IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Felipe Suzuki
09/18/2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Methodologies
  - Data Collection via API
  - Web Scraping
  - Exploratory Data Analysis (EDA) with Data Visualization
  - EDA with SQL
  - Interactive Map with Folium
  - Dashboard with Plotly Dash
  - Predictive Analysis
- Results
  - EDA results
  - Interactive maps and dashboard
  - Predictive results

Introduction

- Project background and context
  - The objective of this project is to predict if the Falcon 9 first stage will successfully land. SpaceX says on its website that the Falcon 9 rocket launch cost 62 million dollars. Other providers cost upward of 165 million dollars each. The price difference can be explained by the fact that SpaceX can reuse the first stage. By determining if the first stage will land, we can determine the cost of a launch. This information is interesting for another company if it wants to compete with SpaceX for a rocket launch.
- Problems you want to find answers
  - What are the main characteristics of a successful or an unsuccessful landing?
  - What are the effects of each relationship of the rocket variables on the success or failure of a landing?
  - What are the conditions which will allow SpaceX to achieve the best landing success rate?

Section 1

Methodology

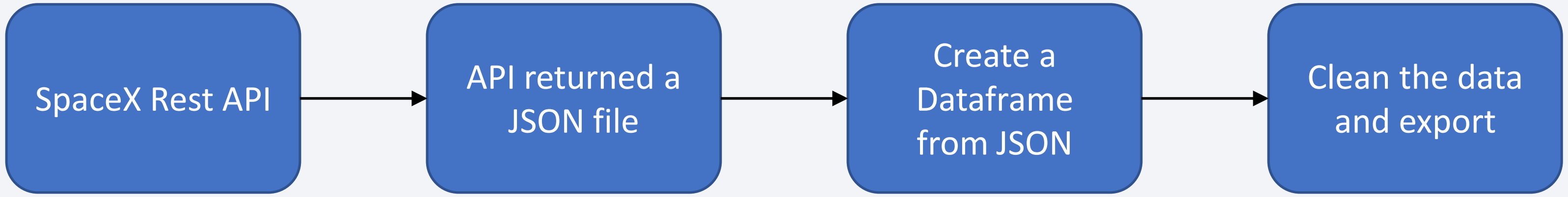
Methodology

Executive Summary

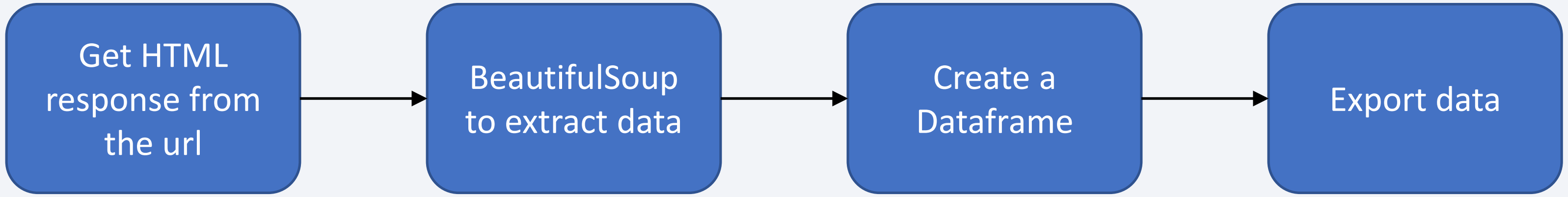
- Data collection methodology:
  - Using SpaceX REST API, and Web Scraping from Wikipedia
- Perform data wrangling
  - Using One Hot Encoding for Classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Use different models and evaluate accuracy on train and test data.

Data Collection

- Datasets were collected from REST SpaceX API and Web Scraping Wikipedia.
  - REST SpaceX API



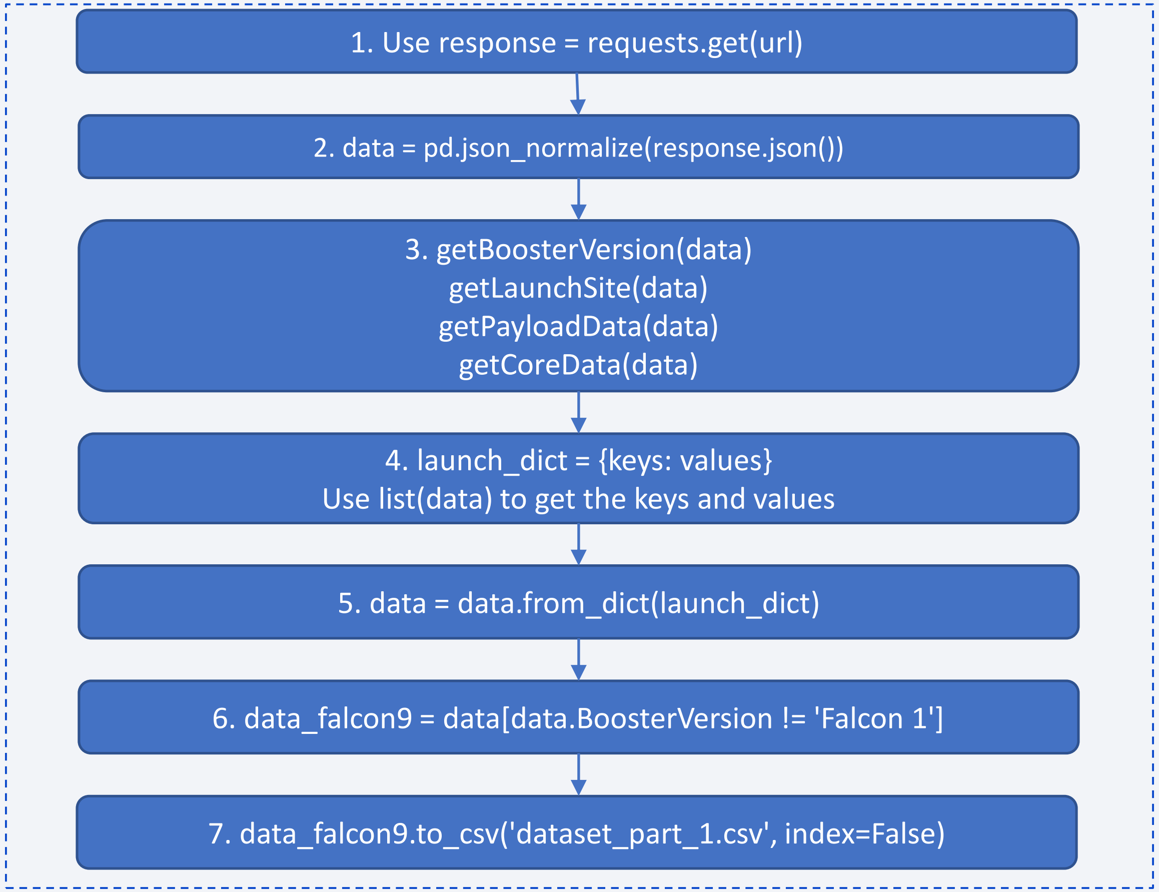
  - Web Scraping Wikipedia:



Data Collection – SpaceX API

- 1. Get Response from API
- 2. Convert the Response to a JSON
- 3. Transform the data
- 4. Create a dictionary with the data
- 5. Create a pandas Dataframe
- 6. Filter the df
- 7. Export to a .csv file

Redirect to the notebook



1. Use response = requests.get(url)

2. data = pd.json_normalize(response.json())

3. getBoosterVersion(data)
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)

4. launch_dict = {keys: values}
Use list(data) to get the keys and values

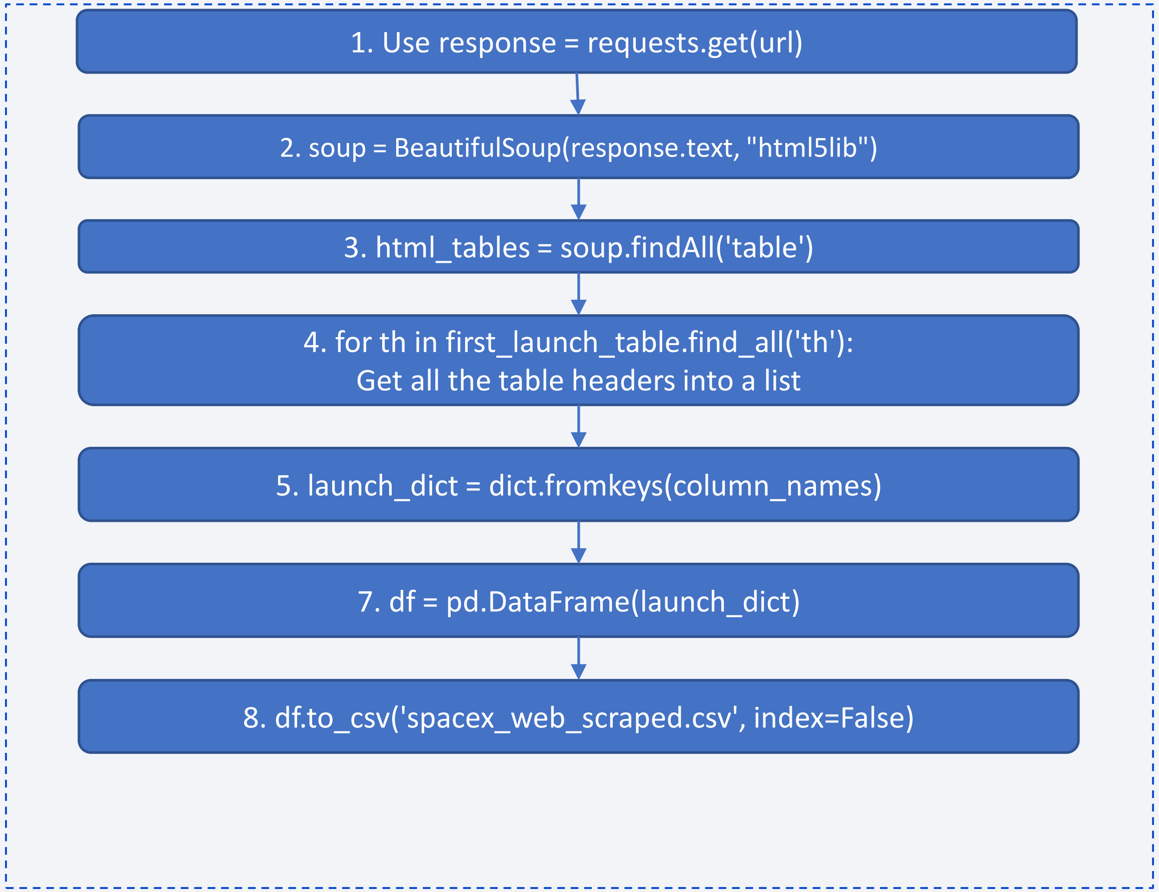
5. data = data.from_dict(launch_dict)

