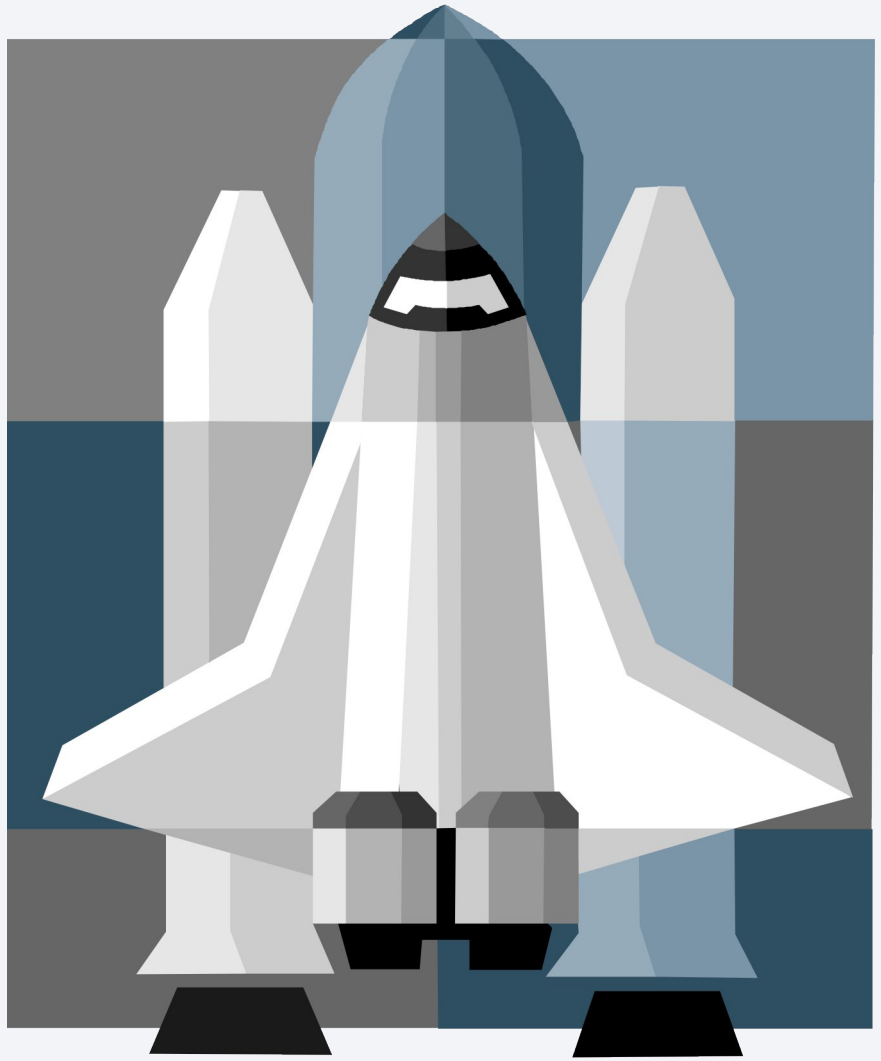
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dimple Rana
September 12, 2025

Outline

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



Executive Summary

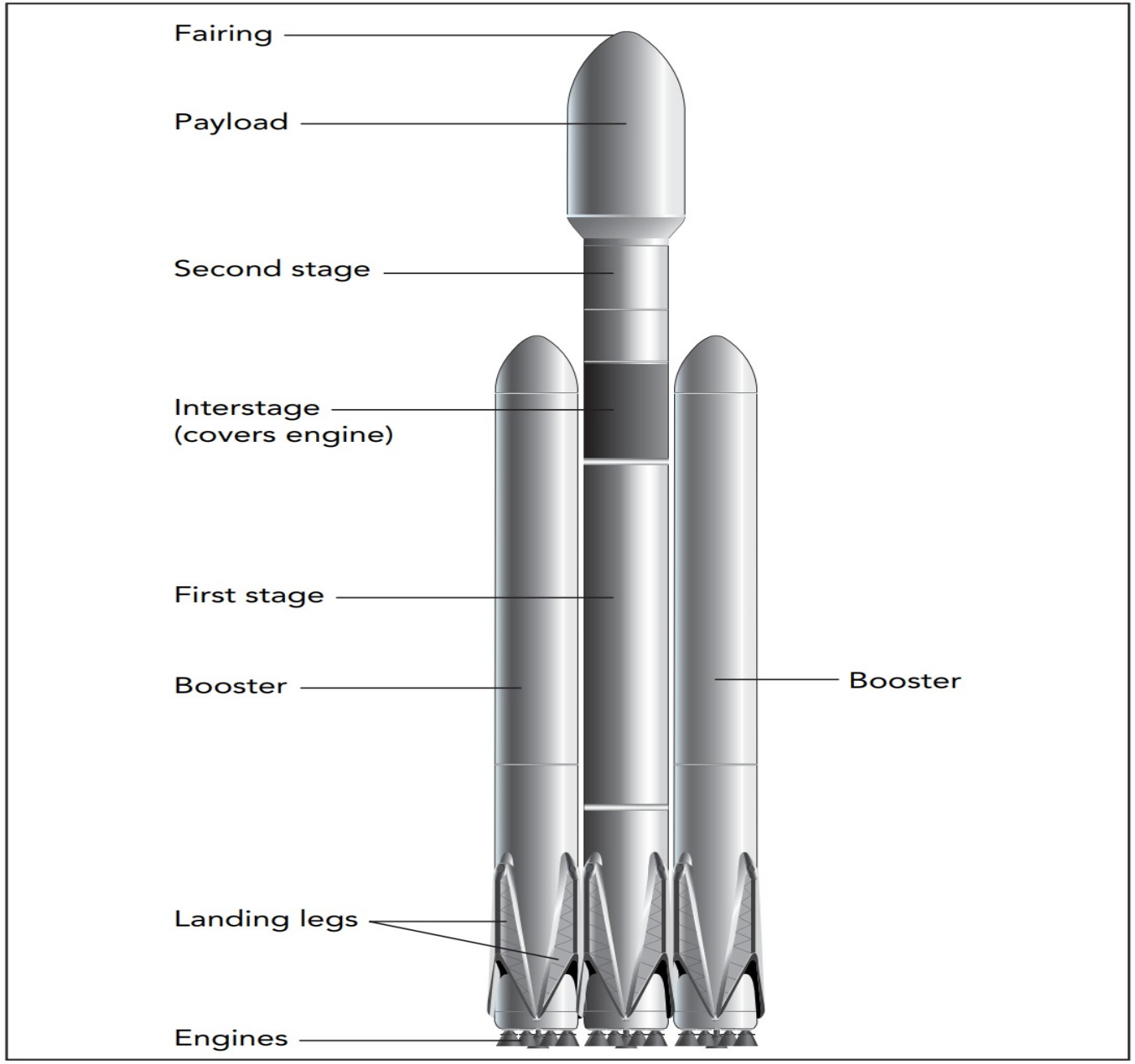
- **Methodologies:**
    1. Collected and preprocessed the SpaceX Falcon 9 launch dataset.
    2. Conducted Exploratory Data Analysis (EDA) to identify key patterns (launch site success rates, effect of payload, booster versions).
    3. Built an interactive dashboard and map using Plotly Dash and Folium for visual analysis.
    4. Developed and evaluated four machine learning models (Logistic Regression, SVM, KNN, Decision Tree) using accuracy and confusion matrix.
- **Results**:
    1. EDA showed that launch success rates vary significantly across sites.
    2. Predictive modeling showed Logistic Regression, SVM, and KNN achieved the highest accuracy (XX%), while Decision Tree performed lower.
- **Conclusion**: Any of the three top models can be chosen depending on context (interpretability, scalability, or simplicity).

Introduction

- The commercial space age is finally here, and the industry is bound to flourish in the upcoming years. In the current era, the major problem businesses face in the space and rocket industry are the overwhelming costs of rocket launches.
- In the past decade a company named SpaceX has reached great heights by advertising their rocket Falcon 9 launches with a cost of 62 million dollars; whereas other providers cost upwards of 165 million dollars each, much of these savings is because SpaceX carefully plans the landing of the first stage of rocket making it viable for reuse.
- Thus, if we can determine if the first stage can land, we can determine the cost of a launch.
- In this project, we understand and gather information about the Falcon 9 launches in the past decades, to understand the conditions required for making the first stage land successfully

Parts of a rocket

- Payload is the major component that is launched in space
- The second stage helps bring the payload to orbit .
- The first stage is the most crucial part of a rocket as it does most of the work. This part is quite large and rather expensive.



