Winning Space Race with Data Science

Anna K. Renner

10/10/2023

Outline

- Executive Summary
- Introduction
- Methodology / Results
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Visualizations
  - Interactive Maps with Folium
  - Plotly Dash Interactive Dashboard
  - Predictive Analysis
- Conclusion
- Appendix

Executive Summary

Research Problem:

The cost of rocket launches is majorly influenced by reusage of the first stage. Thus, via predicting a successful landing of the first stage, companies are enabled to reduce costs.

Research Objective:

This project attempts to identify the factors for a successful first stage landing of SpaceX' Falcon 9 rocket via:

- collecting data using SpaceX REST API and webscraping methods
- wrangle data to create success/fail outcome variable
- explore data with SQL and data visualization techniques
- predict landing outcome via several machine learning techniques

Results:

Significantly landing success-influencing factors are launching success, landing sites and orbits. In detail, launching success increased over time and launch site KSC LC-39A has the highest success rate among landing sites. Additionally, orbits (ES-L1, GEO, HEO, SSO) have a 100 % success rate.

Introduction

Background:

SpaceX was founded in the US and aims to develop technologies to make space travel to Mars possible. After development of the rocket model Falcon 9, SpaceX became a major supplier of the international space station. By pioneering in reusing rocket components e.g. the first stage, Space X could reduce costs by the factor 7 ($62 million per launch). Other providers, which are unable to reuse the first stage, cost upwards of $165 million each.
Therefore, via determination if the first stage will land, we can predict the price of the launch. To do this, we can use public data and machine learning models to predict whether SpaceX – or a competing company – can reuse the first stage.

Methodology

Steps:

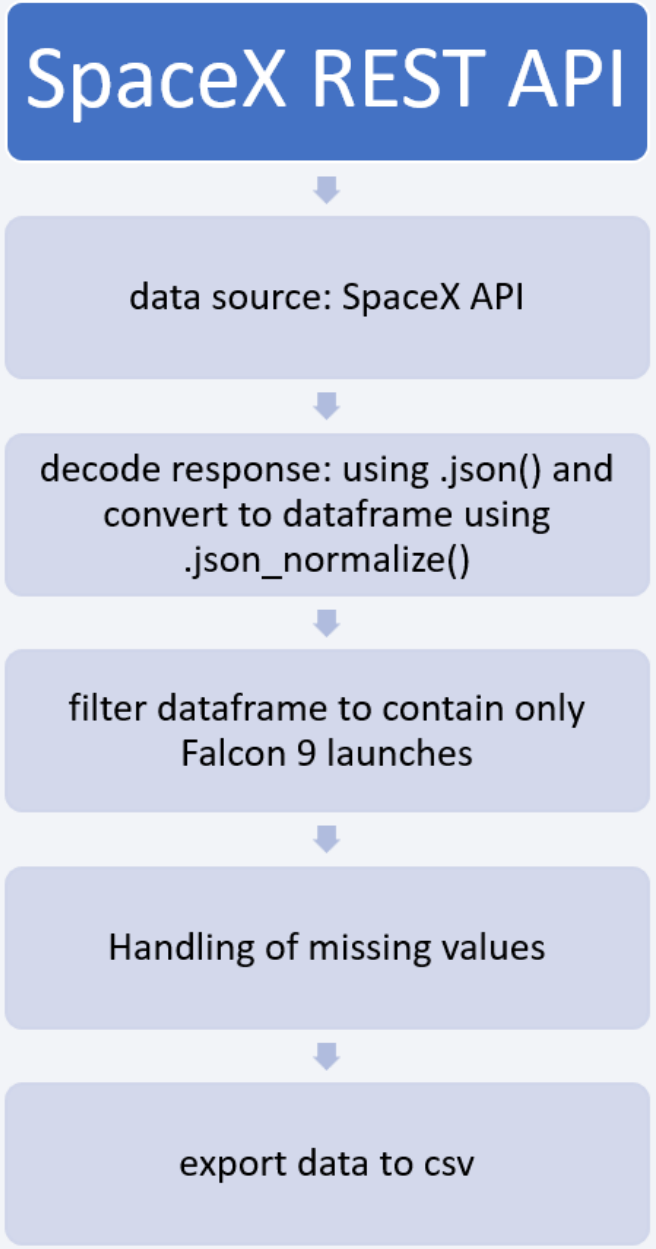
1. **Data collection**
    1. SpaceX REST API
    2. Webscraping
2. **Data Wrangling**
    1. filtering of data and handling of null values
    2. application of one-hot encoding for preparation of modelling
3. **Exploratory Data Analysis**
    1. SQL
    2. Visualization
4. **Interactive Visualization:**
    1. Folium: creation of maps for launching sites
    2. Plotly Dash: creation of interactive dashboard for results
5. **Machine Learning Models:**
    1. prediction of landing outcomes using classification models

Data Collection – SpaceX API

- Request data from SpaceX API (rocket launch data)
- Decode response using .json() and convert to a dataframe using .json_normalize()
- Request information about the launches from SpaceX API using custom functions
- Create dictionary from the data
- Create dataframe from the dictionary
- Filter dataframe to contain only Falcon 9 launches
- Replace missing values of Payload Mass with calculated .mean()
- Export data to csv file

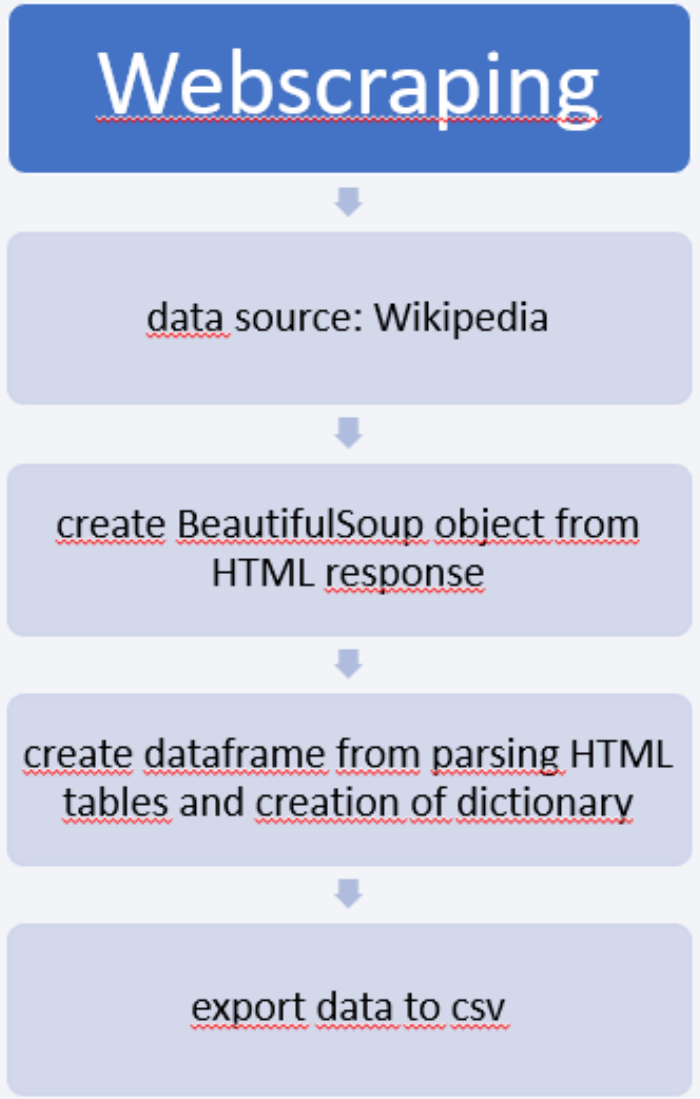
GitHub link to Jupyter Notebook:

https://github.com/rennerak/winning-space-race-with-data-science/blob/d5dd7cb7cef438964bd2d13cee47adde97f955ff/1_Data_collection_with_API.ipynb



