Data Science Capstone Project

Udith.R

15/02/2023

Outline

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

Executive Summary

- Data collection from public SpaceX API and Wikipedia page.
- Data wrangling
- Exploratory Data Analysis with Data Visualization using pandas and matplotlib
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis
- Finding best parameter using GridSearchCV and visualising the accuracy score of the models

Introduction

SpaceX is the most successful company of the commercial space age. They made the space travel very affordable. The company 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. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

Methodology

- Data collection methodology
  - Combined data from SpaceX public API and SpaceX wikipedia page.
- Performed data wrangling
  - Dealt with missing values
  - Used One Hot Encoding
- Performed exploratory data analysis (EDA) to clean the data, and found out interesting insights from it.
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Performed Hyper-parameter tuning with GridSearchCV.

Data collection methodology

Data collection process is a combination of API requests from SpaceX REST API and Web Scraping of data from SpaceX's Wikipedia page.

We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

Data obtained by using SpaceX REST API are

- FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Data obtained by using Wikipedia SPACEX PAGE are

- Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data collection – SpaceX API

1. Requesting rocket launch data from SpaceX API
2. Decoding the response content using .json() and turning it into a dataframe using .json_normalize()
3. Creating a dataframe from the dictionary
4. Filtering the dataframe to only include Falcon 9 launches
5. Replacing missing values of Payload Mass column with calculated .mean() for this column
6. Exporting the data to CSV

Github: data collection

Data collection – Web scraping

1. Requesting Falcon 9 launch data from Wikipedia
2. Creating a BeautifulSoup object from the HTML response
3. Extracting all column names from the HTML table header
4. Collecting the data by parsing HTML tables
5. Constructing data we have obtained into a dictionary
6. Creating a dataframe from the dictionary
7. Exporting the data to CSV

Github:web scraping

Data wrangling

In the data set, there are several different cases where the booster did not land successfully

- True Ocean - the mission was successfully landed to the ocean.
- False Ocean - the mission was unsuccessfully landed to the ocean.
- True RTLS - the mission was successfully landed to a ground pad.
- False RTLS - the mission was unsuccessfully landed to a ground pad.
- True ASDS - the mission was successfully landed on a drone ship.
- False ASDS - the mission was unsuccessfully landed on a drone ship.

Convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful.

Github:data wrangling

EDA with data visualization

Charts were plotted

- Flight Number vs Payload Mass
- Flight Number vs Launch Site
- Payload Mass vs Launch Site
- Orbit Type vs Success Rate
- Flight Number vs Orbit Type
- Payload Mass vs Orbit Type
- Success Rate Yearly Trend

Scatter plots, line charts, and bar plots were used to compare relationships between variables to find out if a relationship exists so that they could be used in training the machine learning models.

Github:eda with pandas

EDA with SQL

- Loaded the dataset into IBM DB2 database
- Performed SQL queries:

1. Displaying the names of the unique launch sites in the space mission
2. Displaying 5 records where launch sites begin with the string ‘CCA'
3. Displaying the total payload mass carried by boosters launched by NASA (CRS)
4. Displaying average payload mass carried by booster version F9 v1.1
5. Listing the date when the first successful landing outcome in ground pad was achieved
6. Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. Listing the total number of successful and failure mission outcomes
8. Listing the names of the booster versions which have carried the maximum payload mass
9. Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
10. Ranking 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:sql_eda

Build an interactive map with Folium

Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.

This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.

Github:folium

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:
  - Added a dropdown list to enable Launch Site selection.
- Pie Chart showing Success Launches:
  - Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.
- Slider of Payload Mass Range:
  - Added a slider to select Payload range.
- Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:
  - Added a scatter chart to show the correlation between Payload and Launch Success.

Github:dashbording

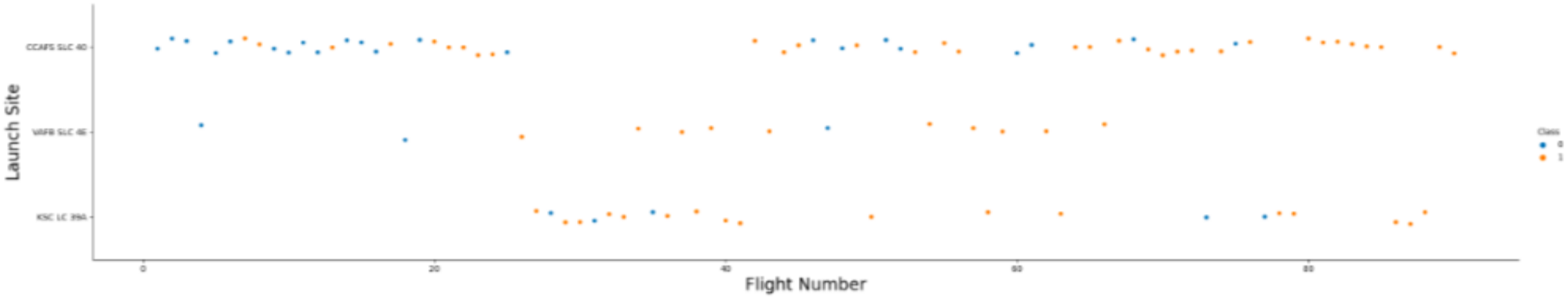
Predictive analysis (Classification)

