IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Usman Zafar Mirza
08-06-2023

Outline

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

Executive Summary

In order to analyze the data, several methodologies were employed, which included:

- Collecting data through web scraping and the use of the SpaceX API.
- Conducting Exploratory Data Analysis (EDA) which encompassed tasks such as data wrangling, data visualization, and interactive visual analytics.
- Utilizing Machine Learning Prediction techniques.

Overall, the following results were obtained:

- Valuable data was successfully collected from various public sources.
- EDA enabled the identification of key features that can be utilized to predict the success of launchings.
- Machine Learning Prediction helped determine the optimal model for predicting which characteristics are essential for driving this opportunity in the most effective manner, by leveraging all the data that was gathered.

Introduction

- The main goal is to assess the feasibility of Space Y, a new company, in terms of its ability to compete with Space X.

Desired Outcomes:

- Identifying the optimal approach to estimating the overall cost of launches, via the prediction of successful landings of the initial stage of rockets.
- Determining the most suitable location for conducting launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Two distinct sources were utilized to procure the data from Space X.
    - Space X API (https://api.spacexdata.com/v4/rockets/)
    - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - To enhance the collected data, we developed a landing outcome label using the outcome data. This label was created through a process of analyzing and summarizing the features of the data.
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - At this stage, the collected data underwent normalization and was divided into two sets: training and test data. To evaluate the data, four distinct classification models were employed, with the accuracy of each model being assessed using various combinations of parameters.

Data Collection

- The data sets used in this study were sourced from two distinct locations. The first was the Space X API (https://api.spacexdata.com/v4/rockets/), while the second was Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches). Web scraping techniques were employed to obtain the data from both sources.

Data Collection – SpaceX API

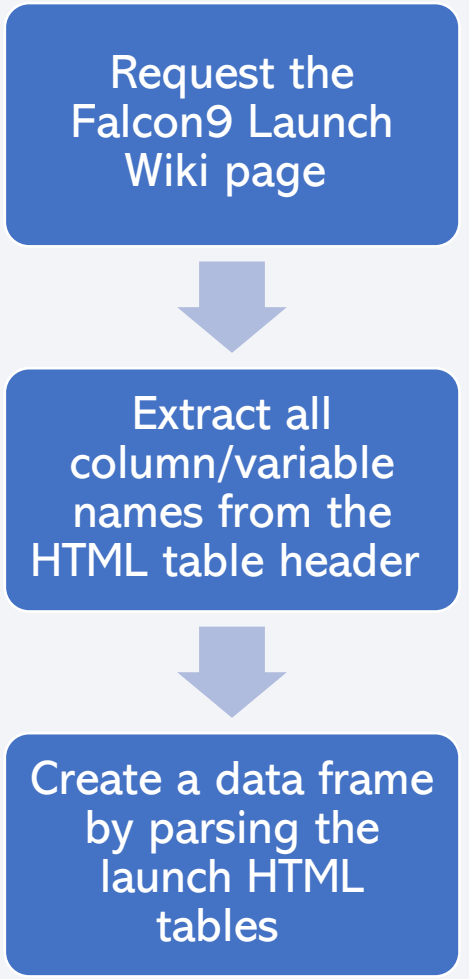
- SpaceX provides an accessible and publicly available API, through which data can be retrieved and utilized.
- Following the provided flowchart, the API was employed to obtain data, which was then stored for future use.
- Source Code: https://github.com/usmanzafarmirza/applied-data-science-capstone/blob/72092e8b75bb1faf7afde121cb4d433bf5851be1/lab-1-collecting-the-data.ipynb

Data Collection - Scraping

- Information about SpaceX launches is also retrievable from Wikipedia.

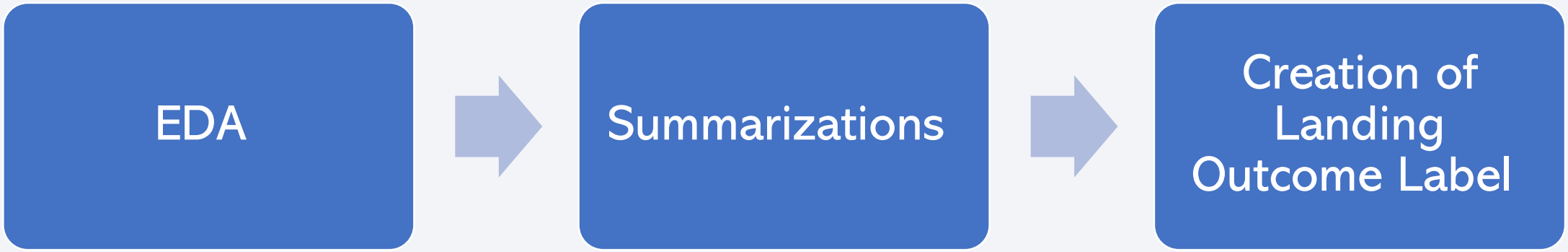
- Data is acquired from Wikipedia based on the instructions given in the flowchart, and subsequently stored for later use.

Source Code: https://github.com/usmanzafarmirza/applied-data-science-capstone/blob/master/Data%20collection%20with%20web%20scraping.ipynb



Data Wrangling

- Firstly, the dataset underwent Exploratory Data Analysis (EDA).
- Following this, calculations were made to determine the number of launches per site, frequency of each orbit, and the frequency of mission outcomes per orbit type.
- Lastly, the landing outcome label was created by extracting information from the Outcome column.

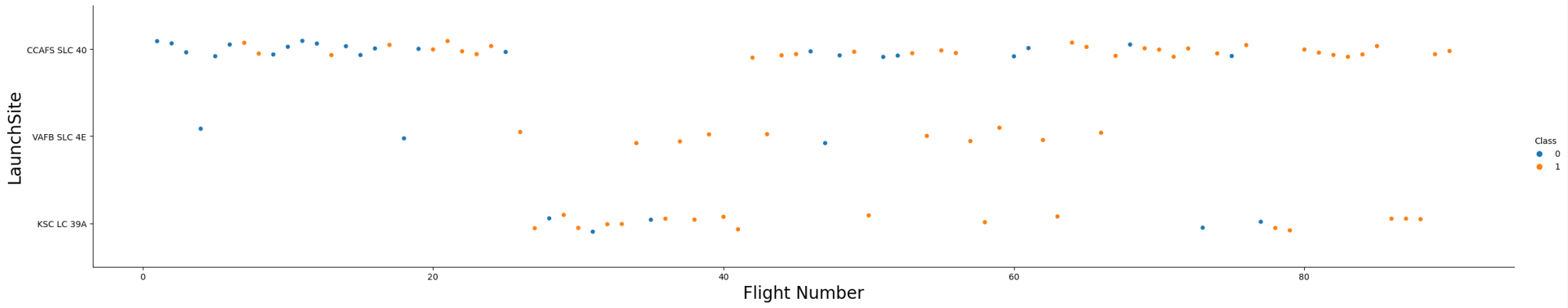


Source Code: https://github.com/usmanzafarmirza/applied-data-science-capstone/blob/72092e8b75bb1faf7afde121cb4d433bf5851be1/lab-2-data-wrangling.ipynb

EDA with Data Visualization

Scatterplots and barplots were utilized to examine the relationship between pairs of features and to investigate the data:

- Flight Number X Payload Mass, Flight Number X Launch Site, Payload Mass X Launch Site, Flight Number and Orbit, Orbit and Payload.



Source Code: https://github.com/usmanzafarmirza/applied-data-science-capstone/blob/baa80d46e72582790627cbfcae575b04e939d200/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

The ensuing SQL queries were executed:

- Identifying distinct launch locations utilized in the space expedition;
- The top 5 launch sites that start with the string 'CCA';
- The overall payload mass transported by NASA (CRS) launch vehicles;
- The mean payload mass conveyed by the F9 v1.1 rocket model;
- The date on which the first successful touchdown on a ground pad was accomplished;
- The names of launch vehicles that have achieved successful landings on drone ships and have a payload mass in the range of 4000 to 6000 kg;
- The cumulative count of successful and unsuccessful mission outcomes;
- Identifying the rocket models that have transported the heaviest payload mass;
- Instances of unsuccessful landings on drone ships in 2015, along with the respective rocket models and launch sites; and
- Tallying the frequency of different landing outcomes (for instance, Failure (drone ship) or Success (ground pad)) within the timeframe from 2010-06-04 to 2017-03-20.

