IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Natalya Matviyenko
2022

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 a Dashboard with Plotly Dash
  - Predictive analysis
- Summary of all 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 SpaceX can reuse the first stage.

- Problems you want to find answers

  The project task is to predicting if the first stage of the SpaceX Falcon 9 rocket will land successfully.

Section 1

Methodology

Methodology

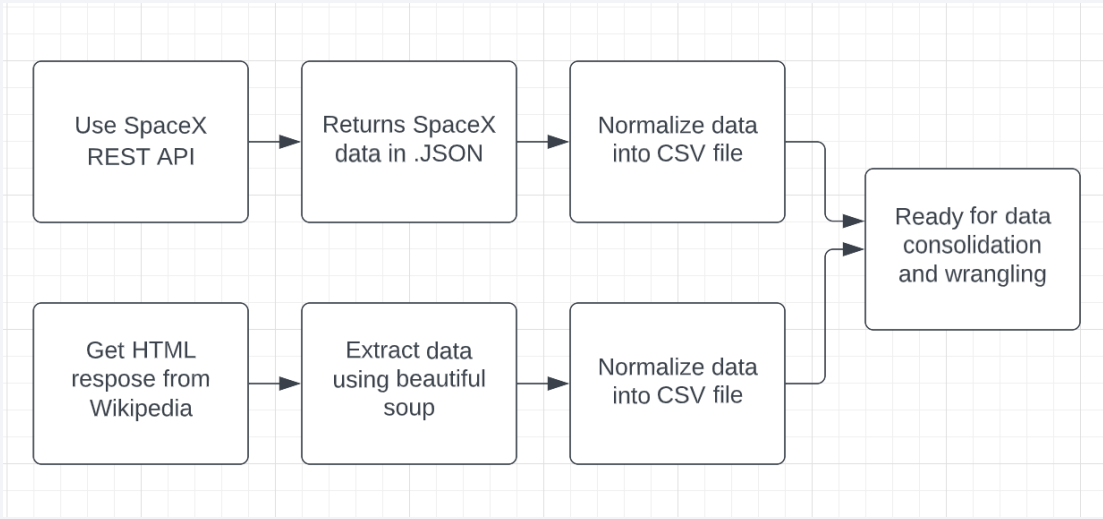
Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - One Hot Encoding data fields for Machine Learning and data cleaning of null values and irrelevant columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - LR, KNN, SVN, DT models have been built and evaluated for the best classifier

Data Collection

- The following data sets was collected:
  - SpaceX launch data that is gathered from SpaceX REST API
  - This API gives us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications and landing outcome.
  - The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/.
  - Another popular data source for obtaining Falcon 9 launch data is web scraping Wikipedia using BeautifulSoup.

SpaceX API



Web Scraping

Data Collection – SpaceX API

1. Getting response from API

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

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

2. Converting responce to a .json file

```
response.json()
data = pd.json_normalize(response.json())
```

3. Apply custom function to clean data

```
In [18]: # Call getLaunchSite
         getLaunchSite(data)

In [19]: # Call getPayloadData
         getPayloadData(data)

In [20]: # Call getCoreData
         getCoreData(data)
```

4. Assign list to dictionary then dataframe

```
In [21]: 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}

df=pd.DataFrame(launch_dict)
```

