IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

Pablo Bagano

OUTLINE

EXECUTIVE SUMMARY

- The data was collected from the public SpaceX API and SpaceX Wikipedia page. A new column called 'class' was created to classify successful landings. The data was explored using SQL queries, visualizations, folium maps, and dashboards. Relevant columns were gathered to be used as features in machine learning models. All categorical variables were converted to binary using one hot encoding. The data was standardized and GridSearchCV was used to find the best parameters for the machine learning models. Finally, the accuracy score of all models was visualized.

- We created four machine learning models: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. Turns out the Decision Tree Classifier did better than the others, scoring around 88% accuracy. The other models were all about the same, with a score of roughly 83.33%. But, they all had a tendency to overestimate successful landings. So, we're thinking we need more data to get a more accurate prediction.

Introduction

- Commercial space industry is expanding. Rapidly.
- Space X leads the market and our project aims to determine the cost of SpaceX launches and predict first-stage reusability to help Space Y compete with SpaceX.
- We will gather information and create dashboards to analyze SpaceX launches and train a machine learning model to predict first-stage reusability using public information.
- The goal is to help Space Y develop a competitive pricing strategy and potentially revolutionize the commercial space industry by predicting the reusability of rocket components.

METHODOLOGY

- Data collection methodology:
  - The data was gathered from the SpaceX REST API targeting a specific endpoint of the API using the ‘requests' library
  - The result of the data gathering will be a list of JSON objects which will later be converted to a Dataframe
- Perform data wrangling:
  - The data wrangling involved reviewing and selecting the most relevant attributes such as Flight Number, Date, Payload mass and Launch Site. The Outcome column was converted to binary values (0 an 1) representing, respectively, unsuccessful and successful landing outcomes. This process was crucial for building a classification model that could predict landing outcomes.

METHODOLOGY

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models:

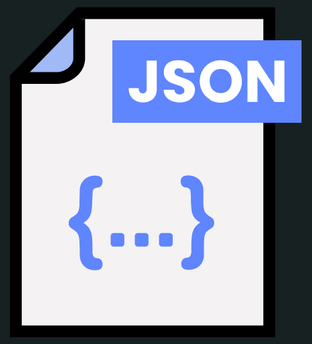
  - The predictive analysis was built using a machine learning pipeline, which reprocessed the data and split it into training and testing sets. The model was trained and tuned using GridSearch, in order to find the optimal hyperparameters for Logistic Regression, Support Vector Machines, Decision Tree Classifier and K-nearest neighbors. The best model was determined using the training data and the evaluation was done by outputting the Confusion Matrix.

DATA COLLECTION

HOW IT WAS DONE

- The Data Collection process involved two fundamental steps:
  - API requests from Space X public API using Python library 'requests'
  - Web Scraping technique using Python library 'BeautifulSoup' on Space X's Wikipedia page
- API Data:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Webscraping Data:
  - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time
- Next, you will see a Flowchart of both processes

API DATA COLLECTION



DATAFRAME CREATION

DATA WRANGLING

LINK

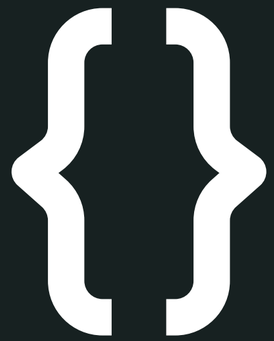
WEB SCRAPING

WIKIPEDIA HTML

BEAUTIFULSOUP

DATA ITERATION

DICTIONARY CREATION



LINK

DATA WRANGLING

- IN ORDER TO DIFFERENTIATE SUCCESSFUL LAUNCHES FROM UNSUCCESSFUL ONES, THE NUMBER 1 AND 0 WERE ASSING TO EACH, RESPECTIVELY

- SUCCESSFUL LAUNCHES (1):
  - True ASDS, True RTLS, & True Ocean

- UNSUCCESSFUL LAUNCHES (0):
  - None None, False ASDS, None ASDS, False Ocean, False RTLS

LINK

EXPLORATORY DATA ANALYSIS

EDA WITH VISUALIZATION

- Chosen plots:
  - Scatterplots
  - Line charts
  - Bar charts

- Scatterplots and line charts were used to express the relation between two numerical variables

- Bar charts were used to express the count of categorical variables

LINK

EDA with SQL

- After the Dataset was loaded into the IBM DB2 Database, the queries were performed using SQL-Python integration.
- This step was fundamental for better understanding of the dataset
- Data retrieved from queries:
  - launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

LINK

FOLIUM AND DASHLY

FOLIUM

- An interactive map was created with Folium, marking Launch Sites, successful/unsucessful landings

- The map was also proven to be an important tool to explain the reasons behind the choice of each location

