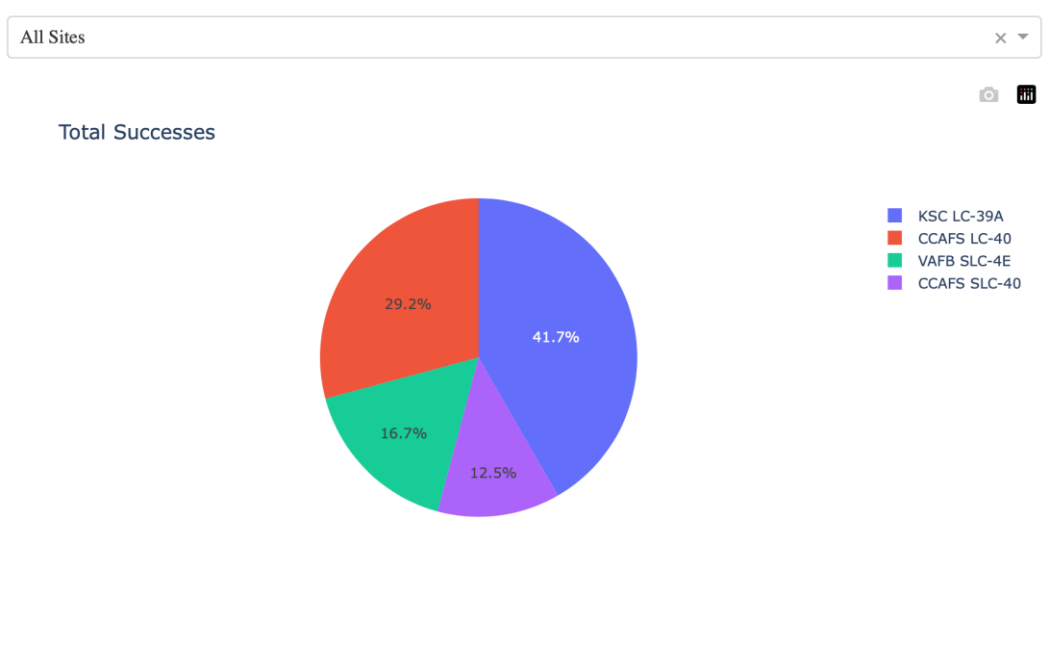
Map of Launch Site Infrastructure

- After ten launches, it was discovered that VAFB SLC-4E was located approximately 1.36 km from the ocean, next to a railway, and 14.78 km from the closest city, Los Angeles.



