Data Science Capstone project

Muhammad Suwaid Aslam

9/1/2021

Outline

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

Executive Summary

- In this project we go through different data science methodologies like, Data collection with Web Scraping, Data wrangling, Exploratory Data Analysis and Machine Learning.

- We explored patterns in the data and created machine learning models using this data to predict whether Falcon 9 first stage will land successfully and what is the probability of landing successfully on the ground. We found out that it has 75 to 83 percent chances of successfully landing on the launch site.

Introduction

- There is a private Space company called SpaceX, which sends rockets to space with very affordable cost. It charges way less than other Space companies because it has rocket which can be reused after the mission. This significantly reduced the cost for the flight. The name of this rocked is Falcon 9 it has the first stage which can land back on the ground and can be reused.

- We want to find out if the Falcon 9 first stage will land successfully on the ground and also what are the chances of successfully landing on the ground.

Methodology

- Data collection methodology:
    - Data was collected using the SpaceX API and data was also collected using web scraping from the Wikipedia page.
- Perform data wrangling
    - Basically we encoded the categorical labels into numerical labels. We also used one hot encoding technique to label the data.
- Perform exploratory data analysis (EDA) using visualization and SQL
    - We performed bunch of SQL queries on our dataset to find different answers.
- Perform interactive visual analytics using Folium and Plotly Dash
    - We compared different columns together to find the relation using Plotly. We also used Folium to create map based on the location of launch sites.
- Perform predictive analysis using classification models
    - We Split the data into Train and Test set and then define different parameters of our model and test them using GridSearchCV and choose the best parameters for the model. And Finally we compare different models and select the one with high test accuracy.

Methodology

In data Science we use different methods to solve our problems. We can apply different methods to get the most accurate results. In this project we used Data collecting, data wrangling, Exploratory Data analysis and Machine learning methodologies to tackle the problem.

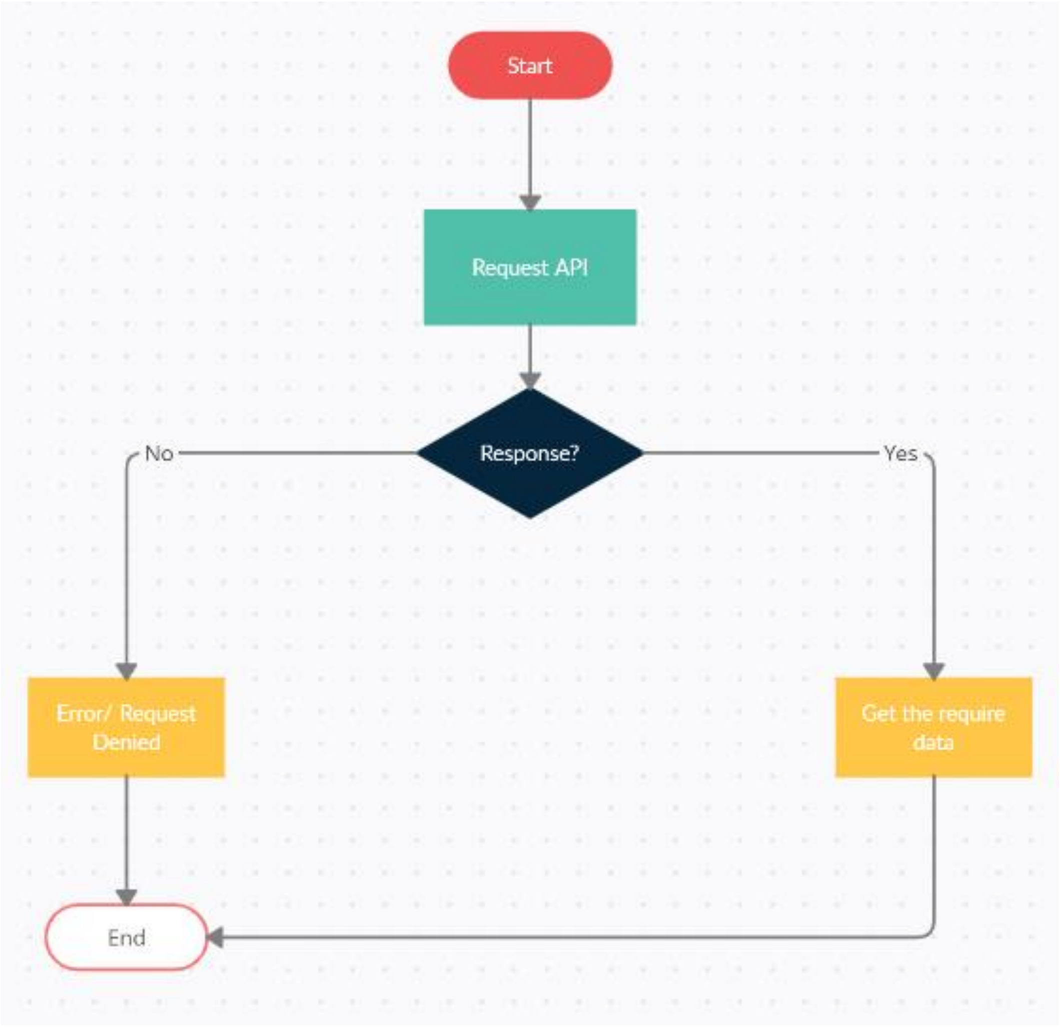
Data collection

- Data was collected using the SpaceX API and data was also collected using web scraping from the Wikipedia page.

- Request the data using request library. Get the data as a response object and convert it into a dataframe and use it.

Data collection – SpaceX API

https://github.com/SuwaidAslam/Applied-Data-Science-Capstone-Project/blob/master/Data%20Collection%20API%20Lab%20Notebook.ipynb

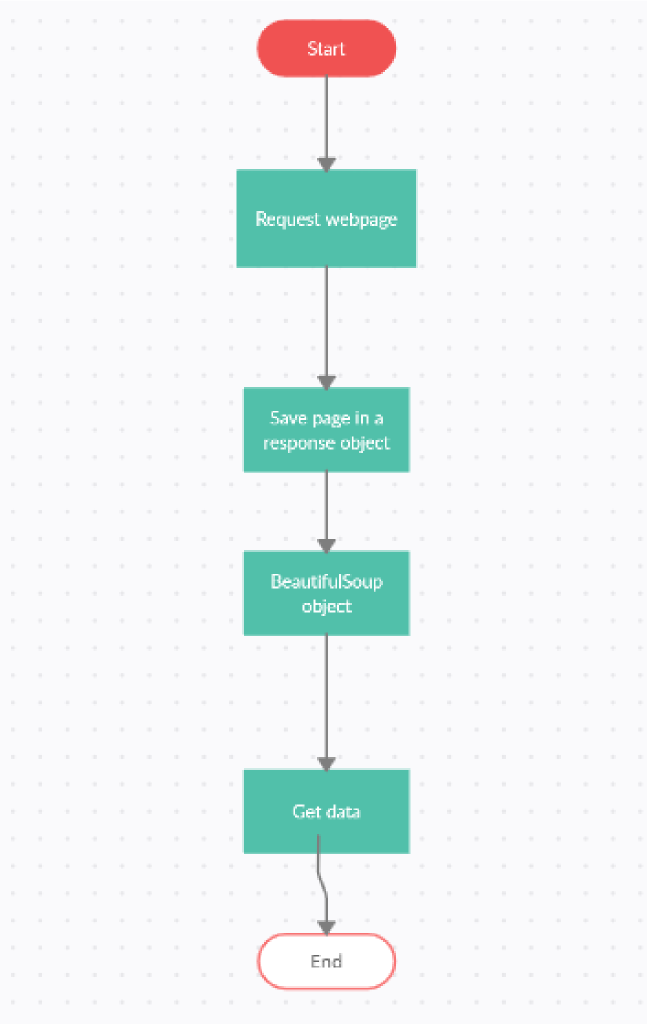


Data collection – Web scraping

Web scrap Falcon 9 launch records with BeautifulSoup:

- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame

https://github.com/SuwaidAslam/Applied-Data-Science-Capstone-Project/blob/master/Data%20Collection%20with%20Web%20Scraping%20lab.ipynb



Data wrangling

- Basically we encoded the categorical labels into numerical labels. We also used one hot encoding technique to label the data.

- Converted Categorical data into numerical labeled data and performed one hot encoding as well.

