IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Maksim Terentev
2024-07-02

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
  - Data collection using web scraping and SpaceX API
  - Data wrangling
  - EDA using SQL, pandas, matplotlib, folium and ploty dash
  - landing prediction using different ML models.
- Summary of all results
  - EDA results
  - Dashboards
  - Landing outcome predictions.

Introduction

- Project background and context

  This project aims to analyze SpaceX launch data to identify factors influencing landing success and build a predictive model to assist competitors in making informed bids against SpaceX. SpaceX's $62 million Falcon 9 launches are significantly cheaper than competitors' $165 million launches due to reusable first stages.

- Problems you want to find answers

  Predict the success of SpaceX Falcon 9 first stage landings to better estimate launch costs.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

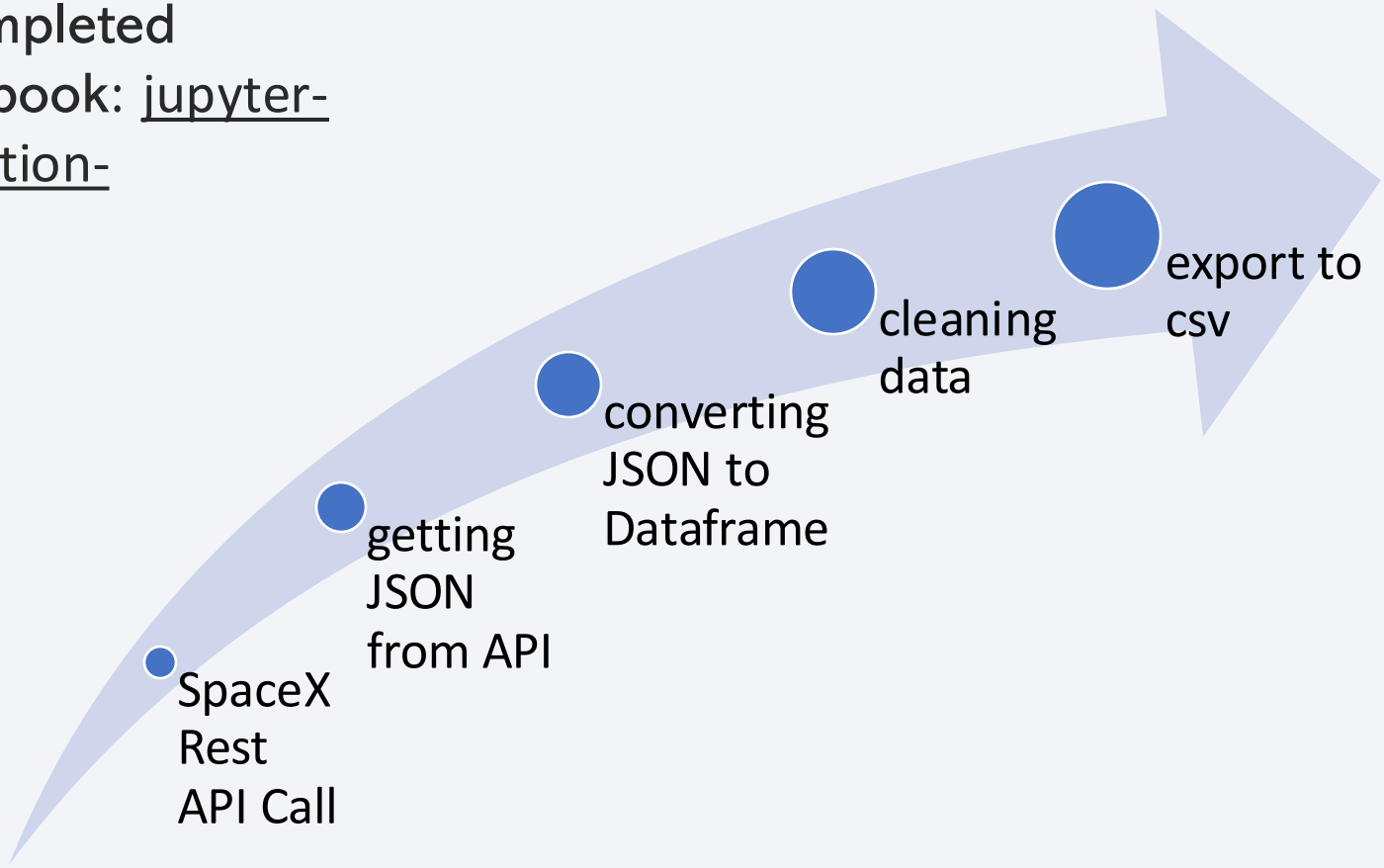
The data collection methodology involves web scraping for SpaceX launch details and utilizing the SpaceX API for real-time and historical data. These sources are integrated and cleaned to prepare a unified dataset for analysis. The API provided detailed information on Falcon 9 launches, including rocket specifics, payload details, launch site coordinates, and core data.

Steps involved:

- **API Requests:** Gathering historical launch data from the SpaceX API.
- **Data Extraction:** Extracting key details such as booster version, payload mass, orbital parameters, launch site coordinates, and core information.
- **Data Wrangling:** Cleaning and structuring the data to handle missing values and filter out irrelevant information.
- **Data Integration:** Consolidating the cleaned data into a unified format suitable for analysis

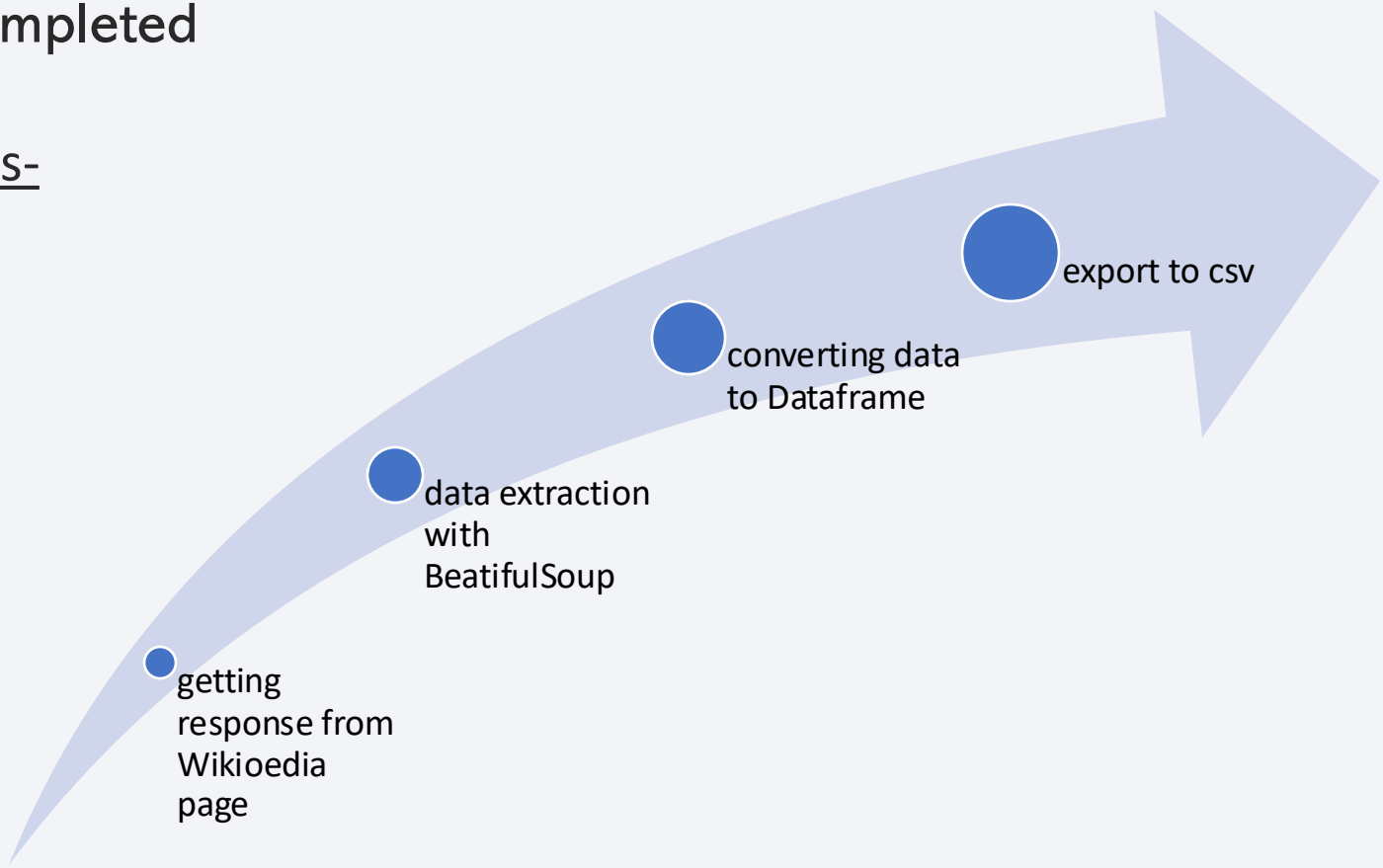
Data Collection – SpaceX API

- GitHub URL of the completed SpaceX API calls notebook: jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- GitHub URL of the completed web scraping notebook: jupyter-labs-webscraping.ipynb



