IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jack B.
2/28/2024

Outline

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

Executive Summary

Summary of methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis with Visualization
- Exploratory Data Analysis with SQL
- Interactive Maps with Folium
- Dashboards with Plotly and Dash
- Predictive Analytics with Scikit Learn

Summary of all results

- Exploratory Data Analysis (EDA)
- Interactive analytics
- Predictive Analytics Results

Introduction

Project background and context

- Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage.
- This cost-saving measure is a cornerstone of SpaceX's strategy to make space travel more affordable and sustainable.

Problems you want to find answers

- Based on the features available via API or other open sources can we determine what affects first stage landings?
- Can we determine the cost of a launch?
- What is the success and failure rates? What attributes affect these conditions?
- Can we use machine learning pipelines to predict if the first stage will land?

Section 1

Methodology

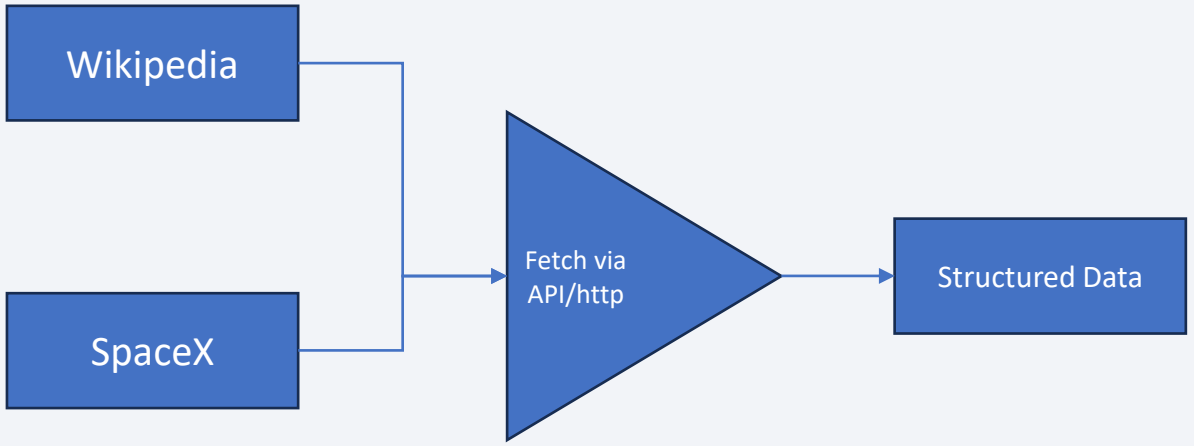
Methodology

Executive Summary

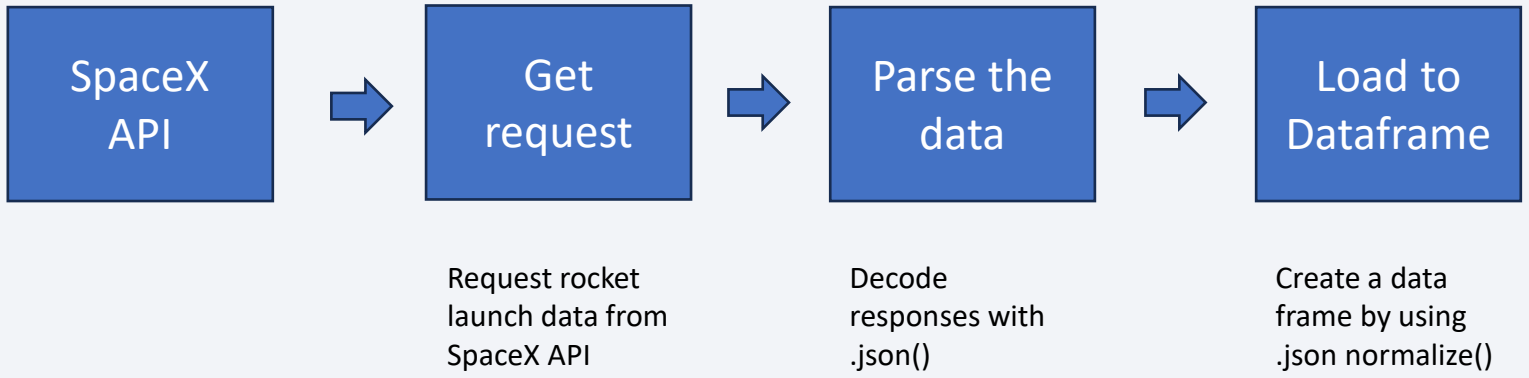
- Data collection methodology:
- Perform data wrangling
  - Data standardization
  - Data cleansing and replacement of non-values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data was collected via the SpaceX API and web scraping Wikipedia
- Over 100 features were extracted for modeling



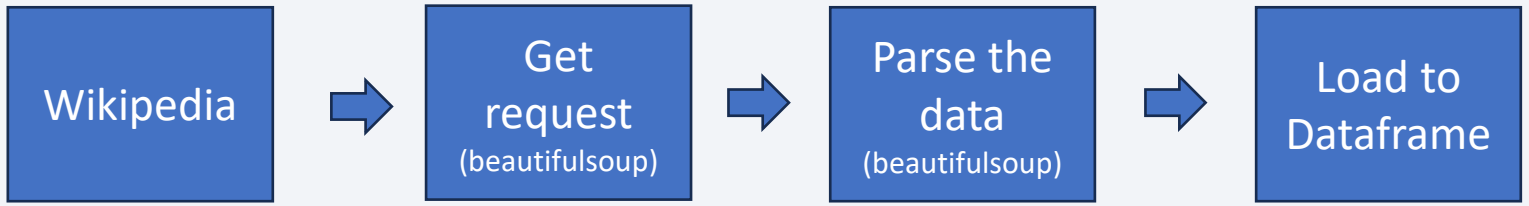
Data Collection – SpaceX API



SpaceX Data via API: Utilizing the SpaceX API for data wrangling involves fetching data related to SpaceX launches, rockets, and missions in JSON format. The wrangling process included filtering this data to obtain specific information (e.g., launches for a particular year), transforming data formats, and cleaning the data (e.g., handling missing values or inconsistencies). This cleaned and structured dataset can then be used for analysis, such as predicting future launch success rates or analyzing the frequency of launches over time.

https://github.com/jack-bil/IBM/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection – Web Scraping



Request the Falcon9 Launch Wiki page from its URL

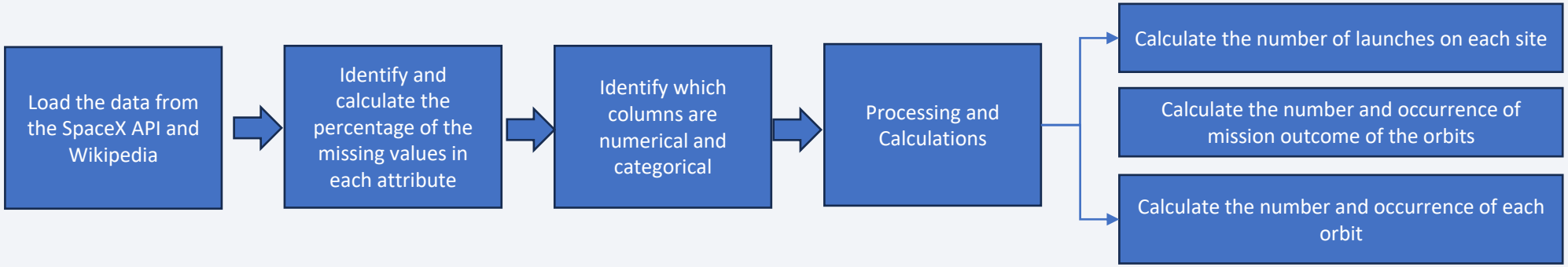
Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Wikipedia Site for Data Extraction: The Wikipedia site can be used to extract structured information from Wikipedia pages. The wrangling process here involves making web calls to retrieve data, parsing the returned html to extract relevant information, and transforming this information into a structured dataset. This dataset could then be analyzed to identify trends in historical events or to compile comprehensive profiles on specific topics or individuals

https://github.com/jack-bil/IBM/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling



