IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Patricio Benavides
25/12/2023

Outline

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

Executive Summary

- Summary of methodologies

  - ✓ Data Collection through API
  - ✓ Data Collection with Web Scrapping
  - ✓ Data Wrangling
  - ✓ Exploratory Data Analysis with SQL
  - ✓ Exploratory Data Analysis with Data Visualization
  - ✓ Interactive Map with Folium
  - ✓ Machine Learning Prediction

- Summary of all results

  - ✓ Exploratory Data Analysis result
  - ✓ Interactive map and dashboards
  - ✓ Predictive Analytics result with Machine Learning models as Logistic Regression, SVM (Support Vector Machine) and KNN (k-nearest neighbors)

Introduction

- Project background and context

  Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars, while other providers had increased their cost up to 165 million. The principal saving is because Space X can reuse the first stage. The goal will predict if the first stage will land and calculate the cost of the launch. The project consists of the creation of a machine learning pipeline to predict if the first stage will land successfully.

- Problems you want to find answers
  - What are the main characteristics of a successful or failed landing?
  - What are the effects of each relationship of variables on a successful landing?
  - What operating conditions need to be in place to ensure a successful landing program?
  - Which machine learning model would work best (have the highest accuracy) to predict the outcome of a Falcon 9 first stage landing from a future launch?

Section 1

Methodology

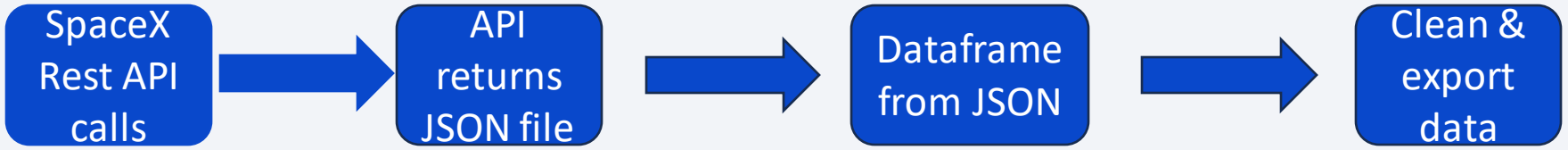
Methodology

Executive Summary

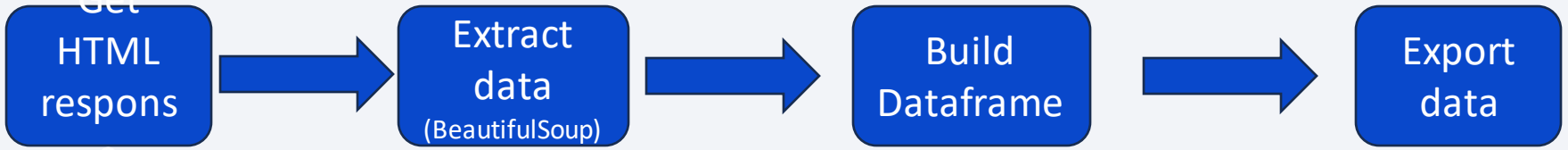
- Data collection methodology:
  - SpaceX REST API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Dropping unnecessary columns
  - One Hot Encoding for classification models
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- There are two sources where data are collected
  - The information obtained by the API are rocket, launches and payload information

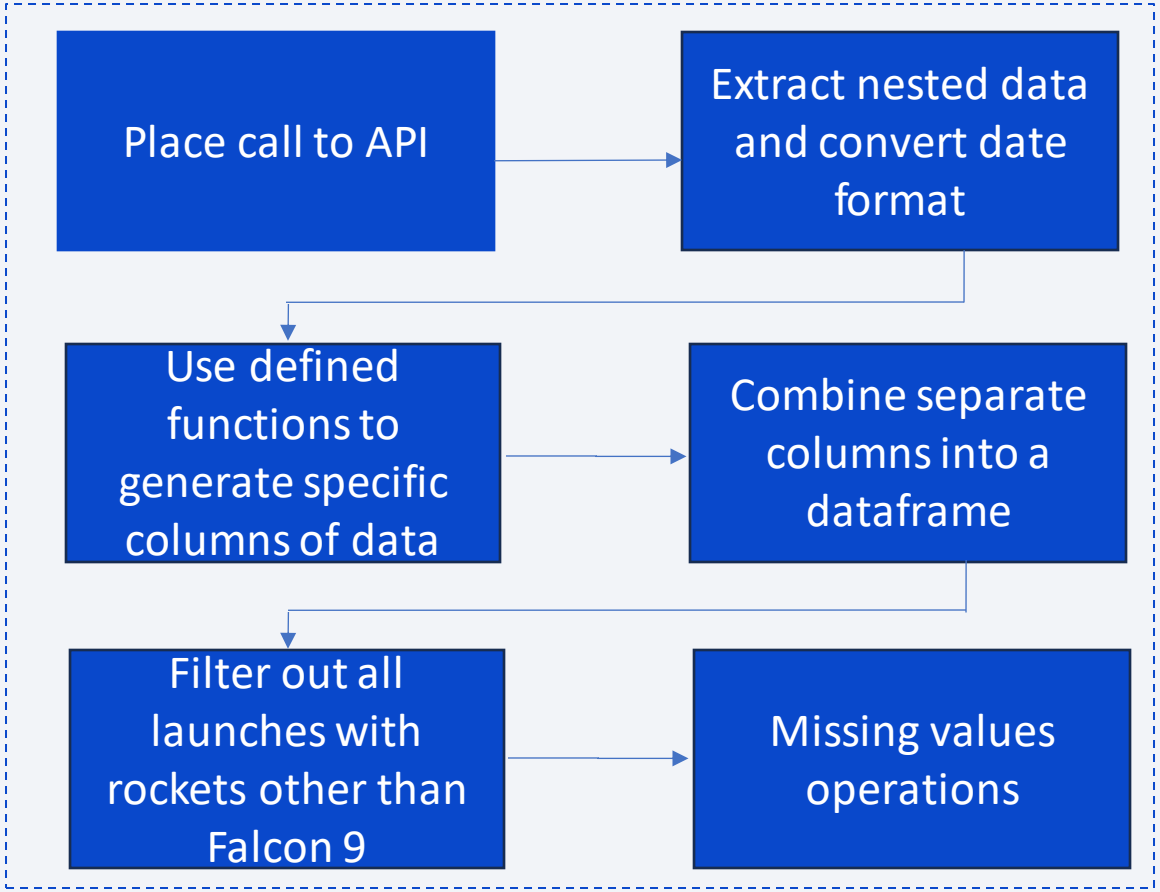


- Wikipedia webscrapping to obtain launches, landing and payload information



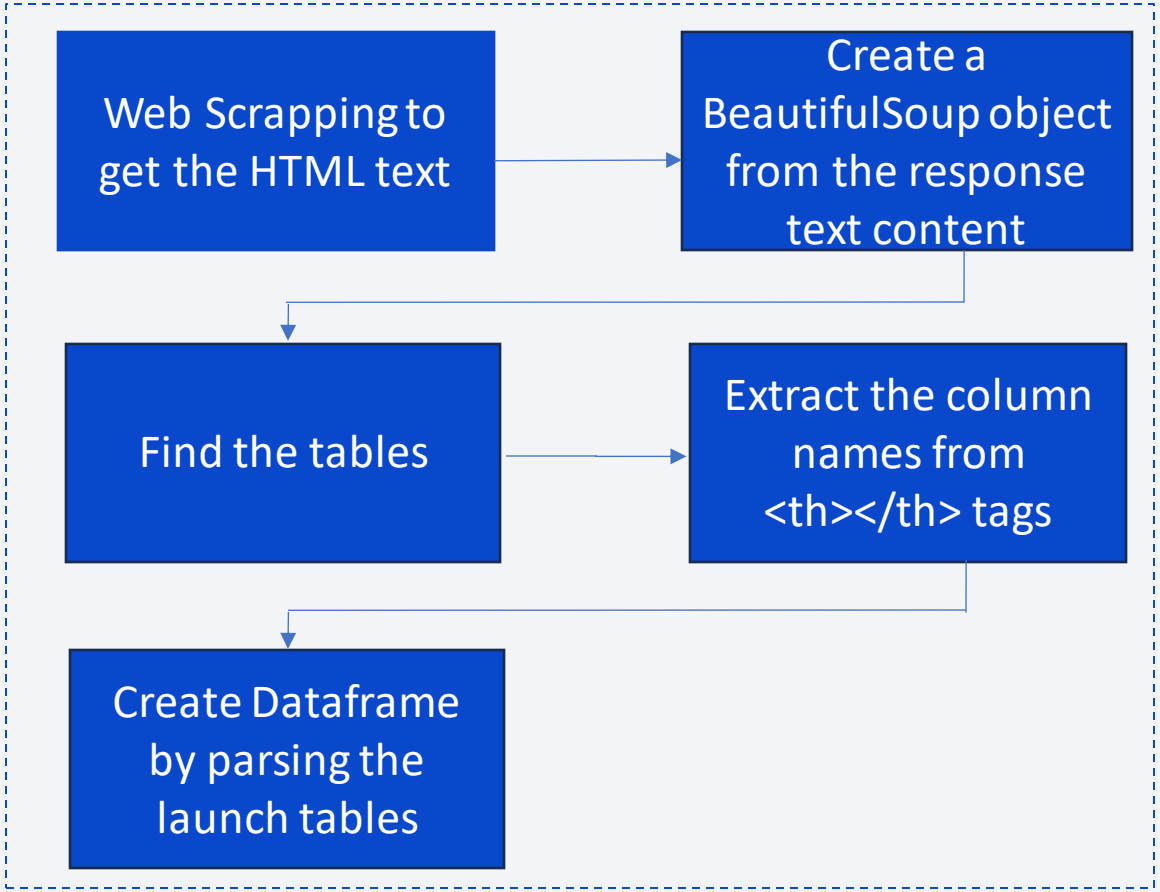
Data Collection – SpaceX API

- Collect data from a public site Space X API.
- Once a GET request has been made to the SpaceX API and the response received, the data can be placed into a Pandas DataFrame for further analysis.
- GitHub URL (Data Collection)



