Winning Space Race with Data Science

DIYA BHUTANI
15.09.2025
GITHUB - https://github.com/DiyaBhutani/SPACE-Y-PROJECT-

OUTLINE

EXECUTIVE SUMMARY

SUMMARY OF METHODOLOGIES

Data Collection

Data Wrangling

EDA with Visualization

EDA with SQL

Building an Interactive Map with Folium

Building a Dashboard with Plotly Dash

Predictive Analysis (Classification)

EDA RESULTS

EDA Results

Interactive Analysis

Predictive Analysis

INTRODUCTION

Project Background and Context :

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 unlike other rocket providers, SpaceX's Falcon 9 can reuse the first stage.

Problem to Solve :

If we can determine if the first stage will land, we can determine the cost of a launch . We will train a machine learning model using the previous data of launches of Falcon 9 rocket to predict if the first stage of the SpaceX will land or not .

Section 1

Methodology

METHODOLOGY

Data Collection

- SpaceX REST API
- Web Scraping from Wikipedia

Data Wrangling

- Data cleaning (e.g. replacing missing payload mass with its mean; removing irrelevant columns)
- Create a landing outcome label
- Data transformation (One Hot Encoding for categorical features, Standardization of numerical features)

Exploratory Data Analysis (EDA) using Visualization and SQL

Interactive Visual Analytics using Folium and Plotly Dash

Predictive Analysis using Classification Models

- Linear Regression, K Nearest Neighbors, Support Vector Machine, and Decision Tree models have been built and evaluated for the best classifier

DATA COLLECTION

SPACEX REST API

- SpaceX launches is gathered from the SpaceX REST

API.

- The url starts with api.spacexdata.com/v4/
- The information gathered include: the rocket used,

payload delivered, launch specifications, landing

specifications, and landing outcome.

WEB SCRAPING FROM WIKIPEDIA

- SpaceX lunches is gathered from Wikipedia:
- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- The launch records are stored in a HTML table

DATA COLLECTION – SPACEX API

1. Make a request to get rocket launch data

```
spacex_url="https://api.spacexdata.com/v4/launches/past"

response = requests.get(spacex_url)

static_json_url='https://cf-courses-data.s3.us.cloud-object-

response.status_code

200
```

2. Normalize to data frame

```
# Decode the response content as a JSON using .json()
json_data=response.json()

# Convert the json response into a dataframe using .json_normalize()
data = pd.json_normalize(json_data)
```

3. Create a dictionary

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

4. Create a Pandas data frame

```
#create a dataframe from launch_dic
data = pd.DataFrame(launch_dict)
```

5. Filter Falcon 9 only

```
# Hint data['BoosterVersion']!='Falcon 1'
data_falcon9= data[data ['BoosterVersion']!='Falcon 1']
```

6. Save to CSV

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

DATA COLLECTION – SCRAPING

1. Get Response from HTML

```
page = requests.get(static_url)
```

2. Create Beautiful Soup Object

```
content = page.text
BeautifulSoup = BeautifulSoup(content, 'html.parser')
```

3. Find tables

```
html_tables = BeautifulSoup.find_all('table')
```

4. Get column names

```
column_names = []
for i in first_launch_table.find_all('th'):
    if extract_column_from_header(i)!=None:
        if len(extract_column_from_header(i))>0:
            column_names.append(extract_column_from_header(i))
```

5. Create a dictionary

```
launch_dict= dict.fromkeys(column_names)
del launch_dict['Date and time ( )']

launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

6. Append data to keys (e.g. below for Date- refer to notebook for more)

```
#Append the date into launch_dict with key `Date`
date = datatimelist[0].strip(',')
launch_dict['Date'].append(date)
```

7. Convert dictionary to dataframe

```
df=pd.DataFrame(launch_dict)
```

8. Save to CSV

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

DATA WRANGLING

Initially some Exploratory Data Analysis (EDA) was performed:

1. Check and calculate null values

```
df.isnull().sum()/df.shape[0]*1
```

```
FlightNumber        0.000000
Date                0.000000
BoosterVersion      0.000000
PayloadMass         0.000000
```

2. Calculate the n of launches on each site

```
df['LaunchSite'].value_counts()
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

3. Calculate the n and occurrence of each orbit

```
df['Orbit'].value_counts()
```

```
GTO     27
ISS     21
VLEO    14
PO       9
```

4. Calculate the n and occurrence of outcome per orbit type

```
landing_outcomes= df['Outcome'].value_counts()
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
```

5. Create a landing outcome label from Outcome column

```
def outcome_to_class(outcome_value):
    if outcome_value in bad_outcomes:
        return 0
    else:
        return 1
landing_class = df['Outcome'].apply(outcome_to_class)
```

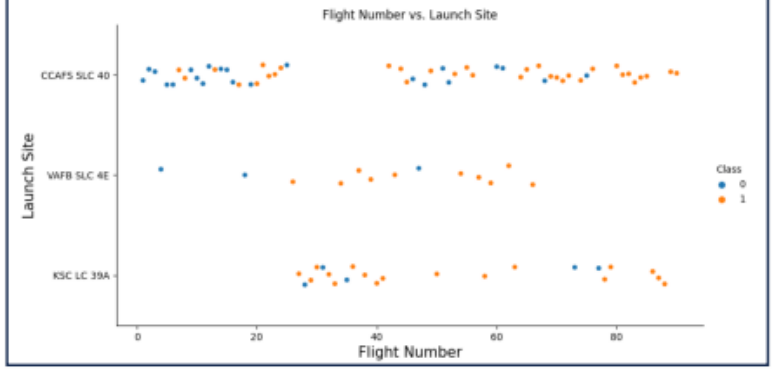
6. Calculate success from Class column

```
df["Class"].mean()
```

