Ad Astra

A data-driven Conquest of Space

By Jordi González de Regàs
February, 2023

Presentation Outline

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

Executive Summary

- Data from Wikipedia (check Annex) has been scraped using SpaceX API and BeautifulSoup.
- The data has been cleaned using Pandas and Numpy and wrangled.
- EDA has been performed through SQL queries and variable dependencies have been determined through Data Visualization techniques using pyplot and seaborn.
- Mission outputs have been classified by location using Folium.
- A Dashboard of the data using piecharts in order to determine success score per launch site and success rate according to payload has been developed.
- Machine learning algorithms have been tested by predictability using Grid Search methods.
- The research has concluded that KNN was the most suitable ML technique, only showing false positives, and that launch sites are located close to the Equator and the coast, far from cities and connected by railroad. Also, success rates have escalated drastically since 2013.

Introduction

- SpaceX is the main comercial spaceflight company globaly.
- Thanks to its reusable rocket technology, cost-of-flight has reduced significantly.
- Smaller payloads can now be launched with reusable rockets, which has democratized access to lower-to-medium Earth orbits.
- Since the Falcon 9 reached technological maturity, SpaceX has had great success and demonstrated reliability both in launch and recovery operations.
- This report aims at analyzing the Falcon 9 missions carried out throughout the last decade and building a ML model that predicts the outcome of the next launch.
- Variables such as launch site, payload, orbit, flight number and date of launch are to be taken into account.

Section 1

Methdology

Methodology

Executive Summary

- Data collection methodology:
  - Through webscraping of the 'List of Falcon 9 and Falcon Heavy Launches' [1] dataset from Wikipedia
- Perform data wrangling
  - Describe how data was processed After studying the datatypes and cleaning the missing values, the information regarding landing outcome was converted into a dummy variable.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Several classification models (Tree Regression, KNN, Logistic Regression, SVM) were tested and optimized using Grid Search.

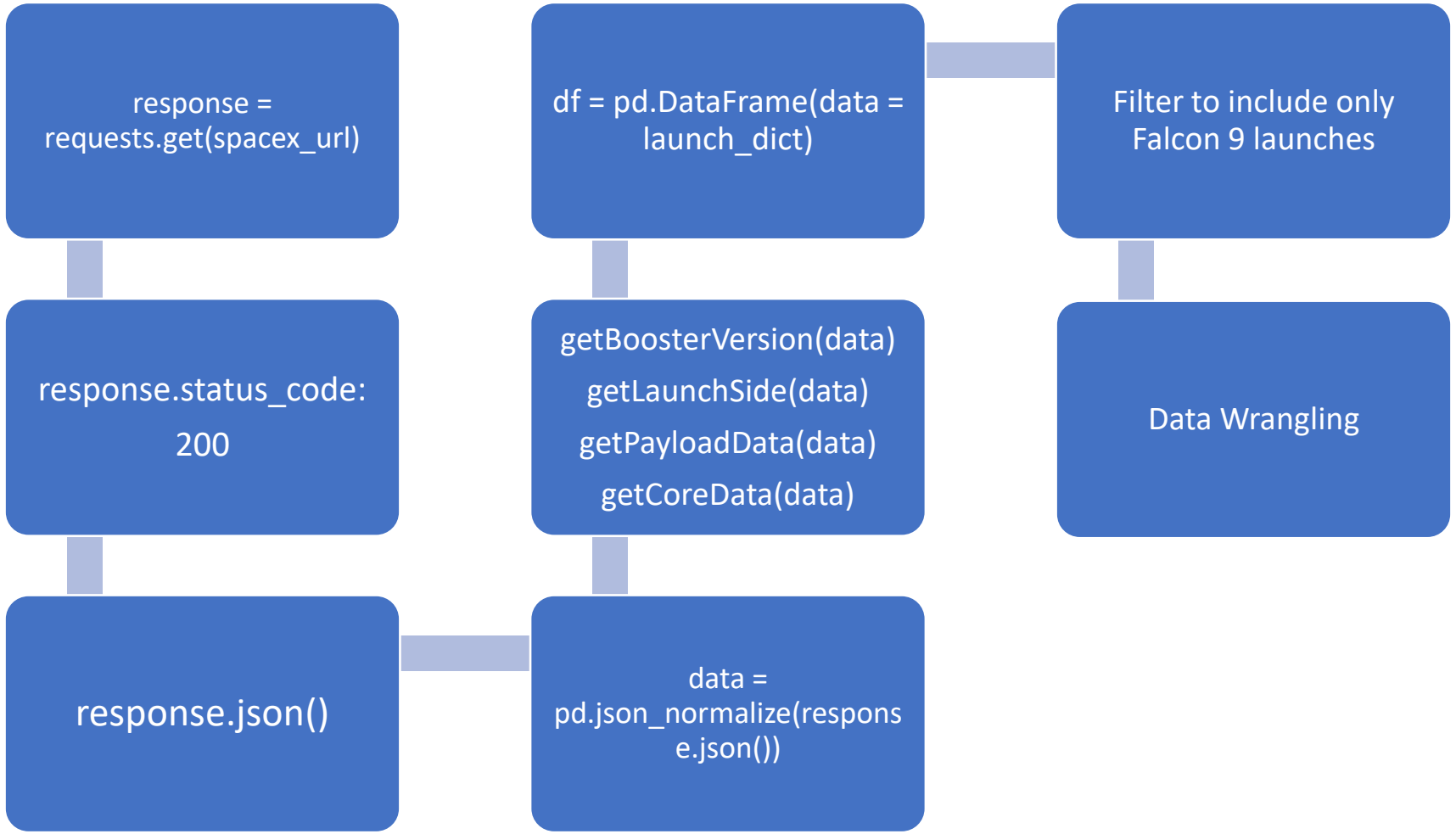
Data Collection using SpaceX API

Using the SpaceX API, a request to retrieve the data from [1] was passed.

The functions getBoosterVersion, getLaunchSite, getPayloadData, getCoreData,
Simplified the data collection of the variables with the same name.

Afterwards, the data was filtered to include only Falcon 9 launches.

Finally, the null elements of Landing_pad were replaced with the average.

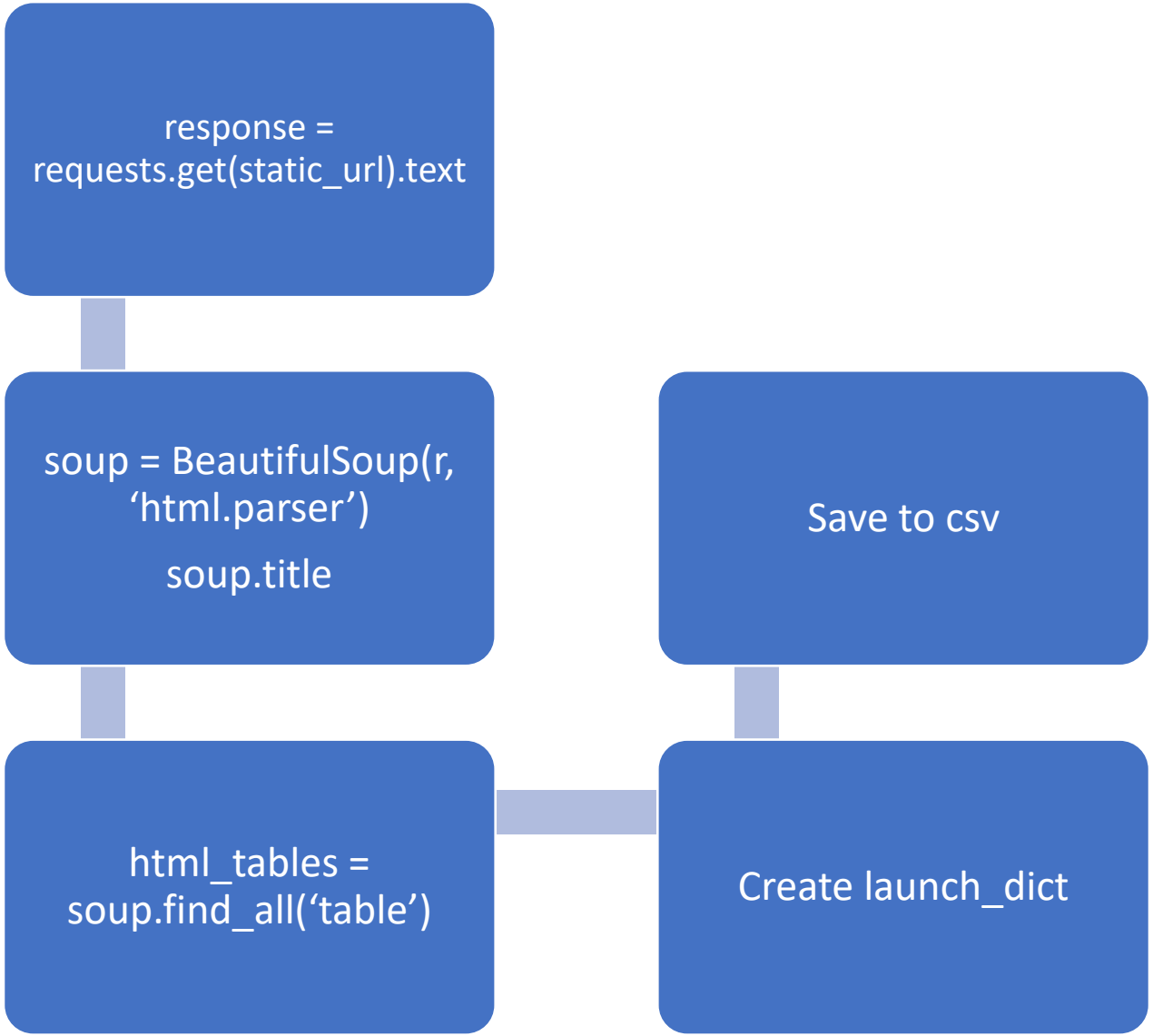


[Link to Notebook on GitHub]

Data Collection using BeautifulSoup

Using the BeaufiulSoup library, a soup object was created to retrieve the data from [1].
The functions date_time, booster_version, landing_status, get_mass, extract_column_from_header simplified the data collection.

