IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Marc-André Hillebrandt
2002/05/05

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection & data wrangling
  - EDA with SQL
  - EDA with data visualization
  - Building an interactive map with Folium
  - Building a dashboard with Plotly Dash
  - Predictive analysis (classification)
- Summary of all results
  - EDA results
  - Predictive analysis results

Introduction

- Project background and context

In this capstone, we will predict if the Falcon 9 first stage 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. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. In this module, you will be provided with an overview of the problem and the tools you need to complete the course.

- Problems you want to find answers

- What influences if the rocket will land successfully?
- How does each relationship with certain rocket variables the success rate of a successful landing?
- What conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate?

Section 1

Methodology

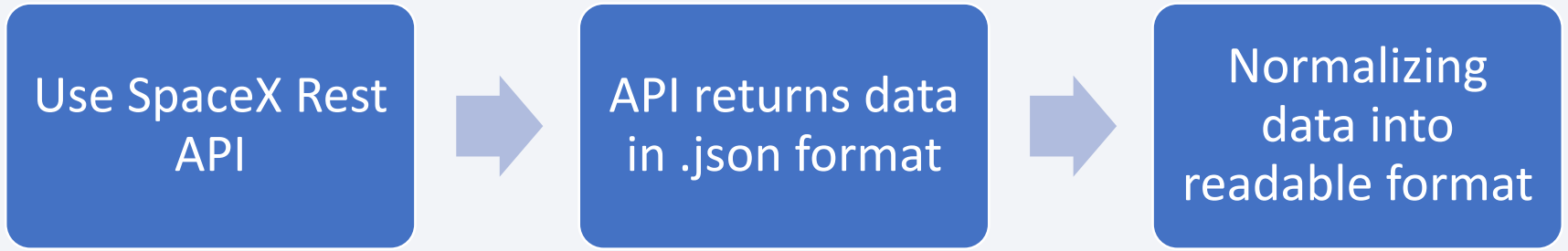
Methodology

Executive Summary

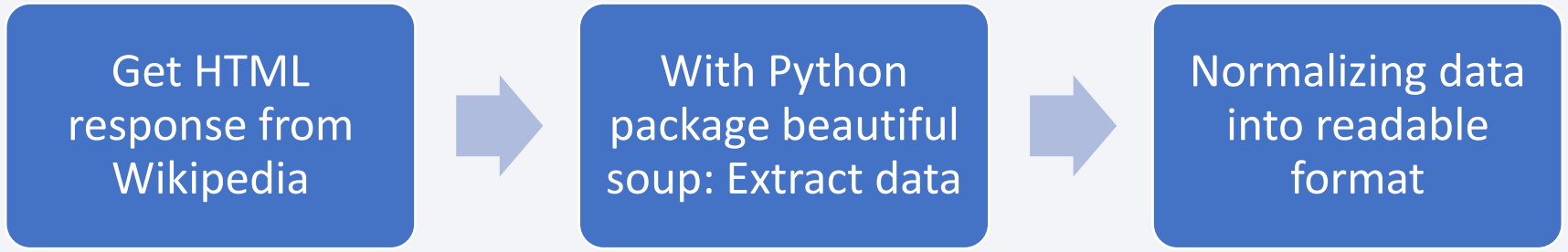
- Data collection methodology:
  - SpaceX Rest API & Web Scraping from Wikipedia
- Perform data wrangling
  - One Hot Encoding data fields for Machine Learning and dropping irrelevant columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- SpaceX Rest API & Web Scraping from Wikipedia
- SpaceX Rest API (Link: https://github.com/m-hillebrandt/data_science_capstone/blob/aa89b8d4b19eaec3def2f84bd83ccba87d19cf13/jupyter-labs-spacex-data-collection-api.ipynb):



- Web Scraping (Link: https://github.com/m-hillebrandt/data_science_capstone/blob/5fd81c462700b5875aa13a846df81382cf8f43ba/Webscraping.ipynb):



Data Wrangling

- Process:

1. Perform EDA on dataset to find some patterns in the data
2. Calculate the number of launches at each site
3. Calculate the number and occurrence of each orbit
4. Create a landing outcome label from Outcome column

- Link: https://github.com/m-hillebrandt/data_science_capstone/blob/5fd81c462700b5875aa13a846df81382cf8f43ba/Data%20wrangling.ipynb

EDA with Data Visualization

The following scatter plots were drawns:

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

The following bar chart was drawn:

- Mean vs. Orbit

The following line chart was drawn:

- Success Rate vs. Year

Link: https://github.com/m-hillebrandt/data_science_capstone/blob/5a1f807ebbf32b0b6076679cf61f97029b72b10d/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

The following SQL queries were performed (Link: https://github.com/m-hillebrandt/data_science_capstone/blob/5a1f807ebbf32b0b6076679cf61f97029b72b10d/EDA.ipynb):

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'KSC’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date where the successful landing outcome in drone ship was achieved.
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- 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

Build an Interactive Map with Folium

- To visualize the Launch Data into an interactive map. We took the Latitude and Longitude Coordinates at each launch site and added a Circle Marker around each launch site with a label of the name of the launch site.
- We assigned the dataframe launch_outcomes(failures, successes) to classes 0 and 1 with Green and Red markers on the map in a MarkerCluster()
- Using Haversine's formula we calculated the distance from the Launch Site to various landmarks to find various trends about what is around the Launch Site to measure patterns. Lines are drawn on the map to measure distance to landmarks
- Link: https://github.com/m-hillebrandt/data_science_capstone/blob/cc1aab616ec2108da8052d67c5072bc6be654f77/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