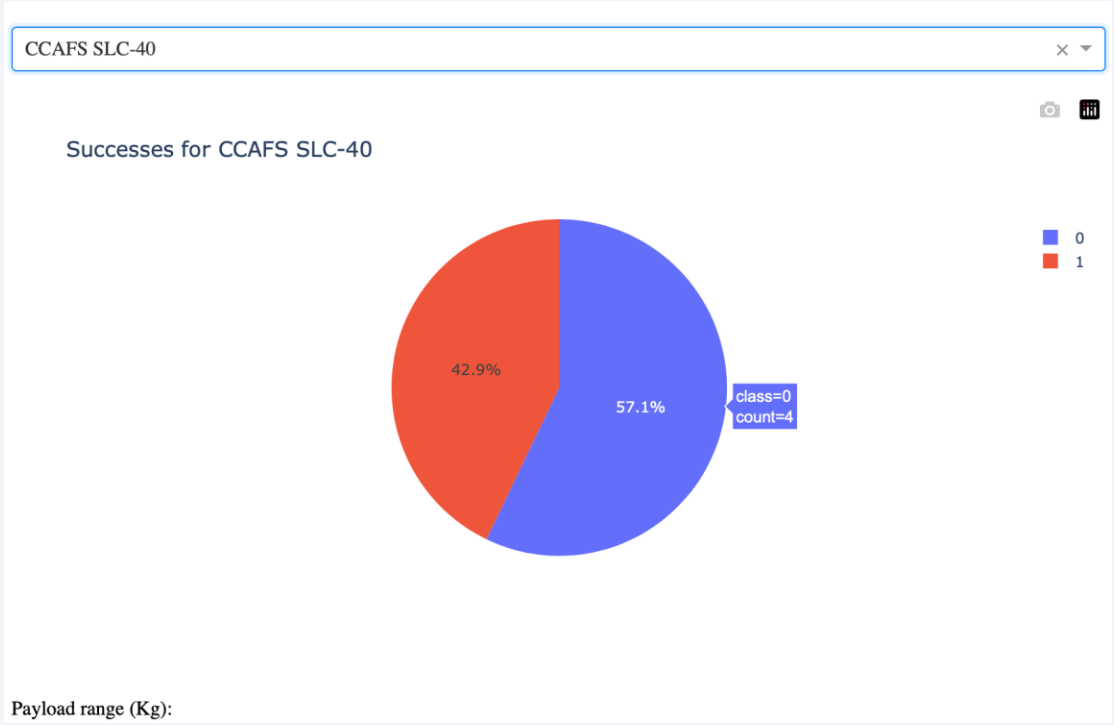

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count On All Sites



- The most successful outcomes were at KSC LC-39A, while the least successful outcomes were at CCAFS SLC-40. CCAFS LC-40 and VAFB SLC-4E were in the centre of the pack in terms of successful outcomes.

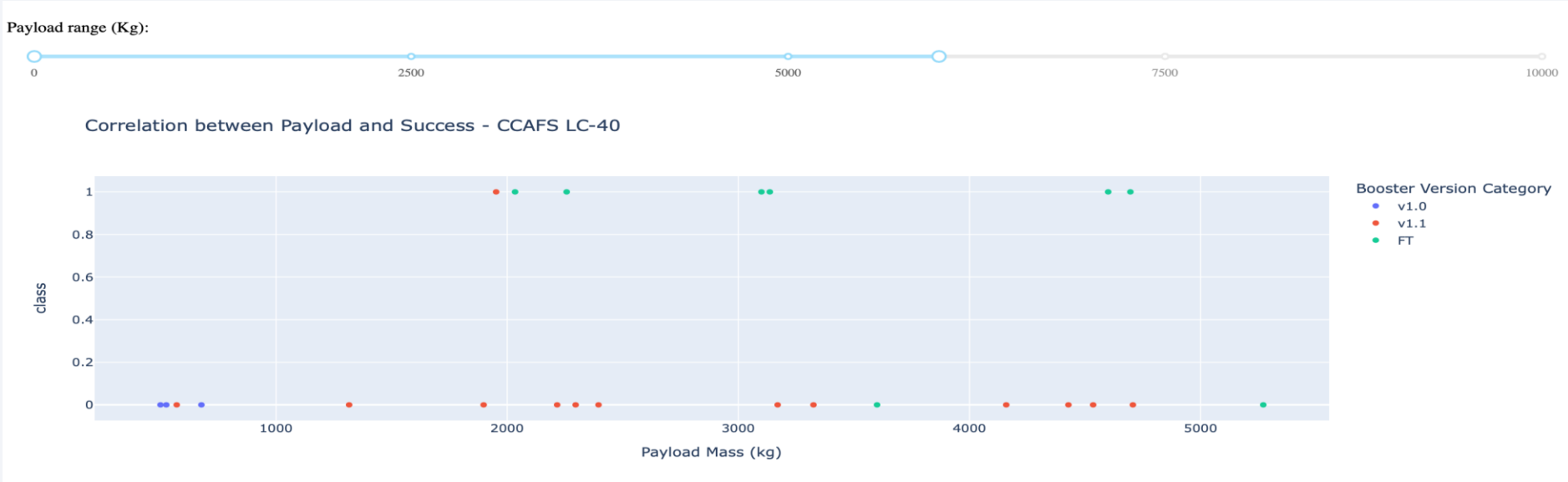
Highest Launch Success Ratio at the Launch Site



- The highest launch success ratio was achieved by CCAFS SLC-40, with 42.9% of launches ending successfully. To put it another way, 3 out of 7 launches were successful and 4 out of 7 launches were unsuccessful.

