IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Pratyush Nandan
1/8/2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data Wrangling
  - EDA with data visualization
  - Built an interactive map with folium
  - Built a Dashboard with Plotly Dash
  - Predictive analysis
- Summary of all results
  - EDA results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- Project background and context
  - SpaceX has become the leading company in the commercial space industry, making space travel more cost-effective. The Falcon 9 rocket launches are advertised on their website for $62 million, significantly less than the $165 million charged by other providers. This price difference is primarily due to SpaceX's ability to reuse the first stage of their rockets. By predicting whether the first stage will successfully land, we can estimate the cost of a launch. Using public data and machine learning models, we aim to predict if SpaceX will reuse the first stage.
- Problems you want to find answers

  •How do variables like payload mass, launch site, number of flights, and orbits impact the success of the first stage landing?
  •Has the rate of successful landings increased over time?
  •What is the most effective algorithm for binary classification in this scenario?

Section 1

Methodology

Methodology

Executive Summary

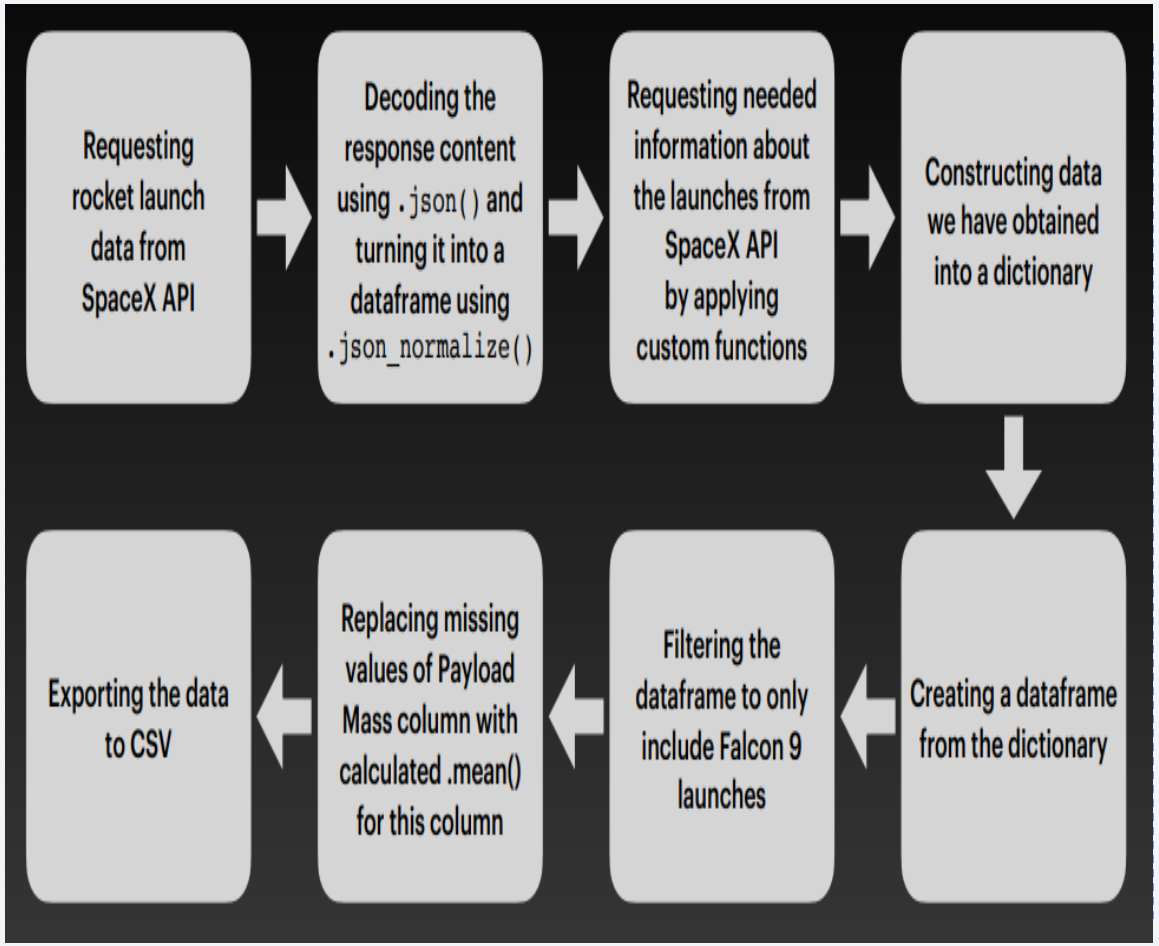
- Data collection methodology:
  - Using API and web scraping
- Perform data wrangling
  - Filtered the data
  - Dealt with missing values and used one Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built , tuned and evaluated the classification models to ensure the best results were gotten

Data Collection

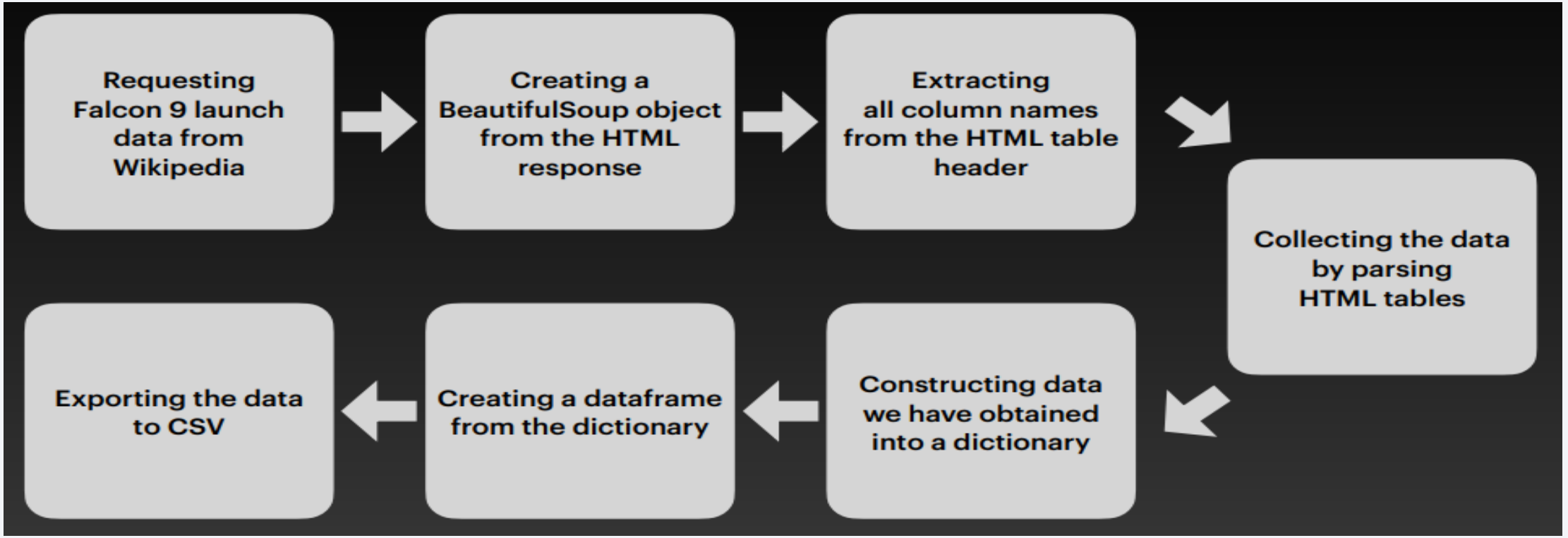
- To collect comprehensive information about the launches for detailed analysis, we used a combination of API requests from the SpaceX REST API and web scraping from SpaceX's Wikipedia page. The data columns obtained from the SpaceX REST API include: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude. The 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

