IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Aditya Pasari

10-05-2023

Outline

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

Executive Summary

SUMMARY OF METHODOLOGIES

Data was collected from SPACEX API and Webscraped from Wikipedia. Data Wrangling was performed followed by EDA using Data Visualization and SQL. Interactive visual analytics and dashboard was created using Folium Map and Plotly. Finally, the data was modeled using various Machine Learning classification algorithms

SUMMARY OF RESULT

Using the Machine Learning classification algorithm we can predict with an accuracy of 83% that the first stage of Falcon 9 will land successfully and therefore the first stage shall be reusable which will reduce the flight cost significantly. We have also discovered relationships between launch site and successful launch, between payload mass and successful launch.

Introduction

- The objective of this project is to 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.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected from the SPACEX API and webscraped from Wikipedia
- Perform data wrangling
  - One hot encoding was performed on the data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Try various classification methods and see which gives the most accurate result

Data Collection

- The following slides show how data was collected using –
  - SPACEX API
  - Webscraping

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- https://github.com/adityapa80/SpaceX-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

https://api.spacexdata.com/v4/launches/past

↓

Normalize JSON

↓

Get data about Rockets, Launchpads, Payloads and Core by calling the relevant API –

- api.spacexdata.com/v4/rockets
- api.spacexdata.com/v4/launchpads
- api.spacexdata.com/v4/payloads
- api.spacexdata.com/v4/cores

↓

Convert to DataFrame

Data Collection - Scraping

- On URL use requests.get
- BeautifulSoup on the response
- Work through the HTML tags

https://github.com/adityapa80/SpaceX-Capstone/blob/main/jupyter-labs-webscraping.ipynb

data = requests.get(static_url).text

↓

soup = BeautifulSoup(data, 'html.parser')

↓

html_tables = soup.find_all('table')

↓

Iterate through the table and extract relevant information

Data Wrangling

The data set contained several different cases where the booster did not land successfully and each of these cases was labelled differently such as “True Ocean”, “False Ocean”, “True RTLS”, “False RTLS”, “True ASDS”, “False ASDS”.

These different cases were converted into 2 values – 1 if booster landed successfully and 2 if it was unsuccessful. This was stored in the DataFrame as a new column called “Class”

https://github.com/adityapa80/SpaceX-Capstone/blob/main/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

The following scatter charts were plotted showing their relationships having successful or unsuccessful launches –

- X=Flight number: Y=Payload
- X=Flight number: Y=Launch Site
- X=Payload: Y=Launch Site
- X=Flight Number: Y=Orbit
- X=Payload: Y=Orbit

Bar Plot – Success Rate of each orbit

Line Plot – Success Rate over the years

https://github.com/adityapa80/SpaceX-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

SQL Queries performed –

- Display the distinct launch sites
- Display the first 5 Launch Sites beginning with "CCA"
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display the 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 with successful drone ship landings and payloads between 4000 and 6000 kg
- List the total number of successful and failure mission outcomes
- List the booster versions which have carried the maximum payload mass
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- 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

https://github.com/adityapa80/SpaceX-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(2).ipynb

Build an Interactive Map with Folium

- First created a Folium map showing location of NASA Space Center, Houston
- Added locations of each of the launch sites on the map by using circles and markers
- Created Marker clusters showing landing outcomes at each site
- Created markers and lines to show distance between launch site and places such as nearest coastline, nearest city, railway and highway

https://github.com/adityapa80/SpaceX-Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite%20(1).ipynb

Build a Dashboard with Plotly Dash

A Dashboard was created using Plotly having 2 inputs –

- Dropdown List for Launch Site selection
- Range Slider to select Payload

The Dashboard had the following 2 outputs –

- Pie-Chart showing the success rate of landing at each of the launch sites
- Scatter Plot showing the successful landings by booster type and payload mass(whose range could be input by user)
- These combinations allow the user to get different insights as to which launch sites has the highest success rate and relationship between the payload mass and successful landing.

https://github.com/adityapa80/SpaceX-Capstone/blob/main/spacex_dash_app%20(1).py

Predictive Analysis (Classification)

The following four Machine Learning Classification models were chosen -

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- Decision Tree (Tree)
- K-nearest Neighbors (KNN)

- All data was standardized and normalized using StandardScaler() and data was split into test sets and train sets.
- Accuracy for each of the models was checked and the accuracies of all the models were compared.

https://github.com/adityapa80/SpaceX-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