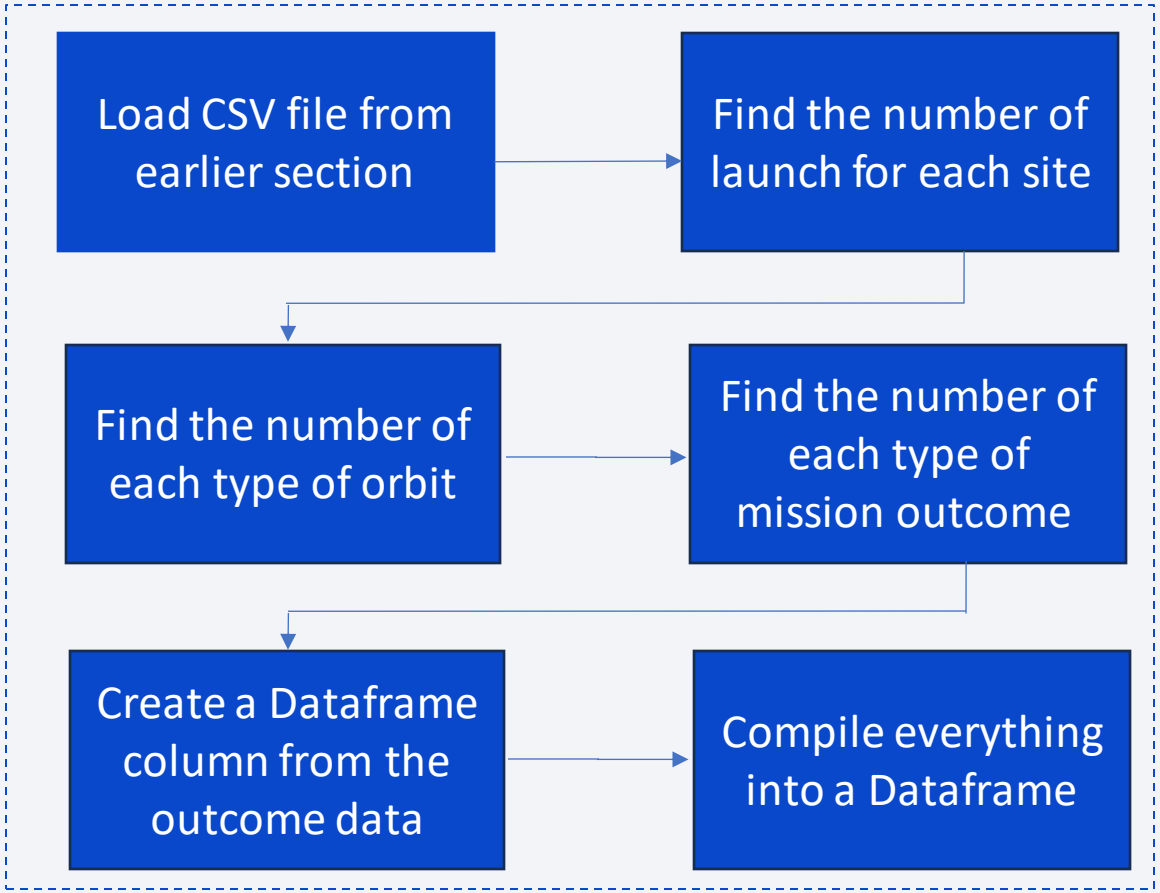
Data Collection - Scraping

- Wikipedia has tables of data about SpaceX launches.
- These tables can be scraped to extract launch data that can be put into a Pandas DataFrame for further analysis.
- GitHub URL (Web Scrapping)



Data Wrangling

- The launch sites, orbit types and mission outcomes were cleaned up.
- The handful of mission outcome types were converted to a binary classification where 1 means that the Falcon 9 first stage landing was a success and 0 means that it was a failure.
- GitHub URL (Data Wrangling)



EDA with Data Visualization

- Scatter Graphs
  - 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
- GitHub URL (EDA with Data Visualization)

- Bar Graph
  - Success rate vs. Orbit
- Line Graph
  - Success rate vs. Year

EDA with SQL

- Queries were written to extract information about:
  - ✓ Launch sites
  - ✓ Payload masses
  - ✓ Dates
  - ✓ Booster types
  - ✓ Mission outcomes
- GitHub URL (EDA with SQL)

Build an Interactive Map with Folium

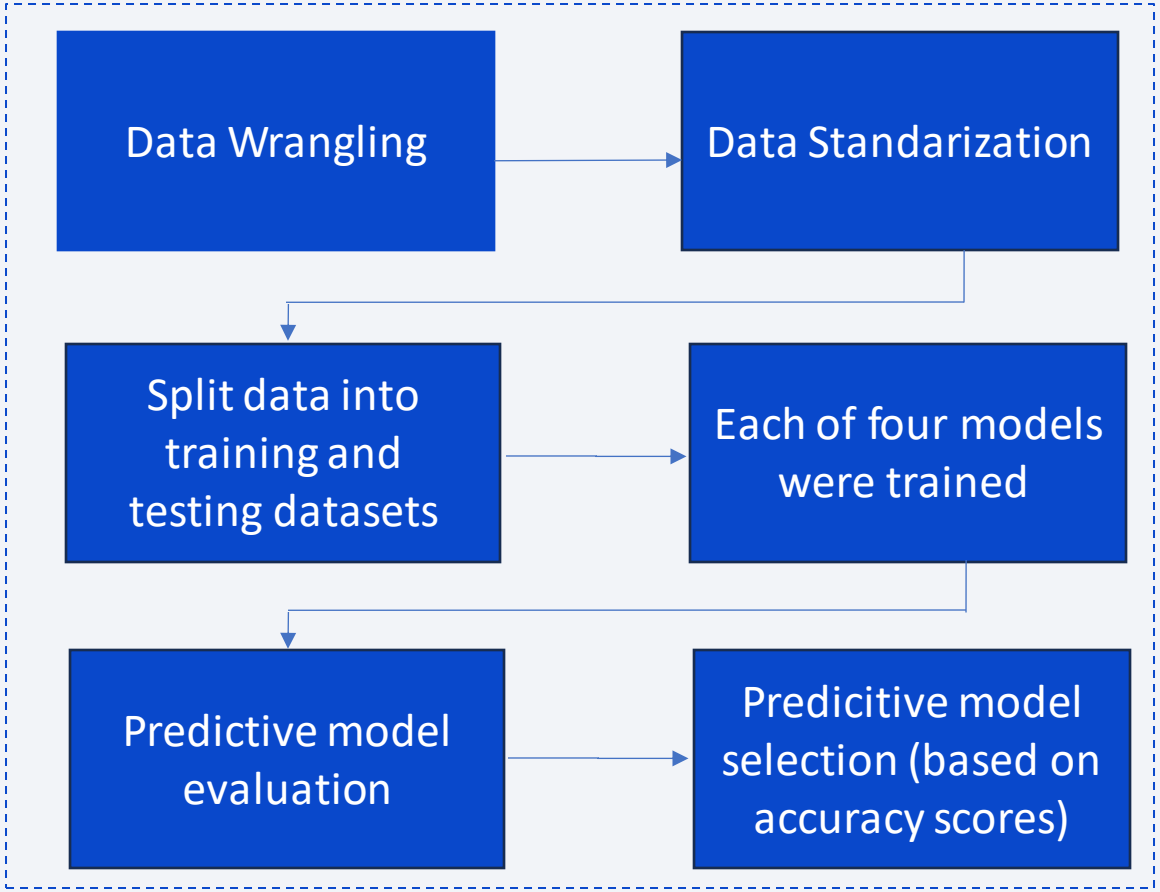
- Mark all launch sites on a map and for the NASA Johnson Space Center.
- Red circles at each launch site coordinates with a label showing its name.
- Mark the success/failed launches for each site on the map. Green for succesful landing and red for unsuccesful landing.
- Calculate the distance between a launch site to its proximities.
- Markers to show distance between launch site to different locations as railway, highway, city and coast. Plot a line between them.
- GitHub URL (Interactive Visual Analytics)

Build a Dashboard with Plotly Dash

- A launch site drop-down input component is used to select one or all launch sites for the pie chart and scatterplot.
- The pie chart based on selected site drop-down:
  - For all sites – the distribution of successful Falcon 9 first stage landings between all the sites
  - For one site – the distribution of successful and failed Falcon 9 first stage landings for the site selected
- A range slicer to filter the payload masses for the scatterplot.
- The scatterplot displays the distribution of Falcon 9 first stage landings split by payload mass, mission outcome and by booster version category
- GitHub URL (Dashboard with Plotly Dash)

