IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bohdan
05 December 2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusions

Executive Summary

This project aims to study the factors affecting the successful landing of the rocket's first stage.

To achieve this goal:

- Data were collected using Requests to SpaceX REST API and Web scraping of related Wikipedia pages
- Data wrangling was performed by filtering the data, handling missing values, feature engineering to prepare the data for analysis and modeling
- Exploratory data analysis (EDA) was performed using SQL and data visualization techniques
- Folium and Plotly Dash were used to build an interactive map and dashboard for interactive visual analytics
- Prediction the landing outcomes were made using four classification models and their cross-validation

Findings show that:

- Success rate increases over time and number of launches
- Launch site KSC LC-39A has the highest number (41.2%) of successful launches
- All launch sites are located close to the equator and a coastline
- Launches with heavy payload (over 9000 kg) are more successful
- DecisionTreeClassifier is a good model for predicting the success of the landing in this project

Introduction

- Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost is 165 million dollars each. The price difference is caused by ability of Space X to reuse the first stage. Therefore, determining if the first stage will land successfully and can be reused, one 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. In this project, a machine learning pipeline will be created to predict if the first stage will land given the collected relevant data.
- Problems to solve are:
    - How factors like payload mass, launch site, number of flights, type of orbit or booster version affect the successful landing of the first stage
    - What is the best predictive model for this project

Section 1

Methodology

Methodology

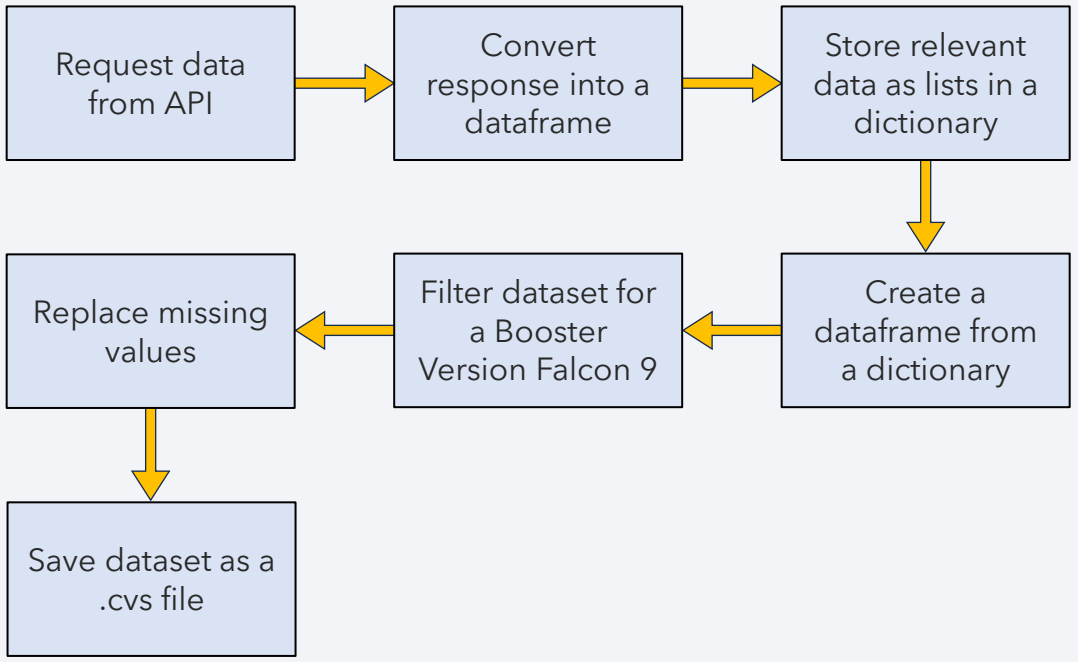
- Data were collected using two approaches:
  - Requests to SpaceX REST API
  - Web scraping of related Wikipedia pages
- Data wrangling was performed by filtering the data, handling missing values, feature engineering (one hot encoding) to prepare the data for analysis and modeling
- Perform Exploratory Data Analysis using SQL and data visualization techniques
- Perform interactive visual data analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Perform standardization of dataset and predict the landing outcomes using four classification models

Data Collection

- Required data were gathered using two methods:
  - Requests to SpaceX REST API with API endpoints starting with *api.spacexdata.com/v4/*
  - Web scraping from a Wikipedia page “List of Falcon 9 and Falcon Heavy launches” at *https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches*

Data Collection – SpaceX API

- Request data SpaceX REST API
- Convert response into a dataframe using pd.json_normalize()
- Store only relevant data as lists for creating a dictionary and
- Create a dataframe from a dictionary
- Filter data for a Booster Version Falcon 9
- Replace missing values
- Save dataset as a .cvs file

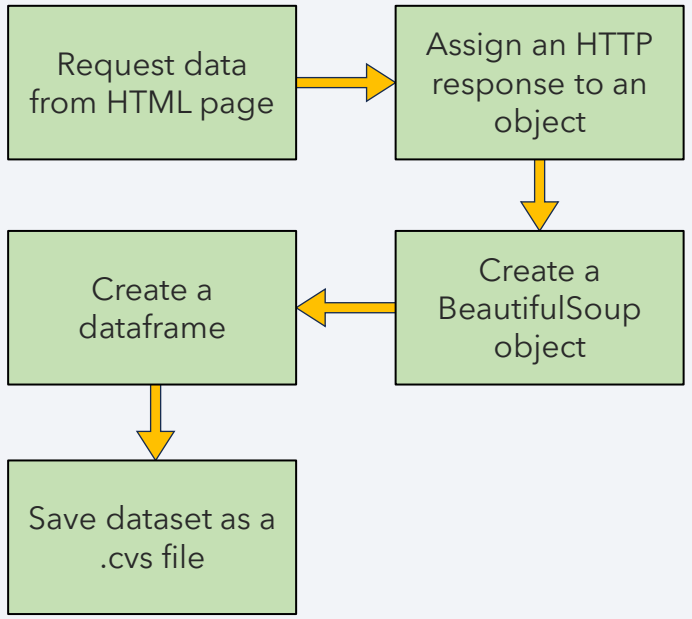


GitHub URL:

https://github.com/oneuser24/IBM-Data-Science/blob/main/course-10-labs-1-data-collection-api.ipynb

Data Collection – Web Scraping

- Perform an HTTP GET method to request the Falcon9 Launch HTML page
- Assign an HTTP response to an object
- Create a BeautifulSoup object from the HTML response text content
- Create a data frame by parsing the HTML tables
- Save dataset as a .cvs file



GitHub URL:

https://github.com/oneuser24/IBM-Data-Science/blob/main/course-10-labs-2-data-collection-webscraping.ipynb

Data Wrangling

- Perform Exploratory Data Analysis (EDA) by calculating:
  - Number of launches on each site
  - Number of launches to each orbit
  - Number of landing outcomes
- Create an Outcome column
- Create a landing outcome label from the Outcome column

GitHub URL:

https://github.com/oneuser24/IBM-Data-Science/blob/main/course-10-labs-3-Data wrangling.ipynb

EDA with Data Visualization

- Perform Exploratory Data Analysis and visualized relationship between :
  - Payload mass and Flight Number
  - Launch Site and Flight Number
  - Launch Site and Payload mass
  - Success Rate and each Orbit type
  - Orbit type and respectively Launch Site and Flight Number
  - Launch success yearly trend

GitHub URL:

https://github.com/oneuser24/IBM-Data-Science/blob/main/course-10-labs-5-eda-data-vizualisation.ipynb

EDA with SQL

- Executed SQL queries to understand the SpaceX dataset and found
  - Names of the unique launch sites in the space mission
  - 5 records where launch sites begin with the string 'CCA'
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first successful landing outcome in ground pad was achieved
  - Names of the boosters which have success in drone ship and have payload mass between 4000 kg and 6000 kg
  - Total number of successful and failure mission outcomes
  - All booster versions that have carried the maximum payload mass
  - Records with the month names in year 2015
  - Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20

