IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Paulo R. M. Yugar
04/27/2023

Outline

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

Executive Summary

- Summary of methodologies:

  Data from Rocket launches were collect by using SpacexApi and Web Scrapping.

  Exploratory Data Analysis(EDA), including Data Visualization, Data wrangling doing cleaning and addressing missing values and interactive visual analytics were made to view starting trends on the data.

  Folium was used to view Geographic insights of the rocket launch locations.

  Dashboards were made to better access relationships between launch location outcomes and relations with payload mass and booster version.

  Lastly, Machine learning algorithms were applied to predict the chances of successful landing of rockets.

- Summary of all results:

  Launch site location, payload mass, type of orbit, and booster version did influence outcome of the landings.

  Machine learning results calculated the rate of success of future landings of Falcon 9 rockets, thus showing how reliable is the company is achieving their results.

Introduction

- Given the advancements of the space industry by SpaceX, how an hypothetical company, SpaceY, would fare against the former, given their highly competitive rockets. Thus the objective of this report is to measure the viability of this new company.
- What is the cost of the each launch?
- If the success of the landing of reusable first stage of a rocket can cut cost of each launch, then which factors and what is the chance of successful landing of the falcon 9 rockets?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  SpaceX REST API was used to collect data, also Web Scraping from Wikipedia About falcon 9 Rockets.

- Perform data wrangling

Data was cleaned and organized, columns of no interest were dropped, like Falcon 1 Rockets data, and missing values were replaced by the means of the columns when needed. Then a new dataframe was created with the Wrangled Data.

Methodology

Executive Summary

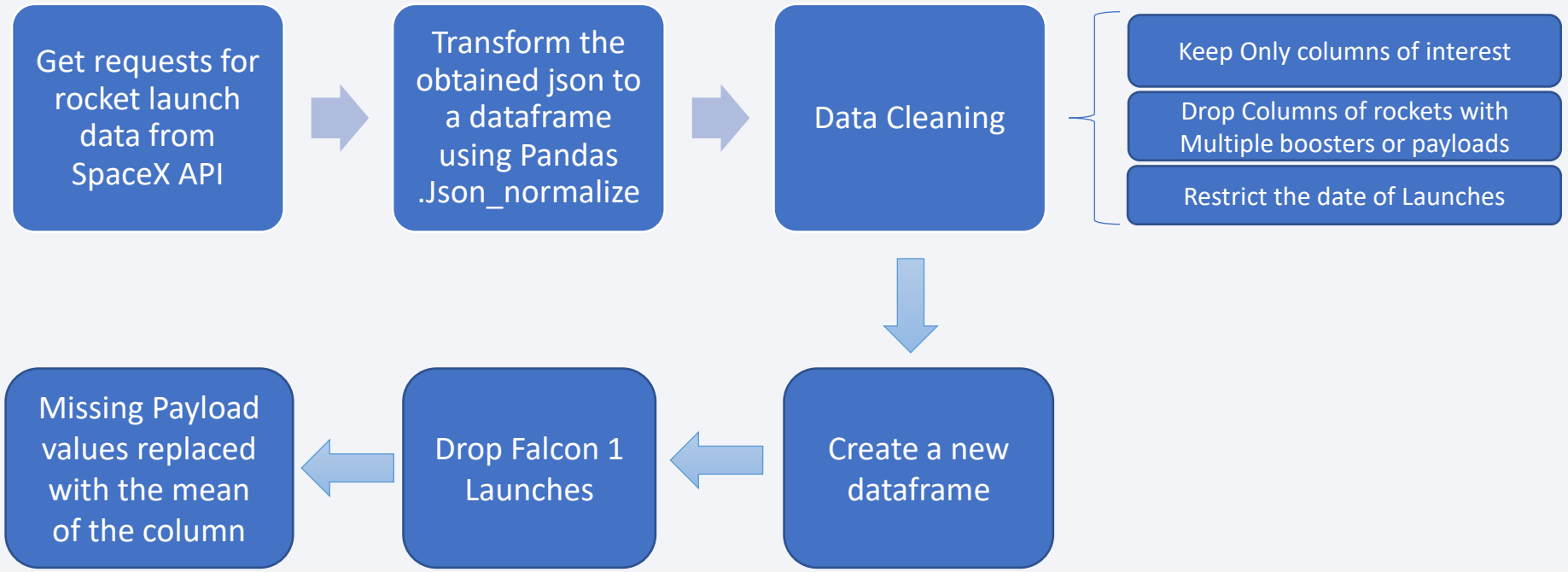
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

  Classification Models were used, requiring a supervised learning approach. Logistic Regression, Support Vector Machine(SVM), Decision Trees, and K-Nearest Neighbors were the algorithms chosen to build a model, tuning parameters for optimal results, and evaluating them with metrics such Precision, Recall and Accuracy. Confusion Matrix were build for each algorithm to better view how their trained prediction fared against test data values.

Data Collection

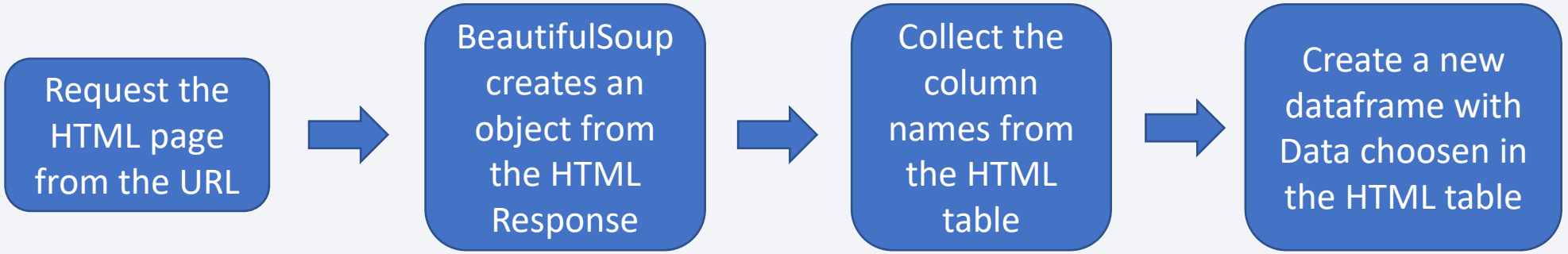
- Data related to SpaceX launches was collected via the SpaceX REST API
  - By requesting the data from the SpaceX API.
- Data regarding Falcon 9 launch records was gathered through web scraping
  - By retrieving a HTML table from Wikipedia.

Data Collection – SpaceX API



https://github.com/PauloYugar/DS_IBM_Capstone/blob/main/Collecting_data.ipynb

Data Collection - Scraping



https://github.com/PauloYugar/DS_IBM_Capstone/blob/main/WebScrapping_lab.ipynb

Data Wrangling

- Since not all Machine learning or statistical methods deal with Categorical data, It was needed to transform it in a binary one.

Outcome	
None none	Class= 0 –> Failure to Land
False ocean	
False ASDS	
False RTLS	
None ASDS	
True ocean	Class = 1 –> Success
True ASDS	
True RTLS	

https://github.com/PauloYugar/DS_IBM_Capstone/blob/main/DataWrangling_lab.ipynb

Data Wrangling

- The first data exploration also yielded other insights such the proportion of rockets in each launch site, and the preferred orbits for the launches.