Section 1

Methodology

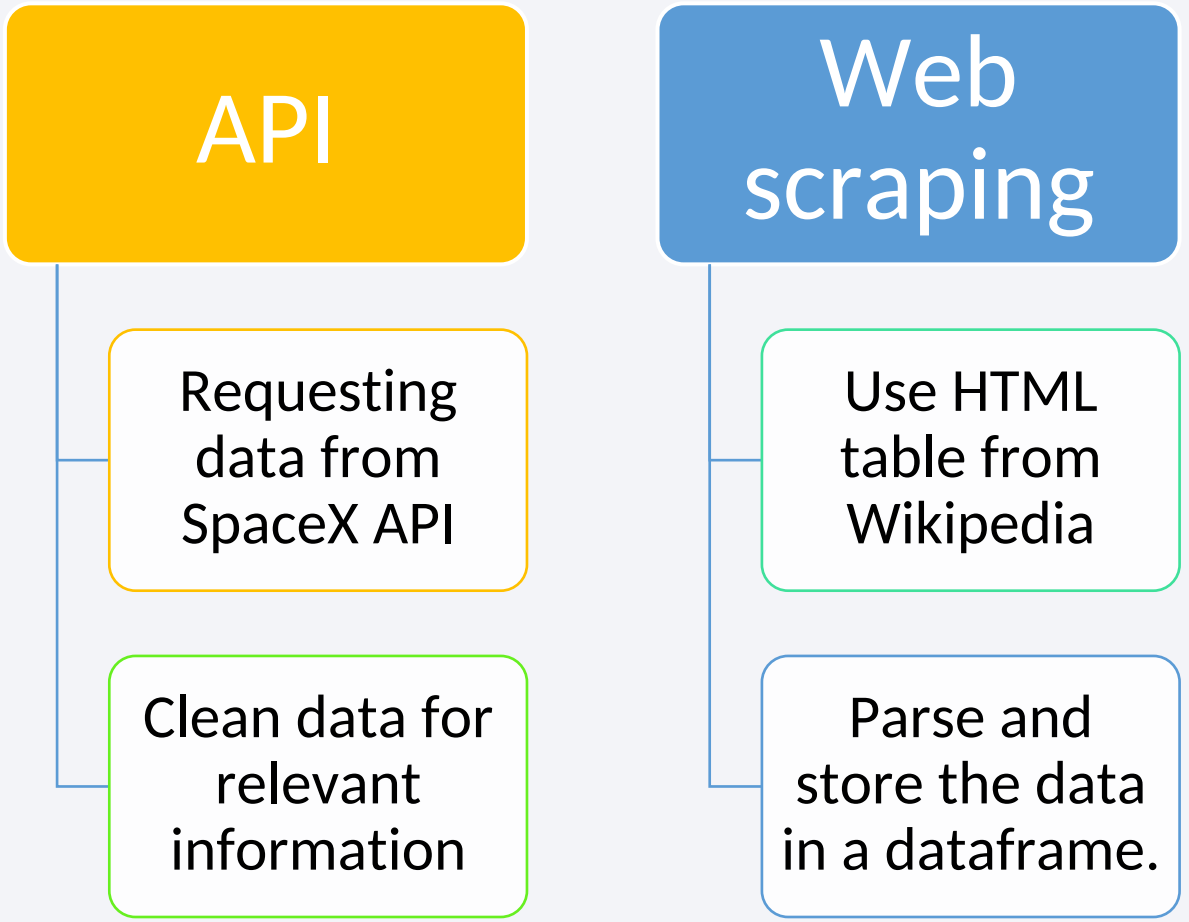
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX launch data gathered from the SpaceX REST API.
  - Web scraping from related Wiki pages.
- Perform data wrangling
  - Processed using an API, Sampling data and dealing with Nulls.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Preprocessing, Standardizing and splitting data into training and testing sets.
  - Train the model and perform Grid Search , to find the best hyperparameters for the

Data Collection

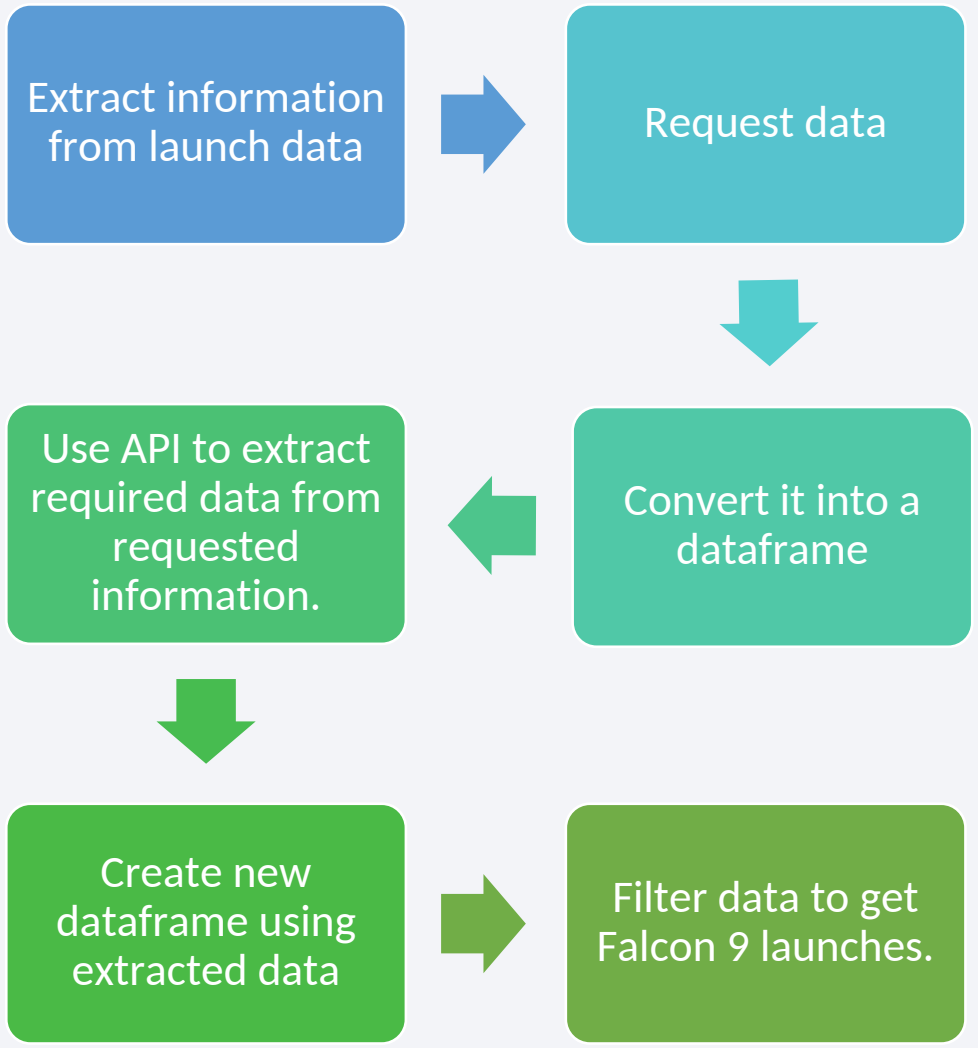
- Two methods were used to collect adequate and beneficial data for the data science project.
- API – The SpaceX API was used to retrieve structured and well-defined data.
- Web Scraping – An HTML table was used to retrieve additional information.



Data Collection – SpaceX API

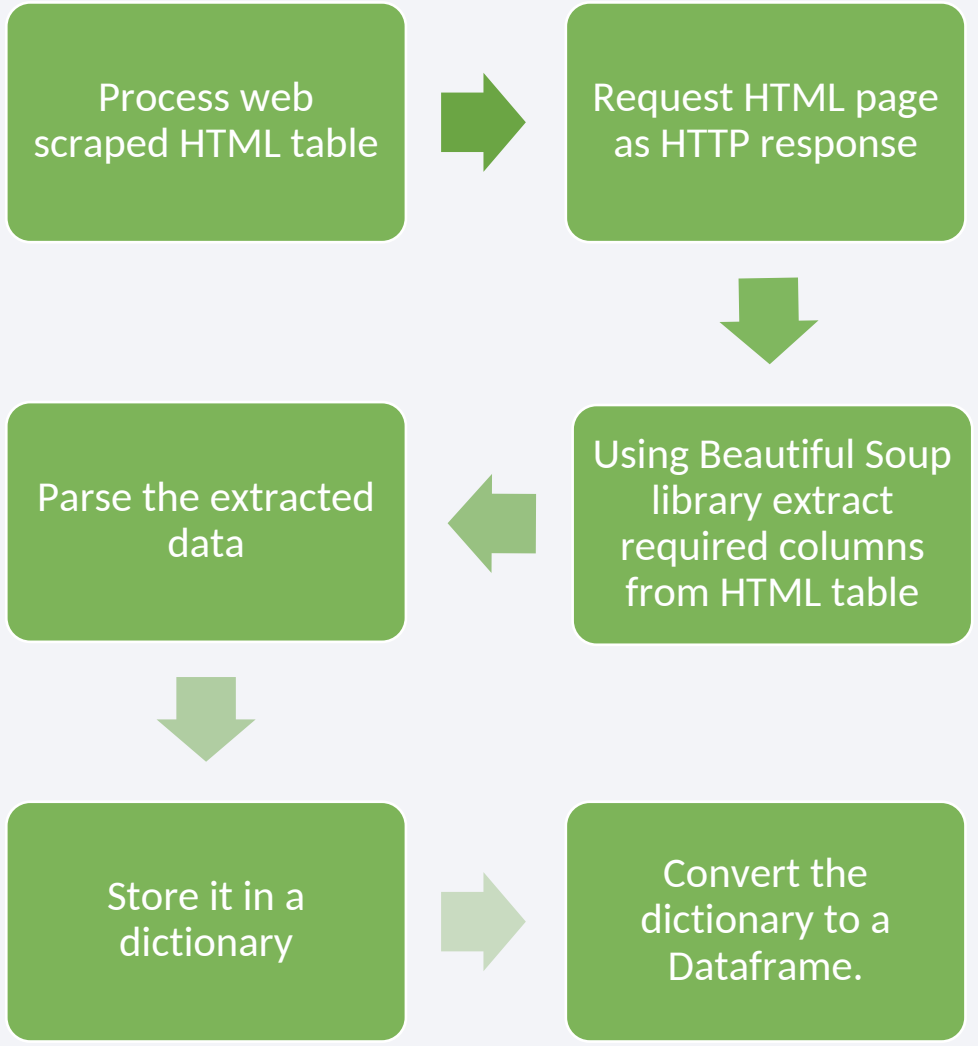
- In the launch data, we use helper functions to extract the required data.
- The requested data has a lot of ID's that we use to extract useful information using API.
- Then the data is further processed and cleaned to make it useful.

- Github URL- API-Data-collection.ipynb



Data Collection - Scraping

- An HTML table from a Wikipedia page was used to extract Falcon 9 launch data.
- Beautiful Soup library was used to parse the data and convert it to dataframe.
- Github URL - Web_scraping.ipynb



Data Wrangling

Data Wrangling involves the process of cleaning and structuring raw data into a desired format for better data analysis

- The pandas library in python is used in this project to process data.
- New variables like ‘Class’ were added indicating the success and failure of each rocket launch in binary form.

- Github URL - Data-Wrangling.ipynb

EDA with Data Visualization

- The exploratory data analysis performed on the dataset using Data visualizations focuses on relationship between different variables and trends of data over time.

In this project, matplotlib and plotly libraries of python are used .

- Key findings are:

➢The success rate of launches has steadily increased since 2013

➢Most of the orbits have launches with payload mass less than 8000 kg.