6. data_falcon9 = data[data.BoosterVersion != 'Falcon 1']

7. data_falcon9.to_csv('dataset_part_1.csv', index=False)

Data Collection - WebScraping

- 1. Get HTML Response
- 2. Create BeautifulSoup object
- 3. Find all tables
- 4. Get the column names
- 5. Create a dictionary
- 6. Add data to keys
- 7. Create a pandas df from dictionary
- 8. Export to .csv file



1. Use response = requests.get(url)

2. soup = BeautifulSoup(response.text, "html5lib")

3. html_tables = soup.findAll('table')

4. for th in first_launch_table.find_all('th'):
Get all the table headers into a list

5. launch_dict = dict.fromkeys(column_names)

7. df = pd.DataFrame(launch_dict)

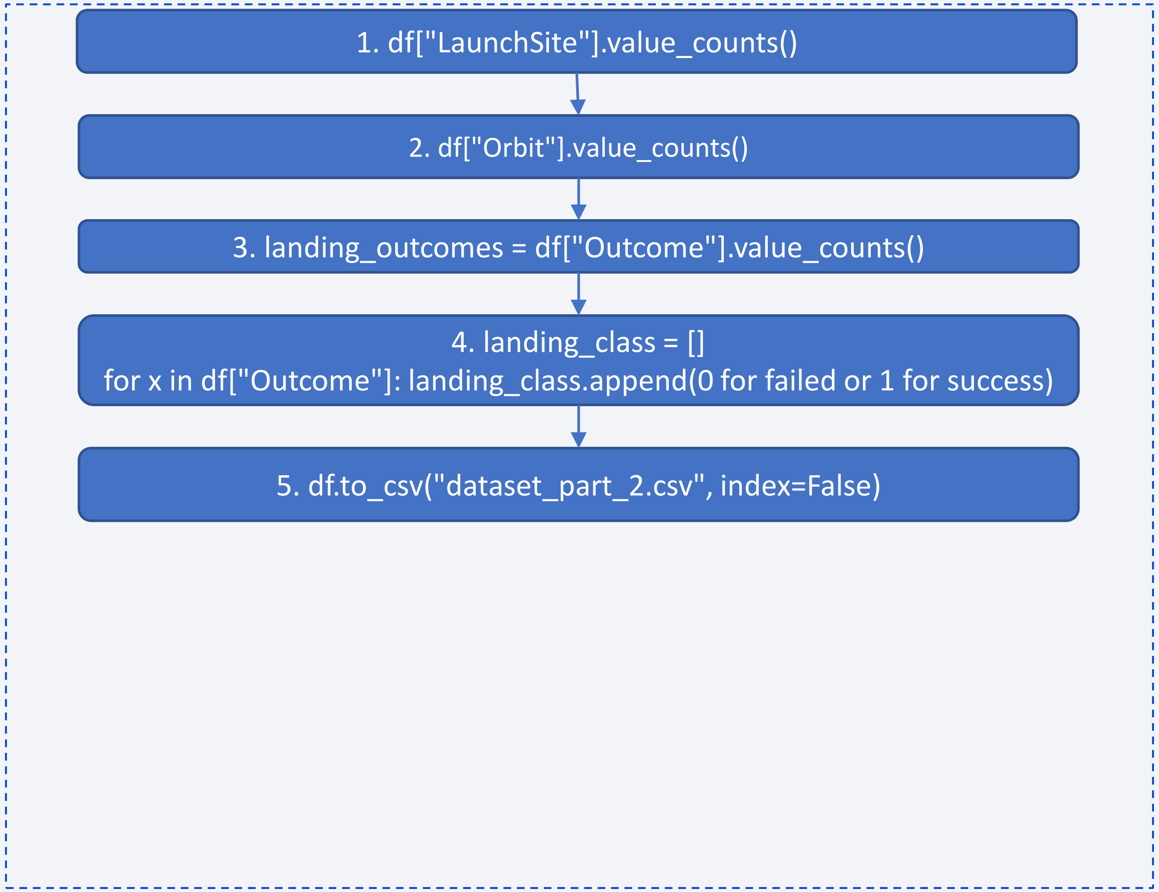
8. df.to_csv('spacex_web_scraped.csv', index=False)

[Redirect to the notebook]()

Data Wrangling

- 1. Calculate the launches number for each site
- 2. Calculate the number and occurrence of each orbit
- 3. Calculate the number and occurrence of mission outcome per orbit type
- 4. Create landing outcome label from Outcome column
- 5. Export to .csv file

Redirect to the notebook



```
1. df["LaunchSite"].value_counts()
2. df["Orbit"].value_counts()
3. landing_outcomes = df["Outcome"].value_counts()
4. landing_class = []
for x in df["Outcome"]: landing_class.append(0 for failed or 1 for success)
5. df.to_csv("dataset_part_2.csv", index=False)
```

EDA with Data Visualization

- Scatter Graphs – shows the correlation between 2 variables
  - Flight Number versus Payload Mass & Flight Number versus Lauch Site
  - Orbit versus Flight Number & Orbit versus Payload Mass
  - Payload versus Lauch Site & Payload versus Orbit Type
- Bar Graph – shows the relationship between numeric and categoric variables
  - Success Rate versus Orbit
- Line Graph – shows data variables and their trends
  - Success Rate versus Year

Redirect to the notebook

EDA with SQL

- SQL queries to get information from the dataset
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA‘
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved.
  - 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 names of the booster versions which have carried the maximum payload mass.
  - List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

Redirect to the notebook

Build an Interactive Map with Folium

- Folium map object is a map centered on NASA Johnson Space Center at Houson, Texas
  - Red circle at NASA Johnson Space Center's coordinate with label showing its name (folium.Circle, folium.map.Marker)
  - Red circles at each launch site coordinates with label showing launch site name (folium.Circle, folium.map.Marker, folium.features.DivIcon)
  - Grouping close points into a cluster (folium.plugins.MarkerCluster)
  - Green marker for successful landing and red for unsuccessful landing (folium.map.Marker, folium.Icon)
  - Markers to show distance between launch site and railway, highway, coastline and city with a plotline between them (folium.map.Marker, folium.PolyLine, folium.features.DivIcon)
- The objects were created to facilitate the visualization of the launch sites, key locations and the successful and unsuccessful landings.

Redirect to the notebook

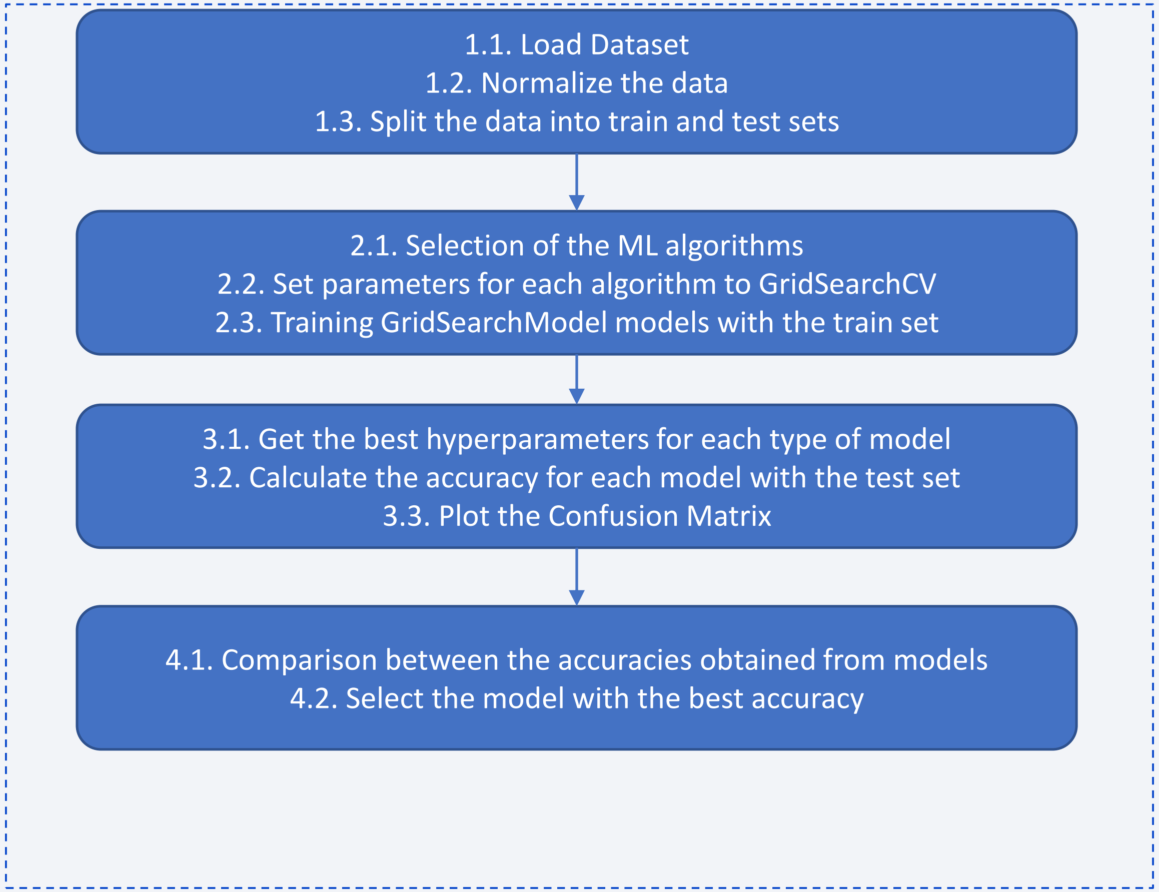
Build a Dashboard with Plotly Dash

- Dashboard has dropdown, pie chart, rangeslider and scatter plot components
  - Dropdown allows a user to choose the launch site or all launch sites (dash_core_components.Dropdown)
  - Pie chart shows the total success and the total failure for the launch site chosen with the dropdown component (plotly.express.pie)
  - Rangeslider allows a user to select a payload mass in a fixed range (dash_core_components.RangeSlider)
  - Scatter chart shows the relationship between two variables, in particular Success vs Payload Mass (plotly.express.scatter)

Redirect to the notebook

Predictive Analysis (Classification)

- 1. Data Preparation
- 2. Model Preparation
- 3. Model Evaluation
- 4. Model Comparison



