IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

S. H. Jonathan TONG
7 May 2023

Outline

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

Executive Summary

- Summary of methodologies
  1. Data Collection from public source
  2. Data Wrangling
  3. Exploratory Data Analysis
     - Visualization and SQL
  4. Predictive Analysis (Classification)
- Summary of all results
  1. Exploratory Data Analysis result
  2. Predictive analysis results

Introduction

Project background

SpaceX is an aerospace manufacturer; they launch rockets frequently. Basically, launching can separate into 2 parts. Stage two helps bring the payload to orbit, but most of the work is done by the stage one. This stage does most of the work and is much larger than the stage two. SpaceX's Falcon 9 Can recover the first stage. Sometimes the first stage does not land. Sometimes it will crash in this clip.

Problems

- Can we gather information that we need from public source

- Can we train a model and use public information to predict if SpaceX will reuse the first stage.

Section 1

Methodology

Methodology

Executive Summary

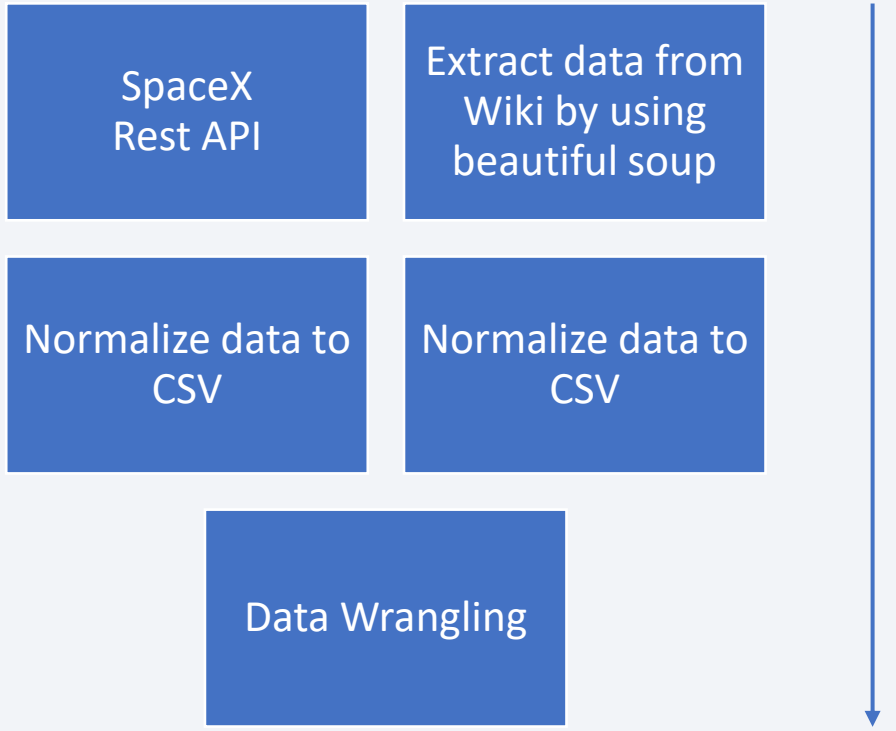
- Data collection methodology:
  - Data was collected from SpaceX API (https://api.spacexdata.com/v4/rockets/)
  - Another part was collected from Wikipedia by using web scraping
- Perform data wrangling
  - Filtering the data (include Falcon 9 only)
  - Dealing with missing value
  - Transforming data to binary classification for further analysis
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - By using 4 different classification models (logistic regression, SVM, decision tree and KNN) to determine the best prediction

Data Collection

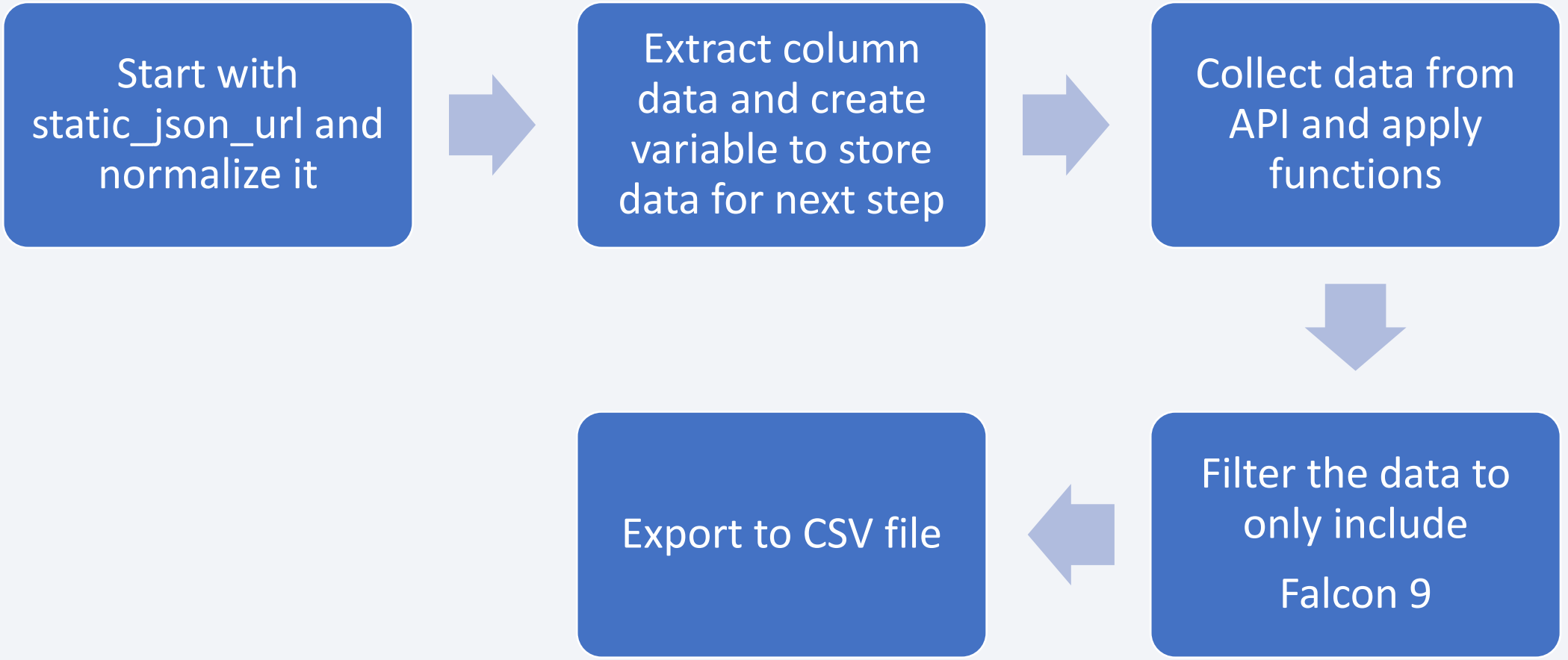
- Data sets were collected from two different source:

