IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Chaitanya Ashok Malagikar
09/04/2025

Outline

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

Executive Summary

- This project explores the analysis of SpaceX's Falcon9 launch vehicle to understand mission trends, assess success rates, and derive actionable insights from data.
- The methodology involves data collection, cleaning, visualization, and predictive modeling to deliver a comprehensive understanding of launch outcomes.

Introduction

- SpaceX, a leader in aerospace innovation, conducts numerous launches aimed at advancing space exploration and commercial satellite deployment. This project leverages SpaceX's publicly available data to analyze launch trends, mission outcomes, and operational efficiency.
- Problems you want to find answers:

1. Which launch sites have the highest number of launches and success rates?
2. How does payload mass affect the success of a launch in various orbit types
3. What trends can be observed in launch success over the years?
4. Which factors, such as flight numbers, orbit types, or payload sizes, influence mission outcomes?
5. How can predictive models be utilized to forecast launch success effectively?

Section 1

Methodology

Methodology

Executive Summary

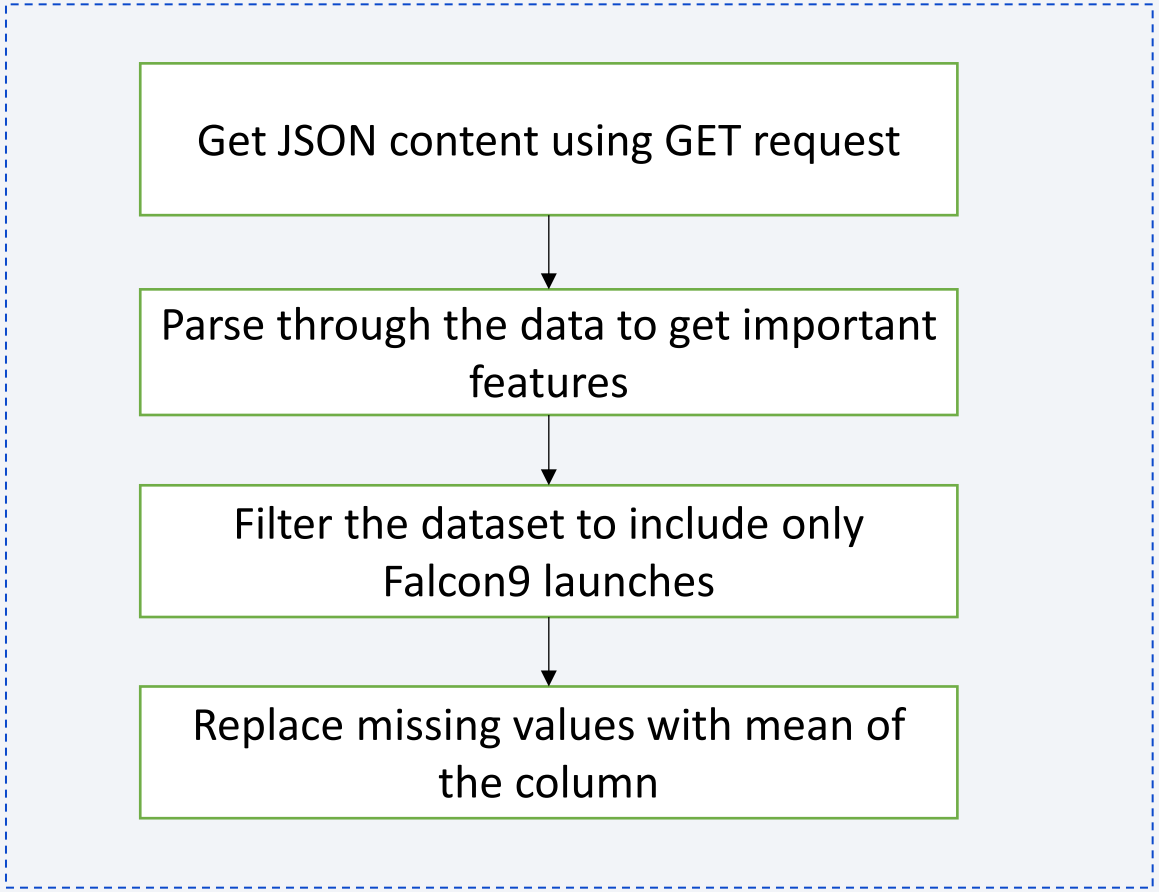
- Data collection methodology:
  - Data has been collected using SpaceX's API and through web scraping
- Perform data wrangling
  - Handle missing values and create a binary numerical column to represent launch outcomes, indicating success or failure
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification models were built to predict launch outcomes, with hyperparameters tuned using Grid Search and cross-validation, and performance evaluated through metrics like accuracy

Data Collection

- The data was collected by using the SpaceX API and through web scraping.
- The details of the data collection process is explained further in the following slides.

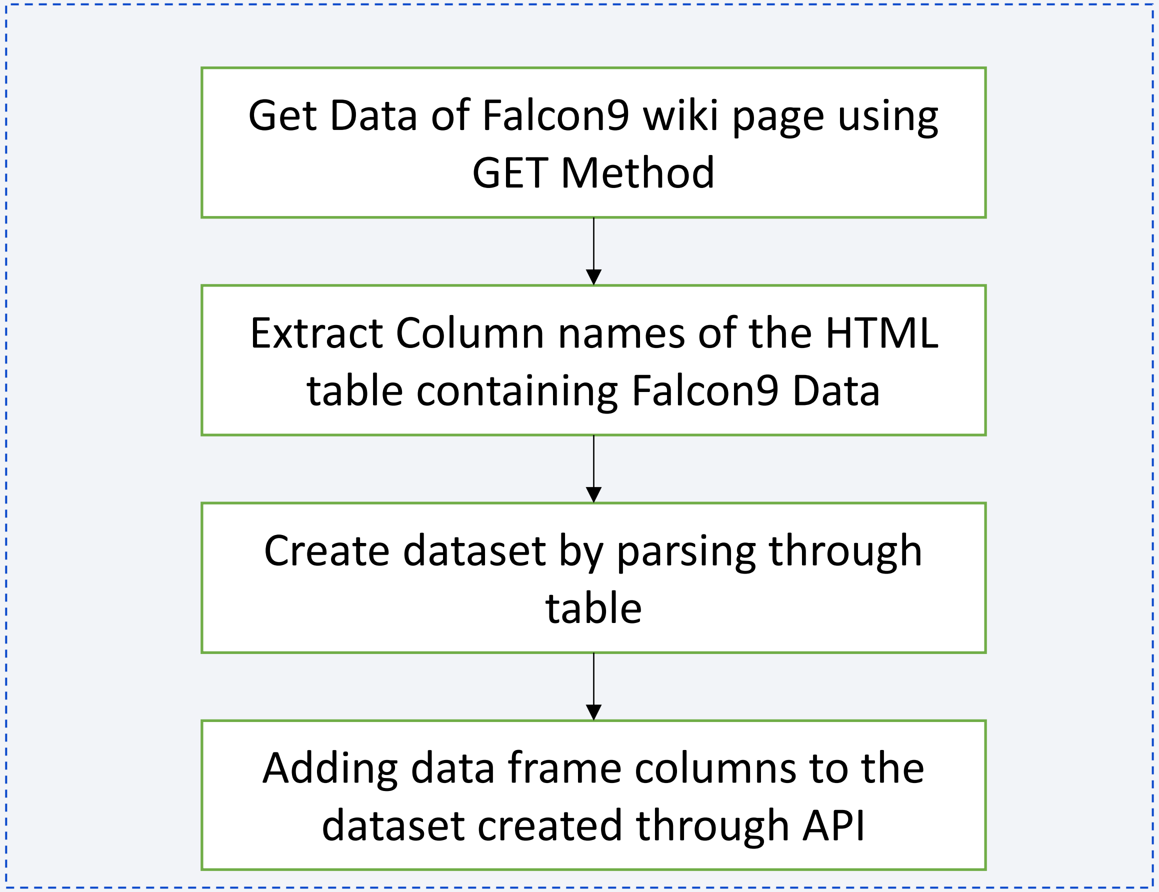
Data Collection – SpaceX API

- The data for this analysis was obtained using the SpaceX API, with the process detailed in the accompanying flowchart
- The code for the project is given [Here]