- An integral part of data wrangling involves cleaning and preparing the data for analysis. This includes addressing missing values (e.g., by imputing missing data based on other observations or removing incomplete records), identifying and converting data types (e.g., converting strings to dates or categoricals to numerical values as appropriate), and preprocessing data (e.g., normalizing or scaling numerical data, encoding categorical variables for machine learning models).
- These steps are crucial for minimizing errors and biases in data analysis and for ensuring that the dataset is in a suitable format for the specific requirements of downstream tasks such as statistical analyses, reporting, or machine learning algorithms.
- For instance, in the SpaceX data, this could involve ensuring that launch dates are in a consistent datetime format, in the Wikipedia data, ensuring that extracted numerical information is correctly typed as integers or floats, and in web-scraped data, ensuring that text data is cleaned of HTML tags and encoded correctly for text analysis.

https://github.com/jack-bil/IBM/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- **Flight Number by Payload**: Ro analyze the distribution of flight numbers associated with different payload types or classes. It helps in understanding which types of payloads are most commonly launched by SpaceX, indicating their client base or focus areas in space missions.
- **Flight Number by Launch Site**: By plotting the number of flights per launch site, this visualization can give insights into the usage frequency and capacity of each SpaceX launch facility. It may also reflect the strategic importance of each site or suggest site specialization for certain mission types.
- **Payload vs. Launch Site**: Comparing payloads with launch sites might reveal preferences or restrictions for launching certain types of payloads from specific sites, possibly due to geographic, logistical, or regulatory reasons.
- **Success Rate by Orbit Type**: This chart is crucial for evaluating the reliability of SpaceX's launches in achieving different orbit types. It can indicate technical proficiency and operational reliability, which are vital for client trust and future contract negotiations.
- **Flight Number by Orbit Type**: This visualization illustrates the frequency of flights to each orbit type, showing which orbits are most commonly targeted by SpaceX missions. This can be indicative of market demand or SpaceX's strategic interests in certain orbits (e.g., Low Earth Orbit for satellite constellations).
- **Success Rate by Yearly Trend**: Displaying the success rate of launches over time allows for the assessment of SpaceX's performance and improvement in launch technology and operations. This can be an indicator of the company's learning curve, technological advancement, and overall mission assurance efforts.

https://github.com/jack-bil/IBM/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

SQL Queries executed to gain more insight:

- Display the names of the unique launch sites in the space mission
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass.
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

https://github.com/jack-bil/IBM/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers for NASA Johnson Space Center using latitude and longitude
- Markers for all launch sites to show their proximity to the equator and other natural landmarks like the coast
- Markers for successful and failed launches using clusters to identify success rates
- Lines to show distances between the launch site and is proximity to other infrastructures like highways and closest cities

https://github.com/jack-bil/IBM/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Launch sites available via dropdown menus
- Pie charts showing successful launch counts
- Slider for payload mass range
- Scatter chart of payload mass by success rate for various booster versions

https://github.com/jack-bil/IBM/blob/main/spacex_dash_app.py

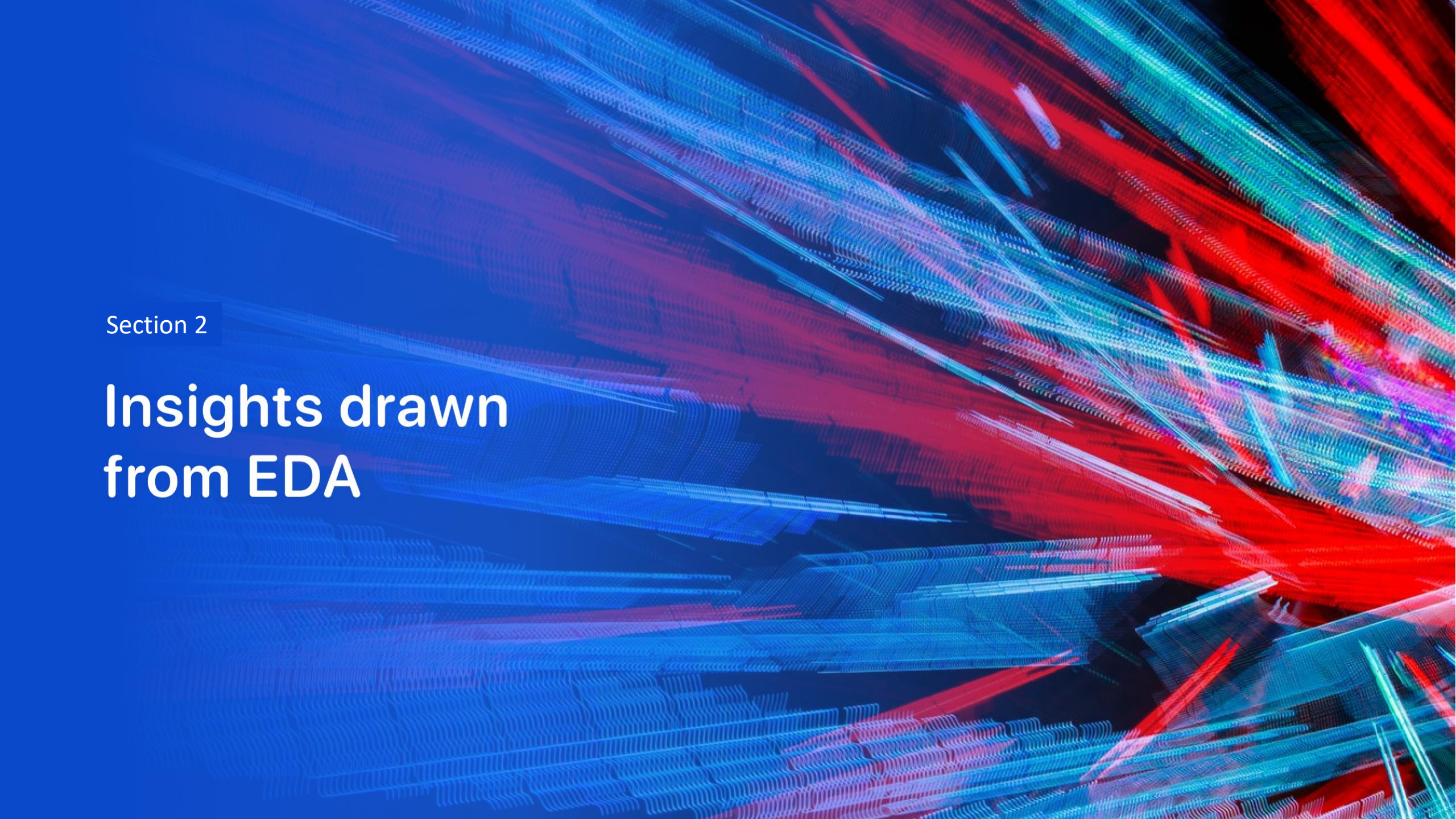
Predictive Analysis (Classification)

- Performed exploratory Data Analysis and determined training labels
  - Create a column for the class
  - Standardize the data
  - Split into training data and test data
- Identified Applicable Models for Classification
- Used Grid Search identify the best Hyperparameter
- Measured the accuracy of each model based on scoring and confusion matrix
- Identified a method to identify the best performing model based on R^2

