IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Salman Al Gahuri
November 20, 2024

Outline

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

Executive Summary

- Summary of used methodologies

  - Data Collection through API
  - Data Collection with Web scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction

- Summary of all results

  - Exploratory Data Analysis results
  - Interactive Analytics in screenshots
  - Predictive Analytics results

Introduction

SpaceX 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 SpaceX 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 SpaceX for a rocket launch. The goal is to create machine learning pipeline to predict if the first stage will land successfully.

We can use Data Science and Machine Learning in order to find:

1. What elements decide whether the first stage will successfully land?
2. What operating conditions needs to be in place to ensure a successful landing program?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from SpaceX API and Web Scraping from Wiki Pages
- Perform data wrangling
  - Data was first converted to a pandas dataframe then One-hot encoding was applied to categorical variables
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Use machine learning to predict if the first stage will land successfully

Data Collection

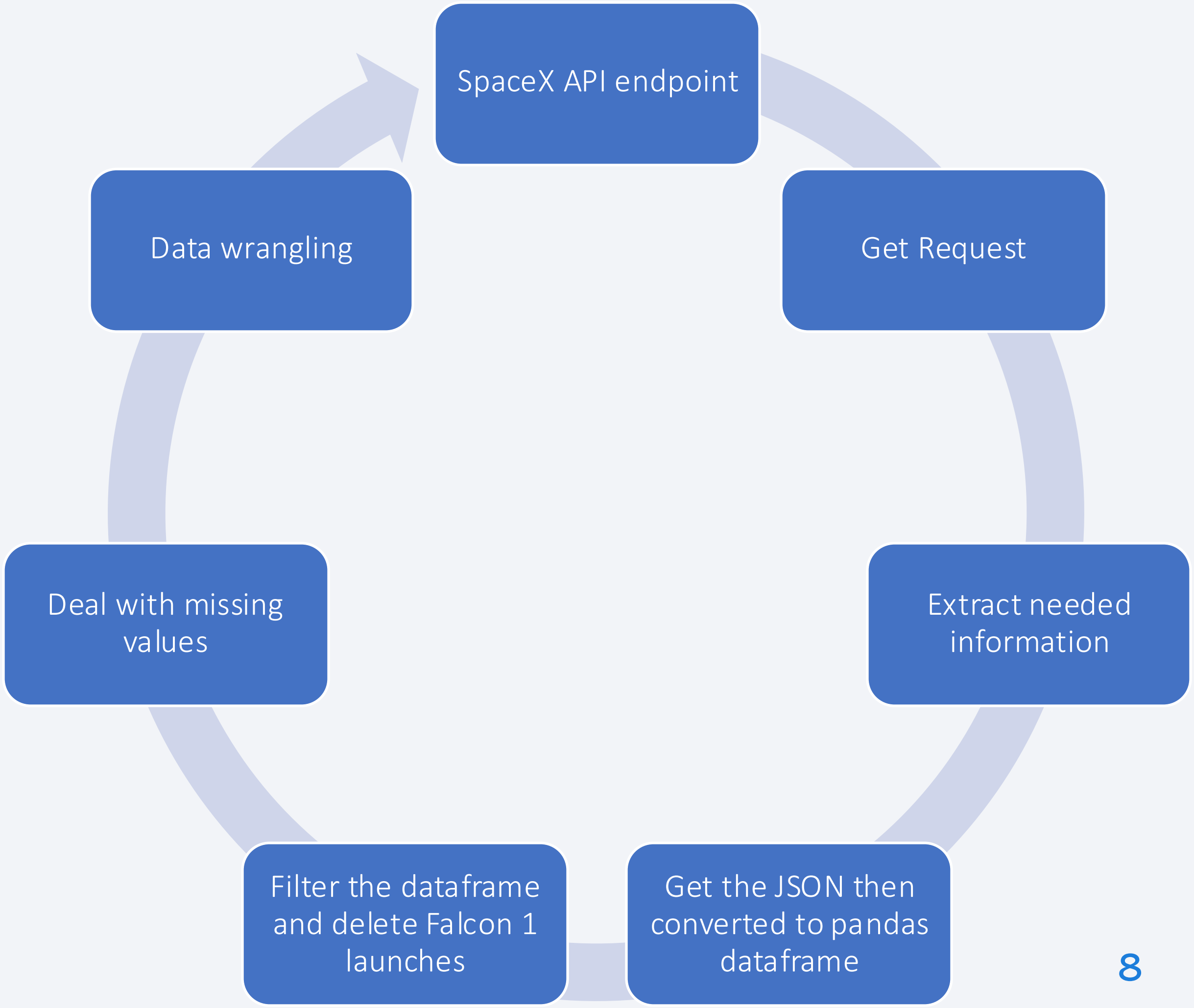
Data was collected from SpaceX API and web scraped from wiki pages.

SpaceX REST API → Get Request → Get data as a JSON object → Convert JSON object to pandas dataframe

Web Scraping Falcon 9 web records → Use BeautifulSoup to web scrape HTML tables → Parse data from tables → Convert tables to pandas dataframe

Data Collection – SpaceX API

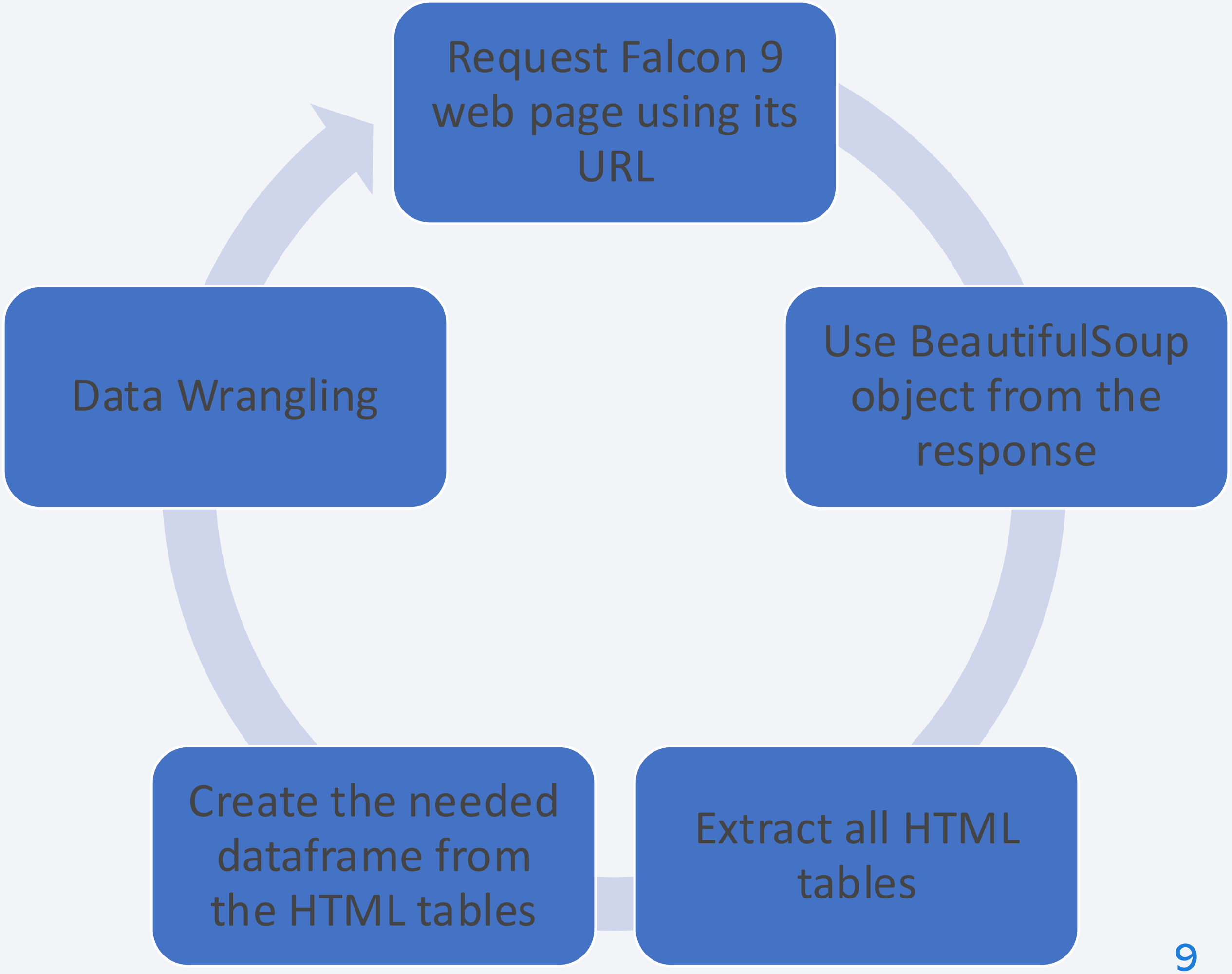
- We used the get request to the SpaceX API to collect and clean the data, as well as wrangling the data and dealing with missing values



