Understanding Space X through Data Science

Gerald Gaitos
02/18/2024

Outline

- Executive Summary
- Table of Contents
- Introduction
- Methodology
- Results
- Discussion
- Conclusion
- Appendix

Executive Summary

This project presents the application of data science in understanding data on Space X using various tools and methodologies in collecting, processing, displaying, and analyzing of the dataset.

Introduction

The project’s main objective is to forecast the likelihood of a successful landing for the Falcon 9 first stage.

SpaceX's website says launching a Falcon 9 rocket costs $62 million, much less than others that charge over $165 million. This is because SpaceX can reuse the first stage. If we can predict if the stage will land properly, we can figure out the total launch cost. This info could be useful for a company trying to compete with SpaceX in rocket launches.

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?

- Does the rate of successful landings increase over the years?

- What is the best algorithm that can be used for binary classification in this case?

Methodology

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

Data Collection

- Datasets are retrieved from REST Space X API and web scraping Wikipedia.

- For the Space X REST API, this url was utilized: api.spacexdata.com/v4/

- For the web scraping from Wikipedia, this url was utilized: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

REST Space X API

1. Requesting rocket launch data from the url.
2. Decoding the response content using .json() and turning it into a dataframe using .json_normalize().
3. Requesting needed information about the launches from SpaceX API by applying the custom functions.
4. Constructing data obtained into a dictionary.
5. Creating a dataframe from the dictionary.
6. Filtering the dataframe to only include Falcon 9 launches.
7. Replacing missing values of Payload Mass column with calculated mean for the column.
8. Exporting the data into CSV file.

Web Scraping

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

Data Wrangling

From the dataset, cases can either land successfully or not successfully.

Successful cases are labeled as True Ocean, True RTLS, and True ASDS.

Failed cases are labeled as False Ocean, False RTLS, and False ASDS.

1. Exploratory data analysis performed and determined training labels.
2. The number of launches on each site is calculated.
3. The number and occurrence of each orbit is calculated.
4. The number and occurrence of the mission outcome per orbit type is calculated.
5. Outcome column is created for a landing outcome.
6. Data exported into a CSV file.

EDA with Data Visualization

Various types of plotting are created and displayed for visualization:

1. Flight Number vs. Payload Mass
2. Flight Number vs. Launch Site
3. Payload Mass vs. Launch Site
4. Orbit Type vs. Success Rate
5. Flight Number vs. Orbit Type
6. Payload Mass vs. Orbit Type
7. Success Rate Yearly Trend

Scatter plots display relationship between variables. Bar charts display comparisons among discrete categories. Line charts show trends in data over time.

EDA with SQL

Various SQL queries are performed to retrieve and understand the data from the dataset.

- To display the unique launch sites
- To display the records of launch sites that begin with the string ‘CCA’
- To display the total payload mass launched by NASA (CRS)
- To display the average payload mass carried by booster version F9 v1.1
- To list the date of the successful landing outcome in ground pad
- To list the boosters that have drone ship success and payload mass of greater than 4000, but less than 6000
- To list the total number of successful and failure mission outcomes
- To list the booster versions that carry the maximum payload mass
- To list the month, failed landing outcomes in drop ship, booster version, and launch site for the year 2015
- To rank the total number of successful landing outcome between 04/06/2010 and 20/03/2017 in descending order

Building an interactive Map through Folium

The map created in this project is centered at the NASA Johnson Space Center in Houston, Texas with the following markers:

1. Red circle at the NASA Johnson Space Center's location with label displaying its name.
2. Red circles at each launch site coordinate with label showing their names.
3. The grouping of points in a cluster to display multiple and different information for the coordinates.
4. Markers to display successful and failed landings.
5. Markers to display line plot with the distance between launch site to key locations, such as railway, highway, coastway, and the city.

Build a Dashboard with Plotly Dash

The dashboard created contains dropdown, pie chart, range slider and scatter plots for more interactive and effective visualization.

The following are the components added to the dashboard:

1. Drop down allowing the user to choose a launch site or all launch sites to display.
2. Pie chart displays the total number of successful and unsuccessful launches for the chosen launch site.
3. Range slider allows the user to select a payload mass in a fixed range.
4. Scatter chart displays the relationship between two variables, such as success and payload mass.

Predictive Analysis

Data Preparation

- Loading dataset
- Normalizing dataset
- Splitting data into training and test sets

Model Preparation

- Selecting the machine learning algorithms
- Setting parameters for each algorithm to GridSearchCV
- Training the GridSearchModel models with training data

Model Evaluation

- Getting the best hyperparameters for each machine learning model
- Computing accuracy for each model with test dataset
- Plotting confusion matrix

Model Comparison

- Comparing the models based on accuracy scores
- Selecting the model with the best accuracy score

