IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Grégoire Bouiller
19.04.2024

Outline

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

Executive Summary

The goal of this research is to study the impact of different factors on the outcome of a spaceship's first stage landing.

- Summary of methodologies
  - **Data collection**: collect data from SpaceX website using data scraping
  - **Data Wrangling**: ready Falcon9's data in pandas dataframe and general statistics analysis
  - **Exploratory data analysis** with Python + SQL: Analysing informations from data previously prepared: Max Payload carried by boosters, impact of payload, type of landing pad on landing outcome
  - **Data visualization**: representation of Flight's Payload and Flight's Launch Site against Outcome of flights, Success rate of each orbits, Flight's Orbit and Payload impact on outcome, success rate of flights over time
  - Representation of launches on the map of the United States with **Folium**
  - **Dashboard** representing success launches by launch site and the correlation between payload and success rate, using Dash
  - Successful landing prediction with **Machine Learning**, testing different models: SVM, KNN, LogReg and decision Tree

- Summary of all results
  - Success rate of launches has improved
  - The launch site KSC LC-39A has the highest success rate
  - Orbits ES-L1, GEO, HEO and SSO have the highest success rates
  - Orbits Polar, LEO and ISS seem more adapted to high payload mass

Introduction

- SpaceX was funded in 2002 by billionaire Elon Musk, judging the aeronautical leader NASA not ambitious enough. His goal was to revolutionize the world of aeronautics by developing a first stage capable of landing, drastically reducing the costs of launches. Starting with many failures with the launcher Falcon 1, SpaceX has reached a very high success rate with their launcher Falcon 9 in the last 10 years, reaching their goals of cutting costs.

- Analyzing available SpaceX data, our goal is to build a machine learning pipeline to predict the outcome of a launch, allowing us to calculate the costs of the launches in advance.

- To do so, we have to analyze and determine the dependency of the landing success rate on many other factors, such as orbit type, site location, payload mass, etc…

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Data was collected using SpaceX REST API
- Perform data wrangling
    - The data was processed using pandas, cleaning and filling missing data. A normalization on all data is also performed.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - How to build, tune, evaluate classification models

Data Collection

Data was collected using two different techniques:

- Data collection using SpaceX REST API. As SpaceX has their own data made public, their API can be used to retrieve the informations we need to build our models.

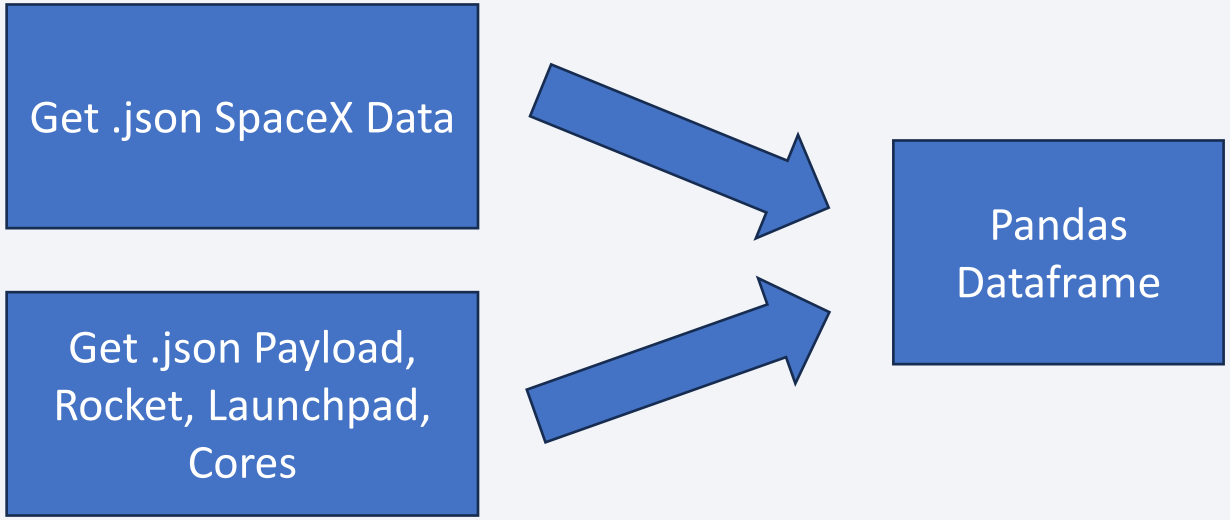
- Data collection using BeautifulSoup and SpaceX wikipedia page

Data Collection – SpaceX API

Data was first collected using SpaceX's REST API. This API permits us to collect data that has been made publicly available and analyze it.

Data was collected in two steps:

1. Get request using REST API for the data, and additional informations on payload, rocket, launchpad and cores
2. .json normalization to transform data into dataframe



Get .json SpaceX Data

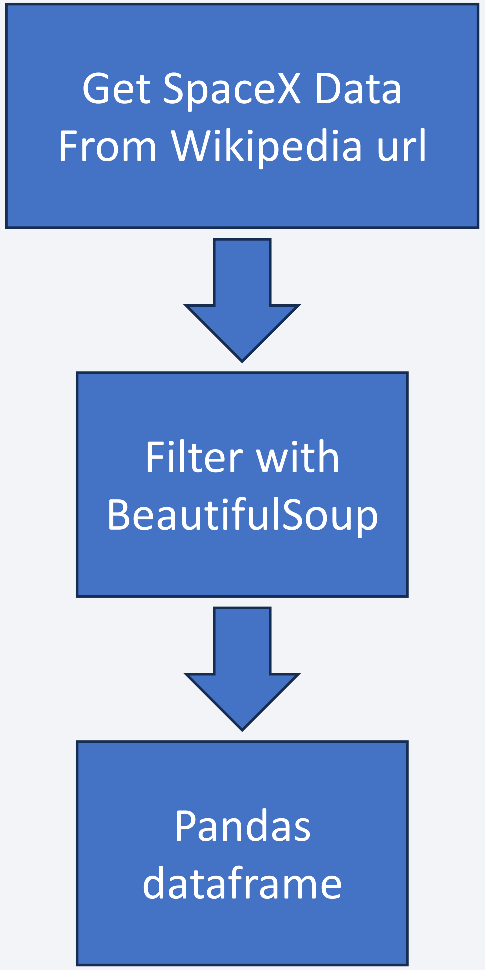
Get .json Payload, Rocket, Launchpad, Cores

Pandas Dataframe

Data Collection - Scraping

The same data collection was performed using BeautifulSoup and SpaceX Wikipedia page to retrieve information on Falcon 9 launcher.

The main advantage of this technique is the necessity of only one .get command, as all data is directly available on the Wikipedia page, but more filtering is needed with BeautifulSoup to select the right data.



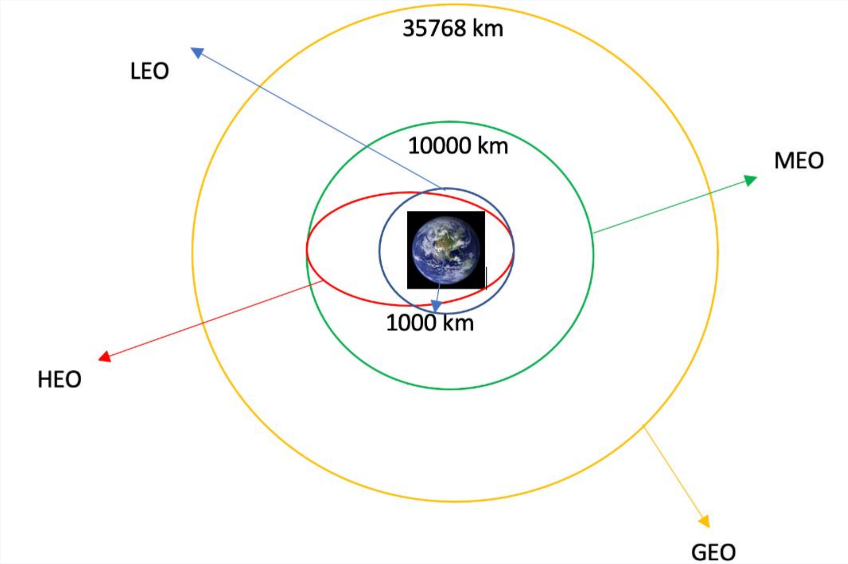
Get SpaceX Data From Wikipedia url

Filter with BeautifulSoup

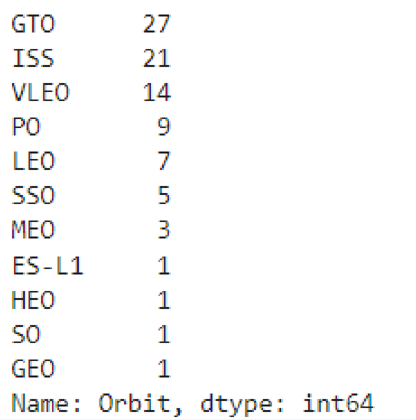
Pandas dataframe

Data Wrangling

- After data collection, we performed normalization and replacement of missing data by the mean of the targeted column. Once the data was ready to be analyzed in a pandas dataframe, we were able to observe the following:
  - Launch Site CCAFS SLC 40 has the highest number of launches, with 55

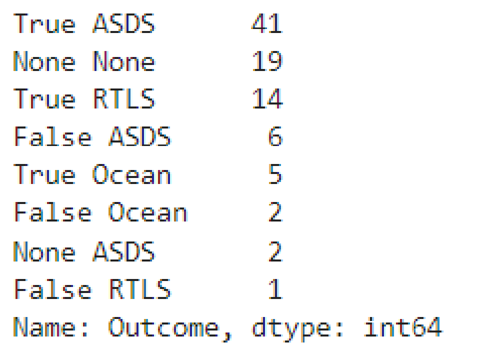


Several orbit types were tested:



```
GTO       27
ISS       21
VLEO      14
PO         9
LEO        7
SSO        5
MEO        3
ES-L1      1
HEO        1
SO         1
GEO        1
Name: Orbit, dtype: int64
```

