IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vijaya Lakshmi Kedari
5-21-2025

Outline

Executive Summary

- The project is about analyzing whether the first stage of Falcon 9 launch is landed successfully.
- In the exploratory data analysis, we have realized that the launches are most successful when launched from the site KSC – LC-39A.
- Success rates at each launch site varies, with KSC-LC-39A launch site achieving the highest success rate at 76.9%.
- The success rate is higher for payloads between 2000 – 6000 kg across all sites.

Introduction

- SpaceX launches rockets at the cost of 62 million dollars, while launching by the other providers cost approximately 165 million dollars. The savings for SpaceX launches comes because they reuse the first stage.
- We want to understand the success of landing the first stage in order to get the actual cost of the launch. Specifically, we would like to understand how the success rate differs across launch sites, booster versions and payloads.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - We have collected data by making requests to SpaceX API. We have also used web scaping to collect data.
- Perform data wrangling
  - Missing values are replaced with the mean of the variable.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four models are fit using the data – support vector machine, logistic regression, random decision trees and KNN

Data Collection

- Data is collected from SpaceX API using web scraping.
- We request the data from several SpaceX API end points using get function.
- Using the identification data in each column of the launch dataset, we call the API again at several end points to extract the required data.
- In this way, we collect the data for BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.
- We have also retrieved data from the html contents using BeautifulSoup. Each table row in the html content is processed and the relevent data is extracted.

Data Collection – SpaceX API

- First we get data in json format from api.spacexdata.com/v4/launches/past.
- Then we use the identification numbers from the launch data as a reference to get the required data about cores, payloads, launches from the SpaceX API
- Github url for the RESEful API code: vijayakedari/Capstone

Get response from api.spacexdata.com/v4/launches/past

Convert the json data into a dataframe

Use the identification information from the dataframe to request for other data from SpaceX API

Data Collection - Scraping

- GitHub URL :

vijayakedari/Capstone

Retrieve the Html table pertaining to Falcon 9 launch

Define Helper functions to extract the data

Each table row is processed and the data is extracted using helper functions.

Data Wrangling

- We understood number of launches per site.
- We understood number of launches per orbit.
- We have identified bad outcomes and good outcomes from the landing_outcomes column.
- Based on which, a new column called 'Class' is created that takes the value 1 when the outcome is a good outcome and it takes 0, otherwise.
- GitHub URL: [vijayakedari/Capstone]()

EDA with Data Visualization

- We plotted the scatter plots between Launch Site and Pay load Mass, flight number and Orbit and Orbit Vs Pay load Mass
- A bar plot of Success Rate at different Orbits and a
- Line plot between year and success rate, which displayed an approximate linear trend
- We explored these graphs because we want to understand how launch success depends on various variables such as Orbit, Pay Load Mass, Flight Number and Launch Site.

EDA with SQL

- Using SQL queries, we looked for different launch sites,
- We understood what is an average payload for a booster version,
- Displayed the date when the landing on ground pad was first achieved.
- We retrieved total number of successes and failures.
- We retrieved Booster versions that have maximum payload.
- We ranked the count of landing outcomes between 2010-06-04 and 2017-03-20.
- GitHub URL : [vijayakedari/Capstone]()

Build an Interactive Map with Folium

- We have placed a marker and a circle at each Launch Site.
- We then created a cluster of makers at each launch site signifying whether a launch is success of failed.
- We have also understood the distance between launch sites and their proximities. This could allow us to understand whether the distance from the launch site proximities can affect the launch outcome.
- All launch sites are located near a coastline. Three launch sites are 75 km further away from a city, while another launch site is 15 KM away from the city. I did not detect any correlation between distance and the success rate of the launch site.

Build a Dashboard with Plotly Dash

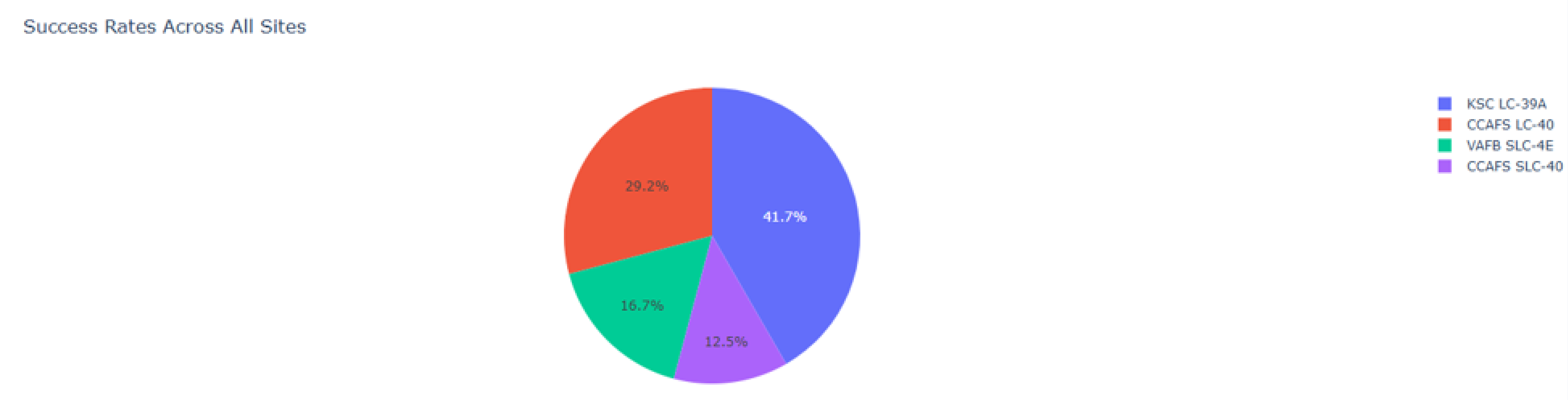
- A pie chart depicting the share of successful launches across all sites is added.
- Depending on the user's choice of the launch site, the success and failure rate at the site is displayed.
- Then we plot the scatter plot between pay load mass and the Class variable, which has success or failure of a launch encoded as 1 or 0. The payload mass is an interactive variable. The user can choose the maximum and minimum payload.
- It is important to understand the success rate of the launch per each launch site and also how the success rate depends on the chosen payload range.
- Add the GitHub URL : vijayakedari/Capstone

Predictive Analysis (Classification)

- We have built Four models - Logistic Regression, Random Decision Trees, Support Vector Machine and K Neighbors Classifier.
- First we divide the dataset into training and test set.
- Create an instance of the model.
- We then specify hyperparameters for tuning. We then use GridSearchCV to search for the best hyperparameters using 10 fold cross validation on the training set.
- We repeat the process for all the other models and choose the model with the best accuracy rate on test data.
- Add the GitHub URL: vijayakedari/Capstone

