IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lakshmi Preetha Kumaraguru
17 Feb 24

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appe ndix

Executive Summary

- Summary of methodologies

  -Data Collection -

  Data Wrangling

  -□□ EDA with SQL

  - EDA with Data Visualization

  -Interactive Map with Folium

  -Interactive Dashboard using Plotly Dash -

  Machine Learning

- Summary of all results

  - Exploratory Data Analysis result
  - Interactive analytics
  - Predictive Analytics

Introduction

- Project background and context

  SpaceX advertises Falcon 9 rocket launches on itswebsite with a cost of 62 million dollars;other providers cost upward of 165 million dollars each, much of the savings is becauseSpaceX can reuse the first stage.

- Problems you want to find answers

  This project aims to predict if the 1st stage of the spaceX Falcon 9 rocket will land successfully

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Web Scrapping from SpaceX Wikipedia
- Perform datawrangling
  - One hot encoding was applied to categorical feature
- Perform exploratory data analysis (EDA) using visualization andSQL
- Perform interactive visual analytics using Folium andPlotlyDash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- How datasets were collected
  - Data is collected by using get request to the SpaceX REST API
  - Thereturned response is in a JSON format, so we converted into a dataframe
  - During the cleaning data, we looked for missing data and replace using the mean
  - Applied web scraping on SpaceX Wikipedia, extracted the launch records and converted into dataframe

Data Collection – SpaceX API

1. Requesting rocket launch data from SpaceX API

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

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

2. Converting Response to .JSON file

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

3. Using helper functions to clean the data

```
# Call getBoosterVersion
getBoosterVersion(data)
```

```
# Call getLaunchSite
getLaunchSite(data)

# Call getPayloadData
getPayloadData(data)

# Call getCoreData
getCoreData(data)
```

4. Combining the columns into a dictionary, and create adataframe

```
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}
```

5. Filteringdataframeand exporting to CSV

```
data_falcon9 = launch_df[launch_df['BoosterVersion'] != 'Falcon 1']
```

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

Github Reference Link

Data Collection - Scraping

1. Getting Response from HTML

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

2. Creating BeautifulSoup Object

```
soup = BeautifulSoup(response.text, "html.parser")
```

3. Finding tables

```
html_tables = soup.find_all("table")
```

4. Getting column names

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

5. Creation of dictionary &Appending data to keys (refer to the link)
6. Converting dictionary todataframe

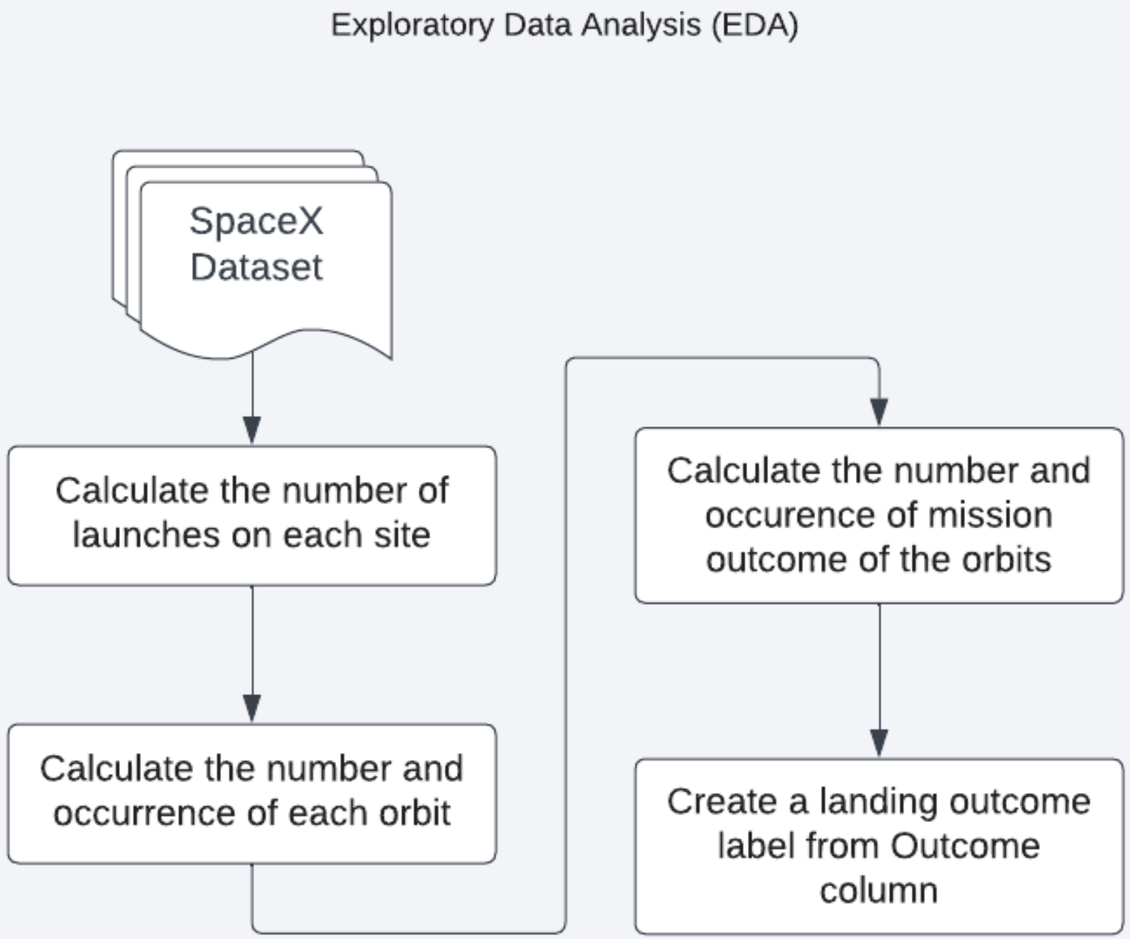
```
df= pd.DataFrame({ key:pd.Series(value) for key, value in launch_dict.items() })
```

