IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Solmaz Vejdani
05-12-2024

Outline

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

Executive Summary

Summary of methodologies

- Data Collection
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning to perform predictive analysis

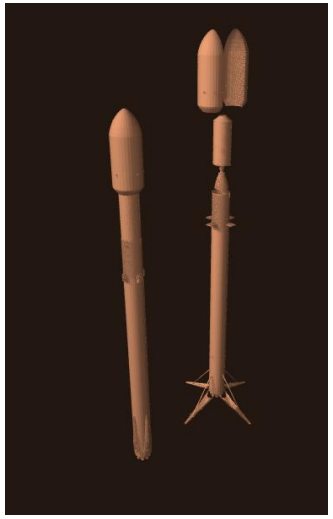
Summary of all results

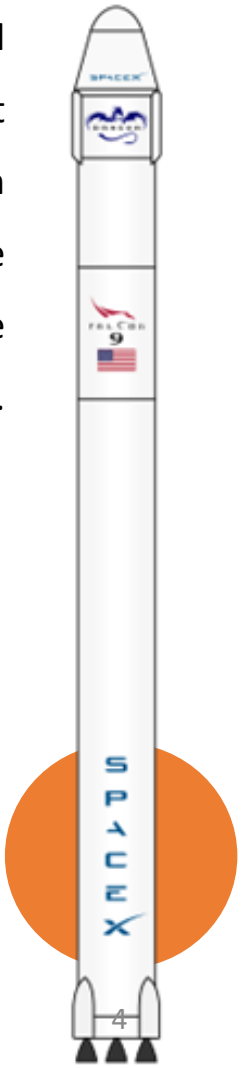
- Exploratory data analysis result
- Interactive analytics demo in screenshots
- Predictive Analytics result from Machine Learning Lab

Introduction

Project background and context

SpaceX was founded by Elon Musk in 2001 with a vision of decreasing the costs of space launches. SpaceX strives to make space travel affordable. Its accomplishments include sending spacecraft to the international space station, launching a satellite constellation that provides internet access and sending manned missions to space. Space X advertises Falcon 9 rocket launches with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, most of this savings is because SpaceX can reuse the first stage of the launch by re-land the rocket to be used on the next mission. Hence, if we can determine if the first stage will land, we can determine the cost of a launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully. 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:

- What factors determine if the rocket will land successfully?
- The interaction amongst various features that determine the success rate of a successful landing.
- How payload mass, launch site, number of flights, and orbits affect first-stage landing success
- What is the best predictive model for successful landing (binary classification)

Section 1

Methodology

Methodology

- **Data collection:** collecting data by using SpaceX REST API and web scraping techniques
- **Data wrangling:** wrangling data: by filtering the data, handling missing values and applying one hot encoding – to prepare the data for analysis and modeling
- **Exploratory Data Analysis:** exploring data via EDA with SQL and data visualization techniques
- **Data Visualization:** Visualizing the data using Folium and Plotly Dash
- **Building Models:** to predict landing outcomes using classification models. Tune and evaluate models to find best model and parameters

Data Collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

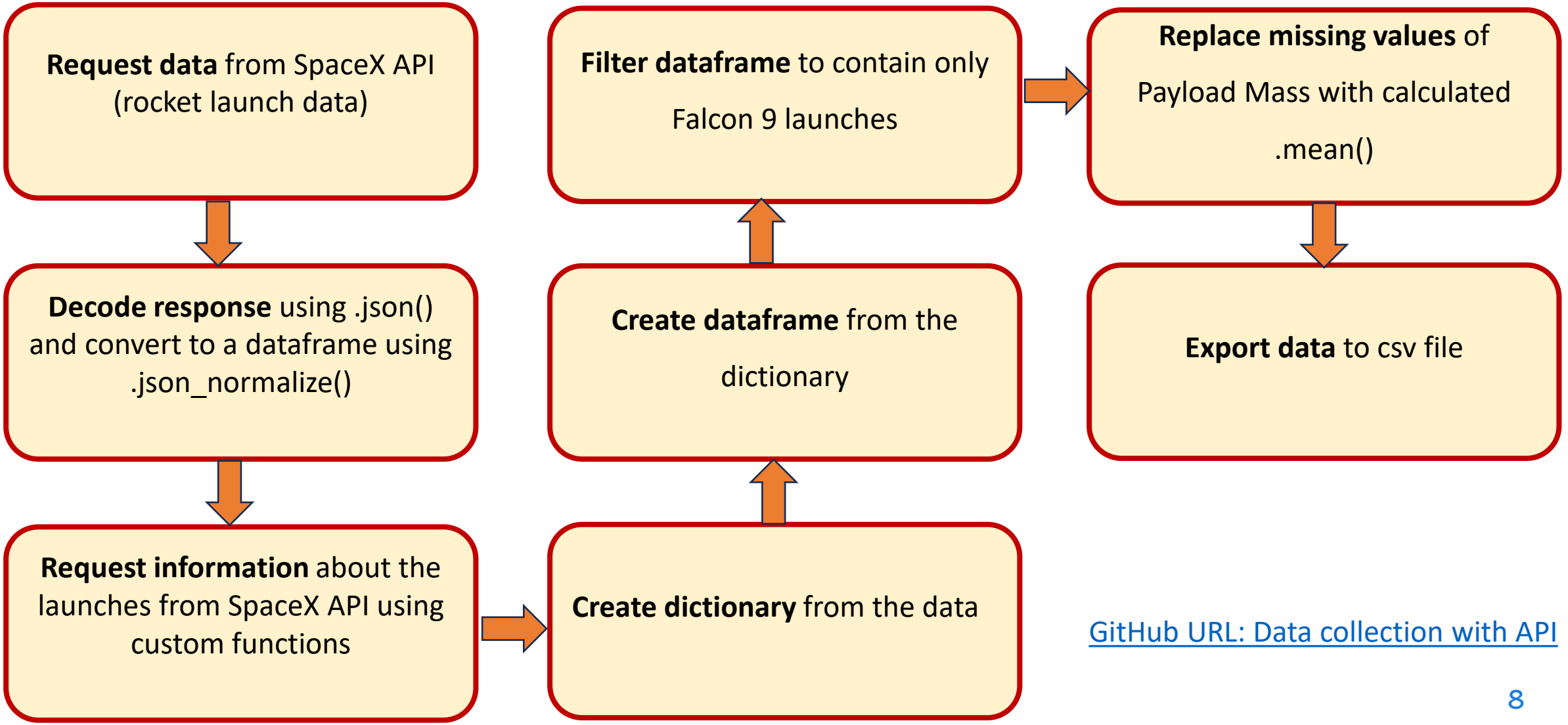
Data Columns obtained by using **API**

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

Data Columns obtained by using **Web Scraping**

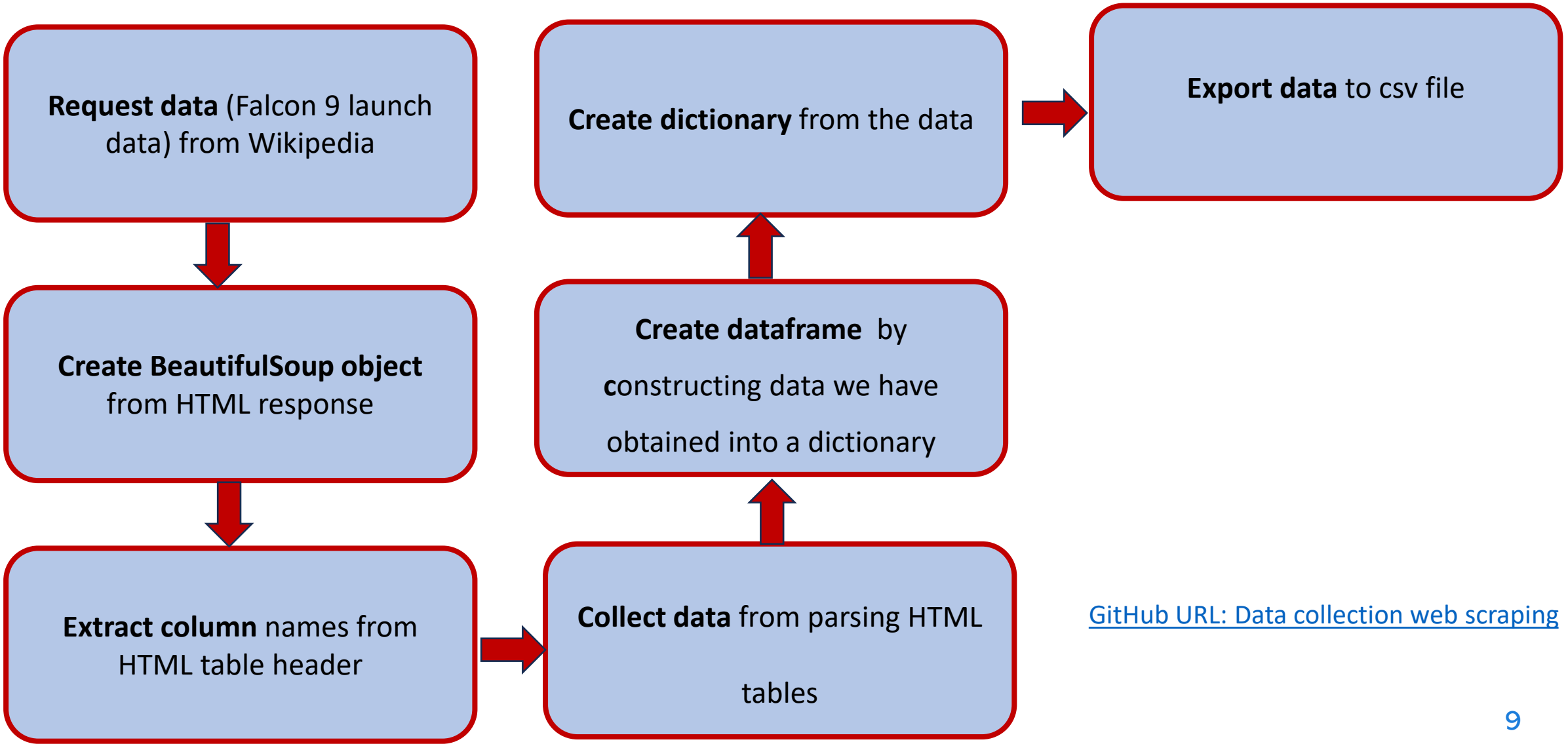
- Flight No
- Launch site
- Payload
- PayloadMass
- Orbit
- Customer
- Launch outcome
- Version Booster
- Booster landing
- Date
- Time