Result

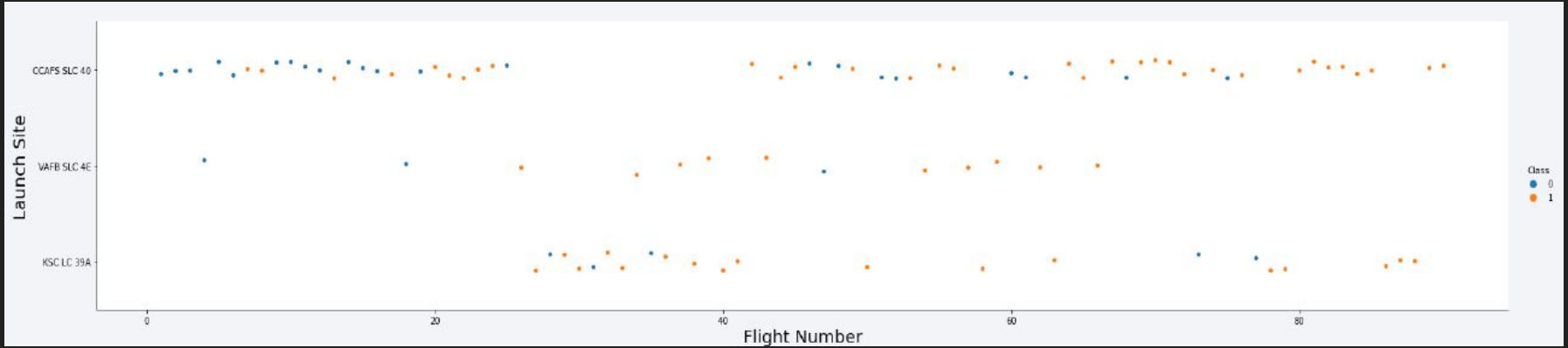
Exploratory data analysis results

Interactive analysis demo in screenshots

Predictive analysis results

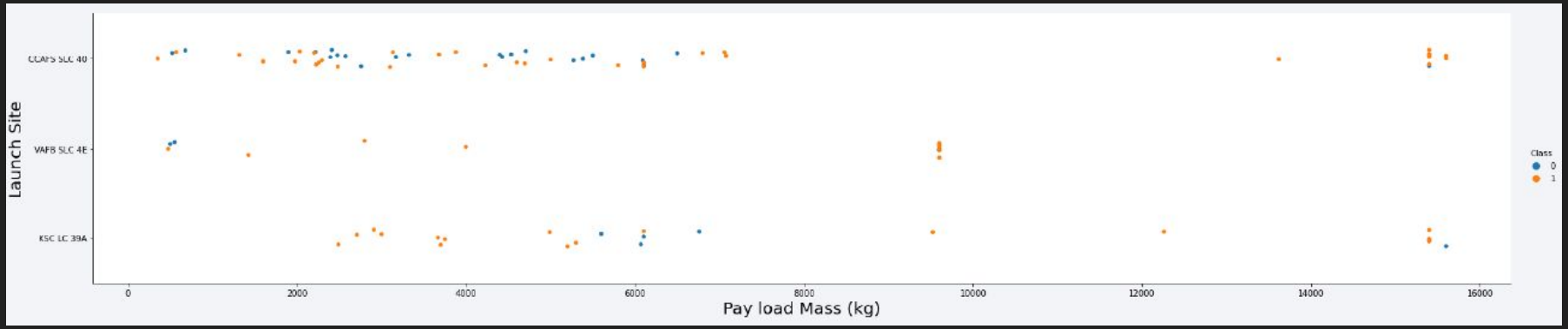
Exploratory Data Analysis Results (Visualization)

Launch Site vs. Flight Number



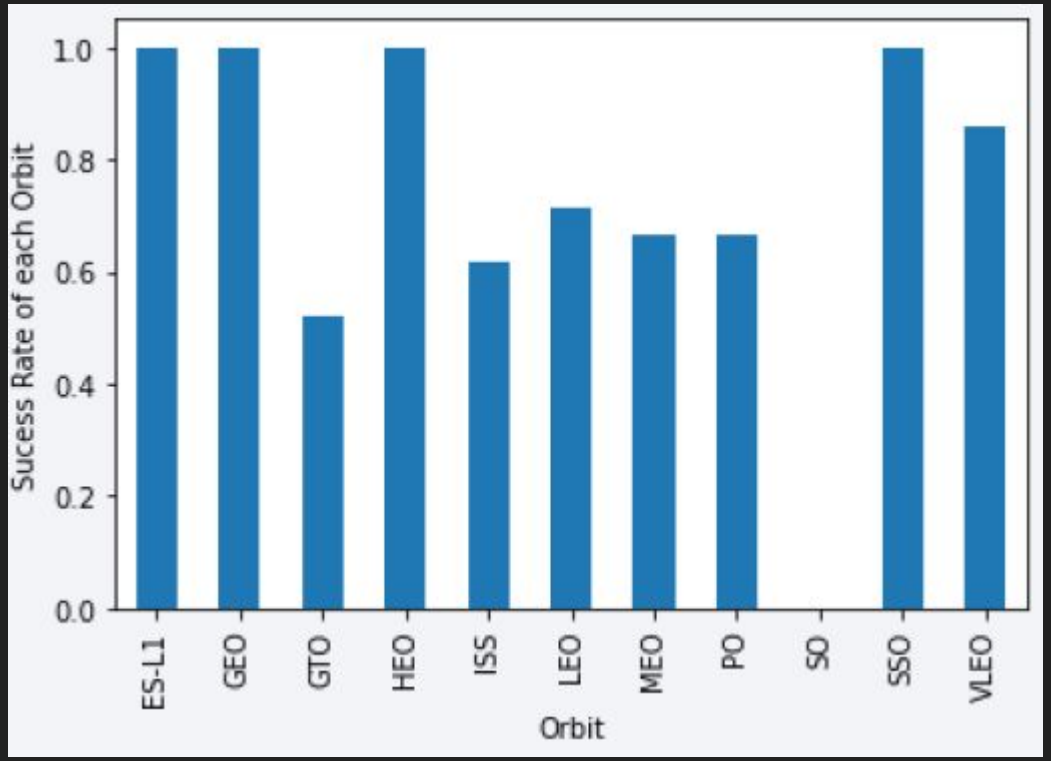
There is an increasing trend of success as the flight number increases or goes above 60.