- https://github.com/SuwaidAslam/Applied-Data-Science-Capstone-Project/blob/master/EDA%20Lab.ipynb

EDA with data visualization

Scatter, bar, and line plots were created to explore:

- Flight number versus launch site location
- Payload mass versus launch site location
- Orbit type versus success rate
- Flight number versus orbit type
- Payload mass versus orbit type

- https://github.com/SuwaidAslam/Applied-Data-Science-Capstone-Project/blob/master/EDA%20with%20Visualization%20lab.ipynb

EDA with SQL

SQL queries performed for extracting useful information from the SpaceX data set include

- Display the names of the unique launch sites in the space mission
- Display some records of launch sites matching criteria
- Display total payload masses carried by boosters launched for specific SpaceX customers
- Display average payload mass carried by booster versions
- List important dates such as first successful landing outcomes at different landing pads
- List the names of boosters having success in drone ship and specified payload mass ranges
- List total number of successful and failed mission outcomes
- List the boosters which have carried the maximum payload mass
- List boosters for specific years with kinging outcome for certain locations and launch sites.
- Recount successful landing outcomes for specific time frames:

//github.com/SuwaidAslam/Applied-Data-Science-Capstone-Project/blob/master/EDA%20with%20SQL%20lab.ipynb

Build an interactive map with Folium

Analyzed existing launch locations for discovering factors describing optimal locations for building launch sites. Some of the items for this analysis include:

- Interactive map mark ups of all launch sites
- Mark up of successful/failed launches for each site
- Distances between launch sites to proximal infrastructure and geographical entities.

github.com/SuwaidAslam/Applied-Data-Science-Capstone-Project/blob/master/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

A Plotly Dash application was created for users to perform interactive visual analytics on SpaceX launch data in real-time. Items in the dasboard include:

- Launch Site Drop-down Input Component
- Launch success pie-chart ‘
- Range Slider to Select Payload
- Success-Payload scatter plot

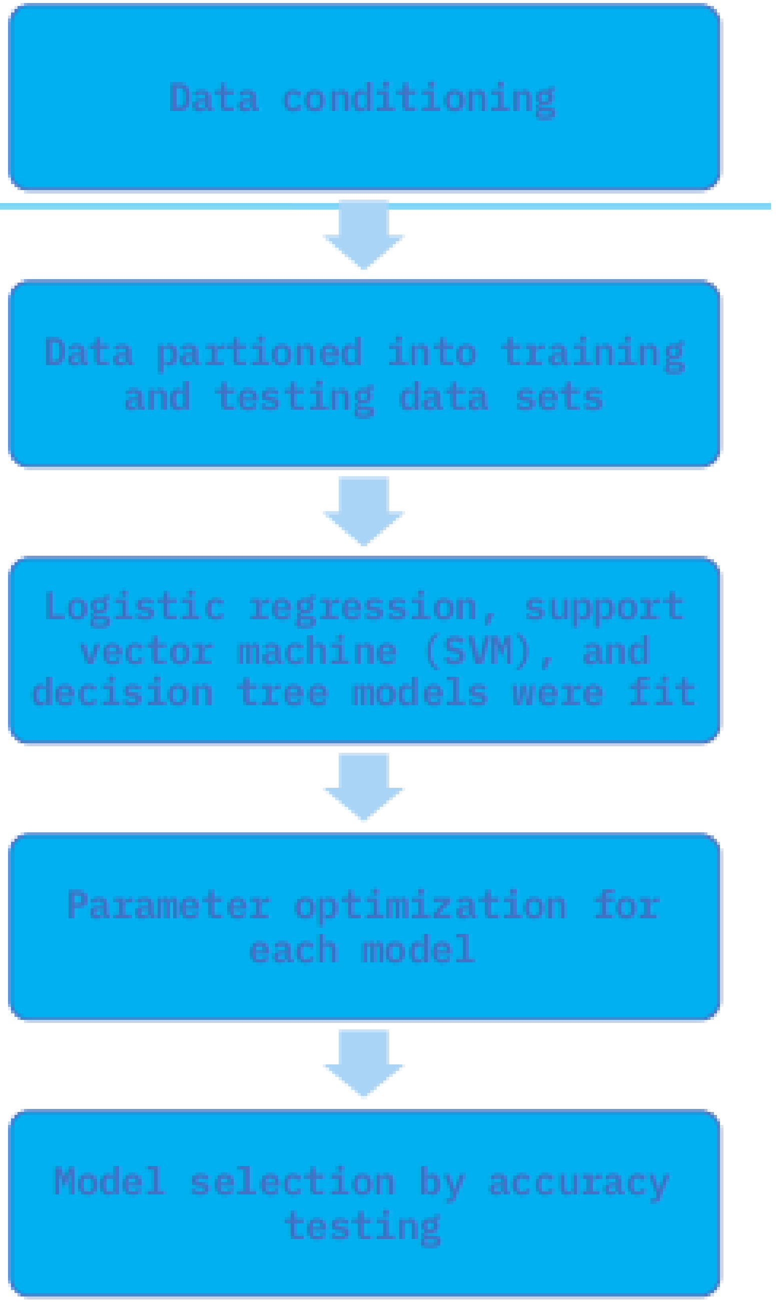
Link:

https://github.com/SuwaidAslam/Applied-Data-Science-Capstone-Project/blob/master/spacex_dash_app.py

Predictive analysis (Classification)

- We Split the data into Train and Test set and then define different parameters of our model and test them using GridSearchCV and choose the best parameters for the model. And Finally we compare different models and select the one with high test accuracy.

https://github.com/SuwaidAslam/Applied-Data-Science-Capstone-Project/blob/master/Machine%20Learning%20Prediction%20lab.ipynb



Data conditioning

Data partioned into training and testing data sets

Logistic regression, support vector machine (SVM), and decision tree models were fit

Parameter optimization for each model

Model selection by accuracy testing

Results

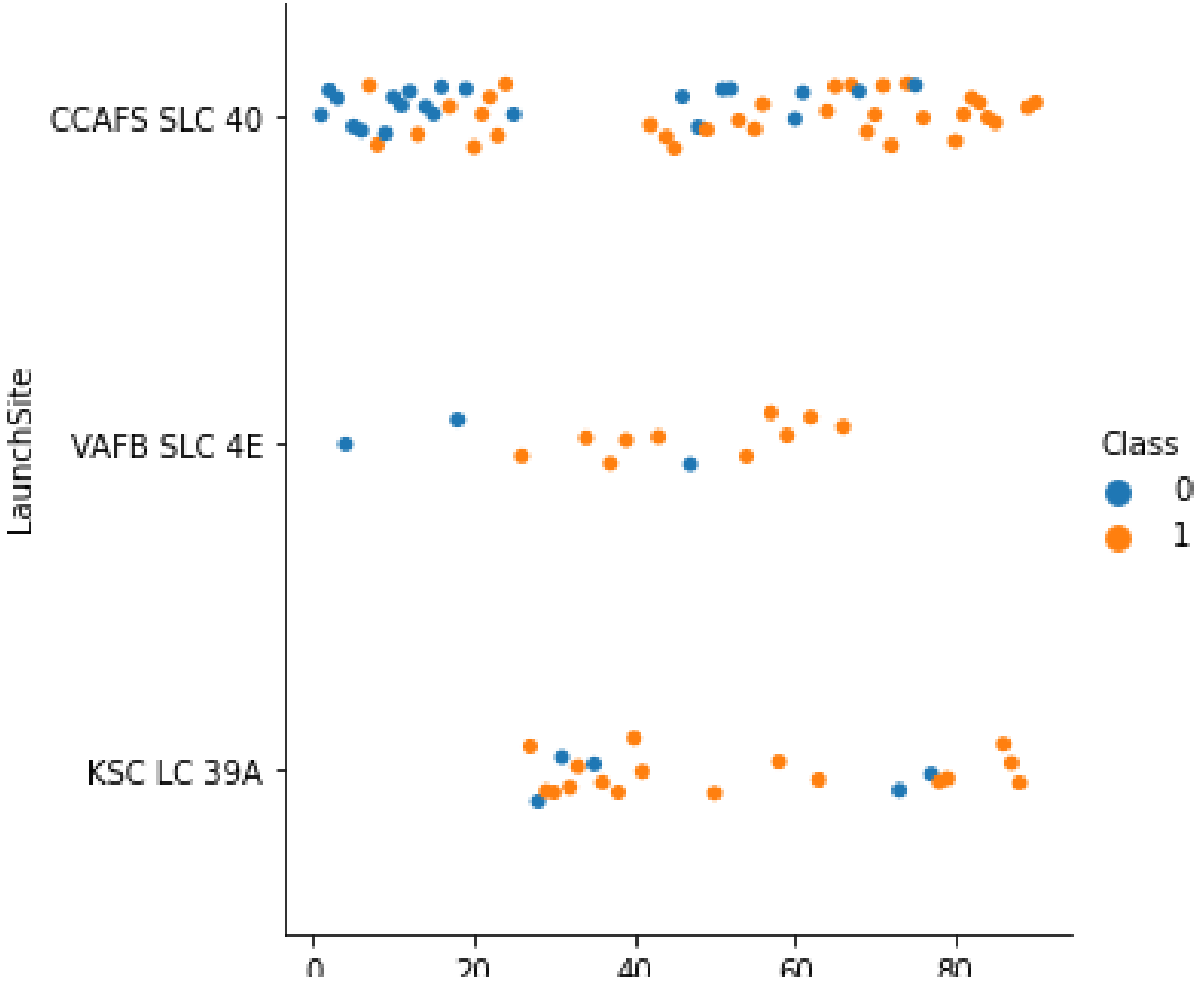
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

EDA with Visualization

Flight Number vs. Launch Site

This scatter plot shows a plot between Launch site and Flight numbers. As we can see that the CCAFS SLC 40 launch site has the highest launches in the history.