Launch site	Launches
CCAFS SLC 40	55
KSC LC 39A	22
VAFB SLC 4E	06

Orbit	Launches
GTO	27
ISS	21
VLEO	14
PO	9
LEO	7
SSO	5
others	7

https://github.com/PauloYugar/DS_IBM_Capstone/blob/main/DataWrangling_lab.ipynb

EDA with Data Visualization

- Seaborn library was used to plot different graphs such bar charts, Scatter plots and line plots to check patterns between different variables, and how they are related with the landing outcome.
- Some examples of graphs that were plotted:
  - Payload Mass x Flight number - Scatter plot.
  - Launch Site x flight number - Scatter plot.
  - Orbit x Landing success - Bar chart.
  - Landing Success x Date - Line chart.
  - More Detailed analysis of each graph will be done late in this presentation.

https://github.com/PauloYugar/DS_IBM_Capstone/blob/main/DataExploring-Lab.ipynb

EDA with SQL

- SQL queries where performed in jupyter notebook with Python SQL integration
- The performed Queries were as follows:
  - Display the names of the unique launch sites in the space mission.
  - Display 5 records where launch sites begin with the string 'CCA'.
  - Display the total payload mass carried by boosters launched by NASA (CRS).
  - Display average payload mass carried by booster version F9 v1.1.
  - List the date when the first successful landing outcome in ground pad was achieved.¶.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - List the total number of successful and failure mission outcomes.
  - List the names of the booster versions which have carried the maximum payload mass.
  - 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.
  - Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

https://github.com/PauloYugar/DS_IBM_Capstone/blob/main/SQL_lab.ipynb

Build an Interactive Map with Folium

- Folium python Library was used for exploratory geographic analysis of the Launch site locations:
  - Circles were used to mark the location of different Launch sites.
  - Markers clusters were used in that locations, to show how many rockets were launched in each site, with different colors to indicate success or failure of each of the launches.
  - Lines were used to show the distance between different points of interest around each site.

https://github.com/PauloYugar/DS_IBM_Capstone/blob/main/Folium_lab.ipynb

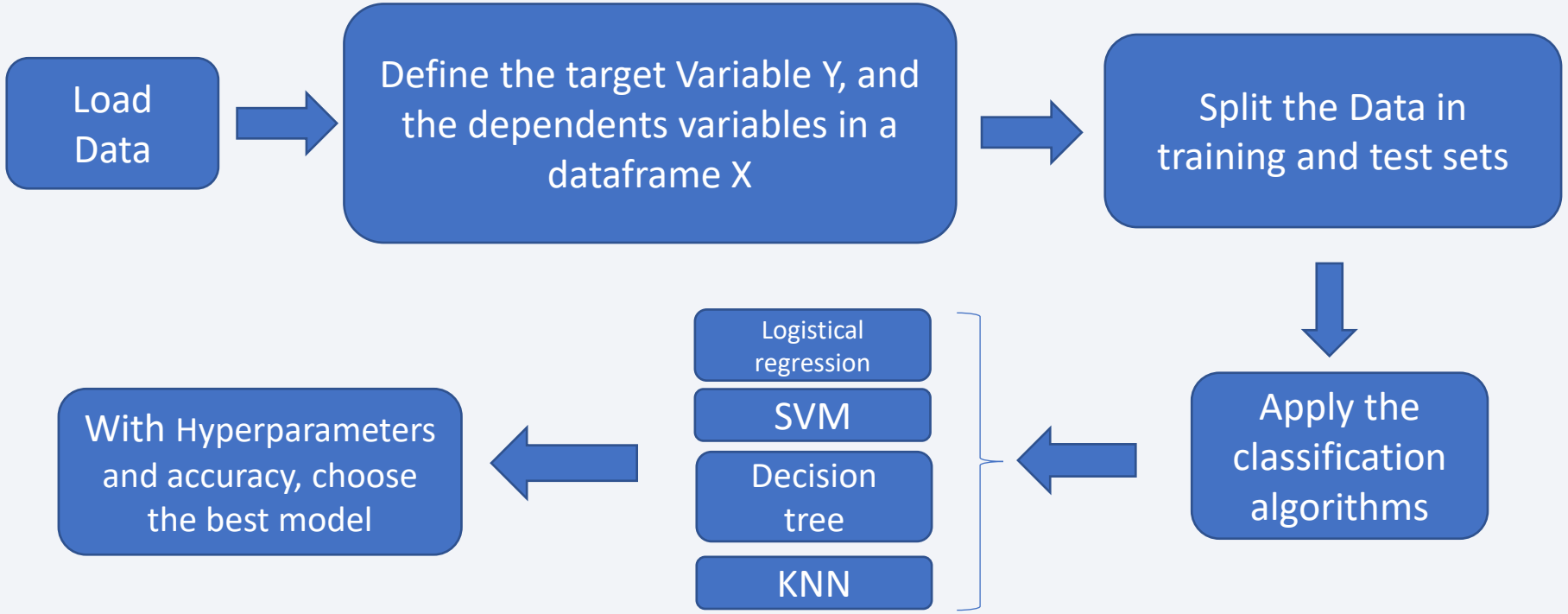
Build a Dashboard with Plotly Dash

- A dashboard was made to quickly show different aspects of the data set, and to facilitate access of the data for people without expertise.
- Certain aspects where added to facilitate user access:
  - A drop down menu to select different launch sites.
  - A callback function to change the pie chart according to changes of the menu.
  - A range slider for payload mass, connected to a Scatterplot, and the respective callback function to update the slider changes made.
- The combination of the pie chart and Scatterplot helps to quickly analyze the best combination for successful launches considering location and booster type used respectively.

https://github.com/PauloYugar/DS_IBM_Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

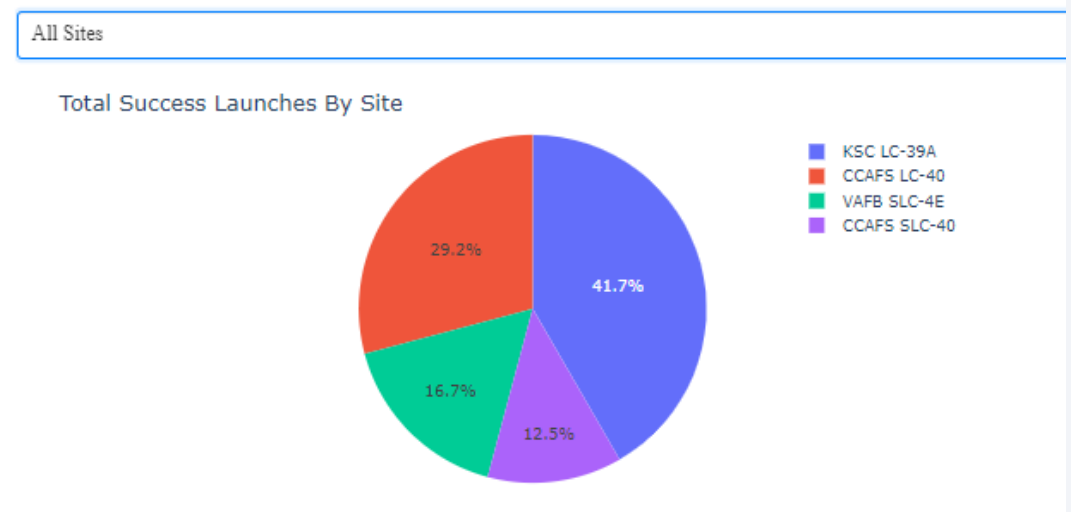
- After EDA was made, different algorithms where applied to the data to test which one could come closer to predict the success of the landings.