GitHub URL:

https://github.com/oneuser24/IBM-Data-Science/blob/main/course-10-labs-4-eda-sql-sqllite.ipynb

Build an Interactive Map with Folium

- To understand the operation of launch sites, we added to Folium map
  - Circles marking and encircling Launch sites and corresponding markers
  - Marker clusters showing the success/failed launches for each site on the map
  - Calculated and marked on the map distances to proximities for a launch site CCAFS-SLC-40
- Using maps, determined that launch sites are usually located close to the highways, railways and coast lines and far from the cities

GitHub URL:

https://github.com/oneuser24/IBM-Data-Science/blob/main/course-10-labs-6_interact_visual_analytics_folium.ipynb

Build a Dashboard with Plotly Dash

- To understand the success rate of each launch site we added an interactive dashboard with
  - Pie plot showing the distribution of successful launches
  - Scatter plot showing relationship between Success of the mission and Payload

GitHub URL:

https://github.com/oneuser24/IBM-Data-Science/blob/main/course-10-labs-7-launch_sites_dash_app.py

Predictive Analysis (Classification)

- Load data
- Perform standardization of data using a Scikit-learn transformer StandardScaler()
- Split input and target data into training and test data
- Fit data with LogisticRegression, Support Vector Machine, DecisionTreeClassifier and k-nearest-neighbors models
- Use GridSearchCV technique for finding the optimal parameter values from a given set of parameters in a grid

GitHub URL:

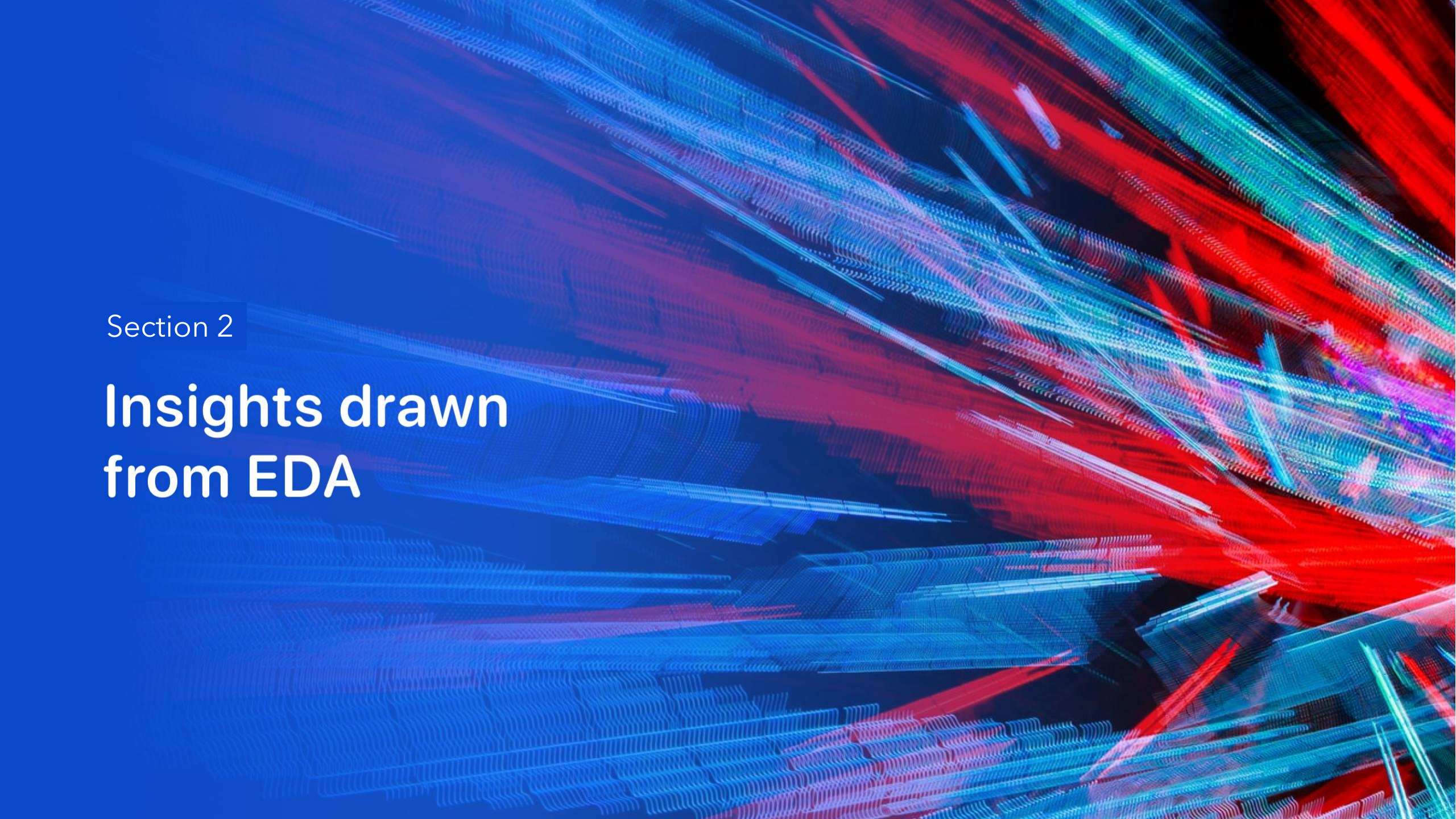
https://github.com/oneuser24/IBM-Data-Science/blob/main/course-10-labs-8-Machine%20Learning%20Prediction.ipynb

Results

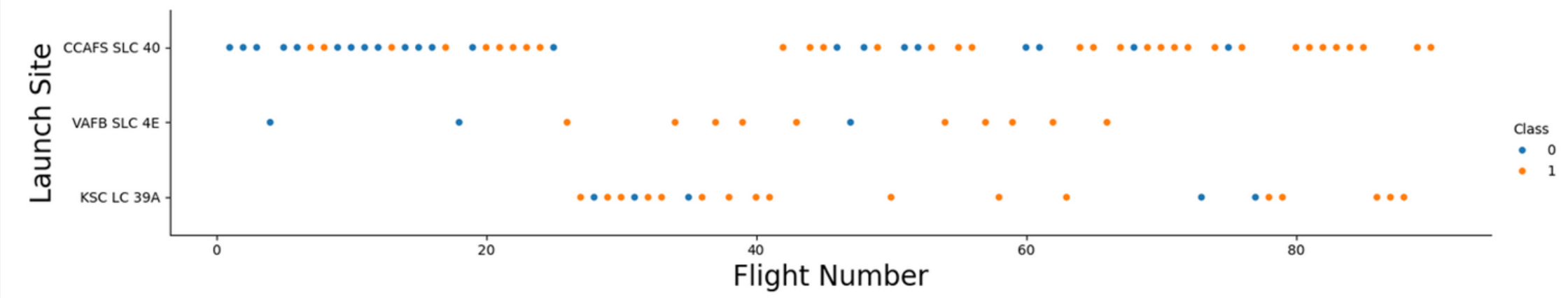
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

