IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

AJIN P JOY
12-03-2022
Github URL:[SpaceX Repository]

Outline

Executive Summary

- **Summary of methodologies**
  - Data Collection through API and Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis using Visualization Tools
  - Exploratory Data Analysis using SQL
  - Building of Interactive Maps using Folium
  - Building a Dashboard using Plotly's Dash
  - Predictive Analysis using Machine Learning Techniques
- **Summary of all results**
  - EDA Results
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- **Project background and context**

  Space X is an American aerospace manufacturer that launch rockets at a relatively lower sum of 62 million dollars, where as other traditional space organizations quote a massive sum of 165 million dollars for the same work. The reason behind this incredible efficiency is due to the fact that SpaceX can reuse the rocket's first stage. Using this data we can predict the cost of a rocket launch and also whether first stage of the rocket will land back successfully.

- **Problems you want to find answers**

  We want to predict if the first stage can be landed successfully.

Section 1

Methodology

Methodology

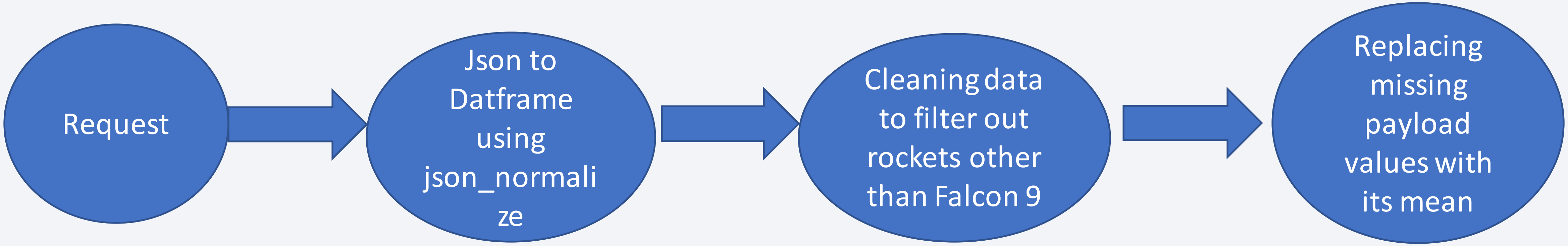
Executive Summary

- Data collection methodology
  - Data was scraped with the help of SpaceX API and from the Wikipedia page of SpaceX.
- Perform data wrangling
  - One-hot encoding was applied to categorical features and data was cleaned to remove empty values and unnecessary features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Several ML models were built and evaluated.

Data Collection

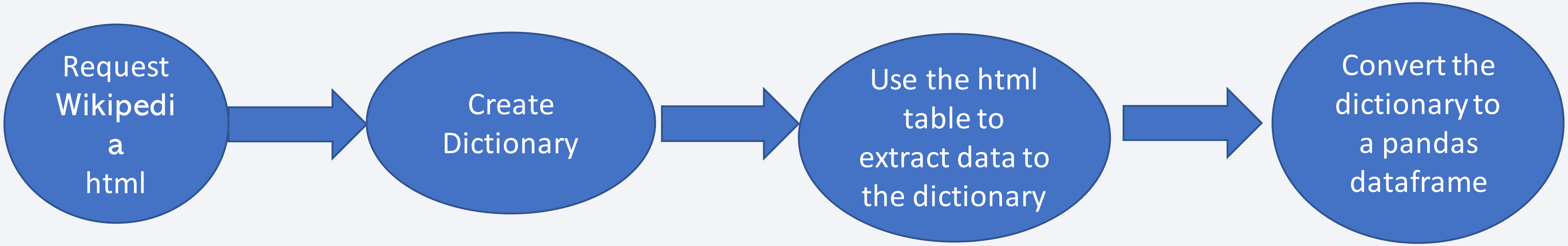
- **The data was collected using various methods**
  - We used a combination of API's and web scrapping methods in order to obtain the required data.
  - Using the SpaceX API we scrapped the booster data, launch site data, payload data and the core data.
  - By the help of beautifulsoup we scrapped the historical launch records of Falcon 9.
  - The response content we obtained was transformed into a useful pandas data frame using Json normalize().