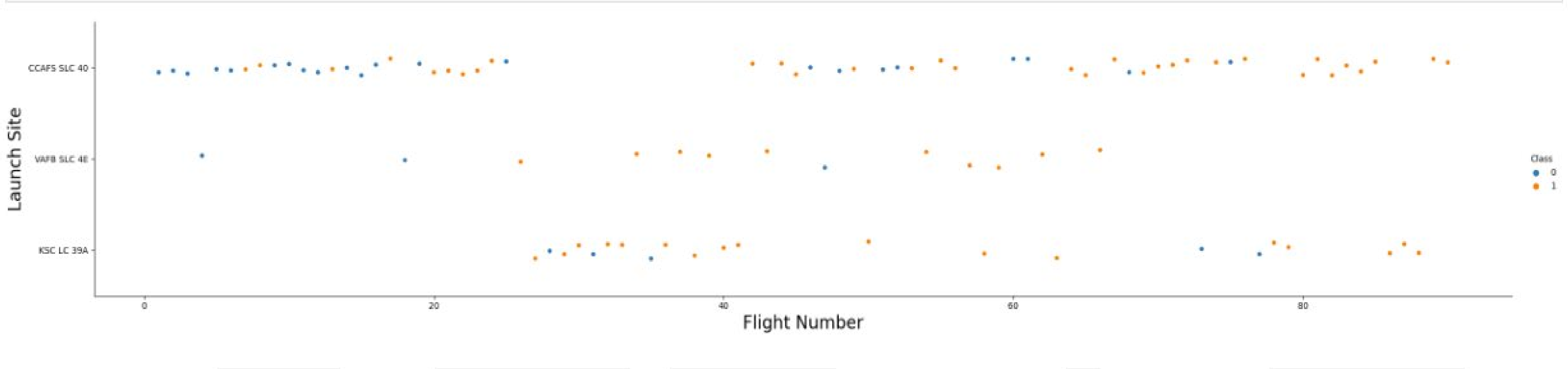
Results

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

Section 2

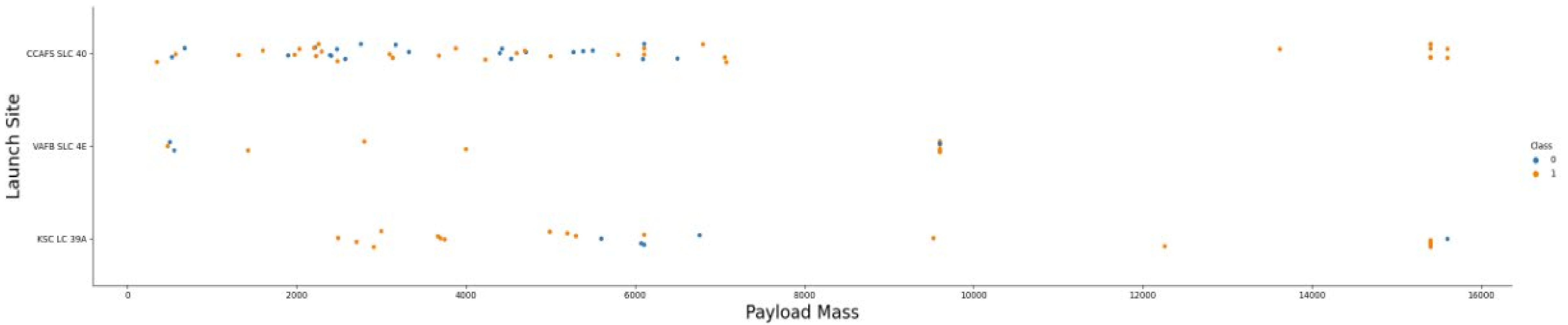
Insights drawn from EDA

Flight Number vs. Launch Site



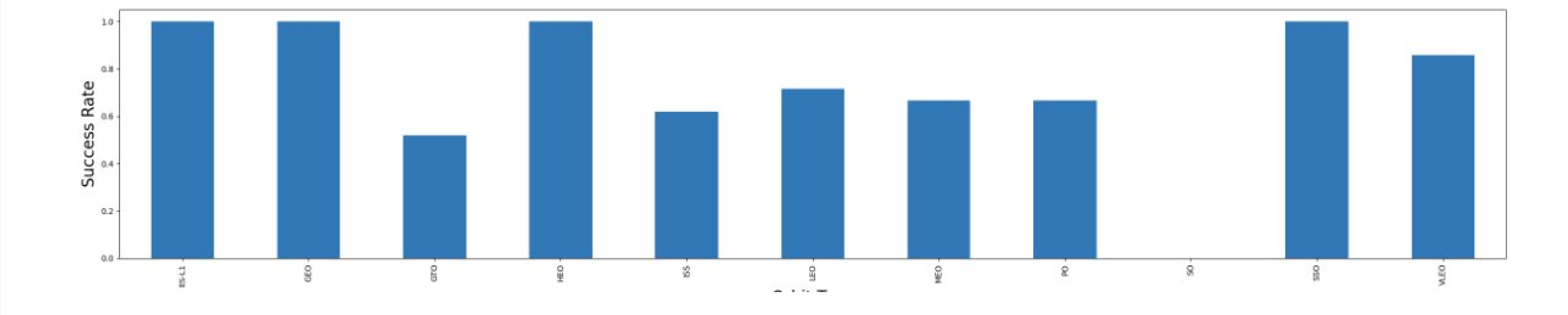
- Successful landing outcomes increased as flights increased.
- Initially only CCAFS SLC40 launch site was used, but the initial launches had high failure rate
- KSC LC39A seems to have had the highest % of successful launches