- https://github.com/Pratyush-bit69/ibmcapstone-datascience/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



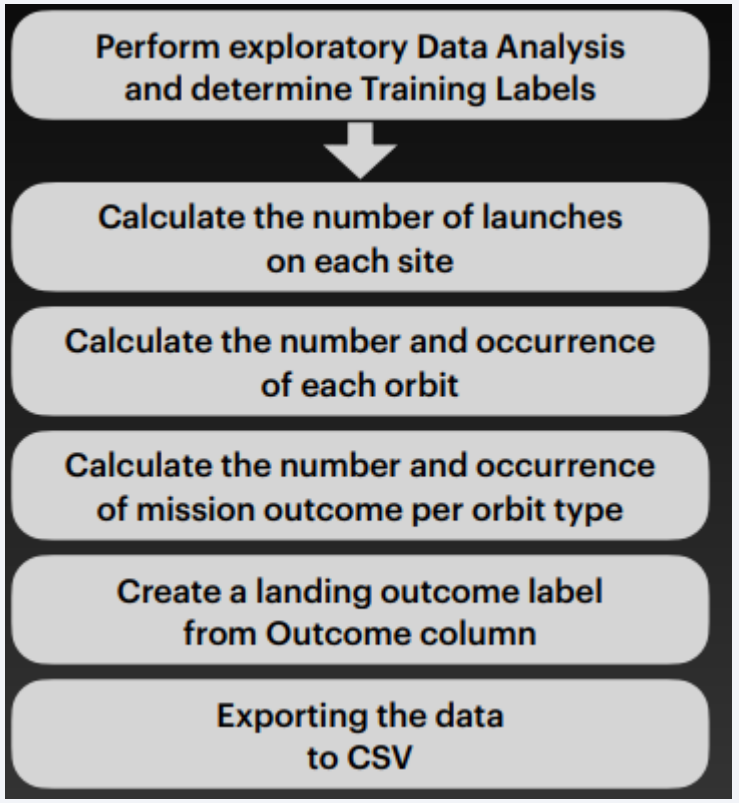
Data Collection - Scraping



- https://github.com/Pratyush-bit69/ibmcapstone-datascience/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- The dataset includes instances where booster landings were unsuccessful. Occasionally, a landing was attempted but failed due to an accident. For example, "True Ocean" refers to a successful landing in a designated ocean region, while "False Ocean" indicates an unsuccessful attempt in the same area. Similarly, "True RTLS" denotes a successful landing on a ground pad, whereas "False RTLS" indicates a failed landing on a ground pad. "True ASDS" represents a successful landing on a drone ship, and "False ASDS" means the landing attempt on a drone ship was unsuccessful. These outcomes are primarily converted into training labels, with "1" indicating a successful booster landing and "0" representing a failed attempt.



https://github.com/Pratyush-bit69/ibmcapstone-datascience/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Charts were generated as follows: Flight Number versus Payload Mass, Flight Number versus Launch Site, Payload Mass versus Launch Site, Orbit Type versus Success Rate, Flight Number versus Orbit Type, Payload Mass versus Orbit Type, and Success Rate by Year.
- Scatter plots were utilized to show relationships between variables, which can be beneficial for machine learning models if a relationship is present. Bar charts were employed to compare various discrete categories, aiming to highlight the relationship between the categories and their associated values. Line charts were used to illustrate trends in data over time, reflecting time series patterns.

- https://github.com/Pratyush-bit69/ibmcapstone-datascience/blob/main/edadataviz.ipynb

EDA with SQL

- SQL queries were run to:
- Display the names of all unique launch sites used in the space missions.
- Retrieve 5 records where the launch sites start with the string 'CCA'.
- Show the total payload mass carried by boosters launched by NASA (CRS).
- Calculate the average payload mass carried by the booster version F9 v1.1.
- Identify the date of the first successful landing on a ground pad.
- List the names of boosters that have successfully landed on a drone ship and carried a payload mass between 4000 and 6000.
- Display the total number of successful and failed mission outcomes.
- Identify the booster versions that have carried the highest payload mass.
- Show failed landing outcomes on drone ships, including their booster versions and launch site names, for the months in the year 2015.
- Rank the count of landing outcomes (e.g., Failure on drone ship or Success on ground pad) between June 4, 2010, and March 20, 2017, in descending order.
- https://github.com/Pratyush-bit69/ibmcapstone-datascience/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

•**Markers for All Launch Sites:**
•A marker with a circle, popup label, and text label was added for NASA Johnson Space Center, using its latitude and longitude coordinates.
•Similar markers were created for all other launch sites to show their geographical locations and proximity to the Equator and coasts.
•**Colored Markers for Launch Outcomes at Each Launch Site:**
•Colored markers indicate successful (green) and failed (red) launches.
•Marker clusters help identify which launch sites have higher success rates.
•**Distances Between a Launch Site and Its Proximities:**
•Colored lines illustrate distances between the KSC LC-39A launch site and nearby features like railways, highways, coastlines, and the nearest city.

https://github.com/Pratyush-bit69/ibmcapstone-datascience/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

•**Launch Sites Dropdown List:**
•Added a dropdown list for selecting different launch sites.
•**Pie Chart Displaying Successful Launches:**
•Included a pie chart to visualize the total number of successful launches across all sites.
•Displays Success vs. Failure counts for a selected site, if applicable.
•**Payload Mass Range Slider:**
•Implemented a slider to select the range of Payload Mass.
•**Scatter Chart of Payload Mass vs. Success Rate:**
•Added a scatter chart to show the relationship between Payload Mass and Launch Success Rate for different booster versions.

https://github.com/Pratyush-bit69/ibmcapstone-datascience/blob/main/spacex_dash_app.py

