IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Zhijie Zhao
2025-03-15

Outline

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

Executive Summary

- Summary of methodologies

  1. Data Collection by Web Scraping and SpaceX API;
  2. Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive dashboard;
  3. Machine Learning Prediction.

- Summary of all results

  1. Data collected from Web Scraping and SpaceX API;
  2. EDA results with visualization. An interactive dashboard is also available;
  3. Machine Learning results with various models.

Introduction

- Project background and context

  SpaceX is a successful company of the commercial space age, making space travel affordable. In this capstone, we are the data scientists working for a new rocket company named SpaceY, which is a competitor of SpaceX. Our job is analyzing SpaceX data to determine the price of each launch. Using the public data and machine learning models, we are going to predict if SpaceX will reuse the first stage of the rocket.

- Problems you want to find answers

  1. How do variables affect the success of the first stage landing?
  2. Is a success launch related to the location or time?
  3. Which machine learning model is the best for prediction?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API and Web Scraping from Wikipedia
- Perform data wrangling
  - Filtering data, dealing with missing values, and using One Hot Encoding to label data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We have built four models: LogisticRegression, SVM, Decision Tree and KNN. These models are tuned by GridSearch. The evaluation metric is accuracy.

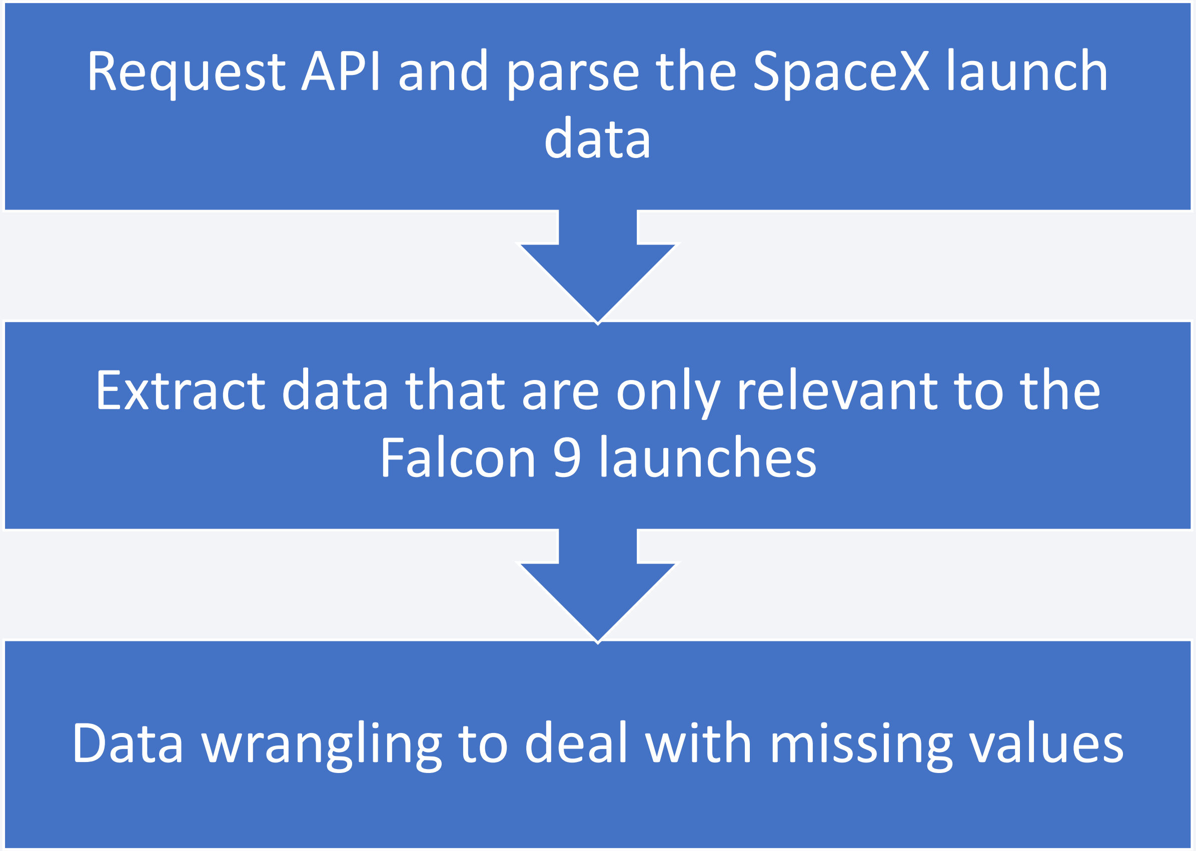
Data Collection

- Data are collected from the SpaceX Rest API (https://api.spacexdata.com/v4/) and the Wikipedia page by using the Web Scraping technique.
- The raw data contain massive information. We have to extract relevant data and clean them for further analysis.

Data Collection – SpaceX API

- SpaceX offers a public API to share their data, so we can write some python codes to obtain these data.
- Raw data contain massive information. We have to extract the relevant data. In this case, we are interested in the Falcon 9 launches.
- Finally, we use the data wrangling technique to deal with missing values.

GitHub Link: Data Collection API



Request API and parse the SpaceX launch data

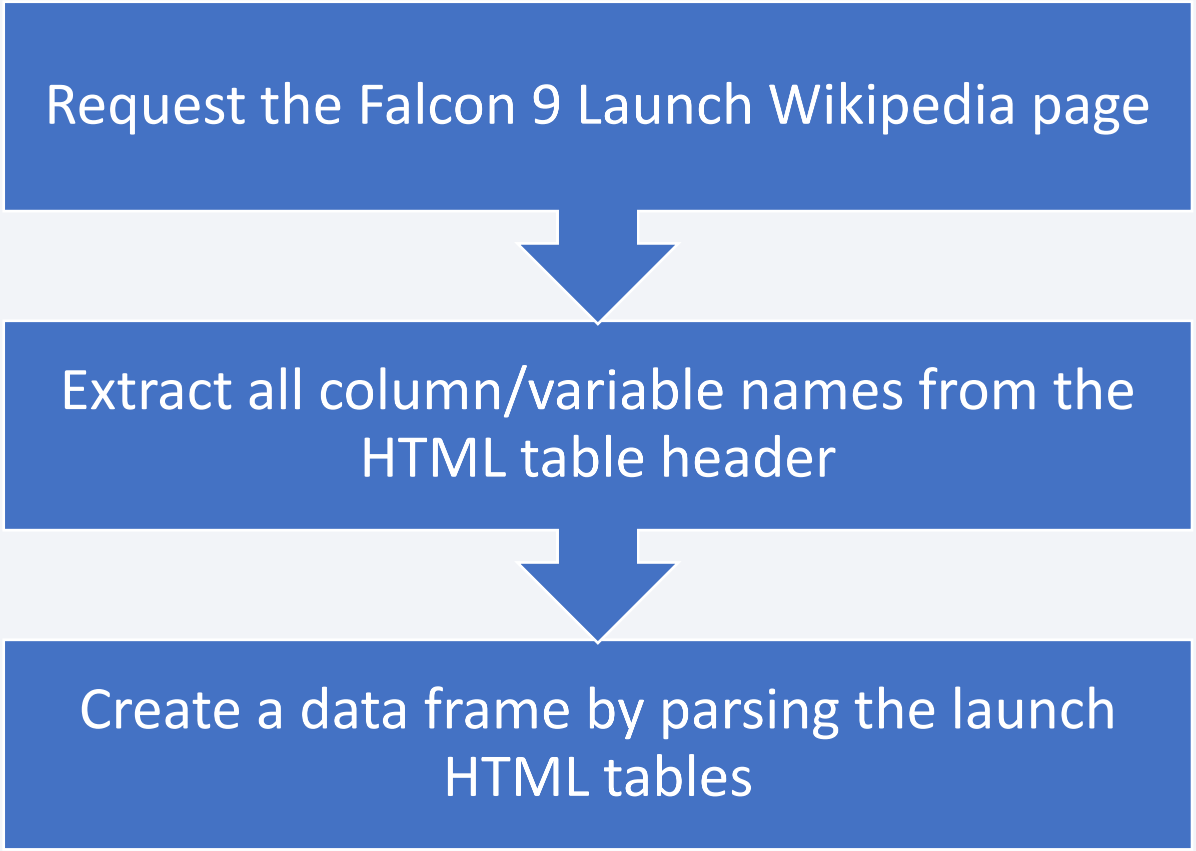
Extract data that are only relevant to the Falcon 9 launches

Data wrangling to deal with missing values

Data Collection - Scraping

- Data can be obtained from the Wikipedia page of SpaceX launches.
- BeautifulSoup is used to scrape data from the Wikipedia page.
- Data are extracted from a table and parsed to a pandas data frame.

GitHub Link: Web Scraping



Request the Falcon 9 Launch Wikipedia page

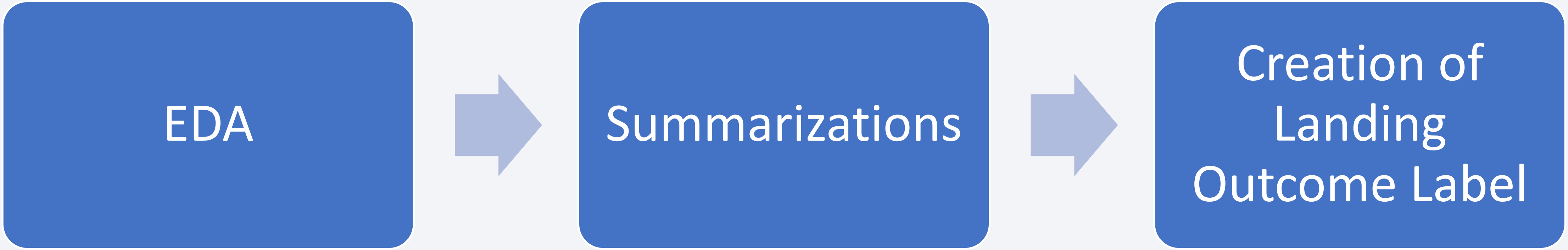
Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Data Wrangling

- First, the initial Exploratory Data Analysis (EDA) is performed understand data.
- Second, we calculate the launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type.
- Finally, the landing outcome label is created from Outcome column.

GitHub Link: Data Wrangling



EDA

Summarizations

Creation of Landing Outcome Label

EDA with Data Visualization

- The dominant charts are scatter plots. These plots are: FlightNumber vs. PayloadMass, FlightNumber vs. LaunchSite, PayloadMass vs. LaunchSite, FlightNumber vs. Orbit, and PayloadMass vs. Orbit. These plots reveal the relations between each two variables.
- In addition, a bar chart for the success rates of each orbits. It is shown that the ES-L1, GEO, HEO, and SSO have the highest success rates.
- Finally, a line chart is created to visualize the launch success yearly trend.