1. SpaceX REST API
   The SpaceX REST API endpoints provided data about rocket used, payload and landing outcome etc.

2. Web Scraping
   Obtaining Falcon 9 Launch data by web scraping from Wikipedia

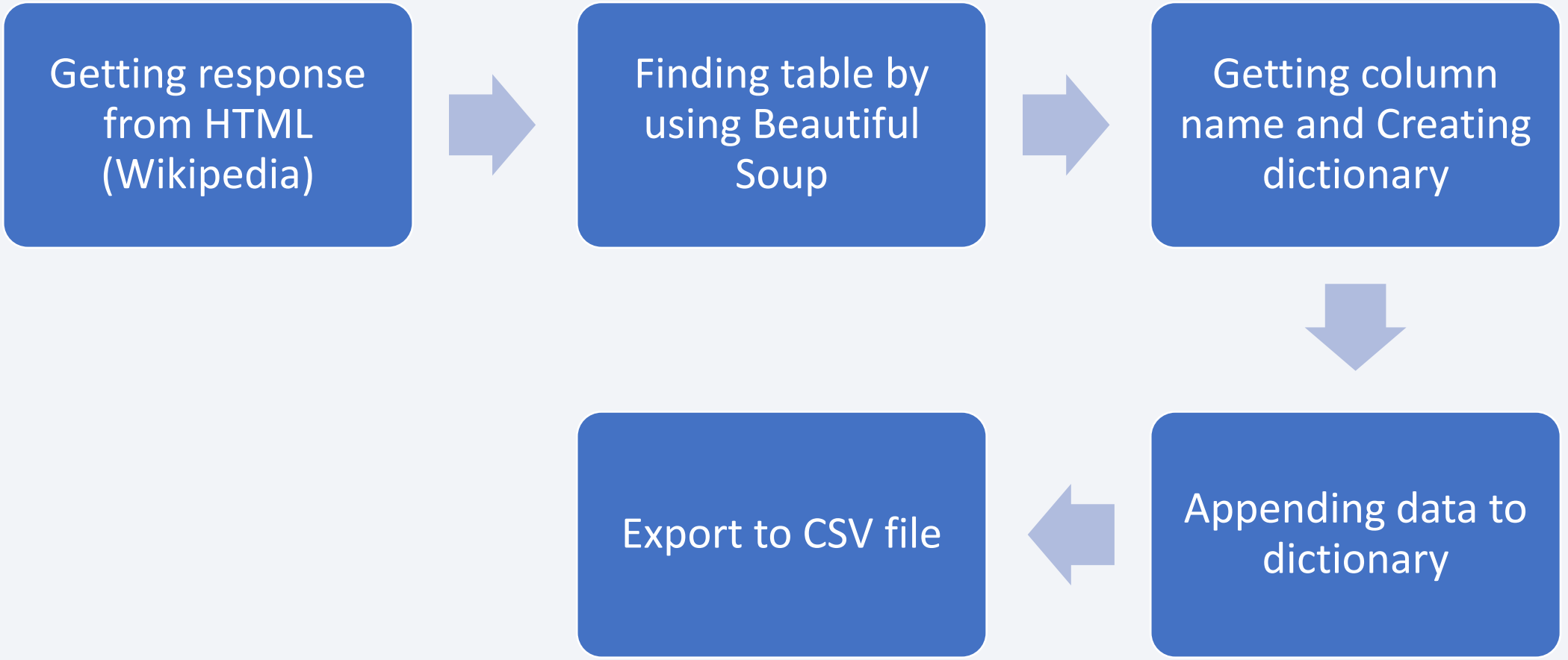


Data Collection - SpaceX API



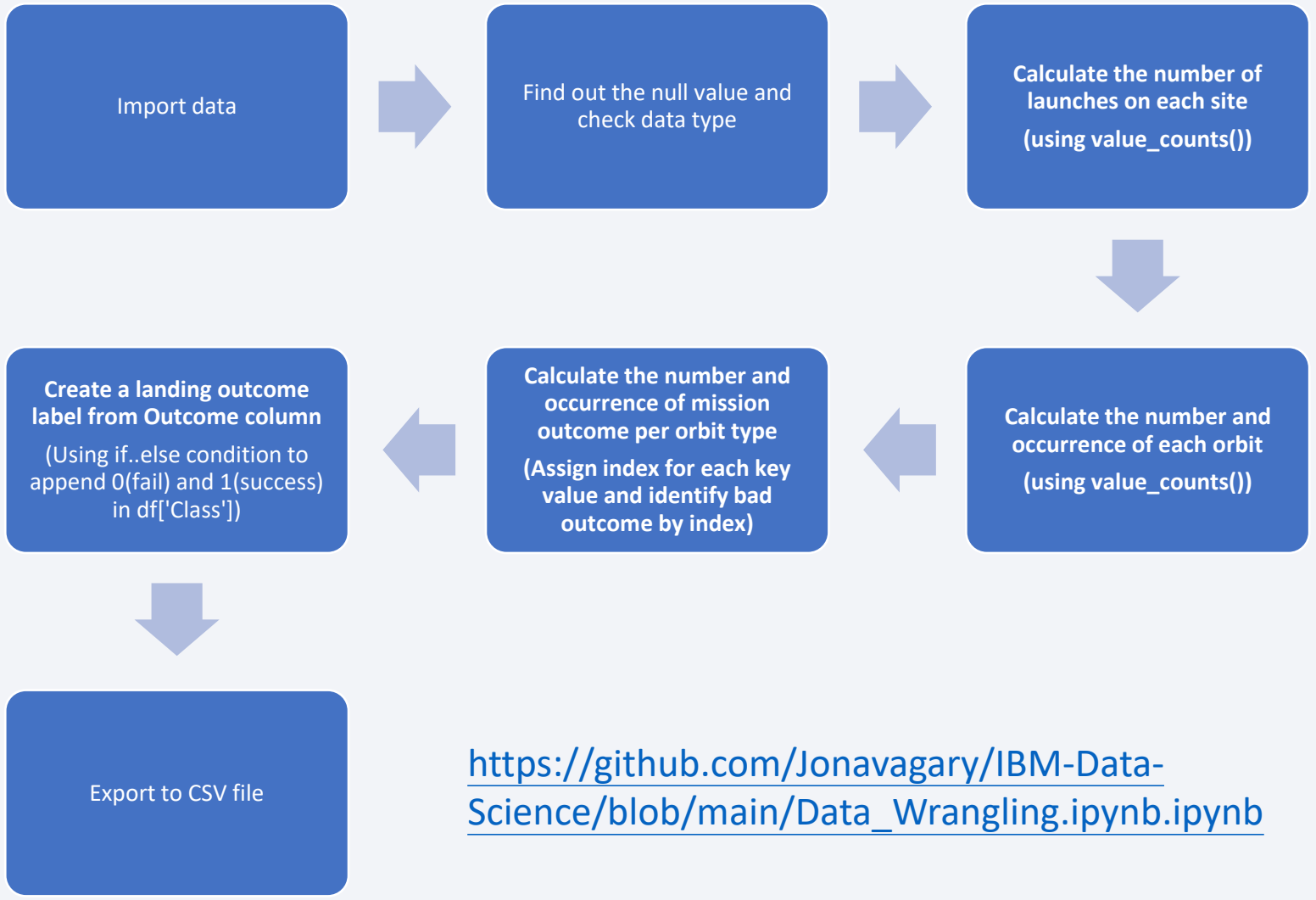
https://github.com/Jonavagary/IBM-Data-Science/blob/main/Data_Collection_API.ipynb.ipynb

Data Collection – Scraping



https://github.com/Jonavagary/IBM-Data-Science/blob/main/Webscraping.ipynb.ipynb

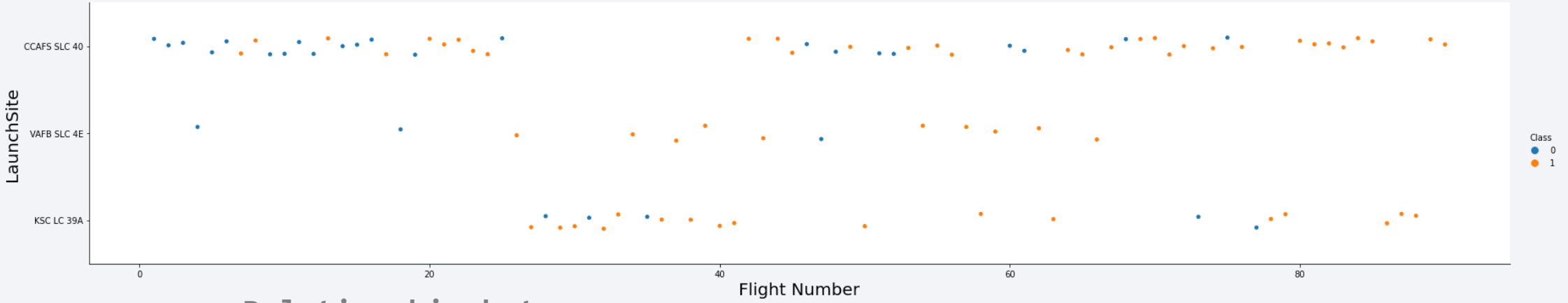
Data Wrangling



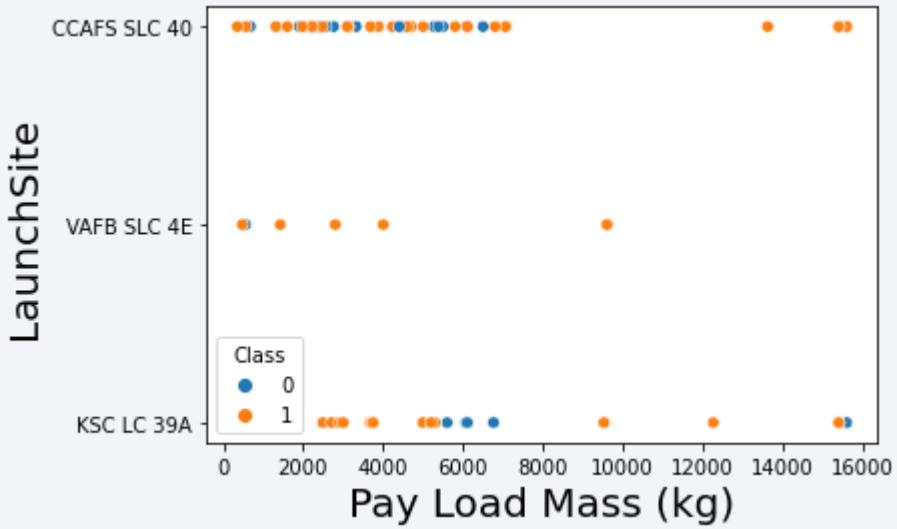
https://github.com/Jonavagary/IBM-Data-Science/blob/main/Data_Wrangling.ipynb.ipynb

EDA with Data Visualization

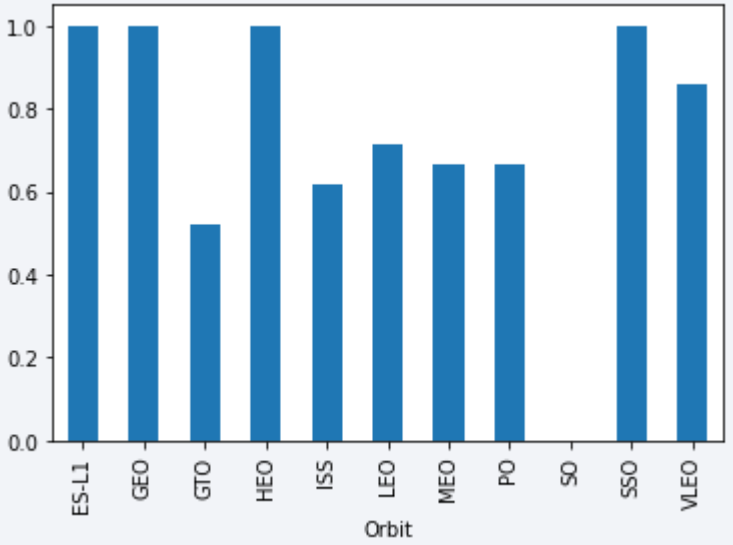
Relationship between Flight Number and Launch Site



Relationship between Payload and Launch Site

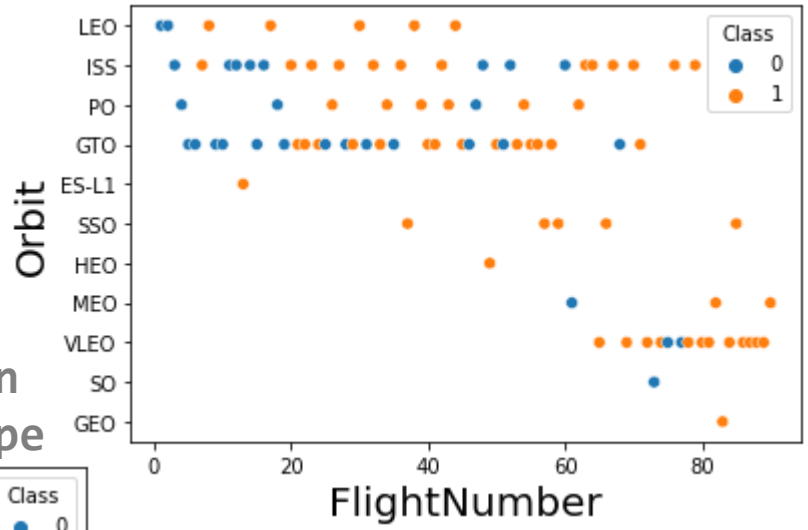


Relationship between success rate of each orbit type

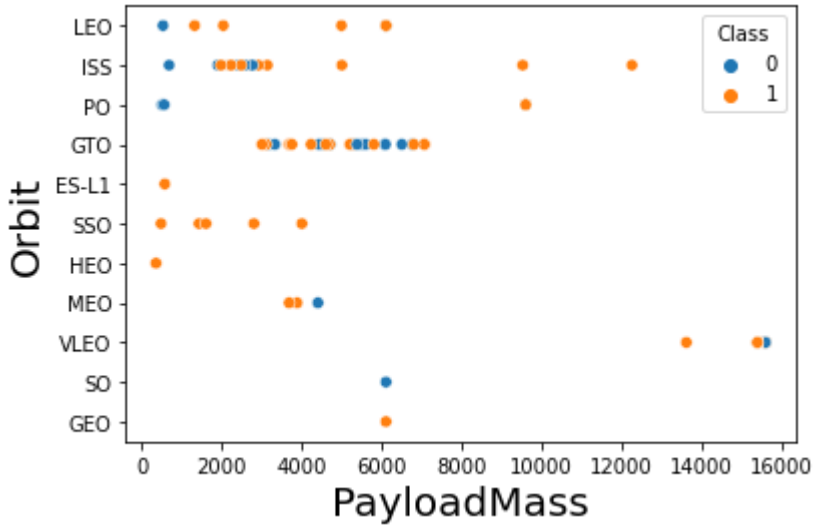


EDA with Data Visualization

Relationship between Flight Number and Orbit type

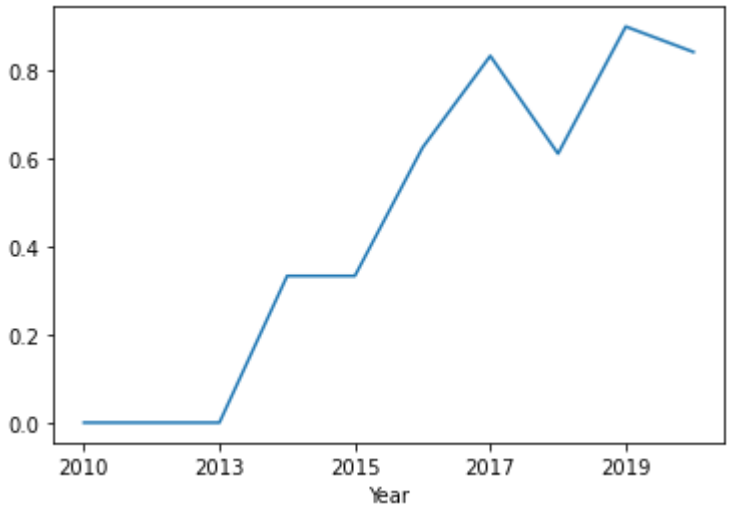


Relationship between Payload and Orbit type



https://github.com/Jonavagary/IBM-Data-Science/blob/main/Data_Visualization.ipynb