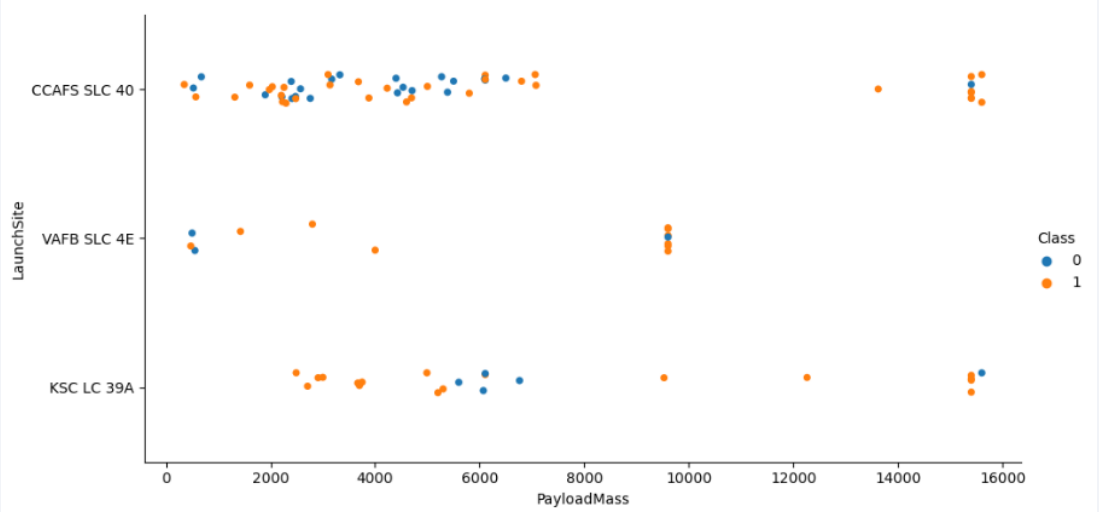


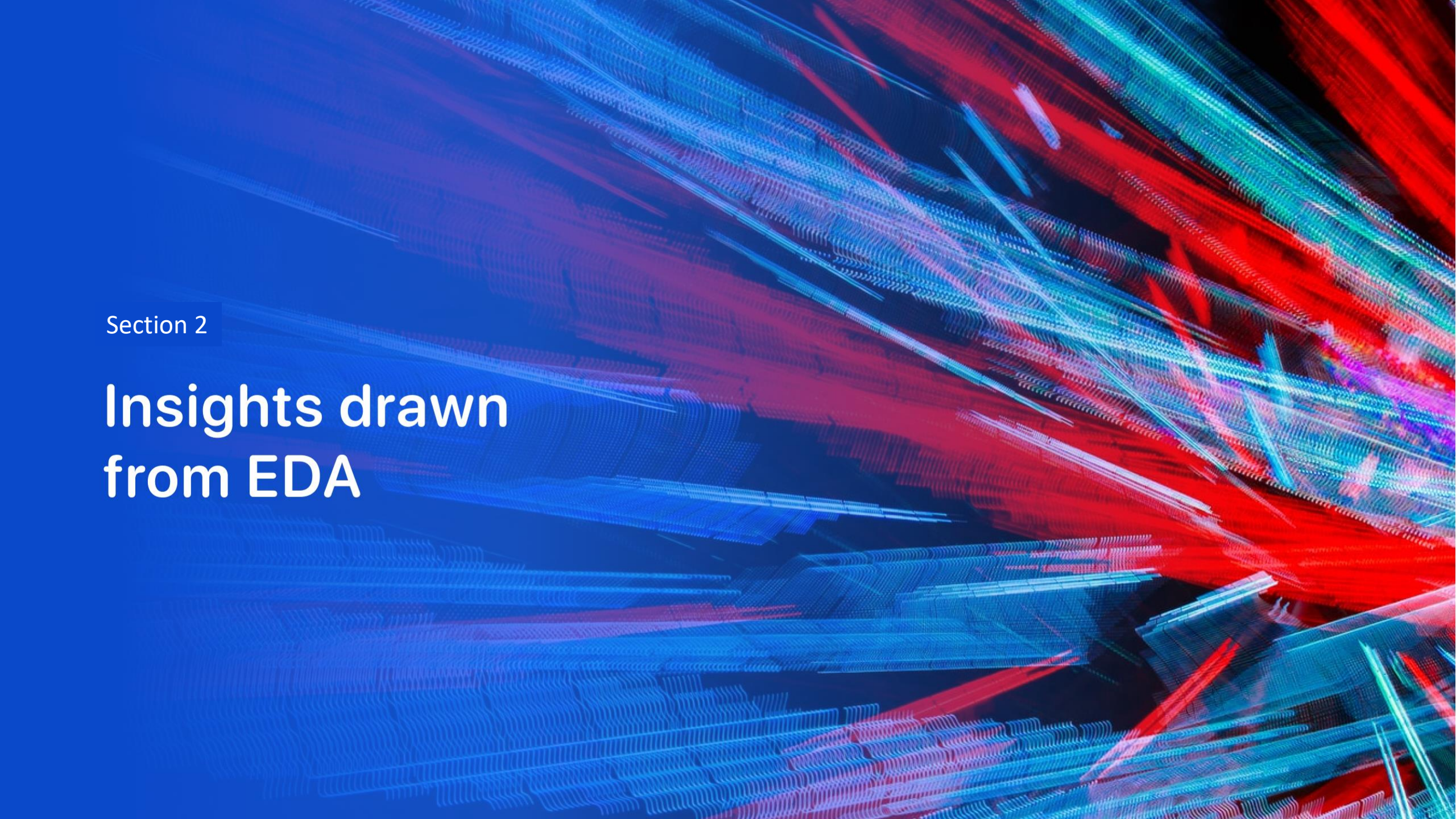
https://github.com/PauloYugar/DS_IBM_Capstone/blob/main/Machine_learning_lab.jupyterlite.ipynb

Results

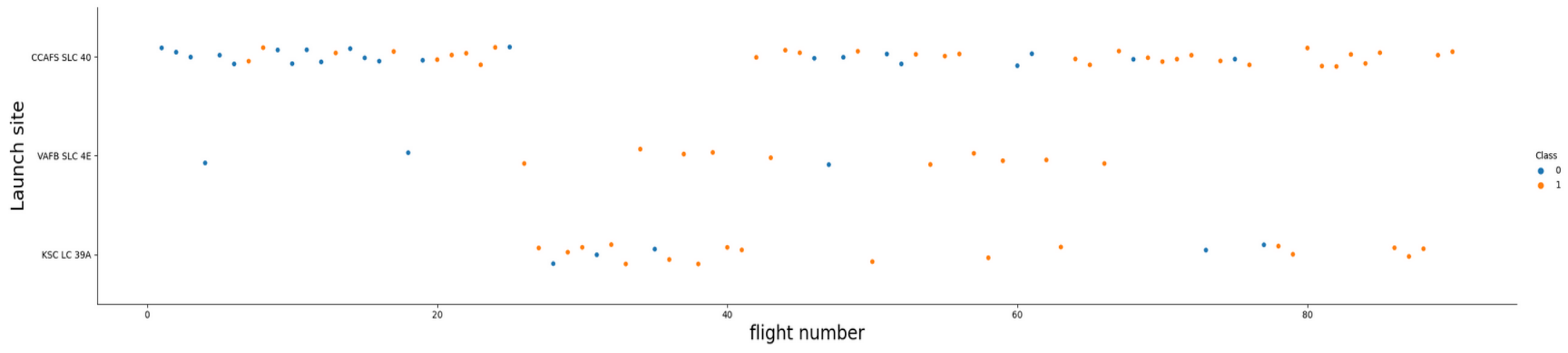
- Some of the findings were:
  - Kennedy space center is the most successful launch site for Landings.
  - Certain Launch sites have a limit of how heavy their payloads are, while other are more successful in certain weight ranges.
  - All algorithms tested had very similar accuracy against the testing sample.





