IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Malek FRIH
1/10/2024

Outline

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

Executive Summary

- **The following methodologies were used to analyze the data :**
  - Data collection using Web Scraping and SpaceX API
  - Explanatory Data Analysis (EDA)
  - Data visualization and Interactive visual analytics
  - Machine Learning predictive analysis (Classification)
- **Summary of all results :**
  - Explanatory Data Analysis results
  - Interactive Analytics dashboard
  - Predictive analysis results

Introduction

- **Project background and context :**

  SpaceX is the most successful company of the commercial space age, making space travel affordable. The company 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. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

- **Problems you want to find answers :**
  - Estimation of the total cost for launches by predicting successfulness of landings in the first stage for rockets.
  - Defining the best Launching sites for rockets.

Section 1

Methodology

Methodology

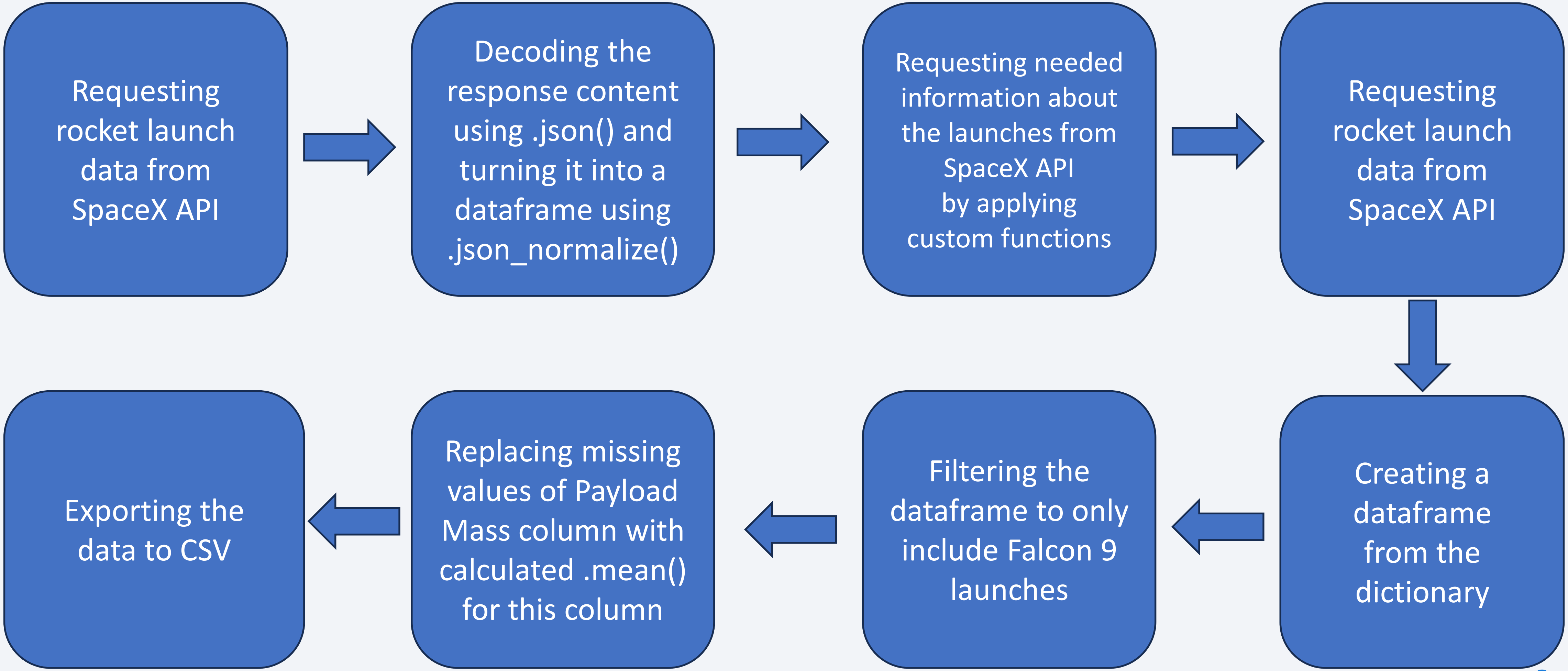
Executive Summary

1. Data collection methodology
    - Using SpaceX Rest API
    - Using Web Scrapping from Wikipedia
2. Performed data wrangling
    - Filtering the data
    - Dealing with missing values
    - Using One Hot Encoding to prepare the data to a binary classification
3. Performed exploratory data analysis (EDA) using visualization and SQL
4. Performed interactive visual analytics using Folium and Plotly Dash
5. Performed predictive analysis using classification models
    - Building, tuning and evaluation of classification models to ensure the best results

Data Collection

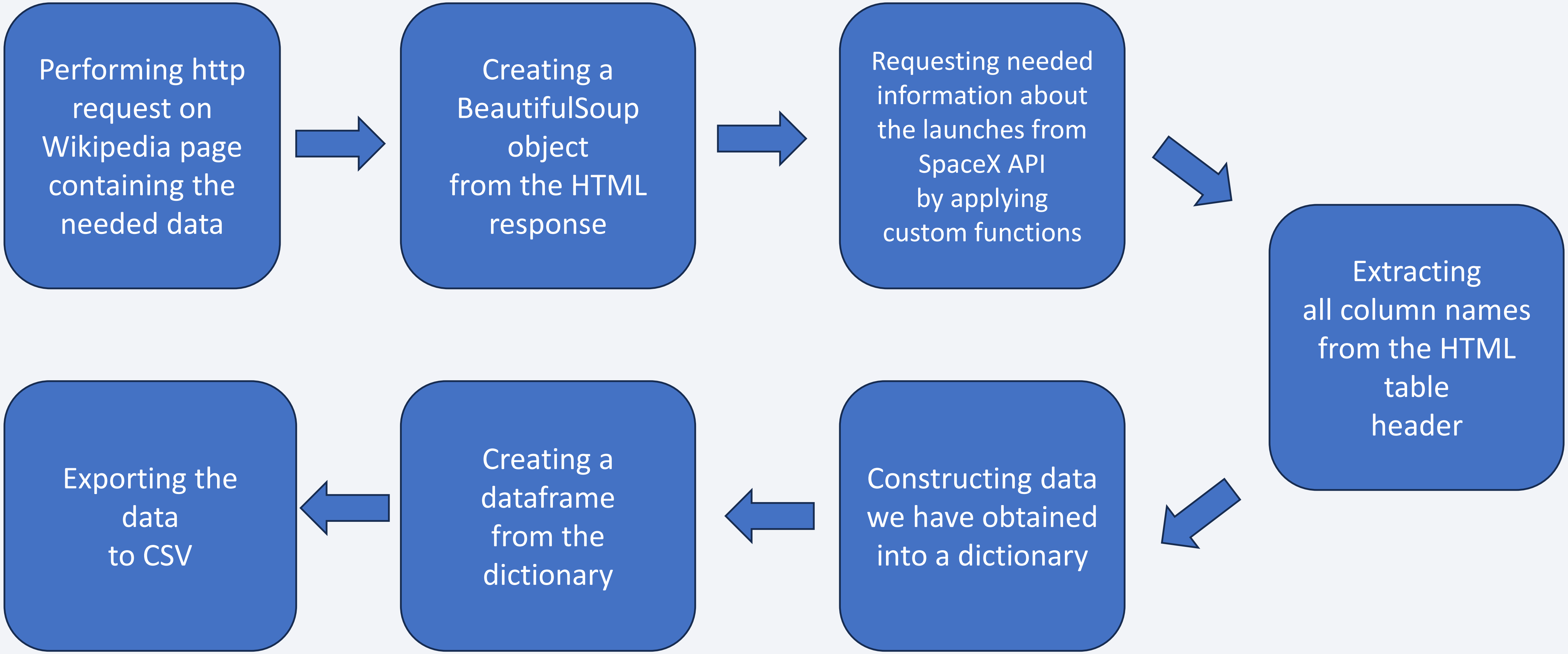
- Data collection process involved a combination of
  - API requests from SpaceX REST API (https://api.spacexdata.com/v4/rockets)
  - Web Scraping data from a table in SpaceX's Wikipedia entry. (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis

Data Collection – SpaceX API



GitHub Notebook URL: Data Collection with API CALLS

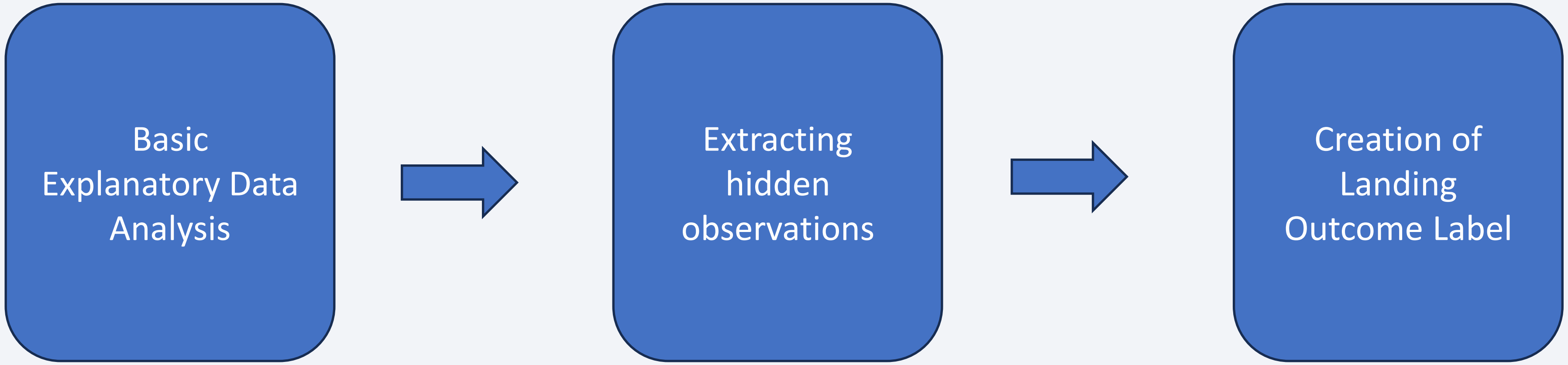
Data Collection – Web Scraping



[GitHub Notebook URL: Data Collection with Web Scraping]

Data Wrangling

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Extracting hidden observations like launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.



GitHub URL: Data Wrangling

EDA with Data Visualization

To explore data, scatterplots and bar plots were used to visualize the relationship between pair of features:

- Payload Mass vs. Flight Number
- Launch Site vs. Flight Number
- Launch Site vs Payload Mass
- Orbit and Flight Number
- Payload and Orbit

GitHub Notebook URL to view charts and code: Data Visualization

EDA with SQL

The following SQL queries were performed:

- Names of the unique launch sites in the space mission;
- Top 5 launch sites whose name begin with the string 'CCA’;
- Total payload mass carried by boosters launched by NASA (CRS);
- Average payload mass carried by booster version F9 v1.1;
- Date when the first successful landing outcome in ground pad was achieved;
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
- Total number of successful and failure mission outcomes;
- Names of the booster versions which have carried the maximum payload mass;
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

GitHub Notebook URL : EDA with SQL

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate points like launch sites
  - Circles indicate highlighted areas around specific coordinates like NASA Johnson Space Center
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site
  - Lines are used to indicate distances between two coordinates.

[GitHub Notebook URL : Interactive Visual Analytics with Folium]

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:

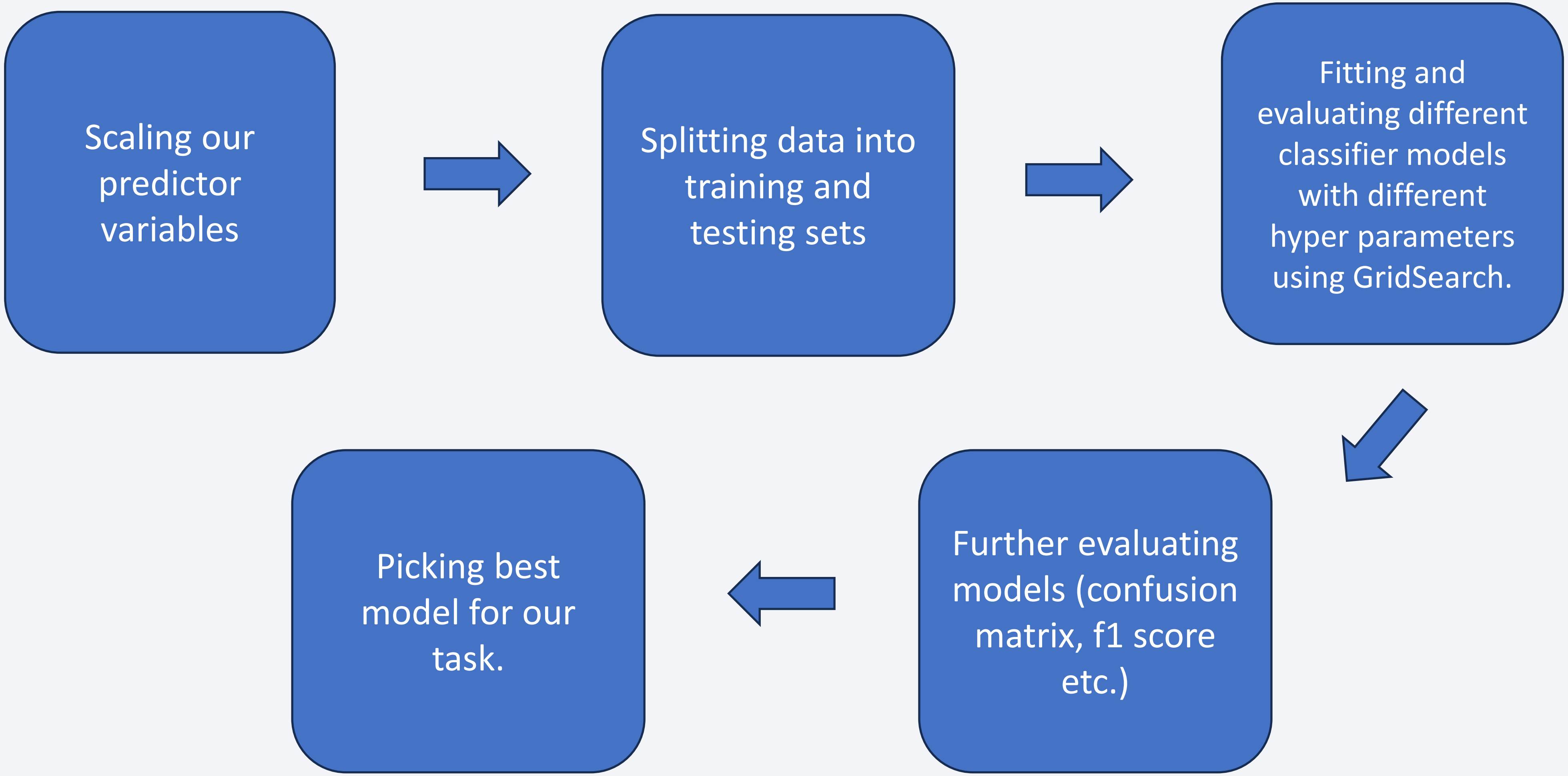
- Added a slider to select Payload range.

