IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jihang Jiang
2022/7/9

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data Collection with API
  - Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- **Summary of all results**
  - Exploratory Data Analysis
  - Interactive Analytics
  - Predictive Analysis

Introduction

- Project background and context

Space X 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 Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we 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.

- Problems you want to find answers

Whether we can create a machine learning pipeline to predict if the first stage will land successfully?

Section 1

Methodology

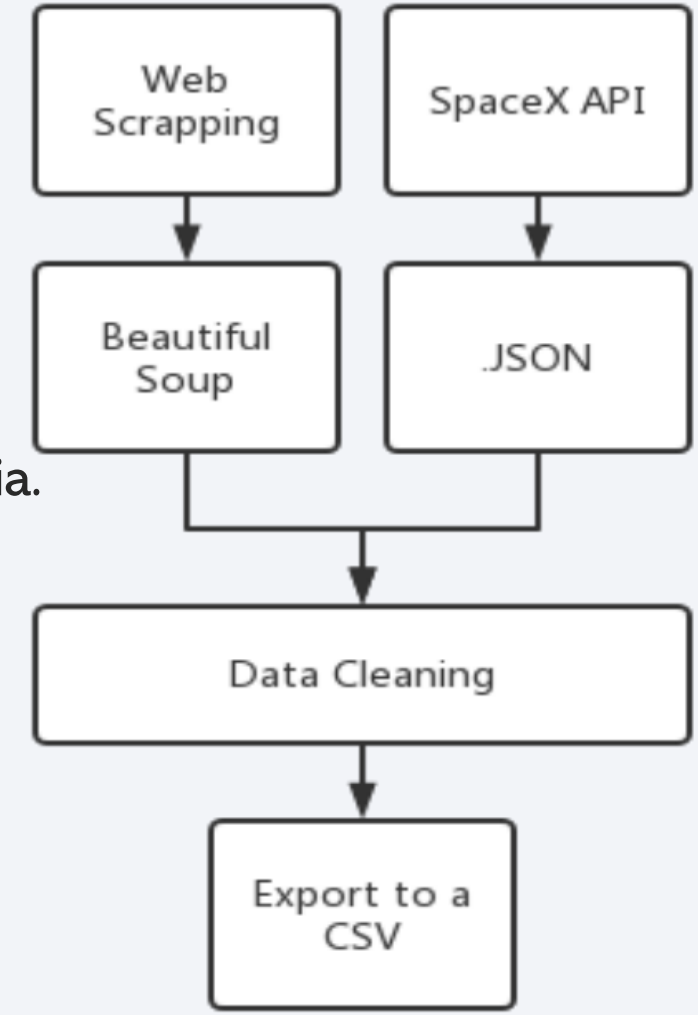
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - web scraping from Wikipedia.
- Perform data wrangling
  - One-hot encoding data was dropped nan values and applied to categorical features then prepared for machine learning model.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build Logistic Regression, SVM, KNN, Decision Tree model then find the best classifier

Data Collection

- Describe how data sets were collected.
- Get SpaceX Launch Data from SpaceX API.
- Using .json() function decoded the response content as a Json.
- Dealing with missing values.
- Get SpaceX Launch Data with Web Scrapping from Wikipedia.
- Using Beautiful Soup to get the launch data then convert the table to dataframe.



Data Collection – SpaceX API

- Data collection with SpaceX API

- GitHub URL: https://github.com/jhjiang1119/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb

Data Collection - Scraping

- Web scraping process

1.request the HTML page from URL and get a response object

```
In [4]: static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
```

```
In [5]: # use requests.get() method with the provided static_url
        # assign the response to a object
        response = requests.get(static_url).text
```

2.Create a BeautifulSoup object from the HTML response

```
In [6]: # Use BeautifulSoup() to create a BeautifulSoup object from a response text content
        soup = BeautifulSoup(response, 'html.parser')
```