Data Collection - Scraping

- Request data (Falcon 9 launch data) from Wikipedia
- Create BeautifulSoup object from HTML response
- Extract column names from HTML table header
- Collect data from parsing HTML tables
- Create dictionary from the data
- Create dataframe from the dictionary
- Export data to csv file



https://github.com/rennerak/winning-space-race-with-data-science/blob/d5dd7cb7cef438964bd2d13cee47adde97f955ff/2_Webscraping.ipynb

Data Wrangling

- Exploratory data analysis:
  - Check null values and data types
  - Calculate number of launches on each site and occurrence of associated orbits
  - Calculate mission outcome dependent on orbit type

- Create binary training labels as dependent variable:
  - Create label from outcome column
    - 1 = good outcome
    - 0 = bad outcome

EDA with Data Visualization

- Charts:
  - Flight Number vs Payload (scatter category plot)
  - Flight Number vs Launch Site (box plot)
  - Payload Mass (kg) vs Launch Site (box plot)
  - Payload Mass (kg) vs Orbit (scatter category plot)

Rationale:

Scatter plots to investigate relationship between two numerical variables

Boxplots to investigate trends of discrete categories like launch site

https://github.com/rennerak/winning-space-race-with-data-science/blob/d5dd7cb7cef438964bd2d13cee47adde97f955ff/5_EDA_with_Visualizations.ipynb

EDA with SQL

Summary of SQL Queries:

- Display unique launch sites
- Display total payload carried by rockets dependent on launch site
- List date when first successful landing was achieved on ground pad
- List the total number of successful and failure mission outcomes

Performed queries and subqueries in sqllite with sqlalchemy

Build an Interactive Map with Folium

Markers:

- Used for characterizing launch sites:
  - **Circle** to denote NASA Johnson Space Center's coordinate as well as other launch sites
  - **Colored markers** to indicate successful (green) and unsuccessful (red) launches at each launch site to show which launch sites have the highest success rate
  - **Lines:** to show the distance between several points of interest e.g. distance to highways, railways, coasts and city

https://github.com/rennerak/winning-space-race-with-data-science/blob/d5dd7cb7cef438964bd2d13cee47adde97f955ff/6_%20LaunchSite_with_Folium.ipynb

Build a Dashboard with Plotly Dash

Dropdown List with lauch sites:

- Allow user to select all launch sites or a distinct launch sites

Pie Chart showing successful launches

- Allow users to see proportional rate of successful launches

Slider of payload mass range

- Allow user to select payload mass range to study the influence of this factor on success rate

Scatter plot showing payload mass vs success rate by booster version

Predictive Analysis (Classification)

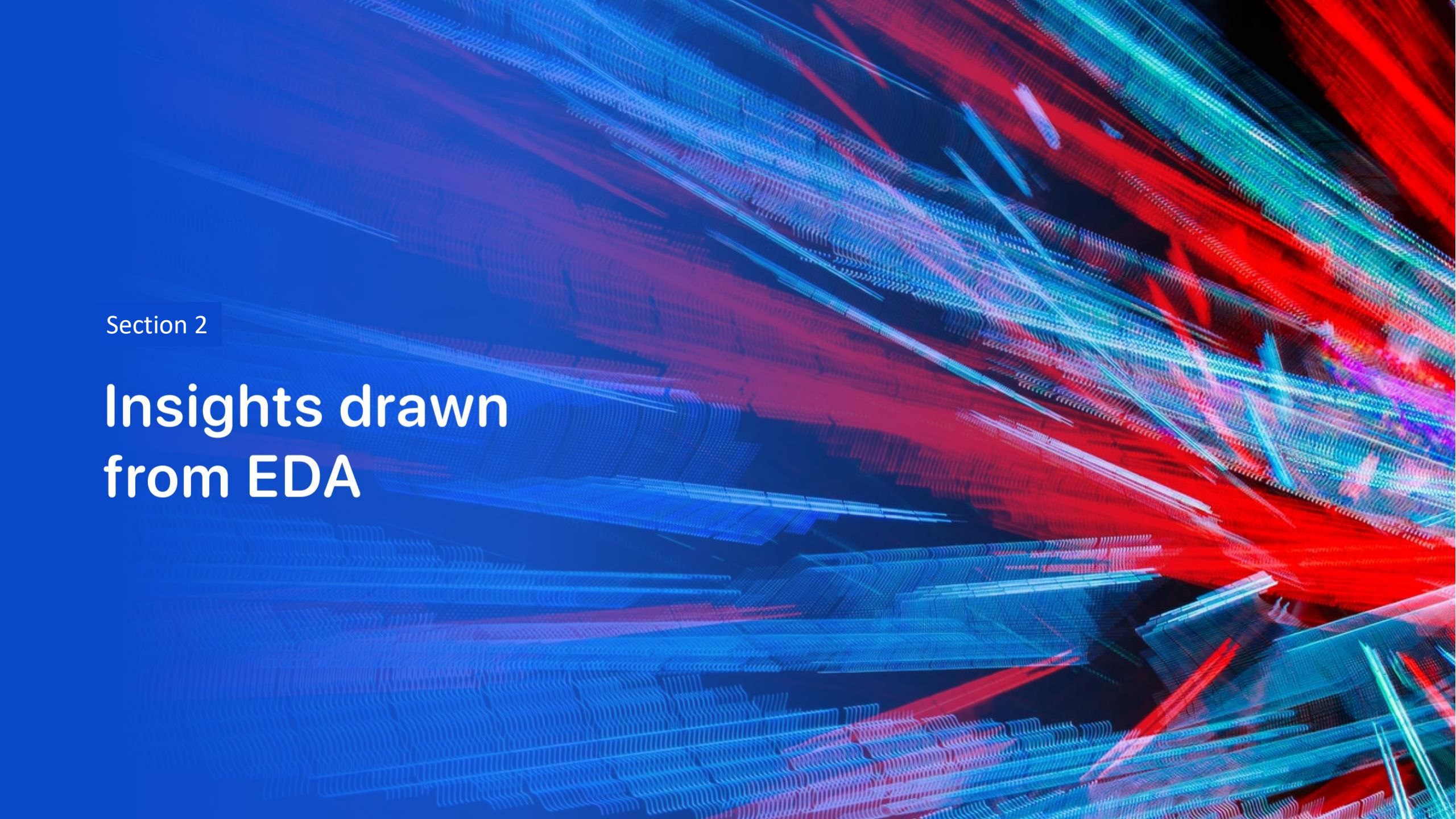
Step process:

1. Create NumPy array from class column = dependent variable
2. Standardize and transform the data
3. Split the data into test and training set using scikit-learn‘s traintestsplit() function
4. Apply GridSearchCV on different machine learning algoritms:
    1. Logistic regression
    2. Suppert vector machine
    3. Decision tree
    4. K-nearest neighbor
5. Calculate accuracy

https://github.com/rennerak/winning-space-race-with-data-science/blob/d5dd7cb7cef438964bd2d13cee47adde97f955ff/7_SpaceX_Machine_Learning_Prediction.ipynb

Results

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


Section 2
Insights drawn from EDA

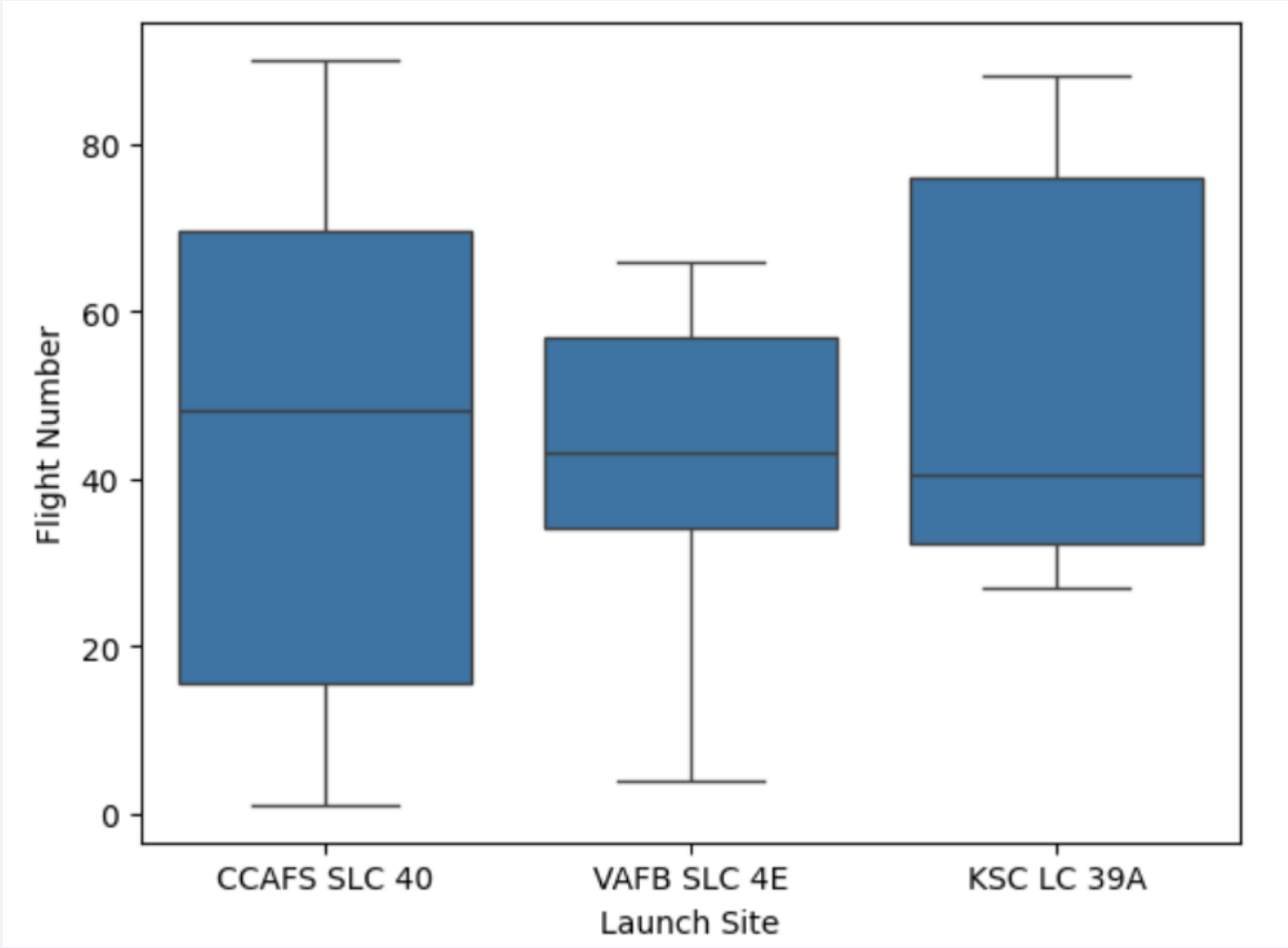
Flight Number vs. Launch Site

Description:

- boxplot shows central tendency of success rate at several investigated launch sites

Conclusion:

launching success increased over time → we can hypothesise that new launches have a higher success rate



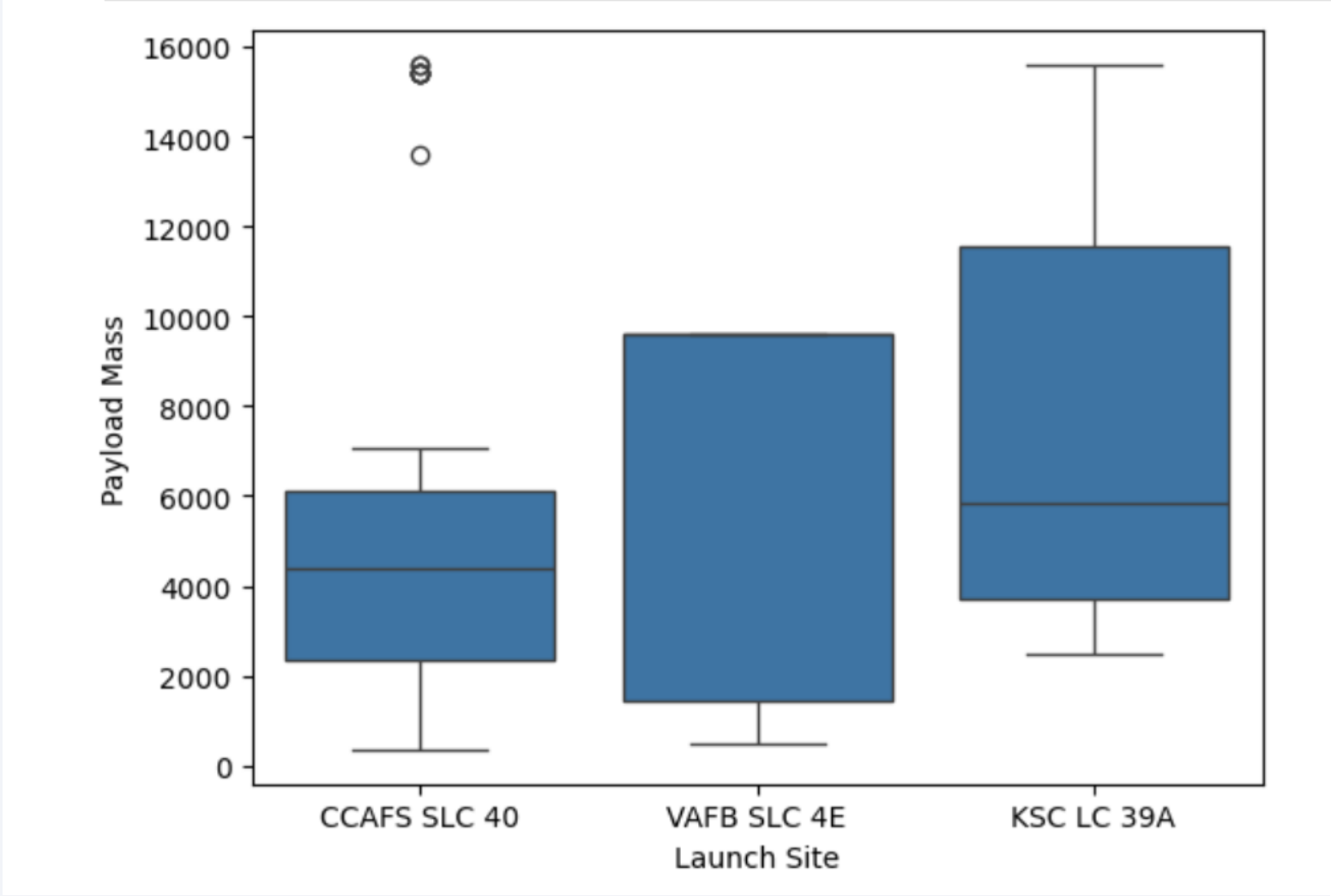
Payload vs. Launch Site

Description:

- boxplot shows central tendency of success rate at several investigated launch sites

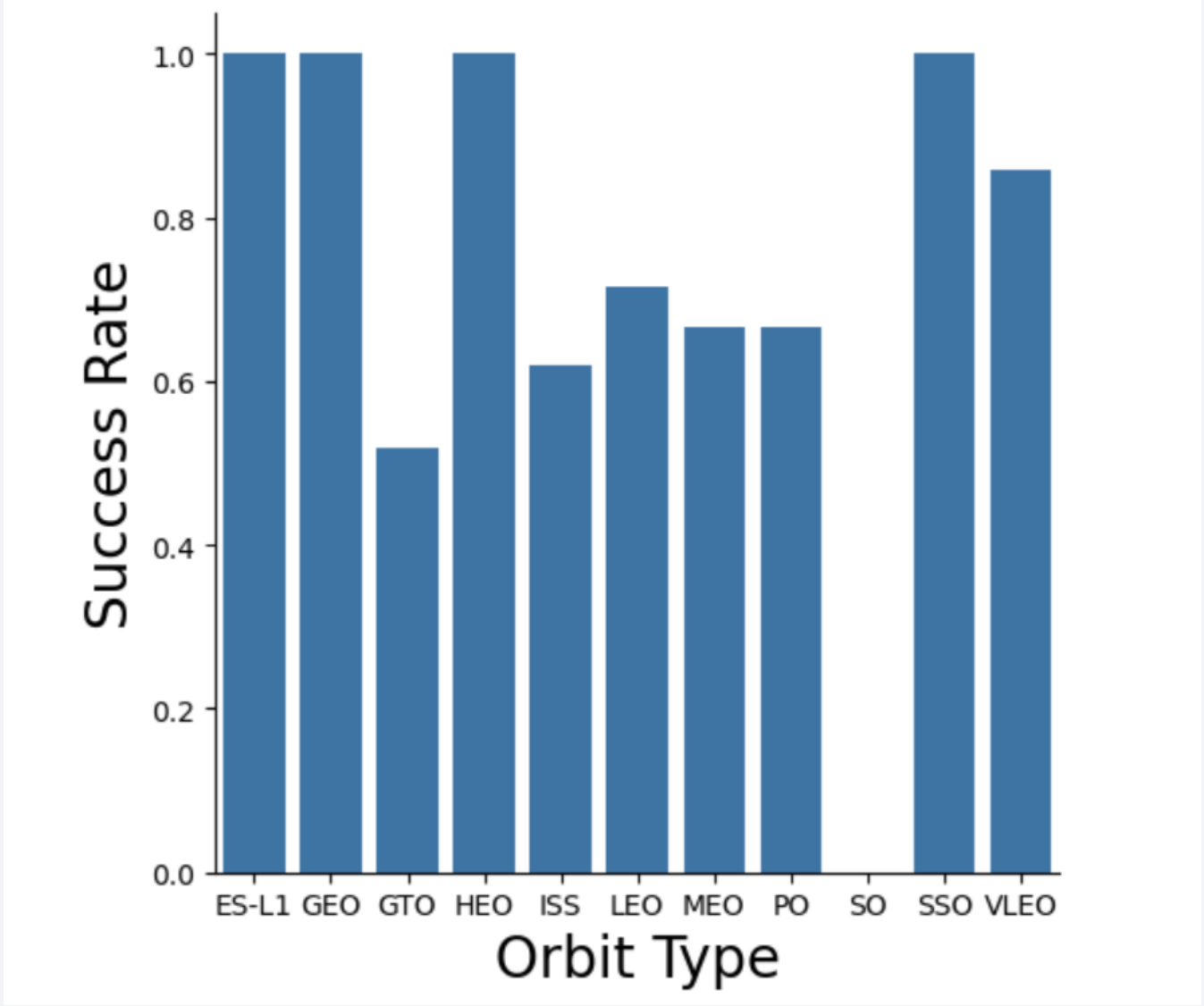
Conclusion:

Carried payload differed depending on launching site → KSC LC 39A loads on average the highest payload mass into space



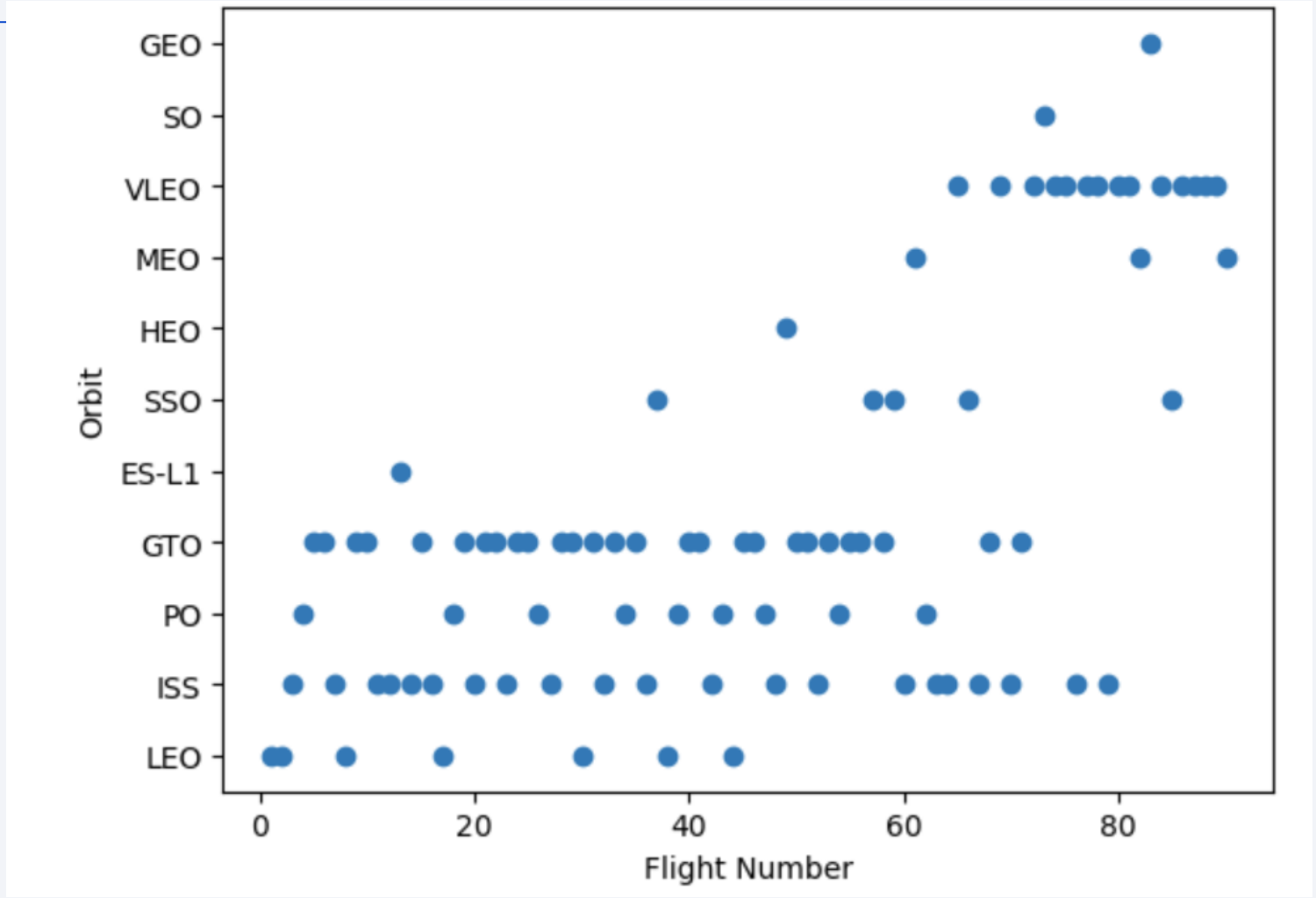
Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO and SSO have a 100 % success rate in contrast to rest of studied orbits