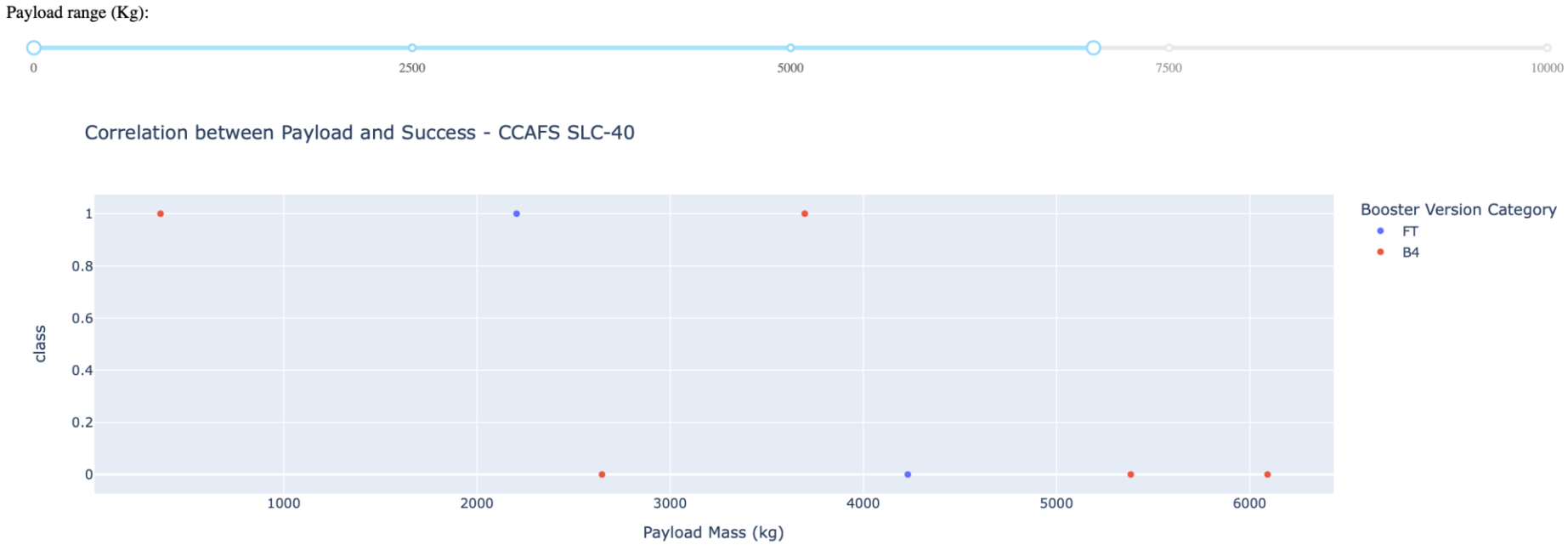
Payloads versus the result of launch: CCAFS LC-40

- The FT Booster performed better at Launch Site CCAFS LC-40, particularly when carrying payload masses of 2,000–5,000 kg.