5. Filter dataframe and export to csv file

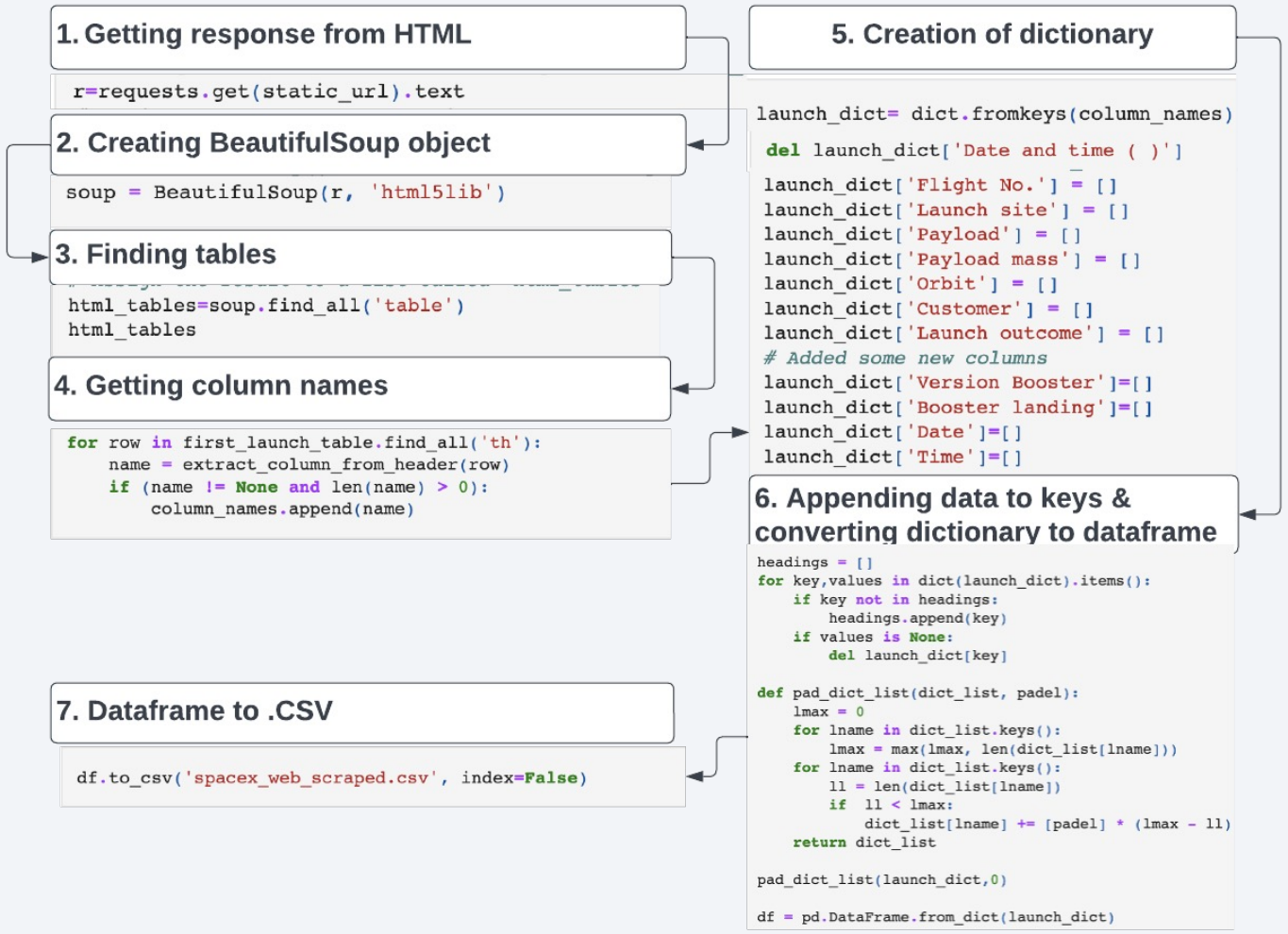
```
data_falcon9 = df[df.BoosterVersion == 'Falcon 9']
data_falcon9

In [33]: data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

Link to the notebook:
https://github.com/Natalya-Ukraine/Applied-Data-Science-Capstone/blob/main/Data Collection API.ipynb

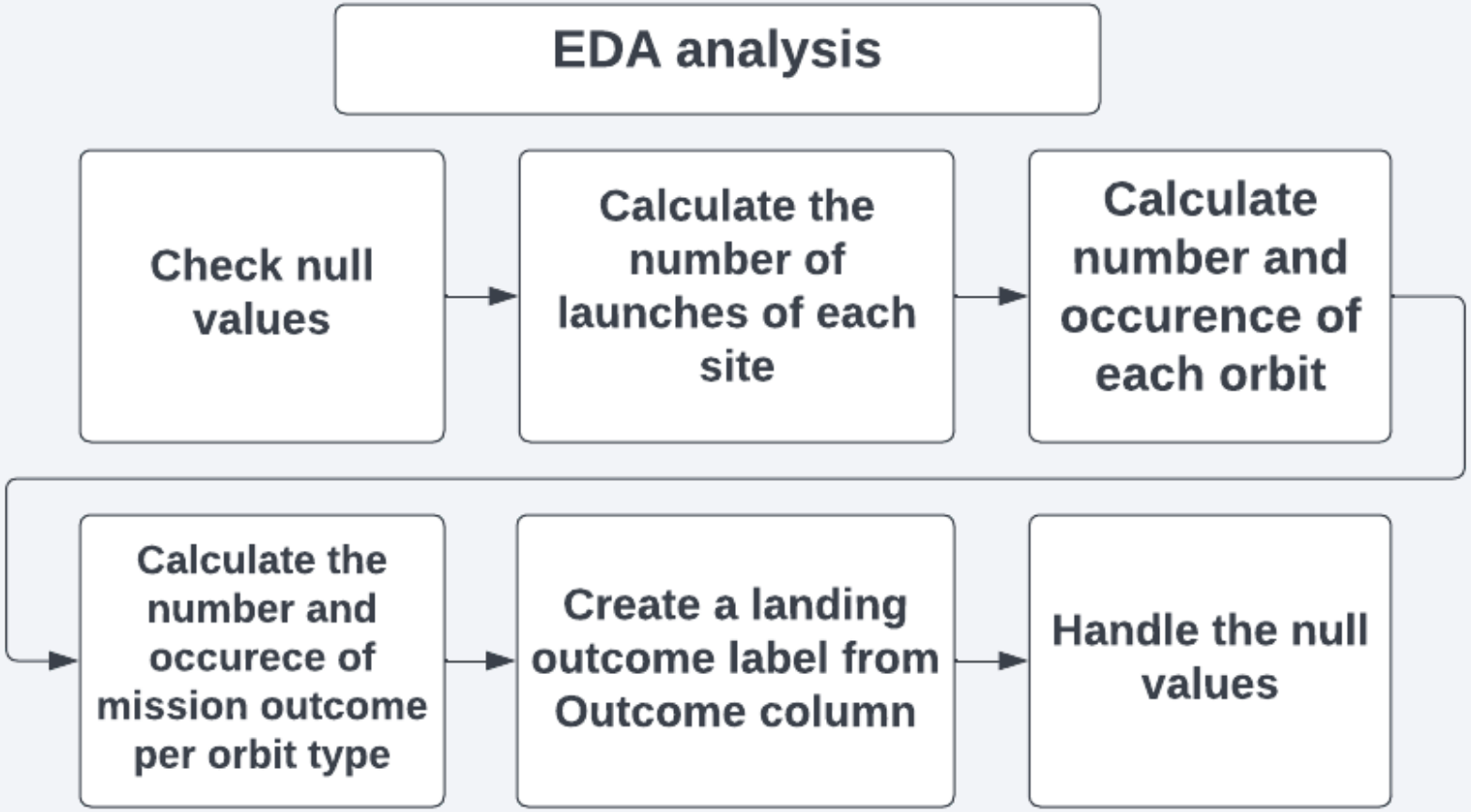
Data Collection - Scraping

- Web Scraping from Wikipedia

- https://github.com/Natalya-Ukraine/Applied-Data-Science-Capstone/blob/main/Data Collection with Web Scraping.ipynb



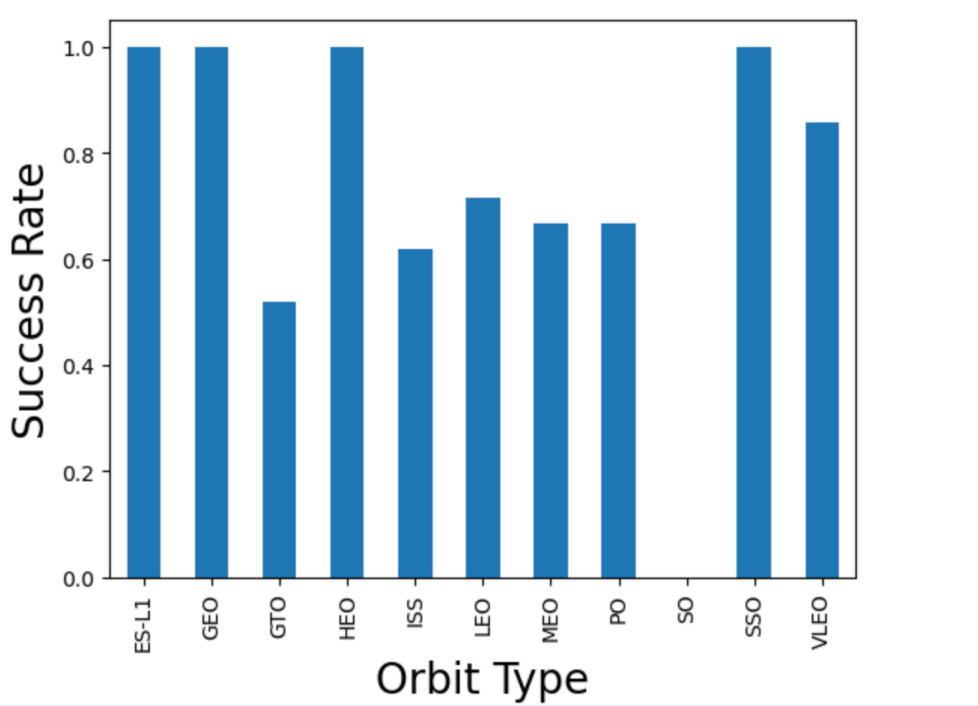
Data Wrangling

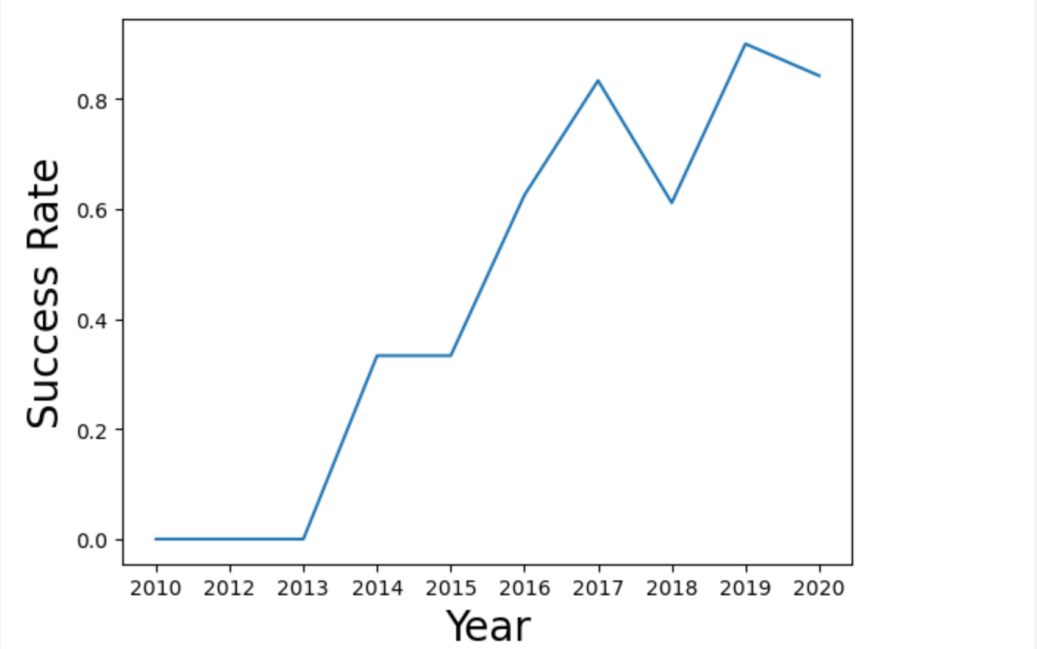


https://github.com/Natalya-Ukraine/Applied-Data-Science-Capstone/blob/main/Data Wrangling.ipynb

EDA with Data Visualization

- We explore the data by visualizing the relationship between flight number and launch site, payload and launch site, success rate of each orbit type, the launch success early trend.





https://github.com/Natalya-Ukraine/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb

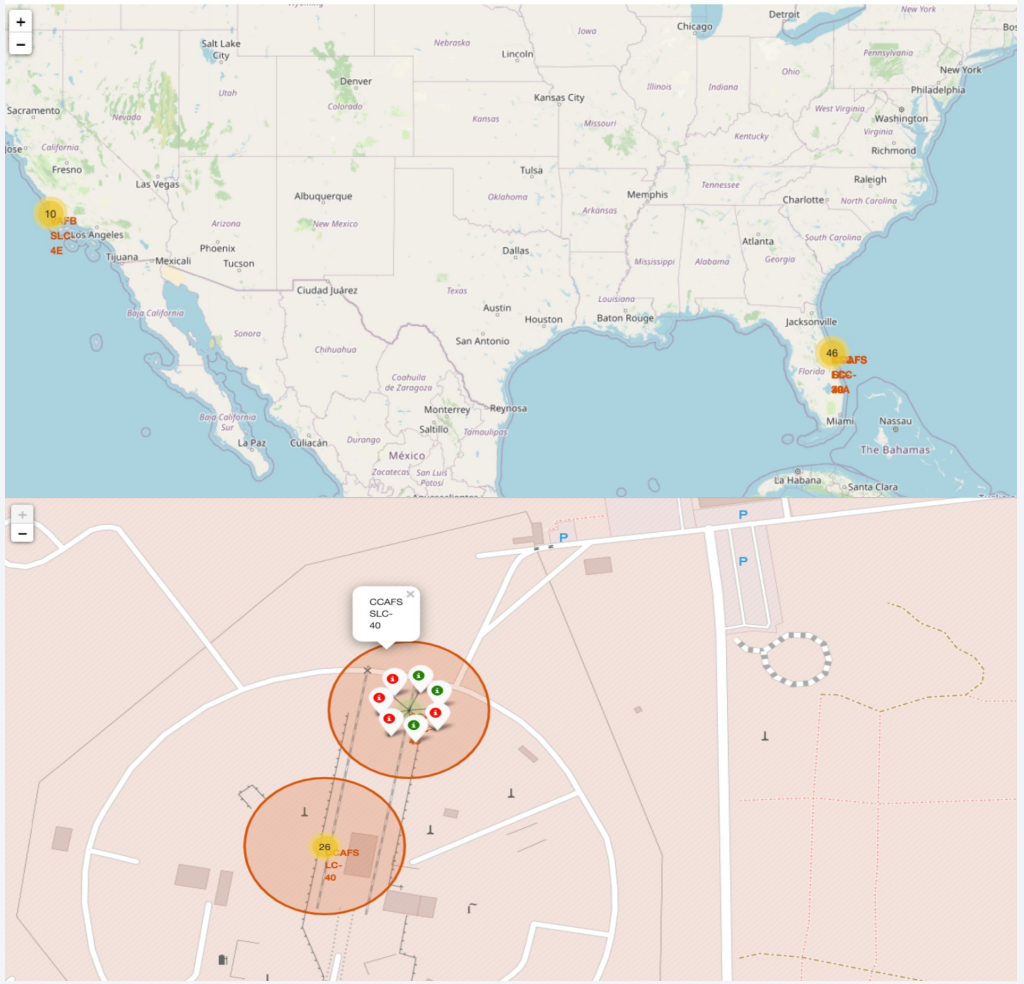
EDA with SQL

- SQL queries performed include:
  - Displaying the names of the unique launch sites in the space mission
  - Displaying five records where launch sites begin with 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 when the first successful landing outcome in ground pad was achieved
  - Listing the names of the boosters which have success in drone ship 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 display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015