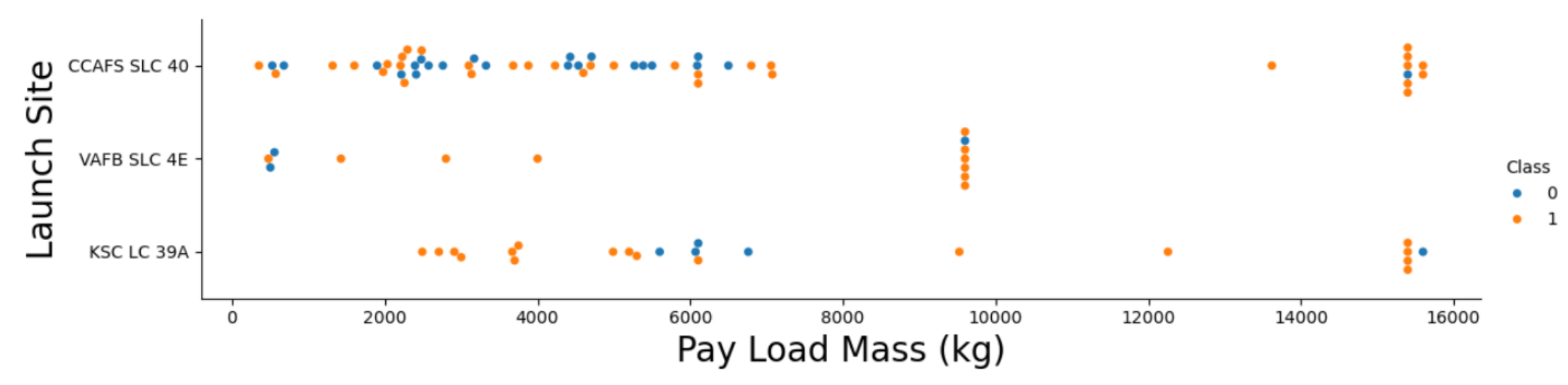


Launch Site vs. Flight Number



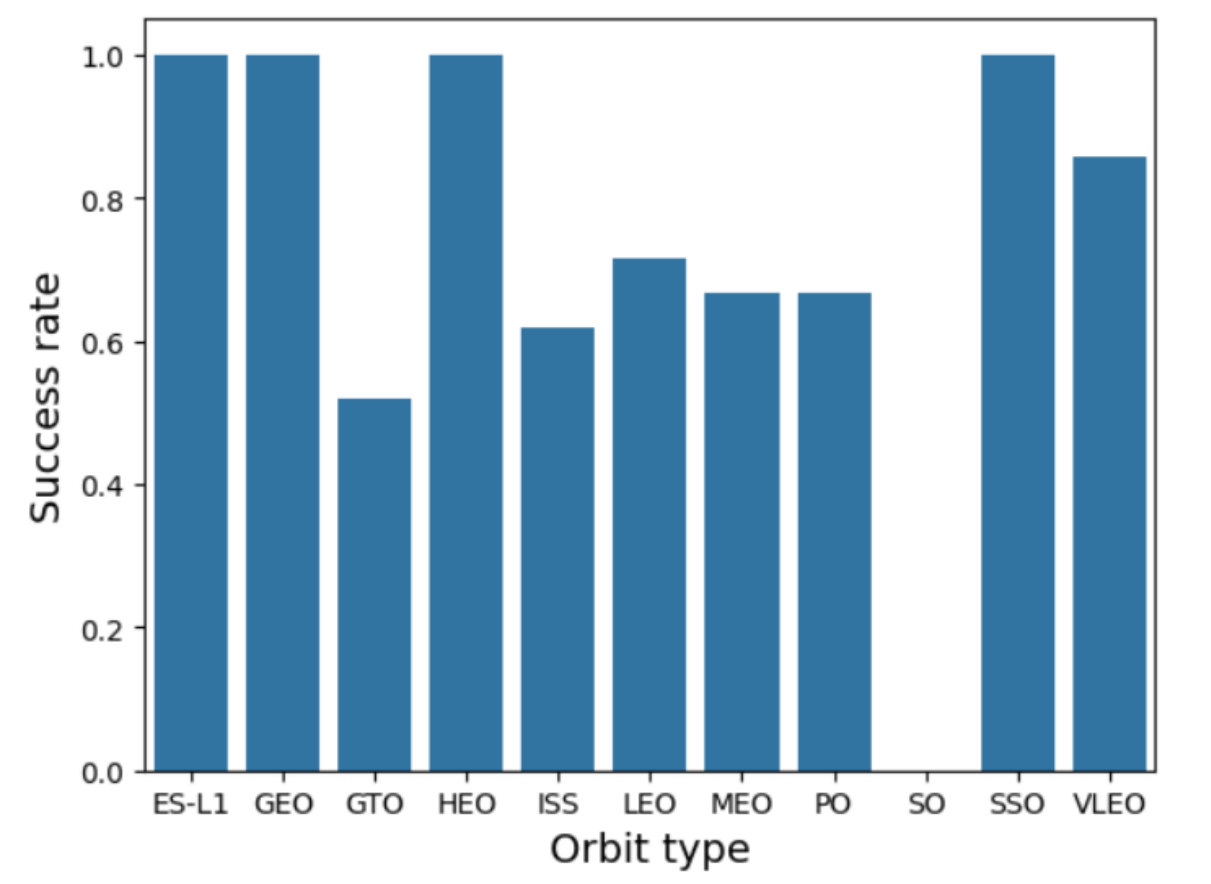
- CCAFS SLC 40 Launch Site has the most of flights
- VAFB SLC 4E Launch site has a fewest number of flights
- All three launch sites have successful and failed launches

Payload vs. Launch Site



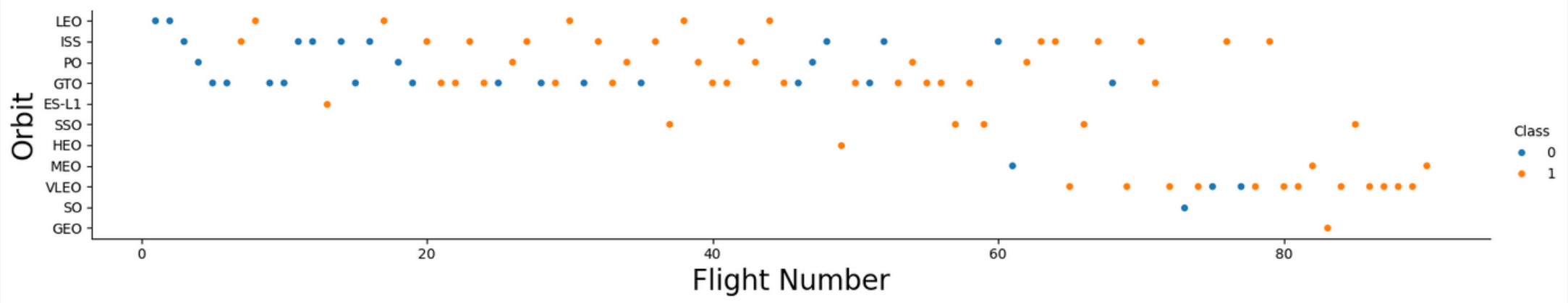
- There are no rockets launched for heavy payload mass (greater than 10000 kg) at VAFB-SLC launch site
- There are less launches for payloads over 7000 kg at all sites
- Heavy (over 7000 kg) payload launches have a better success rate

Success Rate vs. Orbit Type



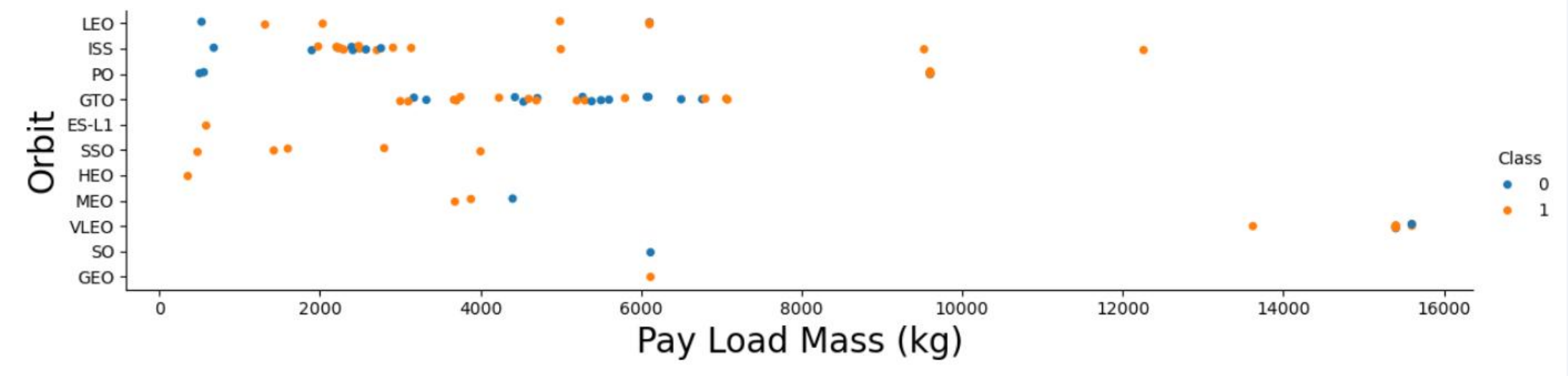
ES-L1, GEO, HEO and SSO type of orbits have the highest success rate

