APPLIED DATA SCIENCE CAPSTONE

COLLEN CRUZ

APRIL , 2023

OUTLINE

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

EXECUTIVE SUMMARY

Rocket launches can costs upwards of 165 million dollars while SpaceX is able to perform their launches at 62 million dollars, a fraction of the cost. The difference in price can be accredited to SpaceX's ability to reuse the first stage of their rockets.

The intention of this study is to utilize machine learning and readily available SpaceX launch data to analyze and determine the success of prior landing outcomes. After highlighting key values and their impacts on a successful landing, engineers can direct there focus in such areas to improve the success rate therefore keeping cost low by continuously recycling the first stage of each rocket.

INTRODUCTION

Project Background

A new company, SpaceY, has begun business in the rocket industry. There goal is to keep the costs of their rocket launches to a minimum and realized that SpaceX is able to advertise cost much lower than other competitors by recycling the first stage of their rockets.

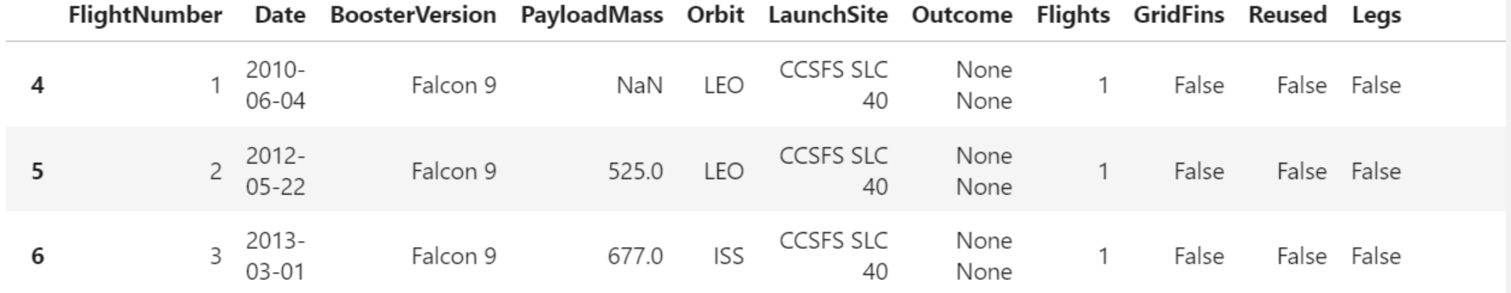
Project Goal

- Identify key indicators that make for a successful landing.
- Determine if a successful landing will occur using machine learning and publicilly provided SpaceX data.

METHODOLGY

Data Collection – SpaceX API

- One of the data sources collected for the use of this project is derived from the SpaceX API (https://api.spacexdata.com/v4/launches/past).
- In order to analyze the data, a request is sent to the SpaceX API and parsed into corresponding elements.
- The data is then filtered into a subset keeping only the features to be analyzed, the flight numbers, and the UTC dates. The subset is then further filtered to only include data pertaining to Falcon 9 launches which can be previewed as follows:



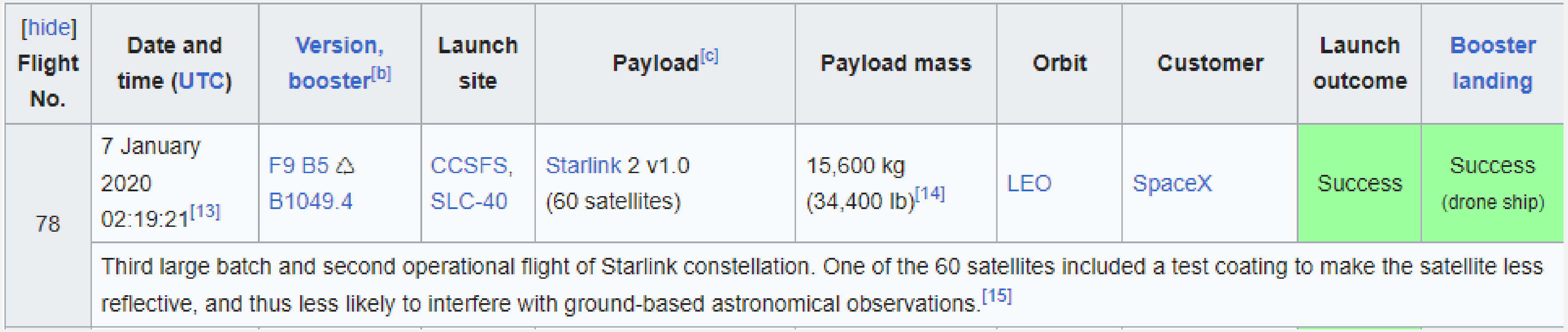
	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False

- The new Falcon 9 dataset contained 5 missing Payload Mass values which were addressed by replacing these values with the calculated mean of Falcon 9's Payload Mass.

Data Collection – Web Scraping

- The second source of data was collected via Web Scraping from the *List of Falcon 9 and Falcon Heavy launches* Wikipedia page (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

- The launch records are stored in an HTML table, the following displays one listing of the table:



[hide] Flight No.	Date and time (UTC)	Version, booster[b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
78	7 January 2020 02:19:21[13]	F9 B5 ♺ B1049.4	CCSFS, SLC-40	Starlink 2 v1.0 (60 satellites)	15,600 kg (34,400 lb)[14]	LEO	SpaceX	Success	Success (drone ship)
	Third large batch and second operational flight of Starlink constellation. One of the 60 satellites included a test coating to make the satellite less reflective, and thus less likely to interfere with ground-based astronomical observations.[15]								

- A request is sent to the URL, all variables are extracted from the columns of the table, and a data frame is created by parsing in the launch HTML table information into their respective columns.

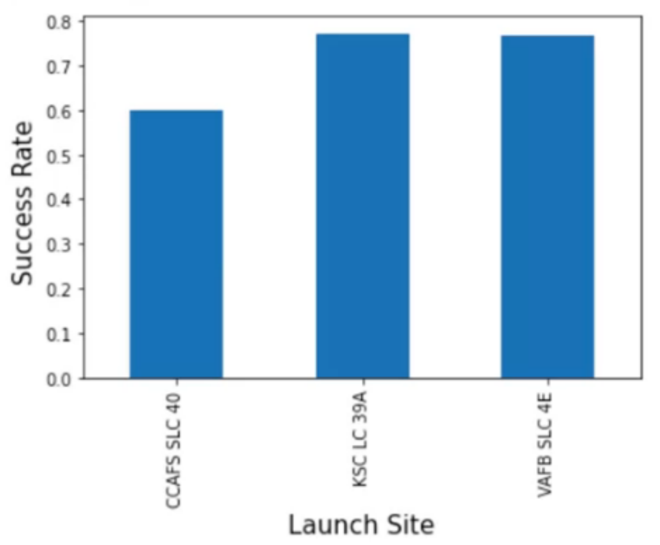
Data Wrangling

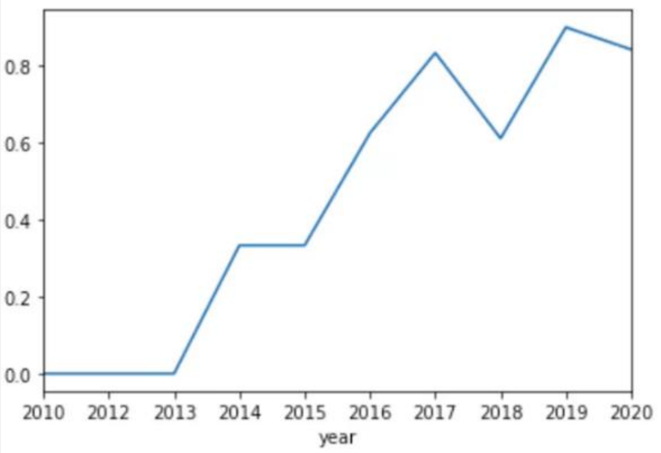
- Initial Exploratory Data Analysis (EDA) is performed in order to identify patterns in the data and determine what will be the label for training supervised models.
- Initial EDA was focused on the launch sites and the number of launches conducted at each site.
- Each launch is set to a dedicated orbit, therefore, the number of occurrences in each respective orbit is then accounted for per launch.
- A list of landing outcomes is created and an additional sublist is then derived keeping only unsuccessful landings. A 'landing_class' variable is initialed and assigned to if the landing outcome is successful or unsuccessful.
- A 'Class' variable is created within the dataset to hold the landing_class outcome, which allows for the calculation of the success rate of the landing outcome.