1. Creating a NumPy array from the column “Class” in data.
2. Standardizing the data with StandardScaler, then fitting and transforming it.
3. Splitting the data into training and testing sets with train_test_split function.
4. Creating a GridSearchCV object with cv = 10 to find the best parameters.
5. Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models.
6. Calculating the accuracy on the test data using the method .score() for all models.
7. Examining the confusion matrix for all models.
8. Finding the method performs best by examining the Jaccard_score and F1_score metrics.

Github:ML

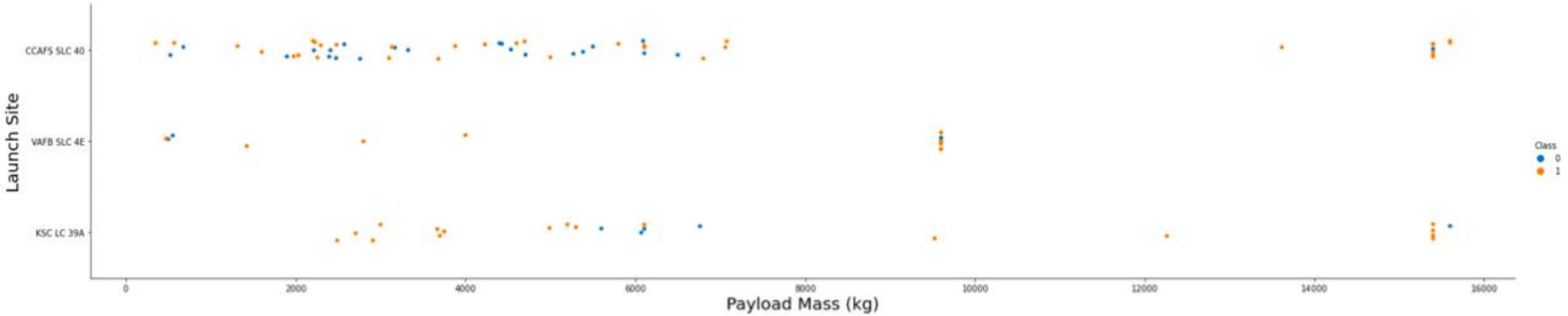
EDA with Visualization

Flight Number vs. Launch Site



Green indicates successful launch and Purple indicates unsuccessful launch.
The success rate in earliest flights were unsuccessful while the latest flights have high success rate.
CCAFS SLC 40 is the main launch site as it has the most flight number.