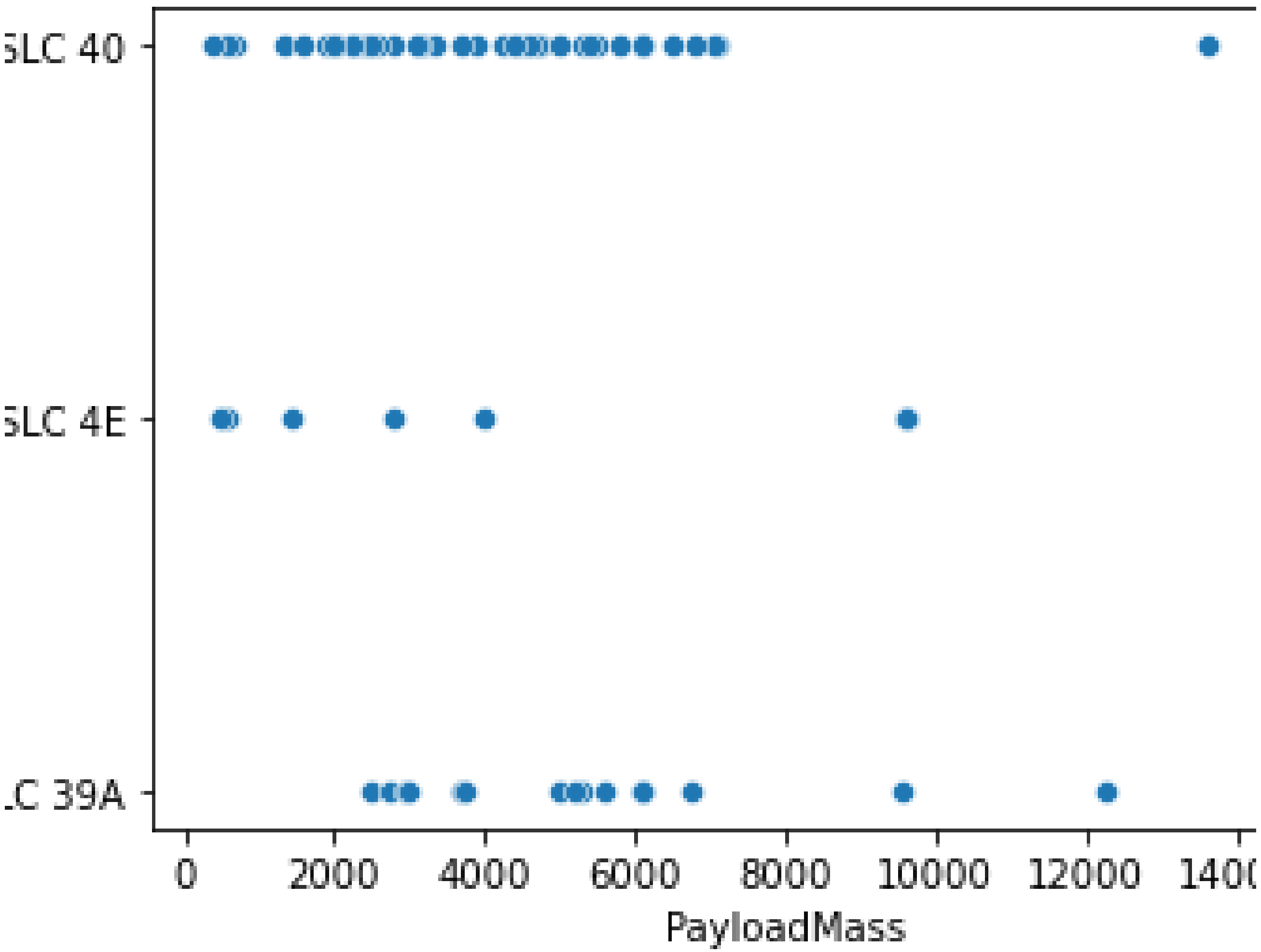
We can also see successfully landed rocket with orange and failed with blue.



Payload vs. Launch Site

This scatter plot shows a plot between Launch site and Payloadmass.

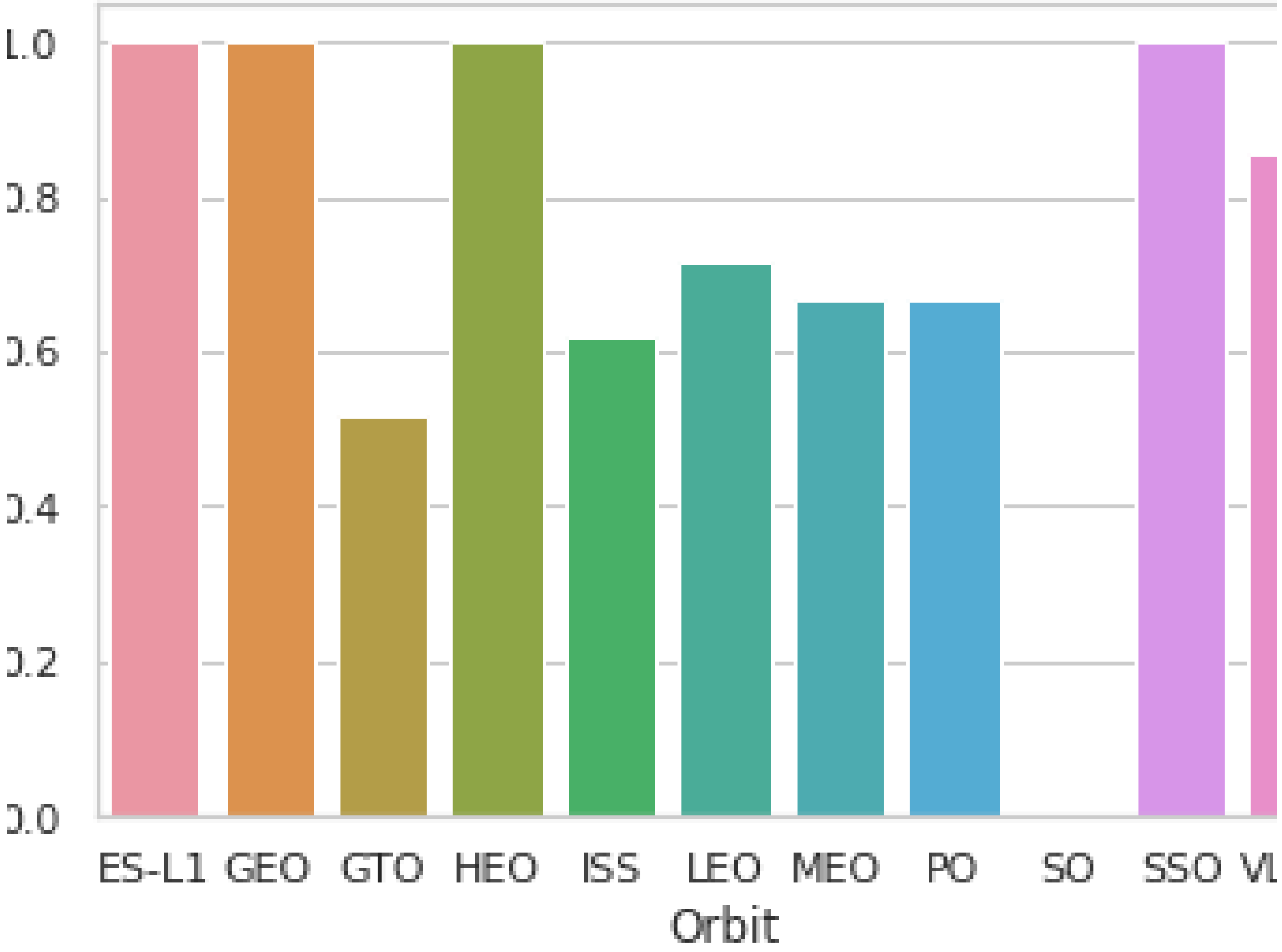
We can observe from this pot that low mass rockets was mostly from the upper location.



Success rate vs. Orbit type

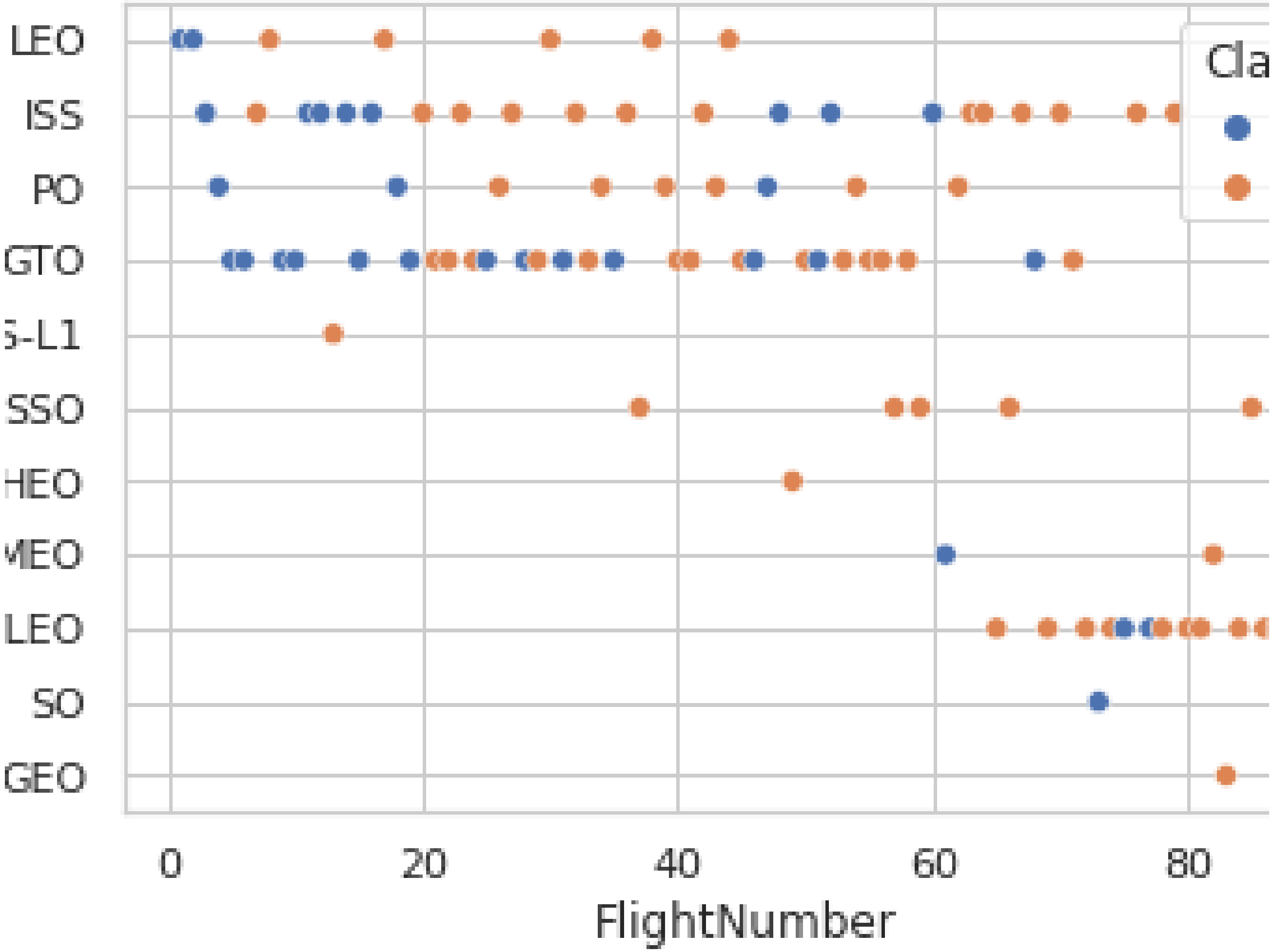
This bar graph tells us the relation ship between success rate vs orbit type.