Print the page title to verify if the `BeautifulSoup` object was created properly

```
In [7]: # Use soup.title attribute
        print(soup.title)
```

```
<title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

3.Extract all column names from the HTML table header

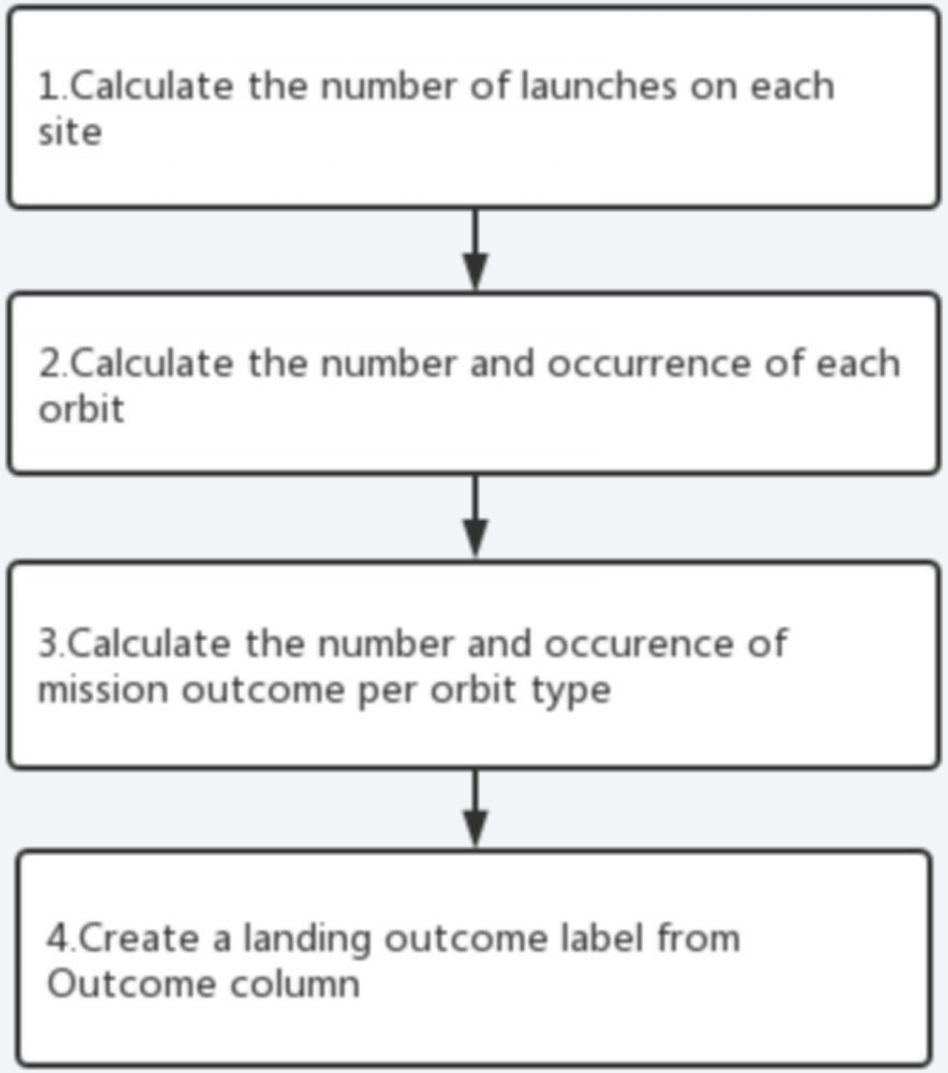
4. Create a data frame by parsing the launch HTML tables

- GitHub URL: https://github.com/jhjiang1119/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping%20(1).ipynb

Data Wrangling

- Data wrangling process

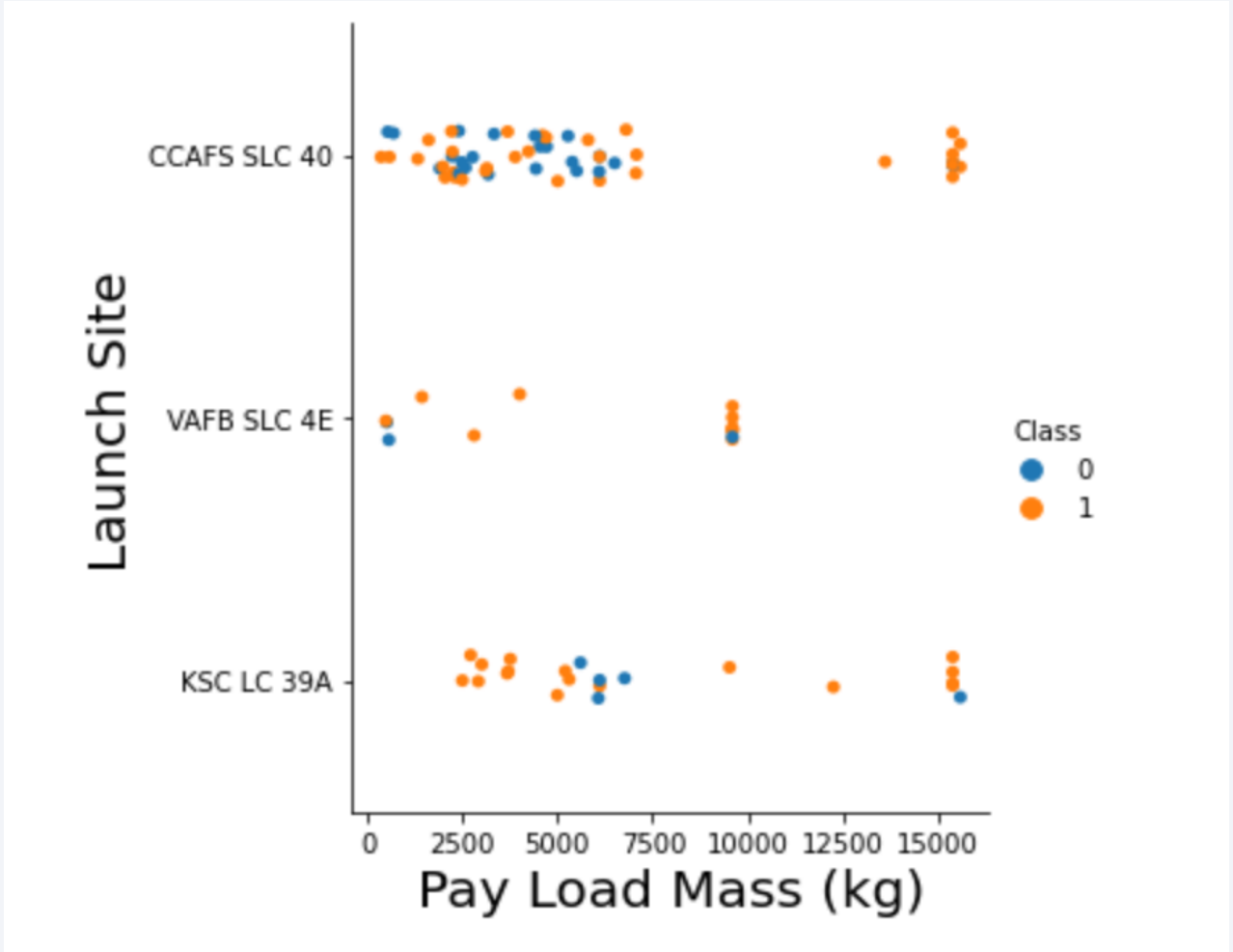
- GitHub URL: https://github.com/jhjiang1119/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb



EDA with Data Visualization

- We can observe if there is any relationship between launch sites and their payload mass.

- GitHub URL:
- https://github.com/jhjiang1119/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb



EDA with SQL

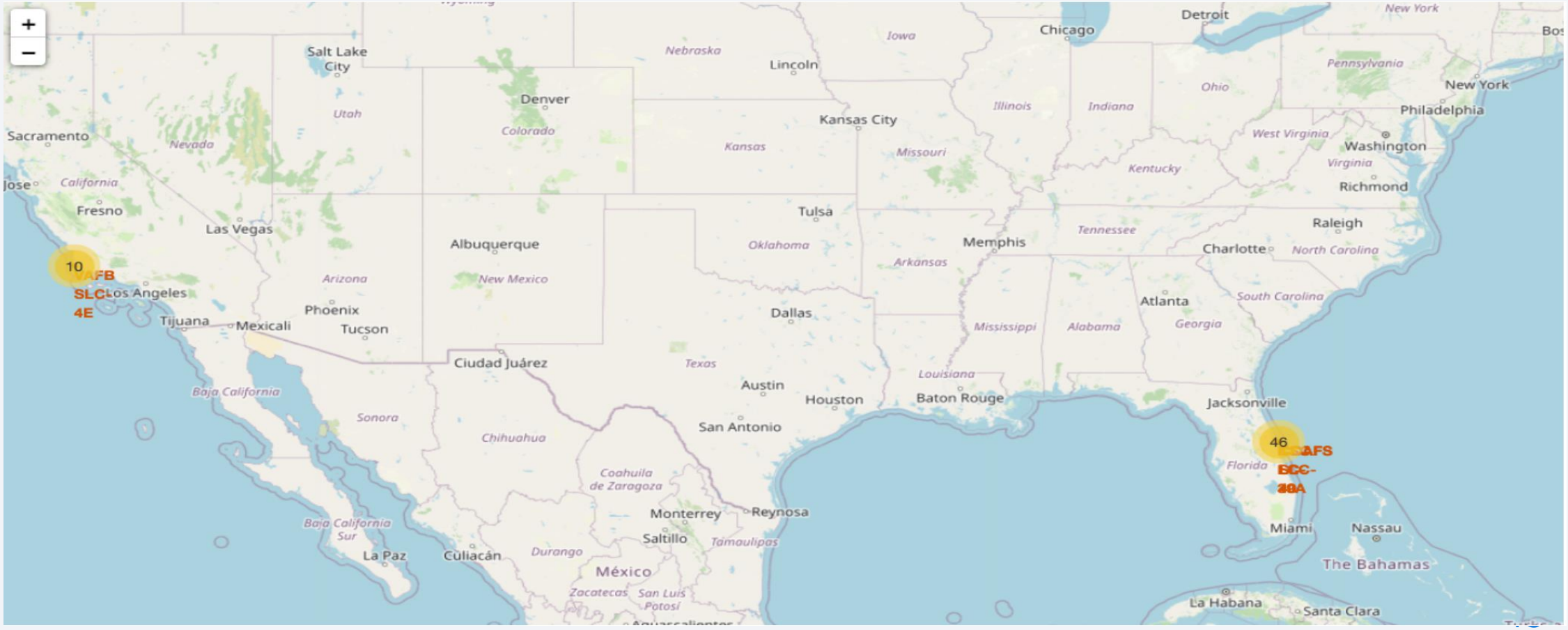
- SQL queries performed

- 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. Use a 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 successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

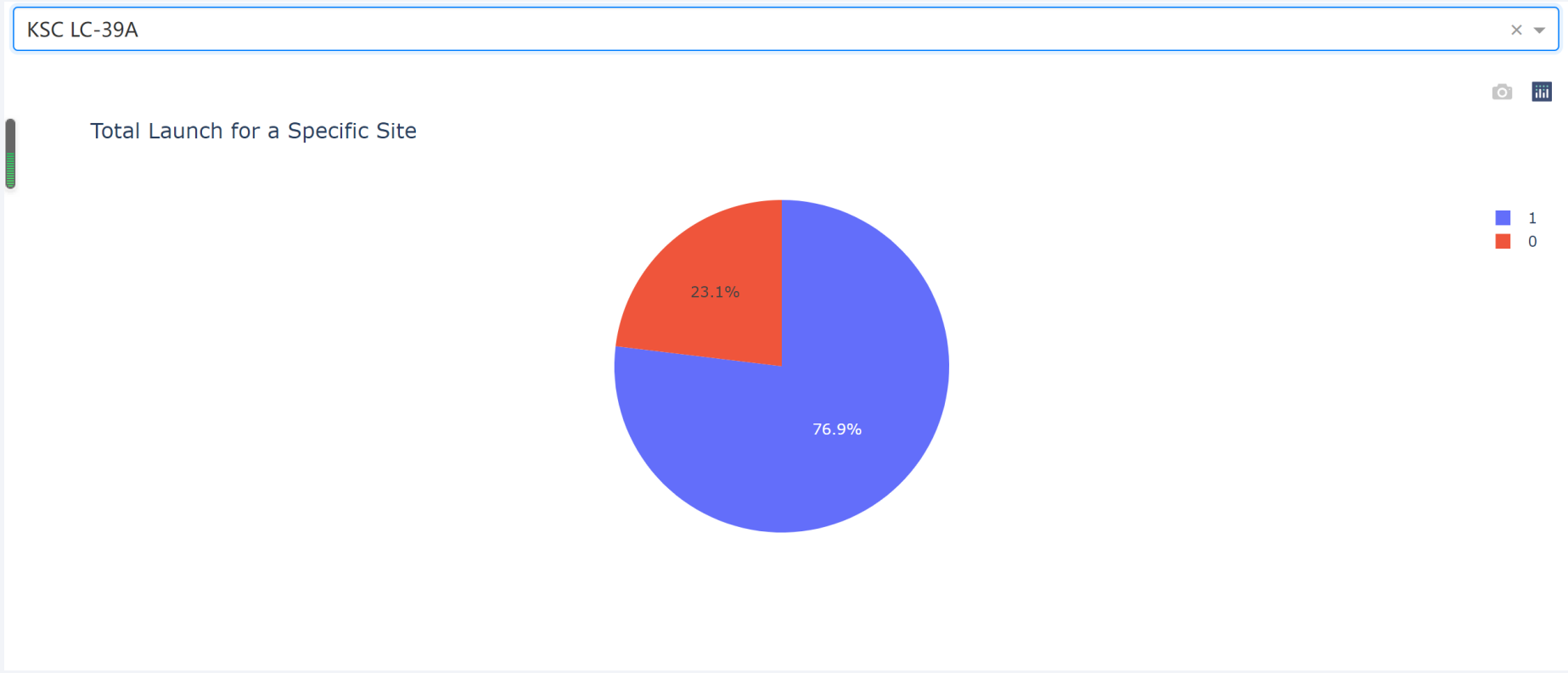
- GitHub URL:https://github.com/jhjiang1119/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

- GitHub URL: https://github.com/jhjiang1119/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb
- Marker clusters can be a good way to simplify a map containing many markers having the same coordinate.



Build a Dashboard with Plotly Dash

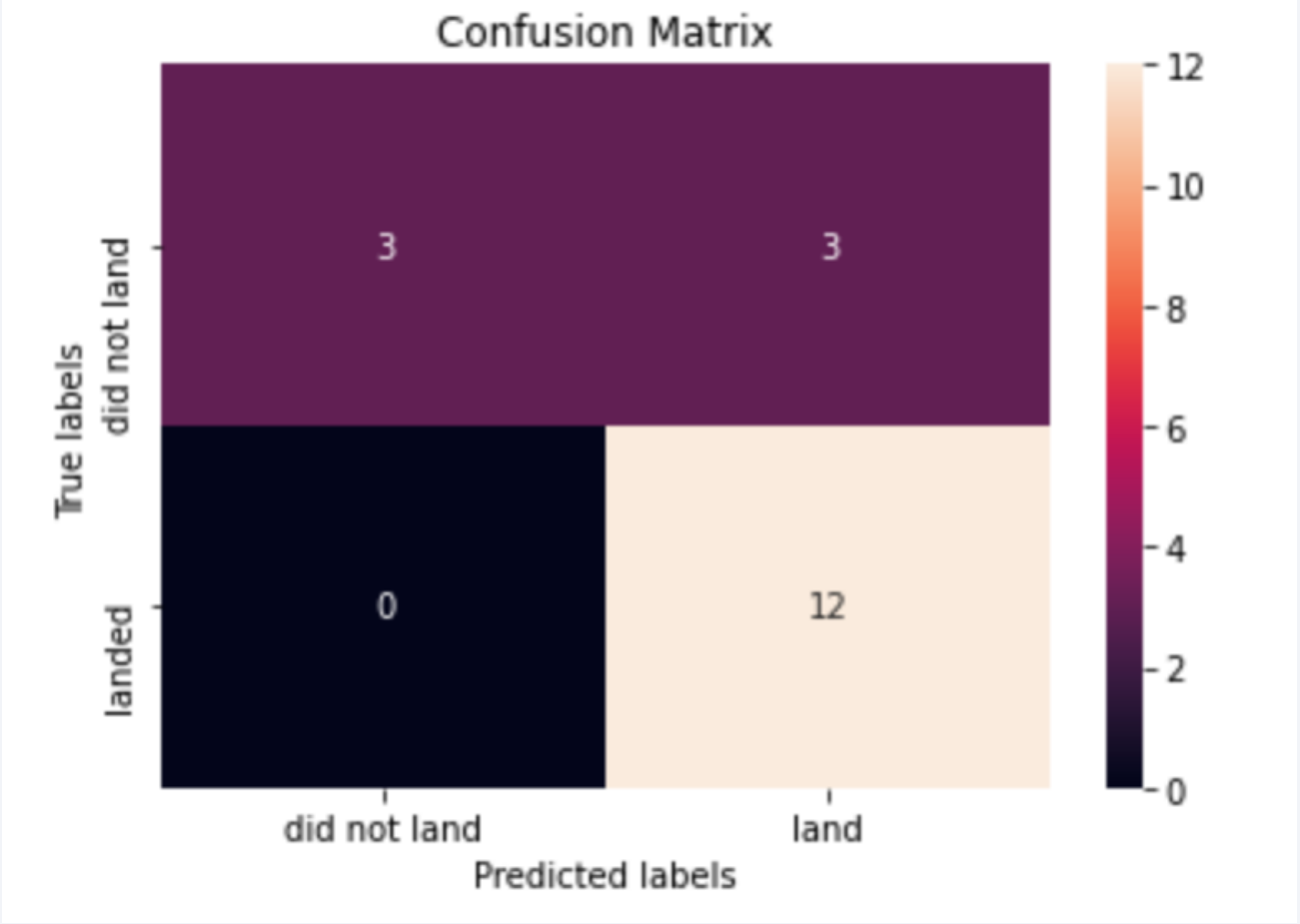


- GitHub URL:https://github.com/jhjiang1119/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Logistic Regression model achieved accuracy at 83.33%

- GitHub URL:https://github.com/jhjiang1119/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