Payload vs. Launch Site



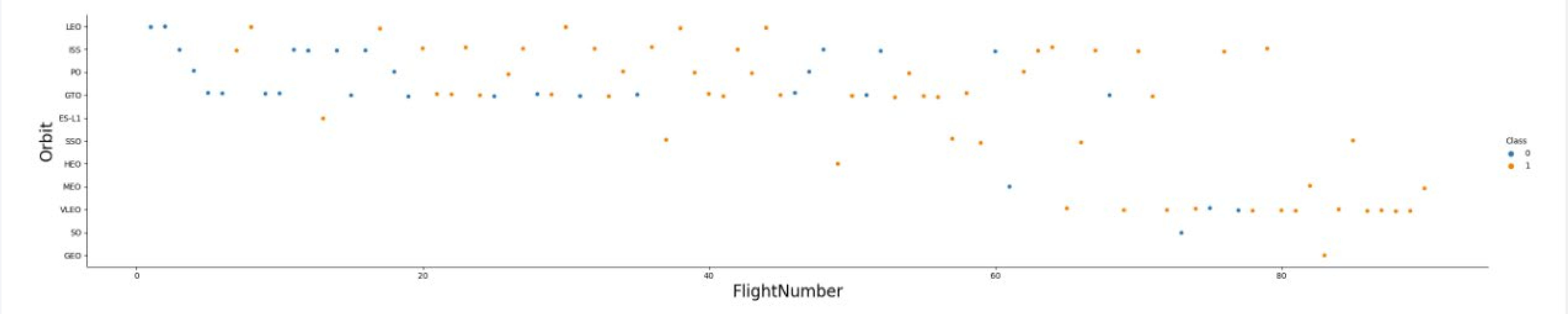
- Payloads over 10000Kg have high probability of success
- CCAFS SLC40 has 100% success rate when payload is more than 8000Kg

Success Rate vs. Orbit Type



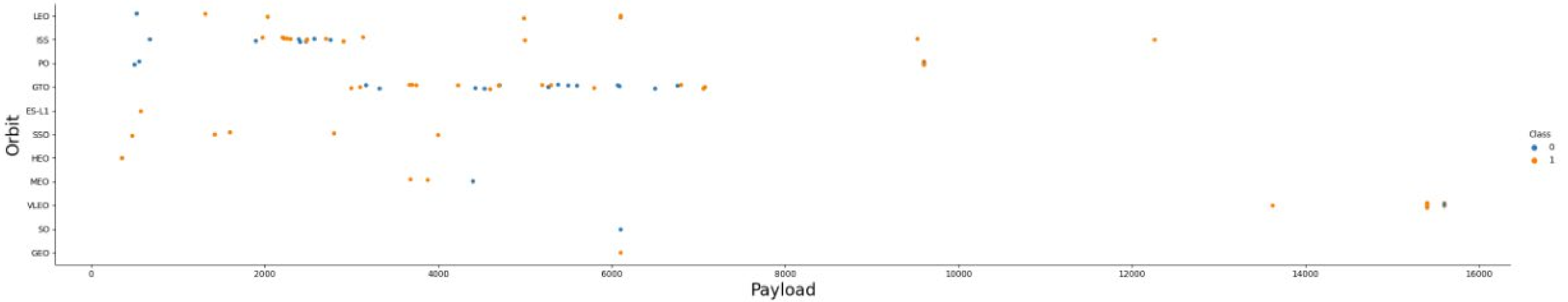
- GTO has the lowest success rate
- ES-L1, GEO, HEO, SSO and VLEO have the highest success rates

Flight Number vs. Orbit Type



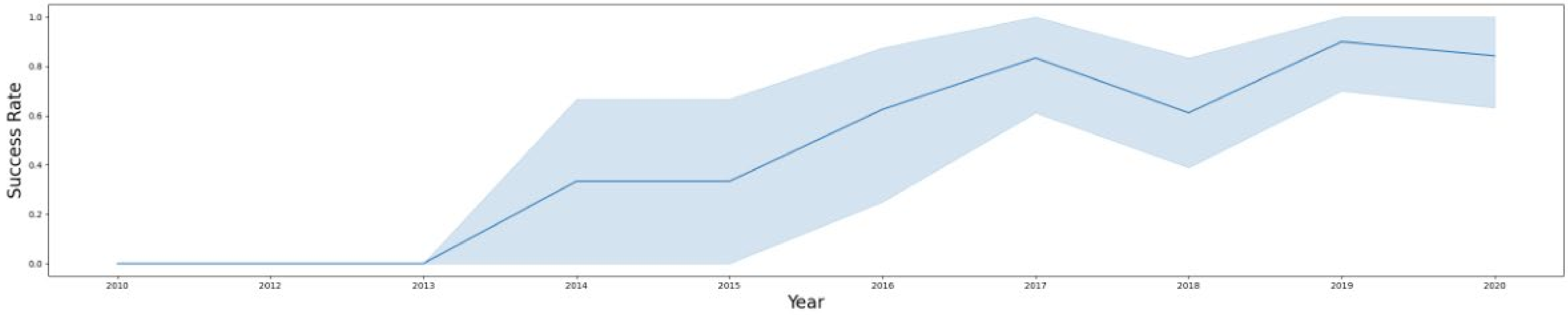
- Initially only 4 orbits were tried (LEO, ISS, PO, GTO) and they all failed.
- As flights increased, the success in each orbit has increased
- GTO Orbit continues to experience failures