➢Flight number does not have much impact on success rate but, payload mass does.

- Github URL - EDA-with_visualization.ipynb

EDA with SQL

- The exploratory data analysis performed on the dataset using SQL focuses on data distribution and identifying key trends.

- Key findings are:
  - The maximum payload pass carried till date is 15600 kg.
  - There are total 4 launch sites used by SpaceX
  - The first successful ground pad landing was accomplished in December 22, 2015

- Github URL- EDA-with-SQL.ipynb

Build an Interactive Map with Folium

- This project includes an interactive map to better understand the geography and the conditions where a launch site is built.

    1. MAP 1: Located all 4 launch sites on the map.
    2. MAP 2: Displayed launch outcomes at each site
    3. MAP 3: Displays proximity to any railway, highway, coastline, etc. by distance.

Build a Dashboard with Plotly Dash

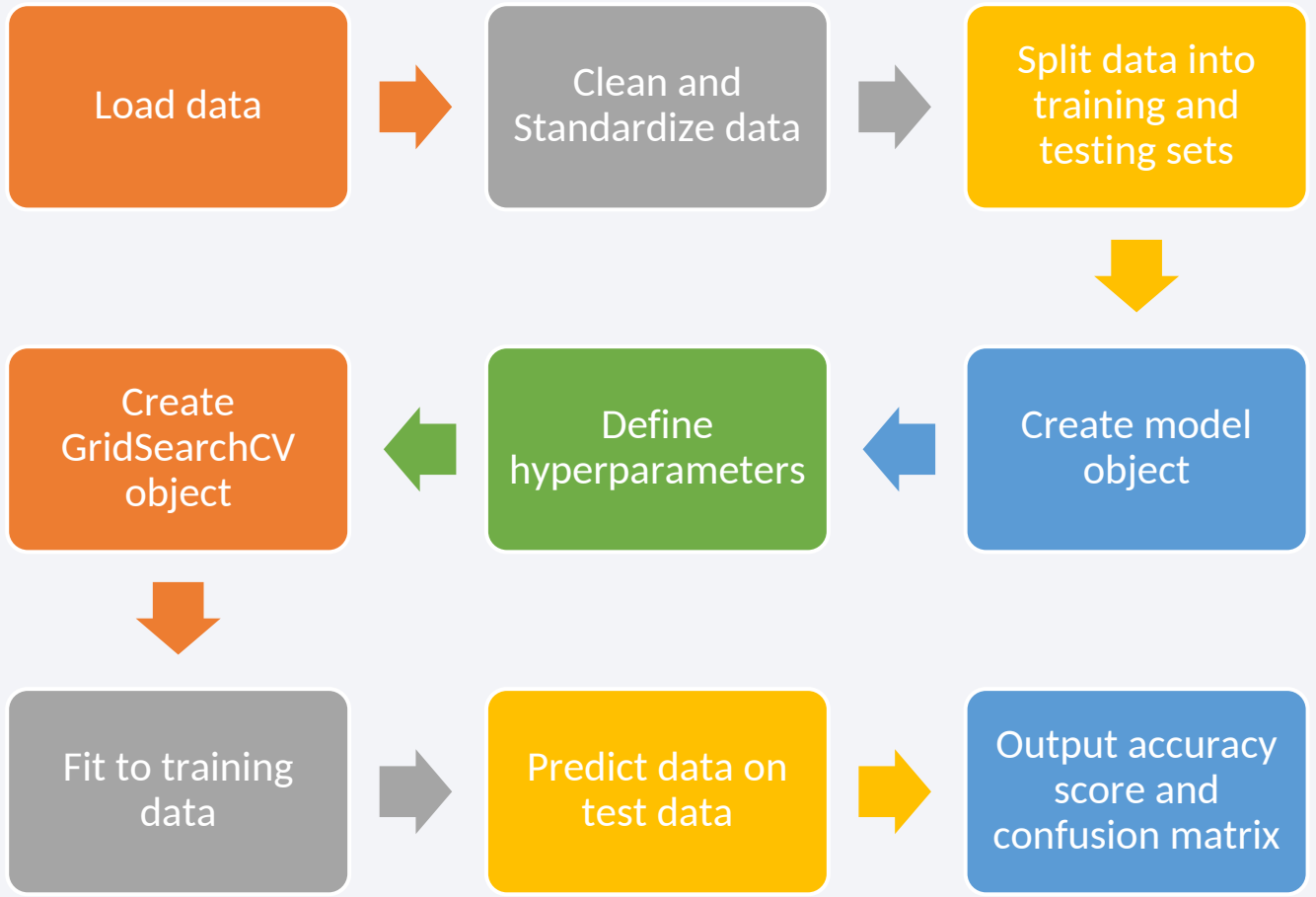
- In this project, an interactive plotly dashboard is included to better understand the data. The dashboard includes:
    1. Pie Chart of all sites – This pie chart displays the success count of all four launch sites. To understand the overall trend.
    2. Pie Chart of individual sites – This pie chart depicts the success and failure count proportion based on the selected launch site from the dropdown menu.
    3. Scatter plot - The plot shows the success count of each booster version based on payload mass that is selected from a range slider.

- GitHub URL - Plotly-Dashboard.py

Predictive Analysis (Classification)

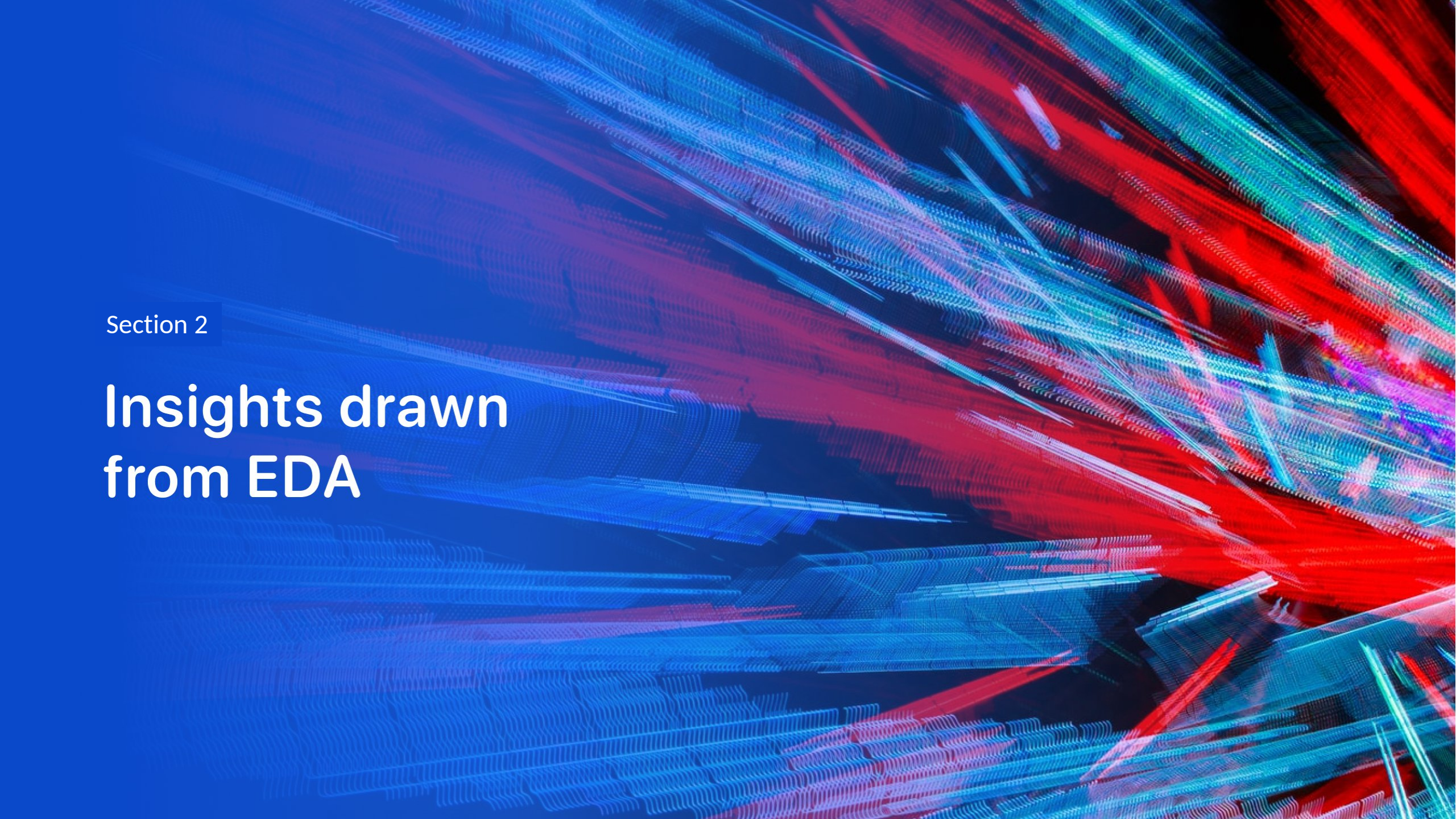
- Four models were created and evaluated :
    1. Logistic Regression model
    2. SVM model
    3. KNN model
    4. Decision tree model

- Accuracy Score and Confusion matrix are used to compare the models.
- Github URL - Machine-learning.ipynb