GitHub Link: EDA with Data Visualization

EDA with SQL

- The following SQL queries are 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 achieved;
    - 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;
    - 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 landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

GitHub Link: EDA with SQL

Build an Interactive Map with Folium

- Markers are added to all launch sites by their latitude and longitude coordinates.
- Circles are added to highlighted circle area around the launch sites.
- Further colored markers are added to show the results of launches.
- Lines are added to show distances between the Launch Site to Coastline, closest city, railway, and highway.

GitHub Link: Interactive Visual Analytics with Folium

Build a Dashboard with Plotly Dash

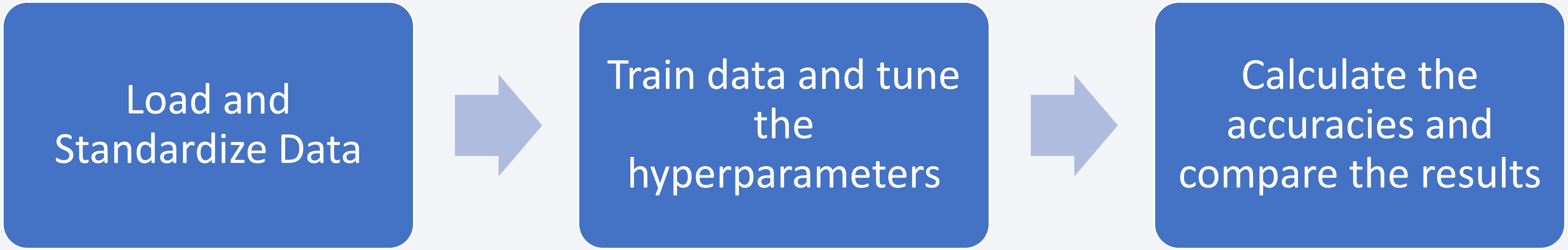
- A Launch Sites Dropdown List is added to select the Launch Site we are interested in.
- A Pie Chart is added to show the success rates of launches at all sites or a certain site.
- A Slider of Payload Mass Range is added to select the range of Payload Mass.
- A Scatter Plot of Payload Mass vs. Launch Site is added to show the success rates of alle sites or a certain site with different Booster Versions.

GitHub Link: SpaceX Dash App

Predictive Analysis (Classification)

- Four classification models are built for comparison: Logistic Regression, Support Vector Machine, Decision Tree, and K Nearest Neighbors.
- The evaluation metric is accuracy in this case.
- Hyperparameter tuning is performed by the Grid Match method.
- Finally, their accuracies are compared to find the best model.

GitHub Link: Machine Learning Prediction



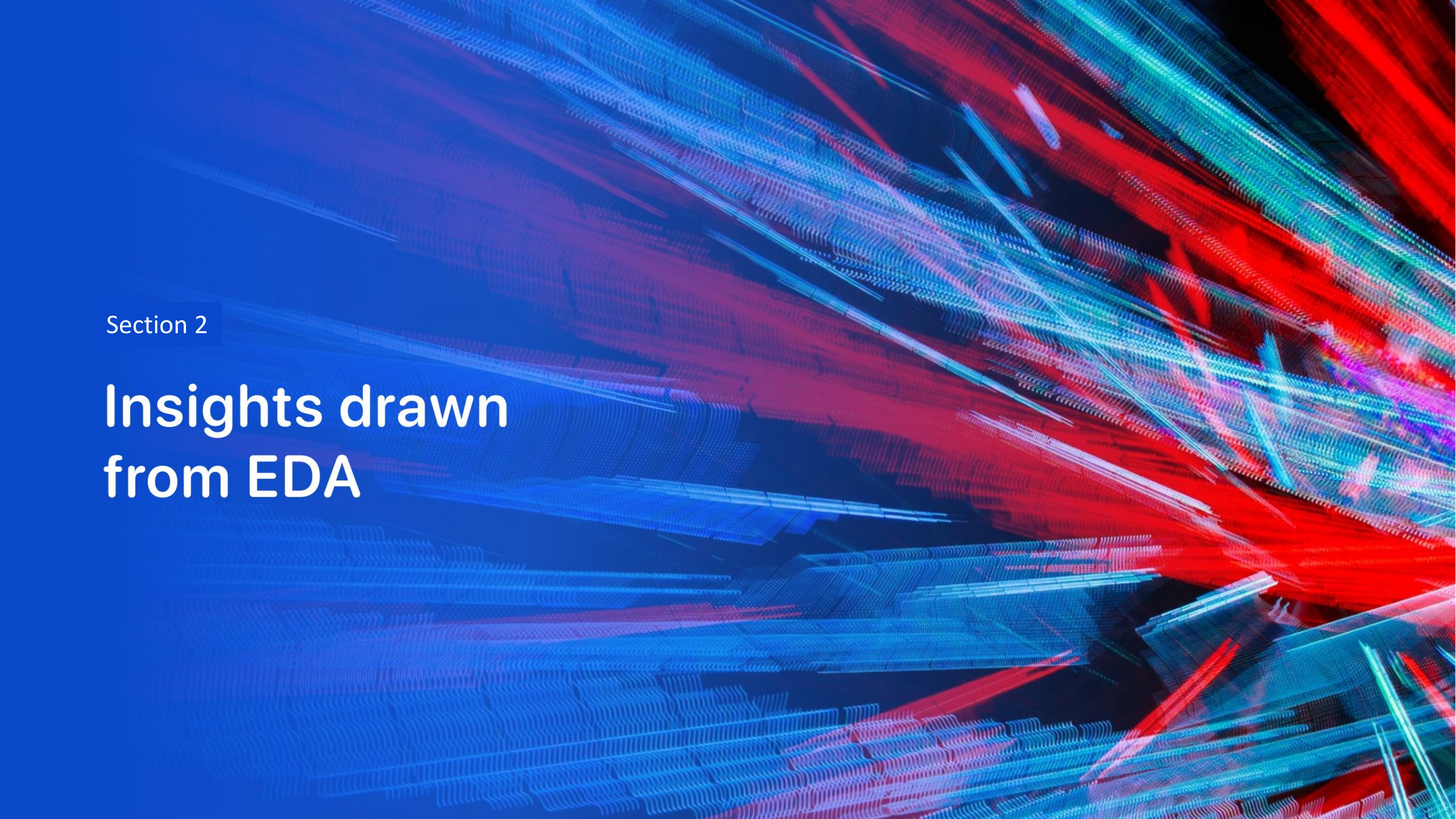
Load and Standardize Data

Train data and tune the hyperparameters

Calculate the accuracies and compare the results

Results

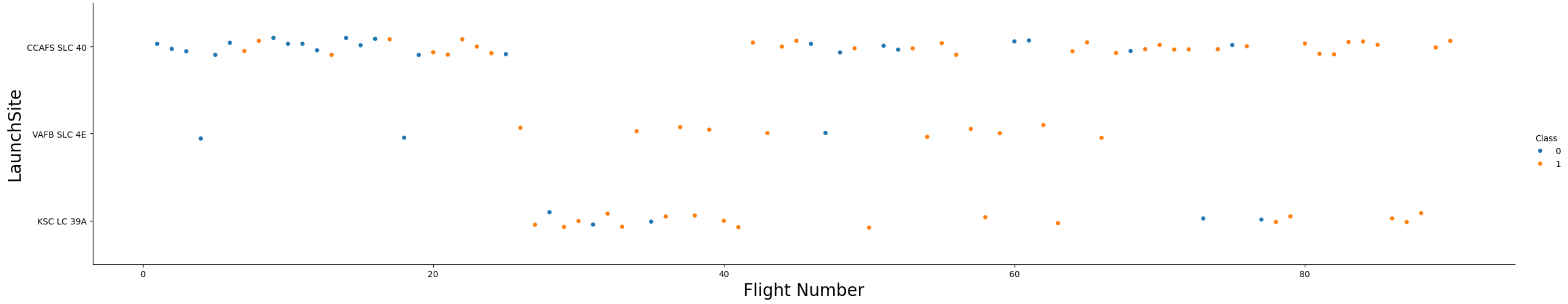
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

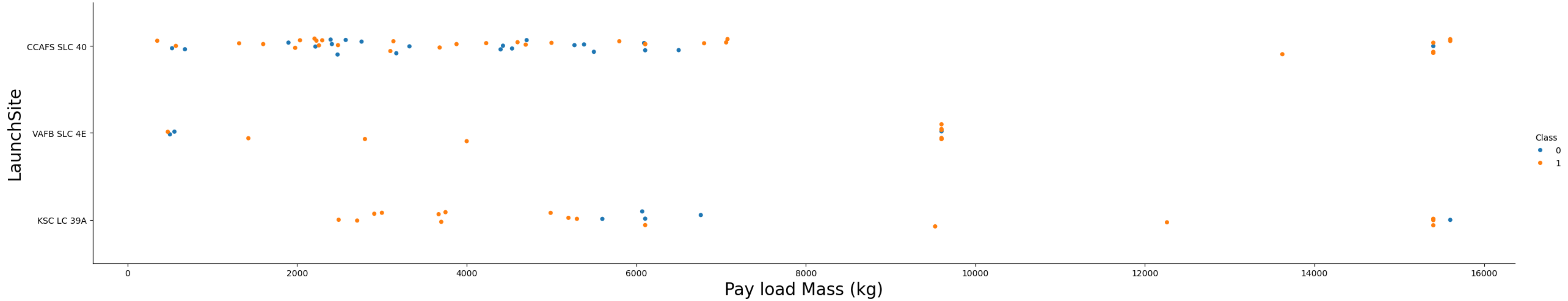
Insights drawn from EDA

Flight Number vs. Launch Site



- The earlier flights are tended to be fail, while the later flights are tended to be success.
- The CCAFS SLC 40 has the most launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.

