IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

BERINYUY BERTRAND MAINIMO

01/01/2023

# Outline

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

# Executive Summary

This capstone project is all about predicting if the SpaceX Falcon 9 first stage will land successfully. If we can determine if the first stage will land, we can determine the cost of a launch. This will be achieved with the use of different data science procedures.

- Summary of methodologies
  - Data Collection with the help of an API
  - Data Collection using Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis using SQL
  - Exploratory Data Analysis using Data Visualization processes
  - Interactive Visual Analytics using Folium
  - Machine Learning Prediction

- Summary of all results
  - Results of Exploratory Data Analysis
  - Screenshots presenting interactive analytics
  - Results of the Predictive Analytics

# Introduction

- **Project background and context**

The main aim of this Data Science project is to allow the company to compete with SpaceX. In order to achieve this goal, it is important to determine if the first stage of the SpaceX Falcon 9 rocket will land successfully.

SpaceX 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.

- **Problems you want to find answers**

Therefore if we can accurately predict the likelihood of the first stage rocket landing successfully, 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, then we can determine the cost of a launch. With the help of the Data Science findings and models, the company can make more informed bids against SpaceX for a rocket launch.

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - Make requests to the SpaceX API.
  - Perform web scraping to collect Falcon 9 historical launch records on the Wikipedia page titled: List of Falcon 9 and Falcon Heavy launches
- Perform data wrangling
  - Clean the data and explore it to find patterns in the data to determine the labels for training supervised models.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Create a machine learning pipeline to predict if the first stage will land given the data.
  - Train the best performing model to make accurate predictions.

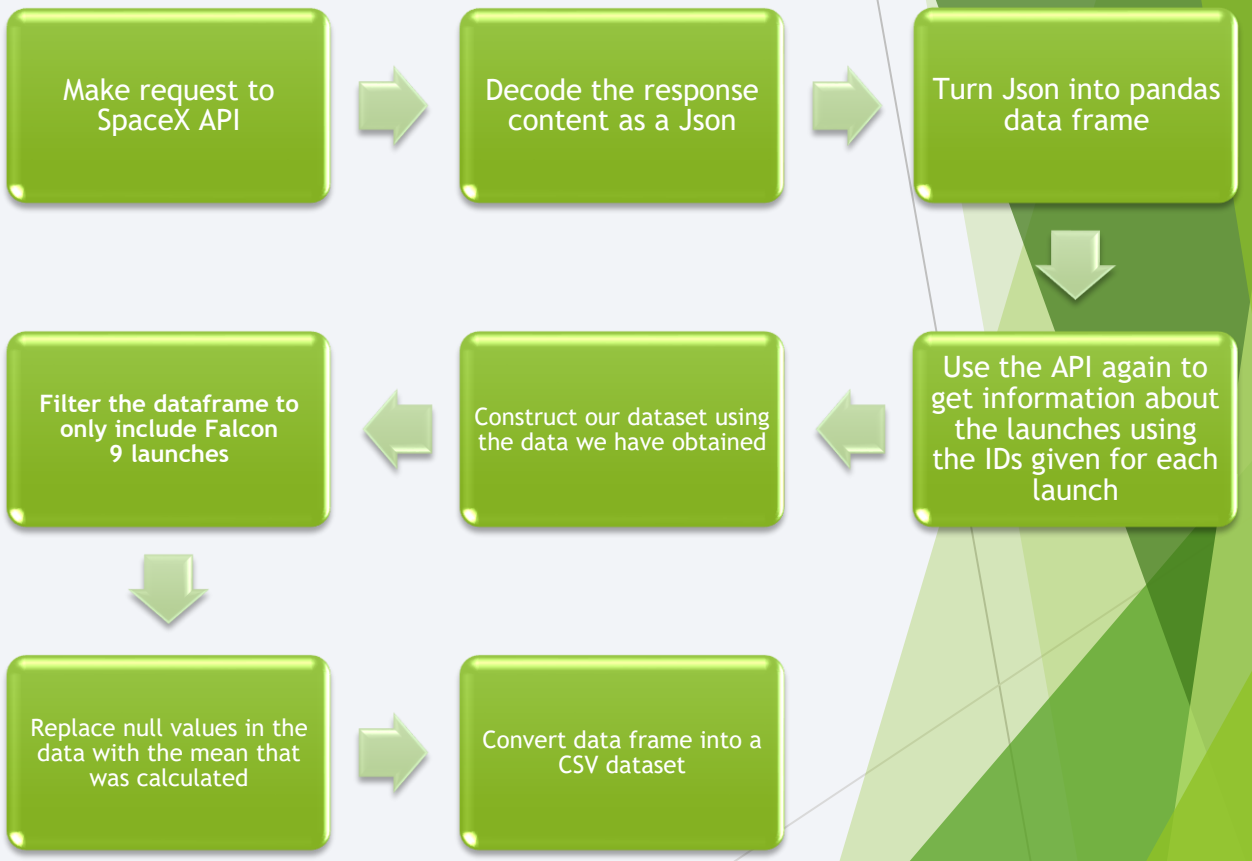
# Data Collection

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 dataframe 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.

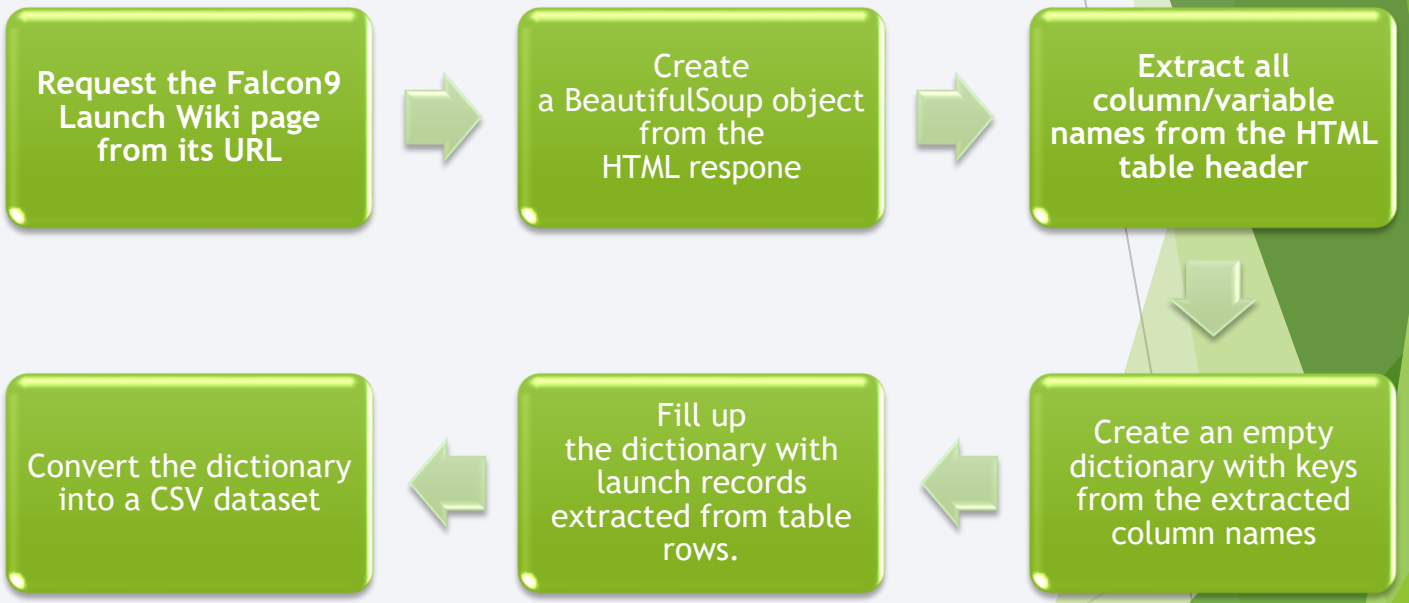
# Data Collection – SpaceX API

- Make a request to SpaceX API and make sure the data is in the correct format.
- Perform some basic data wrangling and formatting in order to clean the requested data.
- Convert our data frame into a CSV dataset.
- URL link: spacex-data-collection-api



