IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jahnvi Trivedi
30 June 2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection – APIs & Webscraping
  - Data Wrangling
  - Exploratory Data Analysis (EDA)
  - Data Visualization
  - Interactive Map
  - Dashboard
  - Predictive Analysis
- Summary of all results
  - 66% success rate of landing, EDA shows relationship between success rate and many features including flight number, orbit, payload mass.
  - Predictive Analysis – 83% accuracy models built

Introduction

- The commercial space industry is growing fast, with companies like Virgin Galactic, Rocket Lab, and SpaceX leading the way. SpaceX has done impressive things, such as sending spacecraft to the International Space Station and launching its Starlink satellite network. The Falcon 9 rocket is one of SpaceX's main rockets. It has two parts, and the first stage does most of the work. SpaceX can reuse this first stage, which helps lower the cost of launches..
- The project involved: using data and machine learning to predict if the Falcon 9's first stage will land successfully. This will help Space Y improve their chances in the space launch business. We wish to answer the following questions:
  - 'How do the different variables such as orbit type, payload mass, and flight number affect the success rate?'
  - 'Can we accurately predict whether the first stage will land successfully?'

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches
- Perform data wrangling
  - Deal with missing values, determine landing outcome
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, tune and evaluate multiple different classification models to determine best model.

Data Collection

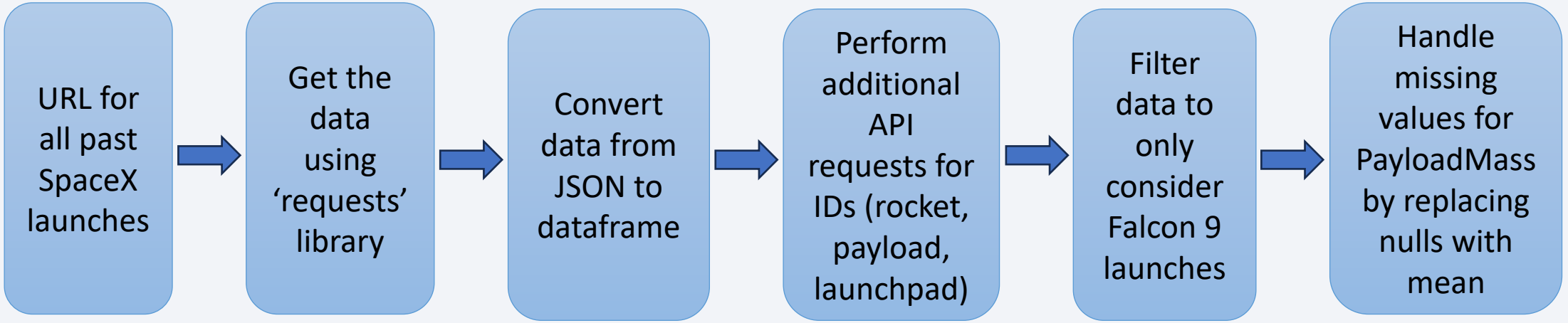
In this project, data about SpaceX launches was collected using the SpaceX REST API. This API provides detailed information about each launch, such as the rocket used, the payload, launch details, landing details, and the landing outcome.

In addition to the API, we also used web scraping to collect Falcon 9 launch data from Wikipedia pages.

Together, this provided a comprehensive set of features that could be used to analyse the data and further determine if it is possible to predict the successful landing of the first stage.

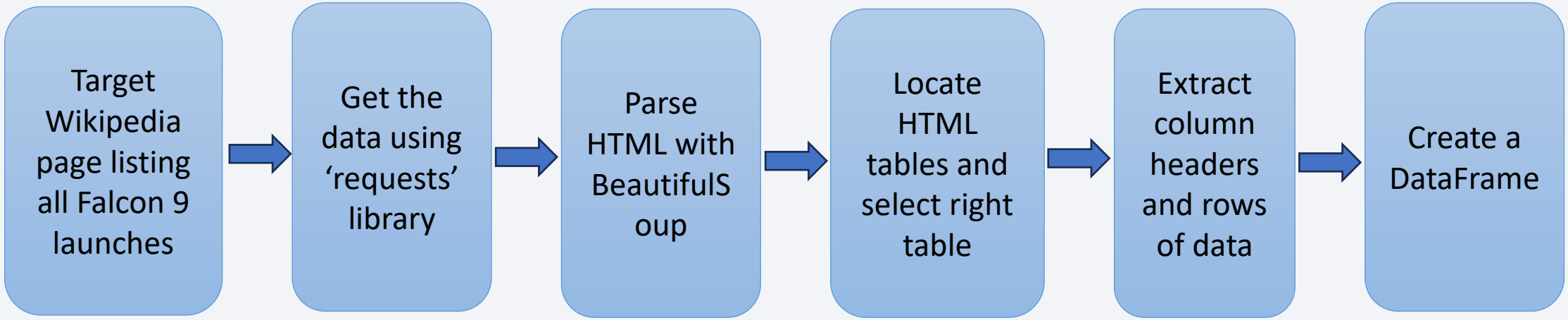
Data Collection – SpaceX API

- A big part of this project involved gathering launch data from the SpaceX REST API. This API is publicly available and provides structured data about launches, rockets, payloads, and more.
- https://github.com/jahnvi-28/IBM-Applied-Data-Science-Capstone/blob/main/1.%20Data%20Collection%20-%20APIs.ipynb



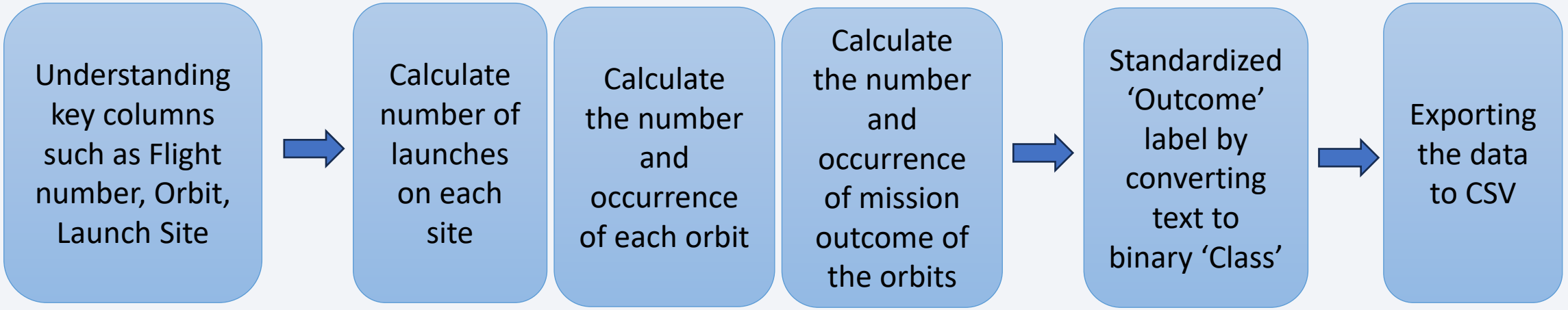
Data Collection - Scraping

- Besides using the SpaceX API, another part of the project involved collecting Falcon 9 launch data by web scraping Wikipedia pages. This is helpful because some launch details are summarized in tables that are easier to read and cross-check.
- https://github.com/jahnvi-28/IBM-Applied-Data-Science-Capstone/blob/main/2.%20Data%20Collection%20-%20Webscraping.ipynb



Data Wrangling

- After collecting the SpaceX launch data, the next step was to clean, organize, and transform it so it could be used for analysis and modeling. This process is called data wrangling.
- https://github.com/jahnvi-28/IBM-Applied-Data-Science-Capstone/blob/main/3.%20Data%20Wrangling.ipynb



EDA with Data Visualization

- In this project, EDA was the first step to understand the SpaceX launch data and spot patterns.
- The following charts were produced:
  - Scatter plot: Visualize the relationship between Flight Number and Launch Site
  - Scatter plot: Visualize the relationship between Payload Mass and Launch Site
  - Bar plot: Visualize the relationship between success rate of each orbit type
  - Scatter plot: Visualize the relationship between FlightNumber and Orbit type
  - Scatter plot: Visualize the relationship between Payload Mass and Orbit type
  - Line chart: Visualize the launch success yearly trend
- Additionally, feature engineering was performed to apply OneHotEncoder to categorical variables and convert entire dataframe to numeric.
- https://github.com/jahnvi-28/IBM-Applied-Data-Science-Capstone/blob/main/5.%20EDA%20-%20Data%20Visualization.ipynb

- Scatter plots were used to visualize if a relationship existed between two variables.
- Bar charts were used to compare categorical variables.
- Line charts were useful in showing trends in data over time.