Flight Number vs. Orbit Type



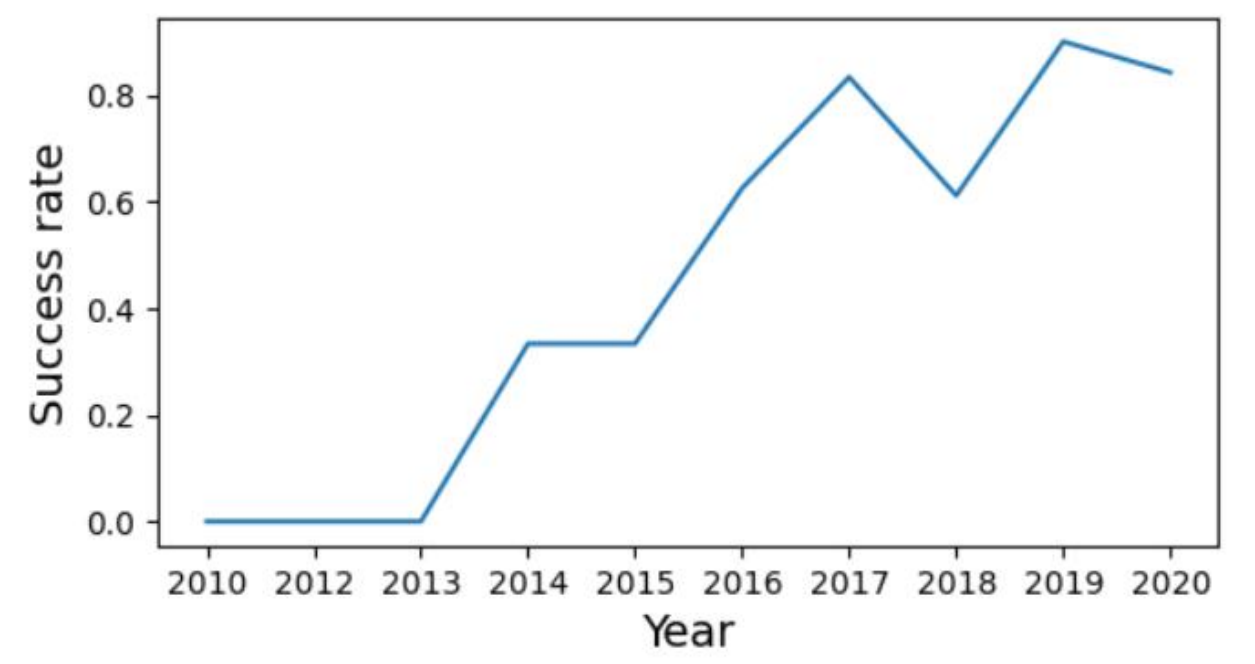
In the LEO orbit, success seems to be related to the number of flights but in the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS orbits. However, for GTO orbit, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



Success rate kept increasing since 2013 till 2020.

All Launch Site Names

```
%sql select distinct "Launch_Site" from SPACEXTABLE
```

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

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

```
%sql select "Launch_Site", count("Launch_Site") from SPACEXTABLE group by "Launch_Site" order by count("Launch_Site") desc
```

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

Launch_Site	count("Launch_Site")
CCAFS SLC-40	34
CCAFS LC-40	26
KSC LC-39A	25
VAFB SLC-4E	16

- A keyword **DISTINCT** was used to get unique names of launch sites
- There are four unique launch sites

Launch Site Names Begin with 'CCA'

```
[24]: %sql select * from SPACEXTABLE where "Launch_Site" like "CCA%" limit 5
```

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

[24]:

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

- A clause **WHERE** and a keyword **LIKE** was used to get names beginning with 'CCA'
- A keyword **LIMIT** was used to display only 5 records

Total Payload Mass

```
[28]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
[28]: sum(PAYLOAD_MASS__KG_)
45596
```

- An aggregate function **SUM()** was used to get a total payload mass

Average Payload Mass by F9 v1.1

Task 4

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

```
[37]: %sql select avg(PAYLOAD_MASS__KG_) as average_payload_mass from SPACEXTABLE where "Booster_Version" = 'F9 v1.1'
```

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

[37]:

average_payload_mass
2928.4

An aggregate function **AVG()** and clause **WHERE** were used to display average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
[42]: %sql select min(Date) from SPACEXTABLE where "Landing_Outcome" = 'Success (ground pad)'
 * sqlite:///my_data1.db
Done.
[42]: min(Date)
2015-12-22
```

MIN() function and a clause **WHERE** were used to list the date when the first successful landing outcome in ground pad was achieved

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
[45]: %sql select "Booster_Version" from SPACEXTABLE where "Landing_Outcome" = 'Success (drone ship)' and "PAYLOAD_MASS__KG_" between 4000 ar
```

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

[45]:

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

A clause **WHERE** and operators **AND** and **BETWEEN** were used to list the names of the boosters which have success in drone ship and have payload mass greater than 4000 kg but less than 6000 kg

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
[50]: %sql select count("Mission_Outcome") from SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

[50]:

count("Mission_Outcome")
101

COUNT() function was used to list the total number of successful and failed mission outcomes

Boosters Carried Maximum Payload

```
[60]: %%sql

select "Booster_Version", "PAYLOAD_MASS__KG_"
from SPACEXTABLE where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS__KG_") from SPACEXTABLE)
```

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

[60]:

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

A subquery with a suitable function **MAX()** was used to list all the booster versions that have carried the maximum payload mass

2015 Launch Records

```
[75]: %%sql

select substr(Date, 6, 2) as "Month", "Landing_Outcome", "Booster_Version", "Launch_Site"
from SPACEXTABLE
where "Landing_Outcome" = "Failure (drone ship)" and substr(Date, 1, 4) = '2015'
```

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

[75]:

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

A function **SUBSTR()** , clause **WHERE** and operator **AND** were used to list the records displaying the month, failure landing outcomes in drone ship, booster versions and launch site for the months in year 2015

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

Task 10

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.

[94]:
```
%%sql

select "Landing_Outcome", count("Landing_Outcome"), rank() over (order by count("Landing_Outcome") desc) as rank
from SPACEXTABLE
where "Landing_Outcome" in ('Failure (drone ship)', 'Success (ground pad)') and Date between '2010-06-04' and '2017-03-20'
group by "Landing_Outcome"
```

