IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vishithraj Shetty
11/04/2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - The dataset pertaining to SpaceX Falcon9 rocket launch was subjected to Data Collection, Data Wrangling, Extrapolatory Data Analysis, Interactive Data Analysis, Visual Data Analysis and Predictive Analysis
- Summary of all results
  - Different features/parameters were tested to see what affects the success rate of the rocket launches and each result was observed to determine the intensity of the effect that the feature had on the outcome
  - Different classification methods, charts and EDA were performed/drawn to get the results

Introduction

- We will predict if the Falcon 9 first stage will land successfully. 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. In this lab, you will collect and make sure the data is in the correct format from an API. The following is an example of a successful and launch.

Methodology

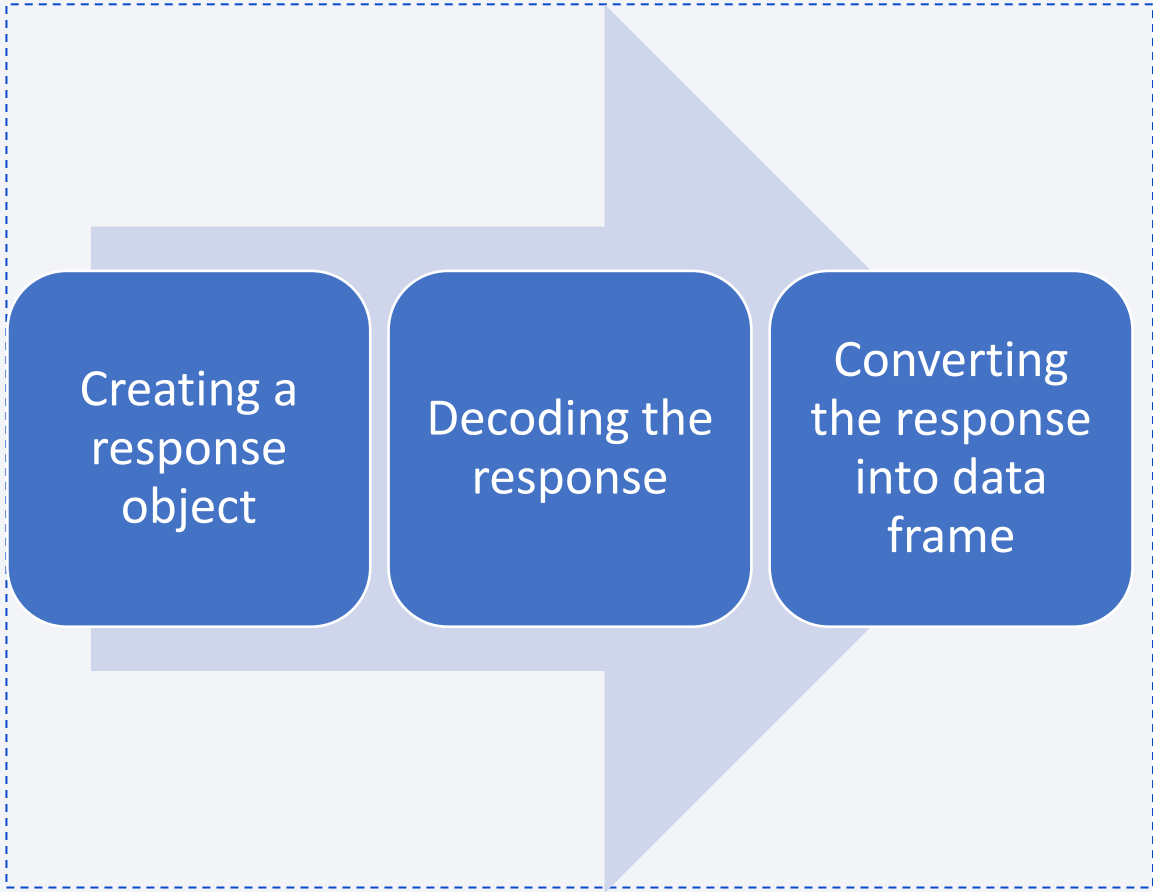
STEPS PERFORMED ON THE SPACEX DATASET:

- Data collection
- Data wrangling
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models

Data Collection – SpaceX API

- Used a static response object
- Converted into json format using .json()
- Converted the now json formatted data into a data frame using pd.json_normalize()
- Checked the column rows and headers using .head()
- GitHub URL for the file:

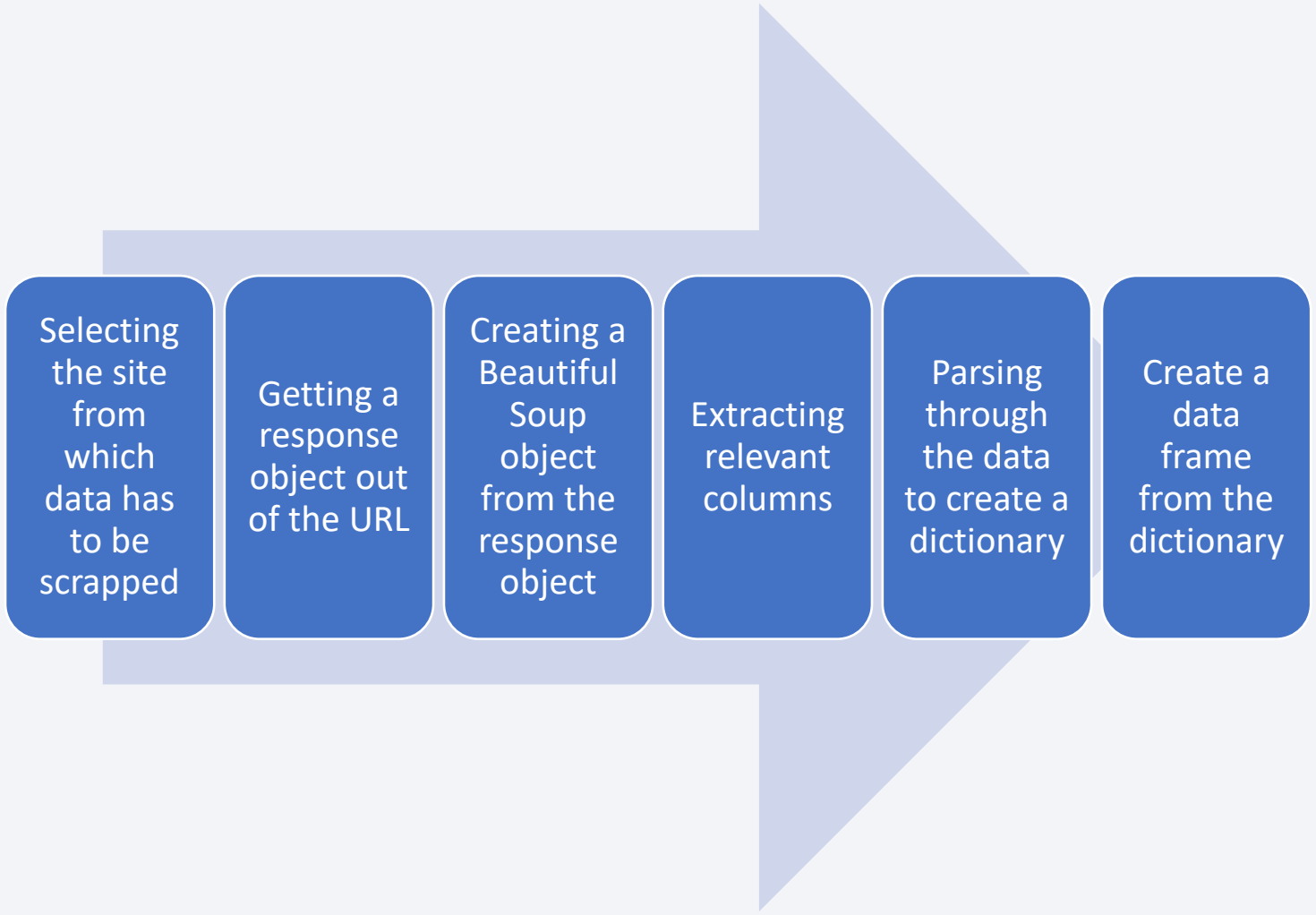
https://github.com/Vishi14/IBM_Capstone_Project/blob/main/Data%20Collection.ipynb