5. ConvertingDataframeto .CSV

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

Github Reference Link

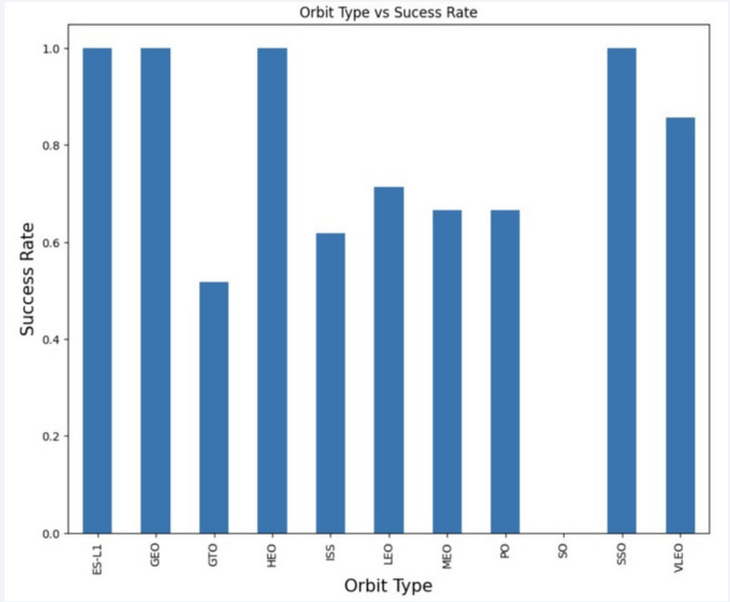
Data Wrangling

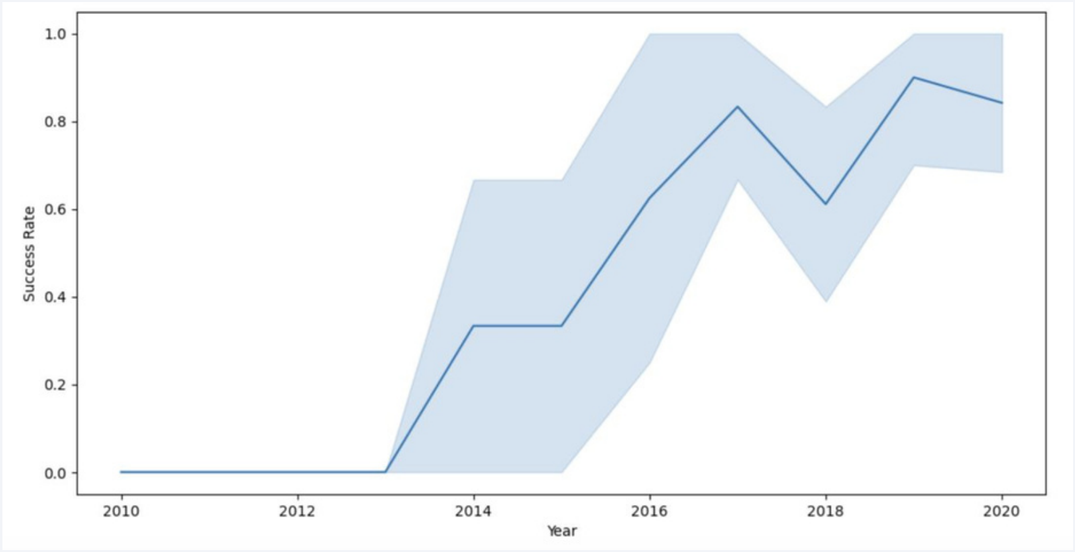


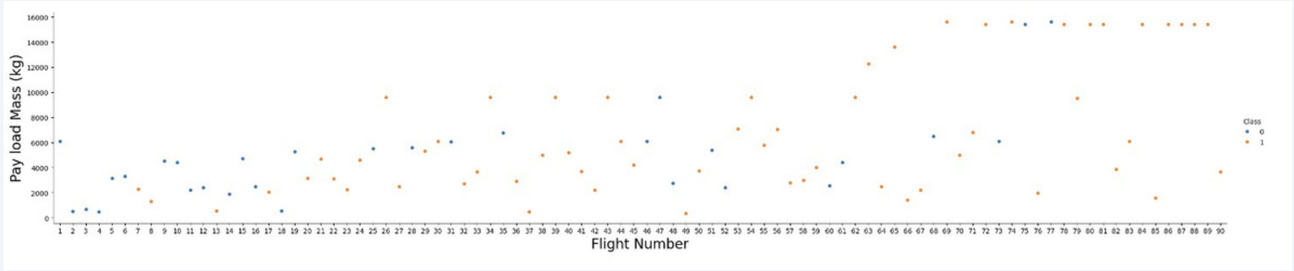
Github Reference Link

EDA with Data Visualization

- In data visualization we used the following charts:
  - Scatter plot: to observe the relationship between numeric variables
  - Bar chart: to check the data distribution
  - Line chart: to check the databehaviorover the time







Github Reference Link

EDA withSQL

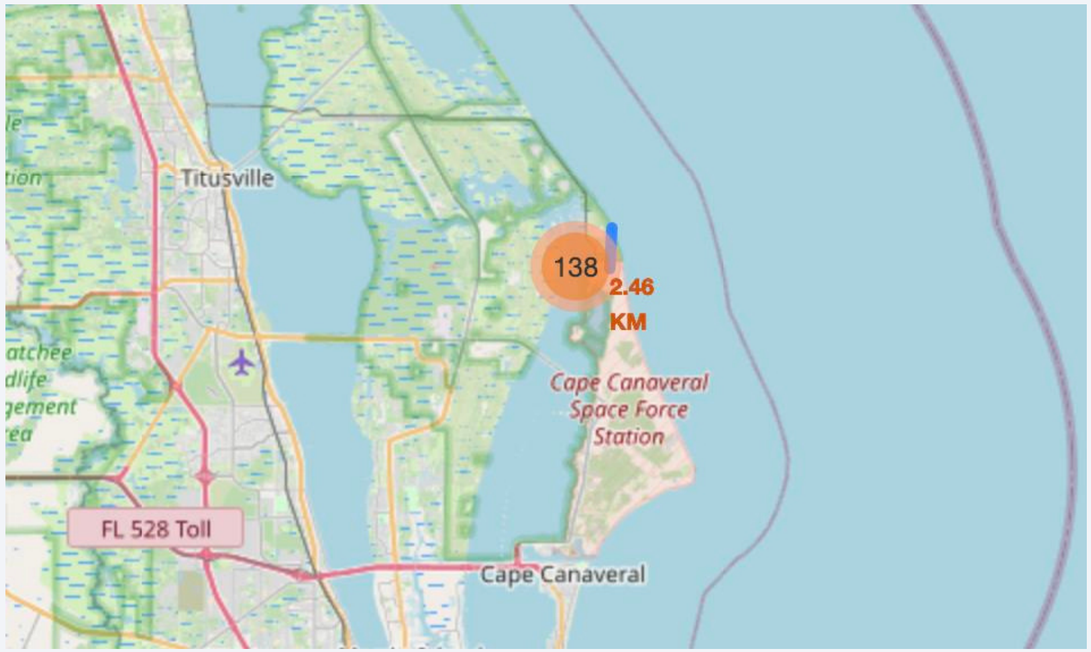
Performed SQL queries:

- 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 succesful landing outcome in ground pad was acheived
- 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 using subquery.
- 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
- 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

Github Reference Link

Build an Interactive Map with Folium

- Using Folium objects such as Markers, Circles and lines. We marked all Successful/Failure Launch sites to analyze launch sites locations and observe patterns during this interactive analytics, the below map is an example:



Github Reference Link

Build a Dashboard with Plotly Dash

- Using Plotly Dash,we created interactive dashboard so we will be able to obtain insights and answer some questions such as:
  1. Which site has the largest successful launches?
  2. Which site has the highest launch success rate?
  3. Which payload range(s) has the highest launch success rate?
  4. Which payload range(s) has the lowest launch success rate?
  5. Which F9 Booster version (v1.0,v1.1,FT,B4,B5, etc.) has the highest launch success rate?

