

A photograph of a Space X Falcon 9 rocket launching at night. The rocket is illuminated by its own engines, creating a bright, fiery trail. A large, billowing cloud of white smoke and fire surrounds the base of the rocket. The launch is taking place over a body of water, which is visible in the foreground.

IBM DATA SCIENCE CAPSTONE PROJECT

**Space X Falcon 9 Landing
Analysis**

**Pooja Karmokar
10th Jan 2023**



OUTLINE

Executive Summary

Introduction

Methodology

Results

- Visualization – Charts
- Dashboard

Discussion

- Findings & Implications

Conclusion

Appendix



EXECUTIVE SUMMARY

Methodologies in use:

The project is followed by collecting data using:

- Data collection using API
- Data Wrangling

Followed by EDA (Exploratory Data Analysis) using:

- SQL
- Pandas and Matplotlib

Performing Data Analysis and Visualization by:

- Folium
- Plotly

And finally applying classification methodologies of:

- Predictive Analysis



Introduction

Falcon 9 rocket launches by SpaceX cost about \$62 million. This is significantly less expensive than other providers (which typically cost more than \$165 million), and a large portion of the savings is due to SpaceX's ability to land and reuse the rocket's first stage.

We can estimate the cost of a launch and use this data to decide whether or not a different company should compete against SpaceX for a rocket launch if we can anticipate whether the first stage will land. In the end, this project will be able to forecast if the Space X Falcon 9 first stage will land safely.



Section 1

Applied Methodology



Applied Methodology

Data Collection

- GET requests to the SpaceX REST API
- Web Scraping

Data Wrangling

Used `.value_counts()` method to determine the following:

- Number of launches on each site
- Number and occurrence of each orbit
- Number and occurrence of mission outcome per orbit type

Creating a landing outcome label that shows the following:

- 0 when the booster did not land successfully
- 1 when the booster did land successfully



Applied Methodology Contd.

Exploratory Data Analysis

- Evaluation of SpaceX datasets using SQL queries
- Making use of Pandas and Matplotlib libraries to determine relationship between patterns and variables

Interactive Visuals

- Using folium for Geospatial Analysis
- Using plotly to prepare interactive Dashboards

Data Model Evaluation

- Plotting confusion matrices for each classification model
- Assessing the accuracy of each classification model
- Using sci-kit learn library to train models

A photograph of a rocket launch at night. The rocket is ascending vertically, leaving a massive, bright orange and yellow plume of fire and white smoke behind it. The launch is taking place on a dark, silhouetted landscape, possibly a coastal area, with the launch site visible in the lower left. The sky is dark, making the bright launch plume stand out prominently.

Data Collection – An overview

Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.

- Make a GET response to the SpaceX REST API
- Convert response into .json file then fitting data to a Pandas DataFrame
- Using custom logic to clean the data
- Define lists to store data sets
- Calling custom functions to retrieve data and fill the lists
- Use defined lists as values in a dictionary and construct the dataset
- Create a Pandas DataFrame from the constructed dictionary dataset
- Filter the DataFrame to only include Falcon 9 launches
- Reset the FlightNumber column
- Replace missing values of PayloadMass with the mean PayloadMass value

Github link:

https://github.com/pooja420/IBM_Data_Science_Capstone.git

Data Collection – Webscraping

Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia to fetch details like Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

- Request the HTML page from the static URL
- Assign the response to an object
- Create a BeautifulSoup object from the HTML response object
- Find all tables within the HTML page
- Collect all column header names from the tables found within the HTML page
- Use the column names as keys in a dictionary
- Use custom functions and logic to parse all launch tables (see Appendix) to fill the dictionary values
- Convert the dictionary to a Pandas DataFrame ready for export

Github link:

https://github.com/pooja420/IBM_Data_Science_Capstone.git



A rocket is shown launching from the water, creating a large, billowing plume of white smoke and orange fire. The rocket itself is a slender, vertical structure with a dark tip. The launch is taking place at night, with the dark water of the ocean in the foreground reflecting the light from the rocket's engines. The background is a solid black, making the bright launch stand out.

Data Wrangling

Perform some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models. Convert below outcomes into Training Labels:

True Ocean - means the mission outcome was successfully landed to a specific region of the ocean

False Ocean - means the mission outcome was unsuccessfully landed to a specific region of the ocean

True RTLS - means the mission outcome was successfully landed to a ground pad

False RTLS - means the mission outcome was unsuccessfully landed to a ground pad

True ASDS - means the mission outcome was successfully landed on a drone ship

False ASDS - means the mission outcome was unsuccessfully landed on a drone ship

Github link:

https://github.com/pooja420/IBM_Data_Science_Capstone.git

Exploratory data analysis (EDA) – SQL

- The SQL queries performed on the data set were used to:
- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string '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 on a ground pad was achieved
- List the names of the boosters which had success on a drone ship and a payload mass between 4000 and 6000 kg
- List the total number of successful and failed mission outcomes
- List the names of the booster versions which have carried the maximum payload mass
- List the failed landing outcomes on drone ships, their booster versions, and launch site names for 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

Github link:

https://github.com/pooja420/IBM_Data_Science_Capstone.git



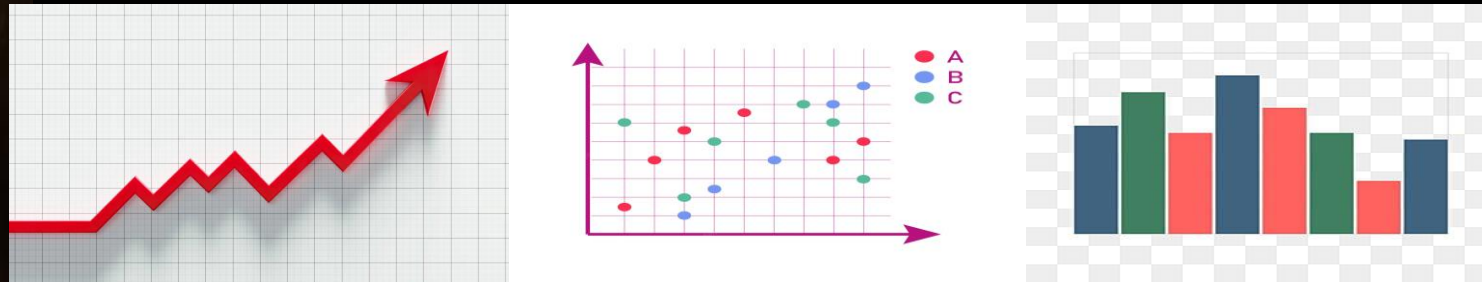
Exploratory data analysis (EDA) – Visual

Use Scatter charts, Bar and Line charts to predict if the Falcon 9 first stage will have land successful or not.

Scatter charts are useful to observe relationships, or correlations, between two numeric variables. Here it was used to visualize the relationship between Flight Number and Launch Site, Payload and Launch Site.

Bar charts are used to compare a numerical value to a categorical variable. Horizontal or vertical bar charts can be used, depending on the size of the data. It was used to visualize relationship between success rate of each orbit type.

Line charts contain numerical values on both axes, and are generally used to show the change of a variable over time. It was used to identify success trends.



Github link:

https://github.com/pooja420/IBM_Data_Science_Capstone.git

Geospatial analysis – folium

Mark all launch sites on map using NASA co-ordinates, initialize the map afterwards add folium circle and marker for each provided launch-site.

Identify and mark success and failure launch-sites by providing dummy variables like 1 for success and 0 for failure, creating clusters, creating icon as text label, and assigning icon colour.

Calculating distance between a launch site and its proximities by exploring launch sites and calculating distance between latitude and longitude, creating folium marker to show the distance and adding these features to map.



