IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Naser Alshakhoori
04 Oct 2023

Outline

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

Executive Summary

In this capstone project, our aim is to predict the successful landing of the Falcon 9 first stage.

- Summary of methodologies
  - Data Collection using API and Web Scrapping
  - Preprocessing data using Data Cleaning and Data Wrangling
  - Exploratory Data Analysis (EDA) to find insights, patterns and identify training labels.
  - Interactive Visual Analytics and Dashboards
  - Predictive Analysis
- Summary of all results
  - Machine learning models exhibited a high accuracy rate, correctly predicting Falcon 9 first stage landing outcomes.
  - Decision Tree model has outperformed other machine learning model with accuracy score of 94.4%
  - The ability to predict landing outcomes empowers alternative companies seeking to bid against SpaceX for rocket launches

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches with a cost of 62 million dollars whereas other providers cost upward of 165 million dollars each.
  - Reusing the first stage reduces SpaceX's launch costs, offering an advantage over competitors.
  - 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.
- Problems you want to find answers
  - Determining factors of successful landing
  - Predicting landing outcome based on various features
  - Identifying correlations between features
  - Evaluating and determining the most effective prediction models

Section 1

Methodology

Methodology

Executive Summary

- Collect the data
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Preprocess the data in preparation for visualization and analysis
  - Data cleaning: handle missing values, inaccurate data, and data standardization
  - Data wrangling: One hot encoding was applied to categorical features
- 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, tune, evaluate different classification models

Data Collection

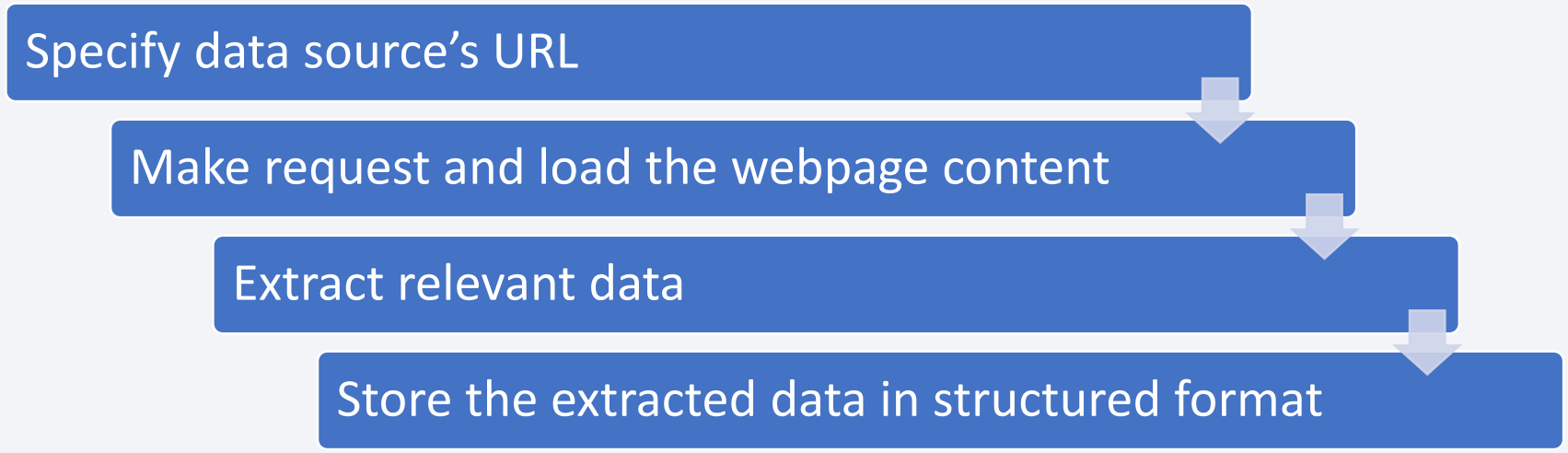
- The data was collected from two sources:
  - SpaceX API: via making HTTP get Requests

URL=https://api.spacexdata.com/v4/launches/past)

  - Wikipedia: via Web Scraping to collect Falcon 9 historical launch records with BeautifulSoup

URL= https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

- You need to present your data collection process use key phrases and flowcharts



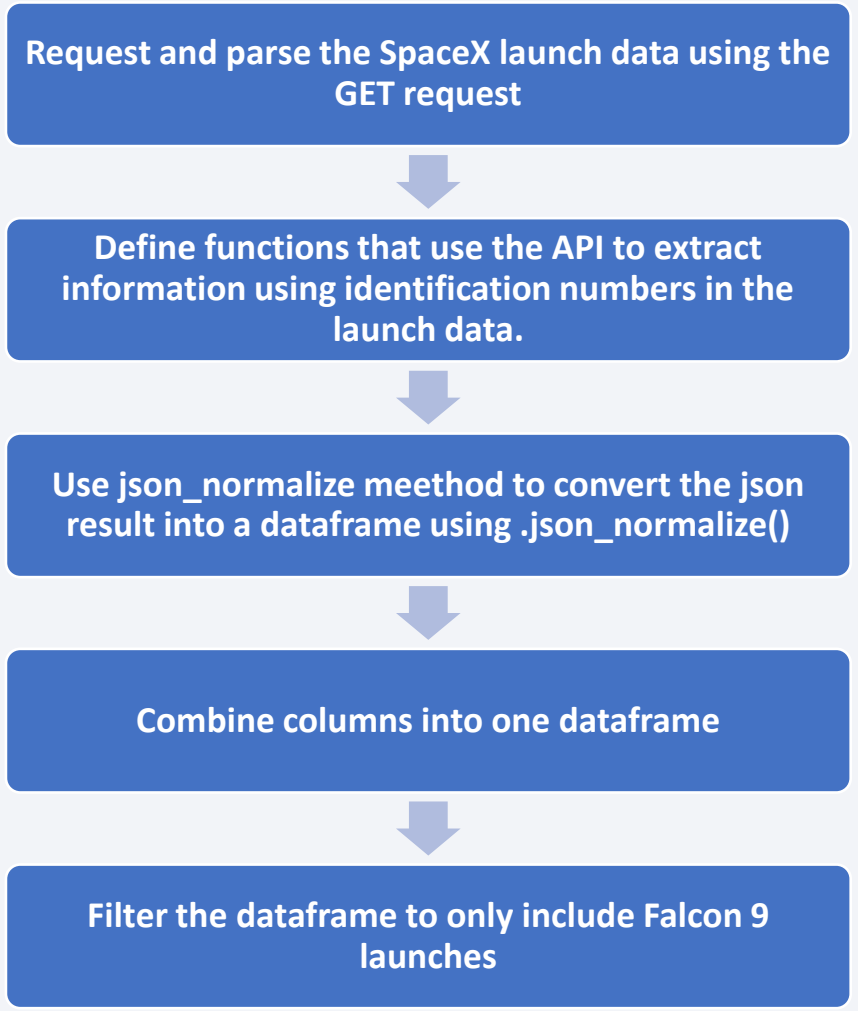
Data Collection – SpaceX API

- SpaceX launch data is accessible to public at https://api.spacexdata.com/v4/launches/past
- The process to extract and store the data is shown in the next flowchart.
- The final dataset includes :

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

- GitHub URL: https://github.com/nalshakhoori/DataScienceCapstone/blob/f08d5cc449fdde3bac21e158b16bce16eab5f555/1_jupyter-labs-spacex-data-collection-api.ipynb

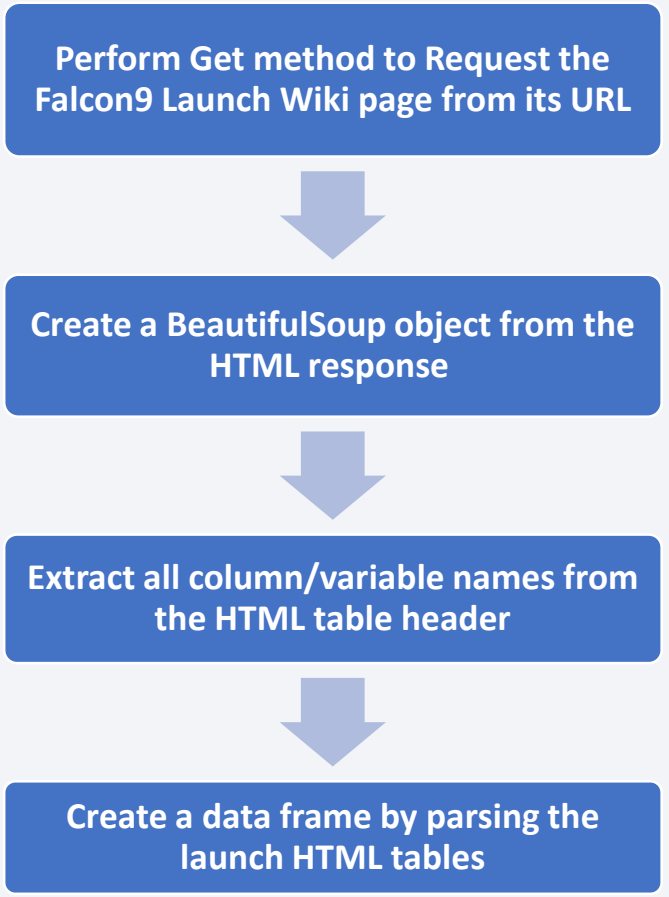
Flowchart of SpaceX API calls



Data Collection - Scraping

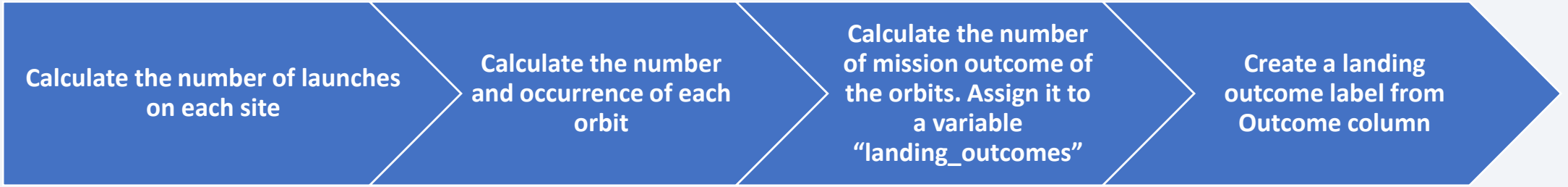
- Falcon 9 historical launch records are stored in HTML table in Wikipedia. The parse the table and convert it into a Pandas dataframe.
- The process to extract and store the data is shown in the next flowchart.
- GitHub URL: https://github.com/nalshakhoori/DataScienceCapstone/blob/f08d5cc449fdde3bac21e158b16bce16eab5f555/2_jupyter-labs-webscraping.ipynb

Flowchart of Web Scraping



Data Wrangling

- Performed Exploratory Data Analysis (EDA) on the final dataset to find some patterns and determine the potential labels for prediction models.
- GitHub URL: https://github.com/nalshakhoori/DataScienceCapstone/blob/f08d5cc449fdde3bac21e158b16bce16eab5f555/3_labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- A Catplot visualizes how FlightNumber and PayloadMass affect Class (launch outcome).
- A Catplot plot visualizes how FlightNumber and LauchSite affect Class.
- A Catplot plot visualizes how LaunchSite and PayloadMass affect Class.
- A bar char visualizes any relationship between Orbit type and the success rate
- A Catplot plot visualizes how FlightNumber and Orbit affect Class.
- A Catplot plot visualizes how PayloadMass and Orbit affect Class.
- A line plot visualizes the launch success yearly trend.
- GitHub URL: https://github.com/nalshakhoori/DataScienceCapstone/blob/f08d5cc449fdde3bac21e158b16bce16eab5f555/5_jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- SQL queries performed include:
  1) Display the names of the unique launch sites in the space mission.
  2) Display 5 records where launch sites begin with the string 'CCA'.
  3) Display the total payload mass carried by boosters launched by NASA (CRS).
  4) Display average payload mass carried by booster version F9 v1.1.
  5) List the date when the first successful landing outcome in ground pad was achieved.
  6) List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  7) List the total number of successful and failure mission outcomes.