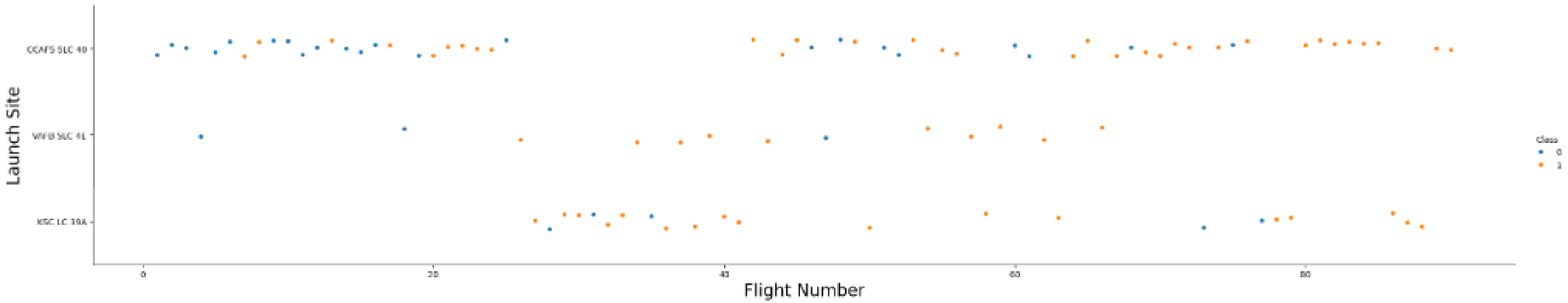
Results

- There is a linear trend with regards to success rate over the years – indicating improvement in technical expertise over time.
- The launch site KSC – LC-39A has the highest number of successful launches.
- Predictive analysis results – The models Linear Regression, Support Vector, machine and K- Neighbors perform equally well with 83.34 % accuracy on the test data.



Section 2

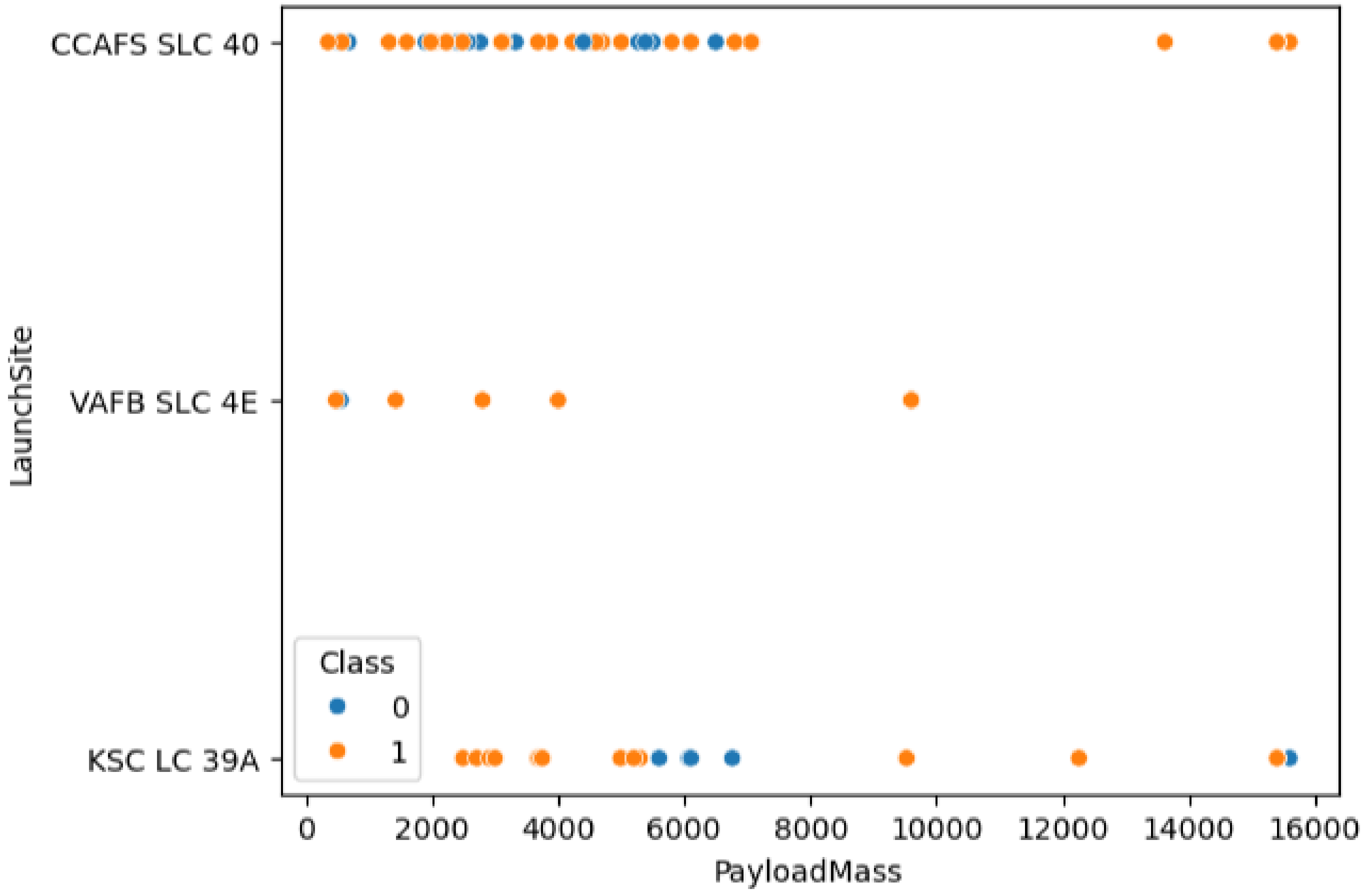
Insights drawn from EDA



Flight Number vs. Launch Site

- According to the graph, CCAFS-SLC40 launches are successful for higher flight numbers.
- Similarly, a few of the higher flight number are successful for other sites – KSC LC-39A and VAFB SLC04E.
- This indicates that the recent launches are more successful in landing the first stage across all sites.