1.1. Load Dataset
1.2. Normalize the data
1.3. Split the data into train and test sets

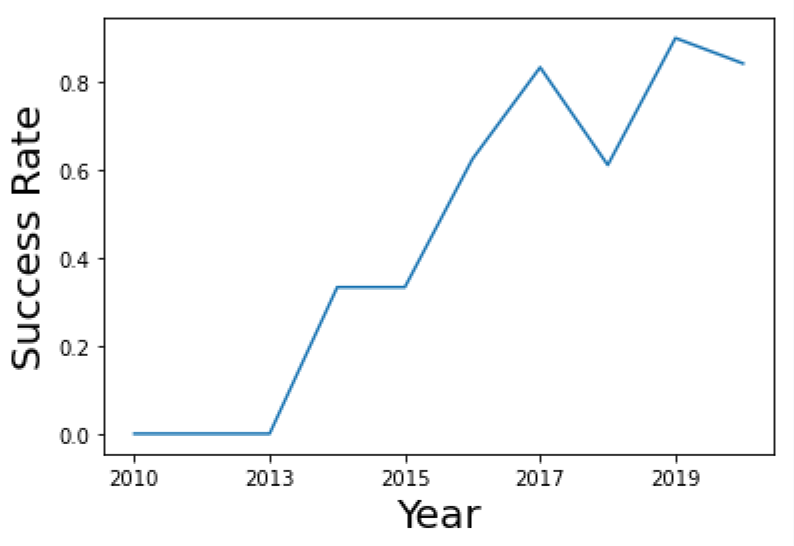
2.1. Selection of the ML algorithms
2.2. Set parameters for each algorithm to GridSearchCV
2.3. Training GridSearchModel models with the train set

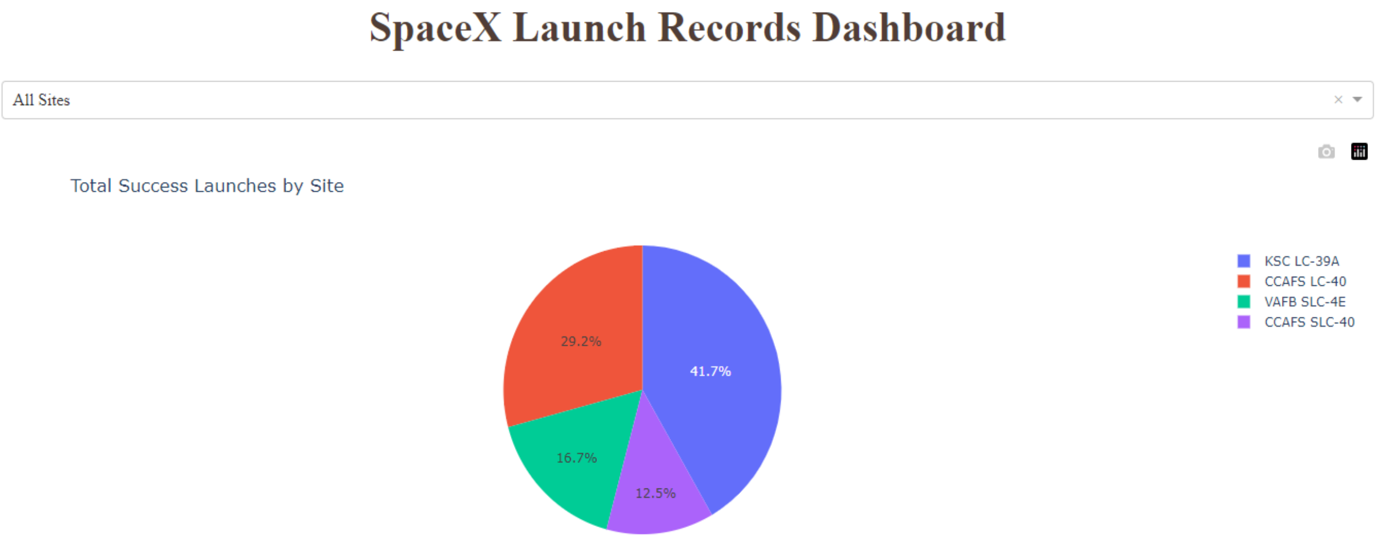
3.1. Get the best hyperparameters for each type of model
3.2. Calculate the accuracy for each model with the test set
3.3. Plot the Confusion Matrix

4.1. Comparison between the accuracies obtained from models
4.2. Select the model with the best accuracy

Redirect to the notebook

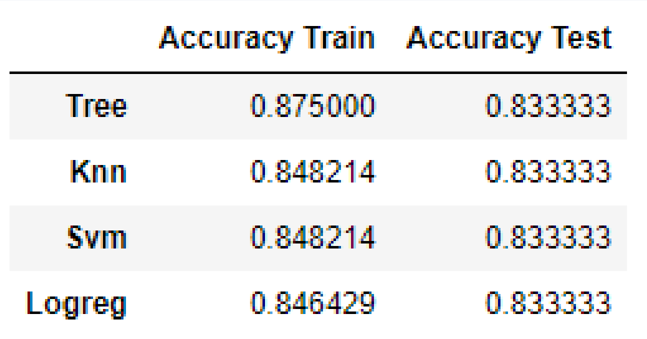
Results





SpaceX Launch Records Dashboard

All Sites

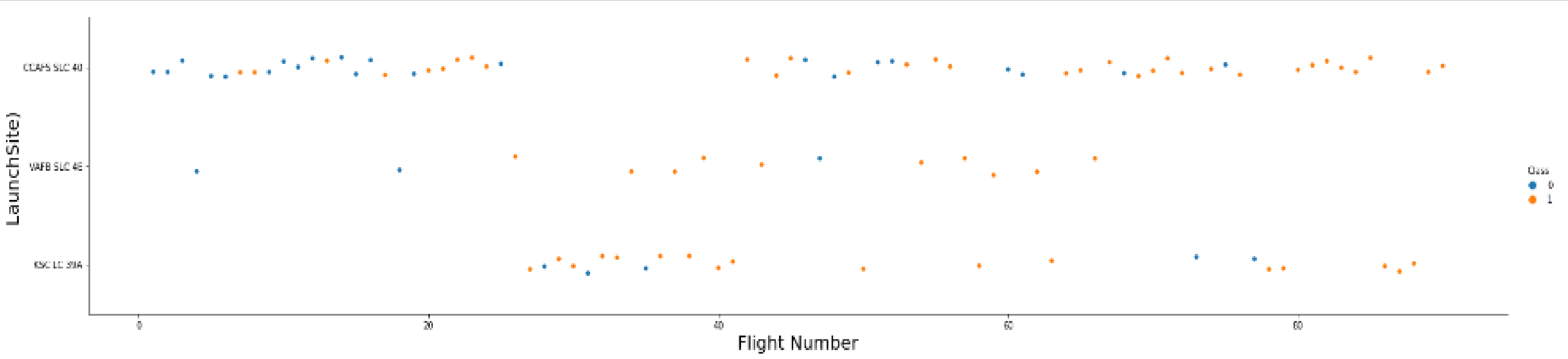


	Accuracy Train	Accuracy Test
Tree	0.875000	0.833333
Knn	0.848214	0.833333
Svm	0.848214	0.833333
Logreg	0.846429	0.833333

Section 2

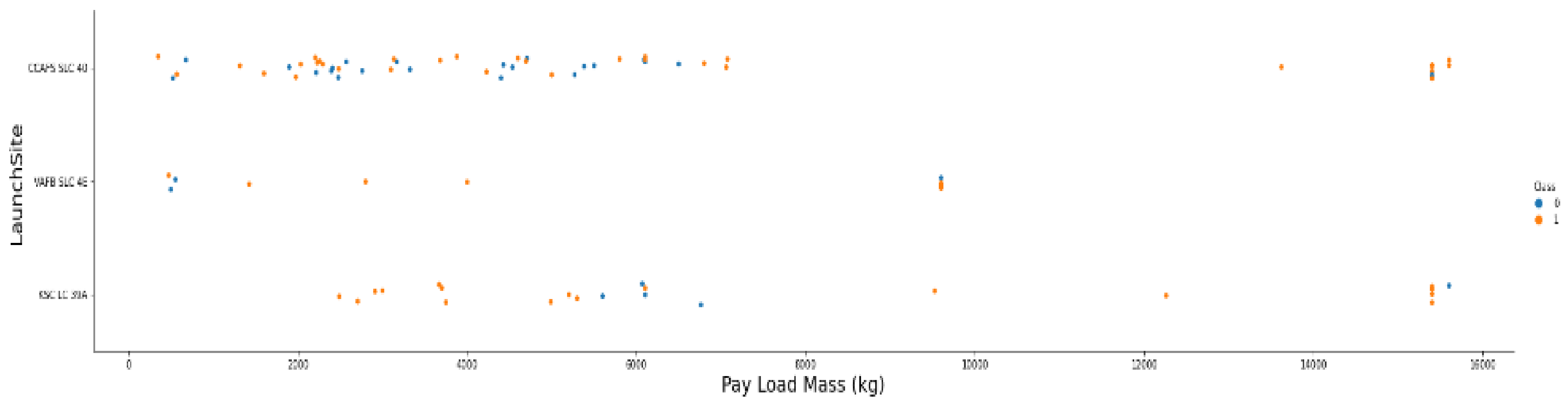
Insights drawn from EDA

Flight Number vs. Launch Site



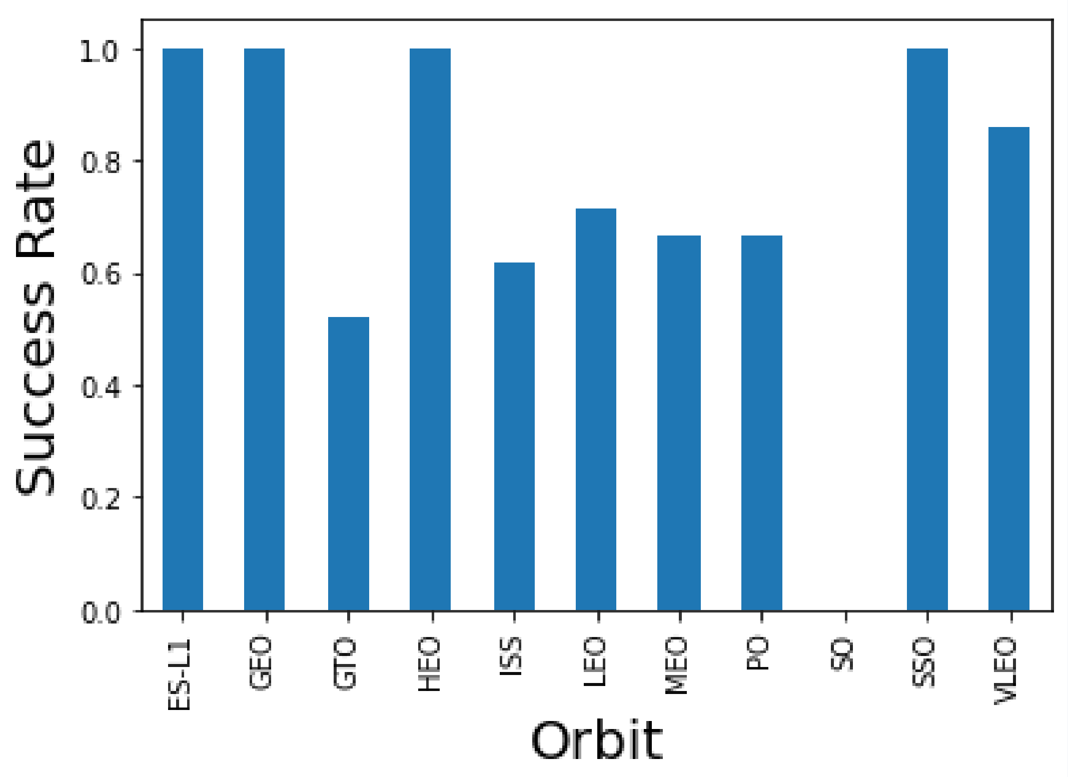
The success rate increases with the flight number for every launch site.

