Winning Space Race with Data Science

Nikki Escamillas
December 17,2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

- Exploratory Data Analysis Results
- Interactive Analytics and Dashboarding
- Predictive Analysis

Introduction

With the advancement of science, improvement and innovation with Space technology has paved its way making space travel and exploration affordable for everyone. SpaceX , one of the leading space companies was able to designed Falcon 9 that can recover the first stage of the rocket making it lesser expensive compared to other companies .

As a startup company, this study will help to determine the price of each launch and determine if Space X will reuse the first stage specifically for Falcon 9 by using different machine learning algorithms . This will served as a basis for future study in order to compete with Space X.

Specifically, this study will try to answer the following questions:

1. What are the main factors affecting the success landing rate while reusing the first stage of the rocket ?
2. How are these factors related to each other ?
3. Can the success rate be predicted using a classification algorithm ?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology
  - API and Web scraping
- Perform data wrangling
  - Filtering , Dealing with missing data and Feature Engineering (Target Label and Encoding Categorical 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
  - Logistic Regression , K Nearest Neighbors, Support Vector Machines and Decision Trees

Data Collection - SpaceX API

1. Request Rocket Launch Data from SpaceX API
2. Parse the response content as .json() and turn it into Pandas dataframe using . Json_normalize()
3. Utilization of other API endpoints using custom functions to get needed data for columns represented as ID
4. Combine columns into one dataframe derive from the API calls through custom functions
5. Filter the data to only include Falcon 9 launches using Pandas sub setting
6. Deal with missing values by means of imputation using mean() and pandas .replace method
7. Export to csv using pandas .to_csv() method

Process the data to filter only rockets with 1 core and 1 payload

Convert date into a datetime object

This will serve as a reference for the custom functions for the next step

GITHUB Link: [Data Collection API]

Data Collection – Web Scraping

Request Falcon9 Launch from Wikipedia

Create a Beautiful Soup object from the HTML response

Extract all column/ variable names from the HTML table header by using the custom function extract_column_from _header ()

Create a data frame by parsing the launch HTML tables by iterating over the HTML tables and applying pre defined functions

Export to csv using pandas .to_csv() method

GITHUB Link: [Data Collection Web Scraping]

Data Wrangling

1. Load the data collected into a dataframe using pandas
2. Identify missing data using the isnull() method and Identify the data types for each feature using the dtypes attribute
3. Extract all column/ variable names from the HTML table header by using the custom function extract_column_from _header ()
4. Calculate the number of launches, occurrences of each orbit ,mission outcome using the value_counts() method of pandas
5. Segregate bad landing outcome from the value_counts () method
6. Create a new feature name "Class" that generalize whether the land is successful or not coding it as 1 or 0
7. Export to csv by using to_csv() method of pandas

GITHUB Link: [Data Wrangling]

EDA with Data Visualization

X axis	X axis data	Y axis	Y axis data	Type of Chart
Flight Number	Discrete	Payload Mass	Continuous	Scatter Plot
Payload Mass	Continuous	Launch Site	Categorical	Scatter Plot
Orbit	Categorical	Class – as success rate	Continuous	Bar Plot
Flight Number	Discrete	Orbit	Categorical	Scatter Plot
Payload Mass	Continuous	Orbit	Categorical	Scatter Plot
Years	Discrete	Class – as success rate	Continuous	Line Plot

GITHUB Link: EDA with Pandas and Matplotlib

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string ‘ CCA’
- Display Total Payload Mass carried by boosters launched by NASA (CRS)
- Display Average Payload Mass carried by booster version F9 V1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass
- List the records which will display the month names, failure landing outcomes in drone ship, booster versions , launch site for the months in year 2015
- Rank the count of landing outcomes between the date 2016-06-04 and 2017-03-20 in descending order

GITHUB Link: [EDA with SQL]

Build an Interactive Map with Folium

- Derived the launch sites' latitude and longitude for the processed data set and added a mark and a circle to each one of it. This is done to locate the launch site locations in the global map in order to identify its position relative to the equator and coast lines
- Added a marker cluster for each location that specifies the number of launch attempts and whether the launch is successful or not by assigning it to specific class with designated colors. In here, the launch site success rate can be assessed and evaluated further against the number of attempts.
- Identified proximities for each launch site such as railways, highways, coastline and cities to determine and to serve as a future reference for choosing launch site locations.

GITHUB Link: [Interactive Visual Analytics with Folium]

Build a Dashboard with Plotly Dash

- Added a launch site dropdown in order for the user to select which launch site to check. This is connected to a callback function that will render a pie chart that will show the success rate for each site
- Created a slider to which the user can select a range for the payload mass. This is connected to a callback function that will render a scatter plot of Payload Mass vs Class with an additional hue of the Booster Type. This will show if there is an existing relationship between success rate and payload mass as well as how to Booster Version affects the success rate.

GITHUB Link: Interactive Visual Analytics with Plotly Dash

Predictive Analysis (Classification)

Import necessary libraries for machine learning modeling → Load the cleaned and processed data as csv to pandas → Use Standard Scaler() to features or explanatory variables → Use train_test_split() to create a train and test set → Define the algorithms to be used for the classification task → Define the parameters needed for GridSearchCV in order to identify the best hyper parameters for the model → Train the model using the training sets → Evaluate the models and identify the best model based on specific metrics

Logistic Regression
Support Vector Machine
K Nearest Neighbors
Decision Tree

GITHUB Link: [Predictive Analysis]