SpaceX Data Collection Notebook

Data Collection - Scraping

- Performed web scraping to Falcon 9 launch records from web pages.



Web Scraping Data Collection Notebook

Data Wrangling

Identify Missing Values

Identify numerical and categorical variables

Calculate the number of launches as well as the occurrence of each orbit

Calculate the number of missing landing outcome in total and for each orbit

Create landing output label as 0 and 1 for failure and success respectively

Data Wrangling Notebook

EDA with Data Visualization

Summary of plotted charts:

- Catplot to visualize how Flight Number and Payload effect the landing outcome.
- Catplot to visualize how Flight Number and Launch Site effect the landing outcome.
- Catplot to visualize how Launch Site and Payload effect the landing outcome.
- Barchart to visualize the success rate of each Orbit type.
- Catplot to visualize how Flight Number and Orbit type effect the landing outcome.
- Catplot to visualize how Orbit type and Payload effect the landing outcome.
- Linechart to visualize the launch success yearly rate.

Data Visualization Notebook

EDA with SQL

SQL Queries Performed:

- Display the names of distinct Launch Sites

```
%sql select Distinct LAUNCH_SITE from SPACEXTBL;
```

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

```
%sql select * from SPACEXTBL where LAUNCH_SITE Like 'CCA%' Limit 5;
```

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
```

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL 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 SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';
```

EDA with SQL

SQL Queries Performed:

SQL Notebook

- 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 SPACEXTBL \
    WHERE Landing_Outcome = 'Success (drone ship)' and (4000 < PAYLOAD_MASS__KG_ < 6000) ;
```

- List the total number of successful and failure mission outcomes

```
%sql SELECT count(Mission_Outcome) FROM SPACEXTBL;
```

- 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);
```

- 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 substr(Date,6,2) as month, Date, [Landing_Outcome],BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL \
where [Landing_Outcome] = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