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

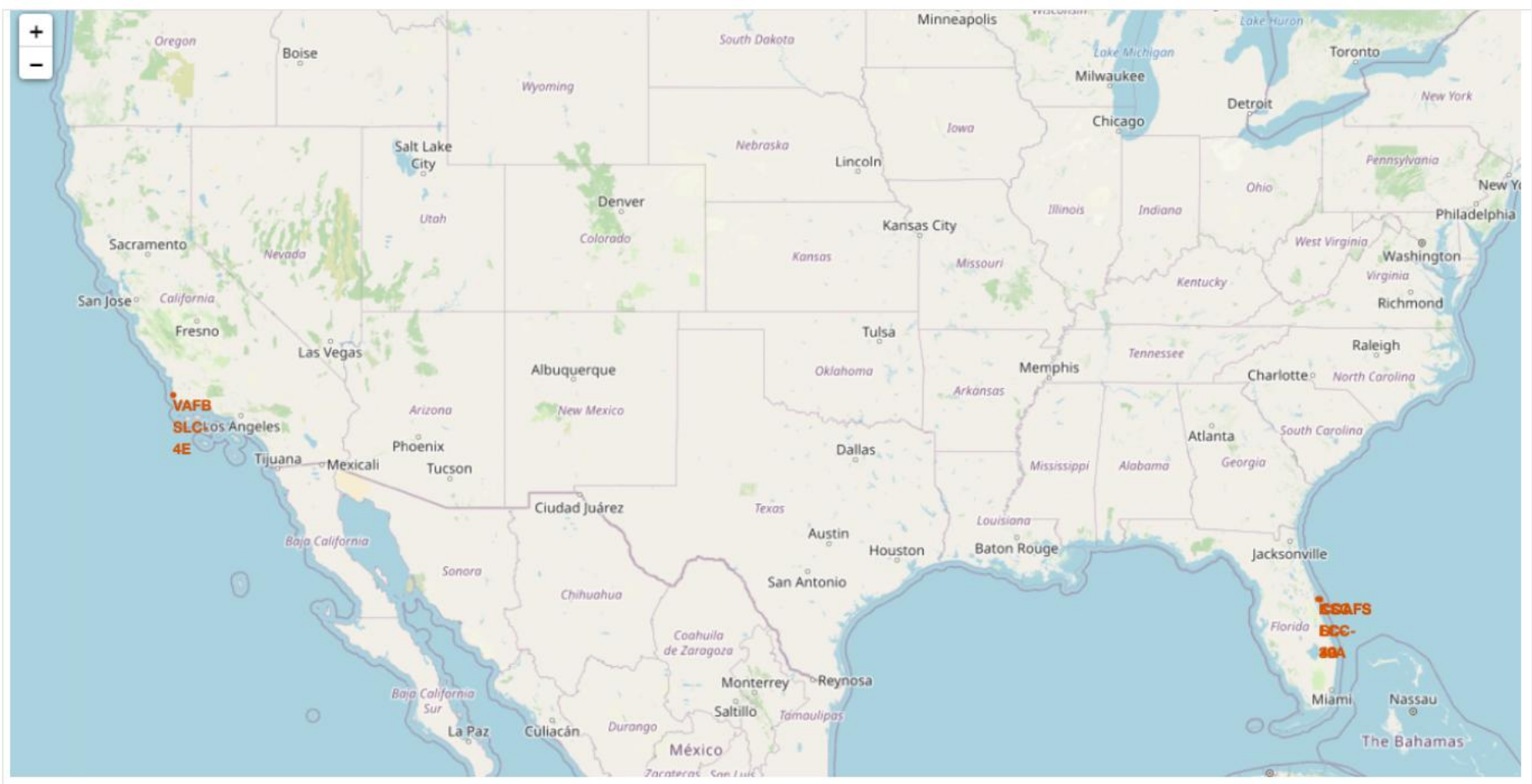
[94]:

Landing_Outcome	count("Landing_Outcome")	rank
Failure (drone ship)	5	1
Success (ground pad)	3	2

A function **RANK()** was used to rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20, in descending order

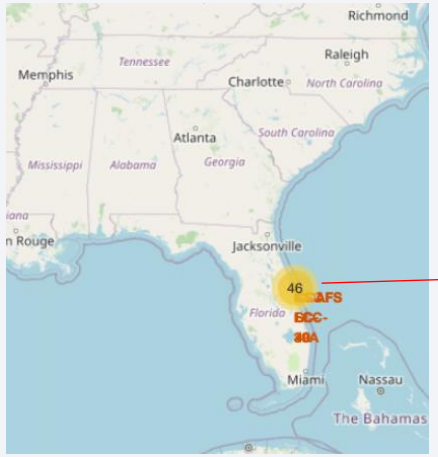

Section 3
Launch Sites
Proximities Analysis

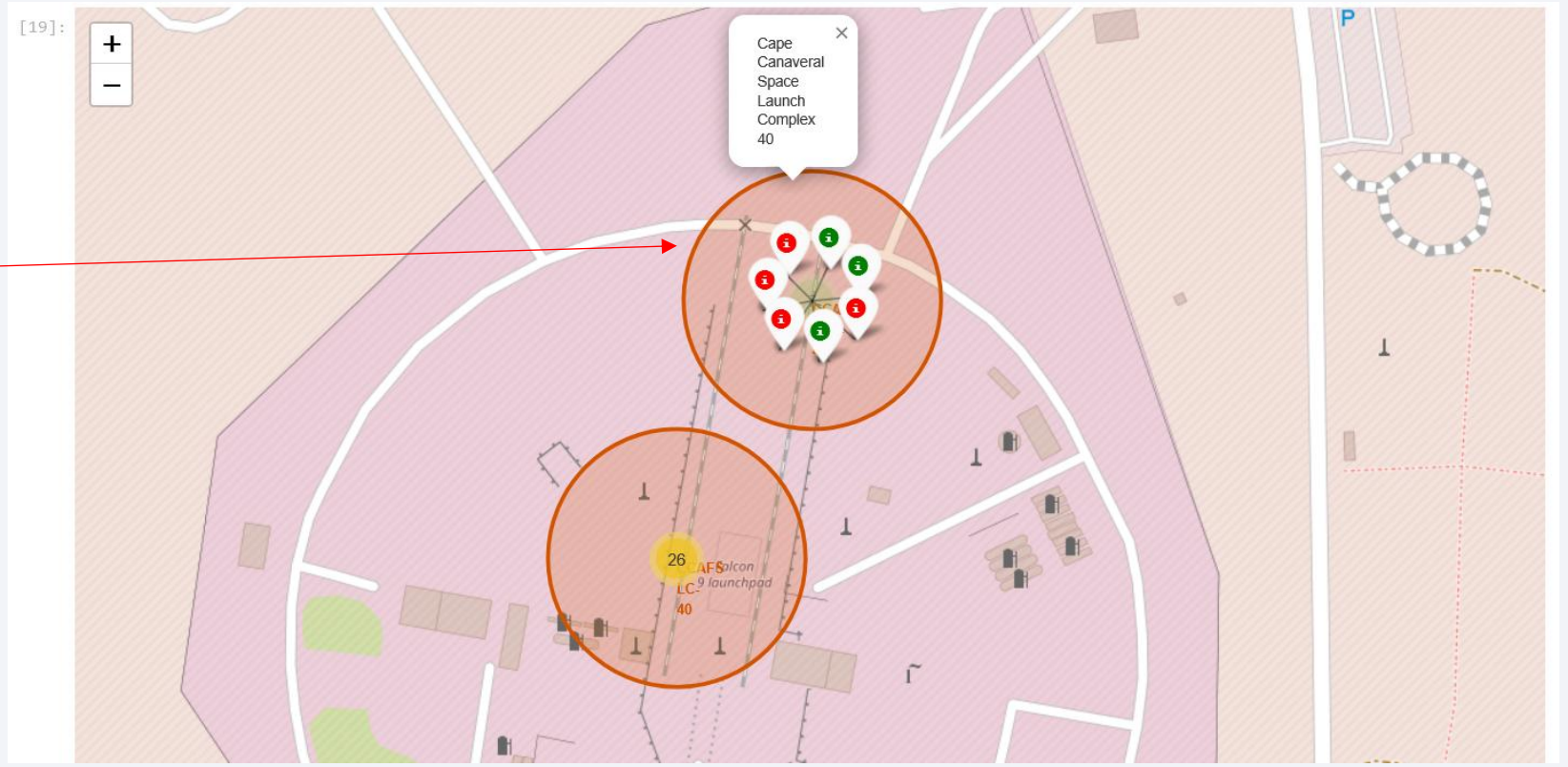
Launch Sites



- All launch sites are located in the United States
- There are three launch sites on east coast in Florida and one launch site on the west coast
- Launch sites in Florida are closer to the equator