  - Ranking the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order
- https://github.com/Natalya-Ukraine/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

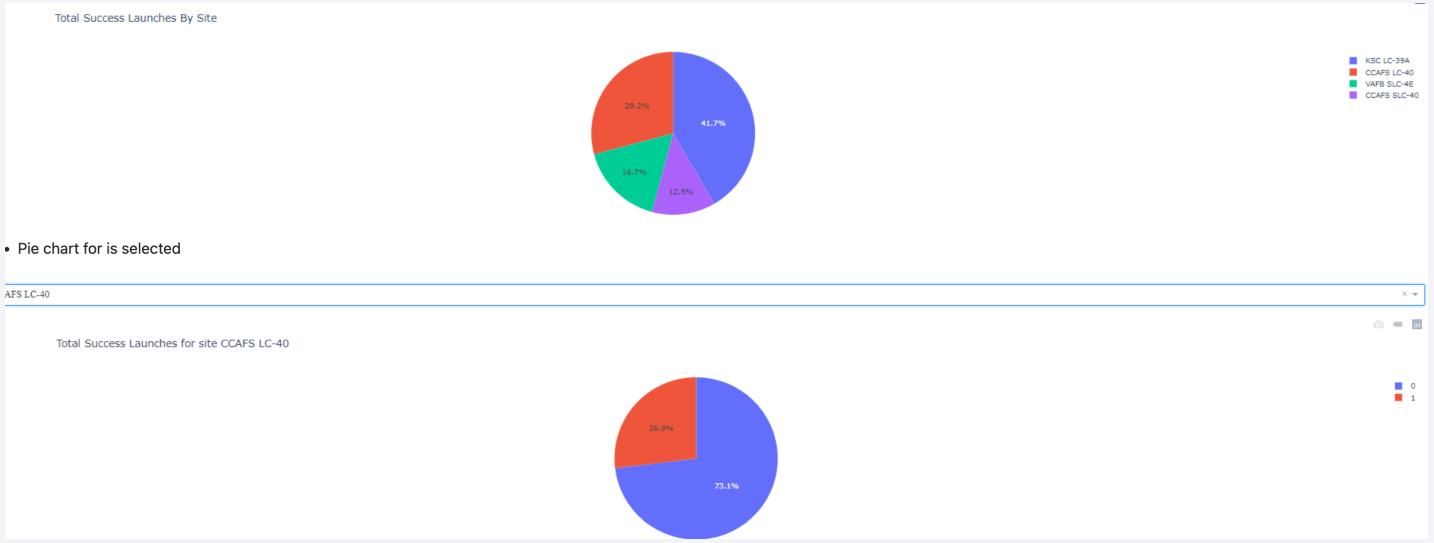
Build an Interactive Map with Folium

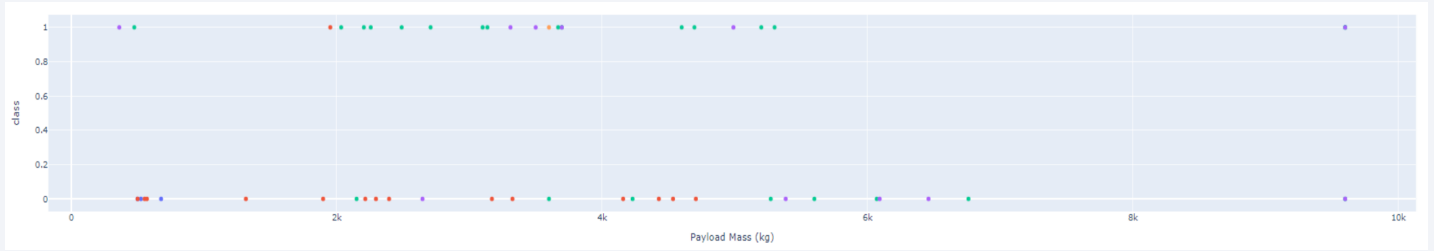
- Marked all launch sites on a map
- Marked the success/failed launches for each site on the map
- Calculated the distances between a launch site to its proximities
- After that we can answer the following questions :
  - •Are launch sites in close proximity to railways?
  - •Are launch sites in close proximity to highways?
  - •Are launch sites in close proximity to coastline?
  - •Do launch sites keep certain distance away from cities?

https://github.com/Natalya-Ukraine/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb



Build a Dashboard with Plotly Dash



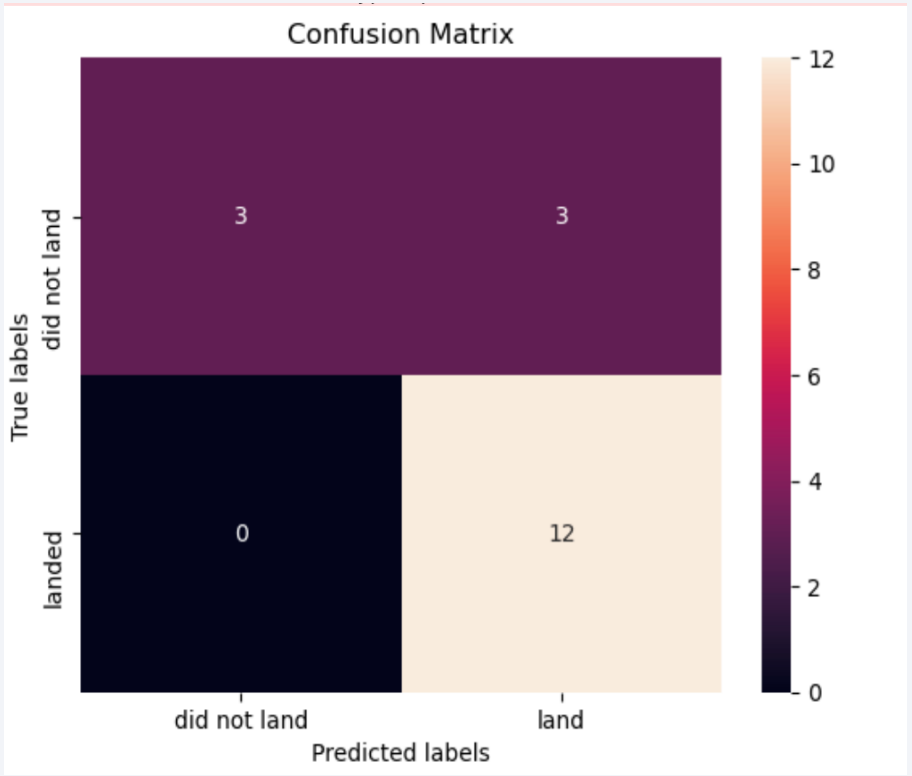


- We plotted pie charts showing the total launches by a certain sites
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.

https://github.com/Natalya-Ukraine/Applied-Data-Science-Capstone/blob/main/python3 spacex_dash_app.py

Predictive Analysis (Classification)

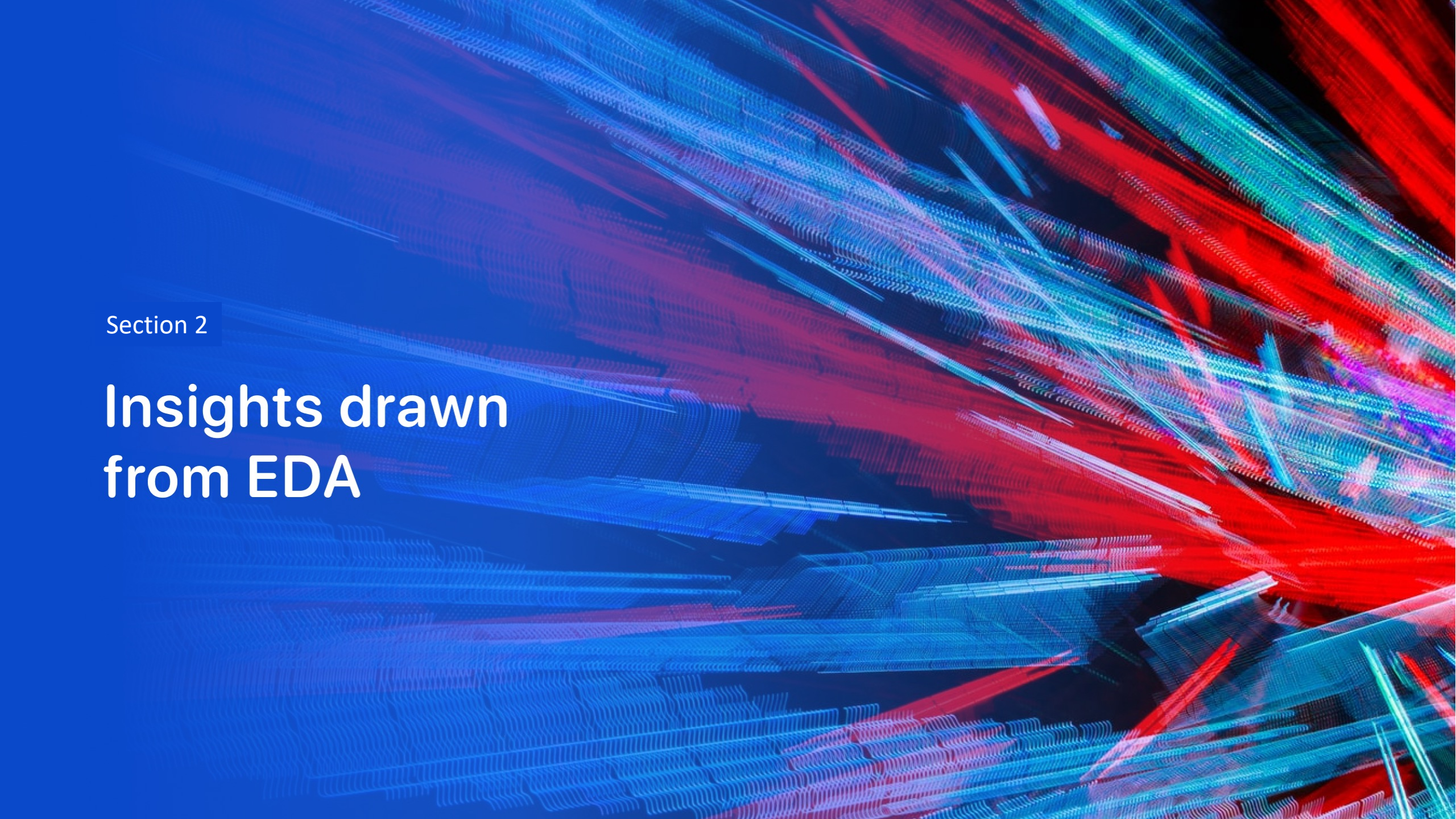
- We loaded the data using NumPy and Pandas, transformed the data, split our data into training and testing.
- We built different machine learning models and tune different hyperparameters using GridSearchCV.
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- We found the best performing classification model.



