IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Fernando Dalai Aguilar Sánchez
December 2022

Outline

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

Executive Summary

Summary of methodologies

- Data Collection
  (Request and BeautifulSoup)
- Data Wrangling
  (Pandas)
- Exploratory Data Analysis
  (SQL, Pandas and Matplotlib)
- Interactive Visual Analytics
  (Dash and Folium)
- Machine Learning
  (Scikit-learn)

Introduction

In this project we will predict if the Falcon 9 first stage will land successfully. 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.

SPACEX

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected by web scraping using the SpaceX API.
- Perform data wrangling
- 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, and evaluate classification models using the machine learning module called scikit-learn.

Data Collection

Process

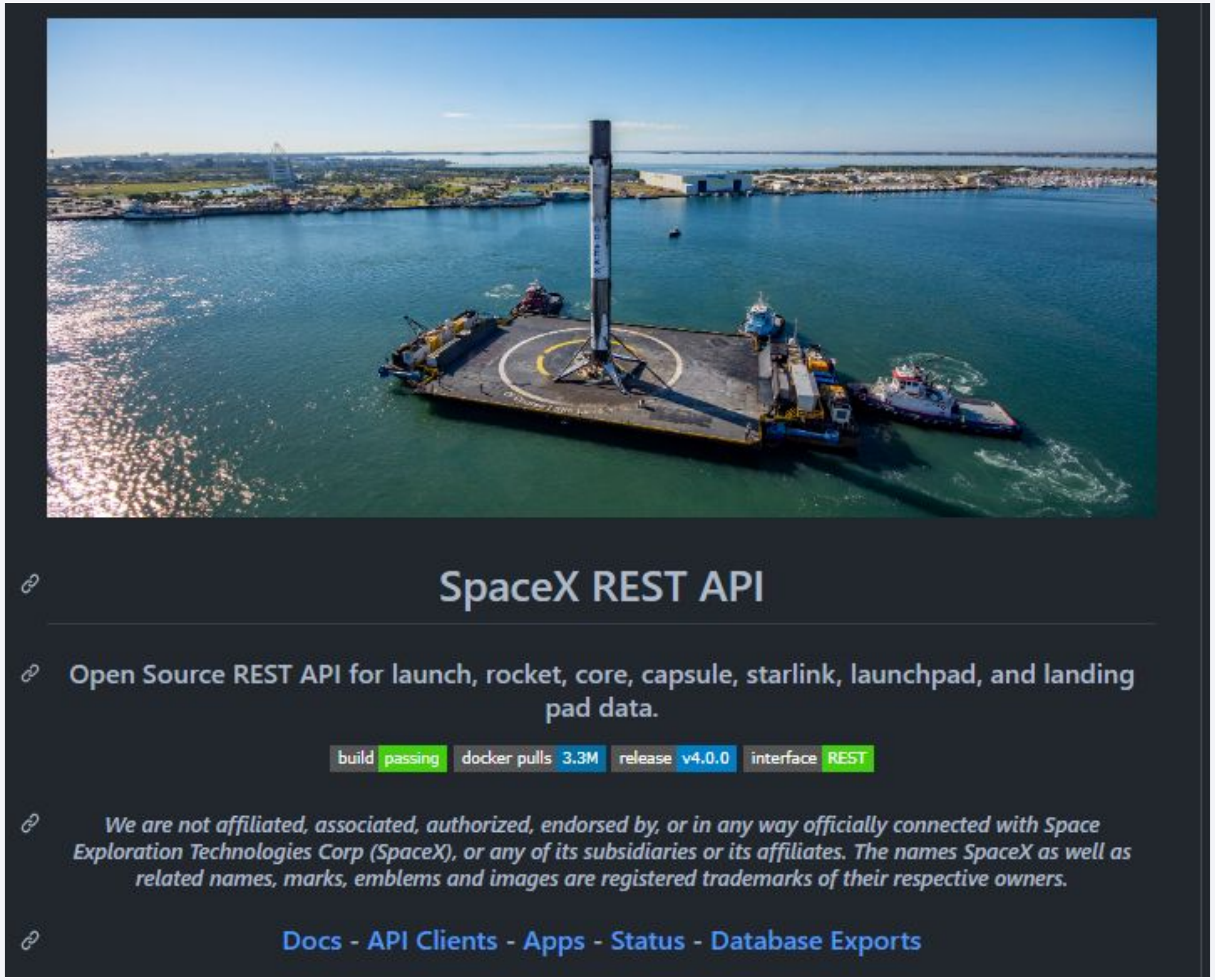
SpaceX API

- Collect the data using the SpaceX API.
- Convert the JSON data to a pandas dataframe.
- Pre-process the data.

Web Scraping

- Use web scraping to obtain the data for Falcon 9 launches.
- Convert to a pandas dataframe.
- Pre-process the data.

Data Collection – SpaceX API



- For the data collection I used the SpaceX API which one is an open source REST API for launch, rocket, core, capsule, starlink, launchpad, and landing pad data from the SpaceX rockets.
- The link to the notebook is: https://github.com/ferdal137/SpaceX/blob/main/Data_Collection/Data%20Collection%20API.ipynb

Data Collection - Scraping

- Furthermore, I undertook the task of collecting data from Wikipedia by employing web scraping techniques facilitated by the Python module Beautiful Soup.

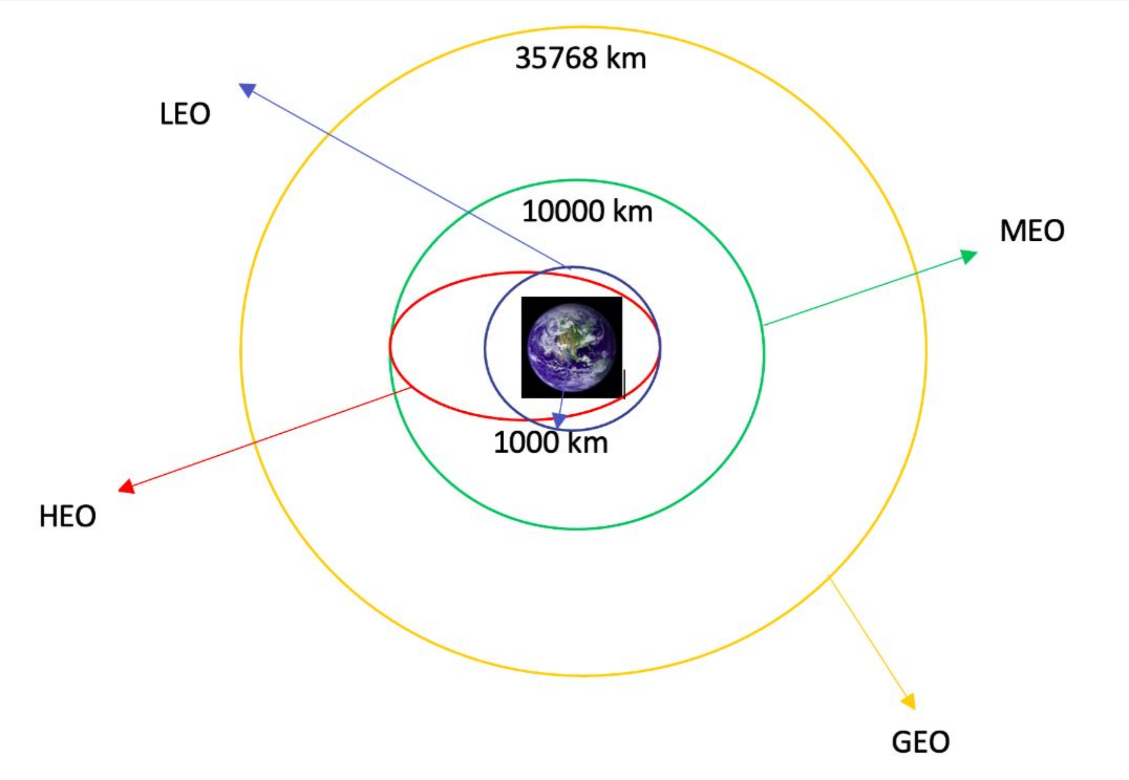
- The link to the notebook is: https://github.com/ferdal137/SpaceX/blob/main/Data%20Collection/Data%20Collection%20with%20Web%20Scraping.ipynb

2020 [edit]

In late 2019, Gwynne Shotwell stated that SpaceX hoped for as many as 24 launches for Starlink satellites in 2020,[490] in addition to 14 or 15 non-Starlink launches. At 26 launches, 13 of which for Starlink satellites, Falcon 9 had its most prolific year, and Falcon rockets were second most prolific rocket family of 2020, only behind China's Long March rocket family.[491]

