IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ridhwaan Mayet
2023 01 13

Outline

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

Executive Summary

Methodologies

- Data Collection via API, web scraping
- Exploratory Data Analaysis (EDA) with Data Viz
- EDA with SQL
- Interactive Map with Folium
- Dashboards with Plotly Dash
- Predictive analysis

Results

-

Introduction

Background

- The aim of this project is to predict if the Falcon 9 first stage will succesfully land. SpaceX says on its website that the launch cost $62m. Other providers cost upward of $165m each. The price difference is explained by the fact that SpaceX can reuse the first stage. By determining if the stage will land, we can determine the cost of a launch. This information is interesting for another company if they wanted to compete with SpaceX for a rocket launch.

Problems to solve

- What are the characteristics of a successful or failed 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

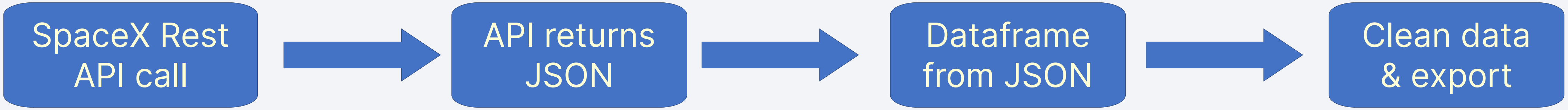
Methodology

Summary

- Data collection methodology:
  - SpaceX RESTAPI
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Dropping unnecessary columns
  - One Hot Encoding for classification models
- 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 and evaluate classification models

Data Collection

Datasets are collected from Rest SpaceX API and ebscrapping Wikipedia. The information obtained by the API are rocket, launches, payload information. The Space X REST API URL is api.spacexdata.com/v4/



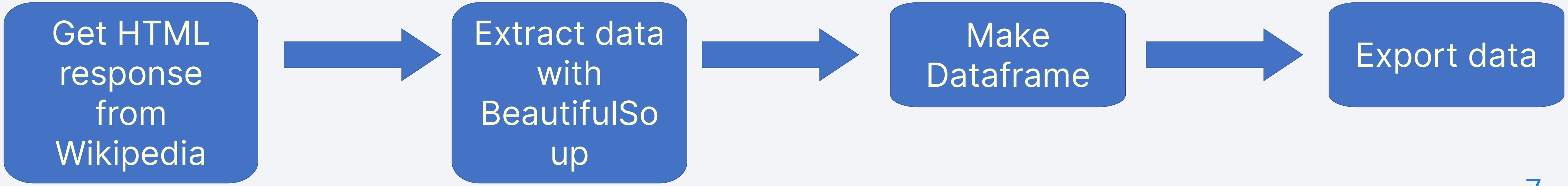
SpaceX Rest API call

API returns JSON

Dataframe from JSON

Clean data & export

The data from web scrapping Wikipedia covers launches, landings and payloads. Wikipedia URL