- https://github.com/Natalya-Ukraine/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine Learning Prediction_Part_5.ipynb

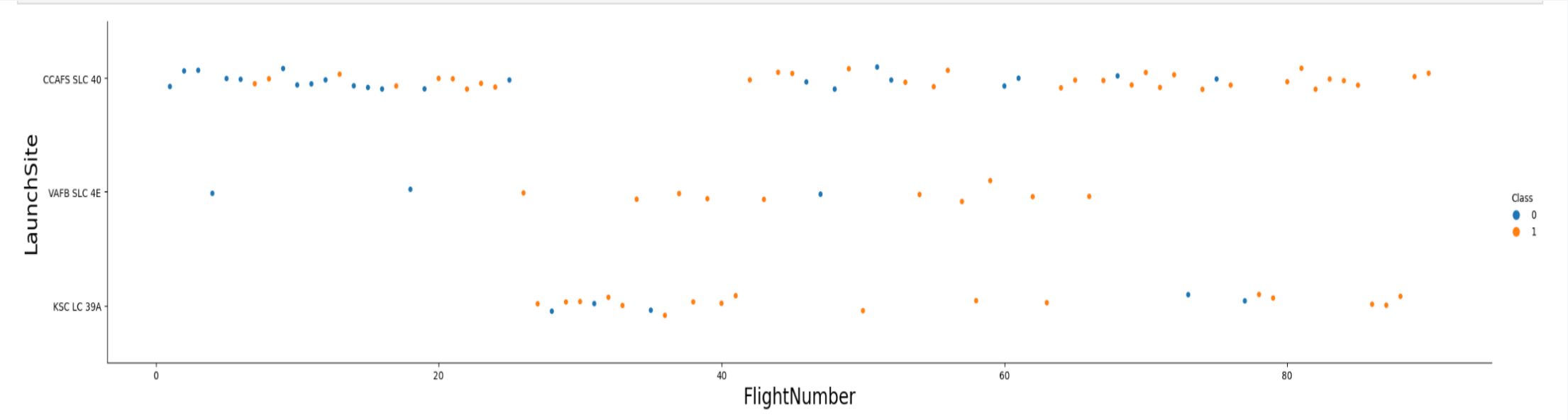
Results

- The CVM, KNN, and Logistic Regression models are the best in terms of predictions accuracy for this dataset.
- Low weighted payloads perform better then the heavier payloads.
- The success rate for SpaceX launches is directly proportional to time, in years they will eventually perfect the launches.
- KSC LC 39A had the most successful launches from all sites
- Orbit GEO, HEO, SSO, ES L 1 has the best success rate.


Section 2
Insights drawn from EDA

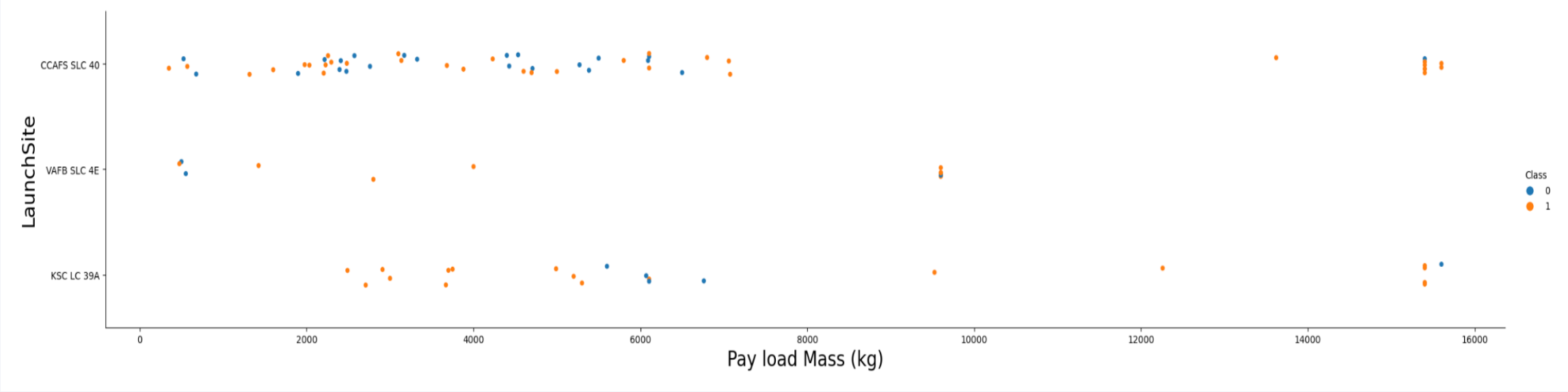
Flight Number vs. Launch Site

- From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.



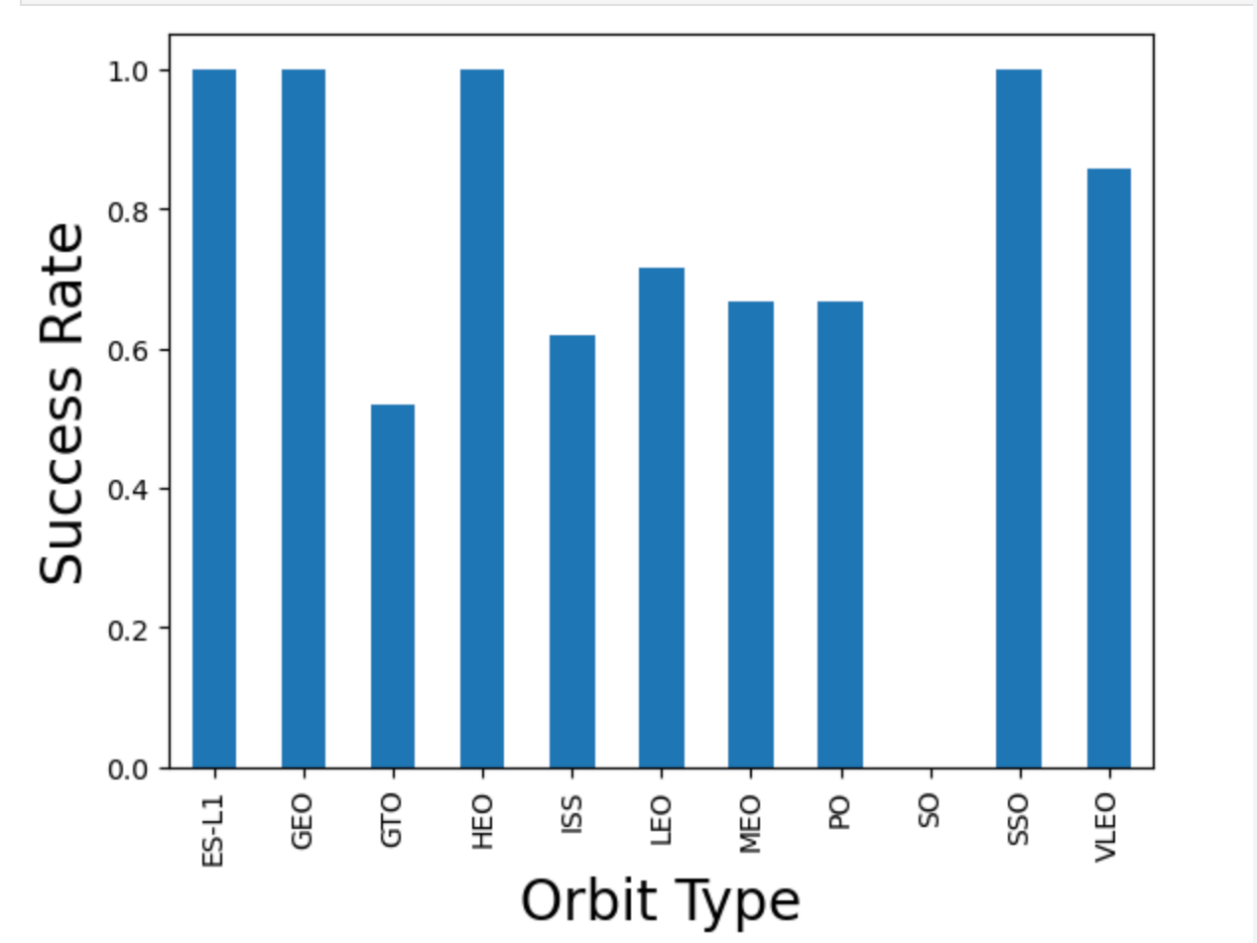
Payload vs. Launch Site

- The majority of low mass loads have been launched from CCAPS SLC 40



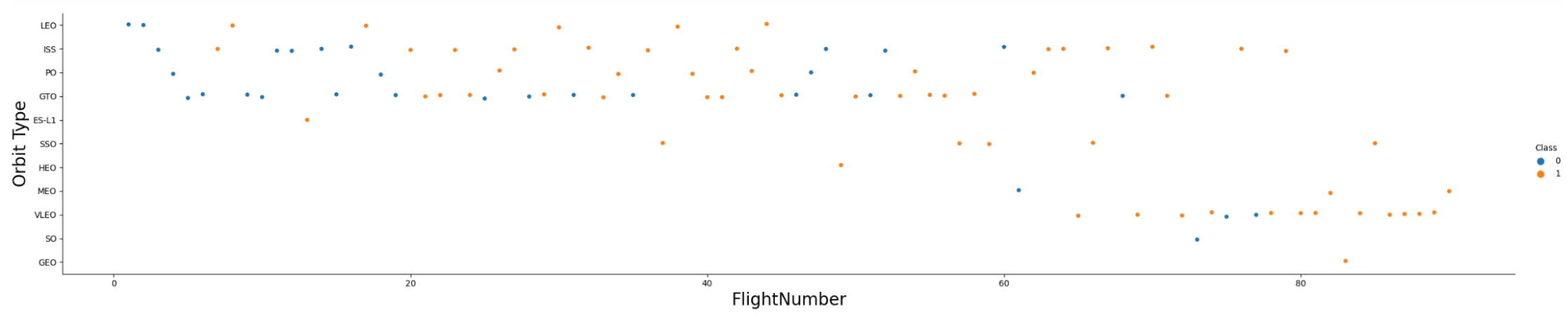
Success Rate vs. Orbit Type

- The orbit types of ES-L1, GEO, HEO, SSO are among the highest success rate.



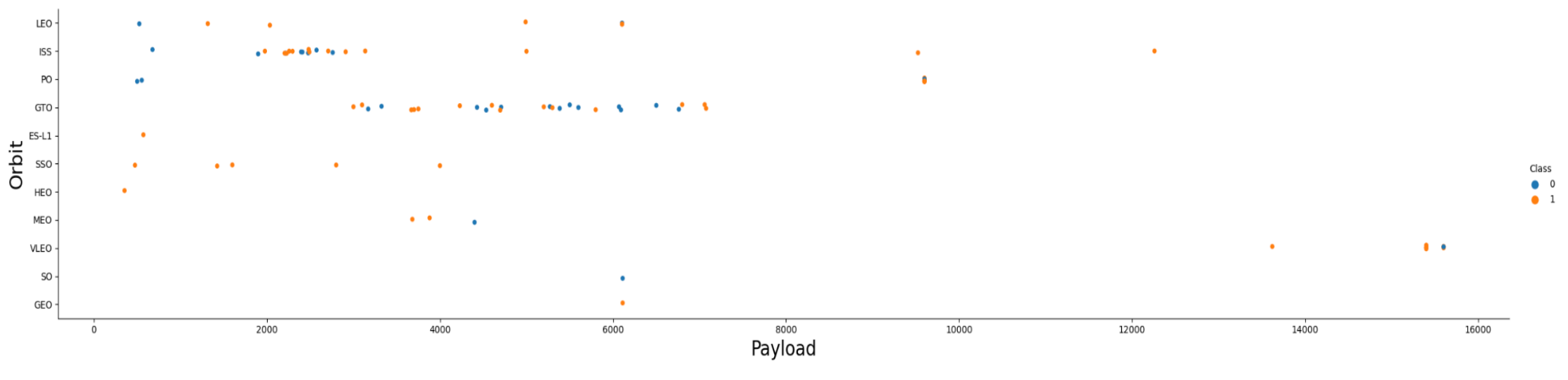
Flight Number vs. Orbit Type

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



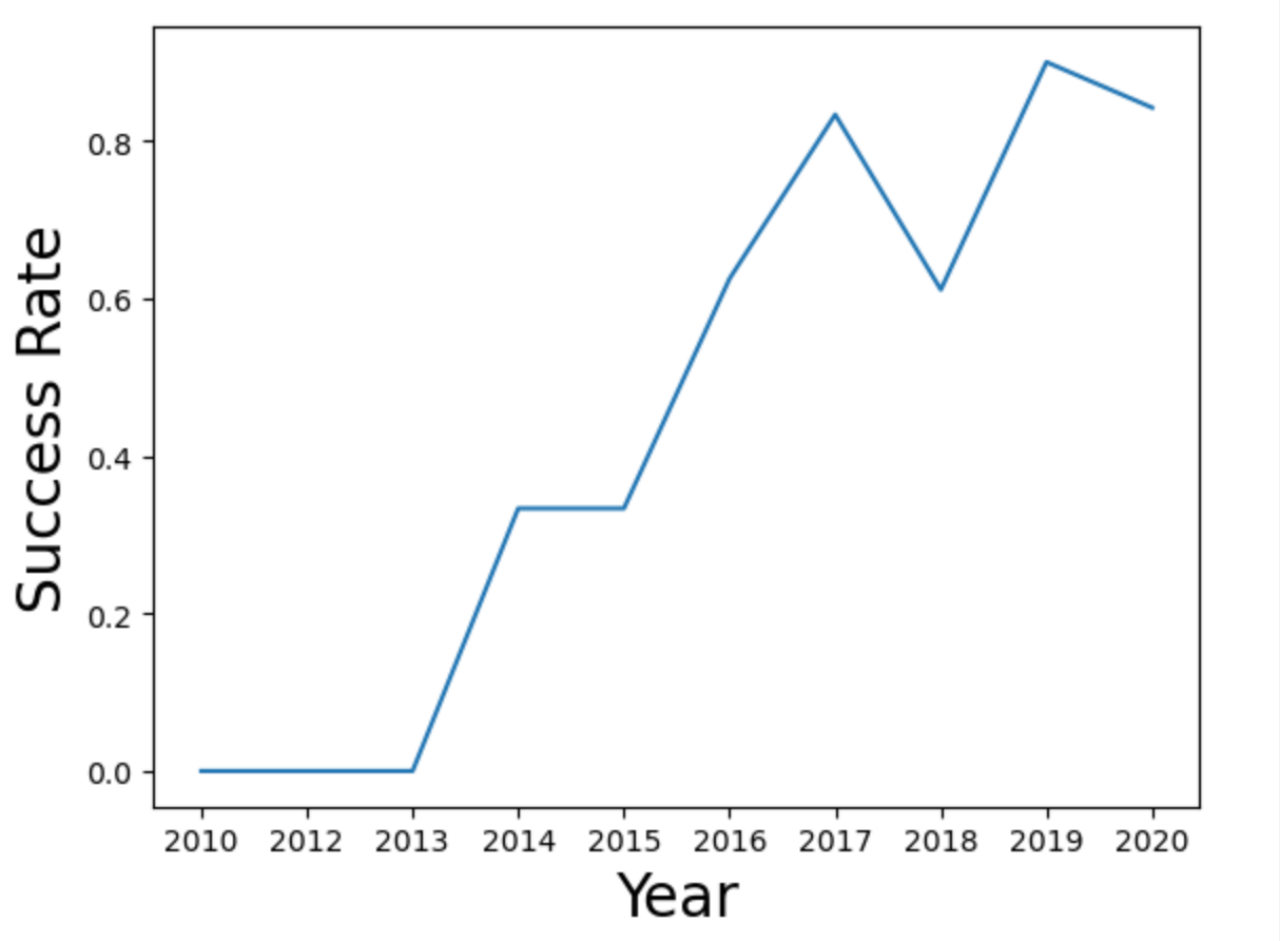
Payload vs. Orbit Type

- There are strong correlation between ISS and payload at the range around 2000, as well as between GTO and range of 4000-8000.



Launch Success Yearly Trend

- Launch success rate since 2013 kept increasing till 2020, potentially due to progress in technology and lessons learned.



All Launch Site Names

- To find unique launch site names in SpaceX data we used DISTINCT key word.