Data Wrangling

- During data wrangling, string-based landing outcomes were converted to numerical labels (1 for successful, 0 for unsuccessful). Null values in the PayloadMass column were replaced with the mean of available data points, ensuring the dataset was standardized for further analysis and modeling.
- GitHub URL of completed data wrangling related notebooks: labs-jupyter-spacex-Data wrangling.ipynb

EDA with Data Visualization

- **PayloadMass vs. FlightNumber** (`sns.catplot`): Visualize the relationship between payload mass and flight number to understand its impact on launch success or failure.
- **LaunchSite vs. FlightNumber** (`sns.catplot`): Compare launch success/failure across different launch sites over multiple flights.
- **LaunchSite vs. PayloadMass** (`sns.catplot`): Examine how payload mass varies with launch site and its effect on launch outcomes.
- **Mean Success Rate by Orbit** (`sns.barplot`): Display the average success rate for different orbital destinations to identify performance variations.
- **Orbit vs. FlightNumber** (`sns.catplot`): Analyze how the choice of orbit influences launch success or failure across different flight missions.
- **Orbit vs. PayloadMass** (`sns.catplot`): Investigate the relationship between payload mass and orbital destinations in relation to launch outcomes.
- **Yearly Success Rate Trend** (`sns.scatterplot`): Track trends in launch success rates over the years to understand historical performance patterns.

GitHub URL of your completed EDA with data visualization notebook: jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Displayed names of unique launch sites involved in space missions.
- Retrieved records where launch sites start with 'CCA' to examine specific launch details.
- Calculated the total payload mass carried by boosters.
- Determined the average payload mass carried by booster version F9 v1.1.
- Identified the date of the first successful landing outcome achieved on a ground pad.
- Listed booster names that successfully landed on a drone ship with payload mass between 4000 and 6000 kg.
- Counted total numbers of successful and failed mission outcomes.
- Identified booster versions that carried the maximum payload mass.
- Ranked landing outcomes in descending order based on frequency.

GitHub URL of completed EDA with SQL notebook: jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Various map objects were added to a Folium map to visualize SpaceX launch sites and their success rates:
- **Markers:** Pinpointed exact locations of each launch site.
- **Circles:** Represented payload range, with size indicating capacity.
- **Lines:** Connected launch sites to landing zones, illustrating flight paths.
- These objects provided an intuitive and interactive way to explore the data, showing the geographical context, payload capacities, and the relationship between launch and landing locations.
- GitHub URL of completed interactive map: lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- **Launch Site Drop-down:** Filters data by specific launch sites or includes all sites.
- **Success Pie Chart:** Shows the percentage of successful launches, adjusting to the selected site.
- **Payload Range Slider:** Allows users to select a payload mass range for in-depth analysis.
- **Success-Payload Scatter Chart:** Illustrates the relationship between payload mass and launch success, updating based on the chosen site and payload range.

GitHub URL of completed Plotly Dash lab: spacex_dash_app.py

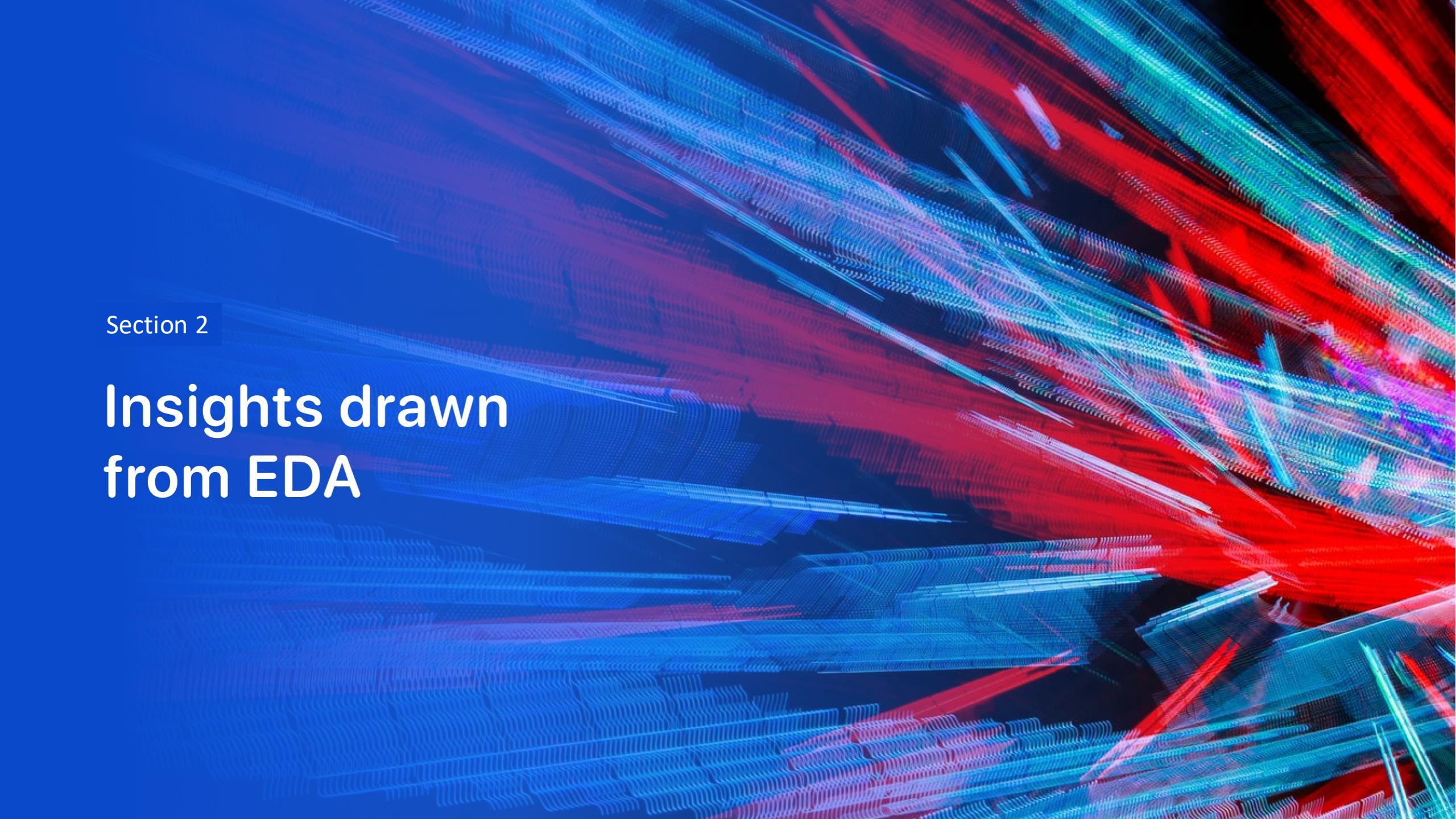
Predictive Analysis (Classification)

- **Data Preparation:**
  - Standardized the data using StandardScaler.
  - Split the data into training and test sets (80/20 split).
- **Model Training and Evaluation:**
  - Trained multiple models (Logistic Regression, SVM, Decision Tree, KNN) using GridSearchCV with cross-validation (cv=10) to find the best parameters.
- **Model Selection:**
  - Evaluated models on the test set.
  - Selected Logistic Regression as a main model due to its interpretability and similar accuracy to other models: accuracy = 0.83.