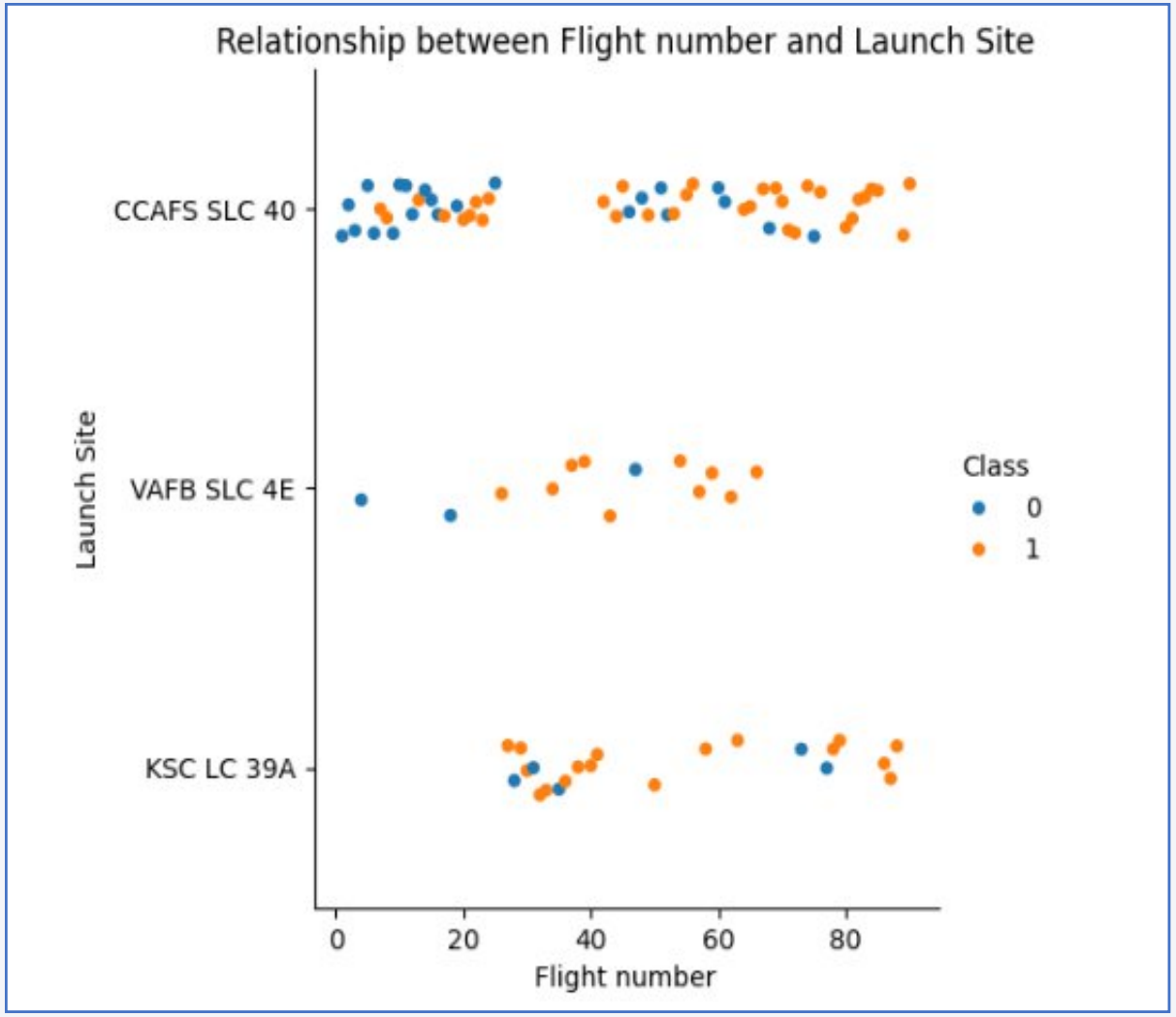
Results

- **Exploratory Data Analysis (EDA) results**:
  1. Launch success rate varies by site: [example: KSC had the highest success rate, CCAFS the lowest].
  2. Payload mass was an important factor — heavier payloads tended to have higher success rates but lower use.
  3. Some booster versions consistently performed better than others.
- **Interactive analytics demo :**
  1. Built a dashboard to visualize launch success by site
  2. Interactive map showed the geographical distribution of launch sites.
- **Predictive analysis results:**
  1. Trained four models: Logistic Regression, SVM, KNN, and Decision Tree.
  2. Logistic Regression, SVM, and KNN achieved the same highest accuracy of 83.34%, while Decision Tree was lower.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

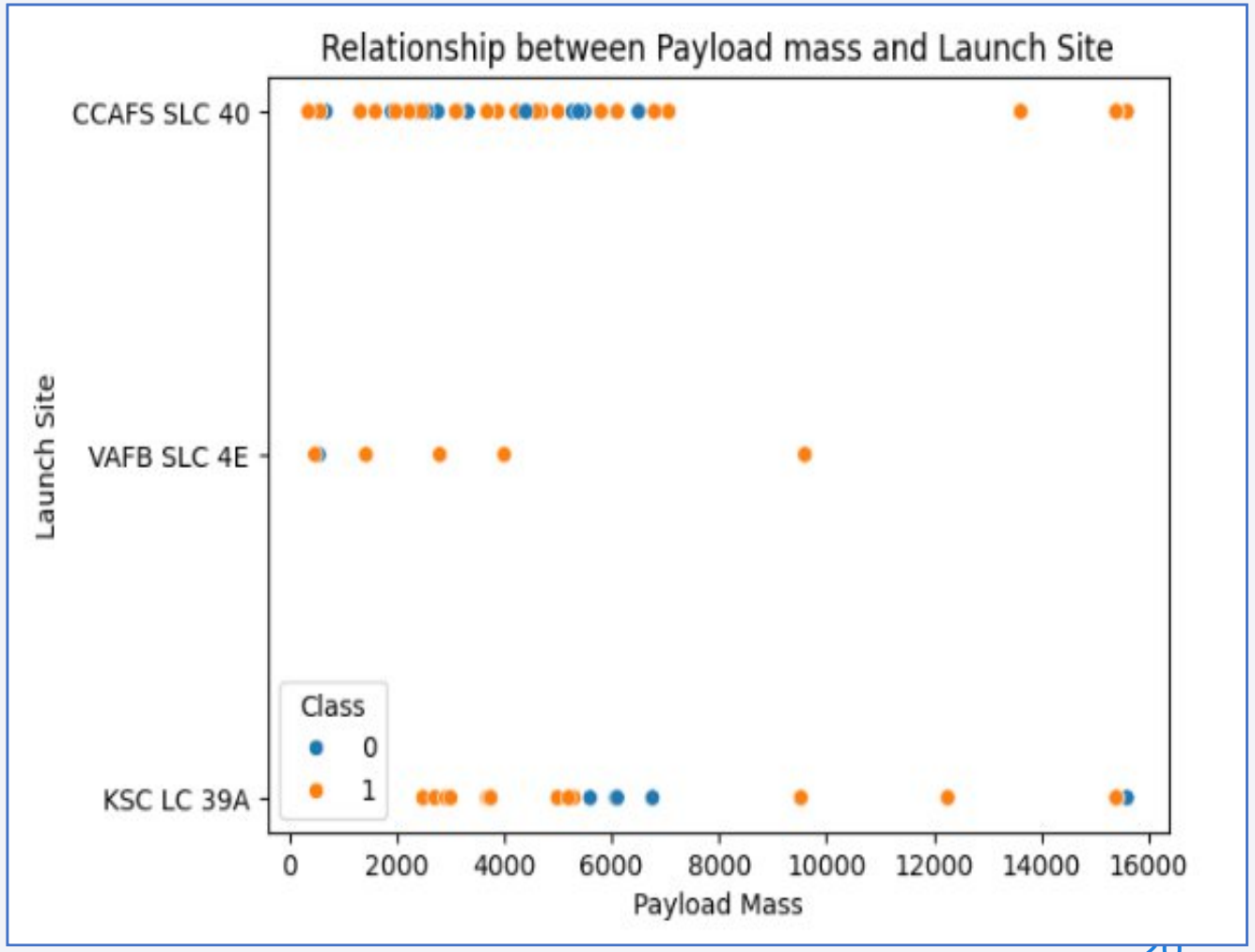
- The image displays the scatter plot of Flight Number vs. Launch Site
- Maximum flights were launched at CCAFFS SLC-40
- There is no significant impact of flight number on the success rate.



Payload vs. Launch Site

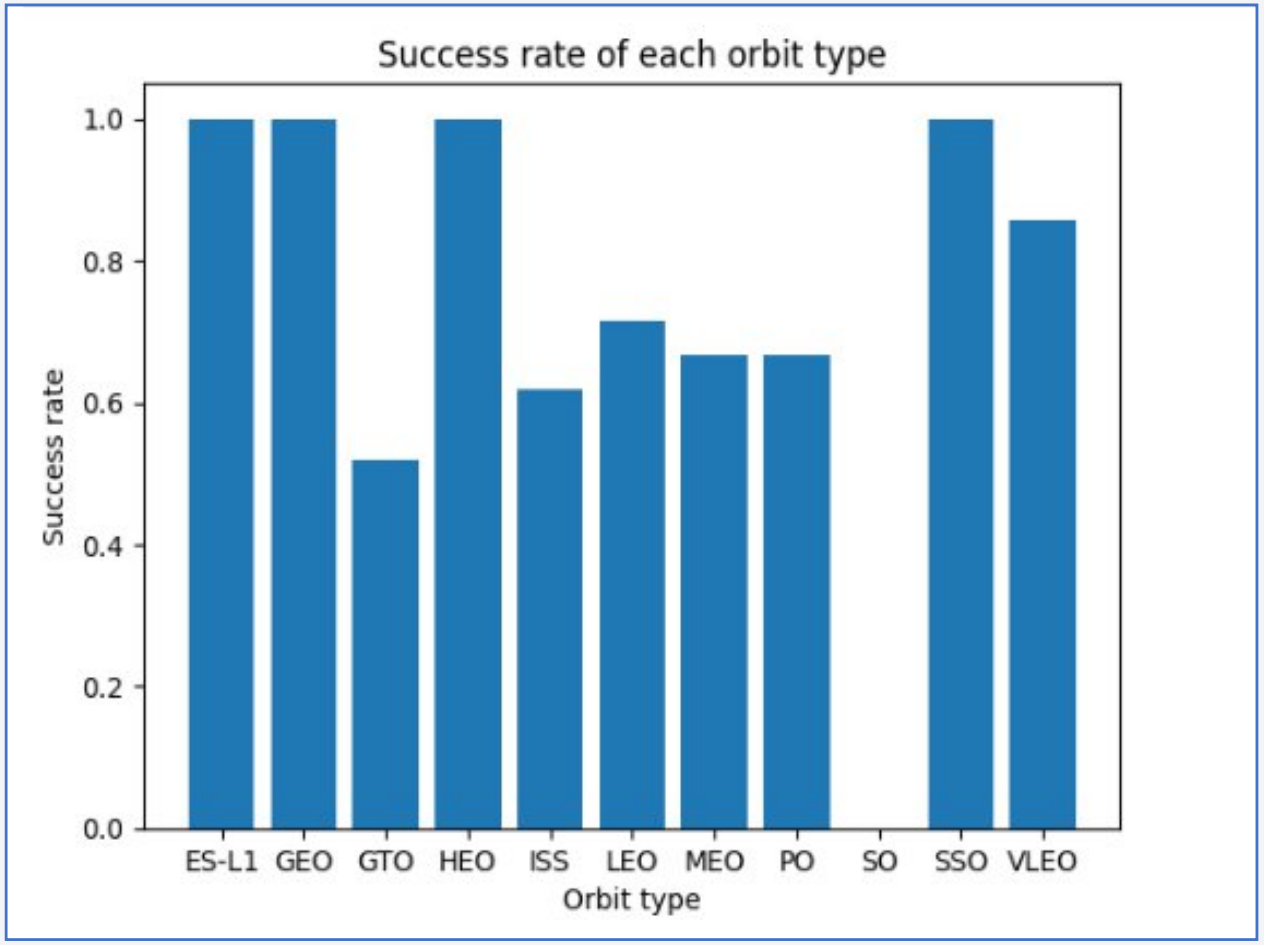
- Image shows a scatter plot of Payload vs. Launch Site.
- CCAFS SLC-40 and