Source Code: https://github.com/usmanzafarmirza/applied-data-science-capstone/blob/master/SQL%20Notebook%20for%20Peer%20Assignment.ipynb

Build an Interactive Map with Folium

Folium Maps employed markers, circles, lines, and marker clusters in the following ways:

- Markers serve to pinpoint locations such as launch sites;
- Circles are utilized to highlight zones around particular coordinates, like the NASA Johnson Space Center;
- Marker clusters are used to signify clusters of activities at each coordinate, like launches from a specific site; and
- Lines are drawn to represent the distances between two given coordinates.

Source Code: https://github.com/usmanzafarmirza/applied-data-science-capstone/blob/master/Launch%20Sites%20Locations%20Analysis%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

The ensuing diagrams and charts were utilized for data visualization:

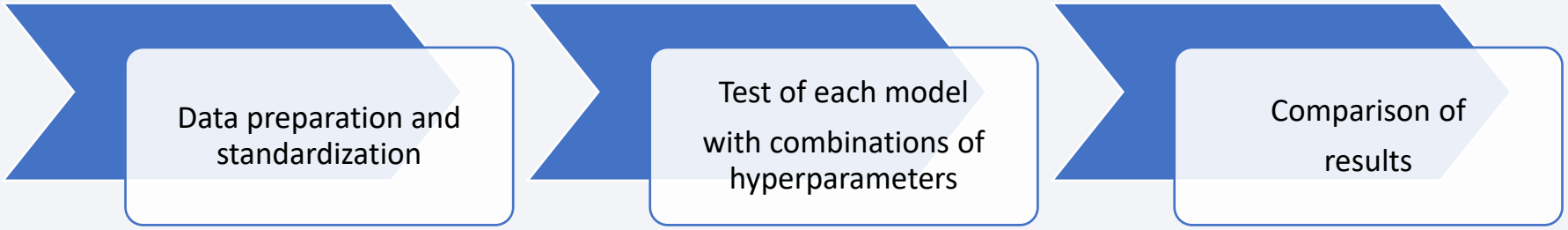
- Proportional representation of launches by site;
- Payload distribution range;

This blend facilitated prompt analysis of the interrelation between payloads and launch sites, assisting in pinpointing the most suitable location for launch based on payload specifics.

Source Code: https://github.com/usmanzafarmirza/applied-data-science-capstone/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

A comparison was made between four classification models: logistic regression, support vector machine, decision tree, and k-nearest neighbors.



Source Code: https://github.com/usmanzafarmirza/applied-data-science-capstone/blob/master/Machine%20Learning%20Prediction.ipynb

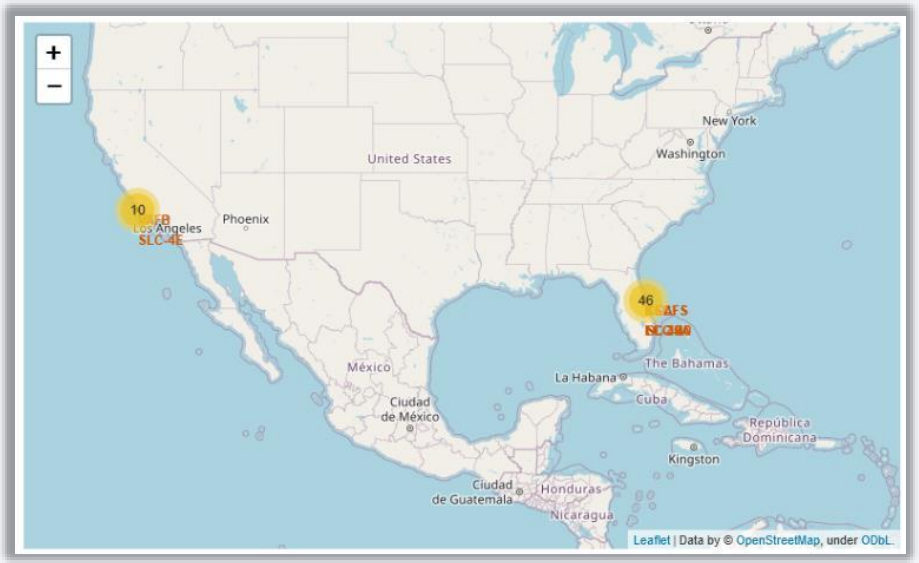
Results

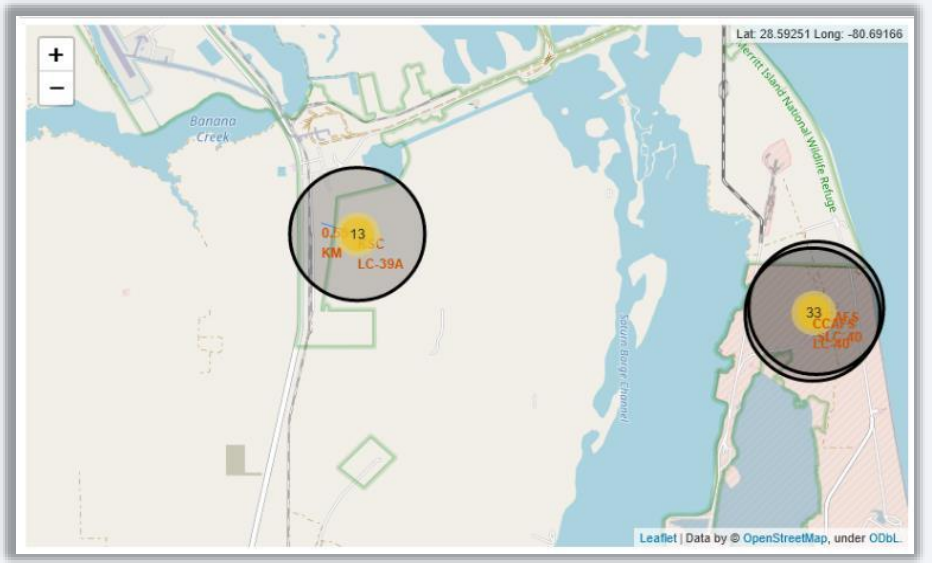
The outcomes of the exploratory data analysis include:

- Space X employs four distinct launch locations;
- Initial launches were conducted by Space X itself and NASA;
- The mean payload for the F9 v1.1 booster is approximately 2,928 kg;
- The inaugural successful landing took place in 2015, five years after the first launch;
- Numerous Falcon 9 booster variants have achieved successful landings on drone ships while carrying payloads above average;
- Nearly 100% of mission outcomes have been successful;
- Two booster versions, namely F9 v1.1 B1012 and F9 v1.1 B1015, failed to land on drone ships in 2015;
- As the years have progressed, the number of successful landing outcomes has improved substantially.

Results

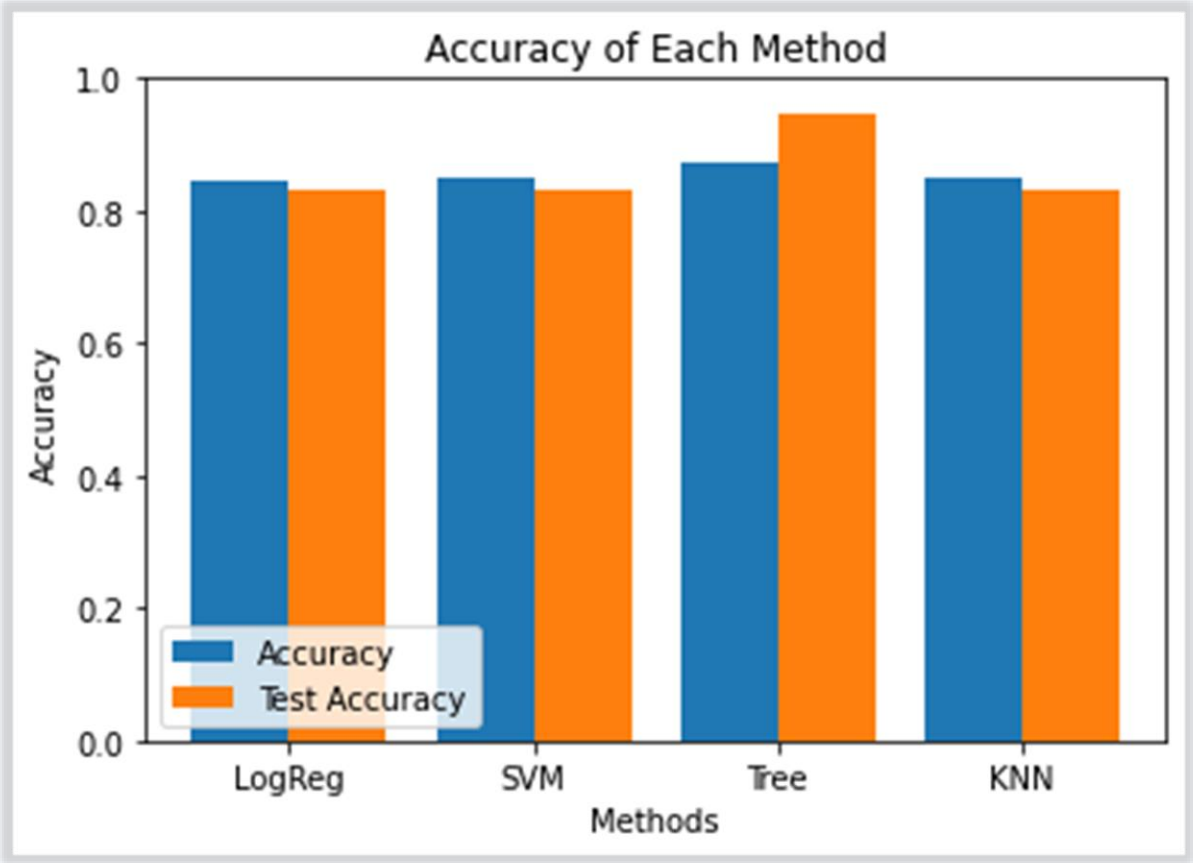
- Interactive analytics revealed that launch sites are typically situated in safe locations near the sea and are surrounded by well-developed logistical infrastructure.
- The majority of launches occur at launch sites on the east coast.