Launch success yearly trend



EDA with SQL

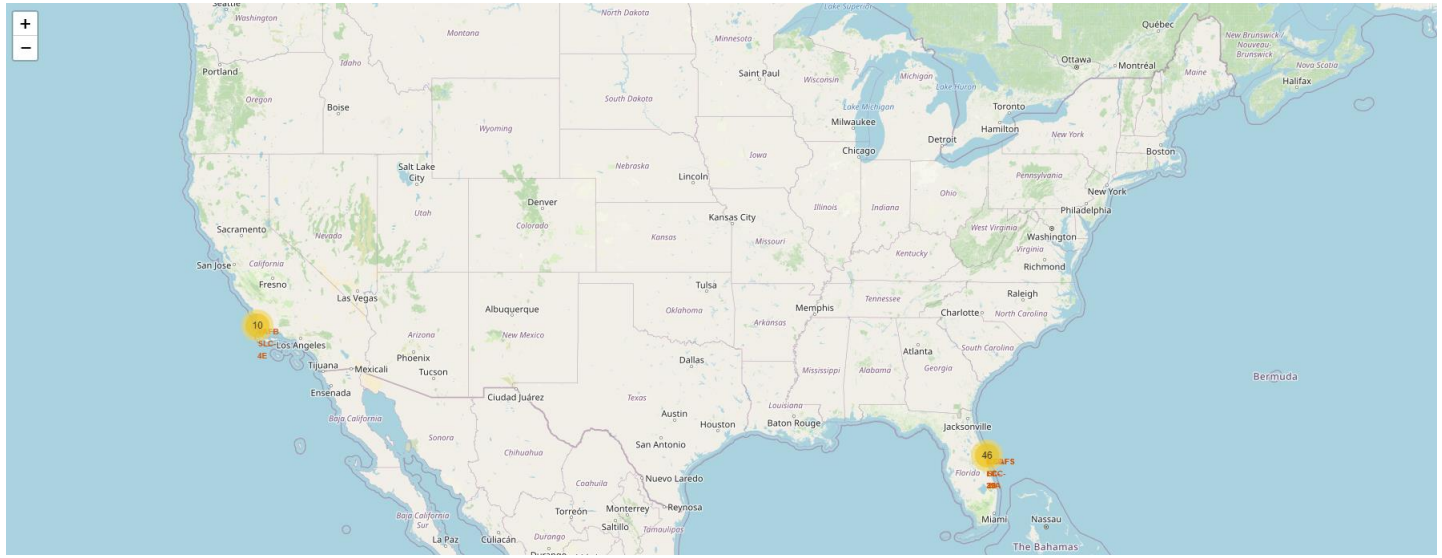
- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA’
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the 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

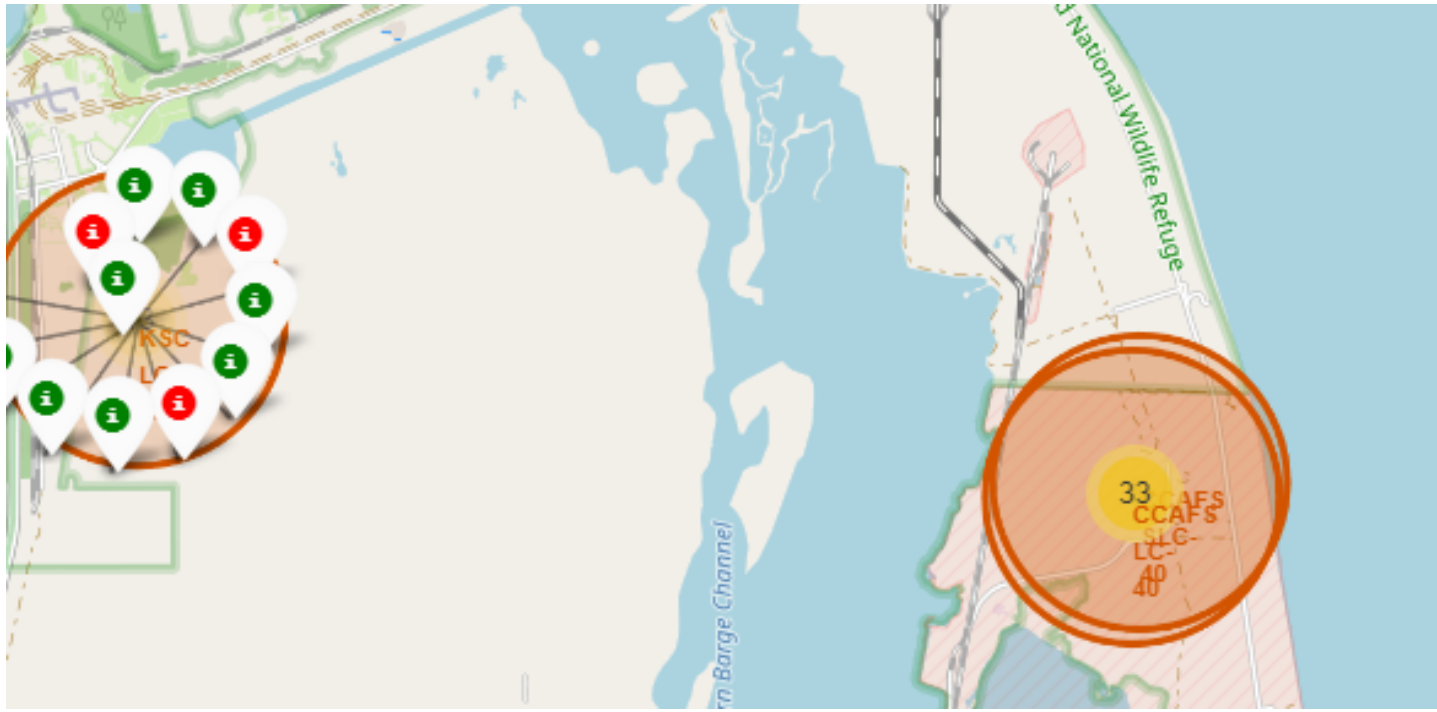
https://github.com/Jonavagary/IBM-Data-Science/blob/main/SQL.ipynb.ipynb

Build an Interactive Map with Folium

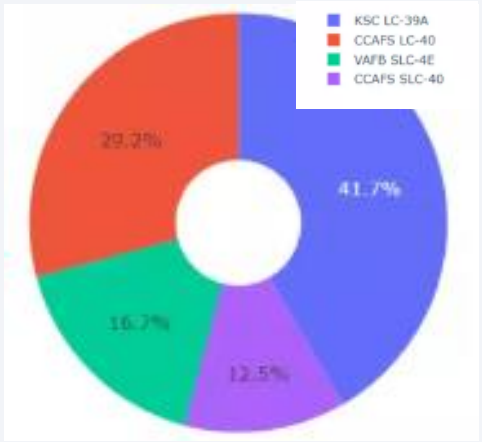
- According to interactive map result. KSC LC-39A has the most successful result for launch