We can see ES-L1, GEO, and SSO orbit has the highest success rate.



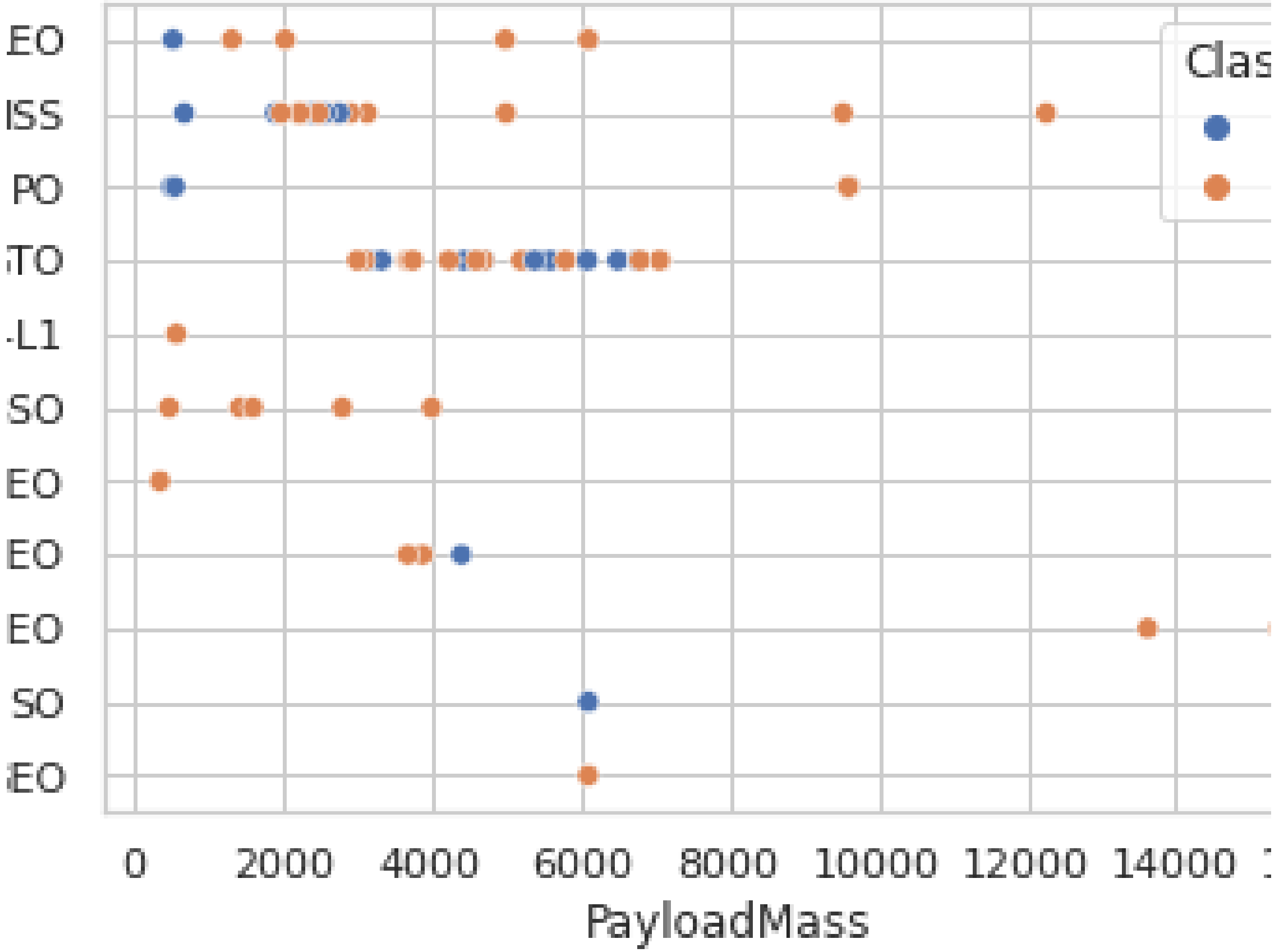
Flight Number vs. Orbit type

This plot shows the successes (1) and failures (0) of launches for the flights from the launch sites per orbit type. It is apparent that higher flight numbers enjoy greater success.



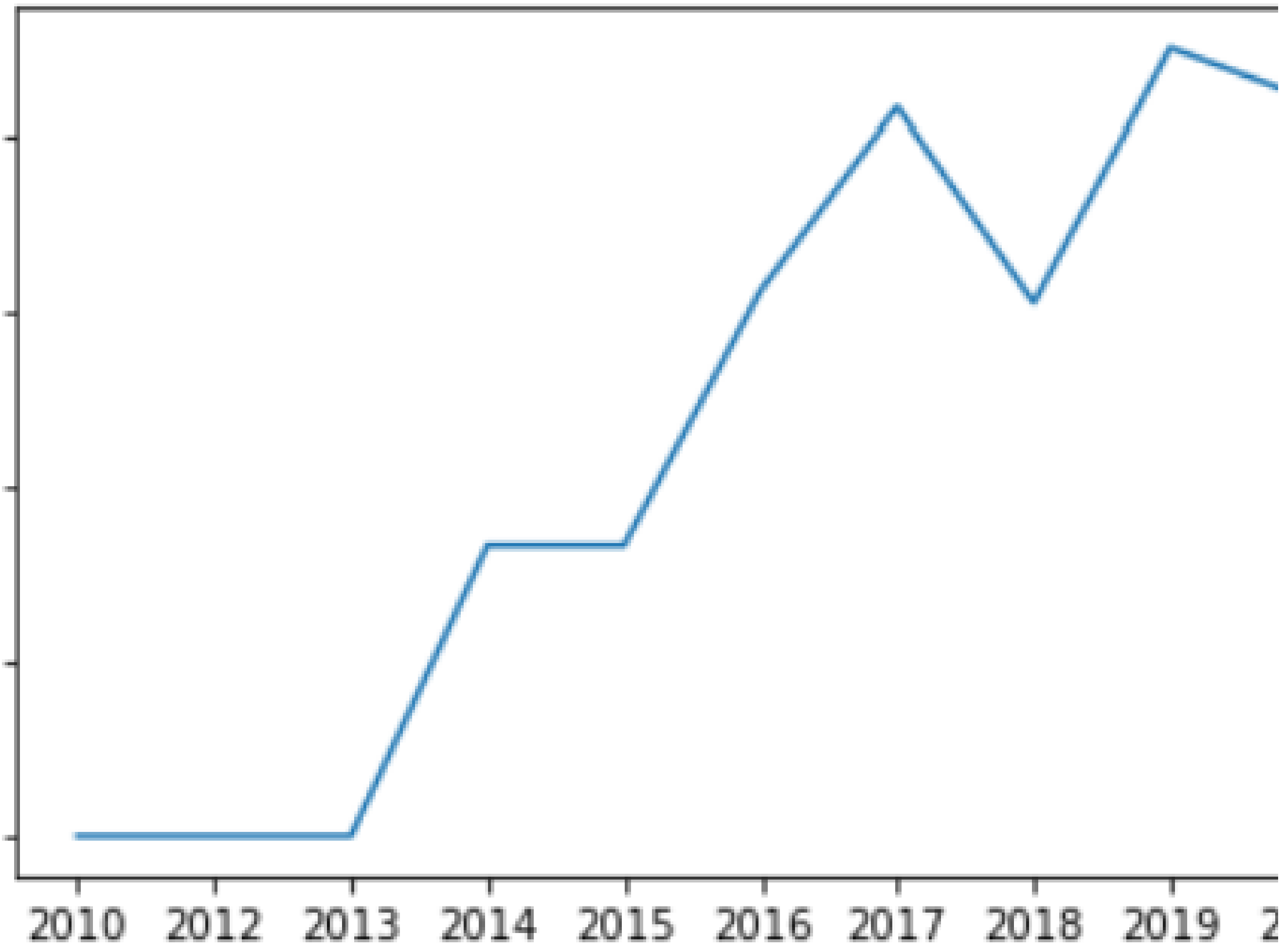
Payload vs. Orbit type

This plot shows the successes '(1) and failures (0) of launches for the orbit types per payload mass. It is apparent that lighter payloads dominate the mass distribution. GTO and MEO orbits have relative narrow payload ranges.