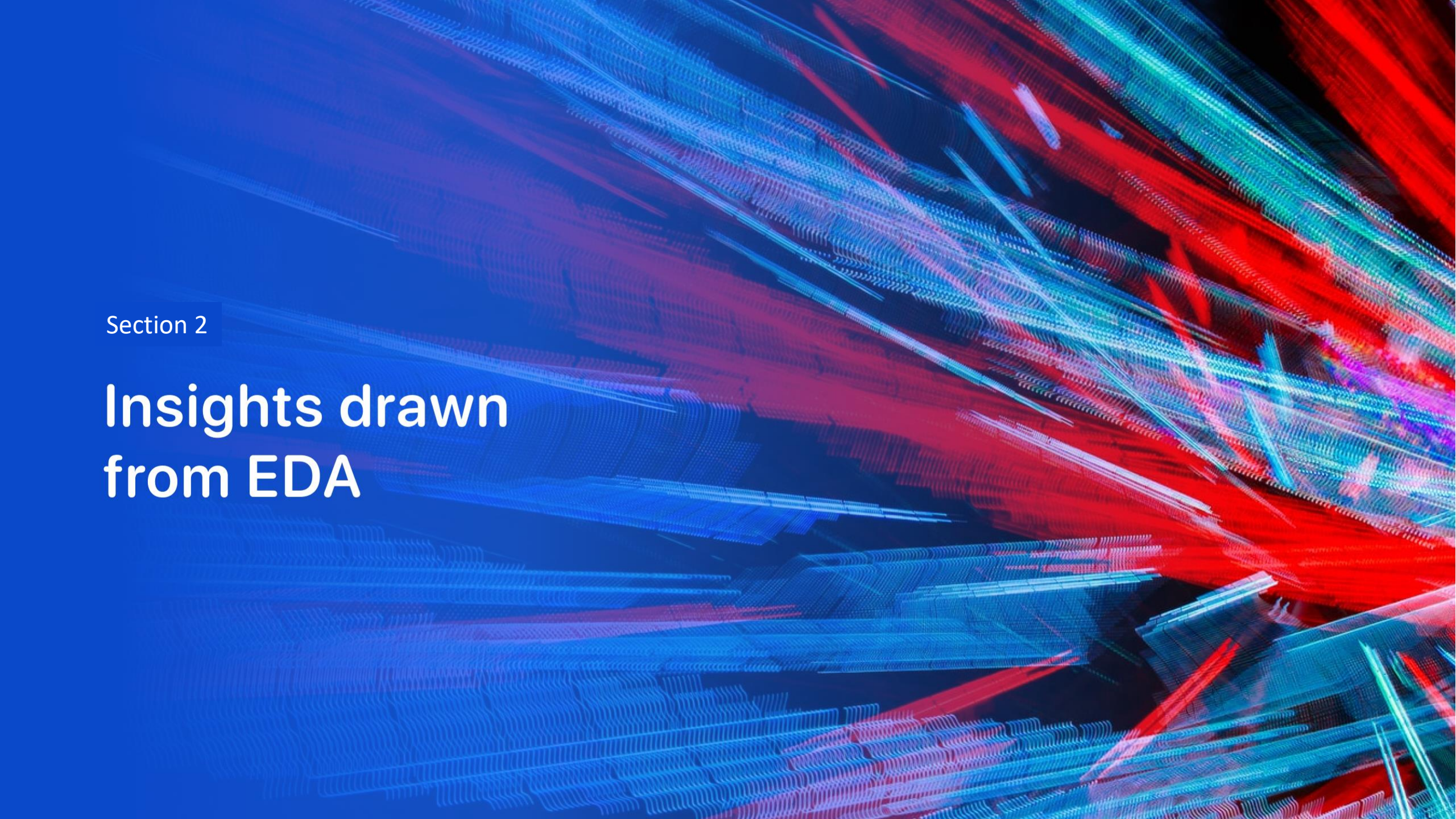
- To visualize the Launch Data into an interactive map. We took the Latitude and Longitude Coordinates at each launch site and added a Circle Marker around each launch site with a label of the name of the launch site.
- We assigned the dataframe launch_outcomes(failures, successes) to classes 0 and 1 with Green and Red markers on the map in a MarkerCluster()
- Using Haversine's formula we calculated the distance from the Launch Site to various landmarks to find various trends about what is around the Launch Site to measure patterns. Lines are drawn on the map to measure distance to landmarks
- Link: https://github.com/m-hillebrandt/data_science_capstone/blob/cc1aab616ec2108da8052d67c5072bc6be654f77/lab_jupyter_launch_site_location.ipynb

Predictive Analysis (Classification)

- **Building model:** Loading data set -> Transforming data -> Splitting data into training and test data -> Checking number of test samples -> Checking which ML algorithm to use -> Setting parameters
- **Evaluating model:** Checking accuracy -> Tuning hyperparameters -> Plotting confusion matrix
- **Improving model:** Feature engineering -> Tuning algorithm
- **Best performing classification model**
- **Link:** https://github.com/m-hillebrandt/data_science_capstone/blob/cc1aab616ec2108da8052d67c5072bc6be654f77/Predictive_Analysis.ipynb

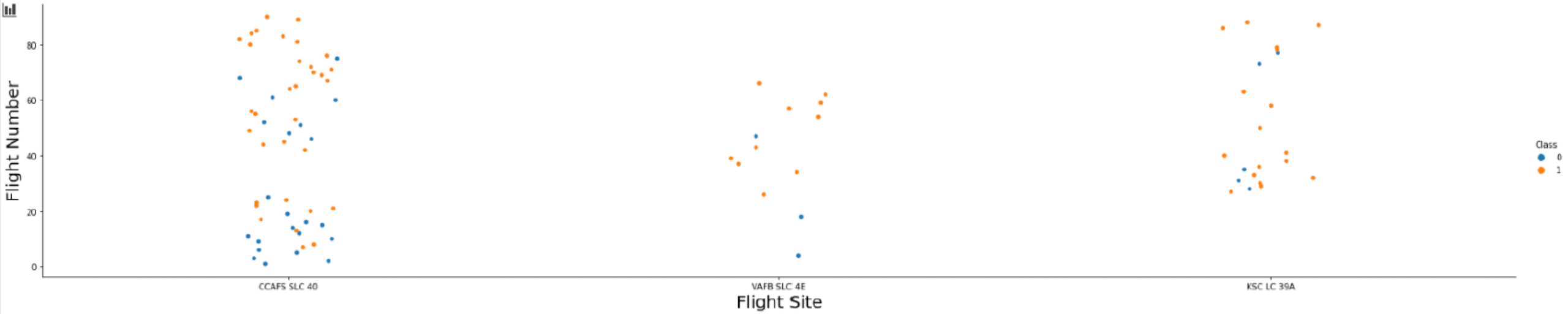
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

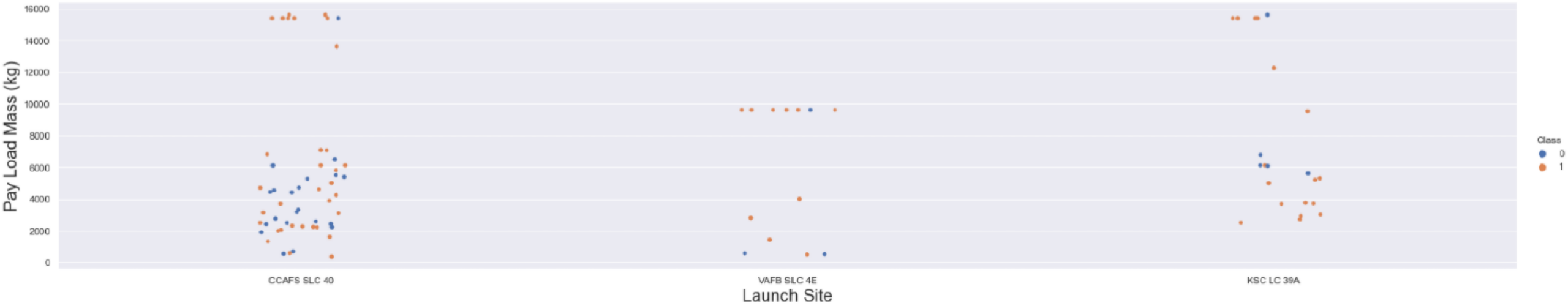
Flight Number vs. Flight Site



The more amount of flights at a launch site the greater the success rate at a launch site.