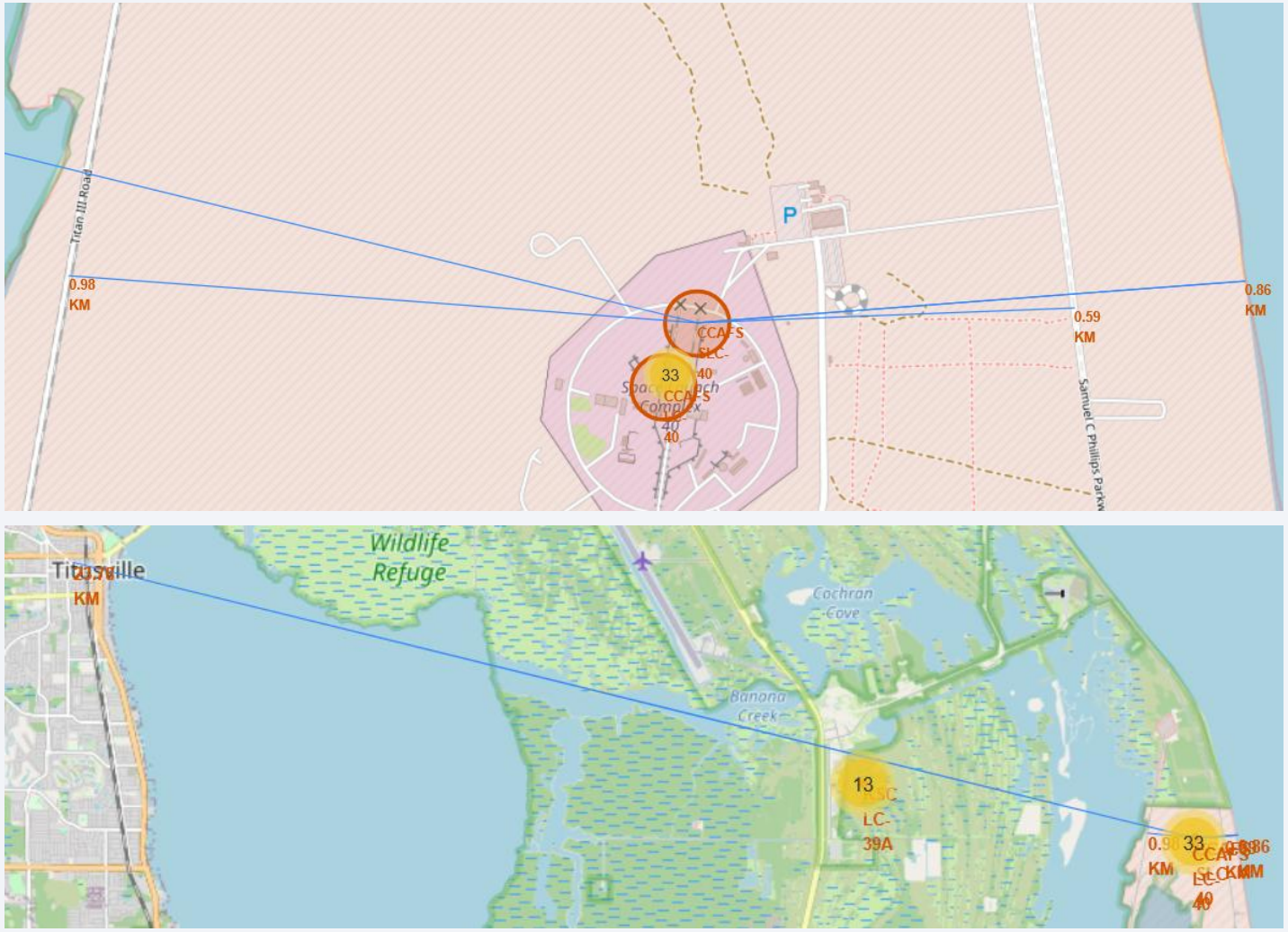
Launch outcomes



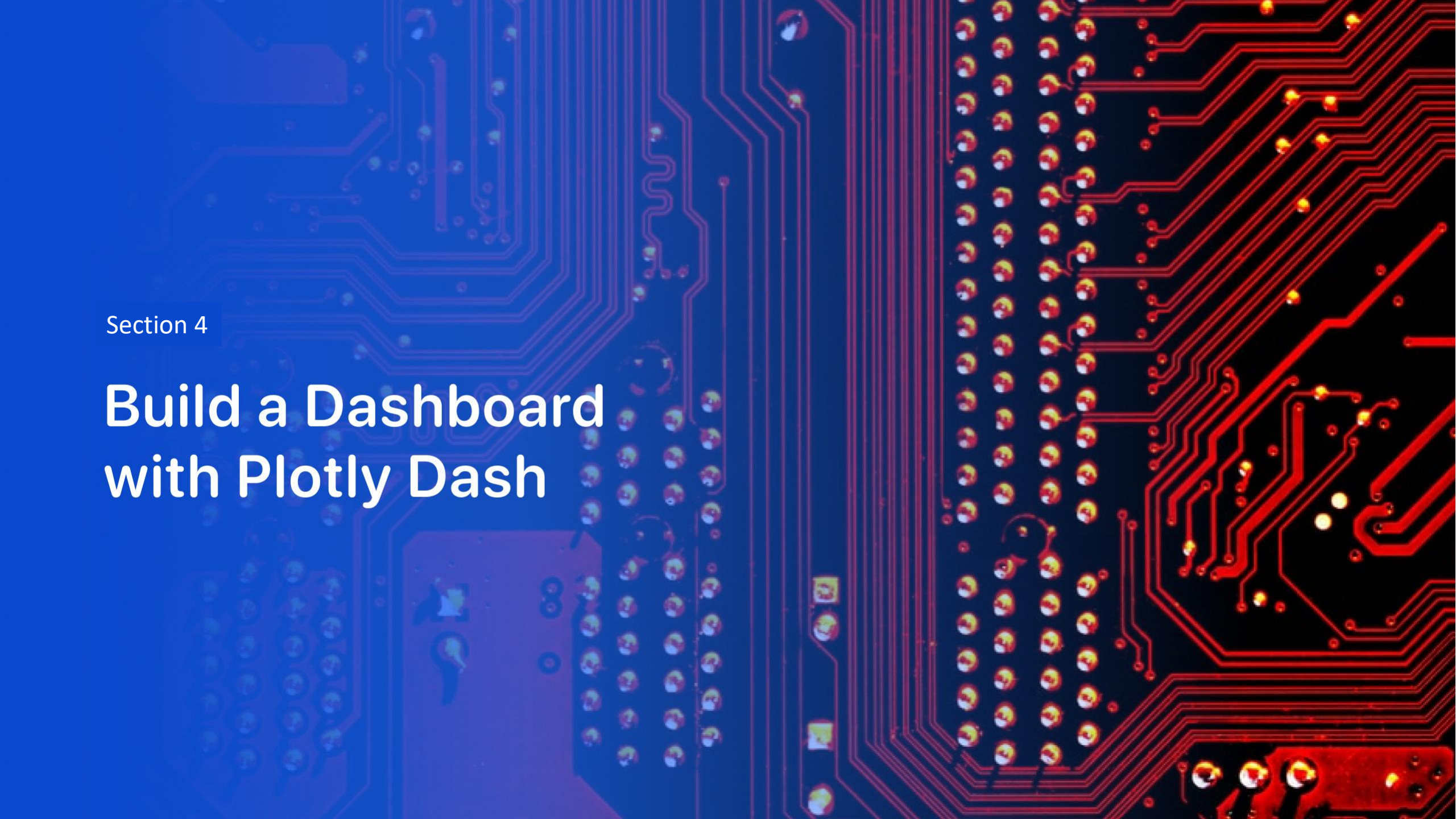


- Successful and failed launches are marked by green and red color respectively
- The success rate at CCAFS-SLC-40 is about 43%