Predictive Analysis (Classification)

- Logistic Regression, SVM (Support Vector Machine), Decision Tree, and KNN (k-Nearest Neighbors) machine learning models were trained on the training data set.
- Hyper-parameters were evaluated using GridSearchCV() and the best was selected using '.best_params_'.
- GitHub URL (Machine Learning Prediction)

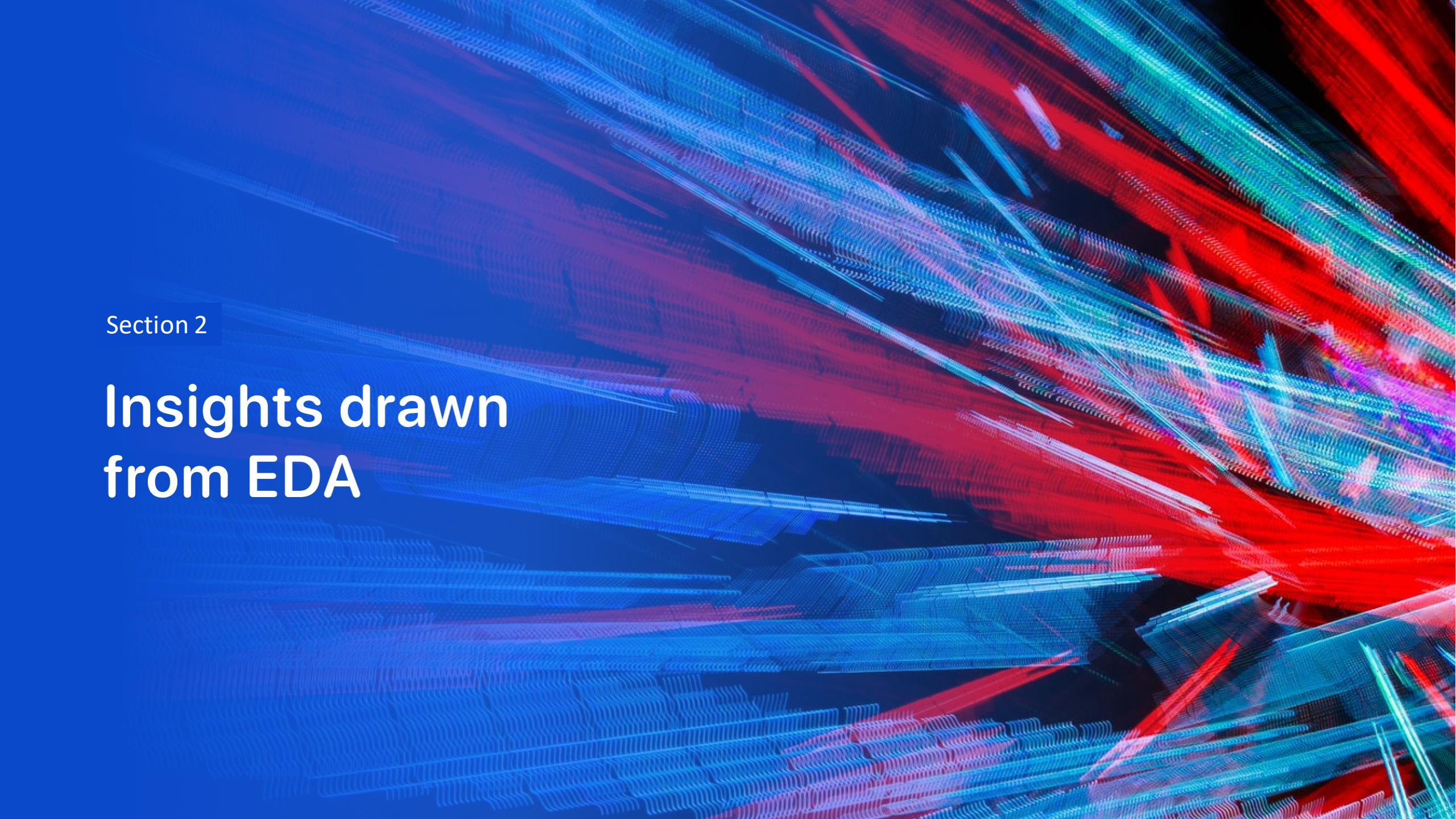


Results

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

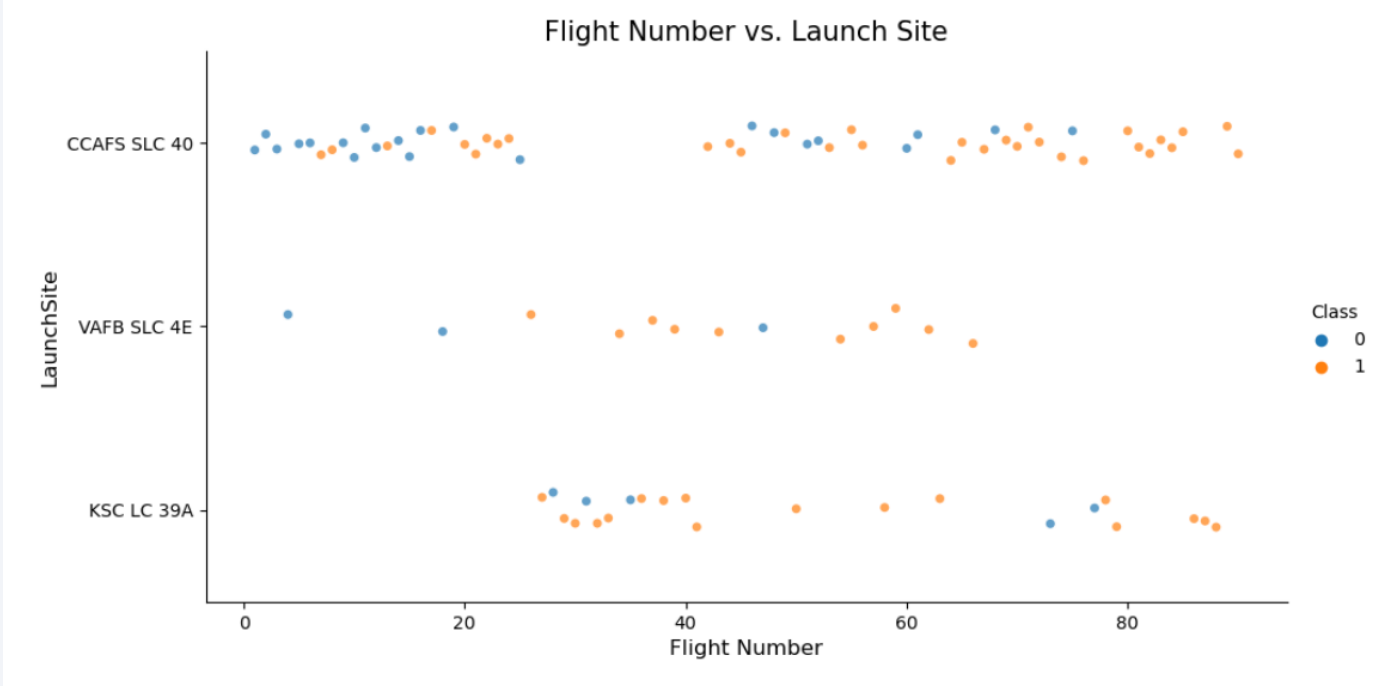
Section 2

Insights drawn from EDA



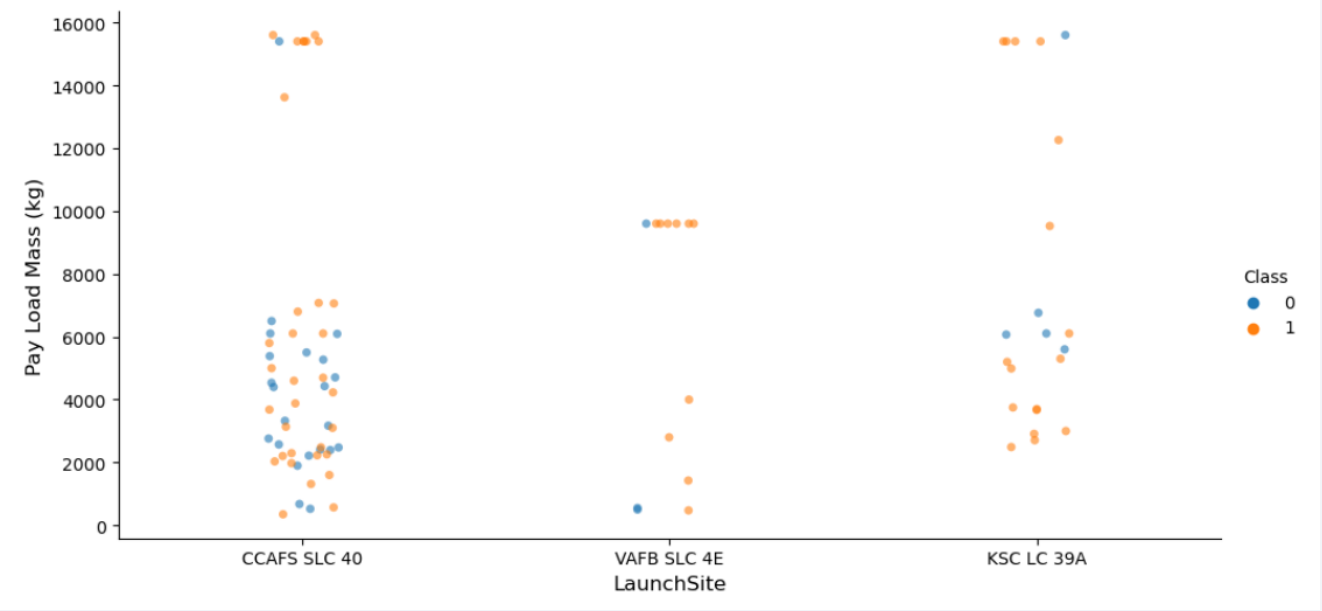
Flight Number vs. Launch Site

- Failed landings are indicated by the '0' Class (blue markers) and successful landings by the '1' Class (orange markers).
- Success rate varies with launch site.
- Successful Falcon 9 first stage landings become more prevalent as the flight number increases.