Launch Site vs. Payload Mass



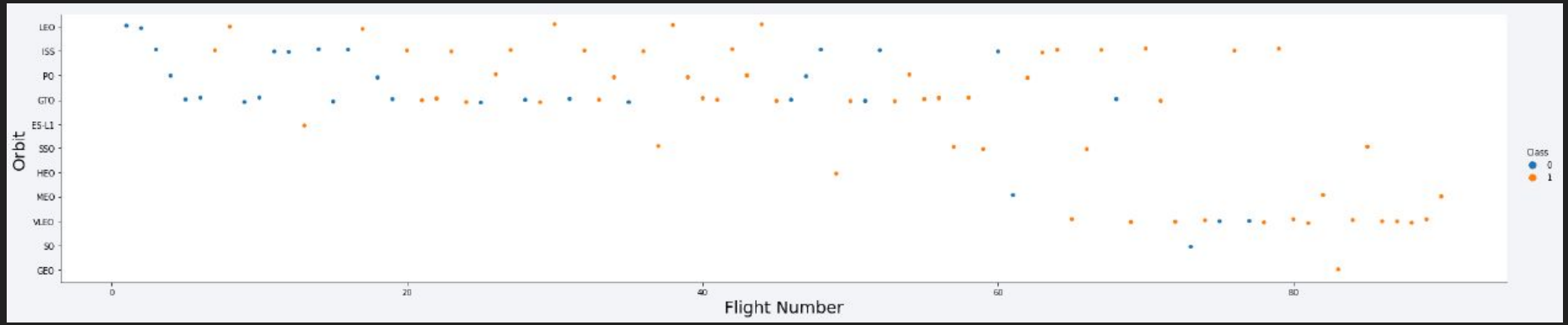
This shows that a heavier payload may be a consideration for a successful landing.

Success Rate vs Orbit



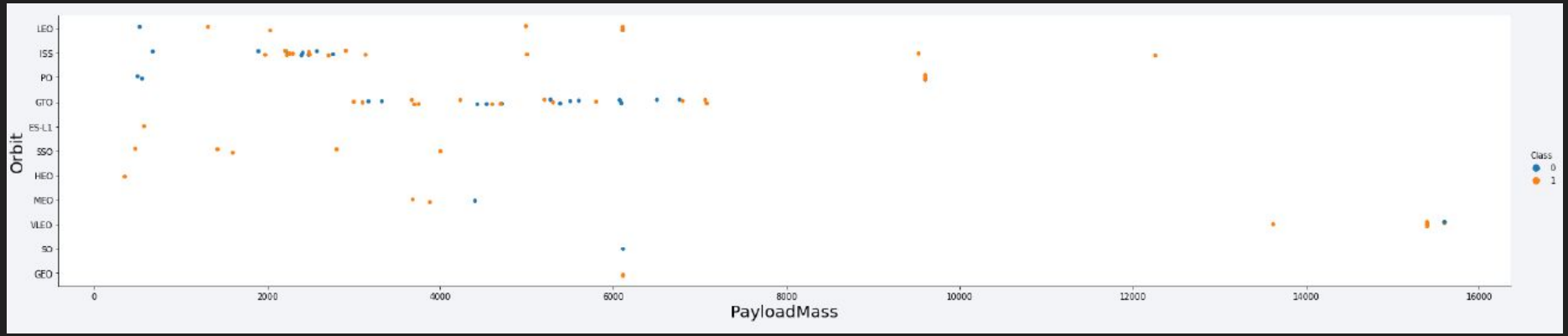
ES-L1, GEO, HEO, and SSO have the highest success rates of 1 (100%)

Orbit Type vs. Flight Number



LEO and VLEO orbits show that an increasing trend of success landing with the increasing trend in flight number. While the remaining orbit types do not show any significant results.