Flight Number vs. Orbit Type

Scatter plot showing the dependency of flight number on orbit types

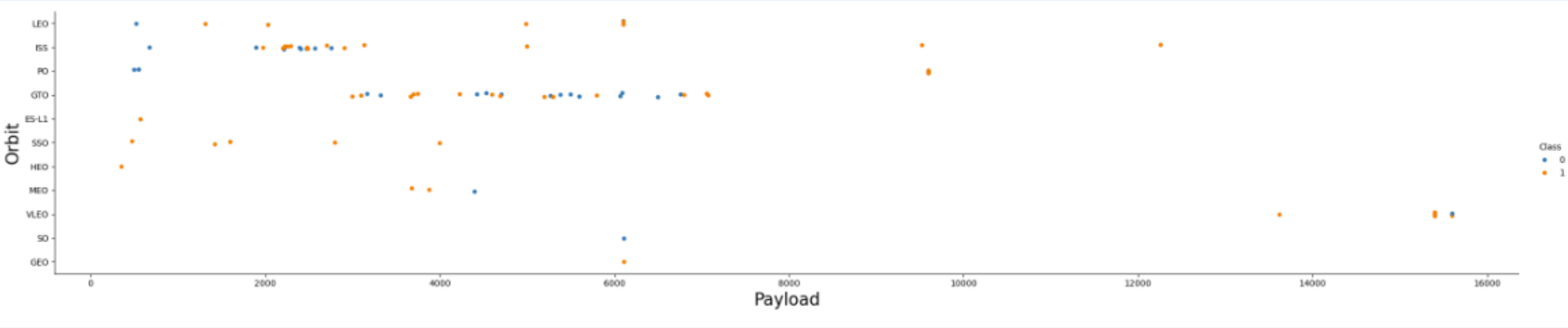


Payload vs. Orbit Type

- Description:
- Blue = fail
- Orange = success

Conclusion:

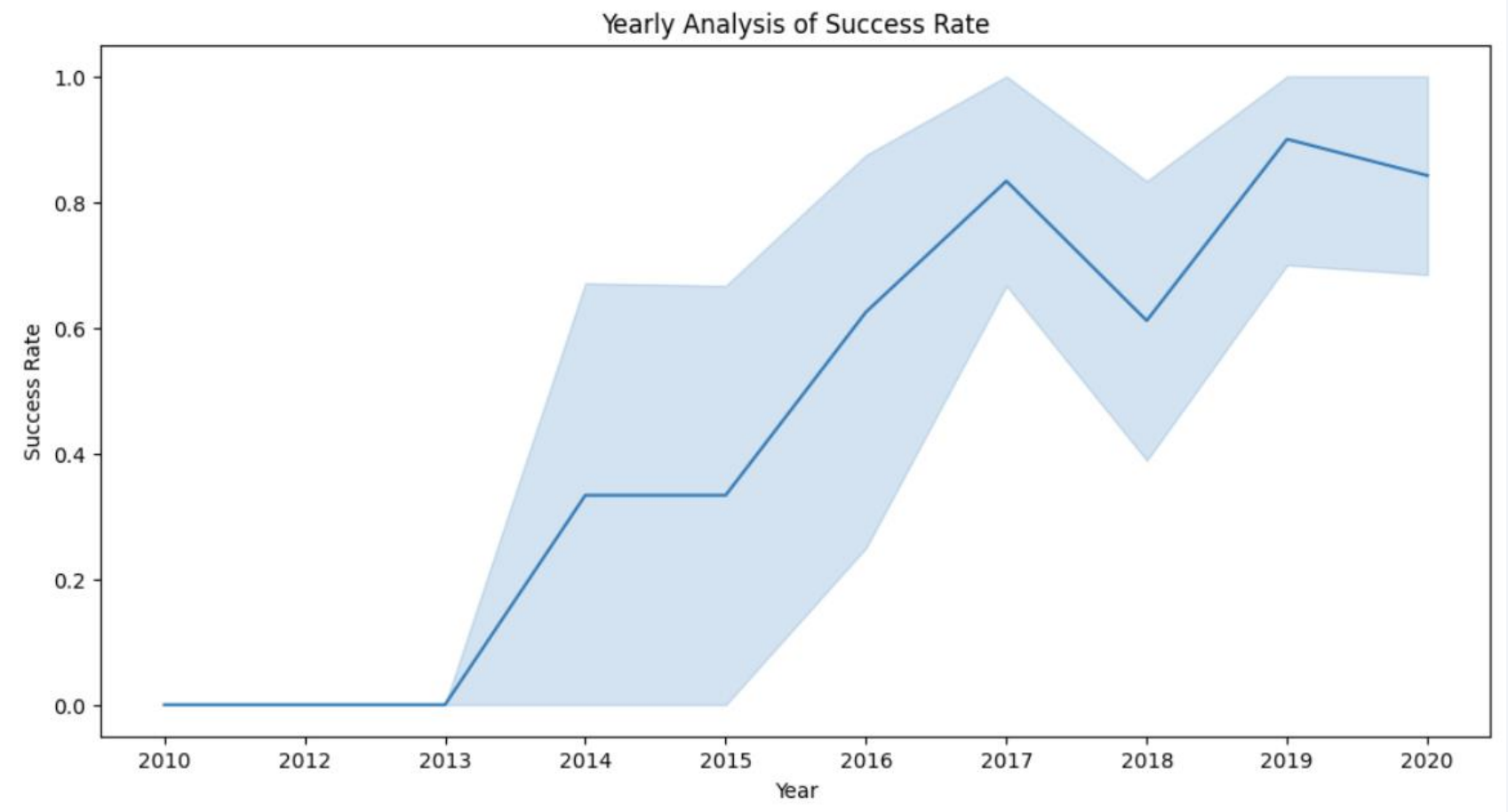
The higher the payload the higher the success rate



Launch Success Yearly Trend

Line chart depicting the yearly analysis of success rate

→ The success rate increased until 2020



All Launch Site Names

- SLQ query was used to list all unique launch sites
- 4 launch sites were used for Falcon 9 rocket launches as listed on the right

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

Launch Site Names Begin with 'CCA'

Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing
18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure
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
07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	
00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	
15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass

- The total carried payload mass is 45,596 kg

In [22]:
```
%%sql
SELECT SUM("PAYLOAD_MASS__KG_") AS Total_Payload
FROM SPACEXTABLE
WHERE Customer = "NASA (CRS)";
```

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

Out[22]:

Total_Payload
45596

Average Payload Mass by F9 v1.1

- The average payload for this respective booster version is 2534.7 kg

Out[24]:

average_payload	Booster_Version
2534.6666666666665	F9 v1.1 B1003

First Successful Ground Landing Date

- The first successful ground landing data was 12/22/2015.