Github link:

https://github.com/pooja420/IBM_Data_Science_Capstone.git

Interactive Dash using PLOTLY

The following plots were added to a Plotly Dash dashboard to have an interactive visualization of the data:

- Pie chart (`px.pie()`) showing the total successful launches per site
- This makes it clear to see which sites are most successful
- The chart could also be filtered (using a `dcc.Dropdown()` object) to see the success/failure ratio for an individual site

Scatter graph (`px.scatter()`) to show the correlation between outcome (success or not) and payload mass (kg)

- This could be filtered (using a `RangeSlider()` object) by ranges of payload masses
- It could also be filtered by booster version

Github link:

https://github.com/pooja420/IBM_Data_Science_Capstone.git



Classification Model – Predictive Analysis

The following steps were taking to develop, evaluate, and find the best performing classification model:

Model Development:

- Load dataset
- Perform necessary data transformations (standardise and pre-process)
- Split data into training and test data sets, using `train_test_split()`
- Decide which type of machine learning algorithms are most appropriate
- Create a `GridSearchCV` object and a dictionary of parameters
- Fit the object to the parameters
- Use the training data set to train the model

Model Evaluation:

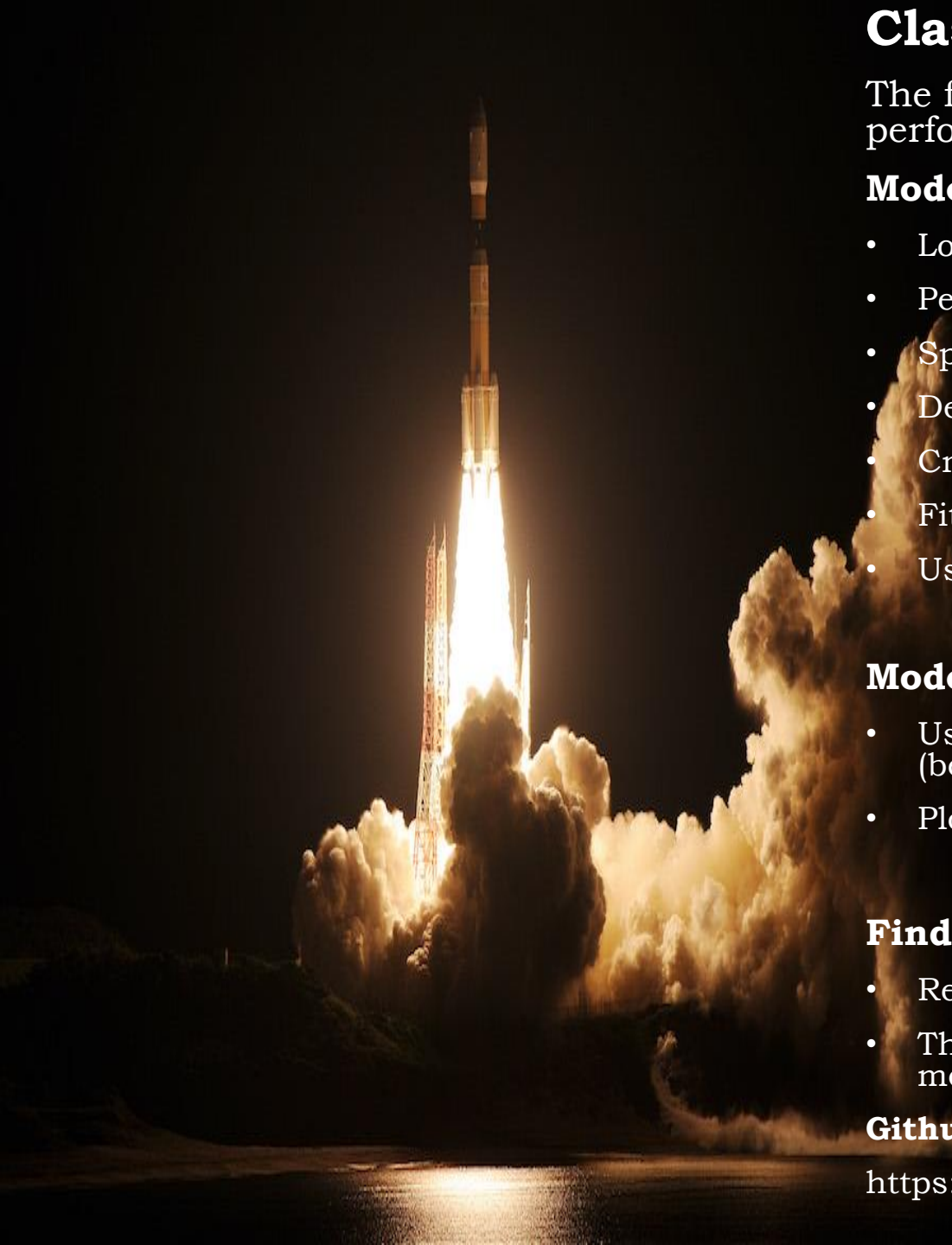
- Using the output `GridSearchCV` object, check the tuned hyperparameters (`best_params_`) and accuracy (`score` and `best_score_`)
- Plot and examine the Confusion Matrix

Finding best FIT Classification model:

- Review the accuracy scores for all chosen algorithms
- The model with the highest accuracy score is determined as the best performing model

Github link:

https://github.com/pooja420/IBM_Data_Science_Capstone.git



Results

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



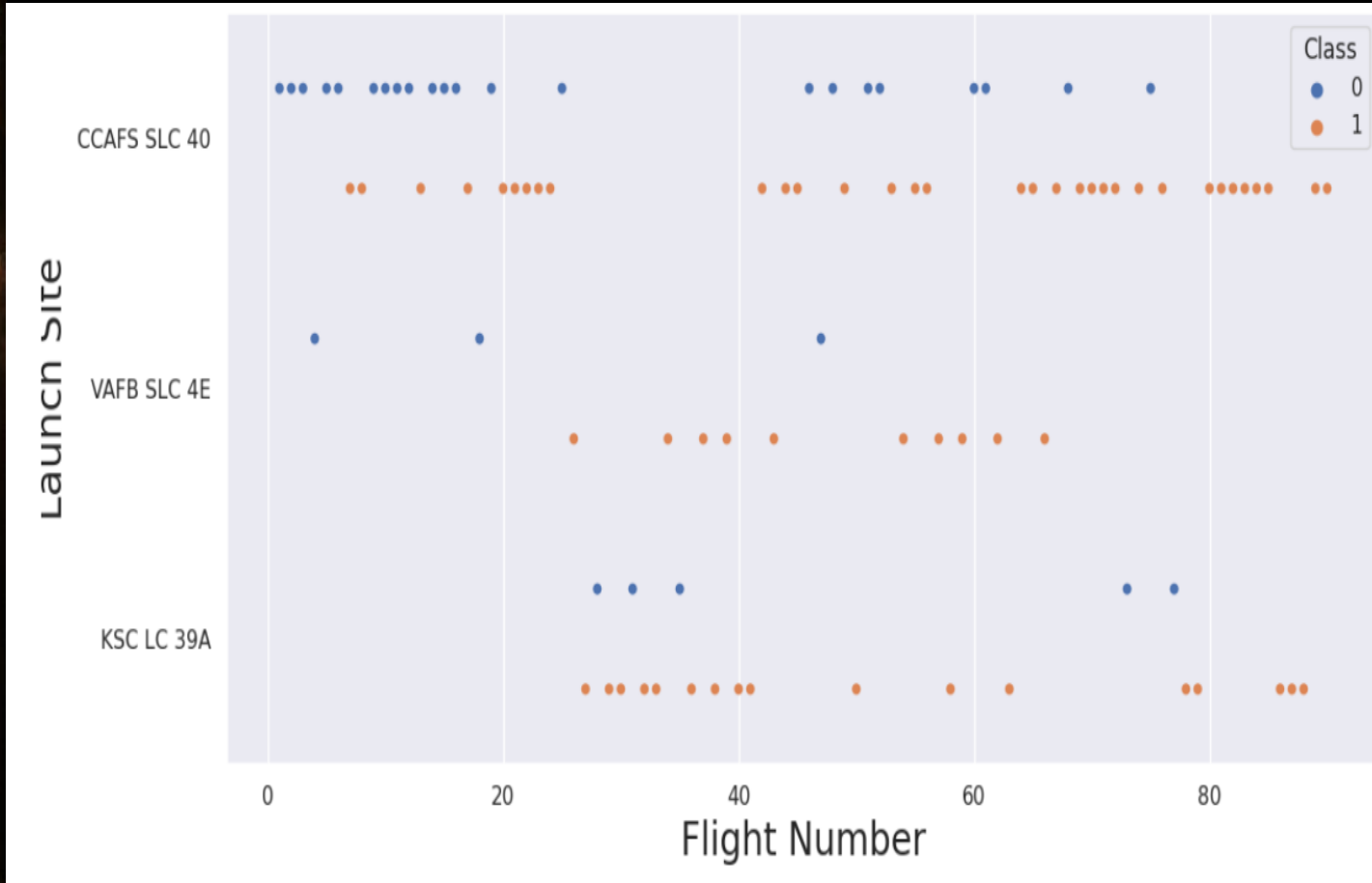


Section 2

Insights drawn from EDA

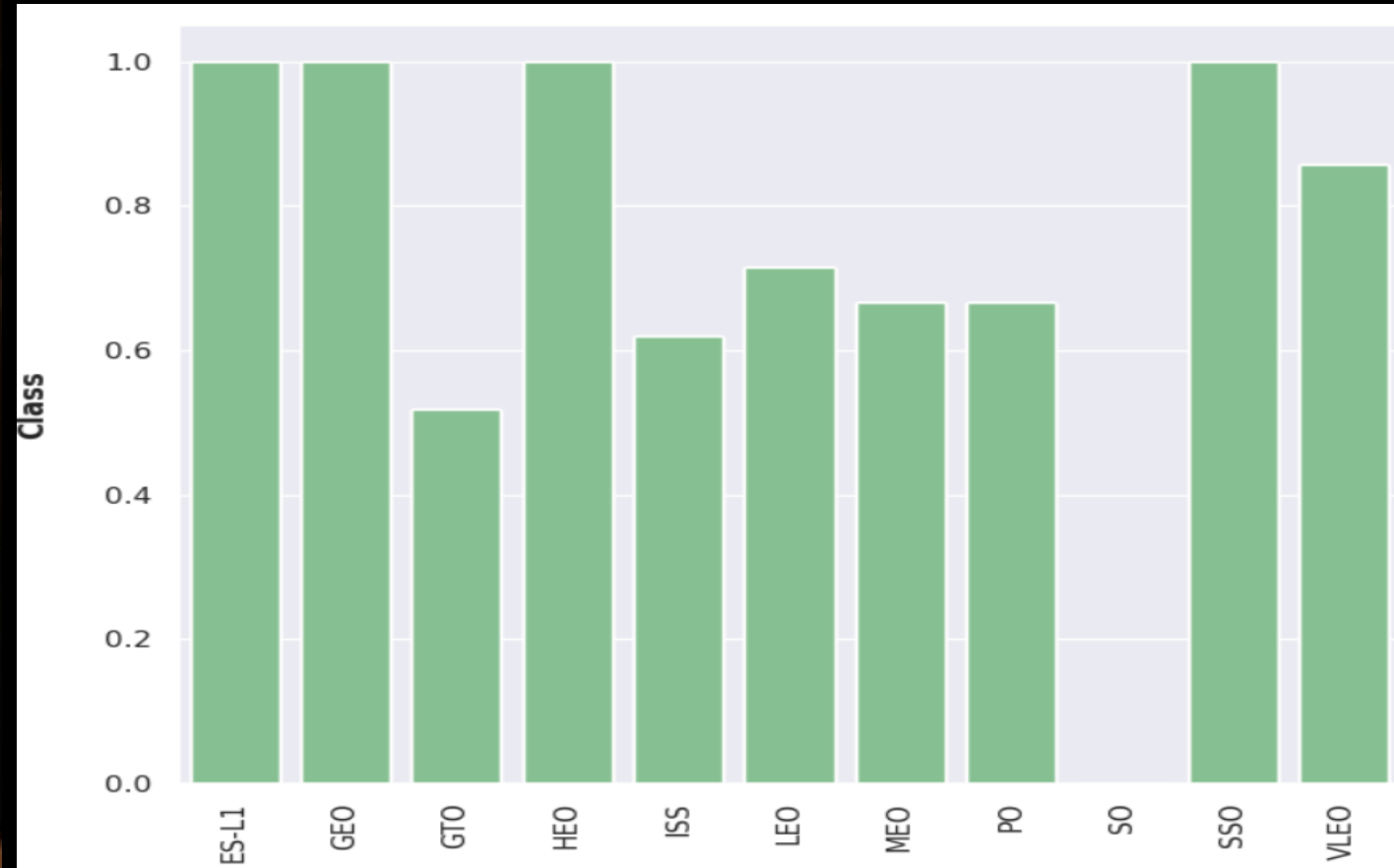
Flight Number vs. Launch Site

With the increase of flight number, the success rate is increasing as well in the launch sites



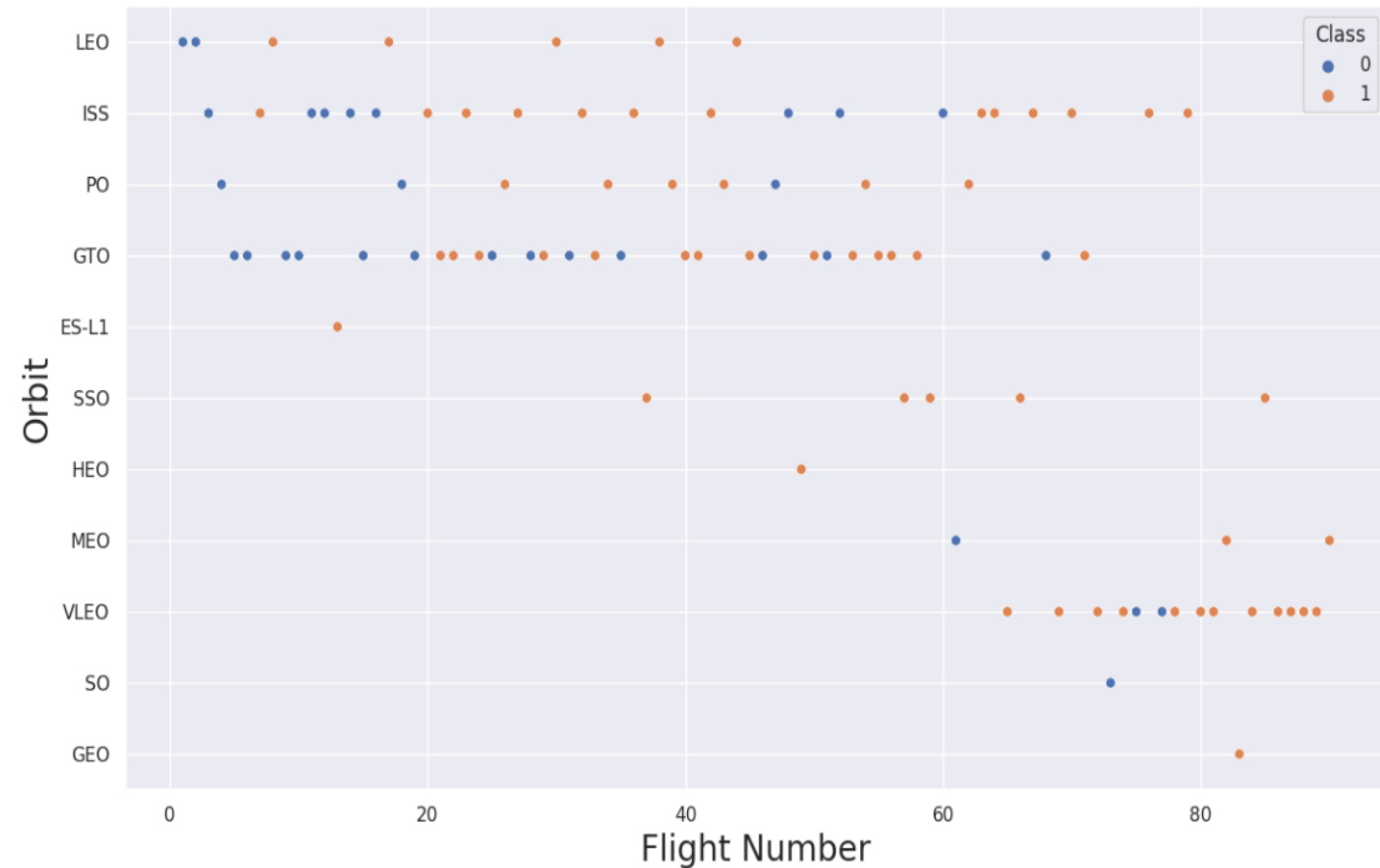
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO have a success rate of 100%
- SO has a success rate of 0%



