IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Olga Sobolevskaya
26/04/2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Methodology:

- Collection the data using SpaceX Rest API and web scraping Wiki page
- Processing the data dealing with missing values, checking column types and determining training labels
- Exploratory Data Analysis using Data Visualization and SQL queries
- Interactive visual analytics using Folium and Plotly
- Predictive analysis using classification models

Results:

- EDA results
- Interactive analysis results
- Predictive analysis results

Introduction

- **SpaceX** is one of the most successful companies that makes space travel affordable for everyone. It advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each. Much of the savings is because **SpaceX can reuse the first stage**.

- If we can determine **whether the first stage will land**, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

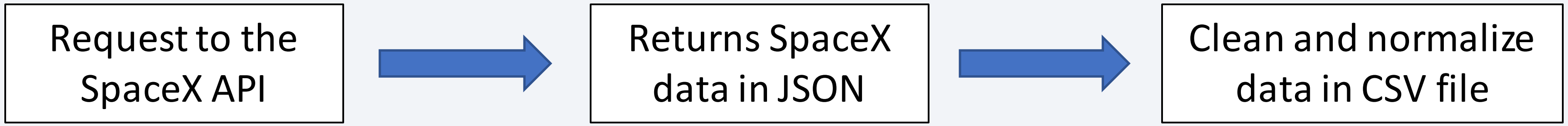
Methodology

Executive Summary

- Data collection methodology:
  - Collection data from the SpaceX Rest API
  - Web Scrapping from Wikipedia
- Data wrangling
  - Performing Exploratory Data Analysis and determining Training Labels
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - SVM, Logistic Regression, Decision Tree, and KNN models were build and evaluated

Data Collection

- Firstly, I collected the Falcon 9 rocket launch data from the SpaceX API, made sure the data was in the correct format, cleaned the data, and dealt with missing values:

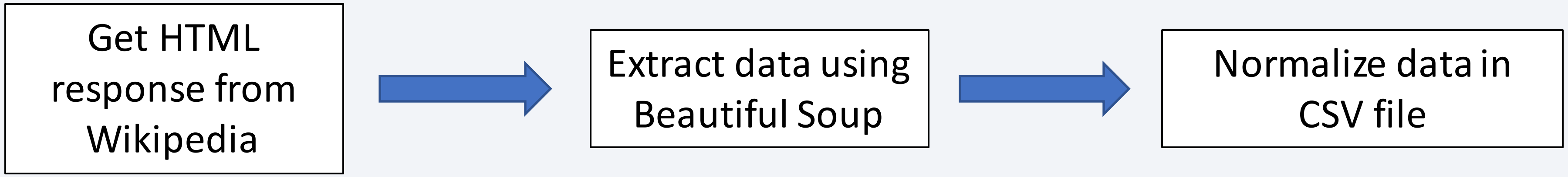


Request to the SpaceX API

Returns SpaceX data in JSON

Clean and normalize data in CSV file

- Secondly, I performed web scrapping to collect Falcon 9 historical launch records from a Wikipedia page:

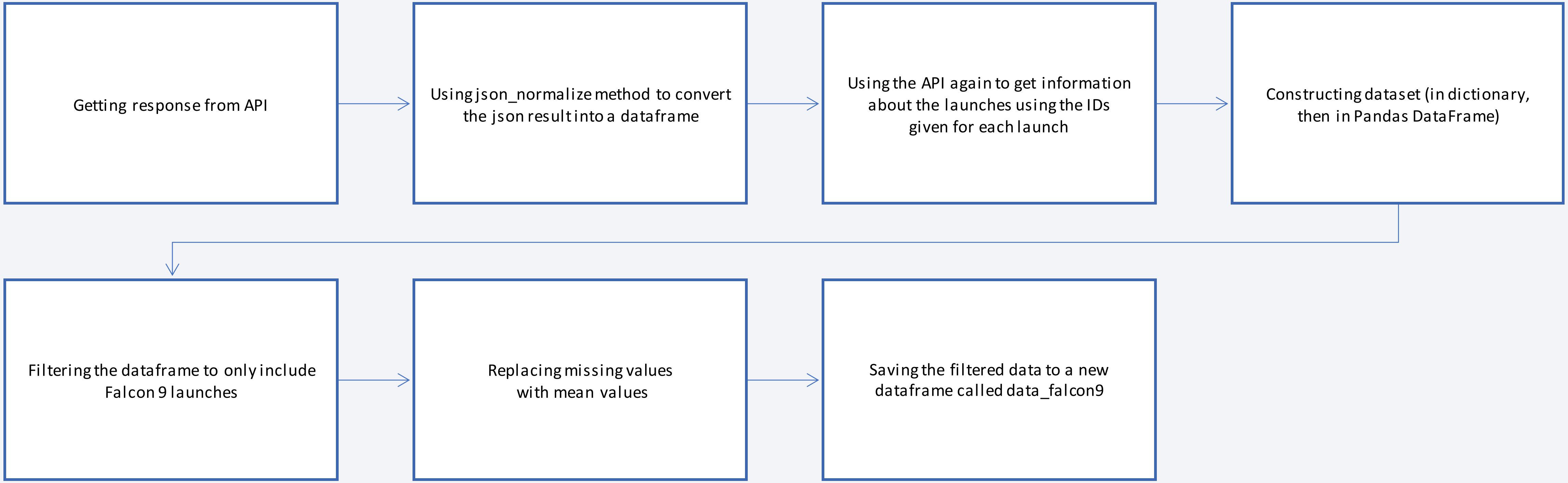


Get HTML response from Wikipedia

Extract data using Beautiful Soup

Normalize data in CSV file

Data Collection – SpaceX API



Getting response from API

Using json_normalize method to convert the json result into a dataframe

Using the API again to get information about the launches using the IDs given for each launch

Constructing dataset (in dictionary, then in Pandas DataFrame)

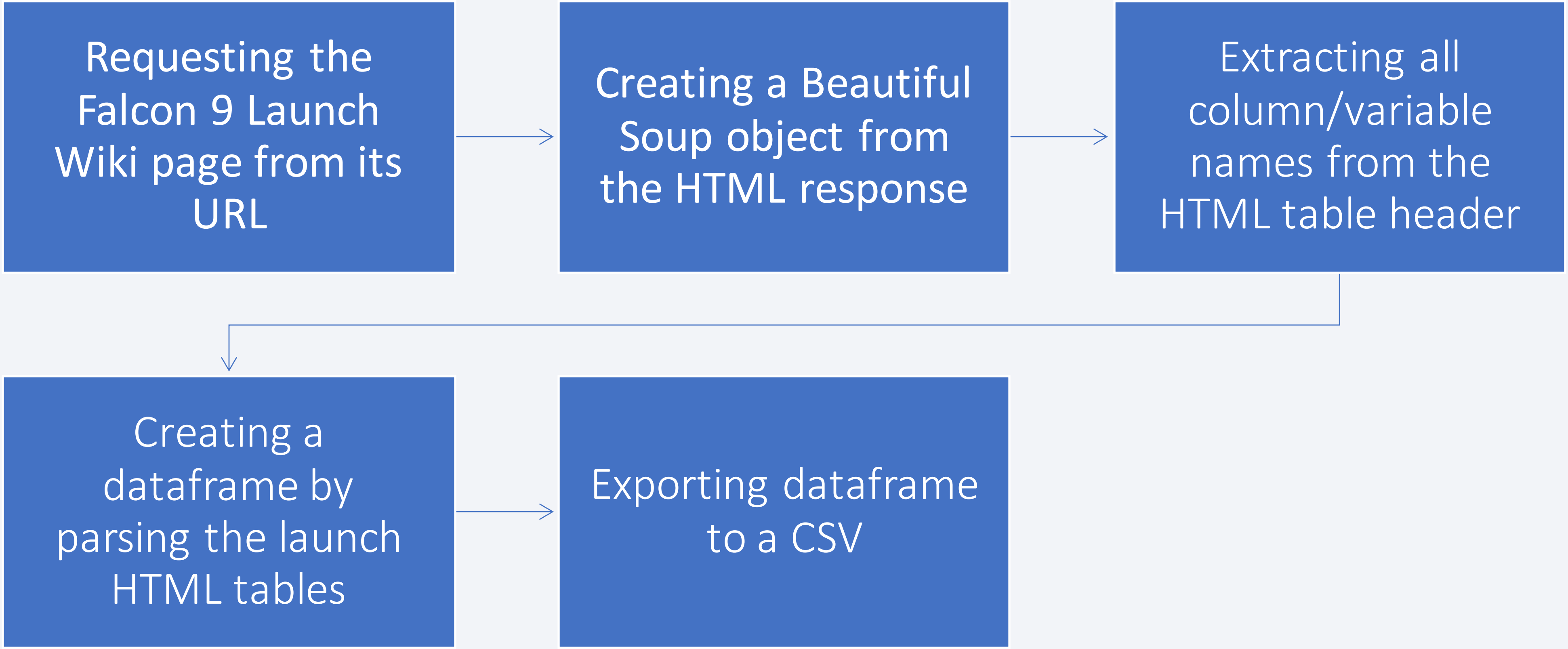
Filtering the dataframe to only include Falcon 9 launches

Replacing missing values with mean values

Saving the filtered data to a new dataframe called data_falcon9

https://github.com/OlyaSobolevskaya/Data-Science-Project-SpaceX/blob/main/capstone-spacex-data-collection-api.ipynb

Data Collection - Scraping



Requesting the Falcon 9 Launch Wiki page from its URL

Creating a Beautiful Soup object from the HTML response

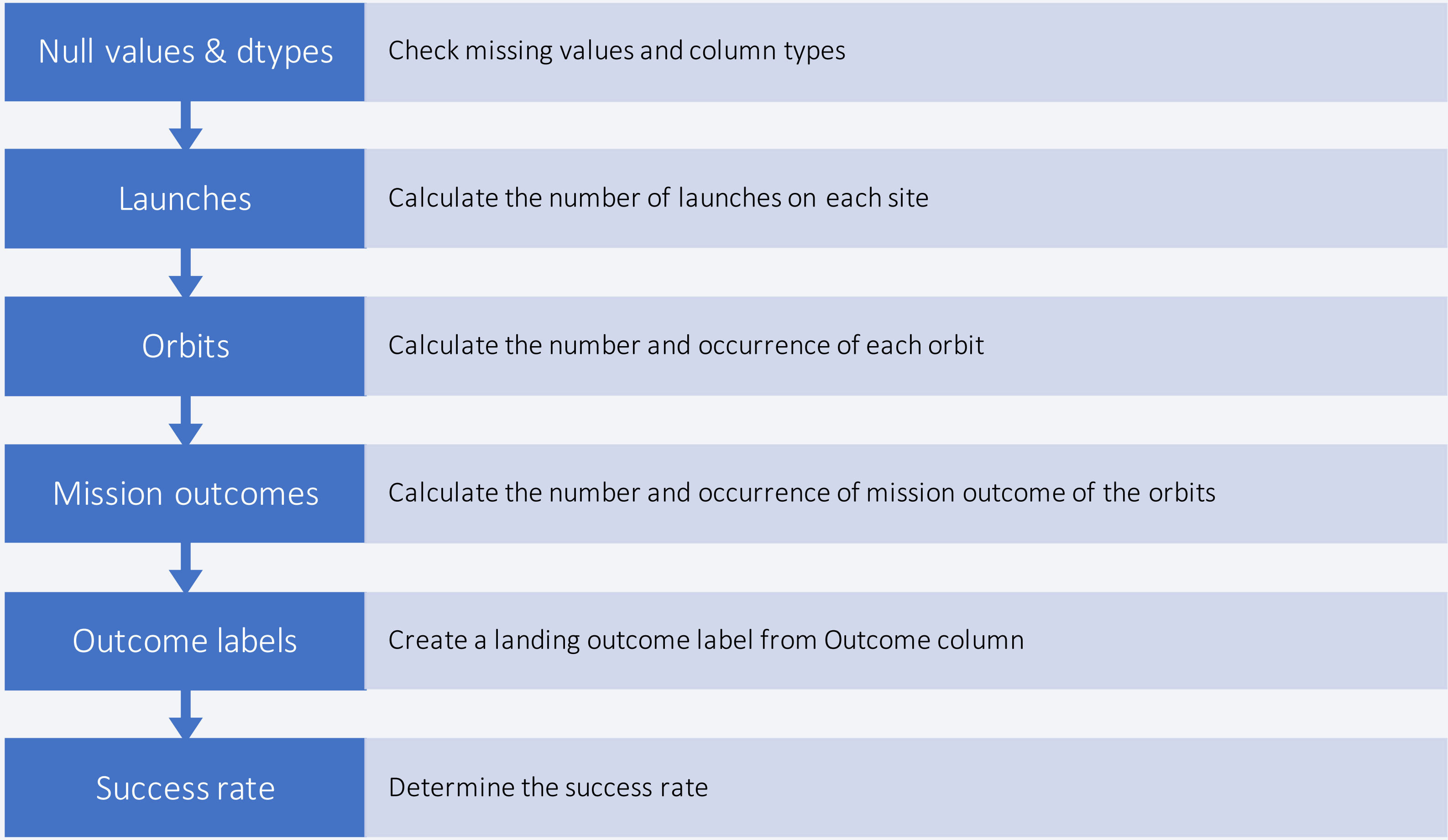
Extracting all column/variable names from the HTML table header

