IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jonh Brian Lemos
2024-02-03

Outline

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

Executive Summary

- The methodologies applied in this project were as follows:
  - Collect SpaceX Falcon 9 data using an API and web scraping.
  - Perform exploratory data analysis (EDA) with data wrangling and visualization to extract crucial information and comprehend the data.
  - Conduct predictive analysis to explore algorithms for predicting rocket landing outcomes.
- The main results were as follows:
  - Analyzed rocket behavior and variables to predict landings with precision.

Introduction

- Launching rockets into space entails significant costs, reaching up to $165 million per mission. The potential for cost reduction lies in successfully landing and reusing the first stage. Predicting the likelihood of a rocket's successful landing is crucial for optimizing launch conditions, reducing costs, and enhancing a company's competitiveness.
- The objective of this project is to develop a predictive model to assess the probability of a successful rocket launch. The goal is to provide companies with actionable insights, enabling them to make informed decisions about the economic viability of each launch. Ultimately, this initiative strives to contribute to a more cost-effective and competitive space exploration industry.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The initial data source utilized was an API from SpaceX: https://api.spacexdata.com/v4/rockets/
  - In addition, web scraping was employed to gather information about launches from Wikipedia's: List of Falcon 9 and Falcon Heavy launches
- Perform data wrangling
  - Analyzed the launch sites and orbits.
  - Introduced a new column to capture information on the outcomes.

Methodology

Executive Summary

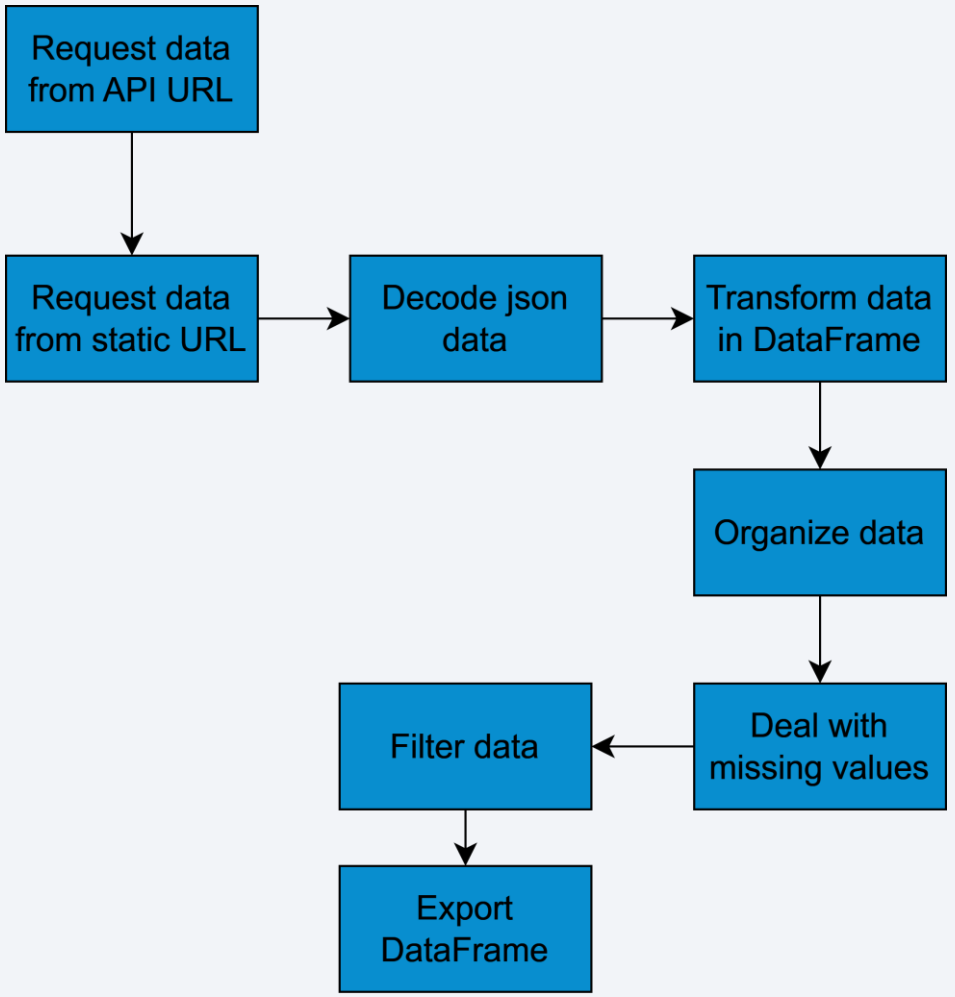
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The collected data underwent transformation using standard scaling.
  - Subsequently, the data was partitioned into training and testing sets for result validation.
  - Grid search methodology was applied to obtain the optimal parameters for the algorithms.
  - For each outcome, a confusion matrix was employed to scrutinize accuracy in each case.

Data Collection

- The initial step involved gathering information using the SpaceX API. This included collecting details about the launch site, rocket, cores, launchpad, outcomes, and more.
- However, it was necessary additional information about the flights was necessary, such as orbit, customer, payload, payload mass, as well as specific time and date details.

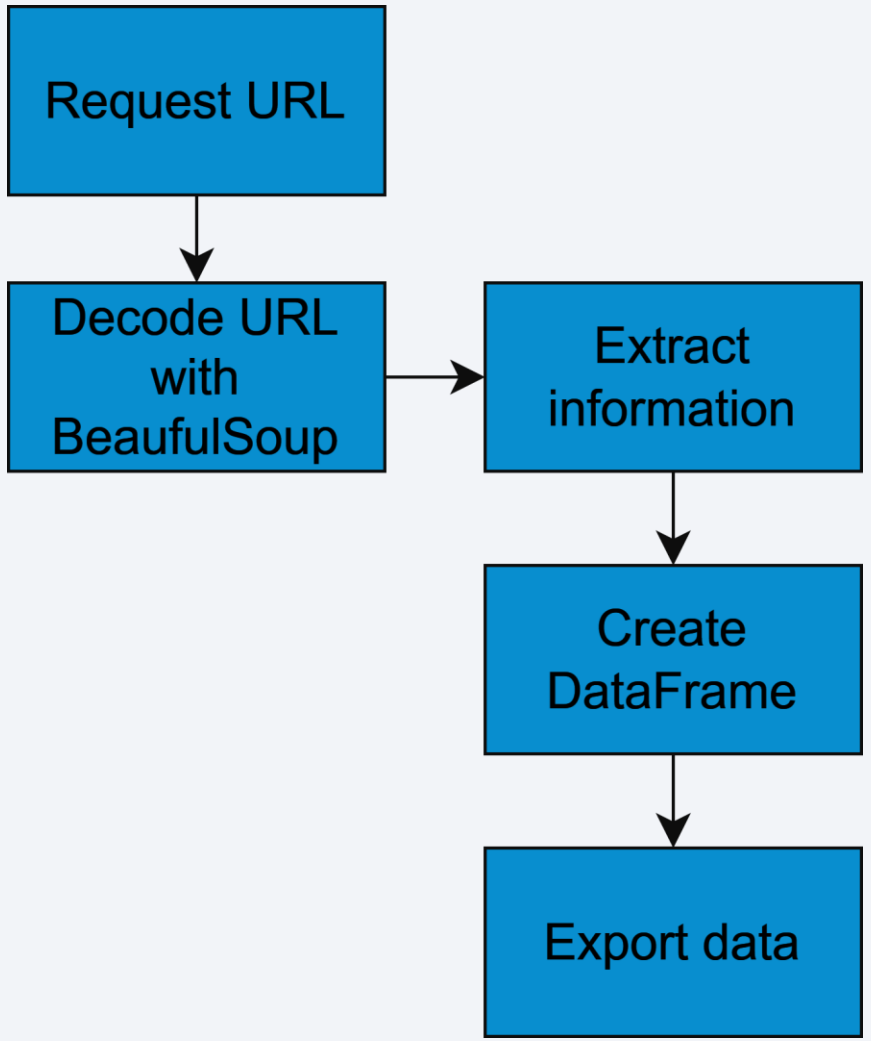
Data Collection – SpaceX API

- The extraction of data involved utilizing an open SpaceX API.
- Source code: https://github.com/jonhbl/ds-coursera-capstone-project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



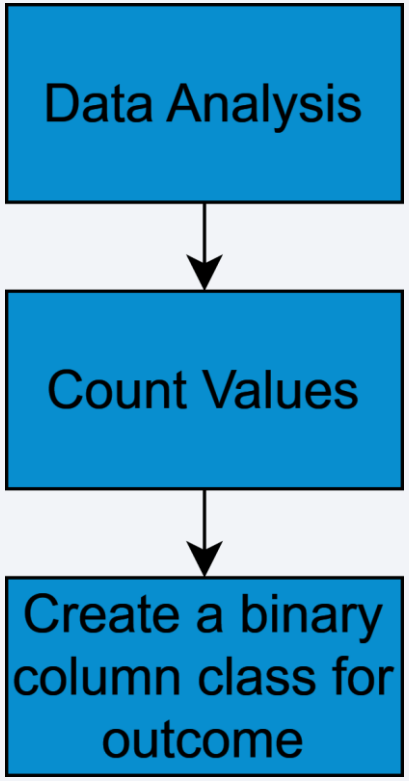
Data Collection - Scraping

- The extraction of data involved utilizing web scraping techniques on Wikipedia.
- Source code: https://github.com/jonhbl/ds-coursera-capstone-project/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

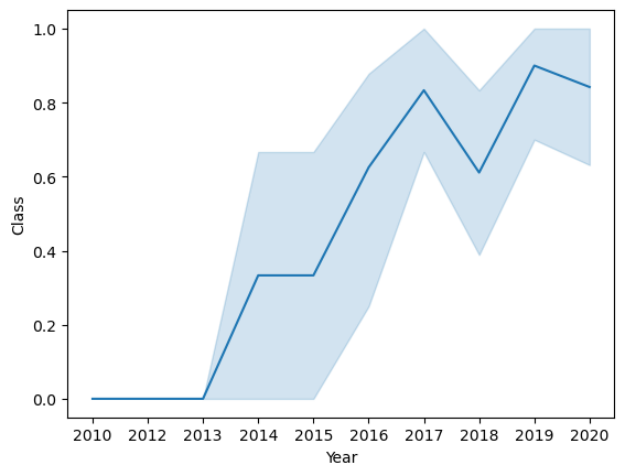
1. The dataset underwent analysis, encompassing the observation of null data quantities and data types.
2. Additionally, the value counts of specific columns, such as launch site, orbit, and outcome, were scrutinized.
3. Leveraging the outcome, a categorization into positive and negative results was achieved.
4. A Boolean column was subsequently generated to capture the outcome class of each flight.