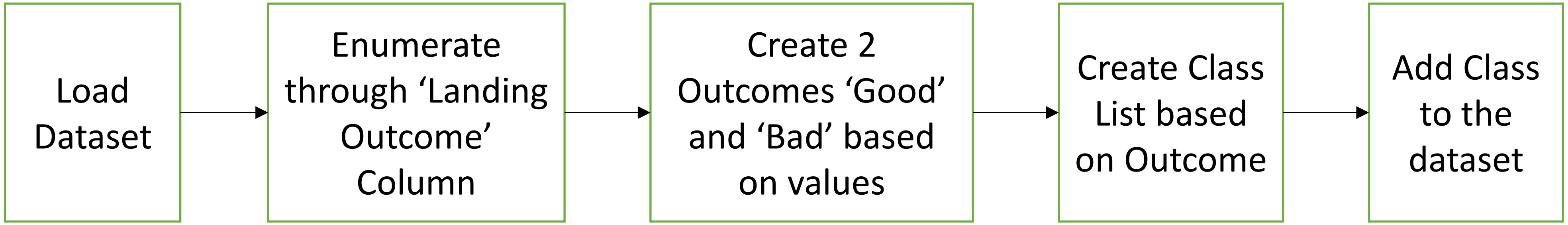
Data Collection - Scraping

- Data has been collected from a Wikipedia page containing details of Falcon9 launch vehicle. The process is detailed in the accompanying flowchart
- The code for the project is given [Here]



Data Wrangling

- To gain meaningful insights into the dataset, an analysis was conducted to examine the number of launches at each site, the frequency of occurrences for different orbit types, and the distribution of various landing outcomes. This study provided valuable trends and patterns across key columns.
- The key task performed here was the addition of a class column to the dataset, which represents a binary classification indicating whether a launch was successful. The steps to accomplish this are outlined below:



- The code for the project is given [Here]

EDA with Data Visualization

- **Flight Number vs. Launch Site:** Plot to check which launch sites were associated with different flight numbers and their corresponding success rates.
- **Payload Mass vs. Launch Site:** To investigate the relationship between payload mass and the launch site. Highlighted that rockets launched from the VAFB-SLC site did not carry heavy payloads
- **Success Rate vs. Orbit Type:** To evaluate the success rate of missions for each orbit type. Orbits like GEO, LEO and SSO have high success rate which SO has no successful missions.
- **Flight Number vs. Orbit Type:** To explore the connection between flight numbers and orbit types. Success seems to be related to the number of flights in LEO but in the GTO orbit, there appears to be no relationship between flight number and success.

EDA with Data Visualization

- **Payload Mass vs. Orbit Type:** To uncover the relationship between payload mass and orbit types. The heaviest payloads are placed in the VLEO orbit. GTO success reduces as the payload increases

- **Launch Site vs. Orbit Type:** To investigate the relationship between launch site and orbit types. VAFB SLC 4E launch site is exclusively utilized for PO and SSO orbits, whereas the CCAFS SLC 40 launch site is versatile and supports nearly all orbit types

- **Launch Success Yearly Trend:** To analyze the yearly trend in launch success rates. Success rates increased significantly from 2014 to 2020, reflecting improvements in rocket technology or launch strategies.

- The code for the project is given Here

EDA with SQL

1. All the Launch Sites from the dataset found using distinct
2. Queried launch sites starting with CCA using Like and showed 5 records using limit
3. Calculated the total payload mass carried by boosters using sum()
4. Calculated the average payload for booster version F9v1.1 using avg()
5. Found the earliest date of successful landing on ground pad using min()
6. Listed booster versions that successfully landed on drone ship with payload mass between 4000 and 6000 kgs

EDA with SQL

7. Counted total number of success and failures with count() and grouped them
8. Retrieved booster version that carried maximum payload mass using max()
9. Displayed how substr() is used by showing failed drone ship landings in 2015
10. Counted landing outcomes between specified dates with count() and grouped them

- The code for this project is given Here

Build an Interactive Map with Folium

- **Launch Site Locations:** The first interactive map uses circles to highlight the geographical locations of the launch sites. Each circle is paired with a marker to display the name of the respective site.
- **Launch Outcomes:** The second map utilizes a marker cluster to effectively visualize the number of successful and unsuccessful launches at each site. This feature makes it easier to analyze launch performance at a glance.
- **Proximity to Key Features:** The third map explores the proximity of each launch site to critical infrastructure and natural features, such as highways, coastlines, and nearby cities.
- The code for this project is given Here

Build a Dashboard with Plotly Dash

- Dropdown menu provides an interactive way to filter the data by launch site. It enables the user to focus on a particular site or aggregated data across all sites
- Pie Chart visualizes the total number of successful launches of all sites. If a particular site is selected, it gives insight of its success and failures
- The range slider is given to filter the data based on payload mass which allows user to gain better insights on correlation between payload mass and launch success
- Provides a scatter plot of launch success and payload mass which provides a way to study the impact of payload mass on successful launch. Includes color coded markers of different booster versions for additional detail
- The code for this project is given Here

Predictive Analysis (Classification)

- Developed a model to predict whether a rocket successfully lands or fails, based on key input features
- Applied multiple classification techniques including Logistic Regression, SVM, Decision Tree and KNN. The workflow is as follows:-

Define Parameters → Classification Method → Grid Search → Fit Model → Retrieve Best Parameters

- We got the best training accuracy with the Decision Tree method. The results are given in the following section.
- The code for this project is given [Here]