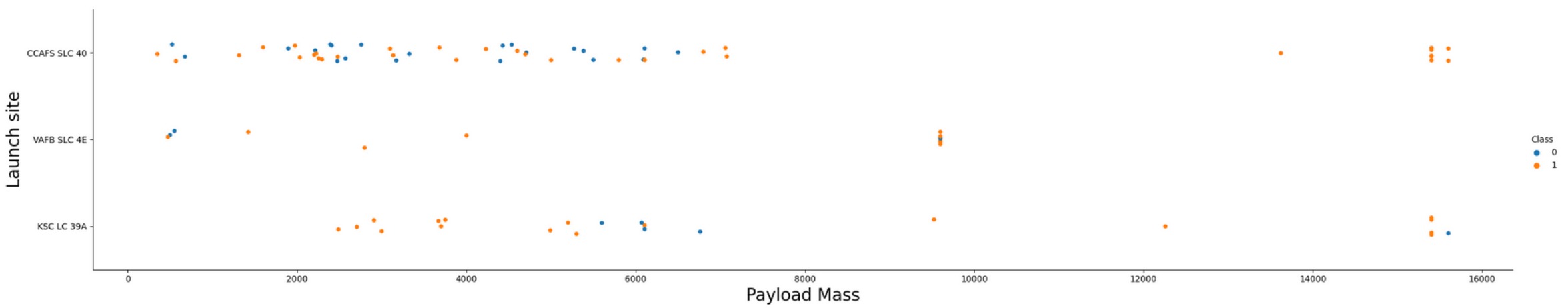

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



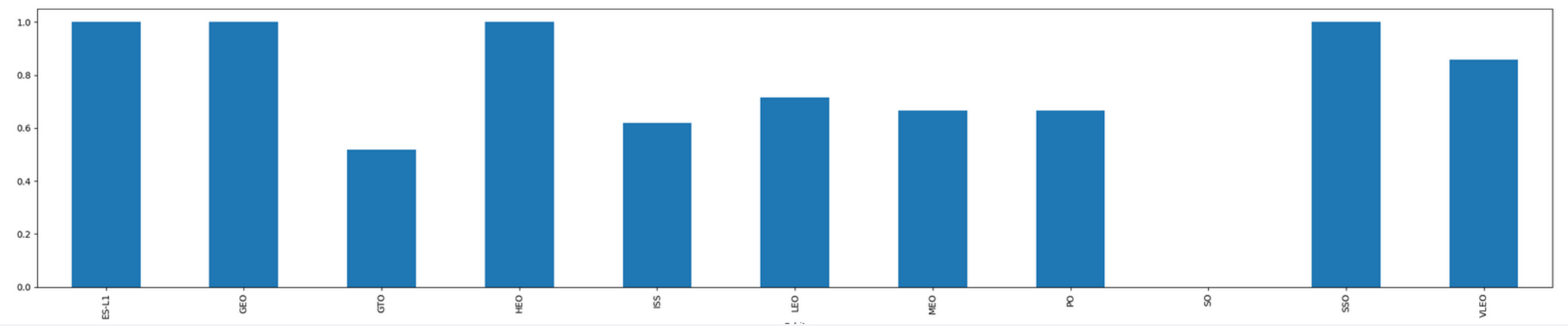
- CCAFS SLC 40 had great failure rate at start of their launches, but that improved over time.
- Generally it seems launch sites successful landings improve over time, but more data is needed in the other locations to be more comparable to CCAFS SLC 40.

Payload vs. Launch Site



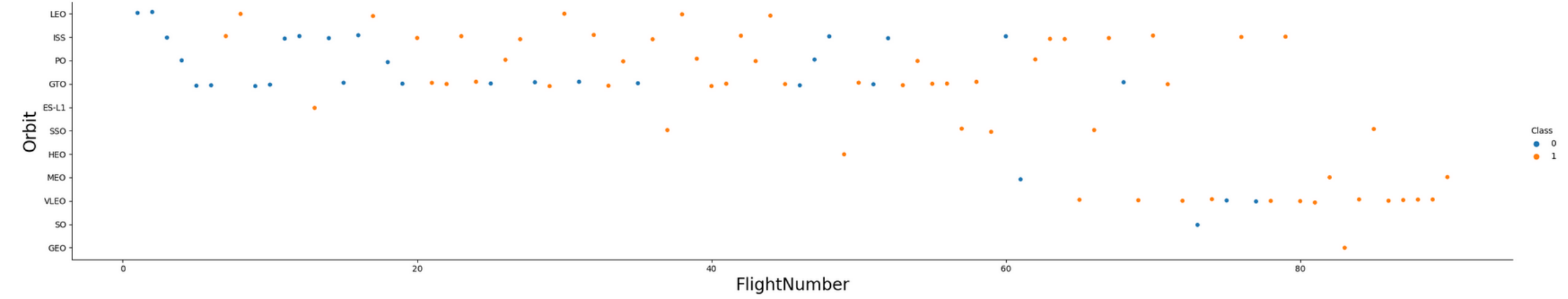
- CCAFS SLC 40 has great success in heavier payload launches(over 13000 kg).
- KSLC LC 39A has great success in lighter payload launches(under 5000kg).
- Overall, Payload over 7000kg have successful launches.
- It seems VAFB SLC 4E only allow payloads with less than 10000kg.

Success Rate vs. Orbit Type



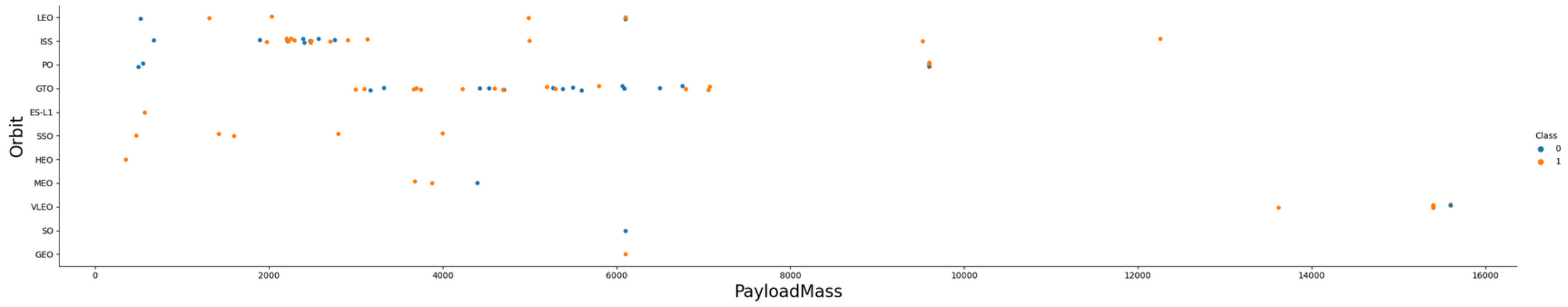
- Some orbits have 100% success in their Landings.
- Orbits with total success were ES-L1, GEO, HEO, and SSO.

Flight Number vs. Orbit Type



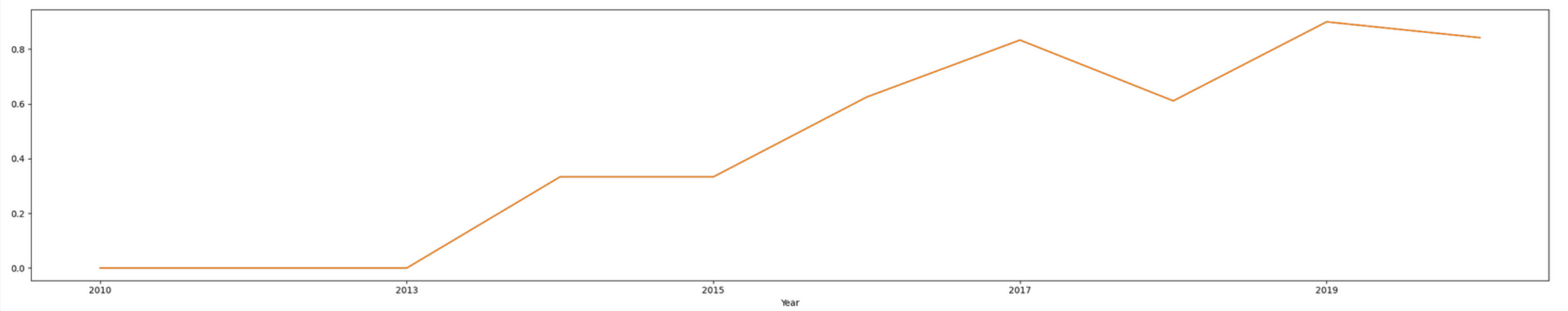
- Different orbits sets were tried over time by company.
- The orbits with great success in landing seems to have low number of flight attempts.

