IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Cayo Betancourt
Mar 22, 2024

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results
  - EDA Results
  - Interactive visual analytics dashboards
  - Predictive analysis

Introduction

- Project background and context

The financial success of SpaceX relies in the first stage rocket re-utilization, increasing the cost savings up to 30% against other competitors.

- Problems you want to find answers

The data provided for Falcon 9 launches is used to predict its first stage successful landing and the author is looking to match with company´s website acclaims

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected from the API website, where the company openly share details about its past performance (https://api.spacexdata.com/v4/launchpads/, https://api.spacexdata.com/v4/payloads & https://api.spacexdata.com/v4/launches/past)
- Perform data wrangling
  - Data wrangling was performed using the mean to replace missing values to increase the data quality and identifying another cases where the booster cannot land successfully.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Preprocessing is used to standardize the data, then after splitting it between training and testing, the author used grid search to find hyper parameters that perform best.

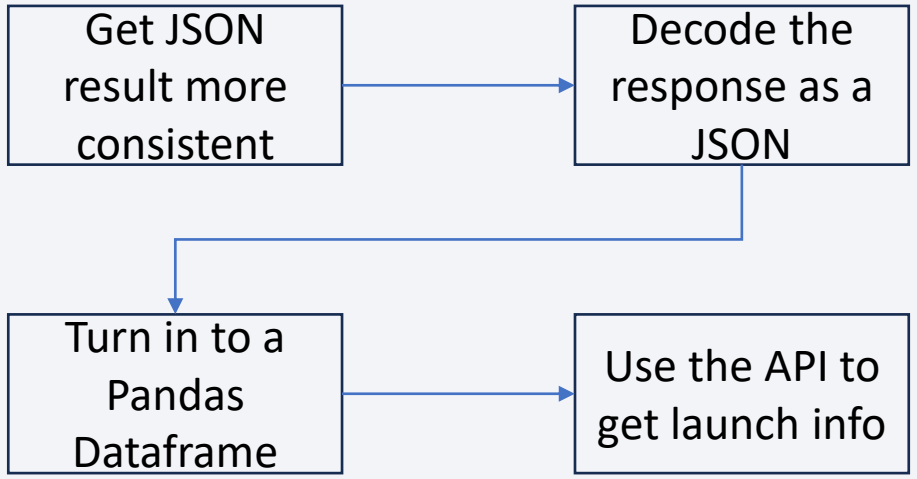
Data Collection

- The data collection process is divided in two main sources:
  - API resting from SpaceX website
  - Web scraping from Wikipedia
- A RESTful API collection is achieved to the SpaceX API, where extracting information from the launch data then matching it with rocket data.
- The list of successful launch records were obtained using Web scraping data collection from the Wikipedia website, this HTML data is parsed then converted to pandas data frame.

Data Collection – SpaceX API

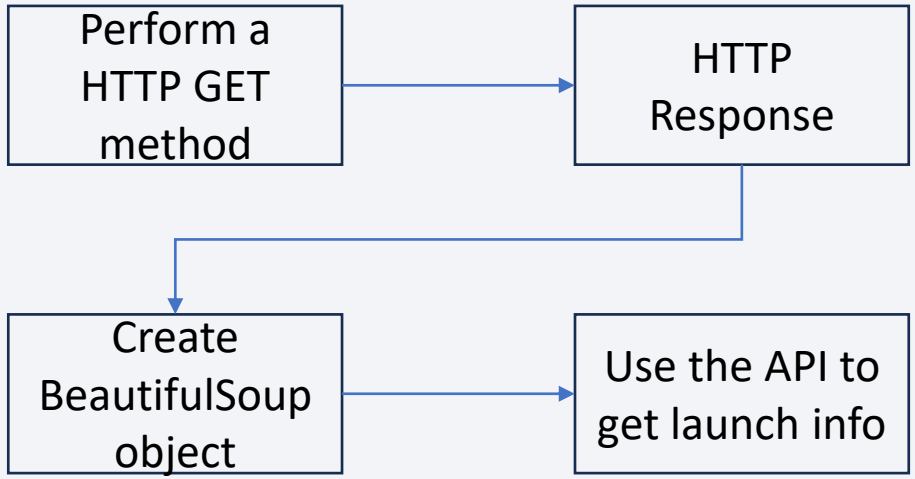
- A static response object is used to make the JSON request more consistent
- A .json_normalize() is used to decode a JSON and turn it into a pandas data frame
- The IDs given for each launch are sued to get information via the API

- SpaceX API calls notebook can be found here



Data Collection - Scraping

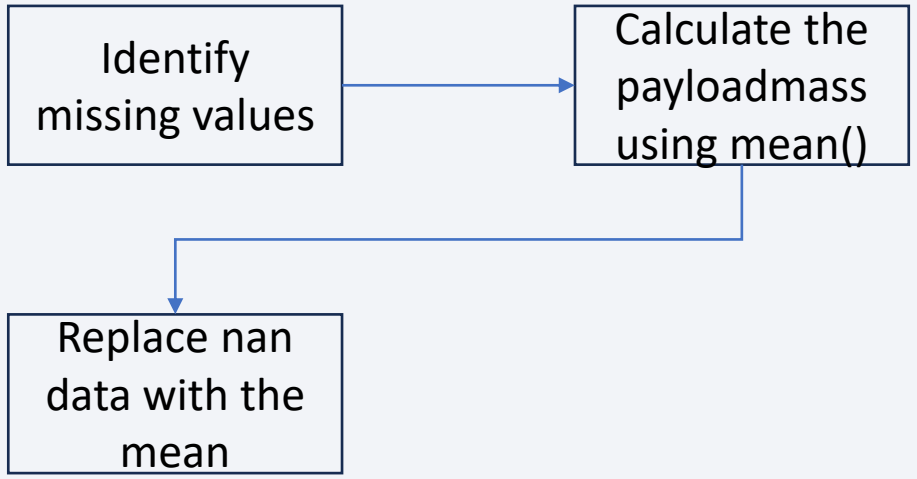
- A HTTP GET method is performed to get a HTTP Response
- Using the HTTP Response a BeautifulSoup object is created
- Finally, the API us used to get the launch info



- The GitHub url for web scraping can be found here

Data Wrangling

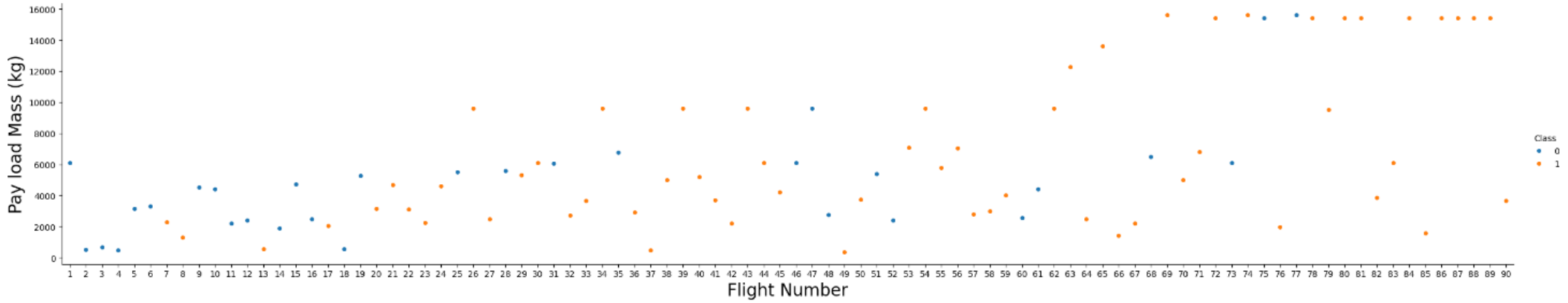
- Identify the missing values using df2.isnull().sum()
- Calculate below the mean for the PayloadMass using the .mean()
- Then use the mean and the .replace() function to replace np.nan values in the data with the mean you calculated.
- The GitHub url for Data Wrangling can be found here



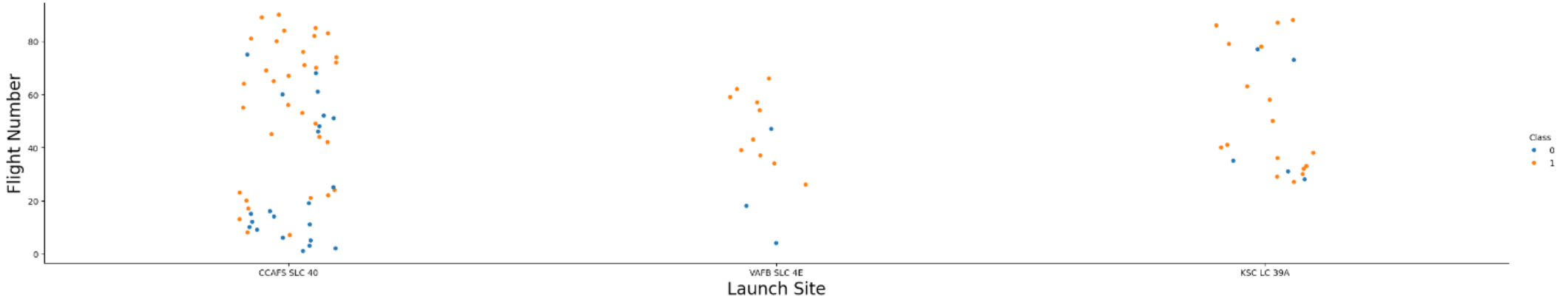
EDA with Data Visualization

- There is a direct relationship between data visualization and understanding.
- Scatter plots are more suitable to see how an independent variable is related to the dependent variable.
- Bar chats become an excellent representation to represent data sets in category axis.
- The space launch visualization increased the data understanding and success visualization supported by data.

