Winning Space Race with Data Science

Rodolpho AKAKPO
05/18/2024

Outline

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

Executive Summary

The methodology used involved several key steps : first, we thoroughly **understood the problem** of predicting the successful landing of the Falcon 9 first stage to estimate launch costs. We then **collected a comprehensive dataset** using SpaceX's API and web scraping techniques. The data collected underwent an **understanding stage (exploratory data analysis)** where visualization, descriptive statistics helped us understand the relationships between several features and feature engineering helped us in selecting relevant features for our analysis. Moving to a **preparation stage** the data has been cleaned, preprocessed to ensure quality and usability. We developed and tested **multiple machine learning models**, including logistic regression, decision trees, support vector machines with cross-validation techniques employed to prevent overfitting. The models were **evaluated** using different metrics.

The decision tree model emerged as the best performer, providing reliable predictions of landing success.

Introduction

In the competitive landscape of aerospace industry, the ability to reduce launch costs is a significant advantage. SpaceX has revolutionized this domain by successfully reusing the first stage of its Falcon 9 rockets, bringing down the cost of launches to 62 million dollars compared to the industry standard of 165 million dollars. This cost-effectiveness is primarily due to the successful landing and reuse of the first stage, a critical component of their operational model.

The focus of this project is to predict the likelihood of a Falcon 9 first stage landing successfully. Accurate predictions can provide essential insights into launch costs, aiding alternative companies in crafting competitive bids against SpaceX. By leveraging data science techniques, this project aims to understand the factors influencing landing success and develop predictive models to forecast this outcome reliably.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Data was collected using SpaceX API and web scraping techniques with python's beautifulsoup, requests modules an pandas library.

- Data wrangling

  - Data has been filtered out to only keep those related to Falcon 9. Then several informations from different sources were aggregated to form a unique dataframe where NaN values were dealed with. Feature engineering also helped drop irrelevant fields for further analysis.

- Exploratory data analysis (EDA) with SQL, Numpy, Pandas

- Data visualization

  Static visualization using matplotlib, seaborn helped us gained relevant informations on correlation between the different variables. interactive visual analytics using Folium and Plotly Dash helped us gain further understanding of the dataset.

- Predictive analysis using classification models

  We used machine learning library scikit-learn to build several classification models. Those models has been evaluated to keep the best performing.

Data Collection

The data collection process primarily involved extracting comprehensive datasets related to all SpaceX launches from the SpaceX API. Using Python's requests library, we made requests to gather detailed information on launch dates, payload mass, launch sites, booster versions, landing outcomes... Further, we utilize wikipedia web page with web scraping techniques to extract more useful data on Falcon 9 launches. The collected data was then processed to ensure consistency and stored both in csv file and sqlite 3 database.

Data Collection – SpaceX API

- Initial data collected from https://api.spacexdata.com/v4/rockets/
- Data has been parsed into DataFrame with pandas (Data content type : Json)
- Based on the ID's obtained from the first API call, call to :
  - https://api.spacexdata.com/v4/launchpads/ to obtain launchpads names and coordinates
  - https://api.spacexdata.com/v4/payloads/ to obtain information about the payload (orbit, mass)
  - https://api.spacexdata.com/v4/cores/ to obtain some useful information about the launch including its outcome

https://github.com/KadAik/IBM-Data-Science/blob/main/Capstone%20SpaceX/01%20Data%20Collection%20using%20SpaceX's%20APIs.ipynb

Data Collection - Scraping

- Get request to https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922 related to SpaceX's Falcon 9 launches
- Response's text content from the website parsed with BeautifulSoup
- All tables on the webpage has been extracted with query on the parsed html Response
- The table of interest has been selected
- A DataFrame has been created based on the table content

https://github.com/KadAik/IBM-Data-Science/blob/main/Capstone%20SpaceX/02%20Data%20Collection%20with%20web%20scraping.ipynb

Data Wrangling

- After being collected, a DataFrame has been built upon collected Data
- Only relevant columns for our analysis has been kept
- All rows on non-Falcon 9 launches has been filtered out
- Missing values has been replaced by the mean of their columns
- Categorical fields are encoded using One-hot encoding
- An extra column called ‘Class’ is added. It equals 0 if a given launch was a failure and 1 if it was successful.
- The final dataframe was casted into float type

https://github.com/KadAik/IBM-Data-Science/blob/main/Capstone%20SpaceX/03%20Data%20wrangling.ipynb

EDA with Data Visualization

- Several plots were made to get how each variables are related to each over and to the target variable :
  - Scatter plot : it helped us in visualizing correlation between different variables and the target variable (the launch outcome); say otherwise, it helped identified how different features influence or impact the target variable and how they are correlated to each other.
  - Bar chart : helped us in gaining quantitative information on categorical variables ; for instance, it helped us visualize the success rate of each orbit
  - Line plot : it helped us to visualize how a specific variable fluctuates or trends over time ; we used it for instance to visualize how the success rate trends over years

https://github.com/KadAik/IBM-Data-Science/blob/main/Capstone%20SpaceX/04%20EDA%20Visualization.ipynb

EDA with SQL

- SQL queries are used to perform the following EDA tasks :
  - Display the names of unique launch sites in the space missions
  - Display launch sites records where the names begin with CCA
  - Display the total payload mass carried by boosters launched by NASA (CRS) : 45 596 Kg
  - Display average payload mass carried by booster version F9 v1.1 : 2 928.4 Kg
  - Get the date of the first successful landing in ground pad : 2015-12-22
  - 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 total number of successful and failure mission outcomes
  - 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

https://github.com/KadAik/IBM-Data-Science/blob/main/Capstone%20SpaceX/05%20EDA%20SQL.ipynb

Build an Interactive Map with Folium

- With Folium, different map object has been created :
  - Circle : to highlight a specific area
  - Markers : to mark a specific site and display it name when point on it
  - Markers Cluster : to group several related markers with the ability to quickly identify which sites have the highest rate of success outcomes
  - MousePosition : to quickly get coordinates of a place of interest ; for instance it helped us get the coordinates of some coastline points, highways, railways…
  - PolyLine : to visualize the distance between a site and a coastline point or some place of interest to state how closest the launch site is to those places.

https://github.com/KadAik/IBM-Data-Science/blob/main/Capstone%20SpaceX/04%20EDA%20Visualization.ipynb

Build a Dashboard with Plotly Dash

- The following elements has been added to the dashboard :
  - A pie chart to visualize the success and the failure rate of a selected site and all sites
  - A dropdown input component to make the site selection
  - A scatter point chart to visualize the correlation between the payload mass and the success for all sites and for a specific site
  - A range slider to select a payload range

https://github.com/KadAik/IBM-Data-Science/blob/main/Capstone%20SpaceX/04%20Interactive%20dashboard%20with%20Plotly%20and%20Dash.py

Predictive Analysis (Classification)

- Given the classification problem, we have built and evaluated several classification models following those steps :
  - We first standize the dataset using standard scaler
  - Split the dataset into train and test set to prevent overfitting
  - Select a given model and a set of hyperparameters
  - Build a model using gridsearchcv and cross validation to select the best parameters
  - Train the model on the train test
  - Evaluate the model on the test set

https://github.com/KadAik/IBM-Data-Science/blob/main/Capstone%20SpaceX/07%20Predicitve%20analysis%20Machine%20Learning.ipynb

Results

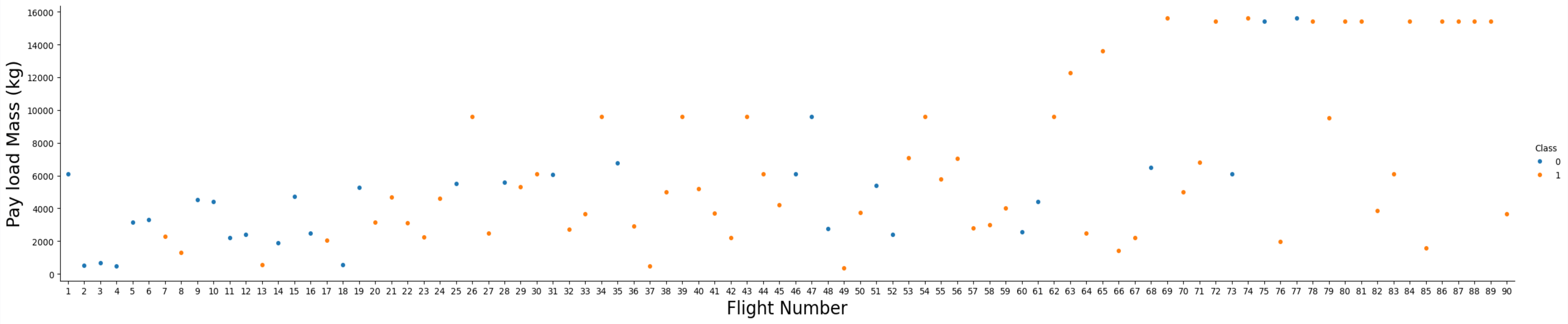
The results section will fall into three sub-sections :

- Exploratory data analysis insights with sql and stactic visualizations
- Interactive analytics
- Predictive analysis results