LINK

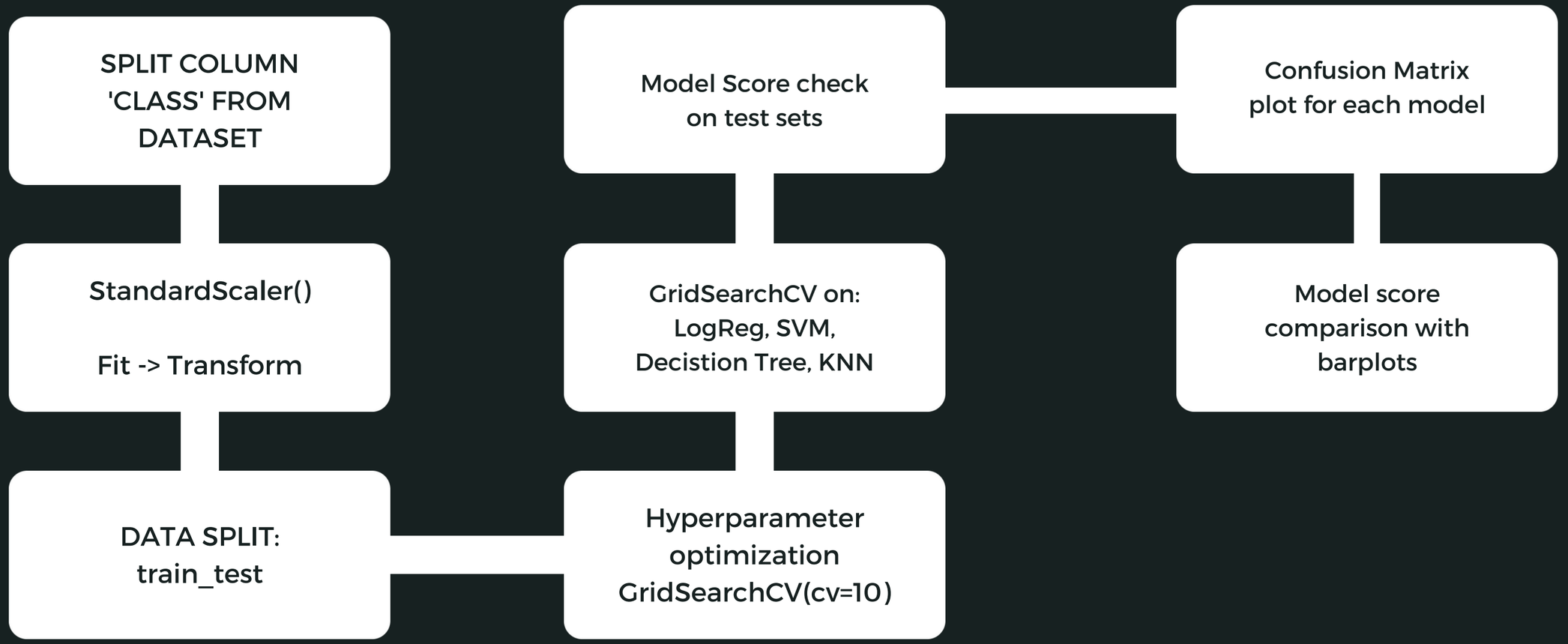
DASHLY

- This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart

- It aims to answer the following questions:
  - Which site has the largest successful launches?
  - Which site has the highest launch success rate?
  - Which payload range(s) has the highest launch success rate?
  - Which payload range(s) has the lowest launch success rate?
  - Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?