EDA with SQL

- The following SQL queries were performed:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first succesful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.
  - 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.
  - 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.
- https://github.com/jahnvi-28/IBM-Applied-Data-Science-Capstone/blob/main/4.%20EDA%20-%20SQL.ipynb

Build an Interactive Map with Folium

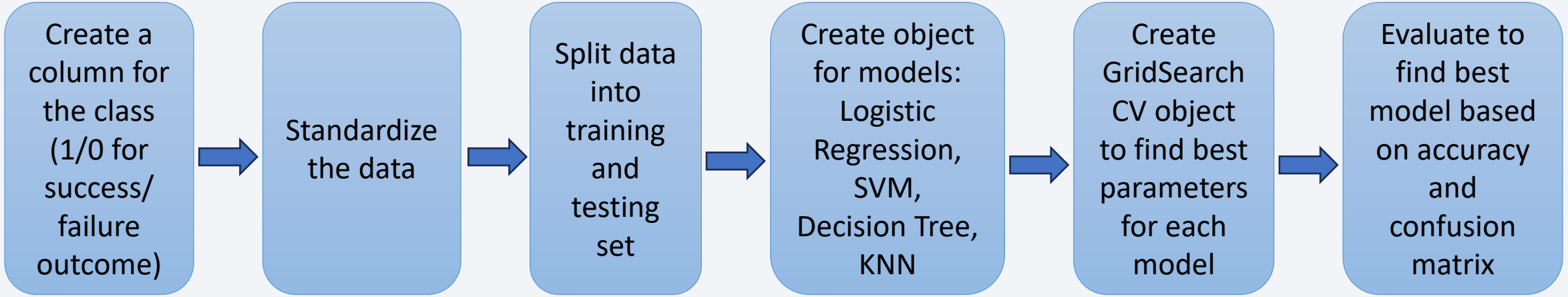
- In this project, we used Folium to create an interactive map displaying SpaceX launch sites and related details. The following Map objects were created and added:
  - Markers – Added markers at each launch site's latitude and longitude to pinpoint exact locations. These markers help users quickly identify where launches happen on the map.
  - Circles – Added circles around each launch site with a certain radius to highlight the area or proximity around each site. This visually emphasizes the geographic influence or coverage of each site.
  - Marker clusters – Used marker clusters to group nearby launch markers together, improving map readability and performance when many markers are close to each other.
- These objects provided a clear visual representation of launch site locations, helped highlight areas around launch sites, and improved user interaction by clustering markets, avoiding clutter and allowing users to zoom and explore the data smoothly
- https://github.com/jahnvi-28/IBM-Applied-Data-Science-Capstone/blob/main/6.%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- In this project, we built an interactive dashboard using Plotly Dash to explore SpaceX launch data dynamically.
- The following interactions were included:
  - Dropdown menu for launch site
  - Range slider for payload mass
- The following plots and graphs were added:
  - Pie chart: showing the proportion of successful vs failed launches. This dynamic chart updates based on the selected launch site, helping users compare success rates across sites or over all sites.
  - Scatter plot: Displaying the correlation between payload and success for the selected site or for all sites, with points colored by success/failure. This helps visualize how different factors influence launch outcomes.
- https://github.com/jahnvi-28/IBM-Applied-Data-Science-Capstone/blob/main/7.%20spacex-dash-app.py

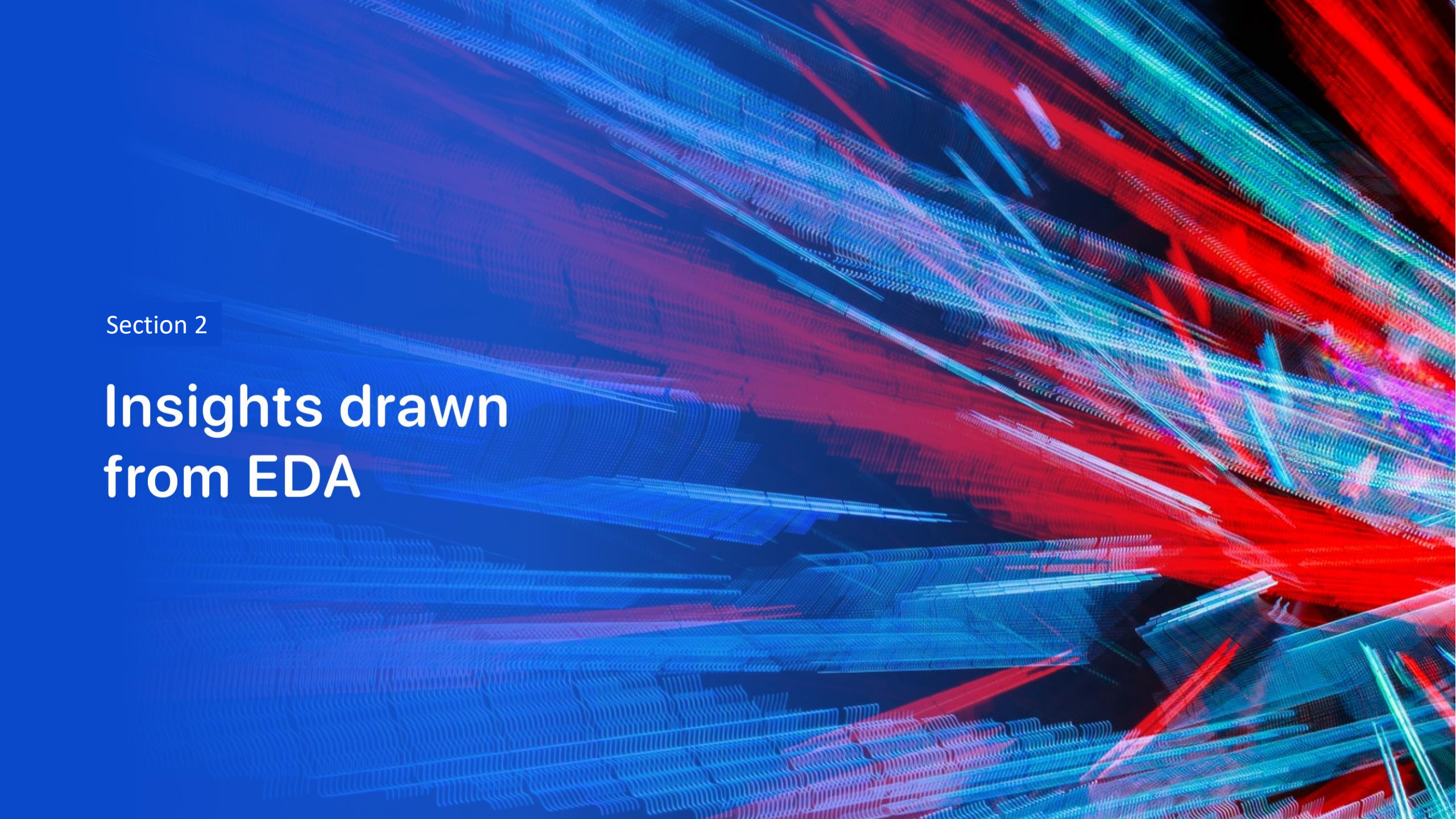
Predictive Analysis (Classification)

