Data Science Capstone Project

Tien Pham

June 23, 2022

OUTLINE

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

EXECUTIVE SUMMARY

- Summary of methodologies:
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- Summary of methodologies:
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

INTRODUCTION

- Project background and context:

SpaceX has gained worldwide attention for a series of historic milestones.

It is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010.

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars whereas other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage.

- Problems you want to find the answer:

- Correlations between each rocket variables and successful landing rate.
- Conditions to get the best results and ensure the best successful landing rate.

METHODOLOGY

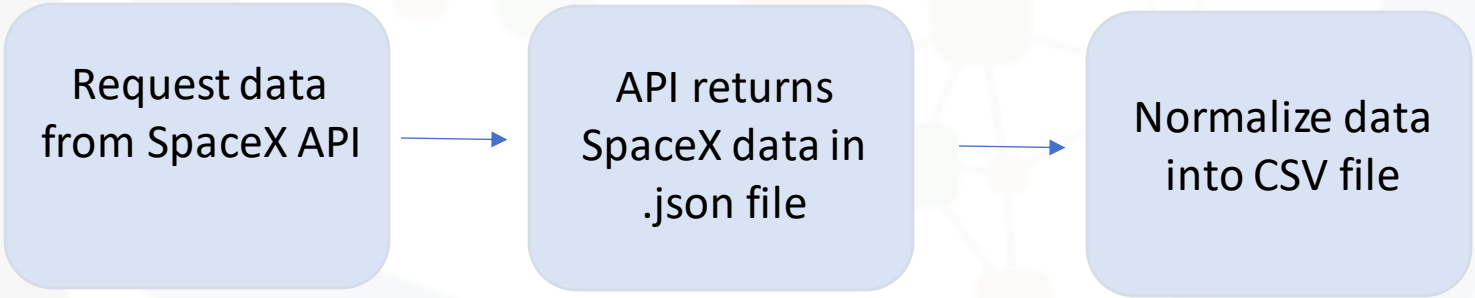
- Data collection methodology:
  - SpaceX API and Web Scraping from api.spacexdata.com/v4/
  - Collect Falcon 9 historical launch records from a Wikipedia

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- Data wrangling:
  - Convert the outcomes into Training Labels to consider the booster successfully or unsuccessfully
- Exploratory data analysis using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

Data Collection SpaceX API

- Making a get request from the SpaceX API.
- Constructing data with columns : FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longtitude, Latitude.
- Dealing with missing values in PayloadMass column using mean value.



Link to GitHub:
https://github.com/dinhlang86/Final_SpaceX_Project/blob/54ece17e56c17792ad18ed55230c3e65a43bafc5/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Making a get request for web scraping data from a table in Wiki page "List of Falcon 9 and Falcon Heavy Launches".
- Constructing data into columns: Flight No., Launch Site, Payload, Payload mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time.

Link to GitHub: https://github.com/dinhlang86/Final_SpaceX_Project/blob/54ece17e56c17792ad18ed55230c3e65a43bafc5/jupyter-labs-webscraping.ipynb

Data Wrangling

- There are several different cases where the booster landed successfully or unsuccessfully:
  - True Ocean: successfully landed to specific region of the ocean.
  - False Ocean: unsuccessfully landed to specific region of the ocean.
  - True RTLS: successfully landed to a ground pad.
  - False RTLS: unsuccessfully landed to a ground pad.
  - True ASDS: successfully landed on a drone ship.
  - False ASDS: unsuccessfully landed on a drone ship.
- Converting the outcomes into Training Labels with 1 means the booster successfully landed and 0 means it was unsuccessful.
- Link to GitHub:

https://github.com/dinhlang86/Final_SpaceX_Project/blob/54ece17e56c17792ad18ed55230c3e65a43bafc5/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter chart:
  - Relationship between FlightNumber and LaunchSite.
  - Relationship between Payload and LaunchSite.
  - Relationship between FlightNumber and Orbit type.
  - Relationship between Payload and Orbit type.
- Bar chart:
  - Relationship between success rate of each Orbit type.

EDA with Data Visualization

- Line chart:
  - Relationship between Year and average launch success trend.
- Link to GitHub:

https://github.com/dinhlang86/Final_SpaceX_Project/blob/54ece17e56c17792ad18ed55230c3e65a43bafc5/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Loading the dataset into the corresponding table in a Db2 database.
  - 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 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 which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass.

EDA with SQL

- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Link to GitHub:

https://github.com/dinhlang86/Final_SpaceX_Project/blob/54ece17e56c17792ad18ed55230c3e65a43bafc5/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Building an interactive map with Folium

- Mark all launch sites to the map.
  - All launch sites are in proximity to the Equator line.
  - All launch sites are in very close proximity to the coast.
- Mark the success / failed launches for each site on the map.
- Calculate the distance between a launch site to its proximities
  - Launch sites are in close proximity to railways.
  - Launch sites are in close proximity to highways.
  - Launch sites are in close to coastline.
  - Launch sites keep certain distance away from cities.

Link to GitHub:

https://github.com/dinhlang86/Final_SpaceX_Project/blob/54ece17e56c17792ad18ed55230c3e65a43bafc5/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The Dashboard application contains:
  - Launch sites dropdown component.
  - Show success pie chart base on selected site dropdown.
  - Range Slider component to select Payload.
  - Show success payload scatter chart base on selected site dropdown and range of payload.

Link to GitHub:

https://github.com/dinhlang86/Final_SpaceX_Project/blob/54ece17e56c17792ad18ed55230c3e65a43bafc5/interactive_dashboard_ploty.py

Predictive analysis (Classification)