Section 2

Insights drawn from EDA

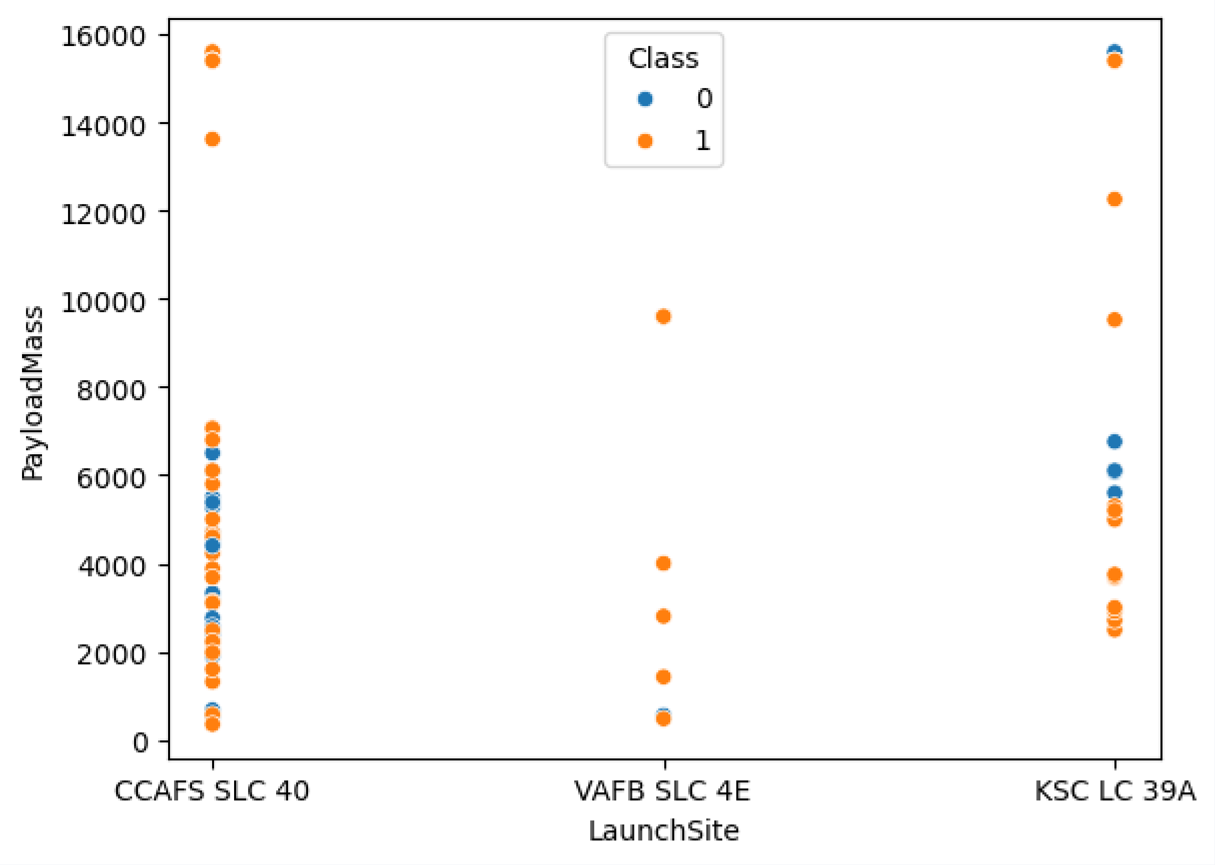
Flight Number vs. Launch Site



Flight Number vs. Launch Site

We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.

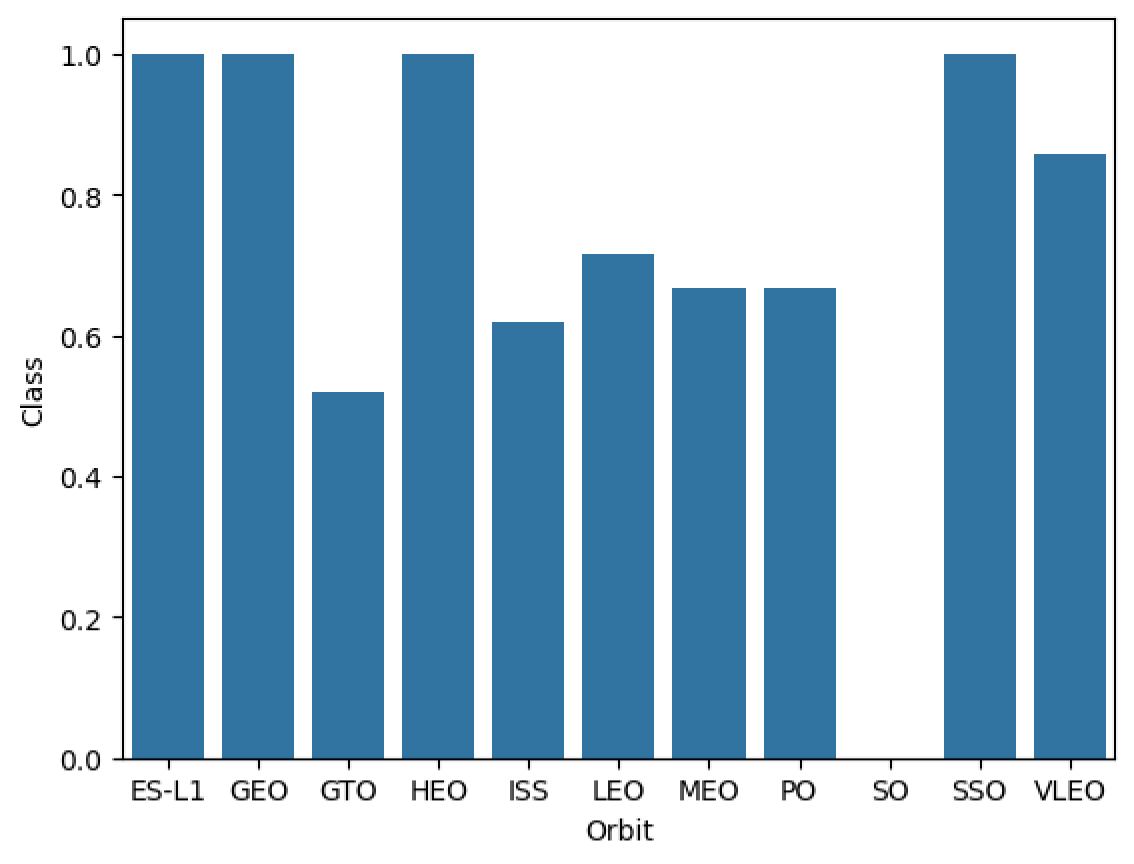
Payload vs. Launch Site



Payload vs. Launch Site

This plot highlights that for the VAFB-SLC launchsite there are
no rockets launched
for heavypayload
mass(greater than 10000).

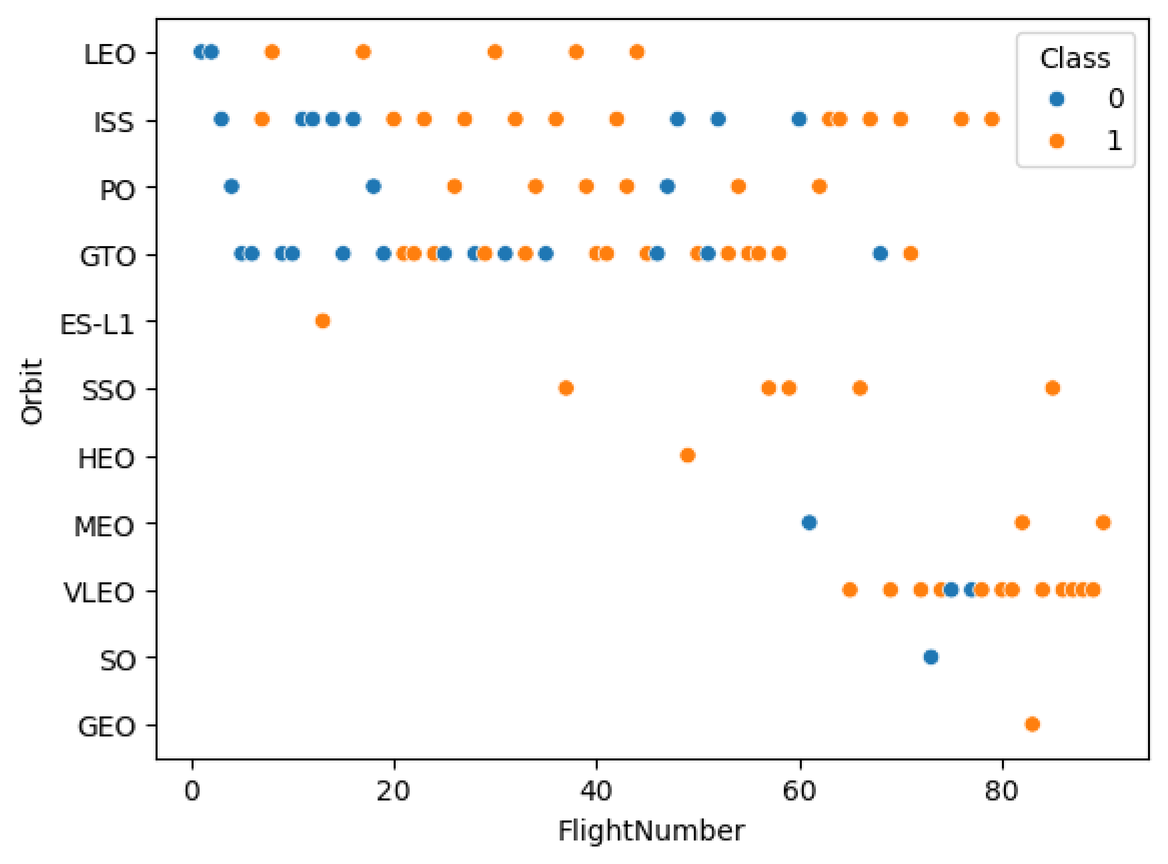
Success Rate vs. Orbit Type



Success rate of each orbit type

The bar chat shows that ES-L1, GEO, HEO and SSO orbits have the highest success rate.

