Space X Falcon 9 Landing Analysis

APPLIED DATA SCIENCE CAPSTONE

AW, June 21, 2024

OUTLINE

- ❖ Executive Summary
- ❖ Introduction
- ❖ Methodology
- ❖ Results
- ❖ Conclusions

EXECUTIVE SUMMARY

Employed methodologies:

- Data Collection and Wrangling
- Exploratory Data Analysis (EDA)
- Interactive Visual Analytics
- Predictive Analysis

Created visualizations and results:

- EDA results
- Geospatial analytics
- Interactive dashboard
- Predictive analysis of classification models

INTRODUCTION

- **Project background**

Launches with a reusable rocket are advertised to be cost significantly less, around 100 million dollars less, than that of competitors. This cost advantage is because the first stage of the rocket can be recovered and flown again. Knowing if the first stage lands successfully is key to estimating launch costs, since a successful landing allows for reuse.

- **Key Inquiry**

Based on historical data, can we determine if the first stage of a SpaceX launch will successfully land and thus be reused?

METHODOLOGY SUMMARY

1. Data Collection
   - Make GET requests to the SpaceX REST API and Web Scraping Wikipedia

2. Data Wrangling
   - Filter the data
   - Determine the number of each of the following:

     launches/site

     Orbit

     outcome per orbit type
   - Create binary classification for outcome

3. Exploratory Data Analysis
   - SQL queries, Pandas, Matplotlib

4. Interactive Visual Analytics
   - Folium and Plotly Dash

5. Predictive Analysis using Classification Models

Data Collection

The raw data was collected using the available SpaceX API as well as Webscraping

SPACE X REST API

Retrieved API data included information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

Request and parse the SpaceX launch data using the GET request

Decode the response content as a Json and turn it into a Pandas dataframe

The data from these requests is stored in lists and used to create a new dataframe.

Filter the DataFrame to only include Falcon 9 launches

Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome
2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None
2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None

API Github Link

WEB SCRAPING

Extracted Falcon 9 launch records from a Wikipedia table using BeautifulSoup (web scraping tool).

1. Perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.
2. Create a BeautifulSoup object from the HTML response
3. Collect all relevant column names from the HTML table header
4. Iterate through the elements and extract column name one by one
5. Create an empty dictionary with keys from the extracted column names.and fill with launch records extracted from table rows.
6. Create a dataframe from the dictionary.

Request the Falcon9 Launch Wiki page from its URL

↓

Extract all column/variable names from the HTML table header

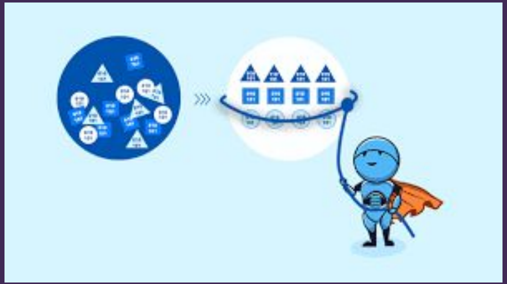
↓

Create a data frame by parsing the launch HTML tables

Webscraping Github

DATA WRANGLING

Performed EDA and determined Training Labels

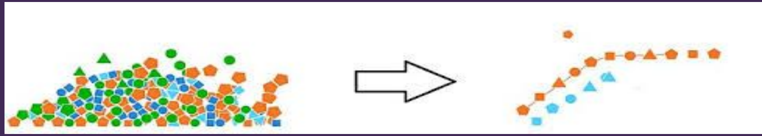


In the data set, there are several different cases where the booster did not land successfully, for example:

- **True Ocean** = the mission outcome was successfully landed to a specific region of the ocean
- **False Ocean** = the mission outcome was unsuccessfully landed to a specific region of the ocean
- **True RTLS** = the mission outcome was successfully landed to a ground pad
- **False RTLS** = the mission outcome was unsuccessfully landed to a ground pad
- **True ASDS** = the mission outcome was successfully landed on a drone ship
- **False ASDS** = the mission outcome was unsuccessfully landed on a drone ship

We converted these outcomes to training Labels with 1 meaning the booster successfully landed. 0 meaning it was unsuccessful.

DATA WRANGLING



The data contains several Space X launch facilities: Cape Canaveral Space Launch Complex, etc

The method value_counts() was used on the column, LaunchSite, to determine the number of launches on each site.

The count was also determined for each of the following variables

- Occurence of Each Orbit
- Number of mission outcomes/orbit

Using the Outcome, a list was created where the element is zero if the corresponding row in Outcome is in the set bad_outcome; otherwise, it was one.

Then it was assigned to the variable landing_class: This variable represented the classification variable for the outcome of each launch.

If the value is zero, the first stage did not land successfully; one means the first stage landed Successfully

Data Wrangling Github Link

EXPLORATORY DATA ANALYSIS (EDA) –

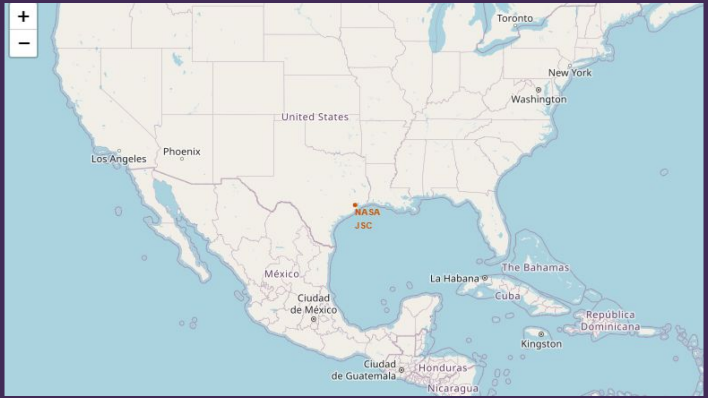
Several tasks are completed in order to perform Exploratory Data Analysis and Feature Engineering.