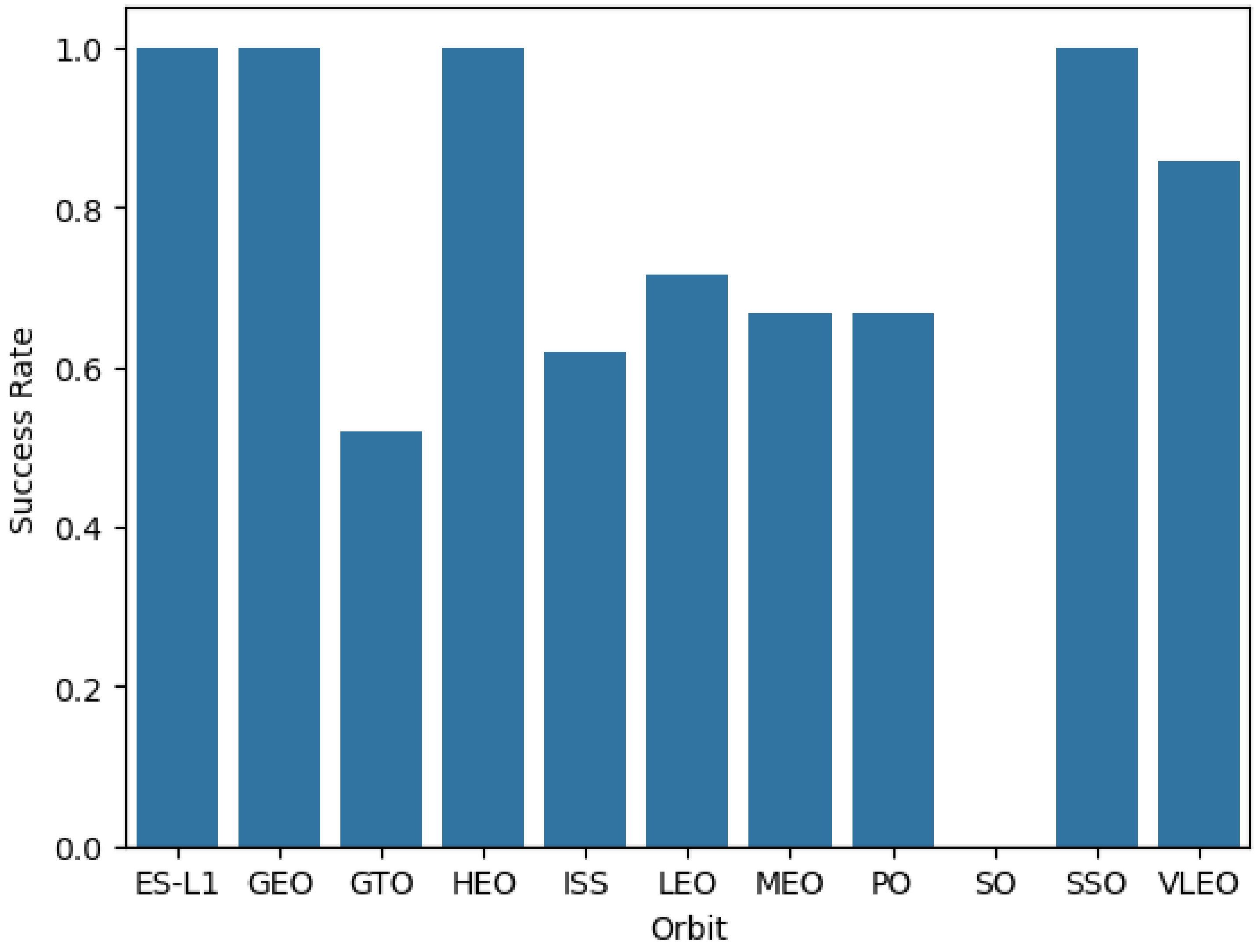
Payload vs. Launch Site



- Launches with higher payloads are successful at the site CAFS SLC 40.
- Similarly, higher payload launches are successful at site VAFB SLC 4E.

Success Rate vs. Orbit Type

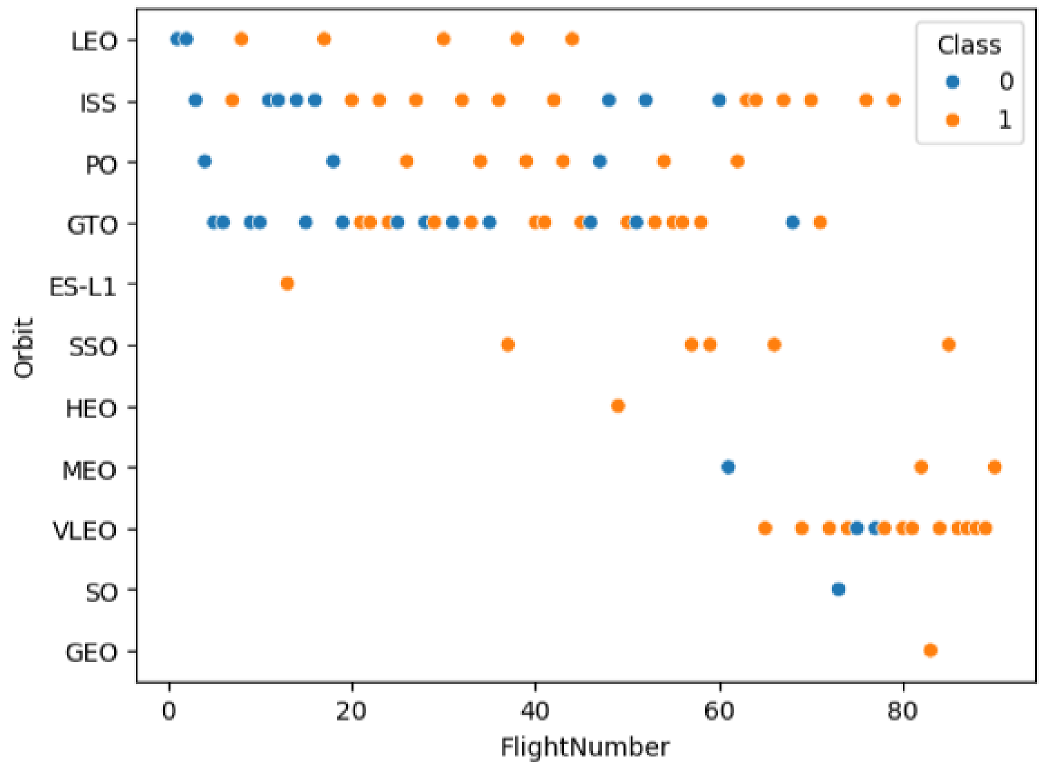
Launches at Orbits ES-L1, GEO, HEO and SSO are 100 % successful