# Data Collection - Scraping

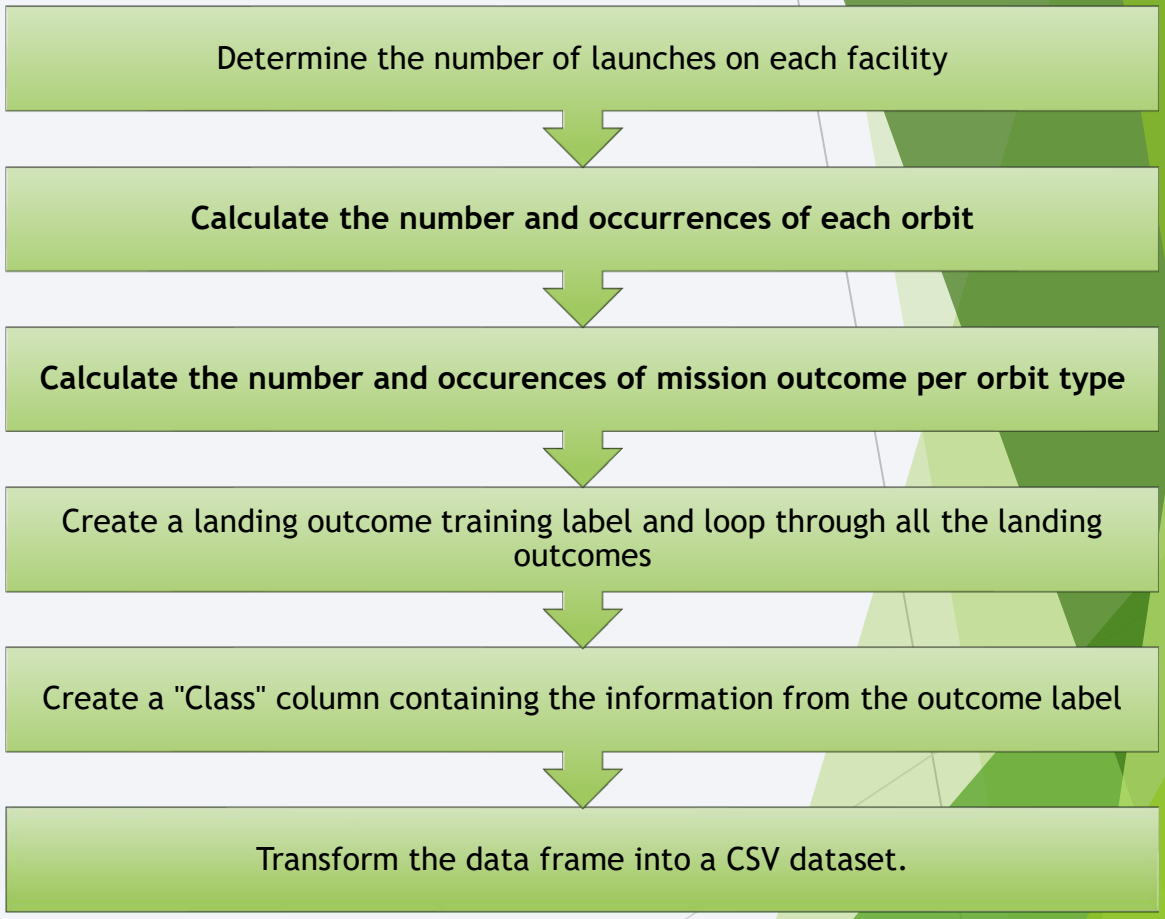
- Data Collection BY Web Scraping process is given in flow chart for an overview. For Completed Notebook link given below
- • Git Hub URL link: [Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia]



# Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

I mainly convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful.



# EDA with Data Visualization

Data visualization helps us understand data by curating it into a form that's easier to understand, highlighting the trends and outliers. Several types of charts were used in the visualization of the data:

- Cat plots and scatter plots were used to view the relationships of categorical variables like *Launch Site* and *Orbit*.
- A bar chart was used to visualize the success rate of each orbit type.
- A line chart was used to visualize the launch success yearly trend.

URL link: eda-dataviz

# EDA with SQL

Summary of SQL queries that were used:

- Display the names of the unique launch sites in the space mission
- Compare the payload mass with boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the total number of successful and failure mission outcomes
- Determine the dates of different landing outcomes

URL link: EDA - SQL

# Build an Interactive Map with Folium

Markers of all Launch Sites:

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Coloured Markers of the launch outcomes for each Launch Site:

- Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

- GitHub URL (Notebook ): Notebook Interactive Map with Folium
- Google colab URL: Google colab Notebook Interactive Map with Folium (use this link for the maps to load)

# 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.

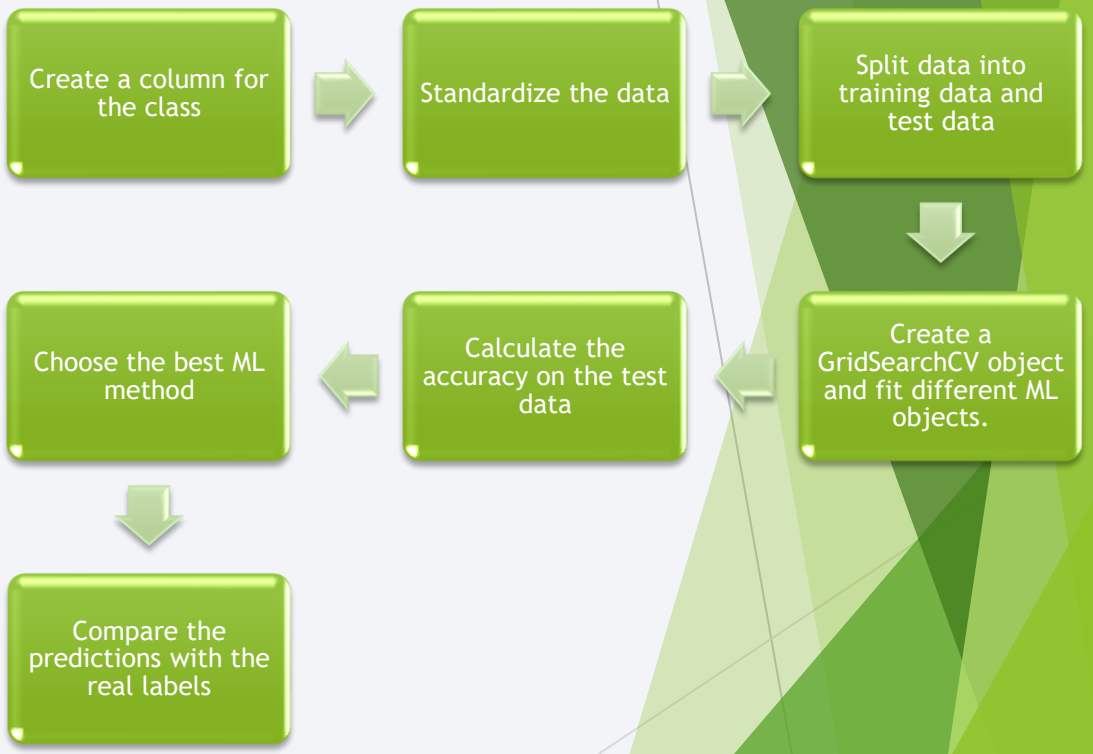
- GitHub URL: Plotly - Dash

# Predictive Analysis (Classification)

Scikit-learn is the primary ML(machine learning) library that was used for predictive analysis. The following took place:

- Created a machine learning pipeline to predict if the first stage will land given the data.
- Using *GridSearchCV*, found the best ML method for predictions.
- Compared the predictions with the real labels.

- GitHub URL: Predictive Analysis



# Results

The exploratory data analysis has shown us that successful landing outcomes are somewhat correlated with flight number. It was also apparent that successful landing outcomes have had a significant increase since the year 2015.

All launch sites are located near the coast line. Perhaps, this makes it easier to test rocket landings in the water.