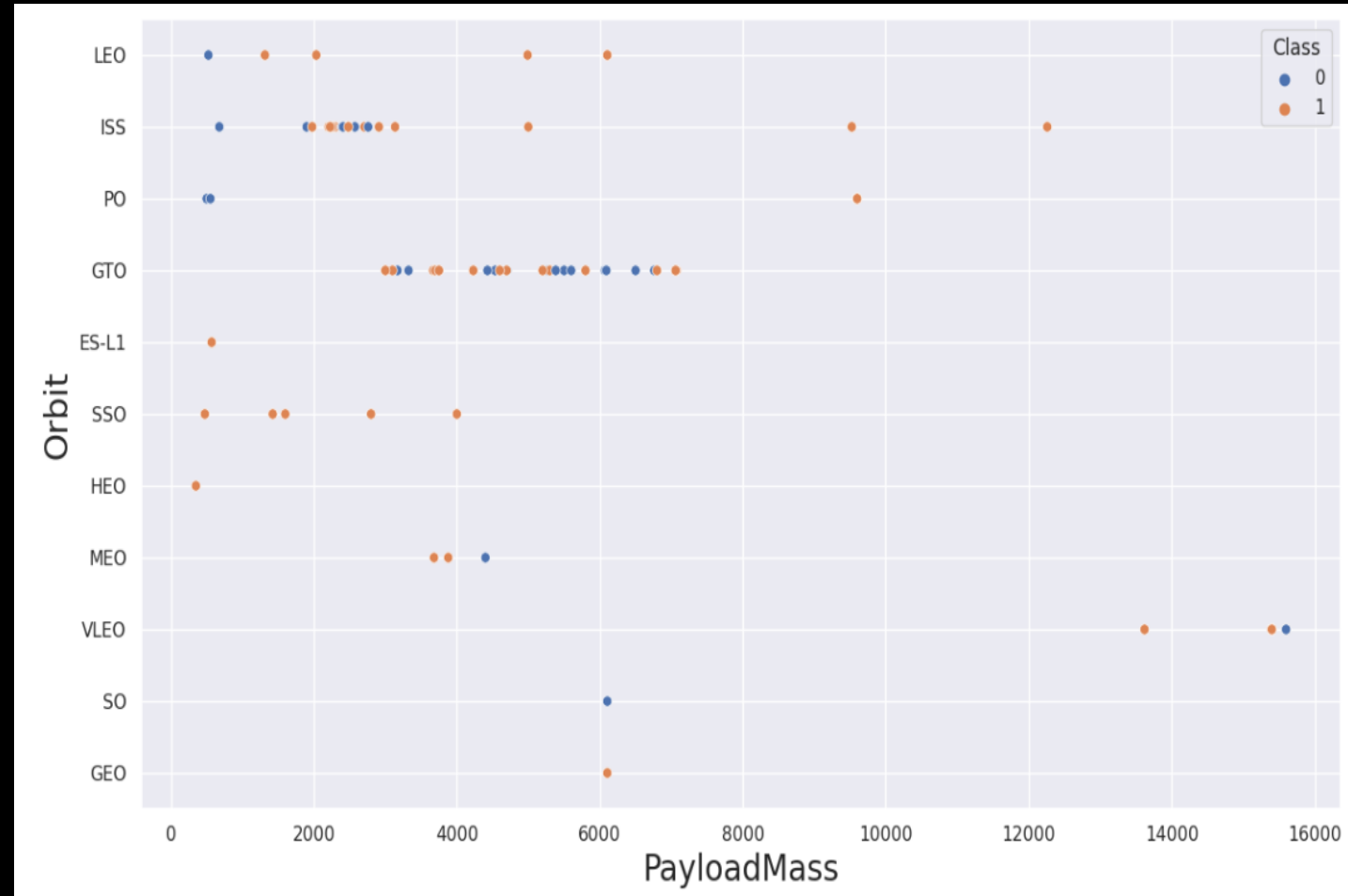
Flight Number vs. Orbit Type

It's hard to tell anything here, but we can say there is no actual relationship between flight number and GTO.



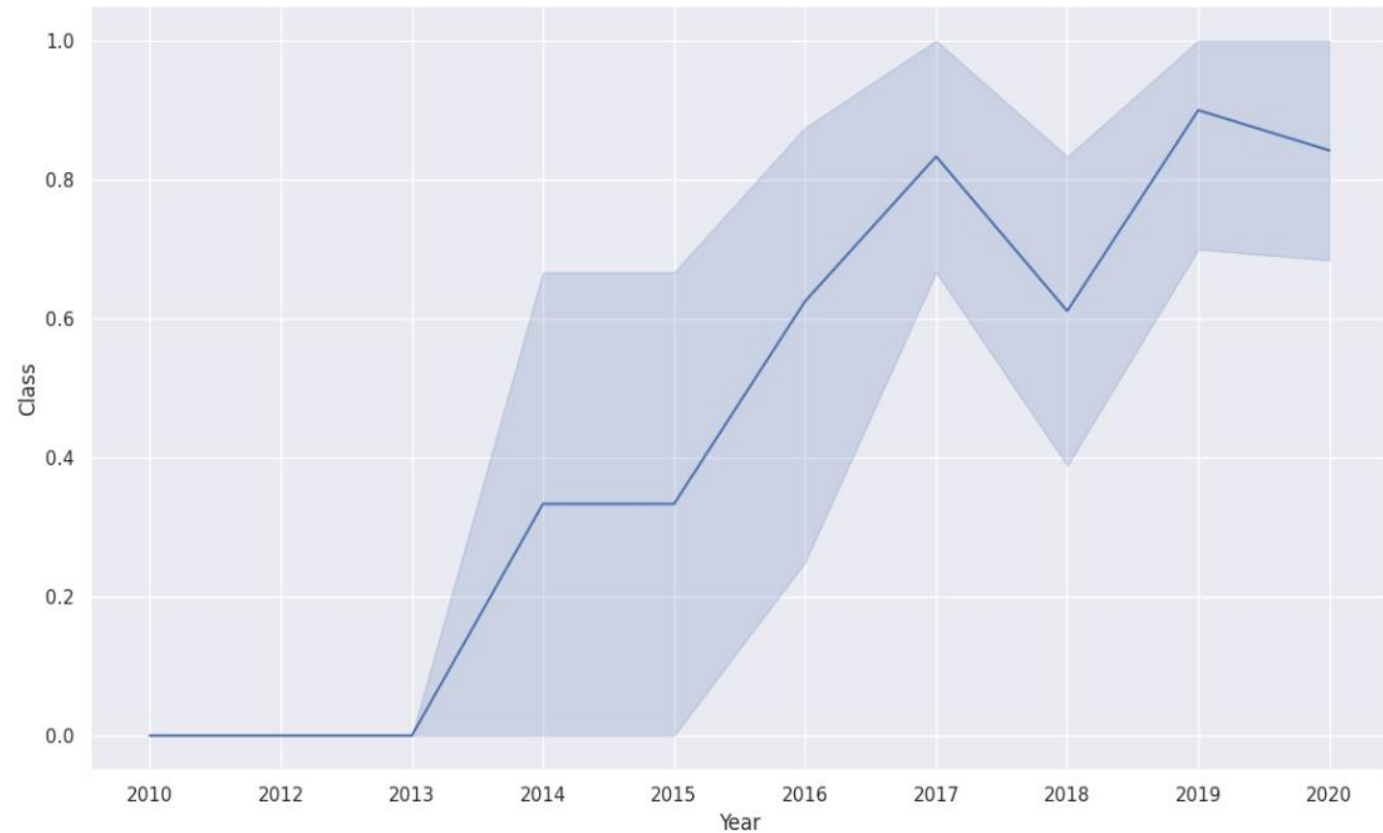
Payload vs. Orbit Type

- First thing to see is how the Pay load Mass between 2000 and 3000 is affecting ISS.
- Similarly, Pay load Mass between 3000 and 7000 is affecting GTO.