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

GitHub Notebook URL : Interactive Visual Analytics with Dash

Predictive Analysis (Classification)



GitHub Notebook URL : Predictive modeling

Results

- Exploratory data analysis results:
    - space X uses 4 different launch sites;
    - The first success landing outcome happened in 2015 fiver year after the first launch;
    - The number of landing outcomes became as better as years passed.

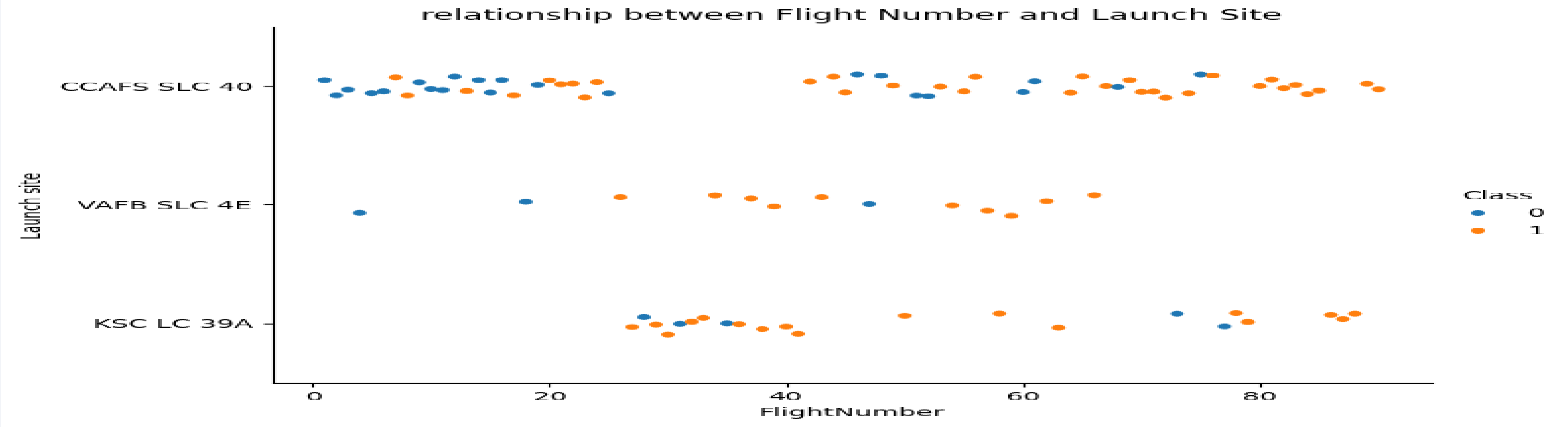
- Using interactive analytics, it was possible to identify that launch sites are located in safe places, near sea, for example and have a good logistic infrastructure around.
- Most launches happens at east cost launch sites.

- As for predictive modeling, while on test data, all models performed the same (partially due to the small size of the test data), on the whole data set, Support Vector Machine performed best with an accuracy of 0.87