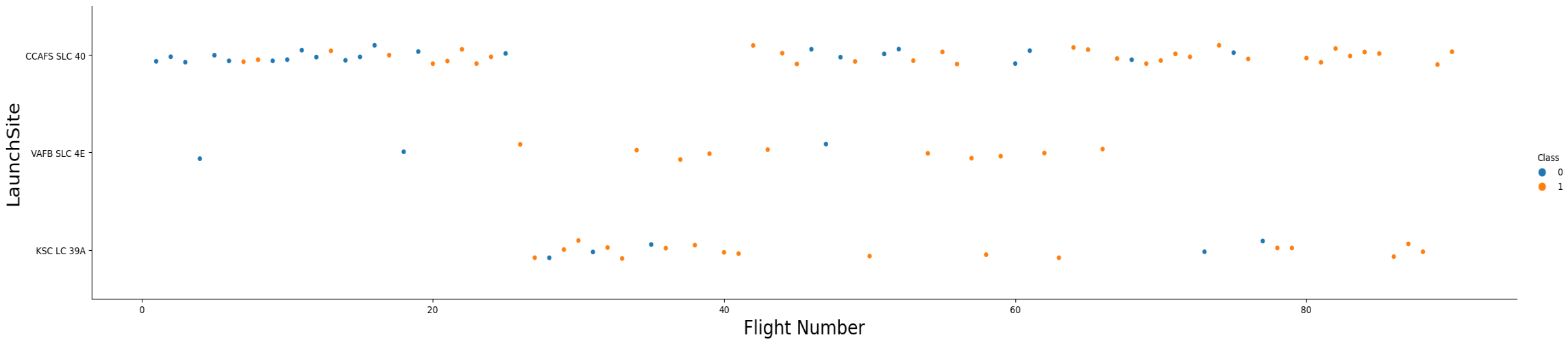
Furthermore, the sites are also located near highways and railways. This may facilitate transportation of equipment and research material.

The machine learning models that were built, were able to predict the landing success of rockets with an accuracy score of 83.33%. This accuracy can be increased in future projects with more data.

Section 2

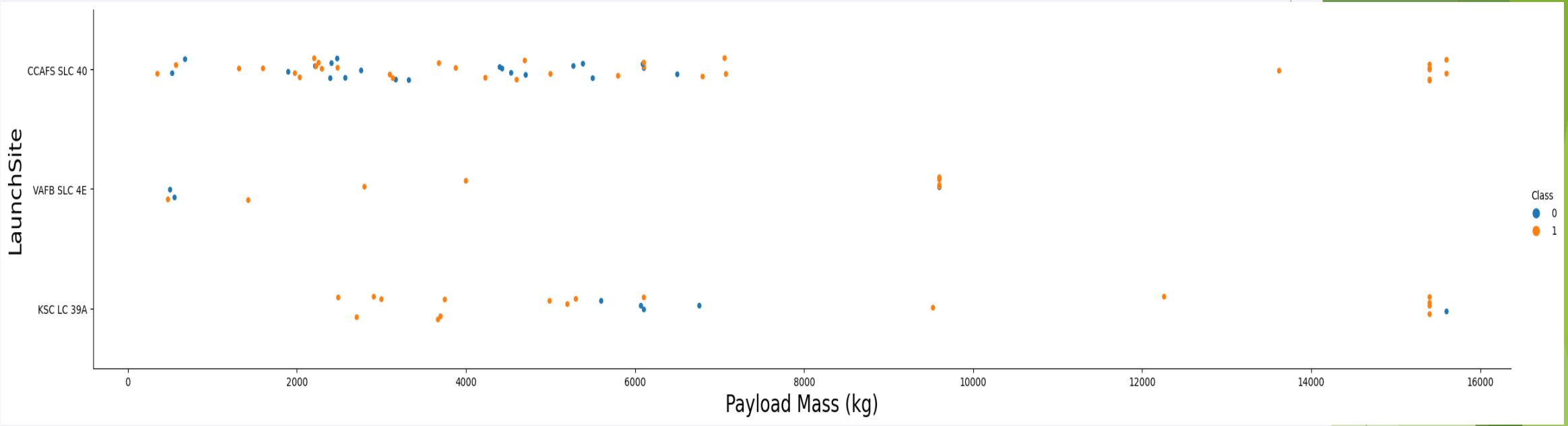
# Insights drawn from EDA

# Flight Number vs. Launch Site



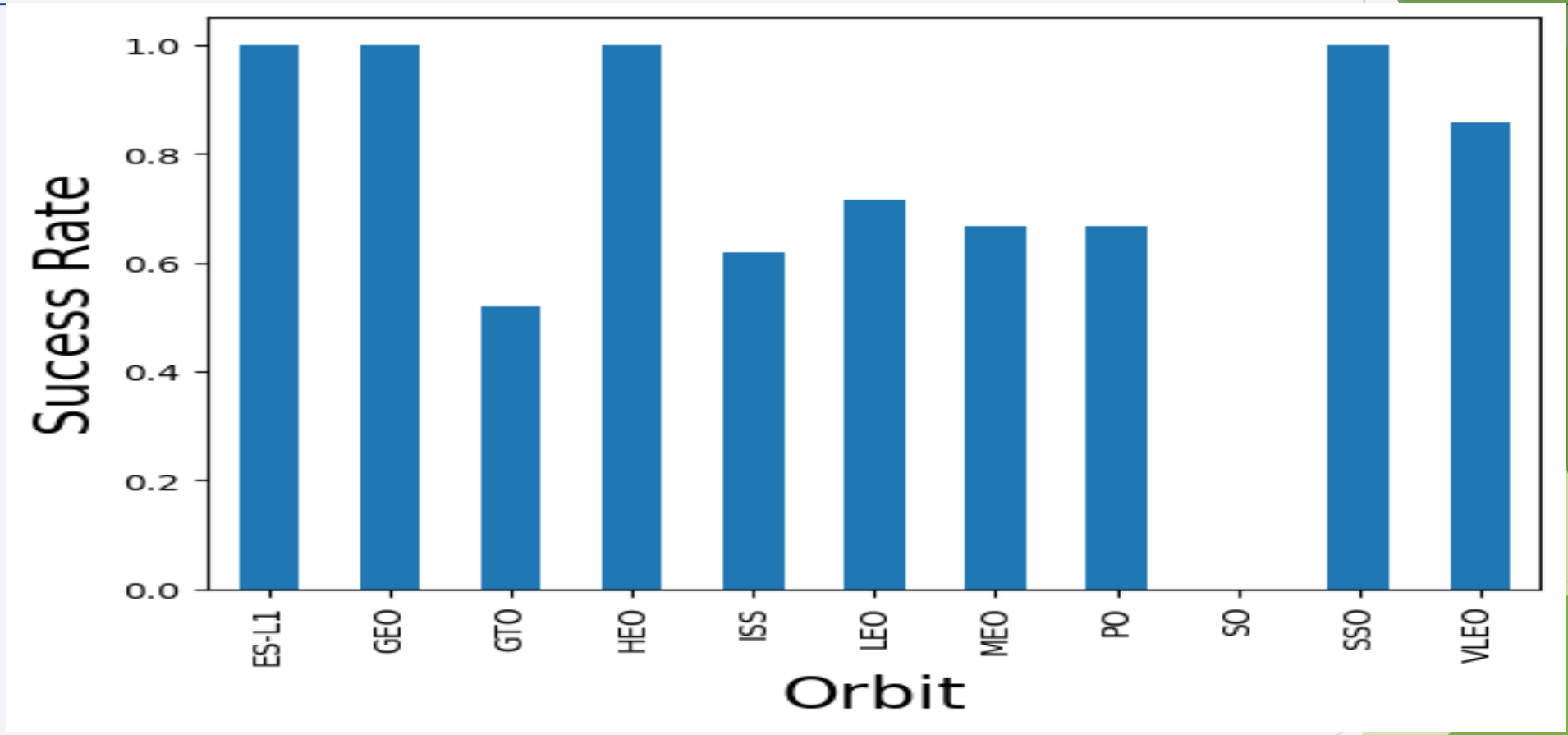

From the plot, it appears that there were more successful landings as the flight numbers increased. It also seems that launch site **CCAFS SLC 40** had the most number of landing attempts while the site **VAFB SLC 4E** had the least number of attempts.