Payload vs. Launch Site



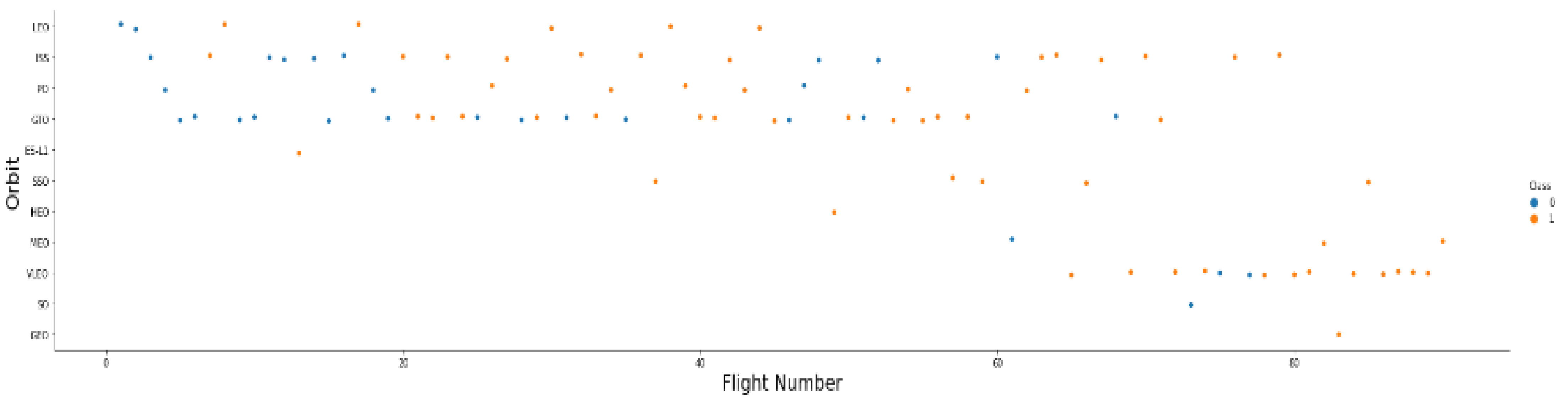
Depending on the launch site, maybe Pay Load Mass can be taken into consideration for a successful landing.

Success Rate vs. Orbit Type



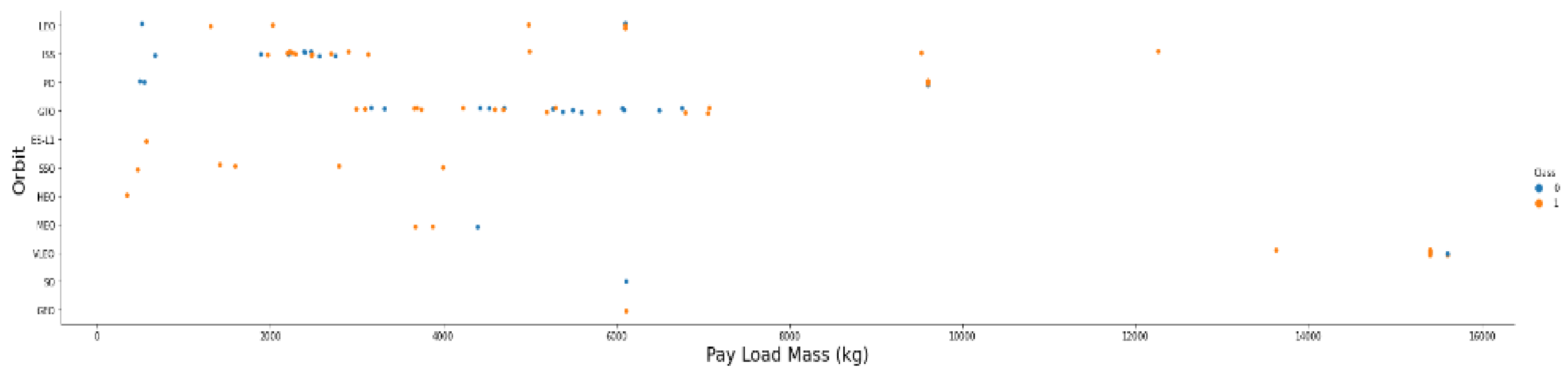
The success rate is different for each Orbit Type, the orbits ES-L1, GEO, HEO, SSO have the best success rates.

Flight Number vs. Orbit Type



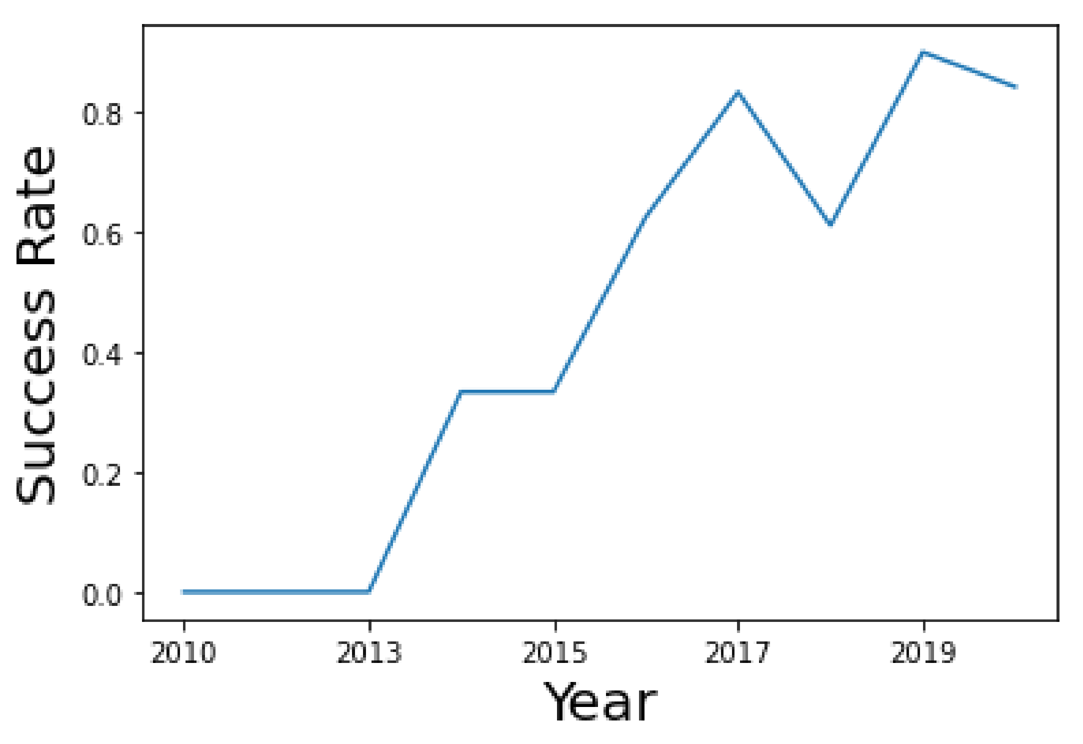
It seems to exist a relationship between the flight number and success rate for most of the orbits.

Payload vs. Orbit Type



For some orbits the payload mass seems to have an impact in success rate.

Launch Success Yearly Trend



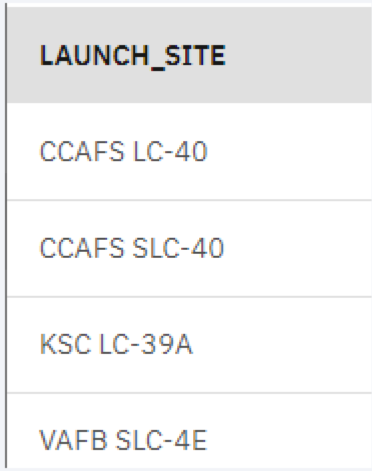
In 2013 is possible to see the start of success rate of the SpaceX, reaching the maximum success rate in 2019.

All Launch Site Names

SQL QUERY

```
SELECT DISTINCT(LAUNCH_SITE)

FROM SPACEXTBL;
```

RESULTS



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

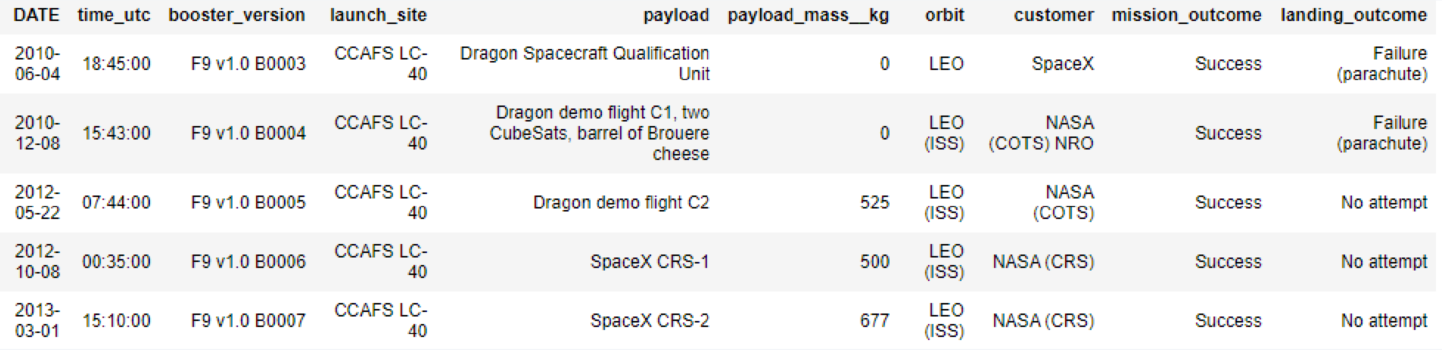
The use of DISTINCT does removes duplicates in LAUNCH_SITE

Launch Site Names Begin with 'CCA'

SQL QUERY

```
SELECT *

FROM SPACEXTBL

WHERE LAUNCH_SITE LIKE 'CCA%'

LIMIT 5;
```

RESULTS



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

The WHERE clause followed by LIKE clause filters launch sites that contain the substring ‘CCA’.