Section 2

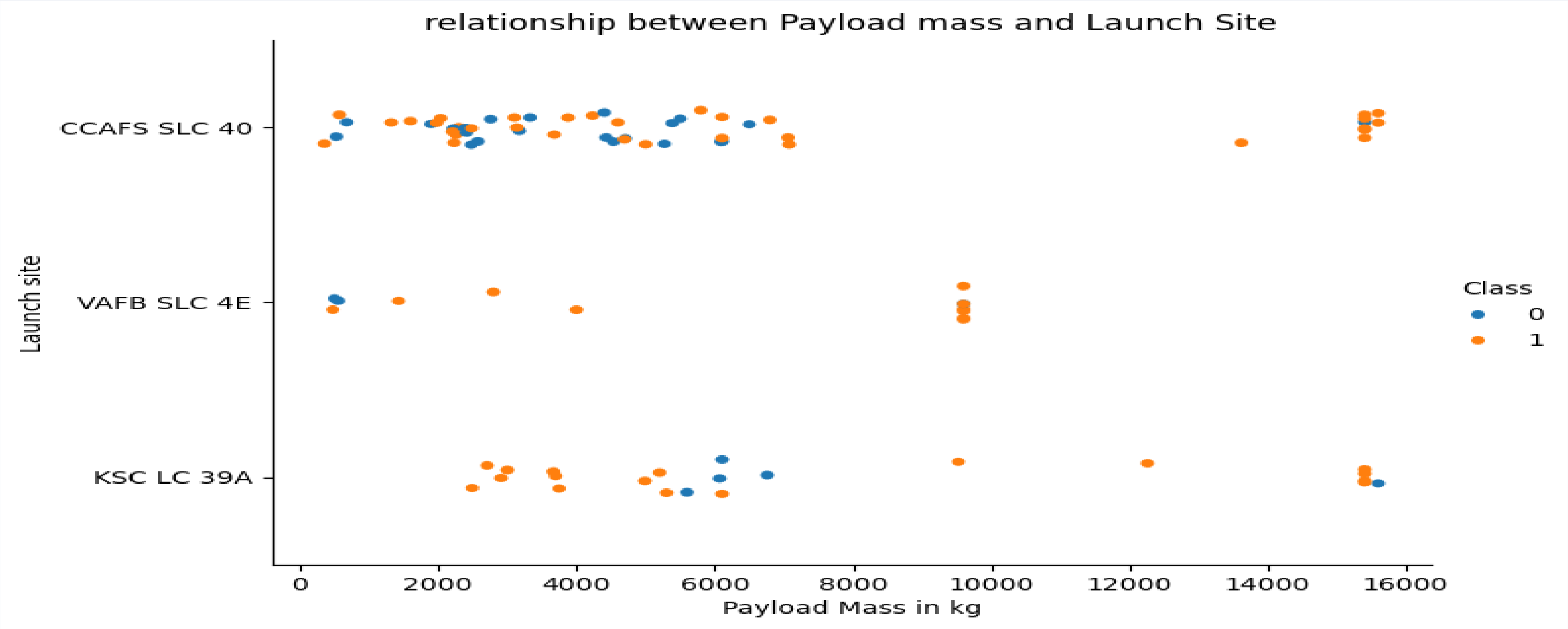
Insights drawn from EDA

Flight Number vs. Launch Site



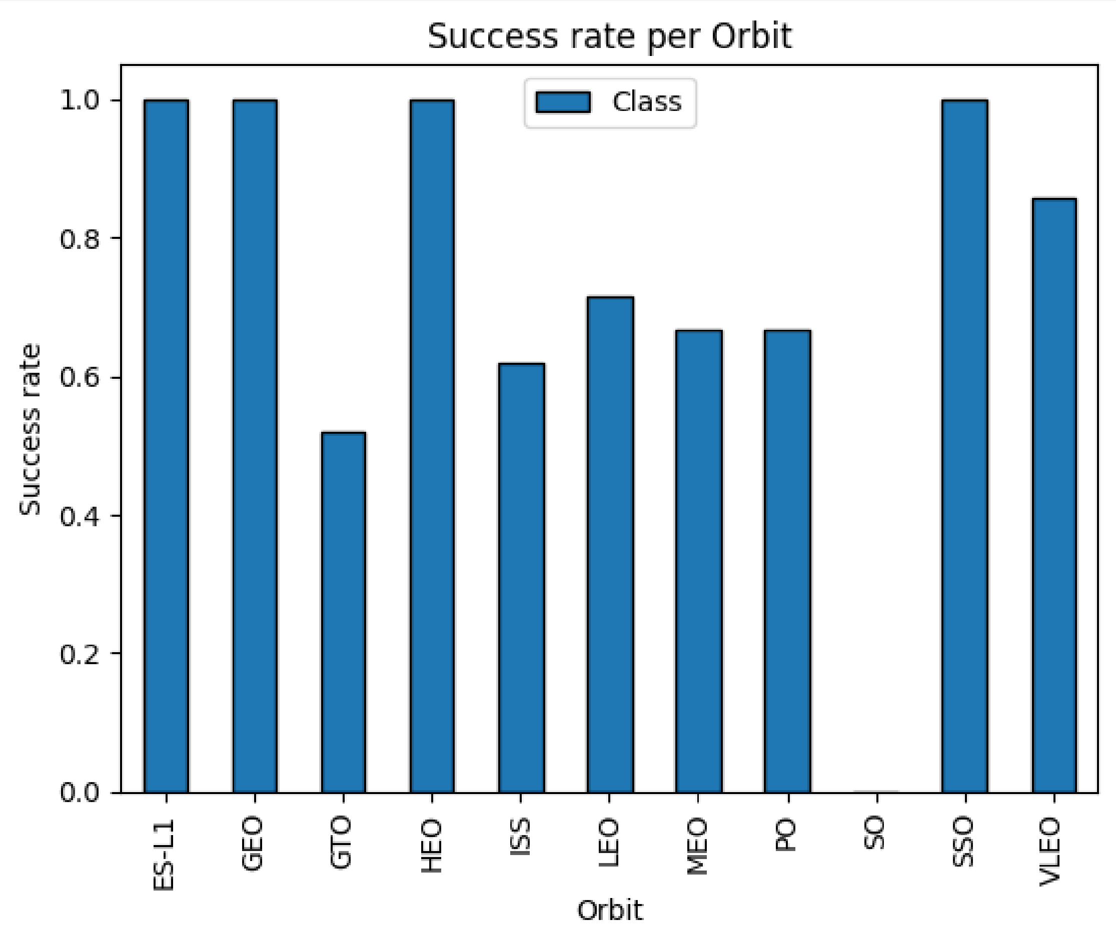
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- Success rate gets better with time,as SpaceX gets more experience with each flight.

Payload vs. Launch Site



- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

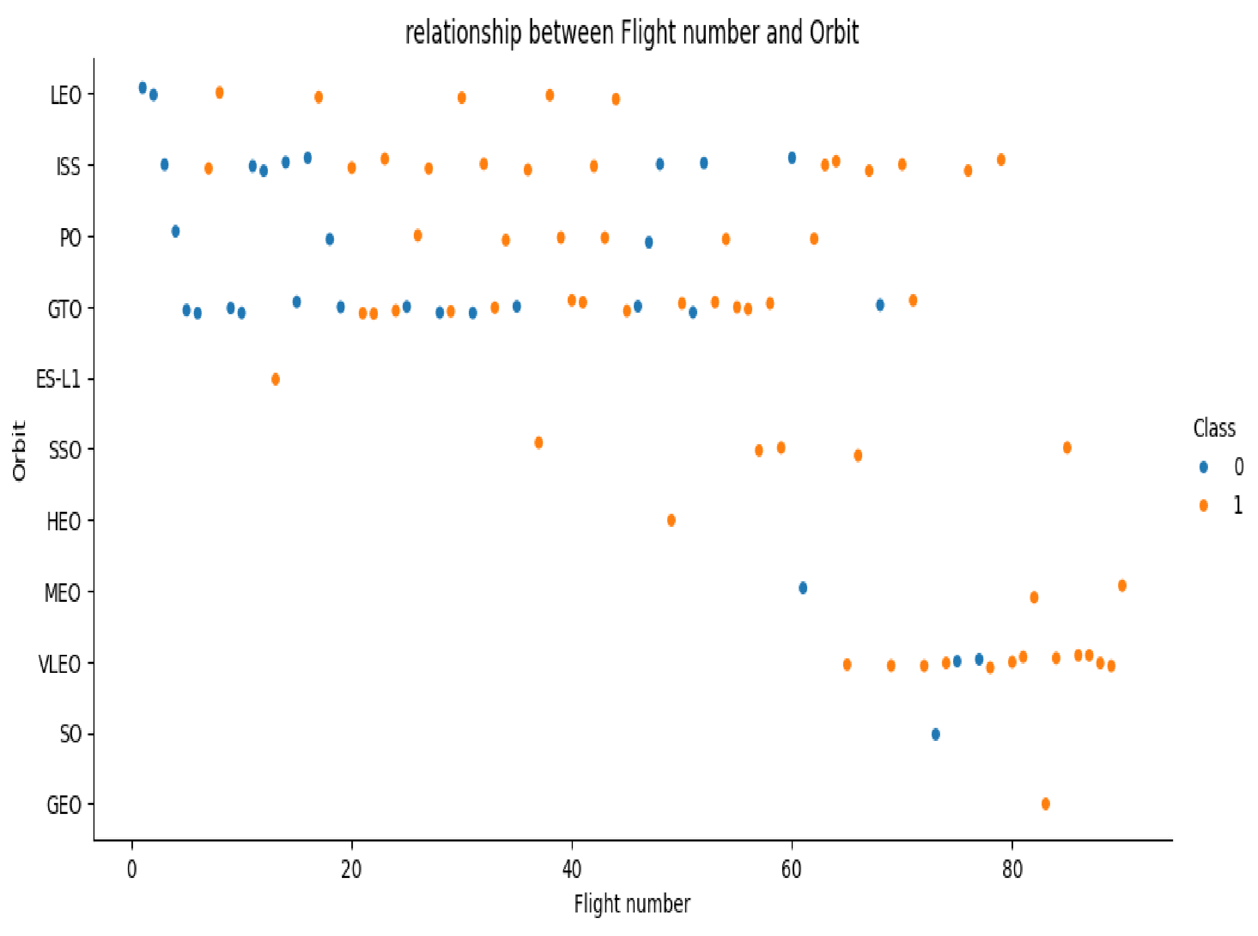
Success Rate vs. Orbit Type



Observations :