Creating a dataframe by parsing the launch HTML tables

Exporting dataframe to a CSV

https://github.com/OlyaSobolevskaya/Data-Science-Project-SpaceX/blob/main/capstone-webscraping.ipynb

Data Wrangling



Null values & dtypes — Check missing values and column types

Launches — Calculate the number of launches on each site

Orbits — Calculate the number and occurrence of each orbit

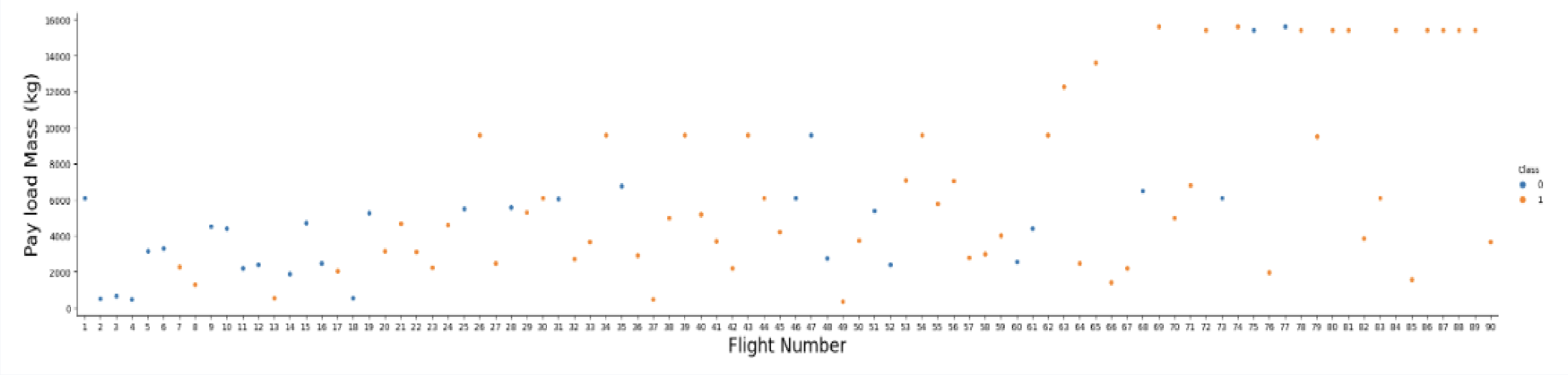
Mission outcomes — Calculate the number and occurrence of mission outcome of the orbits

Outcome labels — Create a landing outcome label from Outcome column

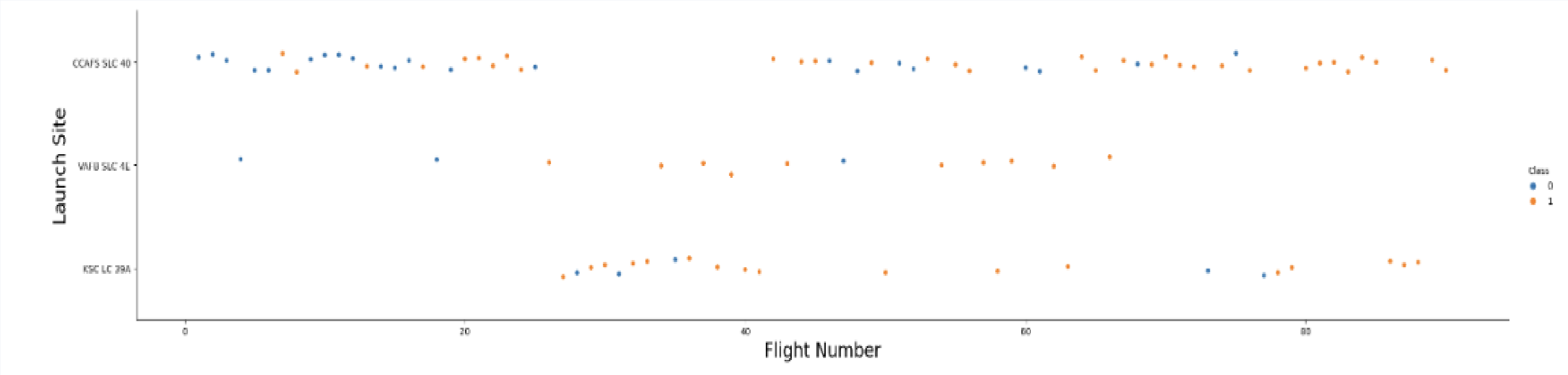
Success rate — Determine the success rate

https://github.com/OlyaSobolevskaya/Data-Science-Project-SpaceX/blob/main/capstone-spacex-Data%20wrangling.ipynb

EDA with Data Visualization



On the plot we can see how the FlightNumber and Payload variables would affect the launch outcome.

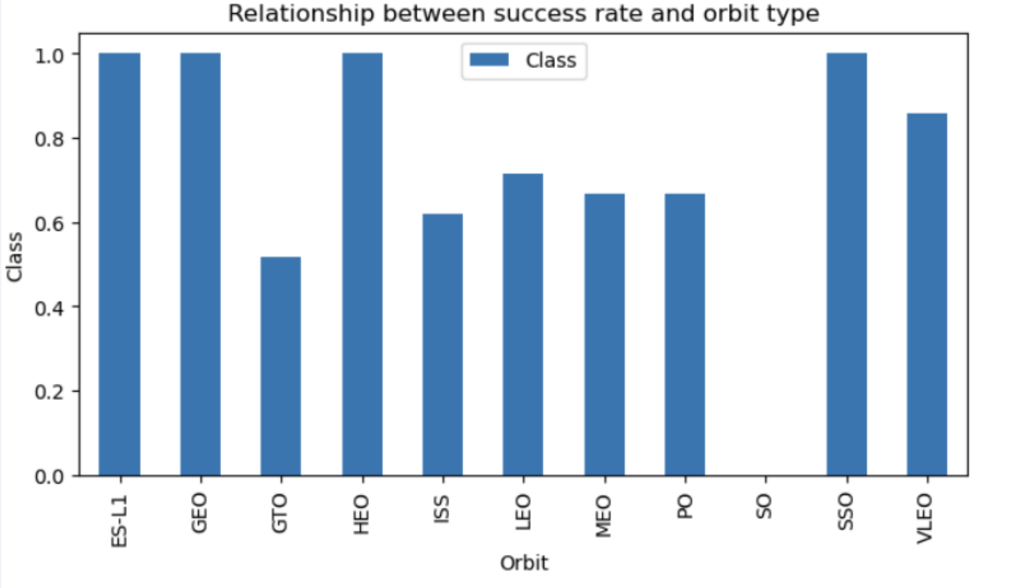


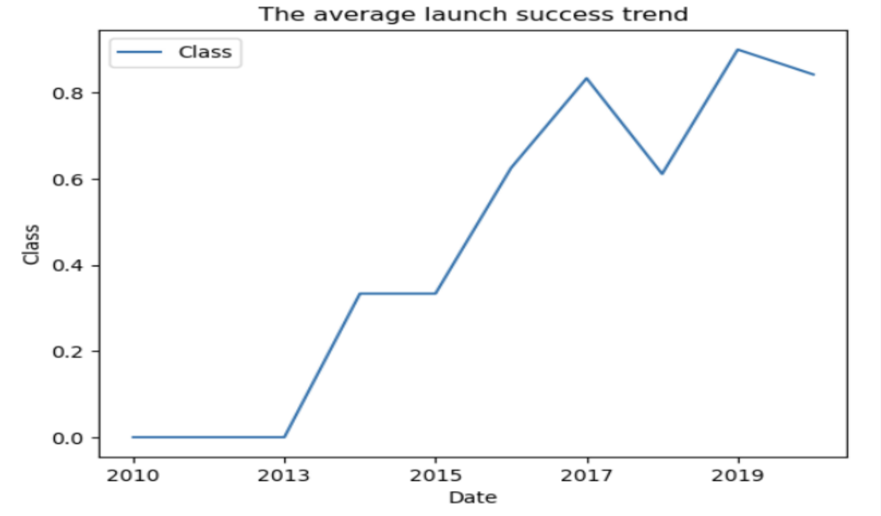
We plot out the FlightNumber vs. PayloadMass and overlay the outcome of the launch.

EDA with Data Visualization

ALso we visualized the relationship between:

- Flight Number and Launch Site;
- Payload and Launch Site;
- FlightNumber and Orbit type;
- Payload and Orbit type;
- and between success rate of each orbit type





Finally, we visualized the launch success yearly trend

https://github.com/OlyaSobolevskaya/Data-Science-Project-SpaceX/blob/main/capstone-eda-dataviz.ipynb

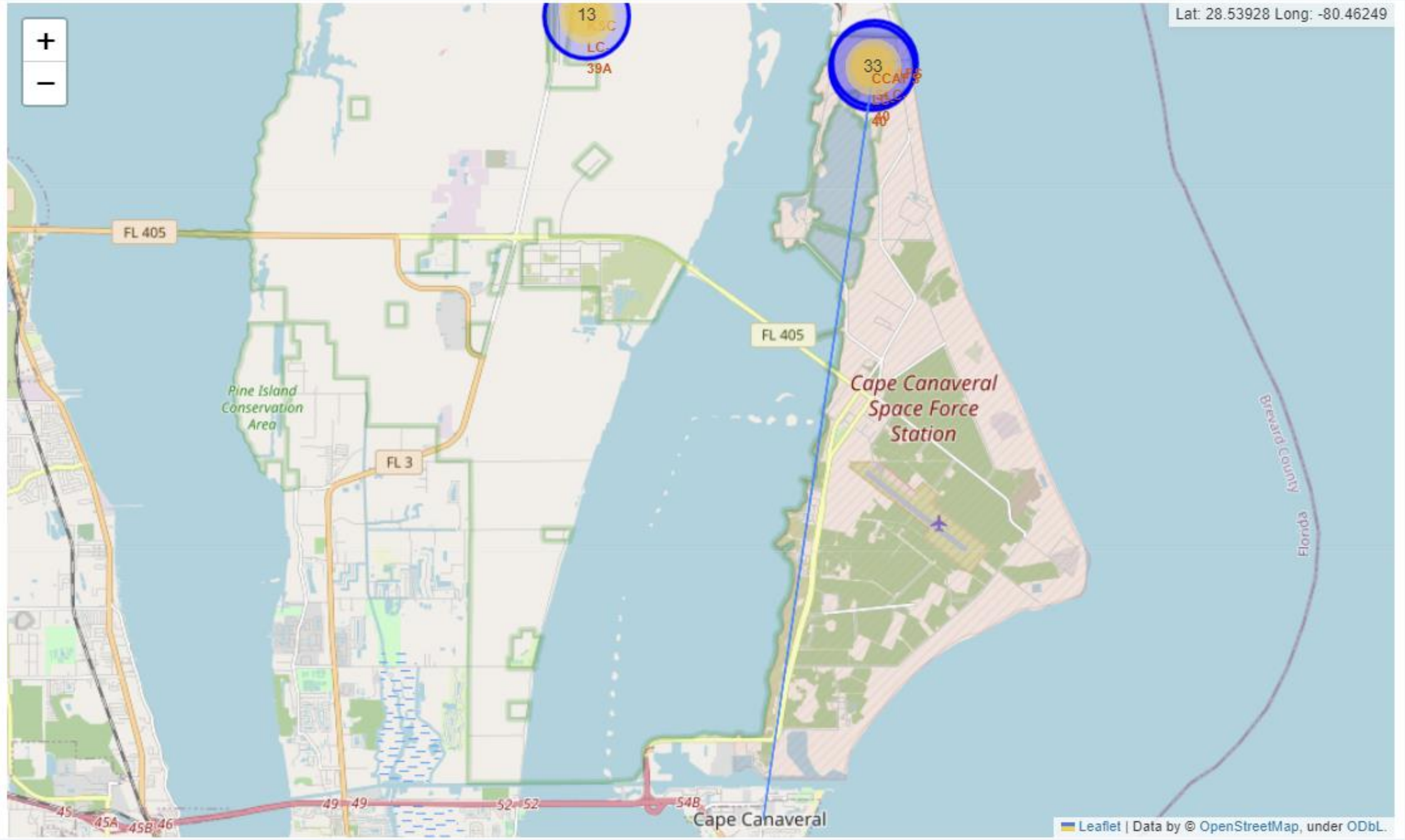
EDA with SQL

- Performed SQL queries:
  - the names of the unique launch sites in the space mission
  - 5 records where launch sites begin with the string 'CCA'
  - the total payload mass carried by boosters launched by NASA (CRS)
  - average payload mass carried by booster version F9 v1.1
  - the date when the first successful landing outcome in ground pad was achieved
  - the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - the total number of successful and failure mission outcomes
  - the names of the booster_versions which have carried the maximum payload mass
  - the records which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015
  - the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order

https://github.com/OlyaSobolevskaya/Data-Science-Project-SpaceX/blob/main/capstone-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