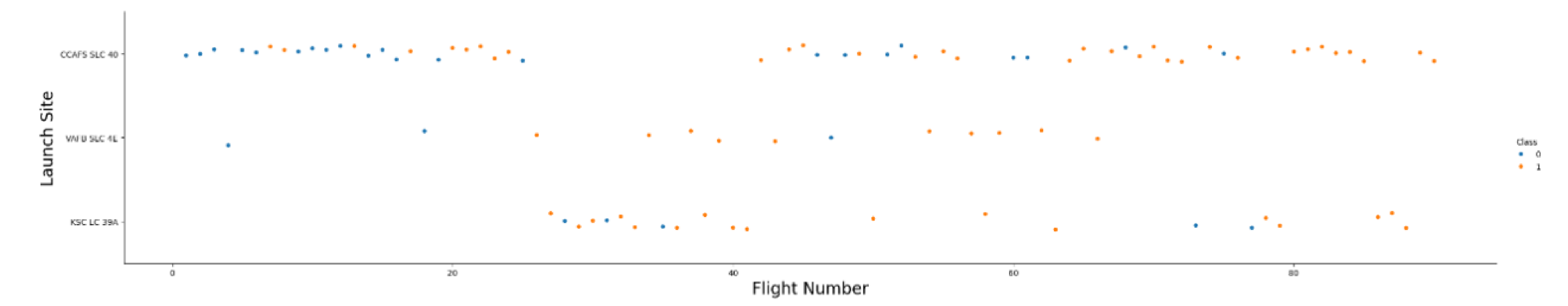
Results

- Exploratory data analysis results has been shown using both the plots and queries to explain the data present in the dataset
- Interactive analytics demo in screenshots. Here the screenshots of both the folium maps and the dash app has been shown
- Predictive analysis results where the accuracy of the findings and the confusion matrix has been displayed

Section 2

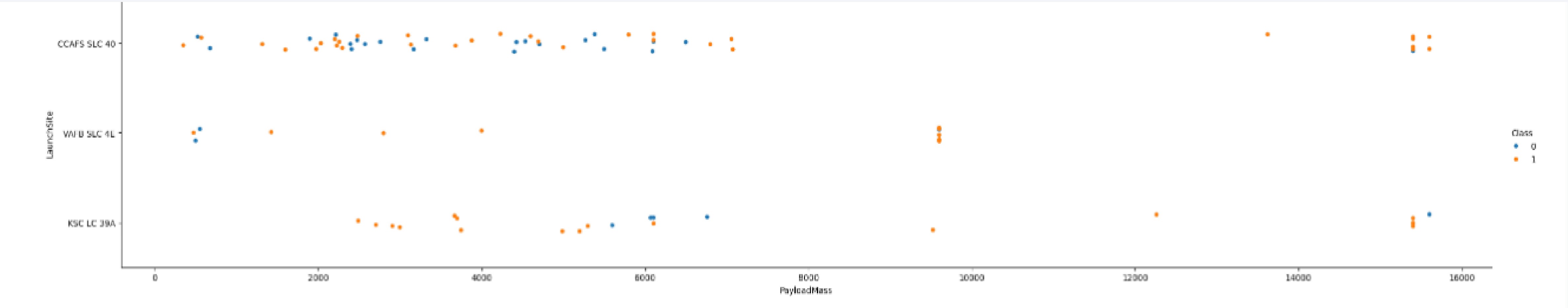
Insights drawn from EDA

Flight Number vs. Launch Site



- There are 3 Launch Sites and the Flight Number of the launches have been plotted.
- The Class represents the number of successful and unsuccessful launches with 1 and 0 respectively.

