Winning Space Race with Data Science

Germán Camargo Ortega
09.12.2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization charts
  - Dashboard
- Conclusion
  - Findings & implications
- Appendix

Executive Summary

- The following slides summarize key methodologies and findings of a study to estimate the total cost for launches, by predicting successful landings of the first stage of rockets of Space X; and where is the best place to make launches. This will help determine the viability of the new company Space Y to compete with Space X

- Summary of methodologies
  - The following methodologies were used to analyze data
  - Data Collection using web scraping and SpaceX API
  - Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive
  - Visual analytics
  - Machine Learning Prediction.
- Summary of all results
  - It was possible to collected valuable data from public sources;
  - EDA allowed to identify which features are the best to predict success of launchings;
  - Machine Learning Prediction showed the best model to predict which characteristics are
  - important to drive this opportunity by the best way, using all collected data.

Introduction

- Project background and context
  - SpaceX is successful because their rocket launches are relatively inexpensive.
  - because SpaceX can reuse the first stage, which is quite large and expensive, compared to stage 2 and 3 lounches.
  - I will take the role of a data scientist working for a new rocket company - Space Y that would like to compete with SpaceX.
  - My job is basically to determine the price of each launch based on SpaceX data.

- Problems I want to find answers = Objectives
  - to evaluate the viability of the new company Space Y to compete with SpaceX.
  - I want to know the best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets
  - Where is the best place to make launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from SpaceX was obtained from 2 sources:
    a. SpaceX API (https://api.spacexdata.com/v4/rockets/)
    b. WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling and EDA using visualization and SQL
  - EDA was done to find some patterns in the data and determine what would be the label for training supervised models.
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing, analyzing and engineering features.

Methodology

Executive Summary

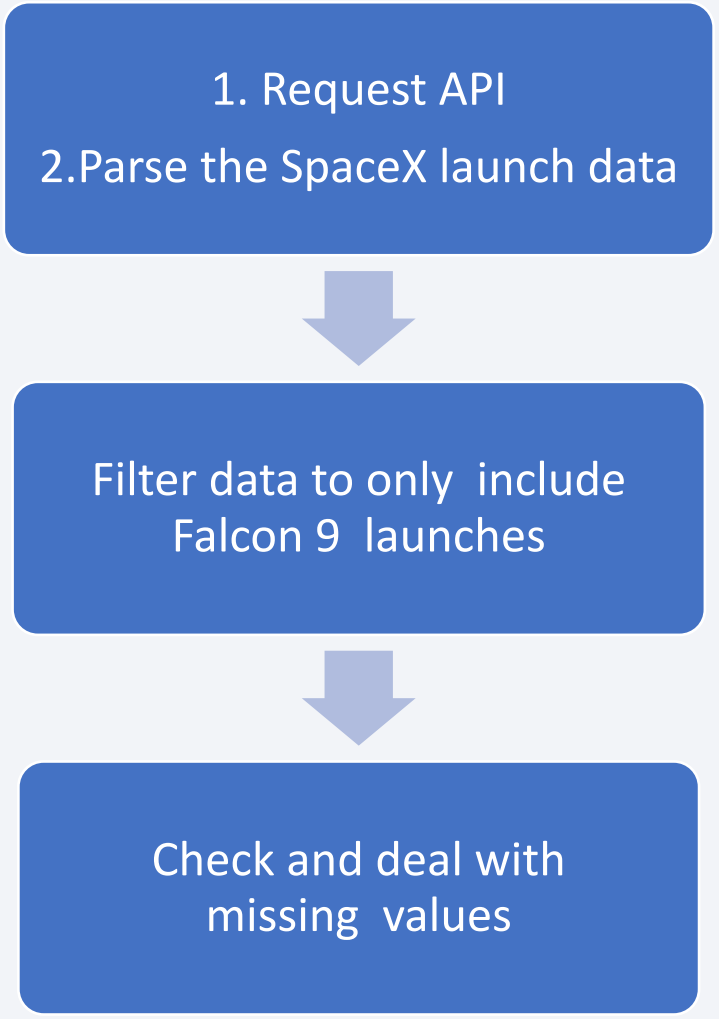
- Perform interactive visual analytics using Folium and Plotly Dash
  - The data was used to visualize the distribution, composition and relationships between features.
- Perform predictive analysis using classification models
  - create a column for the class
  - Standardize the data
  - Split into training data and test data
  - -Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
  - Find the method performs best using test data (i.e. best accuracy score)

Data Collection

- Data from SpaceX was obtained from 2 sources:
  - SpaceX API (https://api.spacexdata.com/v4/rockets/)
  - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

- SpaceX offers a public API.
- Data can be obtained here and used.
- The flowchart beside how this API was used

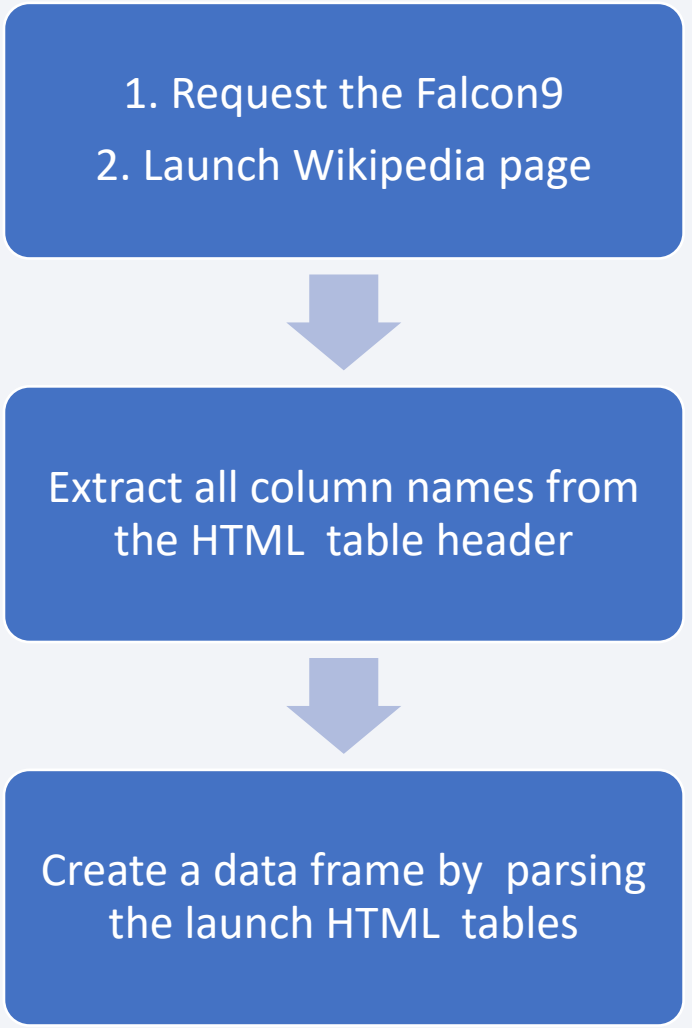


- Source code https://github.com/GermanCamargoOrtega/IBM_Capstone/blob/main/01.%20Hands-on%20Lab_%20Data%20Collection%20API%20Lab.ipynb

Data Collection - Scraping

- SpaceX launches data was obtained from Wikipedia as shown in the flowchart.

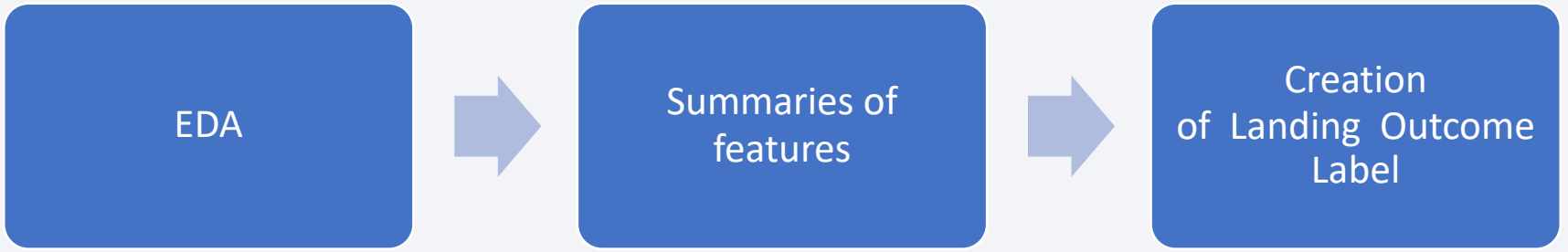
- Source code: https://github.com/GermanCamargoOrtega/IBM_Capstone/blob/main/02.%20Hands-on_%20Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

