Data Science Capstone Project

SpaceX Project

Fatih ÖZGÜL

18.04.2023

Outline

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

Summary of methodologies

- Data collection
- Data wrangling
- EDA with Data Visualization
- EDA with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

- EDA results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

SpaceX is the leading company in commercial space travel, as they have made it more affordable. They offer the Falcon 9 rocket launch on their website for 62 million dollars, which is significantly cheaper than other providers whose prices start at 165 million dollars each. This is because SpaceX is able to reuse the first stage of their rockets. Therefore, by determining whether or not the first stage will land successfully, we can predict the cost of a launch. Using publicly available information and machine learning models, we aim to predict whether SpaceX will be able to reuse the first stage of their rocket.

At this point, some questions come to mind:
In what ways do factors like payload mass, launch location, number of flights, and orbits impact the success rate of the first stage landing?
Is there an upward trend in the success rate over time?
Additionally, which algorithm would be most effective for binary classification in this scenario?

Methodology

Data collection methodology:

- Using SpaceX Rest API
- Using Web Scrapping from Wikipedia

Performed data wrangling

- Filtering the data
- Dealing with missing values
- Using One Hot Encoding to prepare the data to a binary classification

Performed exploratory data analysis (EDA) using visualization and SQL

Performed interactive visual analytics using Folium and Plotly Dash

Performed predictive analysis using classification models

- Building, tuning and evaluation of classification models to ensure the best results

Methodology

Data collection

To gather comprehensive information for a detailed analysis of SpaceX launches, we utilized two methods of data collection: retrieving data through API requests from SpaceX REST API, and extracting information from a table on SpaceX's Wikipedia page by web scraping. Both methods were necessary to ensure that we had all the data we needed.

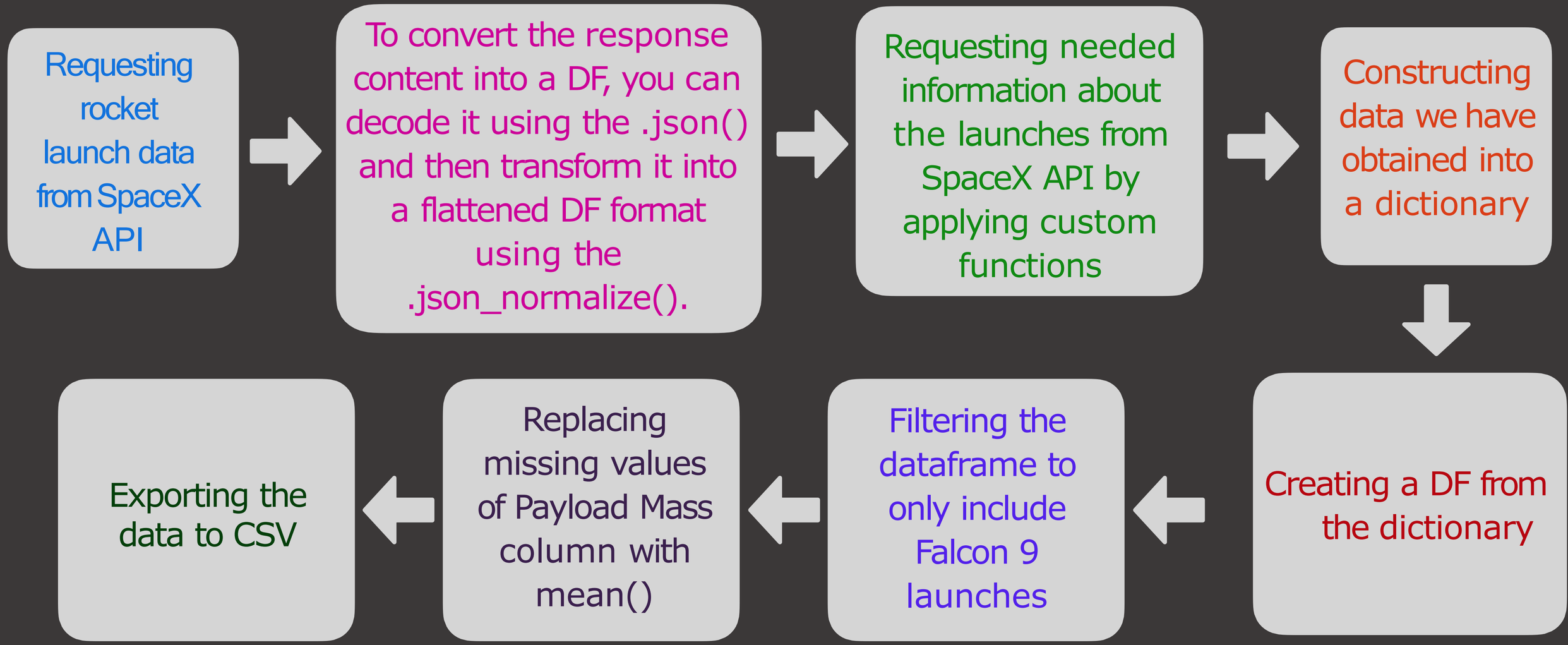
Data Columns are obtained by using SpaceX REST API:

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

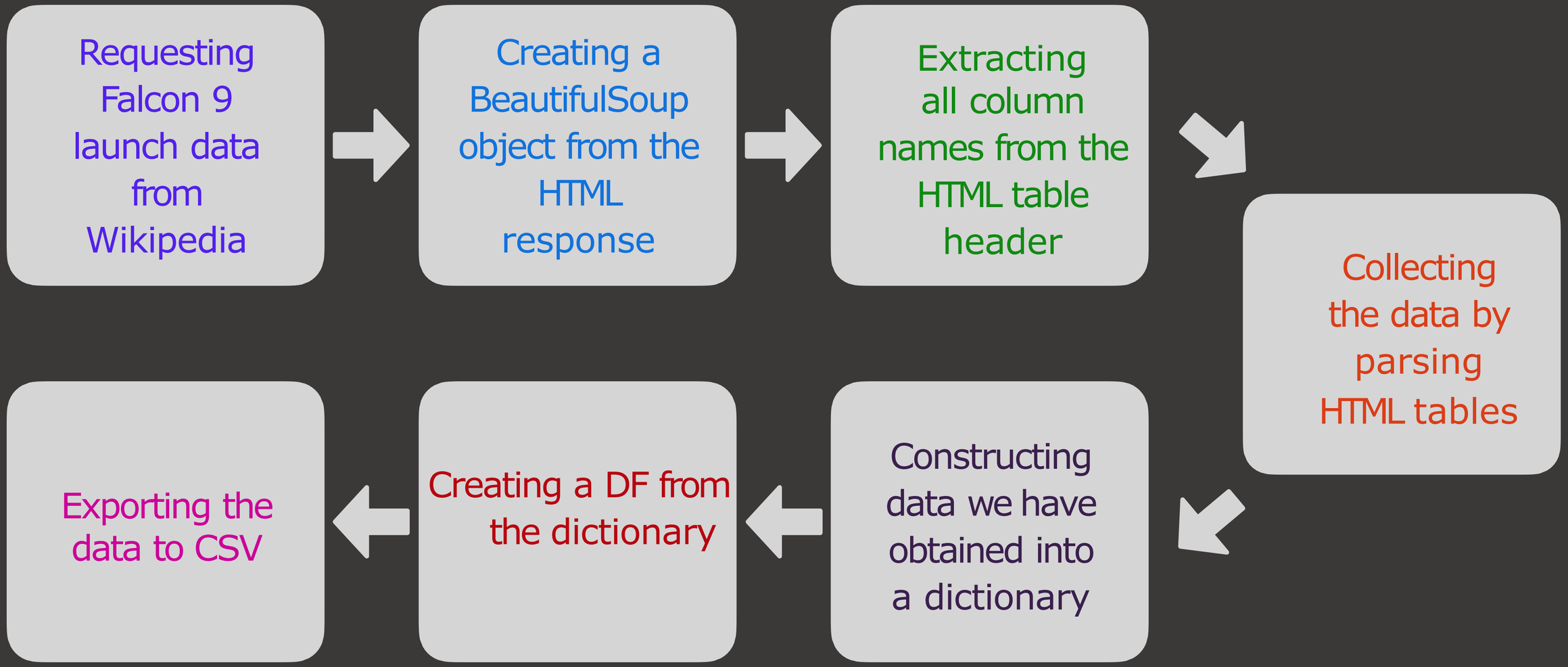
Data Columns are obtained by using Wikipedia Web Scraping:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer,
Launch outcome, Version Booster, Booster landing, Date, Time