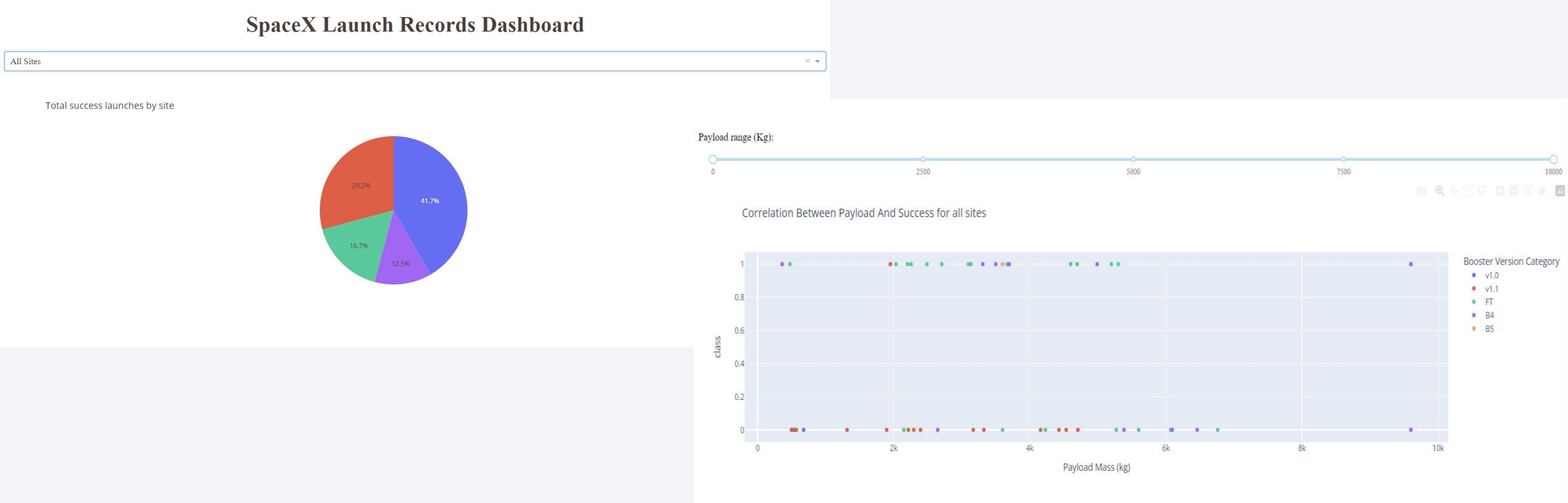
Map markers, circles, lines were added to a folium map to find an optimal location for building a launch site.


+
−
13
LC
39A
33
CCAFS
40
Lat: 28.53928 Long: -80.46249
FL 405
FL 405
Pine Island Conservation Area
FL 3
Cape Canaveral Space Force Station
Cape Canaveral
Brevard County
Florida
Leaflet | Data by © OpenStreetMap, under ODbL.

https://github.com/OlyaSobolevskaya/Data-Science-Project-SpaceX/blob/main/capstone_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

I added to a dashboard a pie chart visualizing launch success counts and a scatter plot to observe how payload may be correlated with mission outcomes for selected site(s).

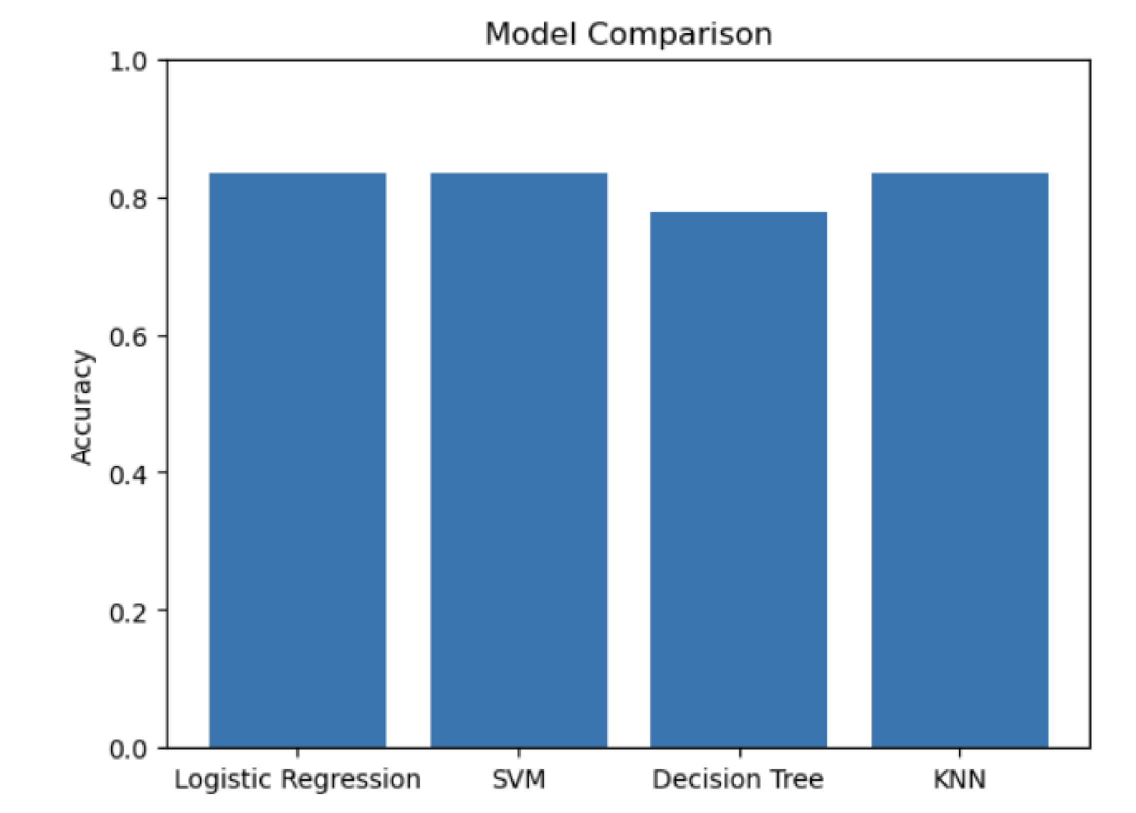


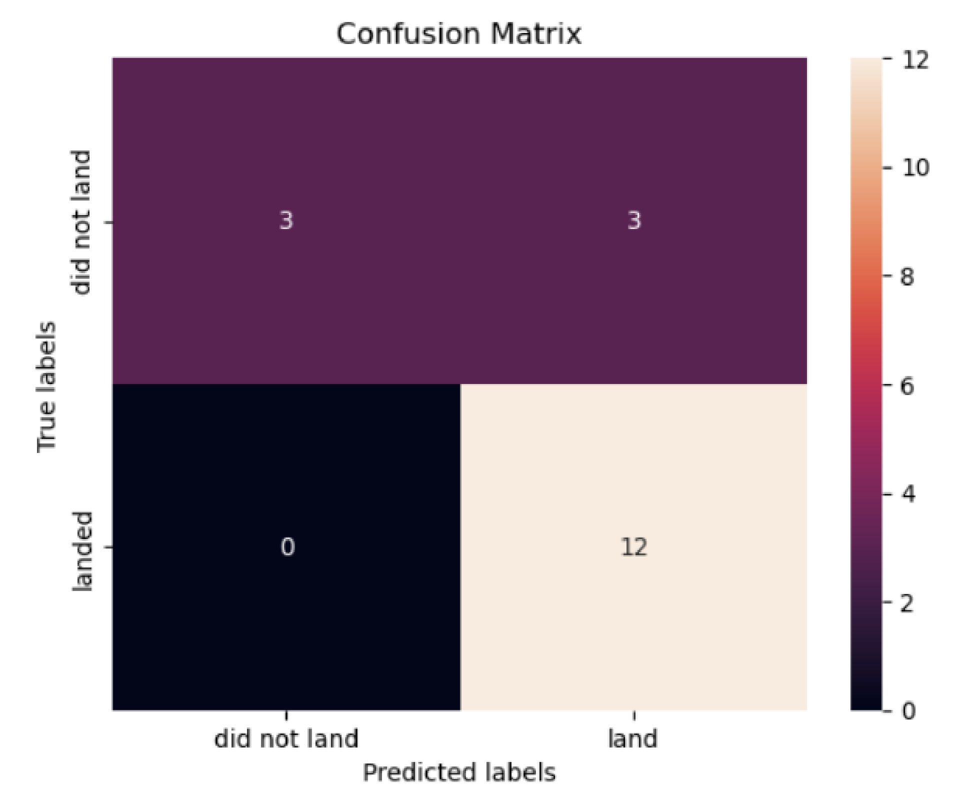
https://github.com/OlyaSobolevskaya/Data-Science-Project-SpaceX/blob/main/capstone-plotly-dash.ipynb

Predictive Analysis (Classification)

We create a machine learning pipeline to predict if the first stage will land.

Results



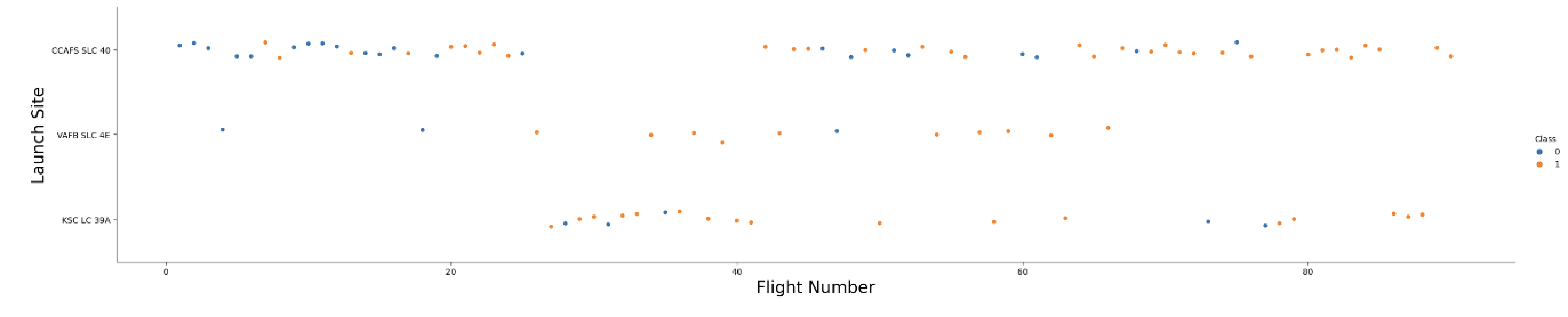


Decision Tree accuracy was slightly worse than other models which had high accuracy at 83.3%.

Section 2

Insights drawn from EDA

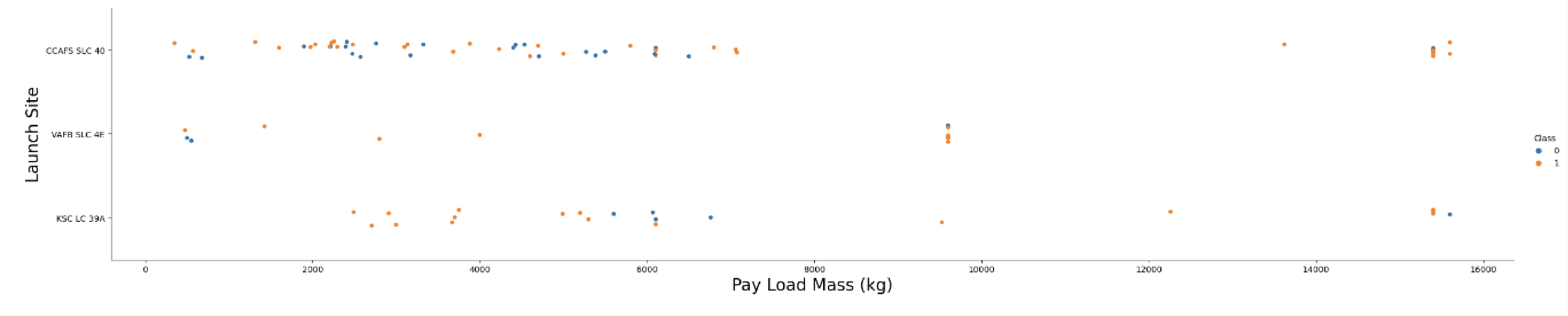
Flight Number vs. Launch Site



As the flight number increases, the first stage is more likely to land successfully.