#	Description
1	Display the names of the unique launch sites in the space mission
2	Display 5 records where launch sites begin with the string 'CCA'
3	Display the total payload mass carried by boosters launched by NASA (CRS)
4	Display the average payload mass carried by booster version F9 v1.1
5	List the date when the first successful landing outcome on a ground pad was achieved
6	List the names of the boosters which had success on a drone ship and a payload mass between 4000 and 6000 kg
7	List the total number of successful and failed mission outcomes
8	List the names of the booster versions which have carried the maximum payload mass
9	List the failed landing outcomes on drone ships, their booster versions, and launch site names for 2015
10	Rank 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

EDA SQL Github

GEOSPATIAL ANALYSIS

FOLIUM



The following steps were taken for launch data visualization:

- Created a map and mark all launch sites

- Clustered launches with the same coordinates and assign a color marker for outcome. Successful = green, Failed = red

- Proximities of launch sites were explored using calculations of distances between points (Latitude and Longitude coordinate values).

Folium Github

EDA with Visualizations

Using Pandas and Matplotlib, variables were tested to see their effects on outcome

To gain preliminary insights about how each important variable would affect the success rate, we selected the features that would be used in success prediction.

- ❖ Flight Number
- ❖ Payload mass overlayed with outcome
- ❖ Launch site
- ❖ Orbit type
- ❖ Year

EDA Github

INTERACTIVE DASHBOARD

PLOTLY DASH

An application for users to perform real-time interactive visual analytics on SpaceX launch data was created

Some insights were obtained from the dashboard to answer the following five 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?

Dash Github

PREDICTIVE ANALYSIS - CLASSIFICATION

Create a column for the class

Standardize the data

Split into training and test data

Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

(Find the best performing method using test data)

Create a NumPy array from the column Class in data, by applying the method to_numpy() then assign it to the variable Y,make sure the output is a Pandas series (only one bracket df['name of column']).

Standardize the data in X then reassign it to the variable X using the transform provided below.

We split the data into training and testing data using the function train_test_split. The training data is divided into validation data, a second set used for training data:

Create a logistic regression object then create a GridSearchCV object logreg_cv
Fit the object to find the best parameters from the dictionary parameters.
We output the GridSearchCV object for logistic regression. We display the best parameters using the data attribute best_params_ and the accuracy on the validation data using the data attribute best_score_.Calculate the accuracy on the test data using the method score.

A similar process is repeated using the following created objects:

- support vector machine
- decision tree classifier
- k nearest neighbors

Github Link

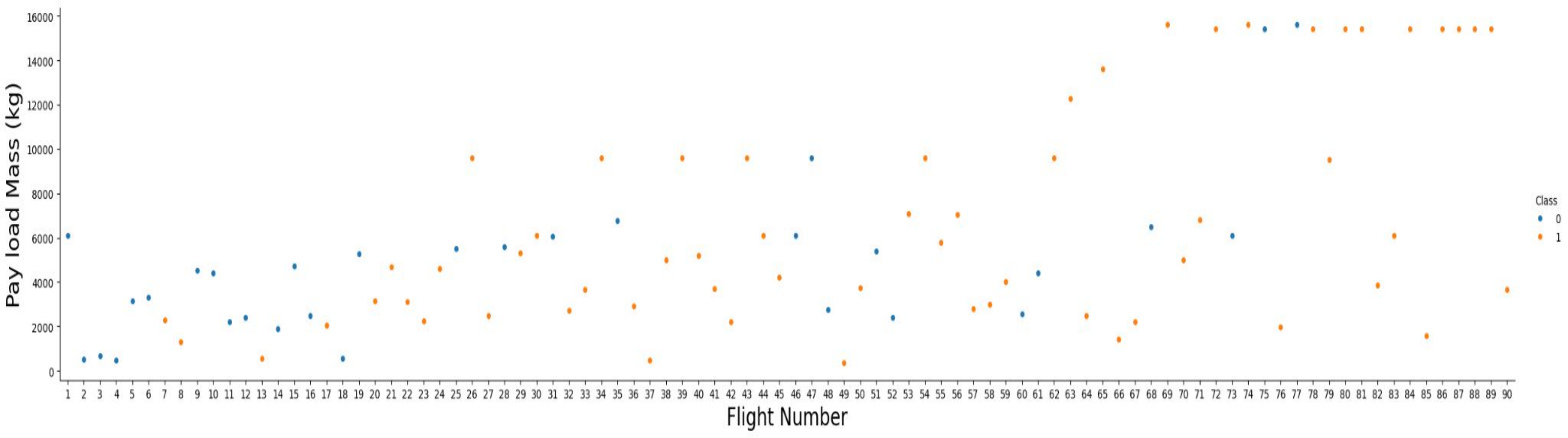
RESULTS


Exploratory
Data Analysis


Interactive Analytics


PREDICTIVE
ANALYSIS

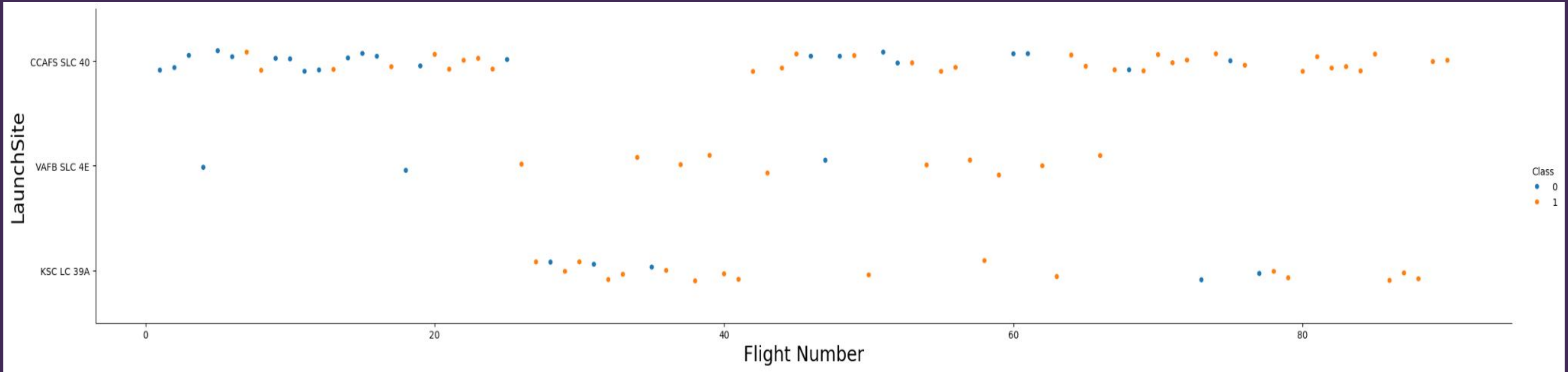
Scatter plot of Flight Number vs Payload Mass with Outcome



We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive payloads have greater success over time.

EDA - Visualizations

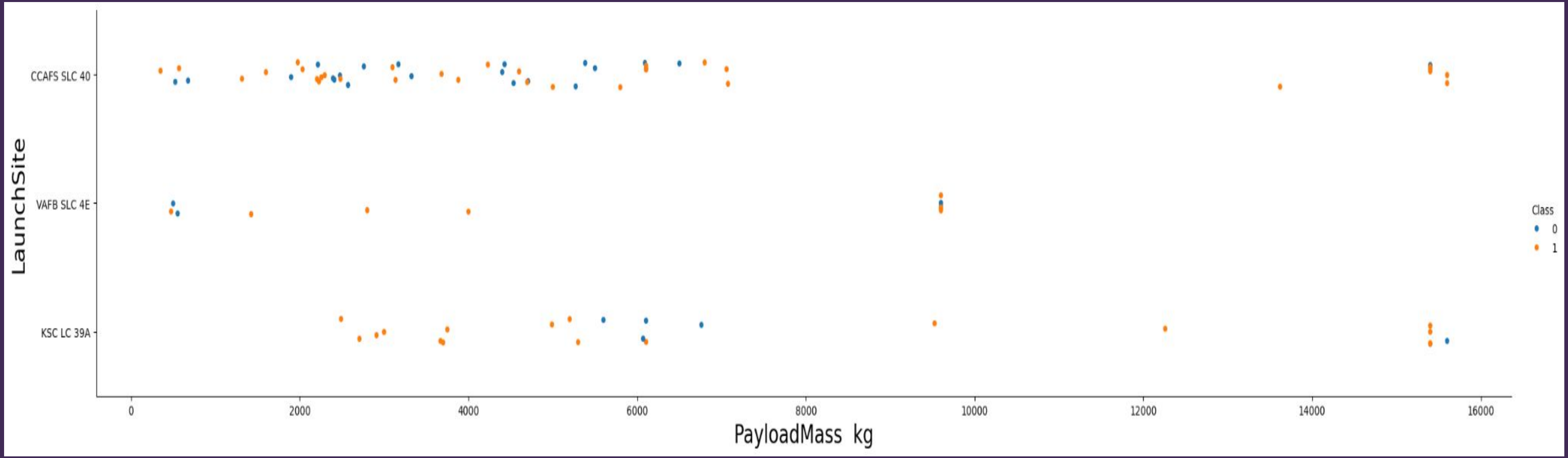
Flight Number vs Launch Site



Different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

EDA - Visualizations

LAUNCH SITE VS. PAYLOAD MASS



Most launches have a payload mass less than 7000kg. Payload mass varied between launches within each site.

