IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mitsu Kansagara
July 30, 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- **Summary of methodologies**
  - ✓ Data Collection using SpaceX API
  - ✓ Data Collection using Web Scraping Wikipedia Pages
  - ✓ Data Wrangling
  - ✓ Exploratory Data Analysis with SQL
  - ✓ Exploratory Data Analysis with Data Visualization
  - ✓ Interactive Visual Analytics with Folium
  - ✓ Machine Learning Prediction
- **Summary of all results**
  - ✓ Leveraged the data collection from public sources
  - ✓ EDA allowed to identify best features for predicting success of launching
  - ✓ ML prediction helped in identifying the optimal model for accurately determining the crucial characteristics that contribute to maximizing the potential of this opportunity

Introduction

- **Project background and context**
  - ✓ The objective is to evaluate the viability of the new company Space Y to compete with Space X.
  - ✓ Space X 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 Space X 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 X for a rocket launch. This goal of the project is to create a machine learning algorithm to predict if the first stage will land successfully.
- **Problems you want to find answers**
  - ✓ What factors determine if the rocket will land successfully?
  - ✓ The interaction amongst various features that determine the success rate of a successful landing.
  - ✓ What operating conditions needs to be in place to ensure a successful landing program.
  - ✓ The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets;
  - ✓ Where is the best place to make launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from SpaceX was retrieved from the following sources:
    - ✓ Web Scraping Wiki Pages (link)
    - ✓ SpaceX API (link)
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome based on outcome data after summarizing and analyzing features
  - One-hot encoding was applied to categorical features

Methodology

Continued Executive Summary

- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - ✓The data collected up to this point has been normalized and split into separate training and test datasets. These datasets were then evaluated using four different classification models, with the accuracy of each model assessed using various parameter combinations.

Data Collection

- The data was collected using various methods
  - ✓Data collection was done using get request to the SpaceX API
  - ✓Decoded the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize()
  - ✓Cleaned the data, checked for missing values and fill in missing values where necessary
  - ✓In addition, I performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup
  - ✓The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for future analysis

Data Collection – SpaceX API

- I used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting

- Source Code: [GitHub Mitsu Kansagara Data Collection API]

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

In [11]:
```
# Use json_normalize method to convert the json result into a dataframe
data = pd.json_normalize(response.json())
```

Calculate below the mean for the `PayloadMass` using the `.mean()`. Then use the mean and the `.replace()` function to replace `np.nan` values in the data with the mean you calculated.

In [27]:
```
# Calculate the mean value of PayloadMass column
payload_mean = data_falcon9['PayloadMass'].mean()
payload_mean
```

Out[27]: 6123.547647058824

In [28]:
```
# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'] = data_falcon9['PayloadMass'].replace(np.nan, payload_mean)
```

Now let's start requesting rocket launch data from SpaceX API with the following URL:

In [6]:
```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

In [7]:
```
response = requests.get(spacex_url)
```

Data Collection – Web Scraping

- Applied web scrapping to web scrap Falcon 9 launch records with BeautifulSoup from Wikipedia page
- Parsed the table and converted it into a pandas dataframe
- Source Code: GitHub Mitsu Kansagara Data Collection with Web Scrapping

In [4]:
```
static_url = "https://en.wikipedia.org/w/index.php?title=List_c
```

In [5]:
```
# use requests.get() method with the provided static_url
# assign the response to a object
#html_data is our response

html_data = requests.get(static_url)
html_data.status_code
```

Out[5]: 200

Create a `BeautifulSoup` object from the HTML `response`

In [6]:
```
# Use BeautifulSoup() to create a BeautifulSoup object from a r
soup = BeautifulSoup(html_data.text, 'html.parser')
```

Print the page title to verify if the `BeautifulSoup` object was created properly

In [7]:
```
# Use soup.title attribute
soup.title
```

In [10]:
```
column_names = []

# Apply find_all() function with `th` element on first_launch_t
# Iterate each th element and apply the provided extract_columr
# Append the Non-empty column name (`if name is not None and le