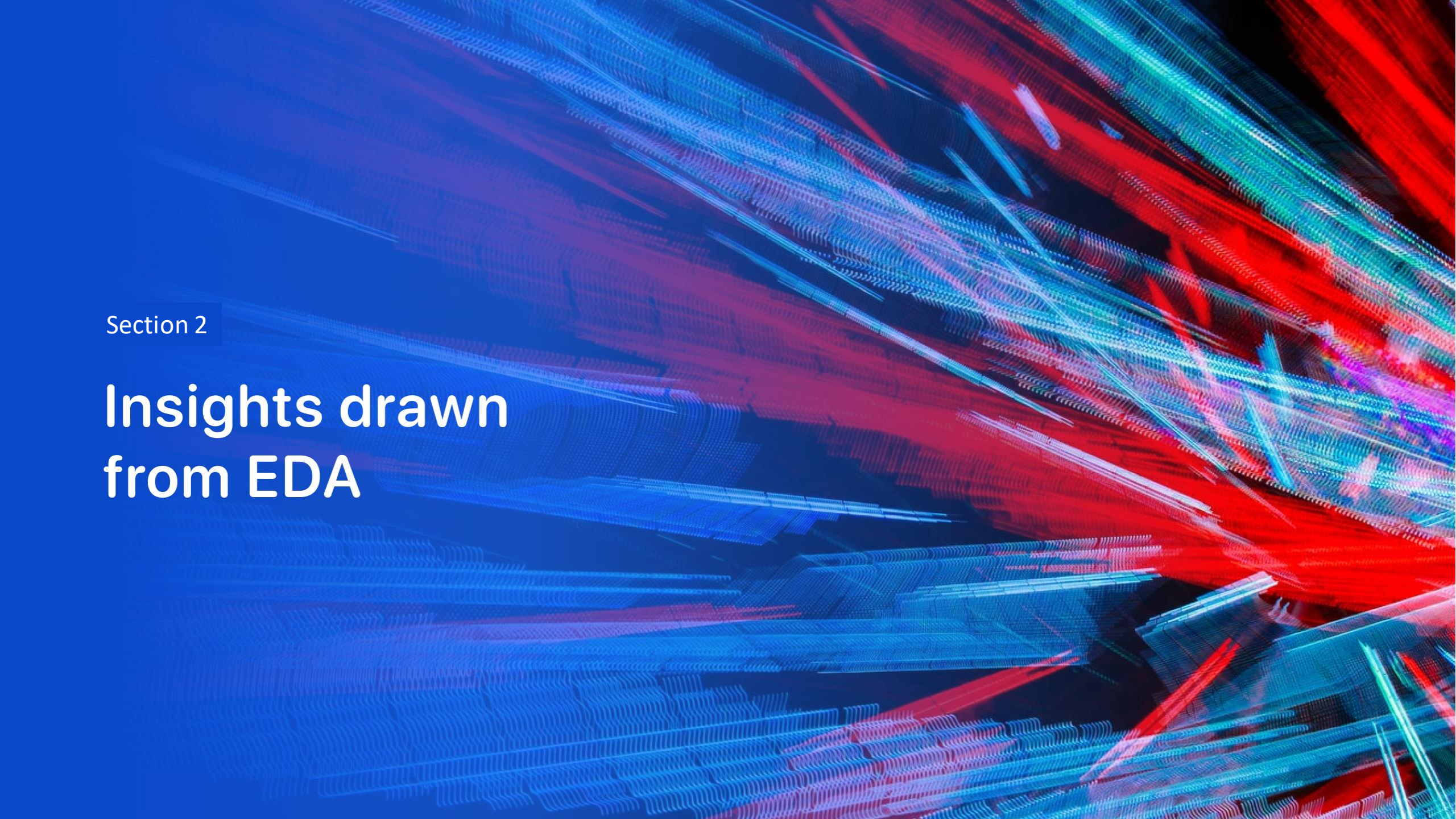


Results

- Logistic Regression, KNN, SVM performed best
- Low weighted payloads perform better than heavier payloads
- F9 Booster verson TF has the highest launch success rate

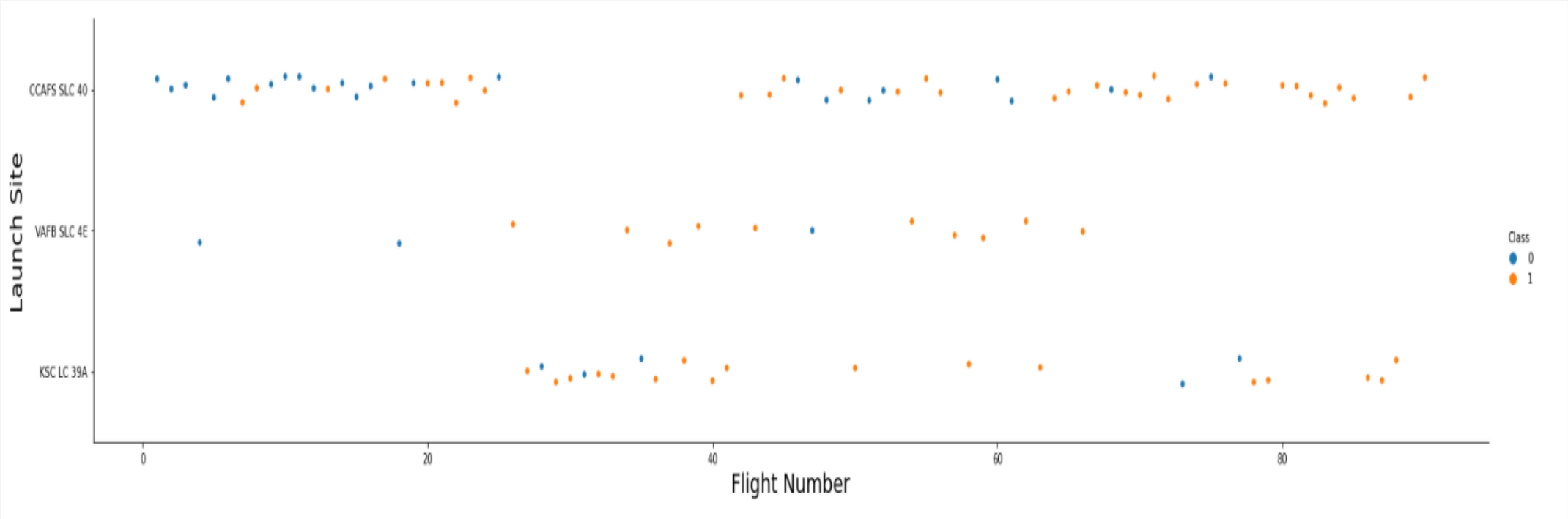
Section 2

Insights drawn from EDA



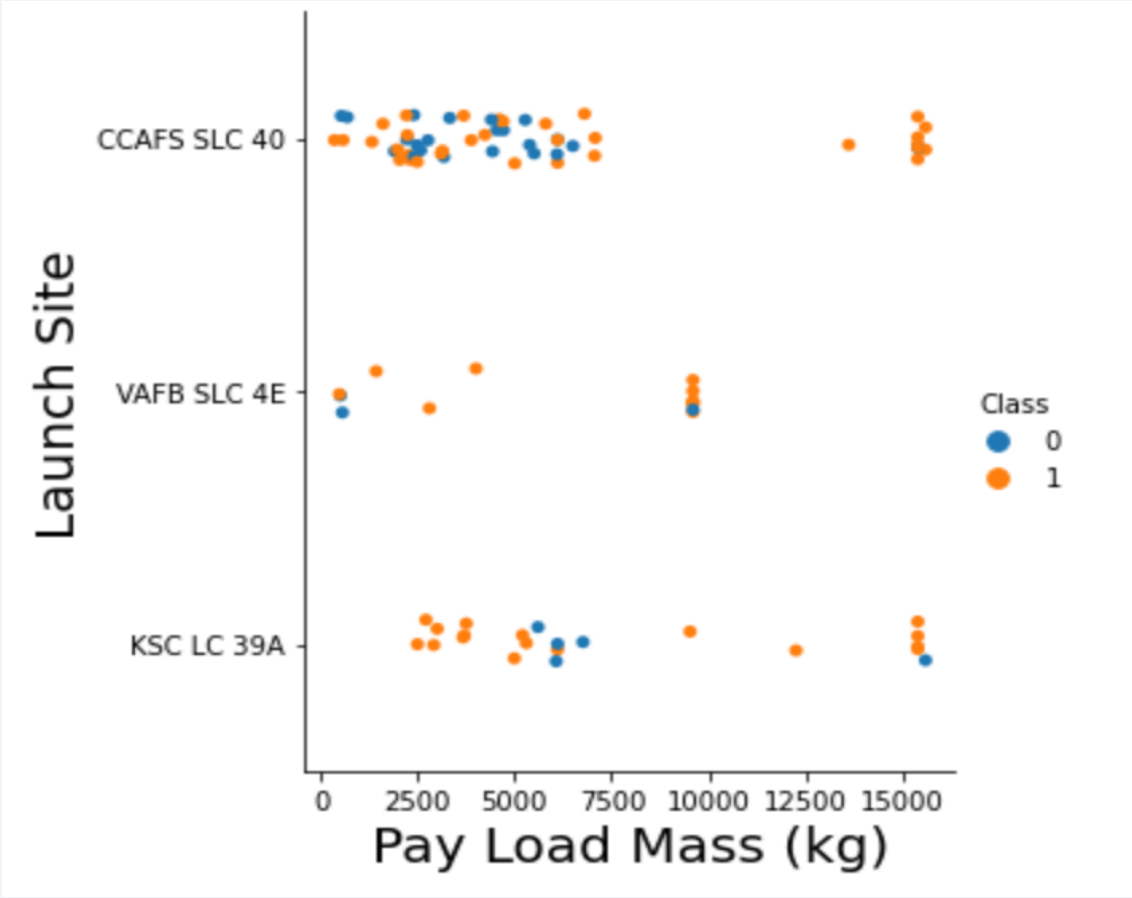
Flight Number vs. Launch Site

- Launches from CCAFS SLC 40 are the highest than other sites



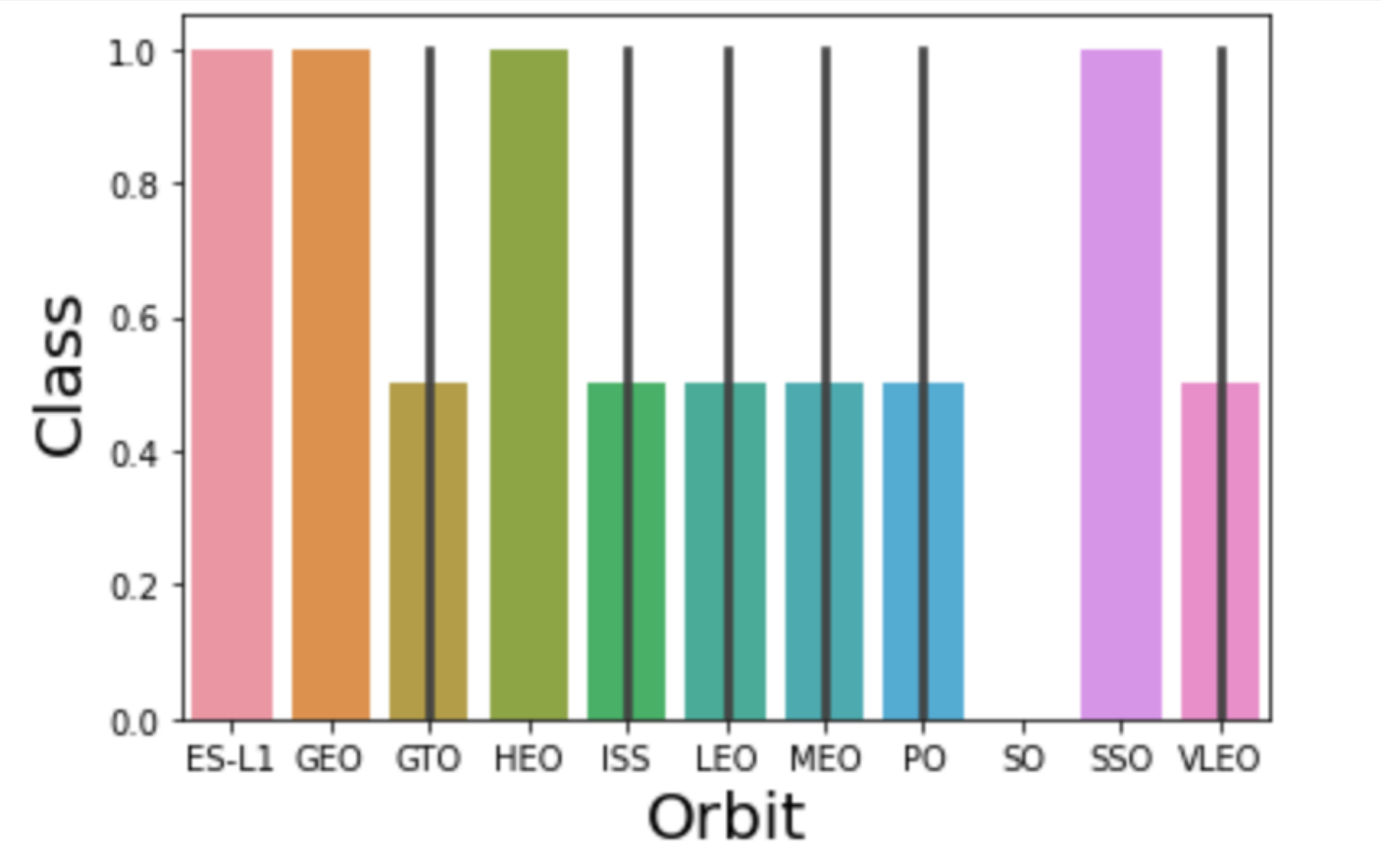
Payload vs. Launch Site

Launches with less payload mass are much more than launches with heavy payload mass.



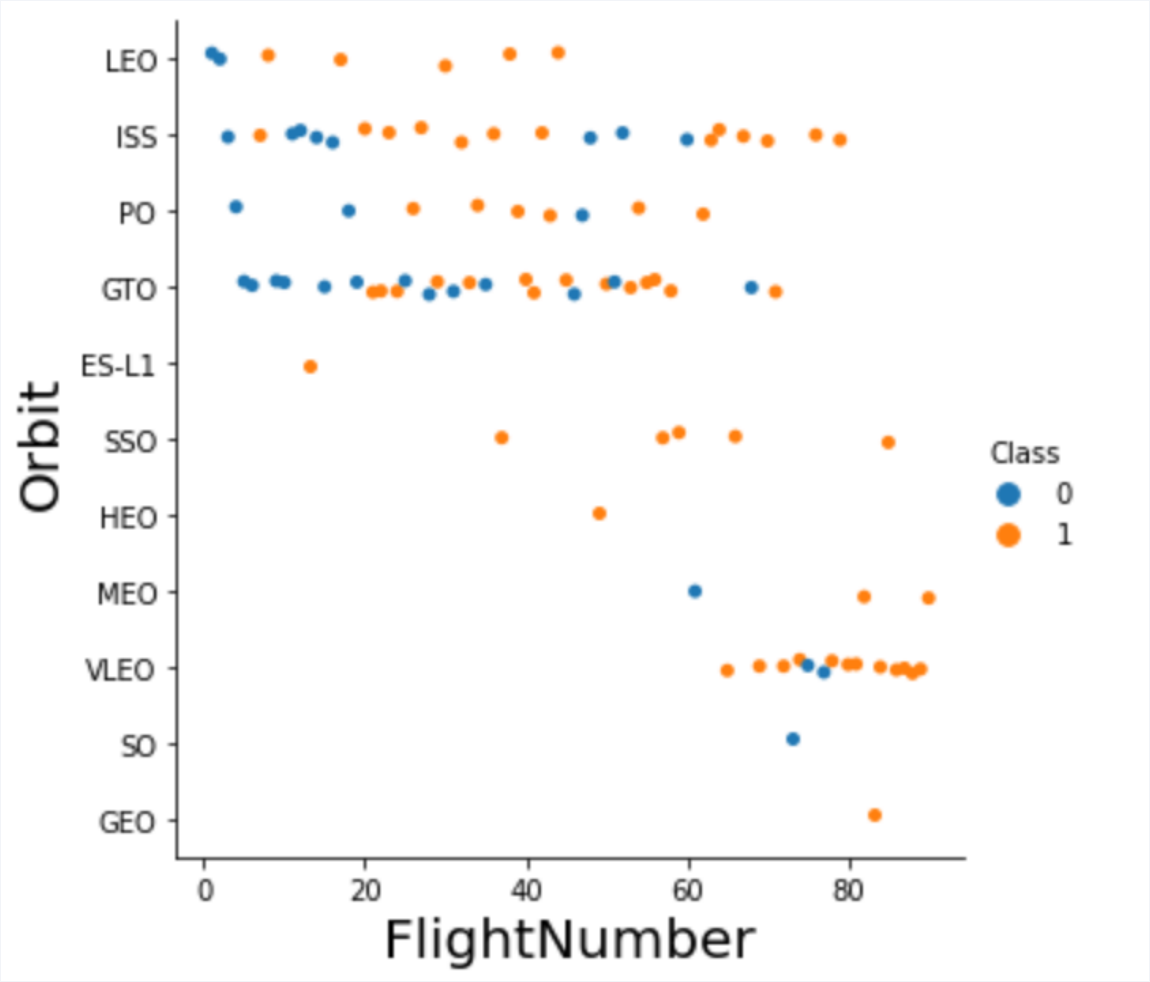
Success Rate vs. Orbit Type

- Orbit type ES-L1, GEO, HEO, SSO have high sucess rate.



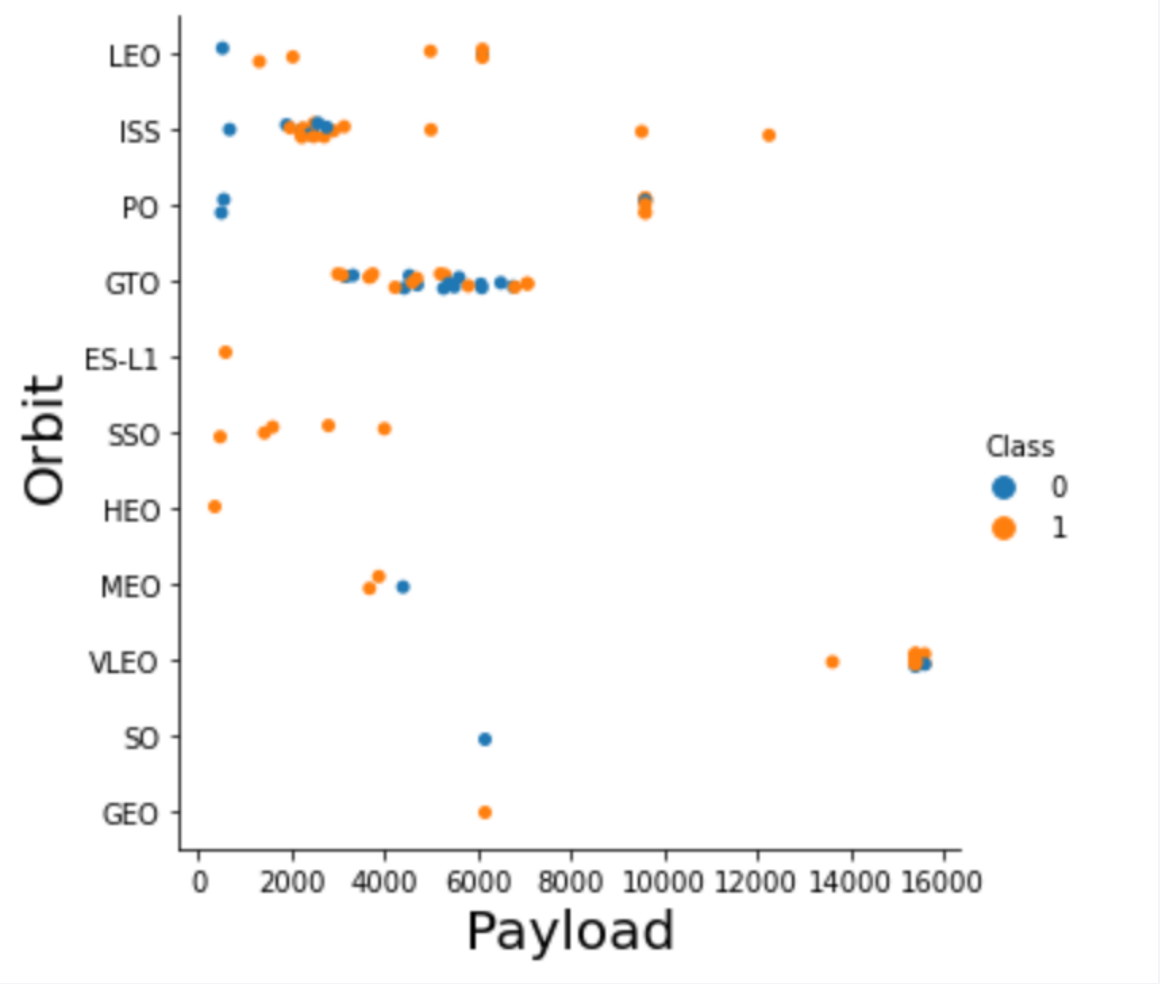
Flight Number vs. Orbit Type

- There is no good relationship between flight number and the orbit.