[hide] Flight No.	Date and time (UTC)	Version, Booster[b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
78	7 January 2020, 02:19:21[492]	F9 B5 ♺ B1049.4	CCAFS, SLC-40	Starlink 2 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third large batch and second operational flight of Starlink constellation. One of the 60 satellites included a test coating to make the satellite less reflective, and thus less likely to interfere with ground-based astronomical observations.[493]								
79	19 January 2020, 15:30[494]	F9 B5 ♺ B1046.4	KSC, LC-39A	Crew Dragon in-flight abort test[495] (Dragon C205.1)	12,050 kg (26,570 lb)	Sub-orbital[496]	NASA (CTS)[497]	Success	No attempt
	An atmospheric test of the Dragon 2 abort system after Max Q. The capsule fired its SuperDraco engines, reached an apogee of 40 km (25 mi), deployed parachutes after reentry, and splashed down in the ocean 31 km (19 mi) downrange from the launch site. The test was previously slated to be accomplished with the Crew Dragon Demo-1 capsule;[498] but that test article exploded during a ground test of SuperDraco engines on 20 April 2019.[419] The abort test used the capsule originally intended for the first crewed flight.[499] As expected, the booster was destroyed by aerodynamic forces after the capsule aborted.[500] First flight of a Falcon 9 with only one functional stage — the second stage had a mass simulator in place of its engine.								
80	29 January 2020, 14:07[501]	F9 B5 ♺ B1051.3	CCAFS, SLC-40	Starlink 3 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third operational and fourth large batch of Starlink satellites, deployed in a circular 290 km (180 mi) orbit. One of the fairing halves was caught, while the other was fished out of the ocean.[502]								
81	17 February 2020, 15:05[503]	F9 B5 ♺ B1056.4	CCAFS, SLC-40	Starlink 4 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Failure (drone ship)
	Fourth operational and fifth large batch of Starlink satellites. Used a new flight profile which deployed into a 212 km × 386 km (132 mi × 240 mi) elliptical orbit instead of launching into a circular orbit and firing the second stage engine twice. The first stage booster failed to land on the drone ship[504] due to incorrect wind data.[505] This was the first time a flight proven booster failed to land.								
82	7 March 2020, 04:50[506]	F9 B5 ♺ B1059.2	CCAFS, SLC-40	SpaceX CRS-20 (Dragon C112.3 ♺)	1,977 kg (4,359 lb)[507]	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
	Last launch of phase 1 of the CRS contract. Carries *Bartolomeo*, an ESA platform for hosting external payloads onto ISS.[508] Originally scheduled to launch on 2 March 2020, the launch date was pushed back due to a second stage engine failure. SpaceX decided to swap out the second stage instead of replacing the faulty part.[509] It was SpaceX's 50th successful landing of a first stage booster, the third flight of the Dragon C112 and the last launch of the cargo Dragon spacecraft.								
83	18 March 2020, 12:16[510]	F9 B5 ♺ B1048.5	KSC, LC-39A	Starlink 5 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Failure (drone ship)
	Fifth operational launch of Starlink satellites. It was the first time a first stage booster flew for a fifth time and the second time the fairings were reused (Starlink flight in May 2019).[511] Towards the end of the first stage burn, the booster suffered premature shut down of an engine, the first of a Merlin 1D variant and first since the CRS-1 mission in October 2012. However, the payload still reached the targeted orbit.[512] This was the second Starlink launch booster landing failure in a row, later revealed to be caused by residual cleaning fluid trapped inside a sensor.[513]								
84	22 April 2020, 19:30[514]	F9 B5 ♺ B1051.4	KSC, LC-39A	Starlink 6 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)

Data Wrangling

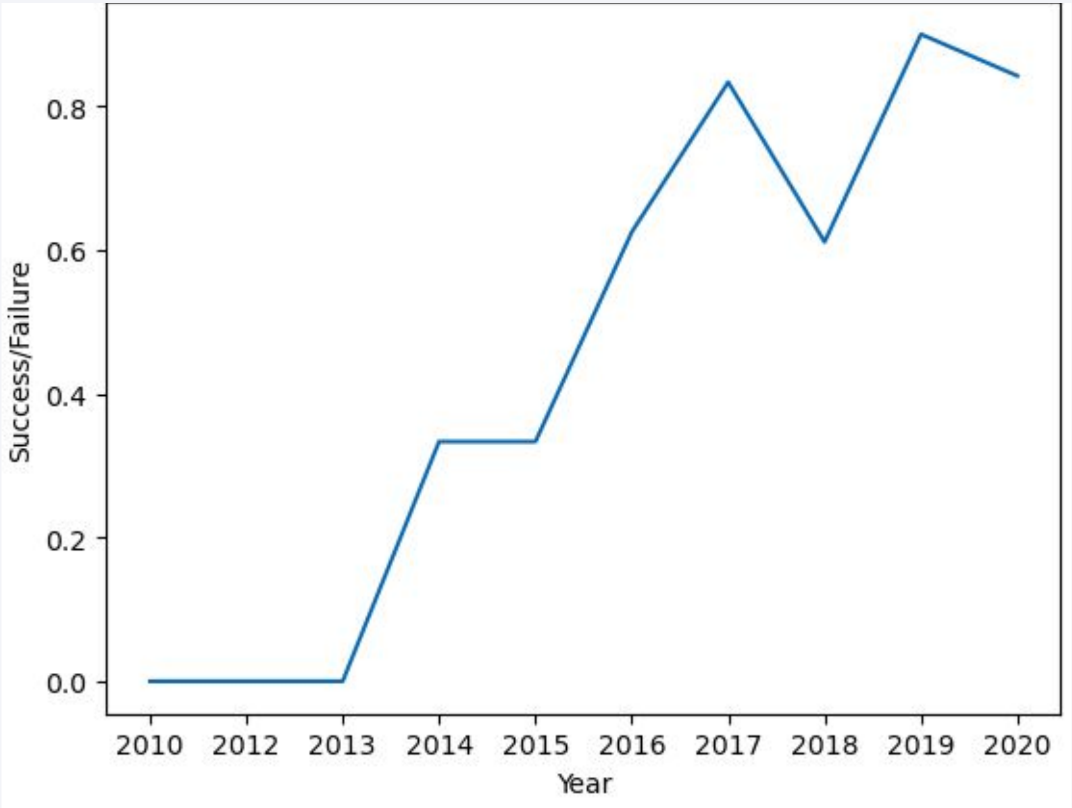


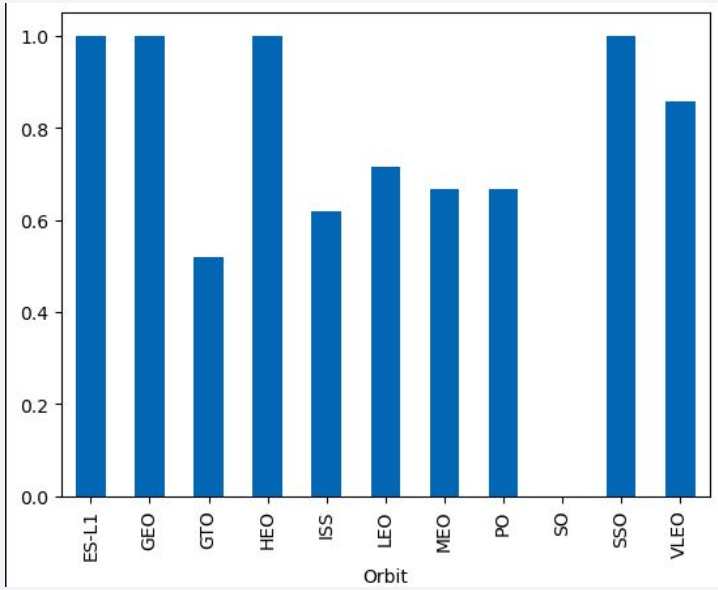
- For the data wrangling, I had to calculate the number of launches from each site, the number and occurrence of each orbit and the number and occurrence of mission outcomes per orbit type

- The link to the notebook is: https://github.com/ferdal137/SpaceX/blob/main/Data%20Wrangling/Data%20Wrangling.ipynb

EDA with Data Visualization

- In the Exploratory Data Analysis, I used pandas for data manipulation in dataframes and matplotlib for the plots.





