IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ravikant Bhurelal Hatwar
30.05.2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - 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 Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

  Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars while other providers cost upward of 165 million dollars each. Space X can reuse the first stage and hence much of the savings can be done. To determine the cost of launch we have to determine if the first stage will land successfully. The objective of the project is to create a machine learning pipeline to predict if the first stage of Space X will land successfully and to evaluate the viability of the new company to compete with Space X.

- 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.
  - What operating conditions needs to be in place to ensure a successful landing program.
  - Where is the best place to make launches?

Section 1

Methodology

Methodology

Executive Summary

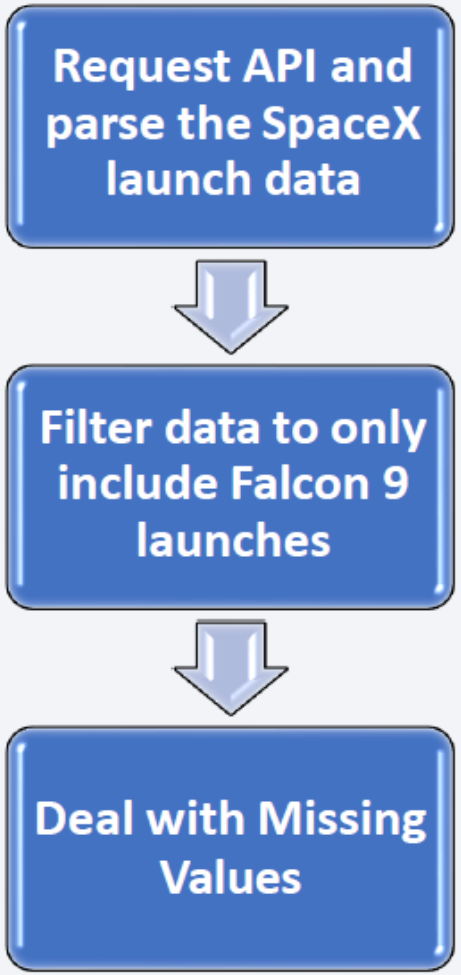
- Data collection methodology:
  - Data of Space X was obtained from following sources.
    1. Space X API
    2. Web Scrapping
- Perform data wrangling
  - Collected data was enriched, One-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The data was collected using various methods
  - Data collection was done using get request to the SpaceX API.
  - Next, we decoded the response content as a Json using .json() function call and turn it into a pandas data frame using .json_normalize().
  - We then cleaned the data, checked for missing values and fill in missing values where necessary.
  - In addition, we performed web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup.
  - The objective was to extract the launch records as HTML table, parse the table and convert it to a pandas dataframe for future analysis.

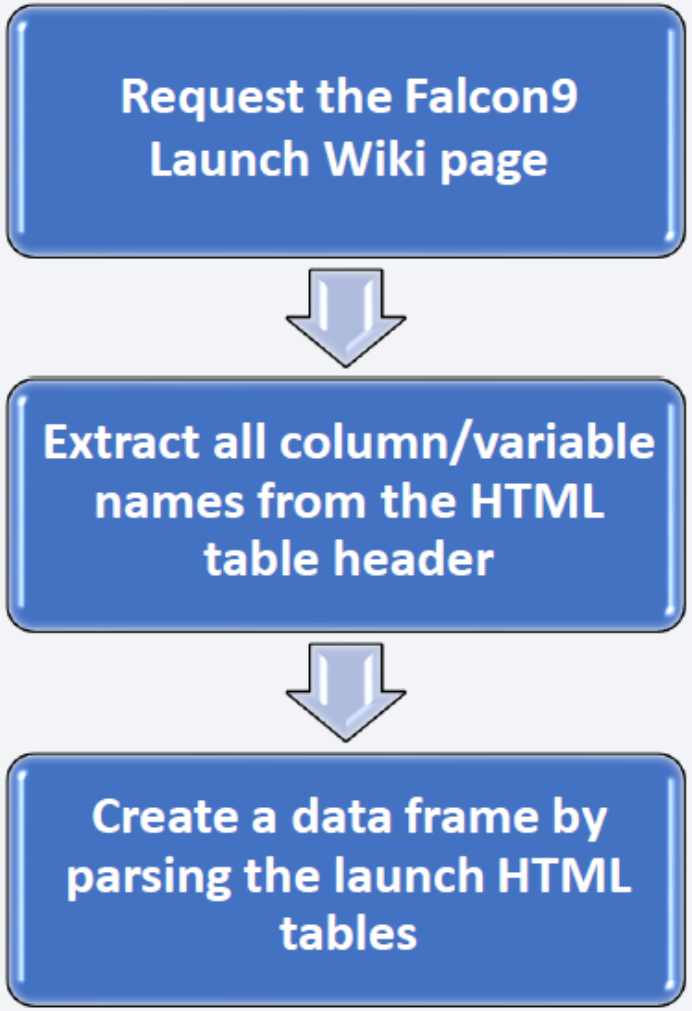
Data Collection – SpaceX API

- The get request has been used to the SpaceX API to collect data. Requested data has been cleaned and some basic data wrangling and formatting done.
- The link to the notebook is https://github.com/RavikantHatwar/CapstoneProject/blob/main/SpaceXCapstoneProject/jupyter-labs-spacex-data-collection-api(1).ipynb



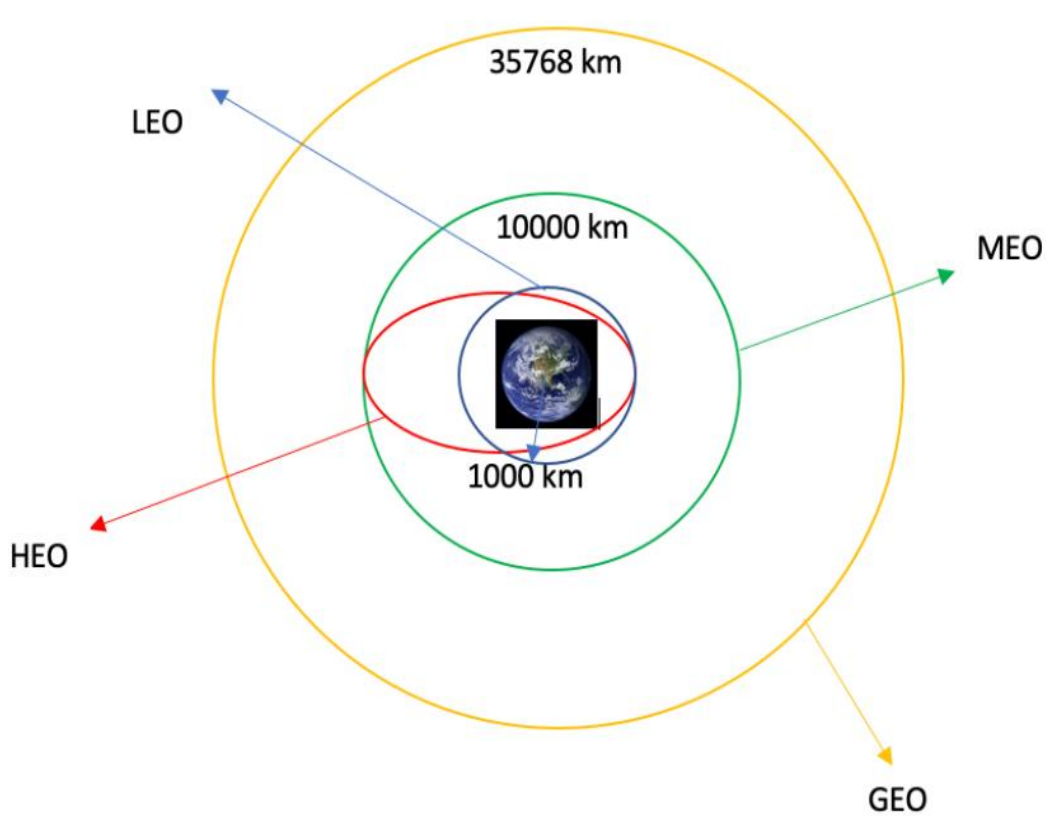
Data Collection - Scraping

- We applied web scrapping to webscrap Falcon 9 launch records with BeautifulSoup
- We parsed the table and converted it into a pandas dataframe.
- The link to the notebook is https://github.com/RavikantHatwar/CapstoneProject/blob/main/SpaceXCapstoneProject/jupyter-labs-webscraping.ipynb