- In the predictive analysis phase, we built a machine learning pipeline to predict whether the Falcon 9's first stage lands successfully. The flow chart below outlines the key steps included in this process.
- https://github.com/jahnvi-28/IBM-Applied-Data-Science-Capstone/blob/main/8.%20SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

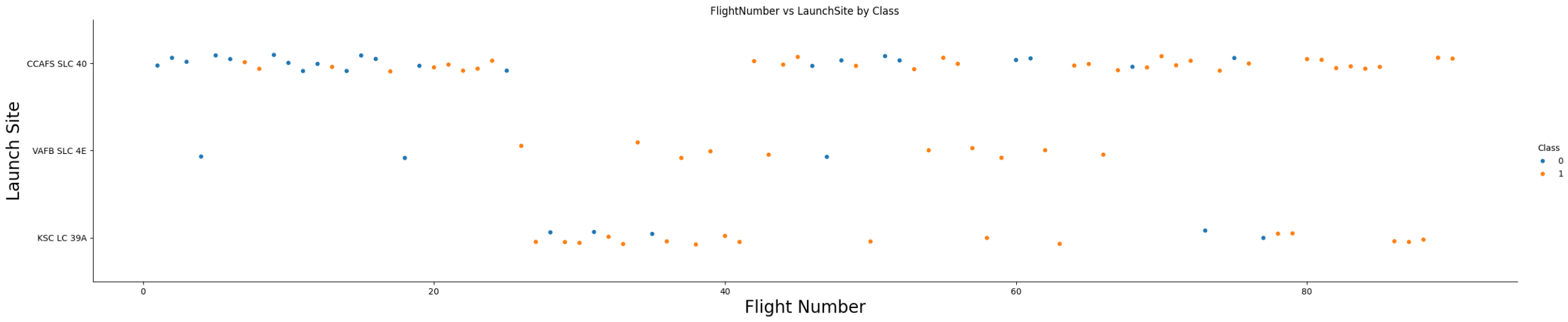


Results

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


Section 2
Insights drawn from EDA

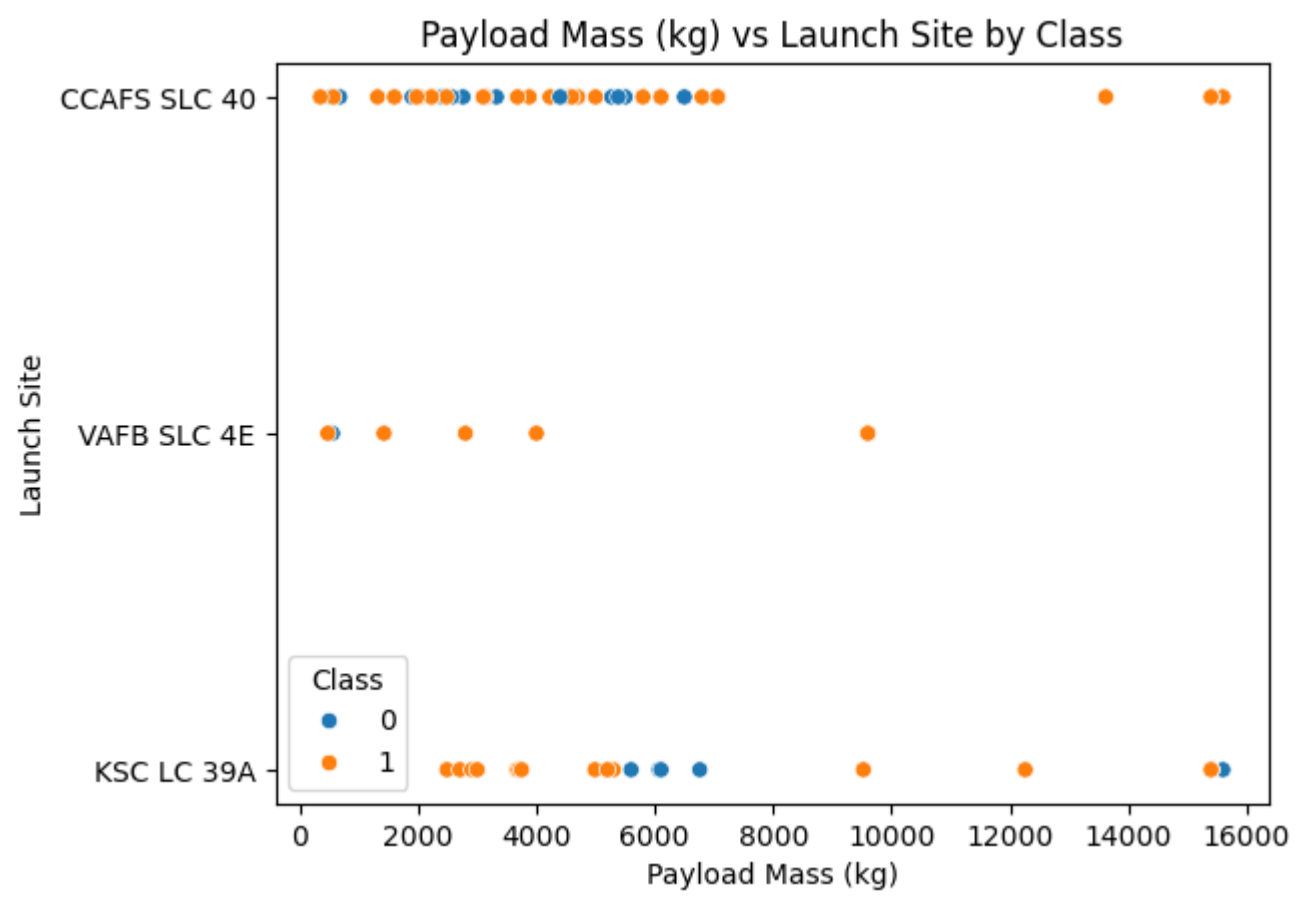
Flight Number vs. Launch Site



- Observations:
  - As the flight number increases, the first stage is more likely to land successfully.
  - CCAFS SLC 40 appears to have more unsuccessful landings, while VAFB SLC 4E and KSC LC 39A have a higher proportion of successful landings.
  - CCAFS SLC 40 has the majority of launches