Launch Success Yearly Trend

Since the year 2013, there was a massive increase in success rate. However, it dropped little in 2018 but later it got stronger than before.



All Launch Site Names

We can get the unique values by using “DISTINCT”

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL
```

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





Name of 5 Launch Site Names starting with “CCA”

We can get only 5 rows by using “LIMIT”

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

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

Calculate – TOTAL PAYLOAD MASS

We can get the sum of all values by using “SUM”

```
%sql select sum(payload_mass__kg_) as sum from SPACEXTBL where customer like 'NASA (CRS)'
```

sum

45596





Calculate – Average payload mass carried by booster version F9 v1.1

We can get the average of all values by using “AVG”

```
%sql select avg(payload_mass__kg_) as Average from SPACEXTBL where booster_version like 'F9 v1.1'
```

Average

2534.6666666666665



To Find out First Successful Ground Landing Date

We can get the first successful data by using “MIN”, because first date is same with the minimum date

```
%sql select min(date) as Date from SPACEXTBL where mission_outcome like 'Success'
```

Date
01-03-2013

To Find out Successful Drone Ship Landing with Payload between 4000 and 6000

The payload mass data was taken between 4000 and 6000 only, and the landing outcome was determined to be “success drone ship”

```
%sql SELECT * FROM SPACEXTBL WHERE mission_outcome = 'Success' and "Landing_Outcome" = 'Success (drone ship)'
and payload_mass_kg between 4000 and 6000
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06-05-2016	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
14-08-2016	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
30-03-2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
11-10-2017	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)



To Find out Total Number of Successful and Failure Mission Outcomes

We can get the number of all the successful mission by using “COUNT” and LIKE “Success%”

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXTBL GROUP by mission_outcome ORDER BY mission_outcome
```

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

To Find out Boosters Carried Maximum Payload

We can get the maximum payload masses by using “MAX”

```
%sql select "booster_version" from SPACEXTBL where payload_mass_kg_ = (select max(payload_mass_kg_) from SPACEXTBL)
```



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

To Find out 2015 Launch Records

We can get the months by using month(DATE) and in the WHERE function we assigned the year value to “2015”

```
%sql select substr(DATE, 4,2) as Month, "landing_outcome", "booster_version",  
launch_site from SPACEXTBL where substr(DATE, 7,4) like '2015' AND "landing_outcome" like 'Failure (drone ship)'
```



Month	Landing_Outcome	"booster_version"	Launch_Site
01	Failure (drone ship)	booster_version	CCAFS LC-40
04	Failure (drone ship)	booster_version	CCAFS LC-40



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

By using “ORDER” we can order the values in descending order, and with “COUNT” we can count all numbers as we did previously