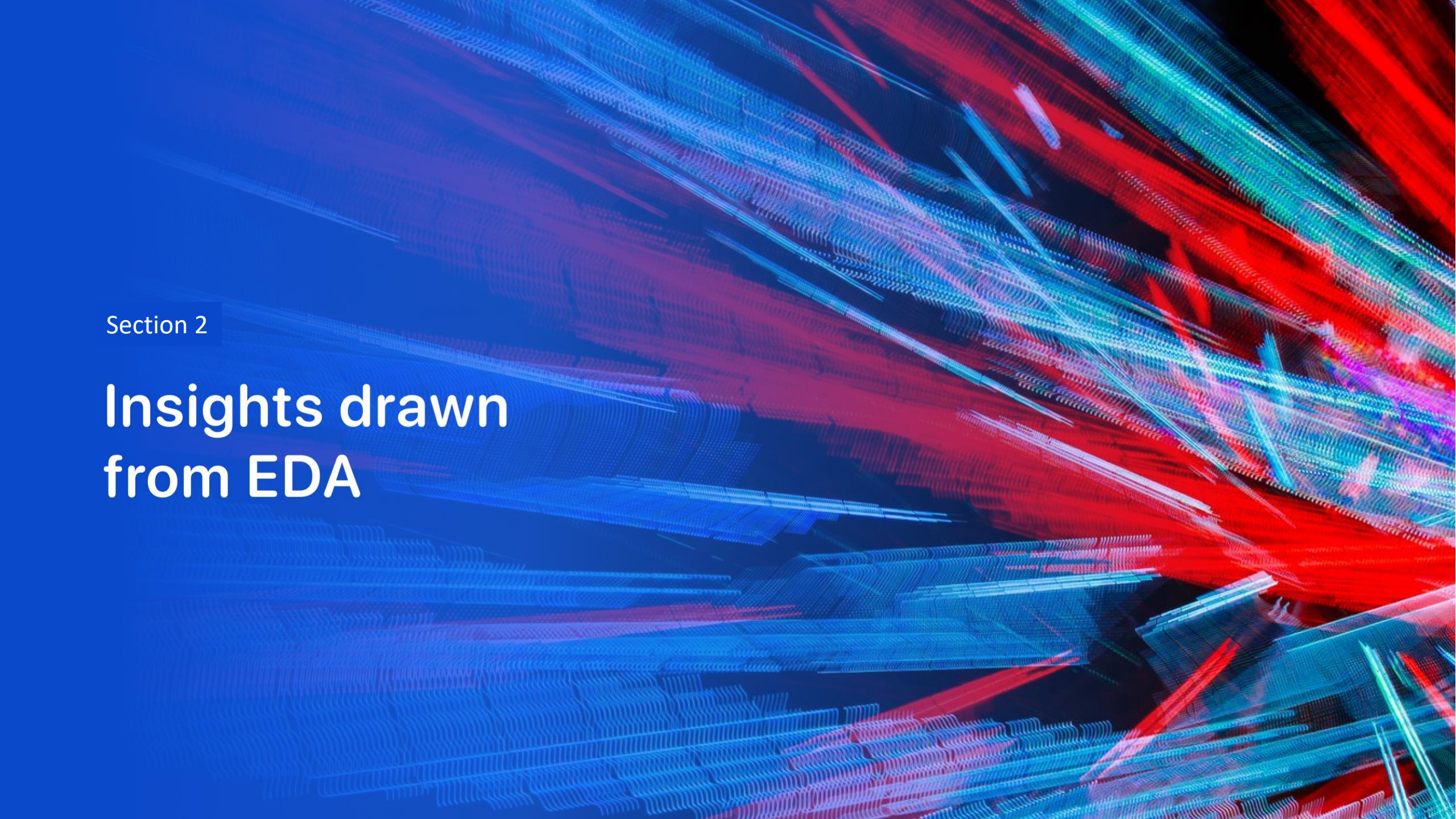




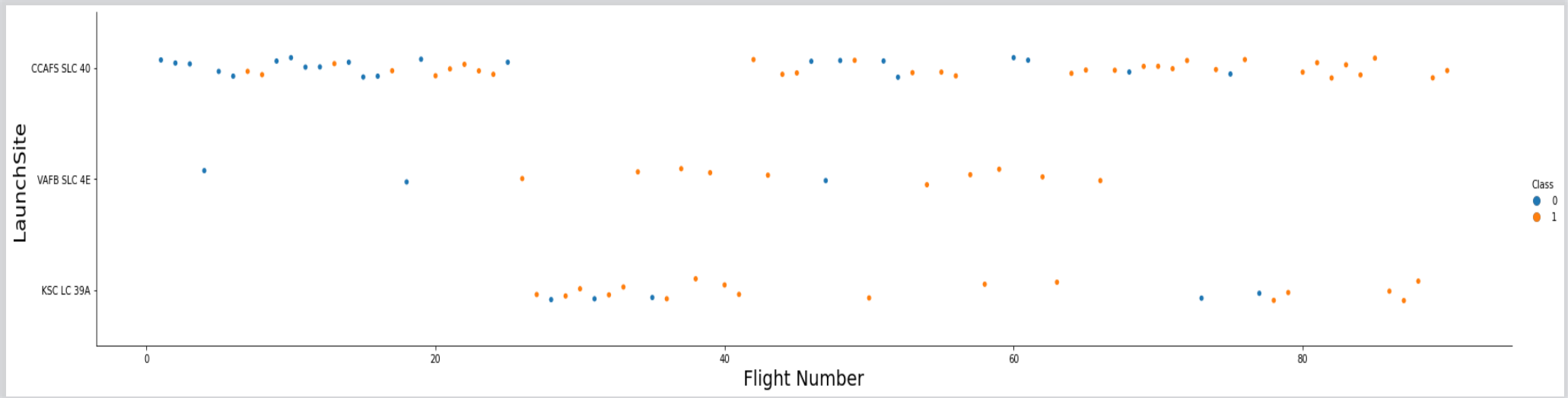
Results

- According to predictive analysis, the Decision Tree Classifier is the most effective model for predicting successful landings, with an accuracy of over 87% and test data accuracy exceeding 94%.