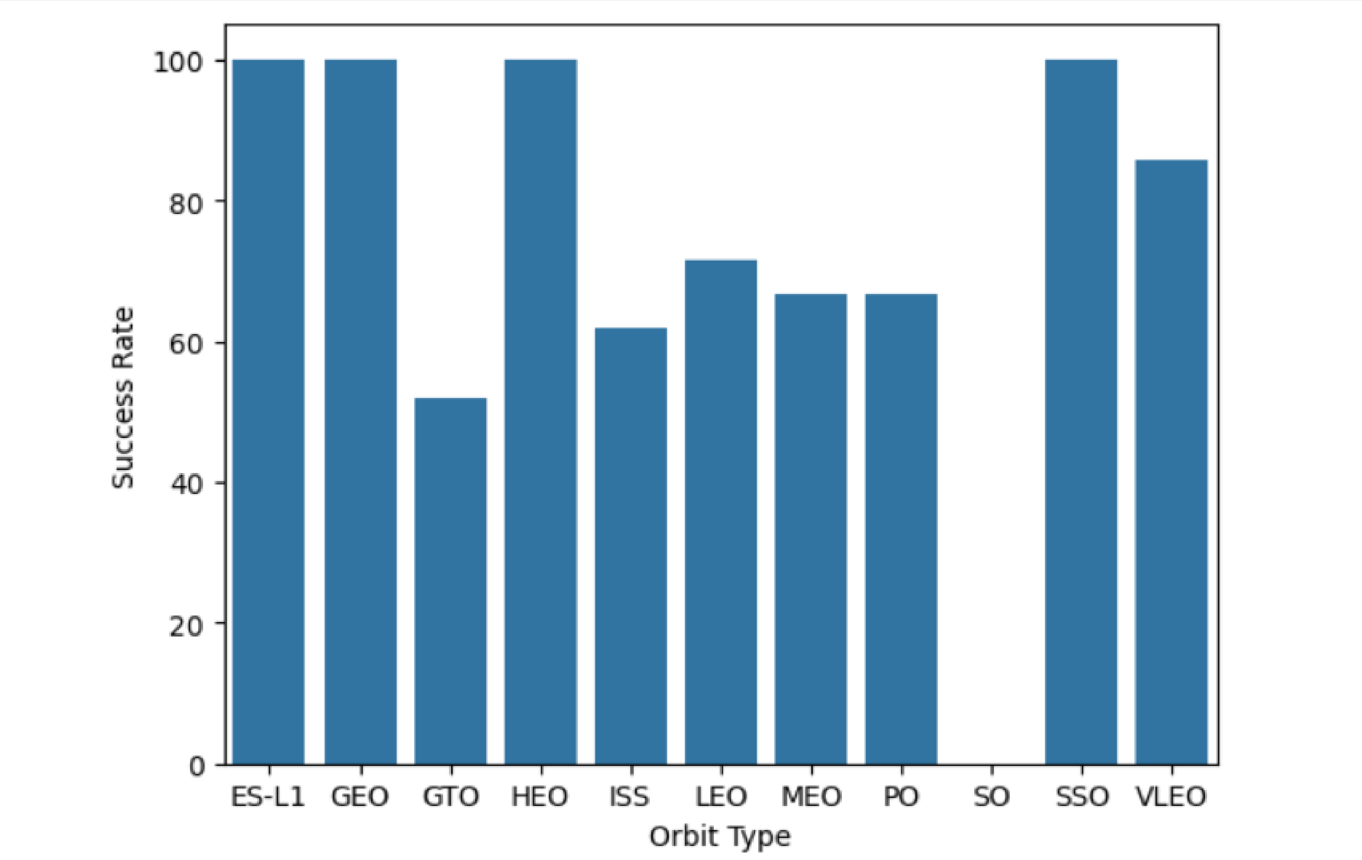
Payload vs. Launch Site



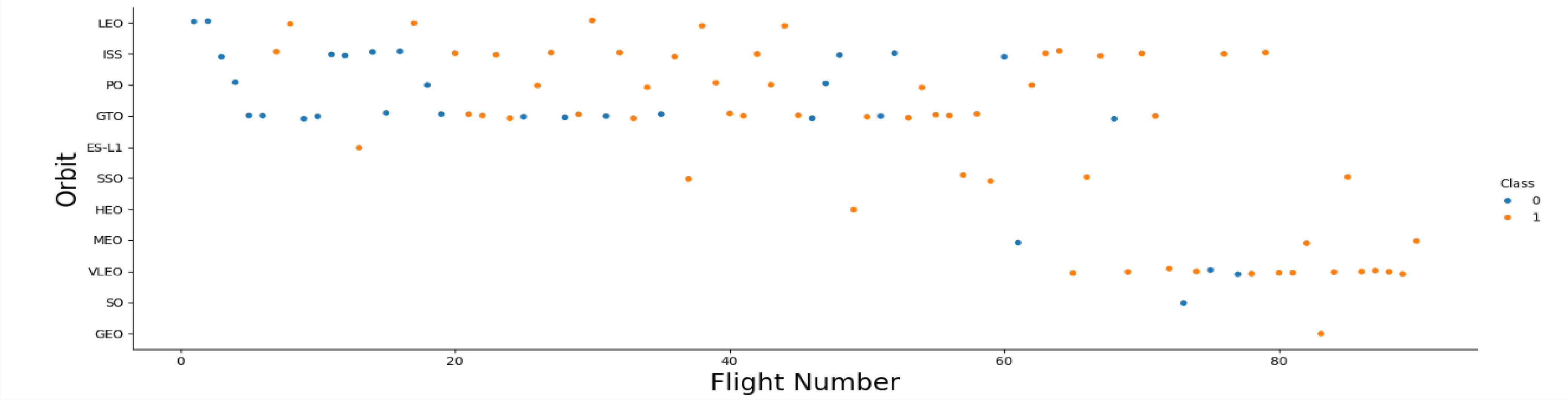
- There are 3 Launch Sites and the Payload Mass of the launches have been plotted.
- The Class represents the number of successful and unsuccessful launches with 1 and 0 respectively.

Success Rate vs. Orbit Type

- The bar chart represents the Success rate of the launches with respect to the orbit type.
- From the plot we can gather important information like the SO orbit having no successful launches and orbits like GEO and HEO having 100% success

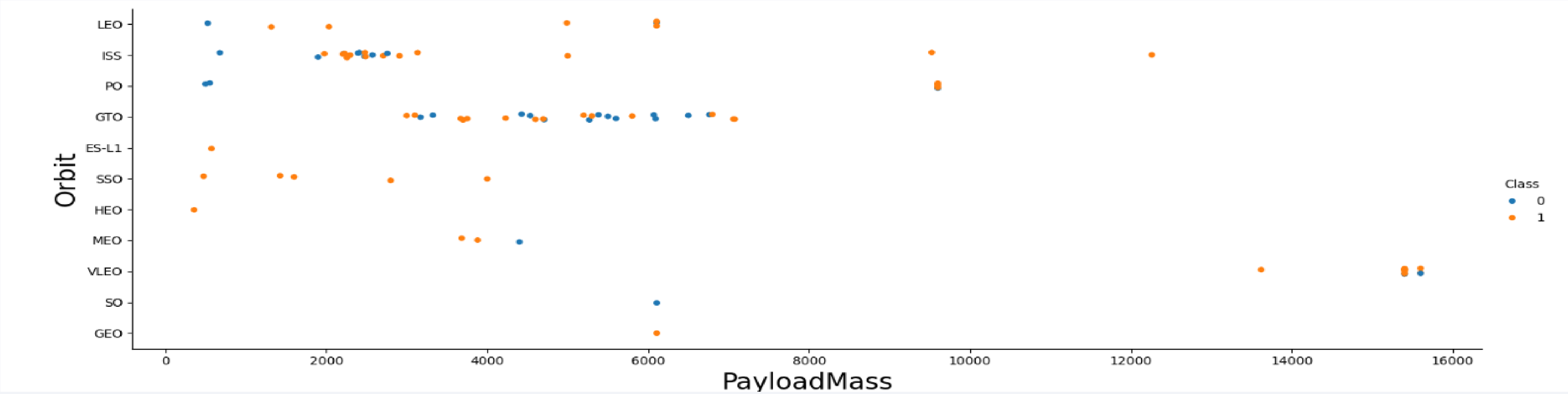


Flight Number vs. Orbit Type



- The various orbits where the satellites were launched and the flight number of the launches have been plotted.
- The Class represents the number of successful and unsuccessful launches with 1 and 0 respectively.

Payload vs. Orbit Type



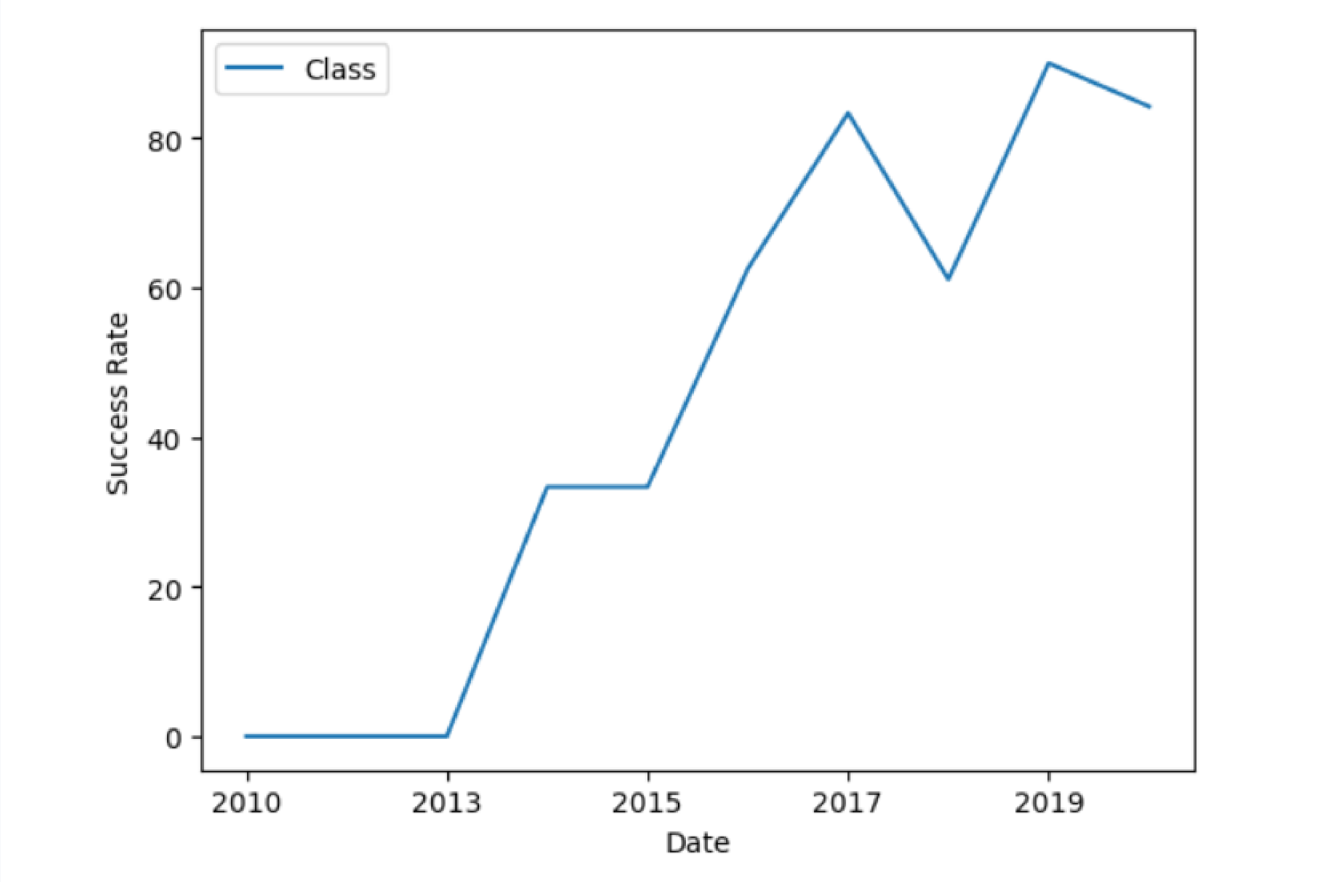
- The various orbits where the satellites were launched and the Payload Mass of the launches have been plotted.
- The Class represents the number of successful and unsuccessful launches with 1 and 0 respectively.