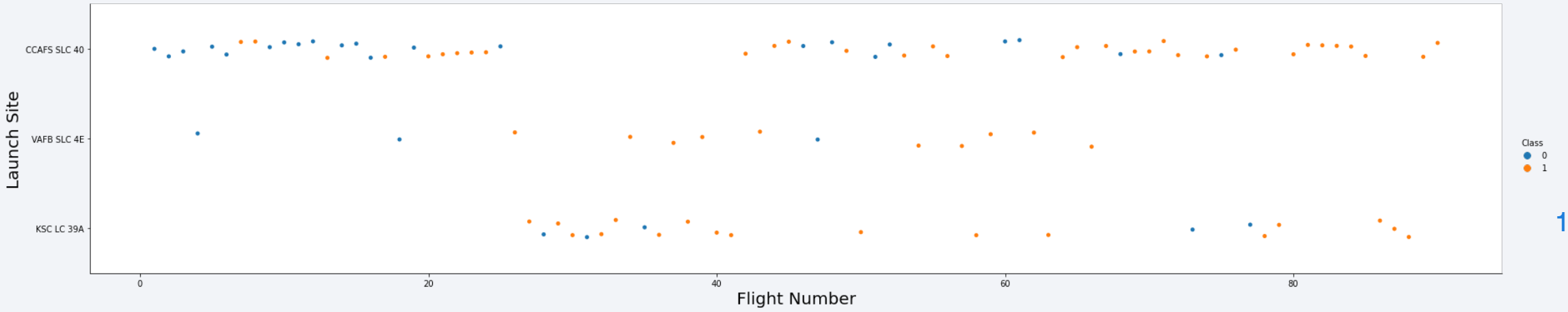
1. Some preliminary EDA was initially done on the data.
2. Following features (summaries) were calculated:
   a. number launches per site
   b. number and occurrences of each orbit
   c. number and occurrences of mission outcome per orbit type
3. The column containing the landing outcome label was created from Outcome column.

- Source code https://github.com/GermanCamargoOrtega/IBM_Capstone/blob/main/03.%20Hands-on%20Lab_%20Data%20Wrangling.ipynb



EDA with Data Visualization

- Relationships between several features were visualized using different scatter and bar plots.
- The relationships between following features were checked:
  - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit
- Source code https://github.com/GermanCamargoOrtega/IBM_Capstone/blob/main/05.%20Hands-on%20Lab_%20EDA%20with%20Visualization.ipynb



EDA with SQL

- SQL queries are found in: https://github.com/GermanCamargoOrtega/IBM_Capstone/blob/main/04.%20Hands-on%20Lab_%20EDA%20with%20SQL.ipynb
- In summary, these queries were done:

1. Names of the unique launch sites in the space mission.
2. Top 5 launch sites whose name begin with the string 'CCA'.
3. Total payload mass carried by boosters launched by NASA (CRS).
4. Average payload mass carried by booster version F9v1.1.
5. Date when the first successful landing outcome in ground pad was achieved.
6. Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000kg.
7. Total number of successful and failure mission outcomes.
8. Names of the booster versions which have carried the maximum pay load mass.
9. Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.

Build an Interactive Map with Folium

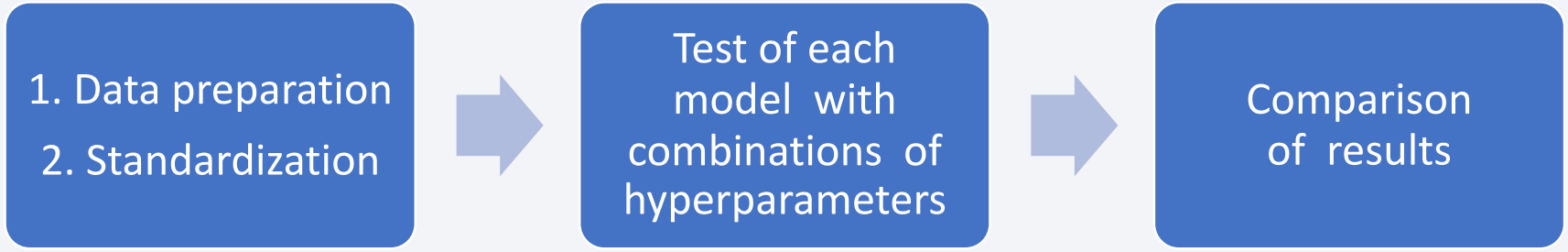
- Summary of map objects created and added to a folium map:
  - Markers, circles, lines and marker clusters were used
  - Markers => points like launch sites
  - Circles => highlighted areas around specific coordinates (e.g. NASA Johnson Space Center)
  - Marker clusters => groups of events in each coordinate (e.g. like launches in a launch site)
  - Lines => indicate distances between two coordinates.
- Source code https://github.com/GermanCamargoOrtega/IBM_Capstone/blob/main/06.%20Hands-on%20Lab_%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Summary of plots/graphs and interactions added to a dashboard :
  - Data was visualized using (1) Percentage of launches by site, and (2) Pay load range.
  - This allowed a quick analysis of relationships between payloads and launch sites.
  - This in turn helps identifying best places to launch (according to pay loads).
- Source code https://github.com/GermanCamargoOrtega/IBM_Capstone/blob/main/07.%20Hands-on%20Lab-%20Interactive%20Dashboard%20with%20Ploty%20Dash

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.
- Source code https://github.com/GermanCamargoOrtega/IBM_Capstone/blob/main/08.%20Hands-on%20Lab:%20Machine%20Learning%20Prediction%20lab.ipynb



Results

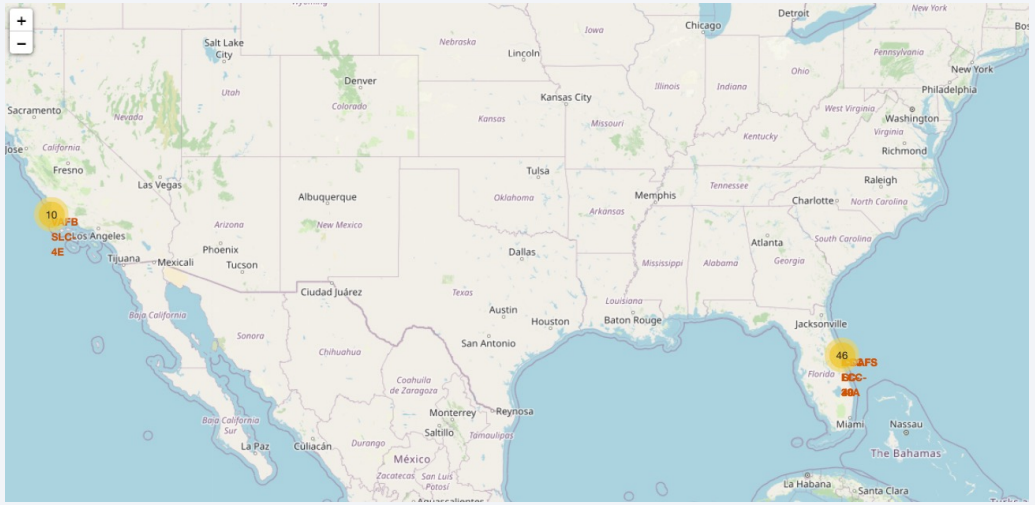
EDA results (summary)

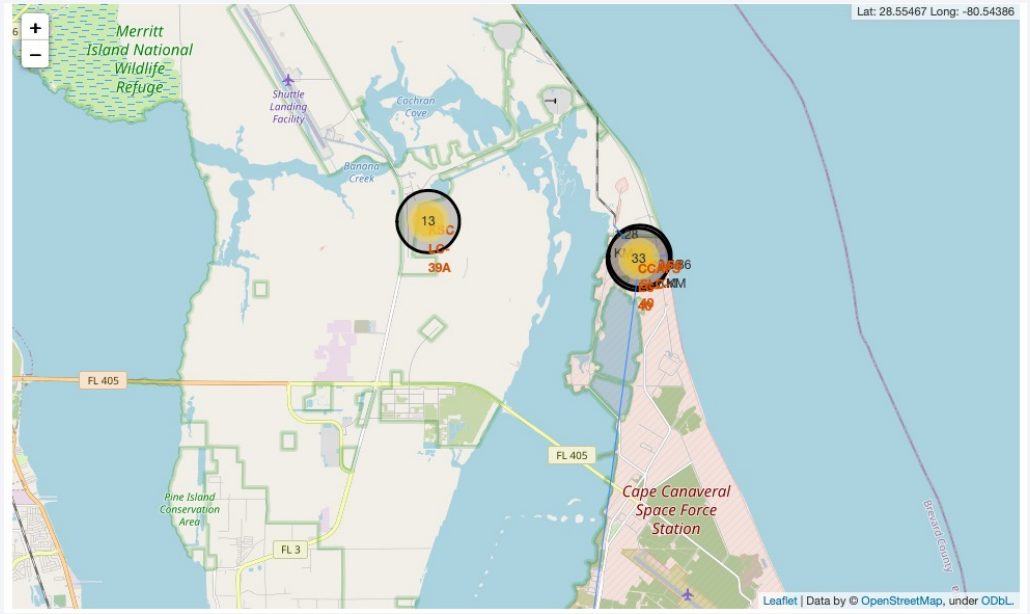
- SpaceX uses 4 different launch sites.
- First launches were done to SpaceX itself and NASA.
- Mean payload of F9 v1.1 booster = 2,928kg.
- First success landing outcome in 2015
  - fiver year after the first launch!!!
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average.
- Almost 100% of mission outcomes were successful.
- Two booster versions failed at landing in drone ships in 2015
  - F9v1.1B1012
  - F9v1.1B1015;
- Number of landing outcomes became better with increasing time.