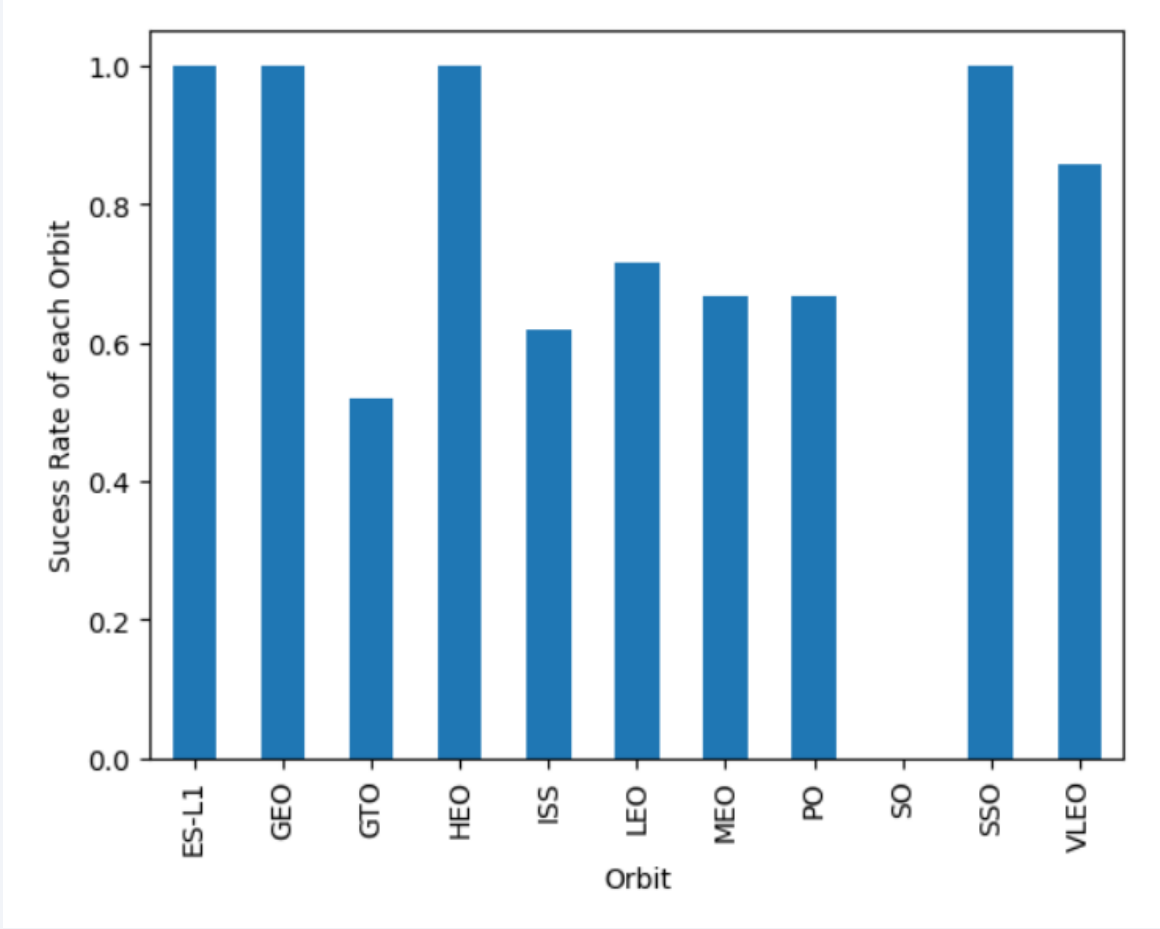
Payload vs. Launch Site

- Failed landings are indicated by the '0' Class (blue markers) and successful landings by the '1' Class (orange markers).
- For launch site CCAFS SLC 40 the Payload Mass and Launch Site appear not to be strongly correlated.
- In the launch site KSC LC 39A the failed landing are all grouped around a narrow band of payload masses.

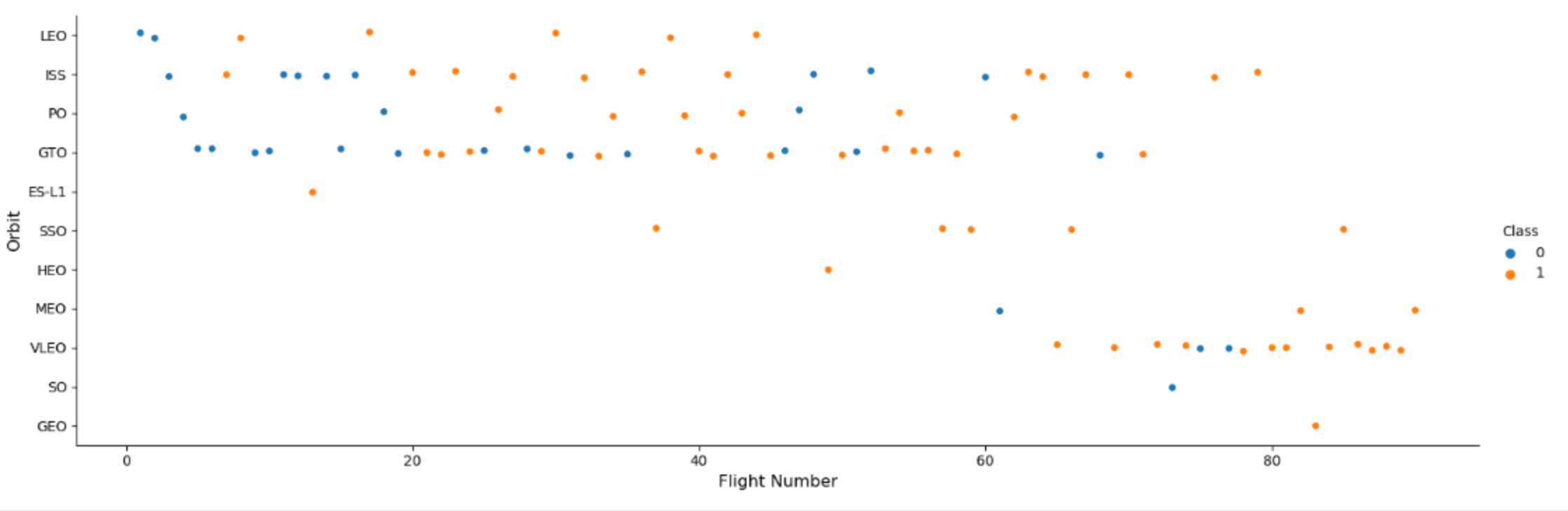


Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO orbits have the highest success rate.
- SO orbit have no succesful first stage landings.