Payload vs. Launch Site



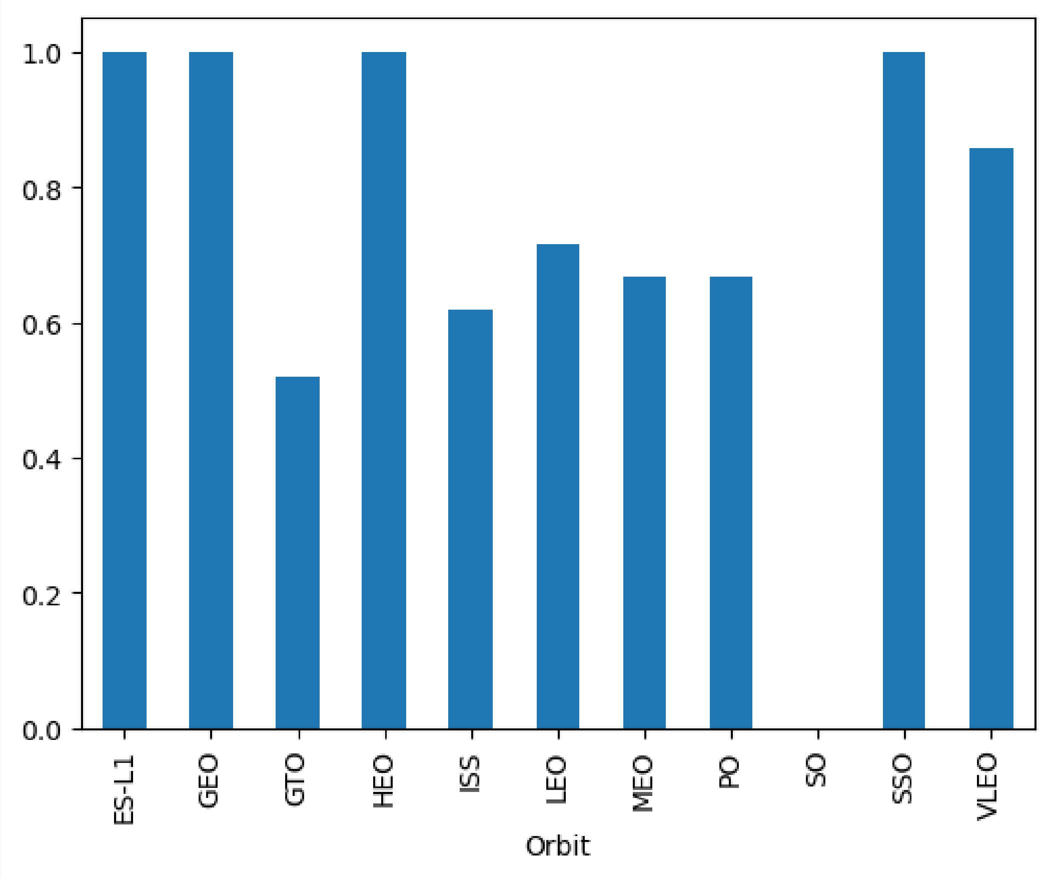
- For every launch site, the higher the payload mass, the higher the success rate is observed.

- For the VAFB-SLC launch site, there are no rockets launched for heavy payload mass (greater than 10000).

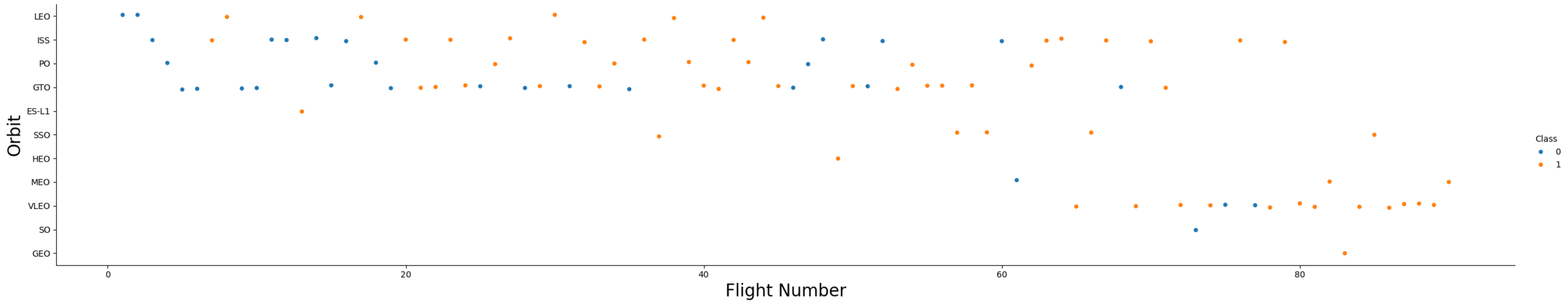
- The light rockets (under 5500 kg) are launched successfully at the KSC LC 39A.

Success Rate vs. Orbit Type

- Orbits with 100% success rate: ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate: SO
- Success rate between 50% and 85%: GTO, ISS, LEO, MEO, PO, VLEO

