IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

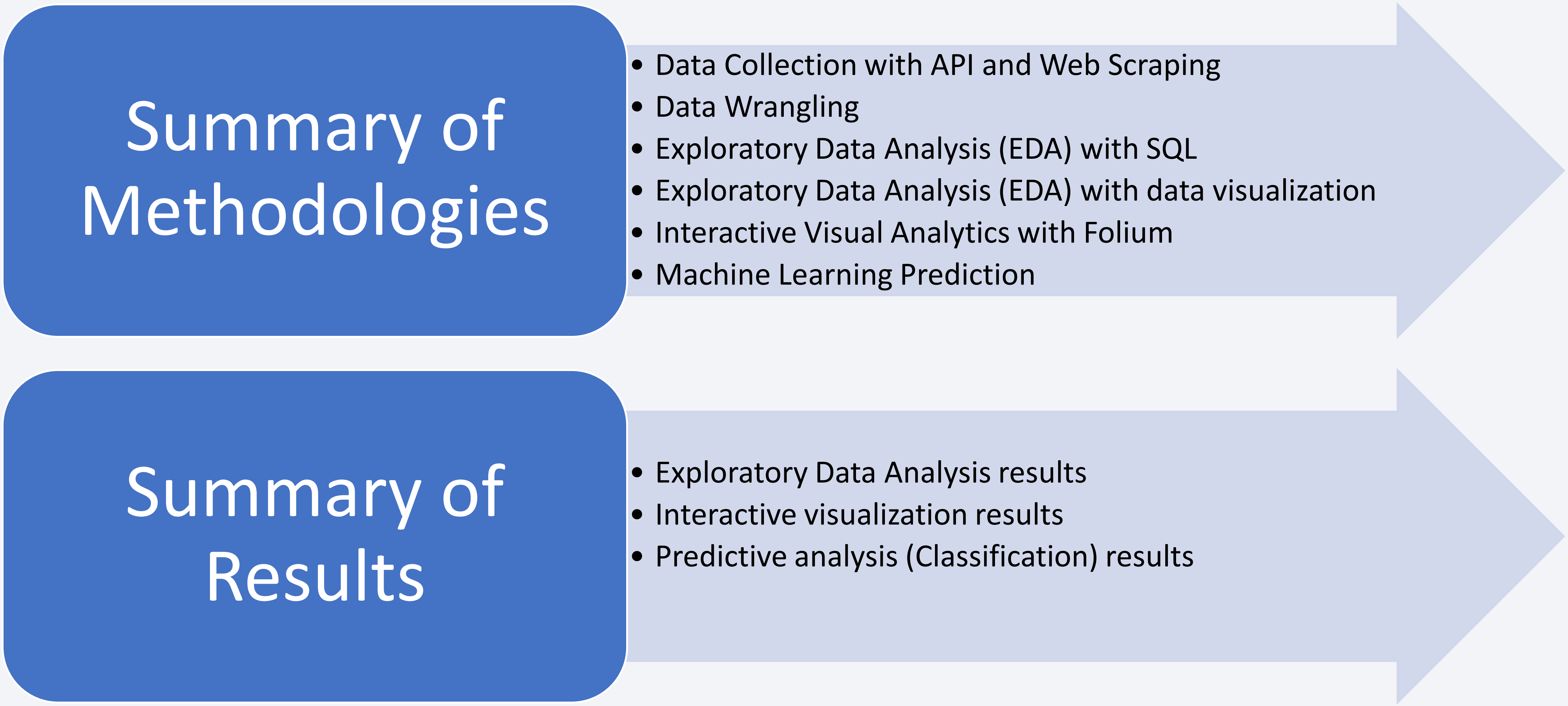
Alberto Segura Sanchez

October 3rd, 2021

Outline

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

Executive Summary



Summary of Methodologies

- Data Collection with API and Web Scraping
- Data Wrangling
- Exploratory Data Analysis (EDA) with SQL
- Exploratory Data Analysis (EDA) with data visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction

Summary of Results

- Exploratory Data Analysis results
- Interactive visualization results
- Predictive analysis (Classification) results

Introduction

Background

The commercial space age is here, companies are making space travel affordable for everyone. SpaceX is the most successful and its accomplishments include: Sending spacecraft to the International Space Station. One reason SpaceX can do this is the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if 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.

Problems we want solve

- Identify the variables that could impact the result of the missions
- First Stage Landing Prediction based on the available data

Section 1

Methodology

Methodology

Executive Summary

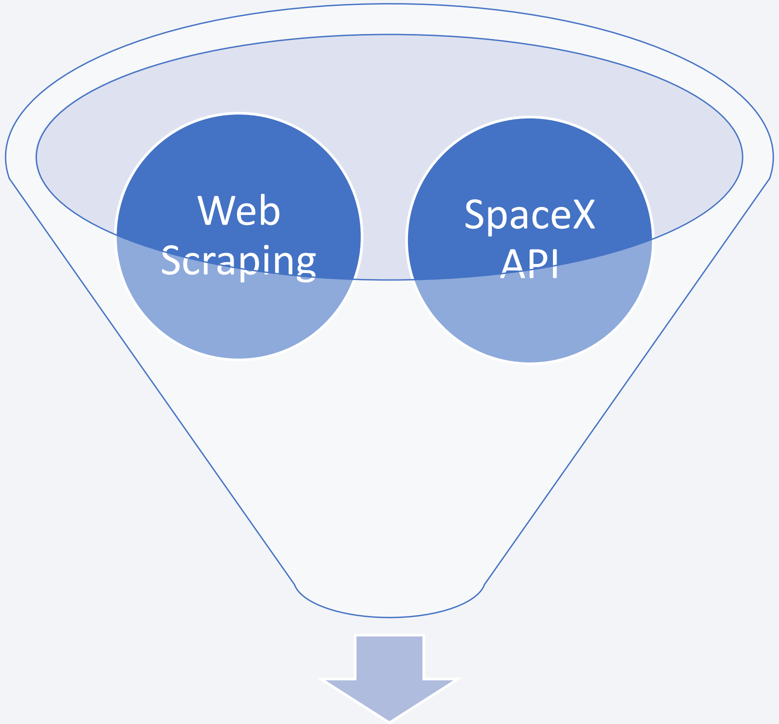
- Data collection methodology:
  - Using SpaceX REST API
  - Doing Web Scraping to Wikipedia pages
- Perform data wrangling:
  - Cleaning data dealing with missing values and irrelevant columns
  - Encoding categorical features into numerical values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

During data collection phase we identified the data sources that we can use as a reference to solve the problem.

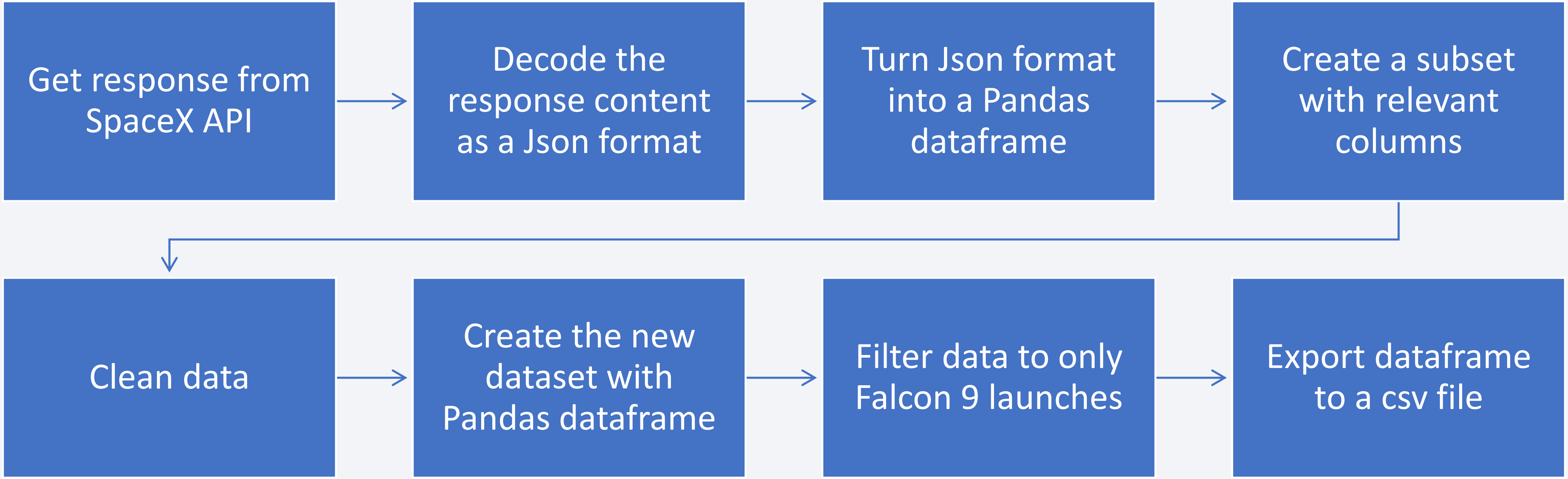
For this project two options were used for data collection: SpaceX API and Web Scraping.

This new data set contain information about each flight and its results.



Data Set for Data Wrangling and EDA

Data Collection – SpaceX API



Get response from SpaceX API

Decode the response content as a Json format

Turn Json format into a Pandas dataframe

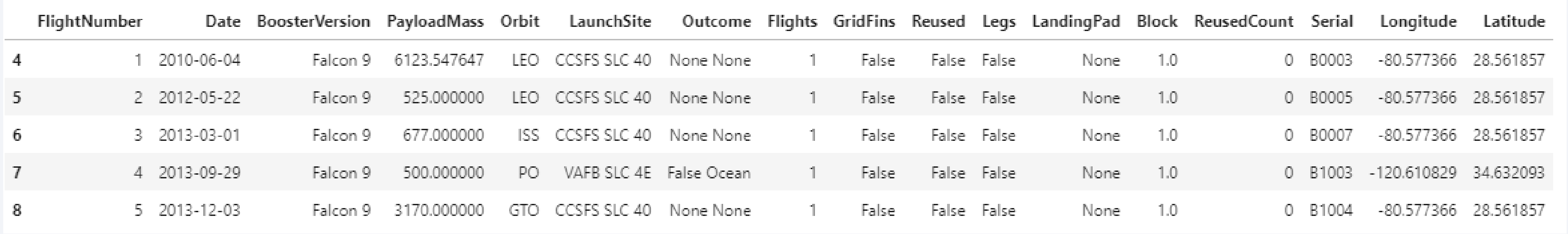
Create a subset with relevant columns

Clean data

Create the new dataset with Pandas dataframe

Filter data to only Falcon 9 launches

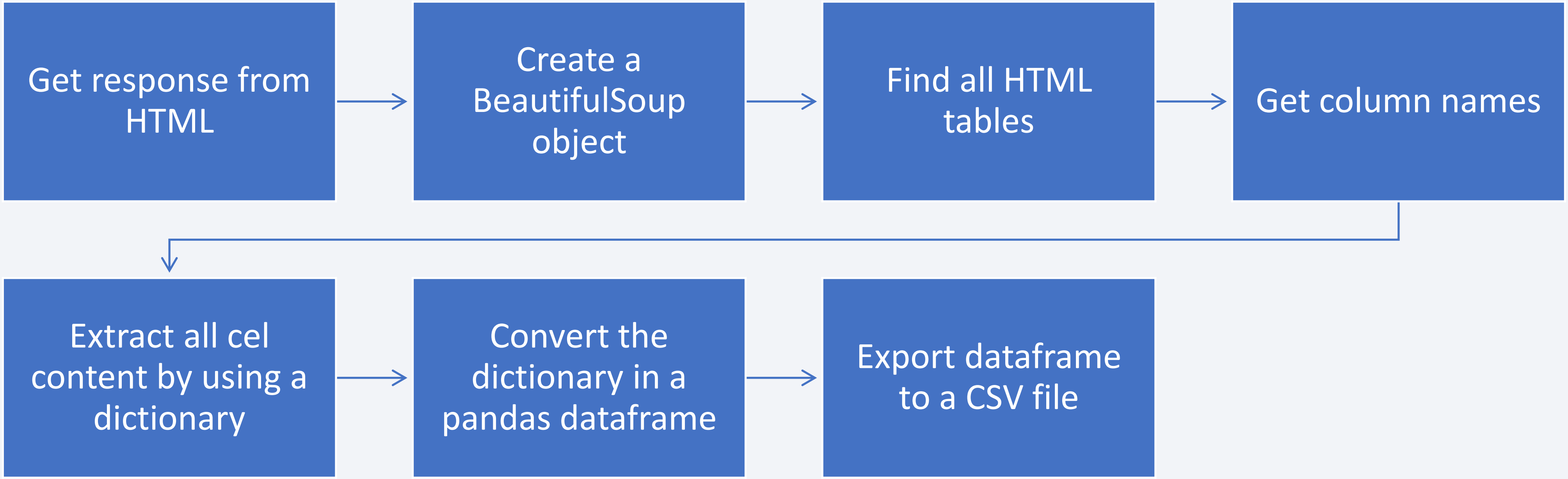
Export dataframe to a csv file



	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	6123.547647	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.000000	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.000000	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.000000	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