Using the .find_all() function, the tables were parsed and iterating through the rows, the data was extracted and stored into a dataframe.



[Link to Notebook on GitHub]

Data Wrangling

- Determination of percentage of missing values per category and data types.
- Use value_counts() to get number of launches per site and occurrences per each type of orbit.
- Classify landing outcomes and convert it to a dummy variable, with 1 being the successful landings and 0 the failed landings.
- Assign them to variable ‘Class’ and export the dataframe.

Link to Notebook on GitHub

EDA with SQL

Queries Performed:

- Names of the unique launch sites of the mision.
- Display 5 entries where location began by ‘CCA’.
- Display total payload mass carried by boosters launched by NASA (CRS).
- Display average payload mass carried by booster version F9 v1.1.
- Date of the first successful landing in groundpad.
- Names of boosters with successful landing in drone ship and 4000 < mass < 6000.
- List of booster versions with máximum payload mass.
- List of total number of successful and failure misión outcomes.
- Names of booster versions that carried máximum payload mass.
- Months, failures in drone ship, booster versions and launch site for months in 2015.

Link to Notebook on GitHub

Build an Interactive Map with Folium

- All launch sites from spaceX were marked using folium.marker and folium.cirlce.
- Successful launches on each launch site were marked on green and failed launches on red using folium.circle, folium.marker and grouped using folium.marker_cluster.
- Distances from launch site to the coast were calculated and lines joining them were drawn using folium.PolyLine and the same was done with railroads, cities…
- This additions were made in order to show that all launch sites are close to the Equator and the coastline and well connected through railroads. This is to maximize launch efficiency, safety and optimize launch operation logistics.

Link to Notebook on GitHub

Build a Dashboard with Plotly Dash

- Using Dash, a dropdown menu was added with dcc.Dropdown, that allowed for launch site selection.
- A callback function that created a pie chart showing success rates per launch site if all launch sites were selected and the percentage of successes and failures if a single launch site was selected using px.pie function.
- A slider that enables the user to select payload range was added using dcc.RangeSlider object.
- A second callback function with the launch site and payload from range slider as inputs outputted a scatter plot (px.scatter) that showed the success rate of each booster category.

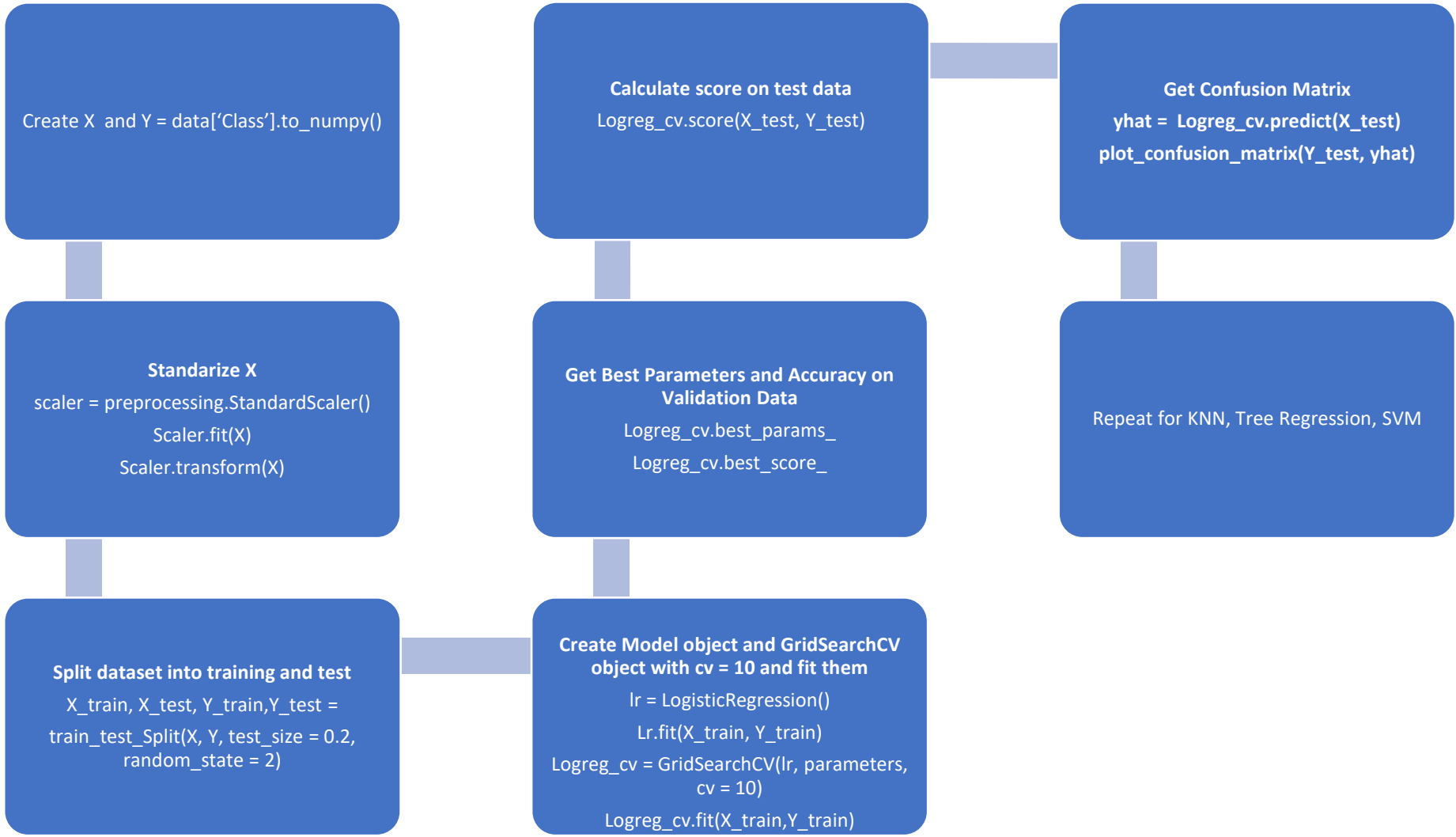
Link to Notebook on GitHub

Predictive Analysis (Classification)

Several models (KNN, Logistic, Tree, SVM) were tested in order to predict whether the next launch will be successful using the 'Class' variable as target.

First, the variable X was standarized, then the dataset was split into training and testing datasets.

ML models were trained and optimized using Grid Search methods.



[Link to Notebook on GitHub]

Results

- EDA concluded that there is a relationship between flight number and launch site and launch success and year.
- Dashboards concluded that KSC LC-39A is the most successful launch site and payloads up to 5400kg report a 100% success rate.
- Predictive analysis concluded that KNN and SVM were the best models for the task, with KNN set to 10 neighbors, p = 1, and auto algorithm, and SVM using a sigmoid kernel, C=1 and Gamma = 0.03.