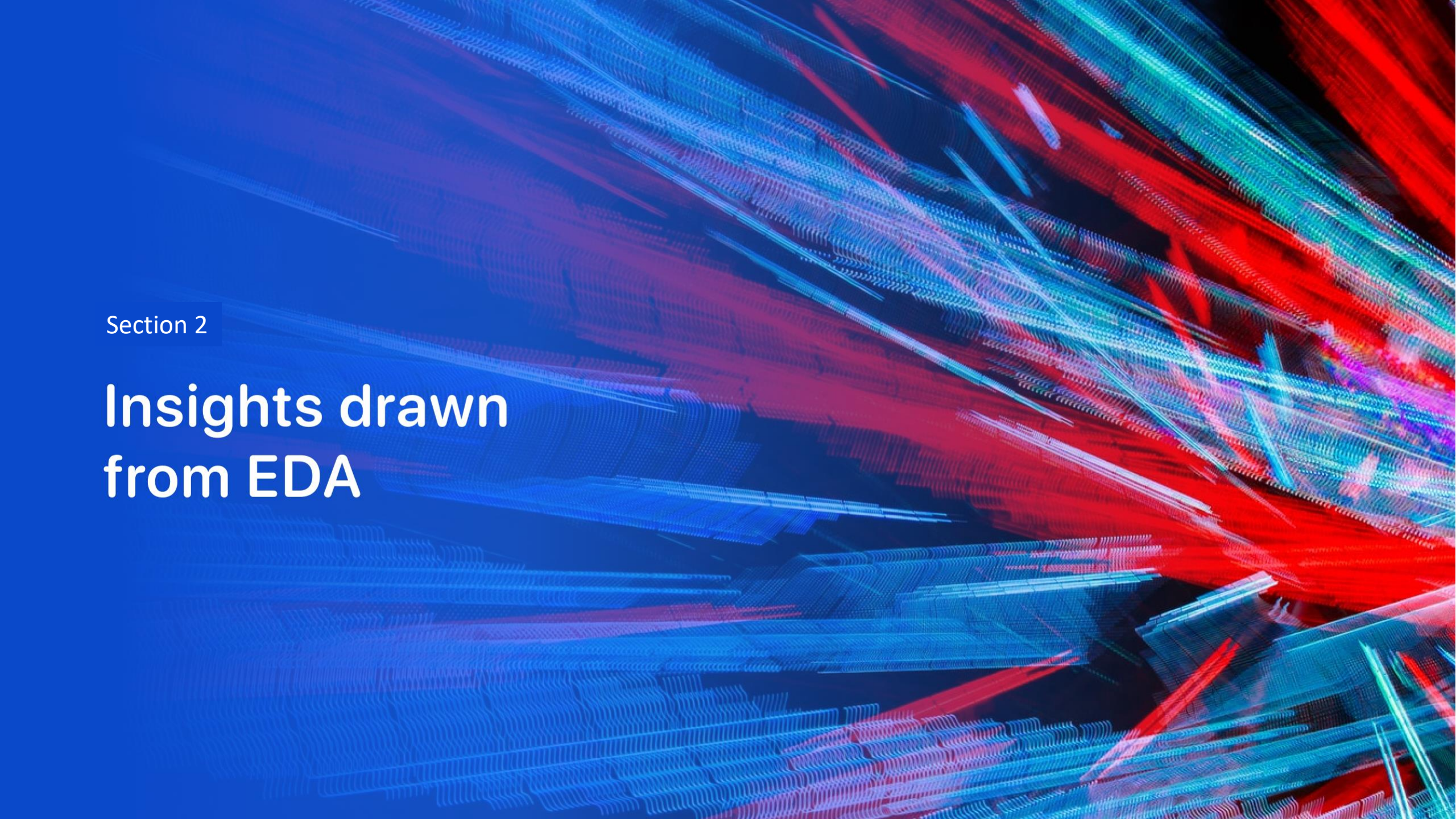
Predictive Analysis (Classification)

- Markers for All Launch Sites: A marker was created with a circle, popup label, and text label for NASA Johnson Space Center, using its latitude and longitude as the reference point. Markers featuring circles, popup labels, and text labels were also added for all other launch sites to show their geographical locations and their proximity to the equator and coasts.
- Colored Markers for Launch Outcomes at Each Launch Site: Colored markers were introduced to represent launch outcomes, with green indicating successful launches and red for failed ones. Marker clusters were used to highlight which launch sites have higher success rates.
- Distances Between a Launch Site and Its Proximities: Colored lines were added to illustrate the distances between the launch site KSC LC-39A (as an example) and nearby features, including railway, highway, coastline, and the nearest city.

https://github.com/Pratyush-bit69/ibmcapstone-datascience/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%5B1%5D.ipynb

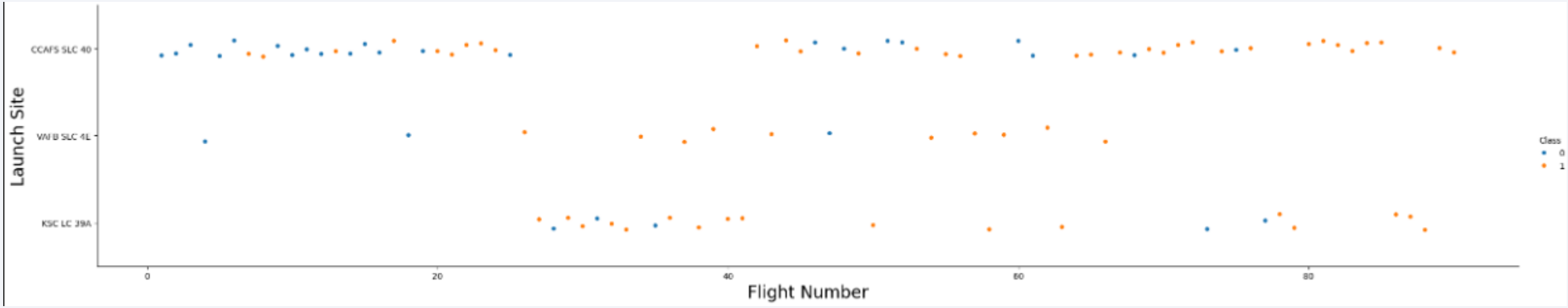
Results

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


Section 2
Insights drawn from EDA

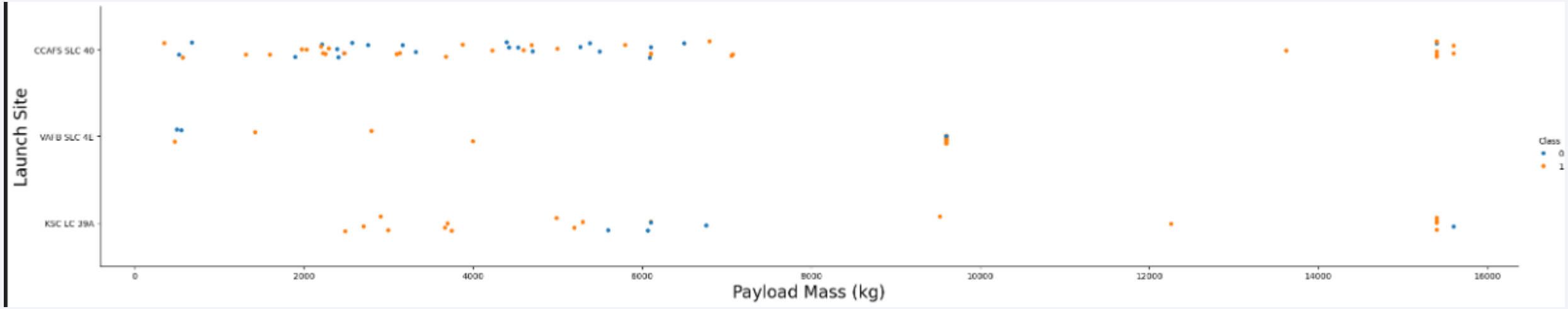
Flight Number vs. Launch Site

- Explanation: In the beginning, many flights were unsuccessful, but recent launches have shown a consistent success rate. The CCAFS SLC 40 launch site is responsible for about half of all launches. Both VAFB SLC 4E and KSC LC 39A have demonstrated higher success rates. It seems that the likelihood of a successful launch increases with each new attempt.