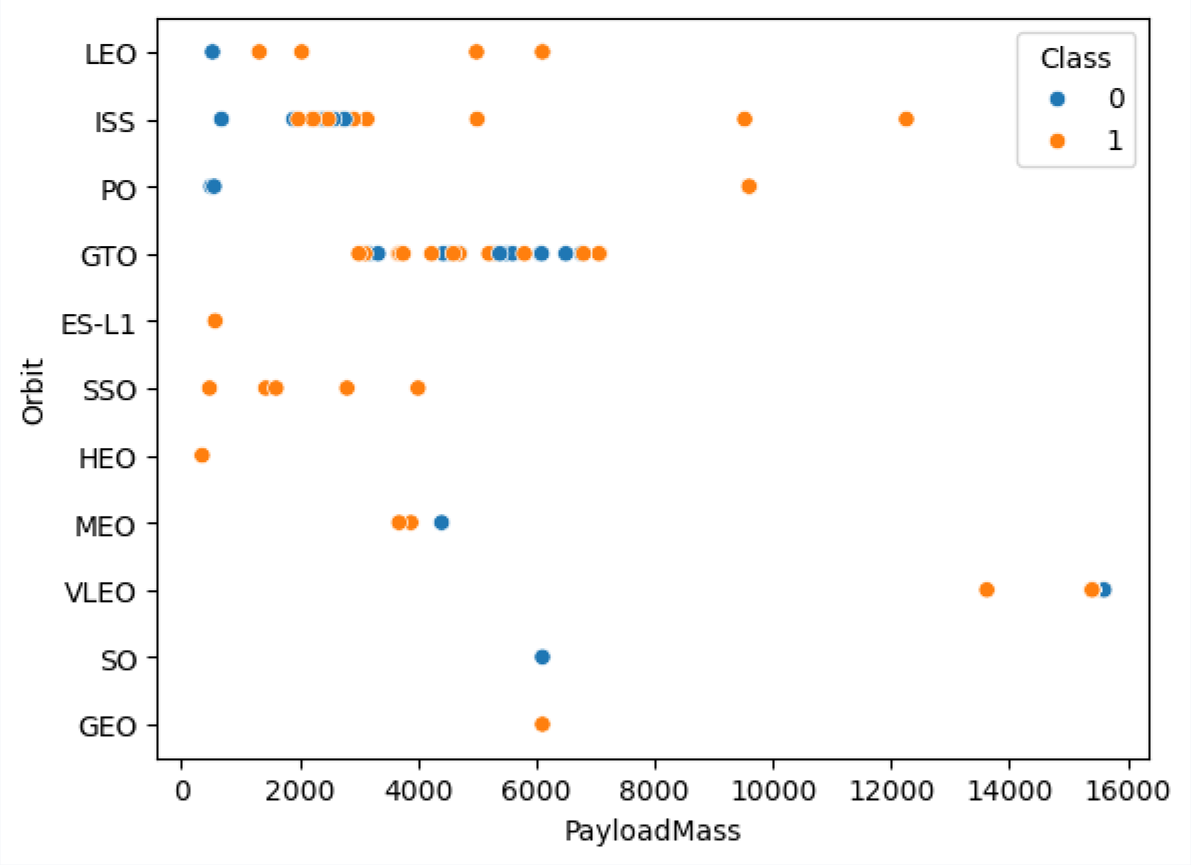
Flight Number vs. Orbit Type



Flight number vs. Orbit type

We should observe that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

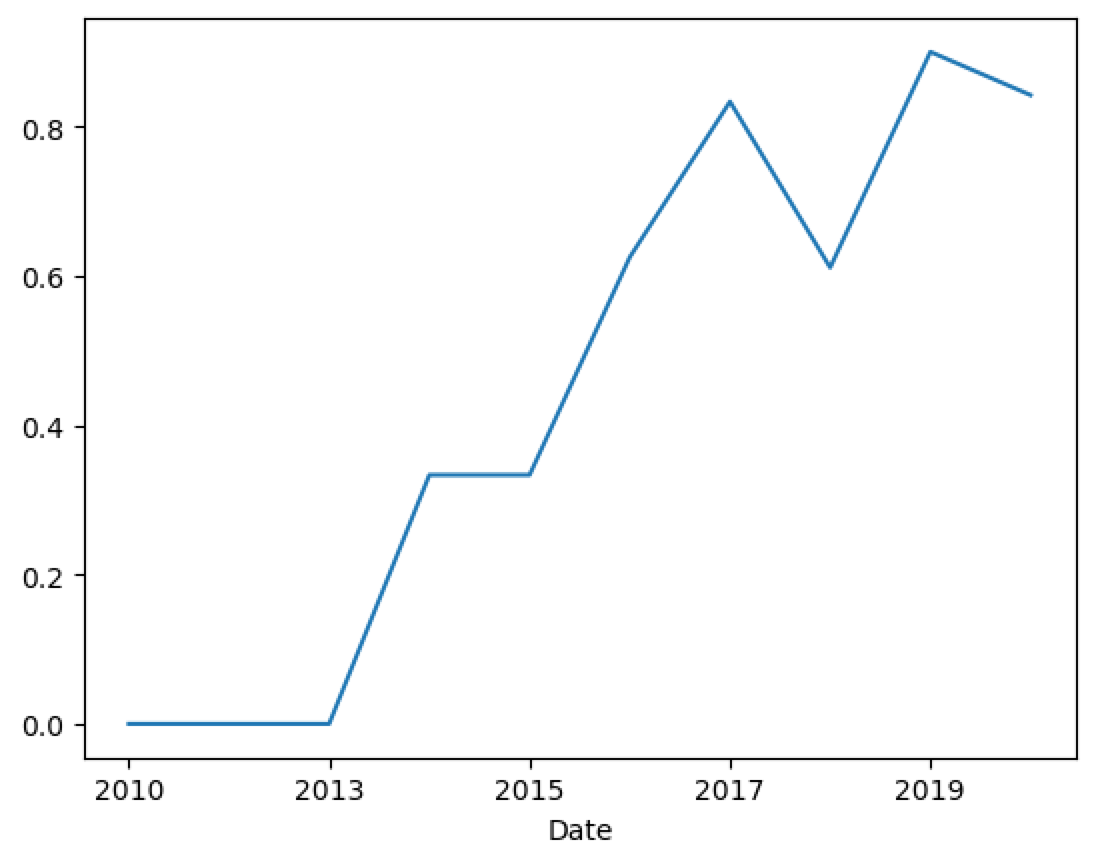


Payload vs. orbit type

With heavy payloads the successful landing or positive landing rate are more for 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.

Launch Success Yearly Trend

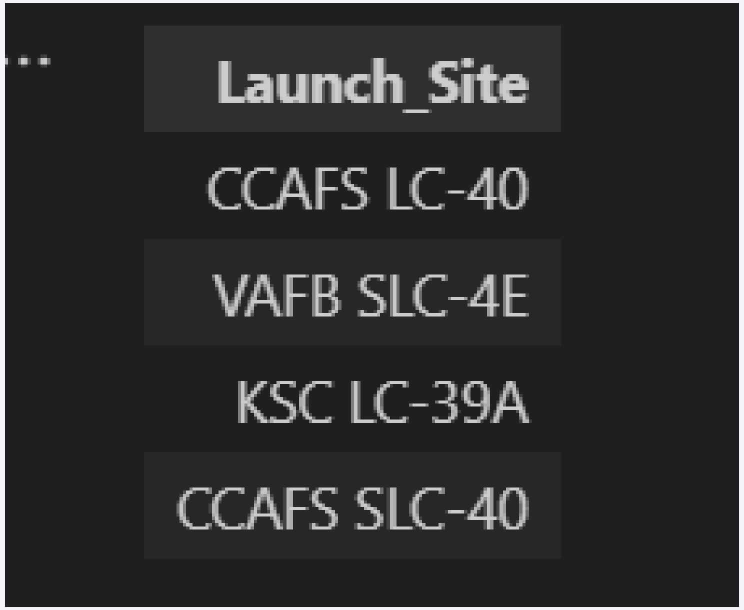


Yearly average success rate

The success rate since 2013 kept increasing till 2020 after a sharp fall around 2018.

All Launch Site Names

Names of the unique launch sites

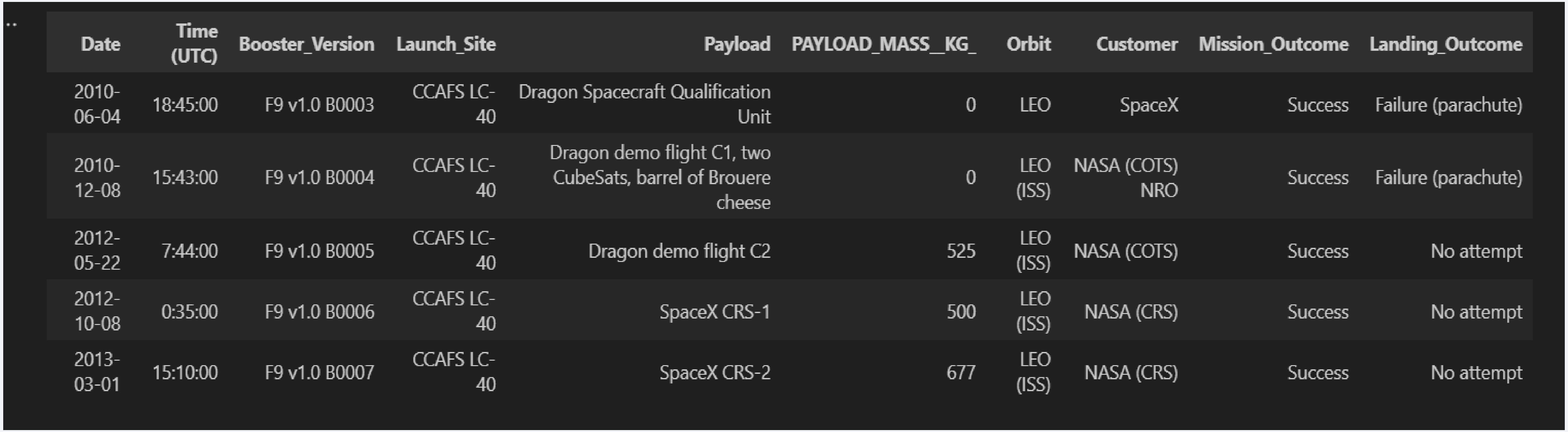


Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Falcon 9 rockets has been being launched on for sites.

