IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Pankaj Kumar Jha

04th March 2024

Outline

Executive Summary

- Summary of methodologies

  To analyze Spacex launches data, we have followed below steps:

  - Business Understanding
  - Analytical Approach
  - Data Requirement
  - Data Collection
  - Data Understanding
  - Data Preparation
  - Modelling
  - Evaluation Deployment
  - Feedback

- Summary of all results: Spacex claim to launch rockets at half the cost of its contemporaries is true. It can be determined if a launch can be successful or not based on launch parameters.

Introduction

- Project background and context:

  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.

- Problems you want to find answers

  If we can determine if the first stage will land, we can determine the cost of a launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Data was collected using public REST API's published by SpaceX as well as WebScraping Wikipedia pages about the Launch.

- Perform data wrangling

  Replacing '?' with NaN
  Replacing NaN with mean, frequency or dropping the whole row
  Converting the data type into proper format
  Data Standardization and Normalization
  Binning the data for categorical evaluation
  One hot encoding categorical data

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

Methodology

- Perform predictive analysis using classification models

  Standardized the data

  Divided the data into Train and Test set

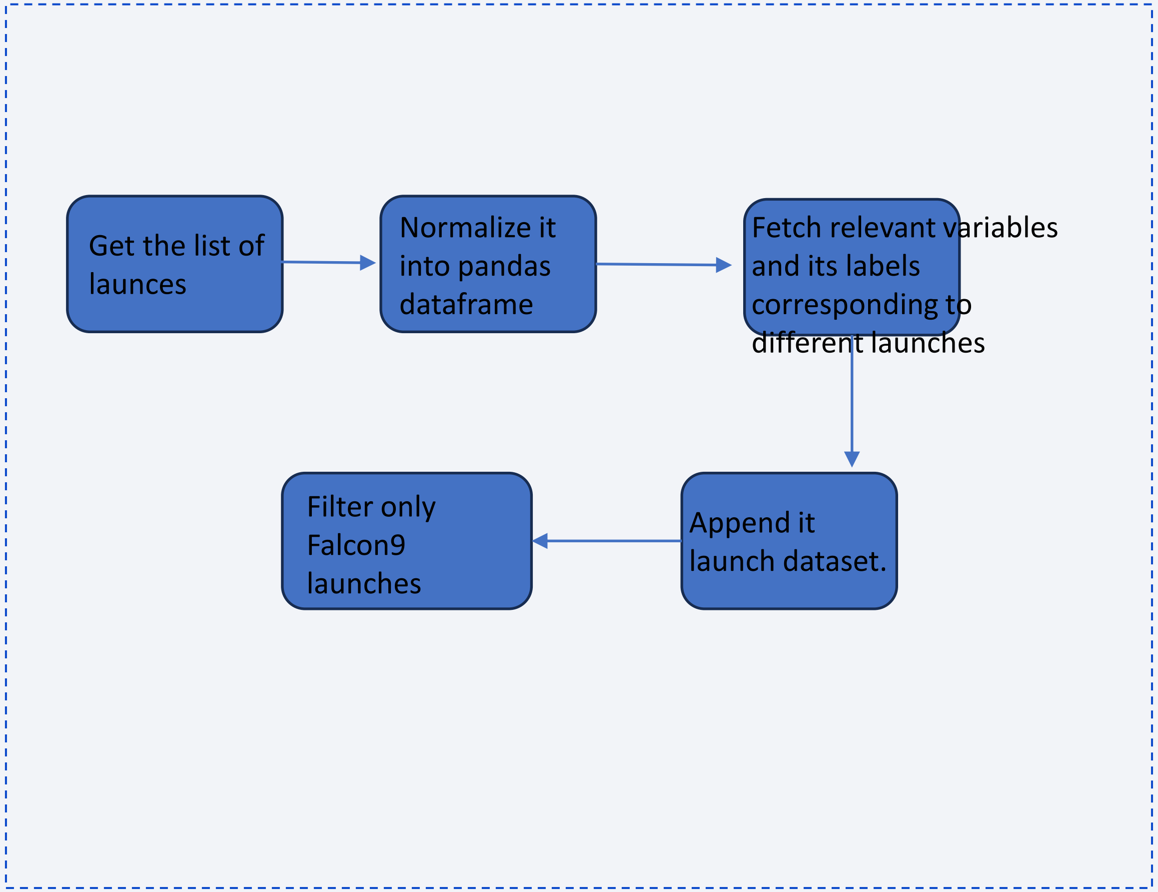
  GridSearched the models with different parameters to find best parameter.

  Fitted the models using train set with best parameter

  Scored the models using test set and compared them for their accuracy.

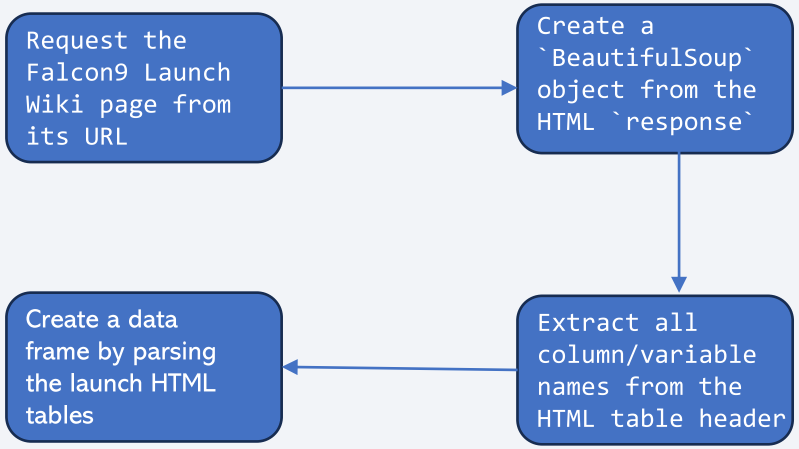
Data Collection – SpaceX API

- Defined a series of helper functions that will help us use the API to extract information using identification numbers in the launch data
- Requesting rocket launch data from SpaceX API
- Convert into Pandas dataframe using json_normalize()
- Got information about the launches using the IDs given for each launch
- Appended it to dataframe
- Filter the dataframe to only include `Falcon 9` launches
  [Spacex data collection API notebook]



Data Collection - Scraping

- `Request the Falcon9 Launch Wiki page from its URL`
- `` Create a `BeautifulSoup` object from the HTML `response` ``
- `Extract all column/variable names from the HTML table header`
- Create a data frame by parsing the launch HTML tables
- [web scraping notebook]()



Data Wrangling

- Identify and calculate the percentage of the missing values in each attribute
- Identify which columns are numerical and categorical

Data wrangling notebook

EDA with Data Visualization

- Summary of charts used for EDA:

    Scatter plot
        To understand the correlation between two variables
    Bar plot
        To understand comparative strength of categorical variables
    Line plot
        To understand the trend of variable over time.