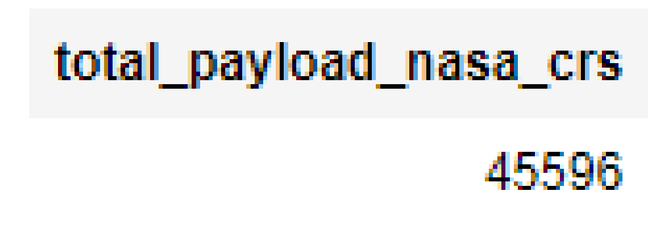
LIMIT 5 shows 5 first records from filtering.

Total Payload Mass

SQL QUERY

```
SELECT SUM(PAYLOAD_MASS__KG) AS

TOTAL_PAYLOAD_NASA_CRS

FROM SPACEXTBL

WHERE CUSTOMER = 'NASA (CRS)';
```

RESULTS



total_payload_nasa_crs
45596

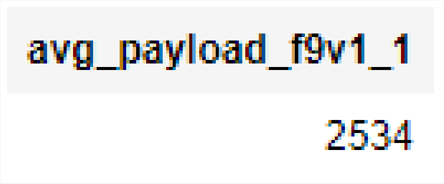
The query returns the sum of all payload masses where the customer is NASA (CRS).

Average Payload Mass by F9 v1.1

SQL QUERY

```
SELECT AVG(PAYLOAD_MASS__KG) AS

AVG_PAYLOAD_F9V1_1

FROM SPACEXTBL

WHERE BOOSTER_VERSION LIKE 'F9 v1.1%';
```

RESULTS



avg_payload_f9v1_1
2534

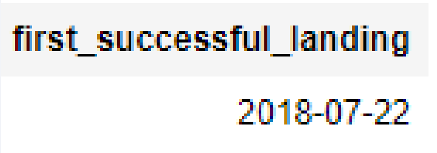
The query returns the average of all payload masses where the booster version contains the substring ‘F9 v1.1’.

First Successful Ground Landing Date

SQL QUERY

```
SELECT MIN(DATE) AS

FIRST_SUCCESSFUL_LANDING

FROM SPACEXTBL

WHERE LANDING_OUTCOME = 'Success';
```

RESULTS



first_successful_landing
2018-07-22

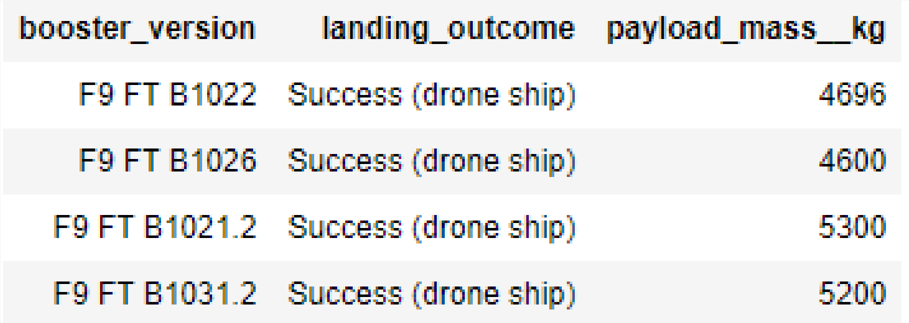
The WHERE clause filters dataset in order to keep only records where landing was successful. With the MIN function, we select the record with the oldest date (the minimum date available).

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL QUERY

```
SELECT BOOSTER_VERSION,

LANDING_OUTCOME, PAYLOAD_MASS__KG

FROM SPACEXTBL

WHERE LANDING_OUTCOME = 'Success (drone ship)'

AND PAYLOAD_MASS__KG > 4000

AND PAYLOAD_MASS__KG < 6000;
```

RESULTS



booster_version	landing_outcome	payload_mass__kg
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

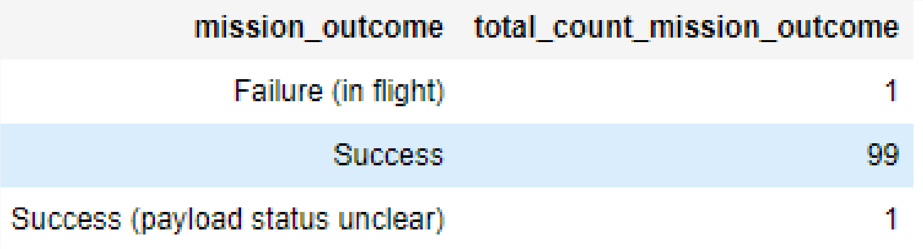
This query returns the booster version where landing was successful and payload mass is between 4000 and 6000 kg. The WHERE and AND clauses filter the dataset.

Total Number of Successful and Failure Mission Outcomes

SQL QUERY

```
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME)

AS TOTAL_COUNT_MISSION_OUTCOME

FROM SPACEXTBL

GROUP BY MISSION_OUTCOME;
```

RESULTS



mission_outcome	total_count_mission_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

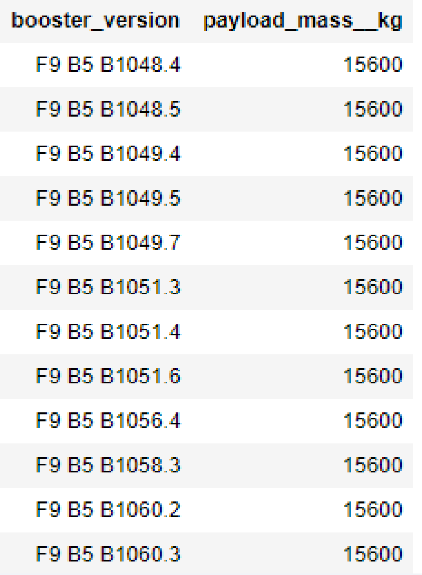
The COUNT function counts records filtered and the GROUP BY statement does group by mission outcomes.

Boosters Carried Maximum Payload

SQL QUERY

```
SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG

FROM SPACEXTBL

WHERE PAYLOAD_MASS__KG = (

    SELECT MAX(PAYLOAD_MASS__KG)

    FROM SPACEXTBL)

ORDER BY BOOSTER_VERSION

;
```

RESULTS



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

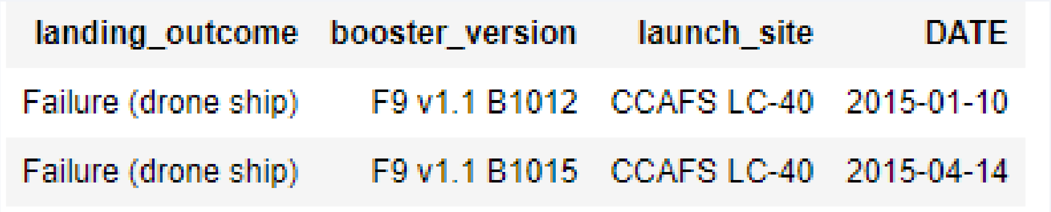
It was used a subquery to filter data by returning only the heaviest payload mass with MAX function. The main query uses subquery results and returns the booster version with the heaviest payload mass ordered by booster_version.

2015 Launch Records

SQL QUERY

```
SELECT LANDING_OUTCOME, BOOSTER_VERSION,

LAUNCH_SITE, DATE

FROM SPACEXTBL

WHERE LANDING_OUTCOME = 'Failure (drone ship)'

AND YEAR(DATE) = 2015

;
```

RESULTS



landing_outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

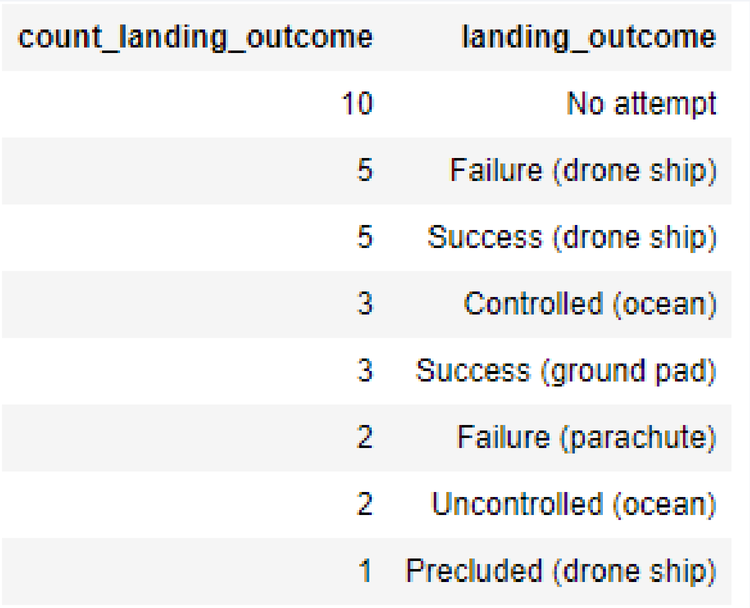
It was used a WHERE clause to filter only the landing_outcome = ‘Failure (drone ship) and the year of 2015.

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

SQL QUERY

```
SELECT COUNT(LANDING_OUTCOME) AS

COUNT_LANDING_OUTCOME, LANDING_OUTCOME

FROM SPACEXTBL

WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'

GROUP BY LANDING_OUTCOME

ORDER BY COUNT(LANDING_OUTCOME) DESC;
```