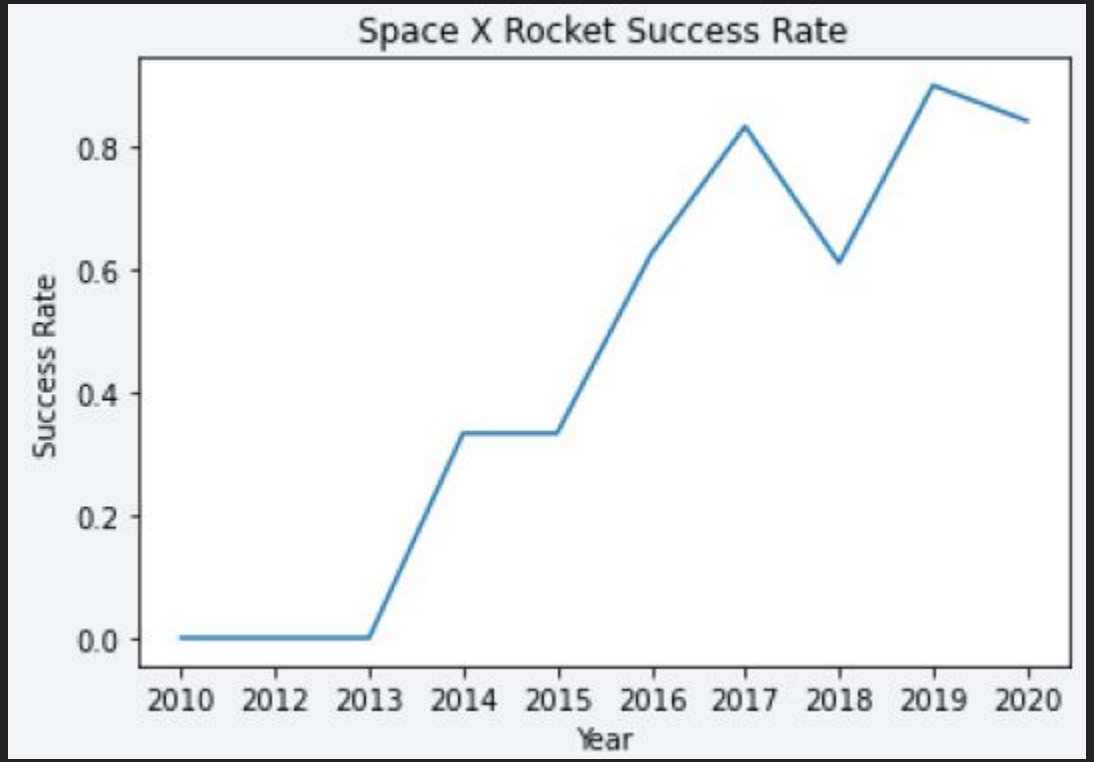
Orbit Type vs. Payload Mass



Payload Mass has a great effect on the success rate of certain launches in certain orbit types.

For example, heavier payloads improve the success rate for the LEO orbit. 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



An increasing trend in the success rate of the launches are shown.

Exploratory Data Analysis Results (SQL)

Launch Site Names

SQL Query:

```
select distinct launch_site from SPACEXDATASET;
```

SQL Result:

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

Launch Site Names with the string ‘CCA’

SQL Query:

```
select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
```

SQL Result:

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

Total Payload Mass

SQL Query:

```
select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
```

SQL Result:

total_payload_mass
45596

Average Payload Mass F9 v1.1

SQL Query:

```
select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
```

SQL Result:

average_payload_mass
2534

First successful ground landing date

SQL Query:

```
select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
```

SQL Result:

first_successful_landing
2015-12-22

Successful drone ship landing with payload mass between 4000 and 6000

SQL Query:

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

SQL Result:

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

Total number of successful and failure mission outcomes

SQL Query:

```
select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

SQL Result:

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

Booster carried maximum payload

SQL Query:

```
select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
```

SQL 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

Launch Records in 2015

SQL Query:

```
select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

SQL Result:

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)

