Capstone Project Final Report

[redacted]

Riabov Petr

May 30, 2024

github.com/ryabovpetr/Capstone_Project

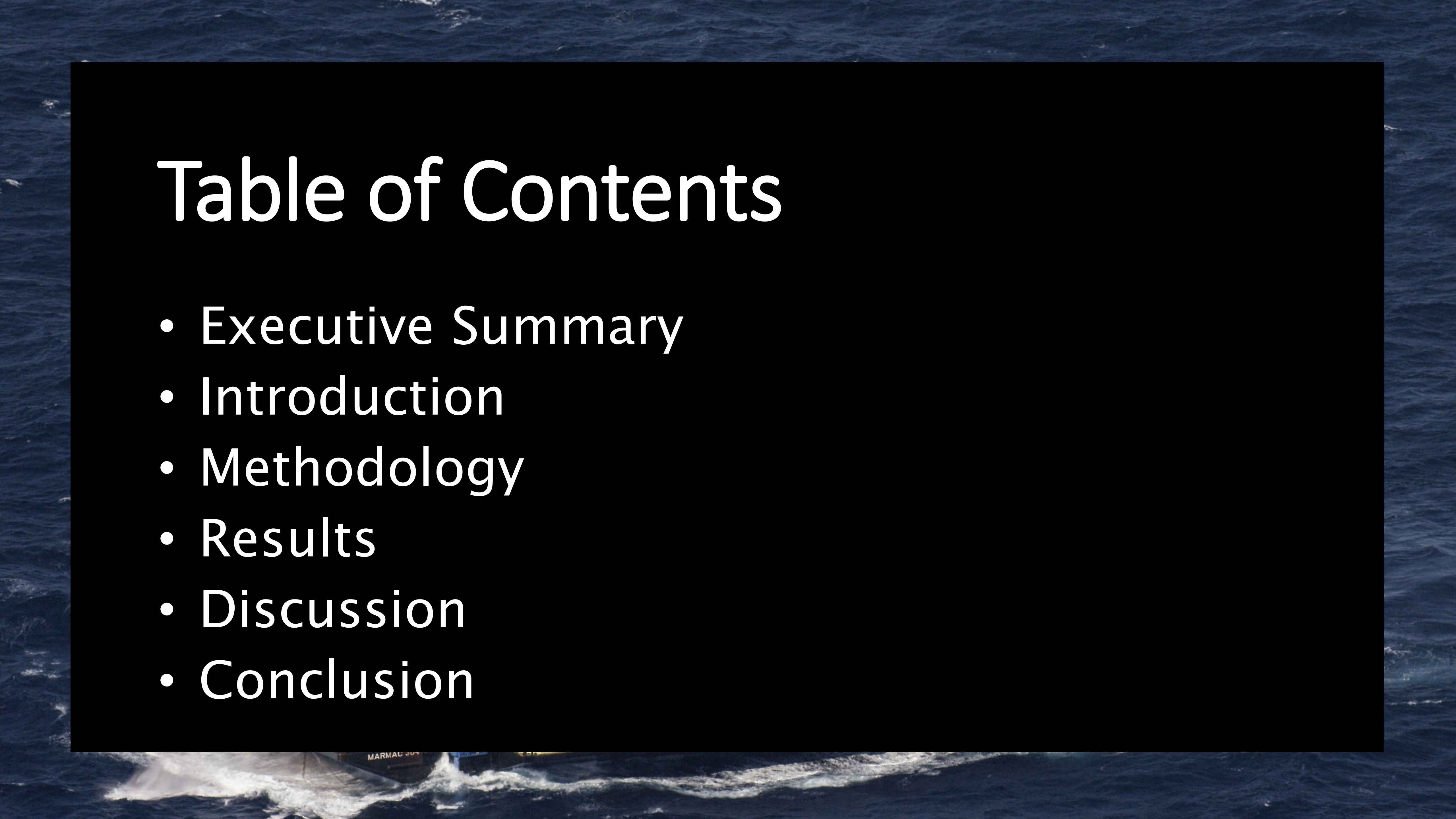


Table of Contents

- Executive Summary
- Introduction
- Methodology
- Results
- Discussion
- Conclusion

Executive Summary

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

The main steps include:

- Data collection, cleaning, and formatting
- Exploratory Data Analysis (EDA)
- Data visualization
- Prediction using machine learning methods

Our graphs show that some features of the rocket launches correlate with the outcomes. We conclude that the decision tree algorithm may be the best for predicting the success of Falcon 9 landings.

Introduction

In this project, we predict whether the first stage of the Falcon 9 rocket will successfully land. SpaceX advertises Falcon 9 launches on their website at a cost of $62 million, which is significantly cheaper than competitors.

This is due to the reusability of the first stage. By predicting the success of landings, we can estimate launch costs. This information is useful for companies competing with SpaceX.

The main question is: given a set of features about a Falcon 9 launch, such as payload mass, orbit type, launch site, etc., can we predict whether the first stage will successfully land?

Methodology

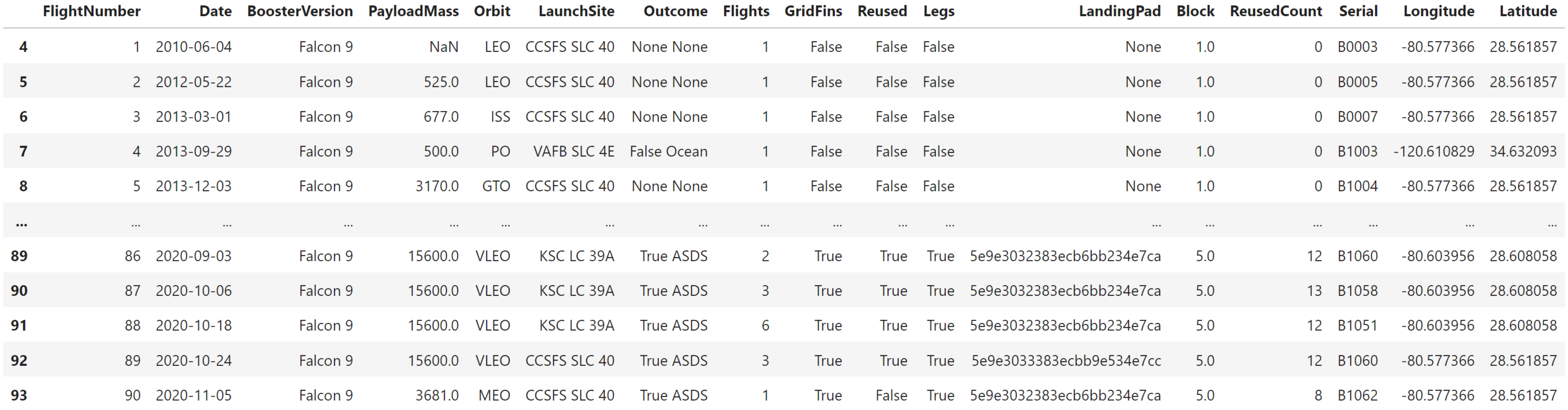
The overall methodology includes:

- Data collection using APIs and web scraping
- Data cleaning and formatting
- EDA using Pandas, NumPy, and SQL
- Data visualization using Matplotlib, Seaborn, Folium, and Dash
- Prediction using logistic regression, SVM, decision tree, and KNN

Data Collection

SpaceX API: Using API https://api.spacexdata.com/v4/rockets/

Data is filtered to include only Falcon 9 launches.
Missing values are replaced with column means.
The final dataset contains 90 rows and 17 columns.

