IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sethuram
Feb 22, 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis through Data Visualization
  - Interactive Visual Analytics with Folium
  - Predictive analysis – Classification
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context
  - Space X 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 Space X 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 space X for a rocket launch. The objective of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- Problems you want to find answers
  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology

Executive Summary

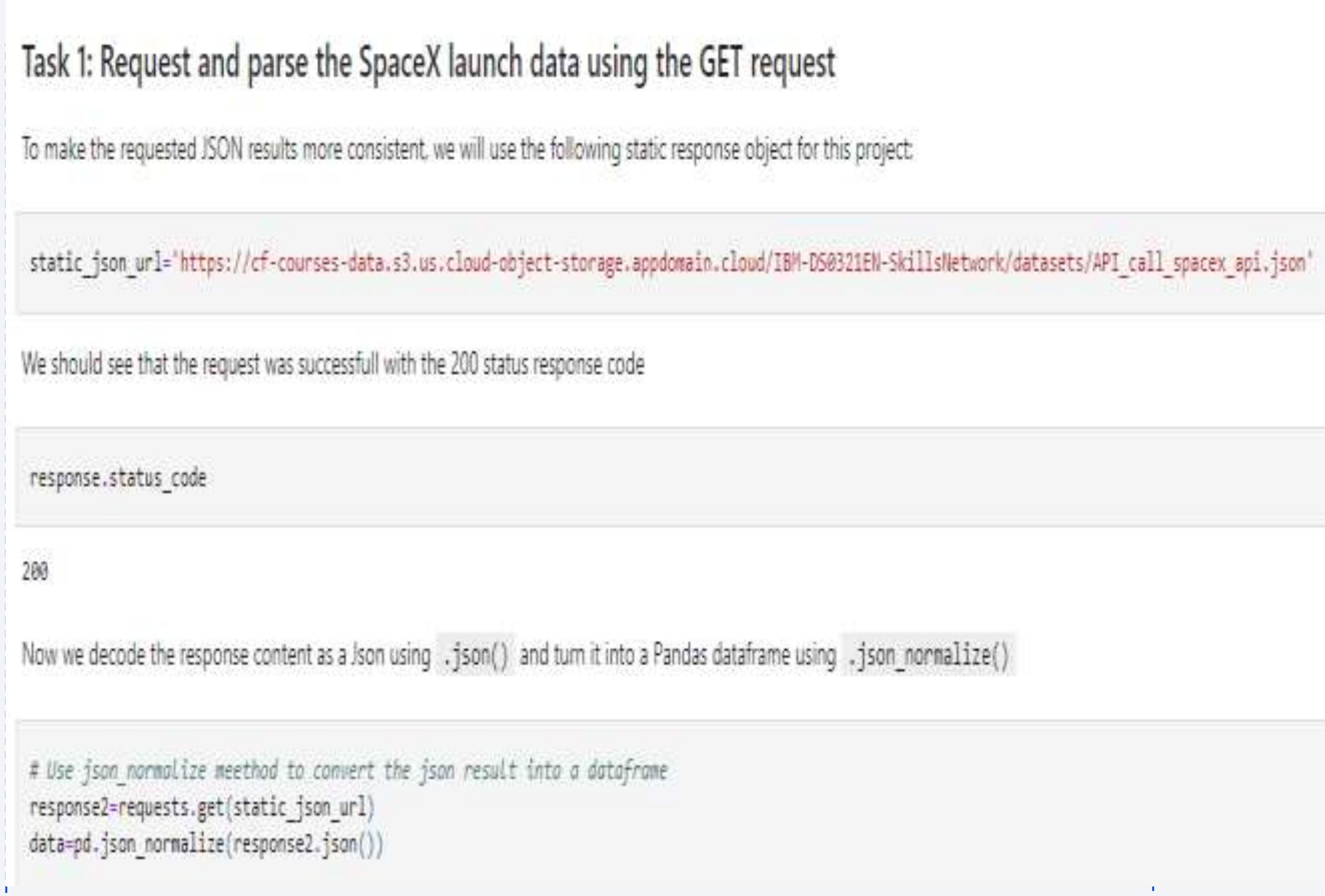
- Data collection methodology:
  - SpaceX REST API
  - Webscrapping from Wikipedia page
- Perform data wrangling
  - Impute missing values
  - One hot coding of categorical variables
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic regression, SVM, KNN, Decision trees models have been built and evaluated to find best classifier

Data Collection

- Data for the project was obtained from two sources:
  - SpaceX API (https://api.spacexdata.com/v4/rockets/)
  - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used;
- This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- Code: https://github.com/sethu470/ibm-datascience-project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Task 1: Request and parse the SpaceX launch data using the GET request

To make the requested JSON results more consistent, we will use the following static response object for this project:

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json'
```

We should see that the request was successfull with the 200 status response code

```
response.status_code
```

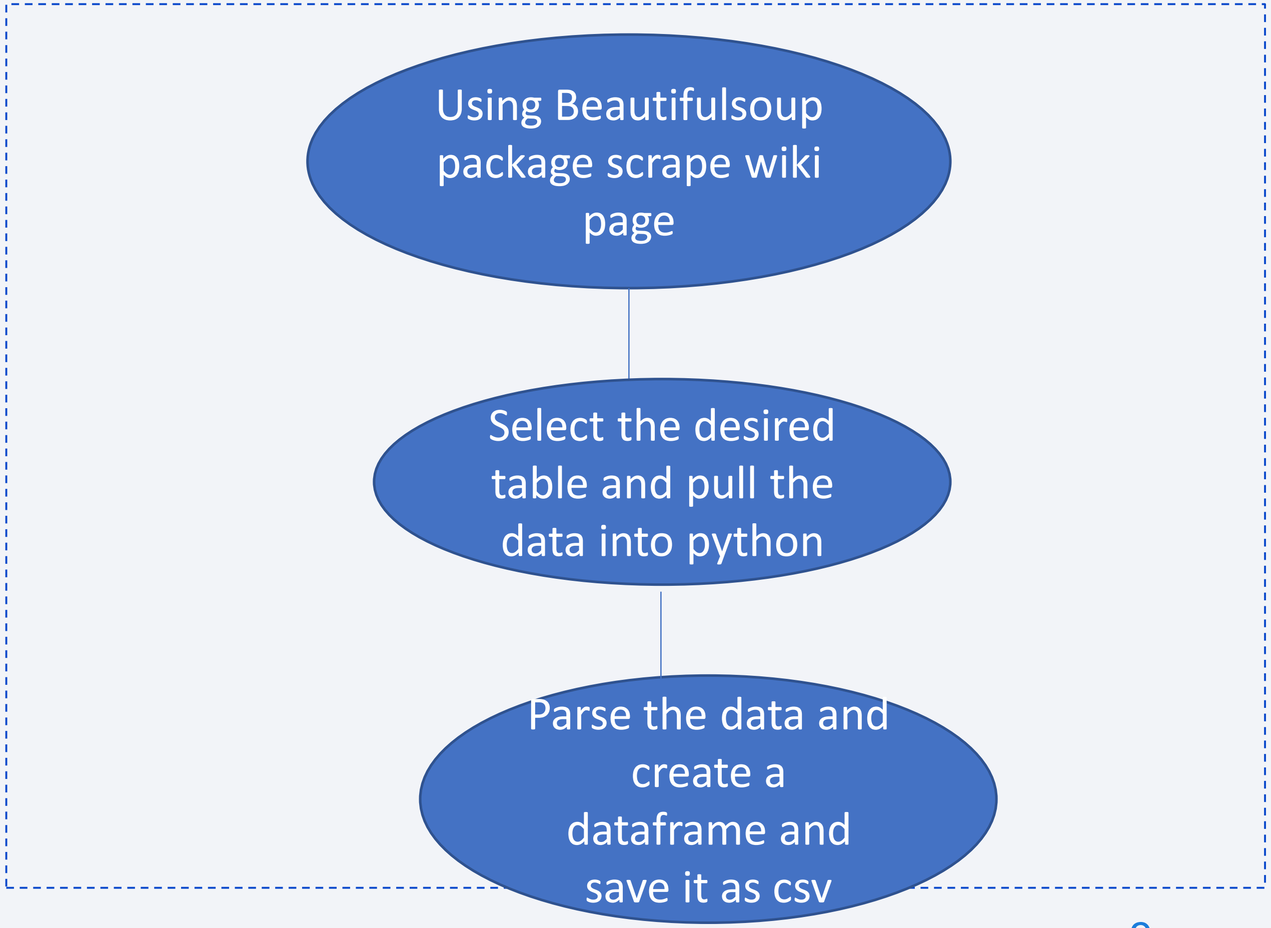
```
200
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

```
# Use json_normalize meethod to convert the json result into a dataframe
response2=requests.get(static_json_url)
data=pd.json_normalize(response2.json())
```

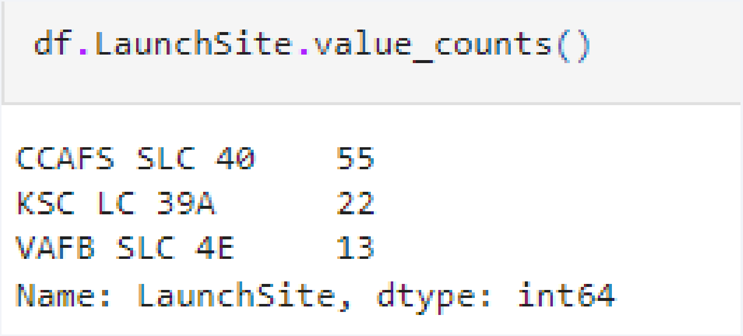
Data Collection - Scraping

- Another source for obtaining Falcon 9 Launch data is web scraping related Wiki pages. Using the Python BeautifulSoup package, scrapped this page (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches) to pull data data from tables that contain valuable Falcon 9 launch records.
- Code: https://github.com/sethu470/ibm-datascience-project/blob/main/jupyter-labs-webscraping.ipynb