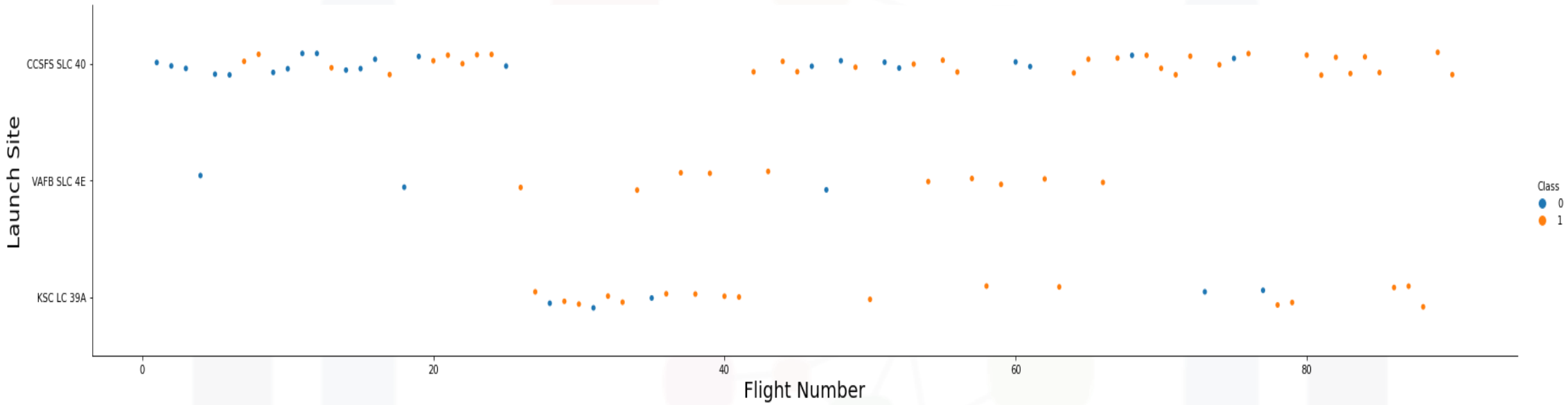
- Perform Exploratory Data Analysis and determine Training Labels:
  - Create a column for the class
  - Standardize the data
  - Split into training data and test data
- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression:
  - Find the method performing best using test data

Link to GitHub:

https://github.com/dinhlang86/Final_SpaceX_Project/blob/54ece17e56c17792ad18ed55230c3e65a43bafc5/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

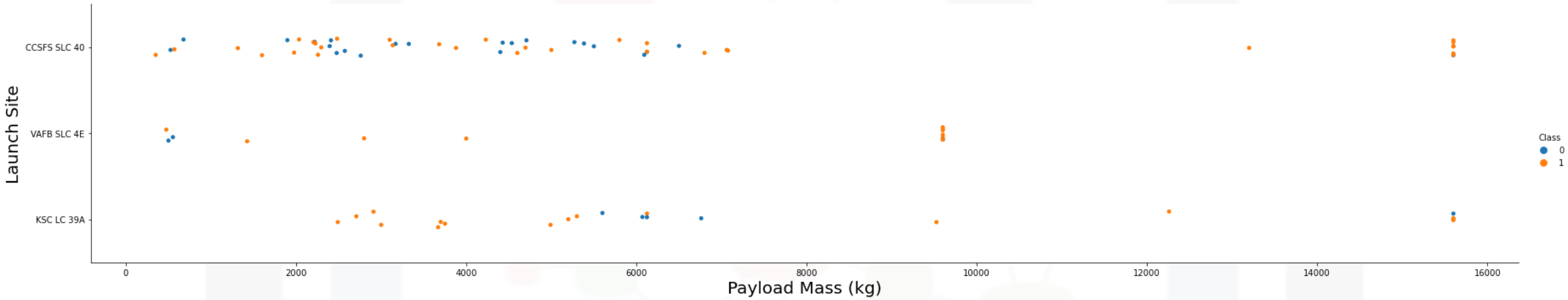
Results from EDA

Relationship between Flight Number and Launch Site



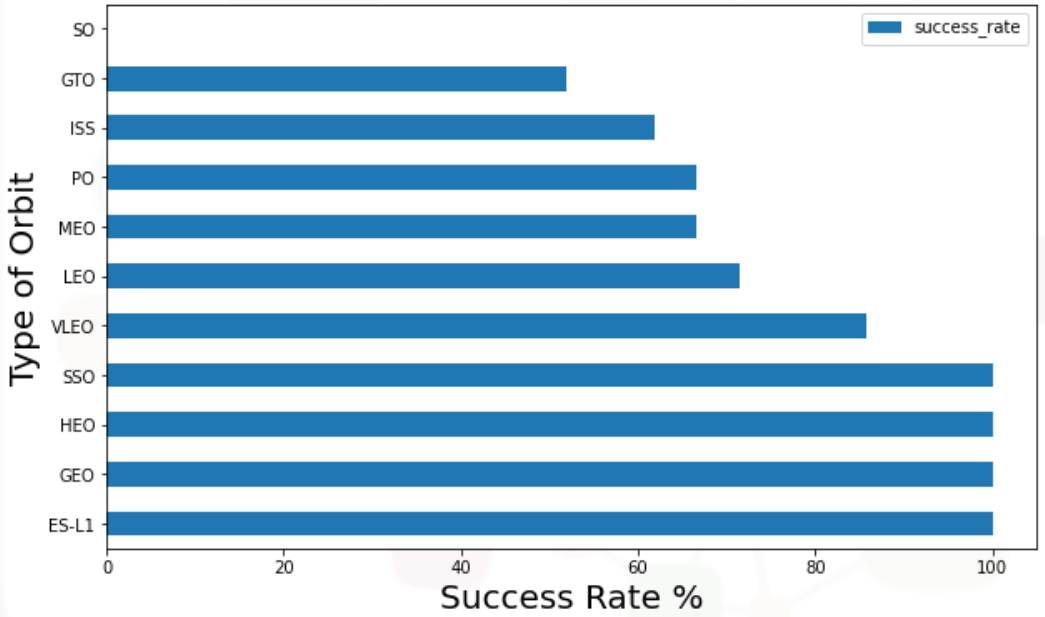
- CCSFS SLC 40 has almost successful with Flight Number over 80
- VAFB SLC 4E has almost successful with Flight Number over 50
- KSC LC 39A has almost successful with Flight Number over 90
- Class 0 (blue) represents unsuccessful launch, class 1(orange) represents successful launch