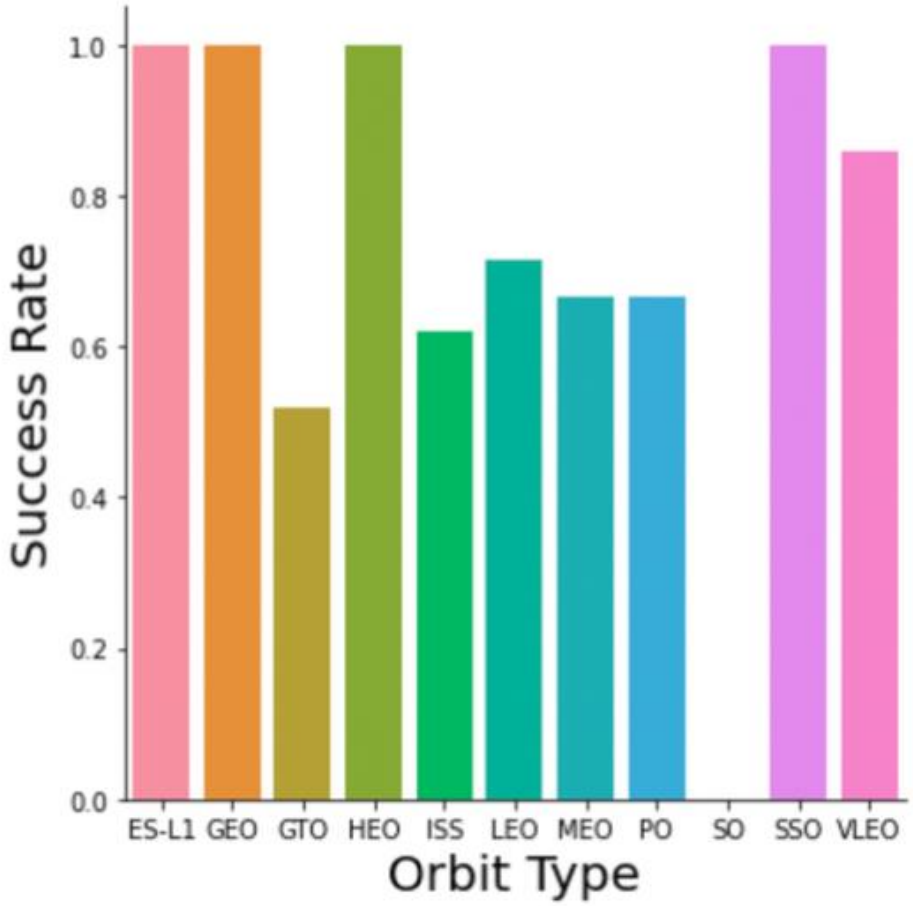
Payload vs. Launch Site



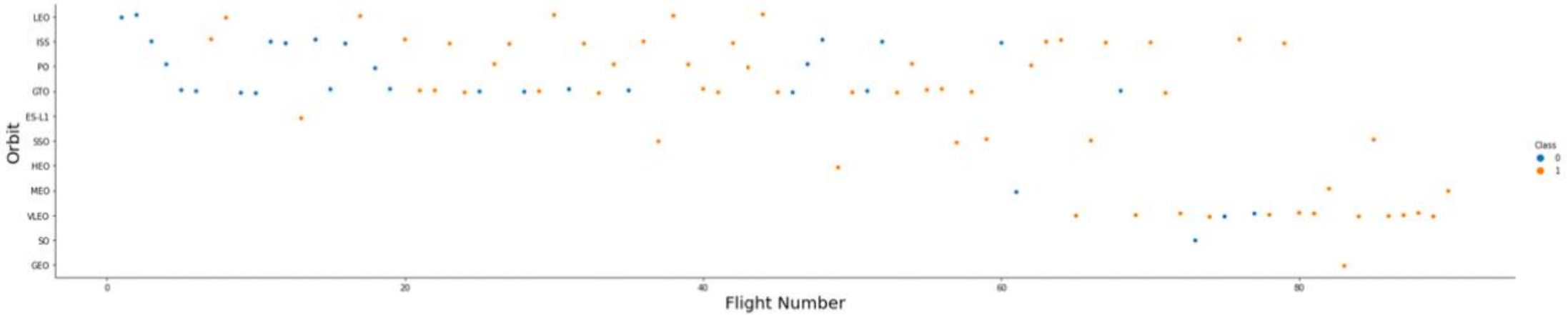
For every launch site the higher the payload mass, the higher the success rate.
Launches with payload mass over 7000 kg have very high success rate.

Success rate vs. Orbit type

ES-L1, GEO, HEO, SSO orbits have 100% success rate
S0 orbit have 0% success rate
GTO, ISS, LEO, MEO, PO have success rate between 50% and 85%

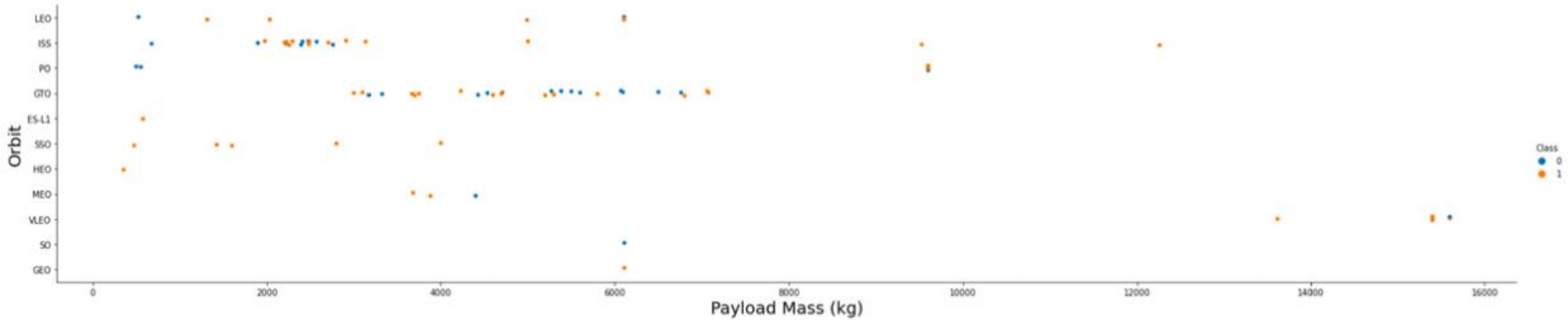


Flight Number vs. Orbit type



Launch orbit preference changed over Flight number.
The success rate was better in low orbits.