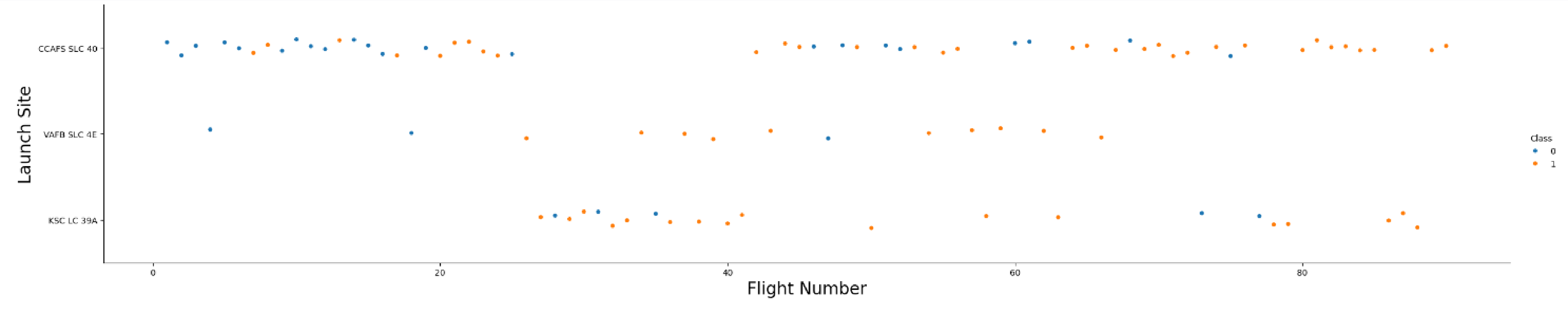
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

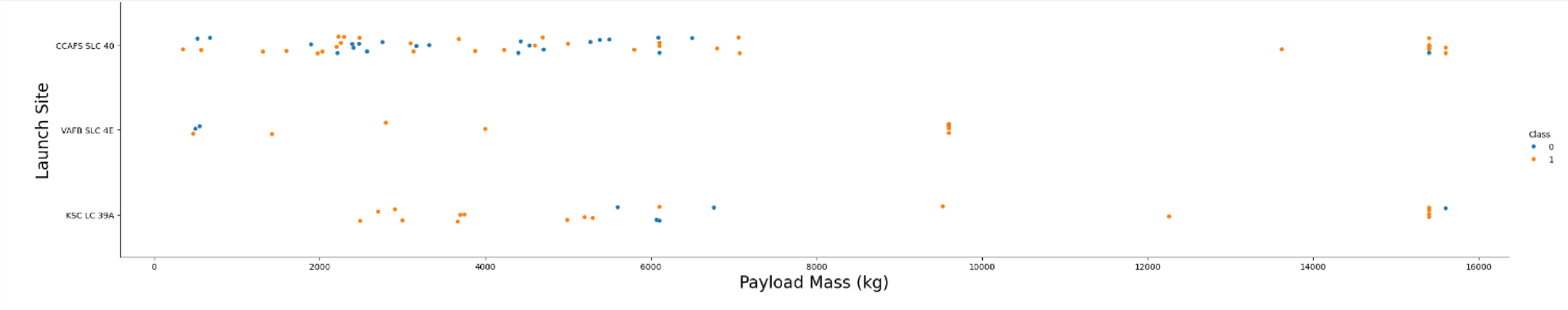
Insights drawn from EDA

Flight Number vs. Launch Site



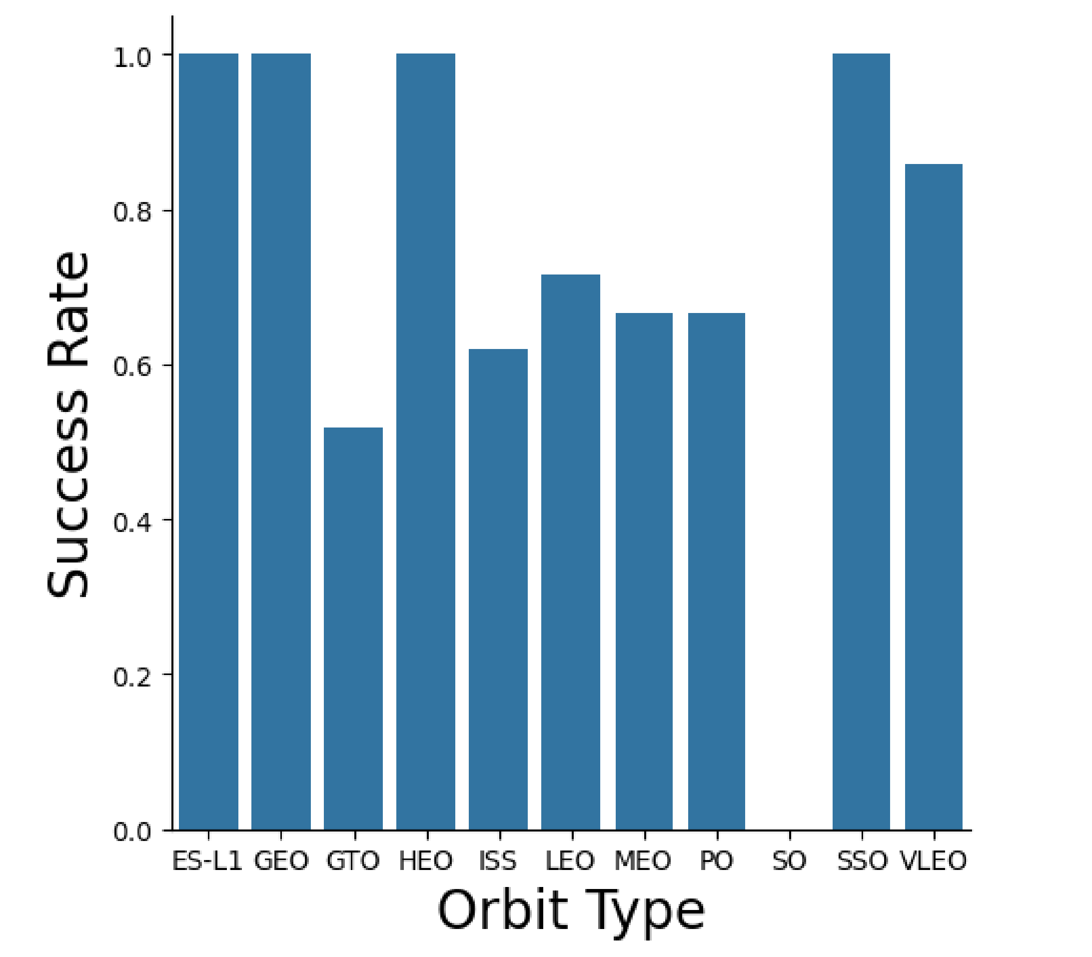
- CCAFS SLC 40 has the most number of flights
- As the flight number increases, the chances of success landing is higher.