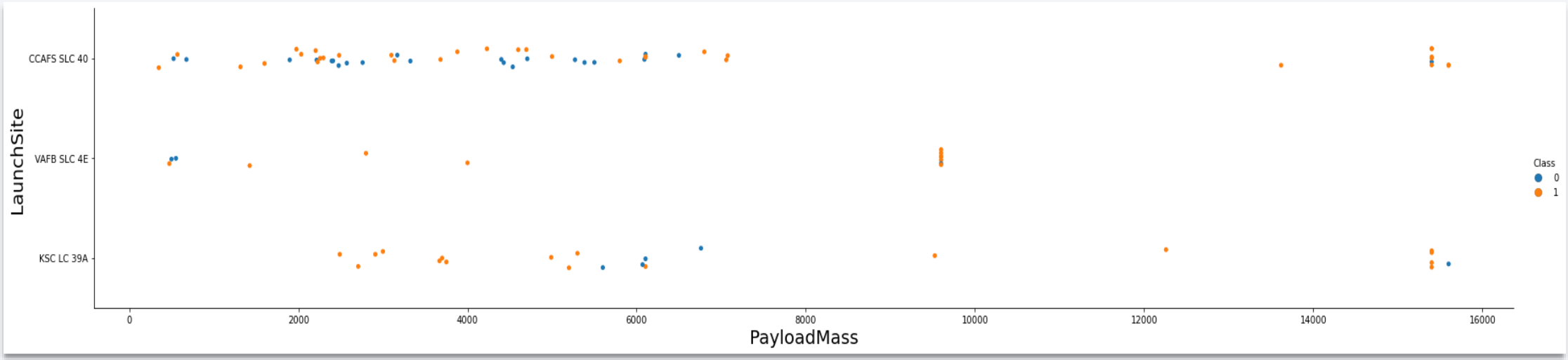

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- As indicated by the above plot, CCAF5 SLC 40 is currently the best launch site, with the majority of recent launches being successful. VAFB SLC 4E and KSC LC 39A are in second and third place, respectively.
- Additionally, the overall success rate has improved over time.

Payload vs. Launch Site



- Payloads exceeding 9,000kg (comparable to the weight of a school bus) boast an exceptional success rate.
- Payloads weighing over 12,000kg appear to be feasible solely at the CCAFS SLC 40 and KSC LC 39A launch sites.

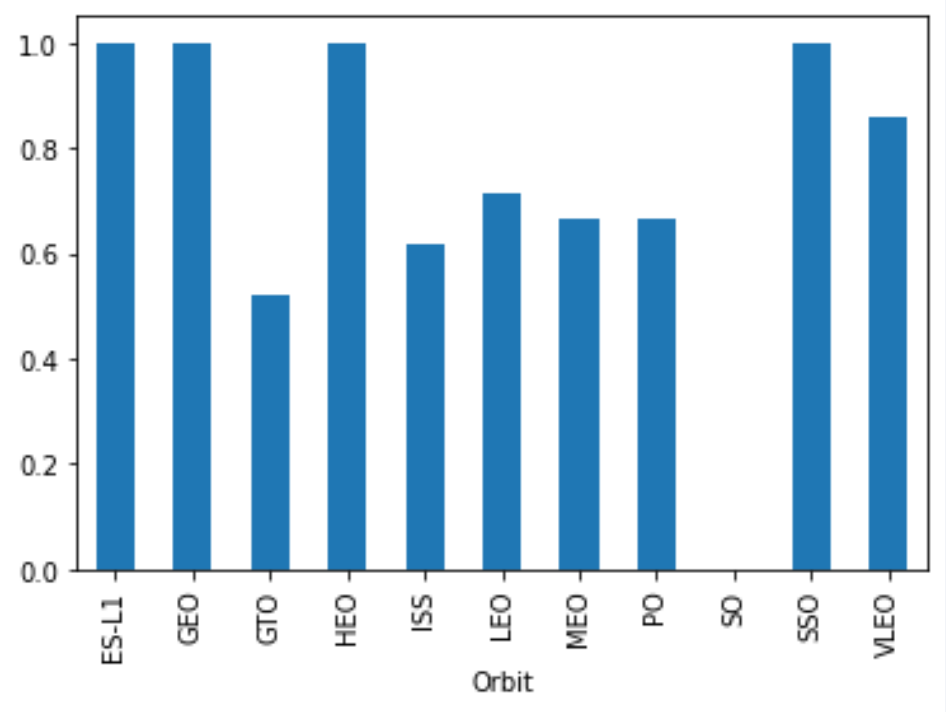
Success Rate vs. Orbit Type

The highest success rates are observed for the following orbits:

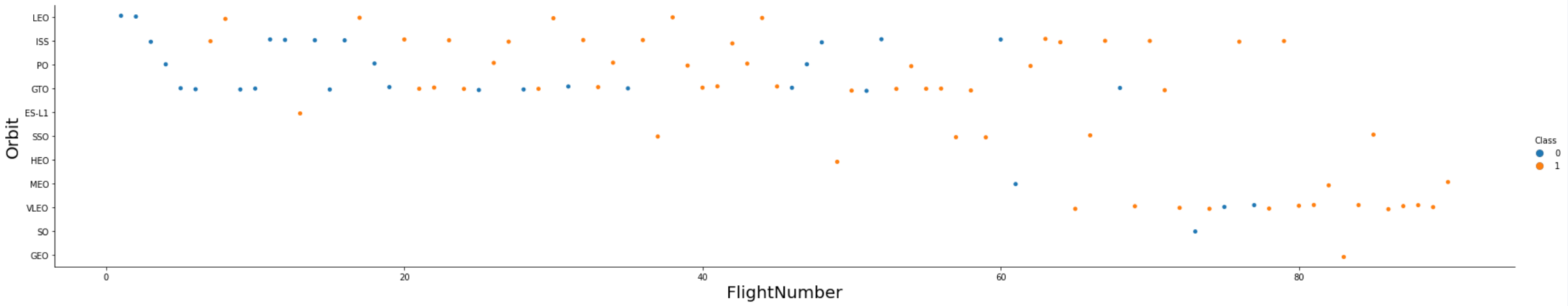
- ES-L1
- GEO
- HEO
- SSO.

These are followed by:

- VLEO (with a success rate above 80%)
- LFO (with a success rate above 70%).

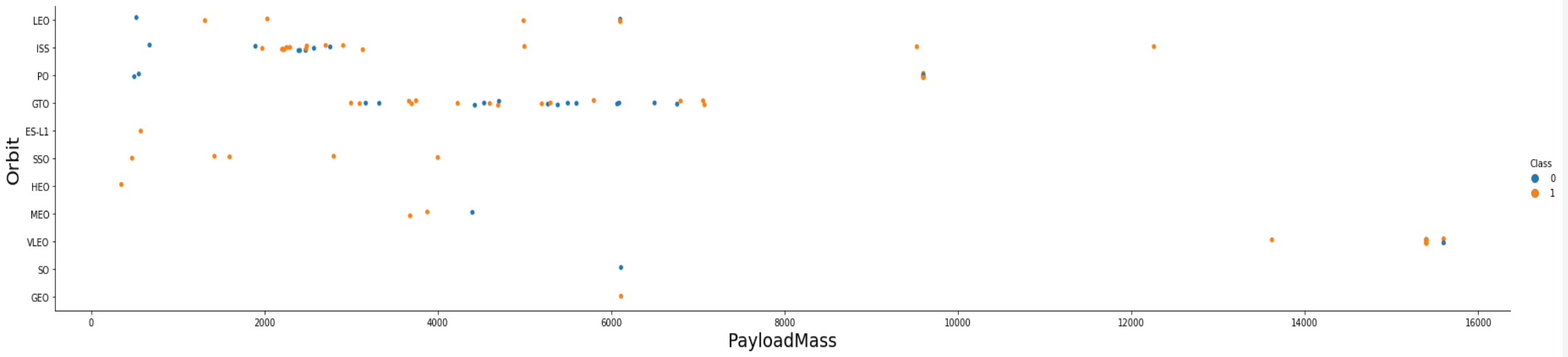


Flight Number vs. Orbit Type



- It appears that the success rate for all orbits has improved over time.
- The VLEO orbit may present a new business opportunity due to its recent increase in frequency.