Data collection – SpaceX API

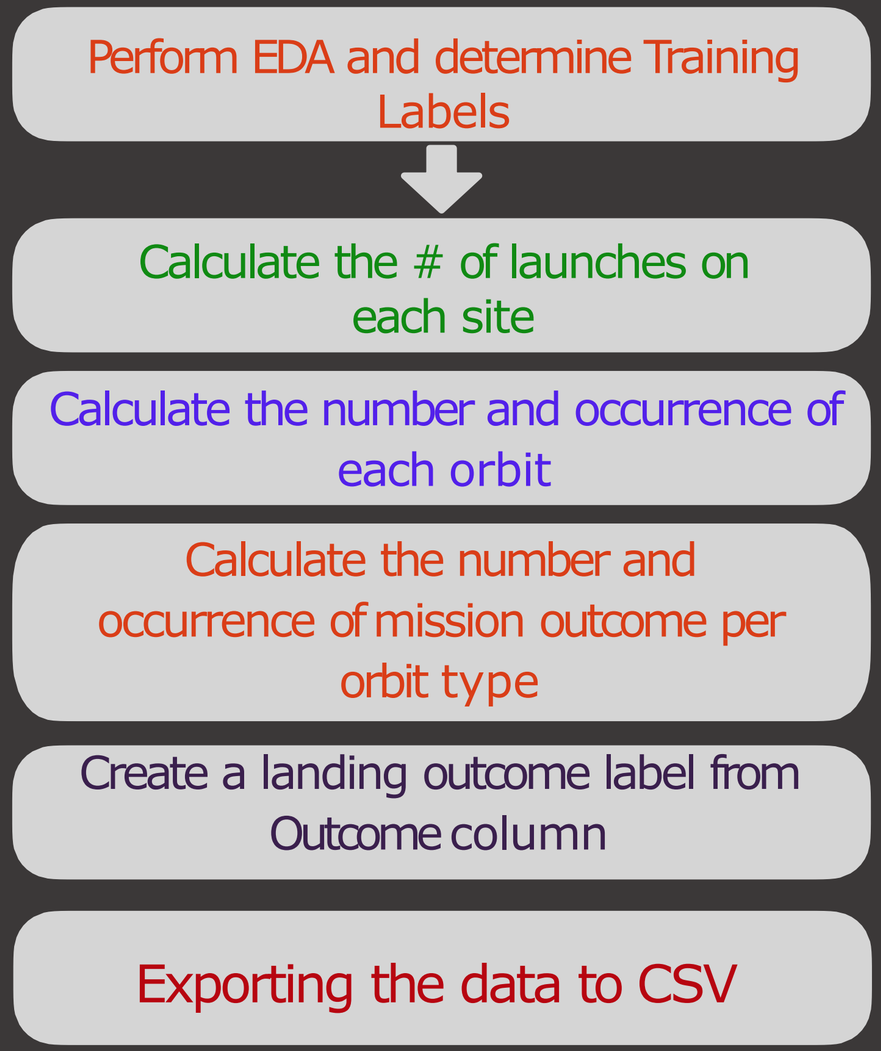


Data collection – Web scraping



Data wrangling

The dataset contains various instances of the booster failing to land correctly. Some landings were attempted but resulted in accidents, such as "True Ocean," indicating a successful landing in a specific area of the ocean, and "False Ocean," indicating an unsuccessful landing in the ocean. Similarly, "True RTLS" denotes a successful landing on a ground pad, while "False RTLS" indicates an unsuccessful landing on a ground pad. "True ASDS" indicates a successful landing on a drone ship, while "False ASDS" indicates an unsuccessful landing on a drone ship. To create training labels from these outcomes, a "1" is assigned for a successful landing and a "0" for an unsuccessful one.



EDA with data visualization

Charts were plotted:

-Flight Number vs. Payload Mass,
-Flight Number vs. Launch Site,
-Payload Mass vs. Launch Site,
-Orbit Type vs. Success Rate,
-Flight Number vs. Orbit Type,
-Payload Mass vs Orbit Type and
-Success Rate Yearly Trend

Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model. Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value. Line charts show trends in data over time (time series).

EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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

Build an interactive map with Folium

Markers of all Launch Sites:

The task involved adding markers with circles, popup labels, and text labels to display the NASA Johnson Space Center's location based on its latitude and longitude coordinates. Additionally, markers with circles, popup labels, and text labels were added to all launch sites to indicate their geographical positions and proximity to the Equator and coasts.

Coloured Markers of the launch outcomes for each Launch Site:

The statement is referring to the addition of colored markers, specifically green for successful launches and red for failed launches, using a Marker Cluster. This addition is intended to help identify which launch sites have a relatively higher success rate.

Distances between a Launch Site to its proximities:

To illustrate the distance between the launch site KSC LC-39A and nearby locations such as the railway, highway, coastline, and closest city, colored lines have been included.