Launch success yearly trend

Trend of obvious increasing success rate over time is apparent. SpaceX learns from each launch to be more success



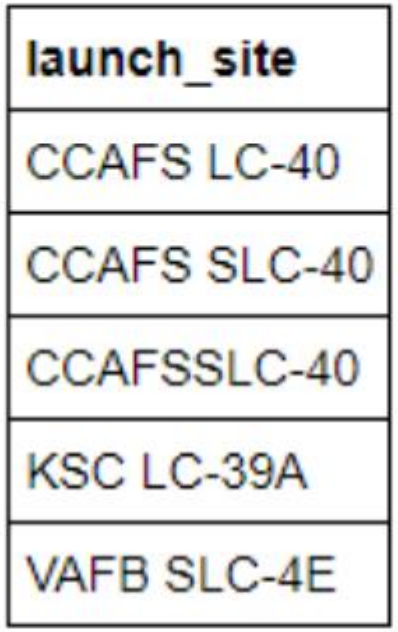
EDA with SQL

All launch site names

SELECT DISTINCT Launch_Site FROM SPACEXTABLE;

Query yields:

This is run against the database created from the main data set that was compiled

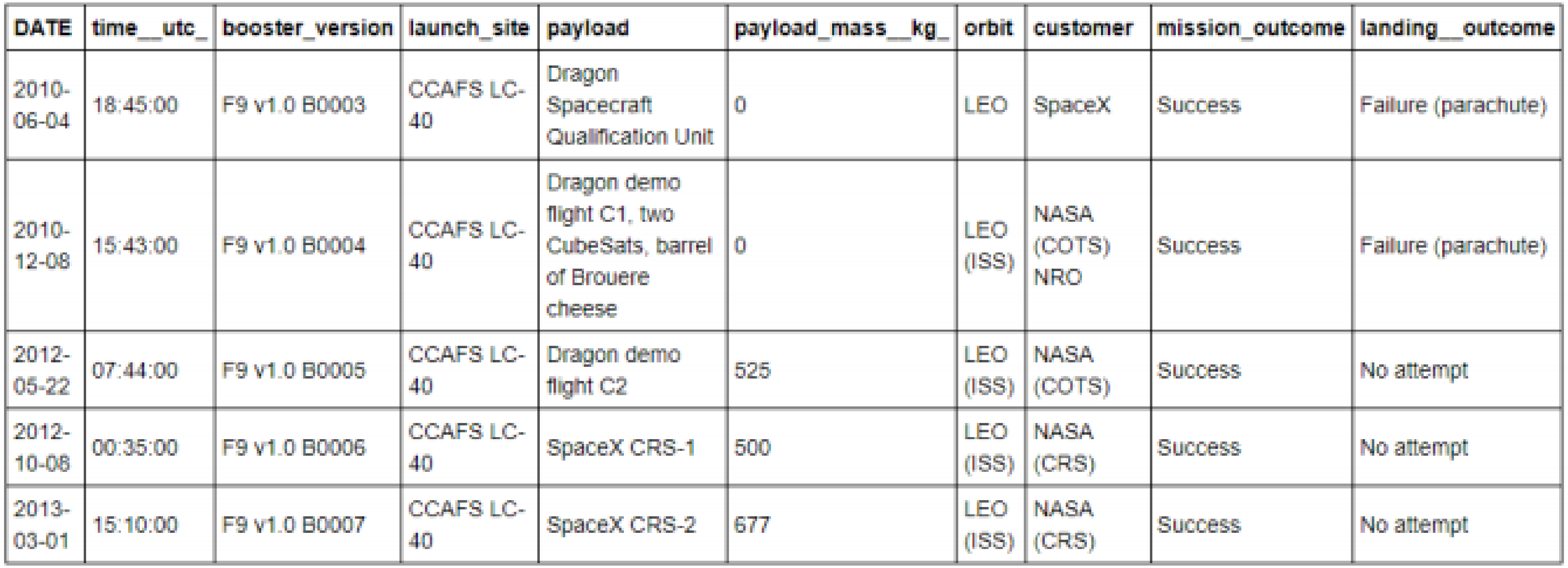


launch_site
CCAFS LC-40
CCAFS SLC-40
CCAFSSLC-40
KSC LC-39A
VAFB SLC-4E

Launch site names begin with `CCA`

SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%’;

Table below contains the first five records of this query

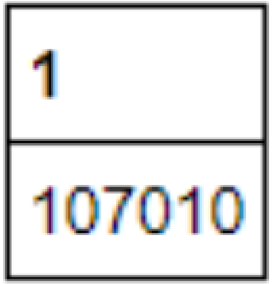


DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

```
SELECT SUM(payload_mass__kg_) FROM SPACEXTABLE WHERE Customer LIKE '%NASA%’;
```

Calculates the total payload mass launced for customer NASA. Equals 107,010 kg

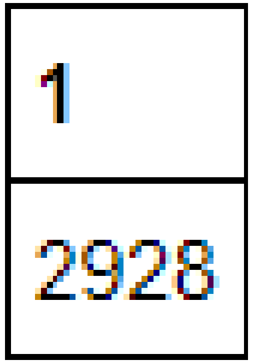


1
107010

Average payload mass by F9 v1.1

SELECT AVG(payload_mass__kg_) FROM SPACEXTABLE WHERE booster_version='F9 v1.1’

Calculates the total payload mass launced by booster F9 v1.1. Equals 2928 kg

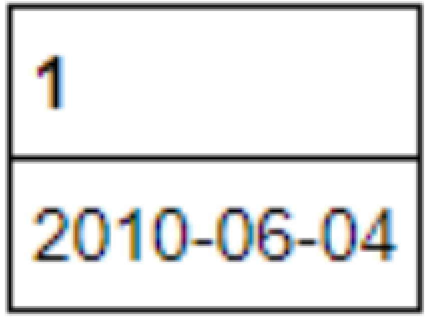


1
2928

First successful ground landing date

SELECT MIN(Date) FROM SPACEXTABLE WHERE mission_outcome='Success’;

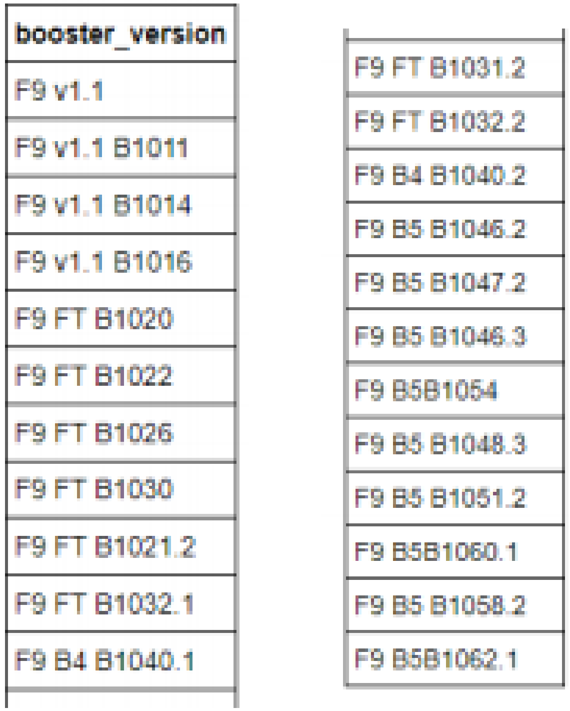
We can see that the first successful mission was launches June 4, 2010.



1
2010-06-04

Successful drone ship landing with payload between 4000 and 6000

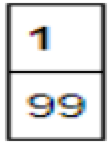
SELECT booster_version FROM SPACEXTABLE WHERE mission_outcome='Success' AND payload_mass__kg_ BETWEEN 4000 AND 6000;



booster_version
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 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5B1054
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

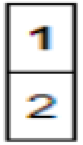
Total number of successful and failure mission outcomes

- SELECT COUNT(mission_outcome) FROM SPACEXTABLE WHERE mission_outcome='Success’;



1
99

- SELECT COUNT(mission_outcome) FROM SPACEXTABLE WHERE mission_outcome!='Success’;

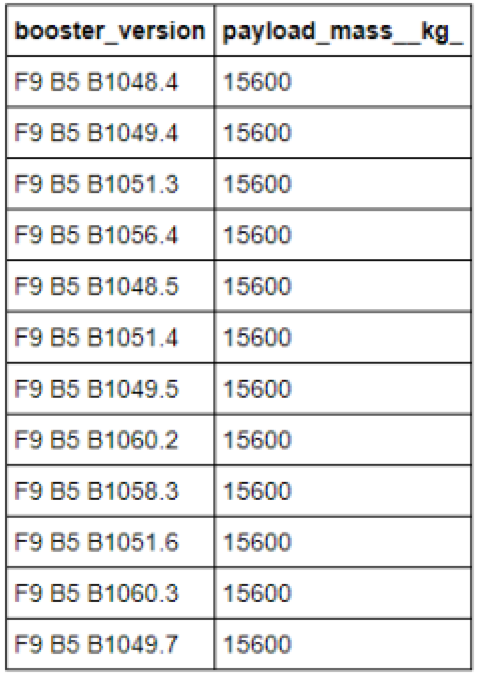


1
2

Boosters carried maximum **payload**

```
SELECT booster_version, payload_mass__kg_ FROM SPACEXTABLE
WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_)
FROM SPACEXTABLE);
```

We can see that all boosters that carried the maximum payload where of the B5 family