Flight Number vs. Orbit Type

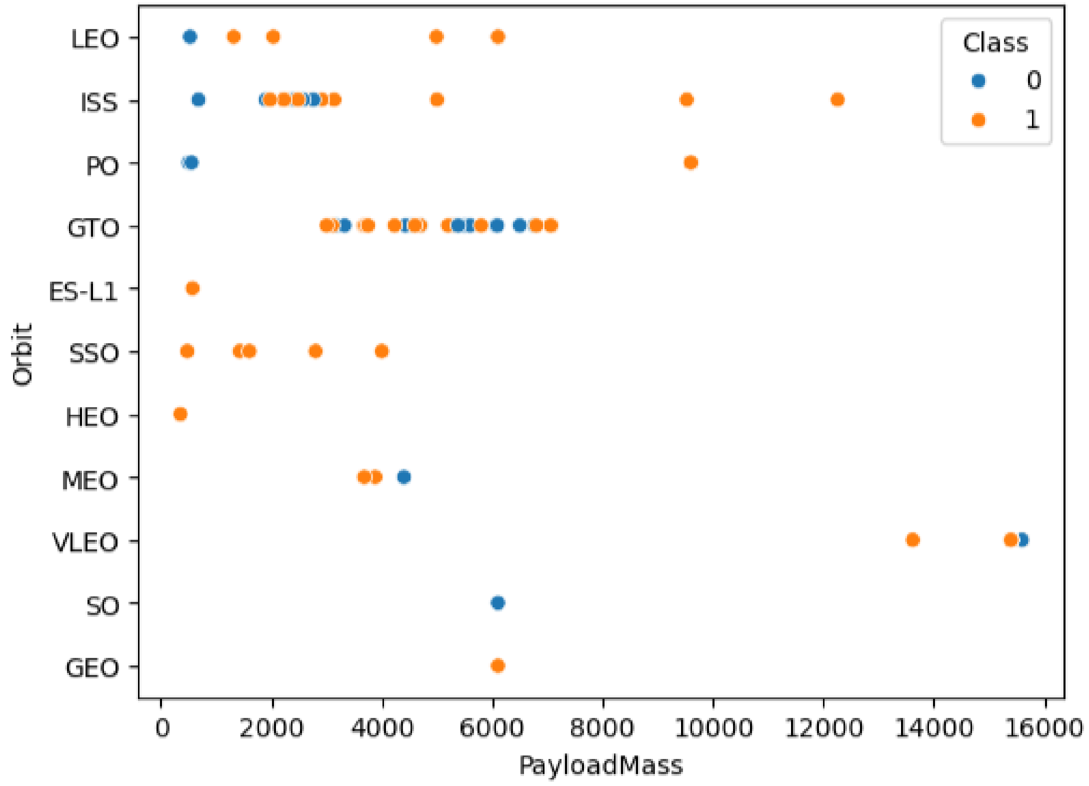
- This is a scatter plot of Flight number vs. Orbit type

- Higher flight numbers are more successful. For example, flight numbers > 80 are mostly successful across all orbits.



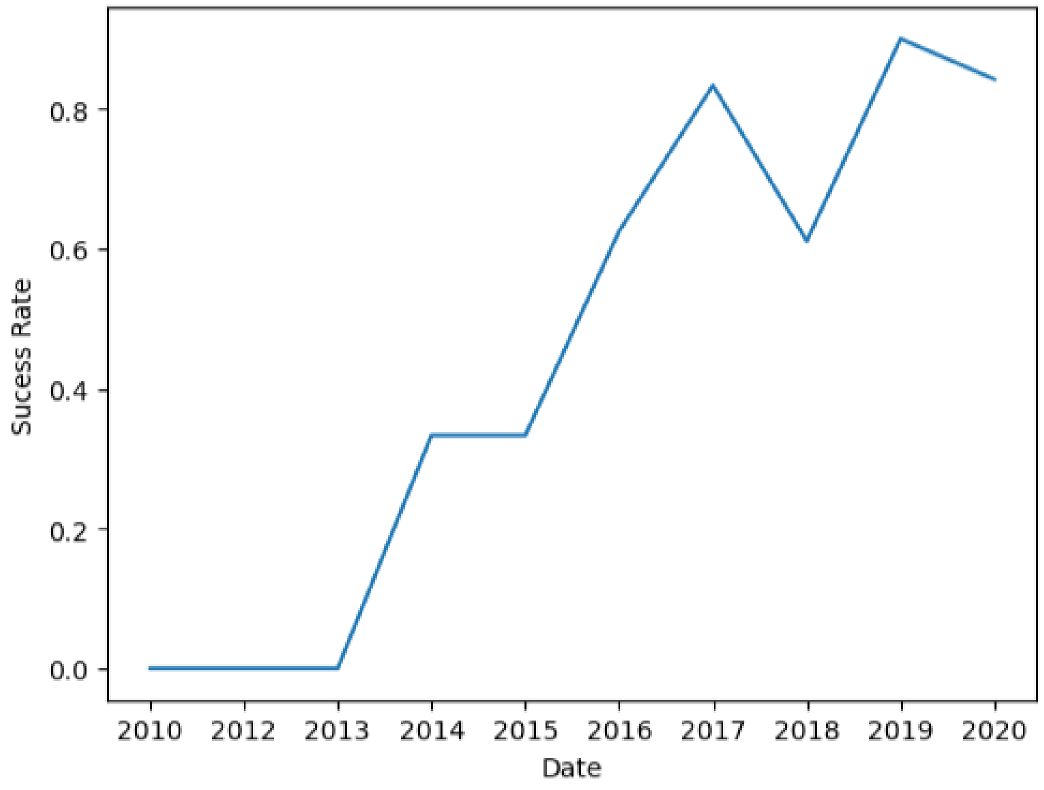
Payload vs. Orbit Type

- This is the scatter plot of payload vs. orbit type
- For Orbits ISS and PO, launches with payload greater than 10000 kg are successful.
- For Orbits ES-L1, SSO and HEO, launches with payloads less than 6000kg are successful



Launch Success Yearly Trend

- This is the line chart of yearly average success rate
- There is an approximate linear trend in the success rate.
- In 2018, there was a decline in the success rate, which increased again from 2019 onward.



All Launch Site Names

- I selected distinct Launch Sites from the table SPACEXTABLE.

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
```

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

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

Launch Site Names Begin with 'CCA'

- 5 records from SPACEXTABLE which belong to the launch sites beginning with the string 'CCA' are displayed.

```
%sql SELECT * FROM SPACEXTABLE 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

Total Payload Mass

- This query calculates total payload carried by boosters from NASA
- The query takes the sum of PAYLOAD_MASS_KG_ from table SPACEXTABLE for the records with NASA (CRS) customer.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
# %sql PRAGMA table_info(spacextable);
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- This query calculates the average payload mass carried by booster version F9 v1.1
- The query computes the average of PAYLOAD_MASS_KG_ for the records with Booster_Version 'F9 v1.1'

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- This query finds the dates of the first successful landing outcome on ground pad
- The query selects the record that has the earliest date among the records with landing outcome as 'Success (ground pad)'

```
%sql SELECT MIN(Date) From SPACEXTABLE WHERE Landing_Outcome ='Success (ground pad)'
```

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query lists the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- The query lists all Booster_Names that carried PAYLOAD_MASS_KG_ between 4000 and 6000 with landing_outcome as 'Success (Drone Ship)'

```
SELECT Booster_Version as Booster_Name FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000 AND Landing_Outcome ='Su
```

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

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

Total Number of Successful and Failure Mission Outcomes

- This query calculates the total number of successful and failure mission outcomes
- The query gets the record count of each Mission_Outcome.

```
%sql SELECT Mission_Outcome, Count(*) as total_outcomes FROM SPACEXTABLE GROUP BY Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

- The query lists the names of the booster which have carried the maximum payload mass

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