- EDA with data visualization notebook

EDA with SQL

Summary of SQL queries performed:

Connecting a database
Creating a database
Saving a csv file as table in the database
Querying the table using SQL DML statements

- EDA with SQL notebook

Build an Interactive Map with Folium

- Summary of map objects used:
  - markers
    - These are graphical objects to denote the location(latitude and longitude)
  - circles
    - These are one of the different type of markers available in folium
  - lines
    - We draw lines to show the distance between two location
- Interactive map with Folium map

Build a Dashboard with Plotly Dash

- Summary of plots/graphs
  - Pie Chart
    - For comparative study of launch site
  - Slider
    - For adjusting payload mass of the analysis
  - Dropdown list
    - For understanding success rate at launch site level
- Plotly Dash lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

- Summary of how model was built, evaluated, improved, and found the best performing classification model

  Load the dataframe
  Identify the label column and perform one hot encoding
  Standardize unlabeled data using ‘standardscaler’
  Split data and label into train and test set
  Create various classification model class
  Using GridSearchCV, find best parameters by fitting training data
  Find the accuracy by calculating score against Test data
  Plot Confusion Matrix
  Choose a model which has highest score.

- predictive analysis lab, as an external reference and peer-review purpose

Results

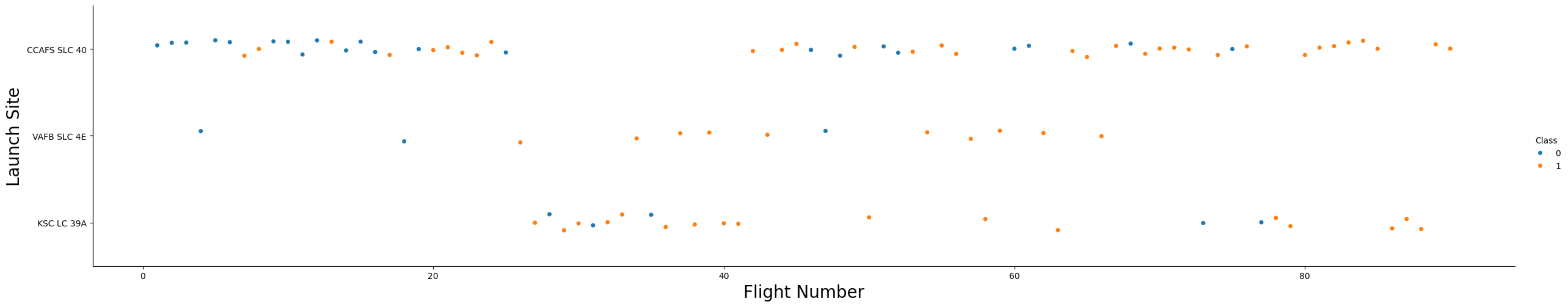
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