Build a Dashboard with Plotly Dash

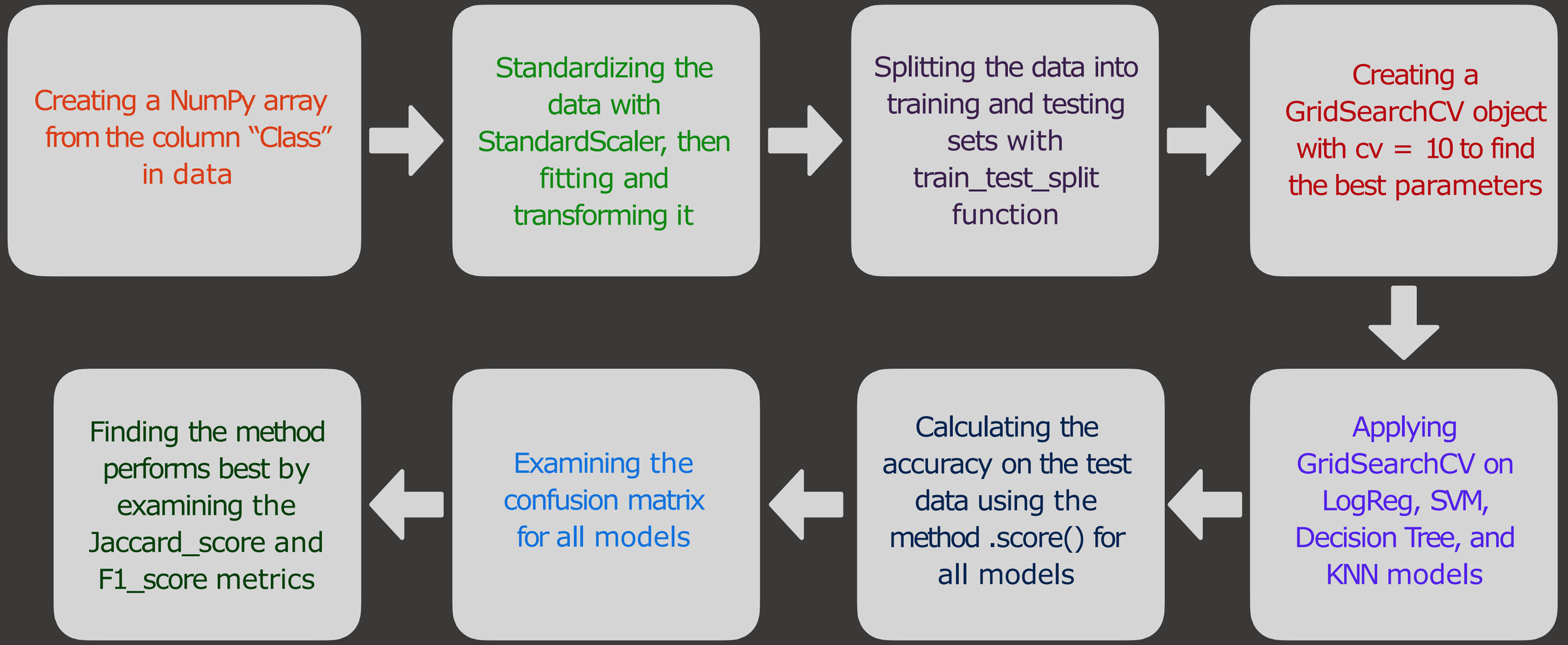
Launch Sites Dropdown List: Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site): Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range: Added a slider to select Payload range

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions: Added a scatter chart to show the correlation between Payload and Launch Success.

Predictive analysis (Classification)

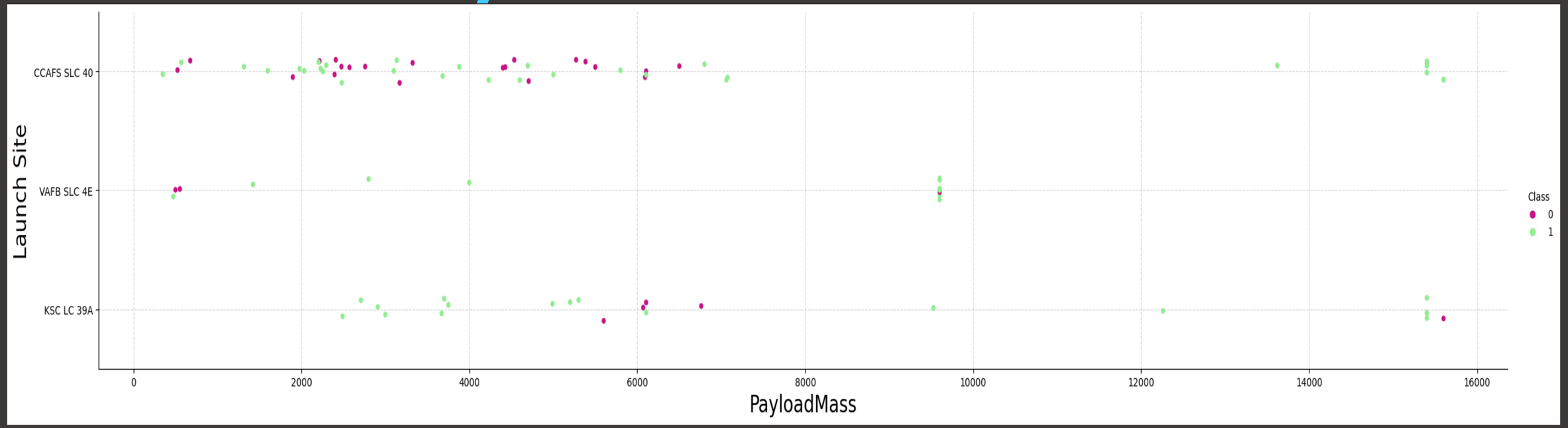


Results

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

EDA with Visualization

Payload vs. Launch Site

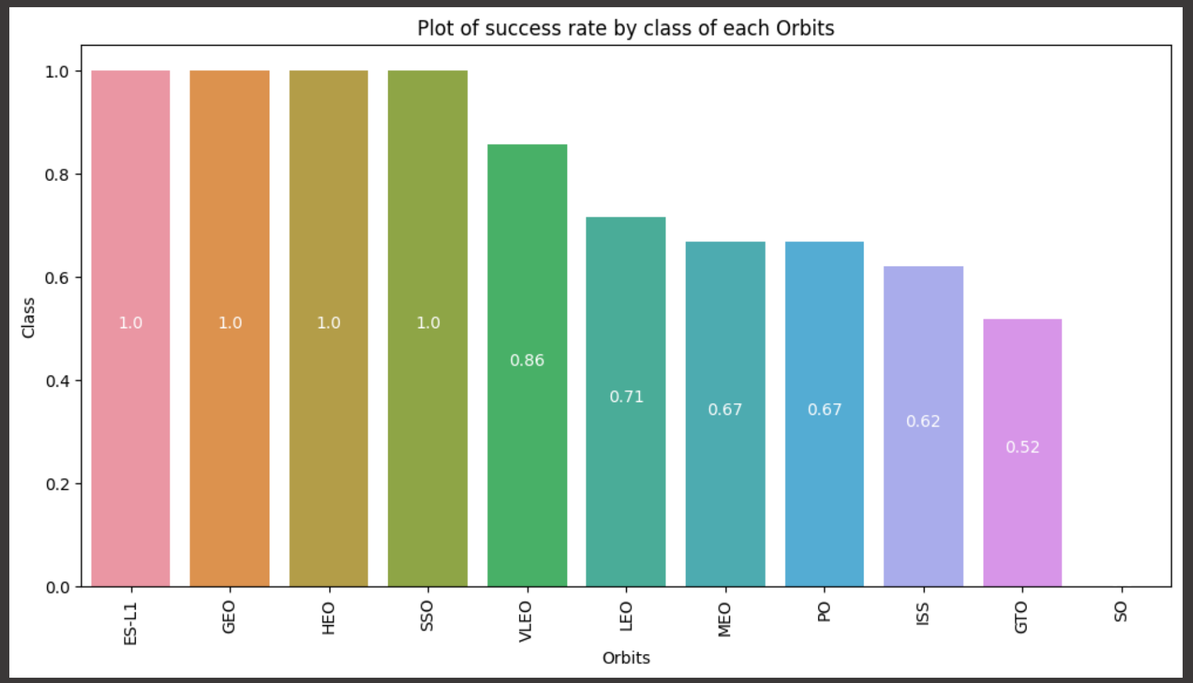


Explanation: The success rate of a launch site increases with the payload mass, and launches with a payload mass greater than 7000 kg were mostly successful. Additionally, KSC LC 39A has a 100% success rate for payload masses under 5500 kg.