Data Collection - Scraping

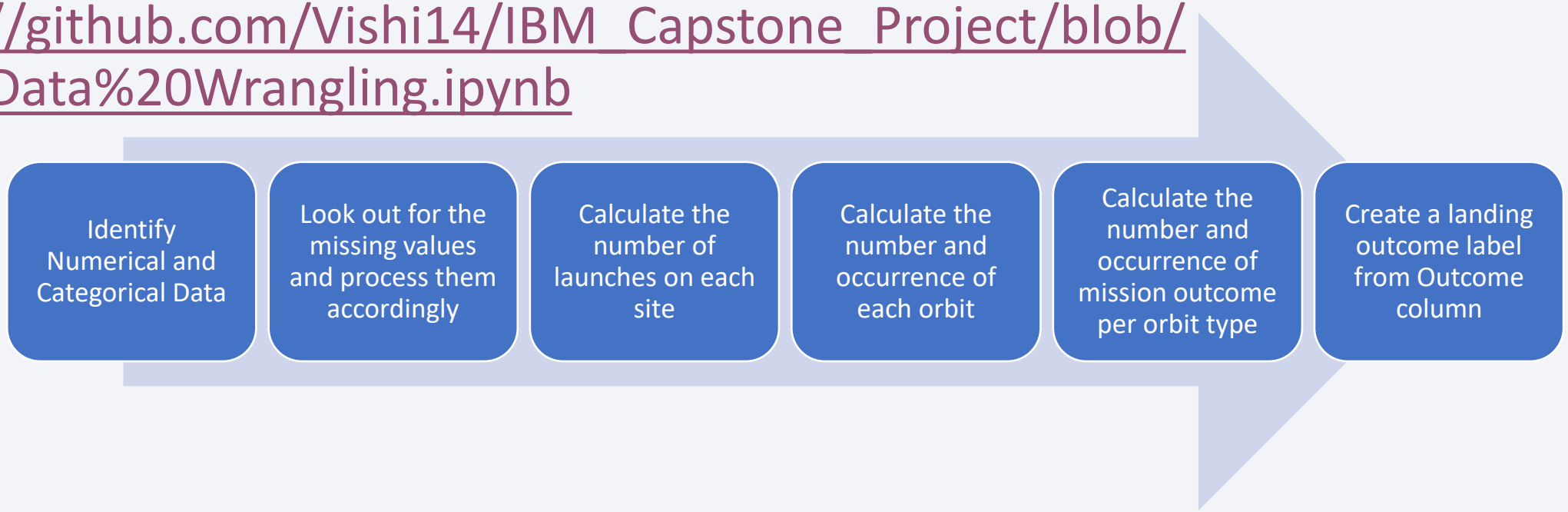
- GitHub URL for webscrapping file:

https://github.com/Vishi14/IBM_Capstone_Project/blob/main/Webscraping.ipynb

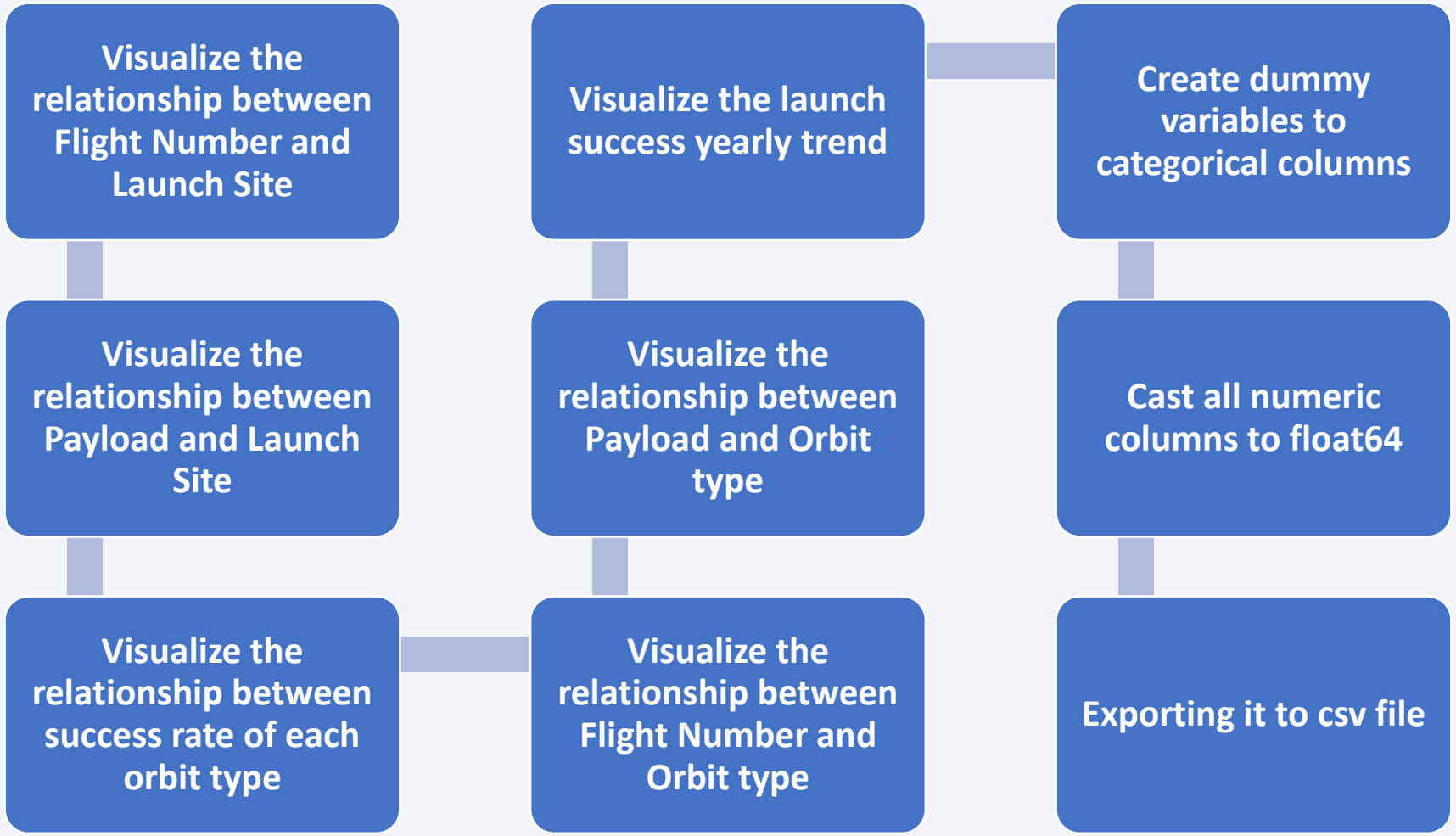


Data Wrangling

- The following flowchart gives a brief outlook on the steps taken to process the data. Various functions were used to bring about this process. It will be more clear in the below provided GitHub URL.
- https://github.com/Vishi14/IBM_Capstone_Project/blob/main/Data%20Wrangling.ipynb



EDA with Data Visualization



EDA with Data Visualization

- The charts that were plotted were:
  - Scatter Chart : Was used to identify patterns and trends in data
  - Bar Chart : Used for showing differences in data between different categories
  - Line Chart : Used to visualize changes in numerical data over time

- GitHub URL for EDA file:

https://github.com/Vishi14/IBM_Capstone_Project/blob/main/Exploratory%20Analysis%20Using%20Pandas%20and%20Matplotlib.ipynb

EDA with SQL

- ***SQL Queries performed:***

- Connecting to the database
- ***Displayed the names of the unique launch sites in the space mission***
- ***Displayed 5 records where launch sites begin with the string 'CCA'***
- ***Displayed the total payload mass carried by boosters launched by NASA (CRS)***
- ***Displayed average payload mass carried by booster version F9 v1.1***
- ***Listed the date when the first successful landing outcome in ground pad was achieved.***
- ***Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000***
- ***Listed the total number of successful and failure mission outcomes***
- ***Listed the names of the booster versions which have carried the maximum payload mass using a subquery***
- ***Listed the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015***
- ***Ranked 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***

- GitHub URL for SQL file:

https://github.com/Vishi14/IBM_Capstone_Project/blob/main/EDA-SQL.ipynb

Build an Interactive Map with Folium

Circle	Marker	Marker Cluster	Polyline
• Used to indicate a particular area of interest, such as a radius around a point of interest.	• A marker is a point on the map that can be customized with an icon, color, and popup text. It can be used to indicate a specific location or point of interest.	• Used to show a large number of markers without cluttering the map.	• Used to show a route or a path on the map.