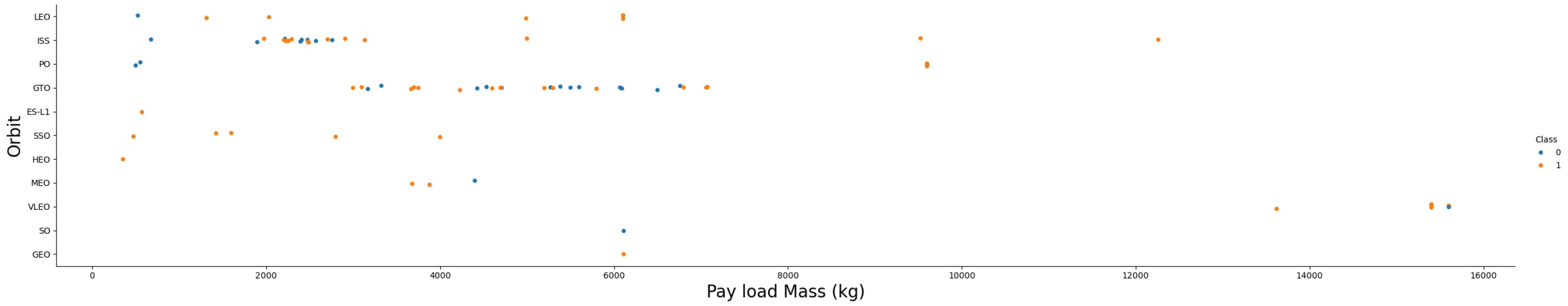


Flight Number vs. Orbit Type



- One can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type

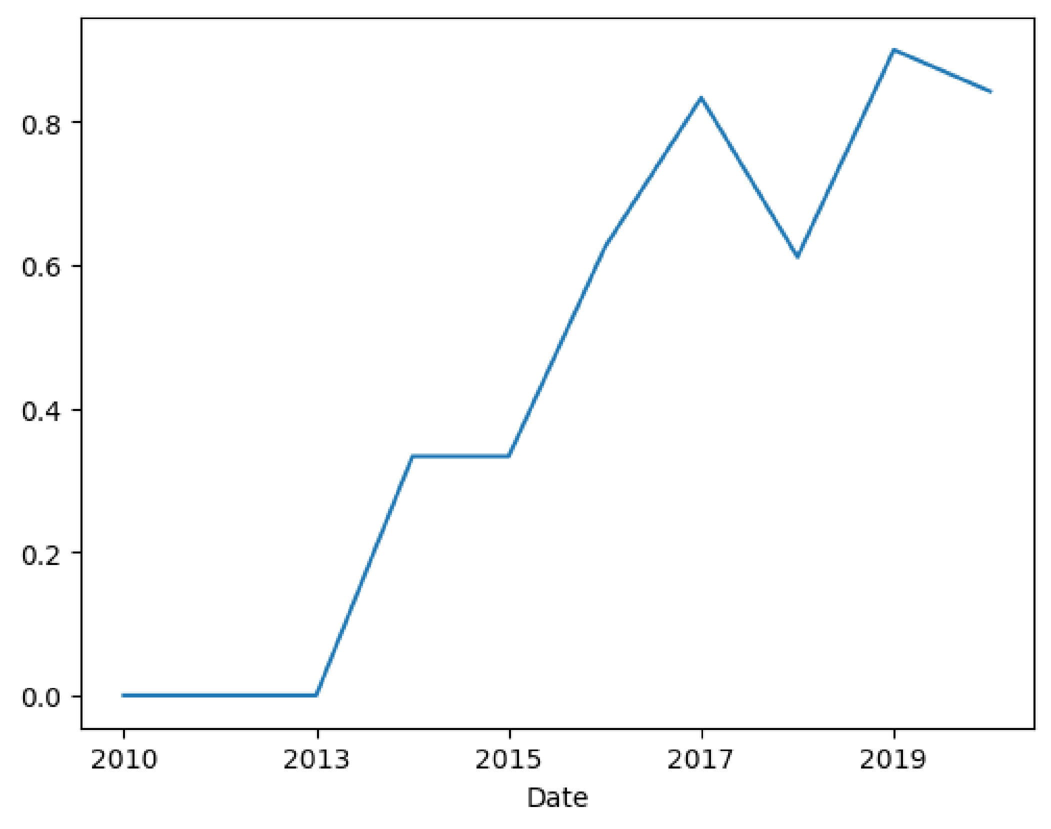


- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

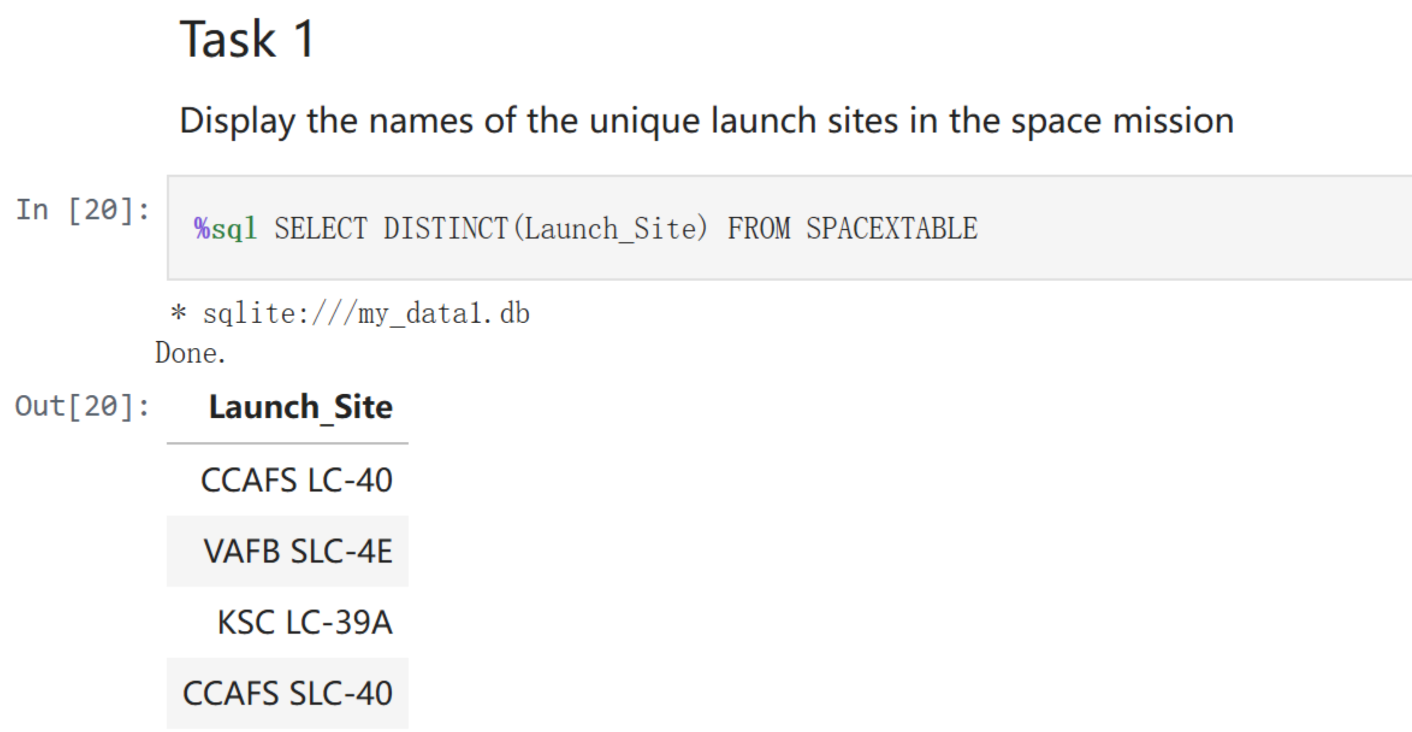
Launch Success Yearly Trend

- One can observe that the success rate since 2013 kept increasing till 2020.
- But there are dropdowns on 2017 and 2019.



All Launch Site Names

- This query gives the names of all Launch Site Names. They are CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.



Task 1

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

```
In [20]: %sql SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE
```

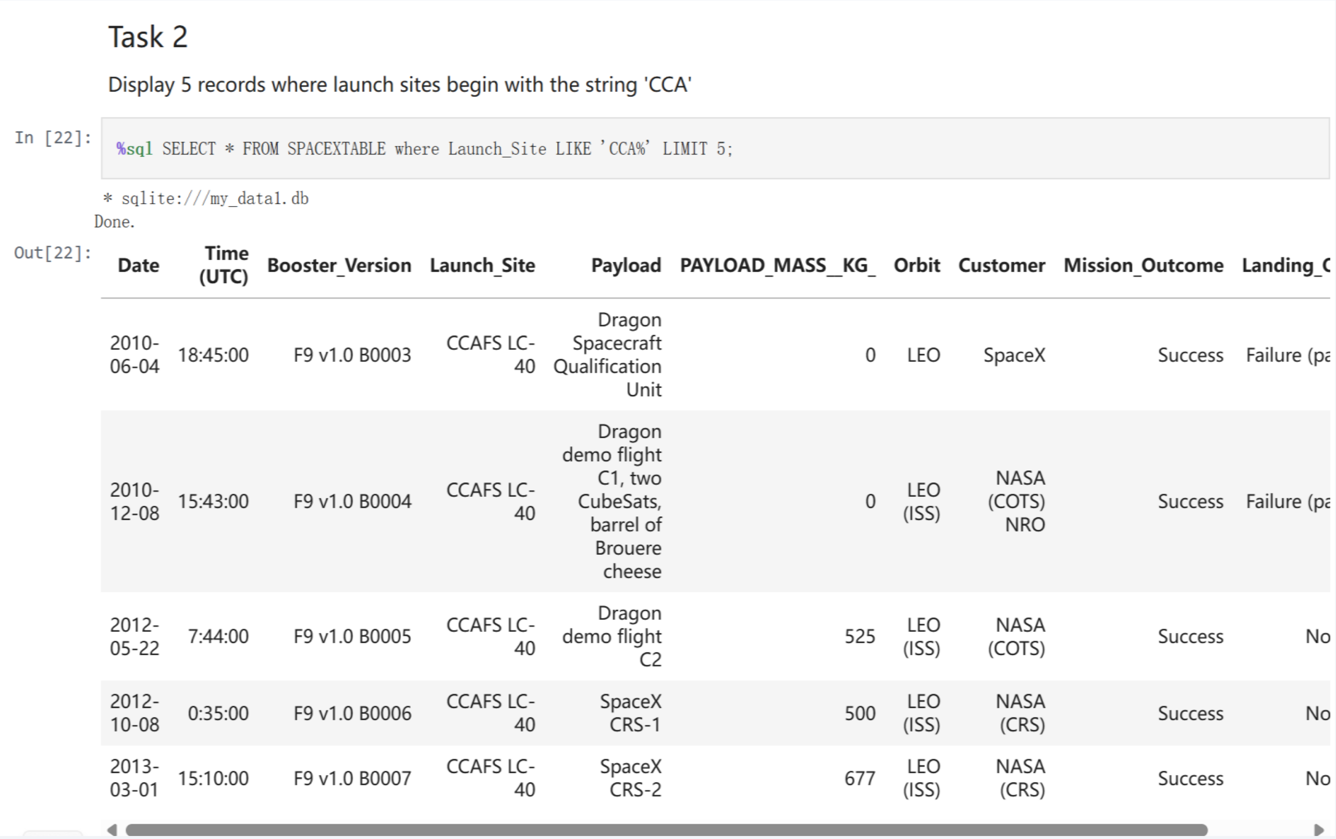
```
 * sqlite:///my_data1.db
Done.
```

Out[20]:

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

Launch Site Names Begin with 'CCA'

- This query finds 5 records where launch sites begin with `CCA`.
- These sample are data of CCAFS LC-40 and the orbits are LEO.



Task 2

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

```
In [22]:  %sql SELECT * FROM SPACEXTABLE where Launch_Site LIKE 'CCA%' LIMIT 5;
```

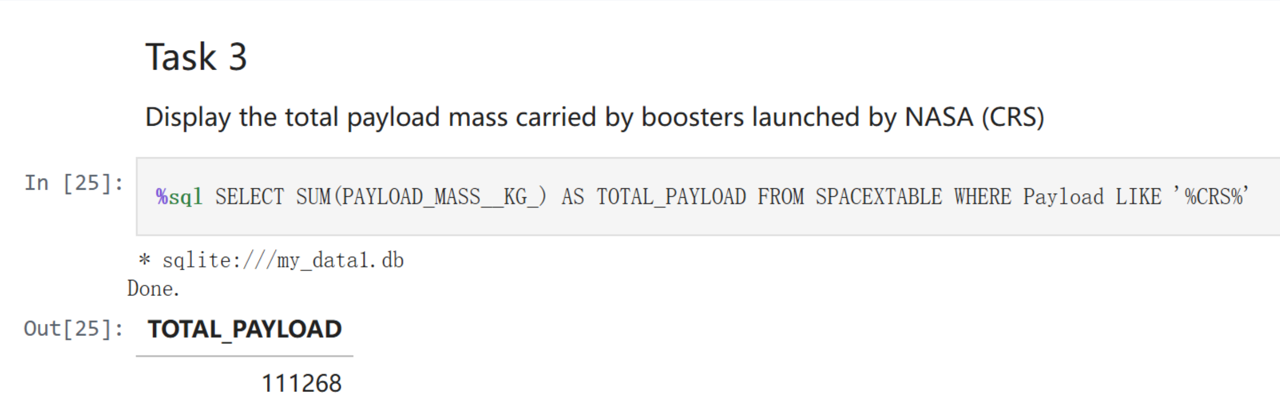
```
 * sqlite:///my_data1.db
Done.
```

Out[22]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_C
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (pa
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 (pa
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No

Total Payload Mass

- This query calculates the total payload carried by boosters from NASA
- Present your query result with a short explanation here



Task 3

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

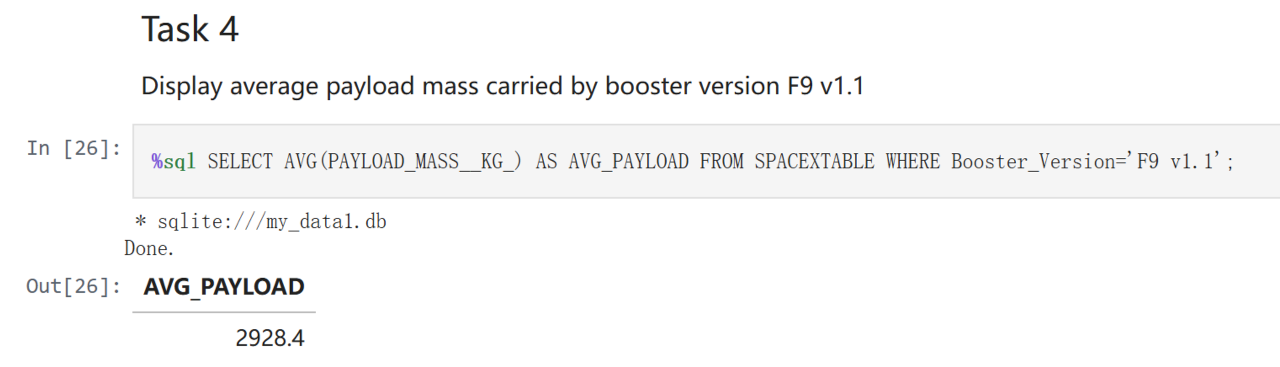
```
In [25]:  %sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTABLE WHERE Payload LIKE ’%CRS%’

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

Out[25]:

TOTAL_PAYLOAD
111268

Average Payload Mass by F9 v1.1

- This query calculates the average payload mass carried by booster version F9 v1.1.



Task 4

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

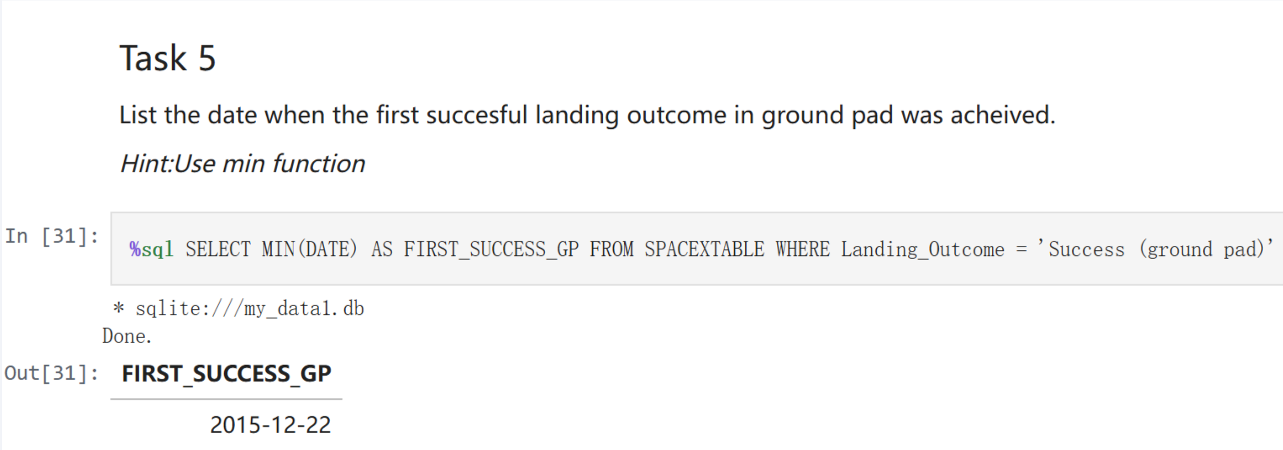
```
In [26]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTABLE WHERE Booster_Version='F9 v1.1';

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

Out[26]:

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

- This query finds the date of the first successful landing outcome on ground pad.
- It is 2015-12-22.