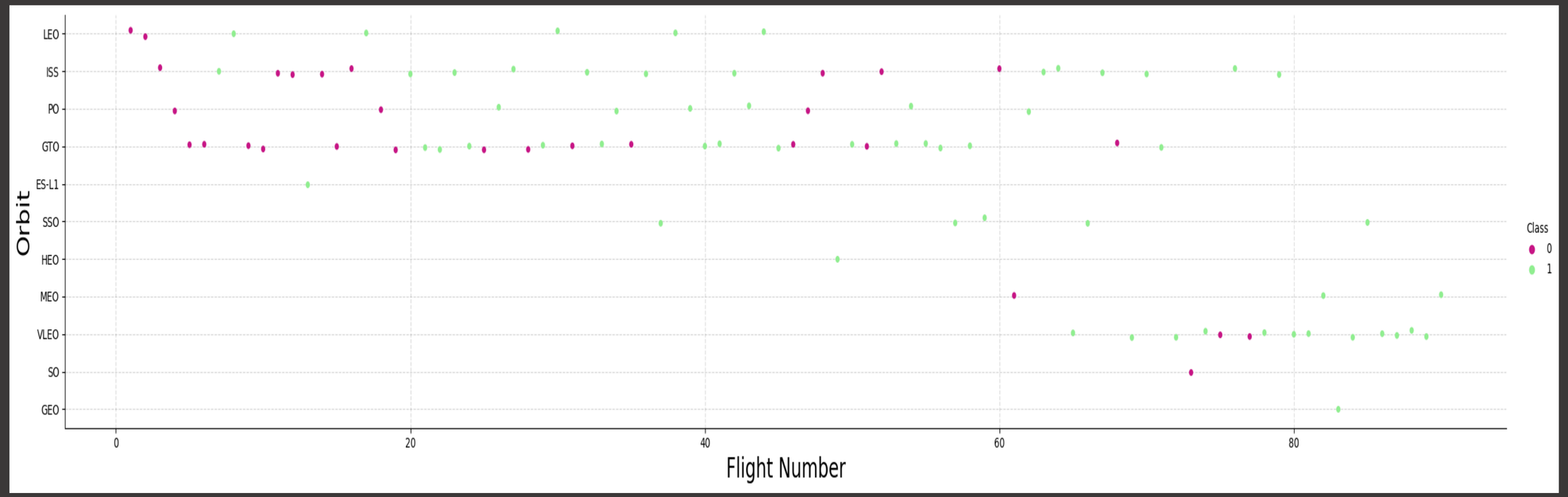
Success rate vs. Orbit type

Explanation:

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

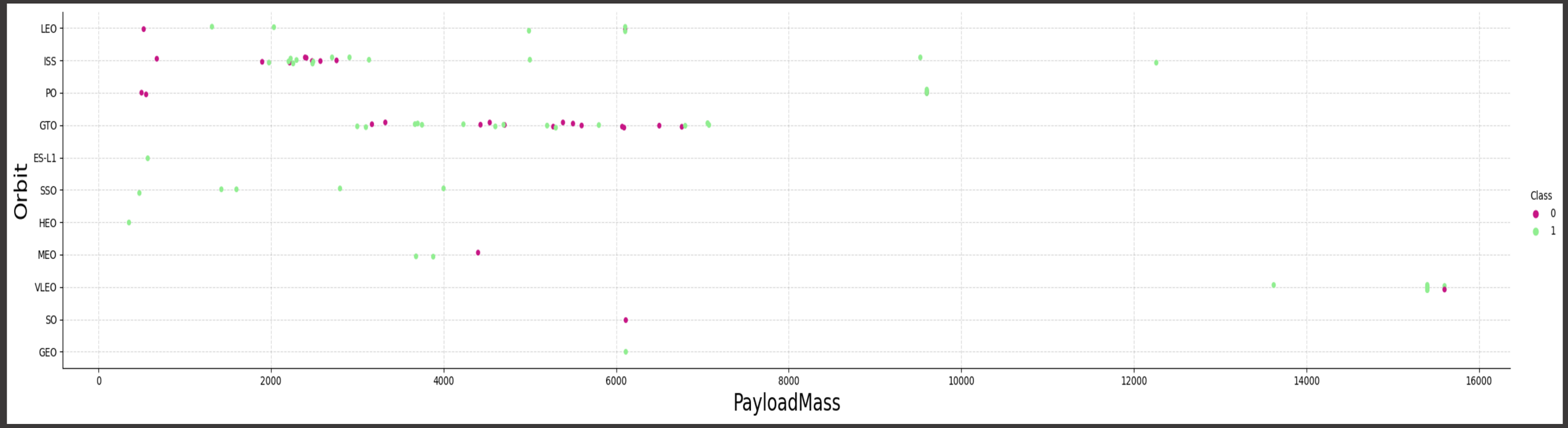


Flight Number vs. Orbit type



Explanation: The number of successful flights in a Low Earth Orbit (LEO) appears to be dependent on the number of flights conducted, whereas there is no correlation between the number of flights and the rate of success in a Geostationary Transfer Orbit (GTO).