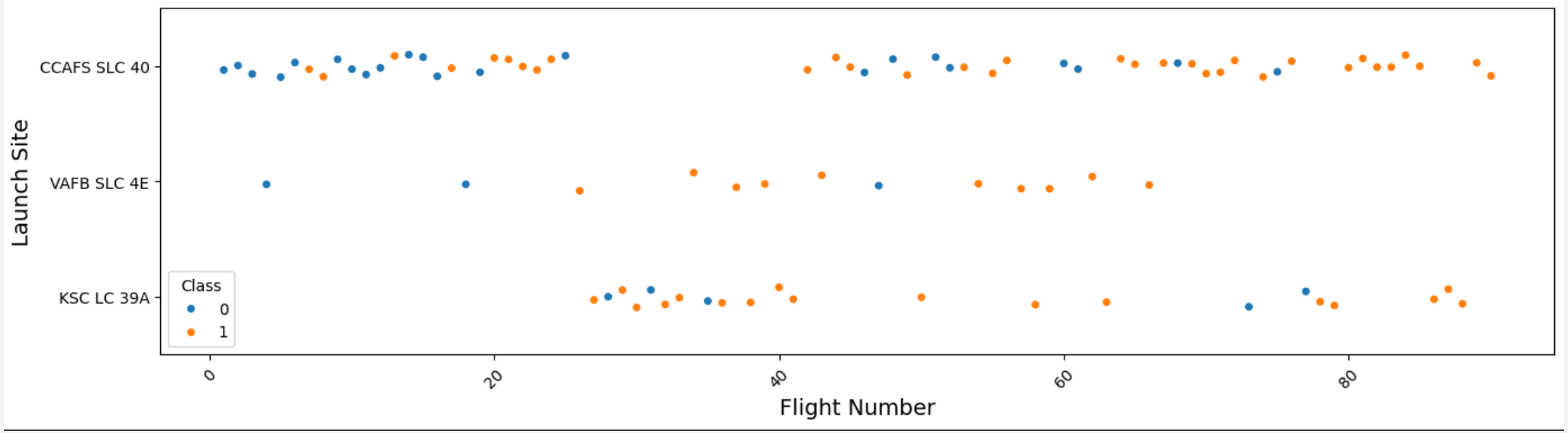
https://github.com/jack-bil/IBM/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

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

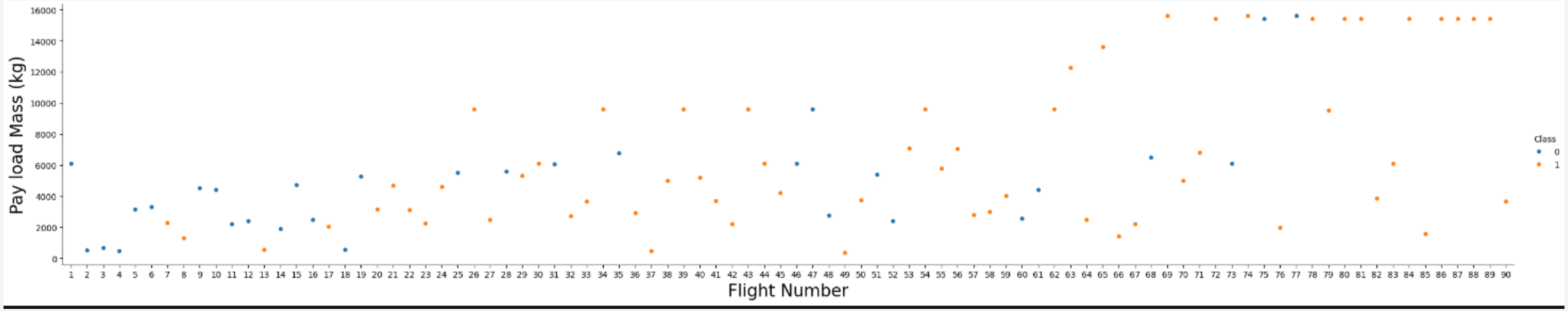

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Three launch sites are depicted: CCAFS SLC 40, VAFB SLC 4E, and KSC LC 39A. CCAFS SLC 40 seems to be the most frequently used site across the range of flights, followed by KSC LC 39A. VAFB SLC 4E has been used less frequently.
- Class 1, which could indicate a successful outcome, is present at all three launch sites. Class 0, possibly indicating a less desirable outcome, also appears at each site but less frequently.
- While Class 1 outcomes are present across all launch sites, there is a visible cluster of Class 0 outcomes at CCAFS SLC 40 in the earlier flights. As the flight number increases, the frequency of Class 0 outcomes seems to diminish, suggesting possible improvements in launch success over time or learning effects.