Task 5

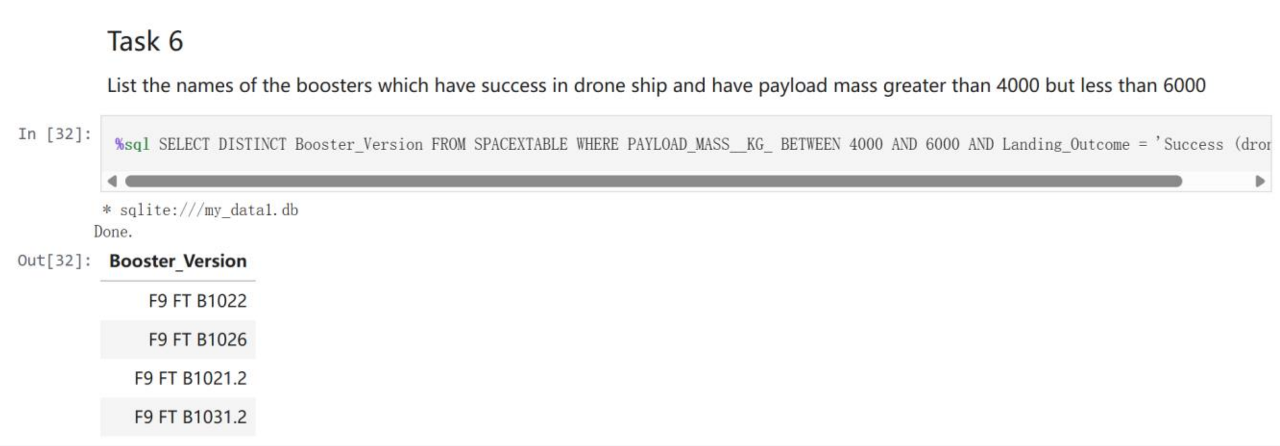
List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
In [31]:  %sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'

           * sqlite:///my_data1.db
          Done.
Out[31]:  FIRST_SUCCESS_GP
                2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query lists the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.
- Only four records satisfies this requirement.



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

```
In [32]: %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND Landing_Outcome = 'Success (dror
```

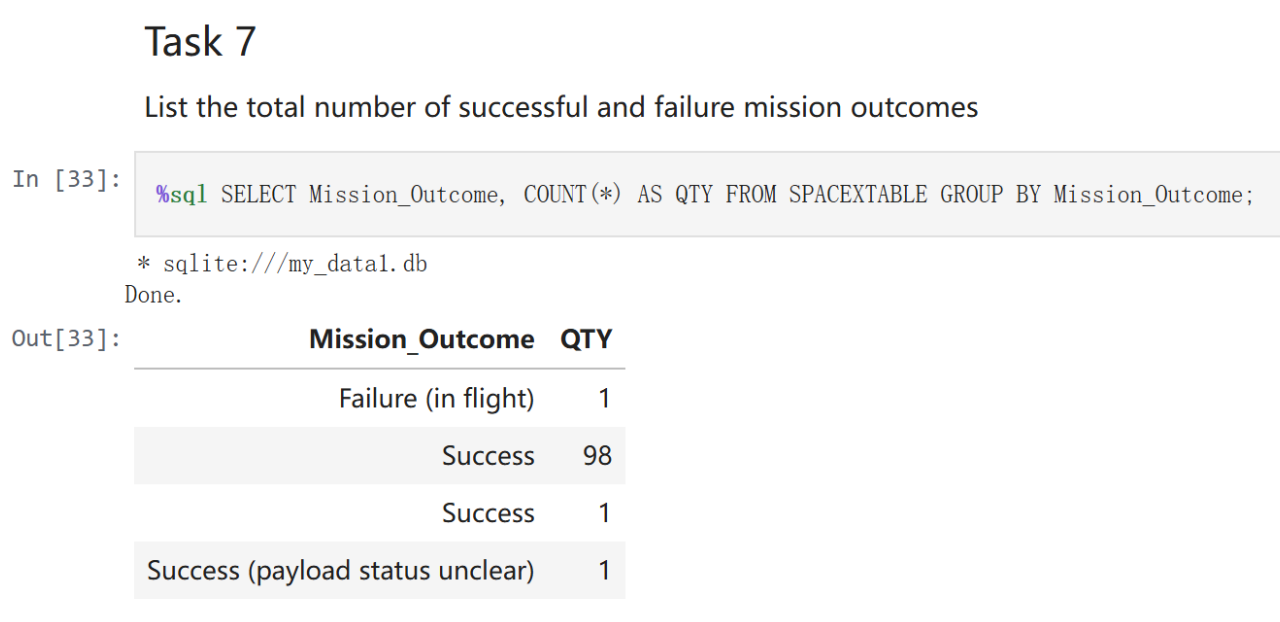
```
 * sqlite:///my_data1.db
Done.
```

Out[32]:

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

Total Number of Successful and Failure Mission Outcomes

- This query calculates the total number of successful and failure mission outcomes.
- There are two success in the tables. An unseen space may be in the data.



Task 7

List the total number of successful and failure mission outcomes

```
In [33]:  %sql SELECT Mission_Outcome, COUNT(*) AS QTY FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

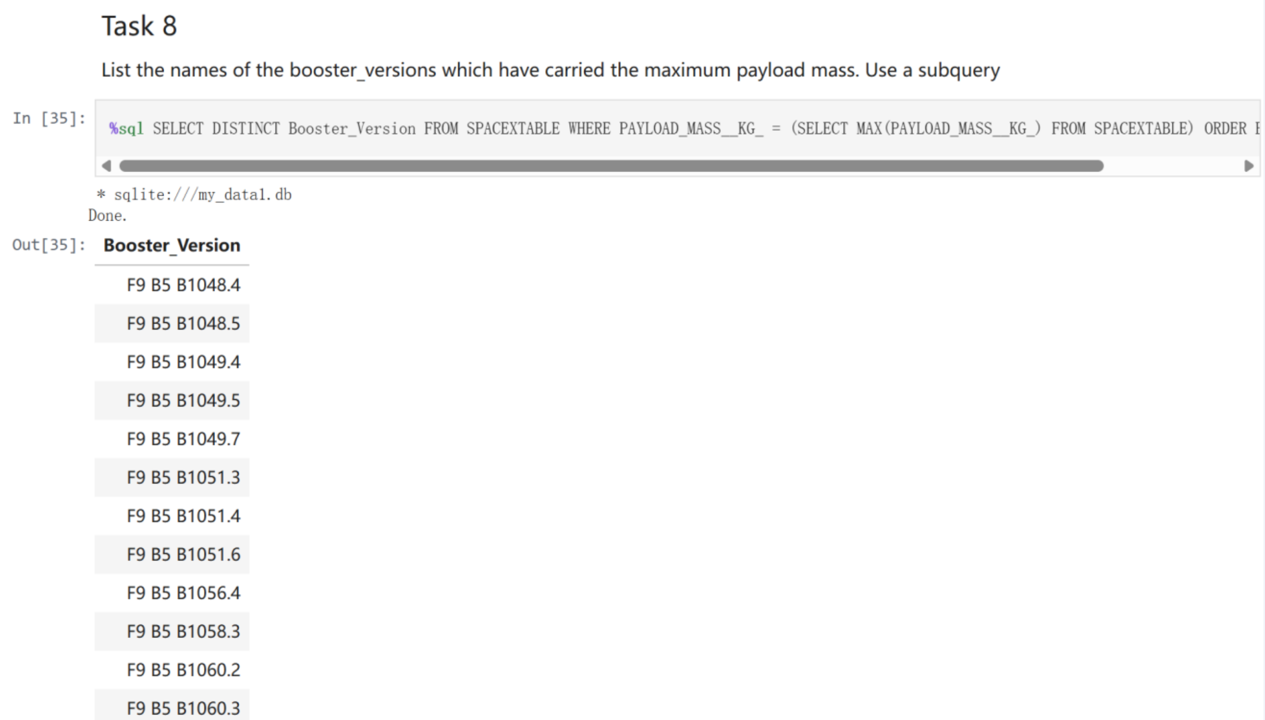
```
 * sqlite:///my_data1.db
Done.
```

Out[33]:

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

Boosters Carried Maximum Payload

- This query lists the names of the booster which have carried the maximum payload mass.
- There are 12 records.



Task 8

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

```
In [35]: %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE) ORDER B
```

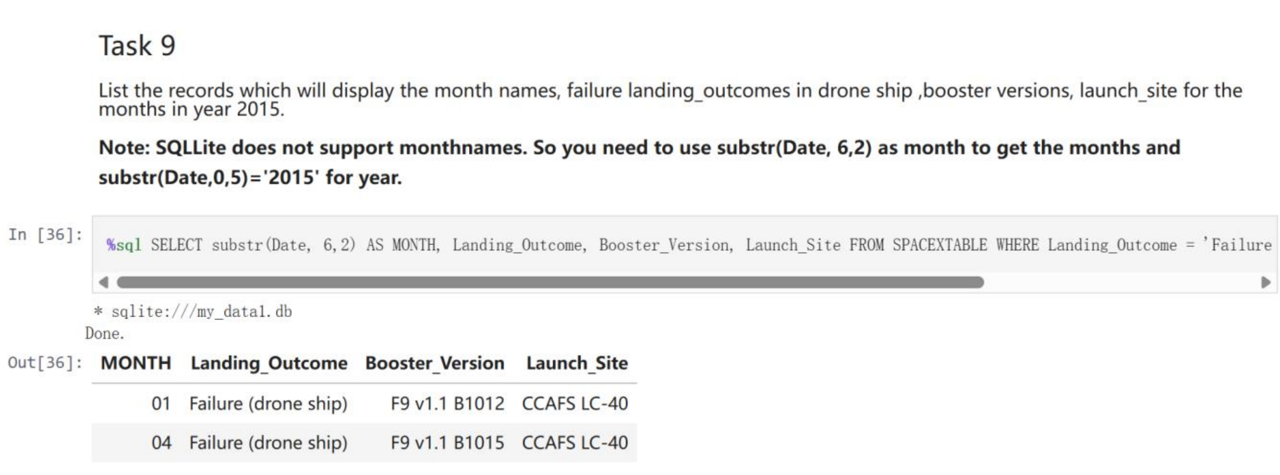
```
 * sqlite:///my_data1.db
Done.
```

Out[35]:

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

2015 Launch Records

- This query lists the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.
- Only two records in this case.



Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
In [36]: %sql SELECT substr(Date, 6,2) AS MONTH, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE Landing_Outcome = 'Failure
```