Data Collection SpaceX API - Github

Data Collection - Scraping



Get response from HTML

Create a BeautifulSoup object

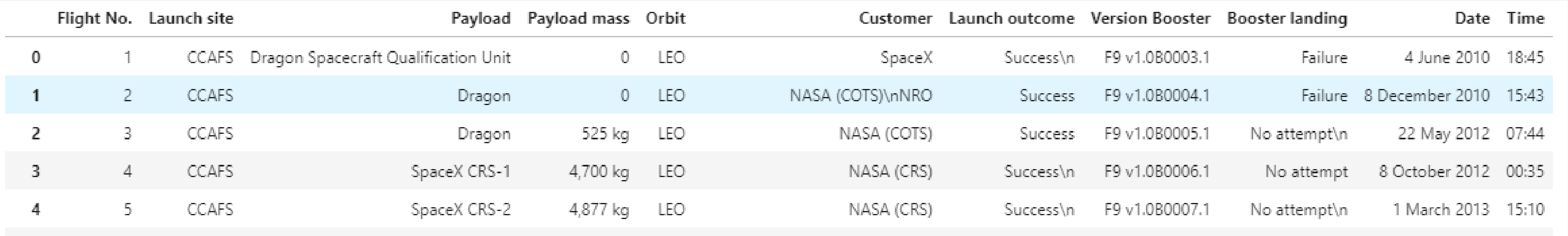
Find all HTML tables

Get column names

Extract all cel content by using a dictionary

Convert the dictionary in a pandas dataframe

Export dataframe to a CSV file

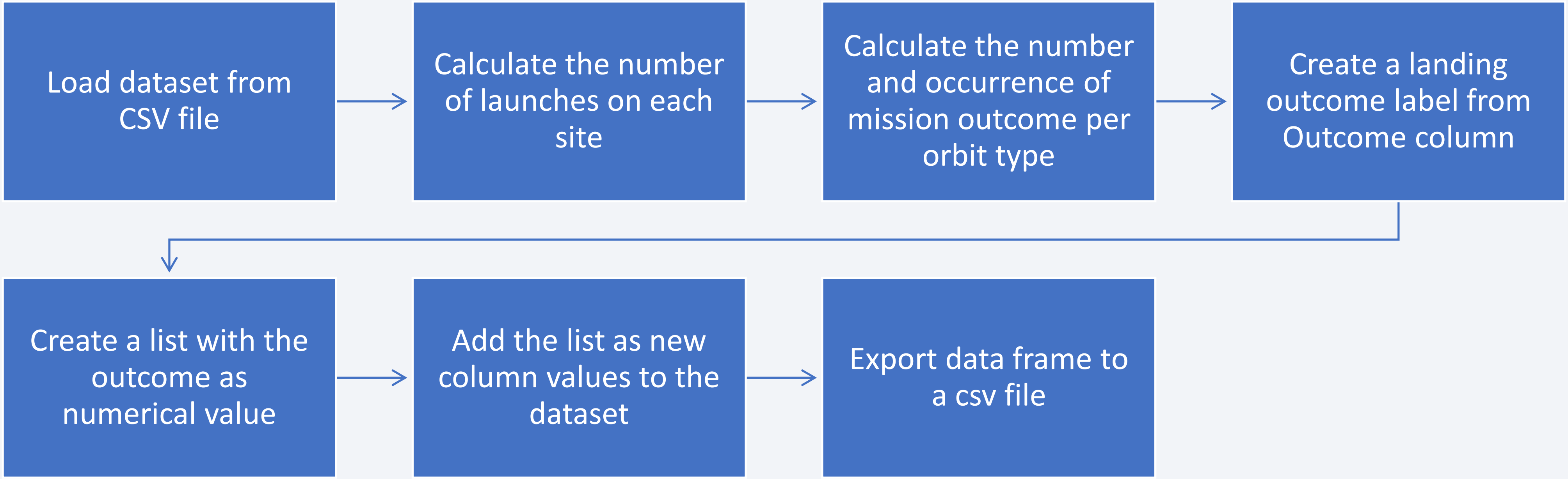


	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA (COTS)\nNRO	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA (COTS)	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Data Collection Web Scraping - Github

Data Wrangling

In data wrangling phase it was performed exploratory data analysis and determined training labels



Load dataset from CSV file

Calculate the number of launches on each site

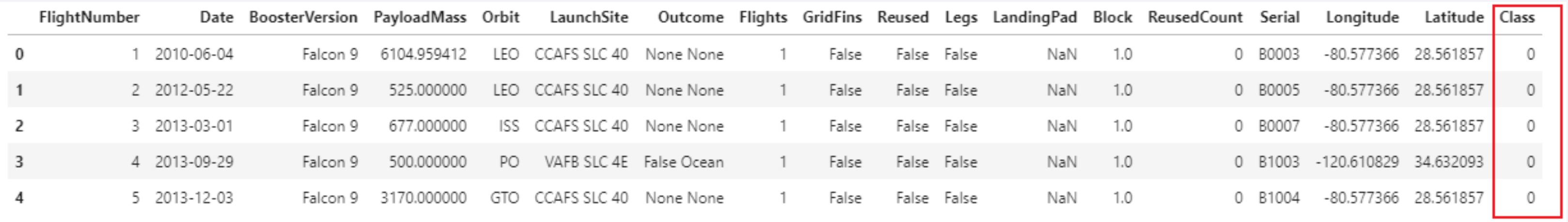
Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column

Create a list with the outcome as numerical value

Add the list as new column values to the dataset

Export data frame to a csv file

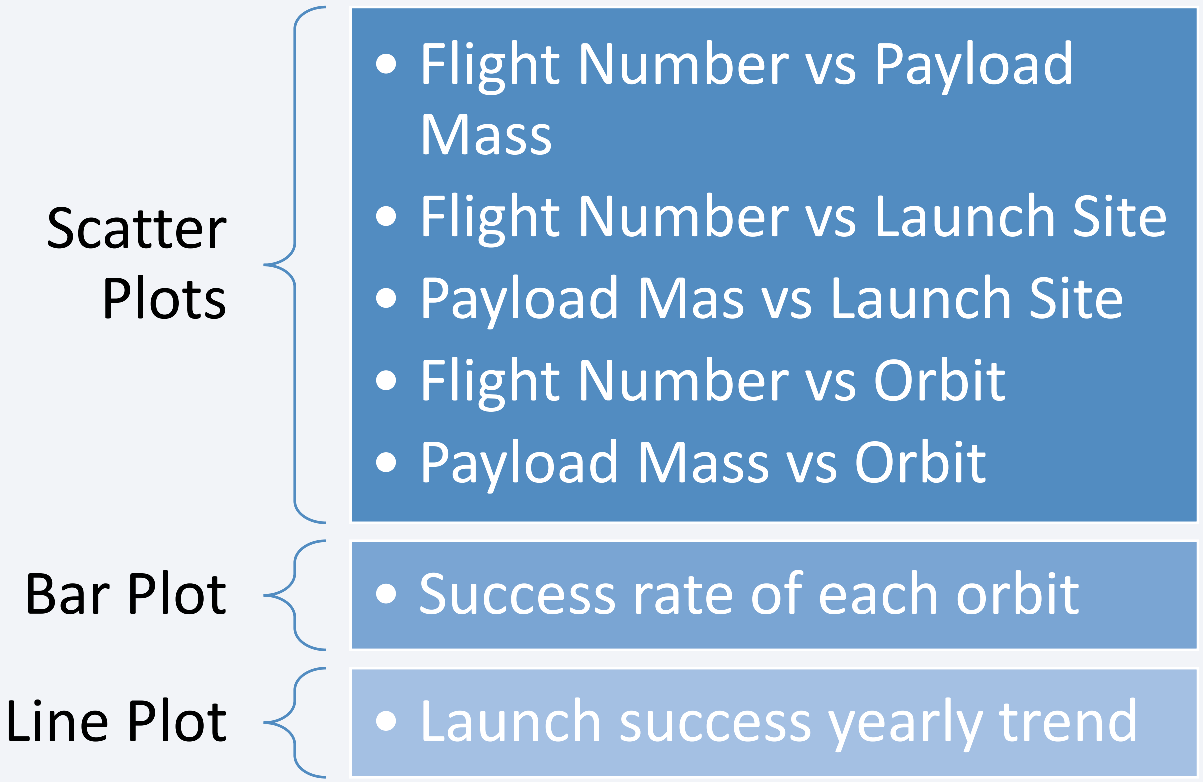


	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

Data Wrangling - Github

EDA with Data Visualization

Data visualization was done using tree different type of plots as shown below:



Scatter Plots

- Flight Number vs Payload Mass
- Flight Number vs Launch Site
- Payload Mas vs Launch Site
- Flight Number vs Orbit
- Payload Mass vs Orbit

Scatter plots were used to determine the correlation between two parameters, how one variable is affected by the change on the other one.

Bar Plot

- Success rate of each orbit

Bar plot was used to visualize the success of the mission grouped by the orbit.

Line Plot

- Launch success yearly trend

Line plot was used to visualize the success rate has change over the years.

[EDA with Data Visualization - Github]

EDA with SQL

SQL Queries

- 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. Use a subquery
- 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 (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

EDA with SQL - Github

Build an Interactive Map with Folium

Using Folium we can visualize geo-locations in an interactive map. For this project we used Folium to determine if the outcome of a mission depend on the location and proximities of a launch site.

Folium is composed by different type of objects and for this exercise we are using the following ones:

- Circle: Used to identify the location of the launch sites based on their coordinates
- Marker: Used to add a label at specific coordinates
- MarkerCluster: Used to simplify a map containing many markers having the same coordinate by grouping them
- MousePosition: Used to identify the specific coordinates of points of interests in the map
- PolyLine: Used to represent the distance between two or more coordinates in the map

Interactive Visual Analytics with Folium- Github

Build a Dashboard with Plotly Dash

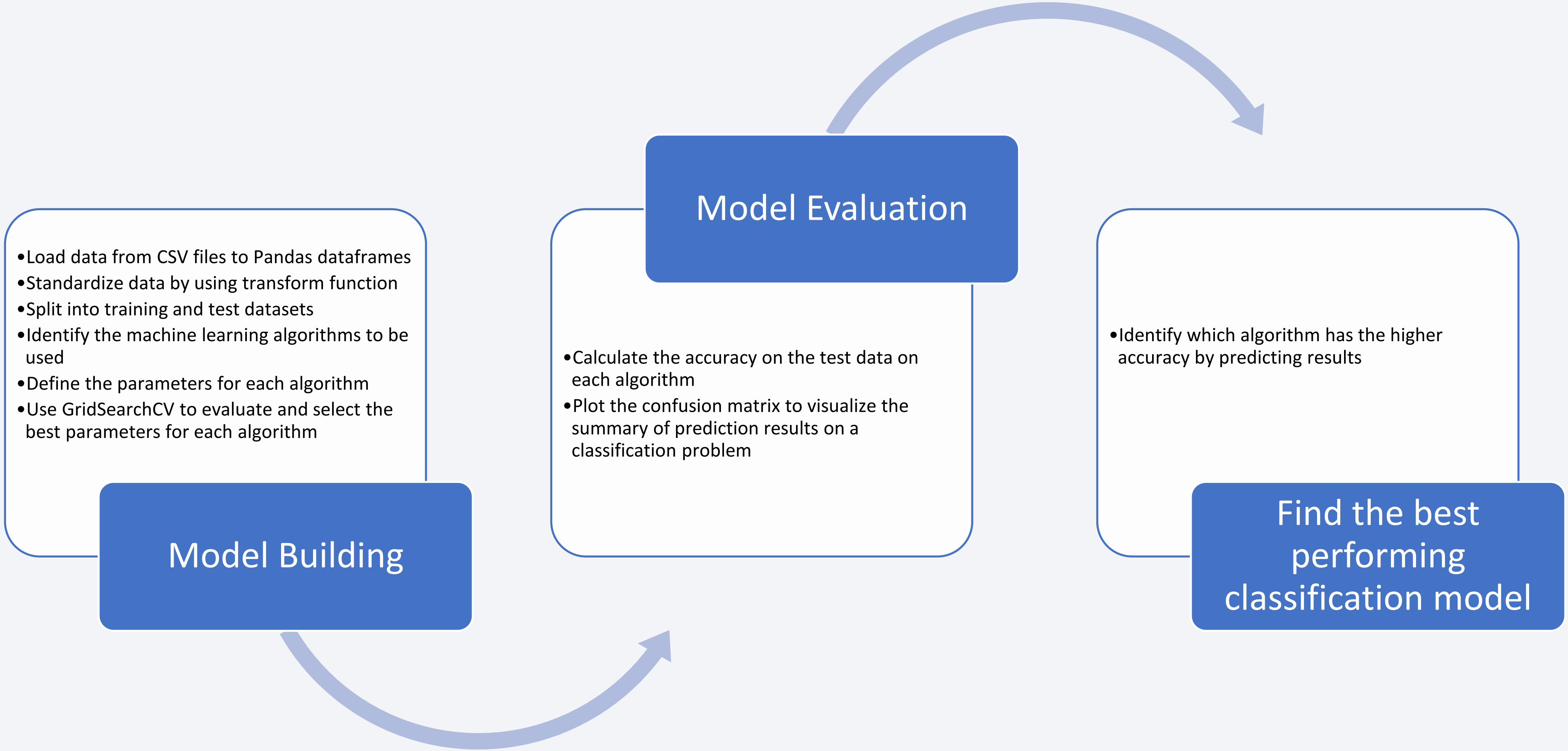
We are using Plotly Dash to create an interactive dashboard where it is possible to plot the interaction of some variables.