Relationship between Payload and Launch Site



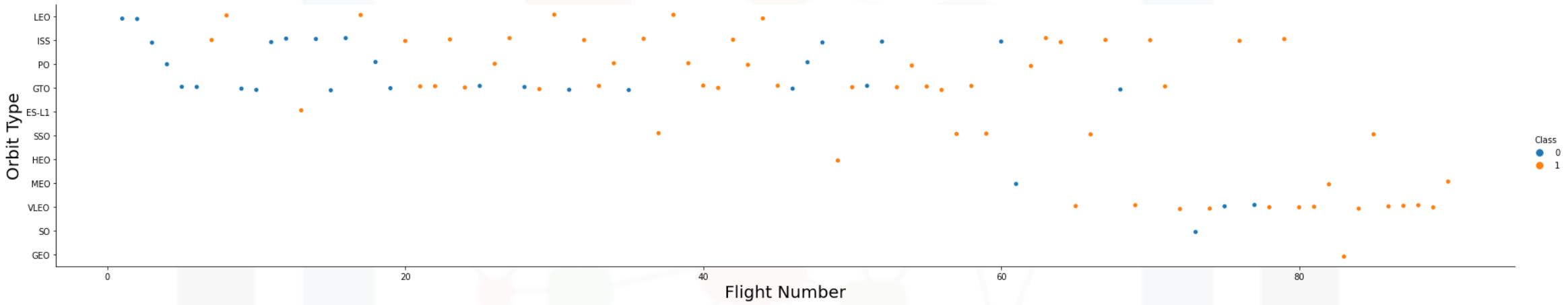
- The VAFB-SLC 4E launch site there are no rockets launched for heavy payload mass(greater than 10000).
- It seems difficult to show the relationship base on this figure because no clear pattern can be found between Payload and Launch Site.
- Class 0 (blue) represents unsuccessful launch, class 1(orange) represents successful launch.

Relationship between success rate of each Orbit type



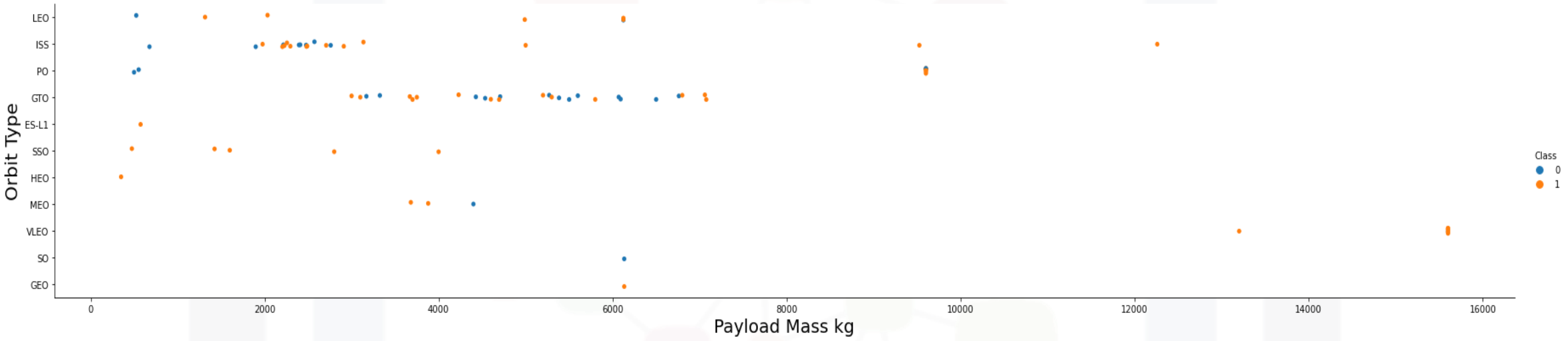
- The high success rate : SSO, HEO, GEO, ES-L1 with 100%.
- SO type is recorded failure in a single attempt.
- GTO is the lowest success rate only 50%.

Relationship between Flight Number and Orbit type



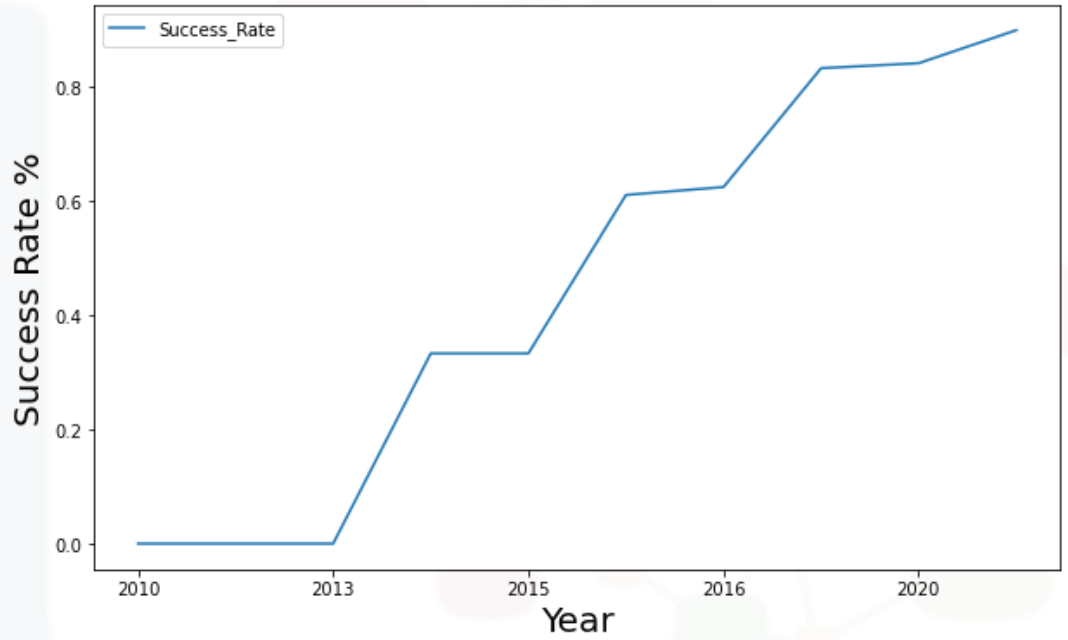
- The LEO orbit the Success appears related to the number of flights.
- There seems to be no relationship between flight number when in GTO orbit.
- Class 0 (blue) represents unsuccessful launch, class 1(orange) represents successful launch.

Relationship between Payload and Orbit type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- For GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.
- Class 0 (blue) represents unsuccessful launch, class 1(orange) represents successful launch.