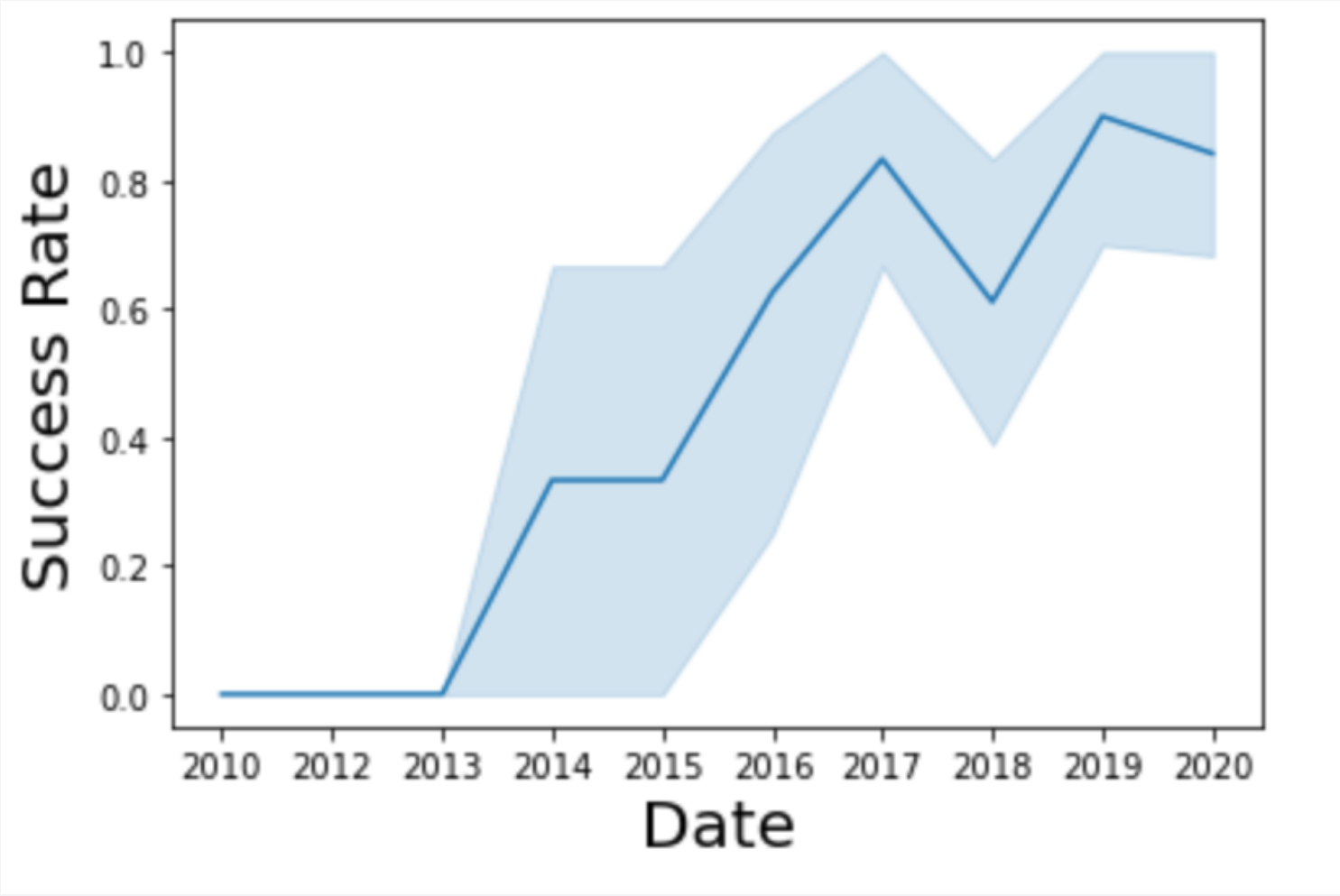
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.



Launch Success Yearly Trend

- The sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- Find the names of the unique launch sites

%sql select distinct(LAUNCH_SITE) from SPACEXTBL

[7]:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- %sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landir _Outcon
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failu (parachut
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	Failu (parachut
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N attem

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'

sum(PAYLOAD_MASS__KG_)
45596

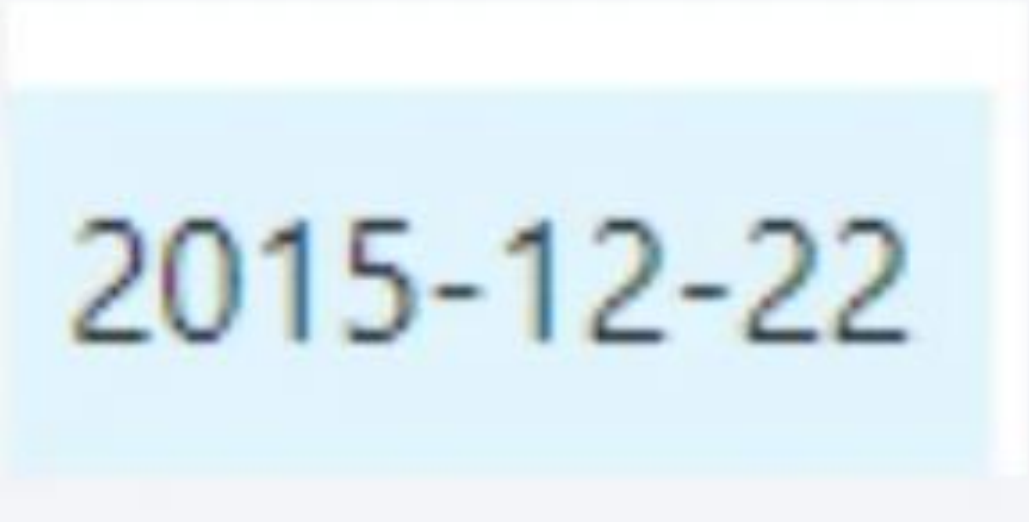
Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- %sql select min(DATE) from SPACEXTBL where Landing__Outcome = 'Success (ground pad)'