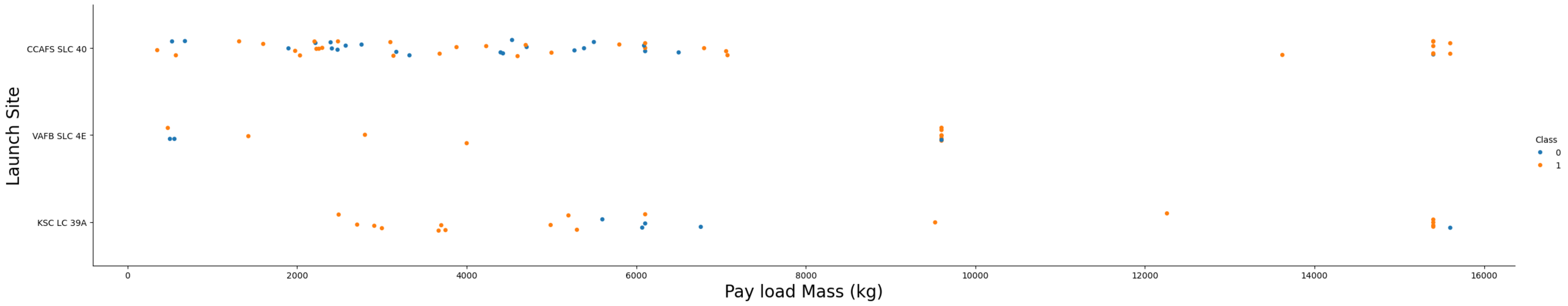
Flight Number vs. Launch Site

- Higher flight numbers are related to CCAFS SLC 40 successful launches



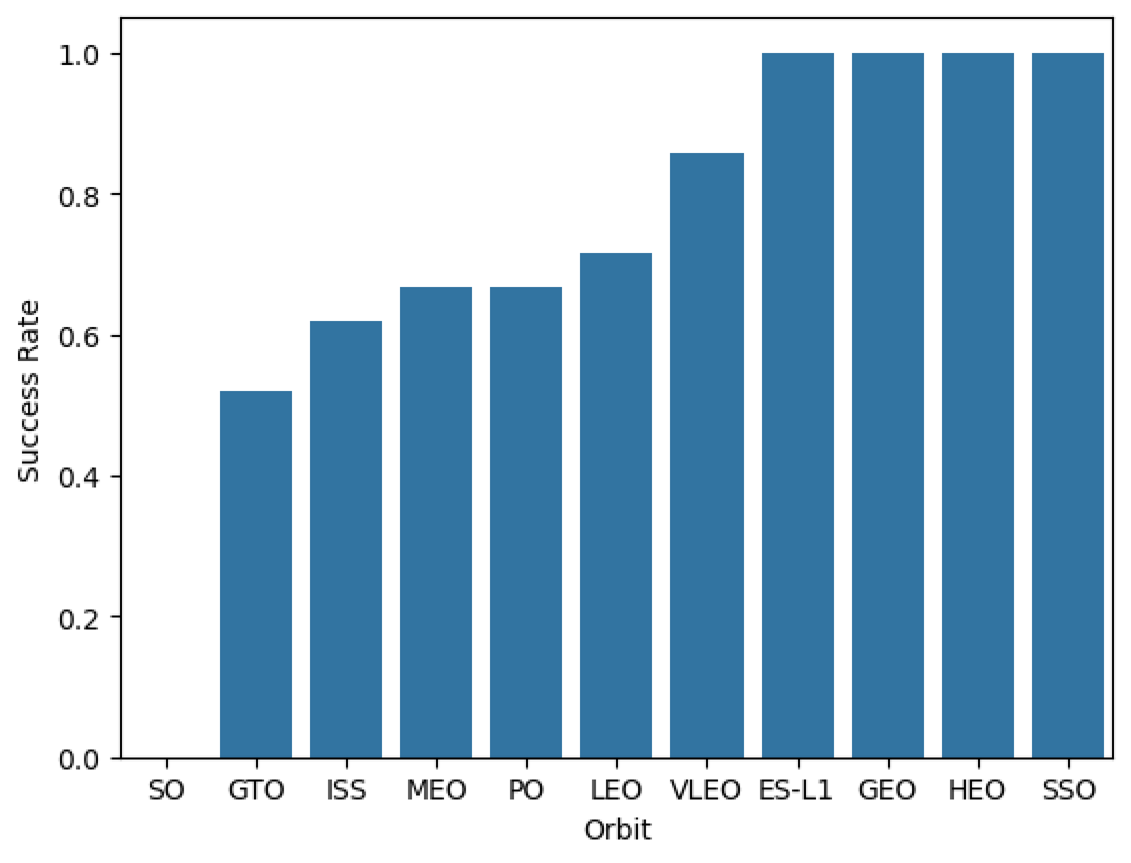
Payload vs. Launch Site

- Lower payloads are frequently used at CCAFS SLC 40



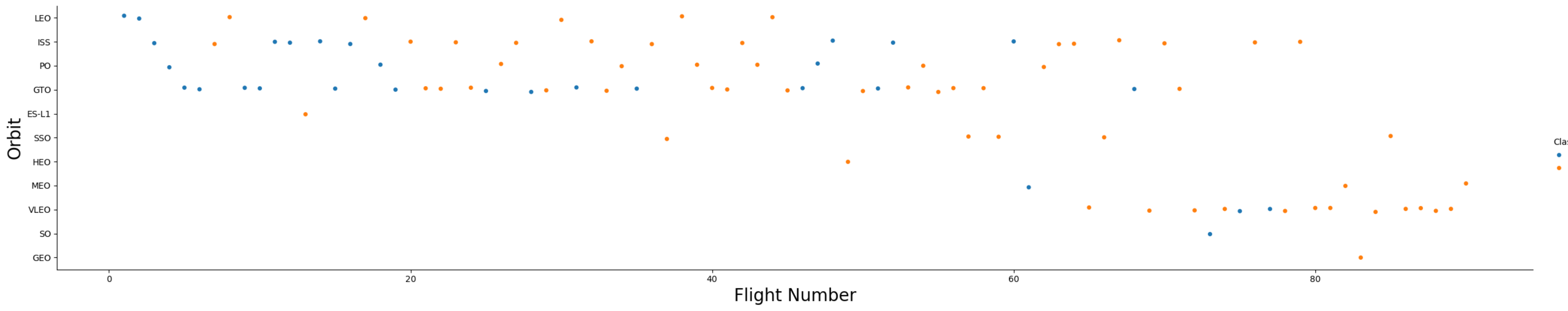
Success Rate vs. Orbit Type

GEO, HEO and SSO orbit type launches are more successful.