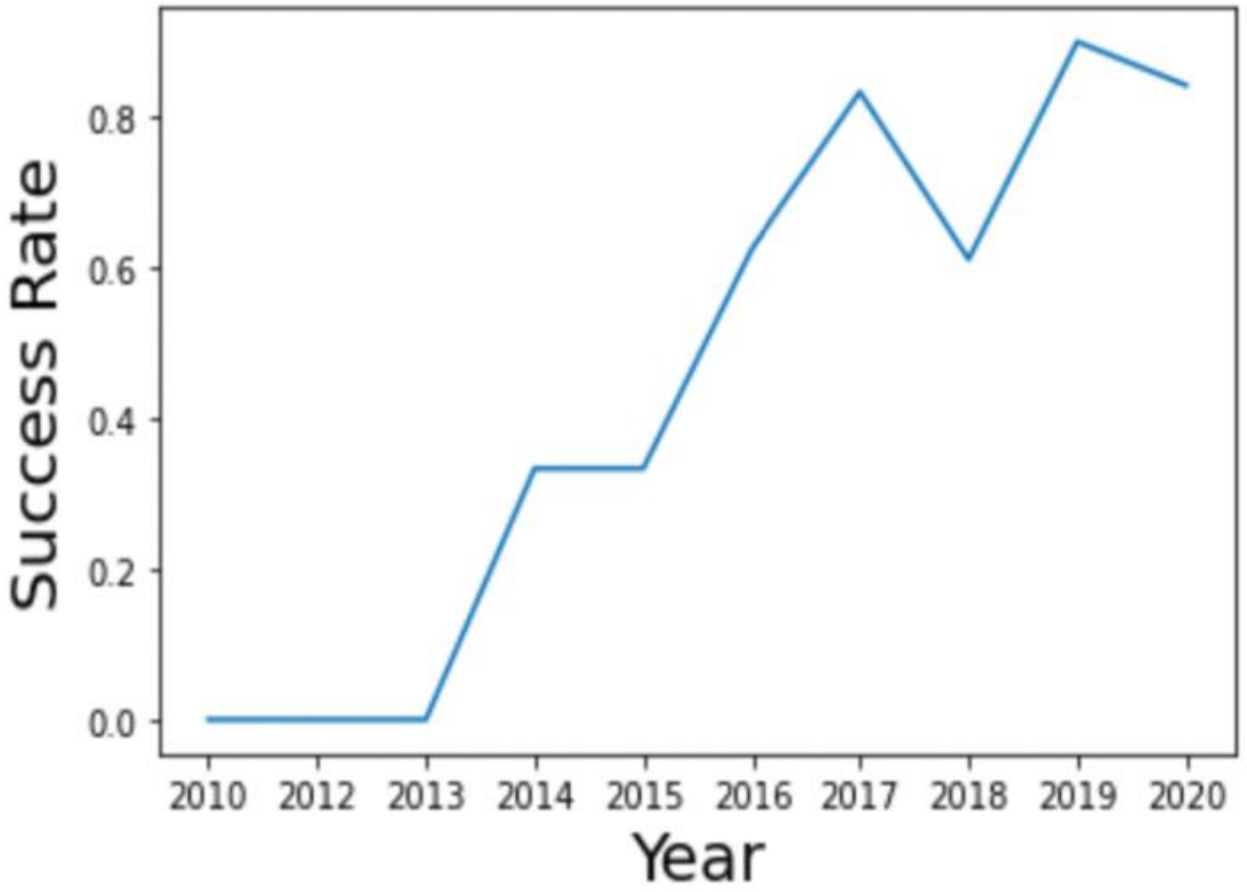
Payload Mass vs. Orbit type



Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch success yearly trend

The success rate since 2013 kept increasing till 2020.succes in recent years is around 80%.



EDA with SQL

All launch site names

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[4]:

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

Displaying the names of the unique launch sites in the space mission.

Launch site names begin with `CCA`

```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Displaying 5 records where launch sites begin with the string 'CCA'.

Total payload mass

```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[6]:

total_payload_mass
45596

Displaying the total payload mass carried by boosters launched by NASA

Average payload mass by F9 v1.1

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[7]:

average_payload_mass
2534

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

First successful ground landing date

```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[8]:

first_successful_landing
2015-12-22

Listing the date when the first successful landing outcome in ground pad was achieved.

Successful drone ship landing with payload between 4000 and 6000

```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[9]:

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

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

Total number of successful and failure mission outcomes

```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[10]:

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Listing the total number of successful and failure mission outcomes.

Boosters carried maximum payload

```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[11]:

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

Listing the names of the booster versions which have carried the maximum payload mass.