And different landing outcomes:



```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: Outcome, dtype: int64
```

RTLS: Ground pad
ASDS: Drone ship

EDA with Data Visualization

- Plotting variables against each other give us many information on which criteria will influence the outcome of a launch. In this part, we plotted the following:
- **Flight nb** vs **Payload mass**, to study the impact of mass on the outcome
- **Flight nb** vs **Launch Site**, to unveil which launch site tends to have more success rate
- **Launch site** vs **Payload mass**
- Success rate of each **orbit**
- **Flight nb** vs **orbit**
- **Payload mass** vs **orbit**
- Success rate trend over the years

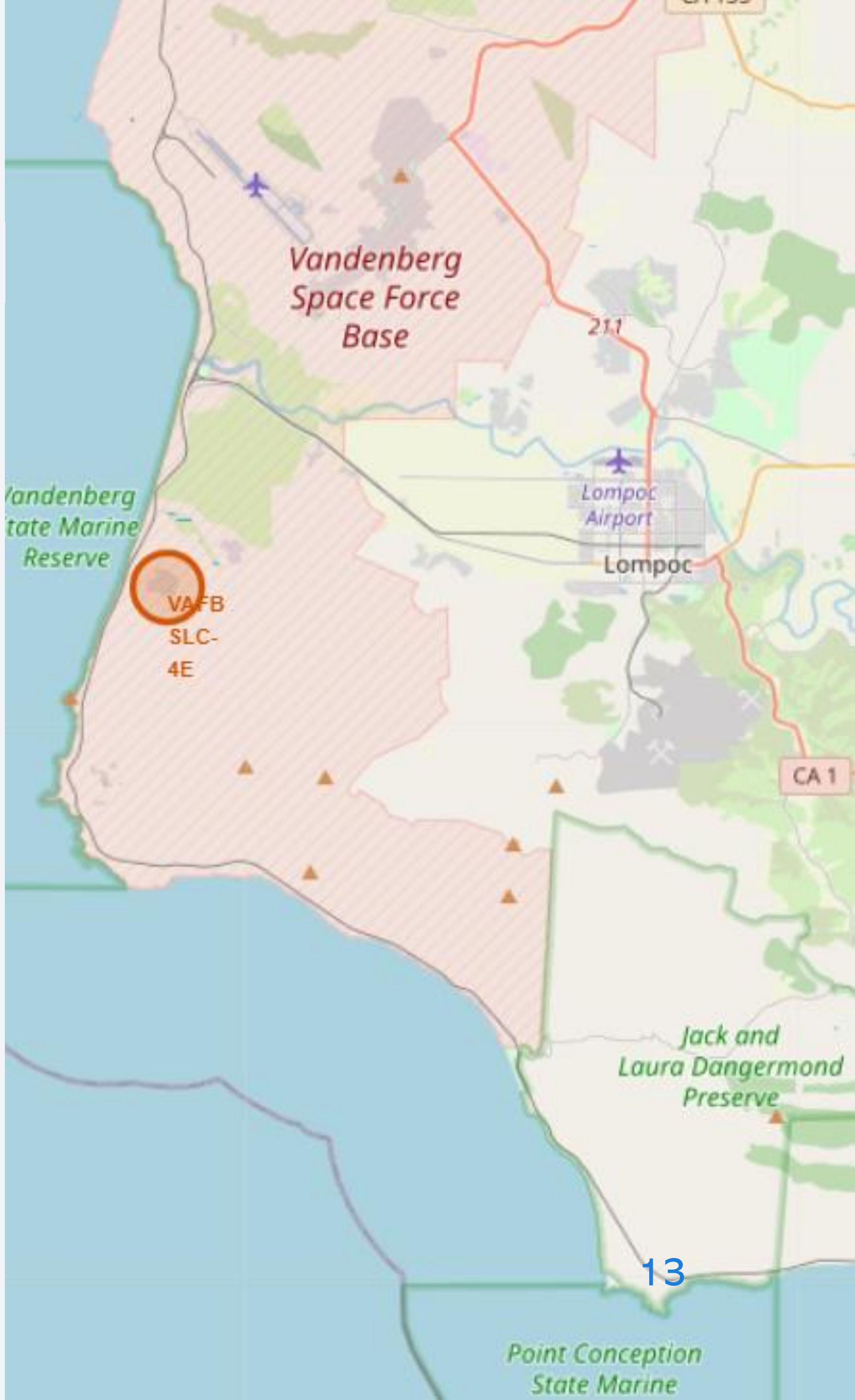
EDA with SQL

Here is a list of the information gained from EDA using SQL:

- Unique launch sites
- Payload mass sum
- Average payload mass per flight
- First successful landing on ground pad
- List of booster versions

Build an Interactive Map with Folium

- Using Folium with Python, we added to the map of the USA markers to show the locations of the launch sites. Information on all the launches are also displayed for all launch sites, including their outcomes
- Adding markers allowed us to get information on the proximity of the sites to key infrastructure, or how the locations of the sites can help with sending stages into orbit.


Vandenberg Space Force Base
211
Lompoc Airport
Lompoc
/andenberg
tate Marine
Reserve
VAFB SLC-4E
CA 1
Jack and Laura Dangermond Preserve
Point Conception State Marine

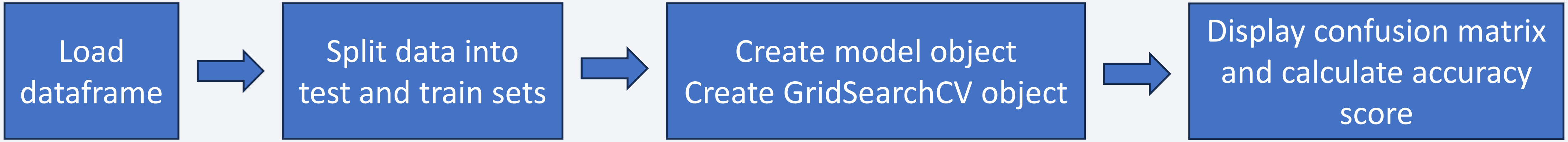
Build a Dashboard with Plotly Dash

- Using Plotly Dash, we created a dashboard to display various charts making data concerning success rate of launch sites and the relation payload mass / success rate per site very accessible and interactive.
- It is possible to compare sites to each other, or display data per site.

Predictive Analysis (Classification)

- We tested several classification model to find the best for our problem:
  - K nearest neighbour
  - Classification decision Tree
  - Logistic Regression
  - Support vector machine

The same development process was used for the four models:



Load dataframe

Split data into test and train sets

Create model object
Create GridSearchCV object

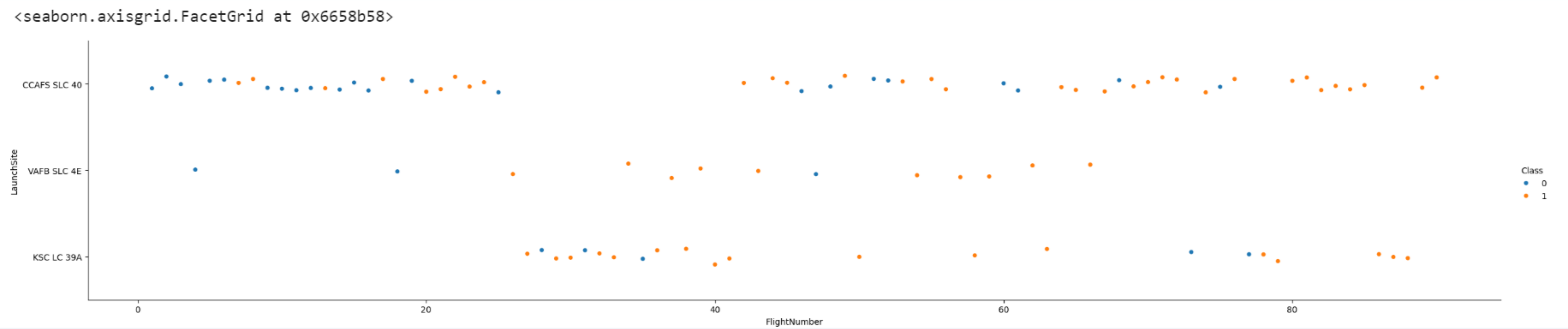
Display confusion matrix and calculate accuracy score

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

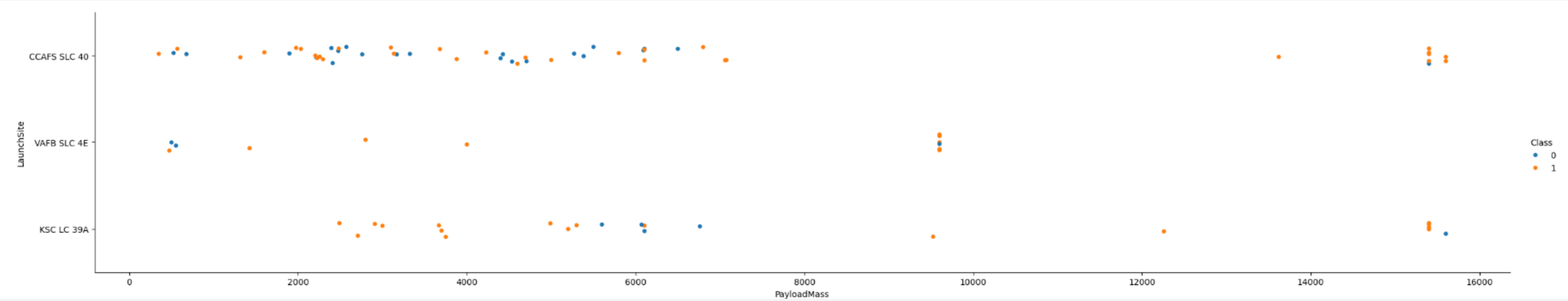
By observing the Flight Number vs Launch Site scatter plot, we can determine that the success rate of the launches have improved over time for all launch sites. More launches were conducted from the CCAFS SLC 40 site than the other two sites, as it seems to have been the first one constructed.