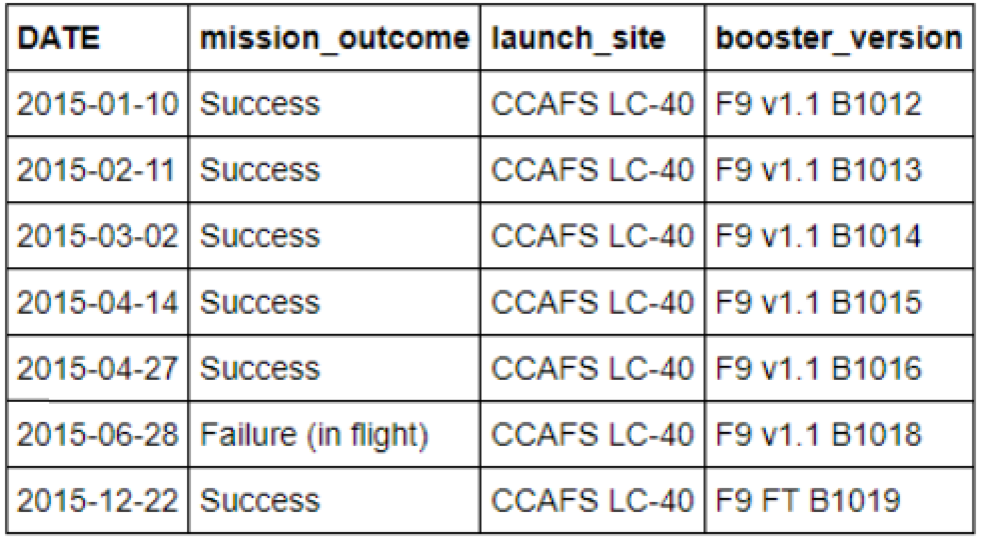


booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 launch records

SELECT Date, mission_outcome, Launch_Site, booster_version FROM SPACEXTABLE WHERE Date LIKE '2015%’;

Only a single failed launch for 2015

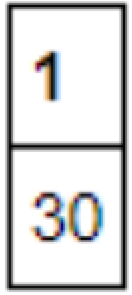


DATE	mission_outcome	launch_site	booster_version
2015-01-10	Success	CCAFS LC-40	F9 v1.1 B1012
2015-02-11	Success	CCAFS LC-40	F9 v1.1 B1013
2015-03-02	Success	CCAFS LC-40	F9 v1.1 B1014
2015-04-14	Success	CCAFS LC-40	F9 v1.1 B1015
2015-04-27	Success	CCAFS LC-40	F9 v1.1 B1016
2015-06-28	Failure (in flight)	CCAFS LC-40	F9 v1.1 B1018
2015-12-22	Success	CCAFS LC-40	F9 FT B1019

Rank success count between 2010-06-04 and 2017-03-20

SELECT COUNT(mission_outcome) FROM SPACEXTABLE WHERE mission_outcome ='Success' AND Date BETWEEN '2010- 06-04' AND '2017-03-20’;

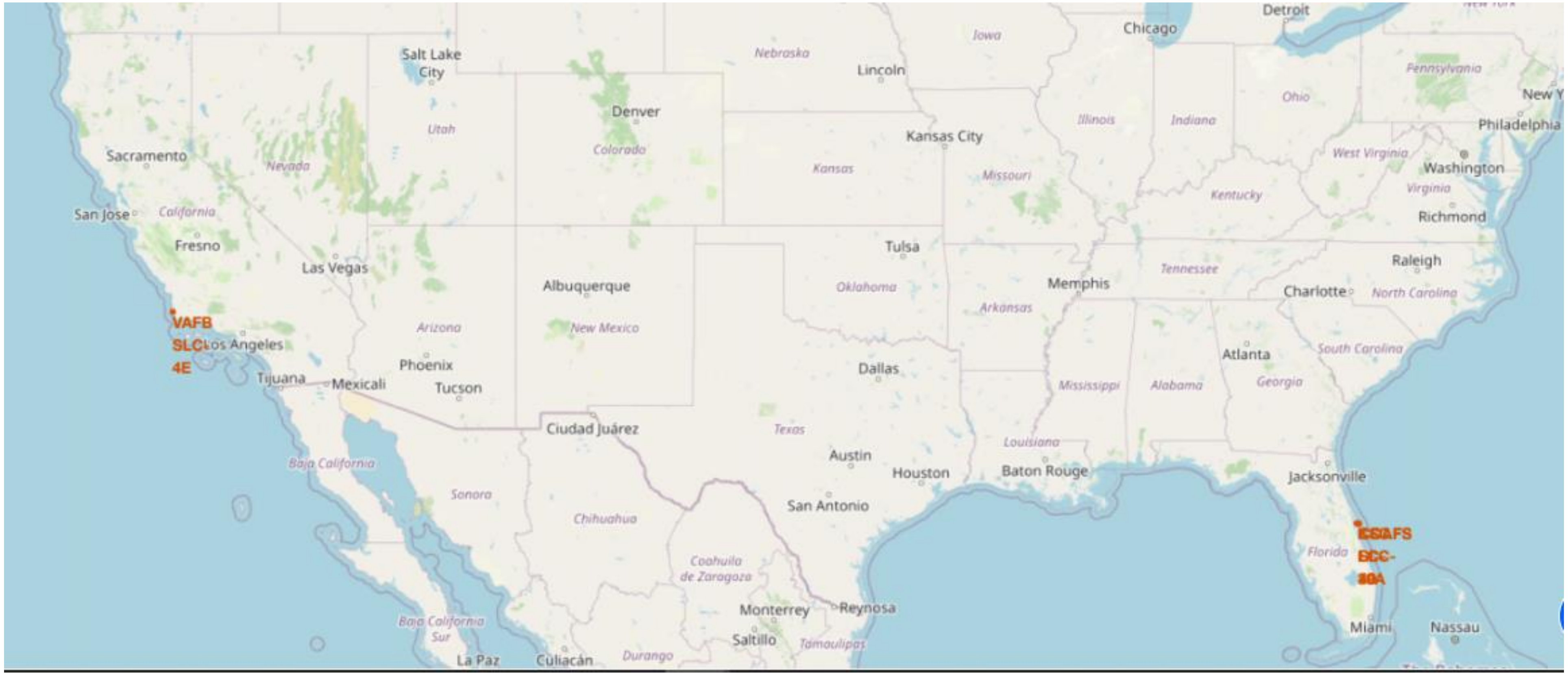
We see that there were thirty successful launches between June 2010 and March 2017.