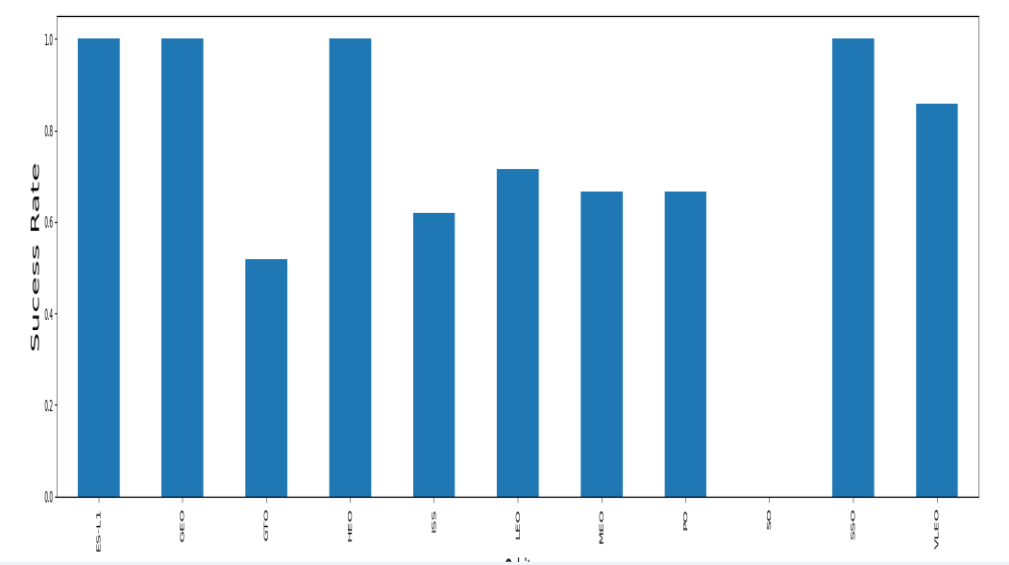
Data Wrangling

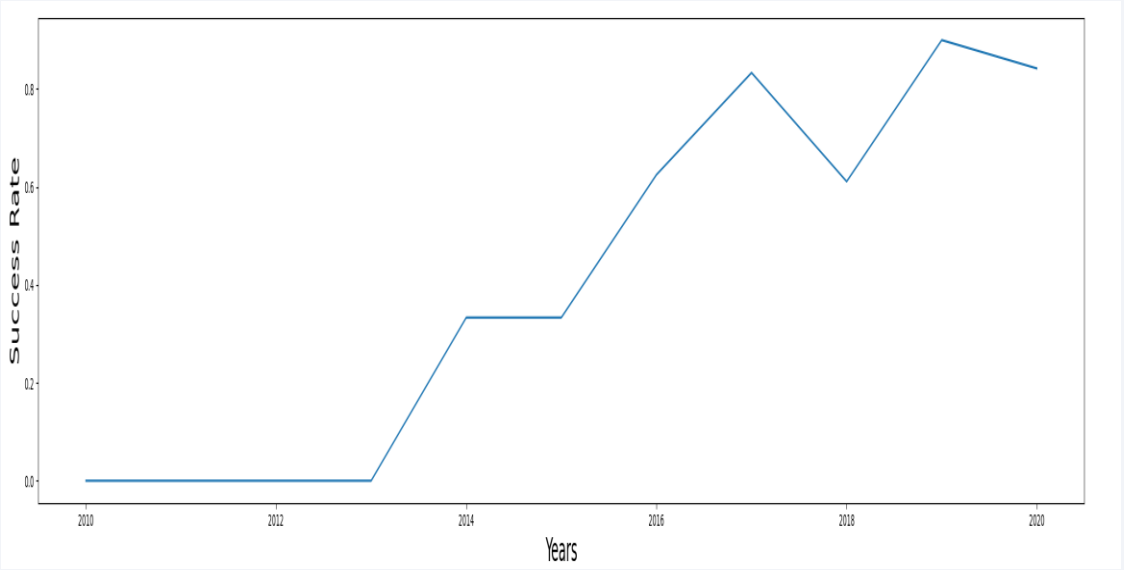
- Exploratory data analysis has been preformed to determined the training labels.
- Number of launches at each site and the number and occurrence of each orbits has been calculated.
- After creating landing outcome label from outcome column data (results) has been exported to the csv file.
- The link to the notebook is https://github.com/RavikantHatwar/CapstoneProject/blob/main/SpaceXCapstoneProject/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- We analyze the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly





- The link to the notebook is https://github.com/RavikantHatwar/CapstoneProject/blob/main/SpaceXCapstoneProject/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Following SQL queries were performed

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

```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

```
sql SELECT SUM (PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER='NASA (CRS)'
```

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

```
sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE booster_version LIKE 'F9 v1.1%'
```

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

```
sql SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

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

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING_OUTCOME = 'Success
```

List the total number of successful and failure mission outcomes

```
sql SELECT MISSION_OUTCOME, COUNT(*) FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

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

```
sql SELECT BOOSTER_VERSION,PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACE
```

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

```
sql SELECT substr(DATE,4,2),MISSION_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where substr(Date,7,4)='2015';
```