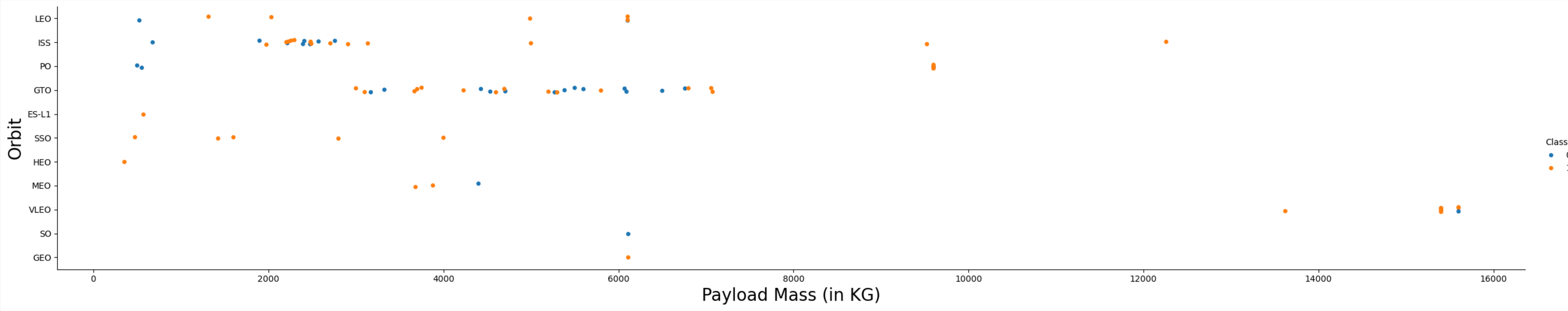
Flight Number vs. Orbit Type

LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



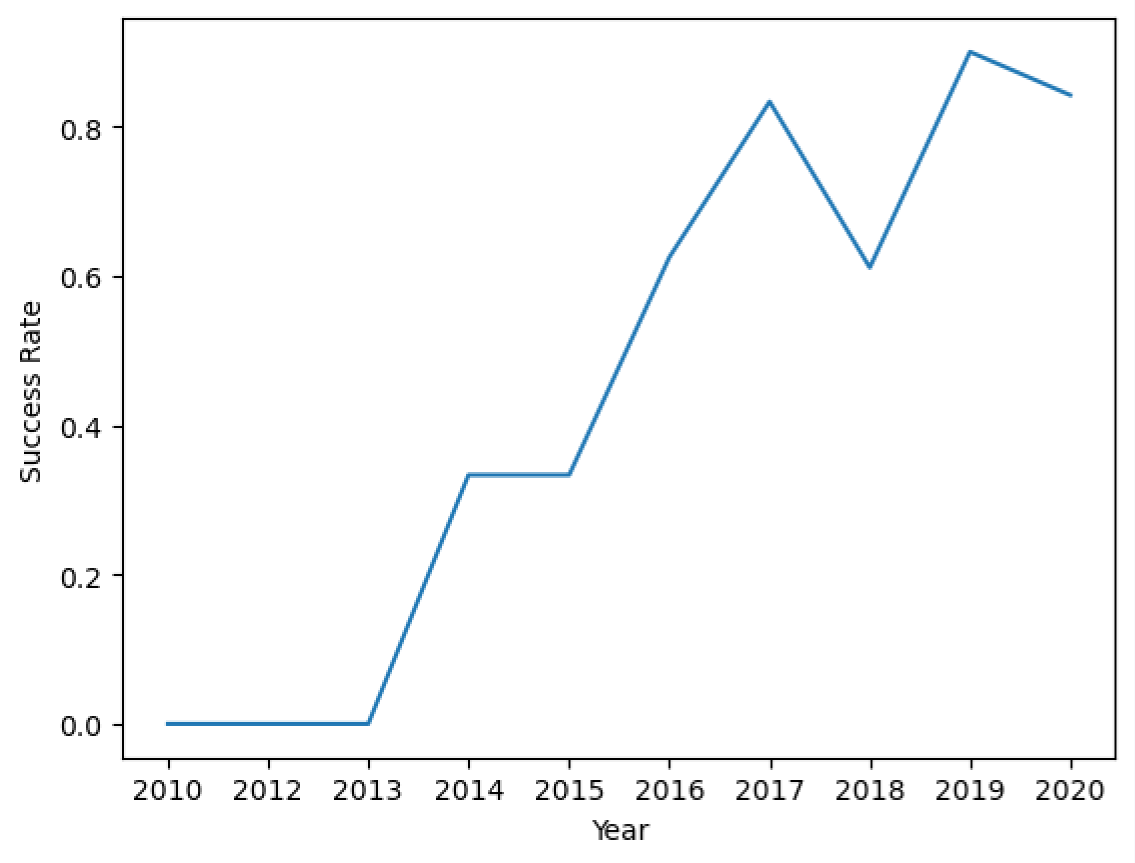
Payload vs. Orbit Type

With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

The sucess rate since 2013 kept increasing till 2020



All Launch Site Names

Two of them lie on east coast and other two lie on west coast

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

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL WHERE "Launch_Site" like 'CCA%' limit 5
```

Python

```
 * 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

Total Payload Mass

Total payload mass carries by booster versions:

```
%%sql
select "Booster_Version",sum("PAYLOAD_MASS__KG_") FROM SPACEXTBL GROUP BY "Booster_Version"
```

* sqlite:///my_data1.db
Done.

Booster_Version	sum("PAYLOAD_MASS__KG_")
F9 B4 B1039.2	2647
F9 B4 B1040.2	5384
F9 B4 B1041.2	9600
F9 B4 B1043.2	6460
F9 B4 B1039.1	3310
F9 B4 B1040.1	4990
F9 B4 B1041.1	9600
F9 B4 B1042.1	3500
F9 B4 B1043.1	5000
F9 B4 B1044	6092

Average Payload Mass by F9 v1.1

```
%sql select avg("PAYLOAD_MASS__KG_") FROM SPACEXTBL where "Booster_Version" = 'F9 v1.1'
```

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

avg("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

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

```
%sql select "Landing_Outcome", min(Date) from SPACEXTBL group by  "Landing_Outcome"
```

 * sqlite:///my_data1.db
Done.

Landing_Outcome	min(Date)
Controlled (ocean)	2014-04-18
Failure	2018-12-05