Payload vs. Launch Site

Payload Mass vs. Launch Site

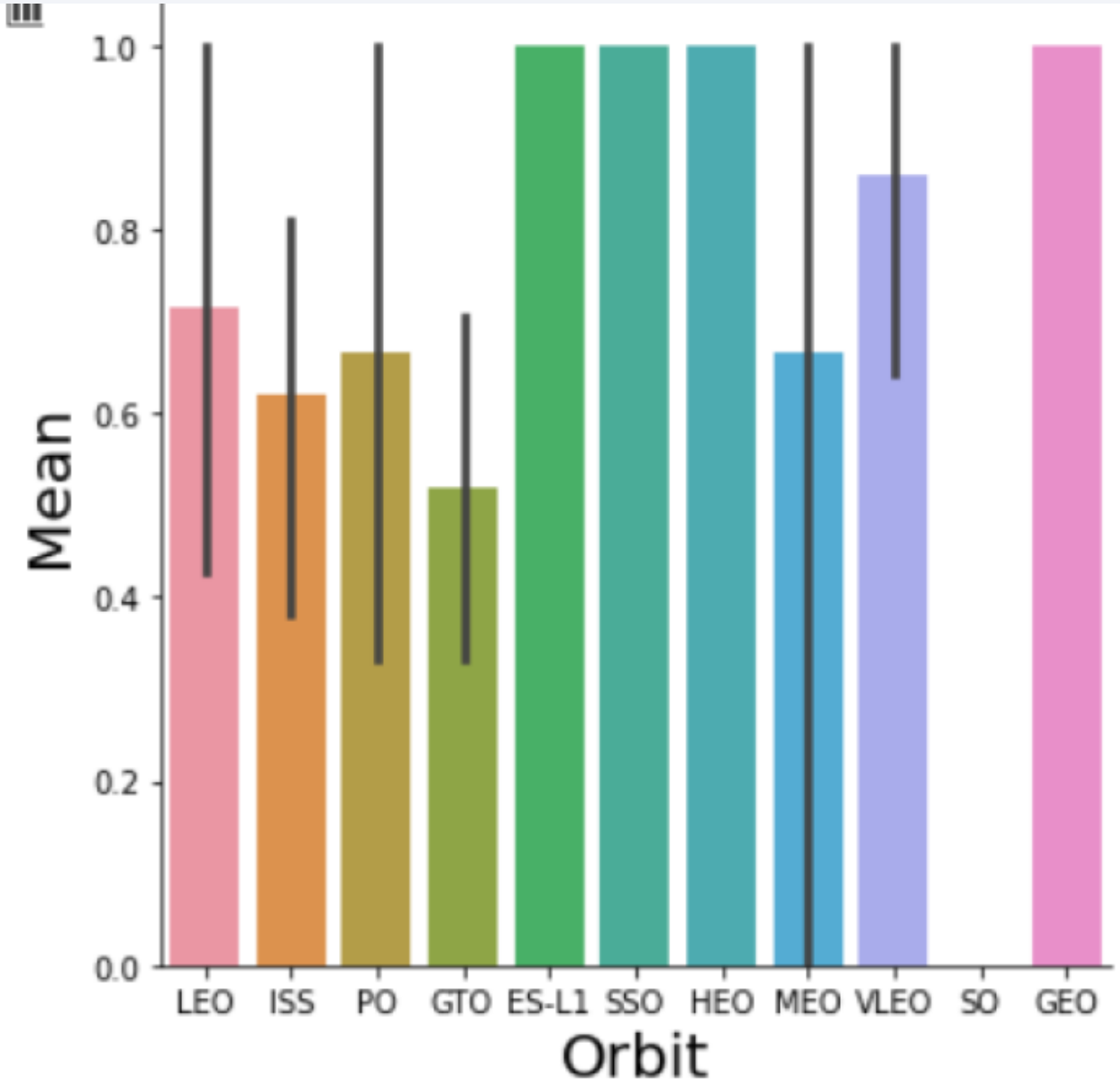


The greater the payload mass for Launch Site CCAFS SLC 40 the higher the success rate for the Rocket.
There is not quite a clear pattern to be found using this visualization to make a decision if the Launch Site is dependant on Pay Load Mass for a success launch.

Success Rate vs. Orbit Type

Success rate vs. Orbit type

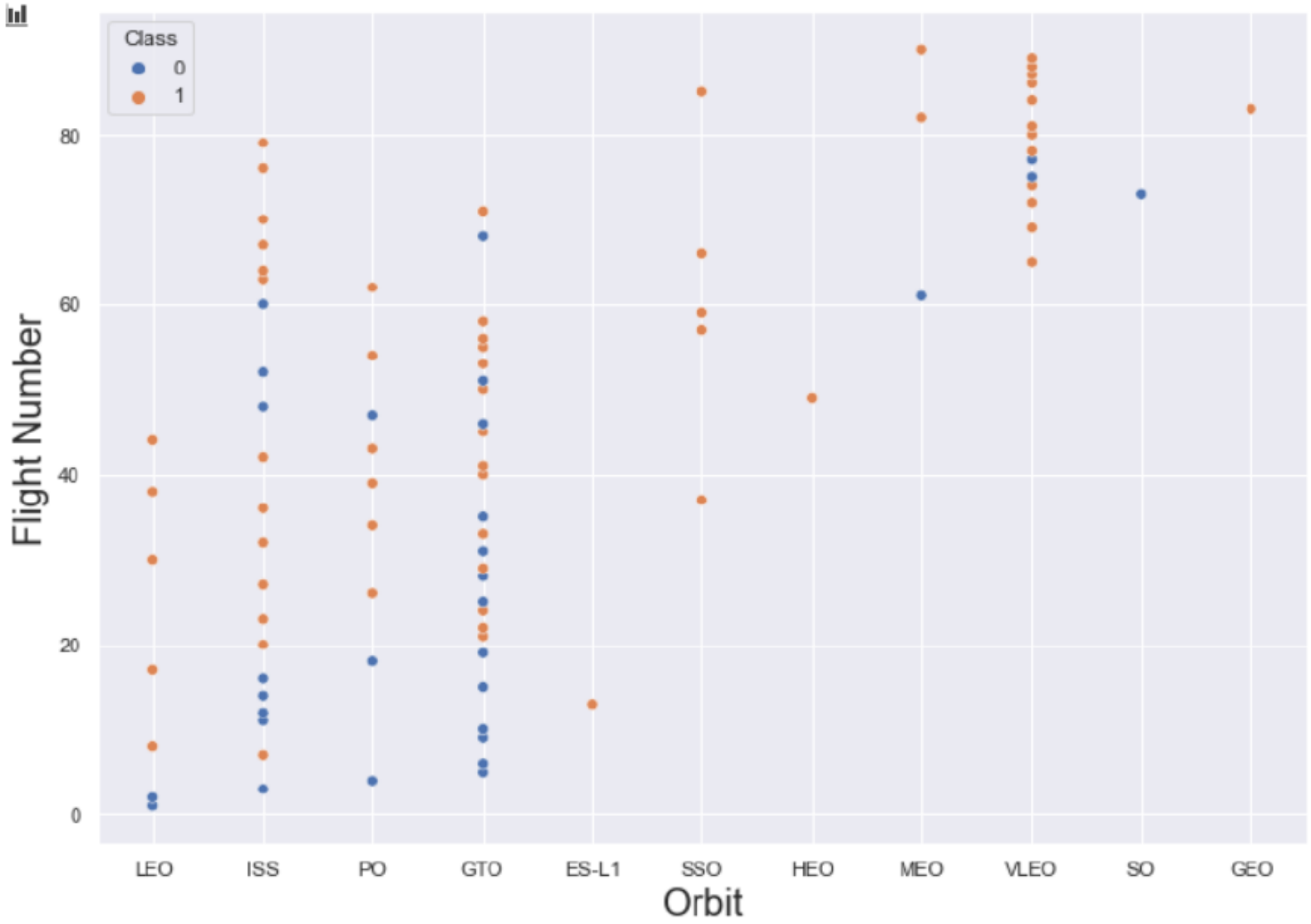
Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate



Flight Number vs. Orbit Type

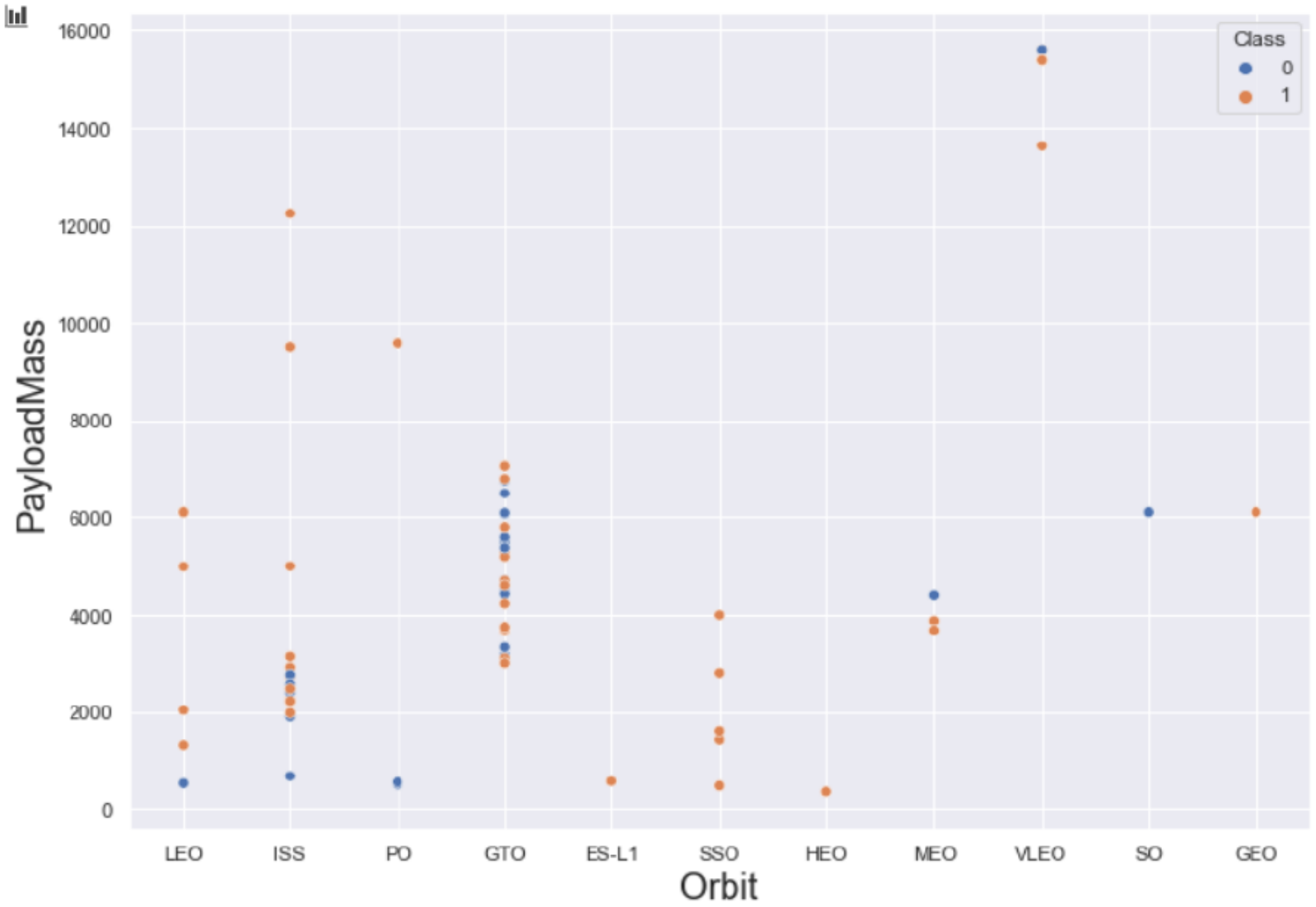
Flight Number vs. Orbit type

You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

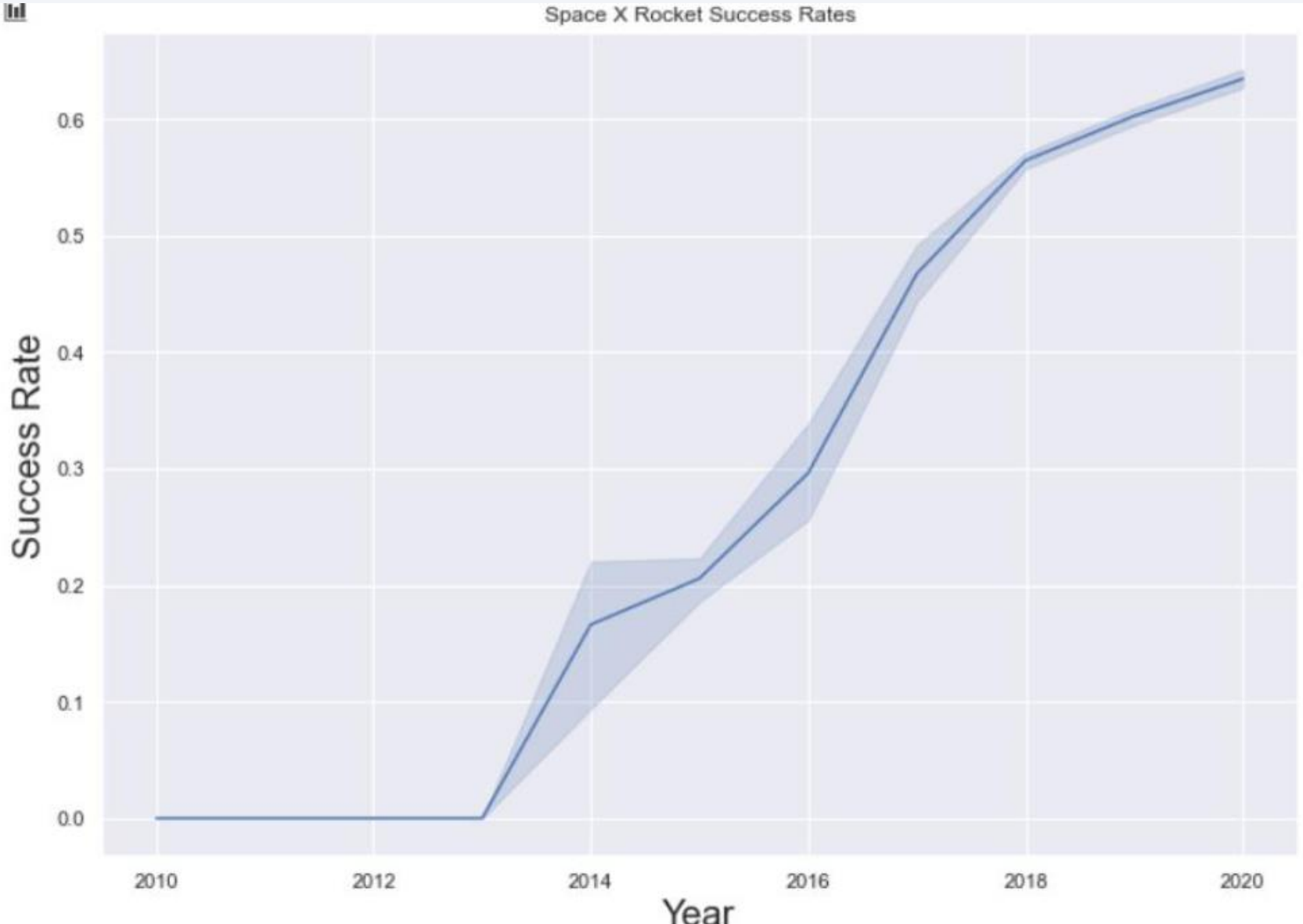
Payload vs. Orbit type



You should observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success Yearly Trend