Results

Interactive analytics demo in screenshots (summary)

- We observed that launch sites used to
  a. be in safety places
  b. be near sea
  c. have a good logistic infrastructure around.
- Most launches happens at east cost launch sites.

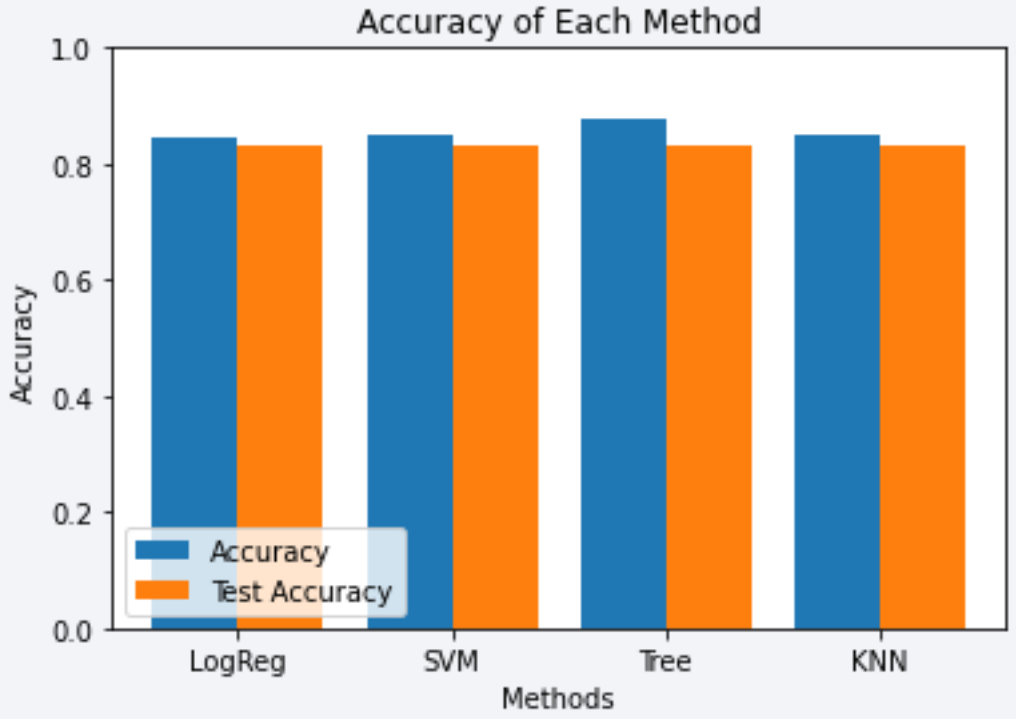


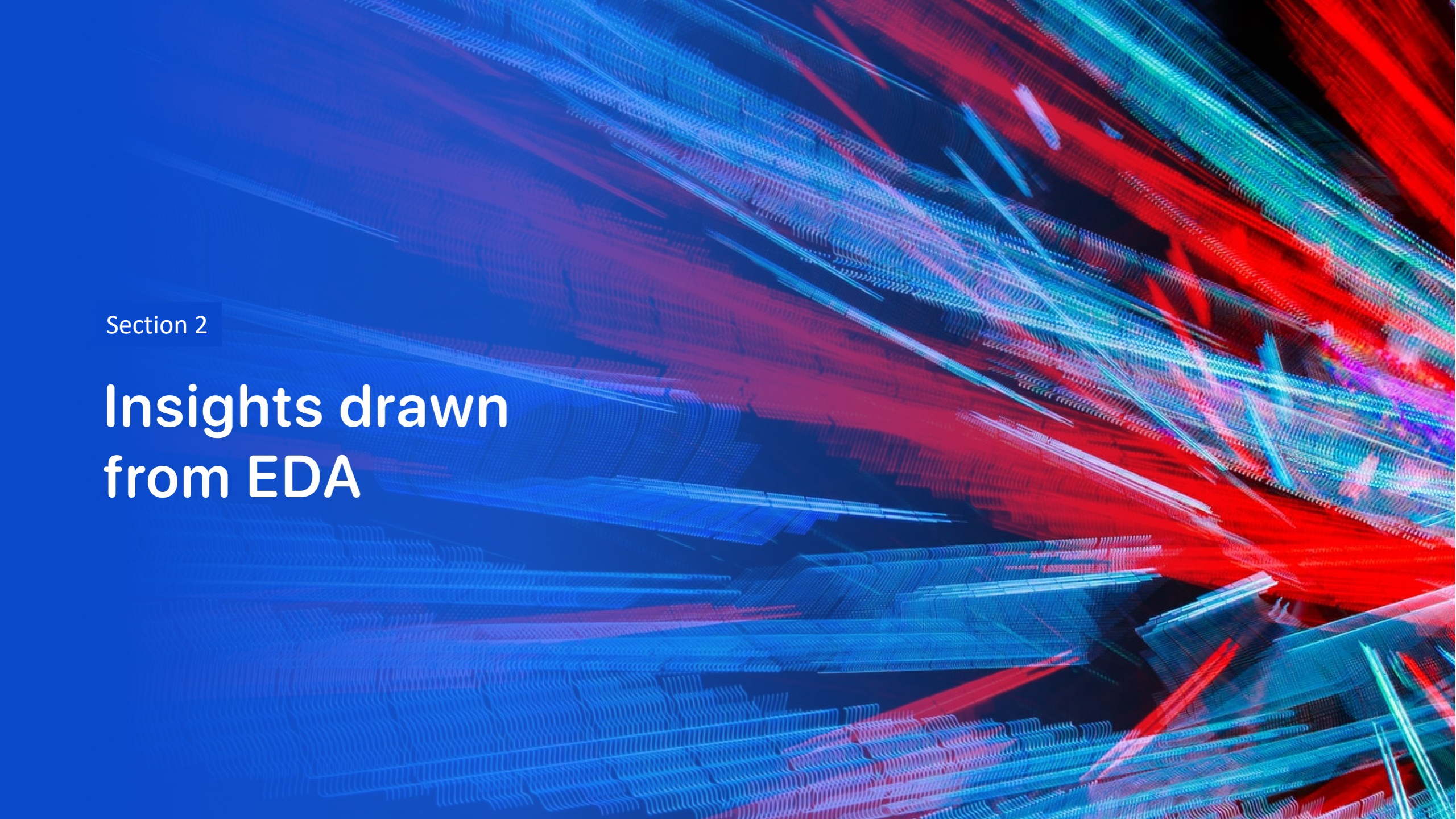


Results

Predictive analysis results (summary)

- All methods preformed basically equivalent.
- The tree model, however, fits train data slightly better but performs on test data worse.
  - accuracy for training over 87%
  - accuracy for test data equals 83%.