EXPLORATORY DATA ANALYSIS

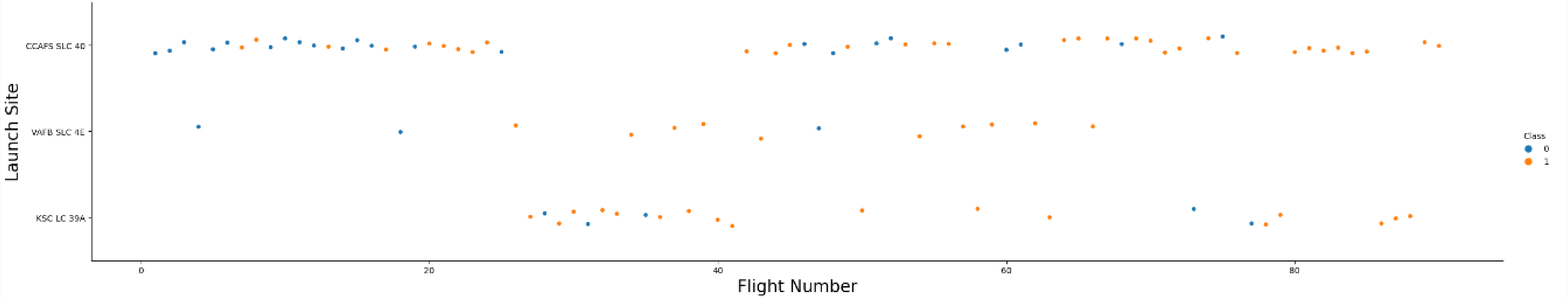
EDA & Interactive Visual Analytics Methodology

- Exploratory Data Analysis is conducted on the SpaceX dataset to further understand the information.
- Feature Engineering is performed to derive which attributes can be used to determine if the first stage of the launch can be reused.
- Select features can be singled to and graphed versus their respective success rate such as the example shown comparing the success rate between different launch sites.
- In addition, visual analytics allows to show patterns within features such as the line chart displaying an increase in success rate pass the year 2013.



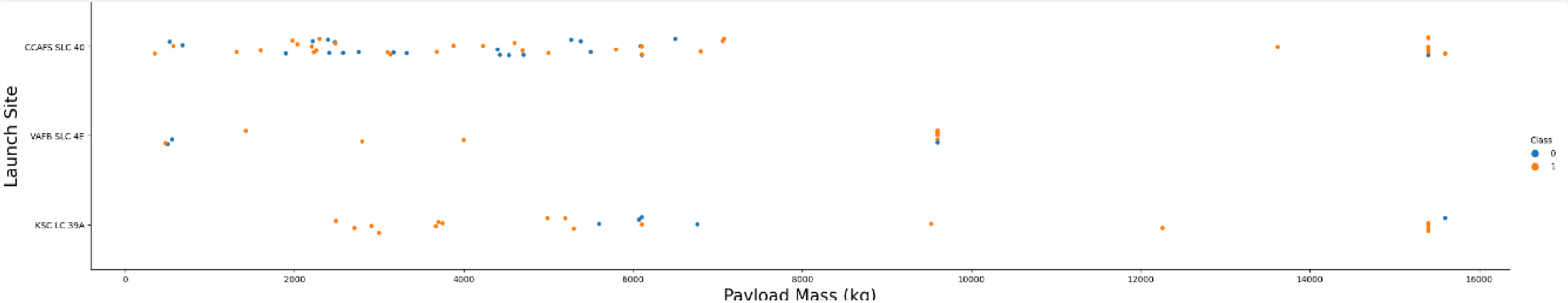


Flight Number vs. Launch Site



- A slight relationship can be noticed that the lower the Flight Number yielded unsuccessful landing results and the higher the Flight Number, the likelihood of a successful landing increases.
- Flight numbers between 0 and 20 have a much lower success rate which greatly improves afterwards

Payload vs. Launch Site

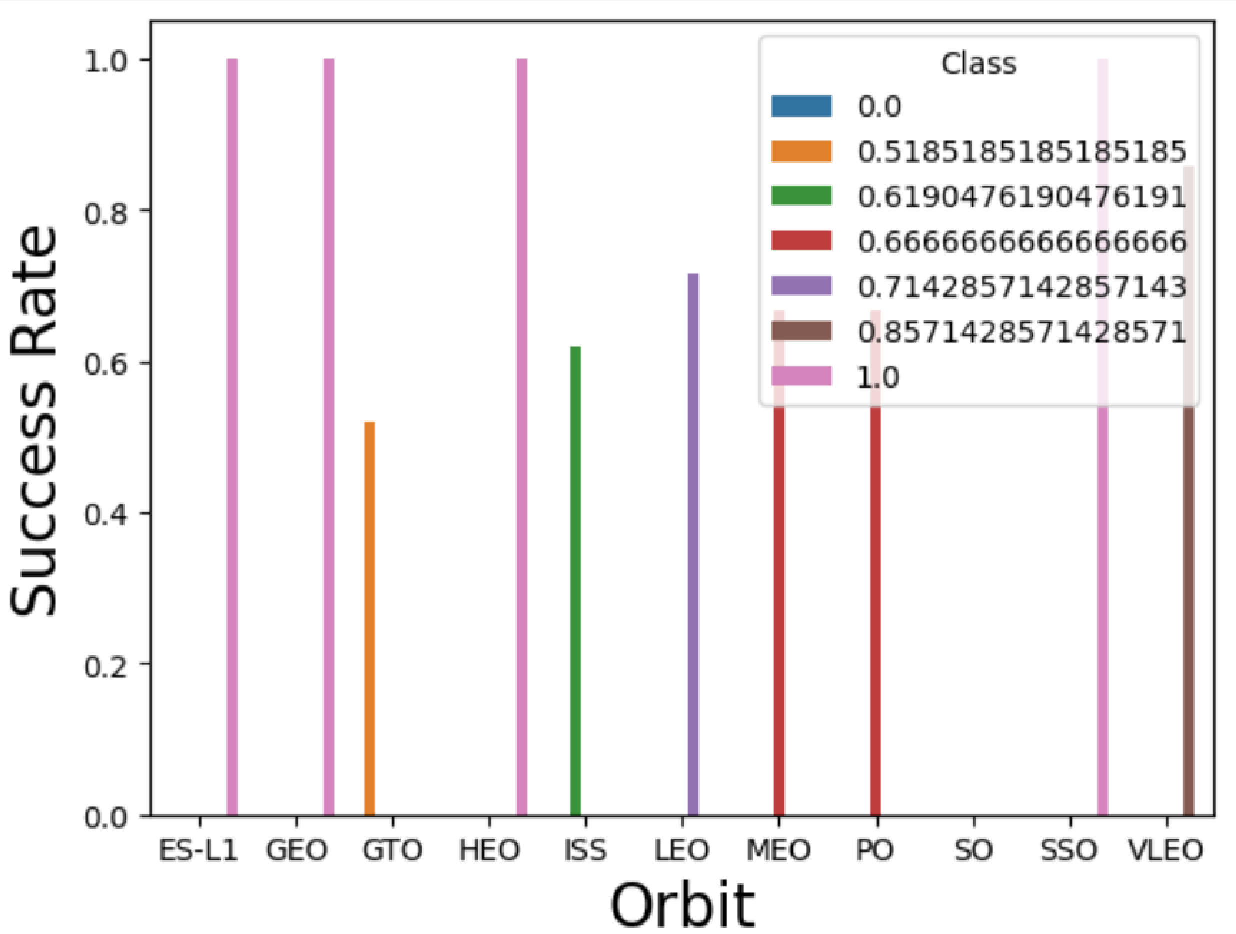


- In the case of Launch Site CCAFS SLC 40, the success rate diminishes between 2000 kg and 5000 kg. This may lead to the conclusion that an increase in payload may lead to unsuccessful results however, success rate greatly improves after 13000 kg.