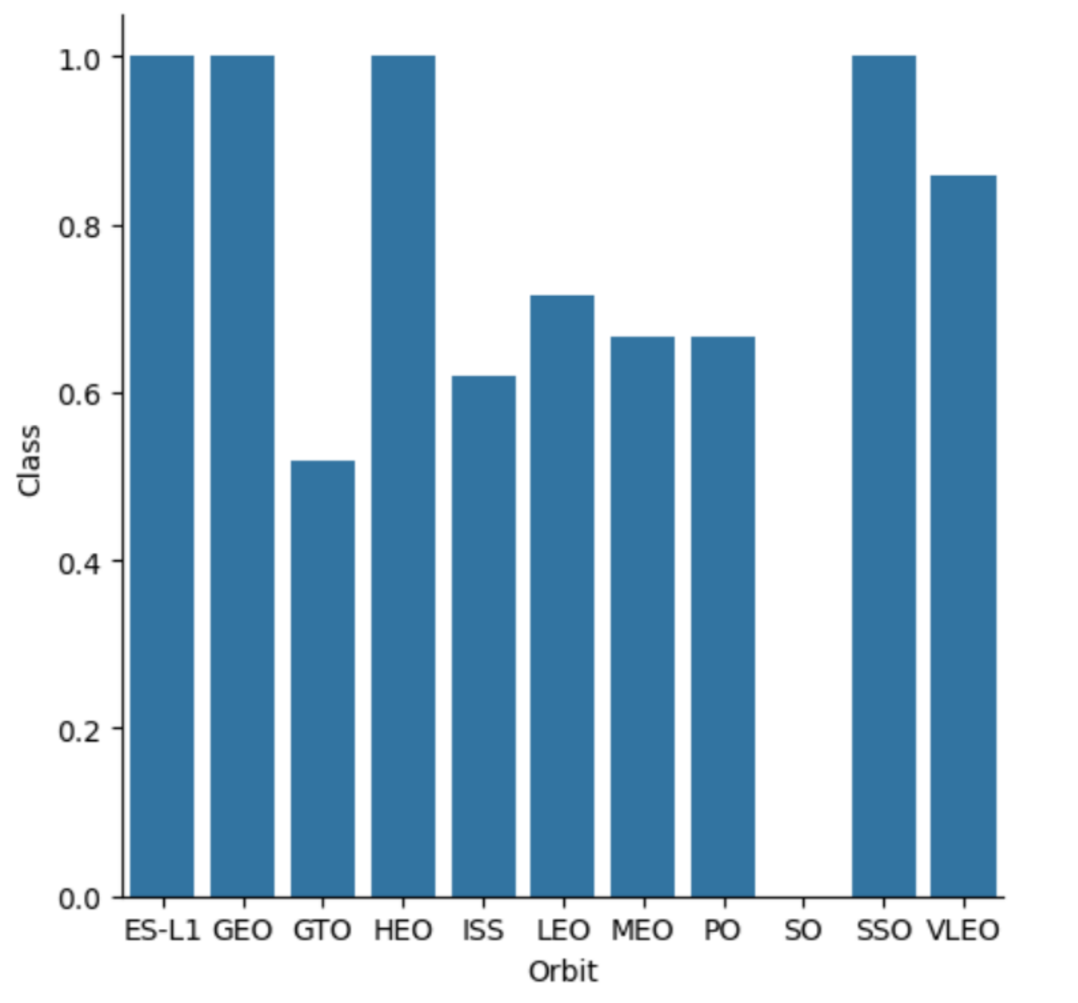
```
<seaborn.axisgrid.FacetGrid at 0x6658b58>
```

Payload vs. Launch Site

The two sites CCAFS SLC 40 and KSC LC 39A had launches with very high payload mass (15000+kg), resulting in a success for most of them. Most of launches were conducted with less than 7000 kg, with a high success rate for VAFB SLC 4^{E} and KSC LC 39A between 1000 and ~5000kg of payload mass.



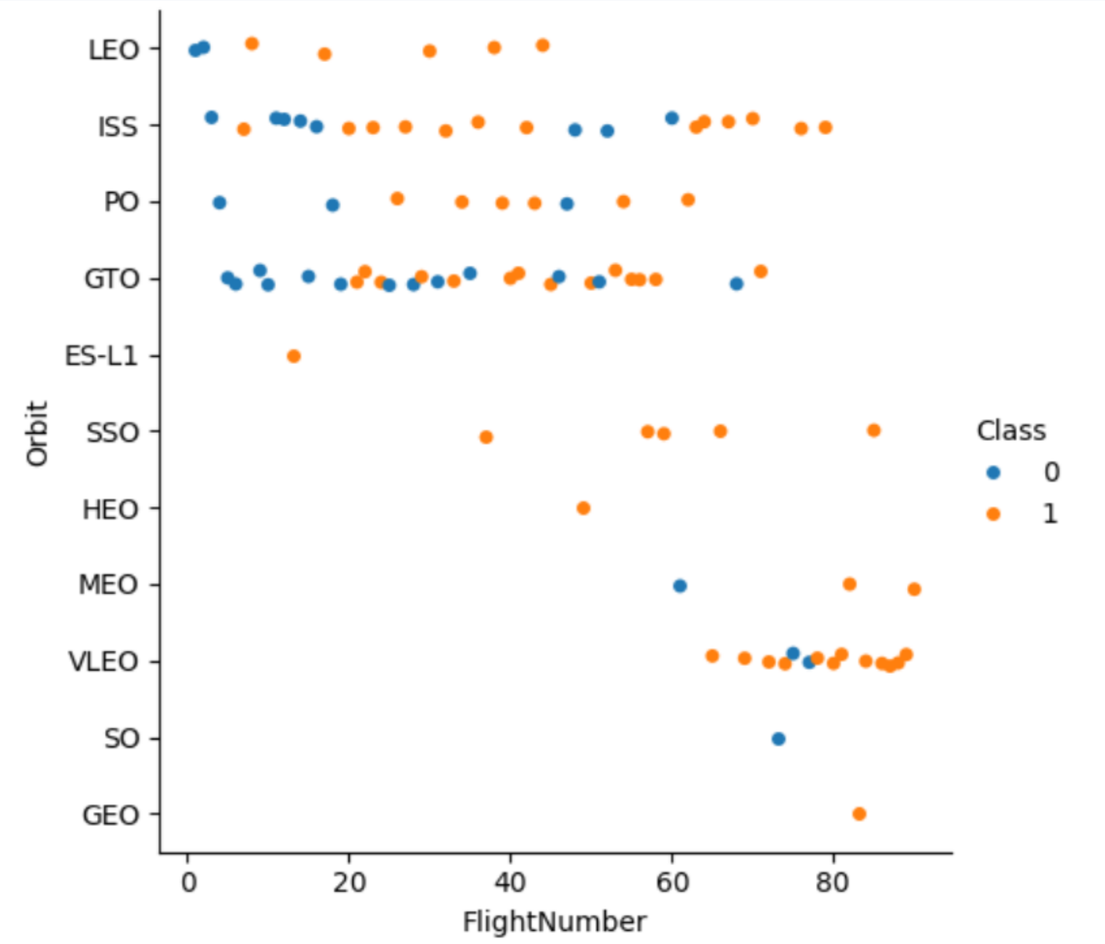
Success Rate vs. Orbit Type



As seen in this bar plot showing the success rate of launches per orbit type, it is clear that the orbits ES-L1, GEO, HEO and SSO have the highest success rate of 1.0, followed closely by VLEO, with over 0.8 and LEO at 0.7, cumulating much more flights than the orbits cited above.

Altough appearing very prone to sucess, the orbits GEO, ES-L1 and HEO only have one reported flight each. More tests should be conducted to draw conclusions.

