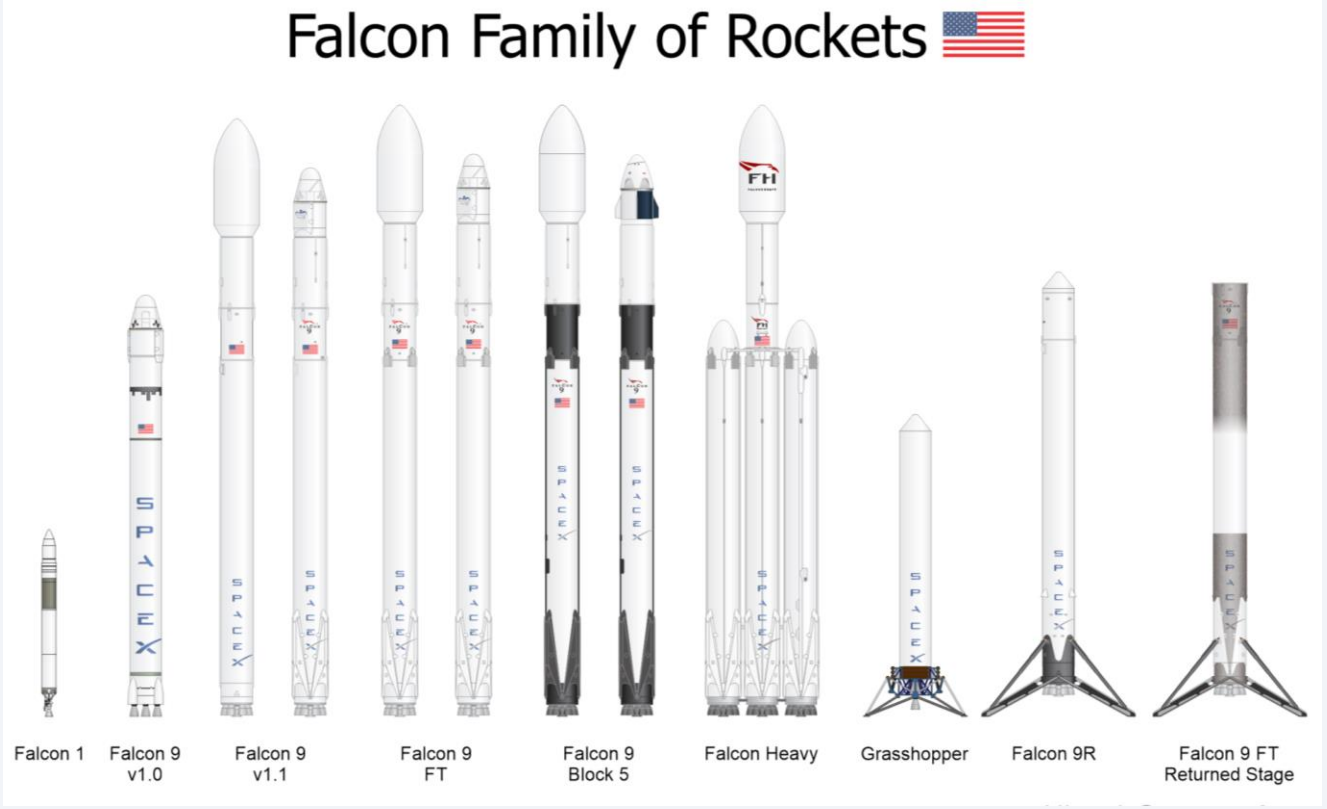
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nokutenda Saungweme
12/02/2024

Table of Contents



Executive Summary

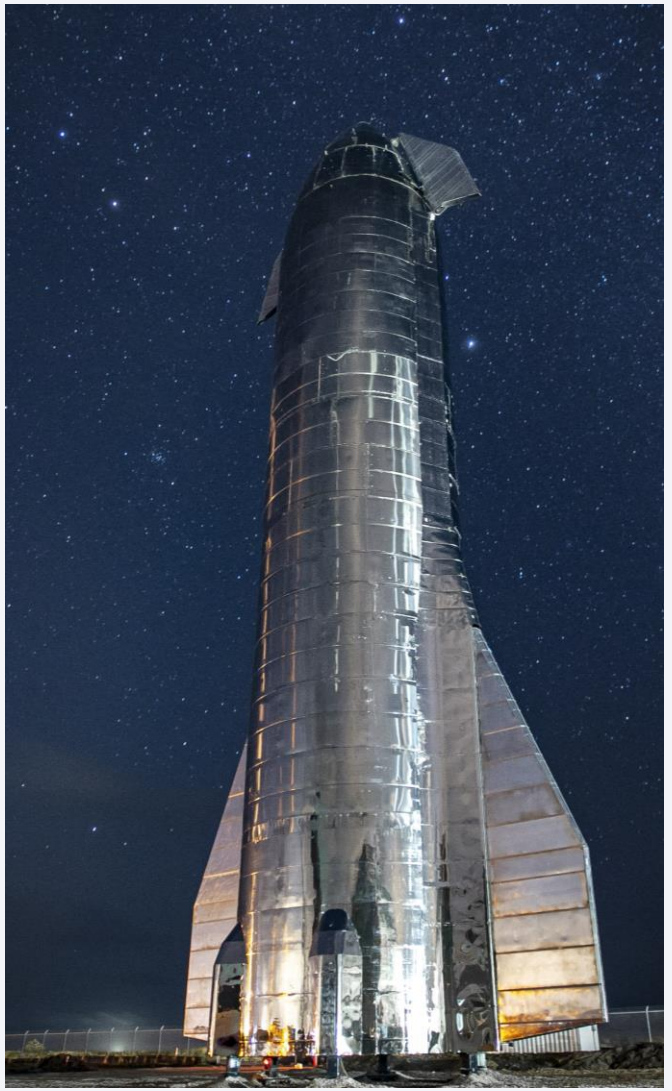
The objective of this data science project is to predict if the SpaceX Falcon 9 first stage will land successfully. By leveraging advanced data analytics techniques and machine learning algorithms, we aim to extract meaningful insights from the available data to drive data-informed decision-making, optimize processes, and we can determine the cost of a launch as much of the savings is because SpaceX can reuse the first stage

Summary of methodologies

- Data Collection using REST API and Web scrapping
- Data Wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building a Dashboard with Plotly Dash
- Predictive Analysis

Summary of all results

- Exploratory Data Analysis Report
- Feature Engineering Documentation
- Trained Machine Learning Models
- Model Evaluation Report



Introduction

SpaceX 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 SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Problems we want to find answers to:

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- What is the best algorithm that can be used for binary classification?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering Data
  - Creating Outcome label
  - Dealing with missing data
  - One Hot Encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using GridSearchCV to find best Hyperparameters for SVM, Classification Trees and Logistic Regression models

Data Collection

The data collection process involved incorporating two different sources of information. API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia page.

By using more than one source of data, we can get a more detailed analysis of the launches and build more effective models.

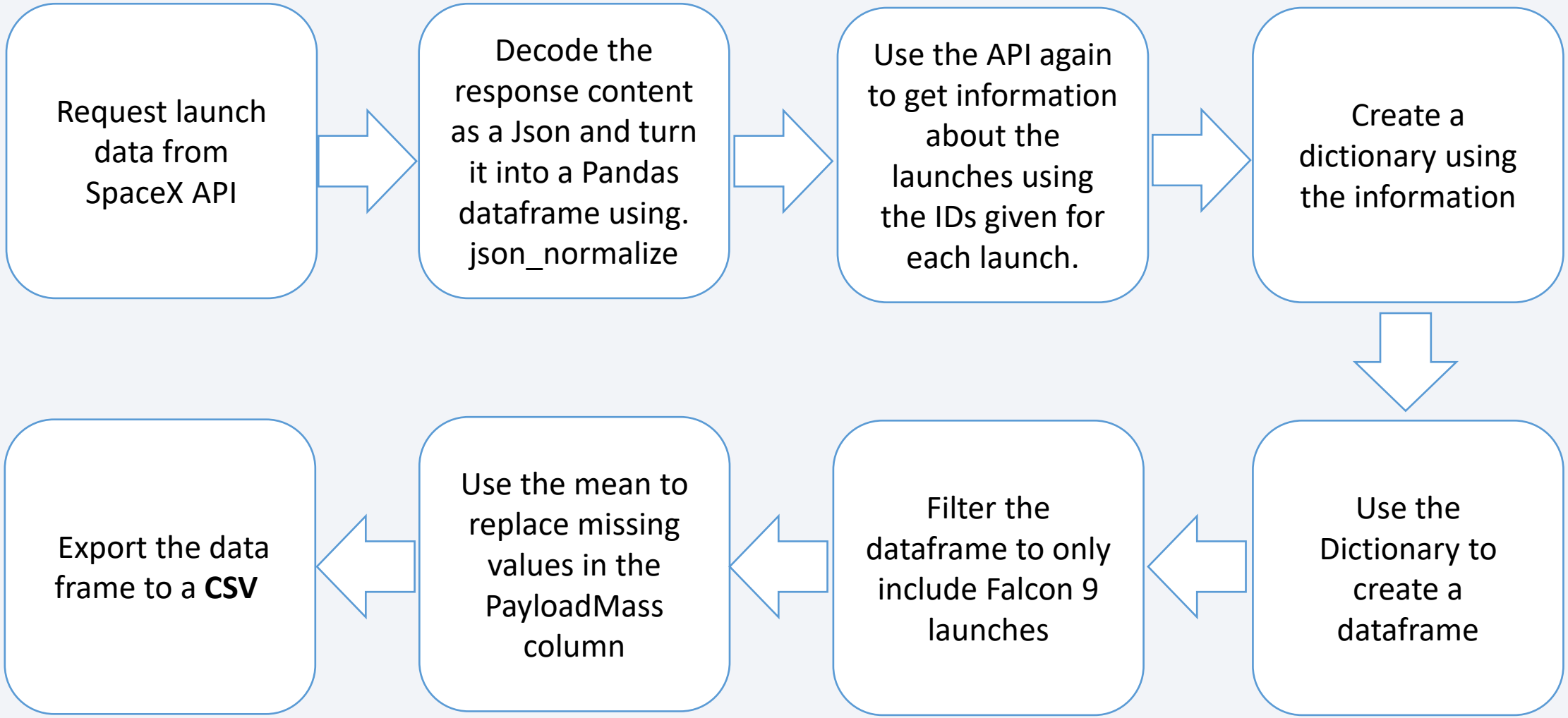
- Data Columns are obtained by using SpaceX REST API:

  FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

- Data Columns are obtained by using Wikipedia Web Scraping:

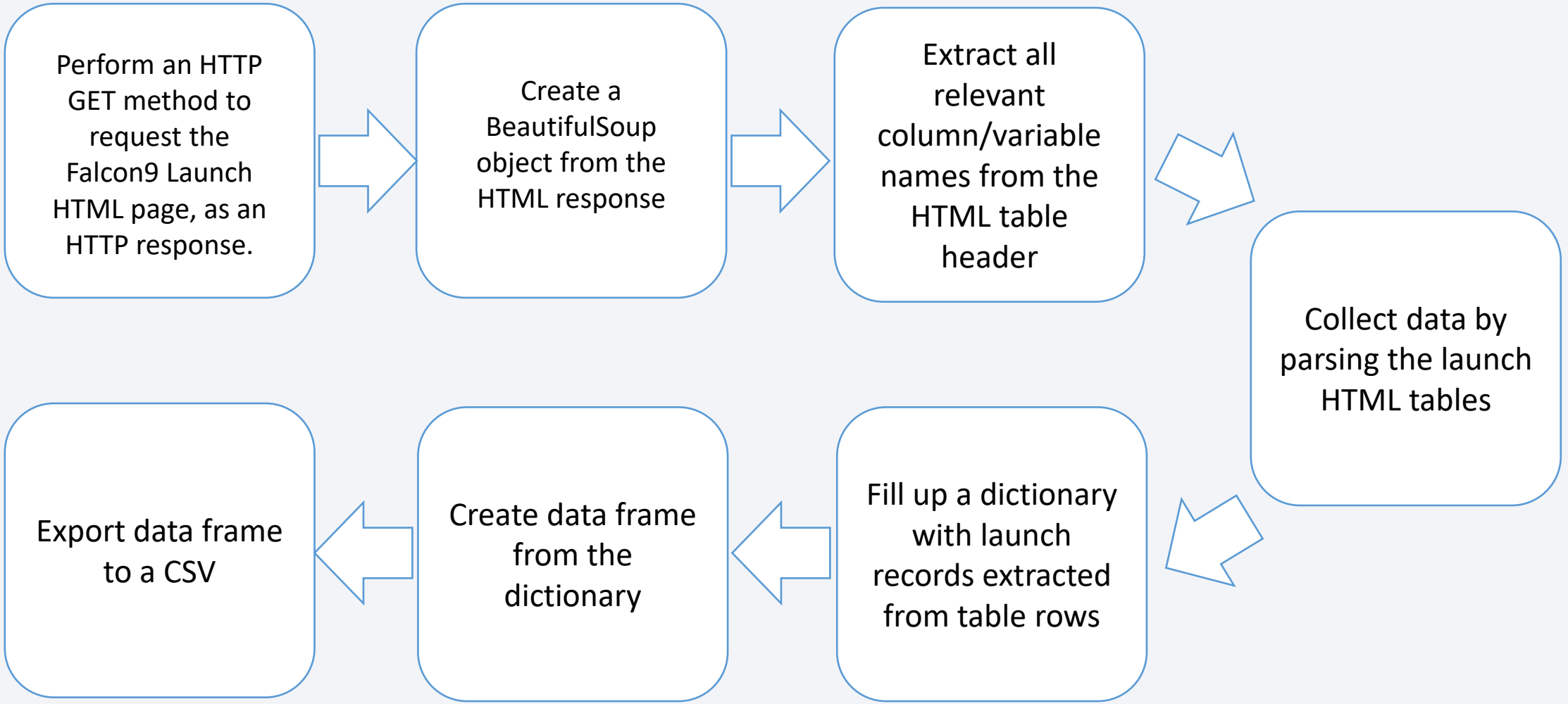
  Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



[Github – SpaceX API]

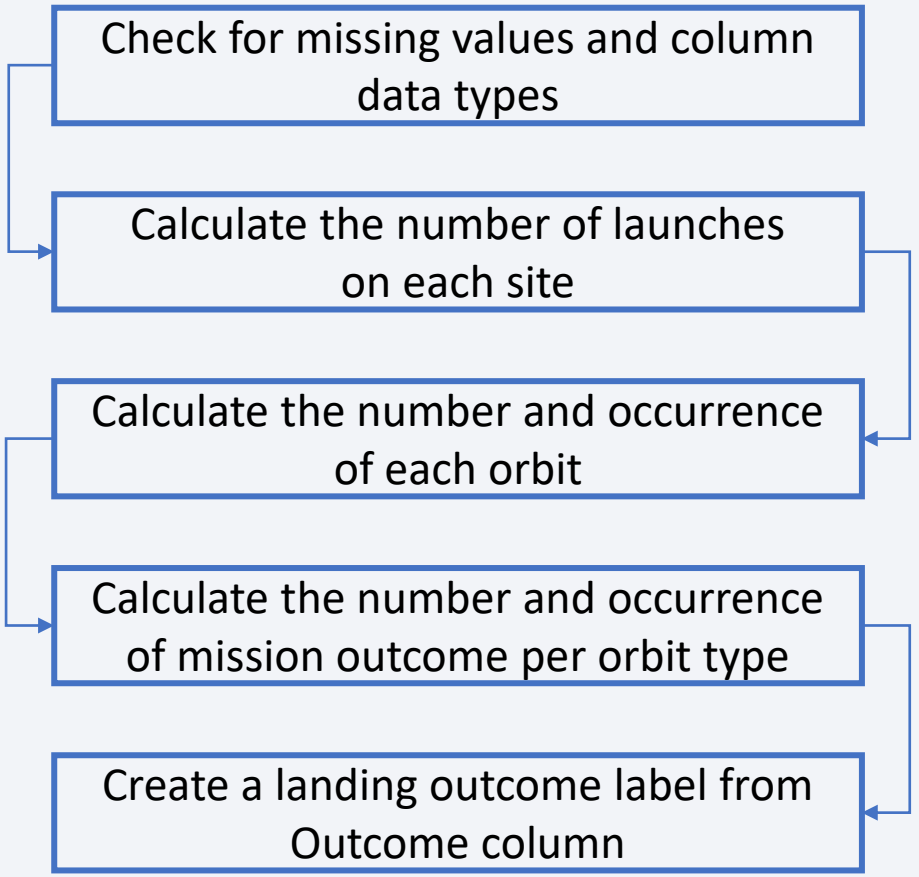
Data Collection - Scraping



Github URL: Web Scrapping

Data Wrangling

The goal of the data wrangling was to conduct the initial exploratory data analysis and determine the training labels. The data analysis involved checking for missing values and data types of different columns. It was followed up by doing a value count check based on the LaunchSite, the Orbit, and the Outcome column. Using the result for the Outcome column value counts, a label encoding was performed. In the raw data there are several different cases for failed landings and successful landings i.e. True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed on a ground pad False RTLS means the mission outcome was unsuccessfully landed on a ground pad. There was a need to separate them into good outcomes and bad outcomes and then encode them such that 0 corresponded to a bad outcome and 1 corresponded to a good outcome and assign this to the Outcome label.



Github URL: Data Wrangling

EDA with Data Visualization

Scatter Plots: FlightNumber vs. PayloadMass, FlightNumber vs LaunchSite, Payload and Launch Site, FlightNumber and Orbit type, Payload and Orbit type.

Reason: We used the scatter plots to try and identify possible relationships between the changes observed between the different variables.

Bar Charts: success rate of each orbit type

Reason: Bar charts make it easy to view discrete variables and compare them. In this case the Orbit types were compared in terms of successful landing rates.

Line Graph: launch success yearly trend

Reason: The line graph makes it easier to view the change of a quantity over a given. In this case we want to see how the success rate of launches changed throughout the years of SpaceX operating.

Github URL: EDA with Data Visualization

EDA with SQL

Queries performed:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA’
- 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 the 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. Use a subquery
- List the records that 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 dates 2010-06-04 and 2017-03-20, in descending order.

Github URL: EDA with SQL

Build an Interactive Map with Folium

Markers of all Launch Sites:

- A marker is created using Latitude and Longitude values to show the launch site locations on the map namely CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E.
- A circle is used for each launch site to highlight the area.

Marking Success and Failed Launches:

- adding the launch outcomes for each site, and see which sites have high success rates.
- The markers a color coordinated to show success as green and failure as red

Calculate the distances between a launch site and its proximities:

- a MousePosition is added on the map to get coordinates for a mouse over a point on the map. It is used to easily find the coordinates of any points of interest by placing a mouse point over a location on the map.
- Color-coded lines are added to show the starting point from the launch site to the proximities (e.g. Railways, Highways etc.). The line is annotated with the distance value.