row = first_launch_table.find_all('th')
for header_name in row:
    name = extract_column_from_header(header_name)
    if name is not None and len(name) > 0:
            column_names.append(name)
```

Data Wrangling

1

Initially some Exploratory Data Analysis (EDA) was performed on the dataset.

2

Summarized launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.

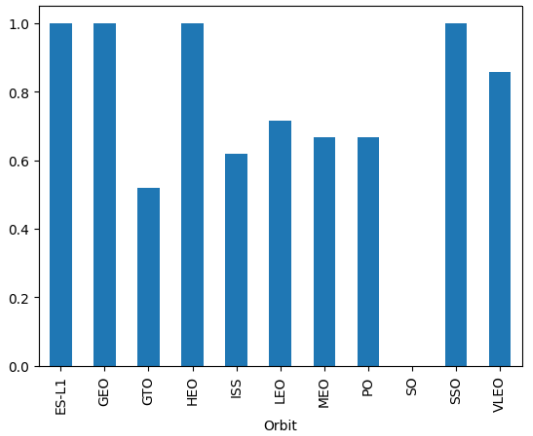
3

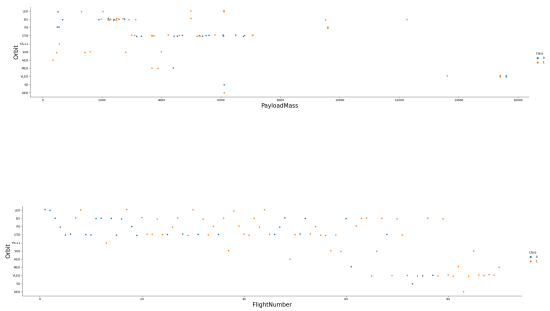
Finally, the landing outcome label was created from Outcome column.

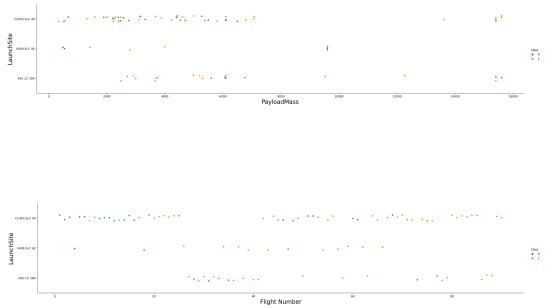
Source Code: GitHub Mitsu Kansagara Data Wrangling

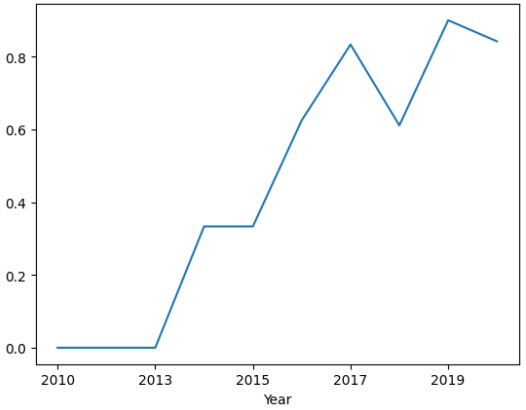
EDA with Data Visualization

- To explore data, scatterplots and bar plots were used to visualize the relationship between pair of features:
- Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit









- Source Code: [GitHub Mitsu Kansagara EDA with Data Visualization]

EDA with SQL

- Loaded SpaceX dataset into Postgre SQL database and performed following queries to get valuable insights from our data:
  - ✓ 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. Use a subquery
  - ✓ List records which will display month names, failure landing outcomes in drone ship ,booster versions, launch site for months in year 2015.
  - ✓ Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.
- Source Code: GitHub Mitsu Kansagara EDA using SQL

Interactive Map with Folium