```
In [7]: %sql SELECT Distinct LAUNCH_SITE FROM SPACEXTBL;
 * sqlite:///my_data1.db
Done.
```

Out[7]:

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

Launch Site Names Begin with 'CCA'

In [8]:
```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

Out[8]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

In [9]:
```
%sql SELECT SUM (PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer="NASA (CRS)";
```

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

Out[9]:

SUM (PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

```
In [10]: %sql SELECT AVG (PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
Out[10]: AVG (PAYLOAD_MASS__KG_)
         2928.4
```

First Successful Ground Landing Date

In [12]:
```
%sql SELECT min(Date) from SPACEXTBL where "Landing _Outcome" = 'Success (ground pad)';
```

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

Out[12]:

min(Date)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ between 4000 and 6000 AND "Landing _Outcome"='Success (drone ship)'
```

```
 * 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 Failure Mission Outcomes

In [15]:
```
%sql SELECT COUNT(*) FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE '%Success%' OR MISSION_OUTCOME LIKE '%Failure%'
```

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

Out[15]:

COUNT(*)
101

Boosters Carried Maximum Payload

In [18]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

Out[18]:

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

In [21]:

```
%%sql
SELECT substr(Date, 4, 2) as month,booster_version,"Landing _Outcome"
from SPACEXTBL where "Landing _Outcome"
='Failure (drone ship)' and substr(Date,7,4)='2015'
```

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

Out[21]:

month	Booster_Version	Landing _Outcome
01	F9 v1.1 B1012	Failure (drone ship)
04	F9 v1.1 B1015	Failure (drone ship)

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

In [22]:

```
%%sql

SELECT "Landing _Outcome",count("Landing _Outcome")as LANDING_OUTCOME_COUNT
from SPACEXTBL where  DATE between '04-06-2010' and '20-03-2017'
group by "Landing _Outcome" order by count("Landing _Outcome") desc
```