https://github.com/Jonavagary/IBM-Data-Science/blob/main/Folium_Map.ipynb.ipynb



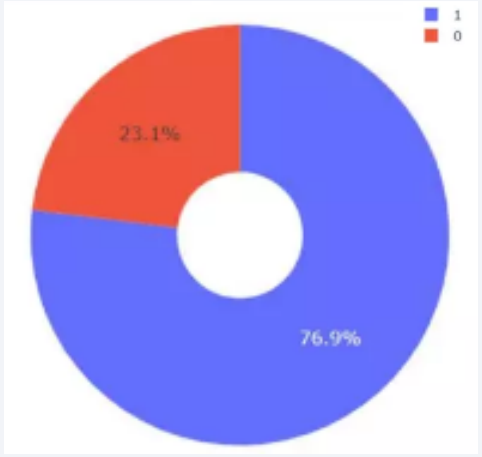


Build a Dashboard with Plotly Dash

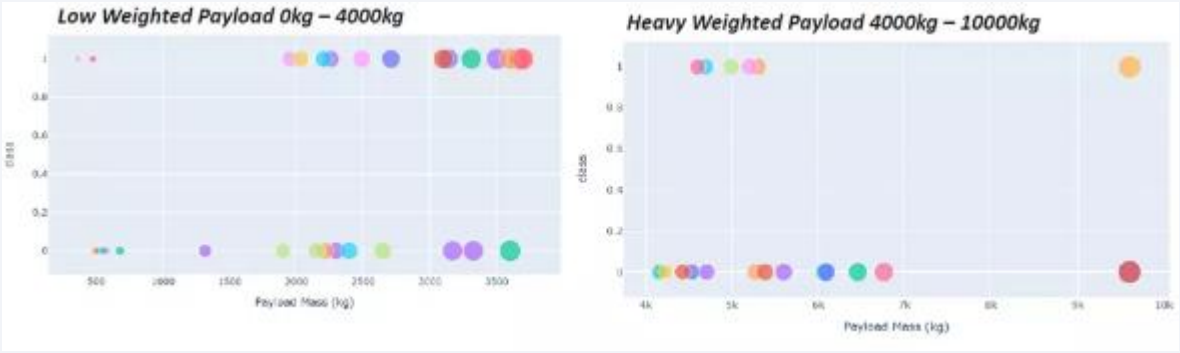


Total Success Launch by sites

KSC LC-39A has the most successful launches

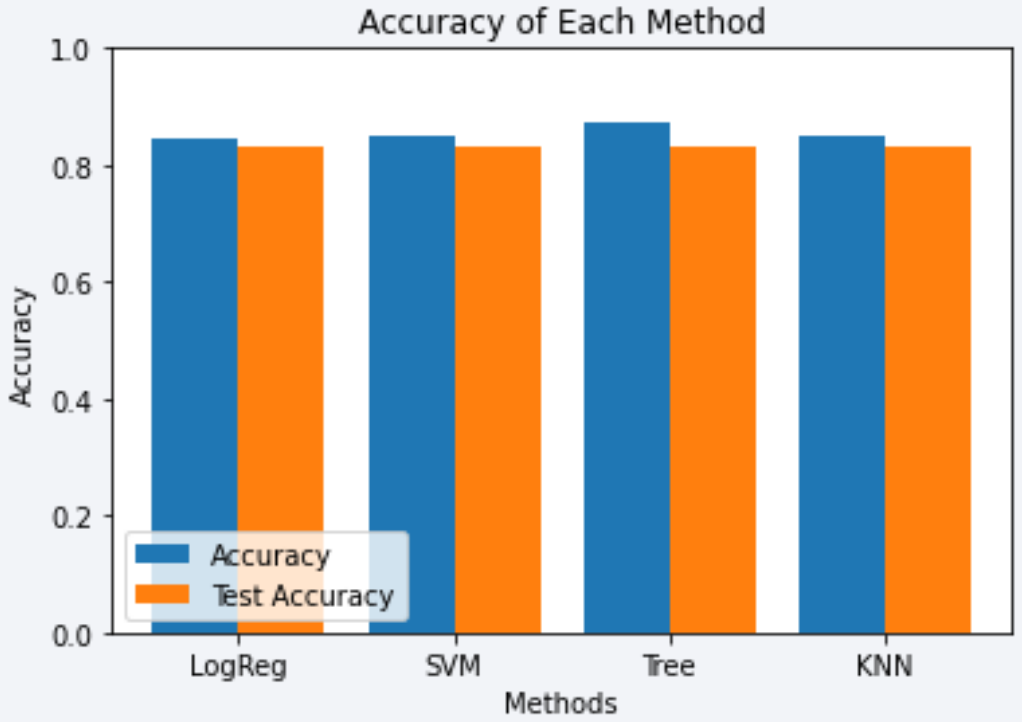


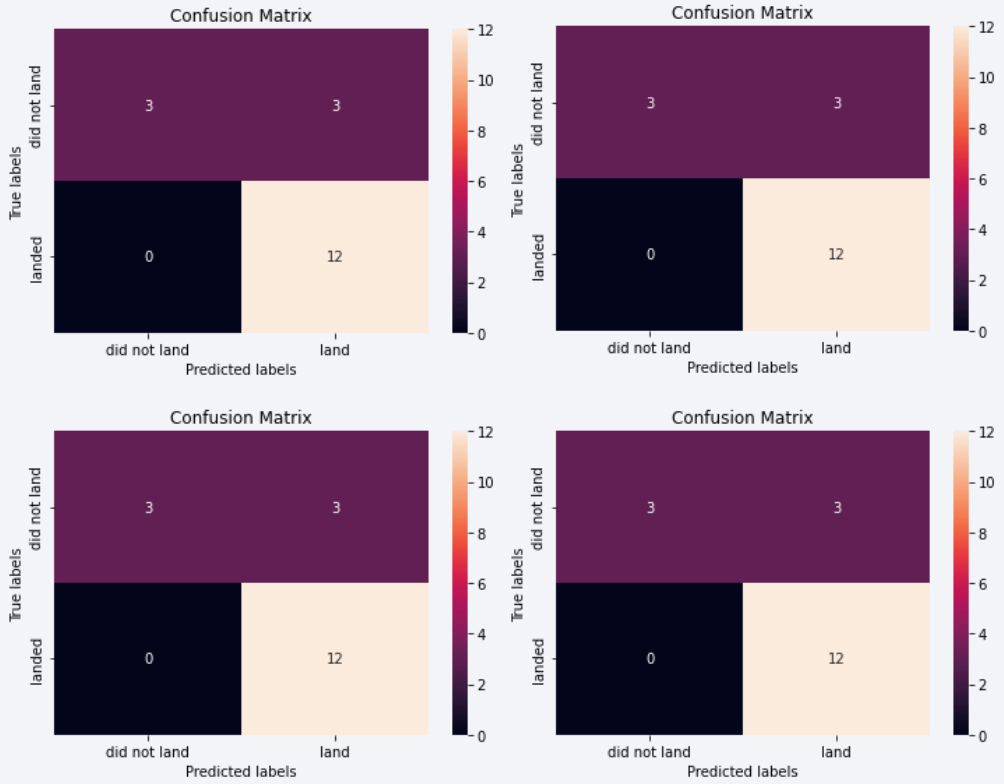
Successful rate of KSC LC-39A



Launch result of Low Weighted Payload and Heavy Weighted Payload

Predictive Analysis (Classification)





The SVM, KMM, Decision Tree and Logistic Regression were used for Predictive Analysis.
All machine learning modal have same accuracy which is 0.833 and Tree model has the highest test accuracy as 0.875.

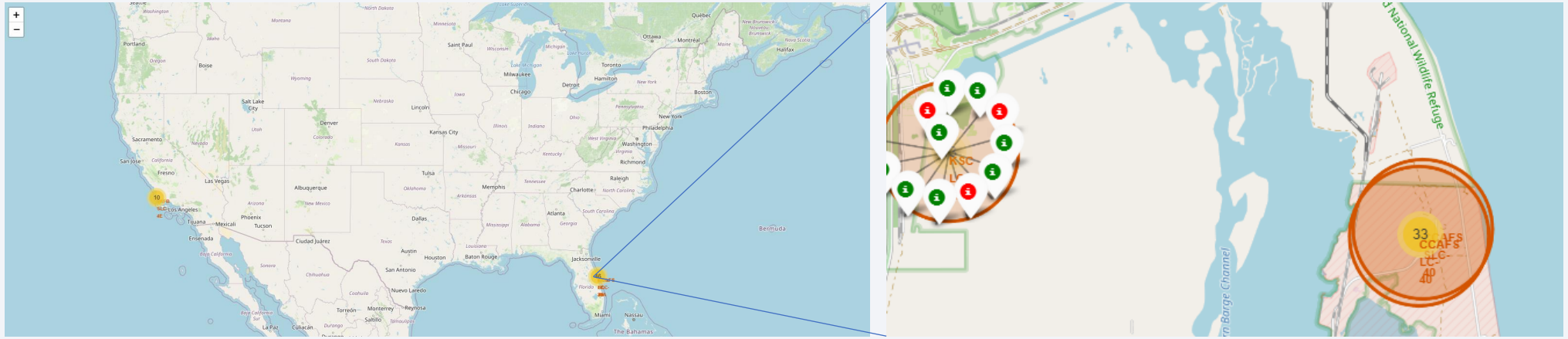
https://github.com/Jonavagary/IBM-Data-Science/blob/main/Machine%20Learning%20Prediction.ipynb

Results

Exploratory data analysis results

- There are 4 different launch sites
- Most Successful launch site is KSC LC-39A
- Average payload for F9 v1.1 is 2,928.4 kg
- Only 1 mission was identified failure
- Result of landing outcome become better over past year

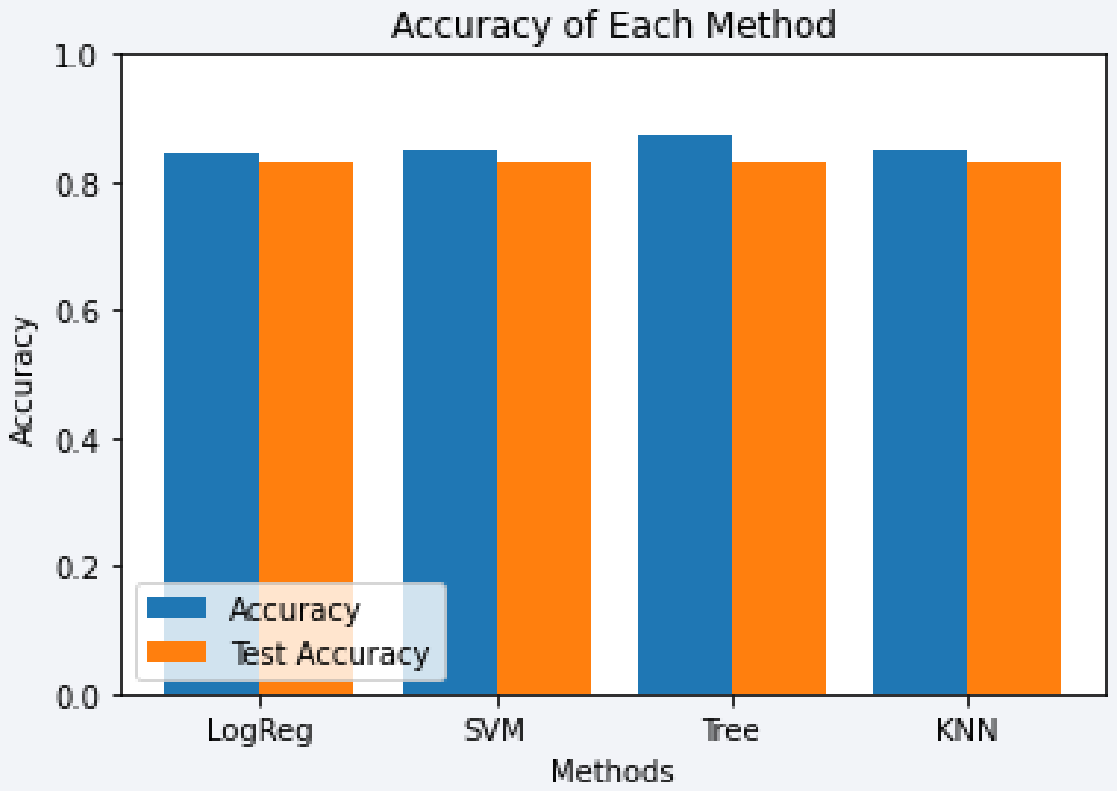
Interactive analytics demo in screenshots