Payload vs. Launch Site



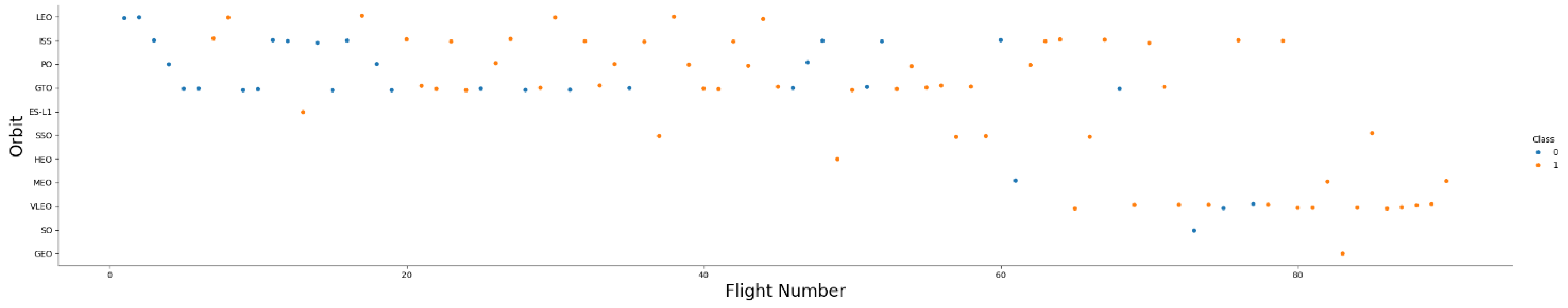
- Most of the launch attempts have a payload of less than 8000 kg
- Payload mass of higher than 8000 kg has only 2 failed landings

Success Rate vs. Orbit Type



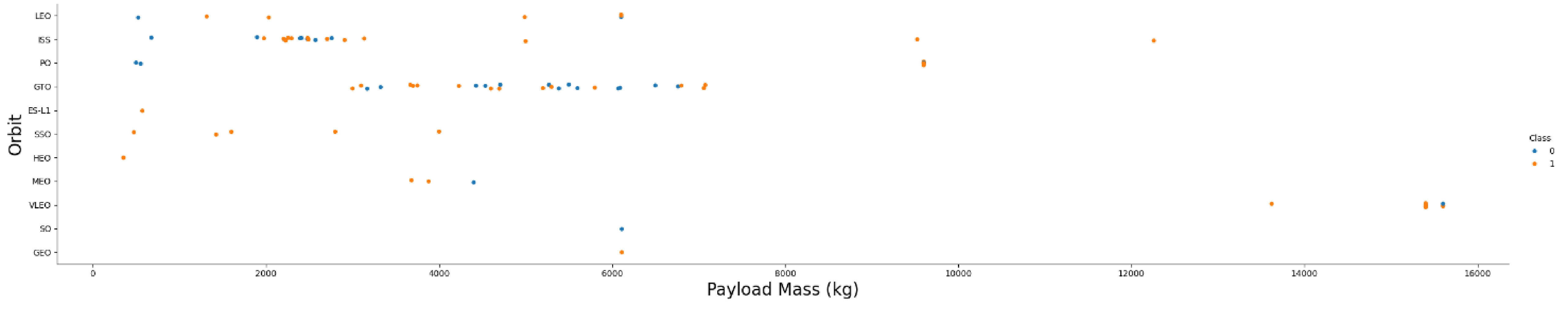
- ES-L1, GEO, HEO, SSO have the highest success rate based on orbit
- SO has 0 % success rate