  KSC LC-39A show 99% success rate with a payload mass greater than 10000.
- VAFB SLC-4F does not depend much on payload mass.



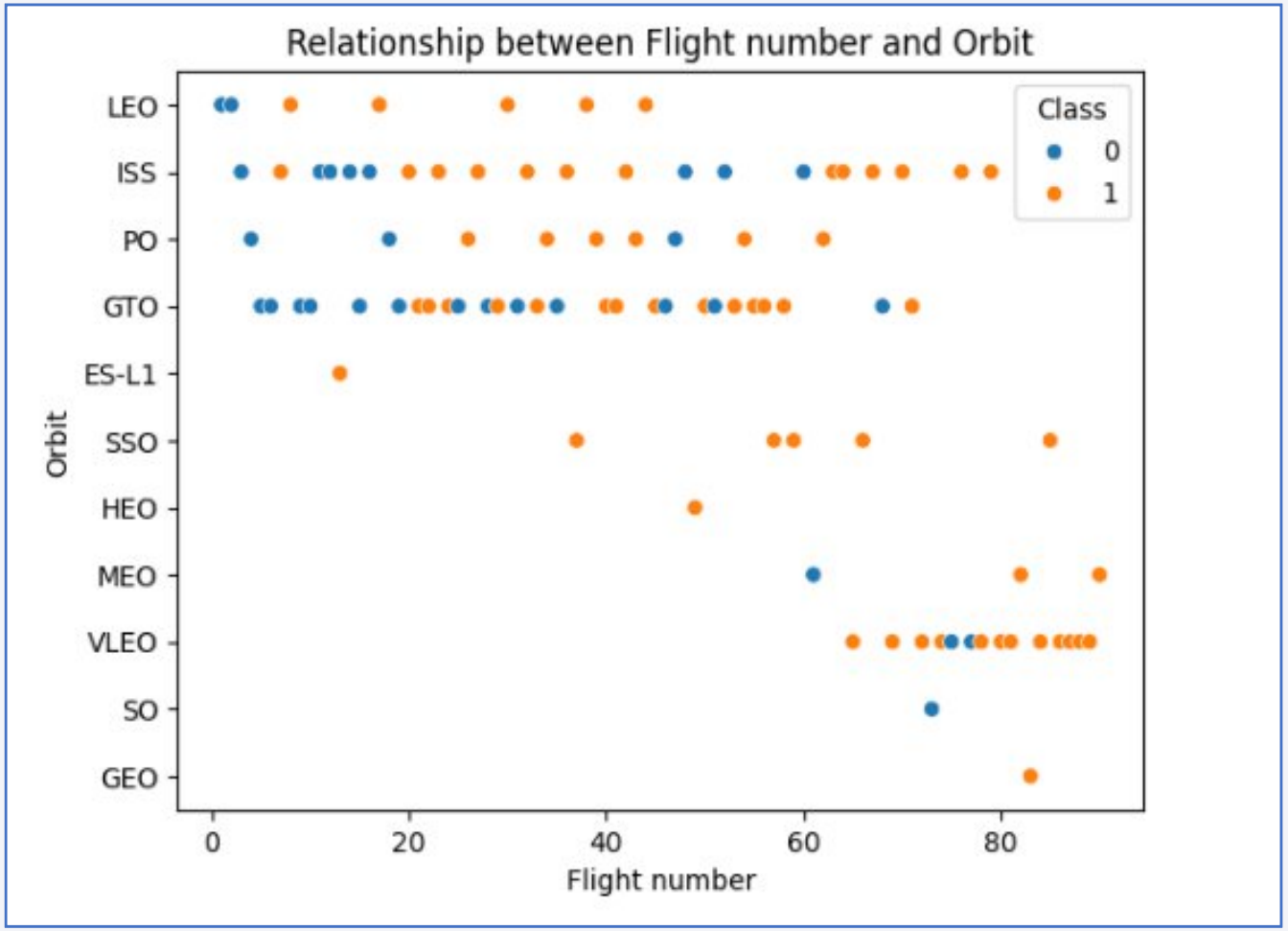
Success Rate vs. Orbit Type

- The picture displays a bar chart for the success rate of each orbit type
- ES-L1, GEO, HEO and SSO orbits have 100% success rate
- SO orbit has the lowest 0% success rate



Flight Number vs. Orbit Type

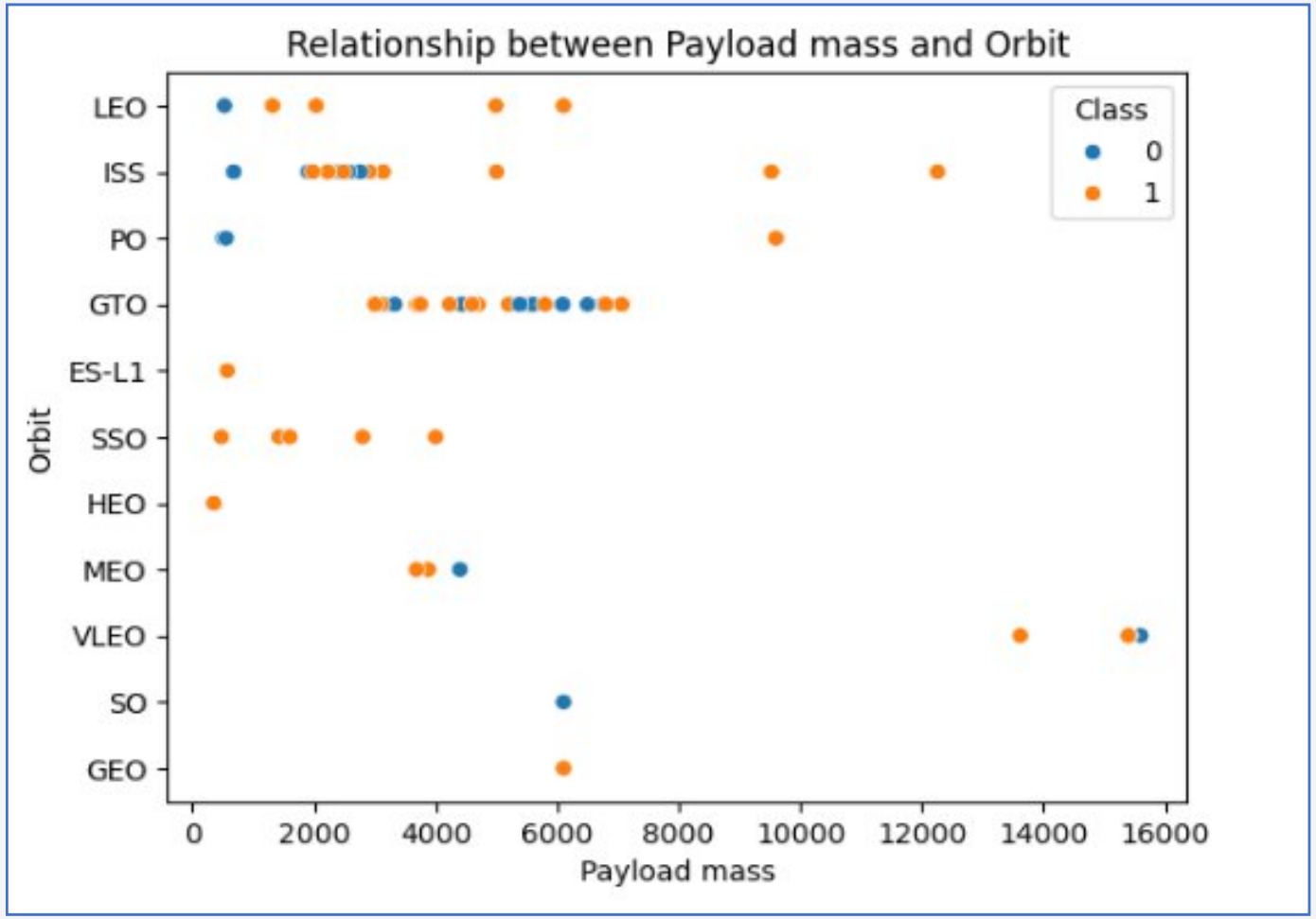
- Image shows a scatter point of Flight number vs. Orbit type
- There is no significant pattern
- The orbits with 100% success rate do not have more than 5 number of launches.



