IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ayham Shaheen
January 20, 2025

OUTLINE

Executive Summary

Introduction

Methodology

Results

Conclusion

Appendix

Executive Summary

In this capstone project, we will predict if the SpaceX Falcon 9 first stage will land successfully using several machine learning classification algorithms.

The main steps in this project include:

- Data collection, wrangling, and formatting
- Exploratory data analysis
- Interactive data visualization
- Machine learning prediction

Our graphs show that some features of the rocket launches have a correlation with the outcome of the launches, i.e., success or failure.

It is also concluded that decision tree may be the best machine learning algorithm to predict if the Falcon 9 first stage will land successfully.

Introduction

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.

The main question that we are trying to answer is, for a given set of features about a Falcon 9 rocket launch which include its payload mass, orbit type, launch site, and so on, will the first stage of the rocket land successfully?

Section 1

Methodology

Methodology

- Data collection
  - SpaceX API
  - Web scraping
- Data Wrangling
  - Exploratory Data Analysis
  - Determine Training Labels
- Exploratory Data Analysis (EDA)
  - Pandas and NumPy
  - SQL
- Data visualization
  - Matplotlib and Seaborn
  - Folium
  - Dash
- Machine learning prediction
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree
  - K-nearest neighbors (KNN)

Data Collection – SpaceX API

- request rocket launch data from SpaceX API with the following URL:
  spacex_url="https://api.spacexdata.com/v4/launches/past"
- The API provides data about many types of rocket launches done by SpaceX, the data is therefore filtered to include only Falcon 9 launches.
- Every missing value in the data is replaced the mean the column that the missing value belongs to.
- We end up with 90 rows or instances and 17 columns or features.

Data Collection – SpaceX API

The picture below shows the first few rows of the data:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

GitHub URL:

https://github.com/AyhamShaheen/IBM-Data-Science-Capstone/blob/main/1-%20Spacex%20Data%20Collection%20API.ipynb

Data Collection – Web Scraping

- The data is scraped from: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- The website contains only the data about Falcon 9 launches.
- We end up with 121 rows or instances and 11 columns or features.
- The picture below shows the first few rows of the data:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Data Collection – Web Scraping

GitHub URL:

https://github.com/AyhamShaheen/IBM-Data-Science-Capstone/blob/main/2-%20SpaceX%20Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- The data is later processed so that there are no missing entries, and categorical features are encoded using one-hot encoding.
- An extra column called 'Class' is also added to the data frame. The column 'Class' contains 0 if a given launch is failed and 1 if it is successful.
- at the end, we end up with 90 rows or instances and 83 columns or features.

GitHub URL:

https://github.com/AyhamShaheen/IBM-Data-Science-Capstone/blob/main/3-%20SpaceX%20Data%20wrangling.ipynb

EDA with Data Visualization

Exploring and Preparing Data:

- perform Exploratory Data Analysis and Feature Engineering

Visualize the relationship between Flight Number vs. Payload Mass:

- We see that as the flight number increases, the first stage is more likely to land successfully

Visualize the relationship between Flight Number vs. Launch Site:

- explain the patterns

Visualize the relationship between Payload Mass vs. Launch Site:

- we find for the VAFB-SLC launch Site there are no rockets launched for heavy payload mass(greater than 10000).

Visualize the relationship between success rate of each orbit type:

- we find that ES-L1, GEO, HEO and SSO orbits have the highest success rates

EDA with Data Visualization

Visualize the relationship between FlightNumber and Orbit type:

- We observe that in the LEO orbit, success seems to be related to the number of flights.
- Conversely, in the GTO orbit, there appears to be no relationship between flight number and success

Visualize the relationship between Payload Mass and Orbit type:

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present

Visualize the launch success yearly trend:

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

GitHub URL :

https://github.com/AyhamShaheen/IBM-Data-Science-Capstone/blob/main/4-%20SpaceX%20EDA%20with %20Data%20Viz.ipynb

EDA with SQL

- 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

EDA with SQL

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 landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

GitHub URL :

https://github.com/AyhamShaheen/IBM-Data-Science-Capstone/blob/main/4-%20SpaceX%20EDA%20with %20SQL.ipynb

Build an Interactive Map with Folium

to find some geographical patterns about launch sites

- Mark all launch sites on a map
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities

GitHub URL :

https://github.com/AyhamShaheen/IBM-Data-Science-Capstone/blob/main/6-%20SpaceX%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

launch success count for all sites

the launch site with highest launch success ratio

Payload vs. Launch Outcome for all sites

GitHub URL :

https://github.com/AyhamShaheen/IBM-Data-Science-Capstone/blob/main/6-%20SpaceX%20DashApp.ipynb

Predictive Analysis (Classification)

Functions from the Scikit-learn library are used to create our machine learning models.

The machine learning prediction phase include the following steps:

- Standardizing the data
- Splitting the data into training and test data
- Creating machine learning models, which include:
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree
  - K nearest neighbors (KNN)
- Fit the models on the training set
- Find the best combination of hyperparameters for each model
- Evaluate the models based on their accuracy scores and confusion matrix

Predictive Analysis (Classification)

The results are split into 5 sections:

- SQL (EDA with SQL)
- Matplotlib and Seaborn (EDA with Visualization)
- Folium
- Dash
- Predictive Analysis

In all the graphs that follow, class 0 represents a failed launch outcome while class 1 represents a successful launch outcome.