- GitHub URL of completed predictive analysis lab:
  SpaceX_Machine_Learning_Prediction.ipynb

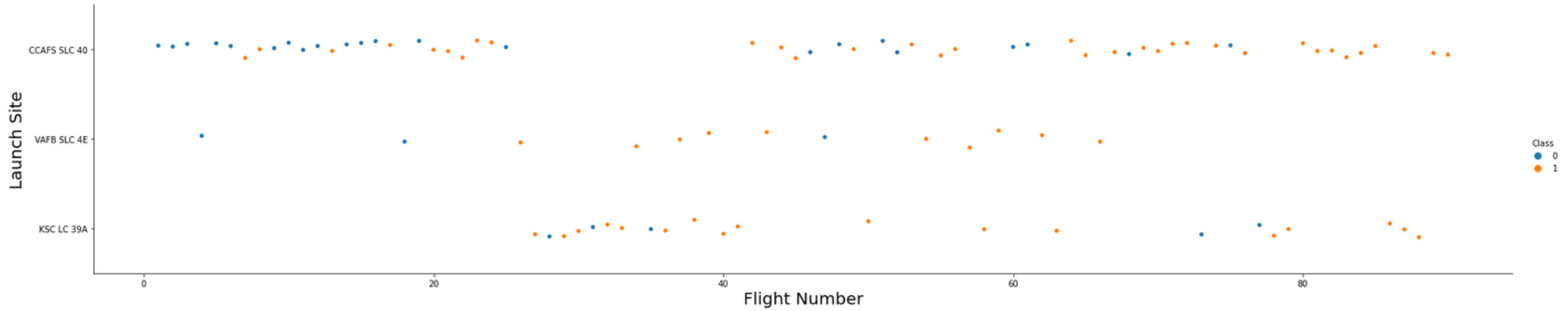
Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

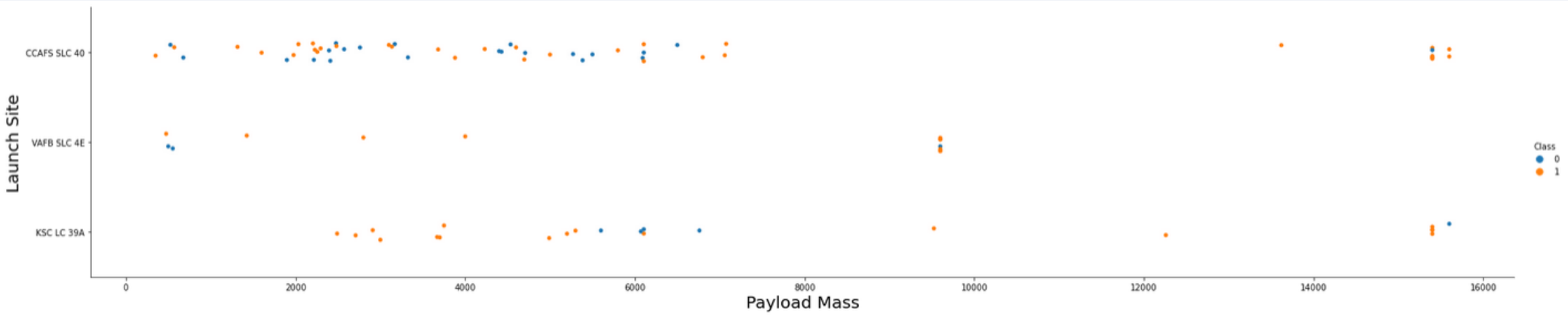
- Scatter plot of Flight Number vs. Launch Site



1. Succes rate for each launch site increases with flight number / time.
2. Launch site "CCAFS SLC 40" has success rate 100% for flight numbers more than 80.

Payload vs. Launch Site

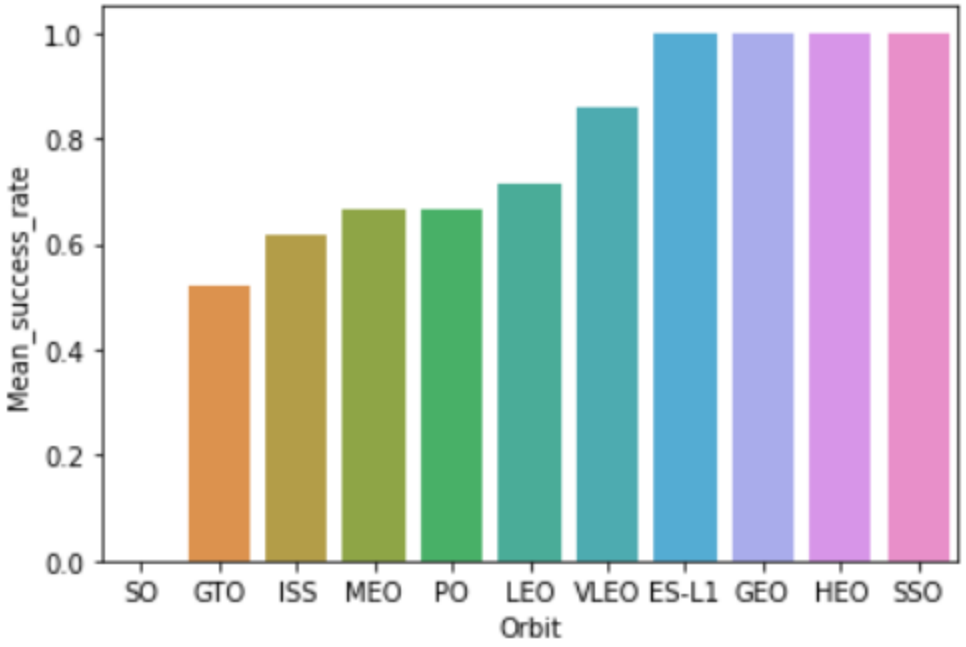
- Scatter plot of Payload Mass vs. Launch Site



For VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

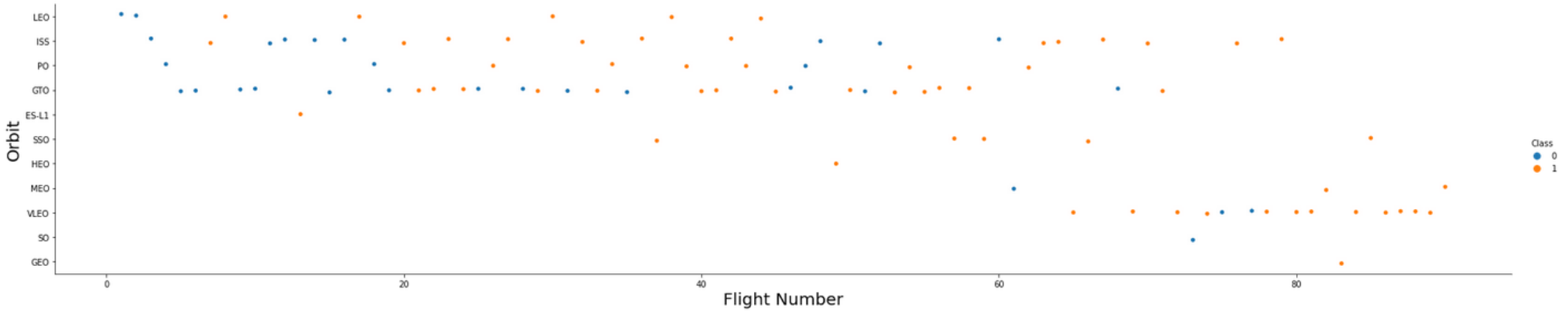
- Bar chart for the success rate of each orbit type



Top 4 orbits with high sucess rate: SSO, HEO, GEO, ES-L1

Flight Number vs. Orbit Type

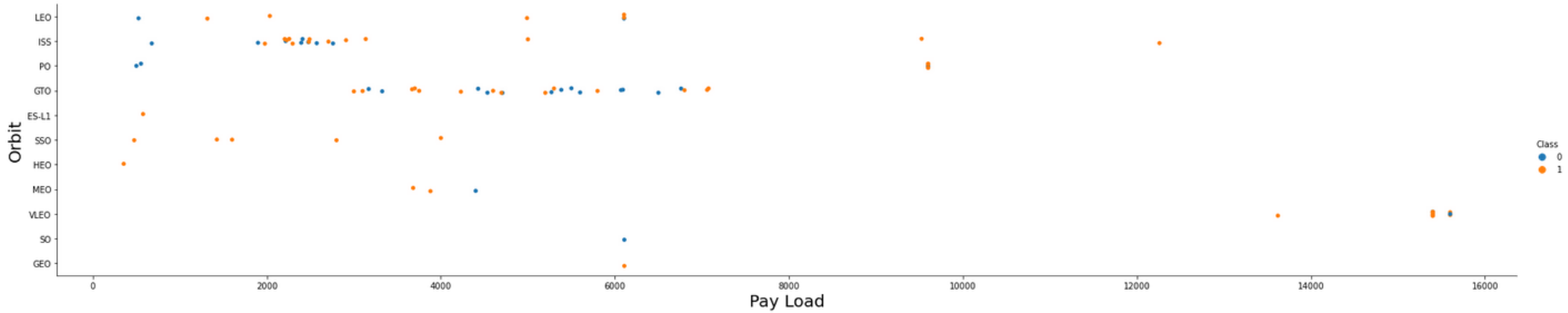
- Scatter plot of Flight number vs. Orbit type