Using Beautifulsoup package scrape wiki page
Select the desired table and pull the data into python
Parse the data and create a dataframe and save it as csv

Data Wrangling

- Perform exploratory Data Analysis and determine Training Labels
    - Exploratory Data Analysis
    - Determine Training Labels
- Explored Launchsite variable to determine the number of launches on each site:



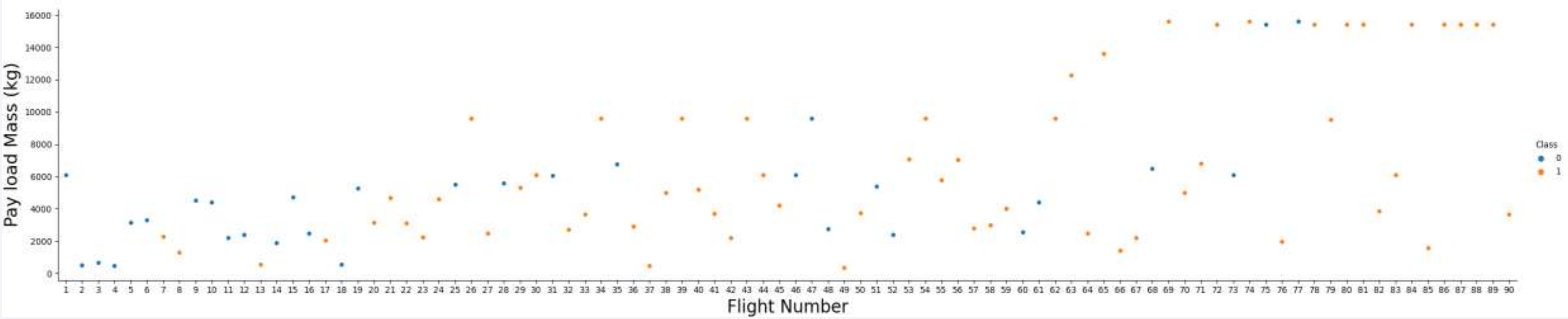
```
df.LaunchSite.value_counts()
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

- Calculate the number of occurrence of each orbit
- Calculate the number and occurrence of mission outcome per orbit type
- Create a landing outcome label from Outcome column
- Code: https://github.com/sethu470/ibm-datascience-project/blob/main/jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

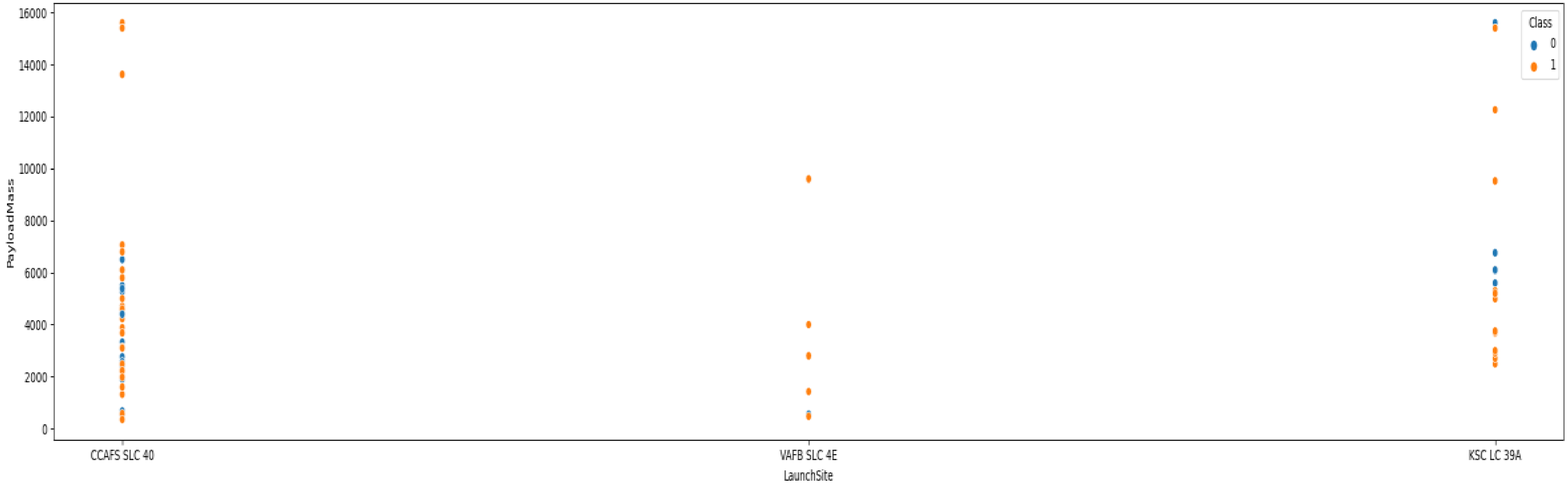
- FlightNumber vs. Payload Mass and overlay the outcome of the launch. We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.



- Code: https://github.com/sethu470/ibm-datascience-project/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with Data Visualization

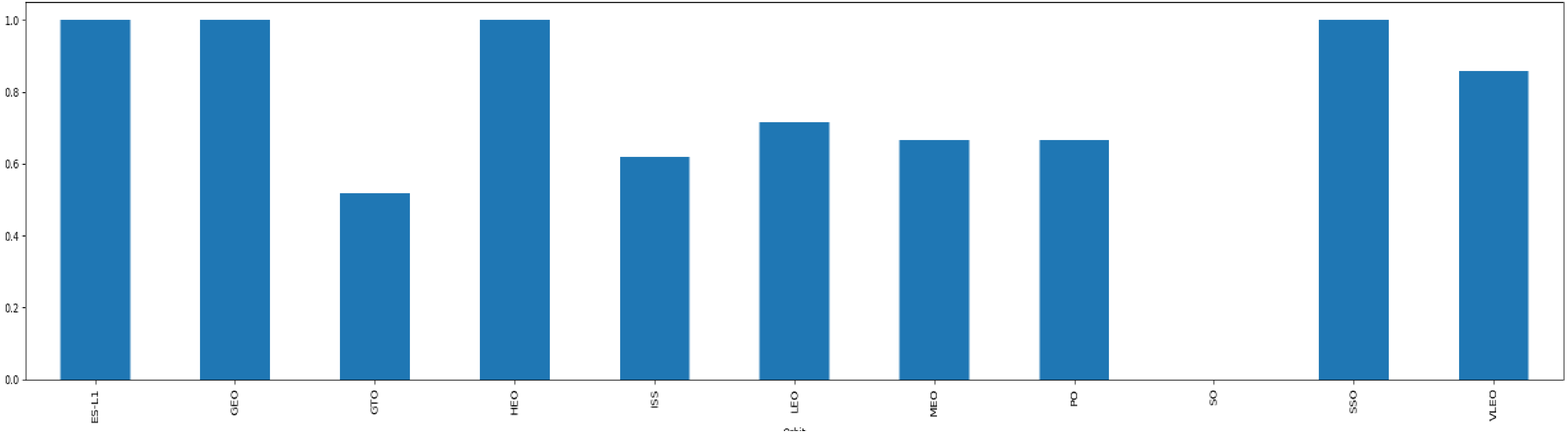
- Launchsite vs. Payload , from this plot we can find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).