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

SQL Query:

```
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;
```

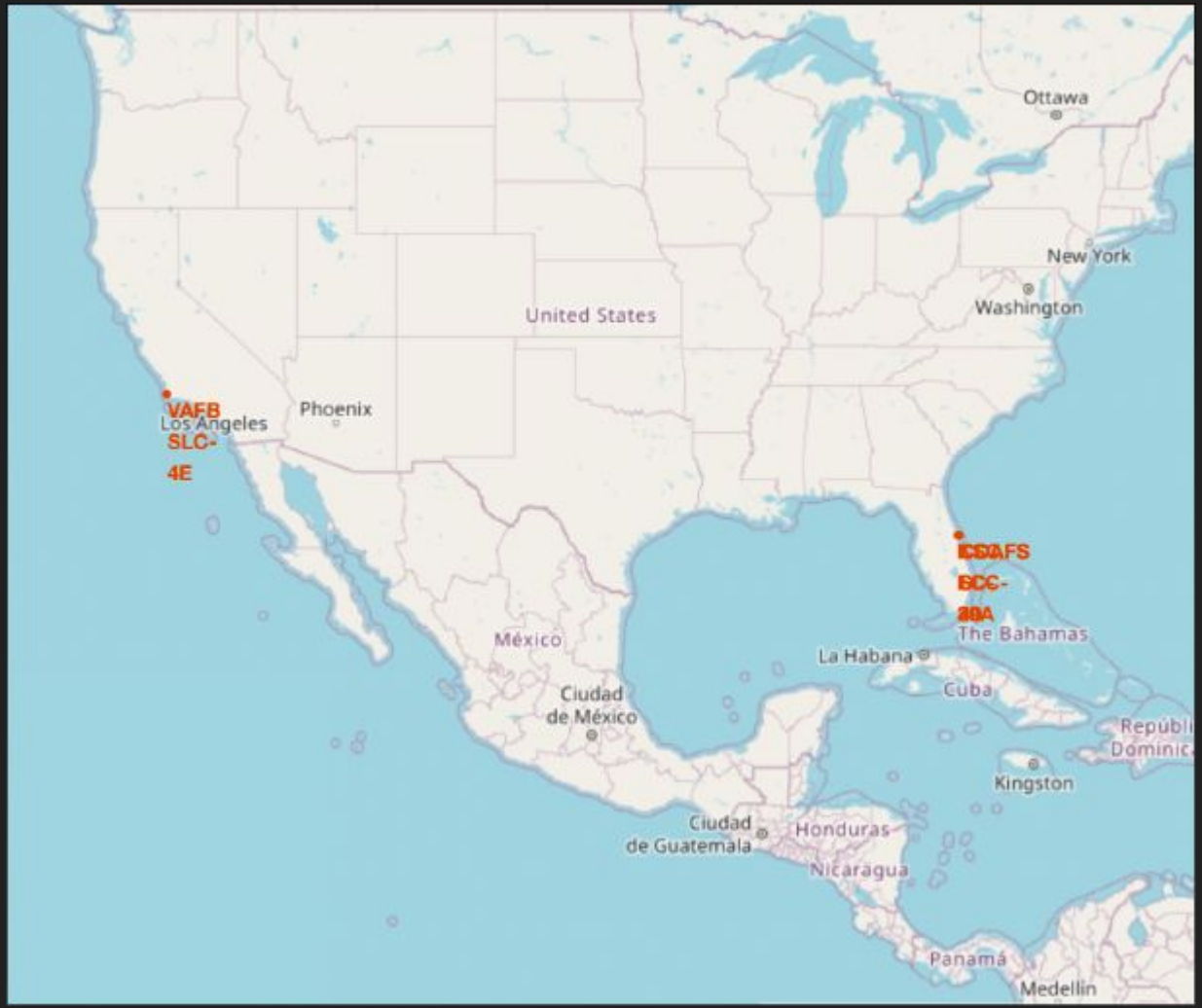
SQL Result:

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