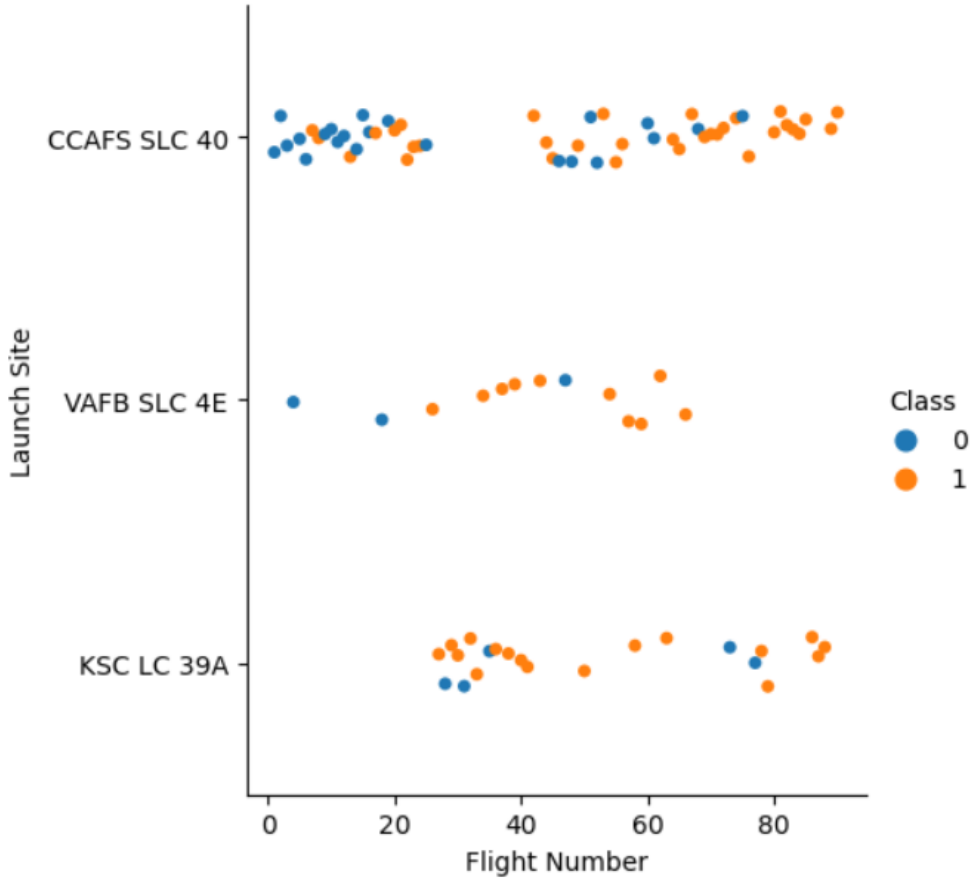
Section 2

Insights from EDA

Flight Number vs Launch Site

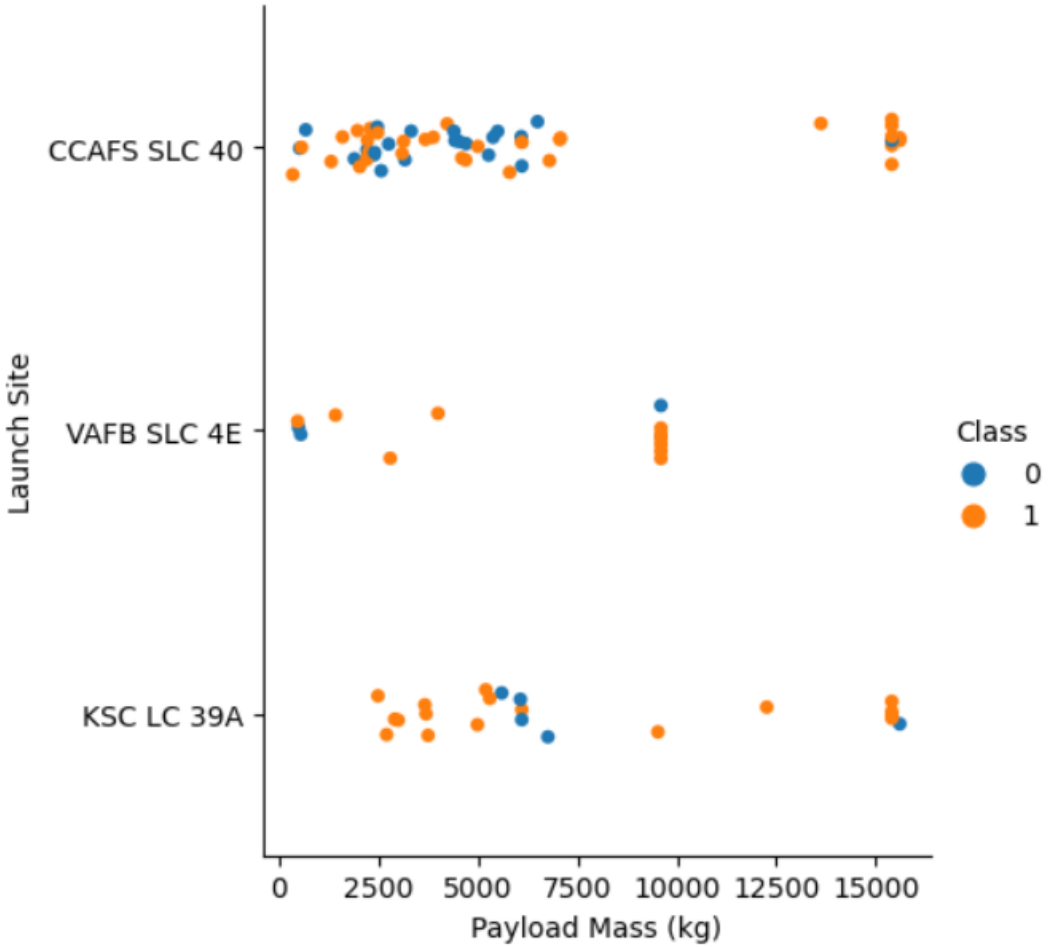
As Flight Number increases, more landings are successful (we see more orange dots indicating class 1 than blue ones, indicating class 0).
The first 20 launches were mostly failed attempts which took place in CCAFS SLC 40, while the latest flights take place mostly in KSC LC 39A and CCAFS SLC 40.

This explains the difference in success ratio per launchsite: The earliest attempts were mostly conducted at the same facility, and as SpaceX mastered the landing technology, the demand increased, and the company diversified its launchsite platforms in order to perform more launches simultaneously.



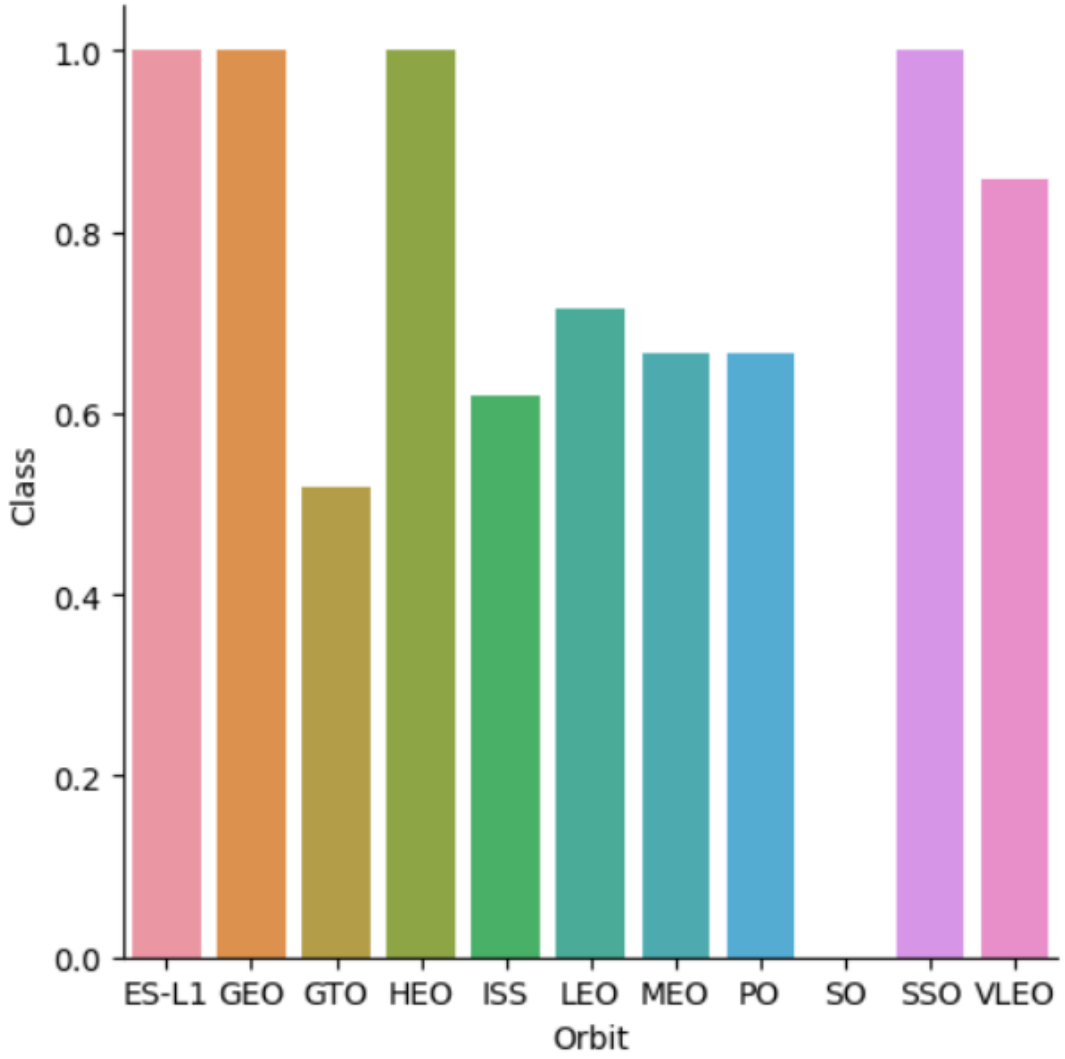
Payload vs Launch Site

for the VAFB-SLC launchsite there are no rockets launched for heavy payload mass (greater than 10000).



