IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alejandro Aguilera
03/30/2024

Outline

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

Executive Summary

The project utilized the following methodologies:

- Data Collection: Leveraging web scraping and the SpaceX API.
- Exploratory Data Analysis (EDA): Including data wrangling, visualization, and interactive analytics.
- Machine Learning Prediction.

Key Findings:

- Successful data collection from public sources.
- EDA identified crucial features for predicting launch success.
- Machine Learning identified optimal characteristics for launch success prediction.
- In summary, our project provides actionable insights for optimizing SpaceX launch success.

Introduction

Our aim is to evaluate the potential of Space Y to rival Space X in the space exploration industry. Central to our assessment are two key inquiries:

- **Cost Prediction:** We seek to devise an accurate method for estimating total launch costs by predicting the success of first-stage rocket landings. This insight will enable Space Y to refine financial planning and competitiveness in the market.

- **Launch Site Selection:** Additionally, we endeavor to identify optimal launch locations for Space Y's operations. Strategic site selection is crucial for minimizing logistical hurdles and maximizing operational efficiency.

Through our analysis, we aim to provide actionable insights that will guide Space Y towards effective competition and success within the space exploration domain.

Section 1

Methodology

Methodology

Executive Summary

Data Collection:

- Gathered launch data through web scraping and SpaceX API.

Data Wrangling & Exploratory Data Analysis (EDA):

- Cleaned and preprocessed data to ensure quality.
- Used visualization and SQL for in-depth data exploration.

Interactive Visual Analytics:

- Created interactive maps with Folium and dashboards with Plotly Dash.

Predictive Analysis:

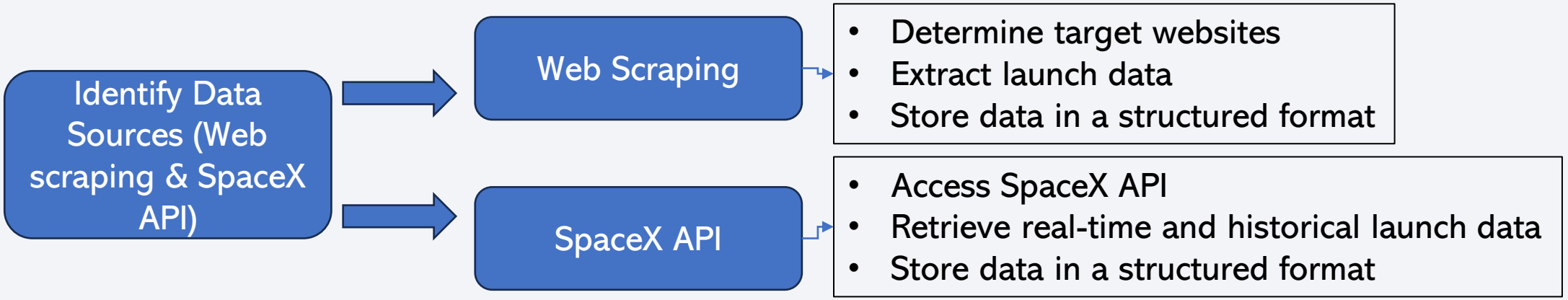
- Employed classification models and tuned them for launch success prediction.
- Evaluated model accuracy, precision, recall, and other metrics to measure capabilities

Data Collection

In the Data Collection phase, we gathered comprehensive launch data through a meticulous process leveraging both web scraping techniques and the SpaceX API.

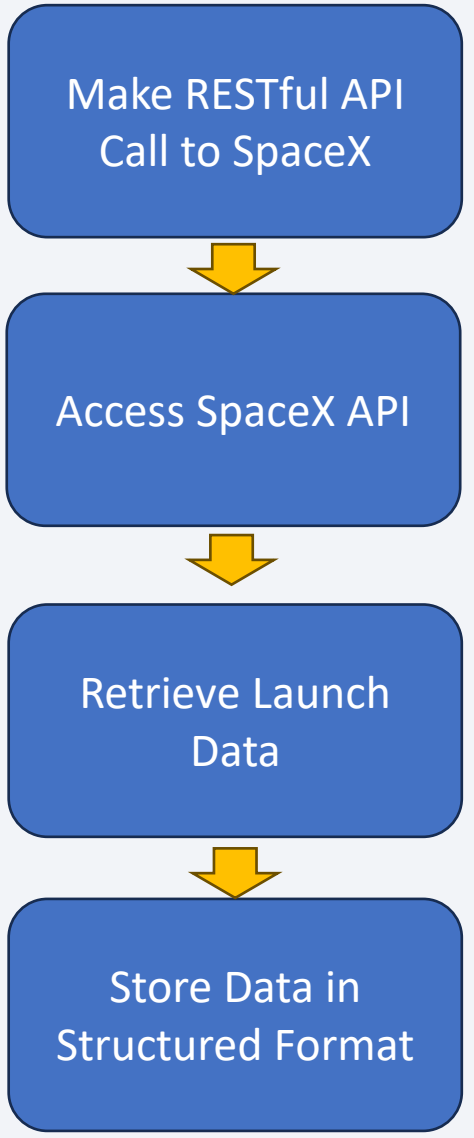
Key Points:

- Utilized web scraping to extract launch records from public sources.
- Integrated SpaceX API to access real-time and historical launch data.
- Collected mission details, outcomes, rocket specifications, and other relevant information.



Data Collection – SpaceX API

- In this phase, we accessed SpaceX's launch data using RESTful API calls to gather real-time and historical information on rocket launches.
- Retrieved details such as mission names, launch dates, outcomes, and rocket specifications.
- Stored the collected data in a structured format for further analysis.
- Source Code: https://github.com/aleagui86/applied-data-science-capstone/blob/main/Data%20Collection%20API.ipynb

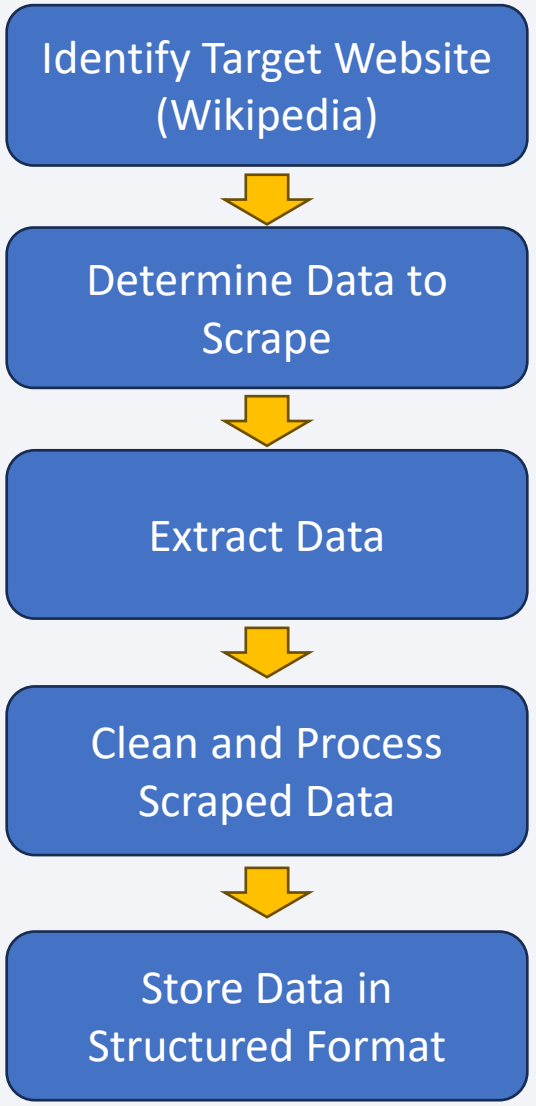


Data Collection - Scraping

- We obtained valuable data on SpaceX launches from Wikipedia through web scraping techniques.
- The data was Downloaded from Wikipedia following a predefined flowchart, and then persisted the obtained data for further analysis.

Source Code:

https://github.com/aleagui86/applied-data-science-capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

During the Data Wrangling phase, we processed the collected data to ensure its quality and prepare it for analysis.