Payload Mass vs. Orbit type

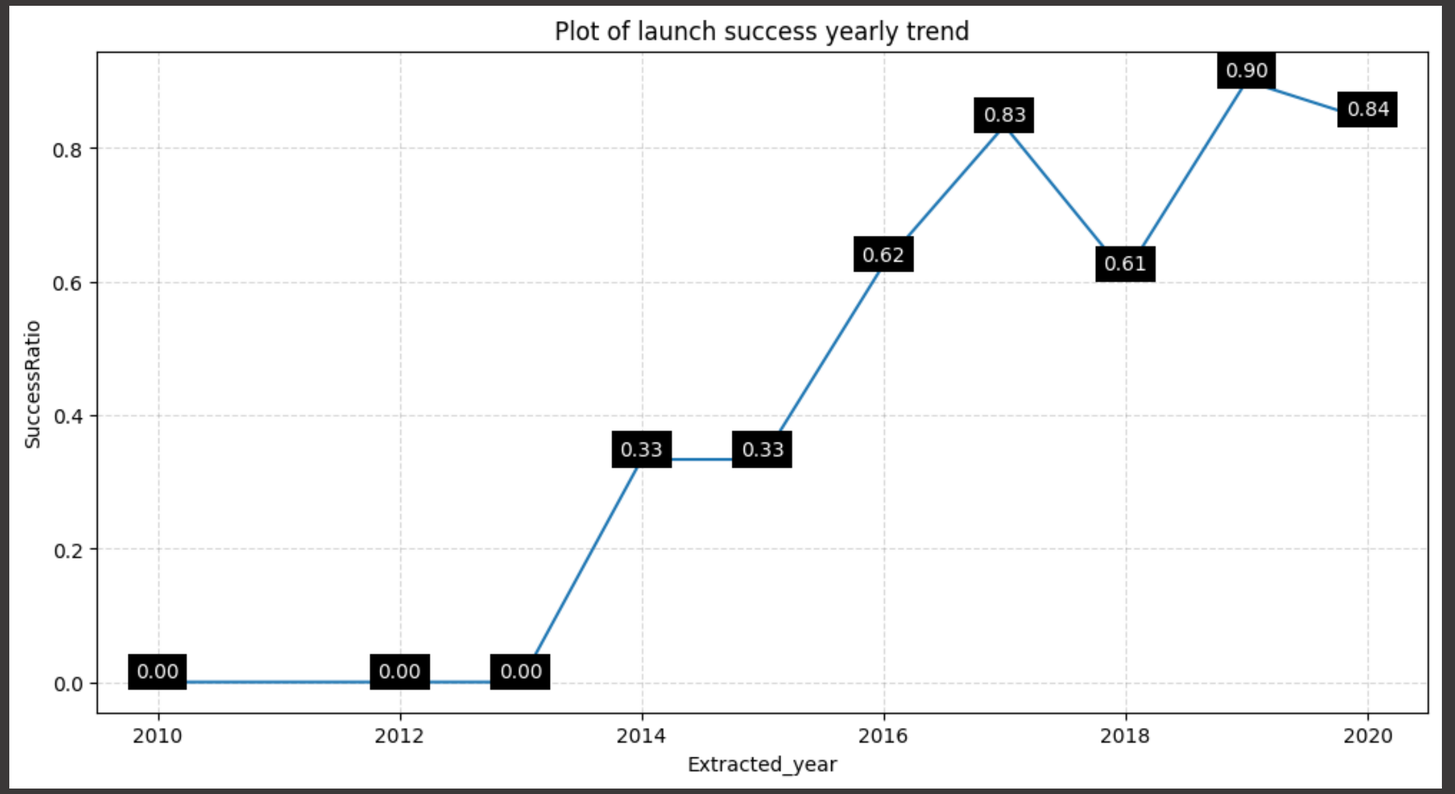


Explanation: Large payloads have a detrimental effect on geostationary transfer orbit (GTO) trajectories but a beneficial effect on GTO and polar low Earth orbit (LEO) trajectories, such as those of the International Space Station (ISS).

Launch success yearly trend

Explanation:

The success rate since 2013 kept increasing till 2020.



EDA with SQL

All launch site names

Task 1

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

```
task_1 = '''
        SELECT DISTINCT LaunchSite
        FROM SpaceX
'''
create_pandas_df(task_1, database=conn)
```

	LaunchSite
0	CCAFS LC-40
1	VAFB SLC-4E
2	KSC LC-39A
3	CCAFS SLC-40

Explanation: Displaying the names of the unique launch sites in the space mission.

Launch site names begin with `CCA`

Task 2

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

```
[ ] task_2 = '''
        SELECT strftime('%Y-%m-%d', substr(Date, 7, 4) || '-' || substr(Date, 1, 2) || '-' || substr(Date, 4, 2)) as Date,Time,BoosterVersion,LaunchSite,Payload,PayloadMassKG,Orbit,
        Customer,MissionOutcome,LandingOutcome
        FROM SpaceX
        WHERE LaunchSite LIKE 'CCA%'
        LIMIT 5
        '''
    create_pandas_df(task_2, database=conn)
```

	Date	Time	BoosterVersion	LaunchSite	Payload	PayloadMassKG	Orbit	Customer	MissionOutcome	LandingOutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	None	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	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: Displaying 5 records where launch sites begin with the string 'CCA'.

Total payload mass

Explanation: Displaying the total payload mass carried by boosters launched by NASA (CRS).

Task 3

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

```
[ ] task_3 = '''
            SELECT SUM(PayloadMassKG) AS Total_PayloadMass
            FROM SpaceX
            WHERE Customer LIKE 'NASA (CRS)'
            '''
    create_pandas_df(task_3, database=conn)
```

	Total_PayloadMass
0	45596

Average payload mass by F9 v1.1

Explanation: Displaying average payload mass carried by booster version F9 v1.1.

Task 4

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

```
task_4 = '''
        SELECT AVG(PayloadMassKG) AS Avg_PayloadMass
        FROM SpaceX
        WHERE BoosterVersion = 'F9 v1.1'
        '''
create_pandas_df(task_4, database=conn)
```

	Avg_PayloadMass
0	2928.4

Successful drone ship landing with payload between 4000 and 6000

Explanation: Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

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

```
task_6 = '''
        SELECT BoosterVersion
        FROM SpaceX
        WHERE LandingOutcome = 'Success (drone ship)'
            AND PayloadMassKG > 4000
            AND PayloadMassKG < 6000
        '''
create_pandas_df(task_6, database=conn)
```

	BoosterVersion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total number of successful and failure mission outcomes

Task 7

List the total number of successful and failure mission outcomes

```
task_7a = ''' SELECT COUNT(MissionOutcome) AS SuccessOutcome FROM SpaceX WHERE MissionOutcome LIKE 'Success%'   '''
task_7b = ''' SELECT COUNT(MissionOutcome) AS FailureOutcome FROM SpaceX WHERE MissionOutcome LIKE 'Failure%'   '''
print('The total number of successful mission outcome is:')
display(create_pandas_df(task_7a, database=conn))
print()
print('The total number of failed mission outcome is:')
create_pandas_df(task_7b, database=conn)
```

```
The total number of successful mission outcome is:
```

	SuccessOutcome
0	100

```
The total number of failed mission outcome is:
```

	FailureOutcome
0	1

Explanation: Listing the total number of successful and failure mission outcomes.

Boosters carried maximum payload

Task 8

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

```
task_8 = '''SELECT BoosterVersion, PayloadMassKG FROM SpaceX WHERE PayloadMassKG=(SELECT MAX(PayloadMassKG) FROM SpaceX) ORDER BY BoosterVersion '''
create_pandas_df(task_8, database=conn)
```

	BoosterVersion	PayloadMassKG
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

Explanation: Listing the names of the booster versions which have carried the maximum payload mass.