Data Collection – SpaceX API



GitHub URL: Data collection with API

Data Collection by using **Web Scraping**



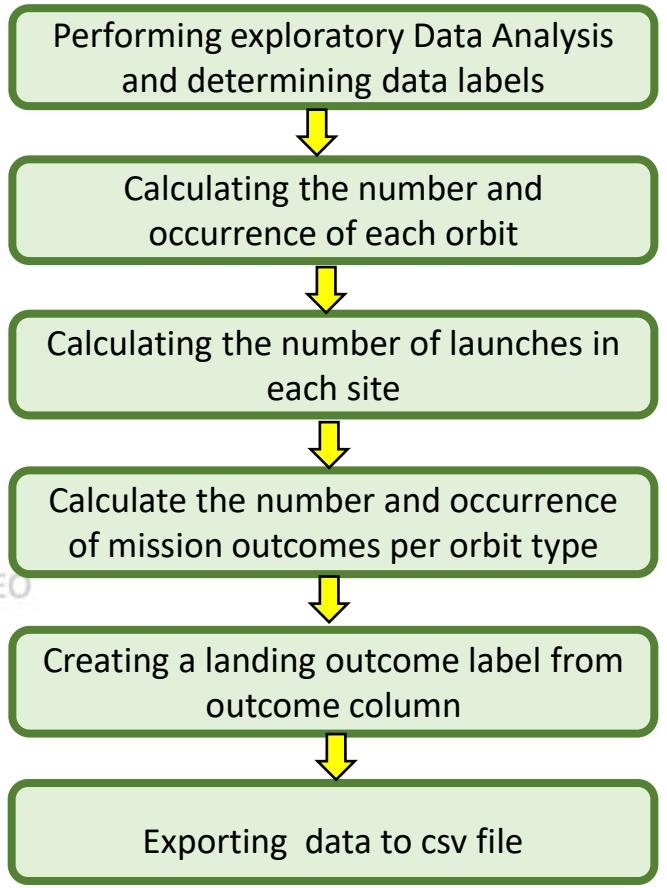
GitHub URL: Data collection web scraping

Data Wrangling

Data Wrangling is the process of cleaning and unifying messy and complex data sets for easy access and Exploratory Data Analysis (EDA). The flow chart represents the steps used for data wrangling.

In the data set, there are several different cases the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident. For example:

- **Landing was not always successful.**
- **True Ocean means mission outcome had a successful landing to a specific region of the ocean. False Ocean represents an unsuccessful landing to a specific region of ocean.**
- **True RTLS means the mission had a successful landing on a ground pad.**
- **False RTLS represents an unsuccessful landing on a ground pad.**
- **True ASDS means the mission outcome had a successful landing on a drone ship.**
- **False ASDS represents an unsuccessful landing on drone ship.**
- **At the end outcomes are converted into 1 for a successful landing and 0 for an unsuccessful. landing**

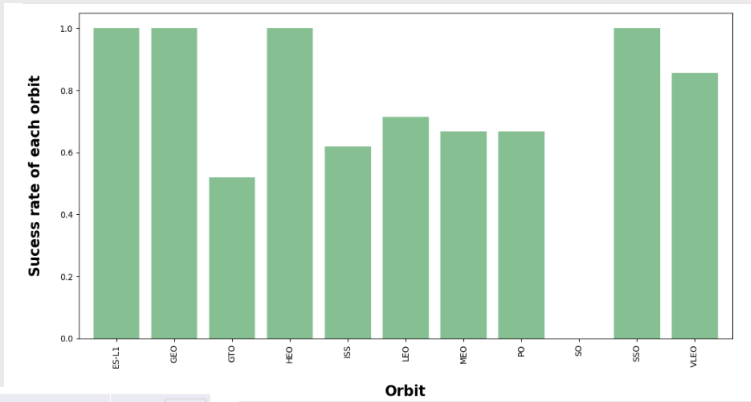


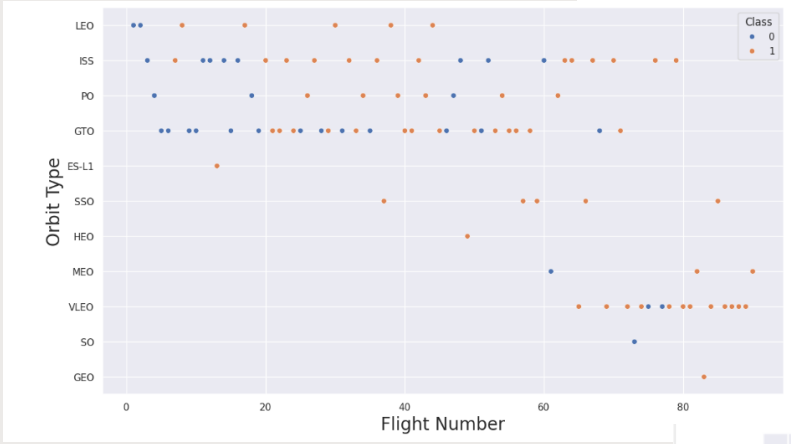
GitHub URL: Data Wrangling

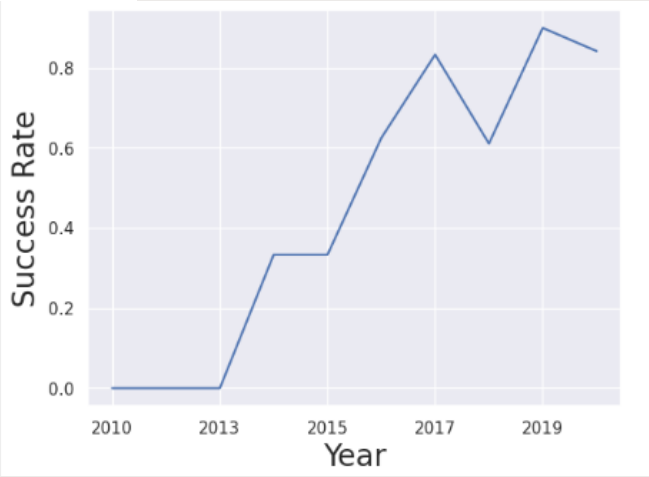
EDA with Data Visualization

Charts & Plots:

- Flight Number vs. Payload Mass -> (Scatter plot)
- Flight Number vs. Launch Site -> (Scatter plot)
- Payload Mass (kg) vs. Launch Site -> (Scatter plot)
- Success rate of each orbit vs orbit type -> (Bar Chart)
- Flight Number vs Orbit type -> (Scatter plot)
- Payload Mass (kg) vs. Orbit type -> (Scatter plot)
- Launch success Rate Yearly Trend -> (Line chart)







Scatter Plots represent the relationship between variables. Existing relationships can be used in machine learning model.

Bar charts show comparisons among discrete categories.

Line charts show trends in data over time.

GitHub URL: EDA with Data Visualization

EDA with SQL

SQL Queries:

- Displaying names of unique launch sites.
- Displaying five records where launch site begins with ‘CCA’.
- Displaying total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1.
- Listing the date of first successful landing on ground pad.
- Listing then names of boosters which had success landing on drone ship and. have payload mass greater than 4,000 but less than 6,000.
- Listing the total number of successful and failed missions.
- Listing the names of booster versions which have carried the max payload.
- Listing the failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015.
- Ranking the count of landing outcomes between 2010-06-04 and 2017-03-20, descending.

GitHub URL: EDA with SQL

Build an Interactive Map with Folium

Adding map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.

Markers Indicating Launch Sites:

- Adding blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates.
- Adding red circles at all launch sites coordinates with a popup label showing its name using its name using its latitude and longitude coordinates.

Colored Markers of Launch Outcomes:

- Adding green colored Markers for successful and red for failed launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

- Adding colored lines to show distances between a launch site (CCAFS SLC-40) and its proximity to the nearest, railway, highway, coastline and city.

GitHub URL: Interactive Visual Analytics with Folium

Building a Dashboard with Plotly Dash

Interactive dashboard with Plotly dash represents:

- **Dropdown List :**

A dropdown list to allow user to select all launch sites or a certain launch site.

- **Pie Charts showing success Launches (All Sites/Certain Site):**

To allow user to see successful and unsuccessful launches as a percent of the total.

- **Slider of Payload Mass Range:**

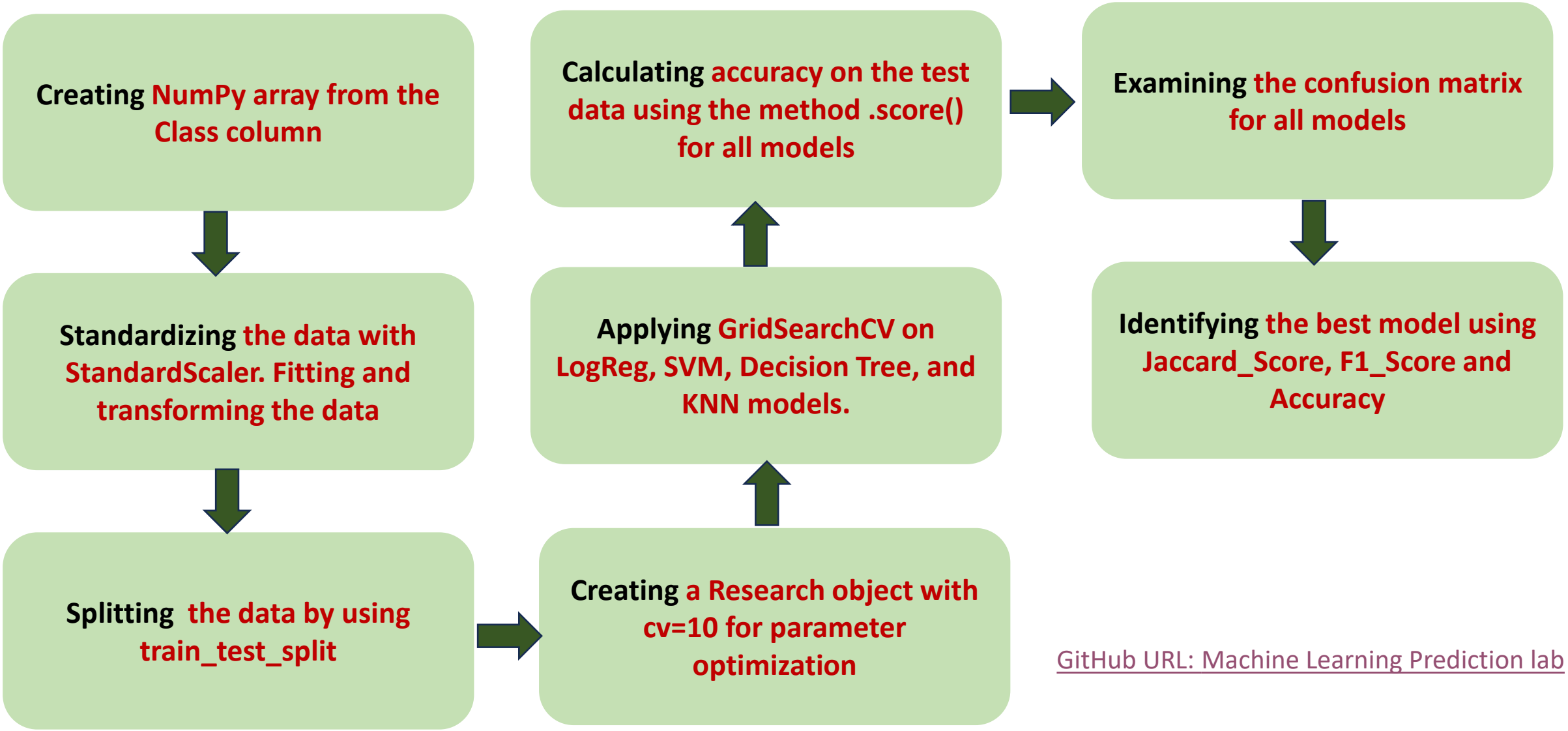
To allow user to select payload mass range.