```
%sql select "landing_outcome" AS LANDING_OUTCOME, count("landing_outcome") AS TOTAL_COUNT  
from SPACEXTBL where date BETWEEN '04-06-2011' and '20-03-2017' group by "landing_outcome"  
order by COUNT("landing_outcome") desc
```

LANDING_OUTCOME	TOTAL_COUNT
Success	19
No attempt	9
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
No attempt	1
Failure (parachute)	1



Section 3

Launch Sites Proximities Analysis

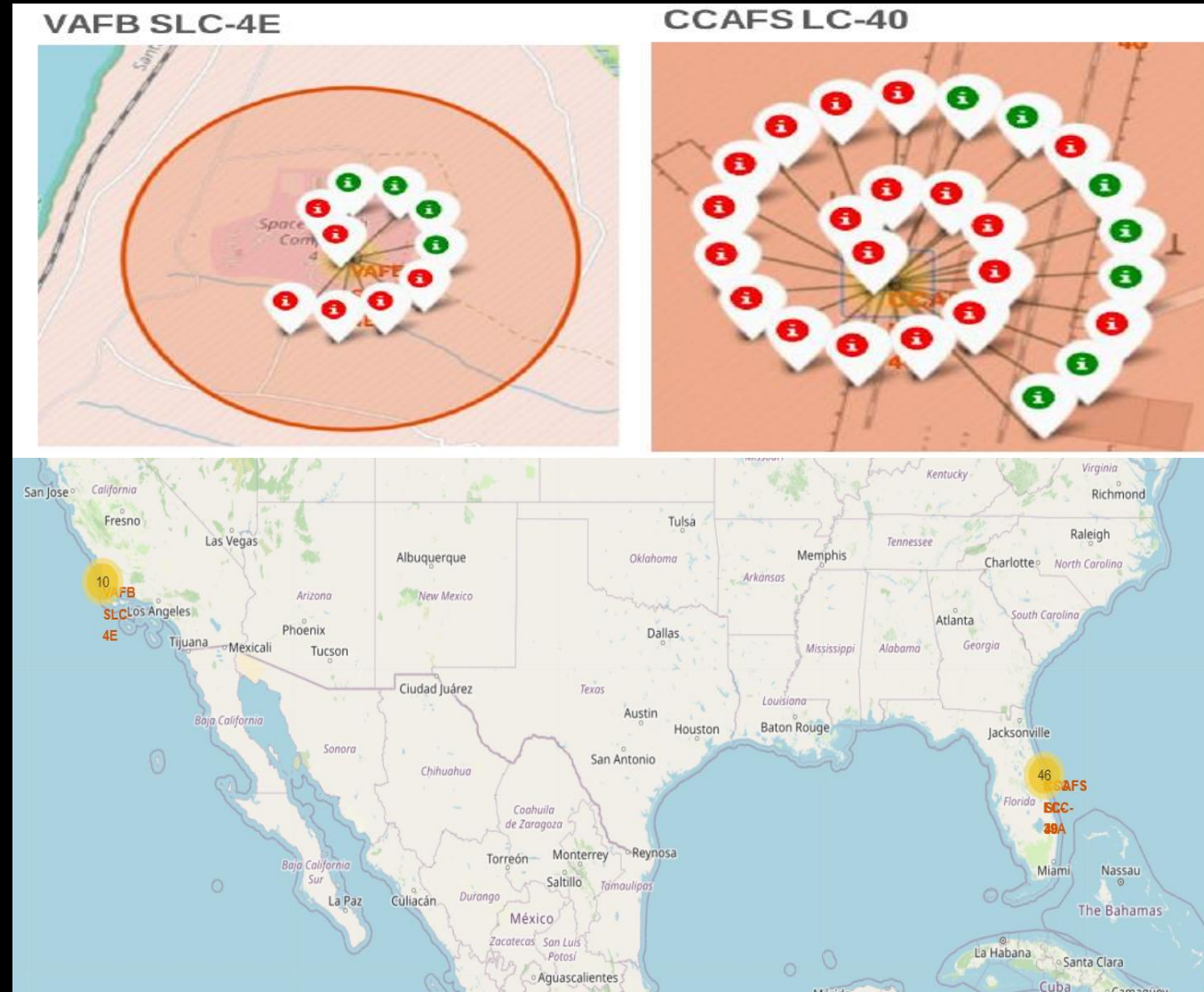
All Launch Sites' Location Markers

All the launches are near USA, Florida, and California



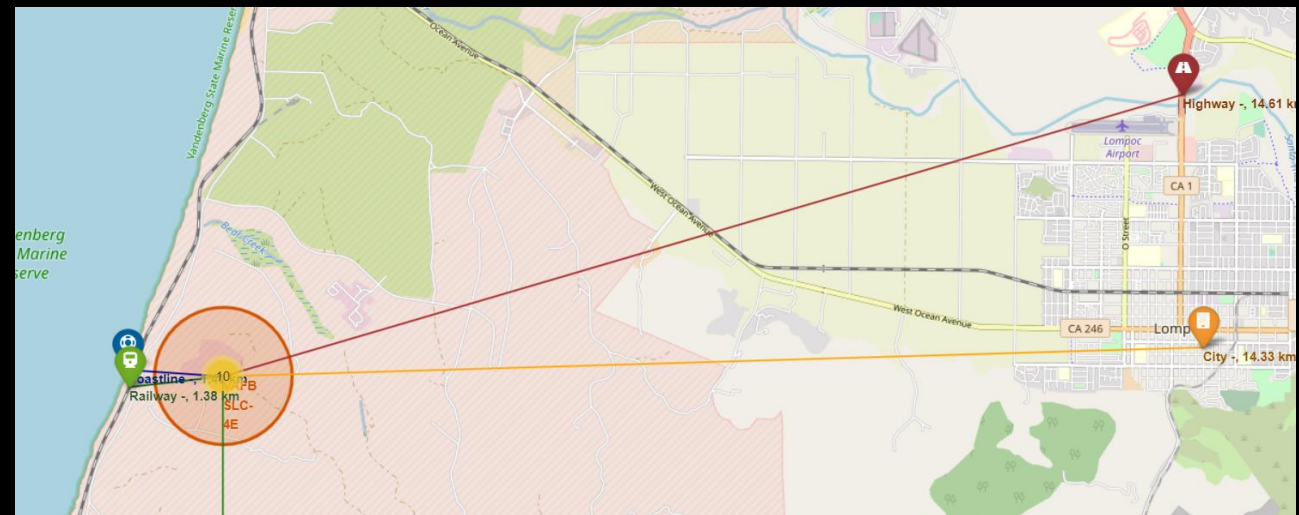
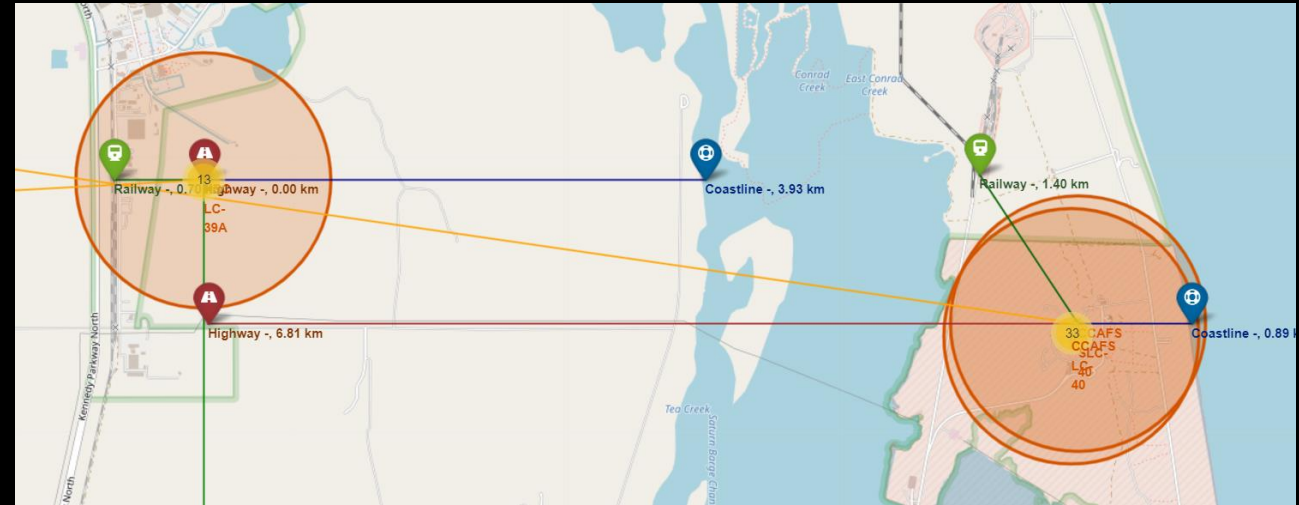
Color-labeled Launch Outcomes

- Green means successful
- Red means Failure



Launch Sites to its Proximities

All distances from launch sites to its proximities, they weren't far from railway tracks.



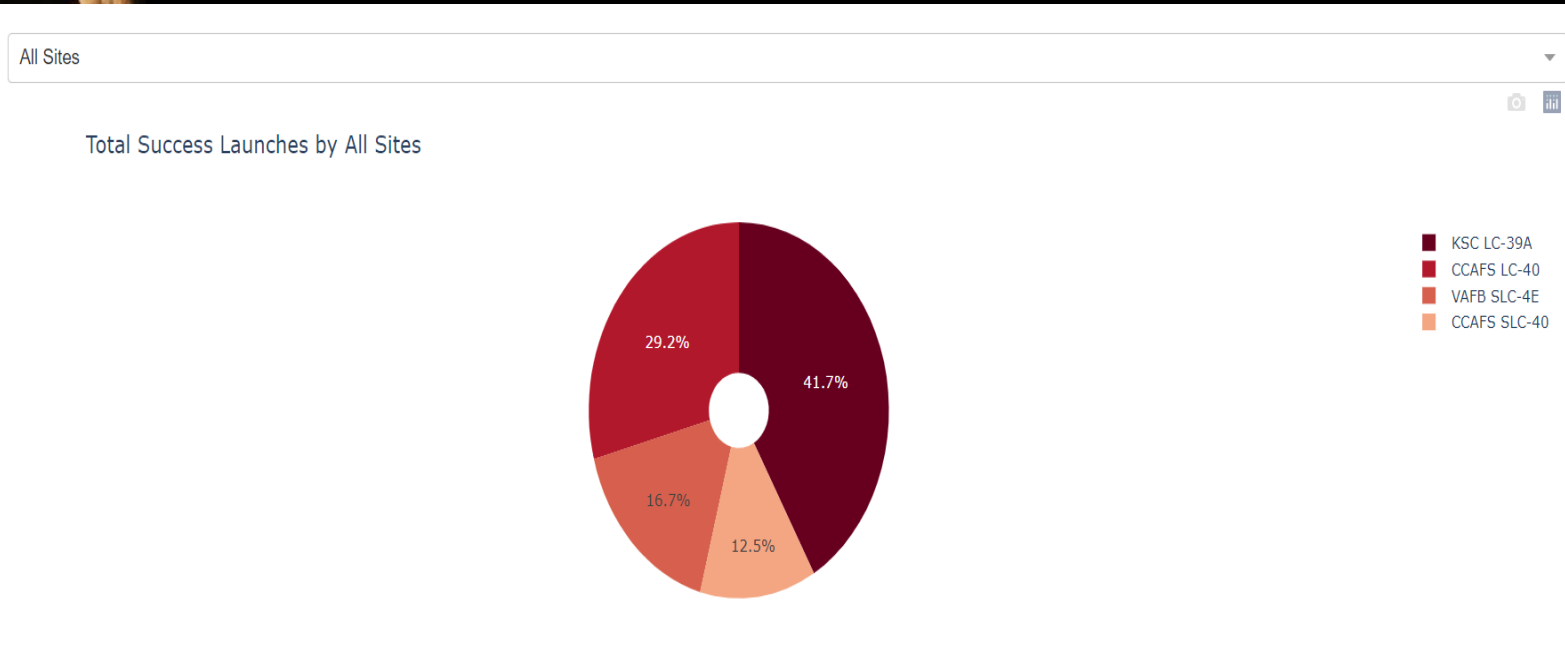