Flight Number vs. Orbit Type



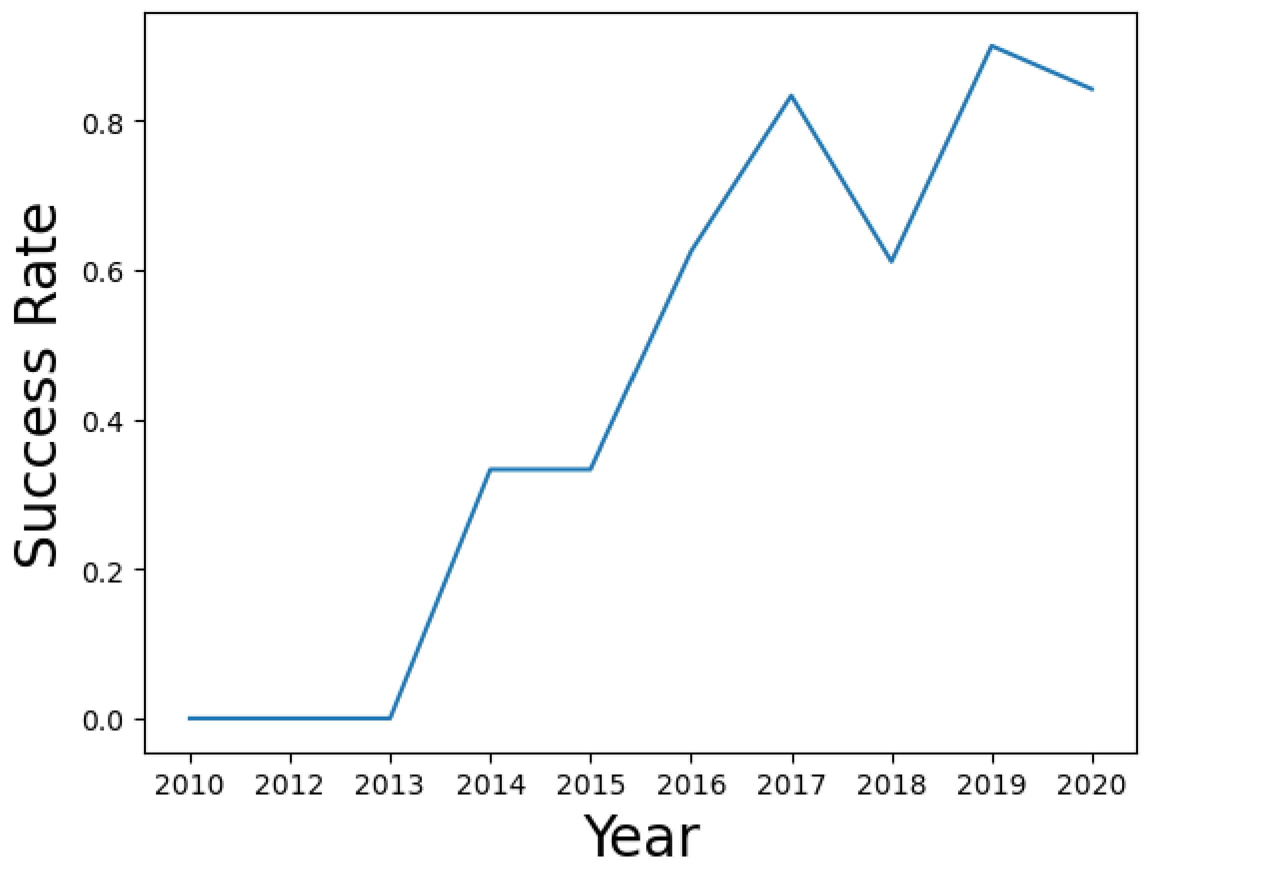
- LEO , ISS, PO, GTO and VLEO has the most number of flights
- SSO has 100 % success landing rate for the 5 attempts it has made
- ES-L1 and GEO each have only 1 flight which are both successful
- SO also has 1 flight but landing attempt was failed

Payload vs. Orbit Type



- As the payload increases for Polar, LEO and ISS increases , the success rate also increases
- There seems to be no relationship with payload mass value for GTO orbit

Launch Success Yearly Trend



- Success Landing Rate spiked up after 2013
- There was a dip in 2018 having 20% decrease
- Overall, the highest success rate is on the year 2019

All Launch Site Names

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

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL
```

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

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

- Identified the Distinct Launch Site

Launch Site Names Begin with 'CCA'

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

```
%sql SELECT * FROM SPACEXTBL 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

- Get at least 5 records where launch site begins with CCA

Total Payload Mass

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

```
%sql SELECT "Customer", SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE "Customer" ="NASA (CRS)" GROUP BY "Customer"
```

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

Customer	SUM(PAYLOAD_MASS__KG_)
NASA (CRS)	45596

- Calculated the total payload mass for customer NASA(CRS)

Average Payload Mass by F9 v1.1

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE "Booster_Version" LIKE "%F9 V1.1"
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

- Calculated the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT MIN("Date") FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (ground pad)"
```

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

MIN("Date")
2015-12-22

- Identified the first successful landing in a ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql SELECT "Booster_Version" FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (drone ship)" AND (PAYLOAD_MASS__KG_ >4000 OR PAYLOAD_MASS__KG_ <4000 )
```

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

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

- Listed the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT "Mission_Outcome",COUNT("Mission_Outcome") FROM SPACEXTBL GROUP BY "Mission_Outcome"
```

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

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

- Calculated the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql select booster_version from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL);
```

```
 * 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

- Listed the names of the booster which have carried the maximum payload mass

2015 Launch Records

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%%sql SELECT SUBSTR(date,6,2) AS month, booster_version, launch_site, landing_outcome FROM SPACEXTBL
      WHERE landing_outcome = 'Failure (drone ship)' AND SUBSTR(date,0,5)='2015';
```

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

month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Listed the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

Rank 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.

```
%%sql SELECT landing_outcome,COUNT(*) AS count_outcomes FROM  SPACEXTBL
      WHERE date BETWEEN '2010-06-04' AND '2017-03-20'
      GROUP BY landing_outcome
      ORDER BY count_outcomes DESC;
```

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

Landing_Outcome	count_outcomes
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