Launch success yearly trend



you can observe that the success rate since 2013 kept increasing till 2020

All Launch Site Names

```
In [8]: %sql SELECT Distinct LAUNCH_SITE FROM SPACEXTBL
 * ibm_db_sa://kcq64325:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
```

Out[8]:

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

Explanation: Using the word ***DISTINCT*** in the query means that it will only show unique values in the ***Launch_Site*** column from ***SPACEXTBL***

Launch Site Names Begin with 'CCA'

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

```
 * ibm_db_sa://kcq64325:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/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	None	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	None	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	None	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	None	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	None	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Explanation: Limiting the query to five entries will only show 5 records from SPACEXTBL and LIKE keyword has a wild card suggesting that the LAUNCH_SITE must start with CCA.

Total Payload Mass

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

```
 * ibm_db_sa://kcq64325:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
```

1
45596

Explanation: Using the function SUM summates the total in the column PAYLOAD_MASS_KG_
The WHERE clause filters the dataset to only perform calculations on the customer "NASA (CRS)".
Result: 45,596

Average Payload Mass by F9 v1.1

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

```
 * ibm_db_sa://kcq64325:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
```

1
2928.400000

Explanation: Using the function AVG averages the total in the column PAYLOAD_MASS_KG_
The WHERE clause filters the dataset to only perform calculations for which the booster version is “F9 v1.1”.
Result: 2,928.40

First Successful Ground Landing Date

```
%sql SELECT min(DATE) FROM SPACEXTBL WHERE LANDING__OUTCOME='Success (ground pad)'
```

```
 * ibm_db_sa://kcq64325:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
```

1
2015-12-22

Explanation: Using the function MIN selects the first date for which the LANDING_OUTCOME equals “Success (ground pad)”
Result: 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ between 4000 and 6000 AND LANDING__OUTCOME='Success (drone ship)'
```

```
 * ibm_db_sa://kcq64325:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
```

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

Explanation: Selecting only the BOOSTER_VERSION for which the PAYLOAD_MASS_KG_ is between 4000 and 6000 and where the LANDING_OUTCOME is “Success (drone ship)”
Result: four versions as seen above

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT COUNT(*) FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE '%Success%' OR MISSION_OUTCOME LIKE '%Failure%'
```

```
 * ibm_db_sa://kcq64325:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
```

1
101

Explanation: Total number for which the MISSION_OUTCOME was either (a) a success or (b) a failure.
Result: 101

Boosters Carried Maximum Payload

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

```
 * ibm_db_sa://kcq64325:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/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

Explanation: Selecting only the BOOSTER_VERSION which have carried the maximum payload mass. Therefore, a subquery had to be implemented which at first selected the maximum payload_mass_kg.

2015 Launch Records

```
%sql SELECT TO_CHAR(TO_DATE(MONTH("DATE"), 'MM'), 'MONTH') AS MONTH_NAME, \
    LANDING__OUTCOME AS LANDING__OUTCOME, \
    BOOSTER_VERSION AS BOOSTER_VERSION, \
    LAUNCH_SITE AS LAUNCH_SITE \
    FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND "DATE" LIKE '%2015%'
```

```
 * ibm_db_sa://kcq64325:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
```

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

Explanation: Selecting those records in year 2015 with their month, failure landing_outcomes in drone ship, booster versions, and launch_site

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

```
%sql SELECT "DATE", COUNT(LANDING__OUTCOME) as COUNT FROM SPACEXTBL \
    WHERE "DATE" BETWEEN '2010-06-04' and '2017-03-20' AND LANDING__OUTCOME LIKE '%Success%' \
    GROUP BY "DATE" \
    ORDER BY COUNT(LANDING__OUTCOME) DESC
```

```
 * ibm_db_sa://kcq64325:***@dashdb-txn-sbox-yp-dal09-04.services.dal.bluemix.net:50000/BLUDB
Done.
```

DATE	COUNT
2015-12-22	1
2016-04-08	1
2016-05-06	1
2016-05-27	1
2016-07-18	1
2016-08-14	1
2017-01-14	1
2017-02-19	1

Explanation: Counting successful landings between 2010-06-04 and 2017-03-20 by date

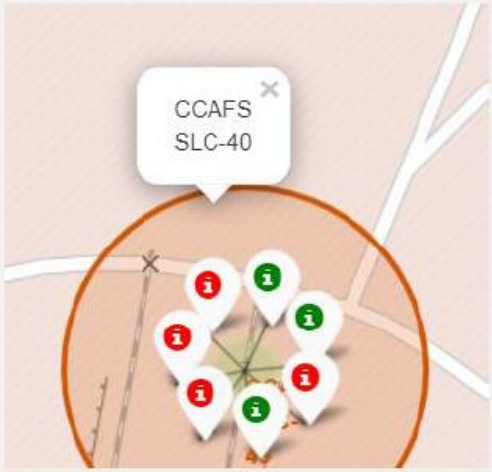
Section 3

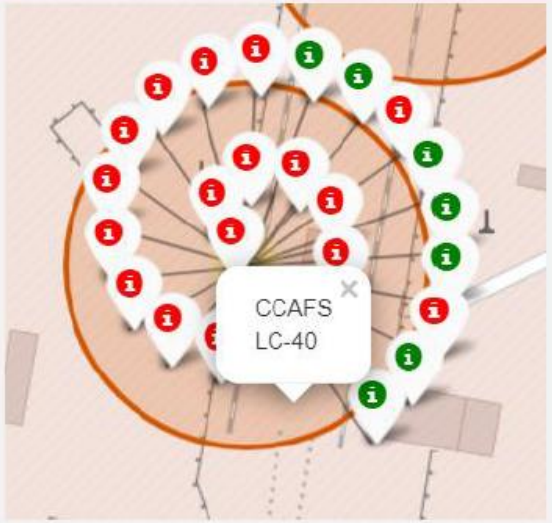
Launch Sites Proximities Analysis

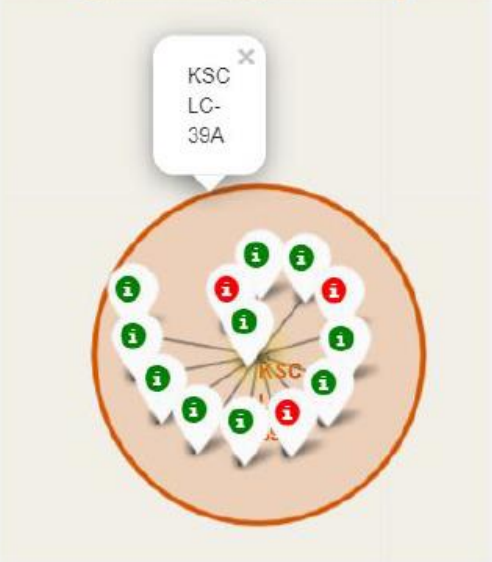
All launch sites global map markers

We can see that the SpaceX launch sites are in the United States of America coasts. Florida and California