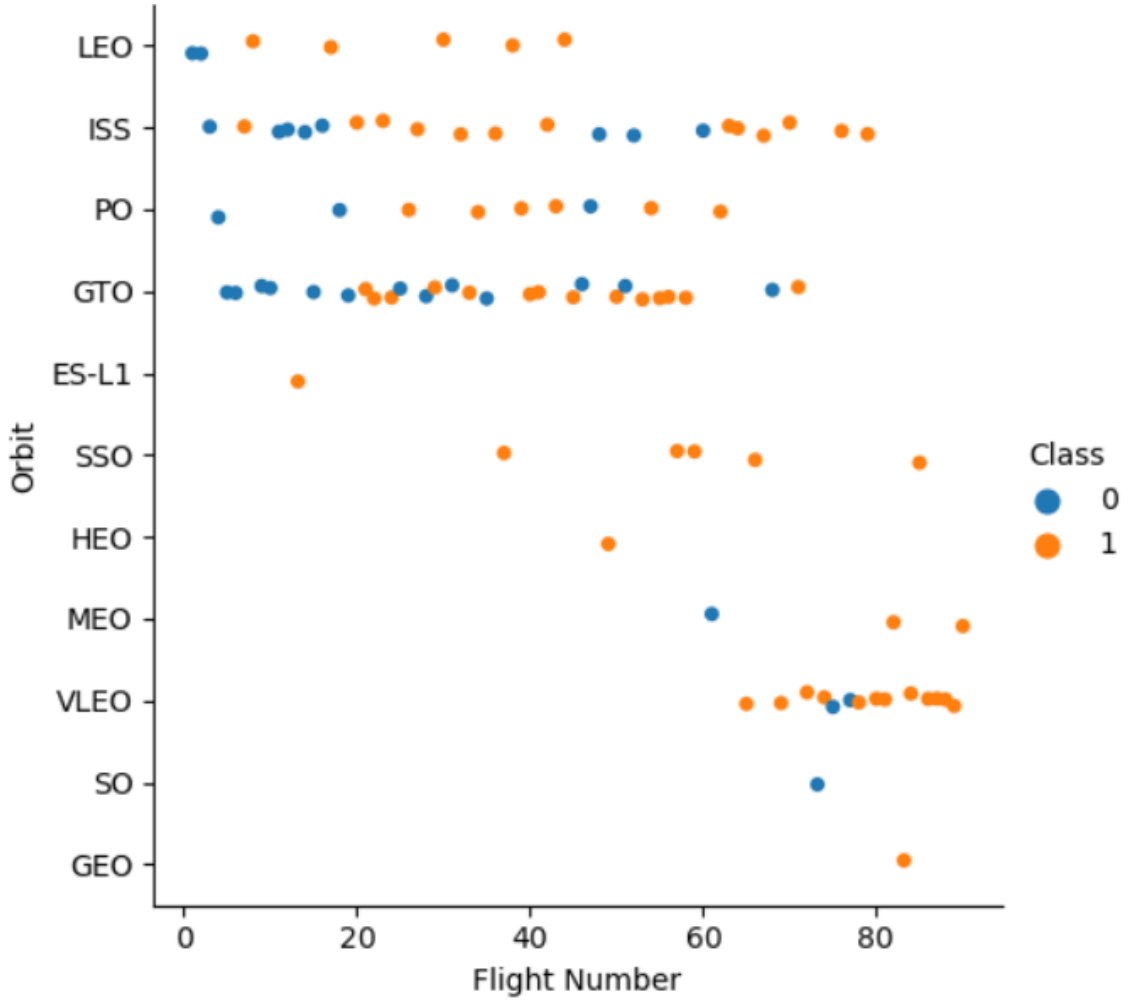
Success Rate vs Orbit Type

ES-L1, GEO, HEO, SSO all have success rates of 1.0, while PO, MEO and ISS have success rates of around 70%, and GTO has a success rate for 50%. There is no data for SO.



Flight Number vs. Orbit Type

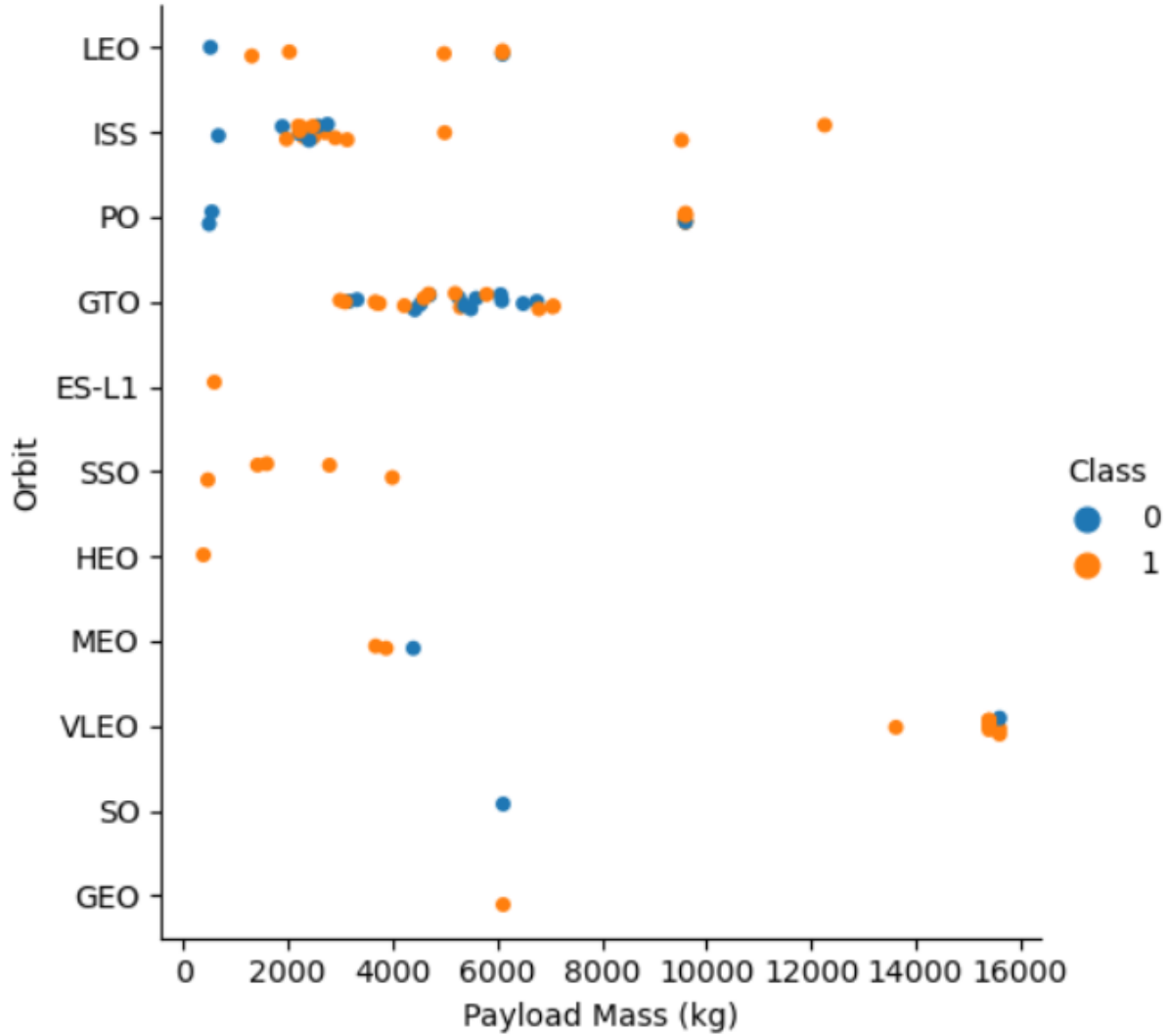
In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