Github Reference Link

Predictive Analysis (Classification)

- Linear Regression Model, Support Vector Machine (SVM) and K Nearest NeighborsAccuracy = 83.3%
- Decision Tree Accuracy = 86%

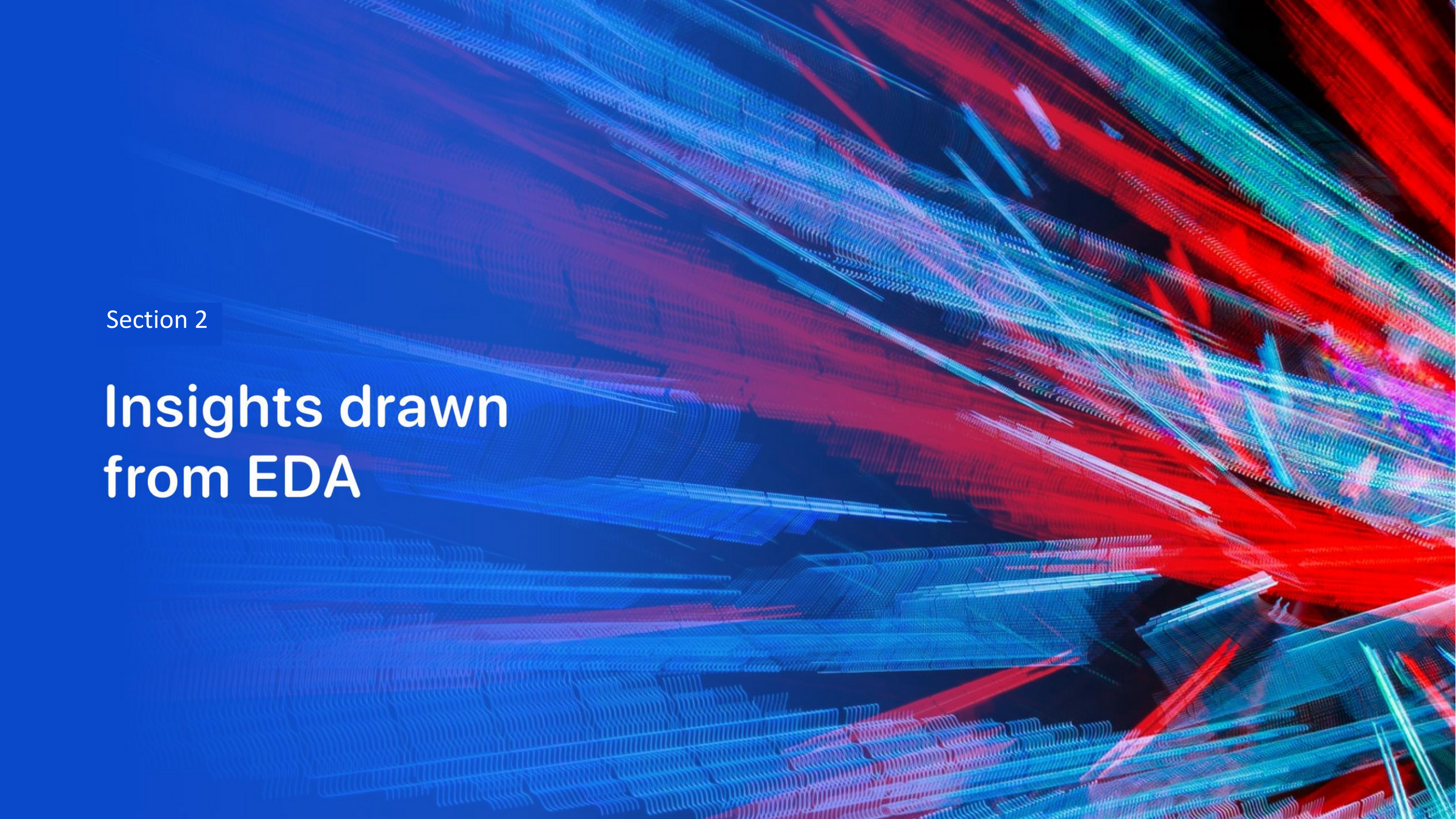
Github Reference Link

Results

- The success rates forSpaceX launches is directly proportional time in yearsthey will eventually perfect the launches
- KSC LC 39A had the most successful launches from all the sites
- Decision Tree Model Achieved the highest Accuracy