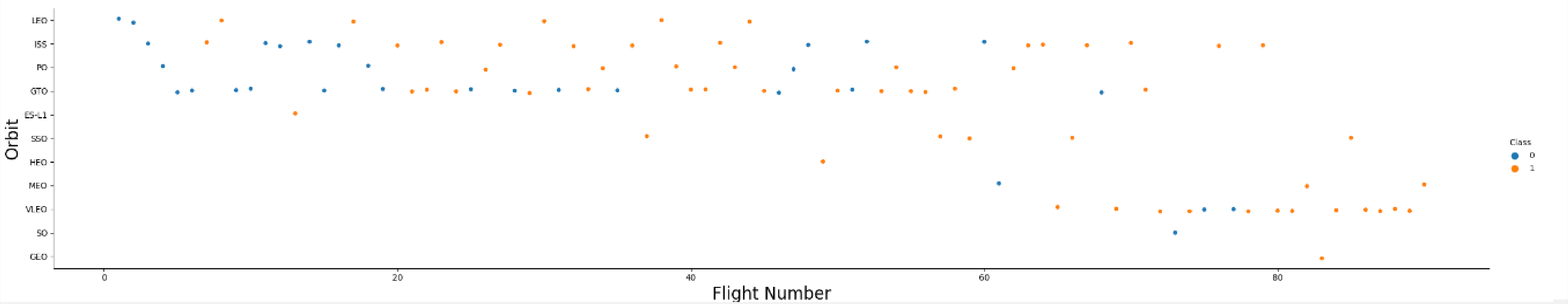
- In opposition to the pattern viewed for CCAFS SLC 40 and VAFB SLC 4E, launch site KSC LC 39A finds success in the earlier launches but if met with unsuccessful results around the 5000 kg mark.

Success Rate of each Orbit Type

- The following orbit types were able to achieve a 100% success rate: ES-L1, GEO, HEO, and SSO
- A pattern can be recognized between the orbits GTO and PO, resulting in a percentile fitting between ~0.5 and ~0.71 with HEO being an outlier.
- The success rate is seen to improve on the outer bounds among the orbit types.



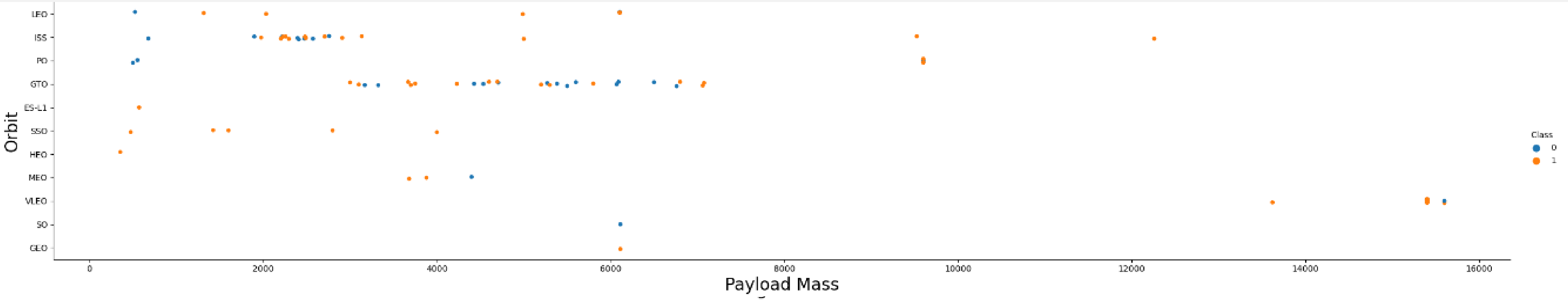
Flight Number vs. Orbit Type



- A cluster of unsuccessful landing points is found at the earlier stages of Flight Numbers with improving success rate as Flight Number increases. Orbit GTO seems to be an outlier, as only little improvement can be seen with the increase of Flight Numbers.

- In pattern can be seen that earlier Flight Numbers prioritized orbits LFO, ISS, PO, and GTO with Flight Numbers pass 60 focusing on orbit VLEO.

Payload vs. Orbit Type

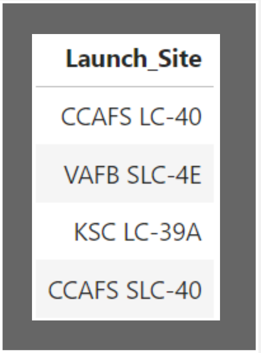


- Excluding orbit GTO, a pattern among the other orbits can be seen after the 2000 kg point where an increase in successful landings occurs with a few slight occurrences of unsuccessful missions for orbit ISS between the payload mass of 2000 to 3000 kg.
- A pattern between success rate vs. payload mass for orbit GTO cannot be recognized as there seems to be an equal amount of successful and unsuccessful points randomly plotted.

EDA with SQL

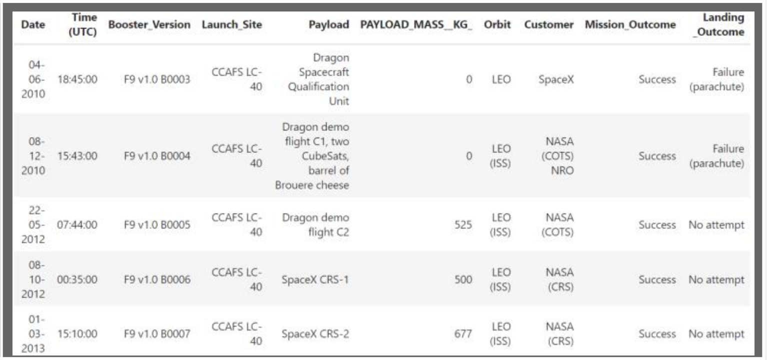
The dataset was loading into a Db2 database and the following task were performed using SQL queries:

1. Display the names of the unique launch sites in the space mission



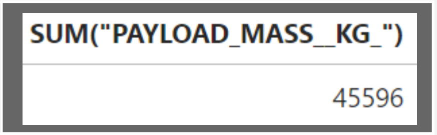
Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

2. Display 5 records where launch sites begin with the string 'CCA'



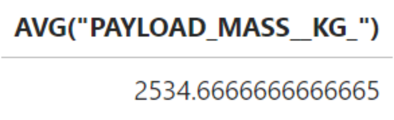
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

3. Display the total payload mass carried by boosters launched by NASA (CRS)



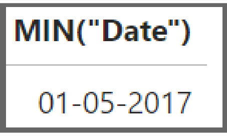
SUM("PAYLOAD_MASS__KG_")
45596

4. Display average payload mass carried by booster version F9 v1.1



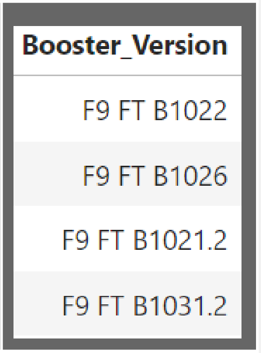
AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