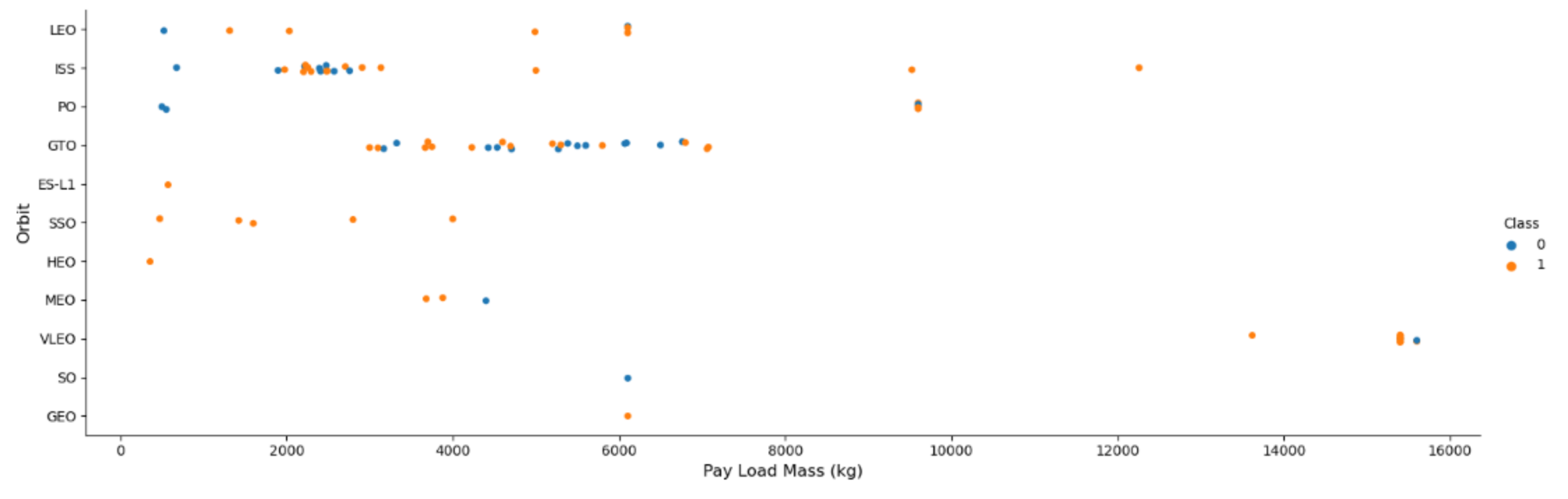


Flight Number vs. Orbit Type



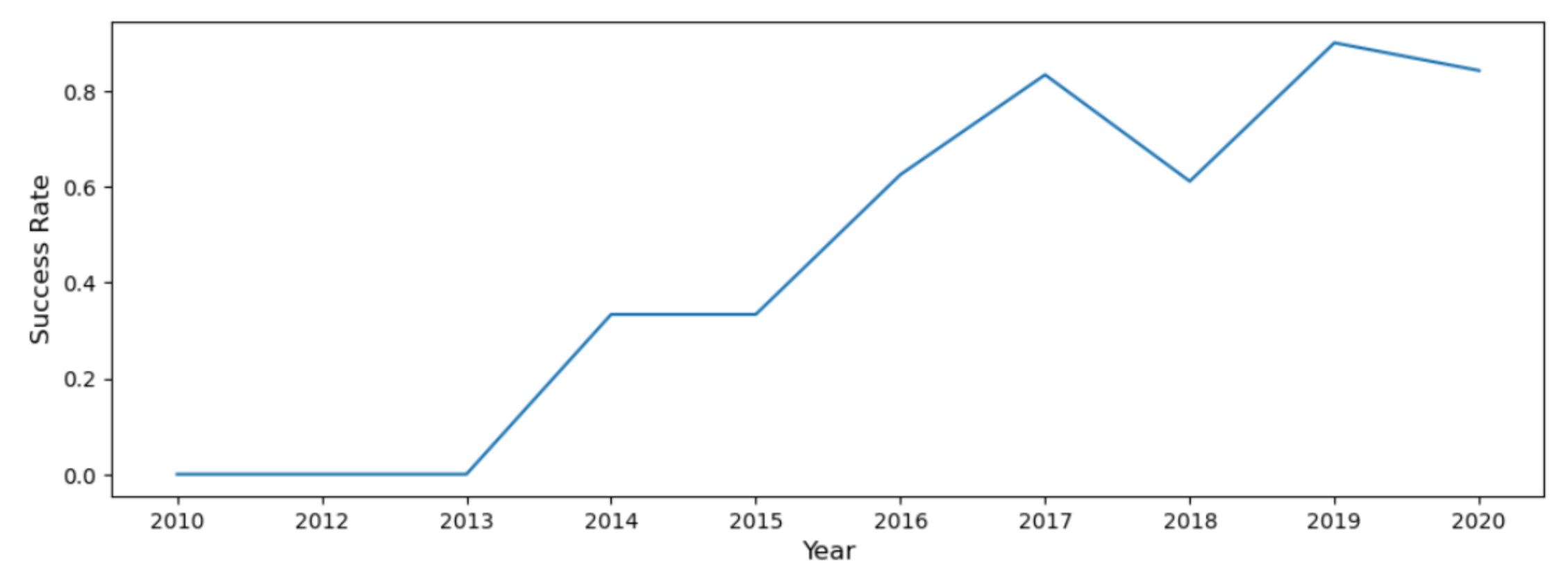
- We can highlight that some orbits as SSO, VLEO, LEO, and PO have a high success rate.
- In certain cases, there is a correlation between the number of flights and a greater number of flights, with a high success rate, as is the case of LEO.

Payload vs. Orbit Type



- For orbits as LEO, ISS and PO have a high success rate with heavy payloads. However, in GTO orbit there is no correlation between orbit and payload mass.

Launch Success Yearly Trend



- Since 2013, the success rate increase.

All Launch Site Names

- Question: What are the names of the unique launch sites?
- Query:

```
%sql select DISTINCT (LAUNCH_SITE) from SPACEXTBL;
```

- Result:

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

- Explanation: There are 4 unique launch sites.

Launch Site Names Begin with 'CCA'

- Question: Display 5 records where launch sites begin with the string 'CCA'
- Query:

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

- Result:

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

- Explanation: The WHERE followed by LIKE clause filter launch sites that contain the substring 'CCA'. The expression LIMIT 5 filter five records.

Total Payload Mass

- Question: What is total payload mass carried by boosters launched by NASA?
- Query:

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

- Result:

SUM(PAYLOAD_MASS__KG_)
45596

- Explanation: The total payload mass is 45.596 kg.

Average Payload Mass by F9 v1.1

- Question: What is average payload mass carried by booster version F9 v1.1?
- Query:

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

- Result:

AVG(PAYLOAD_MASS__KG_)
2928.4

- Explanation: The average payload mass is 2928.4 kg.

First Successful Ground Landing Date

- Question: When the first succesful landing outcome in ground pad was acheived?
- Query:

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

- Result:

MIN(DATE)
2015-12-22

- Explanation: The first succesful landing outcome in ground pad was in December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Question: What are the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000?
- Query:

```
%sql SELECT Booster_Version \
FROM SPACEXTBL \
WHERE LANDING_OUTCOME = 'Success (drone ship)' \
and PAYLOAD_MASS__KG_ between 4000 and 6000
```

- Result:

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

- Explanation: The 4 boosters versions that have successfully landed on drone ship with the payload range between 4000 and 6000 are listed above.

Total Number of Successful and Failure Mission Outcomes

- Question: What was the total number of successful and failure mission outcomes?
- Query:

```
SELECT (SELECT count(*) FROM SPACEXTBL WHERE lcase(landing__outcome) \
        LIKE '%success%') AS "Success", count(*) AS "Failure" FROM SPACEXTBL \
        WHERE lcase(landing__outcome), NOT LIKE '%success%';
```

- Result:

Success	Failure
61	40

- Explanation: There where 61 succesful and 40 failure mission outcomes.

Boosters Carried Maximum Payload

- Question: What were the names of the booster which have carried the maximum payload mass?
- Query:

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

- Explanation: The maximum payload mass was carried by 12 booster versions listed above.

Result:

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

- Question: List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.
- Query:

```
%sql SELECT substr(Date,6,2) as 'month NAMES', LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE \
FROM SPACEXTBL \
where substr(Date,0,5)='2015' and LANDING_OUTCOME='Failure (drone ship)'
```

- Result:

month NAMES	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Explanation: There where 2 failed landing outcomes with a drone ship in 2015, in january and april from CCAFS LC-40.

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

- Question: 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.
- Query:

```
%sql SELECT LANDING_OUTCOME, count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '2010-06-04' and '2017-03-20' group by LANDING_OUTCOME order by count_outcomes DESC;
```

- Result:

Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- Explanation: There most common landing outcome w


Section 3
Launch Sites
Proximities Analysis

Falcon 9 Launch Site Locations

- CCAFS LC-40 (Florida, USA)

Cape Canaveral Air Force Station Launch Complex 40

- CCAFS SLC-40 (Florida, USA)

Cape Canaveral Air Force Station Space Launch Complex 40

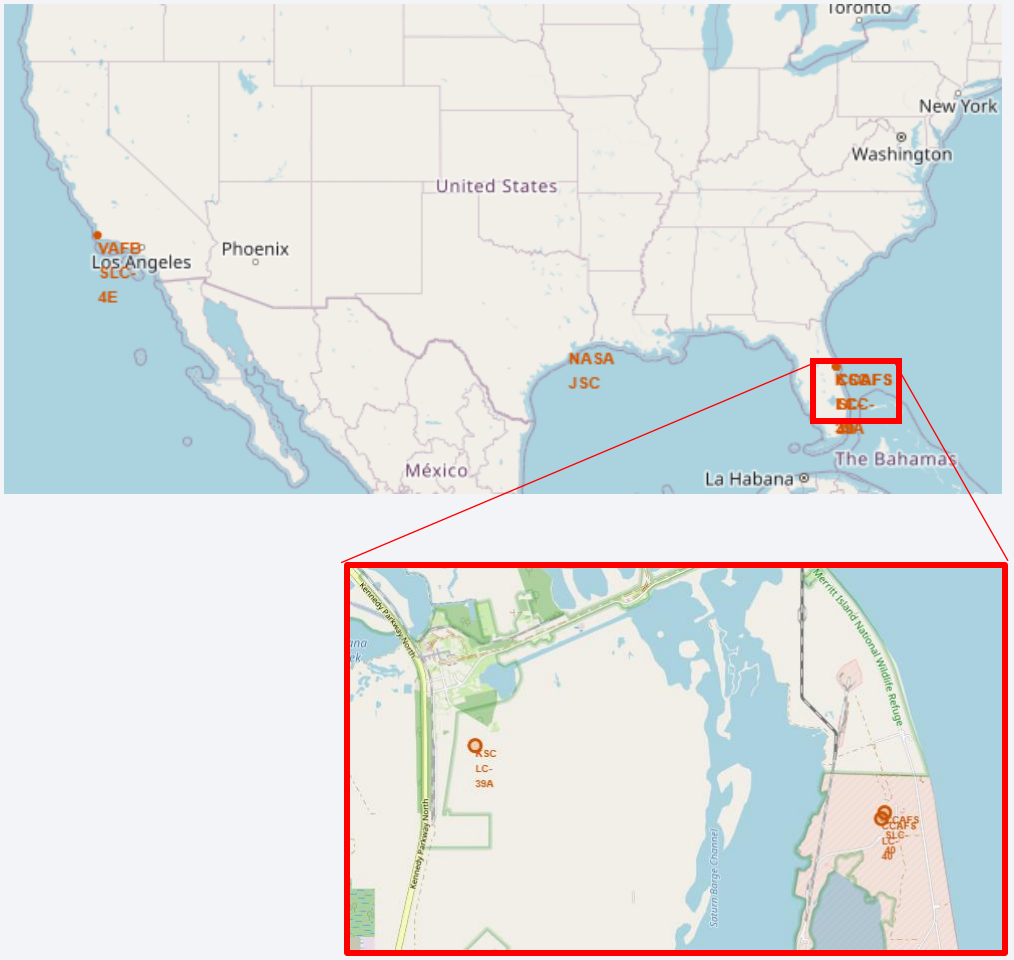
- KSC LC-39A (Florida, USA)