- The link to the notebook is: https://github.com/ferdal137/SpaceX/blob/main/Exploratory%20Analysis/Exploratory_Analysis_Pandas.ipynb

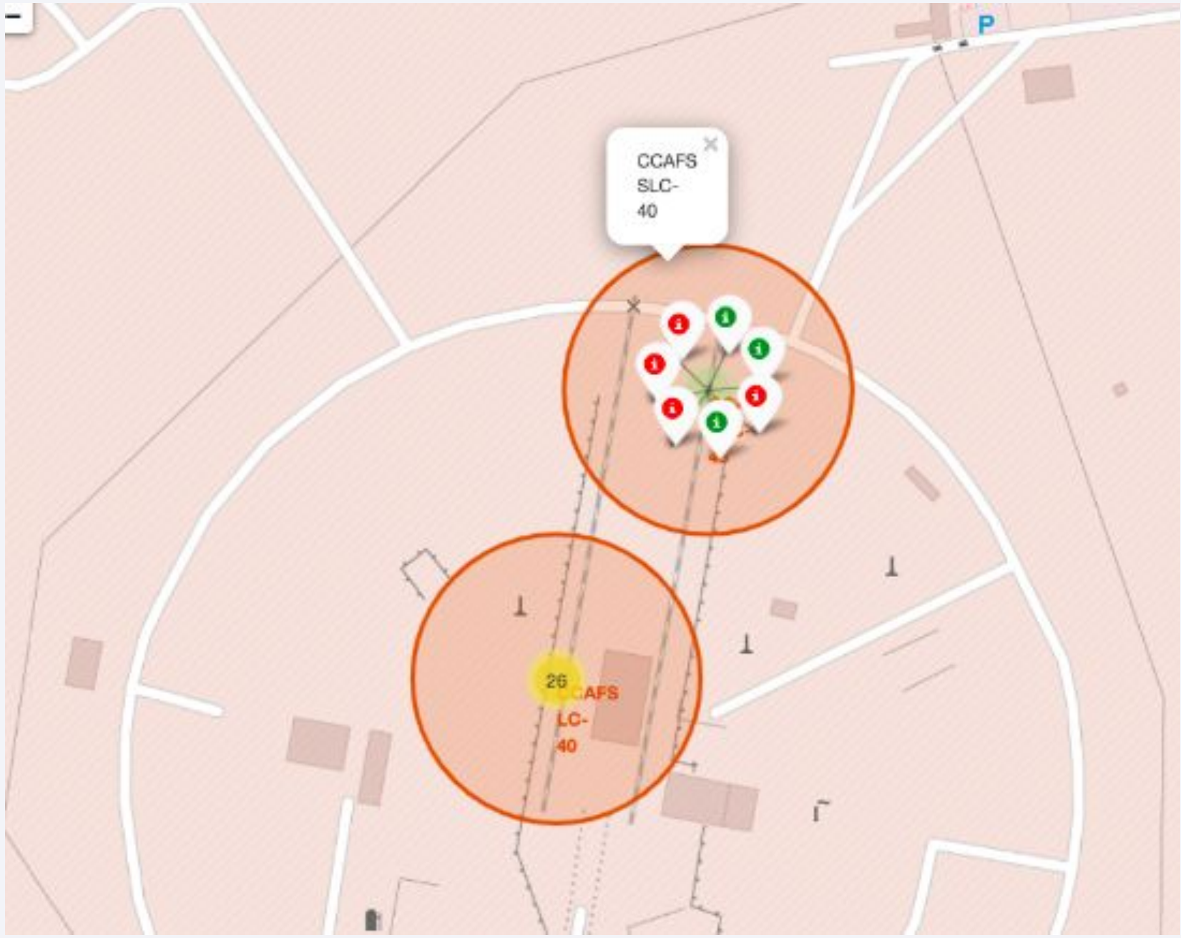
EDA with SQL

For the EDA using SQL, I had to perform the following tasks:

- 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 the average payload mass carried by booster version F9 v1.1.
- List the date when the first successful landing outcome on the ground pad was achieved.
- List the total number of successful and failed mission outcomes.

The link to the notebook is:
https://github.com/ferdal137/SpaceX/blob/main/Exploratory%20Analysis/Exploratory_Analysis_SQL.ipynb

Build an Interactive Map with Folium



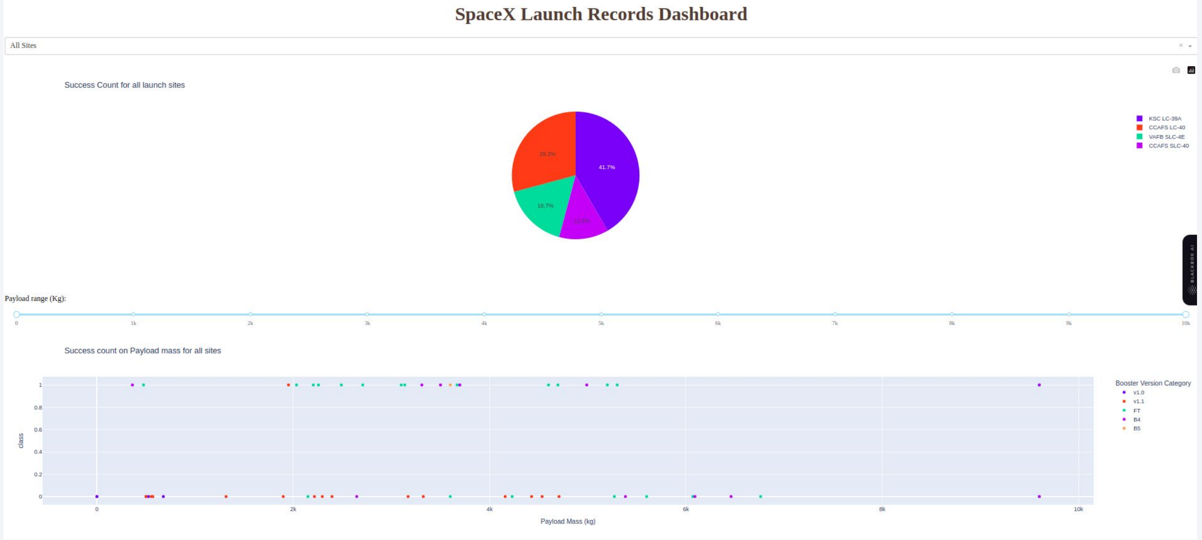
Folium Methods:

- folium.Map(): Creates a map.
- folium.Marker(): Adds a marker.
- folium.Circle(): Adds a circle shape.
- folium.Polygon(): Adds a polygon shape
- folium.Popup(): Creates a popup for markers.

- The link to the notebook is: https://github.com/ferdal137/SpaceX/blob/main/Data%20visualization/Launch_Site_Folium_Map.ipynb

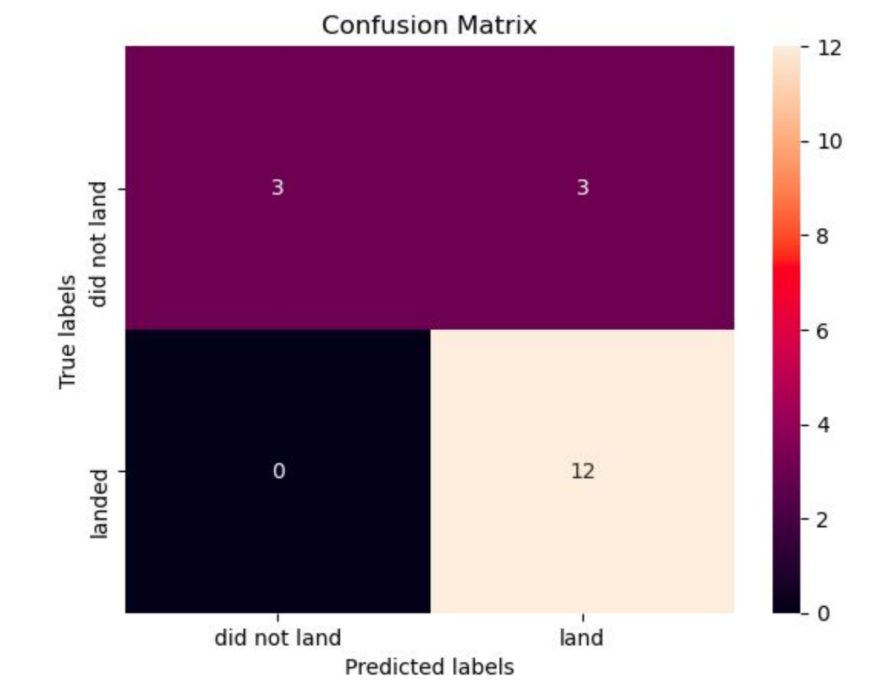
Build a Dashboard with Plotly Dash

- Using Dash, I made an interactive launch dashboard. It shows total successful launches and payload mass success counts for each site. Users can select sites from a dropdown menu, instantly updating the data. It's compact and informative for space launch analysis



- The link to the dash is: https://github.com/ferdal137/SpaceX/tree/main/Data%20visualization/Dash

Predictive Analysis (Classification)