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
sql SELECT Landing_Outcome FROM SPACEXTBL WHERE DATE BETWEEN '06-04-2010' AND '20-03-2017' ORDER BY DATE DESC;
```

EDA with SQL

- The link to the notebook

  https://github.com/RavikantHatwar/CapstoneProject/blob/main/SpaceXCapstoneProject/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

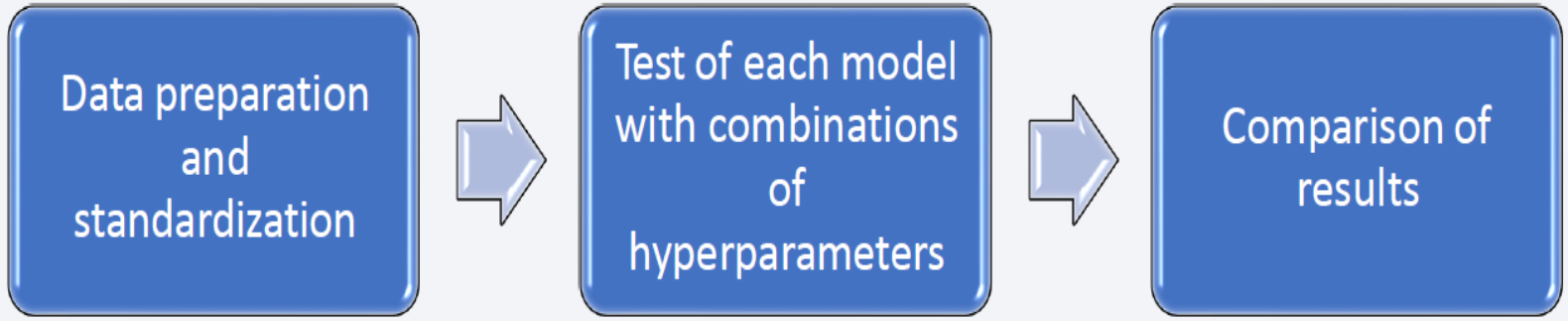
- Map objects such as markers, circles, lines and marker clusters were added to a folium map
  - ➢ Markers to indicate points like launch sites
  - ➢ Circles to indicate highlighted areas around specific coordinates.
  - ➢ Marker clusters to indicates groups of events in each coordinate, like launches in a launch site have relatively high success rate.
  - ➢ Lines to indicate distances between two coordinates and to answer following questions.
    - Are launch sites near railways, highways and coastlines ?
    - Do launch sites keep certain distance away from cities ?
- GitHub URL : https://github.com/RavikantHatwar/CapstoneProject/blob/main/SpaceXCapstoneProject/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- An interactive dashboard with Plotly dash was built and following plots were created.
  - ❖ Pie-charts showing the total launches by a certain sites were plotted.
  - ❖ Scatter chart showing the relationship with Outcome and Payload Mass (Kg) for the different booster version was plotted.
- GitHub URL : https://github.com/RavikantHatwar/CapstoneProject/blob/main/SpaceXCapstoneProject/dash_interactivity.py

Predictive Analysis (Classification)

- Following classification models were modelled and compared, hyper-parameters tuning was done using GridSearchCV, accuracy was the metric used to find the best classification model and model performance was improved using feature engineering and algorithm tuning.
  - Regression
  - Support Vector Machine
  - Decision tree
  - K-nearest neighbour



- GitHub URL : https://github.com/RavikantHatwar/CapstoneProject/blob/main/SpaceXCapstoneProject/IBM-DS0321EN-skillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results

  I. As the flight number increases, the first stage is more likely to land successfully.

  II. More massive the payload, the less likely the first stage will return

  III. Different launch sites have different success rates CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

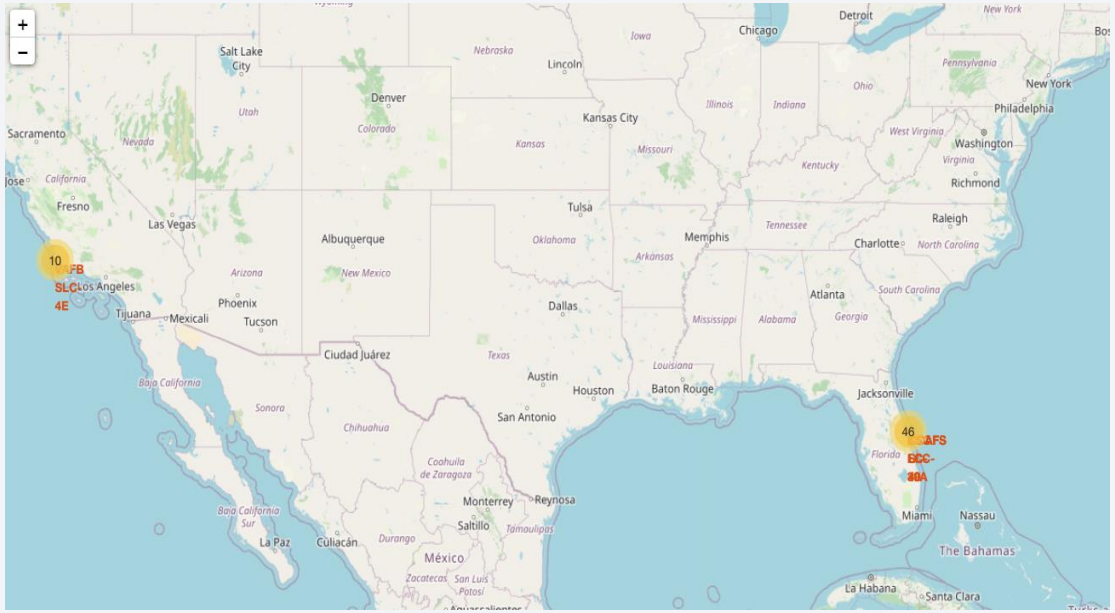
  IV. for the VAFB-SLC launch site there are no rockets launched for heavypayload mass(greater than 10000).

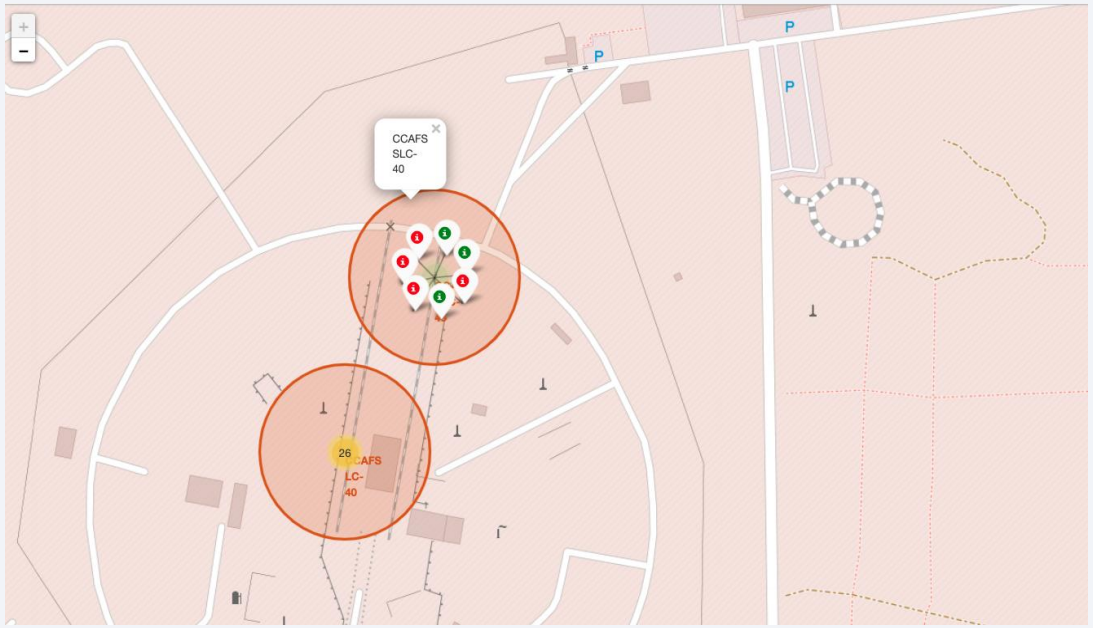