5. Display the total payload mass carried by boosters launched by NASA (CRS)



MIN("Date")
01-05-2017

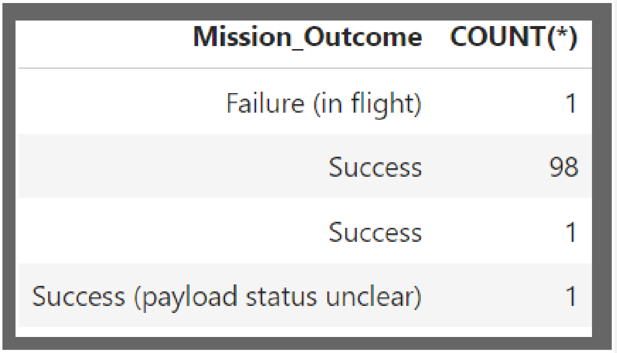
6. Display the names of the unique launch sites in the space mission



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

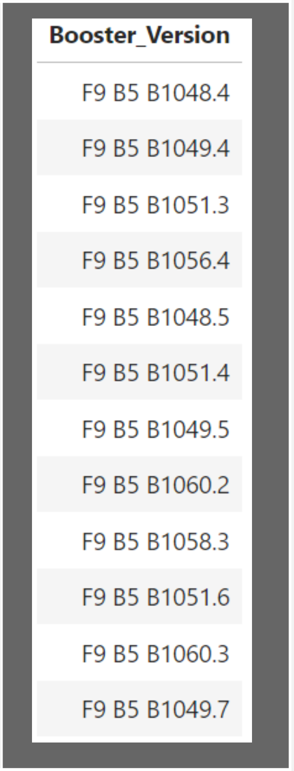
EDA with SQL (Cont…)

7. List the total number of successful and failure mission outcomes



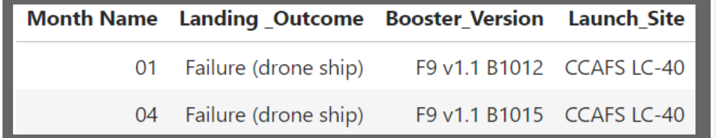
Mission_Outcome	COUNT(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

8. List the names of the booster versions which have carried the maximum payload mass.



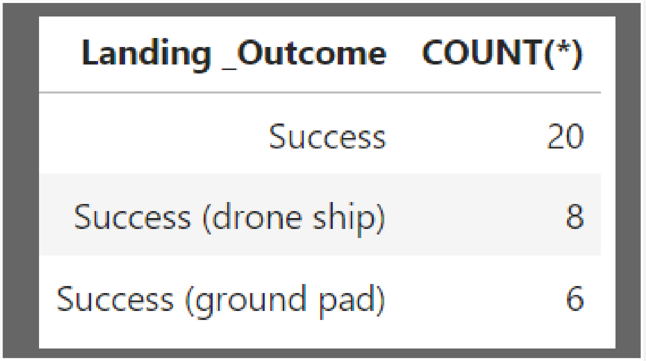
Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

9. List the records which will display the month names, failure landing outcomes in drone ships, booster versions, launch site for the months in year 2015



Month Name	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

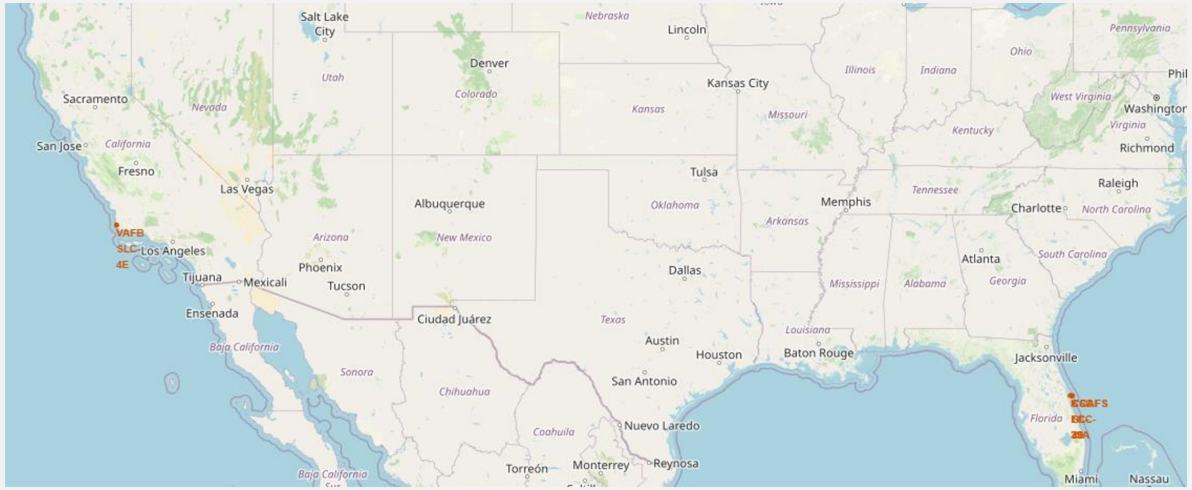
10. Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order

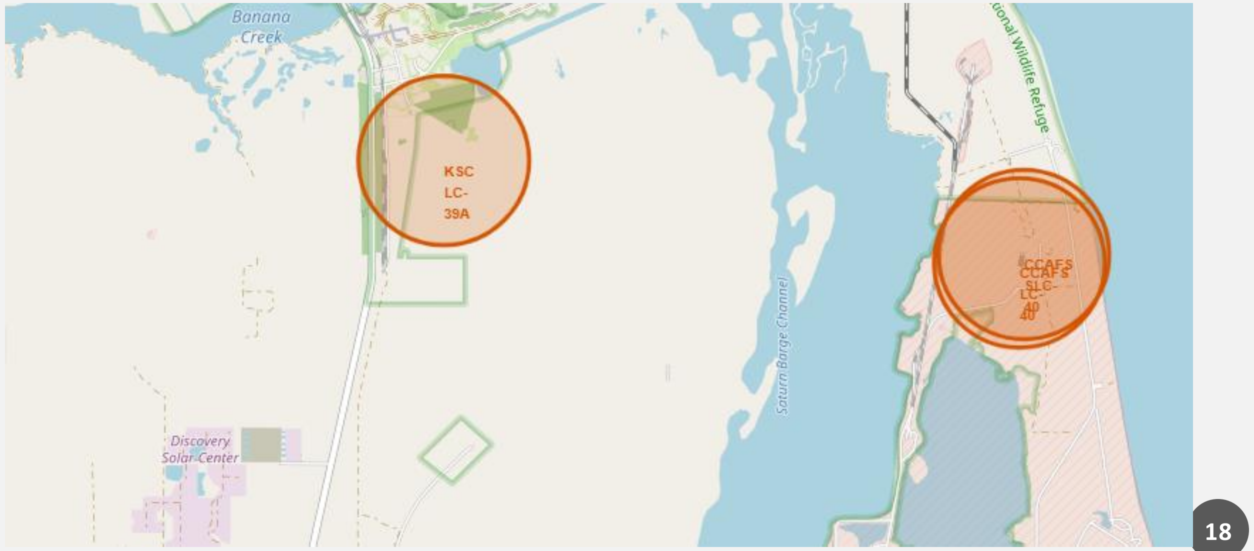


Landing _Outcome	COUNT(*)
Success	20
Success (drone ship)	8
Success (ground pad)	6

Interactive Visual Analytics - Folium

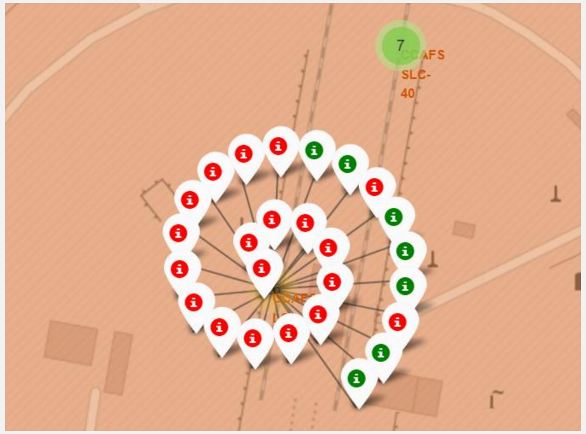
- In addition to analyzing physical attributes of the rocket, environmental attributes were also accounted for such as location and proximities of the launch site.
- The process begins by first gathering the locations of the various launch sites, then plotting each site on an interactive map to visual examine each location and the surrounding environment
- The interactive map allows for in-depth examination, displaying nearby location, railways, bodies of water, etc.