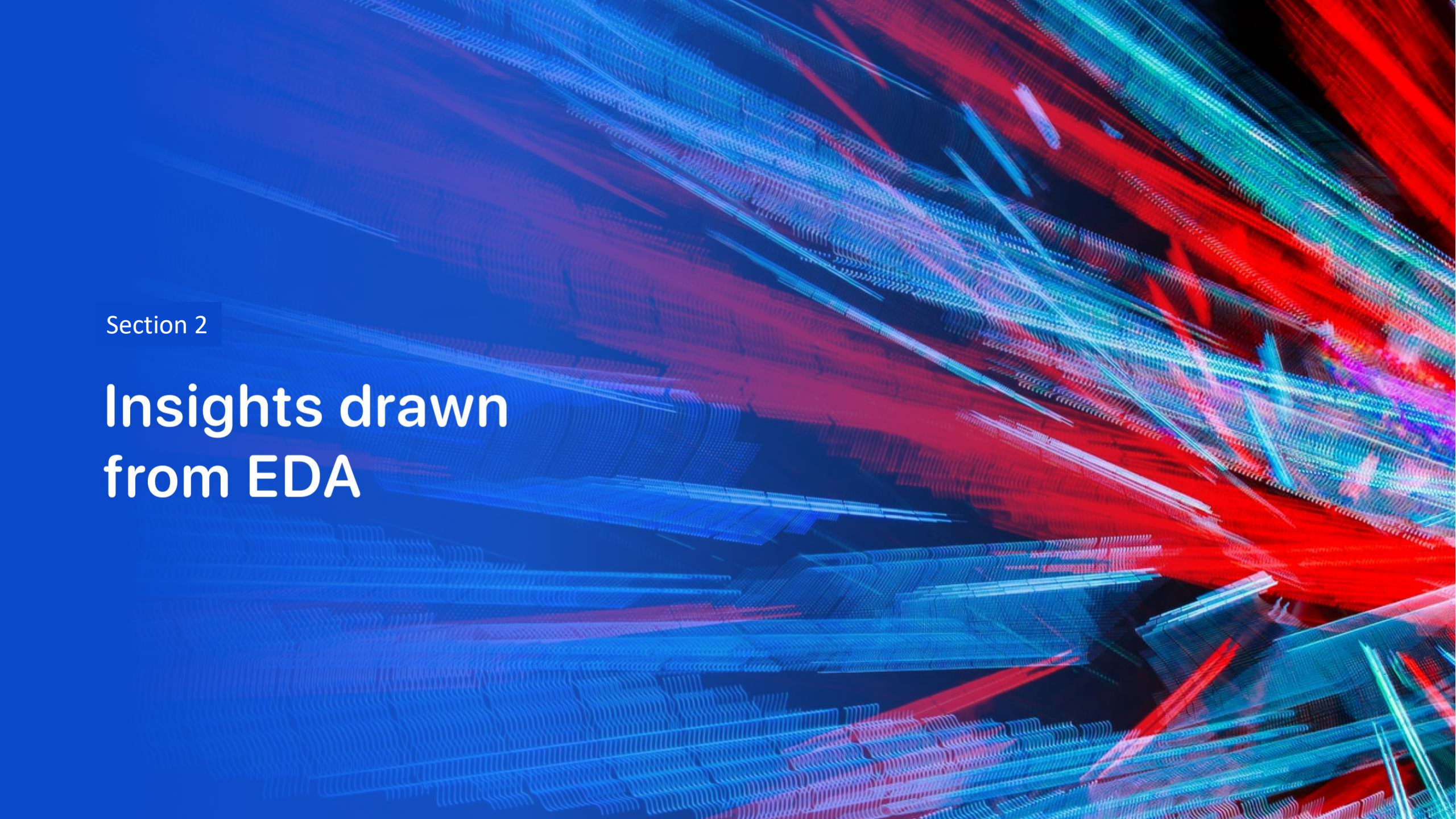


Results

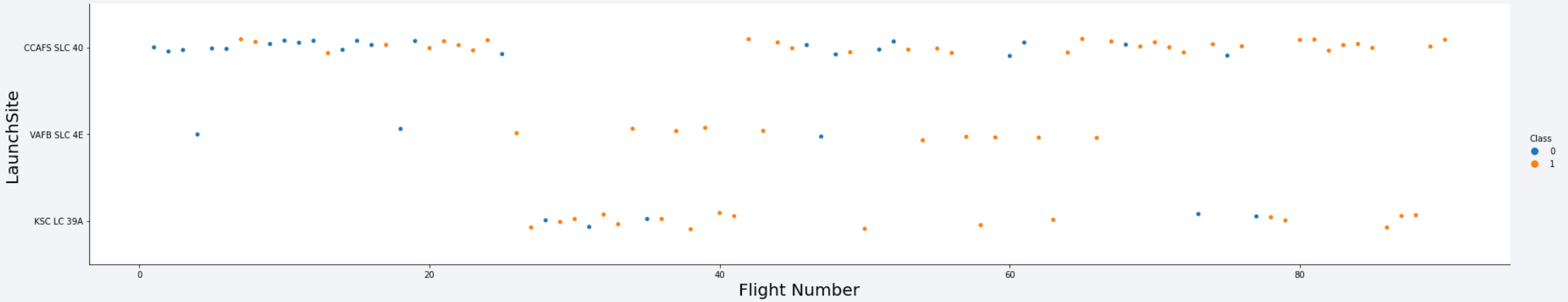
Predictive analysis results

All machine learning modal have same accuracy which is 0.833 and Decision Tree method has the highest test accuracy as 0.875.



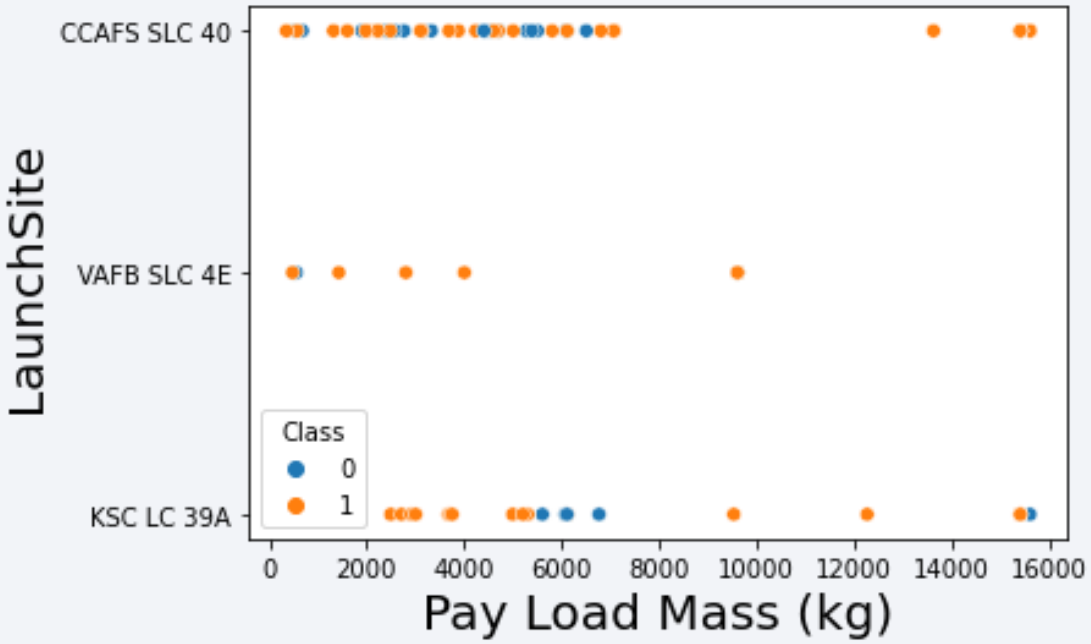

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



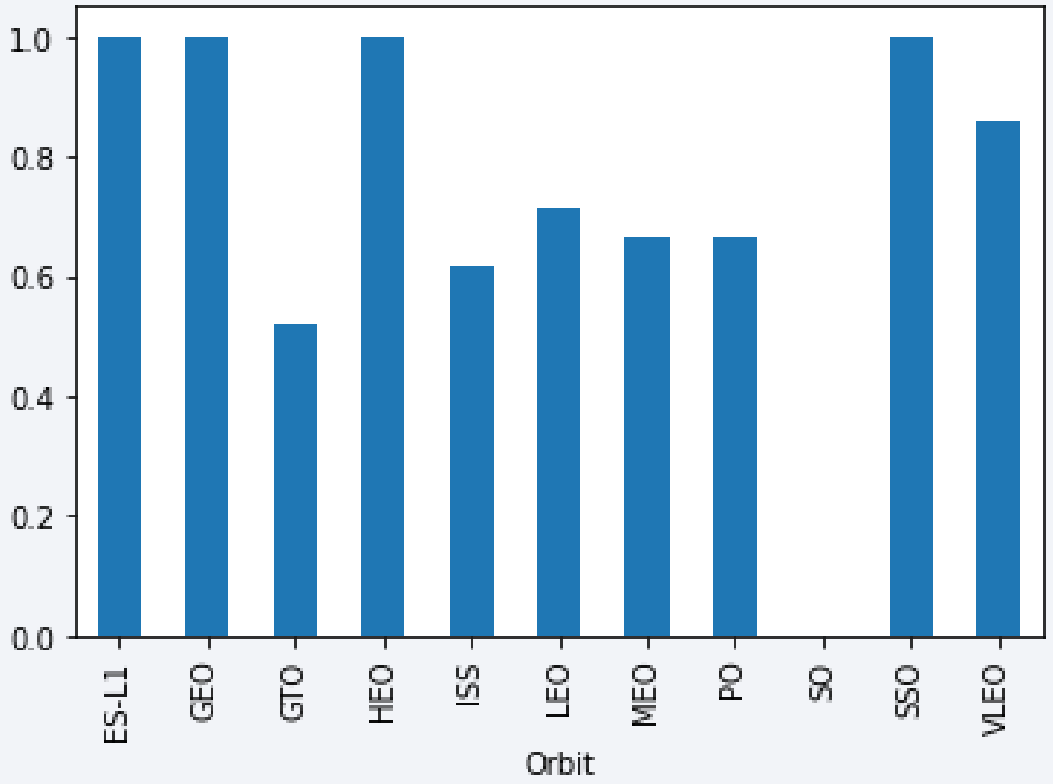
In the beginning, SpaceX used CCAFS SLC 40 as their main launch site. SpaceX start using KSC LC-39A after around 20 launches. Therefore, KSC LC-39A has the highest successful rate.

Payload vs. Launch Site



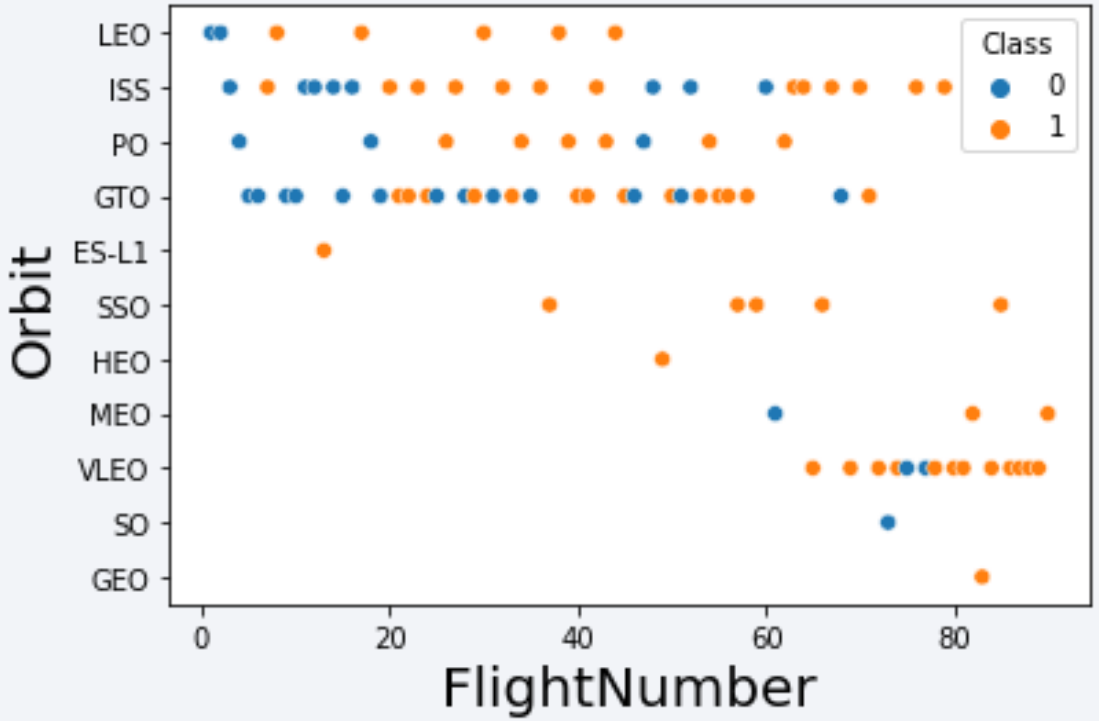
The most payload range is from 0 to 8,000kg. The highest frequency site to launch lower payload is CCAFS SLC 40.

Success Rate vs. Orbit Type



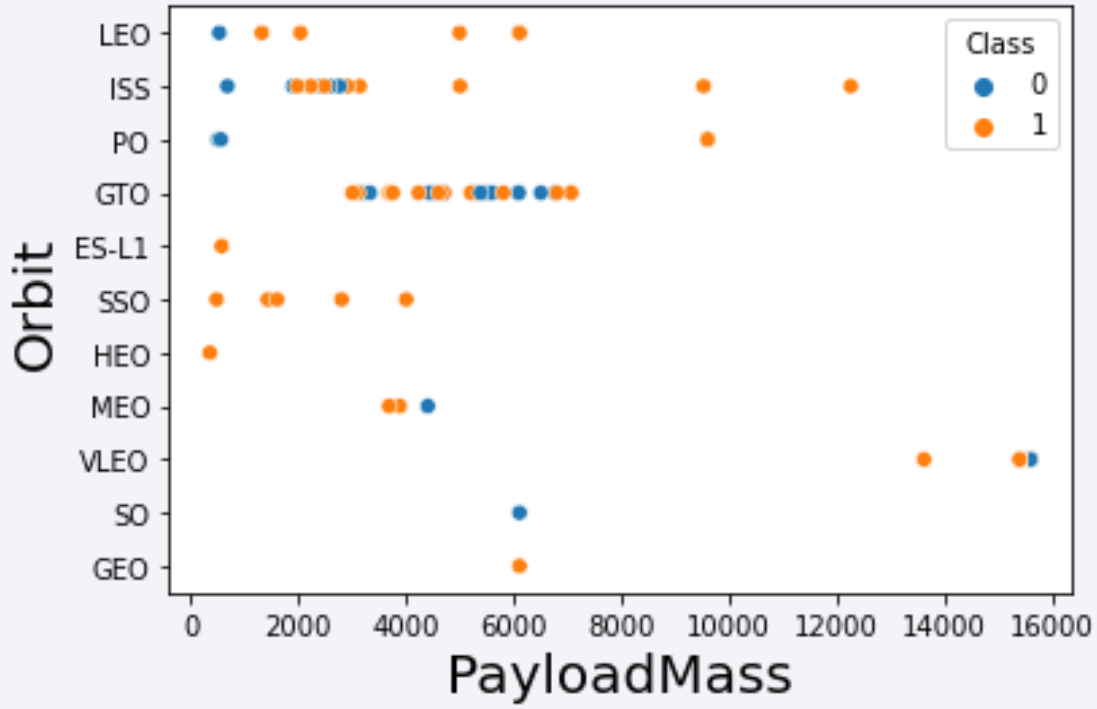
ES-L1, GEO, HEO and SSO have 100% successful rate and SO has 0% as there is one mission only.