  V. With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

  VI. Success rate since 2013 kept increasing till 2020.

Results

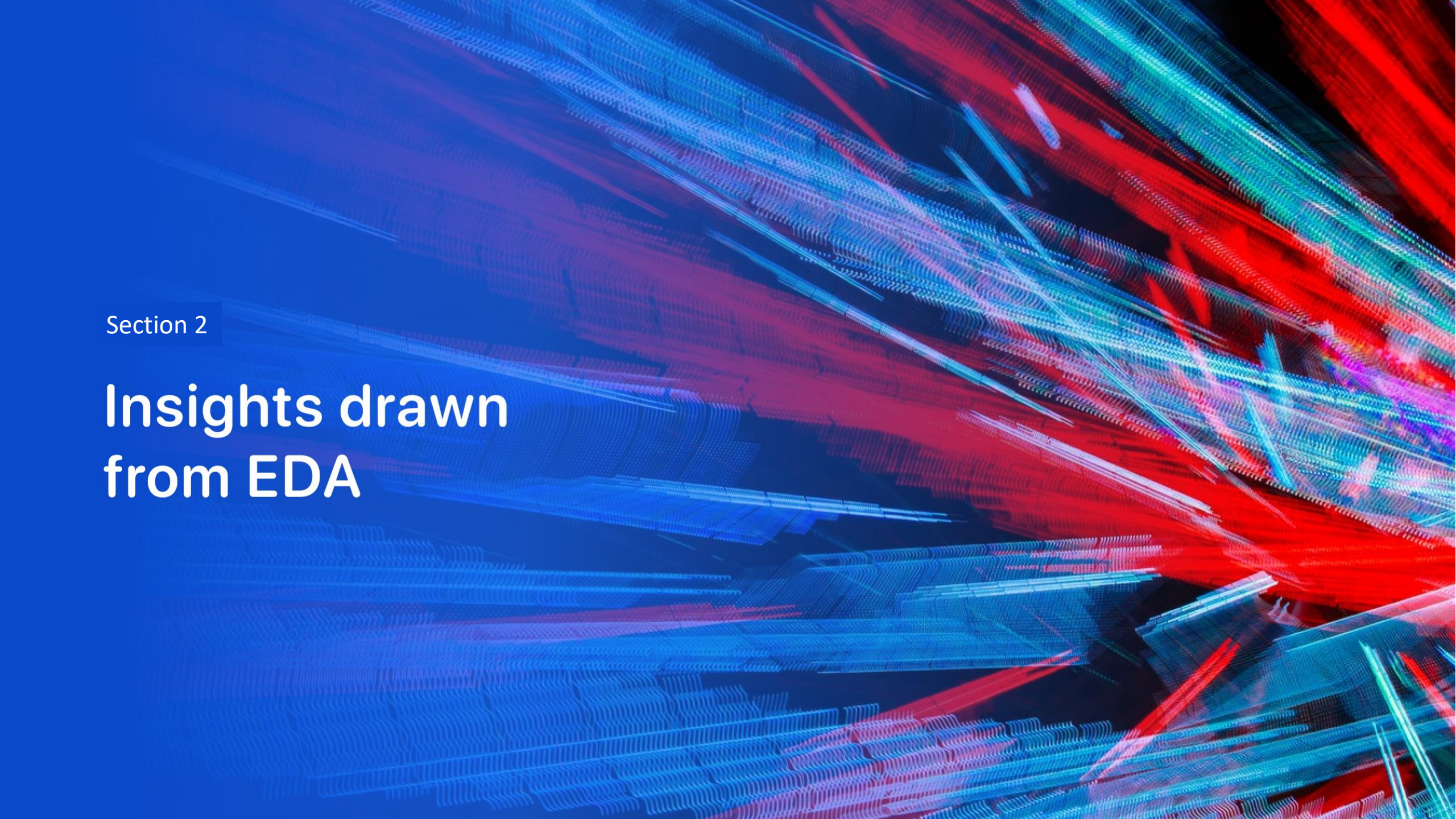
- Interactive analytics demo in screenshots
  I. Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.
  II. Most launches happens at east cost launch sites.



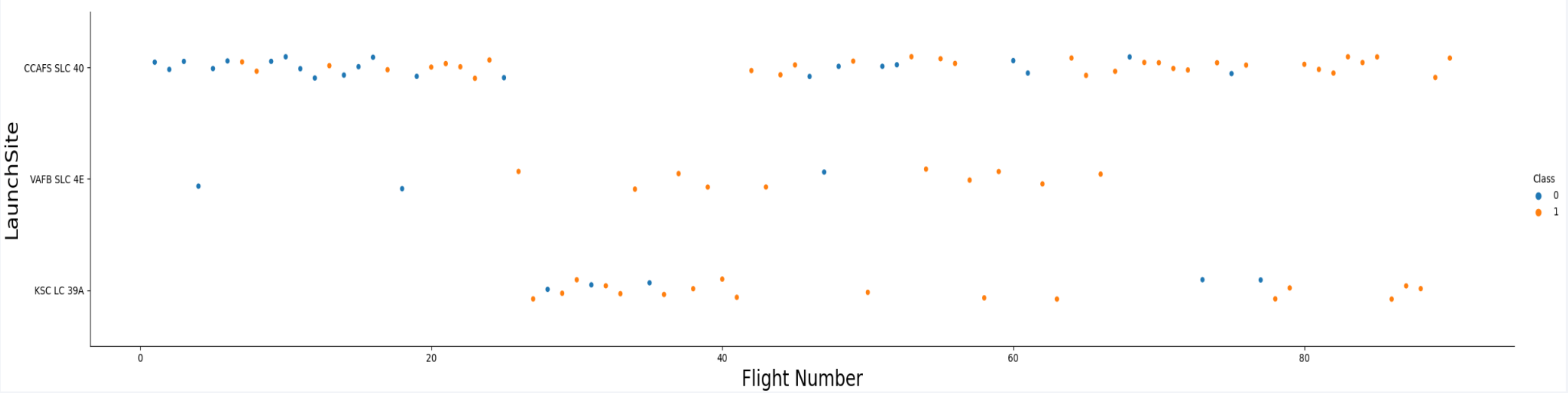


Results

- Predictive analysis results
  - From Predictive Analysis we conclude that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% for training data and accuracy for test data observed over 94%.


Section 2
Insights drawn from EDA

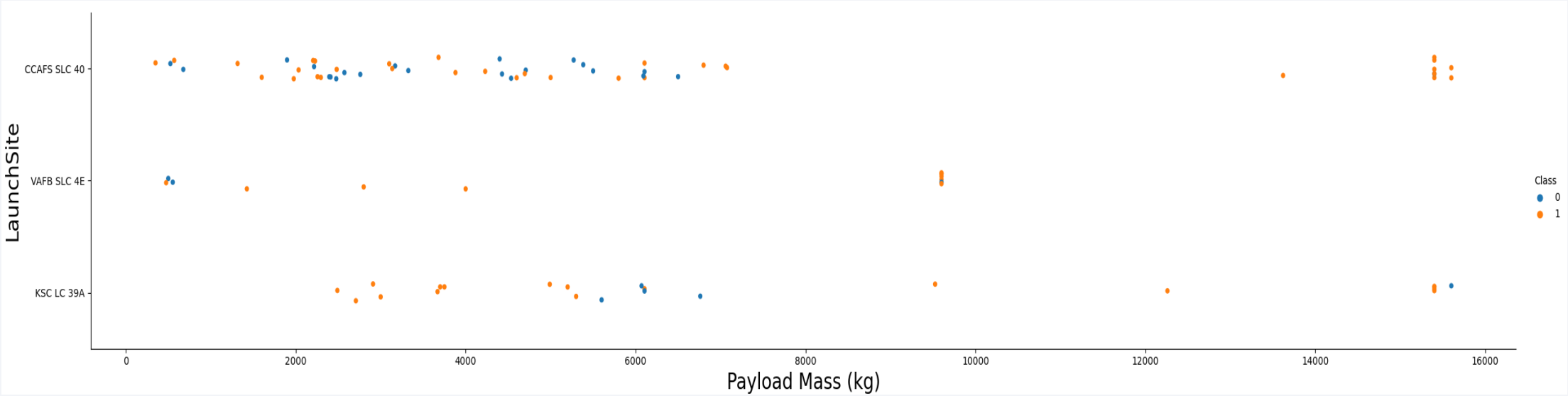
Flight Number vs. Launch Site



From the above plot, we conclude that,

1. The best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful.
2. VAFB SLC 4E and KSC LC39A were the second and third best launch sites respectively.
3. Success rate of launching was improved overtime.

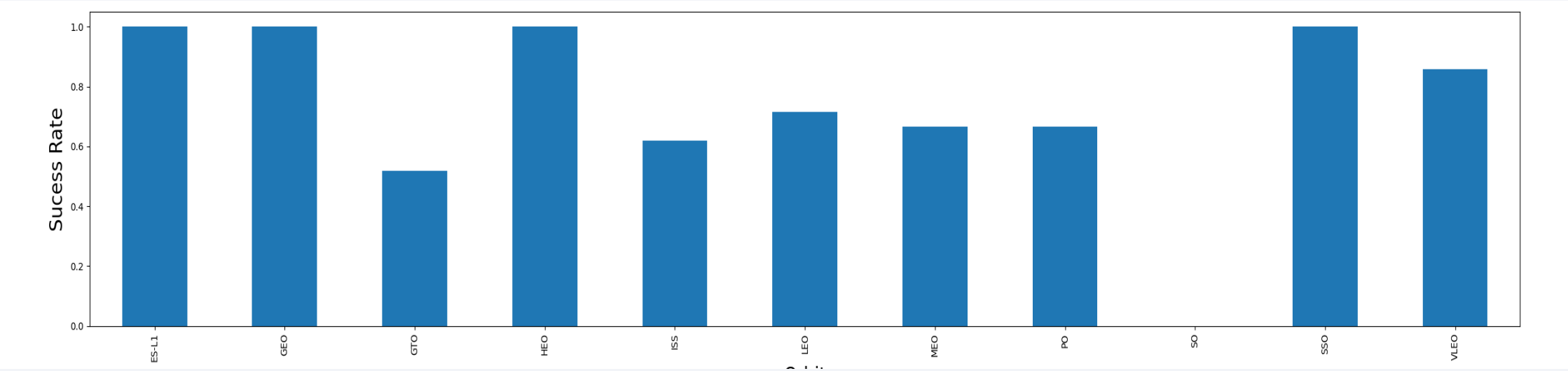
Payload vs. Launch Site



From the above plot we conclude that,

1. Payloads over 9,000 kg (about the weight of a school bus) have excellent success rate;
2. Payloads over 12,000 kg seems to be possible only on CCAFS SLC 40 and KSCLC 39A launch sites.

Success Rate vs. Orbit Type



As it is shown in the graph that,

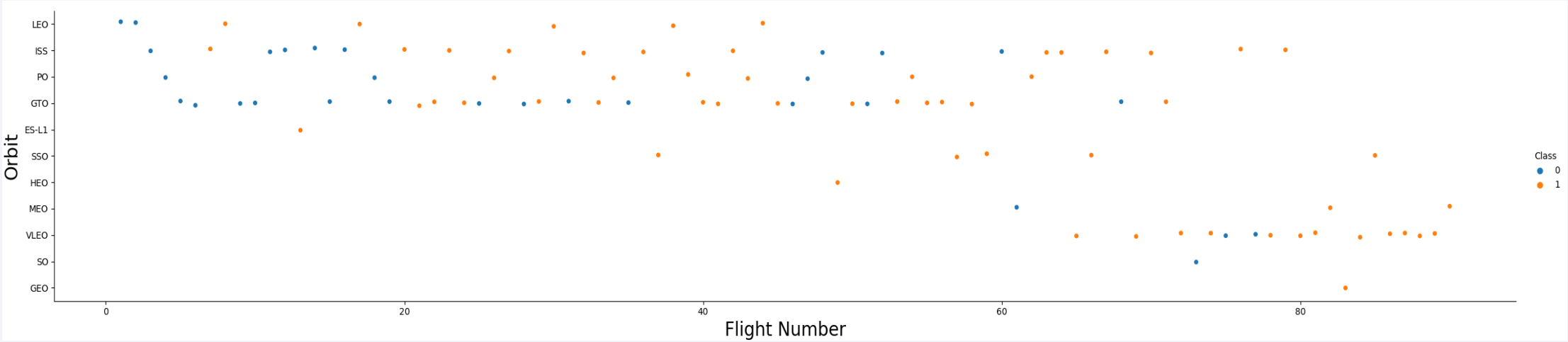
The following orbits have the biggest success rates.

1. ES-L1
2. GEO
3. HEO
4. SSO.

Followed by:

1. VLEO (above 80%) and
2. LFO (above70%).

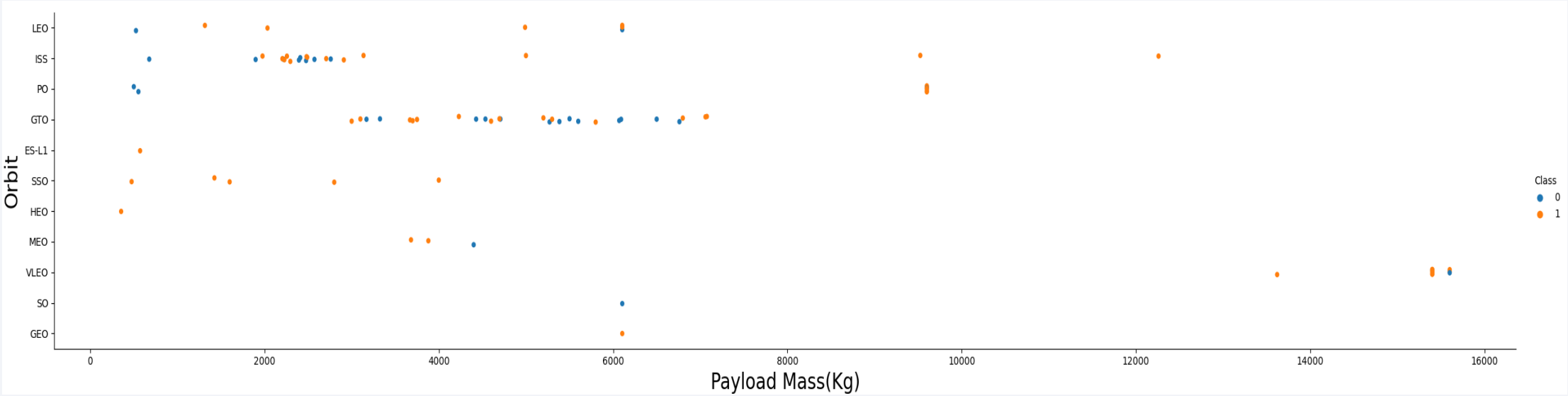
Flight Number vs. Orbit Type



From the above plot we conclude that,

1. There is a strong relationship between LEO orbit and the number of flights.
2. It seems that there is no strong relationship between flight number and GTO orbit.
3. Success rate of all orbits improved over time.
4. Increase in frequency is observed in case of VLEO orbit, it seems a new business opportunity.

Payload vs. Orbit Type



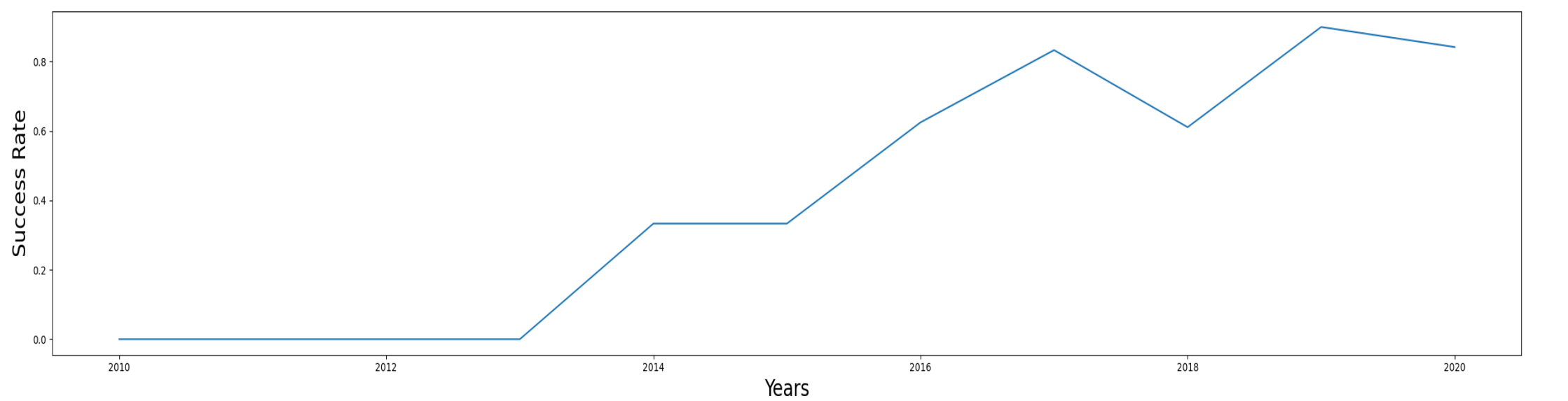
Conclusion from the above plot

I. In case of GTO orbit, it seem there is no relation between payload and success rate.

II. In case of ISS orbit it has the widest range of payload and a good success rate.

III. Few launches have been observed in case of SO and GEO orbits.

Launch Success Yearly Trend



It can be observed that the success rate since 2013 kept increasing till 2020

All Launch Site Names

- There are four different launch sites in the dataset

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

Query to obtain different launch sites