Launch Site vs. Orbit Type

- The plot illustrates that the VAFB SLC 4E launch site is exclusively utilized for PO and SSO orbits, whereas the CCAFS SLC 40 launch site is versatile and supports nearly all orbit types.
- The Class represents the number of successful and unsuccessful launches with 1 and 0 respectively.

Launch Success Yearly Trend

- The line plot shows the successful launches percentage as the years go by.
- The number of successful launches has increased year on year except a blip in 2018.



All Launch Site Names

- %sql Select distinct Launch_Site from SPACEXTBL

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

- There are 4 launch sites used in Space X for launching rockets. Distinct keyword used to extract the sites

Launch Site Names Begin with 'CCA'

- %sql Select * from SPACEXTBL where Launch_Site like 'CCA%' \limit 5

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

- These are 5 records whose launch site starts with CCA. Limit used to see only 5 records

Total Payload Mass

- %sql Select sum(PAYLOAD_MASS__KG_) from SPACEXTBL

sum(PAYLOAD_MASS__KG_)
619967

- The total payload mass is 619,967 kgs. Sum of all weights of the column has been using sum keyword

Average Payload Mass by F9 v1.1

- %sql Select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like '%F9 v1.1%'

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

- The average mass of payload for booster version F9 v1.1 is 2534.67kgs. Avg used to get average mass and like used to filter booster version

First Successful Ground Landing Date

- %sql Select min(Date) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)’

min(Date)
2015-12-22

- First Successful ground landing date was on 22/12/2015. min gives the earliest date and where used to filter the landing outcome

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql Select Booster_Version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000

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

- There are 4 booster versions where drone ship landing was successful between 4000 and 6000 kgs. Where and between used here for filtering

Total Number of Successful and Failure Mission Outcomes

- %sql SELECT Mission_Outcome, COUNT(Mission_Outcome) AS count FROM SPACEXTBL GROUP BY Mission_Outcome;

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

- There has been only 1 mission outcome failure amongst all launches. Count has been used to get the number of success and failure

Boosters Carried Maximum Payload

- %sql Select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)

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

- The above list has all the boosters that have carried maximum payload. Subquery has been used to get the max payload

2015 Launch Records

- %sql Select substr(Date, 6,2) as Month, Booster_Version, Launch_Site from SPACEXTBL where Landing_Outcome = 'Failure (drone ship)' and substr(Date,1,4)='2015'

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

- There has been 2 failures in the year 2015 in the months of Jan and April. Where and substring has been used to filter the database

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

- %sql SELECT Landing_Outcome, COUNT(Landing_Outcome) AS count FROM SPACEXTBL where Date between '2010-06-04' and '2017-03-20' GROUP BY Landing_Outcome;

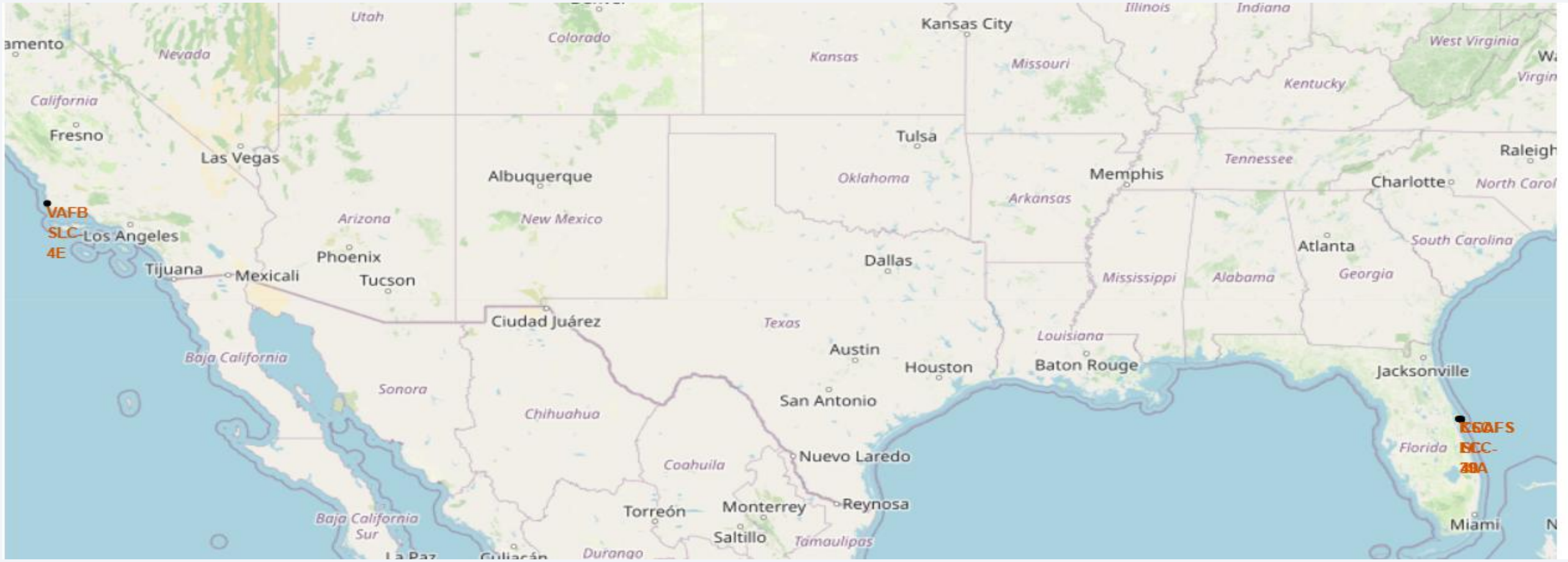
Landing_Outcome	count
Controlled (ocean)	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	10
Precluded (drone ship)	1
Success (drone ship)	5
Success (ground pad)	3
Uncontrolled (ocean)	2

- The above list is the landing outcomes of all launches from 2010-06-04 to 2017-03-20. between used to filter the dates and they have been grouped by landing outcome

