IBM Data Science Capstone Project

Hemant Kumar

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 an interactive dashboard with Plotly Dash

❑Summary of all results

- Exploratory data analysis results
- Interactive analysis demo results with screenshots
- Predictive analysis results

Introduction

Project background and context

We predicted that the Falcon 9 first stage will land successfully. 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. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against Space for a rocket launch.

Problems seeking solutions

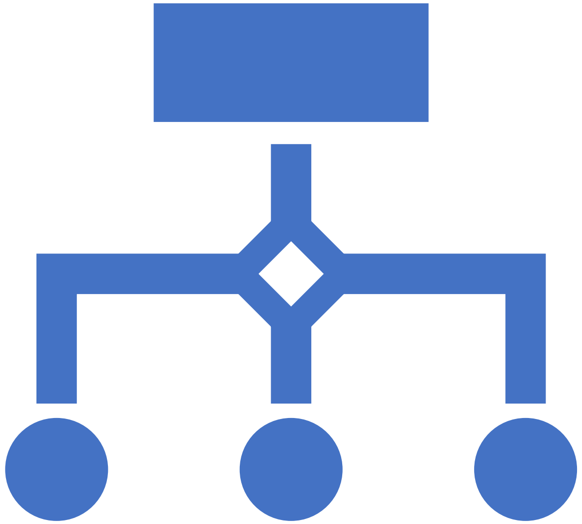
- What influences if the rocket land successfully
- The effect each relationship with certain rocket variables will impact in determining the success rate of a successful landing.
- What conditions does Space have to achieve to get the best results and ensure the best rocket success landing rate?

Section 1

Methodology

Methodology

- Executive Summary
- Data collection methodology:
  - SpaceX REST API
  - Web Scraping from [Wikipedia] using BeautifulSoup
- Perform data wrangling
  - Cleaning the data
  - Dealing with missing values
  - Using one-hot encoding to prepare the data for classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning, and evaluation of classification models to ensure best results.



Data Collection

This process involved a combination of API requests from Space REST API and Web Scrapping data from a table in Space's Wikipedia entry. Both the data collected were used in order to get complete information about the launches for a more detailed analysis. Data columns obtained using

- Space REST API is as follows:

Flight Numbers, Date, Booster Version, Orbit, Payload Mass, Customer, Launch Site, Outcome, Flights, Grid Fins, Reused, Legs, Landing pad, Block, Reused Count, Serial, Longitude, Latitude.

- Data/columns are obtained by using Wikipedia Web Scraping:

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

Data Collection – SpaceX API

Requesting rocket launch data from SpaceX API

Define a series of helper functions that will help us use the API to extract information IDs in launch data.

Getting the response using requests.get() and decoded its content using .json() and turned it into a dataframe using json.normalize().

Using API, requesting needed information using IDS given for each launch.

Constructing data we have into a dictionary using columns and converting it into a dataframe.

Filtering dataframe to contain only Falcon 9 launches and replace values of payload mass using its mean.

Exporting it to CSV file

[Github link for data collection using SpaceX API]

Data Collection – Web Scraping

Requesting Falcon 9 launch data from Wikipedia.

Creating BeutifulSoup object from HTML response.

Extract all column names from the HTML header.

Collecting the data by parsing HTML tables.

Constructing the obtained data into a dictionary.

Create a dataframe from the dictionary.

Export it to CSV file.

[Github link for data collection using web scraping]()

Data Wrangling

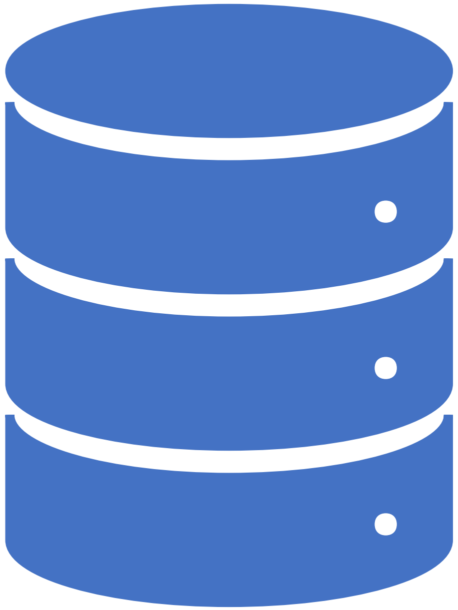
In the data set, there are several different cases where the booster did not land successfully, Sometimes a landing was attempted but failed due to an accident. for example, True Ocean means the mission outcome was successfully landed in a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed in a specific region of the ocean. True RTLS means the mission outcome was successfully landed on a ground pad False RTLS means the mission outcome was unsuccessfully landed on a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship. We mainly convert those outcomes into Training Labels with "1" meaning the booster successfully landed, and “0" meaning it was unsuccessful.

[Github link for 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, and Success Rate Yearly Trend Scatter plots show the relationship between variables. If a relationship exists, they could be used in a machine learning model. Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and/or a measured value. Line charts show trends in data over time.

Github link for EDA with data visualization



EDA with SQL

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

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

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

5. List the names of the boosters which have success in drone ships and have payload mass greater than 4000 but less than 6000.

6. Listing the total number of successful and failed mission outcomes.

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

8. List the failed landing outcomes in drone ships, their booster versions, and launch site names for the months in the year 2015.

9. Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad))between the dates 2010-06-04 and 2017-03-20 in descending order

[Github link for EDA with SQL]()

Build an Interactive Map with Folium

Markers of all Launch Sites:

Ranking the count of landing outcomes (such as Failure (drone ship) or success (ground pad)) between the dates 2010-06-04 and 201703.20 in descending order.

Colored Markers of the launch outcomes for each Launch Site:

Added colored Markers of success (Green) and failed (Red) launches using marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

Added colored lines to show distances between the Launch Site **‘CCAFS FLC-40’** (as an example) and its proximities like Railway, Highway, Coastline, and Closest City.