# Payload vs. Launch Site



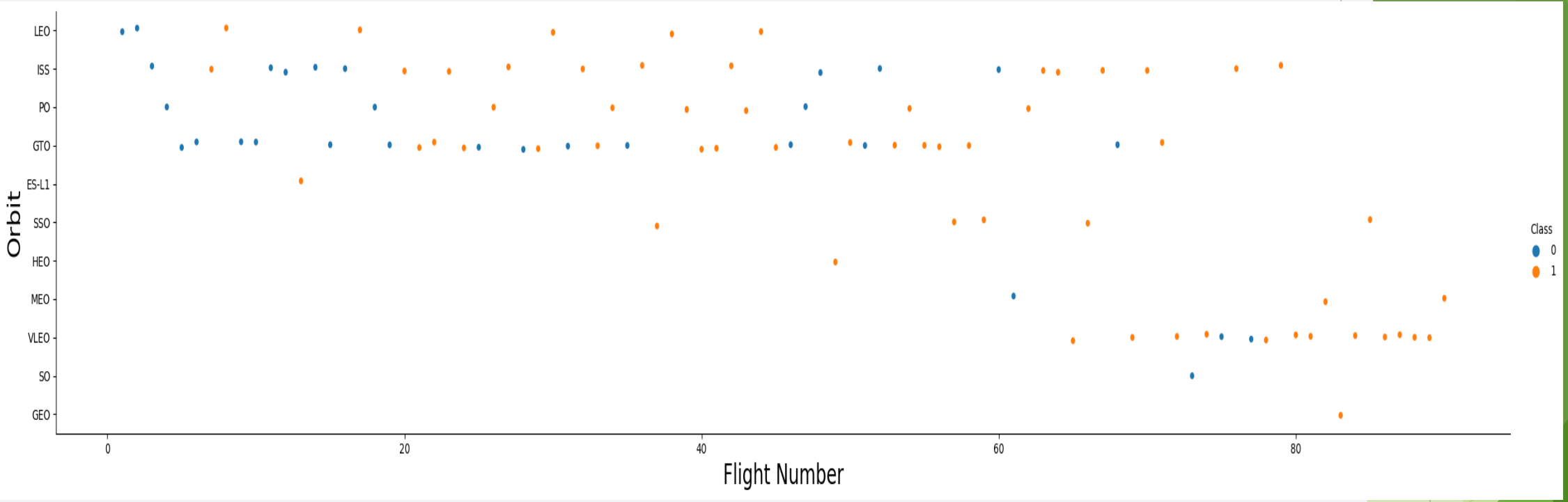

- From the plot as the payload mass increases, the success rate increases as well
- Most of the launches with payload mass over 7000 kg were successful.

# Success Rate vs. Orbit Type




The orbit types **SSO**, **HEO**, **GEO** and **ES-L1** had the highest success rate.

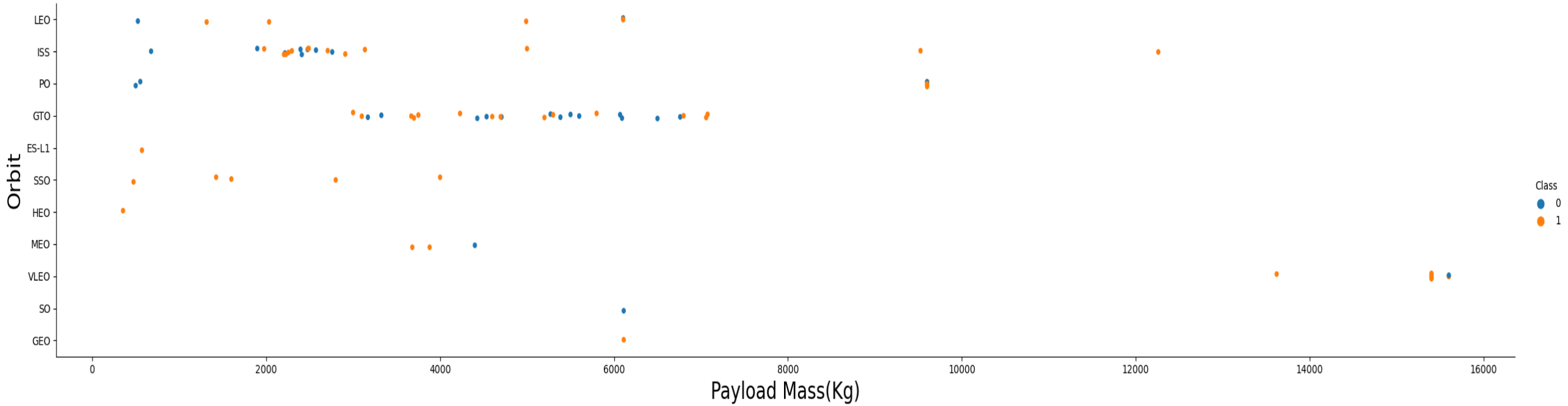
# Flight Number vs. Orbit Type




- In the LEO orbit,  the success is positively correlated to the number of flights.

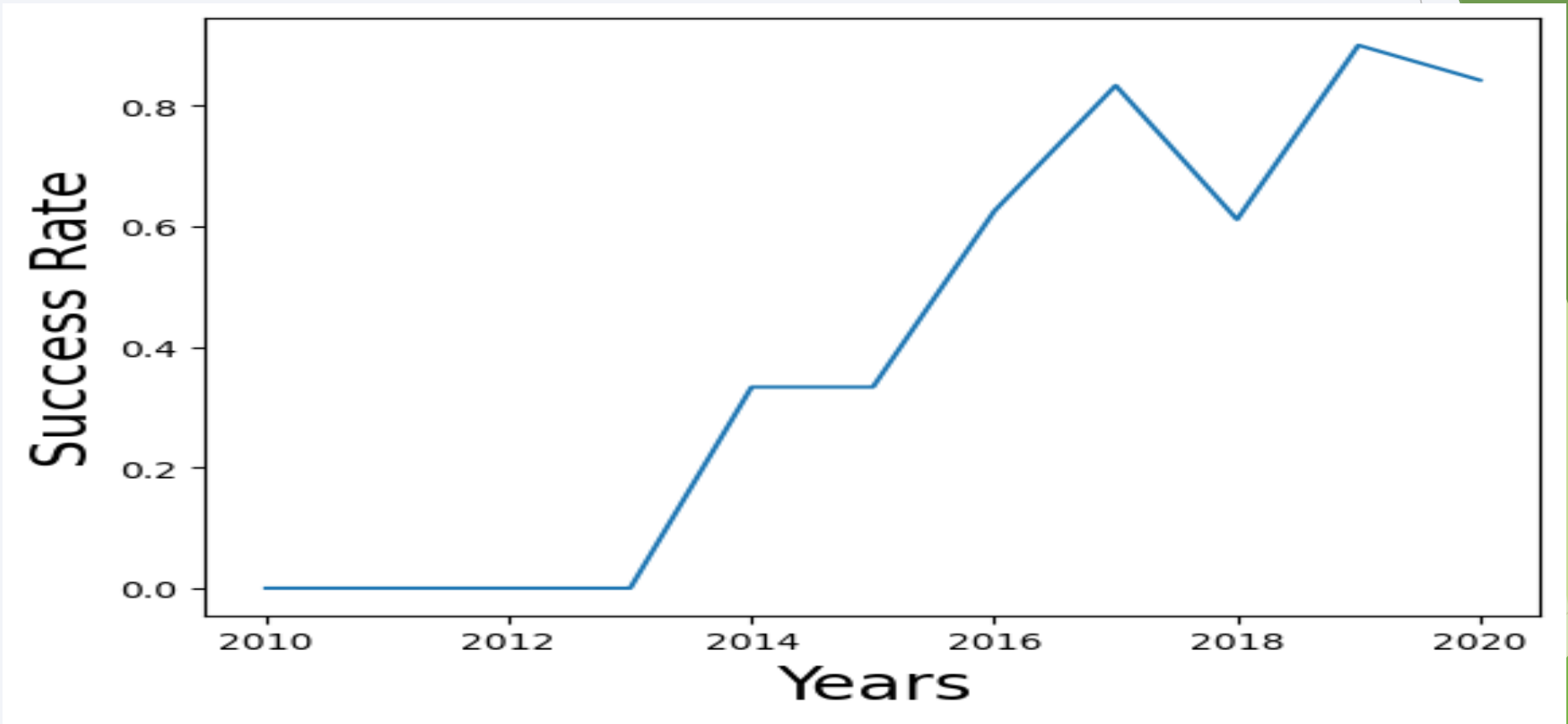
- There seems to be no relationship between flight number in the GTO orbit.