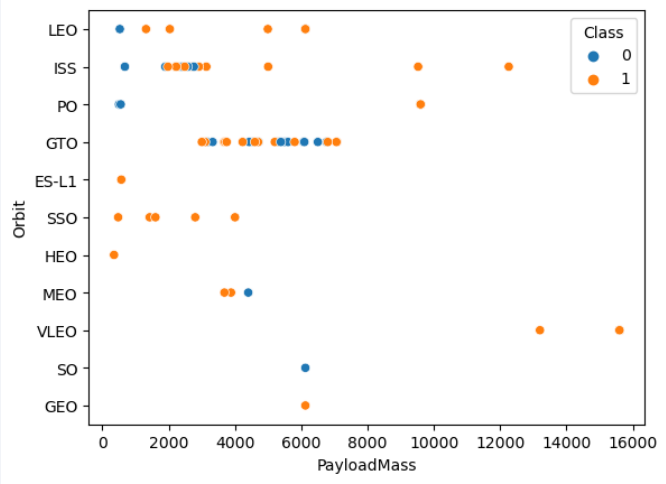


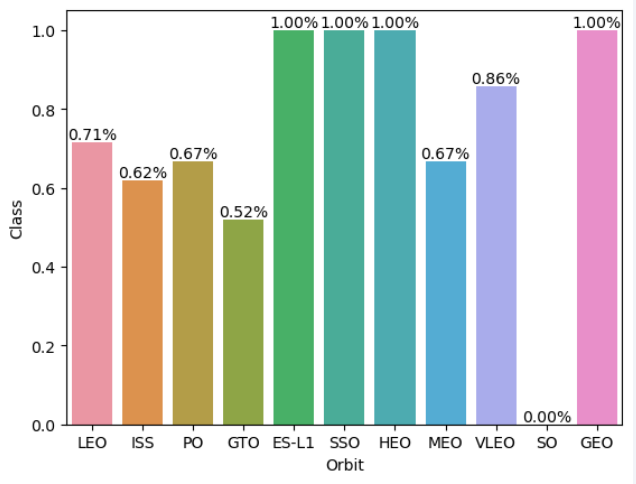
Source code: https://github.com/jonhbl/ds-coursera-capstone-project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Exploratory Data Analysis involved the utilization of scatterplots, line plots, pie charts, and bar plots to delve into the dataset.







- Source code: https://github.com/jonhbl/ds-coursera-capstone-project/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- The following queries were executed to extract information from the data:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Total payload mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Names of the booster versions which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
  - Rank 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.
  - Source code: https://github.com/jonhbl/ds-coursera-capstone-project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Tasks involving marking launch sites, identifying success/failed launches, and calculating distances employed the following objects:
  - Circle: Utilized to highlight specific areas on the map, such as NASA Johnson Space Center.
  - Marker: Grouped events, including launchers and launch sites.
  - Lines: Employed to indicate distances between two coordinates.

- Source code: https://github.com/jonhbl/ds-coursera-capstone-project/blob/main/lab_jupyter_launch_site_location.ipynb

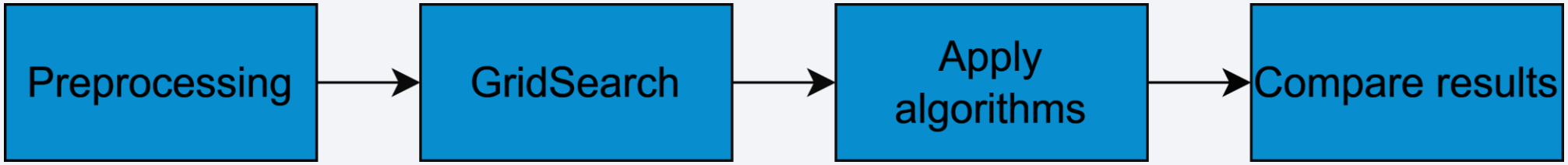
Build a Dashboard with Plotly Dash

- Two graphs were employed for data analysis:
  - A pie chart depicting the "Success Ratio" for all sites or each one individually.
  - A scatter plot illustrating "Payload Mass vs Outcome."
- These visualizations facilitated the examination of outcomes in relation to payload and success. They allowed for the selection of a specific launch site, enabling individual analysis of how payload influences the outcome.

- Source code: https://github.com/jonhbl/ds-coursera-capstone-project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Initially, normalizing the data was crucial to ensure that all attributes carried equal importance in the model. Subsequently, four machine learning models were deployed for predicting the outcome of a rocket landing. To ensure optimal results, a comprehensive range of parameter combinations were tested using grid search.



- Source code: https://github.com/jonhbl/ds-coursera-capstone-project/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

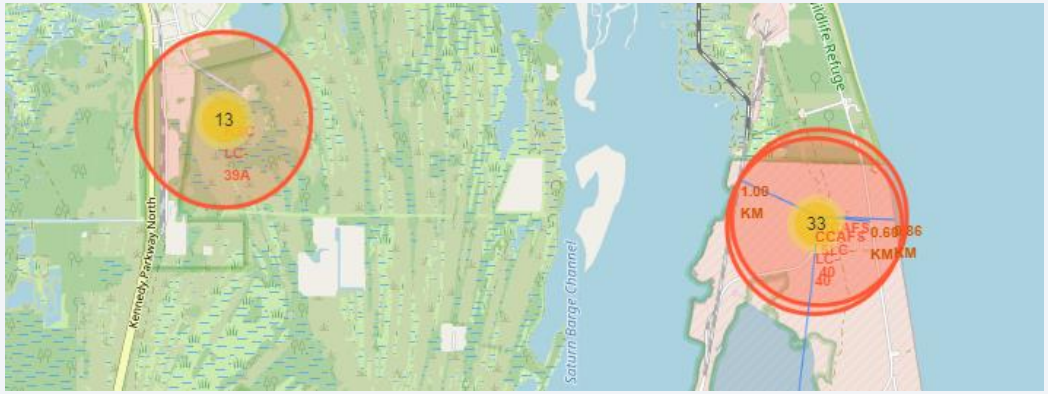
Results

- Exploratory data analysis results:
- SpaceX operates from 4 different launch sites.
- The initial launches were conducted for both SpaceX itself and NASA.
- The average payload of the F9 v1.1 booster stands at 2534 kg.
- The first successful landing occurred in 2013, three years after the inaugural launch.
- As the years progressed, numerous Falcon 9 booster versions demonstrated successful landings on drone ships, particularly those carrying payloads above the average.
- Remarkably, nearly 100% of mission outcomes were successful. However, in 2015, two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, experienced failures in landing on drone ships.
- An interesting trend emerged over time, showing an improvement in the number of successful landing outcomes as the years passed.

Results


Vandenberg
Space Force
Base
Vandenberg
Village
211
211
Lompoc
Airport
Vandenberg
State Marine
Reserve
VAFB
SLC
4E
Lompoc