```
 * 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

2015 Launch Records

- This query lists the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- I get the Month, Landing_Outcome, Booster_Vesion and Launch_Site details for the records from the table when Landing_Outcome is 'Failure (drone ship)' and year of launch is 2015.

```
%sql SELECT substr(Date, 6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE Landing_Outcome = 'Failure (dro
```

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

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

- This query ranks 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
- The query groups by Landing Outcome and orders by the count of landing outcomes.

```
%sql select Landing_Outcome, count(*) as number  from SPACEXTABLE WHERE Date between  '2010-06-04' and '2017-03-20' group by Landing_Outcome
```

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

Landing_Outcome	number
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

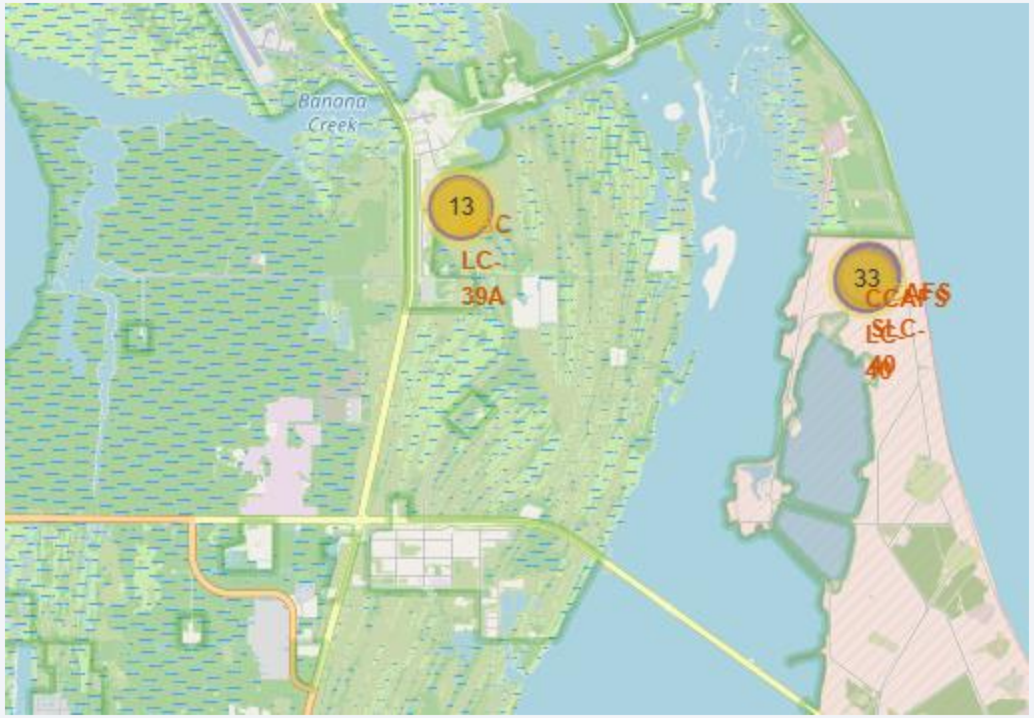
Launch Sites Proximities Analysis

Space X Launch Sites on a Map

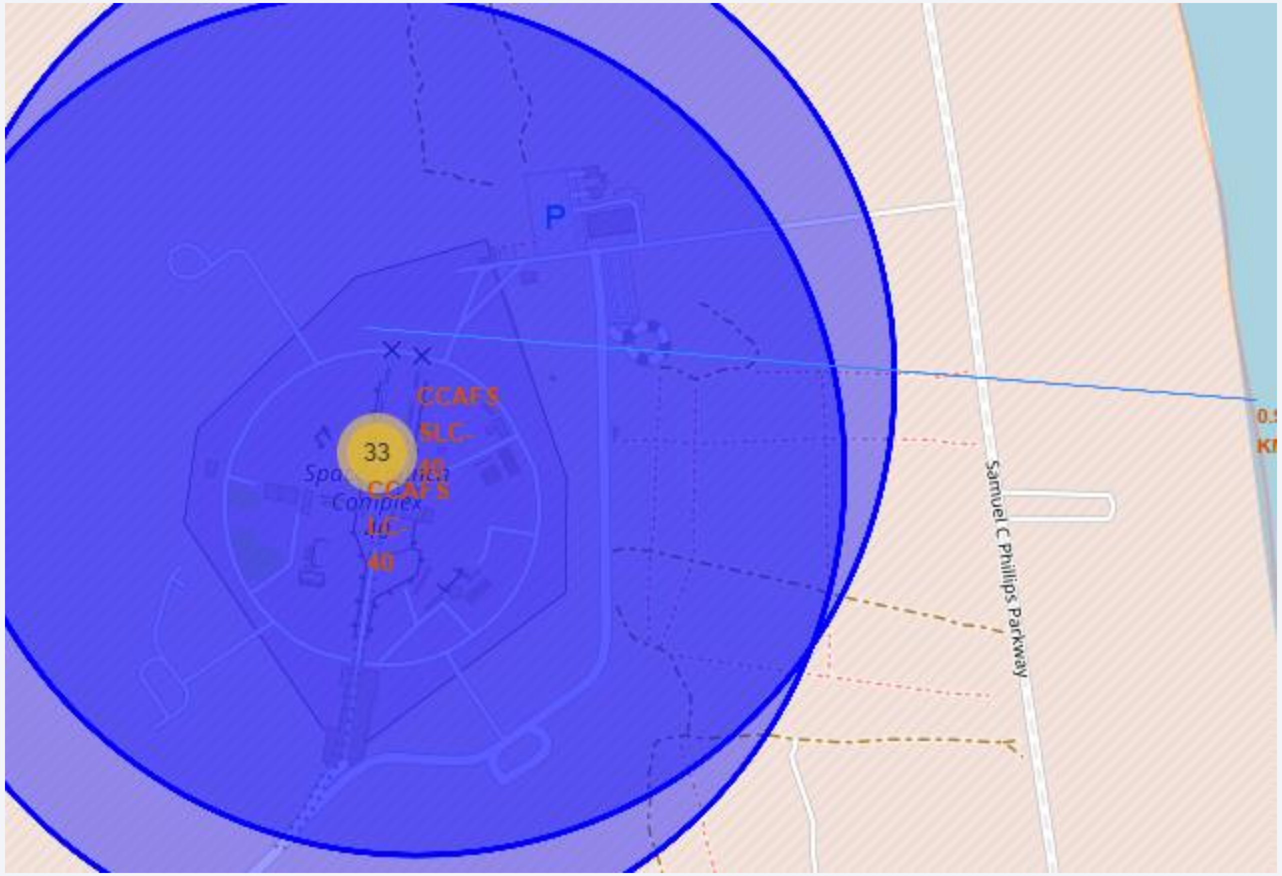
- The map shows locations of all launch sites.
- All launch sites are located are located around the coastal lines of the country.