Key Points:

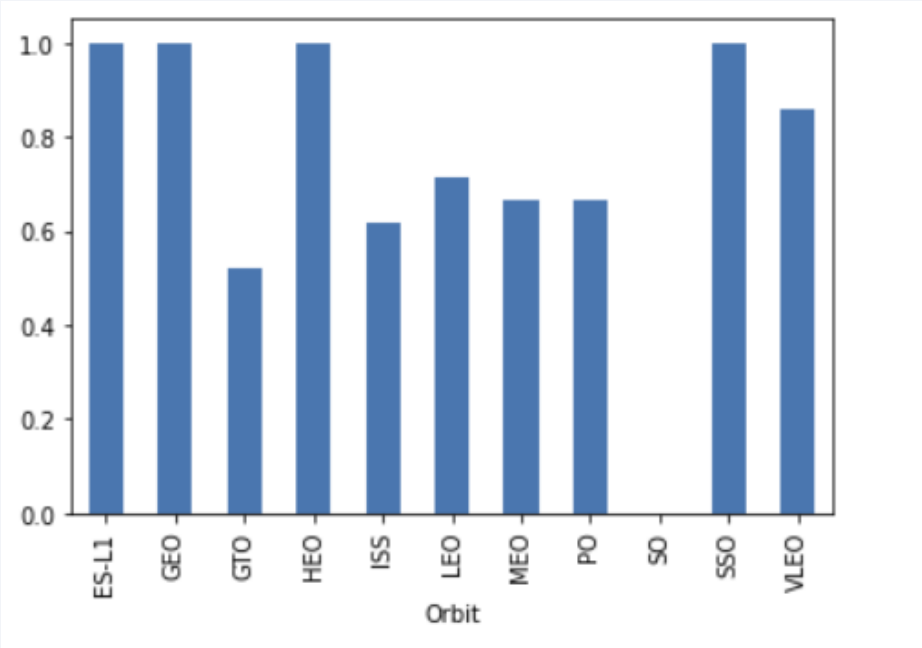
- Cleaned and preprocessed raw data to handle missing values, inconsistencies, and outliers.
- Conducted data transformation and normalization to standardize formats and units.
- Merged and combined multiple datasets to create a unified dataset for analysis.

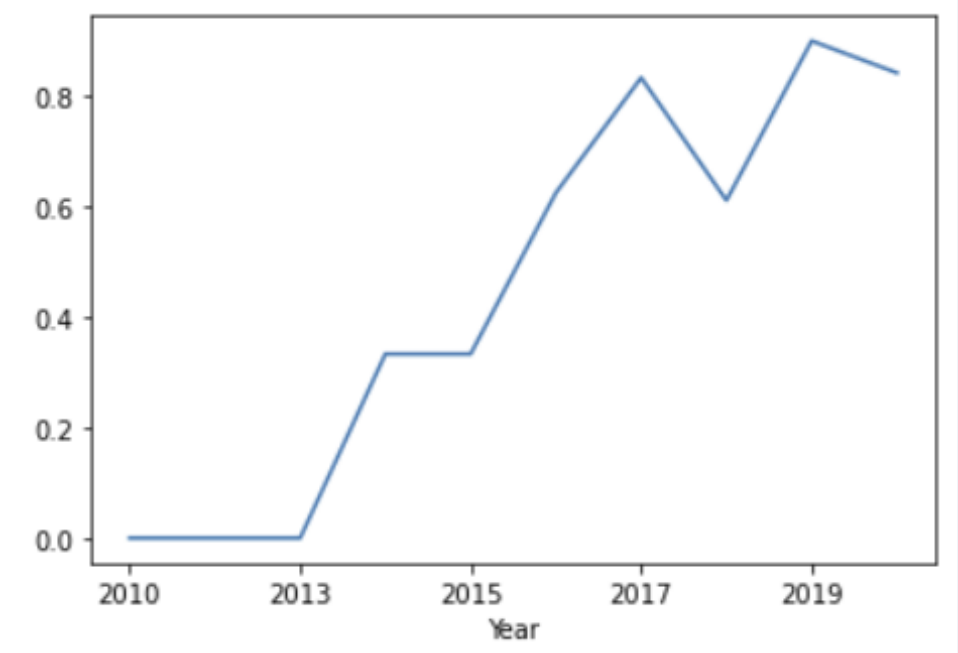
Source Code:

https://github.com/aleagui86/applied-data-science-capstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- To explore data, scatterplots, barplots and line graphs were used to visualize the relationship between pair of features:





Source Code:

https://github.com/aleagui86/applied-data-science-capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

During the exploration of our data using SQL, we conducted various queries to extract insights:

- Unique launch sites identified.
- Top 5 launch sites beginning with 'CCA'.
- Total payload mass by NASA for CRS.
- Average payload mass for booster version F9 v1.1.
- Date of first successful ground pad landing.
- Boosters successful on drone ship with payload between 4000 and 6000 kg.
- Total successful and failed mission outcomes.
- Booster versions with maximum payload mass.
- Failed drone ship landings in 2015 with booster versions and launch sites.
- Ranking of landing outcomes between 2010-06-04 and 2017-03-20.

Source Code: https://github.com/aleaqui86/applied-data-science-capstone/blob/main/EDA.ipynb

Build an Interactive Map with Folium

- Created markers, circles, polylines, and popups.
- Used markers to identify key locations.
- Utilized circles to highlight specific areas of interest.
- Implemented polylines to depict rocket flight paths.
- Added popups for additional details and context.

Source Code:

https://github.com/aleagui86/applied-data-science-capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

We employed a variety of graphs and plots to visualize our data effectively:

- **Percentage of Launches by Site:** This graph provides insights into the distribution of launches across different launch sites.
- **Payload Range:** By visualizing the payload range, we gained a comprehensive understanding of the payloads carried by rockets in our dataset.

This combination of graphs enabled us to analyze the relationship between payloads and launch sites efficiently. Through this analysis, we identified optimal launch locations based on payload considerations.

Source Code: https://github.com/aleagui86/applied-data-science-capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Compared four classification models: logistic regression, support vector machine, decision tree, and k-nearest neighbors.
- Built, evaluated, and improved models using techniques such as cross-validation and hyperparameter tuning.
- Selected the best-performing model based on metrics like accuracy, precision, recall, and F1-score.

Build Classification Models

- Logistic Regression
- Support Vector Machine
- Decision Tree
- K-Nearest Neighbors

Evaluate and Improve Models

- Cross-validation
- Hyperparameter Tuning

Select Best Performing Model

- Compare Model Metrics
- Choose Model with Highest Accuracy

Source Code:

https://github.com/aleagui86/applied-data-science-capstone/blob/main/Machine%20Learning%20Prediction.ipynb

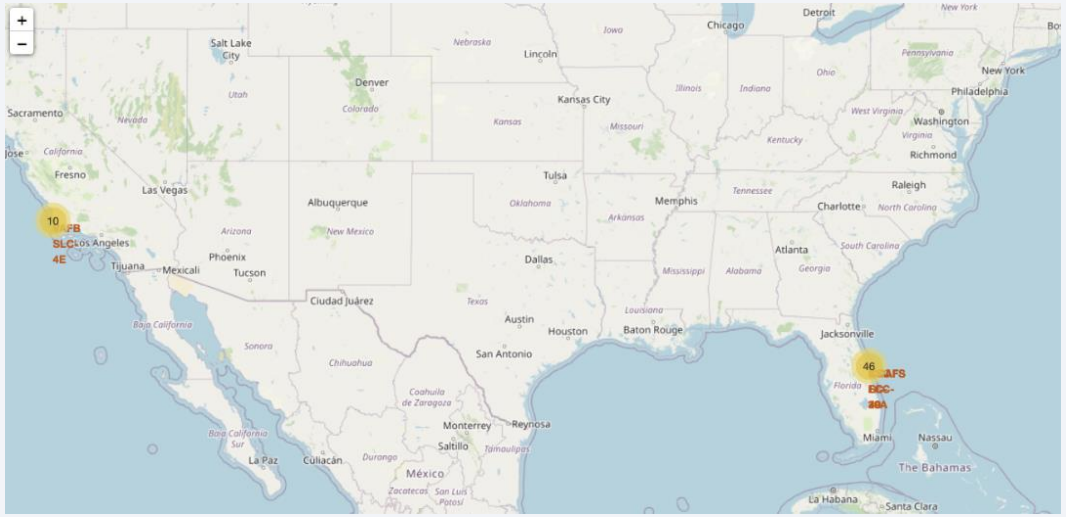
Results

Exploratory Data Analysis Findings:

- Space X operates from 4 distinct launch sites.
- Initial launches targeted Space X and NASA missions.
- The average payload of F9 v1.1 boosters stands at 2,928 kg.
- Successful landing outcomes began in 2015, five years post the first launch.
- Numerous Falcon 9 booster versions successfully landed on drone ships with payloads exceeding the average.
- Mission success rates reached nearly 100%.
- Two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, experienced failed landings on drone ships in 2015.
- Landing outcomes improved over time.