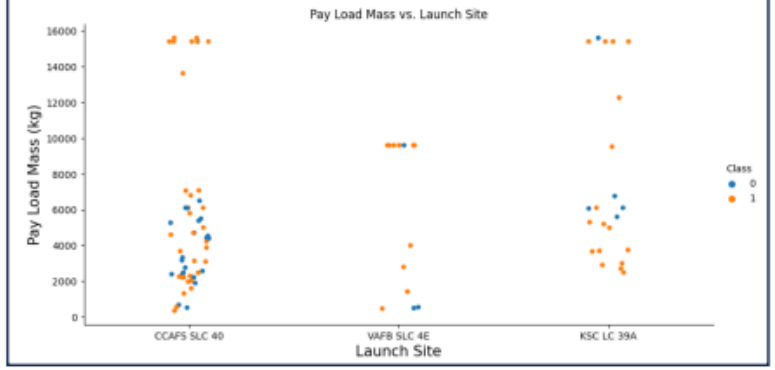
```
0.6666666666666666
```

EDA WITH DATA VISUALIZATION

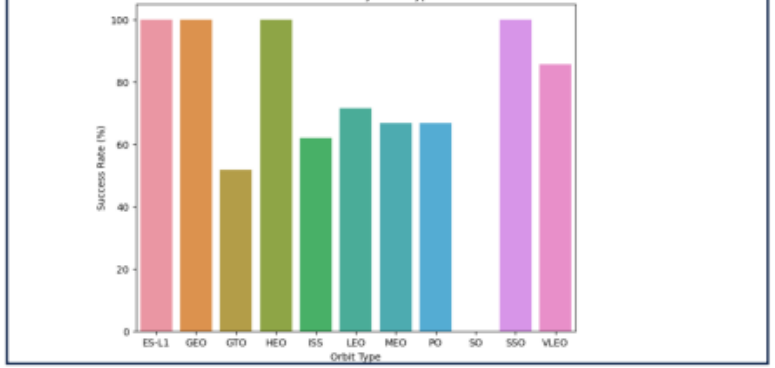
Flight Number and Launch Site



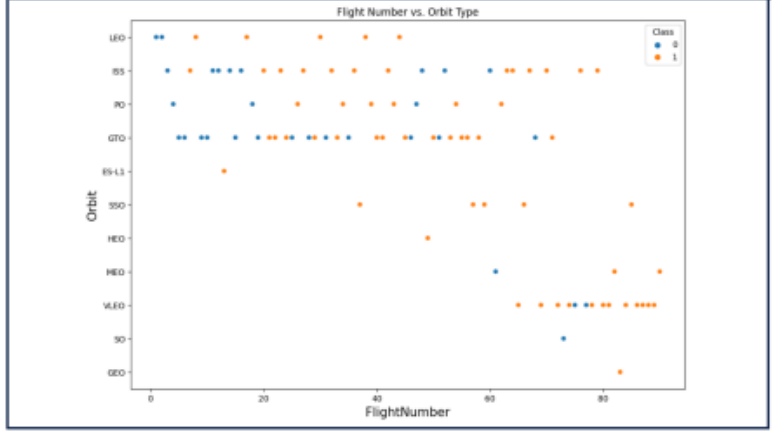
Payload Mass and Launch Site



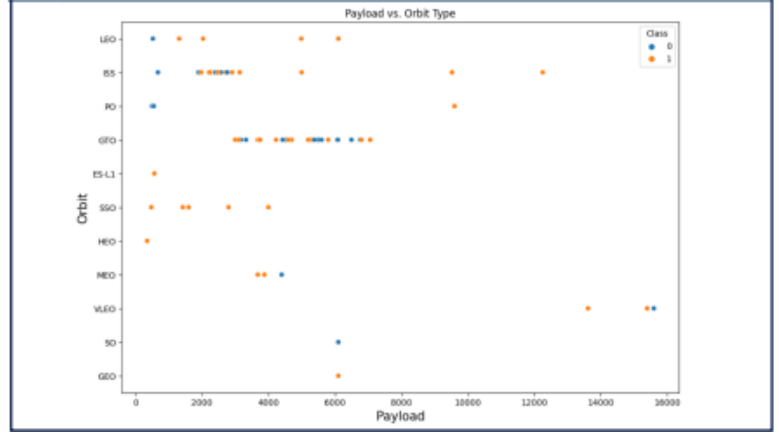
Success Rate and Orbit Type



Flight Number and Orbit Type



Payload Mass and Orbit Type



Trend of success over the years

EDA WITH SQL

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'KSC
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date where the successful landing outcome in drone ship was achieved.
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the records which will display the month names, successful landing outcomes in ground pad booster versions, launch site for the months in

  year 2017
- Ranking the count of successful landing outcomes between the date 2010 06 04 and 2017 03 20 in descending order.

BUILDING AN INTERACTIVE MAP WITH FOLIUM

An interactive map offers a dynamic way to visualize and analyze launch site data, facilitating decision-making and insights into the spatial relationships of launch activities. For instance, the success rate may depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories.

We created three maps for:

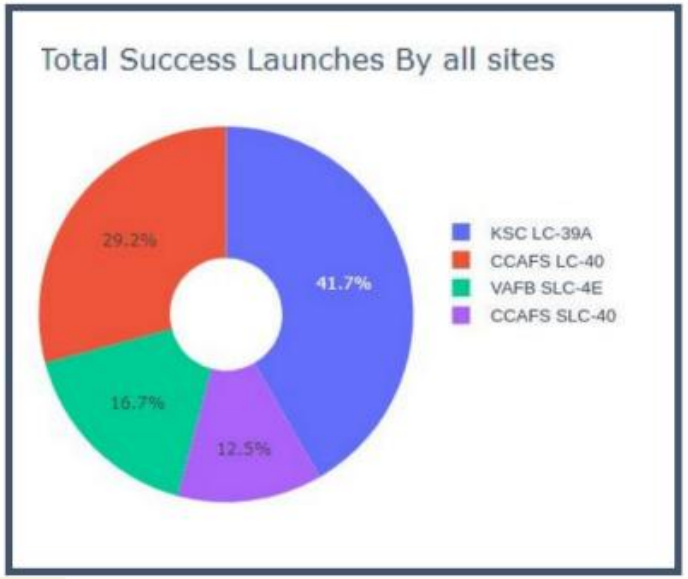
- Visualization of each launch site in an interactive map.
- Visualization of launch sites on the map based on fail or success. This gives us insights into the performance of launch sites.
- Visualization of the geographical context and proximity of the launch sites to railway, highway, coastline, etc

BUILDING A DASHBOARD WITH PLOTLY-DASH

Plotly Dash applications allows us to perform interactive visual analytics on SpaceX lunch data in real time.

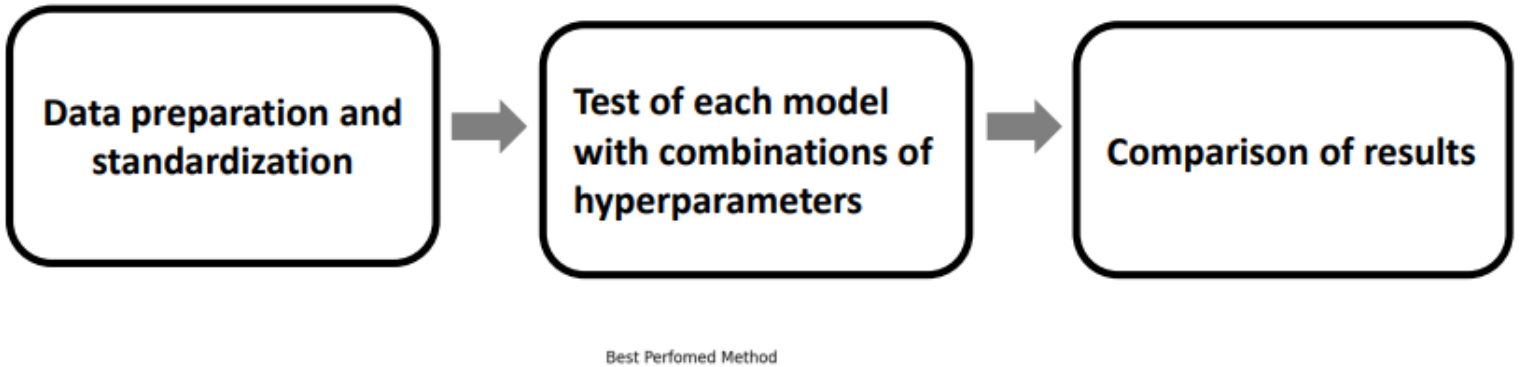
We developed a dashboard that enables us to:

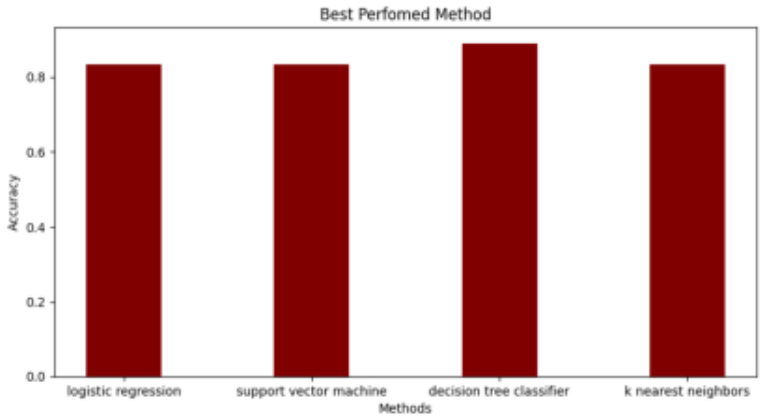
- Visualize a pie chart depicting success based on the selected launch site.
- Explore the correlation between success and payload by utilizing a Range Slider to choose the payload range, presented through a scatter plot.



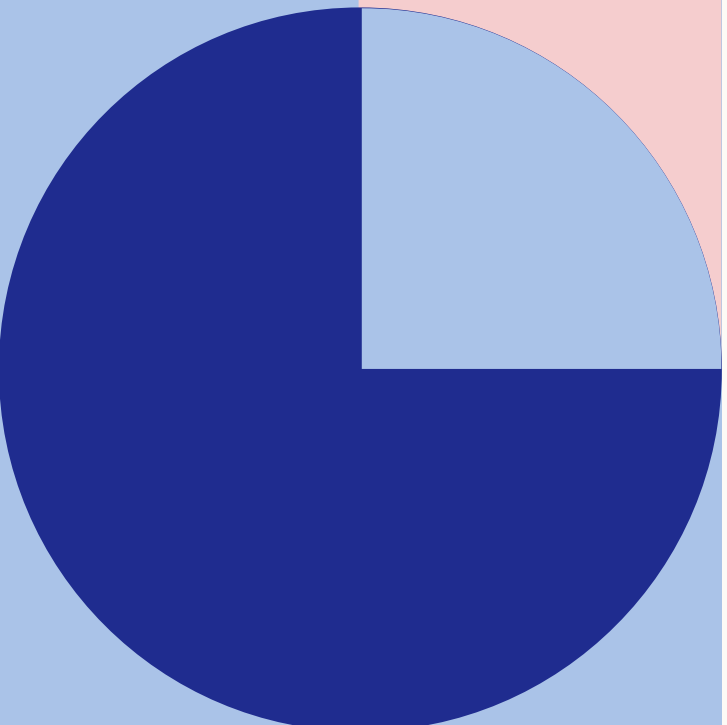
PREDICTIVE ANALYSIS (CLASSIFICATION)

Four classification models including Linear Regression, K Nearest Neighbors, Decision Tree models, and Support Vector Machine, were built and compared.