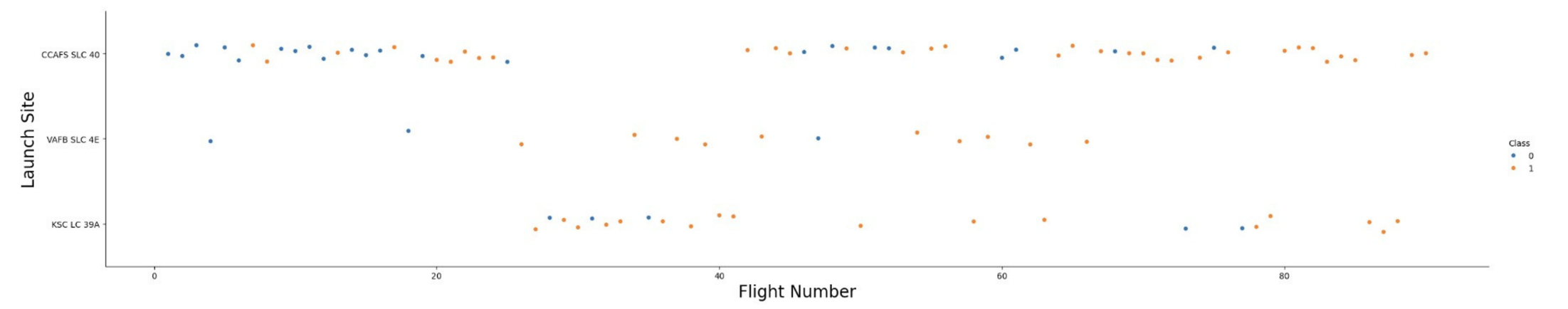
Section 2

Insights drawn from EDA



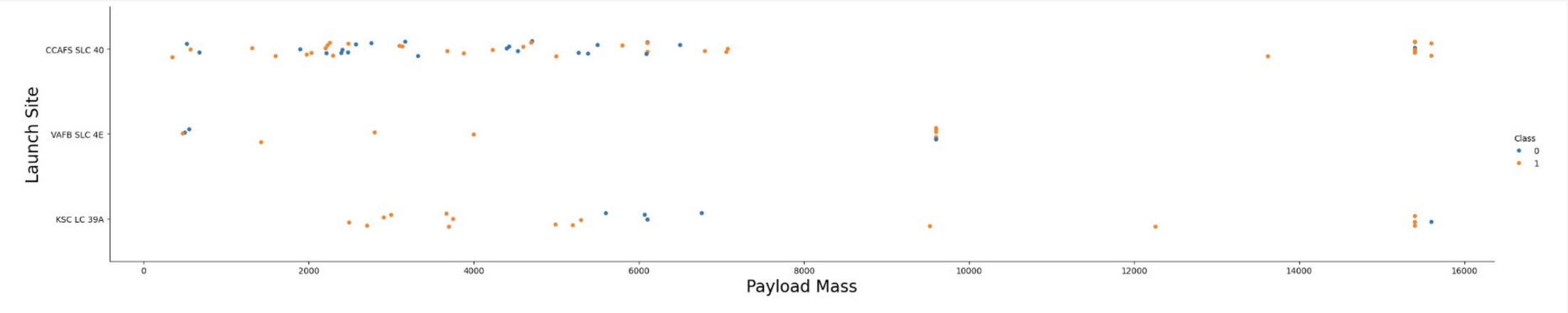
Flight Number vs. LaunchSite

- The below scatter plot shows that the site CCAFS SLC 40 has the greatest numberof launches



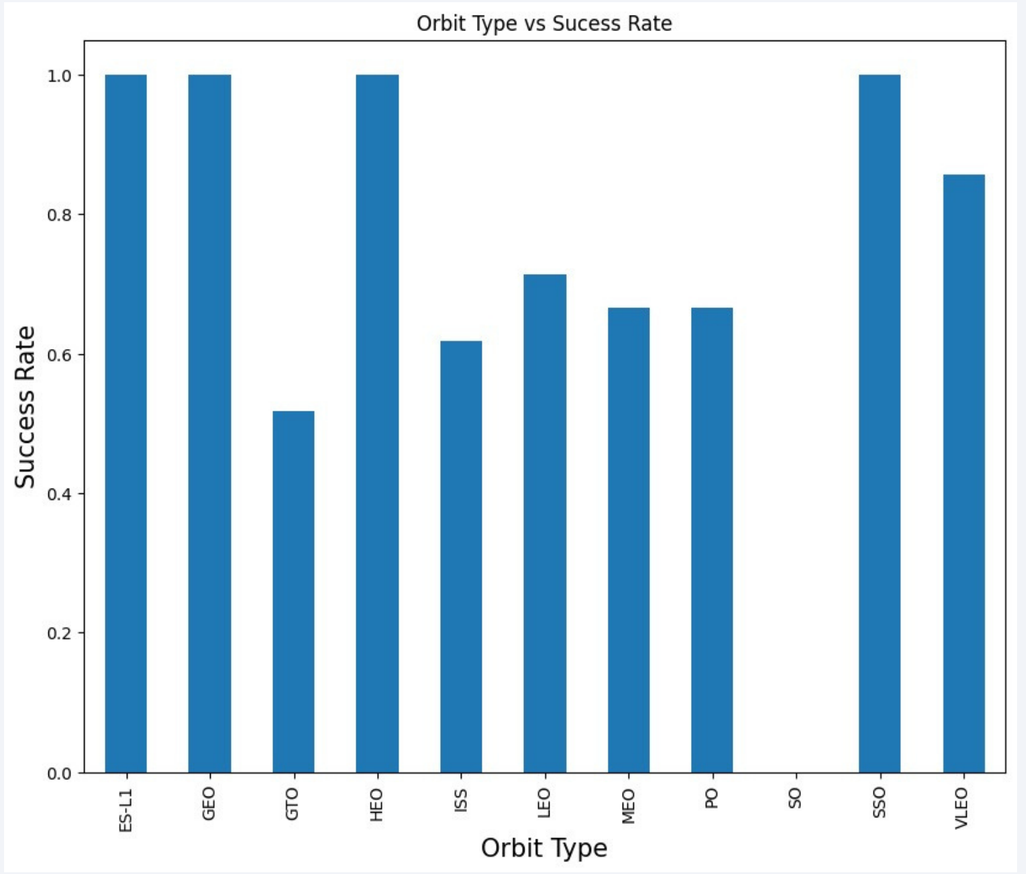
Payload vs. Launch Site

- The below scatter plot shows that the site CCAFS SLC 40has the greatest number of launches with low Payload Mass



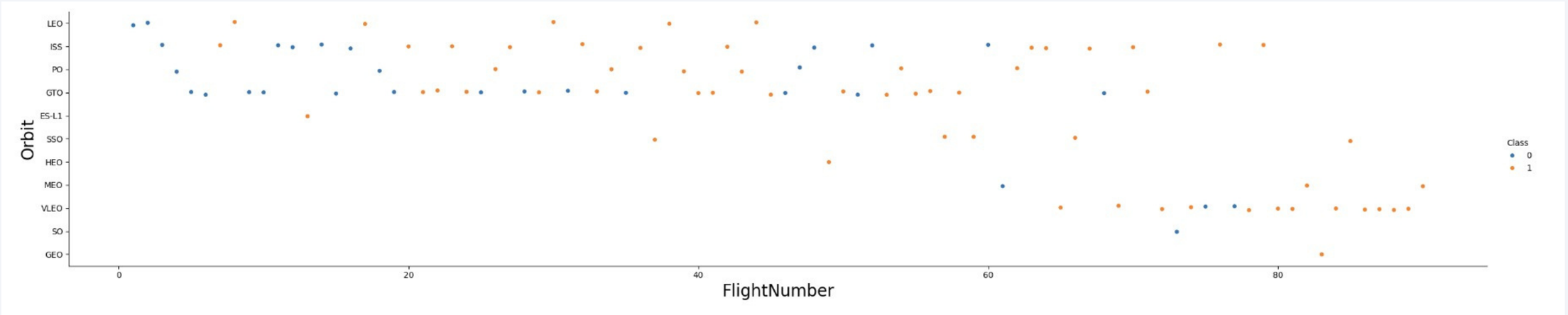
Success Rate vs. Orbit Type

- The following orbit types are the highest success rate:
  - ESL-1
  - GEO
  - HEO
  - SSO