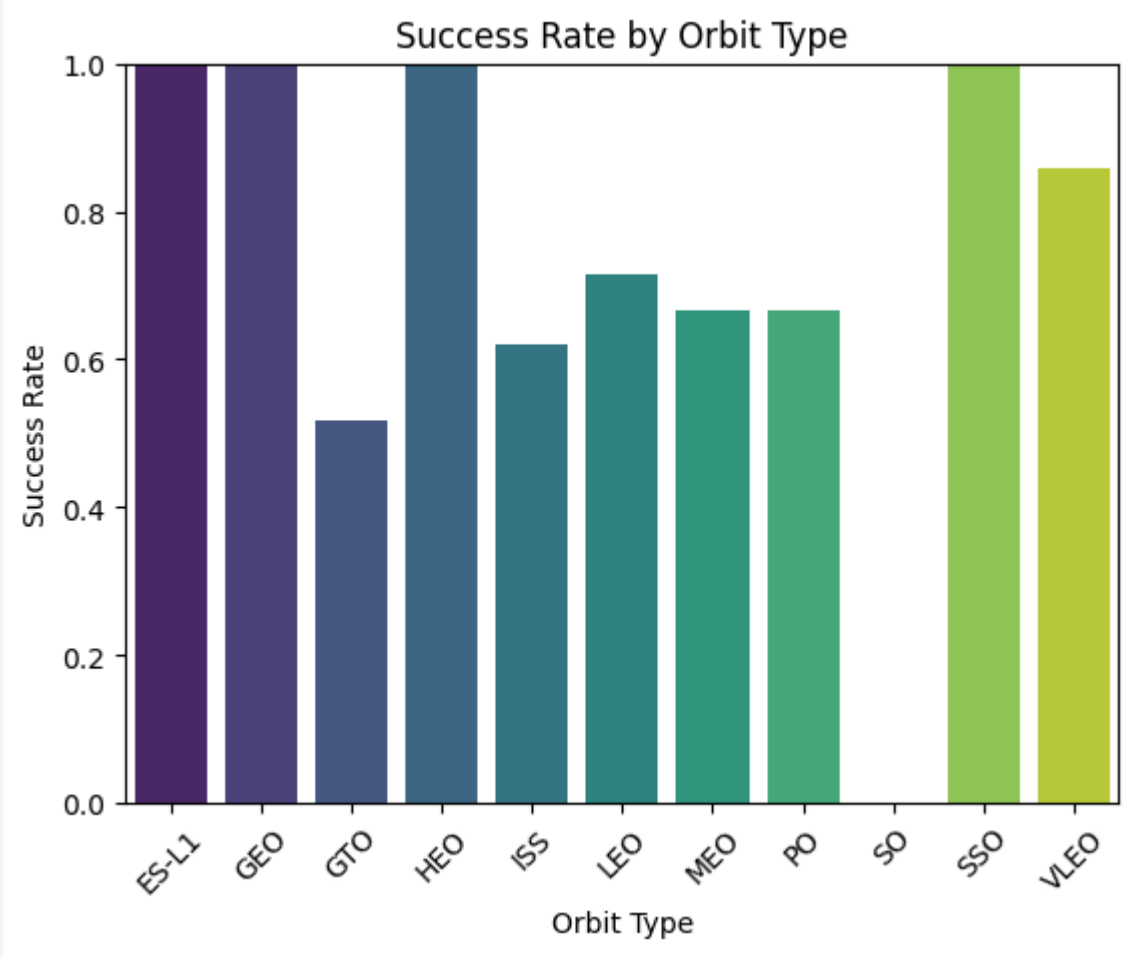
Payload vs. Launch Site

- Observations:
  - VAFB SLC 4E has no rockets launched for heavy payload mass (greater than 10000)
  - There appears again to be more failed landings in CCAFS SLC 40, however this site also appears to have the most number of launches.
  - Higher payload increases the chances of successful landings.



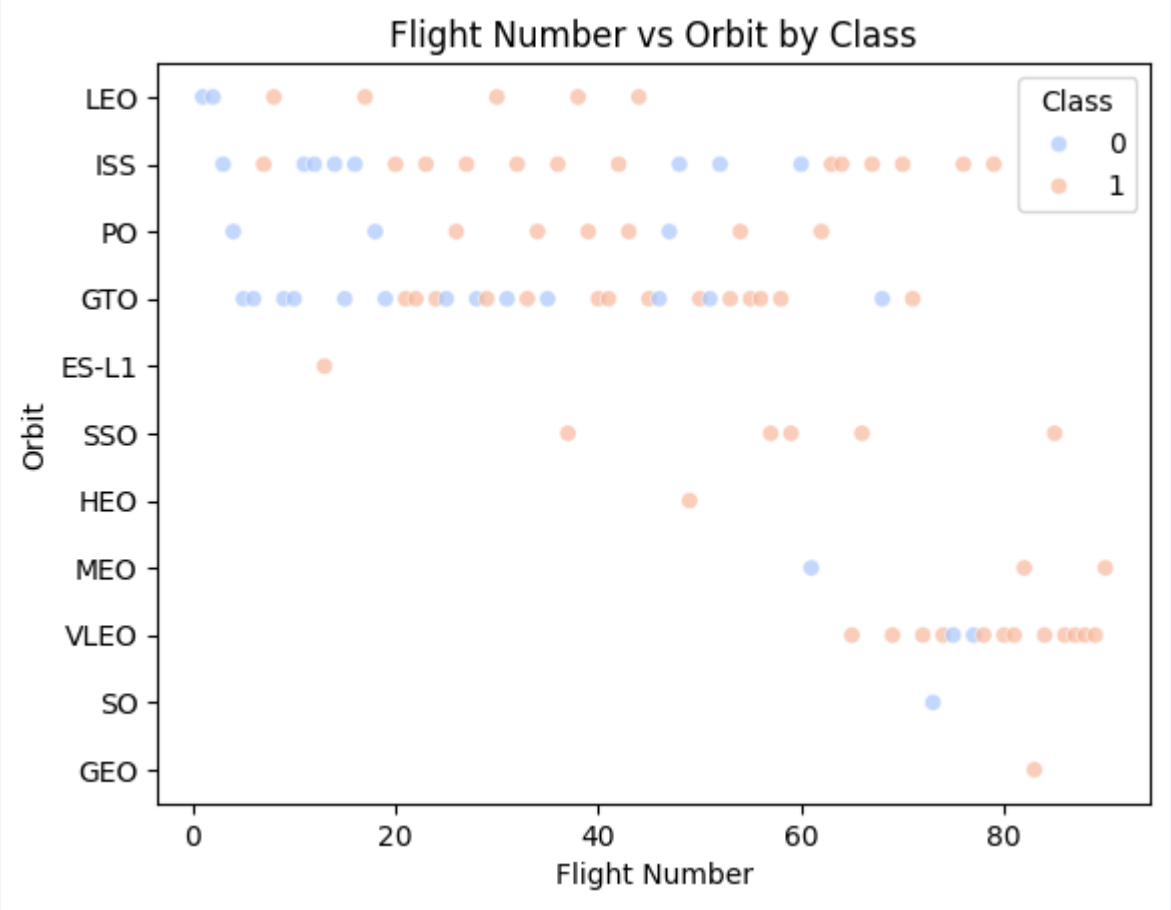
Success Rate vs. Orbit Type

- Observations:
  - The following orbits have 100% success rate: ES-L1, GEO, HEO, SSO
  - SO orbit has 0% success rate
  - The remaining orbits have a success rate between 50% and 85%.