EDA with Data Visualization (Cont)

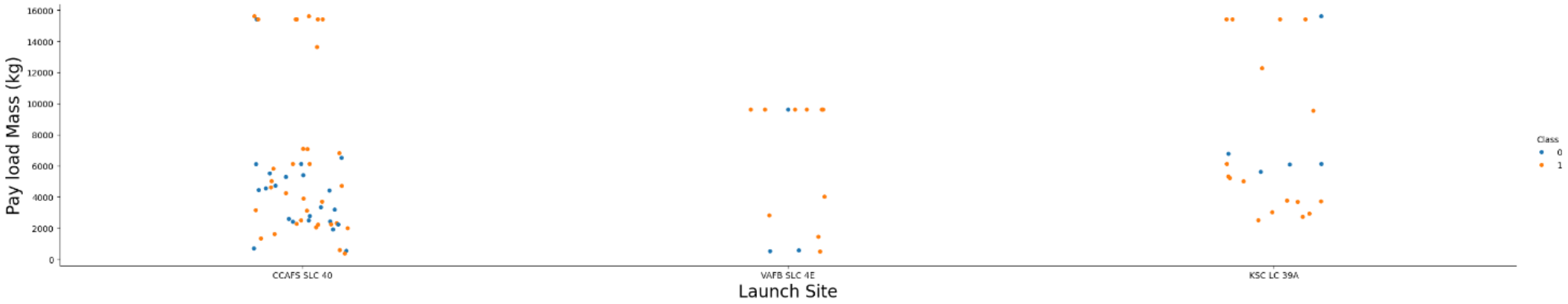


Flight number and payload

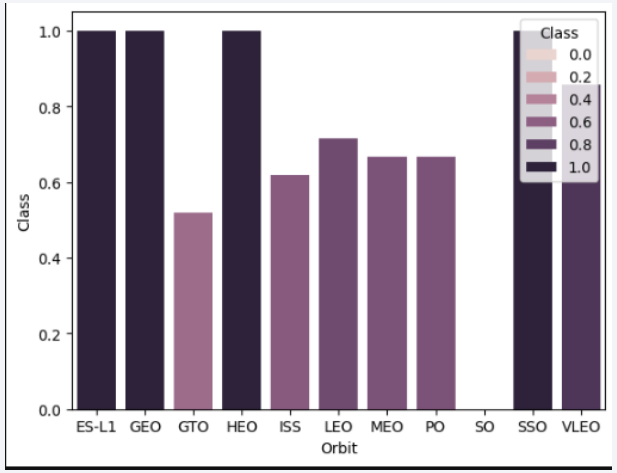


Flight number and launch site

EDA with Data Visualization (Cont)



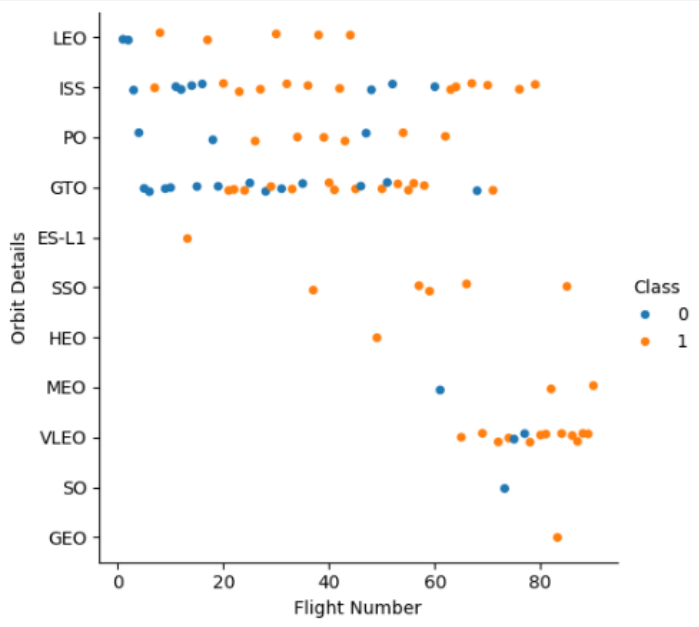
Launch sites and payload



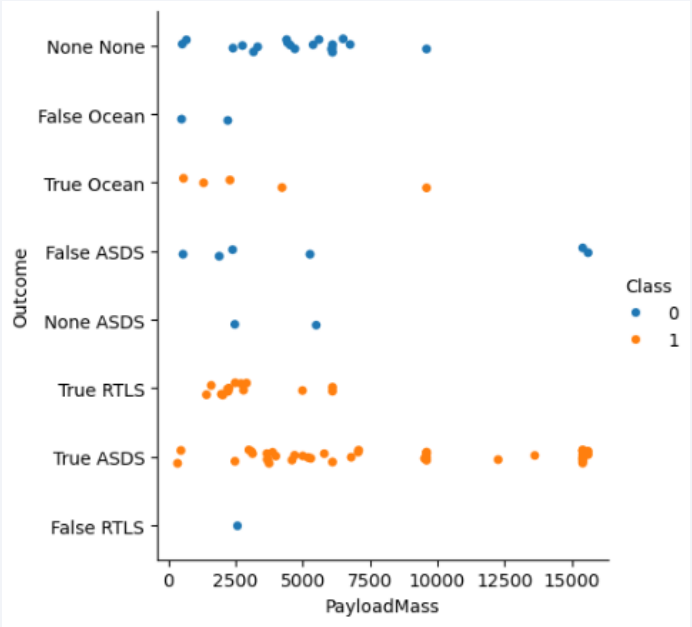
Success rate for each orbit type

- The link for EDA with data visualization can be found here

EDA with Data Visualization (Cont)



Relationship between flight number and orbit type



Relationship between payload and orbit type

- The link for EDA with data visualization can be found here

EDA with SQL

The below queries were executed for EDA

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

%sql select DISTINCT(launch_site) from SPACEXTABLE1

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

%sql SELECT * from SPACEXTABLE1 where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 20

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

- %sql SELECT sum(payload_mass__kg_) as sum_payload from SPACEXTABLE1 where (customer) = 'NASA (CRS)'

EDA with SQL (Cont)

The below queries were executed for EDA

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

%sql SELECT avg(payload_mass__kg_) as average_payload from SPACEXTABLE1 where (booster_version) = 'F9 v1.1'

- List the date when the first succesful landing outcome in ground pad was acheived.

%sql SELECT min(date) from SPACEXTABLE1 where Landing_Outcome = 'Success (ground pad)'

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

%sql select BOOSTER_VERSION from SPACEXTABLE1 where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4001 and 5999

EDA with SQL (Cont)

The below queries were executed for EDA

- List the total number of successful and failure mission outcomes

%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS OUTCOME FROM SPACEXTABLE1 GROUP BY MISSION_OUTCOME

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

%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS OUTCOME FROM SPACEXTABLE1 GROUP BY MISSION_OUTCOME

EDA with SQL (Cont)

The below queries were executed for EDA

- 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.

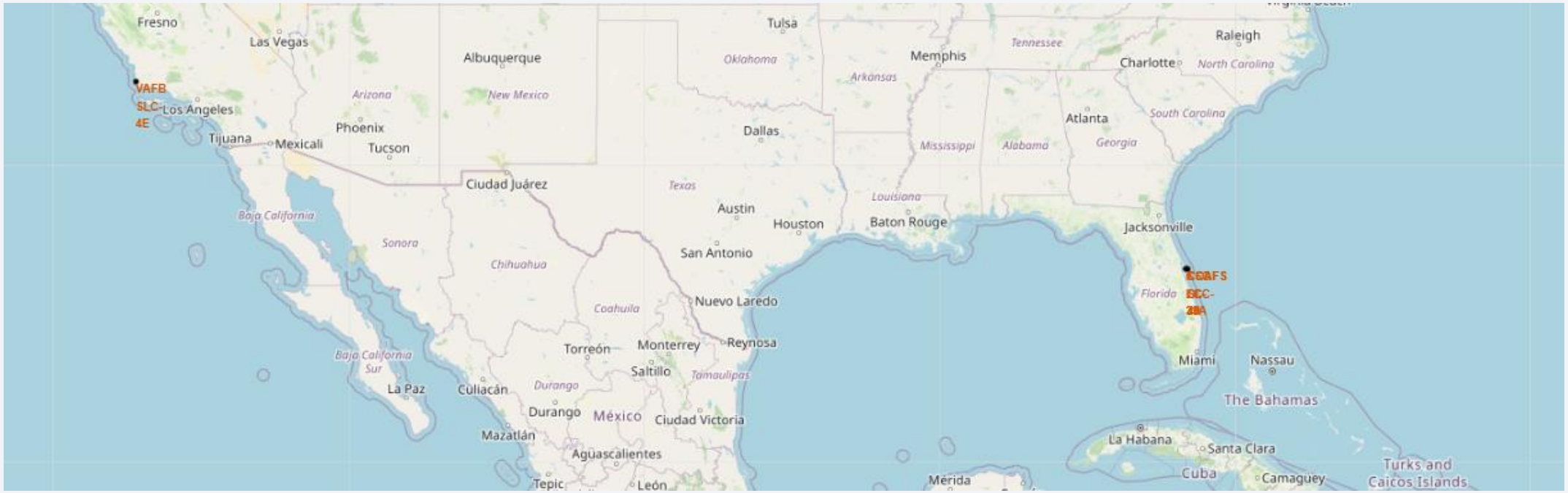
%sql SELECT Date, Booster_Version, Launch_Site, MISSION_OUTCOME, Landing_Outcome FROM SPACEXTABLE1 WHERE Landing_Outcome = 'Failure'

- 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_LAUNCHES FROM SPACEXTABLE1 WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUNT_LAUNCHES DESC;