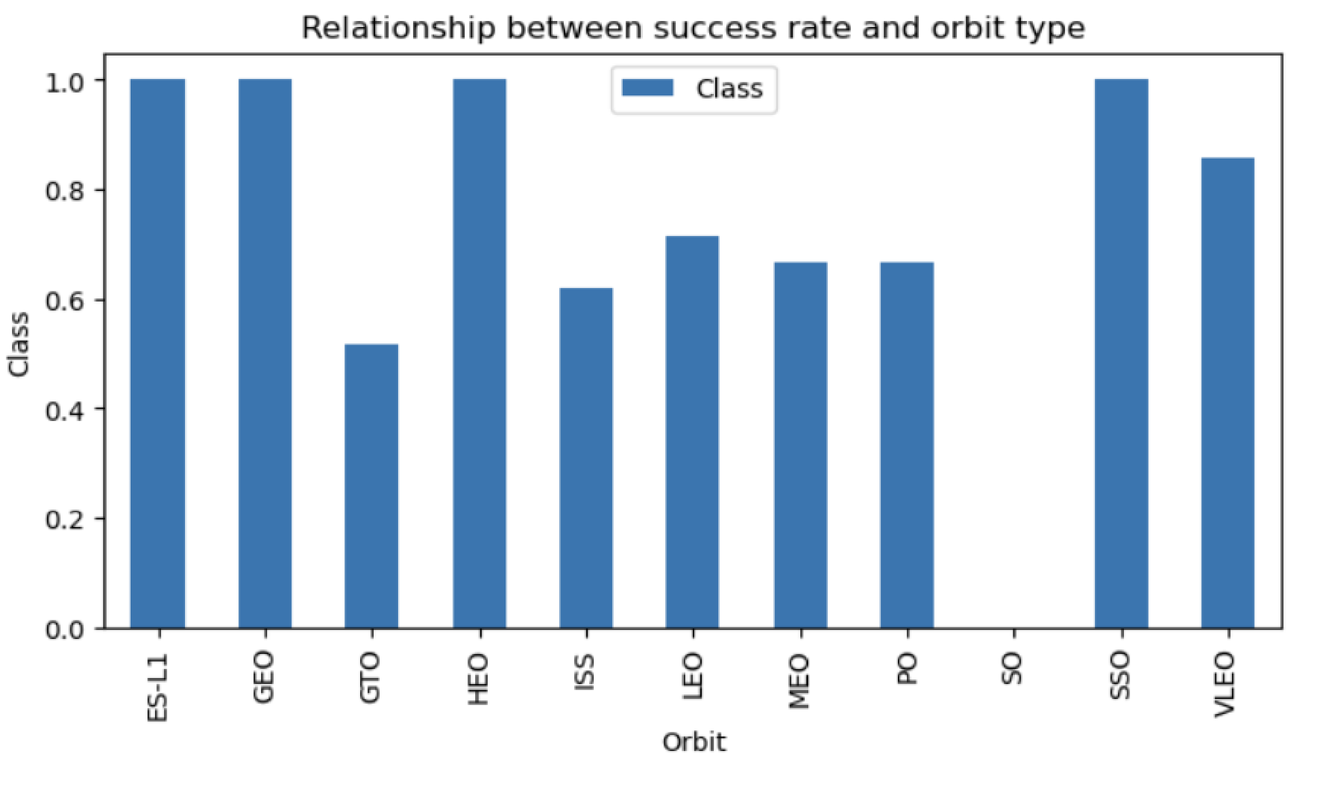
Also we see that different launch sites have different success rates. CCAFS LC-40 has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

Payload vs. Launch Site



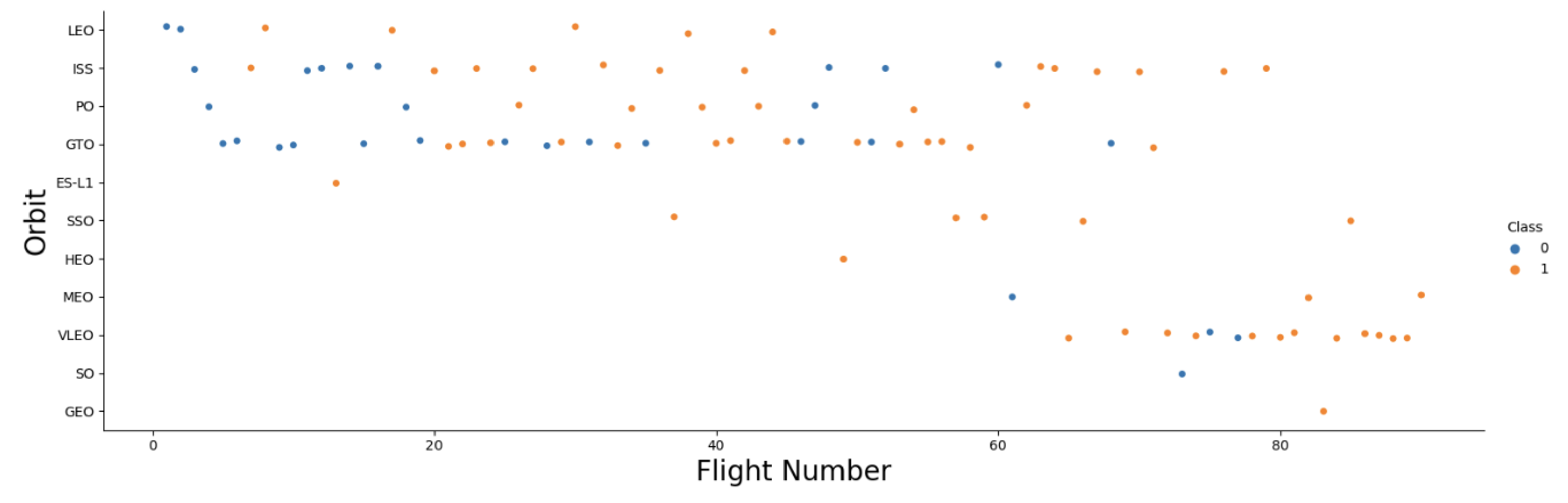
For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000). KSC LC-39A launch site doesn't show well results for payload mass between 5000 and 7000. The majority of rockets with lighter payload mass were launched from CCAFS LC-40 launch site.

Success Rate vs. Orbit Type



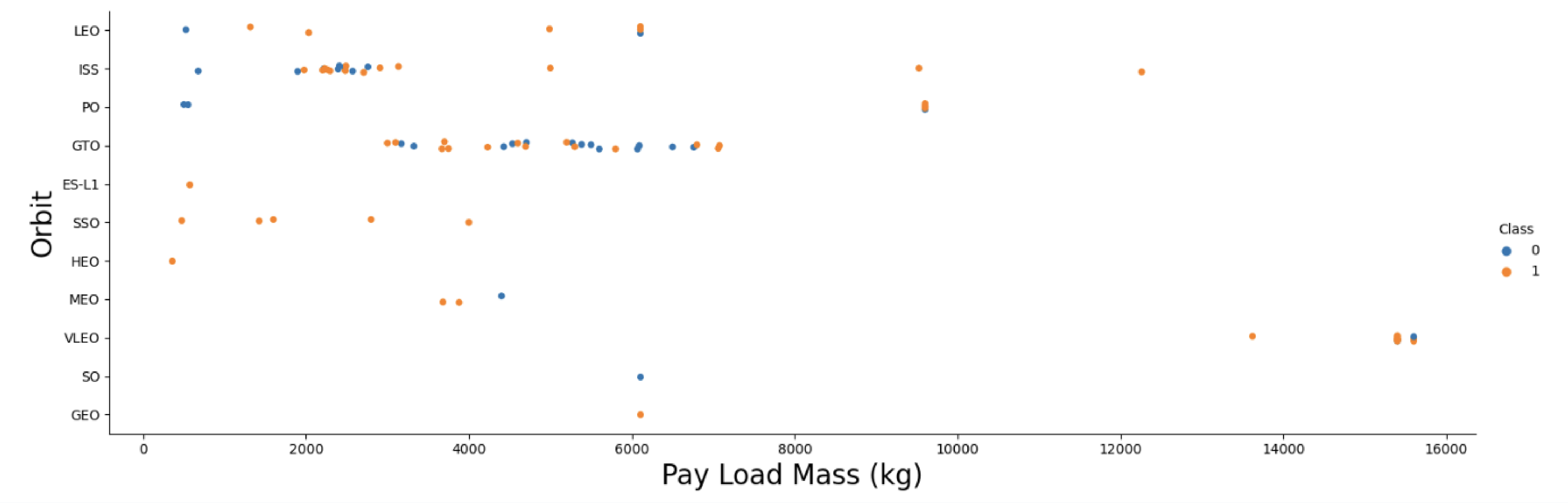
- The highest success rate have FS-L1, GEO, HEO and SSO orbits.

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. Also there's a trend recently to use VLEO orbit more frequently.

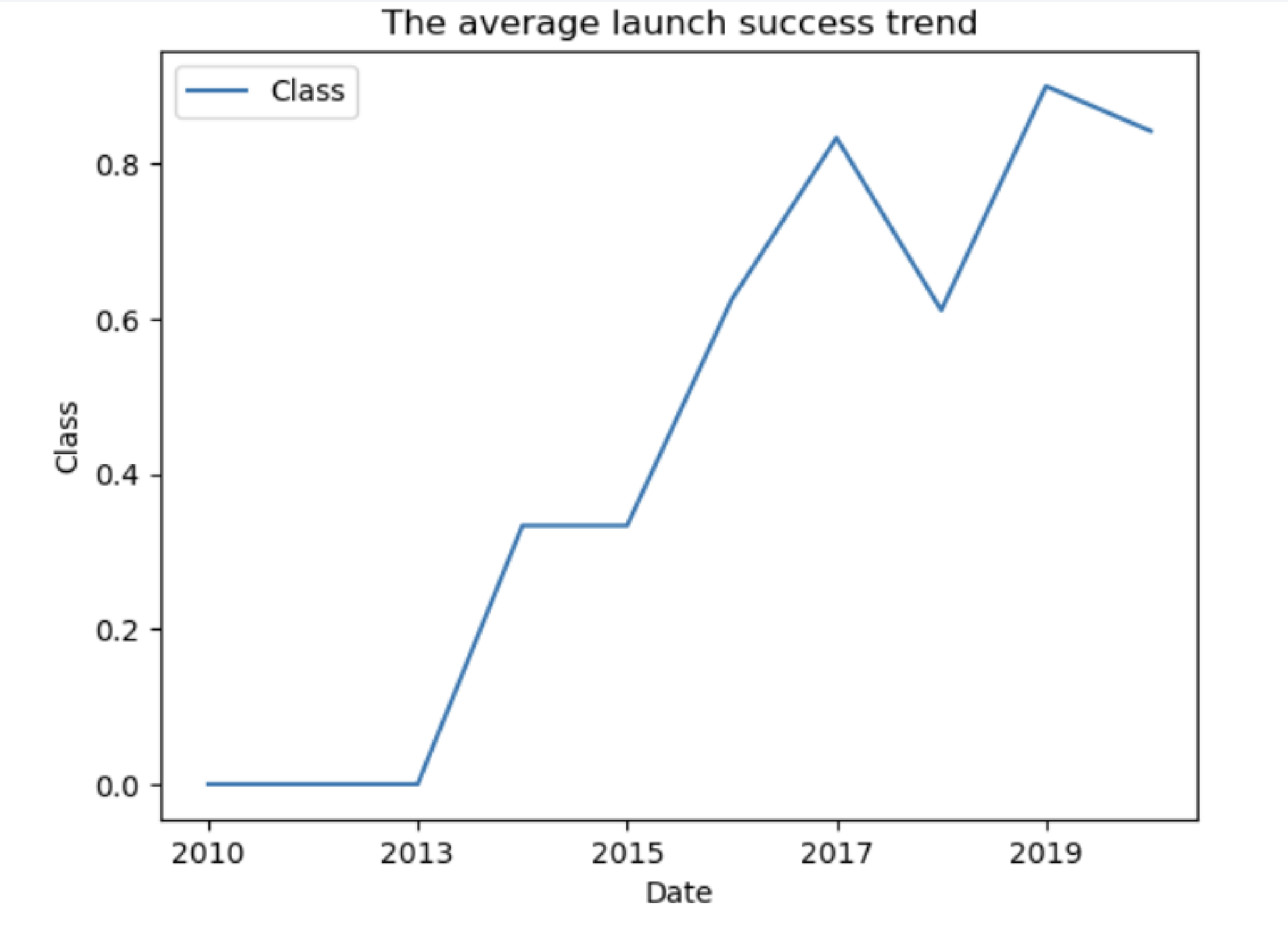
Payload vs. Orbit Type



With heavy payloads the successful landing rate is more for ISS and Po. The majority of light payloads (2000-3000 kg) is for ISS, while payloads between 3000 and 7000 kg were mostly launched on GTO orbit.

Launch Success Yearly Trend

The success rate has increased significantly since 2013.