Github URL: Interactive Visual Analytics with Folium

Build a Dashboard with Plotly Dash

Launch site dropdown:

- A dropdown list to enable Launch Site selection to show records from all sites or from specific launch sites. This allows us to look at the bigger picture or narrow the focus down to specific launch sites and get some insights.

Pie Chart showing Success Launches (All Sites/Specific Site):

- Added a pie chart to show the total successful launches count for all sites
- If a specific launch site was selected, the pie chart shows the Success vs. Failed counts for the site

Payload Mass Range Slider:

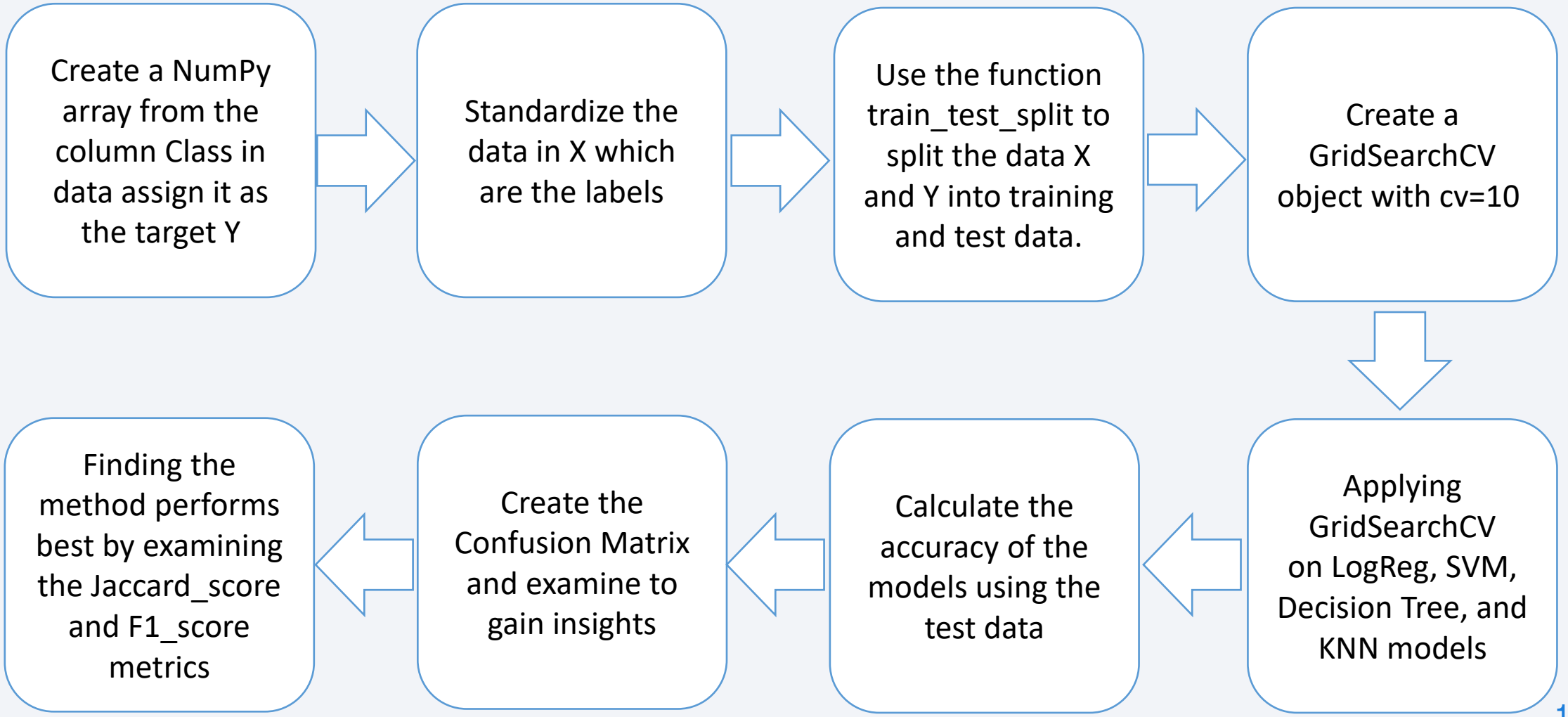
- The slider allows us to select the payload mass range. This changes the information seen in the scatter plot.

Scatter Chart Displaying Relationship between Payload and Launch Success for different Booster Versions:

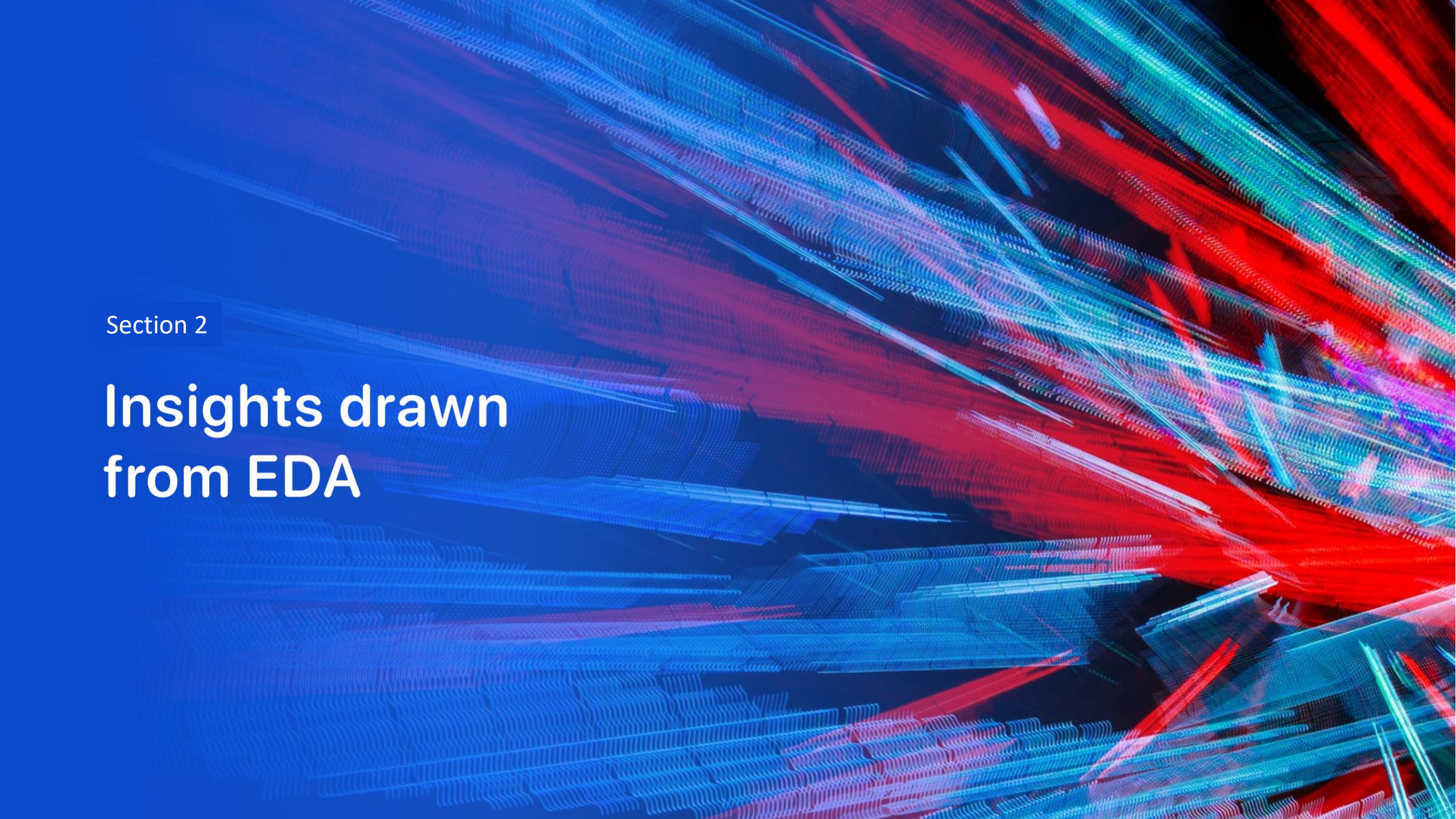
- Added a scatter chart to show the relationship between Payload and Launch Success

Github URL: Plotly Dash App

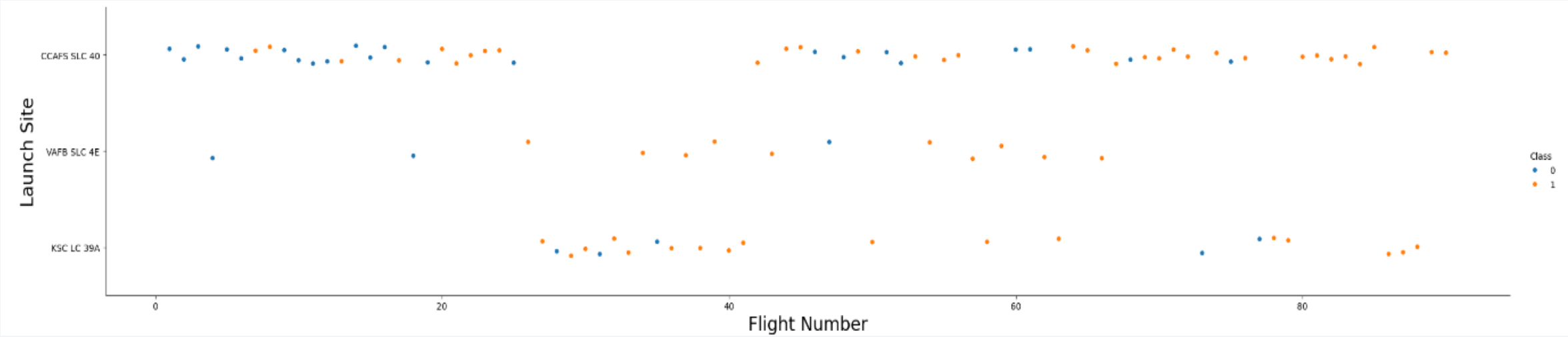
Predictive Analysis (Classification)