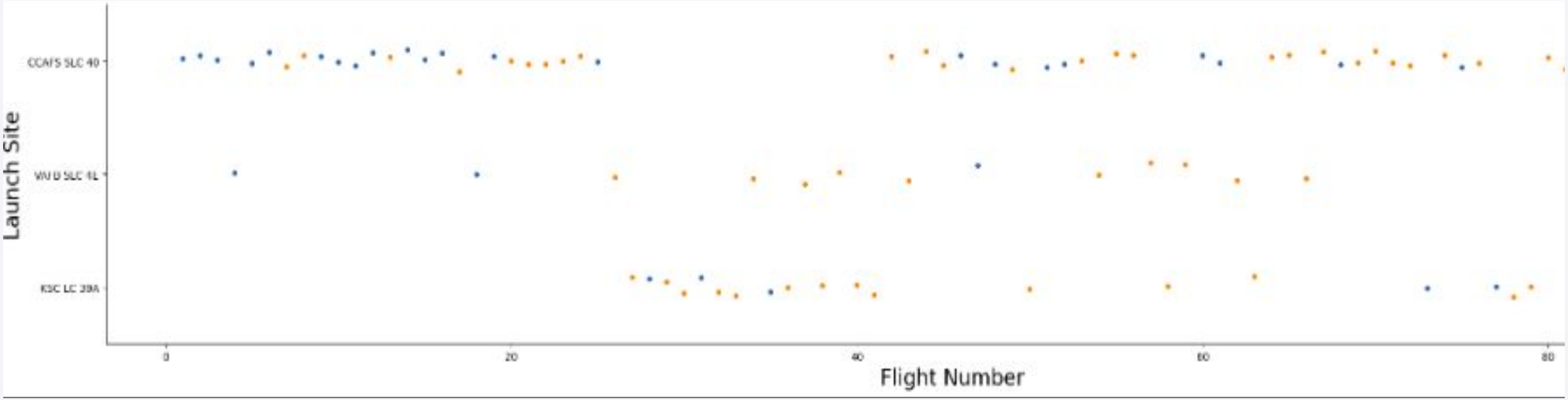
- The link to the notebook is: https://github.com/ferdal137/SpaceX/blob/main/Machine%20Learning/Machine_Learning_Prediction.ipynb

- The data was loaded using numpy and pandas, followed by data transformation and splitting into training and testing sets. Multiple machine learning models were built and various hyperparameters were tuned. The model's performance was evaluated using accuracy as the primary metric, resulting in significant improvements. Ultimately, the best performing classification model was identified through meticulous analysis.

Section 2

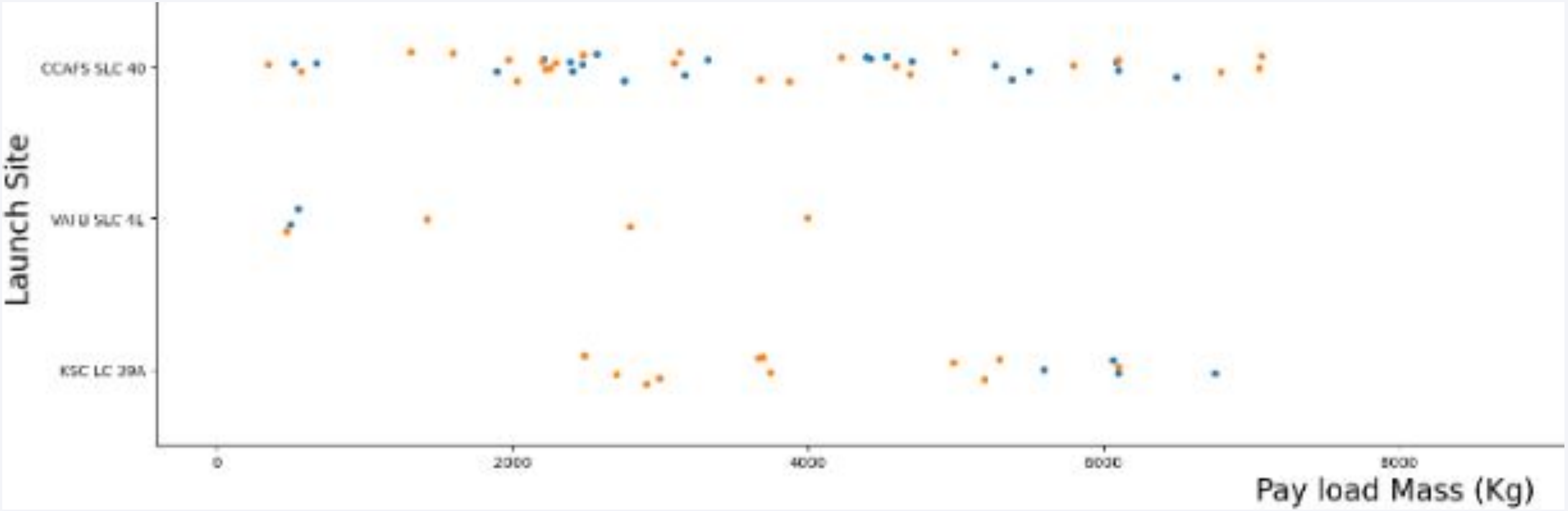
Insights drawn from EDA

Flight Number vs. Launch Site

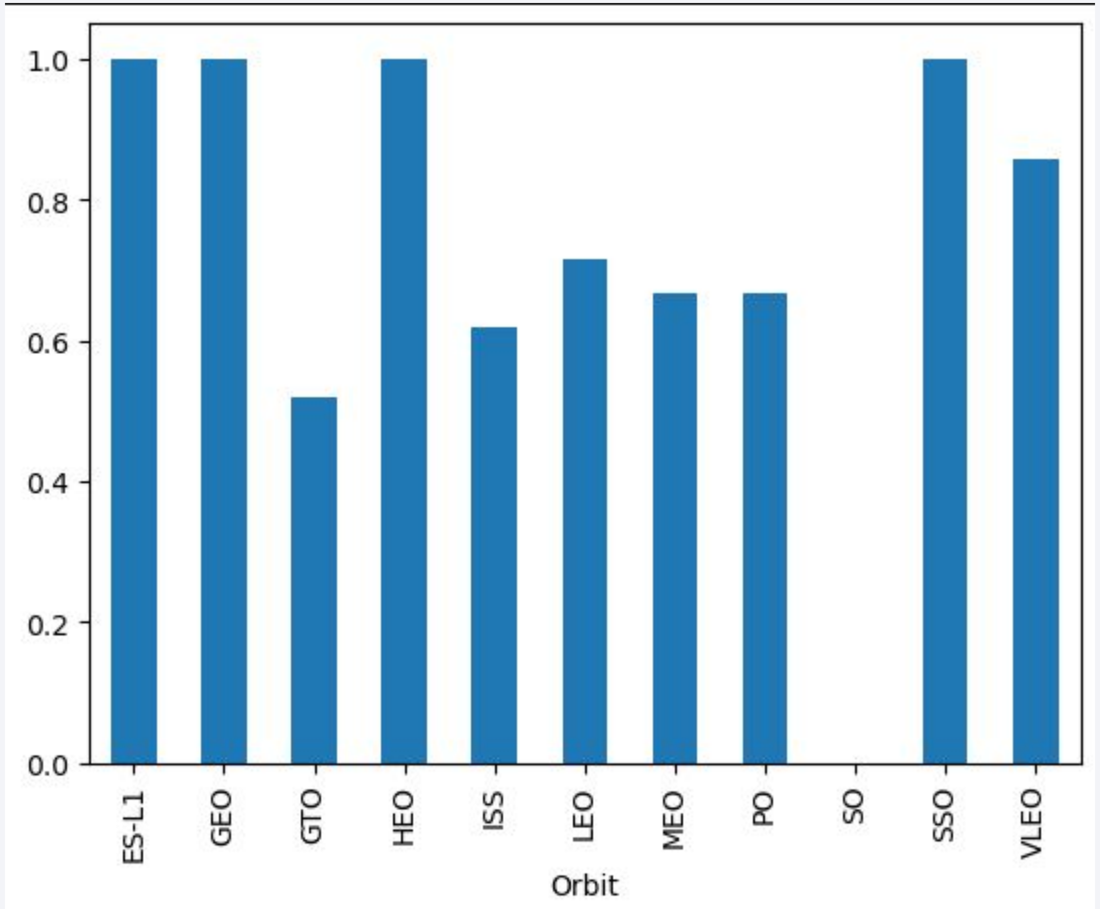


Over the course of time, the rate of success for launches at every launch site has demonstrated improvement. Notably, the Cape Canaveral Air Force Station Space Launch Complex 40 (CCAFS SLC 40) has displayed a particularly noteworthy increase in successful launches, as it has become the primary launch site for the majority of space missions.