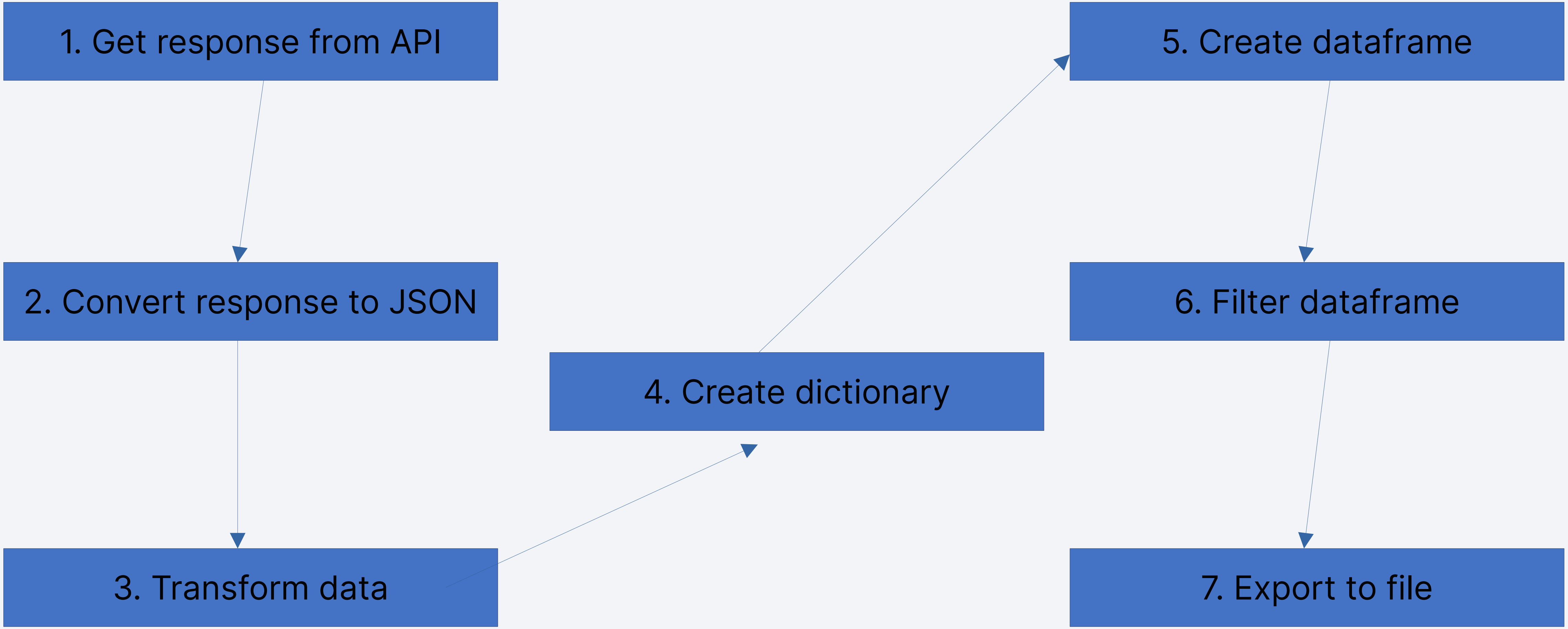
Get HTML response from Wikipedia

Extract data with BeautifulSoup

Make Dataframe

Export data

Data Collection – SpaceX API



1. Get response from API

2. Convert response to JSON

3. Transform data

4. Create dictionary

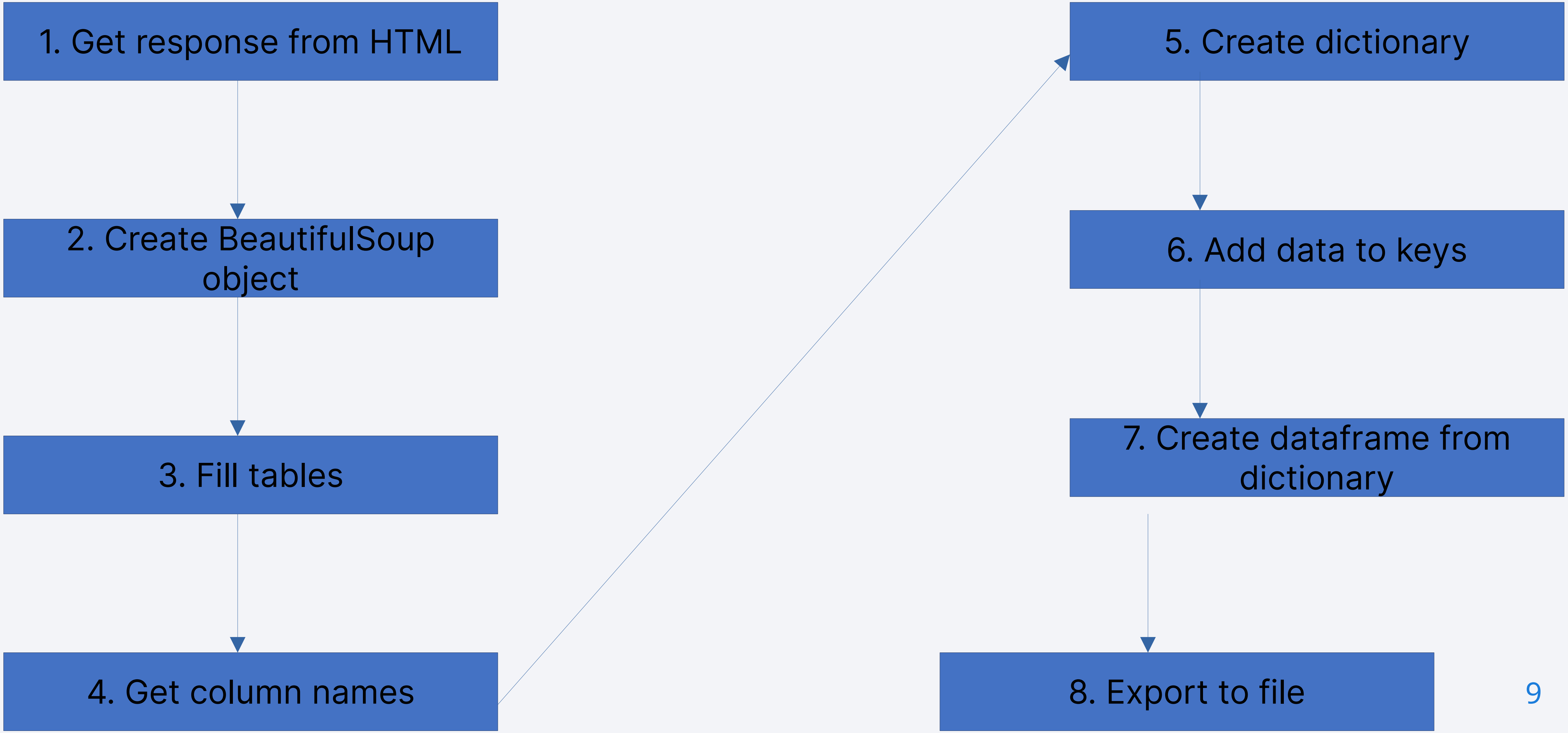
5. Create dataframe

6. Filter dataframe

7. Export to file

Link: Lab1_DataCollectionwitAPI.ipynb

Data Collection - Scraping



1. Get response from HTML

2. Create BeautifulSoup object

3. Fill tables

4. Get column names

5. Create dictionary

6. Add data to keys

7. Create dataframe from dictionary

8. Export to file

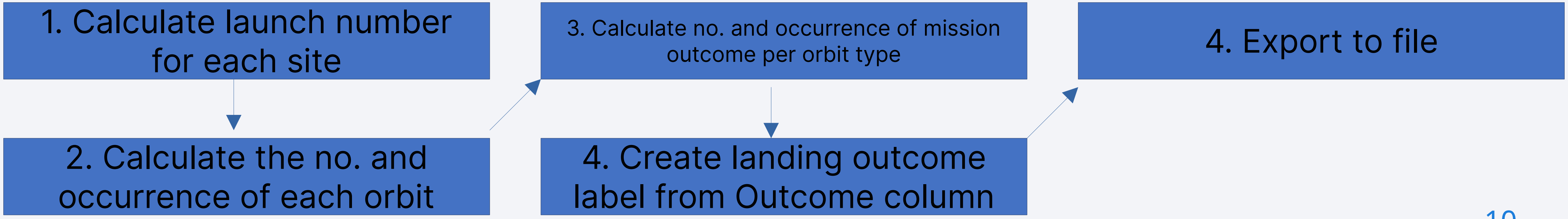
Link Lab2WebscrapinFalco9.ipynb

Data Wrangling

In the dataset, there are several cases where the bosster did not land successfully:

- True ocean, True RTLS, True ASDS means the mission was successful
  False ocean, False RTLS, False ASDS means the mission was not successful

We need to transform string variables into categorical variables where 1 means the mission has been successful and 0 means the mission failed



1. Calculate launch number for each site

2. Calculate the no. and occurrence of each orbit

3. Calculate no. and occurrence of mission outcome per orbit type

4. Create landing outcome label from Outcome column

4. Export to file

Link Lab2DataWrangling.ipynb

EDA with Data Visualization

Scatter plots

- Flight number vs. Payload
- Flight number vs. Launch site
- Payload vs Launch site
- Orbit vs. Flight Number
- Payload vs. Orbit
- Orbit vs. Payload

Scatter plots show the relationship between variables. This relationship is called the correlation

Bar chart

- Success rate vs. Orbit

Bar charts show the relationship between numeric and categorical variables

Line graph

- Success rate vs. year

Linegraphs show variables and their trends. Line graphs can help to show global behavior and make prediction for unseen data

Link EDAwithDataVisualization.ipynb

EDA with SQL

We performed SQL queries to gather and understand data from dataset:

- Displaying 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 records which will display the month names, failure landing outcomes in drone ship, booster versions, launch_site for the months in year 2015.
- Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Link EDAwithSQL_sqllite3.ipynb

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).
- The grouping of points in a cluster to display multiple and different information for the same coordinates (folium.plugins.MarkerCluster).
- Markers to show successful and unsuccessful landings. Green for successful landing and Red for unsuccessful landing. (folium.map.Marker,folium.Icon).
- Markers to show distance between launch site to key locations (railway, highway, coastway, city) and plot a line between them. (folium.map.Marker, folium.PolyLine, folium.features.DivIcon)

These objects are created in order to understand better the problem and the data. We can show easily all launch sites, their surroundings and the number of successful and unsuccessful landings.

Link DataVisualizationwithFolium.ipynb

Build a Dashboard with Plotly Dash

Dashboard has dropdown, pie chart, range slider and scatter plot components

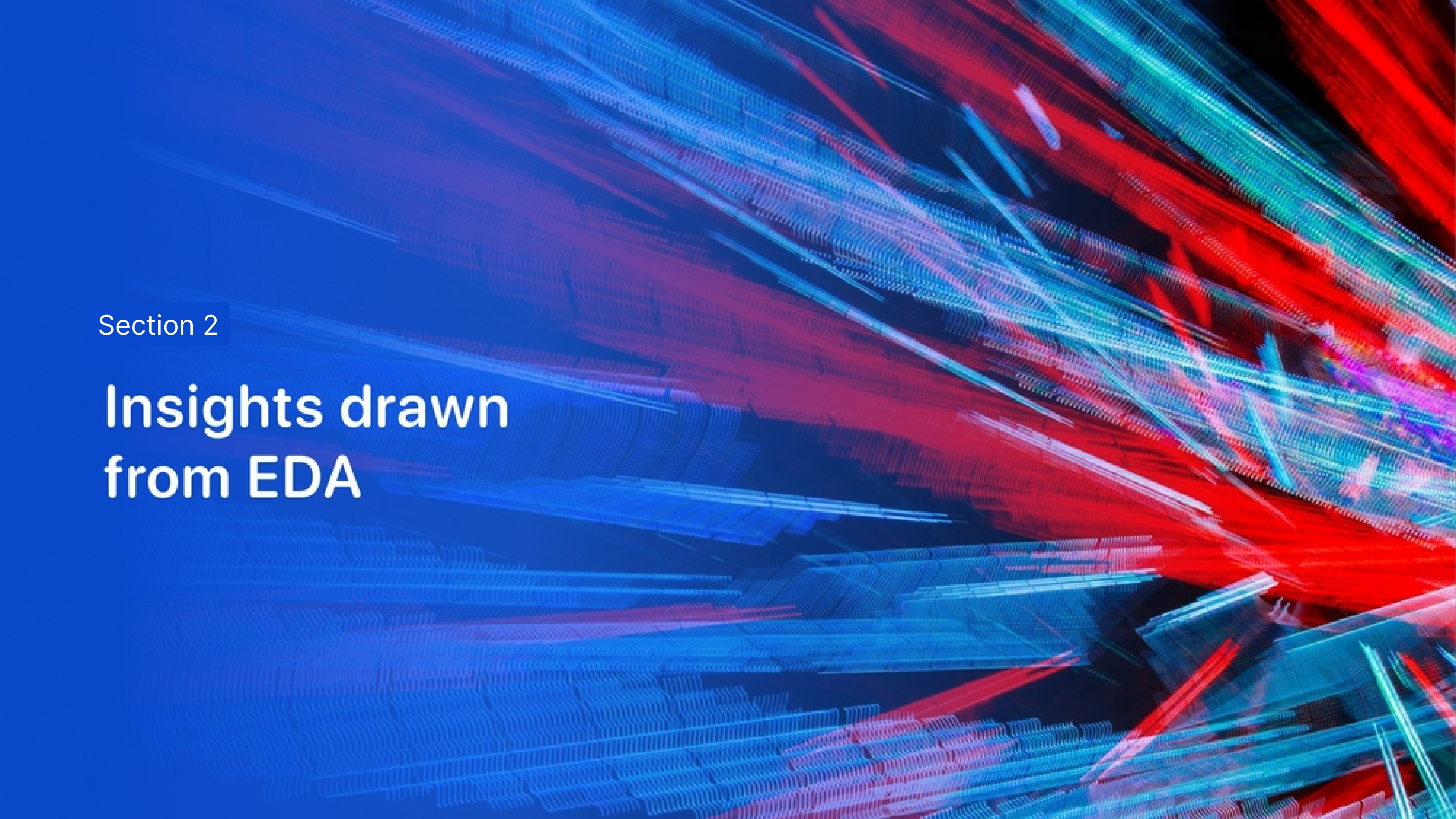
- Dropdown allows a user to choose the launch site or all launch sites (dash_core_components.Dropdown).
- Pie chart shows the total success and the total failure for the launch site chosen with thedropdown component (plotly.express.pie).
- Rangeslider allows a user to select a payload mass in a fixed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

Link DashboardApplicationwithPlotlyDash.ipynb

Predictive Analysis (Classification)

- Data preparation
  - Load dataset
  - Normalize data
  - Split data into training and test sets.
- Model preparation
  - Selection of machine learning algorithms
  - Set parameters for each algorithm to GridSearchCV
  - Training GridSearchModel models with training dataset
- Model evaluation
  - Get best hyperparameters for each type of model
  - Compute accuracy for each model with test dataset
  - Plot Confusion Matrix
- Model comparison
  - Comparison of models according to their accuracy
  - The model with the best accuracy will be chosen (see Notebook for result)