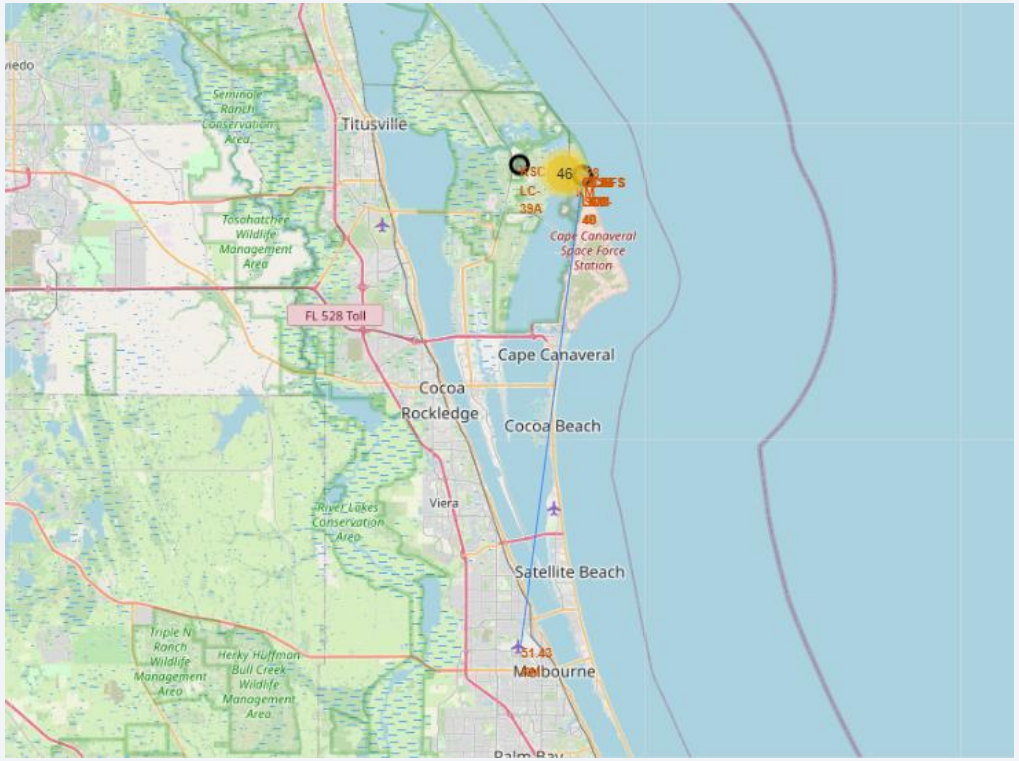
The link for EDA with SQL notebook can be found here

Build an Interactive Map with Folium



- The launch sites are close between them and to the equatorial line, making easier to recover stages and the near shore rescue process
- The circles and colors increases the visibility and the location in the map.
- The link for Interactive MAP with Folium can be found here

Build an Interactive Map with Folium



- Folium markers with distance

Build a Dashboard with Plotly Dash

- Below sites are used to plot and graph success launches for site and the Payload range (Kg):
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCA SLC-40
- Those plots and pie charts creates a visual comparison between the success launches and the associated payload per site
- The link for Dashboard with Ploty Dash notebook can be found here

Predictive Analysis (Classification)

The training data labels were created by:

- Creating Y as the outcome variable assigning the numpy() to it.
- The Sklearn function was used to transform the standardize function dataset.
- Then data was split into training and testing.

The logistic regression, nearest neighbors and classification trees were used to identify the most suitable Machine Learning model.

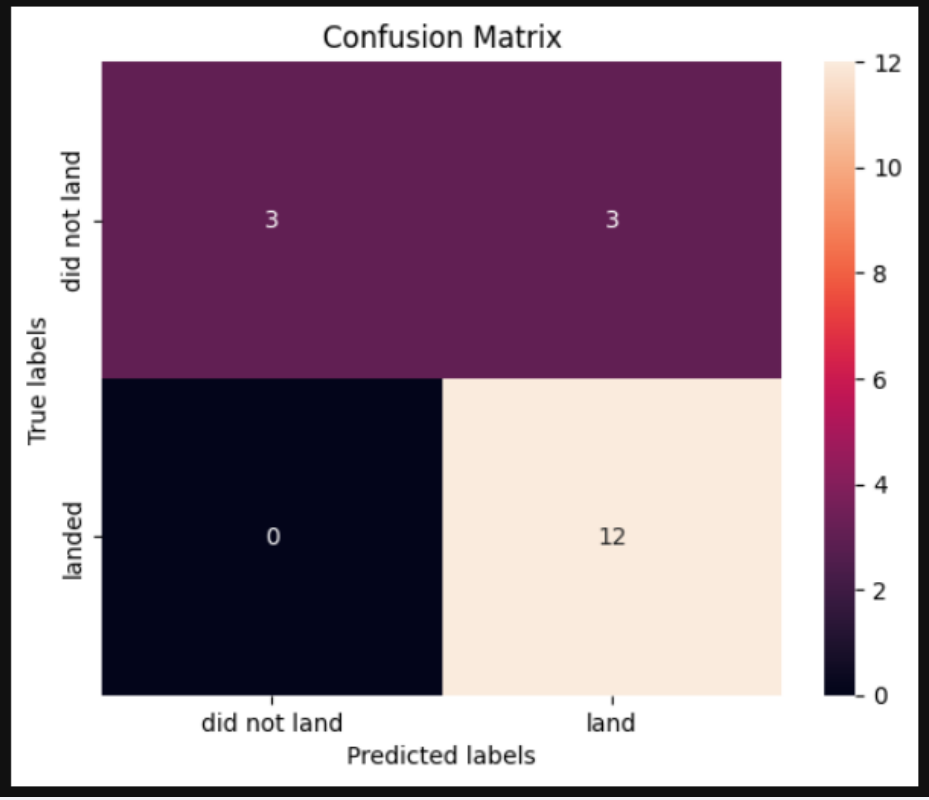
The link for predictive Analysis can be found here

Results

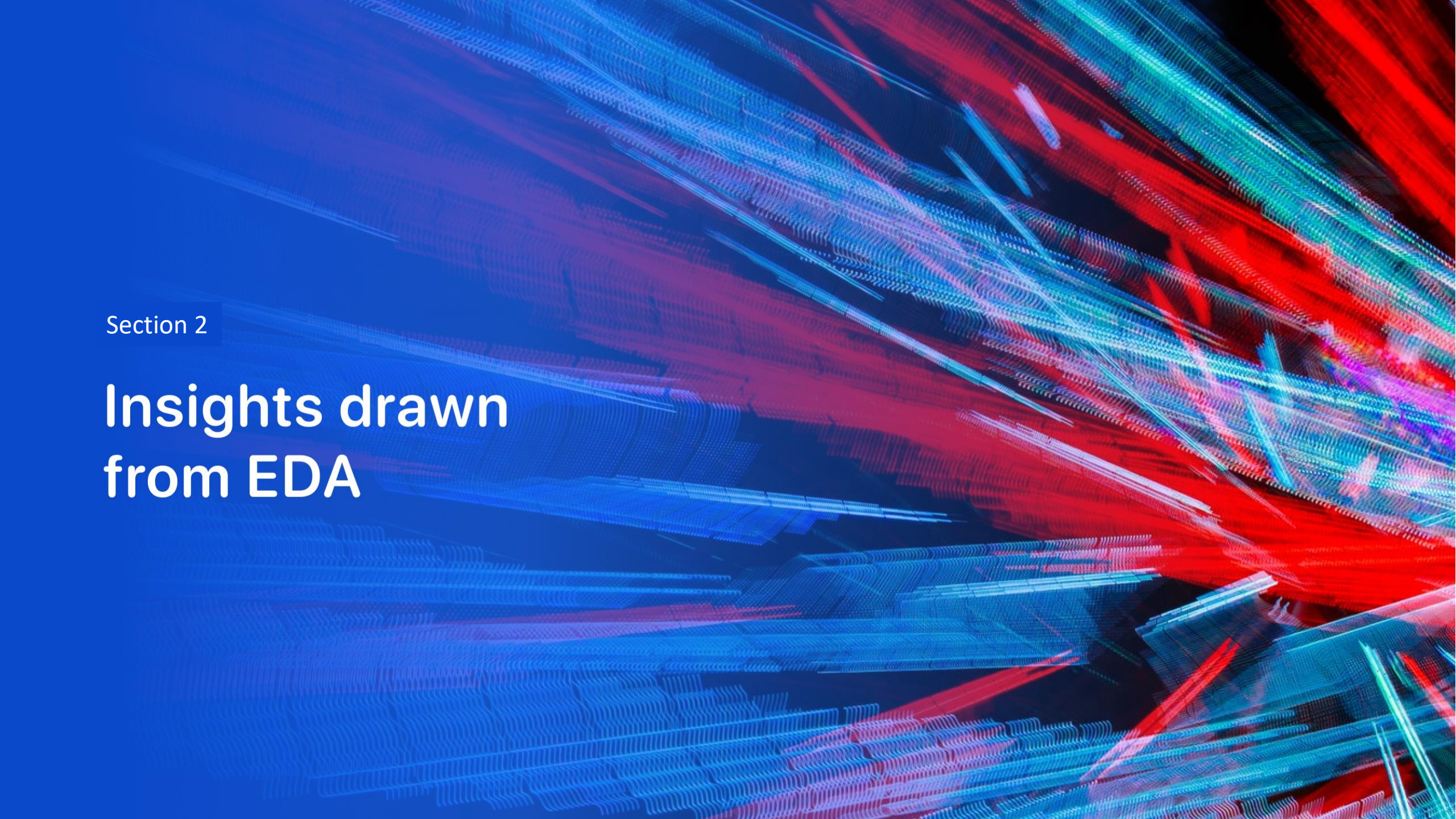
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Results

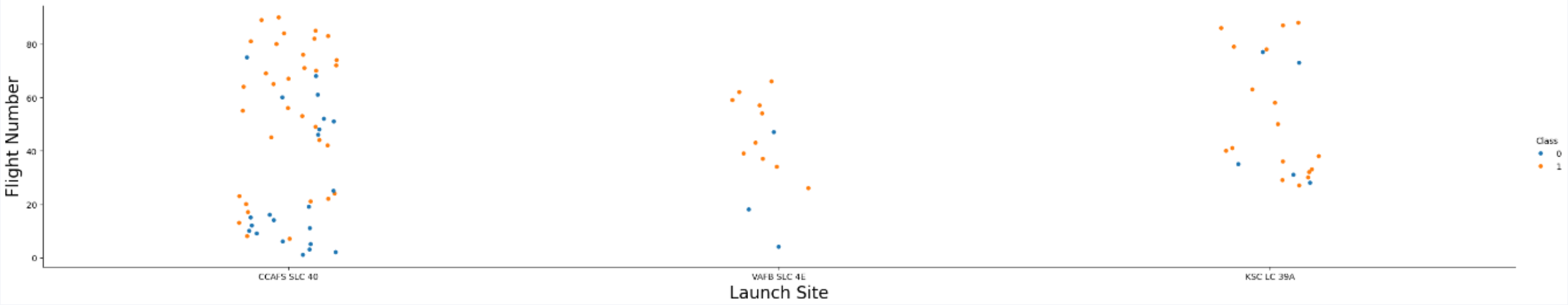
```
[15]: yhat=logreg_cv.predict(X_test)
      plot_confusion_matrix(Y_test,yhat)
      plt.show()
```