Launch Site Names Begin with 'CCA'

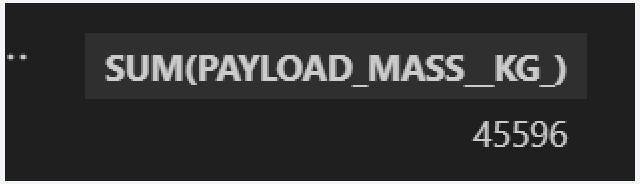
Five records where launch sites begin with `CCA`



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

Total Payload Mass

Total payload carried by boosters from NASA

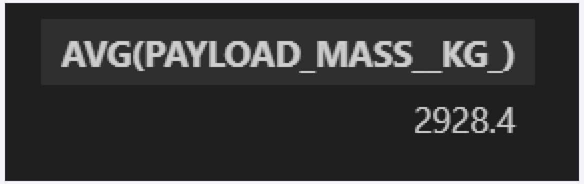


SUM(PAYLOAD_MASS__KG_)
45596

NASA's boosters carried 45 596 Kg of payload in total

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1

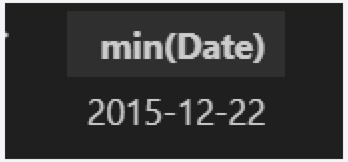


AVG(PAYLOAD_MASS__KG_)
2928.4

Booster version F9 v1.1 carried in average 2 928.4 Kg of payload

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad

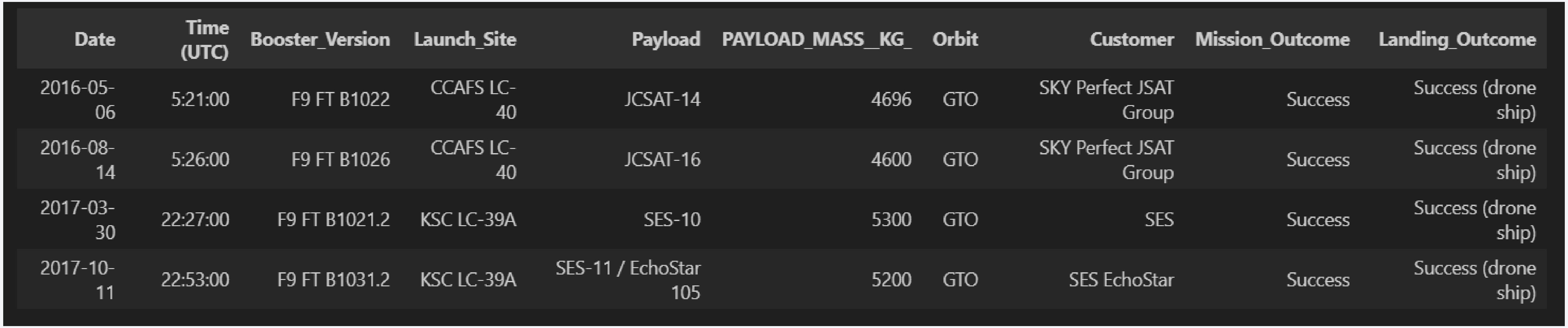


min(Date)
2015-12-22

The first successful landing outcome on ground pad took place in December 22nd, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

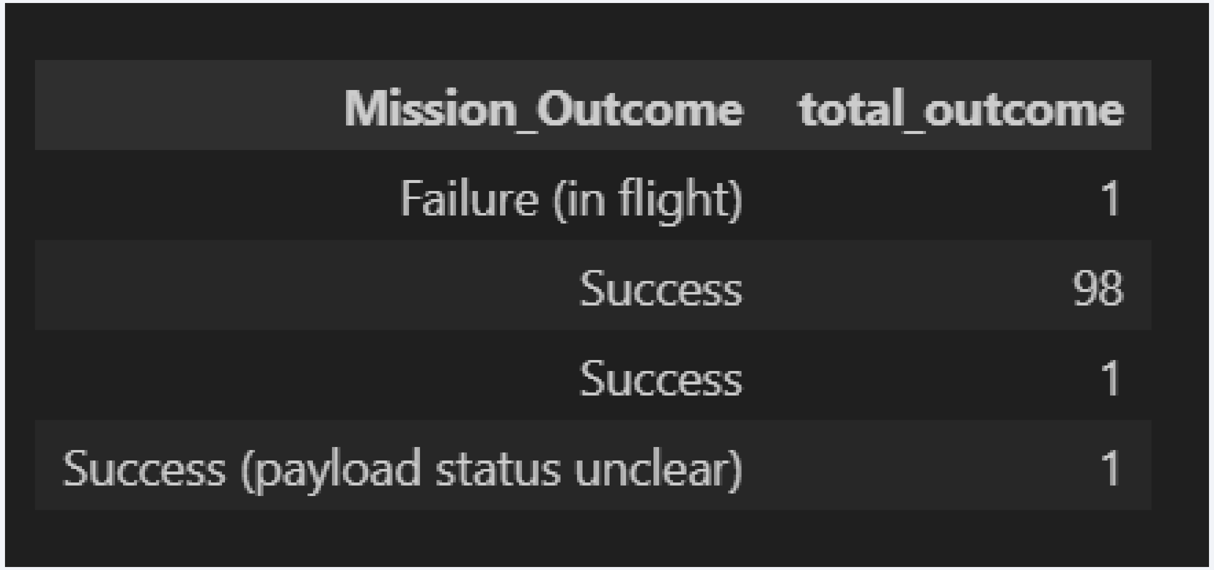


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

Four boosters have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

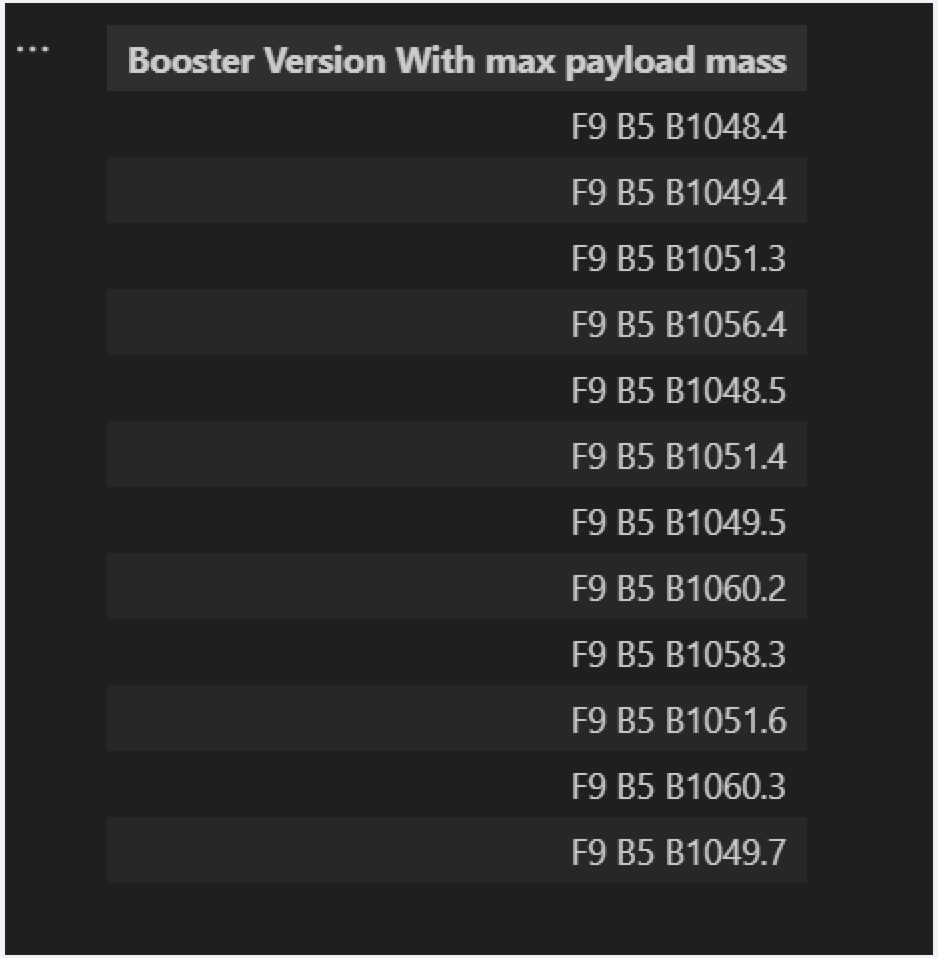
Total number of successful and failure mission outcomes



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

SpaceX carried out 100 successful missions and 1 failure one

Boosters Carried Maximum Payload



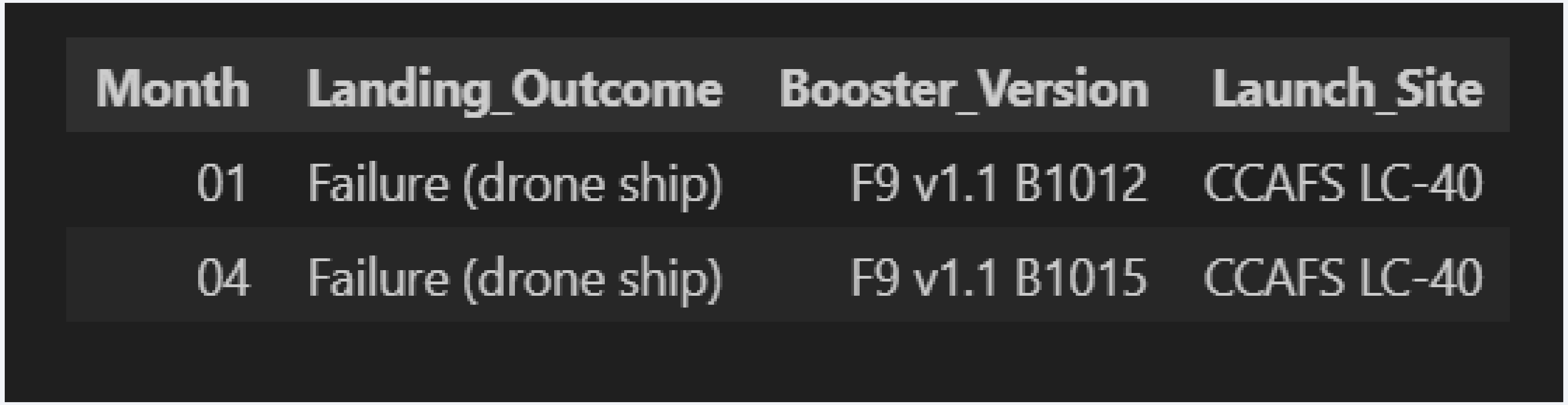
Booster Version With max payload mass
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

12 booster have carried the maximum payload mass

Names of the booster which have carried the maximum payload mass

2015 Launch Records

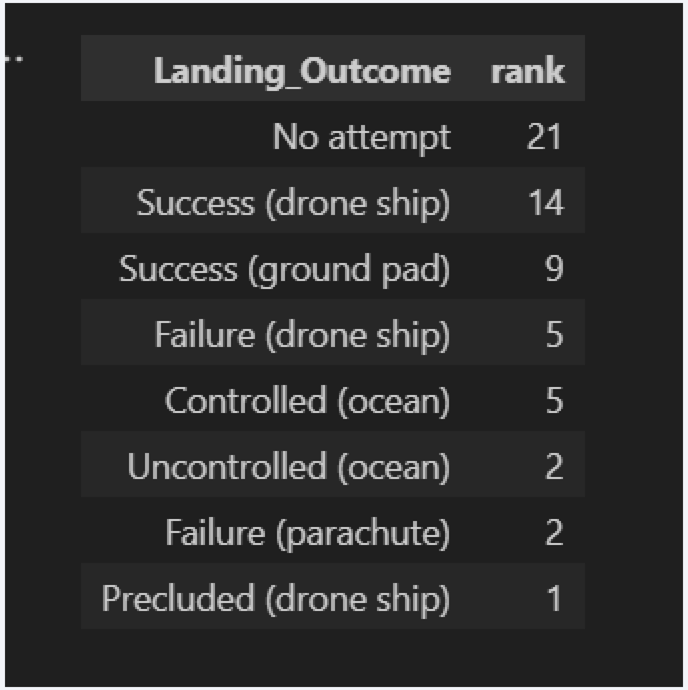
List of failed landing outcomes in drone ship, with booster versions, and launch site names in year 2015