Here are the graphs and interactions included in the dashboard:

- Dropdown list: Enable Launch Site selection
- Pie Chart: To visualize the total successful launches count for all sites, if a specific launch site was selected, show the Success vs. Failed counts for the site
- Slider: To select payload range
- Scatter chart: To show the correlation between payload and launch success

Interactive Dashboard with Ploty Dash - Github

Predictive Analysis (Classification)



Model Building

- Load data from CSV files to Pandas dataframes
- Standardize data by using transform function
- Split into training and test datasets
- Identify the machine learning algorithms to be used
- Define the parameters for each algorithm
- Use GridSearchCV to evaluate and select the best parameters for each algorithm

Model Evaluation

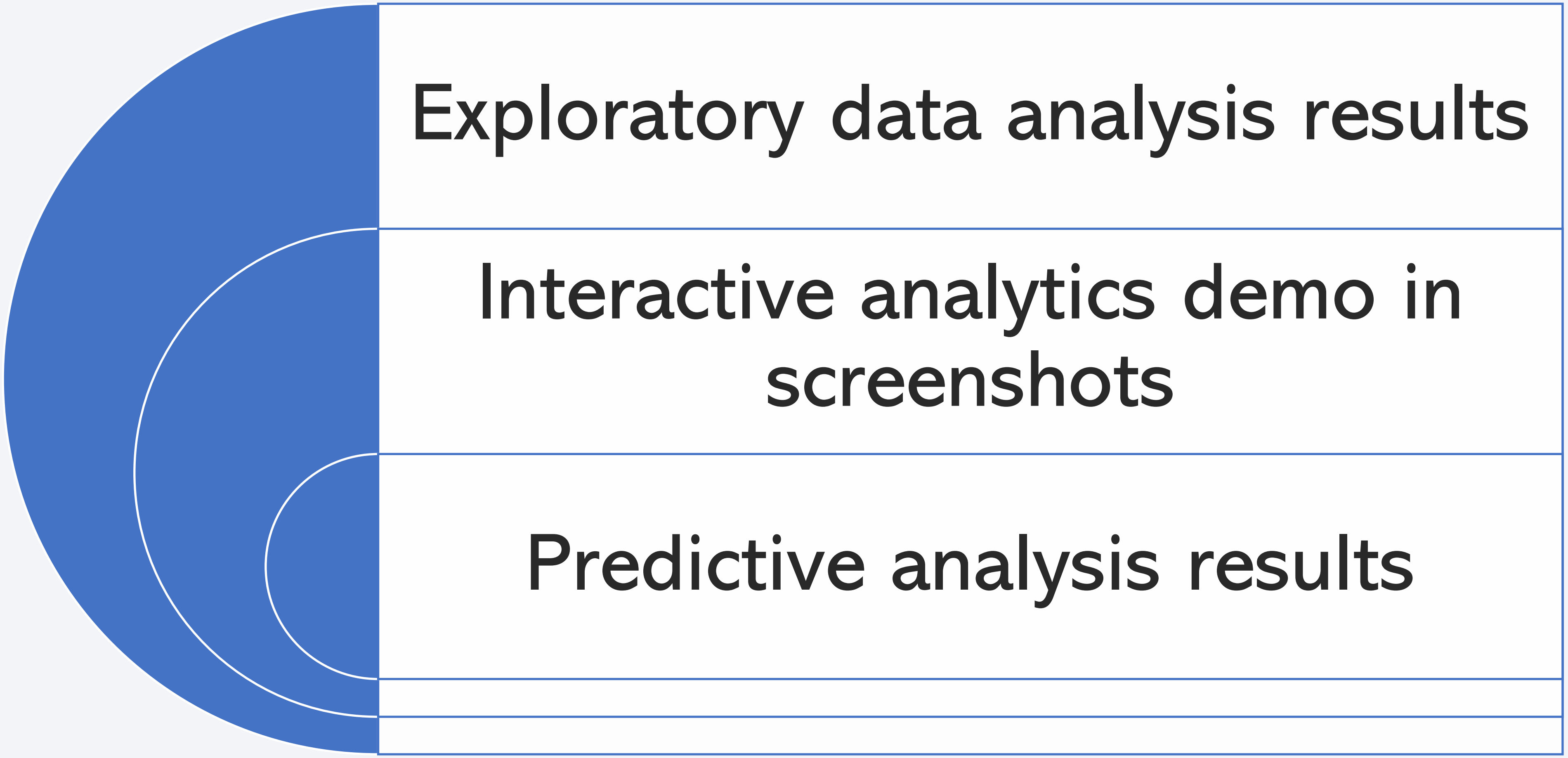
- Calculate the accuracy on the test data on each algorithm
- Plot the confusion matrix to visualize the summary of prediction results on a classification problem

Find the best performing classification model

- Identify which algorithm has the higher accuracy by predicting results

Machine Learning Prediction - Github

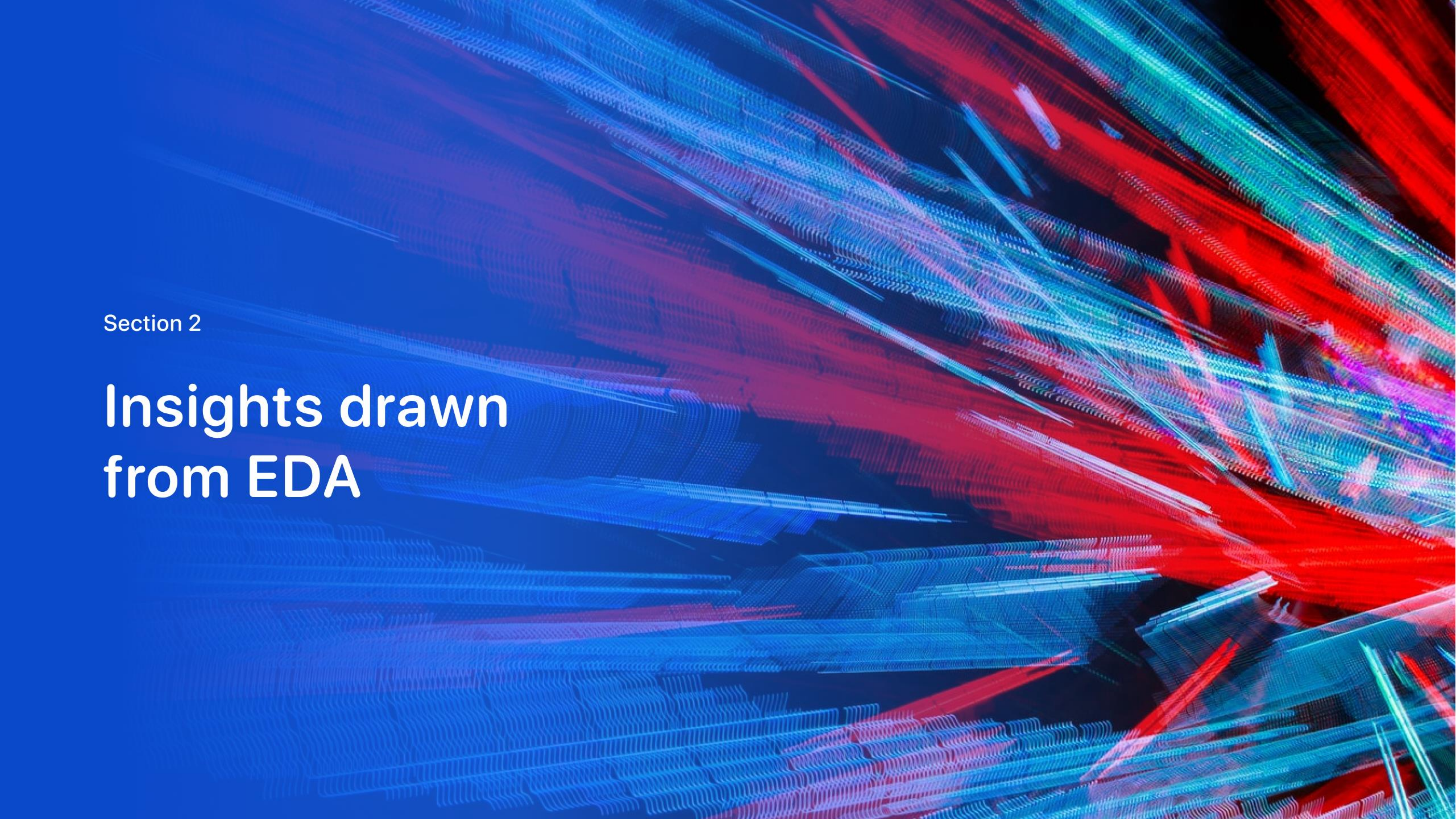
Results



Exploratory data analysis results

Interactive analytics demo in screenshots

Predictive analysis results

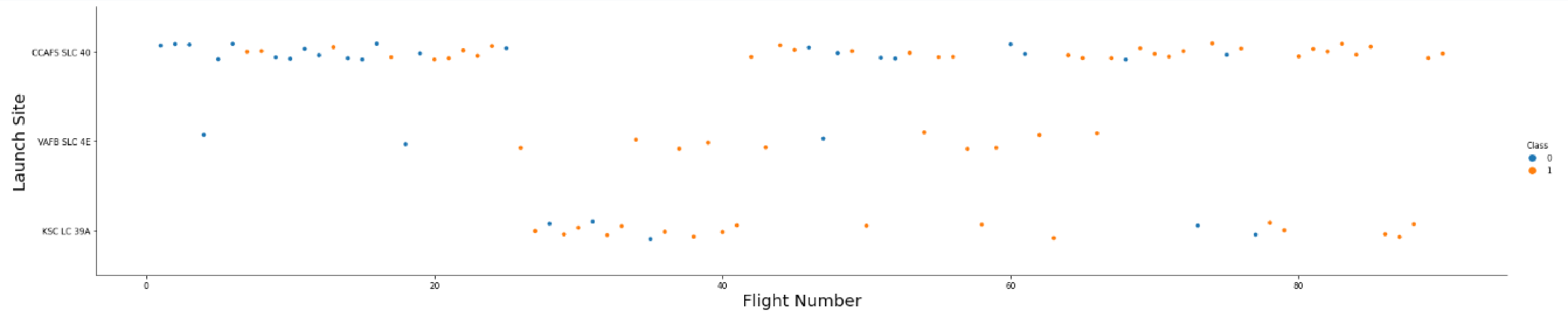


Section 2

Insights drawn from EDA

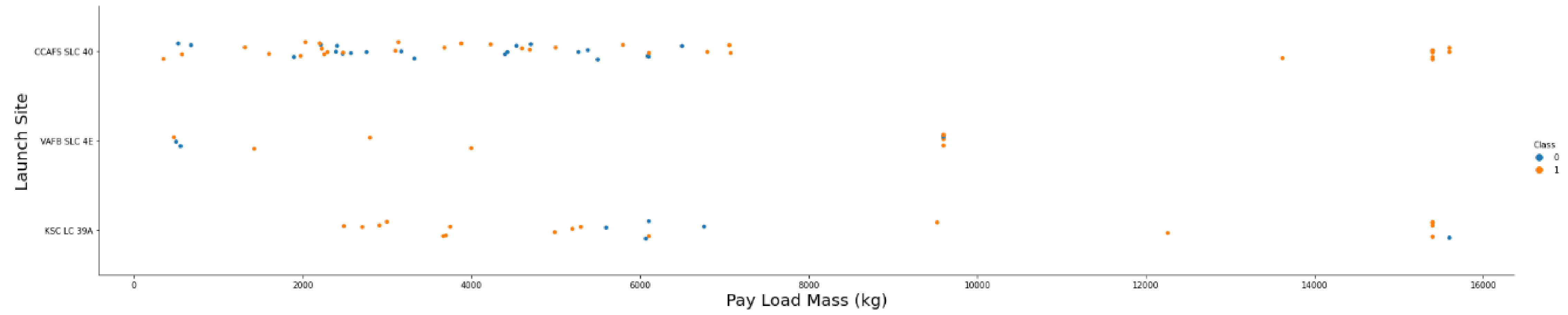
Flight Number vs. Launch Site

We can see in the plot that initially the flights were originated almost in a single site with more fails than successes. The success rate has increased as a greater number of flights and different launch sites.



Payload vs. Launch Site

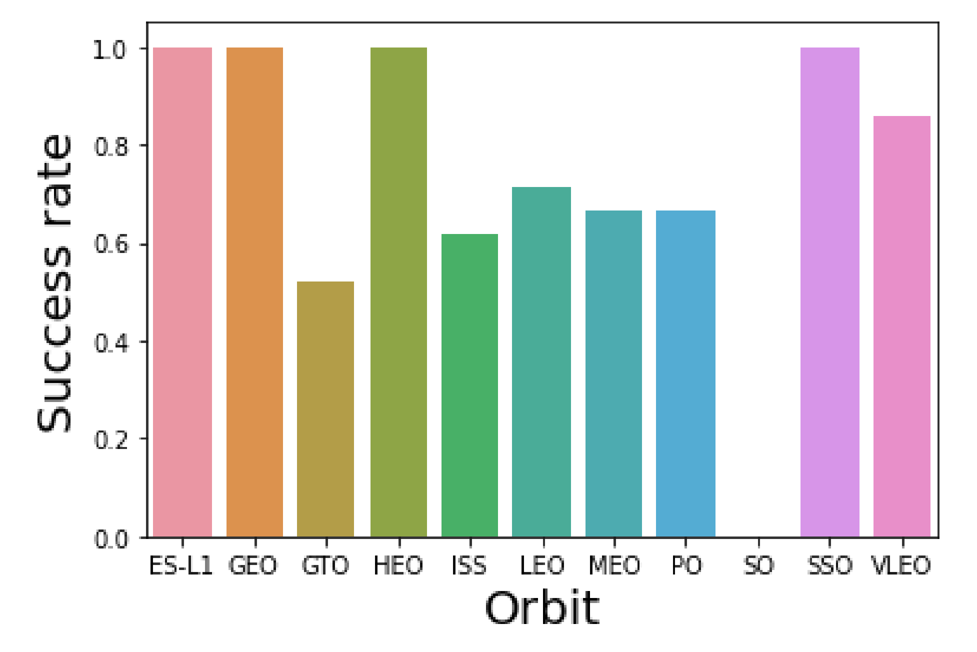
The following plot shows the relation between payload mass and the launch site. The success rate is good with higher payload mass but in the lower rage of payload mass we can see mixed results of outcome, so it is not clear if de success depends on the launch site.