Payload vs. Orbit Type



- Some orbits favor certain payload mass range to have succesful landings.
- SSO seems to have great success with payloads under 4000kg.
- PO, ISS and Leo seems to have better success with payload over 4000kg.

Launch Success Yearly Trend



- The success of landings keep increasing since 2013.
- The 2010 to 2013 period could have been a time of refinement and adjustment of the technology that allowed the use of re-usable rocket parts.

All Launch Site Names

- CCAFS are launch sites in Cape Canaveral; KSC is the Kennedy Space Center, in the same facility; VAFB is the Vandenberg space force base.

```
[8]: %sql SELECT DISTINCT lAUNCH_SITE FROM SPACEXTBL ORDER BY 1
 * sqlite:///my_data1.db
Done.
```

[8]:

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

Launch Site Names Begin with 'CCA'

- The first 5 Cape Canaveral samples found in the dataset.

```
[9]: %sql SELECT * FROM SPACEXTBL WHERE lAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- This query finds the total sum of payloads mass with the code “CRS” that corresponds to NASA.

```
[10]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD LIKE '%CRS%';
 * sqlite:///my_data1.db
Done.
[10]:
```

TOTAL_PAYLOAD
111268

Average Payload Mass by F9 v1.1

- This query calculates the average payload mass, of the rockets with the booster version of v1.1.

```
[11]: %sql SELECT AVG(PAYLOAD_MASS__KG_ ) AS AVG_PAYLOADF911 FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
[11]: AVG_PAYLOADF911
2928.4
```

First Successful Ground Landing Date

- This query finds the date of the first successful landing of the type ground pad.

```
[12]: %sql SELECT MIN(DATE) FROM SPACEXTBL WHERE "LANDING _OUTCOME" = 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
[12]: MIN(DATE)
      01-05-2017
```

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
[13]: %sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND "LANDING _OUTCOME" = 'Success (drone ship)';
 * sqlite:///my_data1.db
Done.
```

[13]:

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

Total Number of Successful and Failure Mission Outcomes

- This query shows the total number of successful and failure mission outcomes.
- The total number of Success is 99, while Failure is 1.

```
[14]: %sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
 * sqlite:///my_data1.db
Done.
```

[14]:

Mission_Outcome	QTY
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- This query list the versions of the boosters which have carried the maximum payload mass.

```
[58]: %sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
 * sqlite:///my_data1.db
Done.
```

[58]:

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

2015 Launch Records

- This query lists the failed landing outcomes in drone ship, their booster versions, and launch site names for in the months of the year 2015.

```
[60]: %sql SELECT substr(Date, 4, 2) AS MONTH, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE "LANDING _OUTCOME"= 'Failure (drone ship)' AND substr(Date,7,4)='2015';
 * sqlite:///my_data1.db
Done.
```

[60]:

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

- This query rank 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.

```
[63]: %sql SELECT "LANDING _OUTCOME", COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '04-06-2010' AND '20-03-2017' GROUP BY "LANDING _OUTCOME" ORDER BY QTY DESC;
 * sqlite:///my_data1.db
Done.
```

[63]:

Landing _Outcome	QTY
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

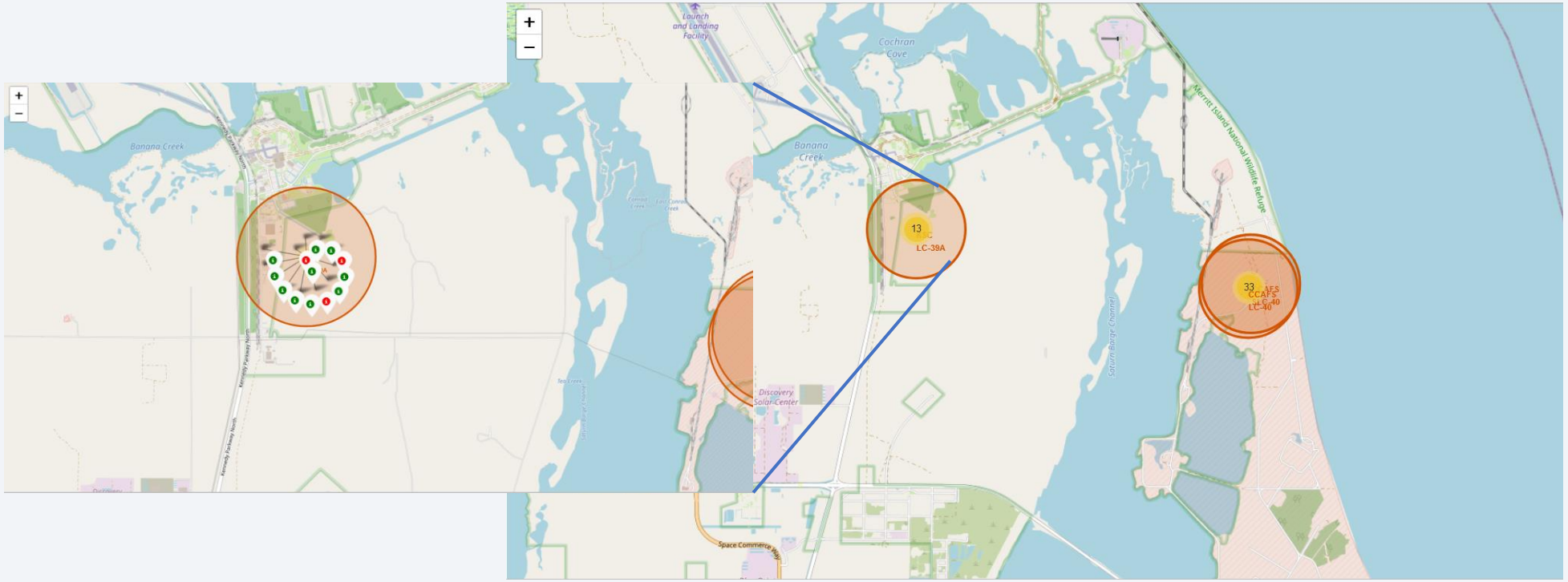
Section 3

Launch Sites Proximities Analysis

Launch sites locations

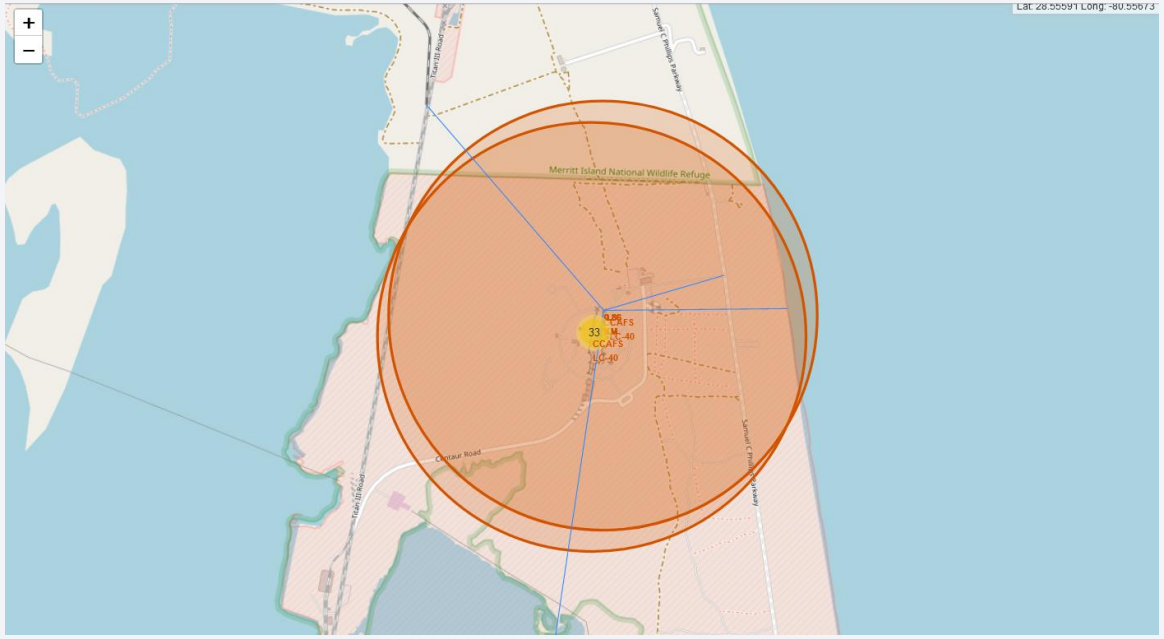
- Launch locations always happens near the sea, in costal regions, probably because part of the rockets retrieval lands at sea, also the abundance of water to deal with explosions of the launches.

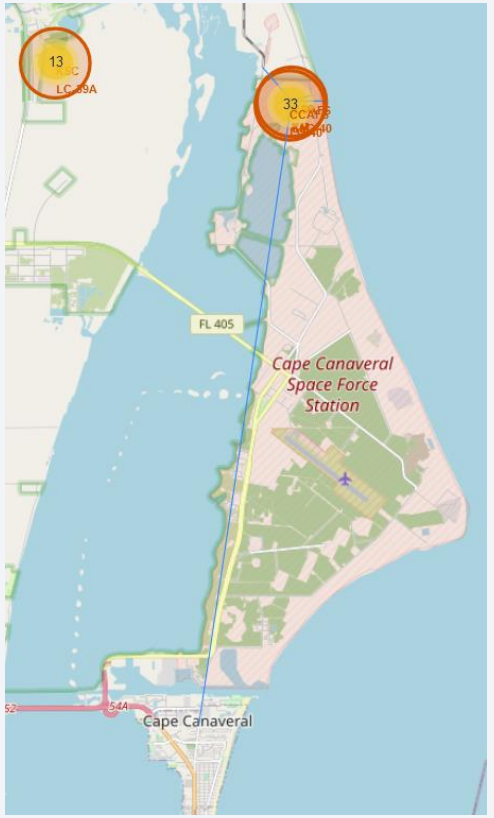
Launch outcomes by site



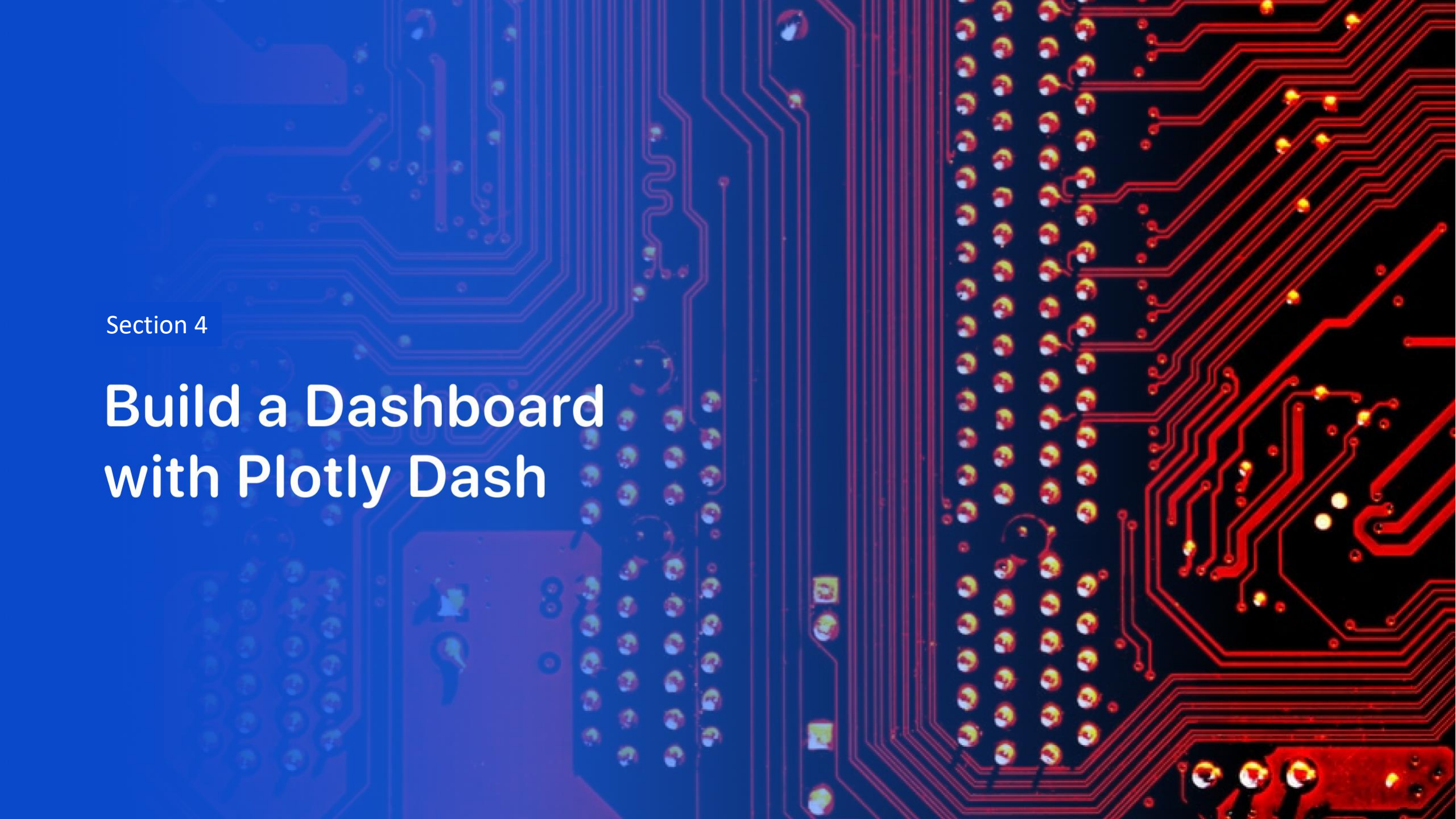
- Cluster of markers where added in each launch site to show how many launches happened, and the color coding to show successful(green) or failure(red) of the landings. The above map show Kennedy Space Center Marker cluster as example.