Payload vs. Launch Site

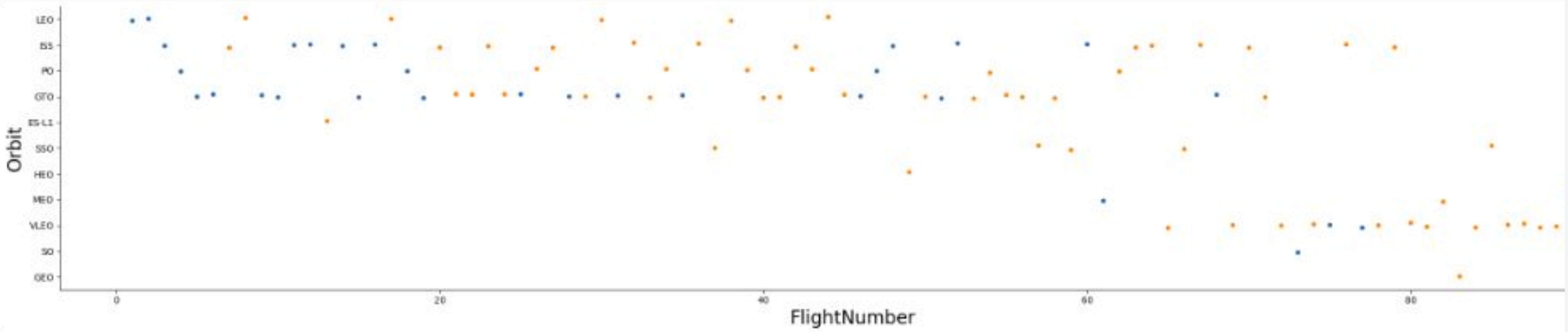


Success Rate vs. Orbit Type



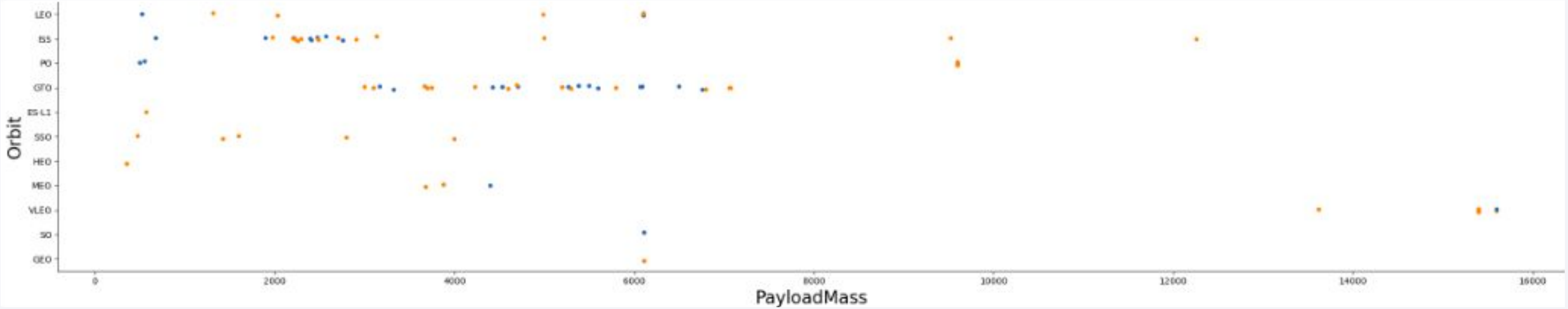
- In the bar chart, we can appreciate that the orbit types ES-L1, GEO, HEO, and SSO have a success rate of 100 percent.

Flight Number vs. Orbit Type

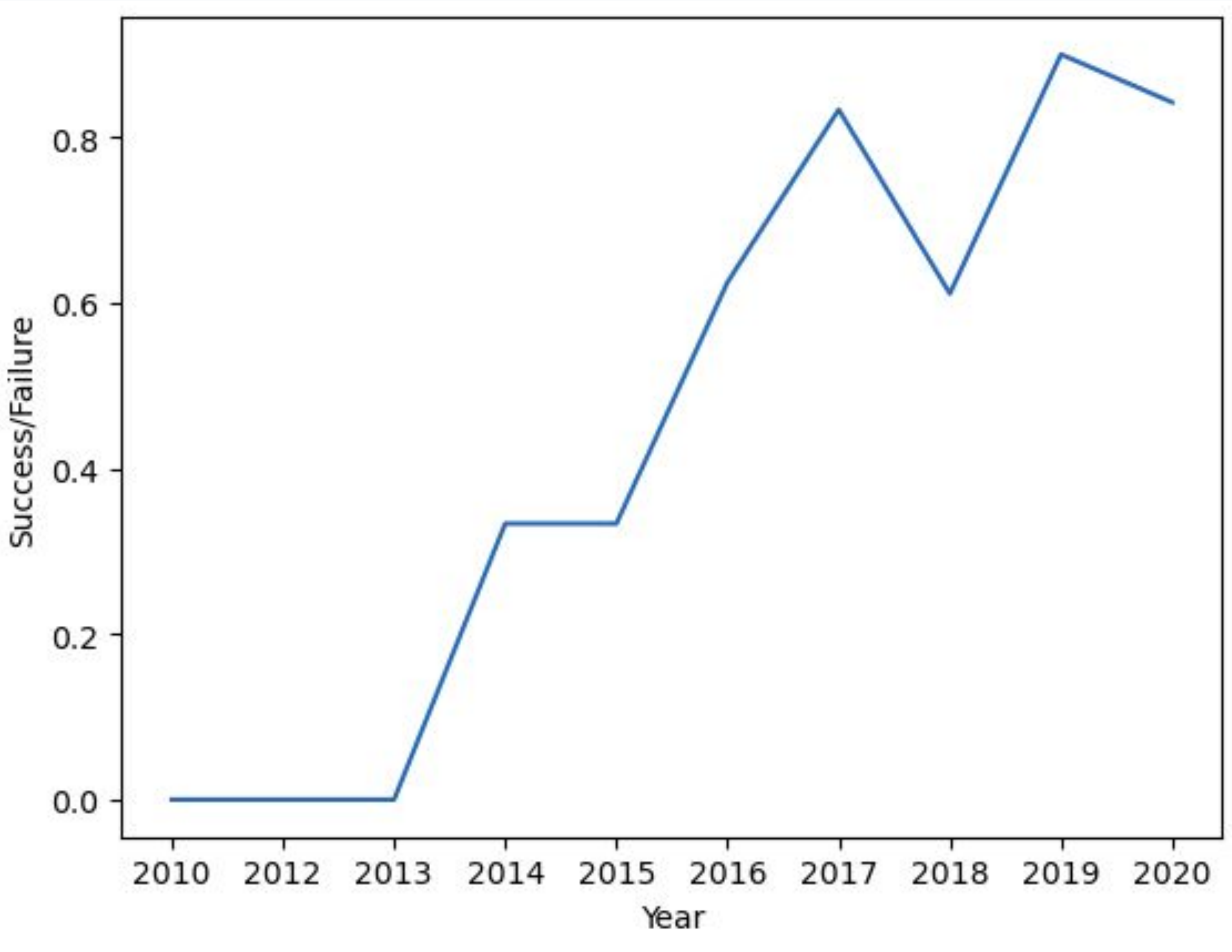


- In this plot, we can see that there is no correlation between the flight number and the orbit. The data points are scattered randomly across the graph, indicating that there is no discernible pattern or relationship between these two variables.

Payload vs. Orbit Type



Launch Success Yearly Trend



- In the line graph, we can see that starting from the year 2013, there was a quite significant increase in the success rate, rising until 2018, where it decreased a bit but then increased again, reaching its peak in 2019.

All Launch Site Names

```
%%sql

SELECT DISTINCT Launch_Site FROM SPACEXTBL
```

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

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

- The given SQL query selects distinct values from the column "Launch_Site" in the table "SPACEXTBL." It retrieves unique launch site names, eliminating any duplicate entries, and presents them as the result of the query.

Launch Site Names Begin with 'CCA'

```
%%sql

SELECT * FROM SPACEXTBL WHERE(Launch_Site) LIKE "CCA%" LIMIT 5;
```

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

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

- This query selects all columns from the table "SPACEXTBL" where the "Launch_Site" column starts with "CCA" and limits the output to a maximum of five rows.

Total Payload Mass

```
%%sql

SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE(Customer) = "NASA (CRS)"
```

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

SUM(PAYLOAD_MASS__KG_)
45596

- This SQL query calculates the sum of the "PAYLOAD_MASS__KG_" column from the table "SPACEXTBL" where the "Customer" column has the value "NASA (CRS)." It retrieves the total payload mass in kilograms for all records that correspond to NASA's Commercial Resupply Services (CRS) missions.

Average Payload Mass by F9 v1.1

```
%%sql

SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE(Booster_Version) LIKE "F9 v1.1%"
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

- This SQL query calculates the average of the "PAYLOAD_MASS__KG_" column from the table "SPACEXTBL" where the "Booster_Version" column starts with "F9 v1.1." It computes the average payload mass in kilograms for all records that match the specified condition, focusing on missions that used the Falcon 9 version 1.1 booster.

First Successful Ground Landing Date

```
%%sql

SELECT MIN(DATE) FROM SPACEXTBL WHERE("Landing _Outcome")="Success (ground pad)"
```

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

MIN(DATE)
01-05-2017

- This SQL query selects the minimum value from the "DATE" column in the table "SPACEXTBL" where the "Landing _Outcome" column has the value "Success (ground pad)." It retrieves the earliest date among records that represent successful landings on ground pads for SpaceX missions.

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
```

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

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

- This SQL query selects the "Booster_Version" column from the table "SPACEXTBL" where the "Landing Outcome" column has the value "Success (drone ship)" and the "PAYLOAD_MASS__KG" column falls within the range of 4000 to 6000 kilograms. It retrieves the booster versions used in missions with a successful landing on a drone ship and a payload mass between 4000 and 6000 kilograms.