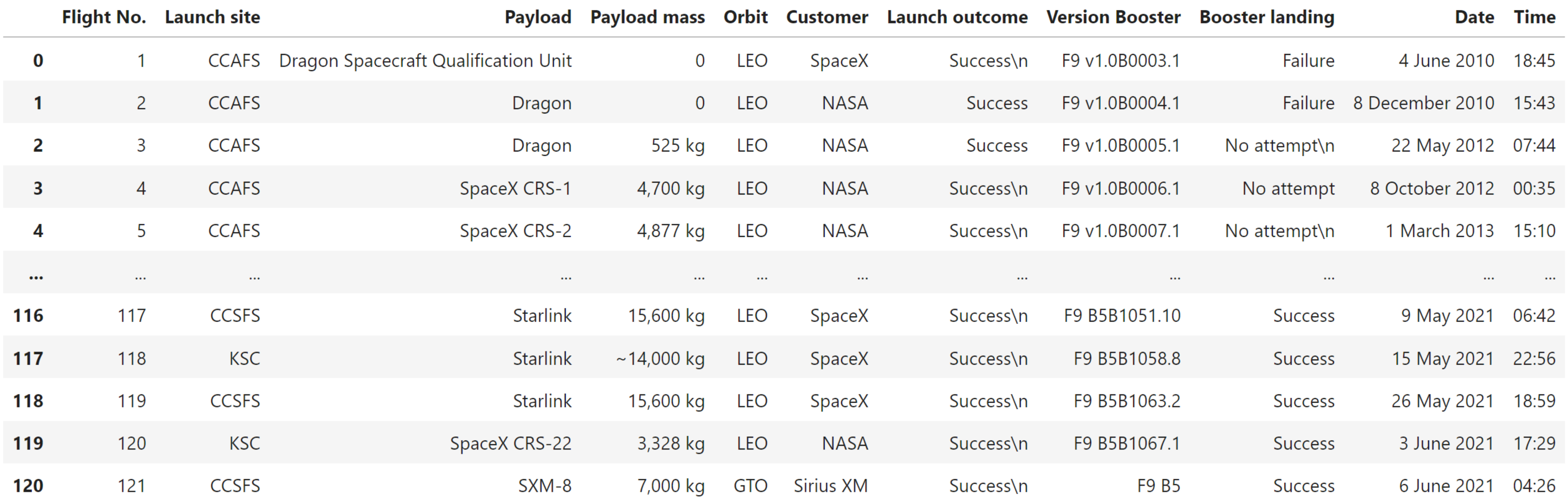


	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1060	-80.603956	28.608058
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	13	B1058	-80.603956	28.608058
91	88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	12	B1051	-80.603956	28.608058
92	89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	12	B1060	-80.577366	28.561857
93	90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	8	B1062	-80.577366	28.561857

Web Scraping

Data Source: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

The site contains data only for Falcon 9 launches.
The final dataset contains 121 rows and 11 columns.



	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
...	...	...	...	...	...	...	...	...	...	...	...
116	117	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1051.10	Success	9 May 2021	06:42
117	118	KSC	Starlink	~14,000 kg	LEO	SpaceX	Success\n	F9 B5B1058.8	Success	15 May 2021	22:56
118	119	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1063.2	Success	26 May 2021	18:59
119	120	KSC	SpaceX CRS-22	3,328 kg	LEO	NASA	Success\n	F9 B5B1067.1	Success	3 June 2021	17:29
120	121	CCSFS	SXM-8	7,000 kg	GTO	Sirius XM	Success\n	F9 B5	Success	6 June 2021	04:26

Data Processing

Missing values were handled, and categorical features were encoded using one-hot encoding.

A "Class" column was added, where 0 indicates a failed launch and 1 indicates a successful launch.

The final dataset contains 90 rows and 83 columns.

Exploratory Data Analysis (EDA)

Pandas and NumPy:

Used to gather basic information about the data, such as the number of launches at each site, orbit frequency, mission outcomes.

SQL:

Queries to obtain unique launch sites, total payload mass, and average payload mass for specific booster versions.

Data Visualization

Matplotlib and Seaborn:

Used to create charts and graphs illustrating relationships between various features.

Folium:

Created interactive maps with markers for all launch sites.

Dash:

Generated an interactive site to display launch success rates and correlations between payload mass and mission outcomes.

Prediction

Using the Scikit-learn library to create machine learning models.

Steps:

- Data standardization
- Splitting data into training and testing sets
- Creating models (logistic regression, SVM, decision tree, KNN)
- Hyperparameters tuning with GridSearchCV
- Model evaluation based on accuracy and confusion matrix

Results

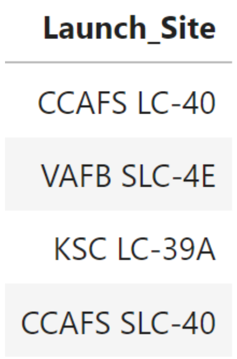
The results are split into 5 sections:

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

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

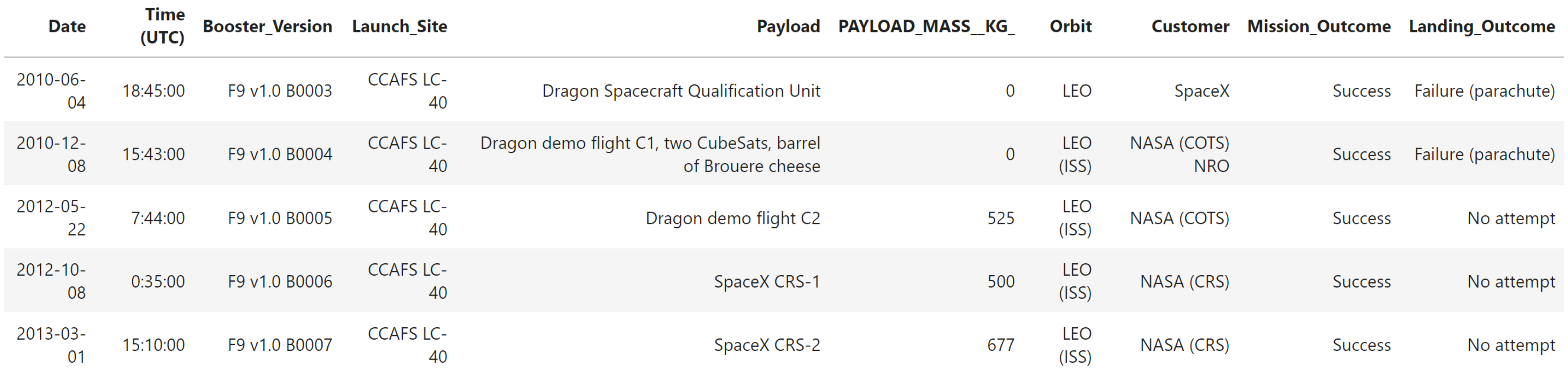
Results (EDA with SQL)

The names of the unique launch sites:



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

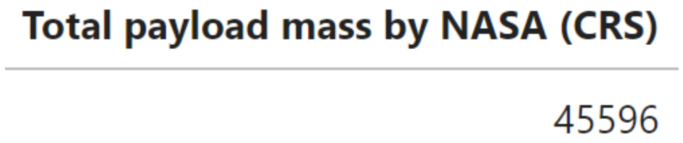
5 records where launch sites begin with “CCA”:



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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

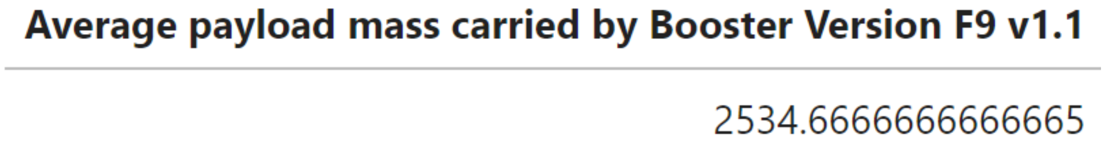
Results (EDA with SQL)

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



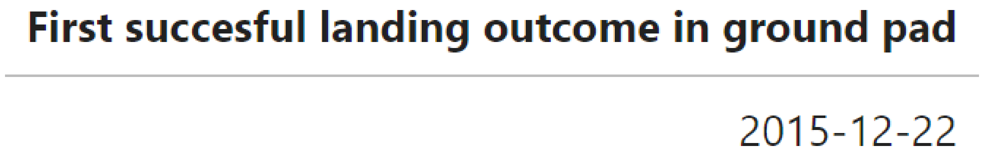
Total payload mass by NASA (CRS)
45596

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



Average payload mass carried by Booster Version F9 v1.1
2534.6666666666665

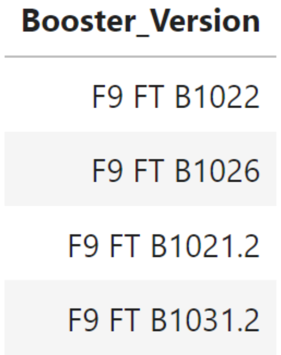
The date when the first successful landing outcome in ground pad was achieved:



First succesful landing outcome in ground pad
2015-12-22

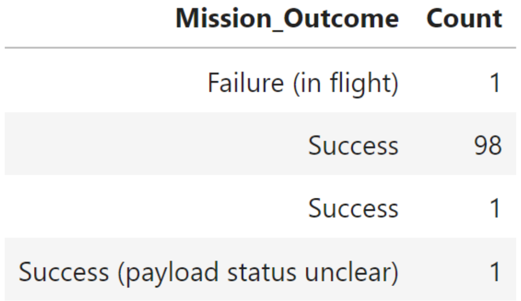
Results (EDA with SQL)

The names of the boosters which have success in drone ship and have 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

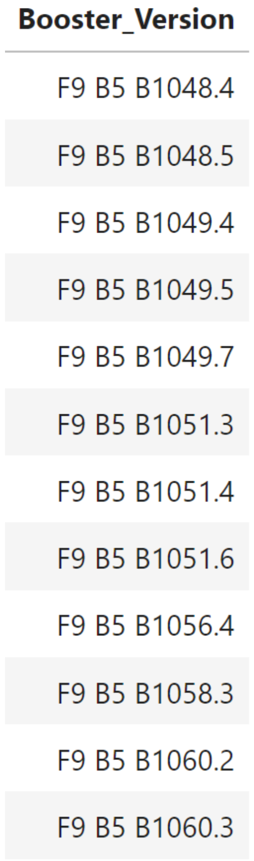
The total number of successful and failure mission outcomes:



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

Results (EDA with SQL)

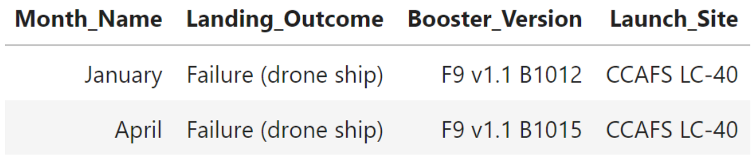
The names of the booster versions which have carried the maximum payload mass:



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

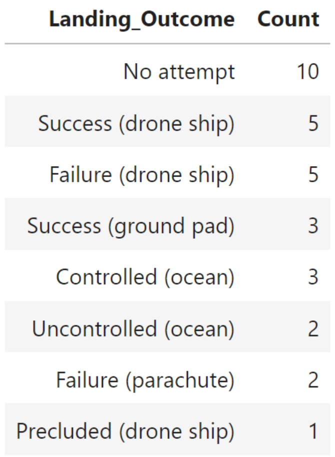
Results (EDA with SQL)

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:



Month_Name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

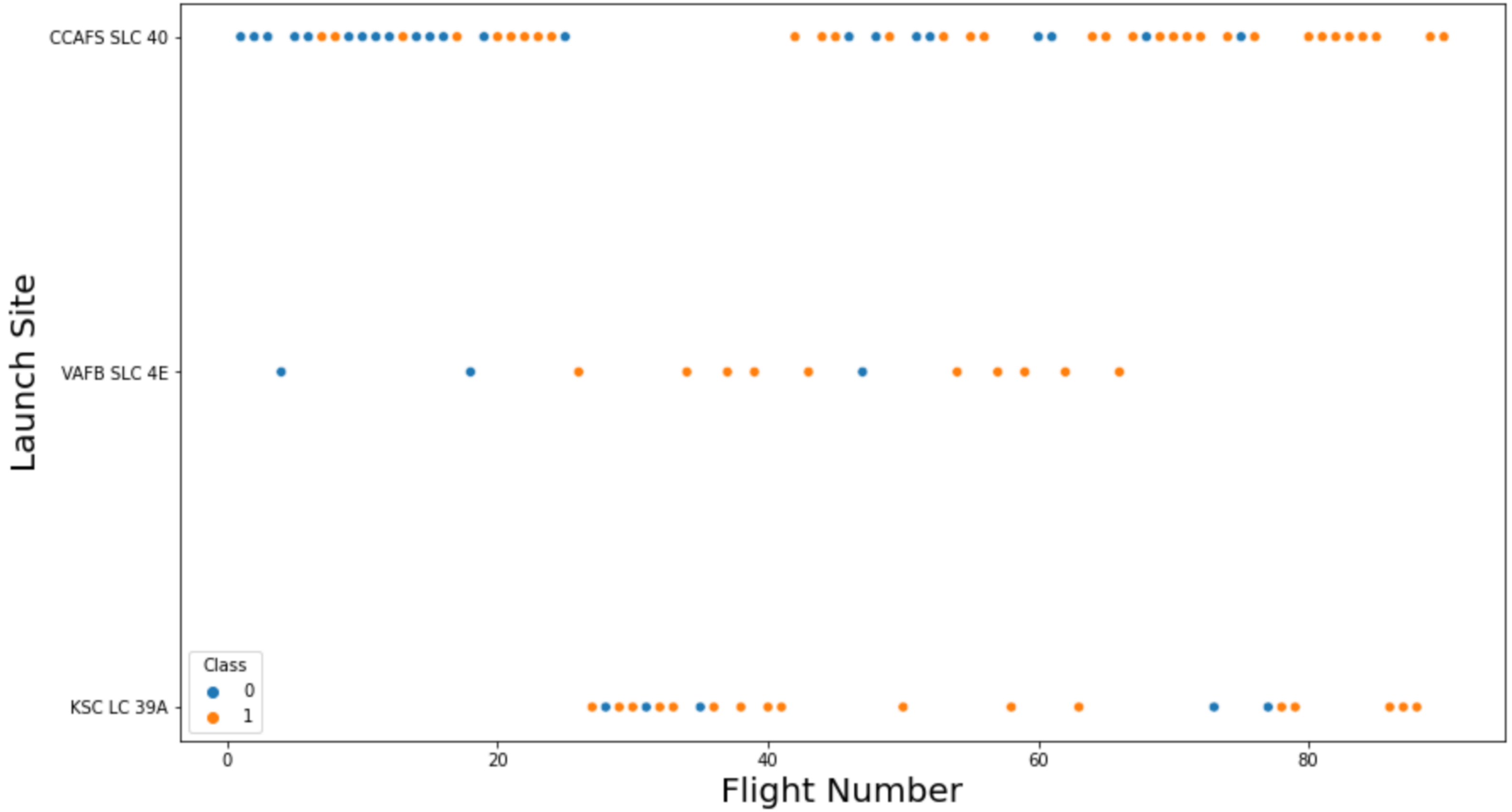
The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order:



Landing_Outcome	Count
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

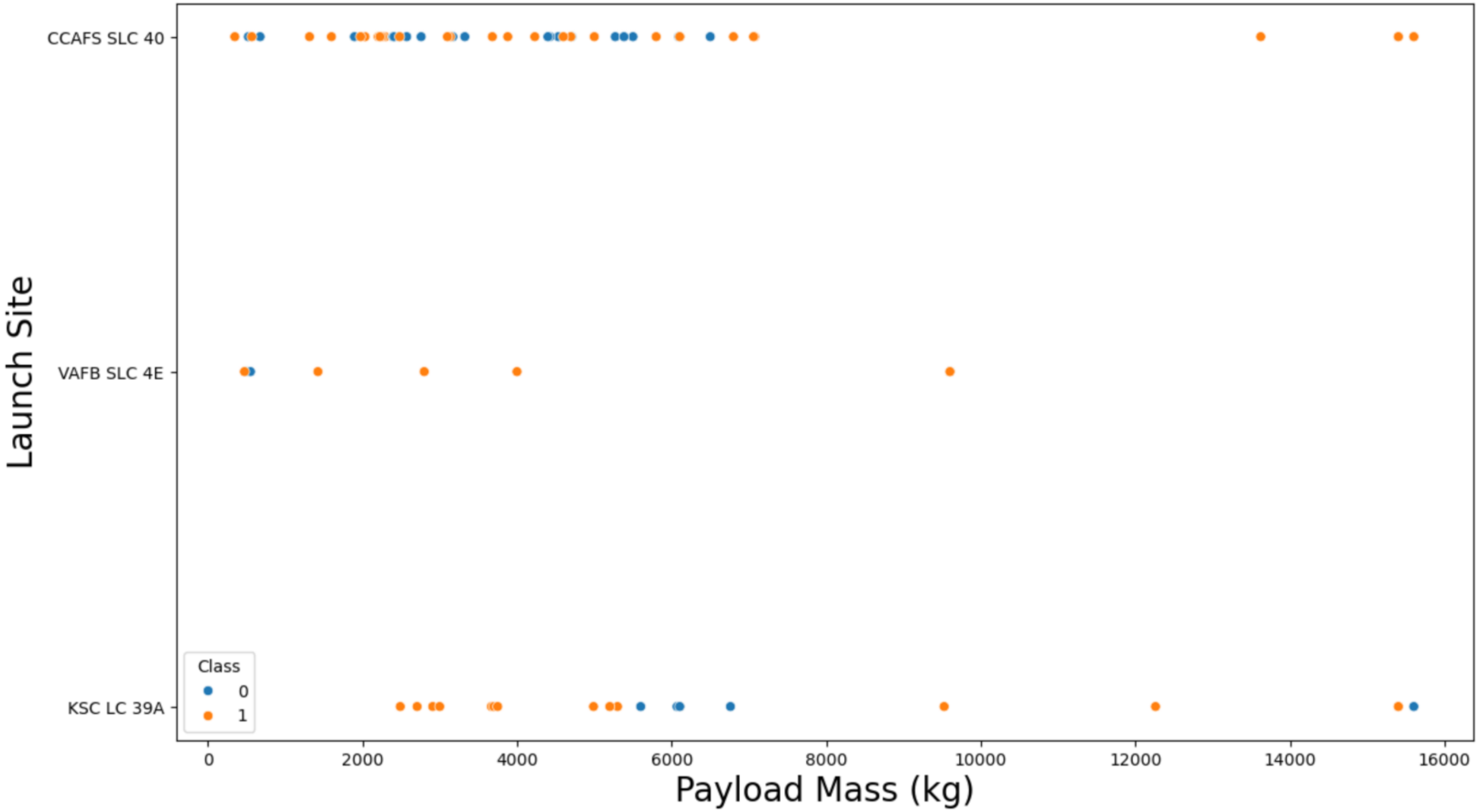
Results
Matplotlib and Seaborn (EDA with Visualization)

The relationship between flight number and launch site:



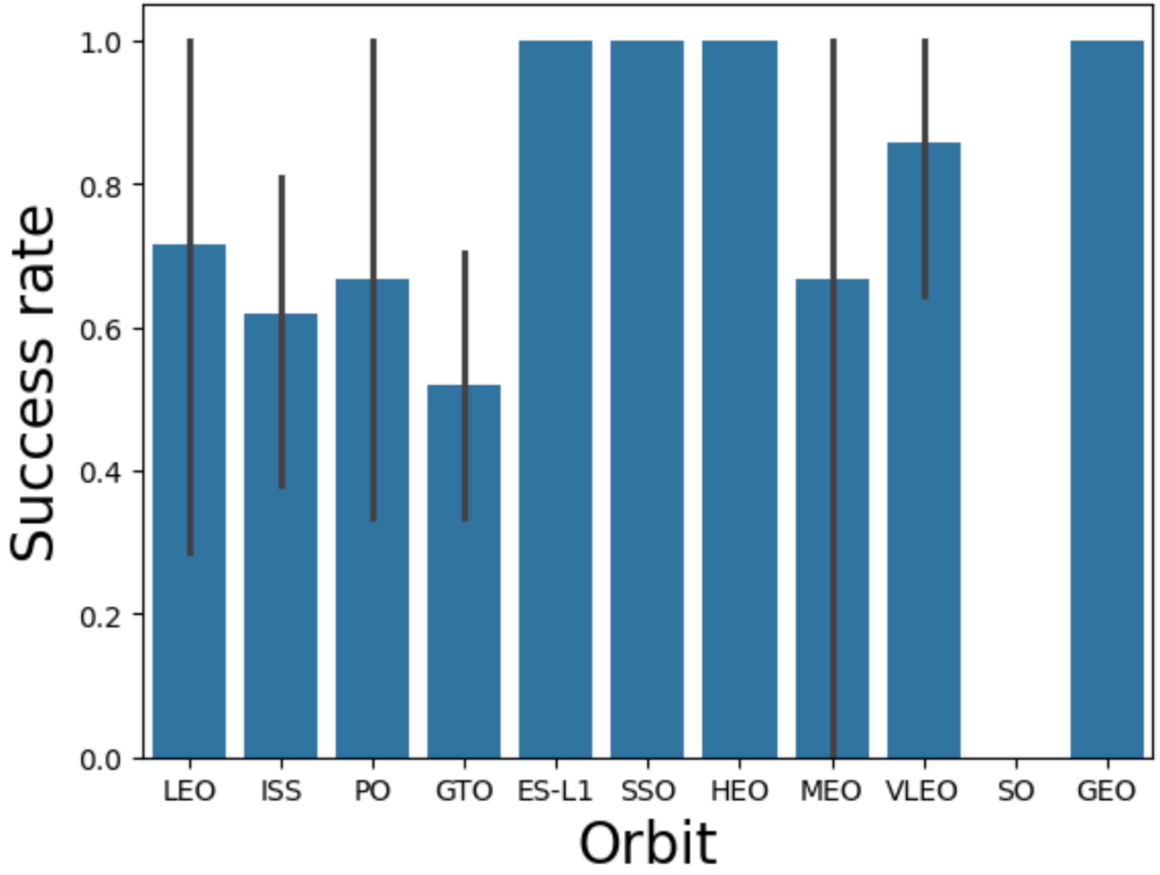
Results
Matplotlib and Seaborn (EDA with Visualization)

The relationship between payload mass and launch site:



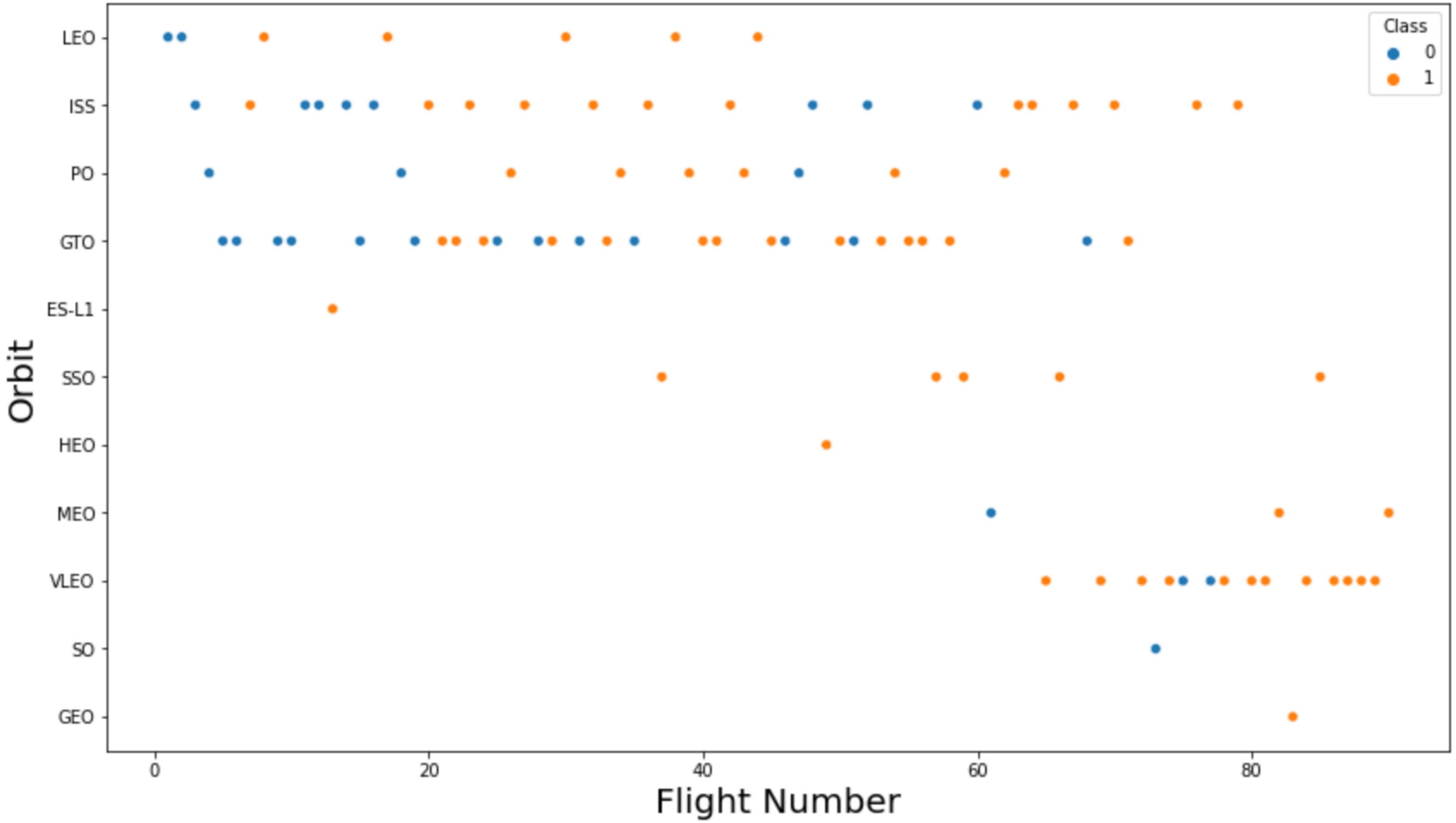
Results
Matplotlib and Seaborn (EDA with Visualization)