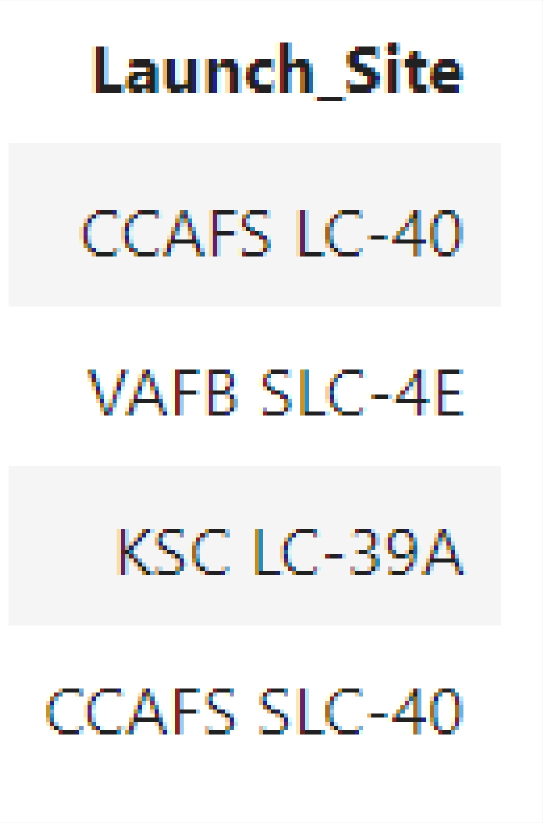


All Launch Site Names

select distinct Launch_Site from SPACEXTBL

There are 4 Launch Site:

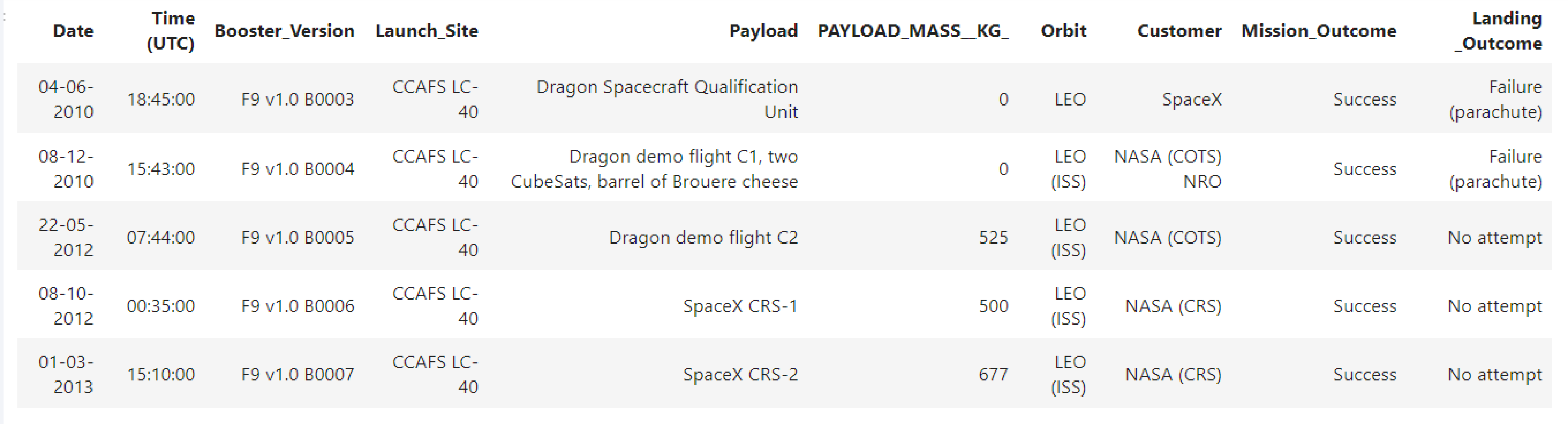
- CCAFS LC-40,
- VAFB SLC-4E,
- KSC LC-39A,
- CCAFS SLC-40



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

Launch Site Names Begin with 'CCA'

- select * from SPACEXTBL where Launch_site like 'CCA%' limit 5

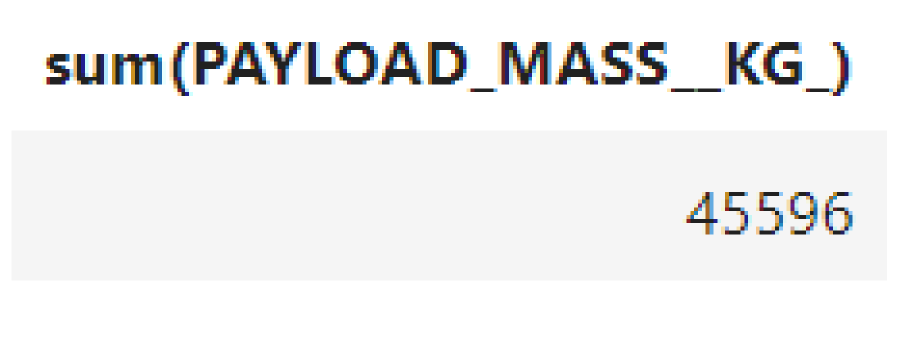


Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

The total payload carried by boosters from NASA is 45,596 kg.

- select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)'

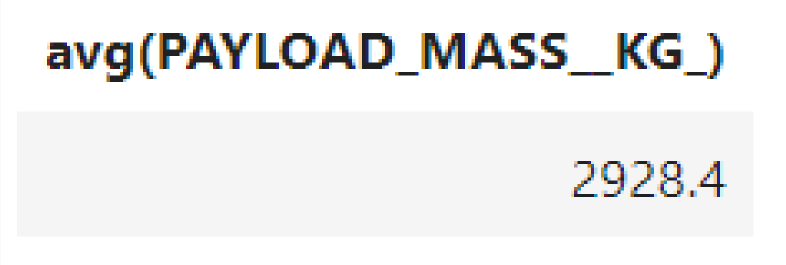


sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2,928.4 kg.

- select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1'

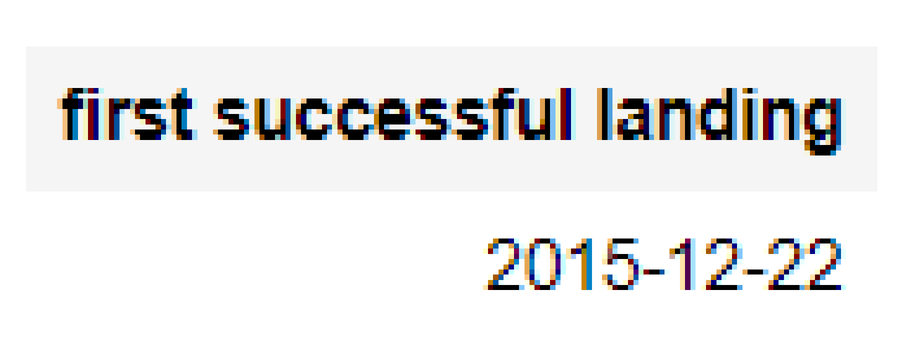


avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

The date of the first successful landing outcome on ground pad: 22/12/2015.

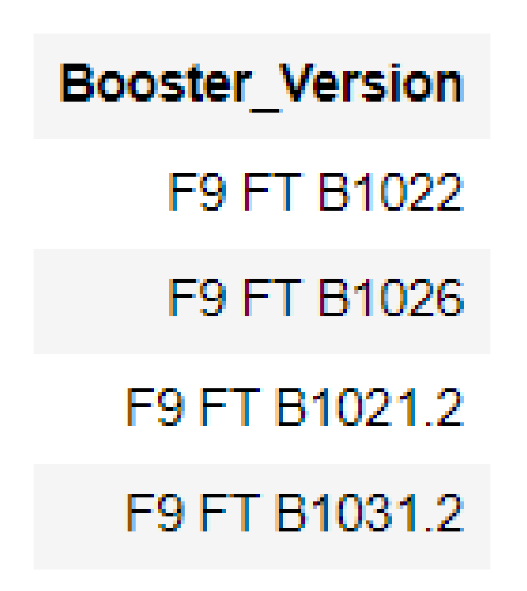
- select min(date(substr(Date, 7, 4) | | '-' | | substr(Date, 4, 2) | | '-' | | substr(Date, 1, 2))) as 'first successful landing' from SPACEXTBL where [Landing _Outcome] = 'Success (ground pad)'



first successful landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

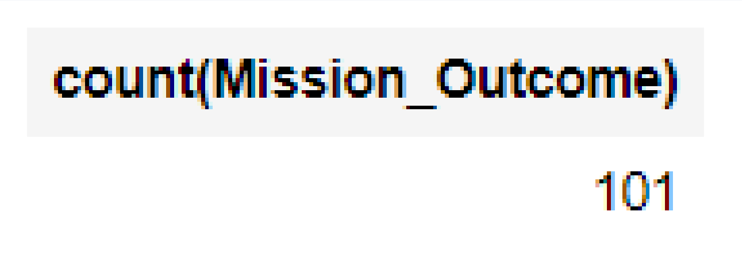
- select distinct Booster_Version from SPACEXTBL where [Landing _Outcome] = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000



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

Total Number of Successful and Failure Mission Outcomes

- select count(Mission_Outcome) from SPACEXTBL



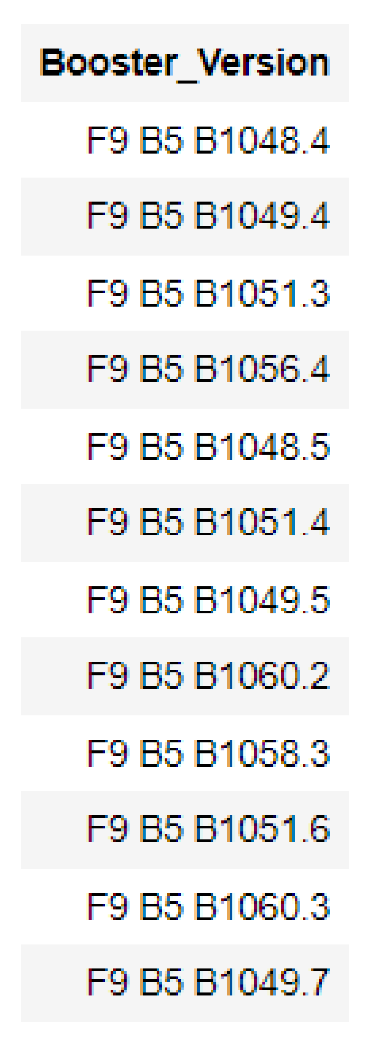
count(Mission_Outcome)
101

The result includes outcomes:

- 'Success',
- 'Failure (in flight)',
- 'Success (payload status unclear)',
- 'Success '.

Boosters Carried Maximum Payload

- select distinct Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)

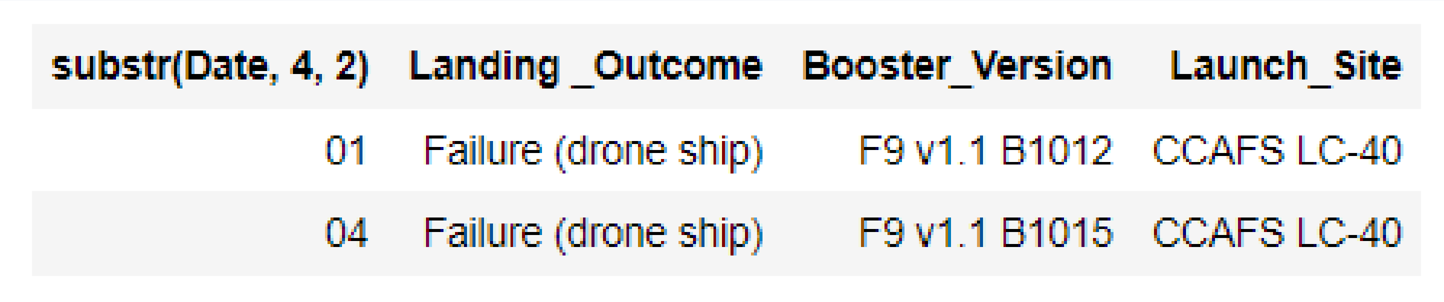


Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

The list of the failed landing outcomes in drone ship for the year 2015.

- select substr(Date, 4, 2), [Landing _Outcome], Booster_Version, Launch_Site from SPACEXTBL where [Landing _Outcome] = 'Failure (drone ship)' and substr(Date,7,4)='2015'

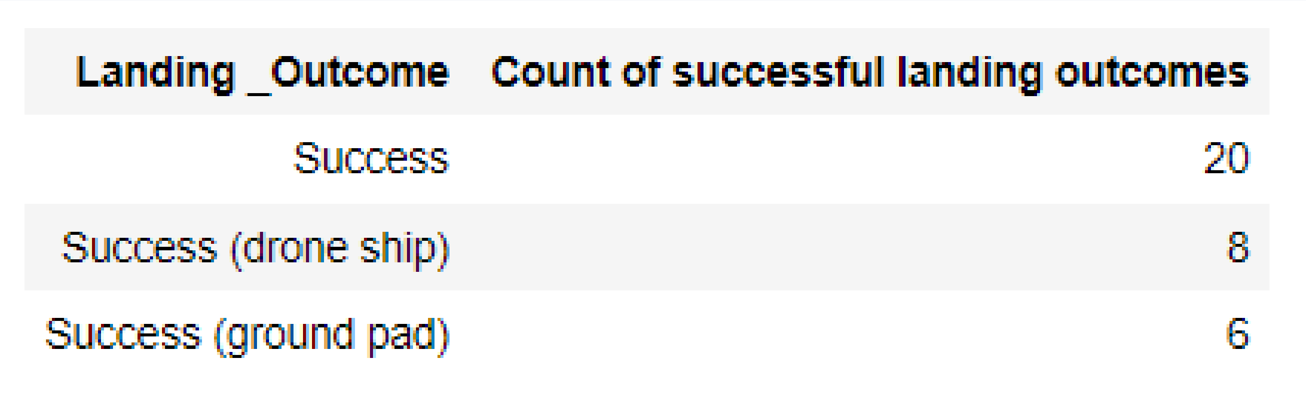


substr(Date, 4, 2)	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

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

- select [Landing _Outcome], count([Landing _Outcome])
  as 'Count of successful landing outcomes' from SPACEXTBL
  where [Landing _Outcome] like 'Success%' and Date between '04-06-2010' and '20-03-2017' group by [Landing _Outcome] order by count([Landing _Outcome]) desc