2015 launch records

In [12]:
```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[12]:

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

Rank success count between 2010-06-04 and 2017-03-20

In [13]:
```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Ranking 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.

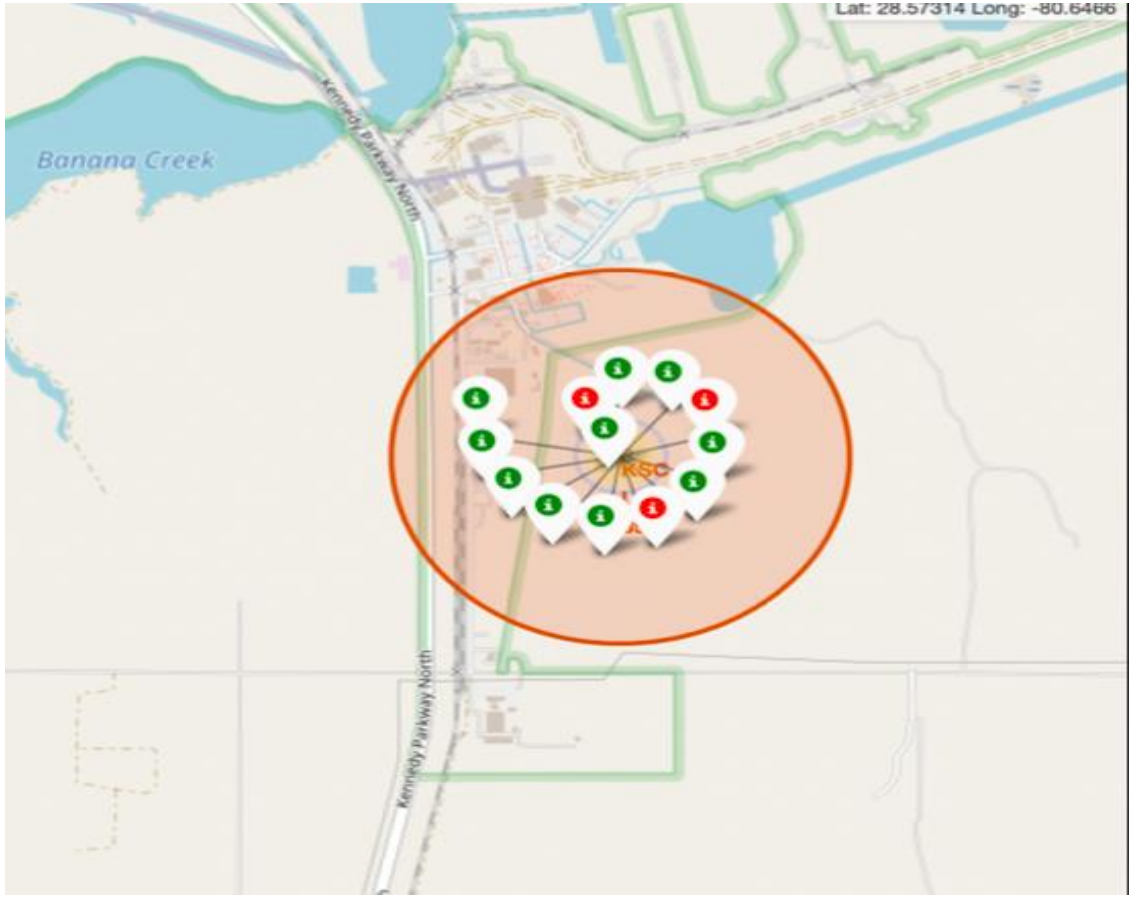
Interactive map with Folium

All launch sites' location markers on a global map

Most of Launch sites are in proximity to the Equator line.
All launch sites are very close proximity to the coast it is because, it minimises the risk of having any debris dropping or exploding near residential areas.