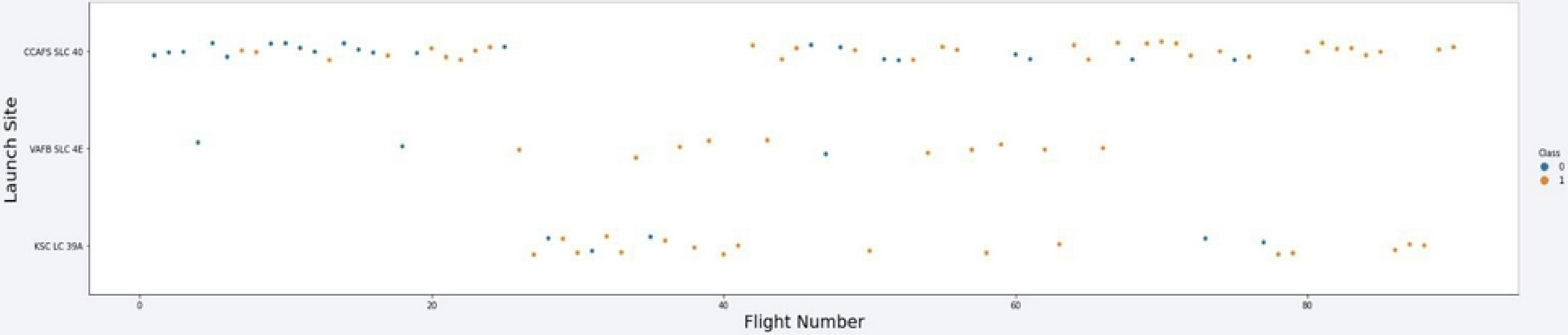
Link MachineLearningPrediction_Part_5.ipynb



Section 2

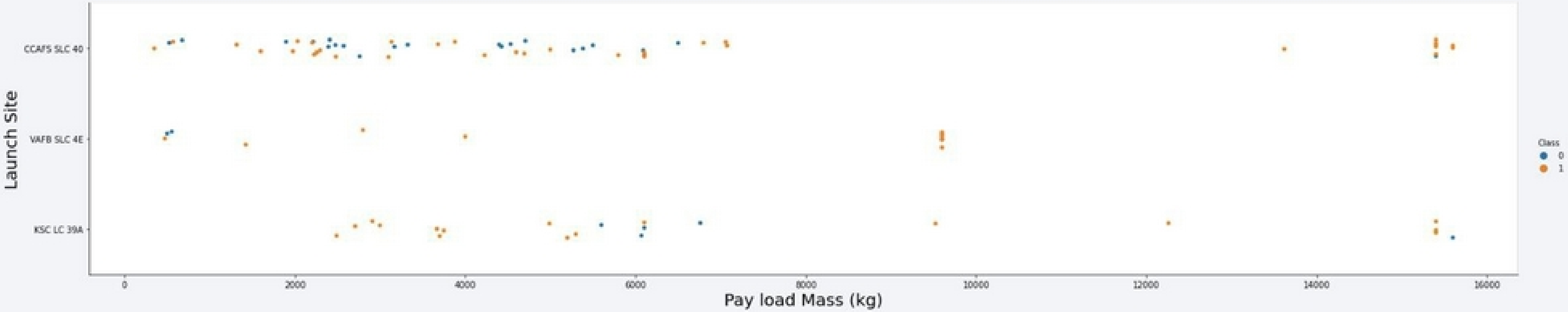
Insights drawn from EDA

Flight Number vs. Launch Site



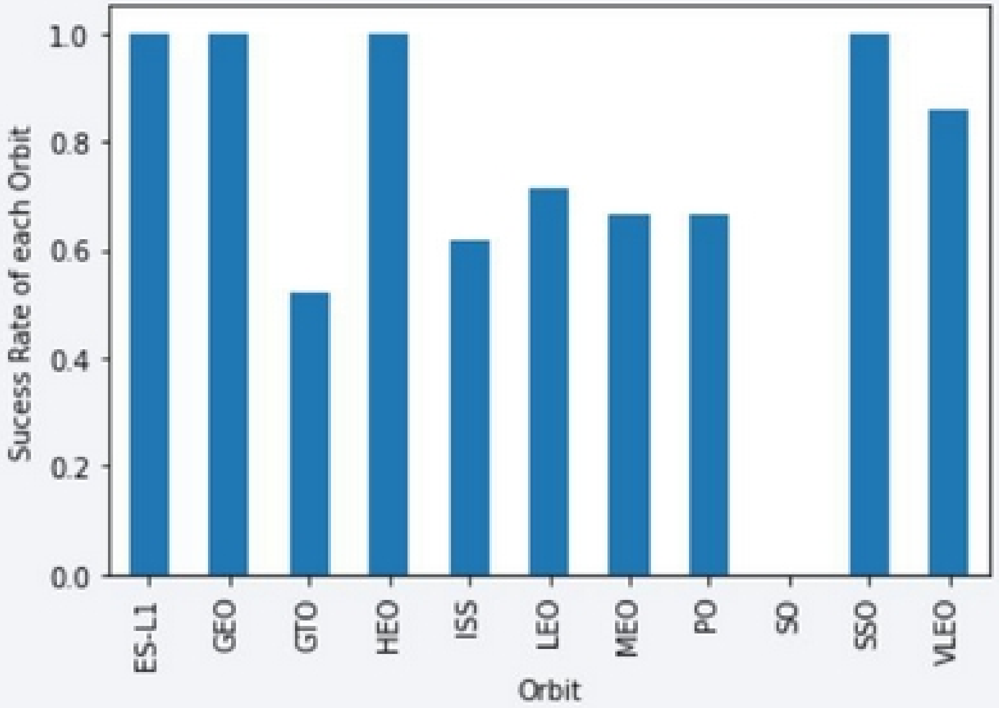
- According to the plot above, it's possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful;
- In second place VAFB SLC 4E and third place KSC LC 39A;
- It's also possible to see that the general success rate improved over time.

Payload vs. Launch Site



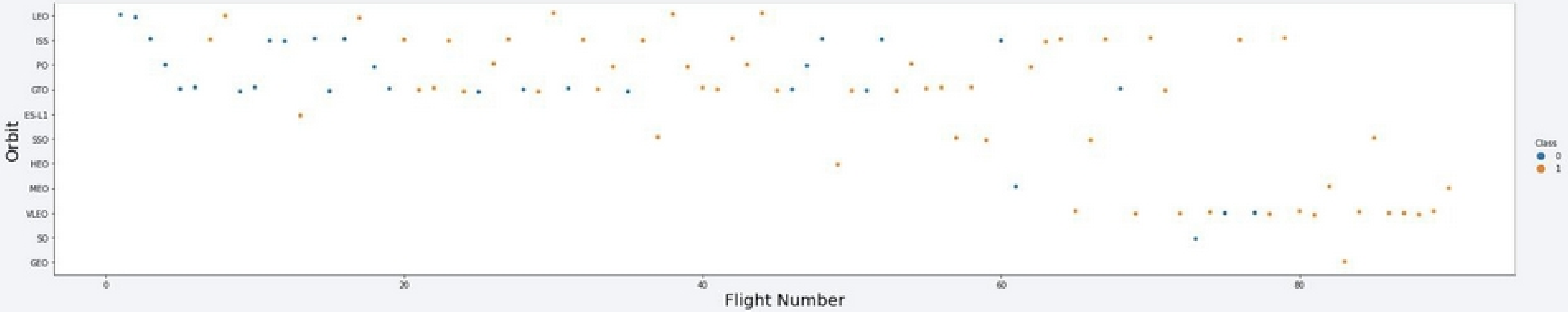
- Payloads below 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.
- Depending on the launch site, a heavier payload may be a consideration for a successful landing. On the other hand, a too heavy payload can make a landing fail.

Success Rate vs. Orbit Type



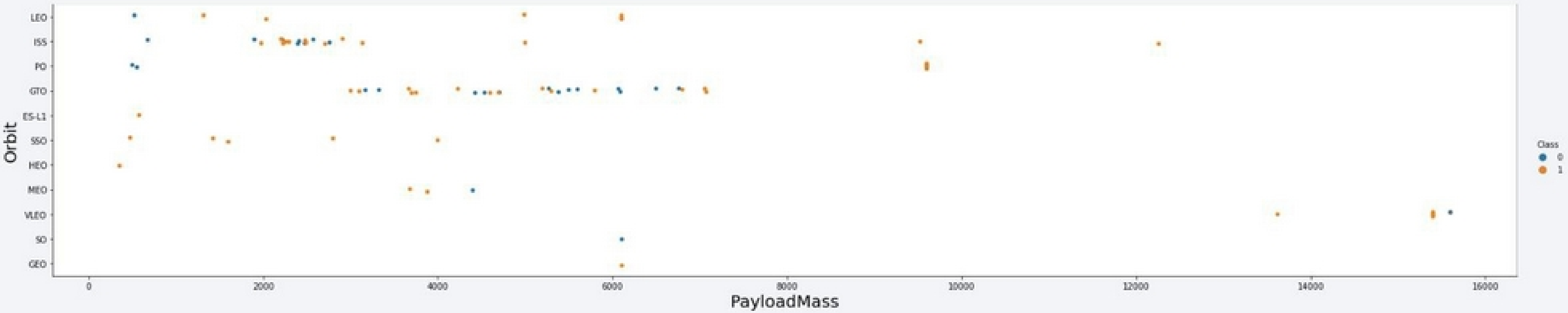
- With this plot, we can see success rate for different orbit types. We note that ES-L1, GEO, HEO, SSO have the best success rate.