- 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_outcomes FROM SPACEXTBL \
WHERE DATE between '2010-06-04' and '2017-03-20' group by [Landing_Outcome] order by count_outcomes DESC;
```

Build an Interactive Map with Folium

Summary of created map objects to the Folium map:

- Created and added `folium.Circle` and `folium.Marker` for each launch site on the site map and added a text labe for each of the.
- MarkerCluster object to simplify the map containing many markers at the same coordinates.
- MousePosition object to show the coordinates when you hover over the map.
- Folium.polyline to draw a line between two coordinates on the map.

[Folium Notebook]

Build a Dashboard with Plotly Dash

Summary of plots/graphs and interactions added to the dashboard:

- The dashboard contains input components such as a range slider and a dropdown list.
- The dropdown list let choose one of the four Launch Sites.
- A Callback function is created to render a Success-PieChart depending on the selected Launch Site.
- The ranger slide able us to easily select a payload range.
- Another Callback function is created to render a Success-Payload-ScatterChart using the payload range as an input.

Dashboard Code Notebook

Predictive Analysis (Classification)

The process to develop the best model to predict the landing outcome:

1. Load the Data
2. Use the data to create column Y representing landing outcome with values 0 and 1
3. Split the data into training and testing sets
4. Create Logistic Regression, SVM, Decision Tree, KNN and GridSearchCV objects
5. Train the different models using the training set
6. Test the models and find the model that perform the best

Model Development Notebook

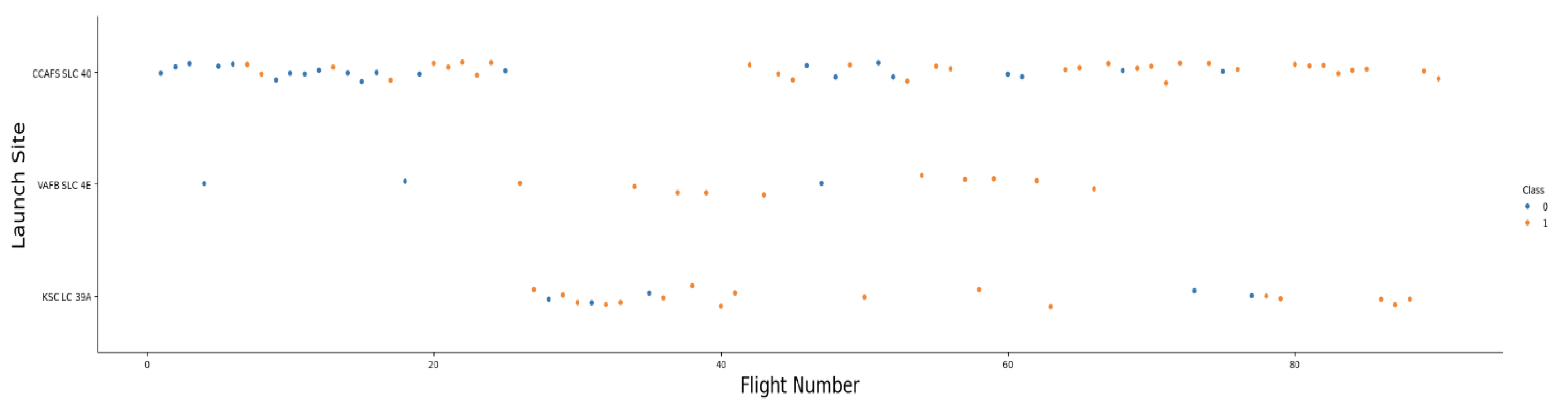
Results

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

Section 2

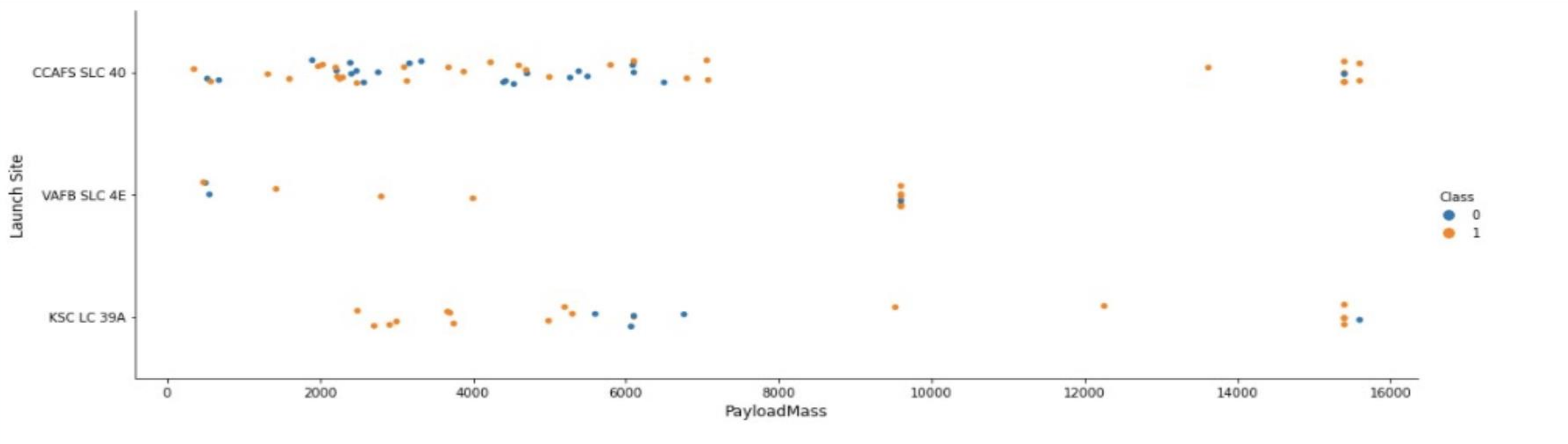