Failure (drone ship)	2015-01-10
Failure (parachute)	2010-06-04
No attempt	2012-05-22
No attempt	2019-08-06
Precluded (drone ship)	2015-06-28
Success	2018-07-22
Success (drone ship)	2016-04-08
Success (ground pad)	2015-12-22
Uncontrolled (ocean)	2013-09-29

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","Landing_Outcome", "PAYLOAD_MASS__KG_"
    from SPACEXTBL
    WHERE "Landing_Outcome" ="Success (drone ship)"
    and "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000
```

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

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

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

[35]

* sqlite:///my_data1.db
Done.

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

Boosters Carried Maximum Payload

```
%%sql
SELECT "Booster_Version"
FROM SPACEXTBL
WHERE "PAYLOAD_MASS__KG_" IN (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

2015 Launch Records

```
%%sql
SELECT substr(Date, 6, 2) as "Month",
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM SPACEXTBL
WHERE "Landing_Outcome" = "Failure (drone ship)"
AND substr(Date, 0, 5) = '2015'
```

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

Month	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

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

```
%%sql
SELECT "Landing_Outcome", COUNT(*)
FROM SPACEXTBL
WHERE Date between '2010-06-04' and '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY COUNT(*) DESC
```

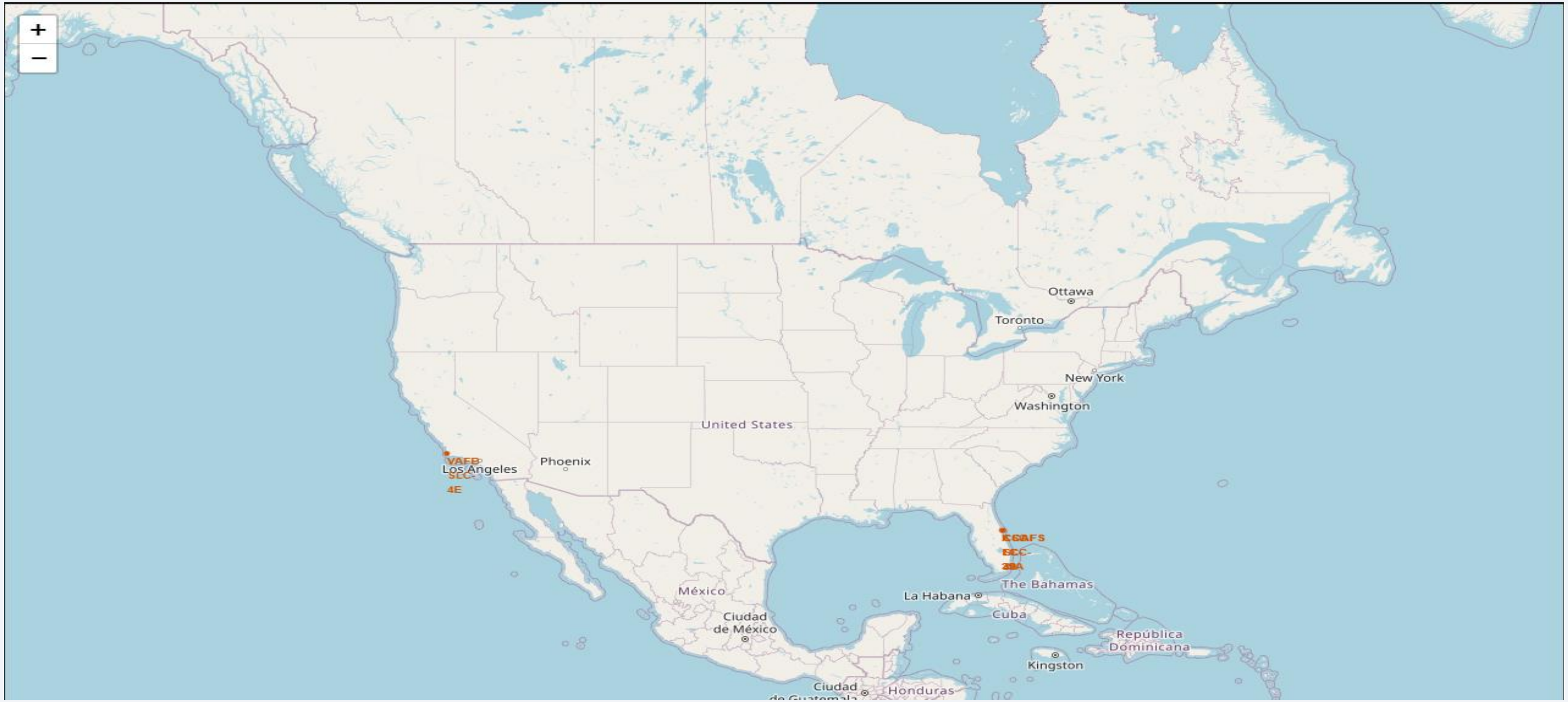
* sqlite:///my_data1.db
Done.