EDA with SQL… Continued

- SQL queries performed include:

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

  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.

  10) Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

- GitHub URL: https://github.com/nalshakhoori/DataScienceCapstone/blob/f08d5cc449fdde3bac21e158b16bce16eab5f555/4_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Folium.Circle and Folium.Marker were added in the map for each launch site. This helps to easily see how far the site from the other objects.
- Marker was used to visualize the success/failed launches for each site
- Folium.PolyLine was used to draw lines between a site and selected points in the map (closest coastline, railway, highway and city)
- GitHub URL: https://github.com/nalshakhoori/DataScienceCapstone/blob/f08d5cc449fdde3bac21e158b16bce16eab5f555/6_lab_jupyter_launch_site_location.ipynb

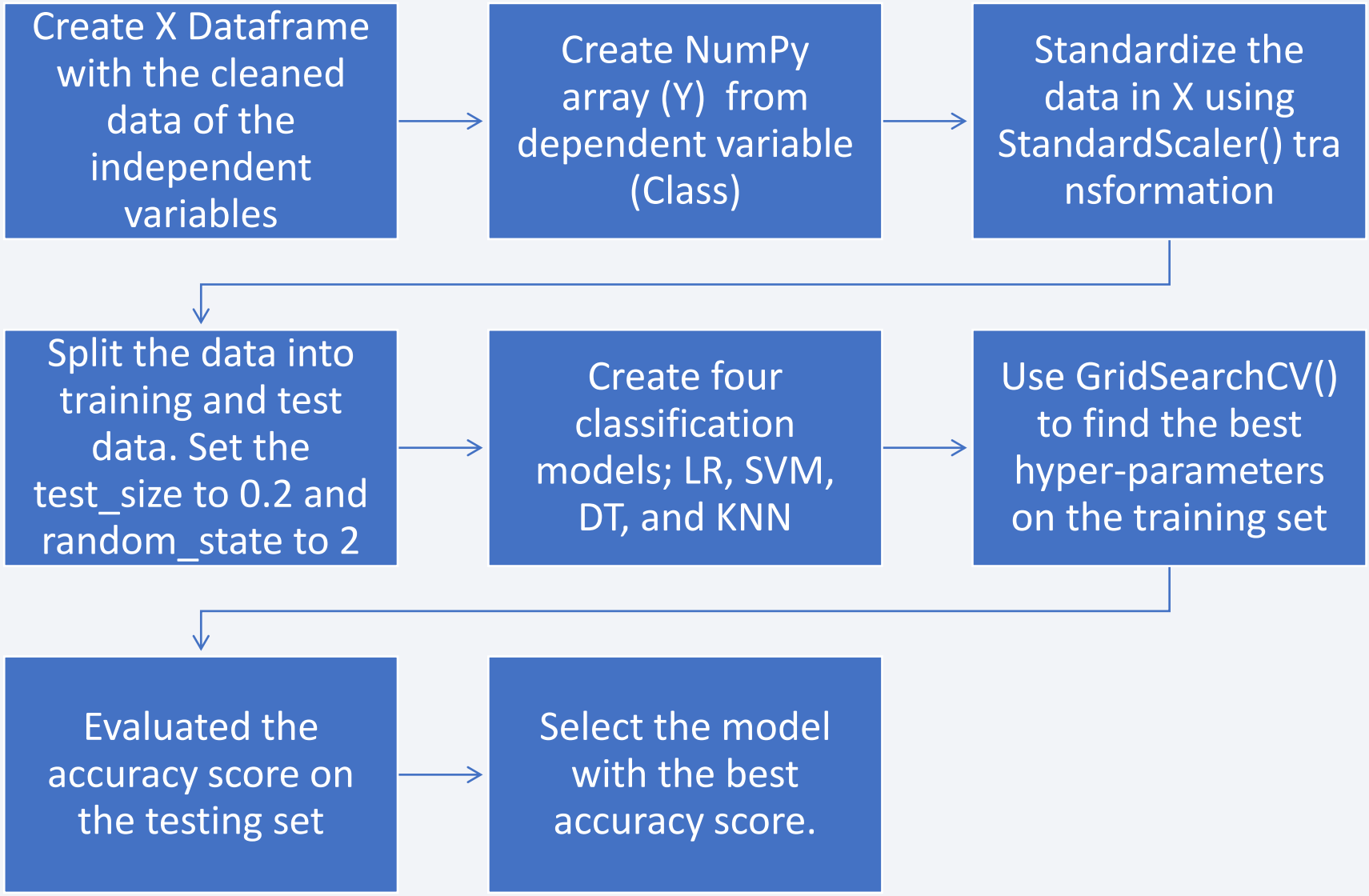
Build a Dashboard with Plotly Dash

- A **pie chart** was used to visualize the success rate by launch site. To make it interactive, the chart can also show the proportion of success and failure of a specific site once selected from the **drop-down** menu.
- Added a **callback function** for `site-dropdown` as **input**, `success-pie-chart` as **output**
- Since the Payload was identified as critical factor for the launch outcome, we added a **scatter** plot to visualize the relationship between PayloadMass and Class. The **color** was set to Booster Version Category to see how outcome differ between Boosters.
- A **slider** was added to allow the user to change the PayloadMass to visualize the impact of different mass on the launch outcome
- Added a **callback function** for `site-dropdown` and `payload-slider` as **inputs**, `success-payload-scatter-chart` as **output**
- GitHub URL: https://github.com/nalshakhoori/DataScienceCapstone/blob/f08d5cc449fdde3bac21e158b16bce16eab5f555/7_spacex_dash_app.ipynb

Predictive Analysis (Classification)

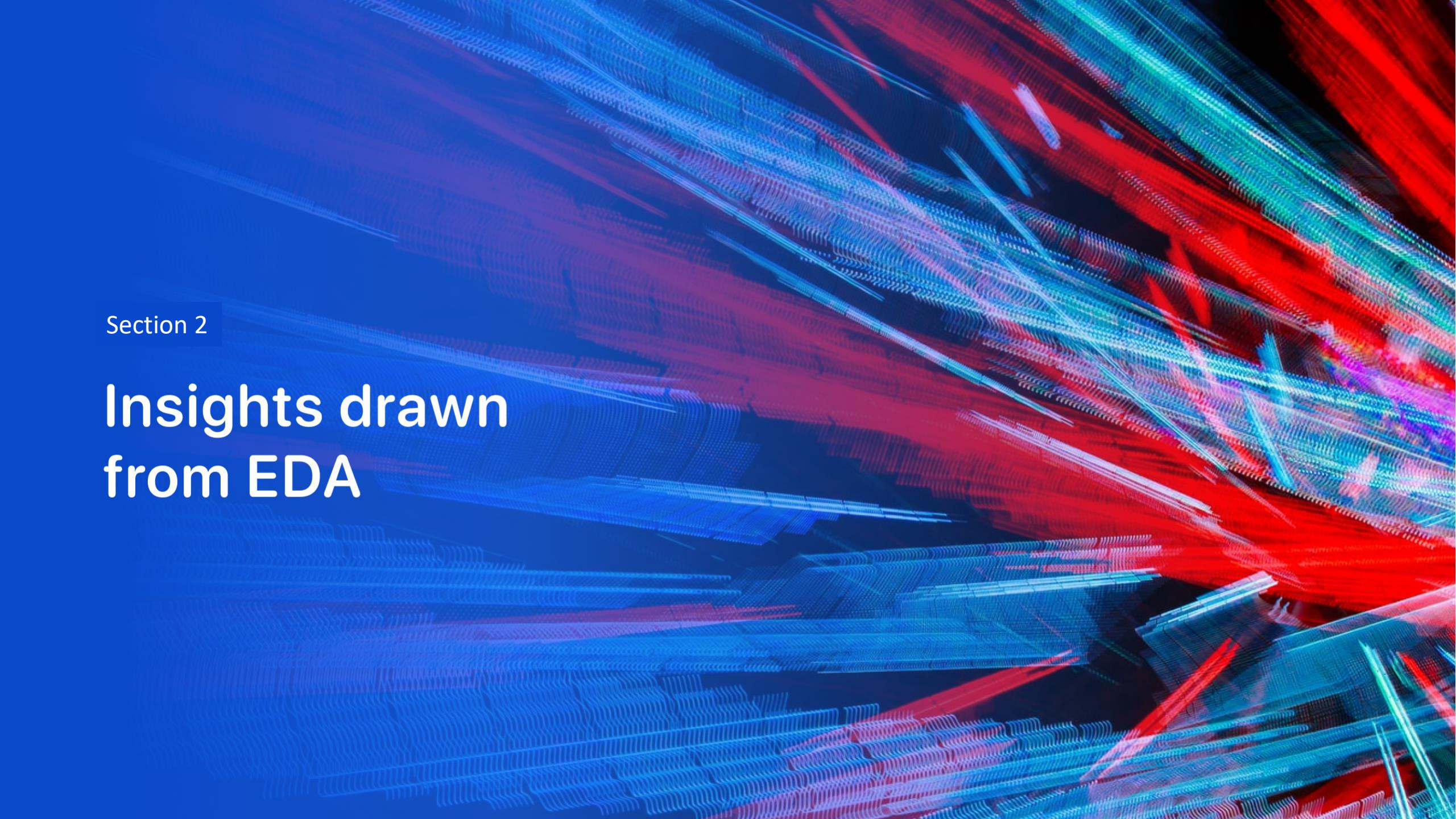
- Four classification models were used to predict the launch outcome (Class):
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearst Neighbors (KNN)
- The next flowchart in the next slide explain the model development and evaluation
- GitHub URL: https://github.com/nalshakhoori/DataScienceCapstone/blob/0a13325c2fd185ff32bed0179c668cad906e84f5/8_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Predictive Analysis (Classification)… Continued