- **Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version**

A scatter chart is added to allow user to see the correlation between Payload and Launch Success.

GitHub URL: SpaceX Dash app

Predictive Analysis (Classification)



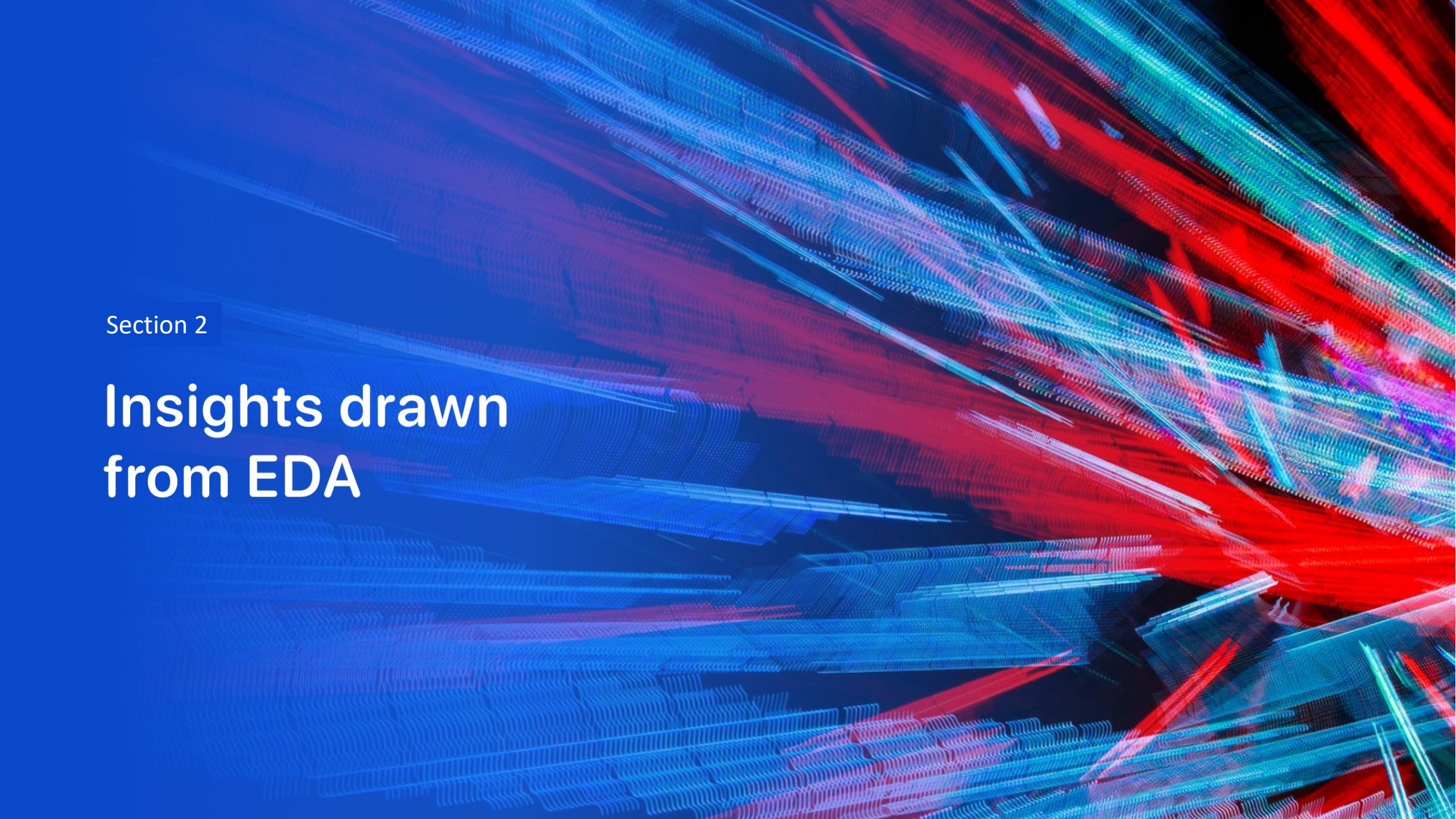
GitHub URL: Machine Learning Prediction lab

Results

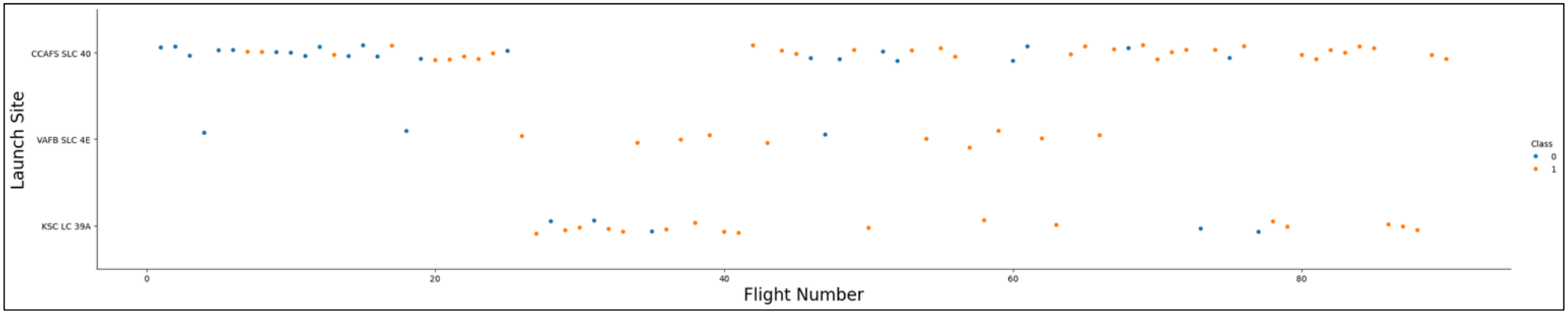
Exploratory data analysis results

Interactive analytics demo in screenshots

Predictive analysis results


Section 2
Insights drawn from EDA

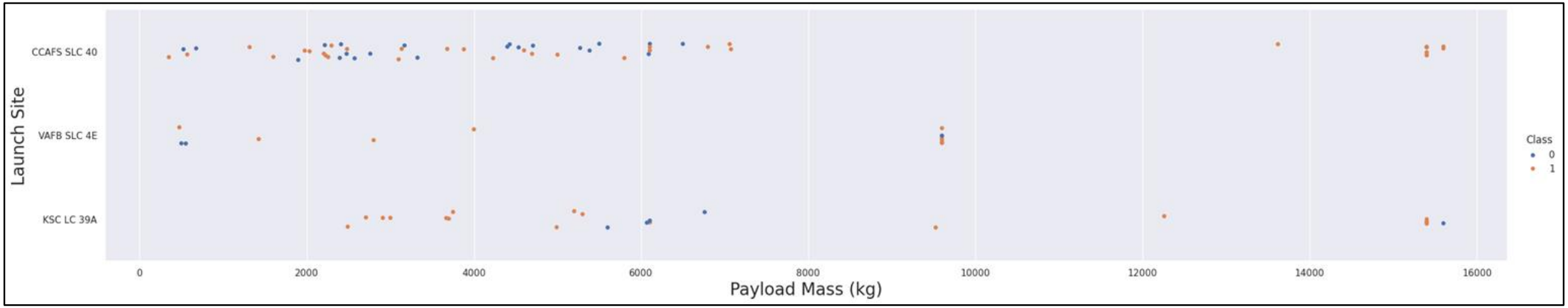
Flight Number vs. Launch Site



Explanations

- Earlier flights had a lower success rate (blue = fail).
- Later flights had a higher success rate (orange = success).
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be inferred that new launches have a higher success rate.