- GitHub URL of folium file:

https://nbviewer.org/github/Vishi14/IBM_Capstone_Project/blob/main/Visual%20Analytics.ipynb

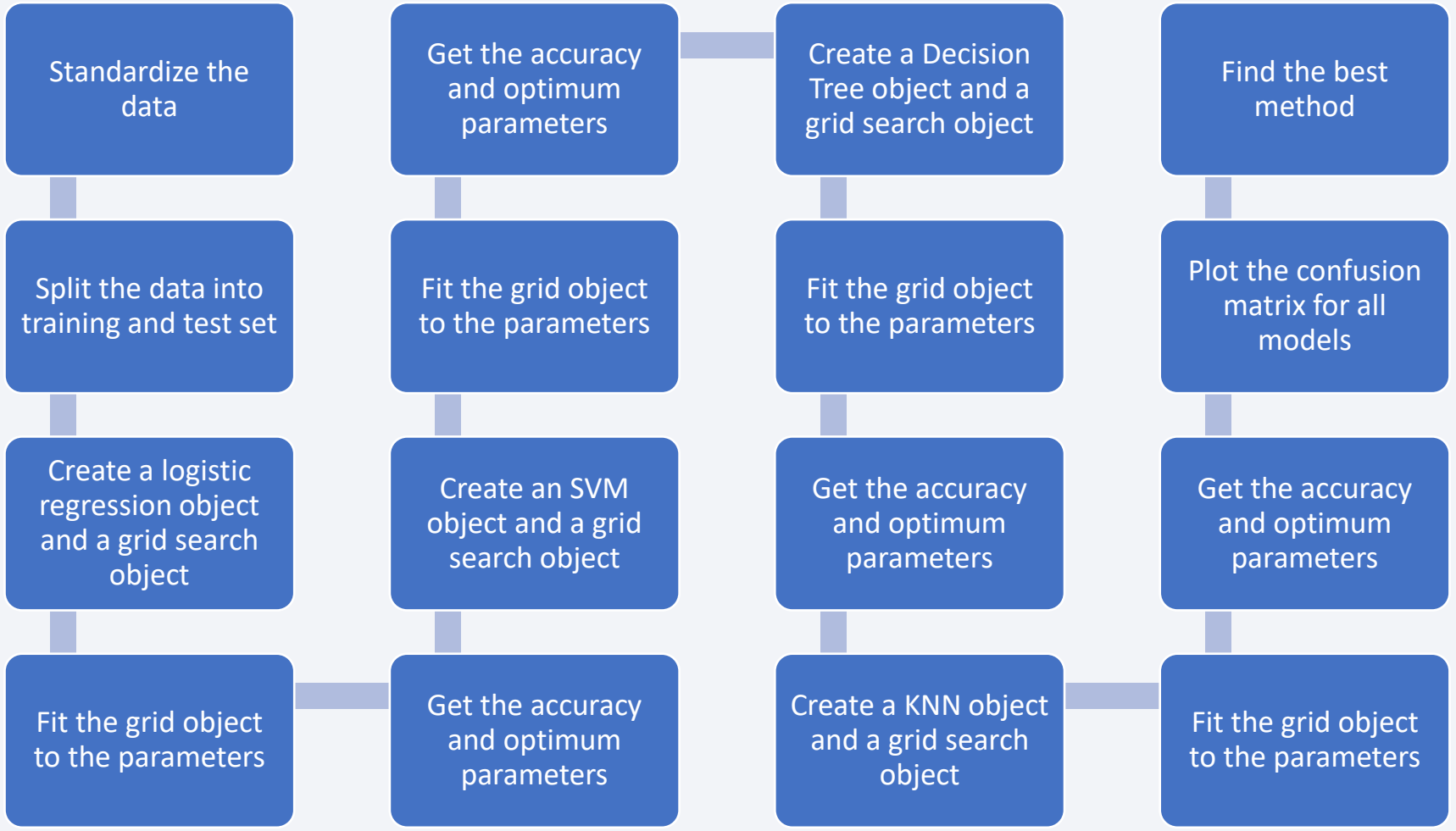
Build a Dashboard with Plotly Dash

- **Elements that were added to the dashboard were:**
  - A dropdown list so that the user can select the launch site
  - A range slider so that the user can select the range of the payload
  - A pie chart to express the success rate in launch of the rockets
  - A scatter plot to see the pattern in which different payload and launch site value affected the success rate of the launch

GitHub URL for the Plotly dash file (This is a python file[.py])

https://github.com/Vishi14/IBM_Capstone_Project/blob/main/Dashboard_plotlyDash.py

Predictive Analysis (Classification)



Predictive Analysis (Classification)

- The observation was, all the models performed the same with an accuracy of 83.33%
- This might be because the dataset was small.
- GitHub URL for Predictive Analysis file:

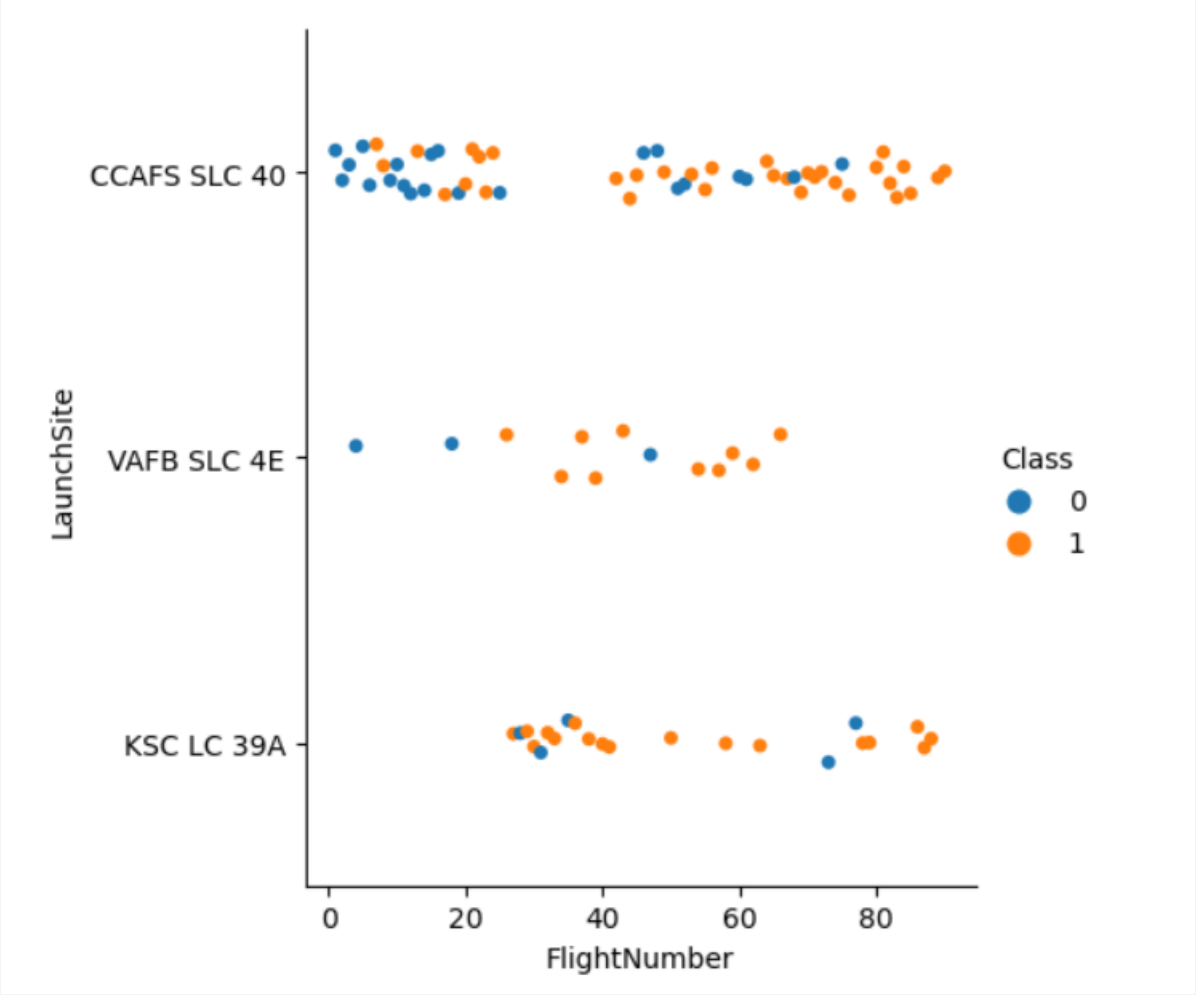
https://github.com/Vishi14/IBM_Capstone_Project/blob/main/Predictive_Analysis.ipynb

Results

- The following slides will contain the results of the data analysis.