Kennedy Space Center Launch Complex 39A

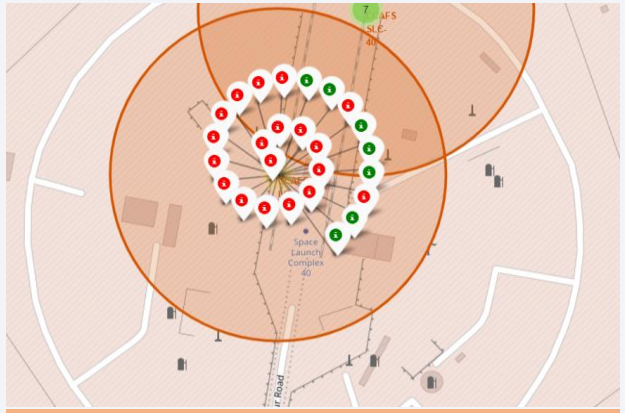
- VAFB SLC-4E (California, USA)

Vandenberg Air Force Base Space Launch Complex 4E

• The map is centered by the NASA Johnson Space Center (Houston, Texas)

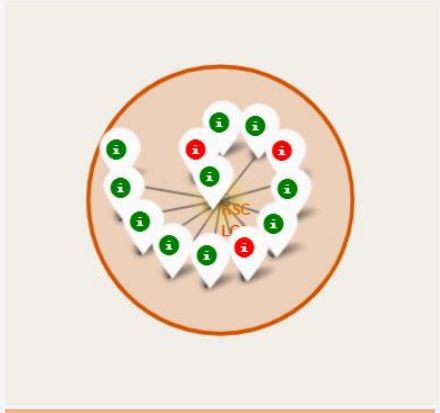


Landing state map markers

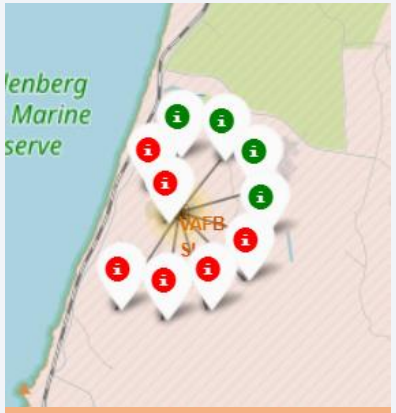


CCAFS LC-40
Success: 7 Fails: 19

CCAFS SLC-40
Success: 3 Fails:4



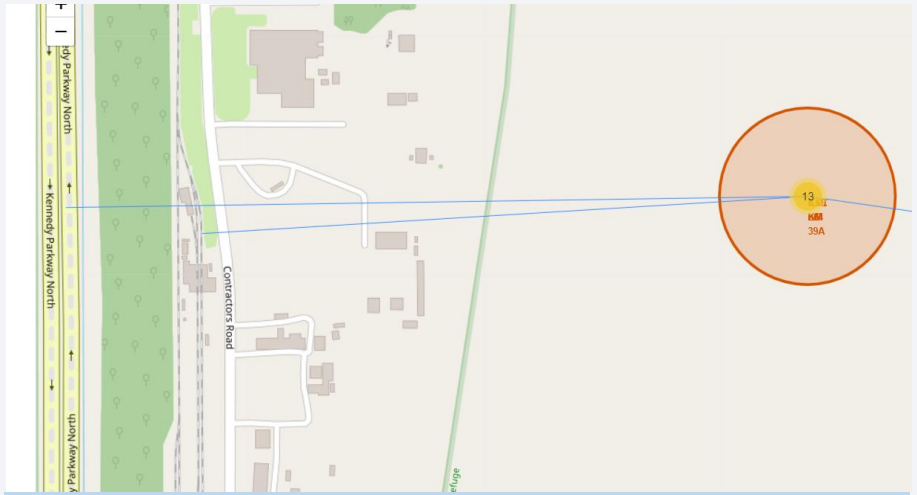
KSC LC-39A
Success: 10 Fails:3



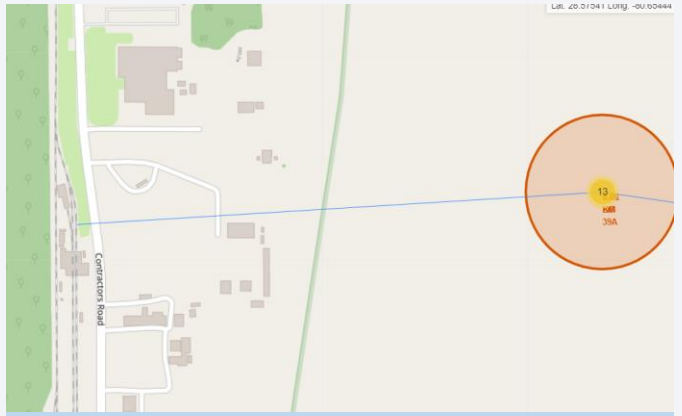
VAFB SLC-4E
Success: 4 Fails:6

- Green marker: landing success.
- Red marker: landing fails
- They are grouped on the map to be associated with the geographical coordinates for the launch site.