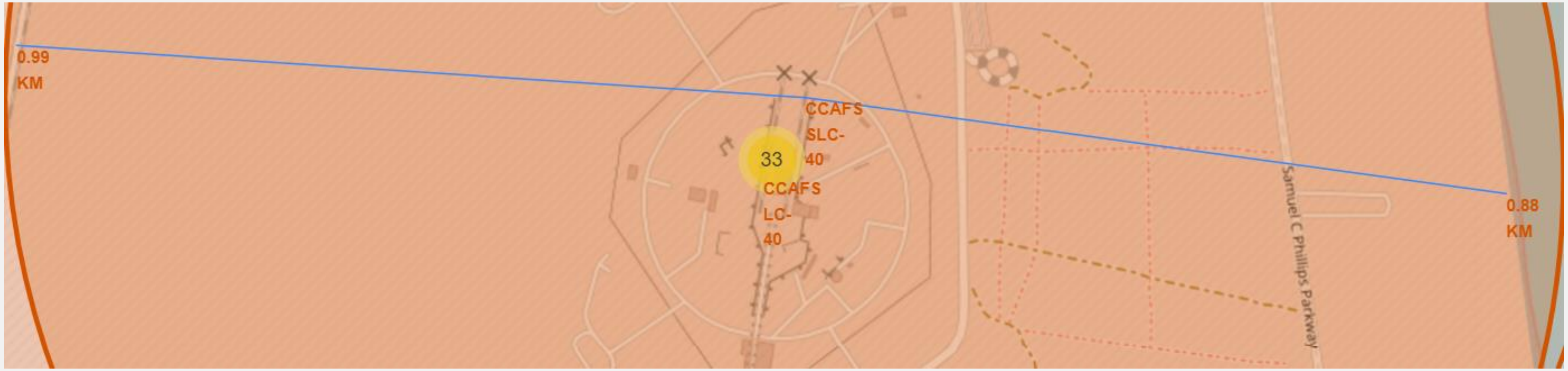




Interactive Visual Analytics – Folium (Cont…)

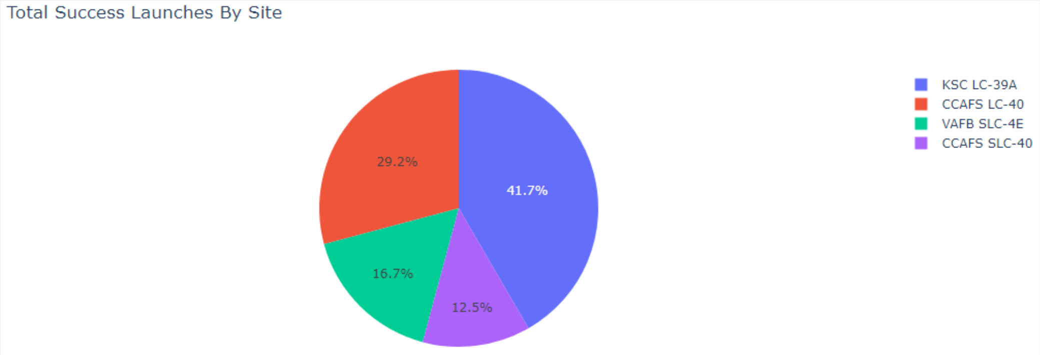
- The details of each launch site were further expanded by markers which highlights both the successful and unsuccessful launches of each site. An initial number is listed as the number of launches in total, which is expanded into red (unsuccessful) and green (successful) markers when the user interacts with the cluster.
- The variable of proximities was then accounted for by marking and calculating distances to land marks within vicinity of the launch sites.

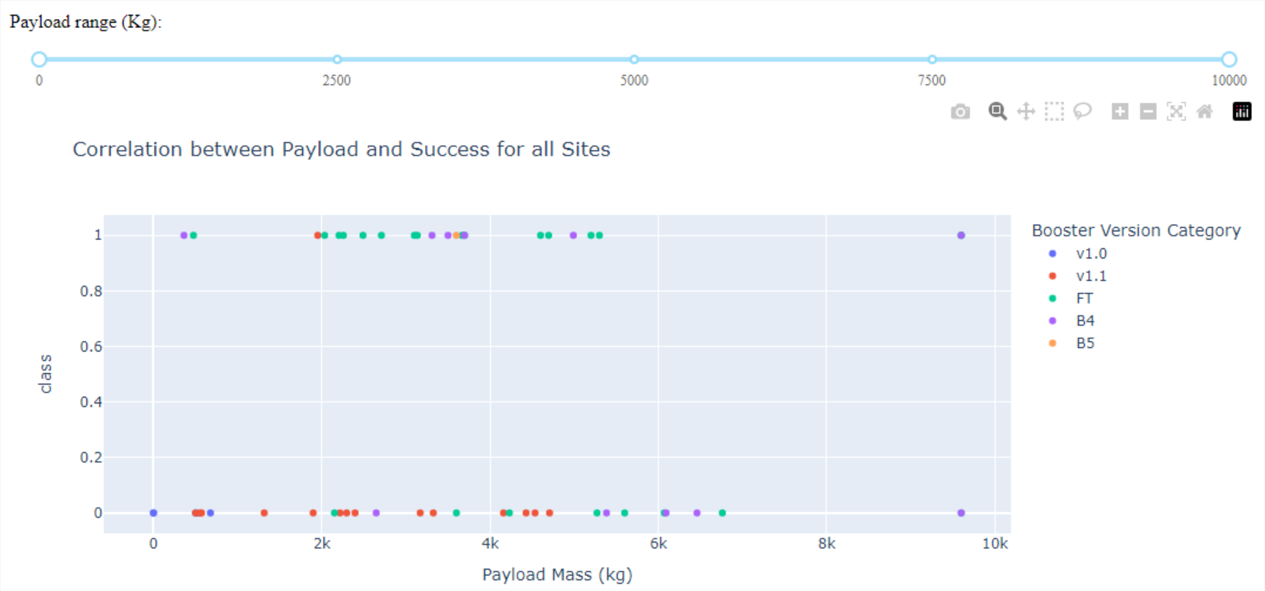




Interactive Visual Analytics - Plotly

- Statistical comparison was made utilizing an interactive dashboard application. One such comparison calculated is a pie chart display each sites respective success rate.