Insights drawn from EDA

Flight Number vs. Launch Site



- We notice the successful landing rate increases with more launches for all launch sites but especially for the CCAFS SLC 40 site.
- Also, we notice VAFB SLC 4E and KSC LC-39A sites has a higher success rate but less launches than CCAFS SLC 40 site.

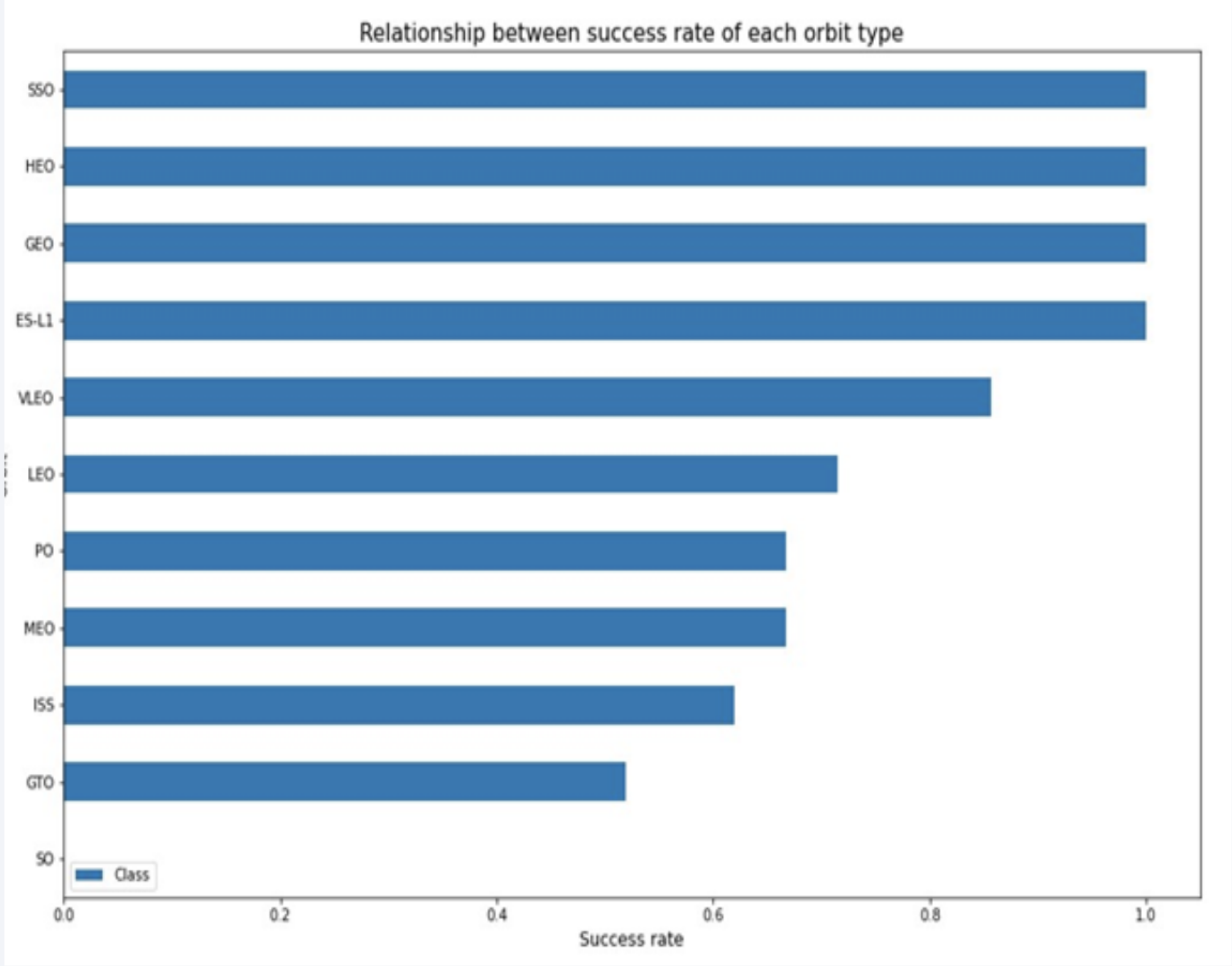
Payload vs. Launch Site



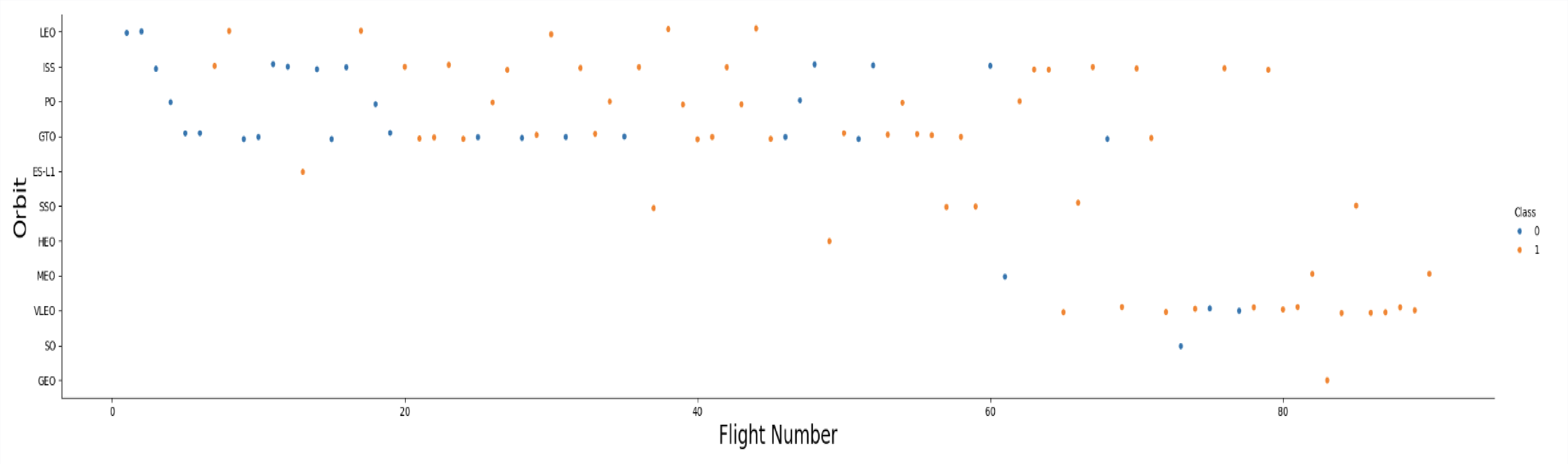
- The VAFB SLC 4E site doesn't have launches with Payload Mass over 10,000 kg.
- Most of the CCAFS SLC 40 site launches has Payload Mass less than 8,000 kg and few launches more than 13,000 kg mass.

Success Rate vs. Orbit Type

- These are the launch sites with the best success rates.
- Notice that some of these sites has fewer launches than others.

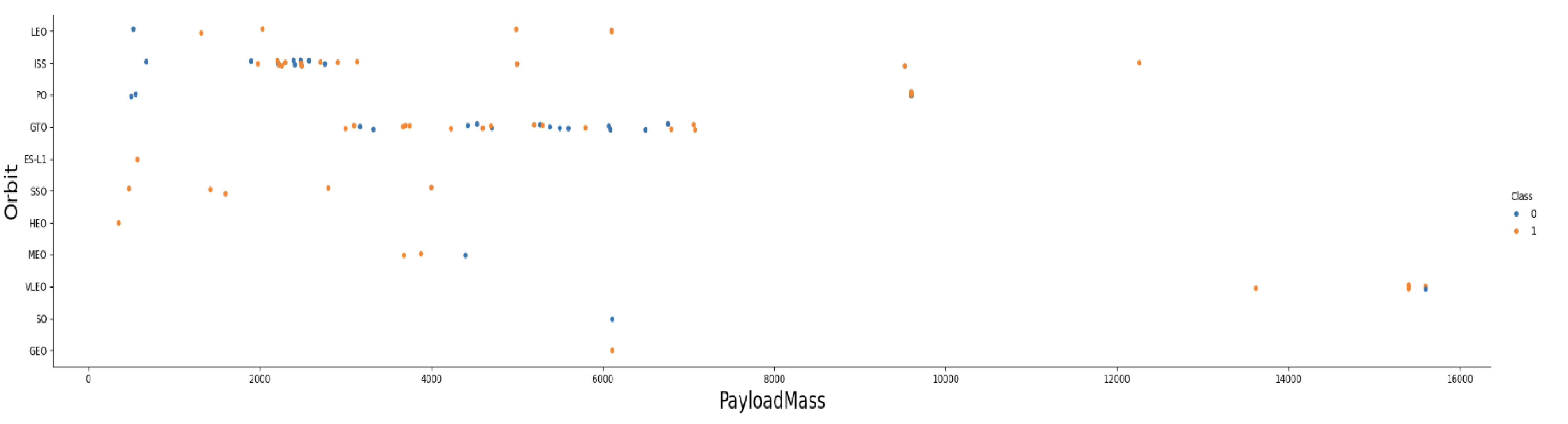


Flight Number vs. Orbit Type



- Success rate increases with more launches for all sites.
- Orbits with fewer launches has better success rate.

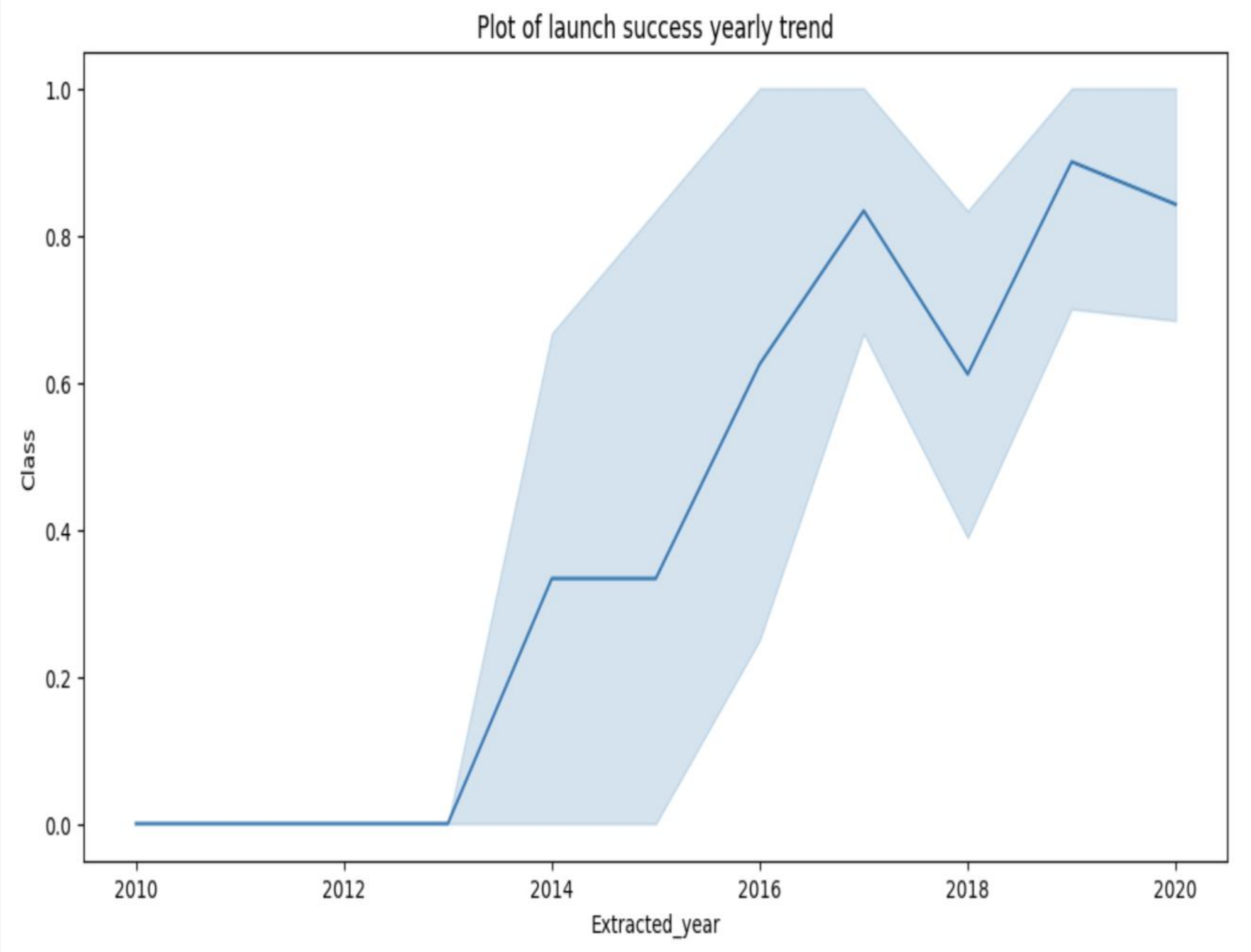
Payload vs. Orbit Type



- Success rate are higher for launches with greater Payload Mass.
- GTO site has mixed outcomes and doesn't have launches with higher Payload Mass.

Launch Success Yearly Trend

- We can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

- There are four distinct Launch Sites in the collected data.

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;
```

```
 * ibm_db_sa://ycy00214:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- 5 records of launch sites starts with the string 'CCA'.

Total Payload Mass

- The total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) \
    FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- 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';
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The date when the first successful landing outcome in ground pad was achieved.