Logistics and safety of launch sites

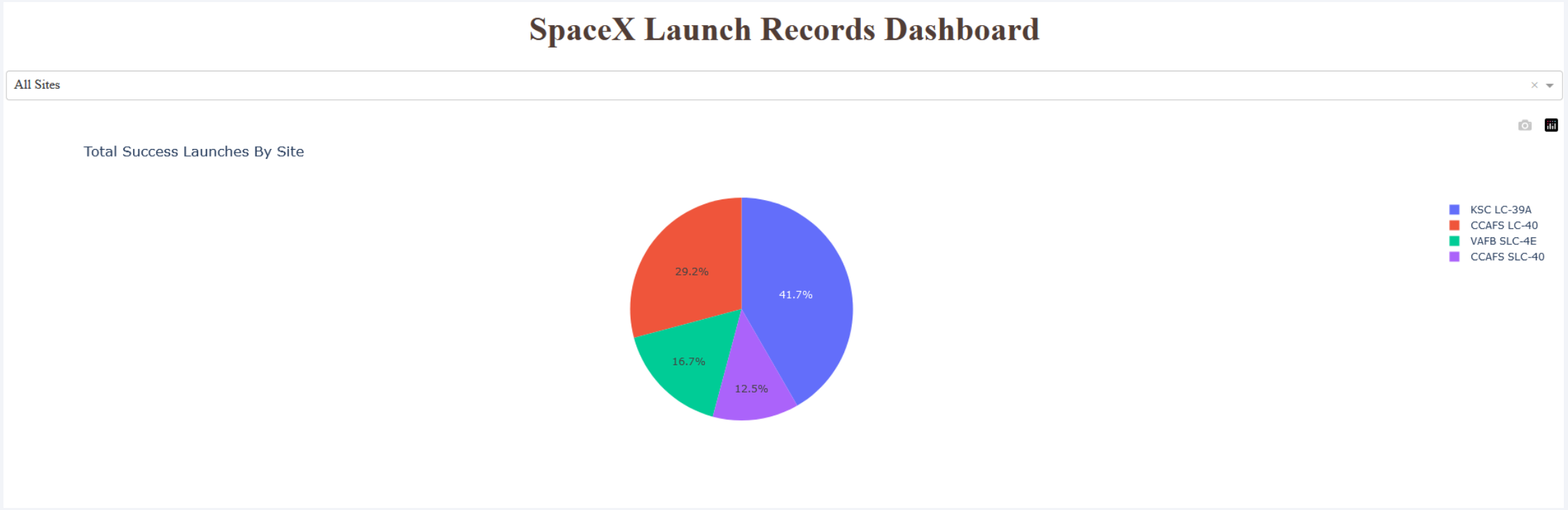




- Launch sites tend to have good access to Logistical points like Highways, railways, and Maritime transportation, marked in the image above with blue lines, also relatively far away from cities for security reasons.

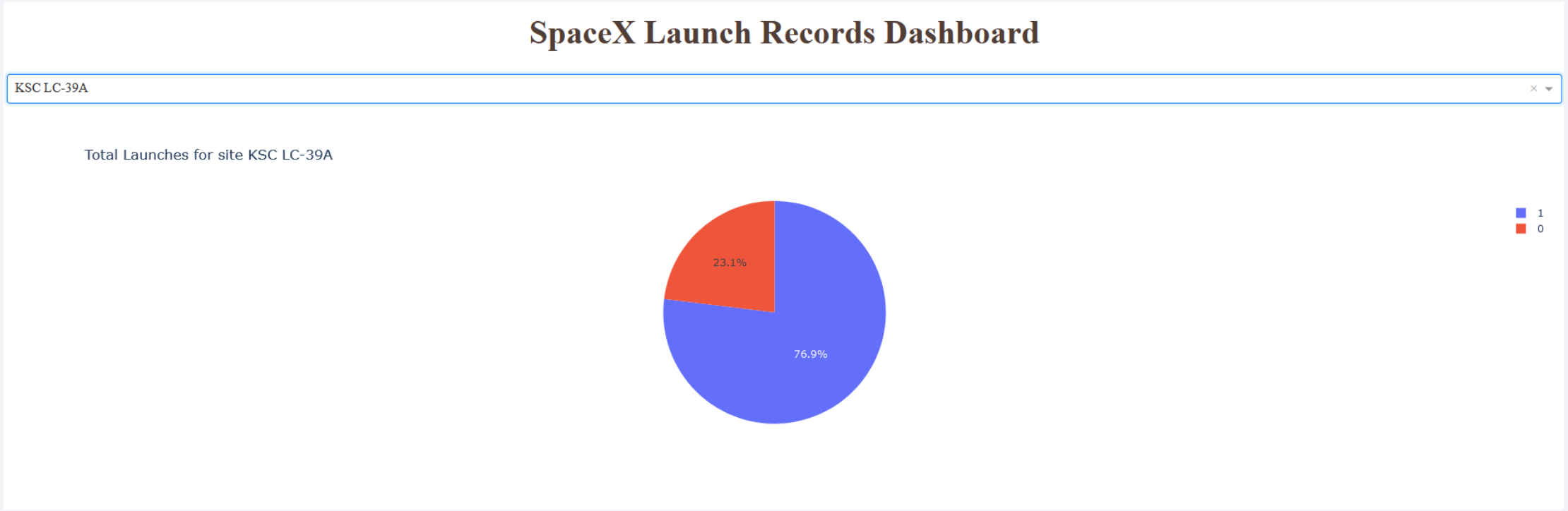

Section 4
Build a Dashboard with Plotly Dash

Successful launches by site



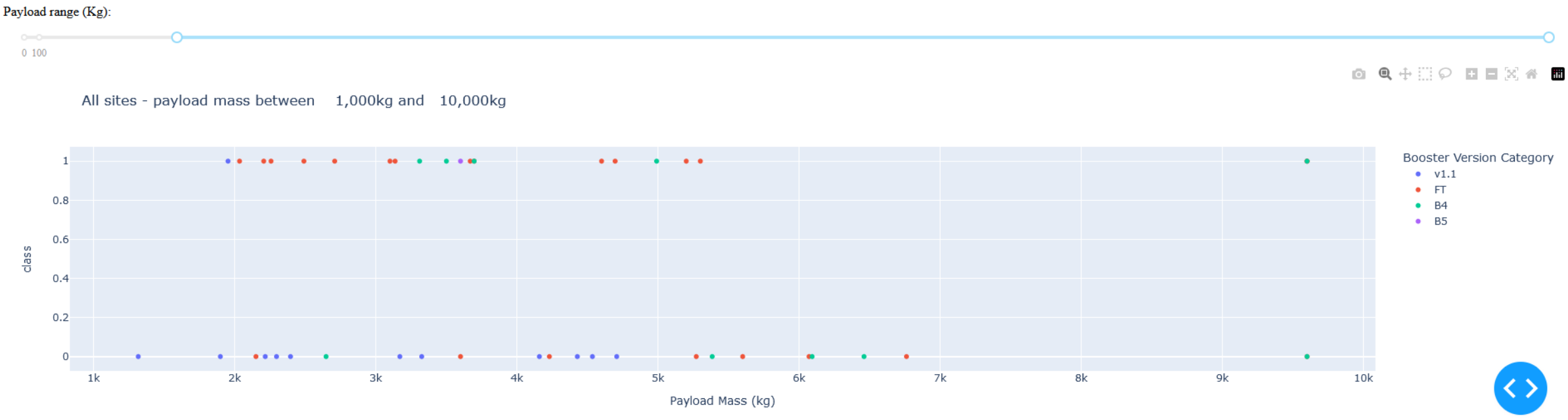
- Selection of the launch site seems to improve significantly the success of the retrieval of the rocket parts, with Kennedy Space Center being the most successful site for a launch.

Launch success ratio for KSC LC-39A



- In this graph we can check the magnitude of the success of Kennedy Space Center site launches.