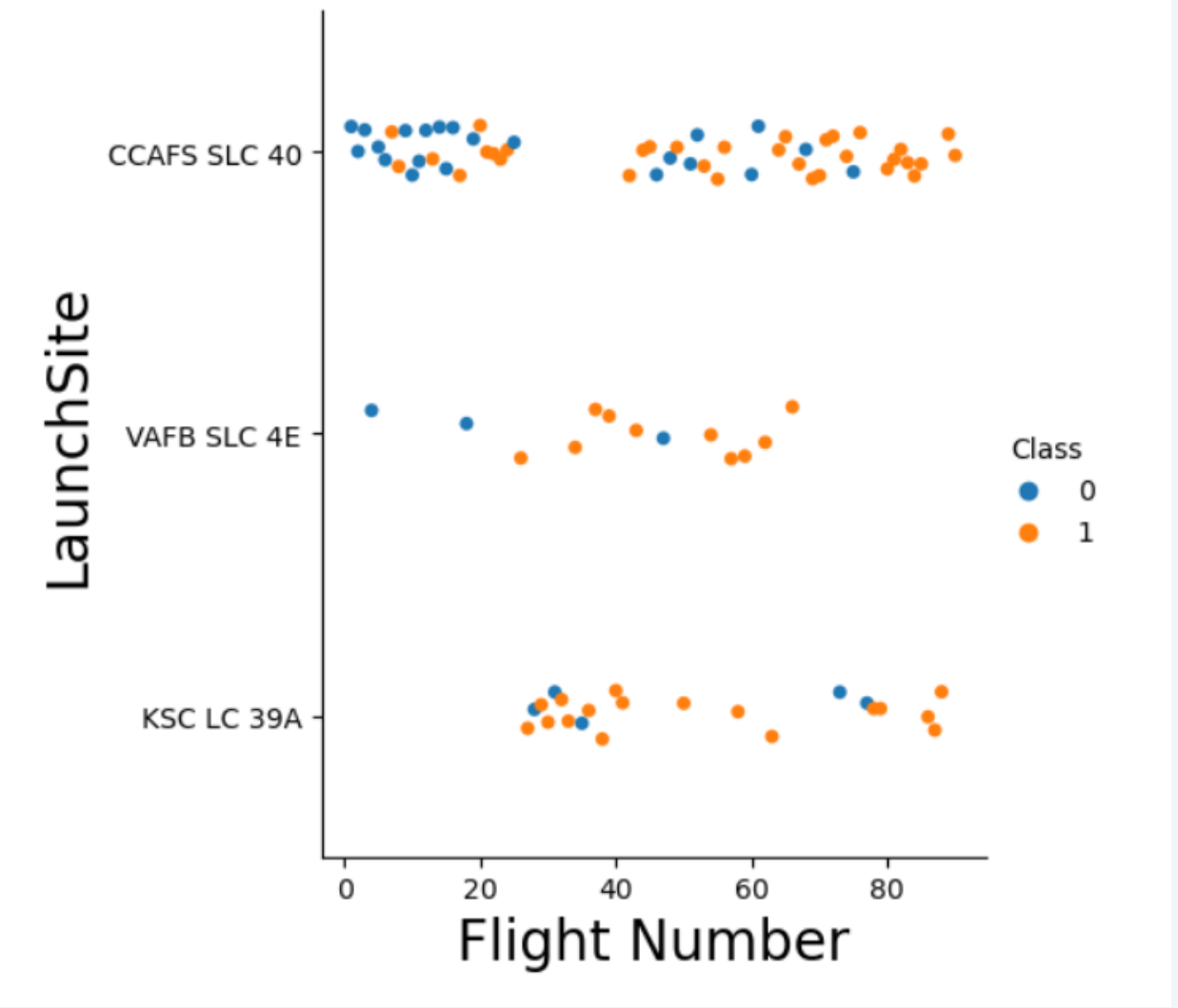
Results

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


Section 2
Insights drawn from EDA

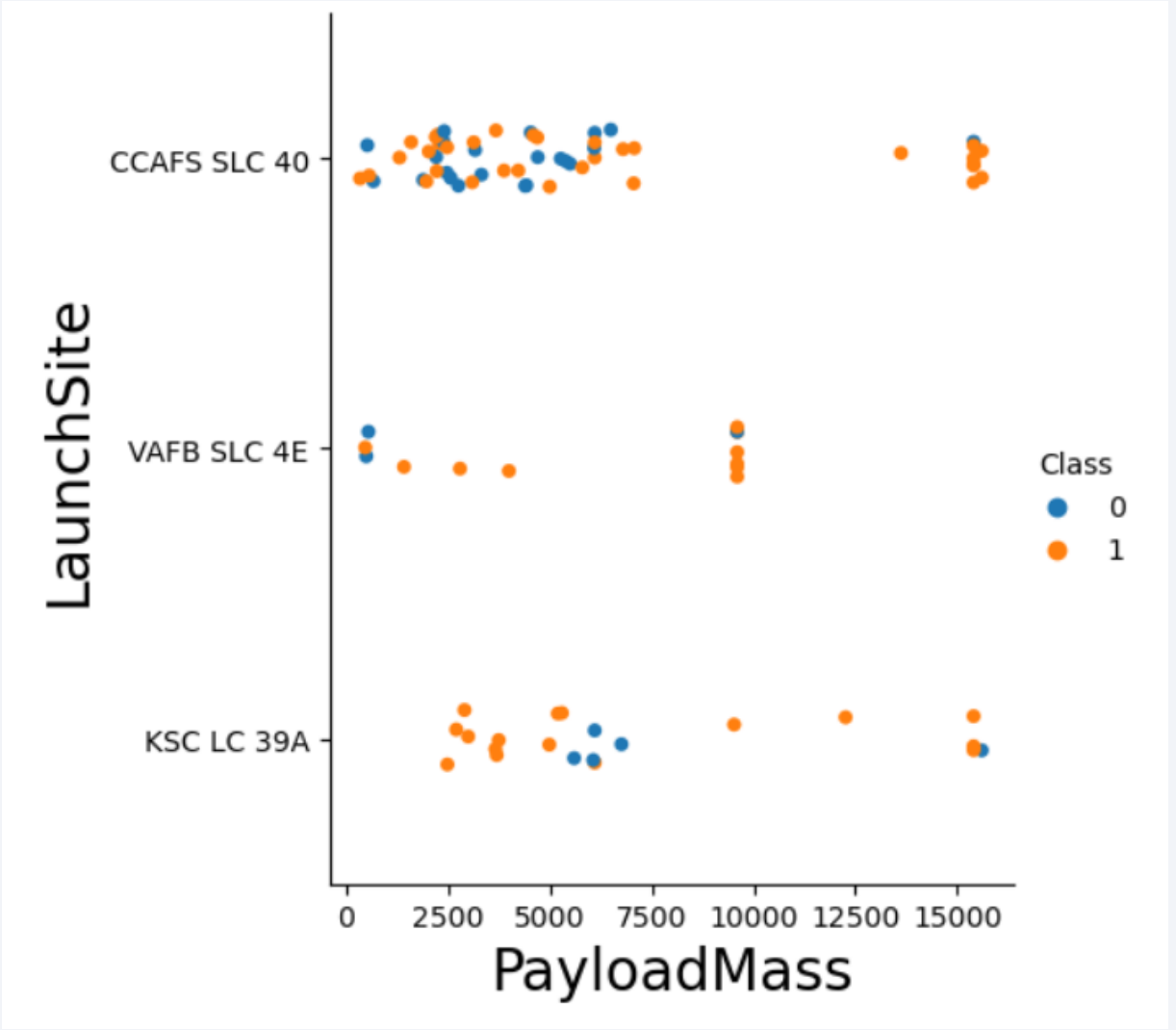
Flight Number vs. Launch Site

- The Successful landing is marked with orange point, and failed landing is marked with blue point.
- We observe that failed landing outcome is associated with small Flight number (e.g. 0 – 20 for CCAFA)
- while there is no clear association with the range of 30-70, CCAFA show mostly successful landing for Flight Number > 70
- VAFB SLC 4E shows mostly successful landing beyond 42 Flight Number



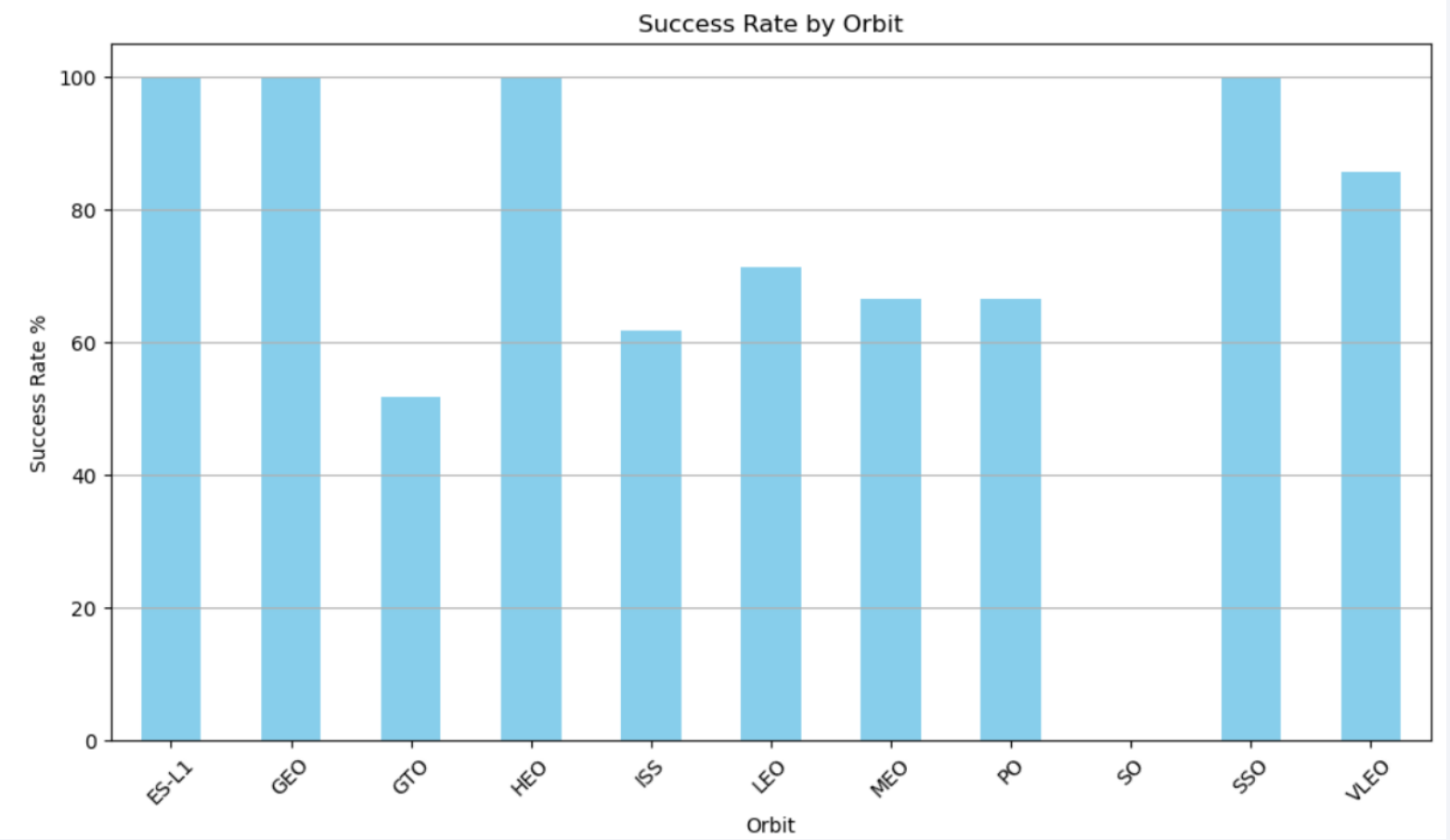
Payload vs. Launch Site

- The Successful landing is marked with orange point, and failed landing is marked with blue point.
- The association is not consistent between the three sites and sometime not consistent for the same site at different loads
- While success in CCAFS SLC 40 is clearly associated with extremely large PayloadMass, such relationship does not hold for loads < 7500 KG
- VAFB SLC 4E show mostly successful landing beyond ~ 1000 KG