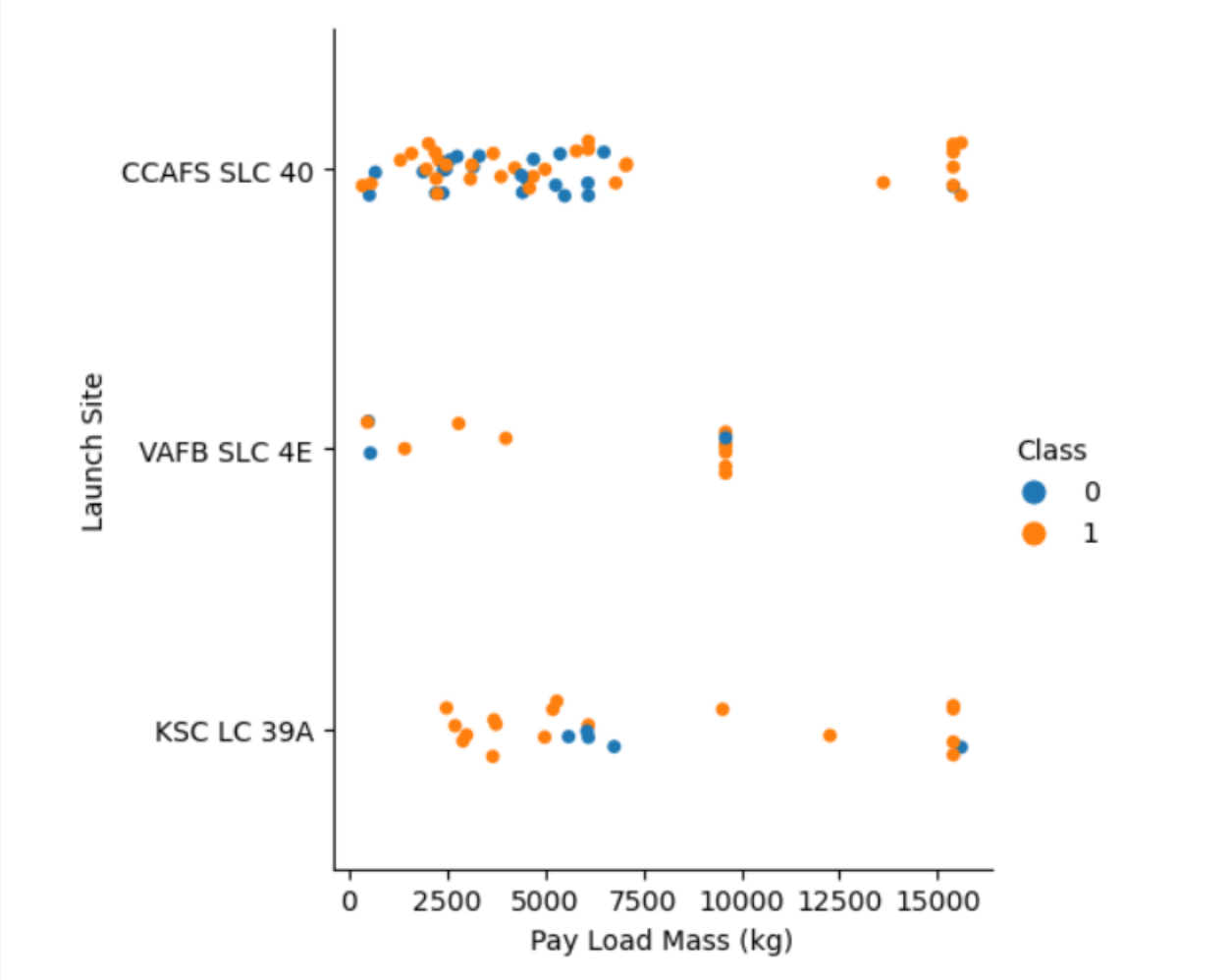
Flight Number vs. Launch Site

- From the image we can see that for FlightNumber more than 80, launch sites KSC LC 39A and CCAFS SLC 40 had successful launches
- Below 80, there is no particular pattern
- But it is fairly possible that as flightnumber increases success rate increases except for a few exceptions



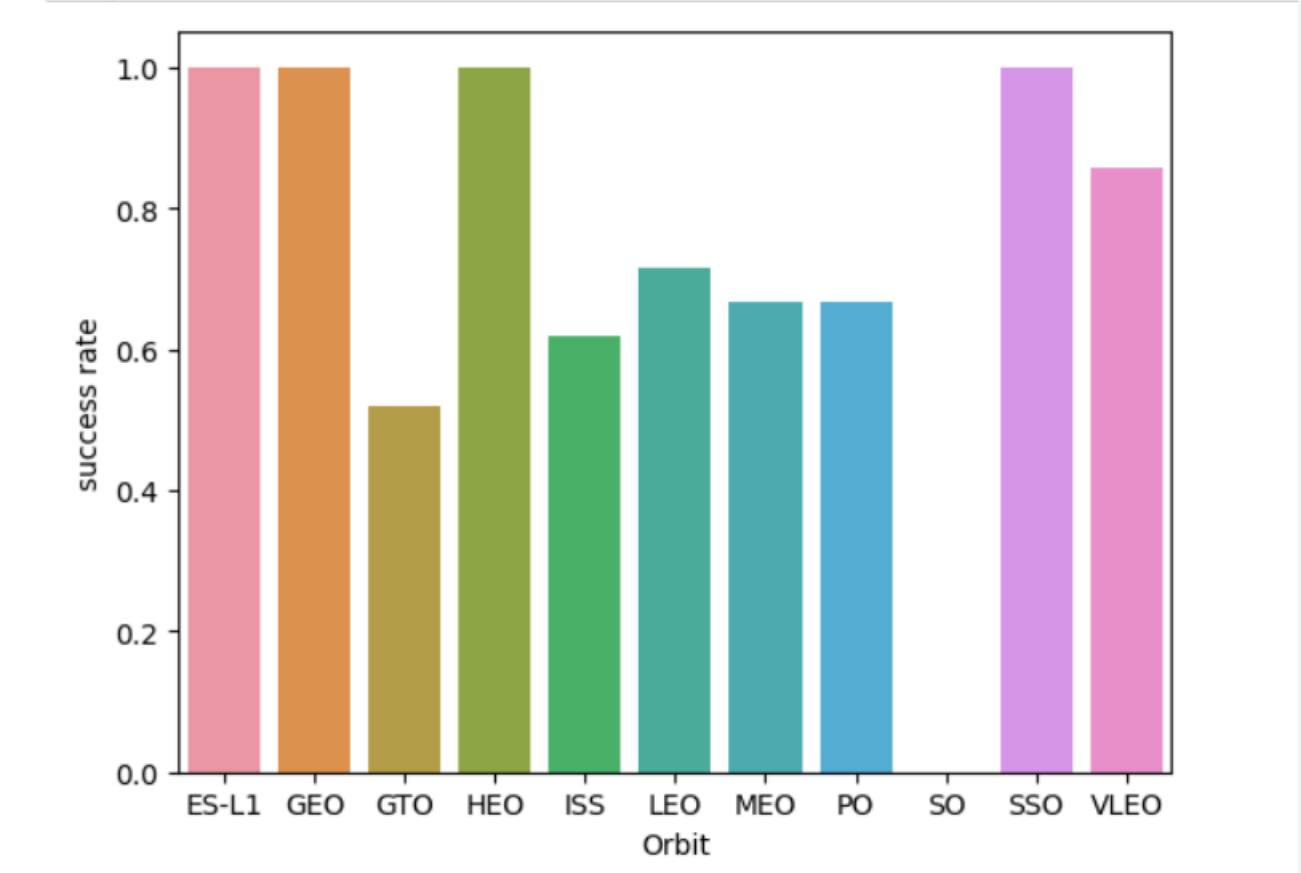
Payload vs. Launch Site

- There is no particular pattern observed
- It is observed that launch sites KSC LC 39A and CCAFS SLC 40 can process heavy payload beyond 12500 kg
- VAFB SLC 4E has not launched rockets with a payload higher than 10000 kg