Total Success Launches By Site

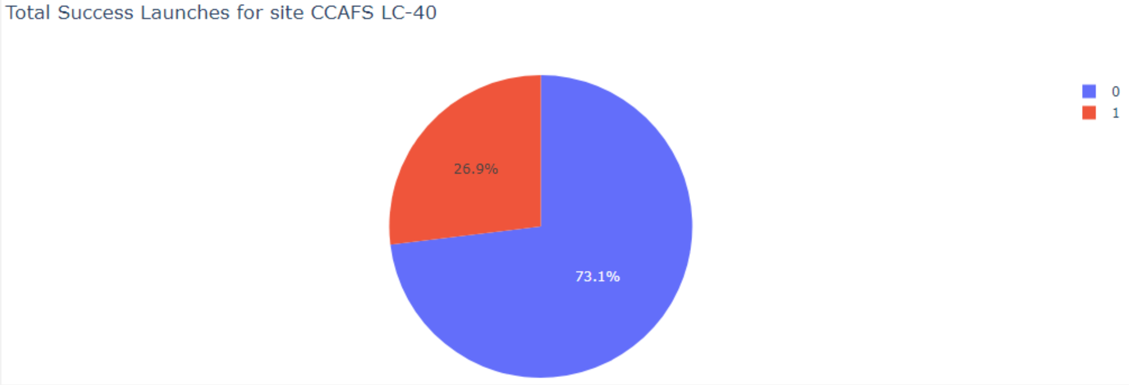

Payload range (Kg):

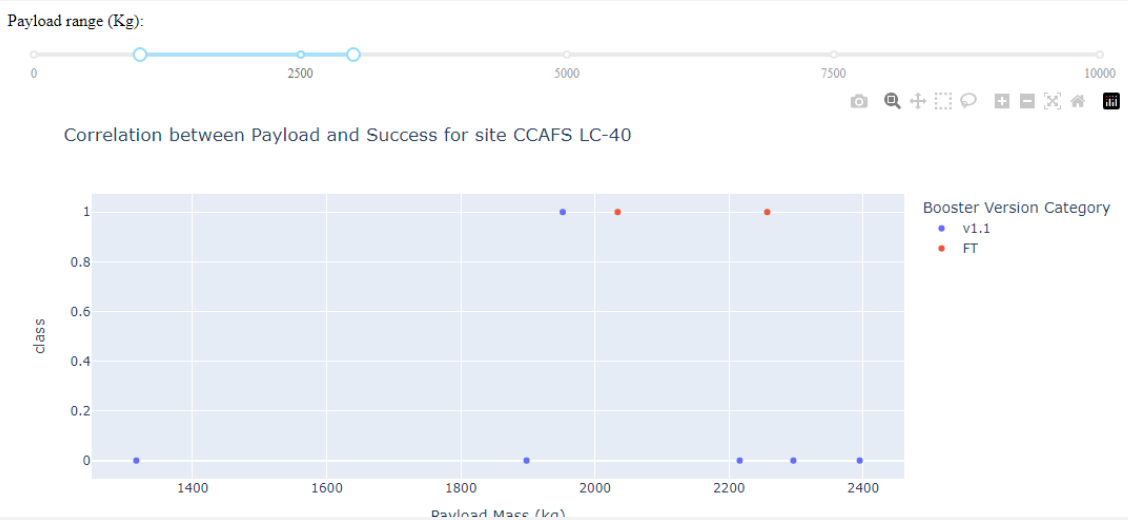
Correlation between Payload and Success for all Sites

- The correlation between Payload Mass and Success rate was plotted and graphed with additional color coded points detailing which particular booster version was involved in each launch.

Interactive Visual Analytics – Plotly (Cont…)

- In depth statistical data was also available by selecting a specific site to view. A payload range slider was also a feature included which allowed for specific ranges to be plotted.
- The pie chart and graph displayed are focused on site CCAFS LC-40. Blue indicates the percentage of unsuccessful launches and red indicates successful launches. The graph displayed is further manipulated to display results within the payload range of 1000 to 3000 kg.