RESULTS

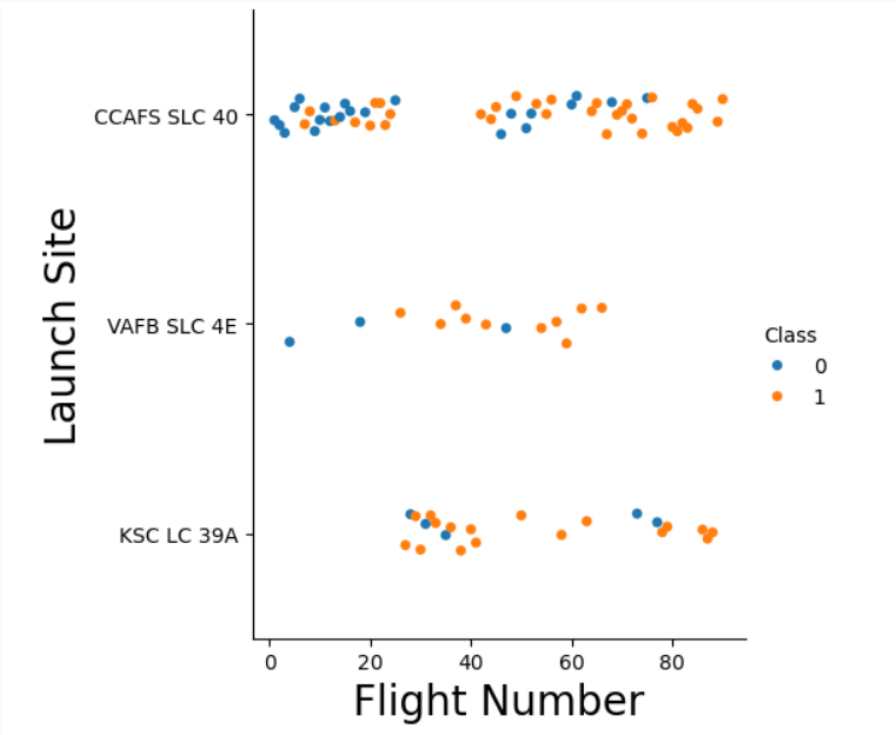


- The first successful ground pad landing took place on December 22, 2015.
- The success rates for SpaceX launches increased over time (2013-2020).
- The location of launch appears to be a significant contributing factor to the success of missions
- KSC LC-39A has the most successful launches compared to other sites.
- Low weighted payloads perform better than the heavier payloads.
- ES-L1, GEO, HEO, and SSO orbits achieved the highest success rate.
- For heavy payloads, the rates of success are higher for VLO and ISS orbits.
- Decision Tree model is the best in terms of prediction accuracy for this dataset.

Section 2

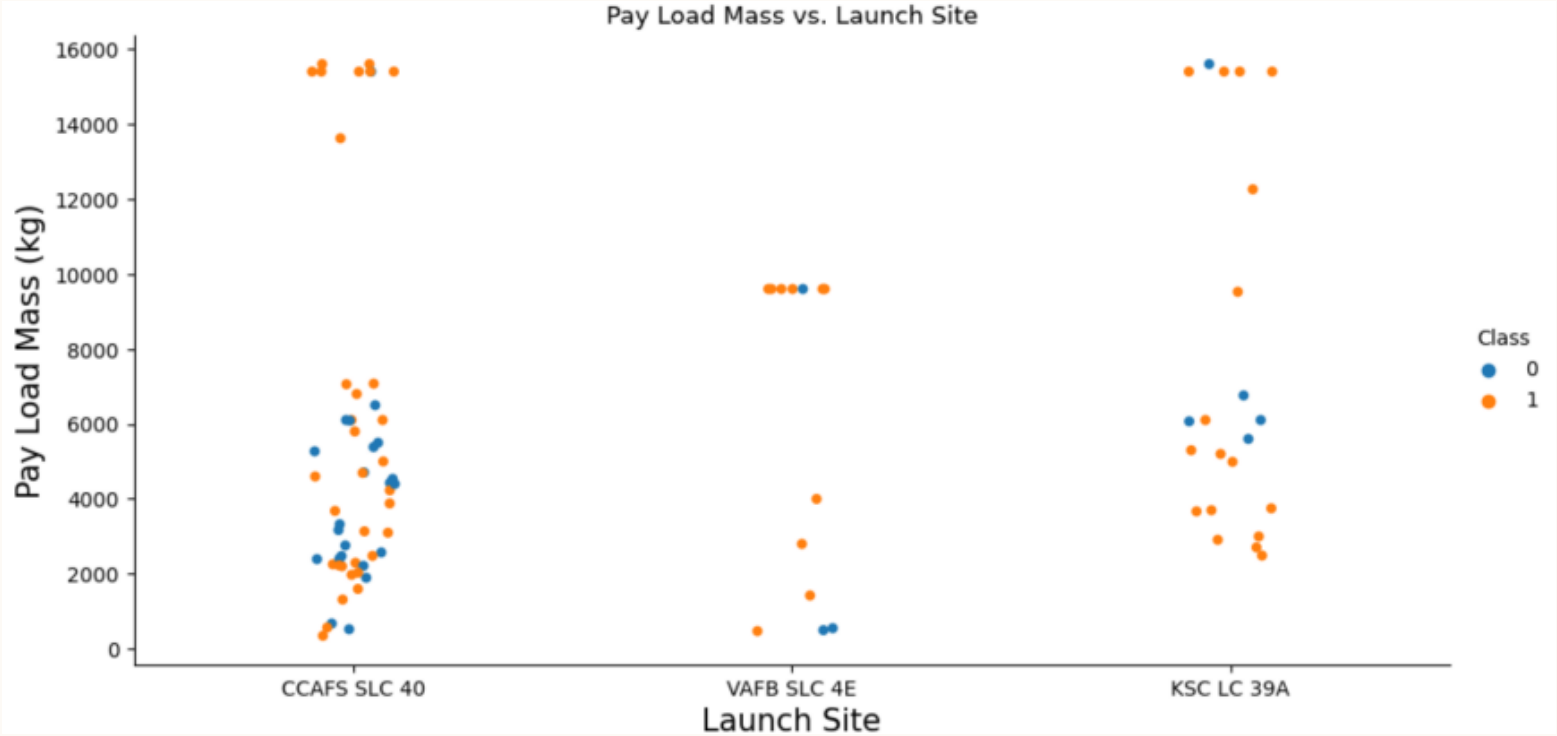
Insights drawn from EDA

FLIGHT NUMBER V.S. LAUNCH SITE



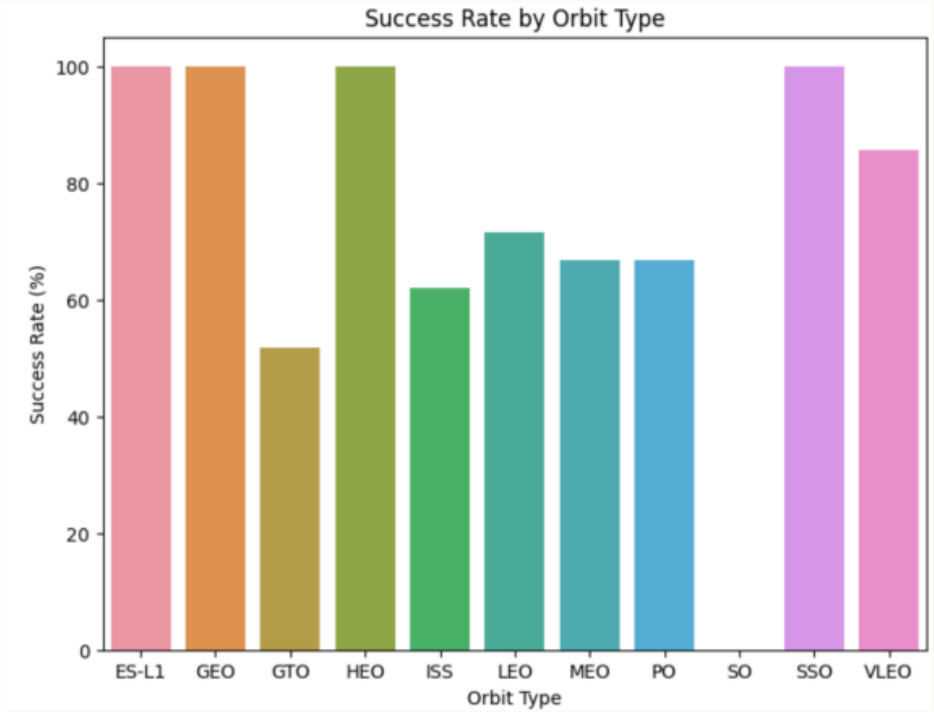
- Most launches took off from CCAFS SLC 40 launch site
- The initial launches were mostly carried out from the CCAFS SLC 40 launch site
- The majority of recent launches from CCAFS SLC 40 resulted in success
- The success rates for all launch sites improved over time.

