IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jayant Arsode
15th February 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

  Space X promotes Falcon 9 rocket launches on its website, boasting a price tag of $62 million per launch, significantly lower than the upwards of $165 million charged by other providers. This substantial cost reduction is primarily attributed to Space X's ability to recycle the first stage of the rocket. Consequently, by accurately predicting whether the first stage will achieve successful landing, we can determine the overall cost of a launch. This insight becomes crucial for potential competitors seeking to bid against Space X for rocket launch contracts. The objective of this project is to develop a machine learning pipeline capable of forecasting the likelihood of a successful first stage landing.

- Problems you want to find answers

  - Determining the factors that influence the successful landing of a rocket.
  - Investigating the interplay between different variables that impact the likelihood of a successful landing.
  - Identifying the operational conditions necessary to guarantee a reliable landing program

Section 1

Methodology

Methodology

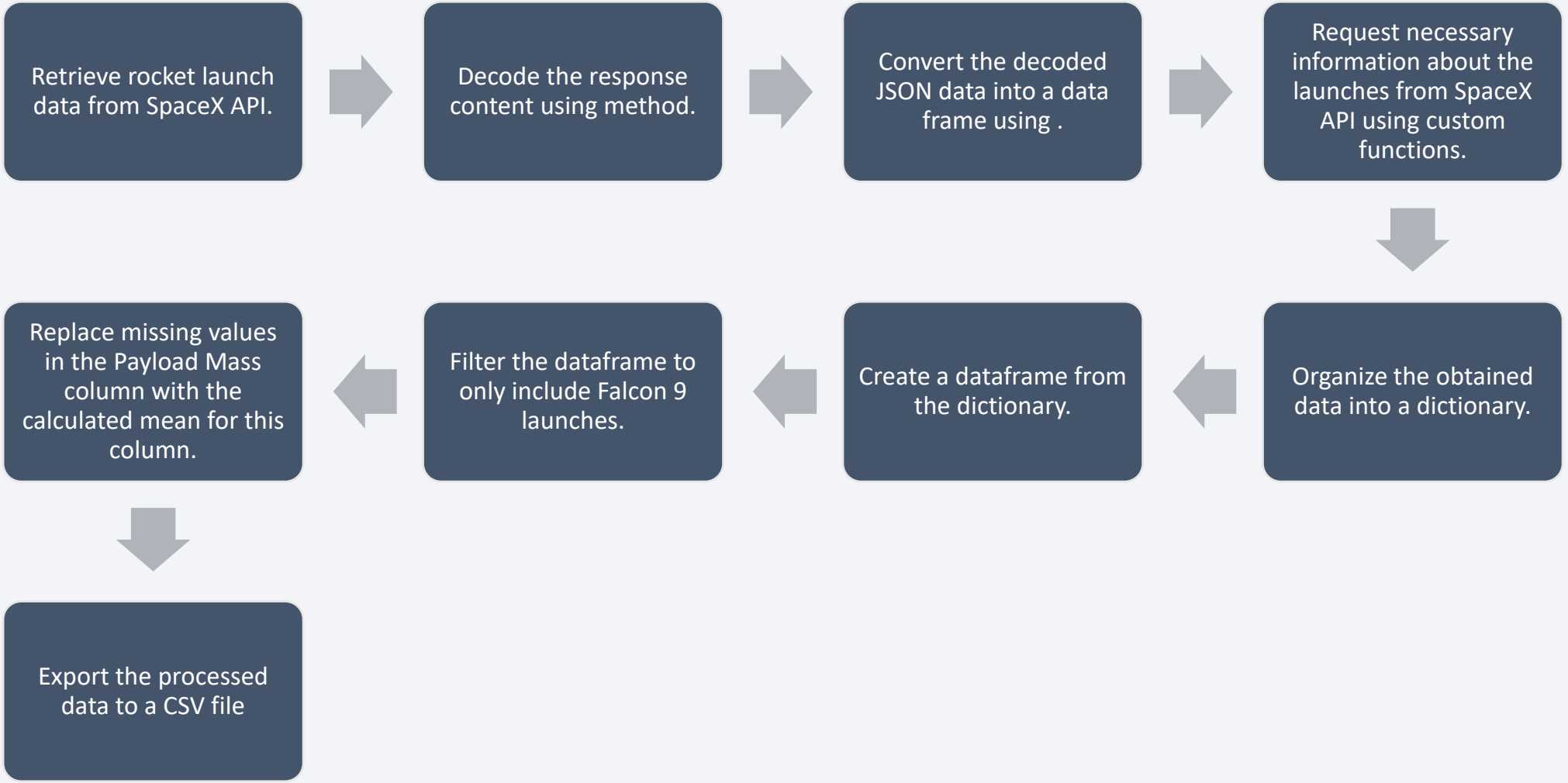
Executive Summary

- Data collection methodology:
  - Utilizing the SpaceX REST API for retrieving real-time data on rocket launches, landings, and other relevant information.
  - Employing web scraping techniques to extract data from Wikipedia pages related to space exploration, rocket technology, and other pertinent topics.
- Perform data wrangling
  - Filtering the dataset to include only relevant features and instances for analysis.
  - Handling missing values in the dataset through imputation, deletion, or other appropriate methods.
  - Employing One Hot Encoding to transform categorical variables into binary format, facilitating binary classification tasks in machine learning algorithms.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Developing, fine-tuning, and evaluating classification models such as Logistic Regression, Random Forest, Support Vector Machines, and Neural Networks to optimize predictive performance.

Data Collection

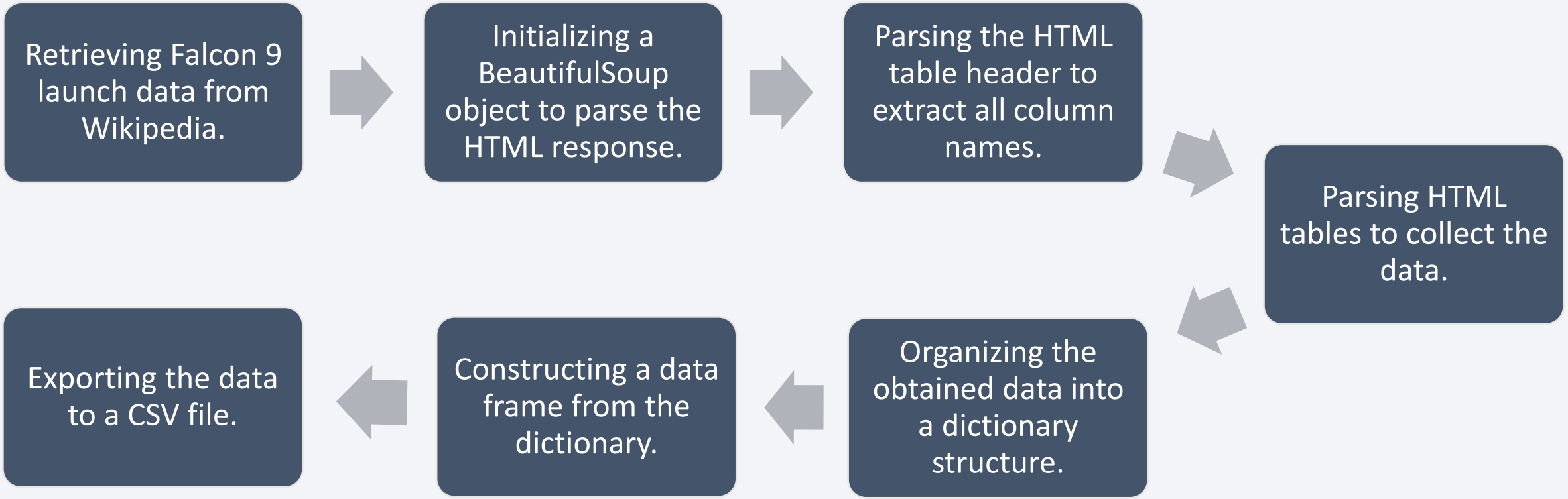
- The data collection process involved a combination of API requests from SpaceX REST API and web scraping from a table in SpaceX's Wikipedia entry.
- Both data collection methods were necessary to gather complete information about the launches for a more detailed analysis.
- Data columns obtained from SpaceX REST API include FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.
- Data columns obtained from Wikipedia web scraping include Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.