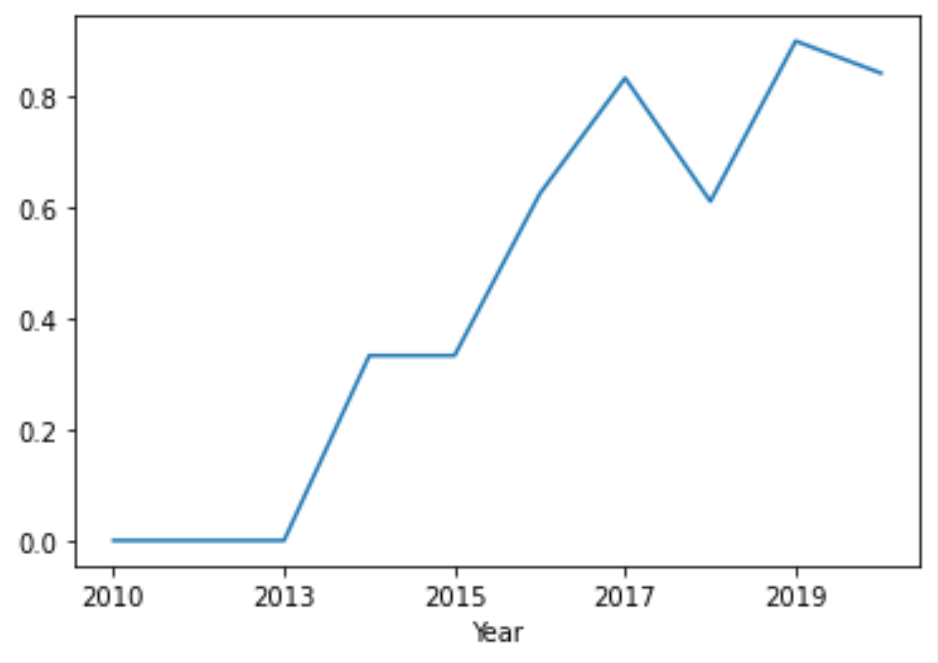
Payload vs. Orbit Type



- There does not appear to be a correlation between payload and success rate for the GTO orbit.
- The ISS orbit has a broad range of payloads and a high success rate.
- There have been few launches to the SO and GEO orbits.

Launch Success Yearly Trend

- The success rate began to rise in 2013 and continued to do so until 2020.
- It appears that the initial three years were dedicated to making adjustments and improving the technology.



All Launch Site Names

- According to data, there are four launch sites:

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- So, the four launch sites were obtained by selecting unique occurrences of the “launch site” values from the dataset.

Launch Site Names Begin with 'CCA'

- The following are five records where the launch sites 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	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

- The following five samples show Cape Canaveral launches.

Total Payload Mass

- Total payload carried by boosters from NASA:

Total Payload (kg)
111.268

- The total payload has been computed by aggregating all payloads whose codes include 'CRS,' which corresponds to NASA.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:

Avg Payload (kg)
2.928

- By applying a filter based on the booster version mentioned earlier and performing calculations, we have determined that the average payload mass is 2,928 kg.

First Successful Ground Landing Date

- First successful landing outcome on ground pad:

Min Date
22-12-2015

- By applying a data filter based on the successful landing outcome specifically on a ground pad, and subsequently determining the minimum value for the date, it is possible to identify the earliest occurrence, which took place on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters with Successful Landing on Drone Ship with Payload Mass between 4000 and 6000:

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

- Based on the aforementioned filters, a careful selection process has yielded a set of four distinct booster versions as the outcome.

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

- The process of categorizing mission outcomes and tallying the corresponding records has resulted in the aforementioned summary.

Boosters Carried Maximum Payload

- Boosters Which Have Carried The Maximum Payload Mass

Booster Version (+)	Booster Version
F9 B5 B1048.4	F9 B5 B1051.4
F9 B5 B1048.5	F9 B5 B1051.6
F9 B5 B1049.4	F9 B5 B1056.4
F9 B5 B1049.5	F9 B5 B1058.3
F9 B5 B1049.7	F9 B5 B1060.2
F9 B5 B1051.3	F9 B5 B1060.3

- The following boosters have successfully transported the maximum payload mass as recorded in the dataset.

2015 Launch Records

- In the year 2015, the dataset indicates instances of failed landing outcomes on drone ships, along with the corresponding booster versions and launch site names.

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- The aforementioned list comprises the only two instances recorded within the dataset.

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

- Below is the ranking of all landing outcomes recorded between the dates June 4, 2010, and March 20, 2017:

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

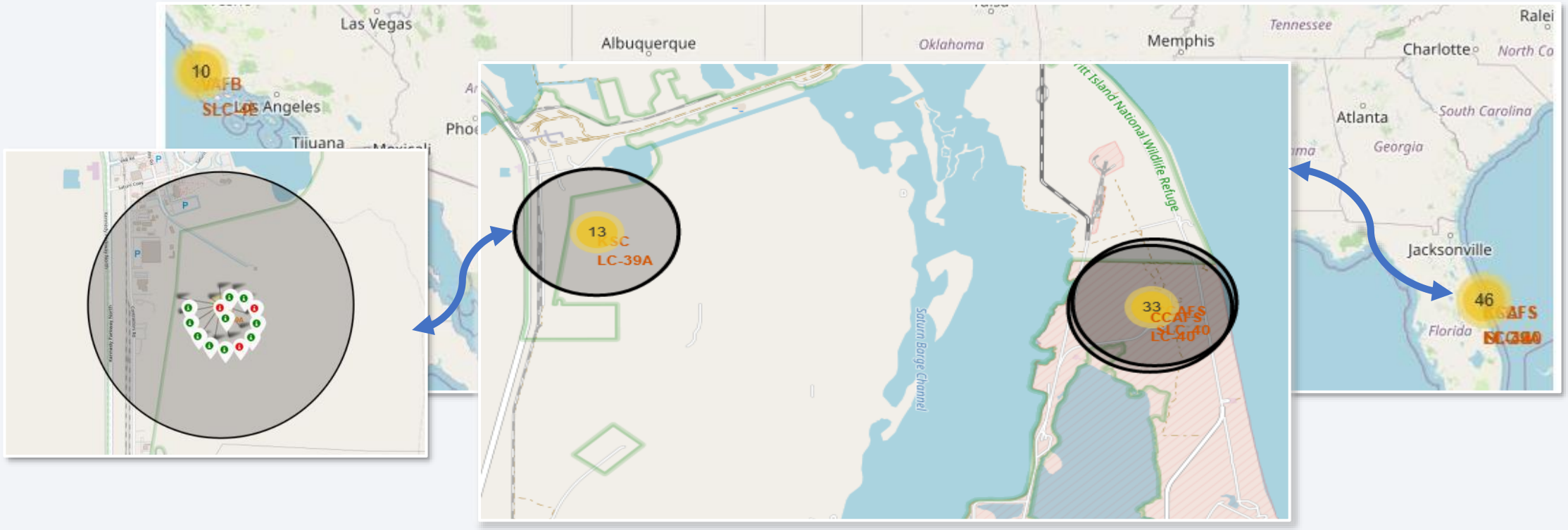
- This particular data view draws attention to the significance of considering the "No attempt" landing outcomes in our analysis.


Section 3
Launch Sites
Proximities Analysis

All launch sites

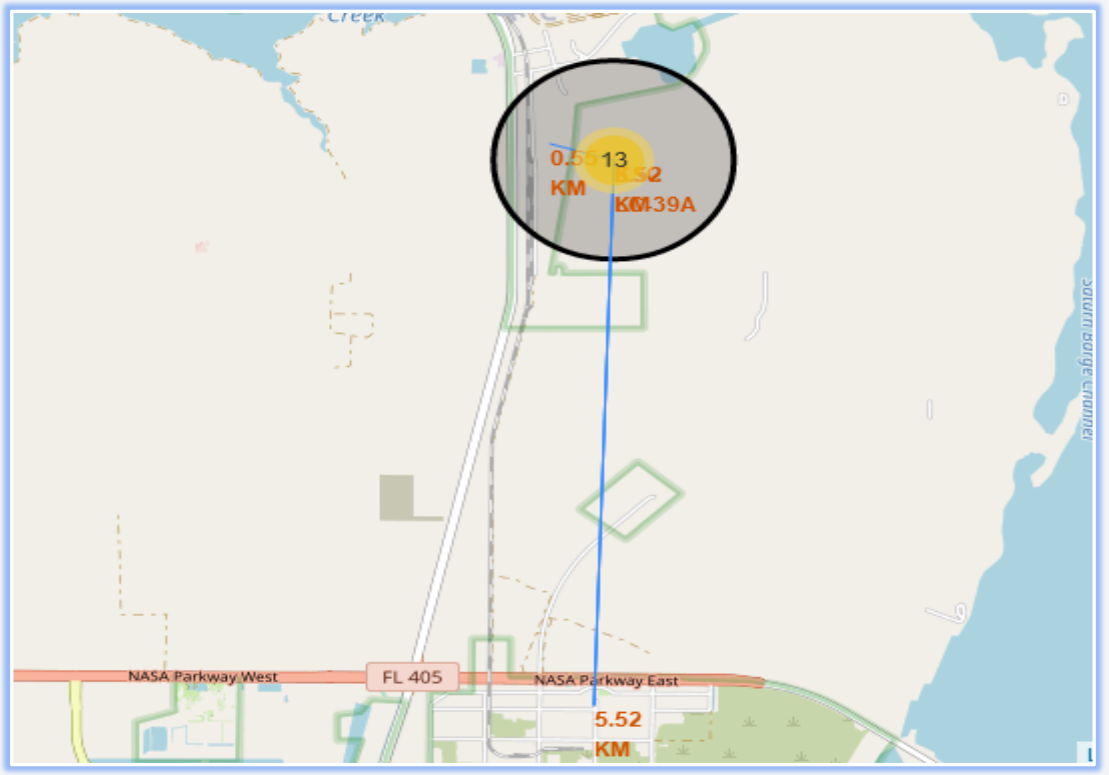
- The launch sites are strategically located in close proximity to the sea, primarily with a focus on ensuring safety. Simultaneously, they are positioned in reasonable proximity to major transportation arteries such as roads and rail networks.