Colour Labelled Markers





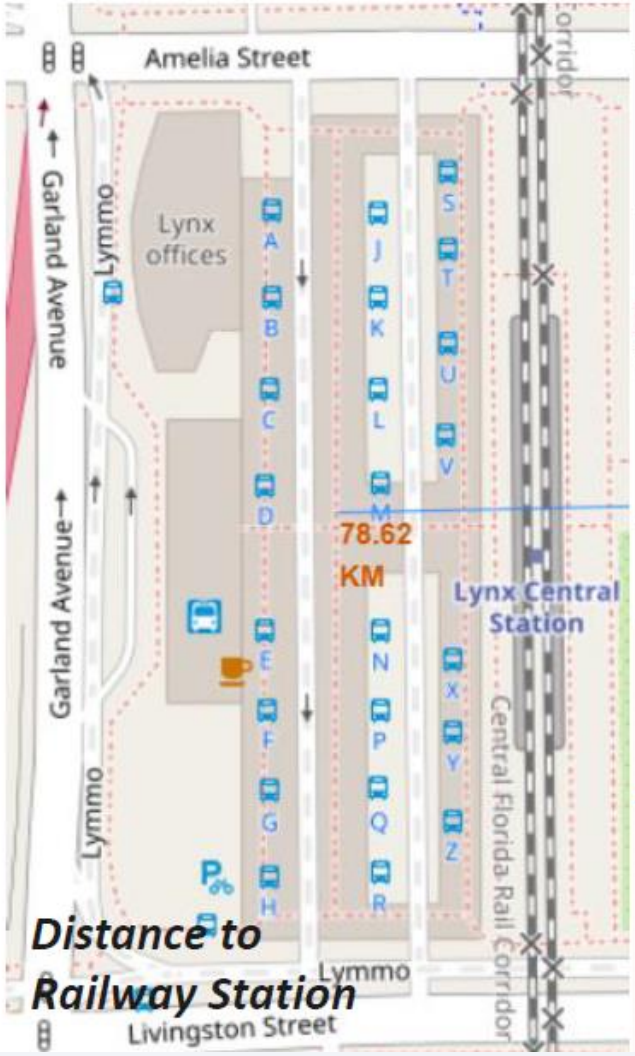


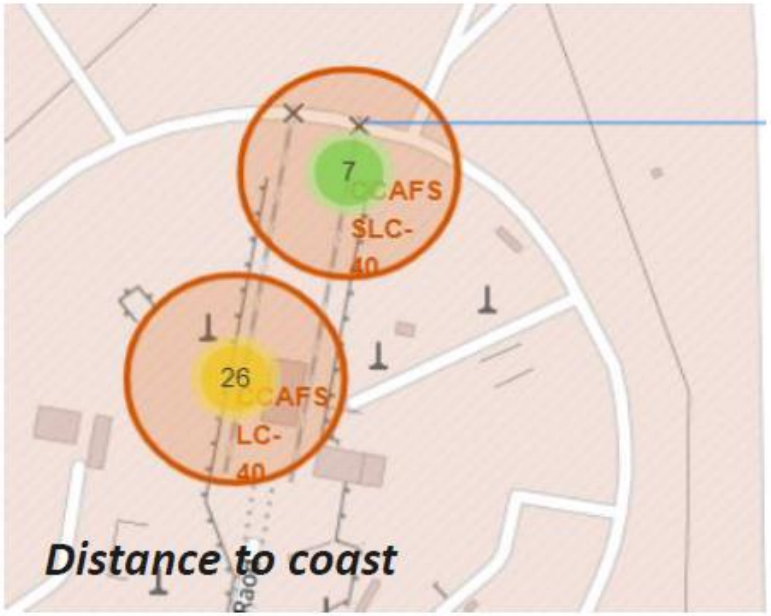
Florida Launch Sites

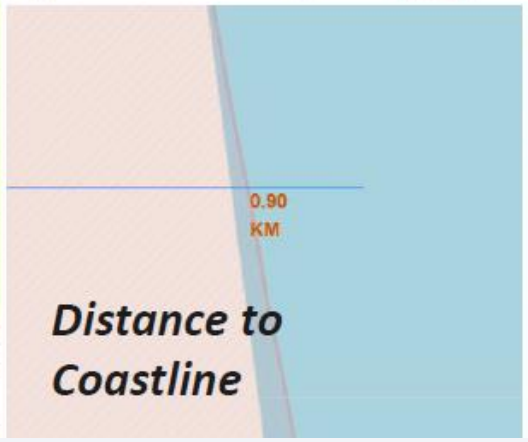
Green Marker shows successful Launches and Red Marker shows Failures