Rank success count between 2010-06-04 and 2017-03-20

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

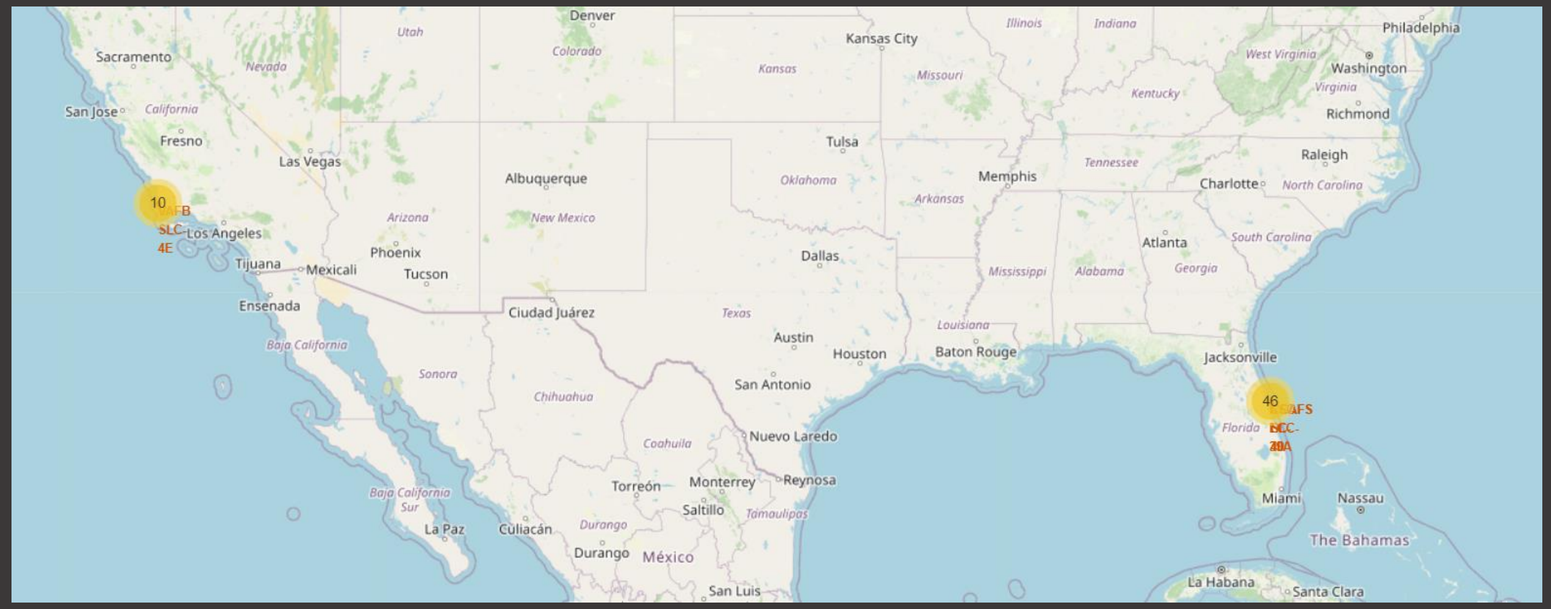
```
[ ] task_10 = '''
        SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE strftime('%Y-%m-%d', substr(Date, 7, 4) || '-' || substr(Date, 1, 2) || '-' || substr(Date, 4, 2)) BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
        '''
    create_pandas_df(task_10, database=conn)
```

	LandingOutcome	COUNT(LandingOutcome)
0	No attempt	7
1	Success (ground pad)	2
2	Success (drone ship)	2
3	Failure (drone ship)	2
4	Failure (parachute)	1
5	Controlled (ocean)	1

Explanation: Reorder the frequency of landing results (such as unsuccessful landing on a drone ship or successful landing on a ground pad) from June 4th, 2010 to March 20th, 2017, starting with the most frequent outcome and ending with the least frequent outcome.

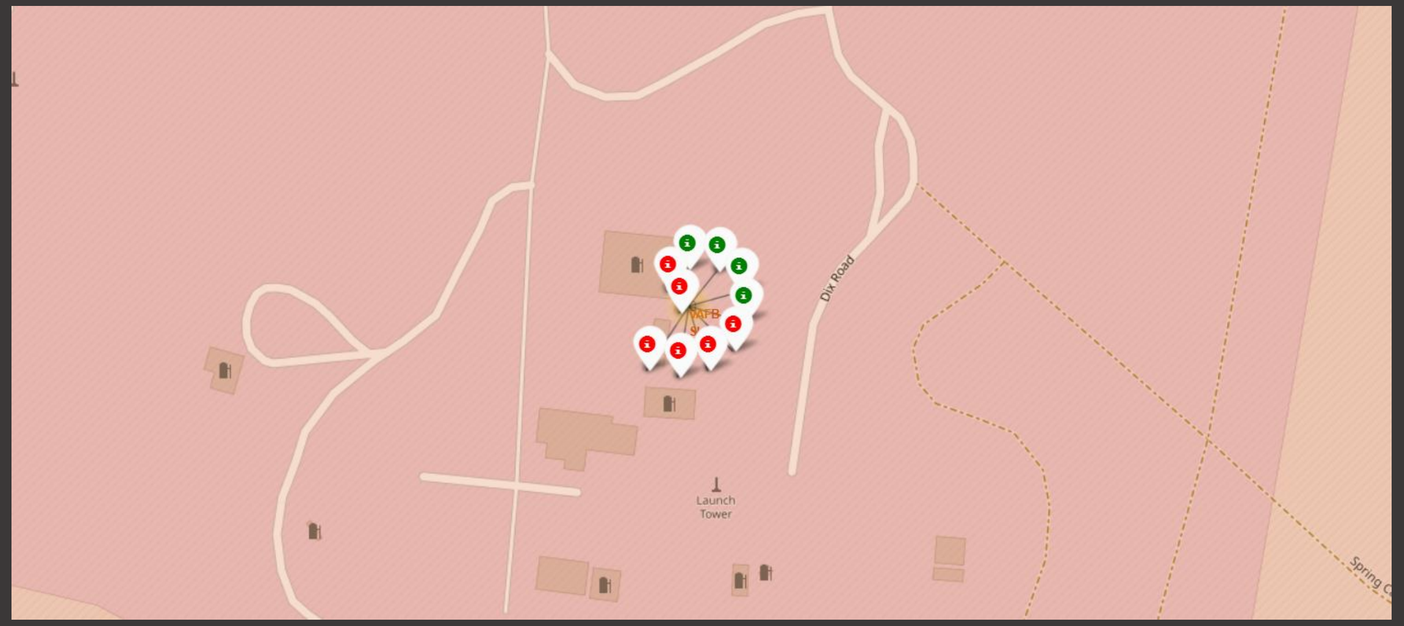
Interactive map with Folium

All launch sites' location markers on a global map



Explanation: 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. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit. All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.

Colour-labeled launch records on the map



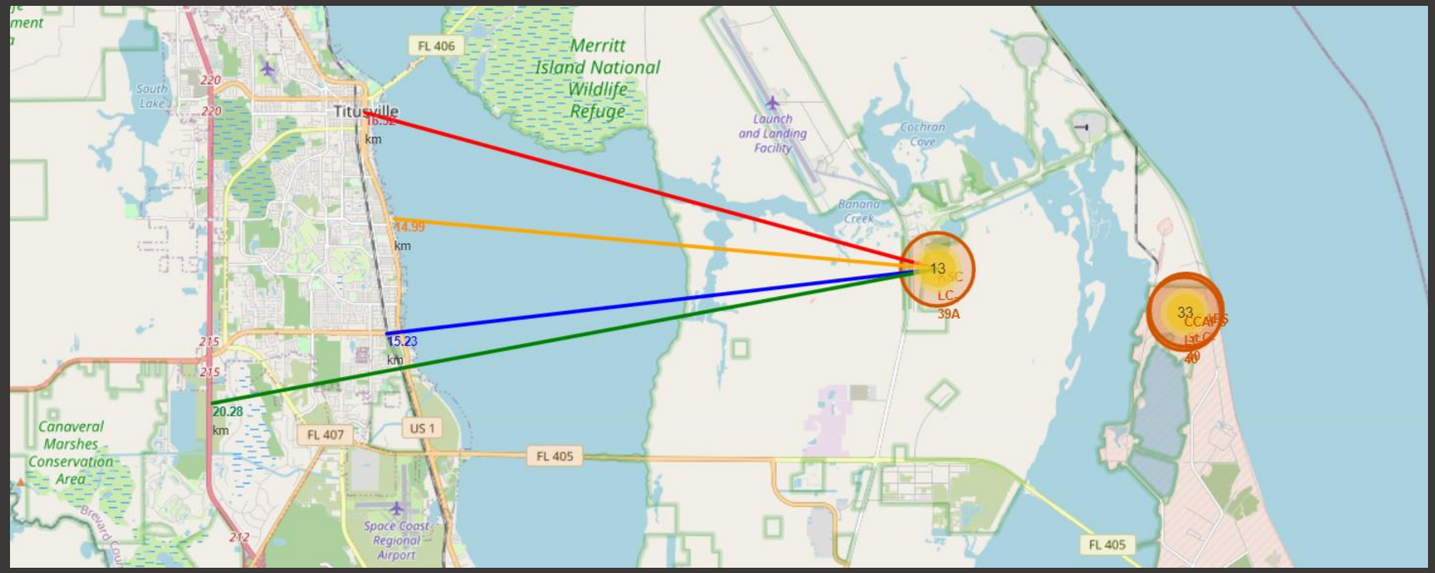
Explanation:

- From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

  Green Marker = Successful Launch Red Marker = Failed Launch

- Launch Site KSC LC-39A has a very high Success Rate.

Distance from the launch site KSC LC-39A to its proximities



Explanation: From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:

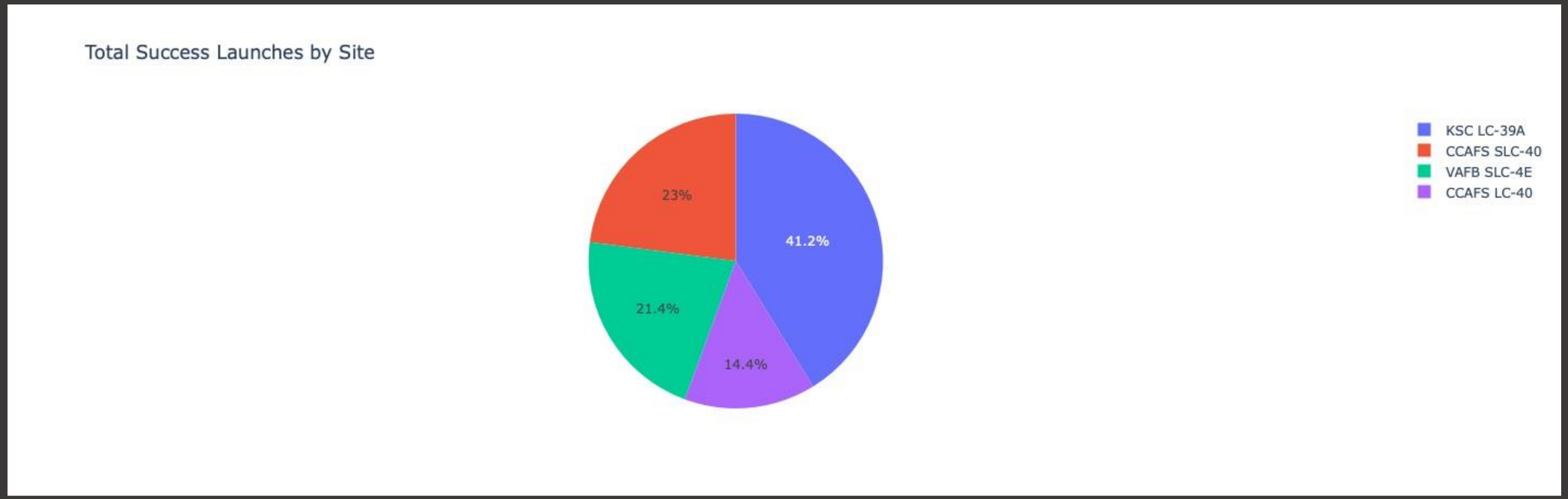
-Relative close to railway (15.23 km)

-Relative close to highway (20.28 km)

-Relative close to coastline (14.99 km)