Data Collection – SpaceX API



GitHub URL: Data Collection – SpaceX API

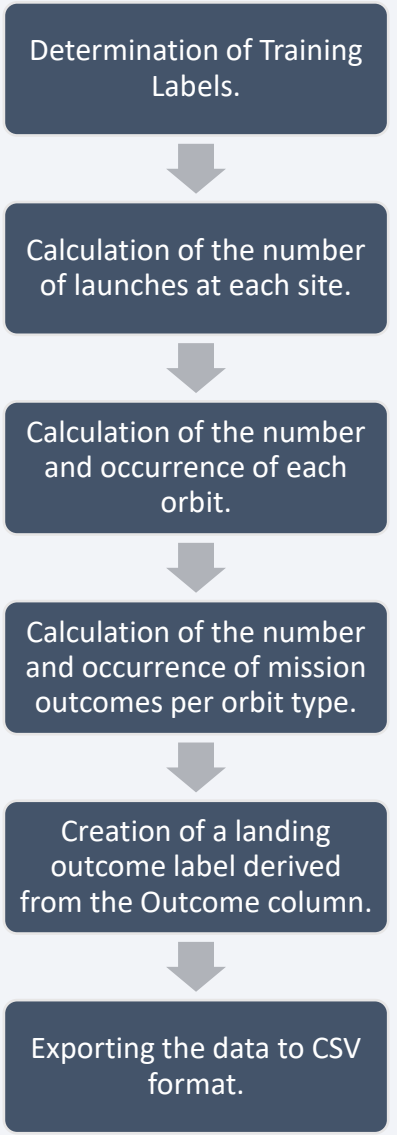
Data Collection - Scraping



GitHub URL: Data Collection - Scraping

Data Wrangling

- The dataset contains instances where the booster failed to land successfully, categorized into different scenarios.
- Examples include cases where a landing attempt was made but failed due to accidents.
- Instances labeled as "True Ocean" indicate successful landings in specific ocean regions, while "False Ocean" denotes unsuccessful landings in those regions.
- "True RTLS" signifies successful landings on ground pads, while "False RTLS" indicates unsuccessful landings on ground pads.
- Similarly, "True ASDS" denotes successful landings on drone ships, while "False ASDS" represents unsuccessful landings on drone ships.
- These outcomes are converted into training labels, where "1" denotes successful booster landings and "0" indicates unsuccessful landings.



GitHub URL: Data Wrangling

EDA with Data Visualization

Charts have been generated to visually represent various relationships and trends within the data:

- Scatter plots illustrate the associations between Flight Number and Payload Mass, Flight Number and Launch Site, as well as Payload Mass and Launch Site. These plots help identify potential correlations that can be utilized in machine learning models.
- Bar charts provide comparisons across discrete categories such as Orbit Type versus Success Rate. They offer insights into the relationships between specific categories and measured values, aiding in the understanding of success rates across different orbit types.
- Line charts depict trends over time, specifically focusing on the Success Rate's yearly trend. These charts allow for the visualization of how success rates have evolved over time, providing valuable insights into the performance of rocket launches.

GitHub URL: EDA with Data Visualization

EDA with SQL

Performed SQL queries for analyzing space mission data:

- Retrieval of unique launch site names from the space mission dataset.
- Display of 5 records where launch sites start with the prefix 'CCA'.
- Calculation of the total payload mass carried by boosters launched by NASA (CRS).
- Determination of the average payload mass carried by booster version F9 v1.1.
- Identification of the date of the first successful landing outcome achieved on a ground pad.
- Listing of booster names that achieved success on a drone ship with payload mass between 4000 and 6000.
- Counting the total number of successful and failed mission outcomes.
- Listing of booster versions that carried the maximum payload mass.
- Identification of failed landing outcomes on drone ships, including booster versions and launch site names, for the months in the year 2015.
- Ranking of landing outcomes (e.g., Failure (drone ship) or Success (ground pad)) between June 4, 2010, and March 20, 2017, in descending order.

GitHub URL: EDA with SQL

Build an Interactive Map with Folium

Markers for Launch Sites:

- Implemented markers with circles, popup labels, and text labels for NASA Johnson Space Center, using its latitude and longitude coordinates as the starting location.
- Incorporated markers with circles, popup labels, and text labels for all launch sites to illustrate their geographical positions and proximity to the Equator and coastlines.

Colored Markers for Launch Outcomes:

- Integrated colored markers, employing a Marker Cluster, to represent success (Green) and failure (Red) outcomes. This aids in identifying launch sites with comparatively high success rates.

Distances between Launch Sites and Proximities:

- Included colored lines to depict distances between the launch site KSC LC-39A (as an example) and its surrounding features such as railways, highways, coastlines, and the closest city. This provides a visual representation of the spatial relationships.

GitHub URL: Interactive Map with Folium

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Implemented a dropdown menu for selecting Launch Sites, facilitating user-friendly site selection.

Pie Chart Displaying Successful Launches:

- Incorporated a pie chart to visualize the total count of successful launches across all sites, with the option to display Success vs. Failed counts for a specific Launch Site when selected.

Payload Mass Range Slider:

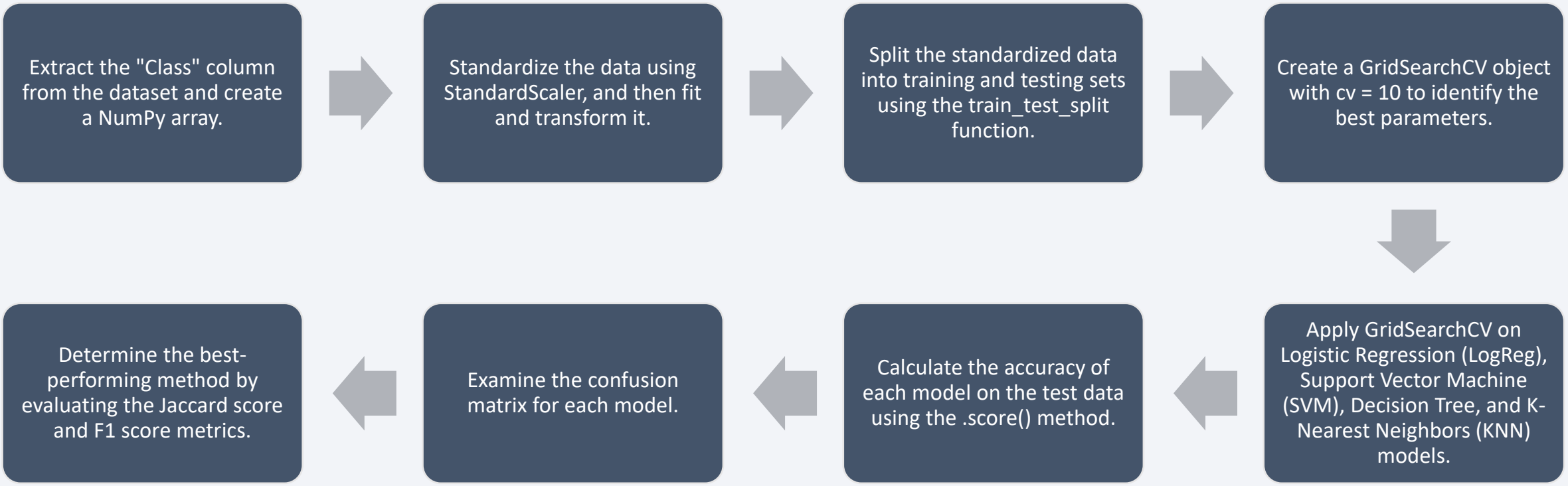
- Introduced a slider component enabling users to select a desired Payload mass range, enhancing flexibility in data filtering.

Scatter Chart of Payload Mass vs. Success Rate by Booster Version:

- Developed a scatter chart illustrating the relationship between Payload Mass and Launch Success rates across different Booster Versions, providing insights into the correlation between these variables.