PAYLOAD V.S. LAUNCH SITE



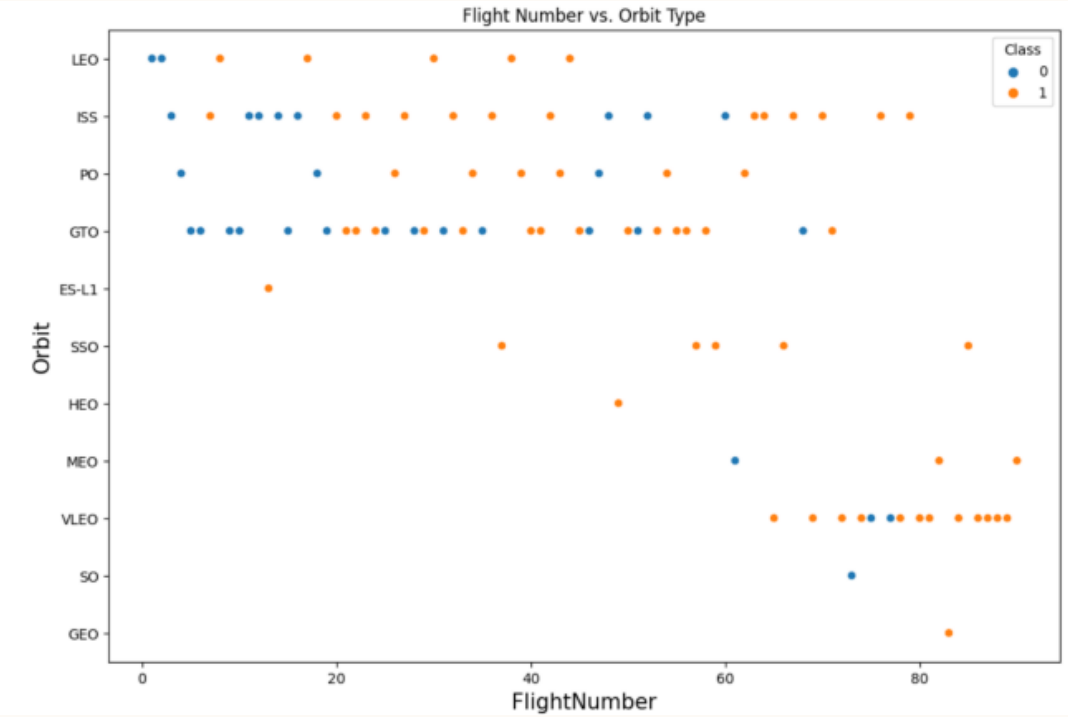
- VAFB SLC 4E launch site conducts launches with lower payloads (zero launches for >10000 kg)
- CCAFS SLC 40 hosts a higher number of launches involving both higher and lower payloads.
- Most payloads weighing over 9000 kg have achieved successful outcomes.

SUCCESS RATE V.S. ORBIT TYPE



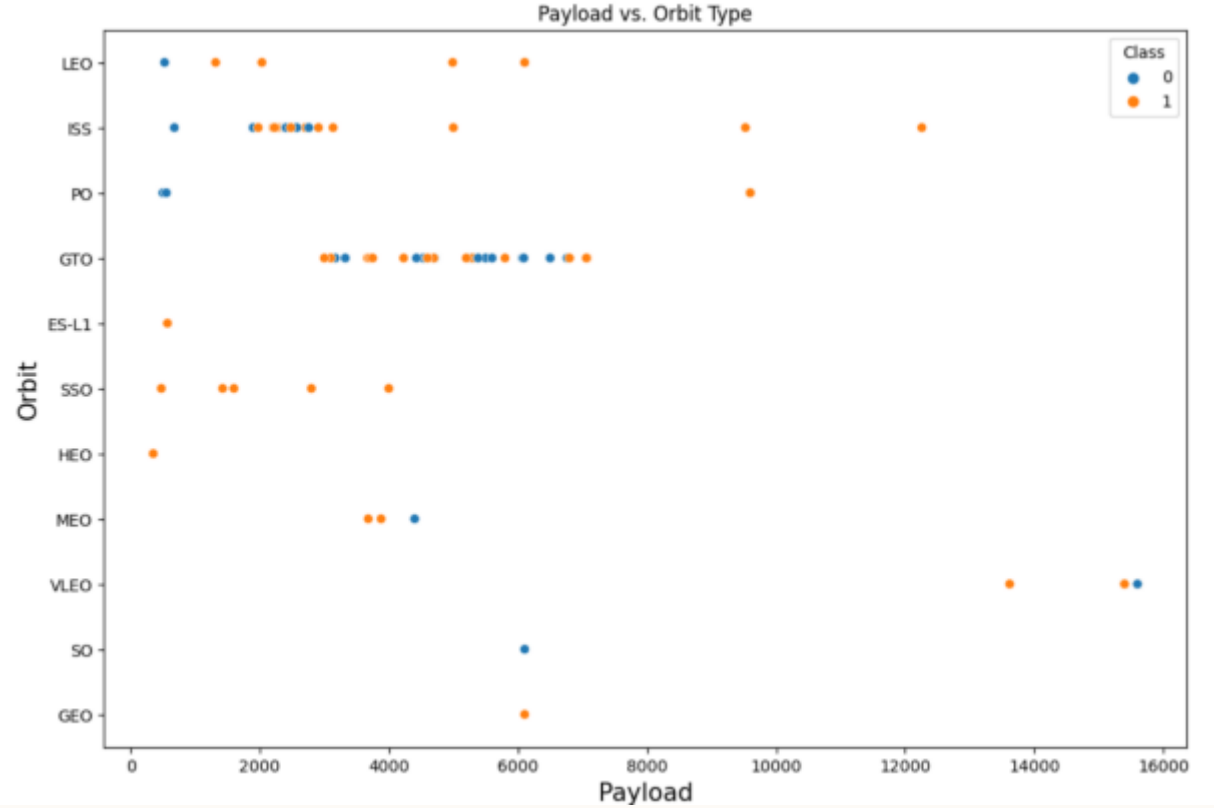
ES-L1, GEO, HEO, and SSO orbits achieved the highest success rate at 100%, followed by VLEO with a success rate >80%, and LEO with a success rate >70%.