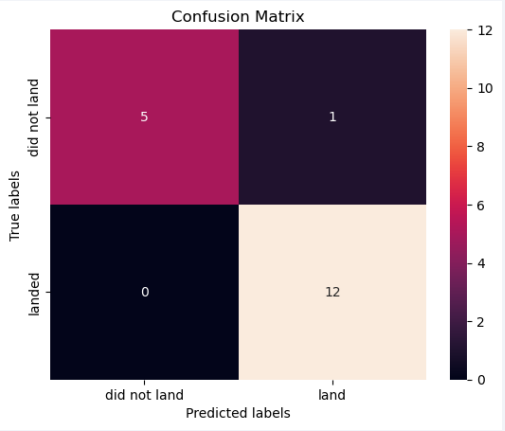

7
AFS
SLC-
40

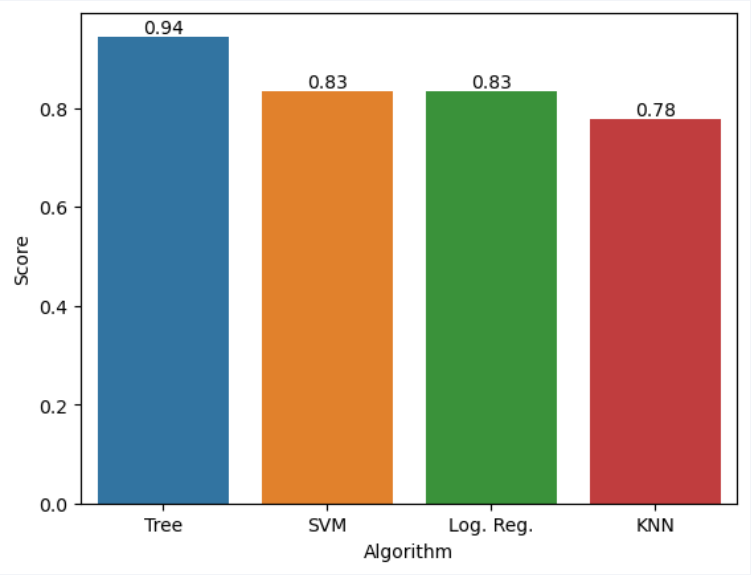

13
LC-
39A
Kennedy Parkway North
Saturn Barge Channel
1.06
KM
33
CCAFS
LC-
40
KM
KM

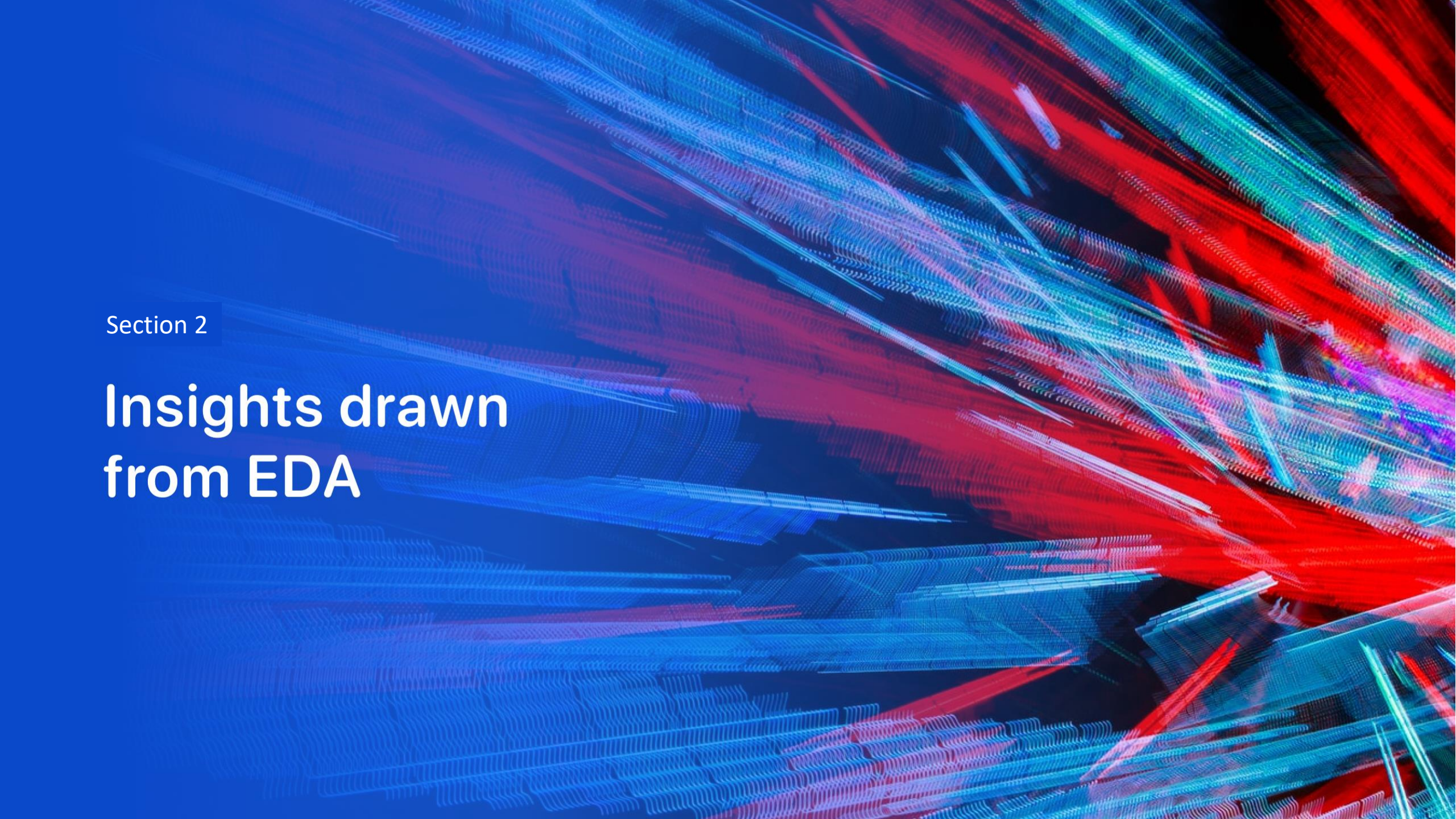
Results

- The Decision Tree Classifier exhibited the most promising results, achieving a 94% accuracy in the test data and an 86% accuracy during the training phase of the grid search.

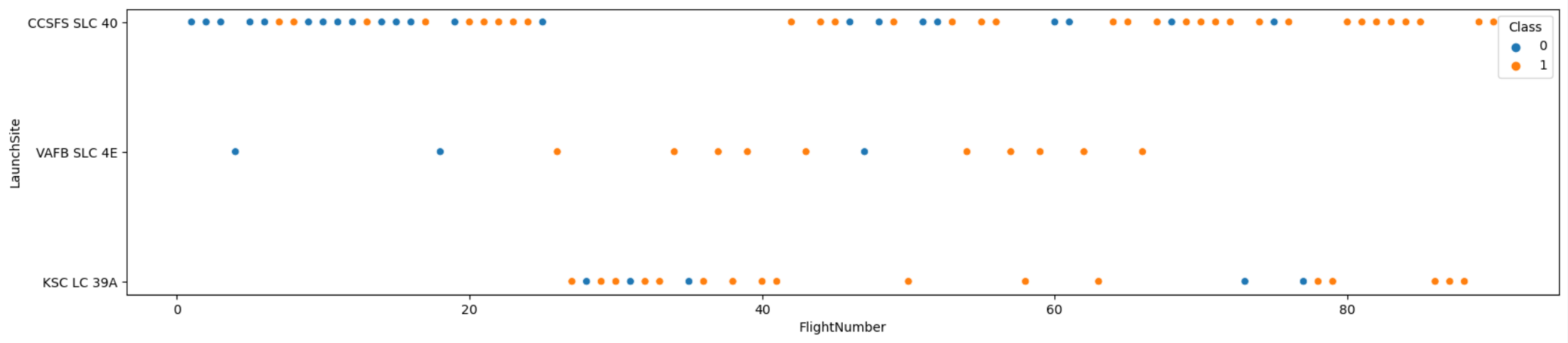
	Algorithm	Best Score	Score
2	Tree	0.862500	0.944444
1	SVM	0.848214	0.833333
3	KNN	0.833929	0.777778
0	Log. Reg.	0.821429	0.833333





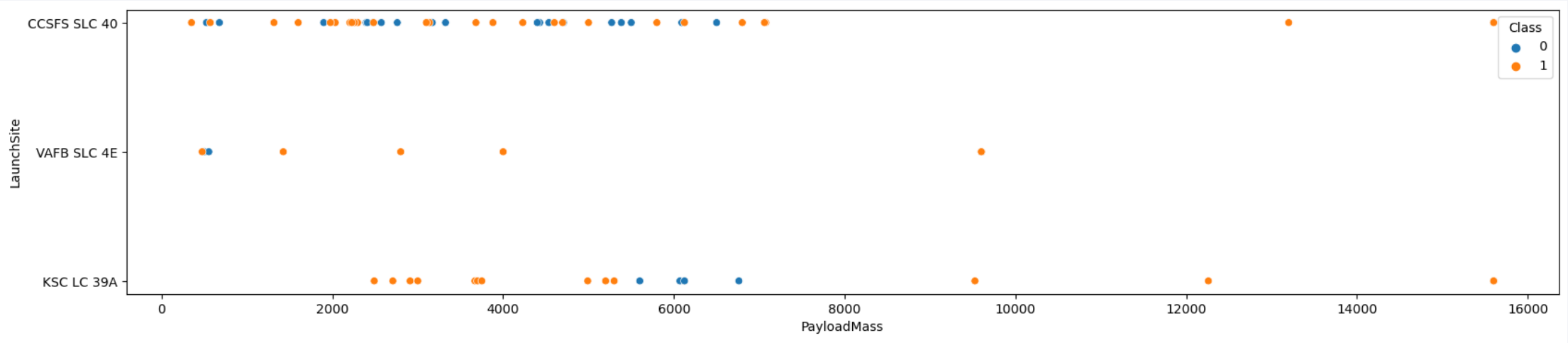

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- An inference can be drawn indicating that the latest flights were more successful than the previous ones, suggesting an improvement over time.
- Additionally, CSSDS SLC 40 emerged as the most frequently utilized launch site.