Payload vs. Launch Site



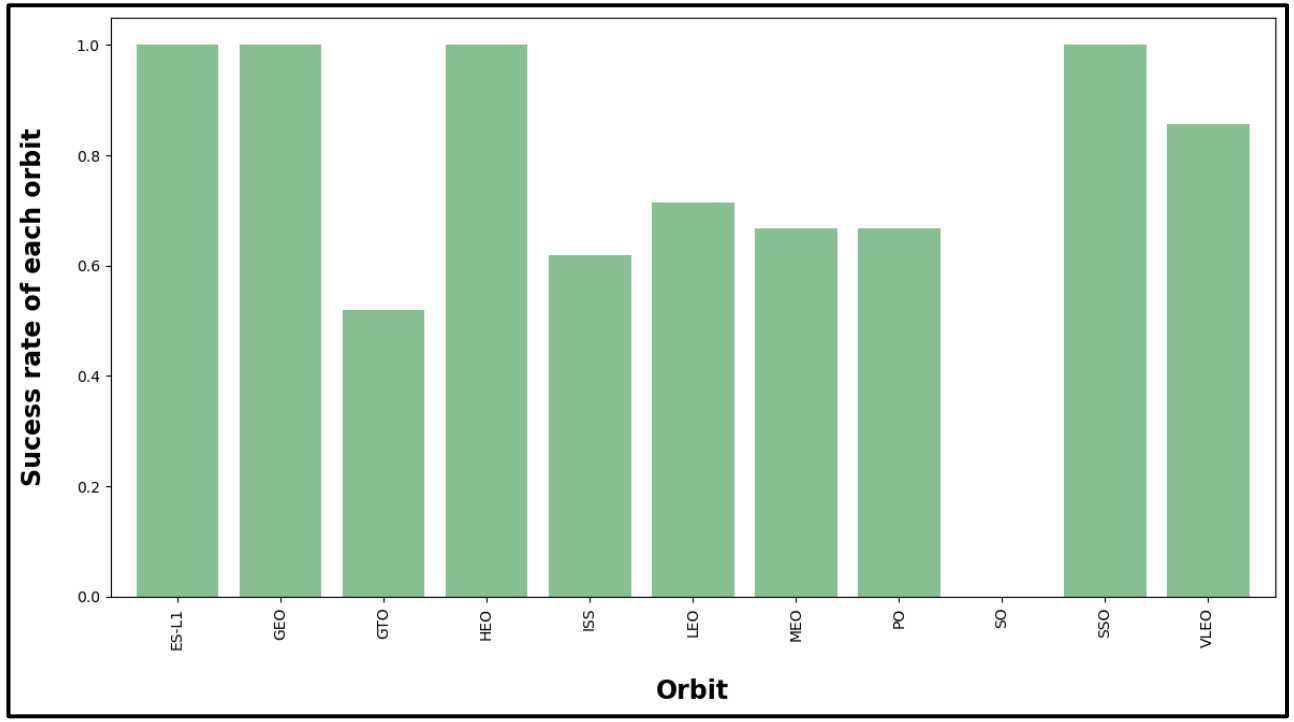
Explanations

- For every launch site the higher the payload mass, the higher the success rate.
- Most launches with a payload greater than 7000 kg were successful
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg.
- VAFB SKC 4E has not launched anything greater than almost 10,000 kg.

Success Rate vs. Orbit Type

Explanations

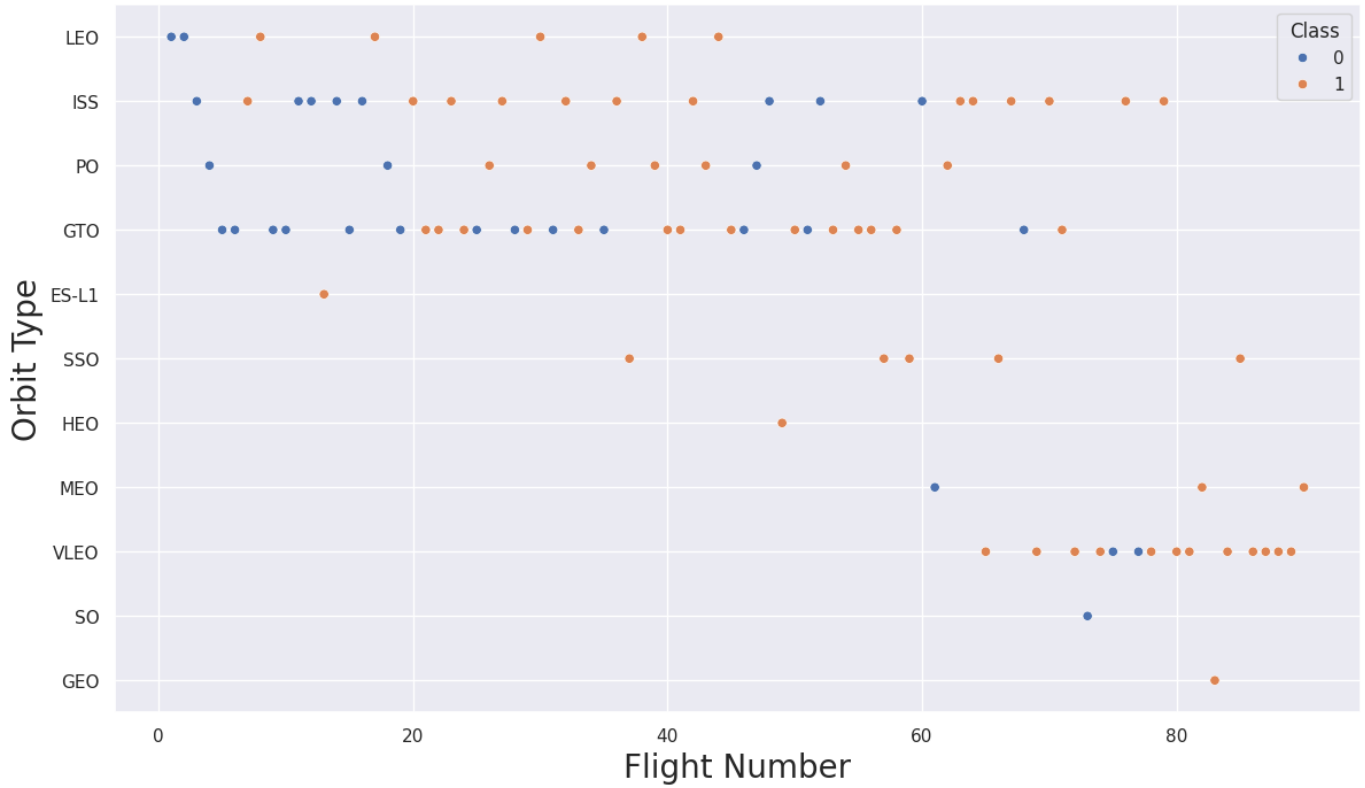
- 100% Success Rate for orbits : ES-L1, GEO, HEO and SSO
- 50%-80% Success Rate for orbits: GTO, ISS, LEO, MEO, PO and VLEO
- 0% Success Rate for orbit: SO



Flight Number vs. Orbit Type

Explanations

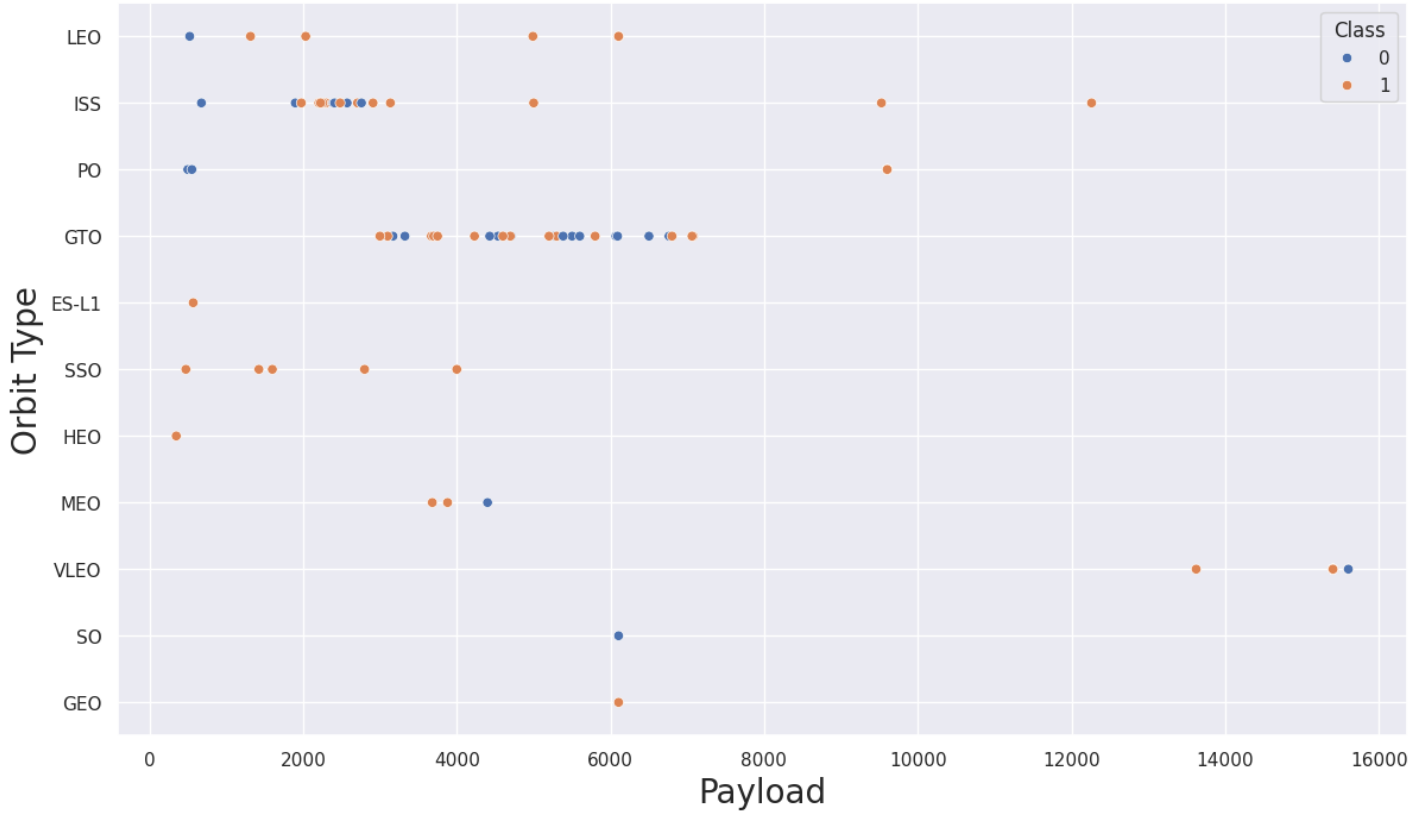
- The success rate typically increases with the number of flights for each orbit.
- For LEO orbit this relationship is highly apparent.
- However, the GTO orbit, however, does not follow this trend.