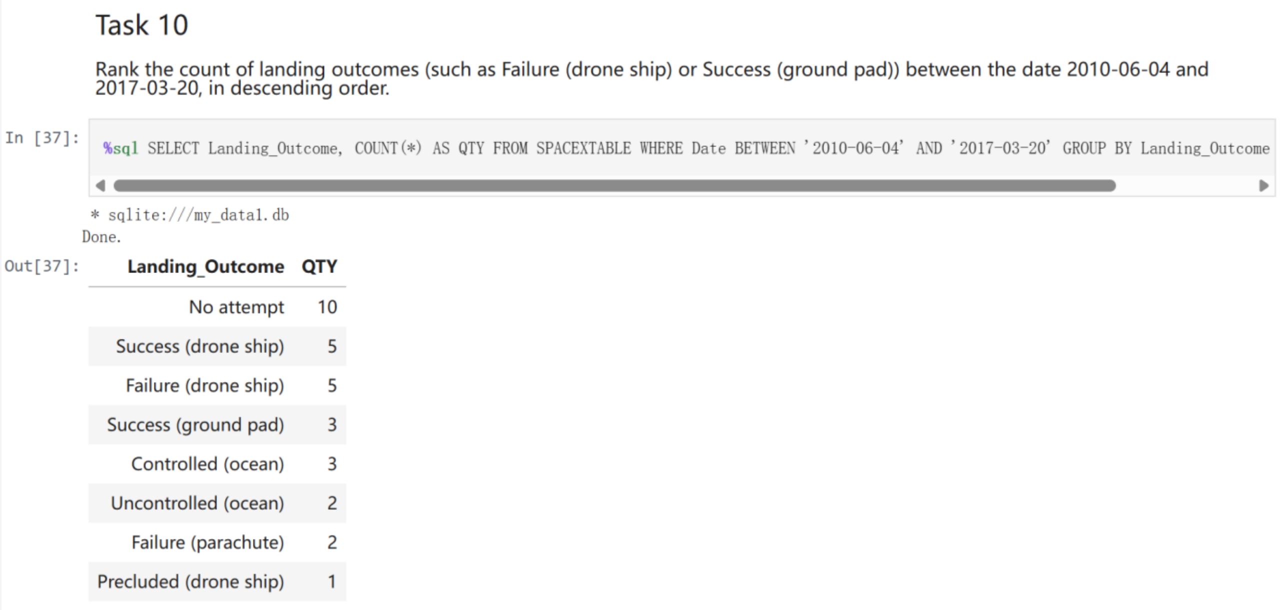
```
 * sqlite:///my_data1.db
Done.
```

Out[36]:

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

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

- This query can 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.
- There are 8 rows. The most is no attempt.



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.

```
In [37]: %sql SELECT Landing_Outcome, COUNT(*) AS QTY FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome
```

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

Out[37]:

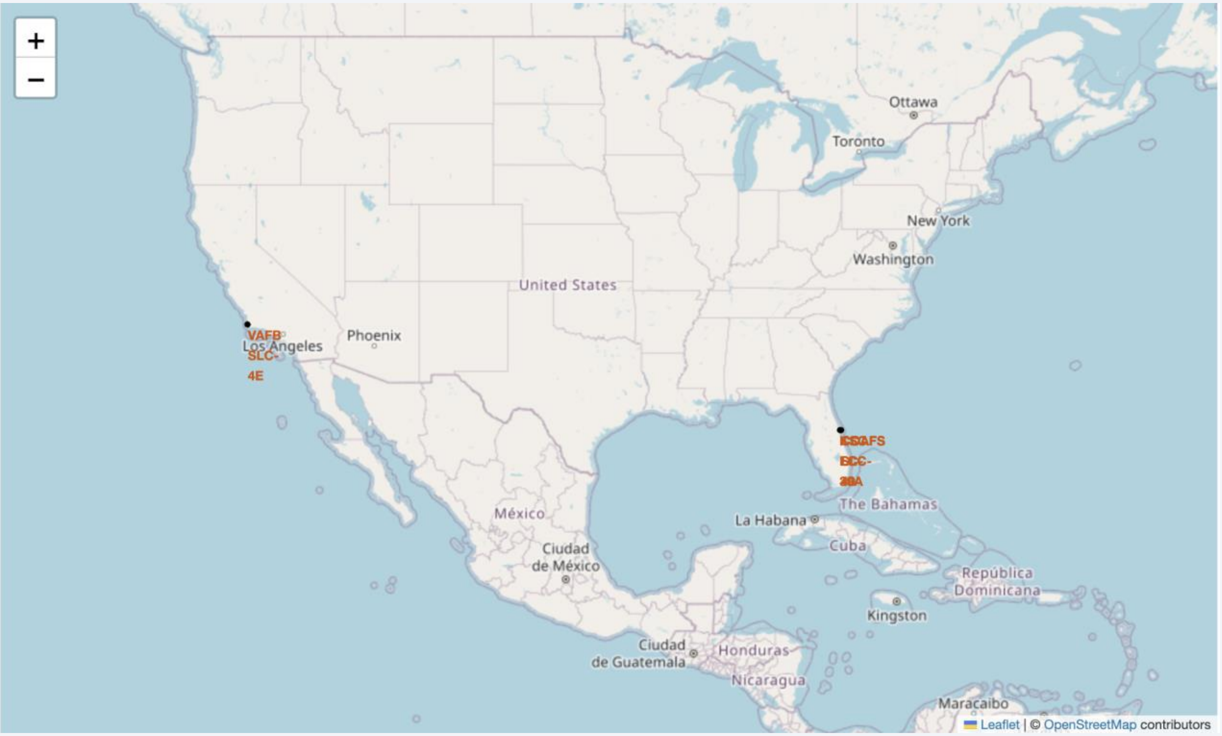
Landing_Outcome	QTY
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

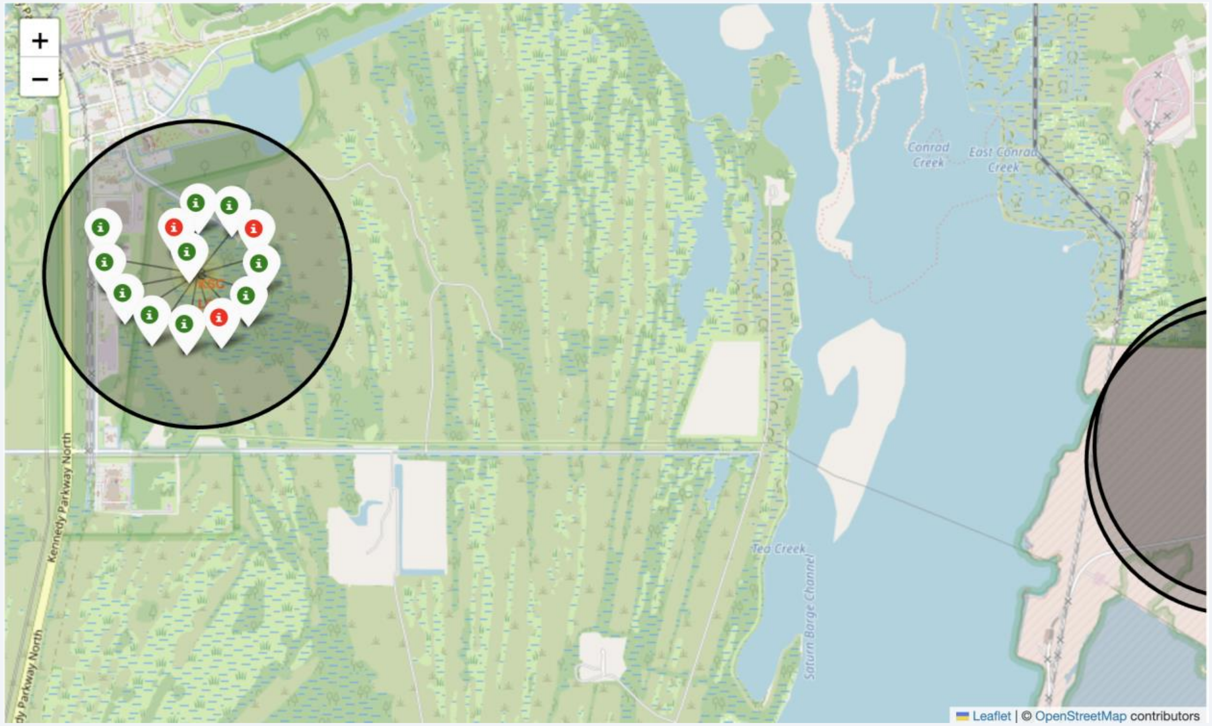
All Launch Sites' Location

- Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth.
- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people.