Payload vs Launch outcome



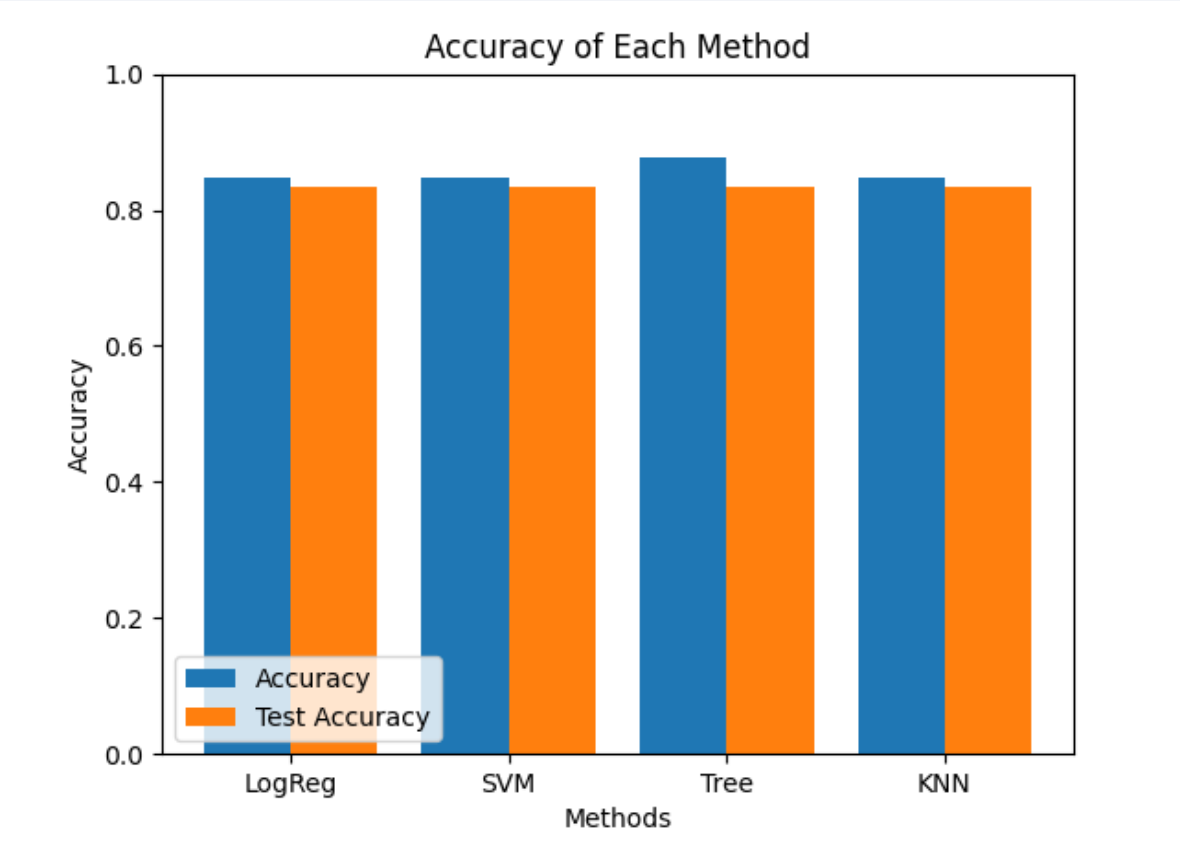
- Payloads under 5.500kg and from FT boosters are the most successful combination
- Avoid v1.1 Boosters, since according to the graph, it has the lowest success of the booster versions

Section 5

Predictive Analysis (Classification)

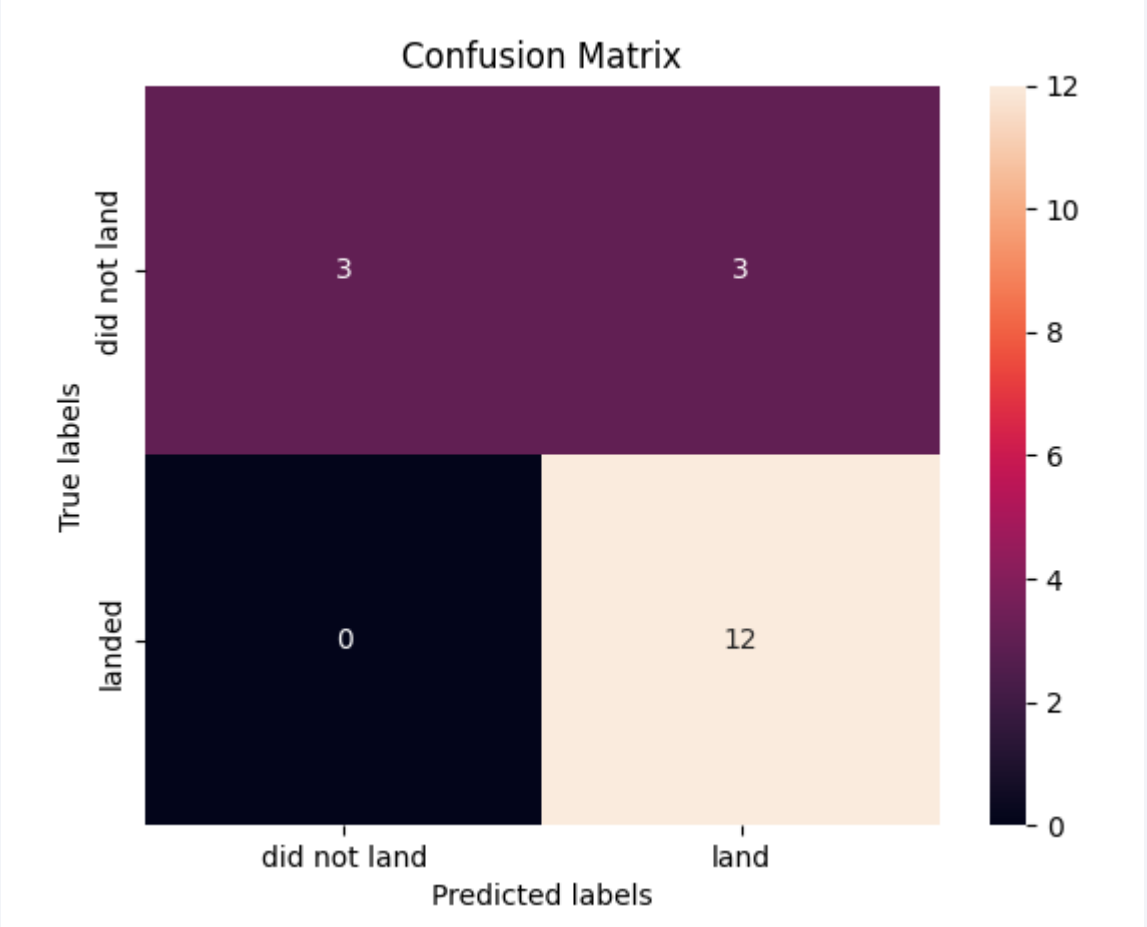
Classification Accuracy

- All methods seems to have similar accuracy while testing the model against the test data.
- Decision tree shows the best accuracy around 87% .



Confusion Matrix

- The confusion matrix of decision tree model shows high accuracy in true positive cases.
- In other hand, it failed to predict half of the cases of failure of the test data, showing them as successful landings, characterizing a type 1 error – false positive.



Conclusions

- Launch site choice is a significant factor of the chance of success of the rocket retrieval, Kennedy Space Center showing the highest ratio of success/failure among all evaluated locations.
- Payload weight also seems to be a important aspect to consider, with payloads over 7000kg having overall higher chance of retrieval, than lighter loads, when analyzing with different orbit choices.
- In the other hand, when booster type is considered, FT boosters has the highest success of the operation, favoring loads under 5500kg.
- Decision tree model showed more precision in predicting successful Landings, but is a weaker model to predict failure of launches.

Appendix

- Data was collected from:
  - https://api.spacexdata.com/v4/launches/past
  - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Resources:
  - Python, Jupyter Notebooks, SQL, Skills Network Labs
  - Python libraries: Numpy, Pandas, Scikit-learn, Matplotlib, Seaborn, Plotly and many others
  - Guidance from IBM and the Coursera team

Thank you!