RESULTS



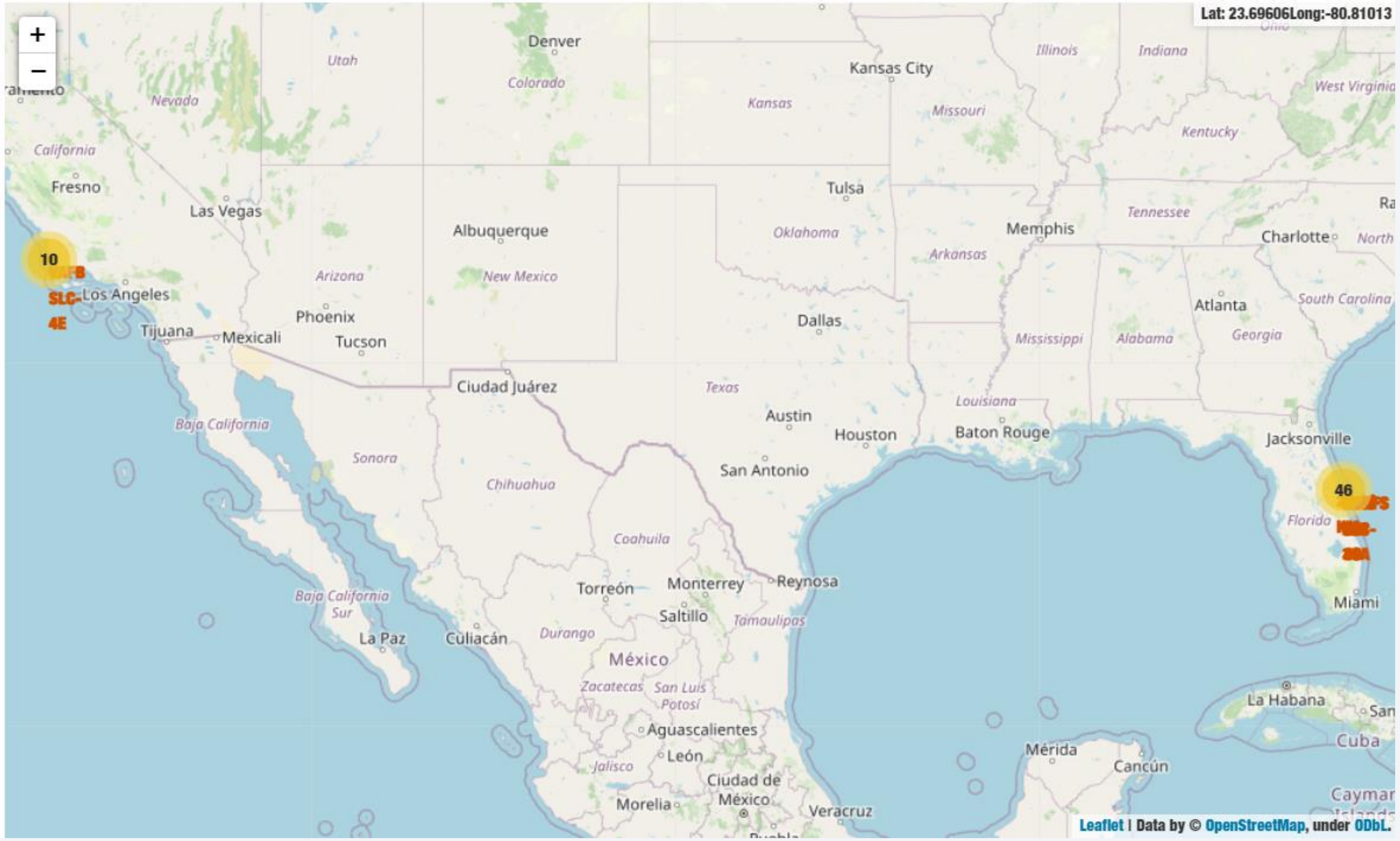
count_landing_outcome	landing_outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

It was used a WHERE clause to filter only the between 2010-06-04 and 2017-03-20, the counting was grouped by landing_outcome, and it was ordered by COUNT(landing_outcomes) DESC (higher to lower count).

Section 3

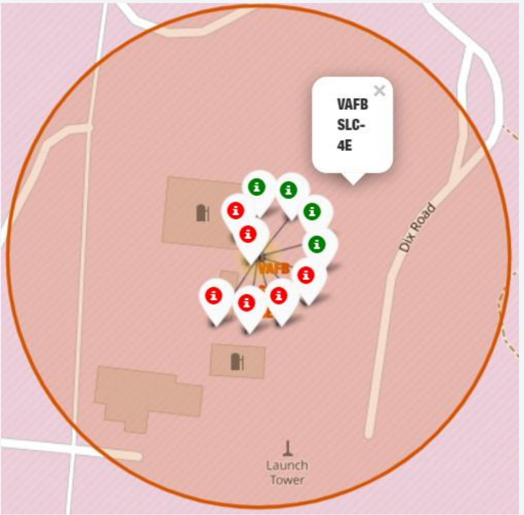
Launch Sites Proximities Analysis

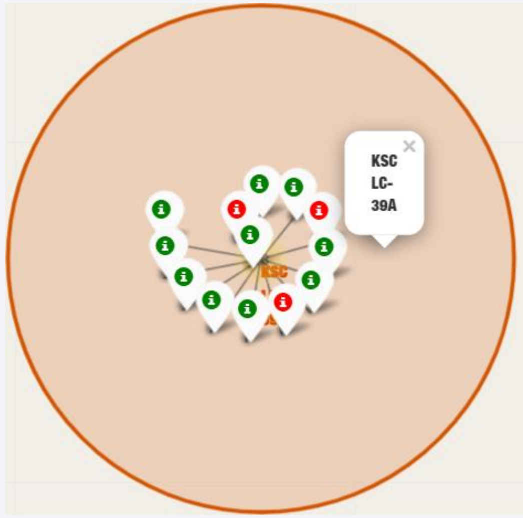
Folium Map – SpaceX Launch Sites

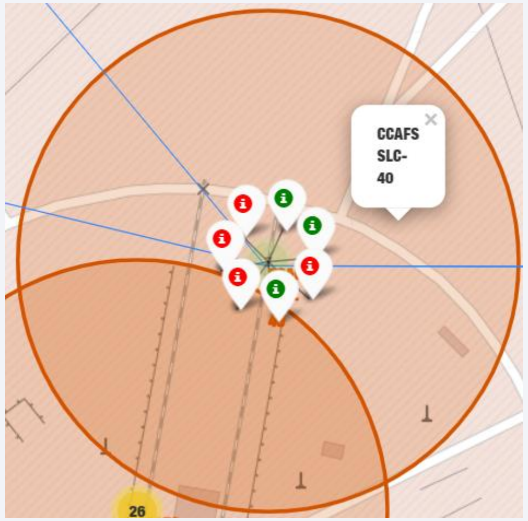

Lat: 23.69606Long:-80.81013
+
−
10
SLC-4E
46
Leaflet | Data by © OpenStreetMap, under ODbL.

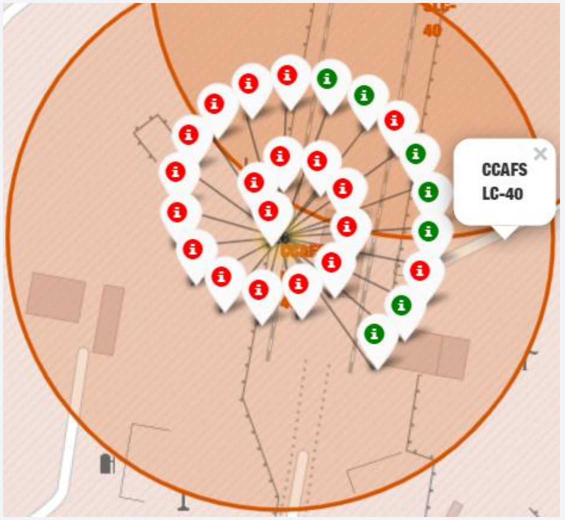
It is possible to check that all SpaceX launch sites are located on the coast of the United States, more precisely in California and Florida.

Folium Map – Color Markers



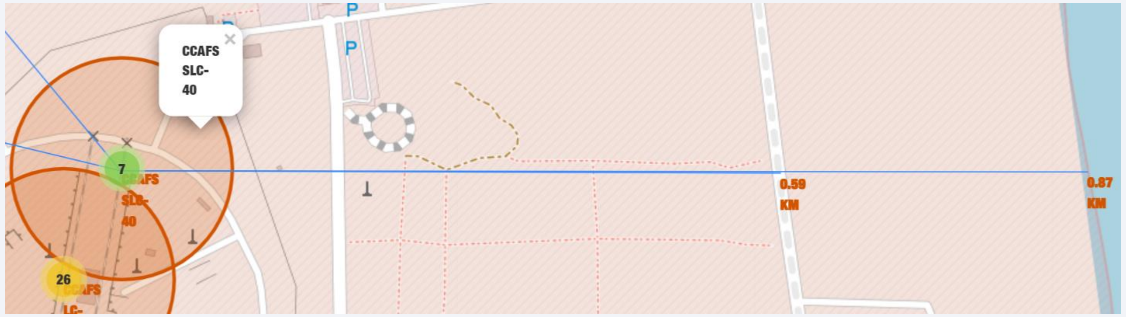


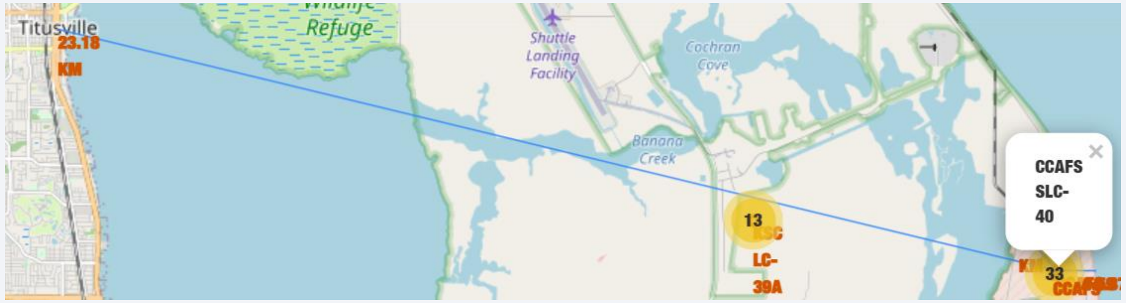


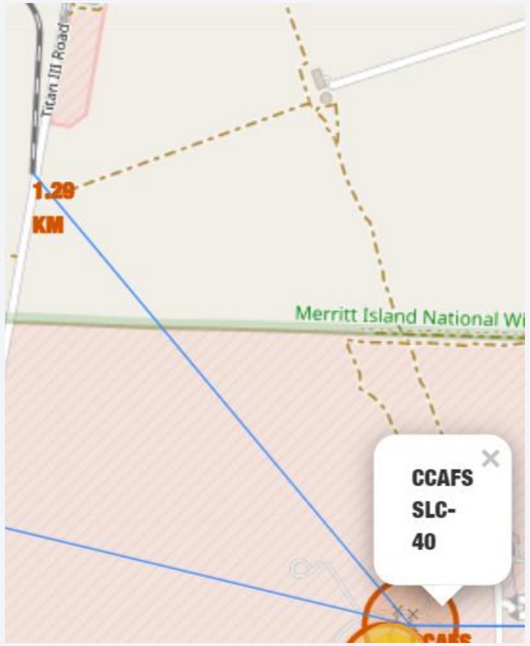


The green markers represent successful launches, and the red markers represent unsuccessful launches. It is possible to verify that KSC LC 39A has a higher launch success rate than the others launch sites.