Data Collection – SpaceX API



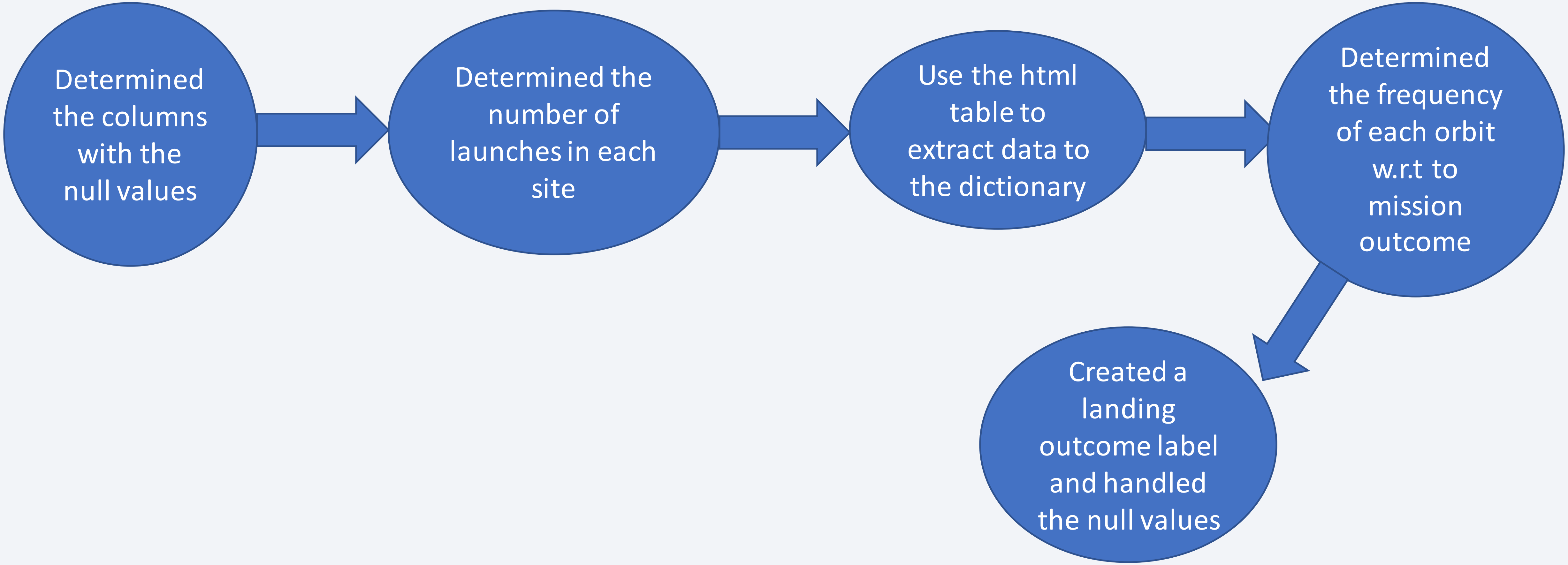
Github URl: https://bit.ly/3pXtXDU

Data Collection - Scraping



Github URL: https://bit.ly/3t5ONCE

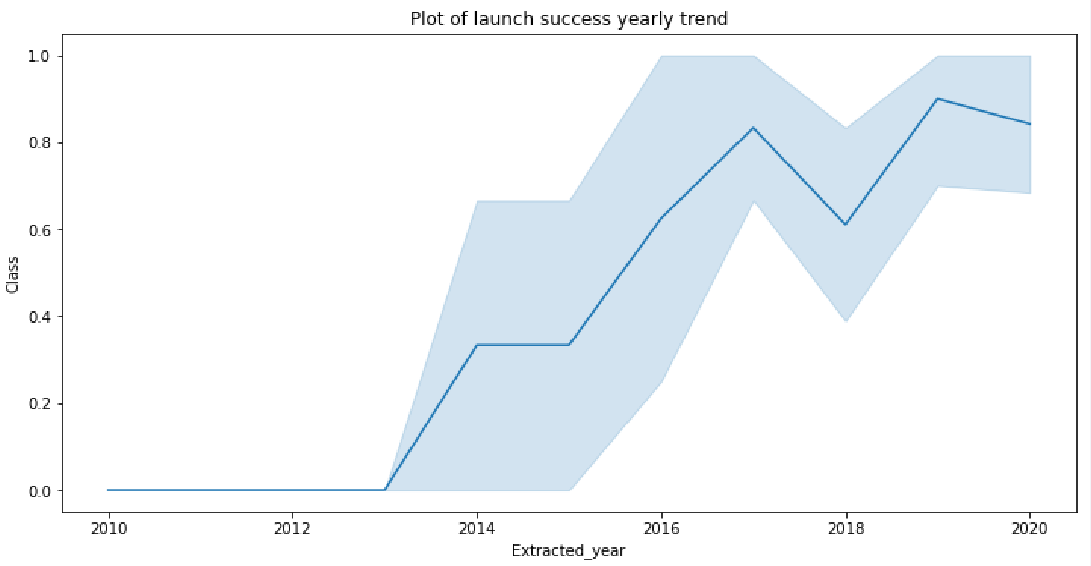
Data Wrangling

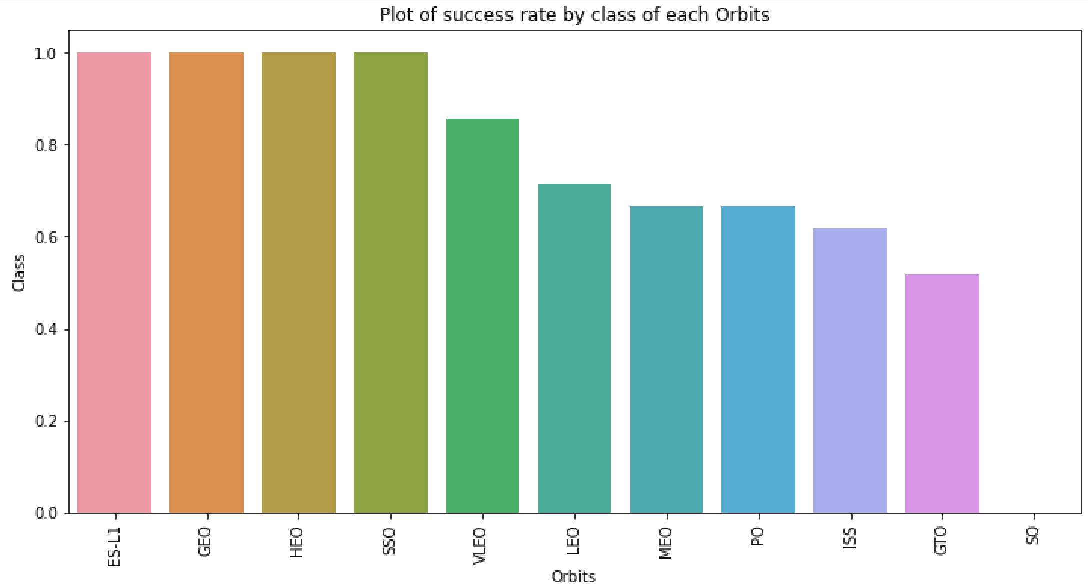


Github URL : https://bit.ly/3tO5pOA

EDA with Data Visualization

- EDA was done among launch site, flight number, payload mass, orbits, class etc.
- Scatter Charts, Line Charts and bar Charts were made to find out the relationship between several variables.





Github URL: [Data Visulization]

EDA with SQL

- We loaded the SpaceX dataset into IBM DB2 Interface.
- Queried the necessary data using SQL magic function in Jupyter notebook.
- Information such as unique launch sites, average payload mass carried, number of successful and failed launches we found out.

Github URL : https://bit.ly/35OgKpO

Build an Interactive Map with Folium

- Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.
- This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.
- Map markers were added to find an optimal launch site.
- Following geographical patterns about launch sites are found:
- Are launch sites in close proximity to railways? Yes
- Are launch sites in close proximity to highways? Yes
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

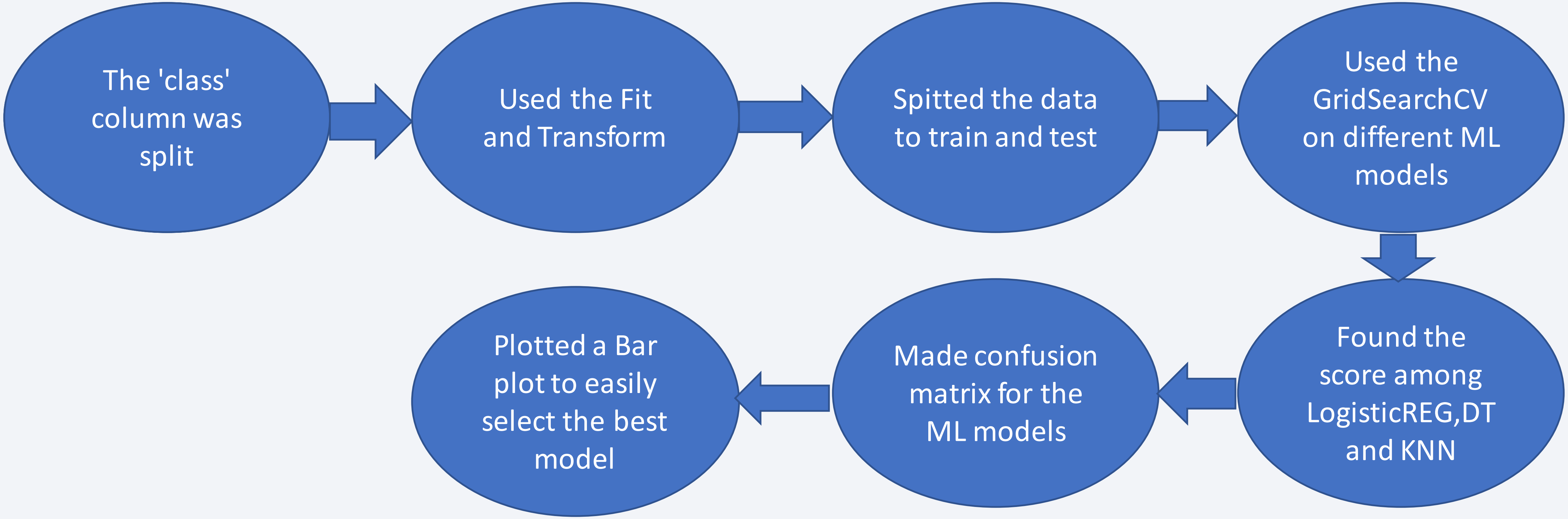
Github URL : https://bit.ly/3i35zwe

Build a Dashboard with Plotly Dash

- We have built an interactive dashboard with Plotly dash
- Pie chart can be selected to show the distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.
- Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.
- The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category.

Github URL : https://bit.ly/3i2S67m

Predictive Analysis (Classification)



Gitub URL : https://bit.ly/3i2BGvL

Results

The key insights we have found out so far are;

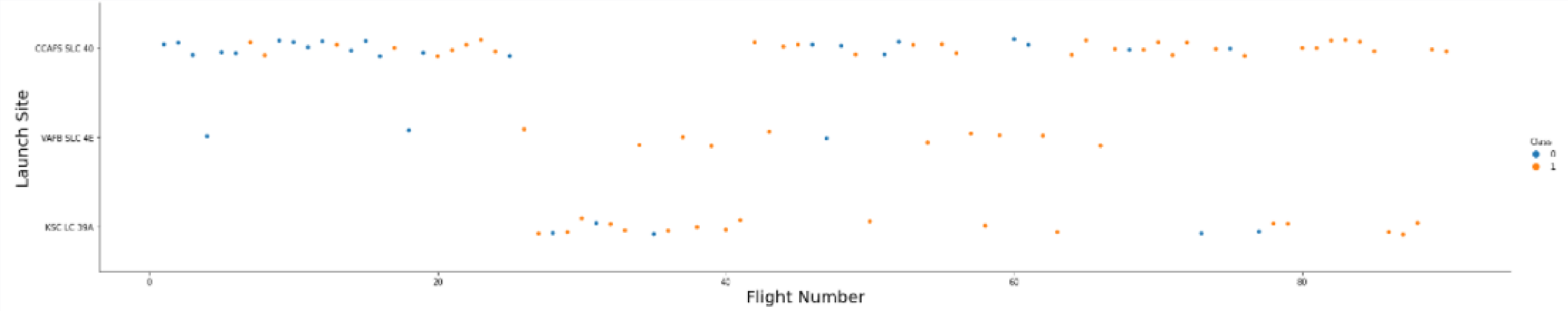
- Light payloads perform better than heavier payloads .
- KSC LC 39A has the highest number of successful launches.
- The best rate of success is in the orbits – GEO,HEO,SSO,ES.
- We also leaned the ideal location for a launch site, closer to coast and near the equator.