False positives are the major problem

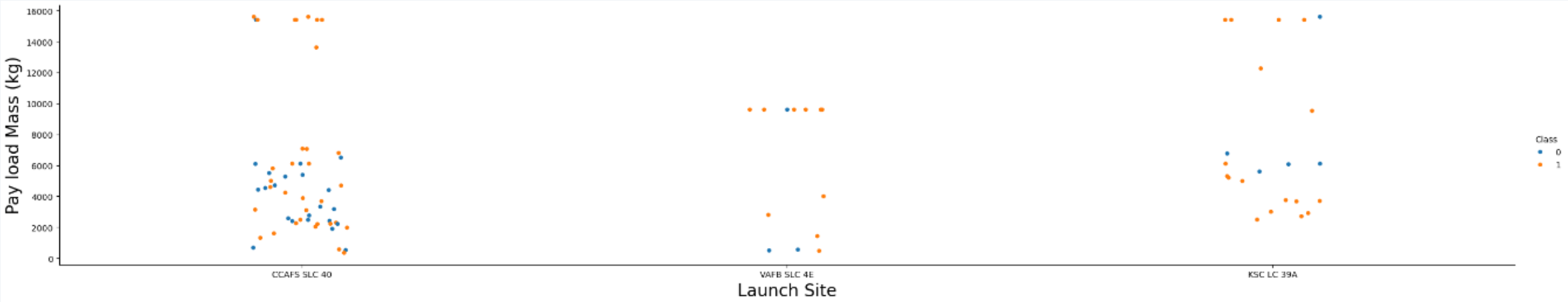

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



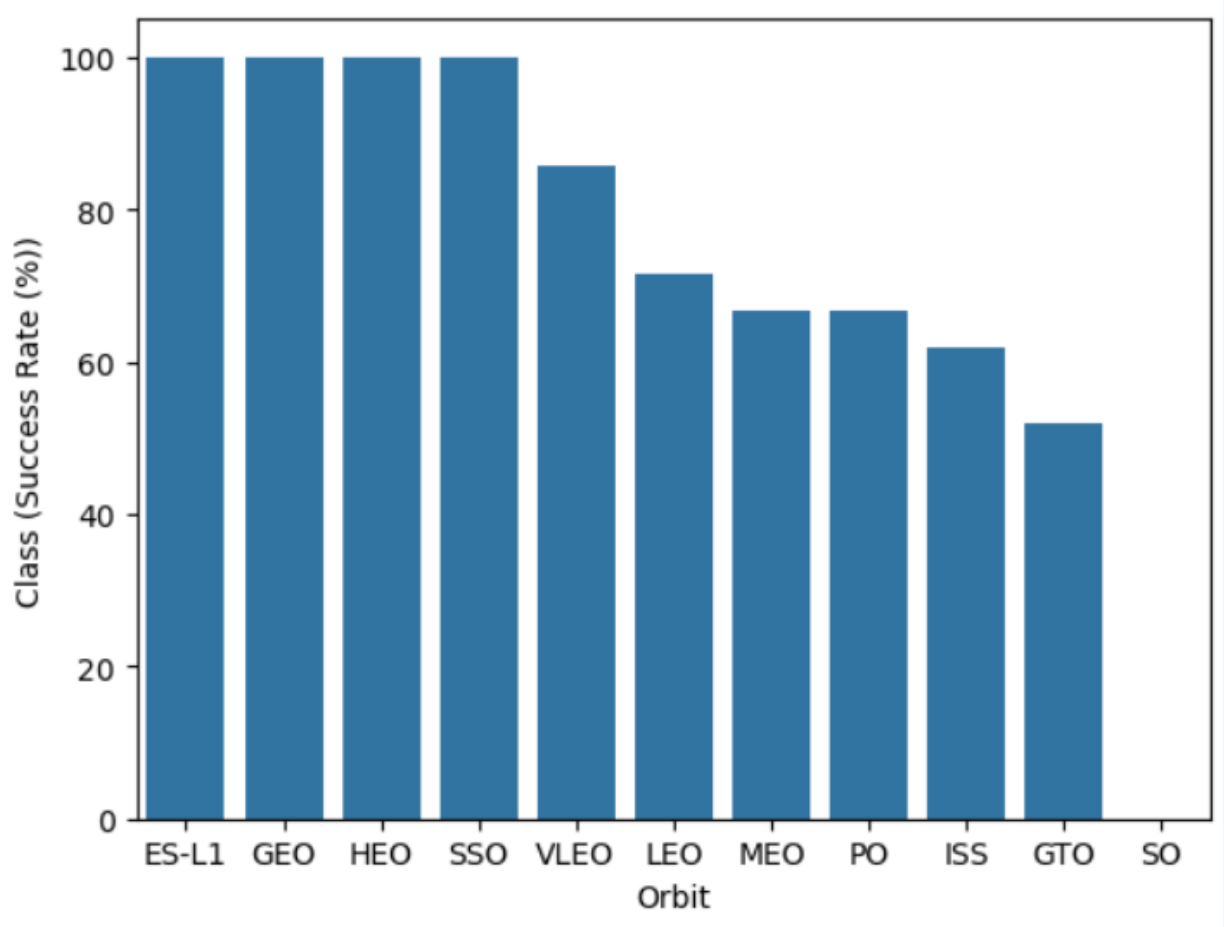
- The class 1 are more prone to have higher number of flight number across all sites

Payload vs. Launch Site



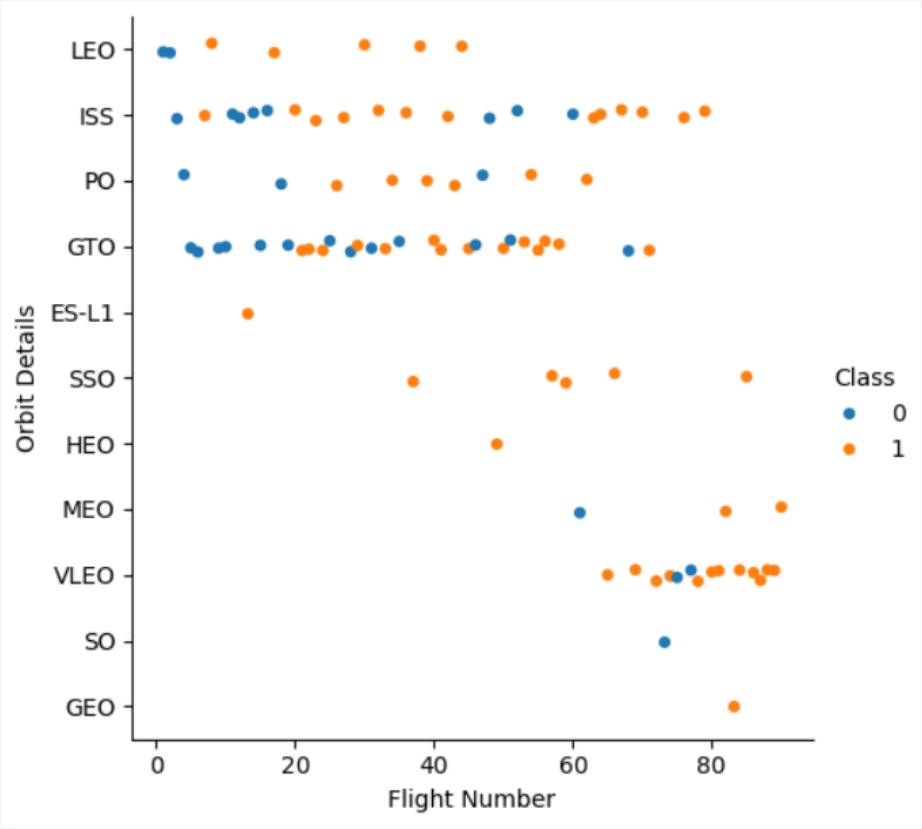
- The CACFLS SLC 40 and KSC LS 39A have the most number of flights and heavy loads

Success Rate vs. Orbit Type



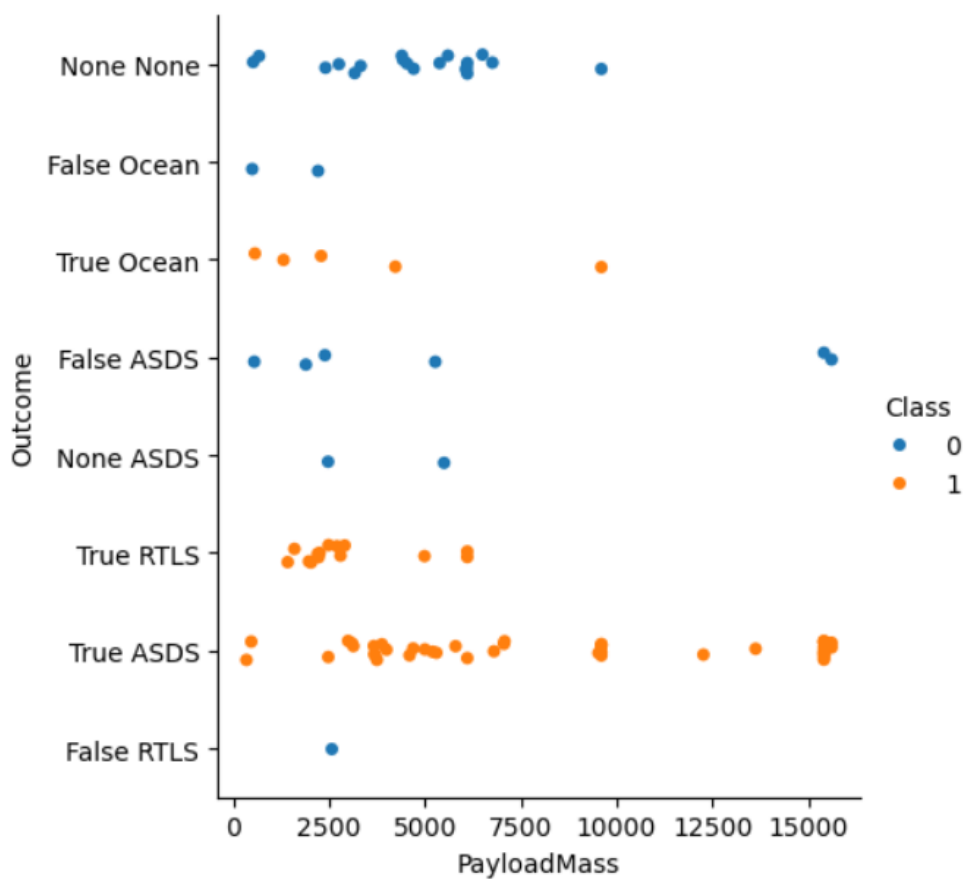
- The ES-L1, GEO, HEO and SSO are the most successful orbits