Month	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

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

Ranking of the count of landing outcomes between the date 2010-06-04 and 2017-03-20



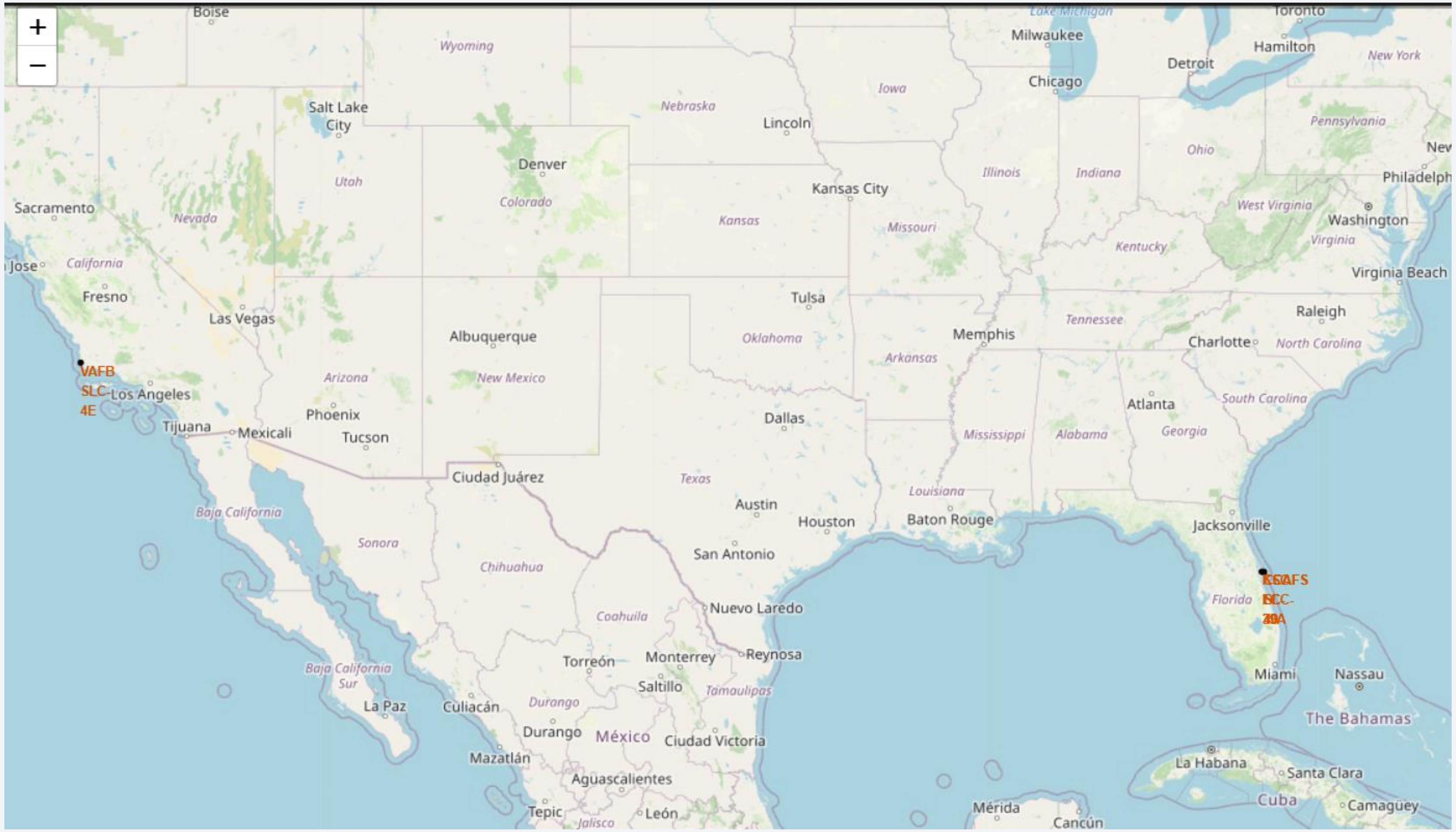
Landing_Outcome	rank
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Apart of no attempts, success in drone ship landing is occurent