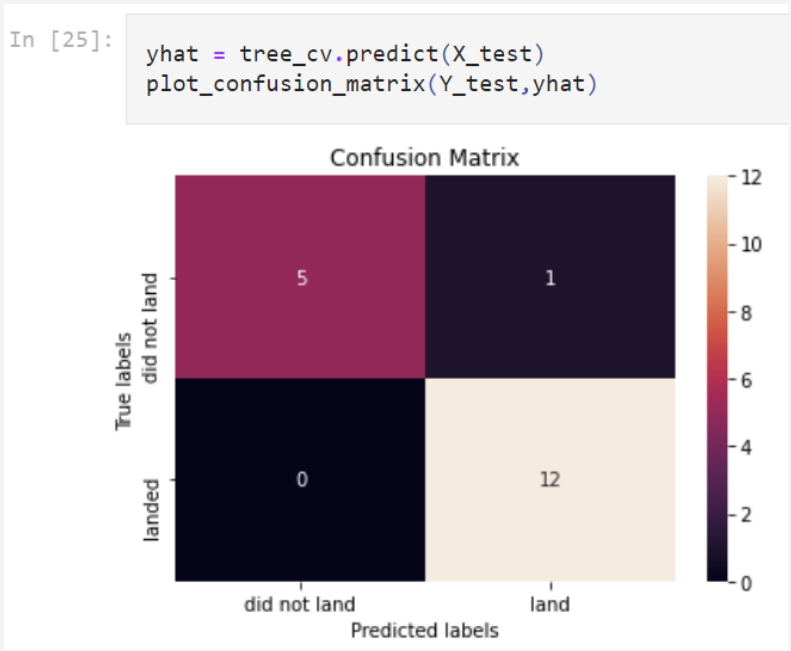
Results

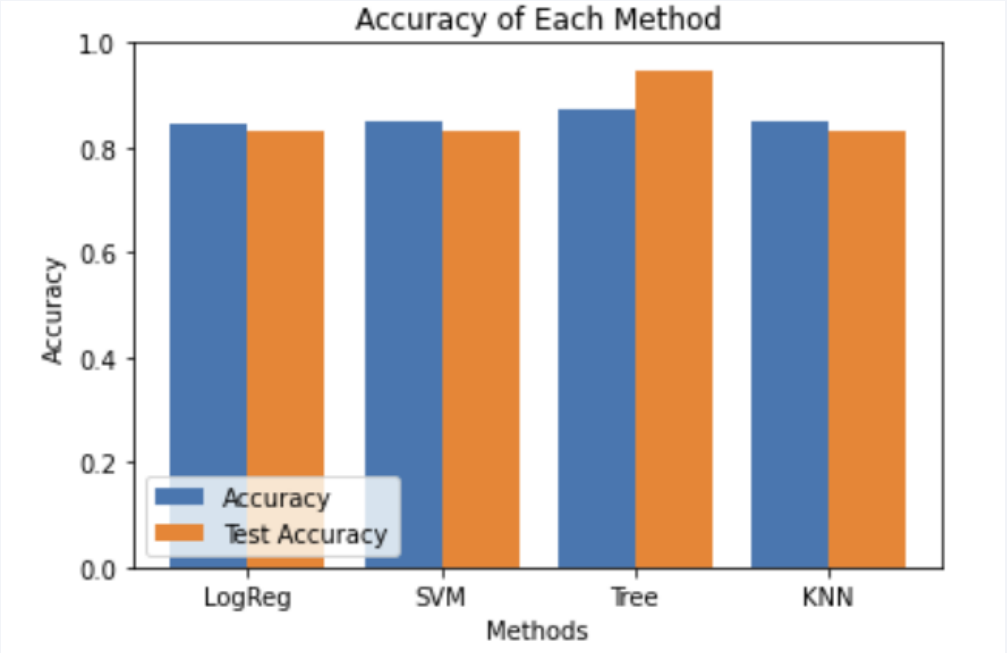
- By leveraging interactive analytics, we discovered that launch sites are typically located in secure areas, often near the coast, with robust logistical infrastructure.
- The majority of launches occur at launch sites along the east coast.