```
%sql SELECT min(Date) \
    FROM SPACEXTBL \
    WHERE Landing_Outcome = 'Success (ground pad)';
```

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT booster_version, payload_mass__kg_, landing_outcome FROM SPACEXTBL \
        WHERE landing_outcome='Success (drone ship)' AND (payload_mass__kg_ BETWEEN 4000 AND 6000) ;
```

```
 * ibm_db_sa://ycy00214:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdc
Done.
```

booster_version	payload_mass__kg_	landing_outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

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

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT mission_outcome, COUNT(mission_outcome) AS TOTAL FROM SPACEXTBL GROUP BY mission_outcome;
```

```
 * ibm_db_sa://ycy00214:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdo
Done.
```

mission_outcome	total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- The total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT(booster_version), (SELECT MAX(payload_mass__kg_) AS "maximum_payload_mass" FROM SPACEXTBL) FROM SPACEXTBL
```

```
 * ibm_db_sa://ycy00214:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

booster_version	maximum_payload_mass
F9 B4 B1039.2	15600
F9 B4 B1040.2	15600
F9 B4 B1041.2	15600
F9 B4 B1043.2	15600
F9 B4 B1039.1	15600

- The names of the booster_versions which have carried the maximum payload mass using a subquery.

2015 Launch Records

```
%sql SELECT substr(Date,6,2) as month, Date, [Landing_Outcome],BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL \
where [Landing_Outcome] = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

month	Date	Landing_Outcome	Booster_Version	Launch_Site
01	2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

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

```
%sql SELECT [Landing_Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '2010-06-04' and '2017-03-20' group by [Landing_Outcome] order by count_outcomes DESC;
```

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

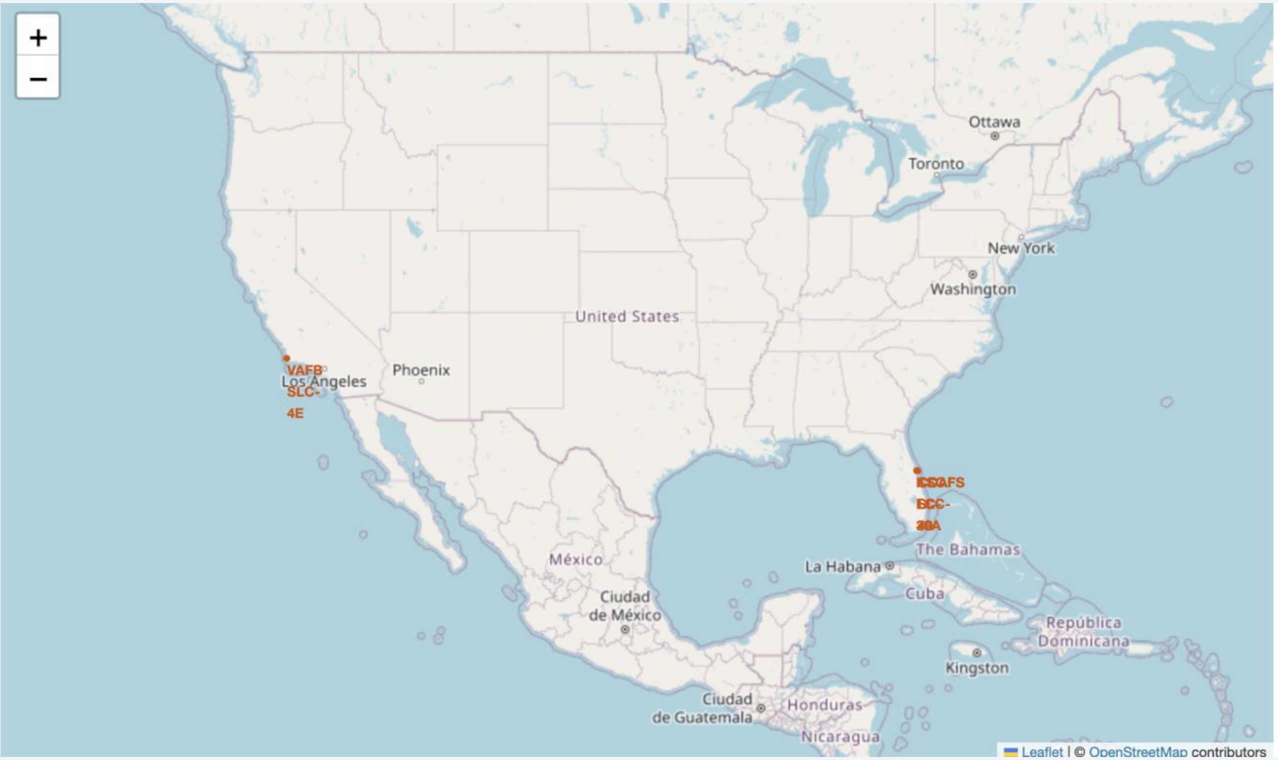
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

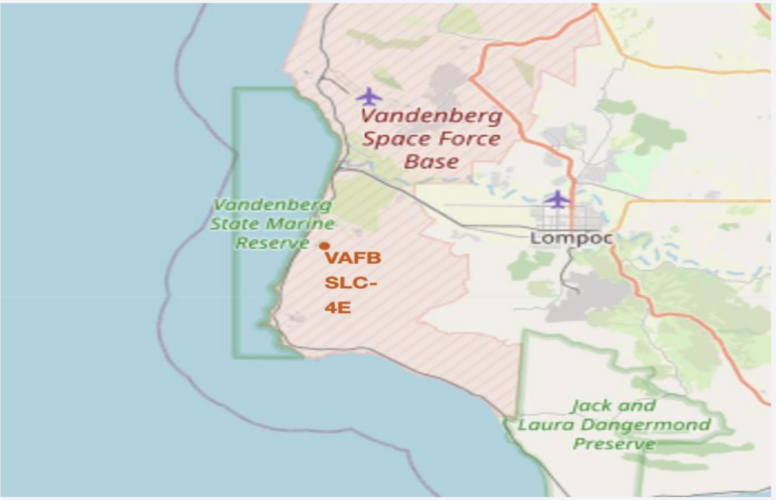
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Section 3