Payload vs. Orbit Type

Explanations

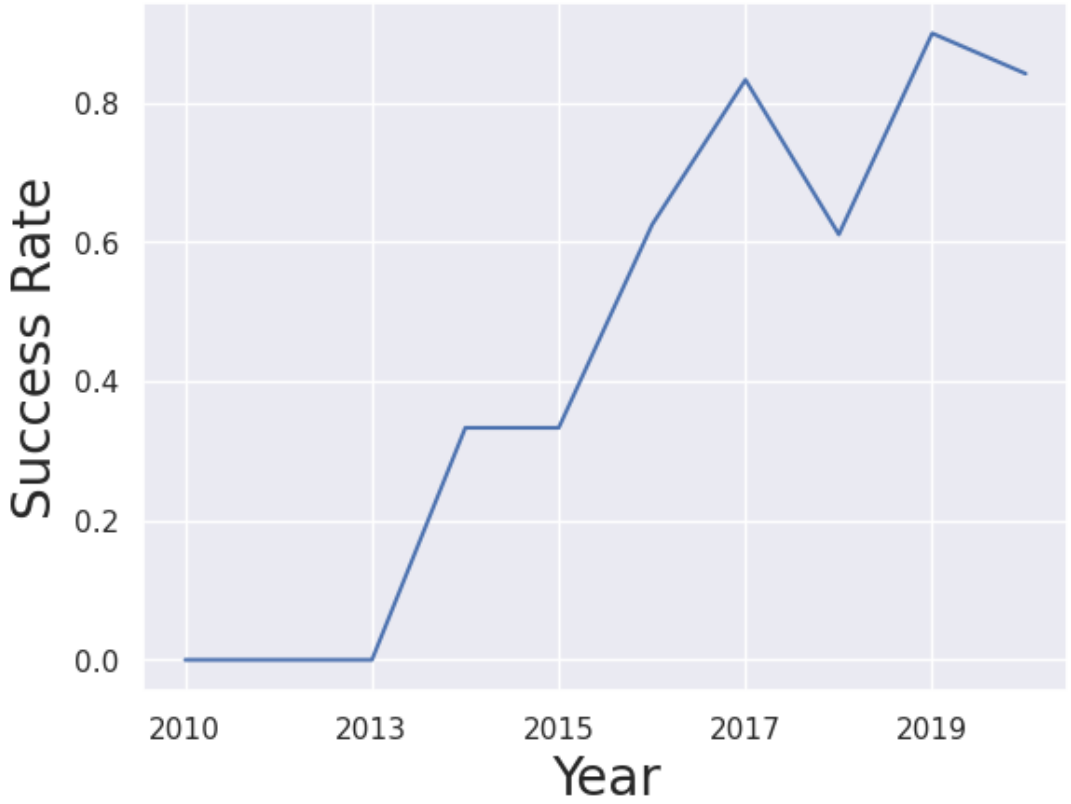
- Heavy payloads have positive impact on LEO, ISS and PO orbits.
- The GTO orbit has mixed success with heavier payloads.



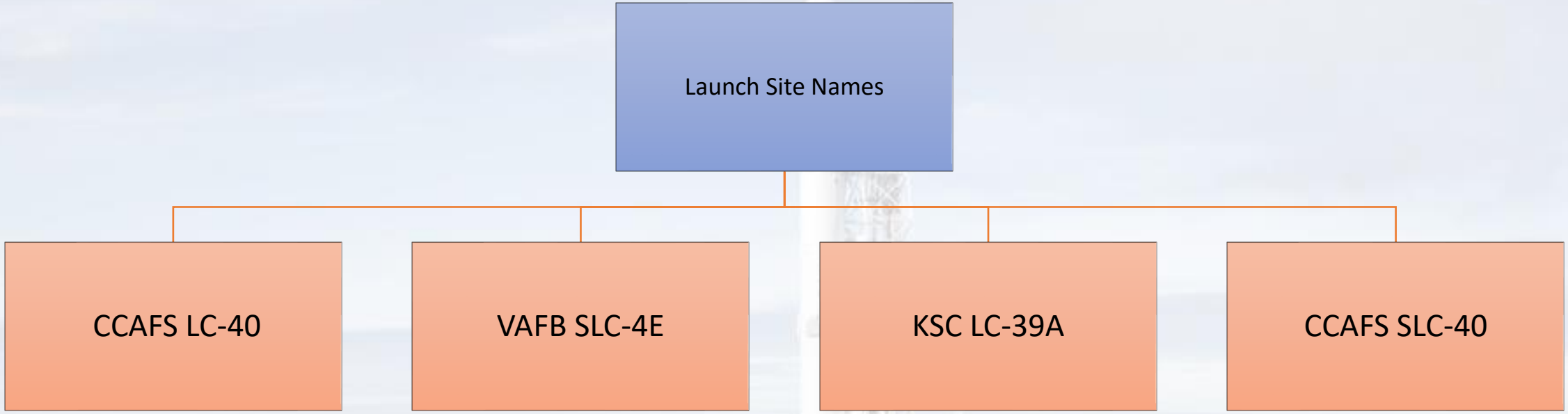
Launch Success Yearly Trend

Explanations

- The success rate decreased from 2017-2018 and from 2019-2020.
- The success rate improved from 2013-2017 and 2018-2019.
- Overall, the success rate has improved since 2013.



All Launch Site Names



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

```
In [12]: %sql select distinct launch_site from SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
```

Out[12]:

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

Launch Site Names Begin with 'CCA’

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

In [13]:
```
%sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
```

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

Out[13]:

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

Total Payload Mass

Total Payload Mass:

45596 kg (total) carried by boosters launched by NASA (CRS.)

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

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

In [14]:
```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';
```

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

Out[14]:

total_payload_mass
45596

Average Payload Mass by F9 v1.1

Average Payload Mass:

2928 kg (average) carried by booster version F9 v1.1

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

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

In [15]:
```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';
```

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

Out[15]:

average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

First Successful Landing in Ground Pad was on 2015-12-22.

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

```
In [17]: %sql select min(date) as first_successful_landing from SPACEXTABLE where landing_outcome = 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
Out[17]: first_successful_landing
         2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

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

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

In [40]:
```
%sql select booster_version from SPACEXTABLE where landing_outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4
```

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

Out[40]:

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

Total Number of Successful and Failure Mission Outcomes

Total Number of Successful and Failed Mission Outcomes

- 1 Failure in Flight
- 99 Success
- 1 Success (payload status unclear)

Listing the total number of successful and failure mission outcomes.

List the total number of successful and failure mission outcomes

[22]:
```
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

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

[22]:

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

2015 Launch Records

Boosters Carrying Max Payload:

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

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

In [41]:
```
%sql select booster_version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);
```

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

Out[41]:

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

2015 launch records

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

In [49]:
```
%%sql select substr(Date, 6,2) as MONTH, DATE, Booster_Version, launch_site, landing_outcome from SPACEXTABLE
      where landing_outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

Out[49]:

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

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

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.

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 [45]:

```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEXTABLE
      where DATE between '2010-06-04' and '2017-03-20'
      group by landing_outcome
      order by count_outcomes desc;
```

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

Out[45]:

Landing_Outcome	count_outcomes
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

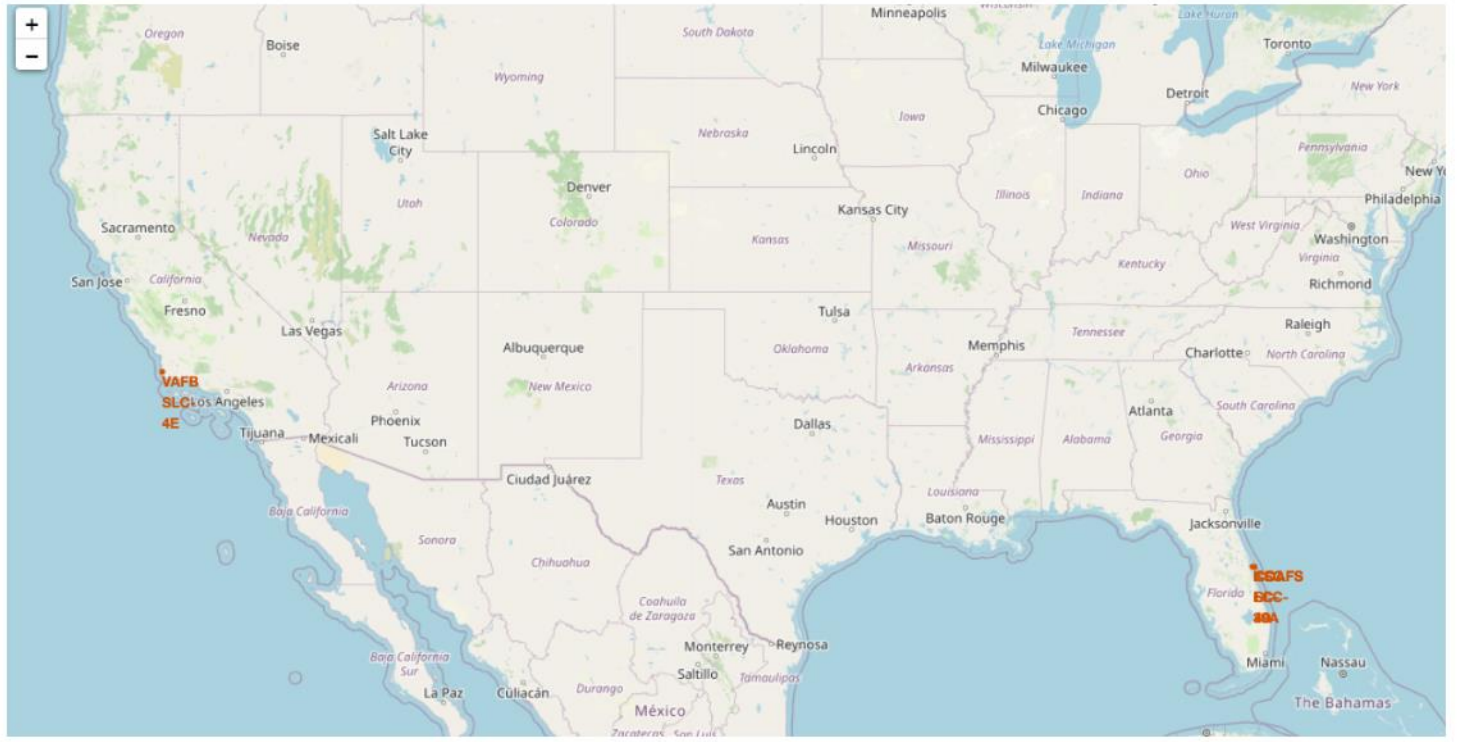
Section 3

Launch Sites Proximities Analysis

Location of all the Launch Sites

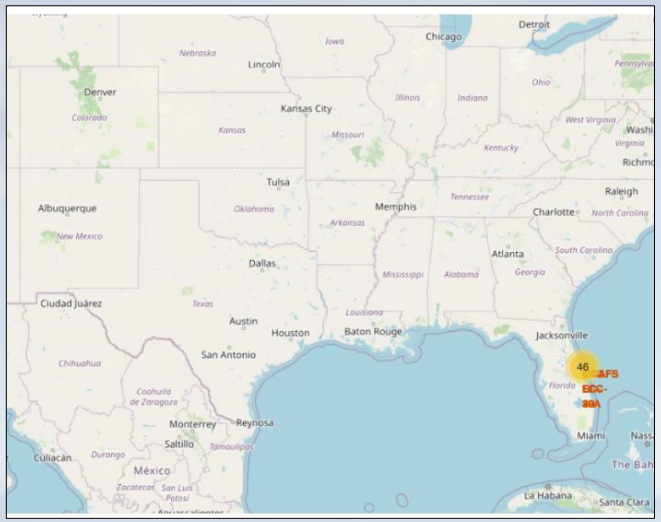
All launch sites' location markers on a global map

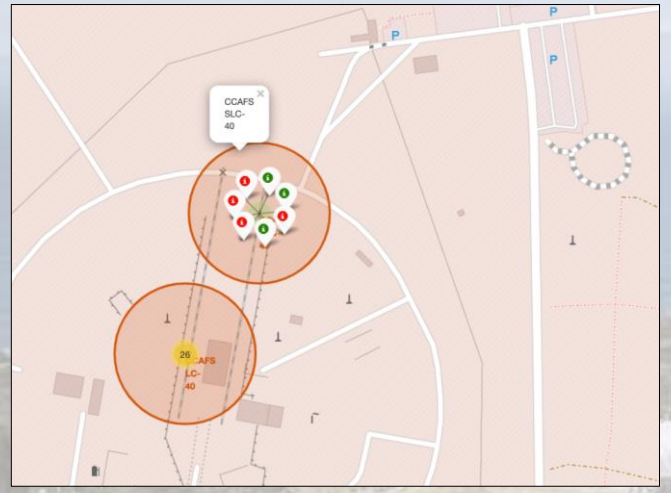
- All the SpaceX launch sites are located inside the United States.
- Most of Launch sites are in proximity to the Equator line.
- All launch sites are in very close proximity to the coast; while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people.

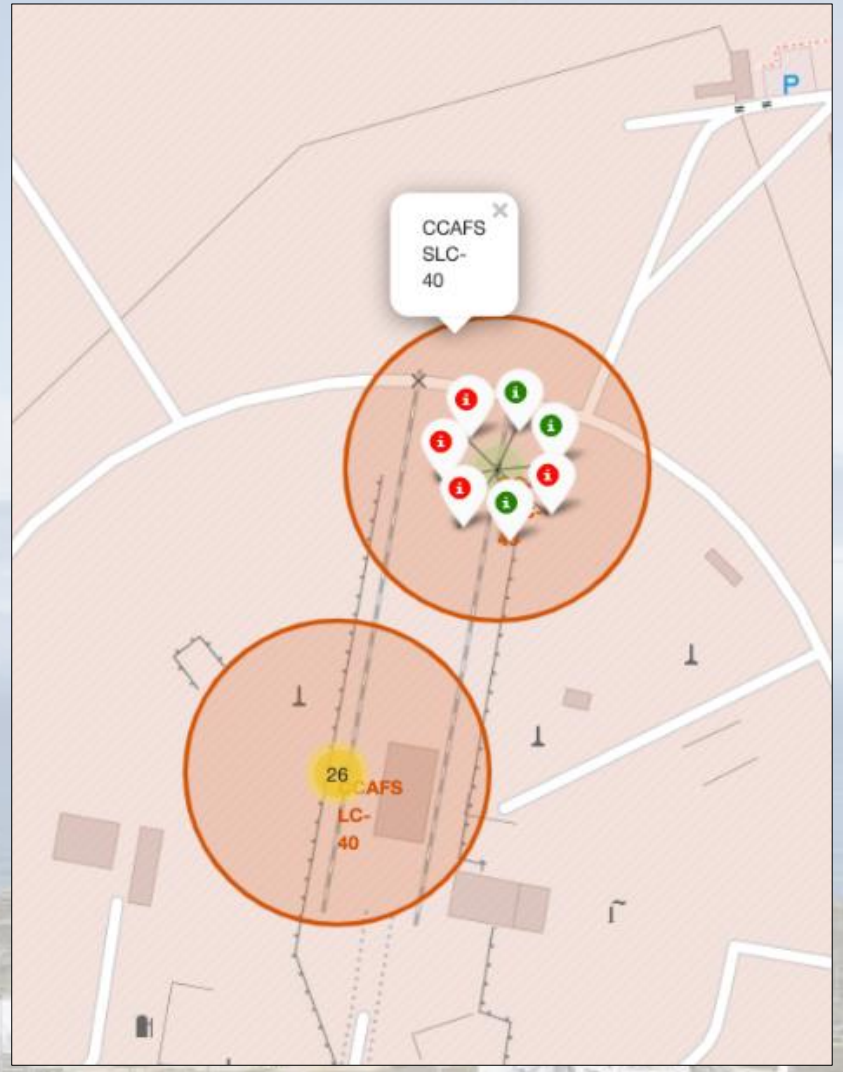


GitHub URL:**Interactive Visual Analytics with Folium**

Markers showing launch sites with color labels







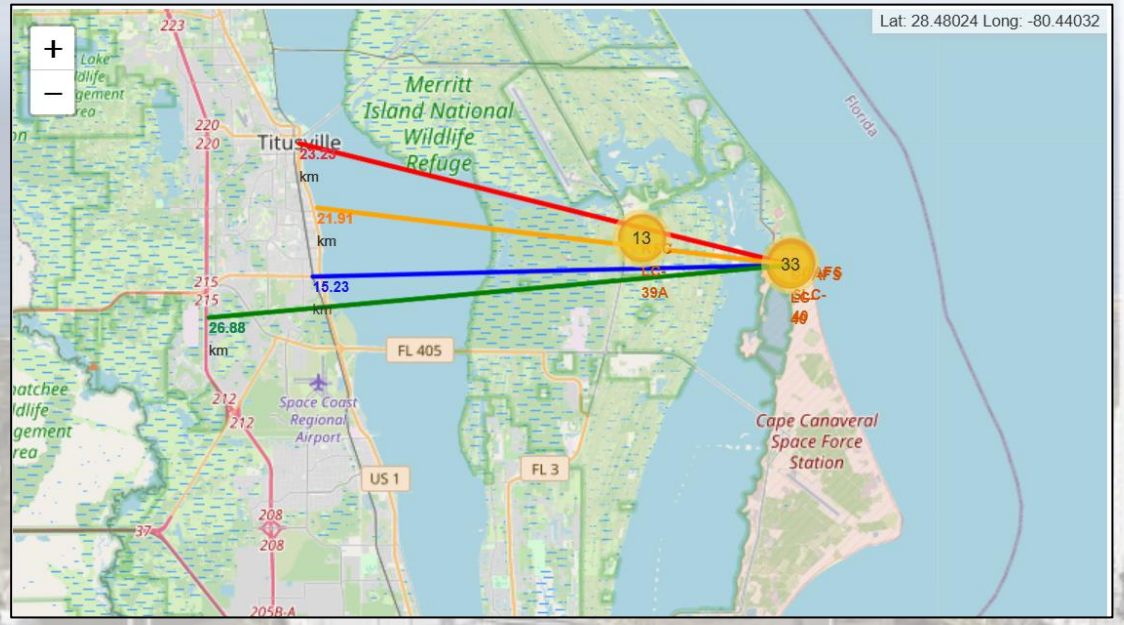
- At Each Launch Site: From the color-labeled markers we can identify relatively high success rates.
- Outcomes:
  - **Green** markers for successful launches.
  - **Red** markers for unsuccessful launches.
  - Launch site CCAFS SLC-40 has a 3/7 success rate (42.9%)

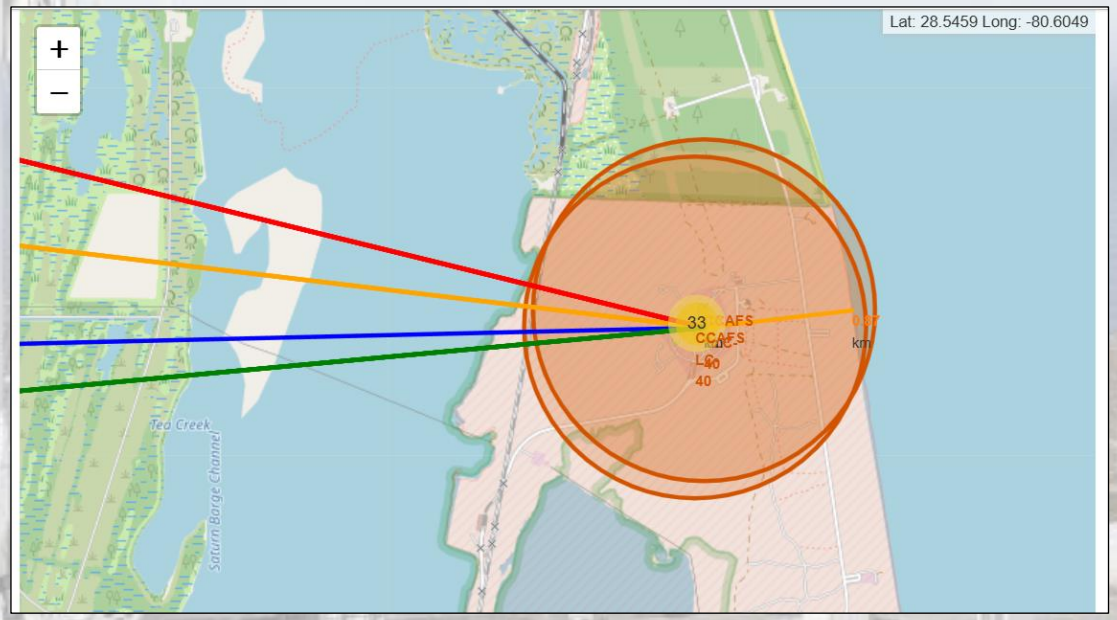
Distance to Proximities

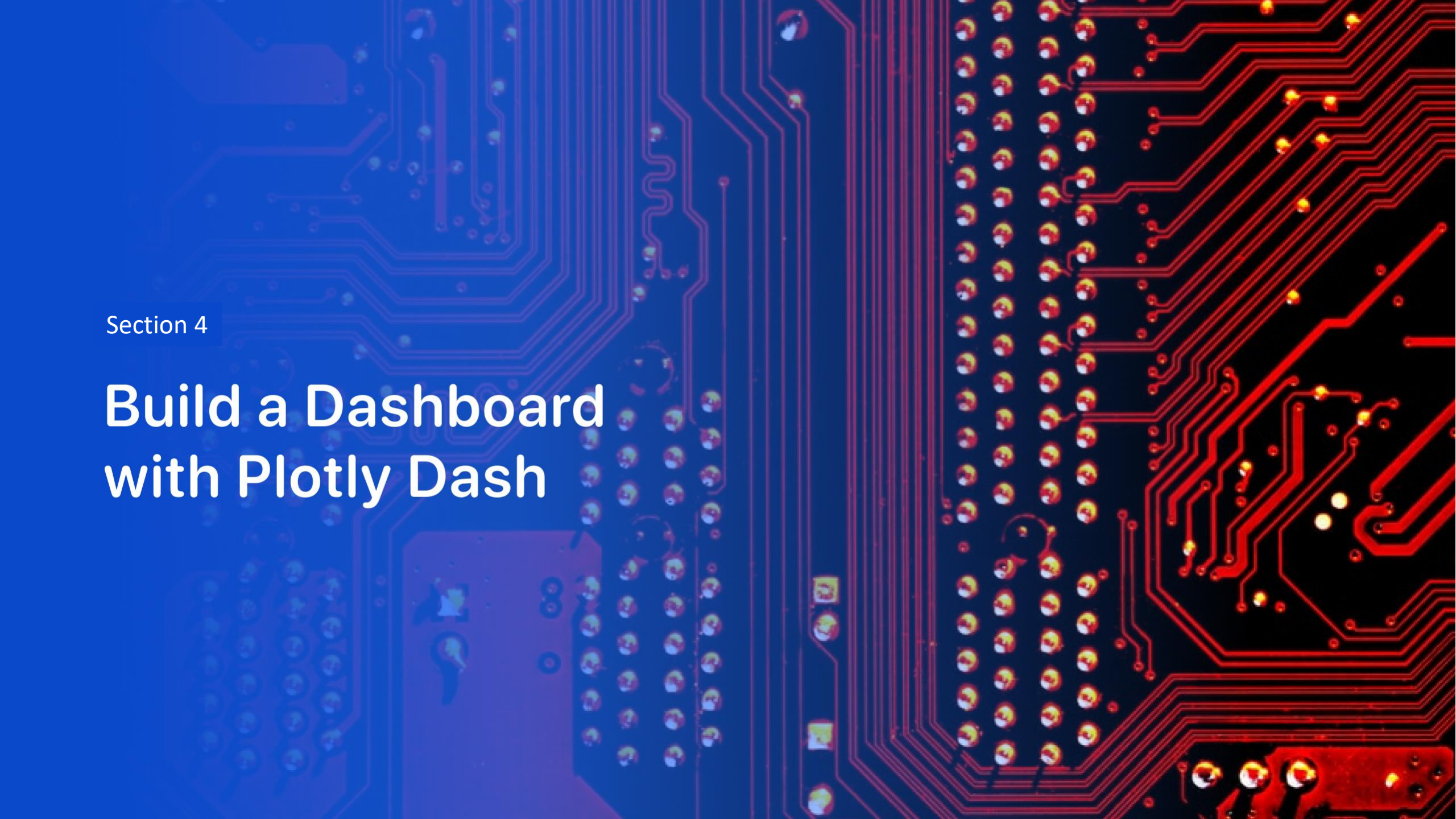