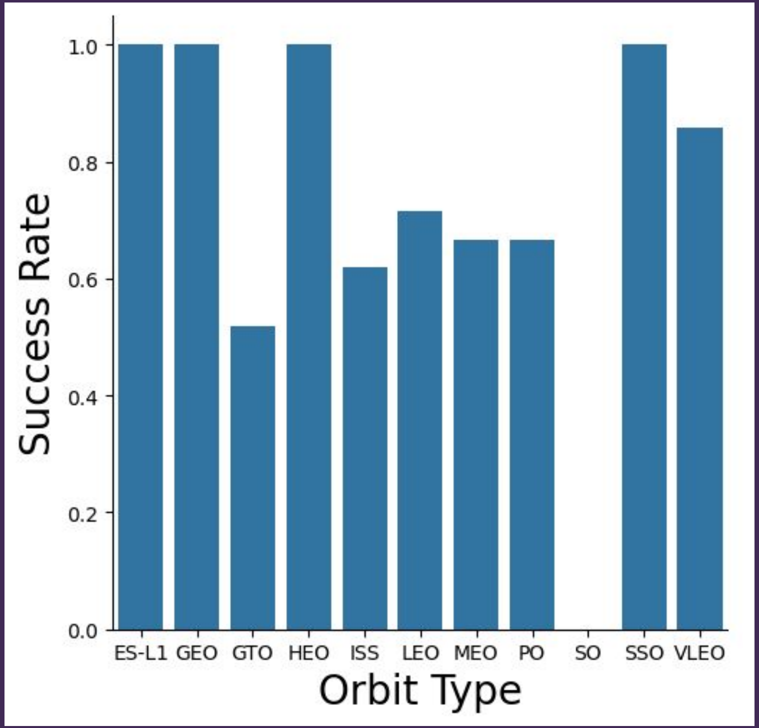
SUCCESS RATE VS. ORBIT TYPE

The bar chart of Success Rate vs. Orbit Type shows that these orbits have 100% success rate:

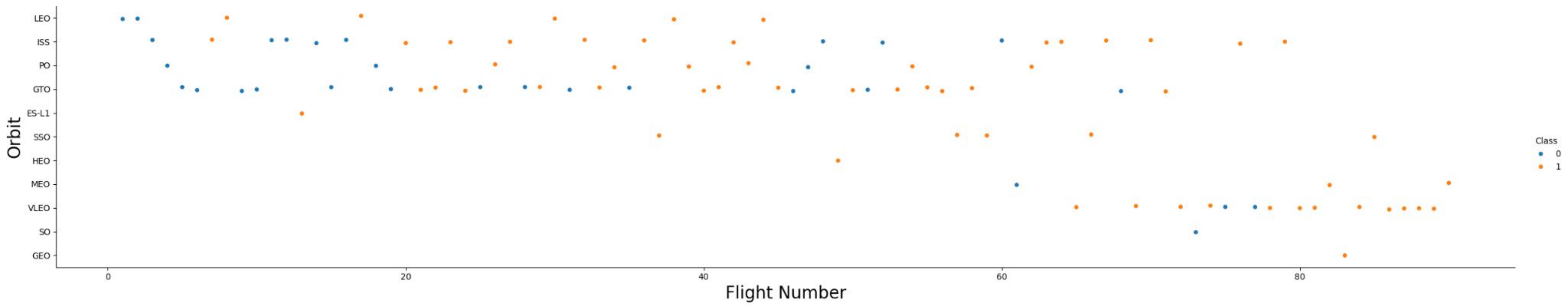
- ES-L1 (Earth-Sun First Lagrangian Point)
- GEO (Geostationary Orbit)
- HEO (High Earth Orbit)
- SSO (Sun-synchronous Orbit)

The orbit with 0% success rate is:

- SO (Heliocentric Orbit)



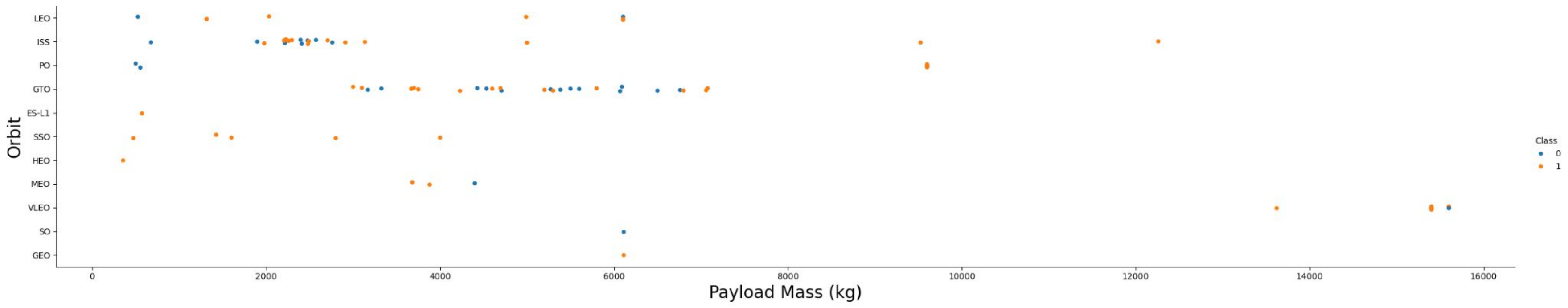
ORBIT TYPE VS. FLIGHT NUMBER



- GEO, HEO, and ES-L1 orbits have such a high success rate because they have only one launch each..
- SSO has been successful 5 out of 5 launches.
- Overall as Flight Number increases, the success rate increases.