# Payload vs. Orbit Type




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

# Launch Success Yearly Trend




From the plot, we can observe that success rate since 2013 kept on increasing till 2020.

# All Launch Site Names

- The DISTINCT clause was used to  return only the unique rows from the launch_site column.
- The names of the launch sites are CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E .

```
%sql select unique(launch_site) from SPACEX_TABLE
```

```
 * ibm_db_sa://qcq21372:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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

# Launch Site Names Begin with 'CCA'

```
%sql SELECT * from SPACEX_TABLE where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://qcq21372:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

| 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 |

These are 5 records where launch sites begin with the letters 'CCA'. As we can see, there are other organizations besides SpaceX that were testing their rockets.

# Total Payload Mass

```
%sql SELECT sum(payload_mass__kg_) as sum_payload from SPACEX_TABLE where (customer) = 'NASA (CRS)'
```

```
 * ibm_db_sa://qcq21372:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

| sum_payload |
| --- |
| 45596 |

- The information in the table displays the total payload mass carried by boosters launched by NASA .
- It seems that *NASA (CRS)* had a significantly higher total payload mass compared to the rest.

# Average Payload Mass by F9 v1.1

```
%sql SELECT avg(payload_mass__kg_) as average_payload from SPACEX_TABLE where (booster_version) = 'F9 v1.1'
```

```
 * ibm_db_sa://qcq21372:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

| average_payload |
|---|
| 2928 |

- The AVG() function was used to the calculate the average payload the average payload mass carried by booster version F9 v1.1
- The WHERE clause was used to filter results so that the calculations were only performed on booster versions only if they were named “F9 v1.1”
- The calculated average payload mass carried by booster version F9 v1.1 is 2928.4

# First Successful Ground Landing Date

```
%sql SELECT min(date) from SPACEX_TABLE where landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://qcq21372:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

| 1 |
| --- |
| 2015-12-22 |

- From the picture given above you can see that the first successful ground pad was in 22 December 2015.

# Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select BOOSTER_VERSION from SPACEX_TABLE where LANDING__OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4001 and 5999
```

```
 * ibm_db_sa://qcq21372:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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

- Only 4 Boosters with a payload mass between 4000 and 6000 experienced a successful drone ship landing
- It is interesting to see that they all had successful landing outcomes.

# Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS OUTCOME FROM SPACEX_TABLE GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://qcq21372:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

| mission_outcome | outcome |
|---|---|
| Failure (in flight) | 1 |
| Success | 99 |
| Success (payload status unclear) | 1 |

- It appears that missions generally tend to be successful with the exception of one failure.

# Boosters Carried Maximum Payload

```
%sql SELECT BOOSTER_VERSION FROM SPACEX_TABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX_TABLE)
```

```
 * ibm_db_sa://qcq21372:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

| 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 |

12 boosters have carried the maximum payload mass of 15600 kg.

# 2015 Launch Records

```
%sql SELECT DATE, BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME FROM SPACEX_TABLE WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND YEAR(DATE) = 20
```

```
 * ibm_db_sa://qcq21372:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

| DATE | booster_version | launch_site | landing_outcome |
|---|---|---|---|
| 2015-01-10 | F9 v1.1 B1012 | CCAFS LC-40 | Failure (drone ship) |
| 2015-04-14 | F9 v1.1 B1015 | CCAFS LC-40 | Failure (drone ship) |

- It appears that 2 boosters failed to land at the beginnnig of the year..
- The first successful landing took place later that year in December as we saw earlier.

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

```
%sql SELECT LANDING__OUTCOME, COUNT(*) AS COUNT_LAUNCHES FROM SPACEX_TABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME
```

```
 * ibm_db_sa://qcq21372:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

| landing__outcome | count_launches |
|---|---|
| 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 |

- If we observe the table, it is apparent that the number of successful landings have increased since 2015.
- Before 2013, it seems that there were no attempts to land the boosters.