Flight Number vs. Orbit Type



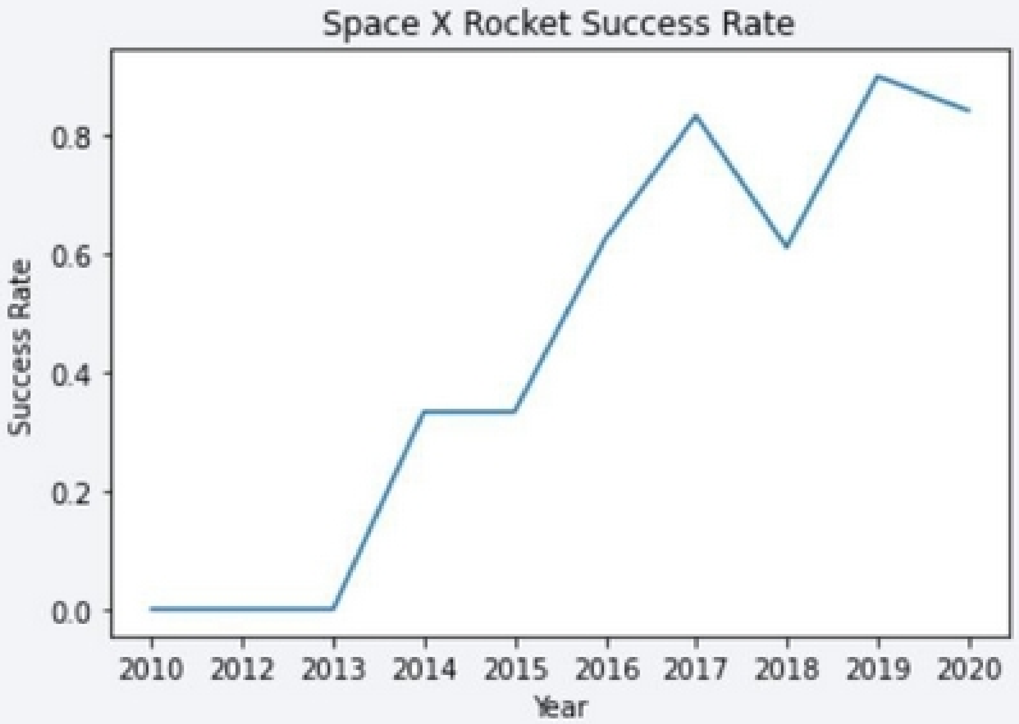
We notice that the success rate increases with the number of flights for the LEO orbit. For some orbits like GTO, there is no relation between the success rate and the number of flights. But we can suppose that the high success rate of some orbits like SSO or HEO is due to the knowledge learned during former launches for other orbits.

Payload vs. Orbit Type



The weight of the payloads can have a great influence on the success rate of the launches in certain orbits. For example, heavier payloads improve the success rate for the LEO orbit. Another finding is that decreasing the payload weight for a GTO orbit improves the success of a launch.

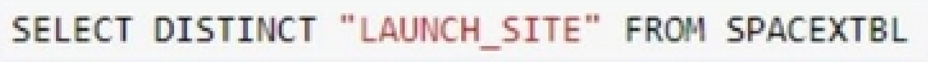
Launch Success Yearly Trend



- Success rate started increasing in 2013 and kept until 2020;
- It seems that the first three years were a period of adjusts and improvement of technology.

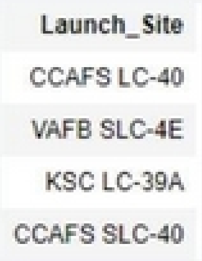
All Launch Site Names

SQLQuery



```
SELECT DISTINCT "LAUNCH_SITE" FROM SPACEXTBL
```

Results



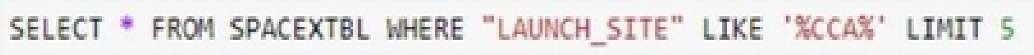
Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Explanation

The use of DISTINCT in the query allows to remove duplicate LAUNCH_SITE.

Launch Site Names Begin with 'CCA'

SQL Query

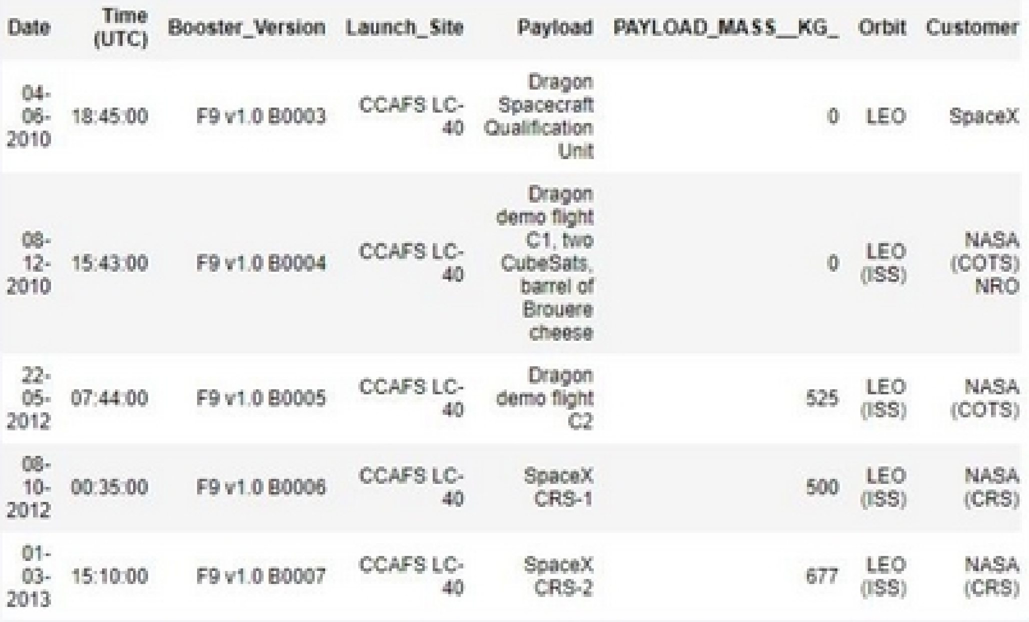


```
SELECT * FROM SPACEXTBL WHERE "LAUNCH_SITE" LIKE '%CCA%' LIMIT 5
```

Explanation

The WHERE clause followed by LIKE clause filters launch sites that contain the substring CCA. LIMIT 5 shows 5 records from filtering.

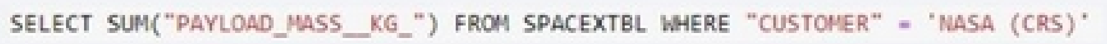
Results



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
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
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

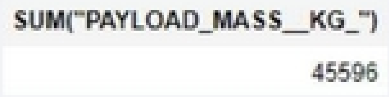
Total Payload Mass

SQLQuery



```
SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "CUSTOMER" = 'NASA (CRS)'
```

Results



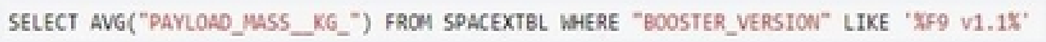
SUM("PAYLOAD_MASS__KG_")
45596

Explanation

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

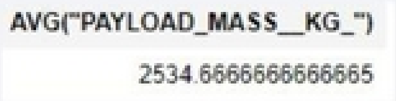
Average Payload Mass by F9 v1.1

SQL Query



```
SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "BOOSTER_VERSION" LIKE '%F9 v1.1%'
```

Results



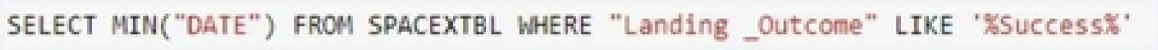
AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

Explanation

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

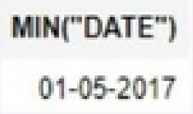
First Successful Ground Landing Date

SQL Query



```
SELECT MIN("DATE") FROM SPACEXTBL WHERE "Landing _Outcome" LIKE '%Success%'
```

Results



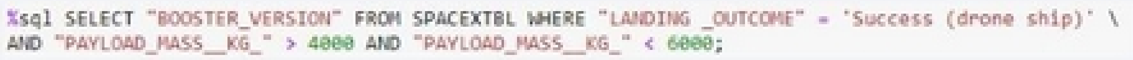
MIN("DATE")
01-05-2017

Explanation

With this query, we select the oldest successful landing. 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.

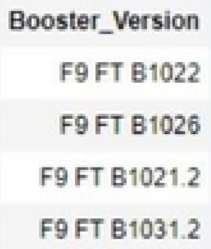
Successful Drone Ship Landing with Payload between 4000 and 6000

SQL Query



```
%sql SELECT "BOOSTER_VERSION" FROM SPACEXTBL WHERE "LANDING _OUTCOME" = 'Success (drone ship)' \
AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;
```

Results



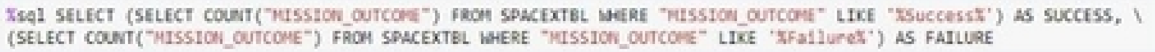
Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Explanation

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.