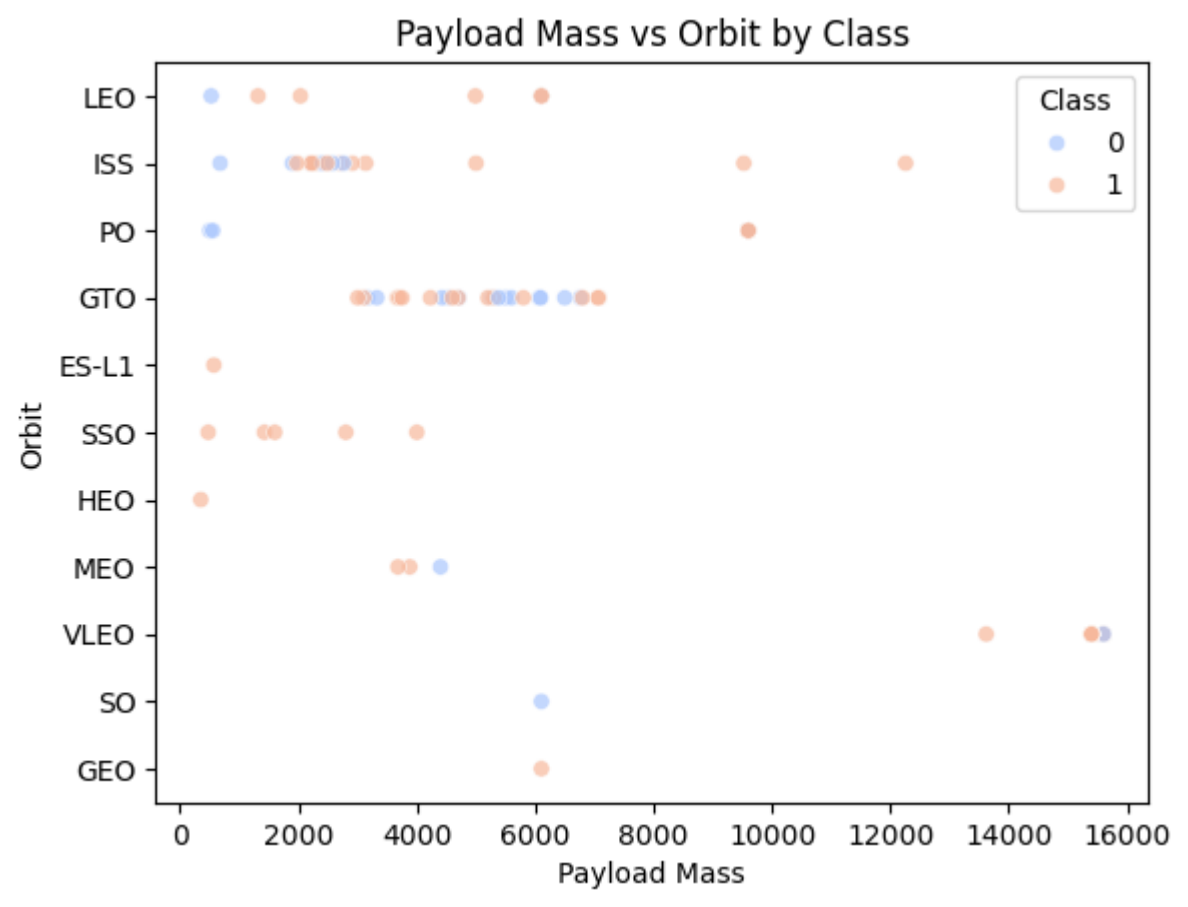
Flight Number vs. Orbit Type

- Observations:
  - In the LEO orbit, success appears to be related to the flight number, while for GTO orbit, this is not the case.
  - Most of the failures are for earlier flight numbers.



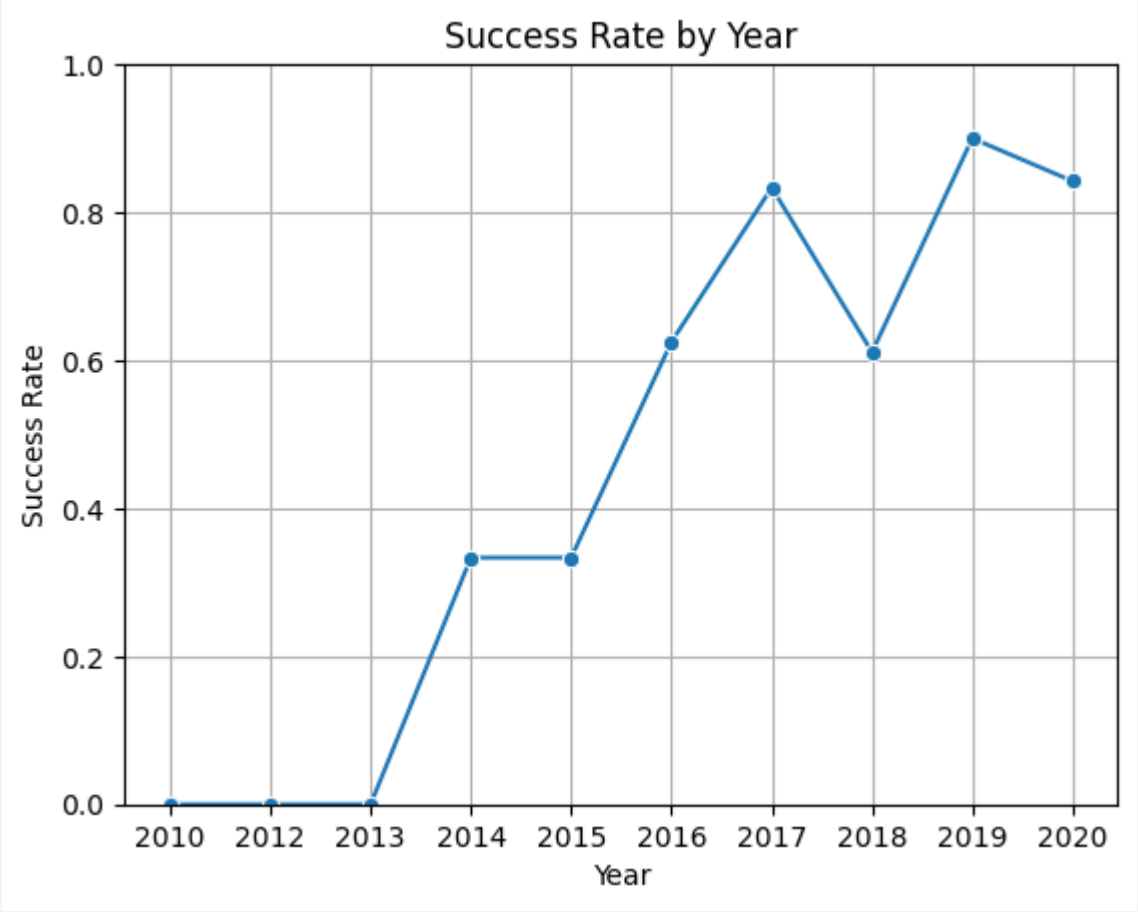
Payload vs. Orbit Type

- Observations:
  - For LEO, ISS, and PO it appears that heavier payloads lead to more successful landings.
  - For GTO there is no clear relationship between success and payload mass.



Launch Success Yearly Trend

- Observations:
  - Evidently, the success rate has been increasing considerably from 2013 to 2020.
  - There is a noticeable dip in success rate in 2018, but this picks back up in 2019, with another slight dip in 2020.



All Launch Site Names

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

[13]:

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

- The above picture outlines the four unique launch site names

Launch Site Names Begin with 'CCA'

```
[14]: %sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

[14]:

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 above picture outlines 5 records where the launch sites begin with string ‘CCA’.

Total Payload Mass

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

[17]:

TotalPayloadMass
45596

- The total payload mass carried by boosters launched by NASA (CRS) is 45596 kgs.

Average Payload Mass by F9 v1.1

```
[19]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AveragePayloadMass FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
[19]:
```

AveragePayloadMass
2928.4

- The average payload mass carried by booster version F9 v1.1 is 2928 kg

First Successful Ground Landing Date

```
[22]: %sql SELECT MIN(Date) AS FirstSuccessfulGroundPadLanding FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
```

[22]:

FirstSuccessfulGroundPadLanding
2015-12-22

- The first successful landing outcome in ground pad was achieved on 22nd December 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

[24]:

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

- The above picture outlines 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

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

[26]:

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

- Total number of successful mission outcomes = 101, failed = 1

Boosters Carried Maximum Payload

- These are the booster versions that have carried the maximum payload mass.

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

[27]:

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

[28]:
```
%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 = 'Failure (drone ship)';
```

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

[28]:

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

- In 2015, the two records which show failed landing outcomes were in months January and April.

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