Launch Outcomes by Site

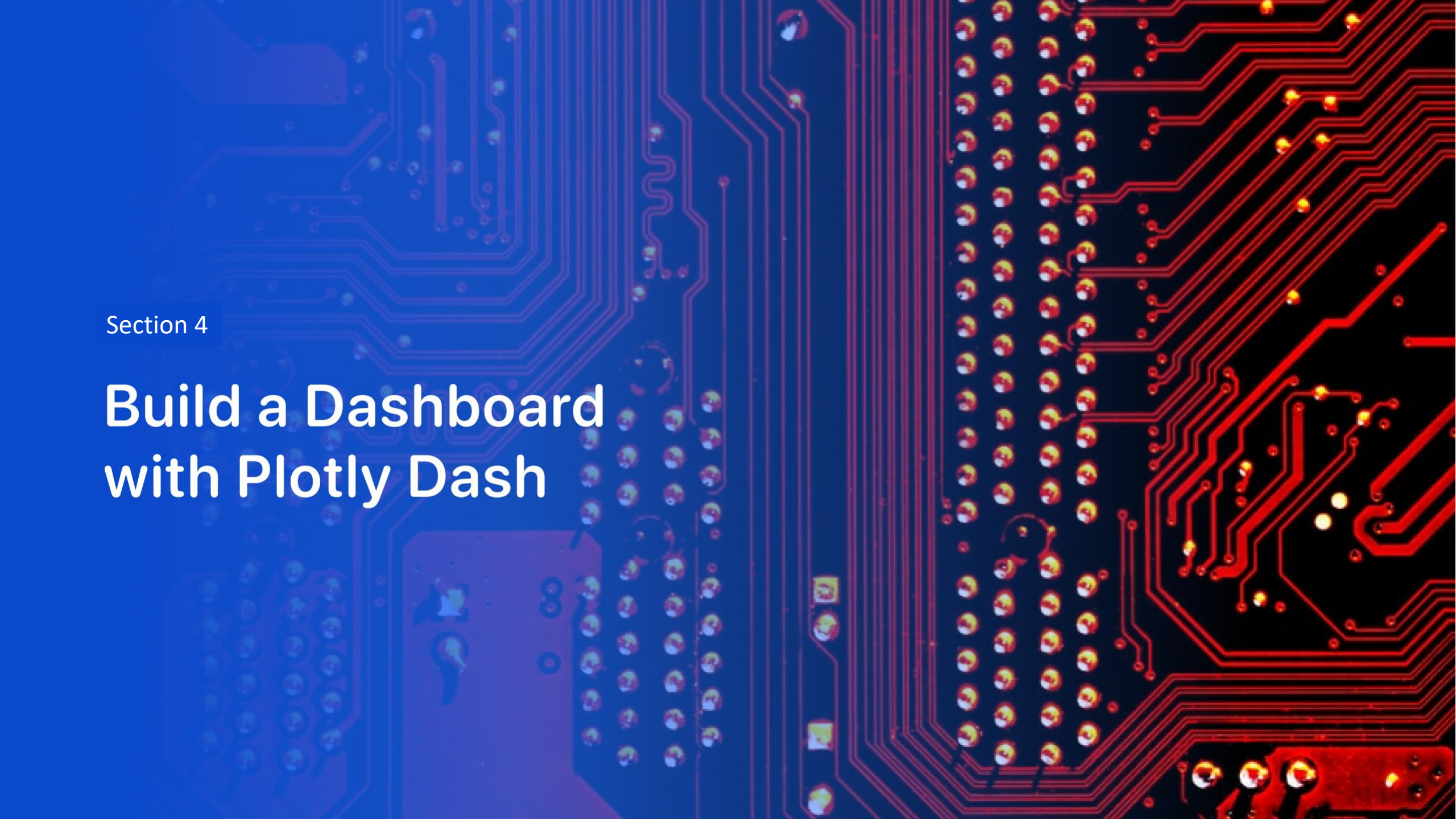


- The green markers serve as indicators of successful outcomes, while the red markers are employed to denote instances of failure.

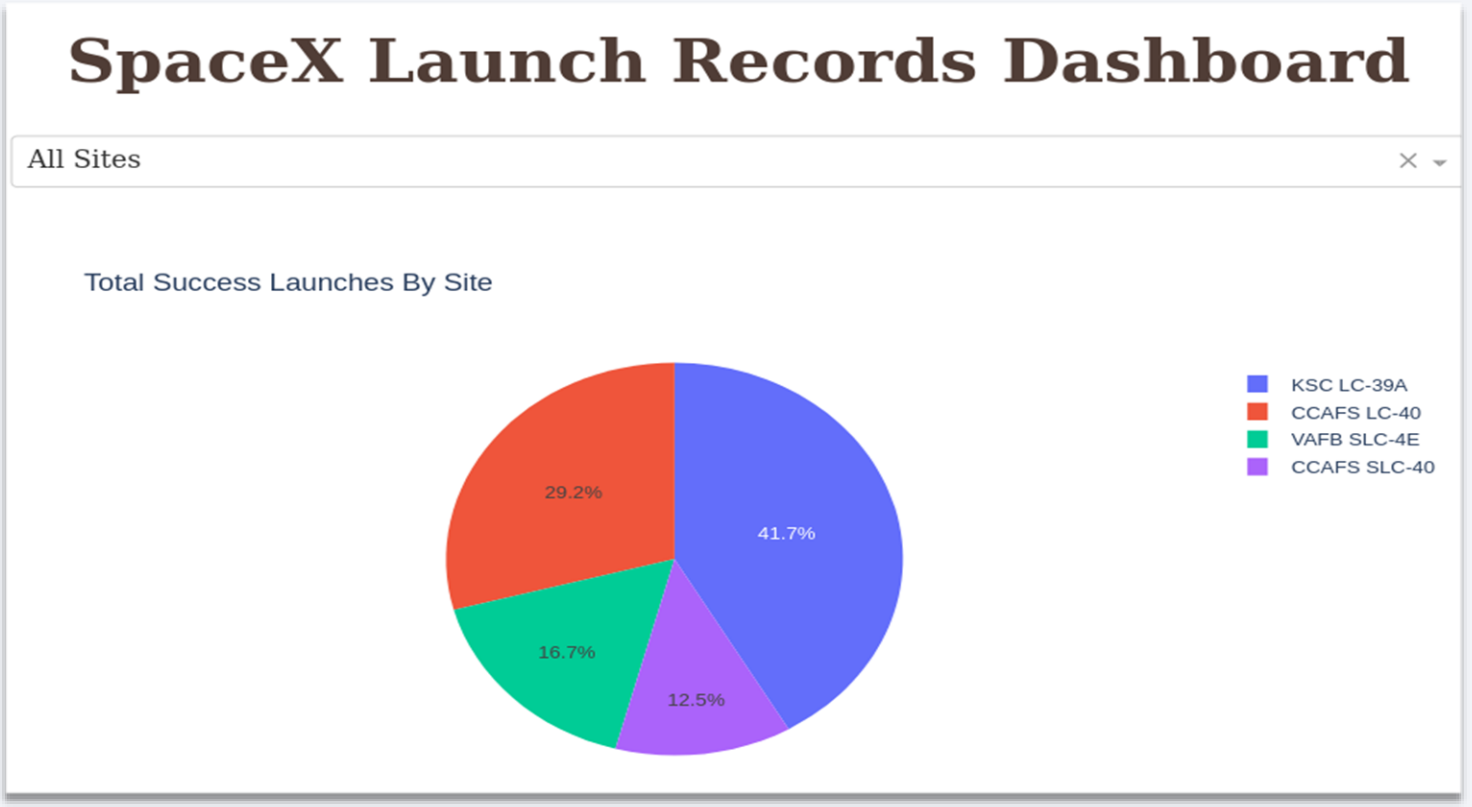
Logistics and Safety



- Launch site KSC LC-39A boasts commendable logistical advantages, as it is strategically positioned in close proximity to both rail and road infrastructure while maintaining a considerable distance from densely populated areas.