Github URL: Predictive Analysis


Section 2
Insights drawn from EDA

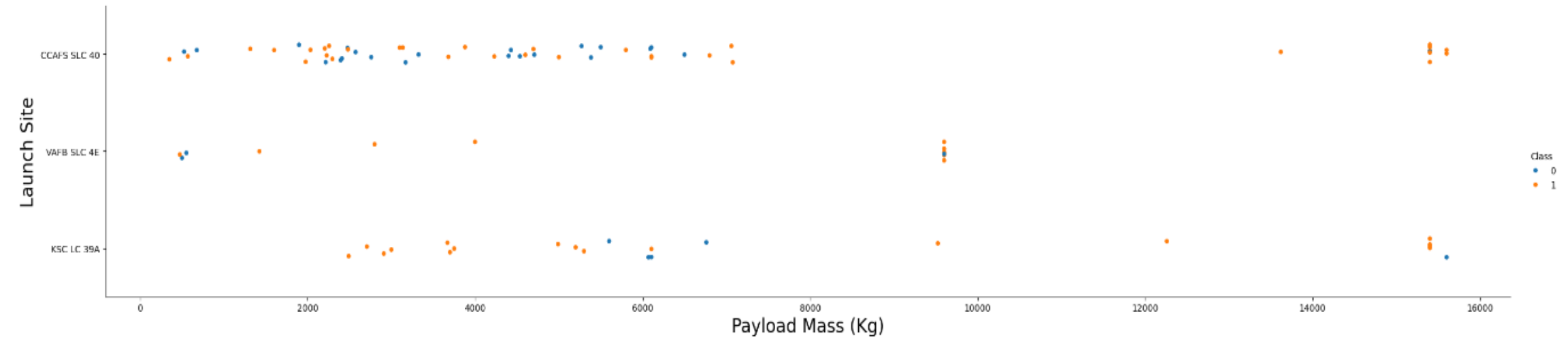
Flight Number vs. Launch Site



Explanation:

- CCAFS SLC 40 has the most launches
- The success rate increased with time across the launch sites
- High number of failures in the first 20 flights
- The latest flights have all been a success

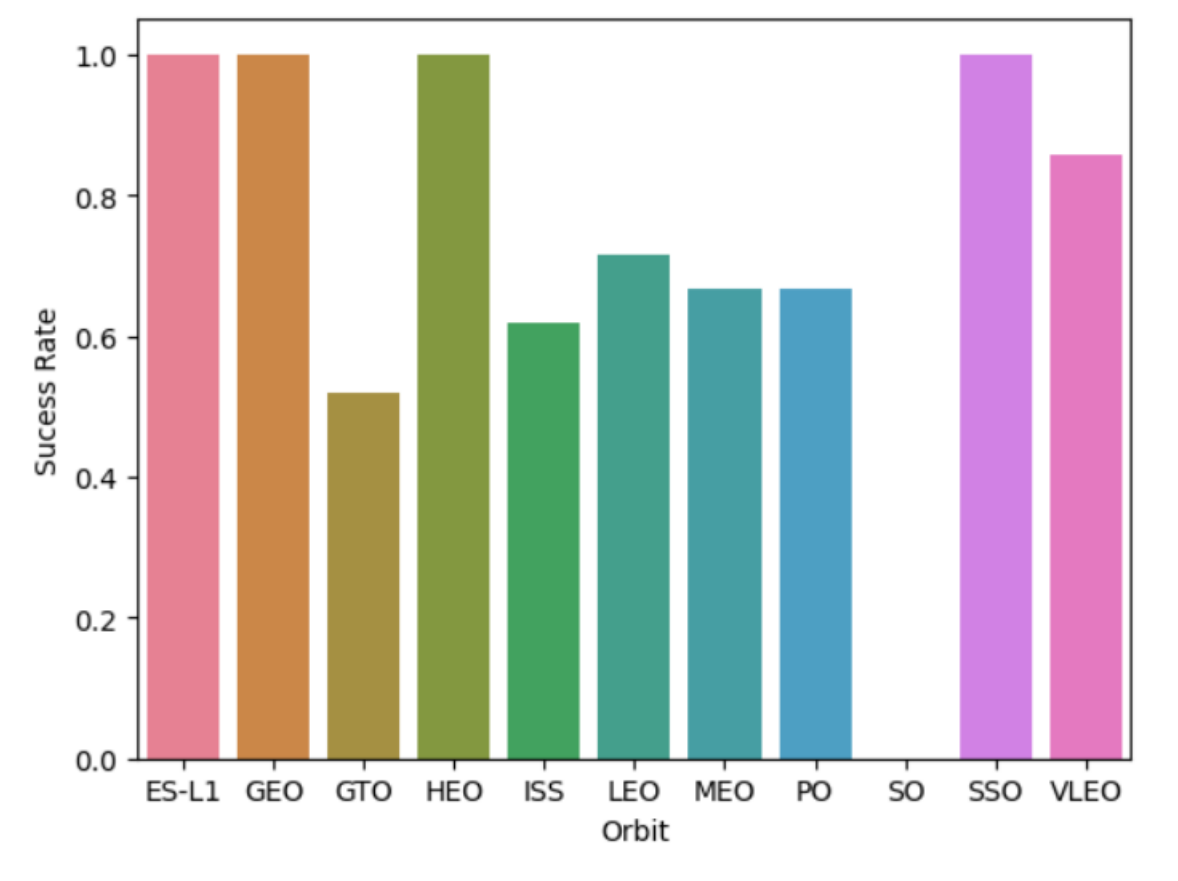
Payload vs. Launch Site



Explanation:

- if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).
- Success rate for payloads above 8000kg is higher than for payloads below 8000kg

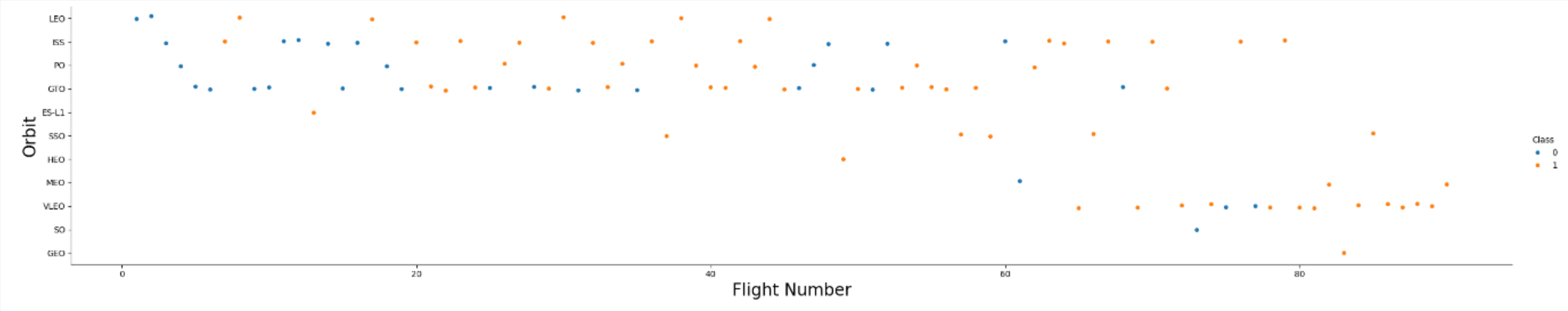
Success Rate vs. Orbit Type



Explanation:

- ES-L1, GEO, HEO and SSO have a success rate of 1
- VLEO has a success rate of about 0.88 which is high.
- The next best is LEO with about 0.7
- SO has a success rate of 0
- GTO, ISS, LEO, MEO, and PO fall within the 0.45-0.75 success rate.

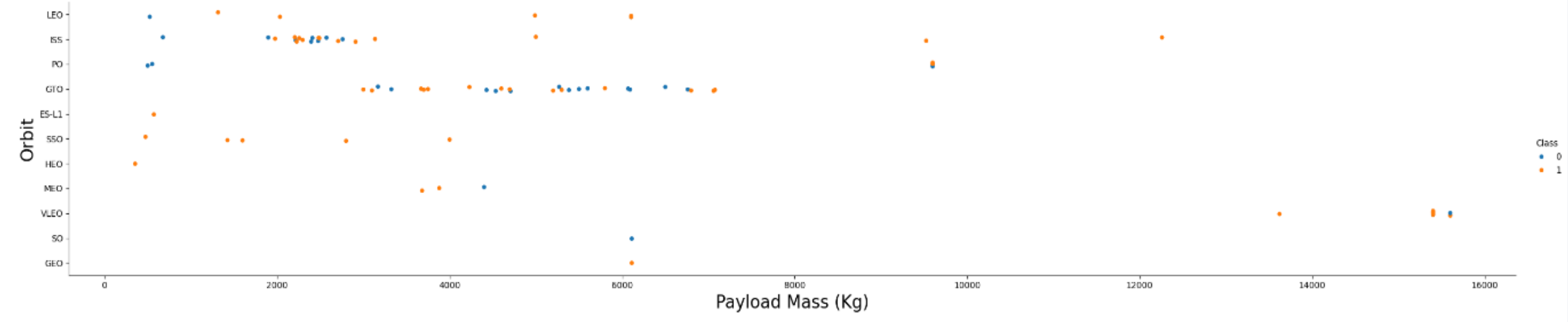
Flight Number vs. Orbit Type



Explanation:

- in the LEO orbit the success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

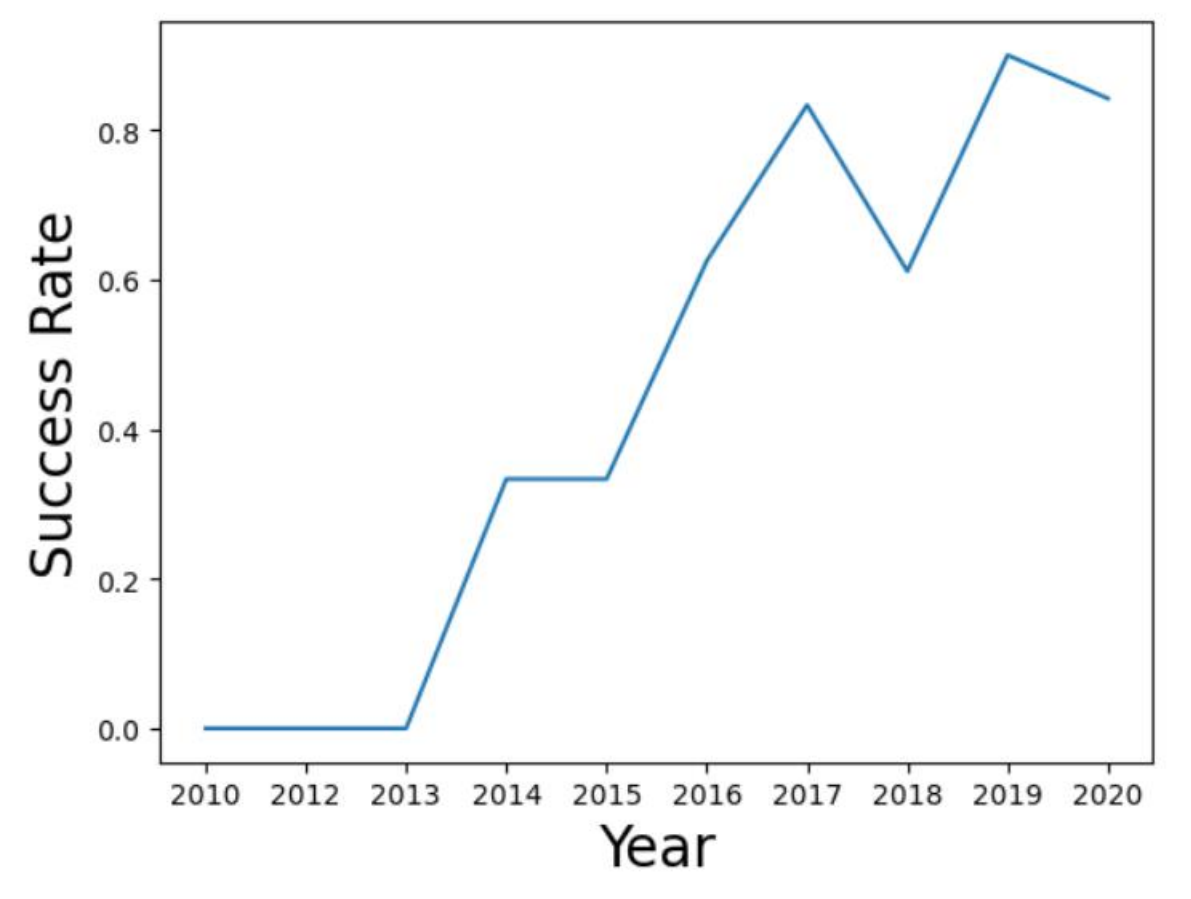
Payload vs. Orbit Type



Explanation:

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.,

Launch Success Yearly Trend



Explanation:

- you can observe that the success rate since 2013 kept increasing till 2020

All Launch Site Names

In [12]:
```
%sql select DISTINCT("Launch_Site") from SPACEXTBL
```

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

Out[12]:

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

Explanation:

- The query is used to find the launch site names

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
```

```
 * 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 the 5 records of the table where launch sites begin with the string 'CCA'

Total Payload Mass

```
%sql select Customer, sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer='NASA (CRS)'
```

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

Customer	sum(PAYLOAD_MASS__KG_)
NASA (CRS)	45596

Explanation:

- Displaying the calculated total payload carried by boosters from NASA

Average Payload Mass by F9 v1.1

```
%sql select Booster_Version ,avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1'
```

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

Booster_Version	avg(PAYLOAD_MASS__KG_)
F9 v1.1	2928.4

Explanation:

- Displaying the calculated average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
%sql select MIN("Date") from SPACEXTBL where Mission_Outcome = 'Success'
```

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

MIN("Date")
2010-06-04

Explanation:

- Listing the date when the first succesful landing outcome in ground pad was acheived.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select Distinct(Booster_Version) from SPACEXTBL where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

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

Explanation:

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
%sql select Distinct(Mission_Outcome), count(*) as Total from SPACEXTBL group by Mission_Outcome
```

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

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

Explanation:

- Calculating the total number of successful and failed mission outcomes

Boosters Carried Maximum Payload

```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ =(select MAX(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

```
 * 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 boosters which have carried the maximum payload mass

2015 Launch Records

```
%sql select substr(Date, 6,2) as month, date, booster_version, launch_site, landing_outcome from SPACEXTBL where landing_outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015'
```

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

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

Explanation:

- Displaying the list the failed landing outcomes in drone ships, their booster versions, and launch site names for in year 2015

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

```
%sql select landing_outcome,count(landing_outcome) as Total from SPACEXTBL where Date between '2010-06-04' and '2017-03-20' group by landing_outcome order by Total desc
```

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

Landing_Outcome	Total
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

Explanation:

- Displaying the rankings of the counts 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

Section 3

Launch Sites Proximities Analysis

Launch Site Locations

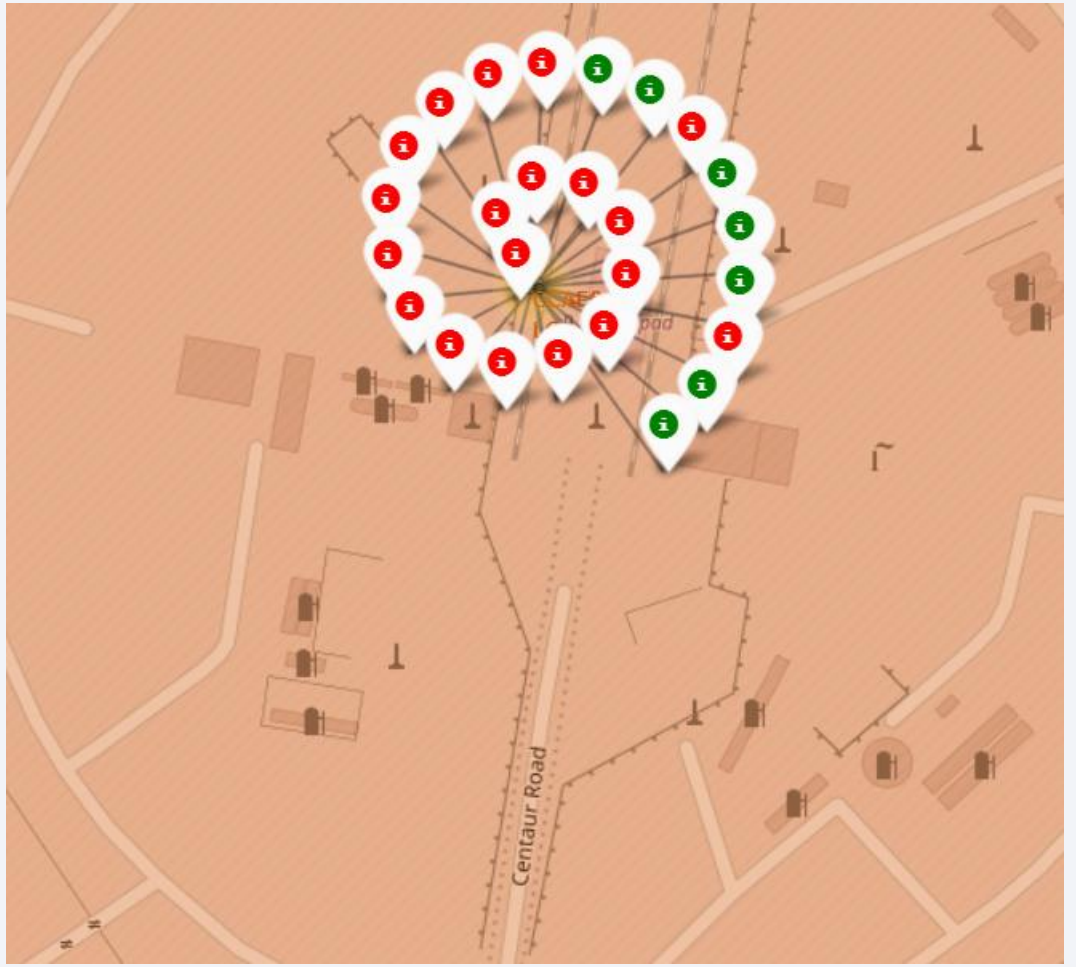
Explanation:

- launch sites are near Earth's equator. It is done to take optimum advantage of the Earth's substantial rotational speed. Sitting on the launch pad near the equator, a spaceship is already moving at a speed of over 1650 km per hour relative to Earth's center.
- The sites are in very close proximity to the coast. Ships are launched toward the ocean to minimize the risk of having any debris dropping or exploding near people.