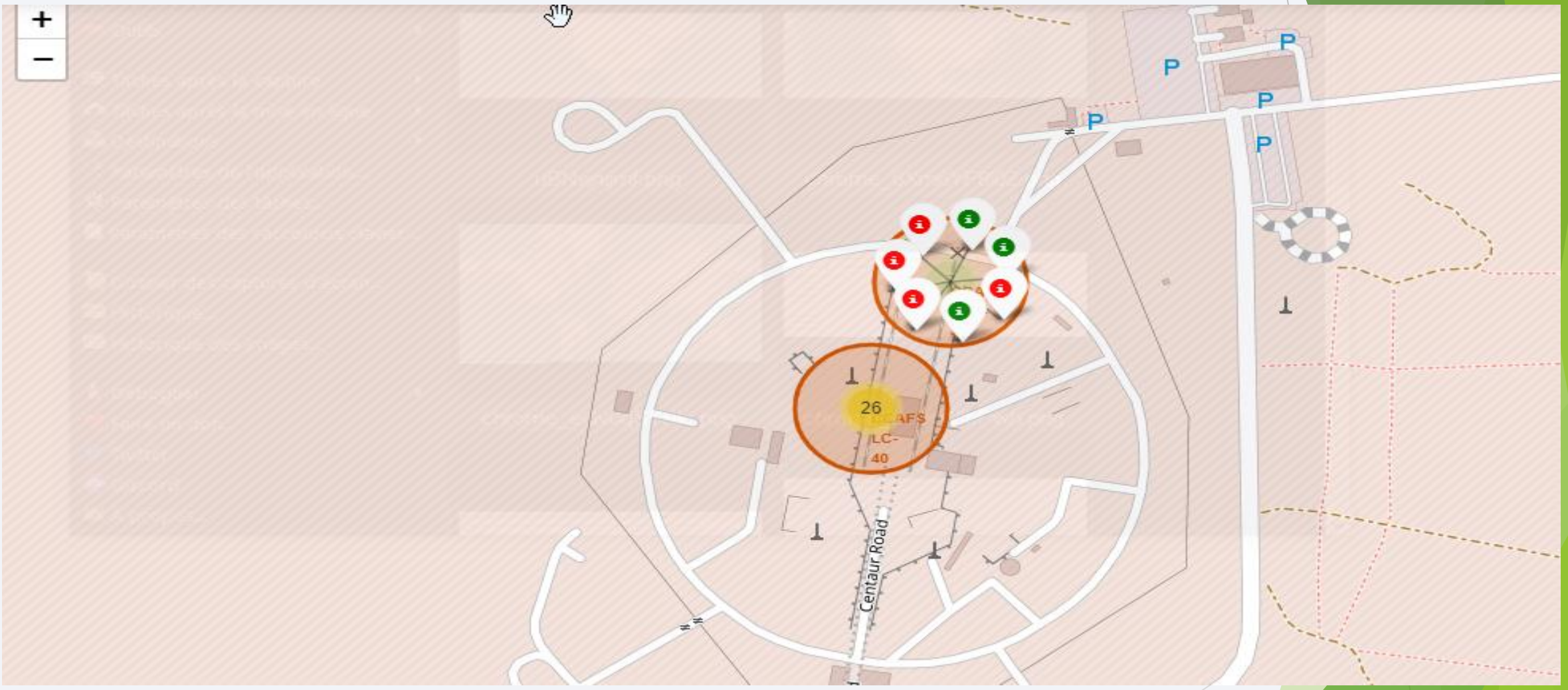
Section 3

# Launch Sites Proximities Analysis

# Launch Site Locations

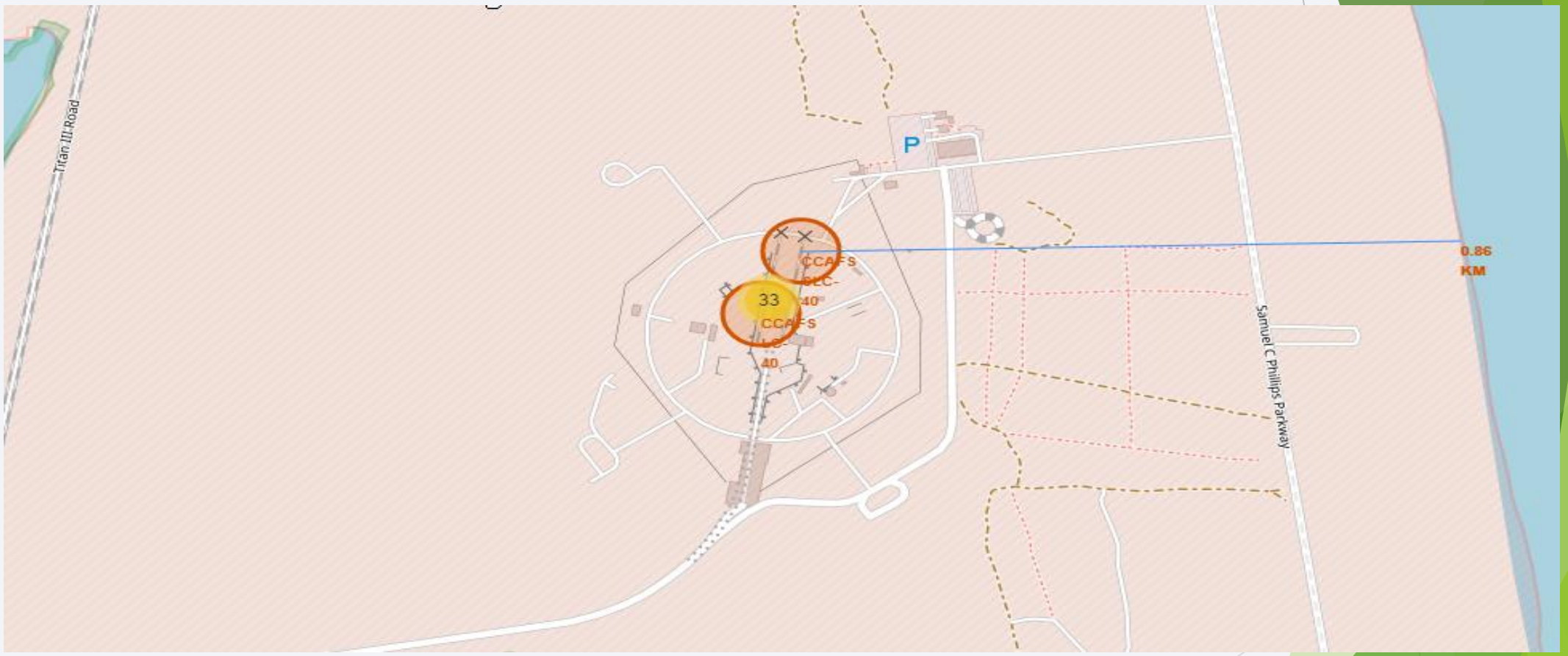

- All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.
- Most of Launch sites are in proximity to the Equator line. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.

# Success vs. Failure Rate of Rocket Launches

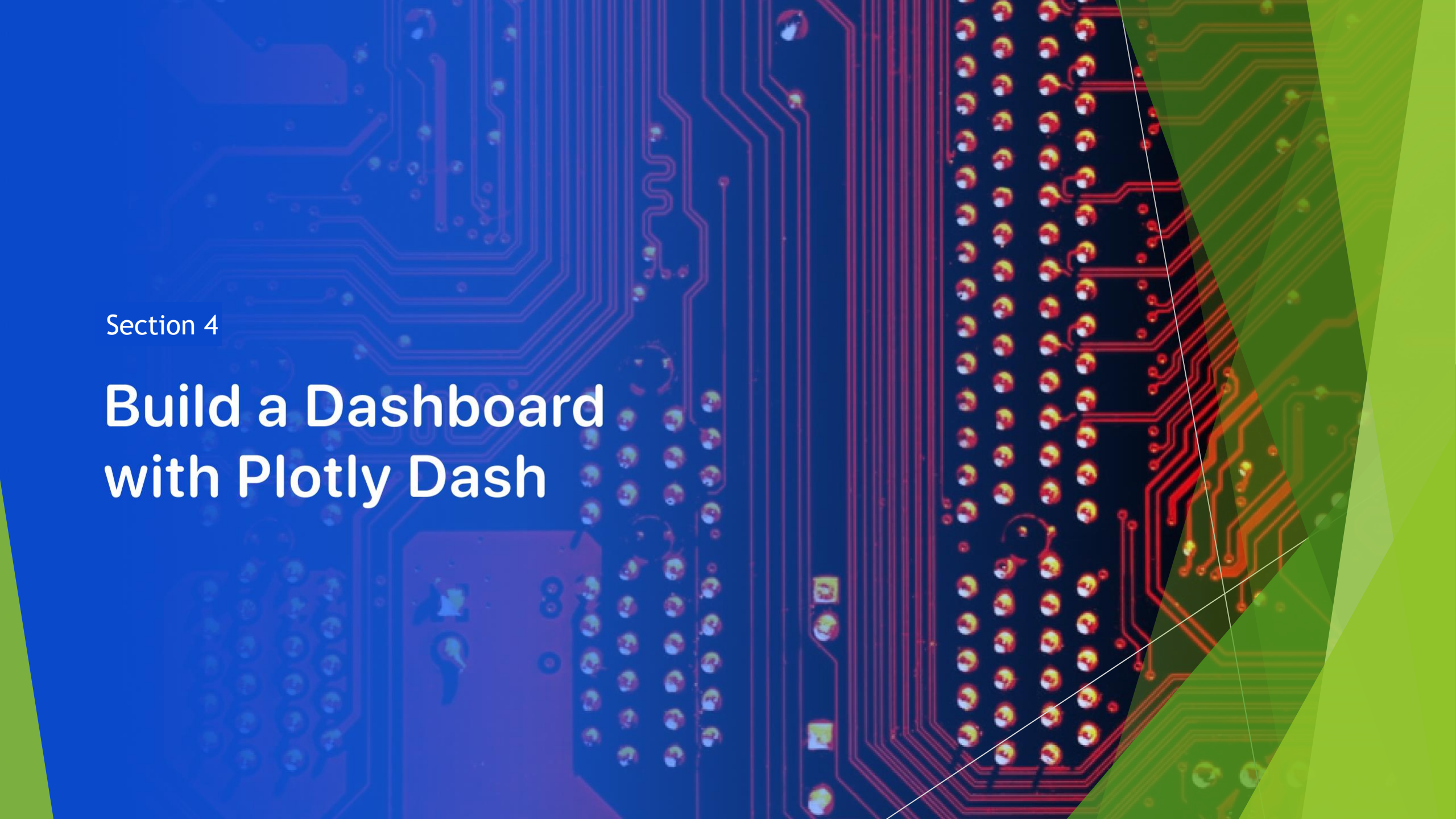



Zooming in on a launch site, we can click on the launch site which will display marker clusters of successful landings (green) or failed landing (red).

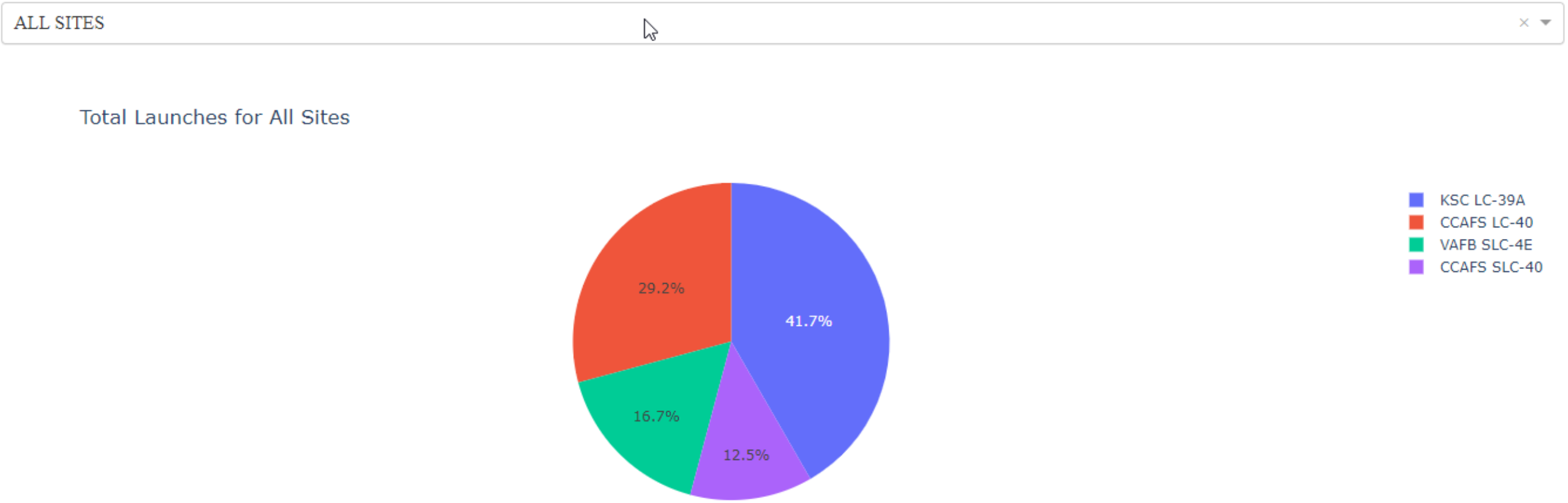
# Launch Site Proximities




The sites are close the coast line. This is evident with the many rocket landing tests on water bodies like the ocean.