FLIGHT NUMBER V.S. ORBIT TYPE



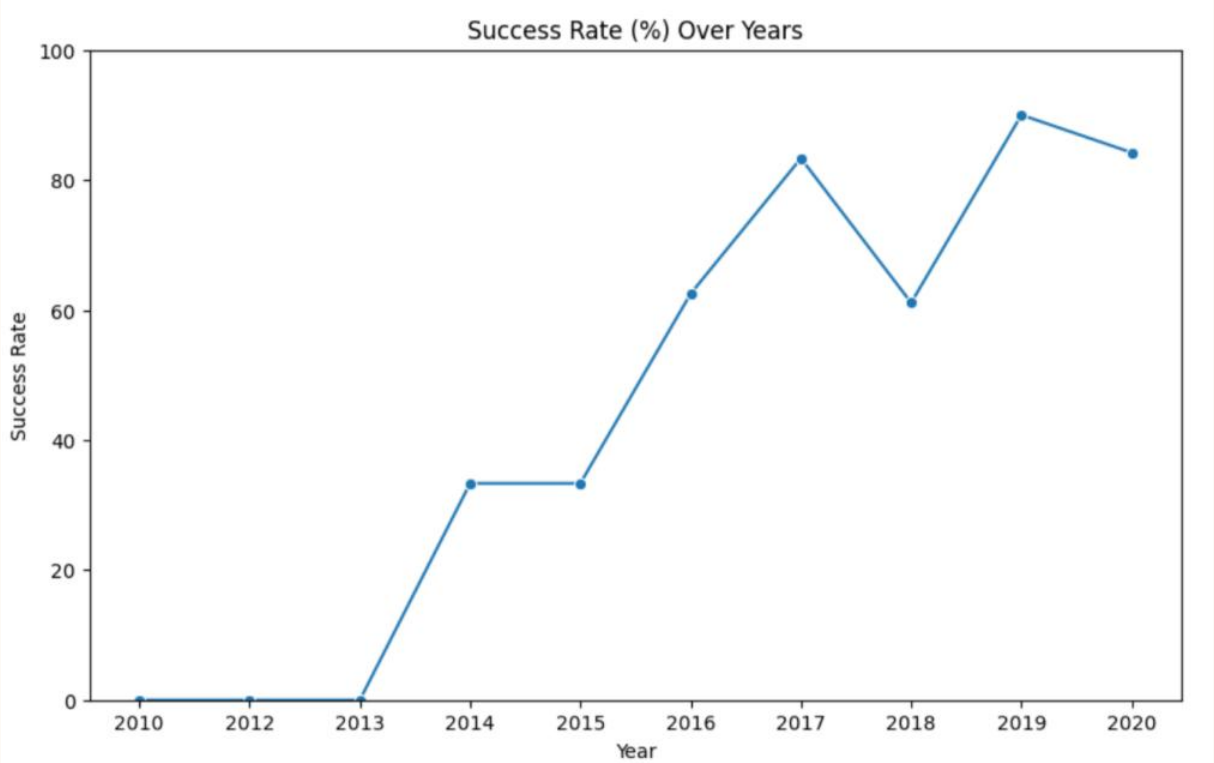
- In the recent years, there has been a transition towards launching missions into Very Low Earth Orbits (VLEO) with a significantly high rate of success.
- While the GTO orbit experiences a low success rate, there appears to be no discernible relationship between flight number and the rate of success in this orbit.

PAYLOAD MASS V.S. ORBIT TYPE



- For heavy payloads, the rates of success are higher for VLO and ISS orbits.
- In the case of GTO, there seems no apparent relationship between payload and the rate of success.

LAUNCH SUCCESS YEARLY TREND



- The rate of success has seen a notable rise since 2013 and continued to rise until 2020, possibly attributed to technological advancements.
- The first three years (2010-2013) seem to have been a phase focused on fine-tuning and technological enhancement.

ALL SITE NAMES

Here are the names of the 4 launch sites-

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

Associated SQL Query -

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

LAUNCH SITES THAT BEGIN WITH CCA

Five records where launch sites begin with the 'CCA' are listed below:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Associated SQL Query -

```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

TOTAL PAYLOAD MASS

The total payload carried by boosters from NASA (CRS) is 45596 (kg)

Associated SQL Query -

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_payload_mass FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

AVERAGE PAYLOAD MASS

The average payload mass carried by booster version F9 v1.1 is 2928.4 (kg)

Associated SQL Query -

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

FIRST SUCCESSFUL GROUND LANDING DATE

The first successful ground pad landing took place on December 22, 2015.

Associated SQL Query -

```
%sql SELECT MIN(DATE) AS First_successful_landing_gound FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
```

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

The boosters that have successfully landed on drone ships and carried payloads with a mass exceeding 4000 but below 6000 are as follows:

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

Associated SQL Query -

```
%%sql SELECT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_>4000 AND PAYLOAD_MASS__KG_<6000;
```

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

The total number of 100 mission outcomes, including both successful and unsuccessful results is presented in the following breakdown:

total_success	total_failure
100	1

Associated SQL Query -

```
%%sql SELECT Mission_Outcome, COUNT(*) AS Total_count
FROM SPACEXTABLE
GROUP BY Mission_Outcome;
```

BOOSTERS CARRIED MAXIMUM PAYLOAD

The boosters which have carried the maximum payload mass include:

Associated SQL Query -

```
%%sql SELECT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTABLE
);
```

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

2015 LAUNCH RECORDS

The unsuccessful landing outcomes on drone ships, along with their corresponding
booster versions and launch site names, during the year 2015, are provided below:

Month	Failure_Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Associated SQL Query -