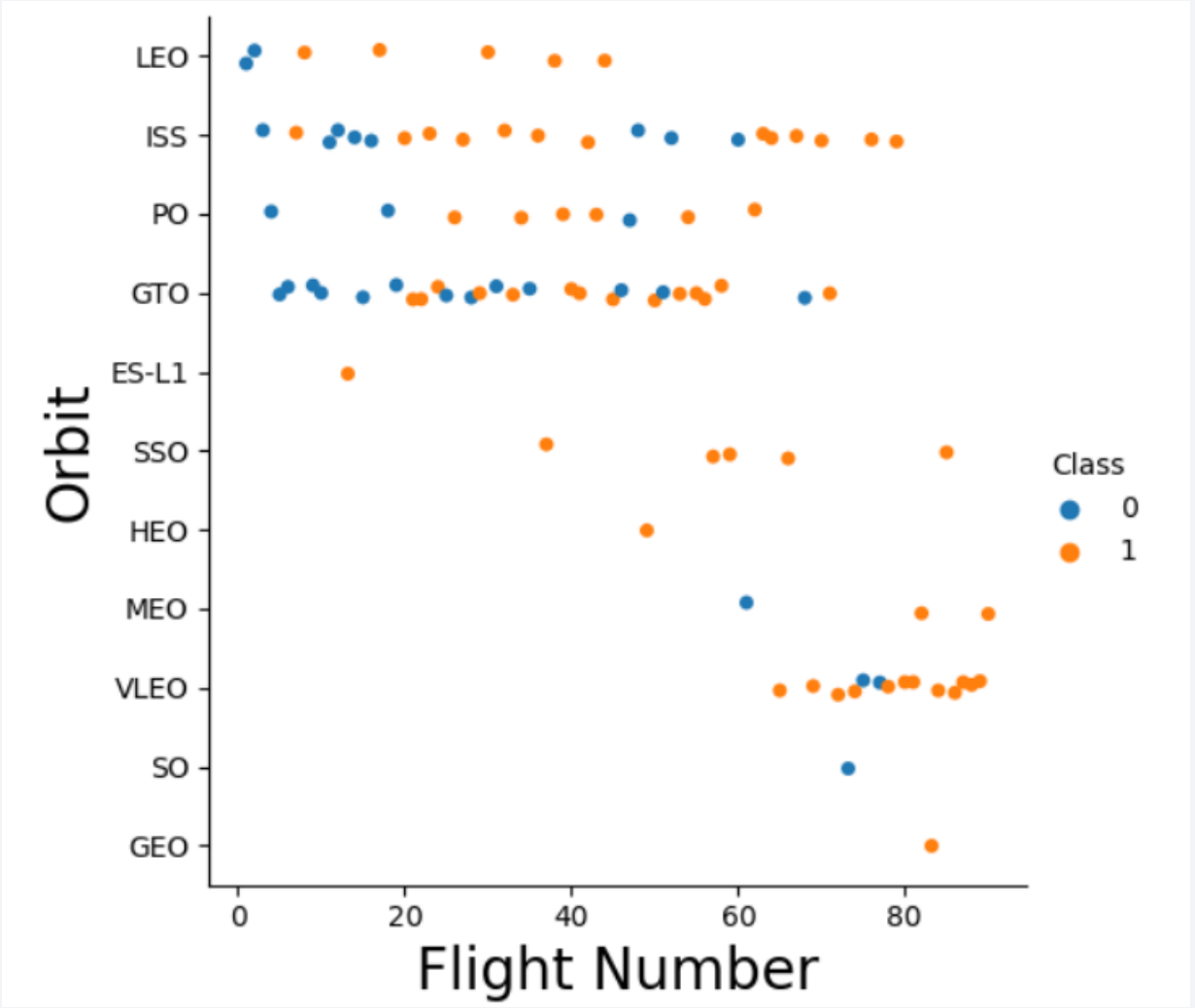
Success Rate vs. Orbit Type

- Four Orbits have no failed first stage landing: ES-L1, GEO, HEO, and SSO
- SO has no successful landing



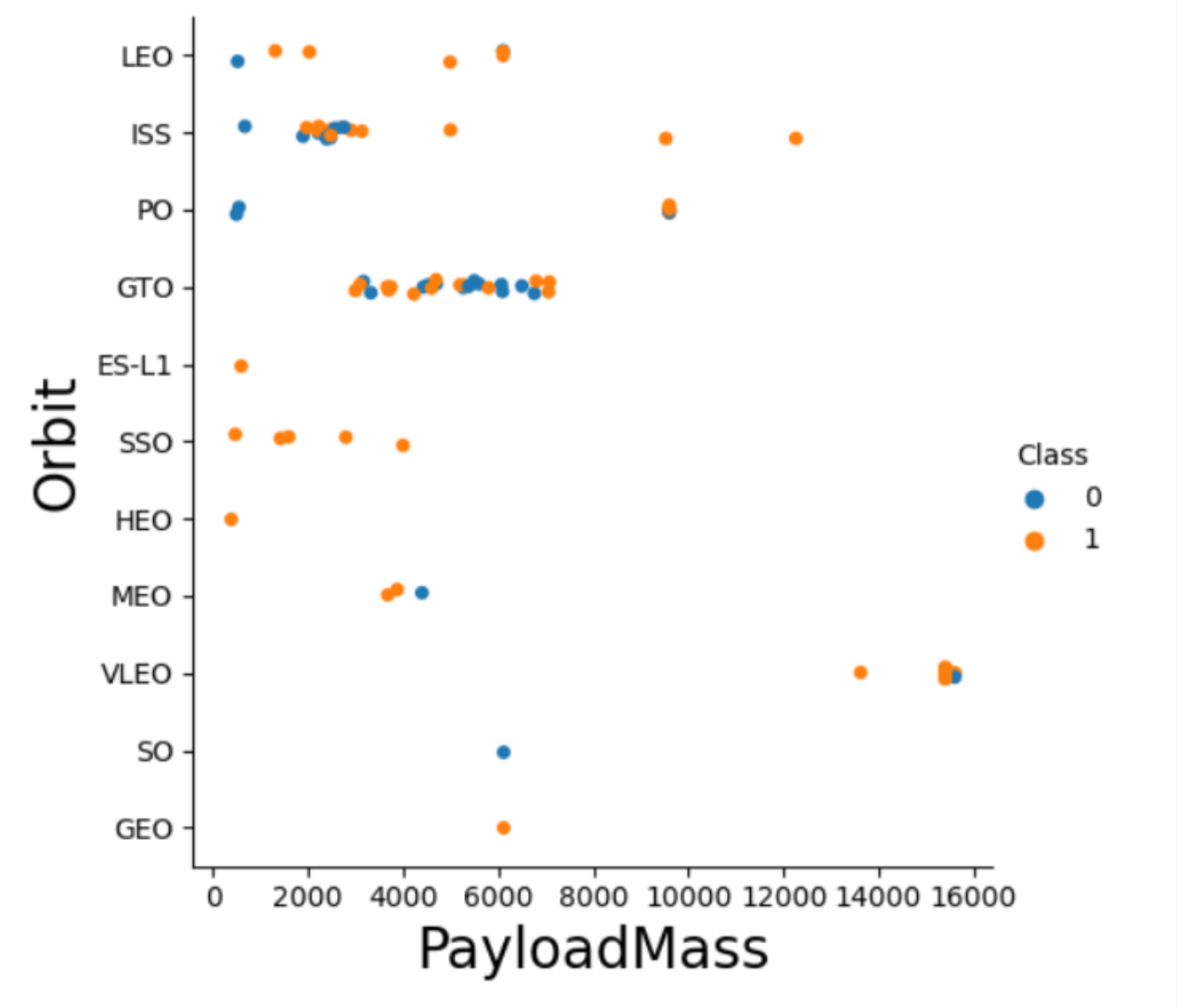
Flight Number vs. Orbit Type

- The Successful landing is marked with orange point, and failed landing is marked with blue point.
- LEO shows successful landing beyond small Flight Number ~ 8.
- GTO has not clear relationship with Flight Number
- All landings were successful for SSO irrespective of the Flight Number
- PO shows mostly successful landing beyond 20 Flight Number

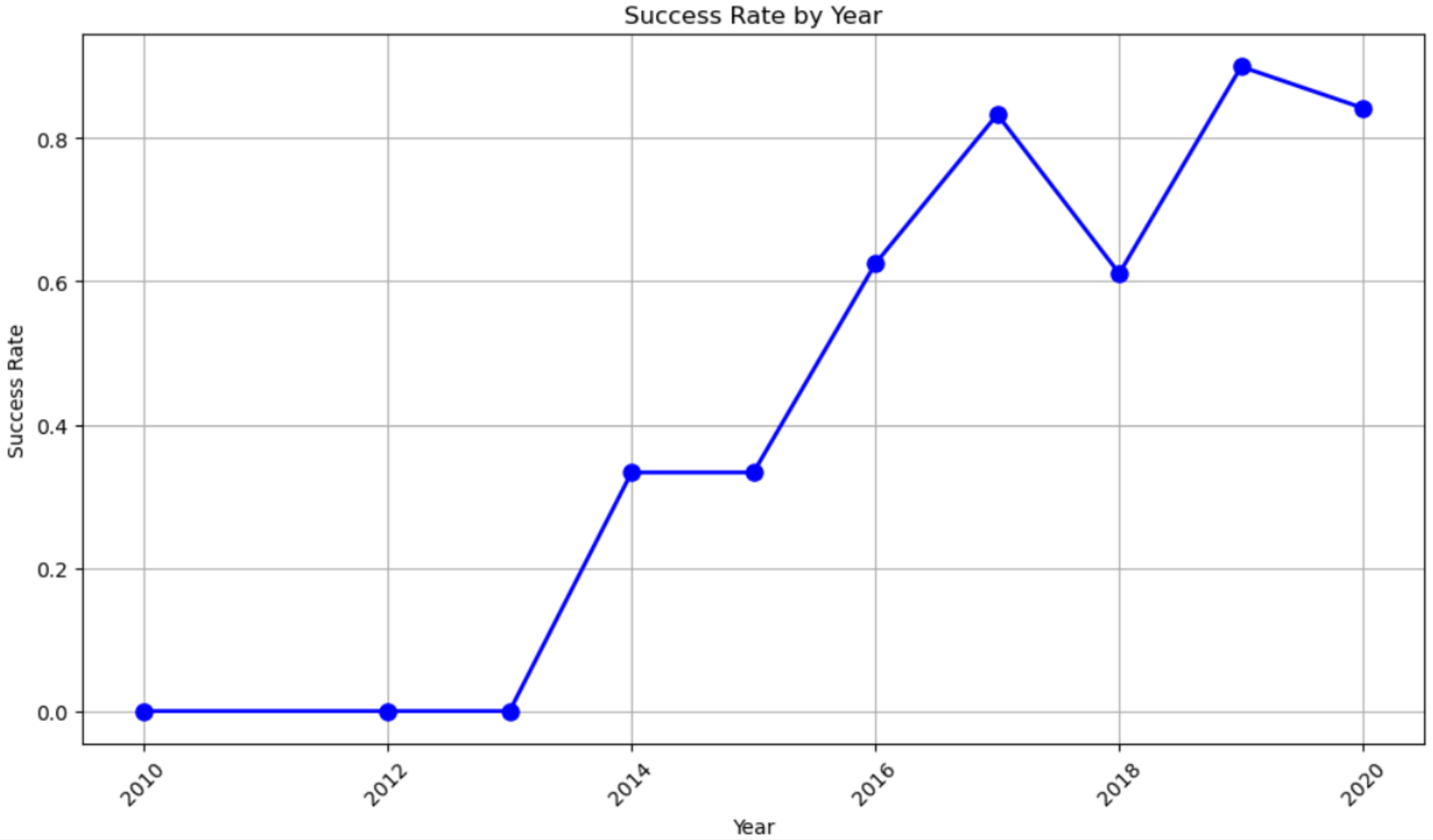


Payload vs. Orbit Type

- The Successful landing is marked with orange point, and failed landing is marked with blue point.
- LEO, ISS, VLEO successful landing is clearly associated with heavier loads which different between the Orbits
- SSO has only successful landings.
- GTO has no clear association as the outcome is mixed at different loads.
- SO has no successful landings



Launch Success Yearly Trend



- We observe that the success rate been increasing since year 2013.
- The highest success rate was in 2019
- Sine 2016, the success rate has surpassed the 60%

All Launch Site Names

Find the names of the unique launch sites

- Query and result:

```
In [10]:  1 %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
```

Out[10]:

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

- Explanation: There are four launch sites

Launch Site Names Begin with 'CCA'

Find 5 records where launch sites begin with `CCA`

- Query and result:

```
In [11]:
1 %sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

Out[11]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Explanation: this is one method of exploring the dataset by searching for specific keyword.

Total Payload Mass

Calculate the total payload carried by boosters from NASA

- Query and result:

```
In [12]:
1 %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
Out[12]:
SUM(PAYLOAD_MASS__KG_)
45596
```

- Explanation: The total payload carried by boosters from NASA is 45,596 KG.

Average Payload Mass by F9 v1.1

Calculate the average payload mass carried by booster version F9 v1.1

- Query and result:

```
In [13]:
1 %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE "Booster_Version" LIKE 'F9 v1.1%';
 * sqlite:///my_data1.db
Done.
Out[13]:
AVG(PAYLOAD_MASS__KG_)
2534.6666666666665
```

- Explanation: the average payload mass carried by booster version F9 V1.1 is 2,534.67 KG

First Successful Ground Landing Date

Find the dates of the first successful landing outcome on ground pad

- Query and result

```
In [14]:
1 %sql SELECT MIN(Date) FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
Out[14]:
```

MIN(Date)
2015-12-22

- Explanation: the first successful landing outcome on ground pad was in 2015-12-22

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

- Query and result:

In [15]:

```
%%sql
SELECT "Booster_Version" FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Success (drone ship)' AND
("PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000);
```

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

Out[15]:

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

- Explanation: There are four boosters as shown above that 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

Calculate the total number of successful and failure mission outcomes

- Query and result:

In [17]:

```
%%sql
SELECT
    CASE
        WHEN "Mission_Outcome" LIKE '%Success%' THEN 'Success'
        ELSE 'Failure'
    END AS "Mission Outcome",
    COUNT(*) AS Total
FROM SPACEXTABLE
GROUP BY "Mission Outcome";
```

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

Out[17]:

Mission Outcome	Total
Failure	1
Success	100

- Explanation: There were 100 successful outcomes and 1 failed mission outcome

Boosters Carried Maximum Payload

List the names of the booster which have carried the maximum payload mass

- Query and result:

In [18]:

```
%%sql
SELECT DISTINCT "Booster_Version"
FROM SPACEXTABLE
WHERE "PAYLOAD_MASS__KG_" = (
    SELECT MAX("PAYLOAD_MASS__KG_")
    FROM SPACEXTABLE
);
```

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

Out[18]:

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

- Explanation: There were 18 boosters who have carried the maximum payload mass.

2015 Launch Records

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- Query and result:

In [19]:

```
%%sql
SELECT
    CASE SUBSTR("Date", 6, 2)
        WHEN '01' THEN 'January'
        WHEN '02' THEN 'February'
        WHEN '03' THEN 'March'
        WHEN '04' THEN 'April'
        WHEN '05' THEN 'May'
        WHEN '06' THEN 'June'
        WHEN '07' THEN 'July'
        WHEN '08' THEN 'August'
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
        ELSE 'Invalid Month'
    END AS Month,
    "Landing_Outcome" AS Landing_Outcome,
    "Booster_Version" AS Booster_Version,
    "Launch_Site" AS Launch_Site
FROM SPACEXTABLE
WHERE SUBSTR("Date", 1, 4) = '2015'
AND "Landing_Outcome" = 'Failure (drone ship)';
```

Out[19]:

Month	Landing_Outcome	Booster_Version	Launch_Site
October	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Explanation: There were two failed landing outcomes

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

- Query and result:

In [20]:

```
%%sql
SELECT "Landing_Outcome", COUNT(*) AS Outcome_Count
FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
AND "Landing_Outcome" IN ('Failure (drone ship)', 'Success (ground pad)')
GROUP BY "Landing_Outcome"
ORDER BY Outcome_Count DESC;
```

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

Out[20]:

Landing_Outcome	Outcome_Count
Success (ground pad)	5
Failure (drone ship)	5

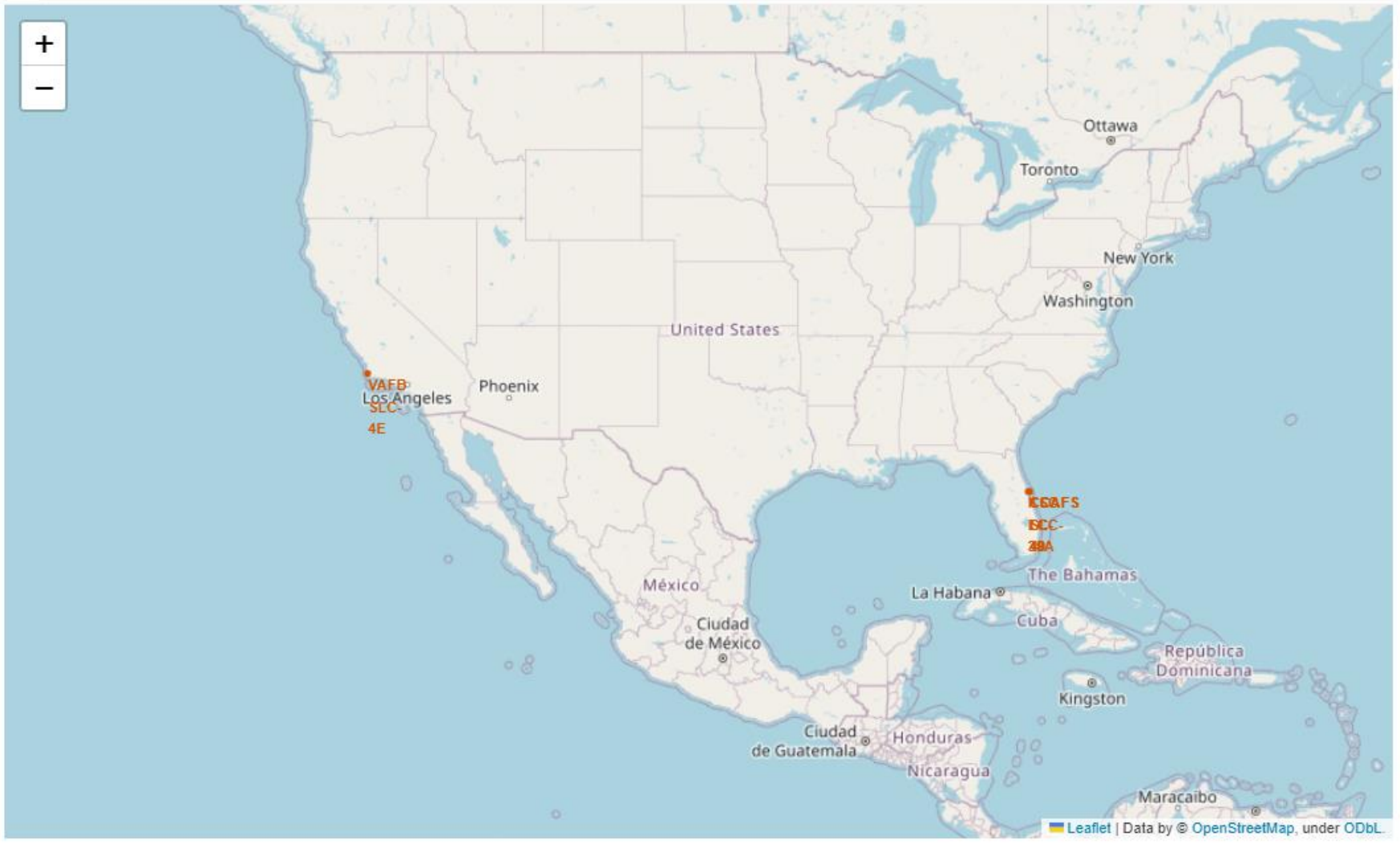
- Explanation: there were five success and failure outcome during the mentioned period

Section 3

Launch Sites Proximities Analysis

Falcon 9 All Launch Sites

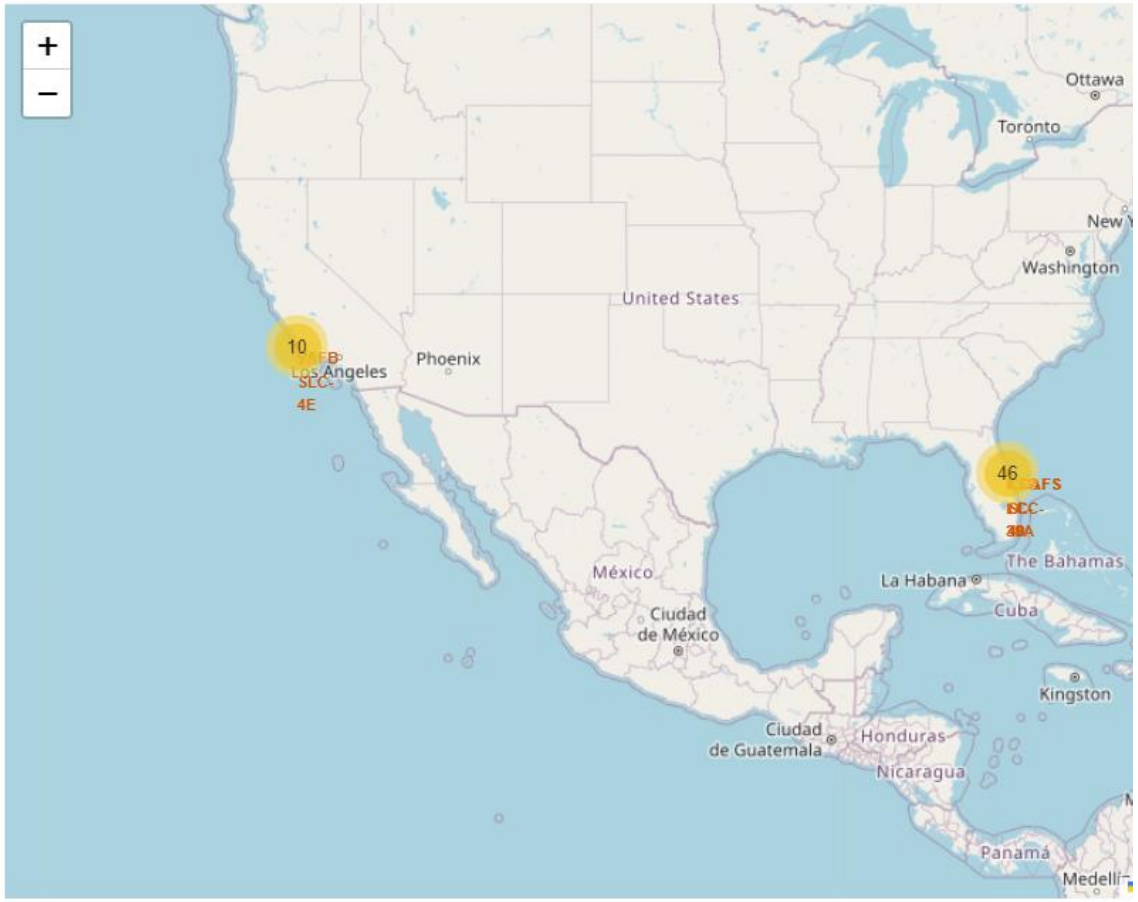
Out[8]:



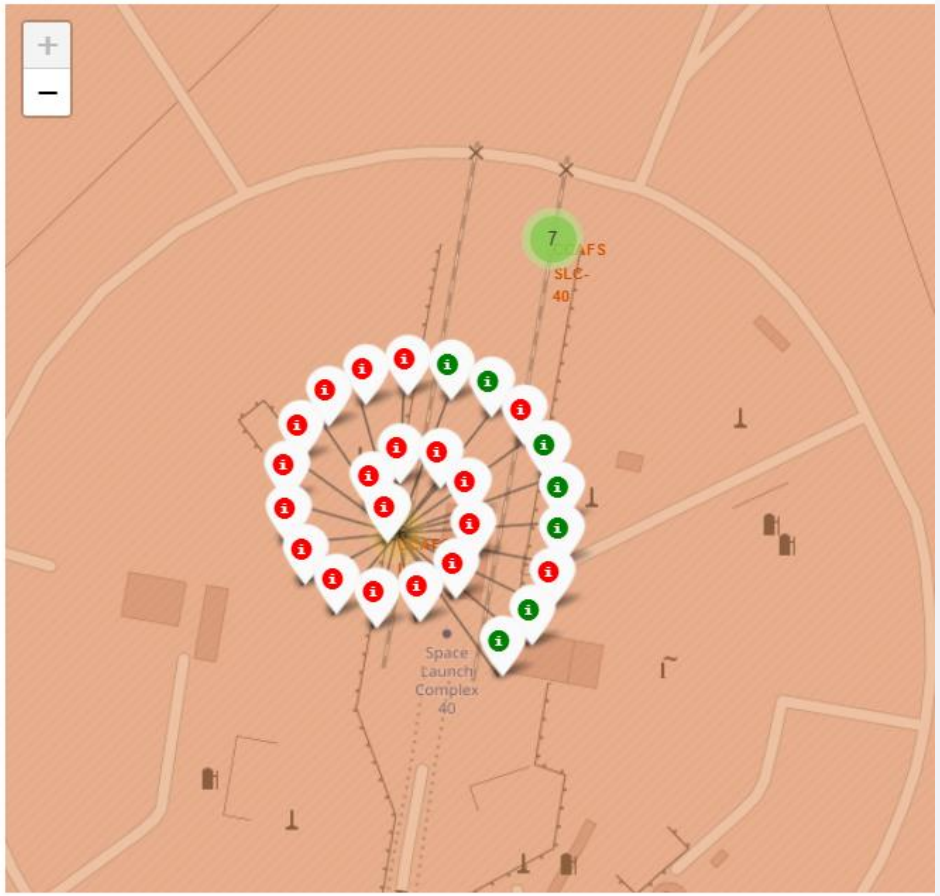
We observe that the four launch sites are located near the coastline

Success/Failed Launches For Each Site

Out[13]:



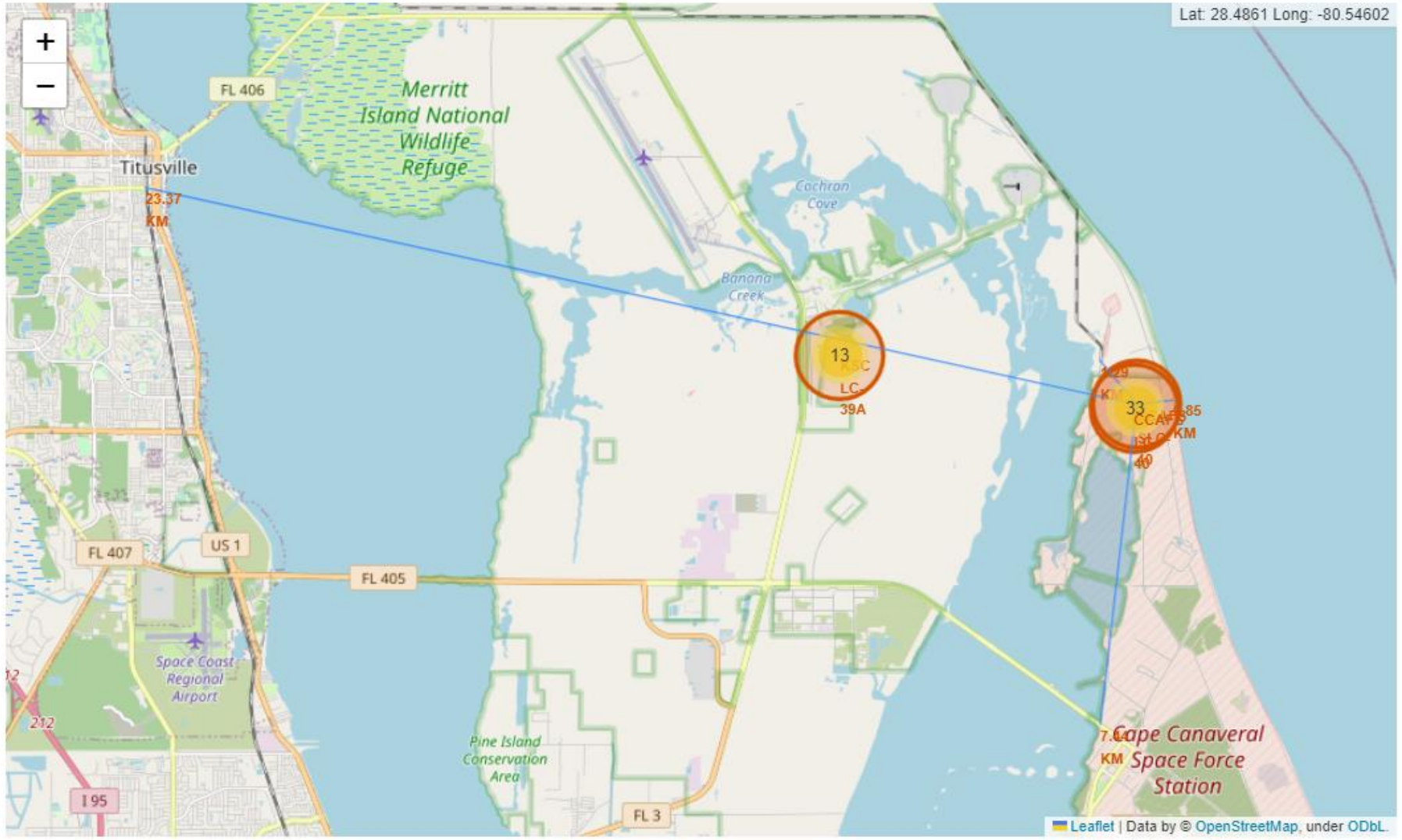
Out[13]:



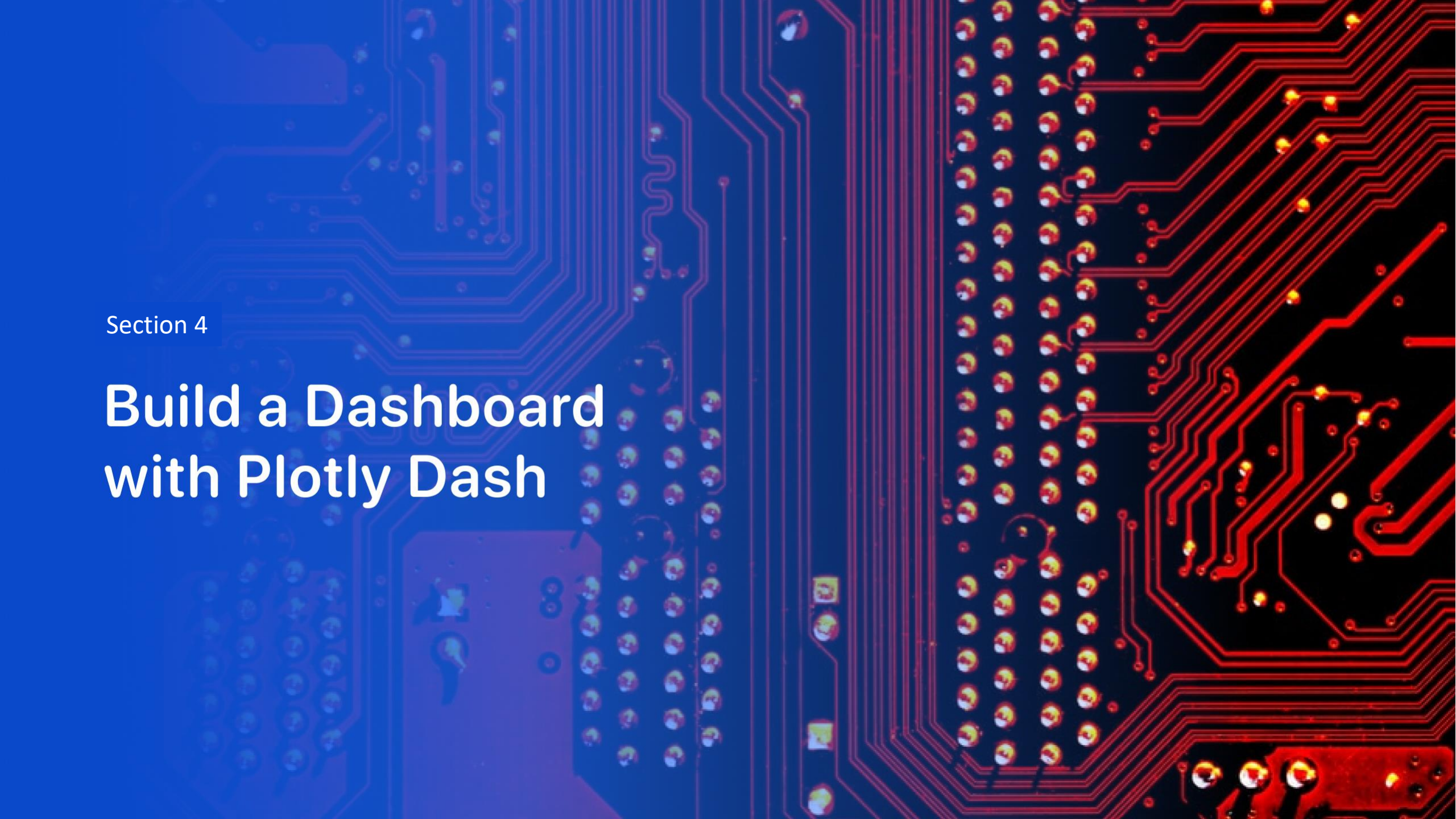
The first map shows the clusters of each site launches. The zoomed in map shows the successful (green marker) and failed (red marker) launches.

Launch Sites And its Proximities

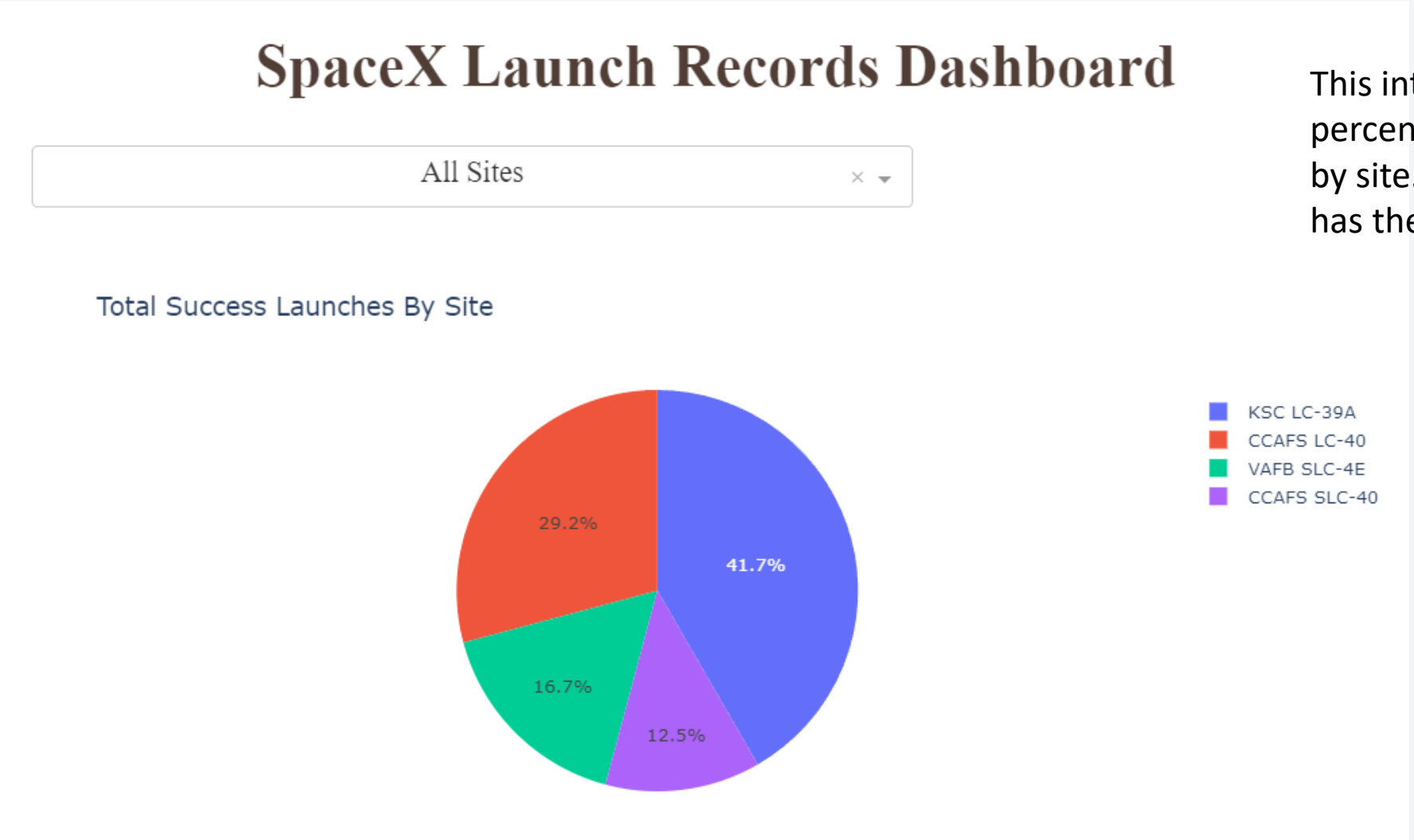
Out[19]:



We notice that the launch site is away from cities, but closest coastline, then railways and highways.

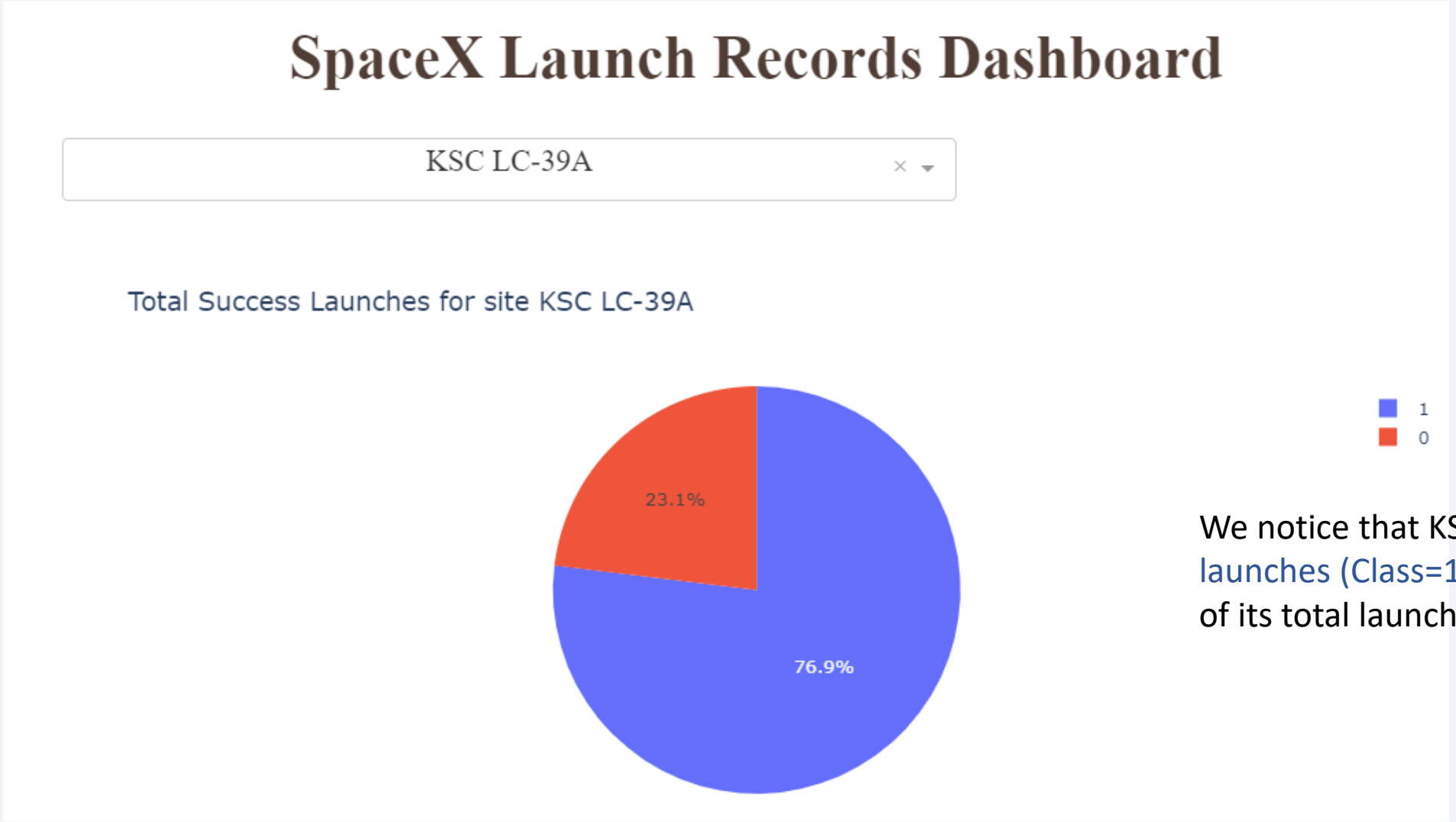

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



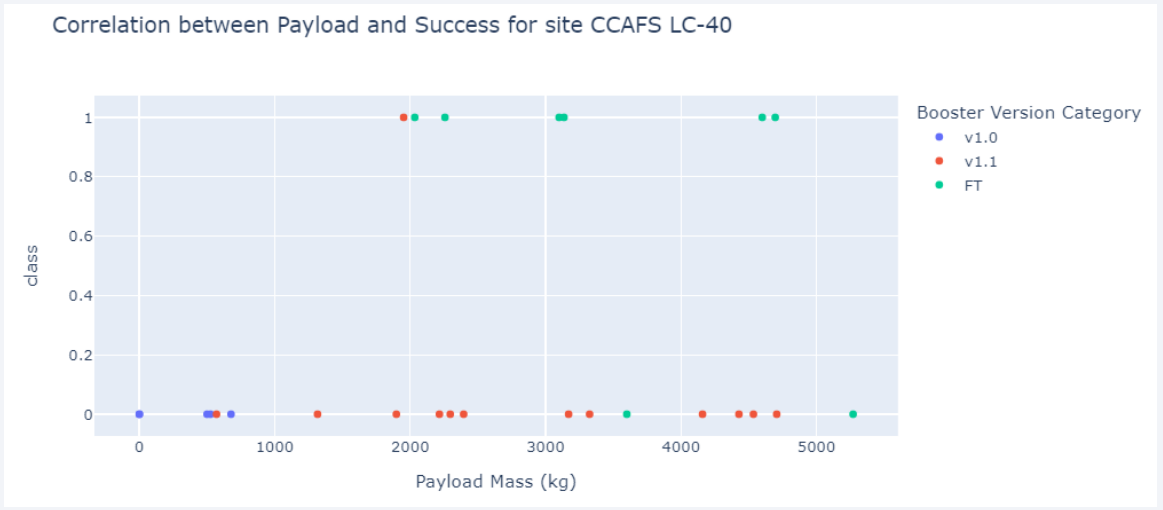
This interactive chart shows the percentage of successful launches by site. We notice that KSC LC-39A has the highest success record.

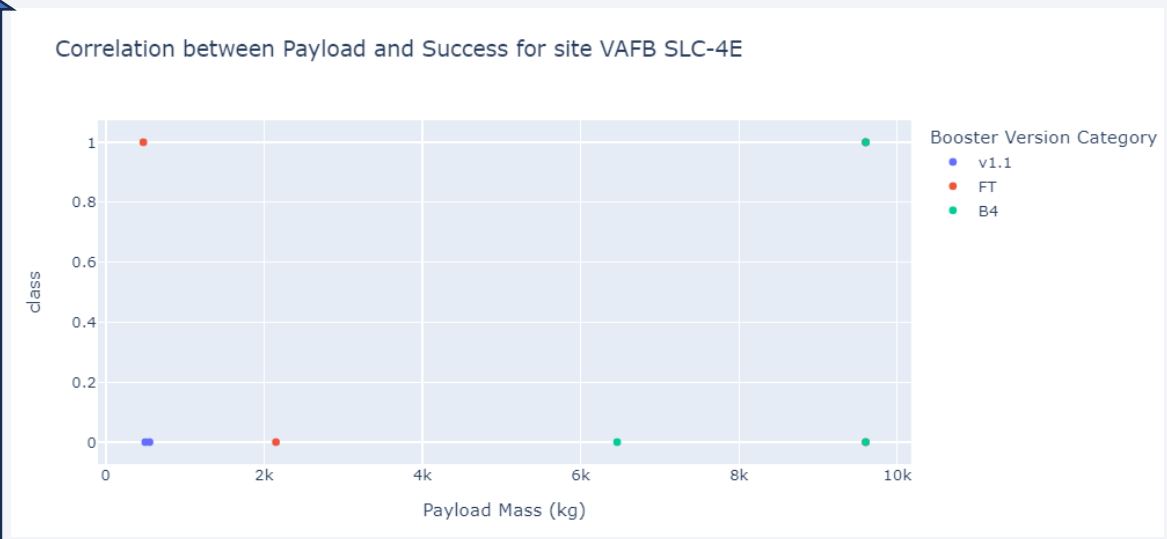
Launch Site With The Highest Success Count

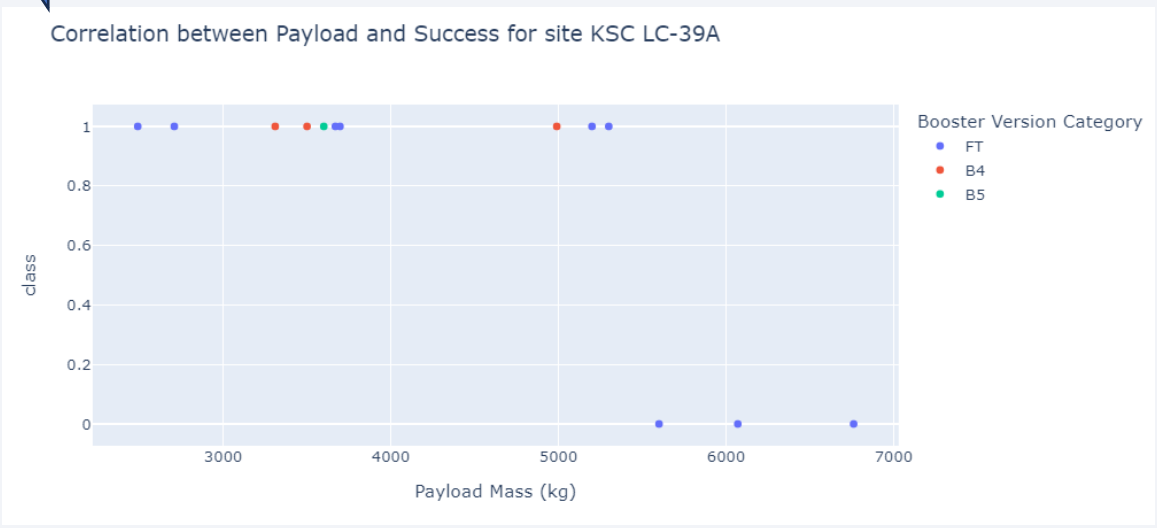


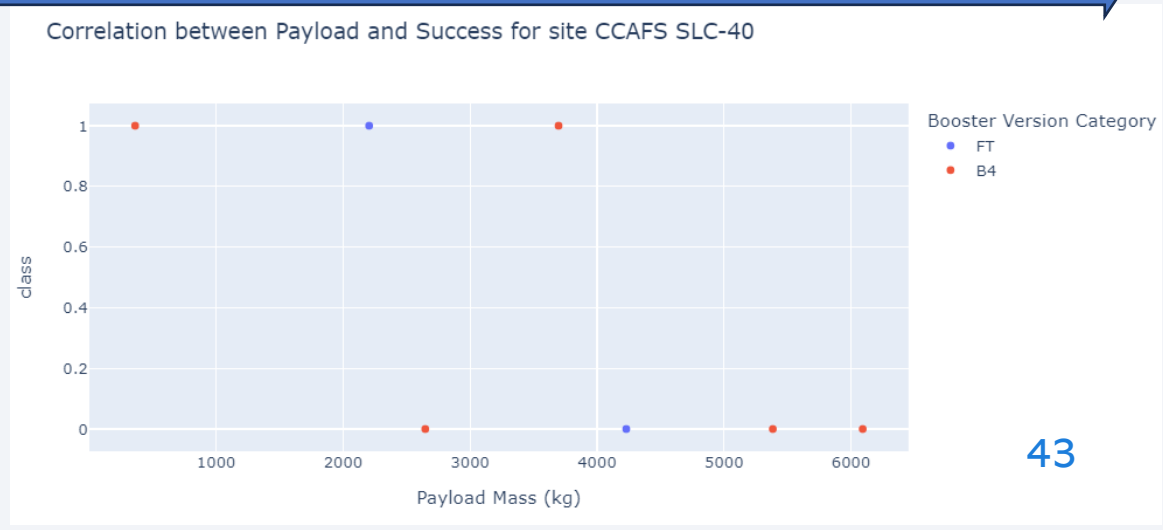
We notice that KSC LC-39A successful launches (Class=1) account for 76.9% of its total launches.

Payload vs. Launch Outcome









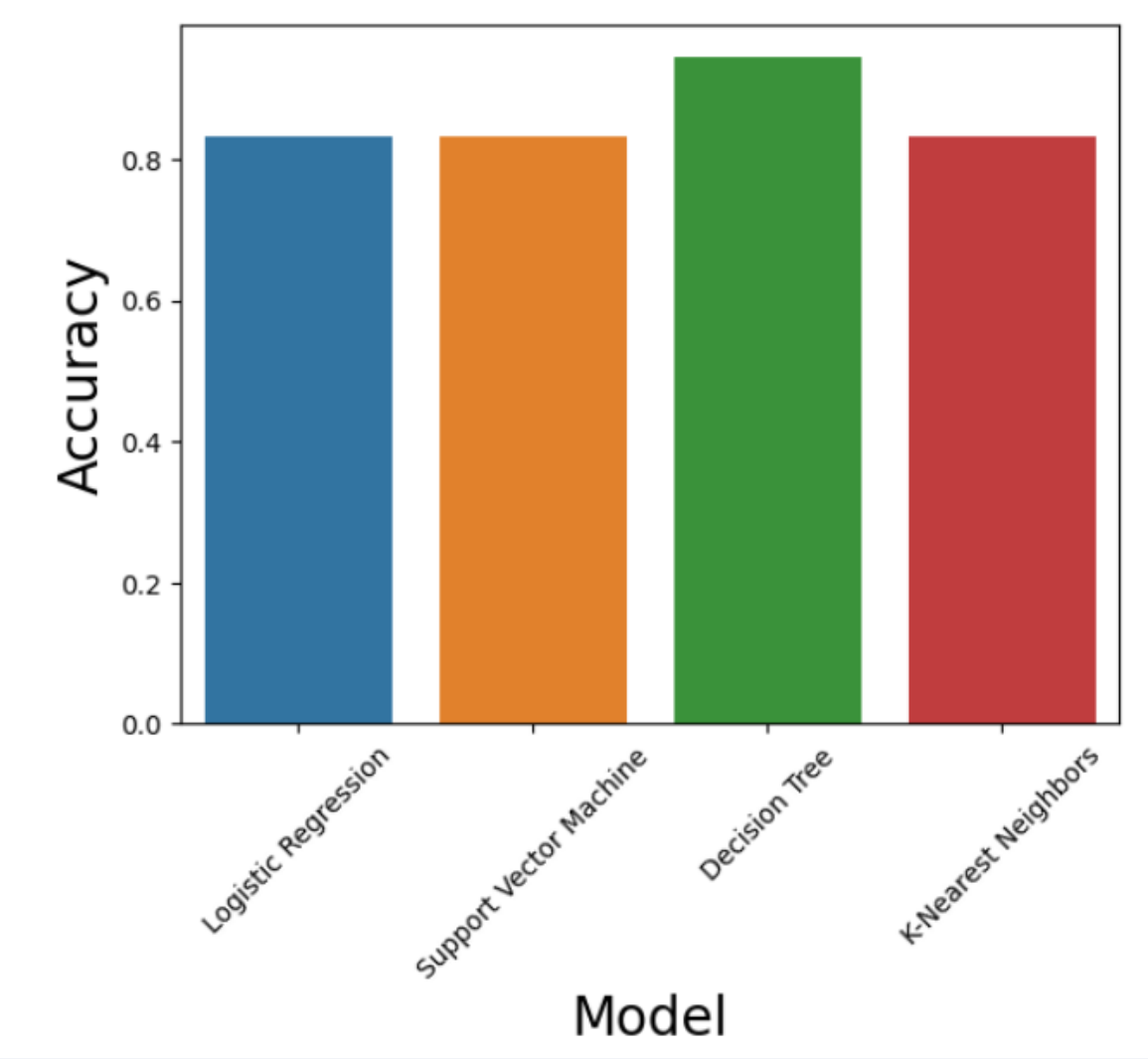
Payload vs. Launch Outcome… Continued

- We noticed that successful launch is associated with lower Payload for KSC LC, but there is no success launches beyond 5500.
- For CCAFS LC-40, successful launch is mostly associated with heavy Payload and specifically with FT Booster.
- VAFB SLC-4E has successful and failed launches at heavy loads, so we do not have clear association. Also, we notice that v1.1 has no successful launch at this site.
- CCAFS SLC-40 has no successful launch beyond 4000 kg payload.