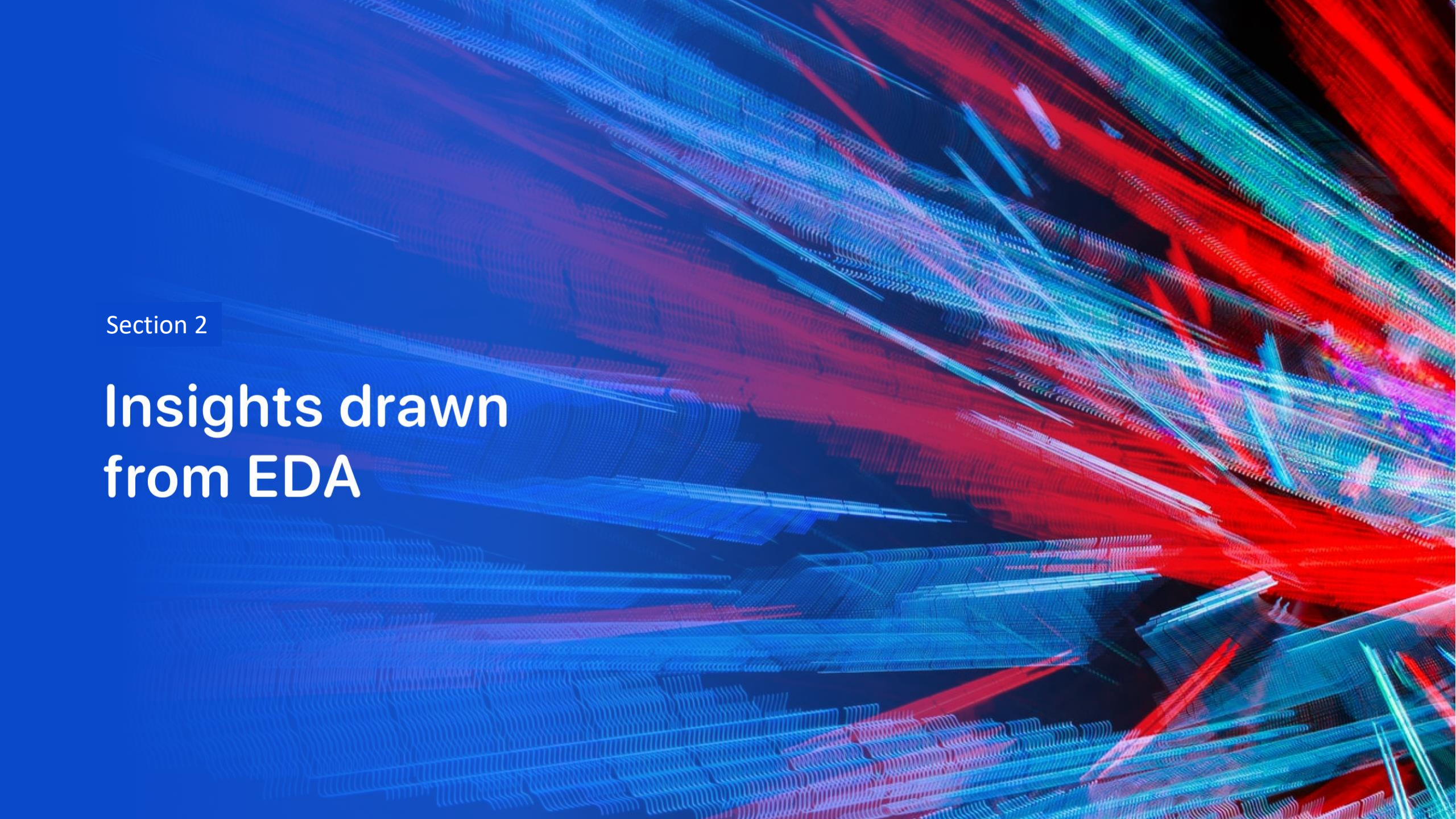


Results

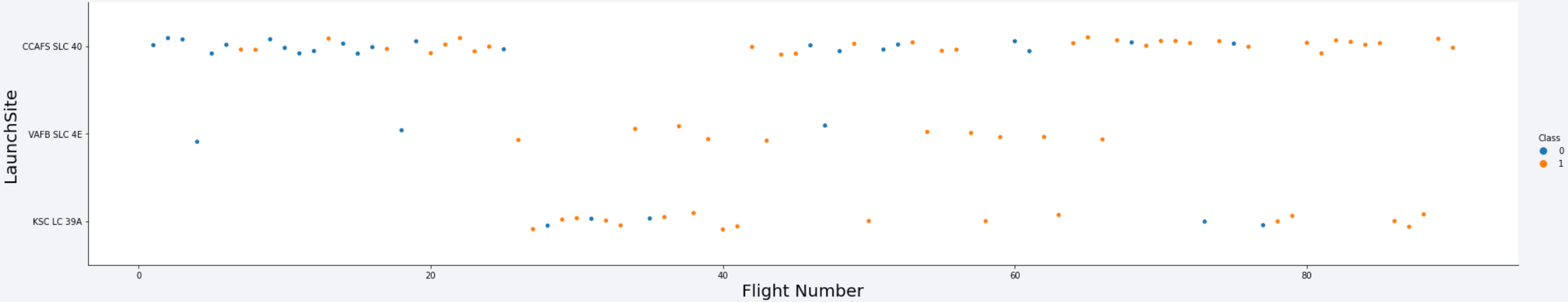
- Based on Predictive Analysis, the Decision Tree Classifier emerged as the most effective model for predicting successful landings, boasting an accuracy exceeding 87% and achieving high accuracy on test data.





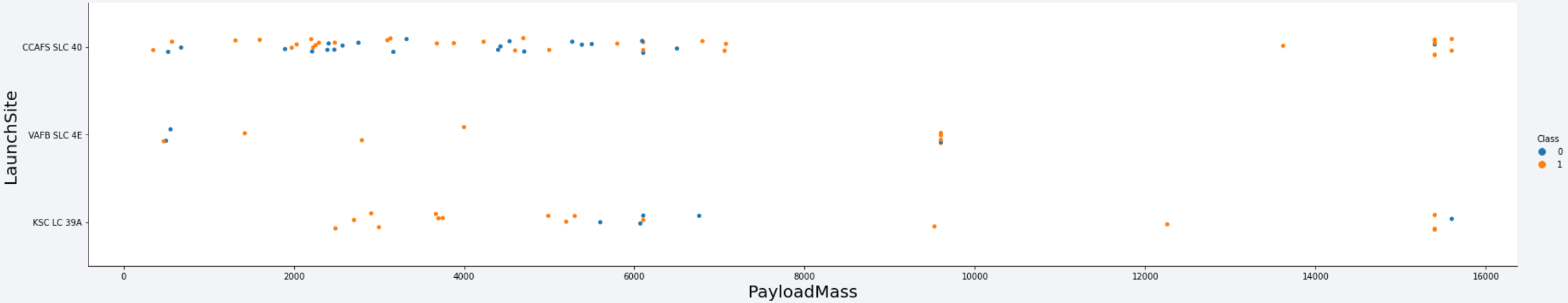

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Analysis of the plot indicates that CCAF5 SLC 40 emerges as the top-performing launch site currently, with a significant number of recent successful launches.
- Following closely in second and third place are VAFB SLC 4E and KSC LC 39A, respectively.
- Additionally, there's a clear trend of improvement in the overall success rate over time.

Payload vs. Launch Site



- Payloads exceeding 9,000kg (approximately the weight of a school bus) exhibit an outstanding success rate.
- Payloads surpassing 12,000kg appear feasible only at CCAFS SLC 40 and KSC LC 39A launch sites.

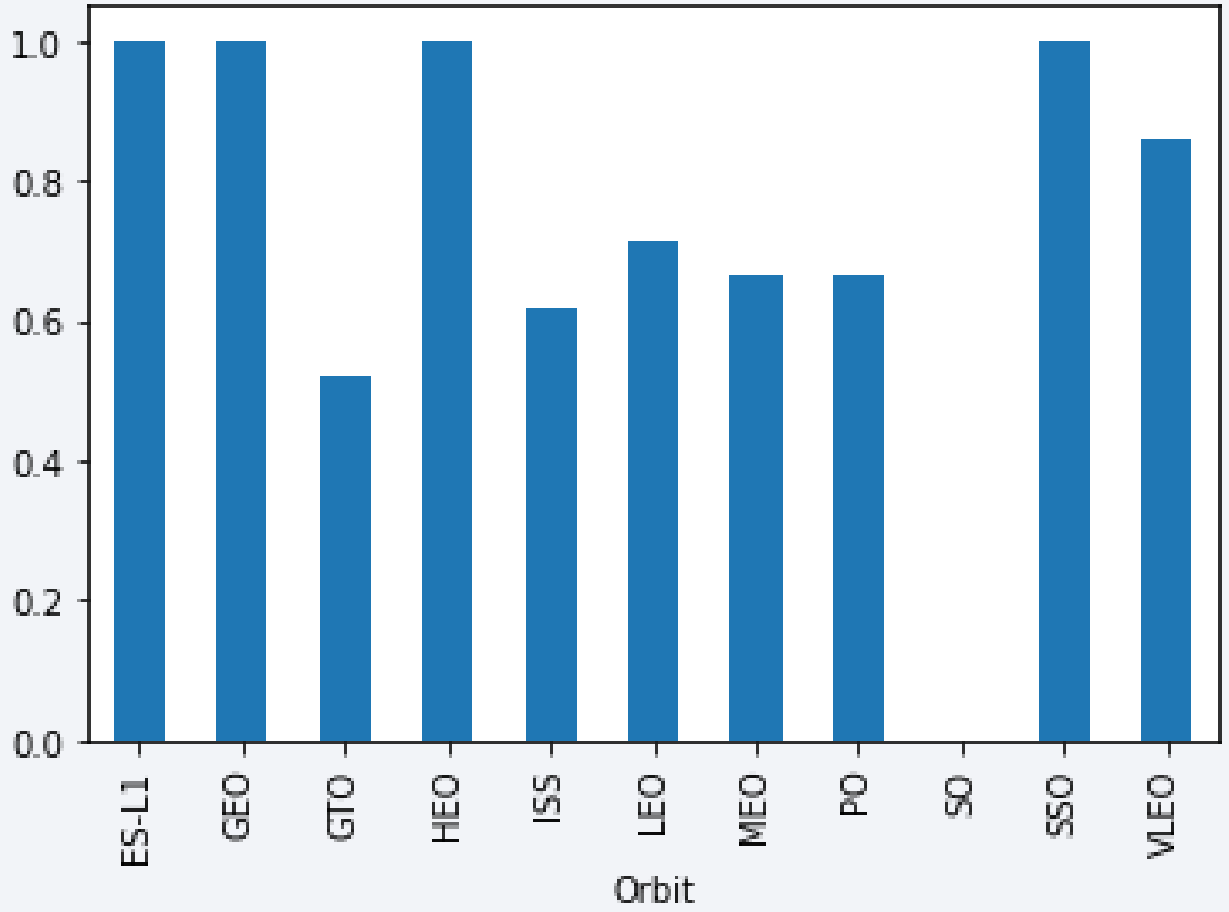
Success Rate vs. Orbit Type

The highest success rates are observed in the following orbits:

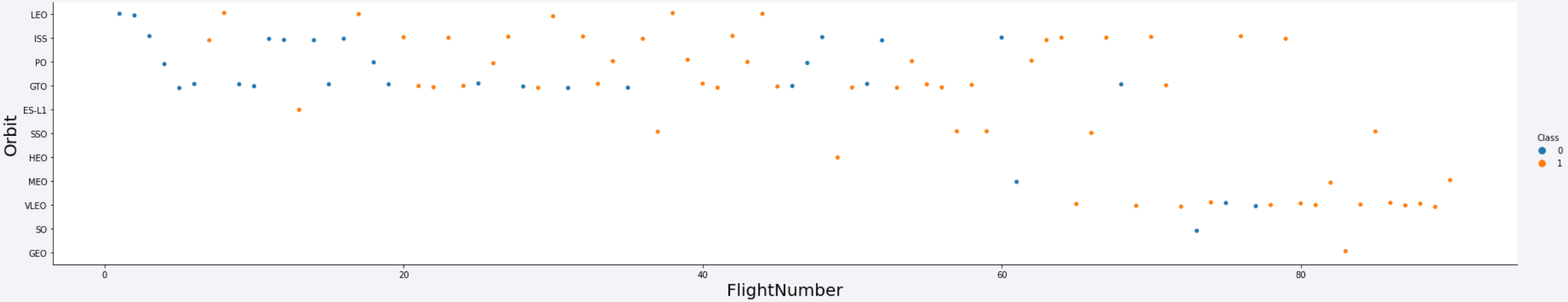
- **ES-L1**
- **GEO**
- **HEO**
- **SSO**

This is followed by:

- VLEO (over **80%** success rate)
- LFO (over **70%** success rate)

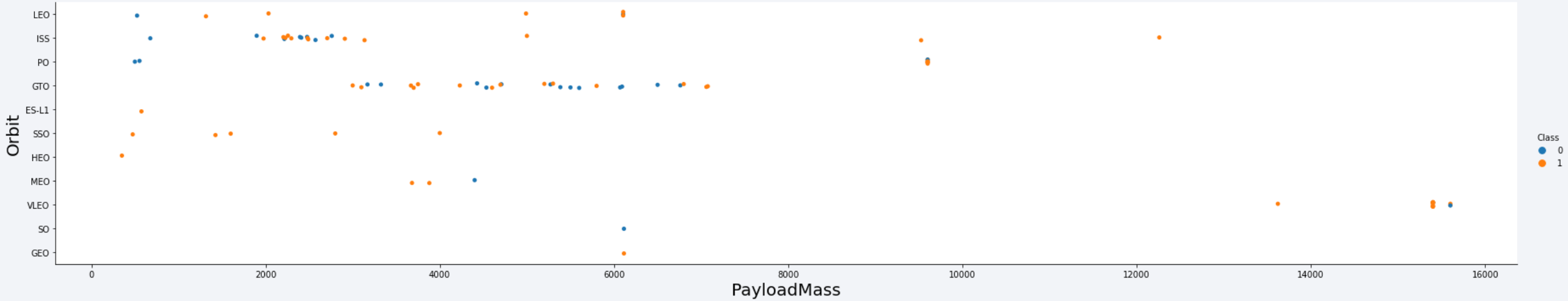


Flight Number vs. Orbit Type



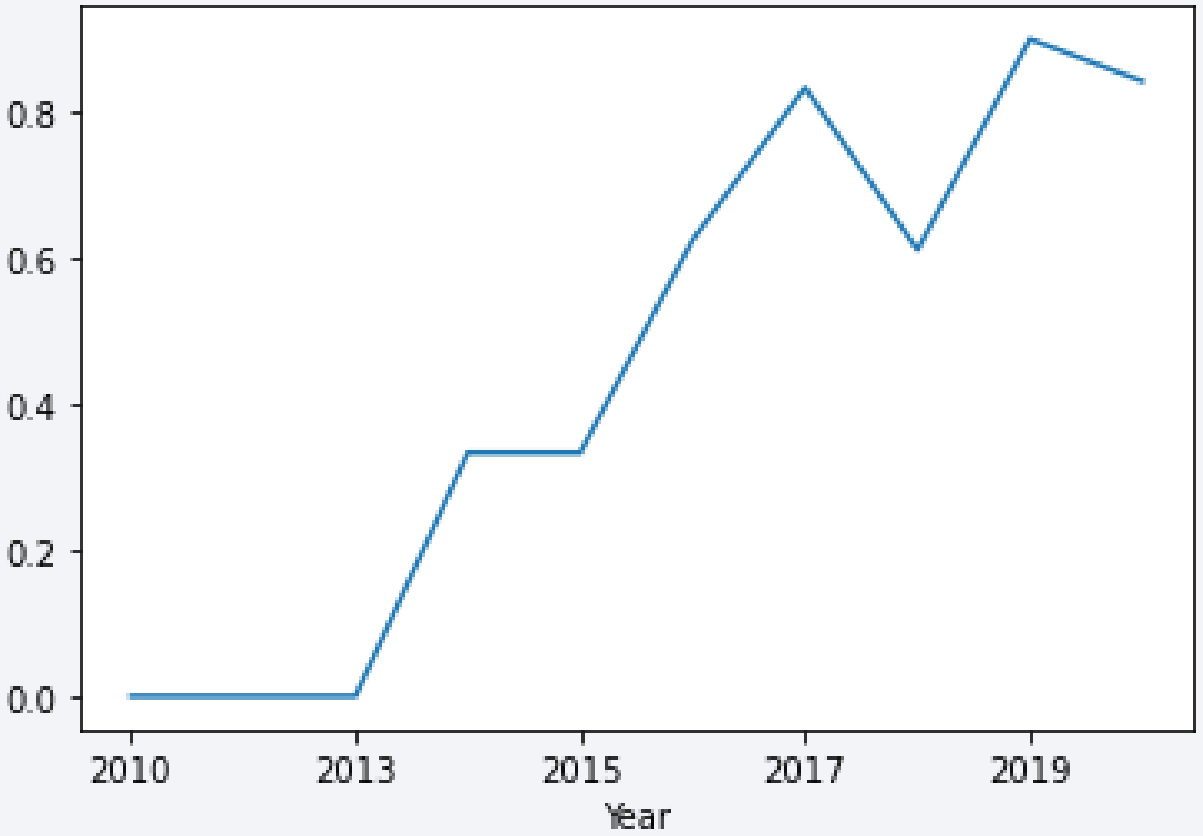
- It appears that the success rate has improved over time for all orbits.
- The VLEO orbit presents a new business opportunity, as its frequency has recently increased.

Payload vs. Orbit Type



- There seems to be no correlation between payload and success rate for the GTO orbit.
- The ISS orbit exhibits the widest range of payloads and a favorable success rate.
- There are relatively few launches to the SO and GEO orbits.

Launch Success Yearly Trend



- The success rate began to increase in 2013 and continued until 2020.
- It appears that the initial three years were characterized by adjustments and technological improvements.

All Launch Site Names

Data indicates the presence of four launch sites:

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

These launch sites were identified by selecting unique occurrences of "launch_site" values from the dataset.

Launch Site Names Begin with 'CCA'

In [9]:
```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[9]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- Here we can see the first five samples of Cape Canaveral launches

Total Payload Mass

- Total payload carried by boosters from NASA:

```
In [10]: sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
Out[10]: total_payload
         111268
```

- The total payload was calculated by summing all payloads associated with codes containing 'CRS', which corresponds to NASA.

Average Payload Mass by F9 v1.1

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

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

In [11]:
```
sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[11]:

avg_payload
2928

- By filtering the data based on the booster version and calculating the average payload mass, we obtained a value of 2,928 kg.