Section 2

Insights drawn from EDA

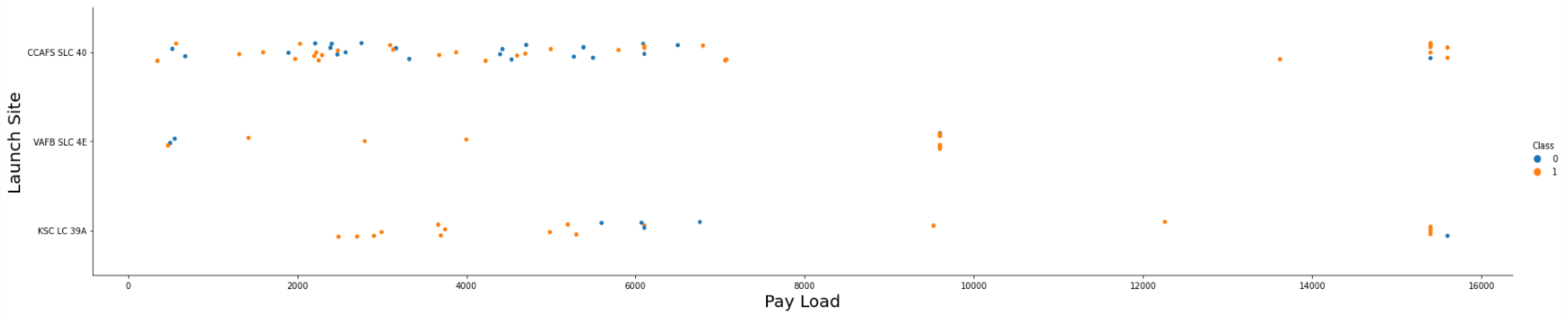
Flight Number vs. Launch Site

- Orange represent successful and blue shows marks shows the failed launches
- The success of the operations has increased over time as more and more flights are tested and conducted.



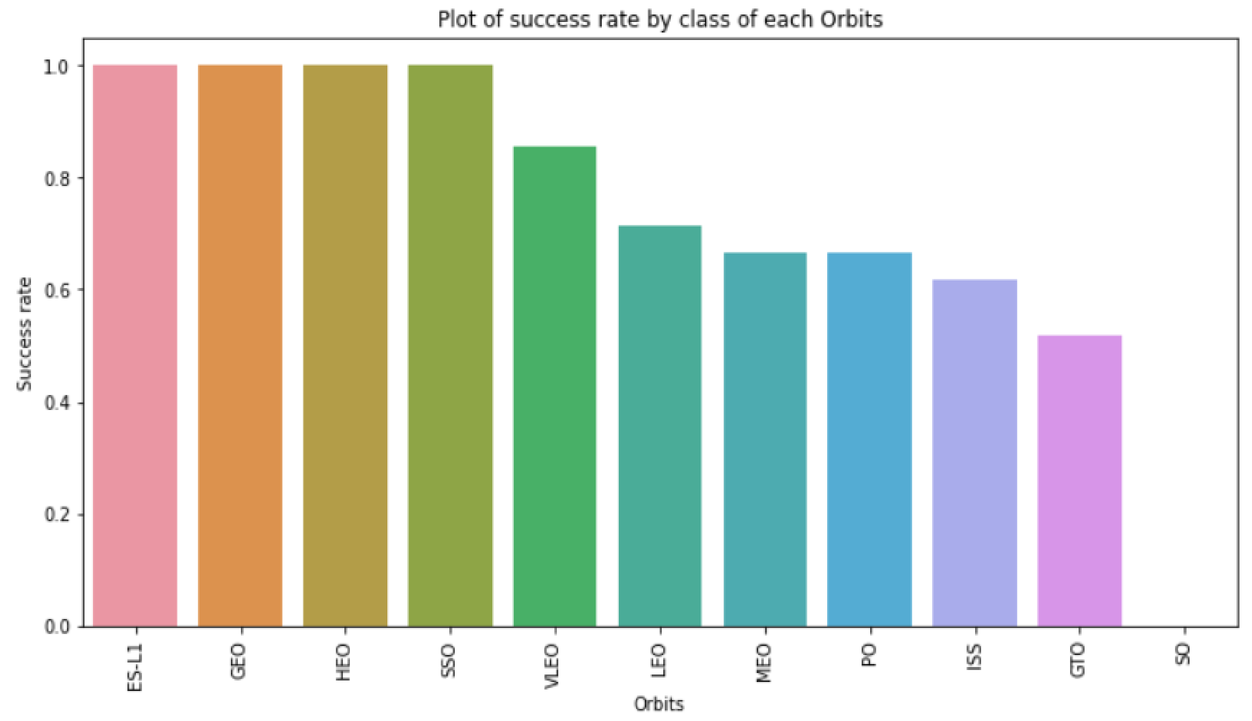
Payload vs. Launch Site

- Orange represent successful and blue shows failed launches.
- Most of the payload are under 6500 Kg and most of them are launched from CCAFS SLC 40.
- The success rate does not really look like depend on the place of launch.



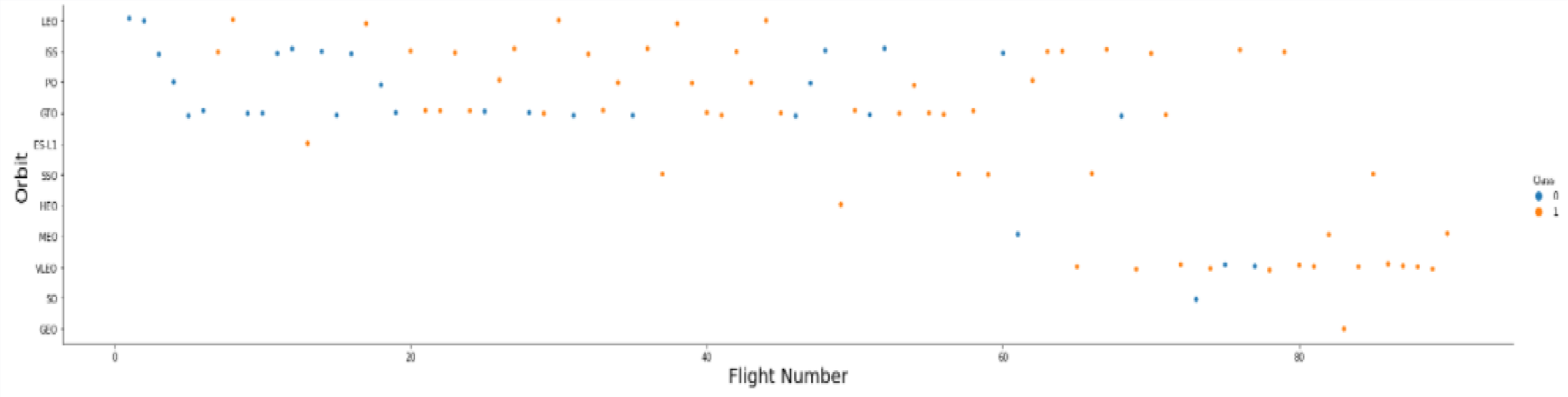
Success Rate vs. Orbit Type

- The orbits ES-L1,GEO,HEO and SSO have an 100 percentage success rate.
- Other orbits excluding SO have a success rate between 50 and 80 percentage.
- SO's zero success rate is due to the failure of a single attempt.



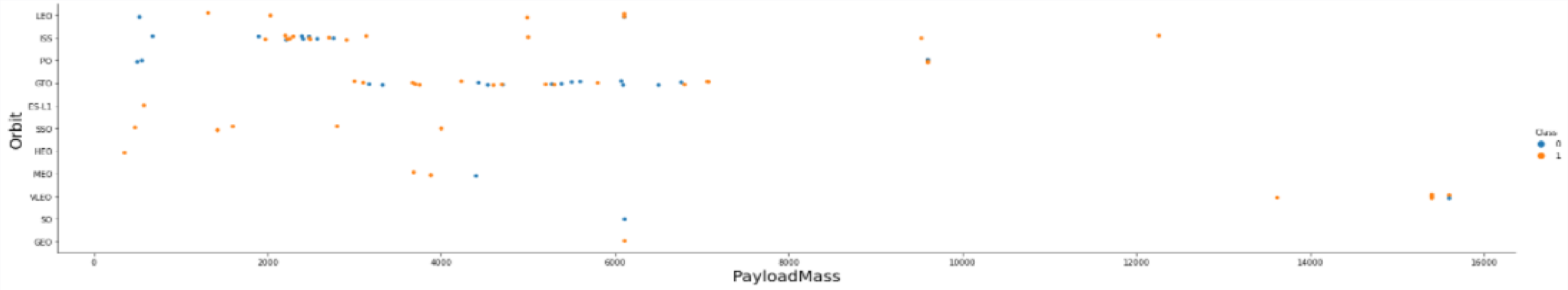
Flight Number vs. Orbit Type

- Orange represent successful and blue shows failed launches.
- The rate of success has increased over time when more flights are tested.
- The launches to VLEO has increased recently.
- LEO,ISS,PO,GEO,VLEO has been the preferred orbits.