Payload vs. Launch Site

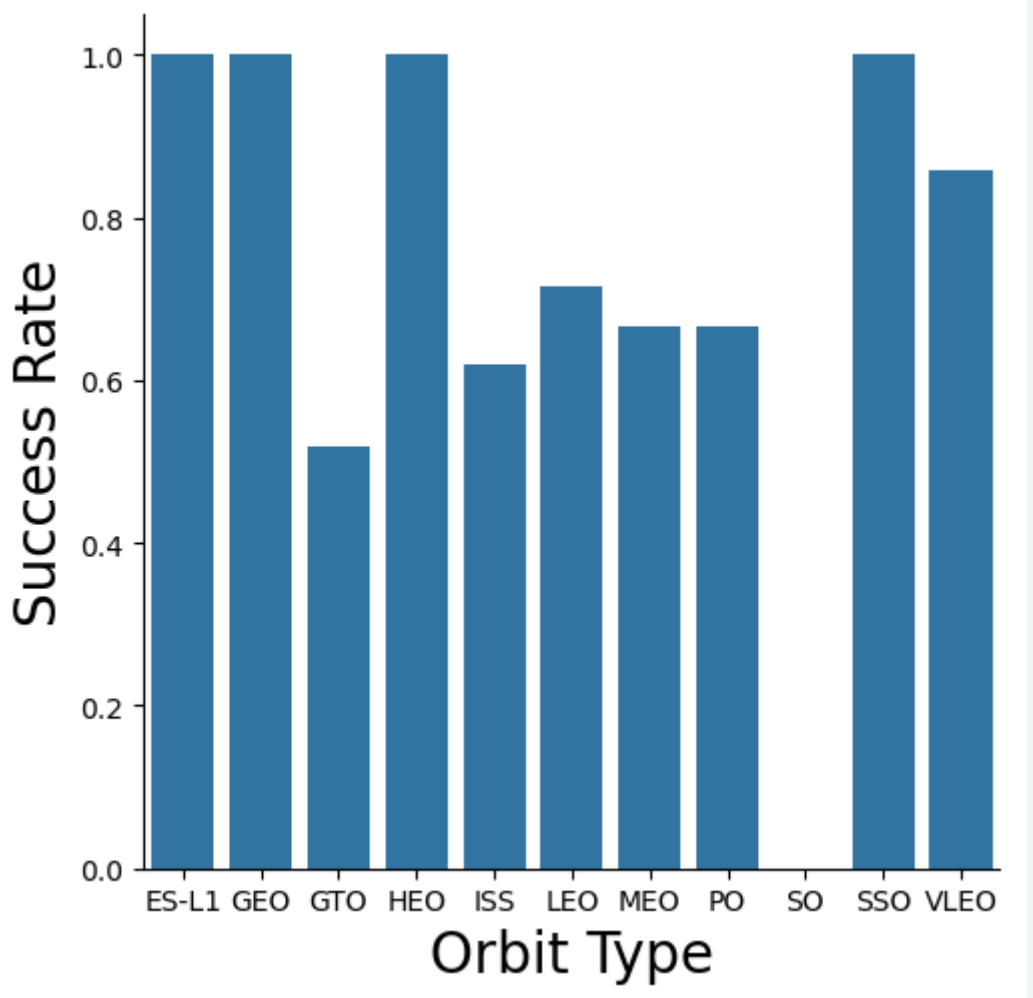
- Explanation: A higher payload mass at any launch site is associated with a higher success rate. Launches with payloads over 7000 kg typically achieved successful outcomes. Additionally, the KSC LC 39A site maintains a perfect success rate for payloads weighing less than 5500 kg.



Success Rate vs. Orbit Type

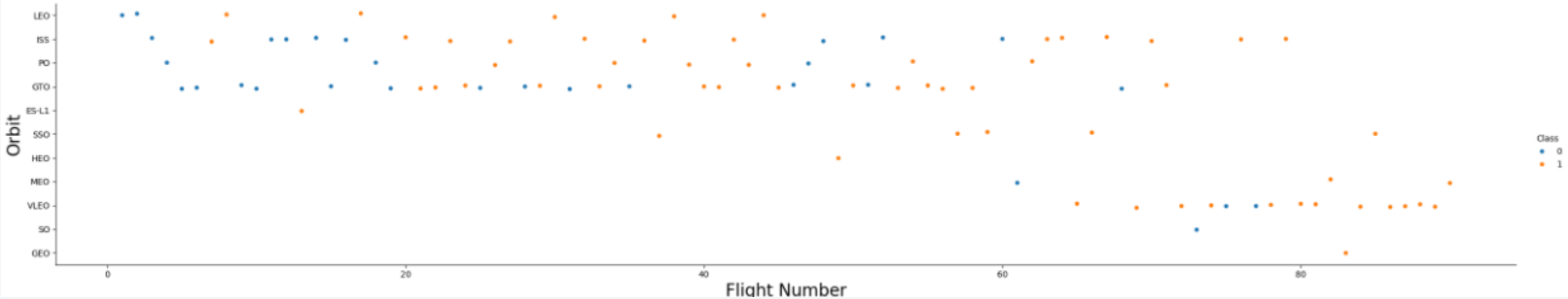
Explanation:

- Orbits with complete success rates include ES-L1, GEO, HEO, and SSO.
- The SO orbit has a 0% success rate.
- Orbits with success rates ranging from 50% to 85% are GTO, ISS, LEO, MEO, and PO.



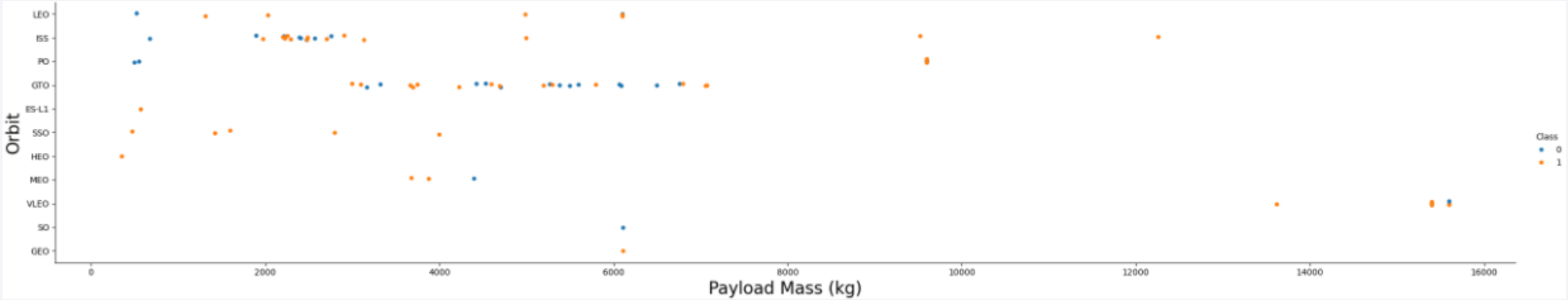
Flight Number vs. Orbit Type

- Explanation: In the LEO orbit, there seems to be a positive correlation between the number of flights and the success rate. On the other hand, in the GTO orbit, no clear relationship between the number of flights and success can be observed.