Success Rate vs. Orbit Type

In the bar chart we can see that there are orbits with a high success rate like ES-L1, GEO, HEO and SSO.

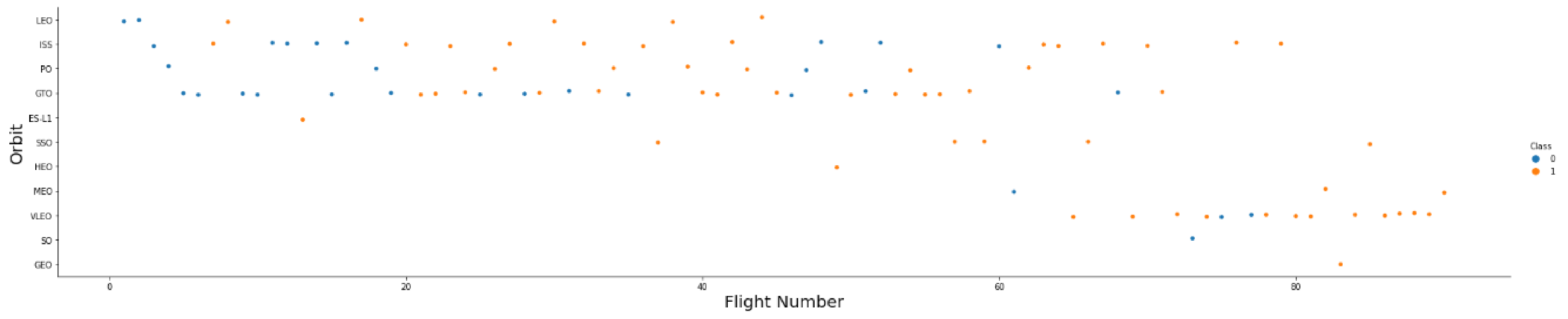
Based on that there is no relation of the distance of the orbit with the success rate.



Flight Number vs. Orbit Type

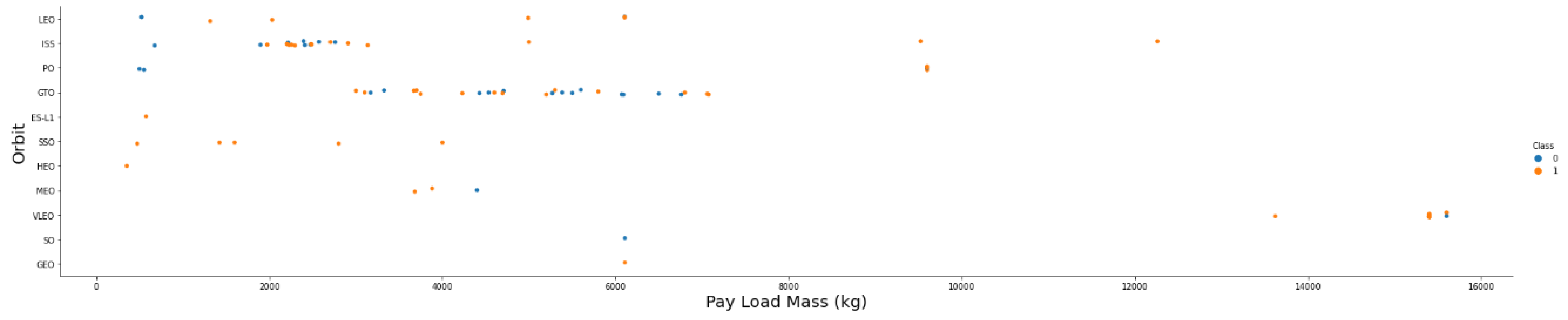
In the scatter plot we can see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

The high success rate for some orbit seems to be related to the low quantity of flights.



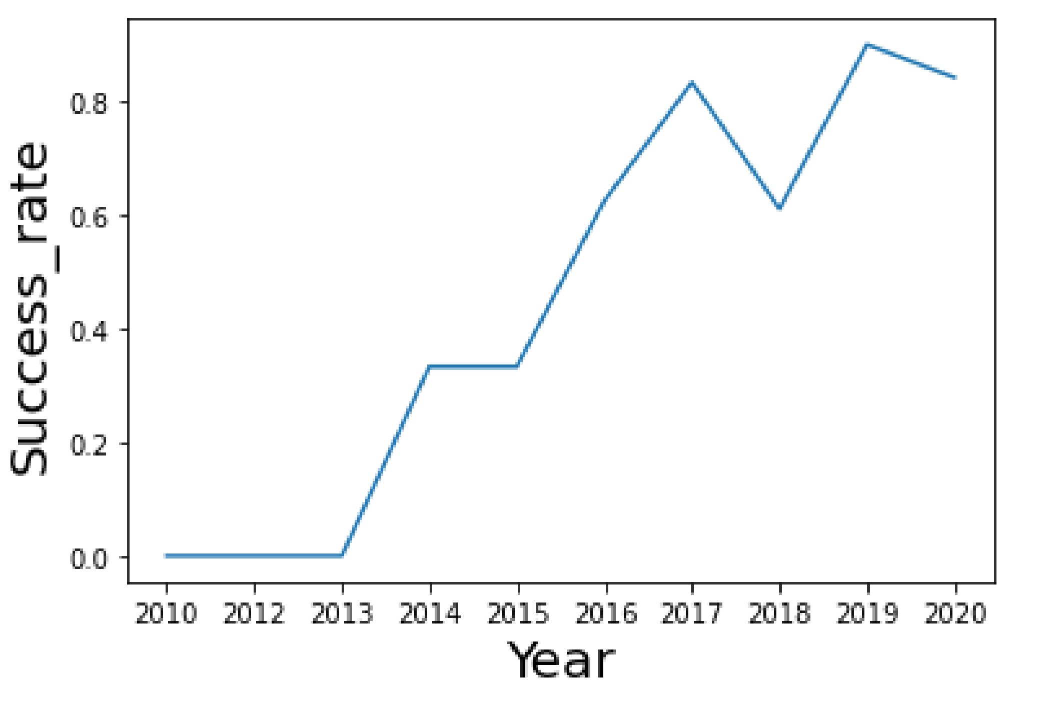
Payload vs. Orbit Type

Heavy payloads have a negative influence on GTO orbits and positive on LEO and Polar LEO (ISS) orbits.



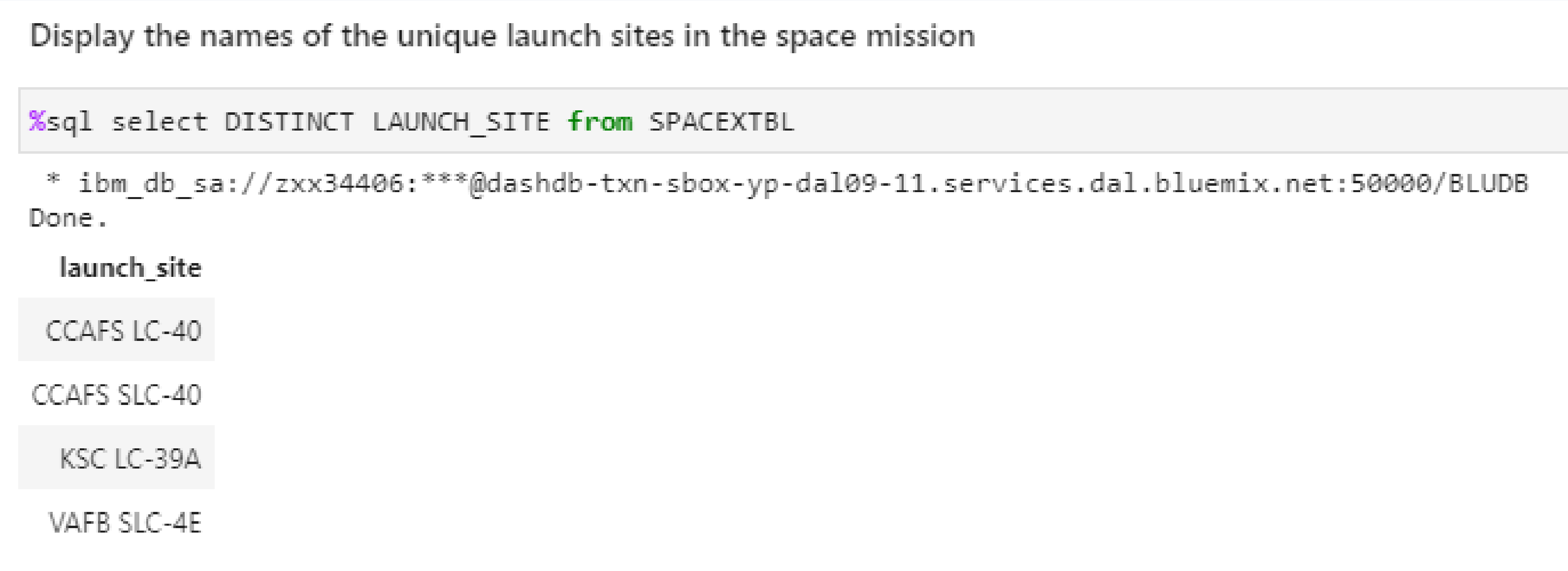
Launch Success Yearly Trend

In the line plot we can see that the success rate since 2013 kept increasing till 2020.



All Launch Site Names

With the following SQL query we can identify 4 different launch sites. DISTINCT in the SQL query allows to get unique values.



Display the names of the unique launch sites in the space mission

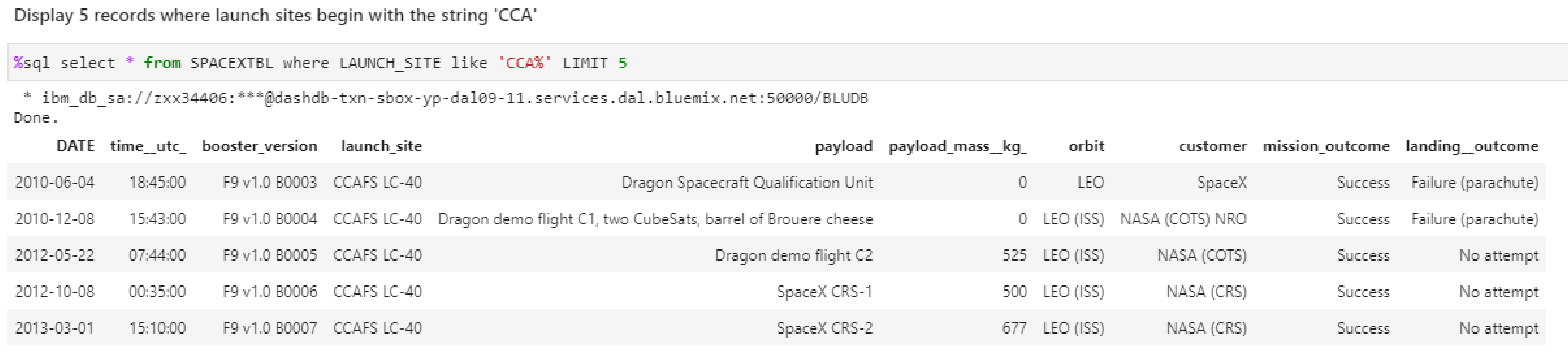
```
%sql select DISTINCT LAUNCH_SITE from SPACEXTBL
```

```
 * ibm_db_sa://zxx34406:***@dashdb-txn-sbox-yp-dal09-11.services.dal.bluemix.net:50000/BLUDB
Done.
```

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

Launch Site Names Begin with 'CCA'

SQL query bellow gets the first 5 records where launch sites begin with `CCA`.



Display 5 records where launch sites begin with the string 'CCA'

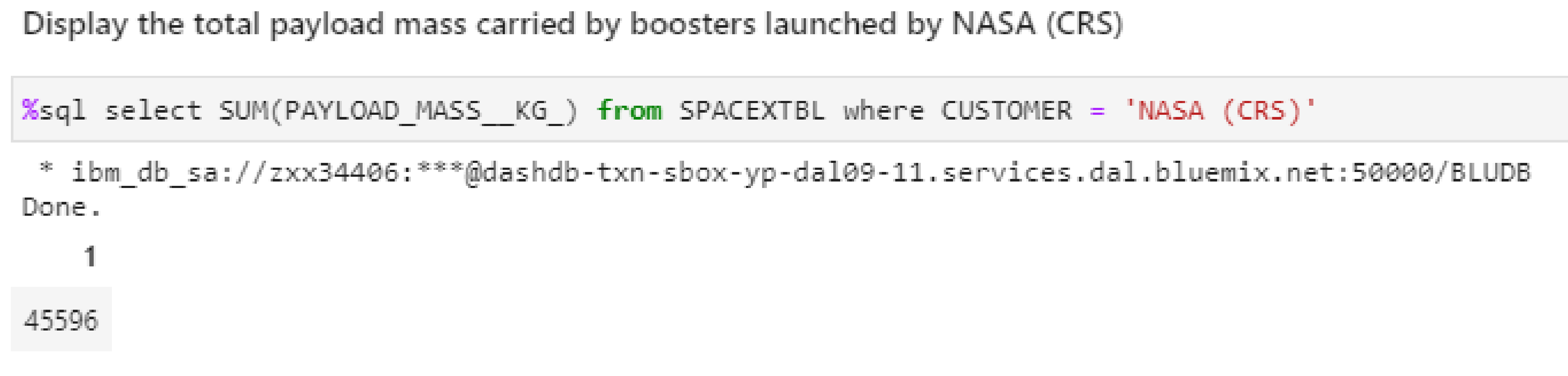
```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://zxx34406:***@dashdb-txn-sbox-yp-dal09-11.services.dal.bluemix.net:50000/BLUDB
Done.
```

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

Usim SUM function in the query allows to calculate the total payload mass for specific customer, in this case for NASA (CRS).