Flight Number vs. Orbit Type



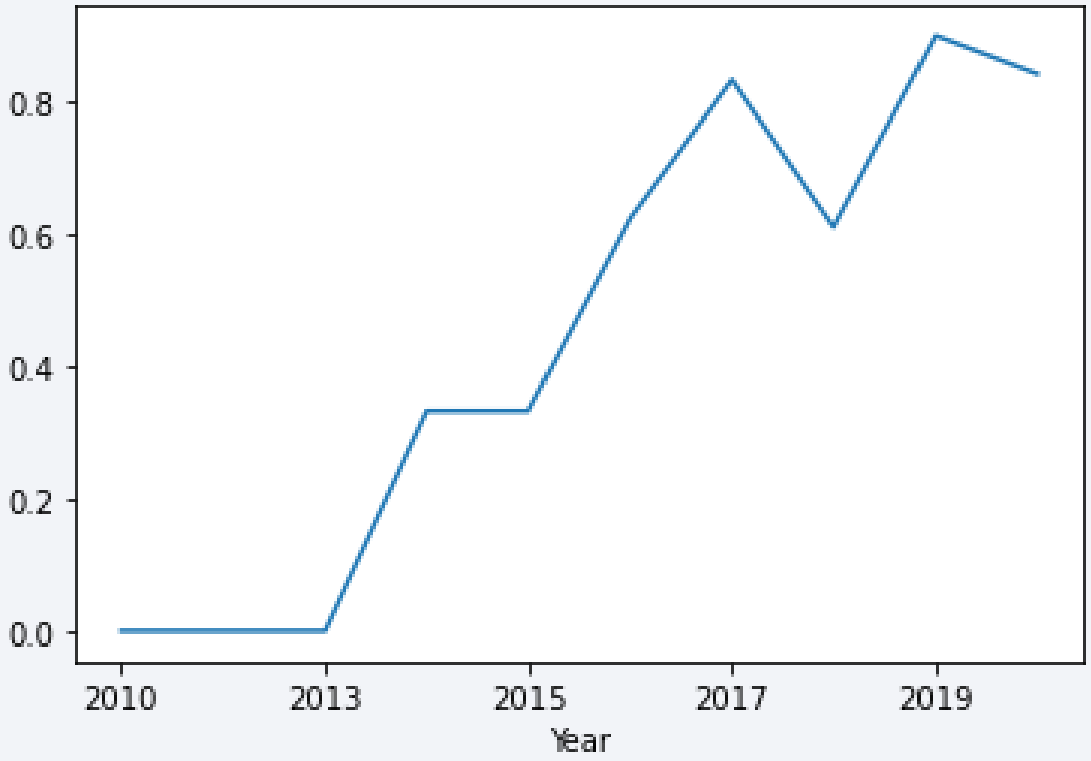
SpaceX launched rocket to LEO, ISS and GTO in early stage and now they shifted to focus on VLEO and ISS.

Payload vs. Orbit Type



The most payload range for ISS is from 2,000 kg to 4,000 kg. For GTO, the most payload range is 3,000 kg to 8,000kg

Launch Success Yearly Trend



Launching successful rate keep increasing since 2013. the successful rate is around 80% in 2020.

All Launch Site Names

```
%sql select distinct launch_site from SPACEXTBL
```

Python

Explain: Select the unique value of launch site.

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

Launch Site Names Begin with 'CCA'

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

Python

Explain: Select 5 data rows if launch sites included word “CCA”.

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

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

Python

Explain: Sum all the payload as total payload if customer is NASA(CRS).

total_payload_mass
45596

Average Payload Mass by F9 v1.1

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

Python

Explain: Average all the payload as average payload mass if booster version included word “F9 v1.1”.

average_payload_mass
2928.4

First Successful Ground Landing Date

```
%sql select min(date) as first_successful_landing from SPACEXTBL where "Landing _Outcome" = 'Success (ground pad)';
```

Python

Explain: Find the first date when rocket successfully land in ground pad.

first_successful_landing
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version from SPACEXTBL where "Landing _Outcome" = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

Python

Explain: Select booster version when rocket successfully land in drone ship and payload is between 4,000 kg to 6,000 kg.

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

Total Number of Successful and Failure Mission Outcomes

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

Python

Explain: Select and count mission outcome.

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

Boosters Carried Maximum Payload

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

Python

Explain: Select booster version if they carried maximum payload.

Booster_Version	
F9 B5 B1048.4	F9 B5 B1049.5
F9 B5 B1049.4	F9 B5 B1060.2
F9 B5 B1051.3	F9 B5 B1058.3
F9 B5 B1056.4	F9 B5 B1051.6
F9 B5 B1048.5	F9 B5 B1060.3
F9 B5 B1051.4	F9 B5 B1049.7

2015 Launch Records

```
%sql select substr(Date,4,2) as Month, date, booster_version, launch_site
from SPACEXTBL where substr(Date,7,4)='2015' and [Landing _Outcome] LIKE 'fail%';
```

0.0s Python

Explain: Select data of month, date, booster version and launch site if launch date was in 2015 and outcome is failed.

Month	Date	Booster_Version	Launch_Site
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40

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

```
%%sql
SELECT "Landing _Outcome",count("Landing _Outcome")as LANDING_OUTCOME_COUNT
from SPACEXTBL where  DATE between '04-06-2010' and '20-03-2017' and "Landing _Outcome"  LIKE 'Success%'
group by "Landing _Outcome" ORDER BY LANDING_OUTCOME_COUNT DESC;
```

✓ 0.0s Python

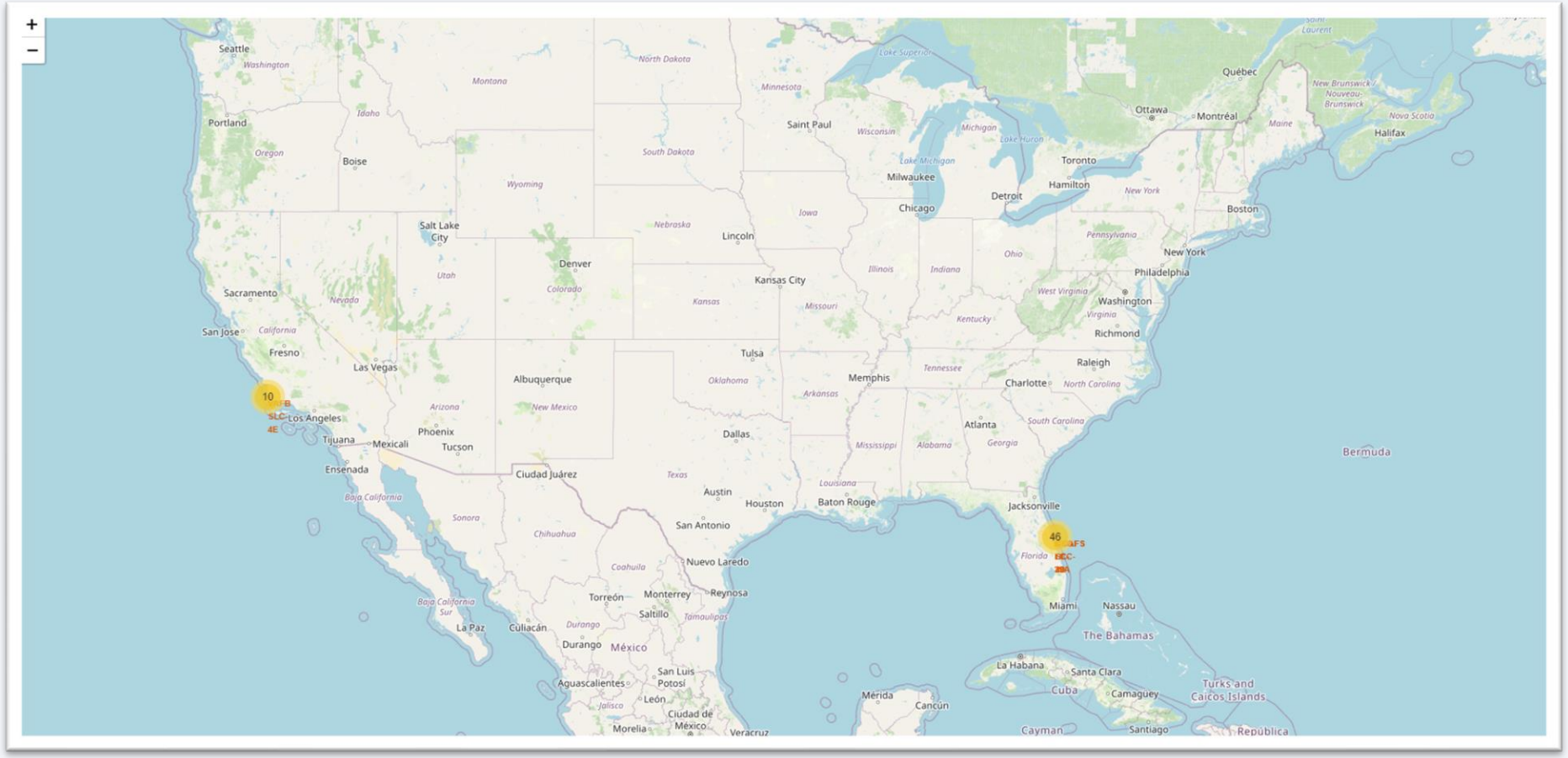
Explain: Select and count successful landing outcome by group if launch date is between 2010-06-04 and 2017-03-20.

anding _Outcome	LANDING_OUTCOME_COUNT
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