Flight Number vs. Orbit Type



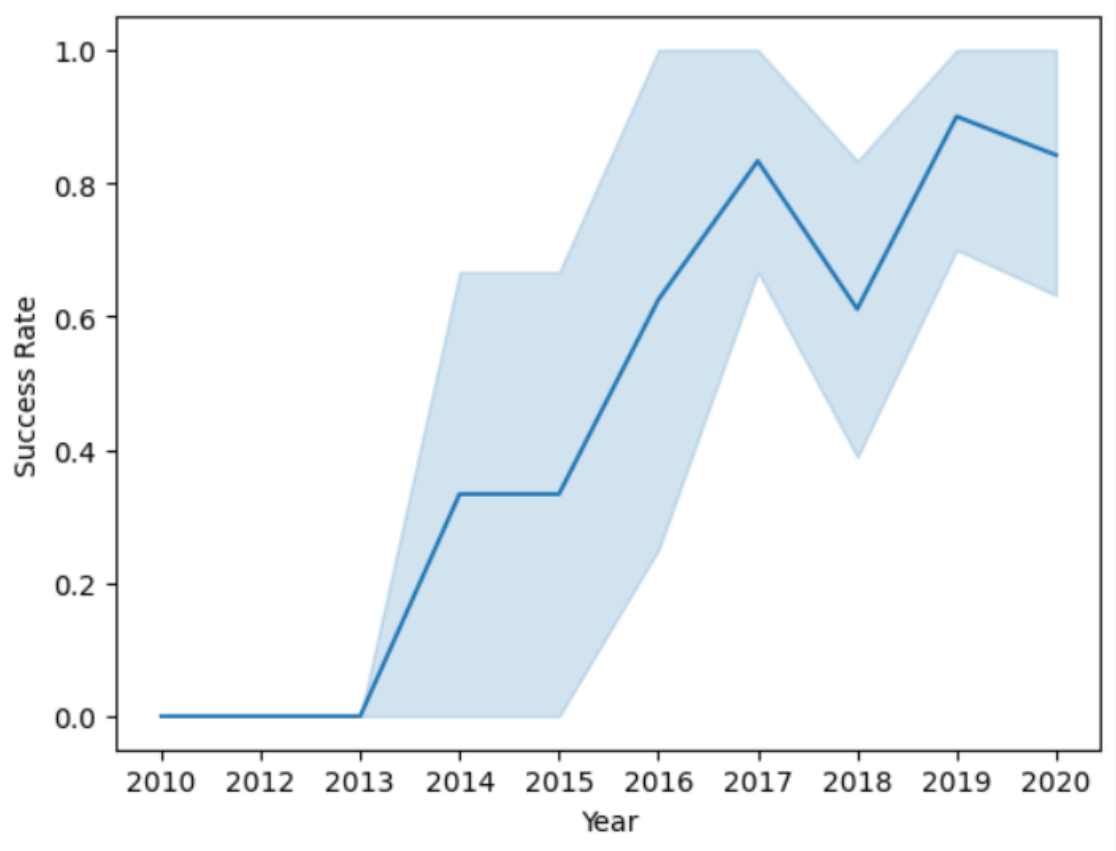
- The VLEO orbit has the most number of flights for class 1

Payload vs. Orbit Type



- There are more success for payload mass less than 10.000 for class 1 and 0

Launch Success Yearly Trend



- 2017 and 2019 has the most success rate. It is required to review why the success rate drop in 2018

All Launch Site Names

Launch Site	Lat	Long
CCAFS LC-40	28.562302	-80.577356
CCAFS SLC-40	28.563197	-80.576820
KSC LC-39A	28.573255	-80.646895
VAFB SLC-4E	34.632834	-120.610745

- This code is used to identify the launch site and its location

```
spacex_df = spacex_df[['Launch Site', 'Lat', 'Long', 'class']]
launch_sites_df = spacex_df.groupby(['Launch Site'], as_index=False).first()
launch_sites_df = launch_sites_df[['Launch Site', 'Lat', 'Long']]
launch_sites_df
```

Launch Site Names Begin with 'CCA'

```
[16]: %sql SELECT * from SPACEXTABLE1 where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5
 * sqlite:///my_data1.db
Done.
```

[16]:

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

- Selecting site names to limit the query with CCA% and limiting to 5 records

Total Payload Mass

```
%sql SELECT sum(payload_mass__kg_) as sum_payload from SPACEXTABLE1 where (customer) = 'NASA (CRS)'
```

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

sum_payload
45596

- The total payload mass is 45596 Kg

Average Payload Mass by F9 v1.1

```
%sql SELECT avg(payload_mass__kg_) as average_payload from SPACEXTABLE1 where (booster_version) = 'F9 v1.1'
```

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

average_payload
2928.4

- The average payload mass by F9 v1.1 is 2,928,4 Kg

First Successful Ground Landing Date

```
%sql SELECT min(date) from SPACEXTABLE1 where Landing_Outcome = 'Success (ground pad)'
```

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

min(date)
2015-12-22

- The first successful ground landing date was Dec 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

```
#To_check
%sql select BOOSTER_VERSION from SPACEXTABLE1 where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4001 and 5999
```

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

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

- This query list the booster versions capable to carry a payload between 4.000 and 6.000 Kg

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS OUTCOME FROM SPACEXTABLE1 GROUP BY MISSION_OUTCOME
```

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

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

- The table shows only one failure and 100 success, it will require further extrapolation tp identify the causes and final status

Boosters Carried Maximum Payload

Booster_Version	Payload	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Starlink 1 v1.0, SpaceX CRS-19	15600
F9 B5 B1049.4	Starlink 2 v1.0, Crew Dragon in-flight abort test	15600
F9 B5 B1051.3	Starlink 3 v1.0, Starlink 4 v1.0	15600
F9 B5 B1056.4	Starlink 4 v1.0, SpaceX CRS-20	15600
F9 B5 B1048.5	Starlink 5 v1.0, Starlink 6 v1.0	15600
F9 B5 B1051.4	Starlink 6 v1.0, Crew Dragon Demo-2	15600
F9 B5 B1049.5	Starlink 7 v1.0, Starlink 8 v1.0	15600
F9 B5 B1060.2	Starlink 11 v1.0, Starlink 12 v1.0	15600
F9 B5 B1058.3	Starlink 12 v1.0, Starlink 13 v1.0	15600
F9 B5 B1051.6	Starlink 13 v1.0, Starlink 14 v1.0	15600
F9 B5 B1060.3	Starlink 14 v1.0, GPS III-04	15600
F9 B5 B1049.7	Starlink 15 v1.0, SpaceX CRS-21	15600

The below code is used to identify the boosters Max payload in Kg

```
%sql SELECT "Booster_Version",Payload, "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE
"PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL);
```

2015 Launch Records

```
# No failures in year 2015_
%sql SELECT Date, Booster_Version, Launch_Site, MISSION_OUTCOME, Landing_Outcome FROM SPACEXTBL WHERE Landing_Outcome = 'Failure'
```

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

Date	Booster_Version	Launch_Site	Mission_Outcome	Landing_Outcome
2018-12-05	F9 B5B1050	CCAFS SLC-40	Success	Failure
2020-02-17	F9 B5 B1056.4	CCAFS SLC-40	Success	Failure
2020-03-18	F9 B5 B1048.5	KSC LC-39A	Success	Failure

The landing outcomes (failure) are presented in the table above

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

Landing_Outcome	COUNT_LAUNCHES
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

- The below code is used to filter the landing outcomes