Section 4

Interactive Dashboard with Plotly

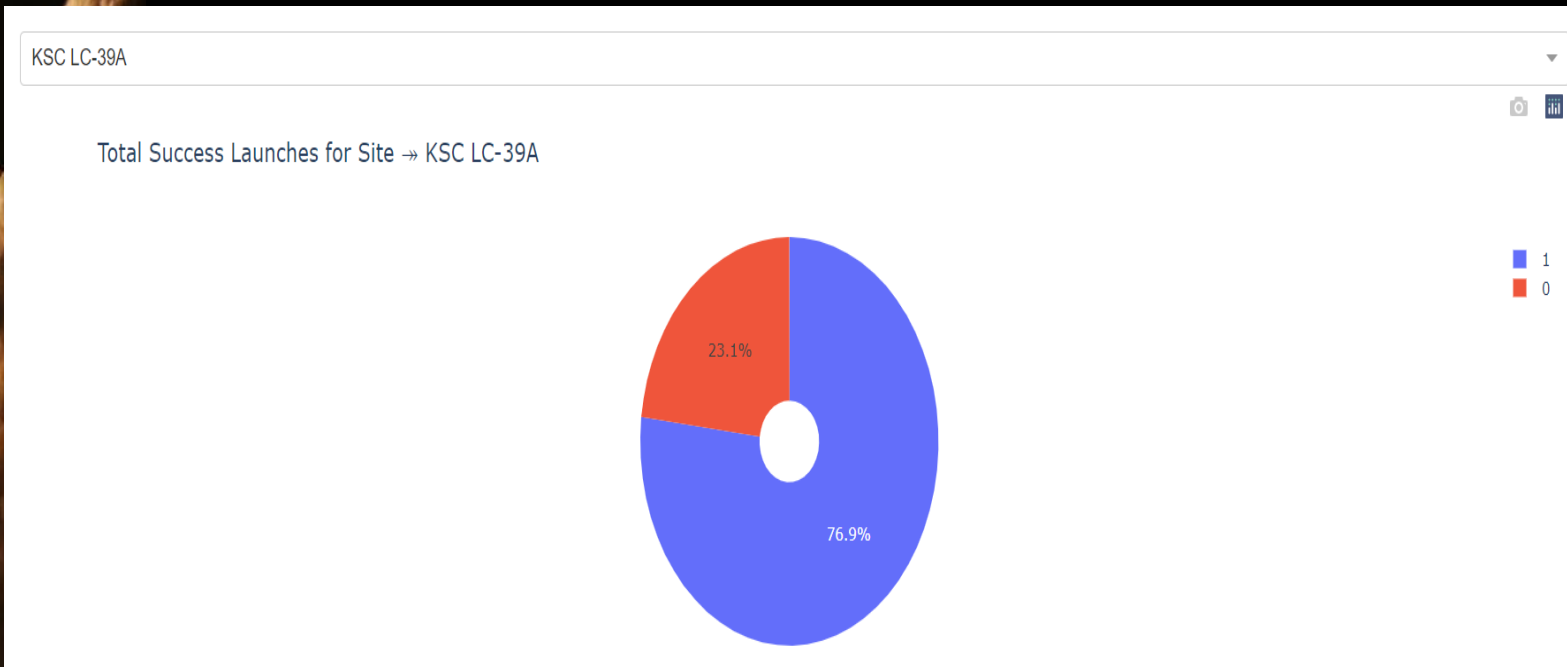
Launch Success Count

- KSC LC-39A has the highest success score with 41.7%
- CCAFS LC-40 comes next with 29.2%
- Finally, VAFB SLC-4E and CCAFS SLC-40 with 16.7% and 12.5% respectively



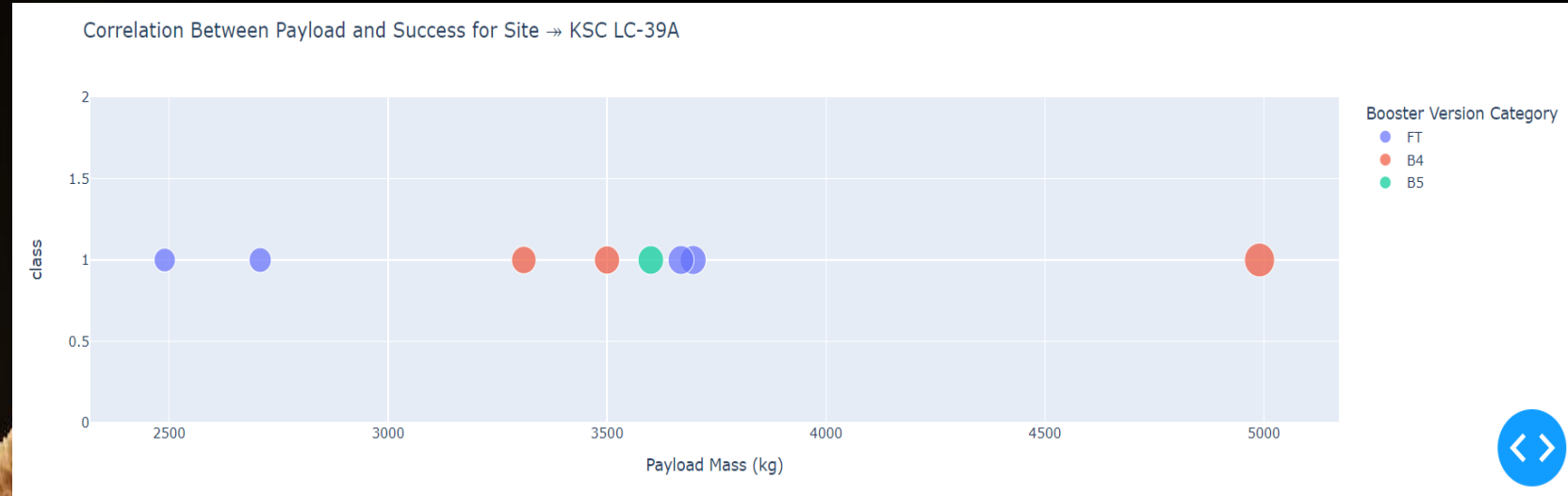
Launch Site with Highest Score

KSC LC-39A has the highest score with 76.9% with payload range of 2000 kg – 10000 kg, and FT booster version has the highest score

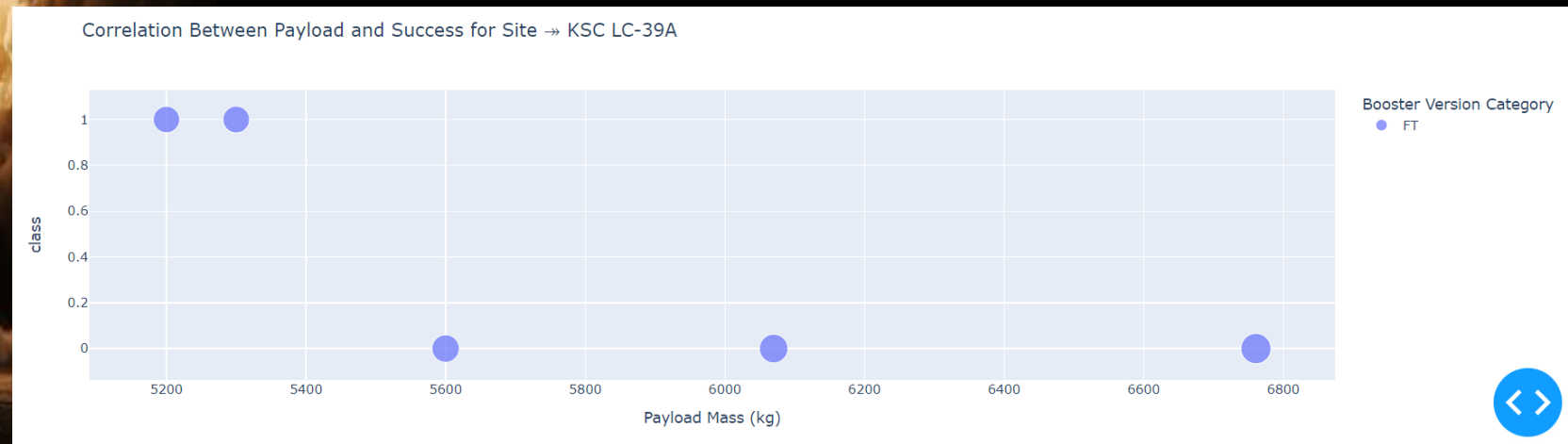


Launch Outcome V/s Payload

Payload 0 kg – 5000 kg (first half)



Payload 6000 kg – 10000 kg (second half)



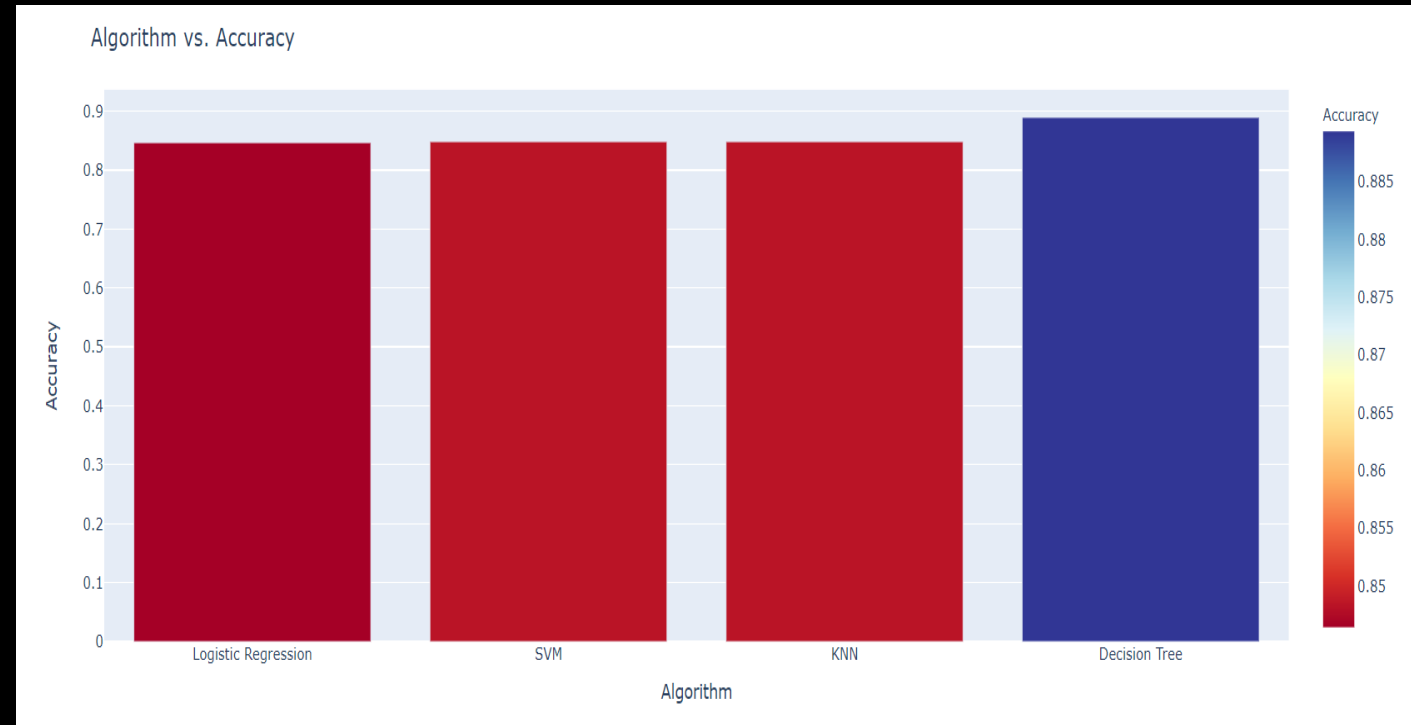


Section 5

Predictive Analysis - Classification

Accuracy of Classification

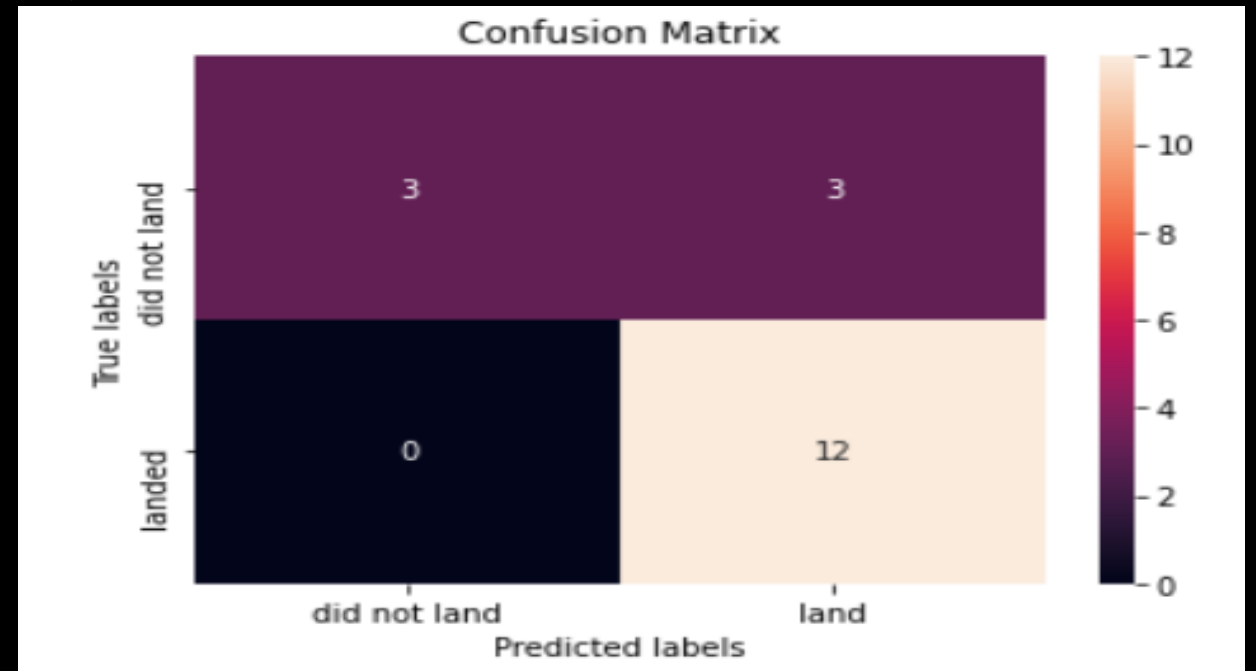
Decision Tree has the highest accuracy with almost 0.89, then comes the remaining models with almost same accuracy of 0.84





Confusion Matrix

- Sensitivity = 1.00, formula: $TPR = TP / (TP + FN)$
- Specificity = 0.50, formula: $SPC = TN / (FP + TN)$
- Precision = 0.80, formula: $PPV = TP / (TP + FP)$
- Accuracy = 0.83, formula: $ACC = (TP + TN) / (P + N)$
- F1 Score = 0.89, formula: $F1 = 2TP / (2TP + FP + FN)$
- False Positive Rate = 0.50, formula: $FPR = FP / (FP + TN)$
- False Discovery Rate = 0.20, formula: $FDR = FP / (FP + TP)$





Conclusions Drawn

- We found the site with highest score which was KSC LC-39A
- The payload of 0 kg to 5000 kg was more diverse than 6000 kg to 10000 kg
- Decision Tree was the optimal model with accuracy of almost 0.89
- We calculated the launch sites distance to its proximities

Appendix



Relevant codes are uploaded in below provided Github link:

https://github.com/pooja420/IBM_Data_Science_Capstone.git