GitHub URL: Interactive Dashboard with Plotly Dash

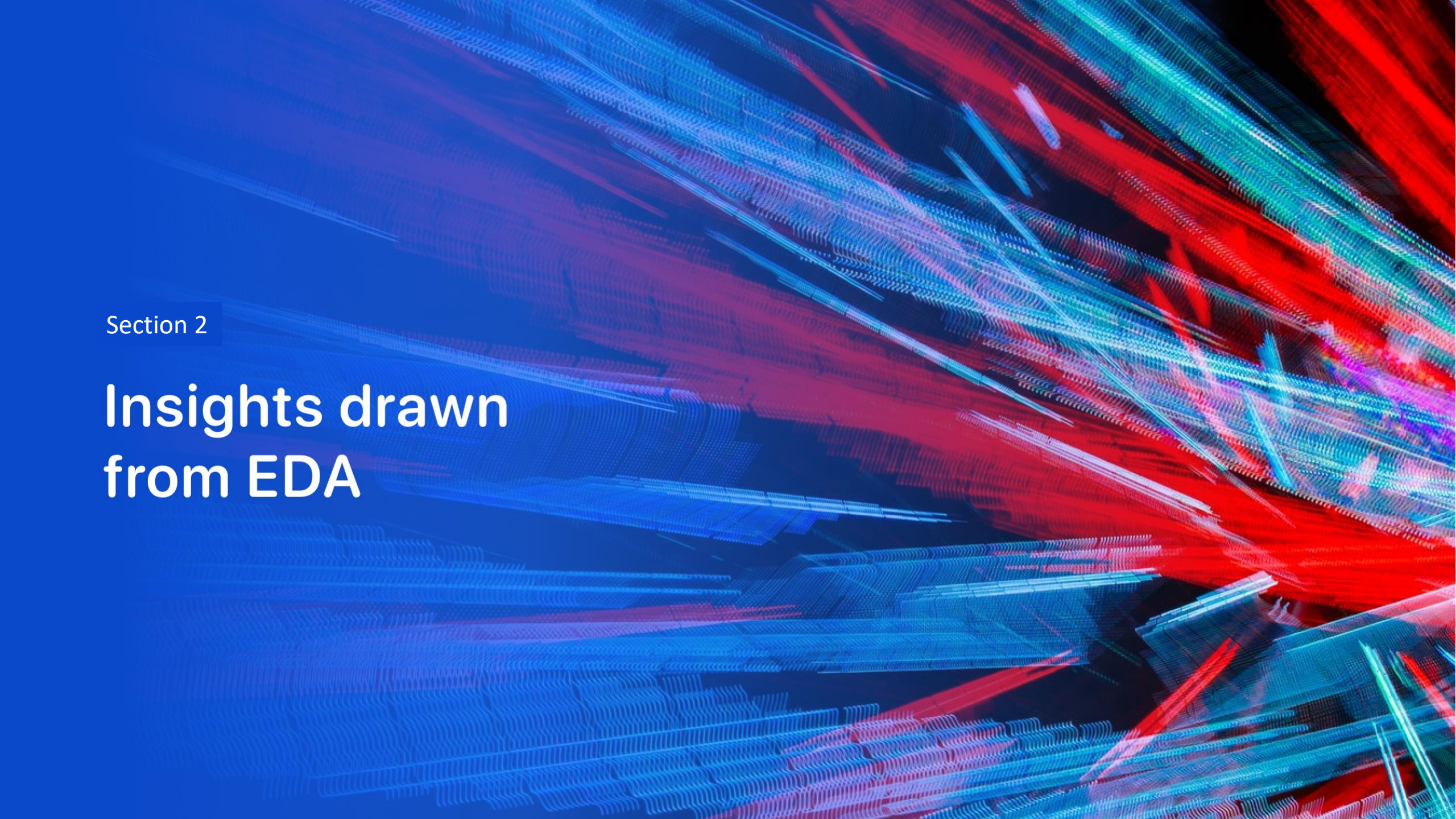
Predictive Analysis (Classification)



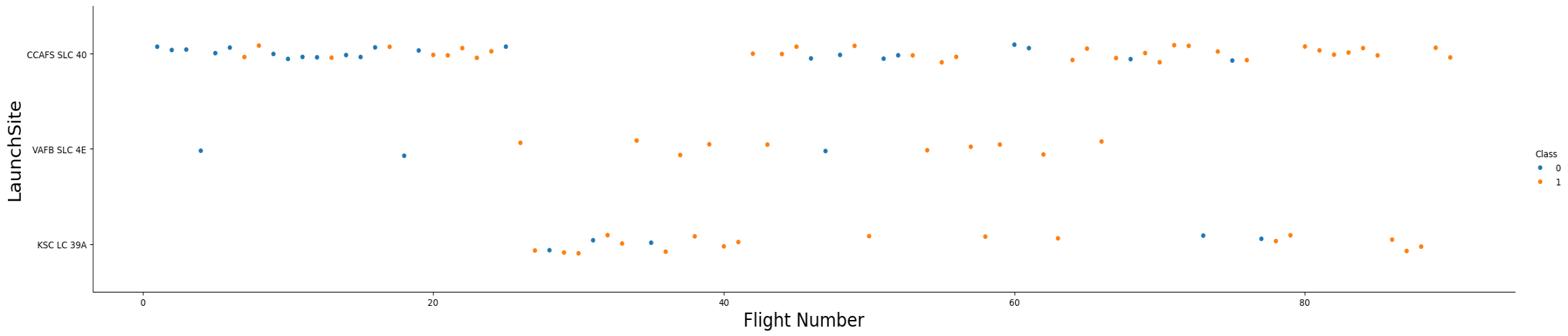
GitHub URL: Predictive Analysis

Results

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


Section 2
Insights drawn from EDA

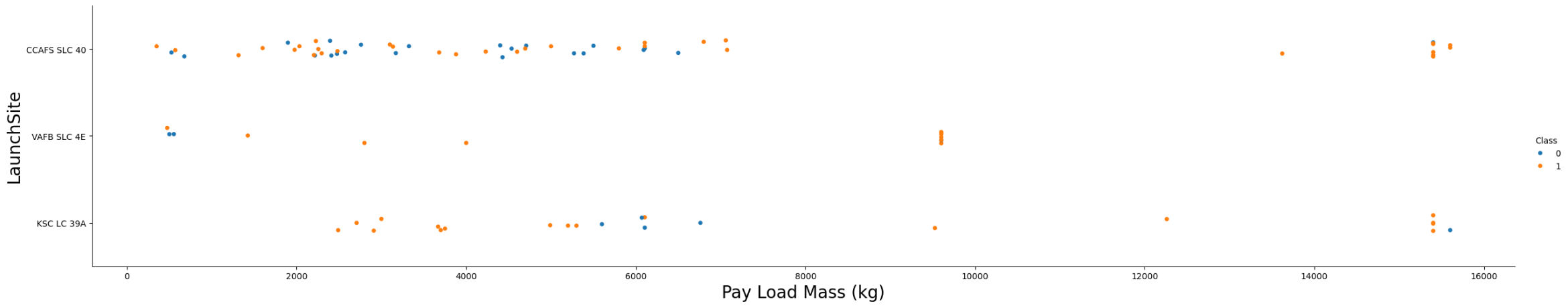
Flight Number vs. Launch Site



Explanation:

- While the initial flights encountered failures, the most recent ones achieved success consistently.
- Approximately fifty percent of all launches occur at the CCAFS SLC 40 launch site.
- Both VAFB SLC 4E and KSC LC 39A exhibit superior success rates.
- It is reasonable to infer that with each new launch, the likelihood of success increases.