```
%sql SELECT Landing_Outcome, COUNT(*) AS COUNT_LAUNCHES FROM
SPACEXTABLE1 WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome ORDER BY COUNT_LAUNCHES DESC;
```

Section 3

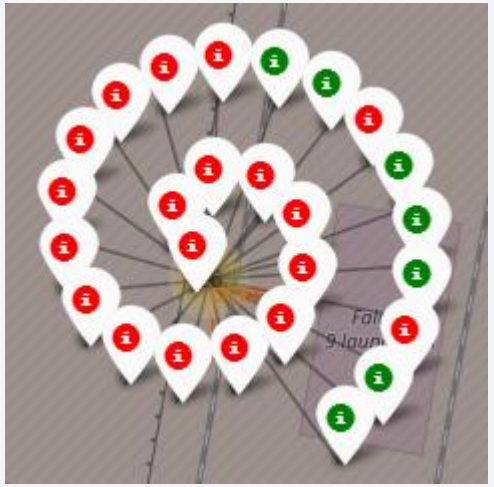
Launch Sites Proximities Analysis

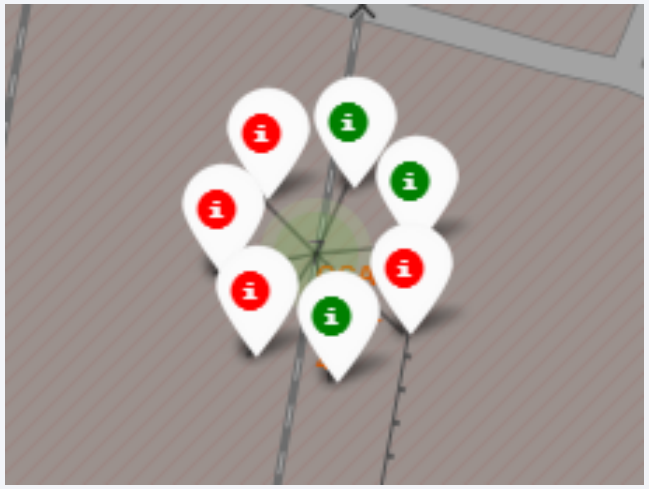
Creata a folium.Circle and folium.Marker for each launch site on the site map

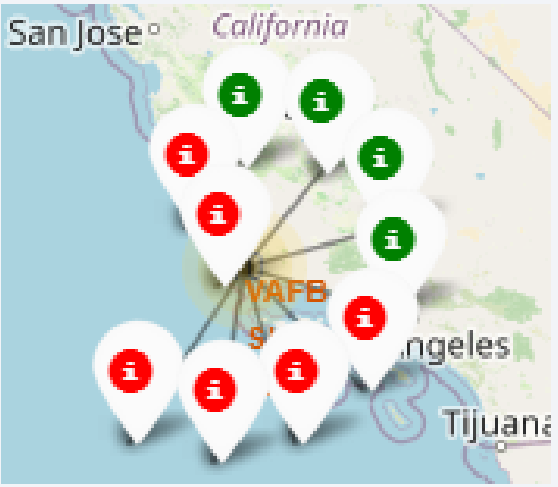
- All site locations are close to the Equator and the shore lines

launch result in spacex_df data frame, add a folium.Marker to marker_cluster

For each launch result in spacex_df data frame, add a folium.Marker to marker_cluster

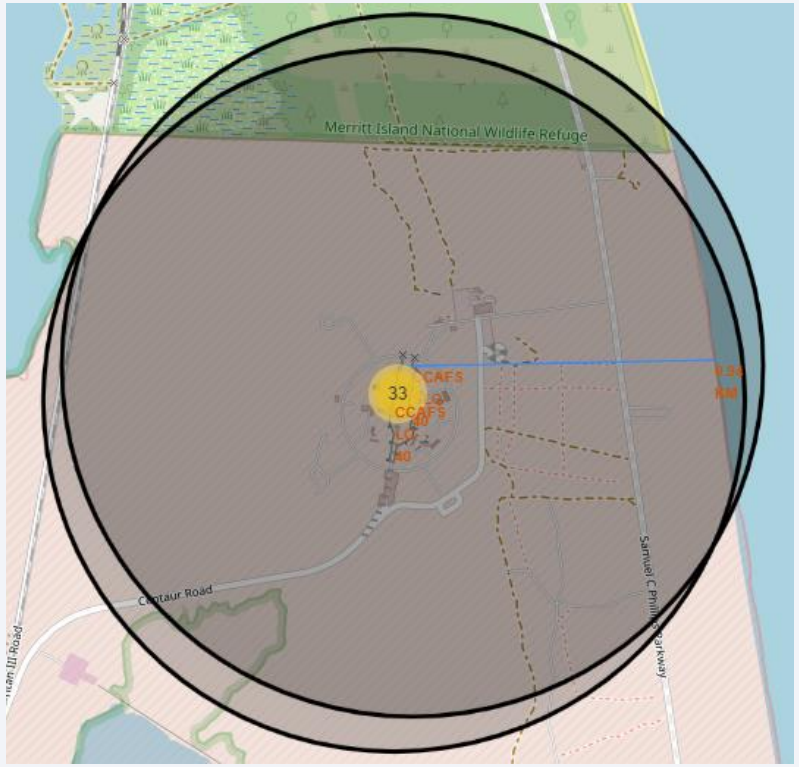






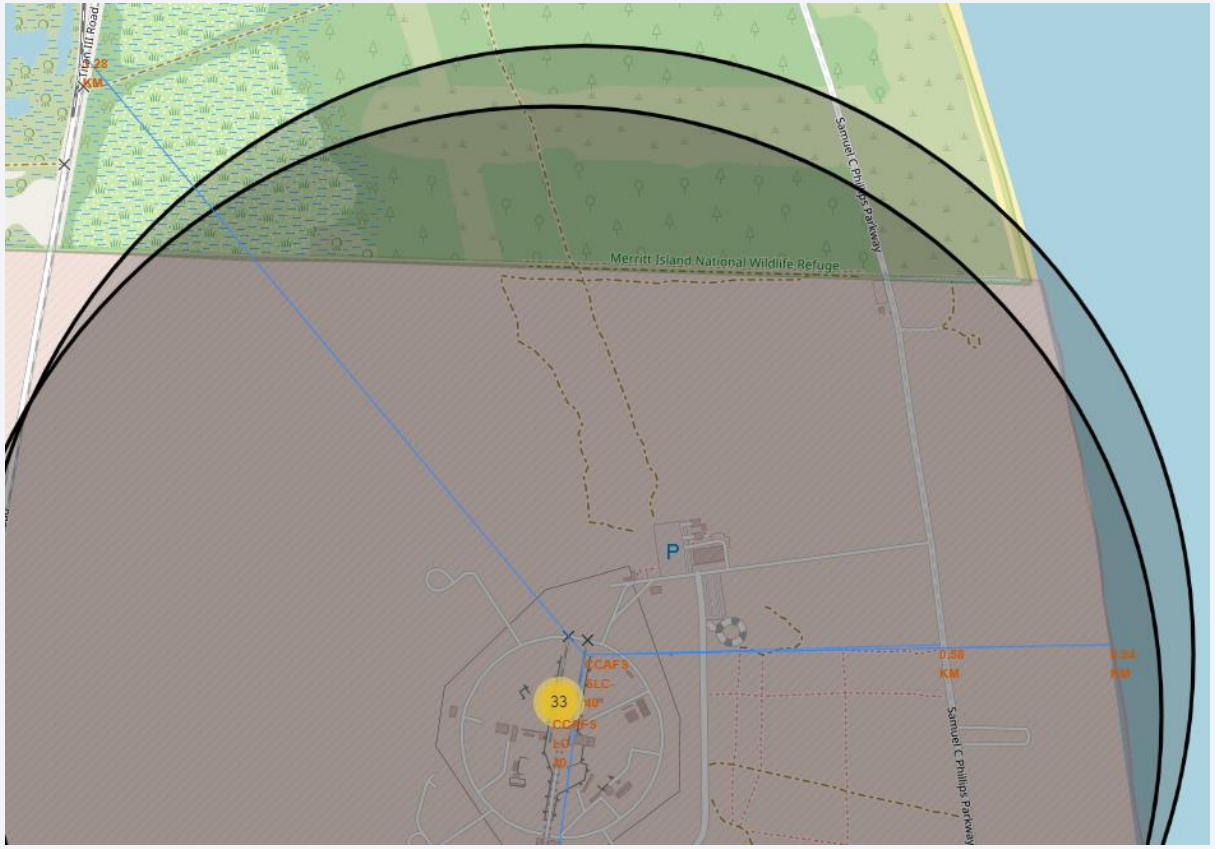
The two Florida sites (26+7 & 13) had more attempts than California site (10) and the success rate is higher for site KSC LC-39A

Distance to coastal line

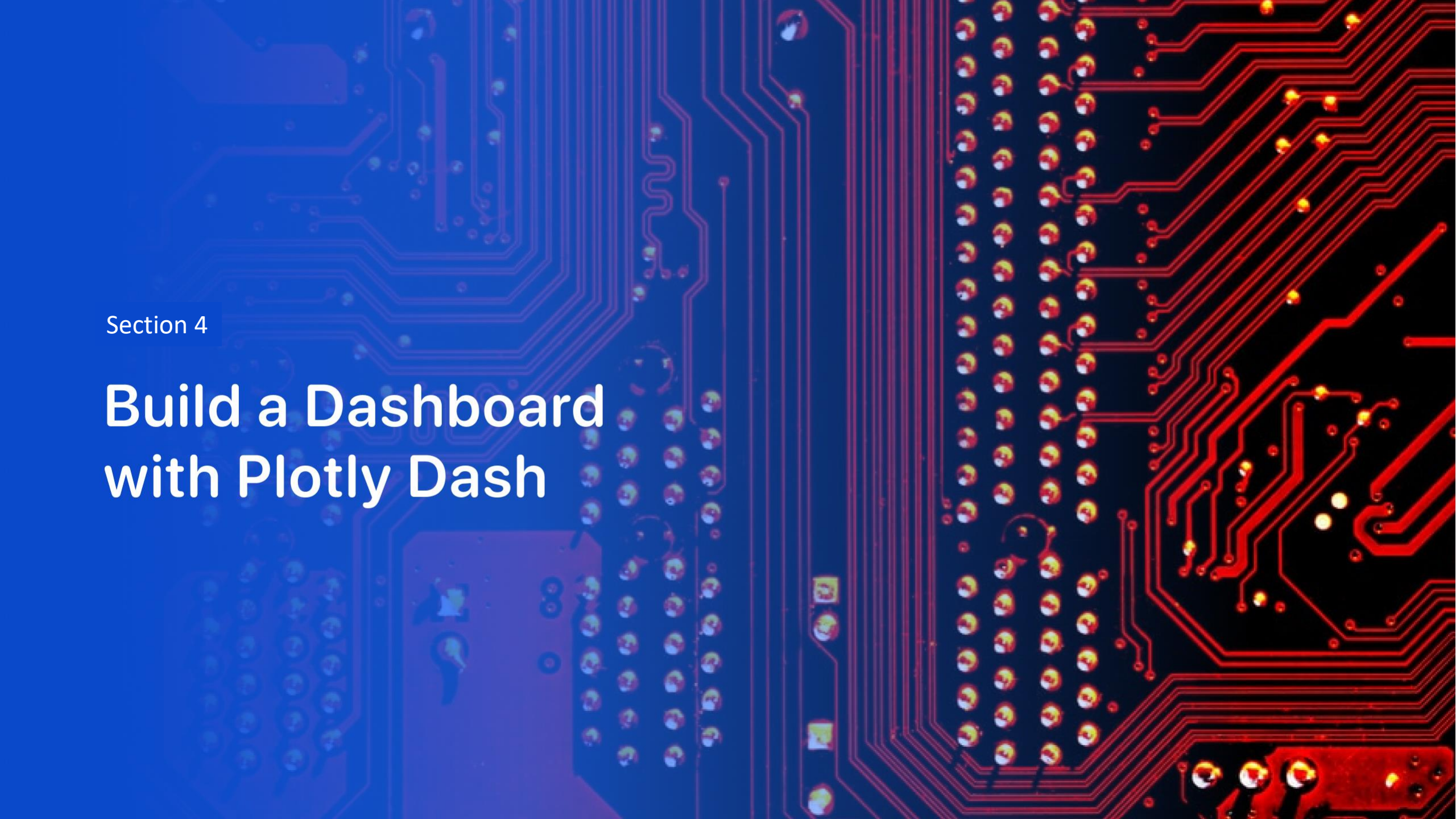