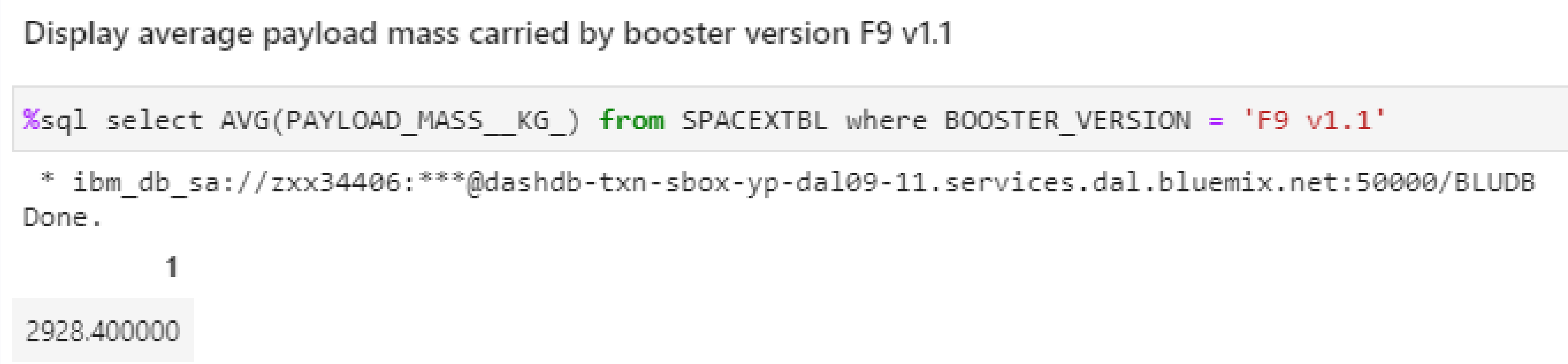
Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://zxx34406:***@dashdb-txn-sbox-yp-dal09-11.services.dal.bluemix.net:50000/BLUDB
Done.
    1
45596
```

Average Payload Mass by F9 v1.1

Following SQL query calculates the average payload mass carried by booster version F9 v1.1 using the AVG function.



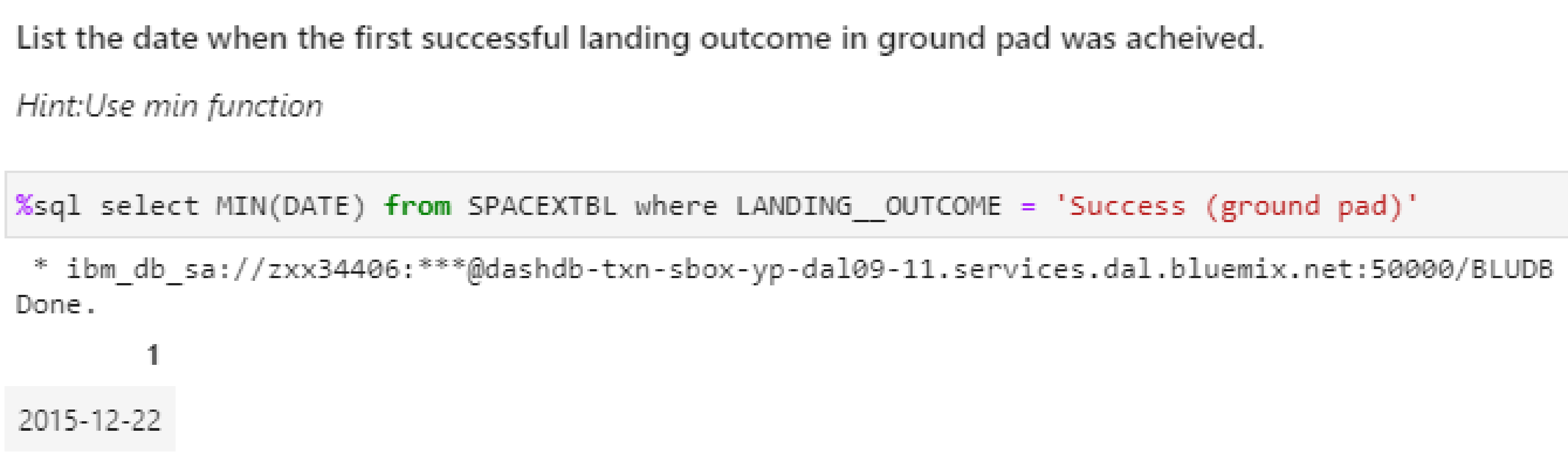
Display average payload mass carried by booster version F9 v1.1

```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

```
 * ibm_db_sa://zxx34406:***@dashdb-txn-sbox-yp-dal09-11.services.dal.bluemix.net:50000/BLUDB
Done.
          1
2928.400000
```

First Successful Ground Landing Date

With the use of MIN function, we can locate the date of the first successful landing outcome on ground pad.



List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

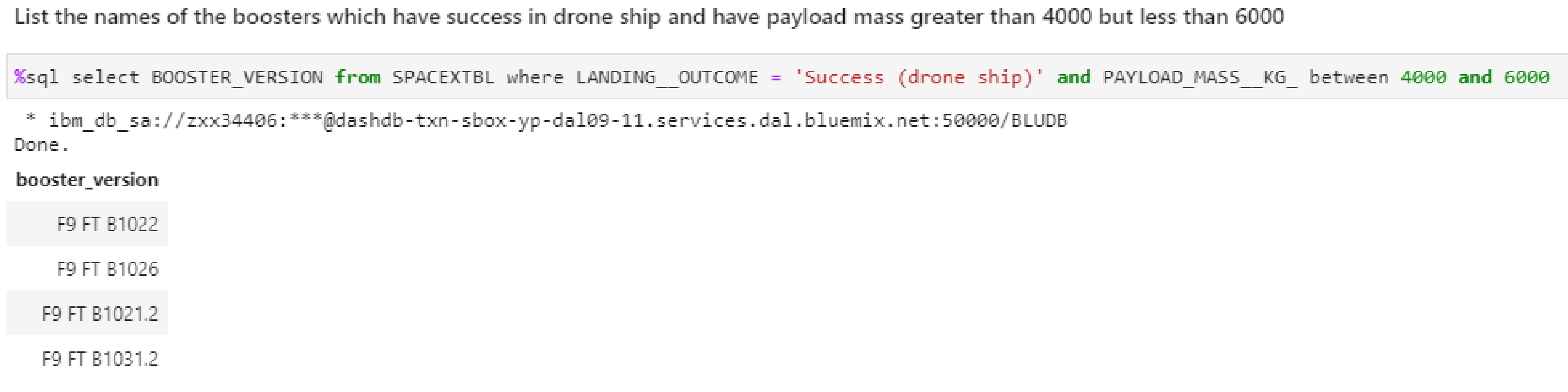
```
%sql select MIN(DATE) from SPACEXTBL where LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://zxx34406:***@dashdb-txn-sbox-yp-dal09-11.services.dal.bluemix.net:50000/BLUDB
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

In the following query we use two conditionals to select the specific booster version. Only when both conditionals are true the query can select a buster version.



List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

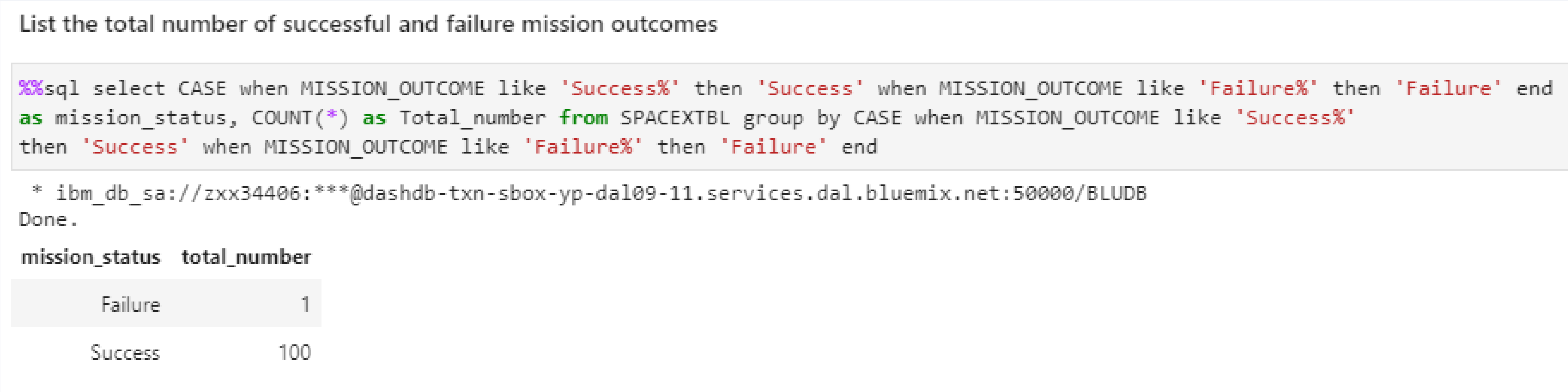
```
%sql select BOOSTER_VERSION from SPACEXTBL where LANDING__OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

```
 * ibm_db_sa://zxx34406:***@dashdb-txn-sbox-yp-dal09-11.services.dal.bluemix.net:50000/BLUDB
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Following SQL query calculates the total number of successful and failure mission outcomes.



List the total number of successful and failure mission outcomes

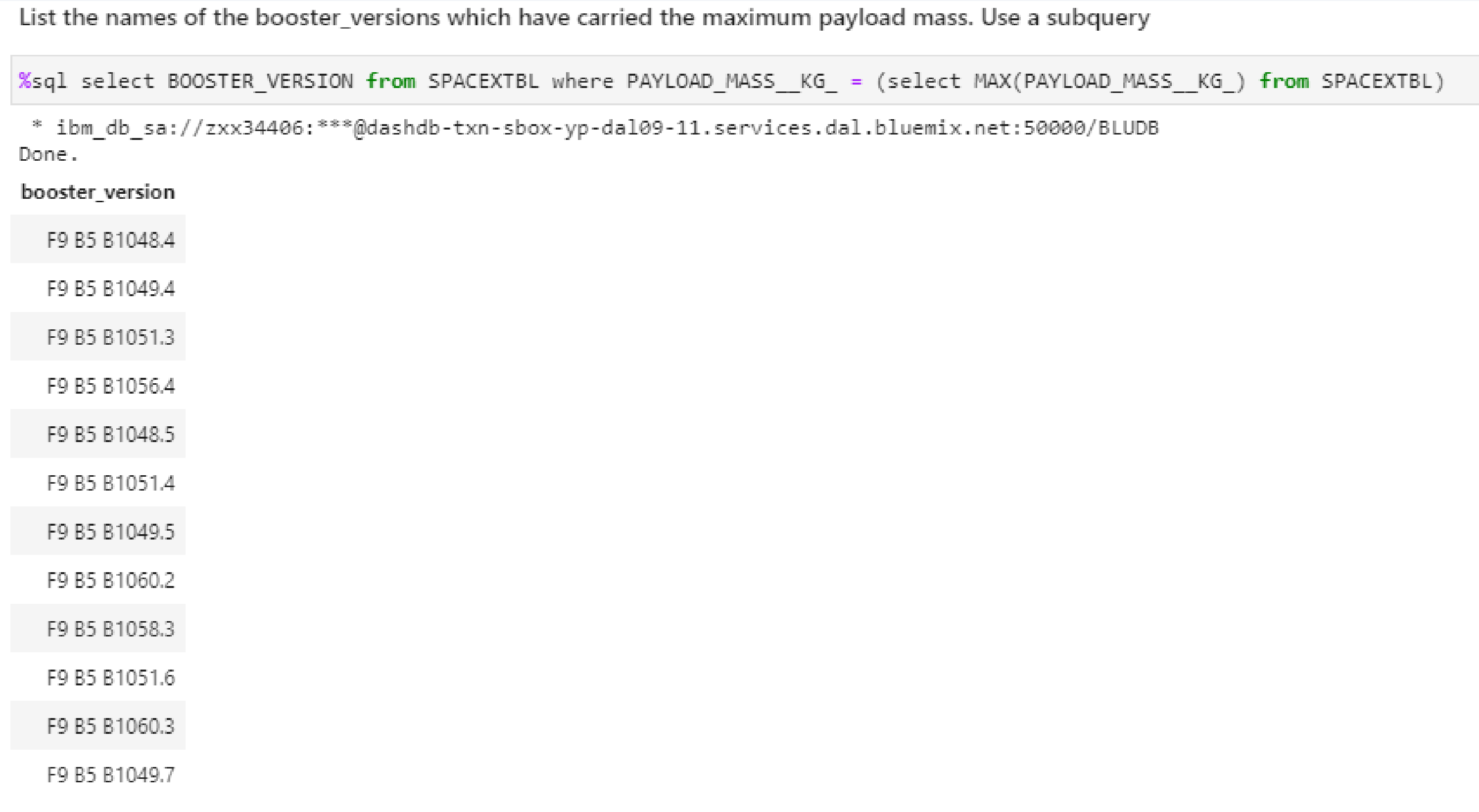
```
%%sql select CASE when MISSION_OUTCOME like 'Success%' then 'Success' when MISSION_OUTCOME like 'Failure%' then 'Failure' end
as mission_status, COUNT(*) as Total_number from SPACEXTBL group by CASE when MISSION_OUTCOME like 'Success%'
then 'Success' when MISSION_OUTCOME like 'Failure%' then 'Failure' end
```