Maps with Folium

Launch Site Locations

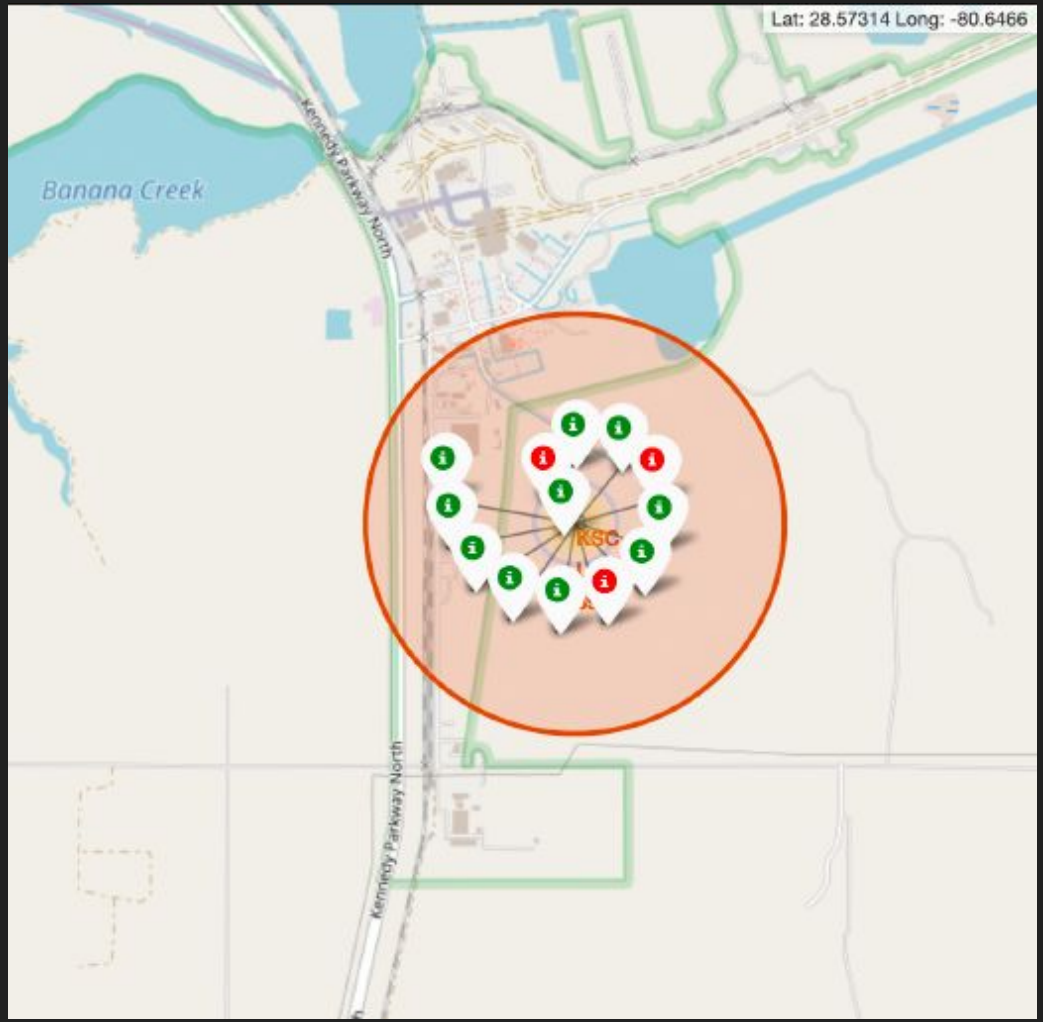
All the launch sites are located at the coastal area and near the equator to minimize the risk of endangering people.