Folium Map – Distance between CCAFS SLC-40 and Key Locations







CCAFS SLC-40 to Highway = 0.59 km

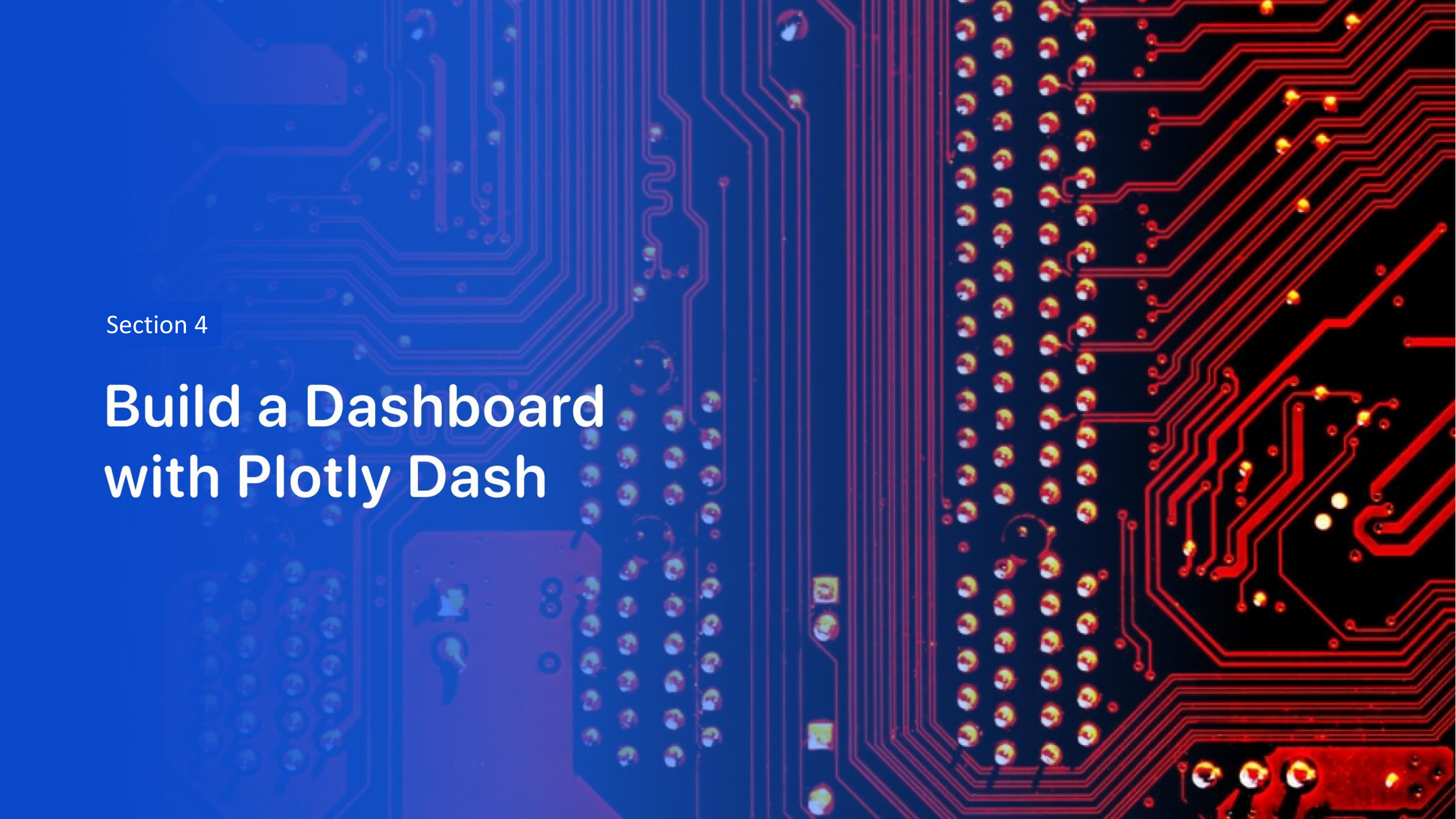
CCAFS SLC-40 to Coastline = 0.87 km

CCAFS SLC-40 to Railway = 1.29 km

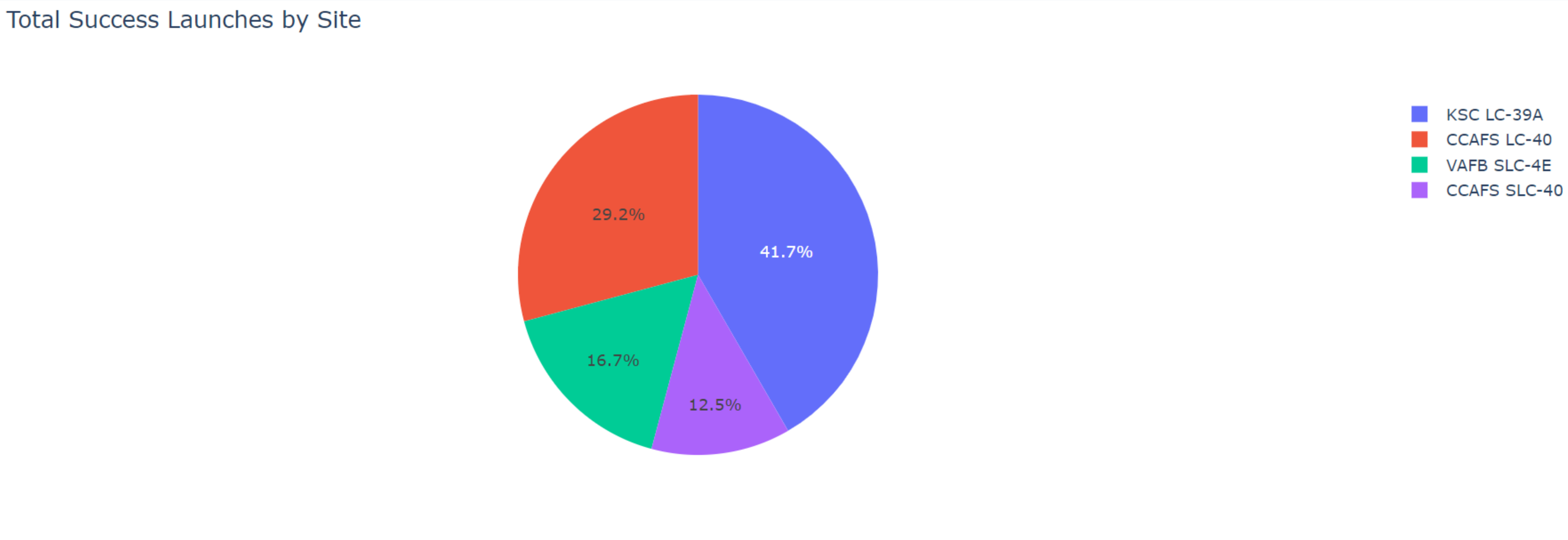
CCAFS SLC-40 to City = 23.18 km

Section 4

Build a Dashboard with Plotly Dash

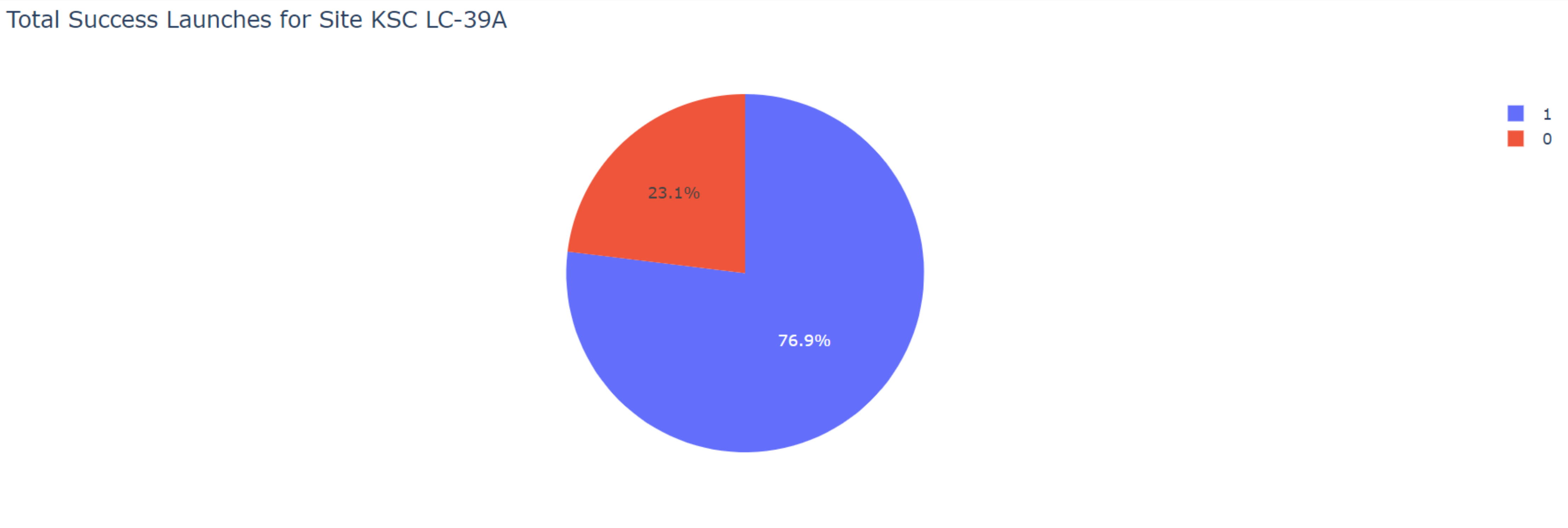


Dashboard - Total Successful launches by site



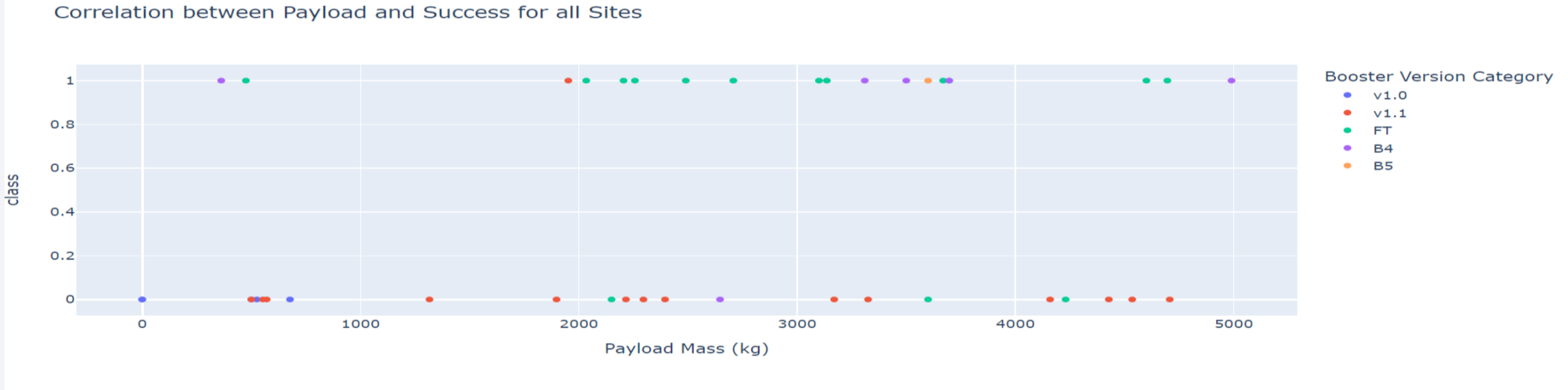
The KSC LC-39A has the most percentage of success, but it should be verified if it has the majority of the launches among all the sites to check if the success rate there is really higher than the other launch sites, or if it is just because of the number of launches.