```
%%sql SELECT
 CASE
        WHEN Date like '%-01-%' THEN 'January'
        WHEN Date like '%-02-%' THEN 'February'
        WHEN Date like '%-03-%' THEN 'March'
        WHEN Date like '%-04-%' THEN 'April'
        WHEN Date like '%-05-%' THEN 'May'
        WHEN Date like '%-06-%' THEN 'June'
        WHEN Date like '%-07-%' THEN 'July'
        WHEN Date like '%-08-%' THEN 'August'
        WHEN Date like '%-09-%' THEN 'September'
        WHEN Date like '%-10-%' THEN 'October'
        WHEN Date like '%-11-%' THEN 'November'
        WHEN Date like '%-12-%' THEN 'December'
    END AS Month,
    Landing_Outcome AS Failure_Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM SPACEXTABLE
WHERE Date like'%2015%' AND Landing_Outcome LIKE 'Failure (drone ship)';
```

RANK LANDING OUTCOMES BETWEEN 2010-06-04 AND 2017-03-20

- The tally of landing results (including Failure on drone ships or Success on ground pads) occurring between June 4, 2010, and March 20, 2017, is presented in a descending order

- As shown, the highest count among the landing outcomes is attributed to "No attempt", totaling 10. This is followed by "Success (ground pad) ", "Success (drone ship) ", and "Failure (drone ship) " each occurring 5 times, while "Controlled (ocean)" is observed 3 times.

Landing_Outcome	Outcome_Count
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

Associated SQL Query -

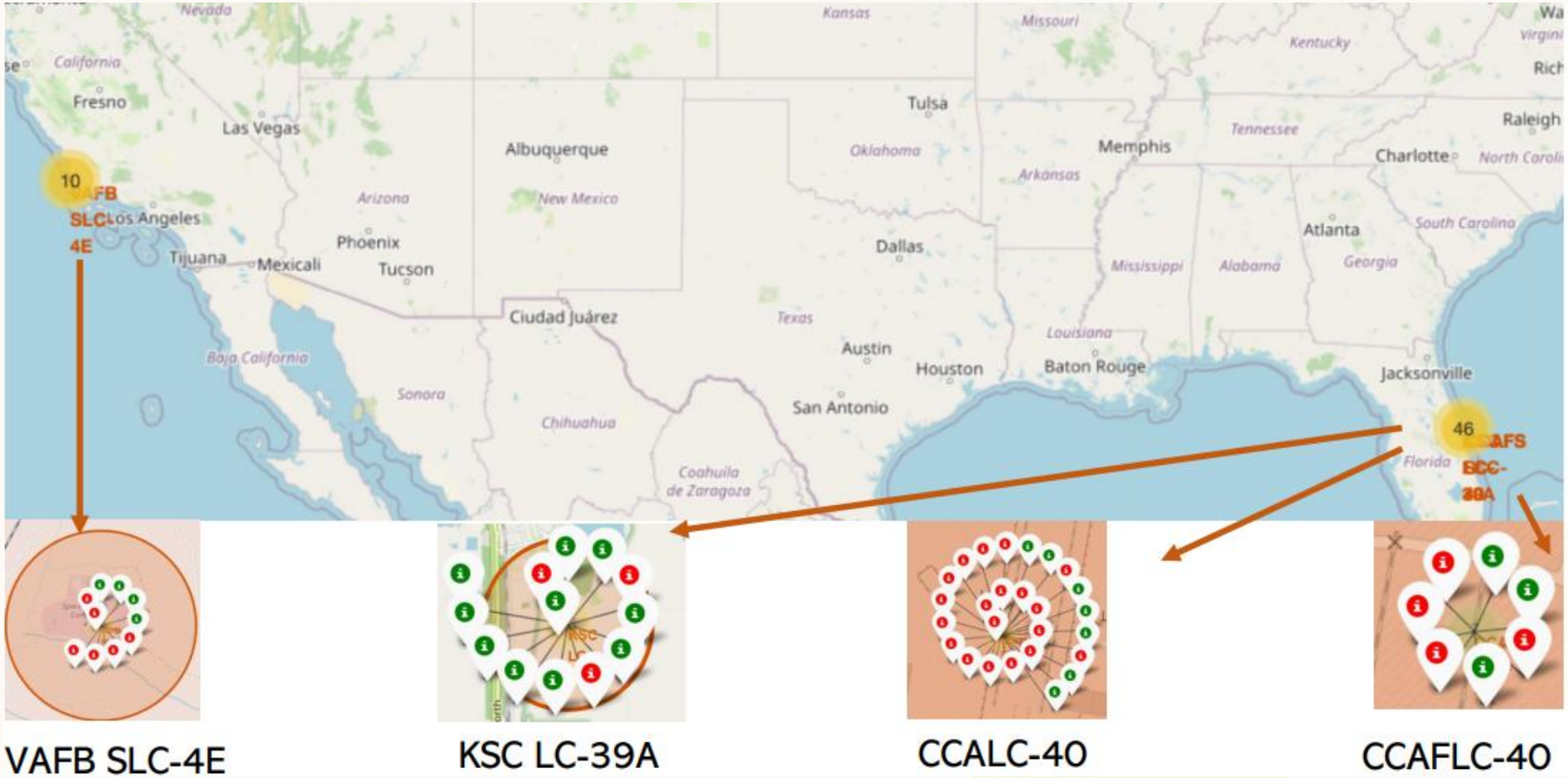
```
%%sql SELECT Landing_Outcome, COUNT(*) AS Outcome_Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Outcome_Count DESC;