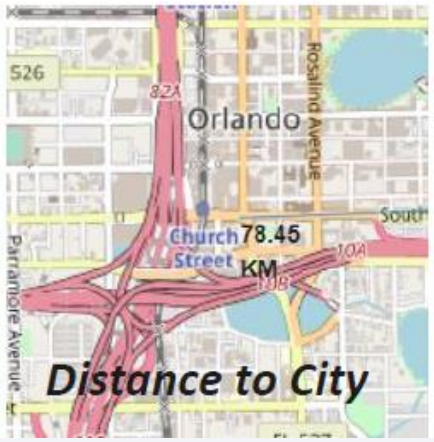
California Launch Site

Launch Sites Distance to Landmarks

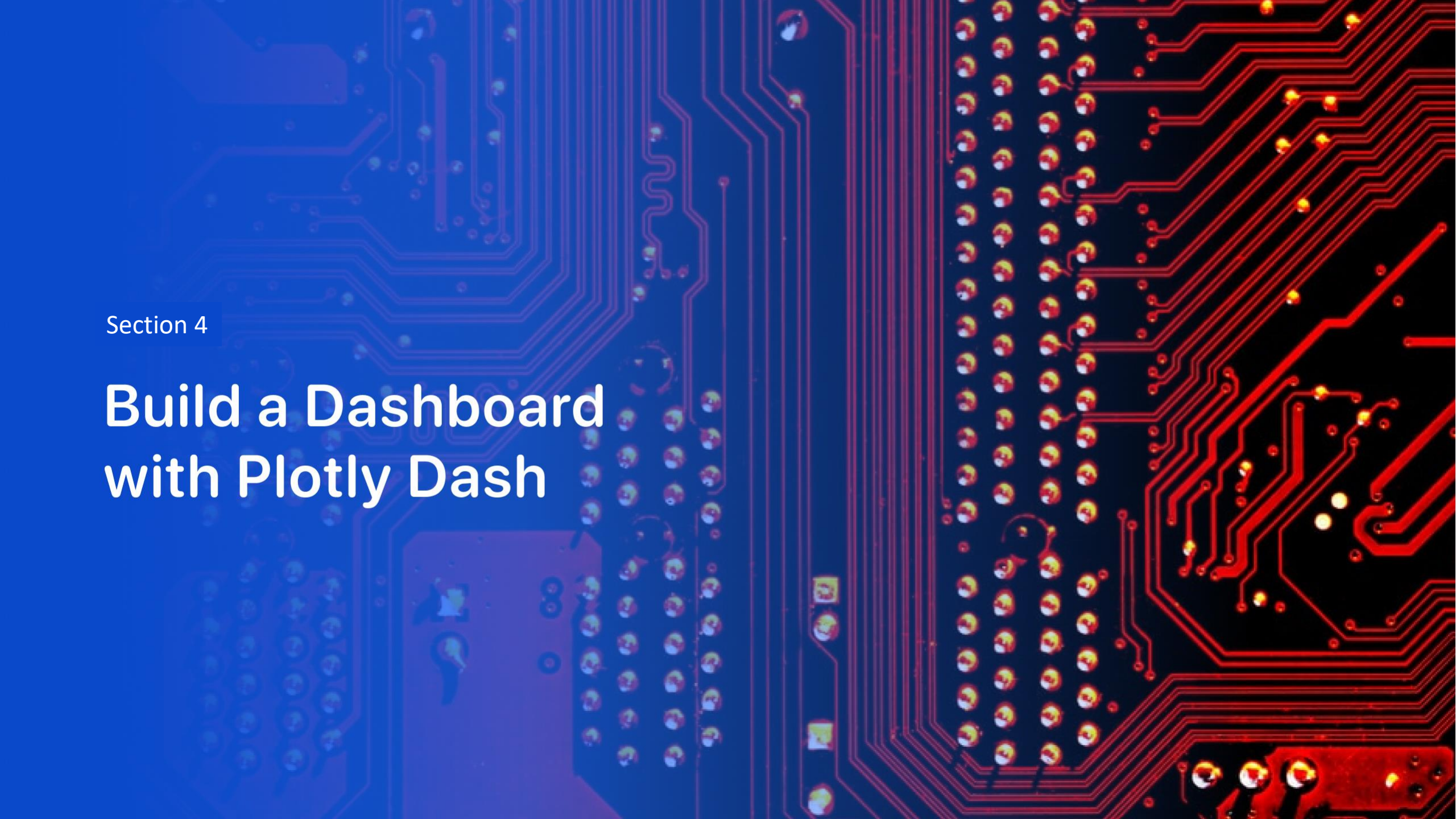




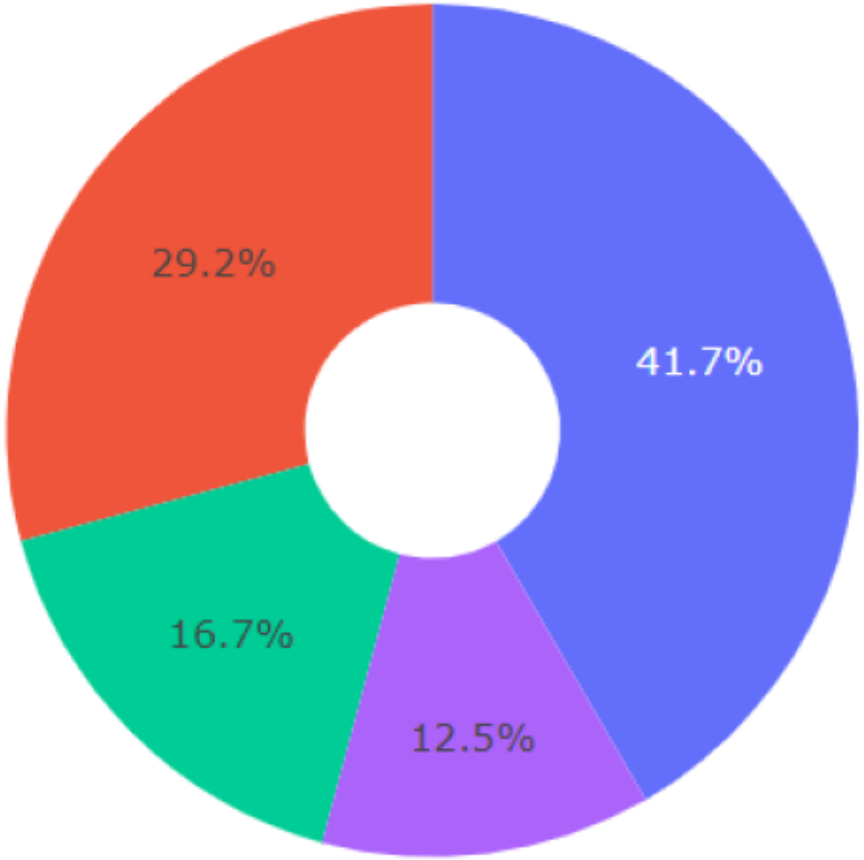




- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes


Section 4
Build a Dashboard with Plotly Dash

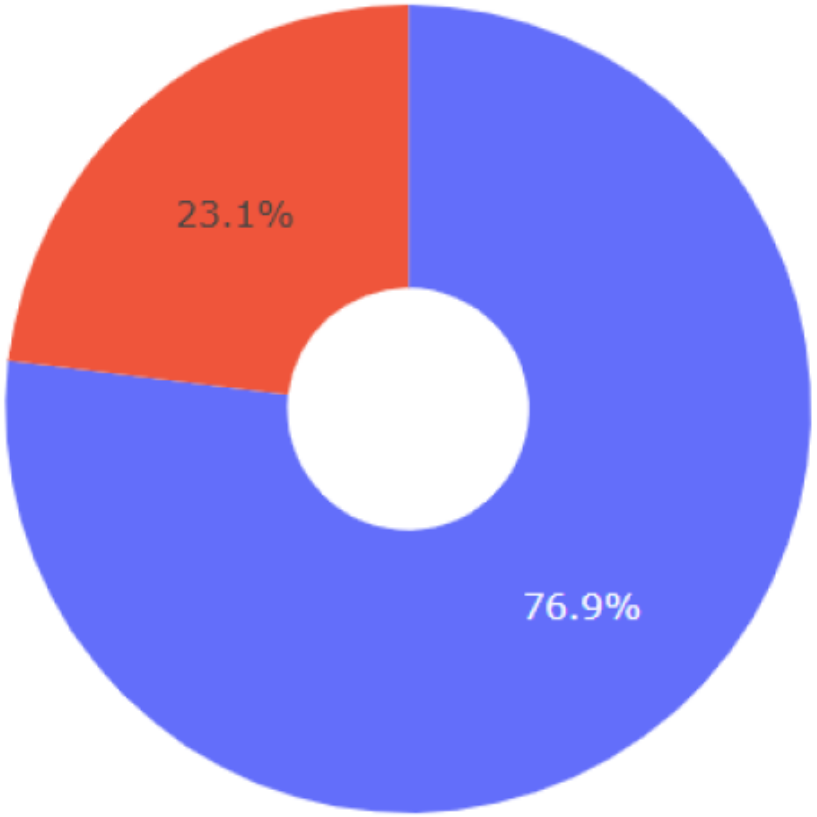
Total Successful Launches by Sites



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

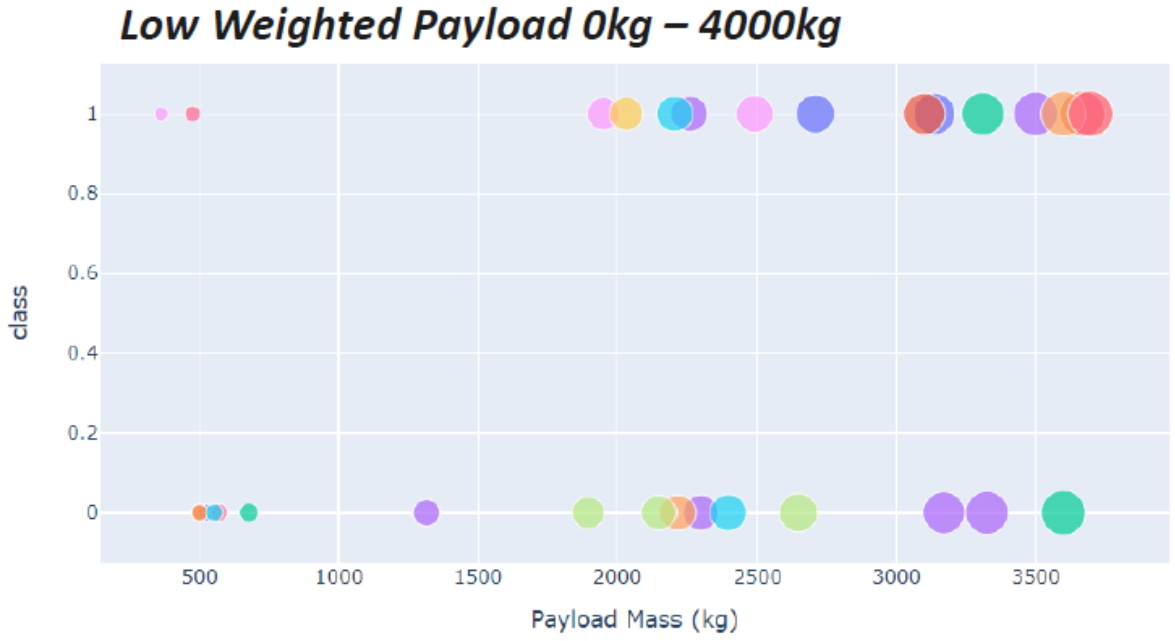
We can see that KSC LC-39A had the most successful launches from all the sites