1. In the LEO orbit the Success appears related to the number of flights
2. There seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

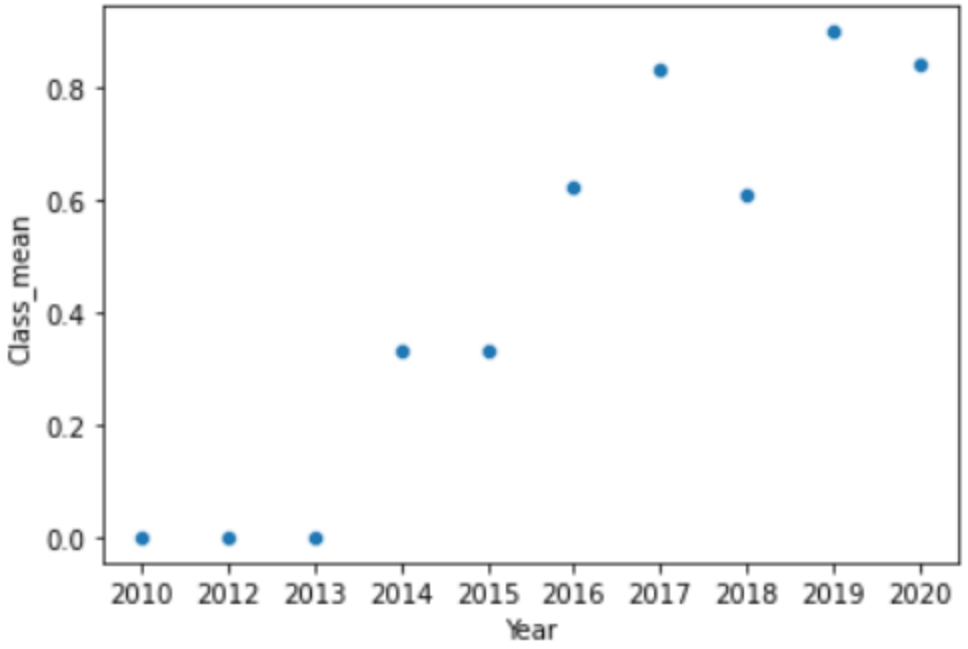
- Scatter plot of Payload vs. Orbit type



1. With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
2. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- Line chart of yearly average success rate



The sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- The names of the unique launch sites

Display the names of the unique launch sites in the space mission

```
%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'

- 5 records where launch sites begin with `CCA`

Display 5 records where launch sites begin with the string 'CCA'

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

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

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

```
%sql select sum(payload_mass__kg_) as sum from SPACEXTBL where customer like 'NASA (CRS)'
```

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

sum
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
%sql select avg(payload_mass__kg_) as average from SPACEXTBL where booster_version like 'F9 v1.1%'
```

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

average
2534.6666666666665

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad

```
%sql select min(date) as date from SPACEXTBL where mission_outcome like 'Success'
```

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

date
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql select booster_version from SPACEXTBL where (mission_outcome like 'Success') \
     and (payload_mass__kg_ between 4000 and 6000) \
     and (Landing_Outcome like '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

- The total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(*) as count from SPACEXTBL group by mission_outcome order by mission_outcome
```

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

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql select booster_version from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)
```

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

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql select substr(DATE, 6,2) as month, Landing_Outcome, booster_version, launch_site from SPACEXTBL where date like '2015%' and Landing_Outco
```

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

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

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

- 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 landing_outcome, count(*) as count from SPACEXTBL where date >= '2010-06-04' and date <= '2017-03-20' group by landing_outcom
```

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

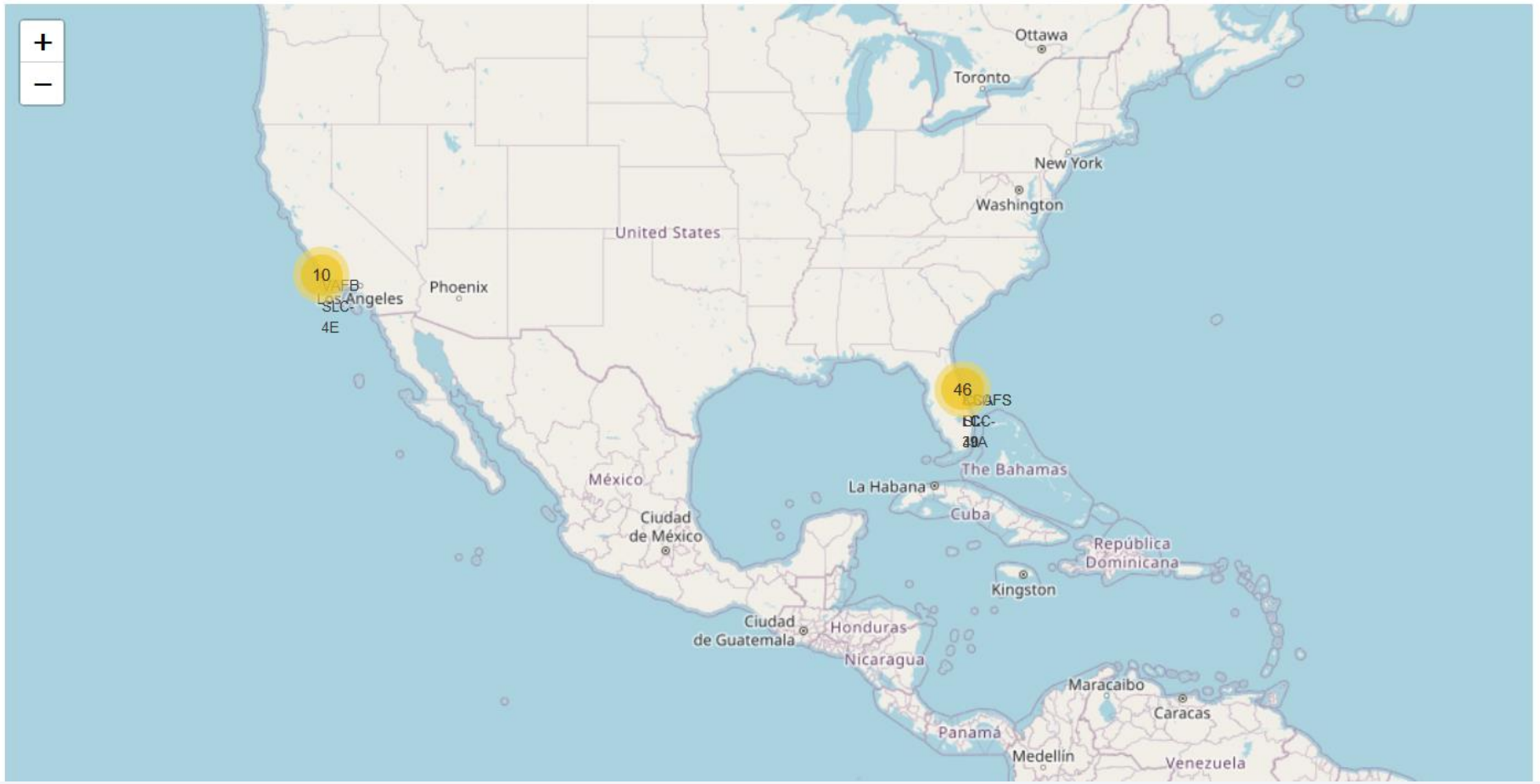
Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