- 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

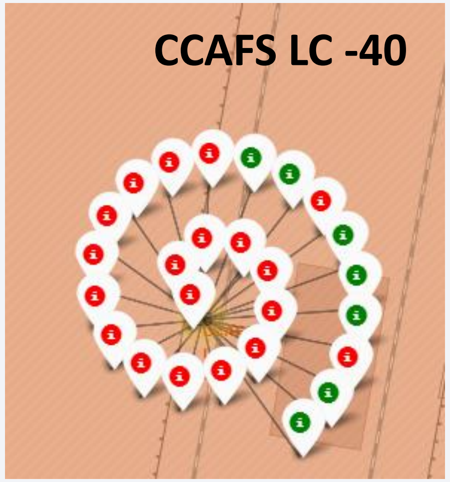
Section 3

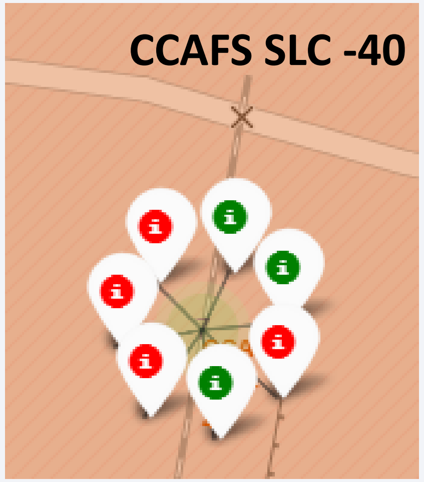
Launch Sites Proximities Analysis

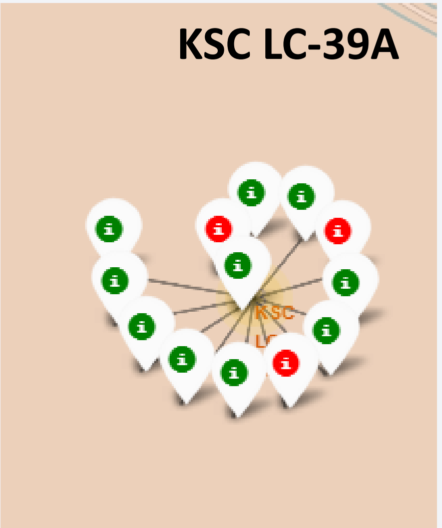
Launch Sites location

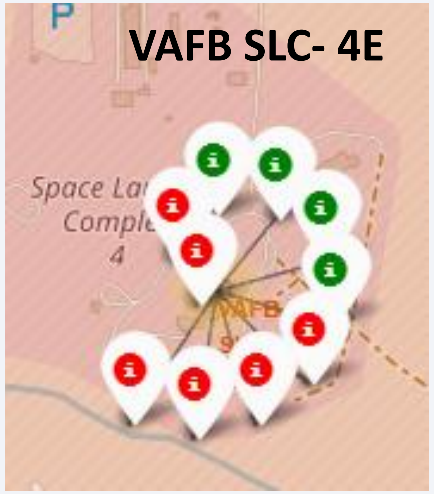
- All launch sites are near the equator and they are all in close proximity to the coast

Success and Failed Launches per Site









- With the markers having green as success and red as failed, launch sites CCAFS LC-40 has the most attempts while CCAFS SLC-40 has the least attempts.
- Overall, site KSC LC-39 A has the highest success rate compared to the other three

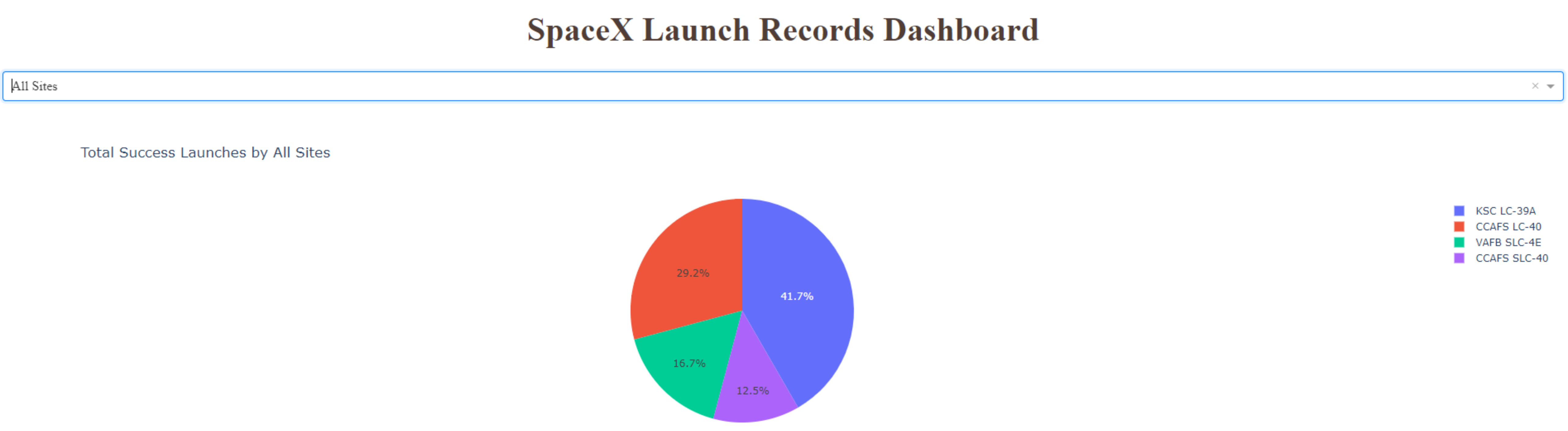
Proximity to Launch Site

- Focusing on launch site KSC LC-39 A, being the site with highest launch success rate . Its proximity to the city is the farthest comparing it to its proximity from highways, railways and coast line hovering to a value of approximately 16 kilometers . This is to ensure that it is far away from populated areas to ensure safety
- Its closest proximity is both for highway and railway which is not greater than 1 kilometer, indicating that the launch site is easily accessible
- For the coastline, its distance is around 3.5 kilometers. As seen on the global marked map as well, the other three sites are also very close to the coastline