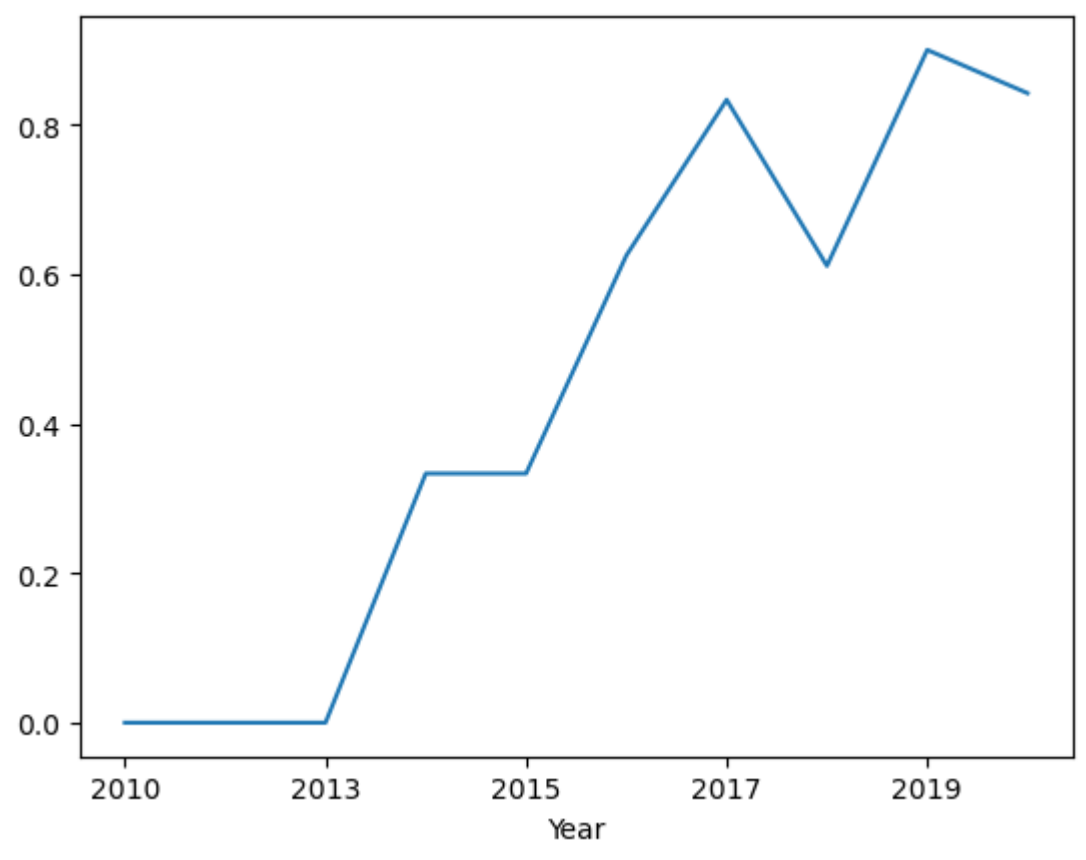
Payload vs. Orbit Type

- Explanation: Heavy payloads tend to have a negative impact on GTO orbits but a positive impact on GTO and Polar LEO (ISS) orbits.



Launch Success Yearly Trend

- The success rate start increasing from 2013 and kept increasing ever since



All Launch Site Names

- Represented all the distinct launch site name from the spacex table

```
In [12]: %sql select distinct launch_site from SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
```

Out[12]:

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

Launch Site Names Begin with 'CCA'

- Displayed 5 records where launch site began with the string like ‘CCA’

```
In [13]: %sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
 * sqlite:///my_data1.db
Done.
```

Out[13]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Lar
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Fai
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	Fai
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass

- Displayed total payload mass where boosters were launched by NASA

```
In [14]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
Out[14]: total_payload_mass
         45596
```

Average Payload Mass by F9 v1.1

- Displayed the average payload mass carried by the booster whose version is F9 v1.1

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';
```

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

average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

- Displayed the first successful landing date from the table whose outcome is success and retrieved the lowest date using min function.

```
In [17]: %sql select min(date) as first_successful_landing from SPACEXTABLE where landing_outcome = 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
Out[17]: first_successful_landing
---------------------------------
2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listed all the booster version which have successful landed in drone ship with payload in between 4000 and 6000

```
In [19]: %sql select booster_version from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
 * sqlite:///my_data1.db
Done.
```

Out[19]:

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

Total Number of Successful and Failure Mission Outcomes

- Displayed the total number of successful and failure mission

```
In [20]: %sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
 * sqlite:///my_data1.db
Done.
```

Out[20]:

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

Boosters Carried Maximum Payload

- Displayed the booster version which carried the maximum load using a sub query which retrieved the max payload first.

```
In [22]: %sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);
 * sqlite:///my_data1.db
Done.
```

Out[22]:

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

2015 Launch Records

- Listed all failed landing outcomes in drone ship,their booster version along with launch site for the months in year 2015

In [29]:
```
%%sql select substr(6,2,date) as month, date, booster_version, launch_site, landing_outcome from SPACEXTABLE
      where landing_outcome = 'Failure (drone ship)' and substr(date,0,5)='2015';
```

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

Out[29]:

month	Date	Booster_Version	Launch_Site	Landing_Outcome
	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- 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 count_outcomes from SPACEXTABLE
      where date between '2010-06-04' and '2017-03-20'
      group by landing_outcome
      order by count_outcomes desc;
```

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

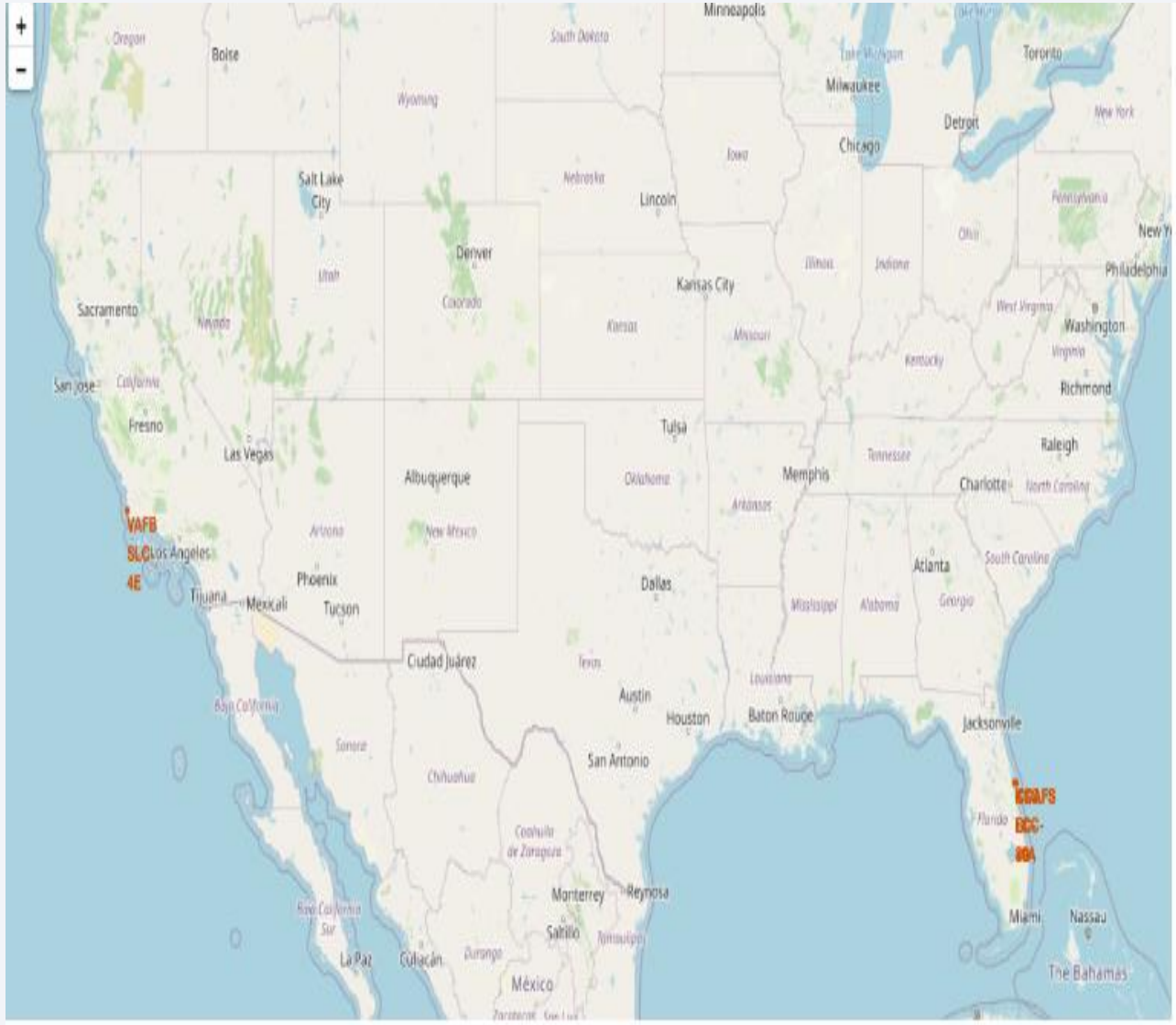
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Location of all launch sites on a world map

- Explanation: Most launch sites are positioned near the Equator. At the Equator, the Earth's rotational velocity is at its peak, moving at 1670 km/hour. This rapid rotation provides an initial boost due to inertia when launching spacecraft, aiding in achieving the required orbital speed.
- Additionally, all launch sites are located near coastal areas. Launching rockets over the ocean reduces the risk of debris or explosions affecting populated regions, thereby enhancing safety.

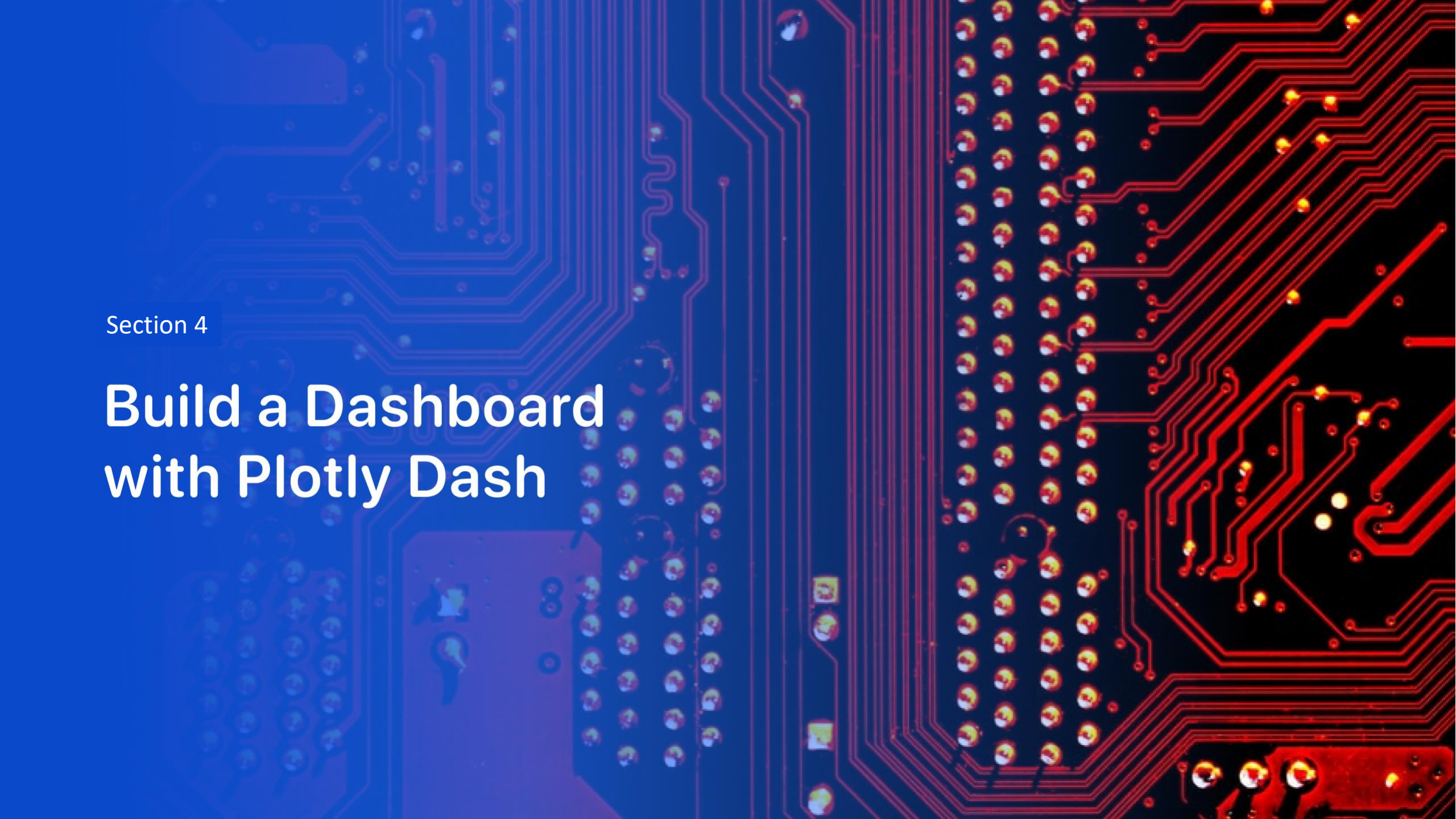


Color labelling records on the map

- Explanation: Color-coded markers are utilized to show the success rates of launch sites. A green marker indicates a successful launch, while a red marker signifies a failed launch.

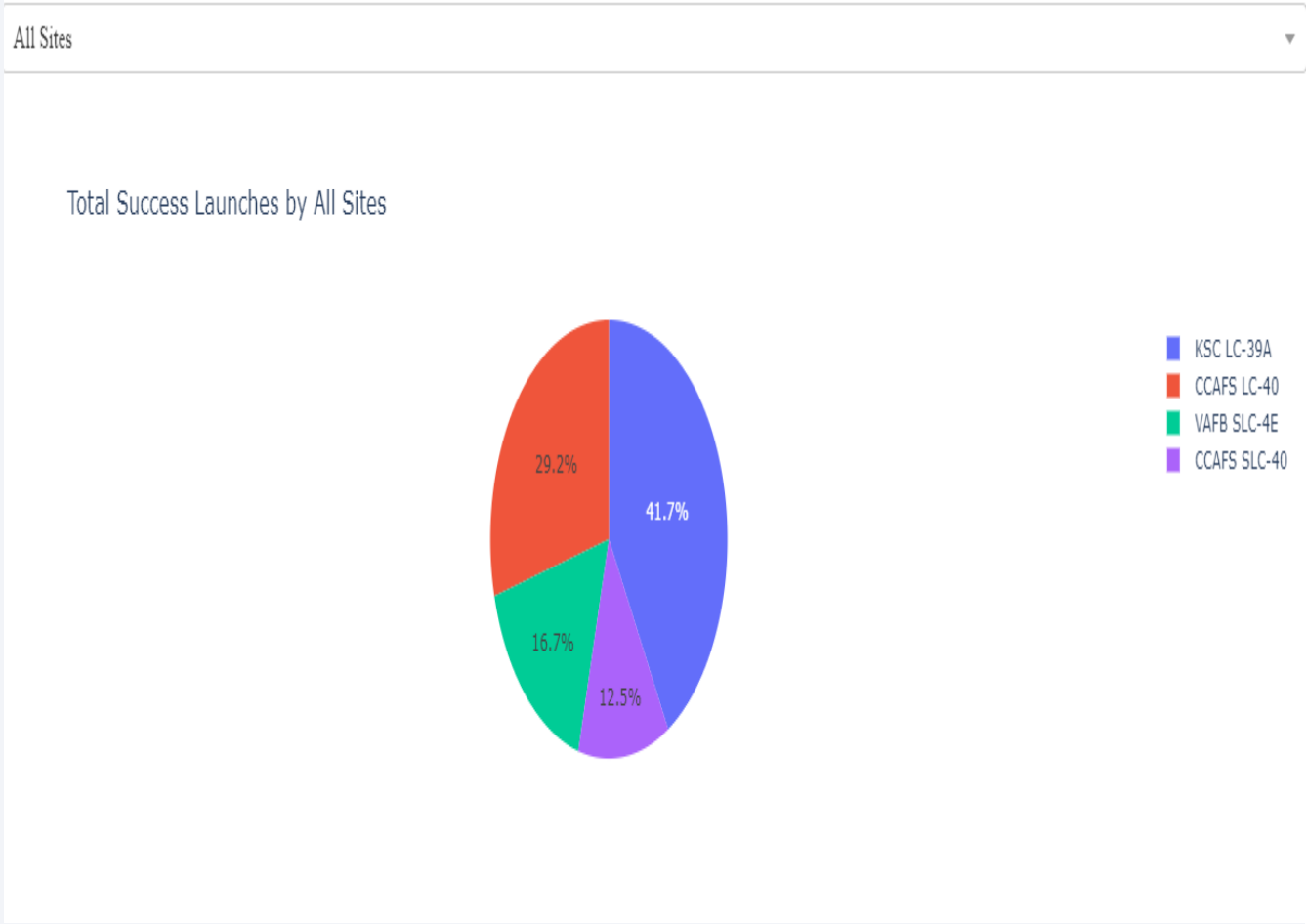