Launch Sites Proximities Analysis

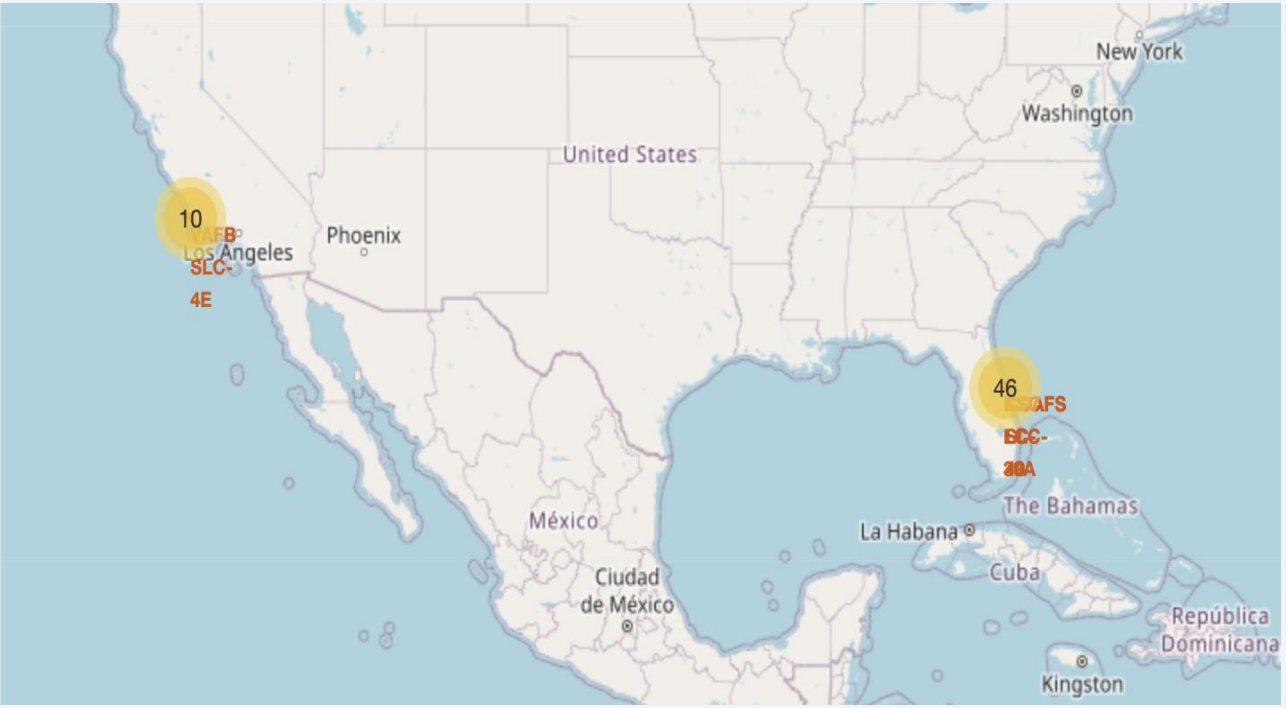
All Launch Sites





Looking at the zoomed in pictures, we notice that there three launch sites very close from each other. All sites are located on the coast.

Launches Outcomes



We can see launches clusters on the map for each site. When we zoom in and click on it, we see the success/fail launches represented by green and red colors.

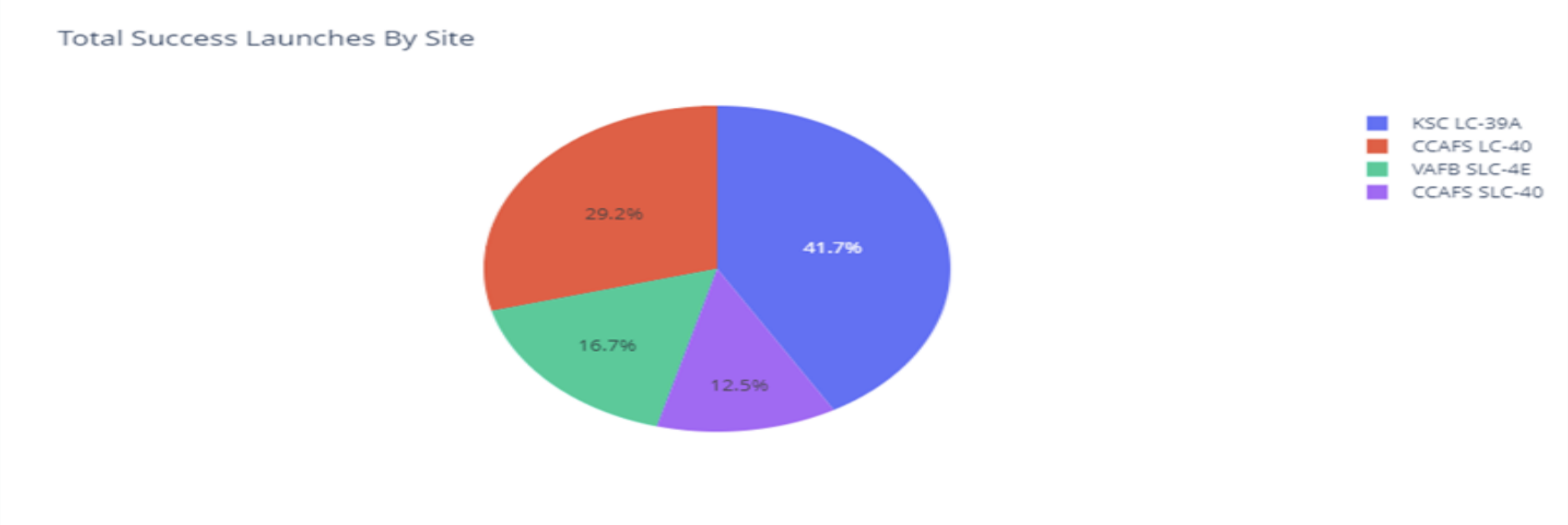
Launch Site to its Proximities

Launch sites are near coastlines, roads and highways. All launch sites maintain safe distance to main cities for safety reasons. The shows a line connecting a launch site to its nearest highway.