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 LANDING__OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- %sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or MISSION_OUTCOME = 'Failure (in flight)'

count(MISSION_OUTCOME)
99

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- %sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)

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 EXTRACT(MONTH, SELECT min(DATE) from SPACEXTBL where Landing__Outcome = 'Success (ground pad)')

time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)
05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
		CCAFS LC-				SKY Perfect JSAT		

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 * from SPACEXTBL where Landing__Outcome like 'Success%' and (DATE between '2010-06-04' and '2017-03-20') order by date desc

2016-05-27	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
2016-05-06	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

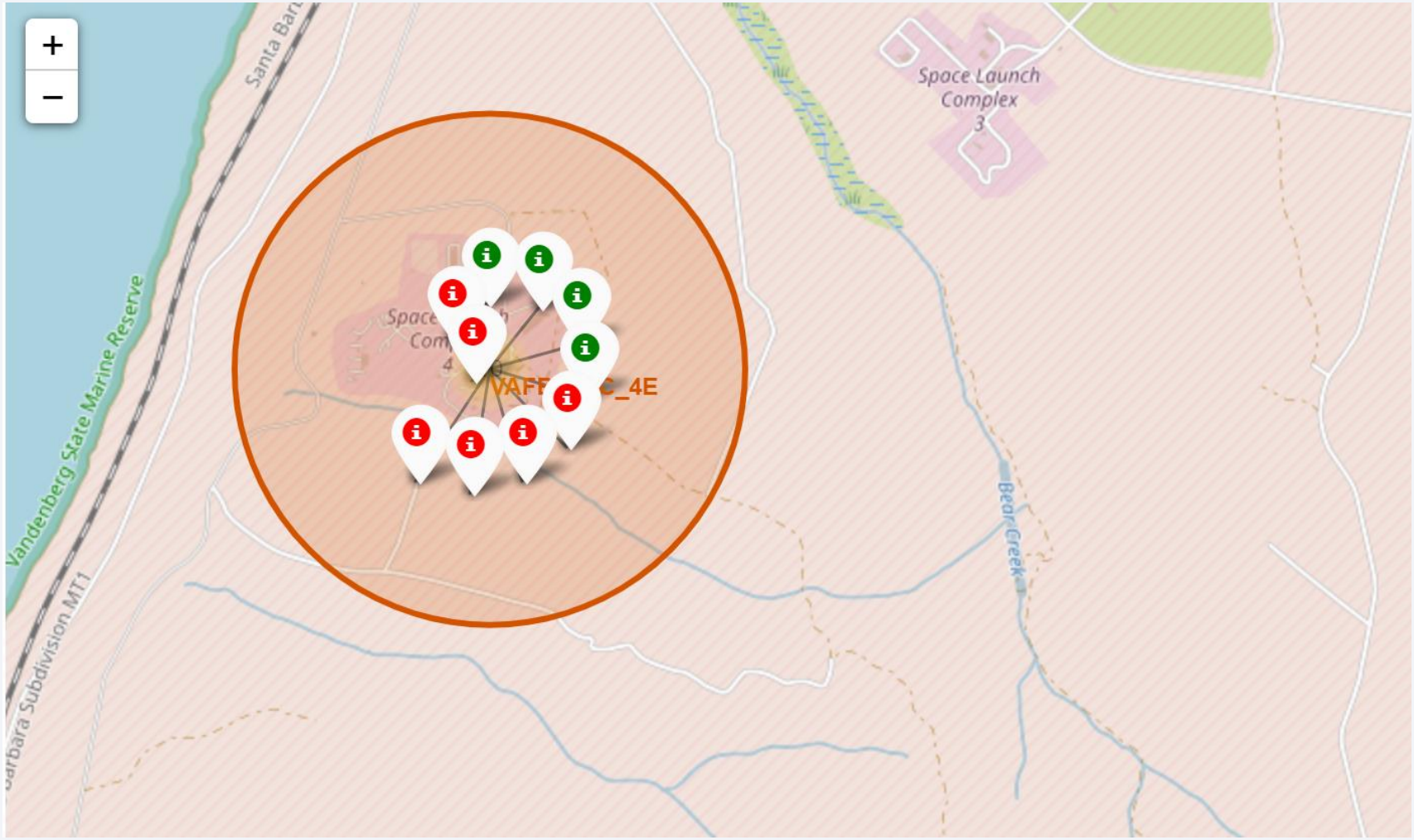

Section 3
Launch Sites
Proximities Analysis

All Launch Site Marker

Use folium.Circle to add a highlighted circle area with a text label on a specific coordinate

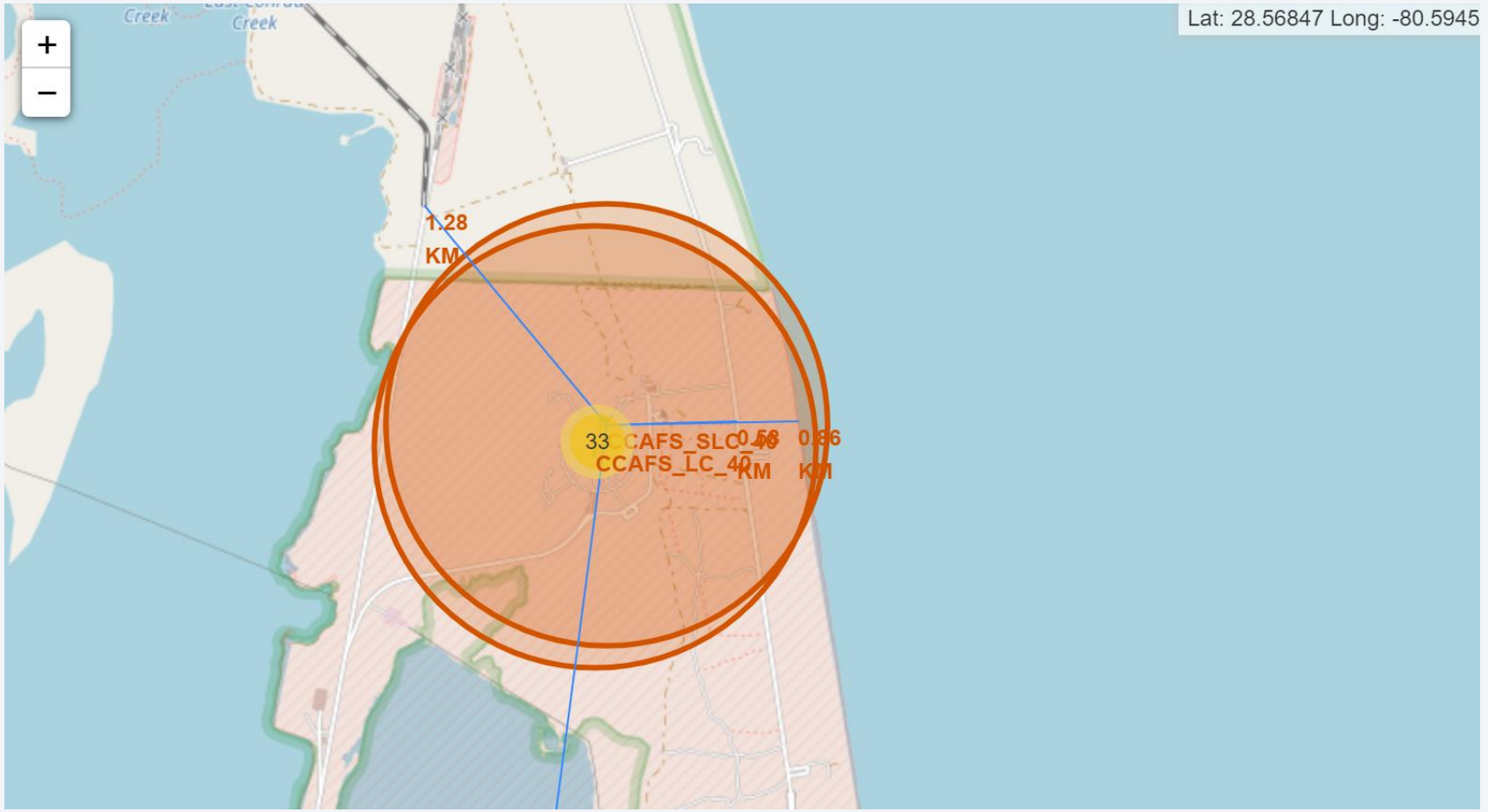
Mark the success/failed launches for each site

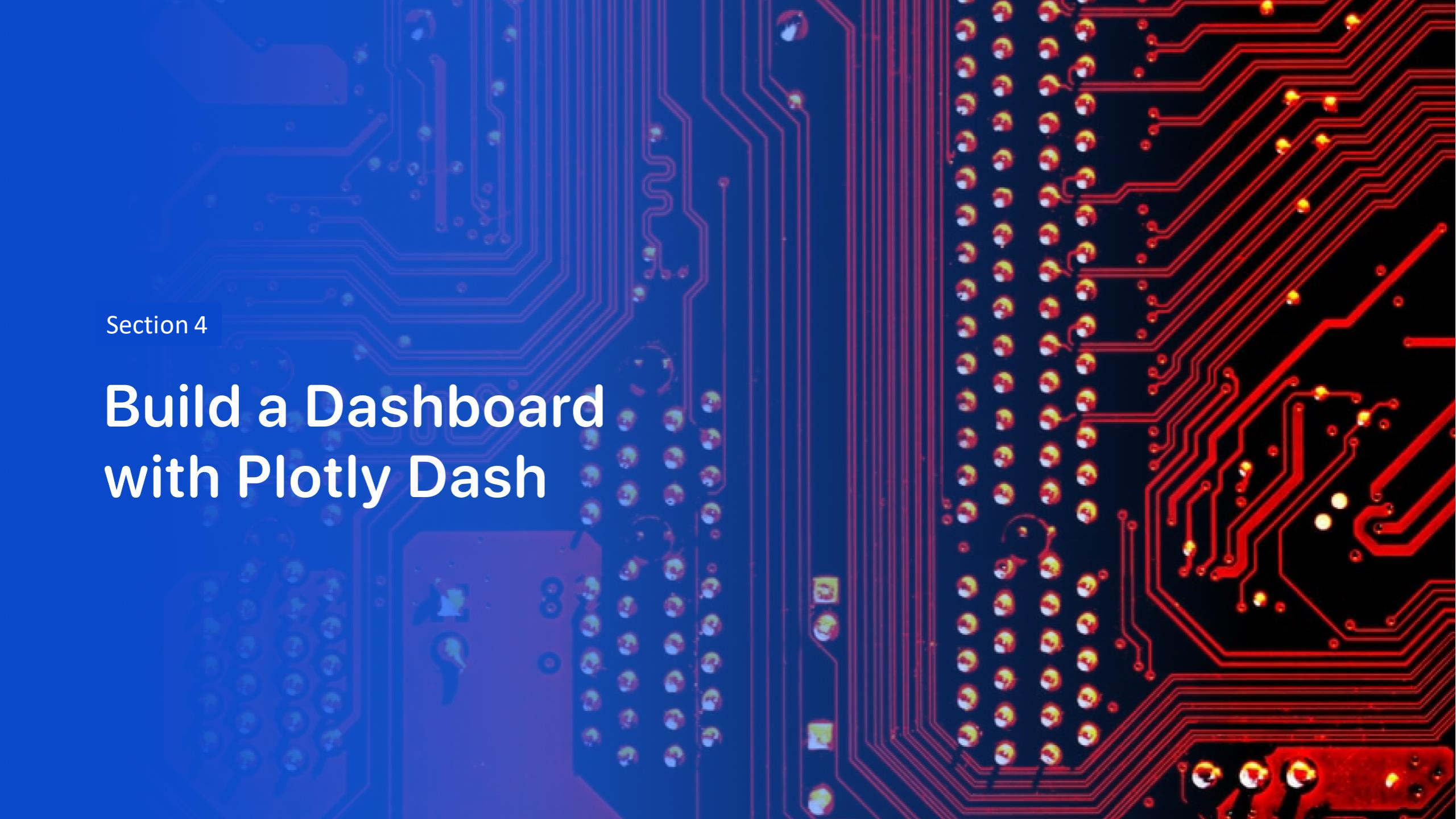
- For each launch result in spacex_df data frame, add a folium.Marker to marker_cluster



distances between a launch site to its proximities

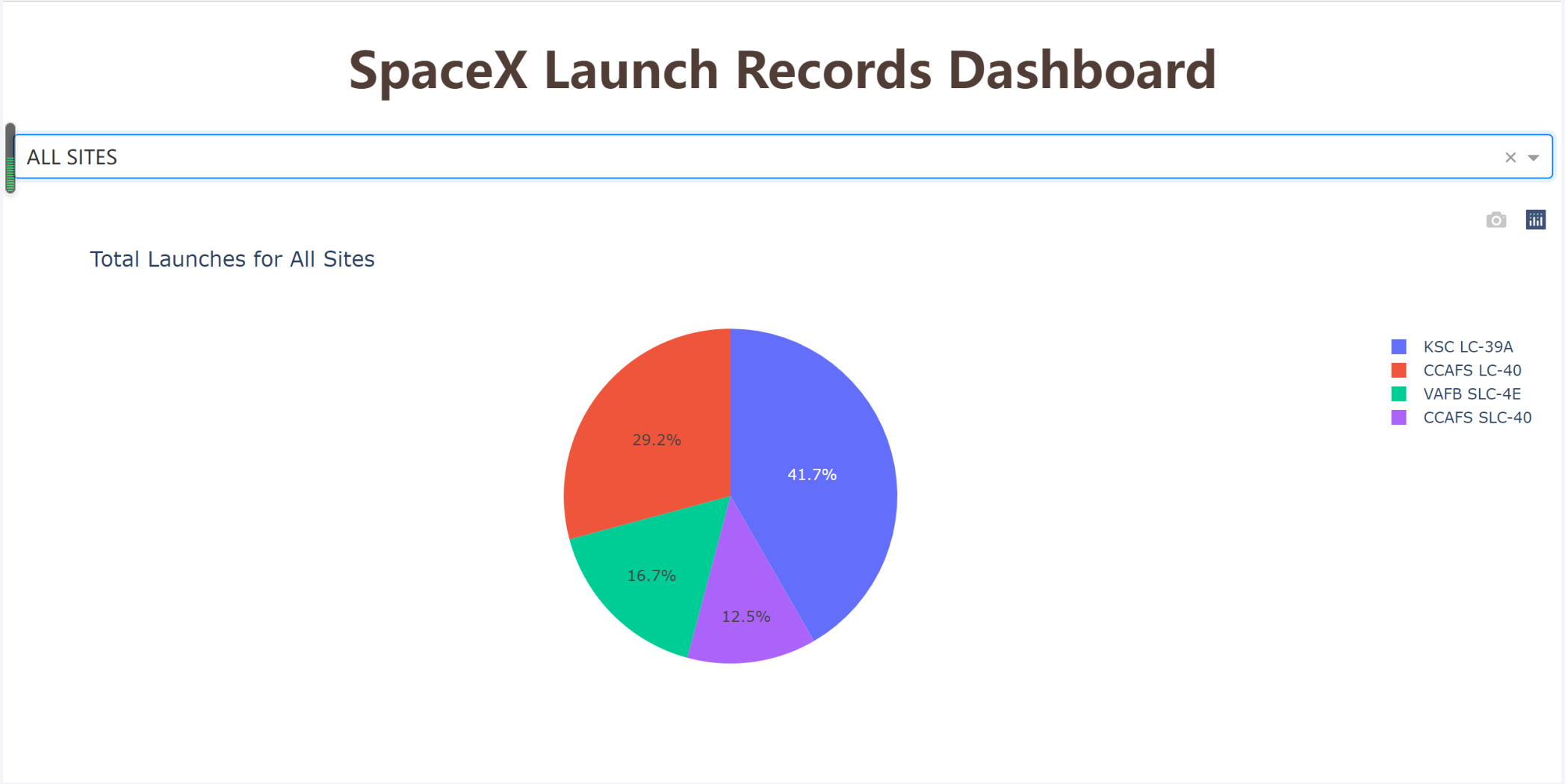
- Create a `folium.PolyLine` object using the coastline coordinates and launch site coordinate




Section 4
Build a Dashboard with Plotly Dash

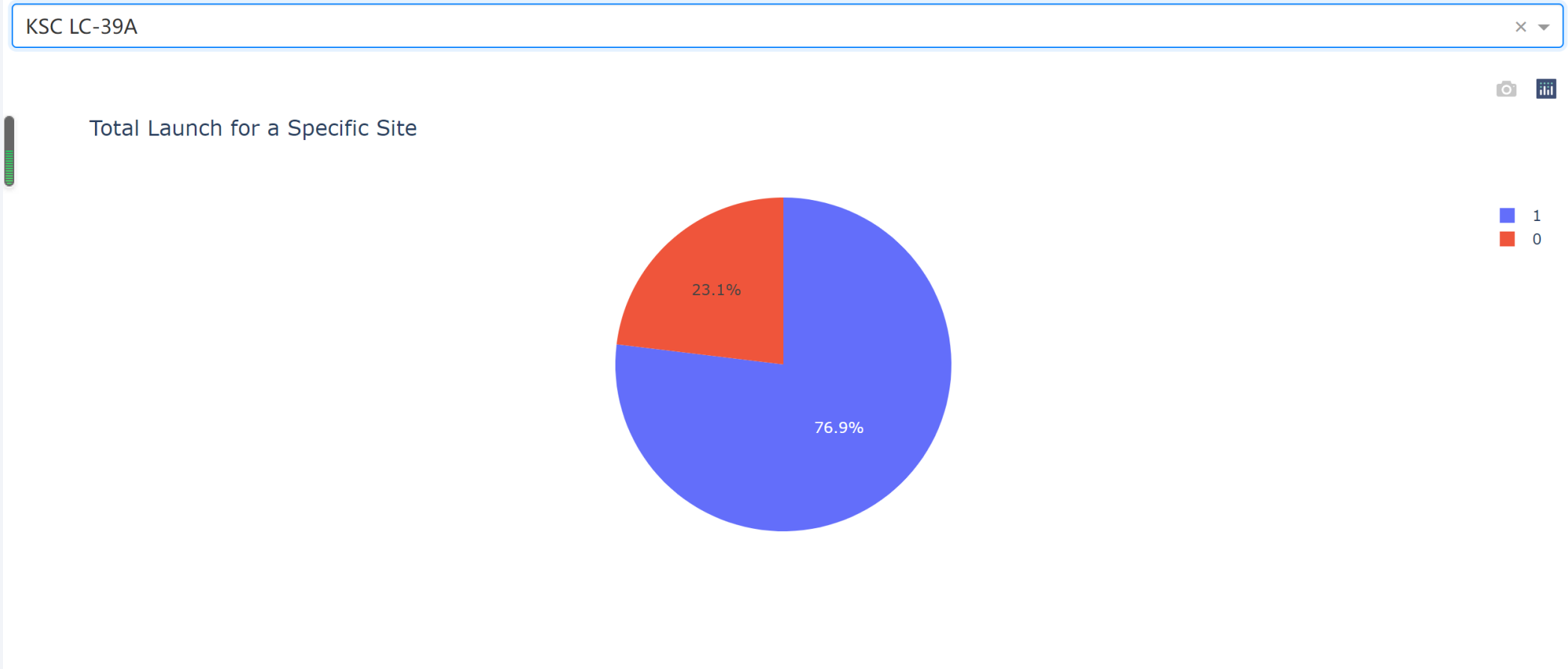
Success Launches by All Sites

- KSC LC-39A had the most successful launches from all launch sites



Highest rate Launch Site

- KSC LC-39A had 76.9% success launch rate