Total Number of Successful and Failure Mission Outcomes

```
4]: %%sql

    SELECT Mission_Outcome, COUNT(Mission_Outcome) AS Total_Number FROM SPACEXTBL GROUP BY Mission_Outcome

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

4]:

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

- This SQL query selects the "Mission_Outcome" column and counts the occurrences of each unique "Mission_Outcome" in the table "SPACEXTBL." The "COUNT" function is used to tally the occurrences, and the result is presented as a new column labeled "Total_Number." The "GROUP BY" clause organizes the results based on the distinct values in the "Mission_Outcome" column, providing a summary of the total count for each mission outcome category.

Boosters Carried Maximum Payload

```
%%sql
SELECT Booster_Version FROM SPACEXTBL WHERE(PAYLOAD_MASS__KG_) = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

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

- This SQL query selects the "Booster_Version" column from the table "SPACEXTBL" where the "PAYLOAD_MASS__KG_" column is equal to the maximum payload mass found in the same column of the "SPACEXTBL" table. In other words, it retrieves the booster version used in the mission with the highest payload mass recorded in the table.

2015 Launch Records

```
%sql SELECT substr(Date,4,2) as month, BOOSTER_VERSION, LAUNCH_SITE, [Landing _Outcome] , DATE \
FROM SPACEXTBL \
where [Landing _Outcome] = 'Failure (drone ship)' and substr(Date,7,4)='2015';
```

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

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

- This SQL query selects the month of the "DATE" column, along with the "MISSION_OUTCOME," "BOOSTER_VERSION," and "LAUNCH_SITE" columns, from the table "SPACEXTBL." It retrieves data only for records from the year 2015, using the "EXTRACT" function to extract the year from the "DATE" column and then filter the results accordingly. The query presents information about the mission outcome, booster version, and launch site for all SpaceX missions that took place in the year 2015.

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

```
ql SELECT [Landing _Outcome], count(*) as count_outcomes \
OM SPACEXTBL \
ERE DATE between '04-06-2010' and '20-03-2017' group by [Landing _Outcome] order by count_outcomes DESC;
```

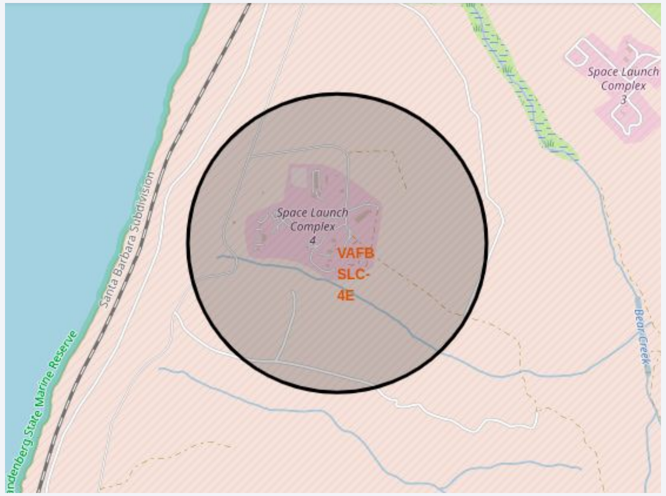
```
 * sqlite:///my_data1.db
Done.
```

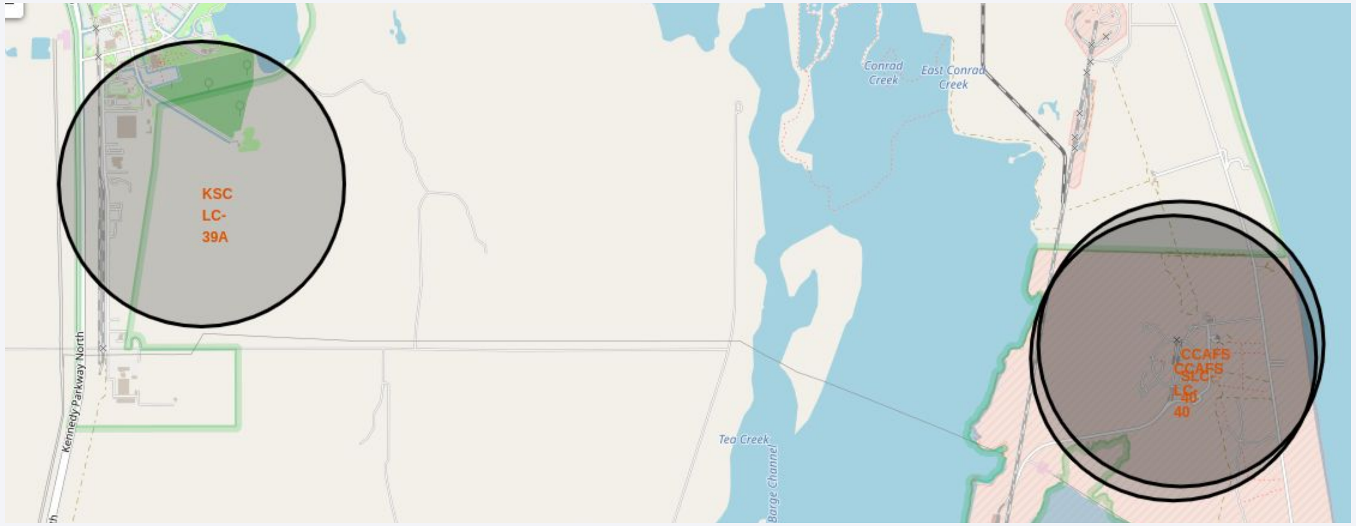
Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