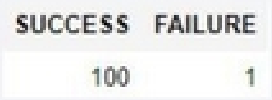
Total Number of Successful and Failure Mission Outcomes

SQL Query



```
%sql SELECT (SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Success%') AS SUCCESS, \
(SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Failure%') AS FAILURE
```

Results



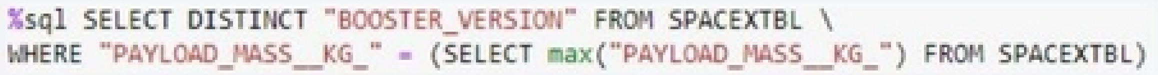
SUCCESS	FAILURE
100	1

Explanation

With the first SELECT, we show the subqueries that return results. The first subquery counts the successful mission. The second subquery counts the unsuccessful mission. The WHERE clause followed by LIKE clause filters mission outcome. The COUNT function counts records filtered.

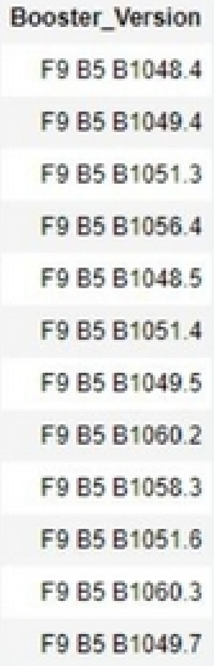
Boosters Carried Maximum Payload

SQL Query



```
%sql SELECT DISTINCT "BOOSTER_VERSION" FROM SPACEXTBL \
WHERE "PAYLOAD_MASS__KG_" = (SELECT max("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
```

Results



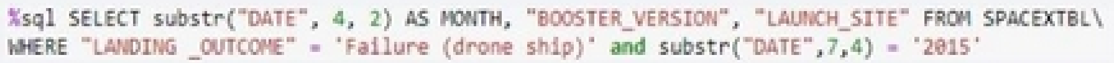
Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

Explanation

We used a subquery to filter data by returning only the heaviest payload mass with MAX function. The main query uses subquery results and returns unique booster version (SELECT DISTINCT) with the heaviest payload mass.

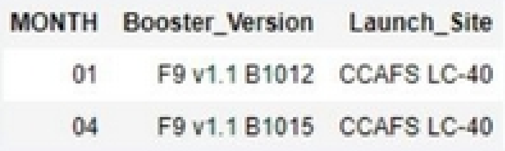
2015 Launch Records

SQLQuery



```
%sql SELECT substr("DATE", 4, 2) AS MONTH, "BOOSTER_VERSION", "LAUNCH_SITE" FROM SPACEXTBL\
WHERE "LANDING _OUTCOME" = 'Failure (drone ship)' and substr("DATE",7,4) = '2015'
```

Results



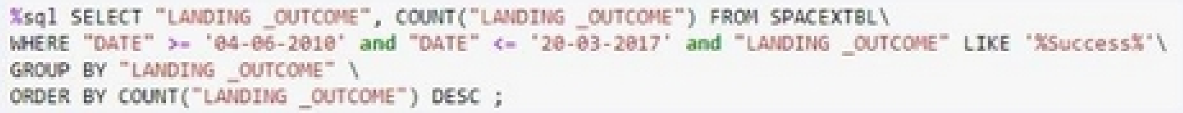
MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

Explanation

This query returns month, booster version, launch site where landing was unsuccessful and landing date took place in 2015. Substr function process date in order to take month or year. Substr(DATE, 4, 2) shows month. Substr(DATE,7, 4) shows year.

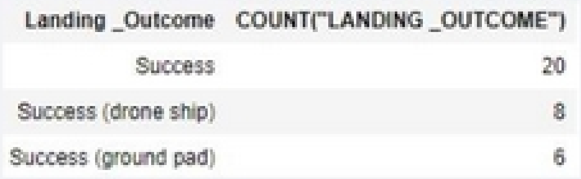
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

SQL Query



```
%sql SELECT "LANDING _OUTCOME", COUNT("LANDING _OUTCOME") FROM SPACEXTBL\
WHERE "DATE" >= '04-06-2010' and "DATE" <= '20-03-2017' and "LANDING _OUTCOME" LIKE '%Success%'\
GROUP BY "LANDING _OUTCOME" \
ORDER BY COUNT("LANDING _OUTCOME") DESC ;
```

Results



Landing _Outcome	COUNT("LANDING _OUTCOME")
Success	20
Success (drone ship)	8
Success (ground pad)	6

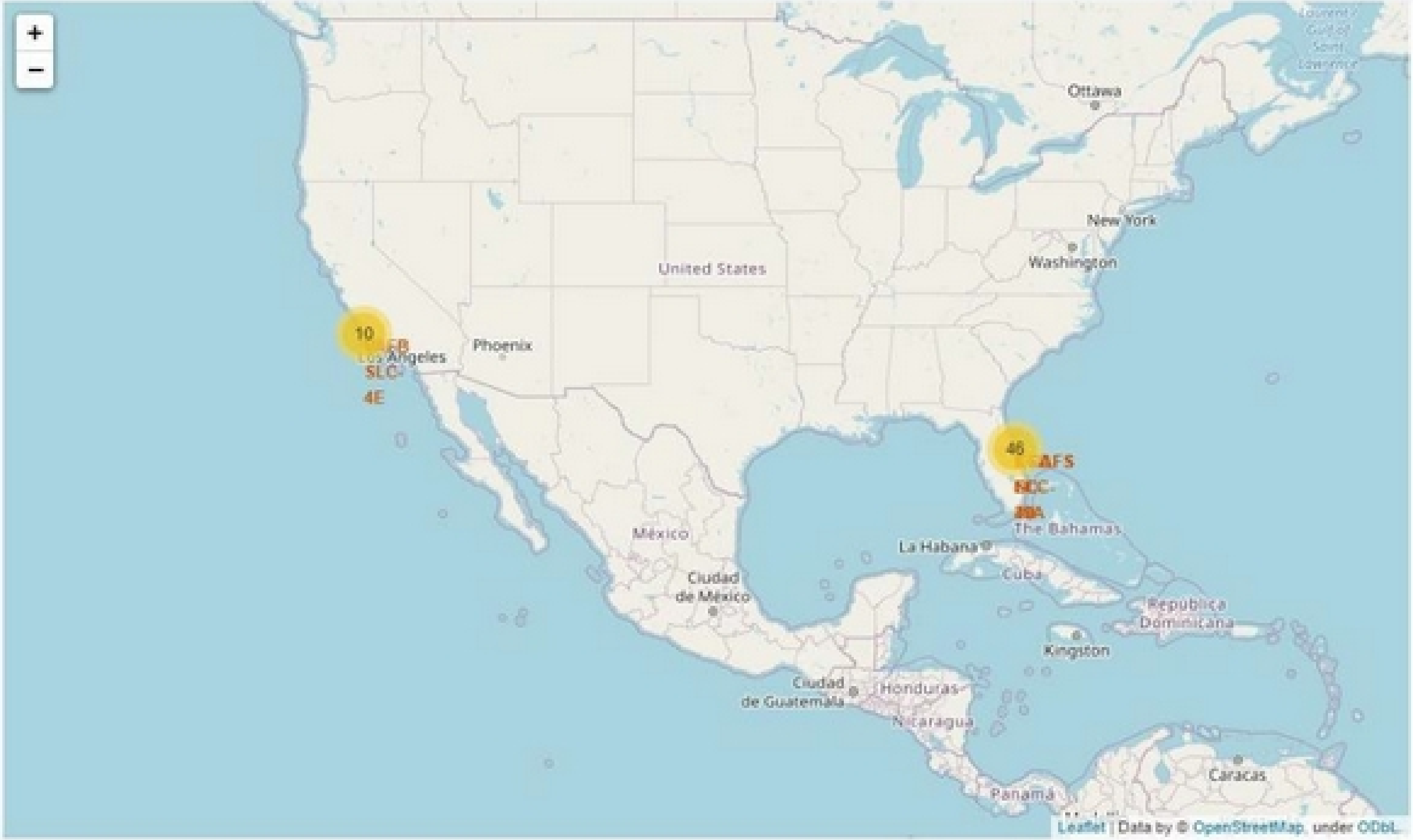
Explanation

This query returns landing outcomes and their count where mission was successful and date is between 04/06/2010 and 20/03/2017. The GROUP BY clause groups results by landing outcome and ORDER BY COUNT DESC shows results in decreasing order.