Relationship between Year and average launch success trend



- The sucess rate since 2013 kept increasing till 2020

EDA with SQL

All Launch Site names

Query

```
%sql select distinct Launch_Site from SPACEXTBL;
```

Result

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

5 records for Launch sites begin with 'CCA'

Query

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

Result

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

Total payload mass by NASA(CRS)

Query

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)';
```

Result

sum(PAYLOAD_MASS__KG_)
45596

Average payload mass by booster version F9 v1.1

Query

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

Result

avg(PAYLOAD_MASS__KG_)
2928.4

The first successful landing outcome in ground pad

Query

```
%sql select min(substr(Date, 7, 4) || substr(Date, 4, 2) || substr(Date, 1, 2)) as Date from SPACEXTBL where "Landing _Outcome" = "Success (ground pad)";
```

Result

Successful Drone Ship landing with payload between 4000 and 6000

Query

```
%sql select Booster_Version from SPACEXTBL where "Landing _Outcome" = "Success (drone ship)" and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

Result

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

Total successful and failure mission outcomes

Query

```
%sql select Mission_Outcome, count(*) from SPACEXTBL group by Mission_Outcome;
```

Result

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

Booster Versions have carried maximum payload mass

Query

```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

Result

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

Failure Drone Ship landing in 2015

Query