Payload vs. Orbit Type

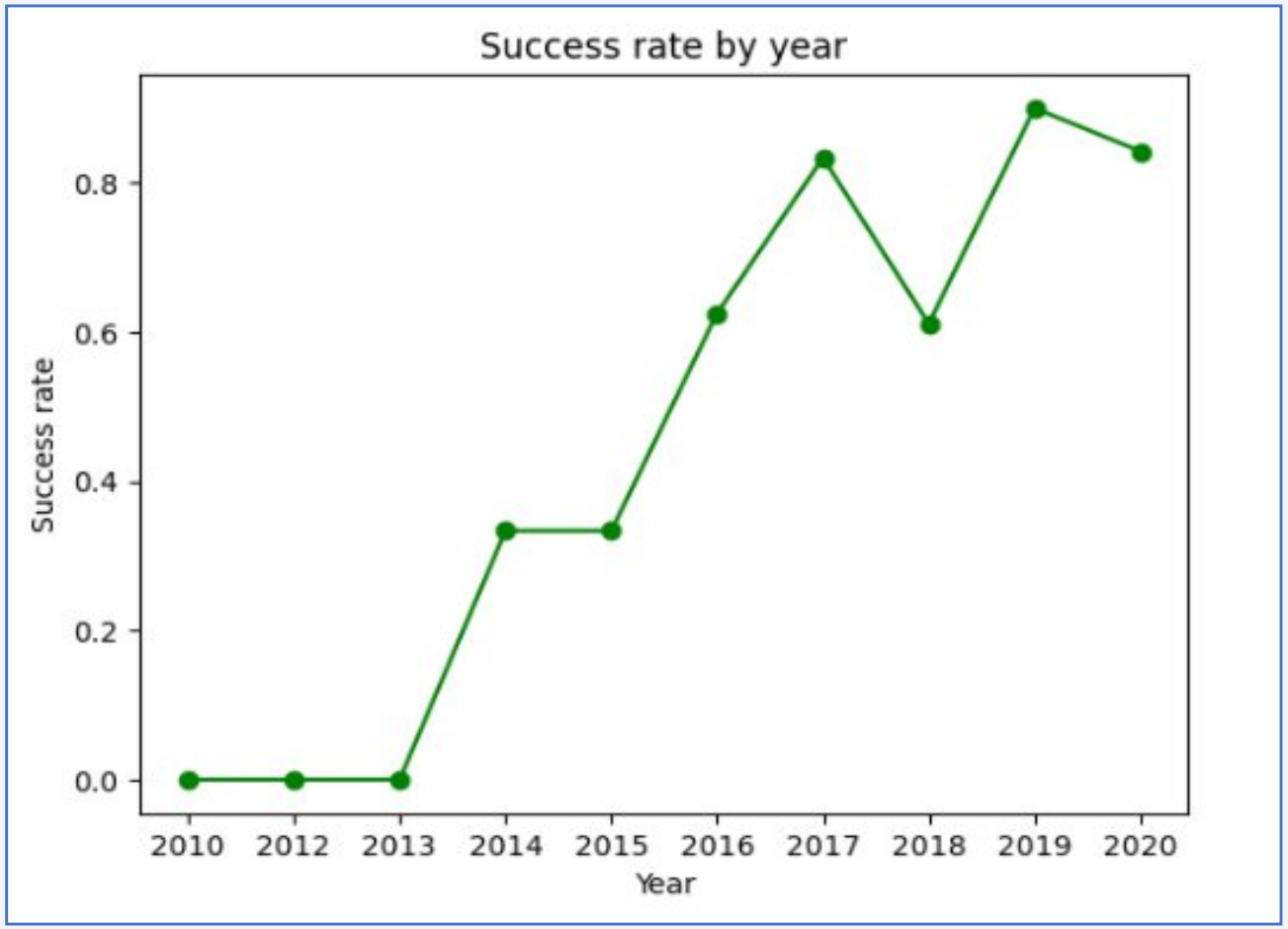
- Image portrays a scatter point of payload vs. orbit type
- Most of the orbits do not have launches with payload mass greater than 8000,

  except from ISS, PO and VLEO with 99% success rate.



Launch Success Yearly Trend

- The picture displays a line chart of yearly average success rate
- The success rate 0 till 2013, followed by a significant increase till mid-2017.
- A drop in success rate was observed in 2018 but it rose again in 2019.



All Launch Site Names

- The unique launch sites are identified by examining the ‘Launch_Site’ column in the data.
- There are 4 distinct launch sites used by SpaceX company

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

Launch Site Names Begin with 'CCA'

- There are two launch sites with names starting from ‘CCA’ ; ‘CCAFS LC-40’ and ‘CCAFS SLC-40’.
- The are 10 columns conveying the information, the query used displays first five rows with launch sites among the selected two sites.

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

Total Payload Mass

- The query displays the total mass of the payload (a part of the rocket) used in rockets launched by NASA (CRS).
- It portrays a total mass of 45596 kg.

Customer	Total_payload_mass
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- The query calculates average payload mass carried by booster version F9 v1.1.
- It portrays a total average of 2928.4.

Booster_Version	avg(PAYLOAD_MASS__KG_)
F9 v1.1	2928.4

First Successful Ground Landing Date

- The query displays the date of the first successful landing outcome on ground pad.
- The date displayed is December 22, 2015 , which was about 10 years ago. Numerous successful landings have took place since.

Date	Landing_Outcome
2015-12-22	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- The query lists the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Only four boosters version have been successful to land on drone ship with a medium payload mass.

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

Total Number of Successful and Failure Mission Outcomes

- The query displays the total number of successful and failure mission outcomes in the data.
- This indicates the time when mission was accomplished , however, it does not state if the first stage landed successfully or not.
- The data shows the mostly the mission has been a success.

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

Boosters Carried Maximum Payload

- The query lists the names of the boosters which have carried the maximum payload mass that is 15600 kg.
- In total, there are 12 boosters that have carried the heaviest payloads.

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

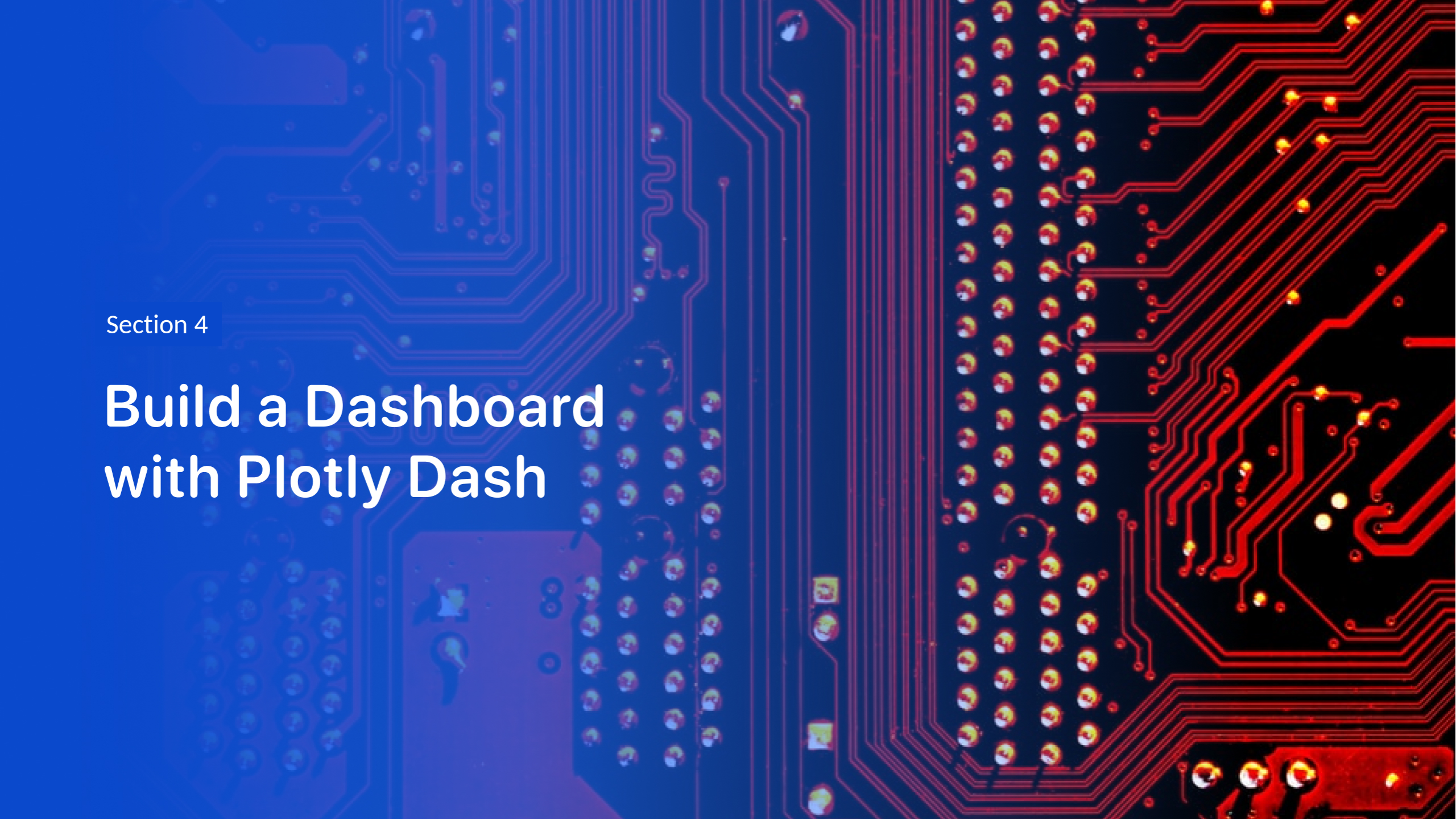
- The query displays the failed landing outcomes in drone ship, their booster versions, and launch site names in the year 2015.
- There were only two failures in drone ship landing at the same site ‘CCAFS LC-40’ but with different booster versions.