```sql
SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Query :

```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

Total Payload Mass

- Total payload carried by boosters from NASA

SUM (PAYLOAD_MASS__KG_)
45596.0

- Query

```
sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE
CUSTOMER='NASA (CRS)';
```

Average Payload Mass by F9 v1.1

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

```
sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE booster_version LIKE 'F9 v1.1%' ;
```

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

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

```
sql SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING_OUTCOME ='Success (ground pad)';
```

MIN(DATE)
01/08/2018

Filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence of landing, that happened on12/22/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

- Query :

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_
BETWEEN 4000 AND 6000 AND LANDING_OUTCOME ='Success';
```

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

Mission_Outcome	COUNT(*)
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Query :

```
sql SELECT MISSION_OUTCOME, COUNT(*) FROM SPACEXTBL GROUP
BY MISSION_OUTCOME;
```

Boosters Carried Maximum Payload

- List of names of the booster which have carried the maximum payload mass

- Query :

```
sql SELECT BOOSTER_VERSION,
PAYLOAD_MASS__KG_ FROM
SPACEXTBL
WHERE PAYLOAD_MASS__KG_
=(SELECT MAX(PAYLOAD_MASS__KG_) FROM
SPACE
```

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

2015 Launch Records

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

substr(DATE,4,2)	Mission_Outcome	Booster_Version	Launch_Site
10	Success	F9 v1.1 B1012	CCAFS LC-40
11	Success	F9 v1.1 B1013	CCAFS LC-40
02	Success	F9 v1.1 B1014	CCAFS LC-40
04	Success	F9 v1.1 B1015	CCAFS LC-40
04	Success	F9 v1.1 B1016	CCAFS LC-40
06	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
12	Success	F9 FT B1019	CCAFS LC-40

Query :

```
sql SELECT substr(DATE,4,2), MISSION_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where substr(Date,7,4) = '2015';
```

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

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

Landing Outcome	Occurrences
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

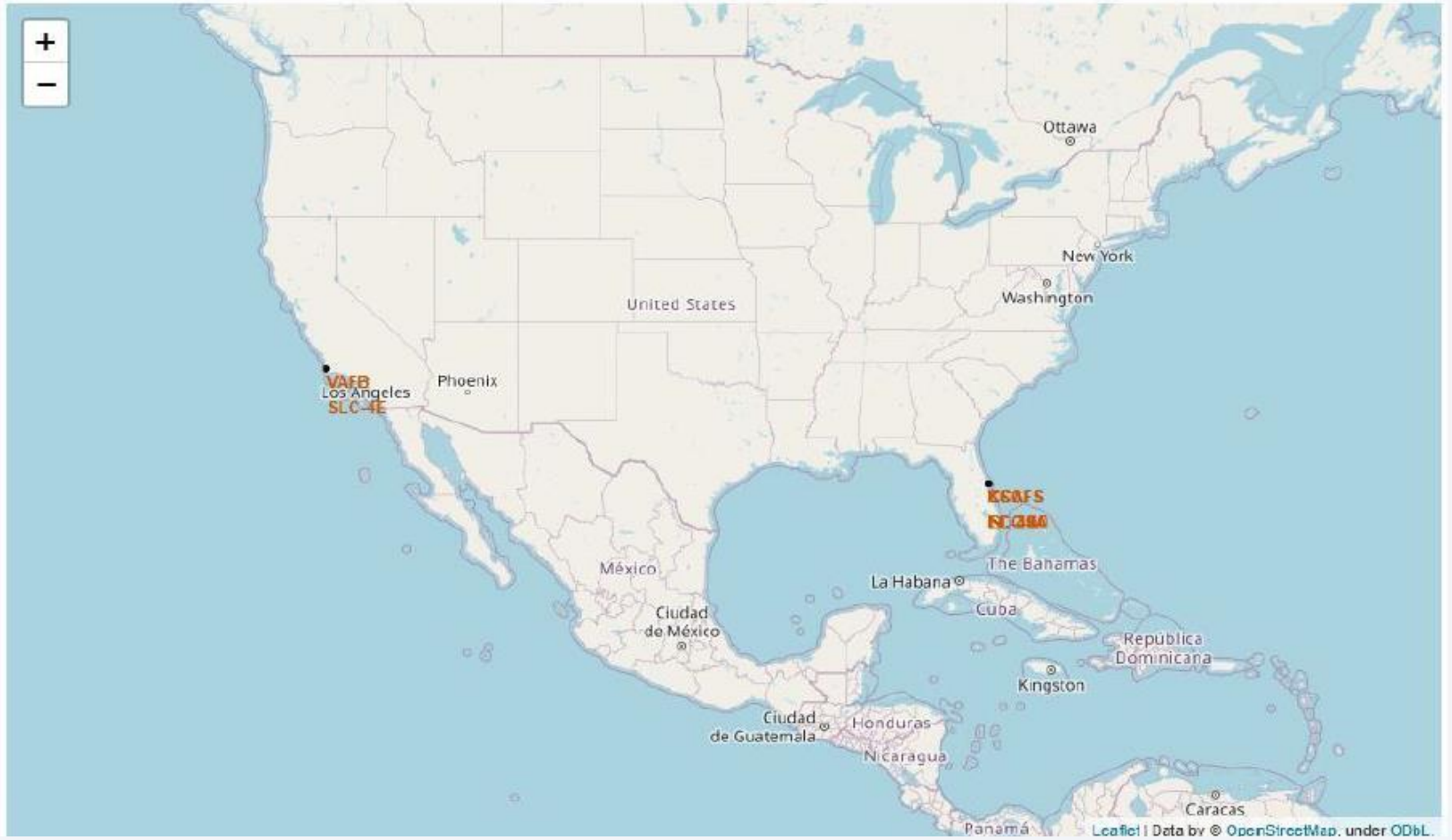
- Query :

```
SELECT LandingOutcome, COUNT(LandingOutcome)
FROM SpaceX
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LandingOutcome
ORDER BY COUNT(LandingOutcome) DESC
```