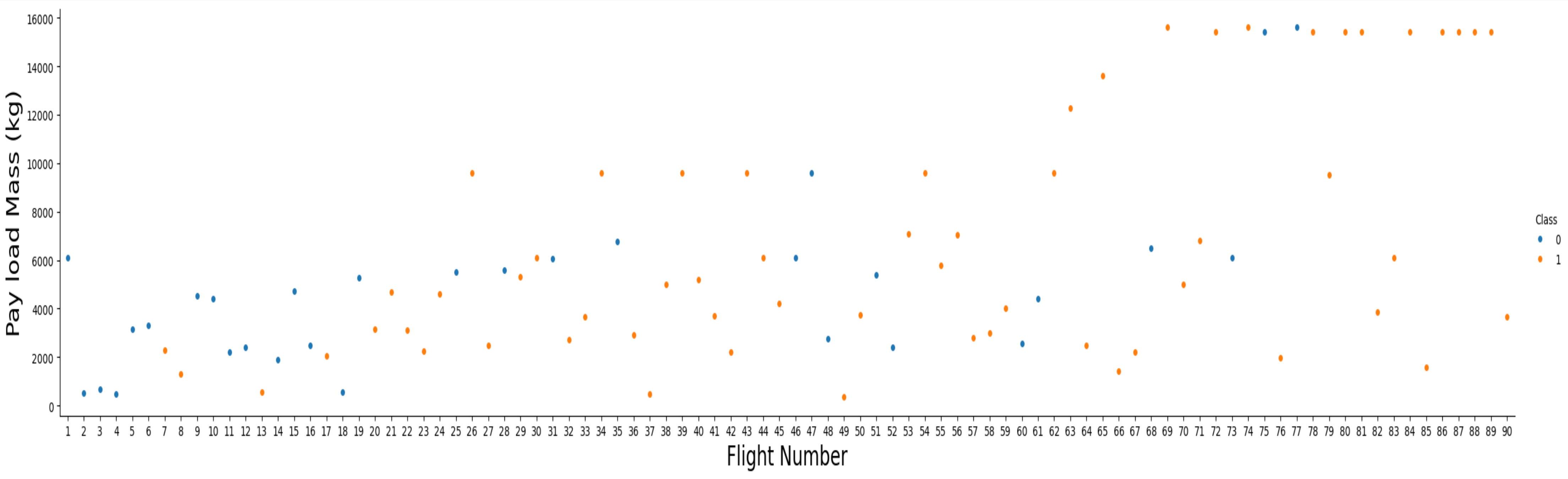
GitHub URL :

https://github.com/AyhamShaheen/IBM-Data-Science-Capstone/blob/main/6-%20SpaceX%20Machine%20Learning%20Prediction.ipynb