EDA - Visualizations

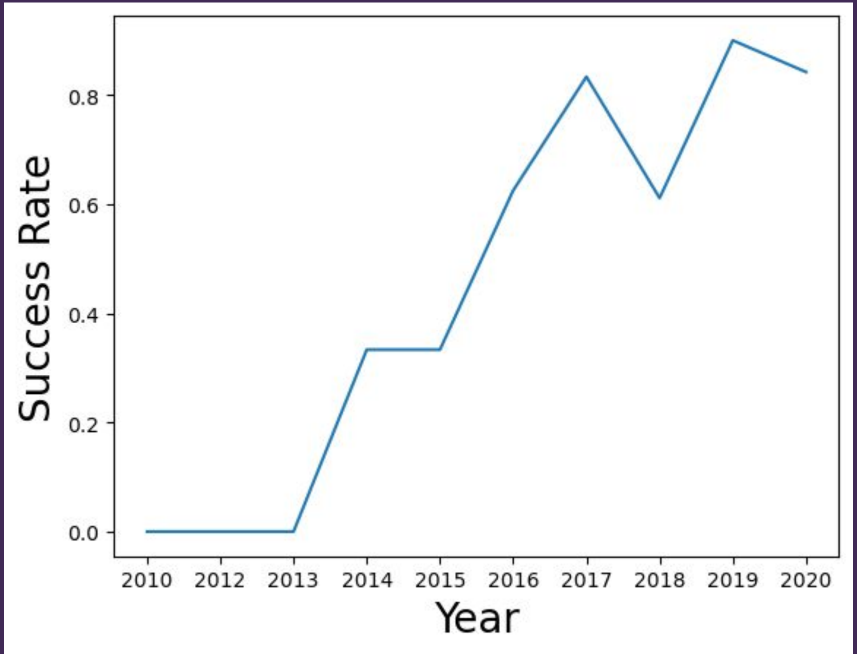
ORBIT TYPE VS. PAYLOAD MASS



- Heavier payloads (>8000 kg) have a greater success rate but fewer launches overall.
- The SSO orbit has a 100% success rate but only low mass launches.
- There is no correlation between mass and success in the GTO orbit.
- VLEO orbit only launched with high payloads.

LAUNCH SUCCESS YEARLY TREND

Overall, launches have become more successful over time.



EDA with SQL

All Launch Sites

```
%sql SELECT DISTINCT launch_site from SPACEXTABLE
```

 * sqlite:///my_data1.db
Done.

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

Listed the names of each unique launch site in the data

Launch Site Names beginning with “CCA”

Display 5 records where launch sites begin with the string 'CCA'

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

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO
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)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)

Viewed 5 records where site name begins with ‘CCA’

Total Payload Mass launched by NASA

Display the total payload mass carried by boosters launched by NASA (CRS)

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

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

total_payload_mass
45596

Calculated the total payload mass carried by NASA boosters

Average Payload Mass by F9

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

```
%sql select avg(payload_mass__kg_) as average_mass from SPACEXTABLE where booster_version = 'F9 v1.1';
```

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

average_mass
2928.4

Calculated the average mass carried by booster version F9 v1.1

First successful outcome (ground pad)

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

Hint:Use min function

```
]: %sql select min(date) from SPACEXTABLE where landing_outcome = 'Success (ground pad)';
```

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

]:

min(date)
2015-12-22

Listing the date in which the first successful landing outcome on a ground pad was achieved.

Successful drone ship landing with payload between 4000 and 6000 kg

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 from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and payload_mass__kg_
```

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

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

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

Total Number of Flights per Mission Outcome

List the total number of successful and failure mission outcomes

```
'''%sql SELECT DISTINCT landing_outcome from SPACEXTABLE'''
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

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

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

Displayed the total number of failure and success outcomes

Booster Versions Carrying Maximum Payload Mass

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);
```

```
 * 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

Lists which boosters have carried the max. Payload mass

2015 Launch Records

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%%sql select substr(Date, 6,2) as month, booster_version, launch_site, landing_outcome from SPACEXTABLE
      where landing_outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

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)

Displayed the failure landing outcomes in drone ship, booster versions, and launch sites for each month in 2015

Success Count Rankings

(2010-06-04 to 2017-03-20)

```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEXTABLE
      where date between '2010-06-04' and '2017-03-20'
      group by landing_outcome
      order by count_outcomes desc;
```

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

Landing_Outcome	count_outcomes
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

Ranked the count of landing outcomes (Failure(drone ship) or Success) in descending order

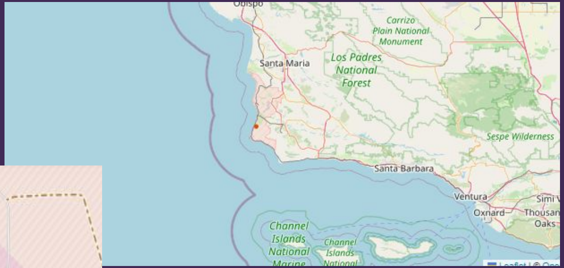
INTERACTIVE MAP

with

Folium

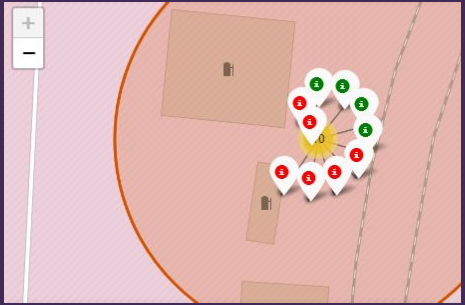
All Launch sites marked on a Map

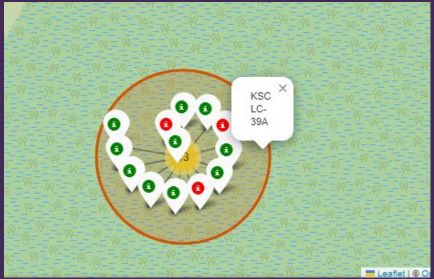
All SpaceX launch sites are on coasts of the United States of America, specifically Florida and California, and in proximity to the Equator.

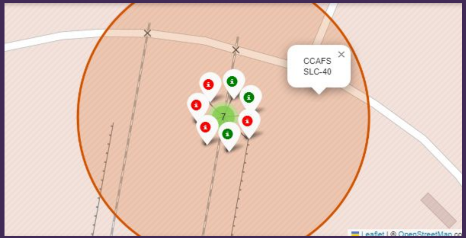


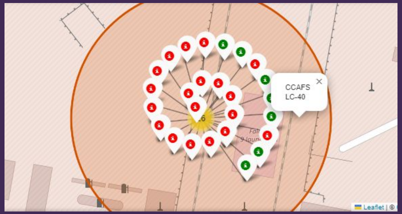
Labeling Launches by color for each site

Launches have been grouped into clusters, and annotated with green icons for successful launches, and red icons for failed launches.