1
30

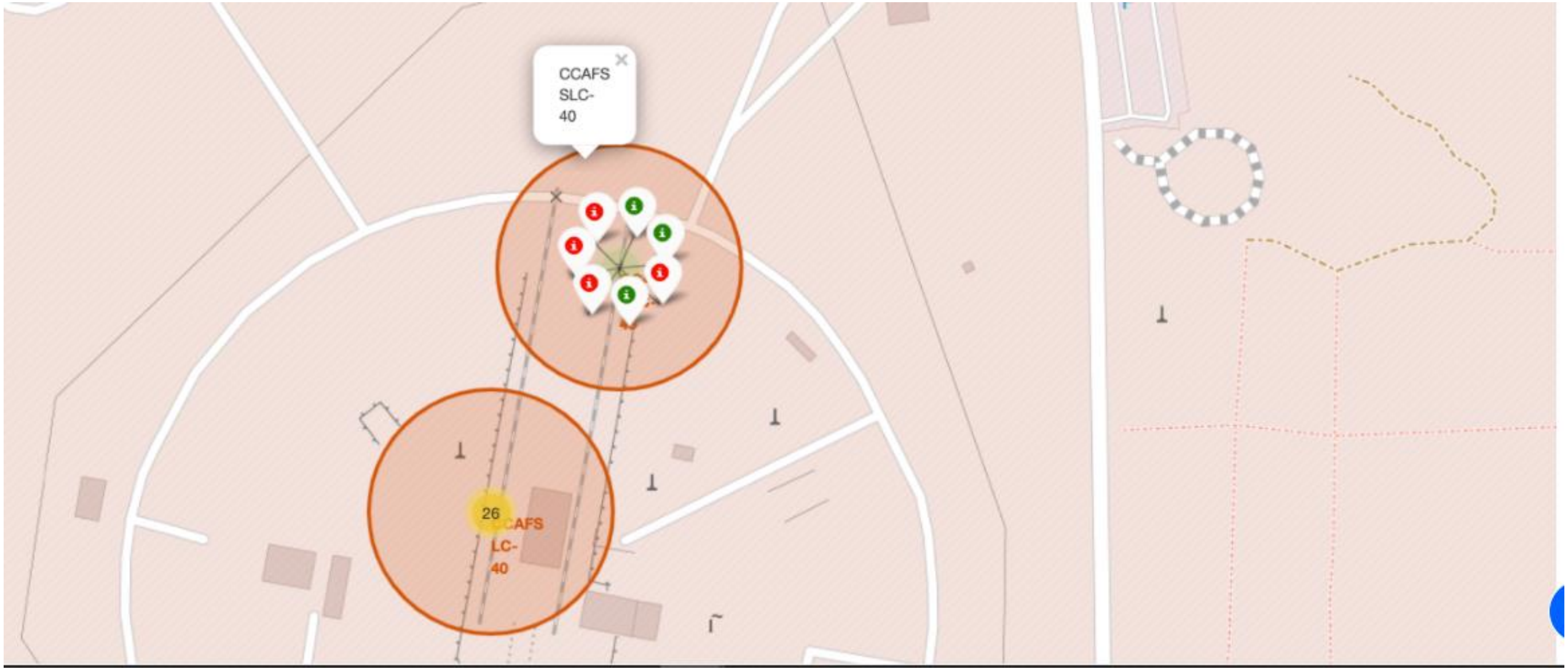
Interactive map with Folium

Launch sites Map locations



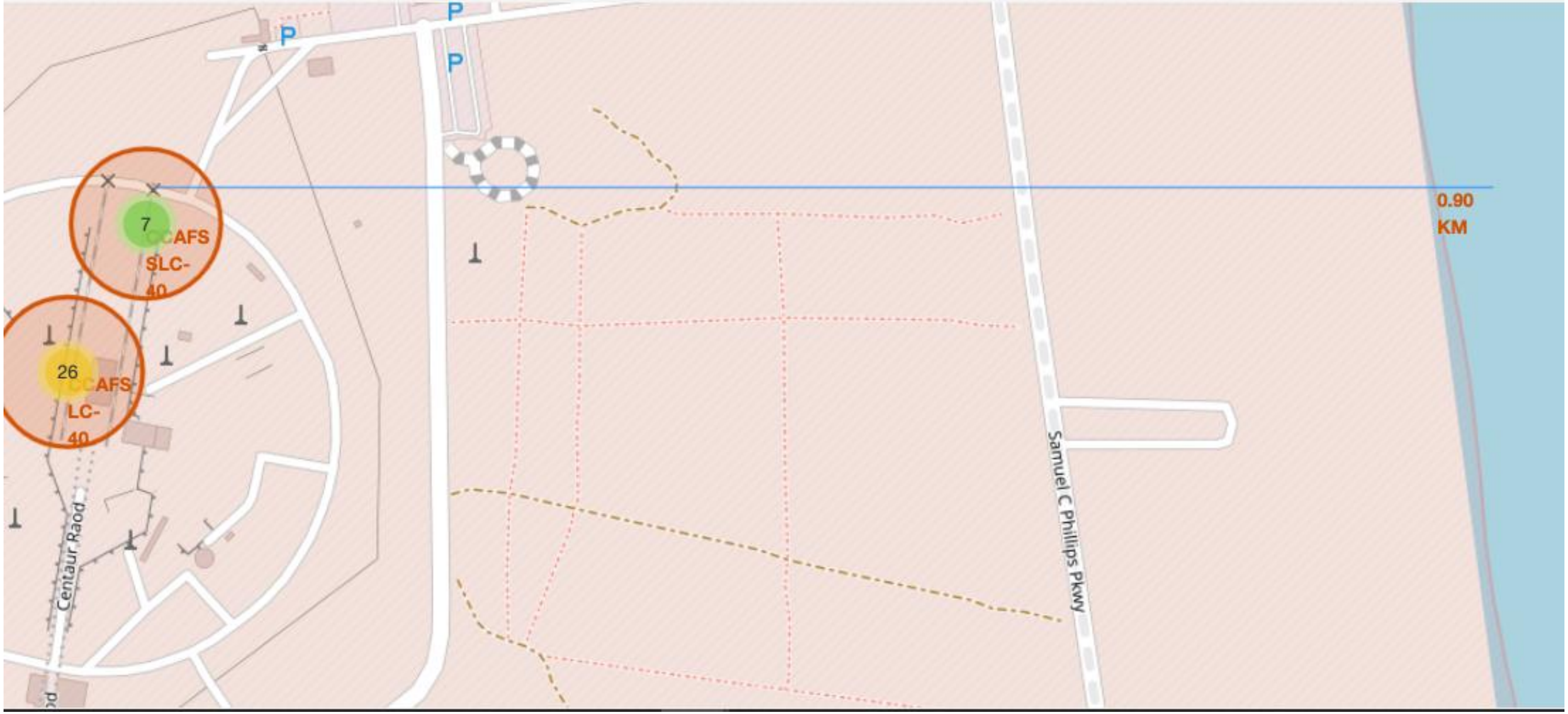
Notice that all SpaceX launch sites are located in the southern US.

Color Labeled Map



Circle object and marker objects were used. And we also colored them.

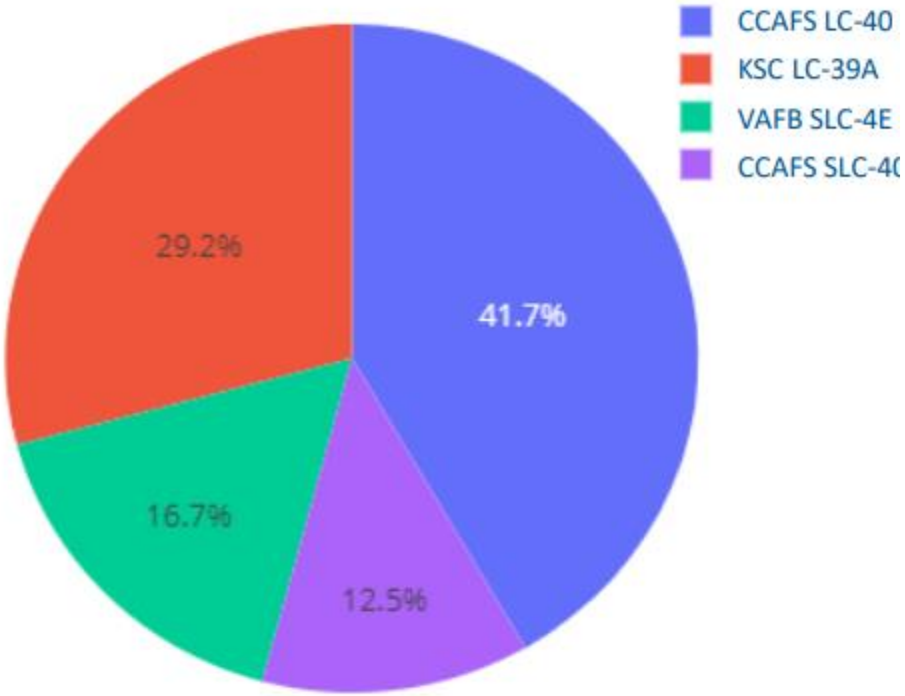
Map of launch site to its proximities



The closest railway to the selected launch site is ~0.1 km and the Atlantic Coast is ~0.9 km away

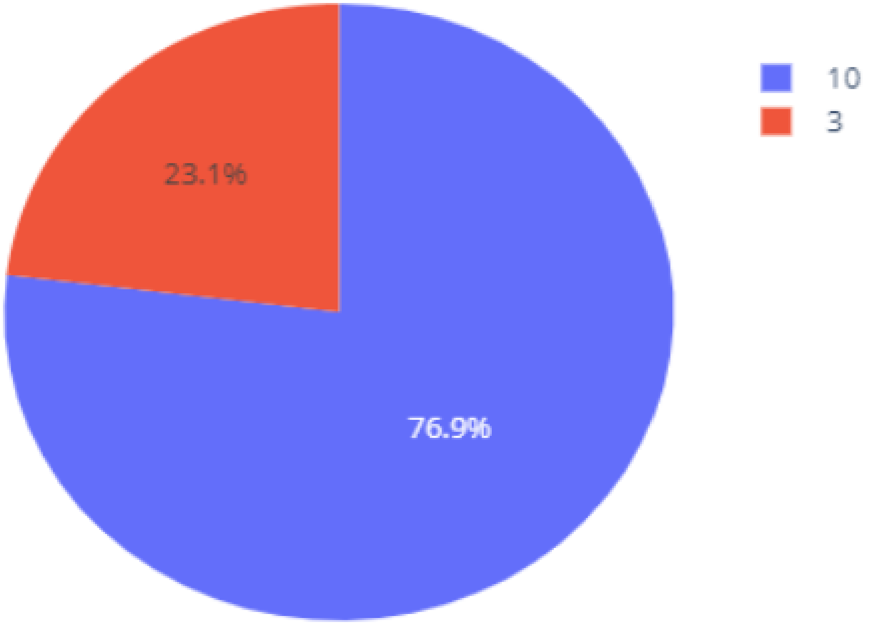
Build a Dashboard with Plotly Dash

Mission Success for All Launch Sites



The chart shows the percent success for each launch site for the total number of mission outcomes.