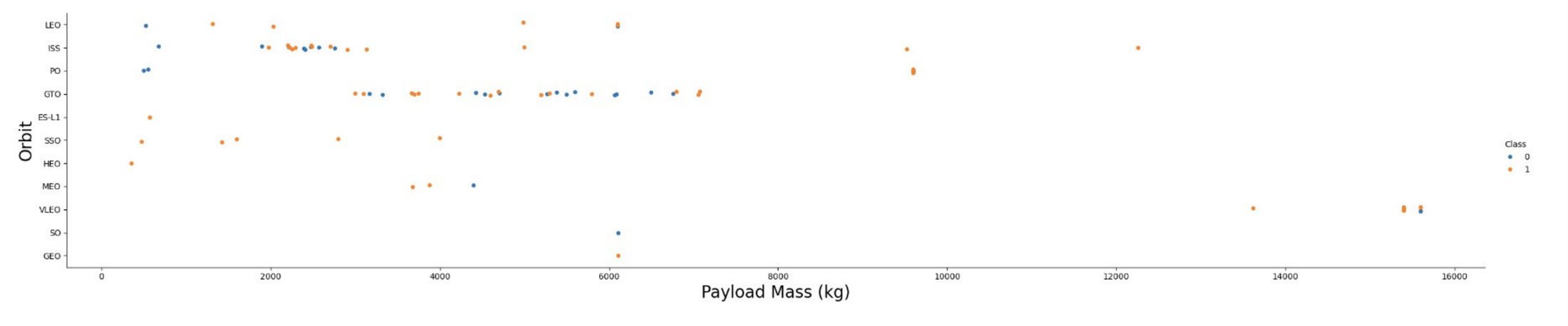
Flight Number vs. Orbit Type

- A trend can be observed of shifting to VLEO launches in recent years



Payload vs. OrbitType

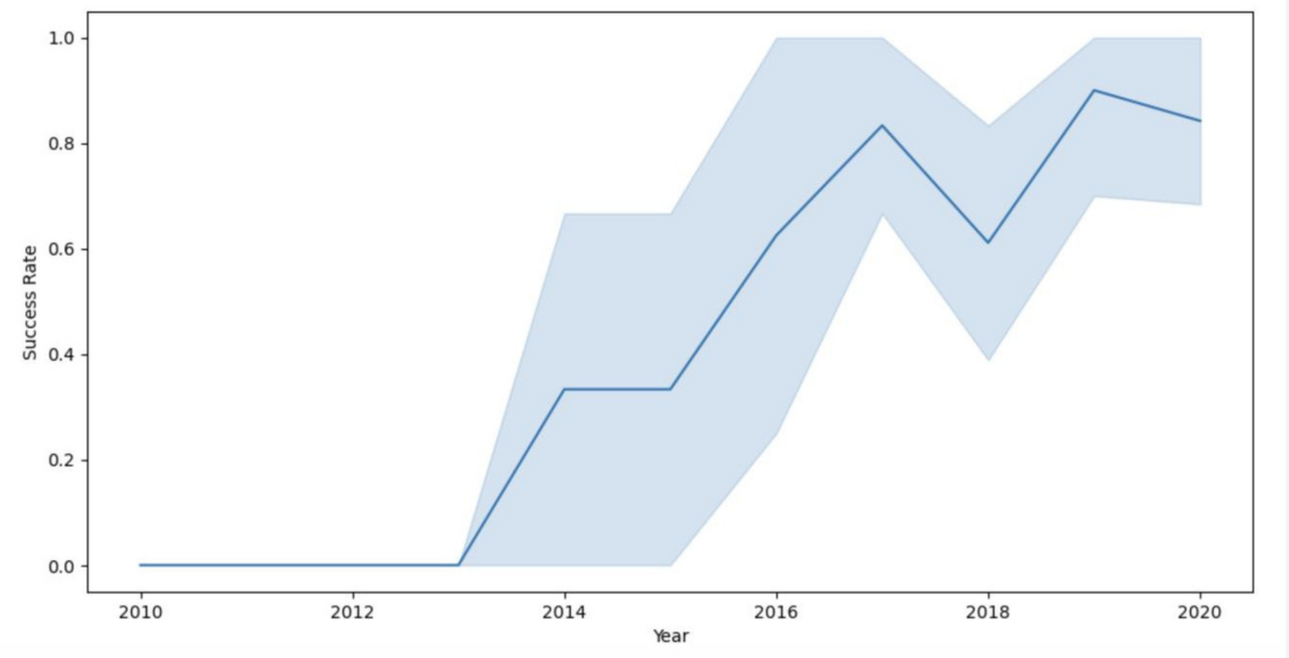
- There's a strong correlation in:
  - ISS & Payload Mass in 2000 kg
  - GTO & Payload Mass in 4000-8000 kg



Launch Success Yearly Trend

- Show a line chart of yearly average success rate

- Show the screenshot of the scatter plot with explanations



All Launch Site Names

- The below query is to display all Launch Sites names using DISTINCT feature

```
%sql select DISTINCT Launch_Site from SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

- For string search, we used 'CCA%' and % sign means any number of characters, We used LIMIT to show 5 rows in the result

```
%sql select Launch_Site from SPACEXTBL where Launch_Site like 'CCA%' limit 5
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- In the below query we used the filtering feature WHERE and aggregation function SUM

```
%sql select sum(PAYLOAD_MASS__KG_) as Payload_mass_total from SPACEXTBL where Customer == 'NASA (CRS)'
```

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

Payload_mass_total
45596

Average Payload Mass by F9 v1.1

- In the below query we used the filtering feature WHERE and aggregation function AVG

```
%sql select avg(PAYLOAD_MASS__KG_) as Payload_mass_average from SPACEXTBL where Booster_Version == 'F9 v1.1'
```

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

Payload_mass_average
2928.4

First Successful Ground Landing Date

- In the below query we used the filtering feature WHERE and aggregation function MIN

```
%sql select min(Date) as First_Successful_LandingOutcome from SPACEXTBL where Landing_Outcome like 'Success%'
```

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

First_Successful_LandingOutcome
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- In the below query we used the filtering feature WHERE and BETWEEN

```
%%sql select Booster_Version from SPACEXTBL
where Landing_Outcome == 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

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

Total Number of Successful and FailureMission Outcomes

- The below query is to show total number of successful and failure mission outcomes

```
%sql select Mission_Outcome, count(Mission_Outcome) as Total_Mission_Outcomes from SPACEXTBL group by Mission_Outcome
```

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

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

Boosters Carried MaximumPayload

- We used subquery to achieve the below result

```
%%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL
where PAYLOAD_MASS__KG_ in (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- In the below query we used the filtering feature WHERE and substr to find the correct date

```
%%sql select substr(Date, 6,2) as month,Booster_Version, Launch_Site from SPACEXTBL
where  Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015'
```

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

month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

- 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

```
%%sql select date,Landing_Outcome, count(Landing_Outcome) as count_ from SPACEXTBL
where date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by date desc
```

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

Date	Landing_Outcome	count_
2016-04-08	Success (drone ship)	5
2015-12-22	Success (ground pad)	3
2015-06-28	Precluded (drone ship)	1
2015-01-10	Failure (drone ship)	5
2014-04-18	Controlled (ocean)	3
2013-09-29	Uncontrolled (ocean)	2
2012-05-22	No attempt	10
2010-06-04	Failure (parachute)	2