LINK

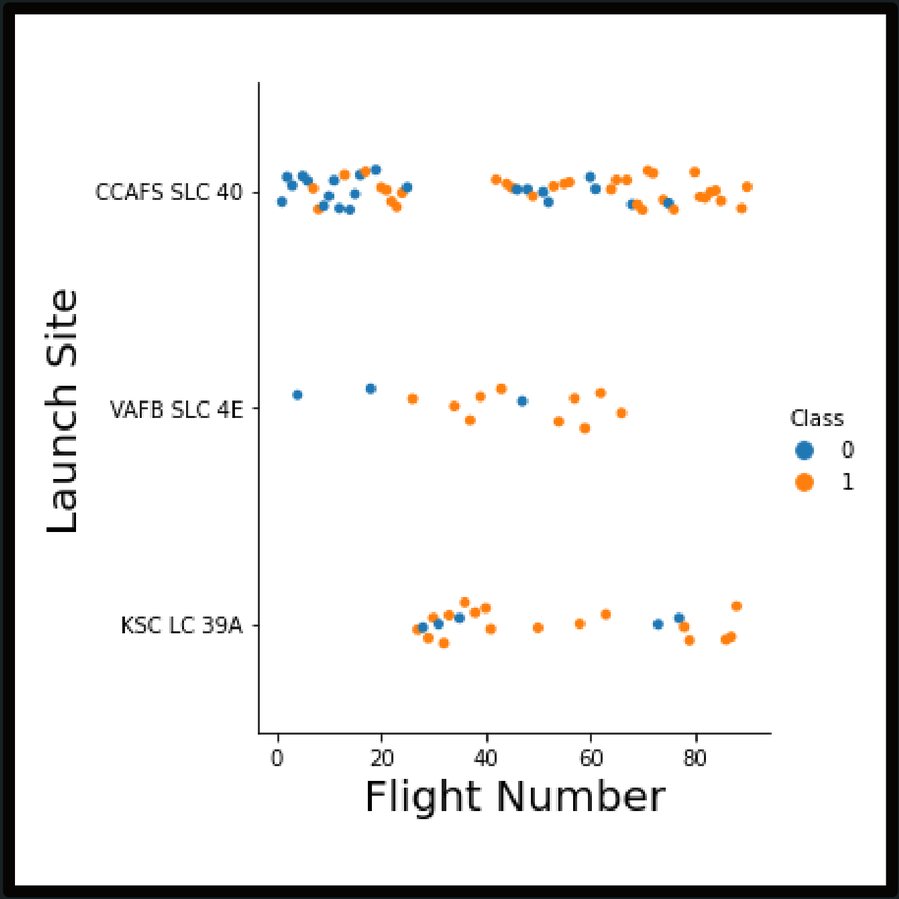
PREDICTIVE ANALYSIS



VISUALIZING THE EDA

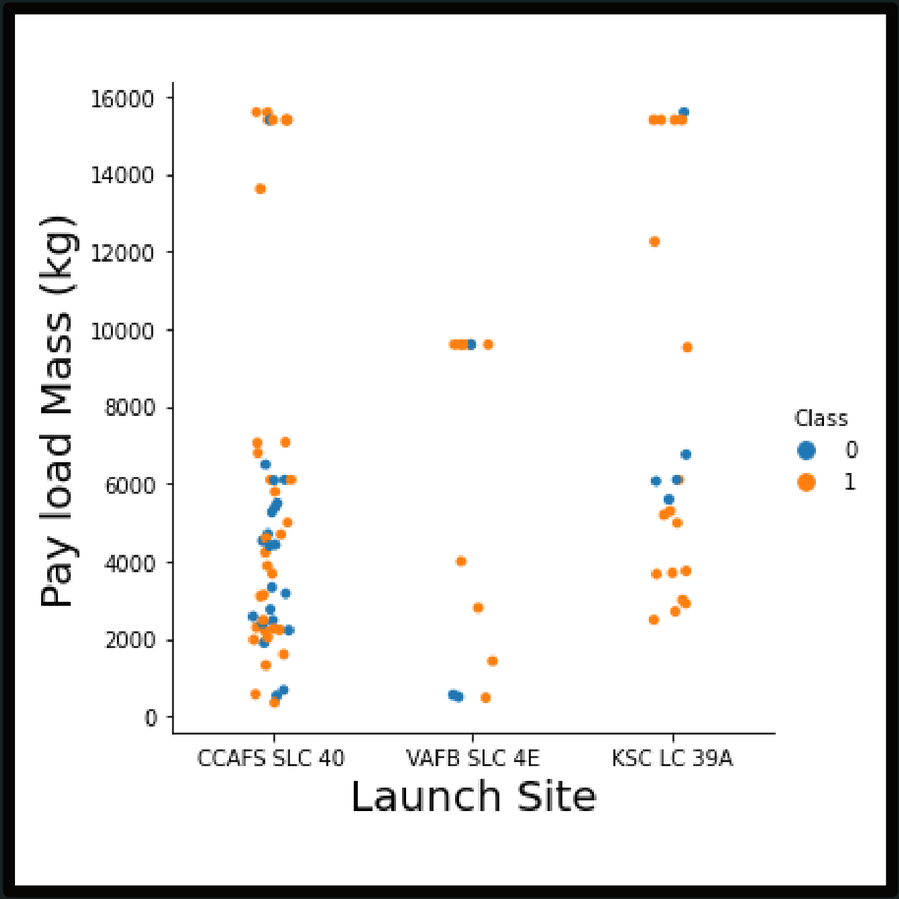
LINK

Flight Number x Launch Site



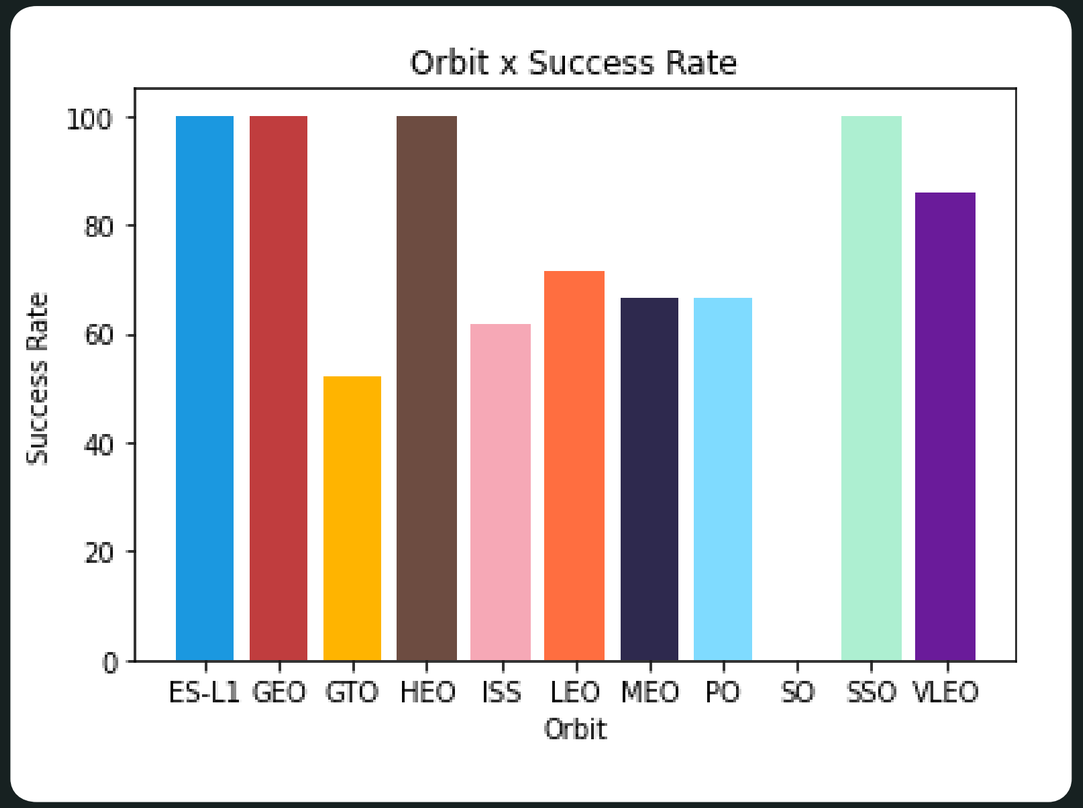
- The plot displays a clear improvement over time when it comes to the success of launches.
- While CCAF has more successful launches in percentages the other two launches seem to present better results.
- Most likely, CCAFS is the main launch base, given the volume of launches

Payload x Launch Site



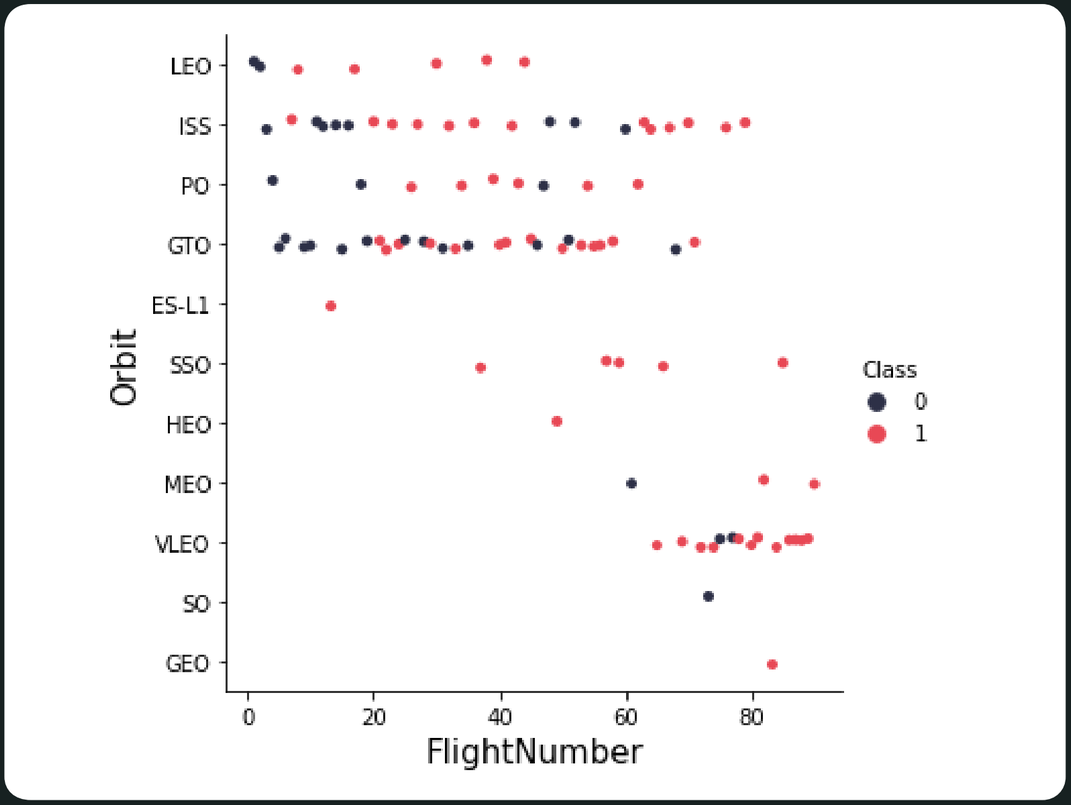
- Payload refers to any cargo or equipment carried by a rocket
- Once again, CCAFS Launch Site has the highest volume of successes.
- VAFB appears to present a threshold as for the payload of rockets launched there
- KSC has the highest rate of successful launch and can perform launches of rockets as heavy as the ones launched at CCAFS