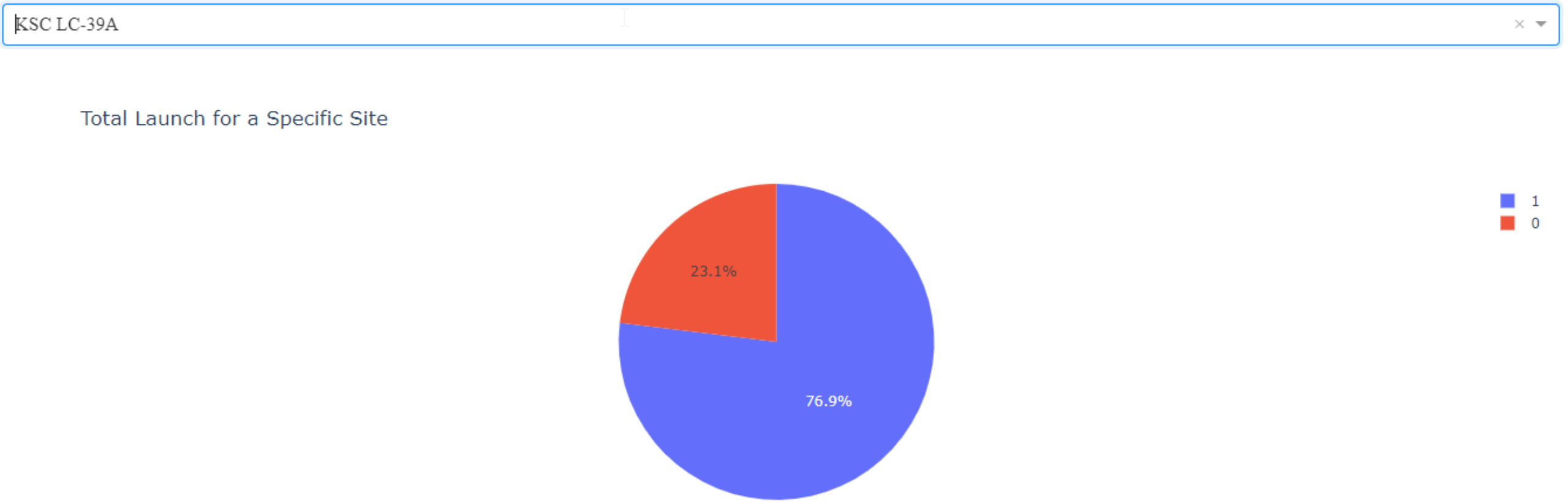

Section 4
Build a Dashboard with Plotly Dash

# <Dashboard Screenshot 1>



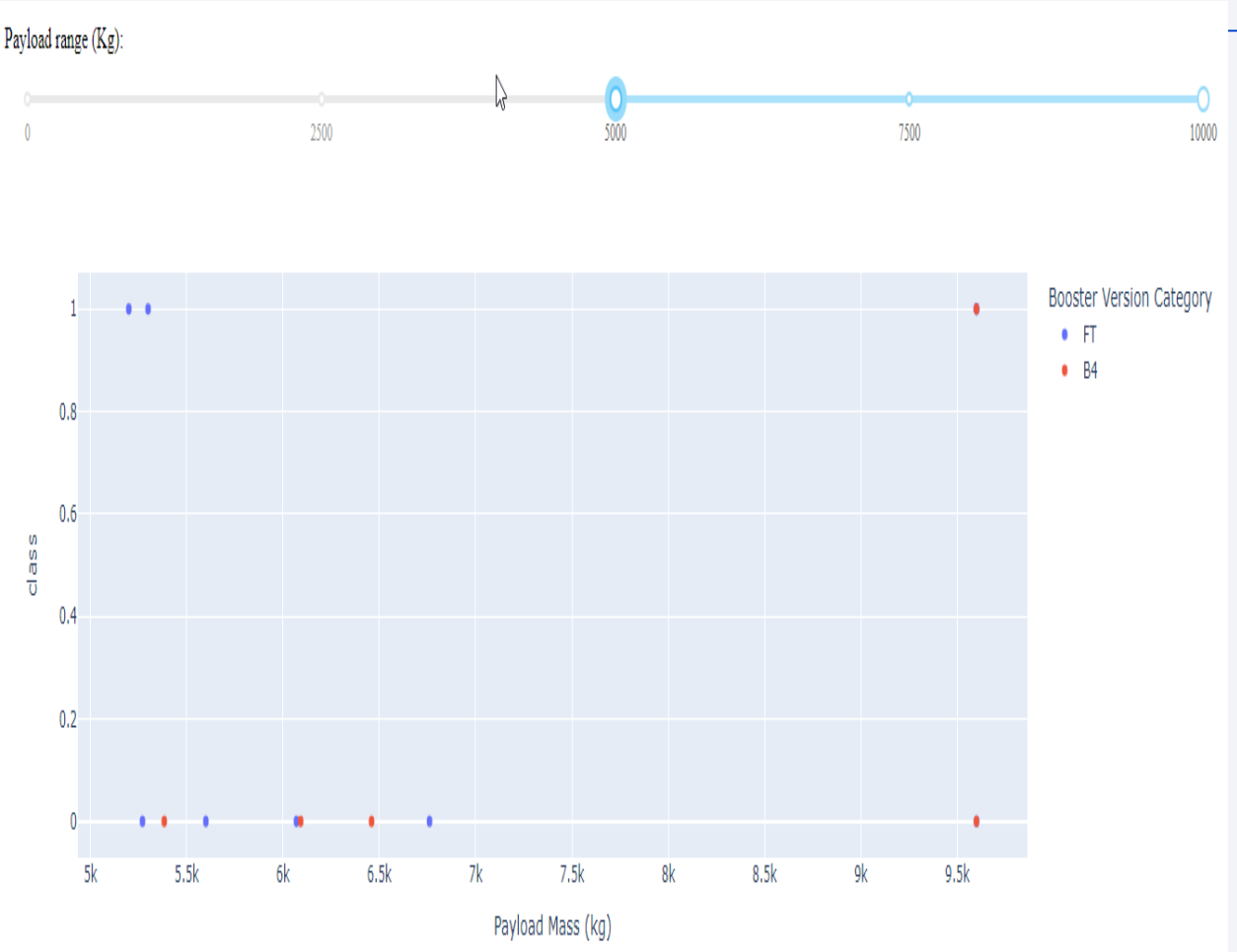

The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches

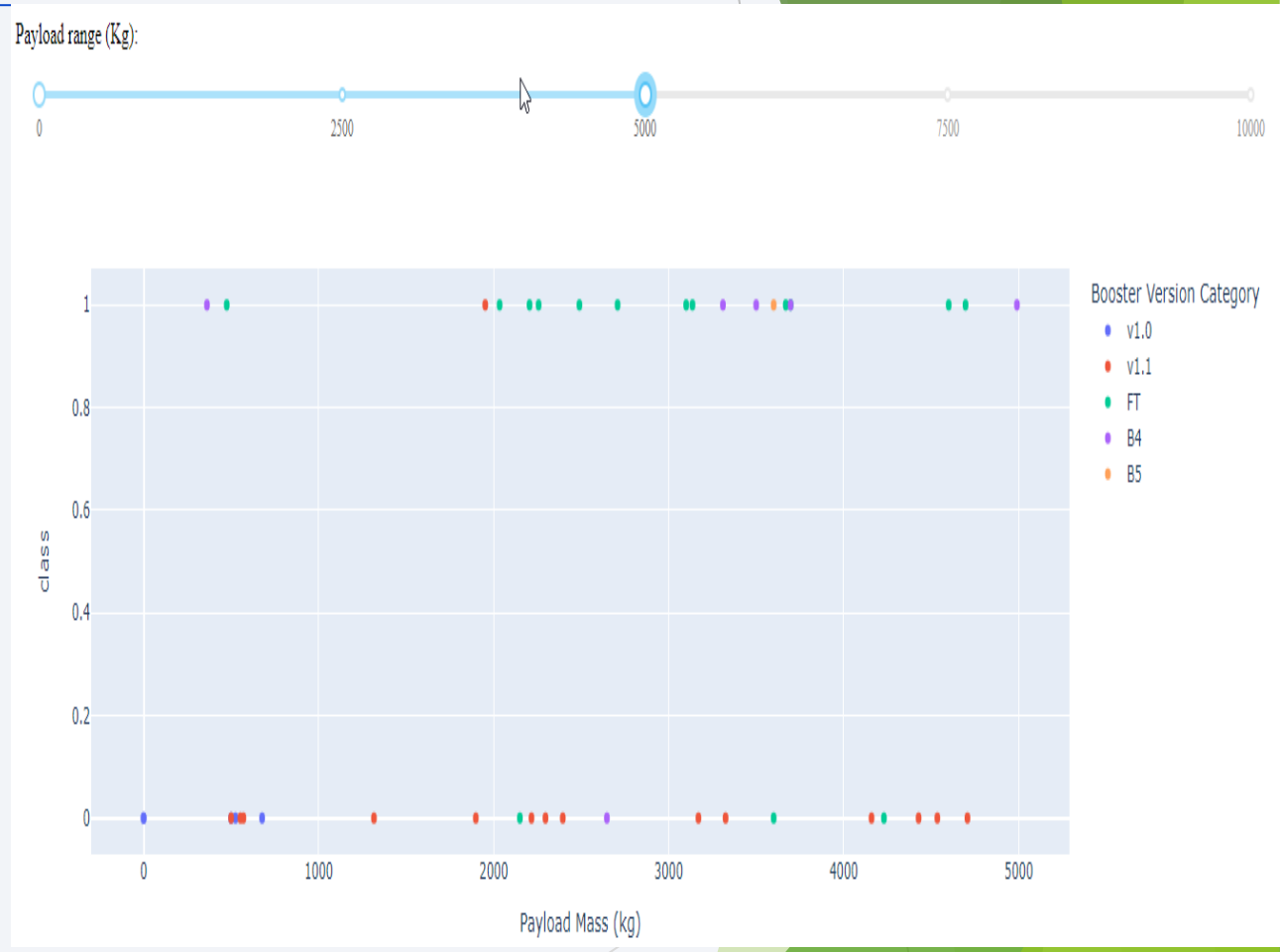
# Pie chart showing the Launch site with the highest launch success ratio




The chart shows that 76.9% of the total launches at site KSC LC-39A were successful. This is a the highest success rate of all the different launch sites.

# Payload vs. Launch Outcome for All Sites



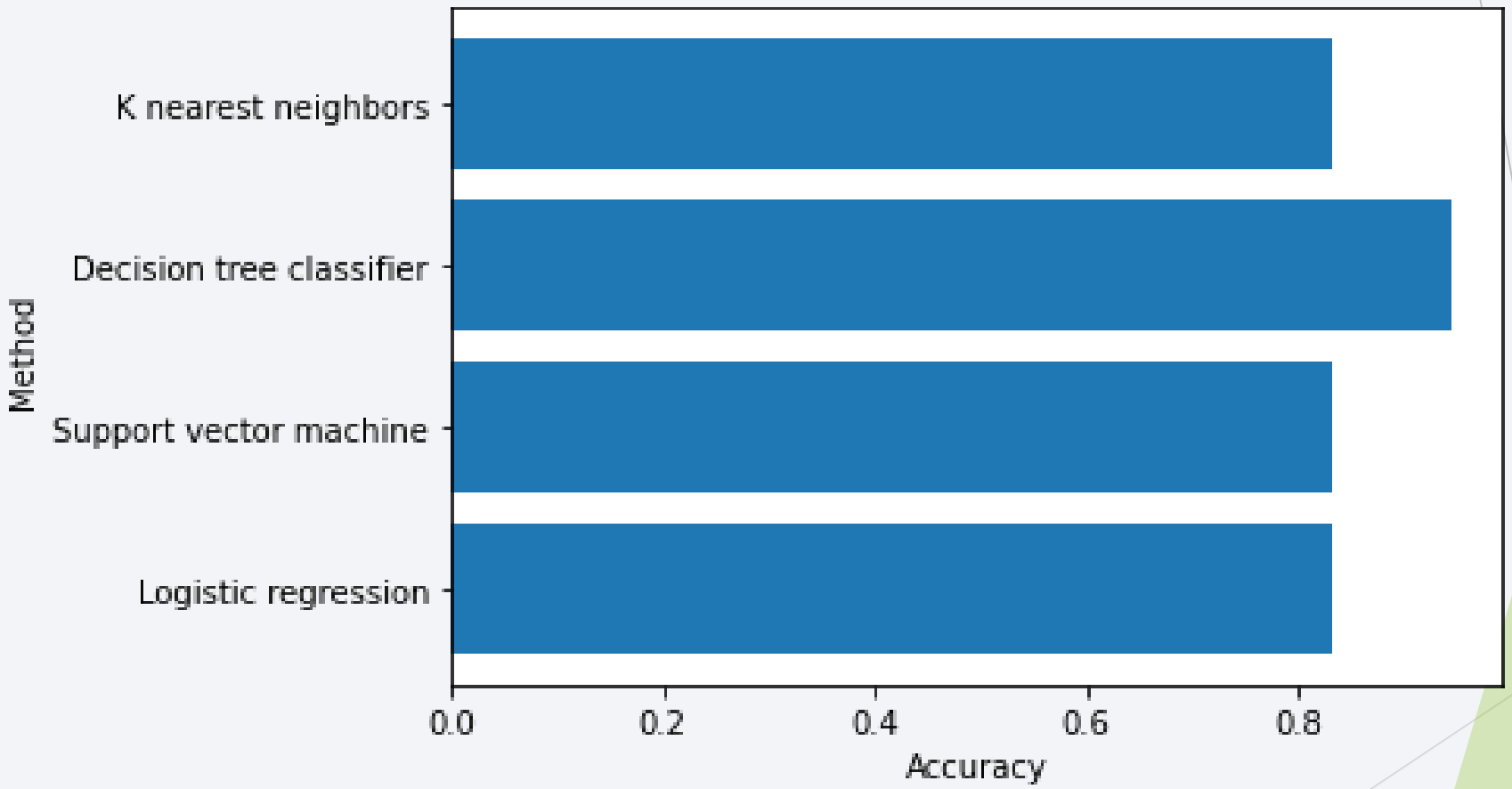




More booster versions acquire a success rate at lower payloads ( from 0 to 5000) as compared to booster versions with higher payloads (5000 to 10000)

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy




From the graph and analysis from the notebook, The decision tree classifier yields the highest accuracy

# Confusion Matrix

• The model predicted 12 successful landings when the True label was successful (True Positive) and 3 unsuccessful landings when the True label was failure (True Negative).

• The model also predicted 3 successful landings when the True label was unsuccessful landing (False Positive).

• The model generally predicted successful landings.

# Conclusions

We can conclude that:

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task.

# References


- Fortune Business Insights (2020). *Space launch services market*. https://www.fortunebusinessinsights.com/industry-reports/space-launch-services-market-101931
- CB Insights. *The Top 12 Reasons Startups Fail*. *https://www.cbinsights.com/research/startup-failure-reasons-top/*
- IBM. *Data Science Professional Certificate*. *https://www.coursera.org/professional-certificates/ibm-data-science*
- *Space.com. SpaceX Lands Orbital Rocket Successfully in Historic First. https://www.space.com/31420-spacex-rocket-landing-success.html*

# Appendix

- All notebooks and relevant materials used are in my github account.

  Github URL: https://github.com/BERI-0094/IBM-CAPSTONE-PROJECT

# Thank you!