- 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

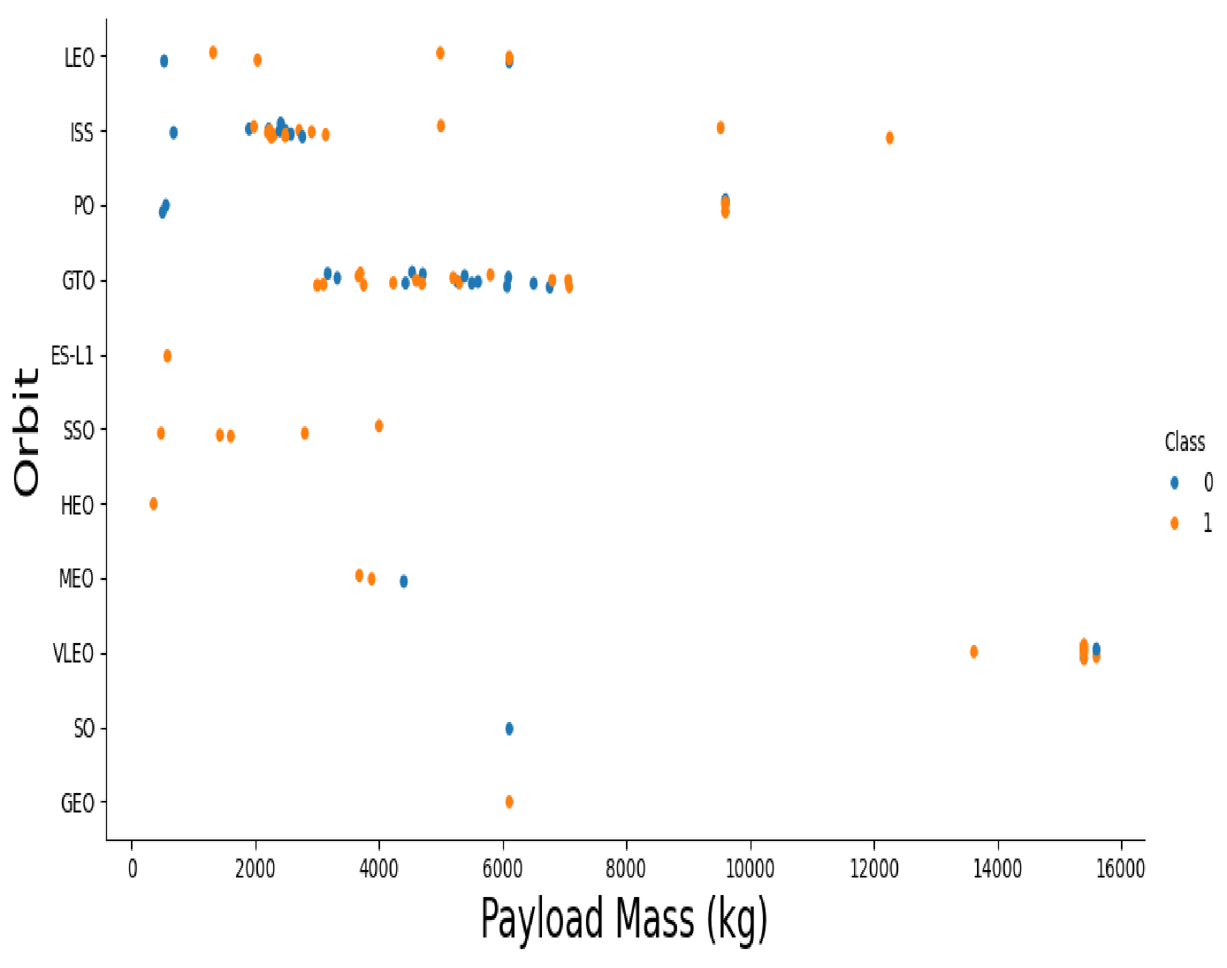
Flight Number vs. Orbit Type



Explanation:

- In the LEO orbit the Success appears related to the number of flights, on the other hand, that does not seem to be the case with the GTO orbit.