First Successful Ground Landing Date

- First successful landing outcome on ground pad was:

```
In [13]: sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)';
 * ibm_db_sa://fvp19040:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
Out[13]: first_success_gp
         2015-12-22
```

- By filtering the data based on successful landing outcomes on the ground pad and obtaining the minimum date value, we identified the first occurrence, which took place on 12/22/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters that achieved successful landings on drone ships and carried payload masses ranging between 4000 and 6000 units:

Out[14]:

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

- Upon applying the filters outlined above, we conducted an analysis to select distinct booster versions. Following this process, we identified four unique booster versions that met the specified criteria.

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

mission_outcome	qty
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Utilizing grouping techniques to categorize mission outcomes and tallying the records for each group, we derived the summarized results presented above.

Boosters Carried Maximum Payload

- We can see the names of the booster which have carried the maximum payload mass

  - **F9 B5 B1048.4**
  - **F9 B5 B1048.5**
  - **F9 B5 B1049.4**
  - **F9 B5 B1049.5**
  - **F9 B5 B1049.7**
  - **F9 B5 B1051.3**
  - **F9 B5 B1051.4**
  - **F9 B5 B1051.6**
  - **F9 B5 B1056.4**
  - **F9 B5 B1058.3**
  - **F9 B5 B1060.2**
  - **F9 B5 B1060.3**

- The above boosters have carried the maximum payload mass recorded in the dataset.

2015 Launch Records

- Failed landing outcomes in drone ship with their booster versions, and launch site names for in year 2015:

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- The following boosters represent those that have carried the maximum payload mass as documented in the dataset.

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

- Ranking 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, in descending order

- This data view emphasizes the importance of considering cases labeled as "No attempt."

landing__outcome	qty
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

Section 3

Launch Sites Proximities Analysis

All Launch Sites

- Launch sites are situated close to the sea, likely for safety reasons, while also maintaining proximity to roads and railroads.

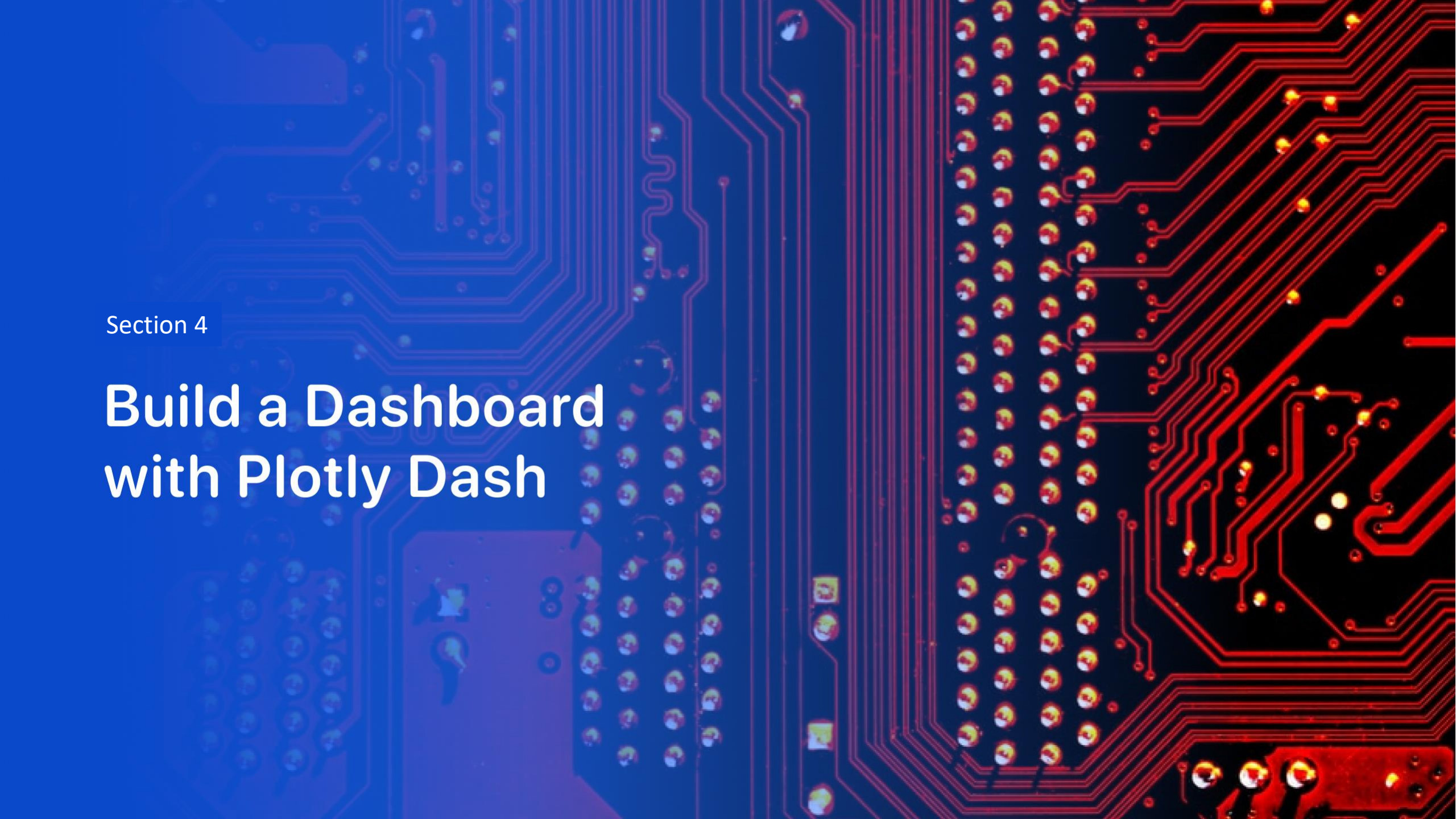
Launch Outcomes by Site

- Example of KSC LC-39A launch site launch outcomes

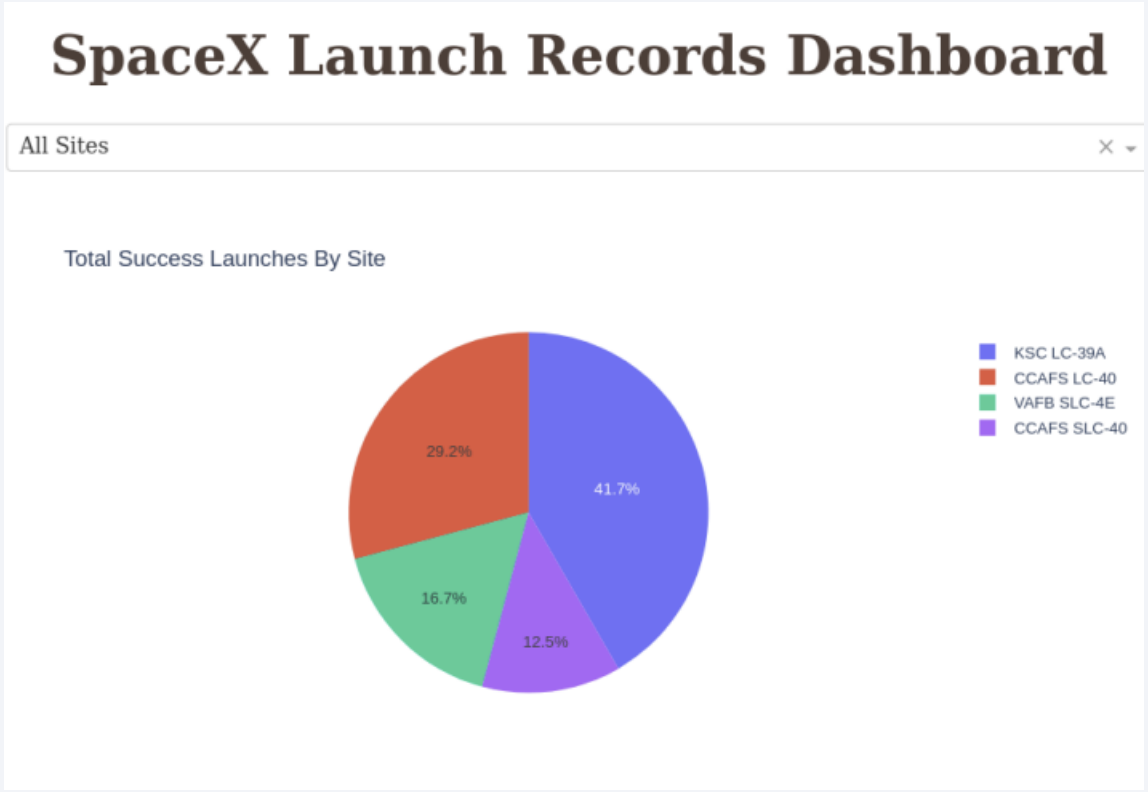
- Green markers denote successful outcomes, while red ones indicate failure.