Depicting Success/ Failure of a Launch at all sites

- These maps depict success and failure of a launch at each site as a cluster.
- The user can interactively explore about the number of successes and failure of each launch



Distance of Launch Sites from their proximities

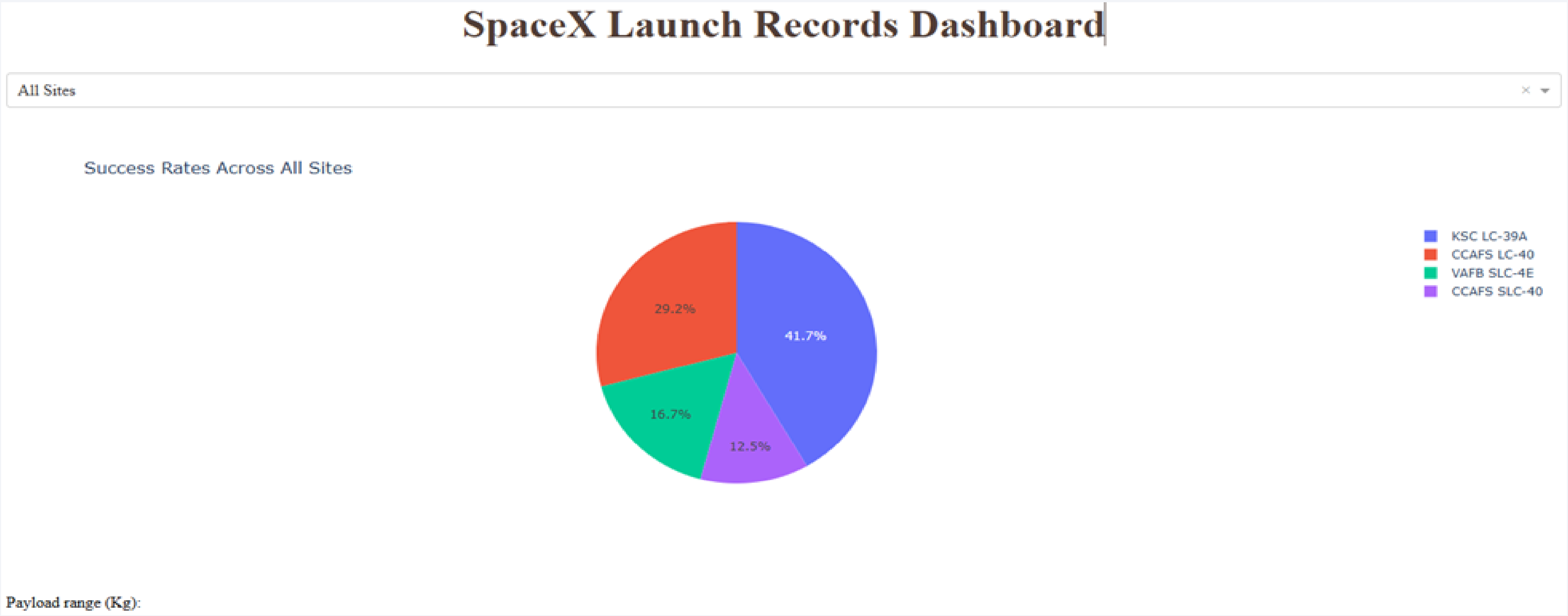


- As discussed before, all the launch sites are located near the coastlines.
- The locations displayed in the picture CCAFS launch sites, they are 0.9km far from the coast line.
- While the launch site VAFB SLC 4E is 14km far from the nearest city, the other launch sites are 78km far from the nearest city Orlando.
- There seems to be no correlation between distance from proximity and the success rate at a launch site.