```


Section 3
Launch Sites
Proximities Analysis

FOLIUM MAP – LAUNCH SITE LOCATIONS

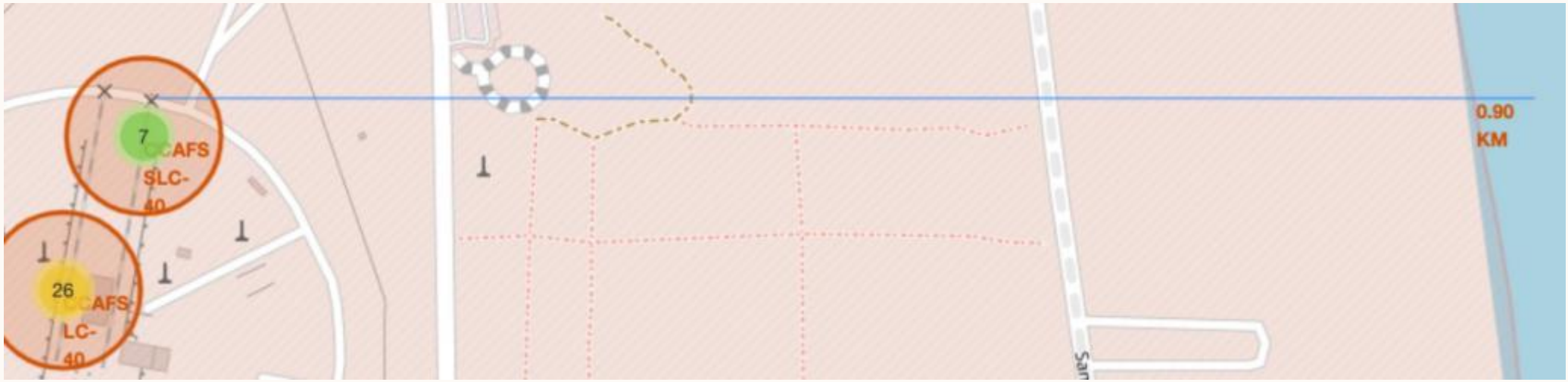
VAFB SLC-4E is situated near the western coastline while KSC LC-39A,CCAFLC-40, and CCAFSLC-40 are positioned along the eastern coastline
Upon zoomin in, it seems both the ,CCAFLC-40 and CCAFSLC-40 are situated in very close proximity to each other

FOLIUM MAP– SUCCESS/FAILED LAUNCHES FOR LAUNCH SITES

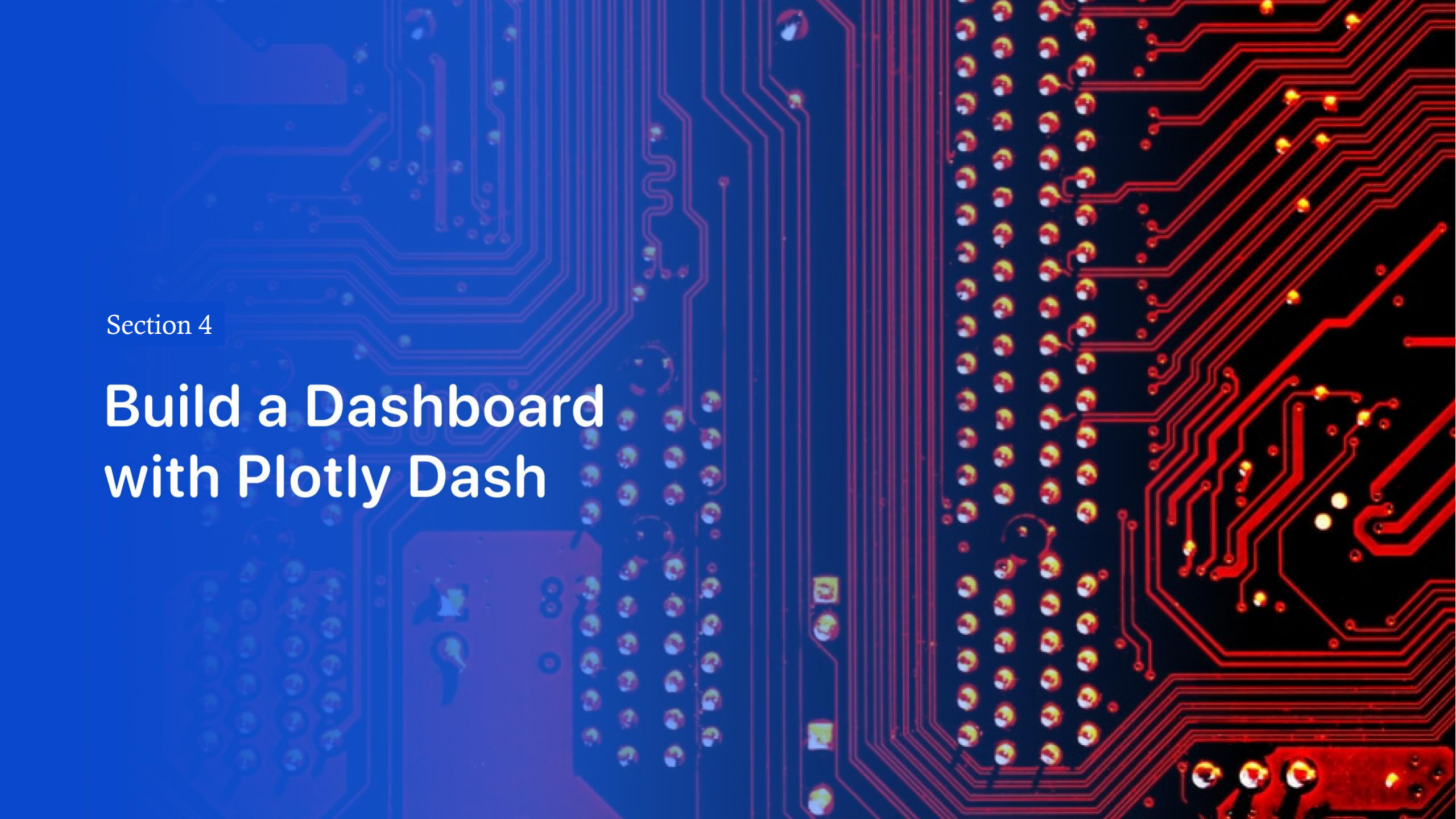


- Upon zooming in, we can see the success (green) and failure (red) marks for each site.
- Out of 13 launches, KSC LC-39A has achieved the highest success rate with 10 successful missions (10/13=76.9%)

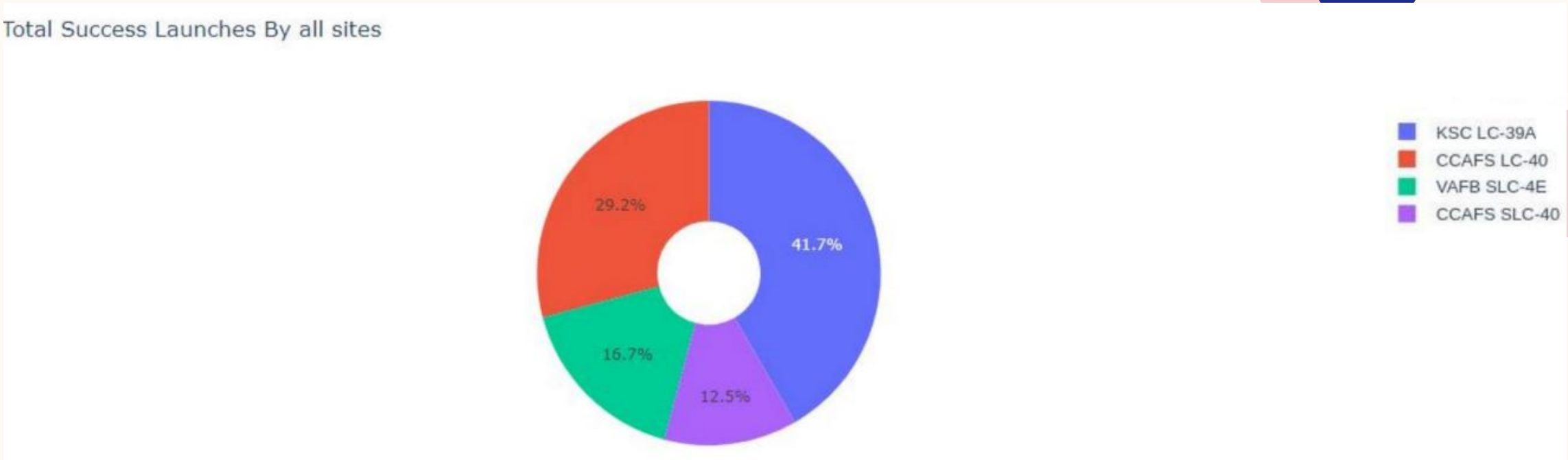
FOLIUM MAP– PROXIMITY OF LAUNCH SITES TO OTHER AREAS



These figures show a PolyLine between CCAFSLC-40 to the selected coastline point, etc.

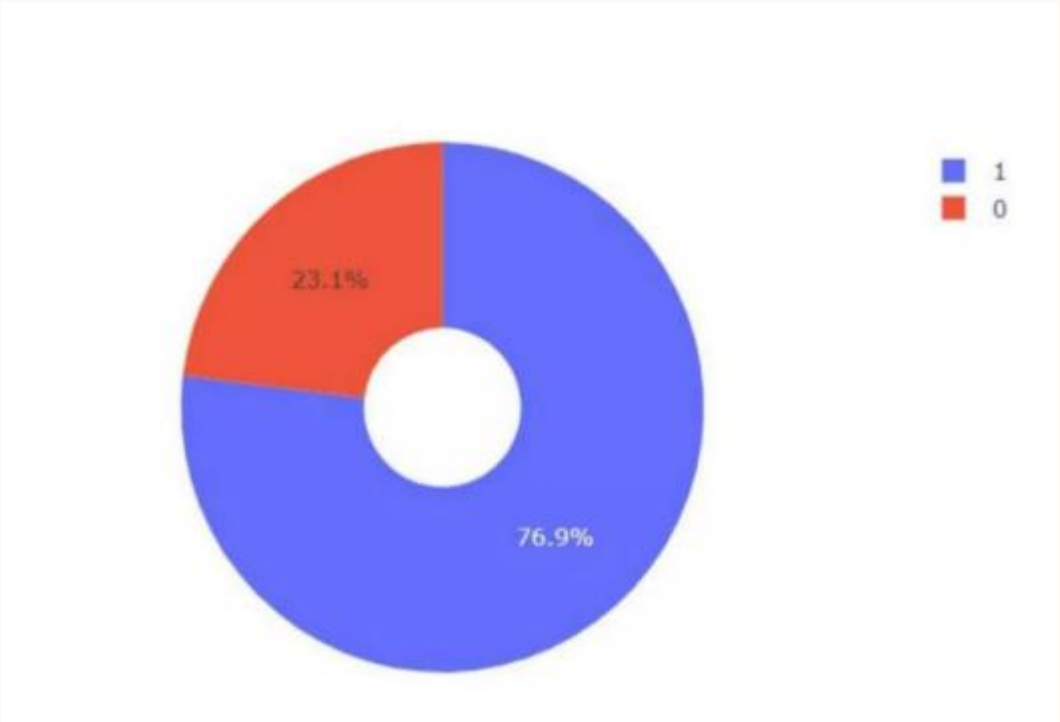

Section 4
Build a Dashboard with Plotly Dash

DASHBOARD – LAUNCH SITE WITH HIGHEST LAUNCH SUCCESS RATIO



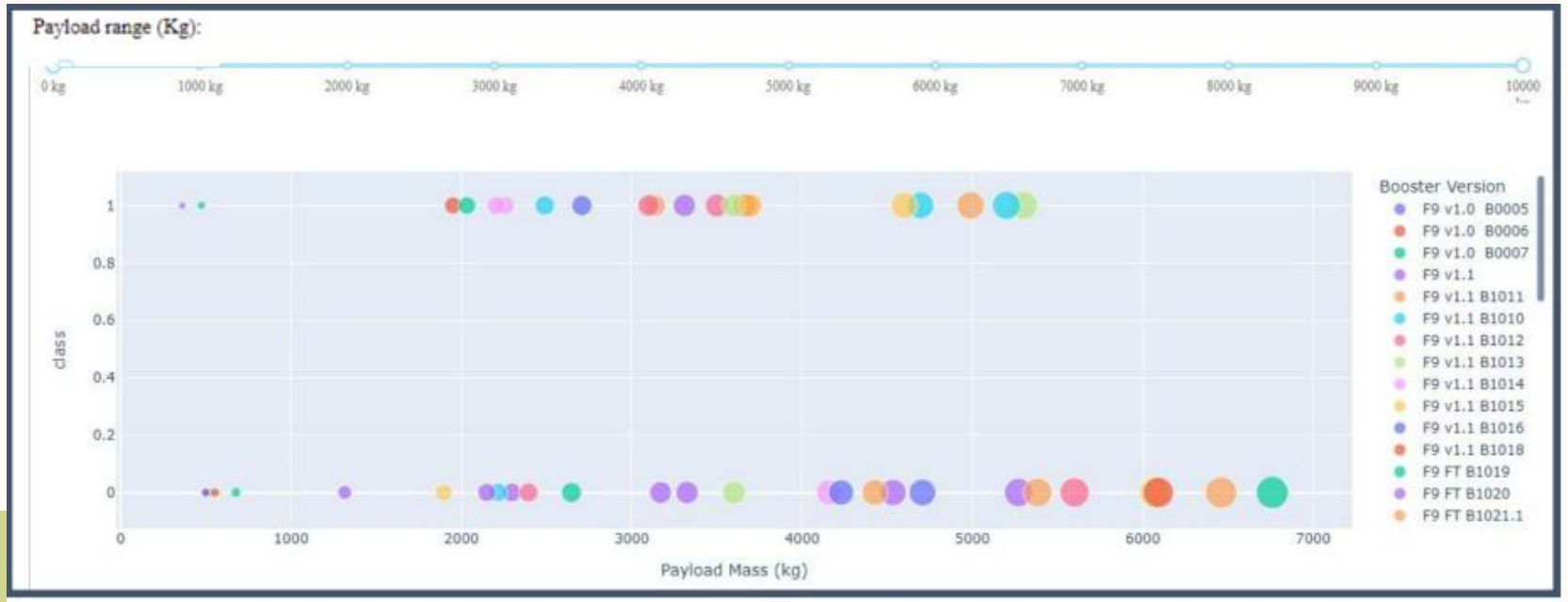
- The location of launch appears to be a significant contributing factor to the success of missions.
- KSC LC-39A has the most successful launches compared to other sites.

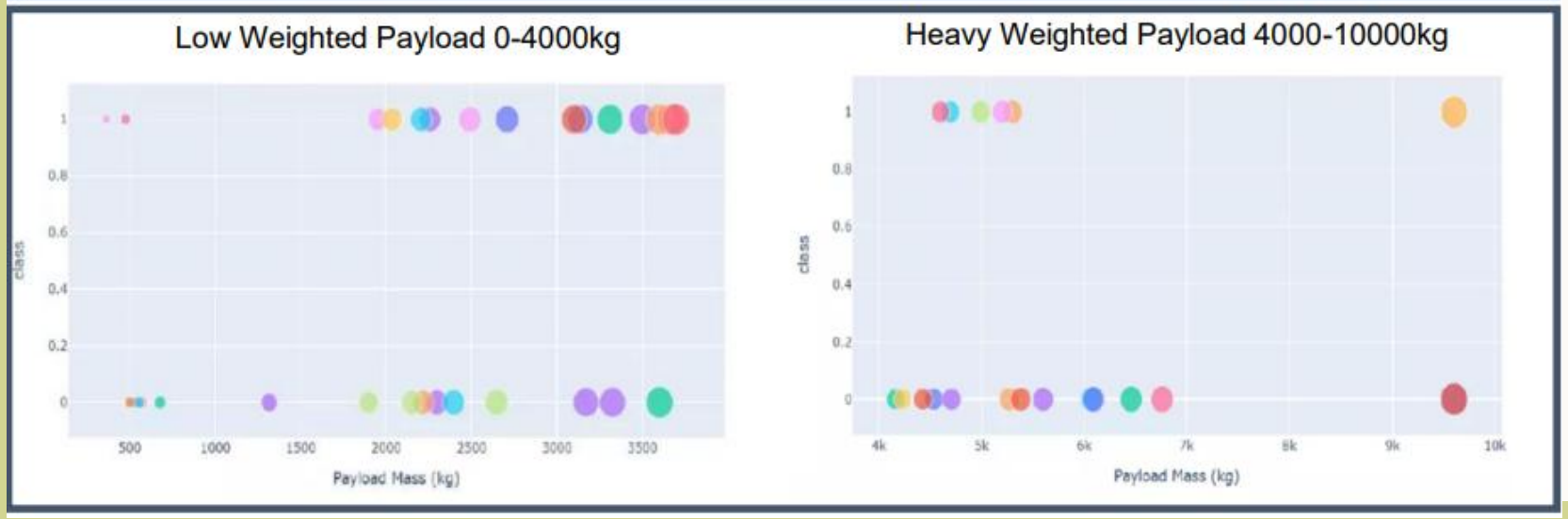
DASHBOARD – TOTAL SUCCESS LAUNCHES BY ALL SITES



- Using the dropdown menu on the dashboard allows for viewing single site launches.
- At KSC LC-39A, 76.9% of the launches resulted in success, while 23.1% experienced failure.