Github link for an interactive map with folium

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List	Dropdown list for Launch site selection.
Pie Chart Showing Success Launches (All Sites/Certain Sites)	Used a scatter plot to showcase the total successful launches cunt for all sites and the Success vs. Failed counts for the site, if a specific Launch Site is selected.
Slider of Payload Mass Range	Slider is used for selecting payload range.
Scatter Chart of PayloadMass vs. Success Rate for the different Booster Versions	Used a Scatter chart to show the correlation between Payload and Launch Success

Github link for Dashboard with Plotly Dash

Predictive Analysis (Classification)

- Creating numPy array from the column ‘Class’ in data.
- Standardizing the data with StandardScaler, then fitting and transforming it.
- Splitting the data into train and test sets with train_test_split function.
- Creating a dictionary with cv=10 to find the best parameters
- Applying GridSearchCV on Logistic Regression, SVM, Decision Tree, and KNN model.
- Calculating the accuracy of the test data using the score() method in all models.
- Examining the confusion matrix for all models.
- Finding which method performs best by examining the F-1 Score, Jaccard Score metrics.

Github link for Predictive Analysis

Results

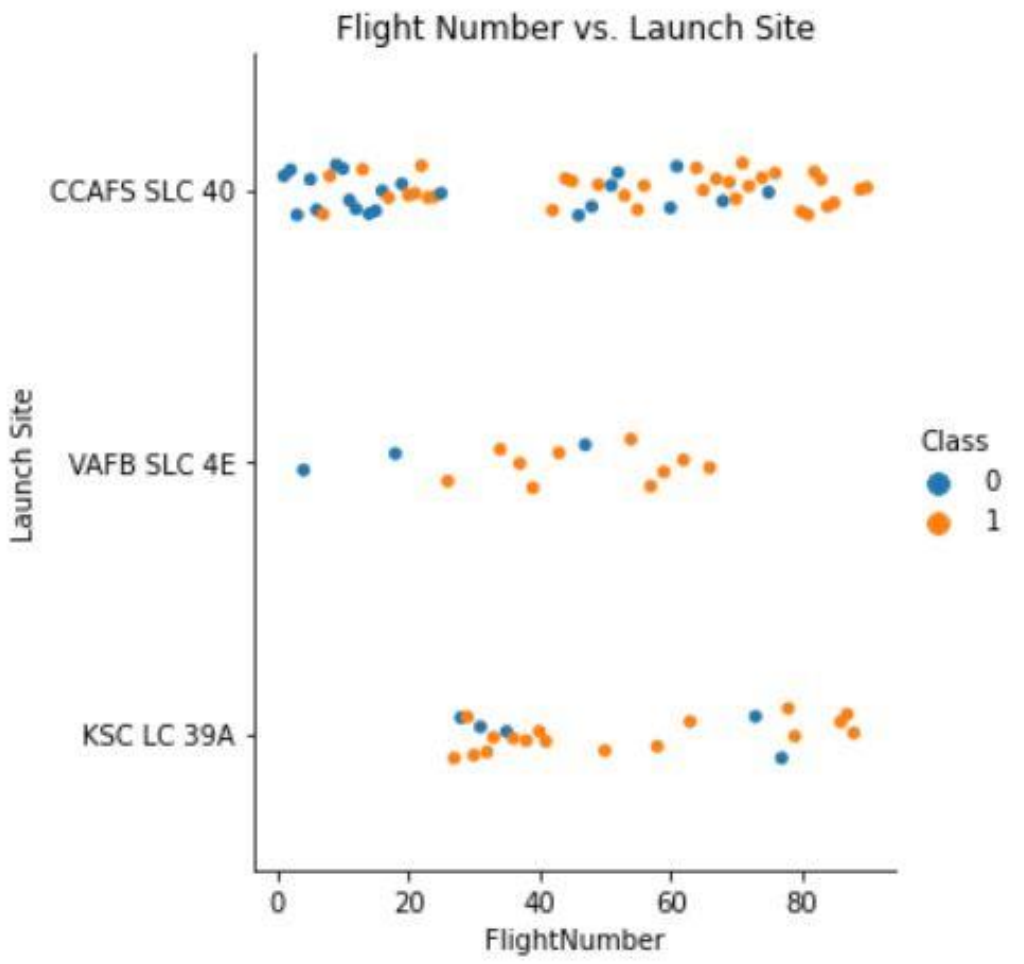
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

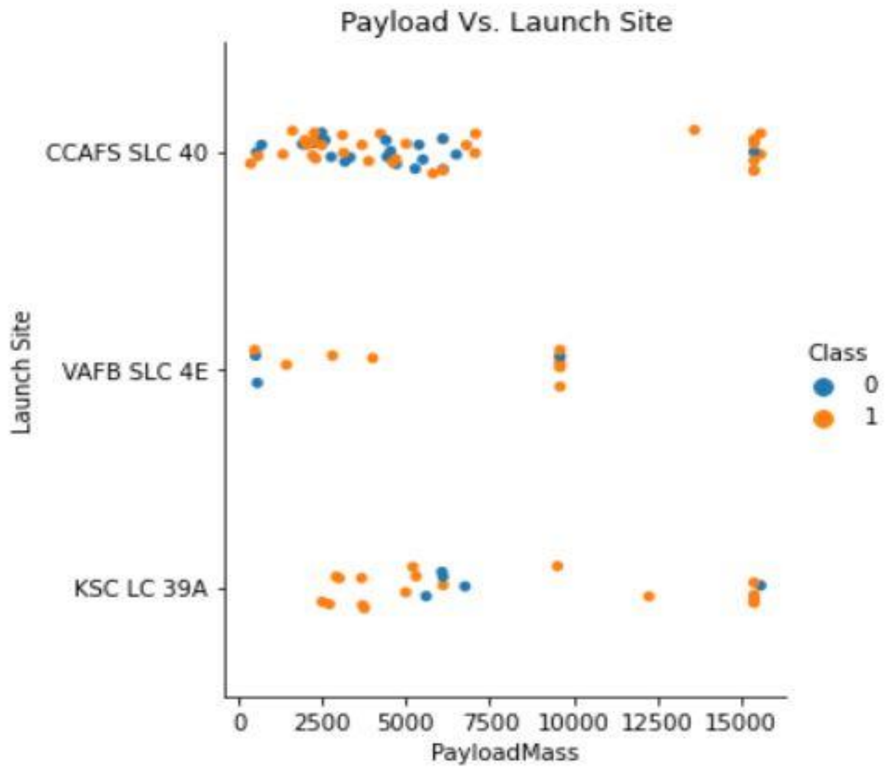
Flight Number vs. Launch Site

- The ‘CCAFS SLC 40’ site has about half of all launches.
- All the earliest flights failed.
- ‘VAFB SLC 4E’ and ‘KSC LC 39A’ both have higher success rates.
- As the Flight Number is increasing the first stage is more likely to predict.