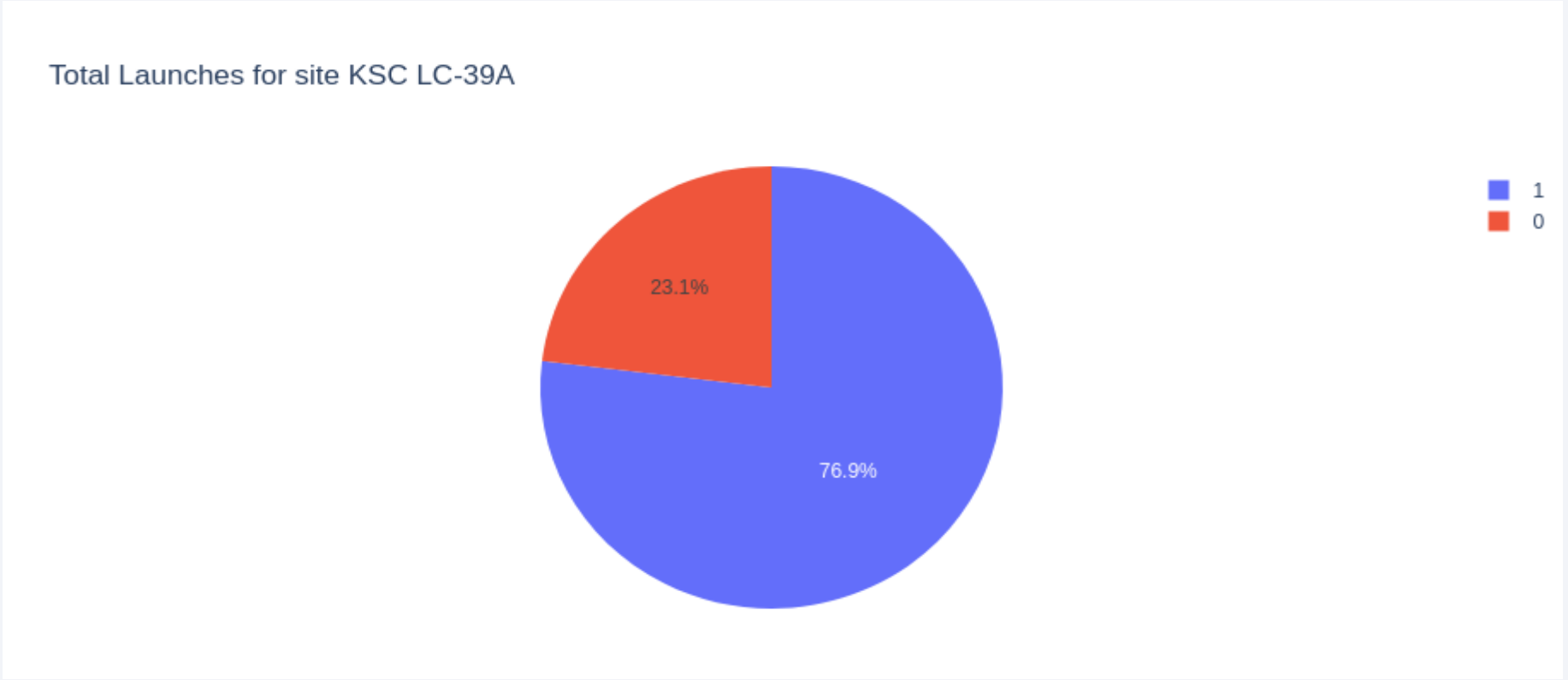

Section 4
Build a Dashboard with Plotly Dash

Successful Launches by Site



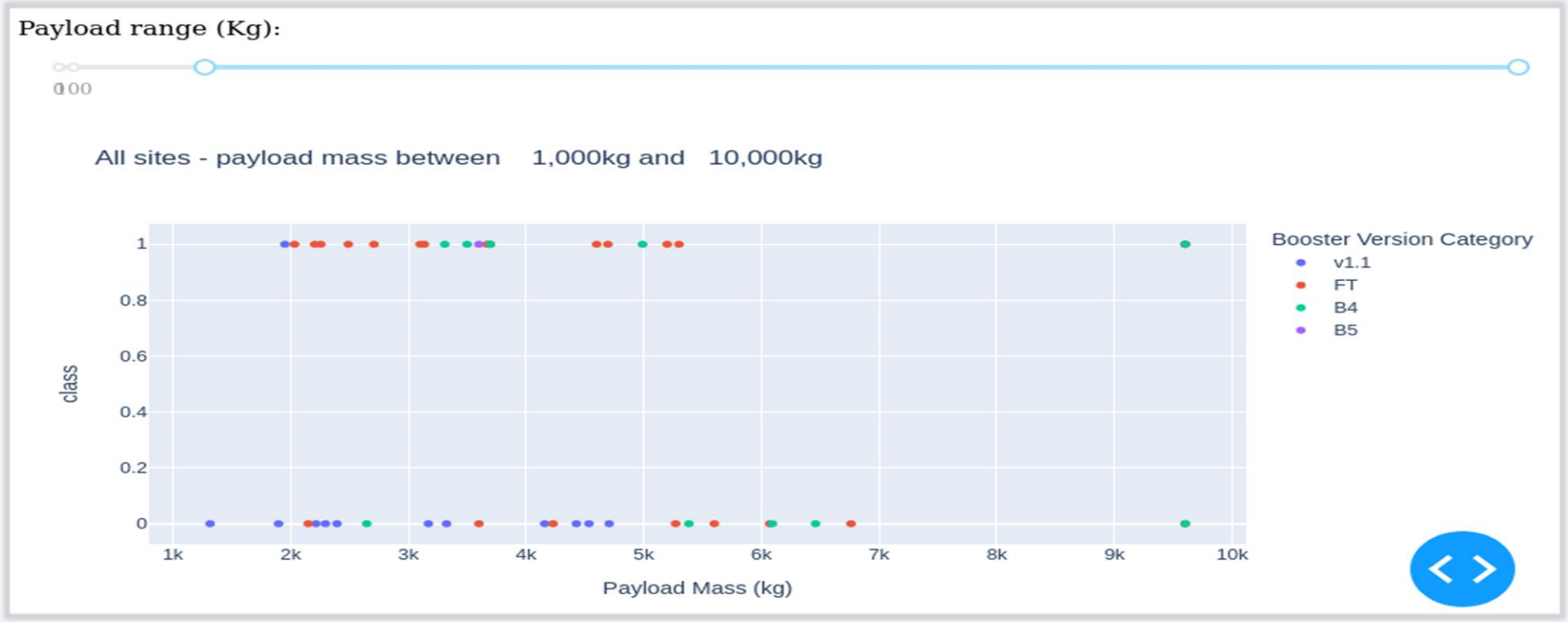
- The selection of the launch site plays a pivotal role in determining the success of missions.