Section 2

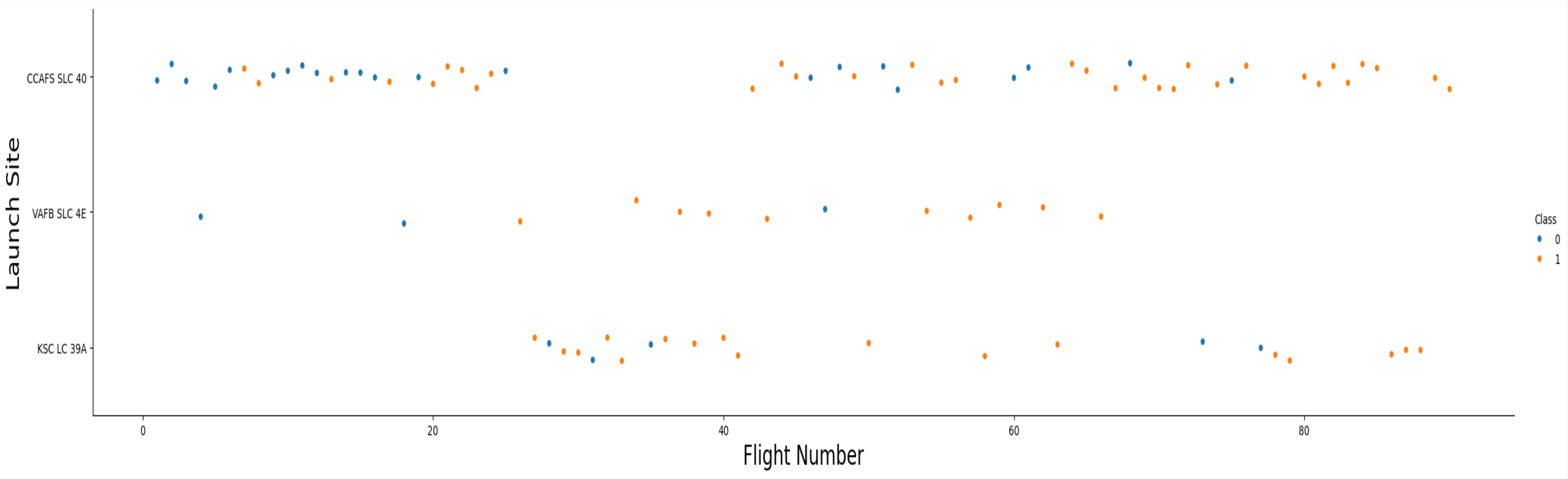
Insights drawn from EDA

Flight Number vs. Payload Mass



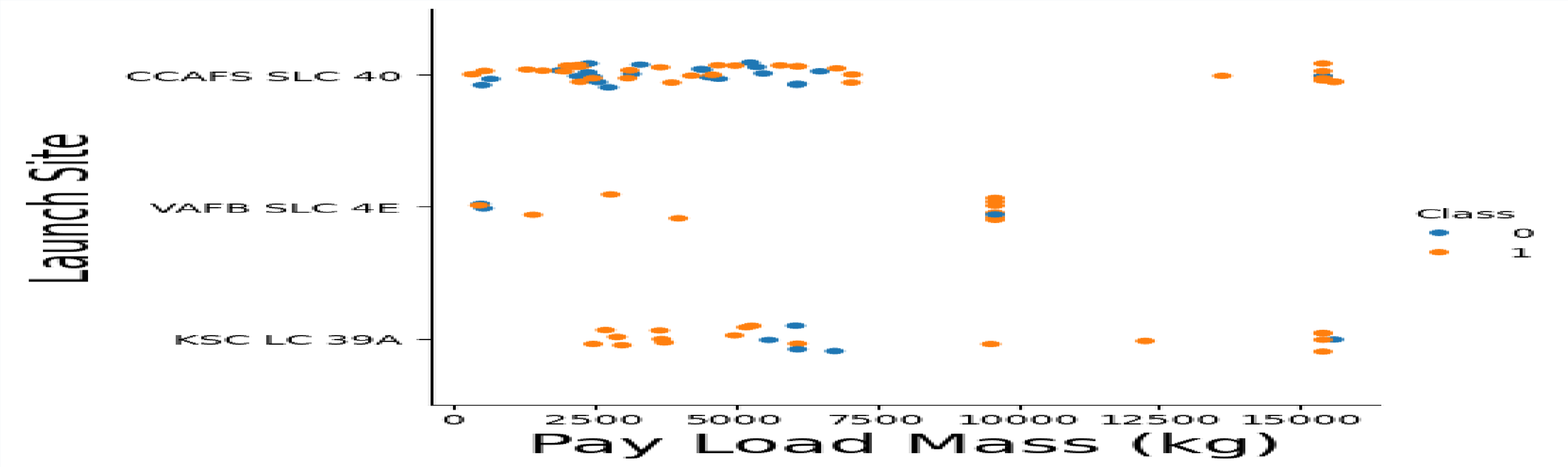
to see how the Flight Number and Payload variables would affect the launch outcome

Flight Number vs. Launch Site



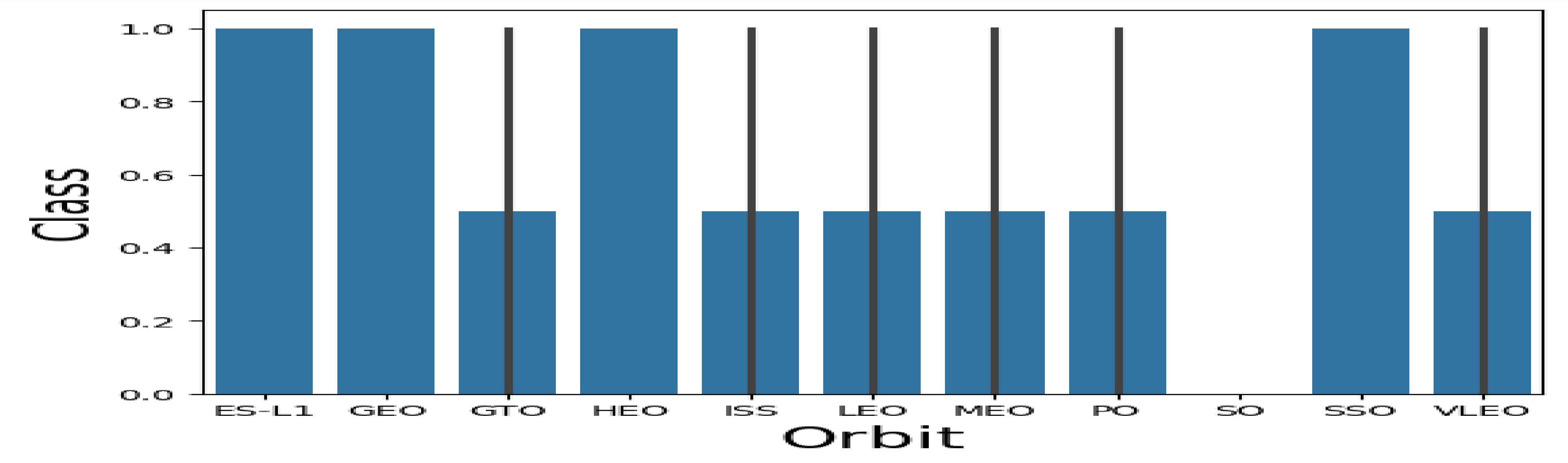
to show relationship between Flight Number and Launch Site