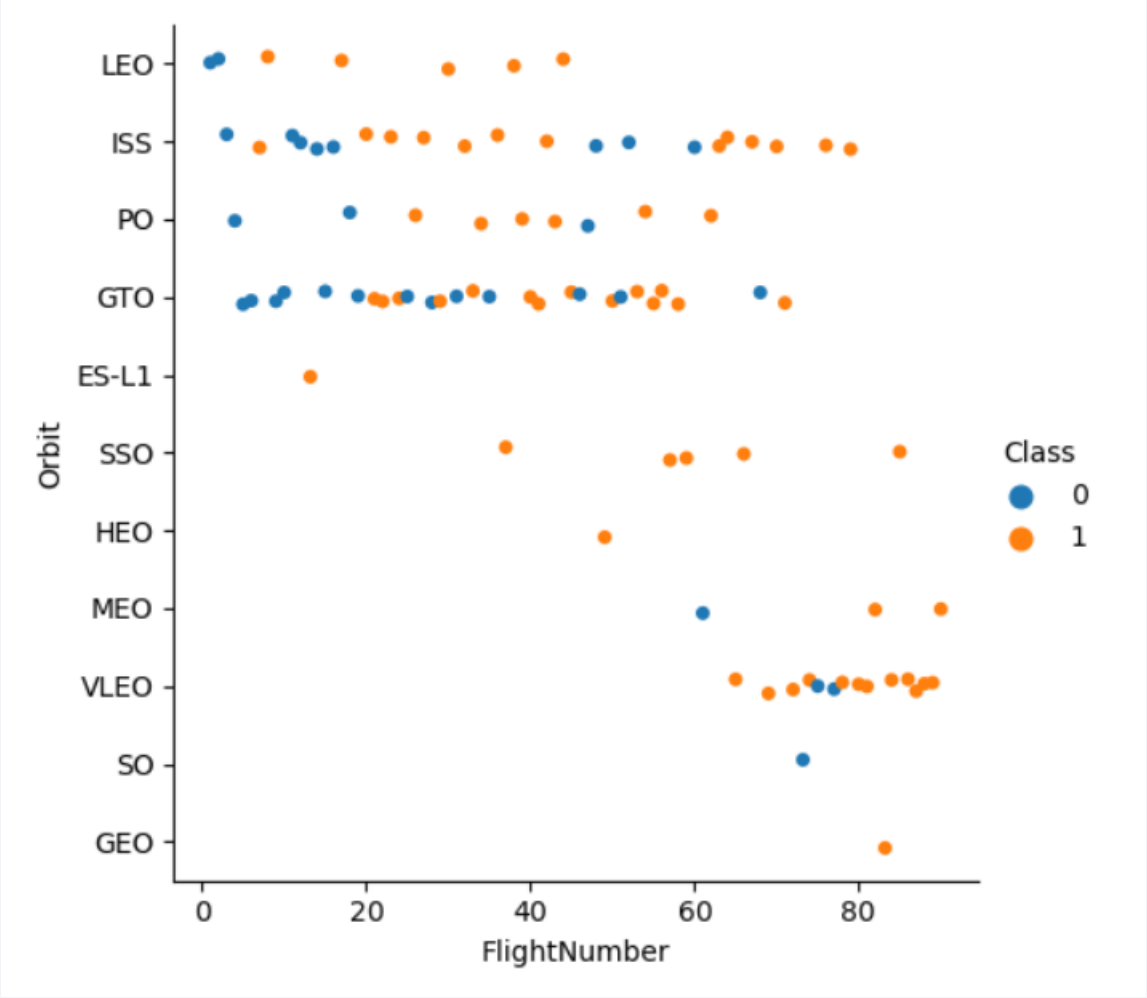
Success Rate vs. Orbit Type

- Orbit types: ES-L1, GEO, HEO and SSO all have very high success rates
- GTO has the lowest success rate



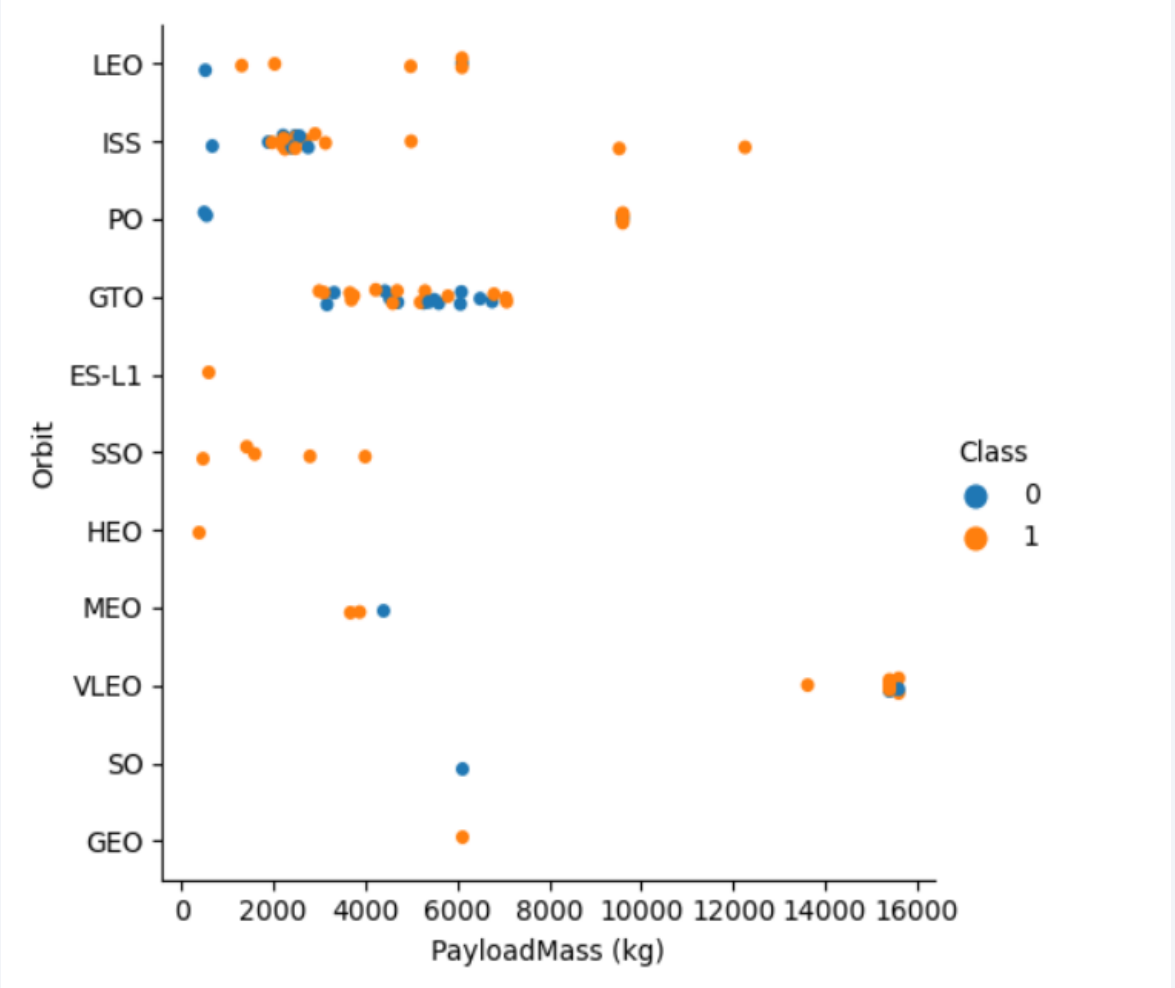
Flight Number vs. Orbit Type

- 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.



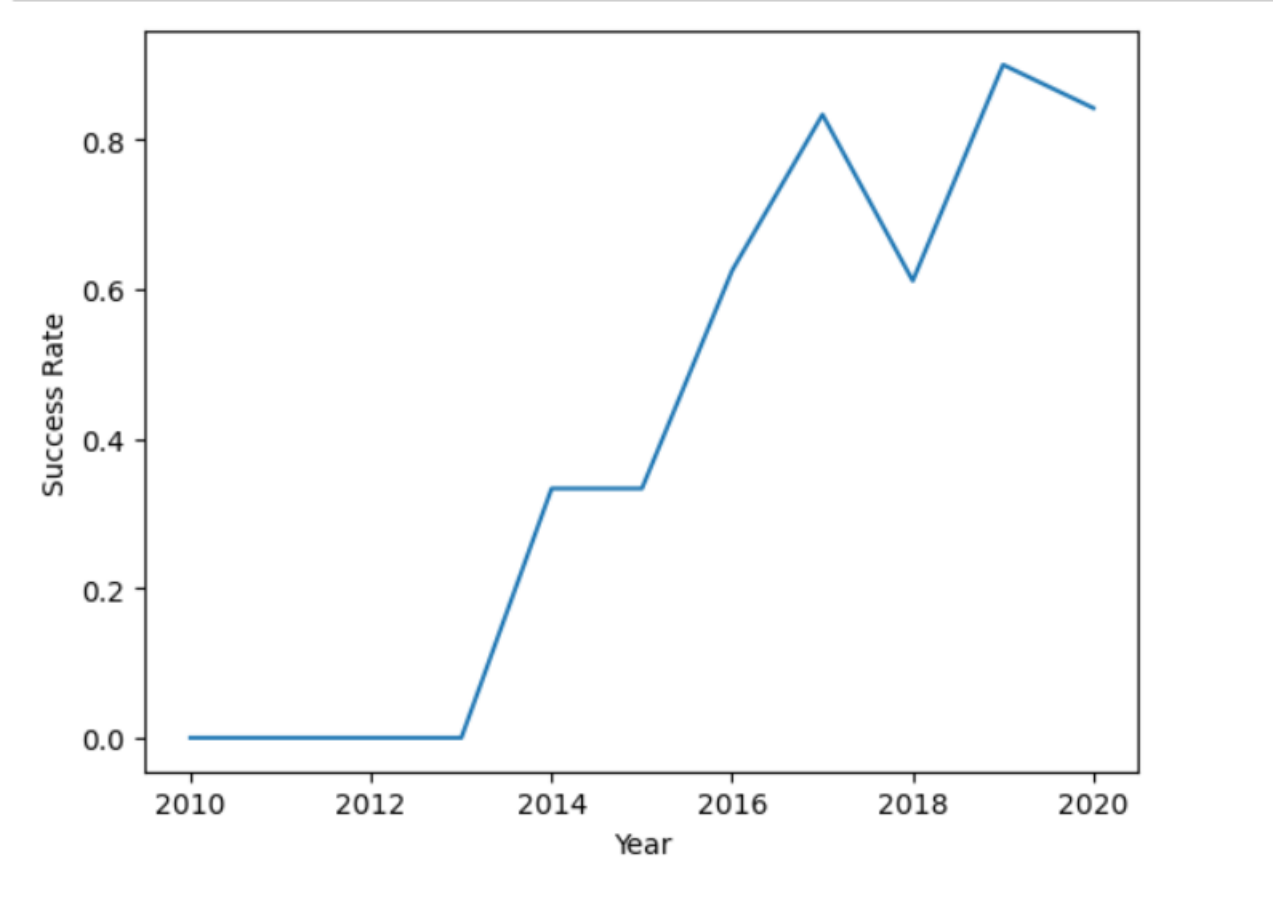
Payload vs. Orbit Type

- 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

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



All Launch Site Names

- The Launch Sites are:
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
  - CCAFS LC-40