- Created and added map objects such as markers, circles, lines, etc. to mark the success or failure of launches for each site on the folium map
  - ✓ Assigned the feature, launch outcomes (class 0 – failure, class 1 – success)
  - ✓ Markers indicate points like launch sites
  - ✓ Circles indicate highlighted areas specific coordinates, like NASA Johnson Space Centre
  - ✓ Marker clusters indicate groups of events in each coordinate, like launches in a launch site
  - ✓ Lines are used to indicate groups of events in each coordinate like launches in a launch site
- Calculated the distance between a launch site to its proximities. Answered following questions:
  - ✓ Are launch sites in-close proximity to railways?
  - ✓ Are launch sites in-close proximity to highways?
  - ✓ Are launch sites in-close proximity to coastline?
  - ✓ Do launch sites keep certain distance away from cities?
- Source Code: GitHub Mitsu Kansagara Interactive Visual Analytics with Folium

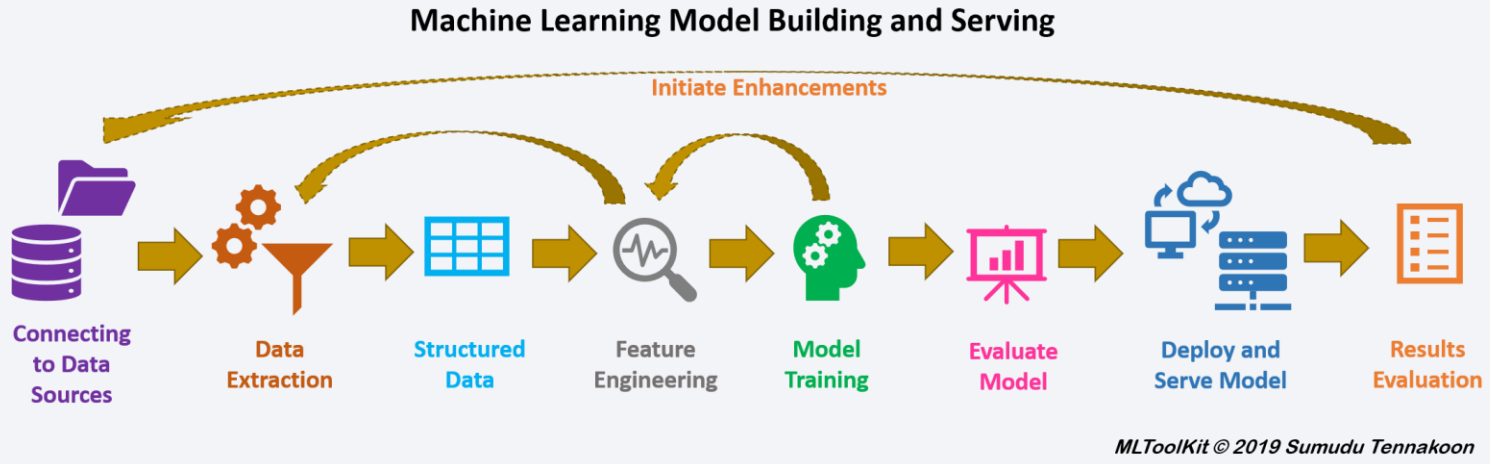
Interactive Dashboard with Plotly Dash

- Built an interactive dashboard with Plotly dash to visualize data:
  - ✓Percentage of launches by site
  - ✓Payload range
- Combination allowed to quickly analyze the relation between payload and launch sites, helping to identify where is the best place to launch according to payloads.

- Source Code: GitHub Mitsu Kansagara Interactive Dashboard with Plotly Dash

Predictive Analysis (Classification)

- Loaded the data using pandas and numpy, transformed the data, and finally split the data into training and testing dataset