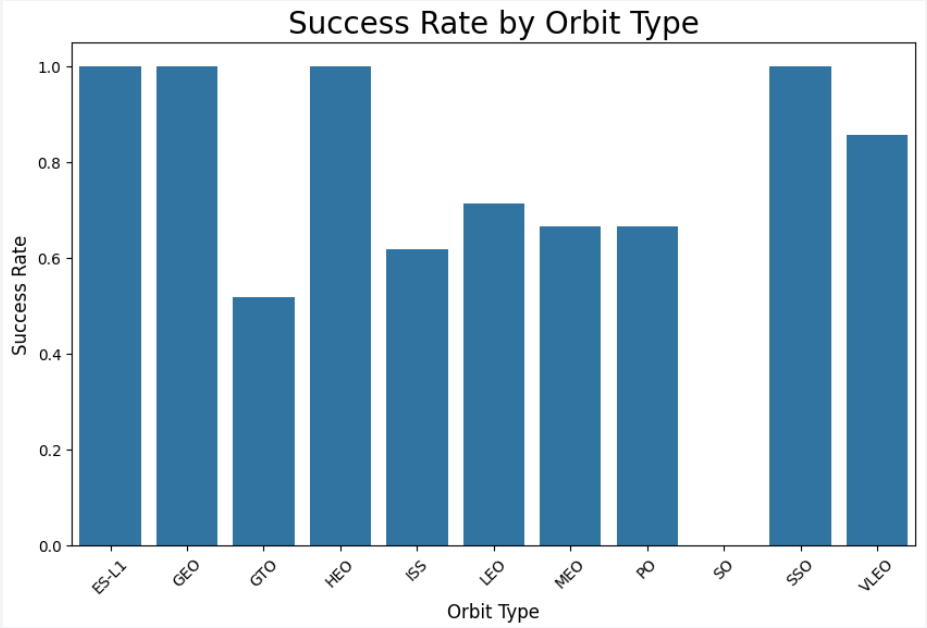
Payload vs. Launch Site



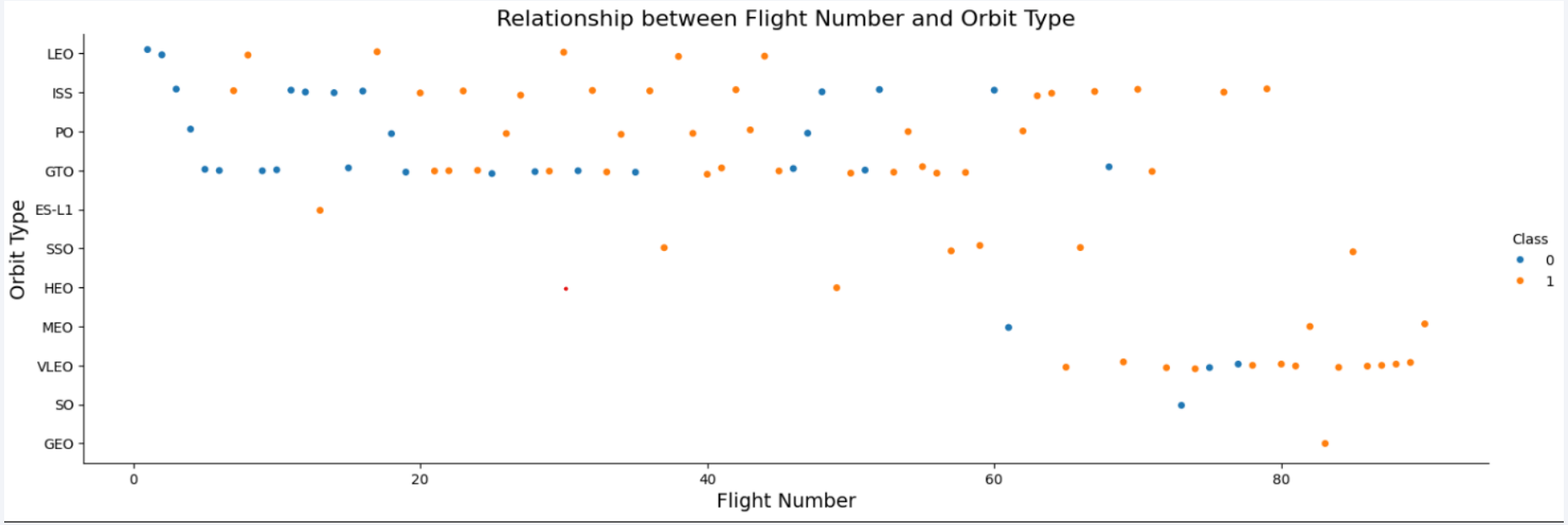
- Class 1, which might represent successful missions or another positive attribute, is represented across a wide range of payload masses and throughout the sequence of flight numbers. Class 0, possibly representing unsuccessful missions or a negative attribute, appears less frequently but is also spread across the entire range of flight numbers and payload masses.
- There seems to be an increase in the range of payload masses over time (as flight numbers increase). In the earlier flights (flight numbers 1-20), payload masses were generally lower. Over time, the scatter plot shows more variation in payload mass, including higher payload masses.
- The increase in maximum payload mass over successive flights might indicate improvements in payload capacity, possibly due to technological advancements or increased confidence in the launch vehicle's capabilities.

Success Rate vs. Orbit Type

- ES-L1, GEO, and VLEO stand out with the highest success rates, which are at or near 100%. This suggests that flights to these orbits are highly reliable or that the conditions and requirements for missions to these orbits are well understood and managed.
- There is a noticeable variability in success rates among different orbit types. GTO, HEO, and ISS have lower success rates compared to others, which could indicate these orbits are more challenging to reach or that the missions to these orbits are more complex.
- Orbits like LEO, MEO, and SSO have success rates that fall in the middle range. This might reflect a balance between the complexity of the missions and the maturity of the technology and procedures used for these flights.

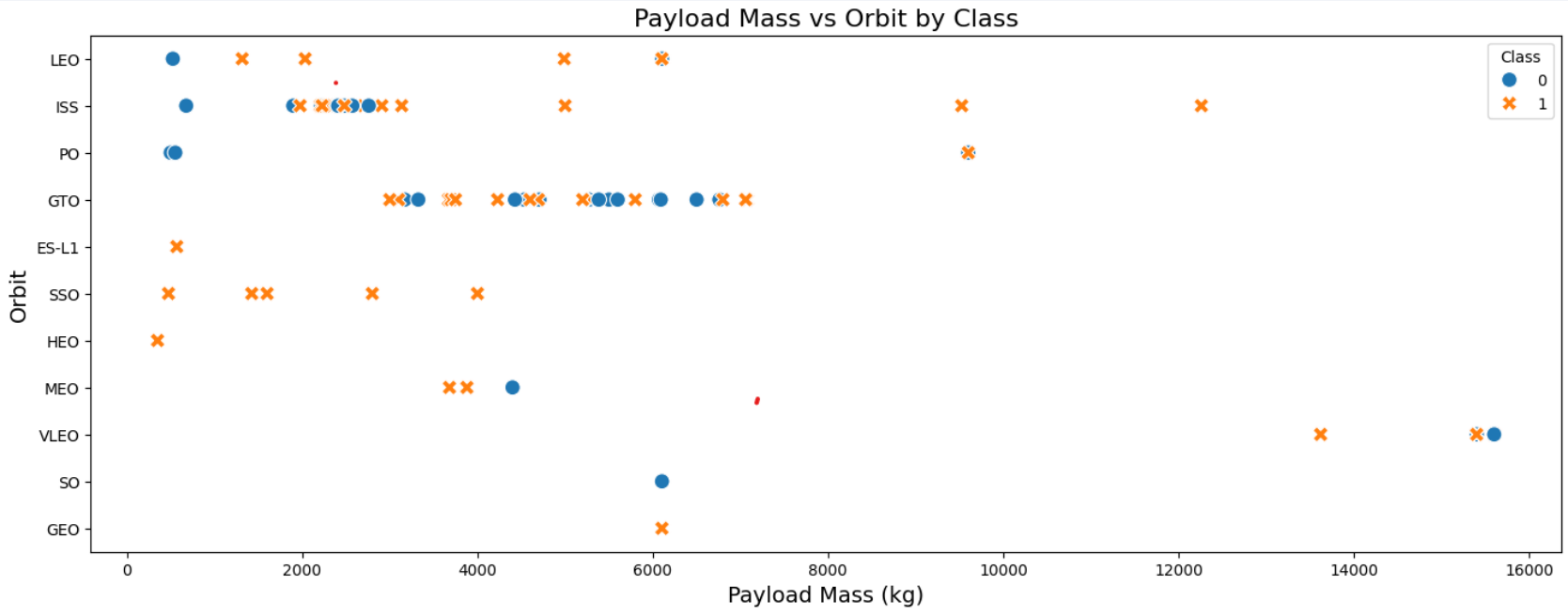


Flight Number vs. Orbit Type



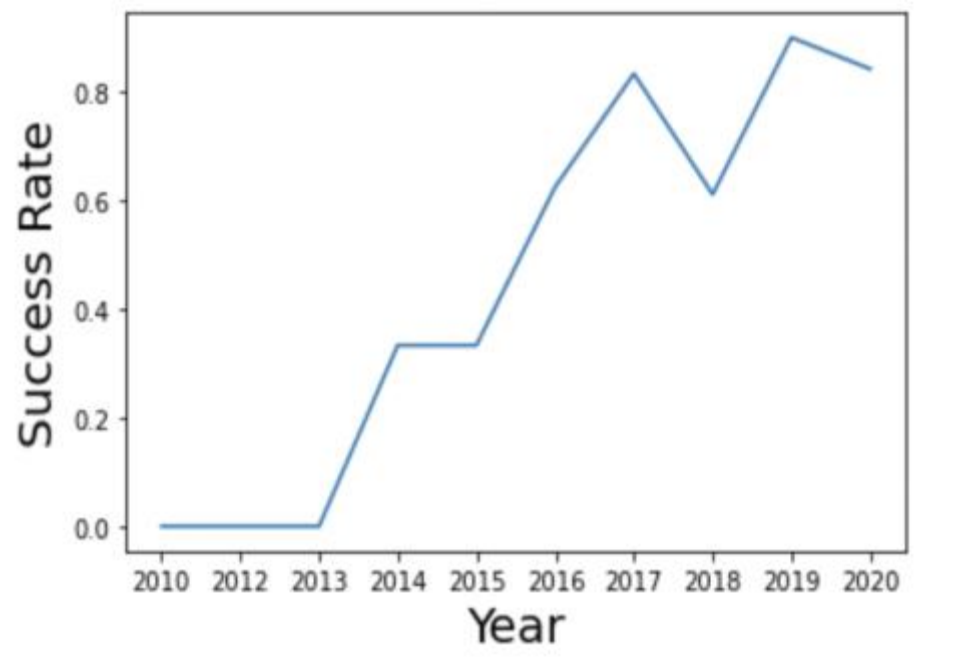
- Certain orbit types, such as LEO (Low Earth Orbit), ISS (International Space Station), and GTO (Geostationary Transfer Orbit), have more flight occurrences than others, like SO (Suborbital) or GEO (Geostationary Orbit). This may indicate a preference or higher demand for certain orbit types, or it might reflect the capabilities or strategic focuses of the flight program.
- Some orbits such as LEO, ISS, and GTO seem to be more commonly targeted compared to others like ES-L1, SSO, and MEO. This could be due to several factors such as the specific needs of payloads, the ease of reaching certain orbits from this launch site, or the commercial demand for these orbits.
- Some orbit types, particularly those at the top and bottom of the plot (like GEO, SO, and VLEO), have fewer flights. These could be more challenging or specialized orbits, or perhaps they are less frequently used for the kinds of missions this program undertakes

Payload vs. Orbit Type



- Different orbits have varying ranges of payload masses. For example, GTO (Geostationary Transfer Orbit) and ISS (International Space Station) orbits have a wide range of payload masses, indicating versatility in the types of missions or variability in the payloads being sent to these orbits.
- Some orbits, such as LEO (Low Earth Orbit), show a high variability in payload mass, which may reflect a broad range of mission profiles that this orbit can accommodate.
- The higher payload masses are predominantly found in GTO, which may indicate that this orbit is commonly selected for heavier payloads, possibly due to the requirements of geostationary satellites.