Payload vs. Launch Site



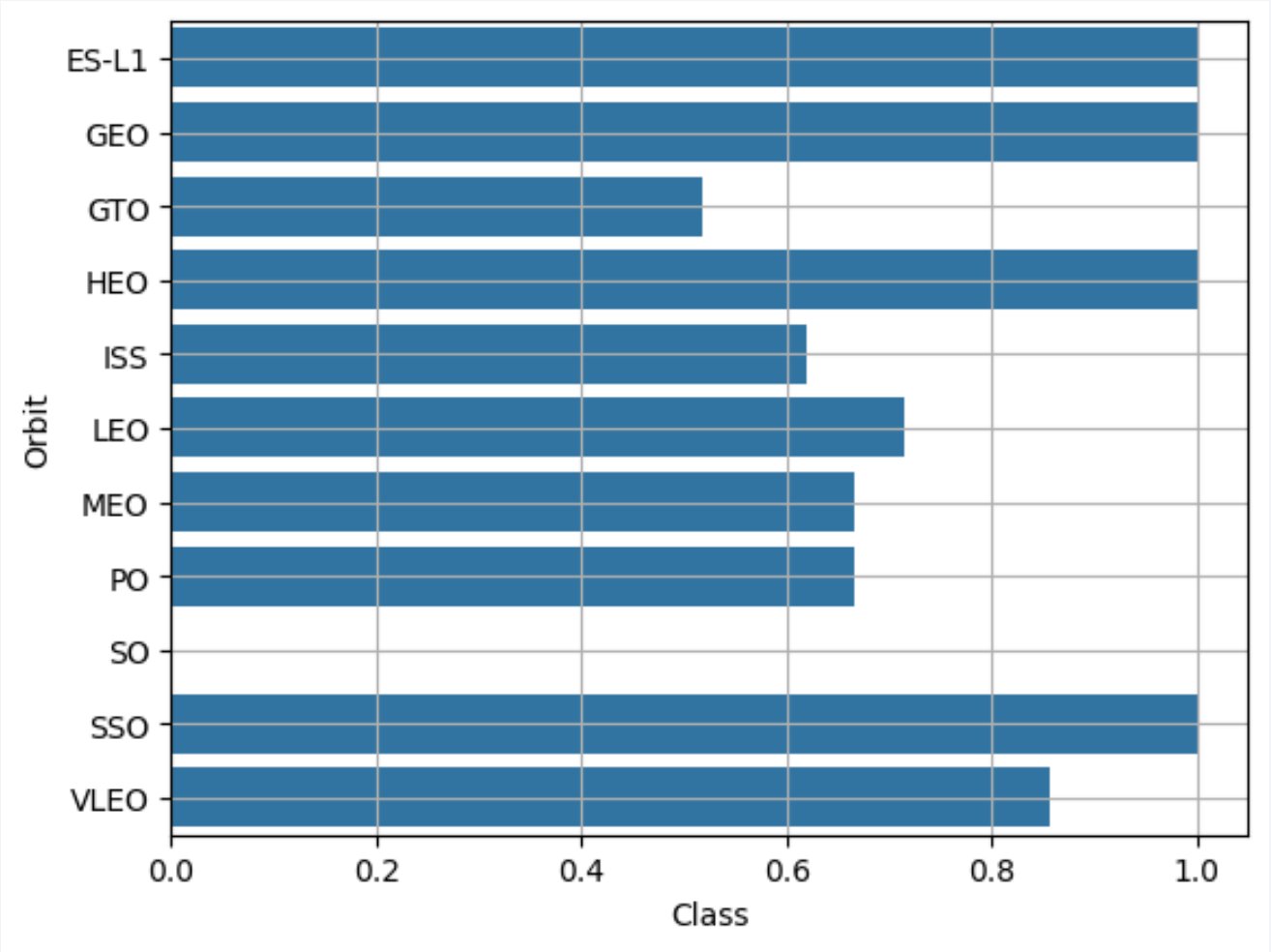
Explanation:

- Greater payload mass correlates with increased success rates across launch sites.
- Successful launches are predominantly observed with payload masses exceeding 7000 kg.
- Additionally, KSC LC 39A exhibits a 100% success rate for payloads under 5500 kg.

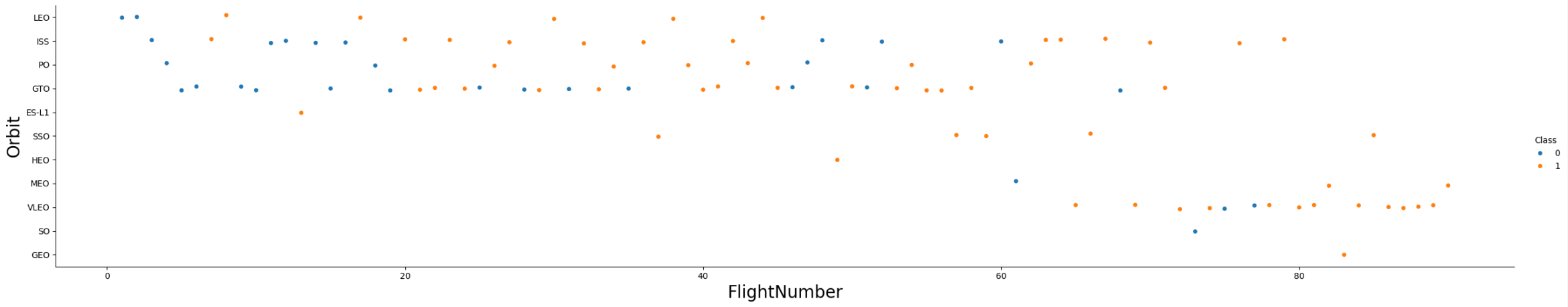
Success Rate vs. Orbit Type

Explanation:

- Orbits with a 100% success rate include ES-L1, GEO, HEO, and SSO.
- Orbits with a 0% success rate comprise SO.
- Orbits with success rates ranging between 50% and 85% consist of GTO, ISS, LEO, MEO, and PO.



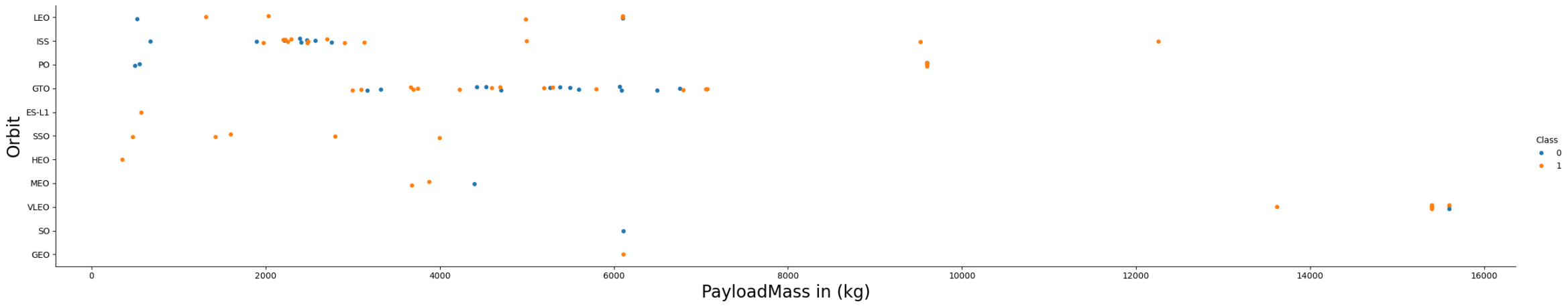
Flight Number vs. Orbit Type



Explanation:

- The correlation between success and the number of flights is evident in the LEO orbit, whereas there is no apparent connection between the flight number and success in the GTO orbit.