DASHBOARD – PAYLOAD VS. LAUNCH OUTCOME





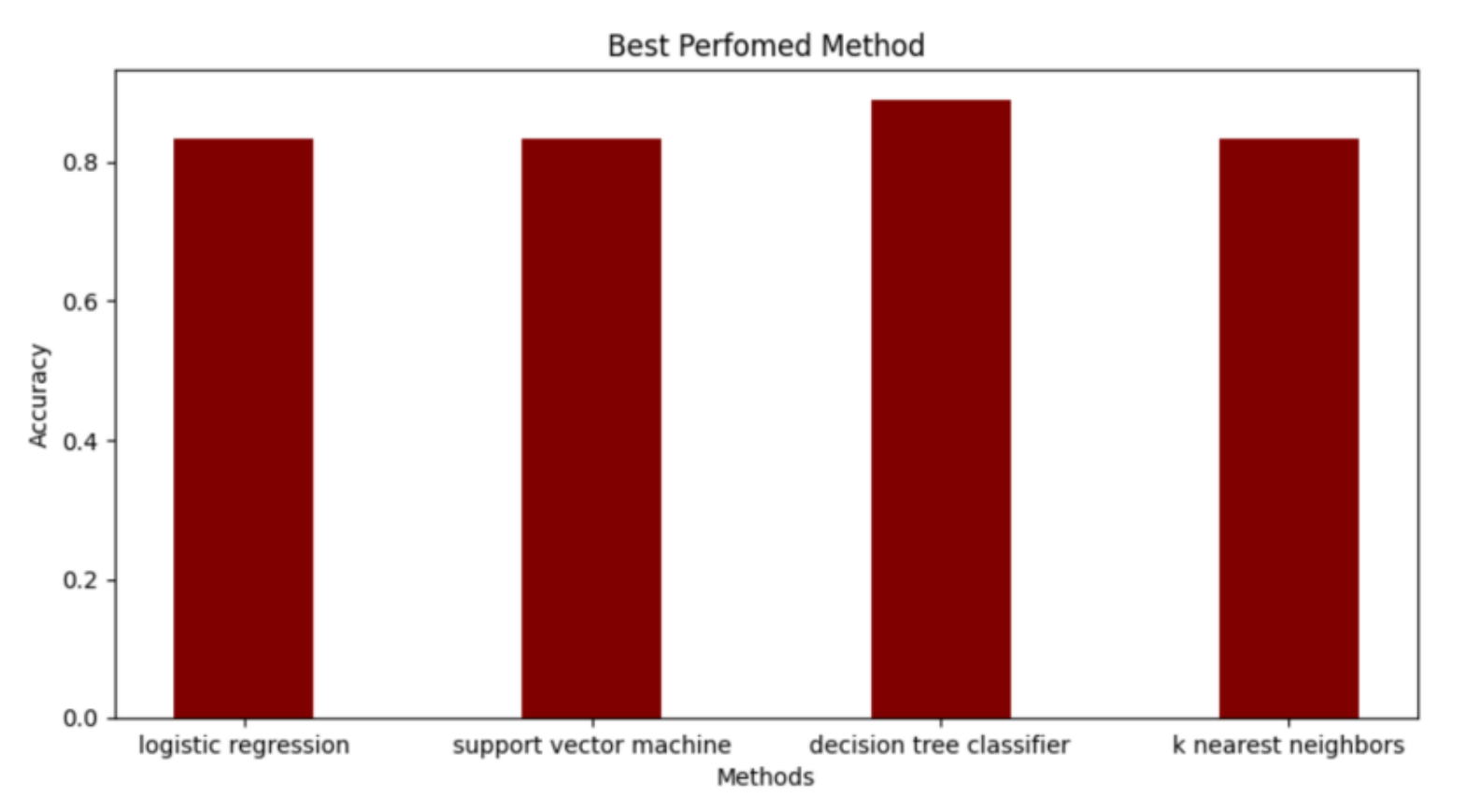
- With the Range Slider, we can observe the outcomes of both successful and failed launches for each booster version, along with the corresponding payload they carried.

- Success rates are higher for lighter payloads compared to heavier ones.

Section 5

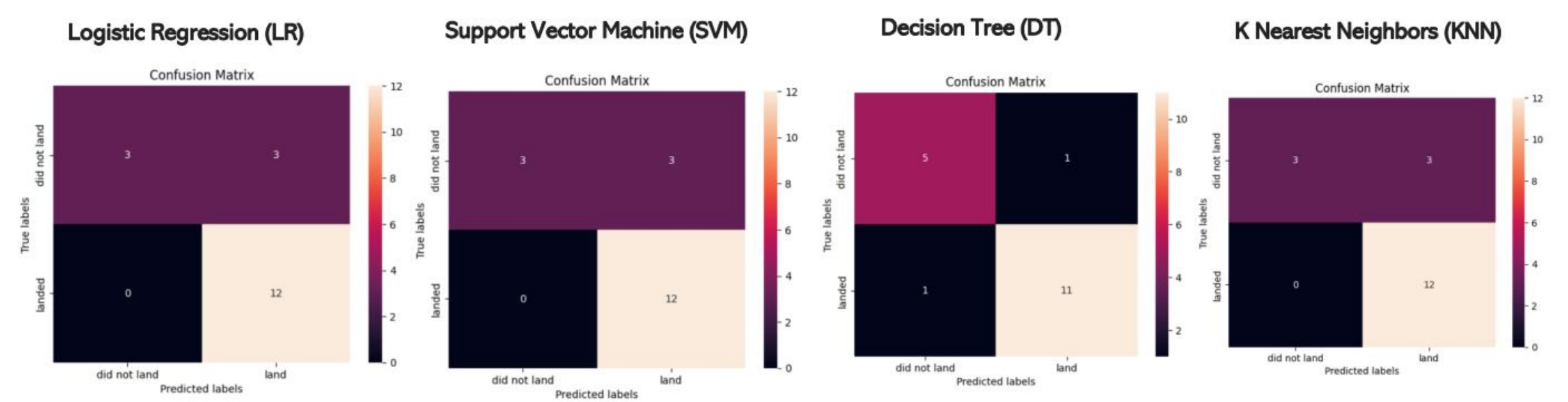
Predictive Analysis (Classification)

CLASSIFICATION ACCURACY



Decision Tree model achieved the highest accuracy at 88 %, while the SVM performs the best in terms of Area Under the Curve at 0.96 (not shown).

CONFUSION MATRIX



- All four models can distinguish between the different classes.
- However, the major problem for LR, SVM, and KNN is False Positives (n=3)
- Decision Tree has the least False Positive (n=1)

CONCLUSIONS

- The success rates for SpaceX launches increased over time (2013-2020).
- KSC LC-39A has the most successful launches compared to other sites.
- Low weighted payloads perform better than the heavier payloads.
- ES-L1, GEO, HEO, and SSO orbits achieved the highest success rate.
- For heavy payloads, the rates of success are higher for VLO and ISS orbits.
- Decision Tree models are the best in terms of prediction accuracy for this dataset.
- Through the utilization of available data and comprehensive analysis, rocket companies can pinpoint the most effective techniques for diminishing launch expenses. This approach averts the risk of losing clients and ensures they remain competitive in the market.

Thank you!