Logistics and safety considerations.

- Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.

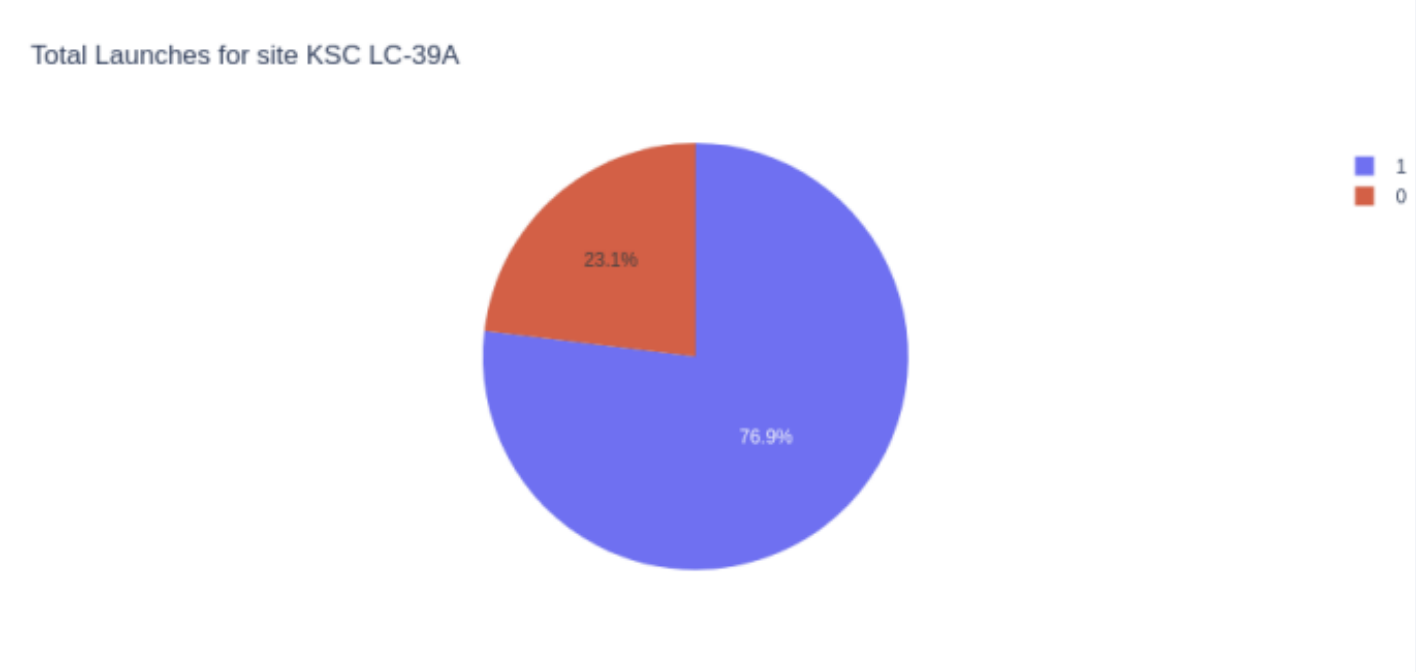

Section 4
Build a Dashboard with Plotly Dash

Successful launches categorized by site.



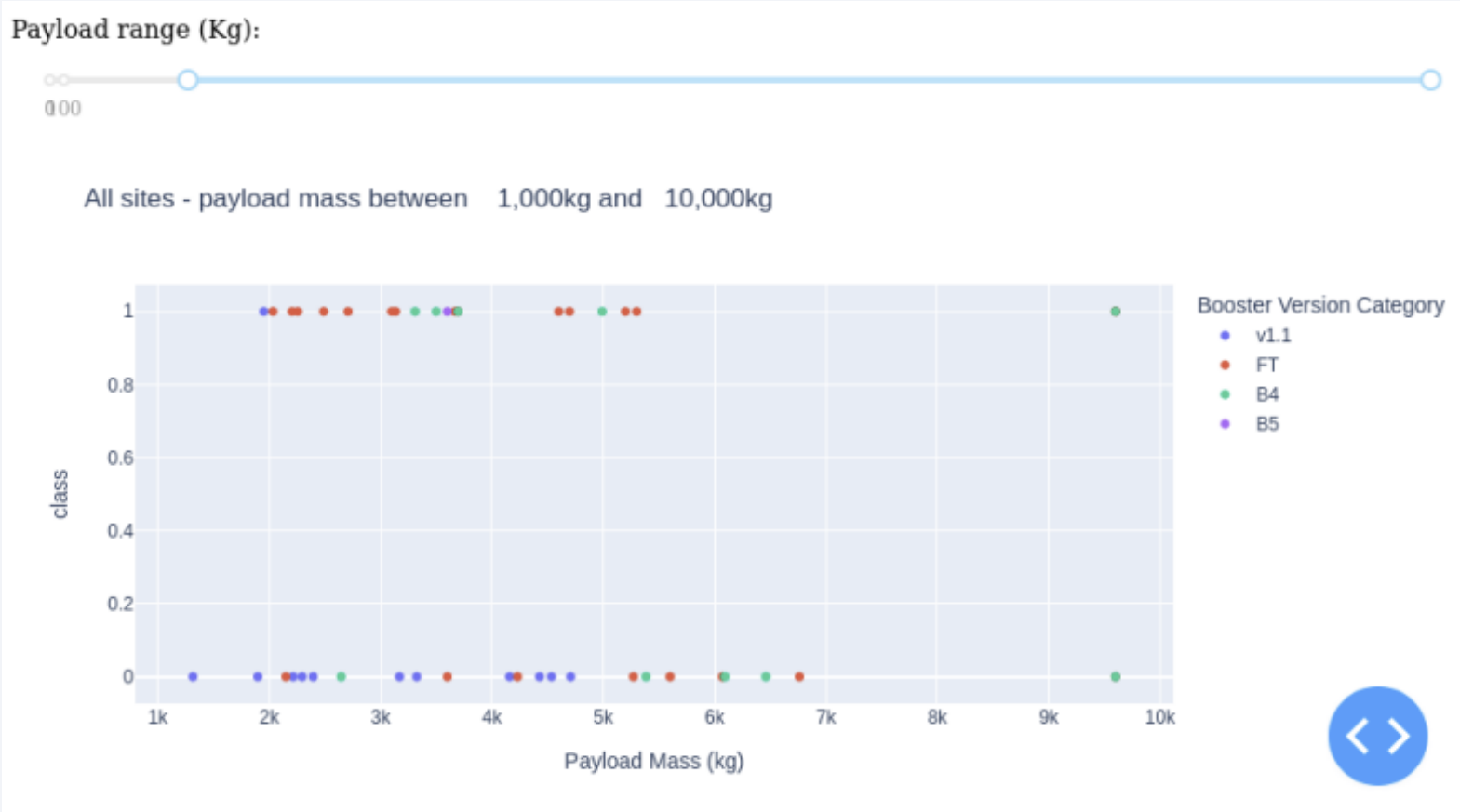
- Launch site selection strongly influences mission success rates, with factors such as geographic location, infrastructure, and safety protocols playing crucial roles.

Success rate of launches from KSC LC-39A.