```
%%sql select substr(Date, 4, 2) as Month, Booster_Version, Launch_Site, "Landing _Outcome"
from SPACEXTBL where substr(Date, 7, 4) = "2015" and "Landing _Outcome" = "Failure (drone ship)";
```

✓ 0.8s

Result

Month	Booster_Version	Launch_Site	Landing _Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Count the successful landing between 04-06-2010 and 20-03-2017

Query

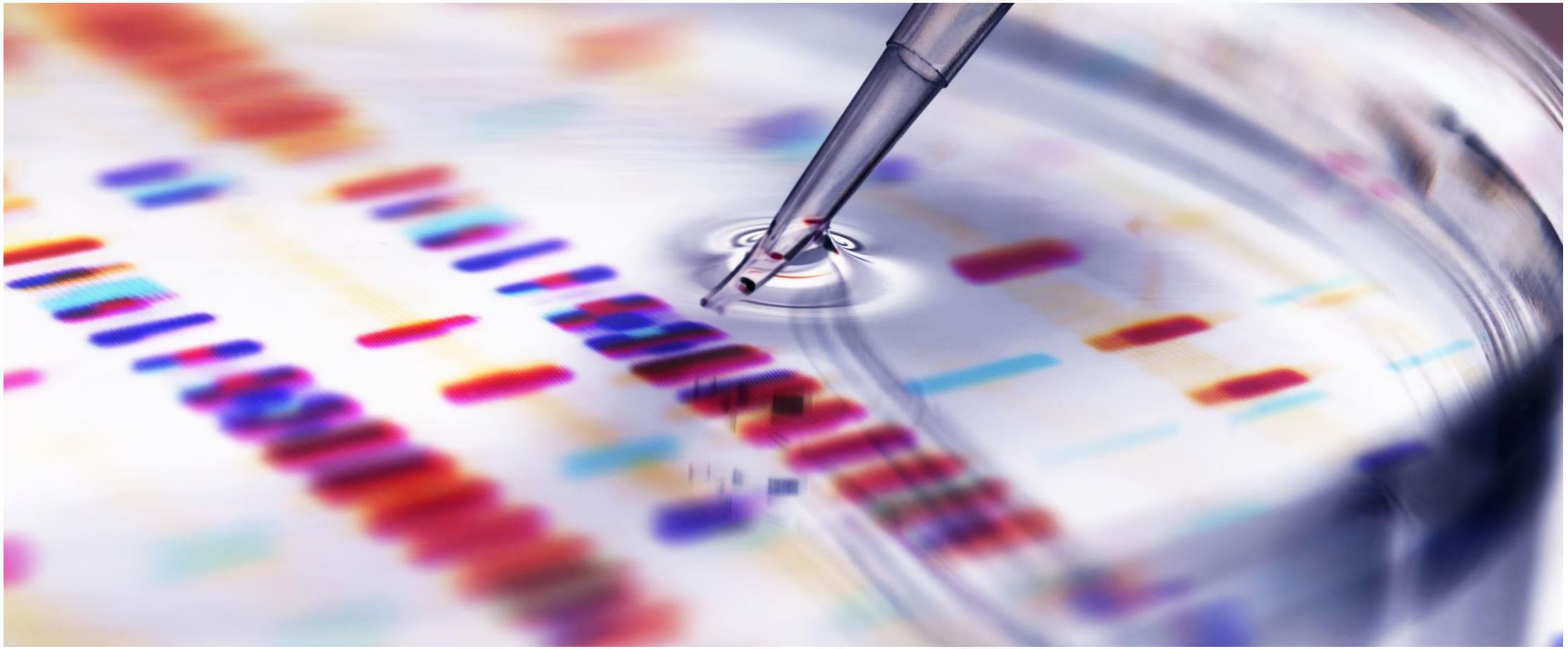
```
%%sql select "Landing _Outcome", count("Landing _Outcome") as Total_Success from SPACEXTBL
where ((substr(Date,7,4) || "-" || substr(Date,4,2) || "-" || substr(Date,1,2)) between "2010-06-04" and "2017-03-20") and "Landing _Outcome" like "%Success%"
group by "Landing _Outcome"
order by Total_Success desc;
```

✓ 0.5s

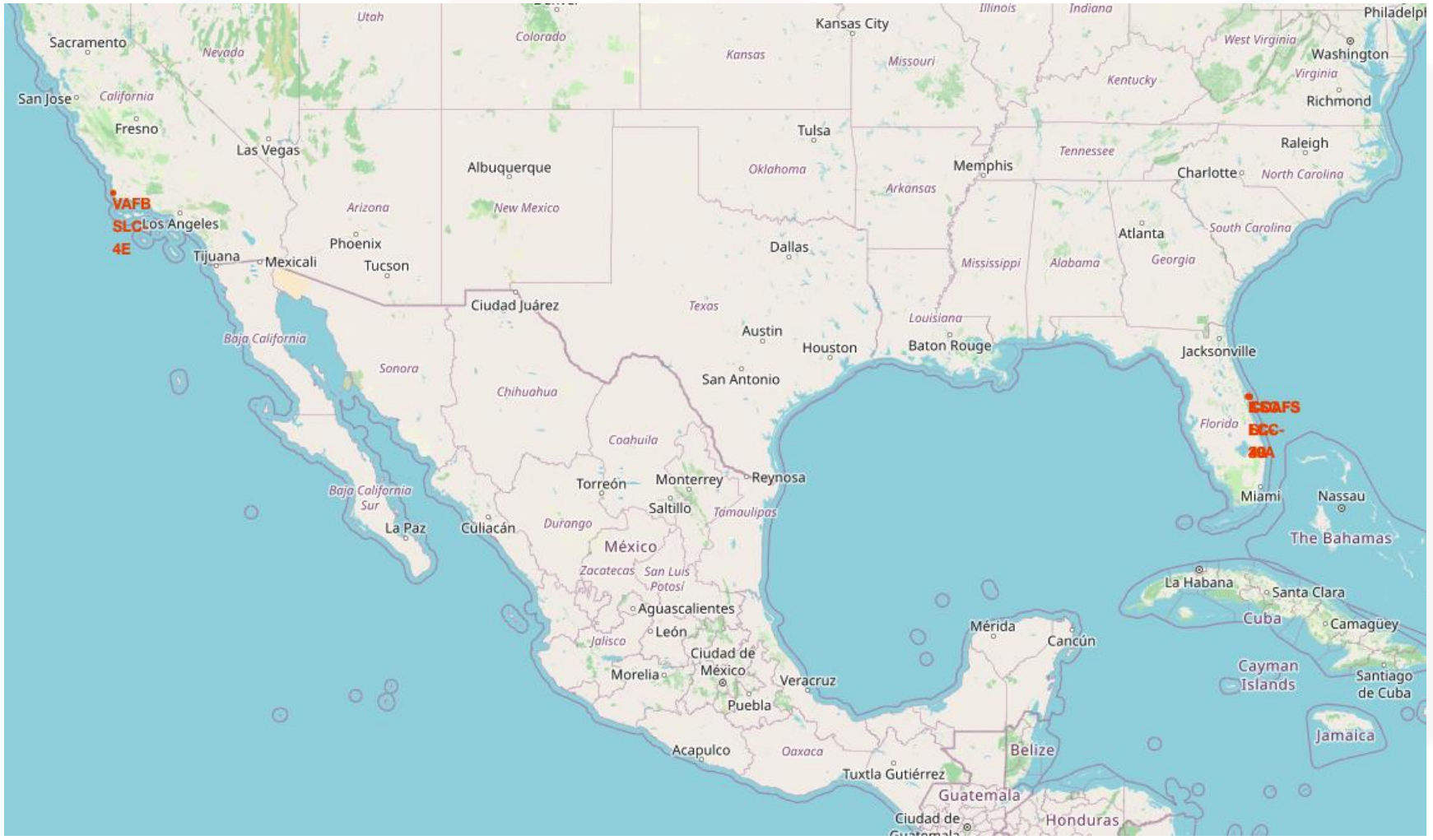