Section 3

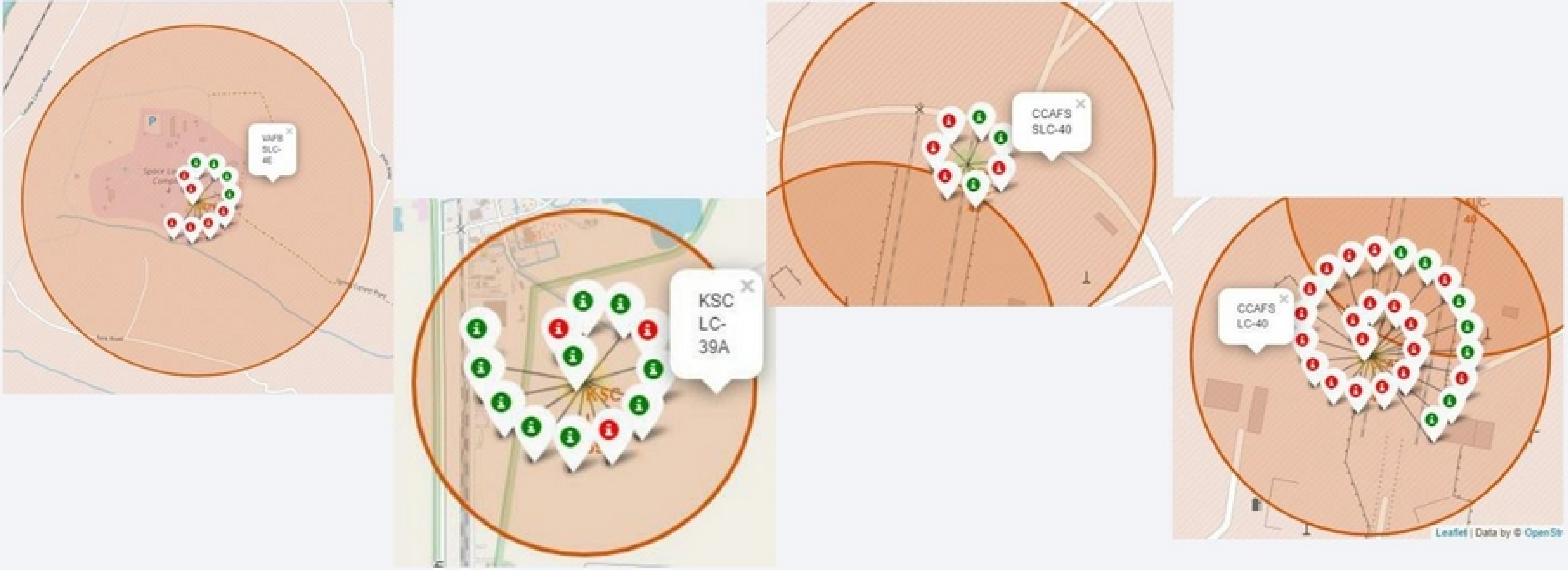
Launch Sites Proximities Analysis

All Launch Sites



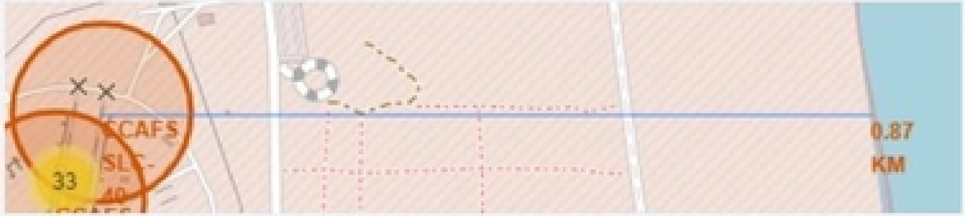
- Launch sites are near sea, probably by safety, but not too far from roads and railroads.
- We see that SpaceX launch sites are located on the United States.

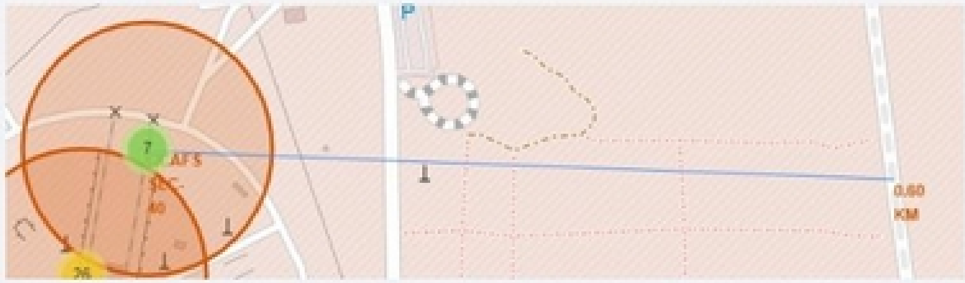
Launch Outcomes by Site

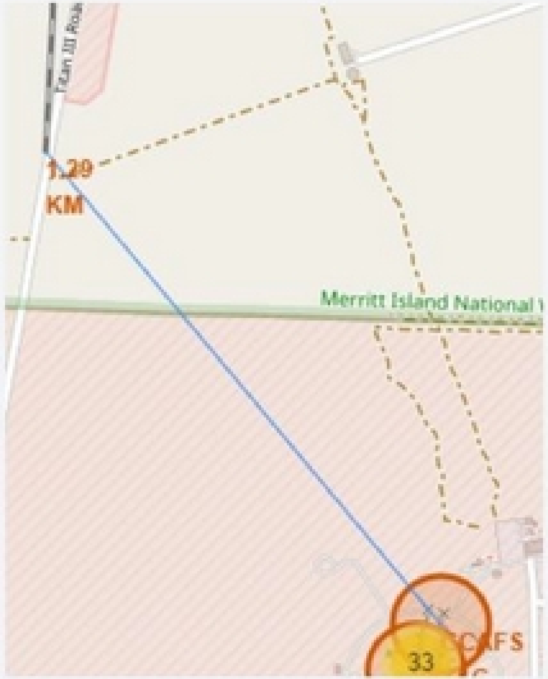


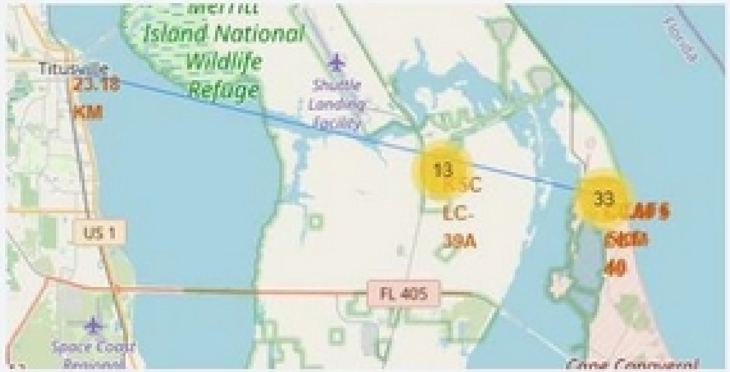
- Green marker represents successful launches. Red marker represents unsuccessful launches. We note that KSC LC-39A has a higher launch success rate.