Orbit Type x Success Rate



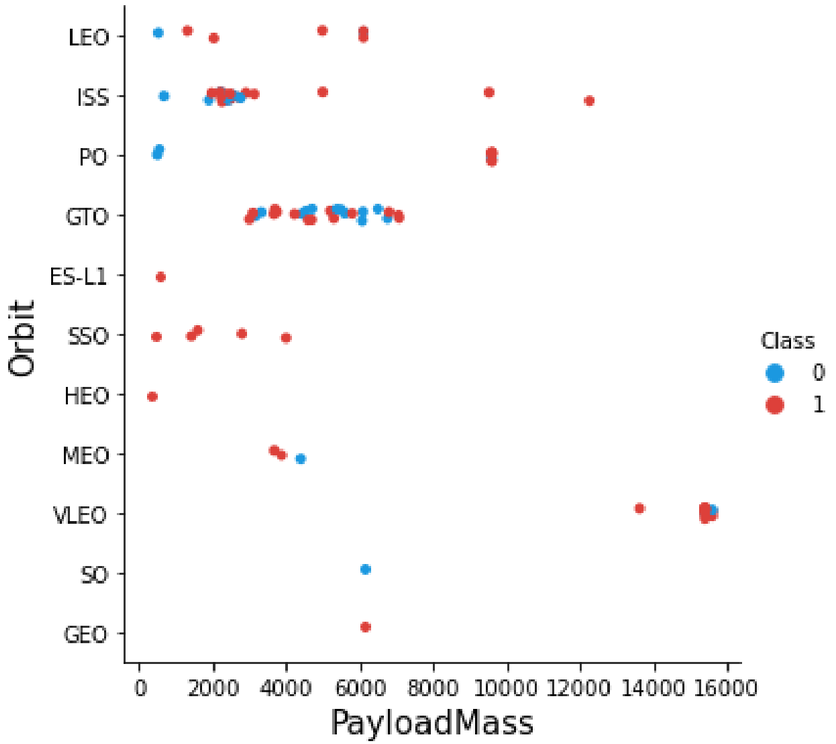
- Orbit types ES L1, GEO, HEO and SSO have around 100% success rate
- Orbyt type SO has 0% succes rate, We will se why further along the presentation
- VLEO has great success rate
- GTO presents the largest sample and 50% of success rate

Flight Number x Orbit Type



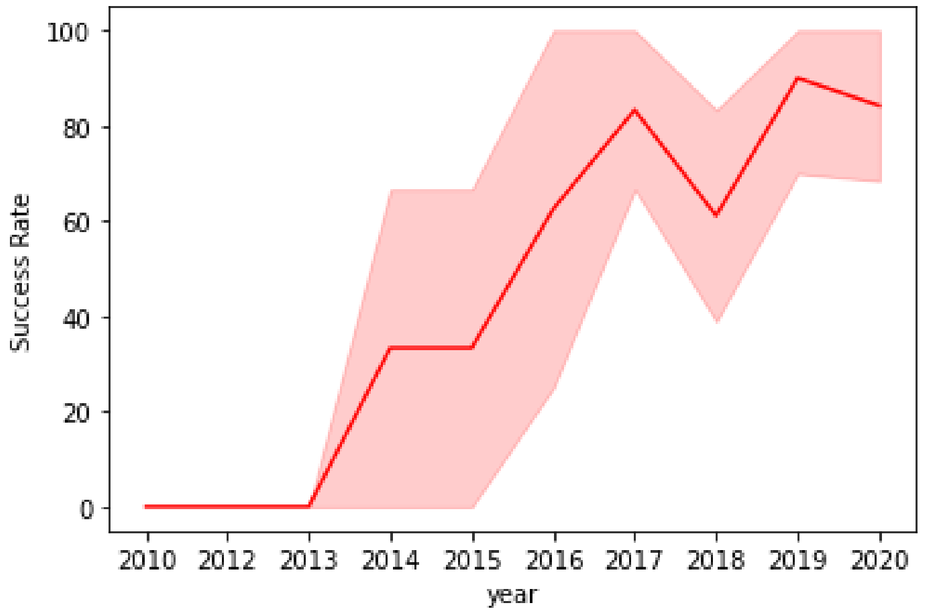
- As the number of launches increased, the preference for orbit types changed, too.
- The launch outcome also seems to play a significant role in the choice of orbit type
- SO orbit had only one launch and it failed
- VLEO is the orbit type that has show best results as for the launch outcome

Payload x Orbit Type



- LEO, SSO and ES-L1 present lower payload mass launches in general and have high rates of success, as we've seen.
- VLEO orbit ty has high percentages of success and higher payload mass values

Launch Success Yearly Trend



- Other than a slight dip in 2018, there has been a consistent improvement in the successful launches over the years
- The success rate is about 80%

EDA WITH SQL

LINK

All Launch Site Names

```
1 %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

- Four unique values were retrieve from the query.

Launch Site Names Begin with 'KSC'

```
1 %sql select * from SPACEXTBL where launch_site like 'KSC%' limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
16-03-2017	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
30-03-2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
15-05-2017	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

As you can see, these are all SpaceX missions with launch sites beginning with "KSC". The results include the flight number, mission name, launch date, and launch site for each of the 5 records returned.

Total Payload Mass

```
1 %sql select sum(payload_mass__kg_) as sum from SPACEXTBL where customer like 'NASA (CRS)'
```

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

sum
45596

- The total total payload carried by boosters from NASA is 45596 kg

Average Payload Mass by F9 v1.1

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

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

Average
2534.6666666666665

The query above calculates the average payload mass carried by booster version F9 v1.1. Rounding it up, the result is 2534.67 kg.

First Successful Ground Landing Date

```
1 %sql select min(date) as Date from SPACEXTBL where mission_outcome like 'Success'
```

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

Date
01-03-2013

- The minimum date where a mission outcome was "Success".
- Date: 01-03-2013

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version from SPACEXTBL
where (mission_outcome like 'Success')
AND (payload_mass__kg_ BETWEEN 4000 AND 6000)
AND ([Landing _Outcome] like 'Success (ground pad)')
```

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

Booster_Version
F9 FT B1032.1
F9 B4 B1040.1

- This query retrieves the names of the booster versions that have achieved a successful landing on a ground pad, and have a payload mass between 4000 and 6000 kg. The result returned are "F9 FT B1032.1" and "F9 B4 B1040.1" (both of them meet the specified criteria)

Total Number of Successful and Failure Mission Outcomes

```
1 %sql SELECT mission_outcome, count(*) as Count FROM SPACEXTBL GROUP by mission_outcome ORDER BY mission_outcome
```

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

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

Above you can the results for the query regarding the total number of successful and failure Mission Outcomes

Boosters Carried Maximum Payload

In [15]:
```
maxm = %sql select max(payload_mass__kg_) from SPACEXTBL
maxv = maxm[0][0]
%sql select booster_version from SPACEXTBL where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXTBL)
```

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

Out[15]:

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

2017 Launch Records

In [16]:

```
%sql SELECT SUBSTR(Date, 6, 2) AS Month, [Landing _Outcome], Booster_Version,
Launch_Site FROM SPACEXTBL WHERE SUBSTR(Date, 7, 4) = '2017'
AND [Landing _Outcome] LIKE 'Success (ground pad)';

```

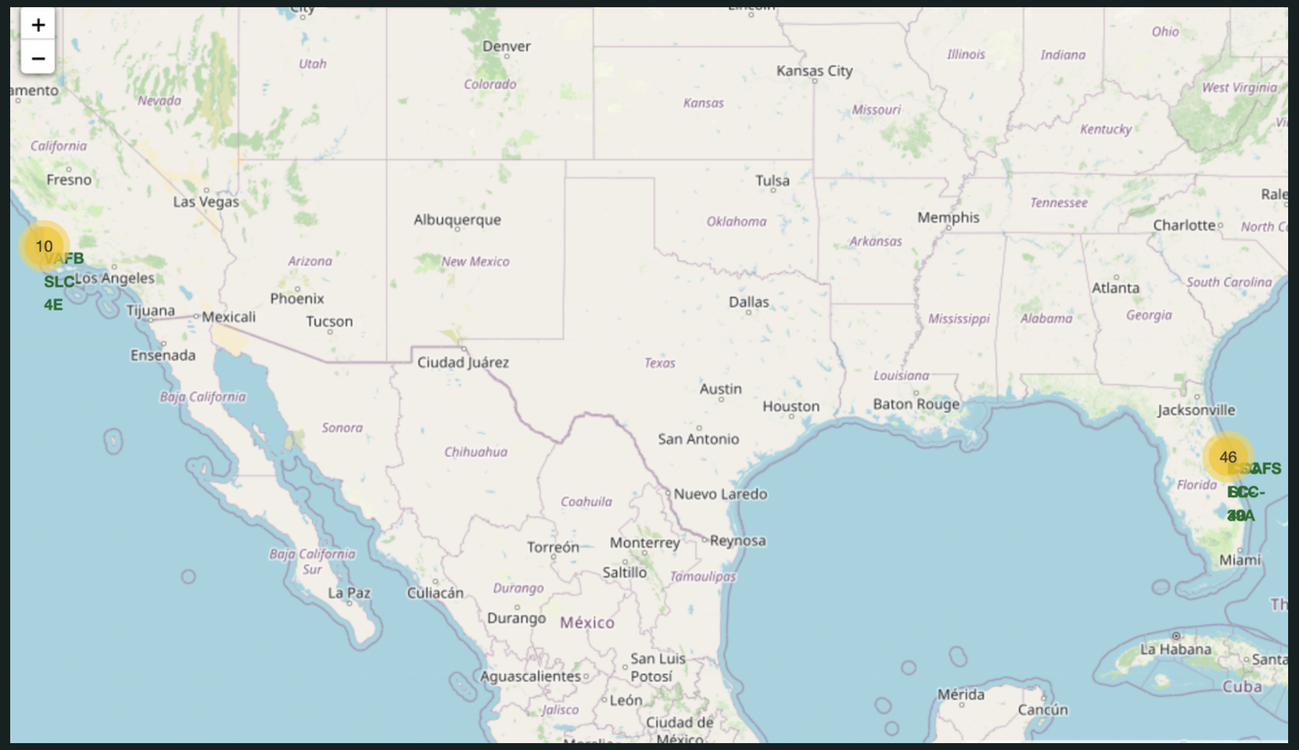
```
 * sqlite:///my_data1.db
Done.
```

Out[16]:

Month	Landing _Outcome	Booster_Version	Launch_Site
-2	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
-2	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
-2	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
-2	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
-2	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
-2	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

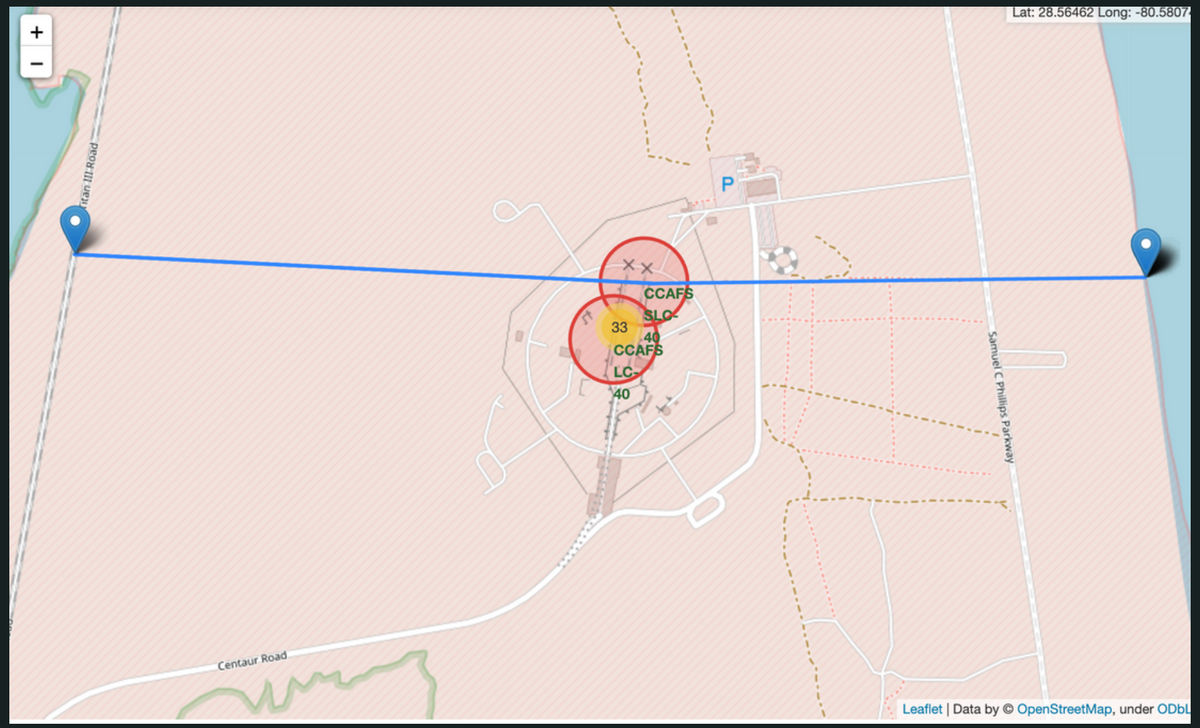
LAUNCH SITE PROXIMITIES ANALYSIS

LAUNCH SITES



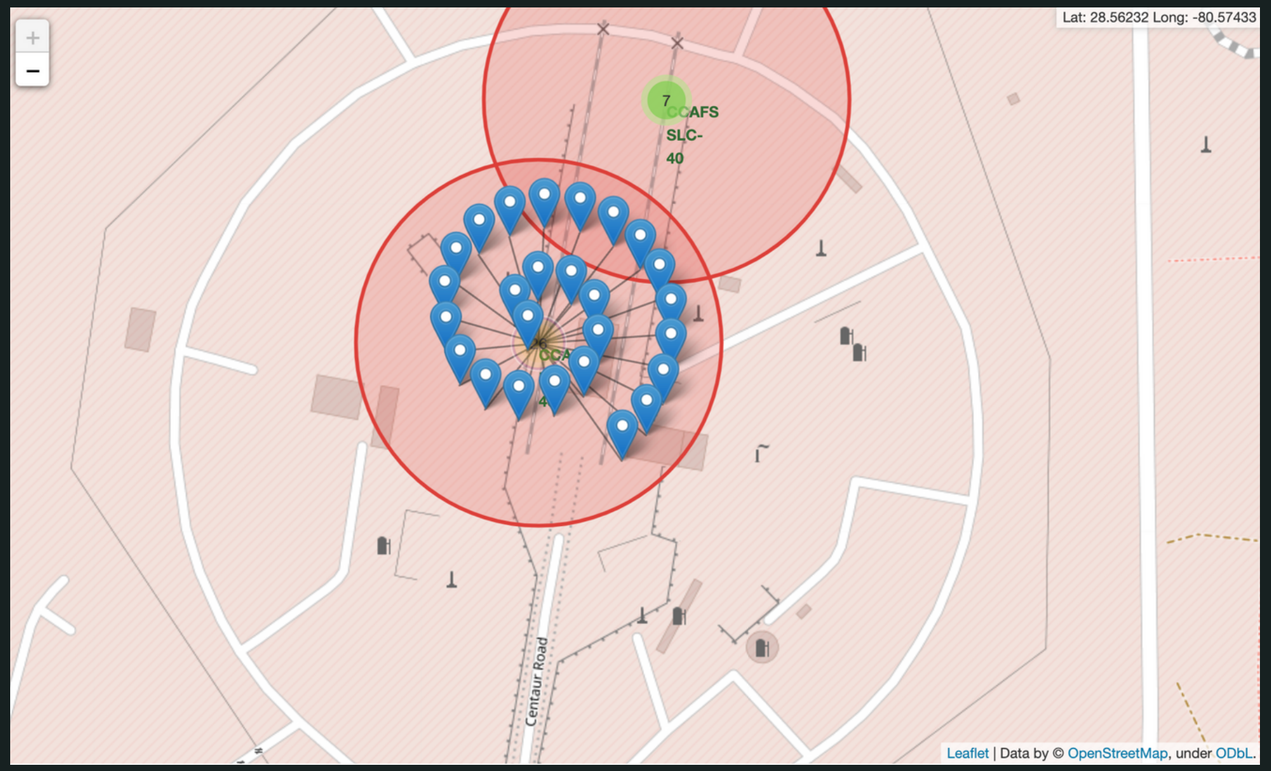
- The launch sites locations can be seen on the picture to the left
- They are located close the Equator line and coastal regions mainly for safety and engineering reasons
- Being close to the Equator can give the rockets an extra "boost", helping to conserve fuel