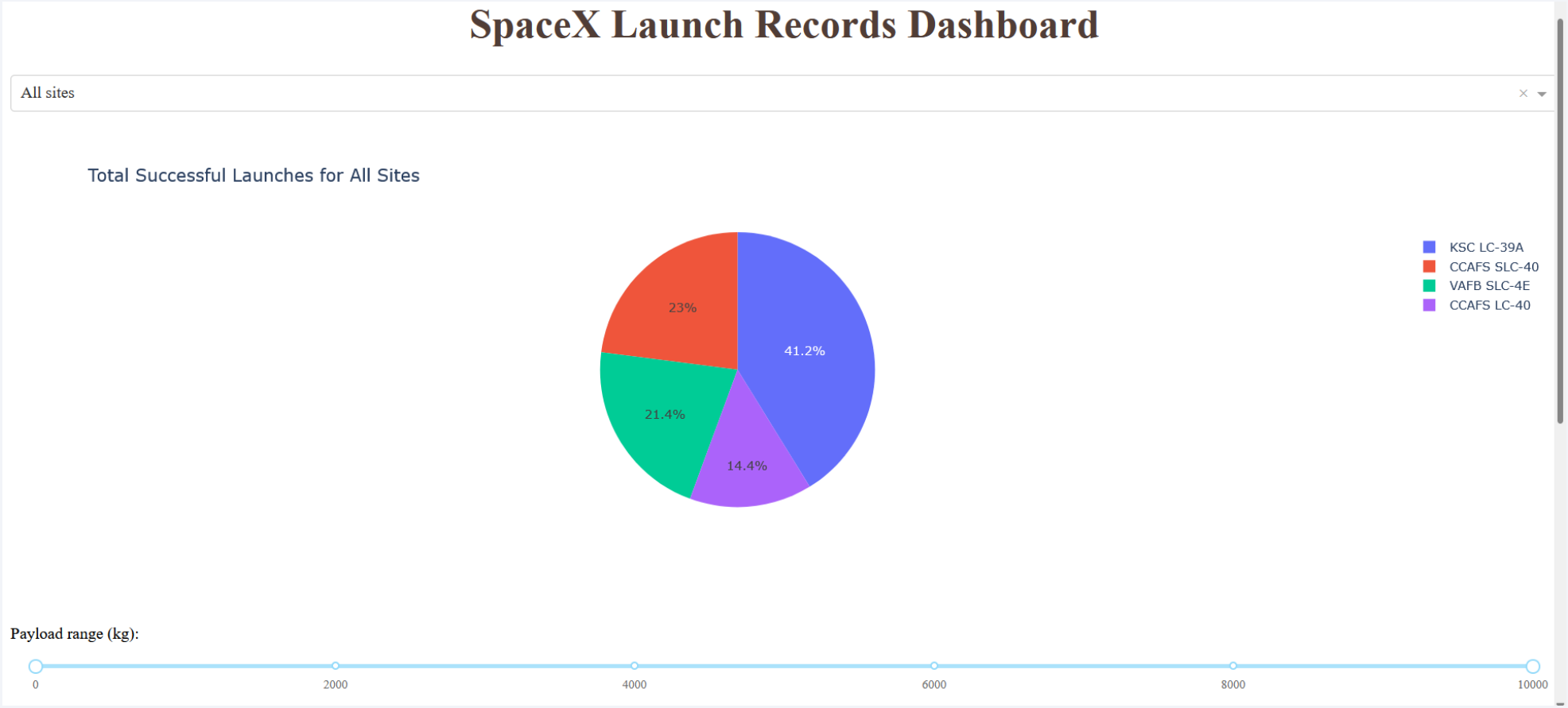
Distance to Proximities



- Distance to the closest highway is **0.58** km
- Distance to the closest railway is **0.98** km
- Distance to the nearest coastline is **0.86** km
- Distance to the nearest city Titusville is **23.78** km

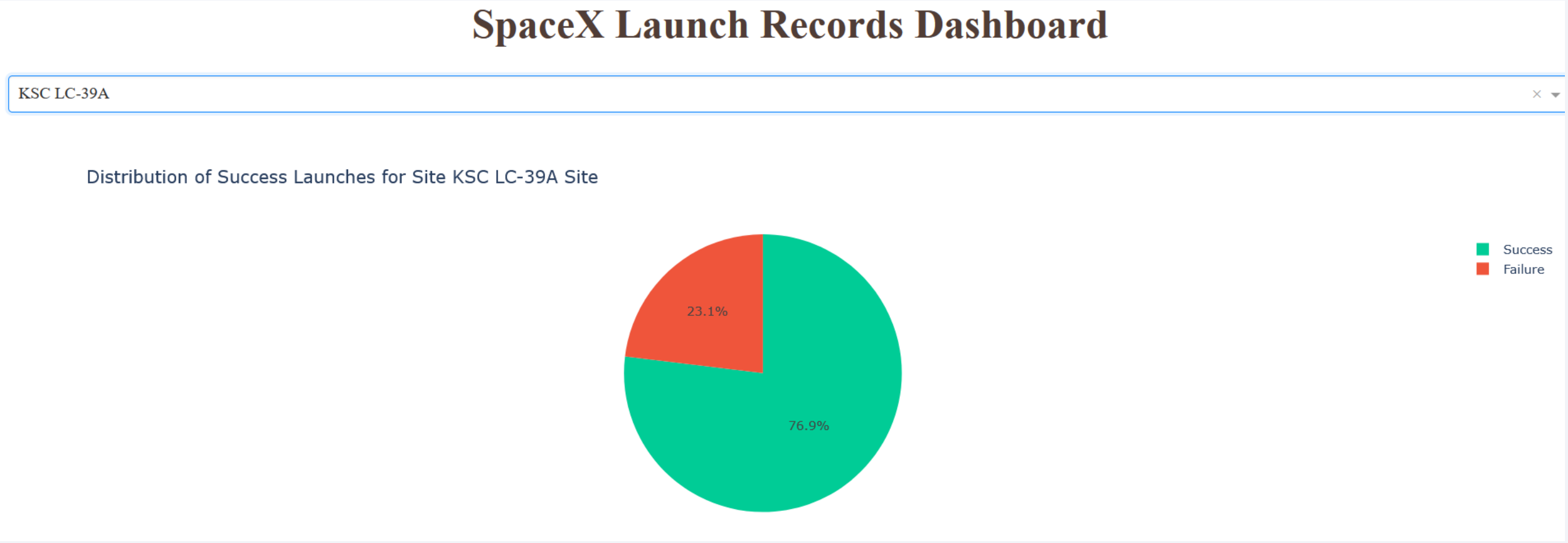

Section 4
Build a Dashboard with Plotly Dash

Success rate by a Launch Sites



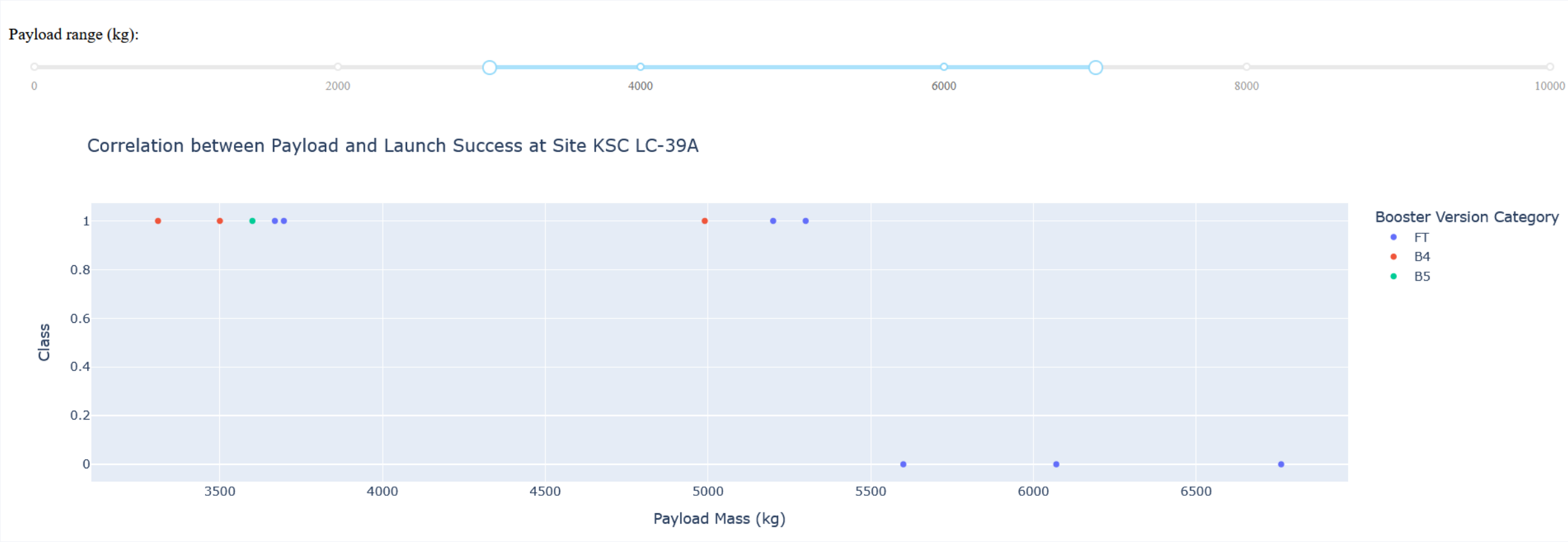
Launch site KSC LC-39A has the highest number of successful launches (41.2%)