Section 4

Build a Dashboard with Plotly Dash

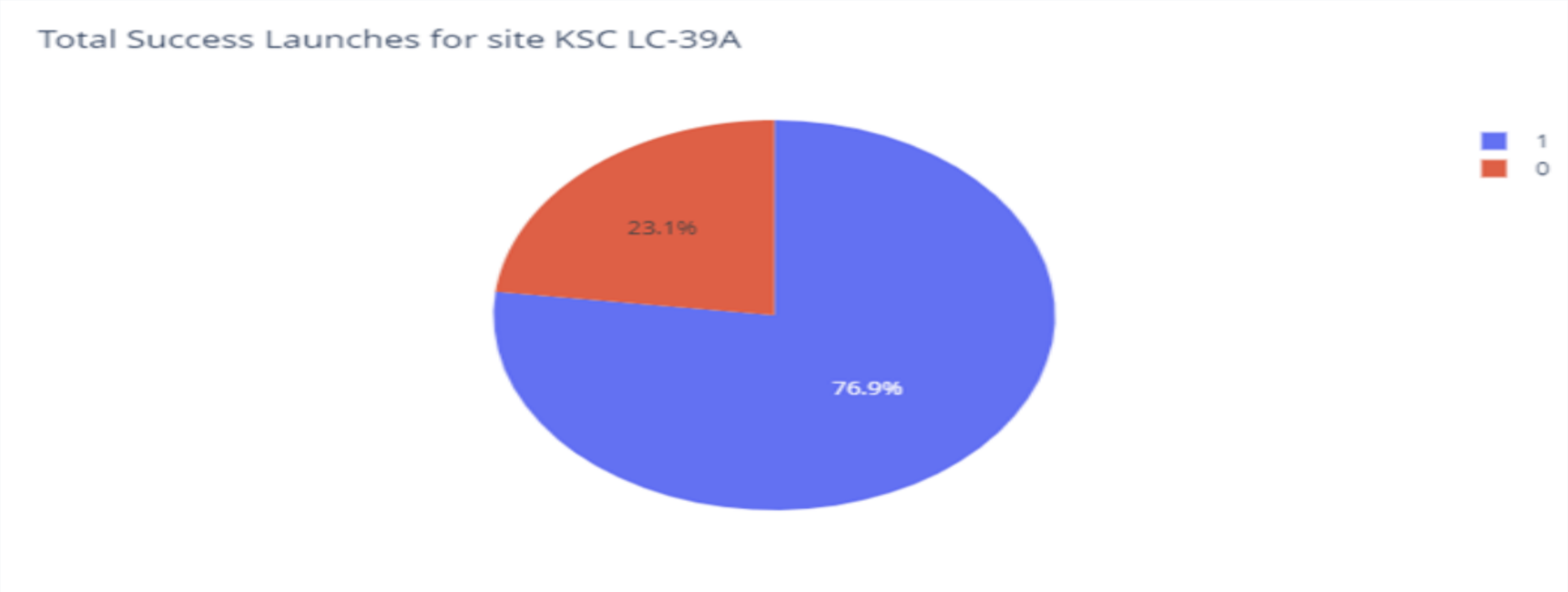
Success Launches



The site KSC LC-39A has highest success launches whils CCAFS SLC-40 has lowest success launches

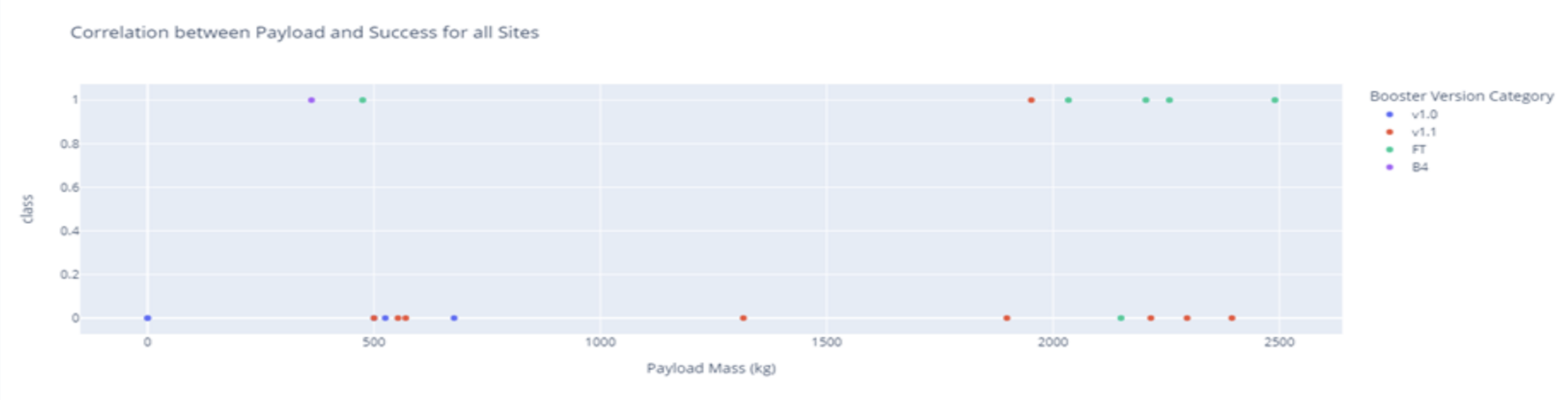
Highest Success Ratio Site

The highest success ratio site KSC LC-39A has 76.9% success ratio



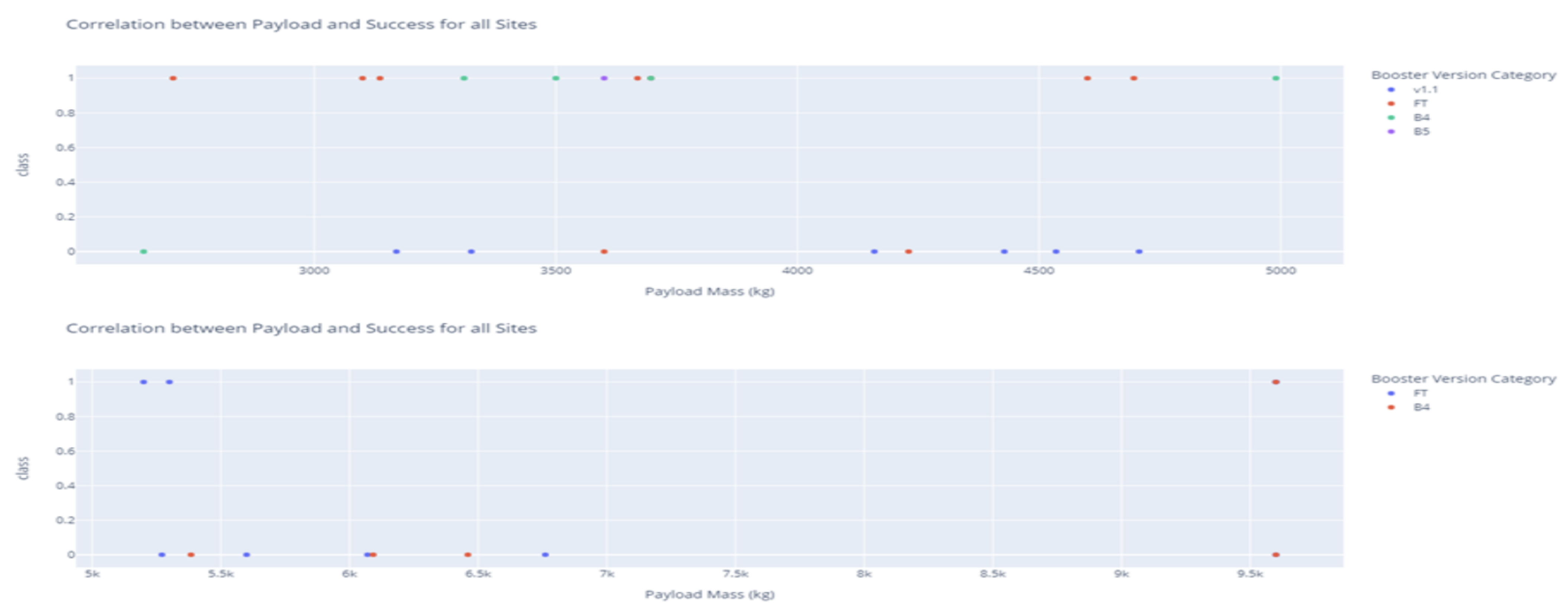
Payload Launch Outcome

- Selecting different Payload Mass ranges we notice that the highest success launches are in between 2500kg – 5000kg payload mass.