Section 3

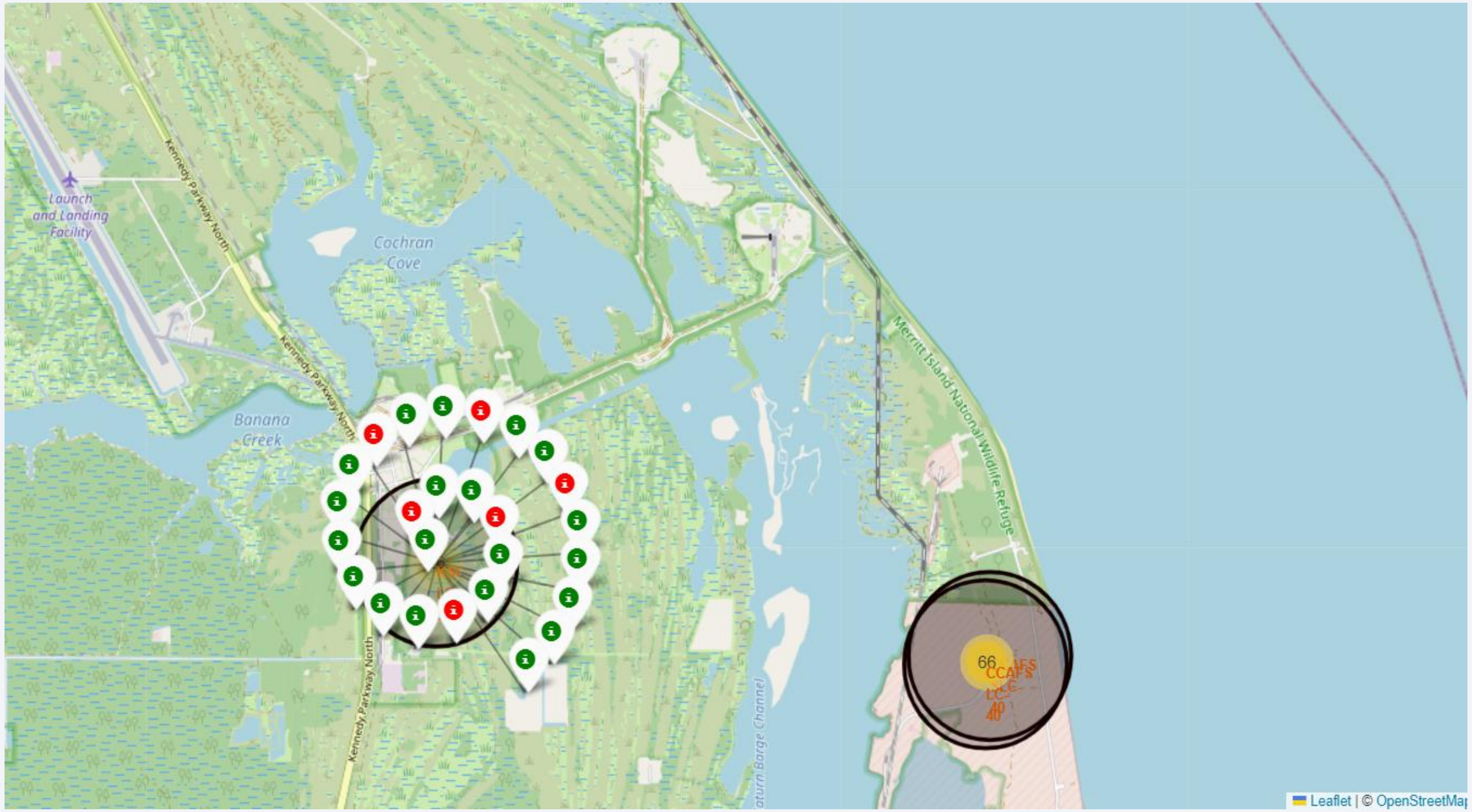
Launch Sites Proximities Analysis

Launch sites locations



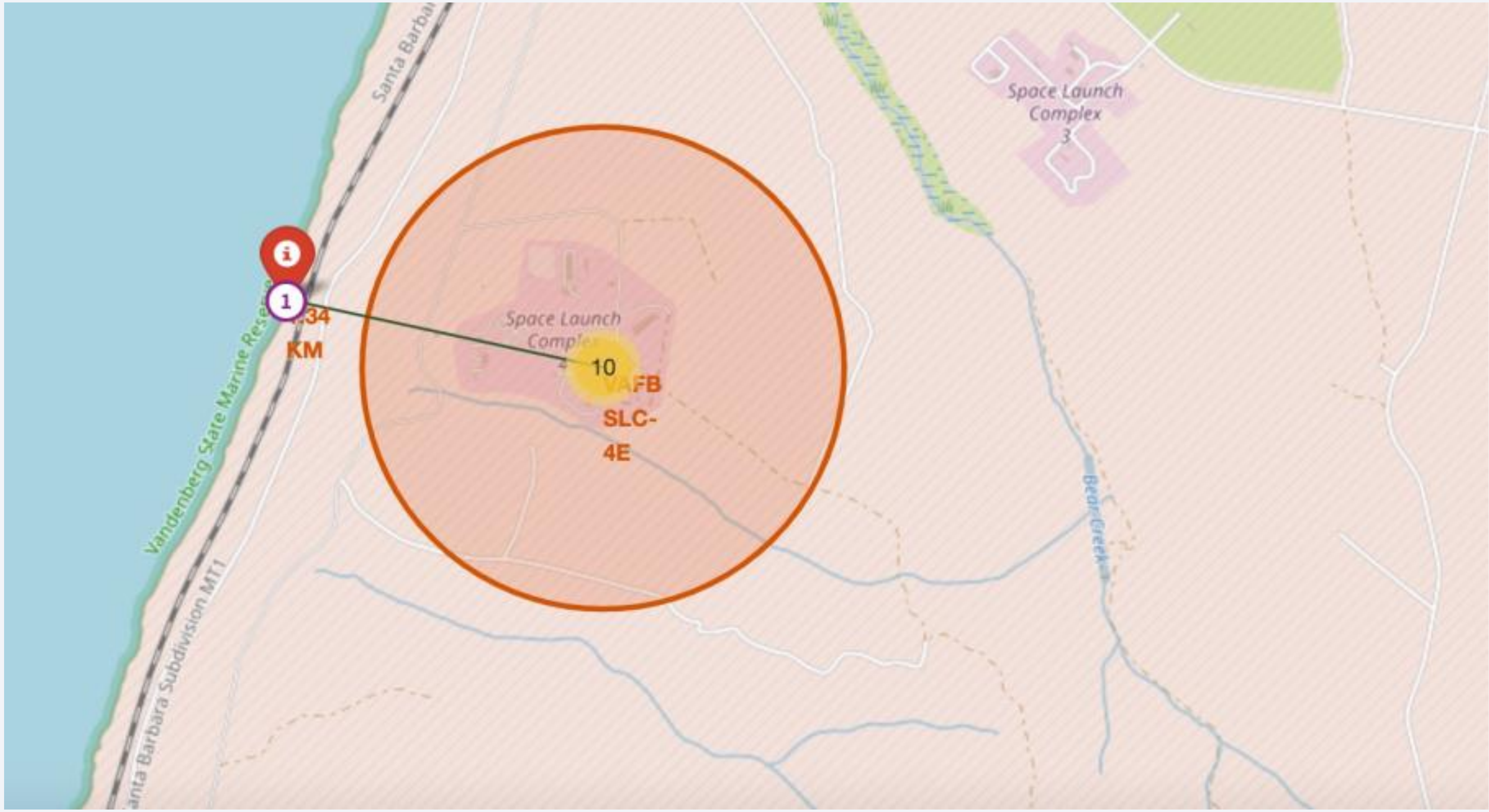
The above map points out that all launch sites are located near a coastline.

Successful and failure launches by site


Launch and Landing Facility
Cochran Cove
Banana Creek
Kennedy Parkway North
Merritt Island National Wildlife Refuge
Saturn Barge Channel
66
CCAFS
LC-40
Leaflet | © OpenStreetMap

The above map shows that for KSC LC-39A site, with 26 launches, most are successful (in green)

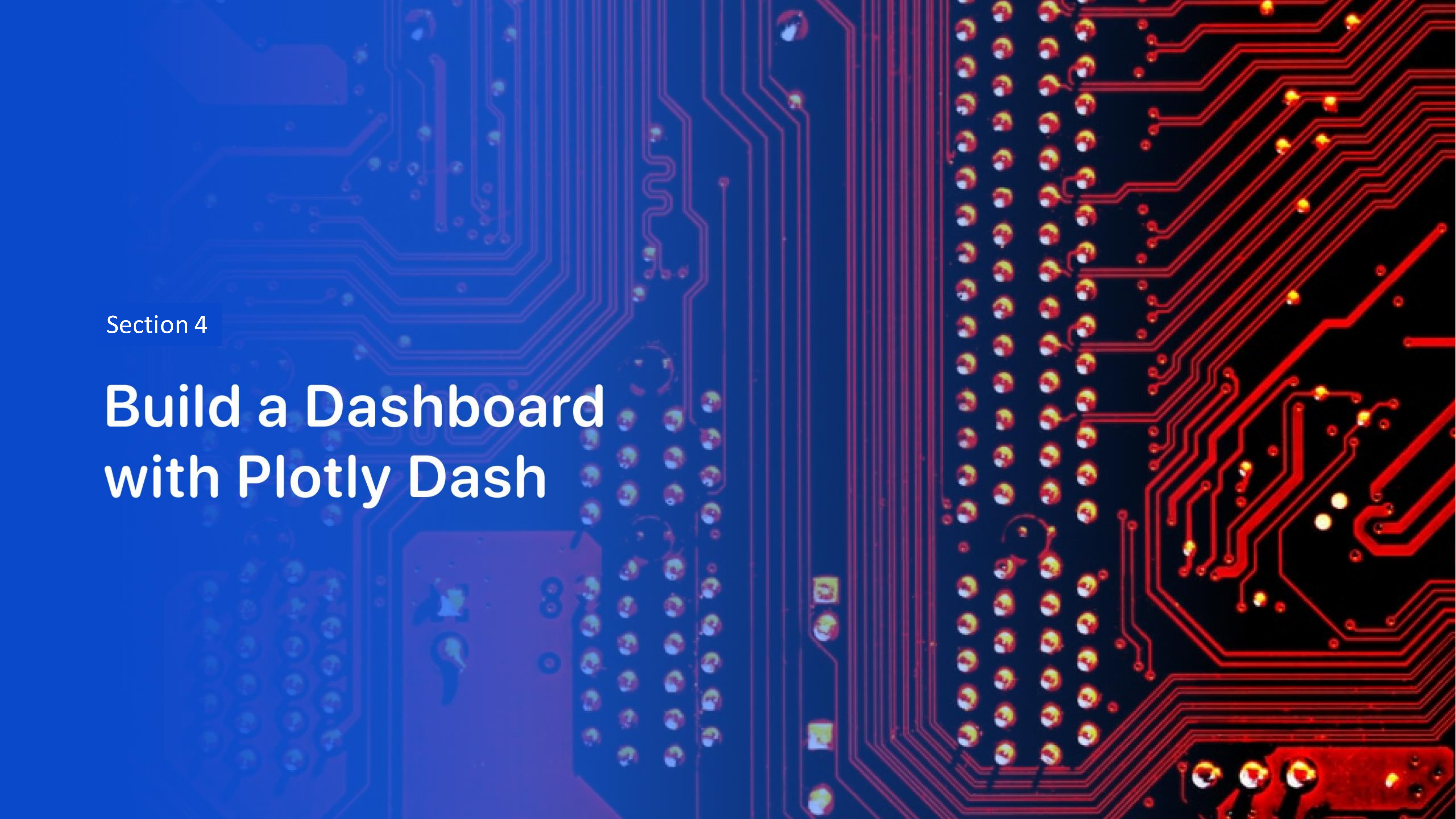
Sites proximity to coastlines



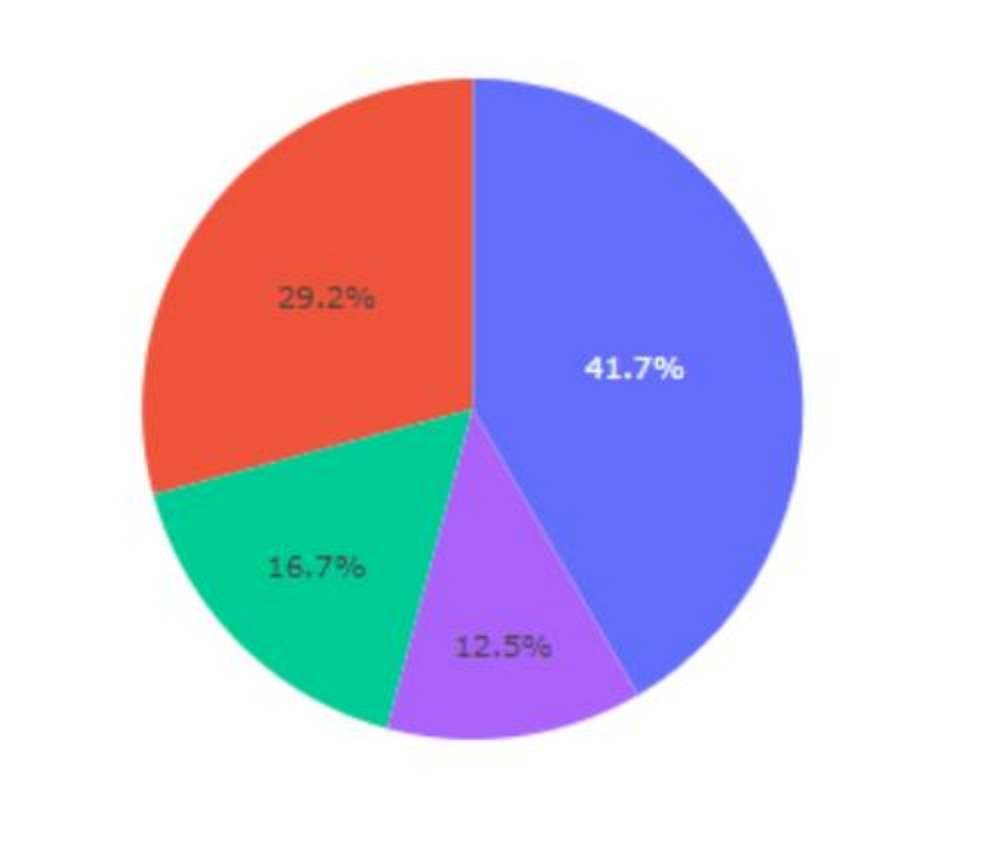
The above map shows that VAFB SLC-4E launch site is 1.34 Km far from a coastline