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

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

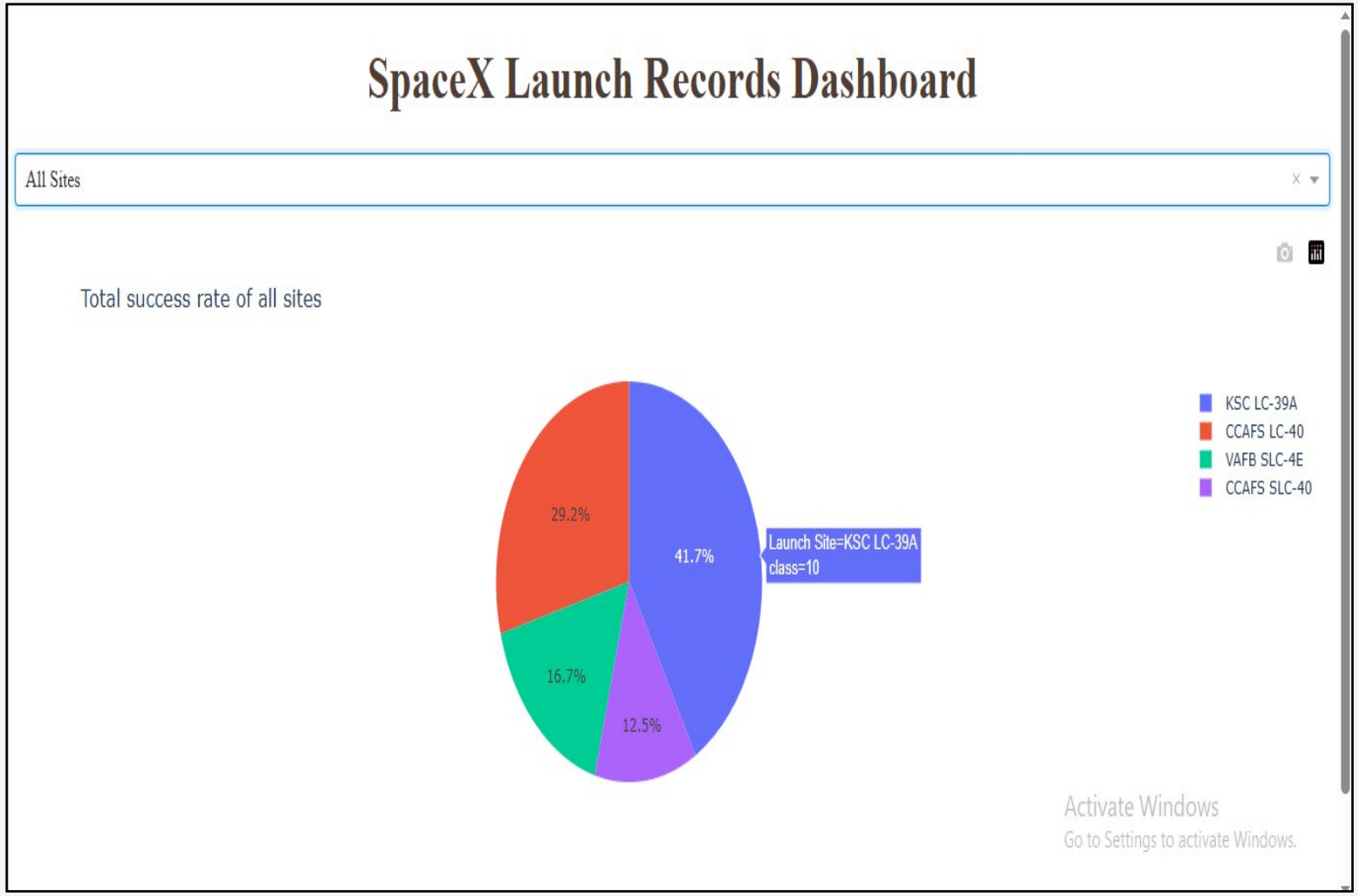
- The query ranks the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.
- There are total 8 landing outcomes, among these most rows have No attempt (data not given)
- Most of the outcomes were success.

Landing_Outcome	Outcome_count
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1


Section 4
Build a Dashboard with Plotly Dash

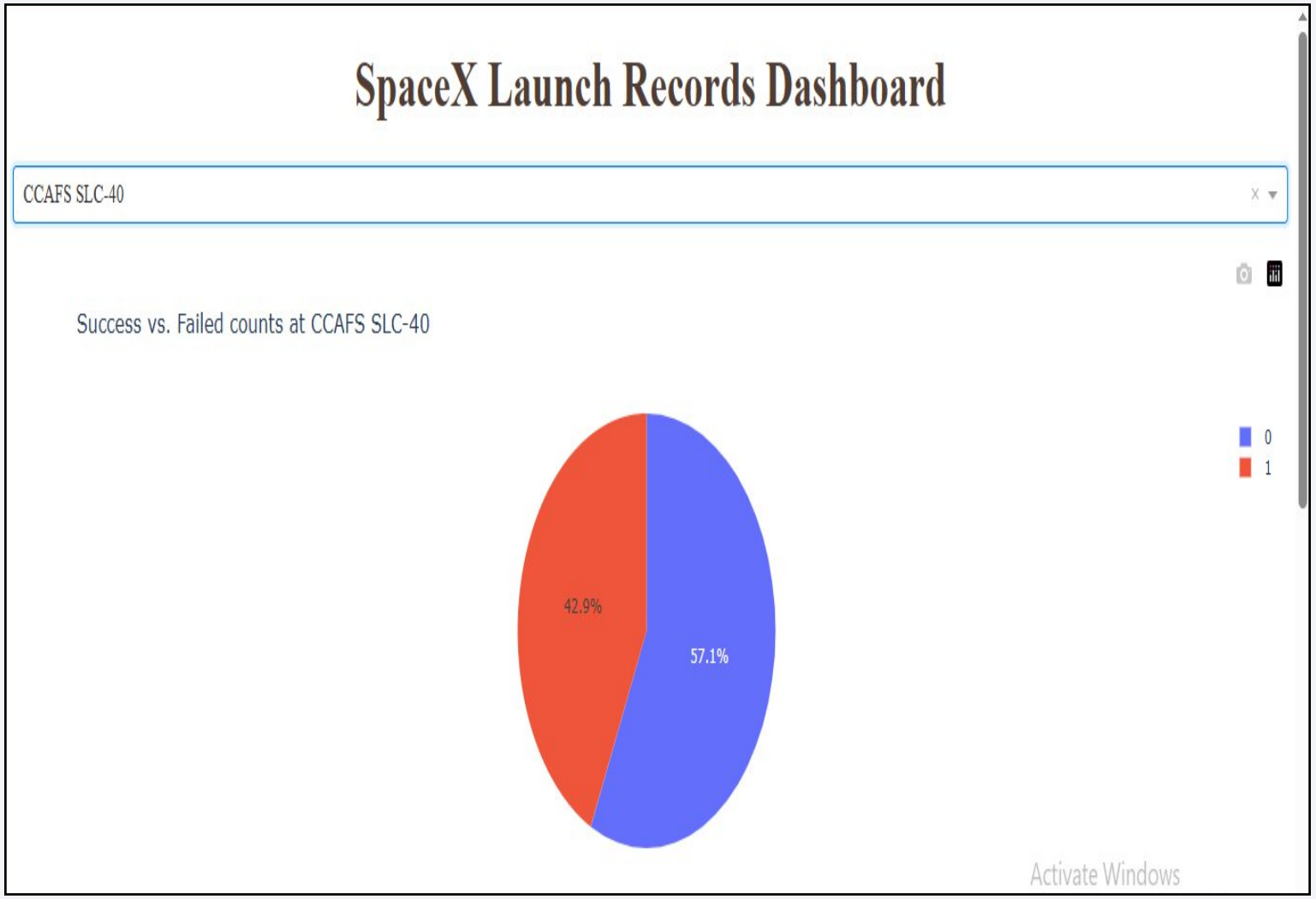
Success count of all Launch Sites

- There are 4 launch sites.
- Site KSC LC-39A has the highest success rate of 41.7%, whereas site CCAFS SLC-40 has the lowest rate of 12.5%.
- It compares absolute count of success across sites.



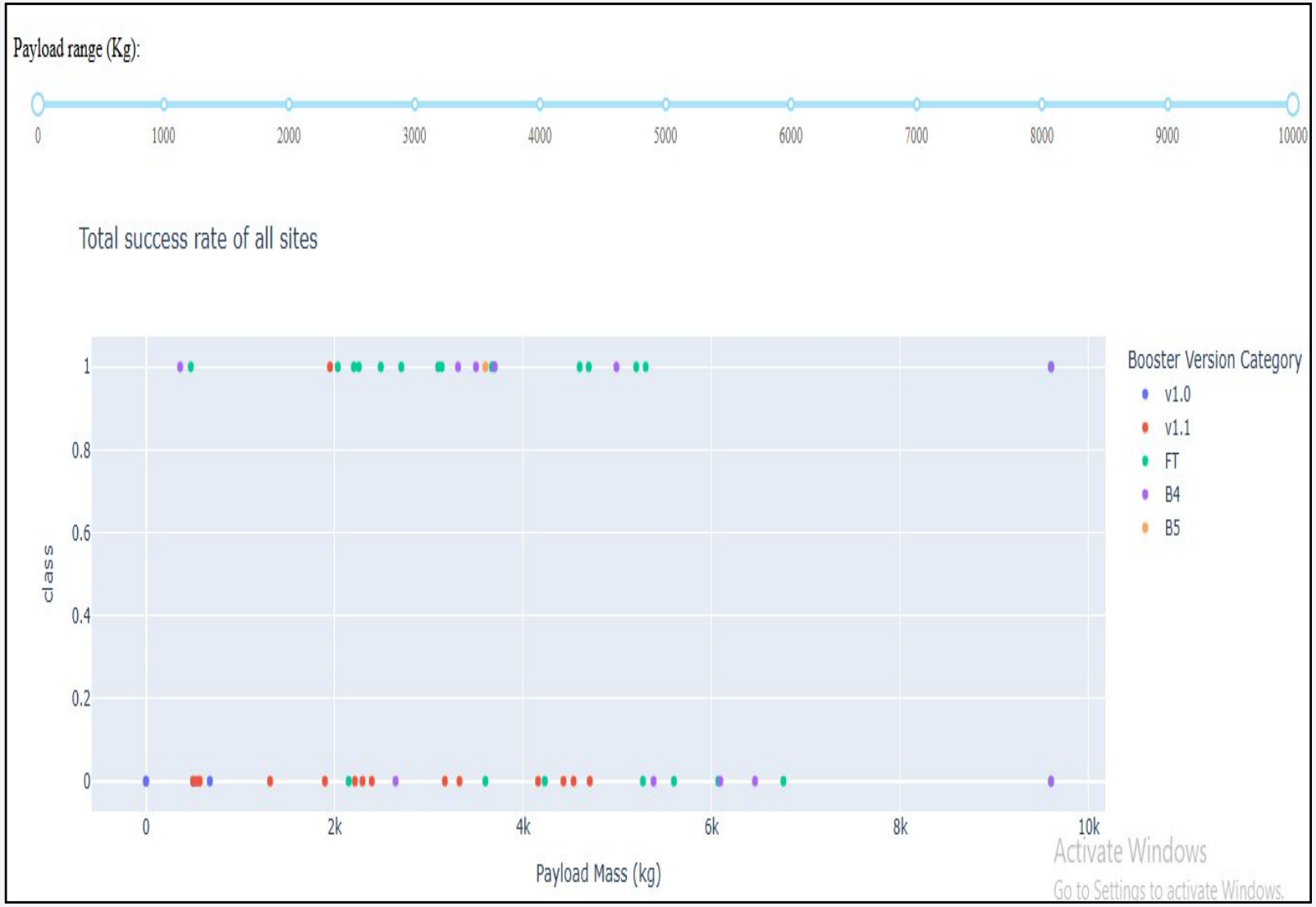
Launch site with highest success ratio

- CCAFS SLC-40 has the highest success ratio with 42.9% success ratio and 52.15 failure count.
- It compares proportions of success and failure within each site.