Launch Success Ratio for KSC LC-39A



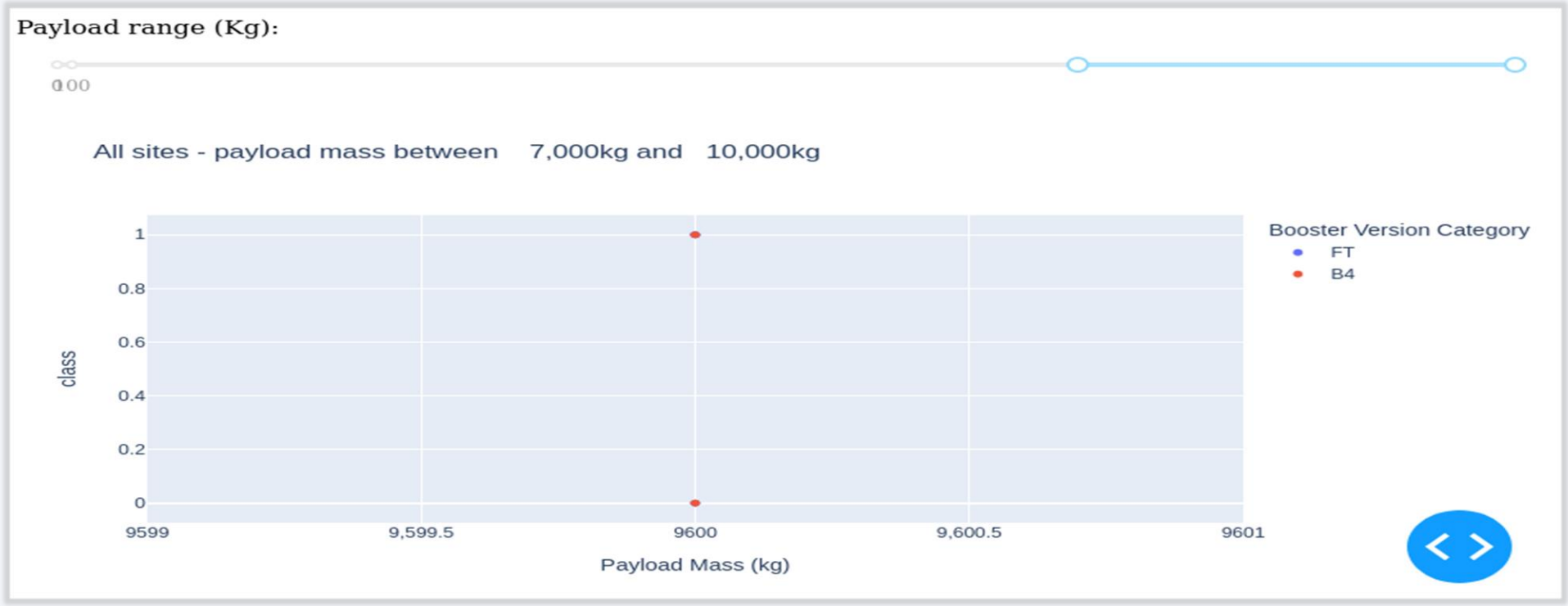
- The success rate of launches at this particular site stands at an impressive 76.9%.

Payload vs. Launch Outcome



- The combination of payloads weighing under 6,000kg and utilizing FT boosters has proven to be the most successful configuration in terms of mission outcomes.

Payload vs. Launch Outcome



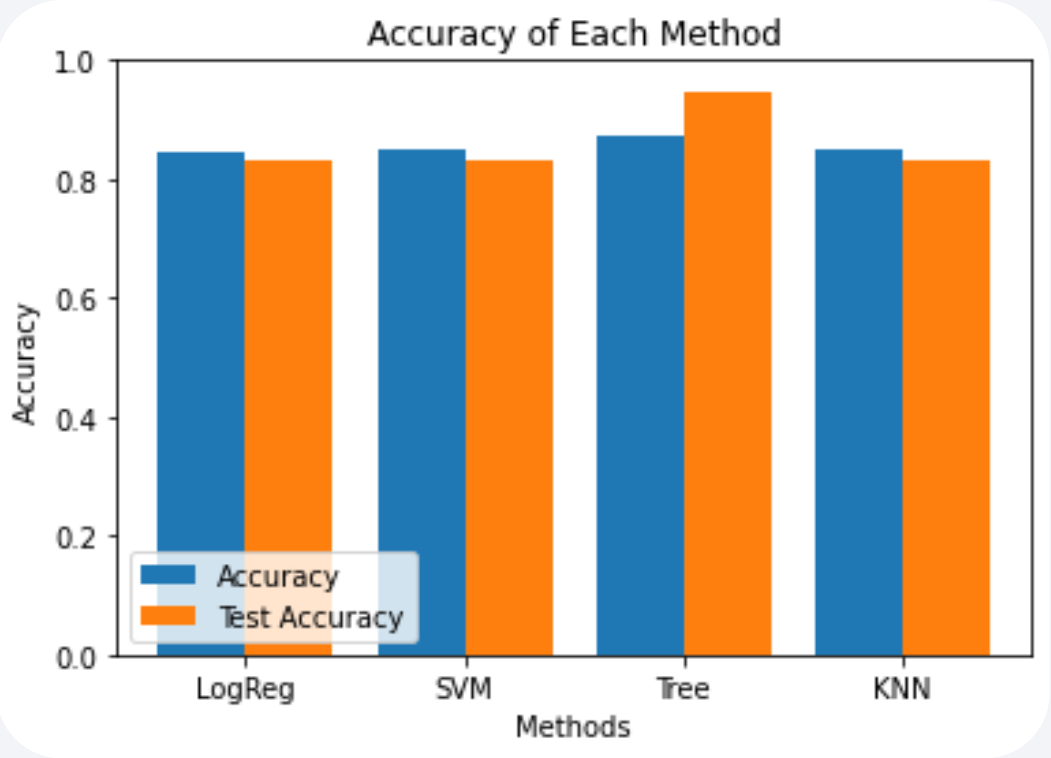
- Due to the limited availability of data, it is challenging to accurately estimate the risk associated with launches exceeding 7,000kg in payload weight.

Section 5

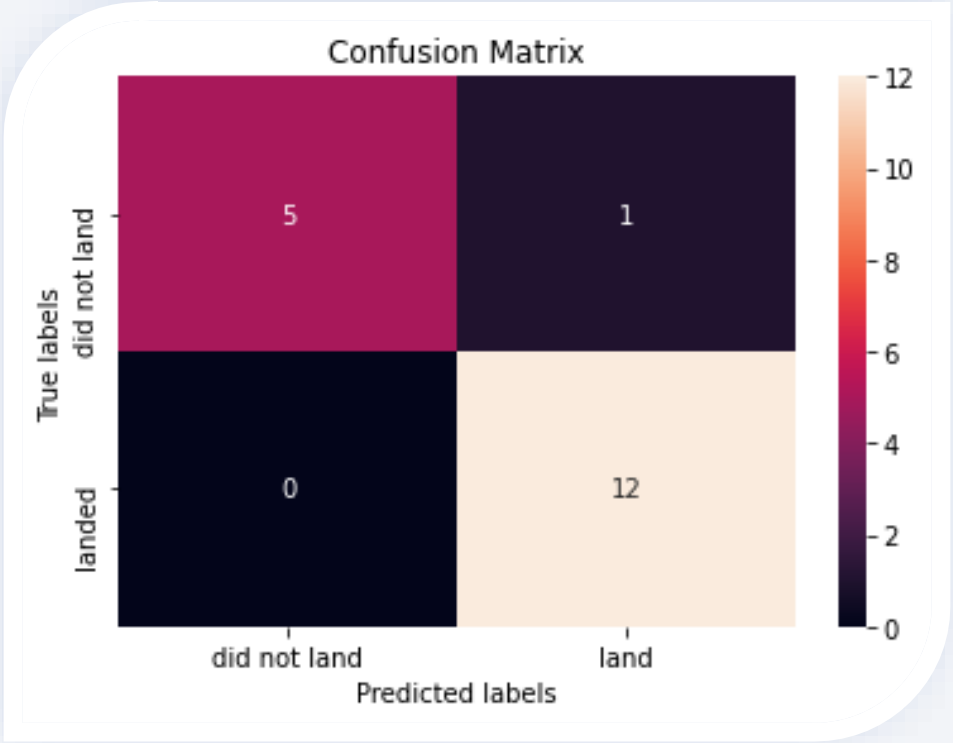
Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were subjected to testing, and their respective accuracies have been graphically represented for comparison.
- The Decision Tree Classifier emerged as the model with the highest classification accuracy, surpassing 87% in its predictive performance.



Confusion Matrix: Decision Tree Classifier



- The Confusion Matrix of the Decision Tree Classifier serves as evidence of its accuracy, as it demonstrates a significant number of true positives and true negatives in comparison to the false positives and false negatives.

Conclusions

- Analysis was conducted across various data sources, with the conclusions being progressively refined throughout the process.
- Undoubtedly, the KSC LC-39A launch site stands out as the premier choice for conducting launches.
- Launches with payloads exceeding 7,000kg were found to carry a lower risk.
- While a majority of mission outcomes are successful, there is a notable trend of improvement in successful landing outcomes over time. This is attributed to advancements in processes and rocket technology.
- The use of a Decision Tree Classifier is recommended to predict successful landings, thereby potentially increasing profits.

Appendix

- To enhance the reliability and consistency of model tests, it is crucial to assign a specific value to the np.random.seed variable. This will ensure that the randomization process remains controlled and reproducible, enabling more accurate comparisons and evaluations of the models.
- Due to limitations with displaying interactive maps from Folium on GitHub, screenshots were taken to capture and present the map visualizations.

Thank you!