- Four classification models were built and compared: logistic regression, support vector machine, decision tree and k nearest neighbors
- Tuned different hyperparameters using GridSearchCV
- Used accuracy as the metric for my model, improved the model using feature engineering and algorithm tuning



- Source Code: [GitHub Mitsu Kansagara Machine Learning Prediction]

Results

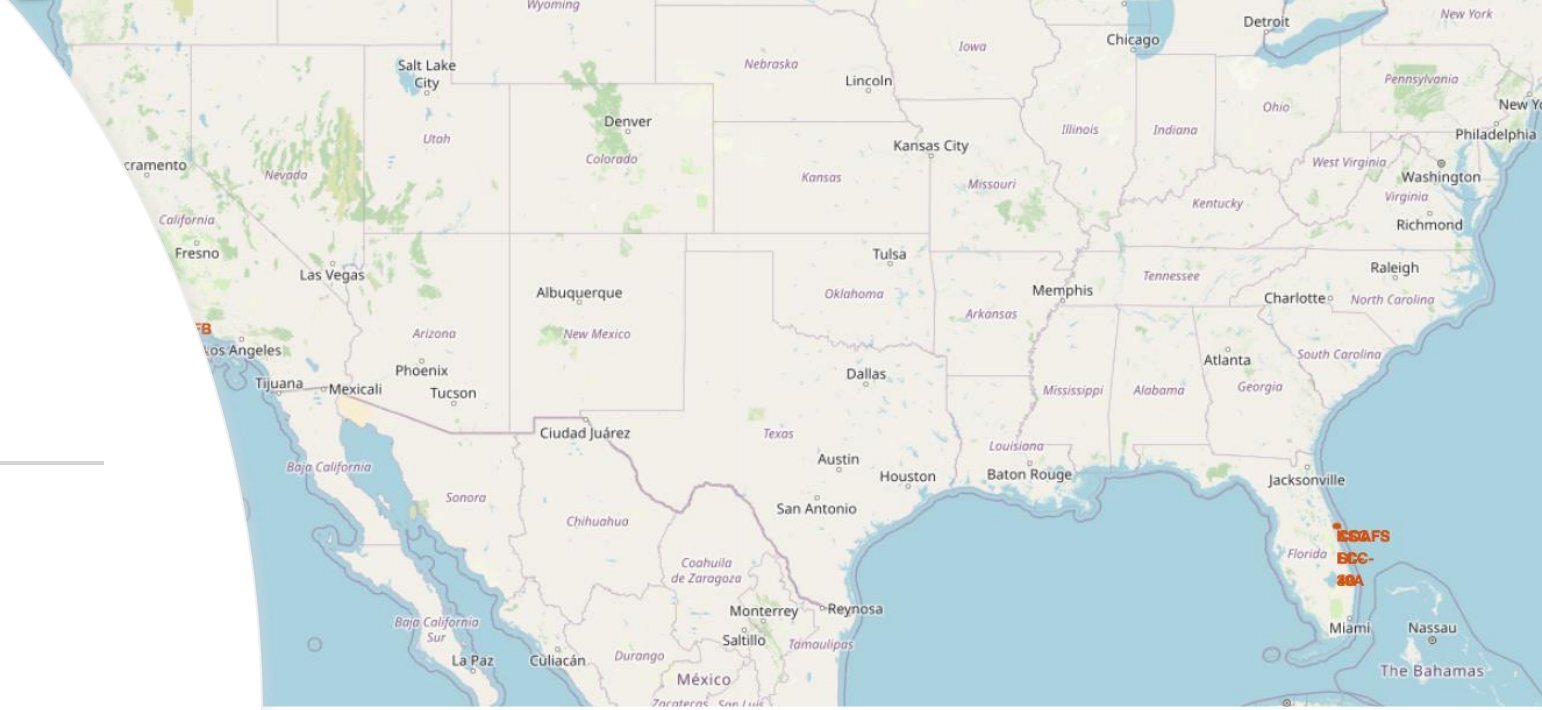
Exploratory data analysis results:

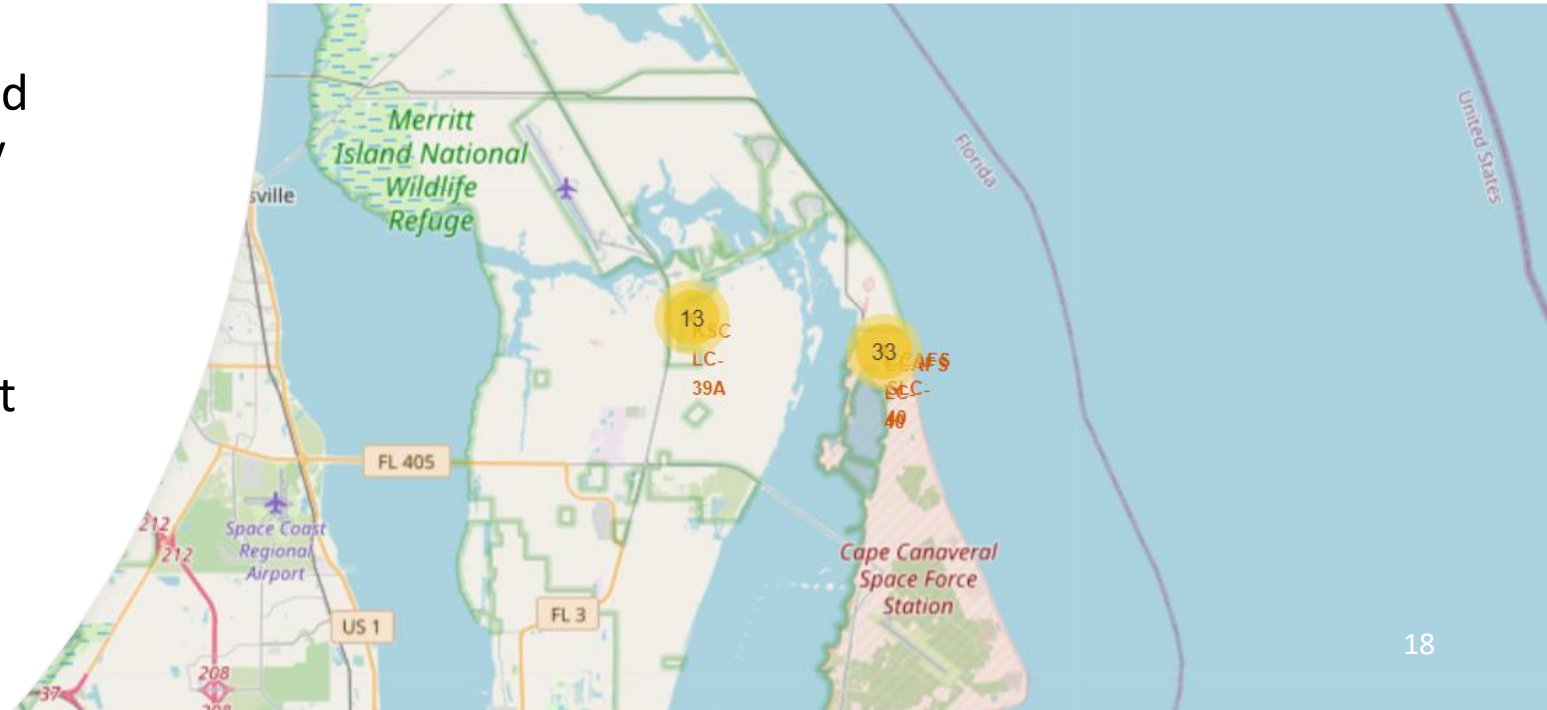
- ✓Space X uses 4 different launch sites
- ✓The first launches were done to Space X itself and NASA
- ✓The average payload of F9 v1.1 booster is 2,928 kg
- ✓The first success landing outcome happened in 2015 fiver year after the first launch
- ✓Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average
- ✓Almost 100% of mission outcomes were successful
- ✓Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015
- ✓The number of landing outcomes became as better as years passed

Results

Interactive analytics demo in screenshots:

- ✓ Using interactive analytics identified that launch sites use to be in safety places, near sea, for instance and have a good logistic infrastructure around.
- ✓ Most launches happens at east cost launch sites

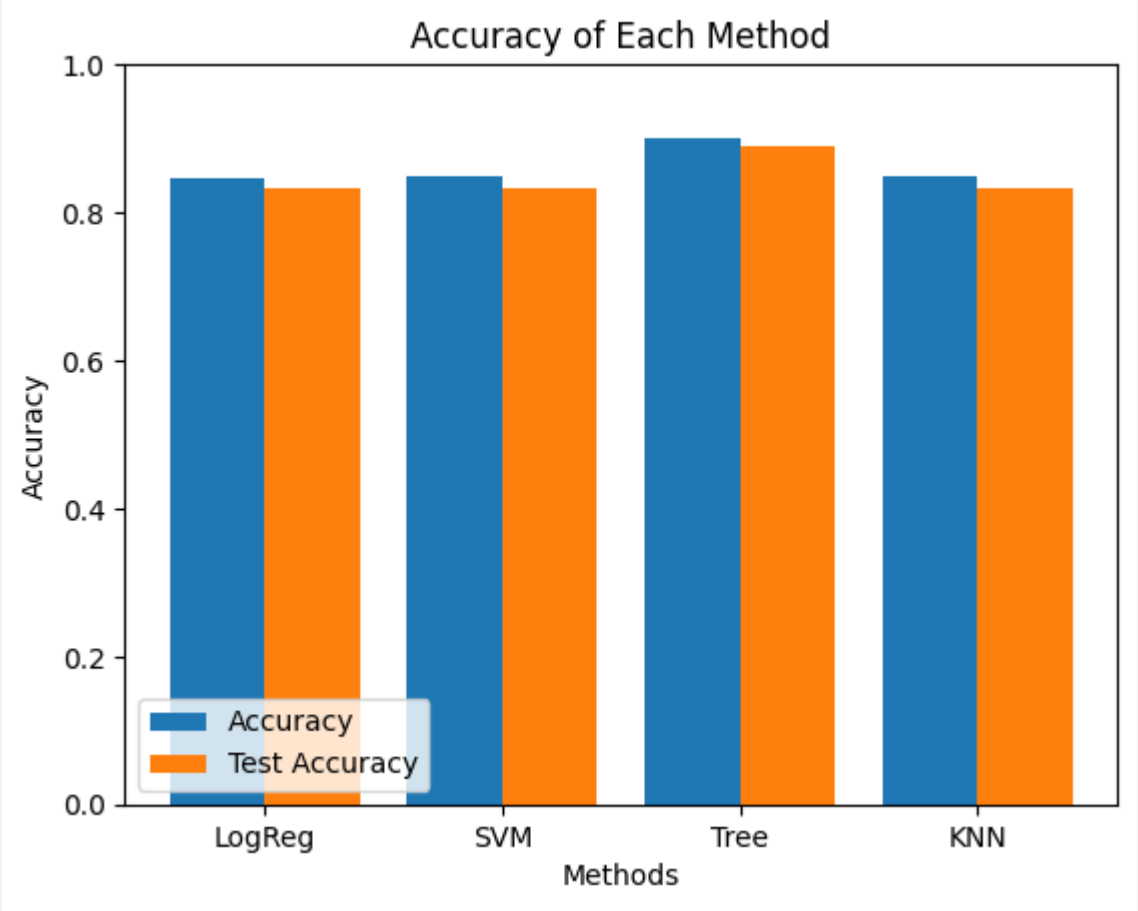


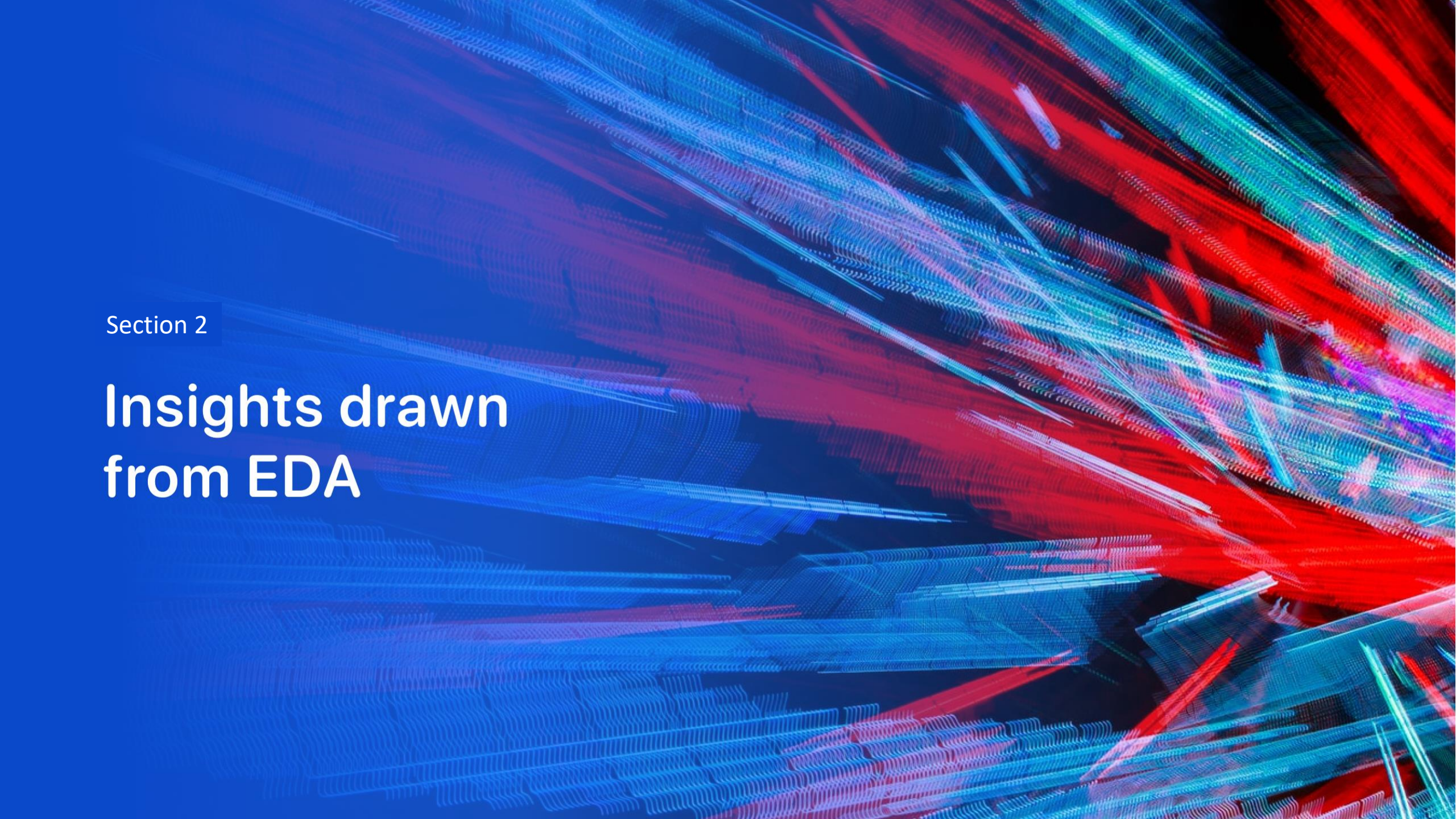


Results

Predictive analysis results

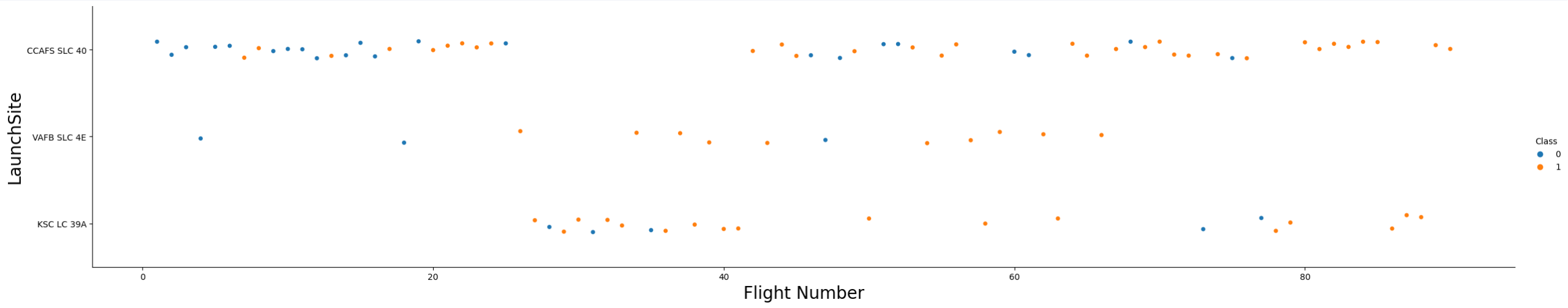