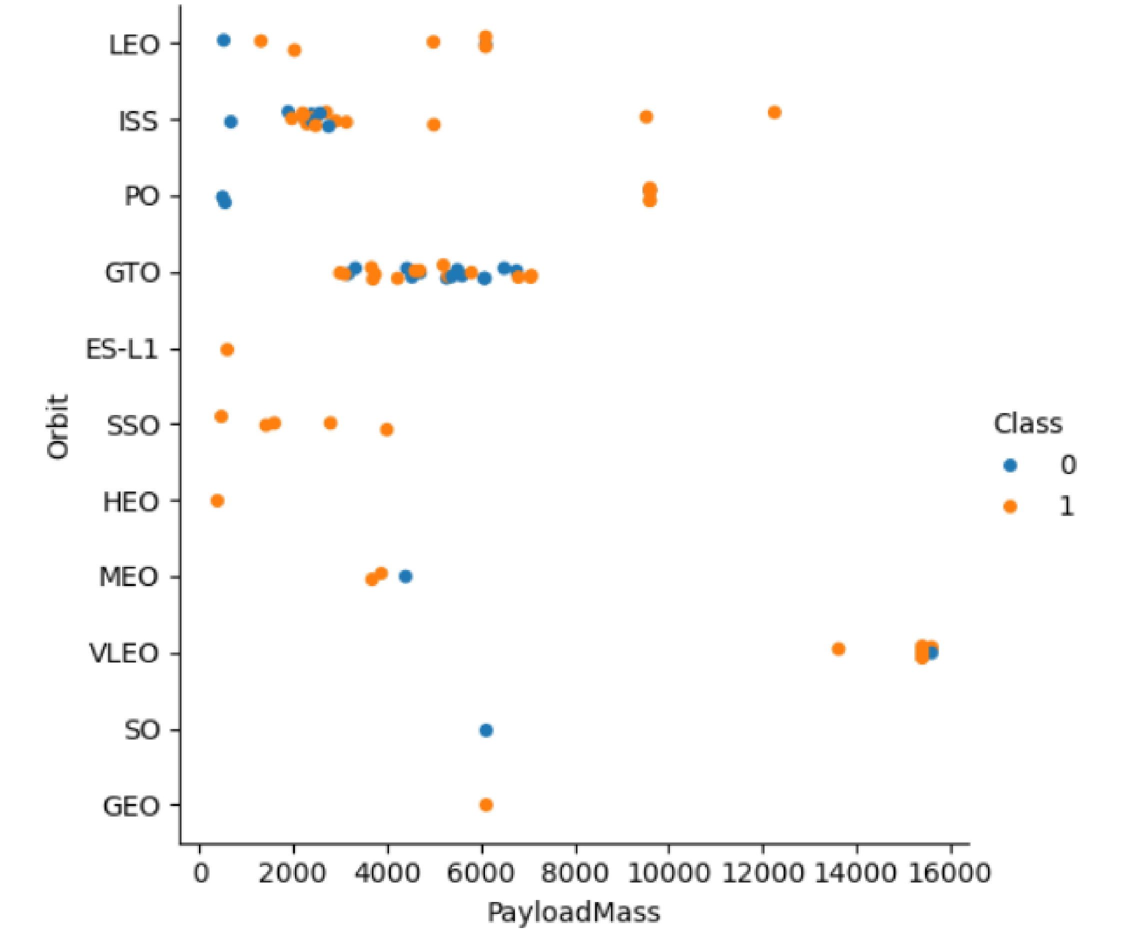
Flight Number vs. Orbit Type



An improvement in successful outcome can be observed for the orbit LEO over time.

The success rate of HEO, GEO and ES-L1 of 1.0 is more obvious in this graph, having too little data to draw any conclusion.

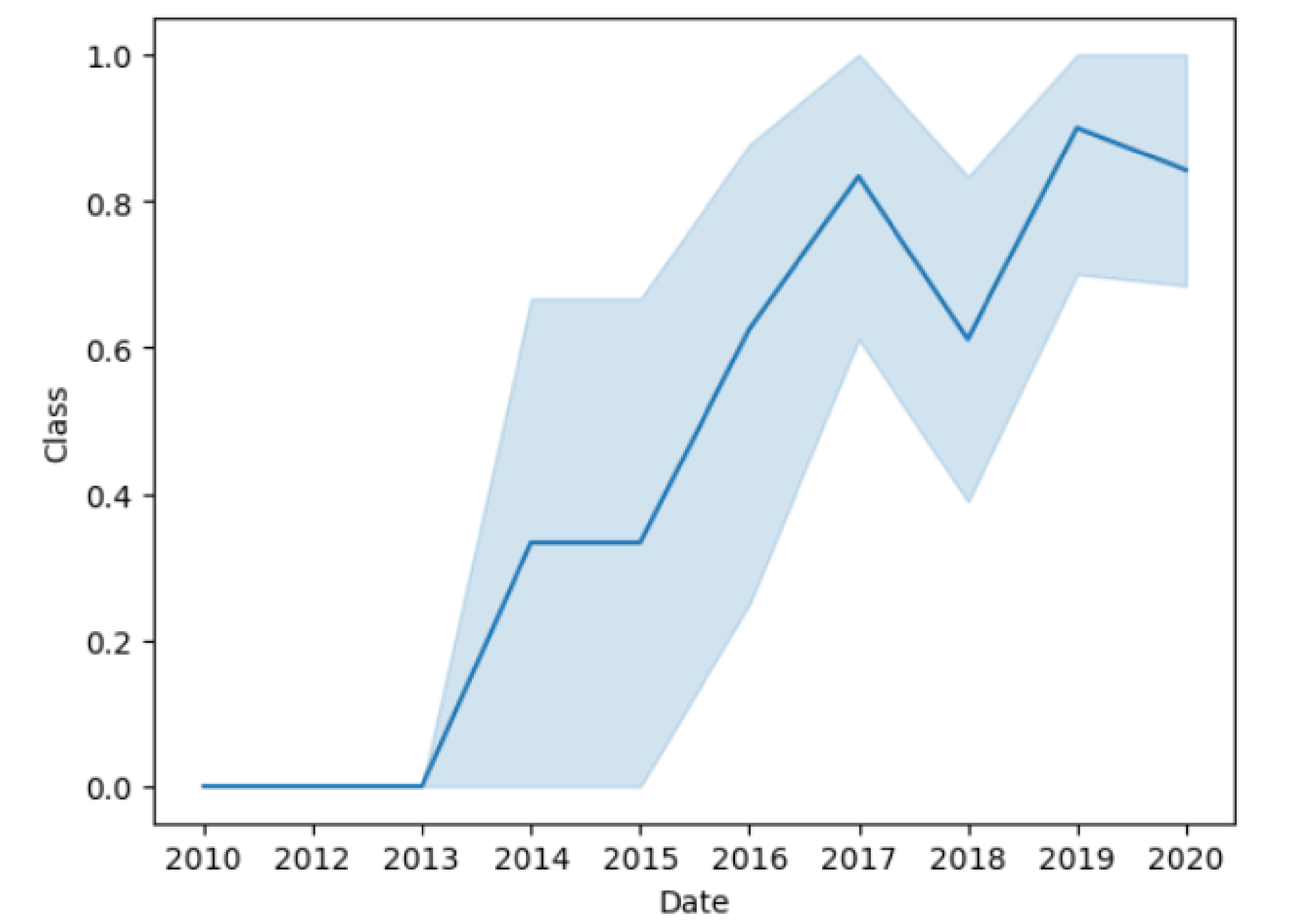
Payload vs. Orbit Type



The scatter plot showing Orbit Type vs Payload mass reveals that the orbits ISS and PO are the best suited for high payloads (>= 10000kg), although more data would be necessary to confirm it.

The orbit VLEO is performing very well with payloads over 13000kg, with only one failure in 5 launches.

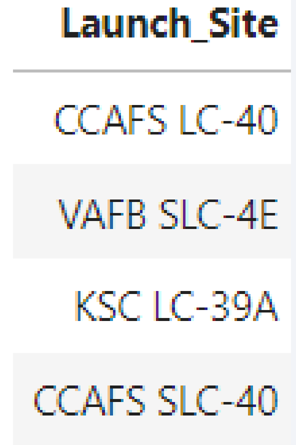
Launch Success Yearly Trend



The increase of launch successes has been drastic over the years, reaching around 80% of success rate in 2020.

All Launch Site Names

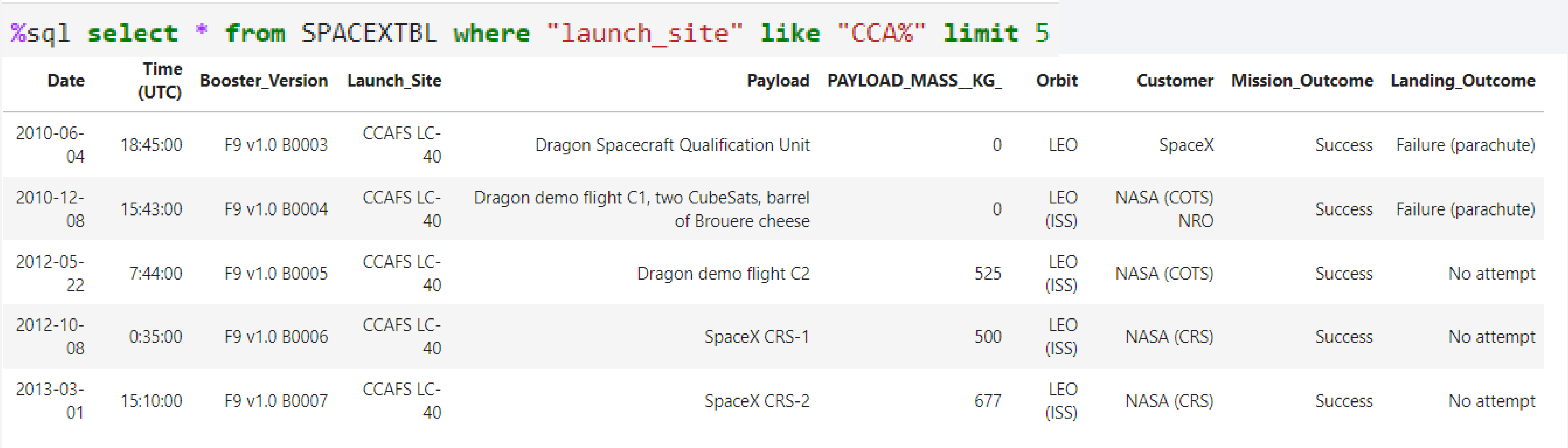
Four launch sites are being studied, all located in the United States of America, with two sites in the east coast and the two other in the west.



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

Launch Site Names Begin with 'CCA'

Here are 5 rows of launches in a site beginning with CCA, queried using SQL.