The relationship between success rate and orbit type:



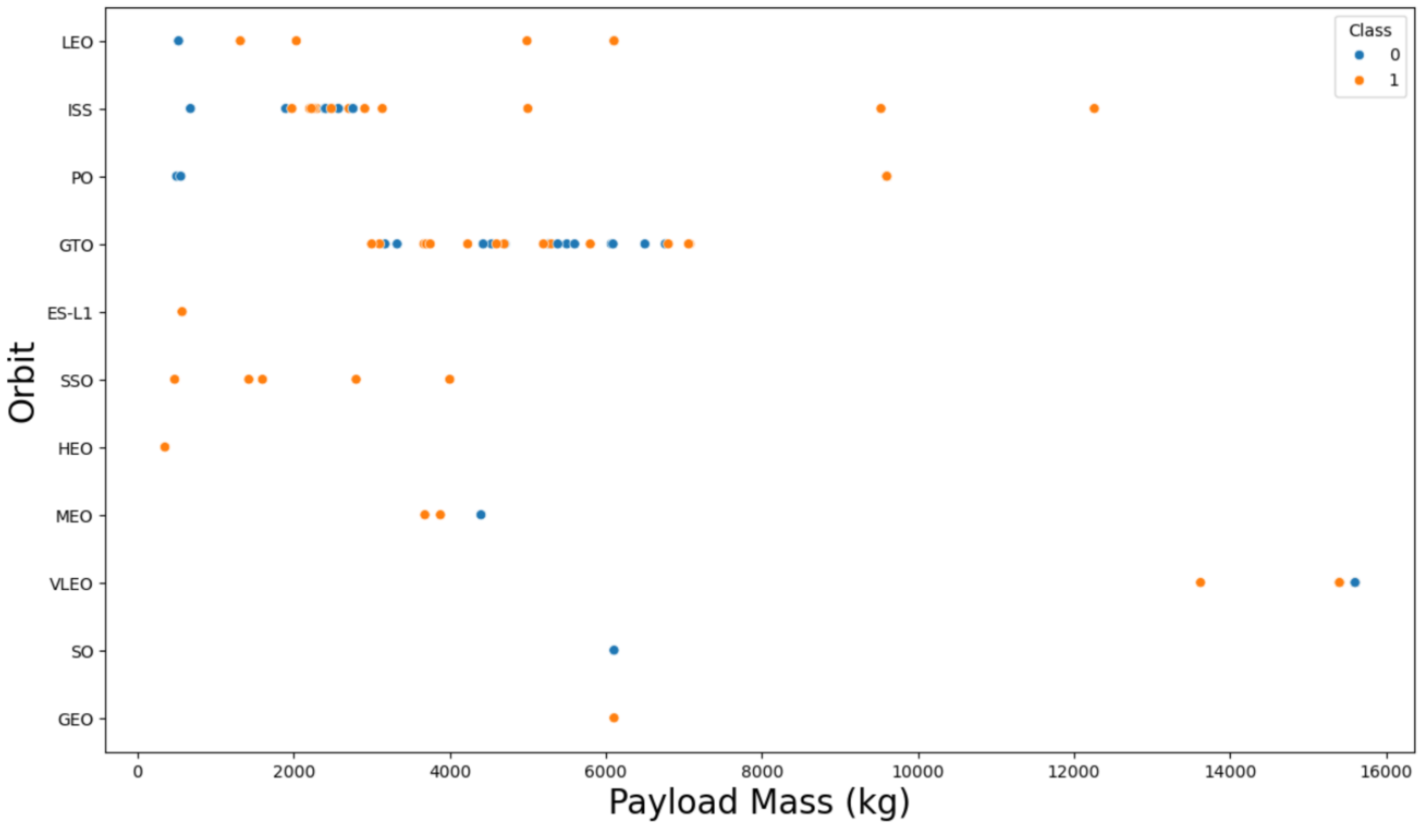
Results
Matplotlib and Seaborn (EDA with Visualization)

The relationship between flight number and orbit type:



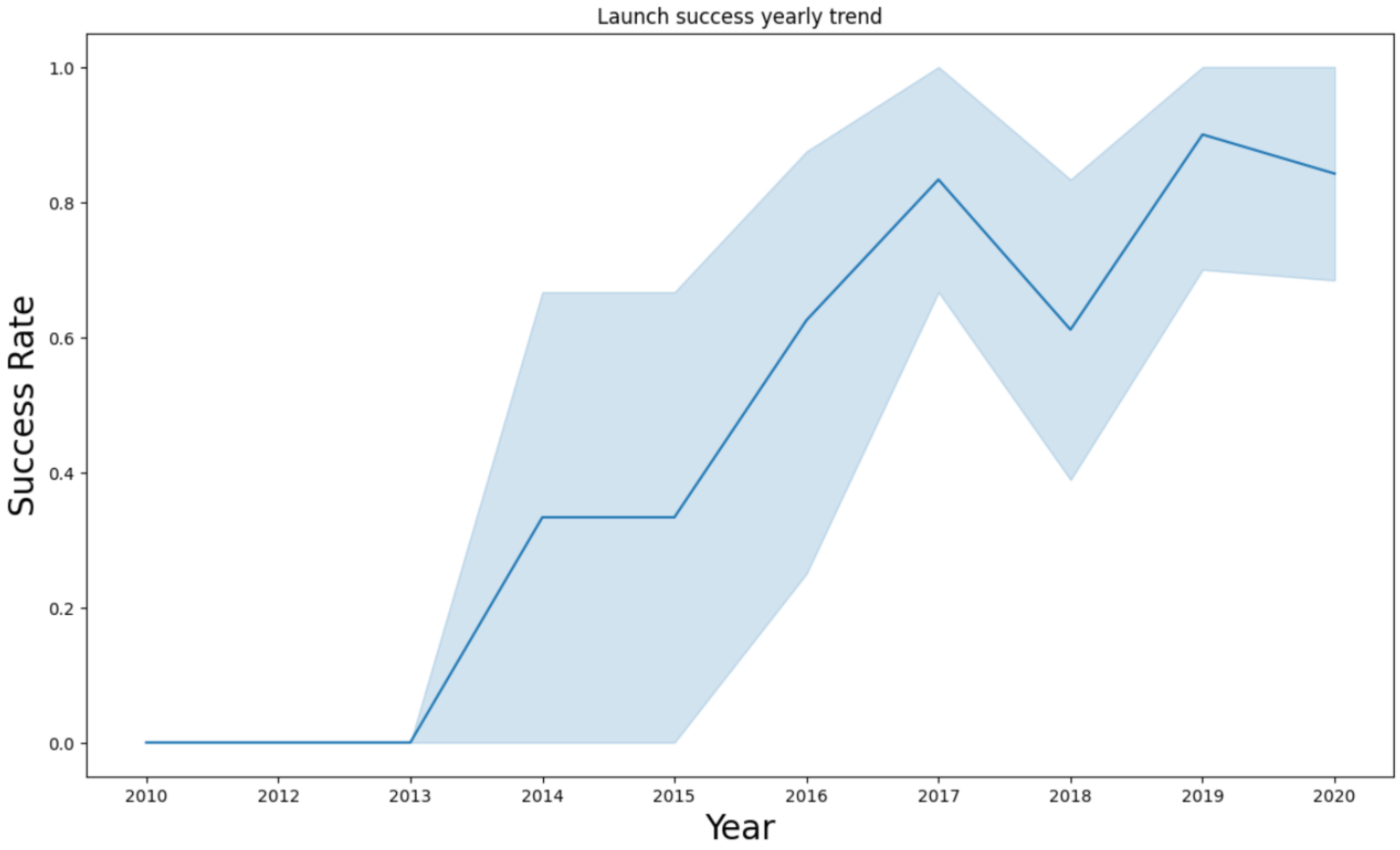
Results
Matplotlib and Seaborn (EDA with Visualization)