Landing_Outcome	COUNT(*)
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

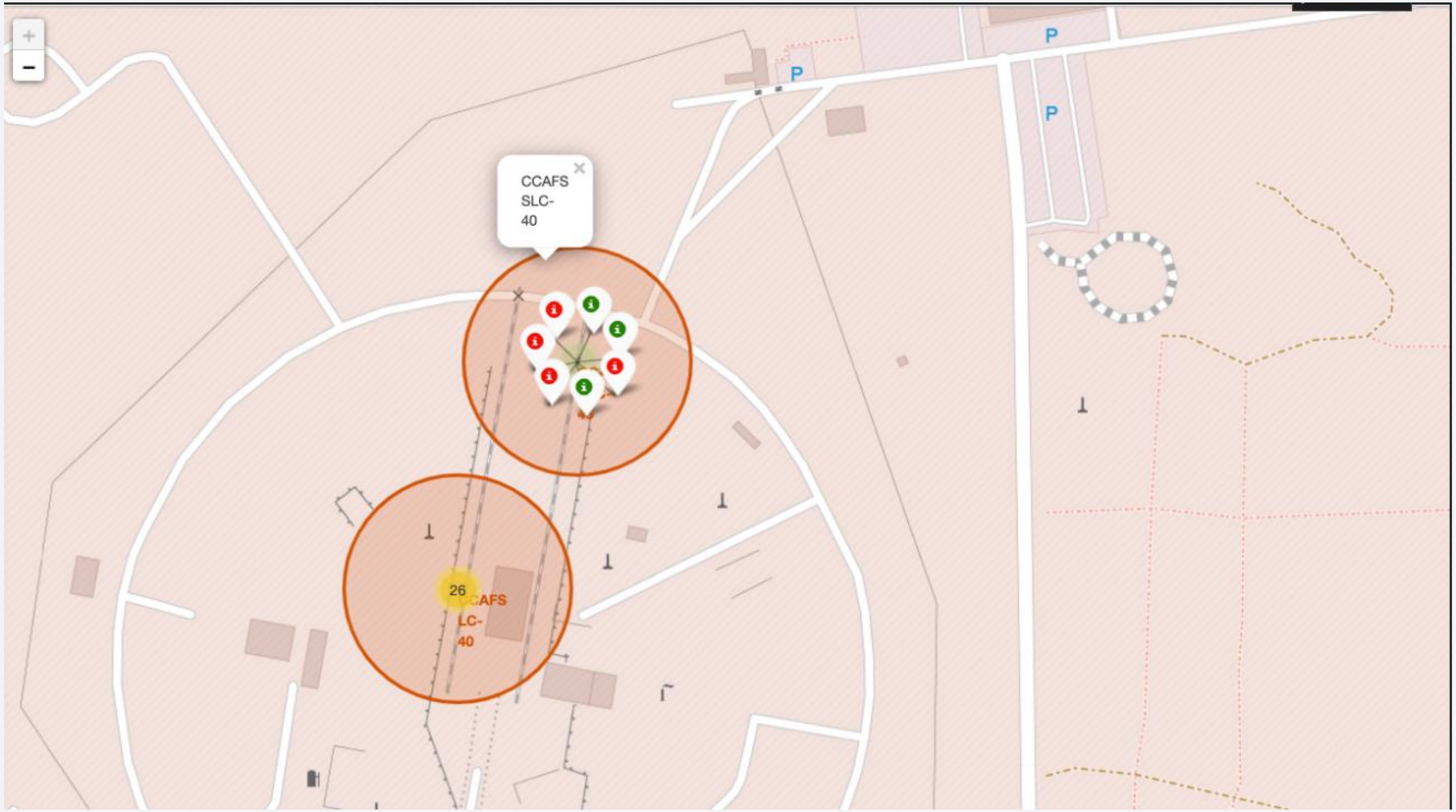
Section 3

Launch Sites Proximities Analysis

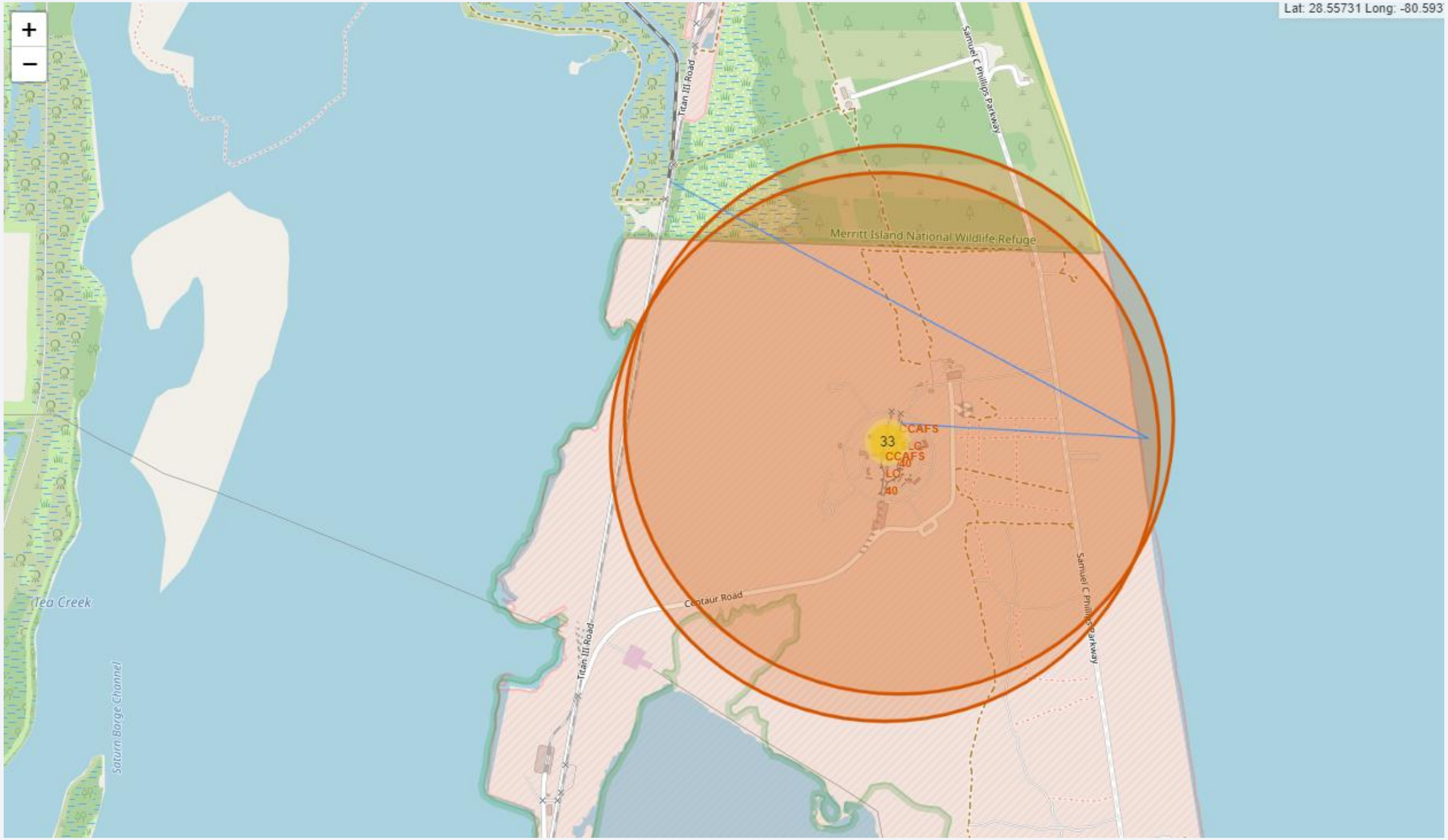
Screenshot of Launch Locations



Color Coded launch outcomes



Distance between two point using line

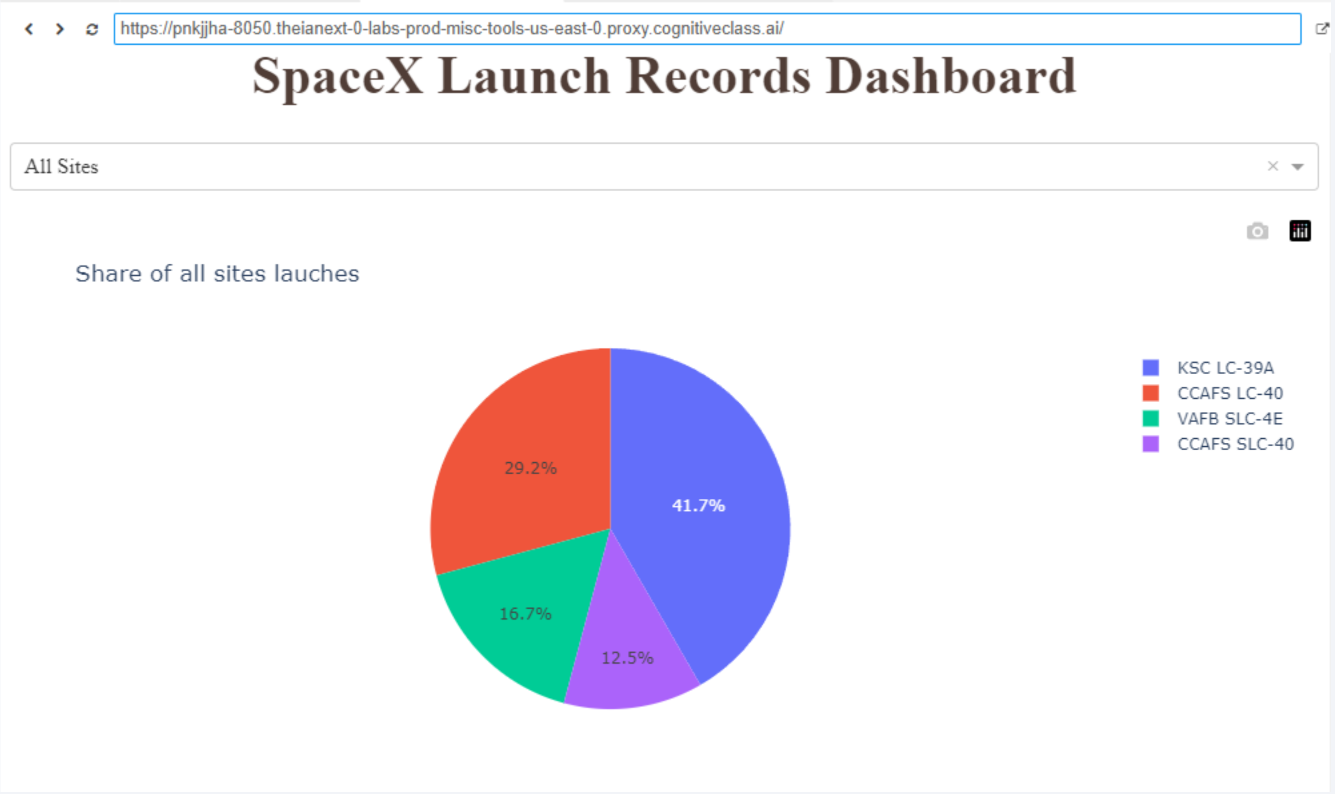