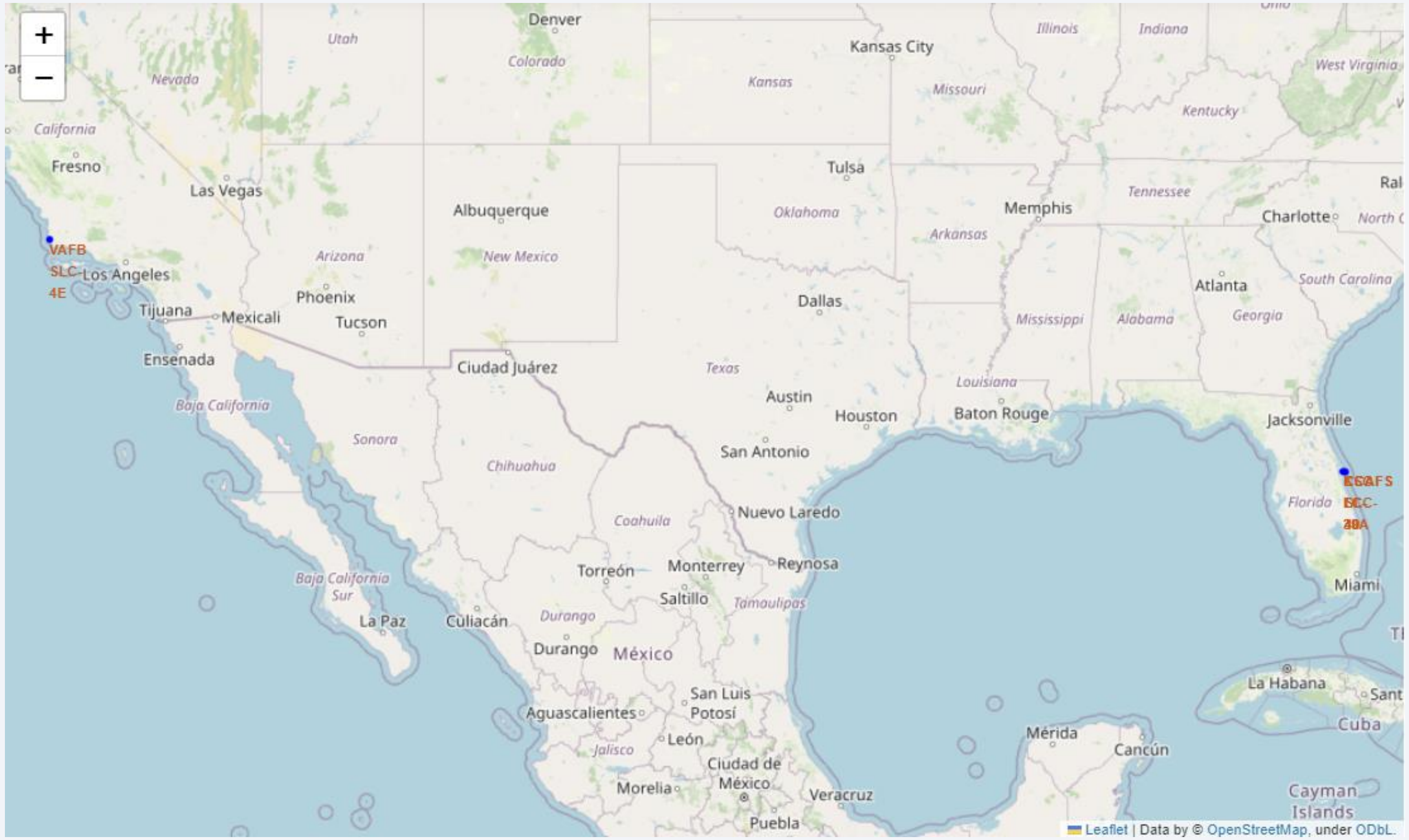


Landing _Outcome	Count of successful landing outcomes
Success	20
Success (drone ship)	8
Success (ground pad)	6

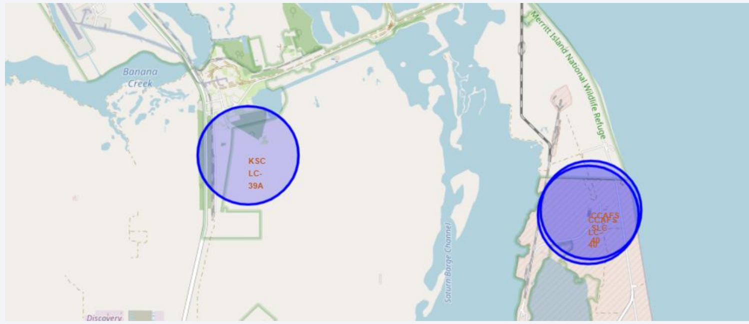
Section 3

Launch Sites Proximities Analysis

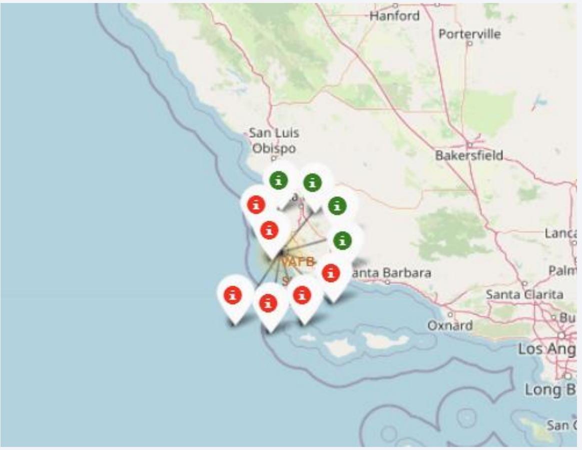
All launch sites on a map



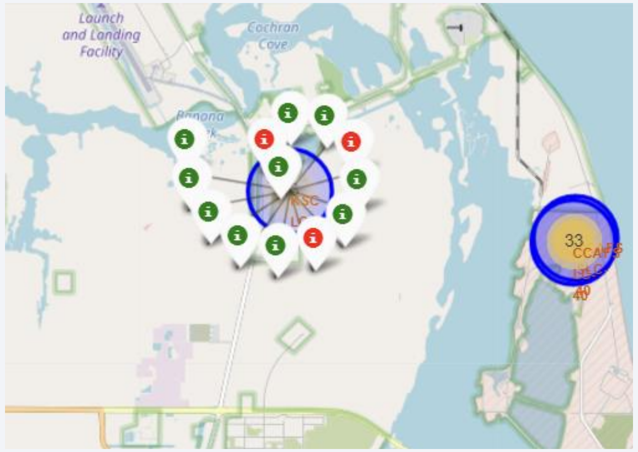
As we can see, CCAFS LC-40, CCAFS SLC-40 and KSC LC-39A are situated on the south-east coast, in Florida,
while VAFB SLC-4E is situated on the west coast, in Los-Angeles.



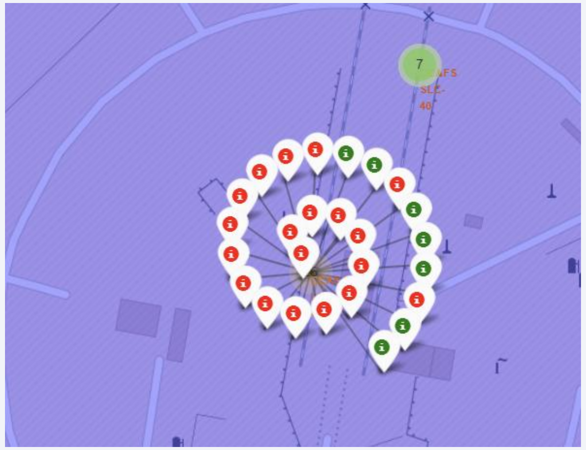
The success/failed launches for each site on the map