The relationship between payload mass and orbit type:



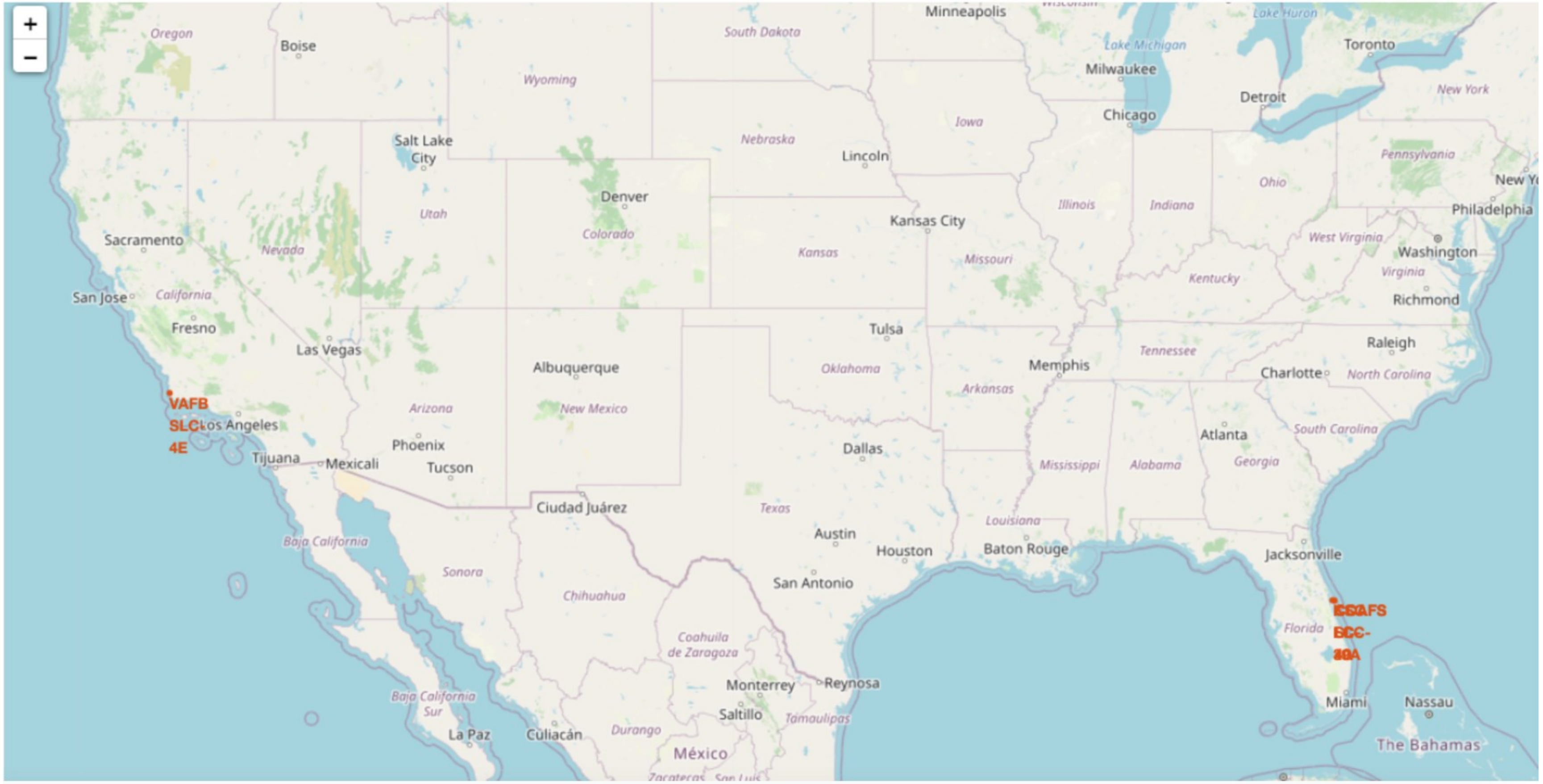
Results
Matplotlib and Seaborn (EDA with Visualization)

The launch success yearly trend:



Results (Folium)

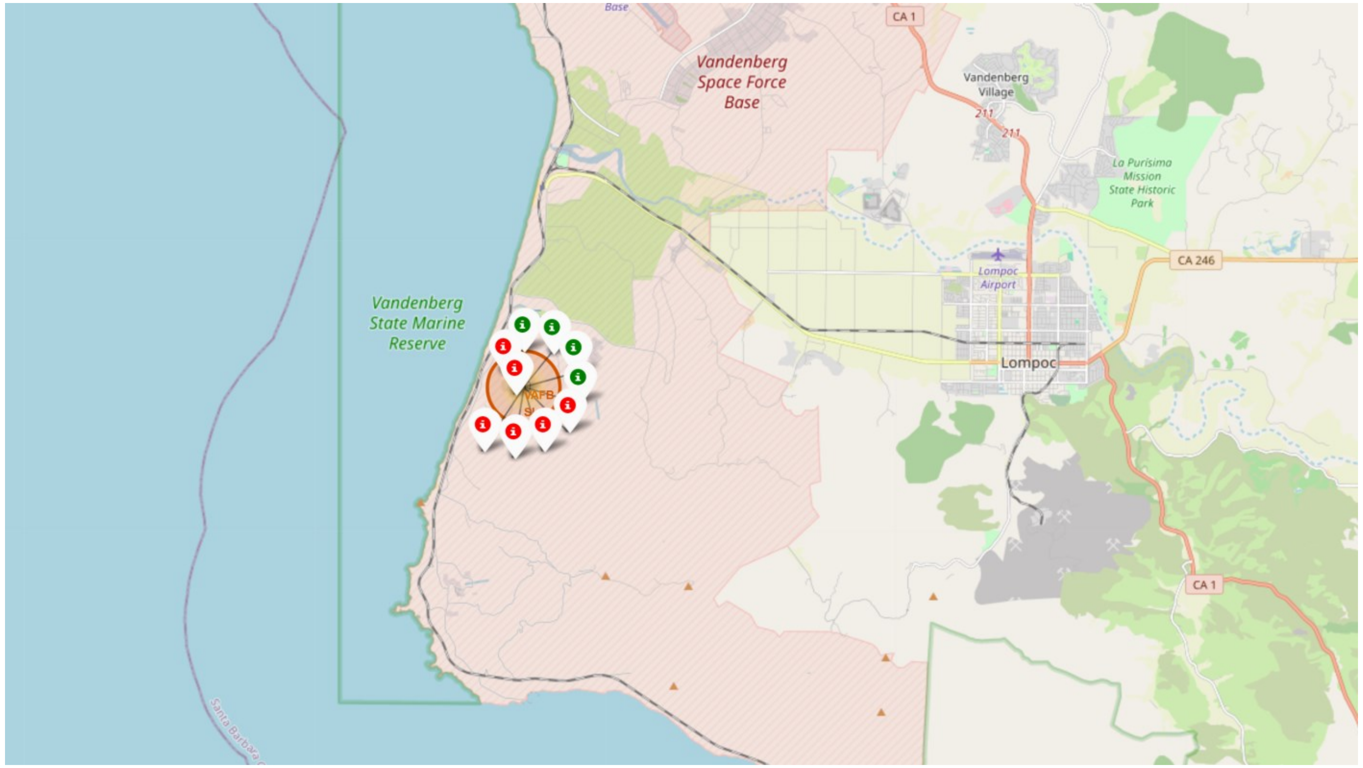
All launch sites on map:



Results (Folium)

The succeeded launches and failed launches for each site on map

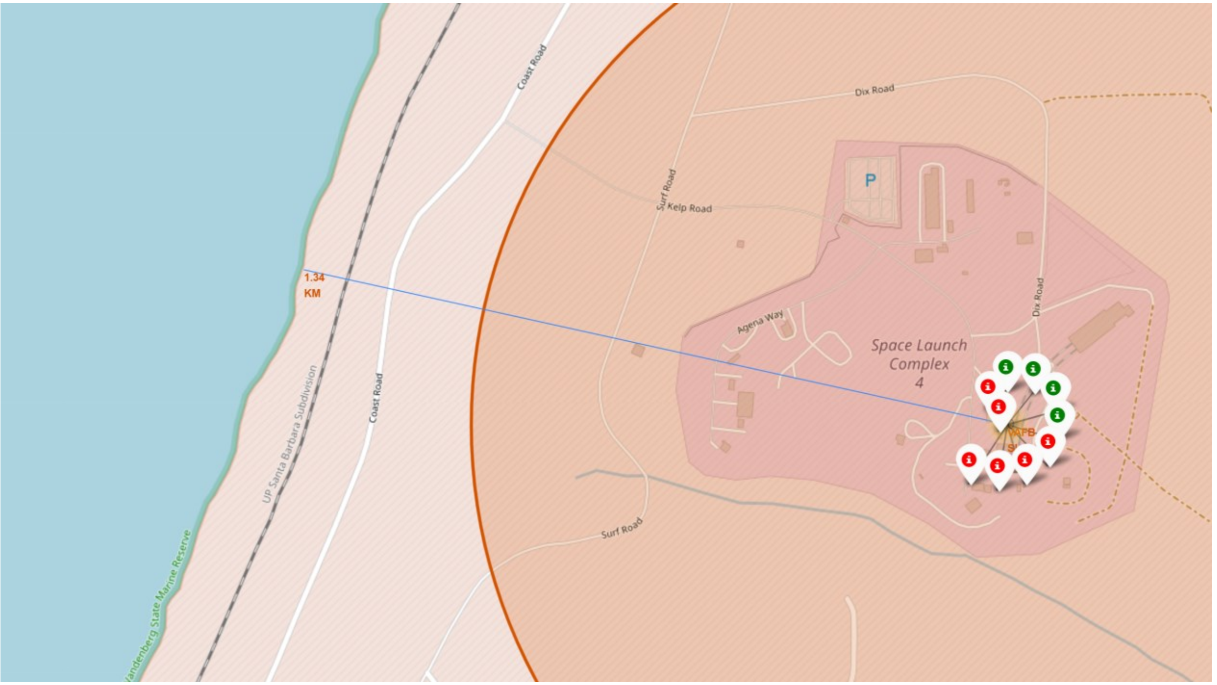
If we zoom in on one of the launch site, we can see green and red tags. Each green tag represents a successful launch while each red tag represents a failed launch


Vandenberg Space Force Base
Vandenberg Village
CA 1
211
La Purisima Mission State Historic Park
Lompoc Airport
CA 246
Vandenberg State Marine Reserve
VAFB
Lompoc
CA 1

Results (Folium)

The distances between a launch site to its proximities such as the nearest city, railway, or highway

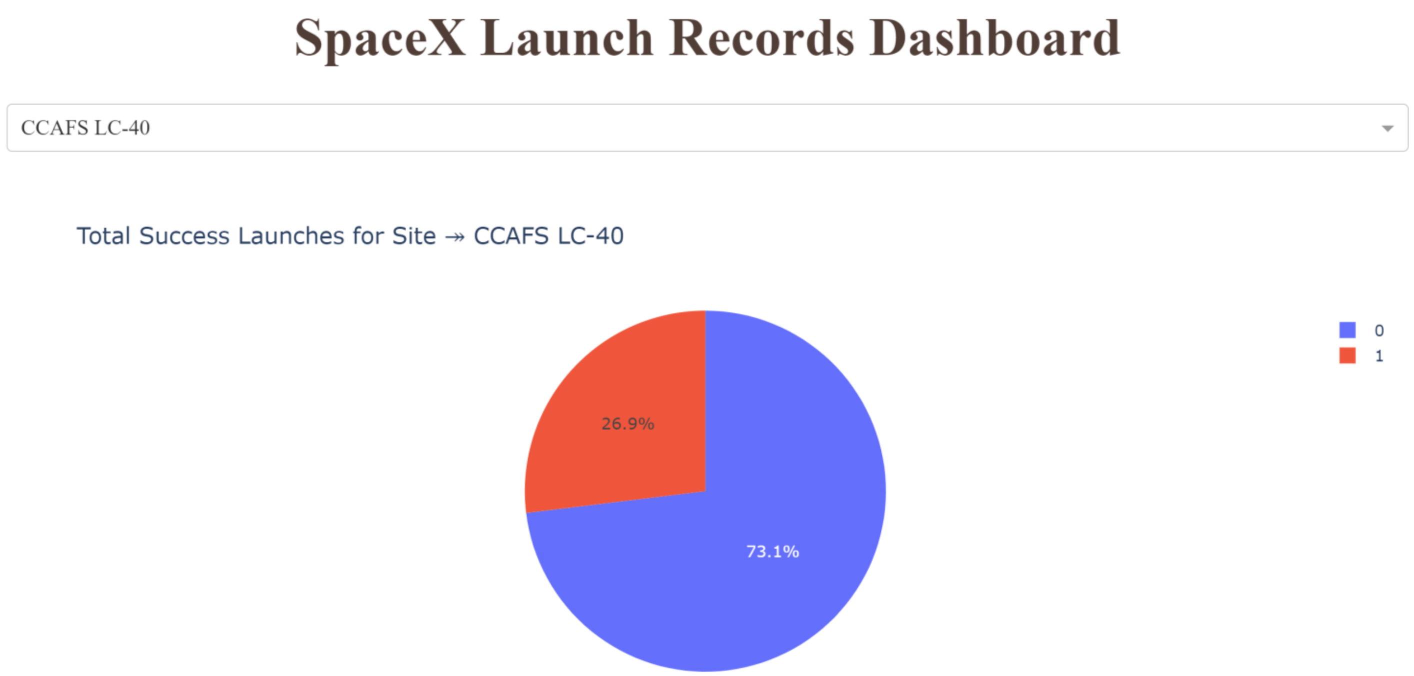
The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline



Results (Dash)

The picture below shows a pie chart when launch site CCAFS LC-40 is chosen.

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.