PROXIMITIES



- The pin points are a road called 'Titan III' to the left and to right, we have the coastline

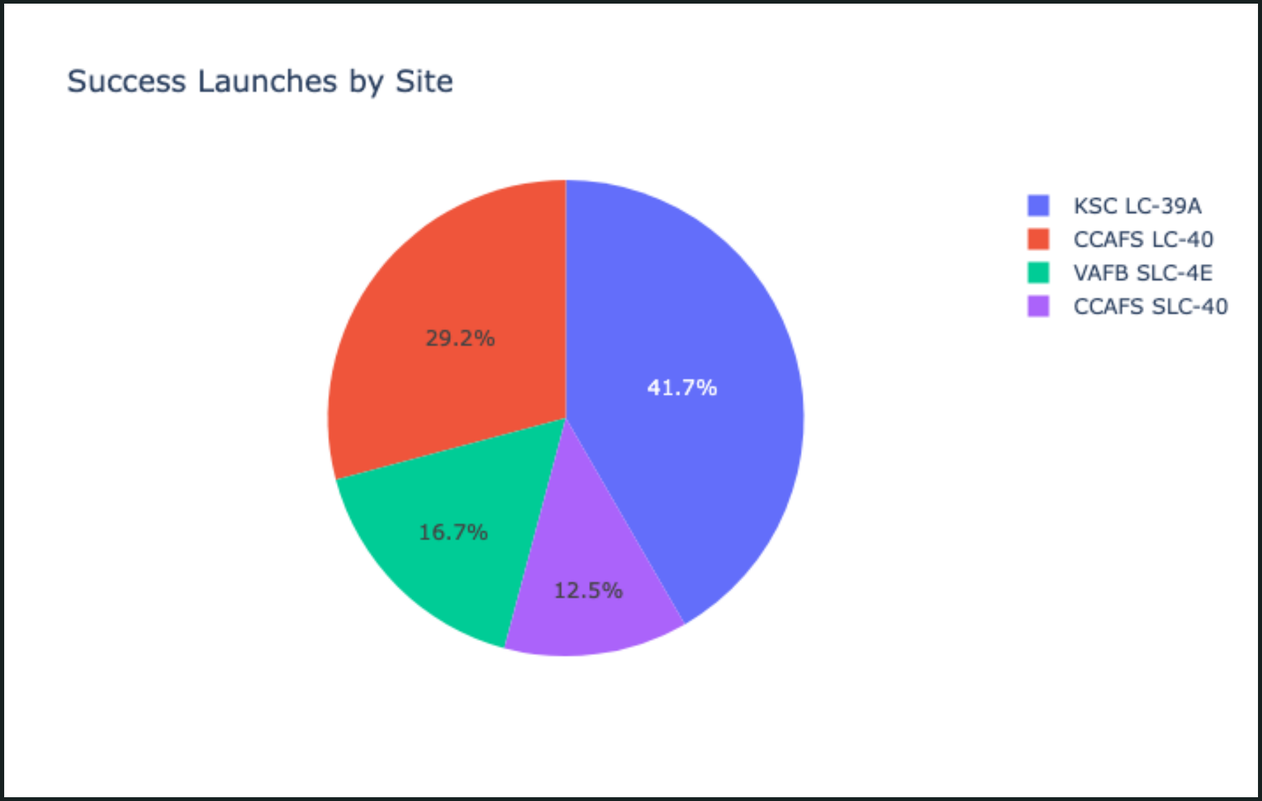
CLUSTERS OF SITES



- A cluster launch sites in a map built with Folium

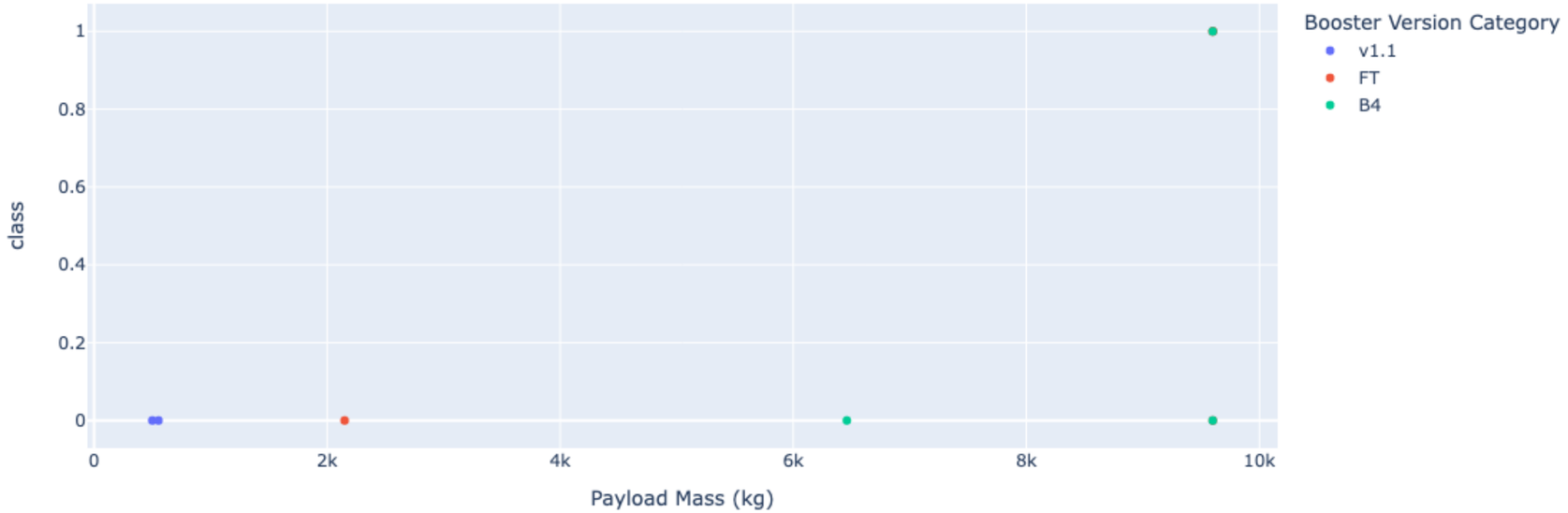
PLOTLY DASH DASHBOARD

DASHLY



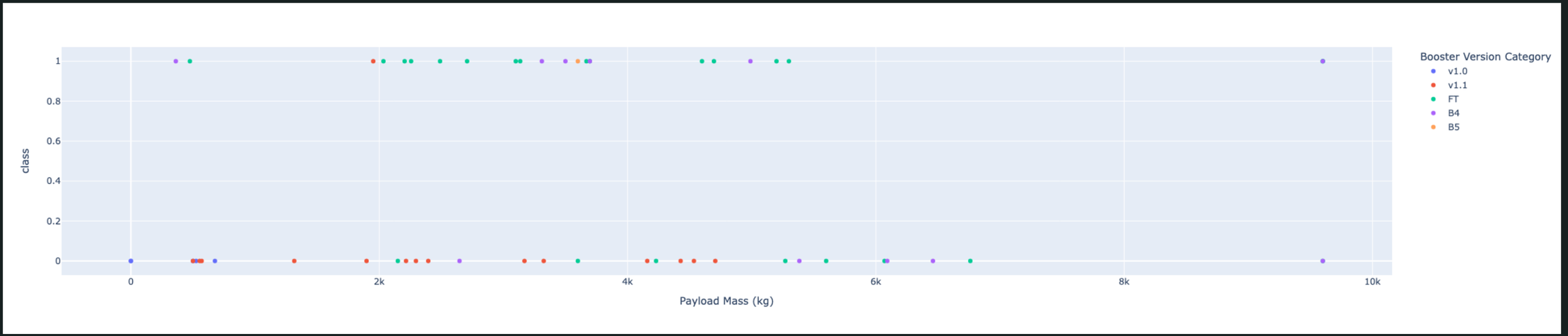
- WHICH SITE HAS THE HIGHEST LAUNCH SUCCESS RATE?
  - As seen on the Pie chart, the site with the highest launch success rate is KSC LC 39A with 41% of the total successful launches

DASHLY



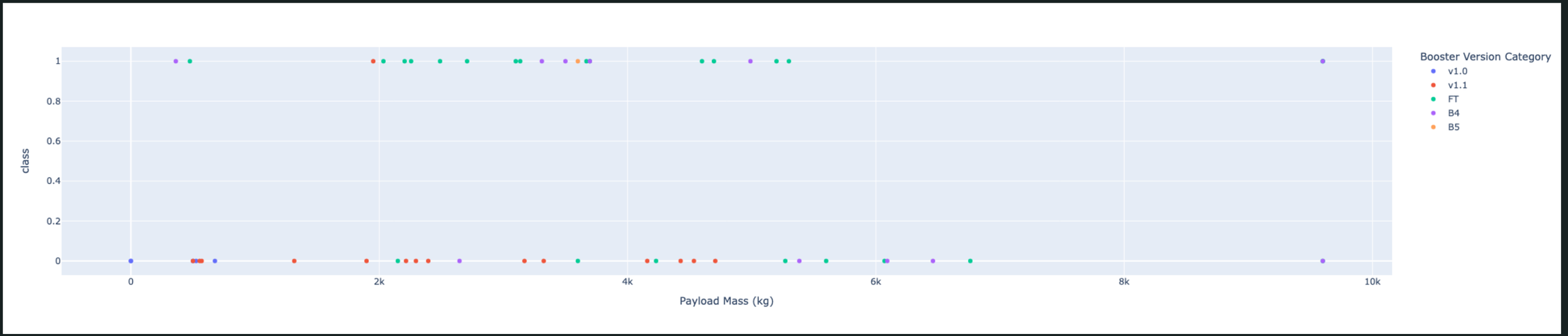
- Which site has the largest successful launches?
    - The plot above represents the launches at VAFB SLC-4E

DASHLY



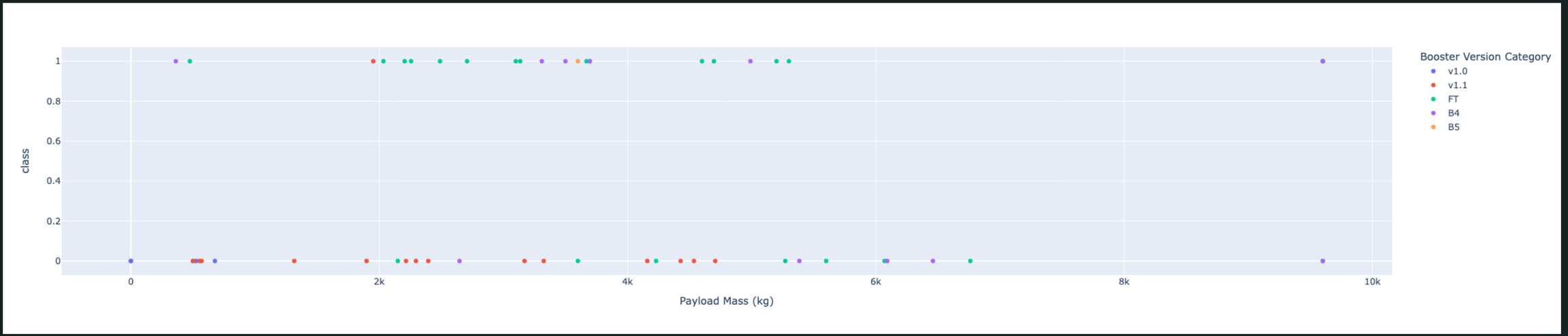
- Which payload range(s) has the highest launch success rate?
    - The payload range with the highest launch success rate is between 2000 kg and around 3.700 kg

DASHLY



- Which payload range(s) has the lowest launch success rate?