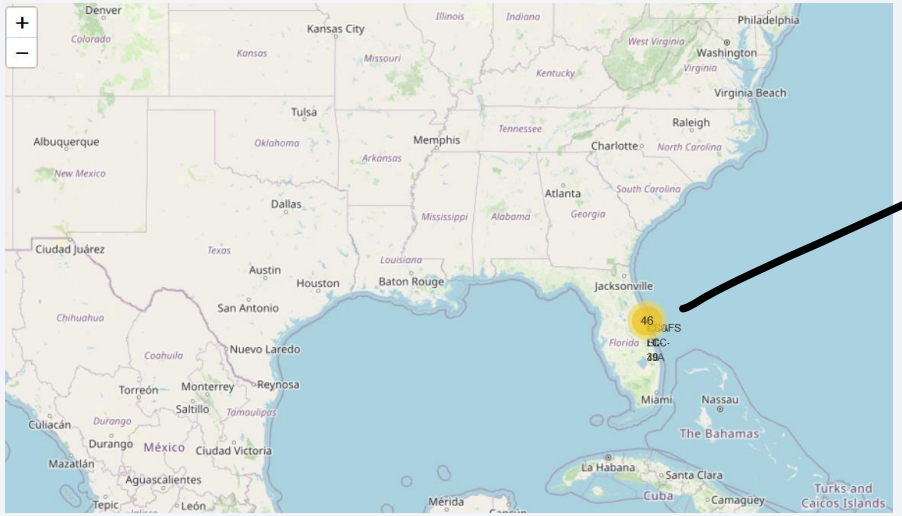
Launch Sites

[13]:

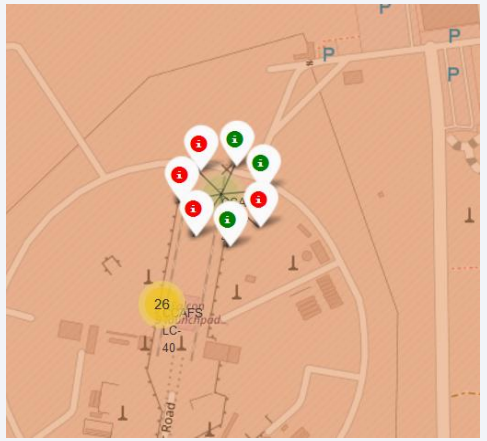


Launch Outcomes

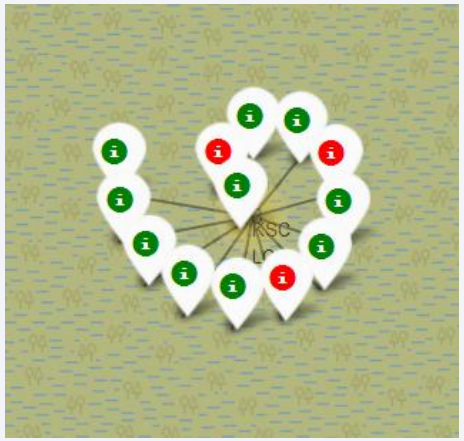
- Green - Succes
- Red - Failure



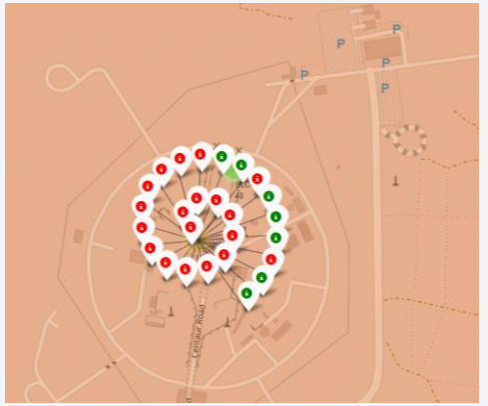
CCAFS SLC-40



KSC LC-39A

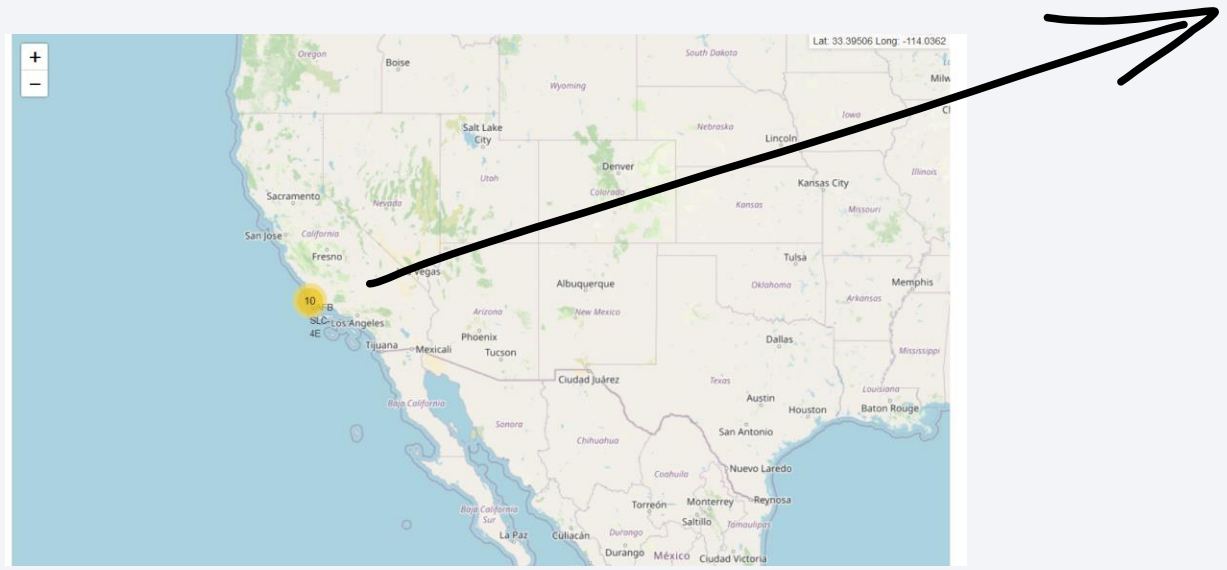


CCAFS-LC-40

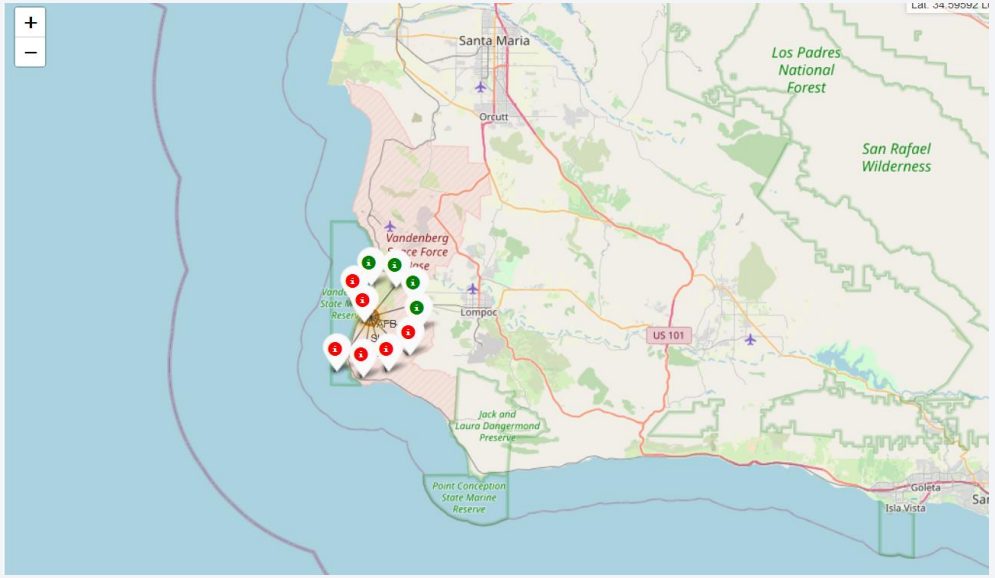


Launch Outcomes

- Green - Succes