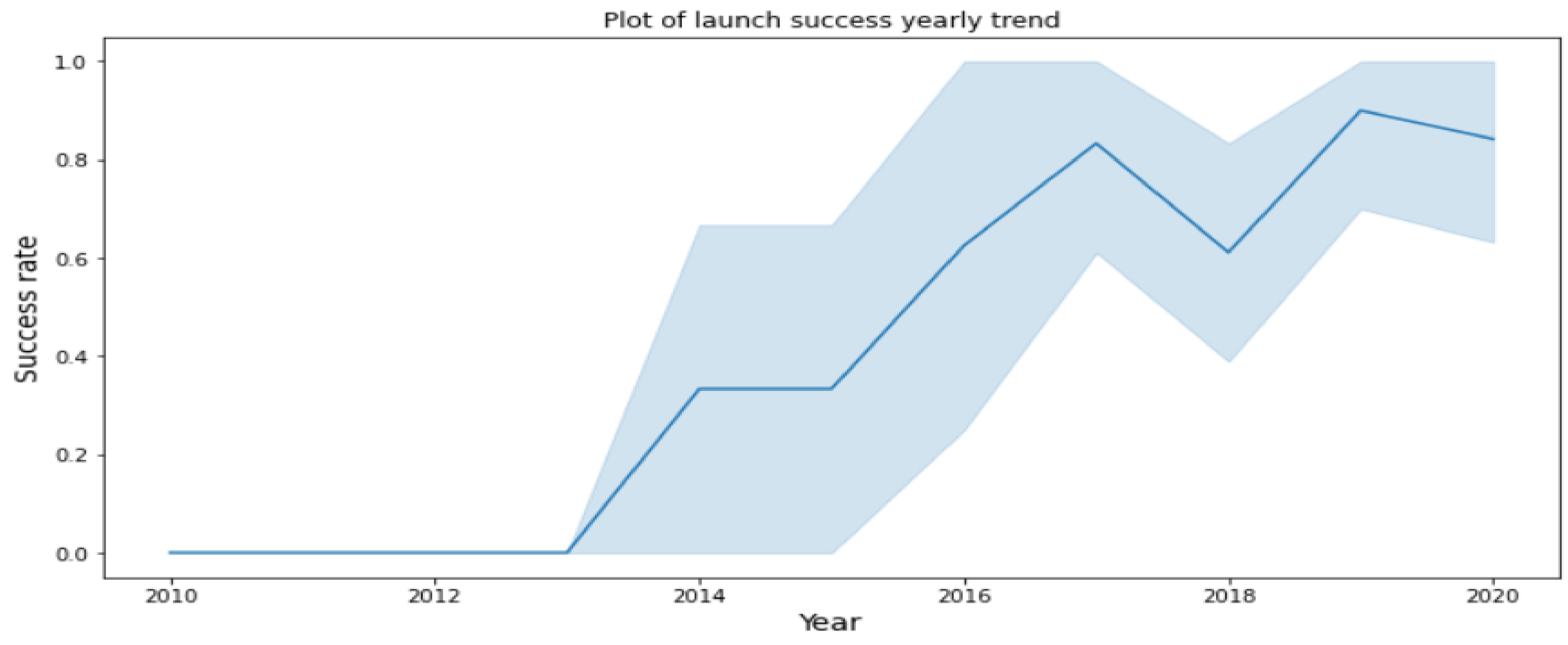
Payload vs. Orbit Type

- Orange represent successful and blue shows failed launches.
- Heavier payloads tends to be more successful.
- But GTO's success rate does not look like it depends on the weight of payload
- LEO and SSO seems to be the preferred orbit for lighter payloads



Launch Success Yearly Trend

- The rate of success doubled from 40 to 80 percentage in a span of 3 years from 2014
- The success rate was at an all-time high at 2019
- Now the rate seems to be stabilizing around 80 percentage at 2020.



All Launch Site Names

The unique launch sites are

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

```
In [23]: %%sql
         SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL

 * ibm_db_sa://jjy33248:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.

Out[23]:
```

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

Launch Site Names Begin with 'CCA'

In [22]:
```
%%sql
SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5
```

```
 * ibm_db_sa://jjy33248:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[22]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

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

In [21]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass_kg FROM SPACEXTBL
WHERE CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://jjy33248:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[21]:

total_payload_mass_kg
45596

- This query shows the total payload mass in Kg when NASA was the customer.
- We used the sum function on 'PAYLOAD_MASS_KG' to find out its total.
- We used the 'WHERE' clause the specify the customer was NASA.

Average Payload Mass by F9 v1.1

```
In [20]: %%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS avg_payload_mass_kg FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 v1.1'

 * ibm_db_sa://jjy33248:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[20]:

avg_payload_mass_kg
2928

- We used the average function to find out the mean of payload
- And the booster version was set as 'F9 v1.1' using the WHERE clause.

First Successful Ground Landing Date

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

In [19]:
```
%%sql
SELECT MIN(DATE) AS first_successful_landing_date FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://jjy33248:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[19]:

first_successful_landing_date
2015-12-22

- We used the combination of min function and WHERE clause as 'Success (ground pad)' to find the first successful landing date.

Successful Drone Ship Landing with Payload between 4000 and 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

In [18]:
```
%%sql
SELECT BOOSTER_VERSION FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (drone ship)'AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

```
 * ibm_db_sa://jjy33248:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[18]:

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

- We used the combination of a WHERE clause and AND operator to find the successful drone ship landing where the payload was between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [17]:
```
%%sql
SELECT MISSION_OUTCOME, COUNT(*) AS total_number FROM SPACEXTBL
GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://jjy33248:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[17]:

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Here we used the group by statement to find out the successful and failed missions.
- From the output it is clear that 99 of the flights have been a success.

Boosters Carried Maximum Payload

In [16]:
```
%%sql
SELECT DISTINCT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * ibm_db_sa://jjy33248:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[16]:

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

- The maximum payload SpaceX takes is 15600 Kg
- They use several versions of F9 B5 to carry their heaviest load.

2015 Launch Records

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

In [15]:
```
%%sql
SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND YEAR(DATE) = '2015'
```

```
 * ibm_db_sa://jjy33248:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[15]:

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- There were 2 failed attempts to land on a drone ship
- Both those failures occurred when the launch site was CCAFS LC-40