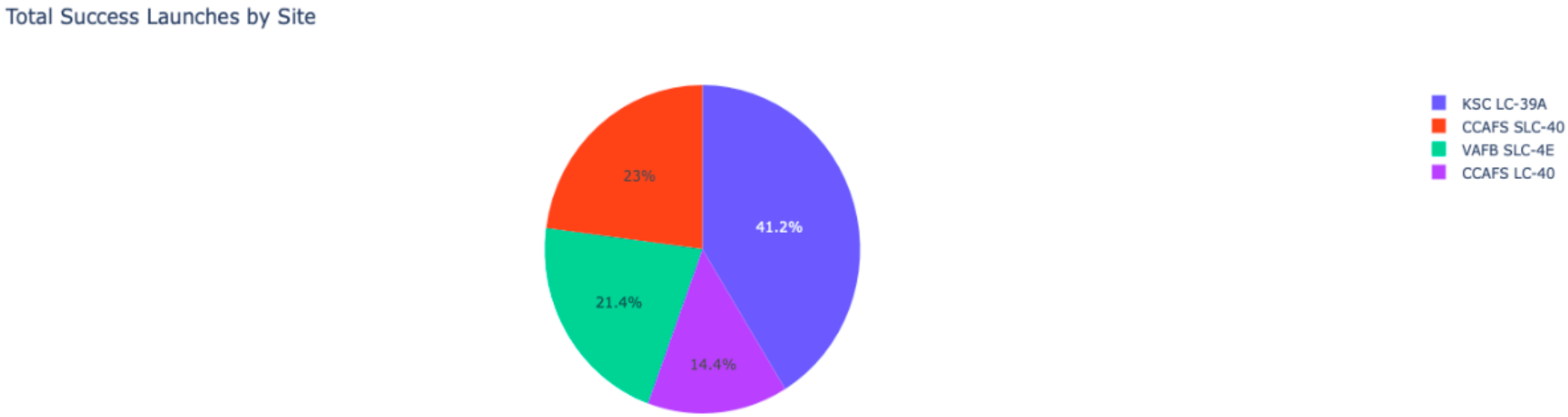
Colour-labeled launch records on the map

Launch Site KSC LC-39A has a very high Success Rate.(green marker denotes successful launch site and red marker denotes unsuccessful launch sites)