VAFB
SLC-4E

KSC
LC-39A

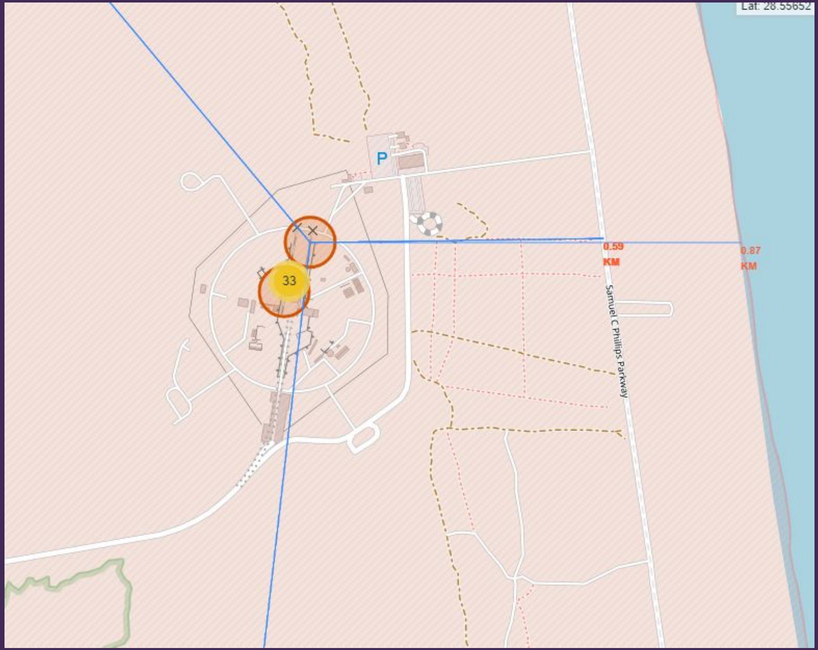
CCAFS SLC-40 and CCAFS LC-40

Launch Sites and Other Points of Interest

We can understand more about the placement of launch sites.

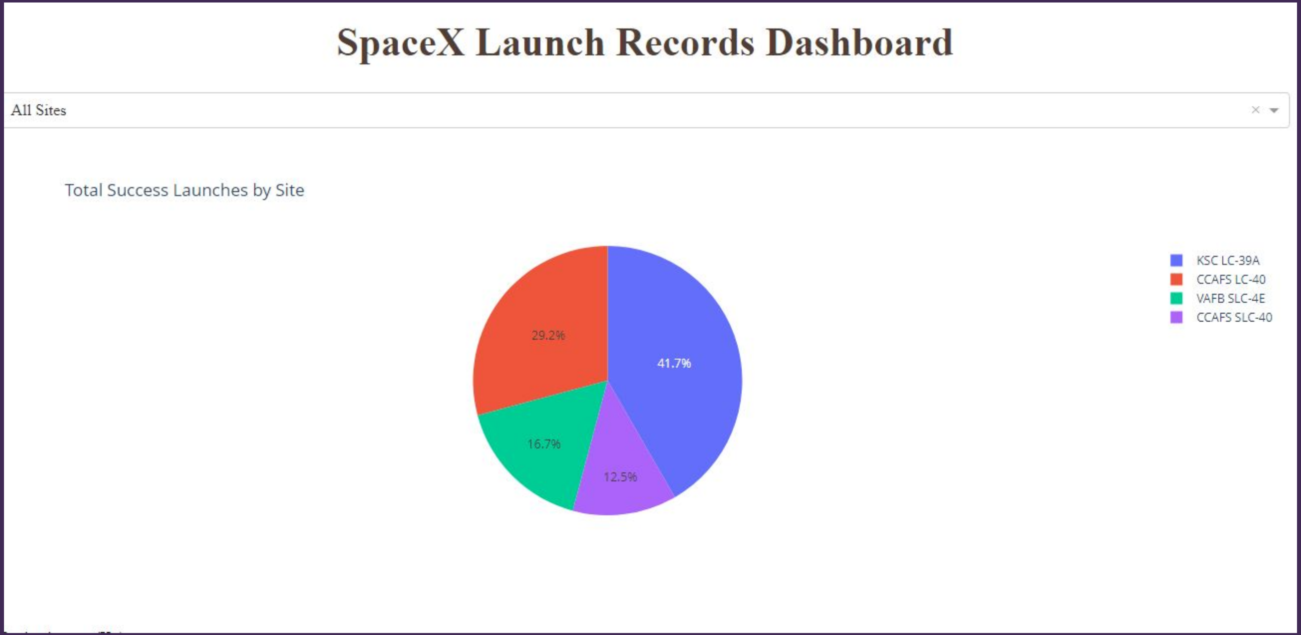
Using SLC-40 as an example, we see

- ❖ The coastline is only 0.87 km due East.
- ❖ Samuel Phillips Parkway, the nearest highway is only 0.59km away.
- ❖ NASA Railroad, the nearest railway is only 1.29 km away.
- ❖ The nearest city is 51.74 km away (Melbourne Int'l Airport)