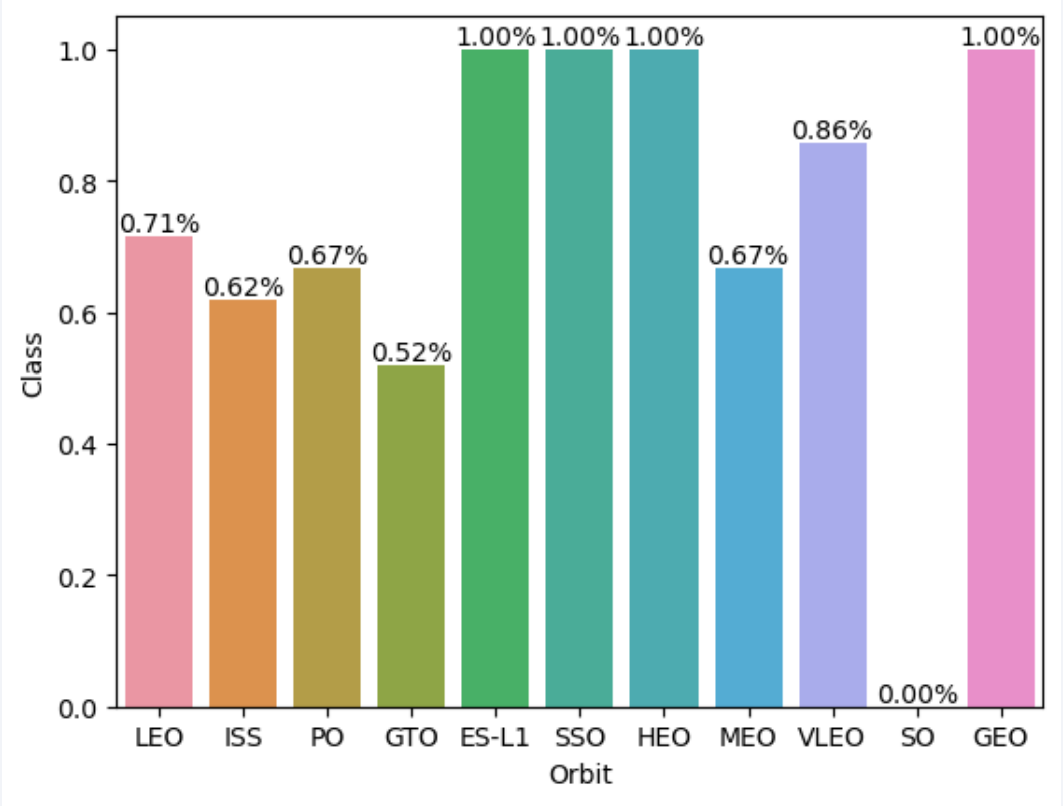
Payload vs. Launch Site



- CCSFS SLC 40 generally handles lighter weights, although there may have been some tests involving heavier payloads.
- The majority of flights don't carry weights exceeding 6000 kg.
- Interestingly, all flights with payloads over 8000 kg were successful.

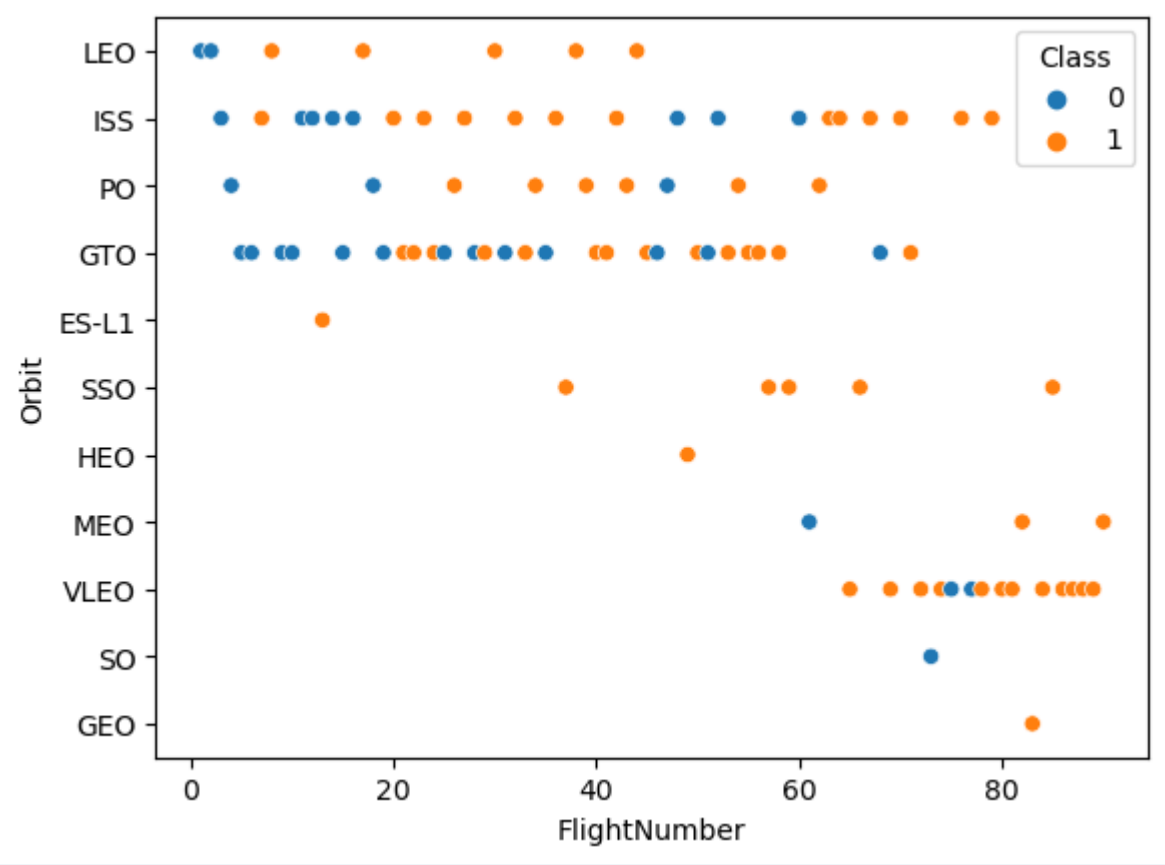
Success Rate vs. Orbit Type

- Certain orbits, such as ES-L1, SSO, HEO, and GEO, boast a 100% success rate in landings.
- Notably, there are no successful landings reported from flights originating from SO.



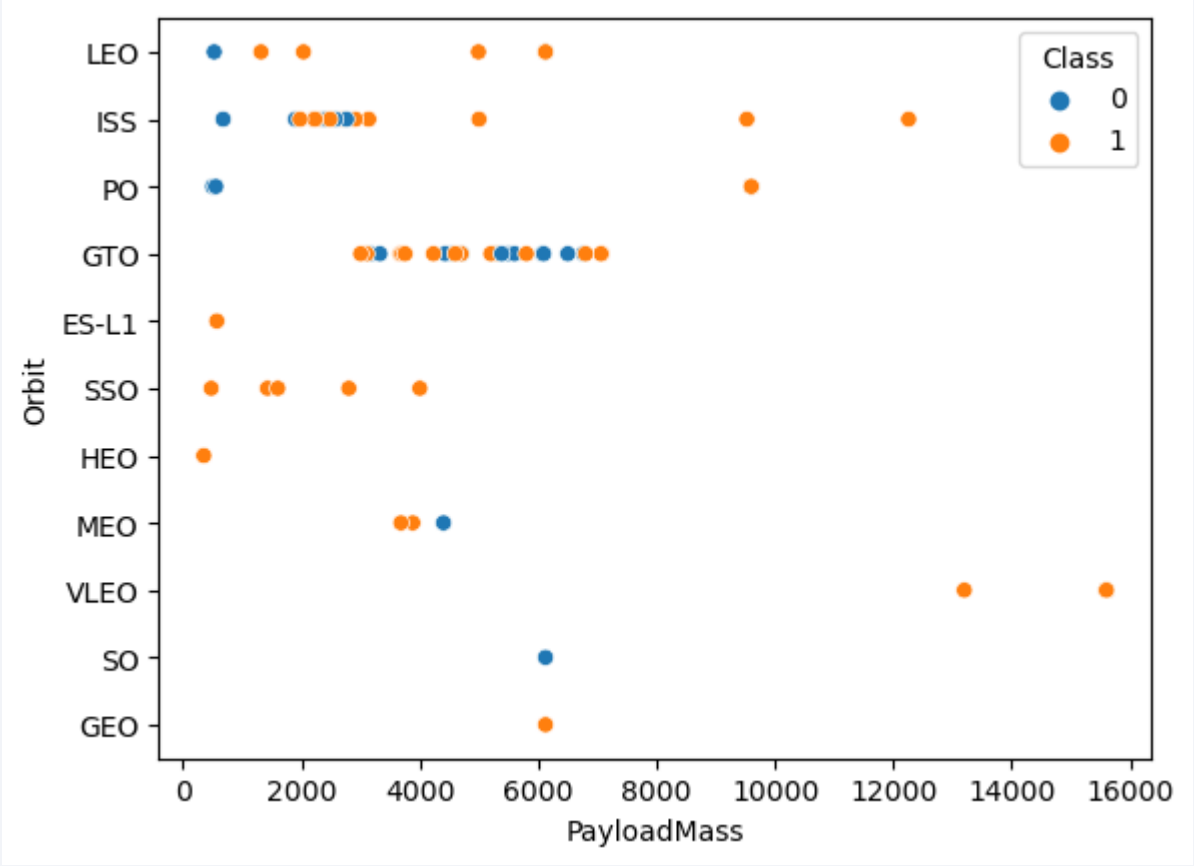
Flight Number vs. Orbit Type

- As mentioned earlier, recent flights demonstrate a higher success rate.
- Opting for Very Low Earth Orbit (VLEO) has proven to be a favorable choice, with a substantial number of flights and an impressive success rate of 86%.