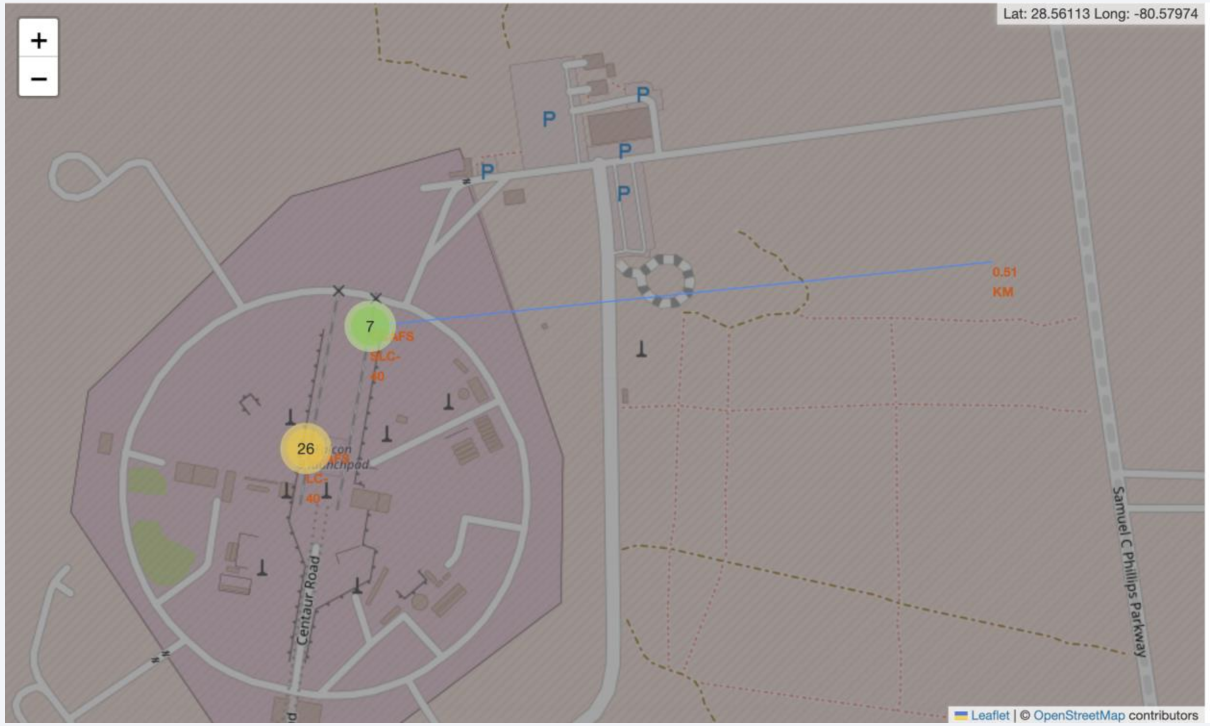
Color-Labeled Launch Outcomes

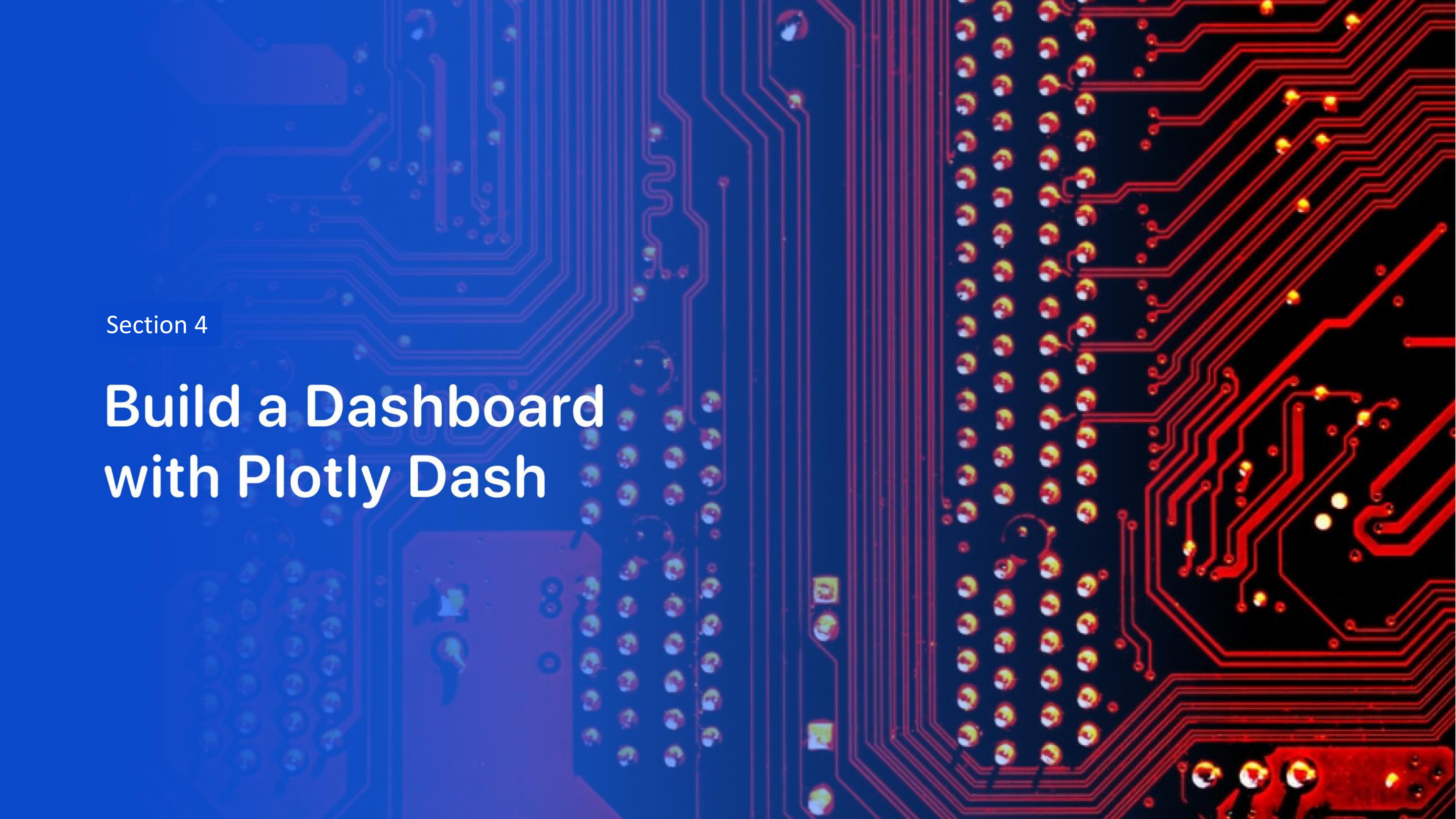
- This screenshot shows the location and the success rates of KSC LC-39A.
- Successful Launches are labeled by green markers, while failed launches are labeled by red markers.
- It shows a high success rate of KSC LC-39A.



Distance from CCAFS SLC-40 to Coastline

- This screenshot shows the distance from CCAFS SLC-40 to Coastline.
- The distance is near 0.51 km. It may be too close.

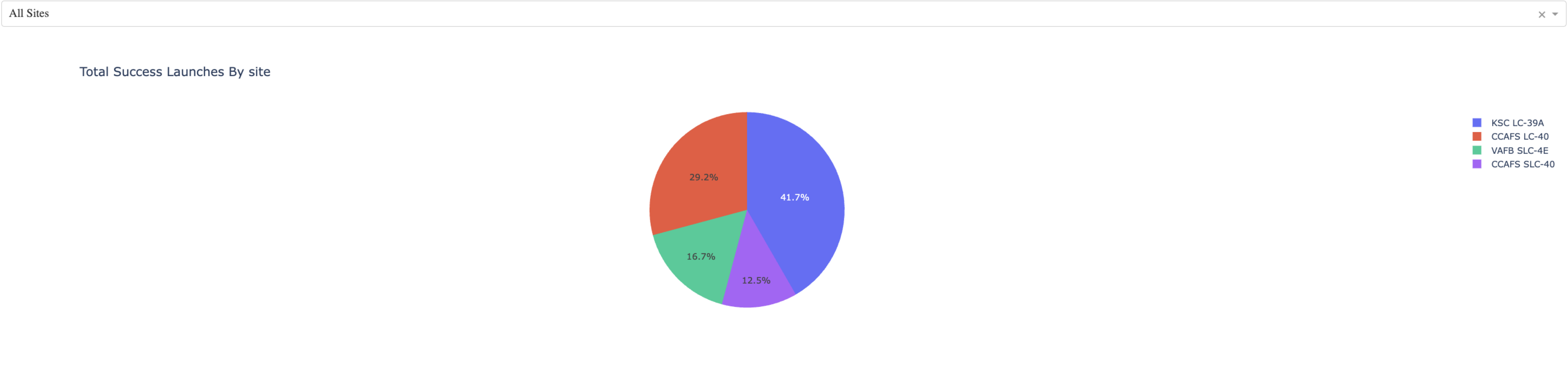

Lat: 28.56113 Long: -80.57974
+
−
P
7
26
0.51 KM
Centaur Road
Samuel C Phillips Parkway
Leaflet | © OpenStreetMap contributors



Section 4

Build a Dashboard with Plotly Dash

Total success Launches By All Sites

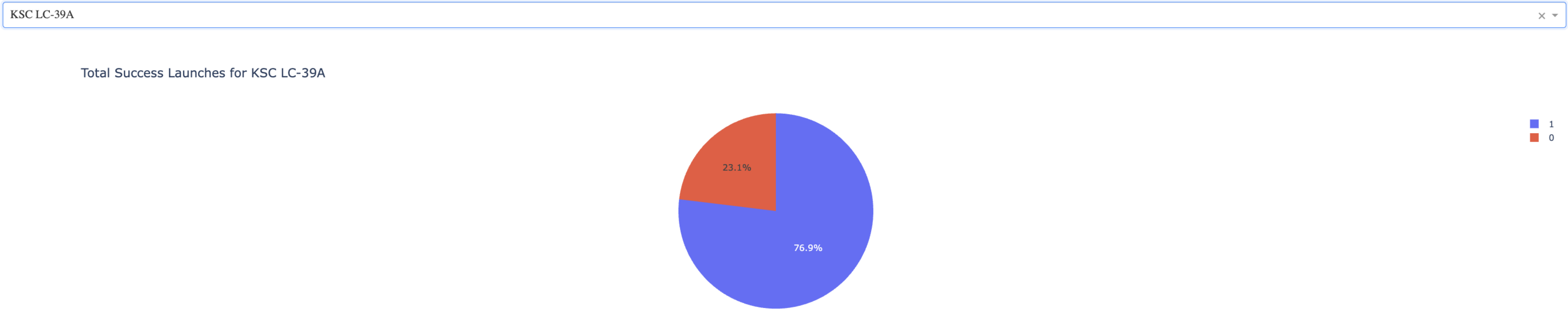


All Sites

Total Success Launches By site

- This pie chart shows the total success launches by all sites.
- It is shown that the KSC LC-39A has the best performance.