Color-labeled Markers

Unique color labels are used to identify the successful and failure launches.

Among the launch sites, KSC has the highest success rate.

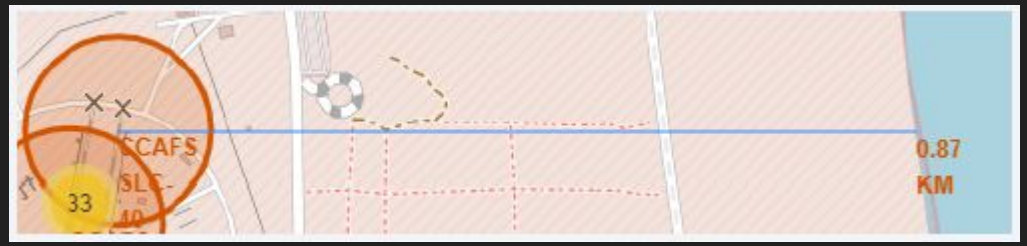


CCAFS Proximities

distance_highway = 0.5834695366934144 km

distance_railroad = 1.2845344718142522 km

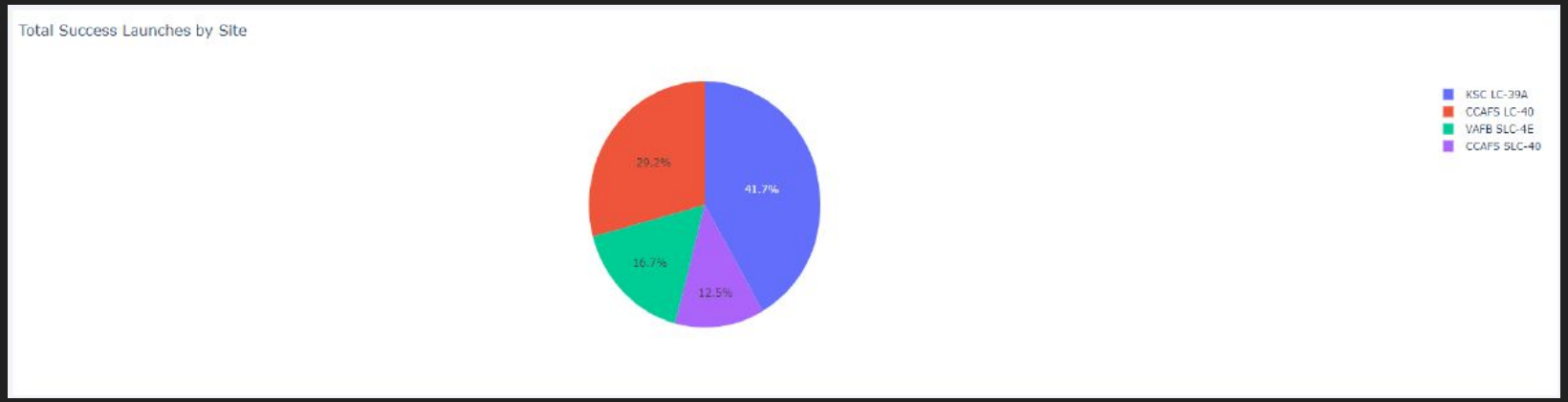
distance_city = 51.434169995172326 km



Maps with Plotly Dash

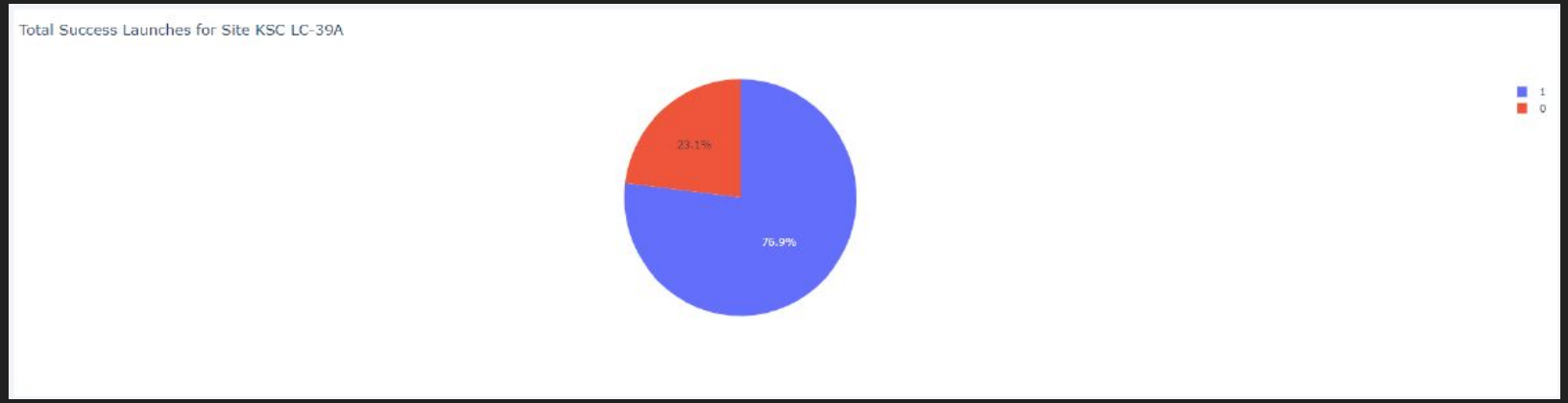
SpaceX Launch Dashboard

Launch success count of all sites



KSC LC-39A has the highest number of successful launches.

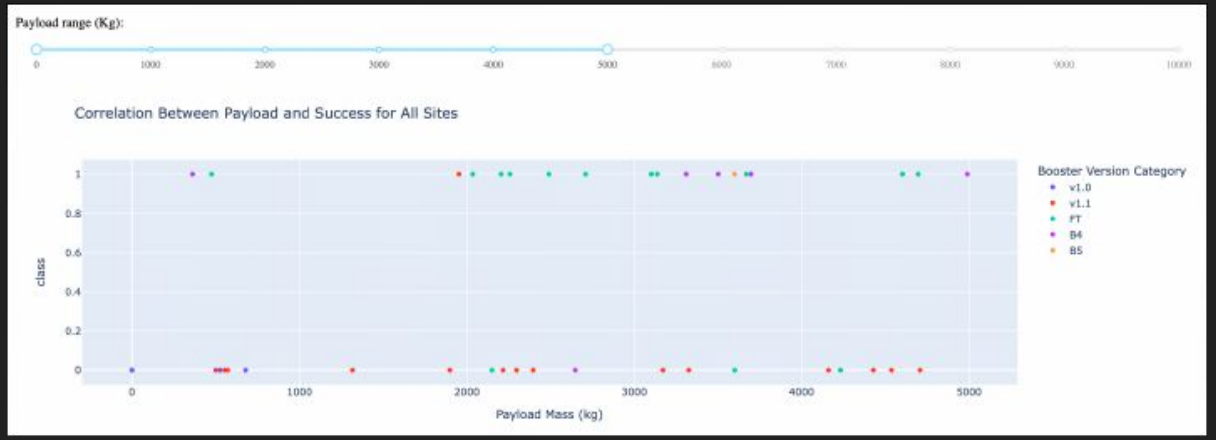
Launch site with the highest launch success ratio