- Code: https://github.com/sethu470/ibm-datascience-project/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with Data Visualization

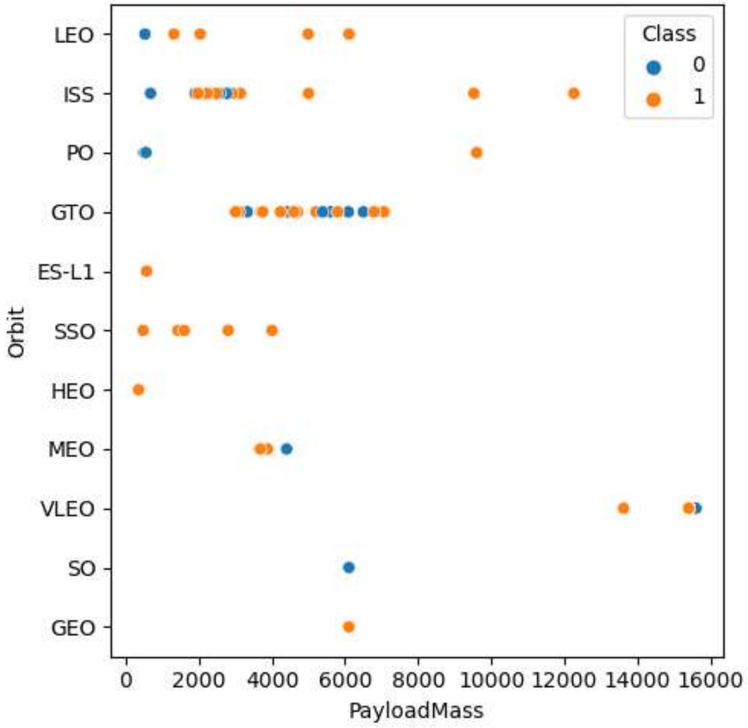
- *Visualize the relationship between success rate of each orbit type*



- Code: https://github.com/sethu470/ibm-datascience-project/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with Data Visualization

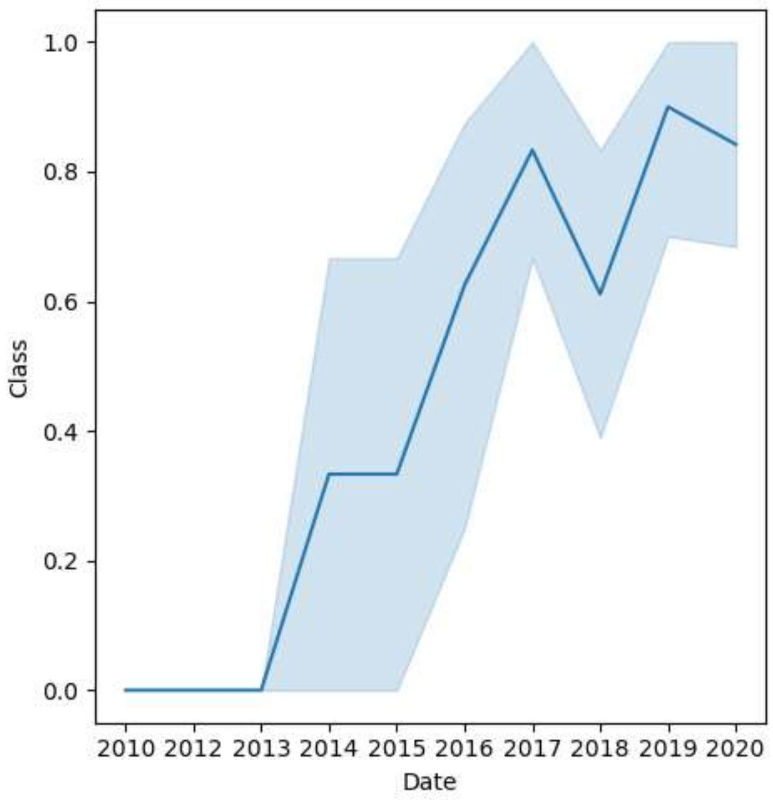
- *Visualize the relationship between Payload and Orbit type*



- Code: https://github.com/sethu470/ibm-datascience-project/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with Data Visualization

- *Visualize the launch success yearly trend*



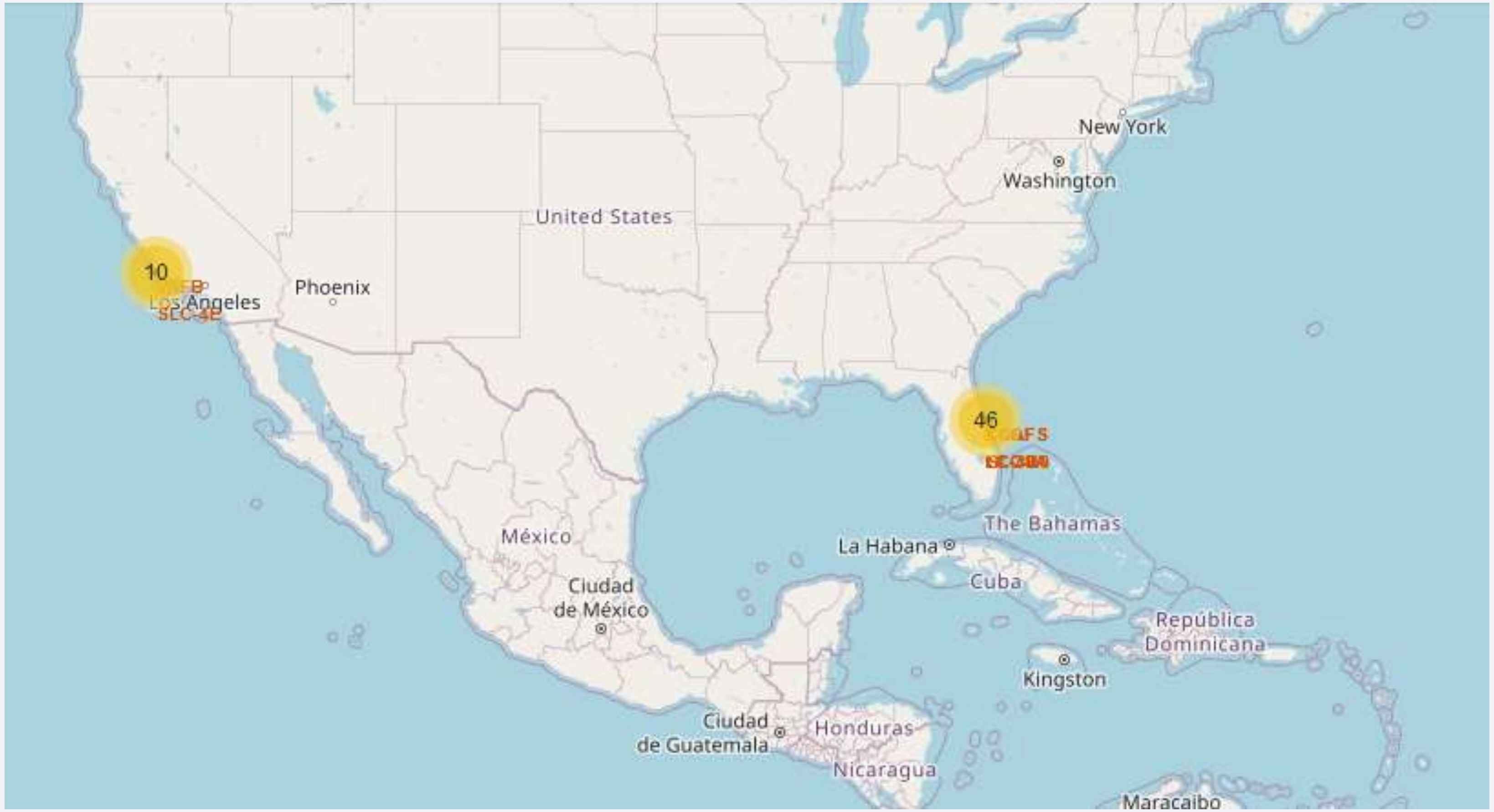
- Code: https://github.com/sethu470/ibm-datascience-project/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- To summarize the SQL queries performed:
    - To display the names of the unique launch sites in the space mission
    - To display 5 records where launch sites begin with the string 'CCA'
    - To display the total payload mass carried by boosters launched by NASA (CRS)
    - To display average payload mass carried by booster version F9 v1.1
    - To list the date when the first succesful landing outcome in ground pad was acheived