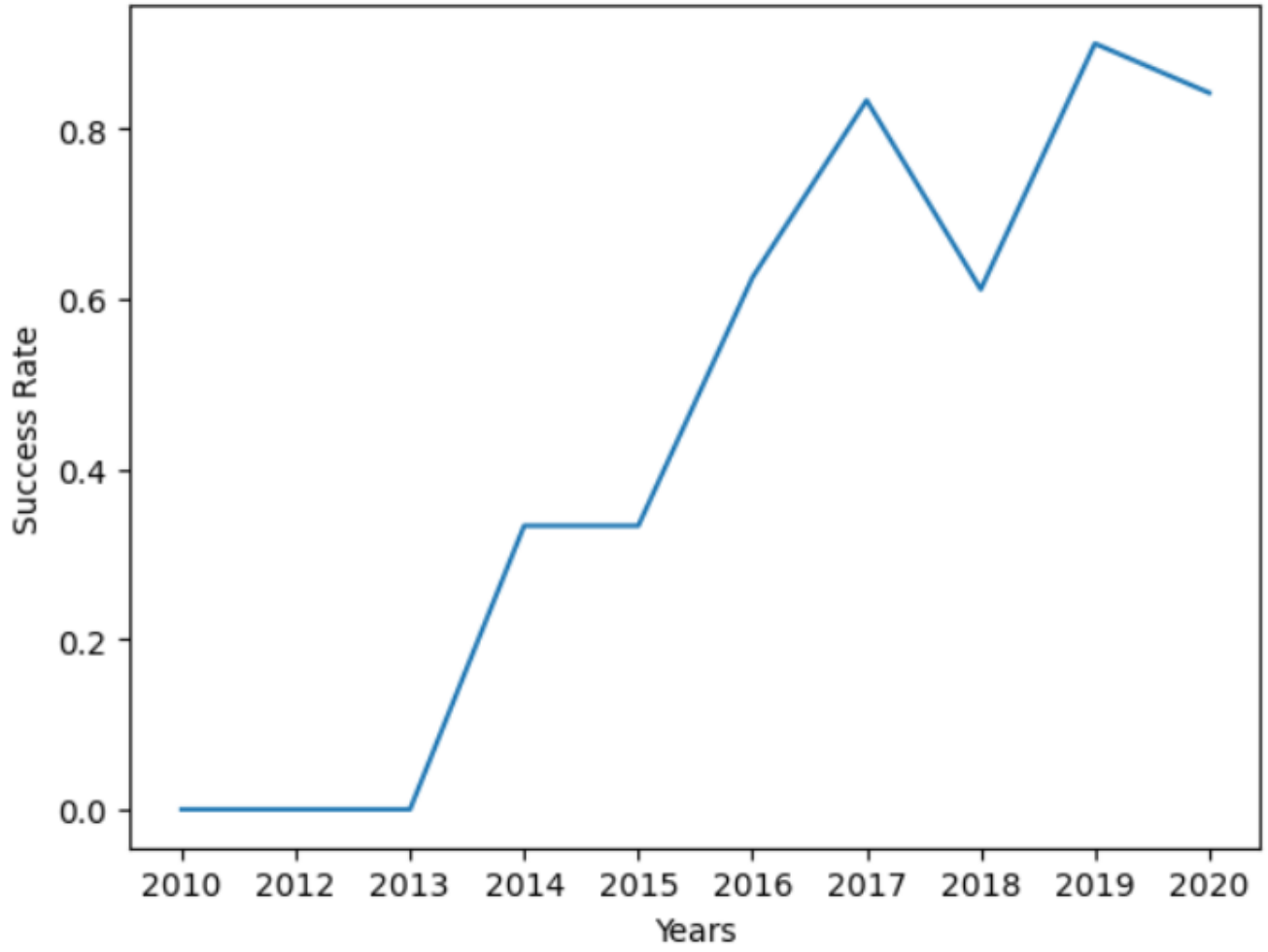
With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

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



Launch Success Yearly Trend

Success rate climbs from 2013 onwards, with a 20% downfall in 2018.



All Launch Site Names

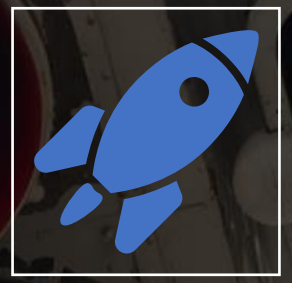
The launch site names are:

- CCAFS LC-40
- KSC LC-39A
- CCAFS SLC-40

Launch Site Names Begin with 'CCA'

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

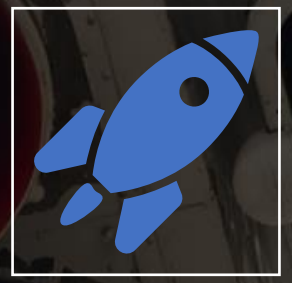
Total Payload Mass



Total Payload mass carried by NASA boosters

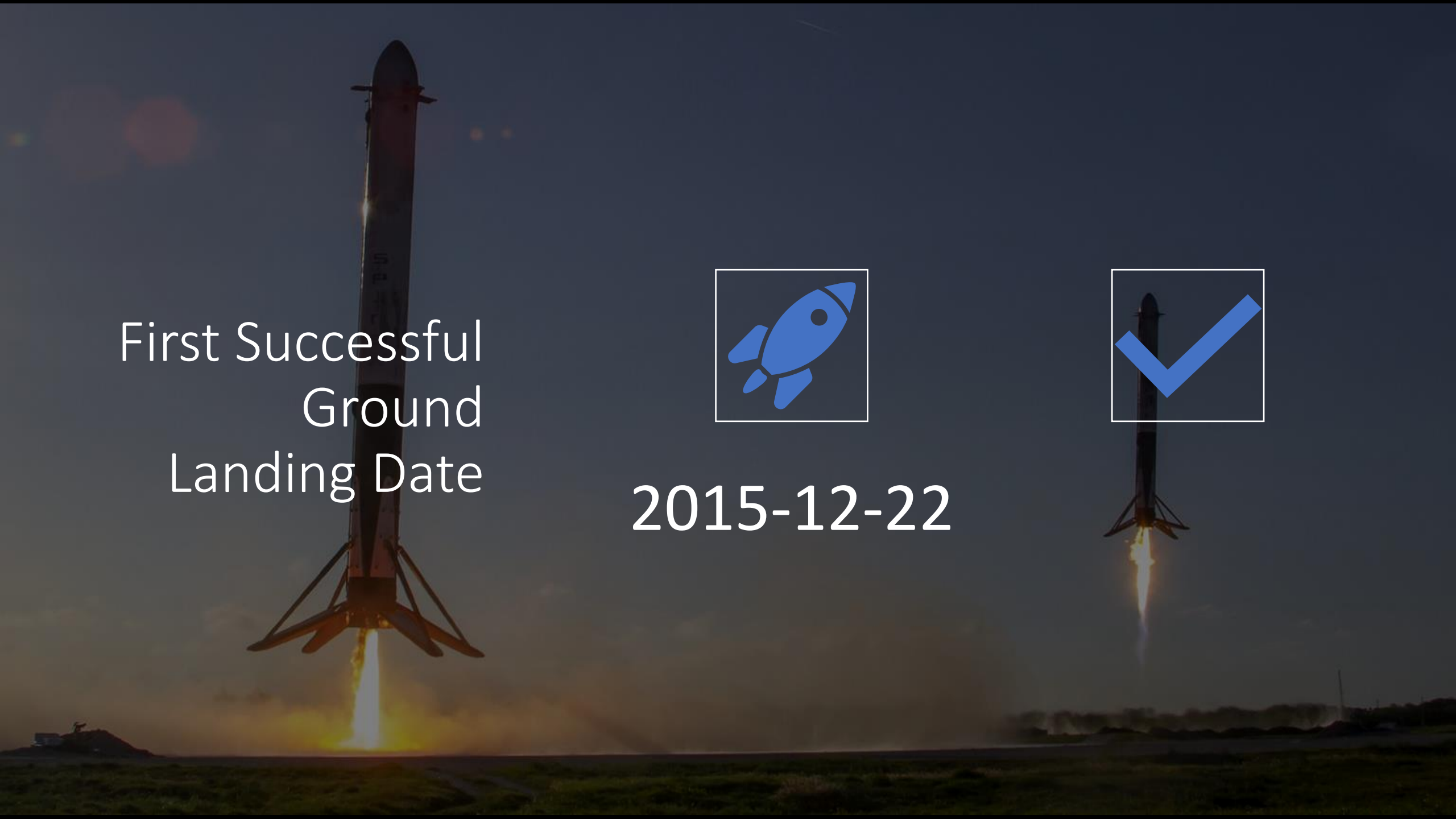
45596 kg

Average Payload Mass by F9 v1.1



Total Payload mass carried by NASA boosters

2928.4 kg



First Successful Ground Landing Date

2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000 kg

'F9 v1.1', 'F9 v1.1 B1011', 'F9 v1.1 B1014', 'F9 v1.1 B1016',
'F9 FT B1020', 'F9 FT B1022', 'F9 FT B1026', 'F9 FT B1030',
'F9 FT B1021.2', 'F9 FT B1032.1', 'F9 B4 B1040.1', 'F9 FT B1031.2',
'F9 B4 B1043.1', 'F9 FT B1032.2', 'F9 B4 B1040.2', 'F9 B5 B1046.2',
'F9 B5 B1047.2', 'F9 B5B1054', 'F9 B5 B1048.3', 'F9 B5 B1051.2 ',
'F9 B5B1060.1', 'F9 B5 B1058.2 ', 'F9 B5B1062.1'