Geographical layout and proximity analysis of CCAFS SLC-40 and LC-40 Launch sites

- The image depicts a map highlighting the locations and proximities of two launch sites labeled CCAFS SLC-40 and LC-40. Key details include:
- **Green Marker (7):** Likely represents successful launches from the CCAFS SLC-40 site.
- **Yellow Marker (26):** Probably denotes the total number of launches, with color variations possibly indicating different outcomes or statuses (such as successful or failed launches).
- The map also illustrates the sites' proximity to significant infrastructure:
- **Proximity to the Coast:** The coastline is marked 0.90 km from the launch sites.
- **Nearby Roadways:** Samuel C Phillips Pkwy is indicated nearby, suggesting ease of access and logistical considerations.
- Circles around the launch sites may denote zones of interest or safety areas, although their specific meanings are not clearly detailed in the image. Overall, the map provides a visual representation of the geographical layout and distances, helping to understand logistics and safety considerations related to the launch activities at these sites.


Section 4
Build a Dashboard with Plotly Dash

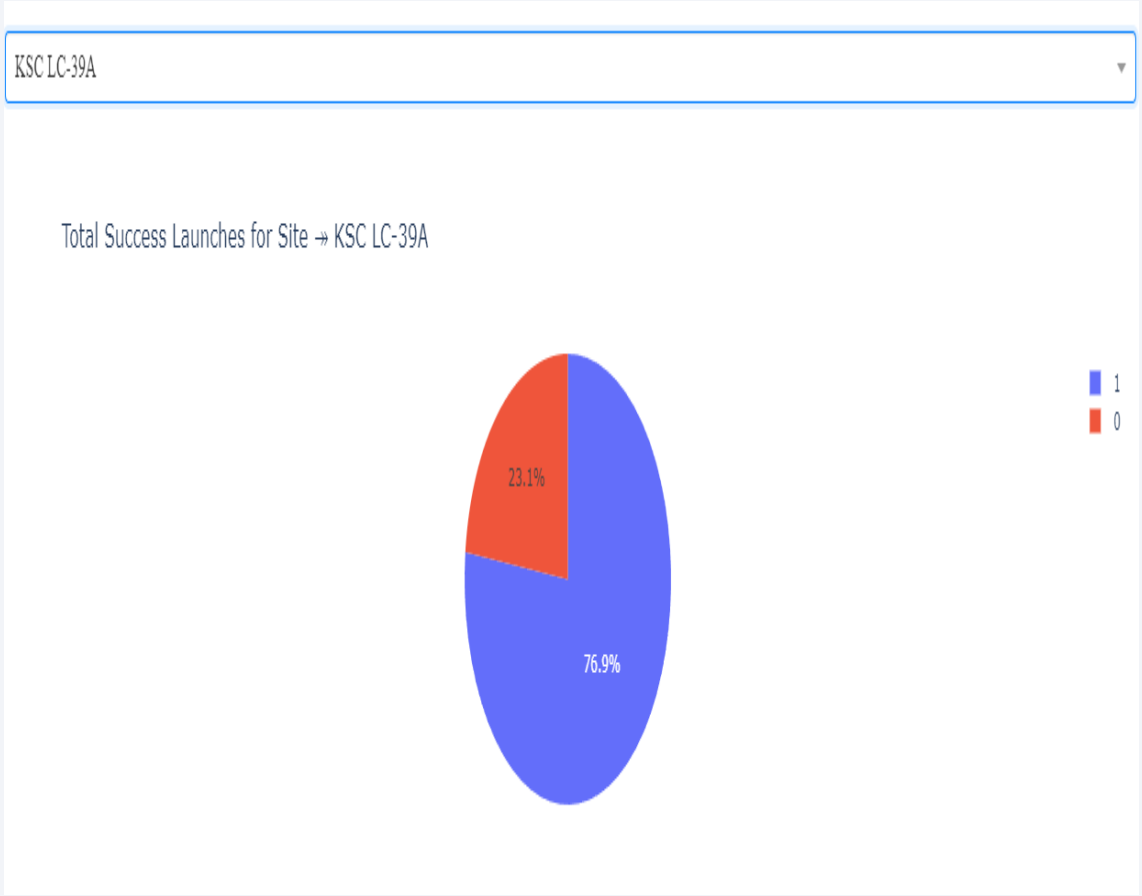
Total success rate by all sites