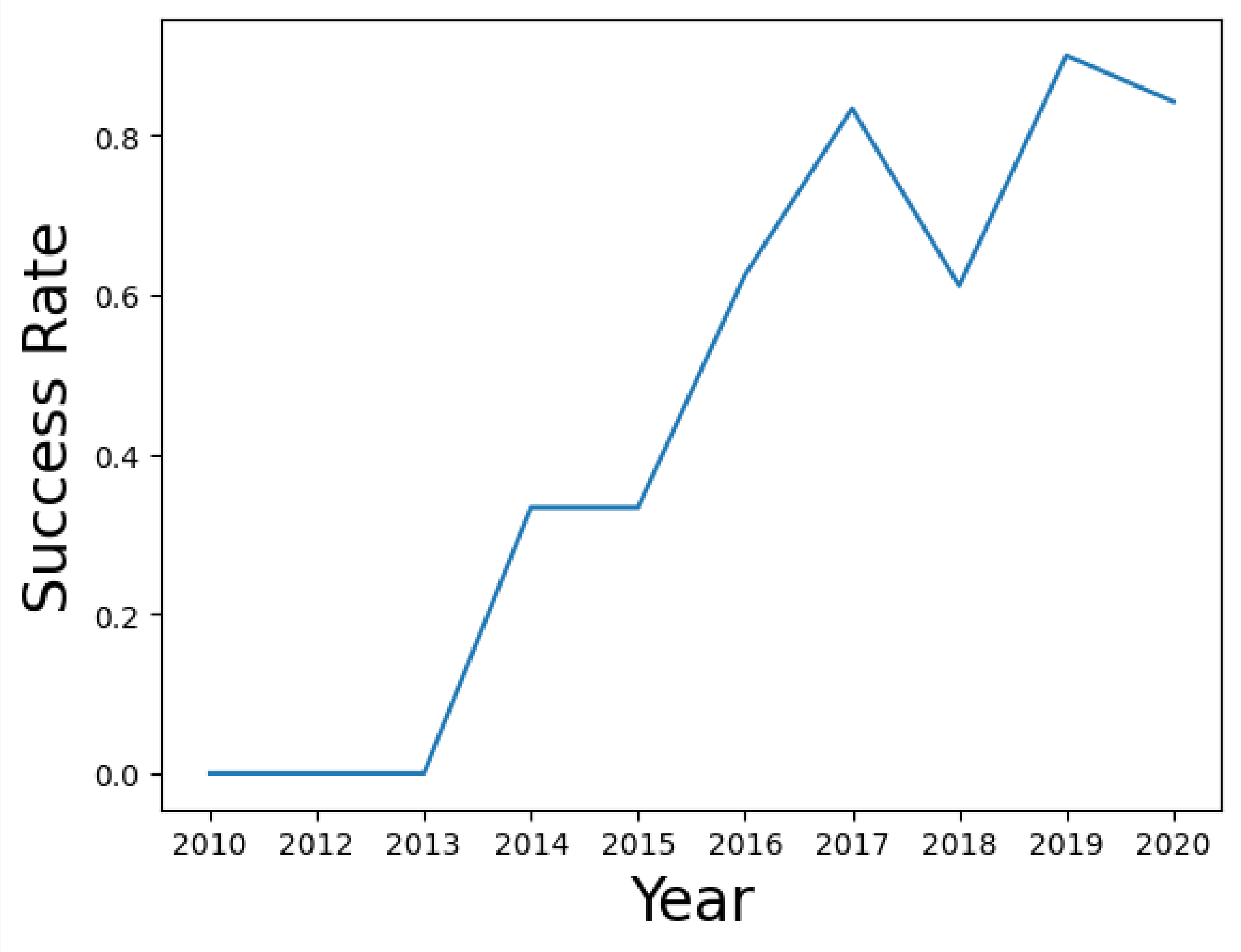
Payload vs. Orbit Type



Observation :

- Heavy payloads have a negative influence on success rate in GTO orbits and positive on LEO and Polar LEO (ISS) orbits

Launch Success Yearly Trend



Observation :

- The success rate since 2013 kept increasing till 2020

All Launch Site Names

Task 1

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

In [8]:
```
%%sql
select DISTINCT(Launch_Site) from SPACEXTABLE
```

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

Out[8]:

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

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

Launch Site Names Begin with 'CCA'

In [9]:
```
%%sql
select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
```

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

Out[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Displaying 5 records where launch sites begin with the string 'CCA'.

Total Payload Mass

```
In [10]: %%sql
         select SUM(PAYLOAD_MASS__KG_) as total from SPACEXTABLE where Customer is 'NASA (CRS)'

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

Out[10]:

total
45596

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

Average Payload Mass by F9 v1.1

```
In [24]: %%sql
select AVG(PAYLOAD_MASS__KG_) as average from SPACEXTABLE where Booster_Version like 'F9 v1.1%'

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

Out[24]:

average
2534.6666666666665

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

First Successful Ground Landing Date

```
In [12]: %%sql
         select MIN(Date) from SPACEXTABLE where Landing_Outcome like '%ground pad%'

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

Out[12]:

MIN(Date)
2015-12-22

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

Successful Drone Ship Landing with Payload between 4000 and 6000

In [13]:

```
%%sql
select Booster_Version from SPACEXTABLE where Landing_Outcome is 'Success (drone ship)' and
PAYLOAD_MASS__KG_ between 4000 and 6000
```

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

Out[13]:

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

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

Total Number of Successful and Failure Mission Outcomes

```
[20]: %%sql
select mission_outcome,count(*) from SPACEXTABLE group by mission_outcome;

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

[20]:

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

Listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

```
In [16]: %%sql
         select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ is (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)

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

Out[16]:

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

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

2015 Launch Records

In [23]:

```
%%sql
SELECT strftime('%Y', date) || '-' || strftime('%m', date) AS month,
       date,
       Booster_Version,
       launch_site,
       Landing_Outcome
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Failure (drone ship)' AND strftime('%Y', date) = '2015';
```

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

Out[23]:

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

Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015

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

In [21]:

```
%%sql
select Landing_Outcome,count(*) as counts from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome order by counts desc
```

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

Out[21]:

Landing_Outcome	counts
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

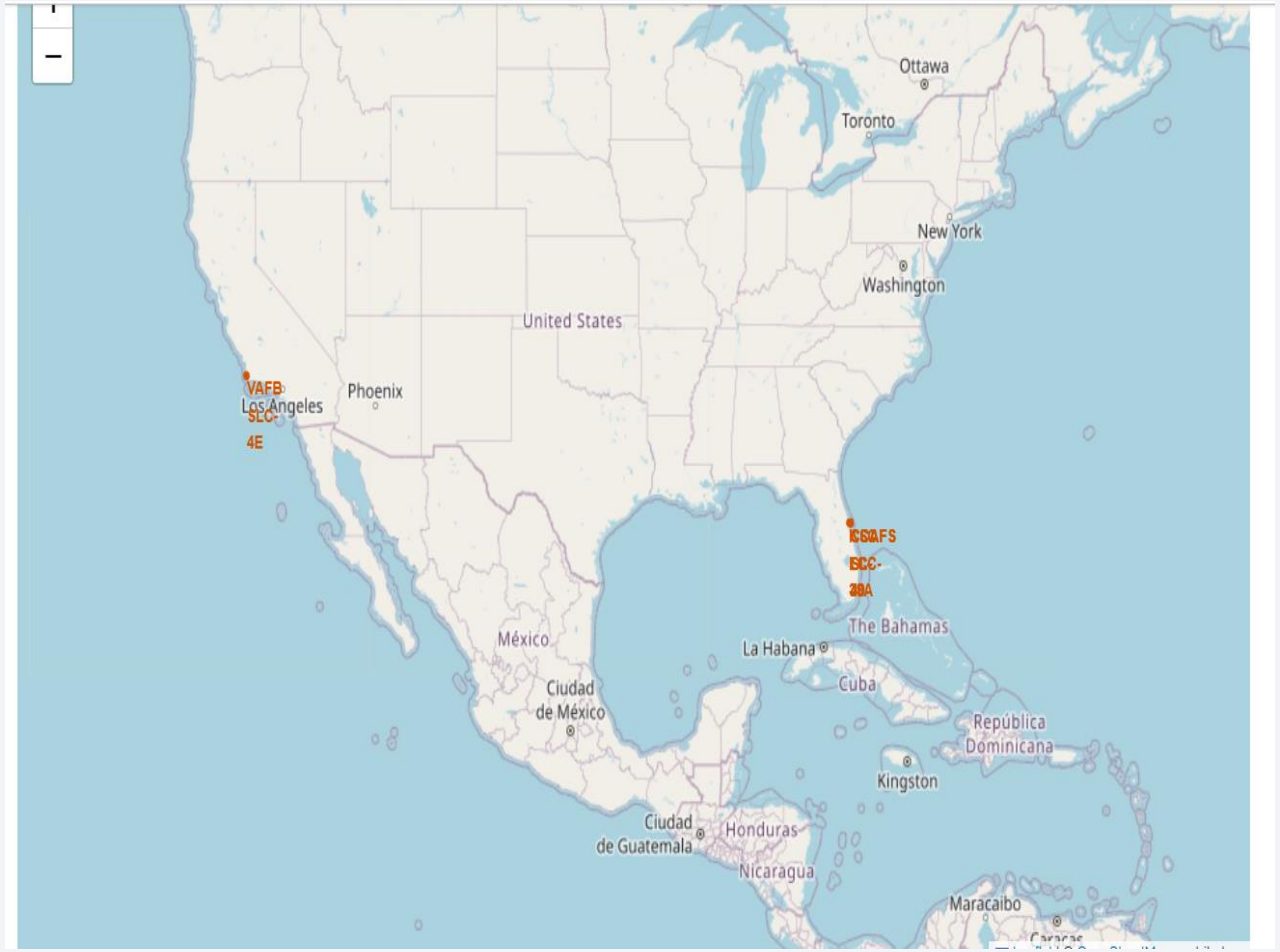
Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order

Section 3

Launch Sites Proximities Analysis

Launch Sites' Locations on Map

- 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.
- Also, most of Launch sites are in proximity to the Equator line where land moves the fastest, helping rockets maintain good enough speed to stay in orbit



Launch records on the map

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

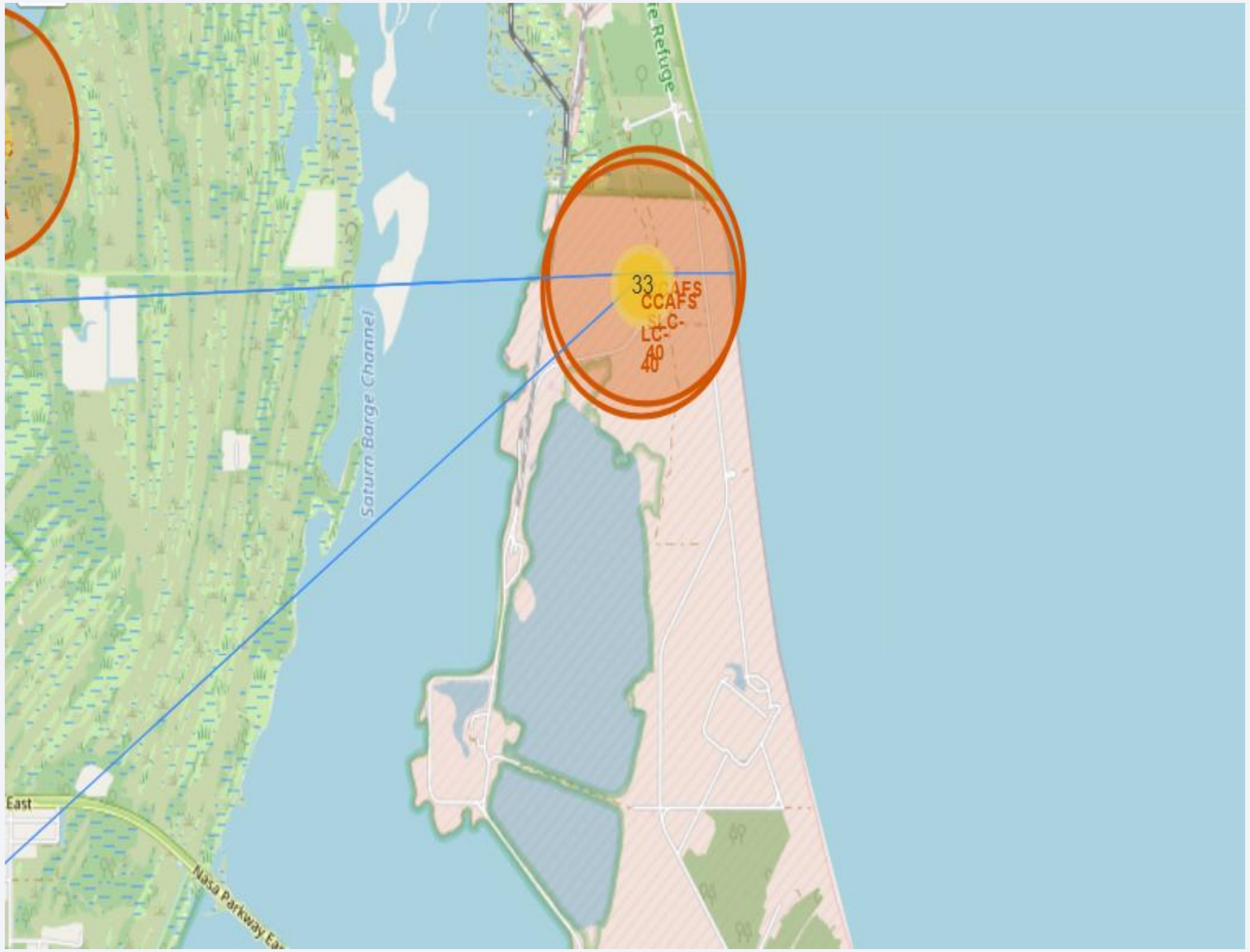
-Green Marker : Successful Launch

-Red Marker : Failed Launch

We can observe that Launch Site KSC LC-39A has a very high Success Rate.

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

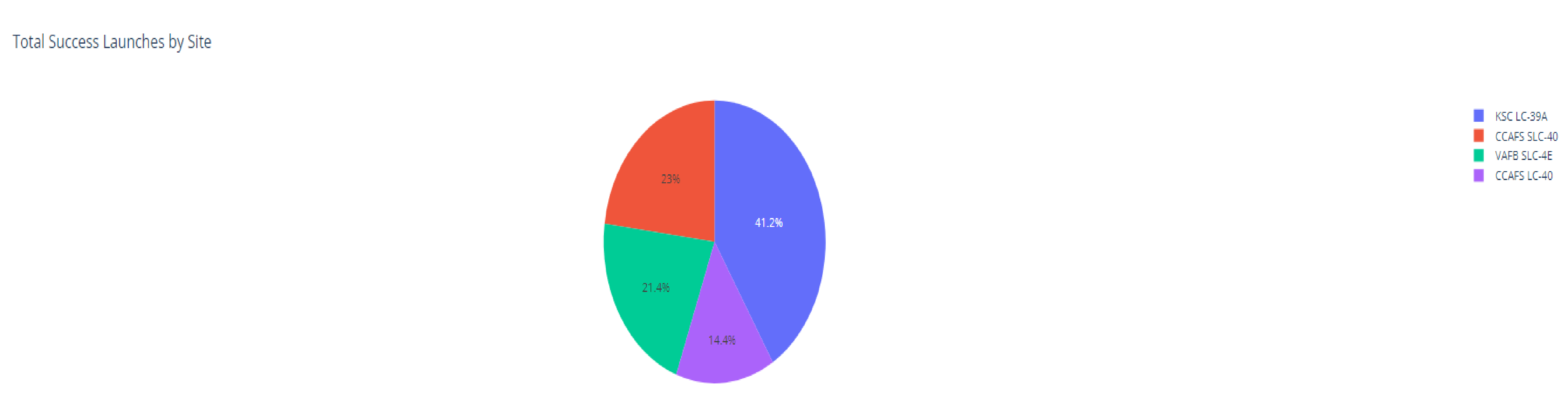
Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.



Section 4

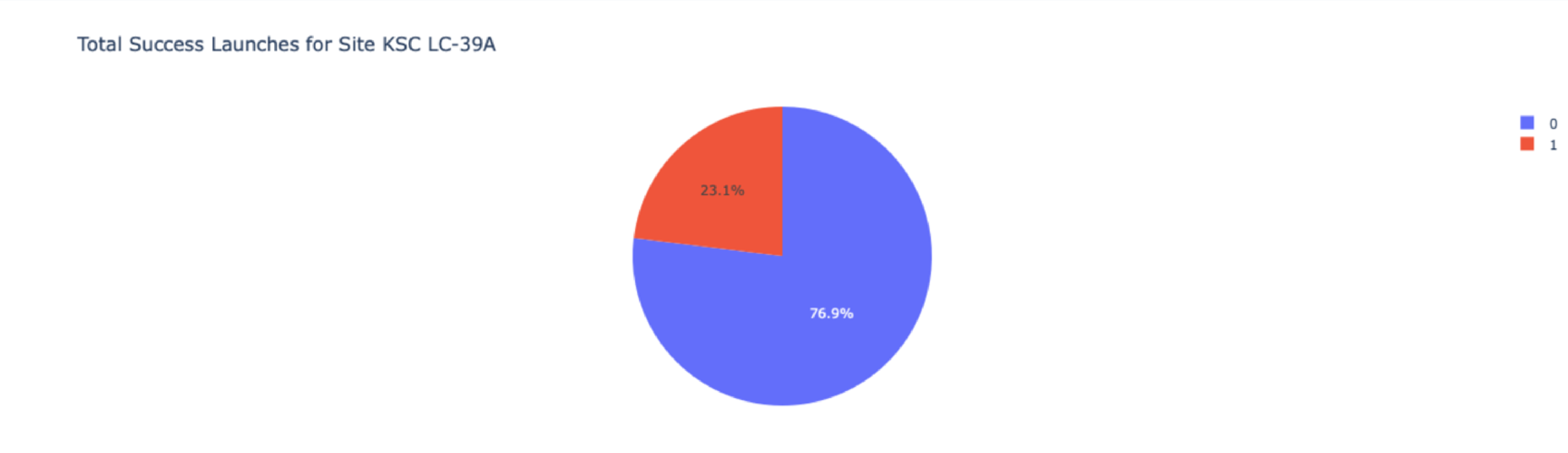
Build a Dashboard with Plotly Dash

Launch success rate for all sites