Total number of Successful and Failure Mission Outcomes

Success : 98

Failure (in flight) : 1

Success (payload status unclear) : 1

Success : 1

Boosters Carried Maximum Payload

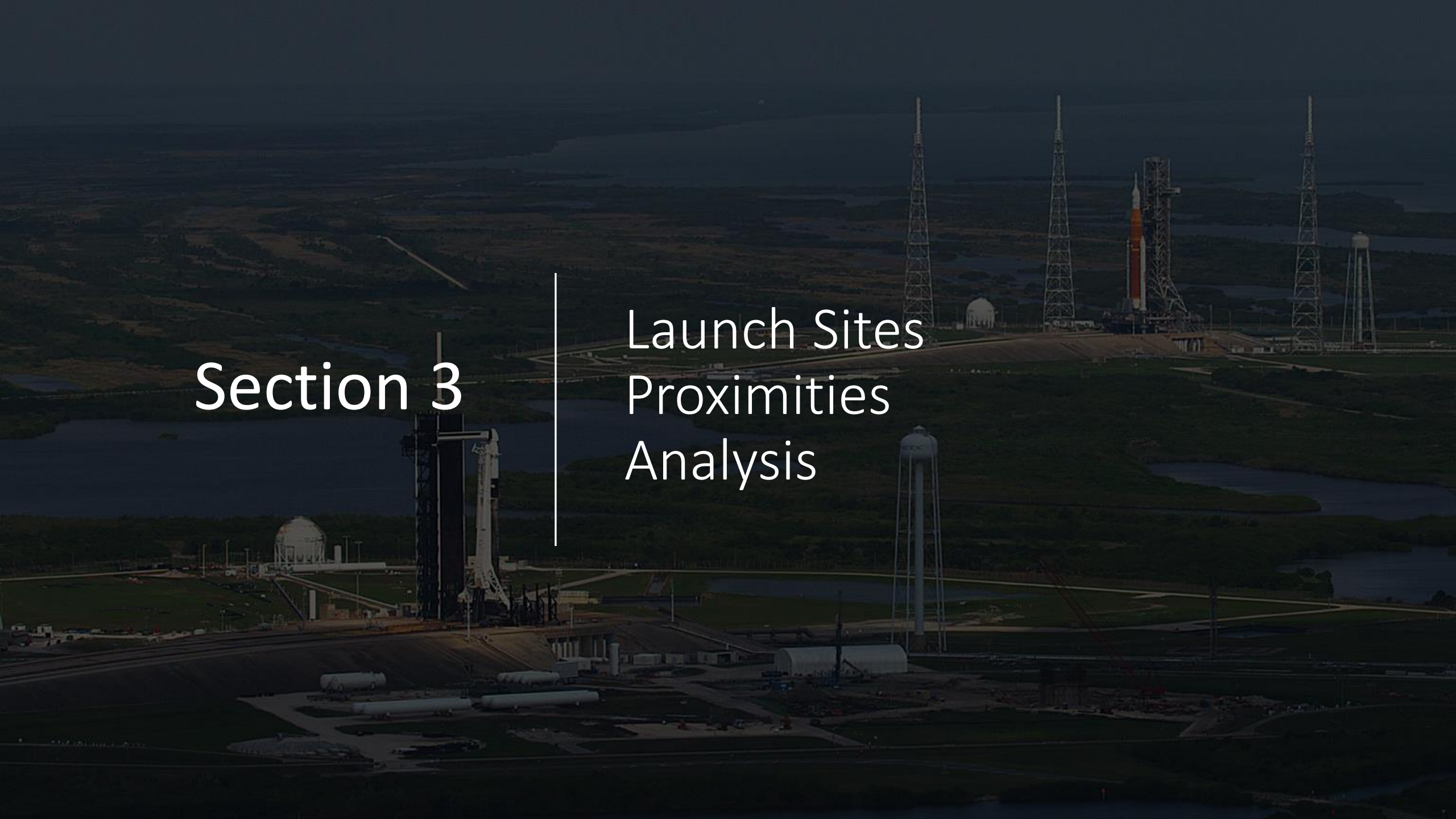
'F9 B5 B1048.4', 'F9 B5 B1049.4', 'F9 B5 B1051.3', 'F9 B5 B1056.4', 'F9 B5 B1048.5', 'F9 B5 B1051.4', 'F9 B5 B1049.5', 'F9 B5 B1060.2 ', 'F9 B5 B1058.3 ', 'F9 B5 B1051.6', 'F9 B5 B1060.3', 'F9 B5 B1049.7 '

2015 Launch Records

	Landing _Outcome	Booster_Version	Launch_Site
13	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
16	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

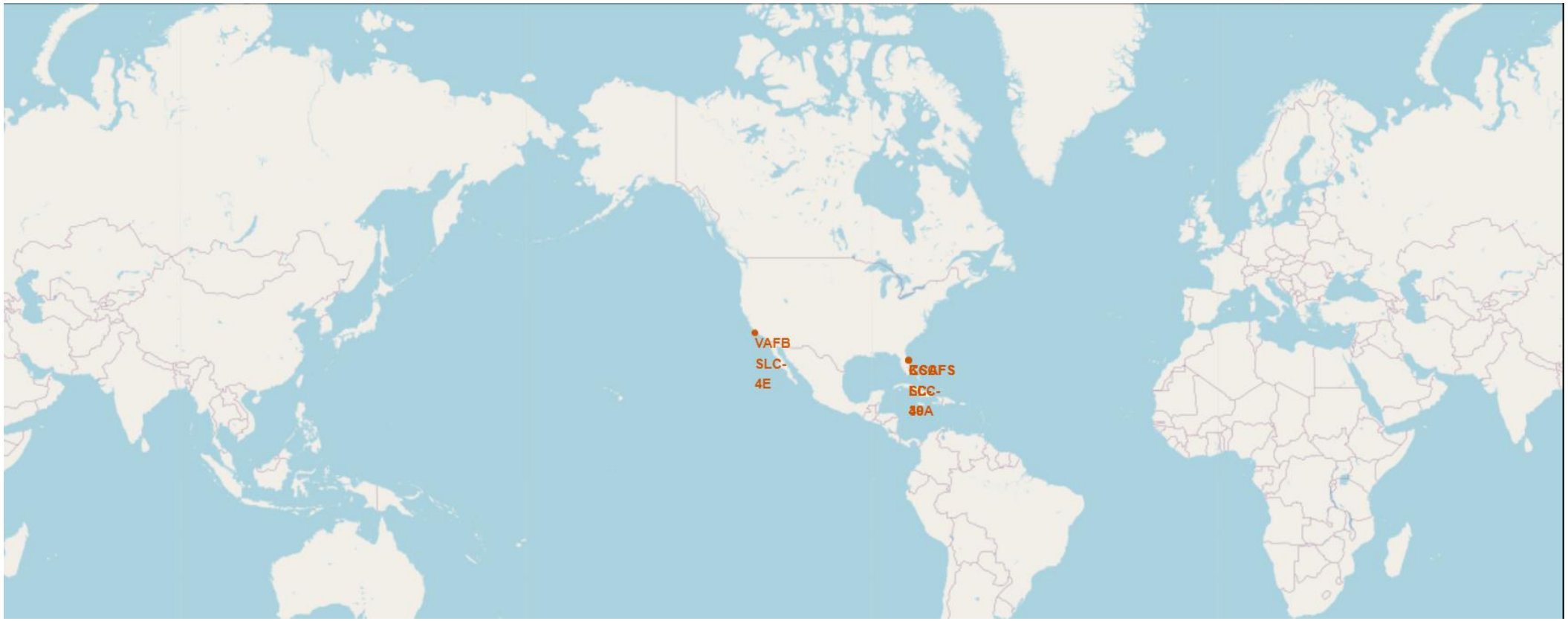
	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
30	2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
34	2017-03-06	21:07:00	F9 FT B1035.1	KSC LC-39A	SpaceX CRS-11	2708	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
29	2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
28	2017-01-14	17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)
32	2017-01-05	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
27	2016-08-14	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
22	2016-08-04	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
26	2016-07-18	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
25	2016-06-15	14:29:00	F9 FT B1024	CCAFS LC-40	ABS-2A Eutelsat 117 West B	3600	GTO	ABS Eutelsat	Success	Failure (drone ship)
23	2016-06-05	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)