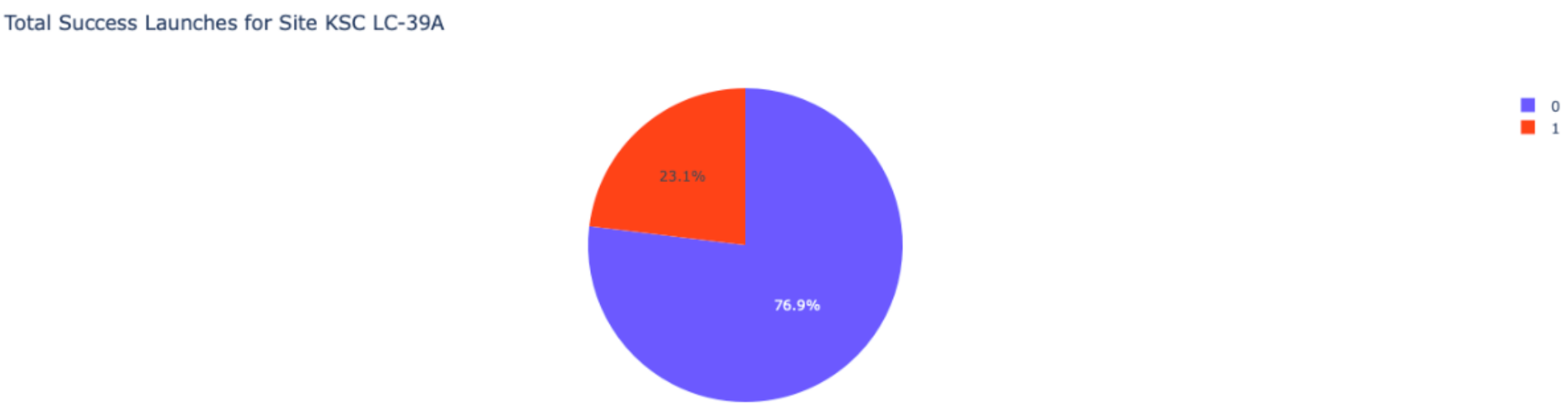
Build a Dashboard with Plotly Dash

Launch success count for all sites



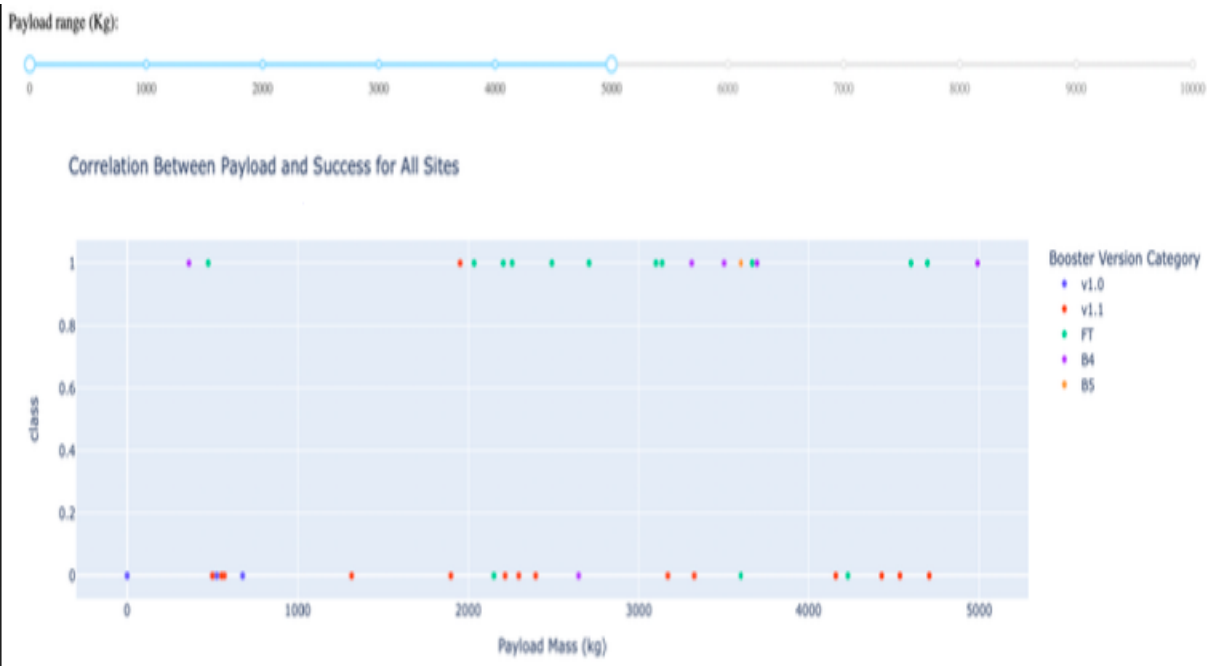
The chart shows that among all the sites, KSC LC-39A has the most successful launches.

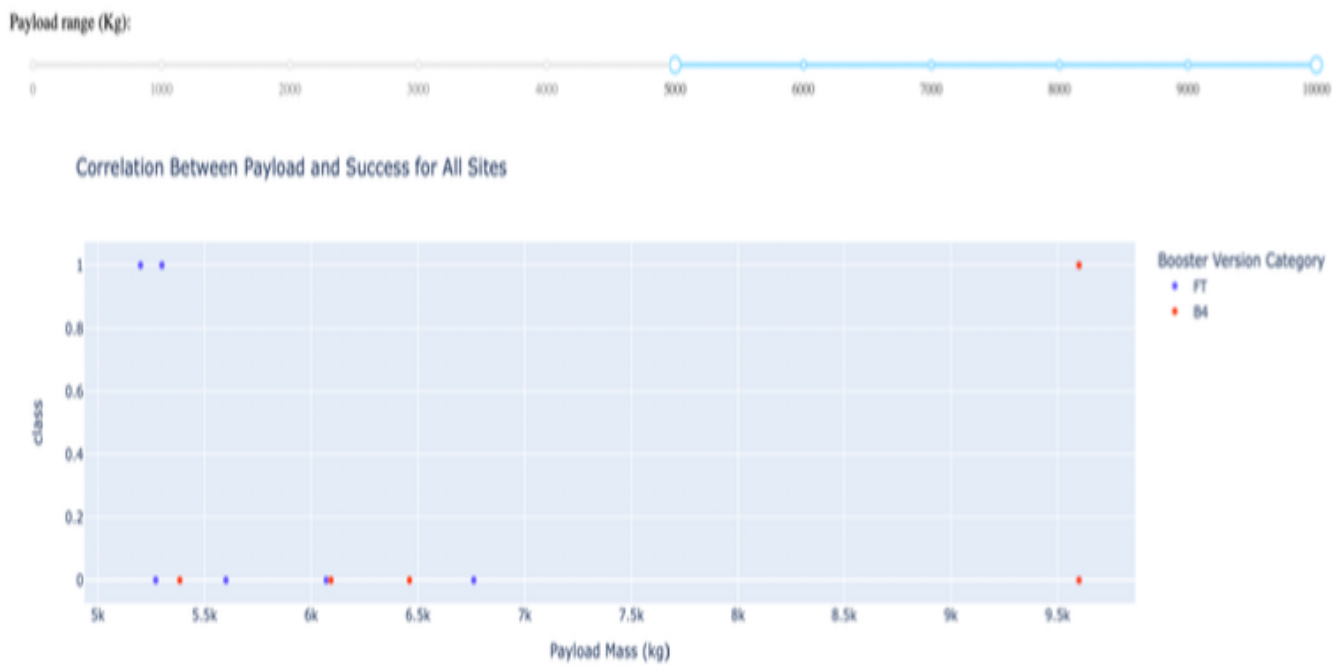
Launch site with highest launch success ratio



KSC LC-39A has the highest launch success rate of 76.9% (10 successful and 3 failed landings).