Result

Landing _Outcome	Total_Success
Success (drone ship)	5
Success (ground pad)	3

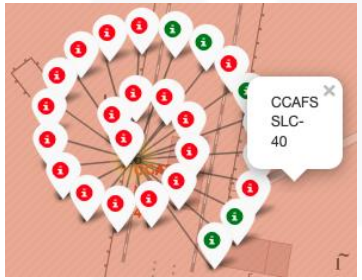
Launch Sites Proximities Analysis

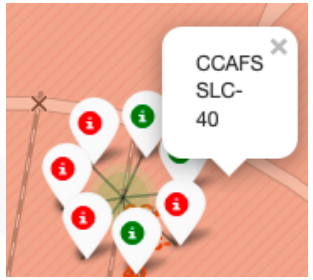


All Launch Sites' Location



Color-labeled Launch Sites



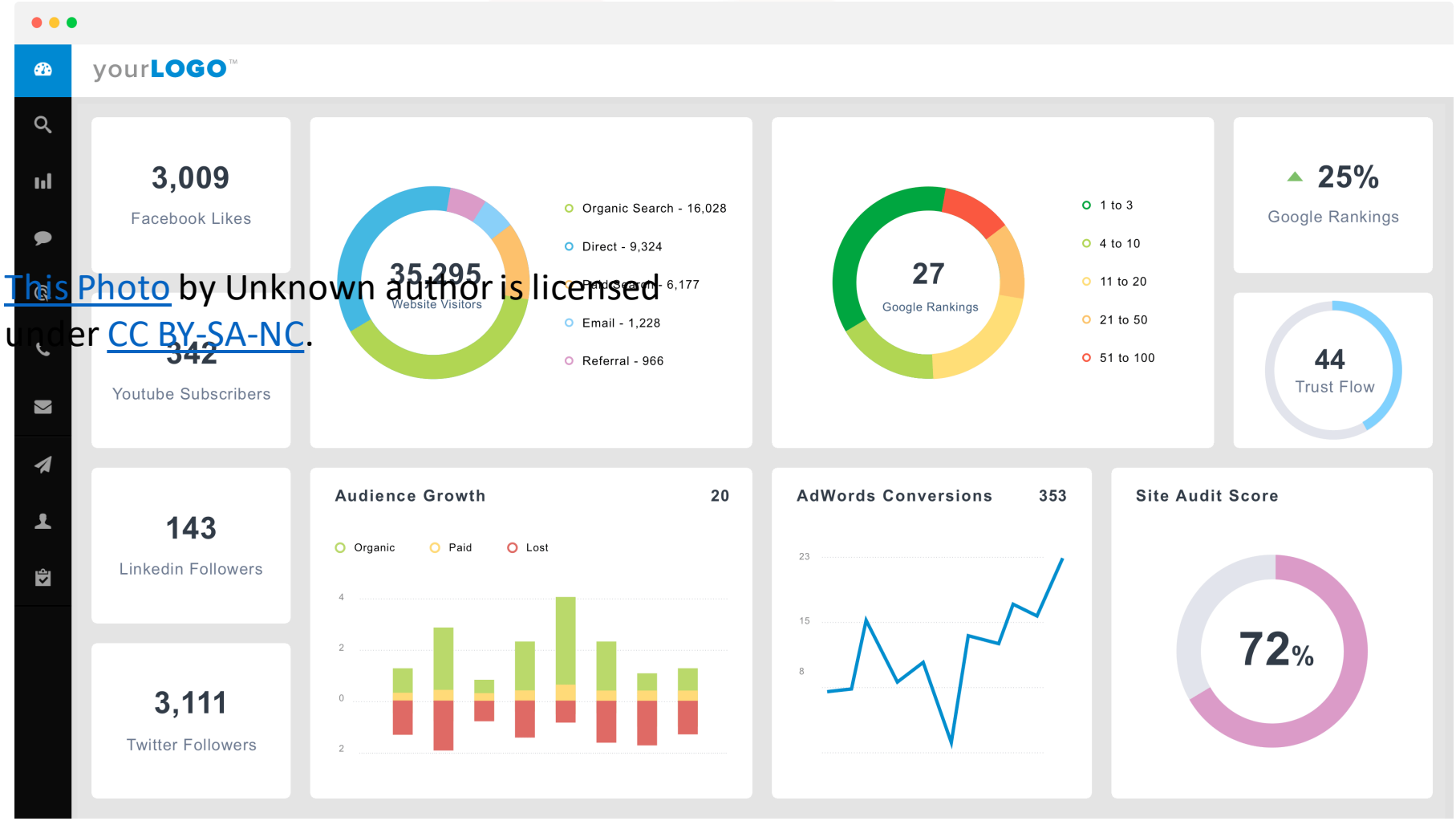


By click the Marker, successful landing and failed landing are shown

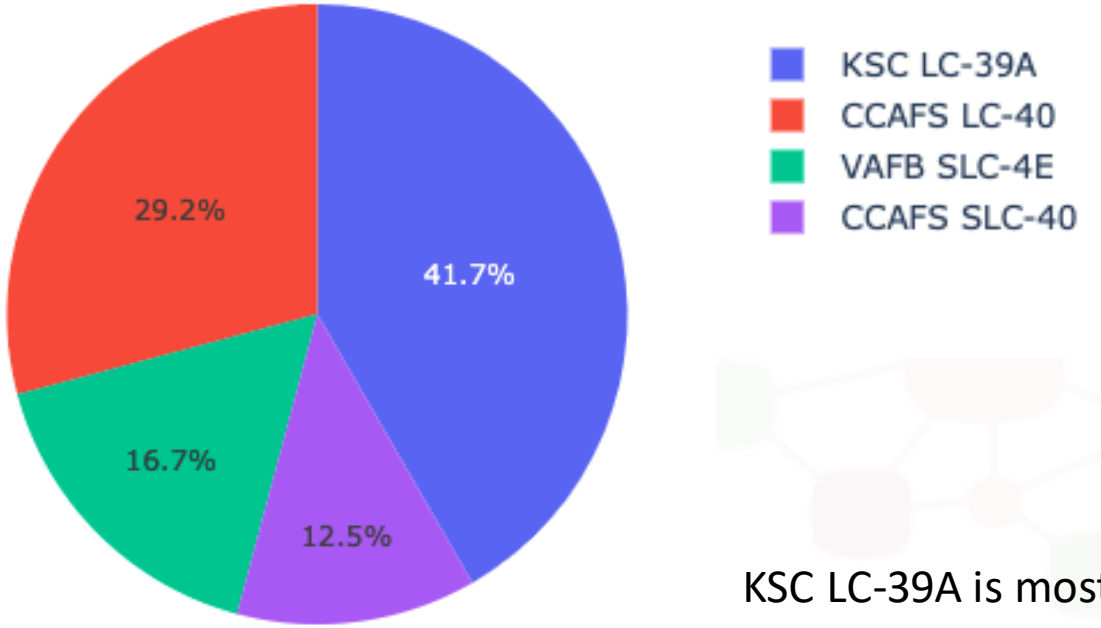
Proximities of Launch Site

The Launch Site is close to coastline, highway, railway for transportation and far away from cities to avoid the threat.

Build a Dashboard with Plotly Dash

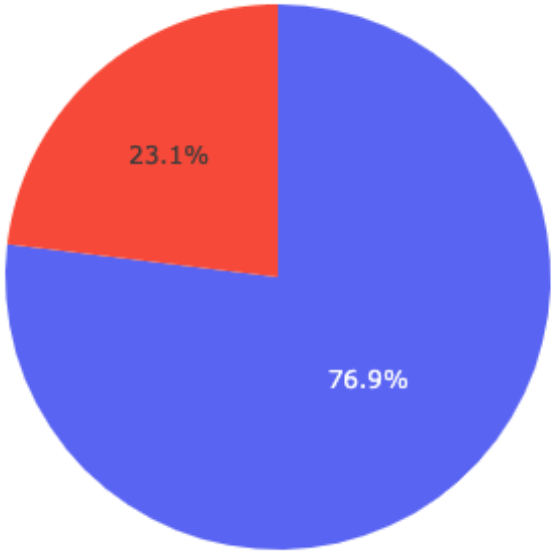


Total Success by all Sites



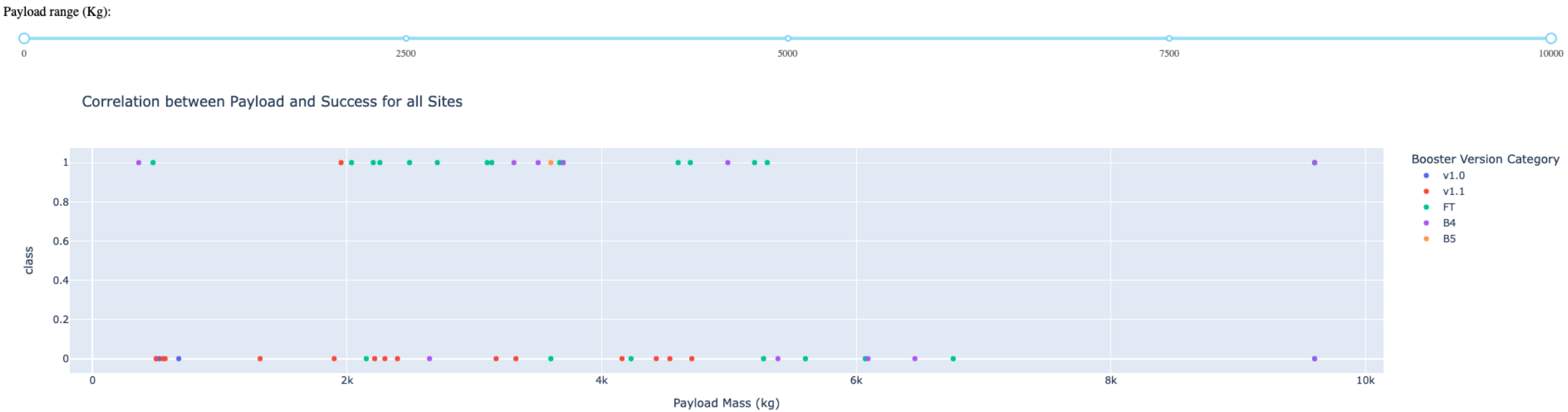
KSC LC-39A is most successful launches than the other ones.