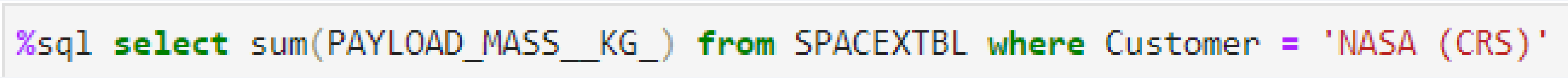


```
%sql select * from SPACEXTBL where "launch_site" like "CCA%" limit 5
```

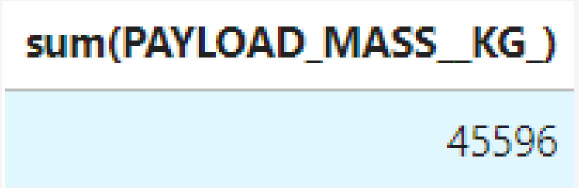
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

Using the following SQL command, we were able to retrieve the total payload mass of every launches cumulated, which sum up to reach around 46 tons.



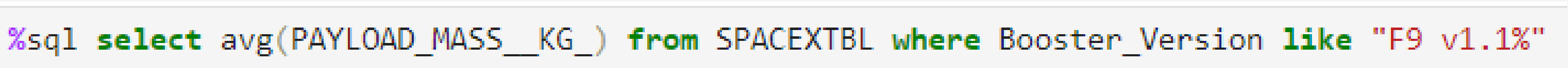
```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)'
```



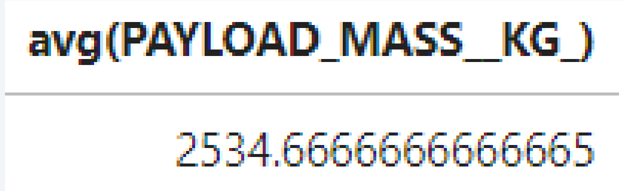
sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

The booster F9 v1.1 average payload is around 2.5 tons. We can deduce that this booster is made for relatively light payload mass, as some boosters can carry up to 16 tons.