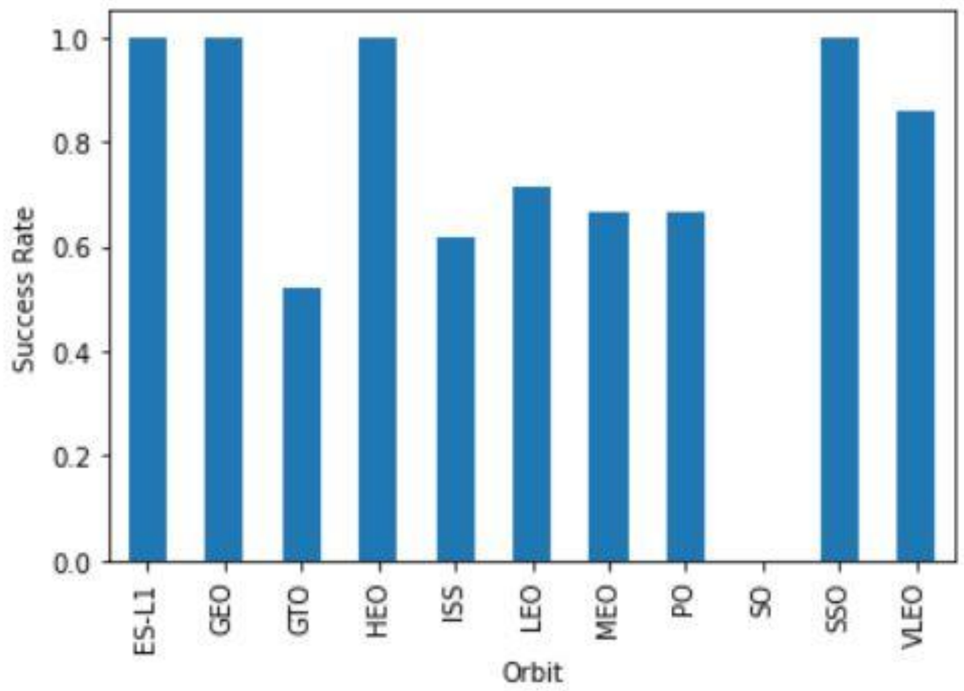
Payload vs. Launch Site

- Higher the Payload Mass, the greater the success rate.
- ‘KSC LC 39A’ has a 100% success rate when Payload Mass is under 500 kg.
- Majority of the launches above 7000 kg were successful.



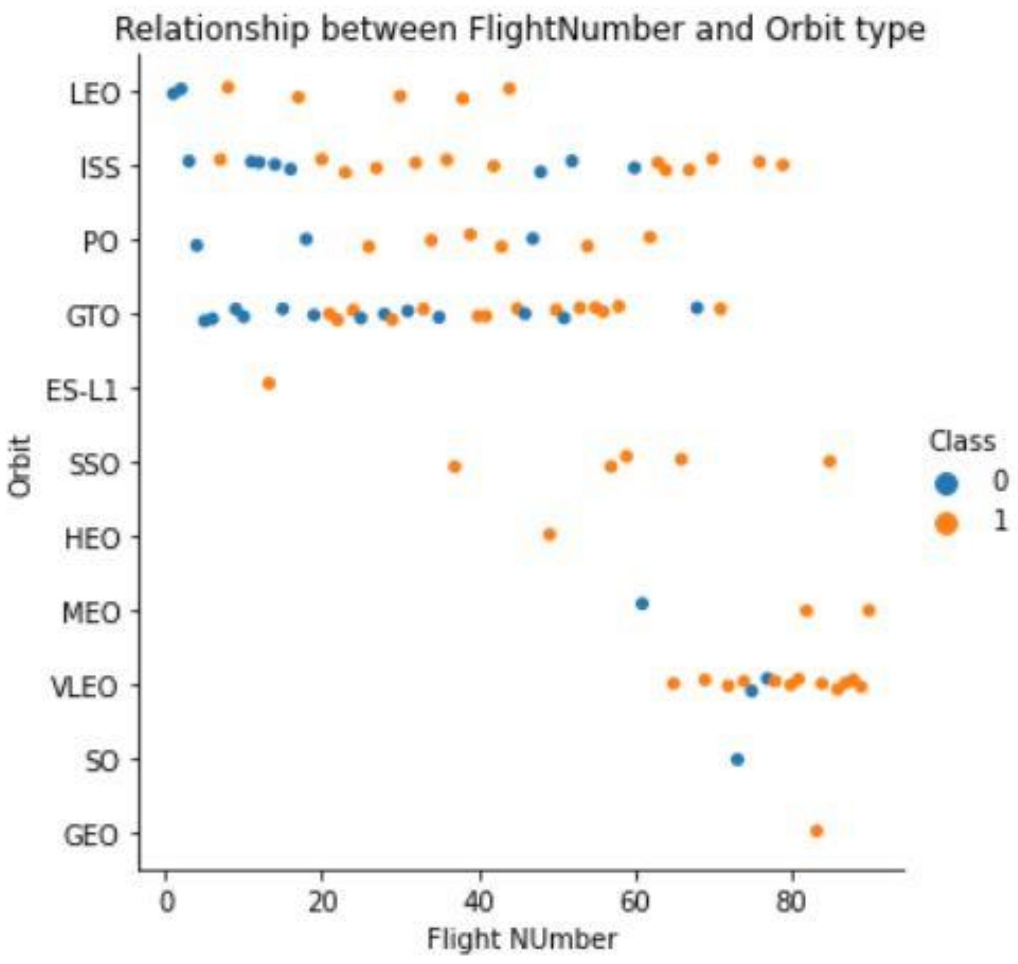
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO has the highest success rate of 100%.
- SO has the poorest success rate of 0%.
- Others have a success rate between 50 -85%.