The highest launch success rate



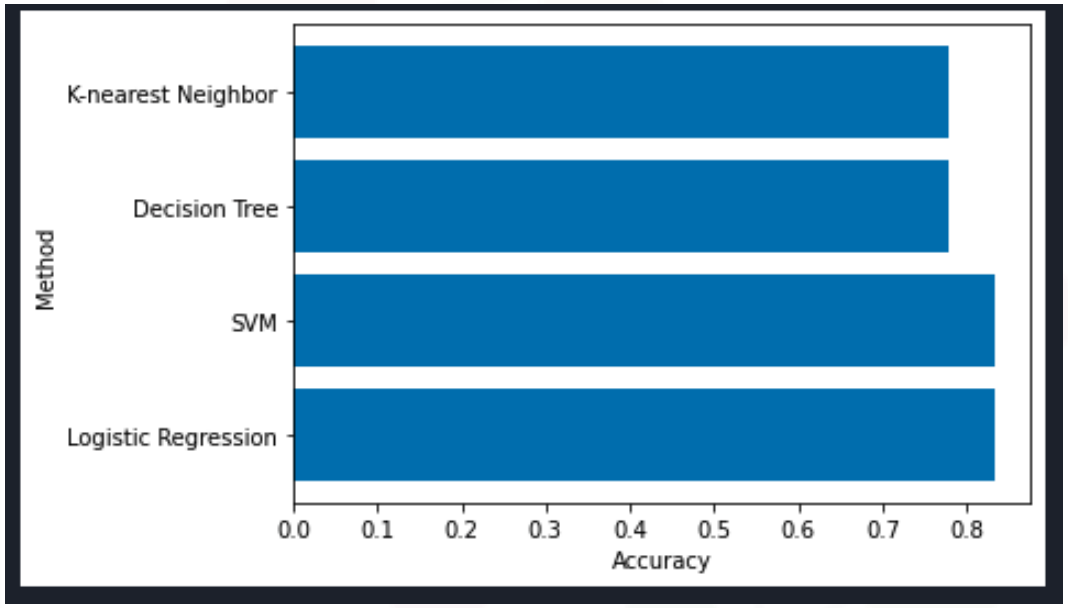
KSC LC-39A has the highest success launch with 10 successful landing (76,9%) and 3 failed landing (23.1%)

Payload and Outcome Launch Scatter chart for all sites



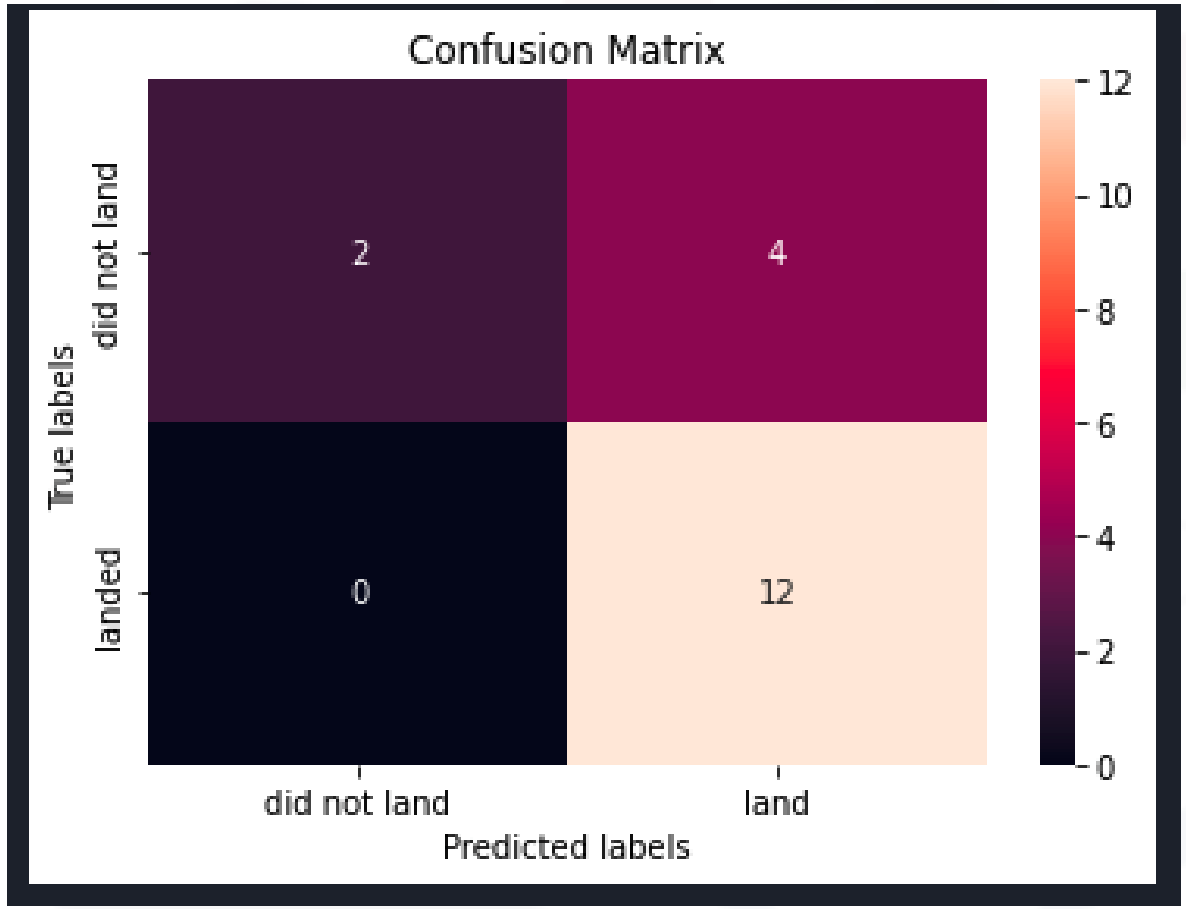
- Range from 2k to 5k has the highest launch success rate.
- Range from 7k to 10k has the lowest launch success rate.
- F9 booster version FT has the highest launch success rate.

Predictive Analysis (Classification)



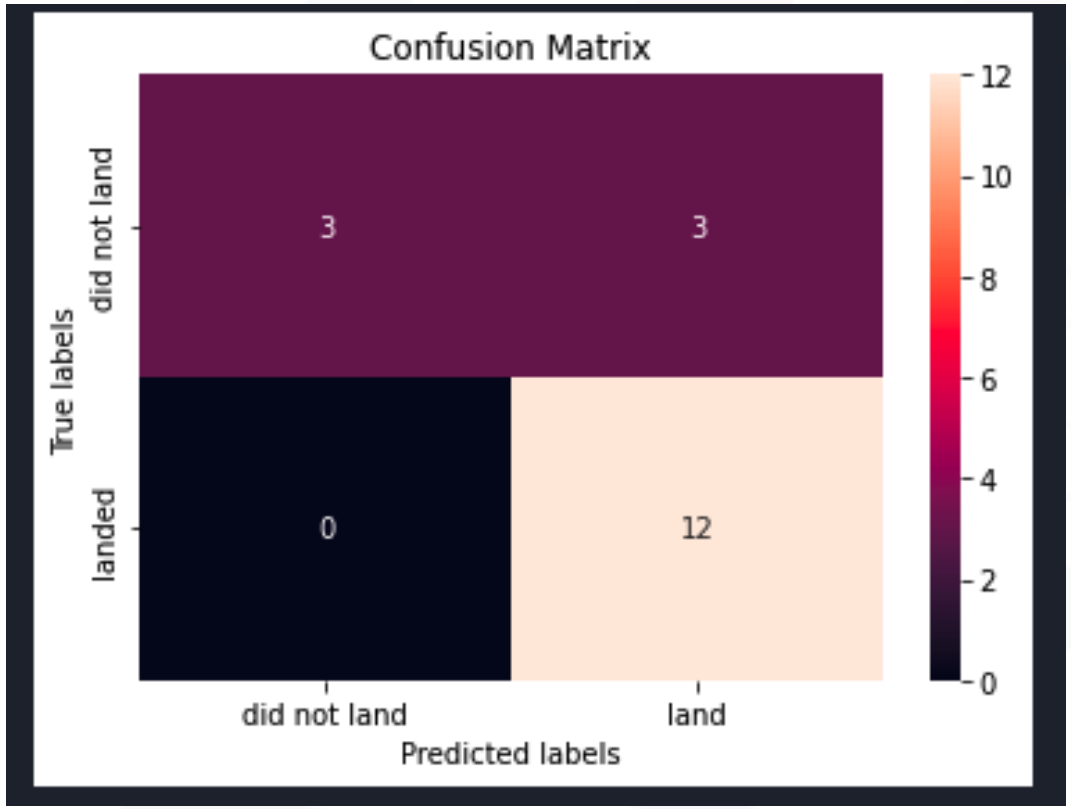
- Only 18 test samples using for test data.
- Logistic Regression and SVM have the accuracy at 0.8333333333333334
- Decision Tree and K-nearest Neighbor have the accuracy at 0.7777777777777778

Confusion Matrix for K-nearest Neighbor