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The successful drone ship landing with a payload between 4000 and 6000 is achieved by the following booster versions:

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

Total Number of Successful and Failure Mission Outcomes

- The dataframe shows the total number of successful and failed mission outcomes

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

Boosters Carried Maximum Payload

- From investigated F9 rocket only rockets with booster version B5 … carried the maximum payload

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

month	Date	Booster_Version	Launch_Site	Landing_Outcome
10	2015-10-01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Most success outcomes were reported on the ground pad

Landing _Outcome	count_outcomes
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

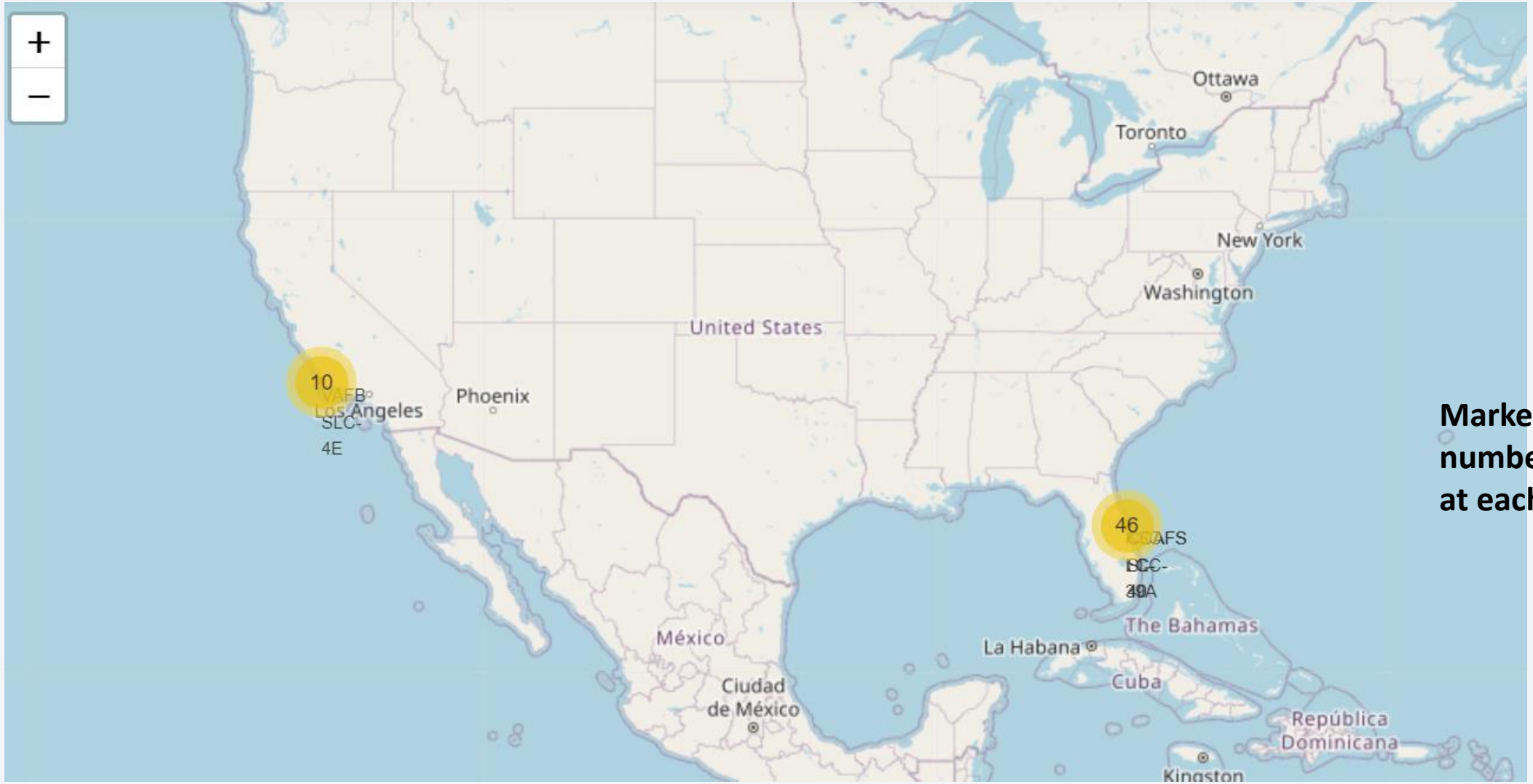
Section 3

Launch Sites Proximities Analysis

Distribution of launching sites in the US

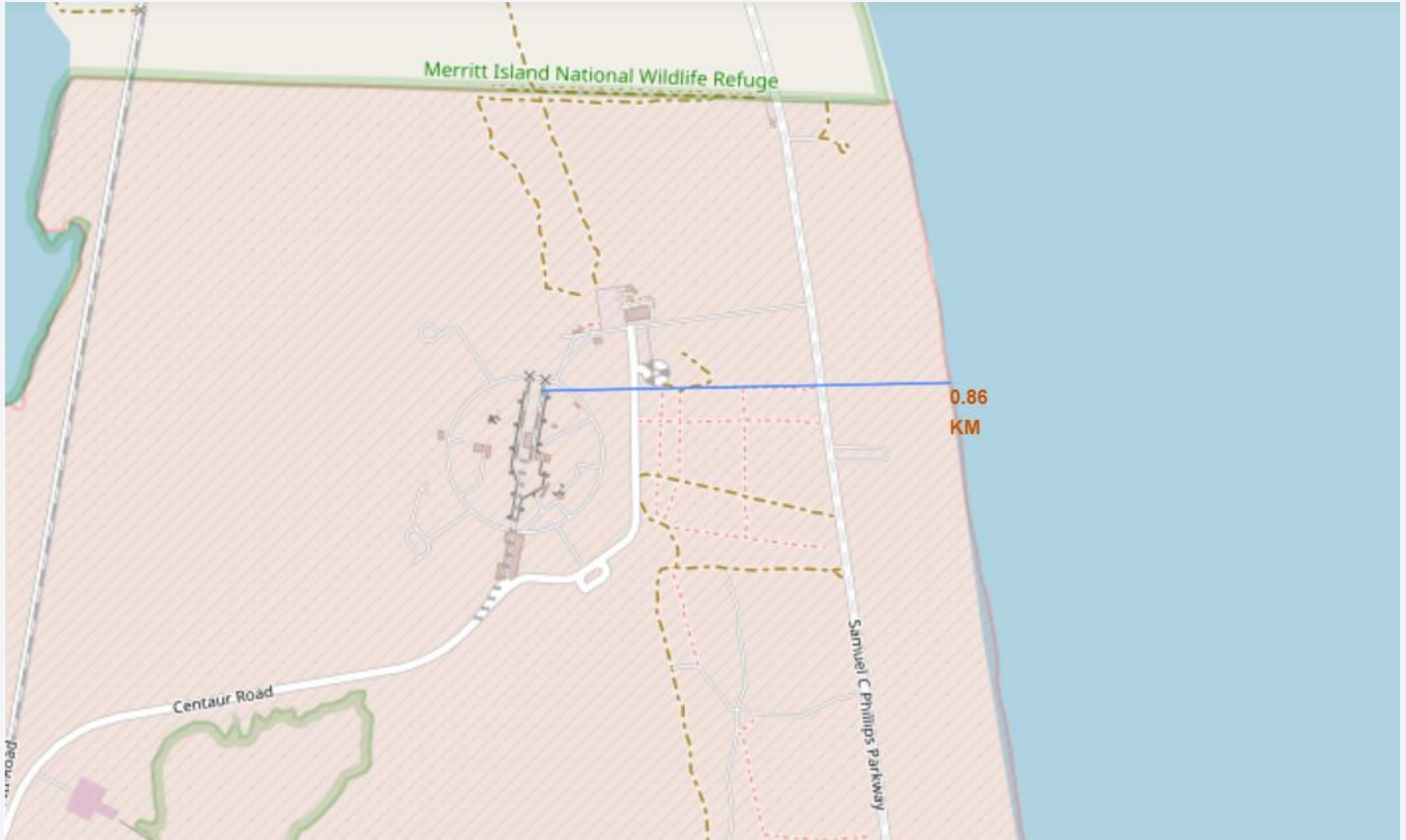
Launch sites are located near the coast in the US