Section 4

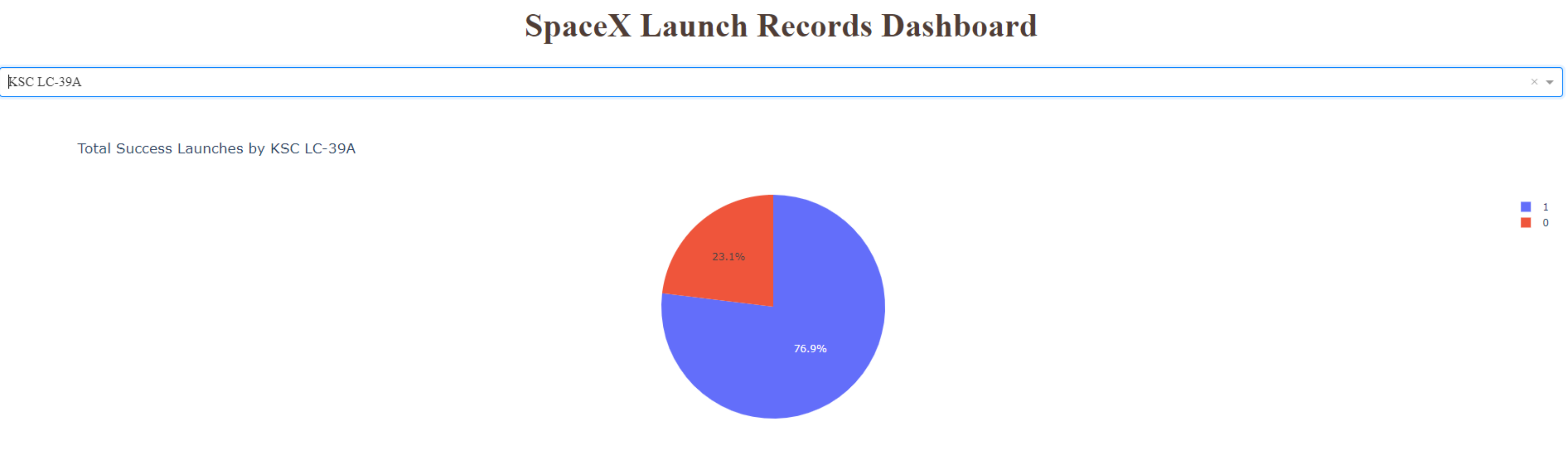
Build a Dashboard with Plotly Dash

Dashboard –Launch Success Count for All Sites



- KSC LC-39 A has the most successful launches

Dashboard- Launch Site with highest launch success ratio

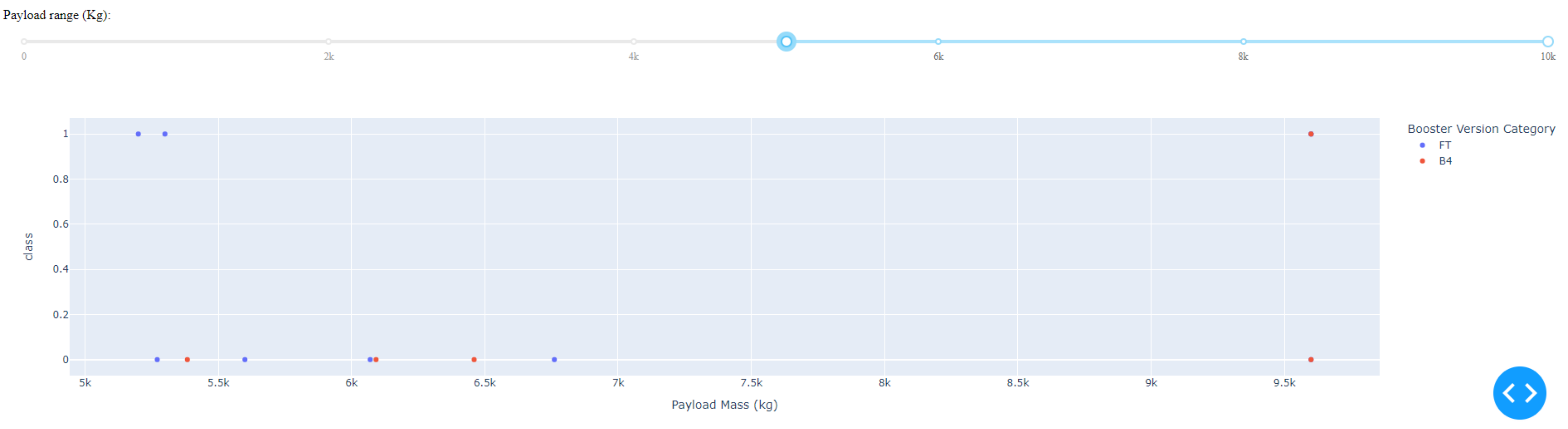


- KSC LC-39 A has 10 out of 13 successful landings

Dashboard –Payload Mass vs Launch Outcome

- Payload of less than 5000 kg has the most number of successful landing with booster version FT having the most count of successes