This was extracted by using the DISTINCT function in SQL

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

Launch Site Names Begin with 'CCA'

- This was extracted by making use of LIKE and wildcard.
- Also Limit function was used to get only 5 values

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Here the GROUP BY function was used to group all the booster versions and then the aggregate function SUM() was used to get the total sum of payloads for each booster version category

booster_version	total_payload_mass_kg
F9 B4 B1039.2	2647
F9 B4 B1040.2	5384
F9 B4 B1041.2	9600
F9 B4 B1043.2	6460
F9 B4 B1039.1	3310
F9 B4 B1040.1	4990
F9 B4 B1041.1	9600
F9 B4 B1042.1	3500
F9 B4 B1043.1	5000
F9 B4 B1044	6092

Average Payload Mass by F9 v1.1

- Here the GROUP BY FUNCTION was used to group the booster version category “F9 v1.1”.
- MEAN function was used to find the mean payload mass
- HAVING clause was used to filter out booster version F9 v1.1

booster_version	mean_payload_mass
F9 v1.1	2928

First Successful Ground Landing Date

- Here the Date was found to be 2015-12-22
- Here the main functions were ORDER BY and LIMIT
- The table was sorted in the descending order by Date and the first value was extracted

DATE	landing__outcome
2015-12-22	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- Here a simple WHERE clause was used
- The conditions were given into the WHERE CLAUSE where AND had to be used because they were 2 conditions to be satisfied

booster_version	payload_mass__kg_	landing__outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- Used the GROUP BY clause to group the mission outcomes together
- The COUNT function was used to give the total count of each outcome
- It could be seen that the success rate is very high

Done.

mission_outcome	count_mission
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Here a subquery was used because of the aggregate function, MAX() to used in the WHERE clause.
- The result is shown as in the picture

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

- Here the WHERE, LIKE clauses were used
- LIKE and wildcard was used to filter the year 2015
- Landing_outcome was set to Failure(drone ship)

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

- The landing outcomes were grouped together with the help of GROUP BY clause.
- COUNT function was used to see the frequency of each outcome
- Then the outcomes were sorted in descending order on the basis of the count
- It could be seen that No_attempt has the highest rank and Precluded(drone ship) has the lowest rank

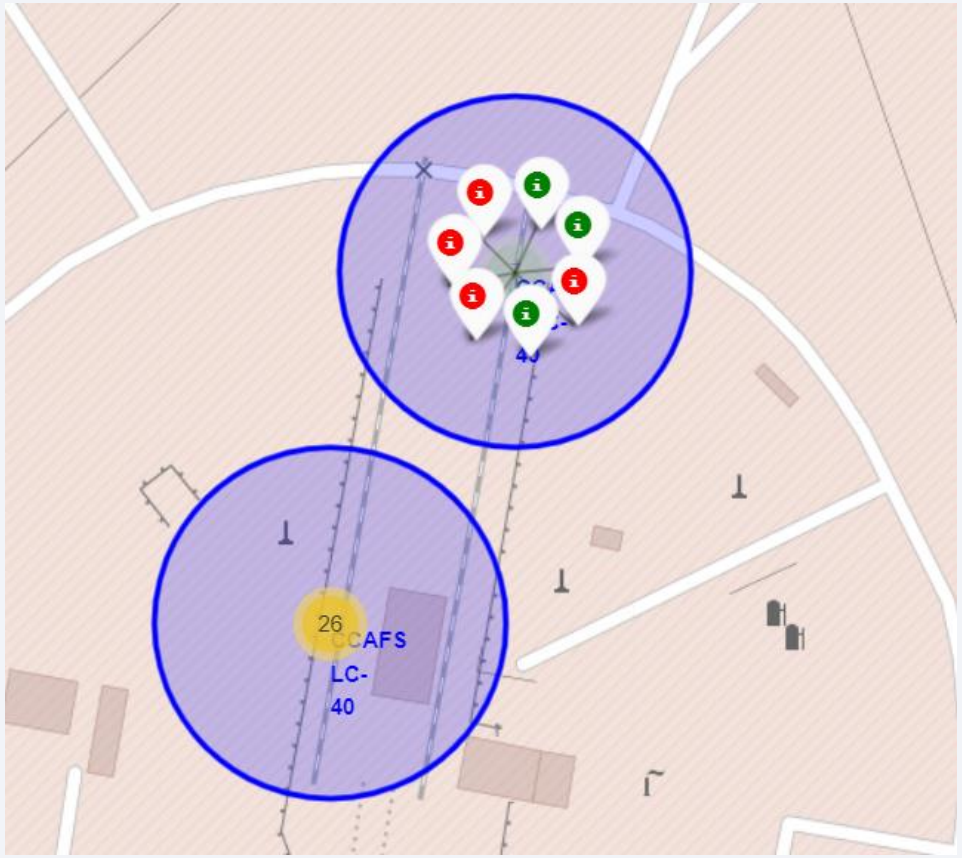
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

LAUNCH SITE LOCATION

- It was observed that VAFB SLLC 4-E was situated on the West coast of USA
- All the other 3 launch sites are located on the East coast of USA

SUCCESSFUL/FAILED LAUNCHES ON EACH SITE

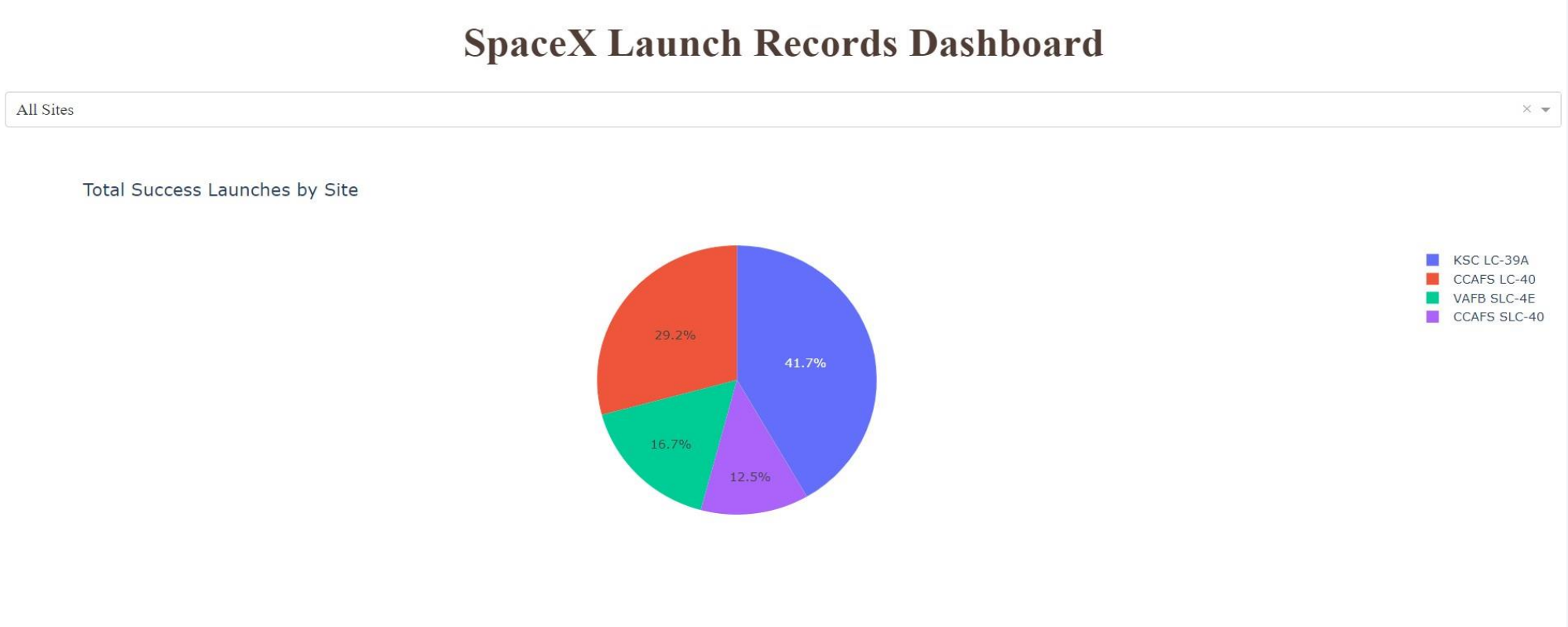
- This image pertains to the CCAFS SLC-40 launch site
- We can infer that there were 3 successful launches out of the 7 rocket launches at this site
- Similarly by exploring the map and markers for different sites we can get the corresponding results.