Payload Mass (kg) vs. Orbit Type



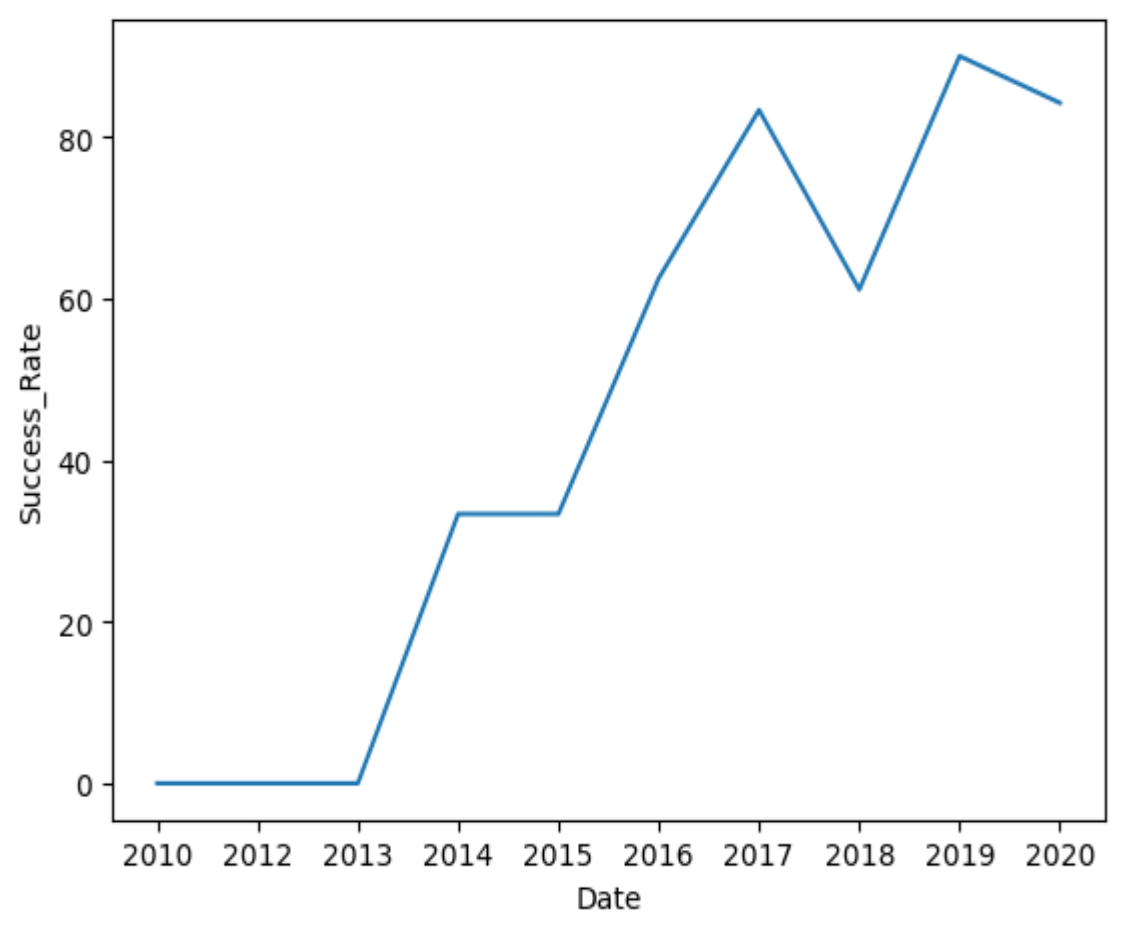
Explanation:

- Heavy payloads exert a detrimental effect on Geostationary Transfer Orbit (GTO) trajectories, while they offer a beneficial impact on GTO and Polar Low Earth Orbit (LEO) paths, such as the International Space Station (ISS).

Launch Success Yearly Trend

Explanation:

- Success rates consistently rose from 2013 to 2020 nut had some some drop between year 2017 and 2018.



All Launch Site Names

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[4]:

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

Explanation:

- Displaying the names of the unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTABLE where "Launch_site" like('CCA%') limit 5;
```

Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Explanation:

- Displaying 5 records where launch sites begin with the string 'CCA'.

Total Payload Mass

```
%sql select sum("PAYLOAD_MASS__KG_") as "total_payload_mass_NASA(CRS)" from SPACEXTABLE where Customer=='NASA (CRS)';
```

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

total_payload_mass_NASA(CRS)
45596

Explanation:

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

Average Payload Mass by F9 v1.1

```
%sql select avg("PAYLOAD_MASS__KG_") as "total_payload_mass_ F9 v1.1" from SPACEXTABLE where "Booster_Version" like 'F9 v1.1%';
```

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

total_payload_mass_ F9 v1.1
2534.6666666666665

Explanation:

- Displaying average payload mass carried by booster version F9 v1.1.
- Also in screenshot total is mass is error it is the average mass.

First Successful Ground Landing Date

```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[8]:

first_successful_landing
2015-12-22

Explanation:

- Listing the date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select distinct("Booster_Version") from SPACEXTABLE where "Landing_Outcome"=='Success (drone ship)' and "PAYLOAD_MASS__KG_"
BETWEEN 4001 and 6000;
```

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

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

Explanation:

- Listing 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

```
%sql select "Mission_Outcome",count("Mission_Outcome") from SPACEXTABLE group by "Mission_Outcome";
```

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

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

Explanation:

- Listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

```
%sql select "Booster_Version" from SPACEXTABLE where "PAYLOAD_MASS__KG_"=(select max("PAYLOAD_MASS__KG_") from SPACEXTABLE);
```

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

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

Explanation:

- Listing the names of the booster versions which have carried the maximum payload mass.

2015 Launch Records

```
%sql select substr("Date", 6,2) as "month","Landing_Outcome","Booster_Version","Launch_site" from SPACEXTABLE where
"Landing_Outcome"=='Failure (drone ship)' and substr("Date",0,5)='2015';
```

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

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

Explanation:

- Listing the failed landing outcomes in drone ship, their booster versions, and launch site names for the months in the year 2015.

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

Explanation:

- Ranking 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.