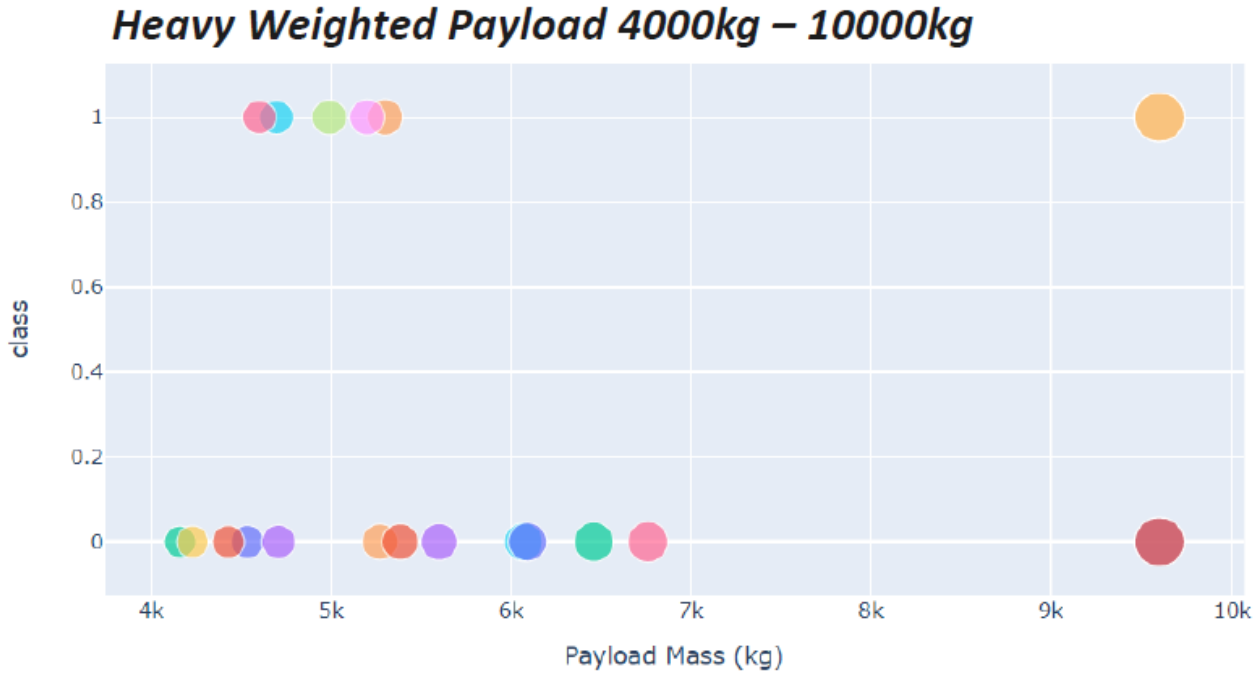
Highest Launch Success Rate



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

Payload vs. Launch Success





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

Section 5

Predictive Analysis (Classification)

Classification Accuracy

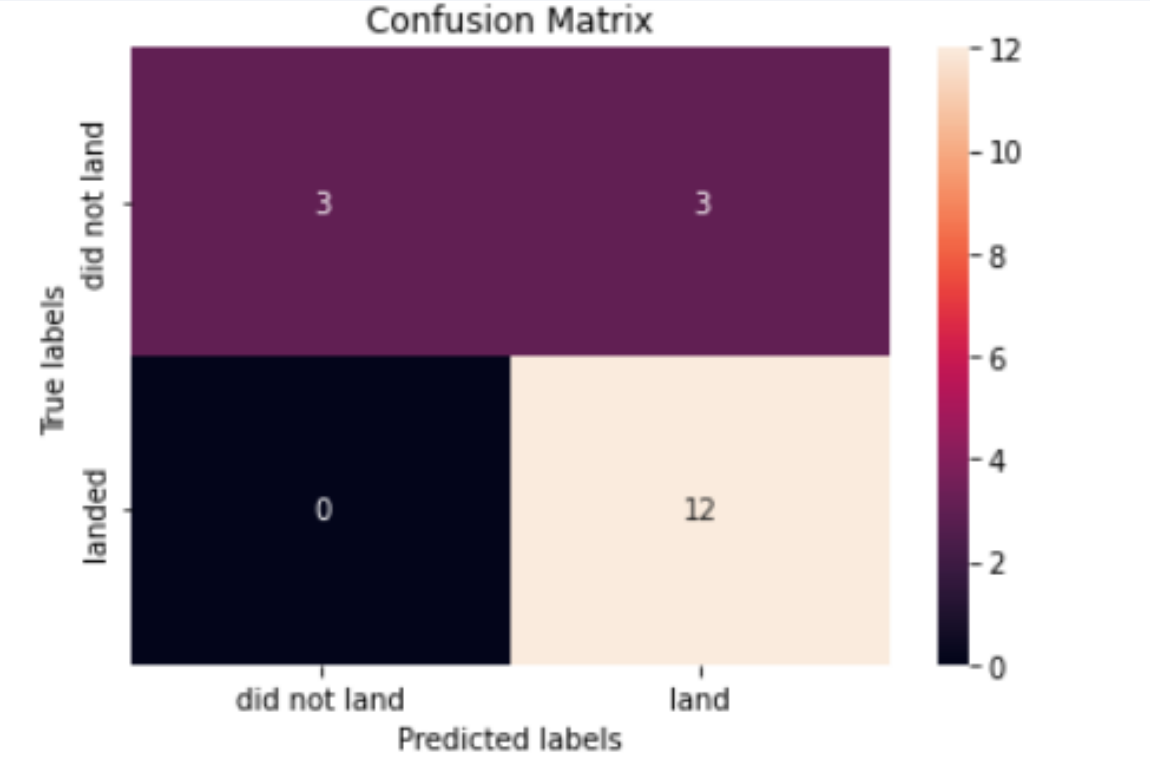
TASK 12

Find the method performs best:

```
print('Accuracy for Logistics Regression method:', logreg_cv.score(X_test, Y_test))
print( 'Accuracy for Support Vector Machine method:', svm_cv.score(X_test, Y_test))
print('Accuracy for Decision tree method:', tree_cv.score(X_test, Y_test))
print('Accuracy for K nearsdt neighbors method:', knn_cv.score(X_test, Y_test))
```

```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.8333333333333334
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```

Confusion Matrix



Conclusions

- Best algorithm for this data set is the Tree algorithm
- Low weighted payloads perform better than heavier payloads