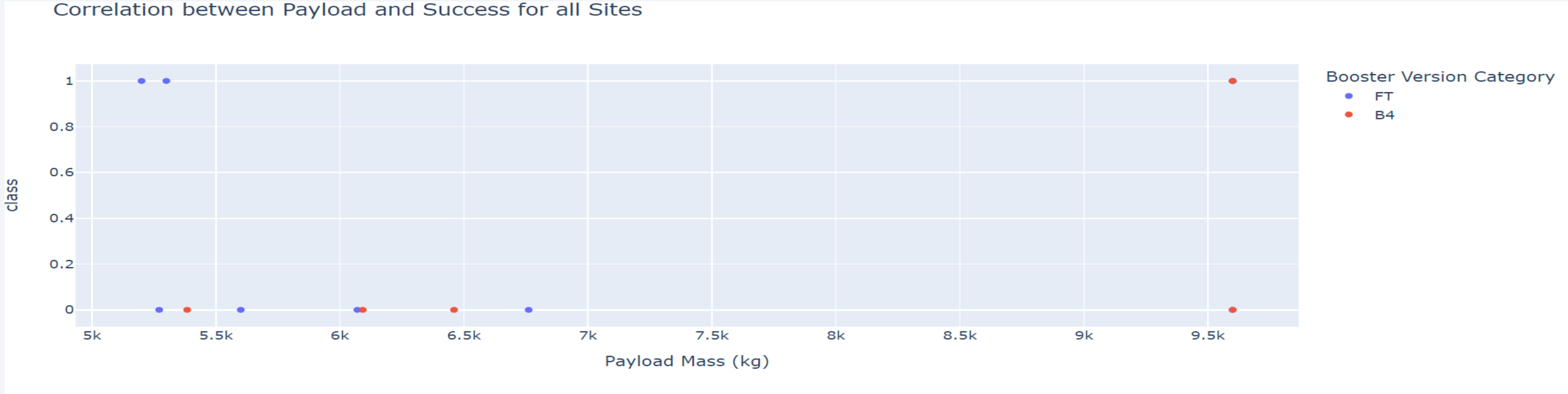
Dashboard – Success rate of the launches at KSC LC-39A



The site KSC LC-39A has indeed the best success rate between the four sites, with 76.9% of the launches being successful.

<Dashboard Screenshot 3>



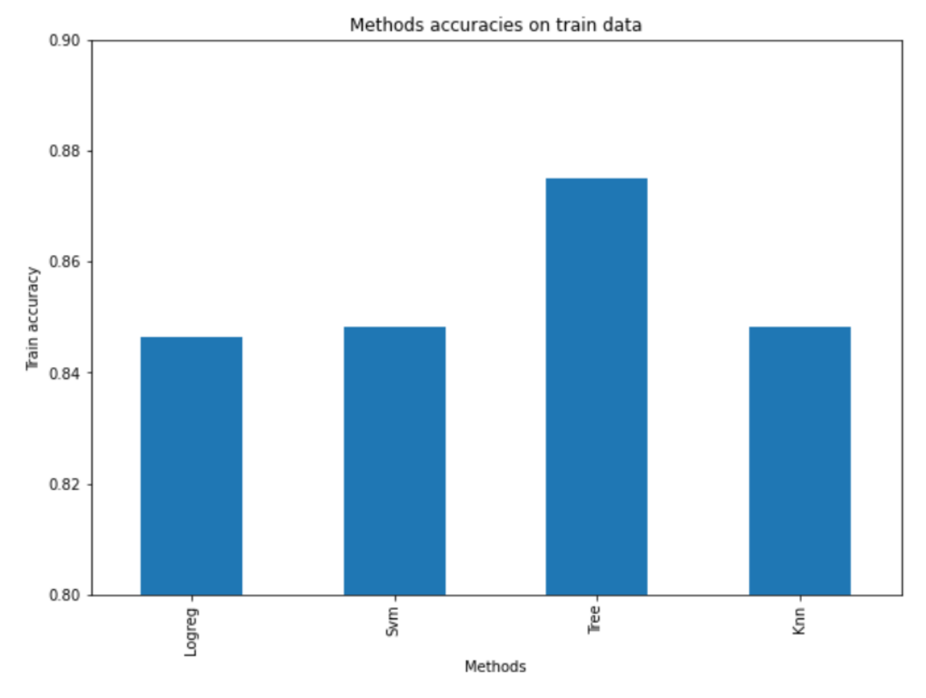


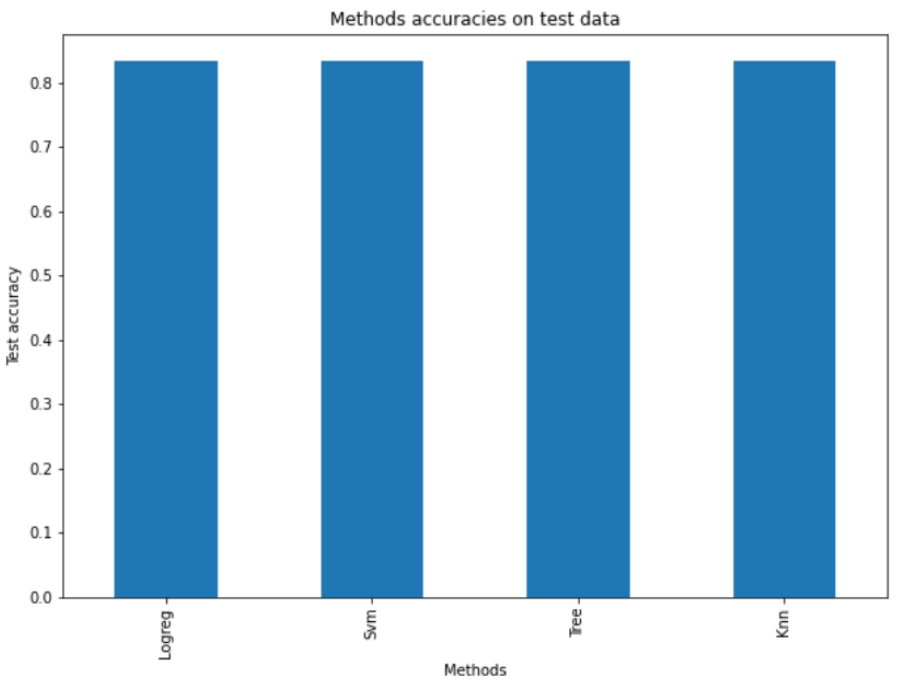
Payloads lower than 5.000 kg have a better success rate than the heavy weighted payloads above 5.000 kg.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

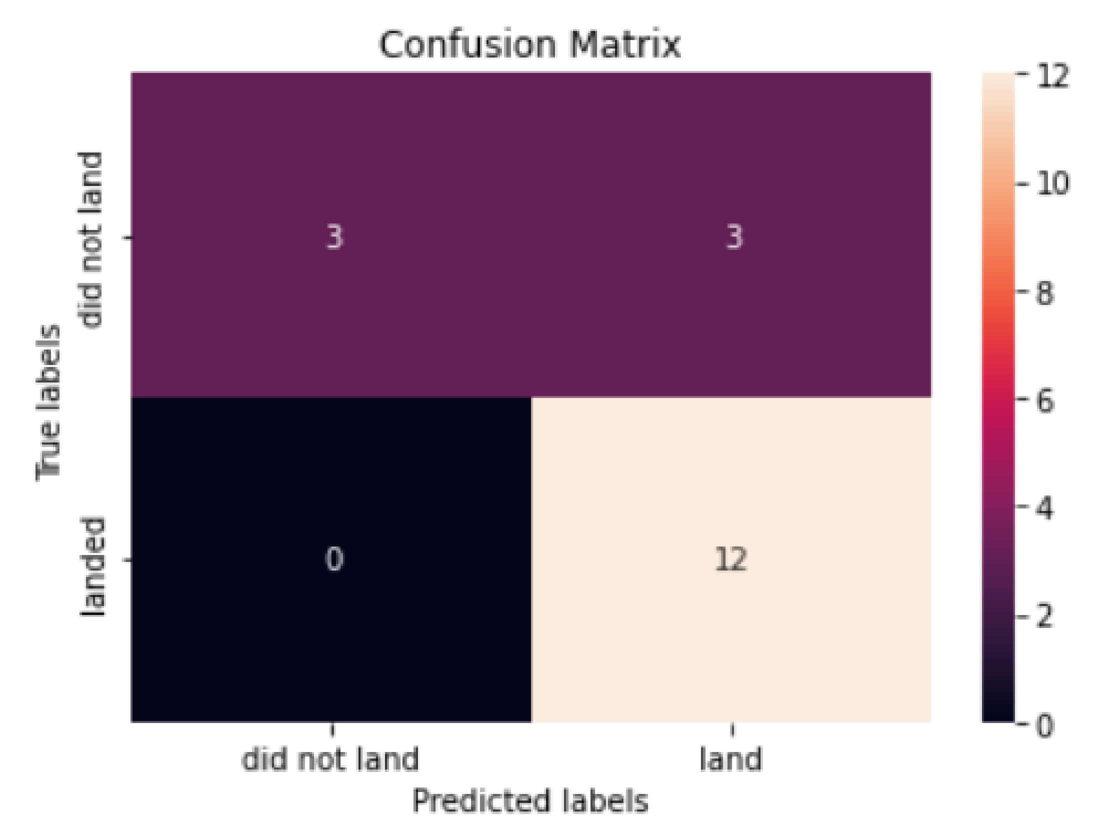




For the accuracy test, all methods were performed similarly. Maybe we could get more test data to decide between them to get a second thought. If we take into consideration the accuracy train, then the decision tree would be the best choice, with the following parameters:

tuned hpyerparameters (best parameters): {'criterion': 'gini', 'max_depth': 14, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 2, 'splitter': 'best'}

Confusion Matrix Decision Tree



There are too many false positives for the Confusion Matrix, it mislabeled 3 out 6 unsuccessful landings as successful.

Conclusions

- Several factors can explain the success of a mission as a launch site, the orbit, and especially the number of previous launches. We probably can assume that there has been an improvement in knowledge between launches that allowed them to get a higher chance of success after each launch.
- The orbits and the payload mass can be a criterion to be considered for the success of a mission. It does seem that some orbits do require a light or heavy payload mass. But generally, payloads below 5.000 kg perform better than heavy-weighted payloads.
- The orbits with the best success rates are GEO, HEO, SSO, and ES L1.
- It would be interesting to conduct another study to check other variables about the launch sites since we could not verify the cause for best performance in specific launch sites compared to the others.
- For this project, the Decision Tree Algorithm was chosen as the best model because it has a better train accuracy. When comparing the accuracy of the test set, all four models performed equally. In the future would be interesting measuring the models again with a larger dataset available.

Thank you!