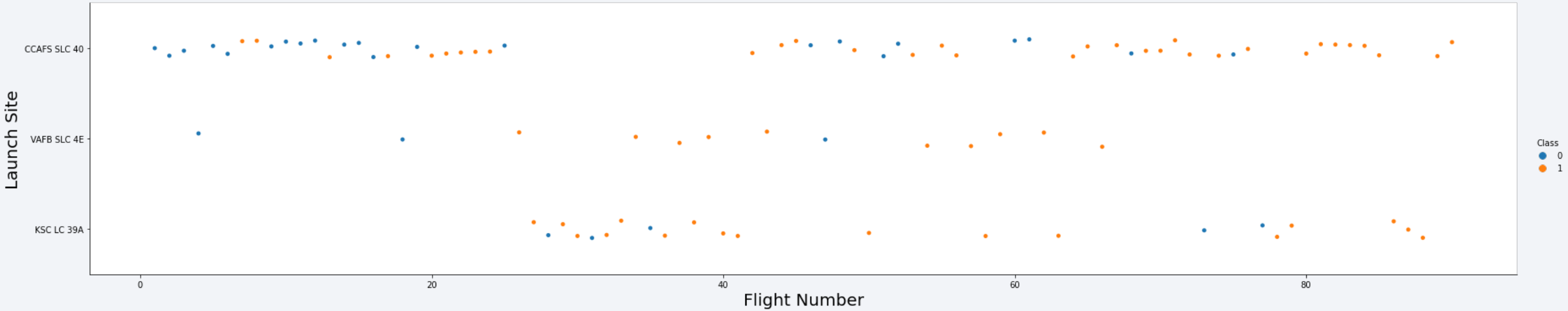



Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Observations

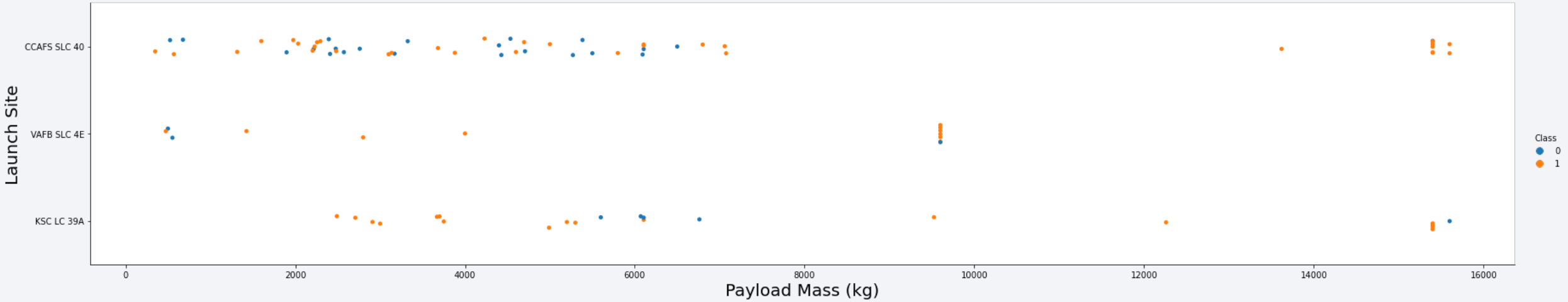
- Current success of launch sites are CCAF5 SLC 40 > VAFB SLC 4E > KSC LC39A
- General success rate improvement overtime is clearly visible.



Payload vs. Launch Site

Observations:

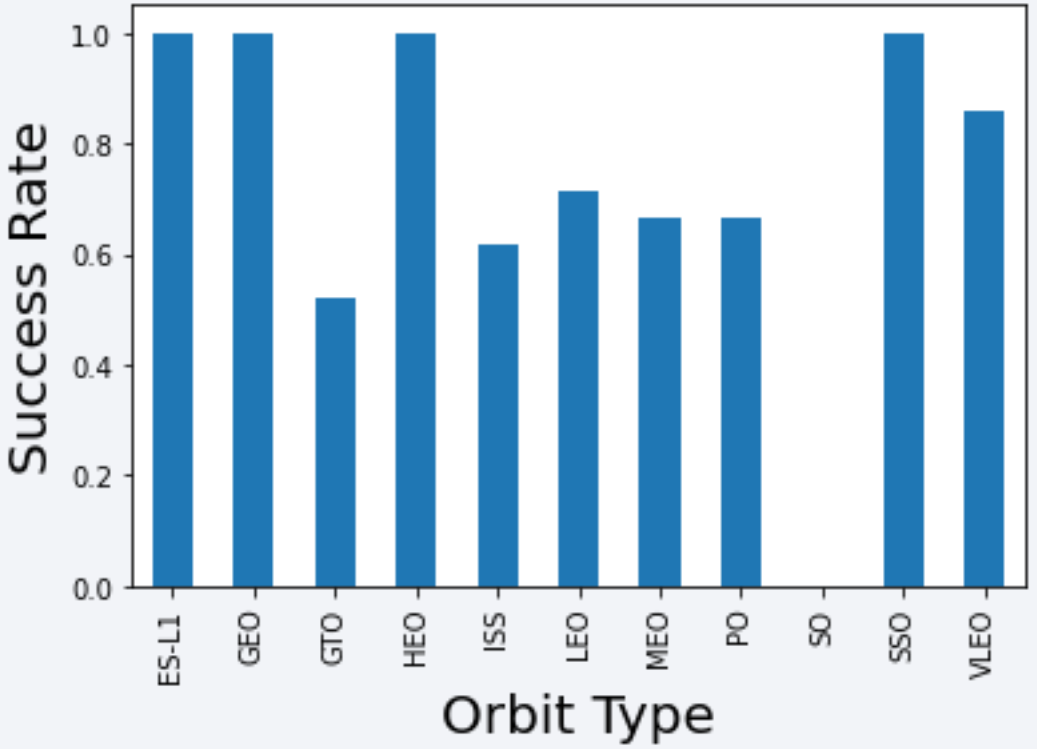
- Payloads over 9,000kg have the best success rate.
- Payloads over 12,000kg done only on CCAFS SLC 40 and KSCLC 39A.



Success Rate vs. Orbit Type

Observations:

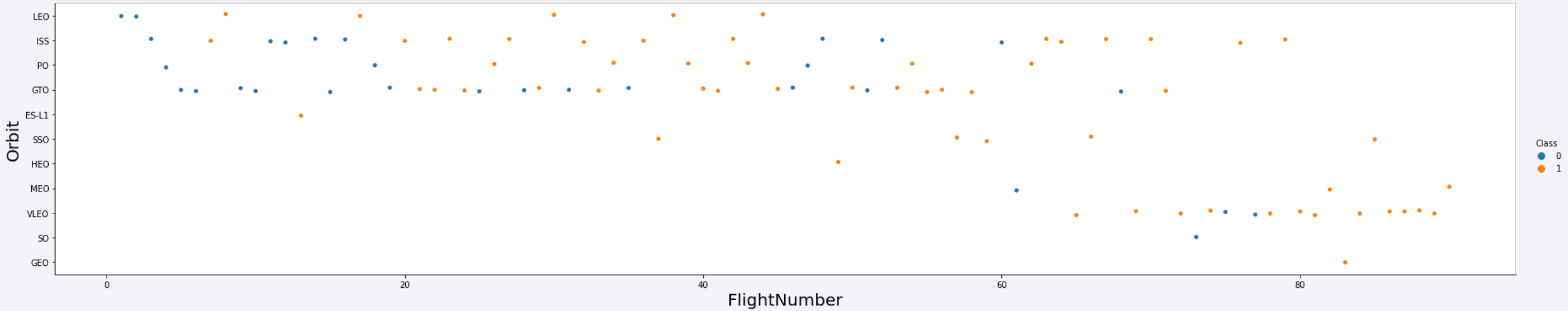
- The highest success rate is achieved in orbits ES-L1, GEO, HEO and SSO.
- The rest of orbits have success rates between 50% and 80%.



Flight Number vs. Orbit Type

Observations:

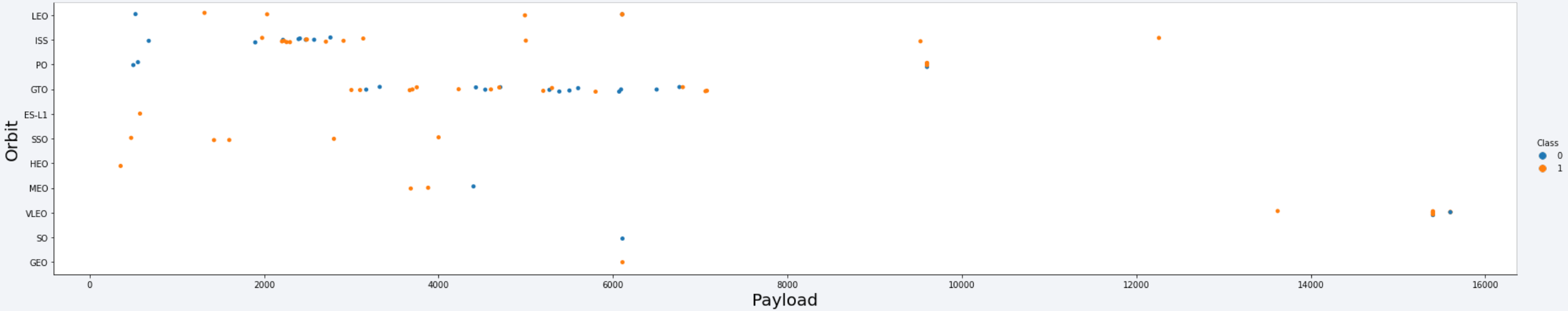
- For all orbits , the success rate improved with increasing flight number.
- Notably, the VLEO orbit recently increase the flight number, and these were quite successful!