- ✓ Comparison between different ML models showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy of **90% on training dataset** and **89% for test dataset**.




Section 2
Insights drawn from EDA

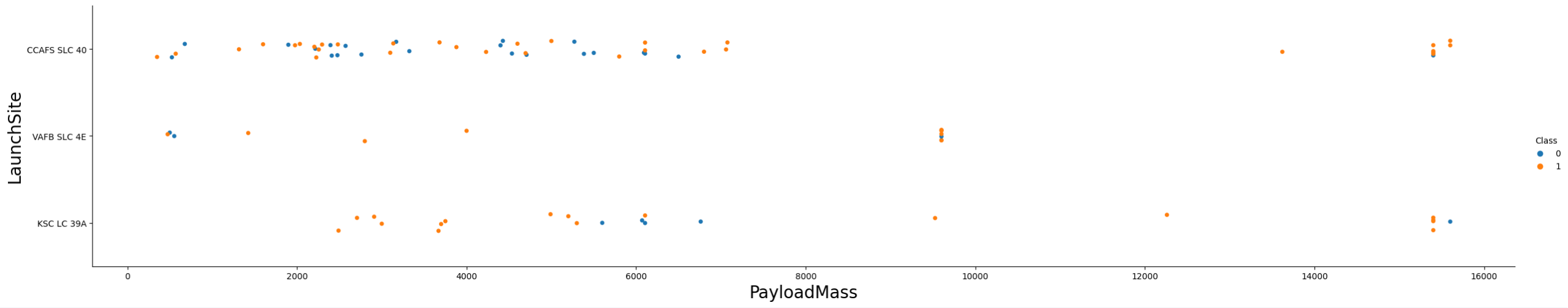
Flight Number vs. Launch Site

- The general success rate improved over time
- Most of recent launches were successful at CCAF5 SLC 40, and hence is the best launch site nowadays
- Following CCAF5 SLC 40 is VAFB SLC 4E and KSC LC 39A



Payload vs. Launch Site

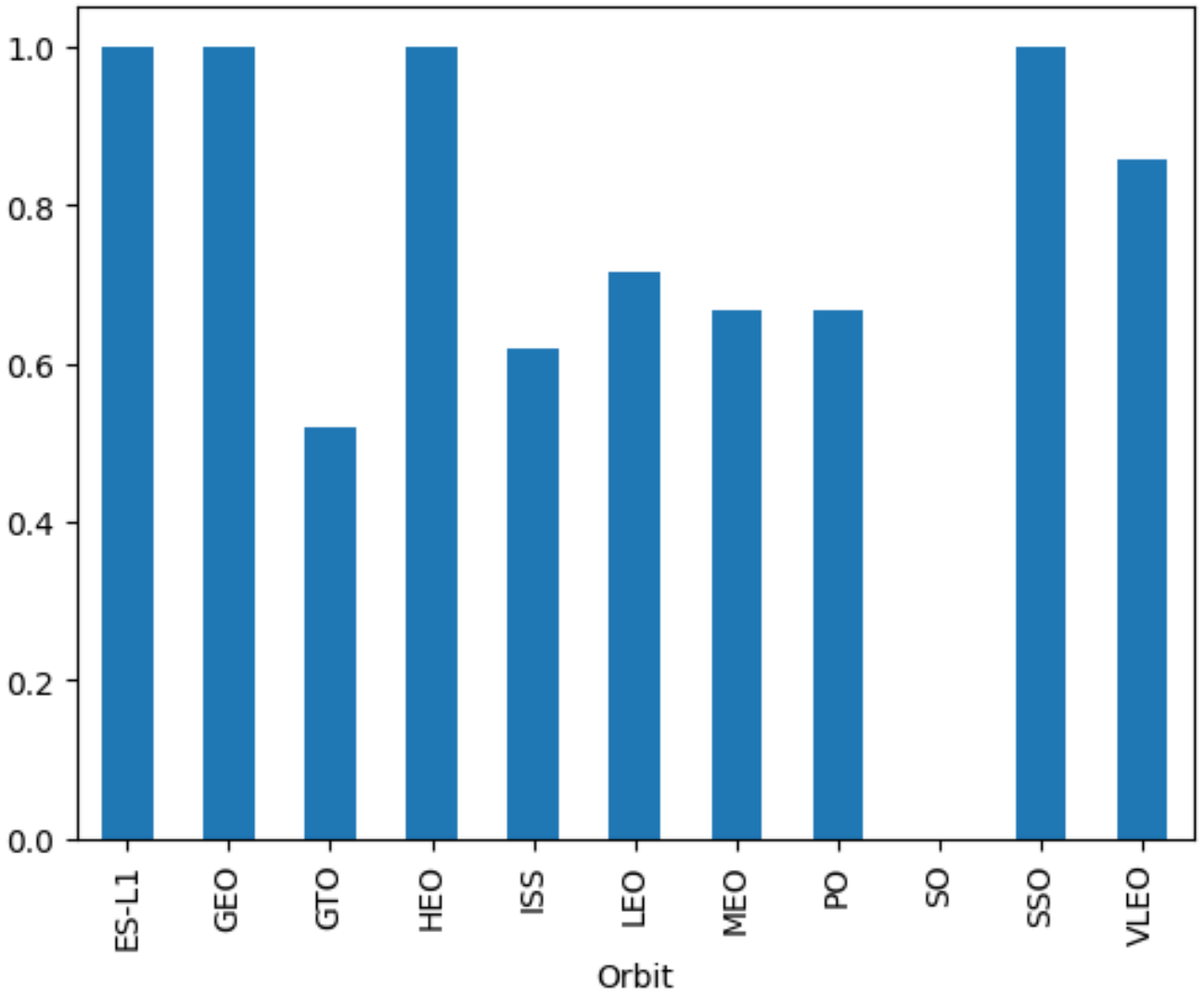
- Payloads over 7000kg have excellent success rate
- Payloads over 12000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites



Success Rate vs. Orbit Type

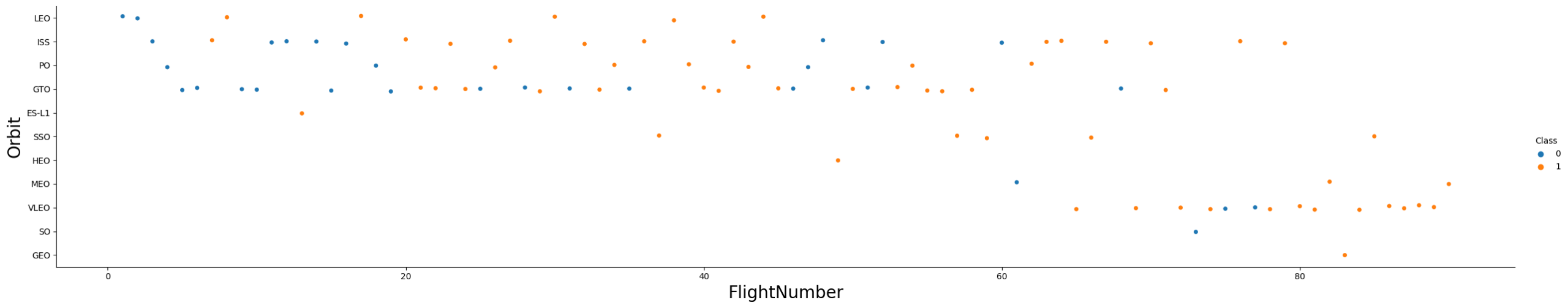
- The biggest success rates happens in the following orbits:
  - ✓ ES-L1
  - ✓ GEO
  - ✓ HEO
  - ✓ SSO

- Followed by VLEO (about 88%) and LFO (about 72%).



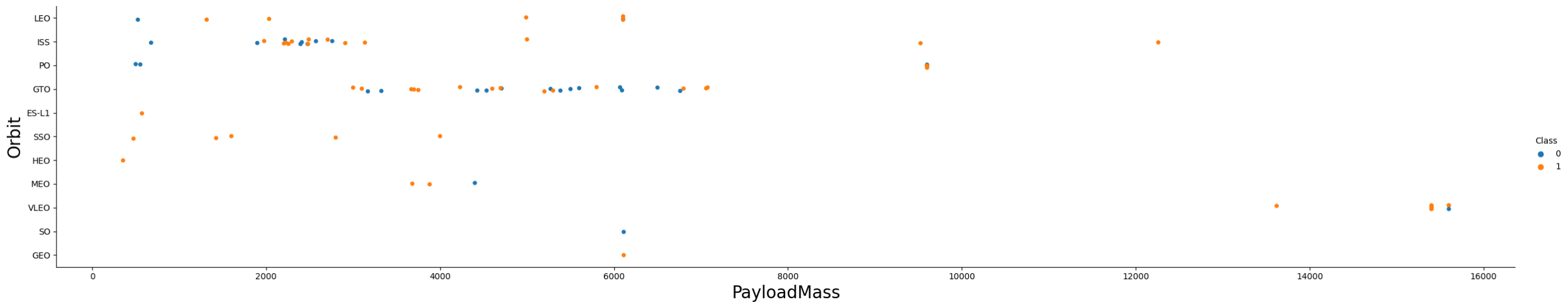
Flight Number vs. Orbit Type

- Success rate has improved over time to all orbits
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency



Payload vs. Orbit Type

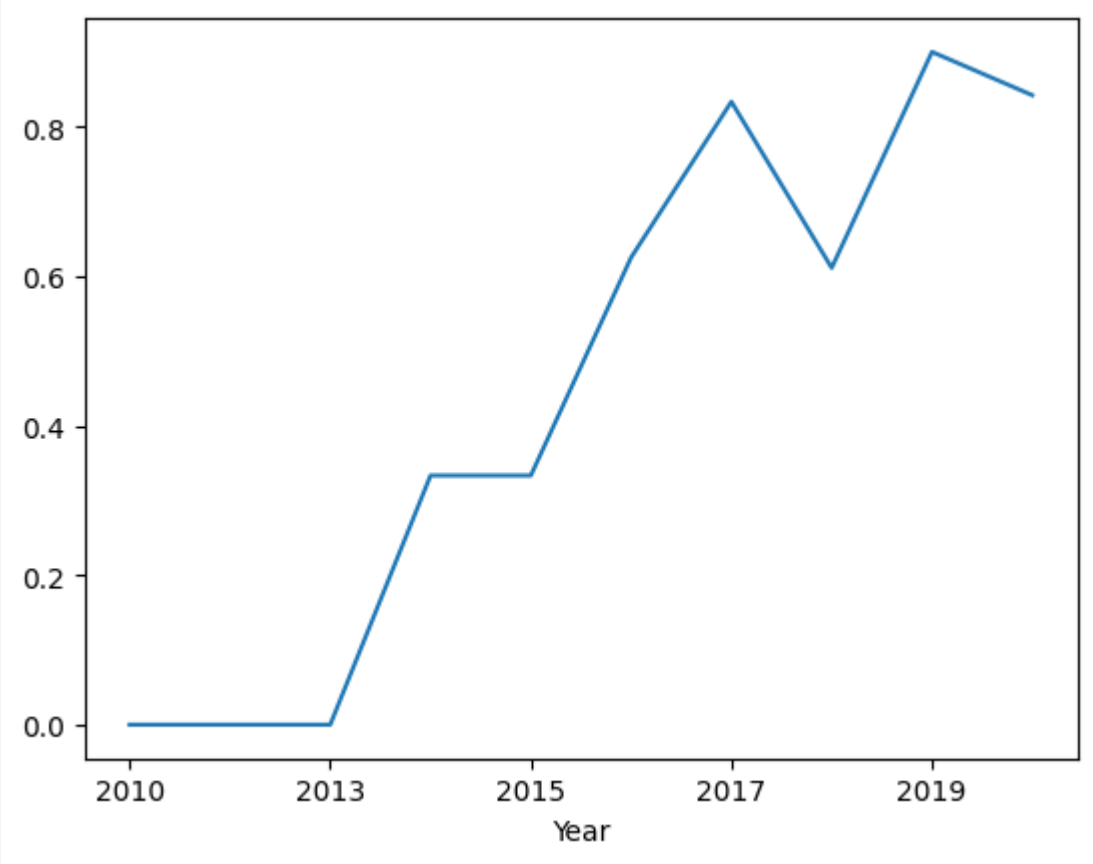
- VLEO handles larger payload mass, GTO and SSO handles lighter payload mass
- ISS orbit has the widest range of payload and a good rate of success
- There are few launches to the orbits SO and GEO



Launch Success Yearly Trend

There was a period of adjusts and improvement of technology in the first three years

Success rate started increasing significantly since 2013



All Launch Site Names