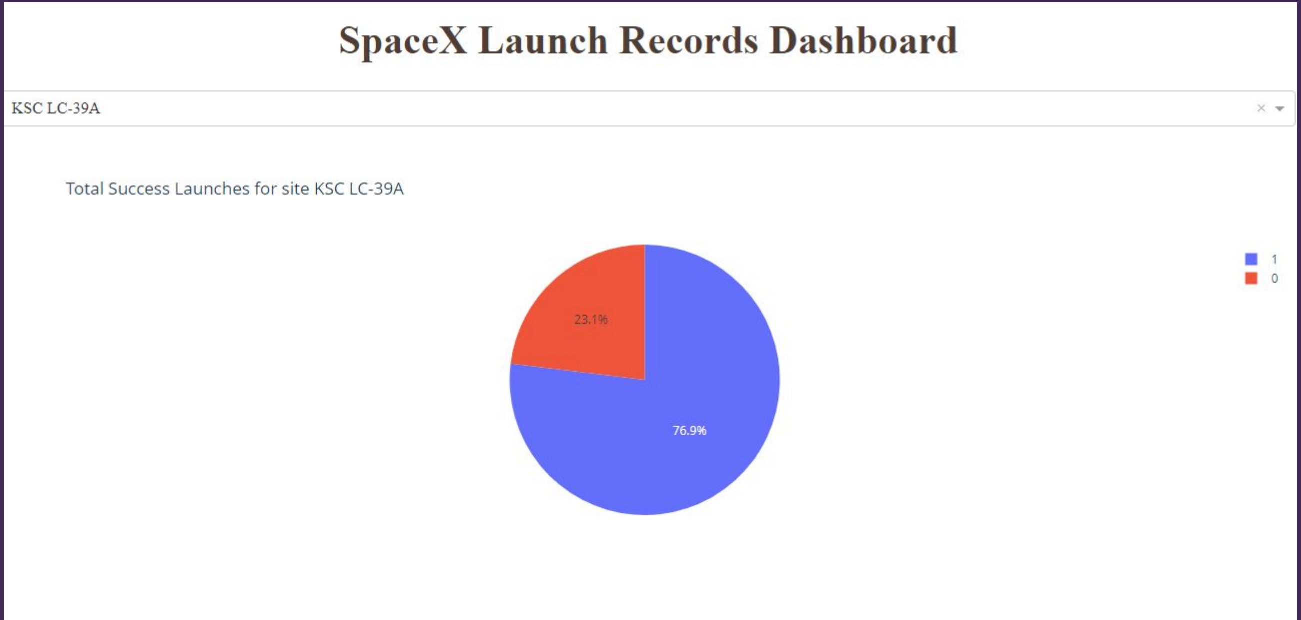
INTERACTIVE DASHBOARD

Launch Success Counts for All Sites



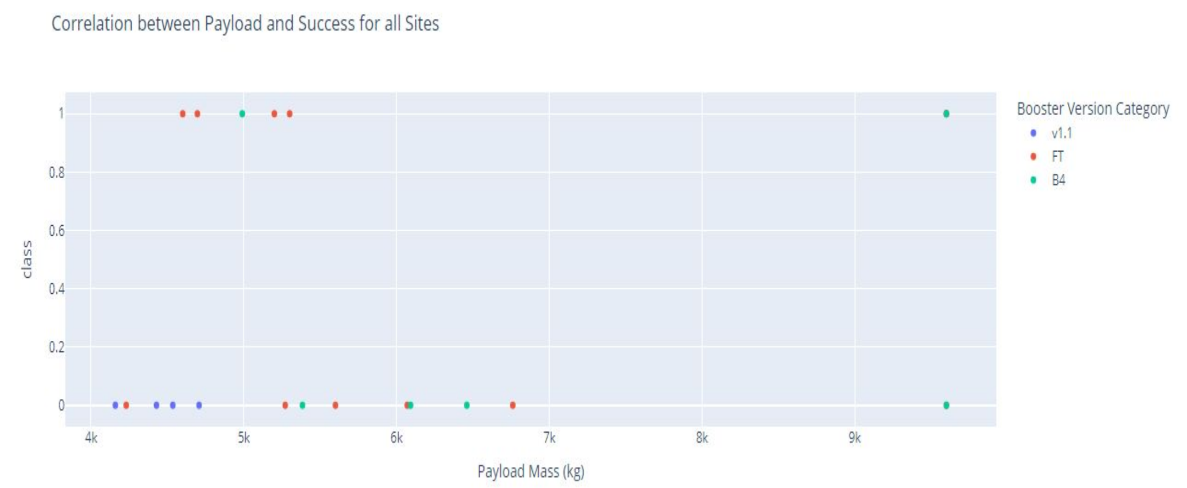
The launch site KSC LC-39A had the most successful launches, with 41.7% of the total successful launches.

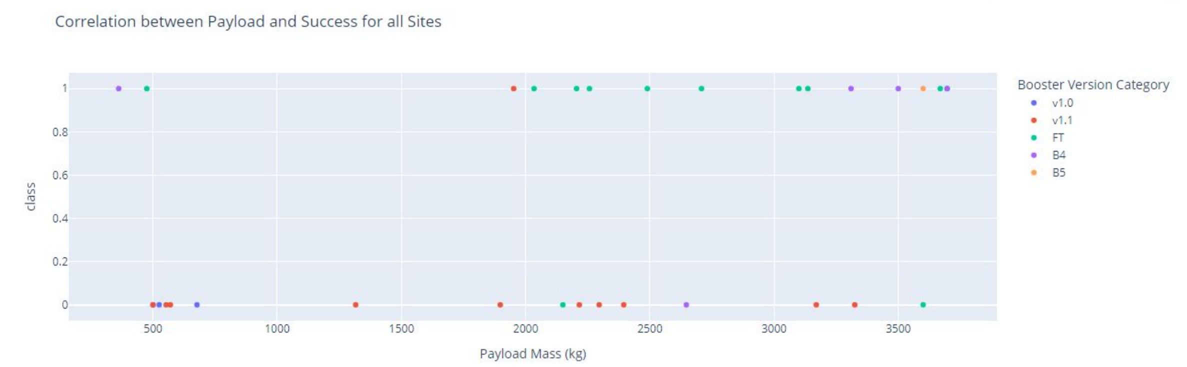
Launch Site with the Highest Launch Success ratio



The launch site KSC LC-39A also had the highest rate of successful launches, with a 76.9% success rate.