Distances between CCAFS SLC-40 and its Proximities

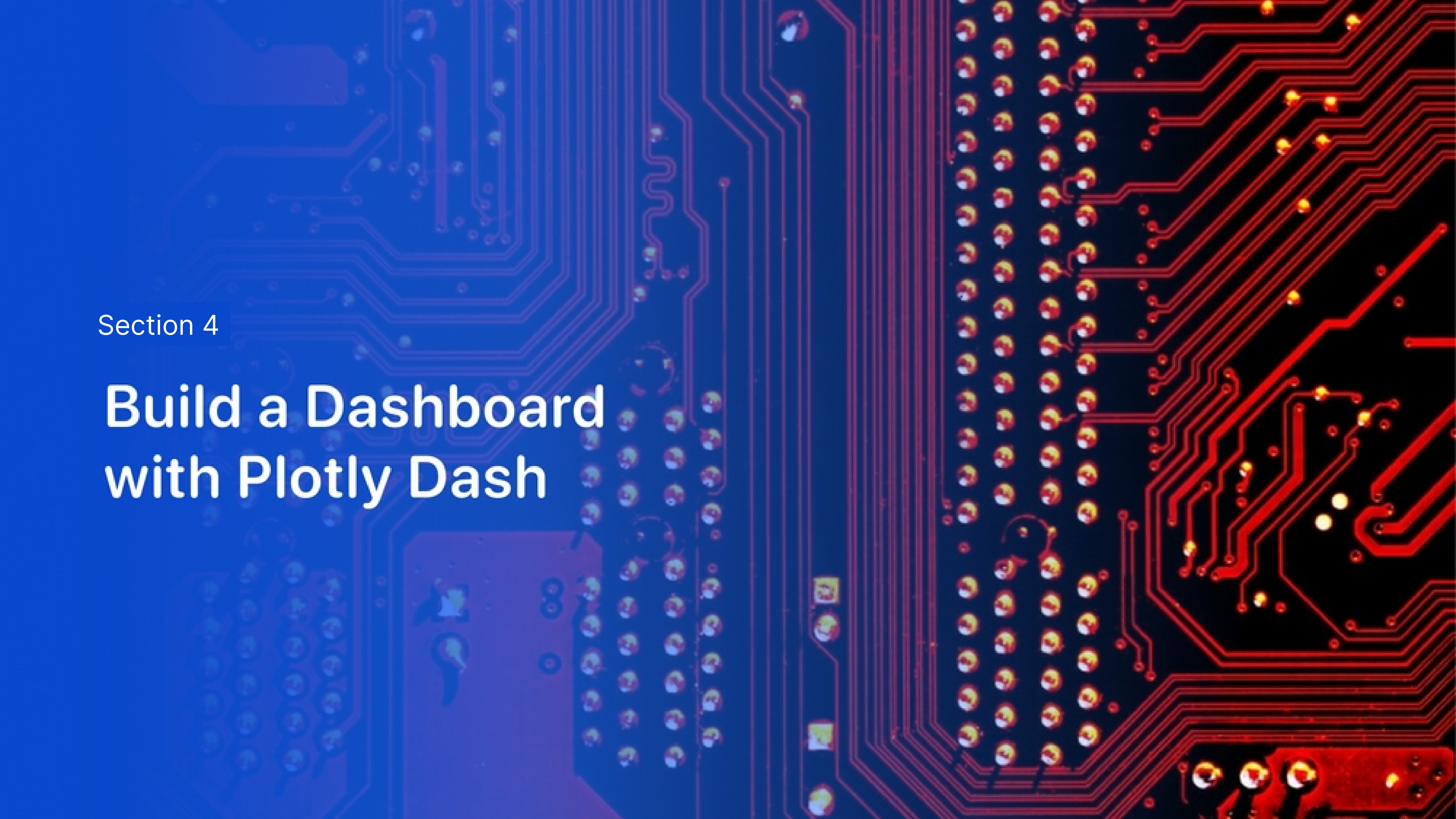








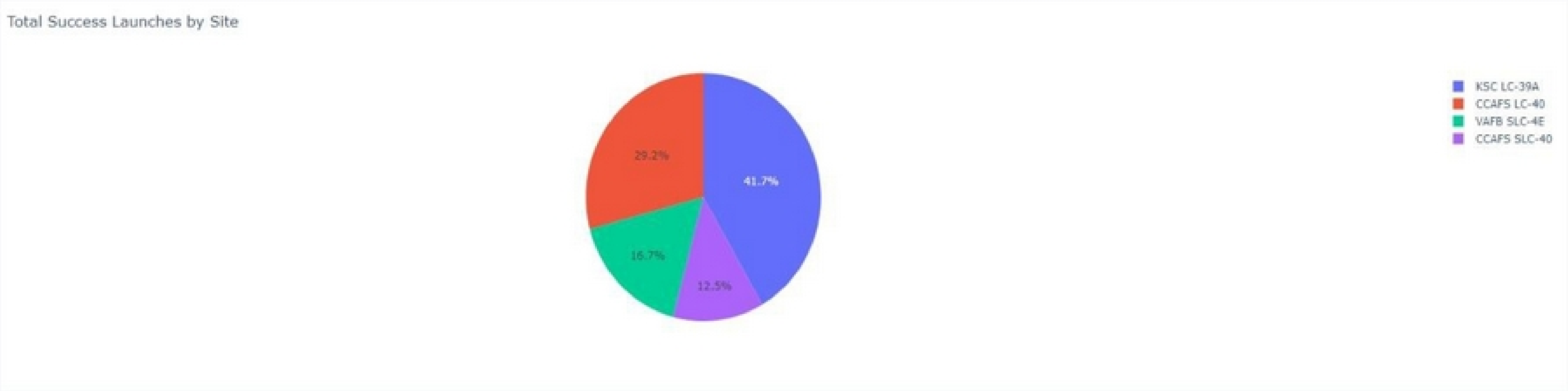
- Is CCAFS SLC-40 in close proximity to railways ? Yes
- Is CCAFS SLC-40 in close proximity to highways ? Yes
- Is CCAFS SLC-40 in close proximity to coastline ? Yes
- Do CCAFS SLC-40 keeps certain distance away from cities ? No



Section 4

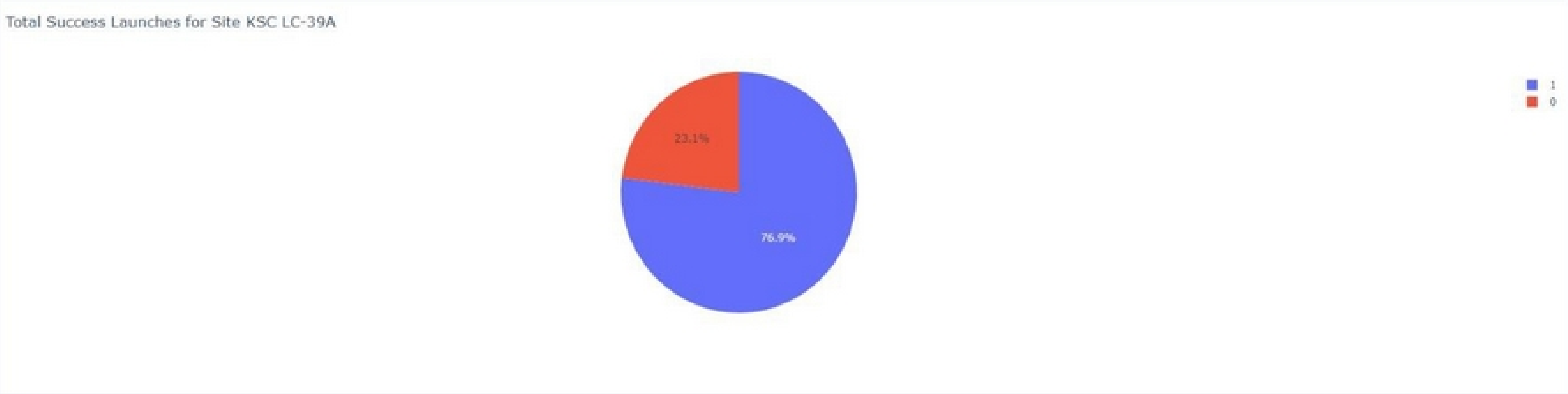
Build a Dashboard with Plotly Dash

Successful Launches by Site



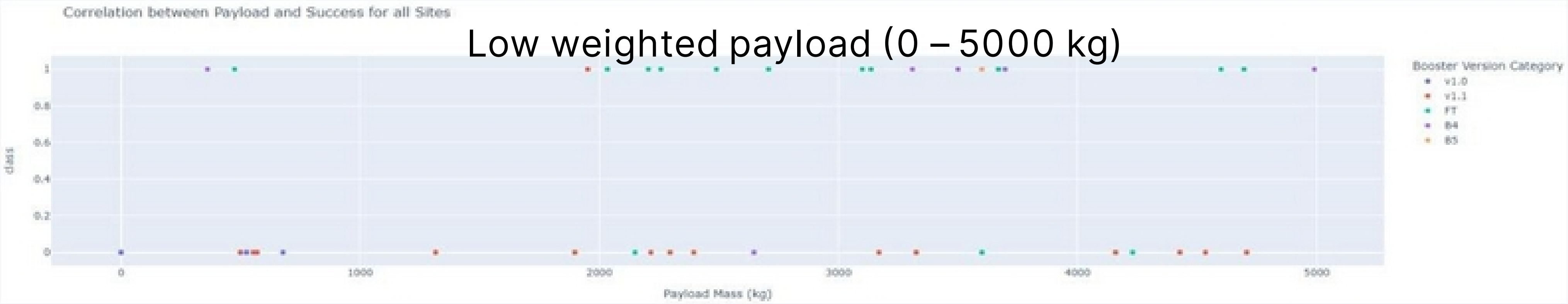
- We see that KSC LC-39A has the best success rate of launches.

Launch Success Ratio for KSC LC-39A

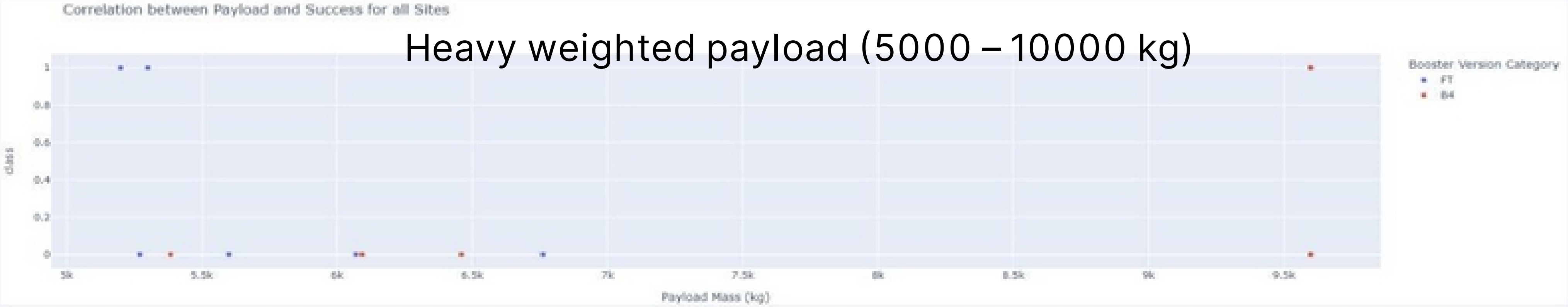


- We see that KSC LC-39A has achieved a 76.9% success rate while getting a 23.1% failure rate.

Payload vs. Launch Outcome



Low weighted payload (0 – 5000 kg)



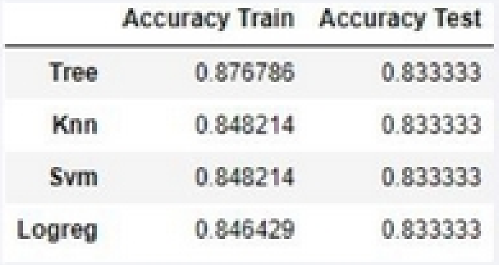
Heavy weighted payload (5000 – 10000 kg)

- Low weighted payloads have a better success rate than the heavy weighted payloads.