Section 4

Build a Dashboard with Plotly Dash

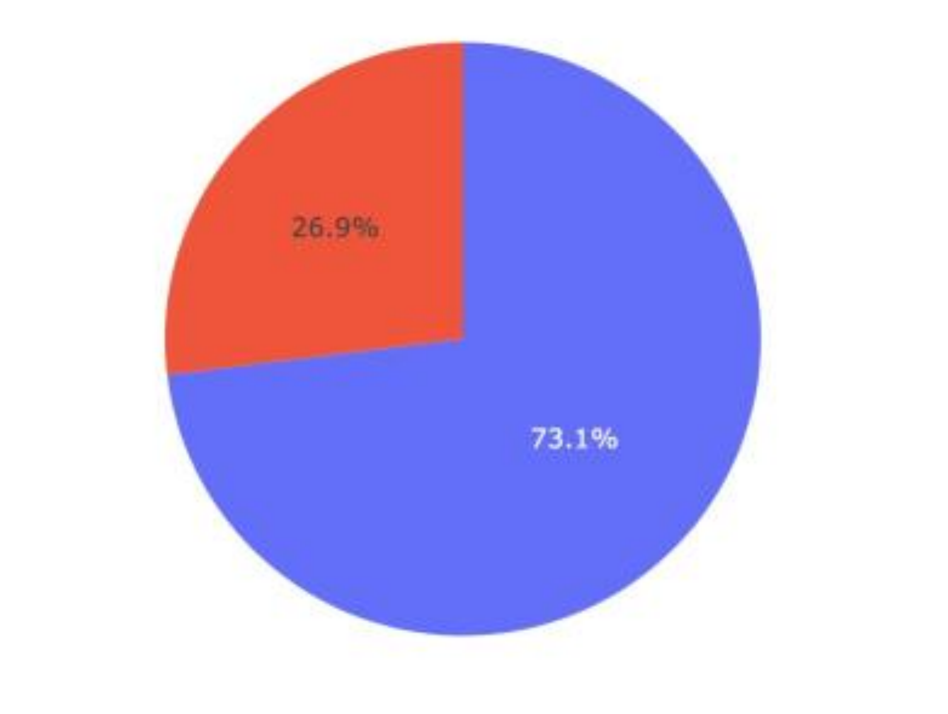


Total success launches by site



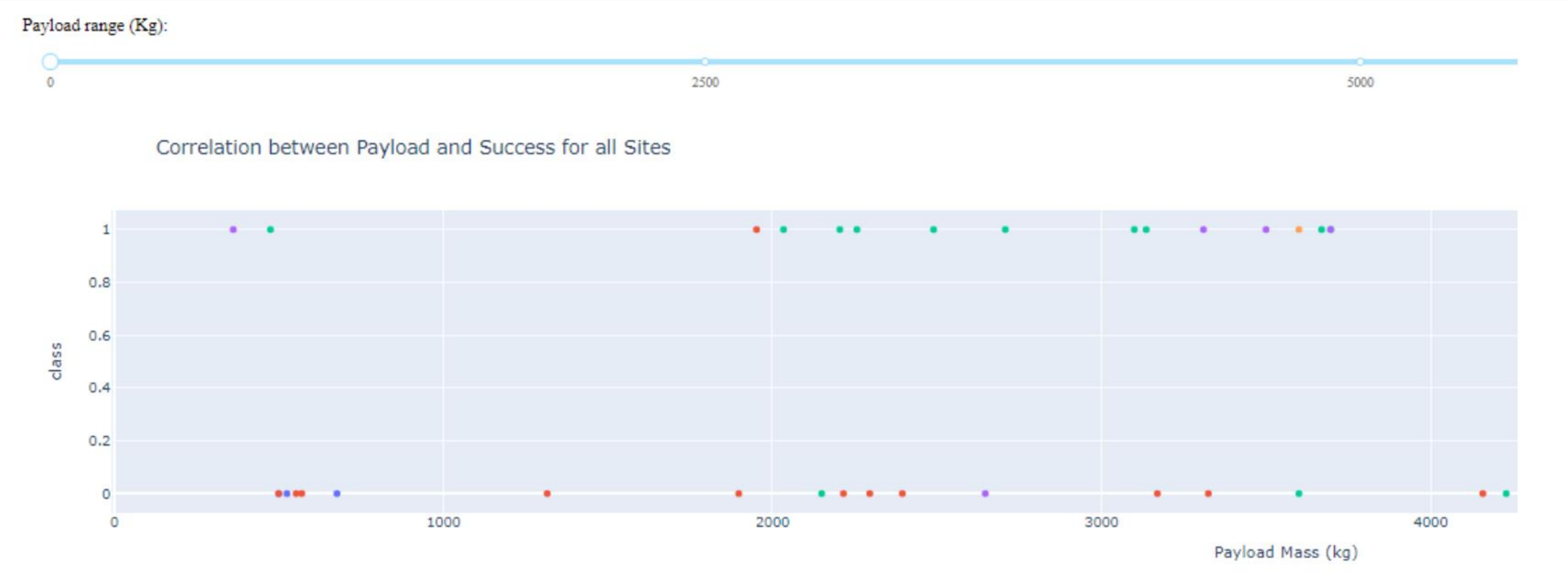
The pie chart shows the successful launches by site with 41.7% the highest ratio

Launches outcome for a specific site



The pie chart shows the success launch ratio for the CCAFS LC-40 site. It is 73.1%

Correlation between payload and launch outcome

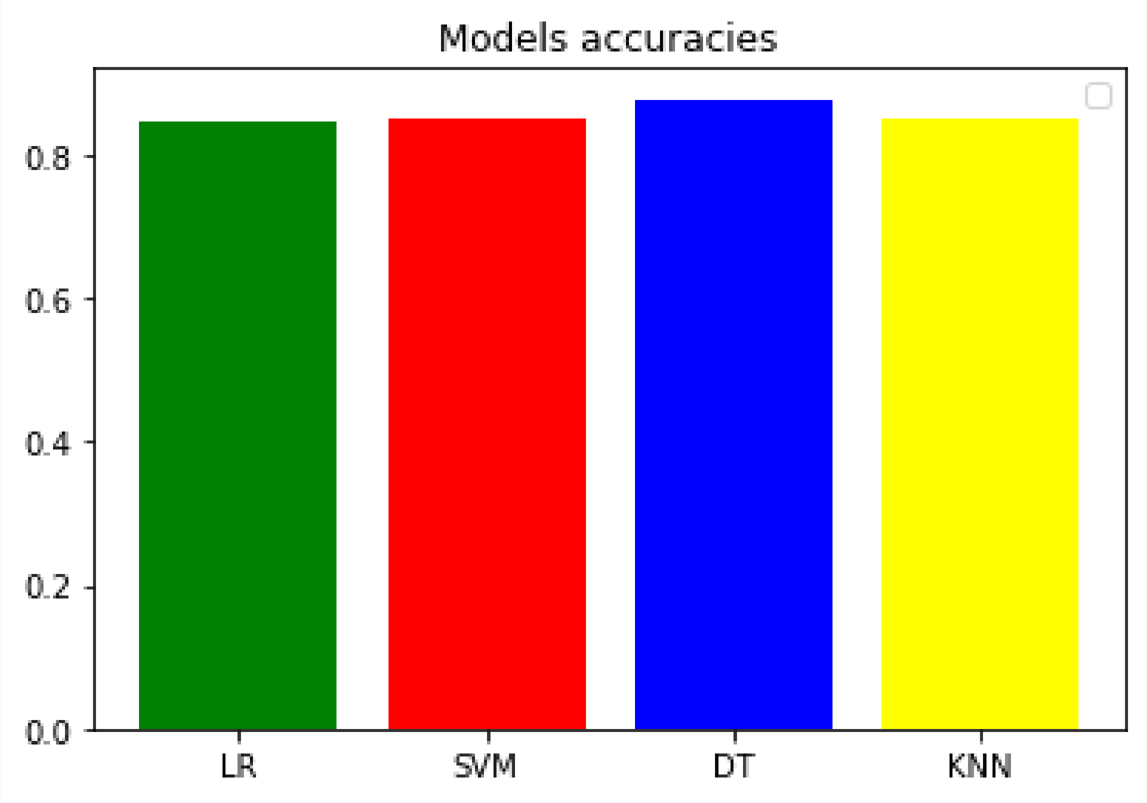


The chart shows correlation between payload and launch outcome for all sites. As the payload increases, CCAFS LC-40 site (in green) gives more successful launches

Section 5

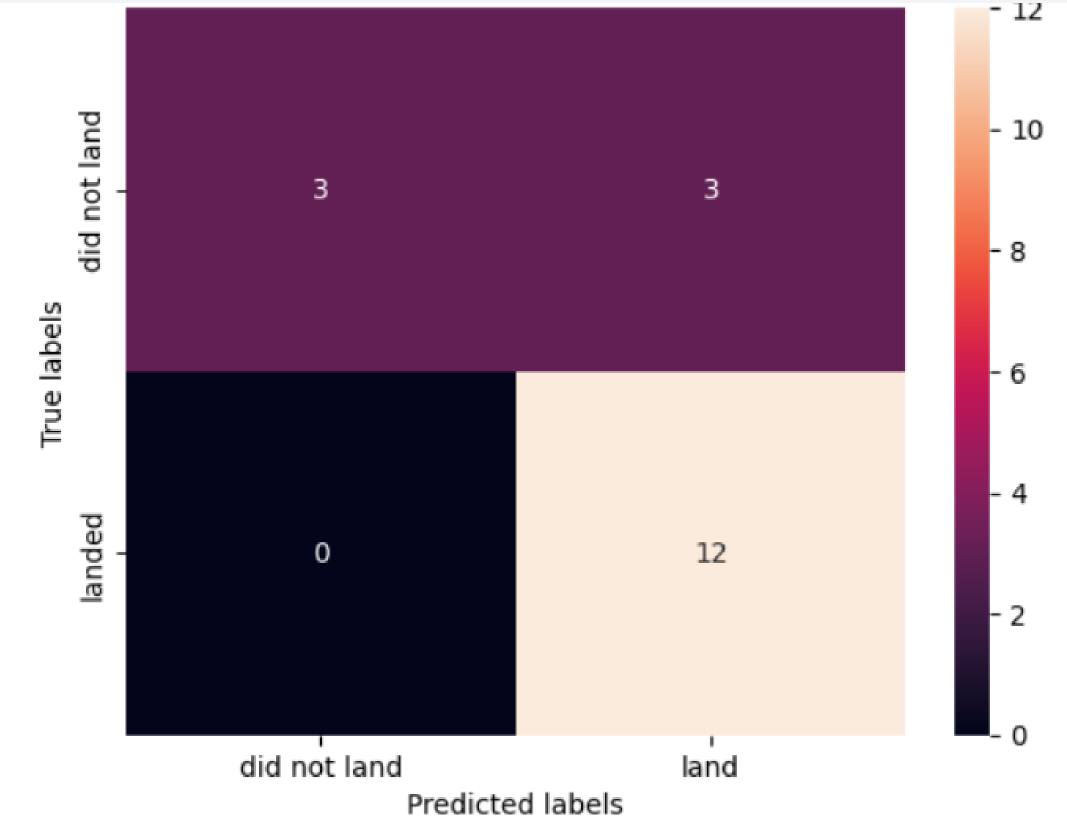
Predictive Analysis (Classification)

Classification Accuracy



The above chart shows the built model accuracy for all built classification models. The decision tree model comes out with the highest accuracy (0.8767).