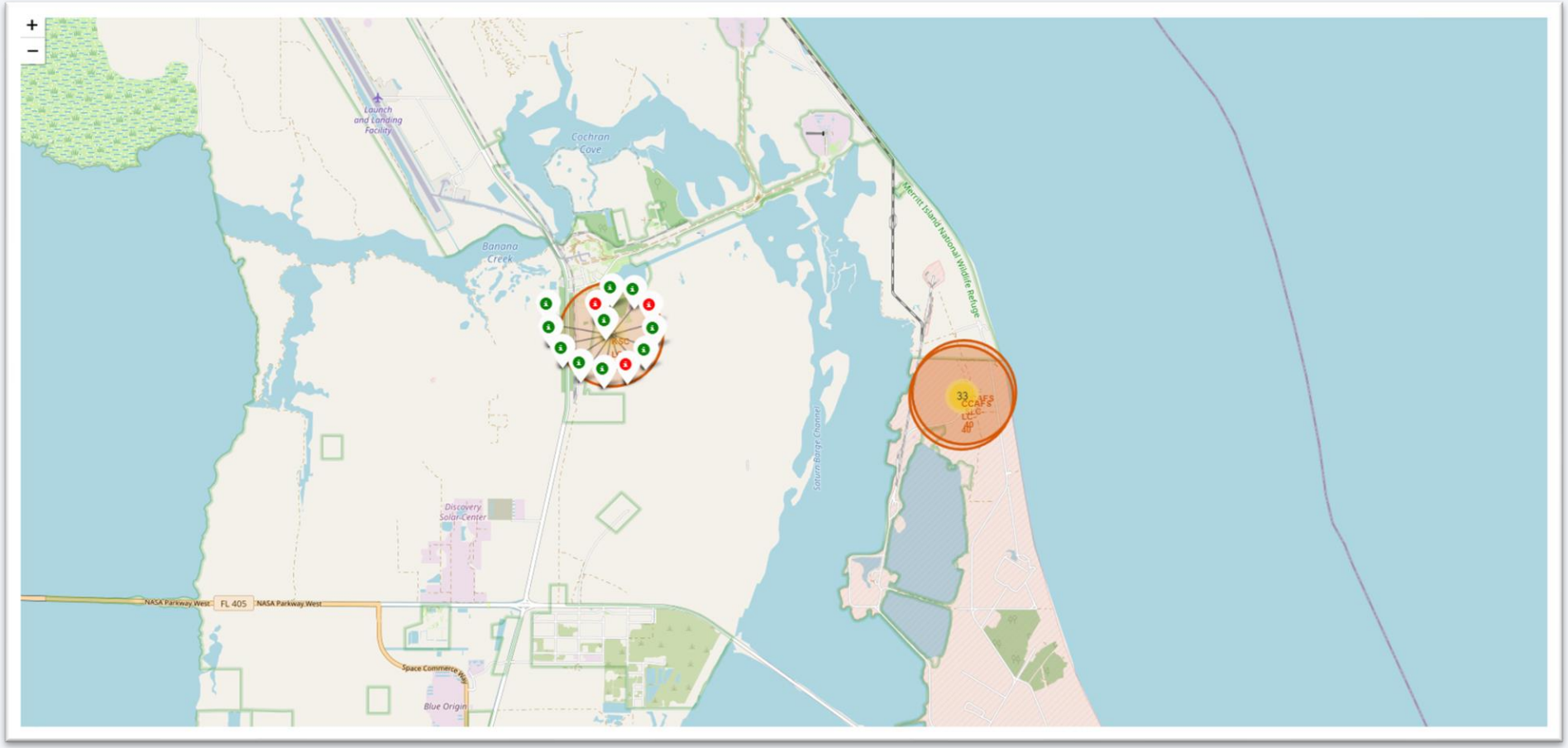
Launch Sites Proximities Analysis

All launch sites



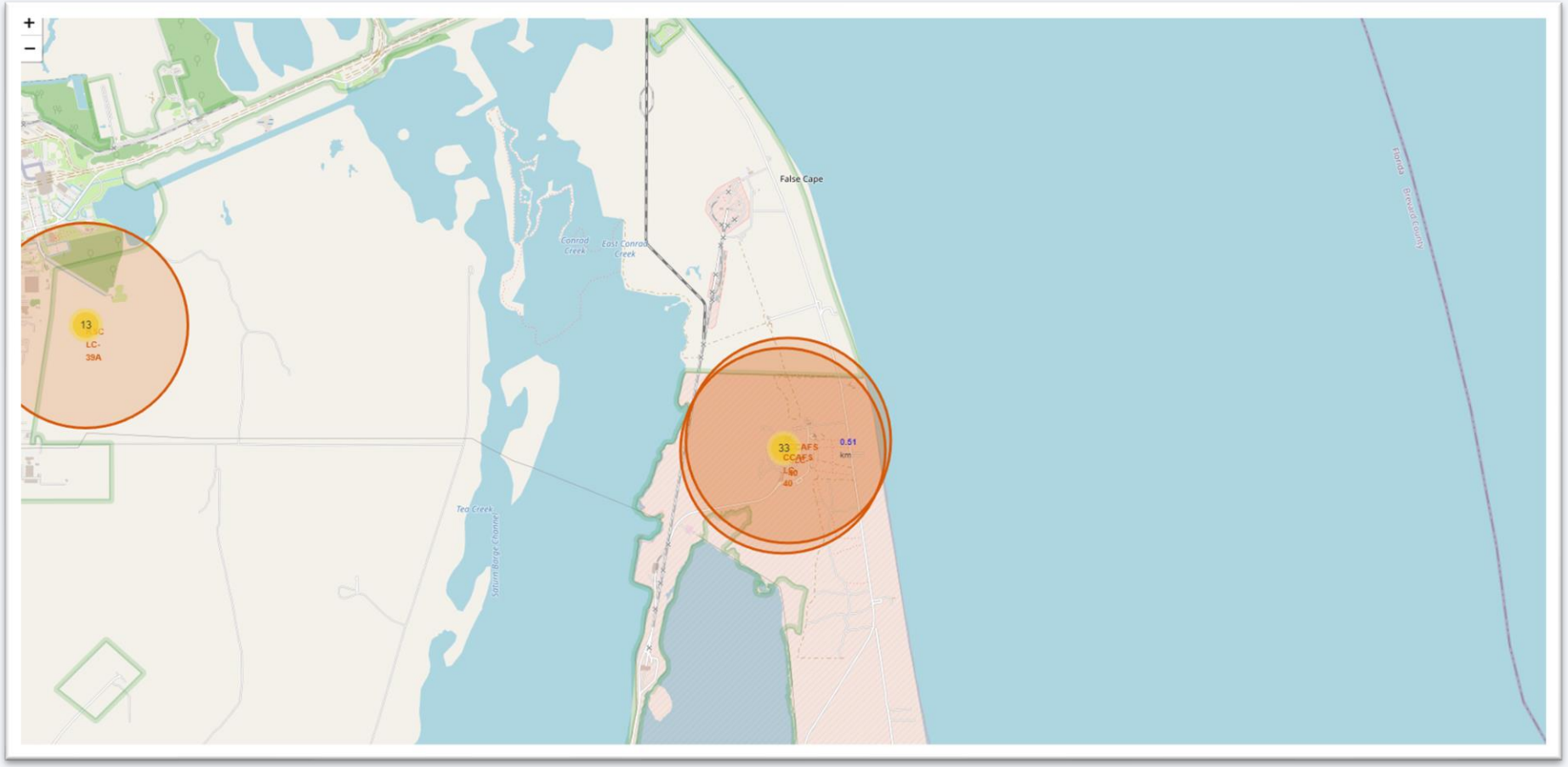
SpaceX has 4 launch sites which all are near sea. One is in west coast and three sites are in east coast

Outcomes by site

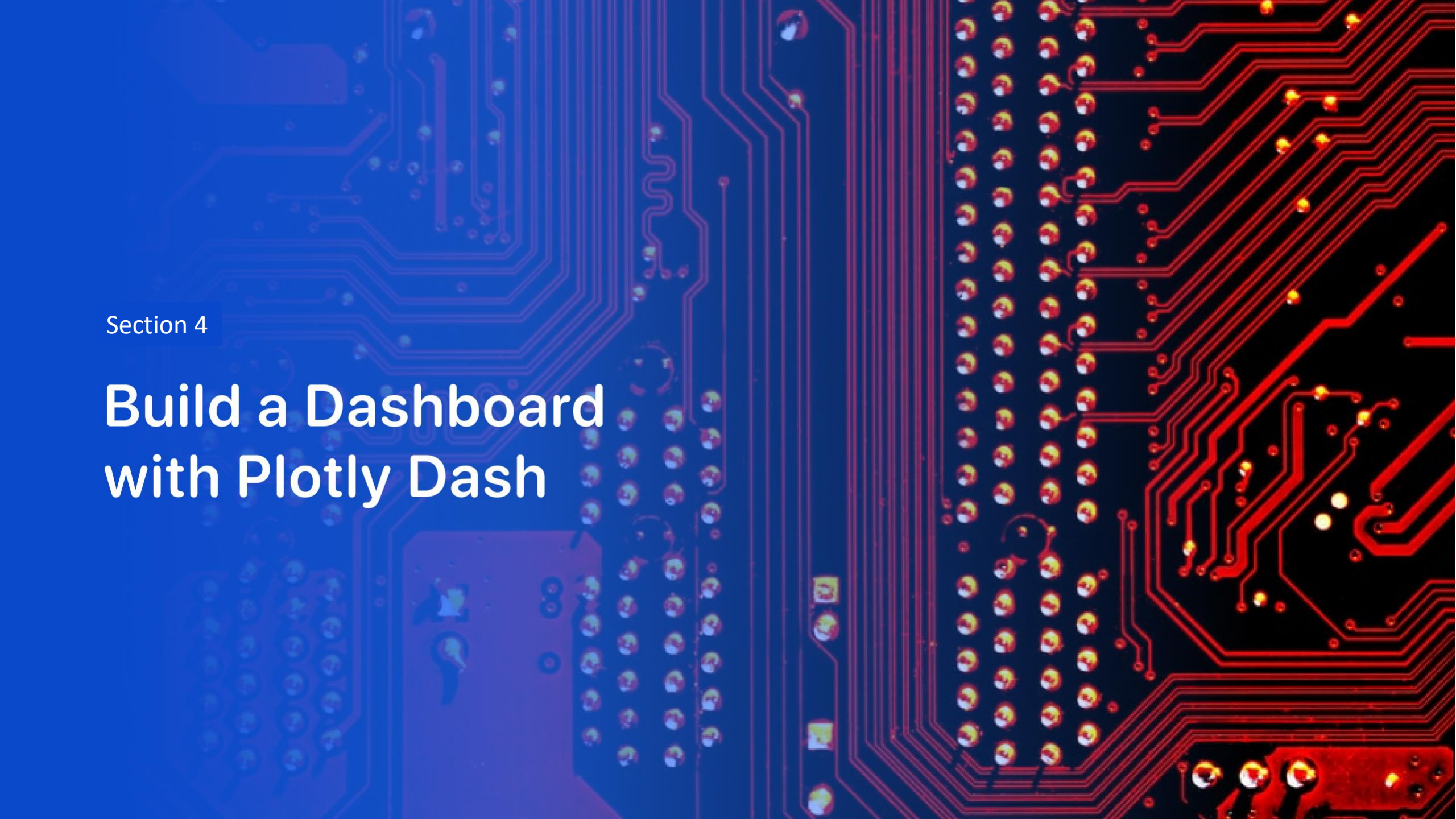


Example of KSC LC-39A launch outcomes. Green maker indicates successful outcome and red maker indicates failure outcome.