Section 5

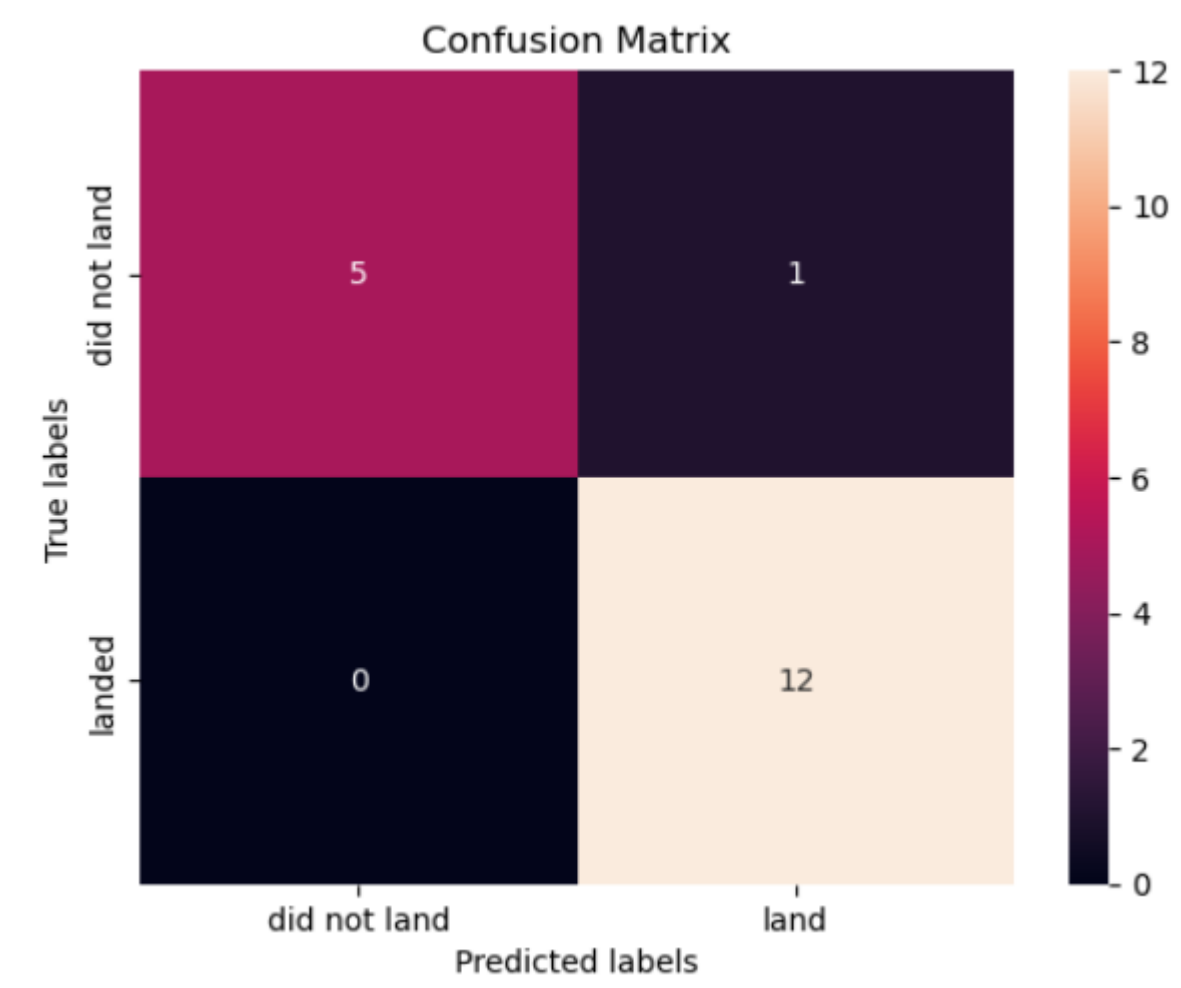
Predictive Analysis (Classification)

Classification Accuracy



We notice that Decision Tree was the best model that achieve the highest accuracy score of 94.4%

Confusion Matrix



The Decision Tree model was able to correctly predict the successful and failed landing with great accuracy. With only one incorrect prediction

Conclusions

- Success rate has been improving sine 2013.
- The best launch site was KSC LC-39A
- Four Orbits have no failed first stage landing: ES-L1, GEO, HEO, and SSO.
- Orbit SO has no successful landing.
- The impact of Flight Number is not consistent across different Orbits
- Generally, successful launch were noticed in heavy Payload range across all sites. However, the impact is not consistent in the lower range.
- Decision Tree was the best classification model to predict the landing outcome based on the dataset we studied.
- We found several factors that impact the success of the Falcon 9 First landing, SpaceX needs to consider them to ensure consistent success landing.

Appendix

Check GitHub for all workings

https://github.com/nalshakhoori/DataScienceCapstone/tree/main

Thank you!