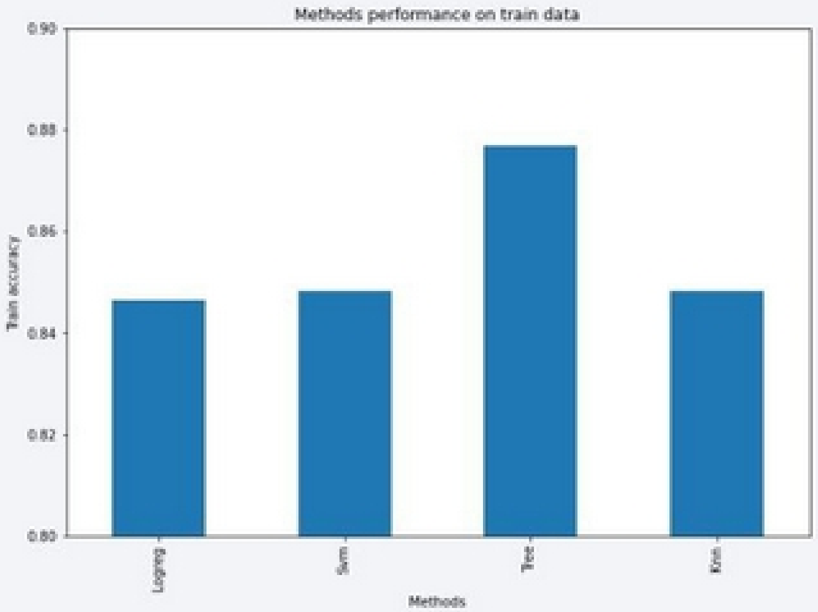
Section 5

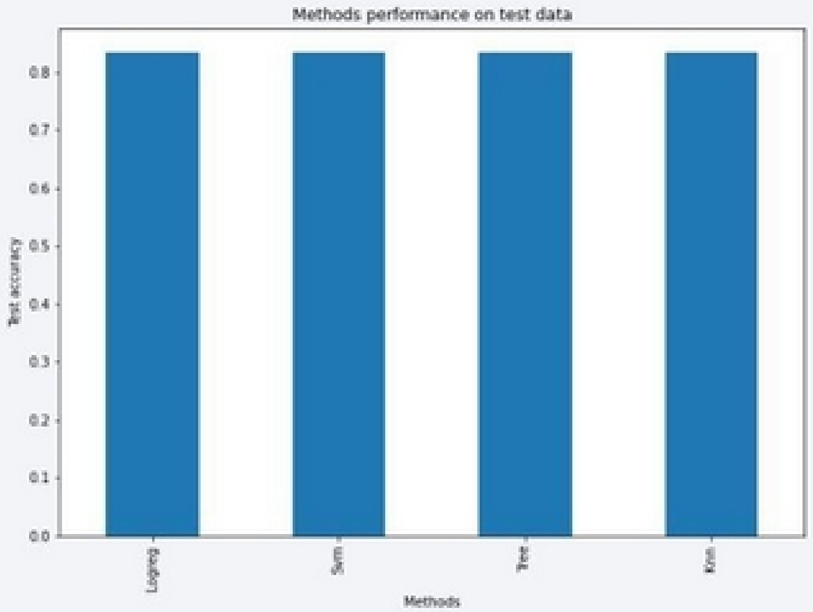
Predictive Analysis (Classification)

Classification Accuracy



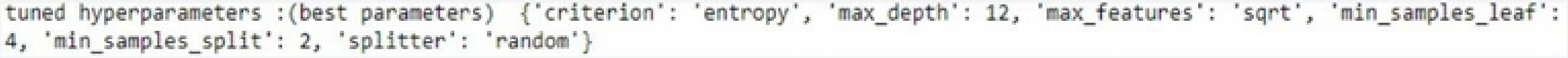
	Accuracy Train	Accuracy Test
Tree	0.876786	0.833333
Knn	0.848214	0.833333
Svm	0.848214	0.833333
Logreg	0.846429	0.833333





For accuracy test, all models performed almost similar. We could get more test data to decide between them. If we really need to choose one right now, we can take the Decision Tree.

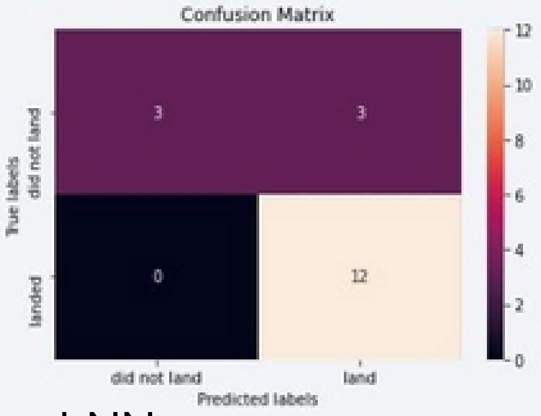
Decision tree best parameters



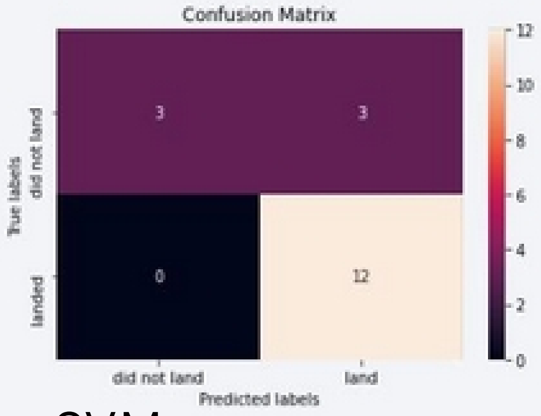
```
tuned hyperparameters :(best parameters)  {'criterion': 'entropy', 'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf':
4, 'min_samples_split': 2, 'splitter': 'random'}
```

Confusion Matrix

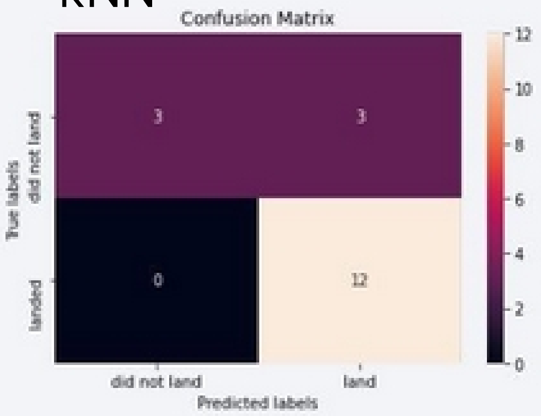
Logistic regression



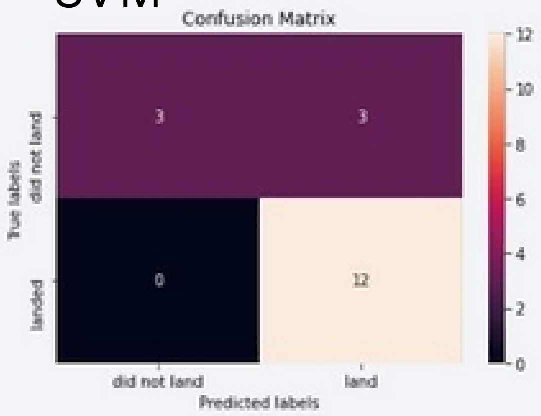
Decision Tree



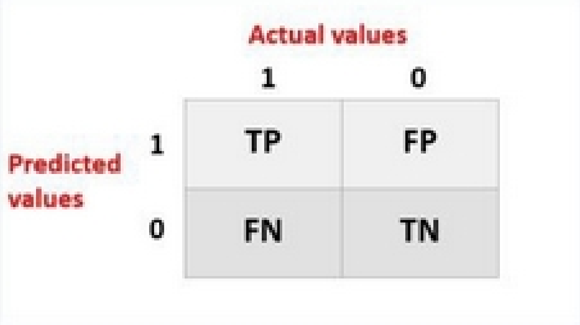
kNN



SVM



As the test accuracy are all equal, the confusion matrices are also identical. The main problem of these models are false positives.



		Actual values	
		1	0
Predicted values	1	TP	FP
	0	FN	TN

Conclusions

- The success of a mission can be explained by several factors such as the launch site, the orbit and especially the number of previous launches. Indeed, we can assume that there has been a gain in knowledge between launches that allowed to go from a launch failure to a success.
- The orbits with the best success rates are GEO, HEO, SSO, ES-L1.
- Depending on the orbits, the payload mass can be a criterion to take into account for the success of a mission. Some orbits require a light or heavy payload mass. But generally low weighted payloads perform better than the heavy weighted payloads.
- With the current data, we cannot explain why some launch sites are better than others (KSC LC-39A is the best launch site). To get an answer to this problem, we could obtain atmospheric or other relevant data.
- For this dataset, we choose the Decision Tree Algorithm as the best model even if the test accuracy between all the models used is identical. We choose Decision Tree Algorithm because it has a better train accuracy.

Thank you!