Screenshot map of launch site proximities

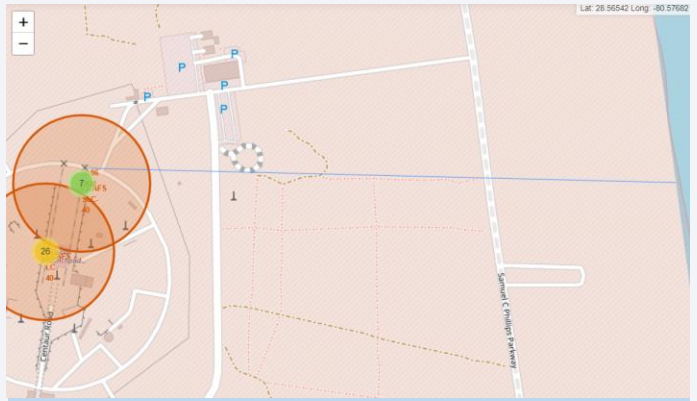


KSC LC-39A
0.84 km to highway

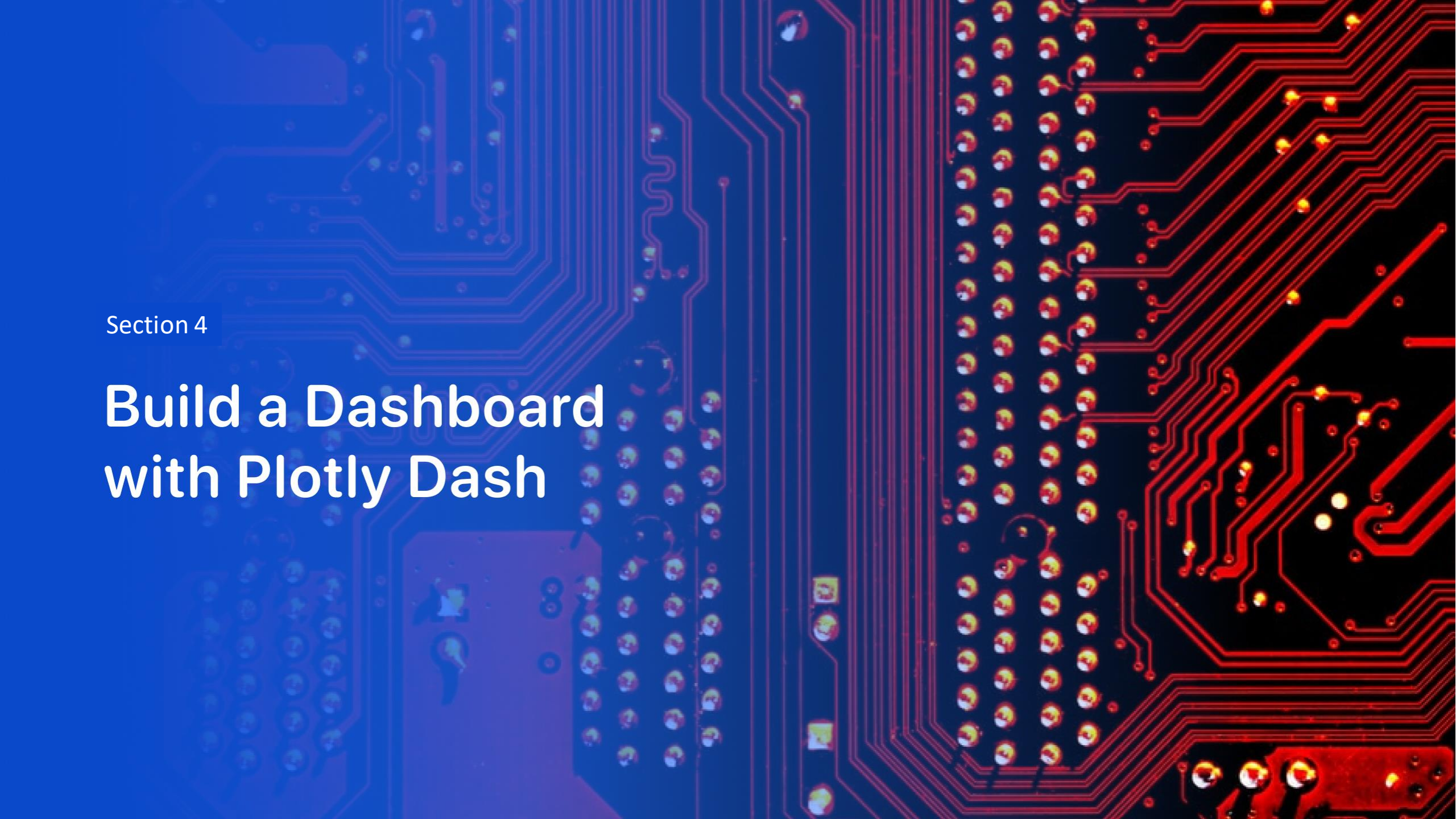


KSC LC-39A
0.69 km to railroad

- The distance to the highway, the coast and the railway lines was measured, and the shortest distance was selected.

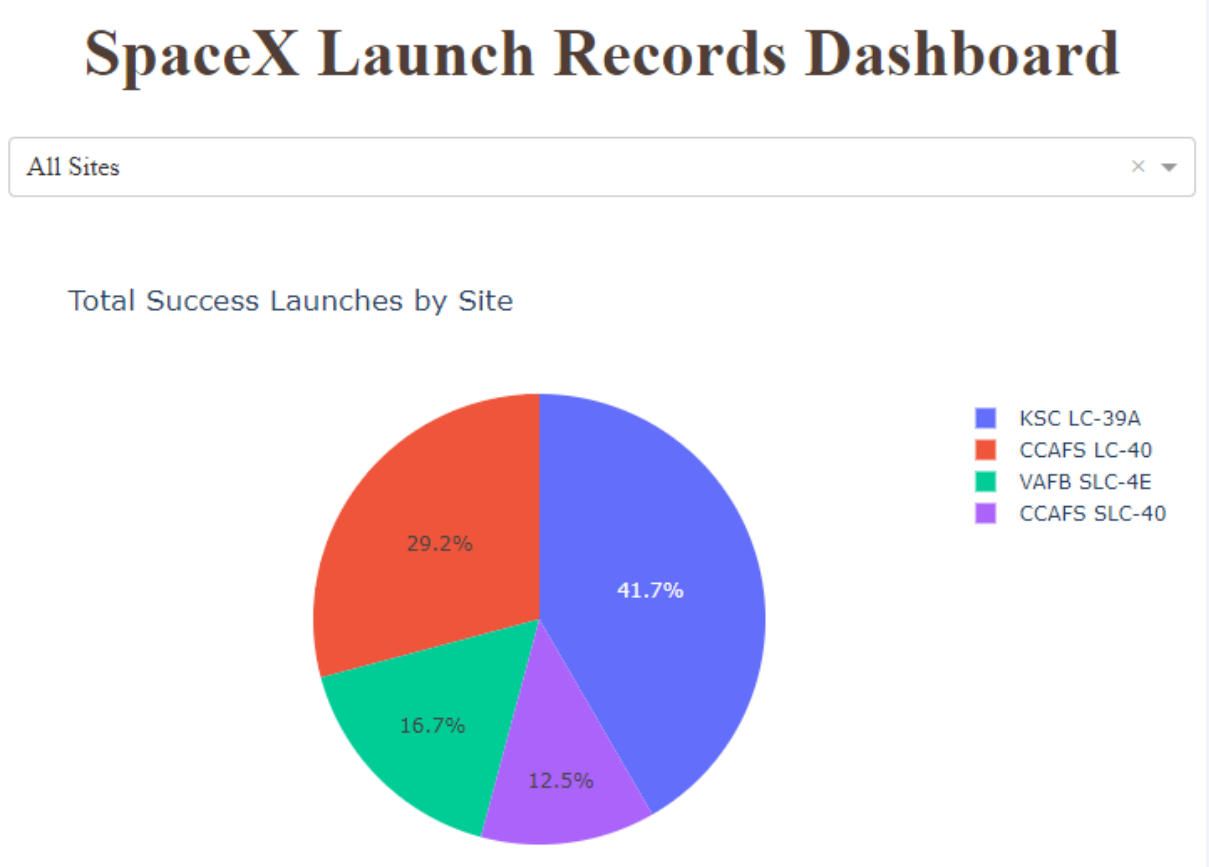


CCAFS SLC-40
0.86 km to coast


Section 4
Build a Dashboard with Plotly Dash

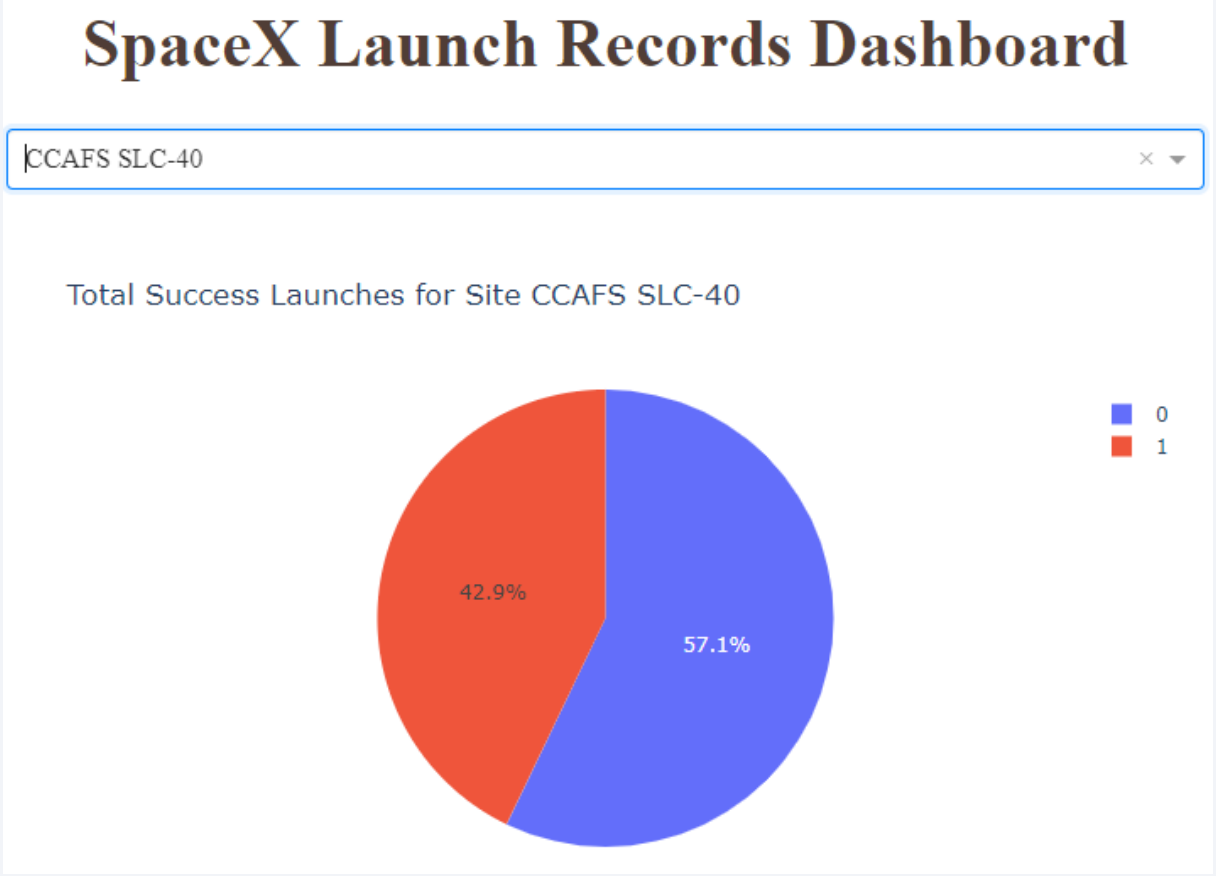
Launch success for ALL sites

- You could select one or all launch sites in the dropdown menu.
- As you can see in the pie chart, there is the distribution of successful first stage landing.
- KSC LC-39A (blue) has the highest rate of successful Falcon 9 first stage landing outcome.
- CCAFS SCL-40 (purple) has the lowest rate.