- According to our data, there are total of 4 launch sites
- These are obtained by selecting unique occurrences of “launch_site” values from the dataset

Launch_Site
None
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

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
111268.0

Total payload carried by boosters from NASA

Total payload is calculated by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA

Average Payload Mass by F9 v1.1

- Filtered data by the booster version Fp v1.1 and calculated the average payload mass

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

FIRST_SUCCESS_GP
01/08/2018

Found the dates of the first successful landing outcome on ground pad

Filtered data by successful landing outcome on ground pad and getting the minimum value for date it’s possible to identify the first occurrence

Successful Drone Ship Landing with Payload between 4000 and 6000

- Distinct boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- Selected distinct booster versions according to the filters above

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

Total Number of Successful and Failure Mission Outcomes

- Grouped mission outcomes and counting records for each group led me to this summary

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

Boosters Carried Maximum Payload

- These are the boosters have carried the maximum payload mass registered in the dataset

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

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

month	Date	Booster_Version	Launch_Site	Landing_Outcome
10	01/10/2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14/04/2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Landing_Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

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

- This view of data alerts us that “No attempt” must be taken in account

Section 3

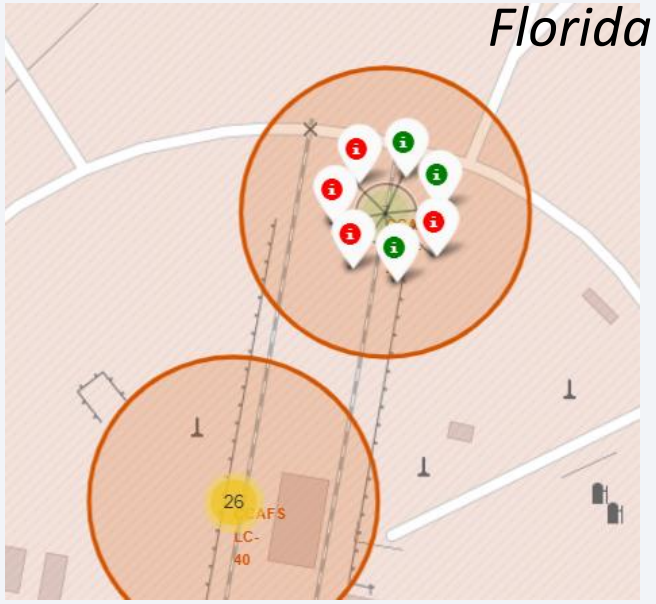
Launch Sites Proximities Analysis

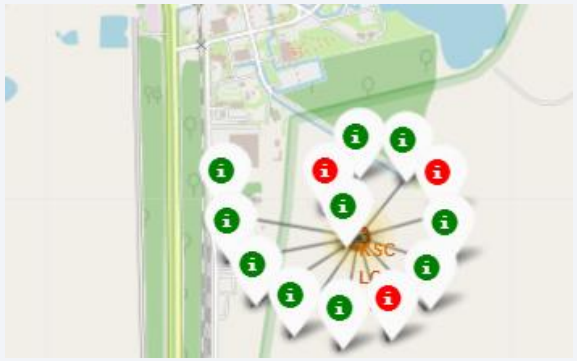
All Launch Sites across US

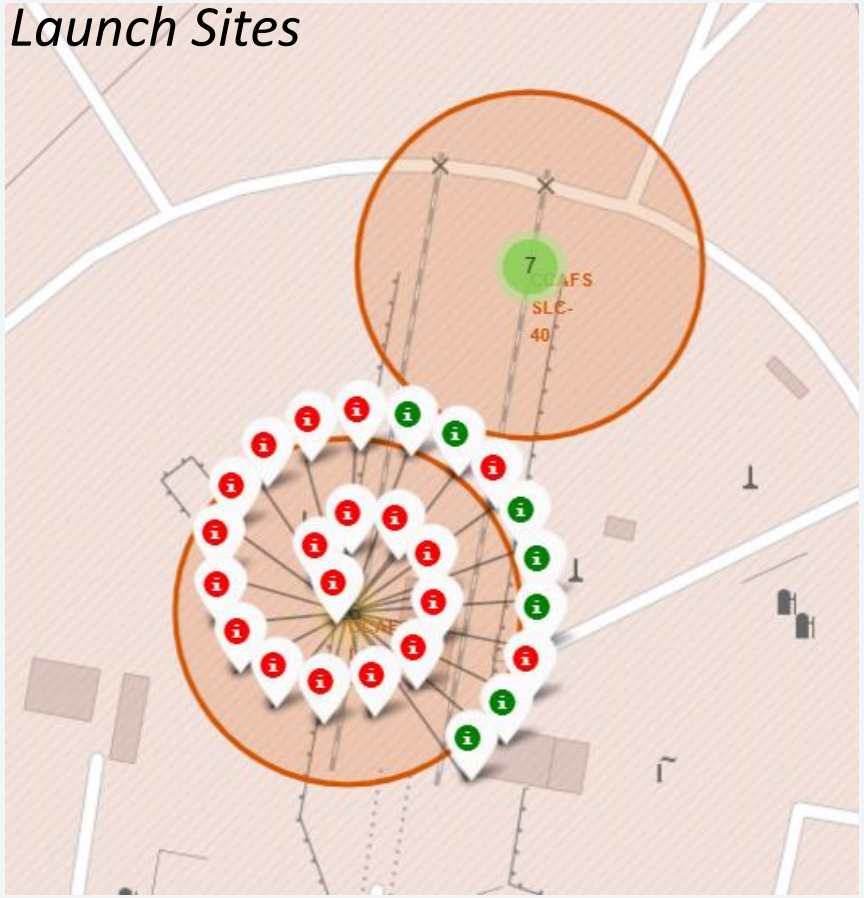
In USA, Launch sites are near sea, probably by safety, but not too far from roads and railroads

Markers showing launch sites with color labels

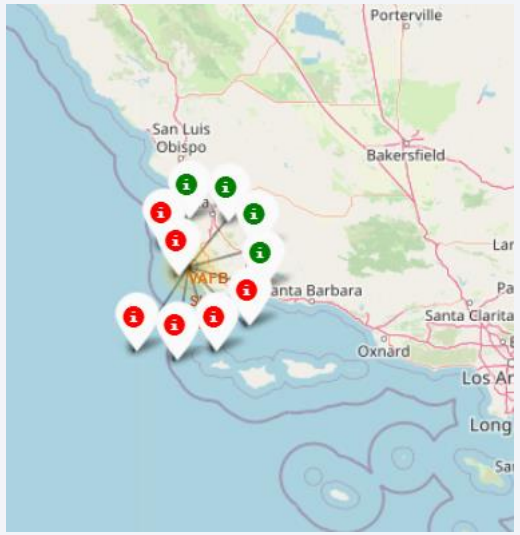
Florida Launch Sites





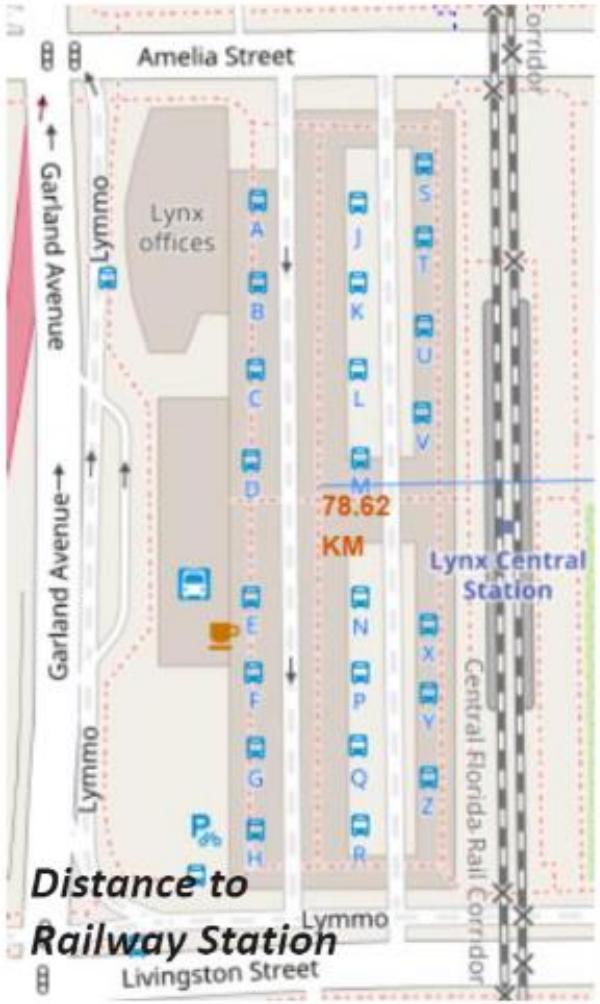


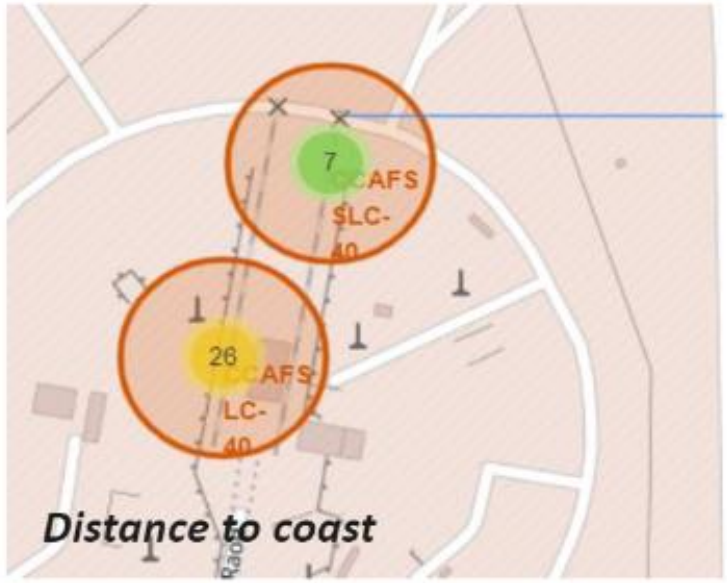
California Launch Sites

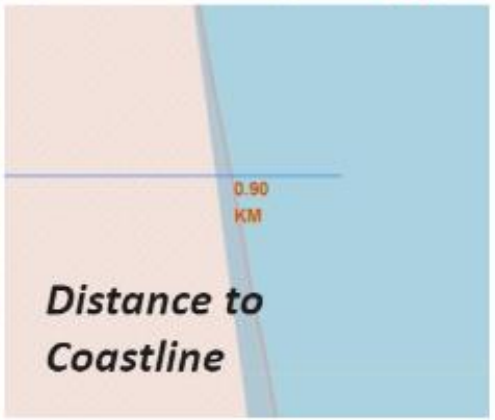


Green Marker shows successful launches and Red Marker shows failure

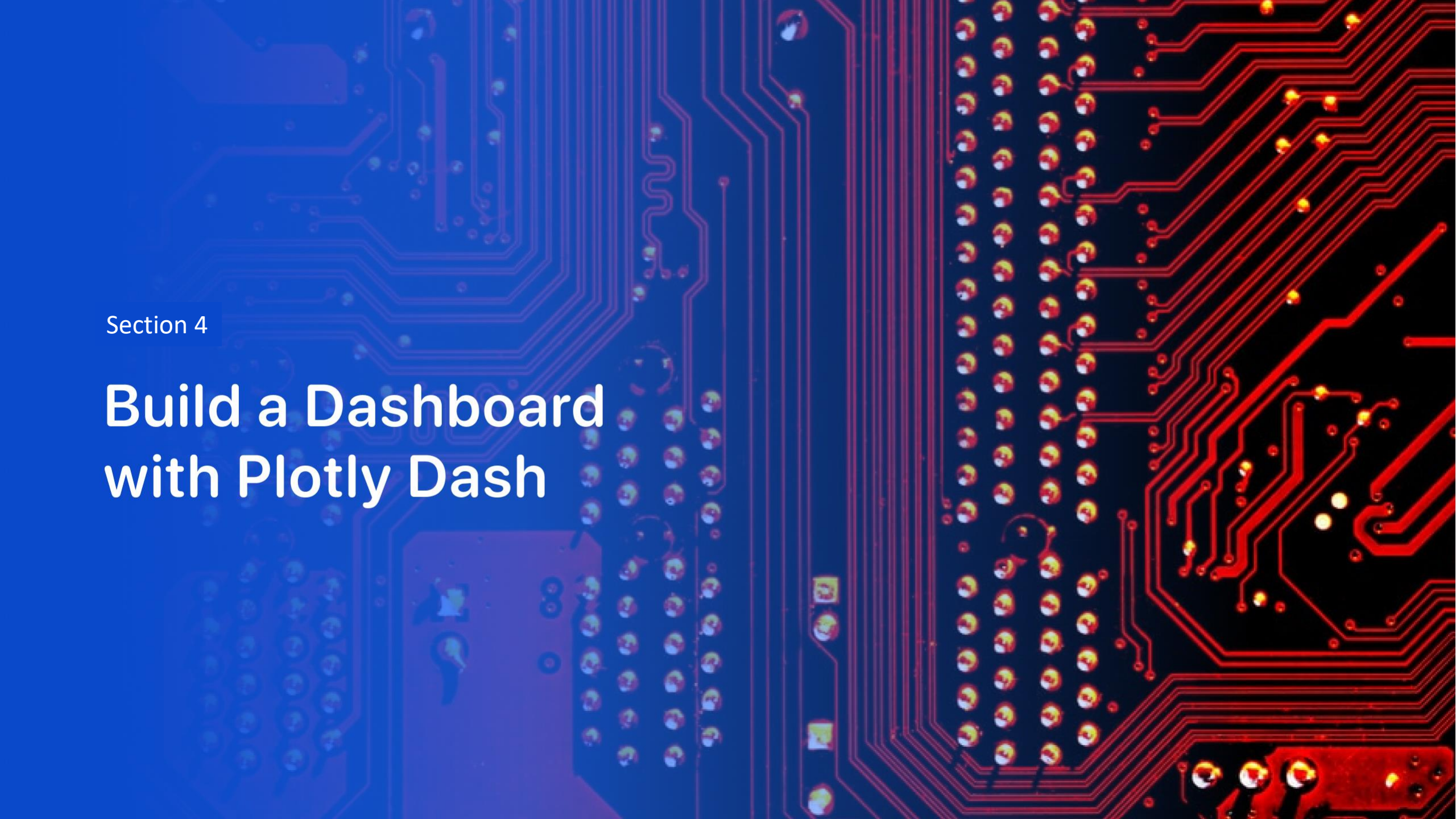
Launch Site distance to Landmarks





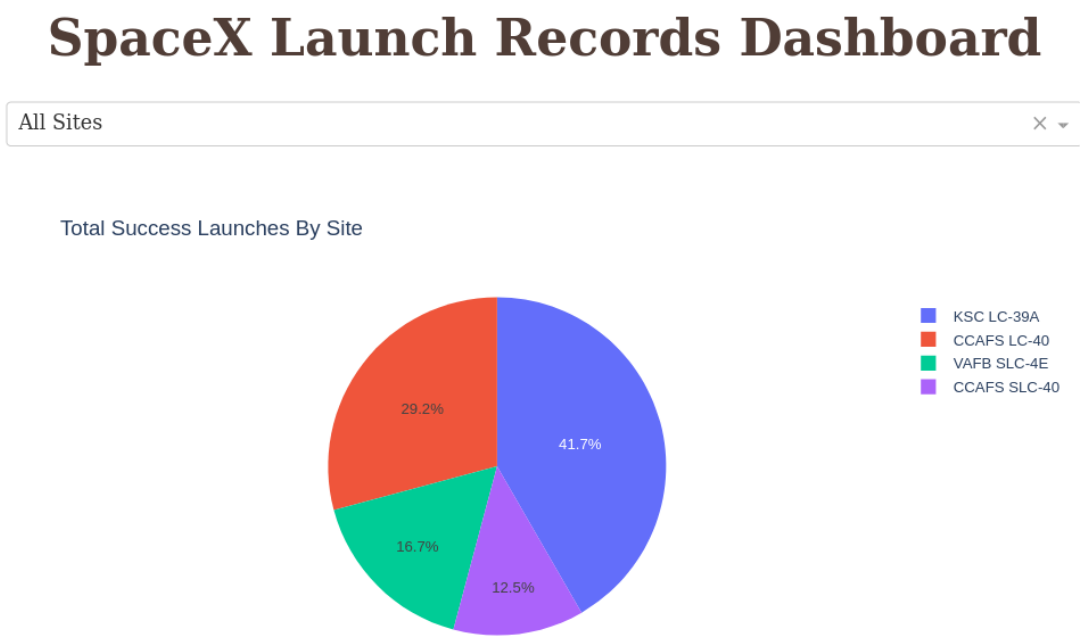


- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes


Section 4
Build a Dashboard with Plotly Dash

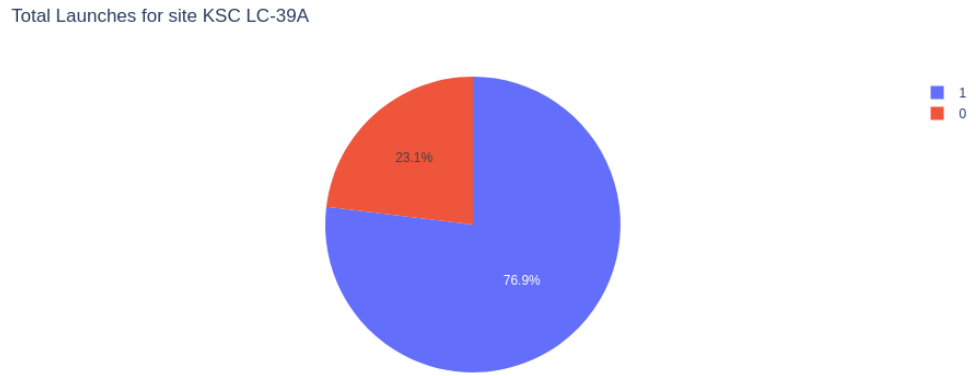
Successful Launches by Site

- The place from where launches are done seems to be a very important factor of success of missions

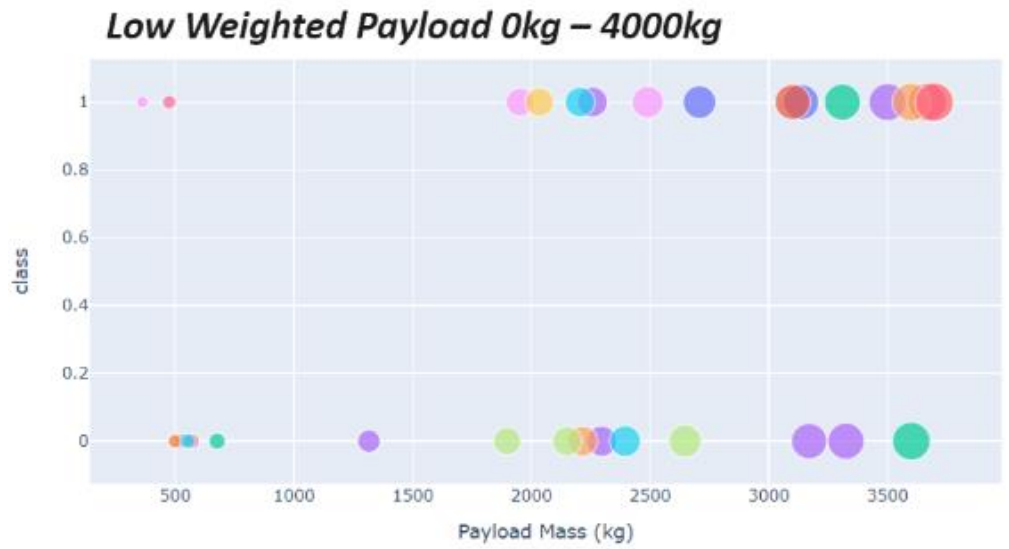


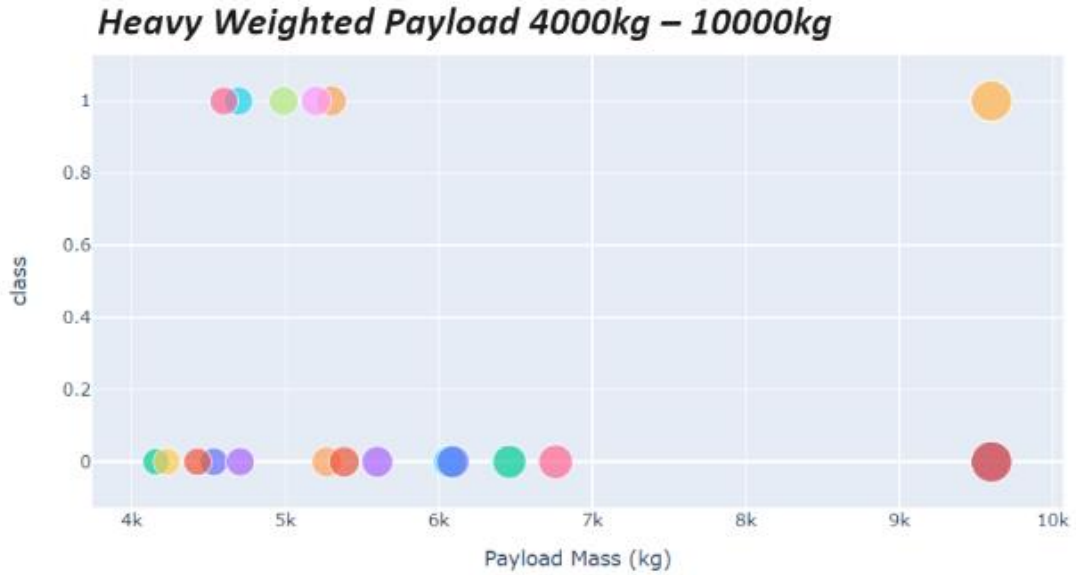