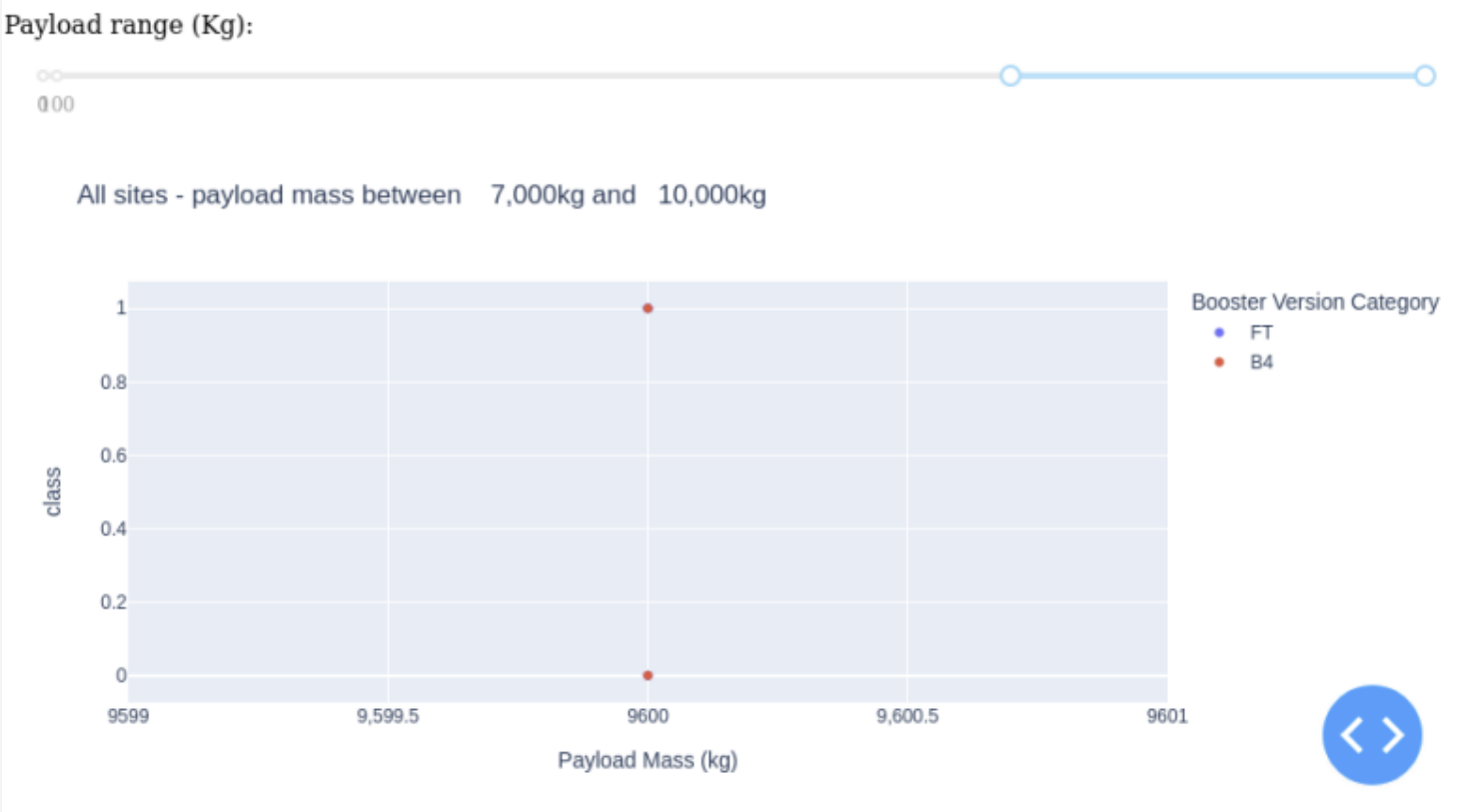
- At this launch site, 76.9% of launches have been successful, indicating a strong track record of mission accomplishments.

Payload vs. Launch Outcome



- Payloads under 6,000kg and FT boosters demonstrate the highest success rate.

Payload vs. Launch Outcome



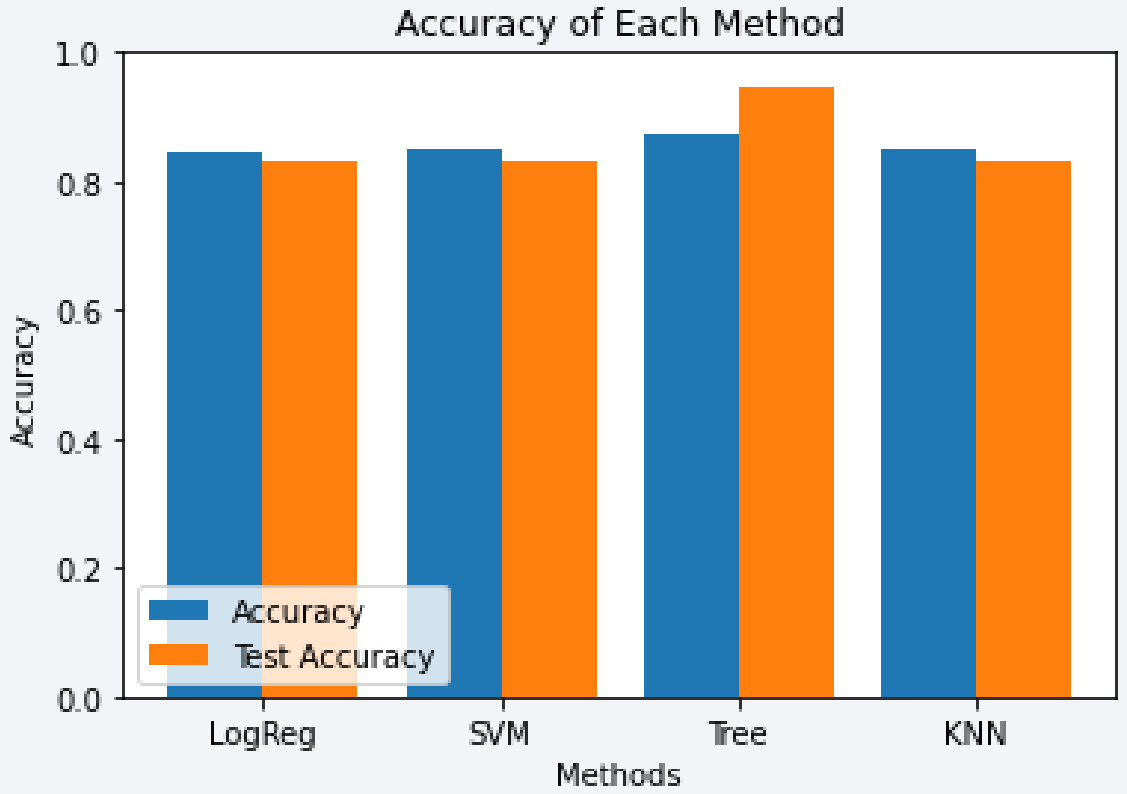
- Due to a lack of sufficient data, it is challenging to accurately estimate the risk associated with launches exceeding 7,000kg.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Four classification models underwent testing, and their respective accuracies are displayed on the graph
- The Decision Tree Classifier emerged as the top-performing model, boasting classification accuracies exceeding 87%.



Confusion Matrix

- The confusion matrix for the Decision Tree Classifier validates its accuracy, highlighting significant numbers of true positives and true negatives in comparison to false instances.

Conclusions

- Through an analysis of various data sources, including refined conclusions, it is evident that KSC LC-39A emerges as the optimal launch site.
- Launches exceeding 7,000kg appear to pose relatively lower risk, offering promising opportunities for payload expansion.
- Despite the overall success of mission outcomes, there is a notable trend of improving successful landing outcomes over time, indicative of advancements in processes and rocket technology.
- Leveraging the Decision Tree Classifier for predicting successful landings presents an avenue for enhancing profitability in space missions.

Appendix

- SpaceX API (JSON): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
- Wikipedia (Webpage): https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- SpaceX (CSV): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321ENSkillsNetwork/labs/module_2/data/Spacex.csv?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetworkChannel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2022-01-01
- Launch Geo (CSV): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_geo.csv
- Launch Dash (CSV): https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_launch_dash.csv

Thank you!