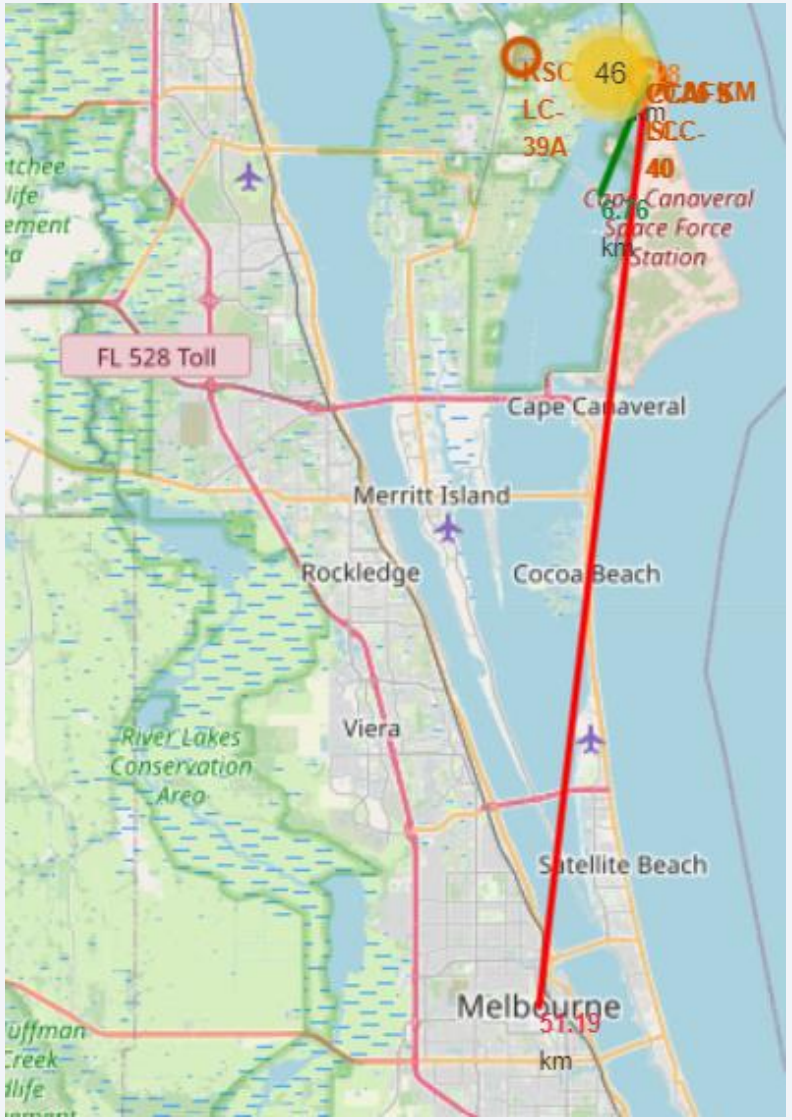
Color-Labeled launch outcomes

Explanation:

- The launch outcomes are color coordinated:
  - Green: Success
  - Red: Failure
- The color-labeled markers make it easy to be able identify which launch sites have relatively high success rates.
- CCAFS SLC-40 has a low success rate

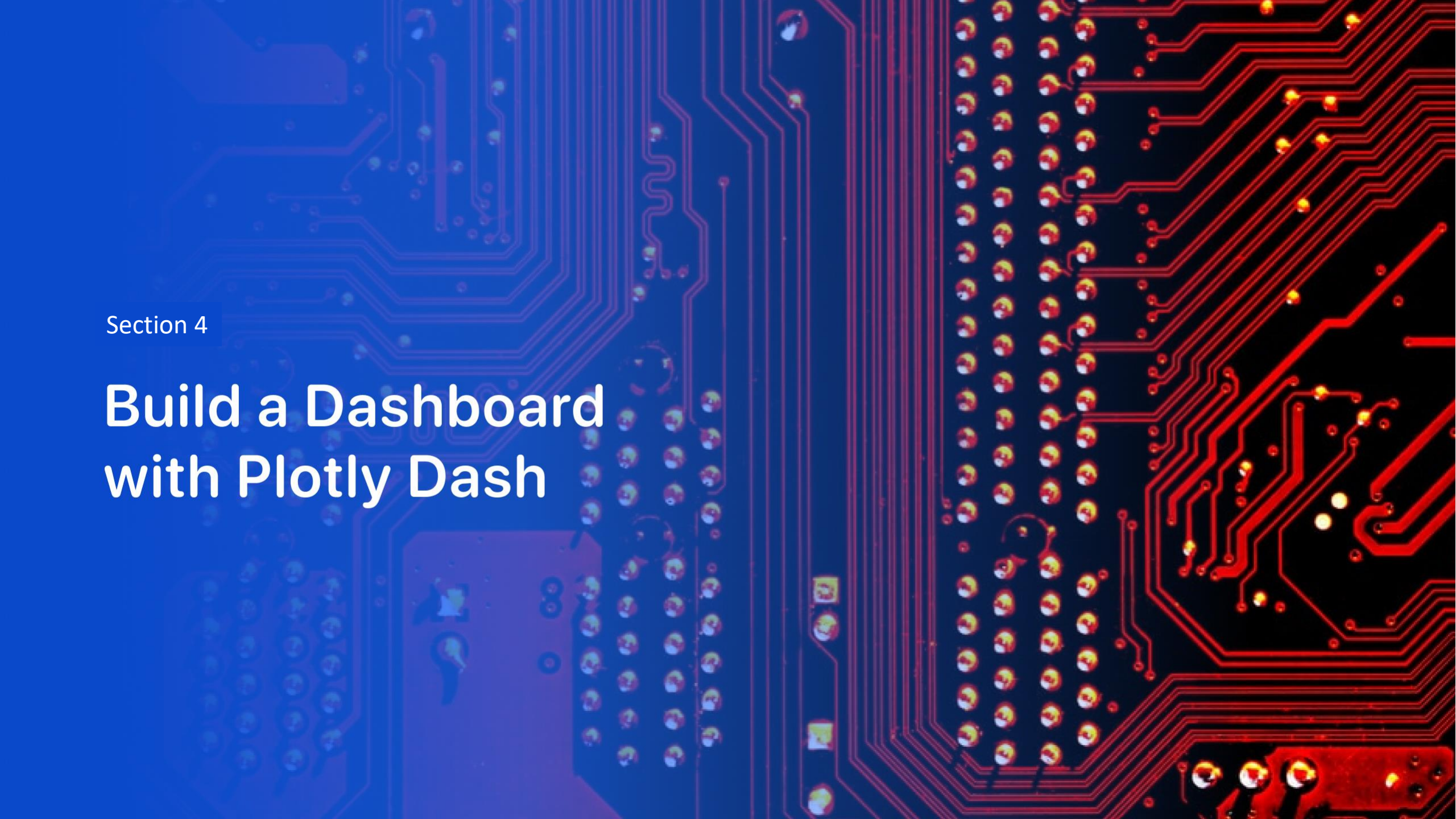


Distance from CCAFS LC-40 to its proximities

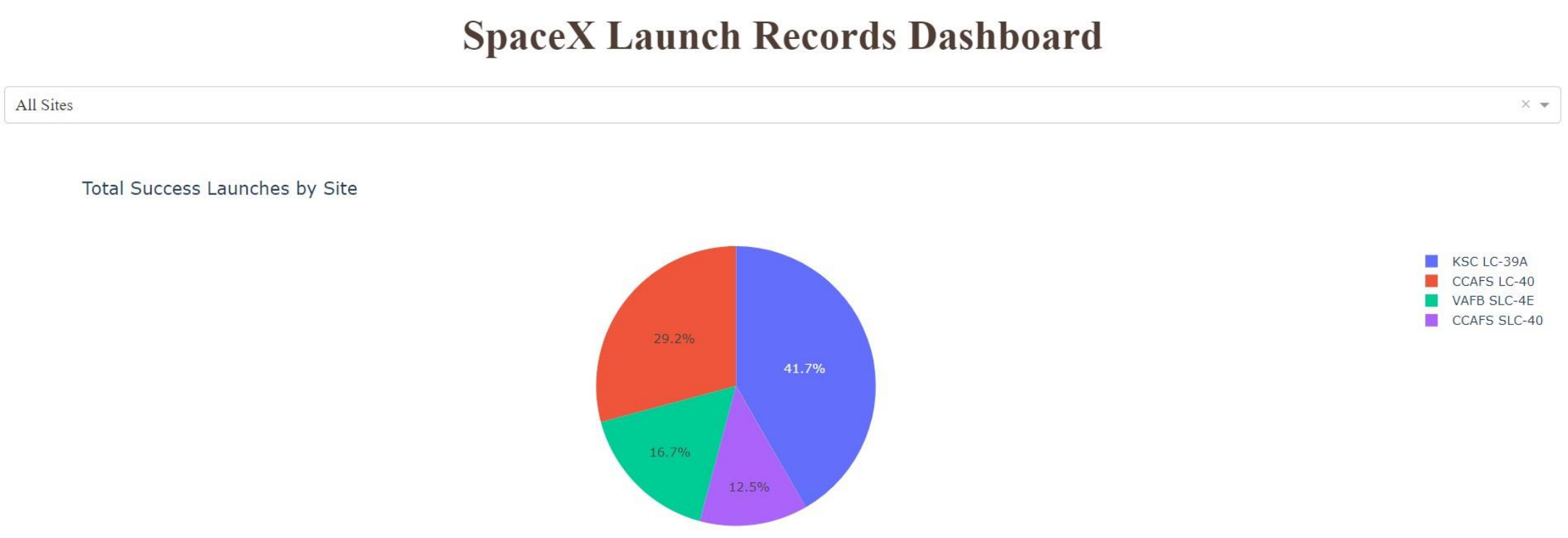


Explanation:

- launch site CCAFS LC-40 is relatively far from its closest major city Melbourne (51.19 Km).
- A rocket with its high speed can cover distances like 15-20 km in a few seconds. It could be potentially dangerous to populated areas
- The launch site is close:
  - to the railway (1.28 km) and a main road (6.76 km) for transport purposes.
  - to the coast (0.86 km) because flying over the ocean, minimizes the risk of having any debris dropping or exploding near people.from


Section 4
Build a Dashboard with Plotly Dash

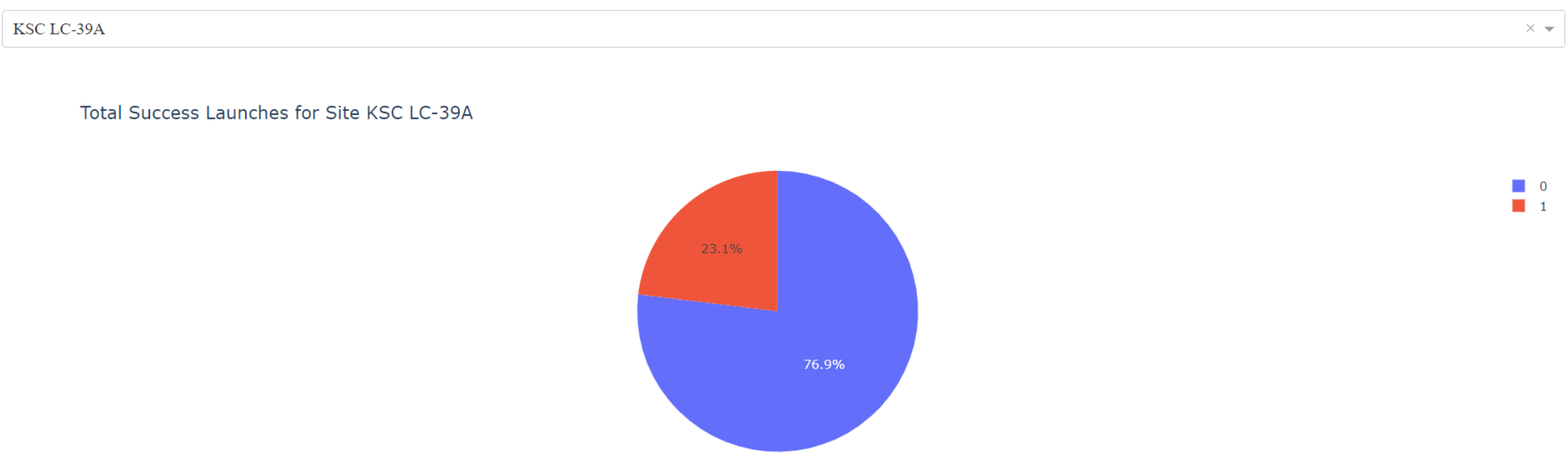
Launch Success Count for All Sites Pie Chart



Explanation:

- From the pie chart, KSC LC-39A has the largest success rates

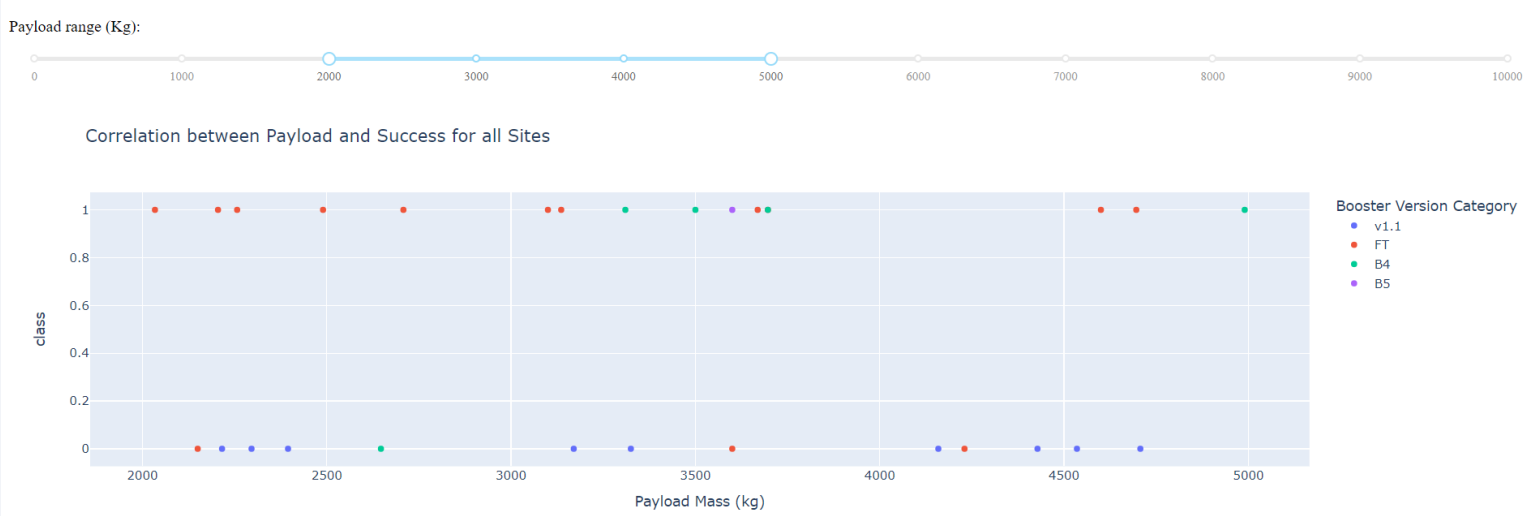
Launch Site with the highest launch success ratio



Explanation:

- KSC LC-39A has the highest success rate for flights with 76.9% of the flights ending in success.

Payload Mass VS Success Rate Scatter Plots



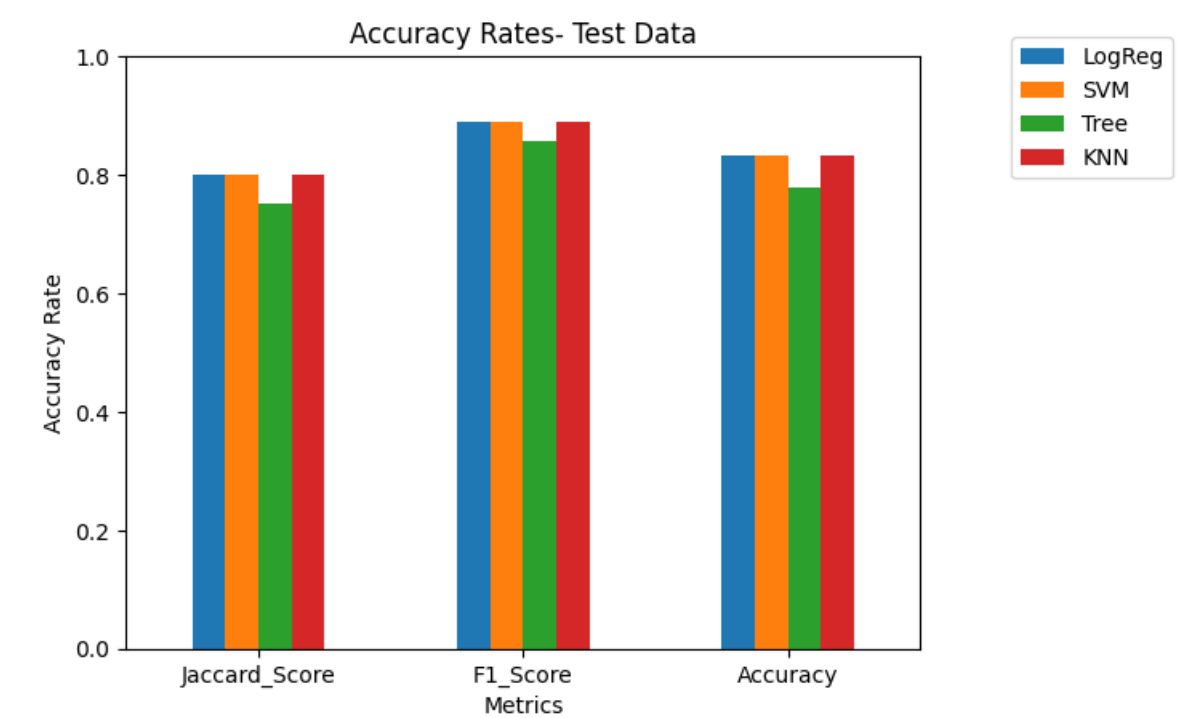
Explanation:

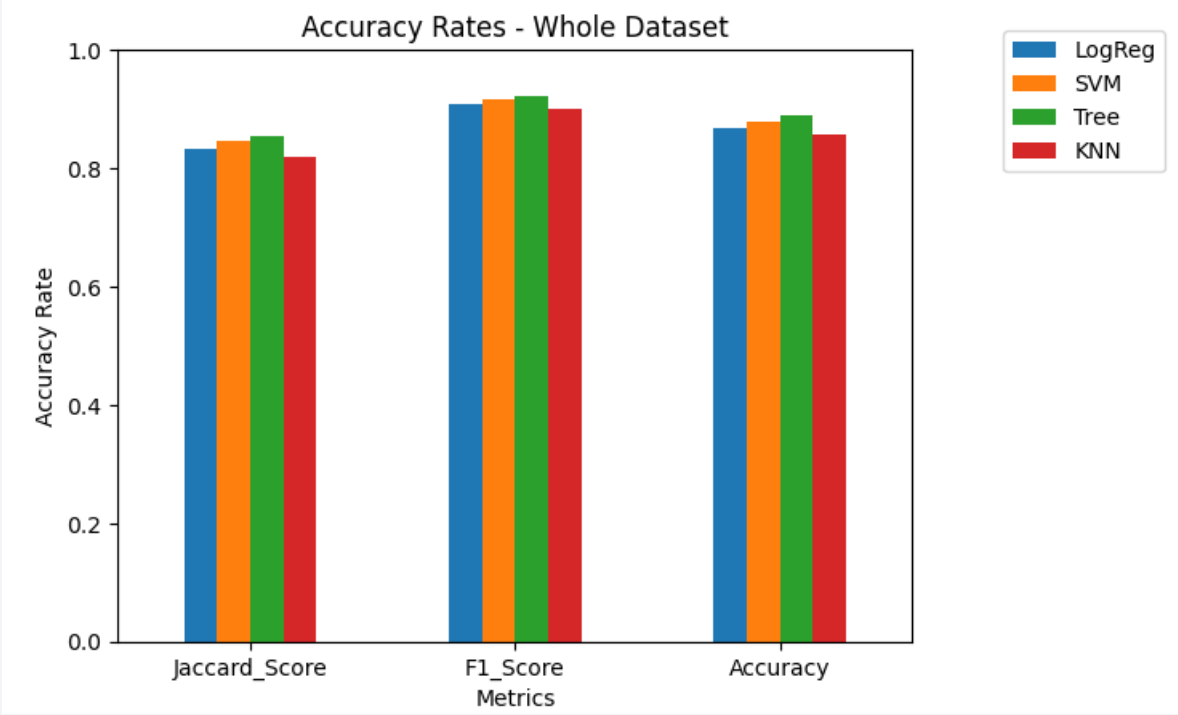
- The scatter plot shows correlation between Payload Mass and Success rate.
- The slider controls the mass range.
- The 2000kg-4000kg mass range has the highest success rate.
- For all masses the FT Booster Version Category has the highest number of successes

Section 5

Predictive Analysis (Classification)

Classification Accuracy

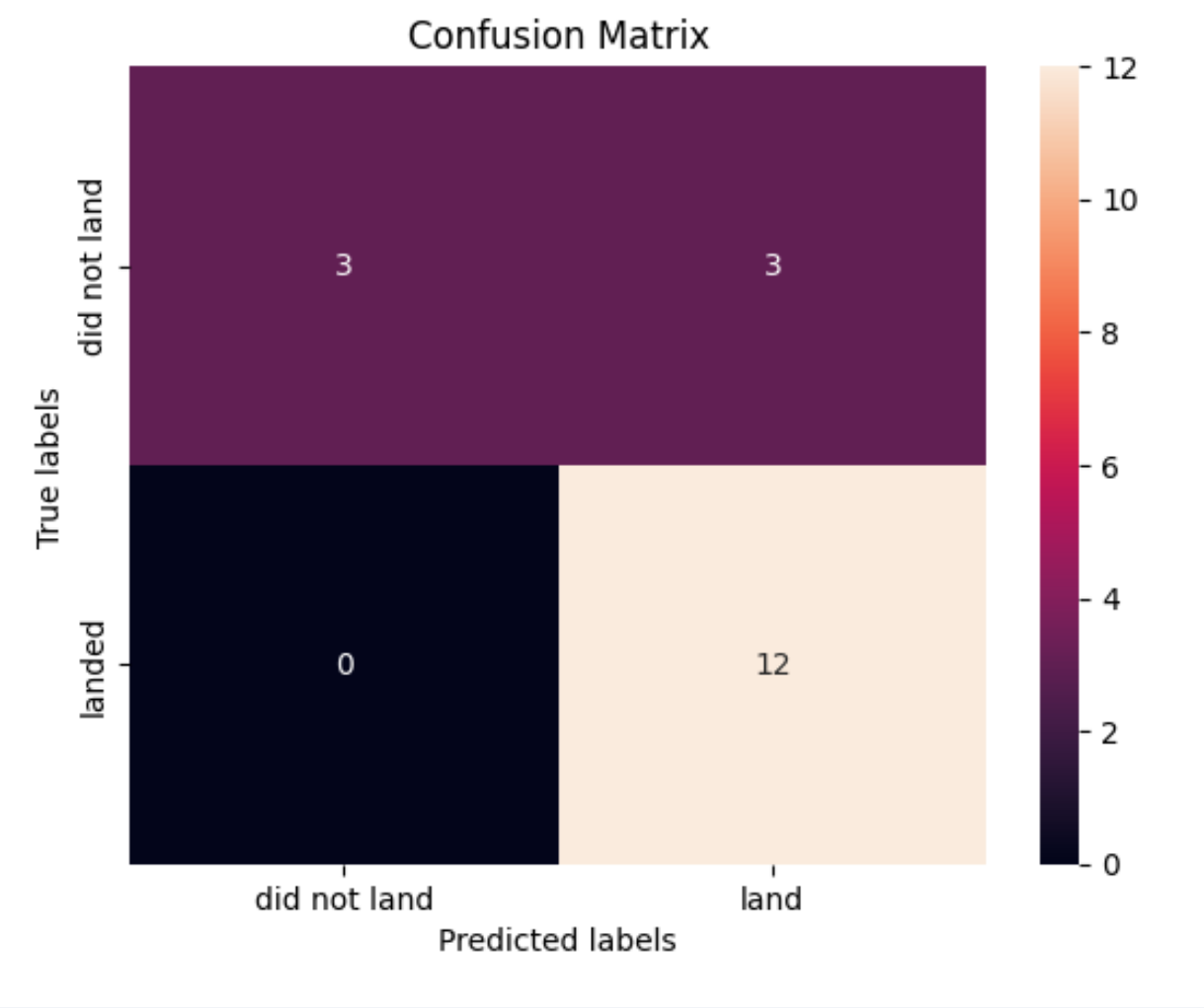




Explanation:

- Using just the test data there is no clear best model. This might be a result of having a small test sample size (18 samples)
- When the whole dataset is used, Decision Tree outperforms the other models

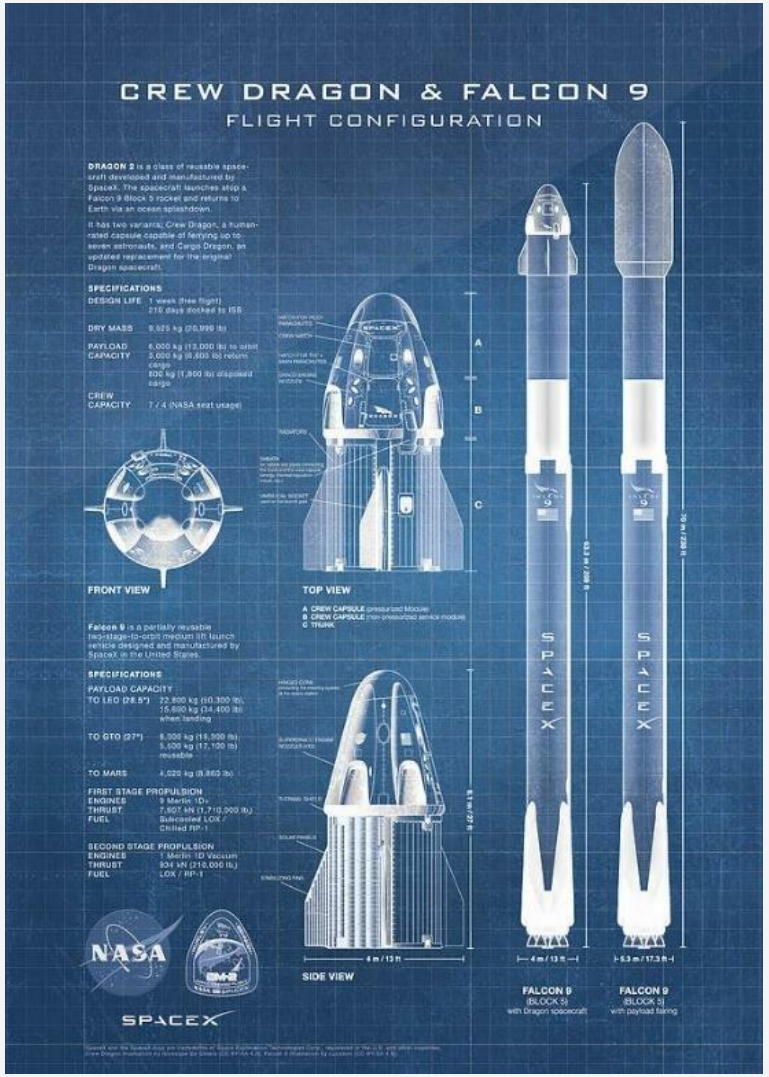
Confusion Matrix



Explanation:

- Examining the confusion matrix, we see that the models can distinguish between the different classes. We see that the major problem is false positives.

Conclusions



- The success rate increased with time across the launch sites
- ES-L1, GEO, HEO and SSO have a success rate of 100%
- Launch sites are near Earth's equator. It is done to take optimum advantage of the Earth's substantial rotational speed.
- KSC LC-39A has the largest success rates.
- The Decision Tree Model outperforms the other models when it comes to this dataset.
- On average Launches with lower payload mass have a higher success rate than those with a higher payload mass.
- The 2000kg-4000kg mass range has the highest success rate.
- For all masses the FT Booster Version Category has the highest number of successes.
- Launch sites are typically closer to railway roads, oceans and major roads. They are further away from major cities for the safety of people in case of launch failure

Thank you!