Section 4

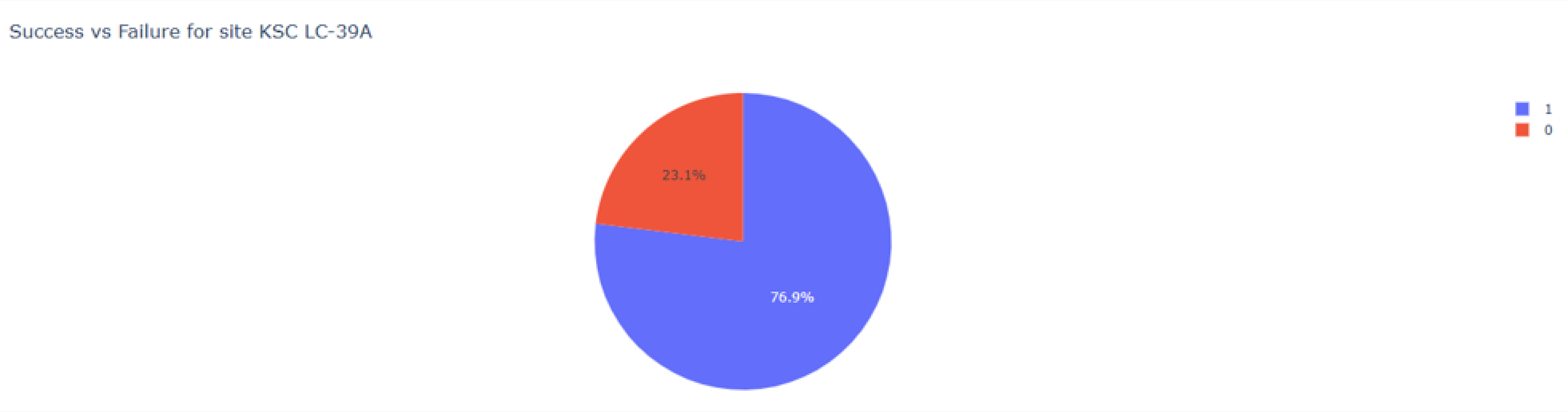
Build a Dashboard with Plotly Dash

Exploring Success Rates with Interactive DashBoard

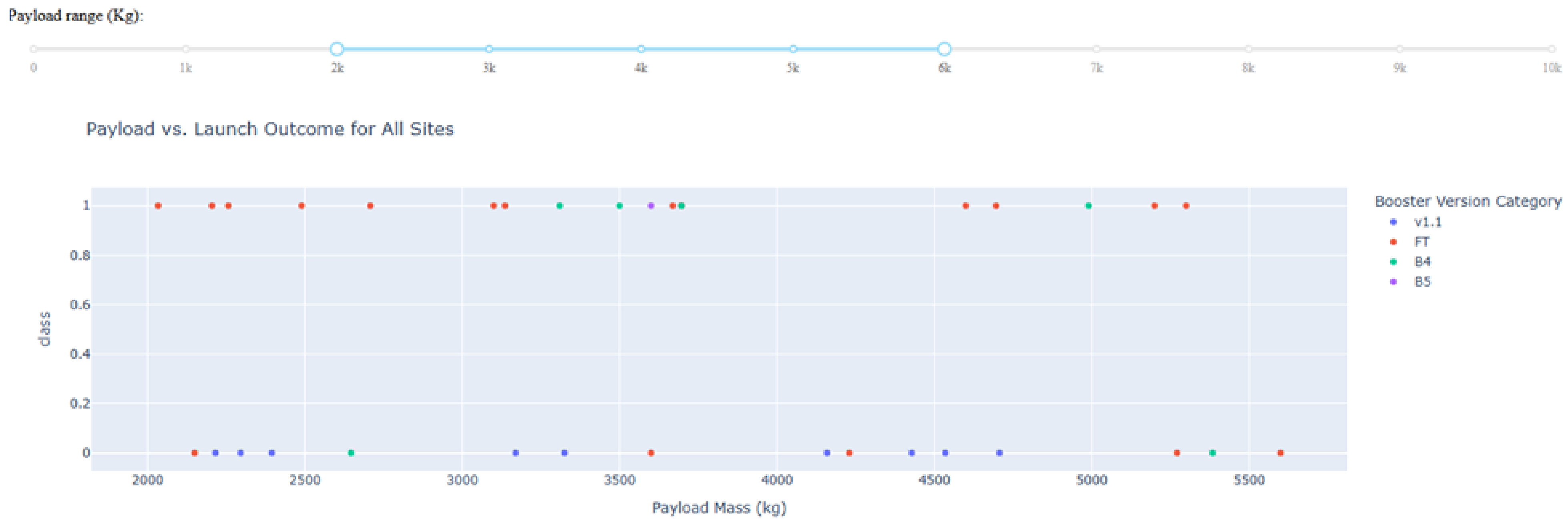


- The launch-site KSC-LC-39A has the highest percentage of successful launches.

Success Rate for KSC LC-39A



- The success rate of a launch at KSC LM-39A is highest at 76.9 percent. That is, 76.9 percent of launches at this site are successful.

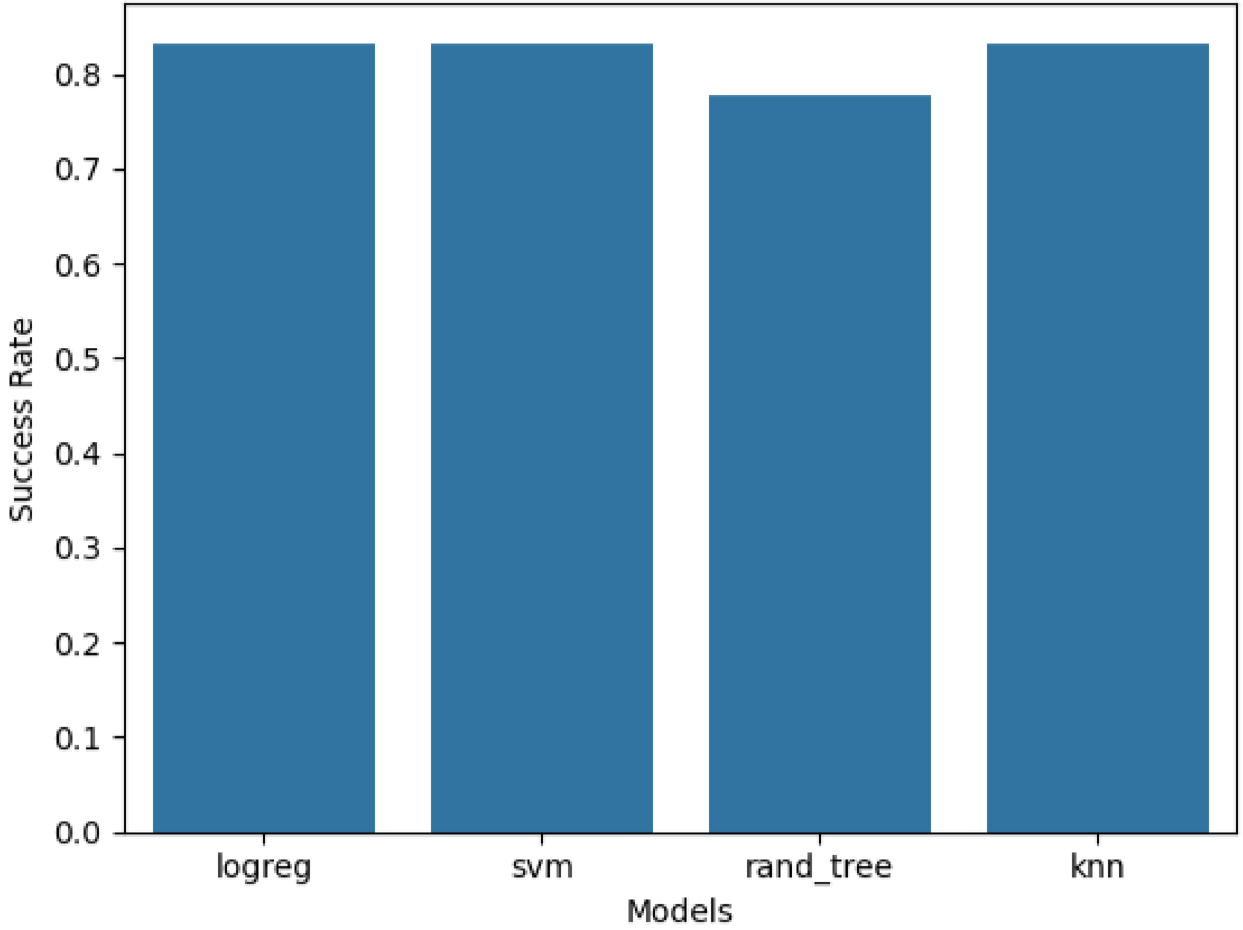


Payload Vs Launch Outcome

- As seen above, the success rate of launches is higher for payloads between 2k - 6k
- Booster version FT has the highest number of success

Section 5

Predictive Analysis (Classification)