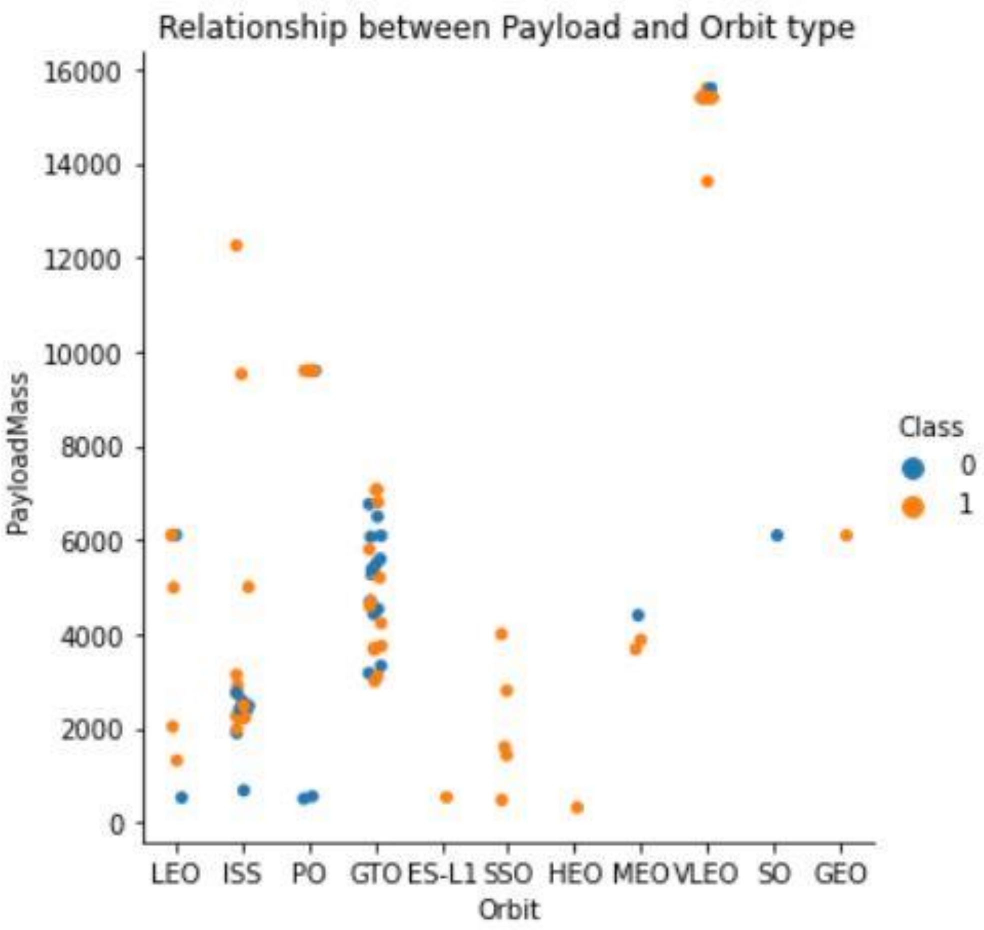
Flight Number vs. Orbit Type

- LEO orbit's success is related to the number of flights.
- No relation when in GTO orbit.



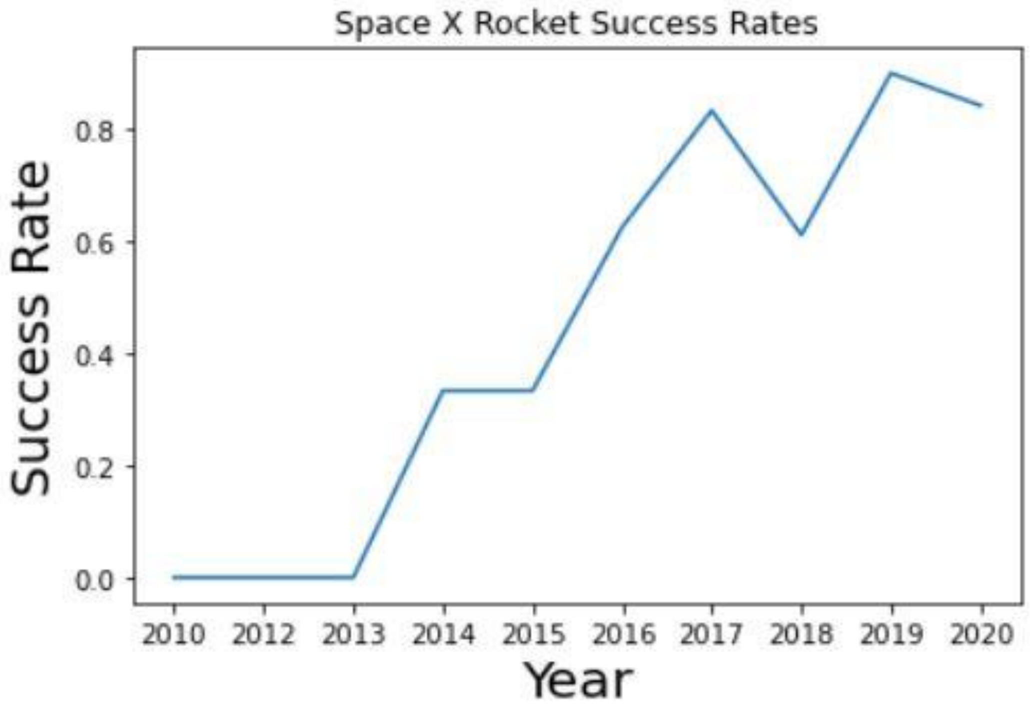
Payload vs. Orbit Type

- With Heavy Payload Polar, LEO, and ISS have a positive impact on landing rate.
- For GTO, no relation with Payload.



Launch Success Yearly Trend

- Since 2013, the success rate kept increasing till 2020.



All Launch Site Names

- Names of the unique launch site in the space mission.

In [10]:
```
%%sql
select distinct(Launch_Site)
from SPACEXTBL
;
```

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

Out[10]:

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

Launch Site Names Begin with 'CCA'

- Display of 5 records where the launch site begins with ‘CCA’.