- Red - Failure

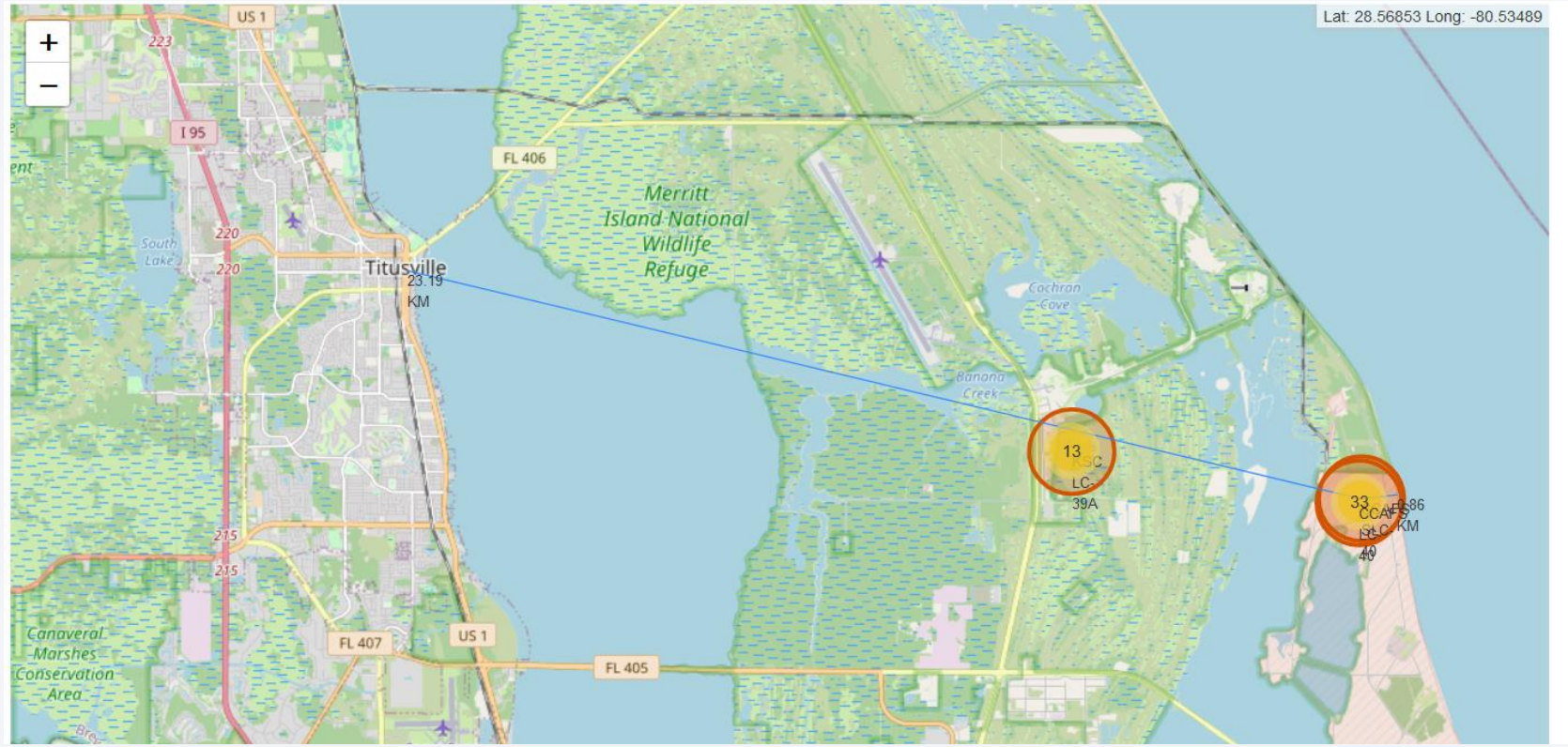


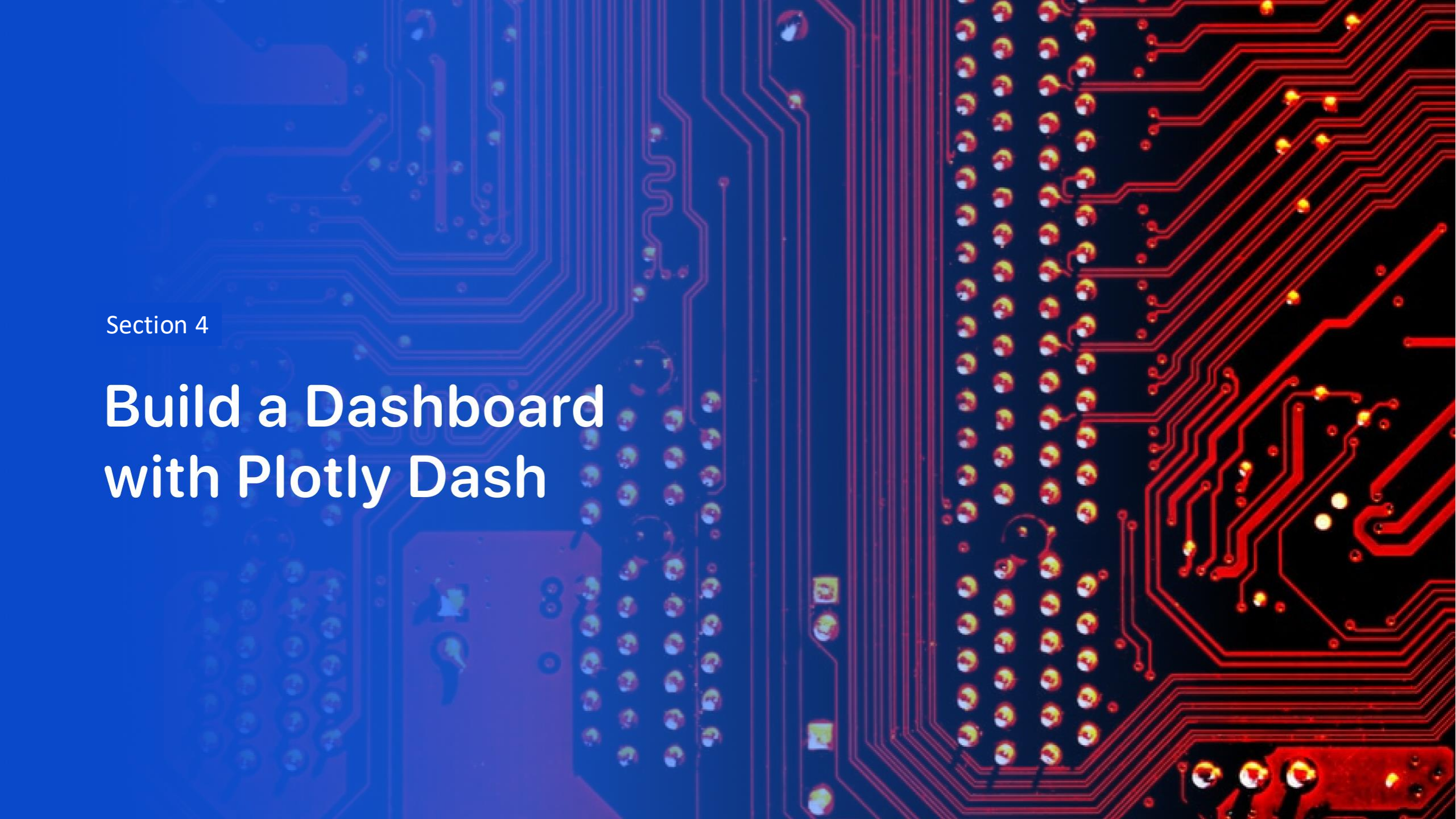
CCAFS-LC-40



Railway Proximity

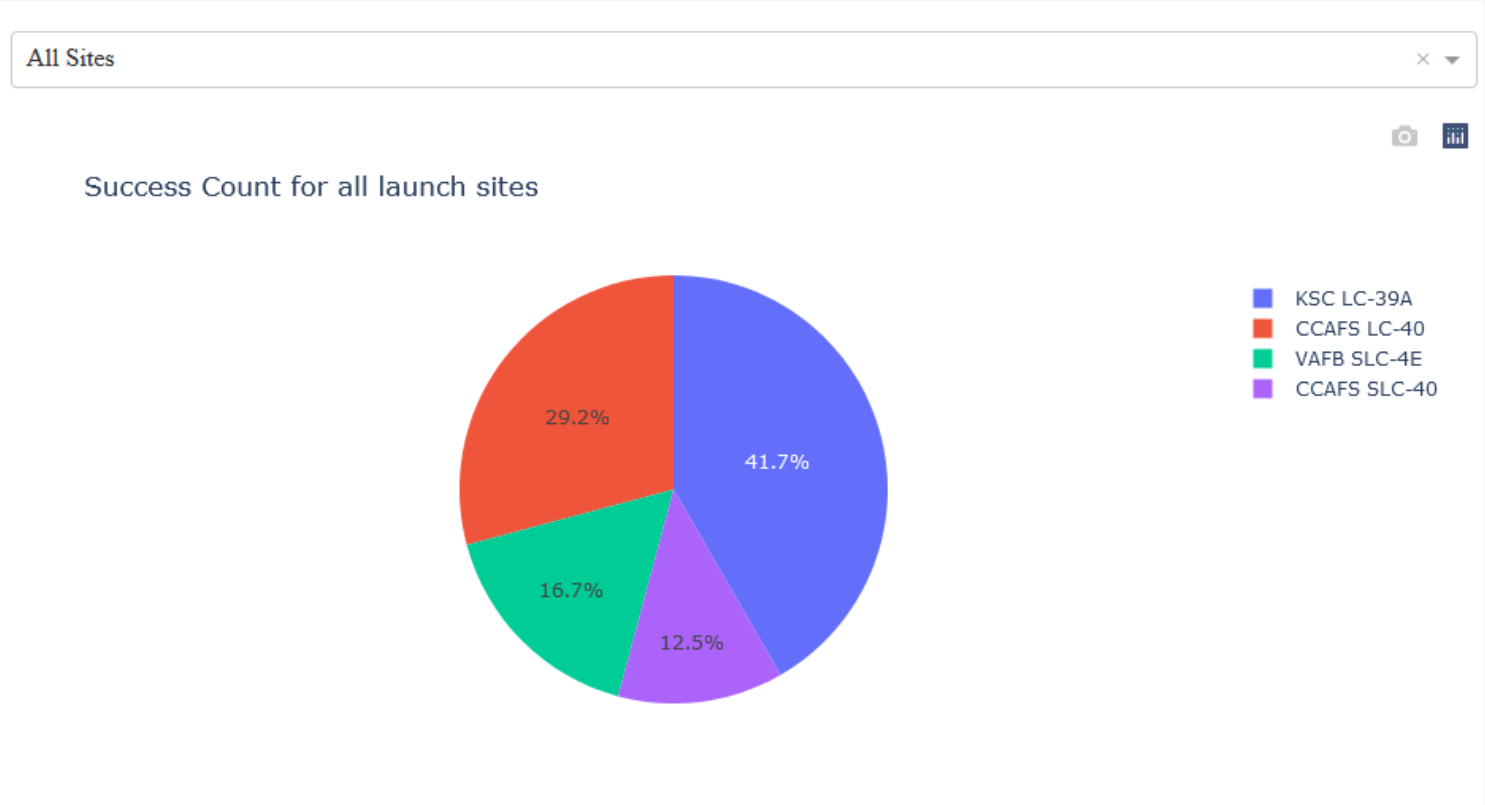
- Launch site to proximity to railway




Section 4
Build a Dashboard with Plotly Dash

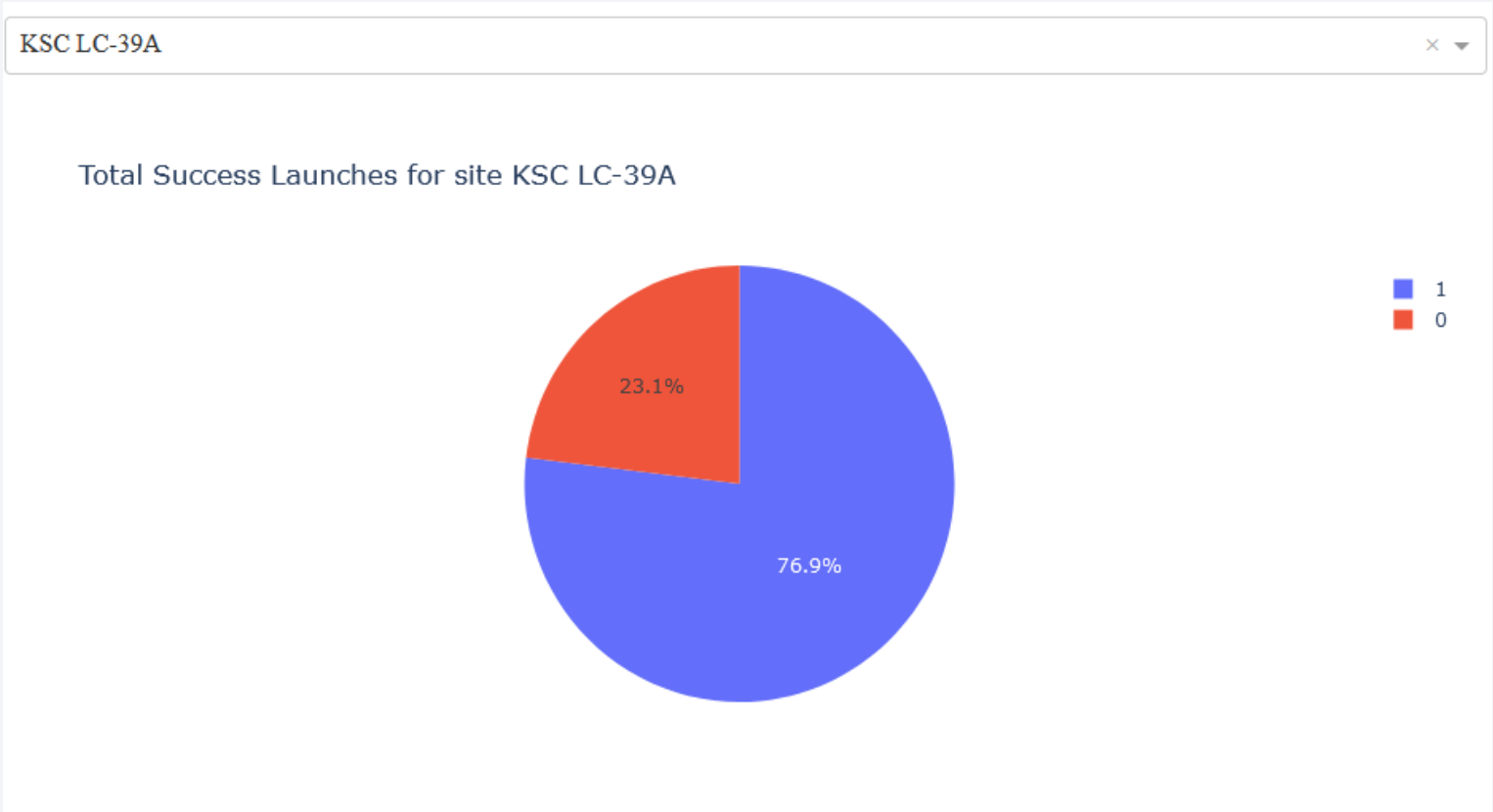
Dashboard. Launch success count for all sites