Success rate for Site KSC LC-39A



- Launch site KSC LC-39A has the highest success rate of 76.9% among all the launch sites
- KSC LC-39A has 10 successful and 3 failed launches

Success vs. Payload Mass

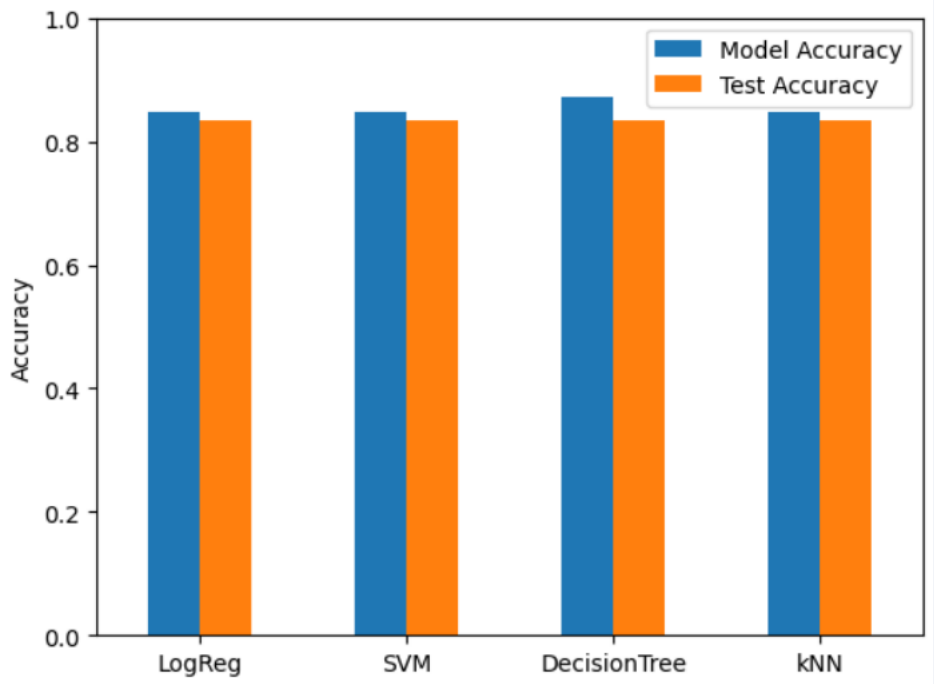


Success vs. Payload mass at KSC LC-39A launch site for payloads between 3000 and 7000 kg

Section 5

Predictive Analysis (Classification)

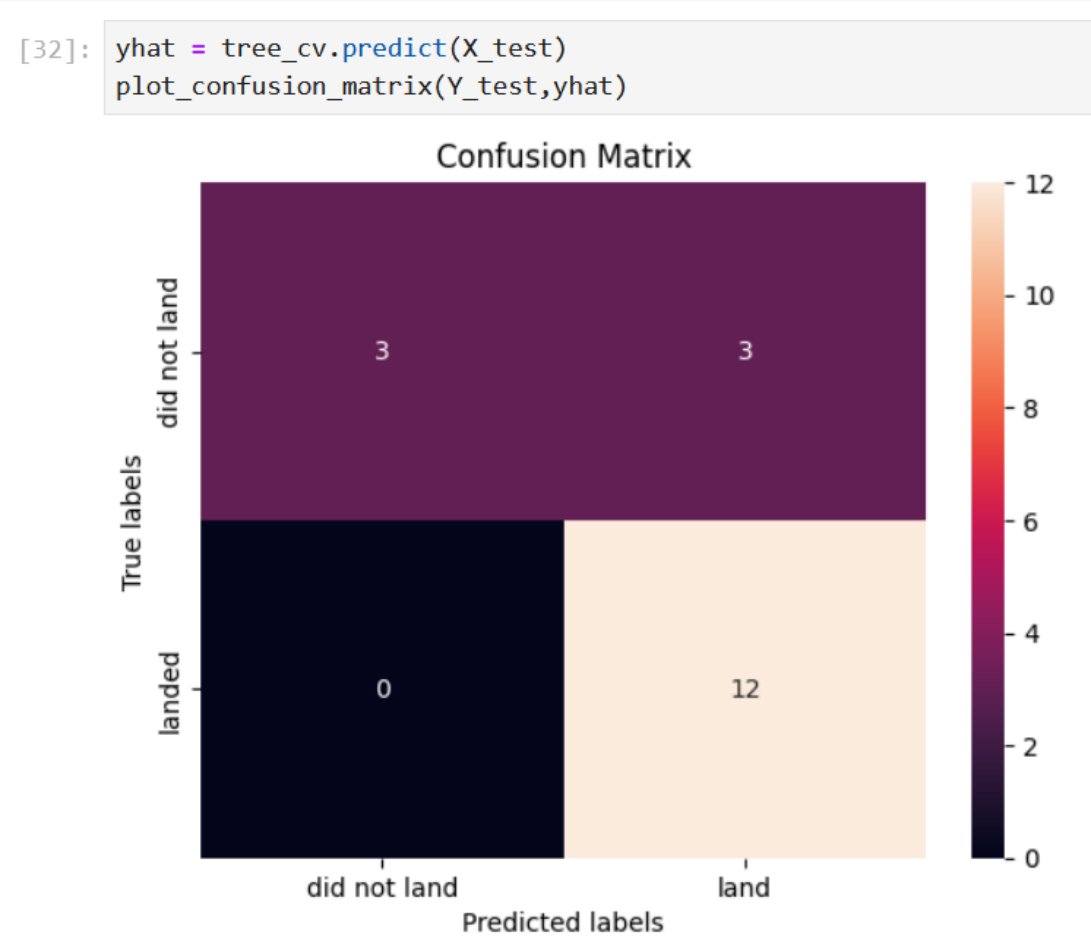
Classification Accuracy



- All methods perform with the same accuracy 0.833 when evaluated on the test data. This is likely due to a small dataset.
- DecisionTreeClassifier was performing slightly better than others when evaluated with .best_score_

Confusion Matrix

```
[32]: yhat = tree_cv.predict(X_test)
      plot_confusion_matrix(Y_test,yhat)
```



- All tried methods produced identical confusion matrix
- Confusion Matrix for DecisionTreeClassifier model is presented because it is performing slightly better than others when evaluated with .best_score_
- True positive = 12
- True negative = 3
- False positive = 3
- False negative = 0

Conclusions

- Success rate increases over time
- Launch site KSC LC-39A has the highest number (41.2%) of successful launches
- All launch sites are located close to the equator and a coastline
- ES-L1, GEO, HEO and SSO type of orbits have the highest success rate
- Launches with heavy payload (over 9000 kg) are more successful
- DecisionTreeClassifier is a good model for predicting the success of the landing in this project

Thank you!