LAUNCH OUTCOME VS. PAYLOAD SCATTER PLOT FOR ALL SITES





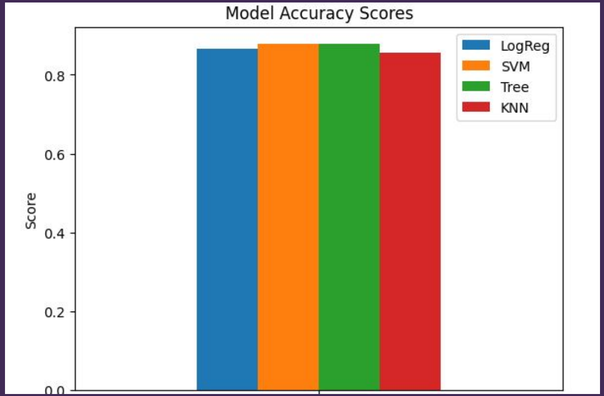
Success ratio for payloads greater than 4000 kg is lower than that for payloads below 4000kg.

PREDICTIVE ANALYSIS

CLASSIFICATION

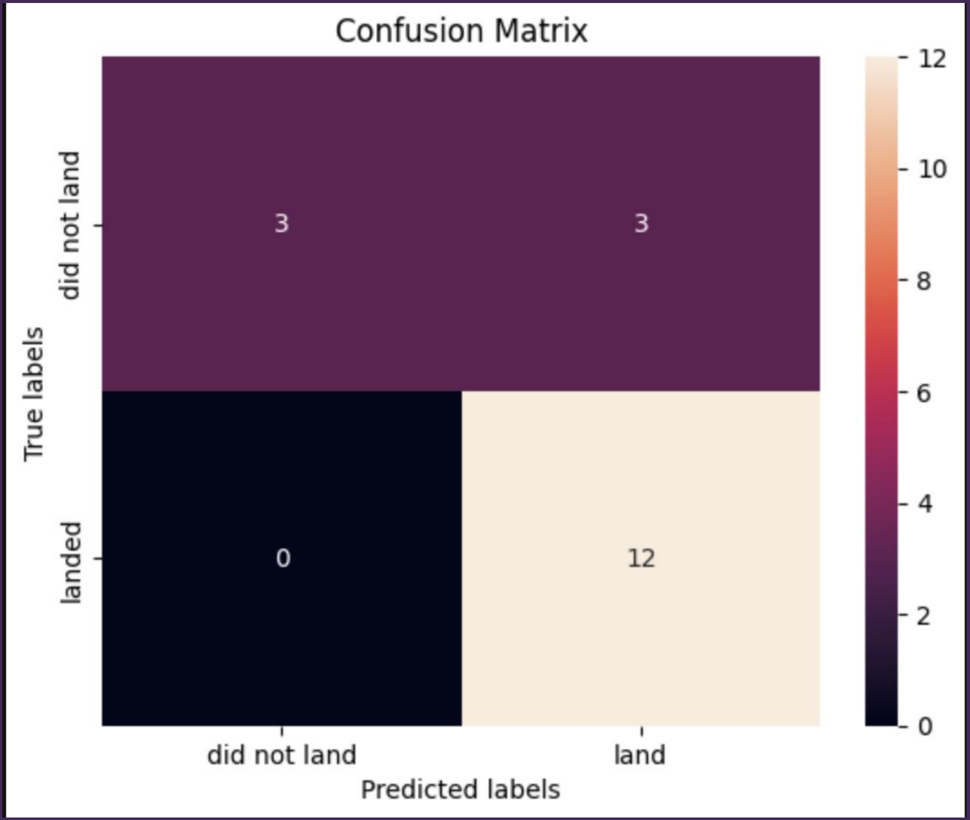
CLASSIFICATION ACCURACY

- The SVM model has the highest classification accuracy
  - The Accuracy Score is 87.78%



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.787879	0.819444
F1_Score	0.909091	0.916031	0.881356	0.900763
Accuracy	0.866667	0.877778	0.844444	0.855556

CONFUSION MATRIX



- As shown previously, best performing classification model is the SVM model, with an accuracy of 87.78%.
-

CONCLUSIONS

CONCLUSIONS

- Launches have become more successful over time. Earlier launches failed more often than current launches.

- Orbit type SSO has been the most successful overall, with a perfect 5 out of 5 launches or 100% success rate.

- Other orbits with 100% success rates only have one launch.

- KSC LC-39A is most successful of launch sites, with 41.7% of the total successful launches, and also the highest rate of successful launches, with a 76.9% success rate.

- The success for massive payloads (over 4000kg) is higher than that for low payloads but have fewer launches overall.

- The best performing classification model is the SVM model, with an accuracy of 87.78%.

Acknowledgments

IBM

Coursera