Section 3

Launch Sites Proximities Analysis

Launch Site Locations



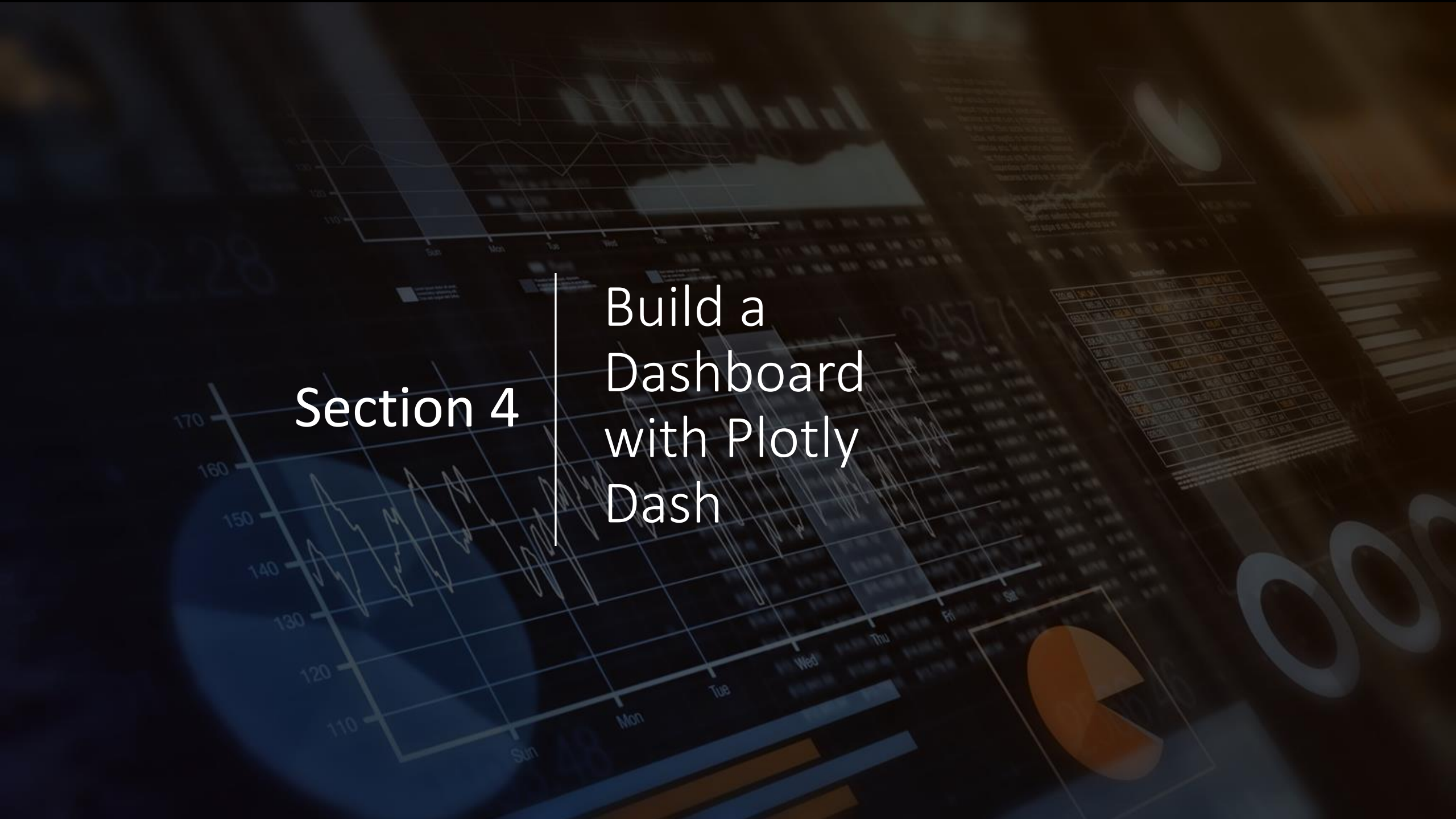
In red we can see the locations of the launch sites. They are all close to the coast and near the Equator to minimize launch costs and to maximize the likelihood of any potential debris falling into de ocean, far away from populated areas.

Launch Results According to Location

In red we can see the failed launches that took place in each site, while the successful launches are displayed in green.

Launch Results According to Location

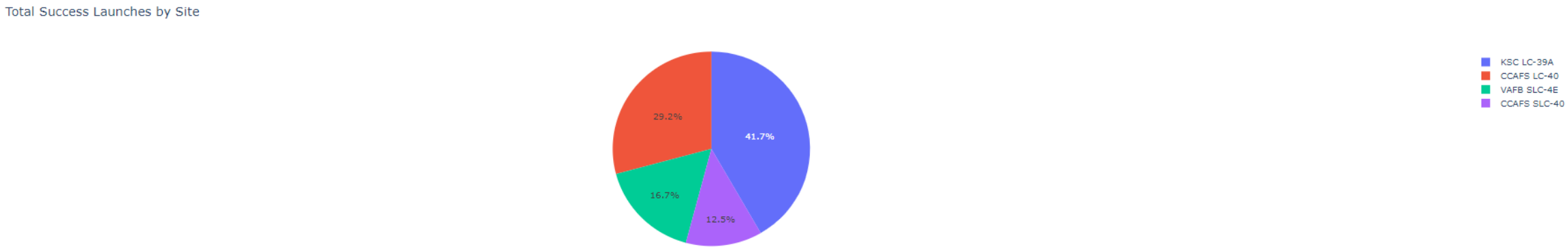
The blue line shows the distance from the launch site to the coast. The distance is displayed in red. From this map, we can infer that all the launch sites are close to the coast and railroads, in order to maximize launch safety and efficiency in transport logistics.



Section 4

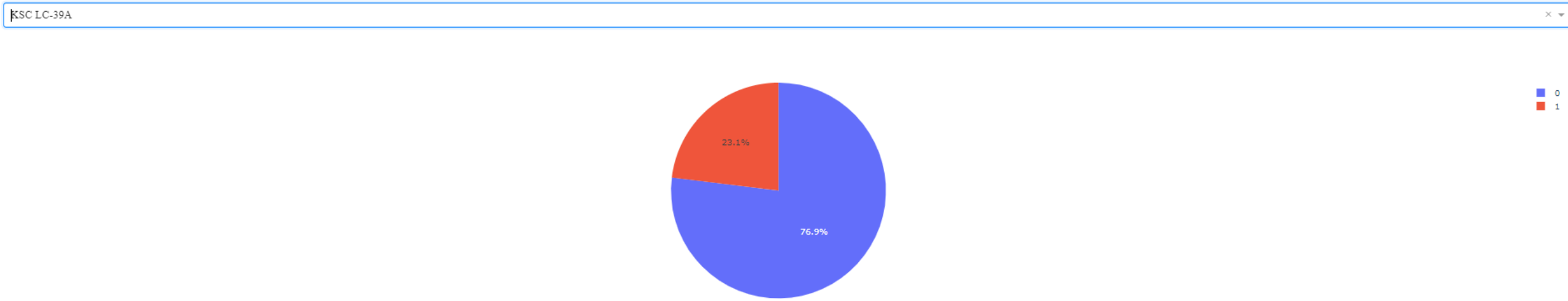
Build a Dashboard with Plotly Dash

Launch Success Count per Site



This screenshot contains a piechart showing the Total success Launches per Site.
As you can see, most successes come from KSC LC-39A, from which comes 41.7% of the total registered success, while CCAFS SLC-40 has the least successful count, representing only 12.5% of the total.

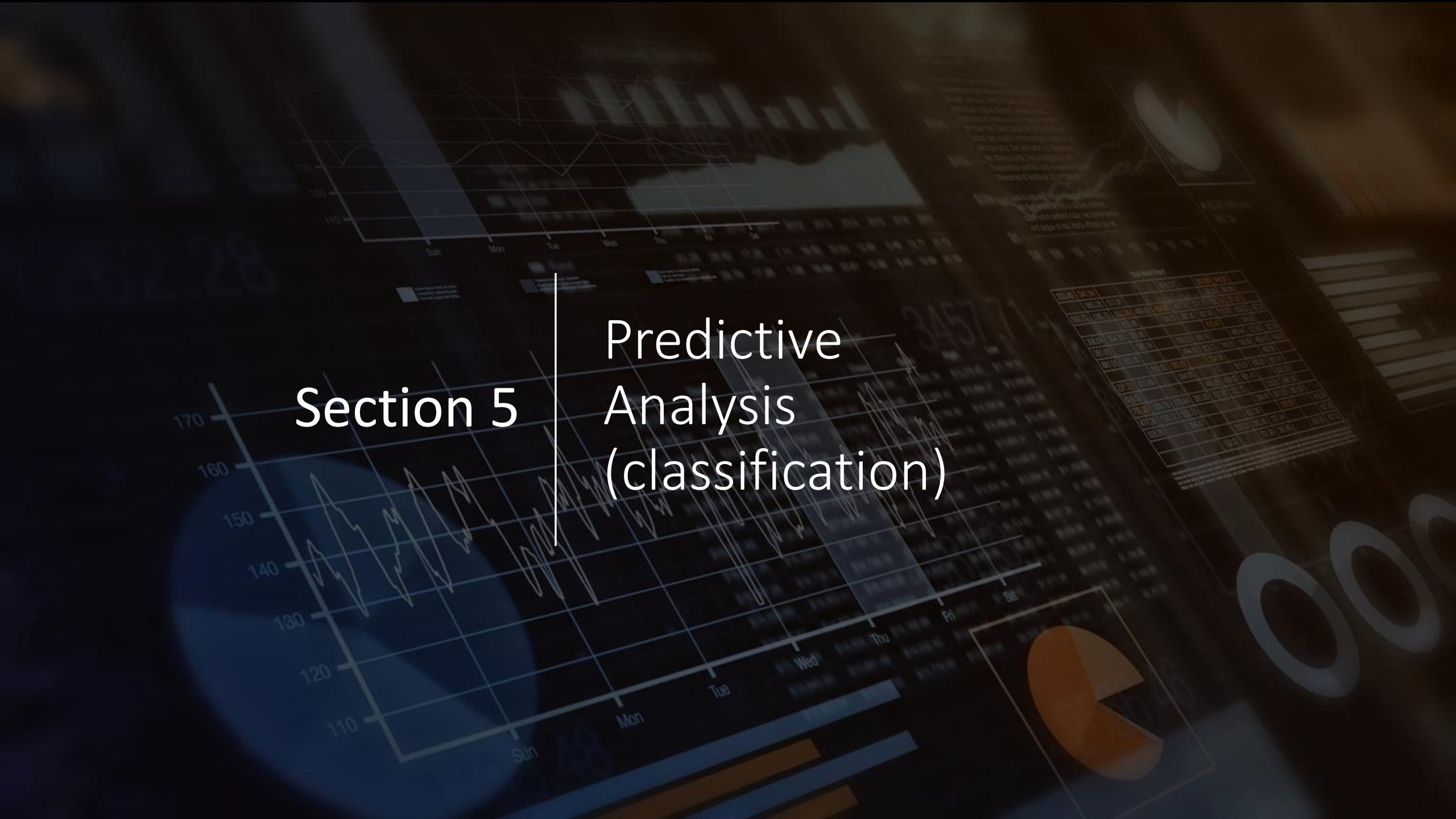
Piechart from Launch Site with highest Success Ratio



This screenshot contains the piechart from the launch site with highest success ratio, which is the blue region, and covers 76.9% of the total and corresponds to the KSC LC-39A launch site.