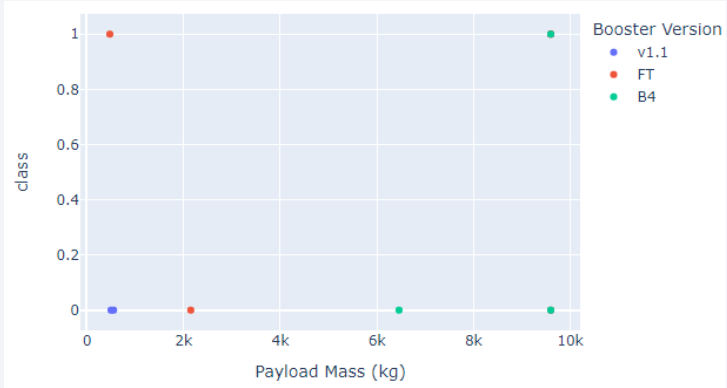
<Dashboard Screenshot 2>

- Successful landings is determined for class 1 (red color).
- CCAFS SLC-40 has 42.9% of landing successful percentage, the highest rate.

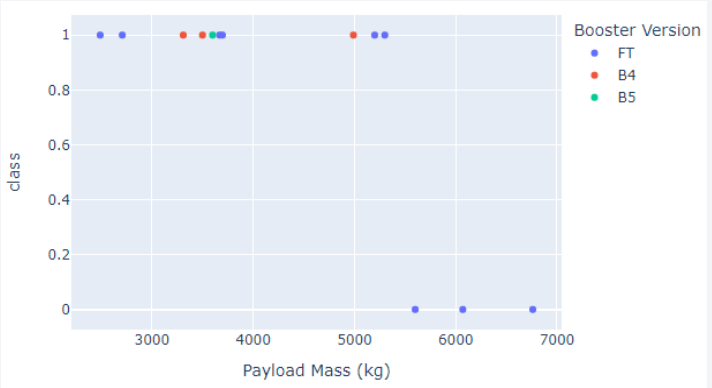


Payload vs. Launch Outcome Scatter Plot

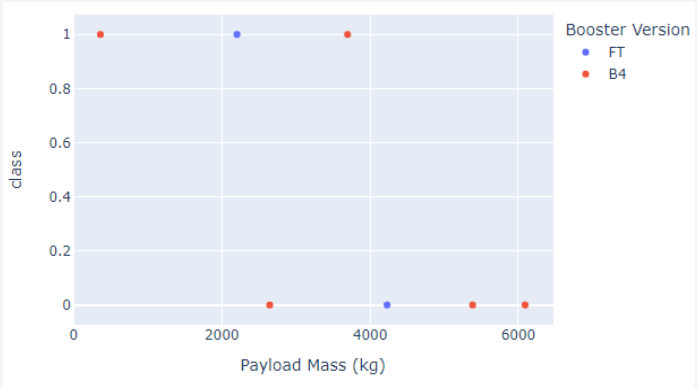
CCAFS LC-40
Best Booster: FT



VAFB SLC-4E
Best Booster: FT & B4



KSC LC-39A
Best Booster: FT



CCAFS SLC-40
Best Booster: B4

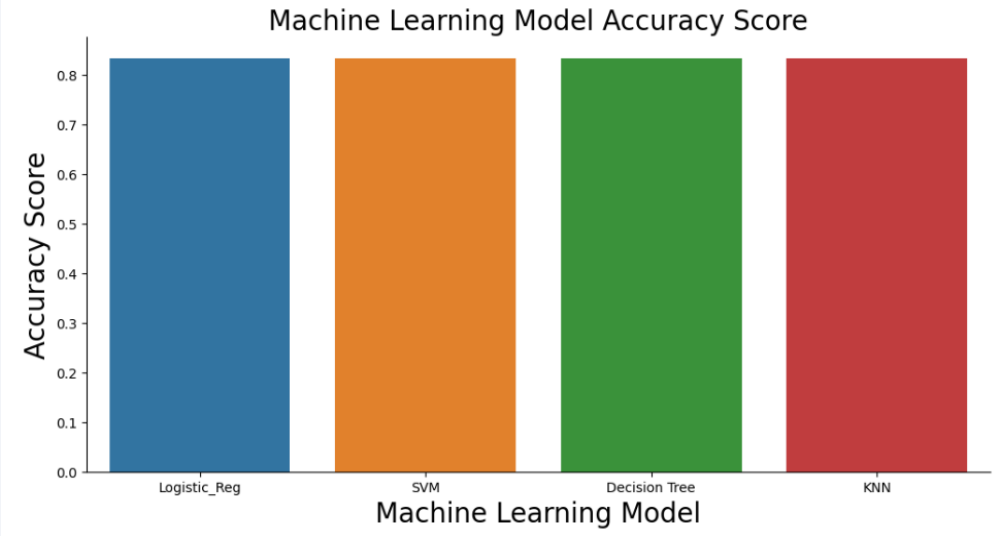
- The FT and B4 booster version has the best success rate.
- Payload range between 2.000 to 6.000 kg.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

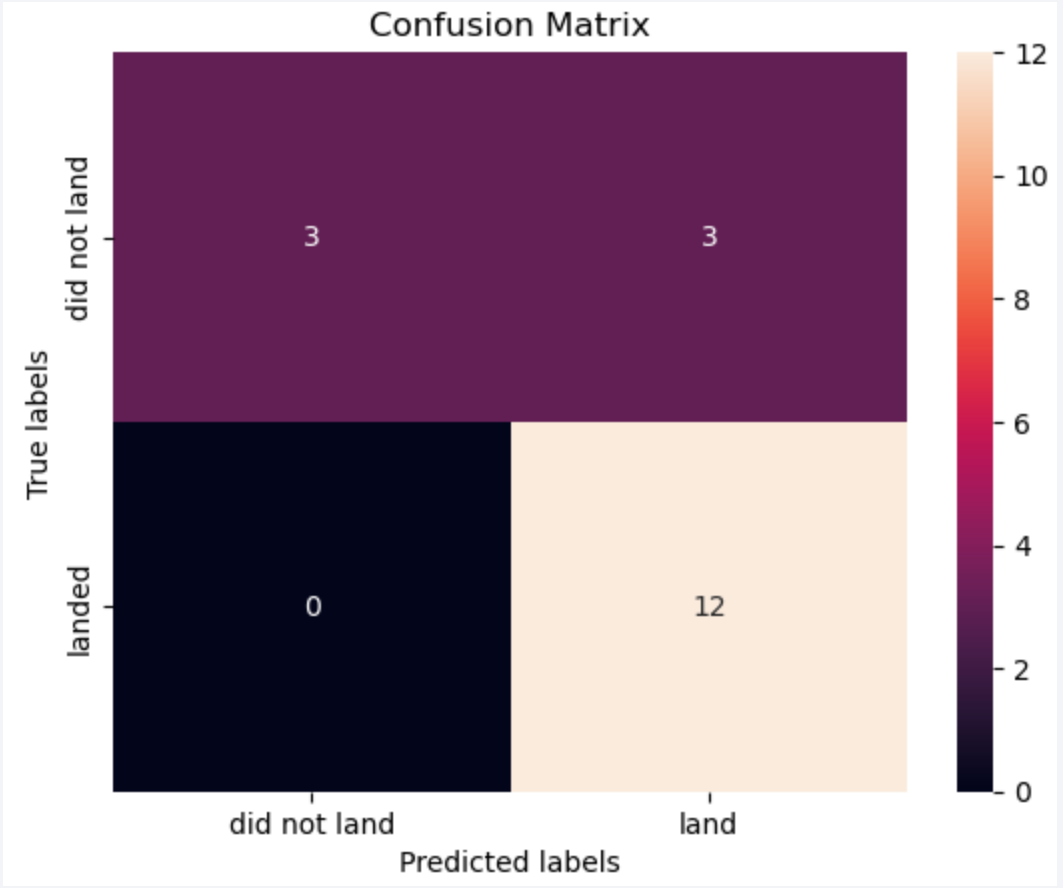
- All four models has the same accuracy score. There is no relevance between them.



Method	Test Data Accuracy
Logistic_Reg	0.833333
SVM	0.833333
Decision Tree	0.833333
KNN	0.833333

Confusion Matrix

- All the models has the same Confusion Matrix result.
- 12 True Positive
- 3 True Negative
- 3 False Positive



Conclusions

- The machine learning models can be used to predict the Space X Falcon 9 first stage outcomes.
- The model let us define the best orbit, the payload mass range and booster version to define which one has a better success rate than others.
- SpaceX's Falcon 9 first stage landing outcomes have been trending towards greater success as more launches are made.

Appendix

Jupyter Notebooks

- LAB_1 Collecting Data
- LAB_2 Webscrapping
- LAB_3 Data wrangling
- LAB_4 EDA with SQL
- LAB_5 EDA with Data Visualization
- LAB_6 Interactive Visual Analytics
- LAB_7 Machine Learning Prediction
- LAB_8 Dashboard with Plotly

Dataset

- Dataset 1
- Dataset 2
- Dataset 3
- Launch GEO
- Webscraped

Thank you!