Launch Success Yearly Trend



- There is a general upward trend in the success rate over the years, which suggests improvement success rates
- Notably, between 2013 and 2016, there is a steep increase in success rate, indicating a period of significant improvement or a series of successful outcomes in those years.
- After a peak in 2017, there is a noticeable decline in the success rate in 2018. This suggests that there may have been issues or challenges that negatively impacted the success rate in that year.

All Launch Site Names

```
%sql select distinct Launch_Site from SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

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

- Four launch sites are listed in the SpaceX table

Launch Site Names Begin with 'CCA'

```
%sql select distinct Launch_Site from SPACEXTABLE where Launch_Site like 'CCA%'
 * sqlite:///my_data1.db
Done.
```

Launch_Site
CCAFS LC-40
CCAFS SLC-40

- Two records begin with `CCA`

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) AS TOTAL from SPACEXTABLE WHERE Customer='NASA (CRS)'
 * sqlite:///my_data1.db
Done.
TOTAL
45596
```

- The total payload carried by boosters from NASA (CRS) is 45,596kg

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) AS TOTAL from SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
```

TOTAL
2928.4

- The average payload for booster version F9 v1.1 is 2928.4kg

First Successful Ground Landing Date

```
%sql select min(date) from SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
```

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

min(date)
2015-12-22

- The first successful landing outcome on ground pad was in 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between '4000' and '6000'
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

- The boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are all F9 FT's
- There are only 4 booster version with an average payload of 5000kg

Total Number of Successful and Failure Mission Outcomes

```
%sql select sum(case when Mission_Outcome = 'Success' then 1 else 0 end) as success, sum(case when Mission_Outcome <> 'Success' then 1 else 0 end) as failed from SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

success	failed
98	3

- The total number of successful outcomes was 98
- The total failure mission outcomes was 3

Boosters Carried Maximum Payload

```
%sql select DISTINCT 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

- Twelve boosters have carried the maximum payload mass

2015 Launch Records

```
%sql select substr(Date, 6,2) as month, Booster_Version, Launch_Site, Landing_Outcome from SPACEXTABLE where substr(Date,0,5)='2015' and Landing_Outcome like '%Fail%Drone%'
 * sqlite:///my_data1.db
Done.
```

month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Above are the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 are listed above
- There were only two cases that took place in January and April

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