- The distance to coast line is 0.94 Km
- The Centaur Road seems the main road to get materials to launch site

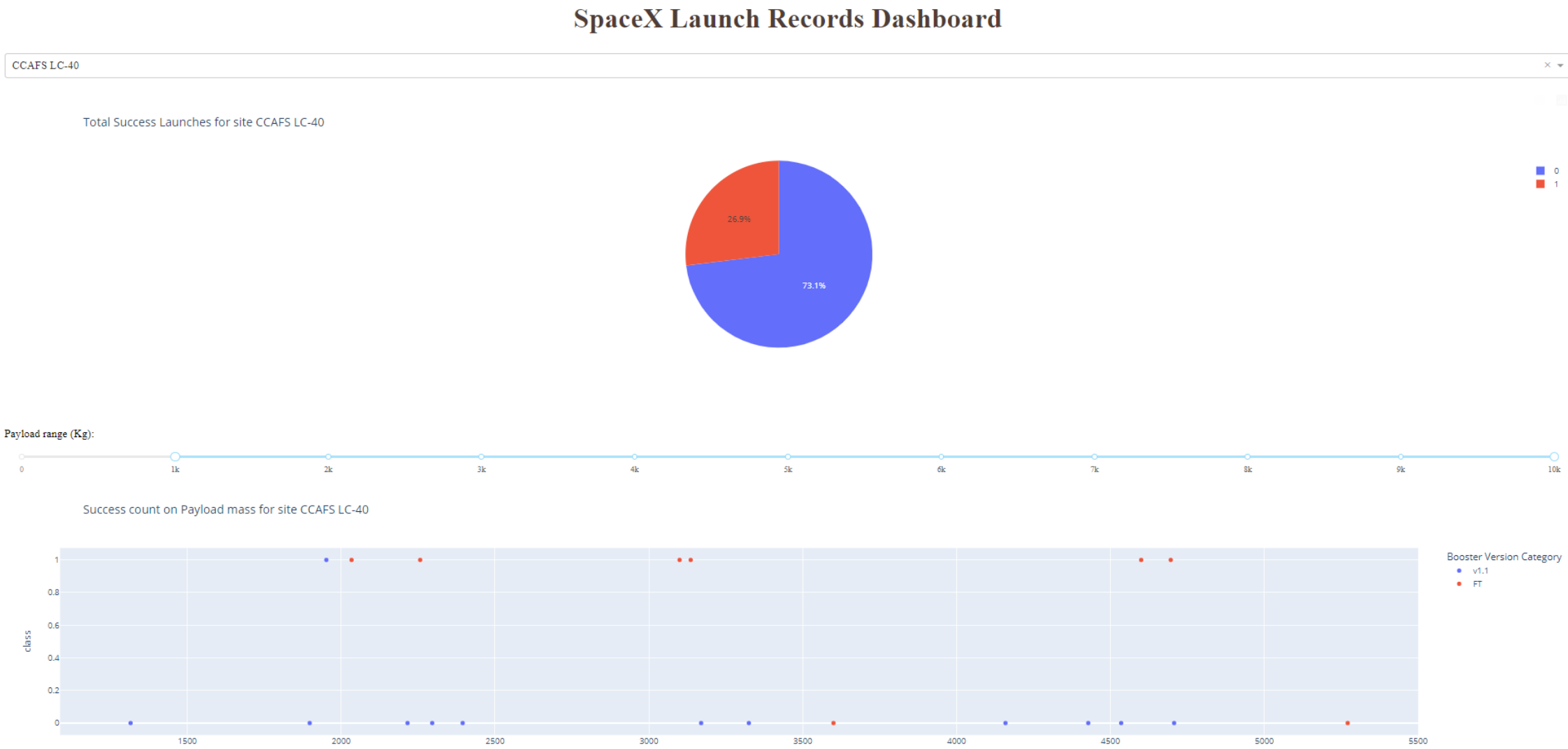
Distance between CACAFS SLC-40 and other elements



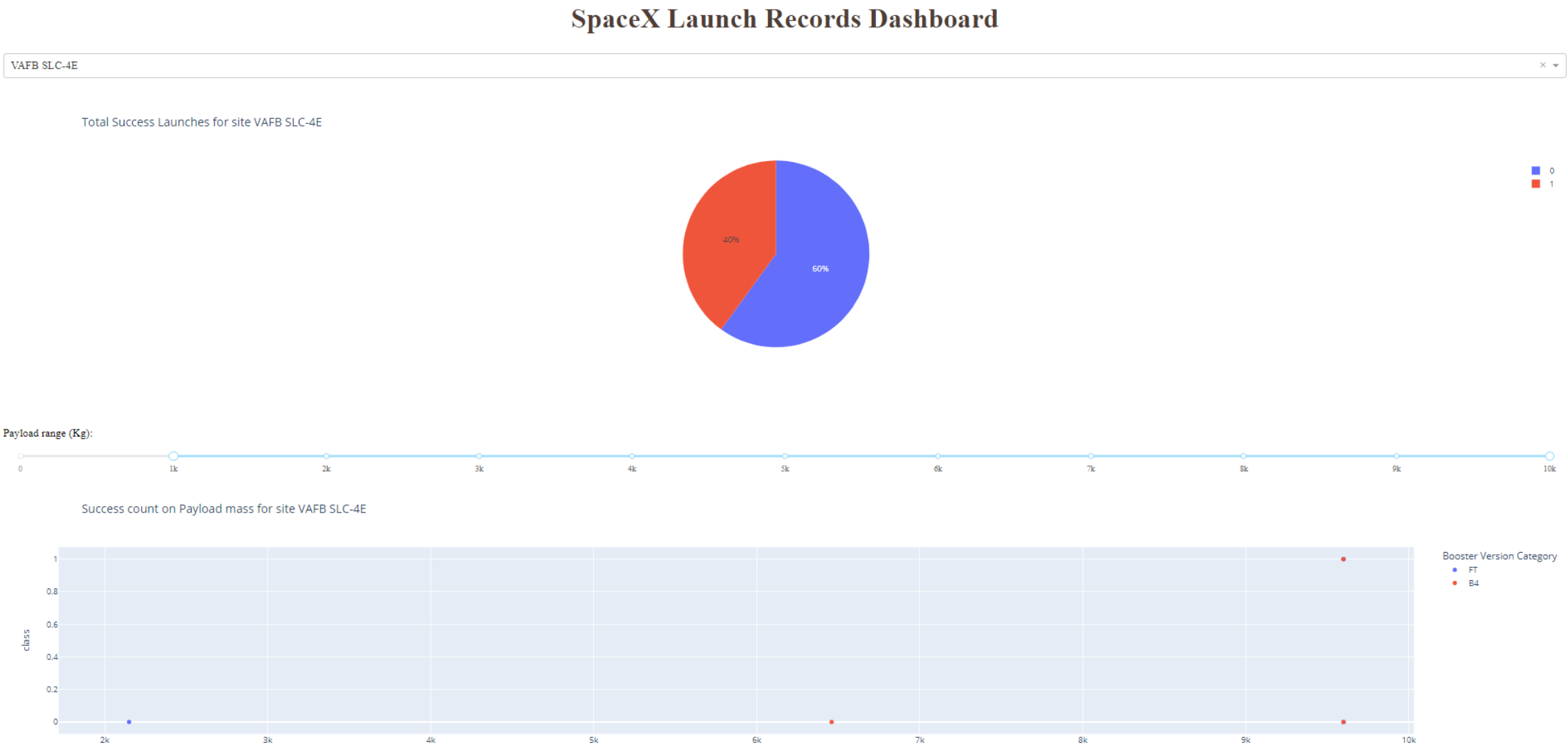
- The distance to airport is 51.43 Km
- The distance to Nasa railroad is 1.28 Km
- The distance to Samuel C Phillips Parkway is 0.58 Km


Section 4
Build a Dashboard with Plotly Dash

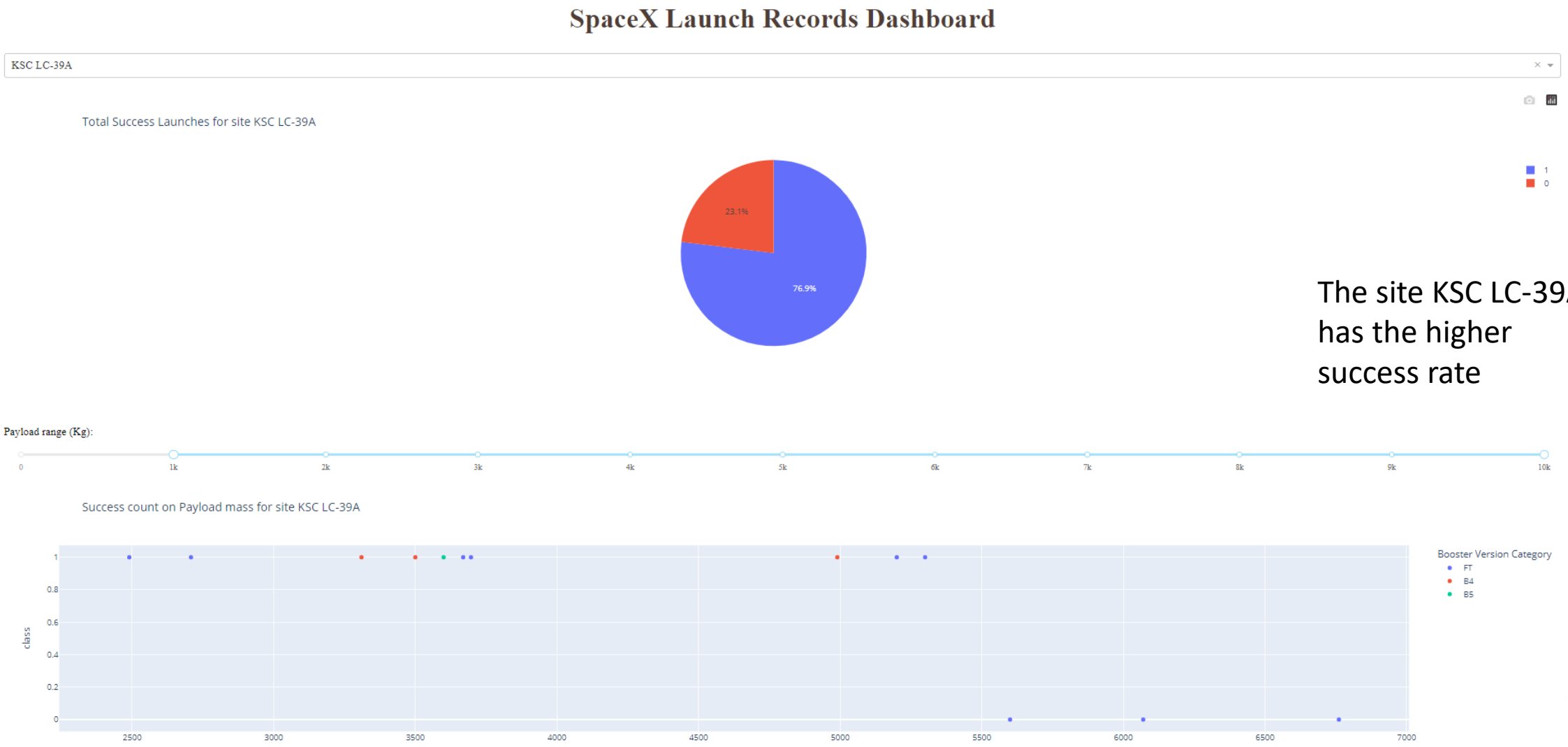
Build a Dashboard with Plotly Dash – CACFLS LC-40



Build a Dashboard with Plotly Dash – VAFB SLC-4E

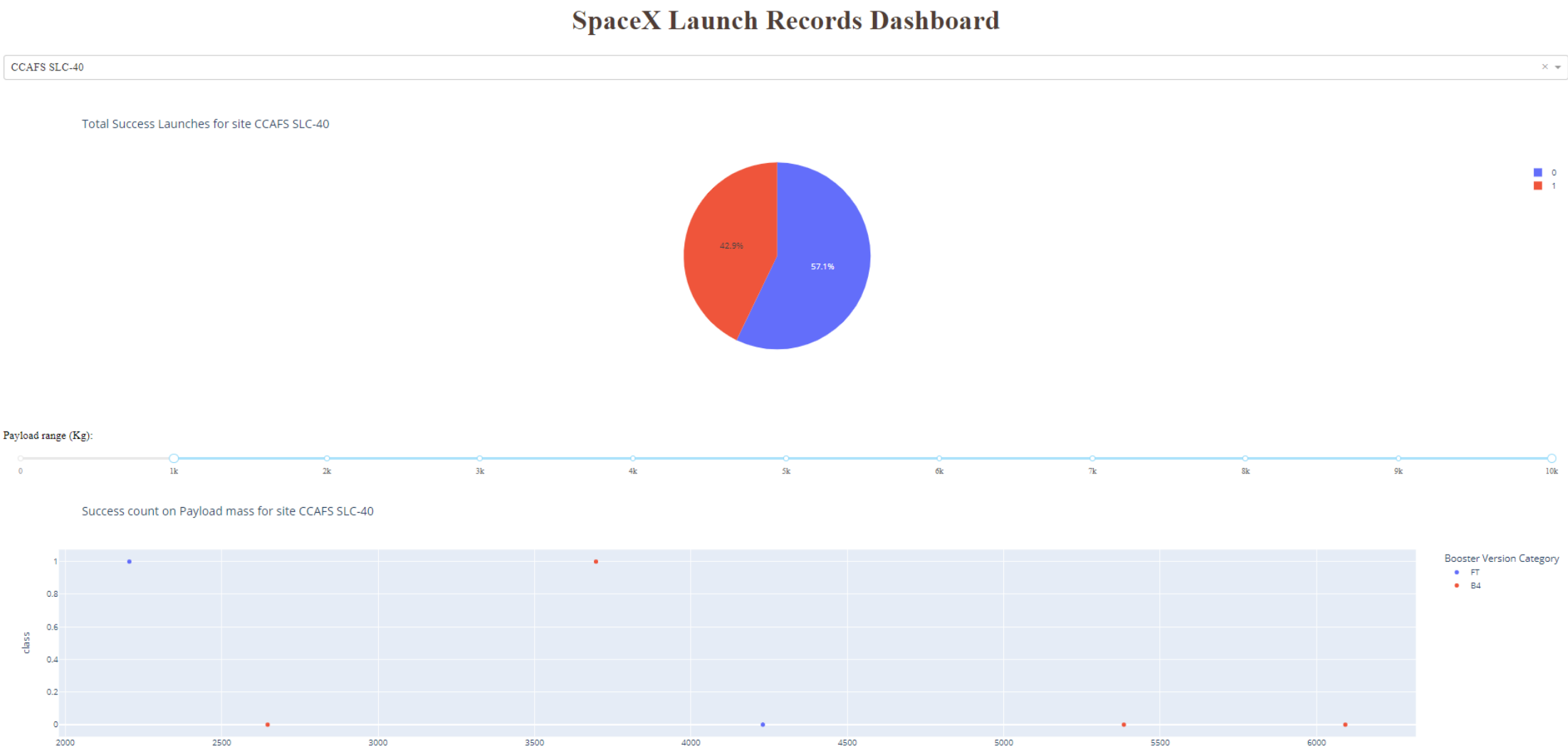


Build a Dashboard with Plotly Dash – KSC LC-39A



The site KSC LC-39A has the higher success rate

Build a Dashboard with Plotly Dash – CCAFS SLC-40



Section 5

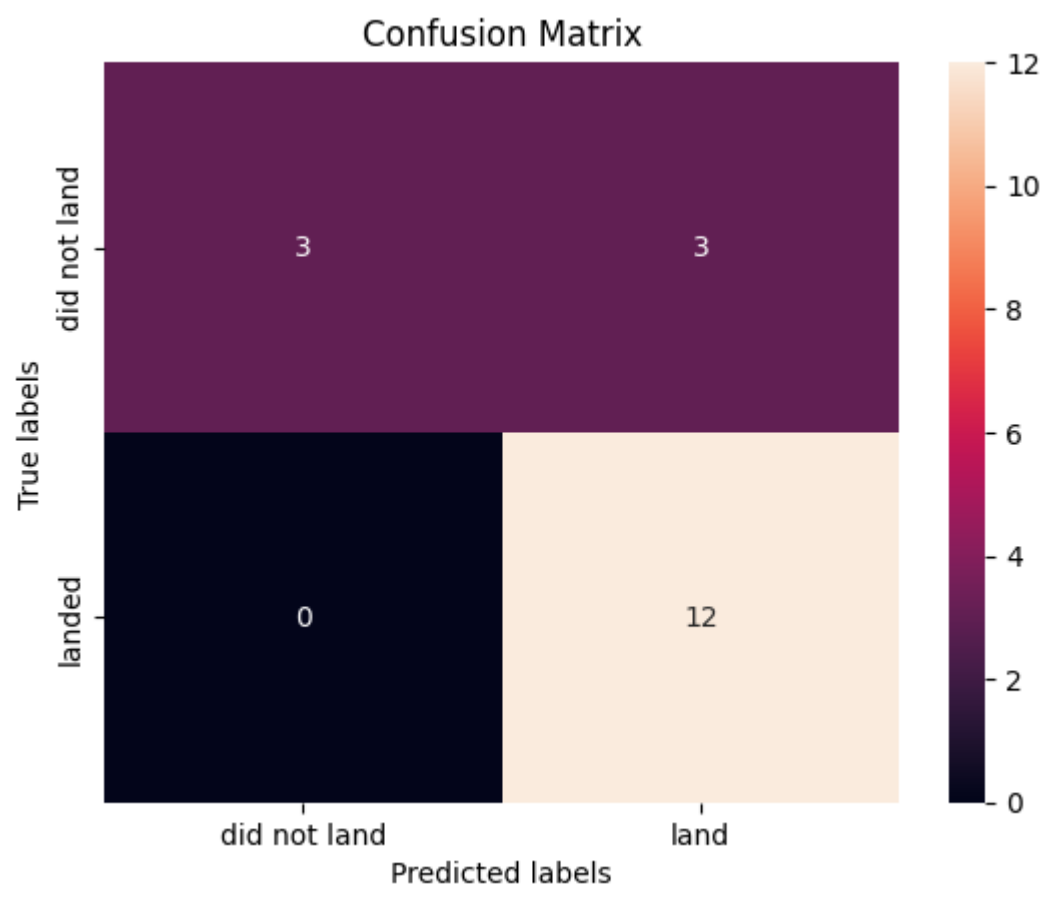
Predictive Analysis (Classification)

Classification Accuracy

Method	Test Data Accuracy
Logistic_Reg	0.833333
SVM	0.833333
Decision Tree	0.666667
KNN	0.833333

- Logistic regression, SVM and KNN have the same data accuracy (0.83) and the lowest one is decision tree with 0.66

Confusion Matrix



- The major problem are the false positives

Conclusions

- The data shows higher success rate in the site KSC LC-39A
- The launch sites are closer to airport and load facilities, reducing the land costs.
- The models used allow the data scientist to identify what is the most suitable, in this case further analysis is required since three models scores the same.
- Since the first success launch (Dec 22, 2015) there are a consistent success rate across the available date to the future.
- The false positives are the major problem to analyze the data.
- The ES-L1, GEO, HEO and SSO are the most successful orbits, furthermore it is recommended to continue using these orbits unless there are external restrictions.

Thank you!