In [12]:
```
%%sql
select *
from SPACEXTBL
where Launch_Site like "CCA%"
limit 5
;
```

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

Out[12]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Total Payload Mass carried by boosters launched by ‘NASA (CRS)’.

In [13]:
```
%%sql
select sum(PAYLOAD_MASS__KG_),Customer
from SPACEXTBL
where Customer == "NASA (CRS)"
;
```

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

Out[13]:

sum(PAYLOAD_MASS__KG_)	Customer
45596.0	NASA (CRS)

Average Payload Mass by F9 v1.1

- Average of Payload Mass carried by boosters with version ‘F9 v1.1’.

```
In [15]: %%sql
select avg(PAYLOAD_MASS__KG_),Booster_Version
from SPACEXTBL
WHERE Booster_Version like "F9 v1.1%"
;

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

Out[15]:

avg(PAYLOAD_MASS__KG_)	Booster_Version
2534.6666666666665	F9 v1.1 B1003

First Successful Ground Landing Date

- Date when the first successful landing outcome in the ground pad was achieved.

```
In [20]: %%sql
select Date,Landing_Outcome
from SPACEXTBL
where Landing_Outcome like "%Success%"
limit 1
;

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

Out[20]:

Date	Landing_Outcome
22/12/2015	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters that have success in drone ships and Payload Mass lie between 4000 and 6000 kg.

In [26]:
```
%%sql
select Booster_Version
FROM SPACEXTBL
where Landing_Outcome like "%Success%drone%)" and PAYLOAD_MASS__KG_ between 4000 and 6000
;
```

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

Out[26]:

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

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failed mission Outcomes.

In [33]:
```
%%sql
select Mission_Outcome,count(*)
from SPACEXTBL
group by Mission_Outcome
;
```

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

Out[33]:

Mission_Outcome	count(*)
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Name of boosters that have carried the maximum Payload Mass.

```
In [41]: %%sql
select Booster_Version,PAYLOAD_MASS__KG_
from SPACEXTBL
where PAYLOAD_MASS__KG_ == (select MAX(PAYLOAD_MASS__KG_) from SPACEXTBL)
;

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

Out[41]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

2015 Launch Records

- Failed landing outcome in drone ships, their booster version, launch site, along with month in the year 2015.

In [44]:
```
%%sql
select substr(Date,4,2),Landing_Outcome,Booster_Version,Launch_Site
from SPACEXTBL
where substr(Date,7,4)='2015' and Landing_Outcome = "Failure (drone ship)"
;
```

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

Out[44]:

substr(Date,4,2)	Landing_Outcome	Booster_Version	Launch_Site
10	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

- Ranking the count of landing outcomes(such as failure or success) between 04-06-2010 and 20-03-2017 in descending order.

In [61]:
```
%%sql
select Landing_Outcome,count(Landing_Outcome)
from SPACEXTBL
where Landing_Outcome like "%Success%" and Date between '04-06-2010' and '20-03-2017'
group by Landing_Outcome desc
;
```

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

Out[61]:

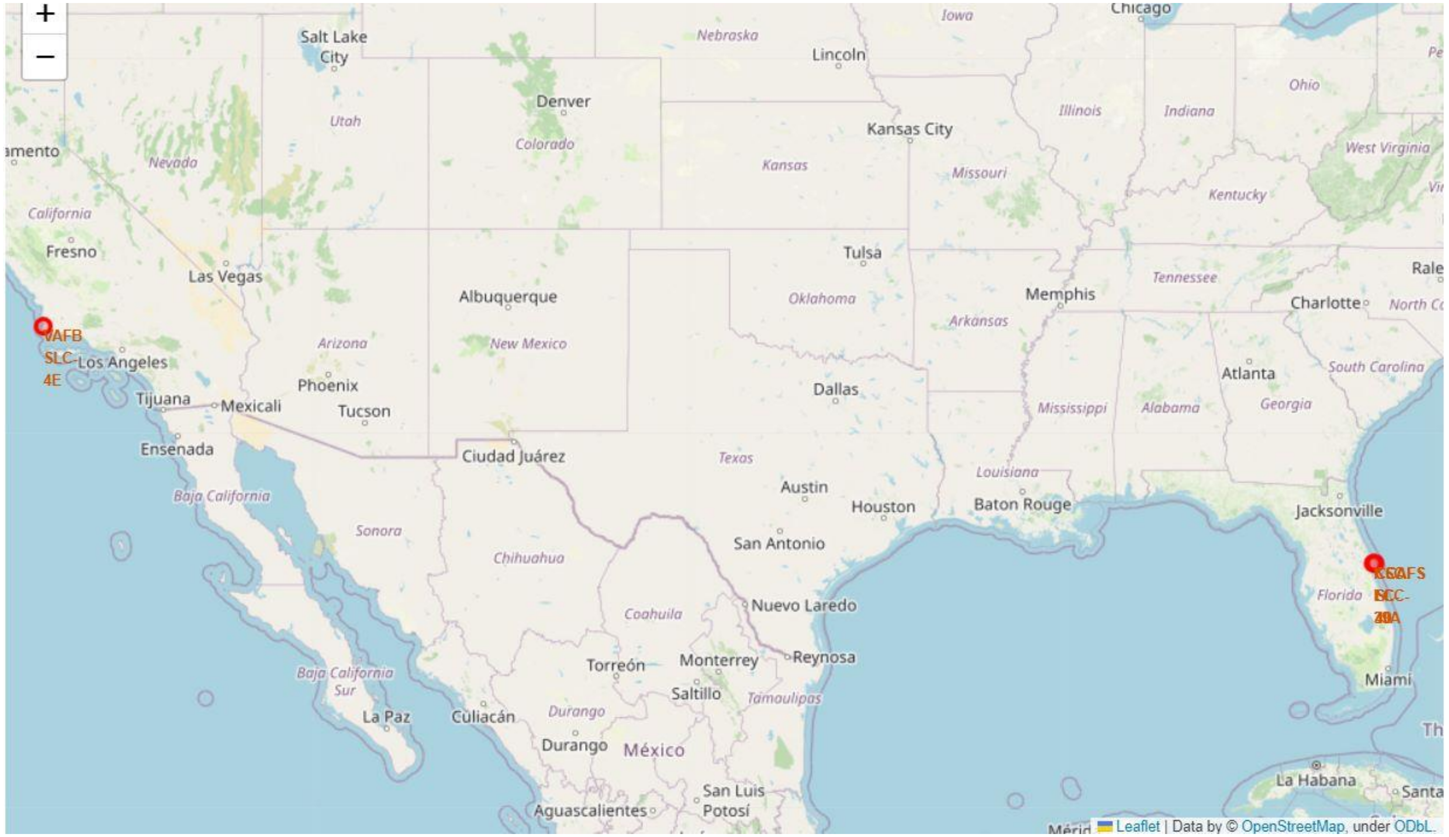
Landing_Outcome	count(Landing_Outcome)
Success	20
Success (drone ship)	8
Success (ground pad)	7

Section 3