Section 4

Build a Dashboard with Plotly Dash

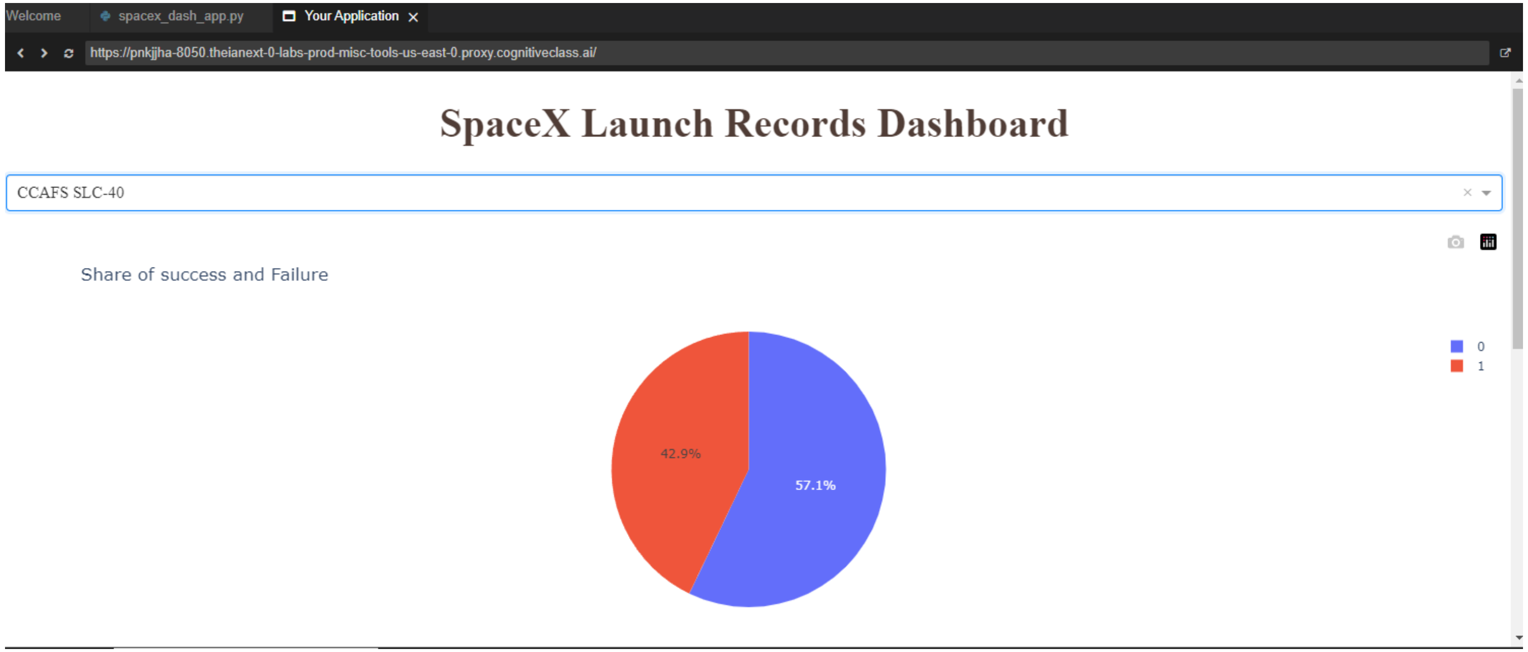
Launches across Sites

- Title of the pie chart
- Legends lists out all Launch Sites
- KSC LC 39A and CCAFS LC-40 contributes to more than 50% of launches.



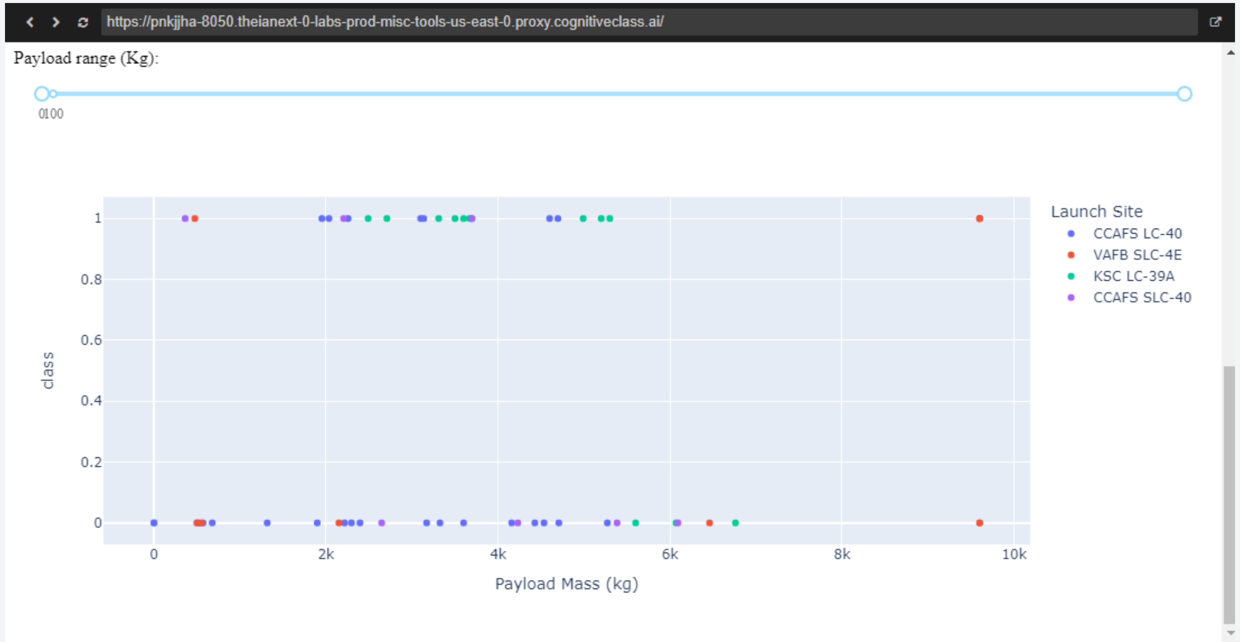
Site CCAFS SLC-40 has the highest success ratio of 42.9%