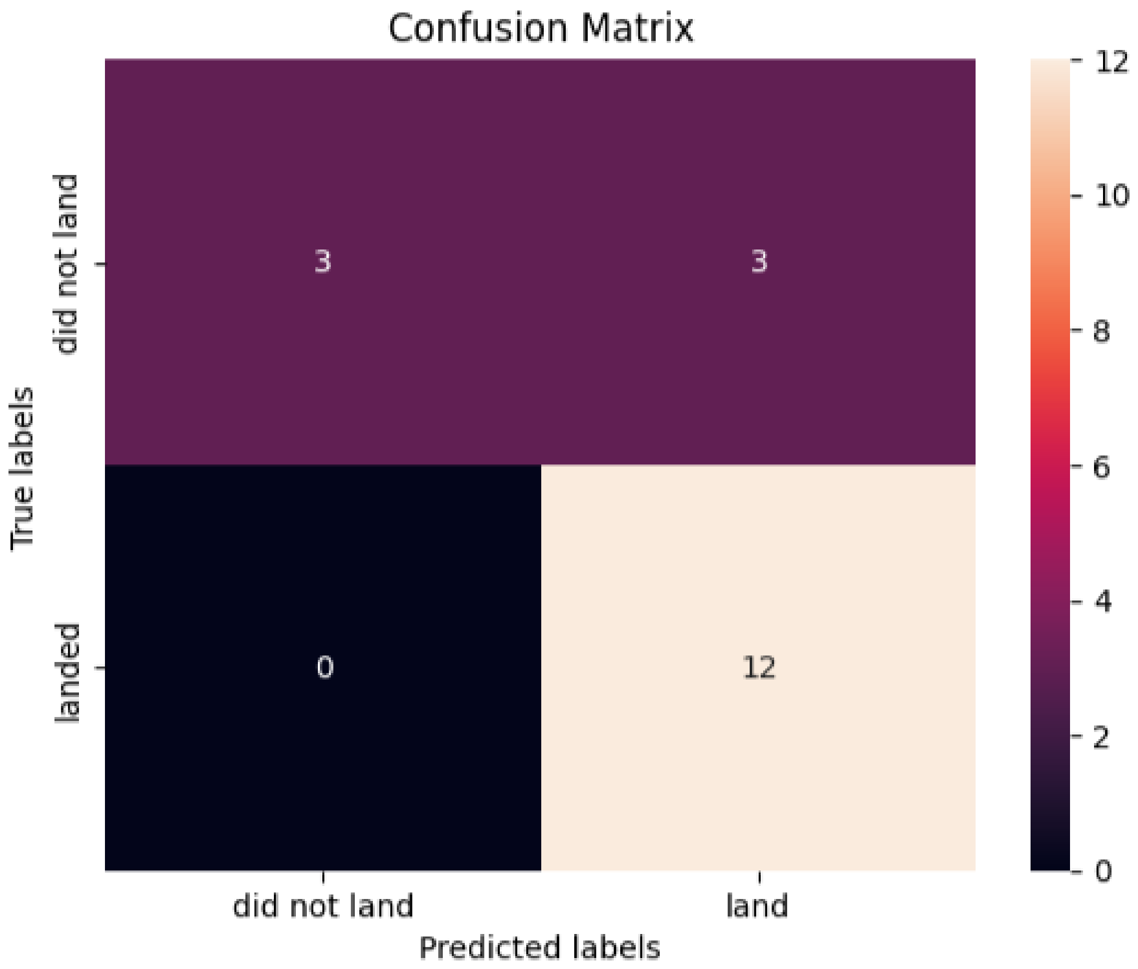
Classification Accuracy

- Four Models are fit to the data – Logistic Regression, Support Vector Machine, Random Decision Trees and K- nearest neighbors

- The accuracy rate of all the three models – Logistic Regression, Support Vector Machine and K-Nearest Neighbors is highest at 0.8333.

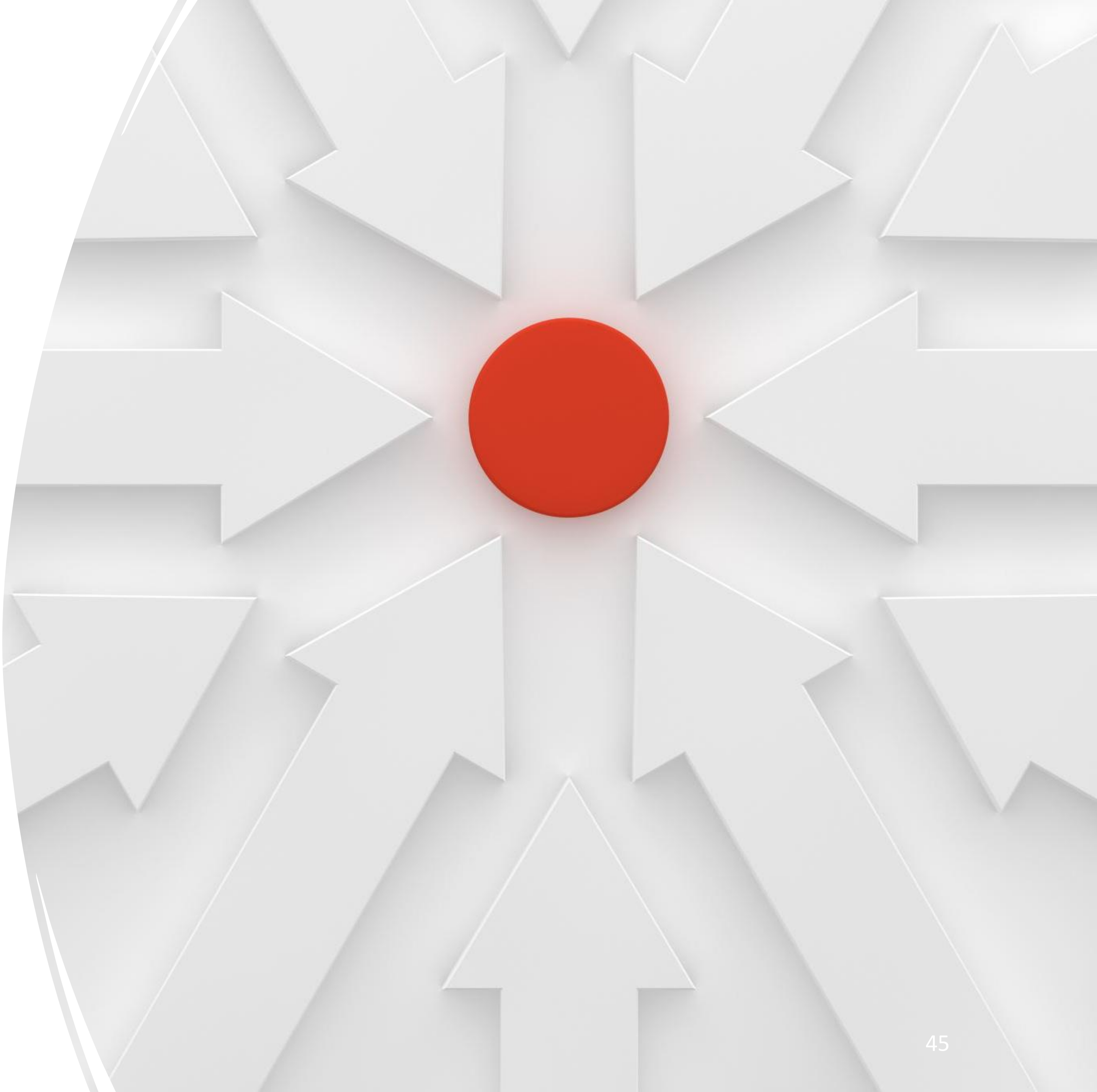
Confusion Matrix

This is the confusion matrix of K-Nearest Neighbors. As seen in the matrix, there are 3 False Positives. There are 3 True Negatives and 12 True Positives.



Conclusions

- The analysis shows that success rates are dependent on launch sites
- Booster versions influence the success rate
- Launch Success Is correlated with Pay load mass and Orbit
- We have fit and tuned a model that predicts the outcome with 83.33 percent accuracy.



Thank you!