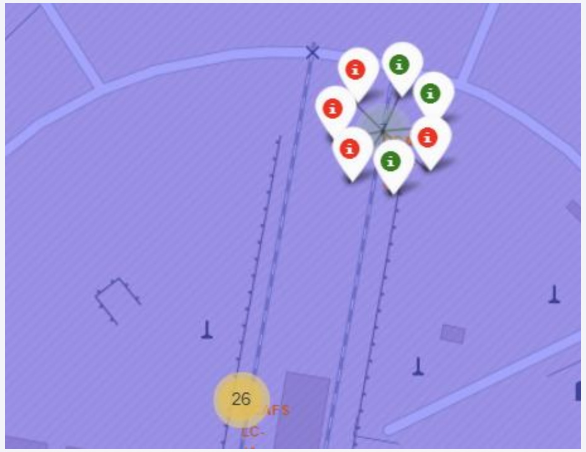
VAFB SLC-4E



KSC LC-39A



CCAFS LC-40

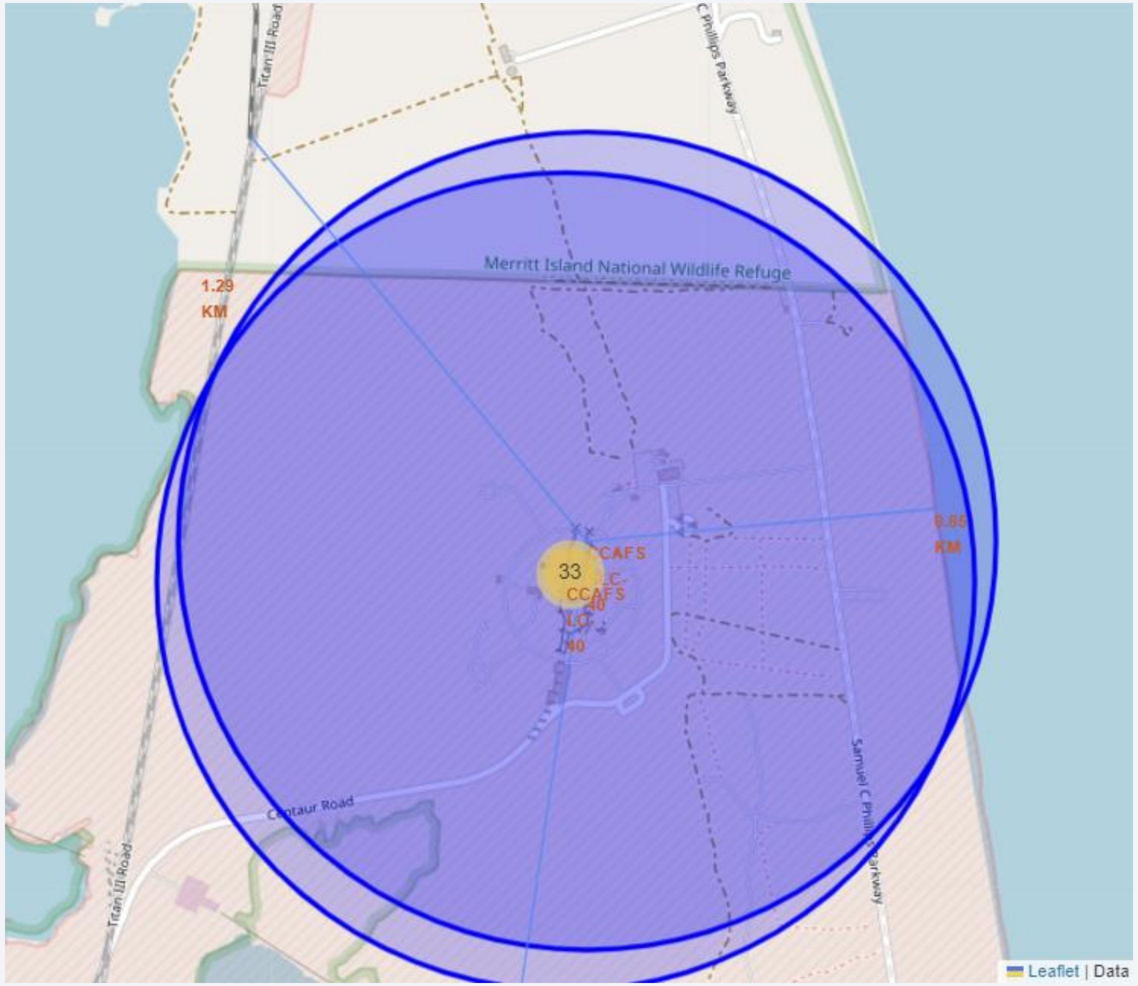


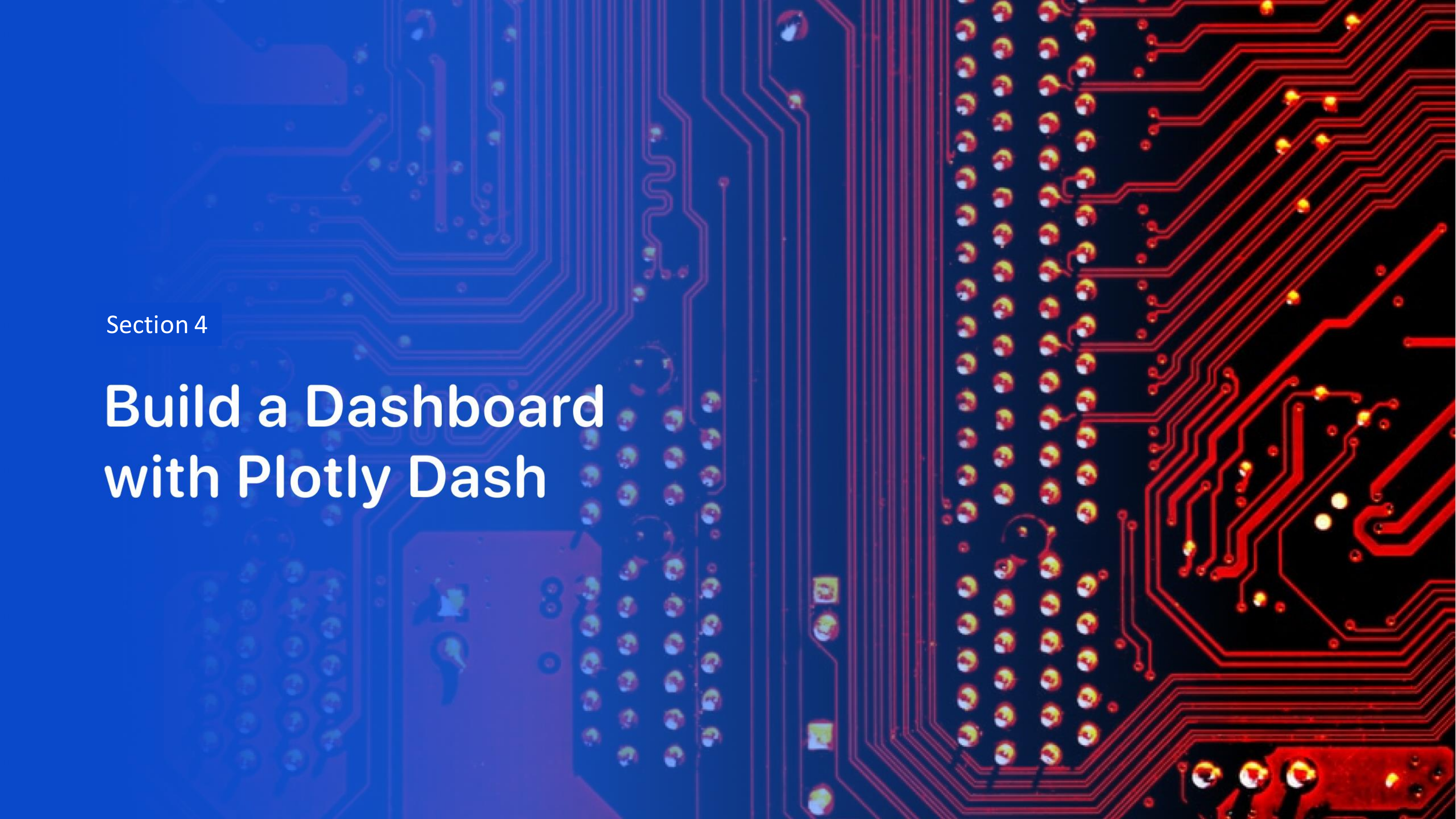
CCAFS SLC-40

As we can see the best successful launches rate has KSC LC-39A launch site, while the worst rate has CCAFS SLC-40.

The distances between a launch site to its proximities

- We can calculate and analyze distances between any launch site and other objects as railway, highway, coastline, city, etc.

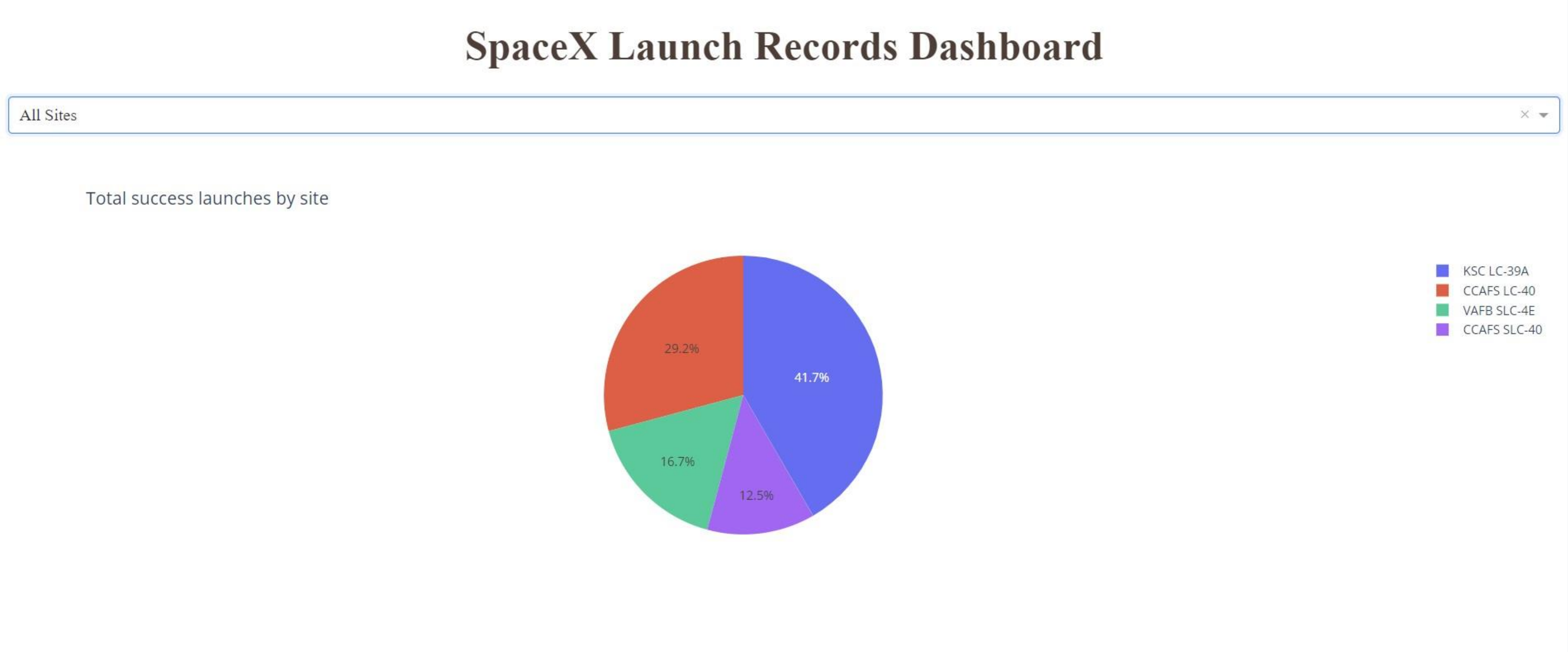




Section 4

Build a Dashboard with Plotly Dash

Total launch success count for all sites



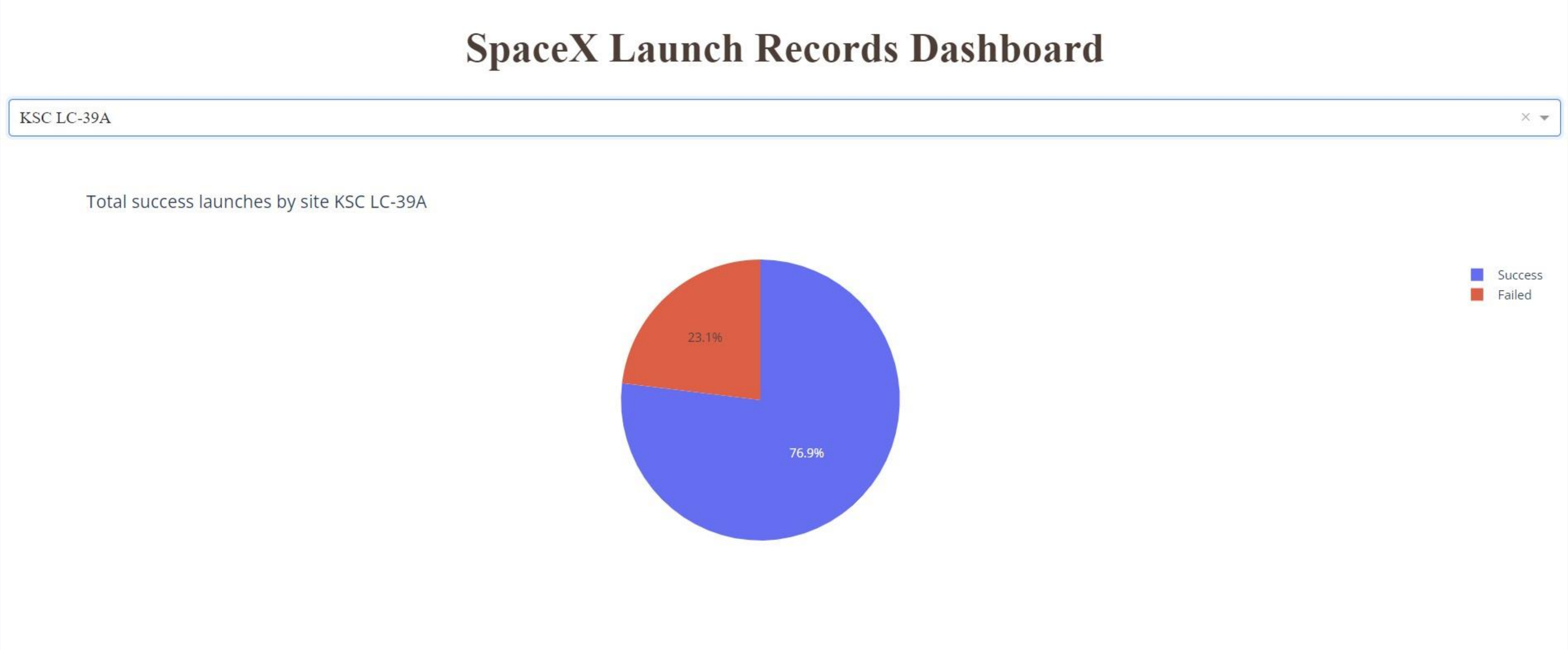
SpaceX Launch Records Dashboard

All Sites

Total success launches by site

KSC LC-39A has the best success launch rate.

The launch site with highest launch success ratio



SpaceX Launch Records Dashboard

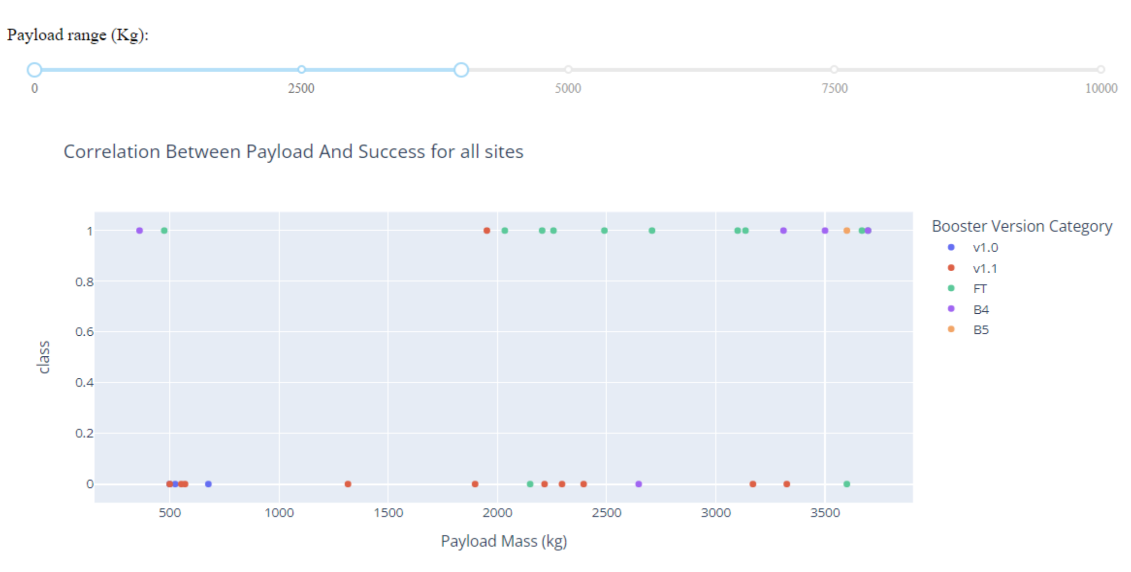
KSC LC-39A

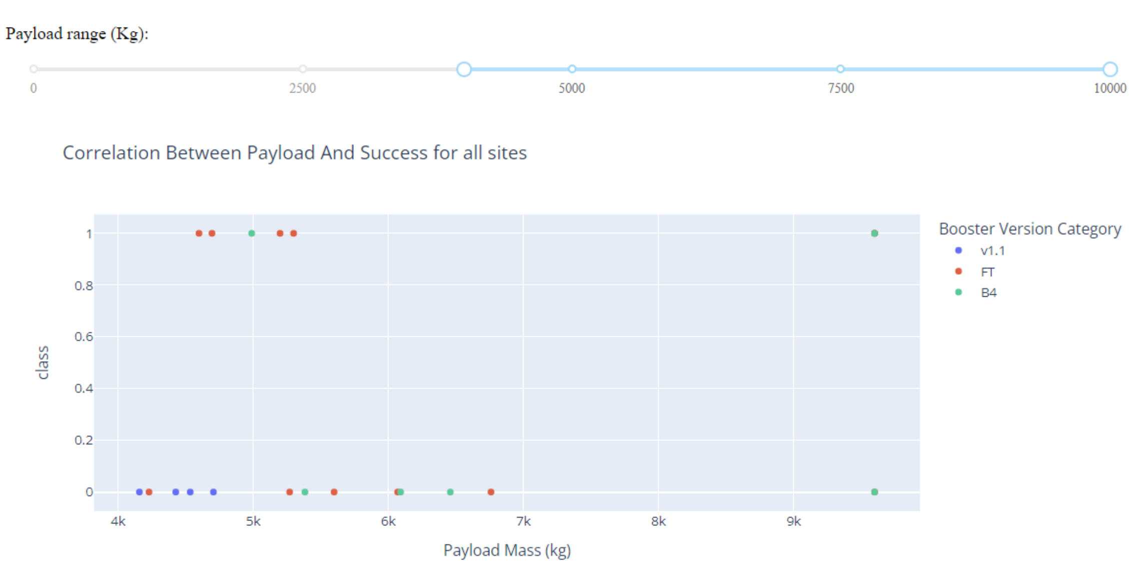
Total success launches by site KSC LC-39A

KSC LC-39A has the highest launch success ratio of 76,9%

Correlation Between Payload And Success for all sites

We can see booster versions FT and B4 has better results with lower payload (up to 4000kg), and FT is better with heavier payload (from 4000 kg). Also only three booster versions were used with heavy payload and success rate for it was lower.