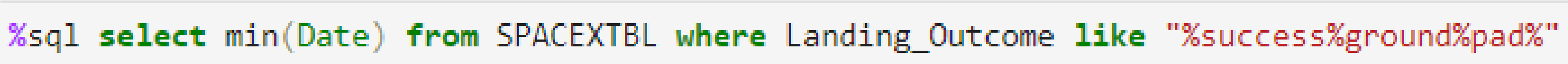
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like "F9 v1.1%"
```



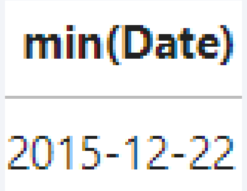
avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

The first succesful landing happened the 22nd december of 2015



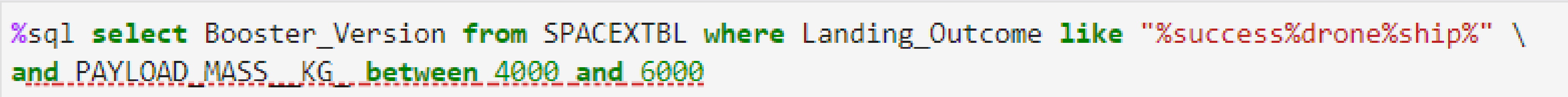
```
%sql select min(Date) from SPACEXTBL where Landing_Outcome like "%success%ground%pad%"
```



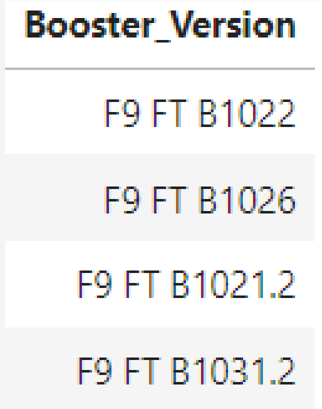
min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Four booster versions succesfully landed in a drone ship carrying 4000 to 6000 kg of payload.



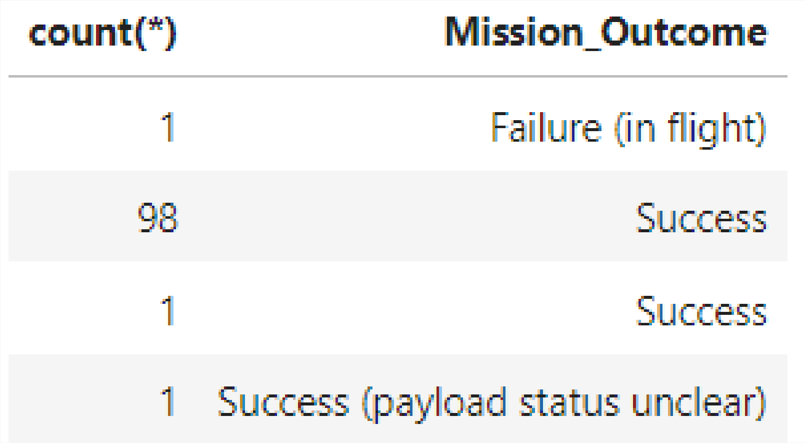
```
%sql select Booster_Version from SPACEXTBL where Landing_Outcome like "%success%drone%ship%" \
and PAYLOAD_MASS__KG_ between 4000 and 6000
```



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

Total Number of Successful and Failure Mission Outcomes

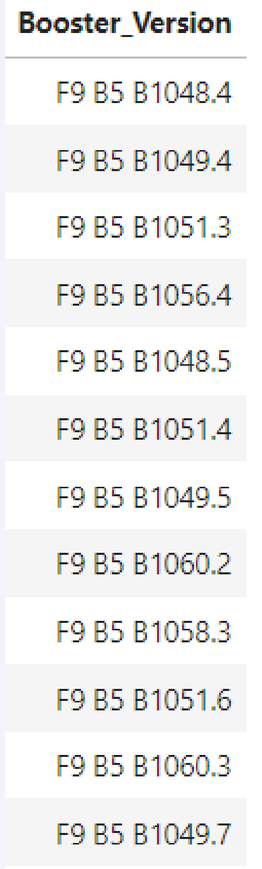
Four booster versions succesfully landed in a drone ship carrying 4000 to 6000 kg of payload.



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

Boosters Carried Maximum Payload

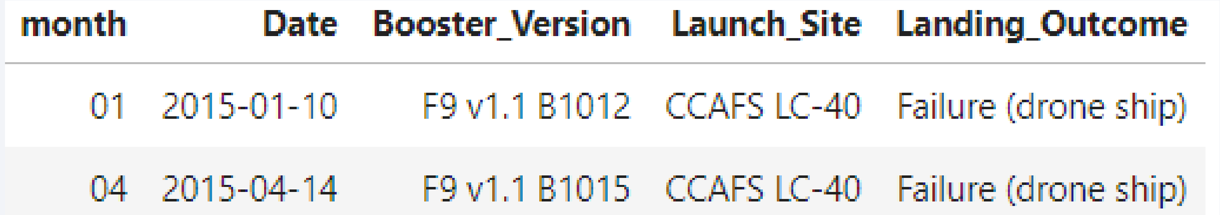
Several versions of boosters were able to carry the maximum payload mass.



Booster_Version
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

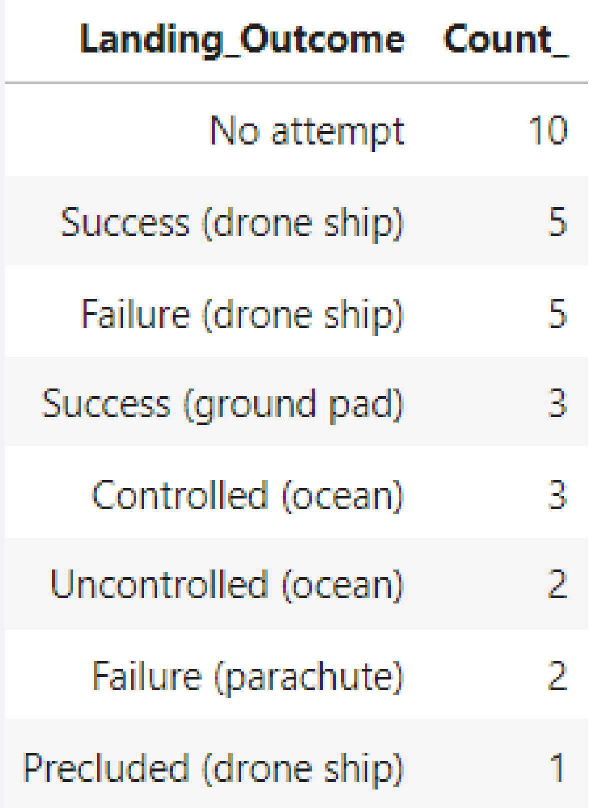
In 2015, two launches failed to land on a drone ship, with booster versions F9 v1.1 B1012 and B1015.



month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	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

Here is a recap of landing outcomes of all launches between 2010-06-04 and 2017-03-20

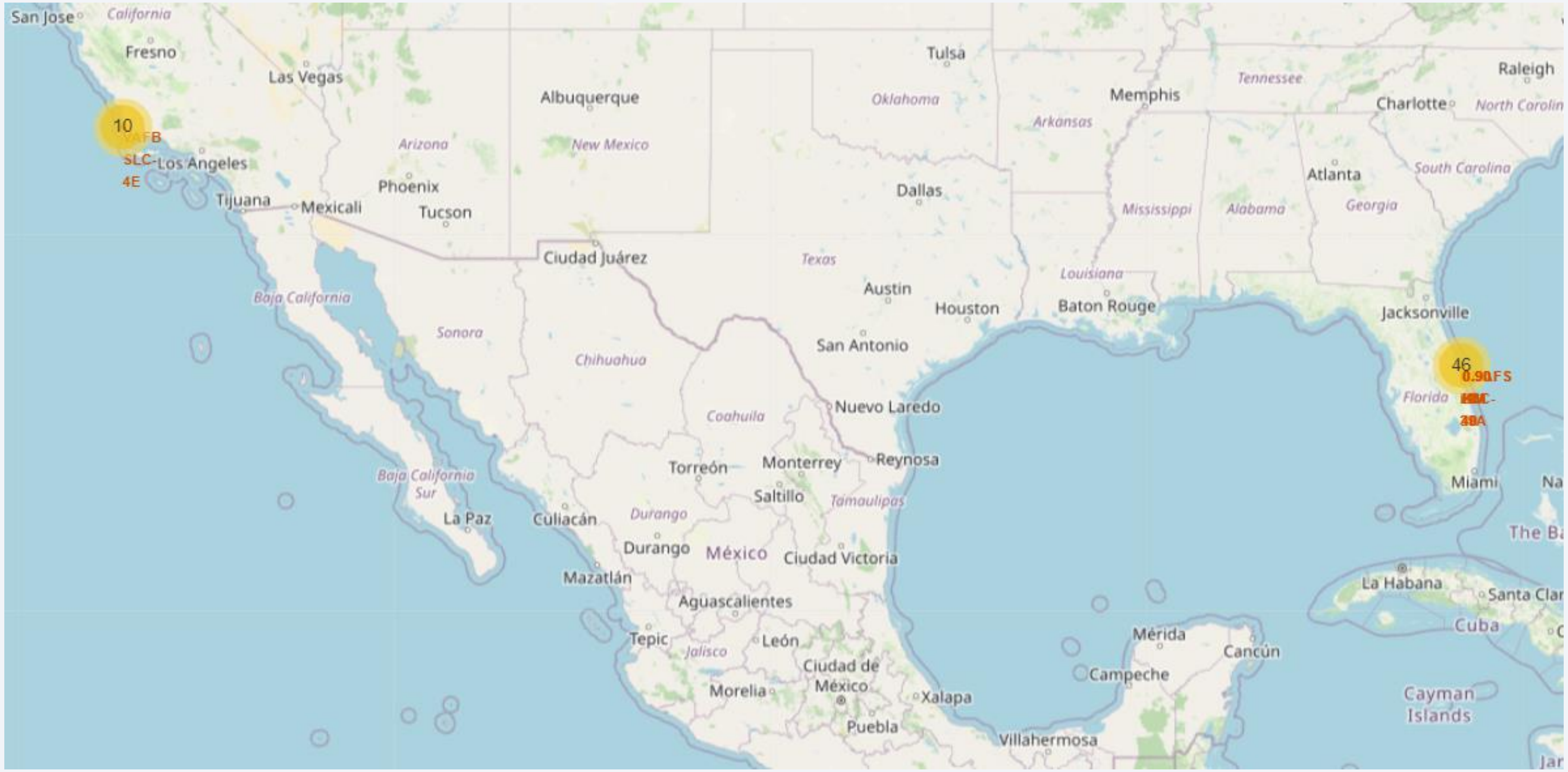


Landing_Outcome	Count_