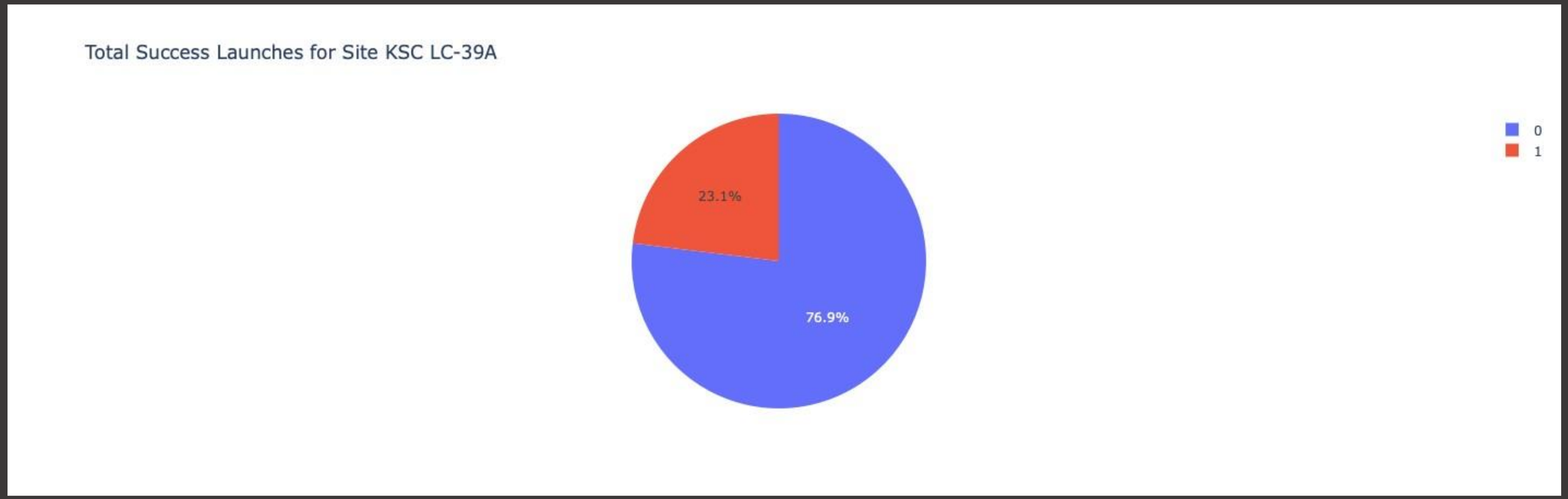
Build a Dashboard with Plotly Dash

Launch success count for all sites



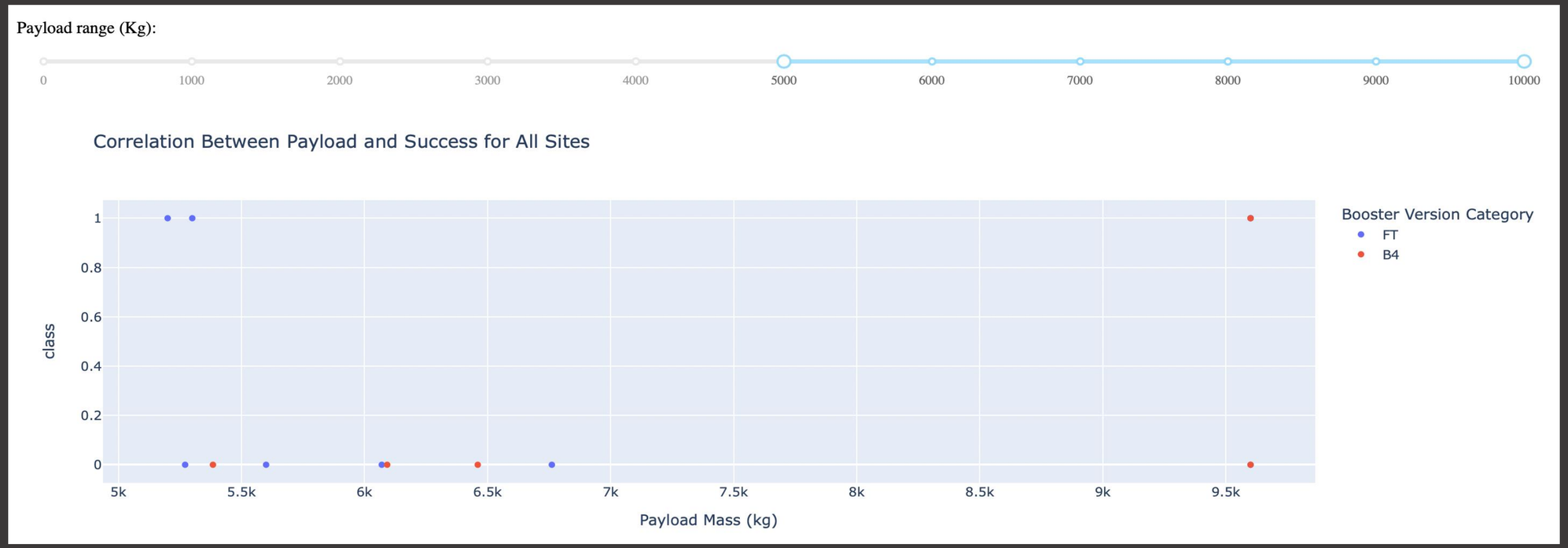
Explanation: The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

Launch site with highest launch success ratio



Explanation: KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

Payload Mass vs. Launch Outcome for all sites



Explanation: The charts show that payloads between 2000 and 5500 kg have the highest success rate.

Predictive analysis (Classification)

Classification Accuracy

Explanation: According to the Test Set results, it is not possible to determine which method performs the best, as the scores may be influenced by the limited sample size of 18. Therefore, to further evaluate all methods, the entire Dataset was tested. The overall Dataset scores indicate that the Decision Tree Model is the most effective, as it has higher scores and the highest accuracy compared to the other models.

Scores and Accuracy of the Test Set

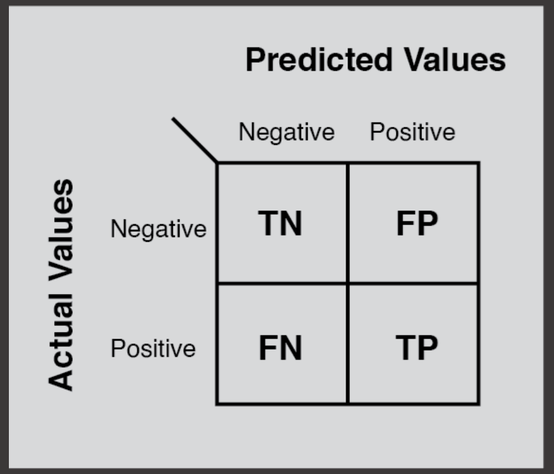
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

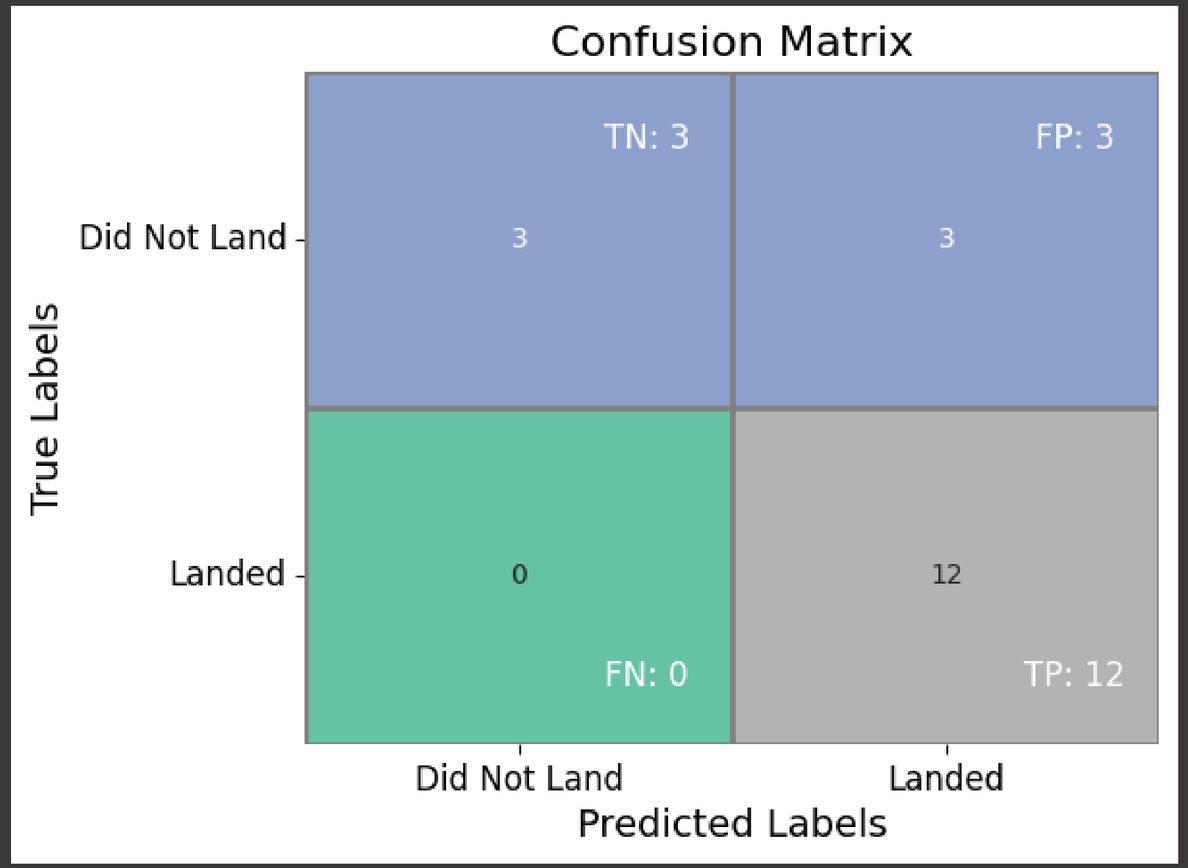
Scores and Accuracy of the Entire Data Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Confusion Matrix

Explanation: Examining confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.





Conclusion

- The optimal algorithm for this dataset is the Decision Tree Model. It has been observed that launches carrying a small payload mass have a higher success rate compared to those carrying a larger payload mass. Additionally, most of the launch sites are located near the Equator and close to the coast. The success rate of launches has been observed to increase over time. KSC LC-39A has the highest success rate among all launch sites. Furthermore, all launches that have orbited ES-L1, GEO, HEO, and SSO have had a 100% success rate.