Confusion Matrix



The chart represents the confusion matrix of the decision tree classifier. The model predicts 3 landing as been successful while they were not (False Positive); this is the only drawback

Conclusions

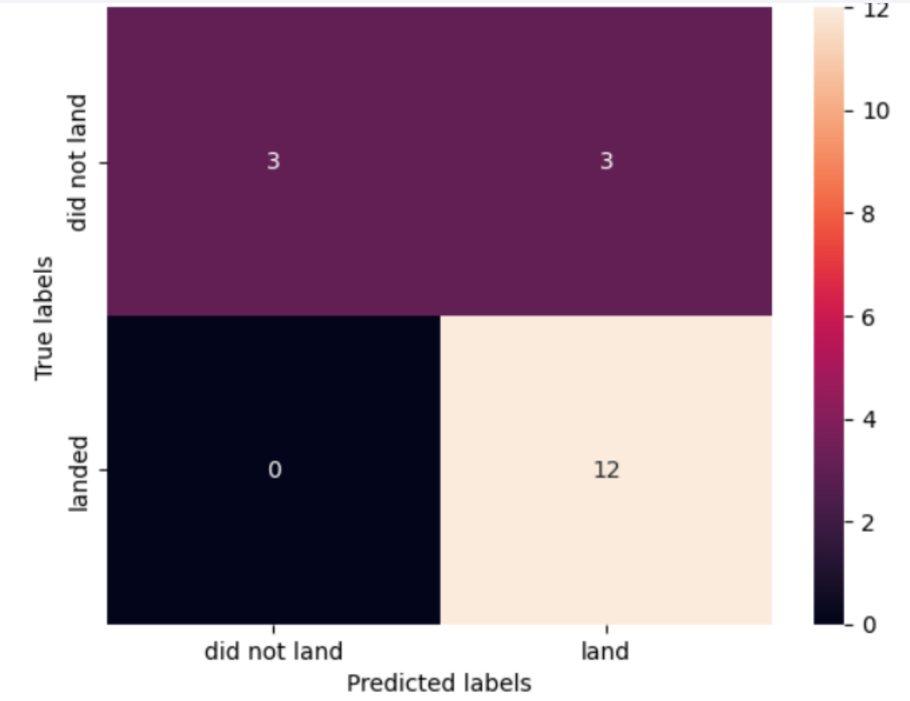
- The objective of this project was to predict the successful landing of the Falcon 9 first stage, a critical factor in SpaceX's cost-effective launch operations.
- Each feature of the Falcon 9 launch, such as the payload mass, the launch site or the orbit type, may affect the mission outcome.
- We collected comprehensive datasets exclusively from the SpaceX API and wikipedia, which provided detailed information on launch parameters such as launch dates, payload mass, launch sites, booster versions, and landing outcomes...
- During the exploratory data analysis (EDA) phase, we examined the collected data to uncover patterns, trends, and relationships among the variables. We identified significant features influencing landing success, such as payload mass, launch site ...
- We developed and tested several machine learning models to predict the successful landing of the Falcon 9 first stage. The models included logistic regression, support vector machine, decision tree, K nearest neighbor. They all came out with the same score and the same confusion matrix during the evaluation phase. We finally keep the best model as the one with the highest accuracy during the cross-validation : the decision tree.

Appendix

Find below some relevant assets for this project

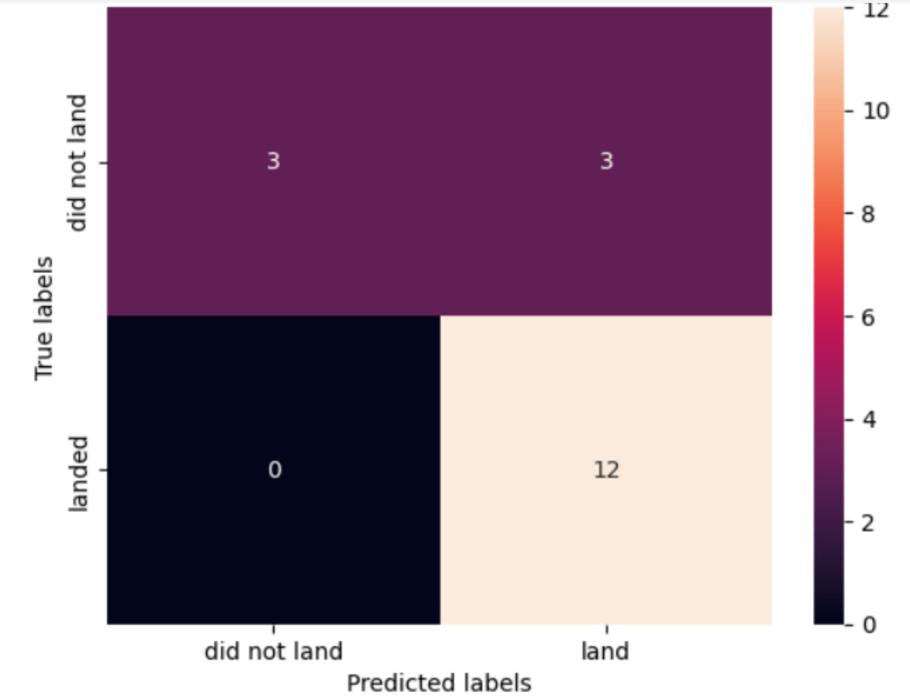
Predictive Analysis (Classification)

- Logistic regression
  - Tuned hyperparameters :(best parameters) {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
  - Accuracy (Cross validation best score) : 0.8464285714285713
  - Score : 0.8333333333333334
  - Confusion matrix :



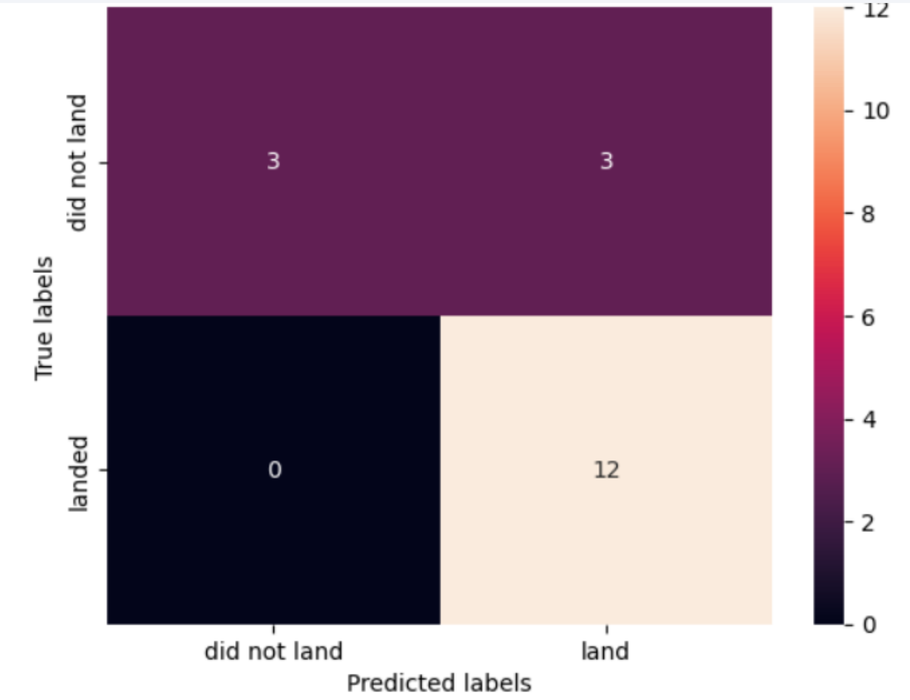
Predictive Analysis (Classification)

- Support vector machine
  - Tuned hyperparameters : (best parameters) {'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'}
  - Accuracy (Cross validation best score) : 0.8482142857142856
  - Score : 0.8333333333333334
  - Confusion matrix :



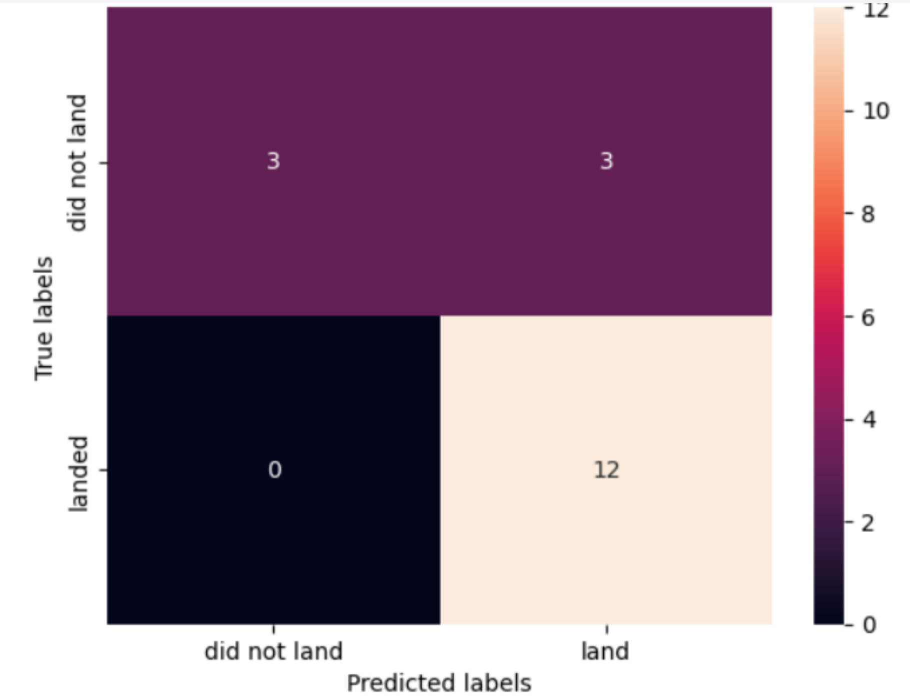
Predictive Analysis (Classification)

- Decision tree classifier
  - Tuned hyperparameters : (best parameters) {'criterion': 'entropy', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5, 'splitter': 'best'}
  - Accuracy (Cross validation best score) : 0.8767857142857143
  - Score : 0.8333333333333334
  - Confusion matrix :



Predictive Analysis (Classification)

- K Nearest Neighbors classifier
  - Tuned hyperparameters : (best parameters) {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
  - Accuracy (Cross validation best score) : 0.8482142857142858
  - Score : 0.8333333333333334
  - Confusion matrix :



Thank you!