```
 * ibm_db_sa://zxx34406:***@dashdb-txn-sbox-yp-dal09-11.services.dal.bluemix.net:50000/BLUDB
Done.
```

mission_status	total_number
Failure	1
Success	100

Boosters Carried Maximum Payload

With the use of a subquerie we can obtain the list the names of the booster which have carried the maximum payload mass.



List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

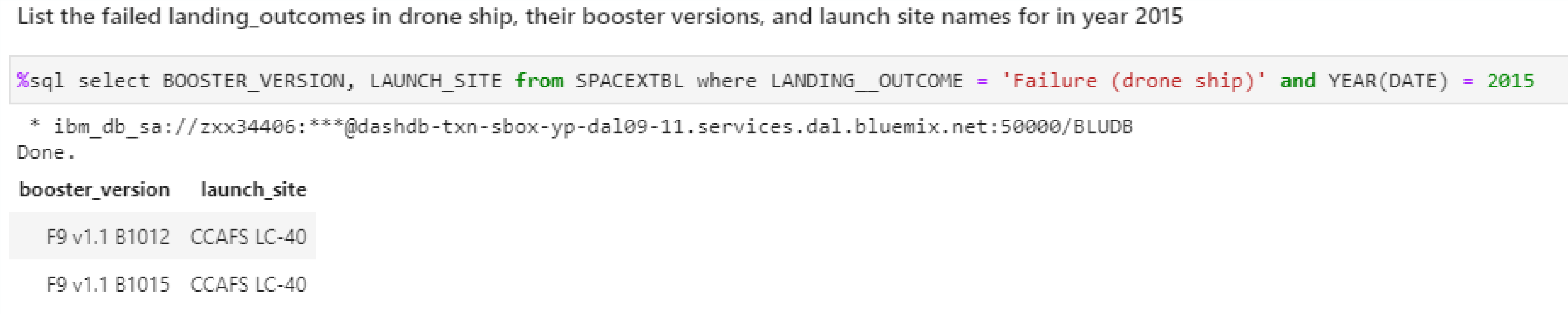
```
%sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

```
 * ibm_db_sa://zxx34406:***@dashdb-txn-sbox-yp-dal09-11.services.dal.bluemix.net:50000/BLUDB
Done.
```

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

By using conditionals, the following query lists the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015



List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

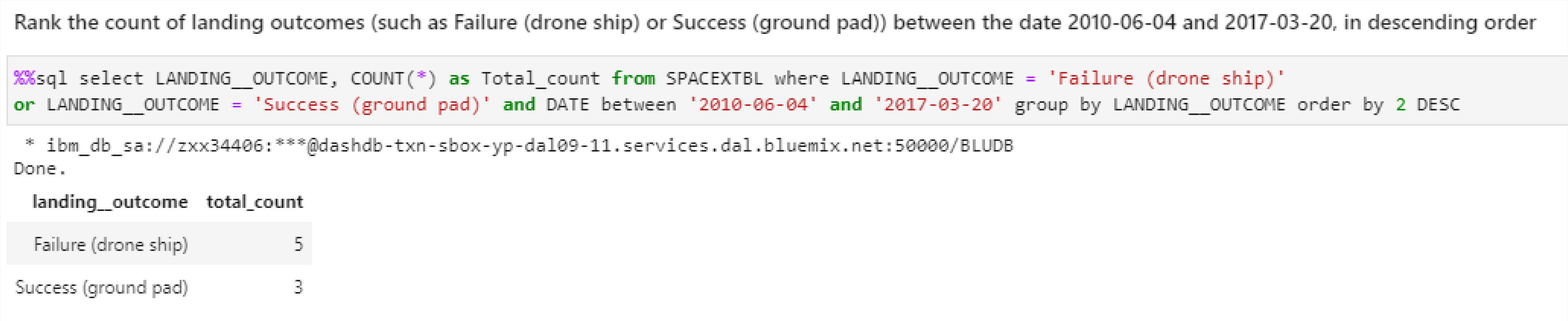
```
%sql select BOOSTER_VERSION, LAUNCH_SITE from SPACEXTBL where LANDING__OUTCOME = 'Failure (drone ship)' and YEAR(DATE) = 2015
```

```
 * ibm_db_sa://zxx34406:***@dashdb-txn-sbox-yp-dal09-11.services.dal.bluemix.net:50000/BLUDB
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

The following SQL query rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order



Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

```
%%sql select LANDING__OUTCOME, COUNT(*) as Total_count from SPACEXTBL where LANDING__OUTCOME = 'Failure (drone ship)'
or LANDING__OUTCOME = 'Success (ground pad)' and DATE between '2010-06-04' and '2017-03-20' group by LANDING__OUTCOME order by 2 DESC
```