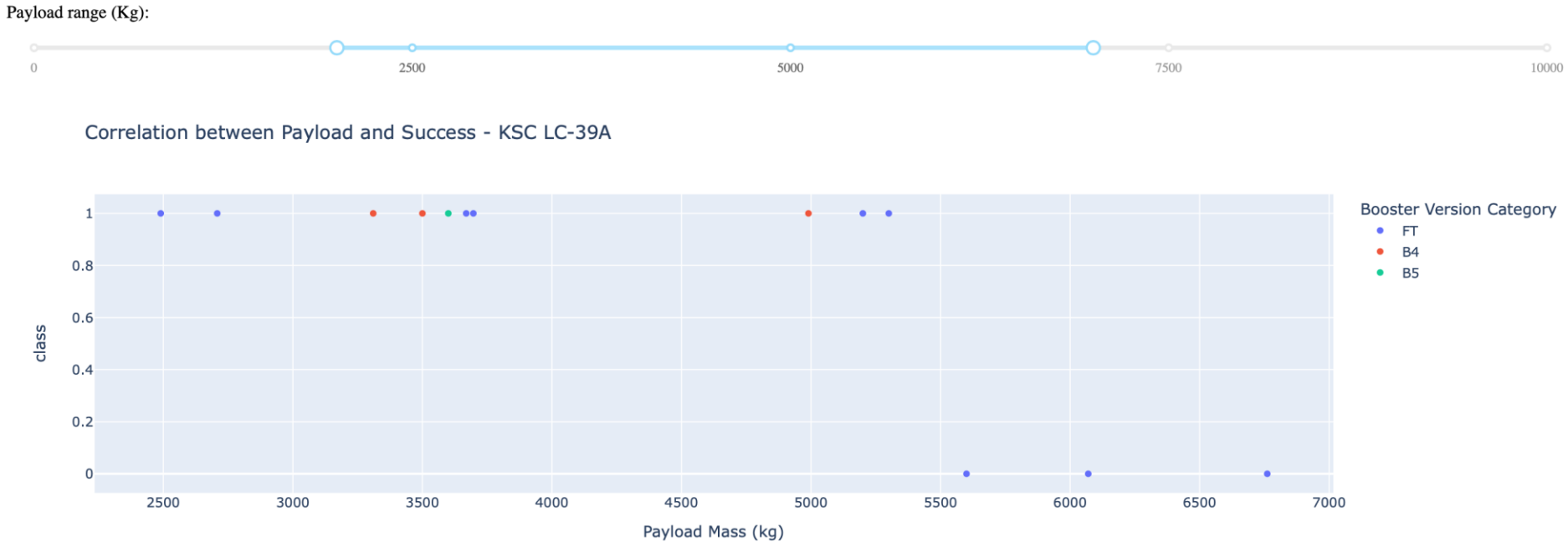
CCAFS SLC-40 Payloads vs Launch Outcome

- The B4 Booster performed better at Launch Site CCAFS SLC-40, particularly with payload masses under 4,000 kg.



Payloads versus the result of launch: KSC LC-39A

- The FT Booster performed well at Launch Site KSC LC-39A, particularly when carrying payload masses of between 2,000 and 5,500 kg.



VAFB SLC-4E Payloads versus Launch Outcome

- The FT and B4 Boosters at Launch Site VAFB SLC-4E achieved a same success rate for the whole payload mass range of 0 to 10,000 kg.

Launch Outcome versus Payloads: Every Site

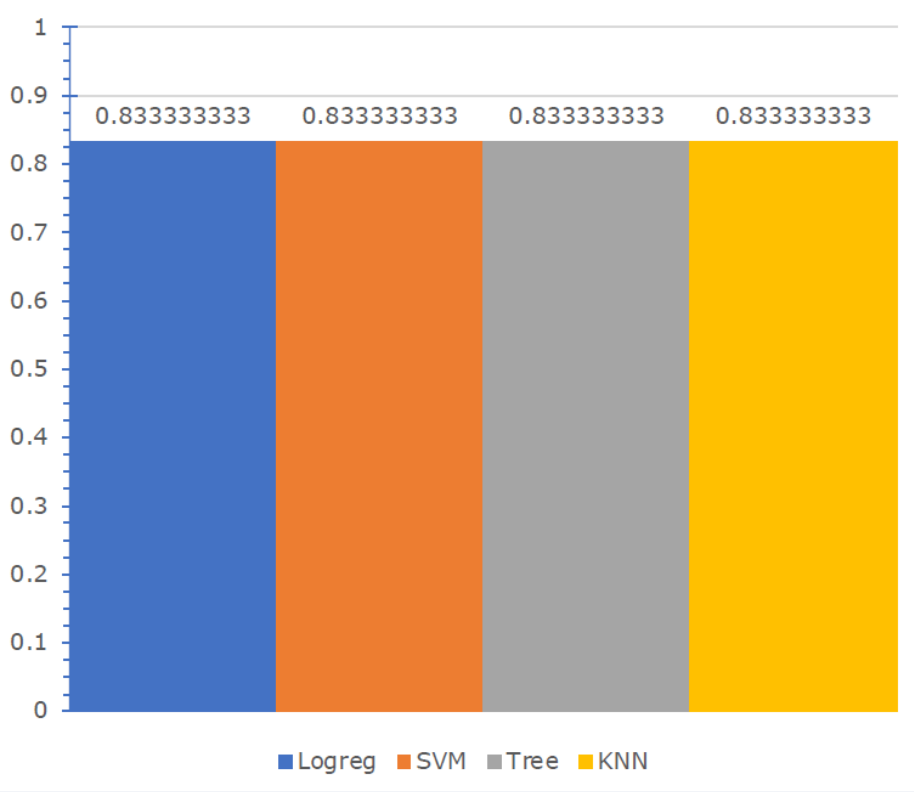
- The FT Booster was the most successful overall across all locations, particularly with payload masses between 2,000 and 5,500 kg; nevertheless, there was one exceptional success with a payload mass closer to 500 kg.