Section 3

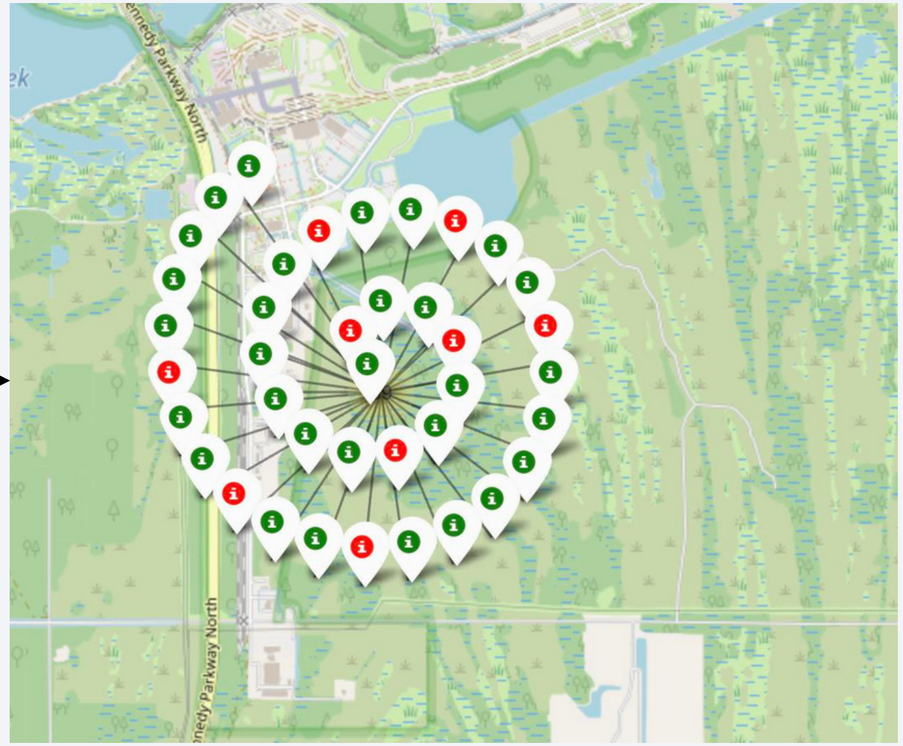
Launch Sites Proximities Analysis

All Launch Sites Location

•As shown in below map, all Launch Sites are in Florida & California States

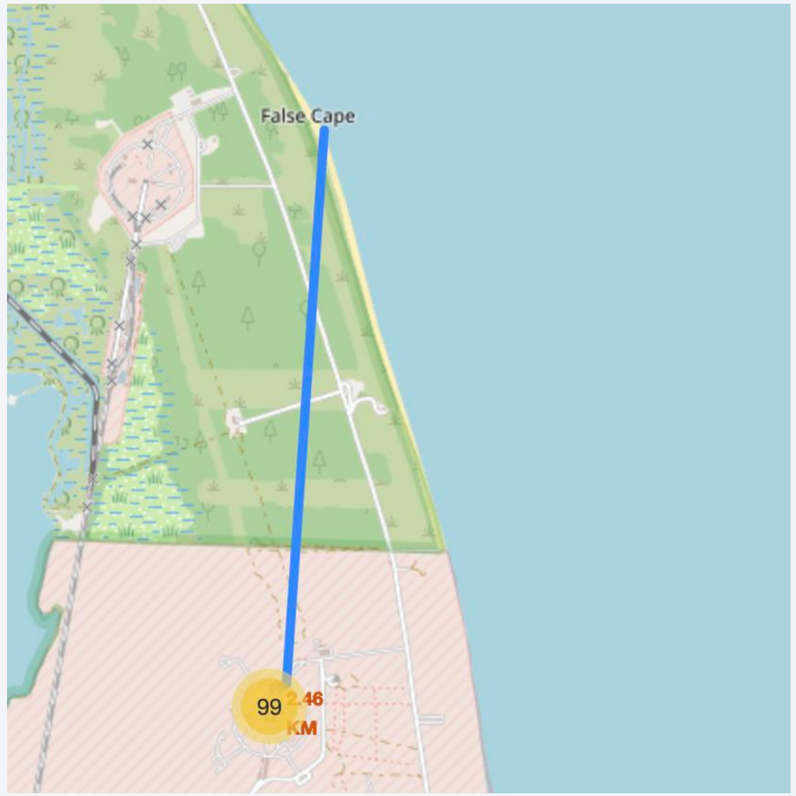
Launch Sites Outcomes With Folium

- If we clicked on the area, Marks will be shown
  - Red: Failure Launch
  - Green: Successful Launch



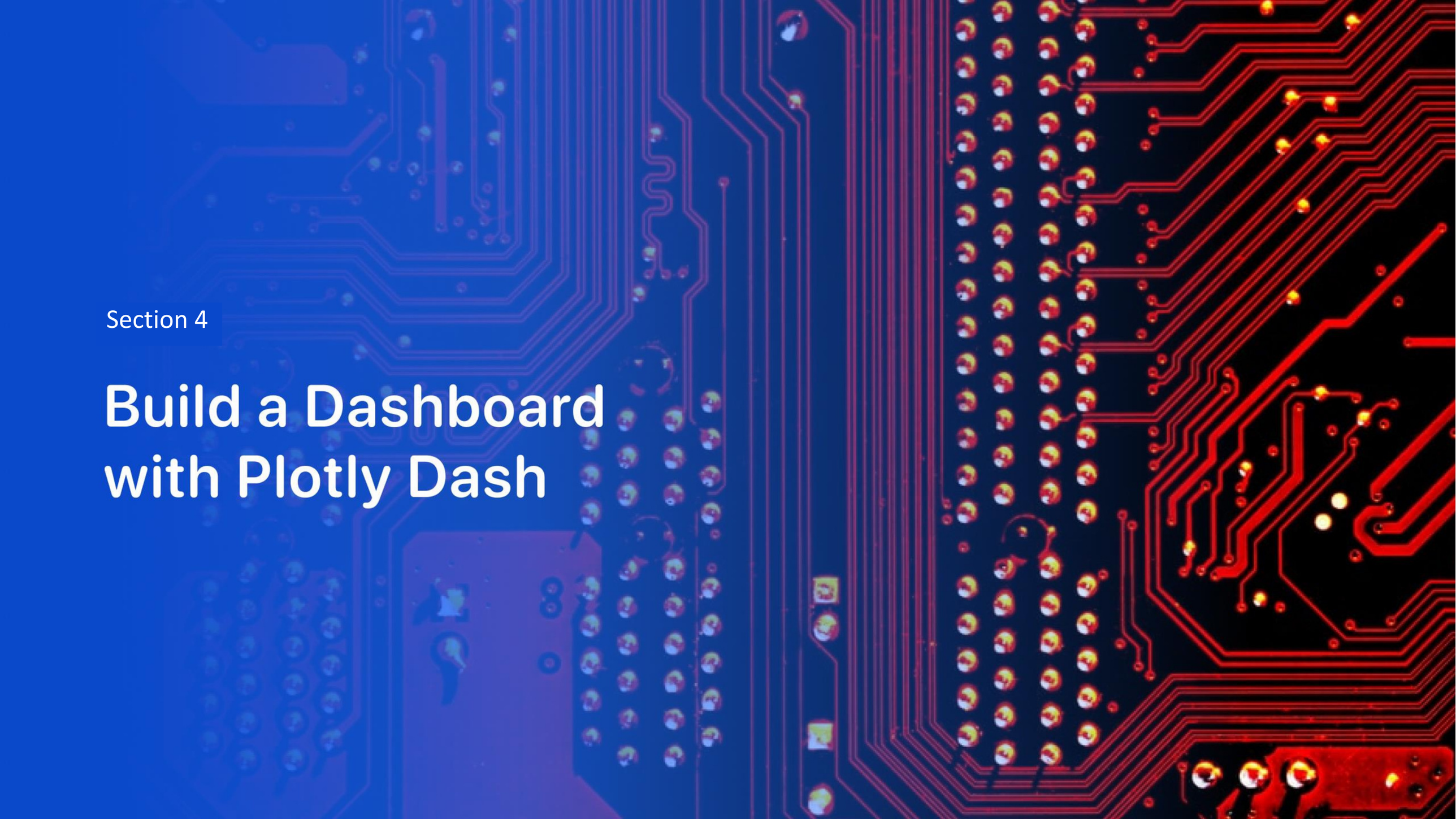
Distance Between A Launch Site to its Proximities