Payload vs. Launch Site



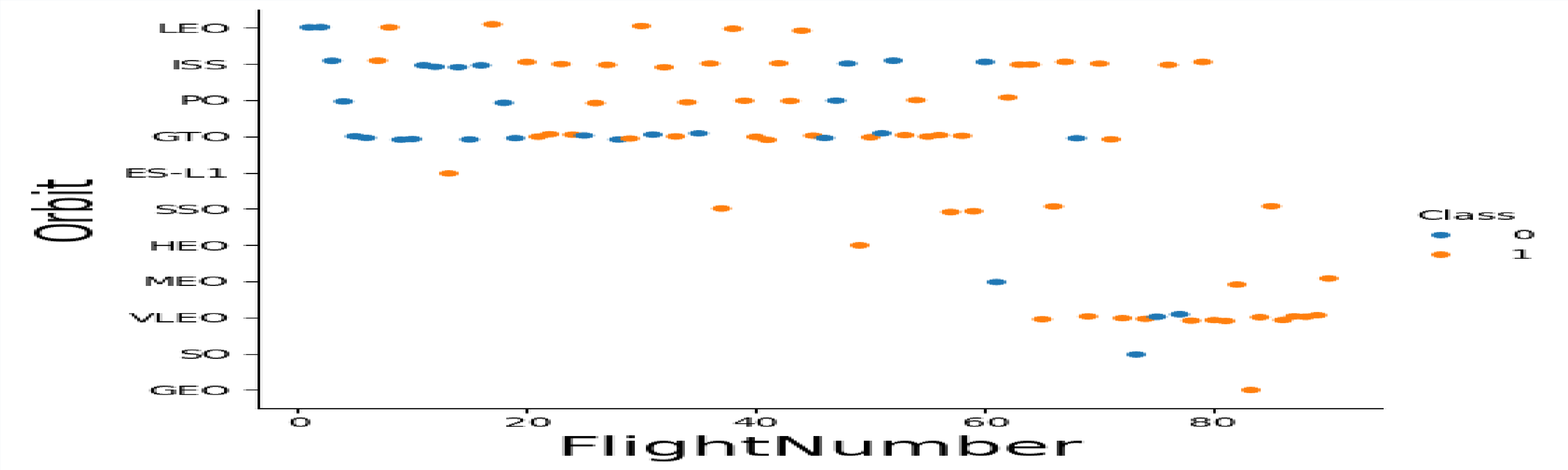
to visually check the relationship between Flight Number and Launch Site

Success Rate vs. Orbit Type



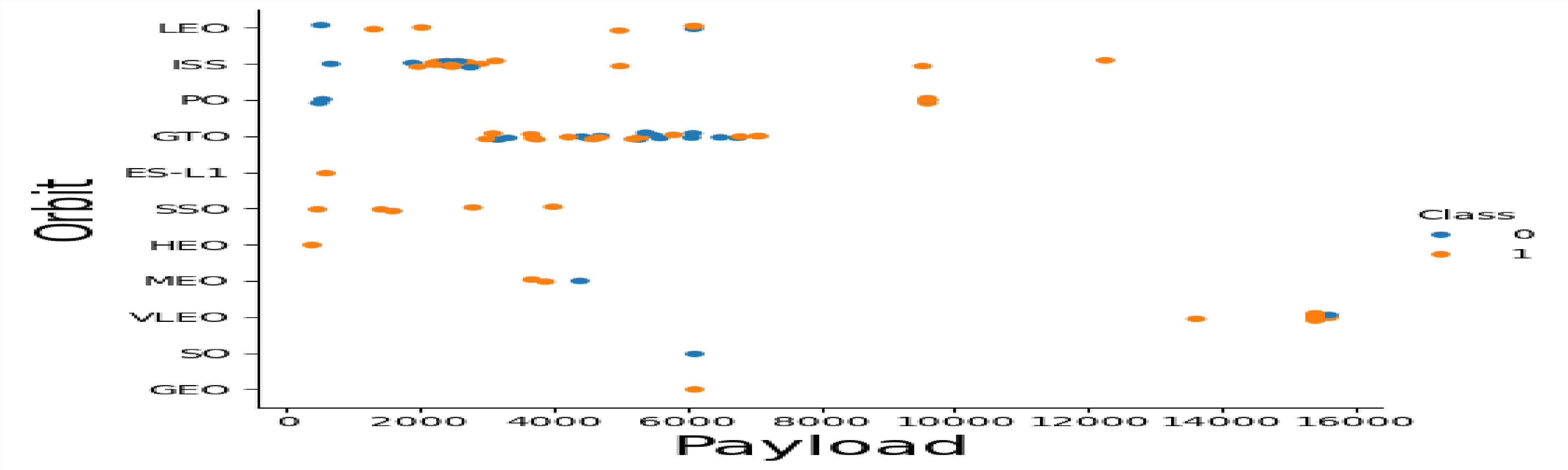
to visually check if there are any relationship between success rate and orbit type