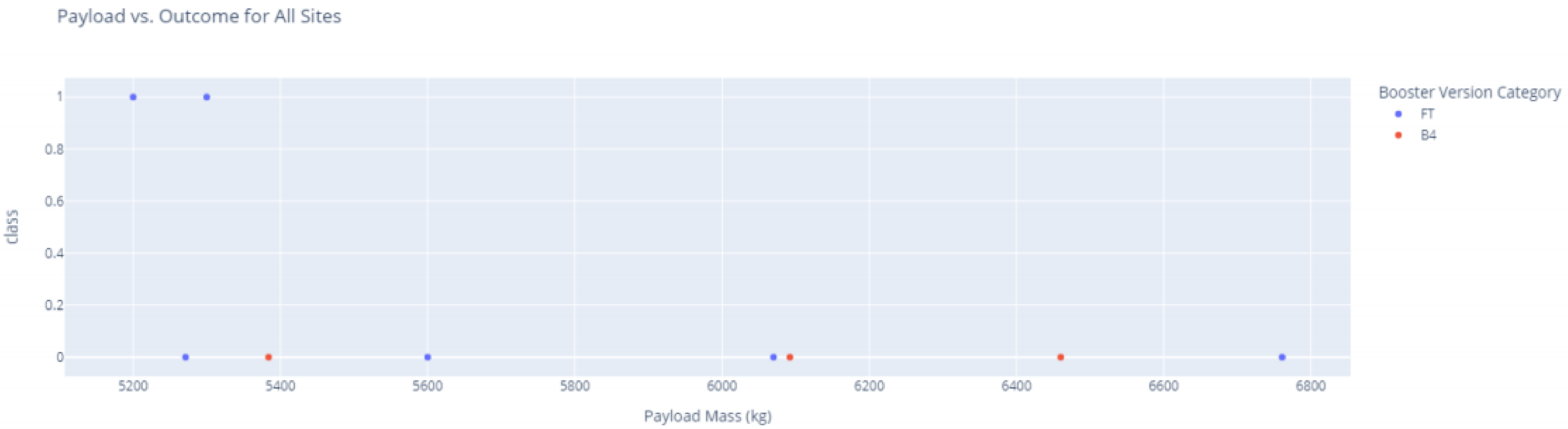
Success Count of KSC LC-39A



This chart shows the success rate of the

Payload Mass and Success Rate

For payload mass range from 5000 to 7500 kg,we see no mission successes for payloads > ~5300 kg for either booster category

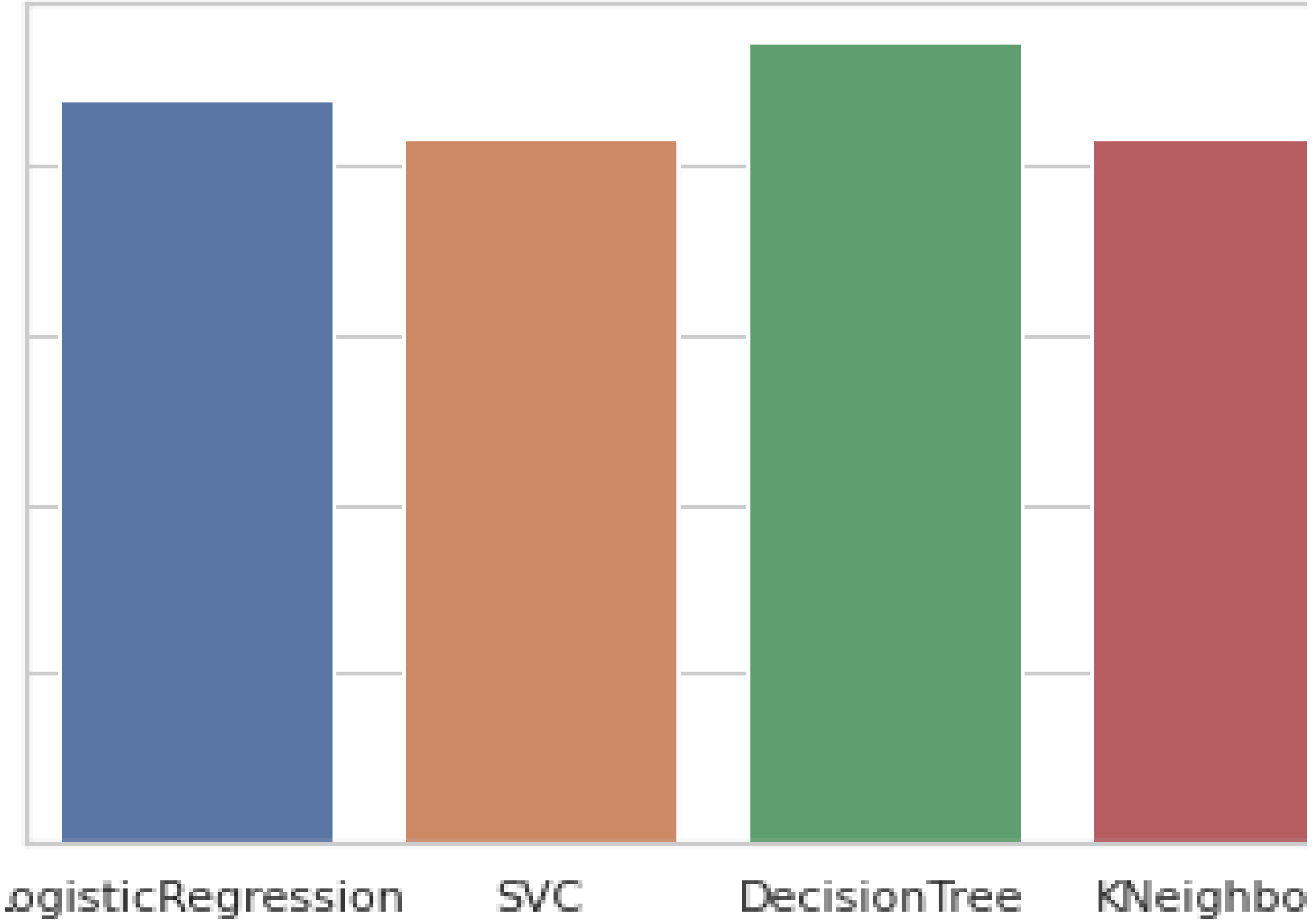


Predictive analysis (Classification)

We Split the data into Train and Test set and then define different parameters of our model and test them using GridSearchCV and choose the best parameters for the model. And Finally we compare different models and select the one with high test accuracy.

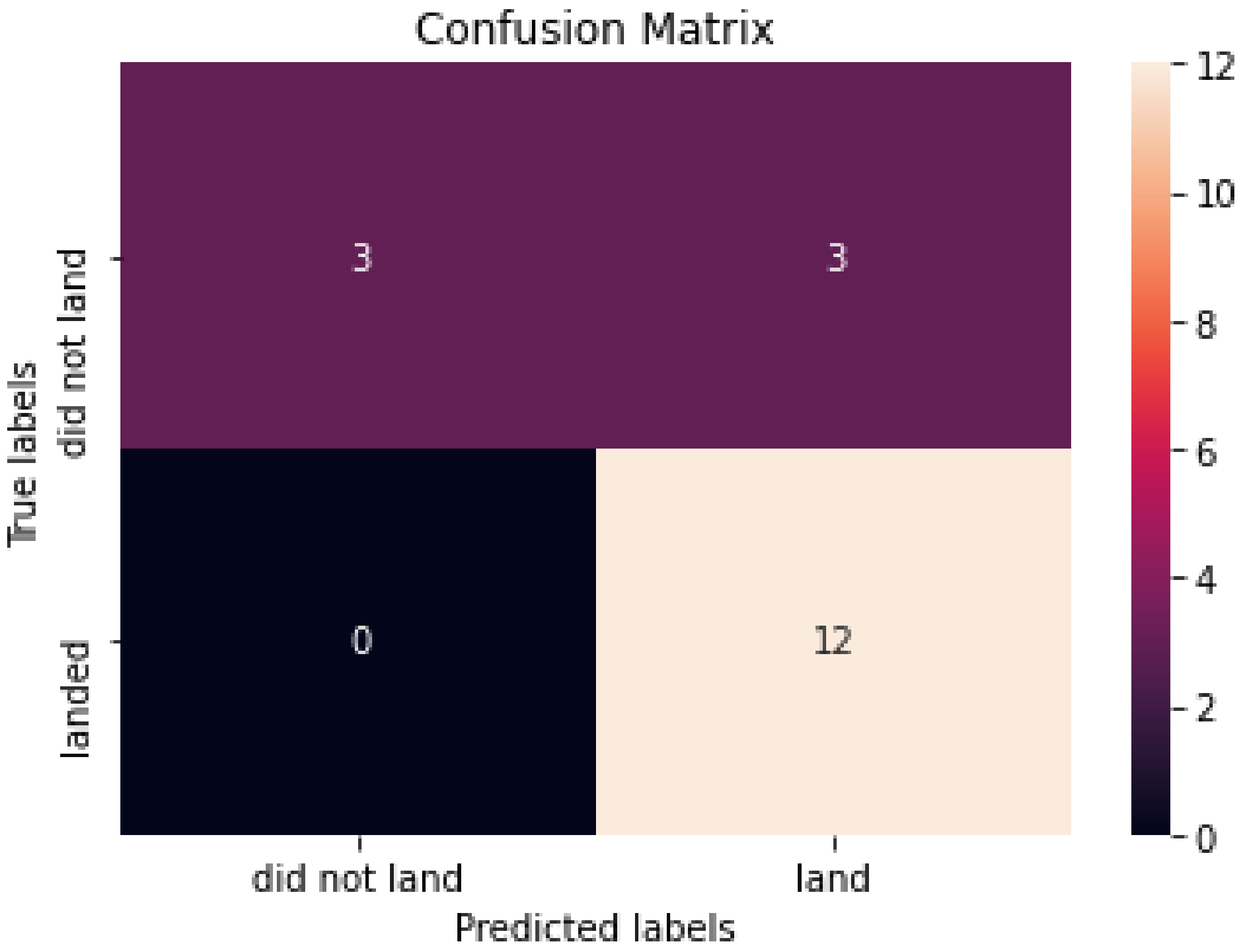
Classification Accuracy

As we can see Decision Tree classifier has the highest accuracy score.



Confusion Matrix

It is the Confusion matrix of Decision tree classifier.



CONCLUSION

- Logistic Regression and Decision tree models all perform equally well in terms of accuracy
- Interactive tools were created for the client to use for launch site and configuration assessment.
- Mission success is greater when payloads are < 5300 kg.
- Mission success rates improve with experience.

APPENDIX

- All code and external information used in the compilation of this project can be found in the git hub repository at this link:

https://github.com/SuwaidAslam/Applied-Data-Science-Capstone-Project/tree/master