Payload vs. Orbit Type

Observations:

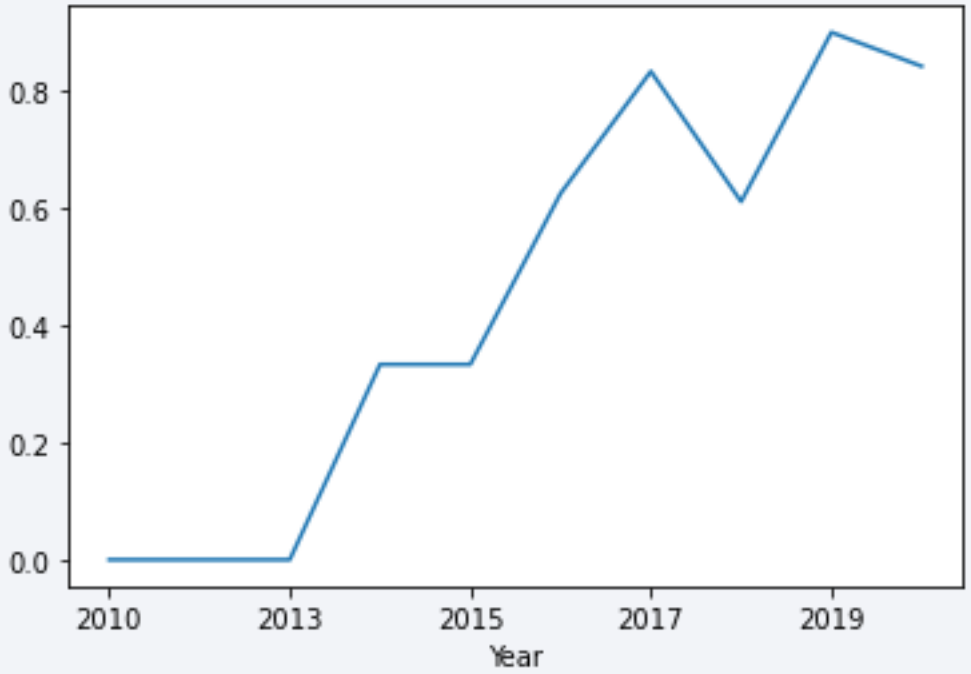
- With heavy payloads the successful landing or positive landing rate are more for Polar, VLEO and ISS.
- For GTO one cannot distinguish this well as both positive and negative landings.



Launch Success Yearly Trend

Observations:

- Success rate since 2013 kept increasing till 2020



All Launch Site Names

Observations:

- Launch Sites Names (see below) are catch by selecting unique occurrences of “launch_site” values from the data set.

```
%%sql SELECT DISTINCT LAUNCH_SITE
      FROM "XXJ77629"."SPACEXTBL";
```

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

Launch Site Names Begin with 'KSC'

- 5 records where launch sites begin with the string 'KSC'
- The source code is

```
%%sql SELECT LAUNCH_SITE
    FROM "XXJ77629"."SPACEXTBL"
    WHERE LAUNCH_SITE LIKE 'KSC%'
    LIMIT 5;
```

launch_site
KSC LC-39A
KSC LC-39A
KSC LC-39A
KSC LC-39A
KSC LC-39A

Total Payload Mass

Observations:

- The total payload mass carried by boosters launched by NASA (CRS) is calculated as following (summing all pay loads whose codes contain ‘CRS’)

```
%%sql SELECT SUM(PAYLOAD_MASS__KG_)
    FROM "XXJ77629"."SPACEXTBL"
    WHERE Customer = 'NASA (CRS)';
```

- The mass is = 45596 KG

Average Payload Mass by F9 v1.1

Observations:

- The average payload mass carried by booster version F9 v1.1 is calculated as following (filtering data by booster version and calculating the average):

```
%%sql SELECT AVG(PAYLOAD_MASS__KG_)

    FROM "XXJ77629"."SPACEXTBL"

    WHERE Booster_Version LIKE 'F9 v1.0%';
```

- average payload mass = 340 KG

First Successful Ground Landing Date

Observations

- The date where the first successful landing outcome in drone ship was achieved is obtained as following (filtering data by successful landing outcome on ground pad and getting the minimum value for date):

```
%%sql SELECT MIN(Date)
    FROM "XXJ77629"."SPACEXTBL"
    WHERE Landing__Outcome = 'Success (ground pad)';
```

- The date is = 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Observations:

- The names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000 are obtained as following:

```
%%sql SELECT BOOSTER_VERSION
    FROM "XXJ77629"."SPACEXTBL"
    WHERE LANDING__OUTCOME = 'Success (drone ship)'
    AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

booster_version
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1029.2
F9 FT B1038.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

Observations:

- The total number of successful and failure mission outcomes (see below) can be checked like this (grouping mission outcomes and counting records for each group):

```
%%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
    FROM "XXJ77629"."SPACEXTBL"
   GROUP BY MISSION_OUTCOME;
```

MissionOutcome	Occurrences
Failure (inflight)	1
Success	99
Success (payload statusunclear)	1

Boosters Carried Maximum Payload

Observations:

- The names of the booster versions which have carried the maximum payload mass can be deducted like this:

```
%%sql SELECT DISTINCT BOOSTER_VERSION
    FROM "XXJ77629"."SPACEXTBL"
    WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM "XXJ77629"."SPACEXTBL");
```

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

2017 Launch Records

Observations:

- The records which will display the month names, successful landing outcomes in ground pad, booster versions, launch site for the months in year 2017 should be obtainable with following code:

```
%%sql SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
    FROM "XXJ77629"."SPACEXTBL"
    WHERE Landing__Outcome = 'Failure (drone ship)'
    AND YEAR(DATE) = 2017;
```

- No records were found for this year though.

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

Observations:

- To rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order one could use this code:

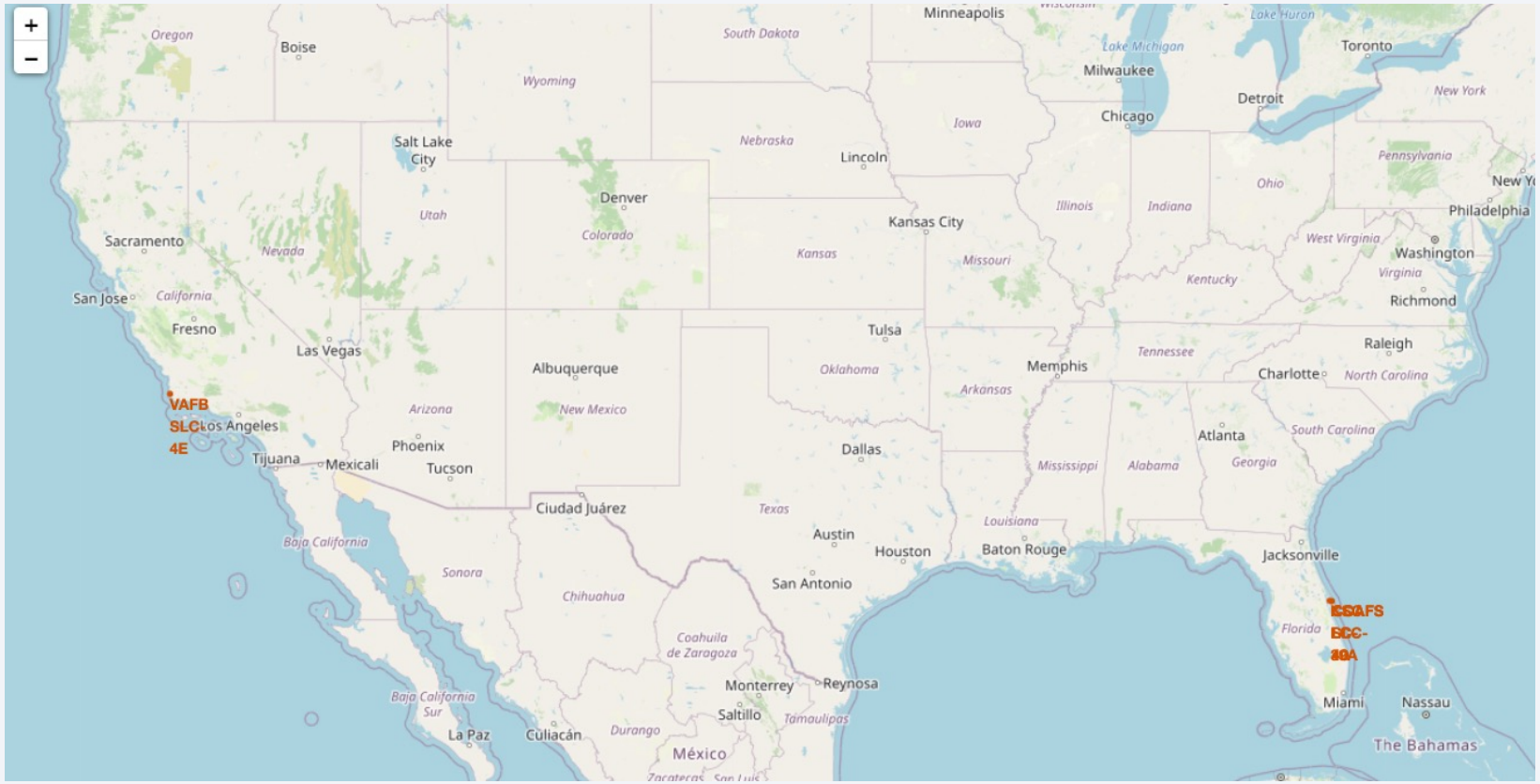
```
%%sql SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER
    FROM "XXJ77629"."SPACEXTBL"
    WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
    GROUP BY LANDING__OUTCOME
    ORDER BY TOTAL_NUMBER DESC
```

LandingOutcome	Occurrences
Noattempt	10
Failure (droneship)	5
Success (droneship)	5
Controlled(ocean)	3
Success (groundpad)	3
Failure(parachute)	2
Uncontrolled(ocean)	2
Precluded (drone ship)1	1

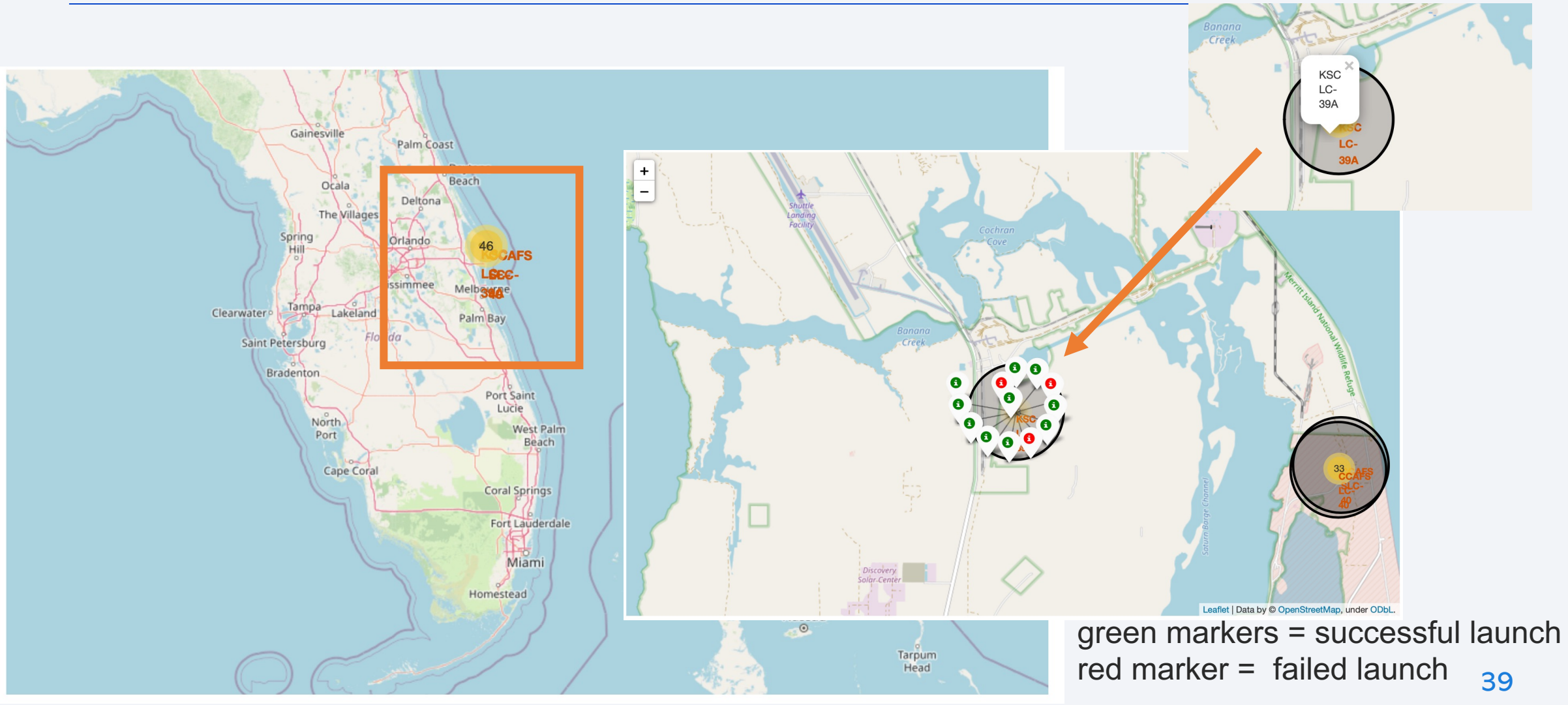

Section 3
Launch Sites
Proximities Analysis

All launch sites

Observations: on both coasts, launch site are in the southern parts of the country and near the sea; near roads and rail roads.

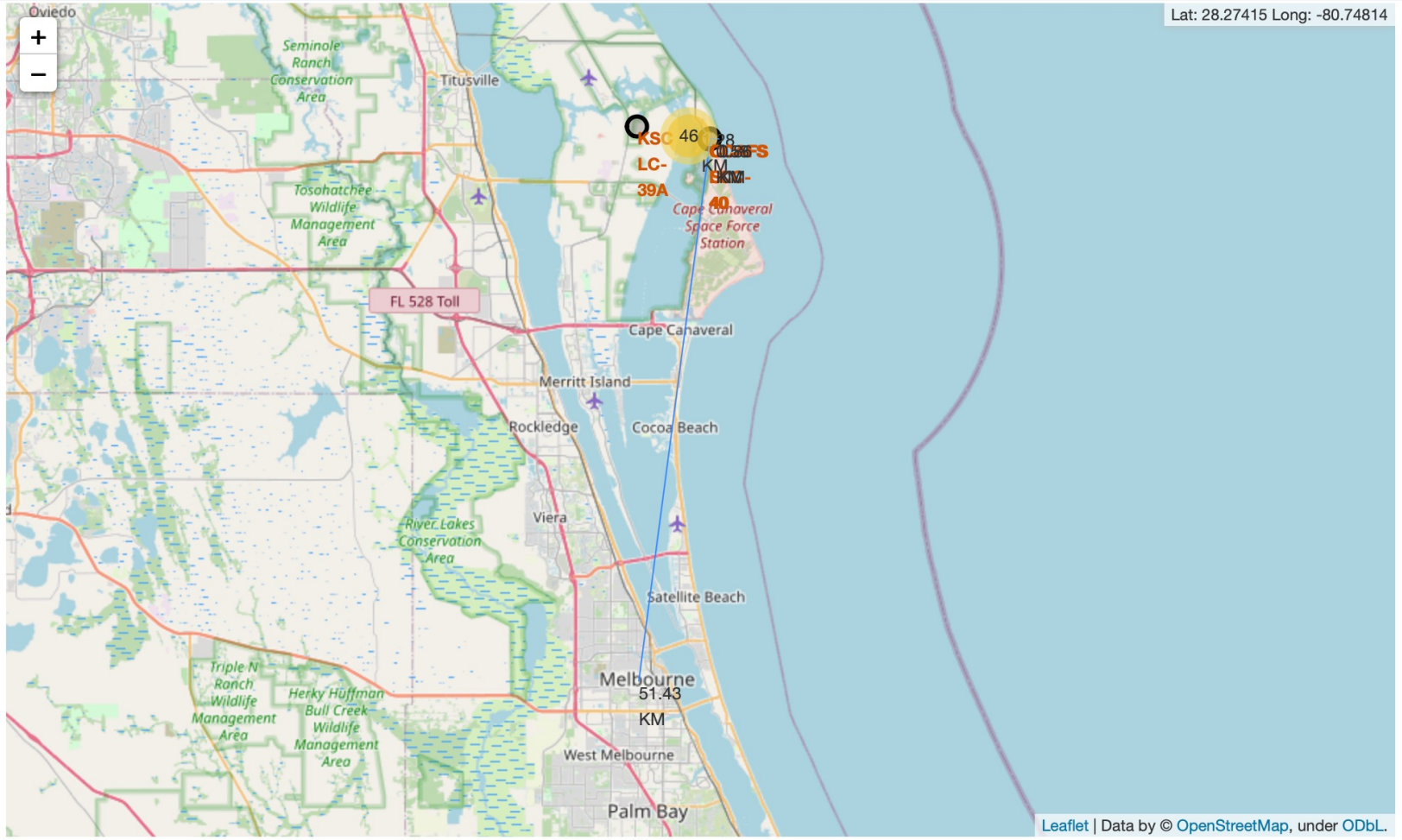


Launch Outcomes by site - KSC LC-39A



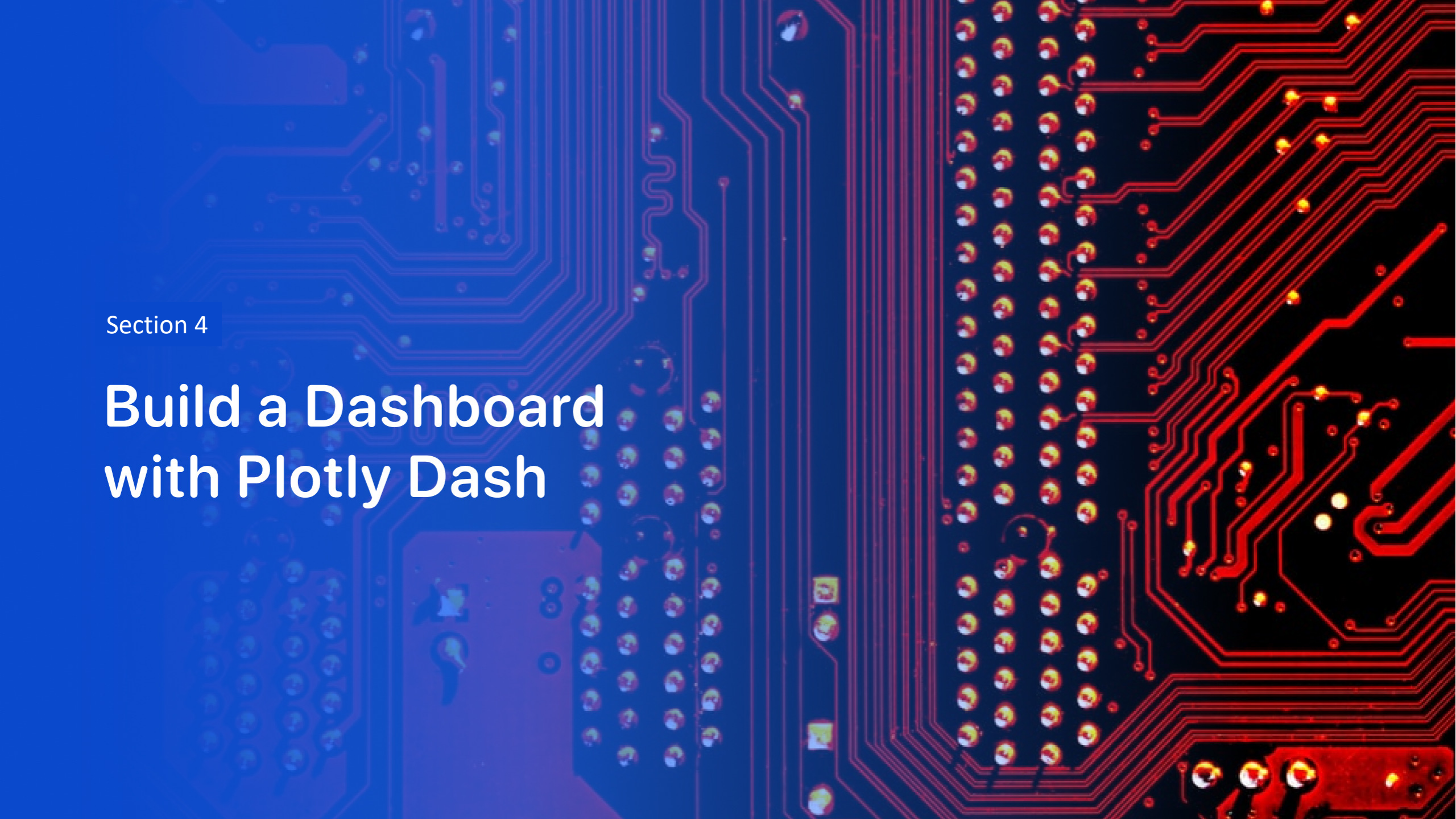
green markers = successful launch
red marker = failed launch