•We noticed that the launch sites are far from railways, highways and cities.
The launch sites are mostly close to coastline, the below figure is an
e xam ple .



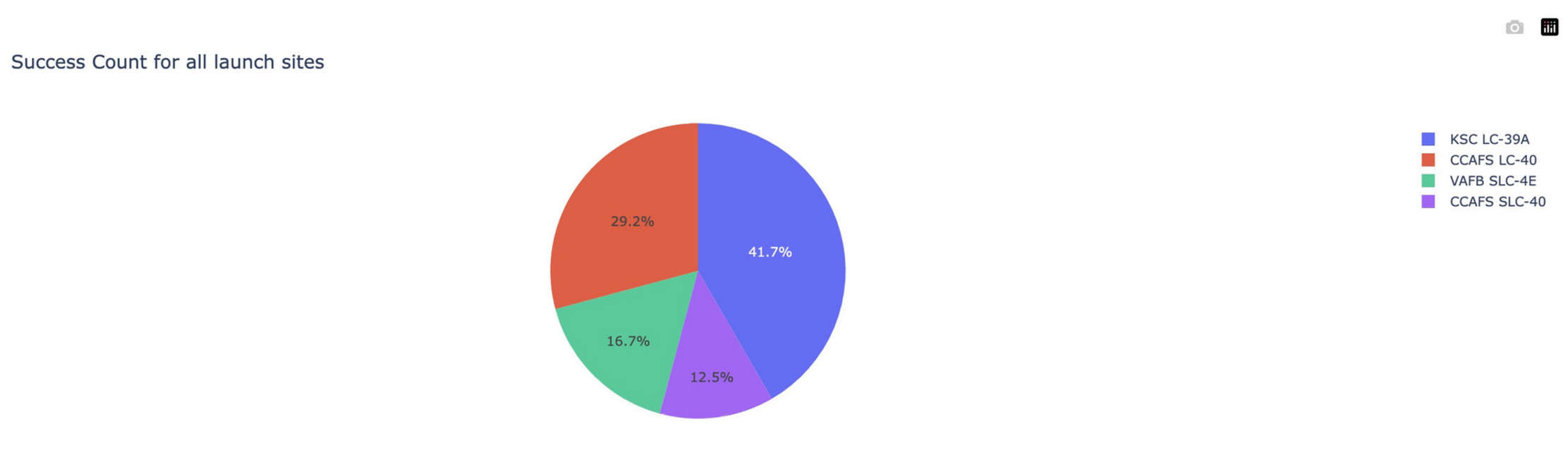
Section 4

Build a Dashboard with Plotly Dash



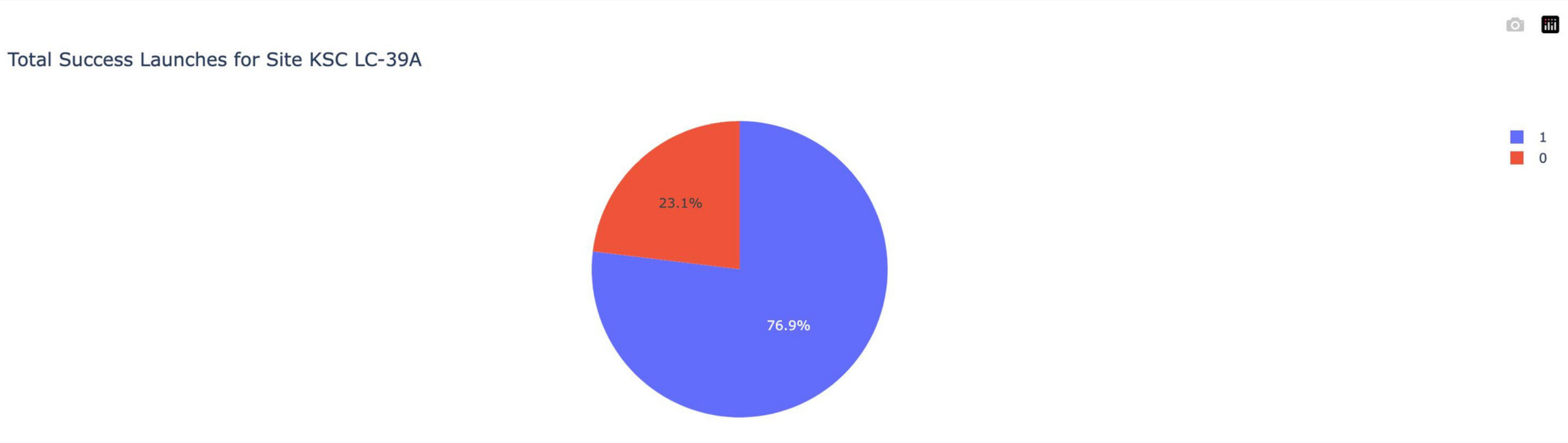
Total Success Launches for all sites

- As shown in the below chart, KSC LC-39A have the top successful rate for all launch sites