Flight Number vs. Orbit Type



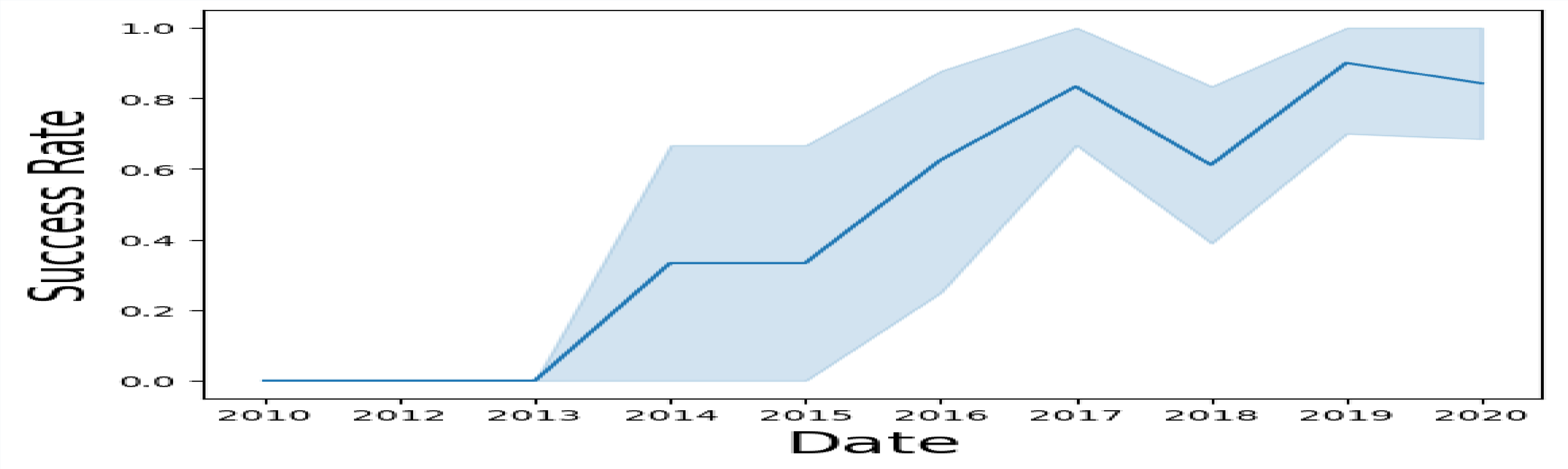
to see if there is any relationship between Flight Number and Orbit type

Payload vs. Orbit Type



to see if there is any relationship between Payload Mass and Orbit type

Launch Success Yearly Trend



to get the average launch success trend

All Launch Site Names

- The names of the unique launch sites in the space mission

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with ‘CCA’

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

Total Payload Mass

- The total payload mass carried by boosters launched by NASA (CRS)

Total payload mass by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1

Average payload mass by Booster Version F9 v1.1
2928

First Successful Ground Landing Date

- The date when the first successful landing outcome in ground pad was achieved

Date of first successful landing outcome in ground pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

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 total number of successful and failure mission outcomes

number_of_success_outcomes	number_of_failure_outcomes
100	1

Boosters Carried Maximum Payload

- the names of the booster which have carried the maximum payload mass

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

2015 Launch Records

- the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

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

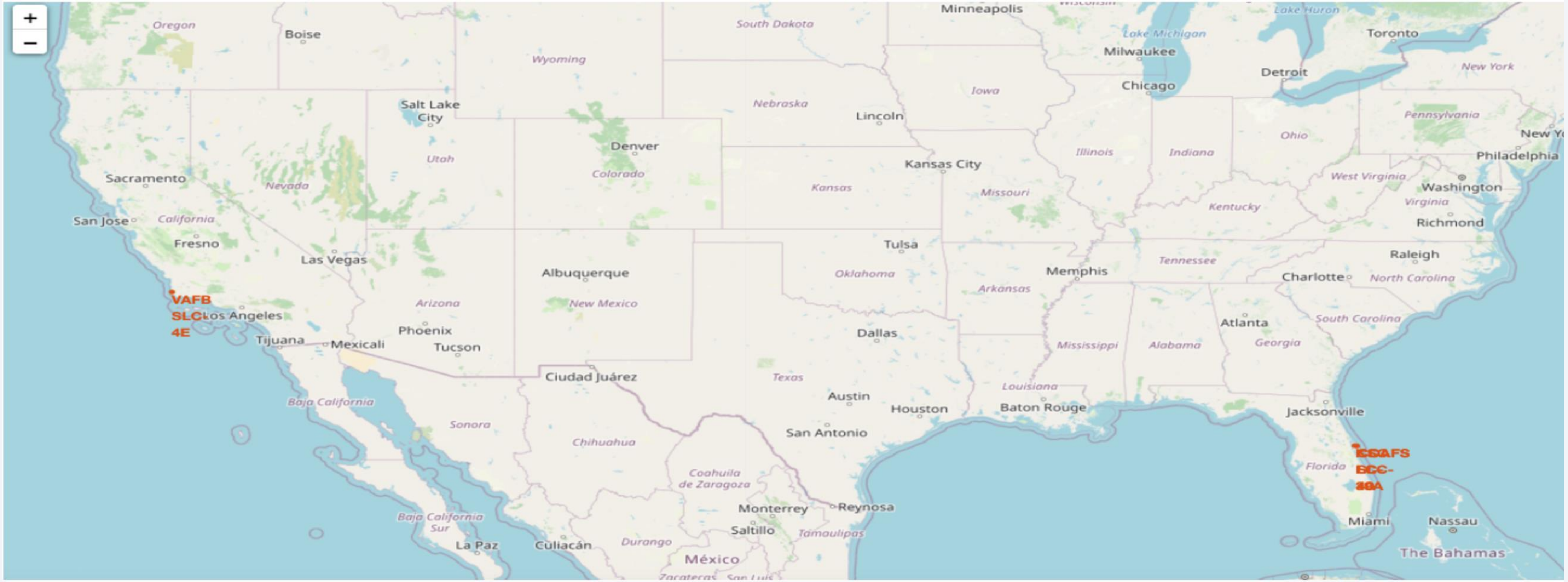
- the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