Launch Sites Proximities Analysis

Launch Sites Location markers on Global map

- The majority of launch sites are in proximity to the Equator line. The land is moving faster at the equator than at any other place on the Earth. If the ship is launched from the equator it goes up into space and it is also moving around the earth at the same speed was moving before punching. Anything on the surface of the earth at the equator is already moving at 1670 km/hr. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit. All launching sites are close to the coast since launching the rocket towards the ocean helps minimize the risk of having any debris falling or exploding near people.

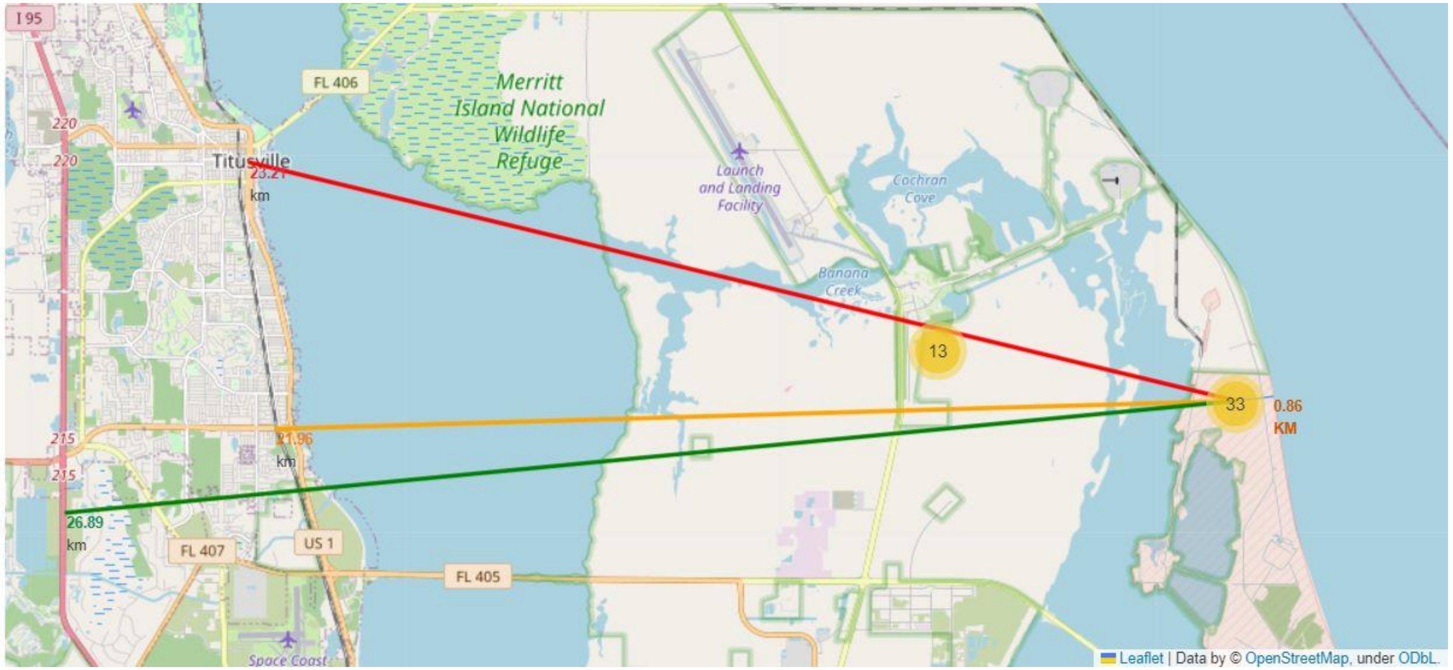


Color-Label Launch record on the map

- We should be able to easily identify which launch sites have success rates.
  - Green marker indicates the success of launched
  - Red marker indicates the failed launch
- Site CCAES SLC-40 doesn't have a high success rate.

Distance from the Launch Site ‘CCAFS SLC-40’ to its proximities

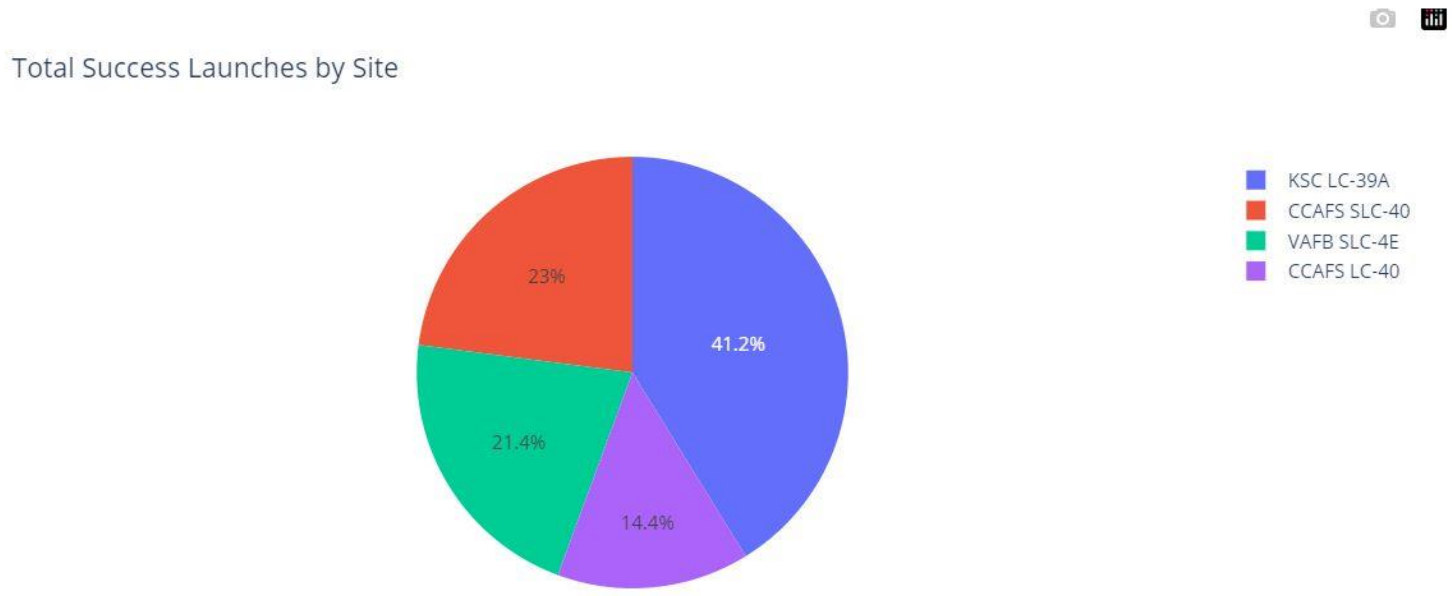
- From the visual analysis of the launch site ‘CCAFS SLC-40’ we can see that it is:
  - At a distance (21.96 km) from the railways
  - At a distance (26.89 km) from the highway
  - At a distance (23.21 km) from the Closest city
- Launch Site ‘KSC LC-39A’ will have a distance nearer than these. So, the Failed rocket with its highest speed can cover distances like 10 - 20 km in a few seconds thus a potential danger to populated areas.