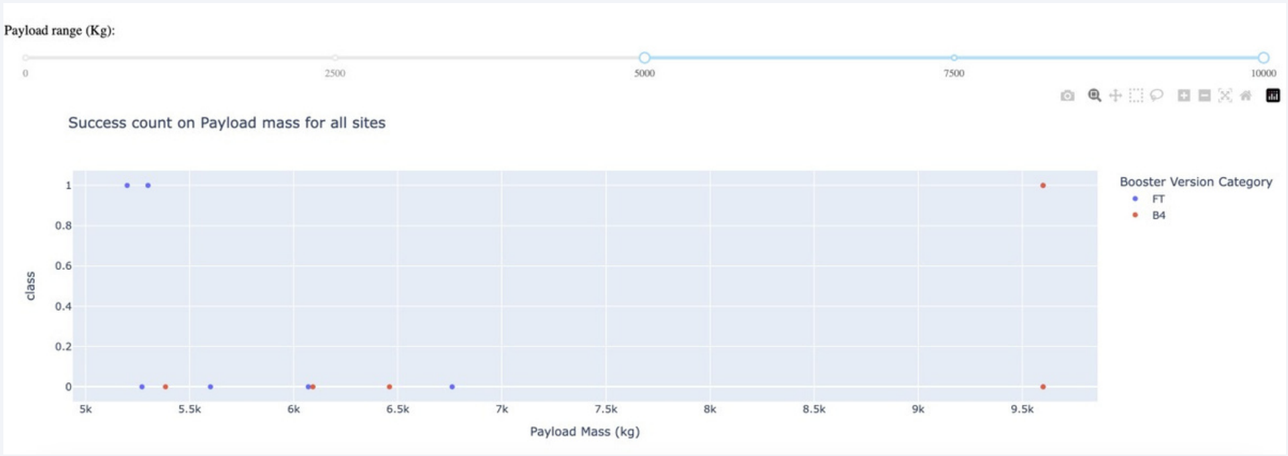
KSC LC-39A Highest Success Ratio

- KSC LC-39A launch site have 23.1% Failure & 76.9% Success



Payload vs. Launch Outcome

- The below charts shows that the low weighted Payload Mass have the largest success rate

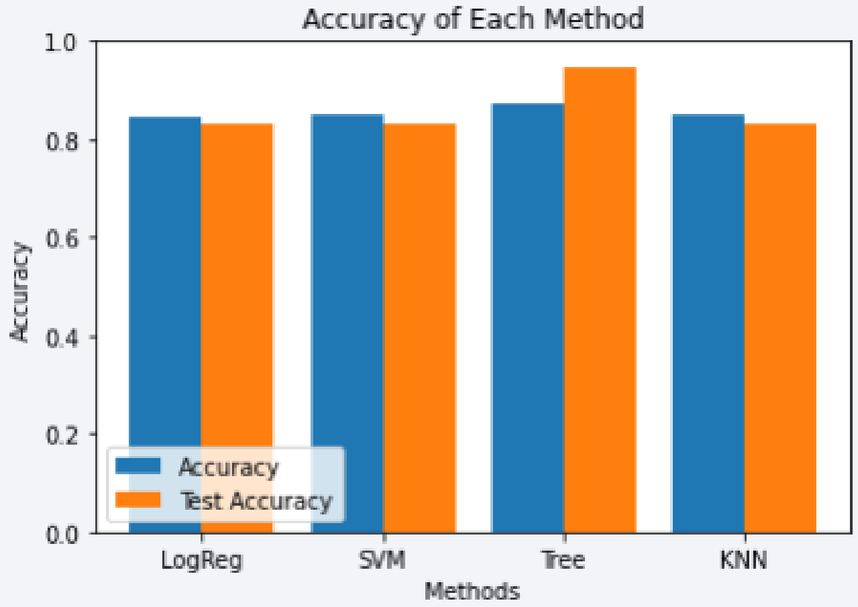


Section 5

Predictive Analysis (Classification)

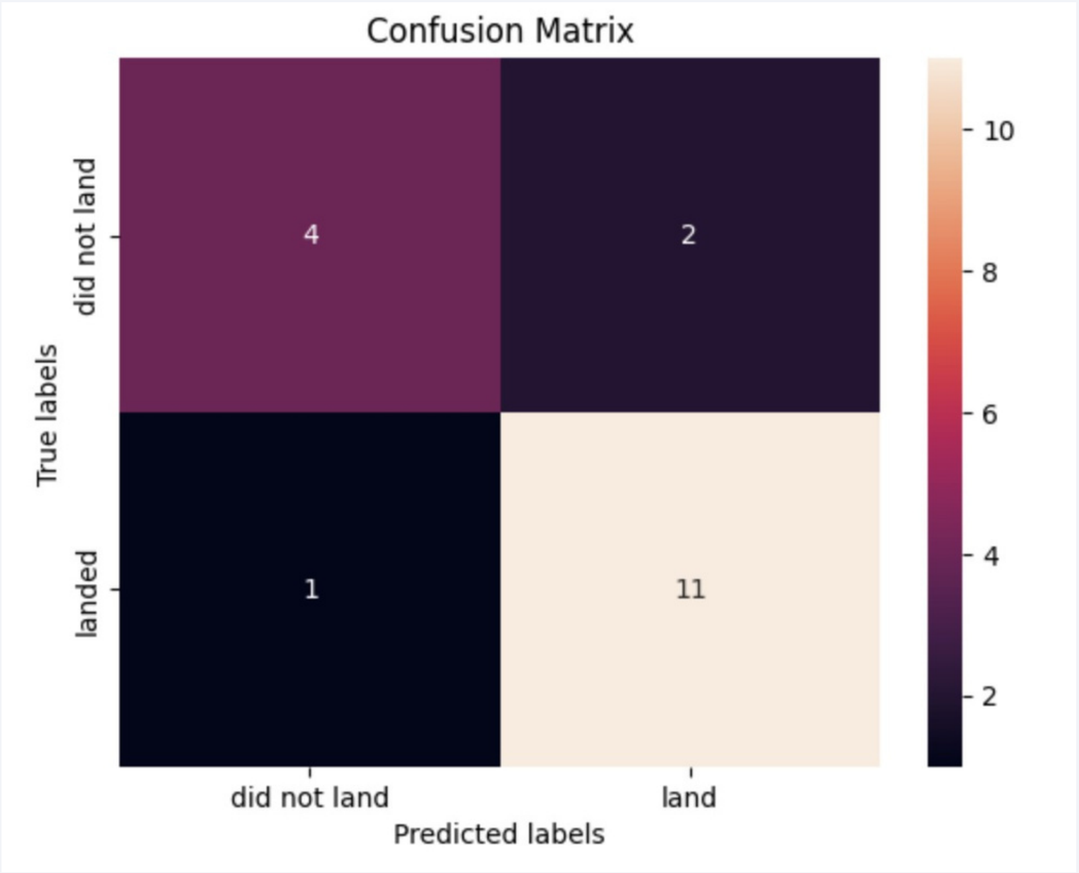
Classification Accuracy

- Decision Tree Model has the Highest Accuracy



Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier is predicted correctly 15 out of 18,
  wherethere're 2 successful landing butthe classifier marked them as unsuccessful landing(false positive)



Conclusions

- To conclude:
- The success rates forSpaceX launches is directly proportional time in yearsthey willeventually perfect the launches
- KSC LC 39A had the most successful launches from all the sites
- Decision Tree Model Achieved the highest Accuracy
- Payloads mass with low weight are more likely to have a successful landing than the heavy weightedPayloads
- Orbit ESL-1,GEO,SSO,HEO has the highest success rates

Thank you!