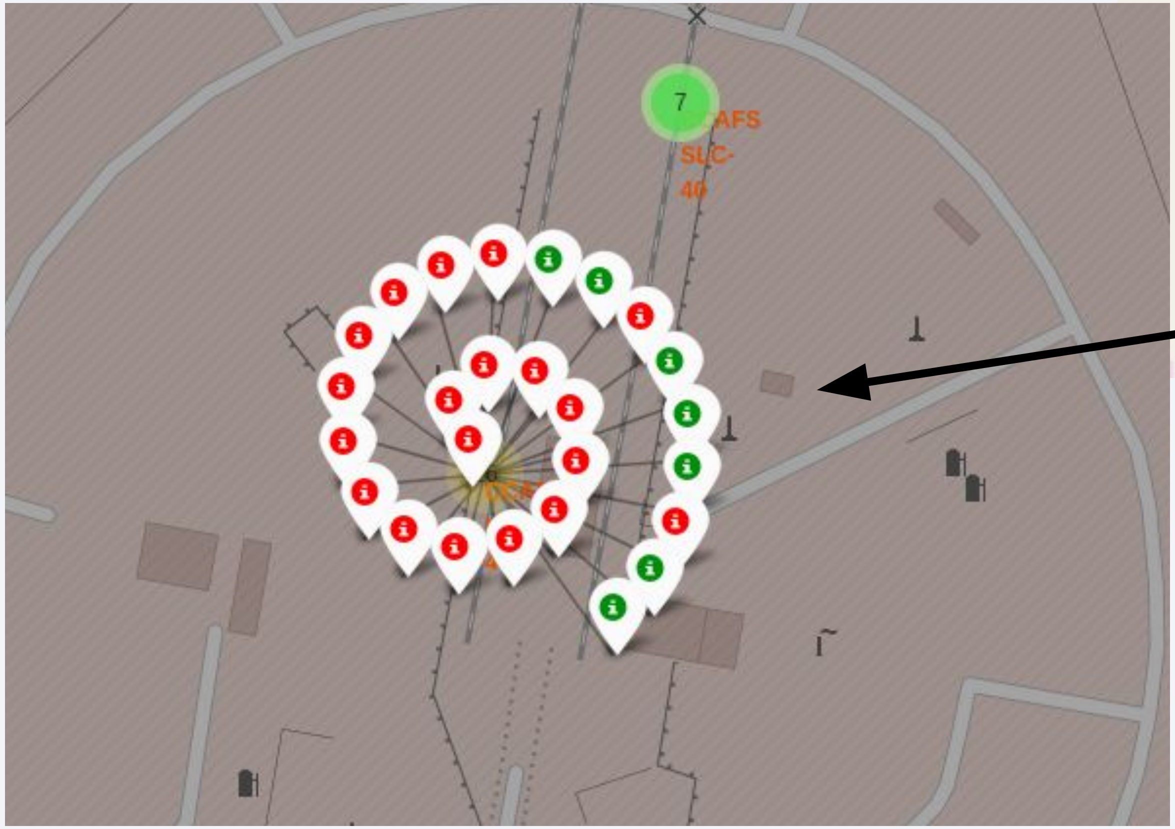
Launch Sites Proximities Analysis

All Launch Sites' Location Markers

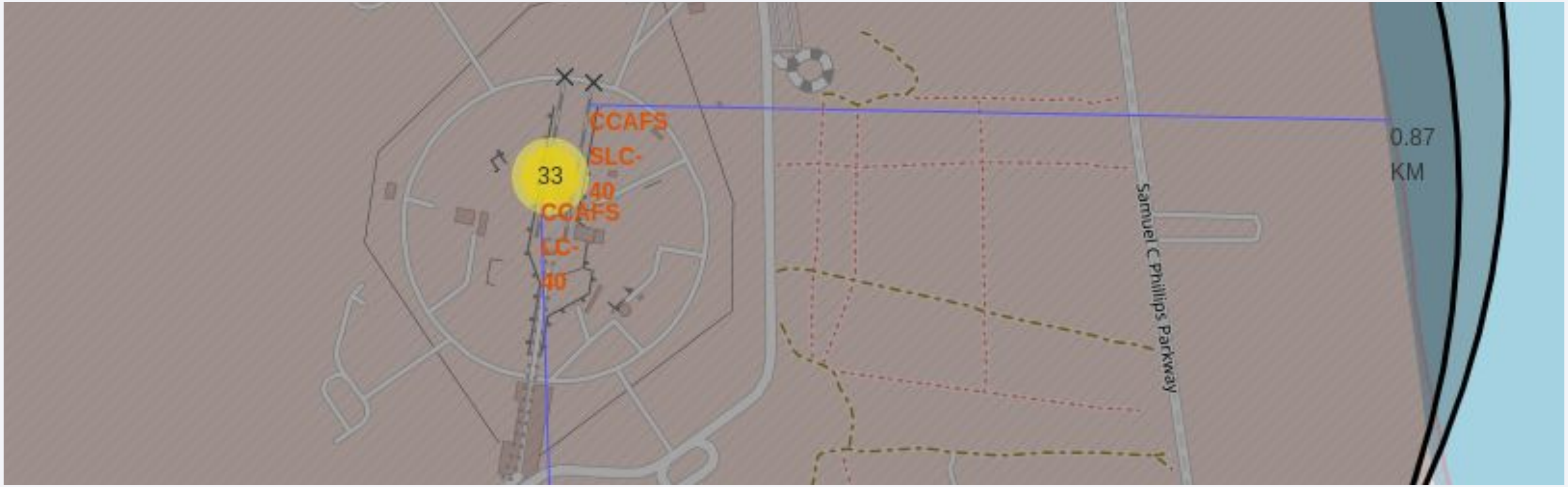




Color-labeled Launch Outcomes



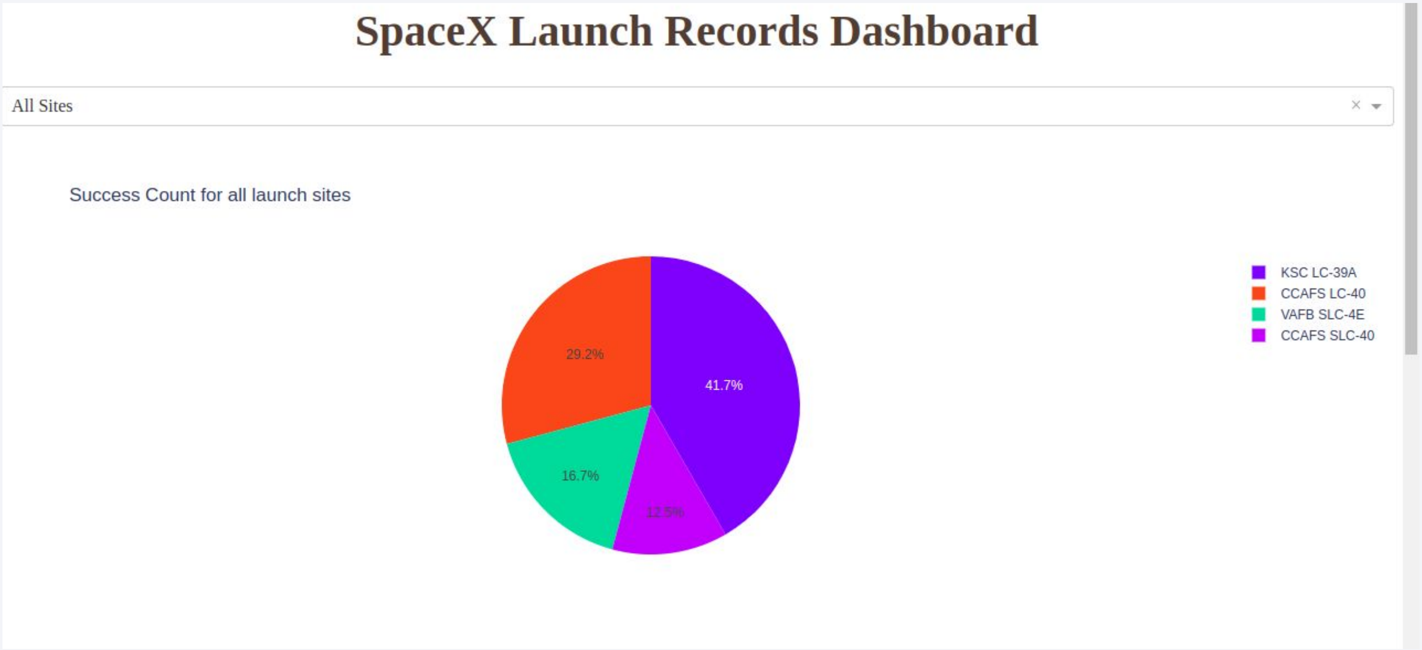
Launch Sites to its Proximities



Section 4

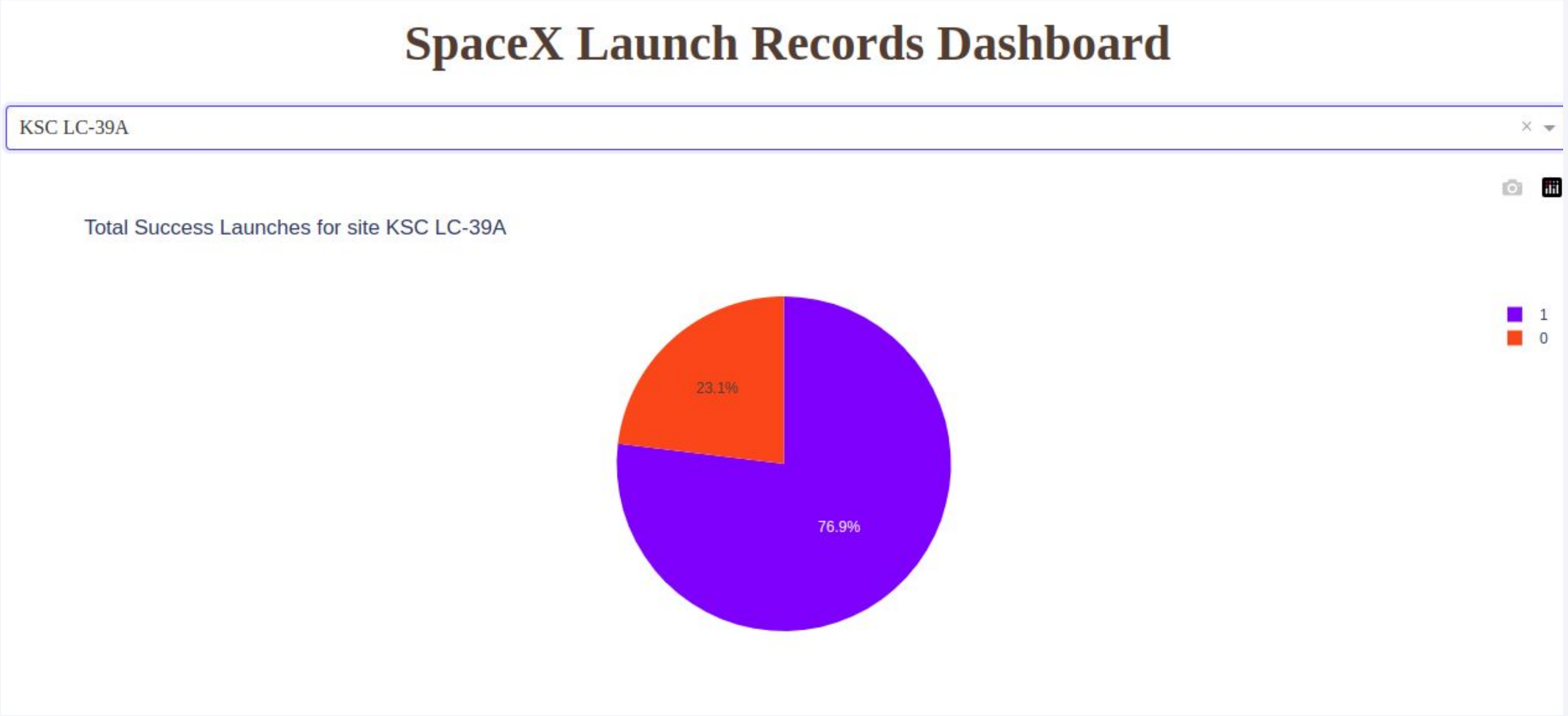
Build a Dashboard with Plotly Dash

Success Count for all launch sites



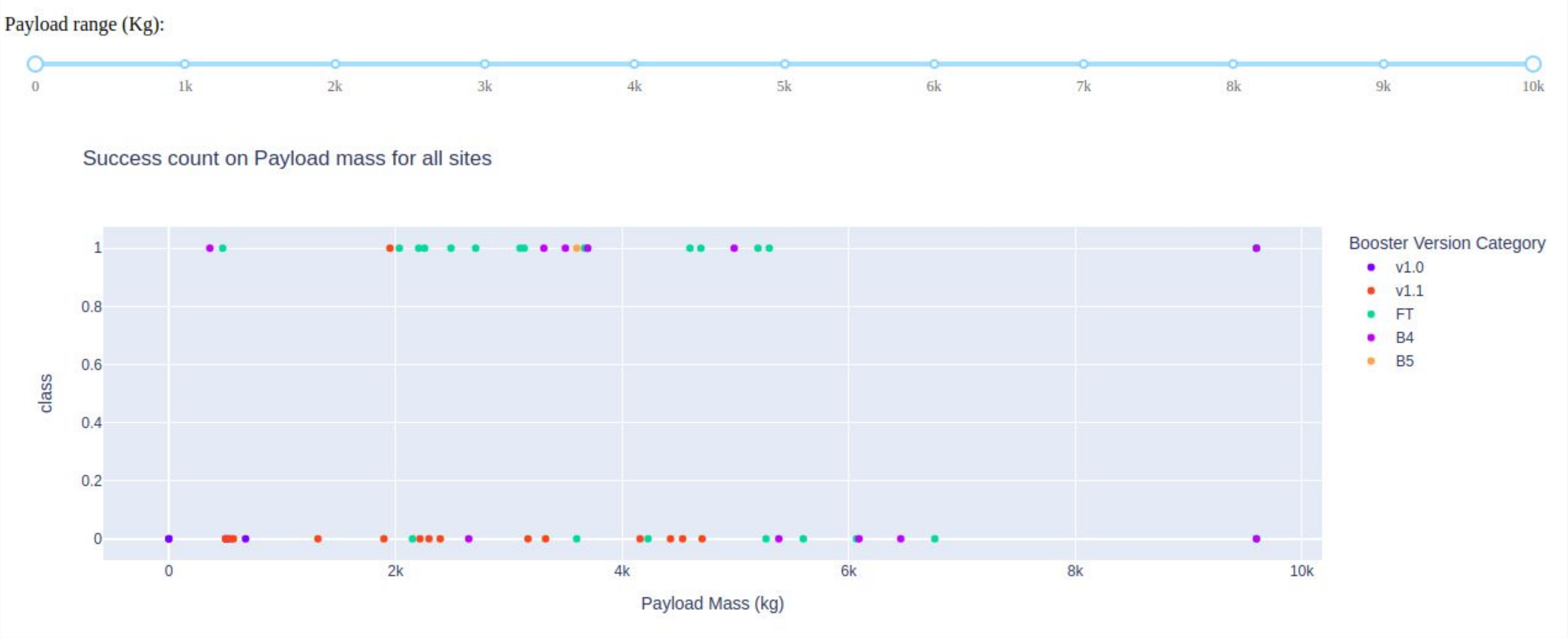
- In the pie chart, we can appreciate the success percentage of each launch site, with KSC LC-39A being the highest at 41.7%.

Total Success Launches for site KSC LC-39A



- In this other pie chart, we can see the percentage of successful and failed launches from the launch site KSC LC-39A, observing that 76.9% have been successful.

Payload vs Launch Outcome



Section 5

Predictive Analysis (Classification)

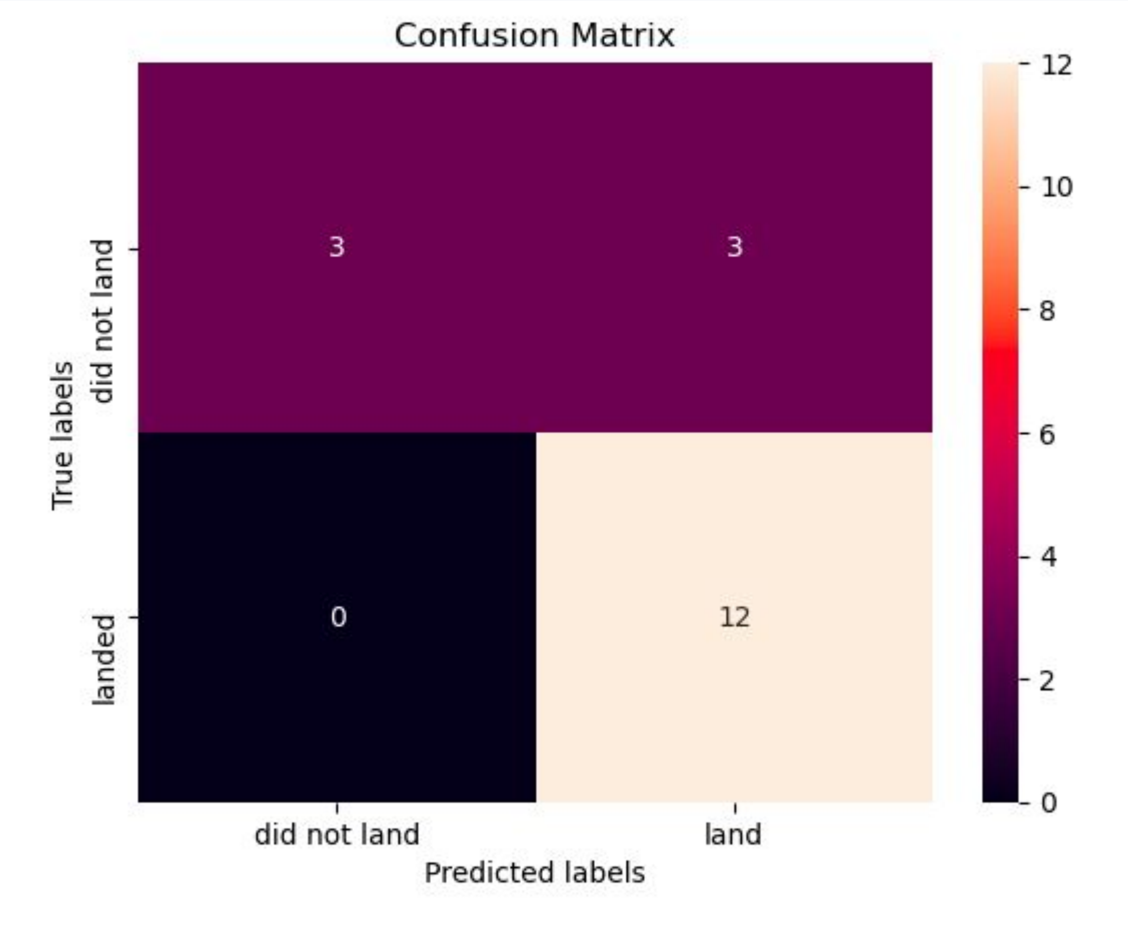
Classification Accuracy

Find the method performs best:

```
print('Accuracy for Logistics Regression method:', logreg_cv.score(X_test, Y_test))
print( 'Accuracy for Support Vector Machine method:', svm_cv.score(X_test, Y_test))
print('Accuracy for Decision tree method:', tree_cv.score(X_test, Y_test))
print('Accuracy for K nearsdt neighbors method:', knn_cv.score(X_test, Y_test))
```

```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.7777777777777778
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```

Confusion Matrix



Conclusions

- The most successful orbits in space have been ES-L1, GEO, HEO, SSO, and VLEO
- From 2013 to 2020, the launch success rate exhibited a steady increase.
- For the comprehension of the data, data visualizations in plots and maps prove to be highly beneficial.
- The worst method of classification was the decision tree, with an accuracy of 0.7777.

Appendix

- In addition, you can visit my project's GitHub repository for other resources: https://github.com/ferdal137/SpaceX

Thank you!