Section 3

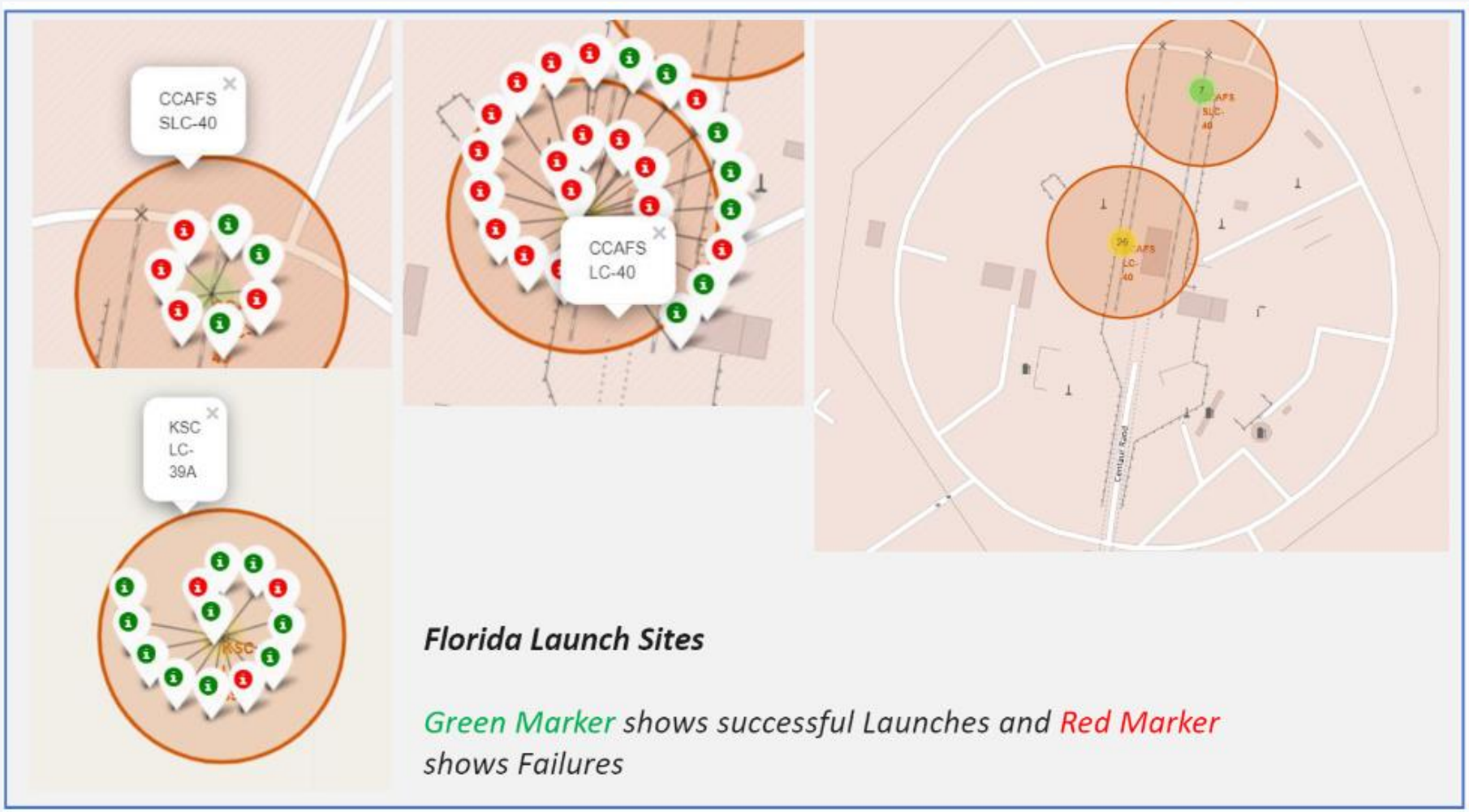
Launch Sites Proximities Analysis

All Launch sites

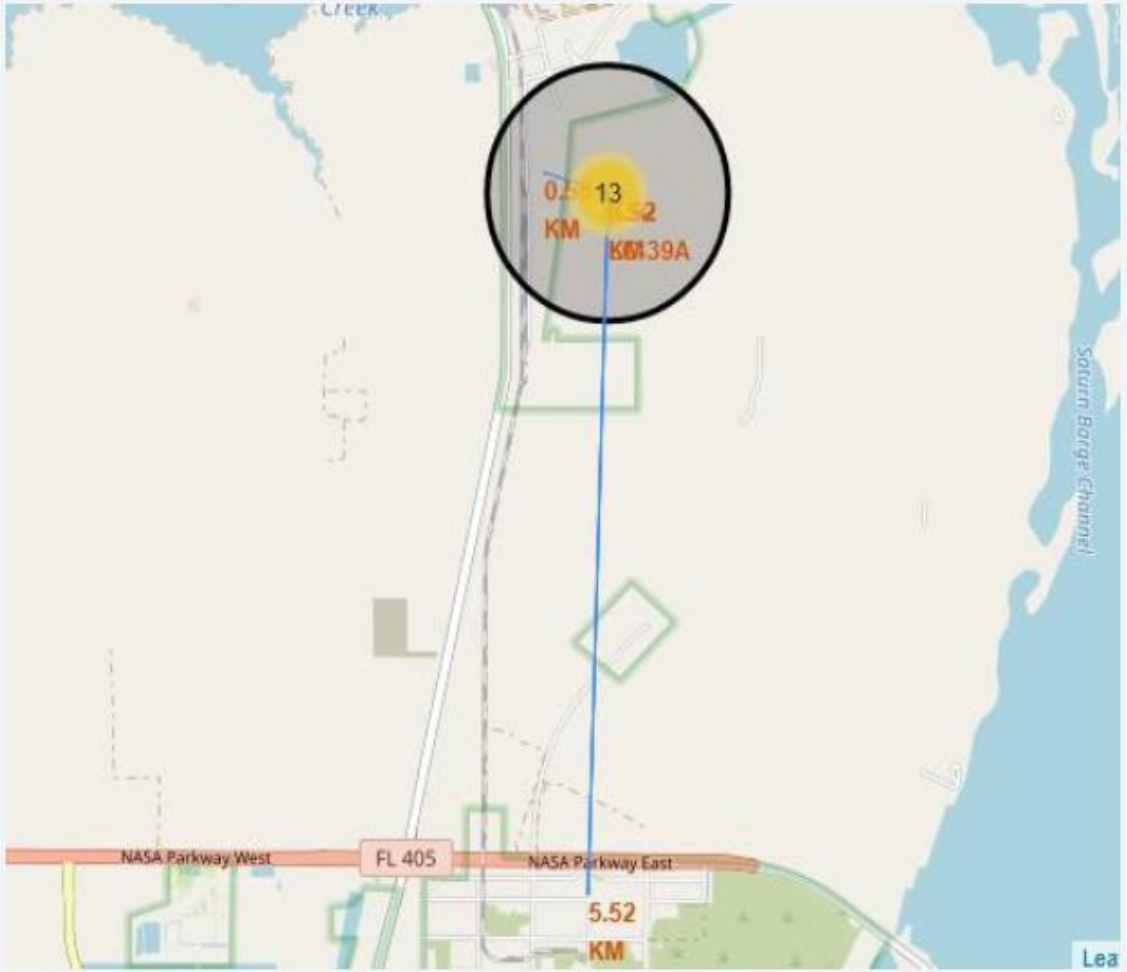


As it can be seen that Launch-sites are near sea at coastal areas considering safety, but not too far from roads and railroads.