Payload vs. Orbit Type



- GTO orbits only have payloads between 2K-8K
- Highest Kg payload goes to VLEO Orbit

Launch Success Yearly Trend



- Success rate since 2013 shows positive trend
- First 3 years there was no success
- 2019 was the best year

All Launch Site Names

In [7]:
```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;
```

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

Out[7]:

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

- 4 distinct Launch Sites selected

Launch Site Names Begin with 'CCA'

```
In [23]: %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;

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

Out[23]:

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

- Five launch sites starting with CCA selected using LIKE and LIMIT

Total Payload Mass

In [10]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';
```

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

Out[10]:

SUM(PAYLOAD_MASS__KG_)
45596

- Use SUM to determine total Payload for Customer NASA (CRS)

Average Payload Mass by F9 v1.1

```
In [12]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1%';

 * sqlite:///my_data1.db
Done.

Out[12]:
```

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

- Use AVG and WHERE clause to determine average payload mass for F9 v1.1

First Successful Ground Landing Date

```
In [31]: %sql SELECT MIN(Date) FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (ground pad)';

 * sqlite:///my_data1.db
Done.

Out[31]:
```

MIN(Date)
01-05-2017

- Use MIN to determine first successful landing outcome in 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

In [32]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE "LANDING _OUTCOME" = 'Success (drone ship)' AND 4000 < PAYLOAD_MASS__KG_ < 6000
```

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

Out[32]:

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

Total Number of Successful and Failure Mission Outcomes

```
In [35]: %sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) FROM SPACEXTBL GROUP BY MISSION_OUTCOME;

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

Out[35]:

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

- Use COUNT and GROUP BY to select total number of successful or failure mission outcomes

Boosters Carried Maximum Payload

```
In [36]: %sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);

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

Out[36]:

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

- By using a subquery in the WHERE clause, we can select the MAX payload as the condition in the subquery

2015 Launch Records

```
In [46]: %sql SELECT "Landing _Outcome", substr(Date, 4, 2), BOOSTER_VERSION, LAUNCH_SITE from SPACEXTBL where "Landing _Outcome" LIKE 'Fa

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

Out[46]:

Landing _Outcome	substr(Date, 4, 2)	Booster_Version	Launch_Site
Failure (drone ship)	01	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	04	F9 v1.1 B1015	CCAFS LC-40