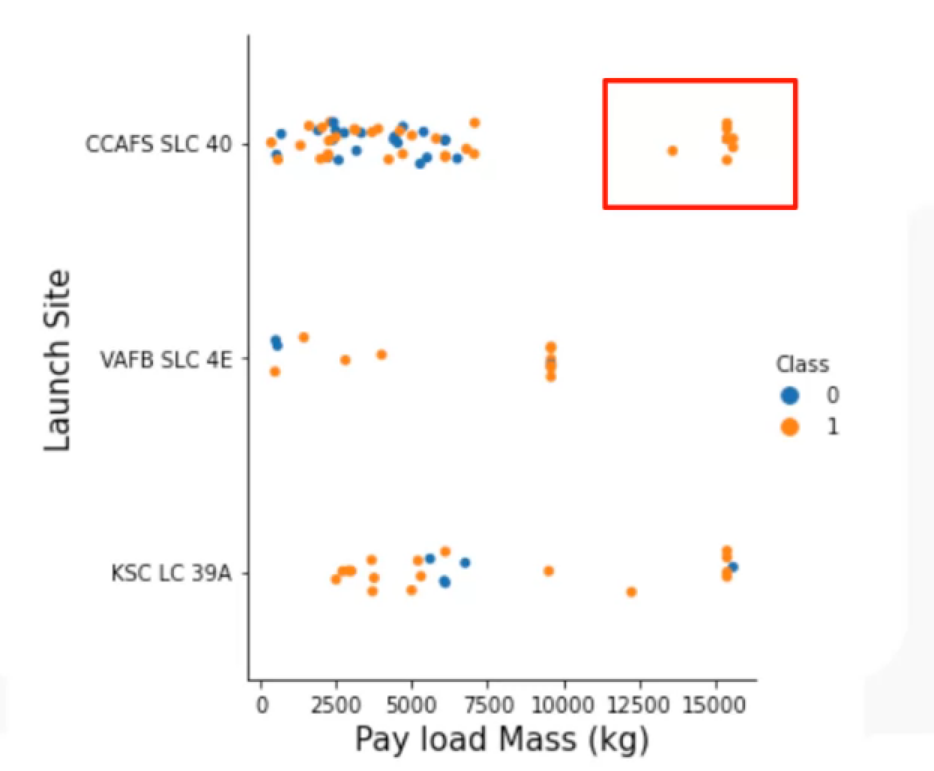




PREDICTIVE ANALYSIS

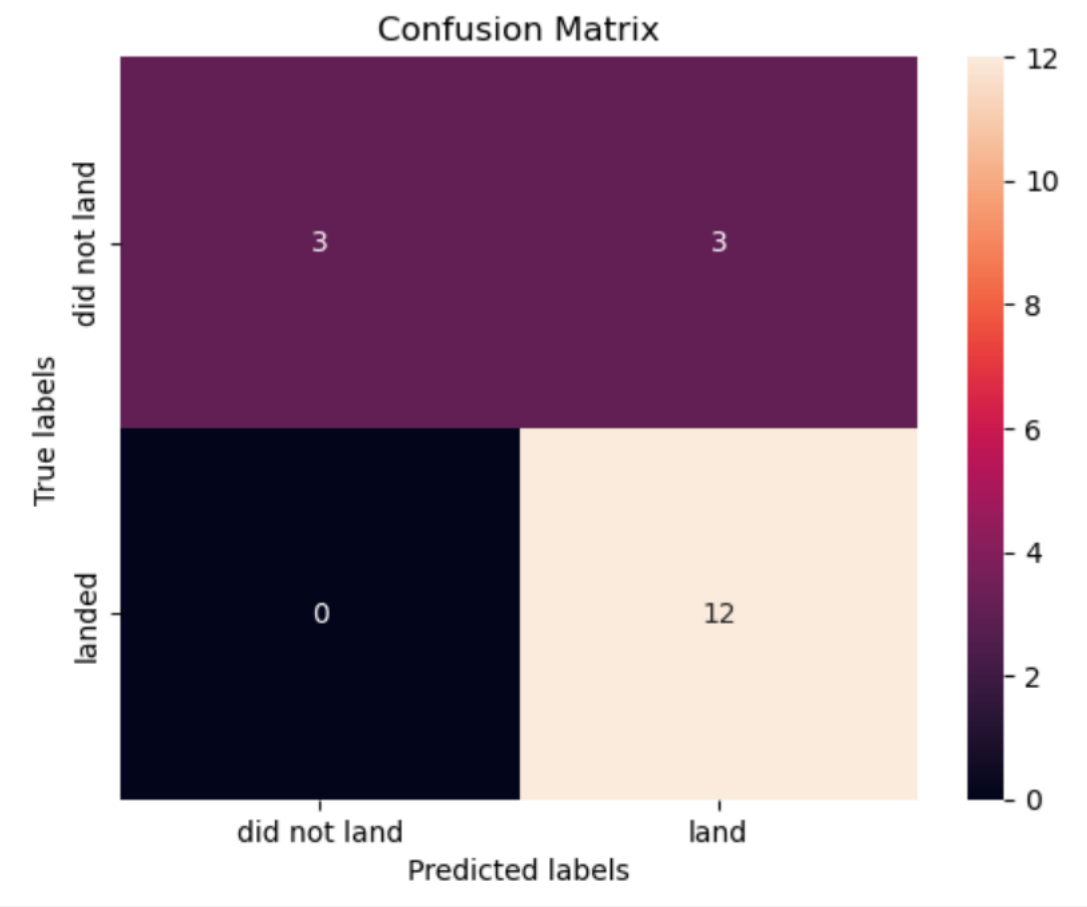
Predictive Analysis Methodology

- Utilizing machine learning and the selected features, a prediction can be made whether the first stage will land successfully.
- An advantage of Feature Engineering is the ability to combine multiple features and compare their effect to the landing success rate. An example of such prediction is shown, accounting for two attribute (Launch Site and Payload Mass), and whether the landing was successful (1) or unsuccessful (0).
- This project will utilize the following machine learning functions: Logistic Regression, Support Vector Machine (SVM), Decision Tree Classifier, and K-Nearest Neighbor



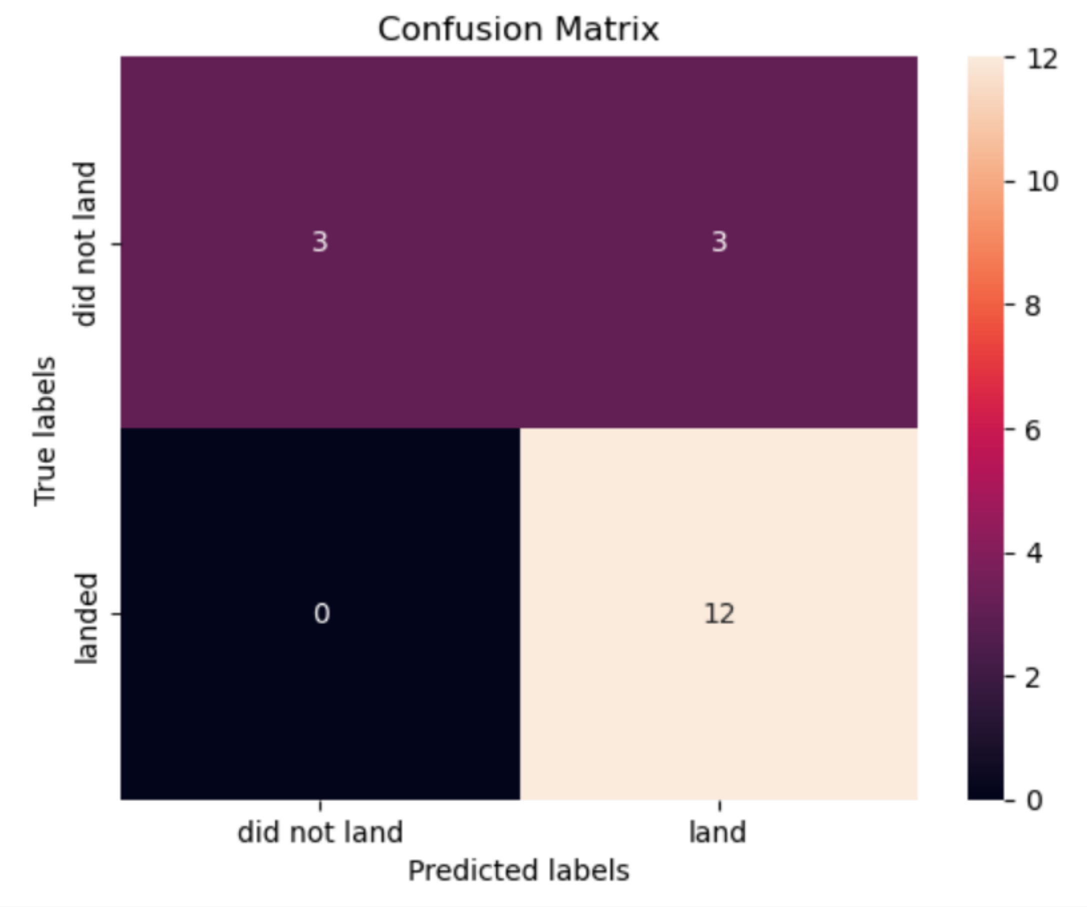
Predictive Analysis Results – Logistic Regression

- Using the score method, the calculated accuracy of Logistic Regression was 0.833333
- The following displays the confusion matrix from the trained Logistic Regression model. A problem of false positives can be noticed when viewing the results of the confusion matrix.



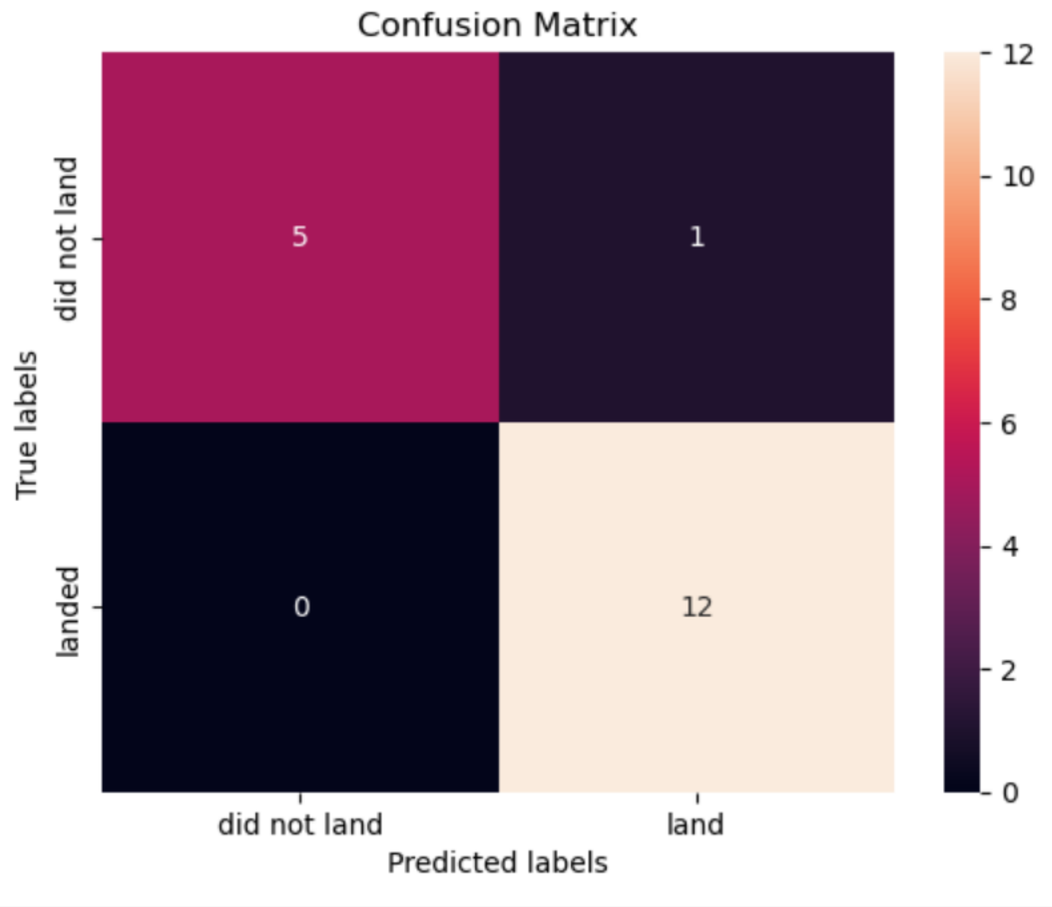
Predictive Analysis Results – SVM

- Using the score method, the calculated accuracy of SVM was 0.848214
- The following displays the confusion matrix from the trained SVM model. The confusion matrix displayed yielded the same results when using Logistic Regression, displaying the issue of false positive results.