```
%%sql SELECT "Landing_Outcome", COUNT(*) AS "Landing_Count"
FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_outcome"
ORDER BY "Landing_Count" DESC;
```

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

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

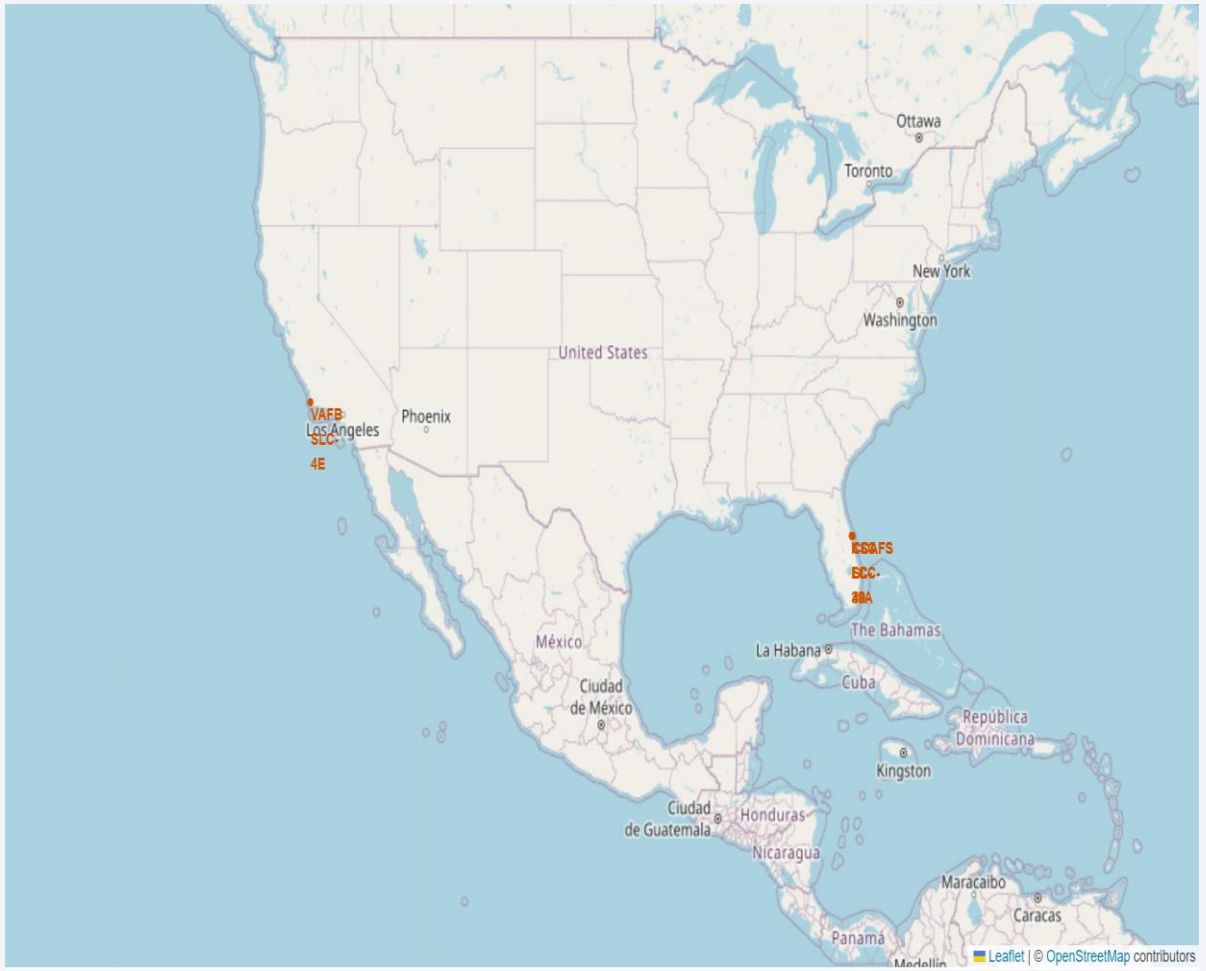
Section 3

Launch Sites Proximities Analysis

Mapping Global Launch Site Locations

Explanation:

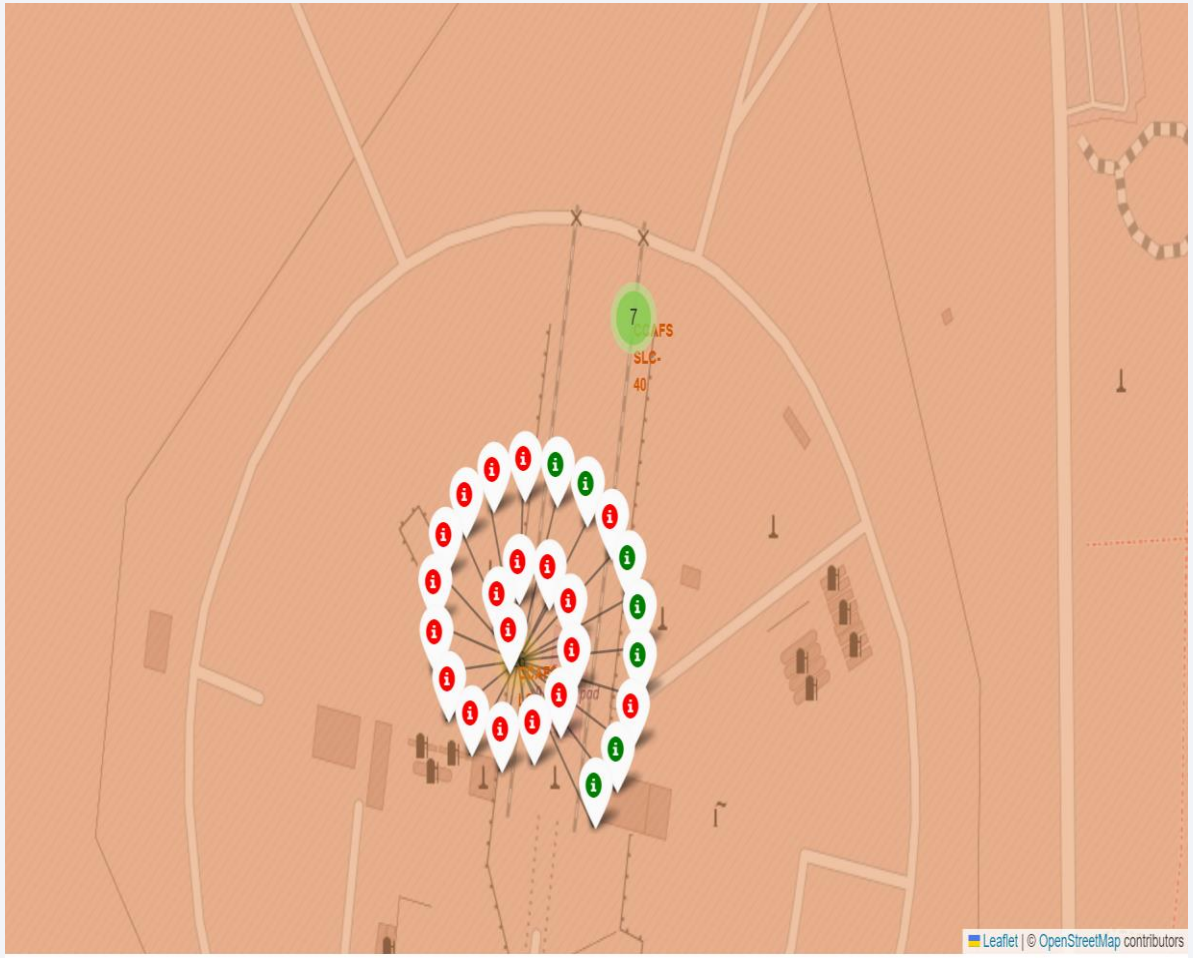
- Launch sites are strategically located near the Equator due to the Earth's faster rotational speed in this region, approximately 1670 km/hour. This inherent velocity aids spacecraft launched from the Equator in maintaining momentum as they enter space, thanks to the principle of inertia.
- Launch facilities are typically situated near coastlines to minimize the potential danger posed by debris or explosions during rocket launches. Launching rockets towards the ocean significantly reduces the risk of harm to populated areas.



Colored Class Labels Of Launch Record On Map

Explanation:

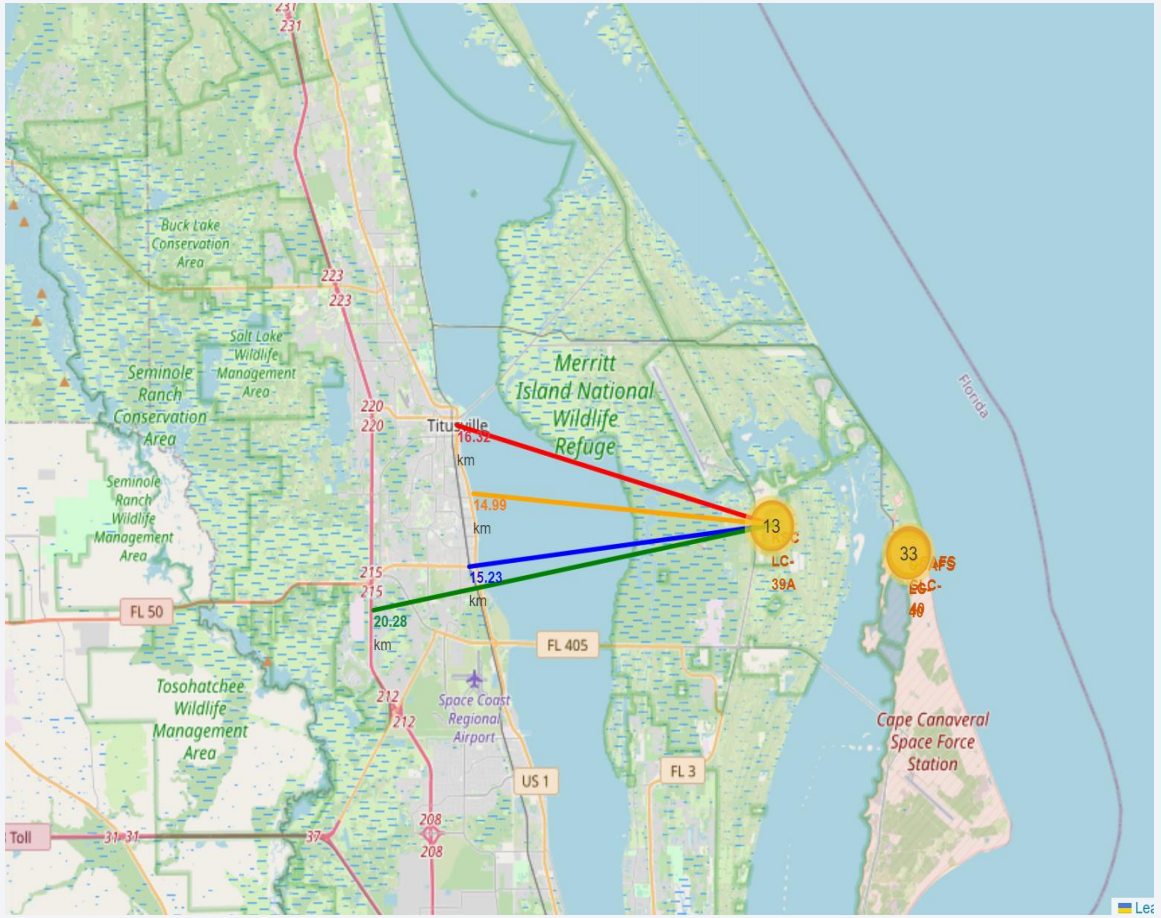
- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
  - Green Marker = Successful Launch
  - Red Marker = Failed Launch
- Launch Site KSC LC-39A has a very high Success Rate

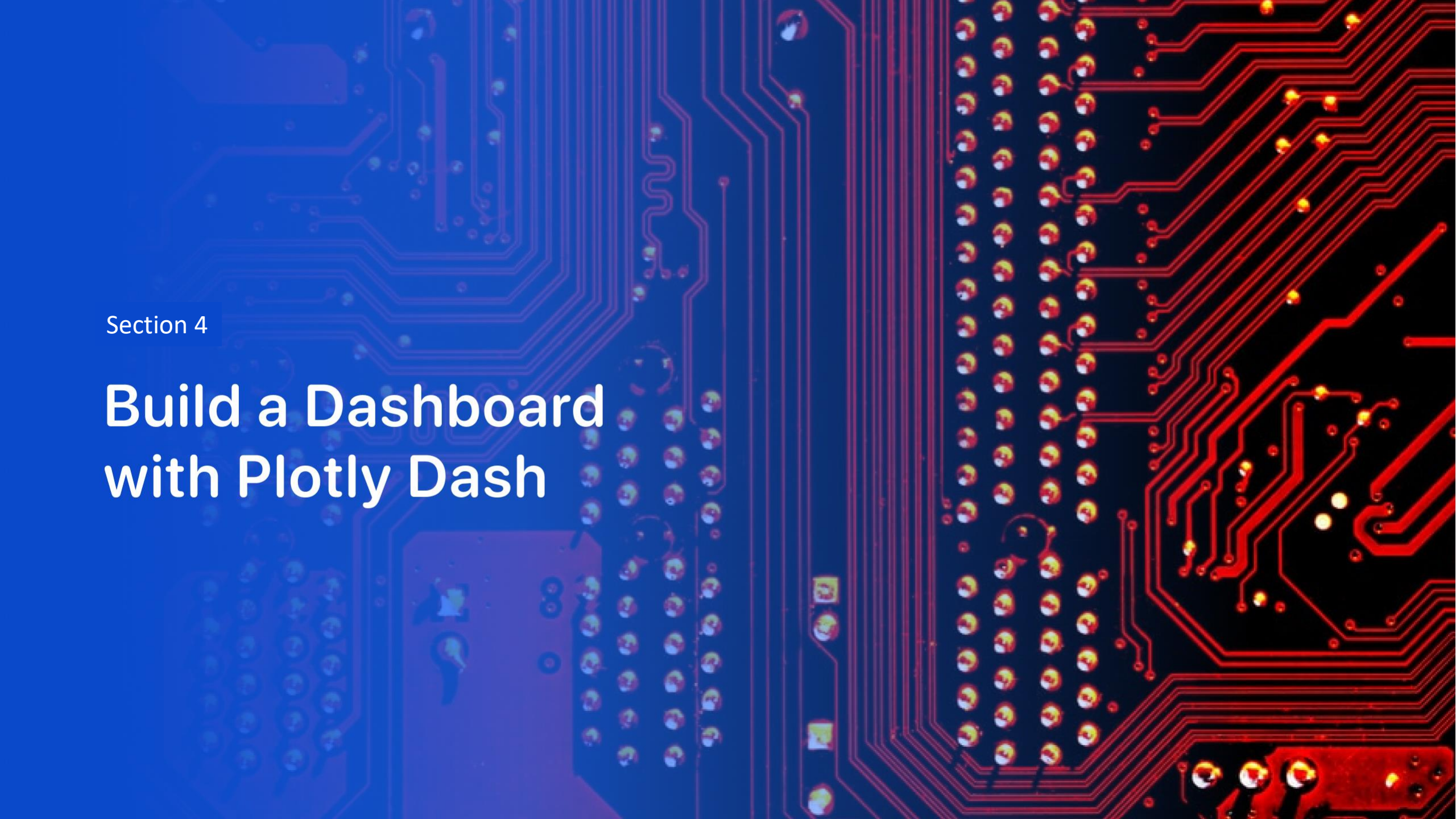


<Folium Map Screenshot 3>

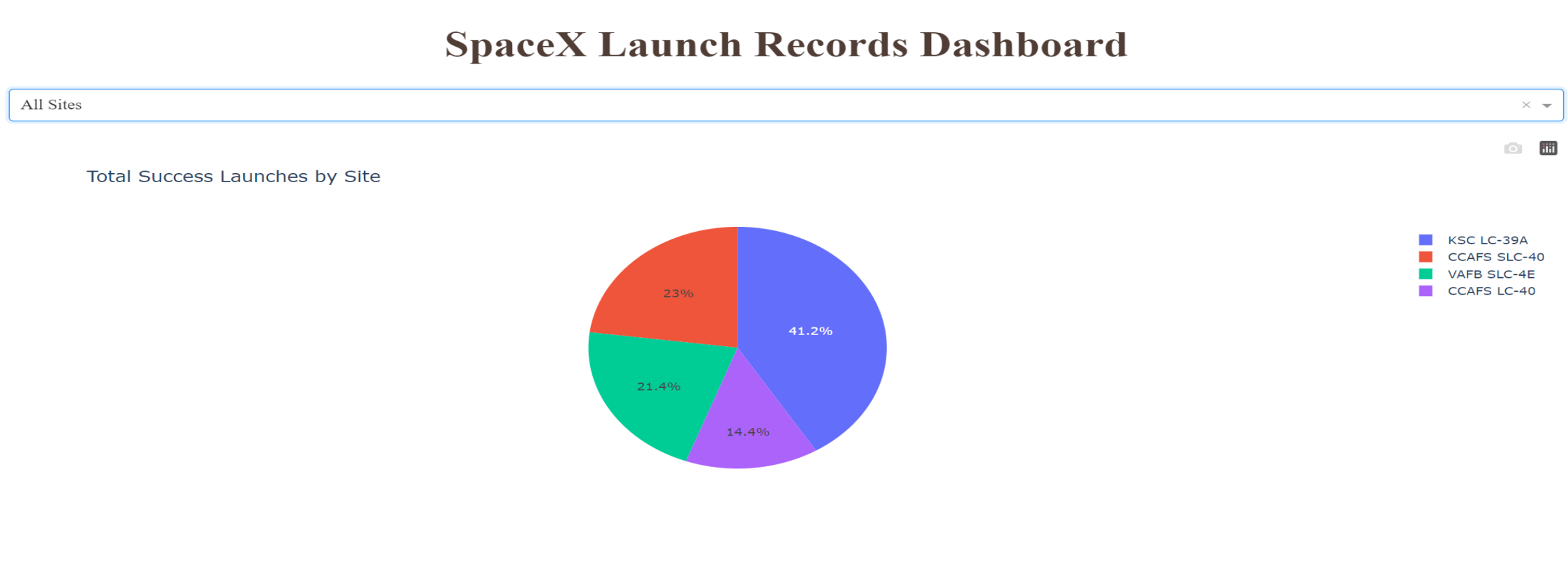
Explanation:

- The visual examination of the KSC LC-39A launch site reveals its proximity to various infrastructural elements:
  - It is situated approximately 15.23 km from a railway line.
  - It is approximately 20.28 km from a major highway.
  - The coastline is relatively nearby, at about 14.99 km from the site.
- Additionally, the launch site, KSC LC-39A, is in close proximity to the city of Titusville, approximately 16.32 km away.