- The pie chart illustrates the percentage distribution of total successful launches among four sites: KSC LC-39A (41.7%), CCAFS LC-40 (29.2%), VAFB SLC-4E (16.7%), and another CCAFS SLC-40 segment (12.5%). KSC LC-39A holds the largest share of successful launches.



KSC LC-39A has highest launch rate

- KSC LC-39A has the highest launch success rate with 76.9 %



Payload VS Launch outcome

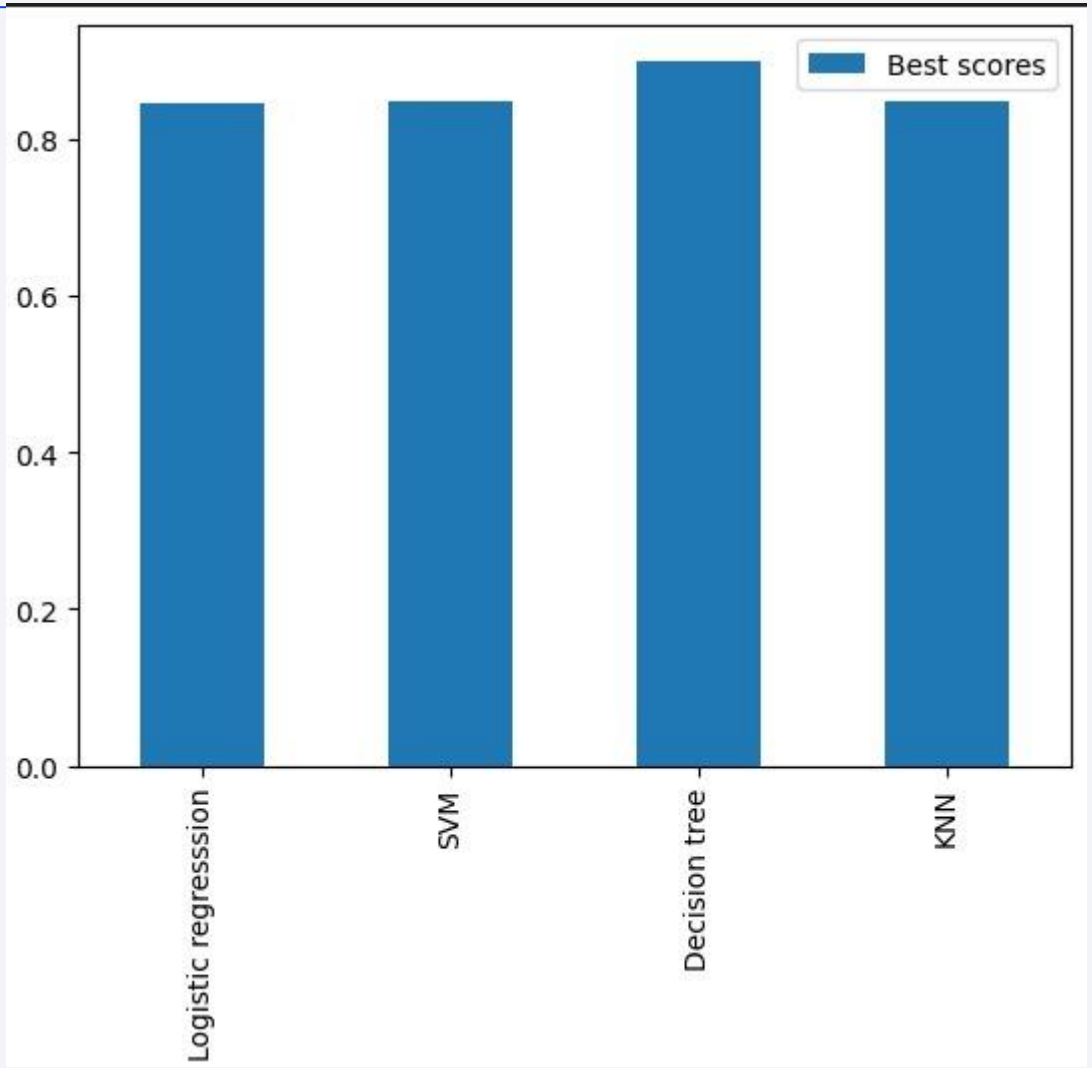
The chart shows the success rate compared to payload mass FT has high success rate with payload mass below 6000kg

Section 5

Predictive Analysis (Classification)

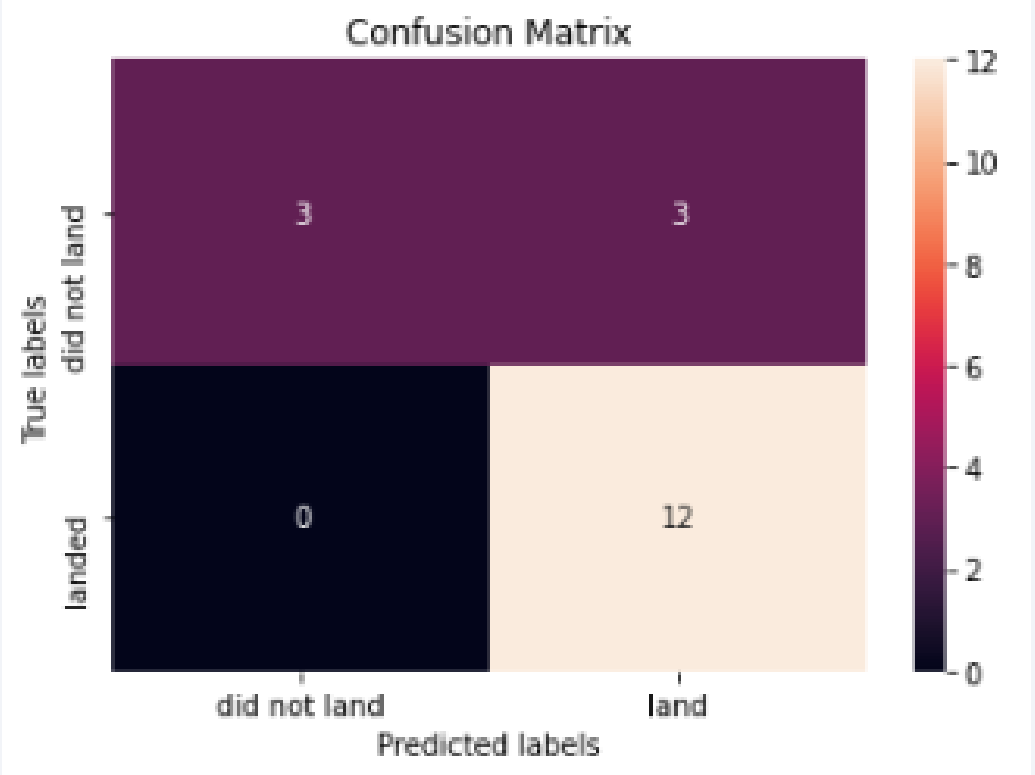
Classification Accuracy

- Visualized all the models via bar graph and took the observation that Decision tree has the highest accuracy



Confusion Matrix

- The confusion matrix for the decision tree model indicates the highest accuracy, with the primary issue being false positives.



Conclusions

- The Decision Tree Model is the most effective algorithm for this dataset.
- Launches with lighter payloads tend to have better outcomes compared to heavier payloads.
- Most launch sites are located near the Equator and close to the coast.
- The success rate of launches has increased over the years.
- KSC LC-39A has the highest success rate among all launch sites.
- The orbits ES-L1, GEO, HEO, and SSO have achieved a 100% success rate.

Thank you!