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

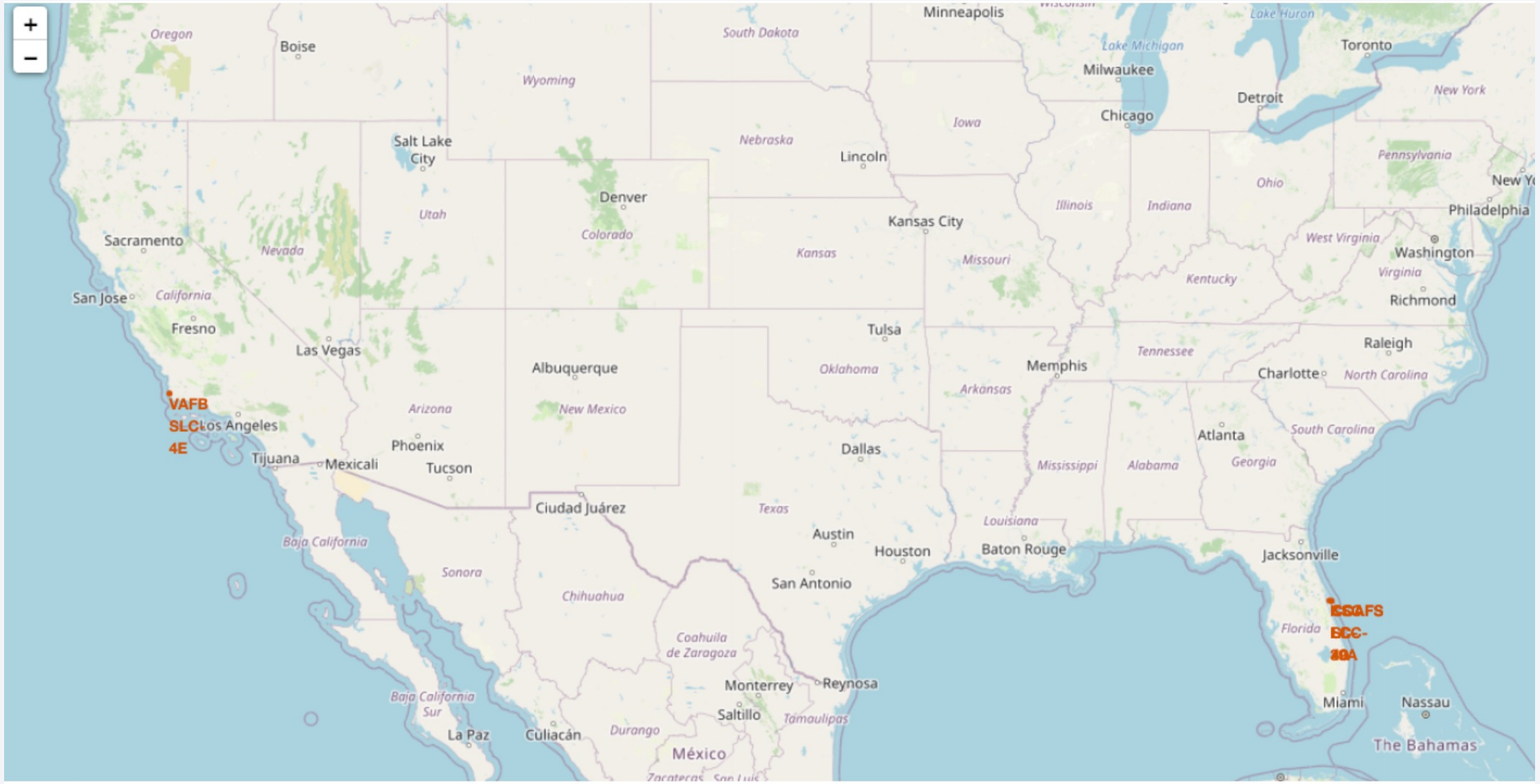
Out[22]:

Landing _Outcome	LANDING_OUTCOME_COUNT
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

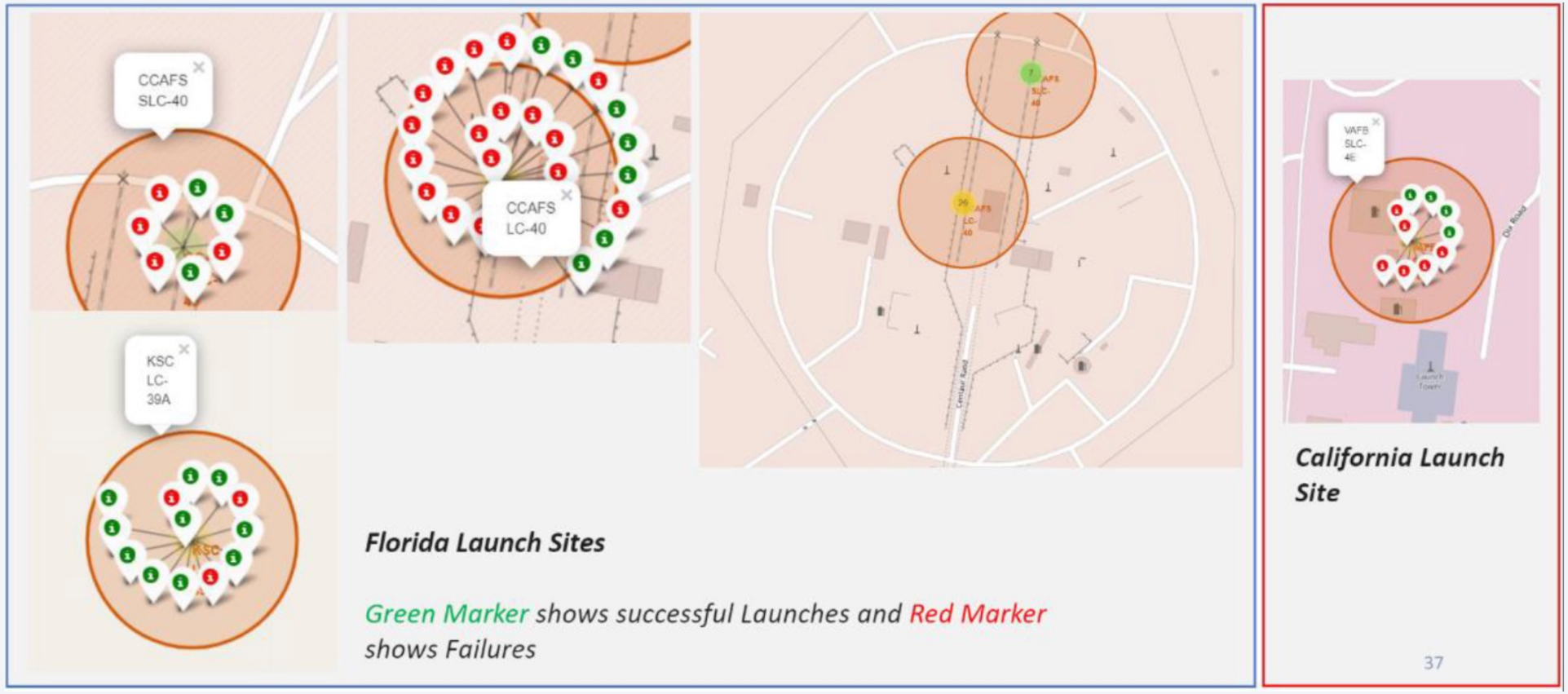

Section 3
Launch Sites
Proximities Analysis

All launch sites global map markers

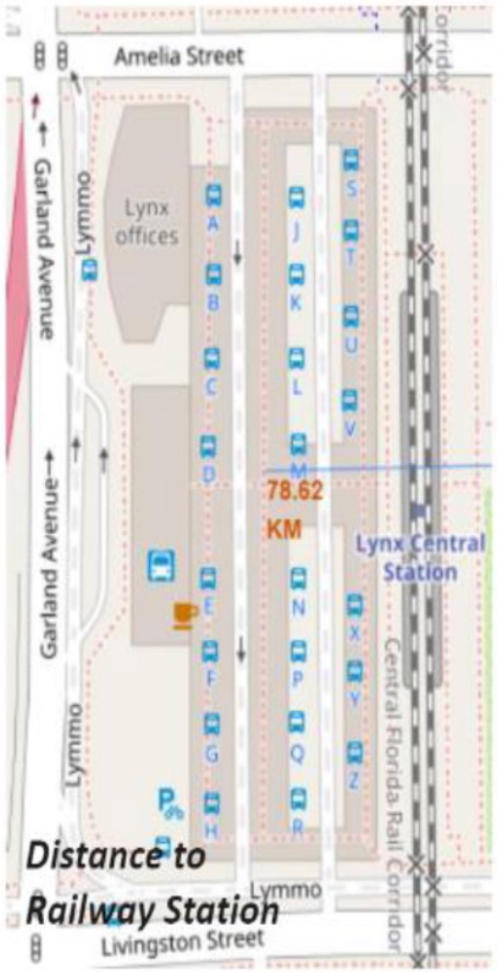
SpaceX launch sites located in USA Atlantic and Pacific coasts