Correlation Between Payload and Success for all Sites

Fig. 1 shows the class (percentage of successful operations) for all sites for payloads in range from 0 to 5000 kg.
Fig. 2 shows the same variables with payloads in range from 5000 to 10000 kg.
We have a full success rate (1.0) for payloads below 5000, and until 5400 kg, but a full failure rate (0.0) for payloads superior to 5400 kg on boosters of type FT. Boosters of type B4 and B3 have success rates of 100%.

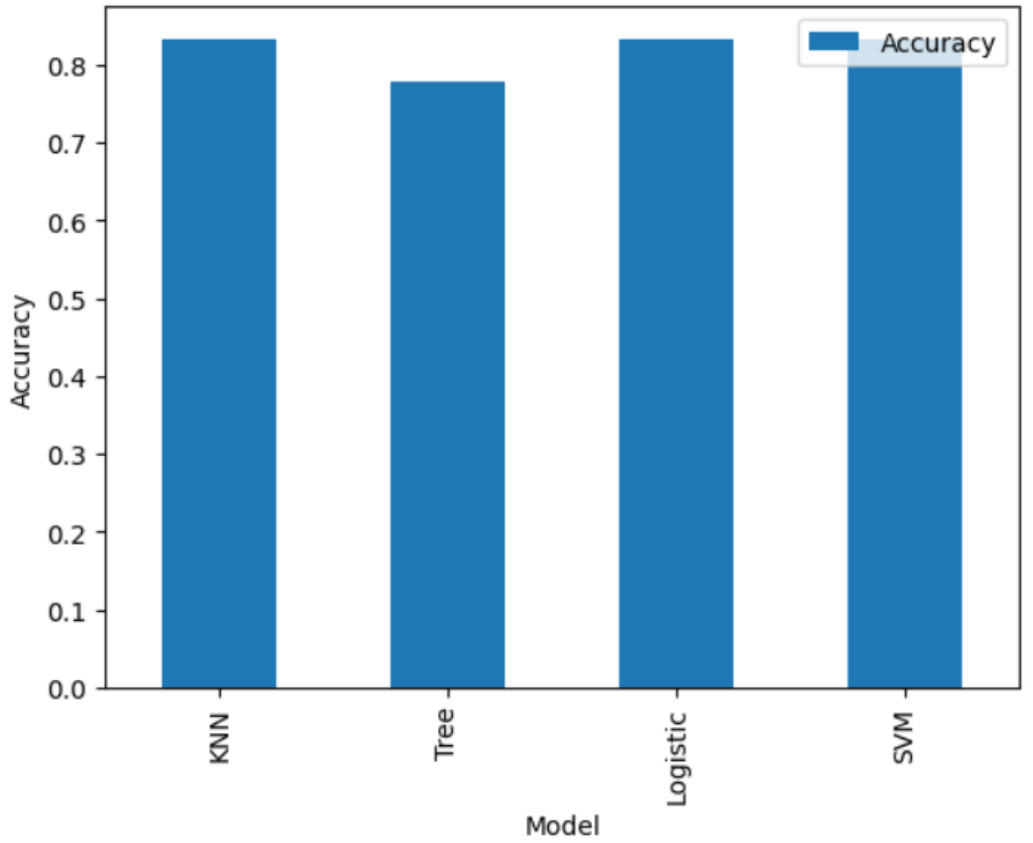


Section 5

Predictive Analysis (classification)

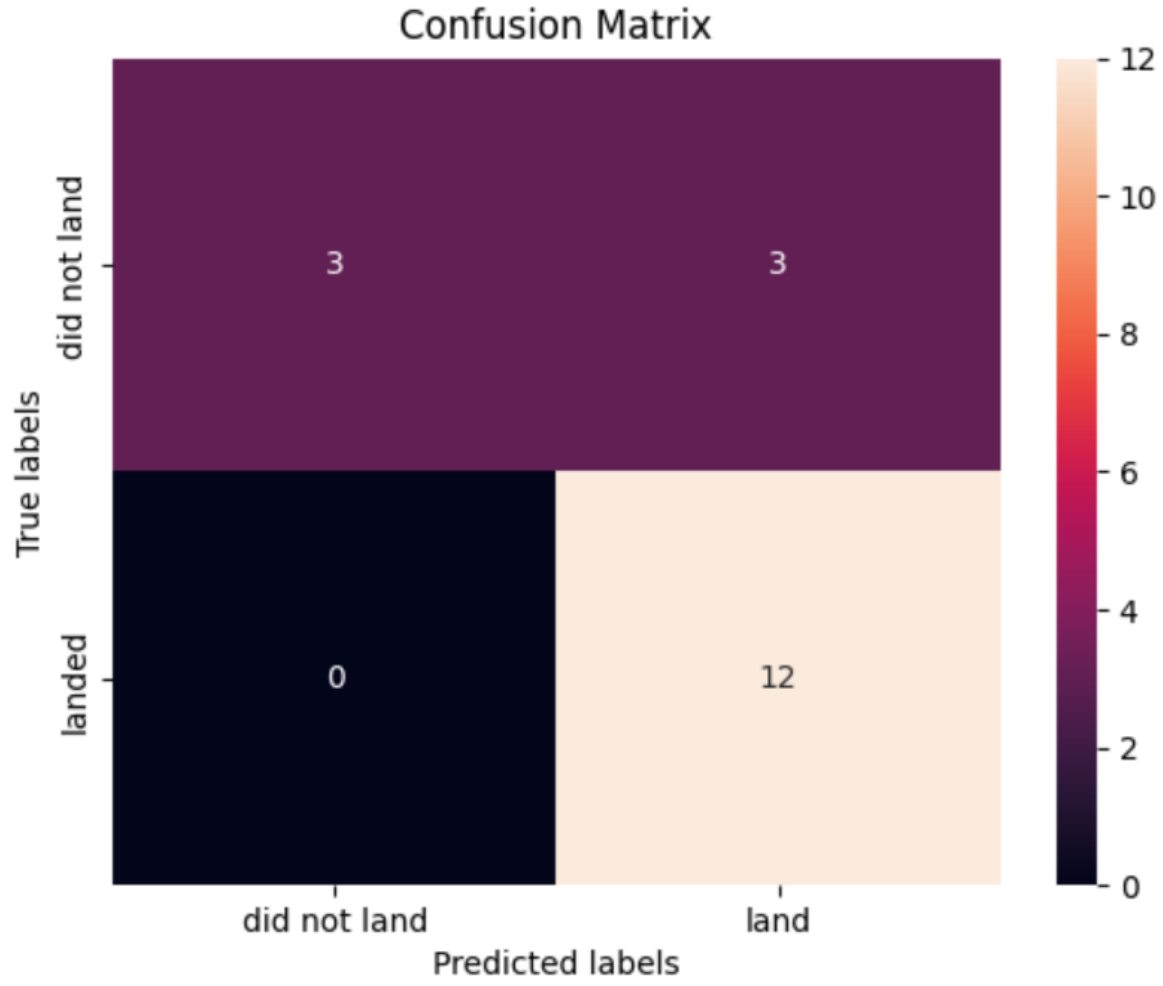
Classification Accuracy

Most accurate models are KNN, Logistic and SVM. However, KNN and SVM have the highest accuracy scores with trained data of 0.84, so they are the best-performing methods.



Confusion Matrix

Confusion matrix of KNN. As can be seen, most of the values land on the correctly predicted (diagonal of the matrix) categories, and there are no false negatives, which means that the model's predictions will fail by delivering false positives.
(i.e., the model predicts that the landing is successful but it fails in real-life).



Conclusions

- KNN with 10 neighbors, p = 1 and auto algorithm provides the most reliable prediction results.
- Launches and landings from spaceX have become more successful since 2013, which means that their recovery technology is now mature.
- Rockets with payloads inferior to 5400 kg report a 100% success rate, whereas superior payloads are still unreliable. This opens a competitive advantage for our company. We recommend that our R+D department starts inmediate development and testing of a reusable heavy-payload rocket, with which we can compete with SpaceX.
- Launch sites are located in coastal zones close to the Equator with railroad connections. Our company should elaborate a list of viable locations in favorable grounds of operation in order to secure operational availability and logistics. The company's building facilities should be located taking these locations in mind, or our scientists will be forced to develop modular rockets that are easy to transport, severely limiting payload size and weight.

- SpaceX is an American-centered company, which means that other space agencies and actors (Virgin Galactic, ESA, JAXA, ISRO, Roscosmos…) would have to deal with transport difficulties from their payloads should they need to purchase a SpaceX launch. In order to exploit this virgin markets regarding commerical reusable rockets, we should aim at expanding in the European market, where there is a lot of movement regarding nanosatellite technology. Perhaps a viability study in conjunction with ESA and their launch facility in French Guayane is in order.

Appendix

Link to Dataset

Thank you!