Payload Mass vs. Launch Outcome for all sites





The charts show that payloads between 2000 and 5500 kg have the highest success rate.

Predictive analysis (Classification)

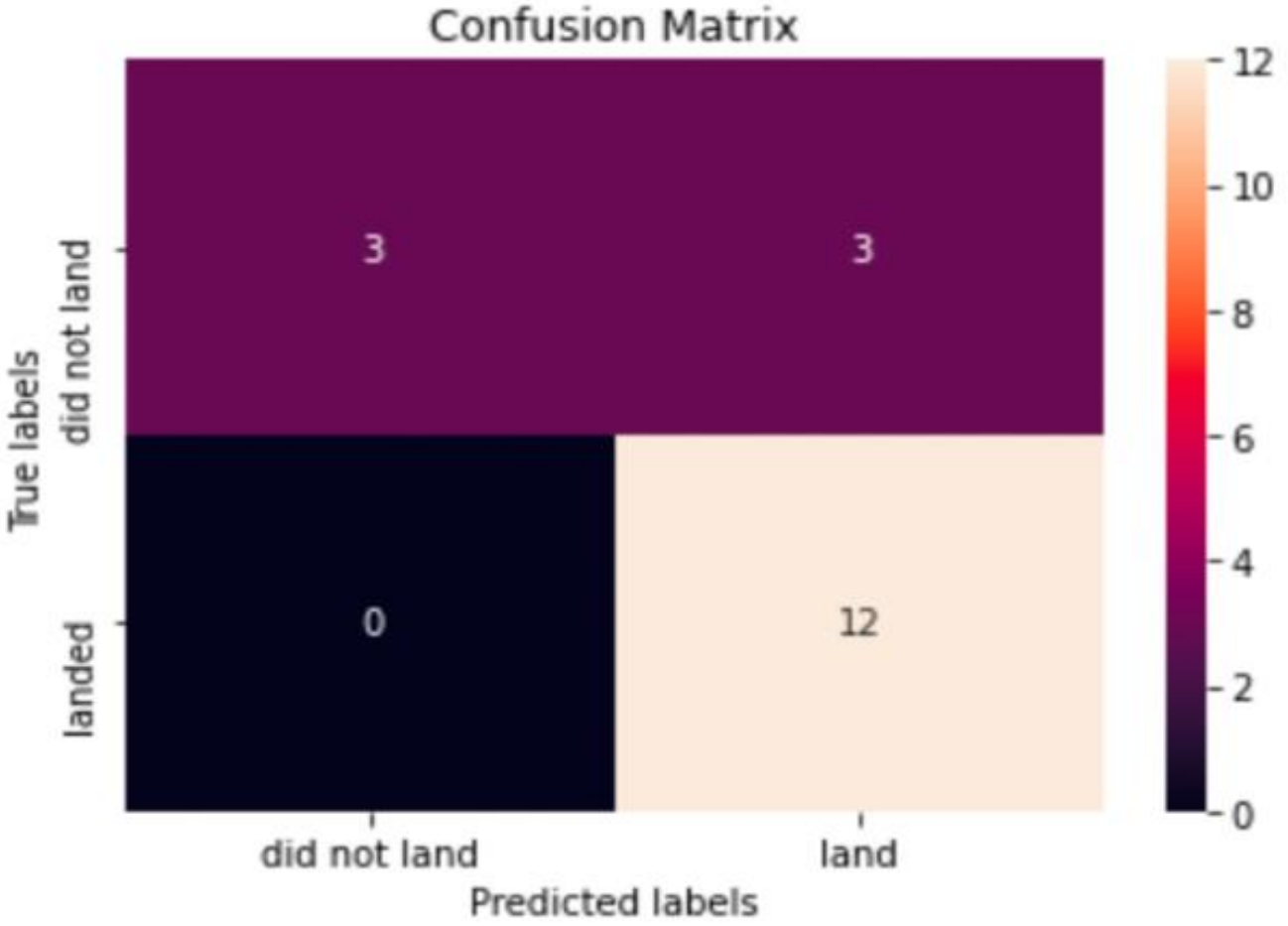
Classification Accuracy

The jaccard_score , F1_Score and accuracy is greater for the decision tree model.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

The false positive is more in our classification model.



Conclusion

1. For this dataset Decision Tree Model algorithm is the best.
2. Flights with a low payload mass has high success rate
3. than flights with a larger payload mass.
4. Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
5. The success rate of launches increases over the years.
6. Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
7. KSC LC-39A has the highest success rate of the launches
8. from all the sites.