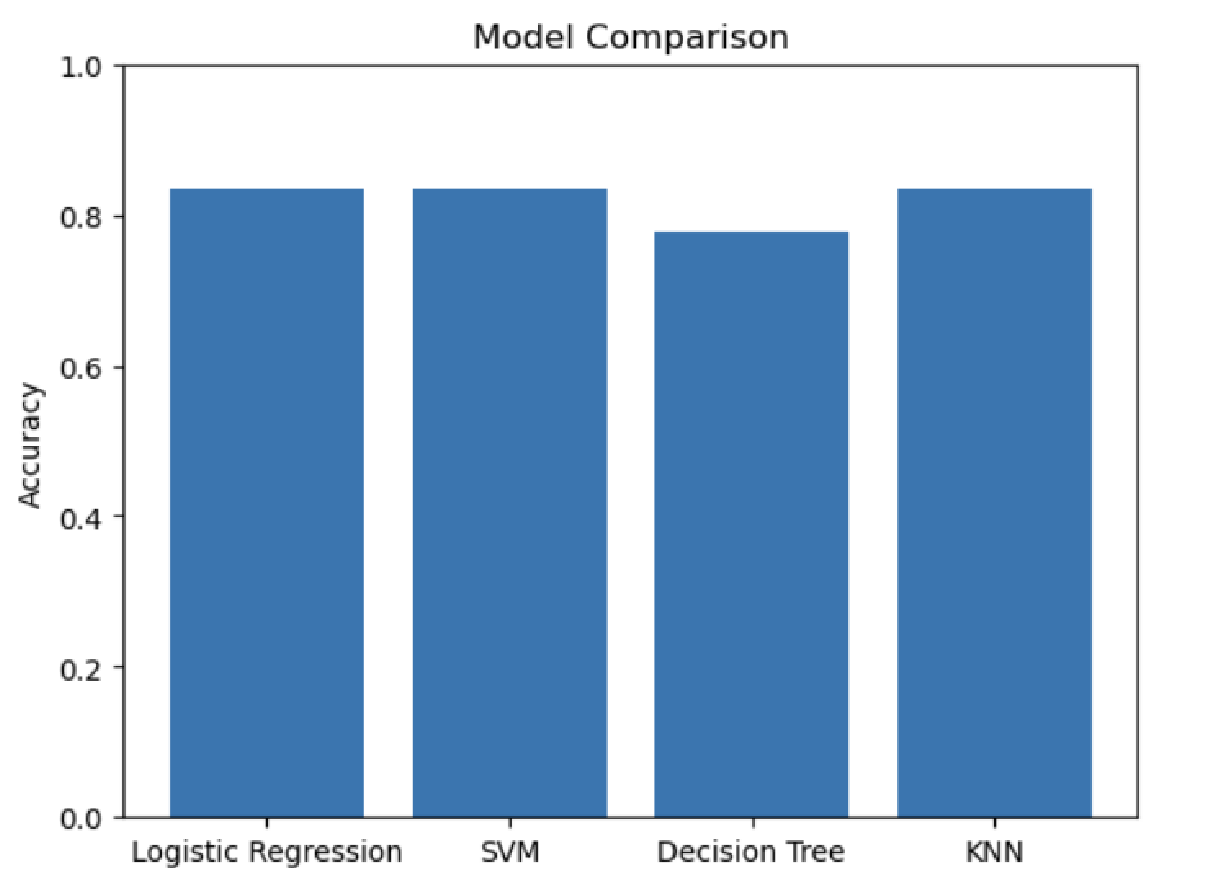


Section 5

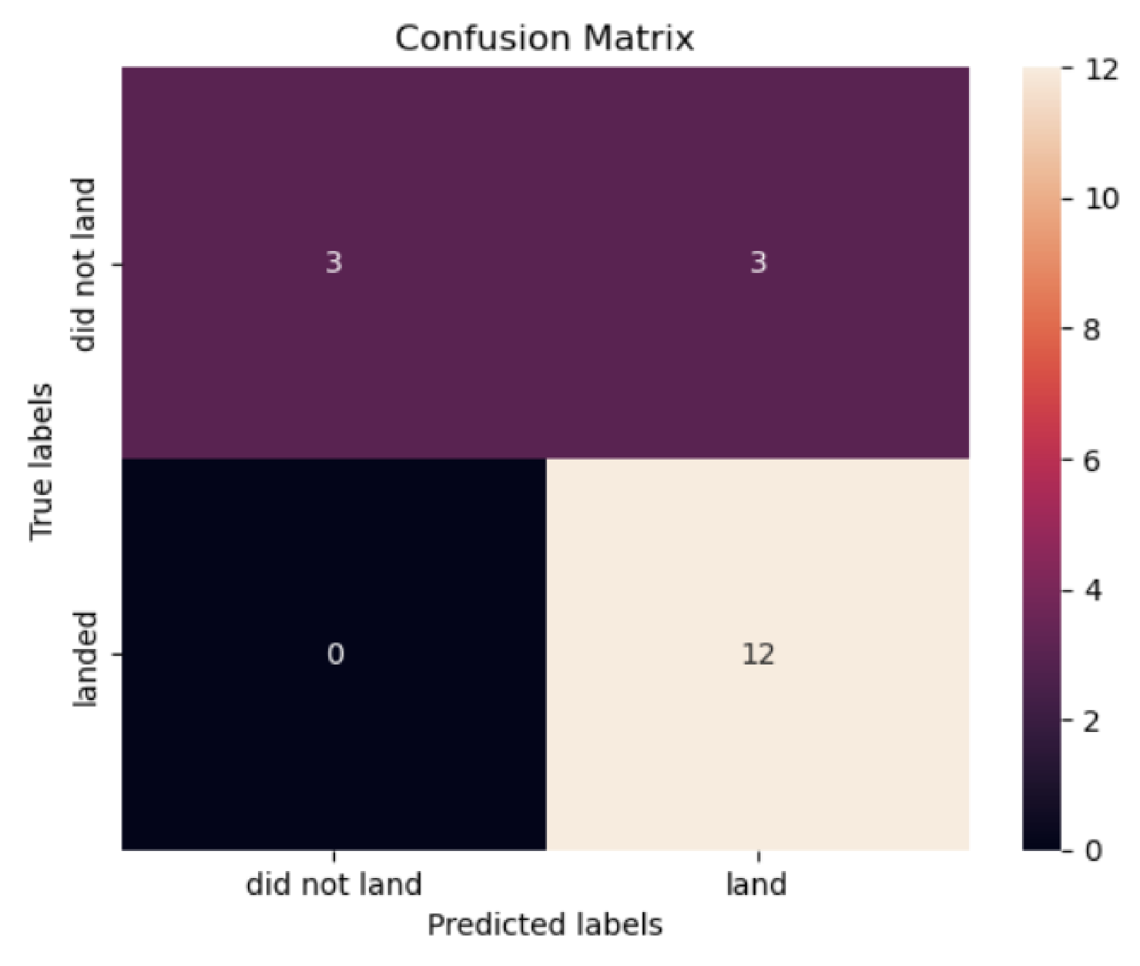
Predictive Analysis (Classification)

Classification Accuracy

Three models – Logistic Regression, SVM and KNN – have the highest classification accuracy of 83.3%



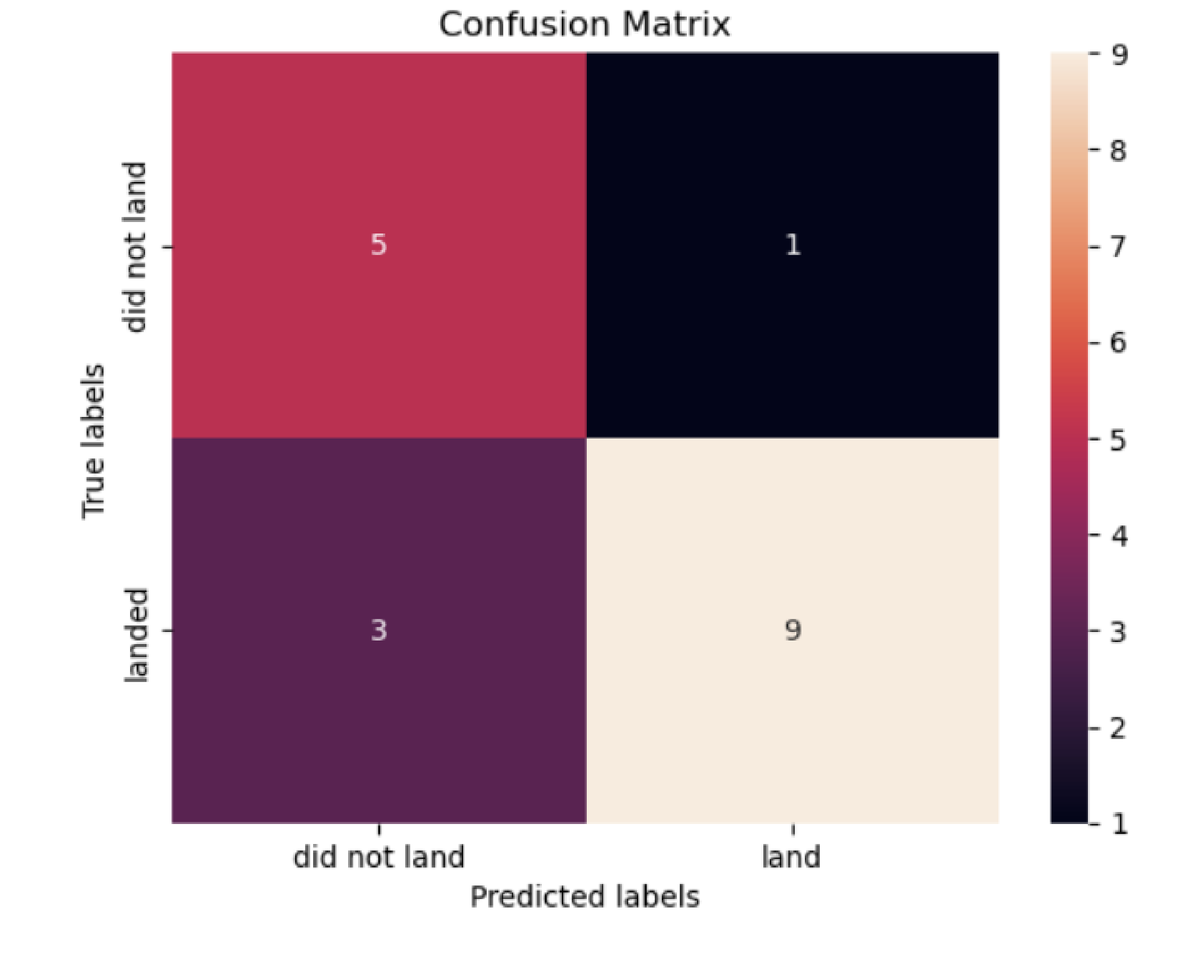
Confusion Matrix



All three best models – Logistic Regression, SVM and KNN – have the same confusion Matrix.

As we can see, the models can predict rather well if the first stage will land.

Confusion Matrix



On the other hand, Decision Tree model is better in predicting if the first stage will not land.

Conclusions

- The success rate has been increasing **since 2013**.
- The highest success rate have **FS-L1, GEO, HEO and SSO orbits**. With heavy payloads the successful landing rate is better for **ISS**.
- **KSC LC-39A launch site** has the highest launch success ratio of 76,9%. But it doesn't show well results for payload mass between 5000 and 7000.
- The **booster version FT** has better results both with lower and heavier payload.
- The success rate of **heavy payload launches** (>4000 kg) is lower.
- **Logistic Regression, SVM and KNN** Models can help to predict if the first stage will load with the accuracy of 83,3%.

Thank you!