- Use substr(Date) to select 2015 launch records

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

```
[21]: %%sql
SELECT "Landing _Outcome", COUNT("Landing _Outcome") AS SUCCESS_LANDS
FROM SPACEXTBL WHERE "Landing _Outcome" LIKE "Success%"
AND date BETWEEN '04-06-2010' AND '20-03-2017'
GROUP BY "Landing _Outcome" ORDER BY SUCCESS_LANDS DESC

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

[21]:

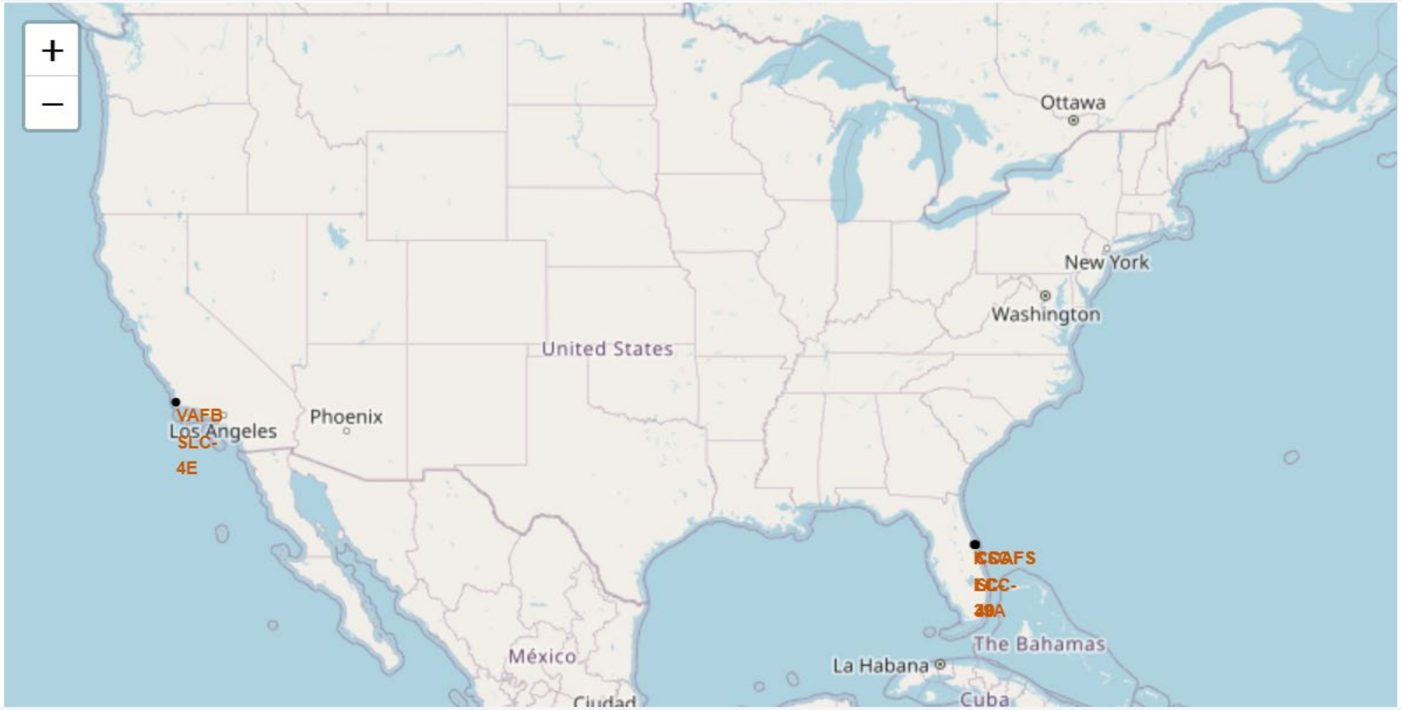
Landing _Outcome	SUCCESS_LANDS
Success	20
Success (drone ship)	8
Success (ground pad)	6

- Use the COUNT(), GROUP BY functions on the landing outcomes and use WHERE to select the date range to rank the successful landing outcomes
- Use DESC to order the list in descending order

Section 3