- Launch success count for all sites, in a piechart



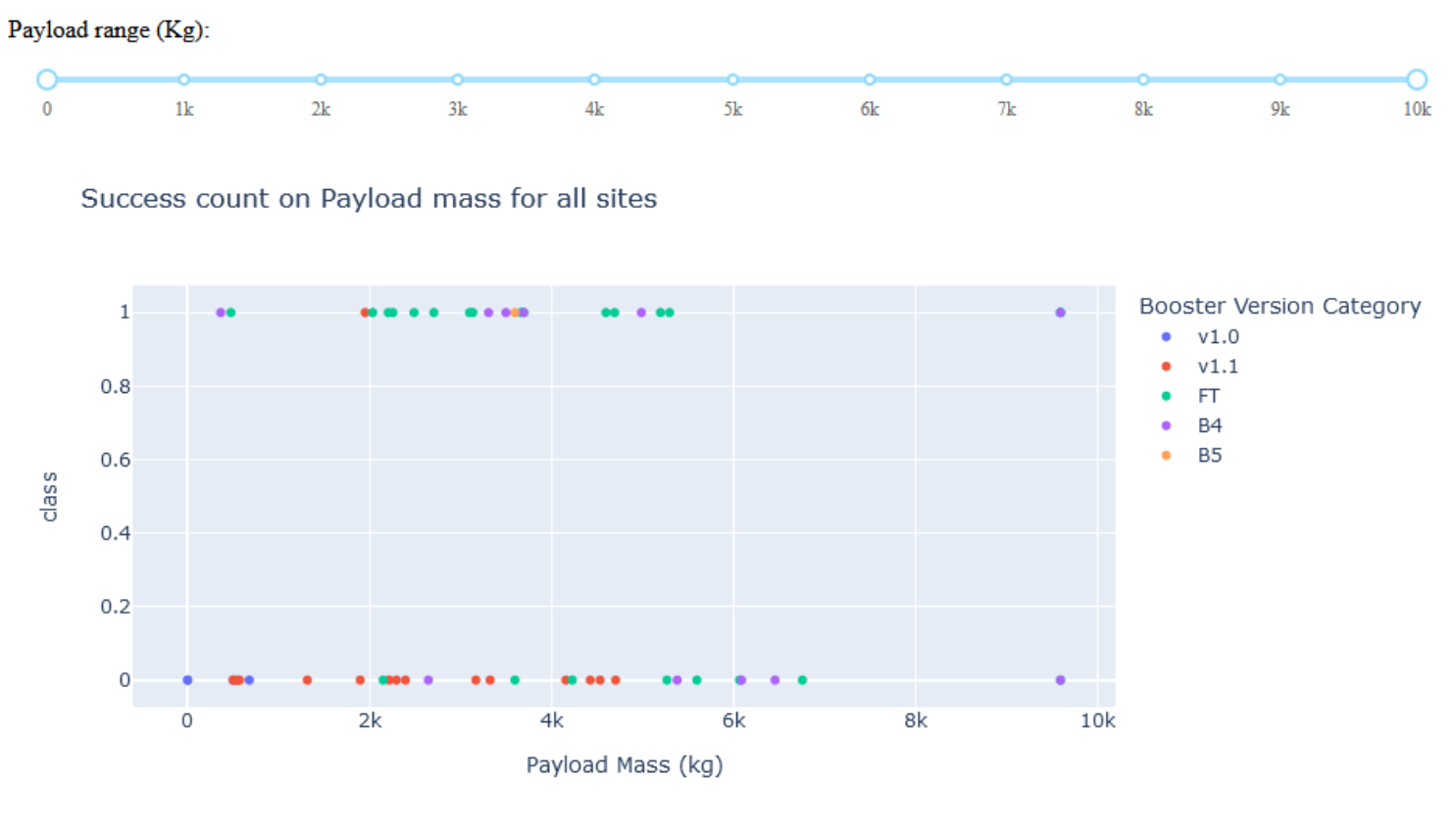
Dashboard. Launch site with highest launch success ratio

- Piechart for the launch site with highest launch success ratio



Dashboard. Payload vs. Launch Outcome scatter plot

- Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

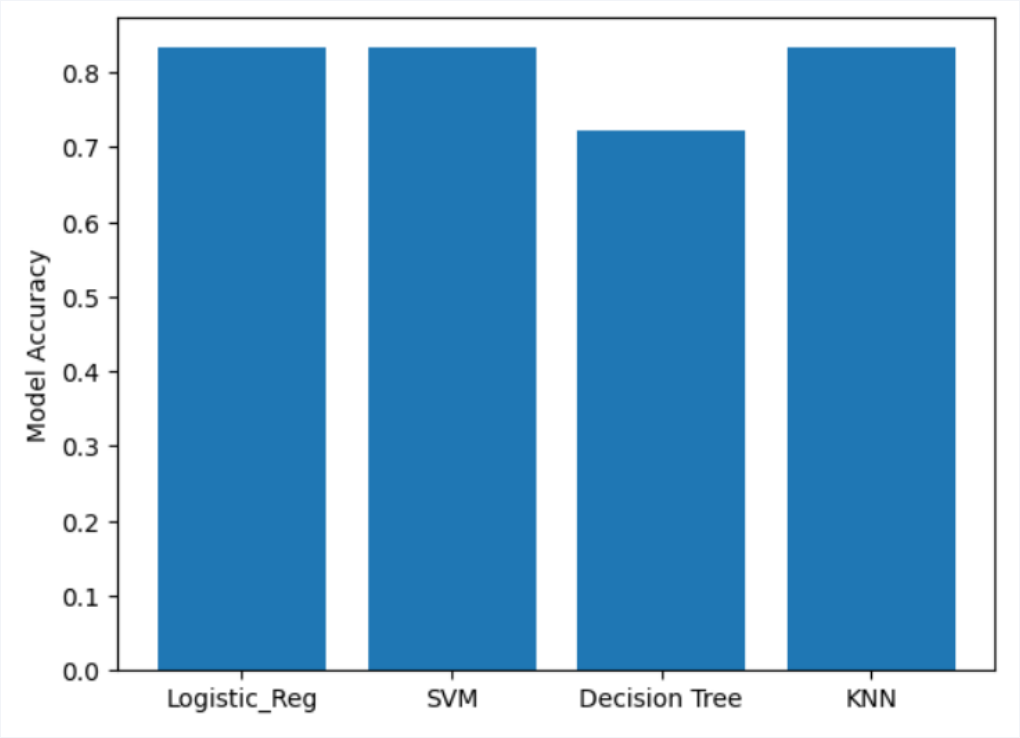


Section 5

Predictive Analysis (Classification)

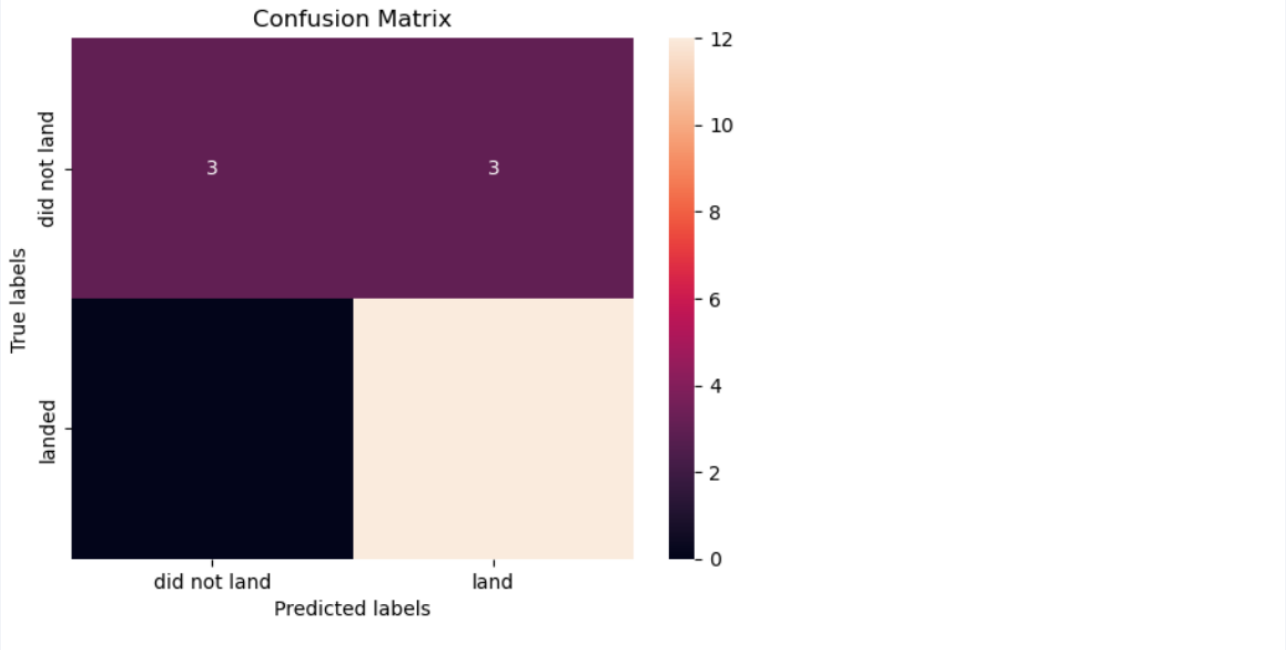
Classification Accuracy

- Accuracy bar chart for all built classification
- Models, with highest accuracy: Logistic regression, SVM, Knn.
- As a base model Logistic regression was chosen due to its interpretability



Confusion Matrix

- Confusion matrix of the best performing model – Logistic regression



Logistic regression can distinguish between the different classes. The major problem is false positives

Conclusions

- Success rates have been increasing steadily since 2013, peaking in 2020.
- Mission success is influenced by launch site, orbit, and number of previous launches.
- GEO, HEO, SSO, and ES-L1 orbits have the highest success rates.
- Lighter payloads generally perform better than heavier ones.
- The superiority of some launch sites remains unexplained; additional data could provide insights.
- Three classification models had the best accuracy of 0.83: logistic regression, SVM, Knn.
- Reducing false positives is crucial for enhancing reliability.
- Logistic regression was chosen for its interpretability.

Thank you!