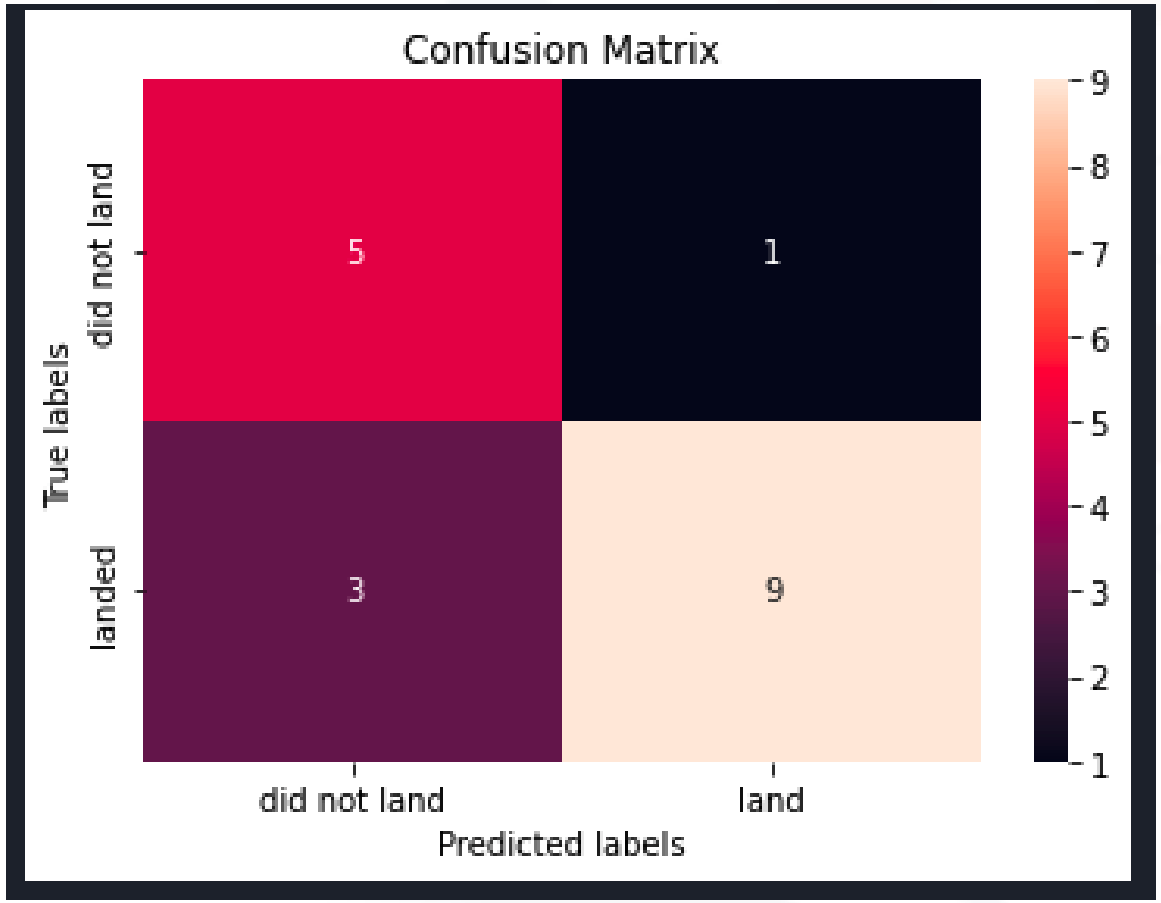
The model predicts well for successful landing

Confusion Matrix for SVM and Logistic Regression



The model predicts well for successful landing.

Confusion Matrix for Decision Tree



The model predicts quite well for successful landing.

CONCLUSION

- As the Flight Number has increased, the success landing has increased.
- The success rate kept increasing from 2013 until 2020.
- Orbit types: SSO, HEO, GEO, ES-L1 have the success rate at 100%.
- The Launch Site is close to railway, highway, coastline but far away from cities.
- The launch success rate at low weighted payload has higher success rate than high weighted payload.
- Logistic Regression, SVM, K-nearest Neighbor, Decision Tree can be considered for training dataset.