```
[29]: %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.
```

[29]:

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

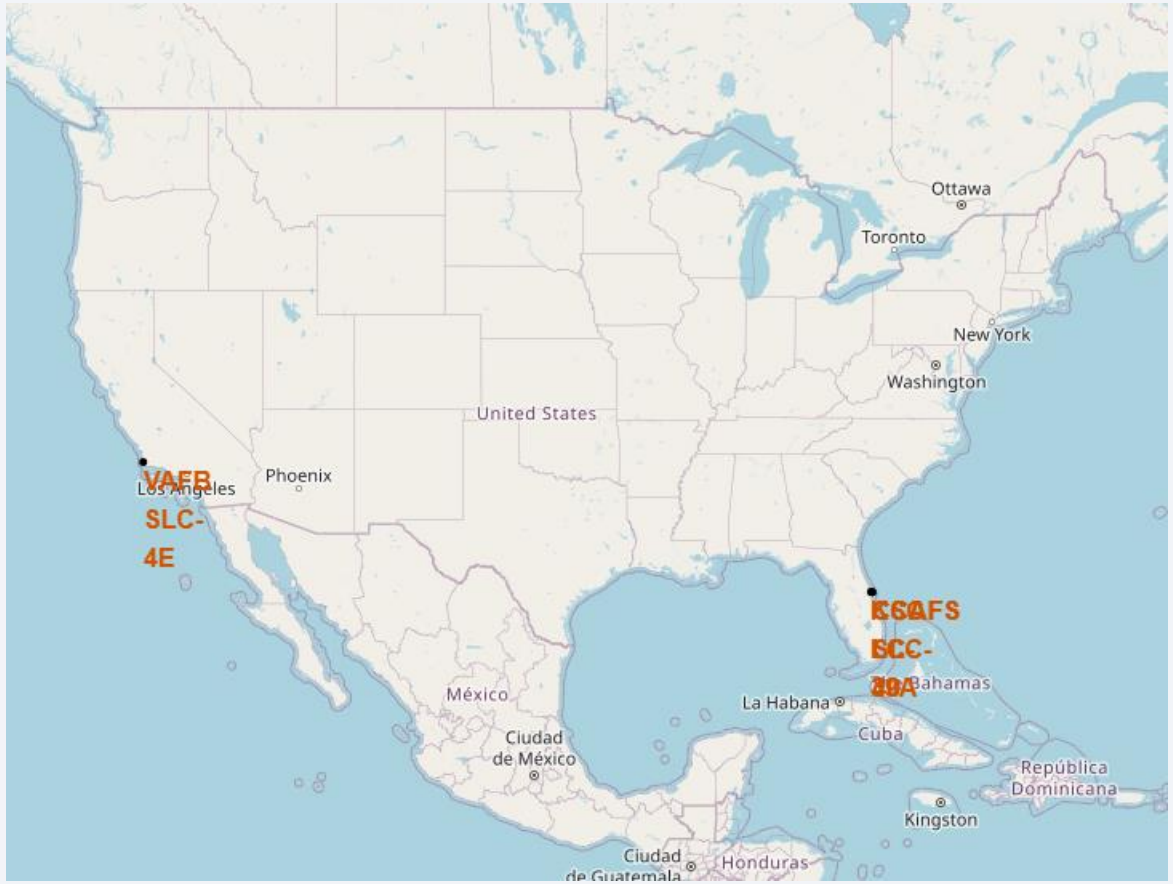
- Between the date 2010-06-04 and 2017-03-20, the common landing outcome was for ‘No attempt’, followed by ‘Success (drone ship)’ and ‘Failure (drone ship)’

Section 3

Launch Sites Proximities Analysis

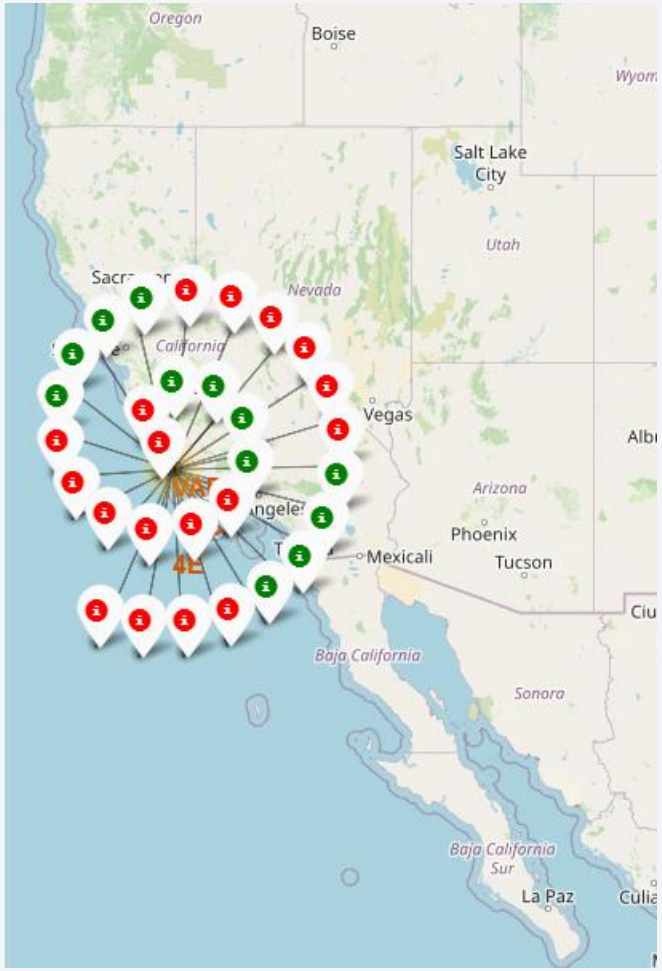
Launch Sites' Location on Map

- This map displays the locations of SpaceX launch sites across the United States.
- Each launch site is marked by a small black circle
- Each site also has a marker containing a label with the launch site's name.
- 3 of the launch sites are clustered in Florida, and 1 is on the West coast in California.



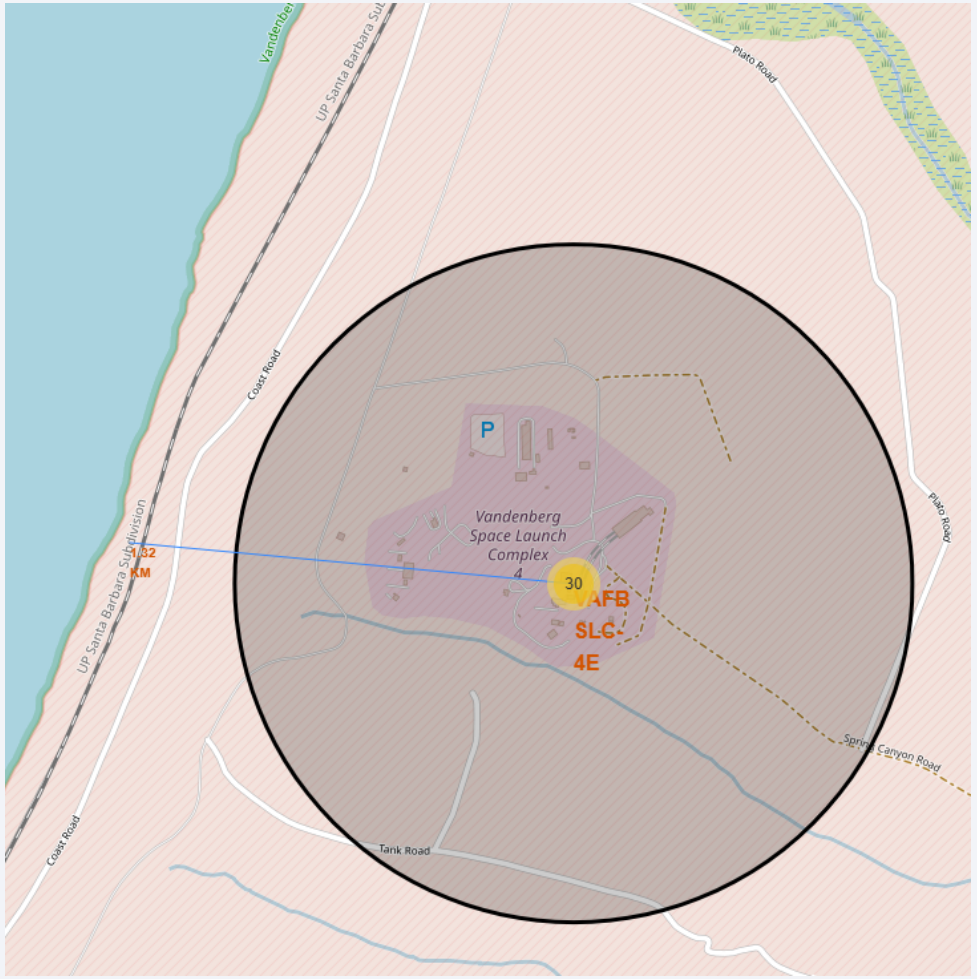
Color-Labeled Launch Outcomes on Map

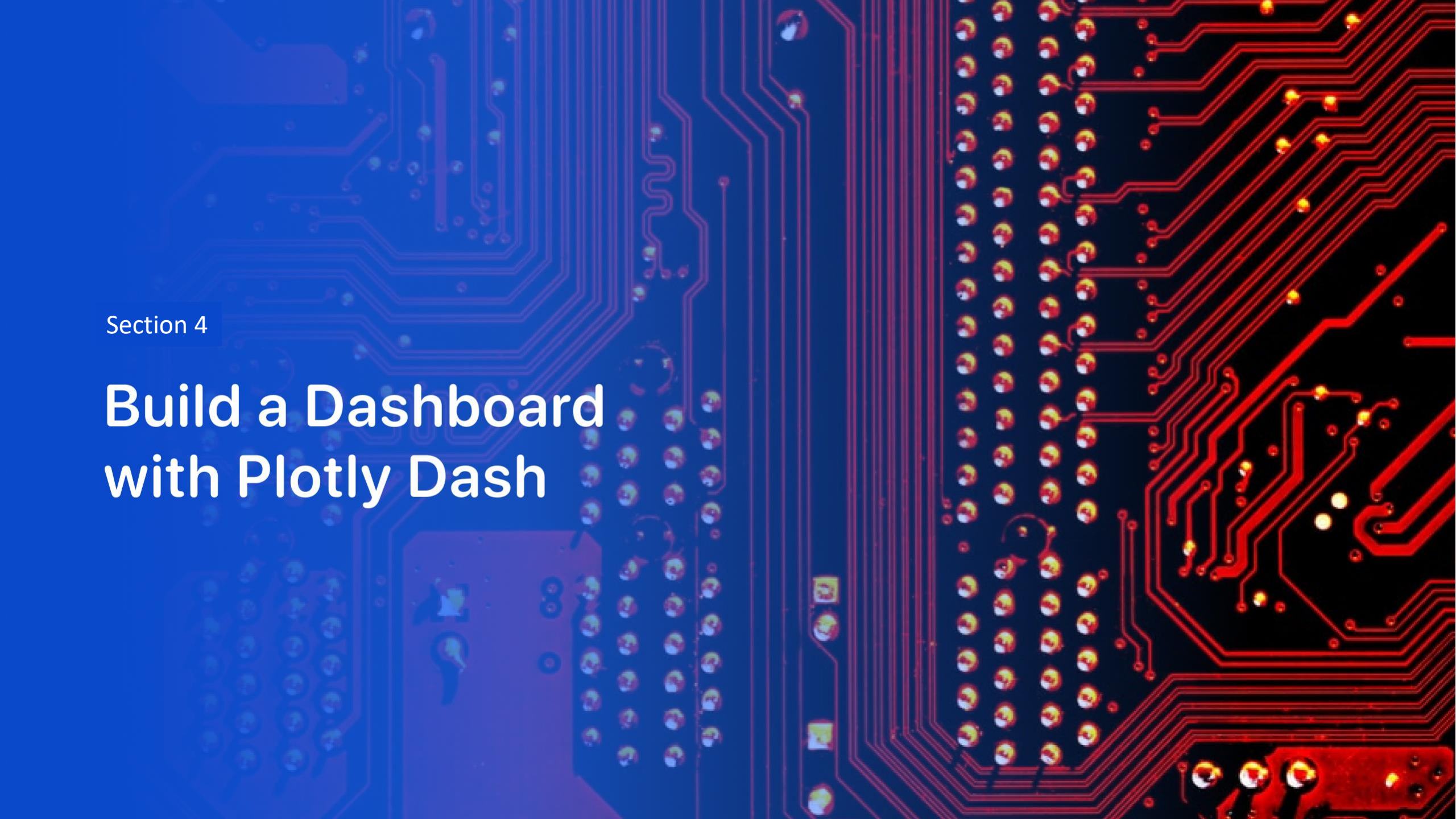
- To enhance the map, markers have been added to show the outcome of each launch – green means a successful launch, while red means a failed launch.
- The marker is shown at the launch coordinates (launch site), and the marker cluster is used to avoid overlap so that markers group together and expand when clicked as shown in the image.