Section 5

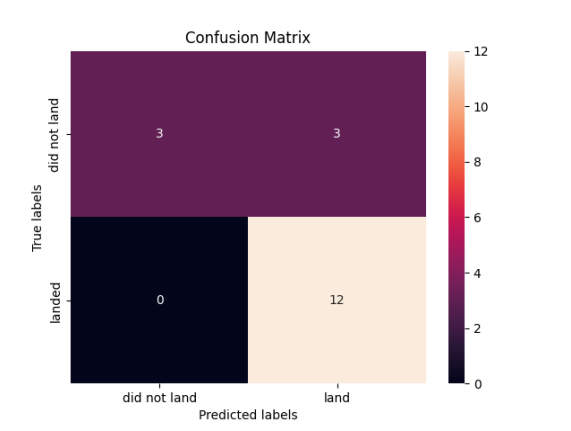
Predictive Analysis (Classification)

Classification Accuracy

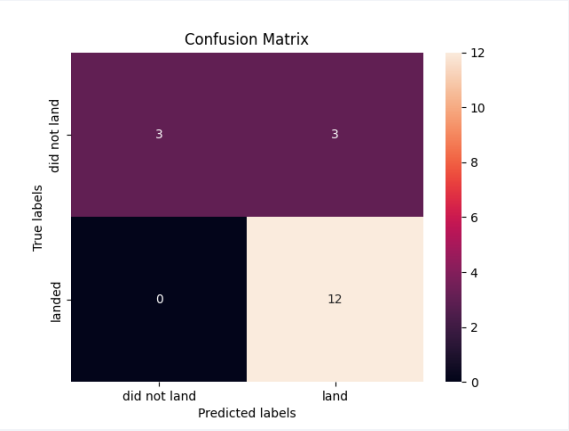
- The classification accuracy score of roughly 0.8333 was shared by the four models (K-Nearest Neighbour, SVM, Classification Tree, and Logistic Regression).



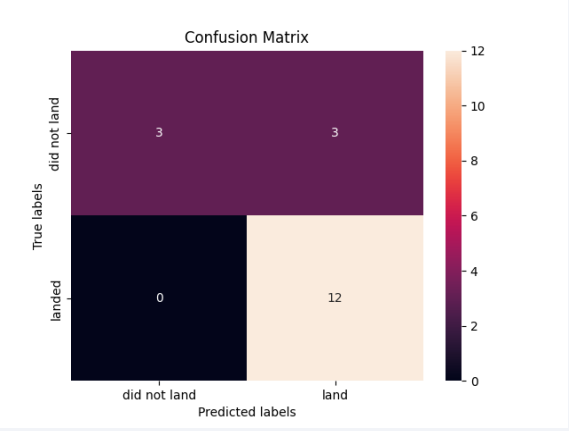
Confusion Matrix

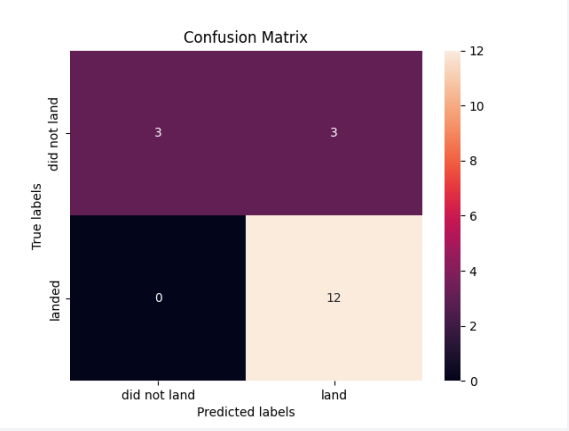


Logistic Regression



SVM





- All four models' confusion matrices seem to be the same, meaning that none of them outperformed the others.
  - Twelve touched down, as predicted.
  - Three were not expected to arrive and didn't.
  - Three were supposed to land, but they didn't.
  - Of the 0 that were expected to land, none did.

Conclusions

- Observations relating to Launch Site, Payload Mass, Launch Number, Orbit Type, and Landing Outcome were obtained from the retrieved Space-X Launch Data.
- Over time, Space-X launches have been increasingly successful overall.
- There are some geographic and infrastructure characteristics that launch sites typically share. For example, launch sites close to coastlines provide safer failure landings, as do those that are far from towns and major thoroughfares.
- In terms of forecasting the landing result of a launch, all four machine learning models fared similarly well. The conclusions could have been different if there had been more data, however the data sets were quite limited.

Thank you!