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

Success rate for the most successful launching site



SC 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

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

Section 5

Predictive Analysis (Classification)

Classification Accuracy

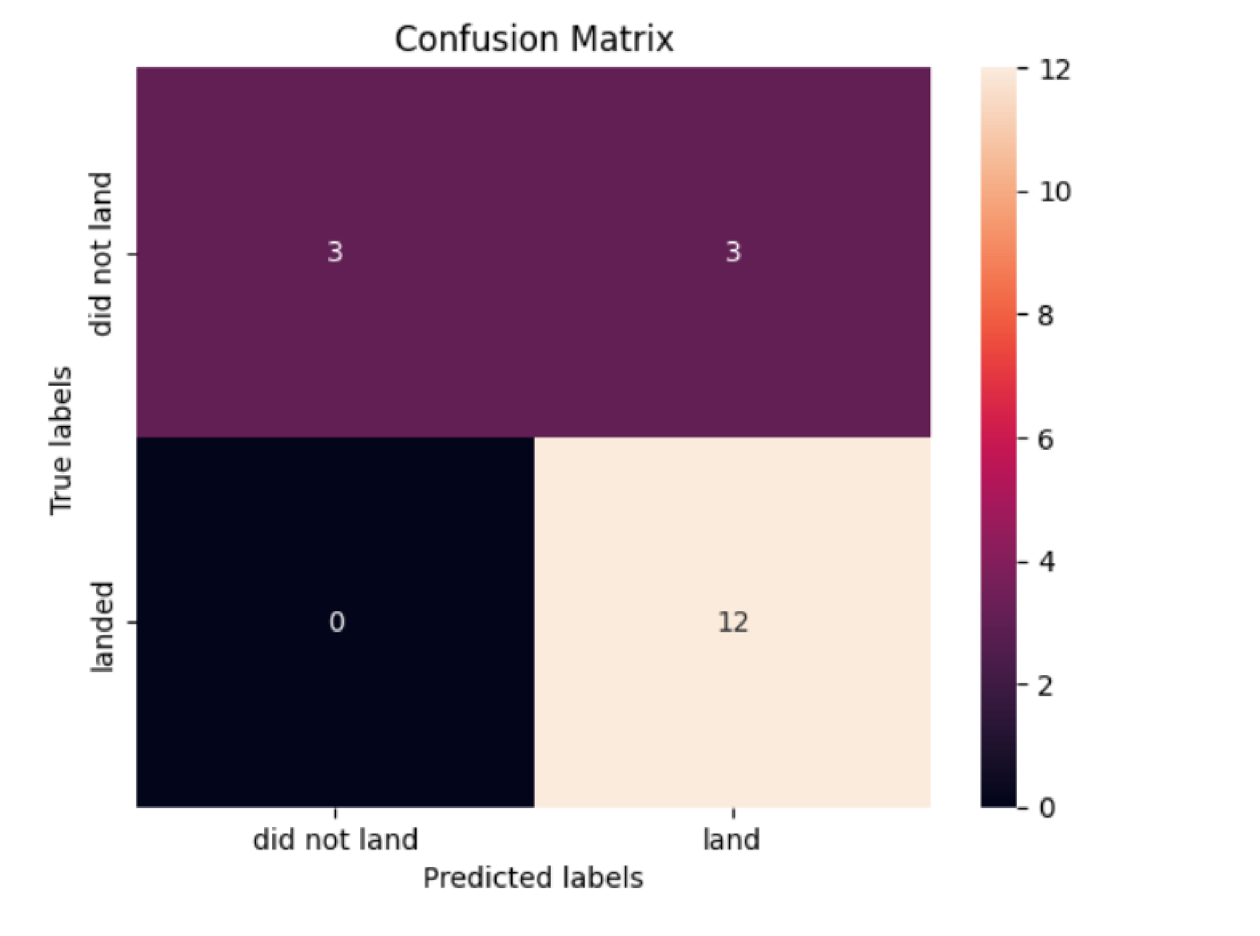
[38]:

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

• Based on the scores of the Test Set, we can not confirm which method performs best.
• Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole dataset.
• The scores of the whole Dataset confirm that the best model is the Support Vector Machine Model. This model has not only higher scores, but also the highest accuracy

Confusion Matrix

Examining the confusion matrix, we see that SVM model can distinguish between the different classes, with perfect accuracy when predicting successful landings, but with worse performance when trying to predict the unsuccessful landings (false positives)



Conclusions

- Support Vector Machine model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate

Appendix

- GitHub repository containing all notebooks and code in case hyperlinks did not work : https://github.com/FrihMalek/IBM-Applied-Data-Science-Capstone

Thank you!