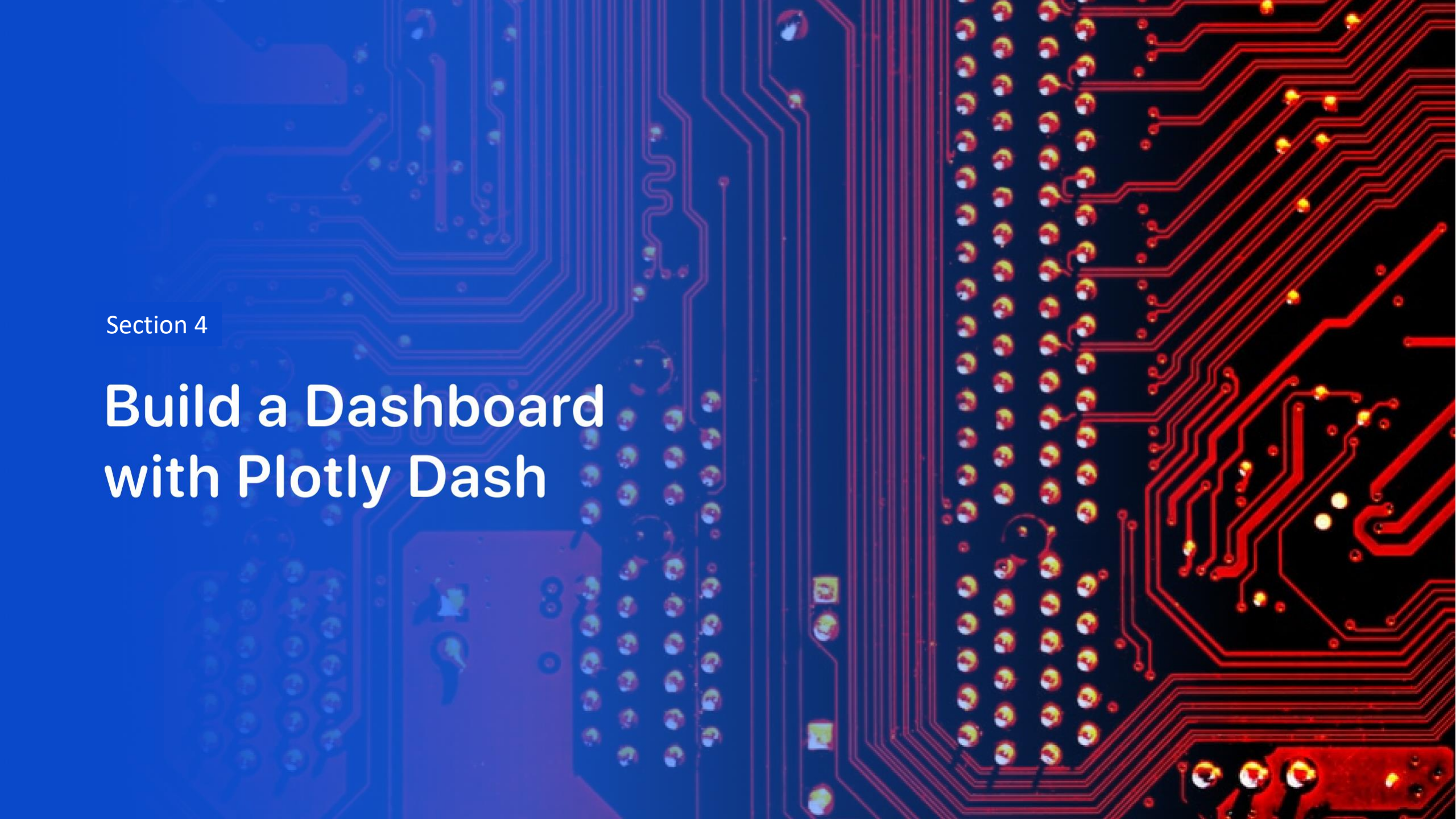
Launch outcomes by sites



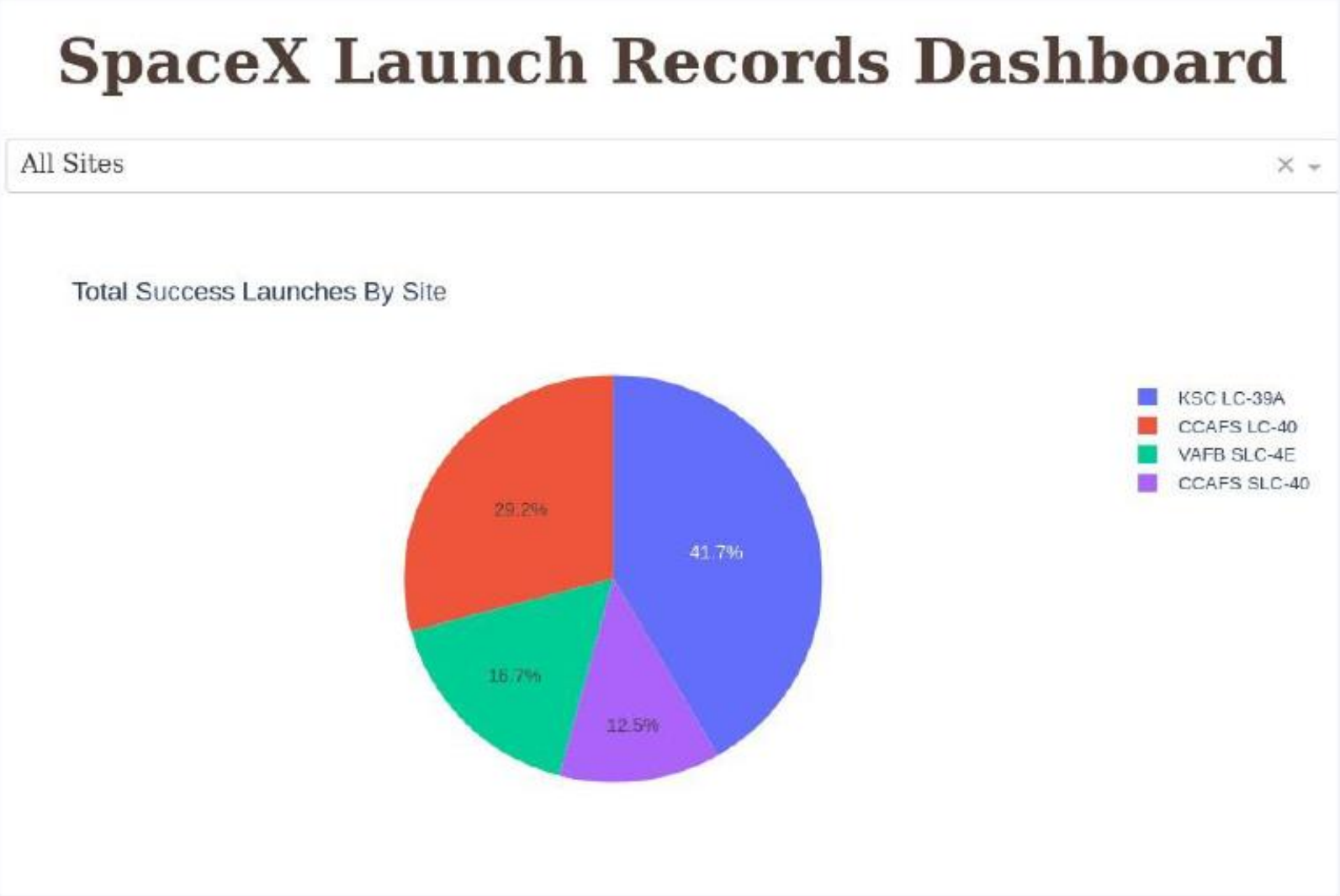
Logistics and Safety



As it can be seen that 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

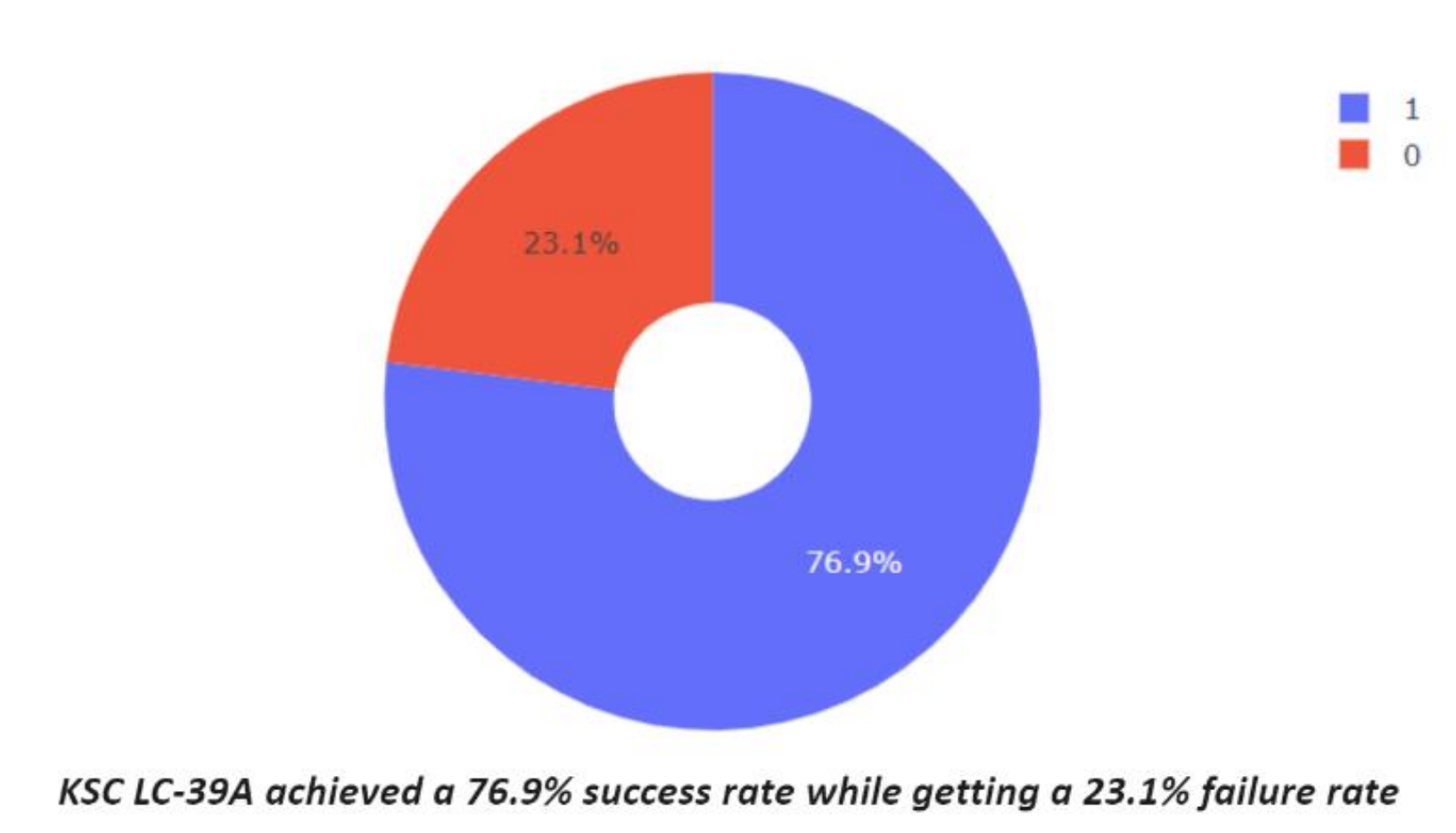
Successful Launches



Very important factor for consideration during launching is the place where launches have been done.