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

In [14]:
```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS total_number FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY total_number DESC
```

```
 * ibm_db_sa://jjy33248:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[14]:

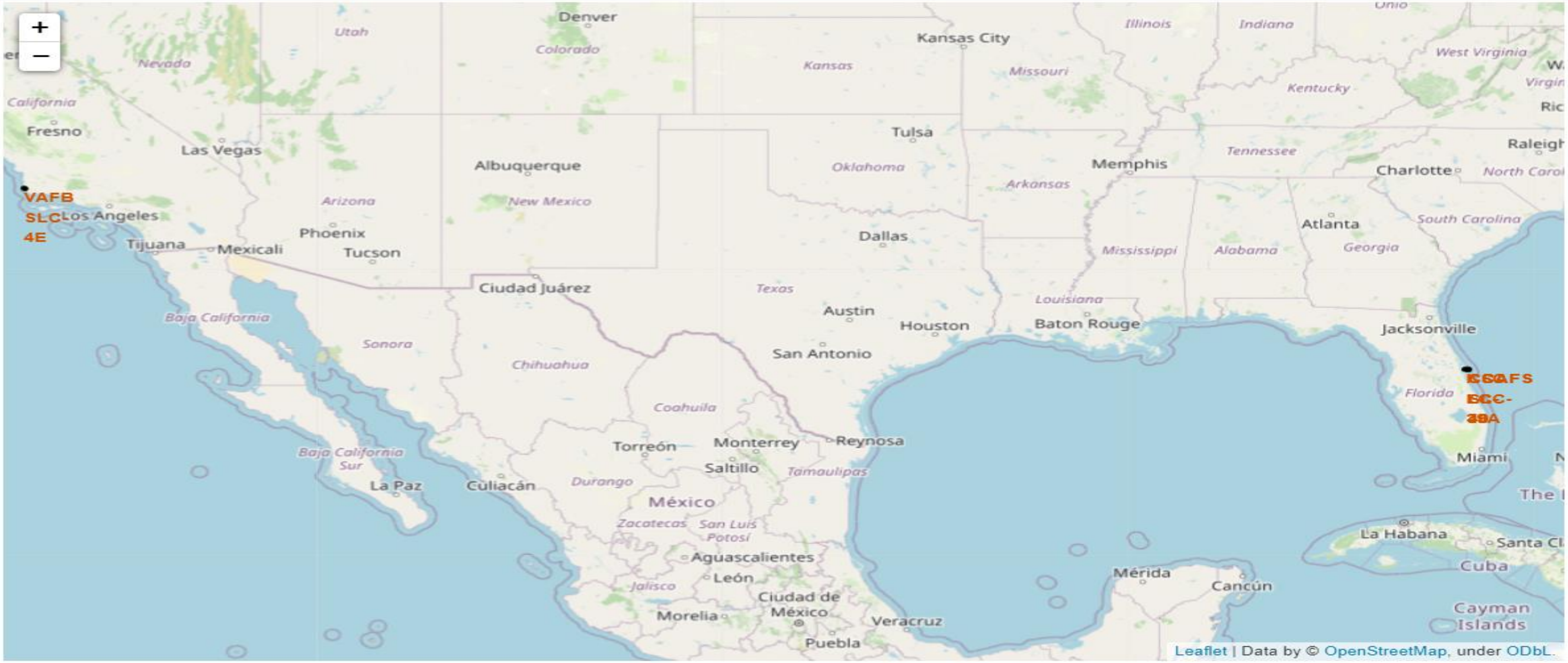
landing__outcome	total_number
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- 8 attempts the land were successful including landing in drone ships and landing in ground pads.
- We used the DESC keyword to sort the records.

Section 4

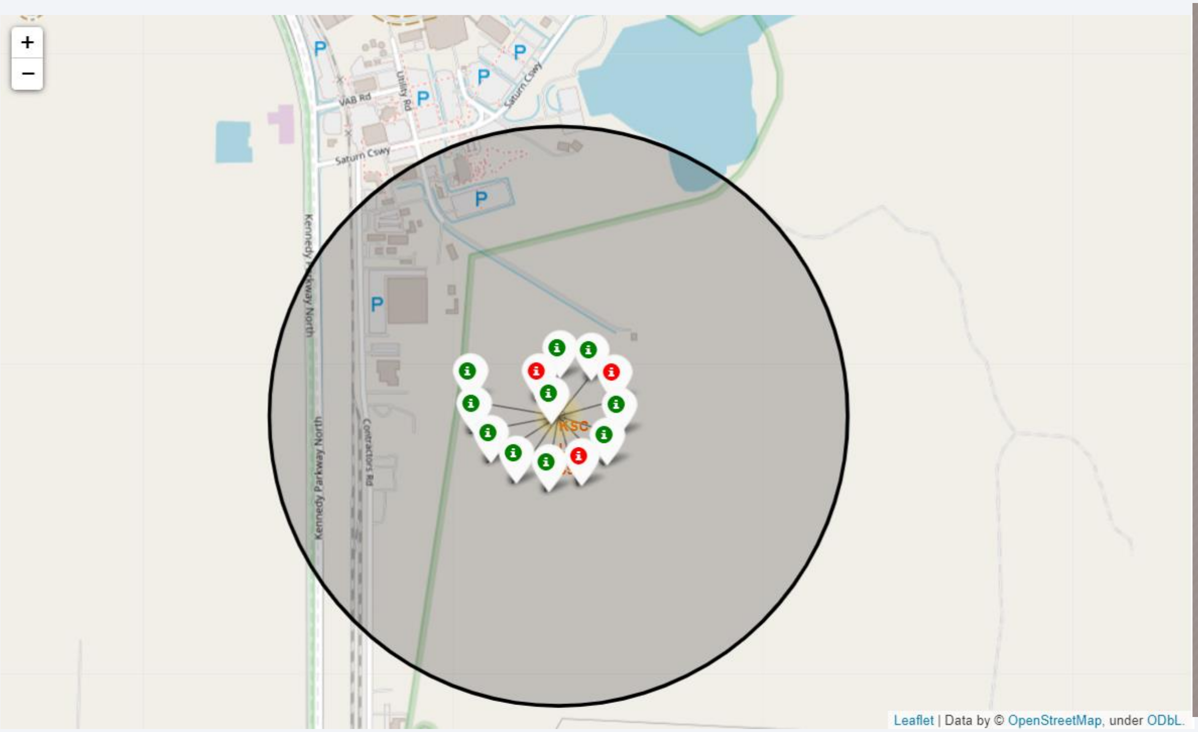
Launch Sites Proximities Analysis

Launch Site Locations

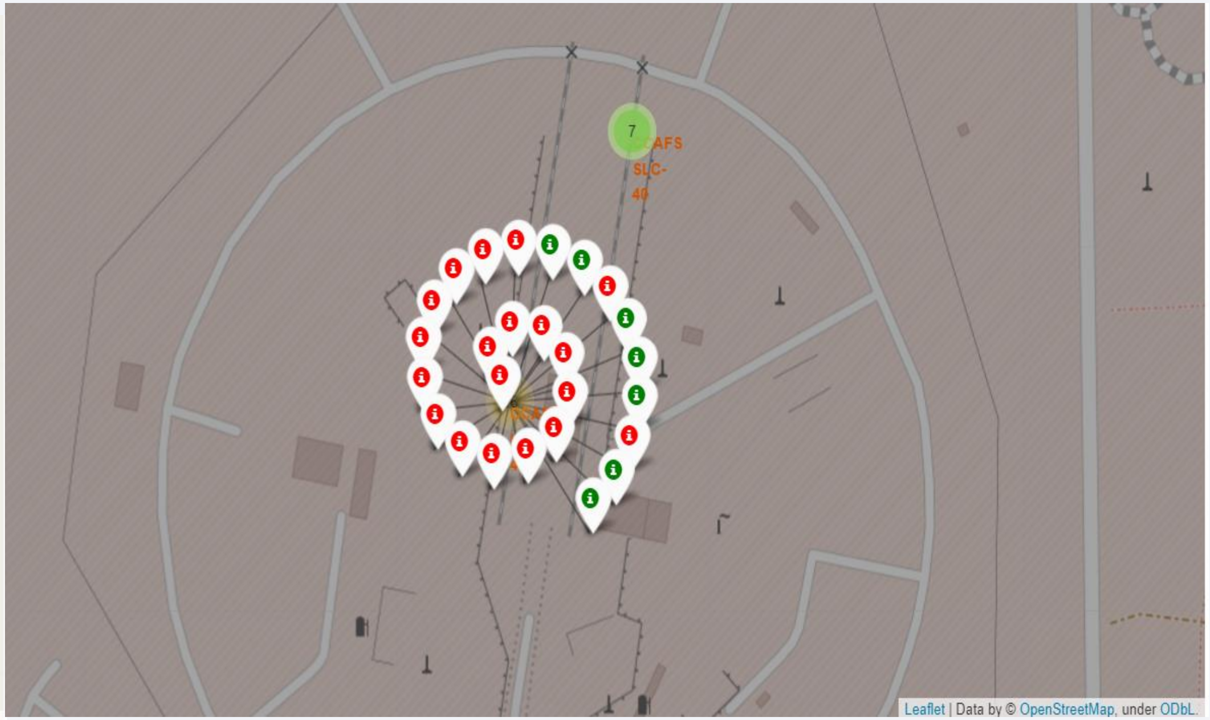


We can see that both launch sites are near coastal areas for better chance of a successful launch and for the safety of the general population.

Successful and Failed Launches in each Site



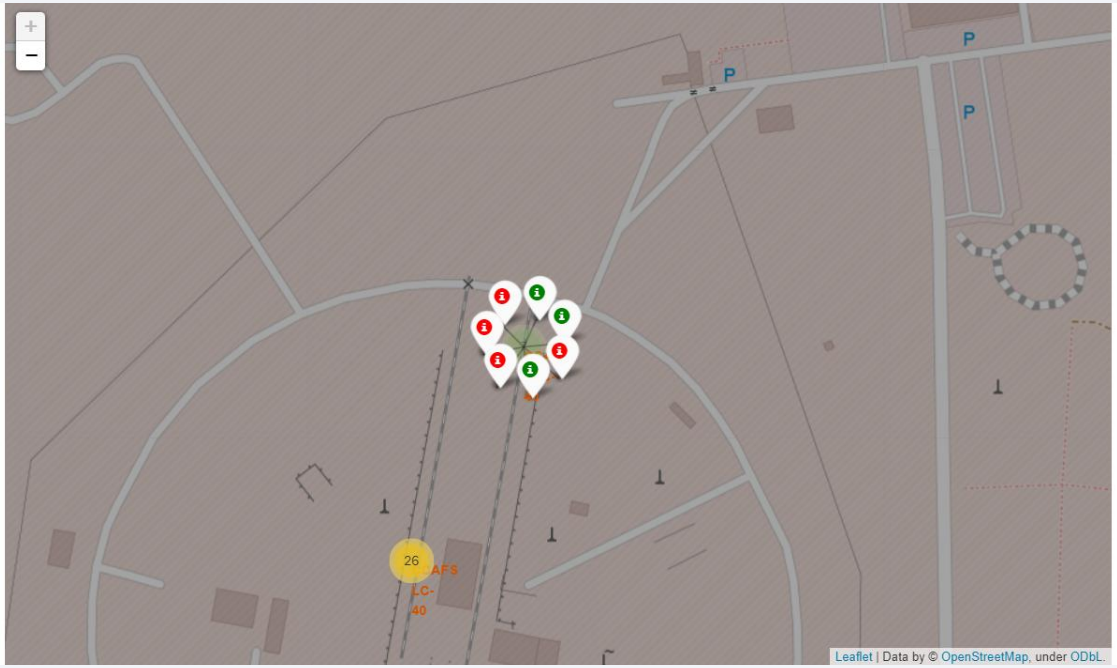
KSC LC-39 A



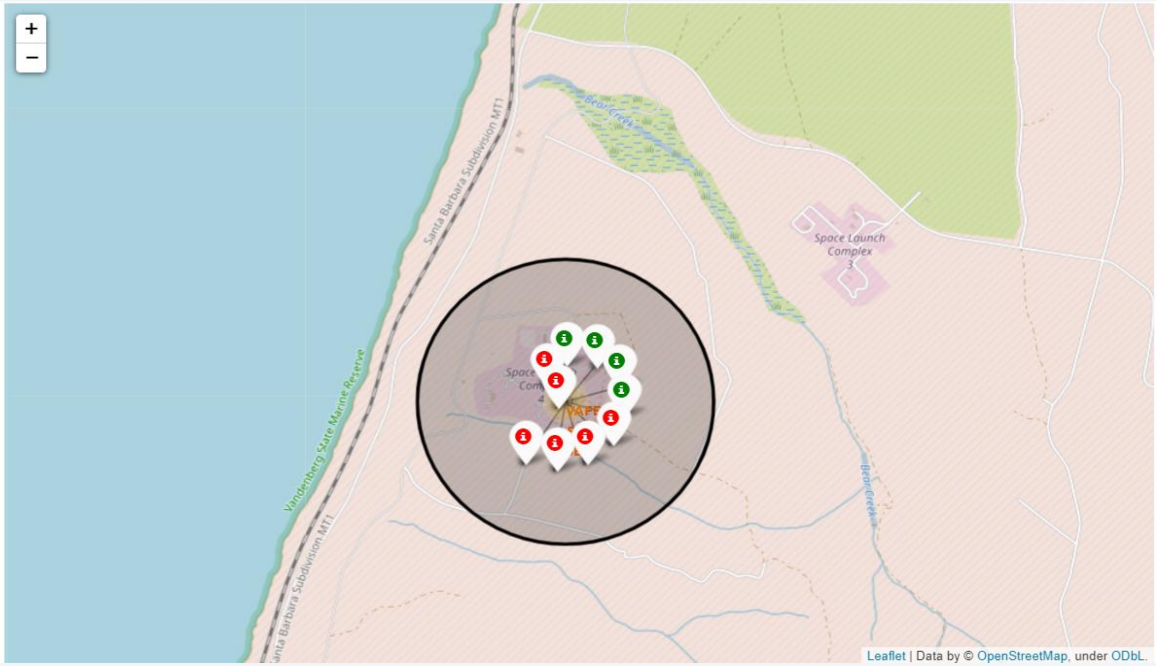