Launch Sites Proximities Analysis

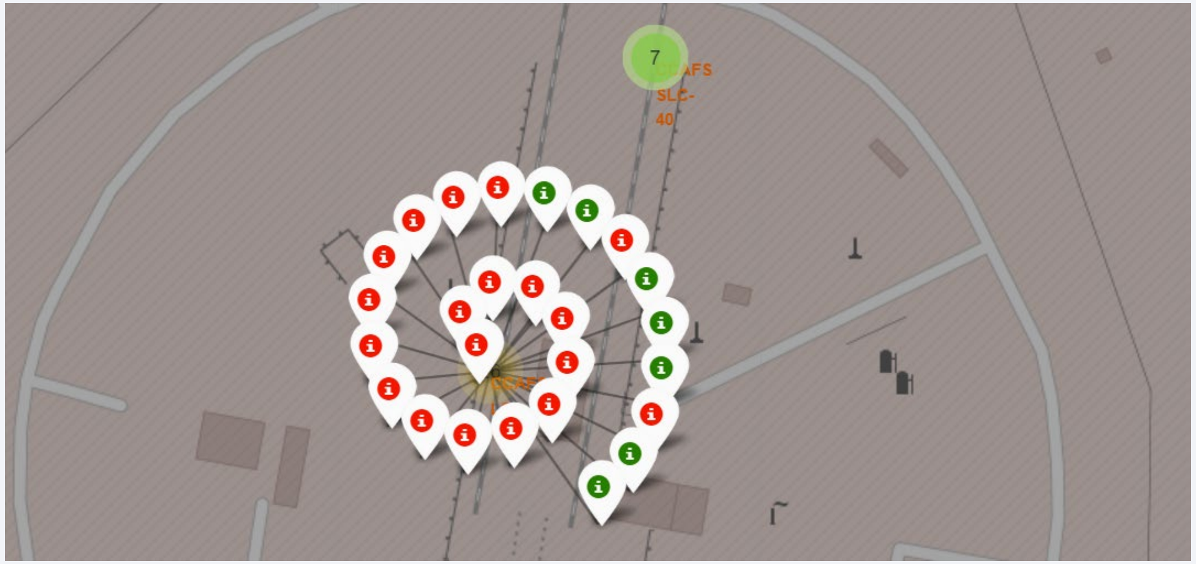
Locations of the Launch Sites



- Map shows the launch sites – in FL and CA

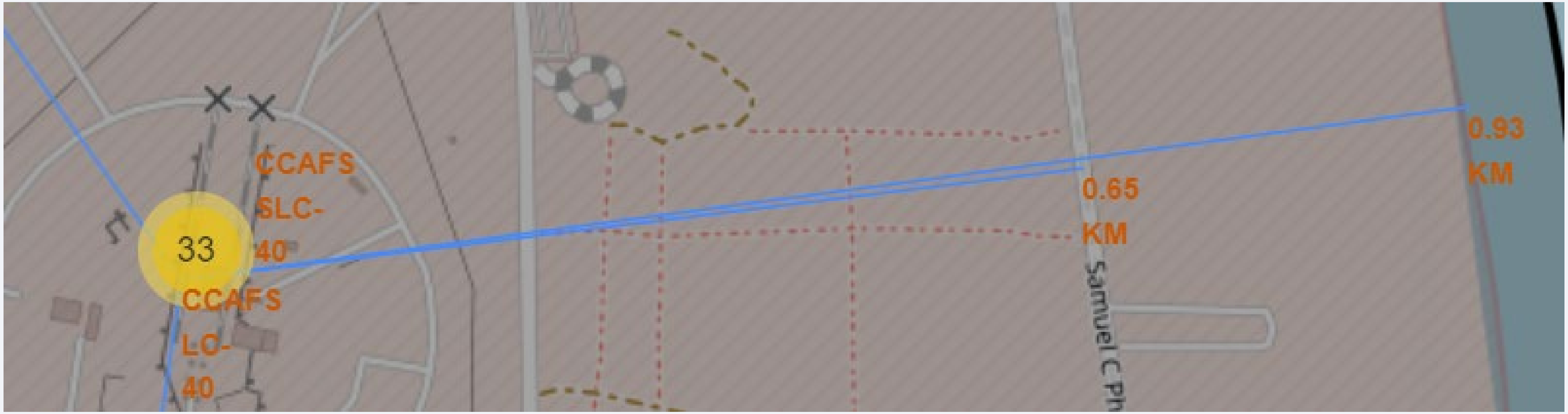
- They are close to the coastline and close to the equator

CCAFS SLC40 Site Landing Outcomes



- Clusters showing the landing outcomes at each site
- Zooming into the cluster shows the number of successful and unsuccessful landing outcomes at the site

Distance from CCAFS Site to highway and coastline



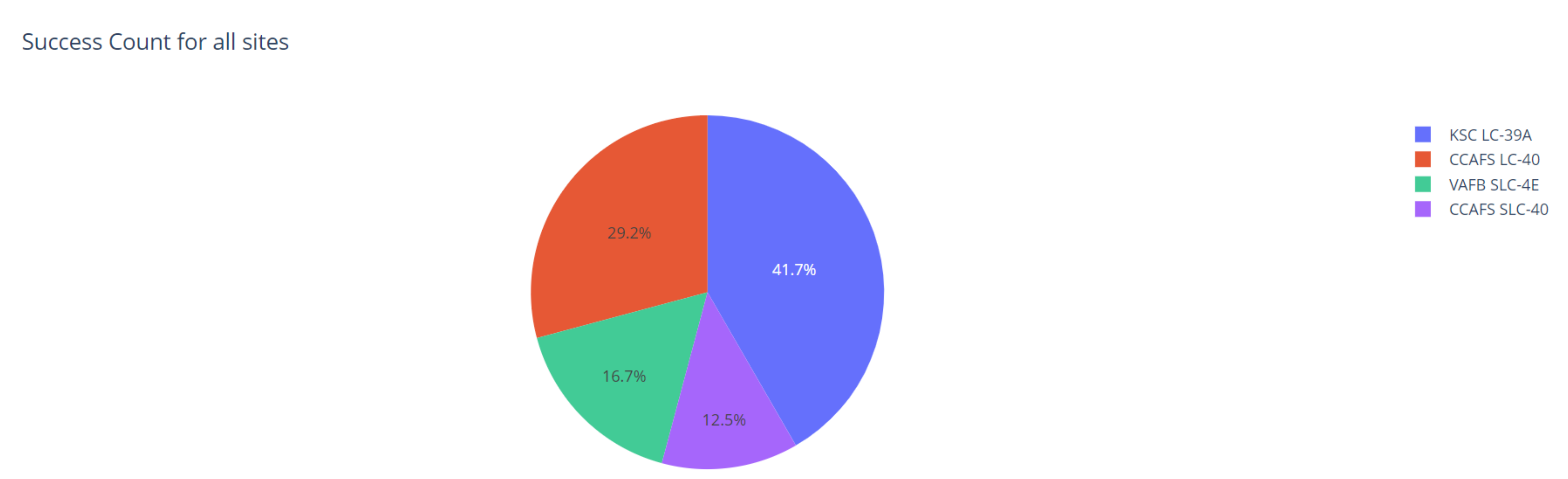
- Map shows distance from CCAFS site to nearest highway (0.65 Kms) and to coastline (0.93 Kms)

- The sites are closer to railways, highway and coastline but further away from city

Section 4

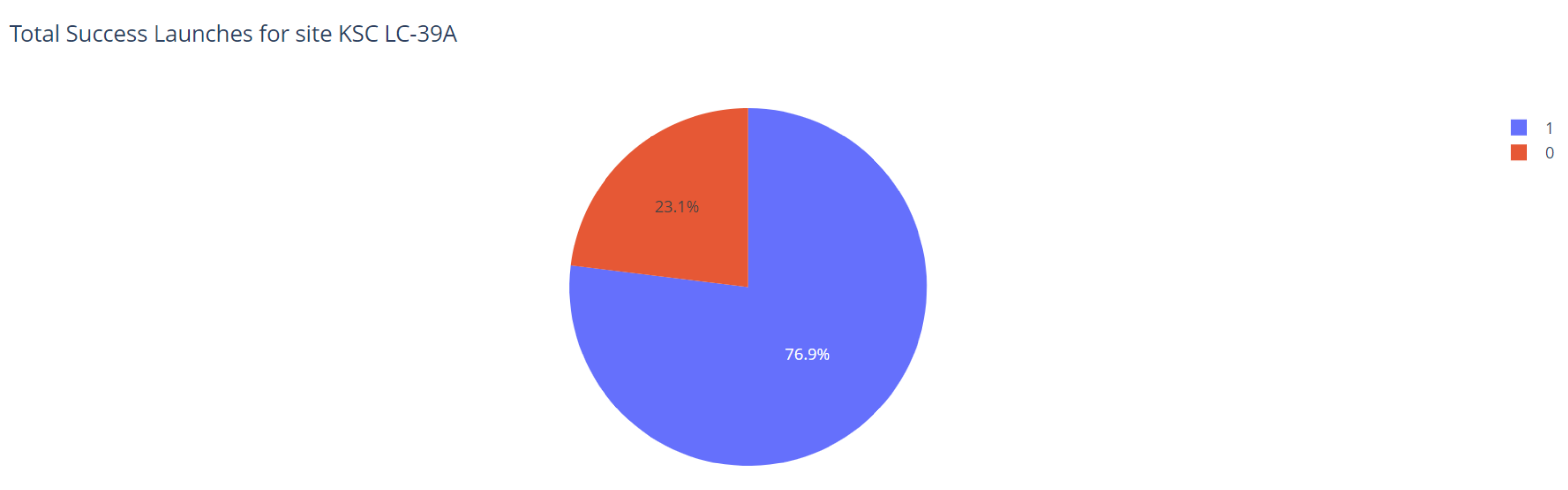