    - To list the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
    - To list the total number of successful and failure mission outcomes
    - To list the names of the booster_versions which have carried the maximum payload mass
    - To 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.

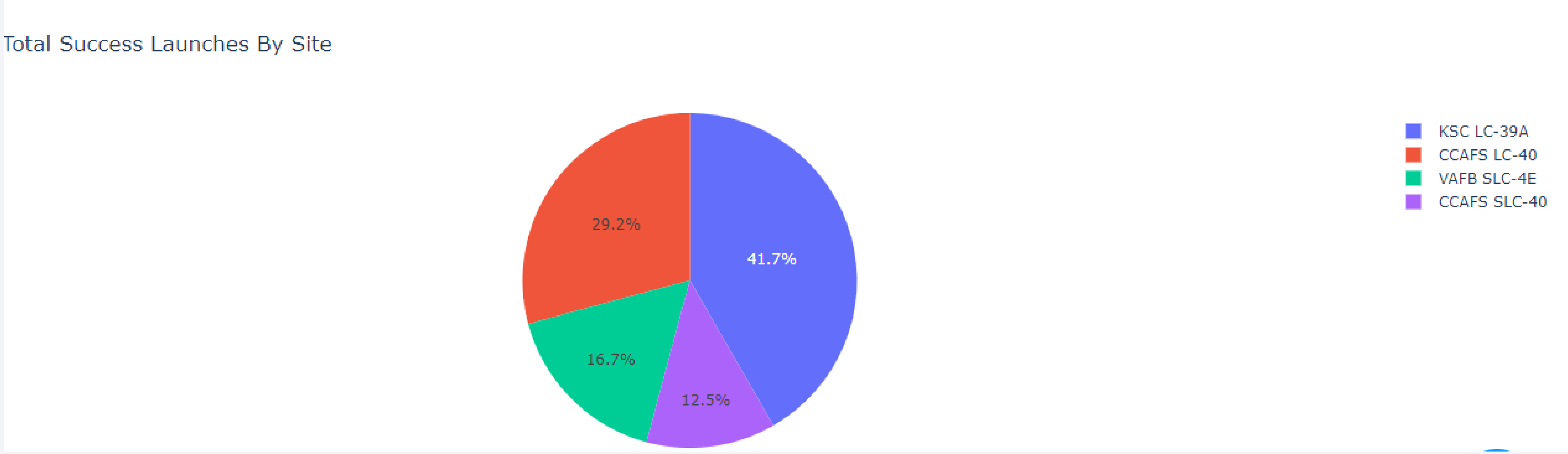
- Code: https://github.com/sethu470/ibm-datascience-project/blob/main/eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

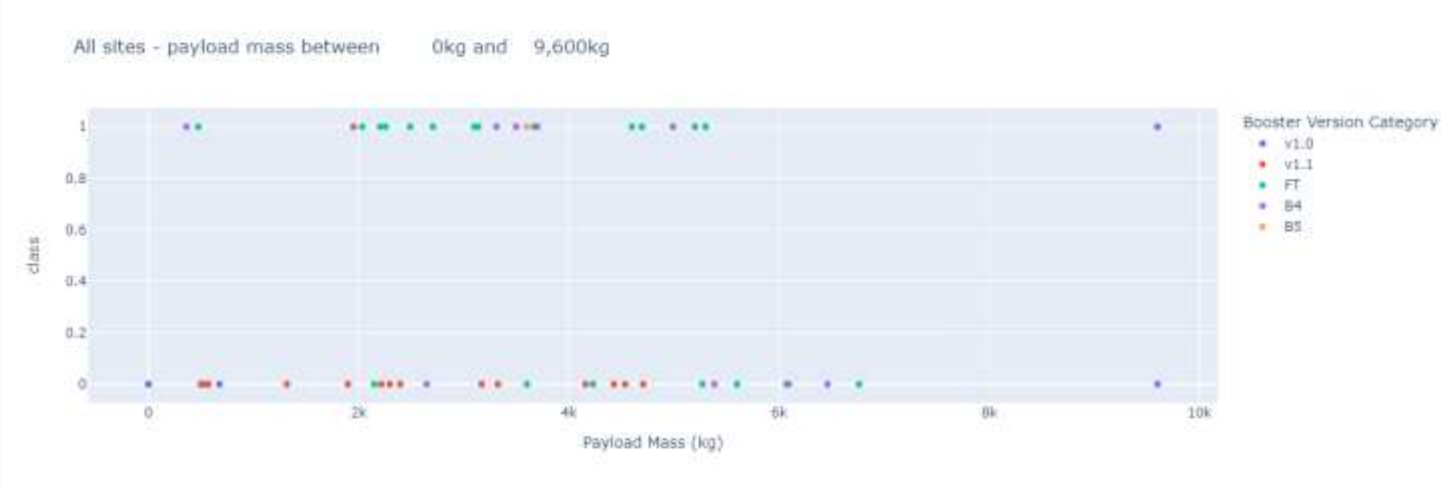


Code - https://github.com/sethu470/ibm-datascience-project/blob/main/jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash



Total Success Launches By Site



All sites - payload mass between 0kg and 9,600kg

Code - https://github.com/sethu470/ibm-datascience-project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Data is collected from various sources (APIs and Webscrapping)
- Data cleaning has been done (imputing missing values)
- A new dataframe is created with selected independent variables and dependent variable
- Through one hot encoding, all categorical variables are converted to numerical columns
- Through standardization, all column values are brought to a common range
- Created an array for Independent variables(X) and one for dependent variable (y)