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

Section 3

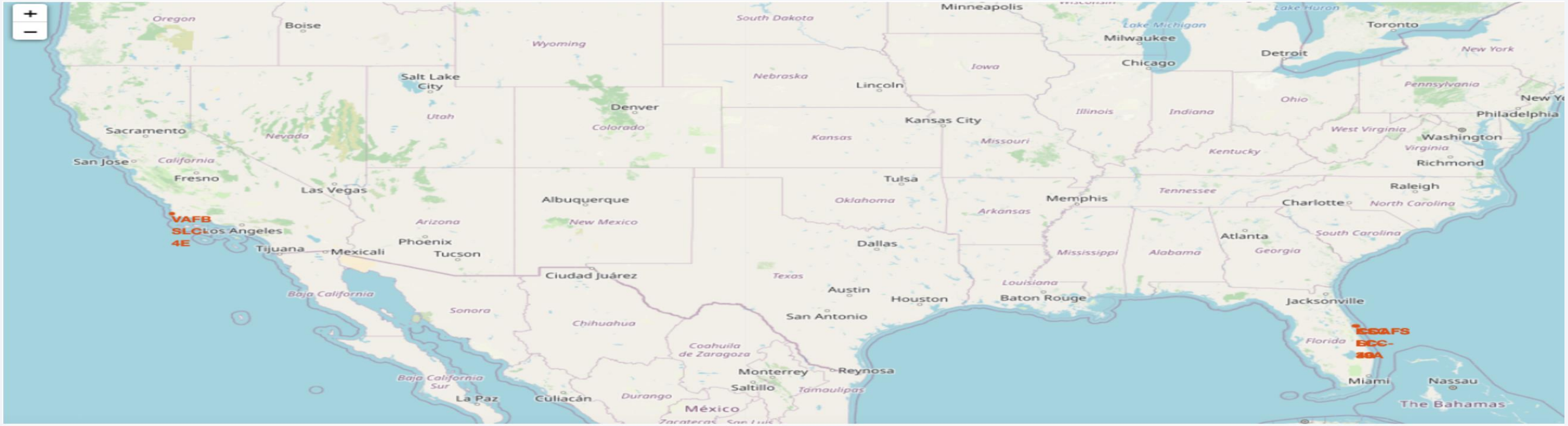
Launch Sites Proximities Analysis

All Launch Sites on Map



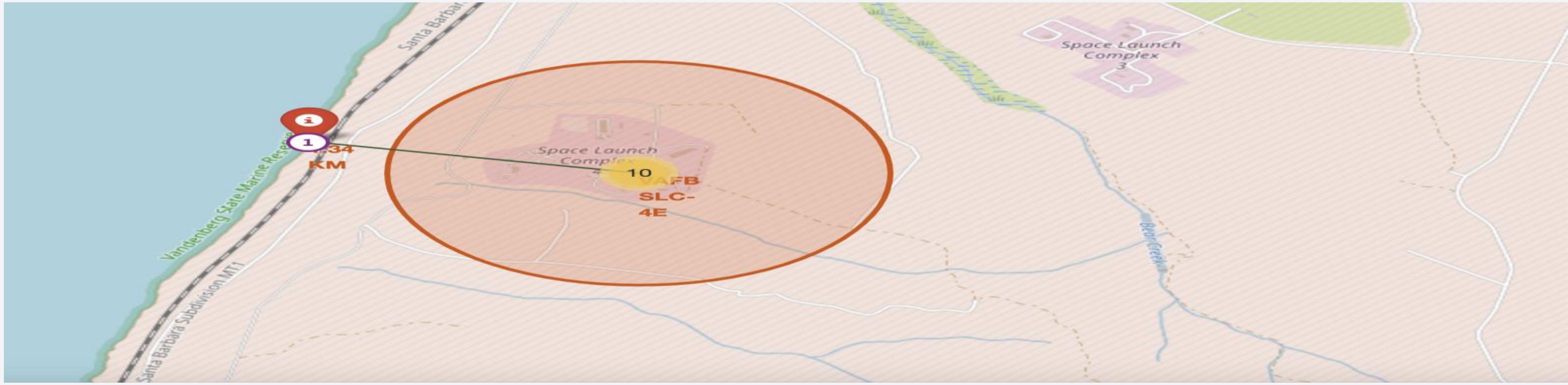
succeeded launches and failed launches for each site on map

- If we zoom in on one of the launch site, we can see green and red tags.
- Each green tag represents a successful launch while each red tag represents a failed launch



The distances between a launch site to its proximities

- The distances between a launch site to its proximities such as the nearest city, railway, or highway
- The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline

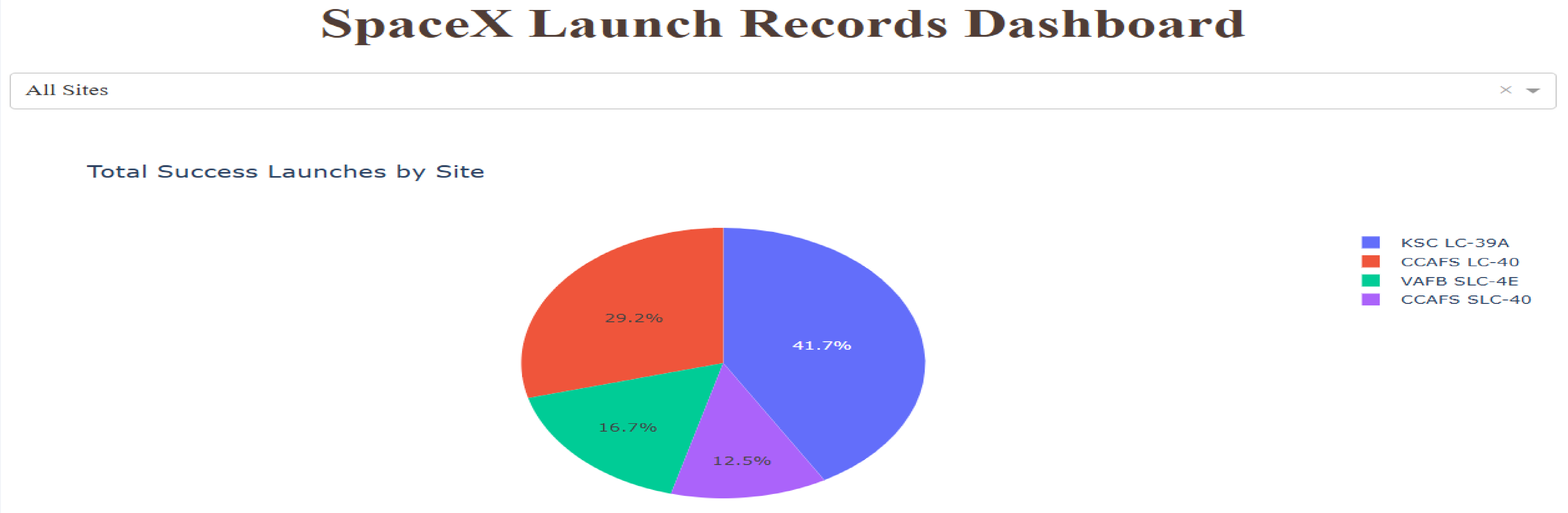


Section 4

Build a Dashboard with Plotly Dash

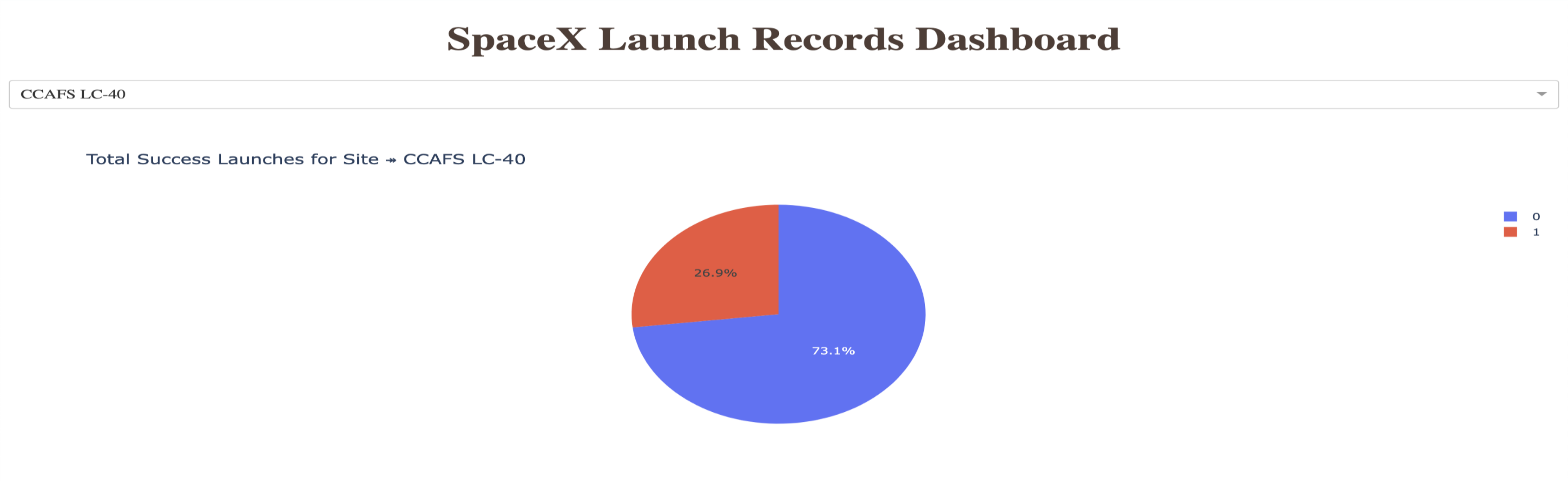
Dashboard

- launch success count for all sites



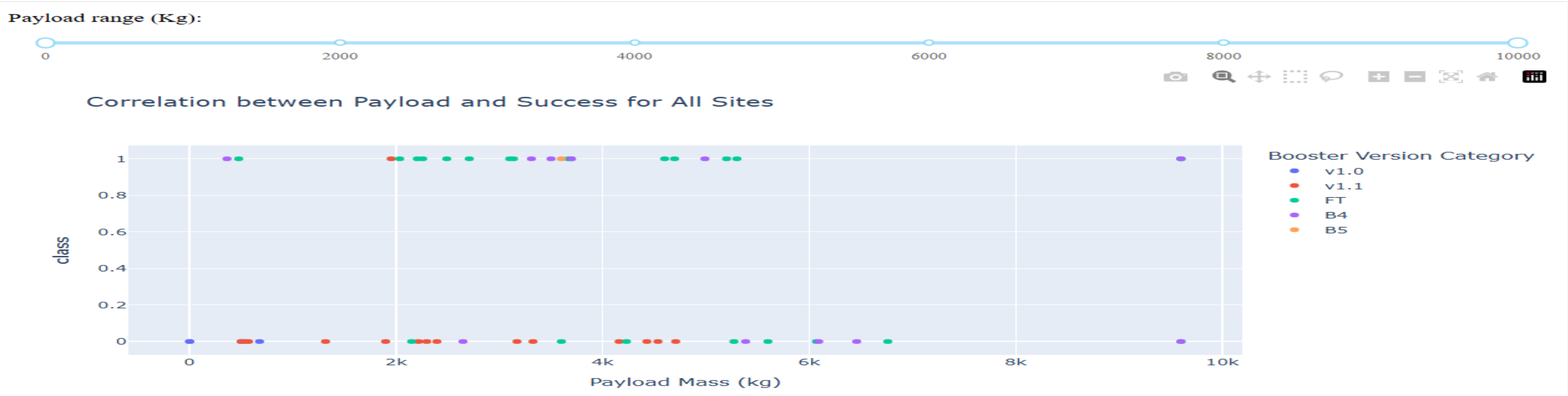
Dashboard

- the launch site with highest launch success ratio **CCAFS LC-40**
- 0 represents failed launches while 1 represents successful launches. We can see that 73.1% of launches done at CCAFS LC-40 are failed launches.



Dashboard

- Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Class 0 represents failed launches while class 1 represents successful launches.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

Putting the results of all 4 models' side by side, we can see that they all share the same accuracy score and confusion matrix when tested on the test set.

Therefore, their GridSearchCV best scores are used to rank them instead.

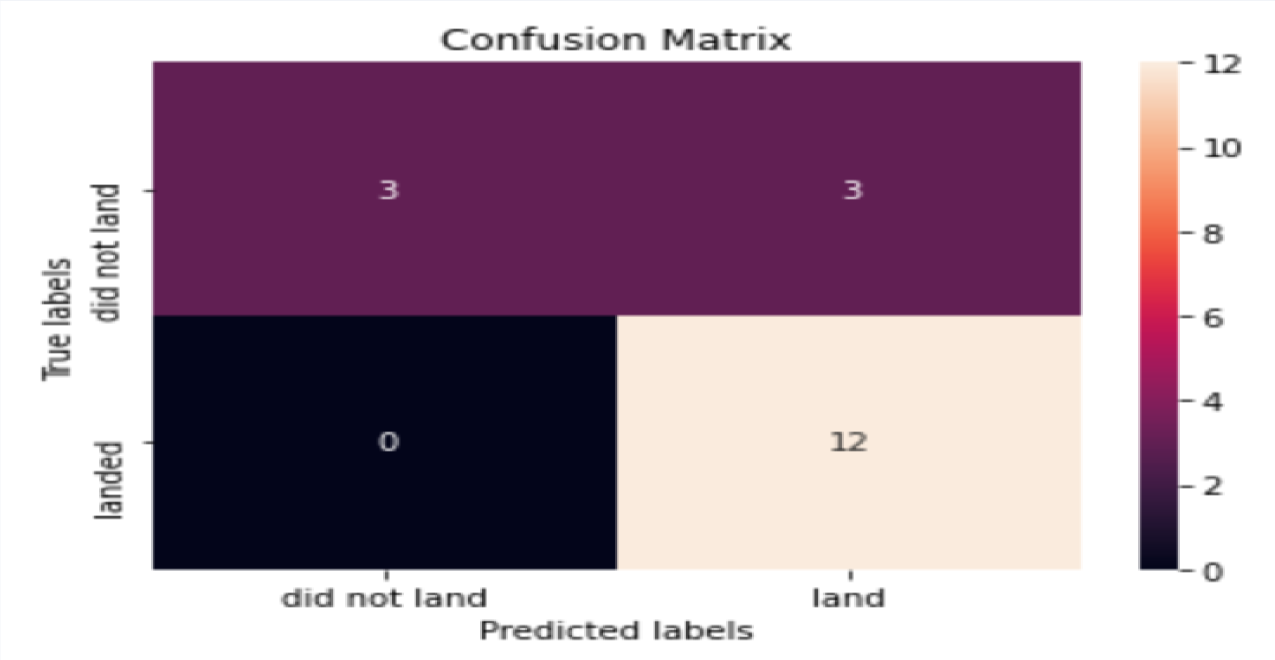
Based on the GridSearchCV best scores, the models are ranked in the following order with the first being the best and the last one being the worst:

- Decision tree (GridSearchCV best score: 0.8892857142857142)
- K nearest neighbors, KNN (GridSearchCV best score: 0.8482142857142858)
- Support vector machine, SVM (GridSearchCV best score: 0.8482142857142856)
- Logistic regression (GridSearchCV best score: 0.8464285714285713)

Confusion Matrix

Decision tree

- GridSearchCV best score: 0.8892857142857142
- Accuracy score on test set: 0.8333333333333334



Conclusions

From the data visualization section, we can see that some features may have correlation with the mission outcome in several ways.

For example, with heavy payloads the successful landing or positive landing rate are more for orbit types Polar, LEO and ISS.

However, for GTO, we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Therefore, each feature may have a certain impact on the final mission outcome.

The exact ways of how each of these features impact the mission outcome are difficult to decipher.

However, we can use some machine learning algorithms to learn the pattern of the past data and predict whether a mission will be successful or not based on the given features.

Appendix

NumPy

SQL

Folium

scikit learn

pandas

Dash by plotly

matplotlib

seaborn

Thank you!