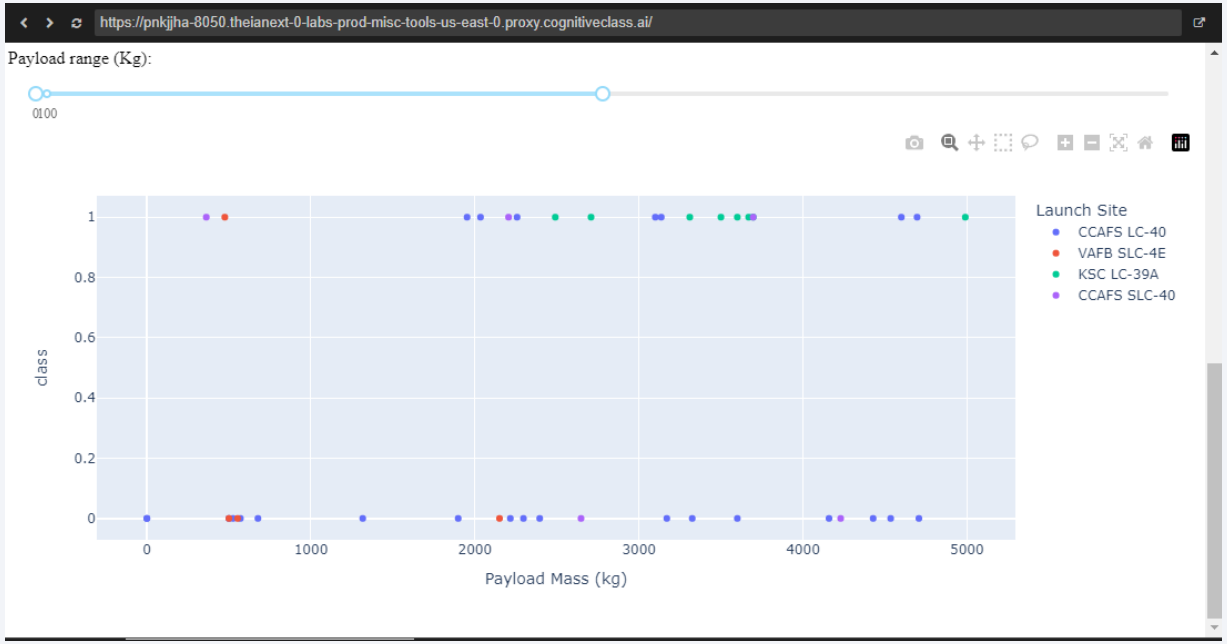
Most successful Launch Site



Payload vs Success ratio across launch sites

Below shows that success ratio increase for lower payload range

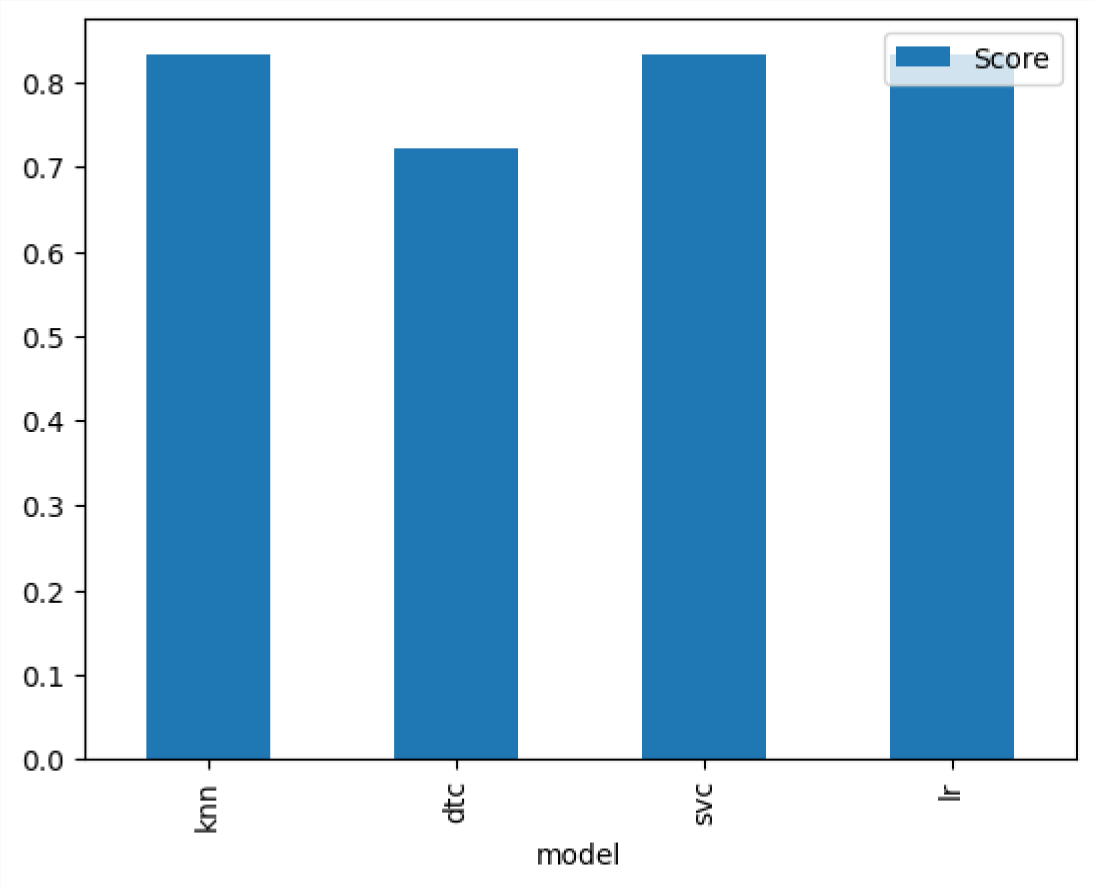




Section 5

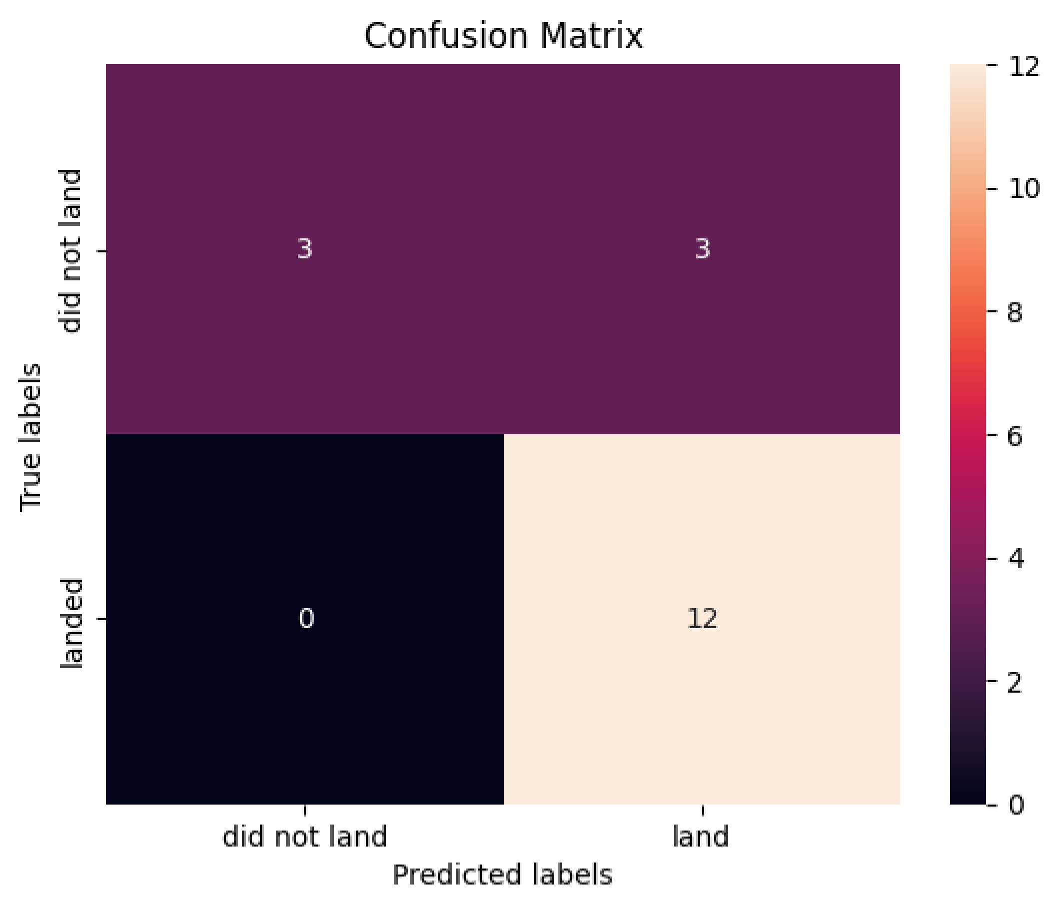
Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix

K – nearest neighbour, support vector machine and logistic regression classifier, all three of them are good classifier with high accuracy of their prediction, i.e high true positives and high true negatives



Conclusions

Successful launch outcomes can be easily predicted by the launch site, Orbit, payload and flight number

Thank you!