Build a Dashboard with Plotly Dash

Success Count for all sites



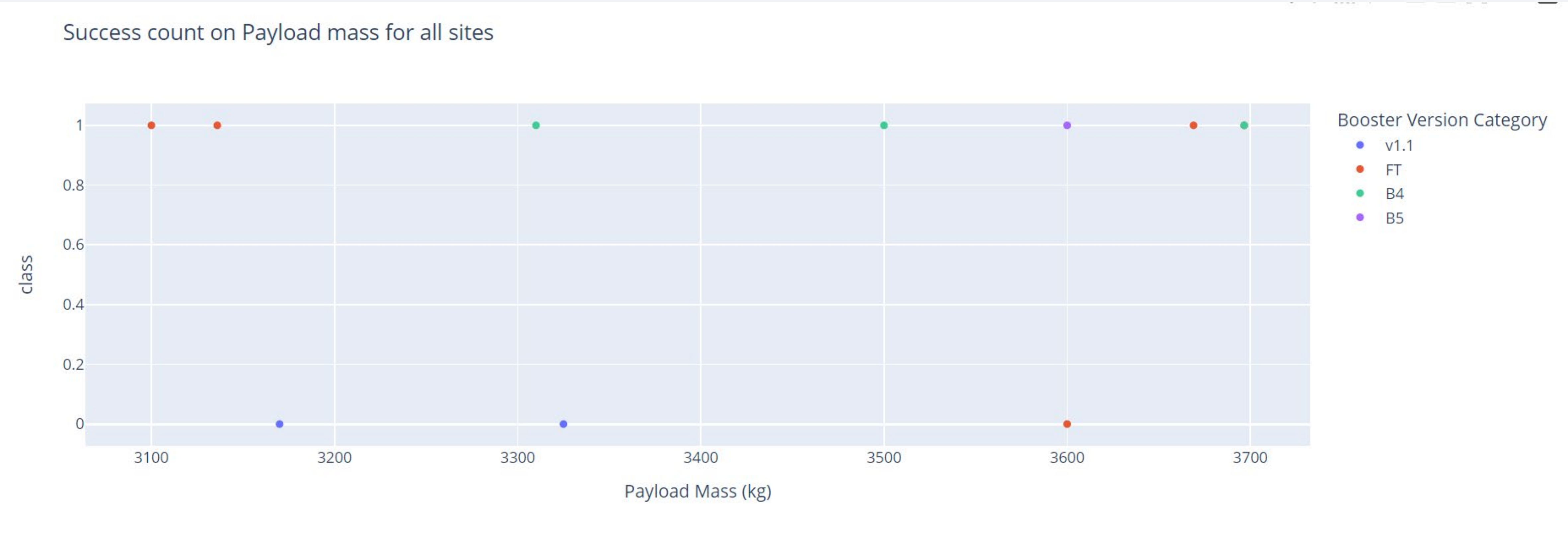
- Chose ALL SITES using the dropdown menu and the resultant pie-chart shows the different success rates of each site
- KSC LC-39A has the highest success rate % (41.7%)

Total Success Launches for KSC LC-39A



- KSC LC-39A has success rate of 76.9% which is highest amongst all sites

Success count on Payload Mass

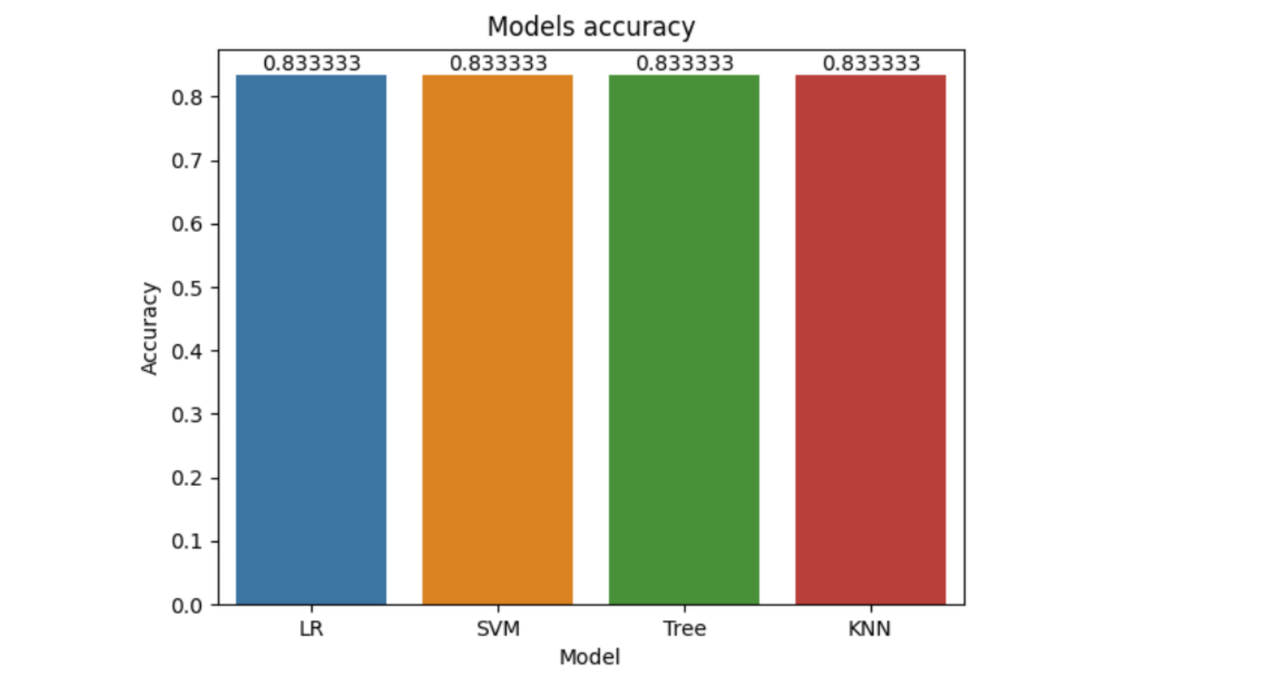


- Using the slider, it is observed that payloads between 3000 and 4000 Kgs have most number of landing successes (7 times)

Section 5

Predictive Analysis (Classification)

Classification Accuracy



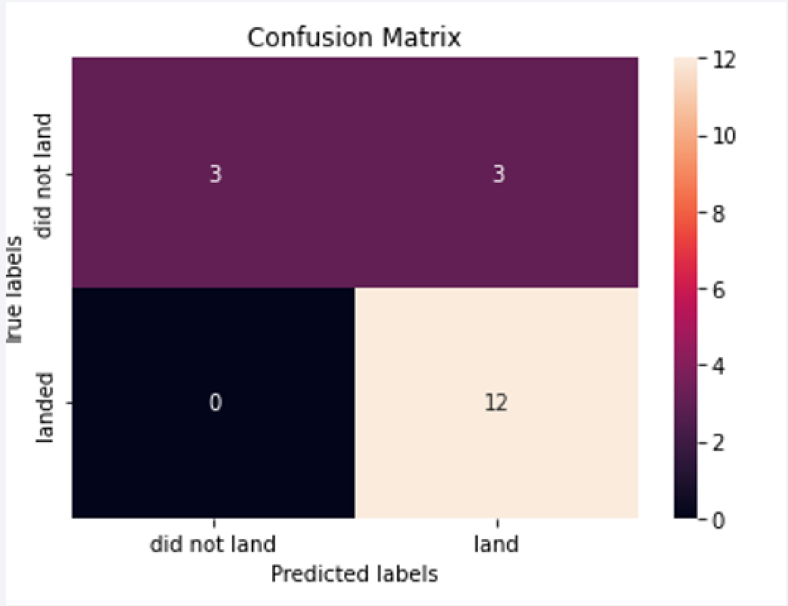
All the models returned an equal accuracy of 0.83

Confusion Matrix

This is confusion matrix for decision tree model

- It shows 0 false negatives

This model can be selected



Conclusions

- Since 2013, SpaceX is improving the landing outcomes whereas 2010-2013 had all failures
- 2019 was the best year for landing outcomes
- Flights from KSC LC39A Launch Site has a 76.9% success rate of landing the booster successfully
- The Launch Sites are strategically chosen – close to coastline, equator, railway and highway but further away from cities
- Highest Kg payload goes to VLEO Orbit
- The Decision Tree Classifier looks to be the best Machine Learning Algorithm for this application – It predicts with 83% accuracy that the first stage of Falcon 9 will land successfully.

Thank you!