- Dataset has been split into training and test datasets
- We built different machine learning models and tune different hyperparameters using GridSearchCV
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.

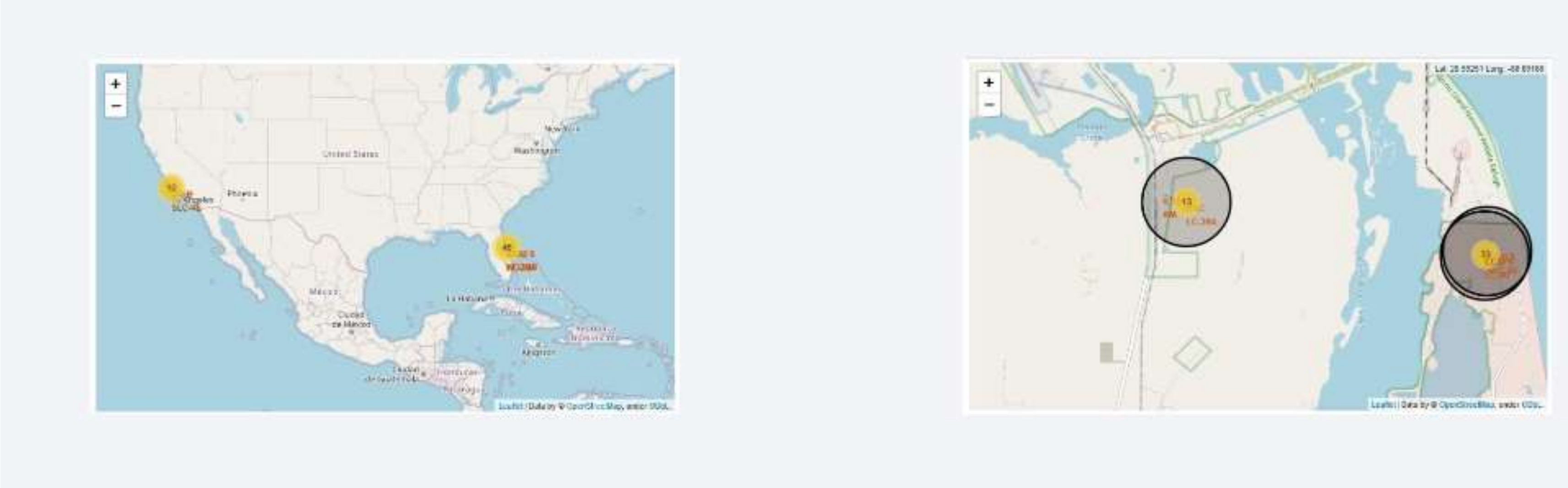
- Code - https://github.com/sethu470/ibm-datascience-project/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results
  - Space X uses 4 different launch sites; CCAFS SLC 40, KSC LC 39A, VAFB SLC 4E, CCAF LC-40
  - The first launches were done to Space X itself and NASA;
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - The first success landing outcome happened in 2015 fiver year after the first launch;
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The number of landing outcomes became as better as years passed

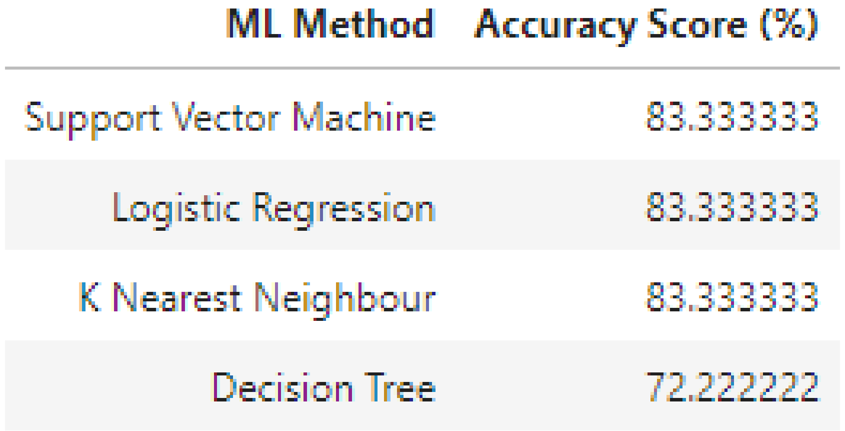
Results

- Interactive analytics demo in screenshots
    - Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.
    - Most launches happens at east cost launch sites.



Results

- Predictive Analytics result
  - Logistic Regression, SVM and KNN models are the best in terms of prediction accuracy (83%) for this dataset.

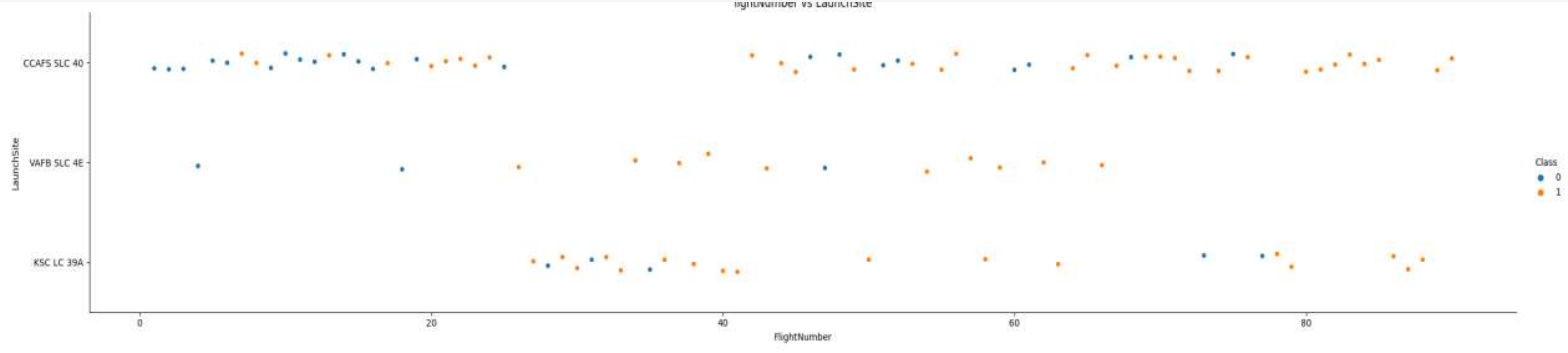


ML Method	Accuracy Score (%)
Support Vector Machine	83.333333
Logistic Regression	83.333333
K Nearest Neighbour	83.333333