Pie chart showing the Launch site with the highest launch success ratio

- About 77% launches are successful in KSC LC-39A



Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider

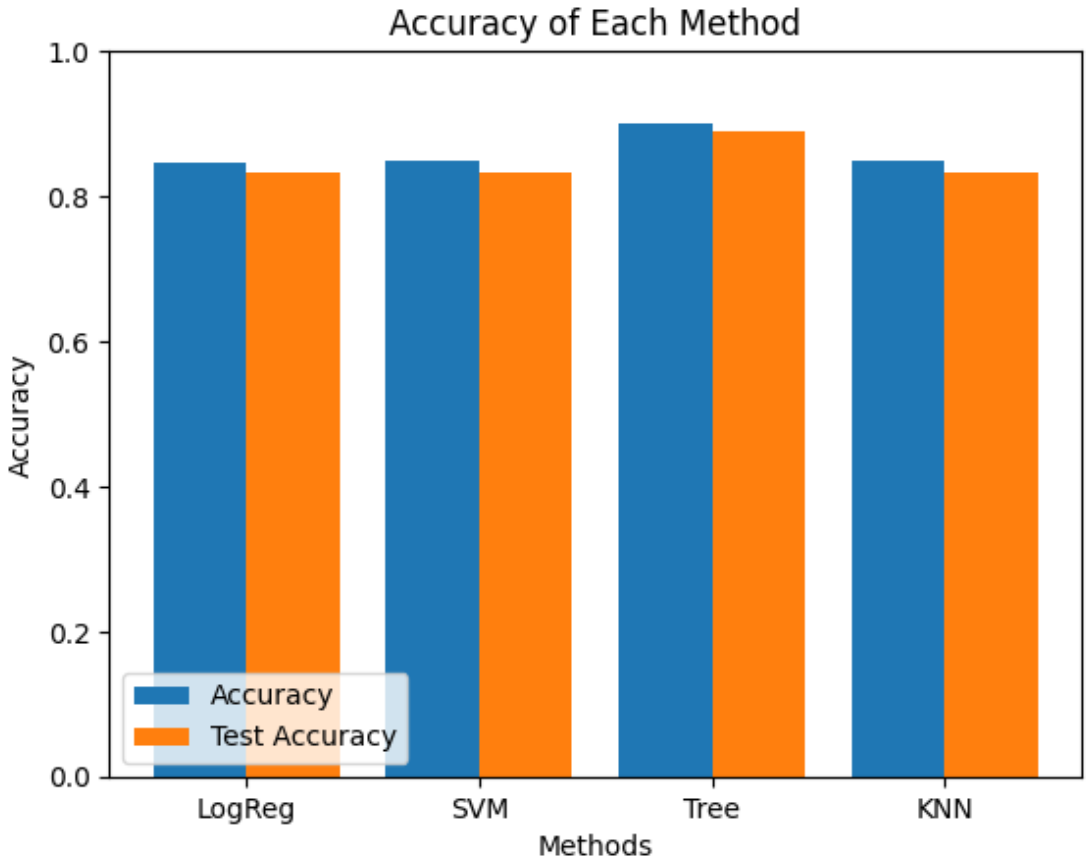




We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

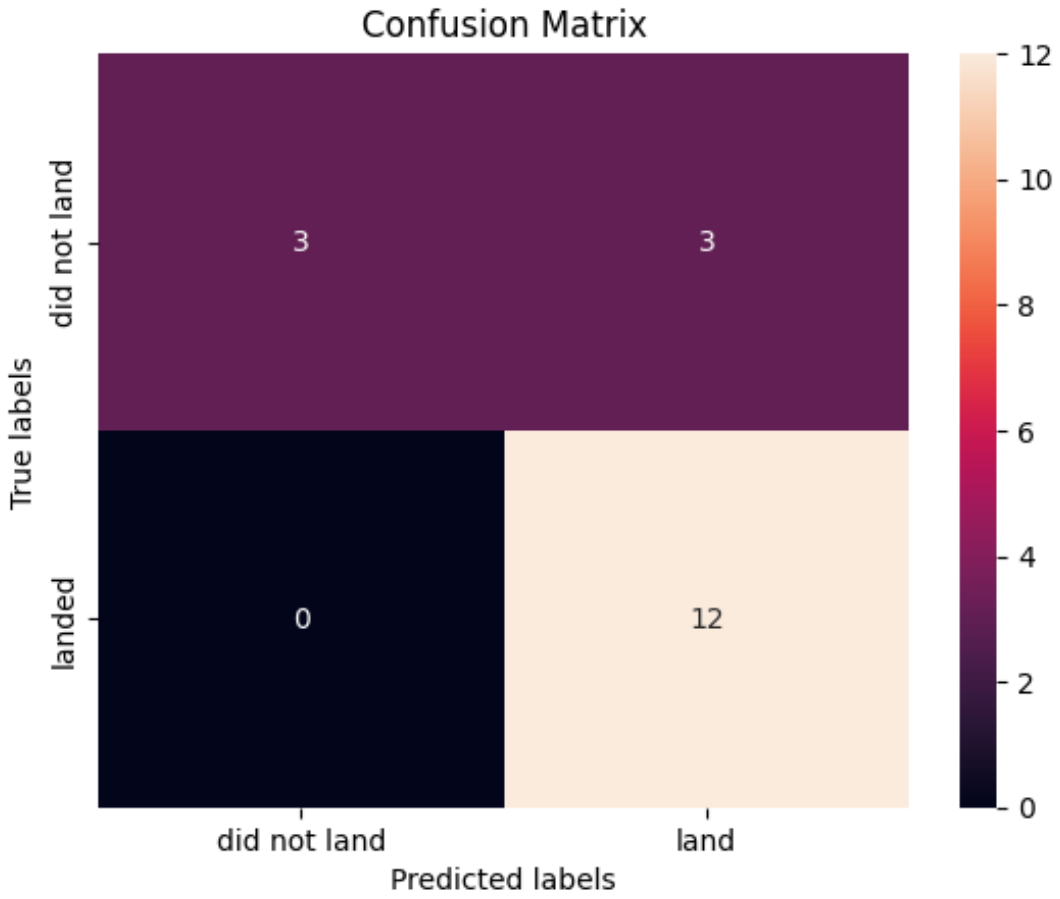
Section 5

Predictive Analysis (Classification)



Classification Accuracy

- Four classification models were trained and tested. Train and Test accuracies are plotted beside
- The model with the highest classification accuracy is Decision Tree Classifier, which has test dataset accuracy of about 89%.



Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes
- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones
- The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier

Conclusions

- Different data sources were analyzed, refining conclusions along the process
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets
- Orbits ES-L1, GEO, HEO, SSO had the most success rate
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency
- VLEO handles larger payload mass, GTO and SSO handles lighter payload mass
- Payloads over 7000kg are less risky and have excellent success rate
- Payloads over 12000kg are possible only on CCAFS SLC 40 and KSC LC 39A launch sites
- The best launch site nowadays is CCAF5 SLC 40
- Decision Tree Classifier is the best Machine Learning Algorithm to predict successful landings and increase profits

Thank you!