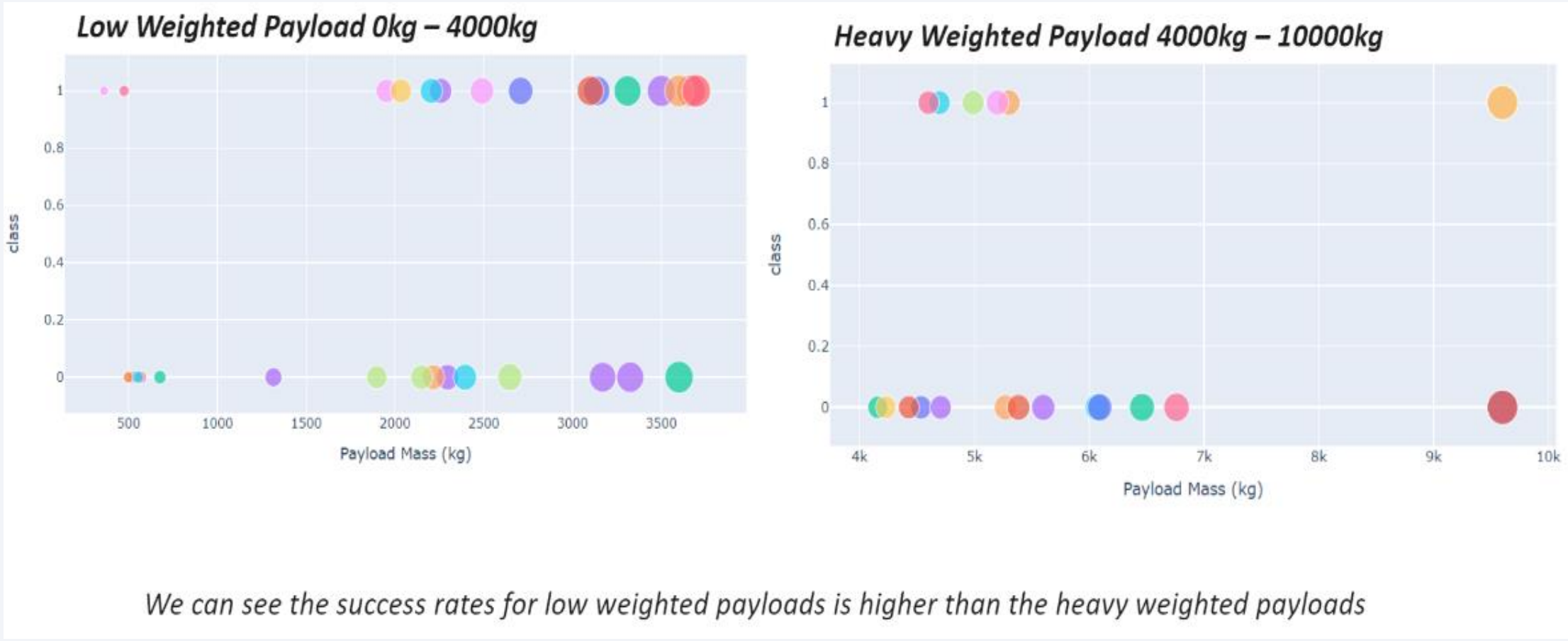
Here KSC LC-39A has highest success rate of 41.7%

Launch site with highest successful ratio



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Payload vs Launch Outcome for all sites



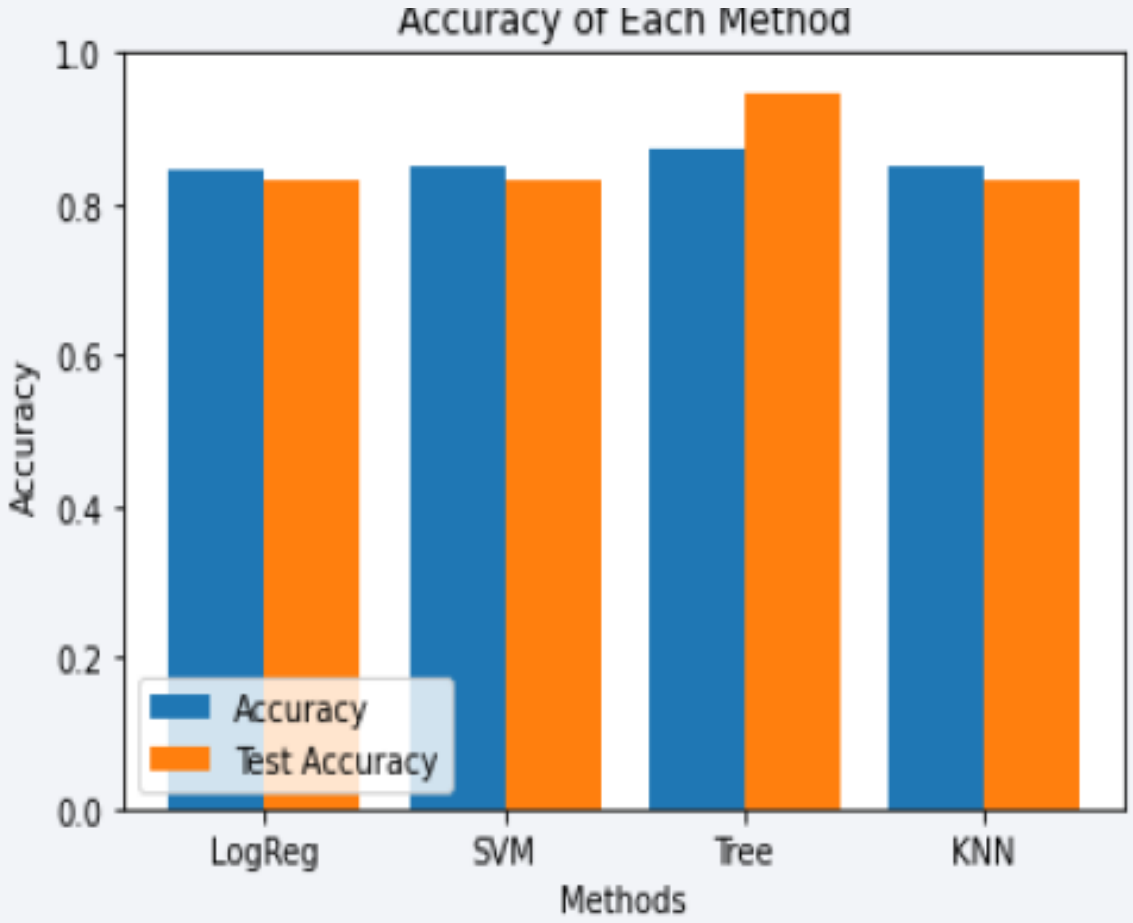
We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

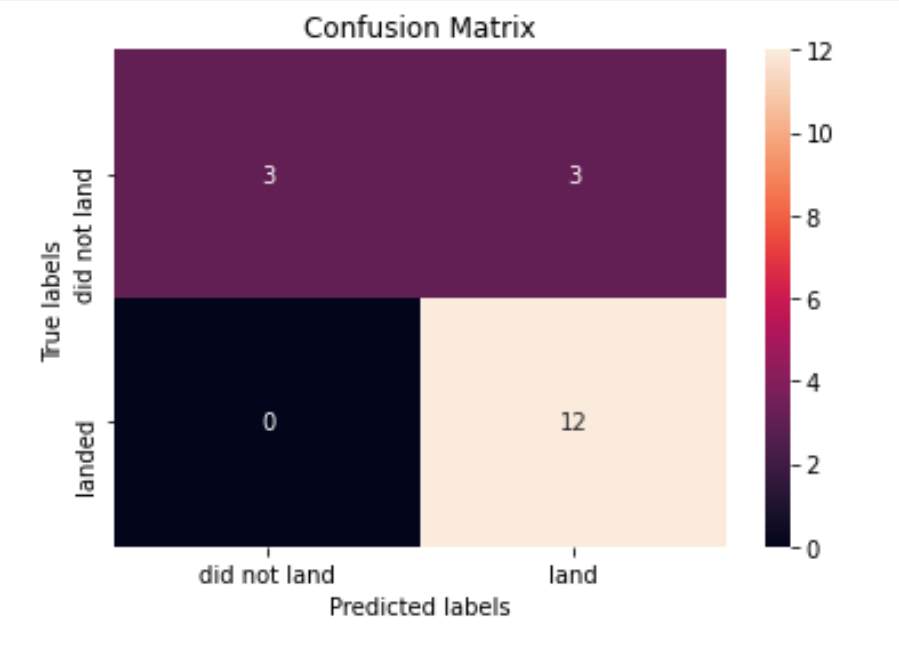
Predictive Analysis (Classification)

Classification Accuracy

Classification models Logistic Regression, SVM, Decision Tree and KNN were tested. The accuracies of different models were plotted. It is fond that the Decision Tree Classifier model had highest accuracy of 87%.



Confusion Matrix



The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.

Conclusions

- The best launch site is KSCLC-39A as it had the most successful launches compared to any other site.
- Orbits having most success rate were ES-L1, GEO, HEO, SSO and VLEO.
- Launches below 7000 kg seems to be more risky in comparison with the launches above 7000 kgs.
- Successful launches improve over time. It is increasing from 2013 till 2020.
- Amongst four models decision tree classifier was found to be the best classifier model with an accuracy of 87%

Appendix

- Folium library didn’t show map on github.
- Screen shots of map has been pasted here.

Thank you!