Payload Launch Outcome

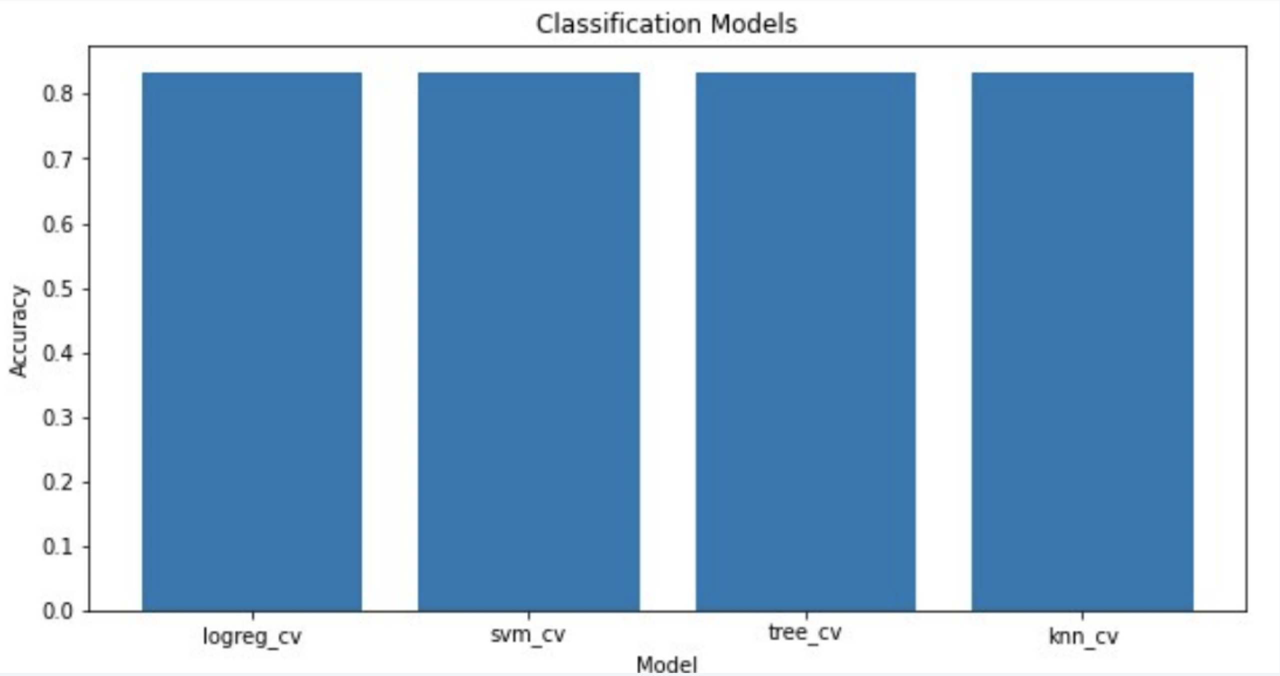
- Selecting different Payload Mass ranges we notice that the highest success launches are in between 2500kg – 5000kg payload mass.



Section 5

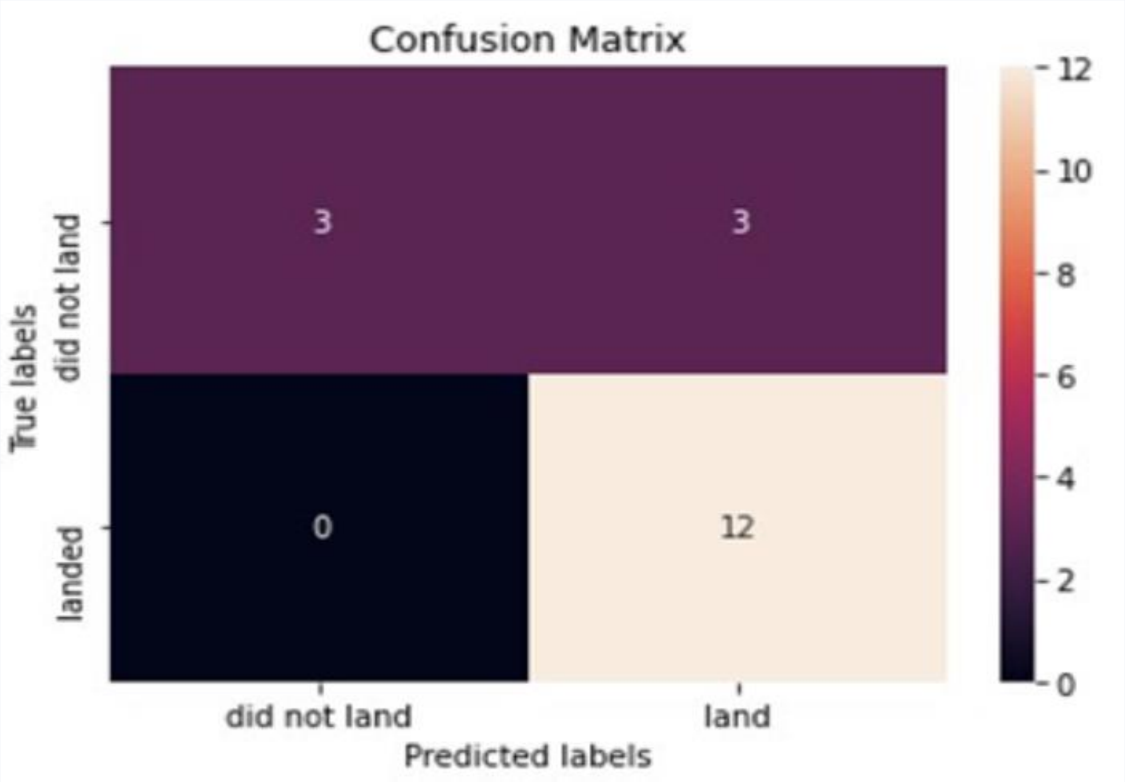
Predictive Analysis (Classification)

Classification Accuracy



- All models have the same accuracy on test data: 0.833
- Decision Tree model had the best_score accuracy of `0.8732142857142856`

Confusion Matrix



- All models had the same confusion matrix.

Conclusions

- Using our machine learning models, we can predict whether a launch will have a successful landing or not.
- The best model to predict the outcome of the landing is the Decision Tree Model.
- We found the Launch site with highest success rate to be KSC LC-39A site.
- All launch sites have a safe distance from cities.

Appendix

- You can check this Github repository to find the notebooks, codes and results found for this work:

[Repository Link]

Thank you!