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

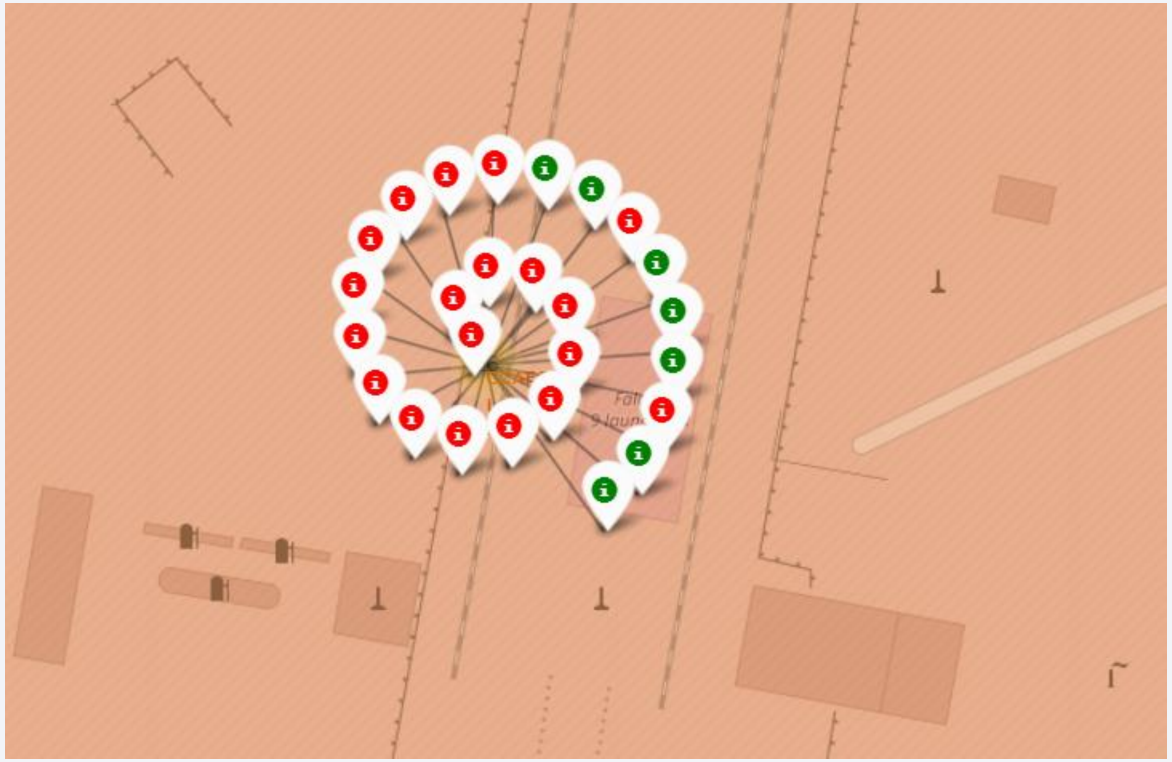
Launch Sites Proximities Analysis

Location of launch sites in the US



The 4 launch sites studied are all placed along the coastline for security measures, clearance of sky and meteorological conditions. The location is as close to the equator as possible within US territory, facilitating the entry into orbit.

Outcomes of launches per site

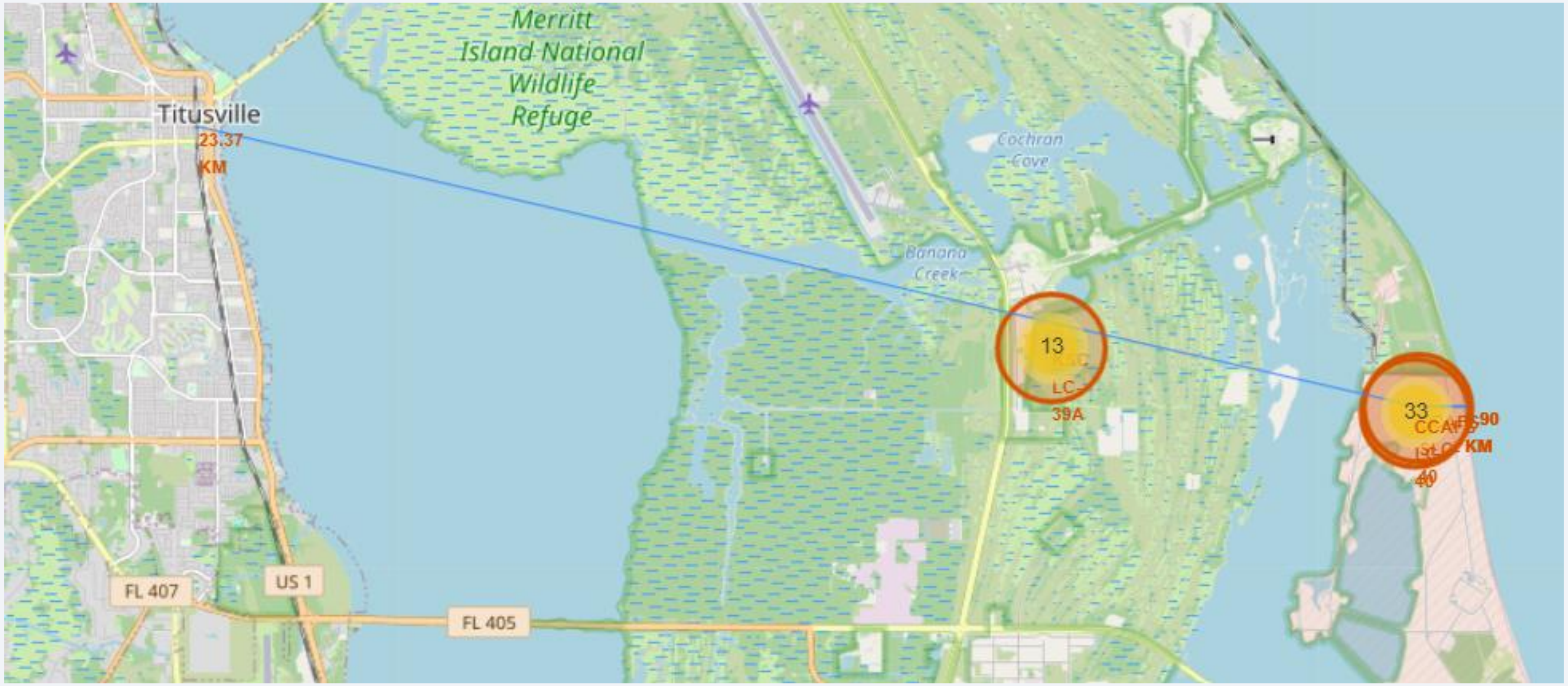


All of the launches outcomes can be observed for each launch site. A red color represent a failure and a green a success.

This map show the outcomes of the launches in CCAFS LC-40

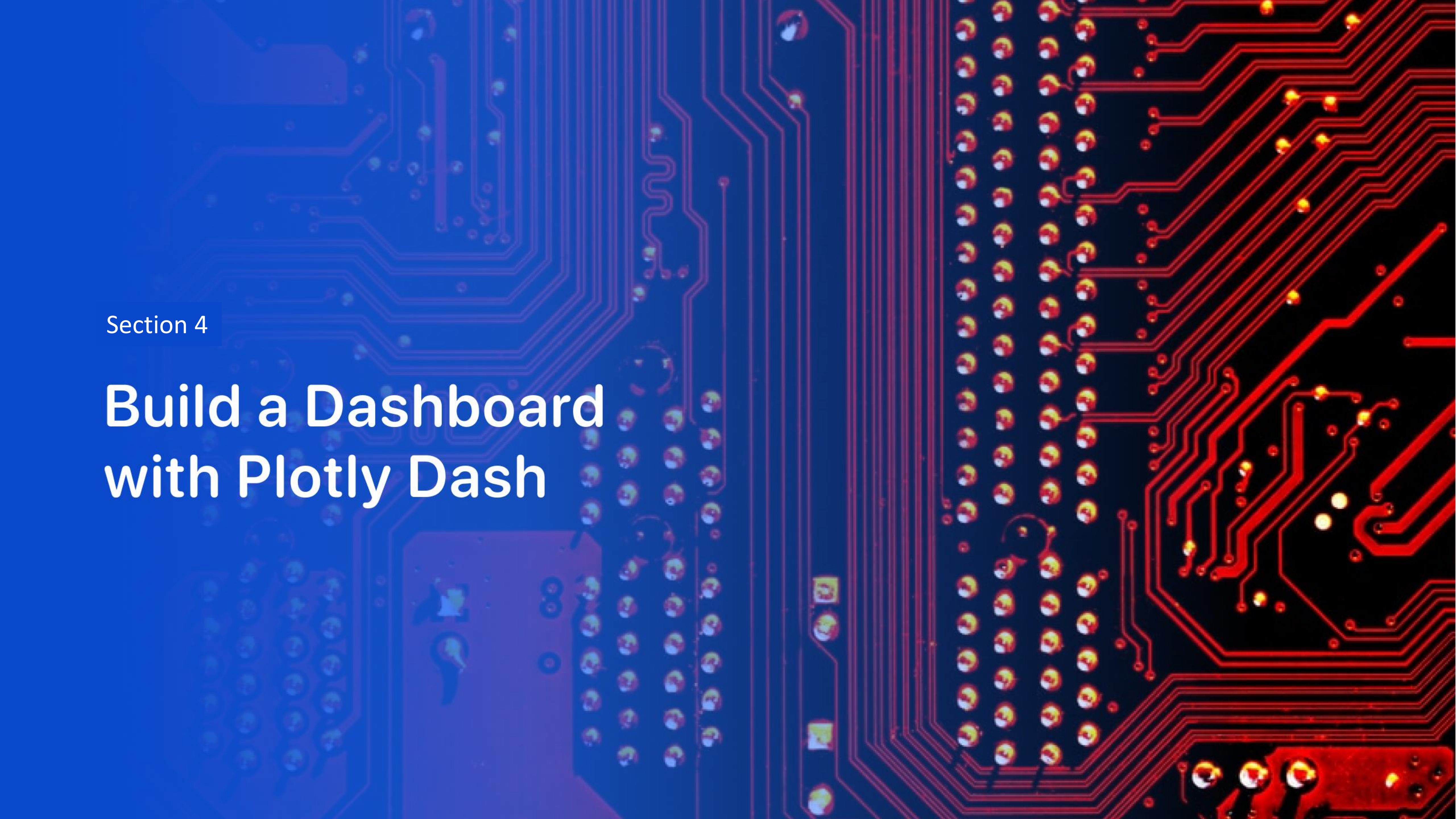
Proximity of launch site to infrastructures

The CCA LS-40 launch site is 23 km away from the nearest city, and 900m from the coast line. When installing a launch site in such area, security measures must be taken to protect any infrastructure from an eventual crash.



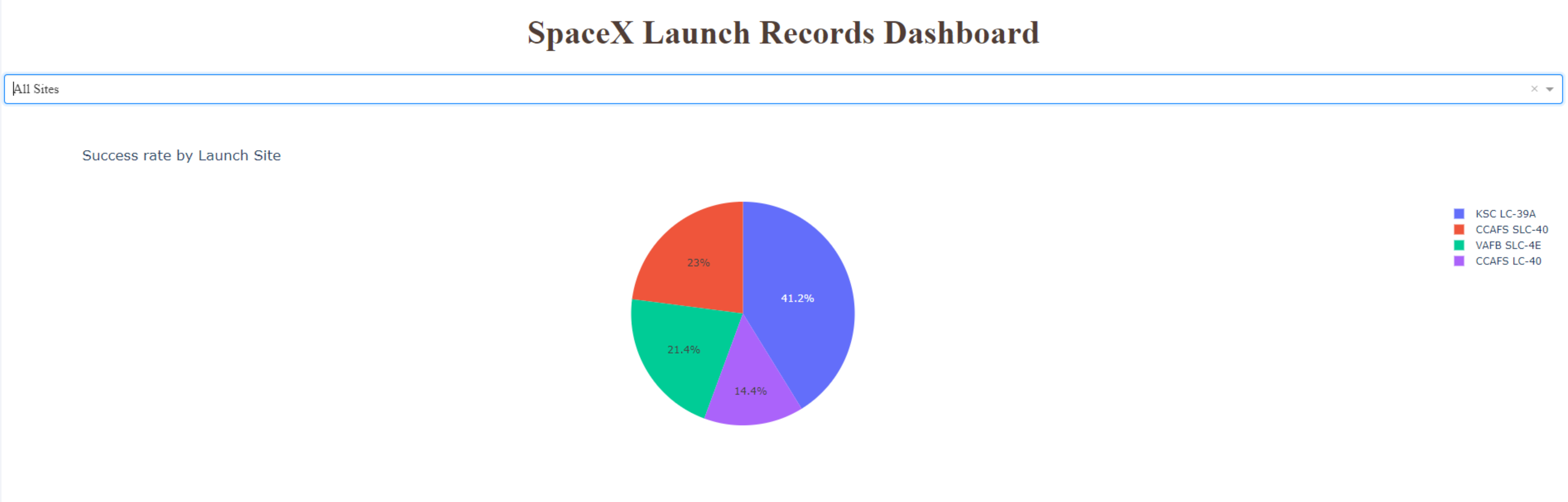
Section 4

Build a Dashboard with Plotly Dash



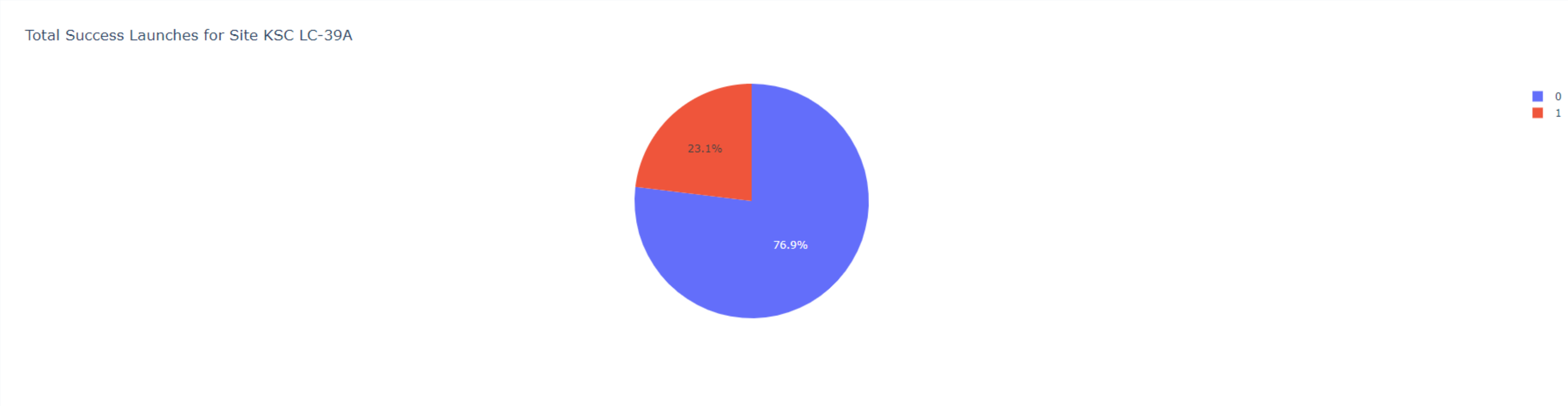
Success rate by launch site

KSC LC-39A has the most successful launches amongst all launch sites, cumulating 41.2% of the launches resulting in a positive outcome. CCAFS LC-40, has only seen 14.4% of the successful launches.



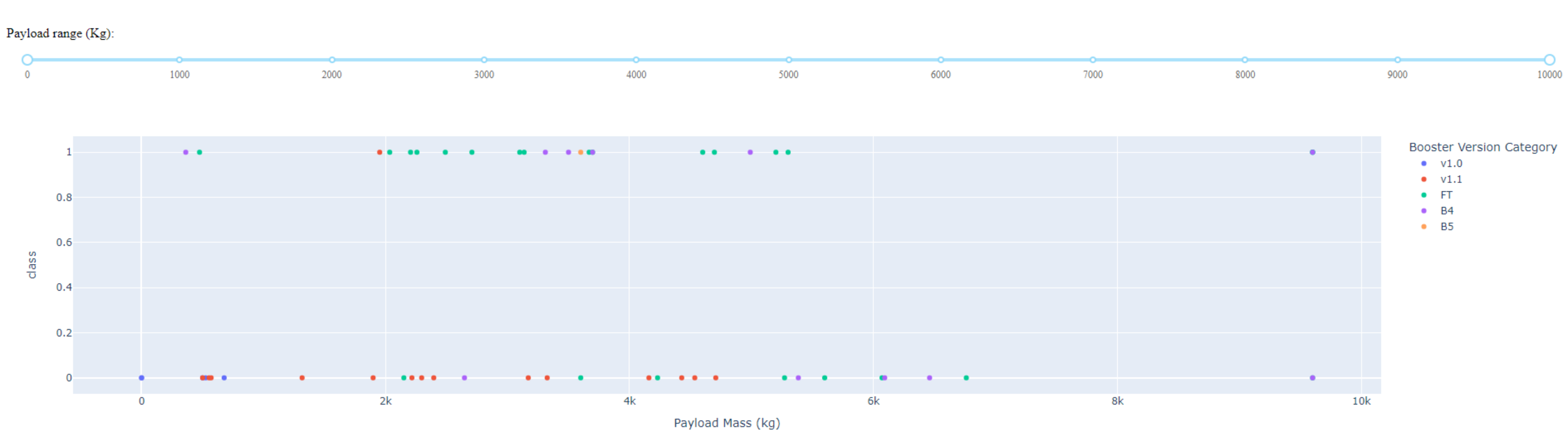
Success rate of KSC LC-39A

Zooming in the launches in KSC LC-39A only, we find that **76.9%** of the launches that happened in this site were successful, which makes **KSC LC-39A** the site with the highest success rate.



Payload mass, booster versions and success rate