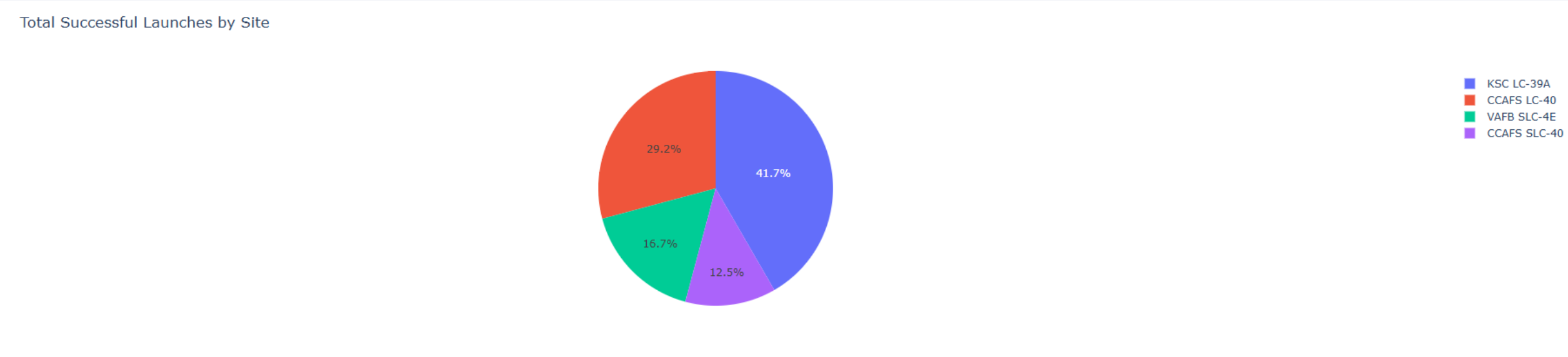
Distance from Launch Site VAFB SLC-4E to Closest Coastline

- The image on the right shows the distance from launch site VAFB SLC-4E to the closest coastline and the distance between them, which is 1.32 km.
- Proximity for launch site to the ocean is important for safety considerations.



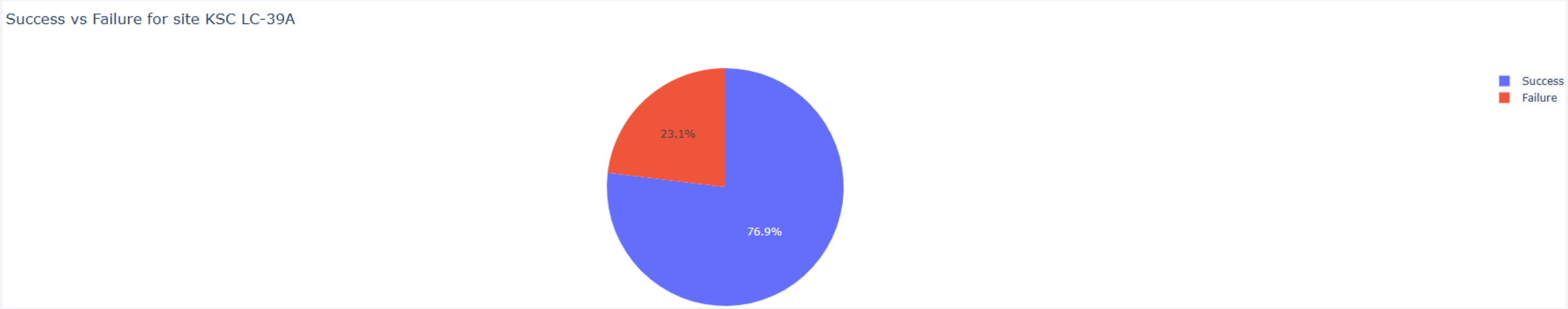

Section 4
Build a Dashboard with Plotly Dash

Total Successful Launches by Site



- KSC LC-39A has the most successful launches, and CCAFS SLC-40 has the least.

Success vs Failure for site KSC LC-39A



- Site KSC LC-39A has 77% successful launches, and 23% failed launches.

Payload vs Launch Outcome for All Sites

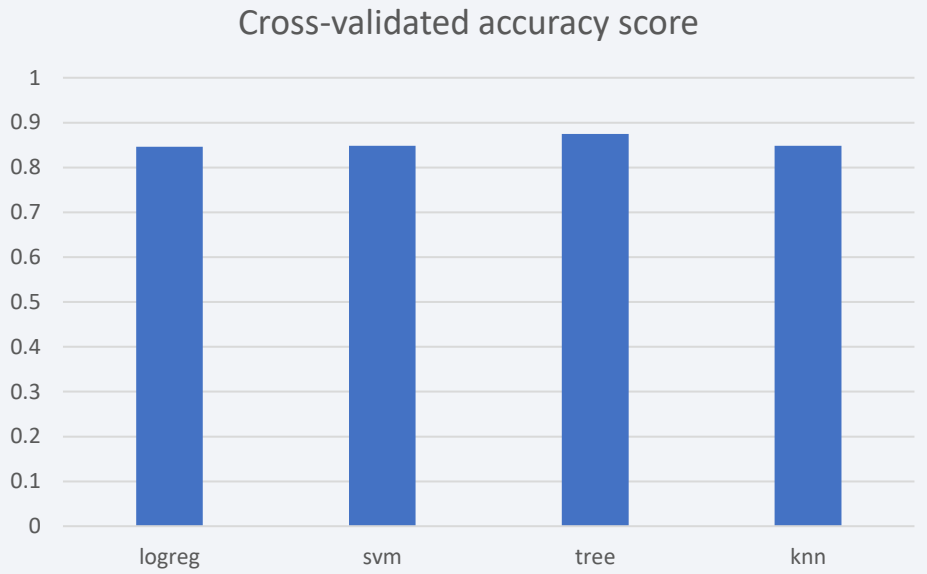
- The image above shows the payload vs launch outcome for payload between 5000 and 10000. There is a lot of failures and little success rate.

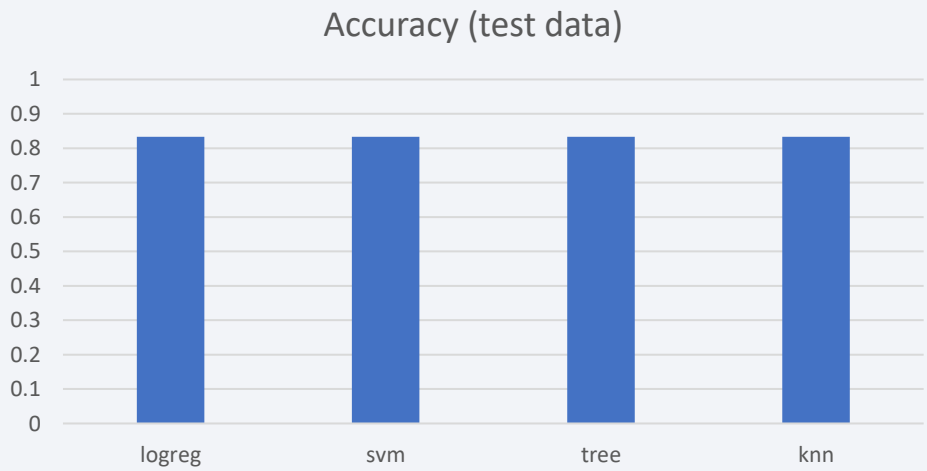
Section 5

Predictive Analysis (Classification)

Classification Accuracy

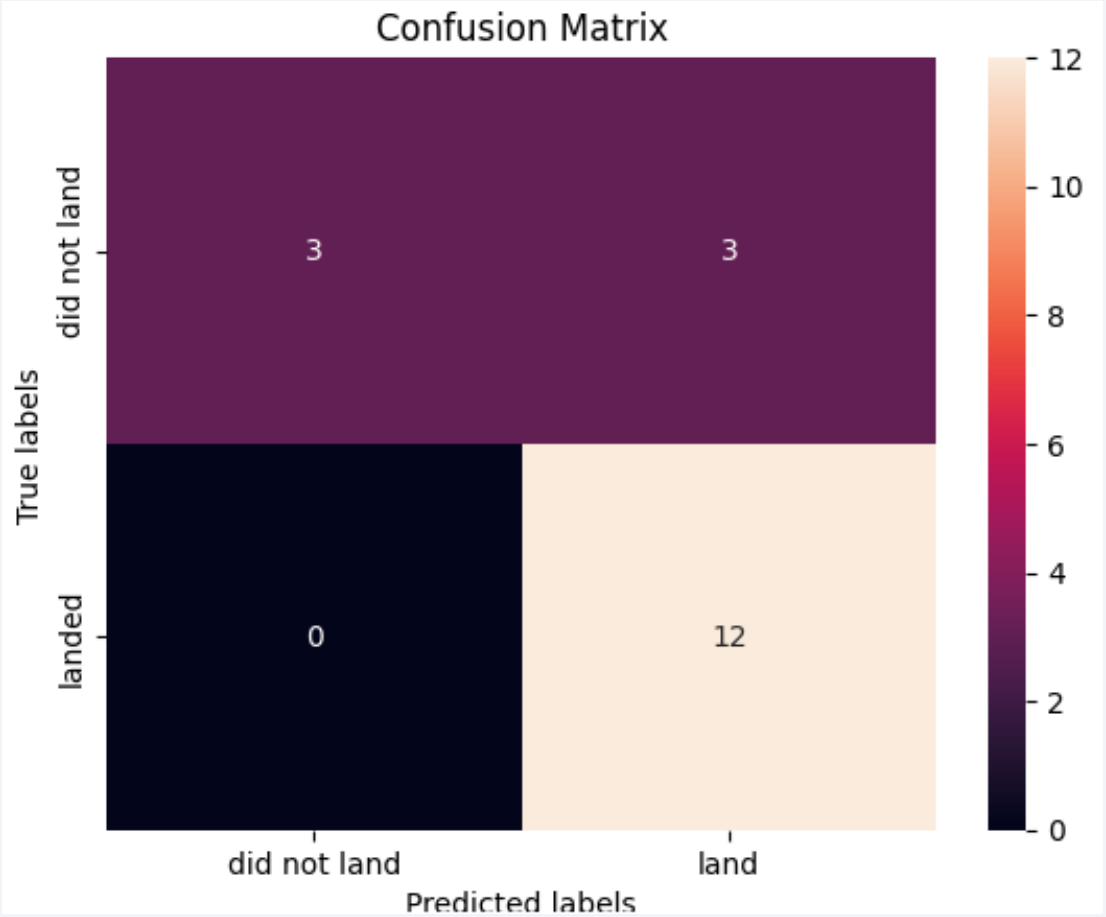
- All four models performed consistently well, with cross-validated accuracies ranging from 84.6% to 87.5%.
- The test accuracy for all models was identical at 83.3%, indicating they generalize to unseen data to the same degree.





Confusion Matrix

- The Decision Tree achieved a cross-validated accuracy of 87.5%, outperforming all other models. This indicates it performed consistently well across different subsets of the training data. The small gap between cross-validation and test accuracy (83%) shows the model is neither overfitting nor underfitting.



Conclusions

- Success rates varied across launch sites, with KSC LC-39A and VAFB SLC-4E showing higher success probabilities than CCAFS LC-40.
- Over time, launch success rates improved, especially after 2013, indicating operational learning and technological improvements.
- All of the launch sites are located in close proximity to the coast.
- Four models were trained on this data: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).
- The Decision Tree Classifier achieved the highest cross-validated accuracy (87.5%) while maintaining stable test performance. This suggests it learned the most informative patterns without overfitting.

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!