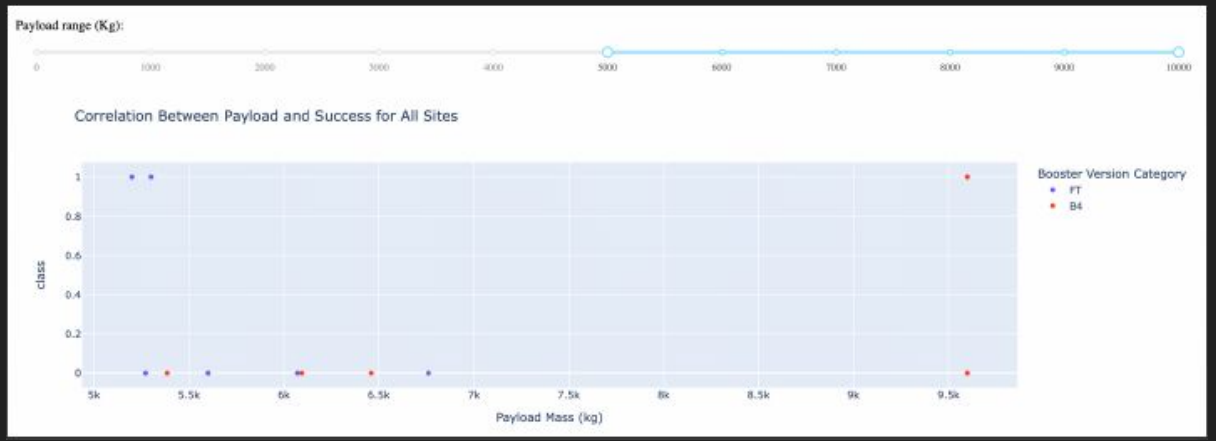


KSC LC-39A has a launch success rate of 76.9%

Payload Mass vs. Launch Outcome for all sites

This shows that most of the successful launch outcomes are in between payload mass of 2000 and 5500 kg.



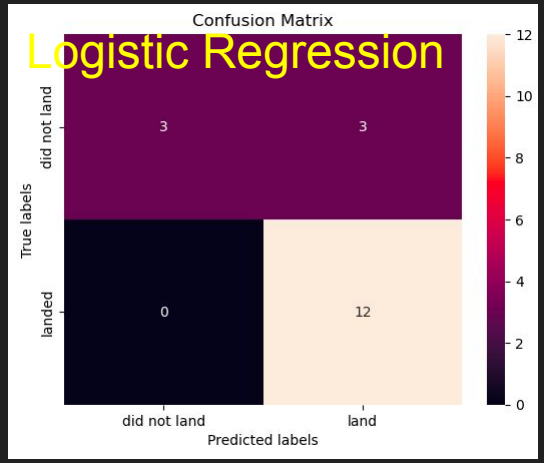


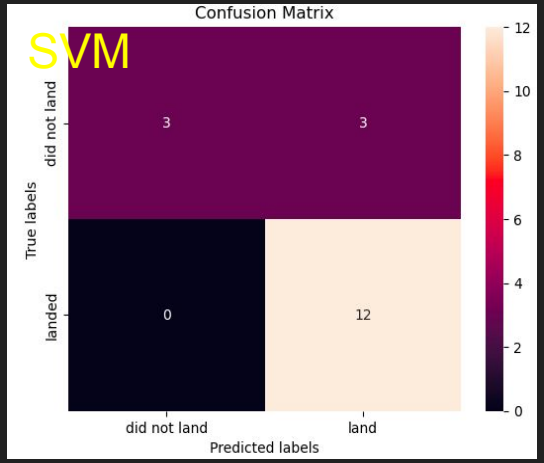
Predictive Analysis

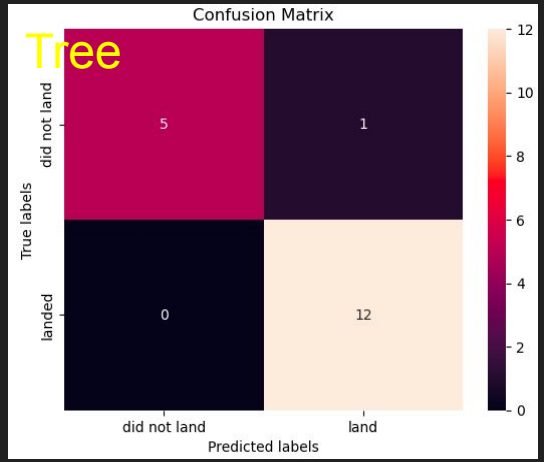
Classification Accuracy

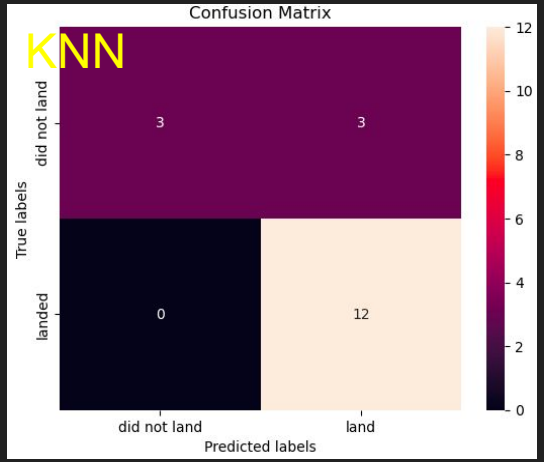
Among the four algorithms, Decision Tree shows the highest accuracy score.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.923077	0.800000
F1_Score	0.888889	0.888889	0.960000	0.888889
Accuracy	0.833333	0.833333	0.944444	0.833333









Conclusion

In conclusion

- Orbits ES-L1, GEO, HEO and SSO have 100% success rate. KSC LC-39A has the highest success rate of the launches from all the sites. The successful launch rate is increasing over the years.
- Launches with a low payload mass show better results than launches with a larger payload mass. Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- Decision Tree Model is the best algorithm for this dataset for predicting the landing outcomes.