Dashboard –Payload Mass vs Launch Outcome



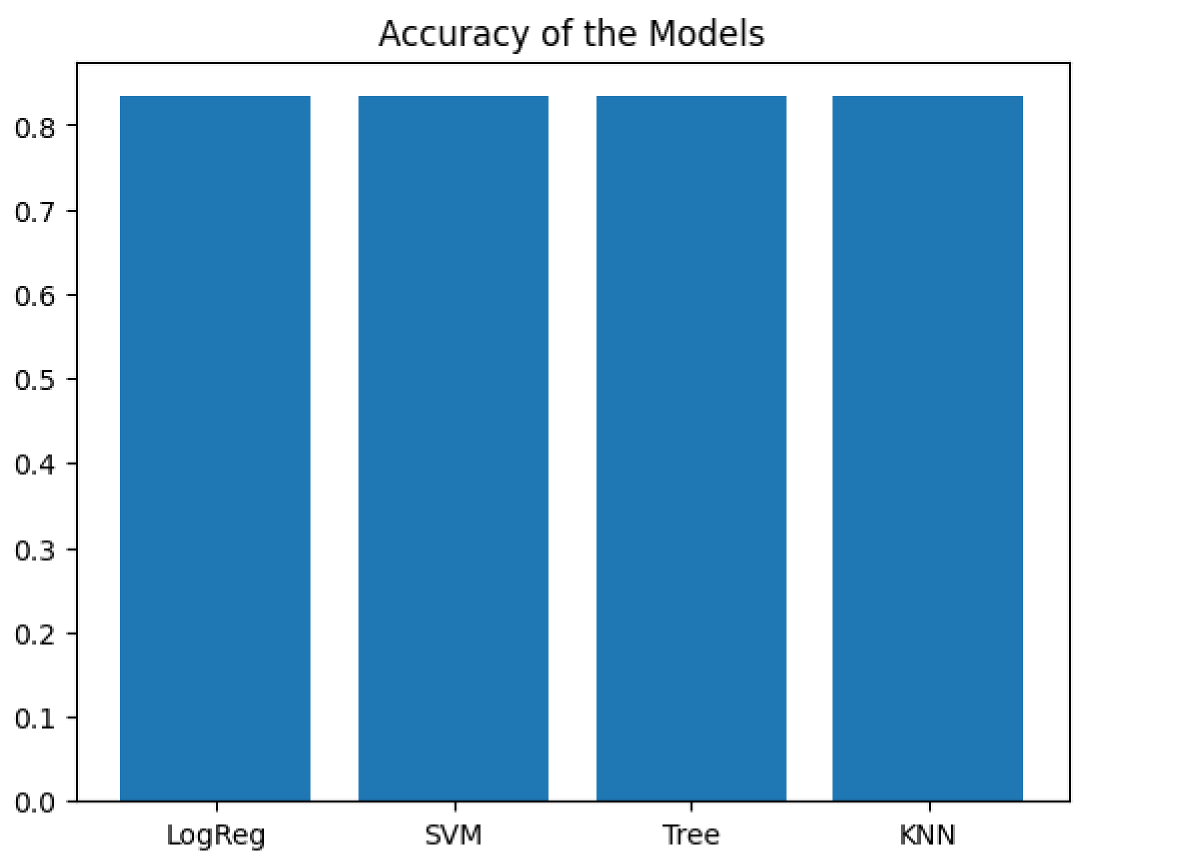
- Only FT and B4 booster version have a payload mass of greater than 5000 kg

Section 5

Predictive Analysis (Classification)

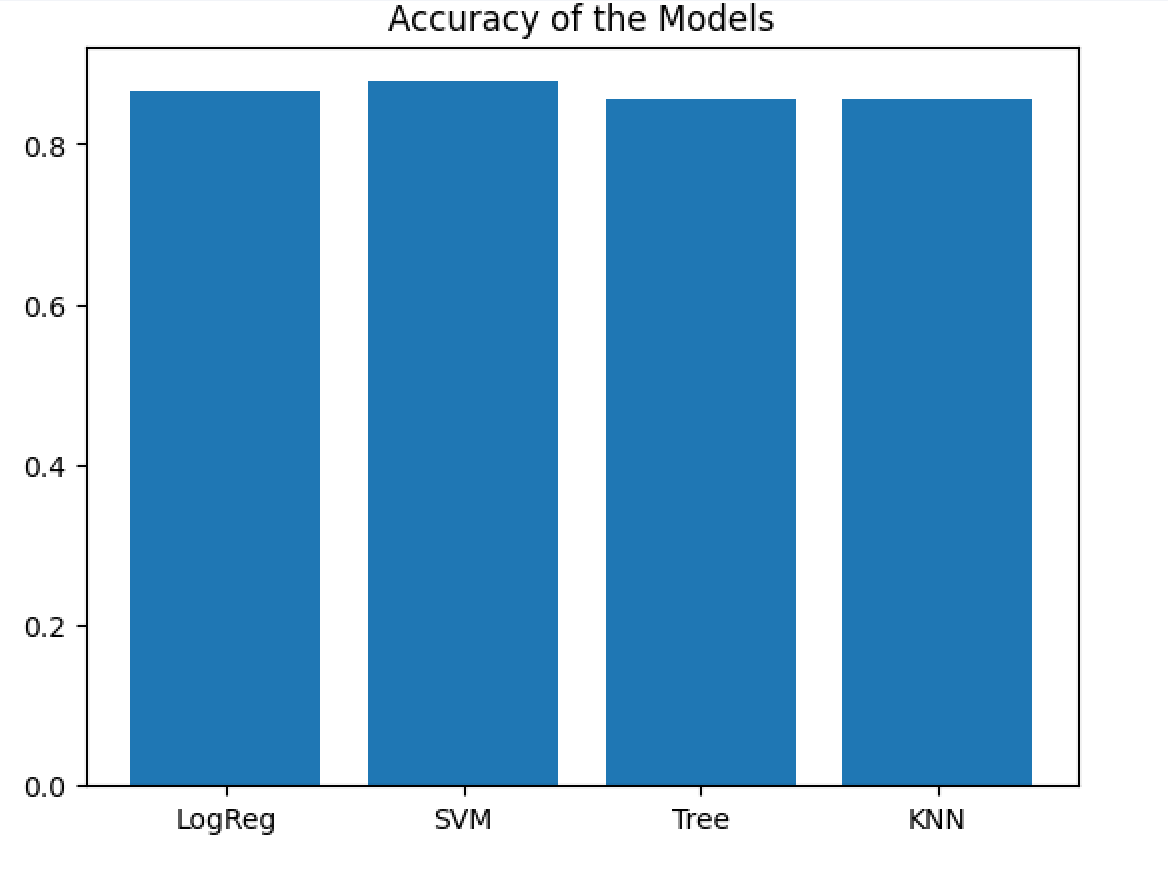
Classification Accuracy

- Same Test Accuracy of 83.33 % for all models. This can be attributed to the few test set accounting to only 18 instances