```
%sql select Landing_Outcome, count(*) from SPACEXTABLE where date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by count(*) desc
```

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

Landing_Outcome	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

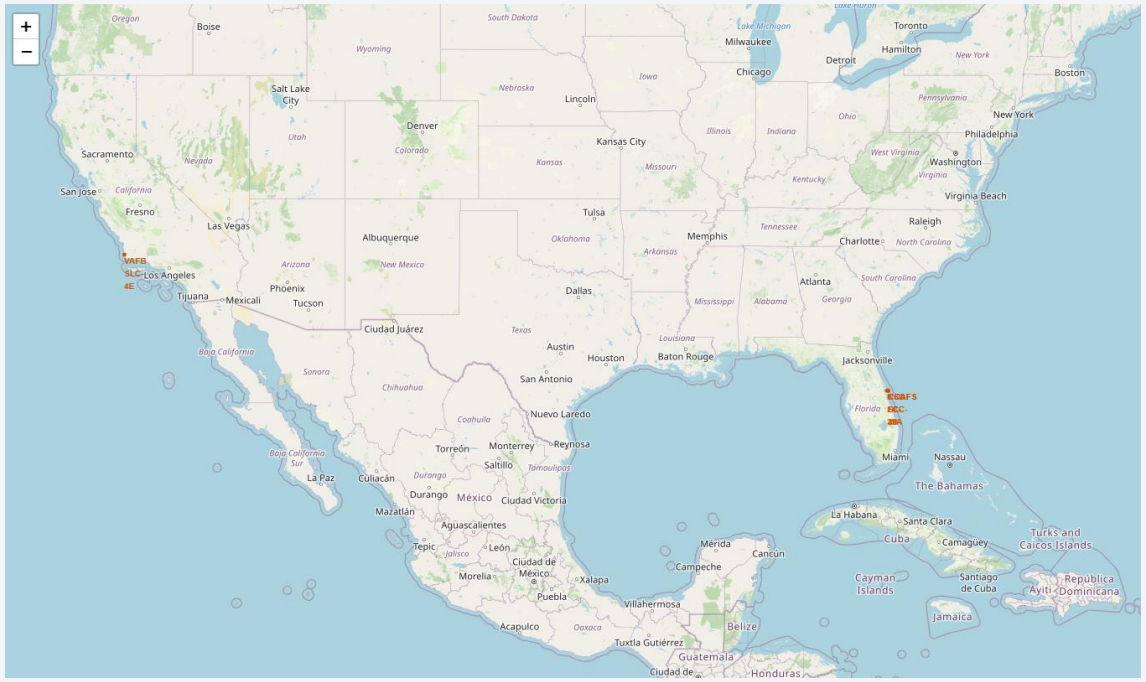
- There are a total of 8 different outcomes with No attempt taking the lead with 10 instances
- There were only 8 successful cases

Section 3

Launch Sites Proximities Analysis

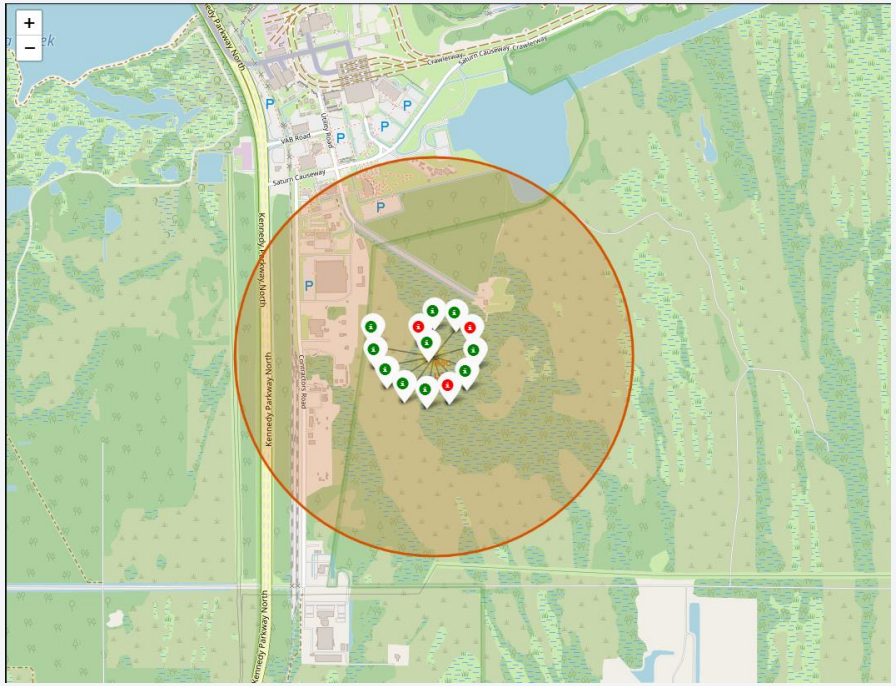
Launch Site Locations

- The launch sites are situated near coastal areas. This is typical for launch sites because it allows for stages of rockets and debris to fall into the ocean, minimizing risks to populated areas.
- Launch sites near the ocean are also chosen to mitigate the environmental impact on terrestrial ecosystems. The ocean's vastness offers a buffer zone for safety and environmental protection.
- Launch sites need to be accessible for transport of large rocket components, which is why they are often located near the sea to facilitate shipping. They also tend to be near infrastructure like roads and railways for overland transport.



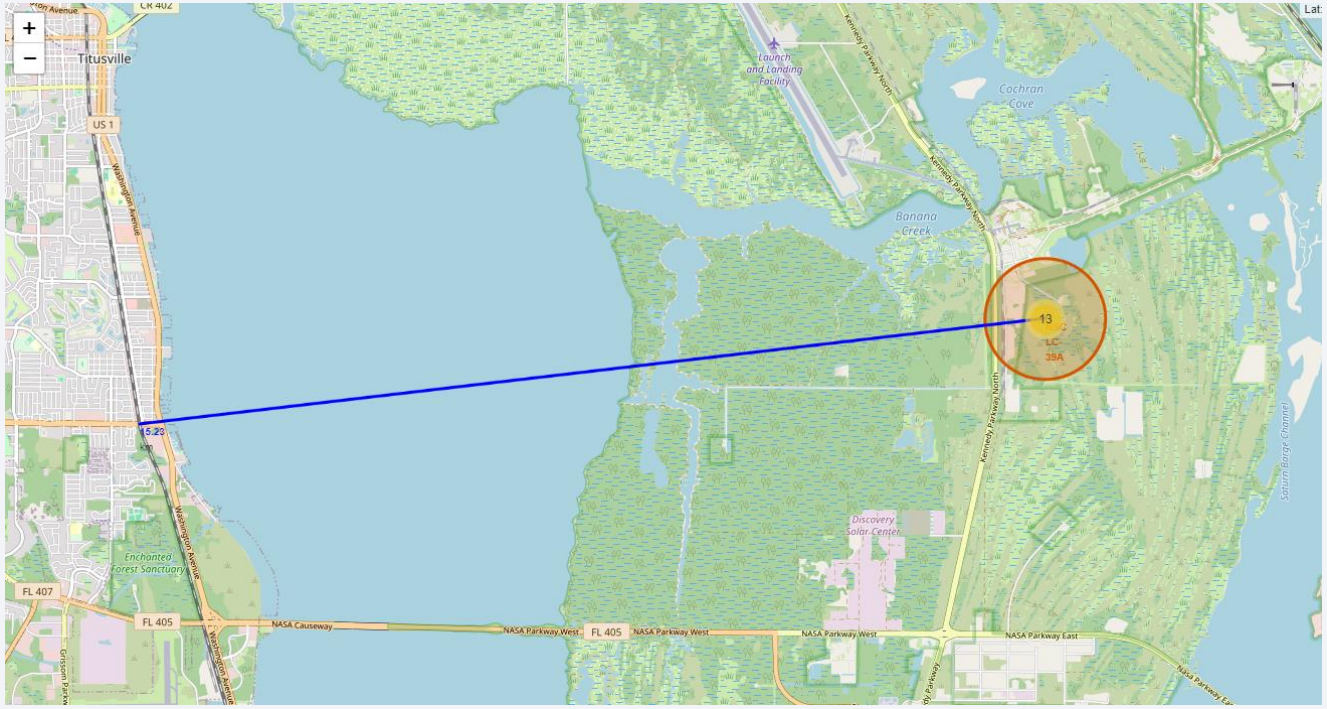
Launch Outcomes KCS LC 39A

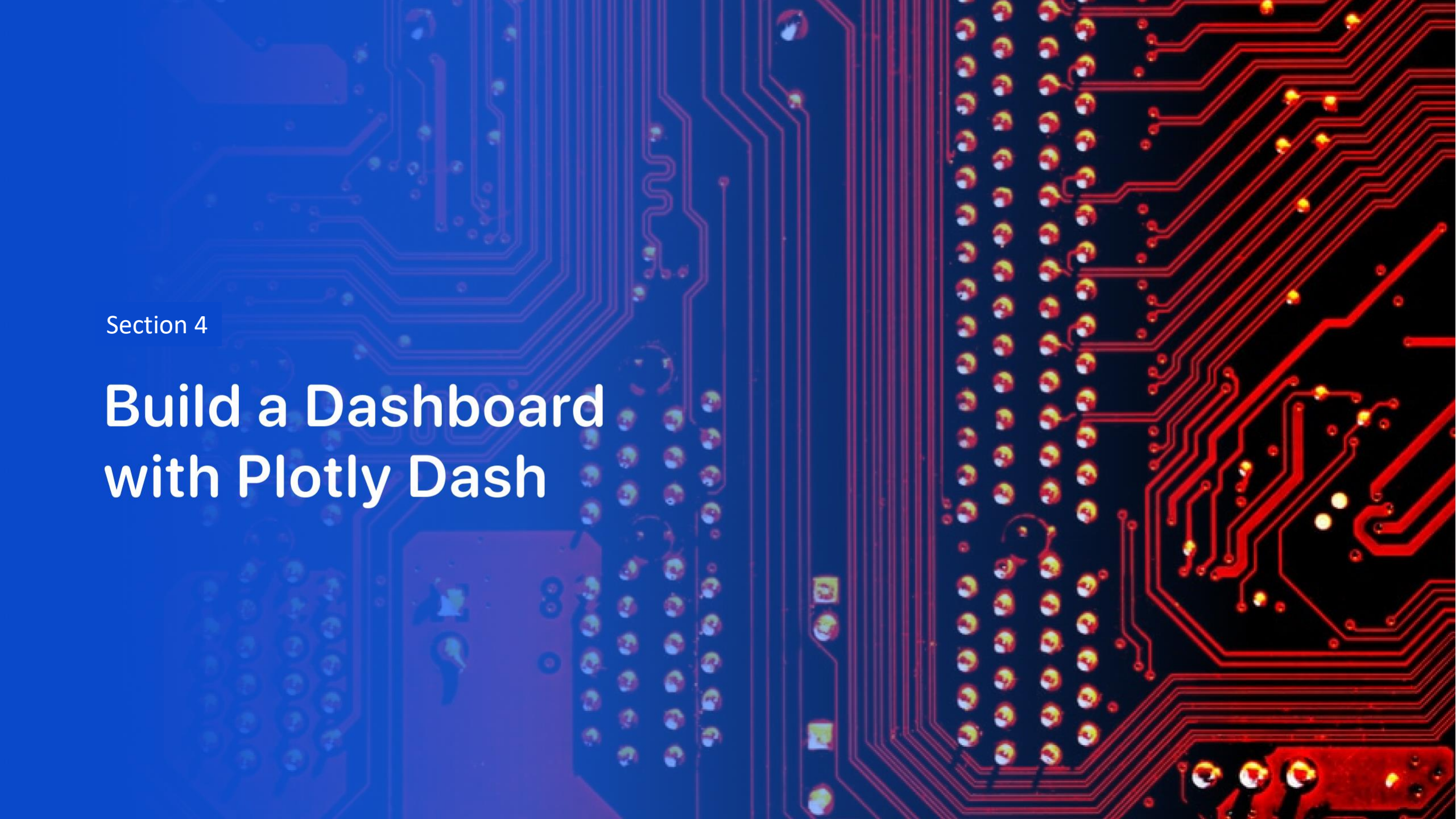
- There is a high concentration of successful launches (green markers) close to the launch site. This suggests a strong record of success for launches from this particular site.
- The red markers, indicating failures, are relatively few compared to the successes and are also close to the launch site. This could imply that if failures occur, they tend to happen shortly after launch, which might be indicative of issues during the initial launch phase.
- The circle represents a safety buffer zone or an area of interest for tracking launch outcomes. The fact that all markers are within this circle suggests that this area is well-monitored and controlled for safety and data collection purposes.
- The distribution of markers does not show a pattern pointing to a particular direction from the launch site, which suggests that failures are not biased towards a specific trajectory.