Logistics and safety



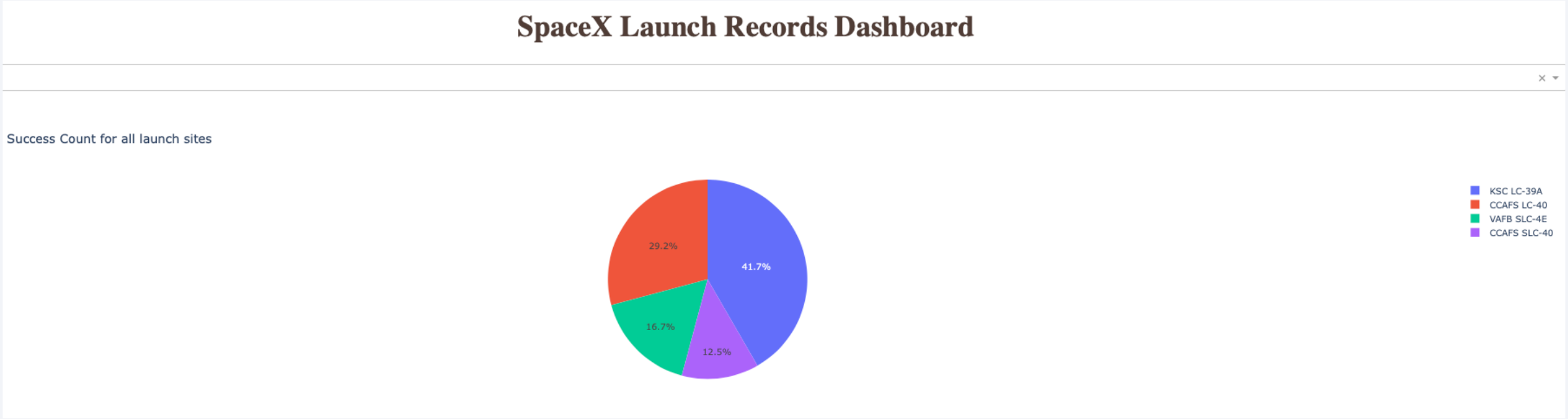
Here we can see that launch sites in the east coast are indeed have indeed optimal logistics aspect

- near railroads and roads
- far from inhabited areas (51 km from Melbourne)


Section 4
Build a Dashboard with Plotly Dash

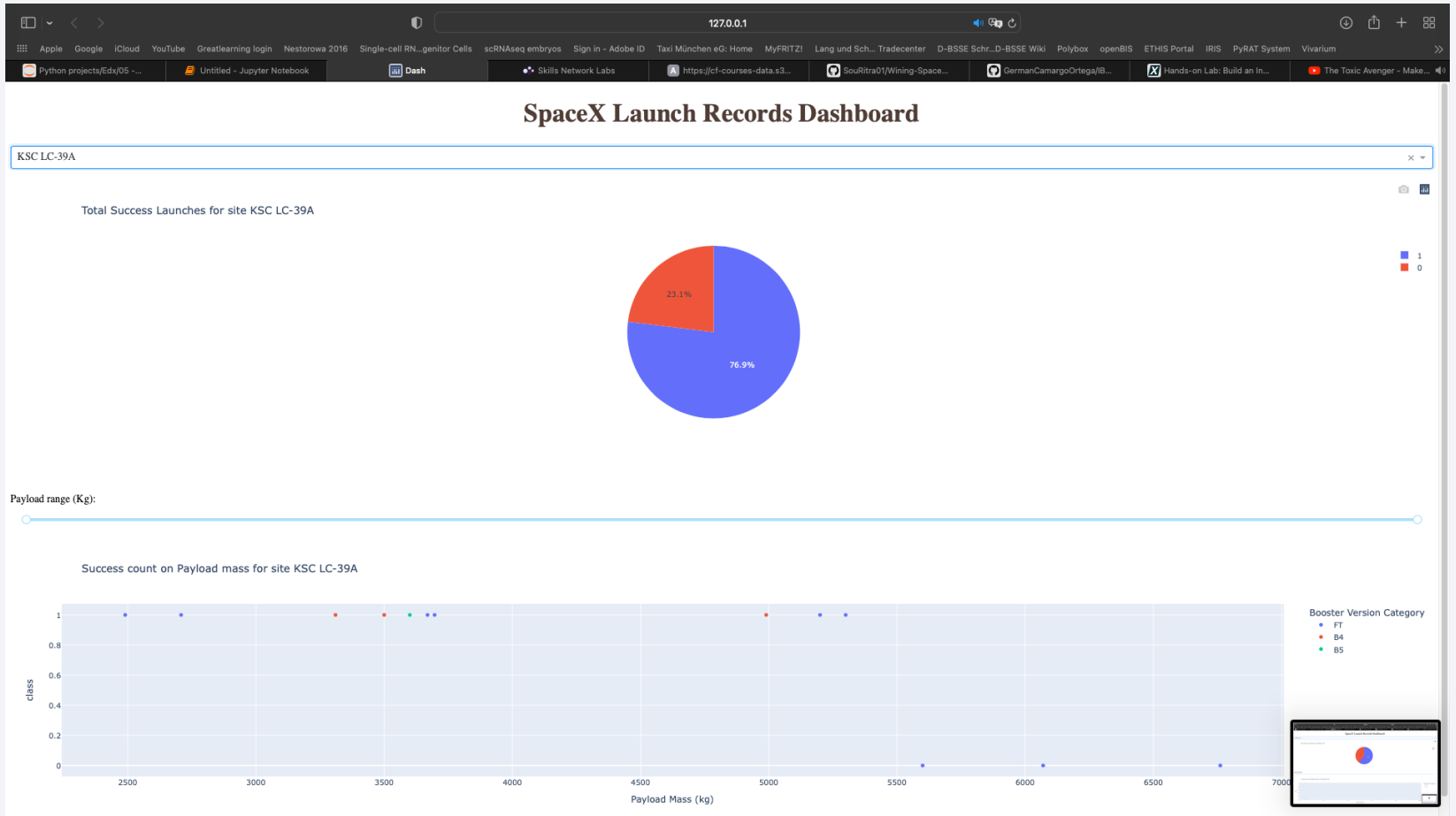
Launch success count for all sites

- There is an obvious link between the launch place to success of missions.
- KSC LC-39A and CCAPS LC-40 have the highest success counts, accounting for 70% of the total success counts.



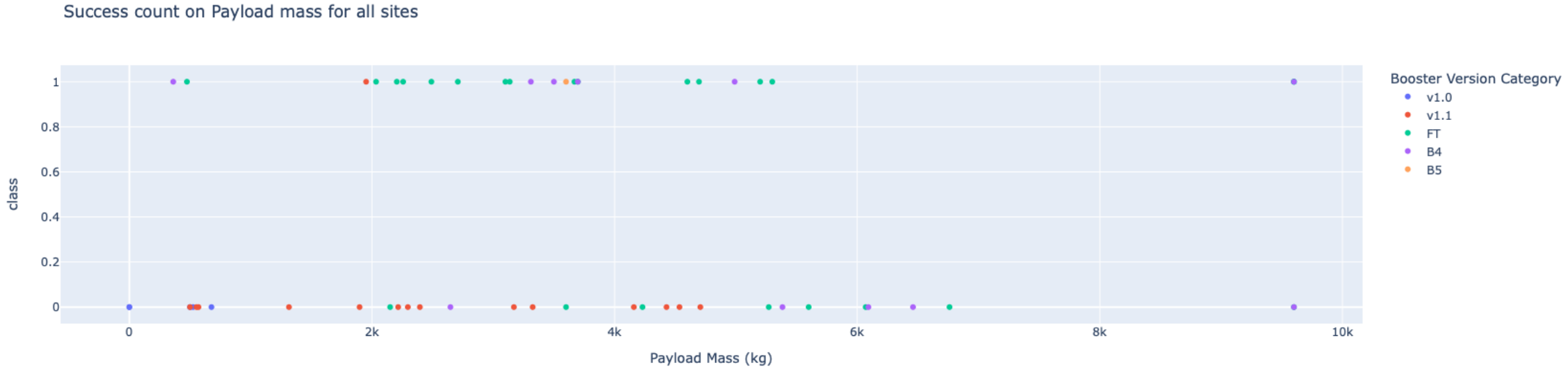
KSCLC-39A: the launch site with highest launch success ratio

- Almost 3 our 4 launches (76.9%) in this site were successful.



Payload vs. Launch Outcome

- Payloads under 6,000kg and FT boosters are the most successful combination.

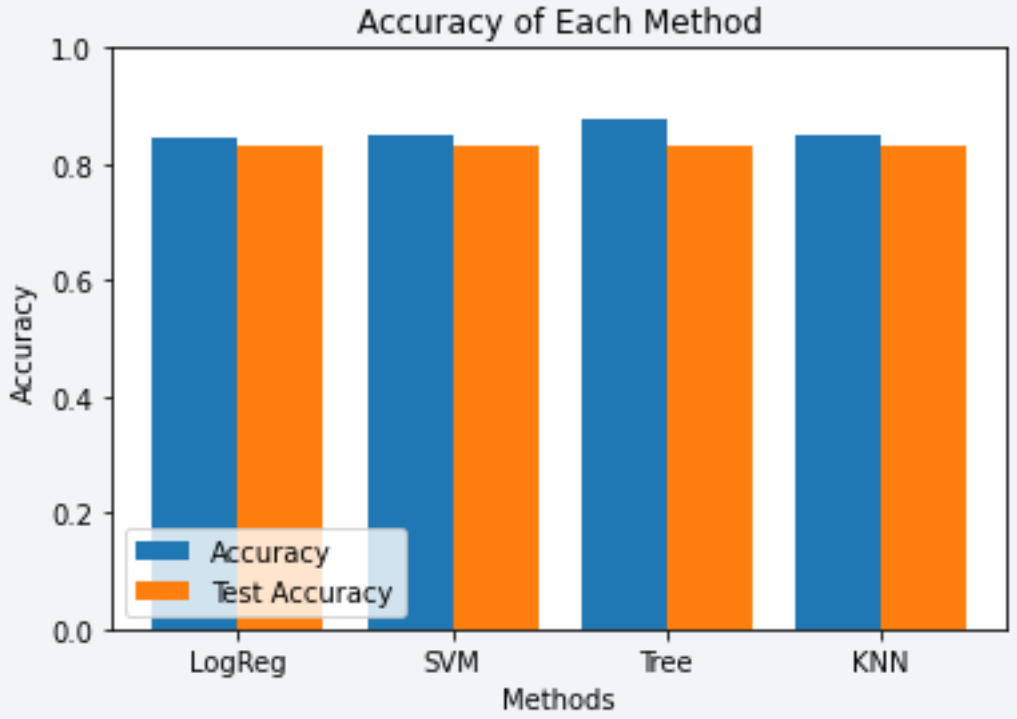


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- I I tested four classification modes:
  - logistic regression
  - SVM
  - Decision Tree Classifier
  - Kean's Nearest Neighbour.
- All methods preformed basically equivalent.
- The tree model, however, fits train data slightly better but performs on test data worse.
  - accuracy for training over 87%
  - accuracy for test data equals 83%.



Confusion Matrix

- The confusion matrix of the Decision Tree Classifier describes the performance of the model.
- It compares the number of true positive and true negative values vs. the predicted true and false values.

Conclusions

1. I analyzed two different data sources. These can help refining the analysis and hence the conclusions.
2. Launches are done on both coasts, in southern parts of the country, near the sea, roads and rail roads.
3. The best/most successful launch site seems to be *KSCLC-39A*.
4. Launches above 7,000kg are less risky (have higher success rates/chances).
5. The success of the missions increase over the years. The optimization of processes and rockets could have certainly played a role in increasing the success rates.
6. Orbits ES-L1, GEO, HEO and SSO have the best success rates.
7. Decision Tree Classifier can be used to predict successful landings. However, any other of the tested models can give a comparable result.

Appendix

- The tree model could be improved by optimizing hyperparameters, such as the cross-validation settings and algorithm used, the number of trees and levels (pruning) or by increasing the size of the data.
- Alternatively, one could use neural networks to check if they could perform better.

Thank you!