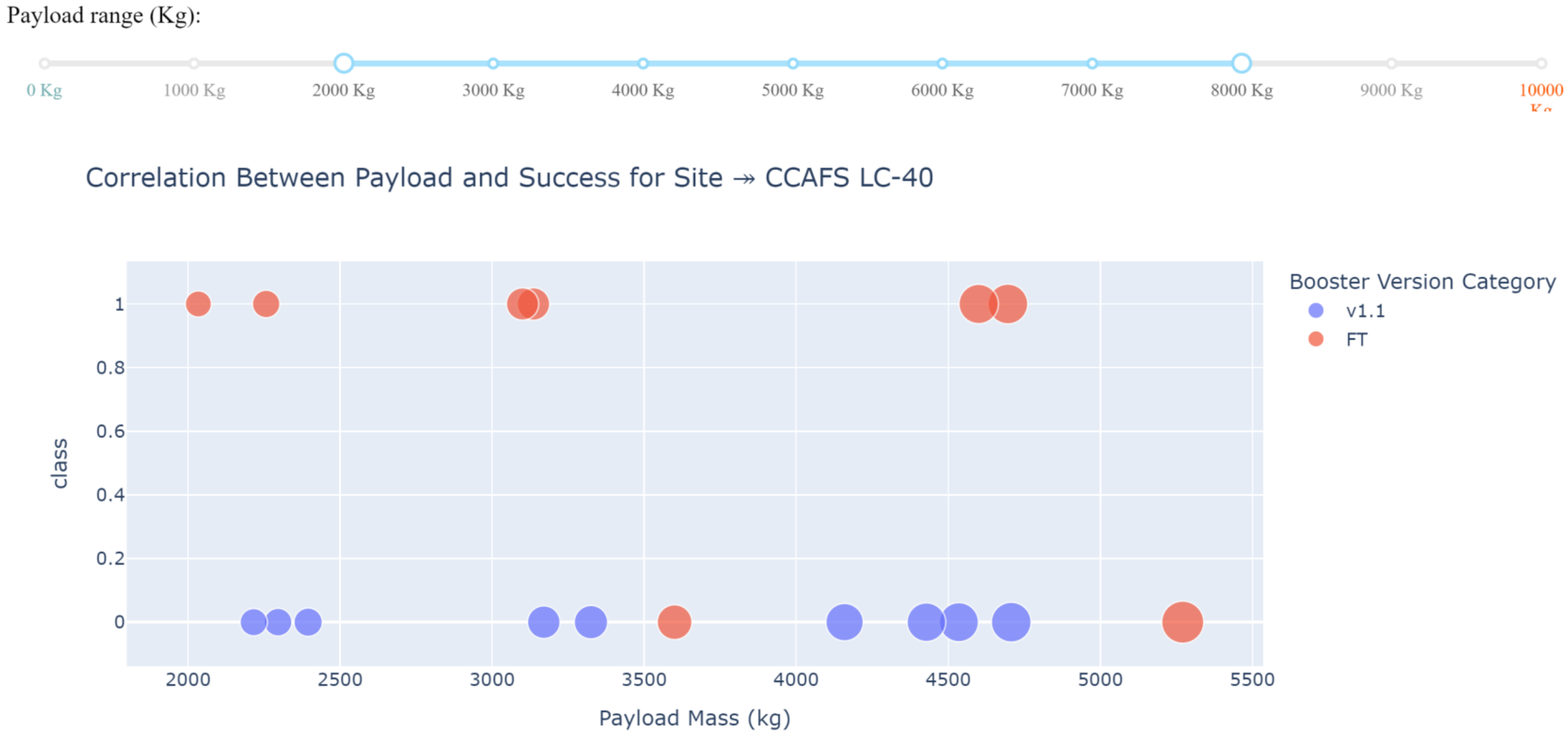
SpaceX Launch Records Dashboard

CCAFS LC-40

Results (Dash)

The picture below shows a scatterplot when the payload mass range is set to be from 2000kg to 8000kg.

Class 0 represents failed launches while class 1 represents successful launches.



Payload range (Kg):

Correlation Between Payload and Success for Site → CCAFS LC-40

Results (Predictive Analysis)

Logistic regression:

GridSearchCV best score: 84.64 %

Accuracy score on test set: 83.33 %

Support vector machine (SVM):

GridSearchCV best score: 84.82 %

Accuracy score on test set: 83.33 %

Decision tree:

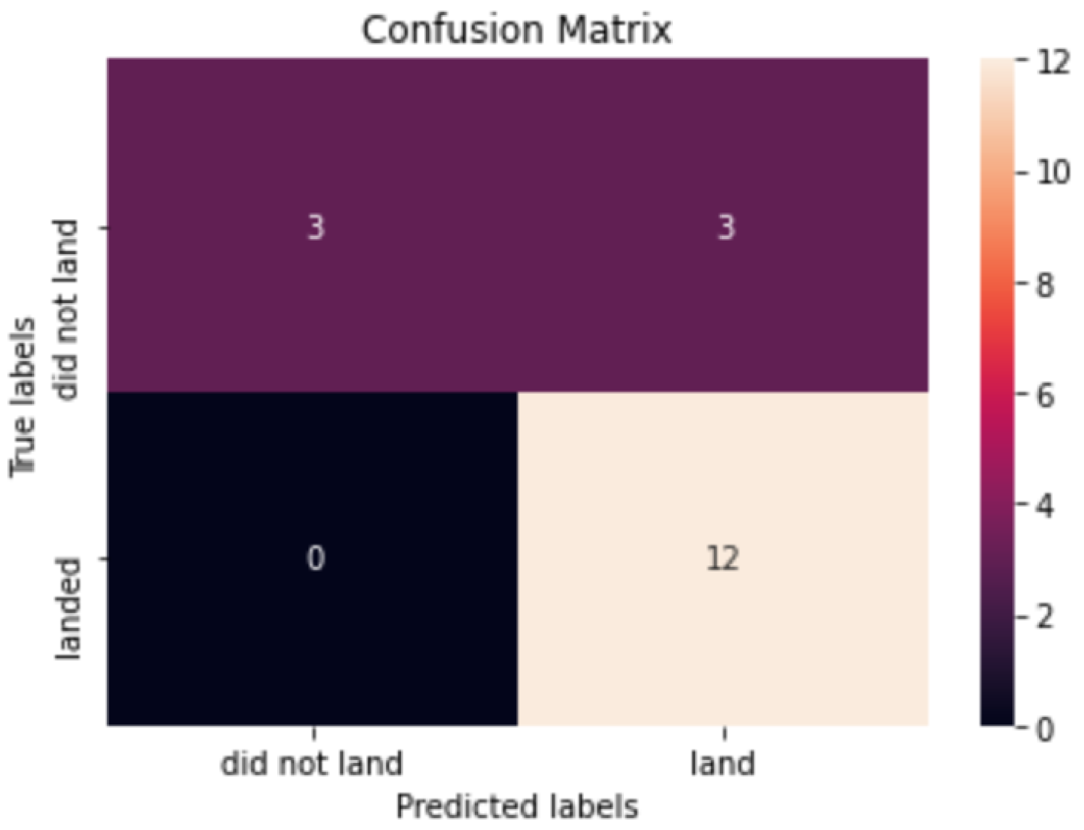
GridSearchCV best score: 88.93 %

Accuracy score on test set: 83.33 %

K nearest neighbors (KNN):

GridSearchCV best score: 84.82 %

Accuracy score on test set: 83.33 %



Results (Predictive Analysis)

When comparing the performance of all four models side by side, it is evident that they achieve identical accuracy scores and confusion matrices on the test set.

Therefore, we utilize the best scores from GridSearchCV to rank these models. Based on their GridSearchCV best scores, the models are ranked as follows, from best to worst:

1. Decision Tree: Best GridSearchCV score: 88.93 %

2. K-Nearest Neighbors (KNN): Best GridSearchCV score: 84.82 %

2. Support Vector Machine (SVM): Best GridSearchCV score: 84.82 %

4. Logistic Regression: Best GridSearchCV score: 84.64 %

Discussion

In the data visualization section, it's evident that certain features may influence the mission outcome in different ways. For instance, heavy payloads tend to correlate with higher success rates for orbit types like Polar, LEO, and ISS.

However, distinguishing such trends becomes challenging for GTO orbits, where both successful and unsuccessful landings occur. Hence, each feature likely plays a role in determining the mission outcome, although understanding precisely how each feature impacts it proves challenging.

Nonetheless, employing machine learning algorithms enables us to analyze historical data patterns and forecast the success of future missions based on these features. The analysis underscores the significance of various launch features in influencing mission outcomes, with machine learning facilitating the identification of predictive patterns.

Among the considered models, the decision tree algorithm stands out for its superior performance in this regard.

Conclusion

In this project, our aim is to forecast whether the first stage of a given Falcon 9 launch will successfully land, which is crucial for determining launch costs.

Each aspect of a Falcon 9 launch, such as payload mass or orbit type, potentially influences the mission outcome uniquely. To achieve this, we utilize various machine learning algorithms to analyze historical Falcon 9 launch data and construct predictive models.

Among the four algorithms tested, the decision tree model outperformed the others. This project showcases the viability of predicting the success of Falcon 9 first-stage landings.

The developed models offer a means to estimate launch expenses and formulate competitive bids, with the decision tree algorithm emerging as the most suitable tool for this purpose.