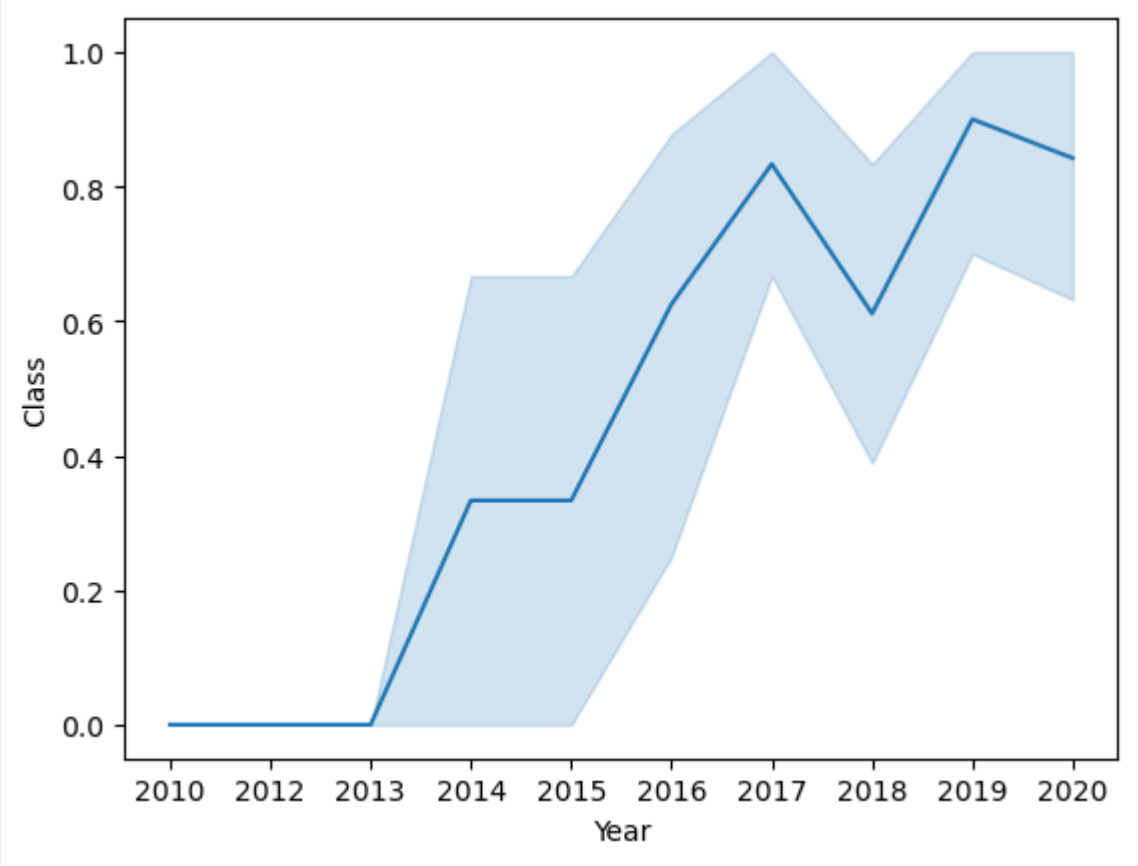
Payload vs. Orbit Type

- VLEO is typically chosen when dealing with heavier weights, while SSO maintains a 100% success rate with lighter payloads.



Launch Success Yearly Trend

- There were no successful flights until 2013.
- Success rates increased steadily until 2017, and post-2018, the trend shifted towards consistently successful landings.



All Launch Site Names

- The launch sites:

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

- The distinct operator was employed to filter and identify the unique launch sites in the dataset.

Launch Site Names Begin with 'CCA'

- Five records that the lauch site begin with ‘CCA’

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD MASS_KG	Orbit	Customer	Mission_Outcome	Landing_Outcome	Date
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)	2010-06-04
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)	2010-12-08
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	2012-05-22
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt	2012-10-08
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt	2013-03-01

- To retrieve five records where the launch site begins with 'CCA', the query likely involved using the "LIKE" operator

Total Payload Mass

- Total payload carried by boosters from NASA

Total Payload (Kg)
99980

- The payload represents the sum of all payloads carried by all customers, not solely NASA (CRS).

Average Payload Mass by F9 v1.1

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

Average Payload (Kg)
2534

- For this query, the "avg" operator was applied, specifically filtering only for the booster version F9 v1.1.

First Successful Ground Landing Date

- First successful landing outcome on ground pad

Date
2015-12-22

- We observed that the first successful landing took place in 2013, but the initial ground pad landing occurred two years later.

Successful Drone Ship Landing with Payload between 4000 and 6000

- These were the booster that drop shipped a payload greater than 4000, but less 6000 kg.

Booster version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- It is necessary to select outcomes as successful, exclude "drop ship," and incorporate mass boundaries for the analysis.

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

Booster version	Count
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- For this result, the "GROUP BY" operator was utilized, and subsequently, a "COUNT" operation was also applied.

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass:

- In this scenario, a subquery is required to identify the maximum payload, and then the query should select the booster version that carried this mass.

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

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

- Both failures originated from CCAFS LC-40 with version F9 v1.1.
- To select the month in this query, it was necessary to utilize the "substr" function to extract the relevant information.

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

- Ranking of the count of landing between the date 2010-06-04 and 2017-03-20:

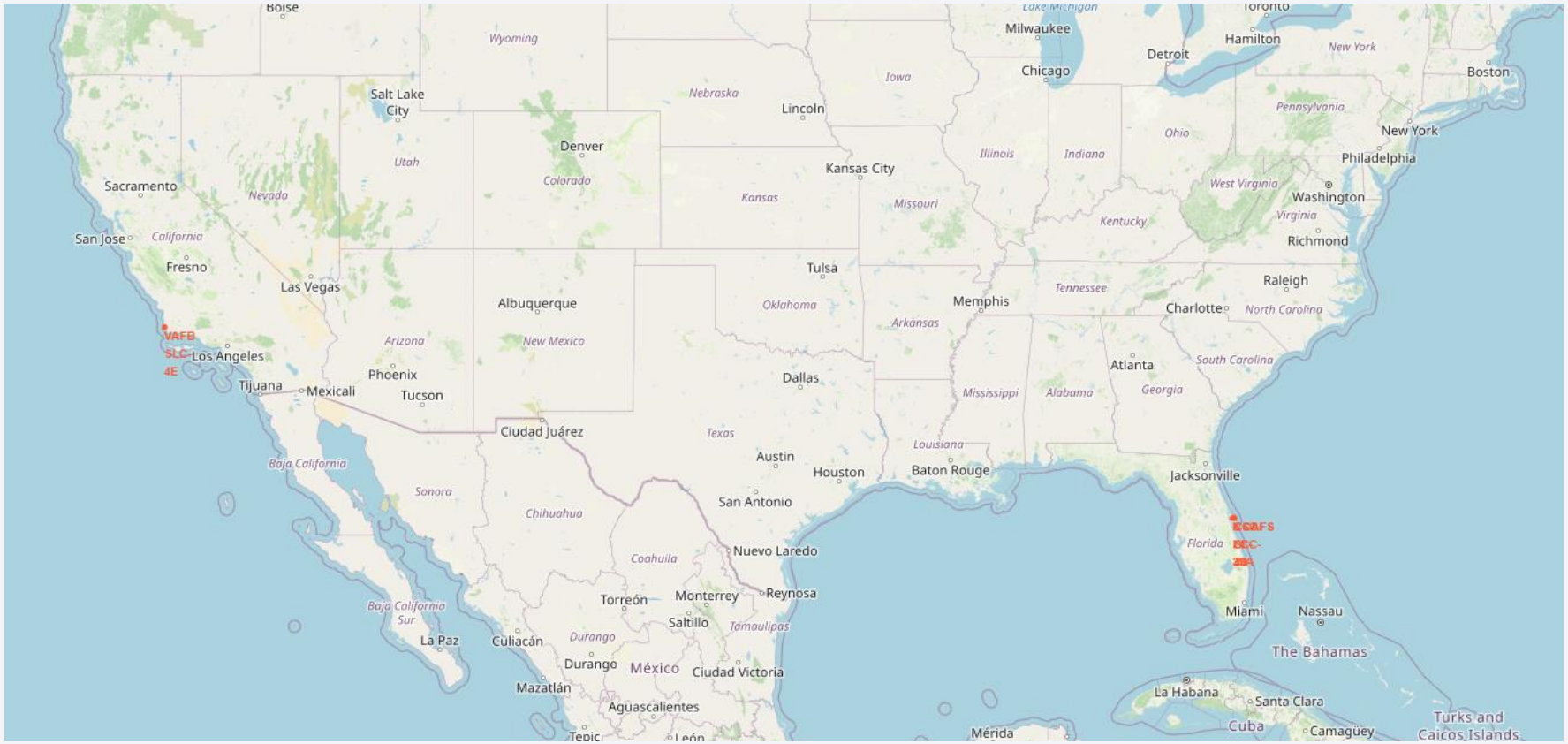
Landing_Outcome	counter
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

- During this period, the most prevalent outcome was "No Attempt."