Payload vs. Launch outcome

- The image shows a scatter plot of Success rate and Payload mass based on booster version.
- FT booster has the highest count of success and v1.1 booster has maximum count of failure.
- Most boosters have payload mass less than 8000 kg

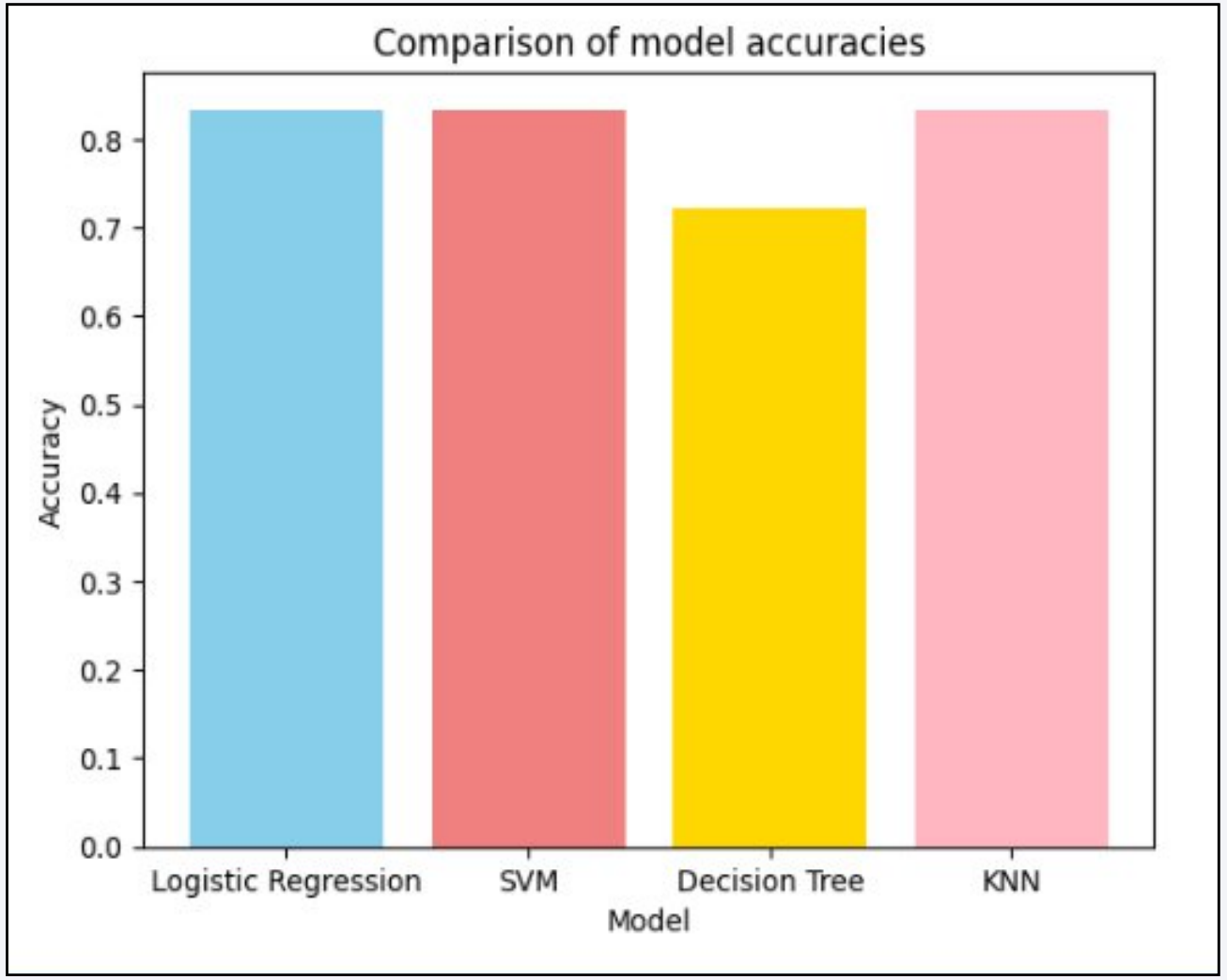


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The bar graph depicts the accuracy scores of the 4 models: Logistic Regression, SVM, KNN, Decision tree.
- Logistic Regression, SVM, and KNN have the same accuracy score of 83.34%., whereas, decision tree has the lowest accuracy.



Confusion Matrix

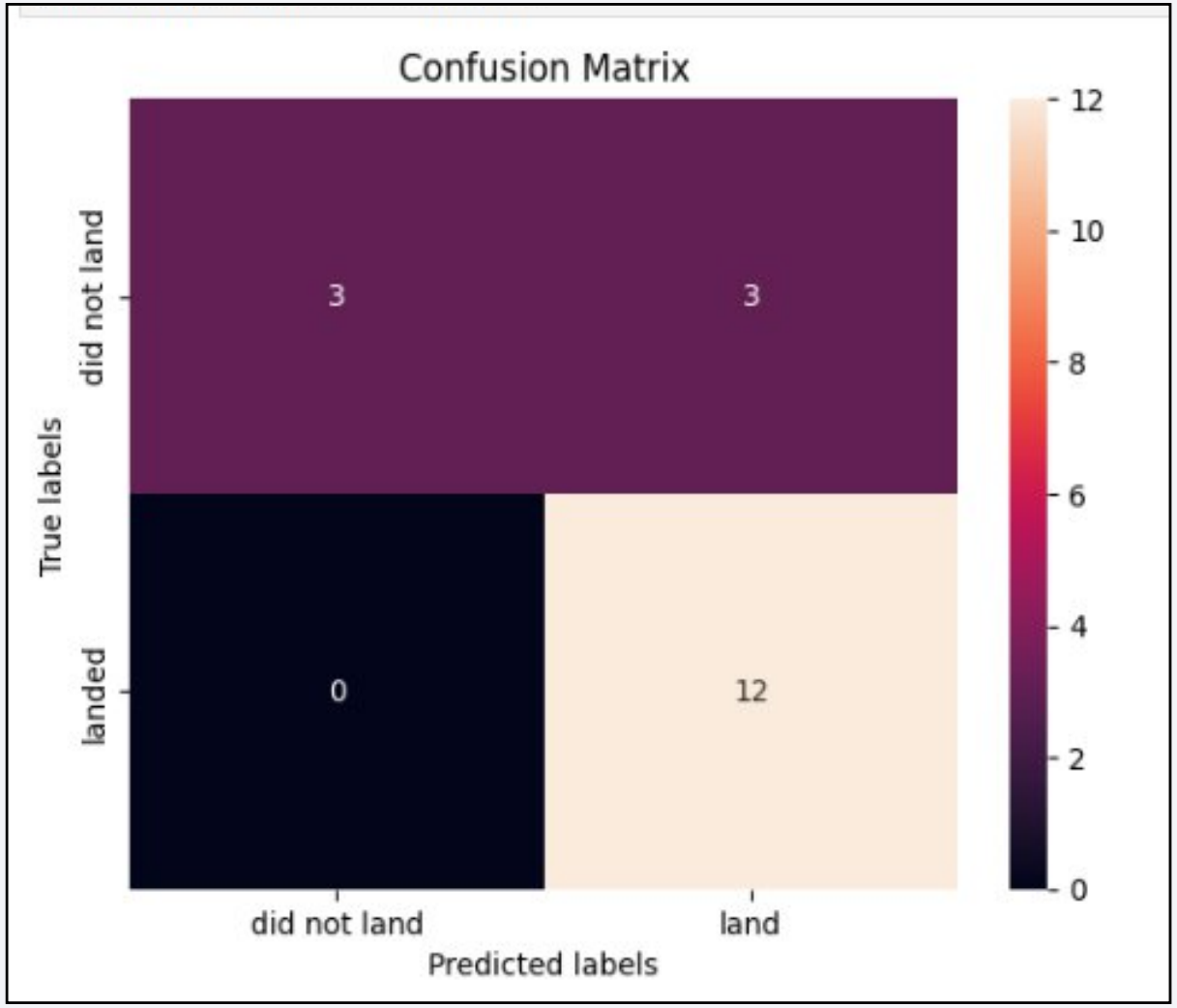
- The image shows the confusion matrix of all three models: Logistic Regression, SVM, KNN.

✓True positives- 12

✓False positives- 3

✓True negatives- 3

✓False negatives- 0



Conclusions

- Exploratory analysis showed that launch success rates vary by site, payload mass, and booster version.
- Interactive dashboard provided visual insights and allowed users to filter and explore launch outcomes.
- Predictive modeling with Logistic Regression, SVM, and KNN all achieved the highest accuracy of 83.34%.
- Since the three models performed equally well, any of them can be chosen depending on whether interpretability (Logistic Regression), scalability (SVM), or simplicity (KNN) is preferred.

Thank you!