```
 * ibm_db_sa://zxx34406:***@dashdb-txn-sbox-yp-dal09-11.services.dal.bluemix.net:50000/BLUDB
Done.
```

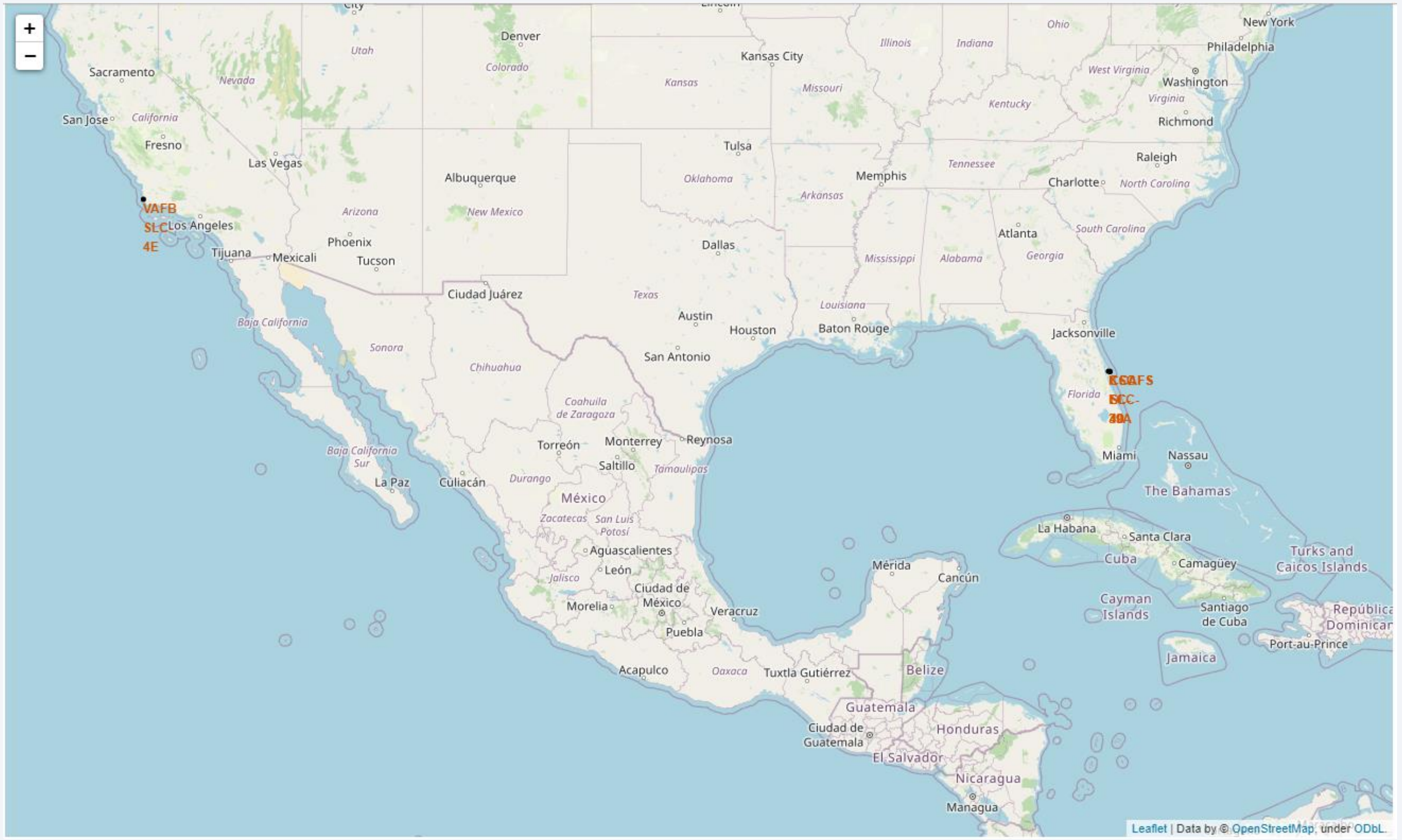
landing__outcome	total_count
Failure (drone ship)	5
Success (ground pad)	3

Section 4

Launch Sites Proximities Analysis

Location of all launch sites with Folium

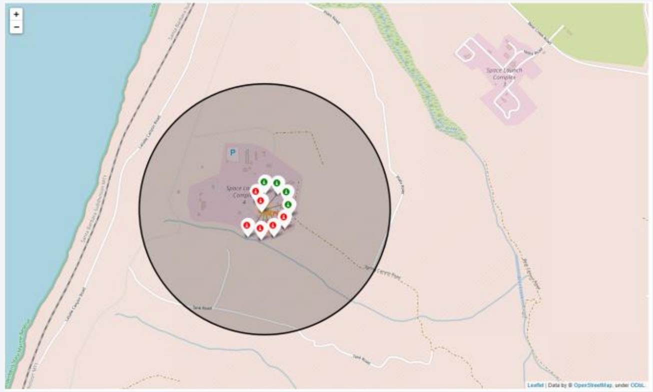
In the following map we can see that all launch sites are located in the coasts of United States of America.



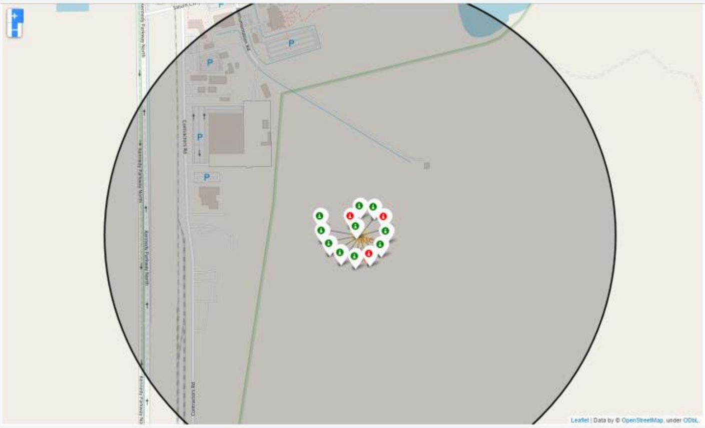
Success and failed launches for each site

By using MarkerCluster in Folium we are able to show all the successful and failure missions on each site.

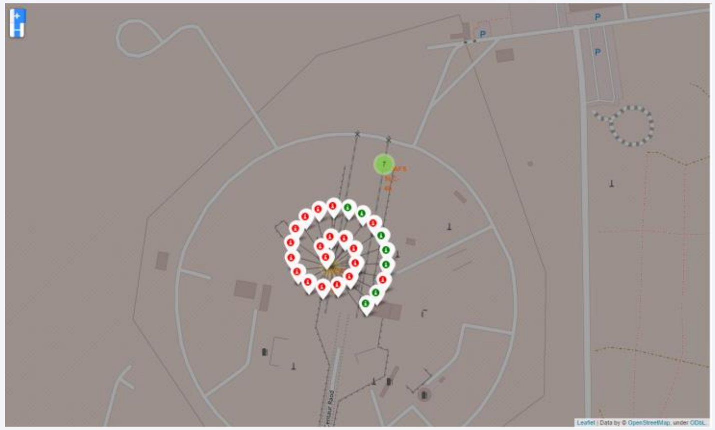
Green markers represent the successful launches and the red markers the failures.



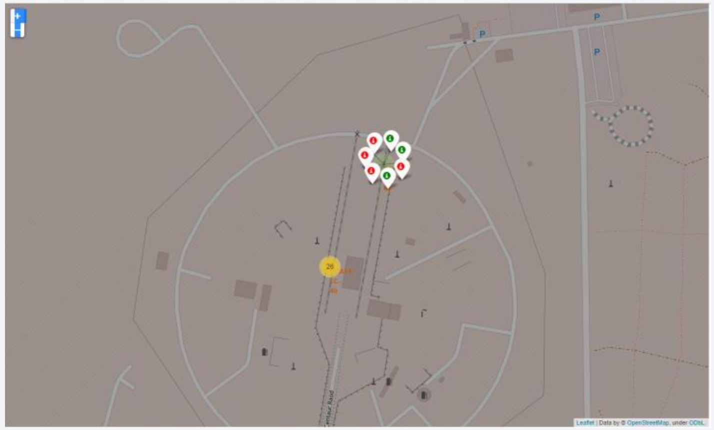
VAFB SLC-4E



KSC LC-39A



CCAFS LC-40



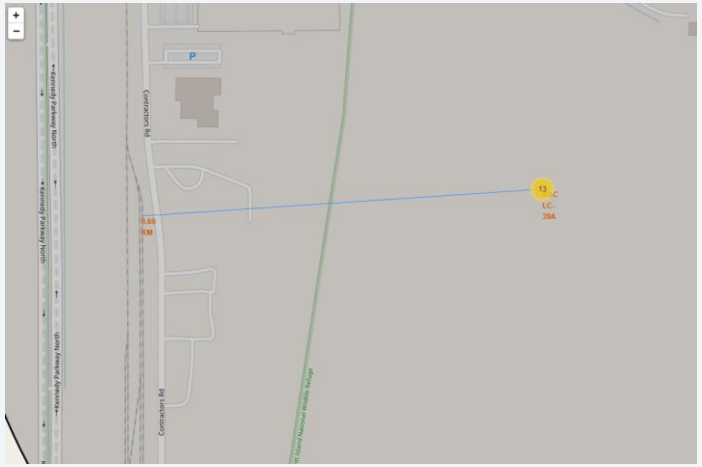
CCAFS SLC-40

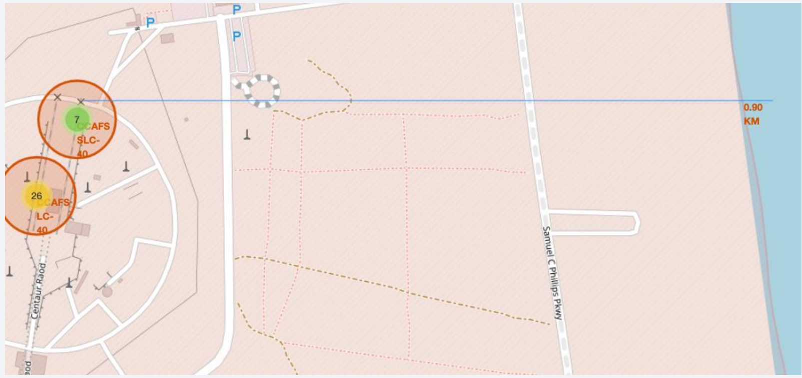
Launch sites and its proximities

The interactive maps generated with Folium allows to identify some characteristics about the launch sites.

We can see the proximity of the sites to railway, highway, coastline and the distance considerable distance from populated cities.

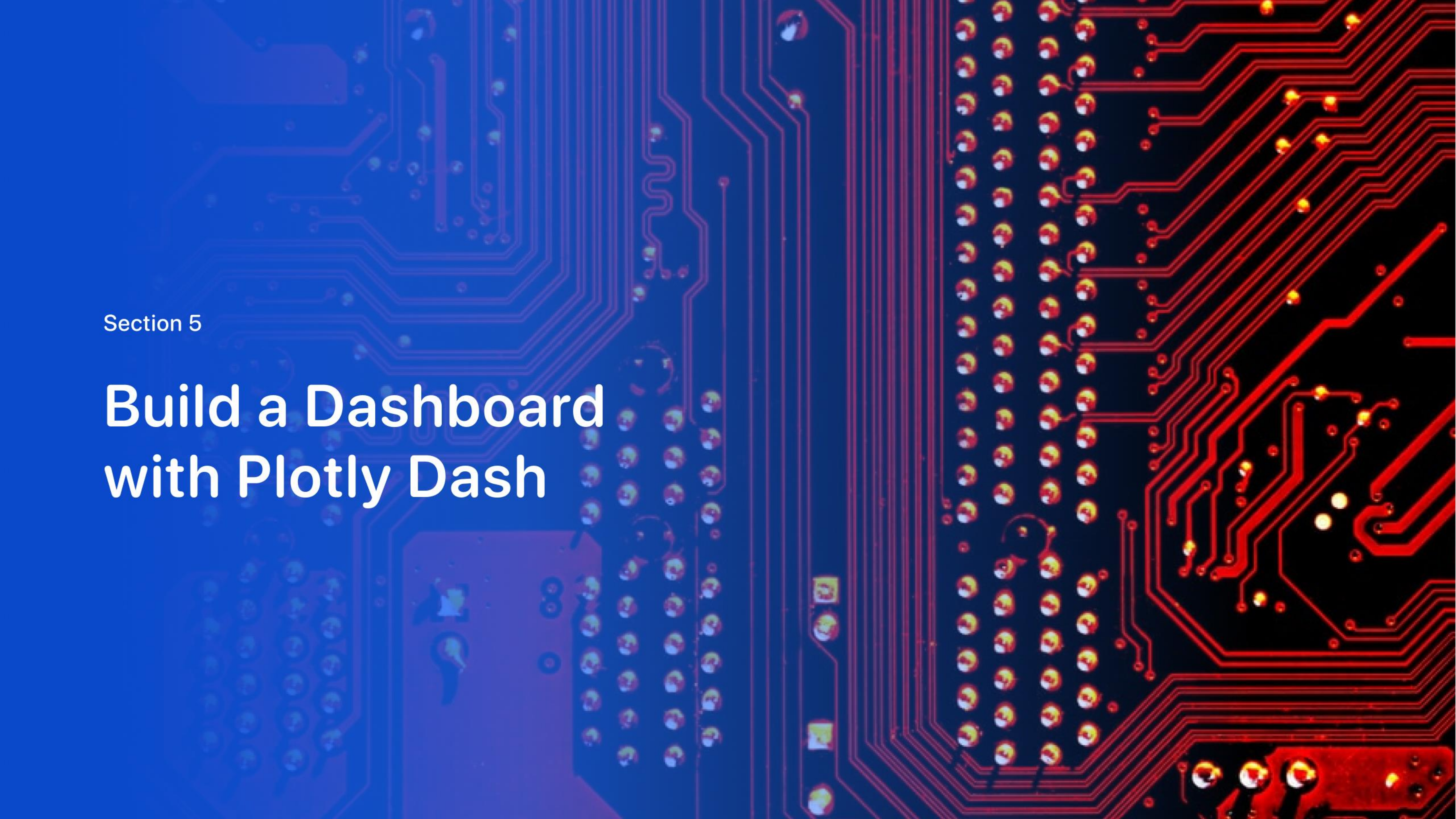
Folium allows to calculate the distance from two different cordenates and also to add polylines over the map.





Section 5

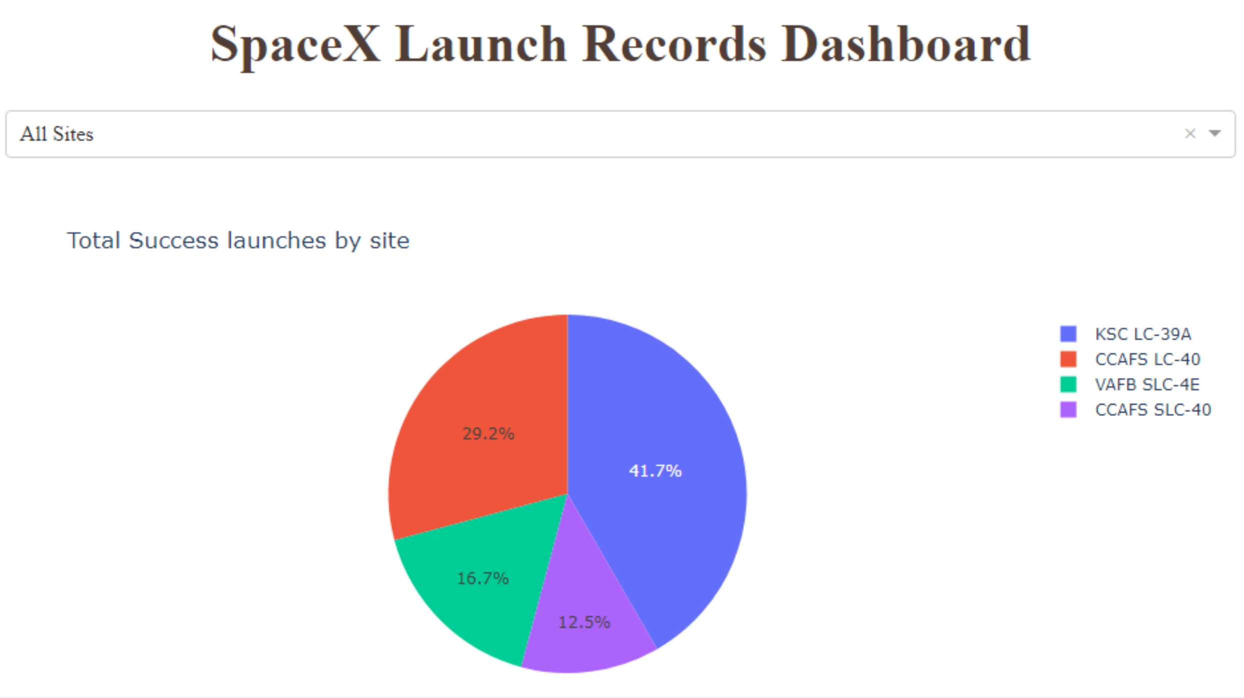
Build a Dashboard with Plotly Dash



SpaceX Launch Records Dashboard

Interactive dashboard created with Dash in Python is very flexible for the layout and the elements that can be included.

In the following image we can see that KSC LC-39A has the highest success rate.



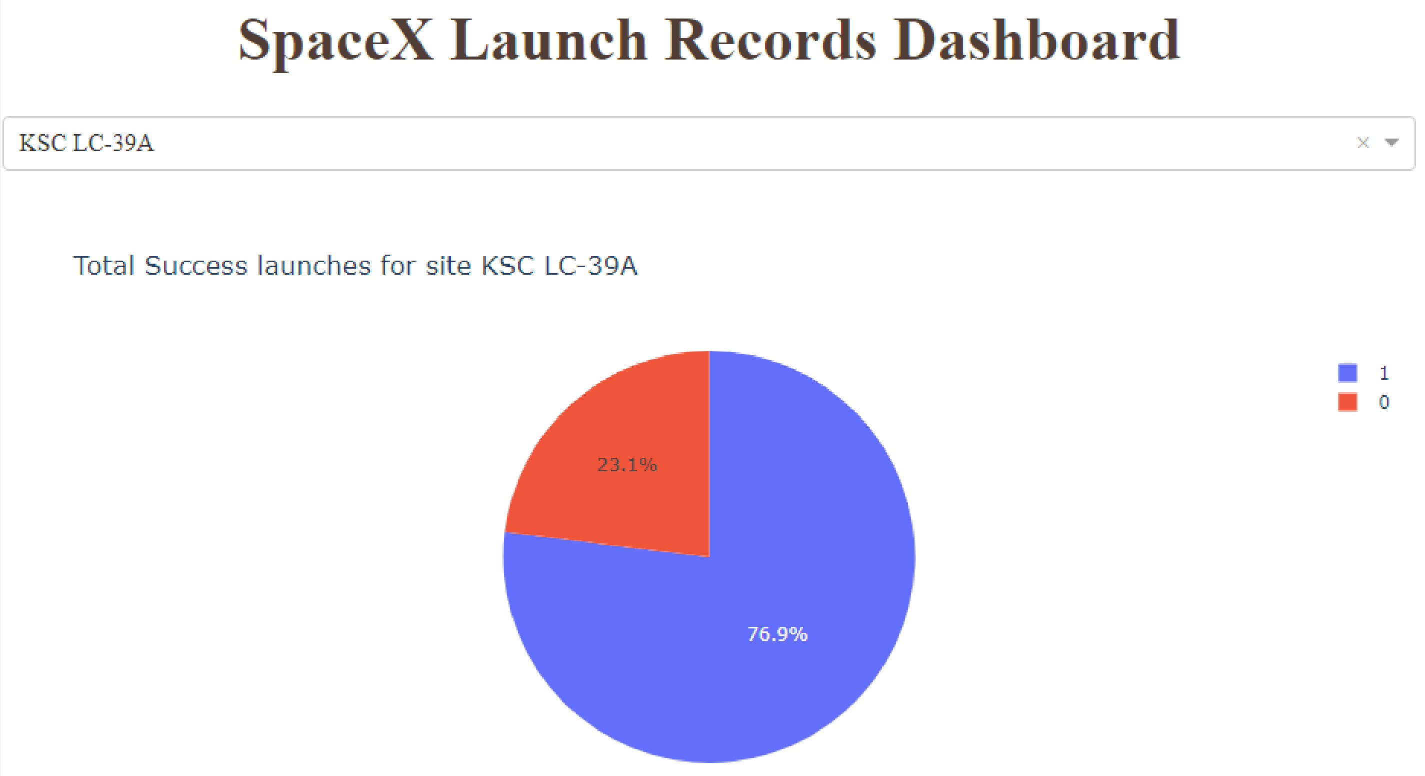