- Given the high speed of a failed rocket, it has the potential to cover distances of 15-20 km within seconds, posing significant risks to populated areas.




Section 4
Build a Dashboard with Plotly Dash

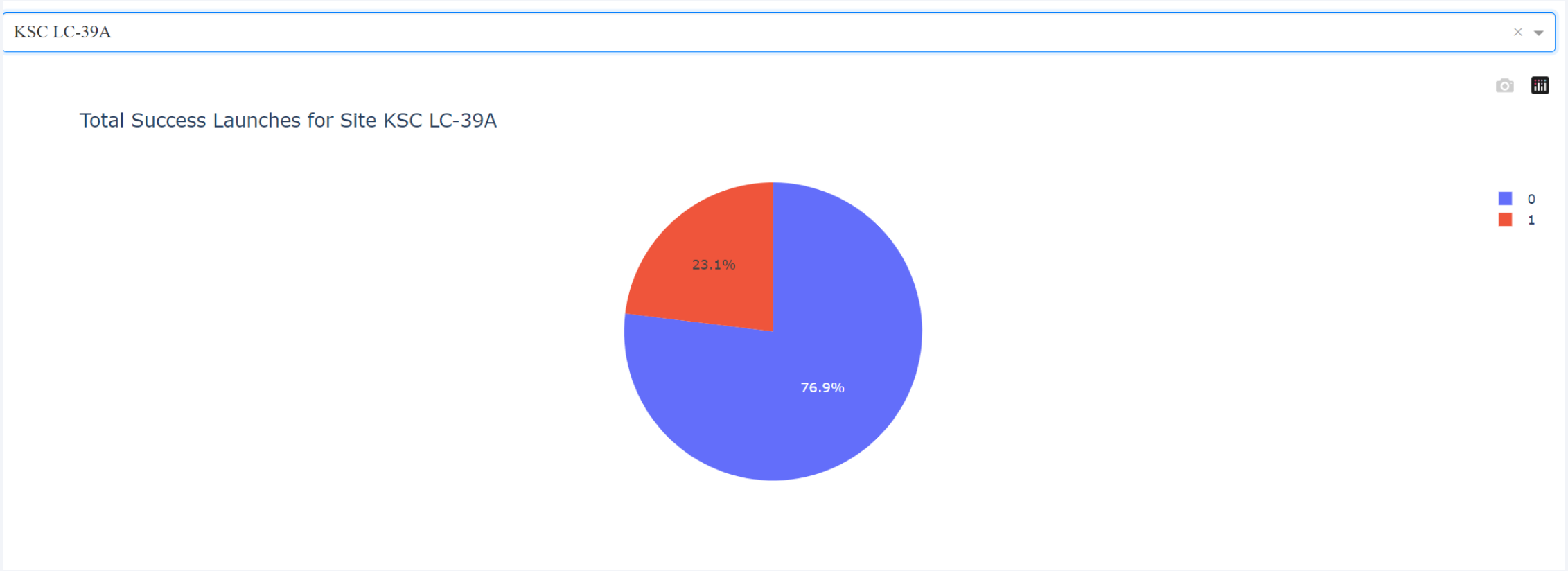
Total Success Launches Distribution By Site



Explanation:

- The data depicted in the chart highlights that KSC LC-39A stands out as the site with the highest number of successful launches among all locations.

Site with the Highest Launch Success Ratio



Explanation:

- KSC LC-39A boasts the highest launch success rate at 76.9%, marked by 10 successful landings and just 3 failures.

Comparative Analysis of Payload Mass and Launch Outcomes Across All Sites

Explanation:

- Payloads ranging from 2000 to 5500 kg exhibit the highest success rates, as indicated by the charts.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Explanation:

- The assessment of the Test Set scores does not conclusively determine the superior method.
- Consistent Test Set scores might stem from the limited sample size (18 samples). Consequently, we evaluated all methods using the entire Dataset.
- Examination of the complete Dataset scores affirms that the Decision Tree Model emerges as the most effective. This model not only exhibits superior scores but also achieves the highest accuracy.

Examining the scores on Test sets

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

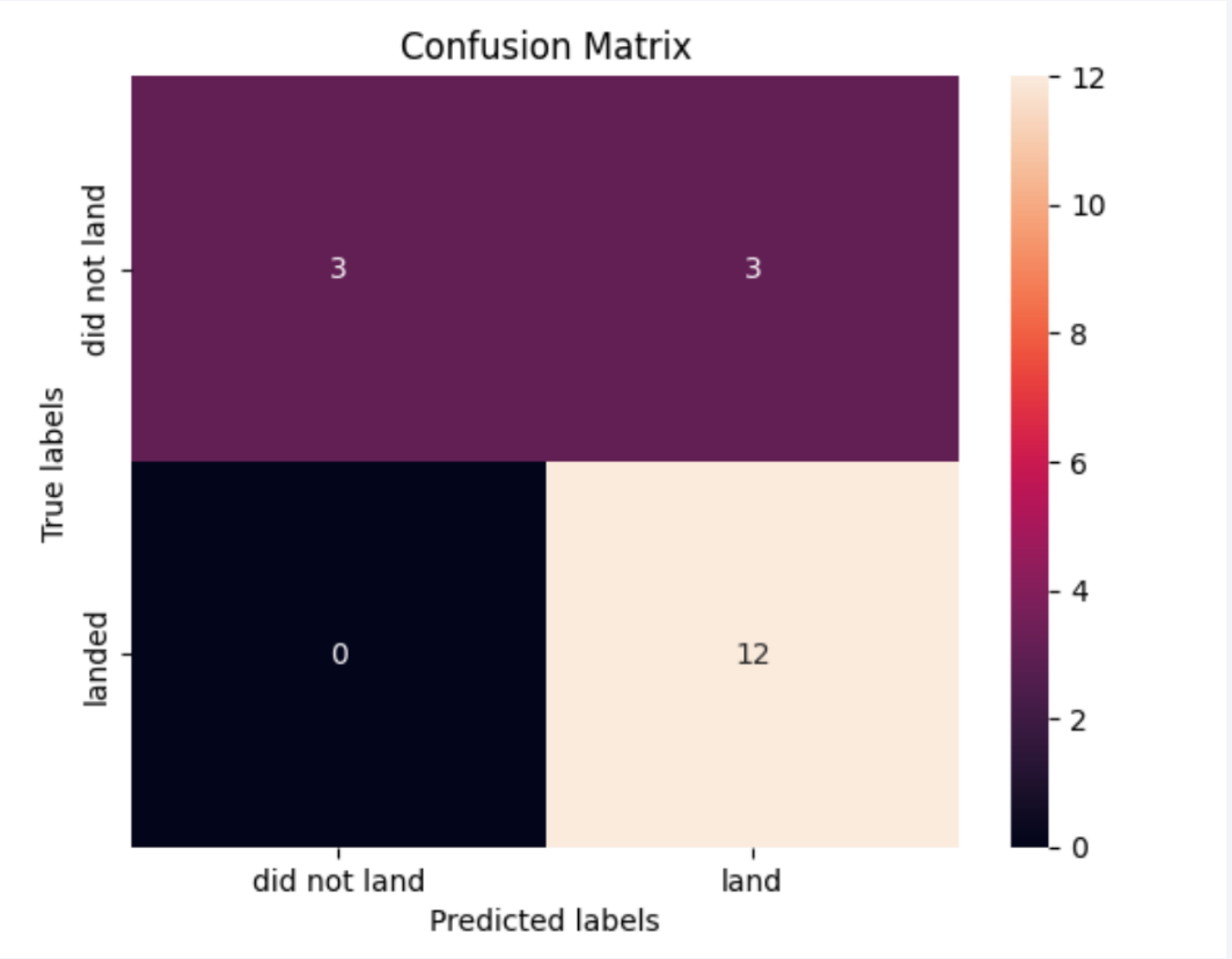
Examining the scores on entire Dataset

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.865672	0.819444
F1_Score	0.909091	0.916031	0.928000	0.900763
Accuracy	0.866667	0.877778	0.900000	0.855556

Confusion Matrix

Explanation:

- We can see that we get same confusion matrix for test data on decision trees, svm and logistic regression model thus we get accuracy of around 83.33%.
- Also only 3 samples are predicted by our model as landed while they do not land.



Conclusions

- The application of a machine learning pipeline enables precise anticipation of favorable outcomes regarding initial stage landings, promising a substantial reduction in launch expenditures and thereby amplifying competitiveness within the space launch sector.
- The developed pipeline furnishes invaluable strategic foresight for prospective contenders vying against SpaceX for rocket launch agreements. This tool empowers them to evaluate the likelihood of prosperous landings and make well-informed judgments during contractual negotiations.
- A positive correlation exists between the volume of flights at a launch site and the corresponding success rate, suggesting that higher flight frequencies contribute to heightened success probabilities.
- The trend in launch success rates has exhibited a consistent ascent from 2013 to 2020, reflecting advancements and improvements in launch technologies and operational methodologies.
- Orbits designated as ES-L1, GEO, HEO, SSO, and VLEO have demonstrated the highest success rates, highlighting their feasibility and reliability for various space missions.
- Among launch sites, KSC LC-39A stands out with the highest number of successful launches, underscoring its reliability and operational efficiency.
- The Decision Tree classifier emerges as the optimal machine learning algorithm for this specific task, leveraging its ability to discern intricate patterns and relationships within the data to facilitate accurate predictions and informed decision-making.

Thank you!