Distances between a launch site to its proximities

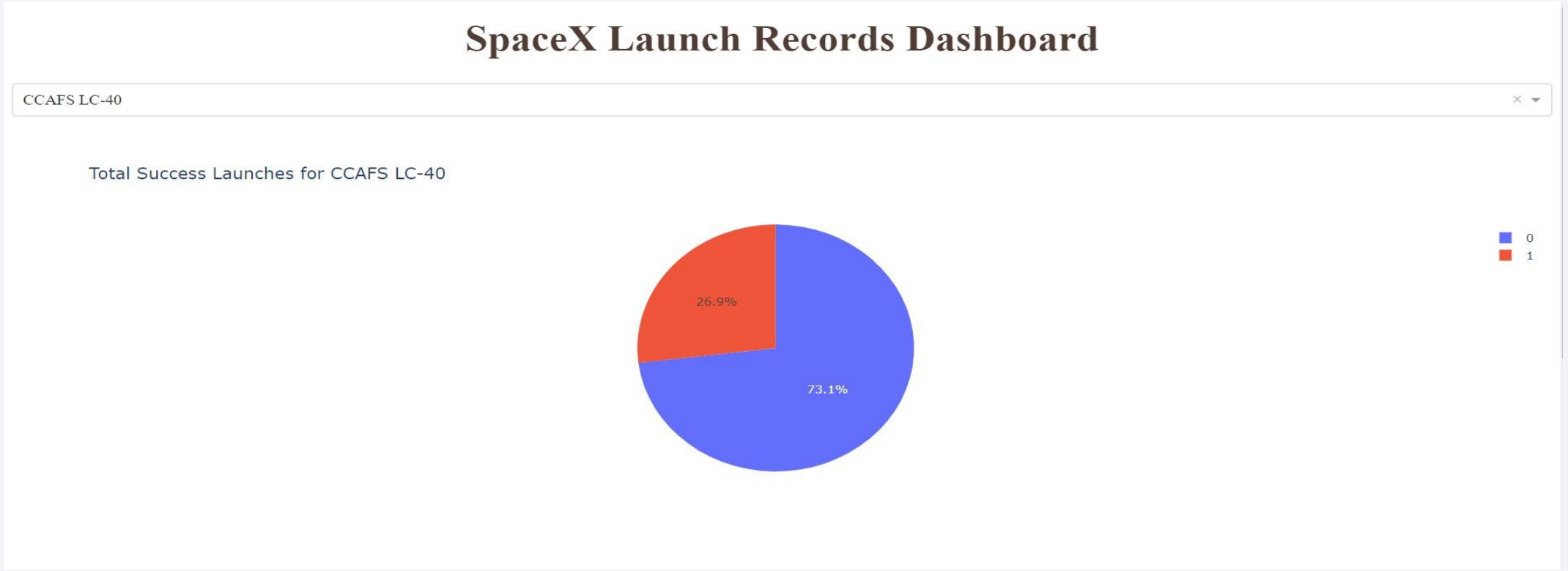
- From the image it can be seen that CCAFS SLC-40 launch site is 0.86 km from the nearest coast
- Its 0.62 km away from the nearest highway
- Its 0.98 km away from the nearest railway.
- Also it is 18.18 km away from the nearest city. This info is not visible in this image because its too zoomed in.

Successful Launches by all Launch Sites



- From the above pie chart it can inferred that Launch Site KSC LC 39-A had the most successful launches of any other launch sites
- On the other hand, CCAFS LC-40 had the least launch success

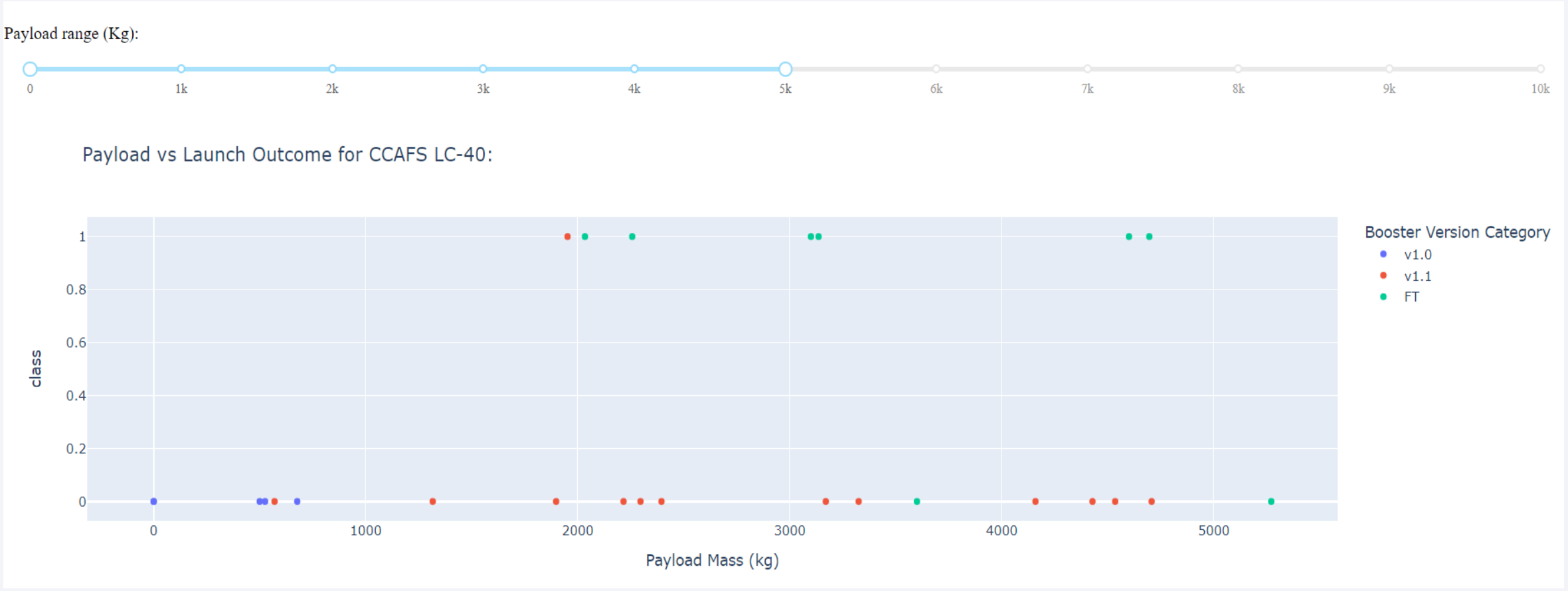
Launch Outcome in site CCAFS LC-40



- From the above pie chart it could seen that CCAFS LC-40 had majority launch successes
- Similar trend follows for rest of the launch sites except for KSC LC 39-A

Payload vs Launch Outcomes

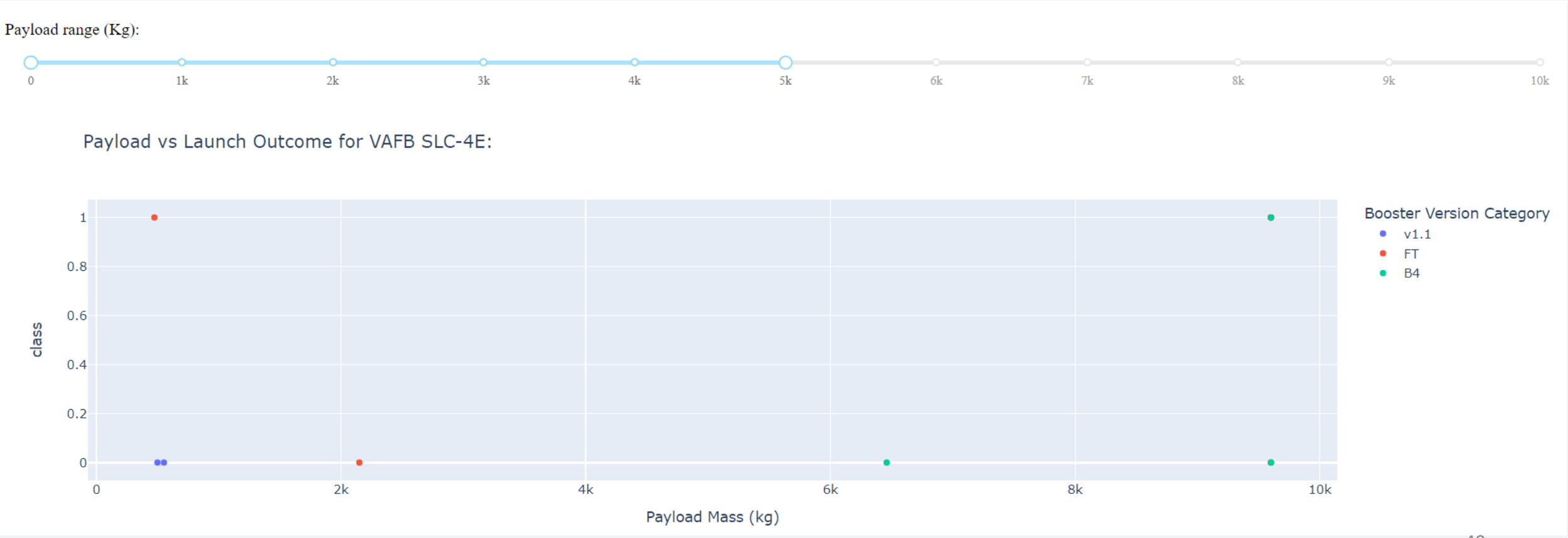
- Let us consider 0-5000 kg as the lower payload range and 5000-10000 kg as the upper payload range
- The below scatter plot is for the launch site CCAFS LS-40



CONT….