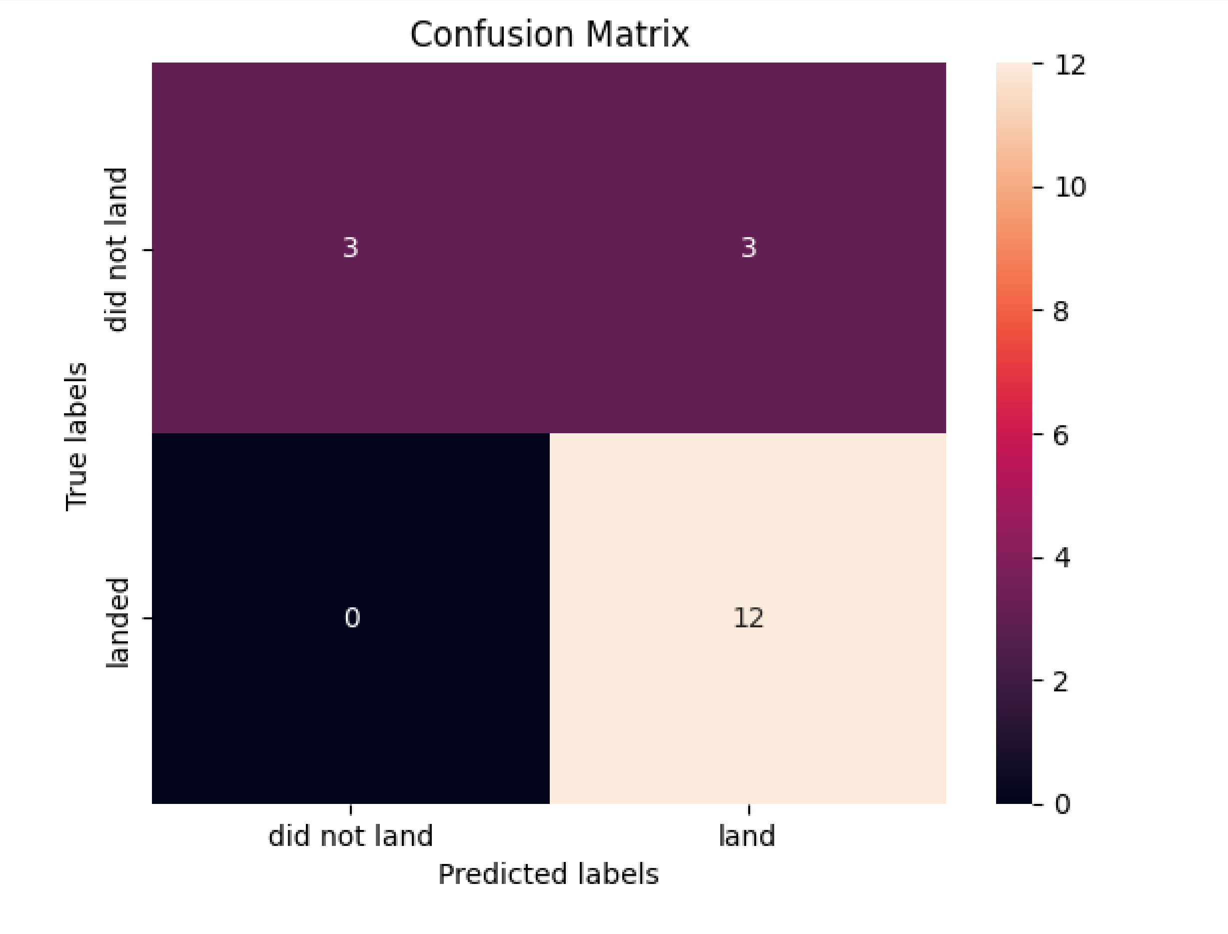
Classification Accuracy

- SVM has the highest accuracy (94.44%) when using the whole dataset.
- This is also reflected in its Jaccard score of 92 % and F1 score of 96 % which is also the highest comparing the four models which was tested.



Confusion Matrix

- The model was able to classify all classes of successful landing correctly
- 3 instances were classified as False Positives indicating that the model is having a hard time identifying those that did not land correctly



Conclusions

- From the period of 2013, the launch success rate is increasing.
- With the number of flights increasing , the success rate is also increasing. This indicates that further optimization and improvements are being made based on the experience from previous launches.
- As there are more launch attempts in lighter payload mass ,it cannot be concluded that lighter payload mass yields to higher chance of success rate as there are fewer attempts on heavier payload.
- Among the orbits , SSO can be seen as the most successful since it has a 100 % success rate for all the attempts it has made.
- FT and B4 booster version is being used for heavier payload with FT having the most count of successes.
- The launch sites are in close proximity to the Equator Line and are near to coast lines, having KSC LC-39A to have the highest success rate compared to the other sites .Its success is reflected on payload mass under 5000 kg
- For the predictive analysis, SVM yielded the best results based on the specific metrics implemented. However, this is evaluated on the whole dataset. Implementing it solely on the test data set yielded similar results to the other model. Additional data is needed to increase the test size that might affect and change this initial result.
- This additional data can also help increase the True Negative Rate or Specificity of the model, as the model has 50 % FP for all data that is labeled as TN

Thank you!