Section 3

Launch Sites Proximities Analysis

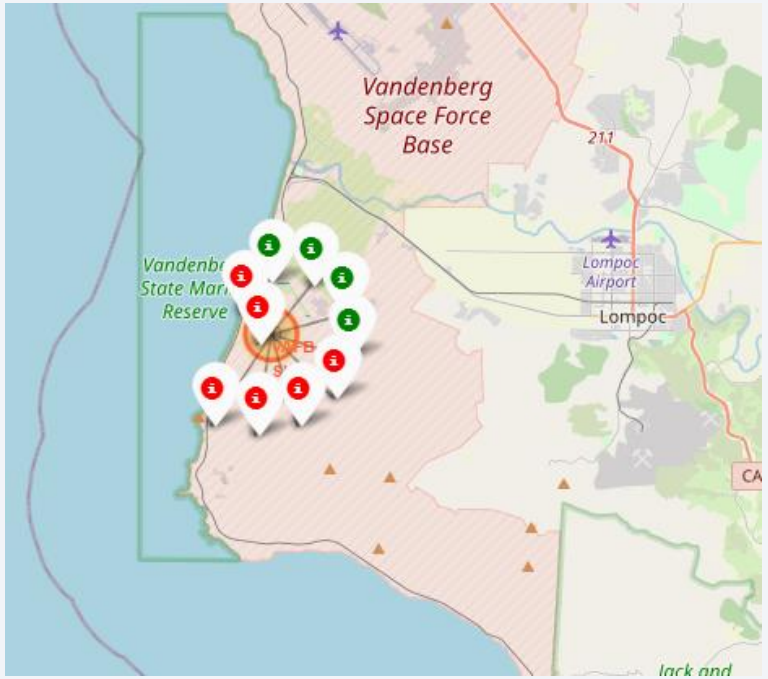
Launch sites on US map



- All launch sites are situated very close to the coast.
- However, the launch sites are not in close proximity to the Equator line.

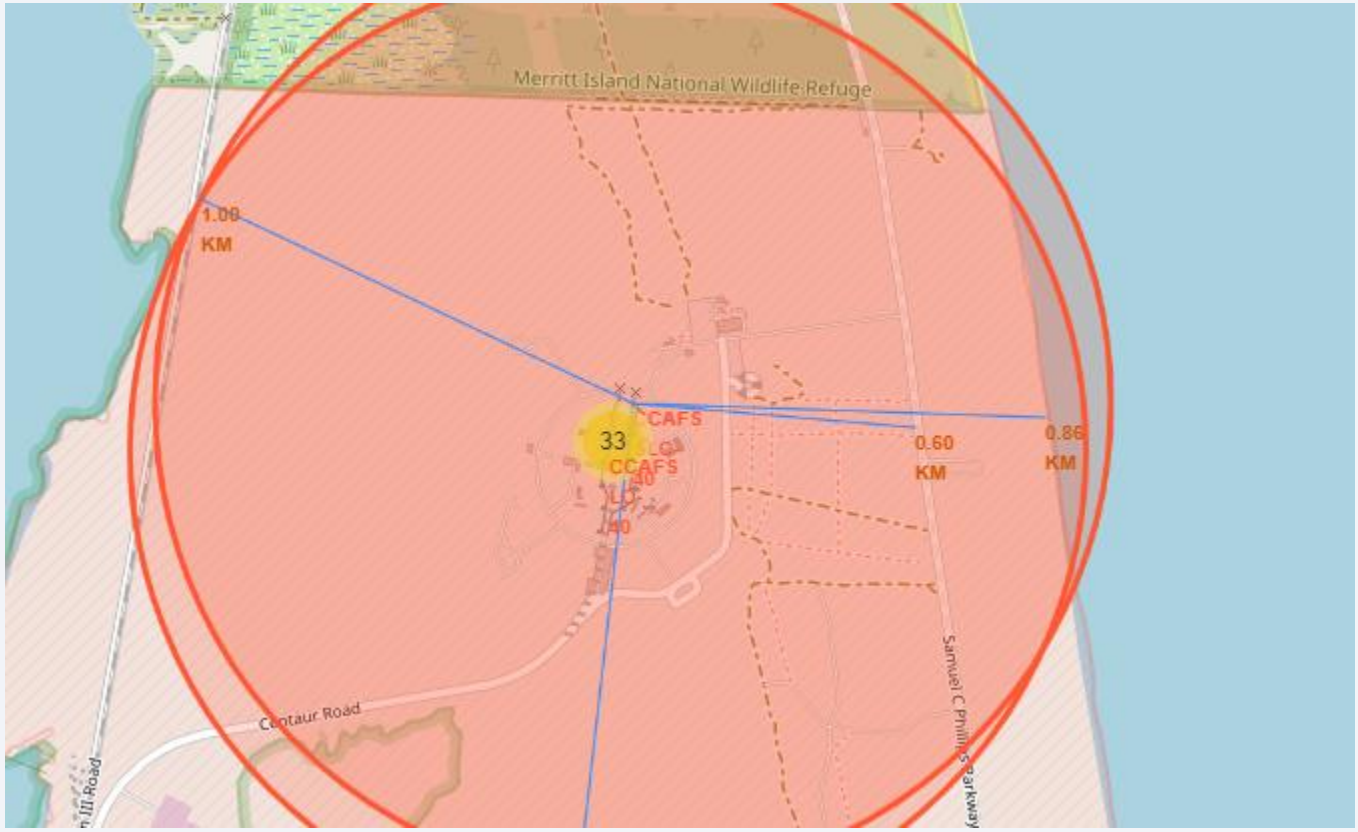
Launch sites and outcomes

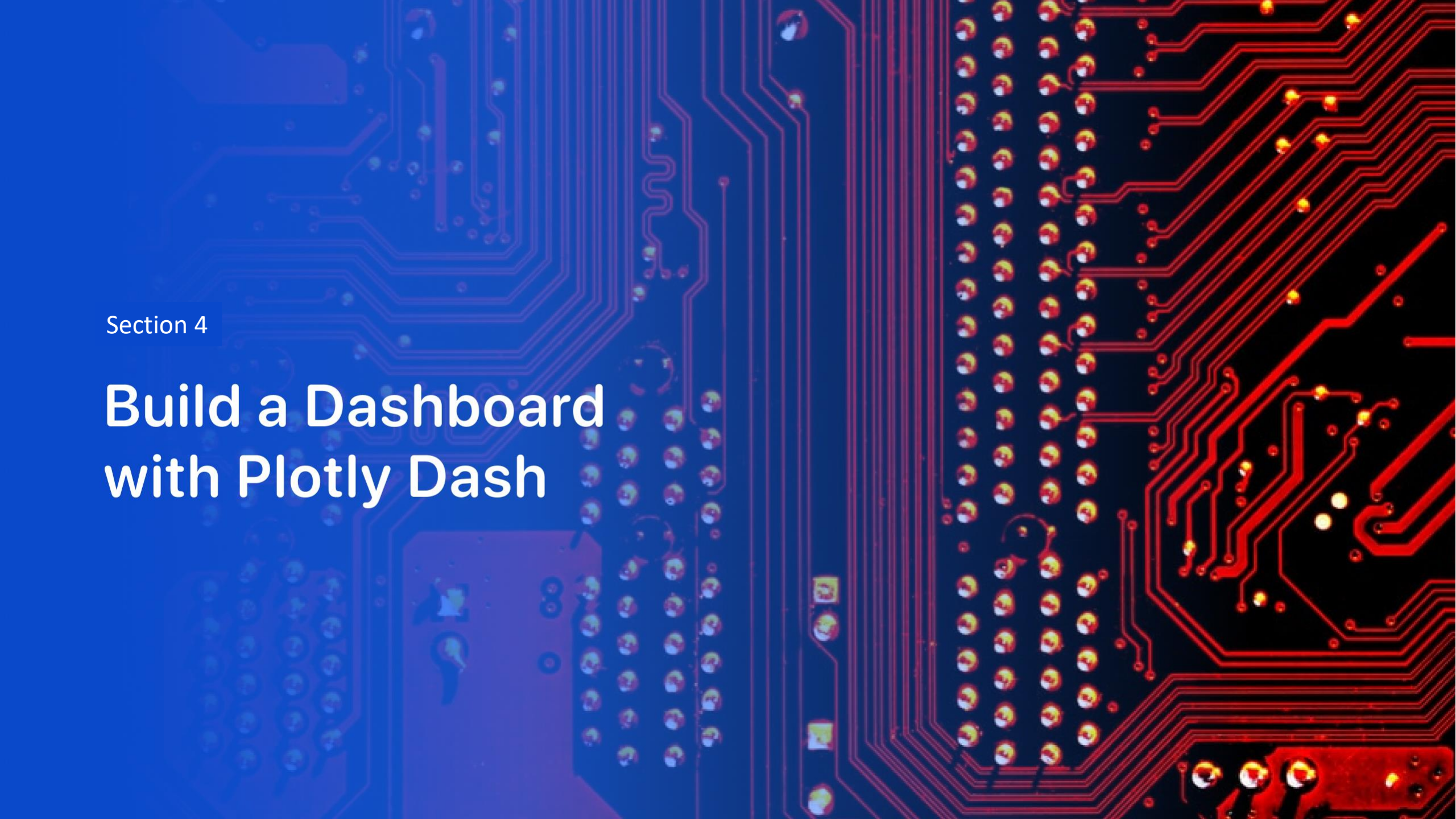
- VAFB SLC 4E in California has experienced more failures than successes.
- Despite being very close to each other, CCAFS LC-40 has more flights than CCAFS SLC-40.