- From the previous slide it can be inferred, for the launch site CCAFS LC-40, Booster Version FT performed well for payloads between 2000-5000 kg
- Rest other versions did not perform.
- Also it can be observed that the maximum payload mass for this site may be approximately around 6000 kg

LAUNCH SITE VAFB SLC-4E

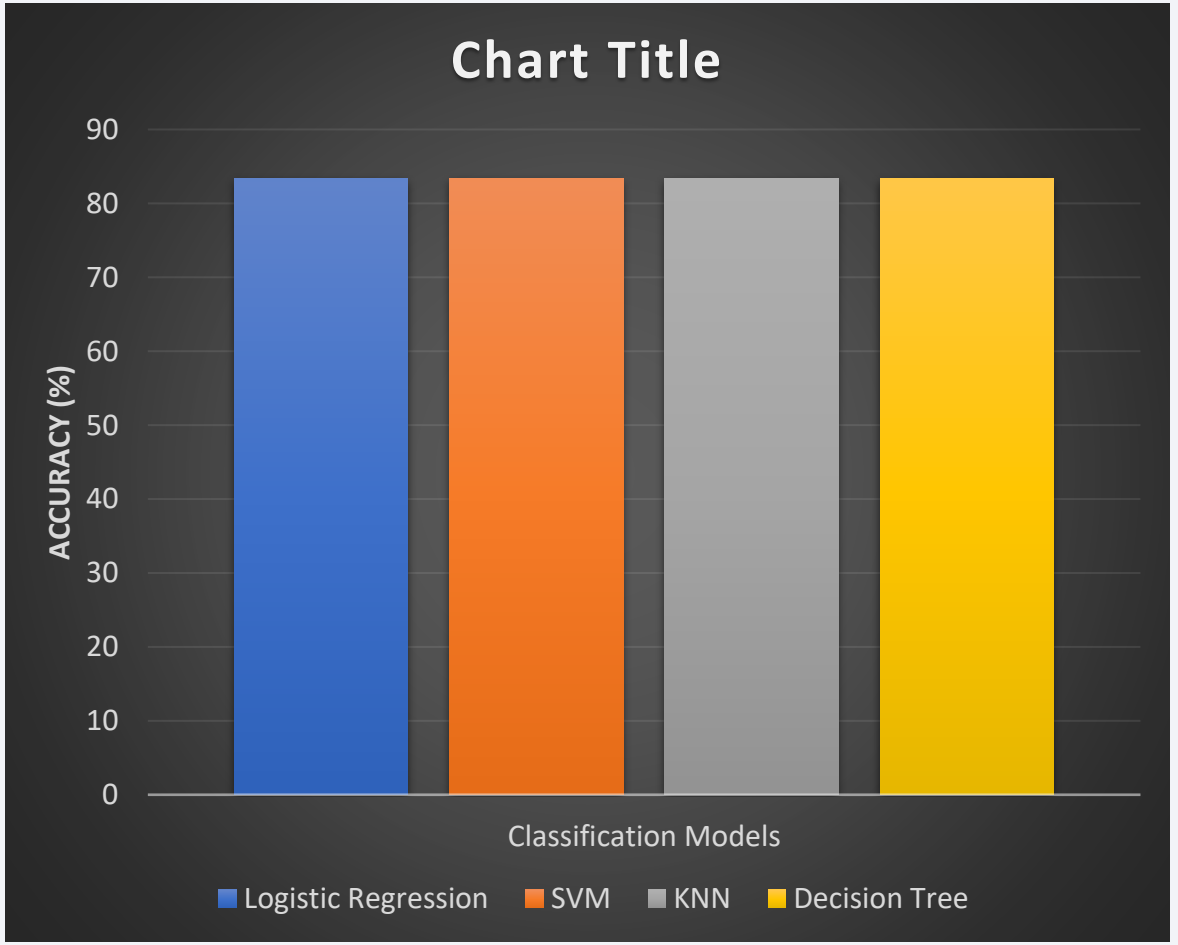


For this we can come up to the following conclusions:

- This site worked with heavy payloads
- The success rate seems to be low
- Booster version B4 seems to have to have succeeded and failed at the same heavy payload
- Booster Version FT had succeeded at lower payload

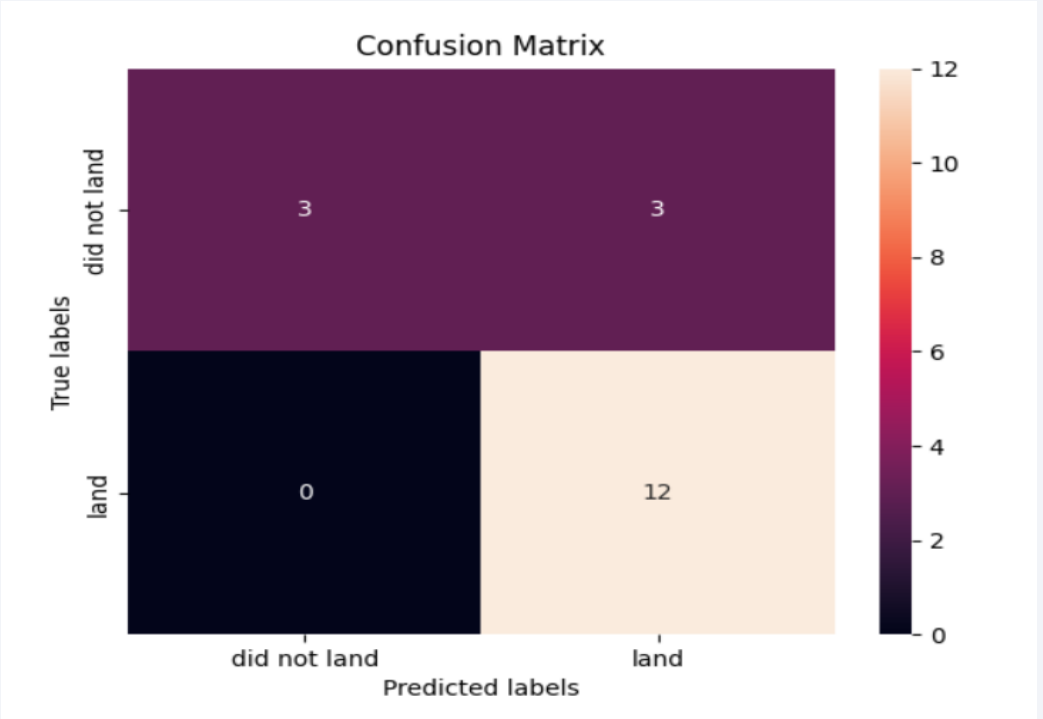
Classification Accuracy

- By performing the accuracy test on all the classification models, i.e. Logistic Regression, Support Vector Machine, K-nearest Neighbors and Decision Tree, it was found that all the models had the same accuracy.
- The accuracy was 83.33%



Confusion Matrix

- Since the accuracy for all classification models were the same, the confusion matrix for all the models were also same.
- We can observe that the problem lies with the False Positive.



Conclusions

- Features that were having the most effect on the success rate of the launch were:
  - ❑ Flight Number
  - ❑ Payload Mass
  - ❑ Orbit
  - ❑ Launch Site
  - ❑ Booster Versions
- It was observed that since the data set was small, any classification model can be used for predictive analysis.
- Launch Sites are mostly situated on the coasts away from cities
- Success rates increased with time, but also it completely relies on the launch site and the rocket payload. Thus if we have to increase the success rate of the launch outcomes, we have to make sure to choose the appropriate launch site, orbit, payload and booster versions. Optimizing all these parameters will help in increasing the success rate of rocket launch