  - The payload range with the lowest launch success rate is between 2.100 kg and around 5.600 kg

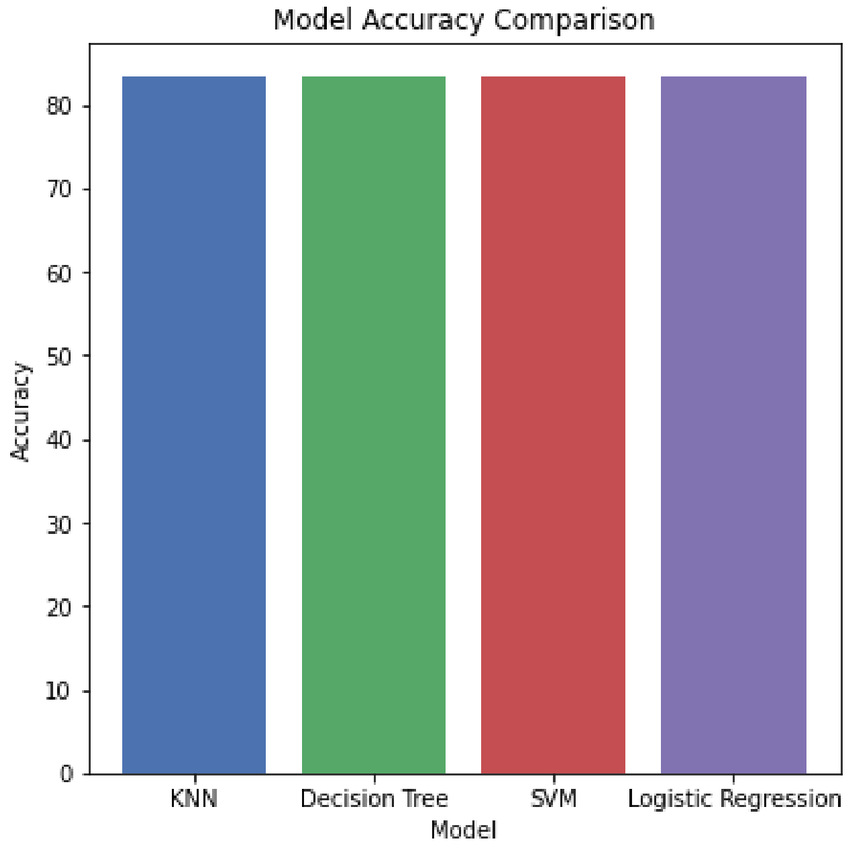
DASHLY



- Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?
  - The Booster Version with highest success rate is FT

PREDICTIVE ANALYSIS (CLASSIFICATION)

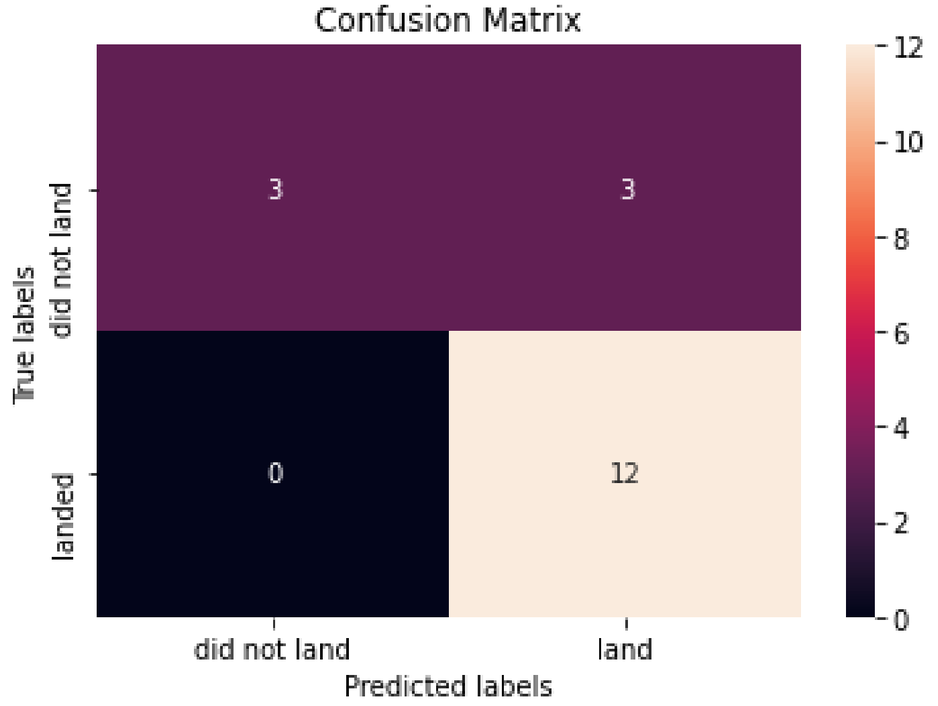
MODEL ACCURACY



- THE ACCURACY FOR ALL THE MODELS USED CAN BE SEEN ON THE PLOT.
- VIRTUALLY, THERE'S NO DIFFERENCE BETWEEN THEM. WHEN THE VALUES ARE ROUNDED, THEY ALL PRESENT 83.33% OF ACCURACY, AS CAN BE ALSO SEEN ON THE NOTEBOOKS.

	Scores
KNN	83.33
Decision Tree	83.33
SVM	83.33
Logistic Regression	83.33

CONFUSION MATRIX



- GIVEN THE SIMILARITY OF THE PERFORMANCES IT WAS EXPECTED THAT THE CONFUSION MATRIXES WOULD PERFORM THE SAME WAY.
- YOU CAN SEE THE CONFUSION MATRIX FOR THE LOGISTIC REGRESSION MODEL

Conclusion

- In this project, we examined several factors that are critical for successful space launches. Our findings indicate that the location of a launch site is a crucial determinant of success. Specifically, our analysis shows that launch sites located near the Equator and close to water have a higher likelihood of successful launches. Therefore, it is recommended that space agencies prioritize such locations when selecting a launch site for their missions.

- We also analyzed the success rate of launches over the years and found that it has been consistently increasing. This is a promising trend that demonstrates the continuous advancements and improvements in the field of space exploration. It is crucial for space agencies to maintain this trend by ensuring that their technology and infrastructure keep pace with the growing demands of space missions

- Finally, we compared the accuracy scores of four different models (LogReg, SVM, Decision Tree, KNN) and found that they had the same accuracy score. This suggests that each of these models is equally effective in predicting space launch outcomes. Our model achieved an accuracy score of 83.33%, indicating that it is a reliable tool for predicting the success of space launches.

THANK YOU!