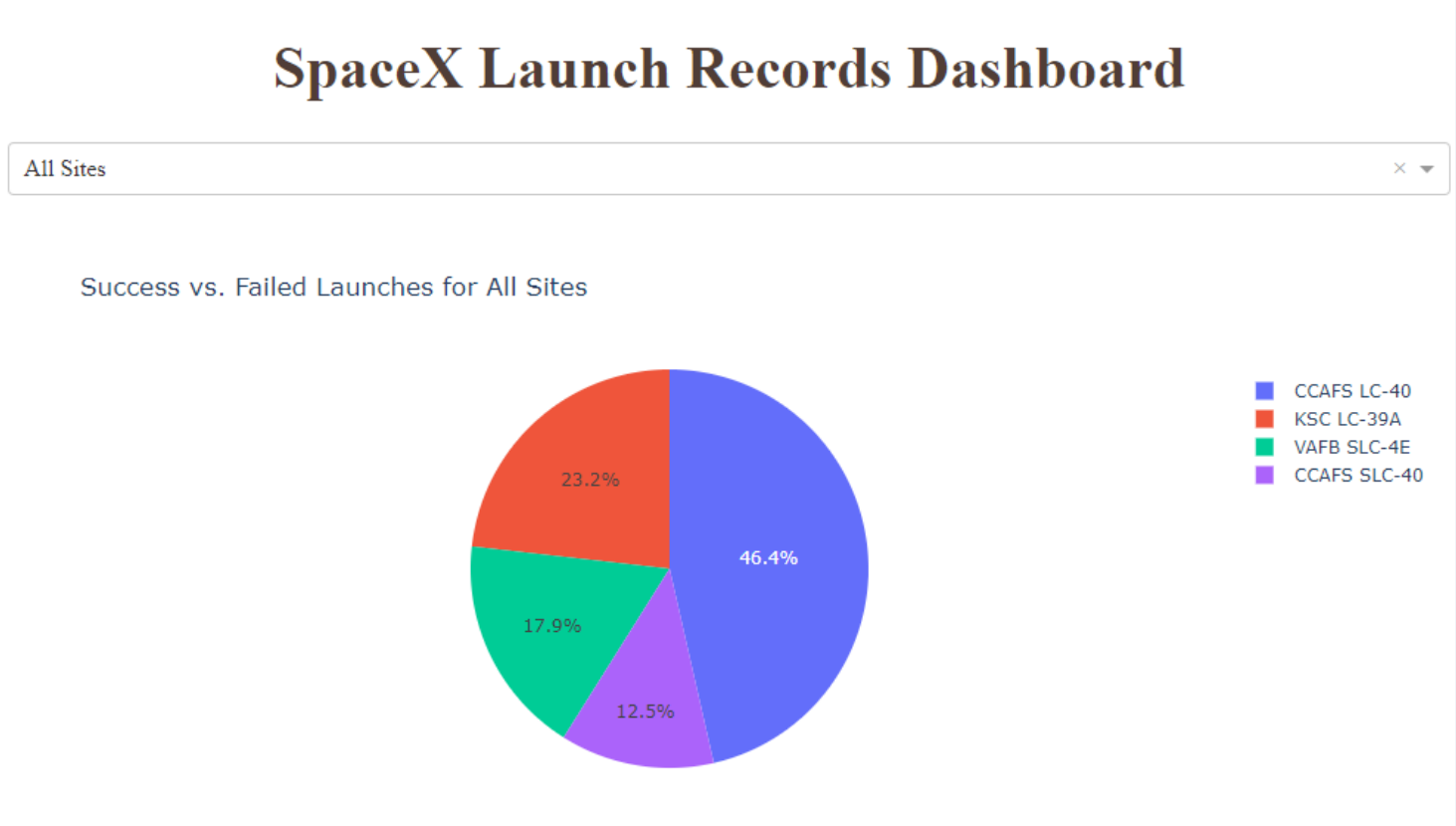
CCAFS SLC-40/LC-40 Proximites

- These launch sites are in close proximity to railways, highways, and the coastline. However, they are distant from the nearest city, Melbourne, which is approximately 51.21 km away.



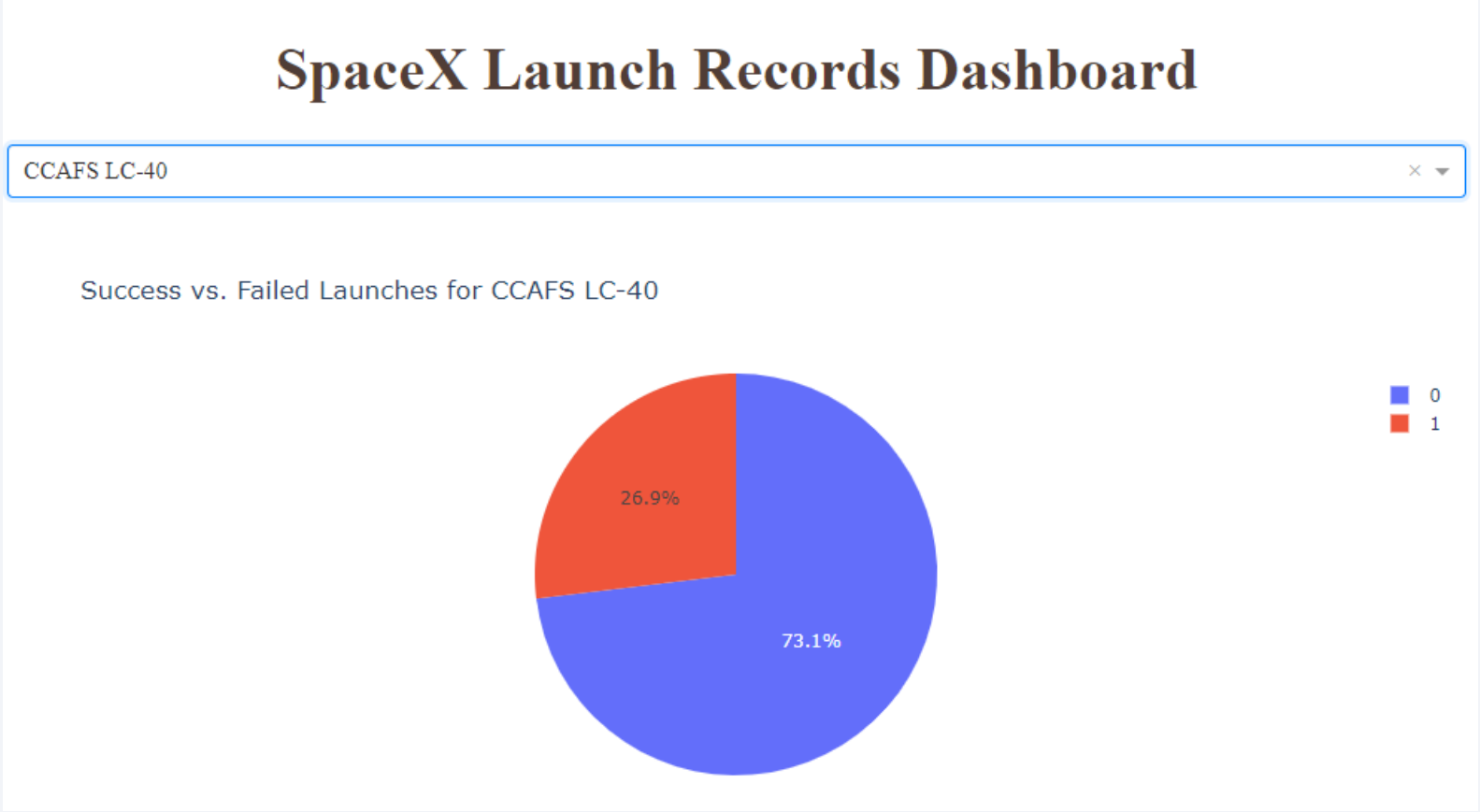

Section 4
Build a Dashboard with Plotly Dash

Launch success for all sites



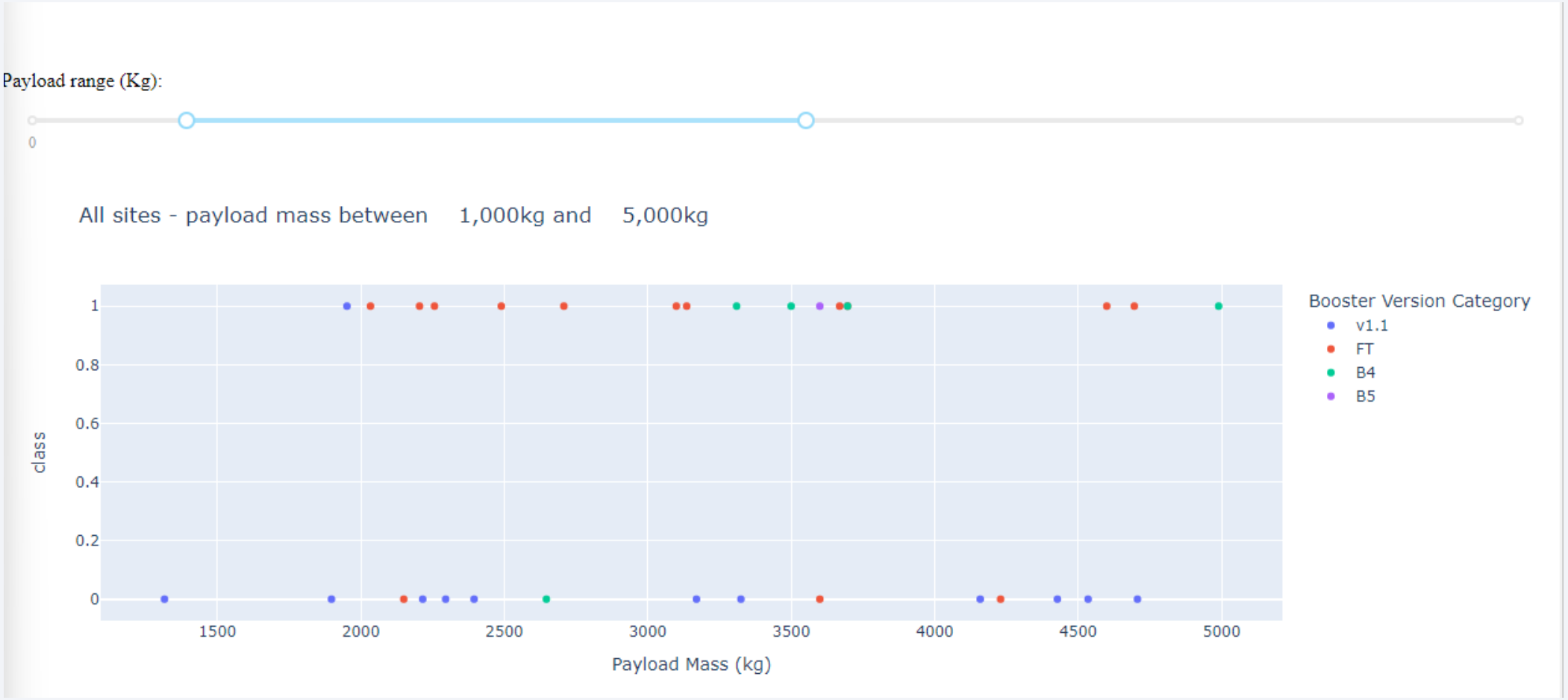
- It is noticeable that CCAFS LC-40 has a high success rate, indicating that the launch site may be an important contributing factor to the success of launches.