CCAFS LC-40

Green marker indicate successful launch and red marker shows the failed launch

Successful and Failed Launches in each Site



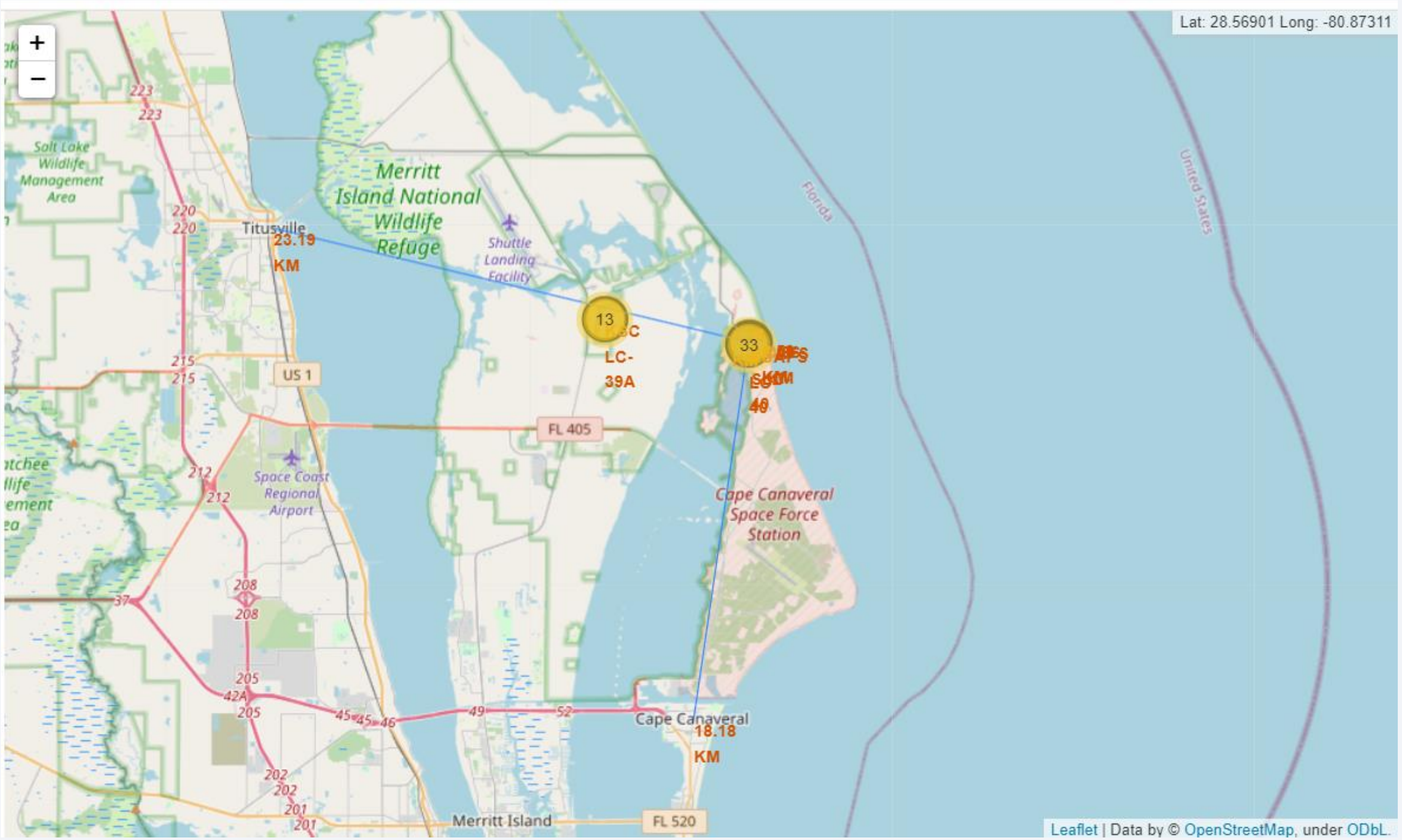
CCAFS SLC-40



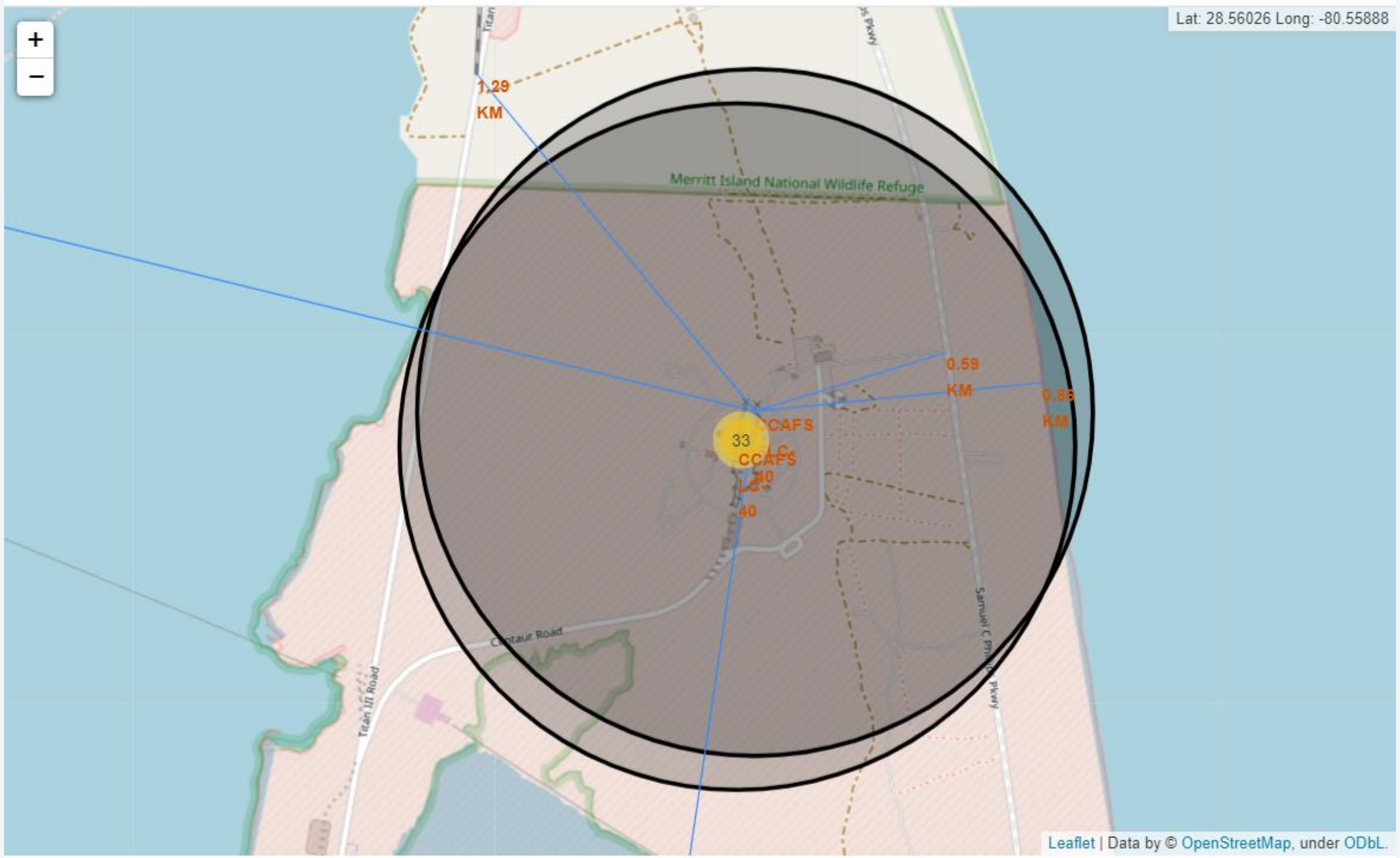
VAFB SLC-4E

Green marker indicate successful launch and red marker shows the failed launch

Proximity of Landmarks near Launch Site



Proximity of Landmarks near Launch Site

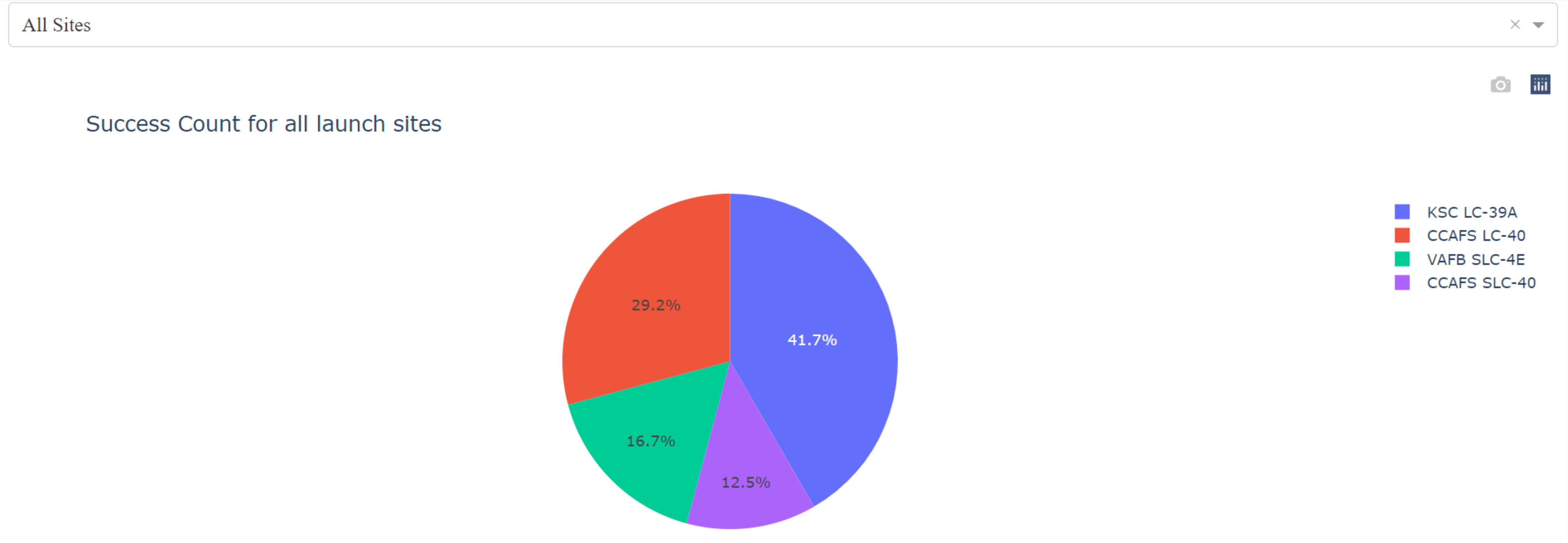


Both the CCAFS launch sites are very close to railway lines and coastal line. This helps In the movement of required goods. All the launch sites also tend to avoid landmarks where general population visit the most in order to avoid any threats to the public.

Section 5

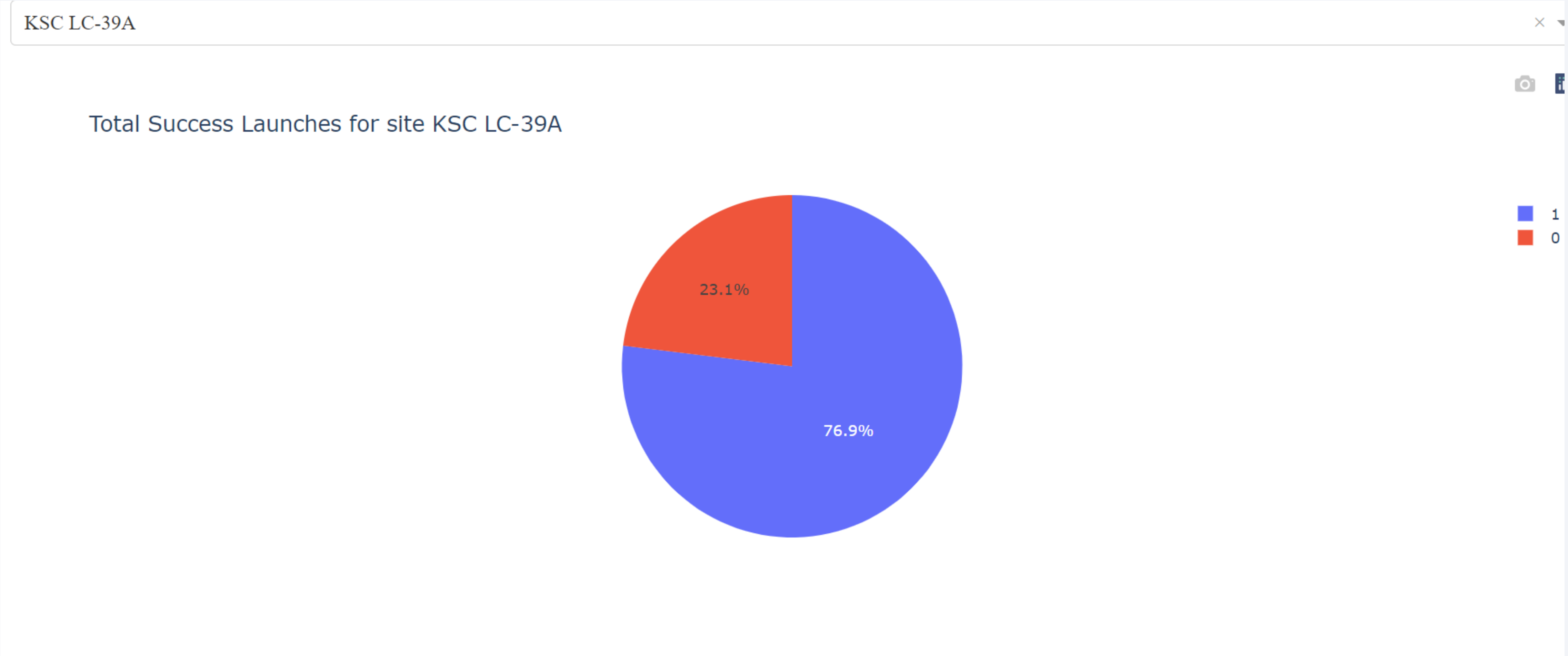
Build a Dashboard with Plotly Dash

Success Percentage of each Launch Site



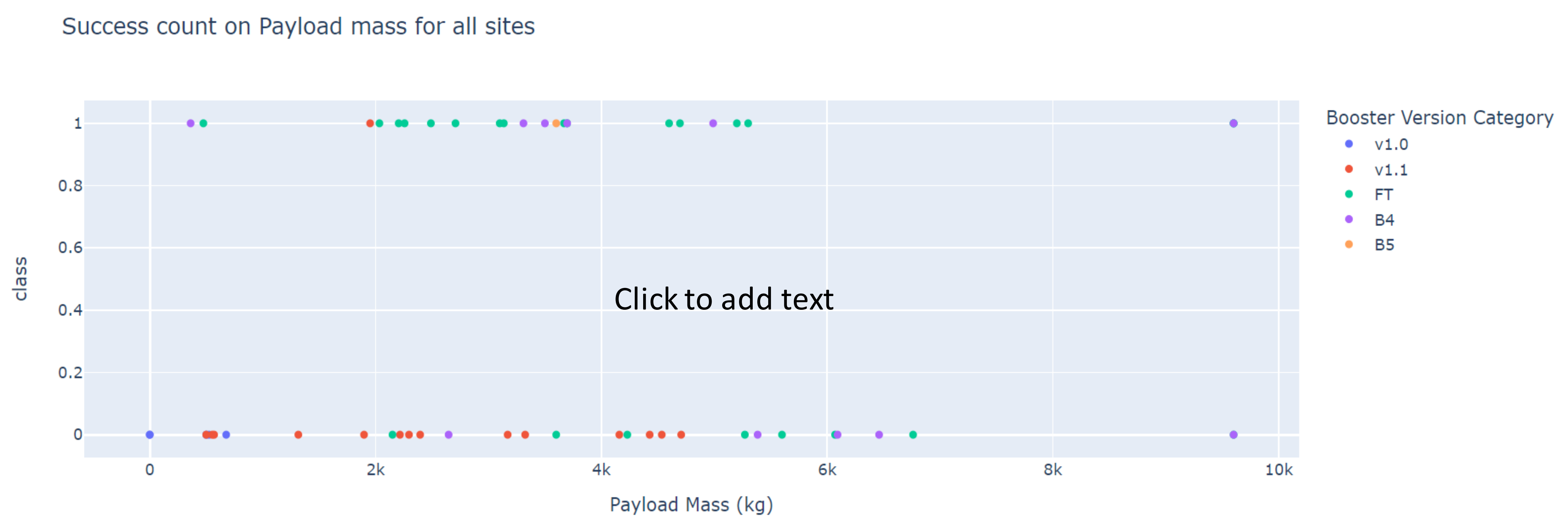
KSC LC-39A has the best success percentage of nearly 42 percentage.
Both CCAFS LC and SLC have a combined success rate of around 42 percentage.