Section 4

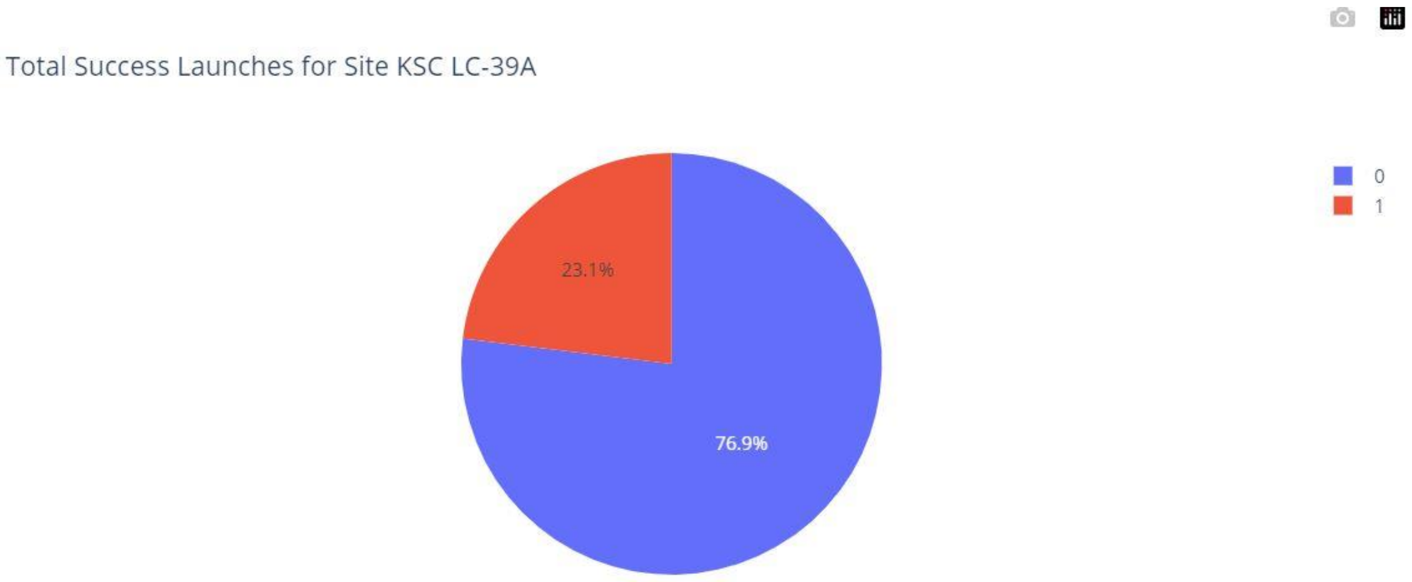
Build a Dashboard with Plotly Dash

Launch success counts for all sites



- The chart showcases that across all the sites, ‘KSC-LC 39A’ has the most successful launches.

Launch site with the highest launch success rate



- ‘KSC-LC 39A’ has the highest launch success rate (76.9%), while getting a (23.1%) failure rate.

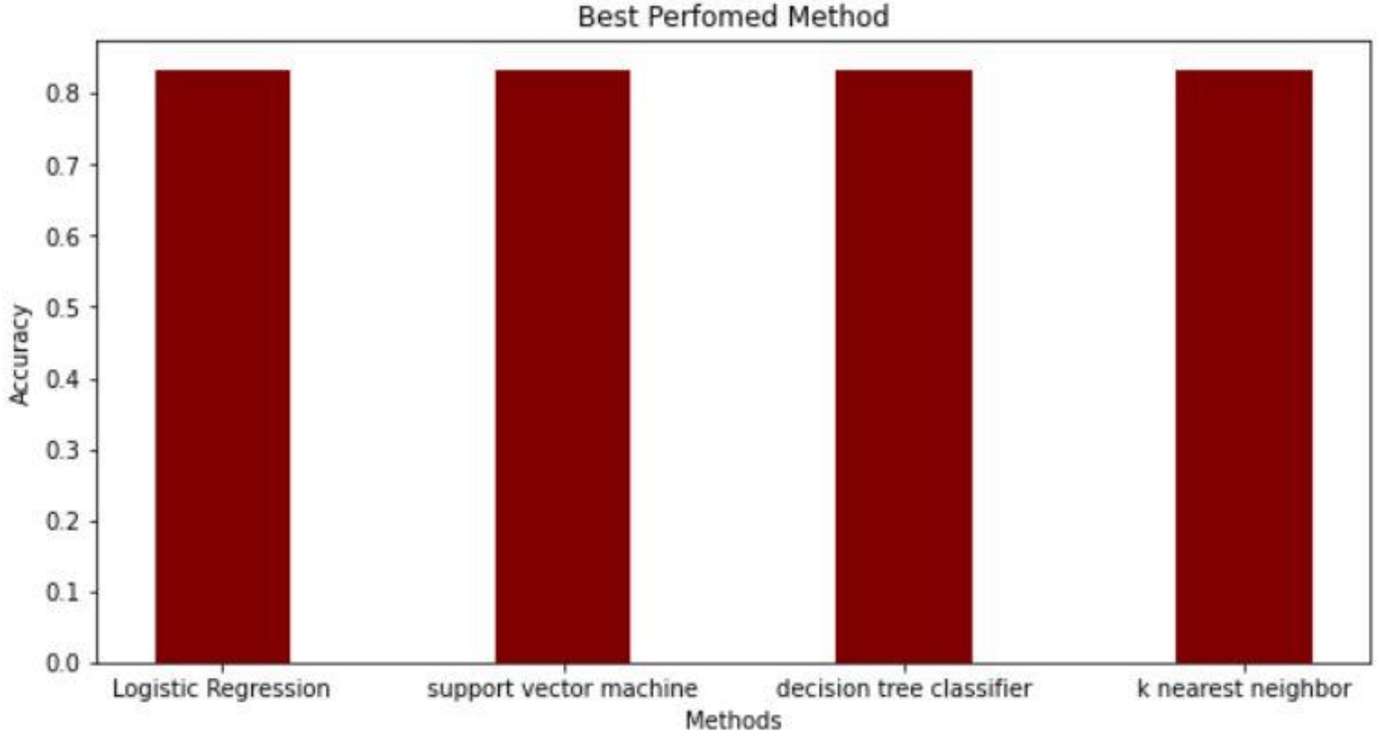
Payload Mass v/s Launch Outcome for all sites

- Payload Mass between 2000 and 5500 kg has the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



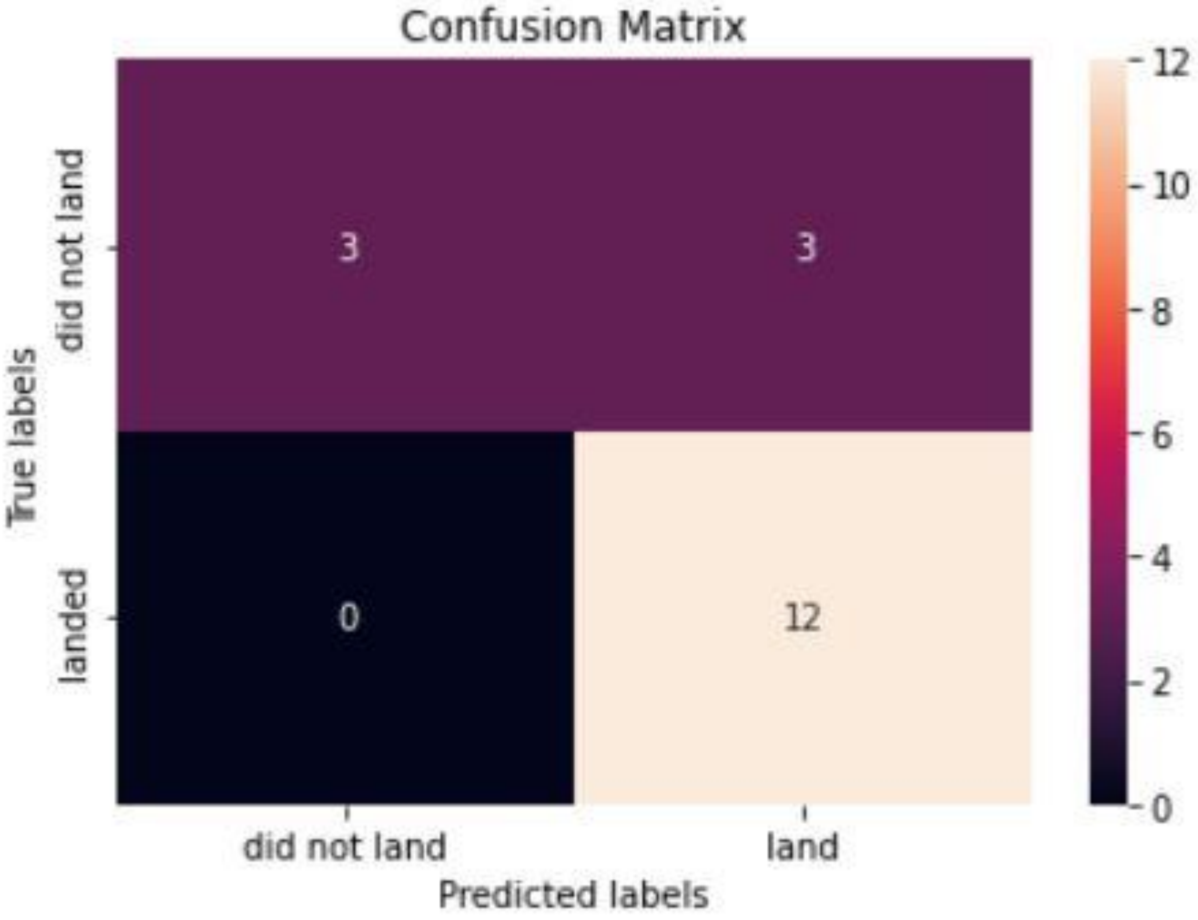
We are getting the same accuracy for all the models. So, we calculated metrics on the whole dataset.

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

- So, the SVM algorithm wins. It has higher scores as well as higher accuracy,

Confusion Matrix

- We see that SVM can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- The SVM model algorithm is the best for this dataset.
- Launches with low Payload Mass perform better than launches with higher Payload mass.
- The success rates are directly proportional to times in years.
- ‘KSC LC-39A’ has the most successful launches from all sites.
- GEO, SSO, HEO, and ES-L1 have the best success rate.

Appendix

Thank you,

Instructors who imparted their knowledge and expertise in this course journey with exceptional skill and dedication.

Thank you!