Launch success for CCAFS LC-40



- Approximately 75% of the launches are successful.

Payload Mass (kg) vs Outcome

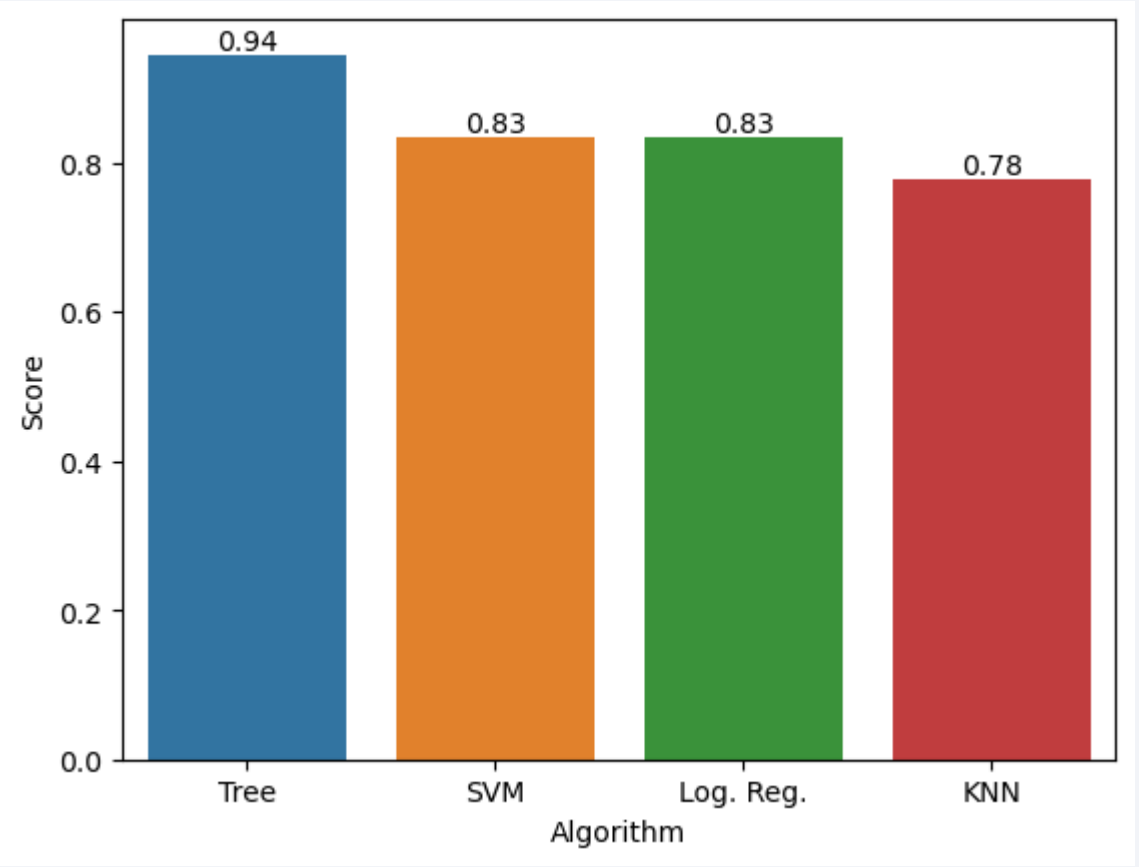


Section 5

Predictive Analysis (Classification)

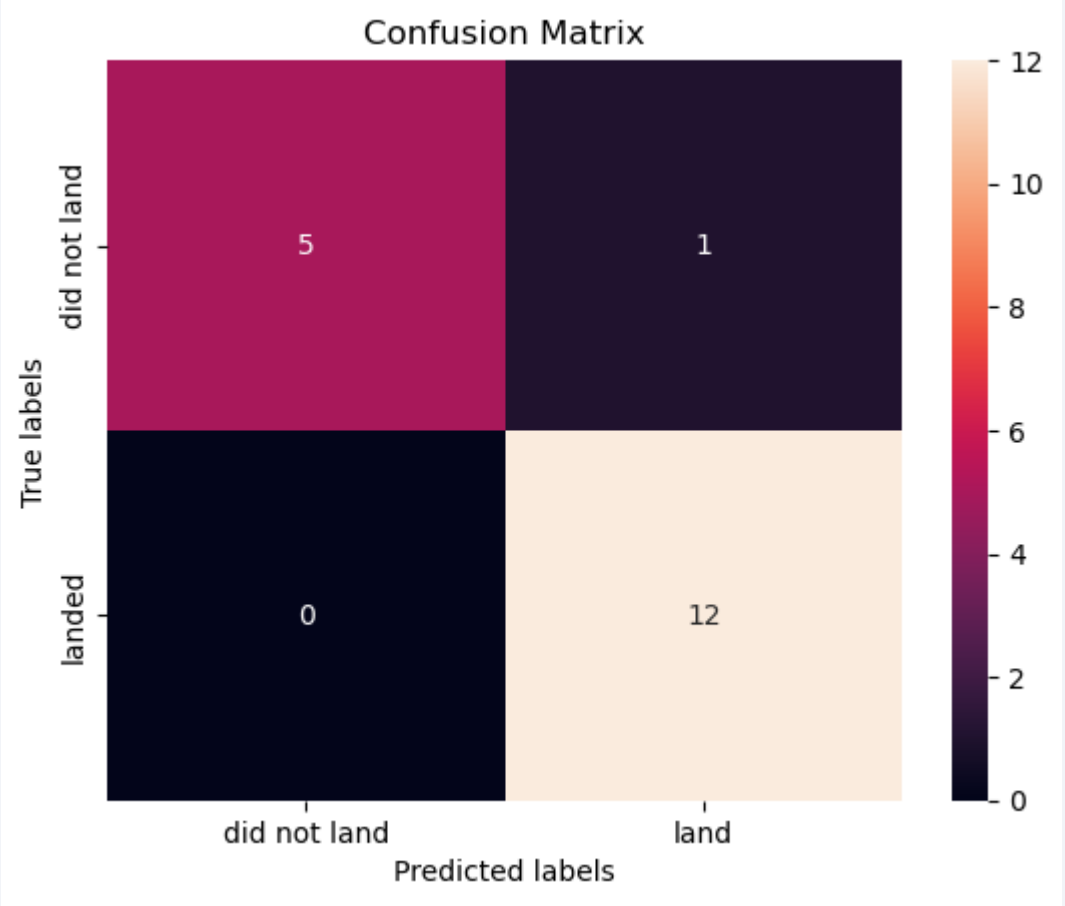
Classification Accuracy

- The most effective model was the Decision Tree Classifier, achieving an accuracy of 94% in the test data.



Confusion Matrix

- Decision Tree Classifier Confusion Matrix:
- It indicates that the algorithm made only one incorrect prediction, misclassifying a case as "land" when it was supposed to predict "did not land."



Conclusions

- Aggregating data from multiple sources is crucial to create a more robust model. Exploratory Data Analysis (EDA) is extremely necessary to extract important insights and generate useful information.
- Dashboards prove to be very useful for dynamically analyzing data, exploring multiple options rapidly.
- The ability to predict with high accuracy whether a rocket can successfully land or not is noteworthy.
- Technology is continually improving, and there is a noticeable trend towards more common successful outcomes over time.

Appendix

- Data Collection:
  - Python: SpaceX API extraction.
  - Web Scraping: Falcon 9 and Falcon Heavy launches.
- Data Analysis:
  - Jupyter: EDA outputs.
  - SQL: Wrangling, analysis, visualization.
  - Python: Scatter, line, pie, bar plots.
- Predictive Modeling:
  - Python: Normalization, splitting.
  - ML Models: Decision Tree, etc.
  - Grid Search: Parameter tuning.
  - Confusion Matrix: Evaluation.
- Visualizations:
  - Python: "Success Ratio" Pie Chart.
  - Scatter Plot: "Payload Mass vs Outcome."
- Queries:
  - SQL: Filtering, aggregating.
  - Subqueries: Max payload.
- Observations:
  - Key findings, insights.
  - Charts: Success trends.
  - Python: Launch site success rates.
- Technology:
  - List of tools, libraries.
- Data Sets:
  - Source links.
  - Processed data sets.

Thank you!