VAFB SLC-4E have a low success percentage of 16.7 percentage

Most Successful launch Site



KSC LC-39 A is the most successful launch site with a success percentage 77.

Success per Payload Category



- The class 1 (successful launches) are more for payloads less than 4000 kg.
- But the failure rate is equally divided among light and heavy payloads.
- FT and B4 are the best boost version to launch heavy payloads.
- Falcon 9 v1.1 has the most records of failure.
- FT booster does well in medium payloads(between 2000 and 6000Kg)

Section 6

Predictive Analysis (Classification)

Classification Accuracy

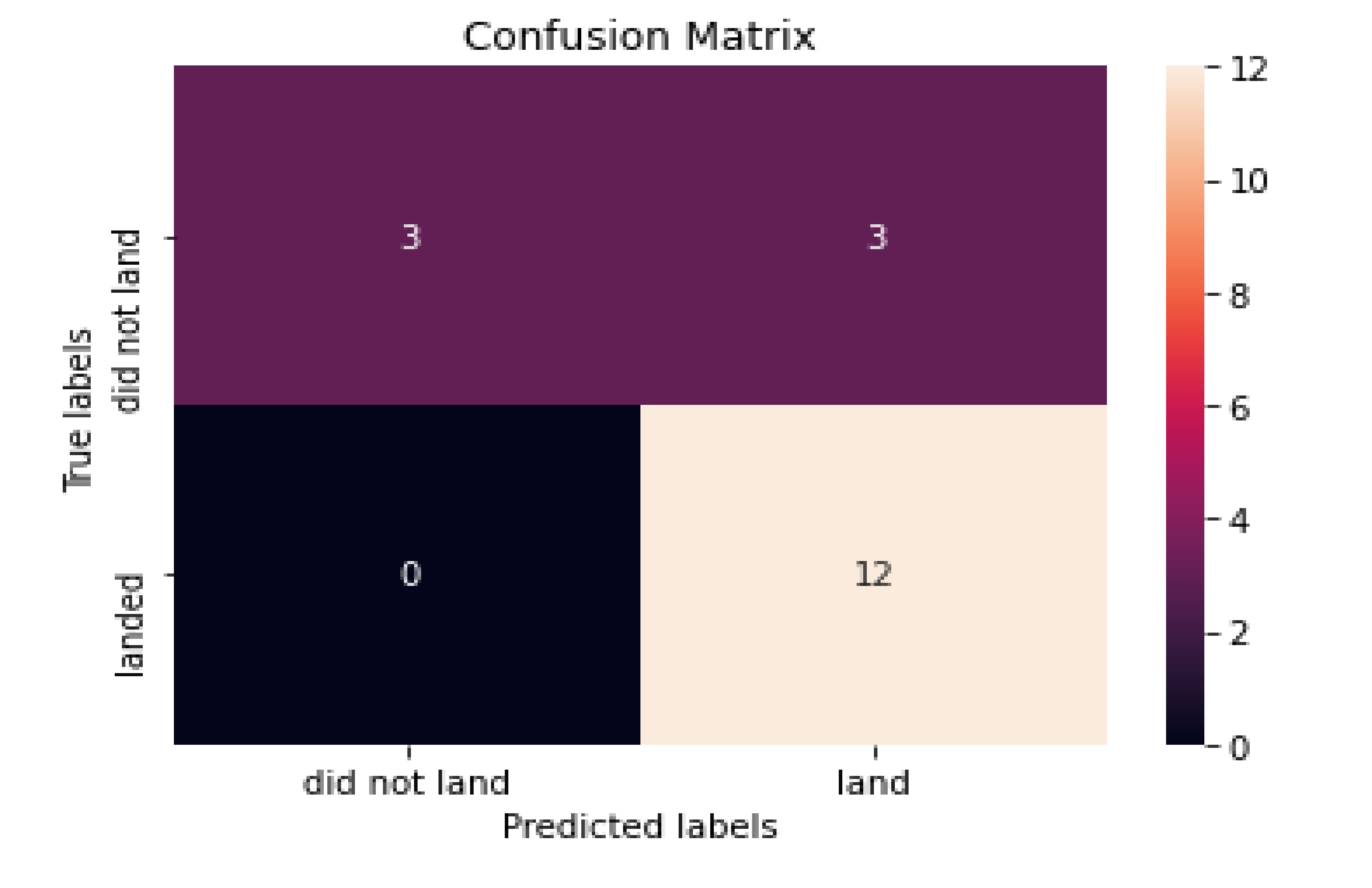
In [30]:
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8732142857142856
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'random'}
```

- The decision tree has better classification accuracy.
- Other models that is, KNN, SVM, and Logistic Regression has a slightly lower accuracy which is around 82 to 83.

Confusion Matrix

- From the confusion matrix we can predict a successful landing
- There are 12 successful landings predicted when the label was true.
- The confusion matrix is similar for the all the models we have tested.



Conclusions

- The success of the launch will increase with time as more and more flights are conducted
- Probability of success is higher if we launch the flights into orbits such as GEO,HEO,SSO and ES-L1.
- Lighter payloads have a better chance of success compared to heavier ones.
- There is a better chance of success if we launch from KSC LC 39A
- Closer proximity to the equator can increase the chance of a successful flight.
- Closer proximity to railways and coastal lines can help in transportation of goods and personnel.
- Launch sites should be built away from a general population.
- Using Falcon 9 rockets gives a better chance of landing the first stage back and thus reducing the overall cost.
- Using a FT booster is best choice if we want to launch with a payload between 2000 and 6000 Kg.

Thank you!