Decision Tree	72.222222

Section 2

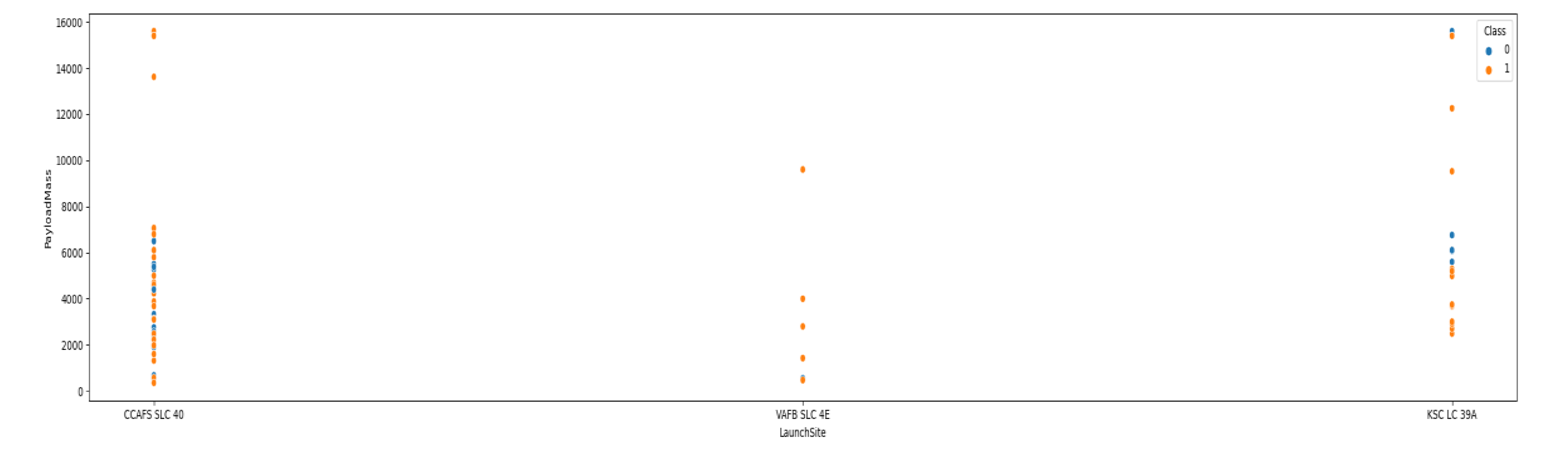
Insights drawn from EDA

Flight Number vs. Launch Site



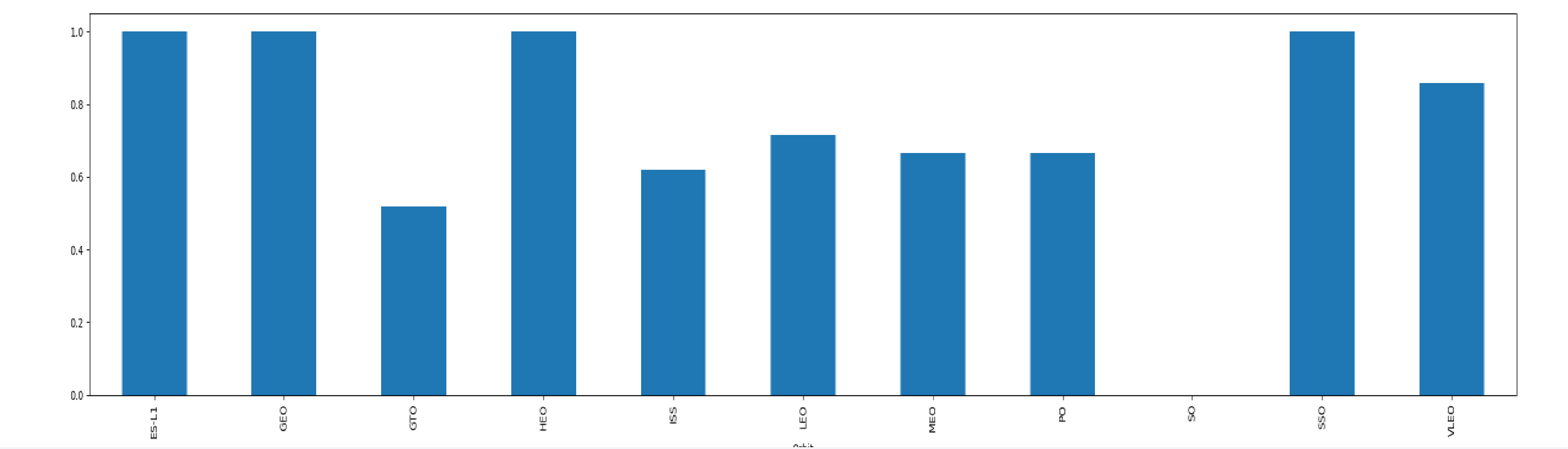
- From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.

Payload vs. Launch Site



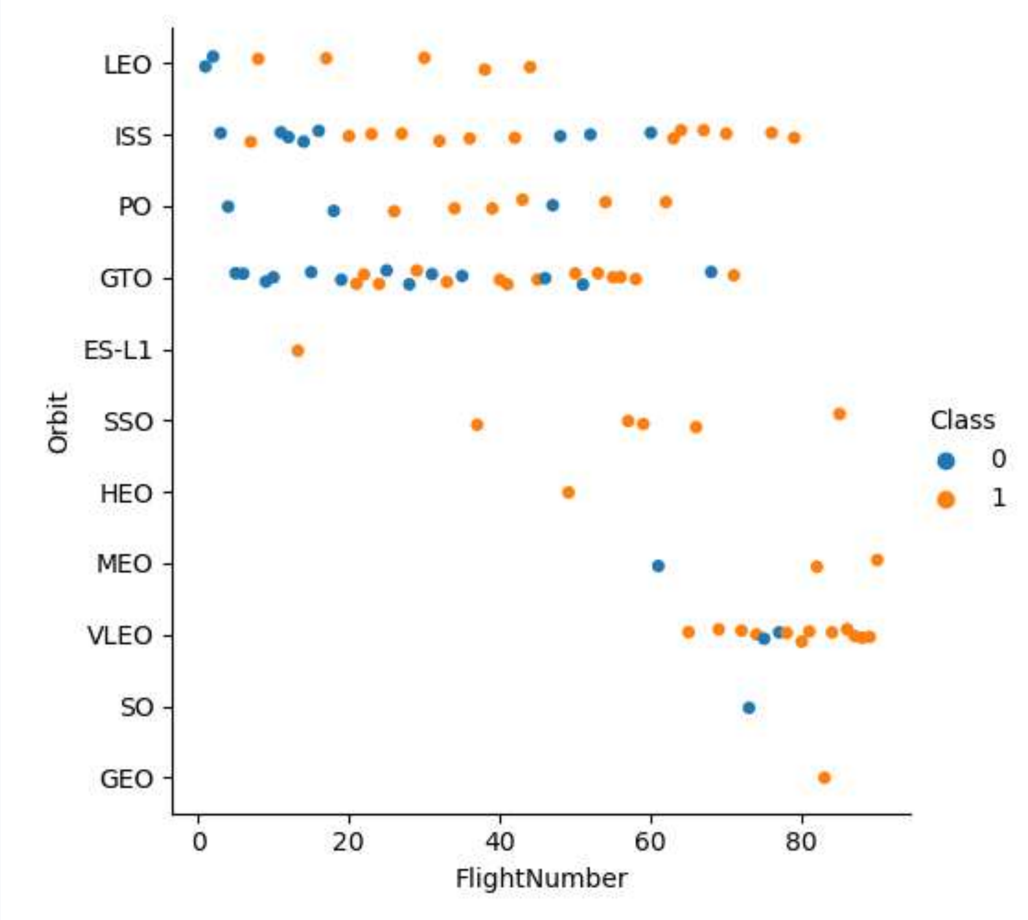
- The greater the payload mass for launch site CCAFS SLC 40 the higher the success rate for the rocket

Success Rate vs. Orbit Type



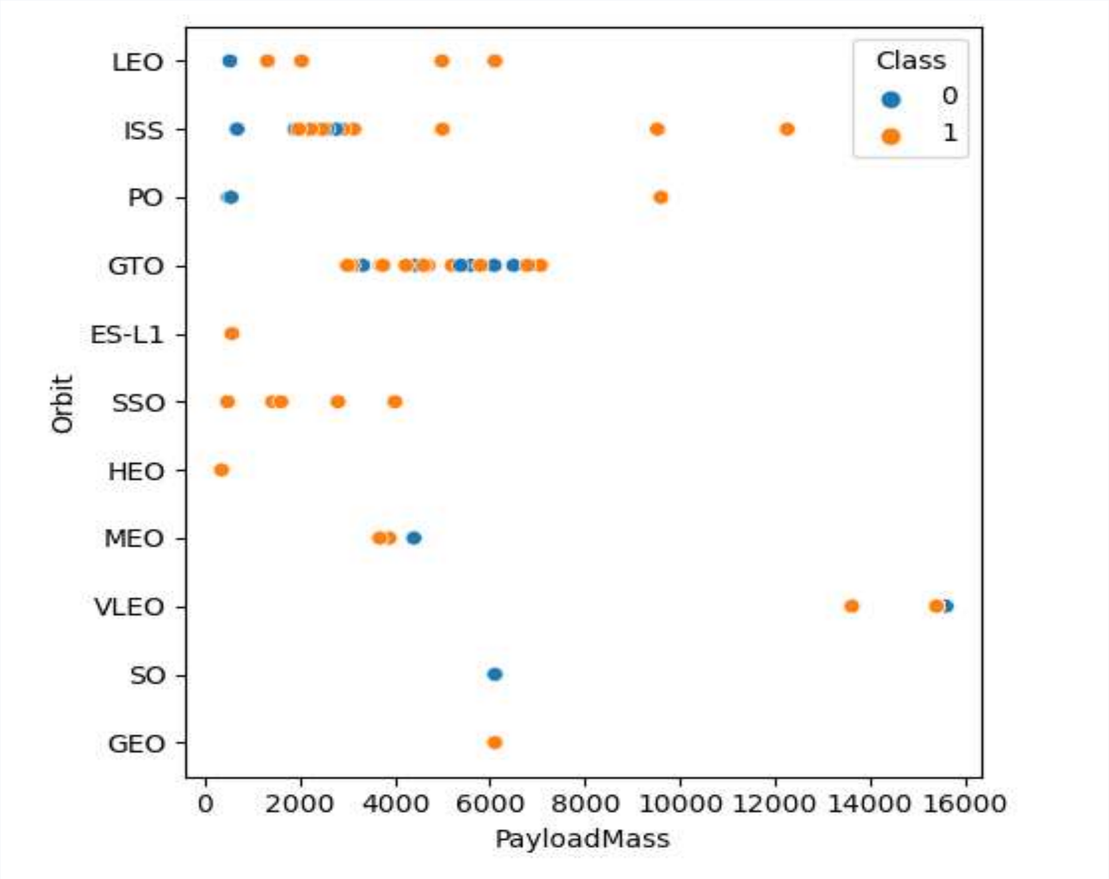
- From the plot, we can see that ES-L1, GEO, HEO, SSO, VLEO had the most success rate.

Flight Number vs. Orbit Type



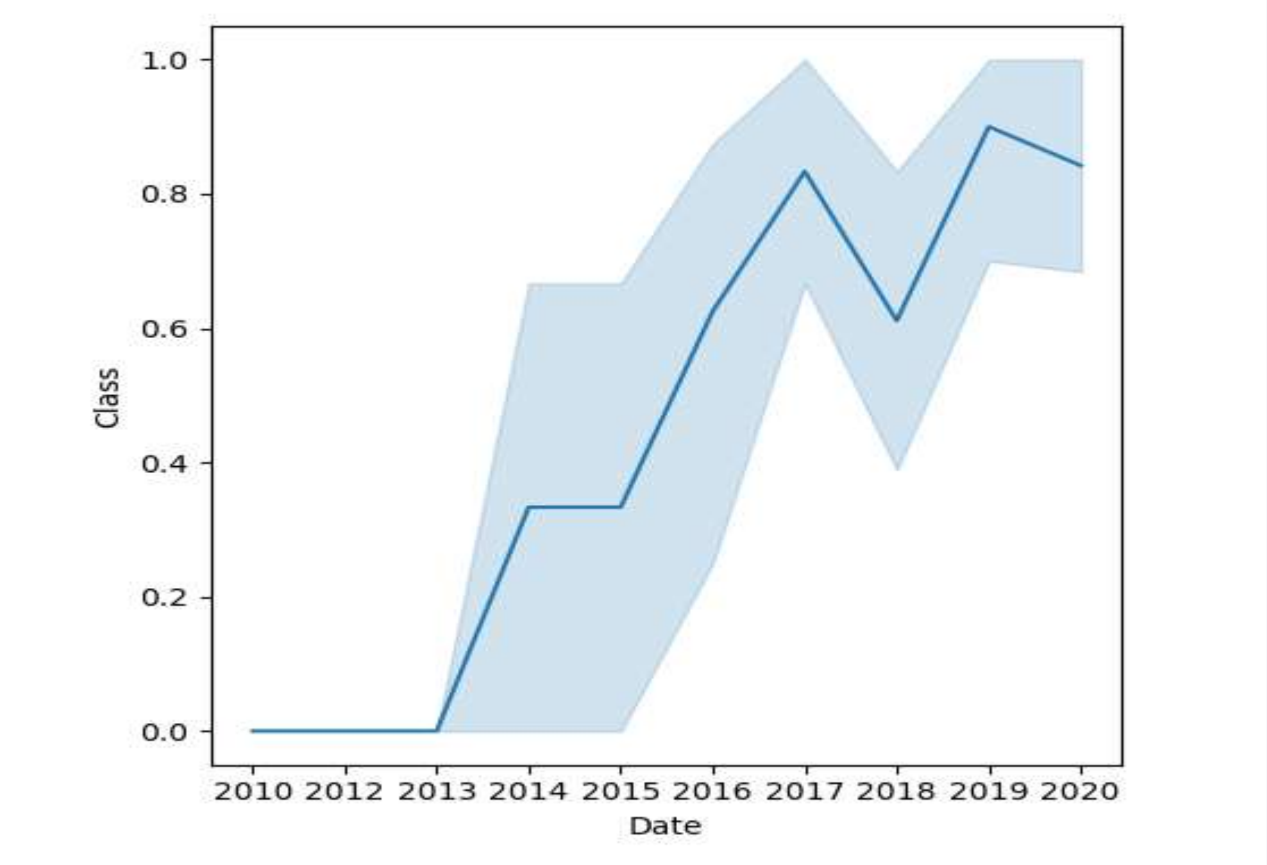
- We observe that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.

Payload vs. Orbit Type



- We can observe that with heavy payloads, the successful landing are more for PO, LEO and ISS orbits.

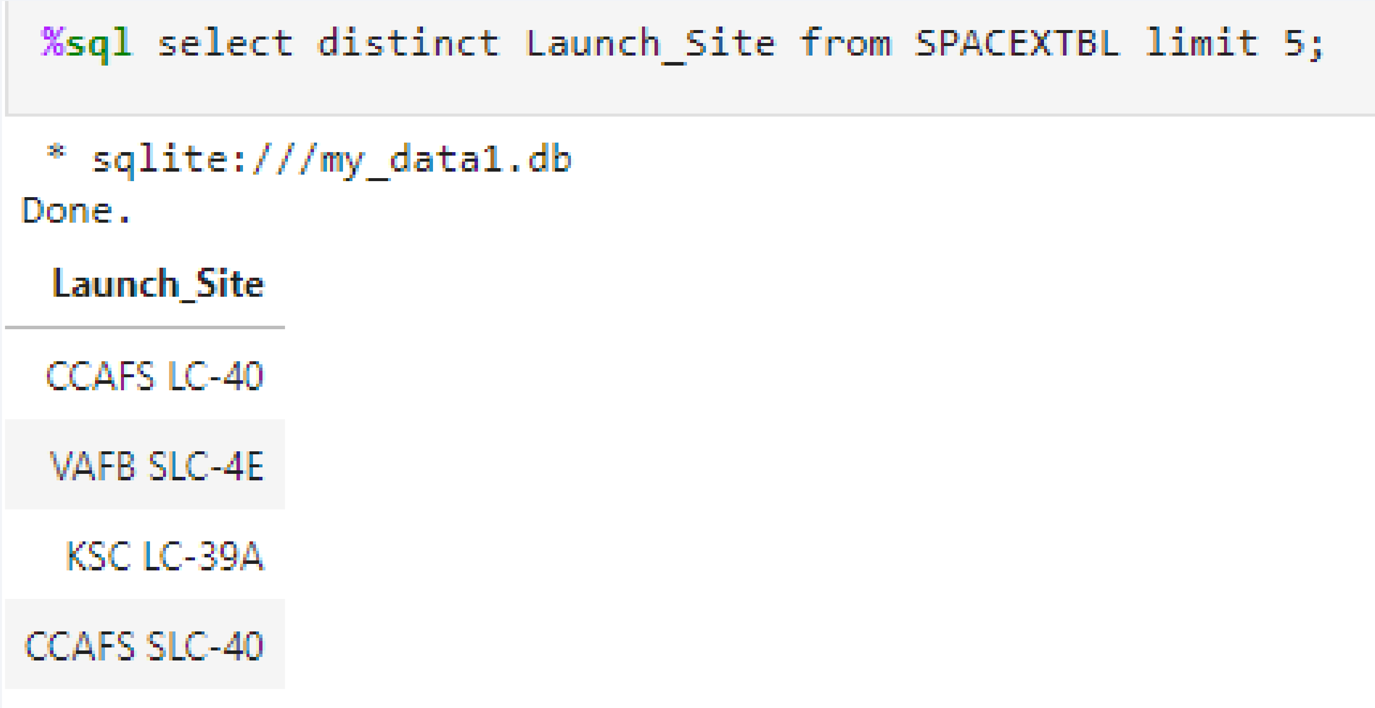
Launch Success Yearly Trend



- From the plot, we can observe that success rate since 2013 kept on increasing till 2020.

All Launch Site Names

- Using SQL DISTINCT command, we can pull unique launch sites from the SpaceX data.



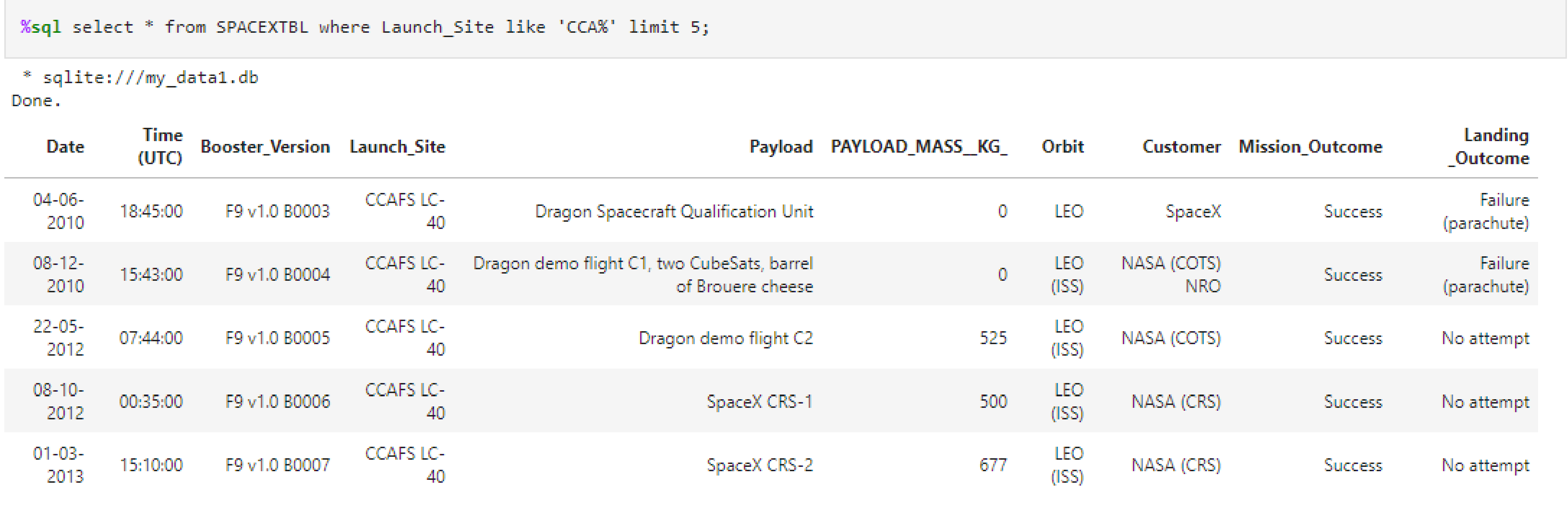
```
%sql select distinct Launch_Site from SPACEXTBL limit 5;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