KCS LC 39A Distance to Railway

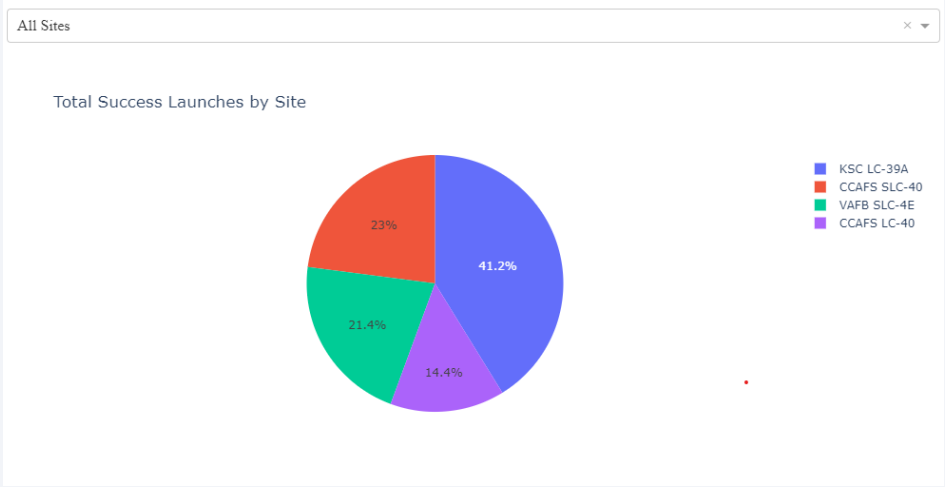
- The blue line may represent the distance between launches from LC 39A and the Florida East Coast Railway.
- There may be safety protocols in place to coordinate launch times with the railway's schedule to minimize the risk of any incident that could affect nearby railway operations or the required distance between the two ~45 miles



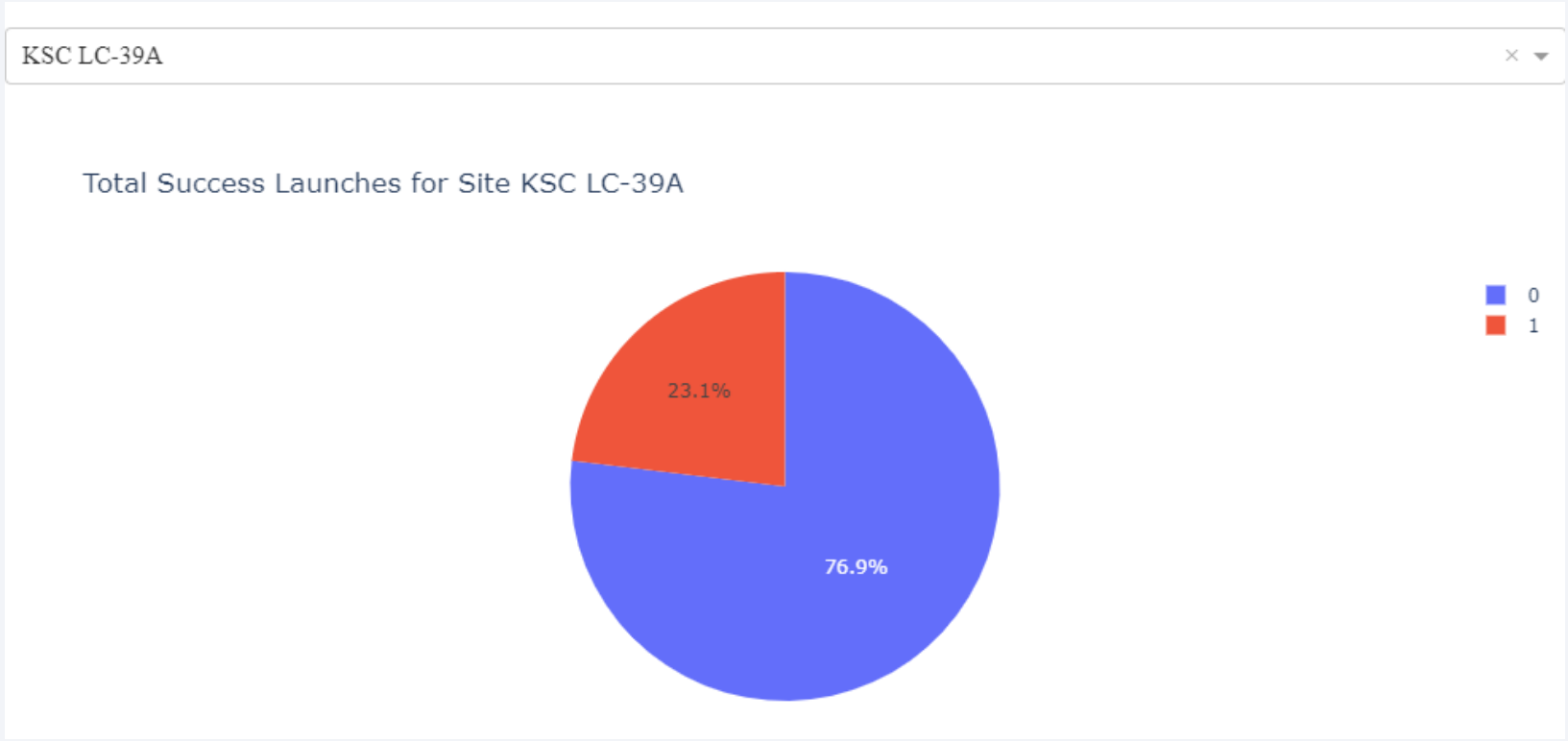

Section 4
Build a Dashboard with Plotly Dash

Success Launches by Site

- KSC LC-39A has the highest proportion of successful launches, making up 41.2% of the total. This suggests that LC-39A is a significant and possibly the most used site for successful launches among the sites listed.
- CCAFS SLC-40 holds the second-largest share with 23%, followed by VAFB SLC-4E with 21.4%. This indicates that these sites are also actively involved in launch operations and have a substantial number of successful launches.
- CCAFS LC-40 has the smallest slice with 14.4%, implying it has the fewest successful launches compared to the other sites.
- The differences in percentages might also reflect the types of missions each site is typically used for. For instance, specific sites might be preferred for certain types of payloads or orbits, affecting the number of successful launches recorded.
- The chart may also reflect the operational history and development of each site. A newer site might have fewer launches and therefore a smaller proportion of the total, or a site might have been upgraded over time to handle more or more complex missions.



Most Successful Launch Site by Ratio



- KSC LC-39A has a high success rate with more than three-quarters of the launches being successful
- The large portion of successful launches indicates that KSC LC-39A is a reliable site for space missions, which may influence decisions for future launch site selections.

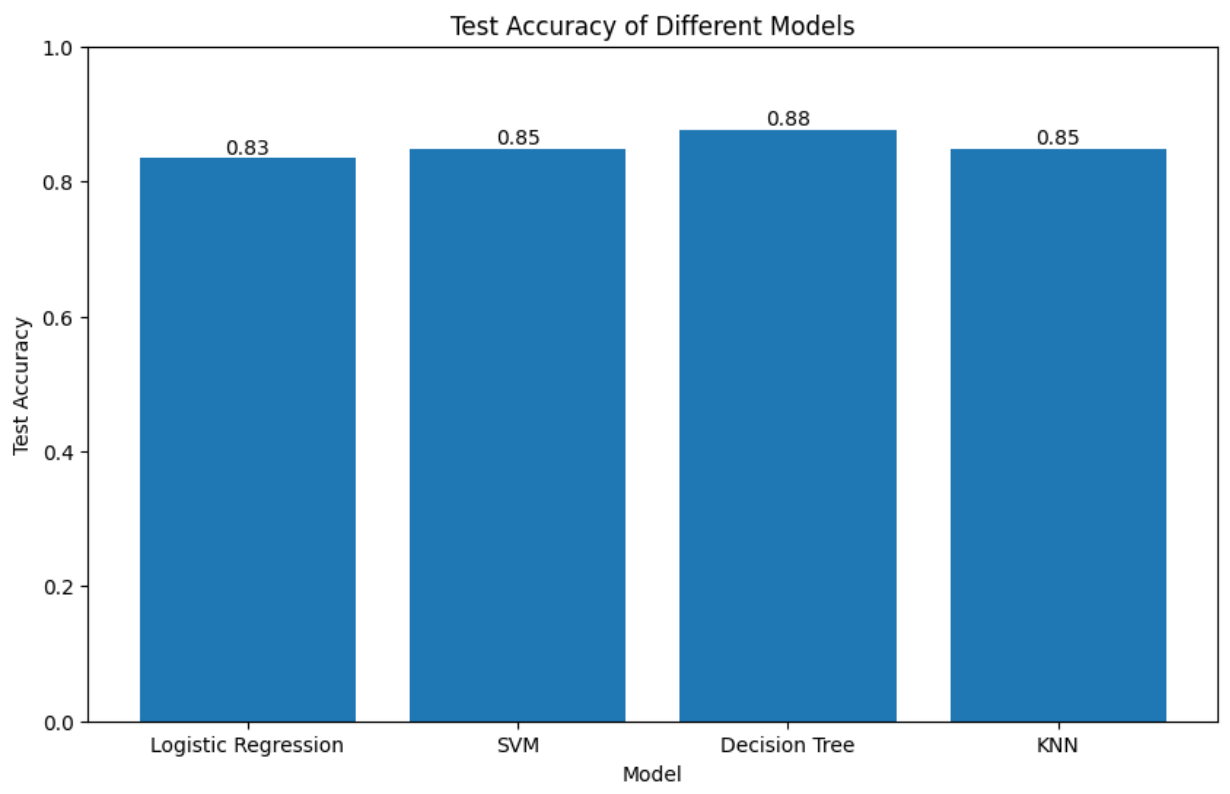
Correlation Between Payload and Success

- Successful launches (class 1) occur across a wide range of payload masses, from very light payloads to those approaching 10,000 kg. This suggests that success is not strictly limited by the mass of the payload.