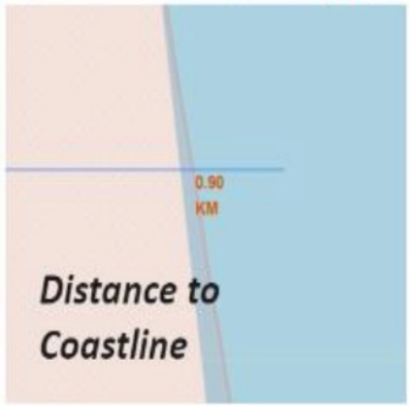


Success/failed launches marked on map

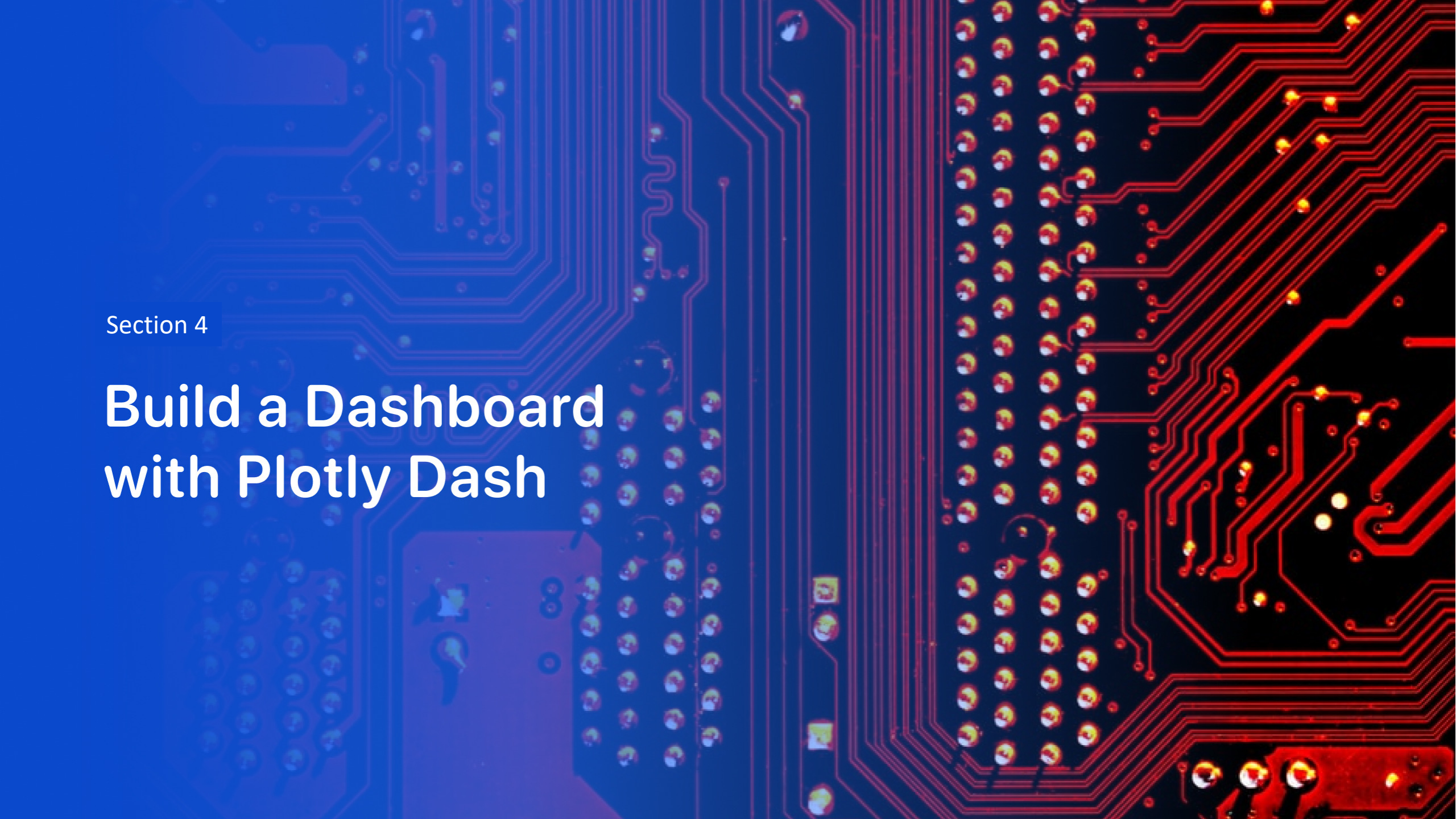


Distances between launch sites and its proximities

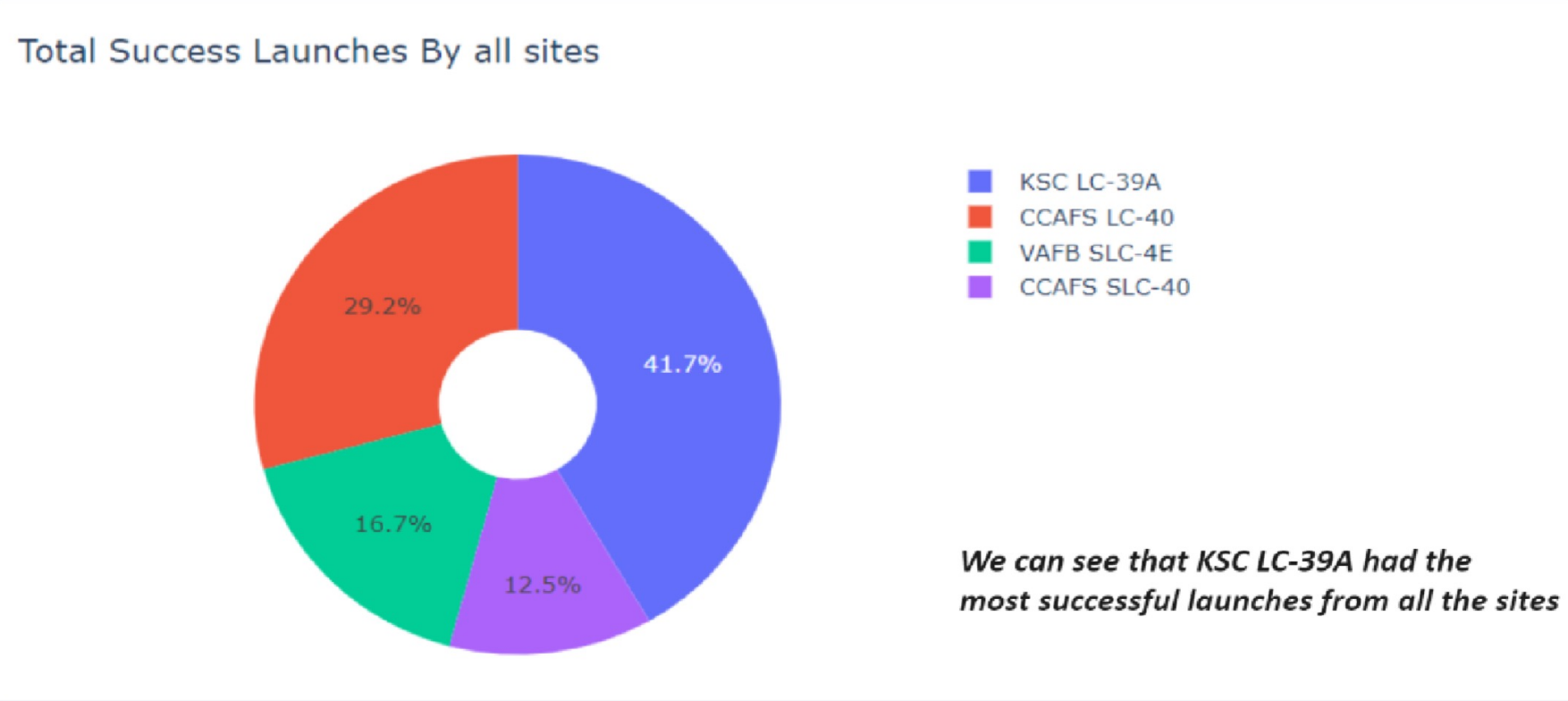




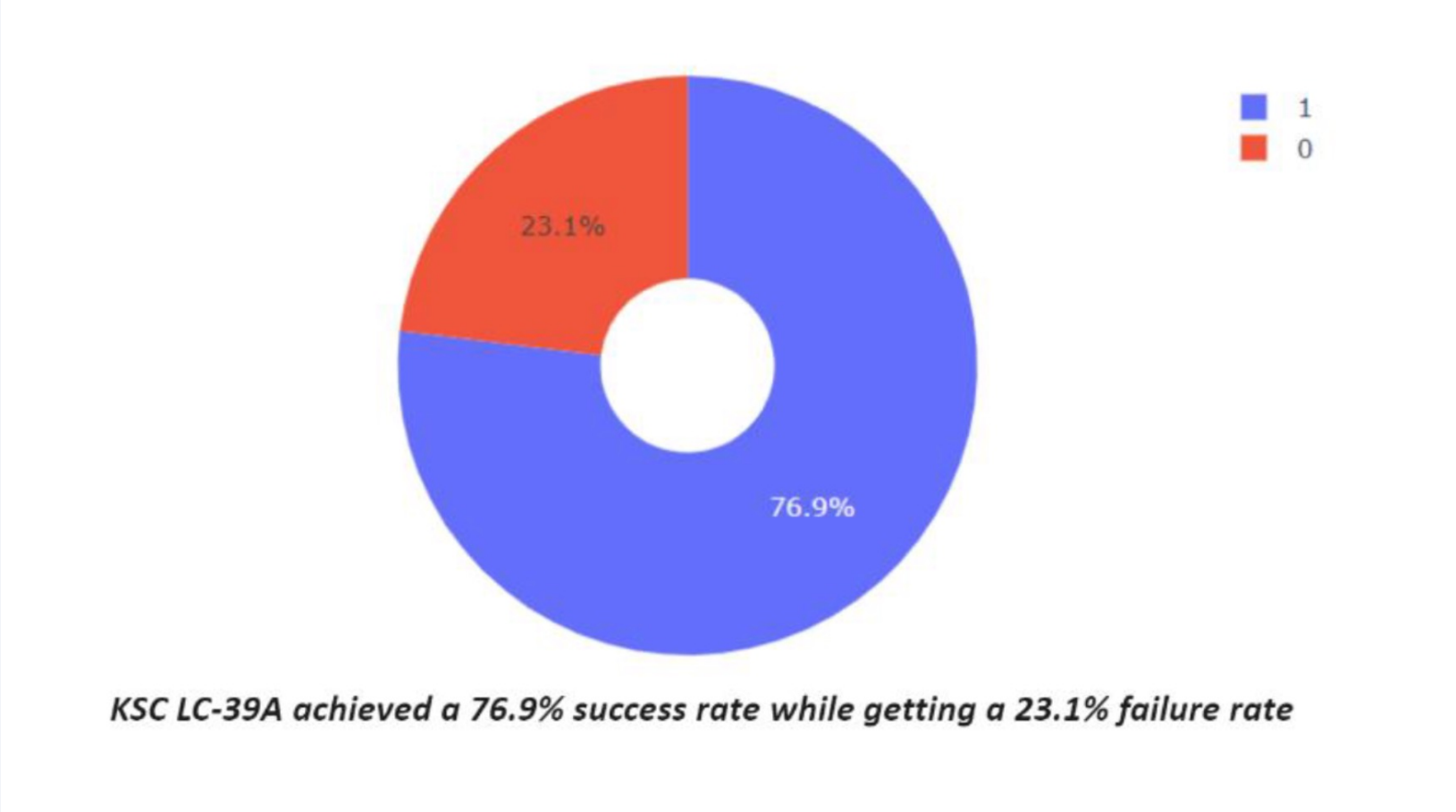
- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes


Section 4
Build a Dashboard with Plotly Dash

Total success launches by all sites

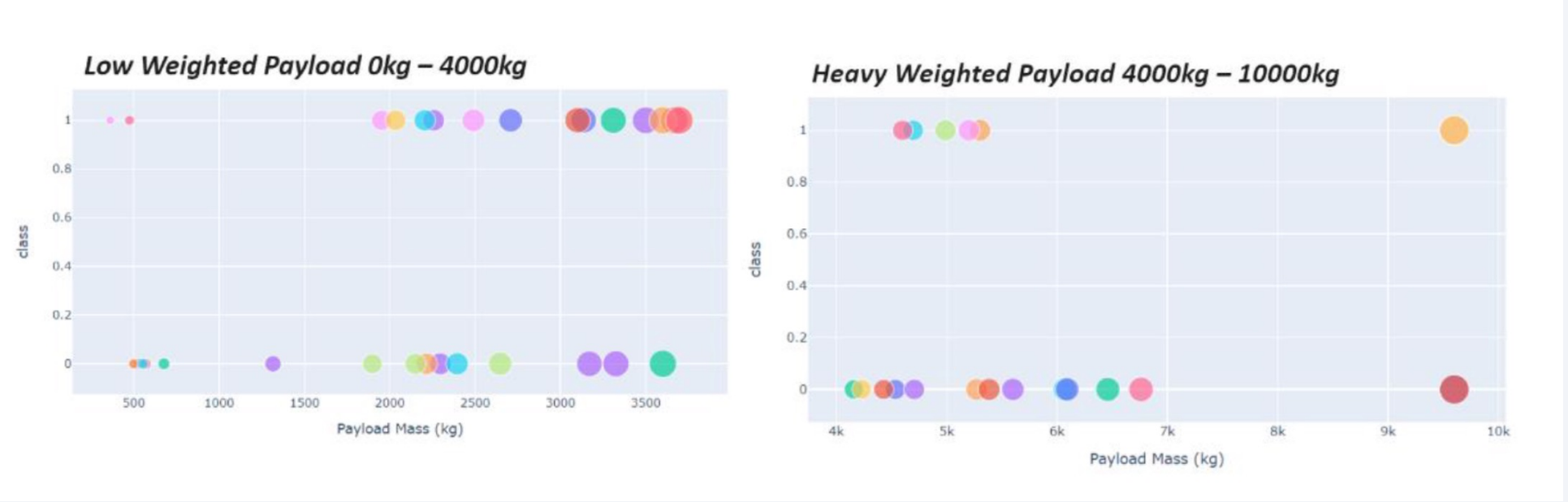


Launch site with the highest success rate



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Payloads vs Launch Outcome



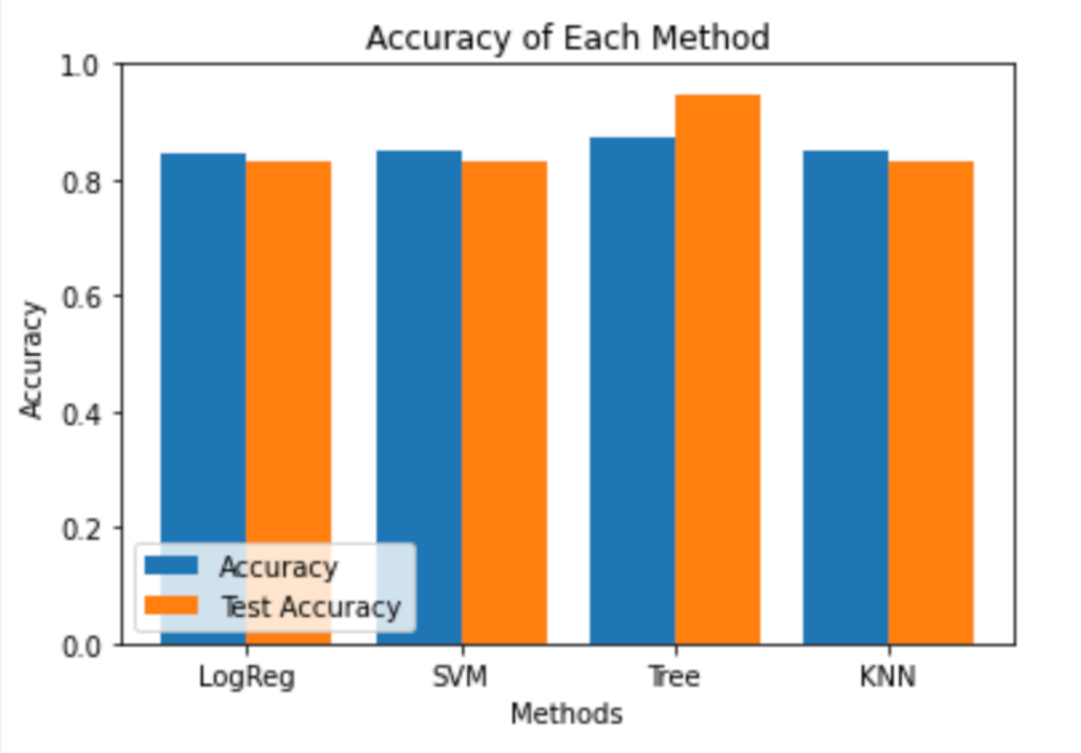
The success rate is higher for low weighted payloads.

Section 5

Predictive Analysis (Classification)

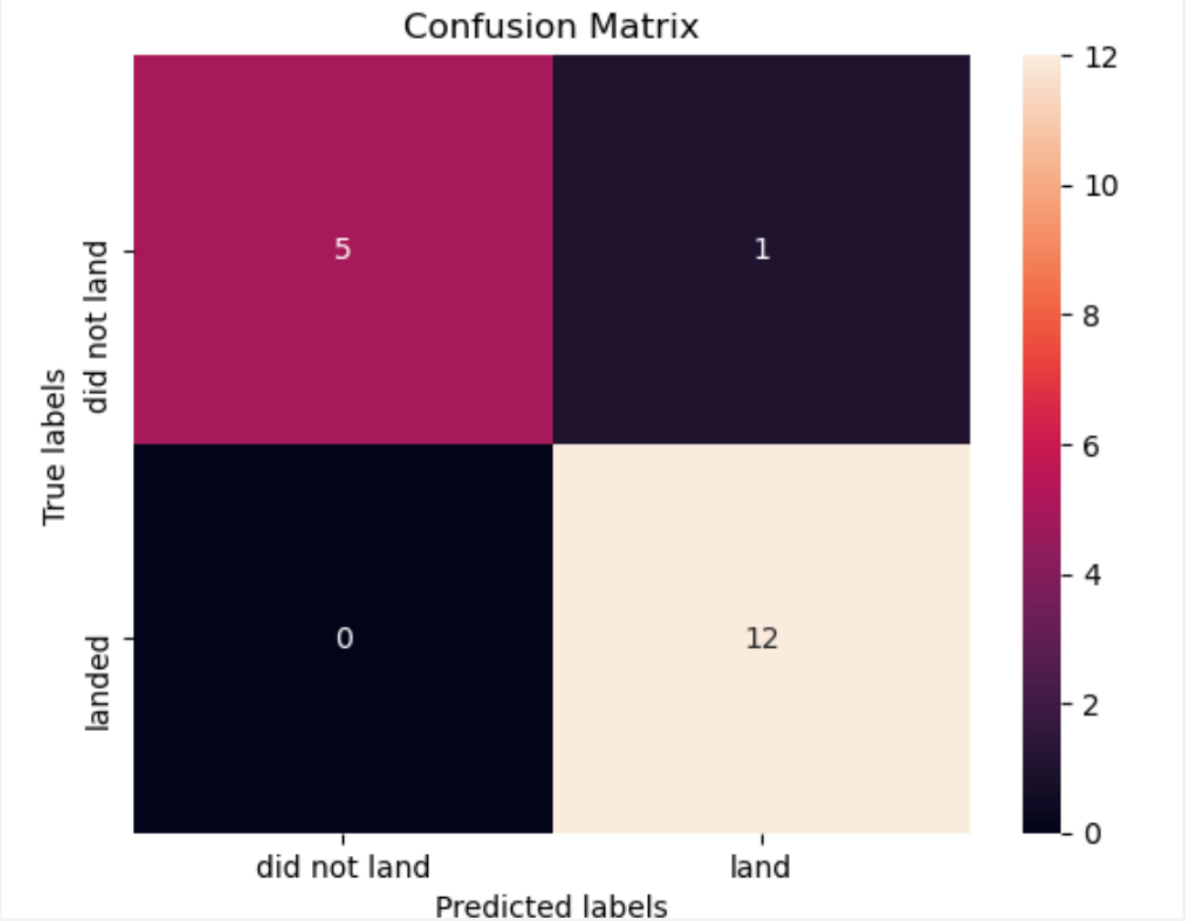
Classification Accuracy

- The Decision Tree Classifier is the model with highest classification accuracy .



Confusion Matrix

- Confusion Matrix for Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.



Conclusions

1. The success rate for SpaceX launches seems to improve over time according to evolution of process and rockets.
2. The most successful launch site is KSC LC-39A.
3. Orbits ES-L1, GEO, HEO, SSO had the most success rate.
4. Launches above 7,000 kg are less risky.
5. The Decision Tree Classifier is the best Machine Learning Algorithm for this task.

Appendix

Thank you!