Section 3

Launch Sites Proximities Analysis

Location of Launch Sites

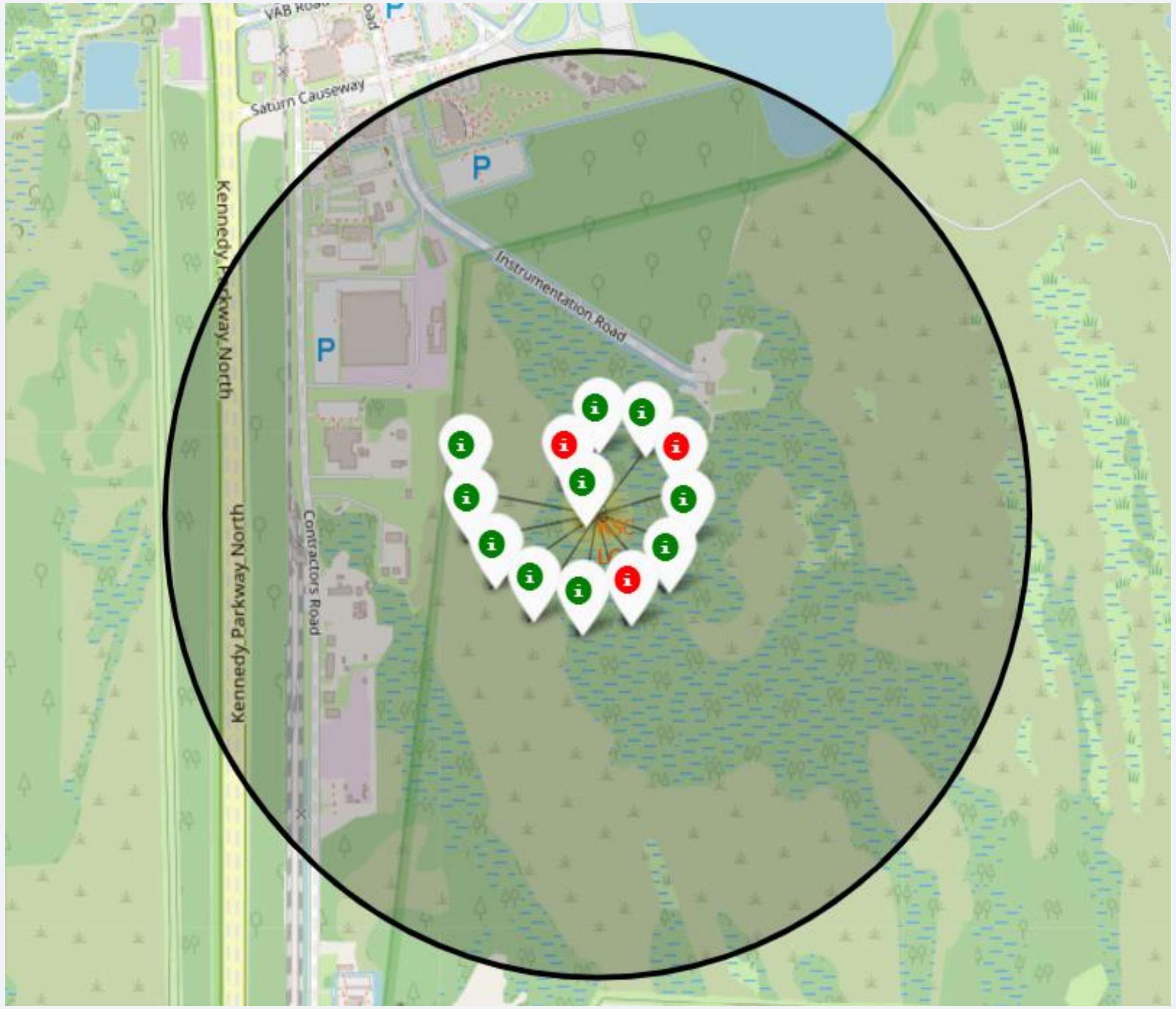


- There are 4 launch sites used by SpaceX. 3 of them are located in the state of Florida and 1 is in California.

Success/ Failed Launches of each Site

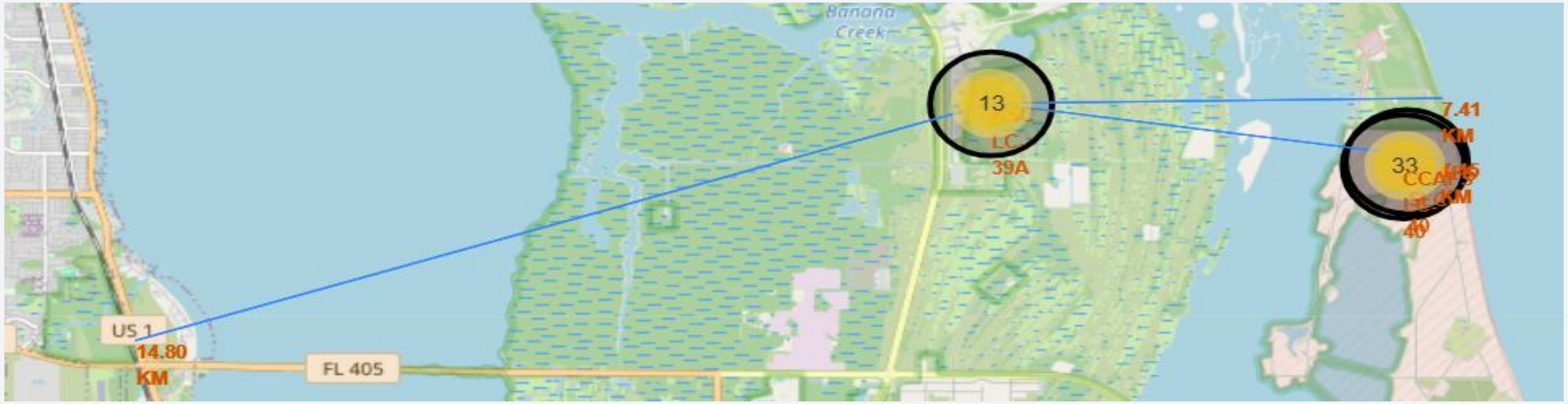
- The number of successful and unsuccessful launches of each site has been given using Marker Cluster

Success/ Failed Launches of each Site



- This is the zoomed in image of the Cluster for the site KSC LC-39A.
- The green and red icons show the number of successful and unsuccessful launches of SpaceX satellites in this particular launch site.