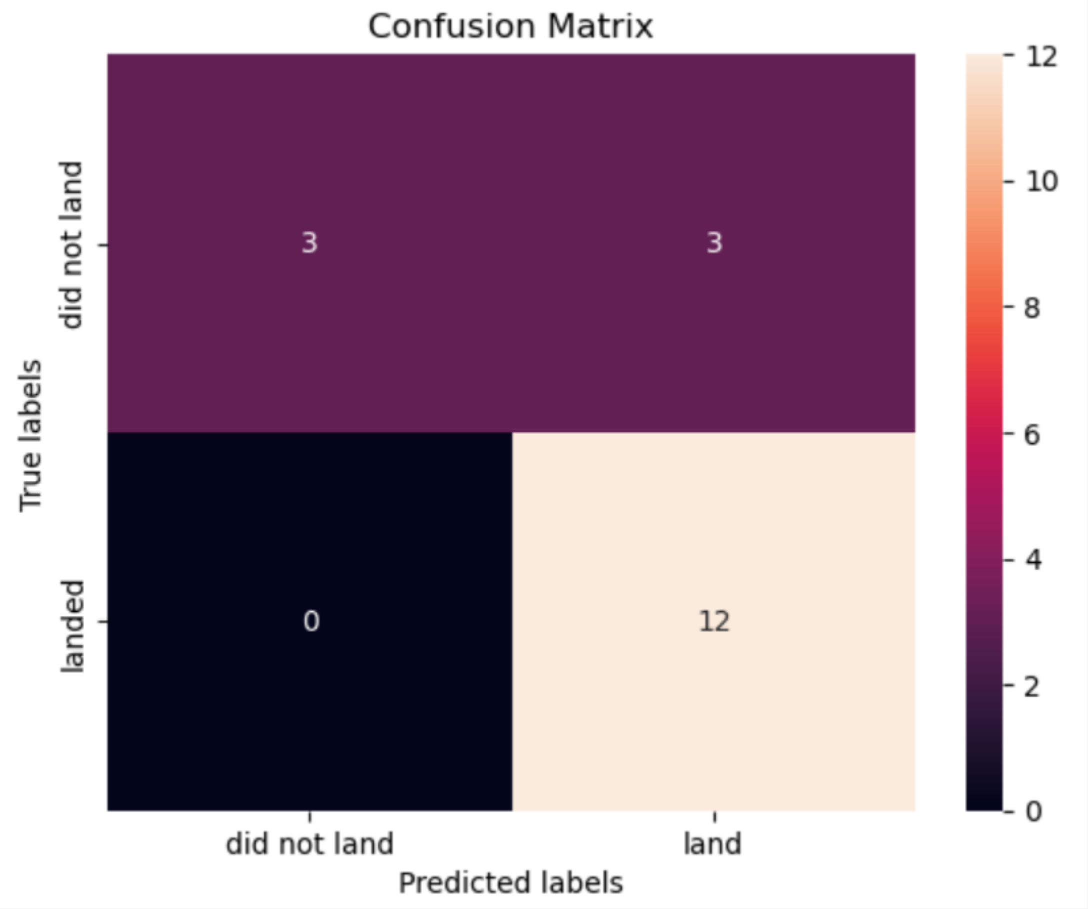
Predictive Analysis Results – Decision Tree Classifier

- Using the score method, the calculated accuracy of Decision Tree Classifier was 0.944444
- The following displays the confusion matrix from the trained Decision Tree Classifier model.



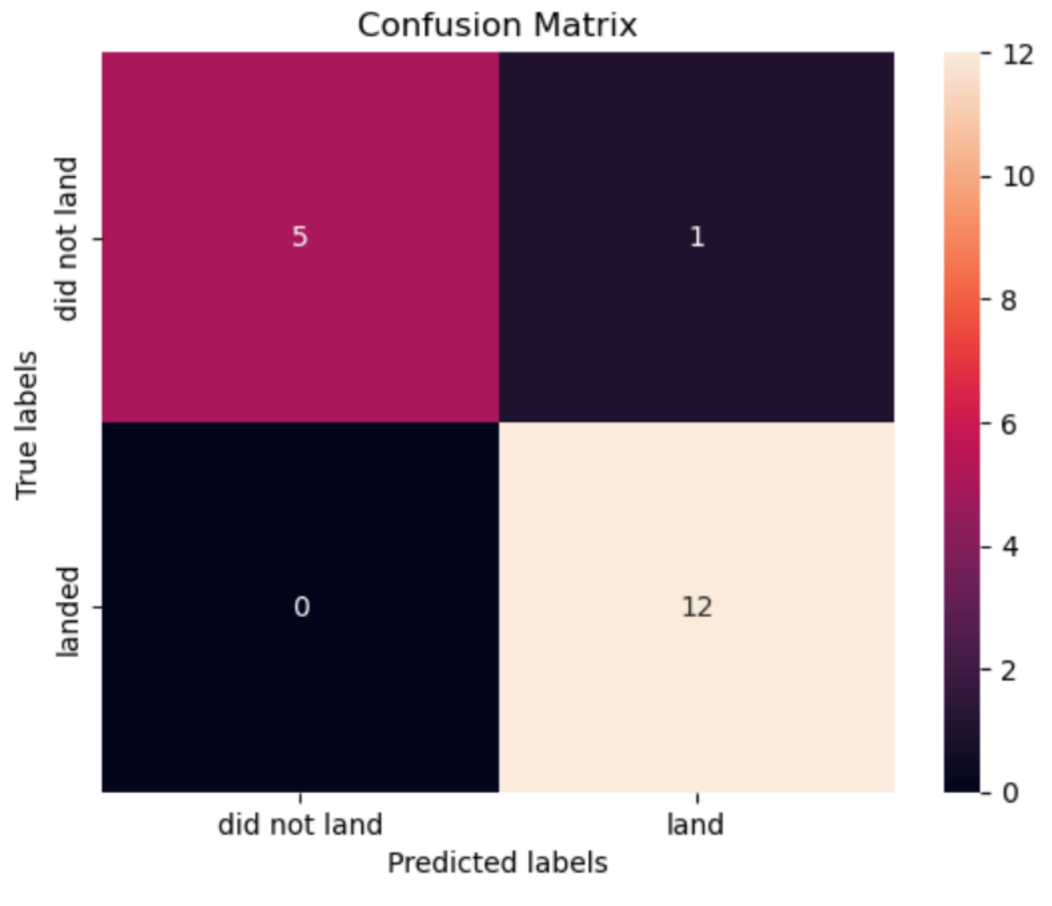
Predictive Analysis Results – KNN

- Using the score method, the calculated accuracy of KNN was 0.833333
- The following displays the confusion matrix from the trained KNN model. The confusion matrix displayed yielded the same results when using Logistic Regression and SVM, displaying the issue of false positive results.



Predictive Analysis Results – Final Results

- A function was created to account for all the total scores and compare them among each machine learning function. The result yielded the best method of prediction as Decision Tree Classifier with an accuracy of 0.944444
- It should be noted that the confusion matrix for Decision Tree Classifier is the only Confusion Matrix that yielded a different result from the other machine learning methods.



CONCLUSION

Utilizing the publicly provided SpaceX launch data and machine learning technology. SpaceY is able to place funding on the critical attributes that enable a successful landing outcome, therefore reducing cost by recycling the first stage of their Rocket launches.

In addition to highlighting the key attributes of a successful landing, the machine learning method, Decision Tree Classifier, allows for current rocket features to be tested without having to launch as landing predications can be made with an accuracy of 94 percent.