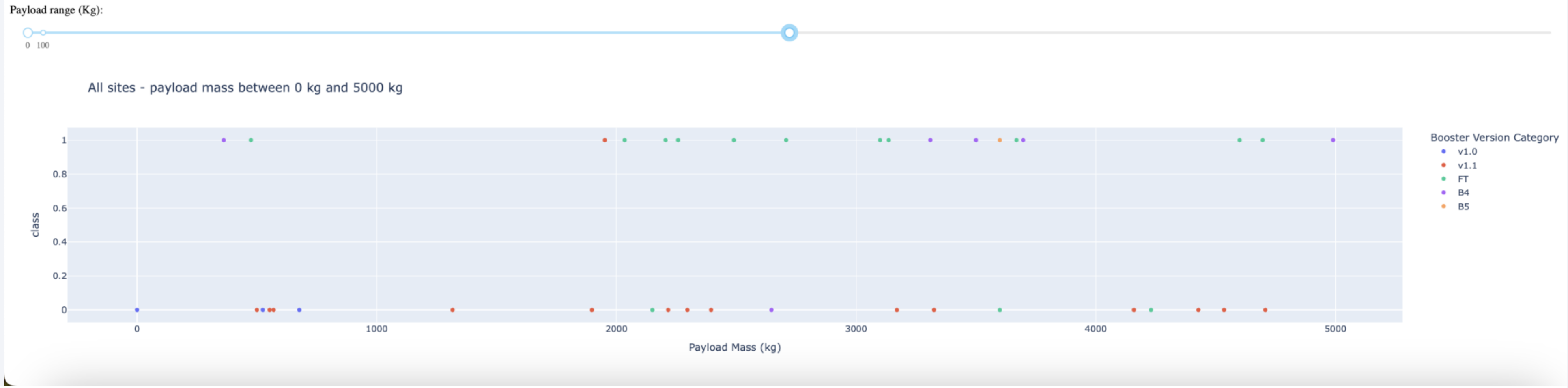
Launch site with highest launch success ratio

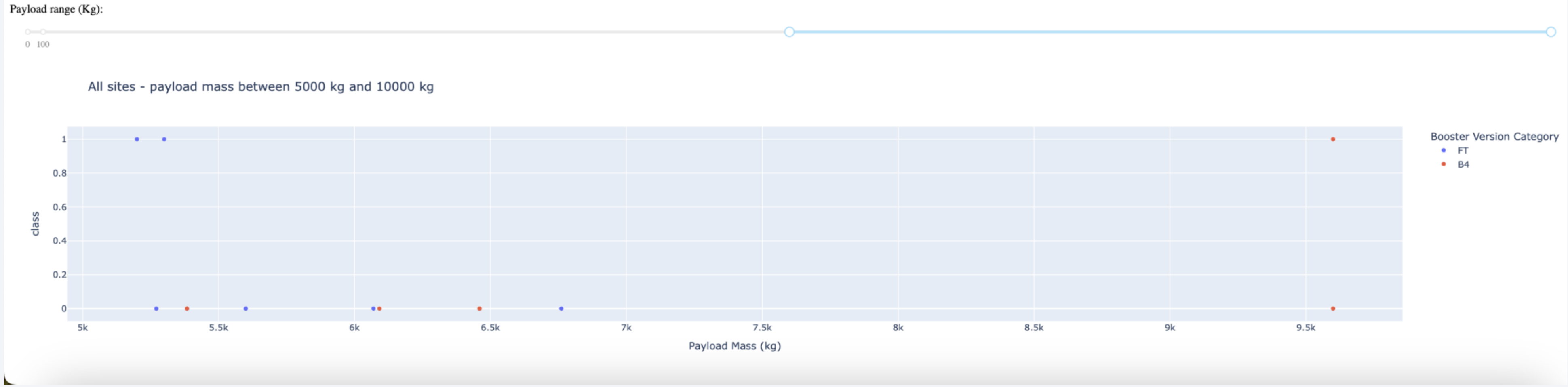


KSC LC-39A

- KSC LC-39A has the highest launch success rate (76.9%).

<Dashboard Screenshot 3>





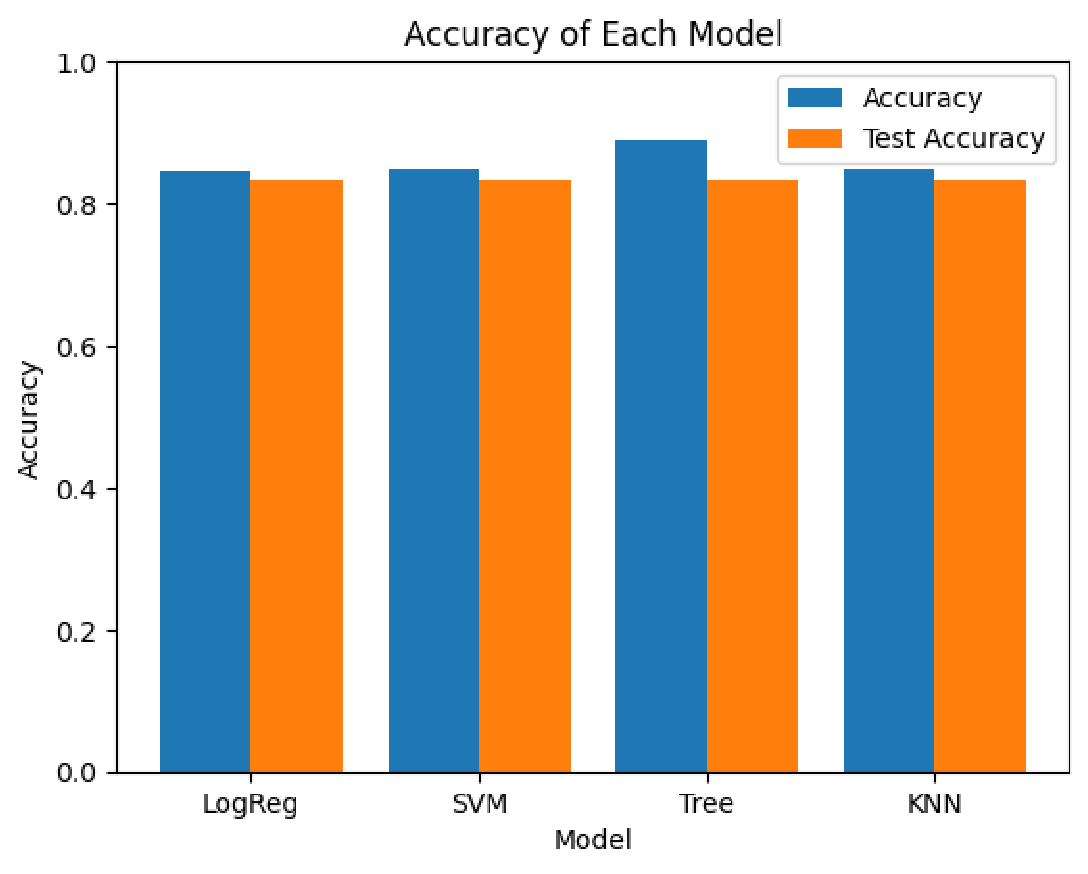
- The Payloads between 2000 and 5000 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

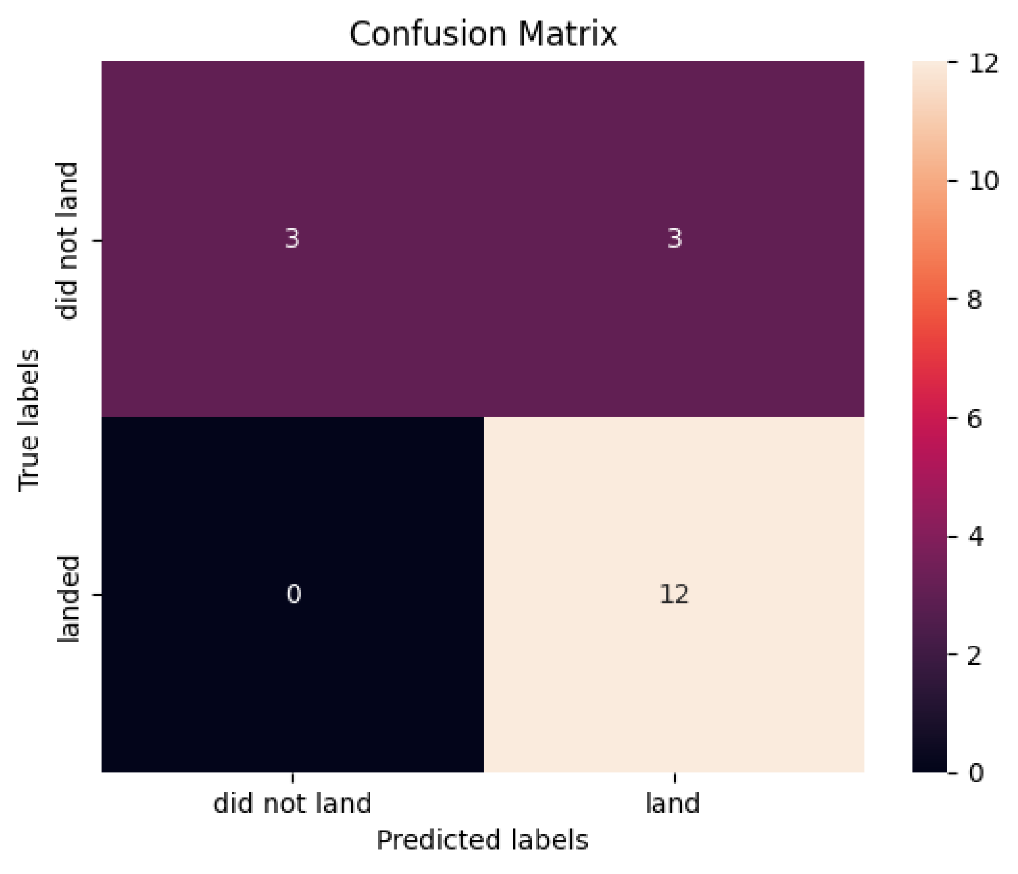
Classification Accuracy

- This bar chart shows the accuracies of four models.
- They have similar results. The accuracy of Decision Tree model is slightly better than other.
- We can say Decision Tree is the best model in this context.



Confusion Matrix

- This is the confusion matrix of the Decision Tree model.
- We can see the true positive rate is high, and other values are low, which leads to a high accuracy.
- We have to point out that the confusion matrix is the same for all models. A possible reason is the small test data sample.



Conclusions

- The SpaceX Launches data have been collected and analyzed in this project.
- We find the best launch site is KSC LC-39A.
- The success rate of launches increases over year.
- The launches to orbits ES-L1, GEO, HEO and SSO have 100% success rates.
- Most of the launch sites are in proximity to the Equator line and all the sites are in very close to the coast.
- The dashboard shows that the Payload Mass between 2000 and 5000 kg has good success rate. However, data over 7000 kg is not sufficient.
- Decision Tree is the best machine learning model in this context.

Appendix

- An error occurs while training the Decision Tree model. Debugging is necessary.
- The test data set is too small, which may affect the performance of the machine learning models.

Thank you!