Number of launches at each site

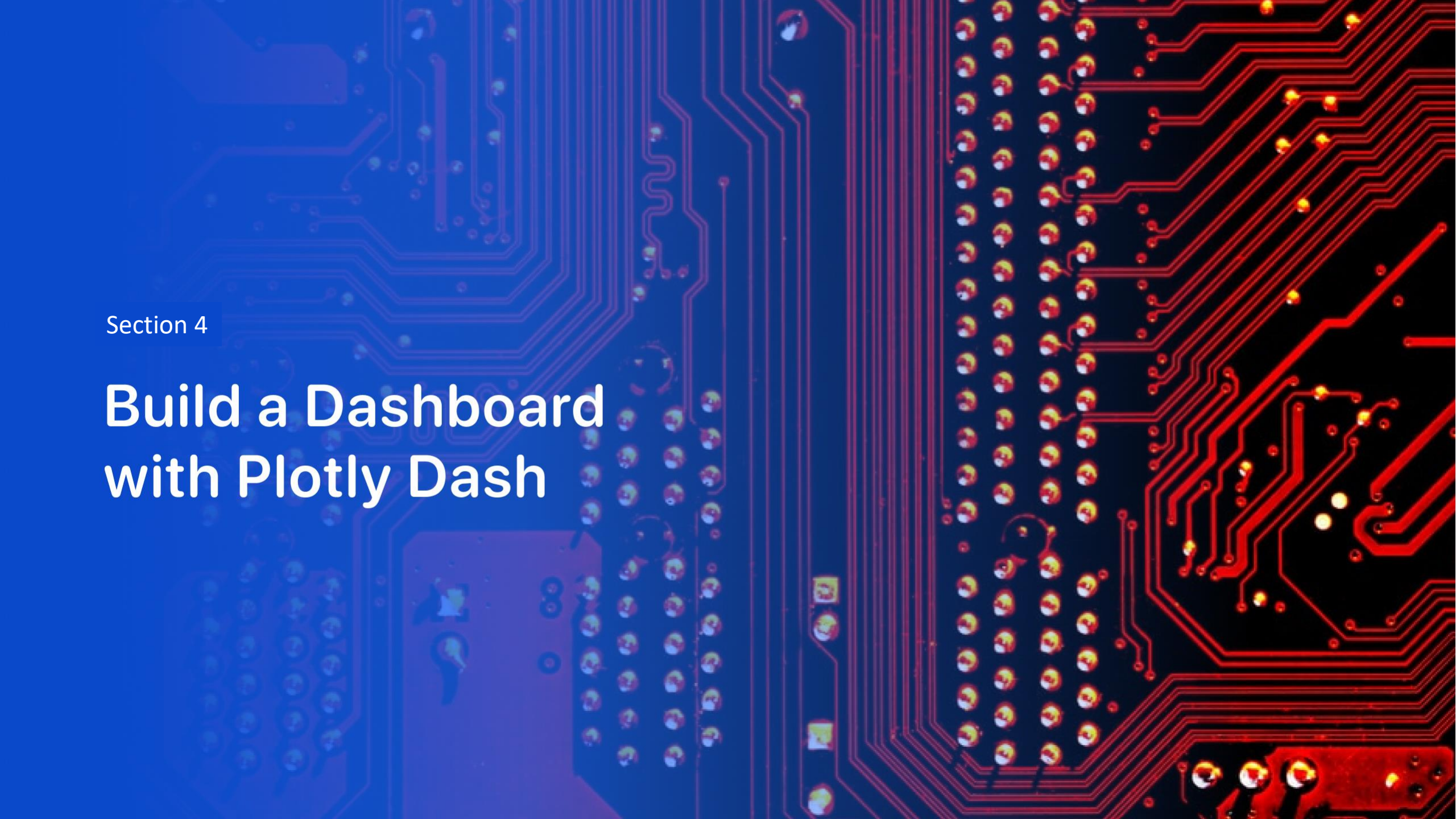


Markers depict number of launches at each site

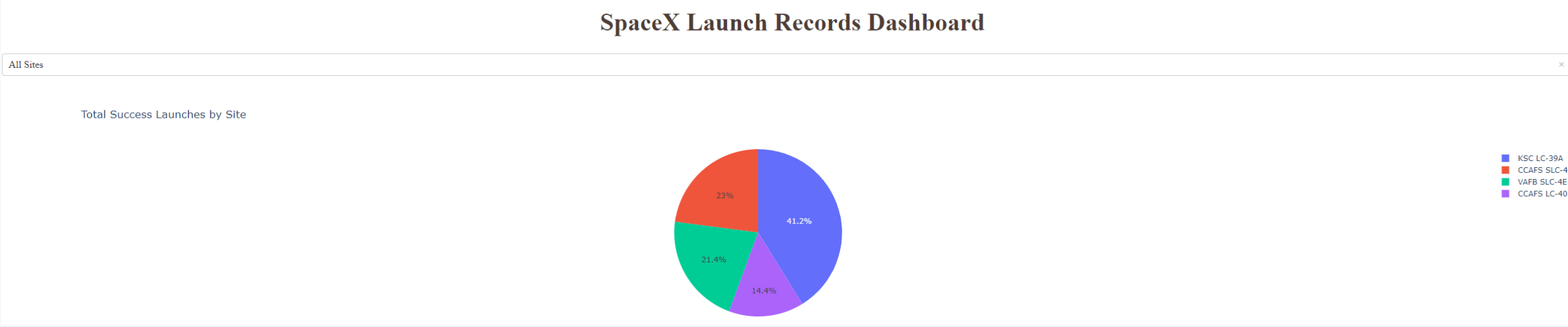
Distance of Launch Site to Coast



Launch sites are located near the coastal area

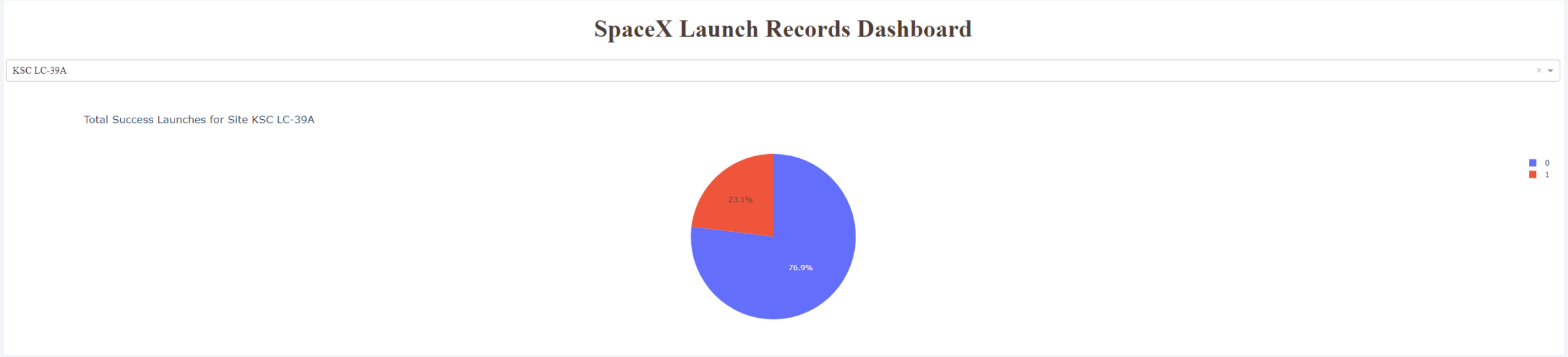

Section 4
Build a Dashboard with Plotly Dash

Launch Success Evaluation by Site



KSC LC 39 A shows the highest success rate of launches

Success Rate of Launch Site KSC LC 39A



Influence of Payload on Success Rate

Description:
Color coding for booster version

Conclusion:
We can see that FT carried a high diversity of payloads with several mission outcomes

V1.0 only carried minimal payloads

Section 5

Predictive Analysis (Classification)

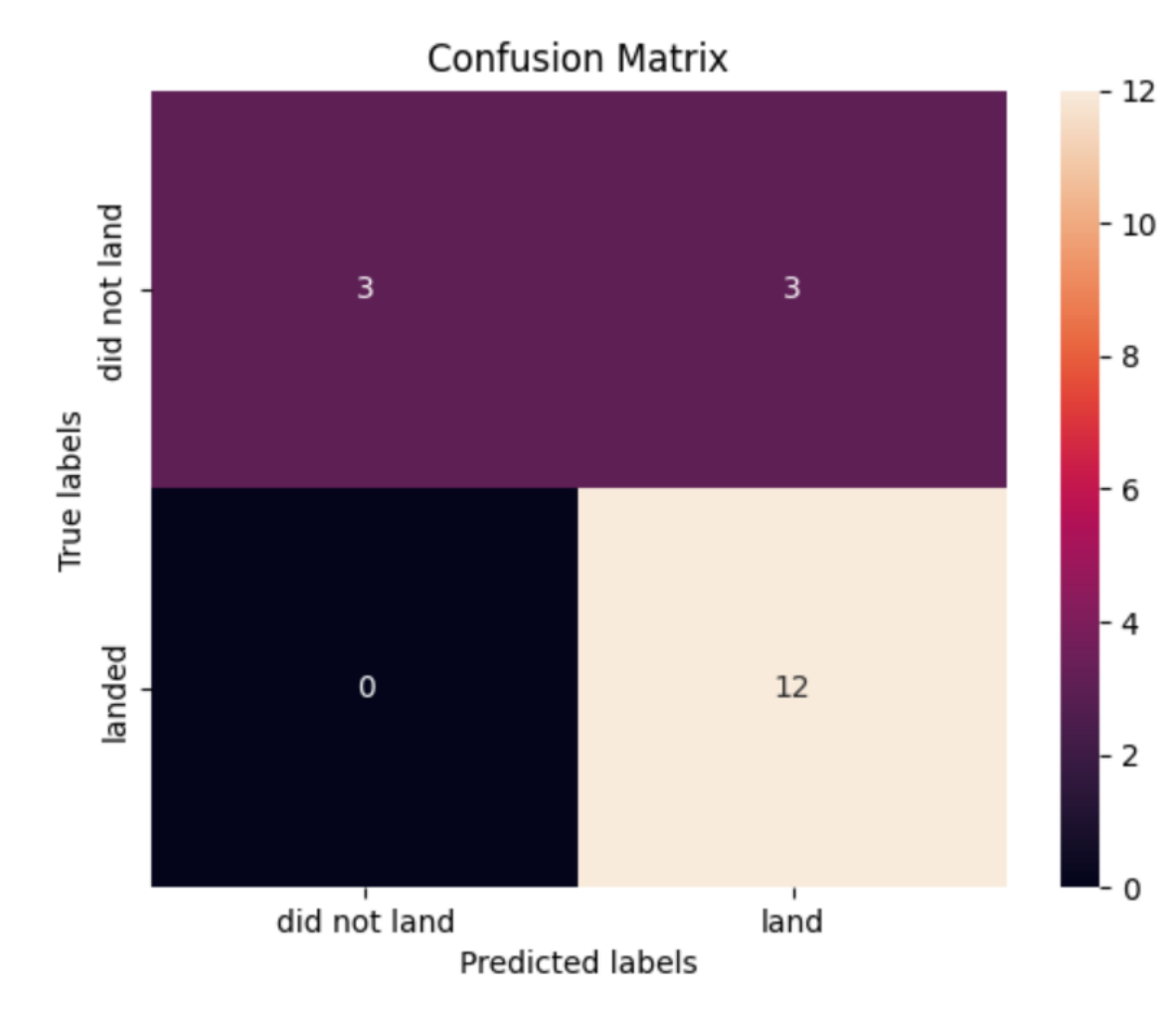
Classification Accuracy

	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	83.333333
3	Decision Tree	83.333333

All evaluated models showed similar accuracy thus no statement can be made which method is best suitable for classification

Confusion Matrix

- The confusion matrix is the derived from KNN classification
- We can see that this model has high accuracy with detecting true positives but performs less well in detection of true negatives since 50 % fall into false negatives for no landing



Conclusions

To predict the cost of rocket launches, this project aimed to evaluate the success rate of first stage landing

Conclusions:

- Rocket launch success rate increased over time
- Orbits ES-L1, GEO, HEO and SSO have a 100 % success rate in contrast to rest of studied orbits
- Most launches are located near the coast
- All evaluated classification algorithms showed an accuracy score of 83 %

Future work

- In order to improve the model accuracy, more classification models should be evaluated
- Since there was a decline in success rate of rocket launches after 2020, it will be interesting to repeat this analysis with an up-to-date dataset

Thank you!