SpaceX Launch Records Dashboard

All Sites

Total Success launches by site

Dashboard success rate per site

Pie chart bellow shows the success rate for KSC LC-39A which has the higher rate among the sites.



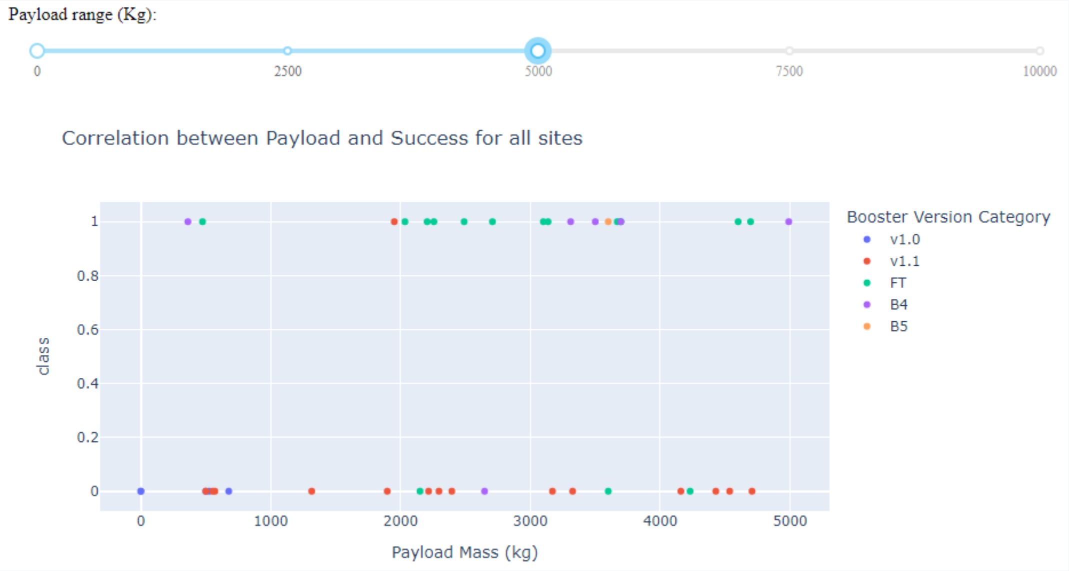
SpaceX Launch Records Dashboard

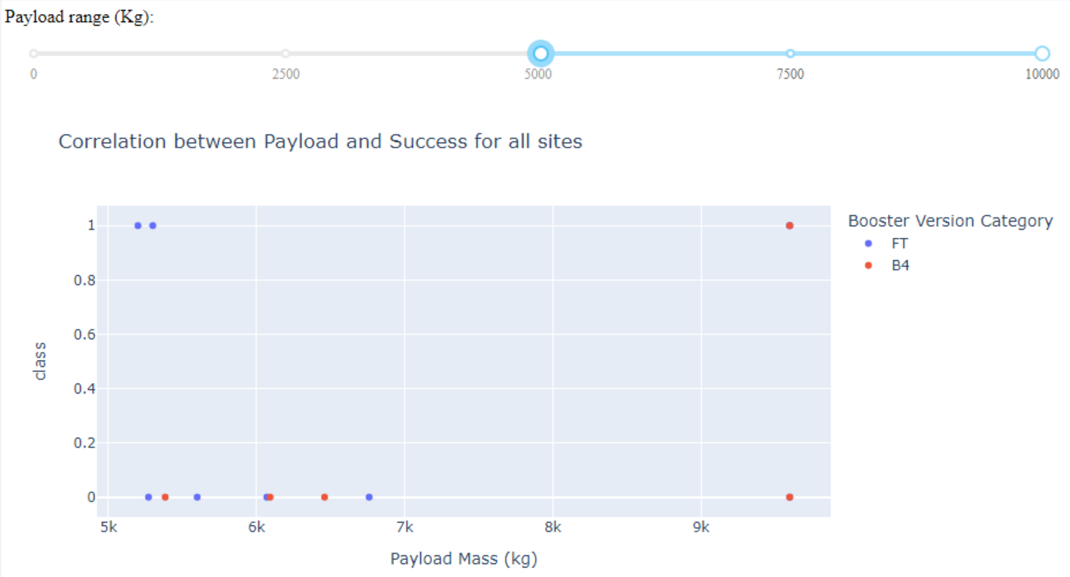
KSC LC-39A

Total Success launches for site KSC LC-39A

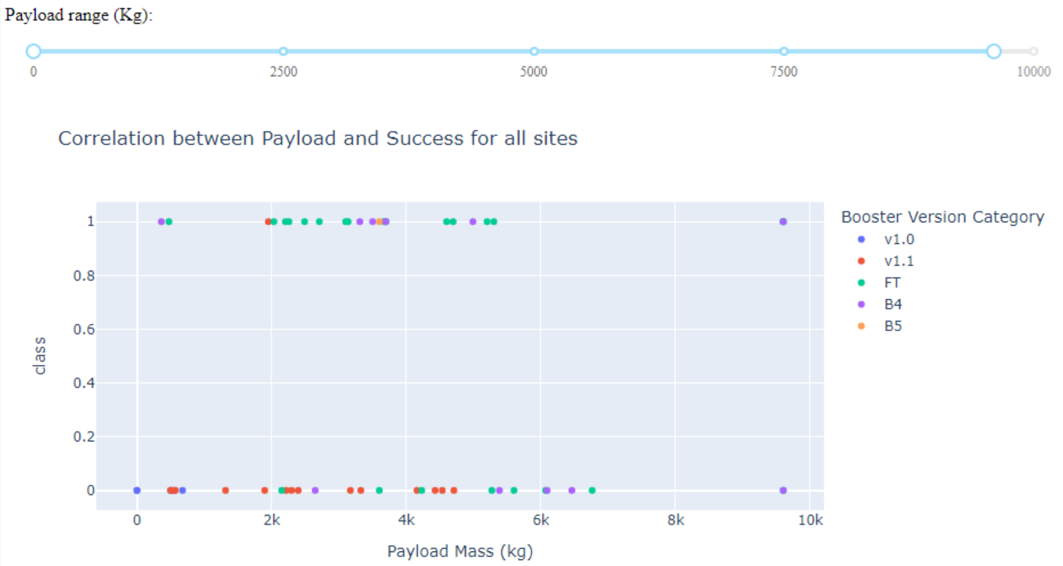
Correlation between Payload vs. Launch Outcome

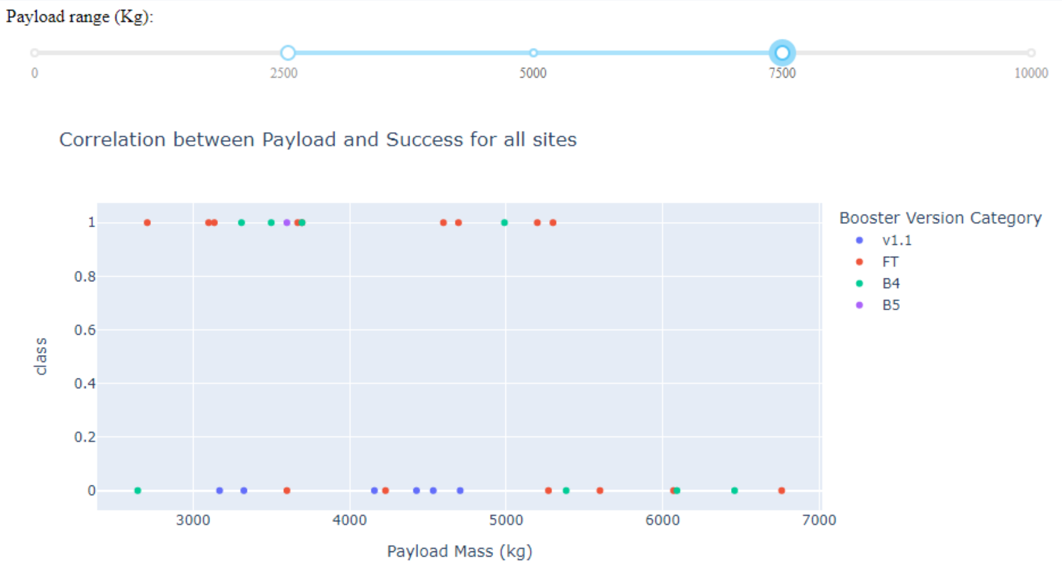
Scatter plot shows that success rate is higher for the low payload mas than for the weighted payloads.





Correlation between Payload vs. Launch Outcome



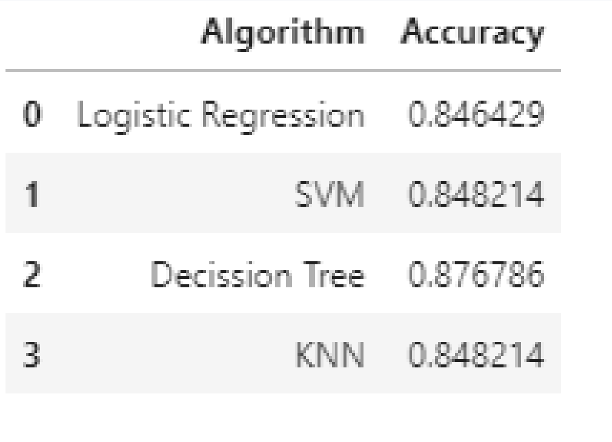


Section 6

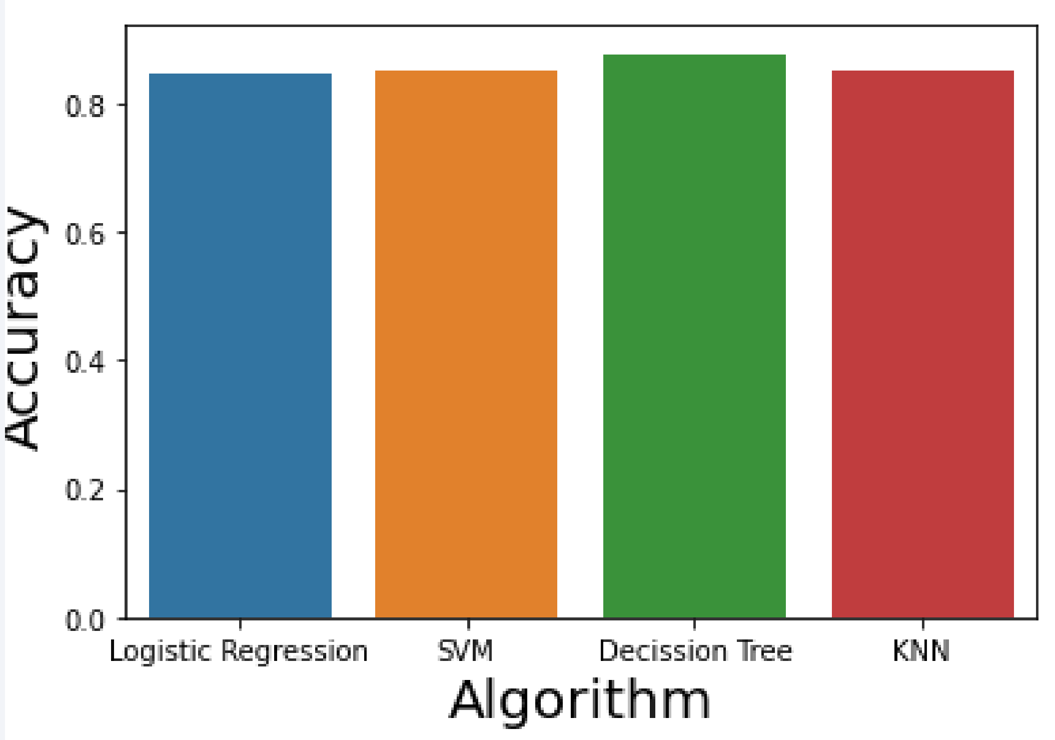
Predictive Analysis (Classification)

Classification Accuracy

According to the results Decision Tree is the best algorithm for our problem. But this could be due to the small data set because in general all algorithms performs very well with almost the same accuracy.

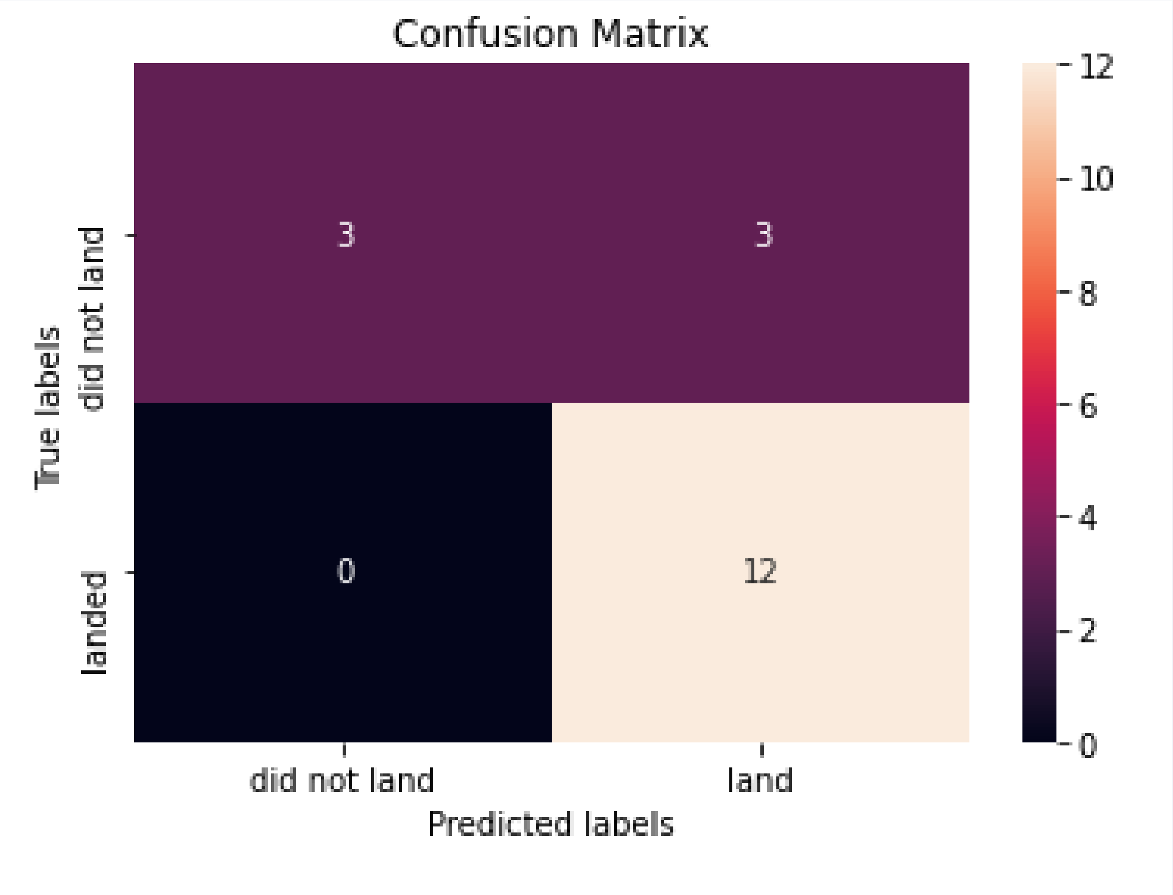


	Algorithm	Accuracy
0	Logistic Regression	0.846429
1	SVM	0.848214
2	Decission Tree	0.876786
3	KNN	0.848214



Confusion Matrix

As we can see in the confusion matrix that Decision Tree can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- The success rate has increased over the years
- ES-L1, GEO, HEO and SSO orbits have a high success rate
- KSC LC-39A launch site has the highest success rate
- Logistic regression, SVM, Decision Tree and KNN models have similar high accuracy
- Decision Tree algorithm shows the best accuracy results
- False positives are the most considerable problem for the classification model with the current dataset

Thank you!