Booster Version with highest success rate

- Booster Version FT had the best success rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy

LR, KNN, SVM model have the highest classification accuracy at 83.33%

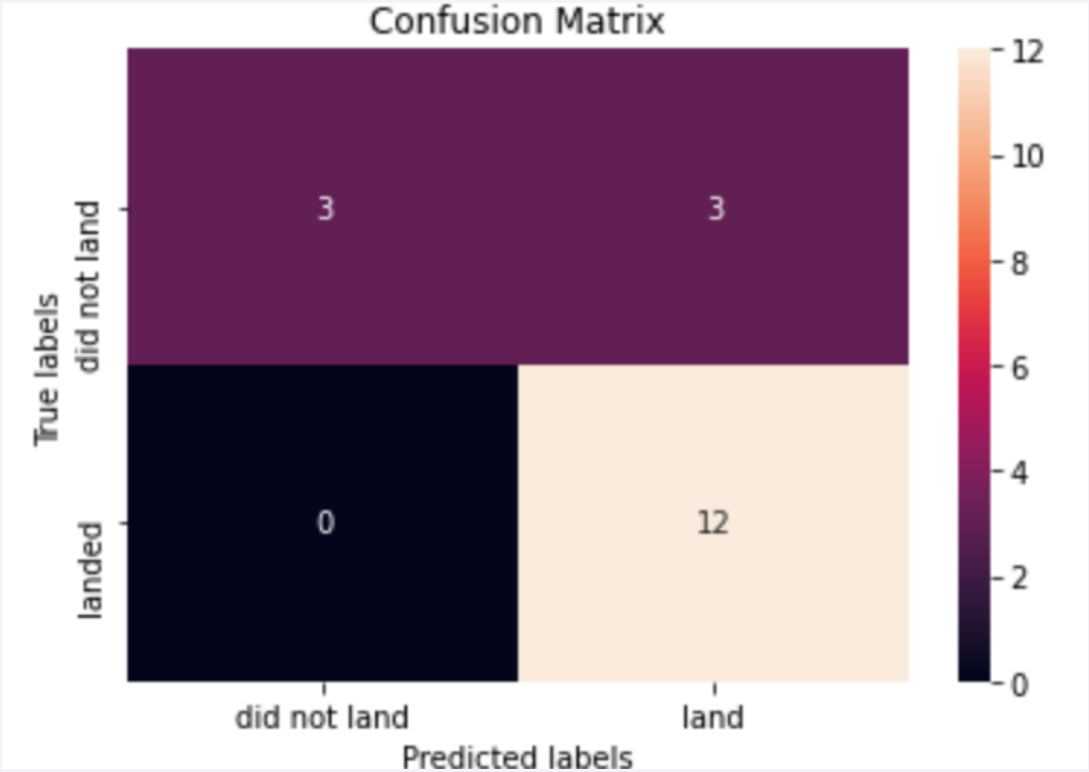
[37]:
```
predictors = [knn_cv, svm_cv, logreg_cv, tree_cv]
best_predictor = ""
best_result = 0
for predictor in predictors:

    predictor.score(X_test, Y_test)
print("best predictor:",predictor.score(X_test, Y_test))
```

```
best predictor: 0.8333333333333334
```

Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes.



Conclusions

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Low weighted payloads perform better than heavier payloads
- F9 Booster verson TF has the highest launch success rate
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The LR, KNN, SVM is the best machine learning algorithm for this task.

Thank you!