- Failures (class 0) seem to occur primarily at the lower end of the payload mass range. This could suggest that failure is not necessarily due to the payload being too heavy, but rather to other factors.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



- The best model is Decision Tree with a test accuracy of 0.8767857142857143 given that the hyperparameters are tuned with grid search.

Confusion Matrix – Decision Tree

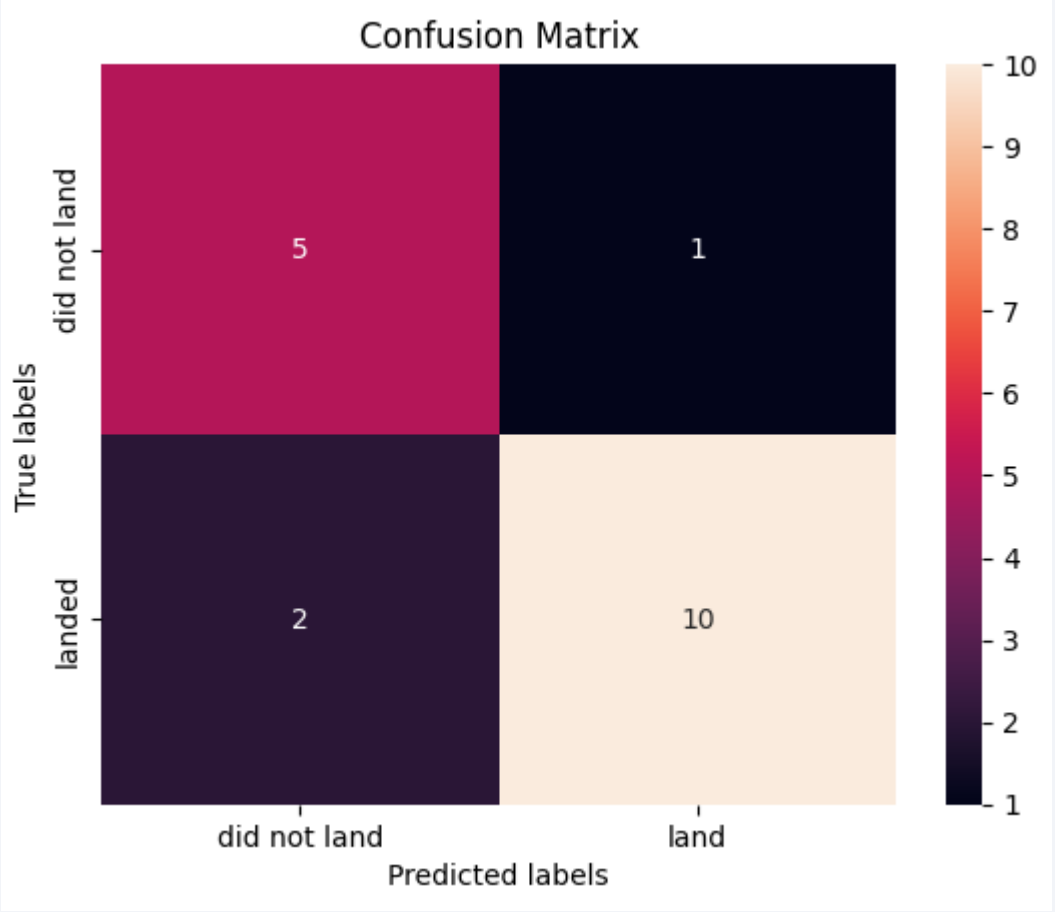
- Overall, the model seems to perform reasonably well, with relatively high precision and recall.
- There are more instances where the model correctly identifies landings than when it misses or falsely identifies them. However, the small sample size (18 instances) makes it difficult to generalize these results without knowing the context or the application where this model is used.
- One might prefer to optimize for either precision or recall. For instance, if the cost of a false negative is very high, one might want to have a model with higher recall even at the cost of precision.
- Below are the calculated values based on the matrix:

  Accuracy: (10 + 5) / (10 + 5 + 1 + 2) = 15 / 18 ≈ 0.833 or 83.3%

  Precision: 10 / (10 + 1) = 10 / 11 ≈ 0.909 or 90.9%

  Recall: 10 / (10 + 2) = 10 / 12 ≈ 0.833 or 83.3%

  F1 Score: 2 * (Precision * Recall) / (Precision + Recall) = 2 * (0.909 * 0.833) / (0.909 + 0.833) ≈ 0.869 or 86.9%



Conclusions

- The initial performance metrics of all models were similar prior to hyperparameter tuning, indicating that out-of-the-box, the models have comparable baseline effectiveness for the given classification task.
- Hyperparameter optimization using grid search significantly improved the Decision Tree model's performance over others, suggesting that this model is more sensitive to hyperparameter changes and has a higher potential for performance gains with fine-tuning.
- Exploring further hyperparameter optimization techniques such as Optuna could potentially yield even better-tuned models, indicating the need for a more thorough and possibly automated search for optimal model parameters.

Thank you!