Lunch sites need to have close distance to coasts helps to ensure that spent stages dropped along the launch path or failed launches don't fall on people or property, basically they need to be away from anything a failed launch can damage, but still close enough to roads/rails/docks to be able to bring people and material to or from it in support of launch activities

From the visual analysis of the launch site CCAFS SLC-40, we can clearly see that it is:

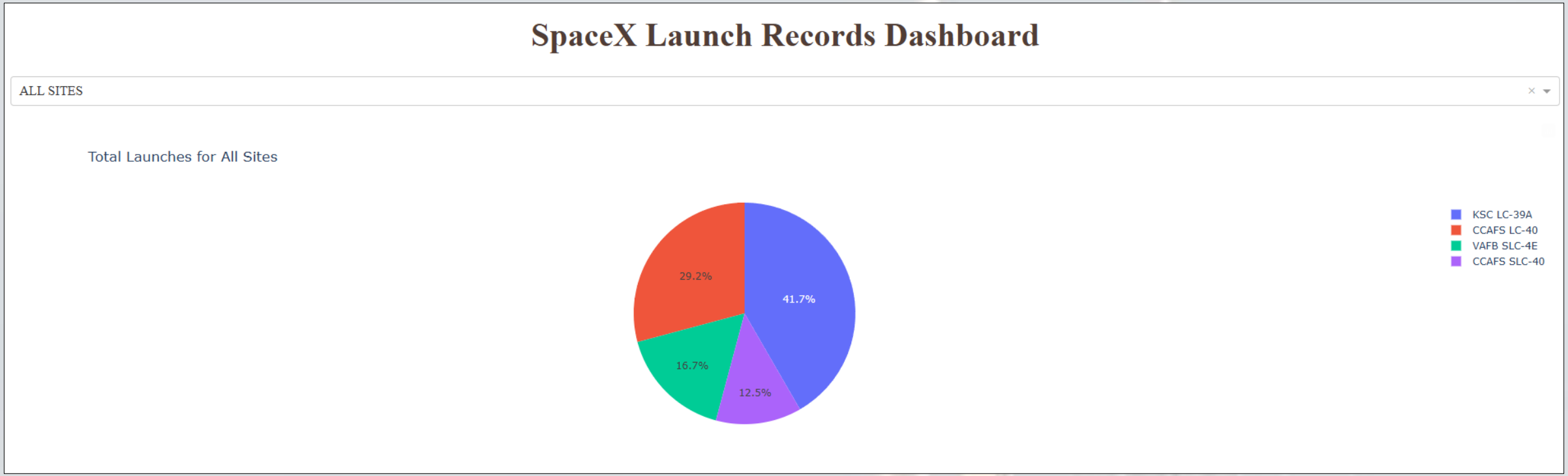
- **Relatively close distance to railway (15.23 km)**
- **Relatively close to highway (26.88 km)**
- **Relatively close to coastline (0.86 km)**
- **Launch sites do not keep certain distance away from cities, the launch site CCAFS SLC-40 is relatively close to its closest city Titusville (23.23 km).**





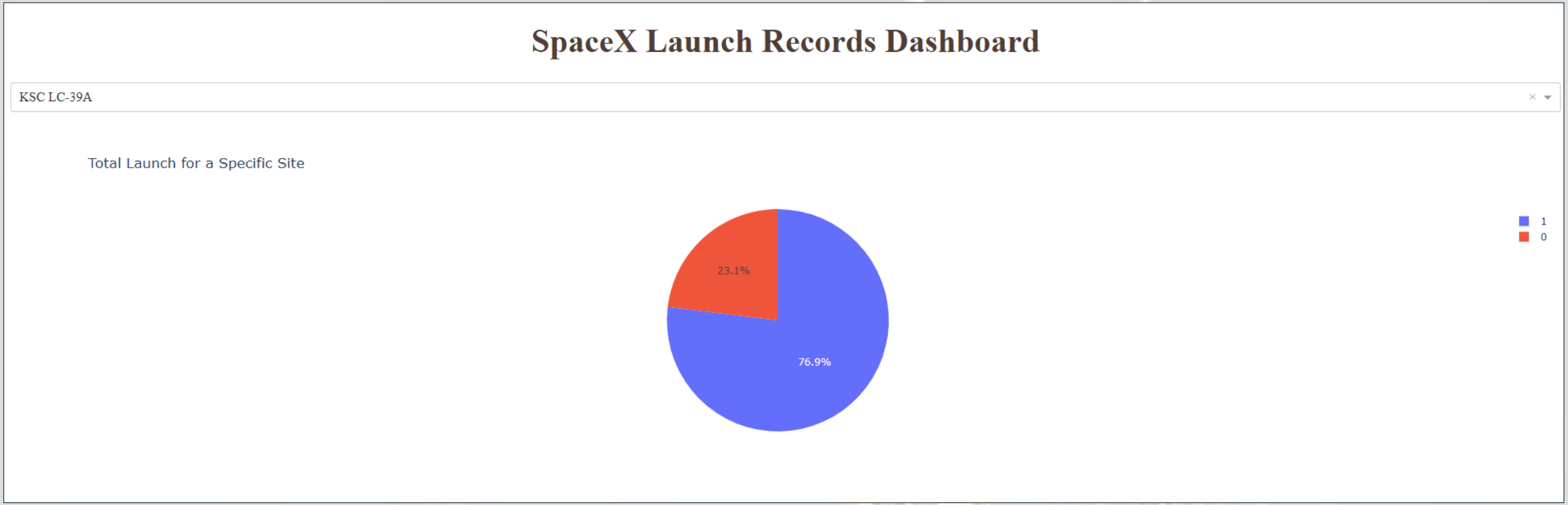

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches. KSC LC-39A has the most successful launches amongst launch sites (41.7%)

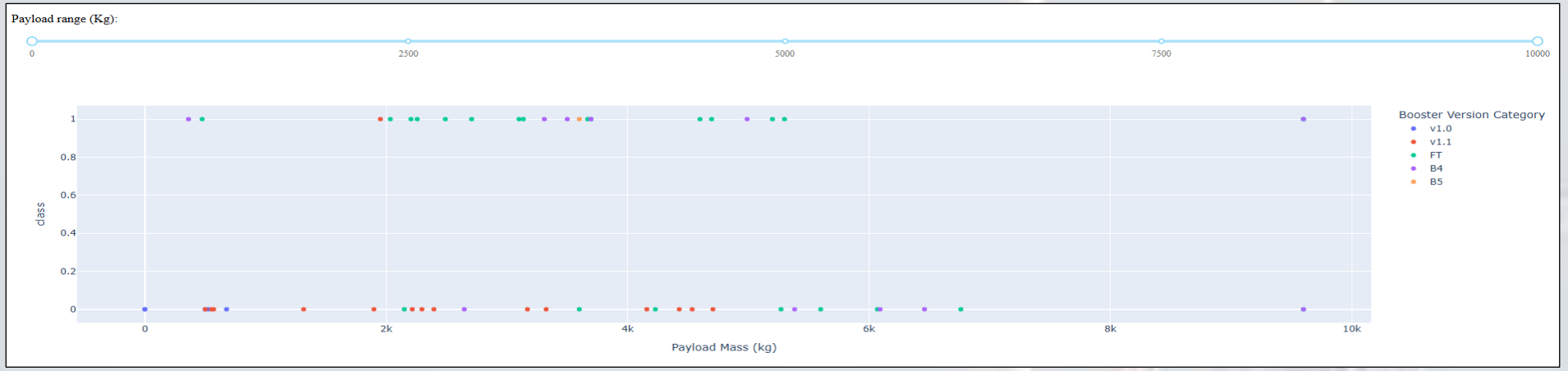
Launch site with highest launch success ratio

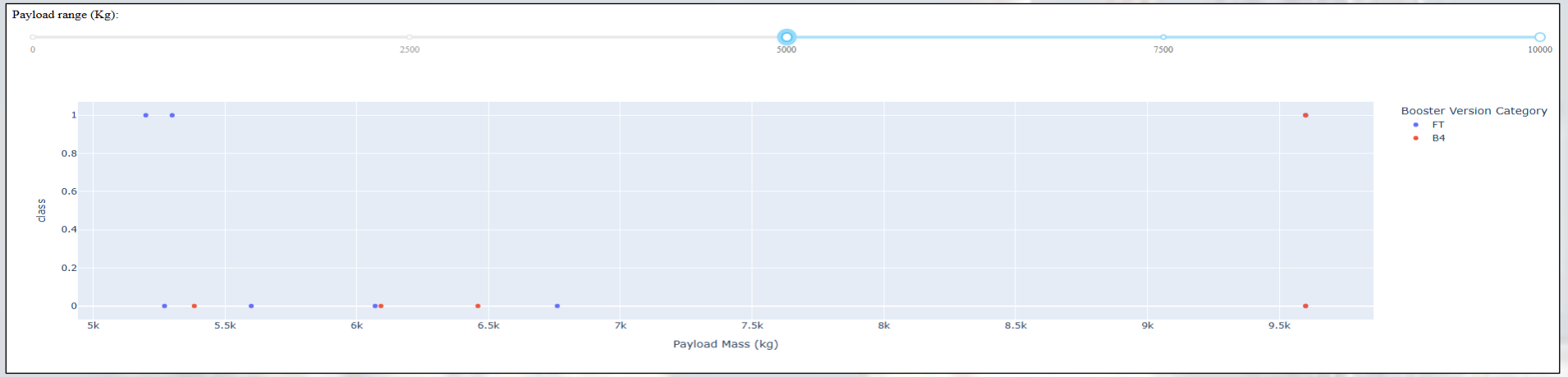


KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

Payload vs Launch Outcome Scatter Plot: (Payload Mass and Success)

By Booster Version: Payloads between 2,000 kg and 5,000 kg have the highest success.





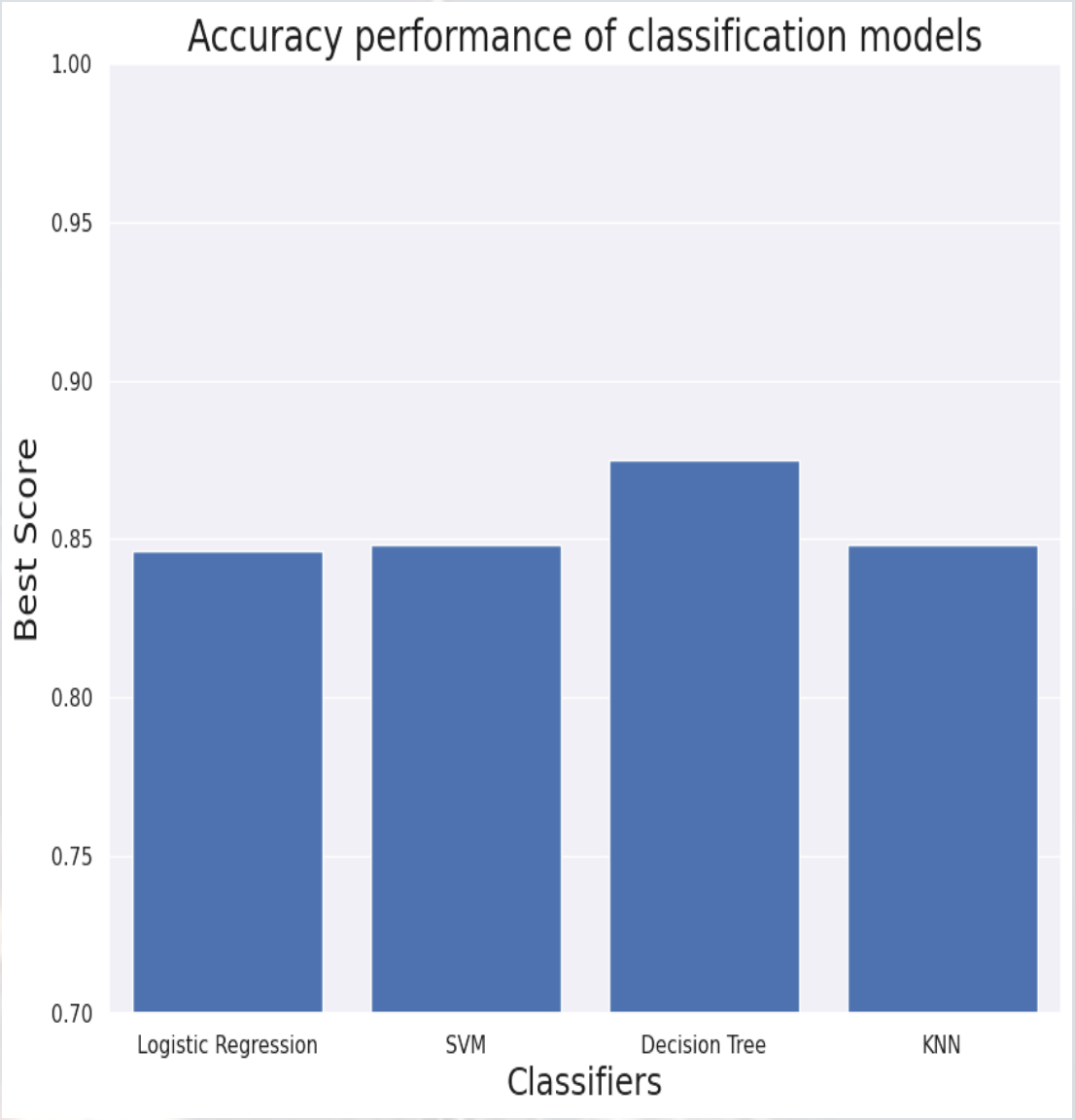
In the plots: 1 is indicating successful outcome and 0 is indicating an unsuccessful outcome.

Section 5

Predictive Analysis (Classification)