Proximity of Launch Site to other Locations

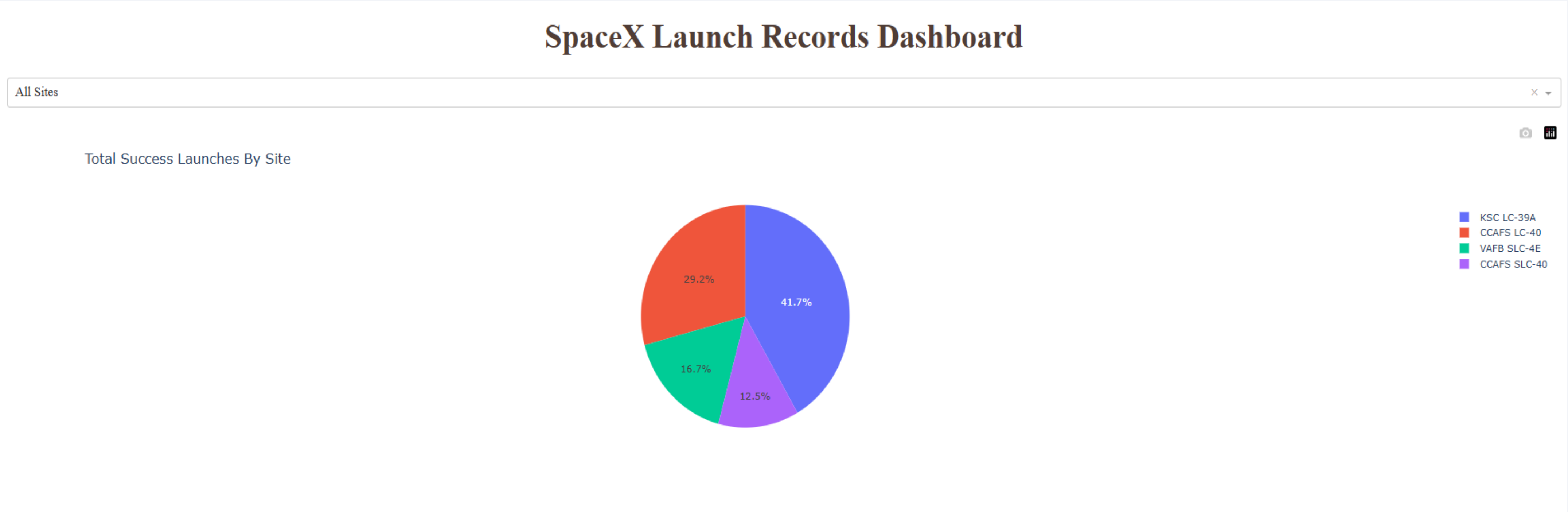


- The launch site is close to the coastline. All 4 launch sites are close to the coastline.
- They have highways and other roads pretty close to the launch site which helps in logistics
- There are placed on the outskirts of a city. This launch site 15kms away from the nearest city

Section 4

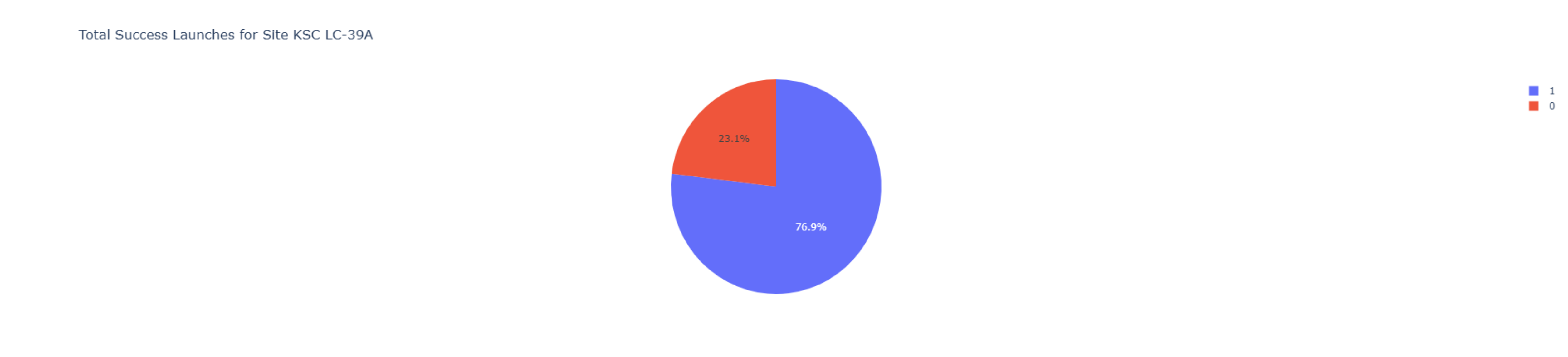
Build a Dashboard with Plotly Dash

Total Success Launches by Site



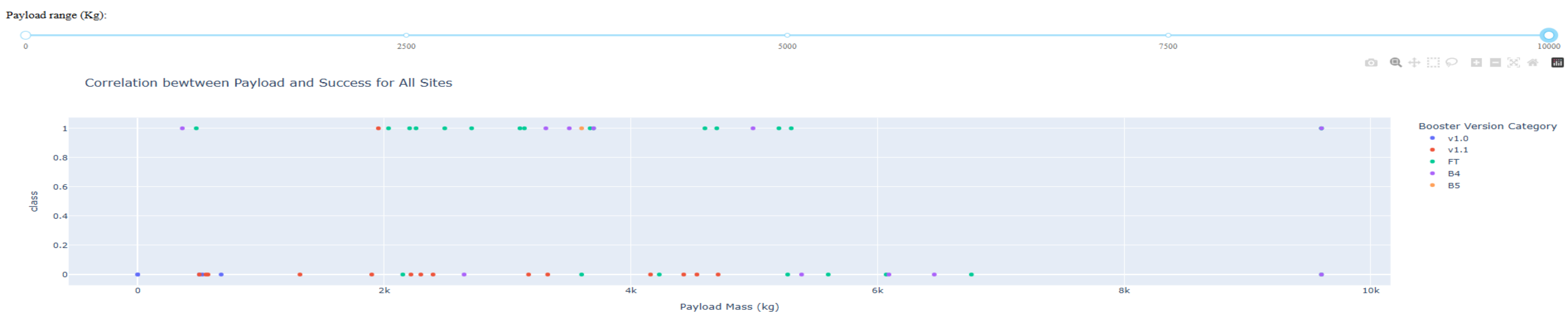
The highest number of successful launches is from the site KSC LC -39A accounting for about 42% of total successful launches. The share of other sites has been visualized in the pie chart with the names given in the side.

Site with the highest launch success ratio



KSC LC -39A site has the highest launch success ratio of 77% compared to all other sites. In the legend, 1 represents successful launches and 0 failures