Location of CCAFS SLC-40



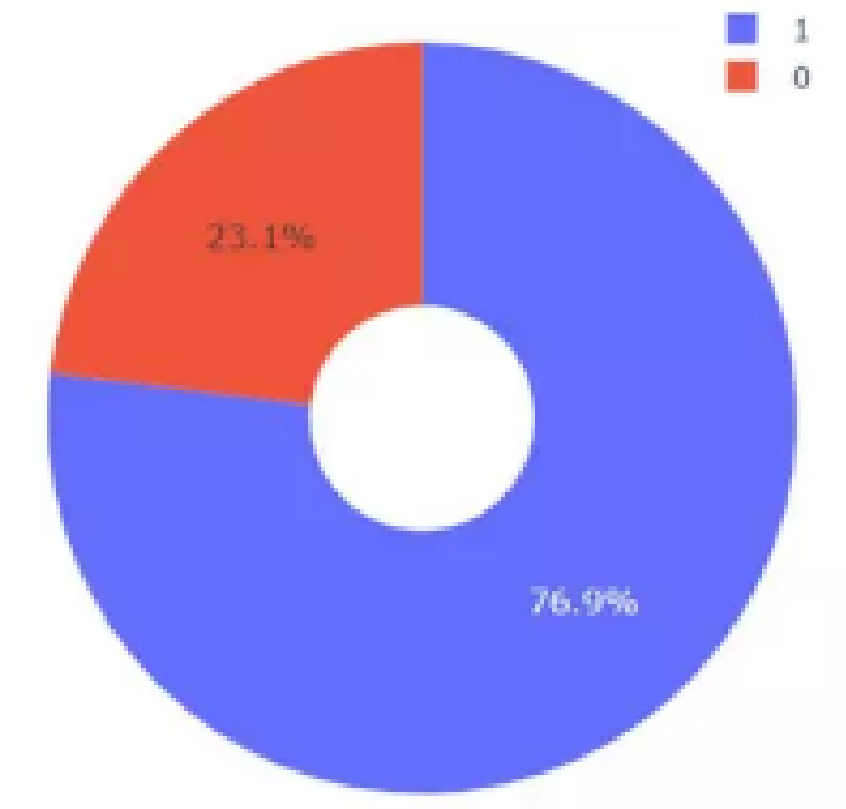
CCAFS SLC-40 launch site has good location. It is near the coastline (0.5 km).


Section 4
Build a Dashboard with Plotly Dash

Successful Launches by site

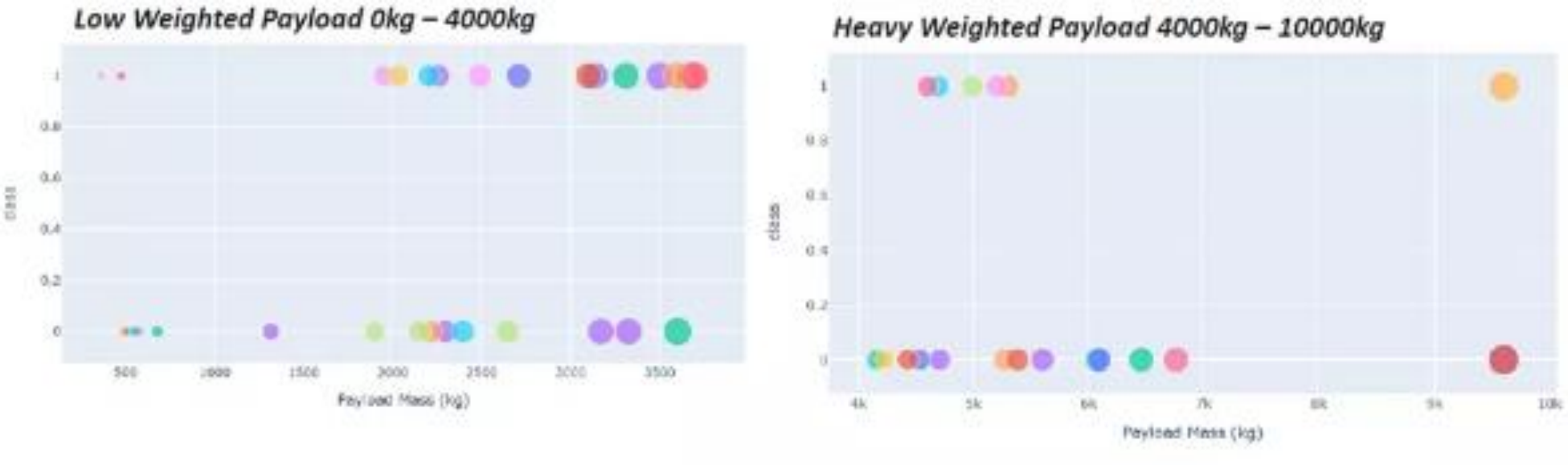
Most successful launches come from KSC LC-39A. No doubt, launching location maybe a factor of successful rate. However, we should also consider the total number of launches in each site and the mission detail.

Launch Success Rate for KSC LC-39A



76.9% of launches are successful in KSC LC-39A

Payload vs. Launch outcome

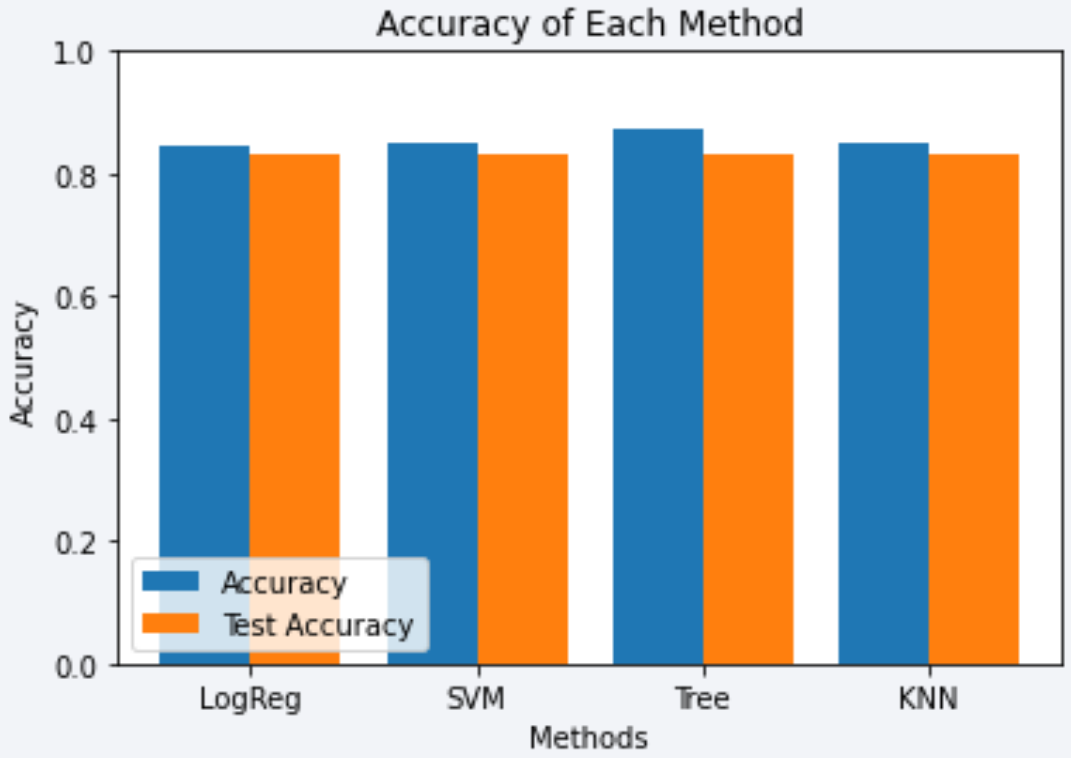


According to the graph, the launch has chance to be successful if the weighted payload is between 2,000 kg and 4,000 kg.

Section 5

Predictive Analysis (Classification)

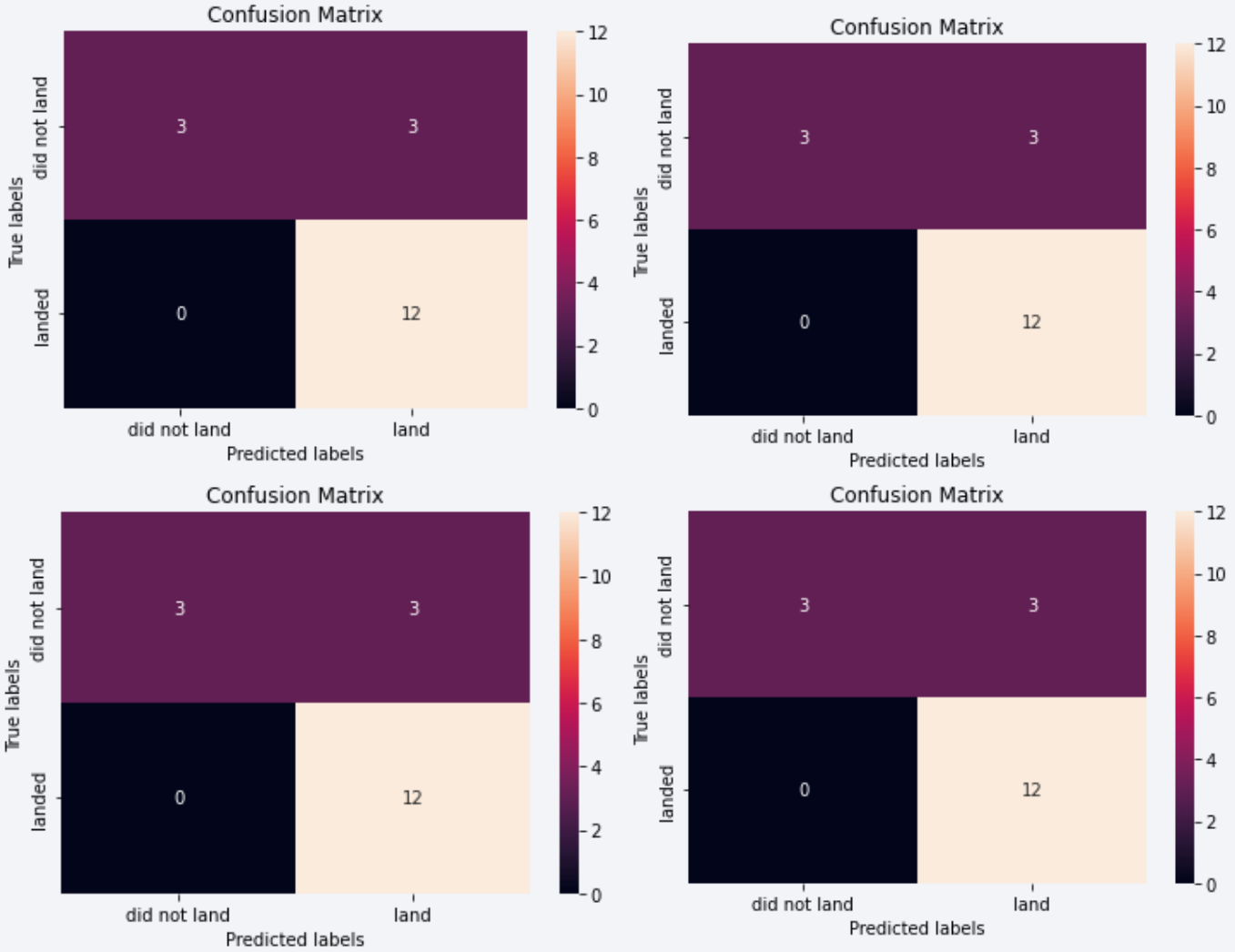
Classification Accuracy



Four classification methods had been tested in this case.

According to the result, decision tree methos has a better accuracy than others. Also, all methods have same test accuracy.

Confusion Matrix



All methods have same test accuracy (0.833) and they are facing type II error which is false negative.

Conclusions

- SpaceX data can be extract from public sources;
- Rocket technology keep improving;
- Successful rate is around 80% in 2020;
- SpaceX has100% successful rate for GEO, HEO and SSO;
- Total payload for NASA is 45,596. this figure will increase over time as SpaceX obtained NASA contract for ISS transportation;
- Launch site is near coastline and railway. The consideration should be transportation and safety;
- All machine learning method performed well, accuracy is 83% for 80 training data and 20 testing data. I believe the accuracy will improve if we input more data.

Thank you!