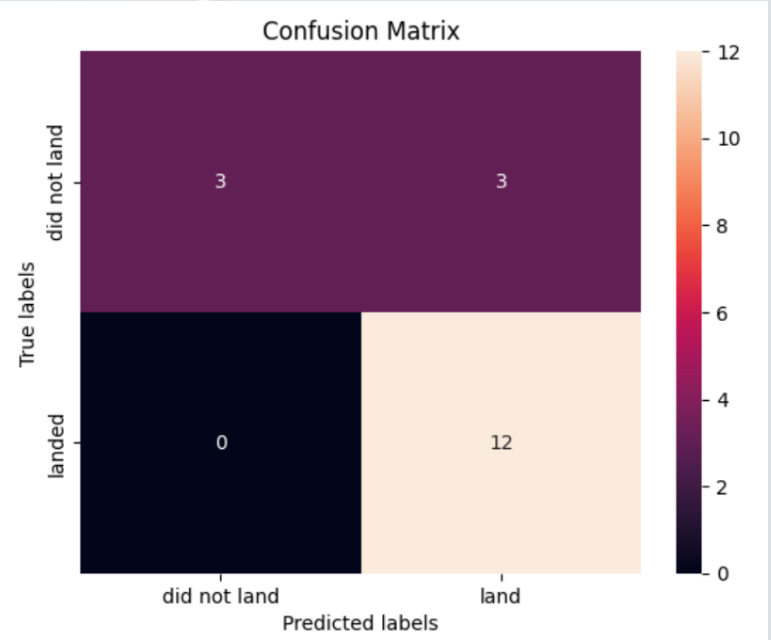
Classification Accuracy

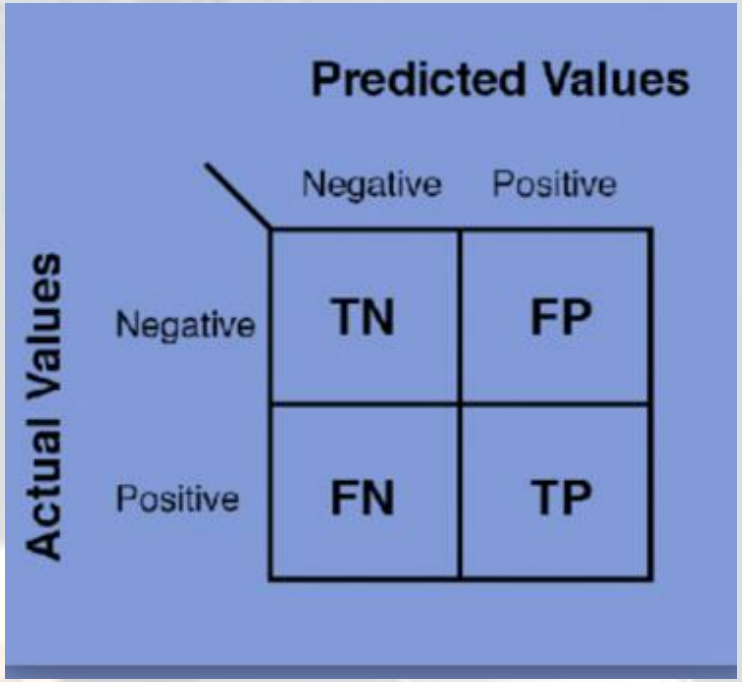
- All the models performed at about the same level and had almost the same scores and accuracy. This is likely due to the small dataset.
- The Decision Tree model slightly outperformed.
- Decision tree classifier performed the best with an accuracy score of approximately 0.875



Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes.
- From the test set, the decision tree classifier was able to correctly predict 12 observations that landed (12 True positives) and 3 observations that did not land (3 True Negatives).
- The classifier also had 0 false negatives because it did not wrongly predict any successful landings.
- The major problem is the false positives, The classifier had 3 false positives as it predicted wrongly for 3 observations that the outcome was successful.





Conclusions

- The Decision Tree Model is the best Machine Learning approach for this dataset.
- Launches with lower payload masses show higher success rates compared to those with larger payloads.
- Most launch sites are situated near the equator line and near coastlines.
- The success rate for SpaceX launches is increased since 2013 kept increasing till 2020.
- Over time, there is a noticeable upward trend in the success rate of launches.
- KSC LC-39A has the highest success rate among all launch sites.
- Orbits ES-L1, GEO, HEO, and SSO could successfully achieve a 100% success rate.

Thank you!