Studying the relation between payload mass and success rate, we can observe that the payload range with the **highest success rate is from 2000 to ~5000kg**. The booster version FT has the most successful flights.



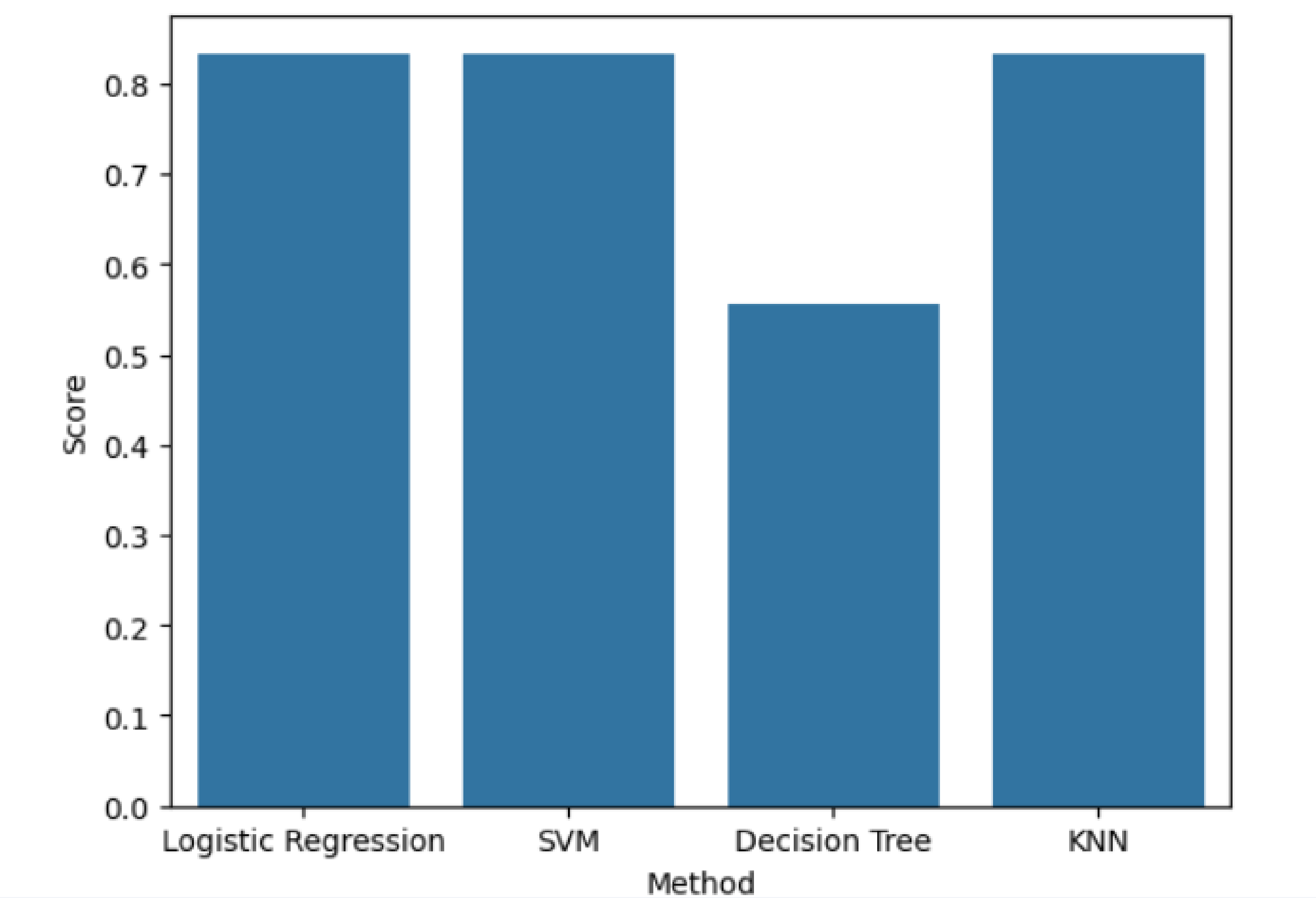
Section 5

Predictive Analysis (Classification)

Classification Accuracy

It was found that Logistic regression, Support vector machine and K nearest neighbours had a similar accuracy score of 0.84. The decision Tree classification method is behind in accuracy, with a score of 0.56.

It would be interesting to calculate jaccard index and f1-score to determine which method is truly the best.

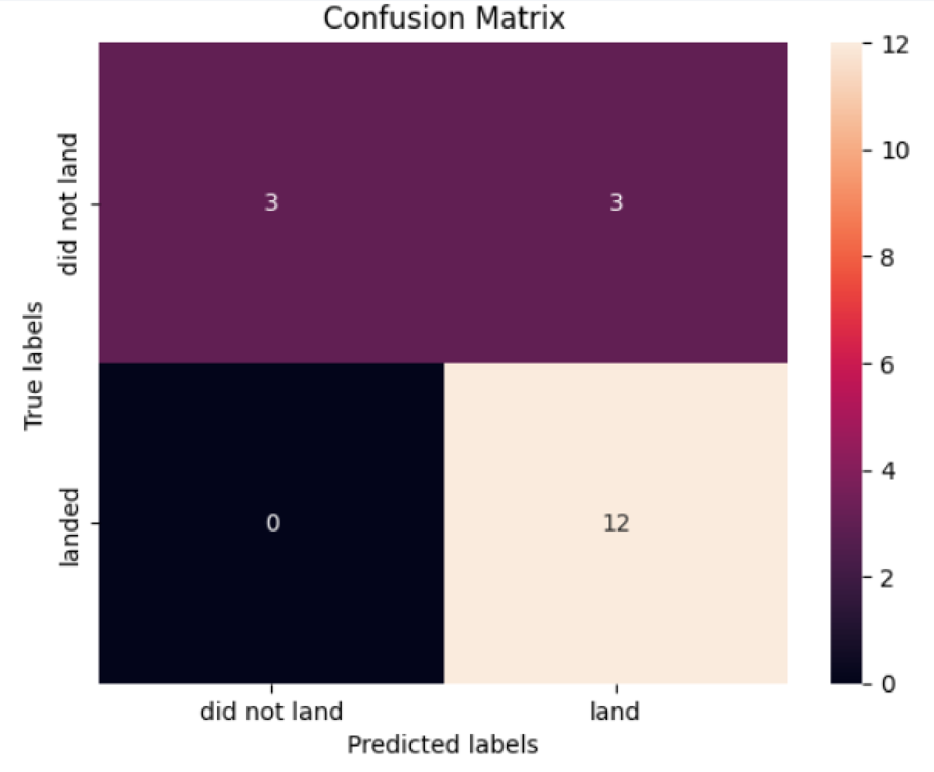


Confusion Matrix of KNN classification method

The confusion matrix of the KNN classification show perfect classification of true positive, but didn’t perform so well in the negative side. Half of launches that failed to land were classified as false positive.

We can calculate by hand the precision, recall and f1-score of the KNN model:

Precision = True Pos / (True Pos + False Pos) = 0.8
Recall = True Pos / (True Pos + False Neg) = 1
F1 score = 0.89



Conclusions

- **Exploratory data analysis and visualization**
  - The success rate of launches has only improved since 2013
  - A higher success rate is observed for higher payloads
  - The orbit VLEO has a high success rate with high payloads, orbits ES-L1, GEO, HEO and SSO have only had one trial flight but all resulted in a success
  - The launch site KSC LC-39A has the highest success rate of all
- **Folium analysis**
  - Launch sites are all located along the coast, far enough from nearest cities for security measures
- **Classification**
  - KNN, SVM and Logistic Regression all have an accuracy score of 0.84, out performing Decision tree

Thank you!