Payload vs Launch Outcome Scatter Plot



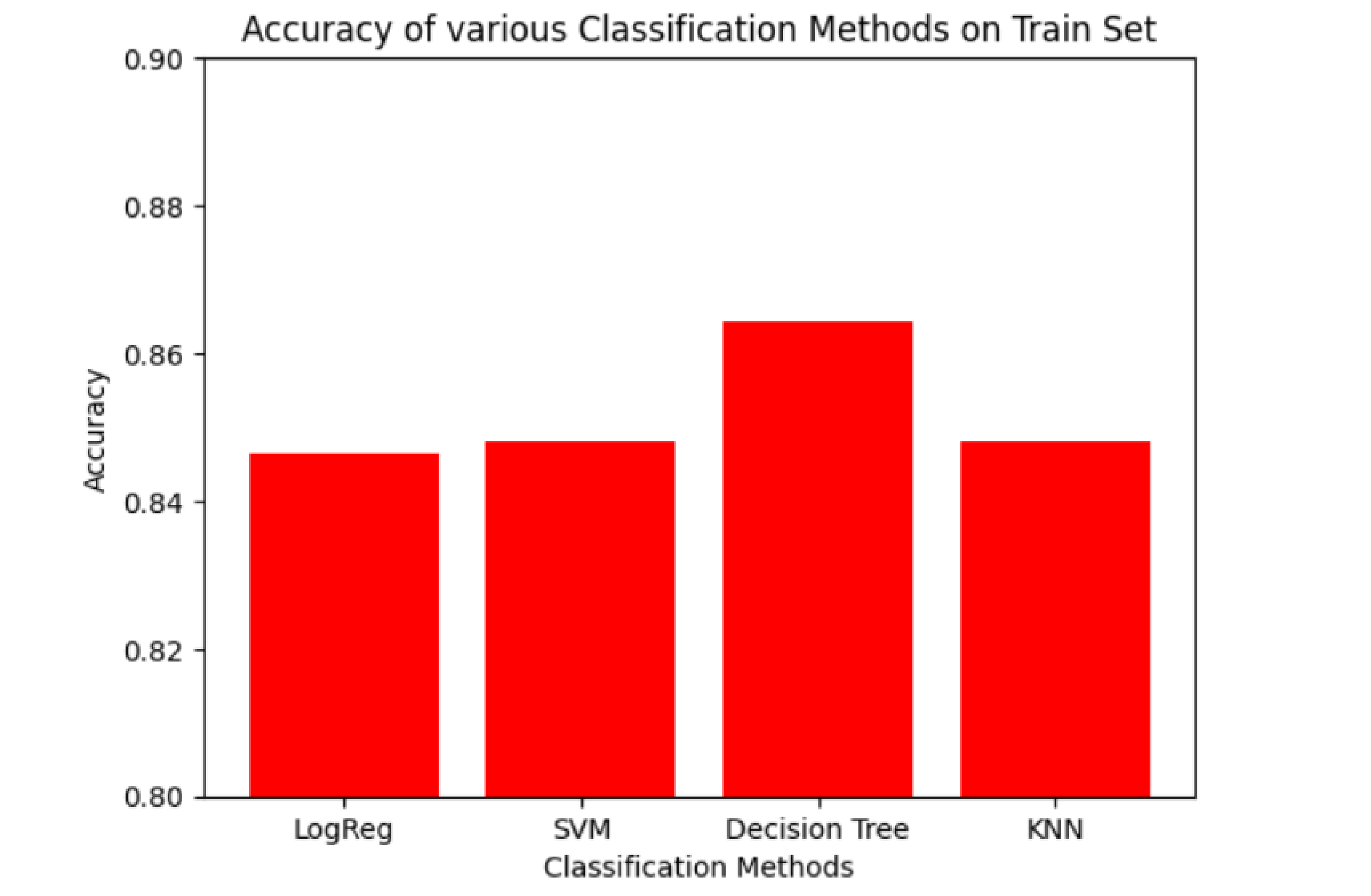
- This is a scatter plot between payload mass and launch outcome. The points at the bottom(class 0) are failures and ones at the top(class 1) are successful launches.
- The colors represent the booster version used for the rockets and we can see how different boosters are used for different mass and their success and failures

Section 5

Predictive Analysis (Classification)

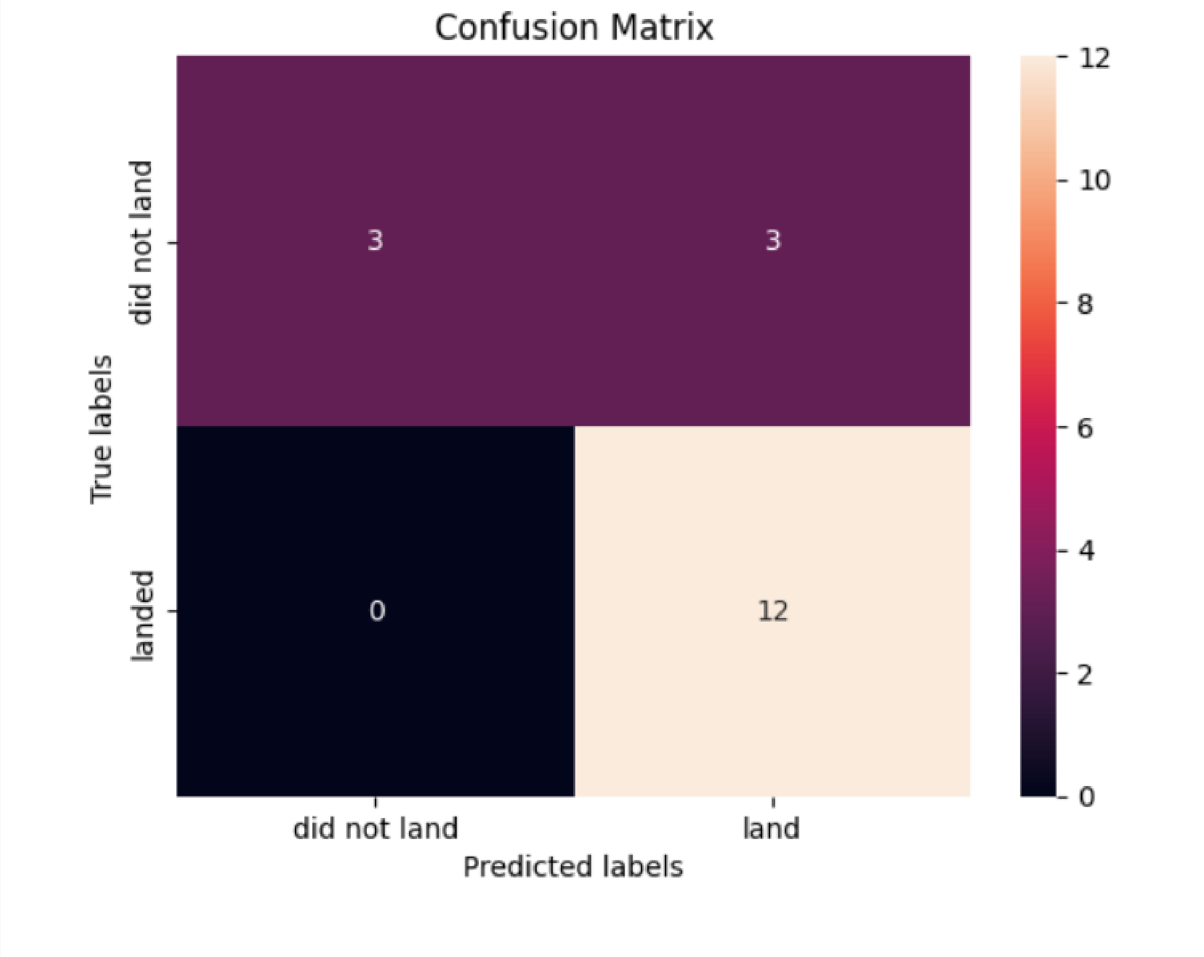
Classification Accuracy

- The bar chart represents how accurately the models classify the training set
- From the chart we can conclude that Decision Tree is the best method to use for this dataset.



Confusion Matrix

- This is the confusion matrix of the test set done using Decision Tree classification method
- From the confusion matrix, we can see that there are 18 samples in the test set.
- The model is accurately predicting when the satellite is successfully landing
- The accuracy is less when it comes to unsuccessful landings where it classifies 3 unsuccessful landings as successful.



Conclusions

- All launch vehicles have a specific payload that they can carry with VAFB SLC 4E site used for lower payloads
- Success of a launch depends on the orbit where it is launched with orbits such as GEO, LEO, and SSO demonstrated higher success rates, whereas certain orbits like SO had no successful missions.
- Heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. Payload greater than 10000kgs have been placed either in the ISS or VLEO orbit.

Conclusions

- A clear upward trend in launch success rates was observed between 2010 and 2020, indicating improvements in SpaceX's technology and operational efficiency
- Among the applied models, the Decision Tree method achieved the best training accuracy, offering reliable predictions for launch outcomes.
- The test accuracy, precision and recall are 83.3%, 80% and 100%.

Appendix

- The link to the SpaceX API can be accessed Here.
- The link to the Falcon9 Launch Vehicle Data used for web scraping is provided Here.
- All code snippets are available on GitHub via the following link Here.

Thank you!