- Display 5 records where launch sites begin with `CCA`



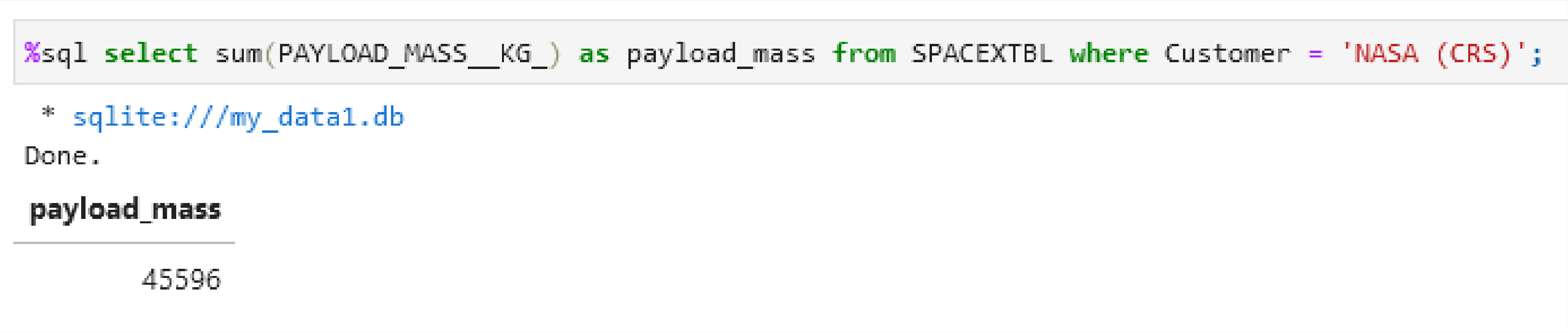
```
%sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5;
```

```
 * sqlite:///my_data1.db
Done.
```

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

- **Query to calculate the total payload mass carried by boosters launched by NASA (CRS) :**



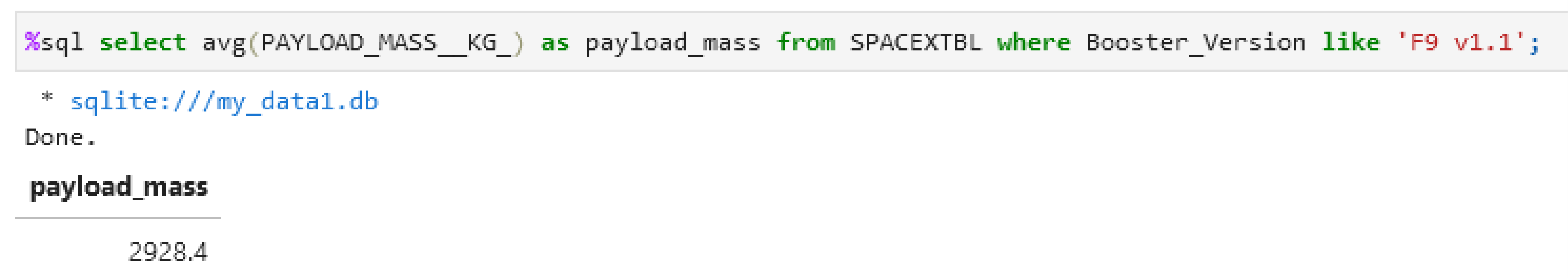
```
%sql select sum(PAYLOAD_MASS__KG_) as payload_mass from SPACEXTBL where Customer = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

payload_mass
45596

Average Payload Mass by F9 v1.1

- Query to calculate the average payload mass carried by booster version F9 v1.1



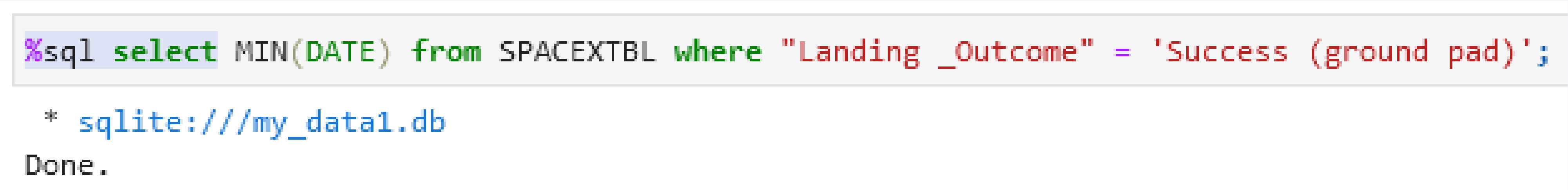
```
%sql select avg(PAYLOAD_MASS__KG_) as payload_mass from SPACEXTBL where Booster_Version like 'F9 v1.1';
```

```
 * sqlite:///my_data1.db
Done.
```

payload_mass
2928.4

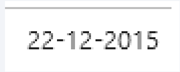
First Successful Ground Landing Date

- We observed that the dates of the first successful landing outcome on ground pad was 22nd December 2015



```
%sql select MIN(DATE) from SPACEXTBL where "Landing _Outcome" = 'Success (ground pad)';
```

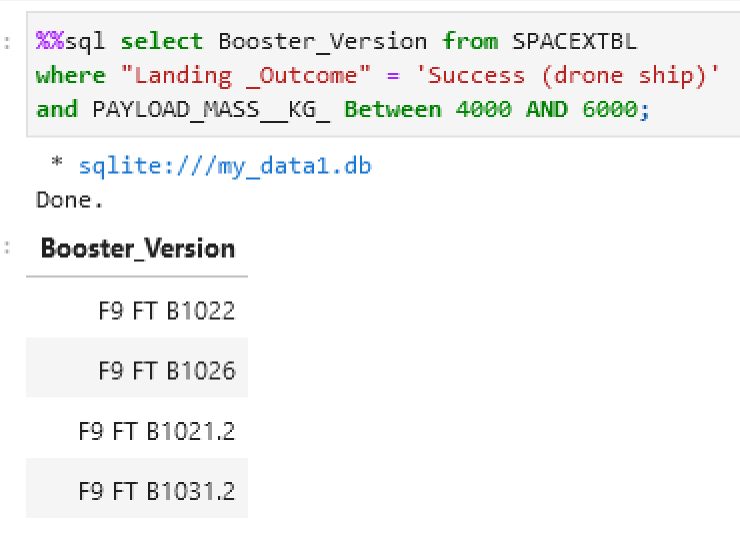
```
 * sqlite:///my_data1.db
Done.
```



```
22-12-2015
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000



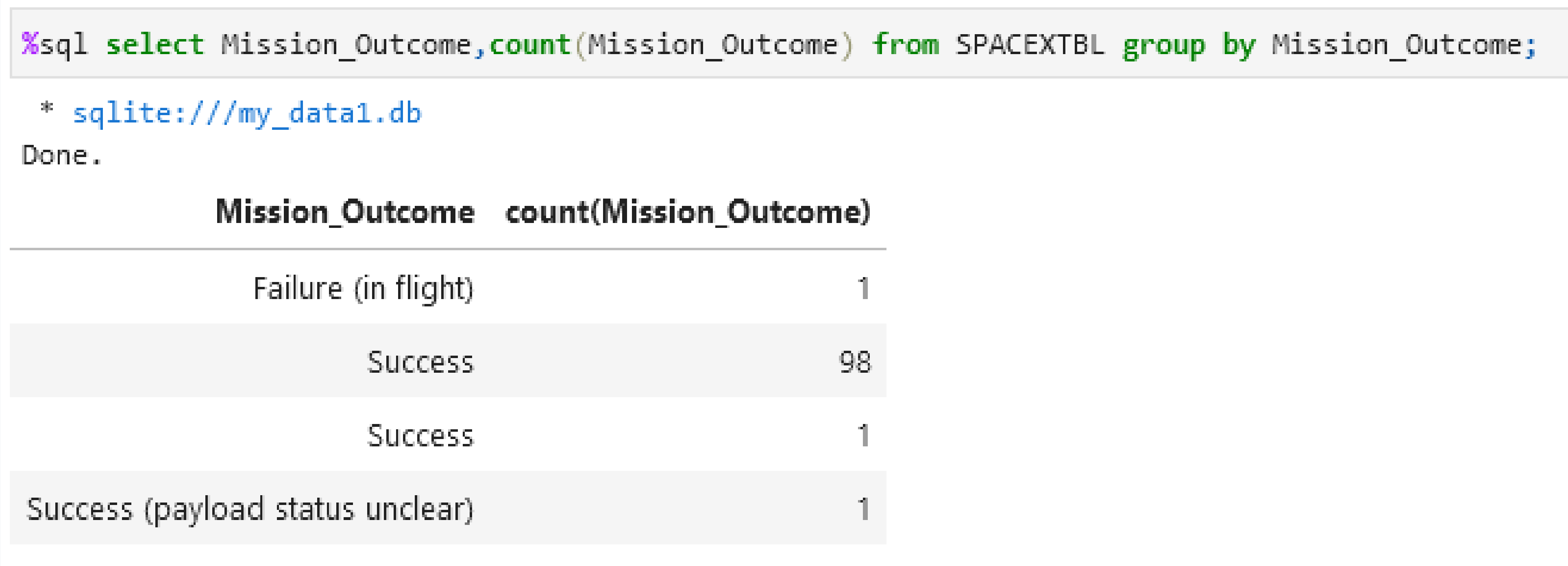
```
%%sql select Booster_Version from SPACEXTBL
where "Landing _Outcome" = 'Success (drone ship)'
and PAYLOAD_MASS__KG_ Between 4000 AND 6000;
```

```
 * sqlite:///my_data1.db
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

- Calculate the total number of successful and failure mission outcomes



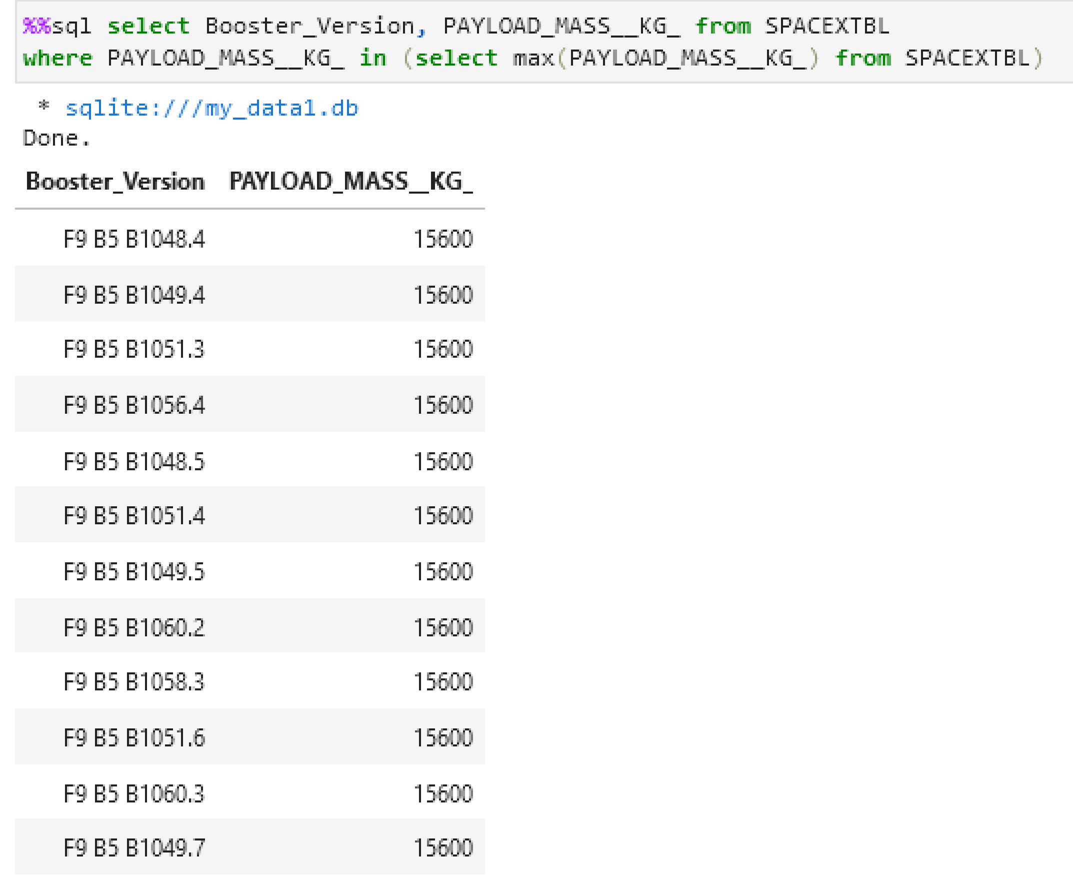
```
%sql select Mission_Outcome,count(Mission_Outcome) from SPACEXTBL group by Mission_Outcome;
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	count(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass



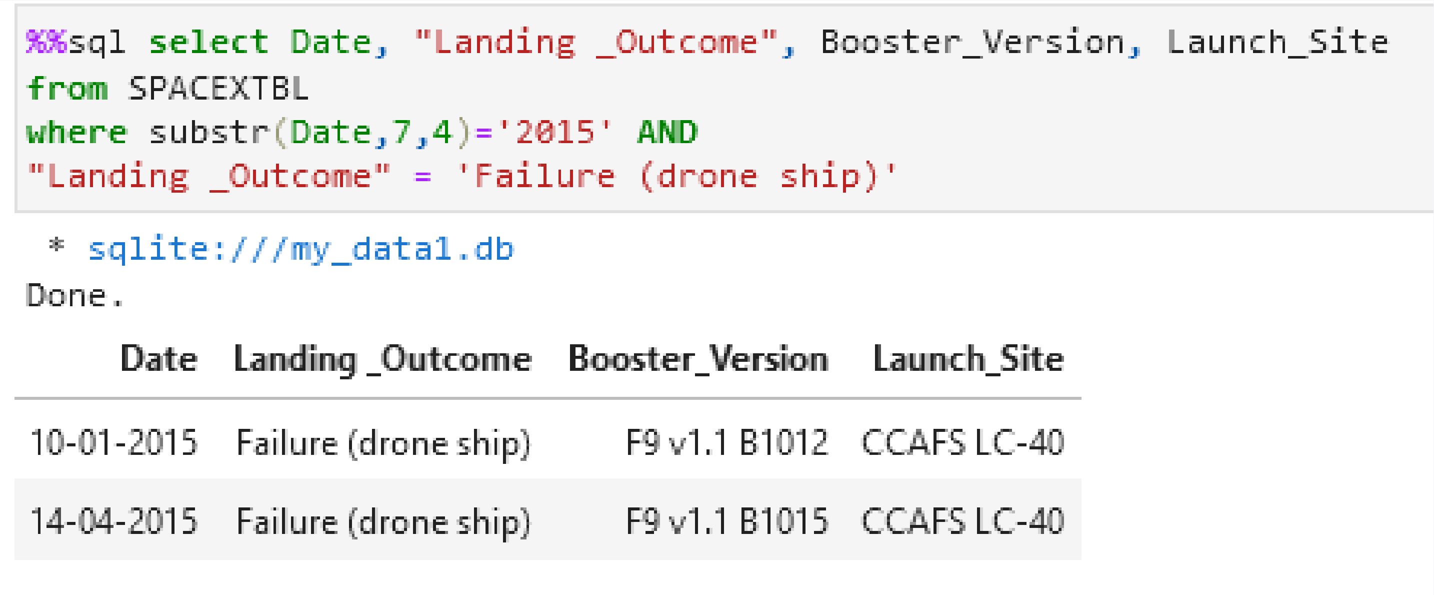
```
%%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL
where PAYLOAD_MASS__KG_ in (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

```
 * sqlite:///my_data1.db
Done.
```

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

2015 Launch Records

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



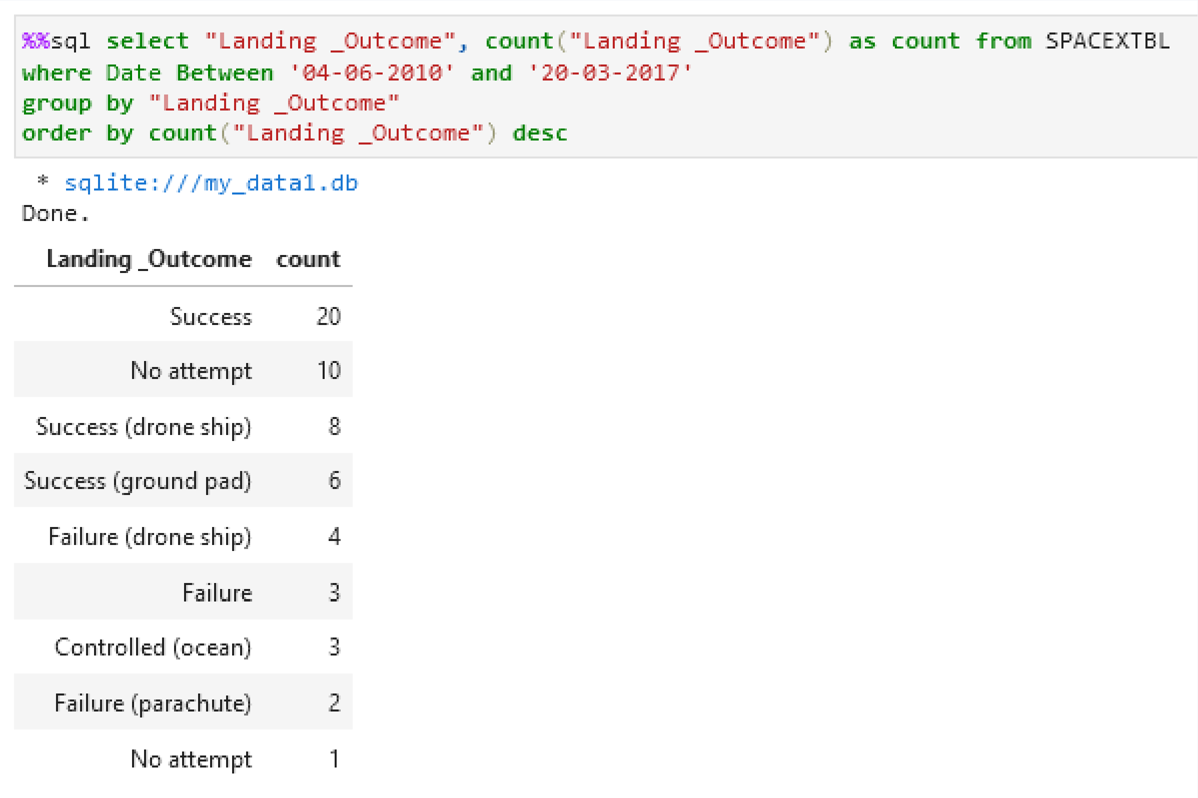
```
%%sql select Date, "Landing _Outcome", Booster_Version, Launch_Site
from SPACEXTBL
where substr(Date,7,4)='2015' AND
"Landing _Outcome" = 'Failure (drone ship)'
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Landing _Outcome	Booster_Version	Launch_Site
10-01-2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
14-04-2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- 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("Landing _Outcome") as count from SPACEXTBL
where Date Between '04-06-2010' and '20-03-2017'
group by "Landing _Outcome"
order by count("Landing _Outcome") desc
```

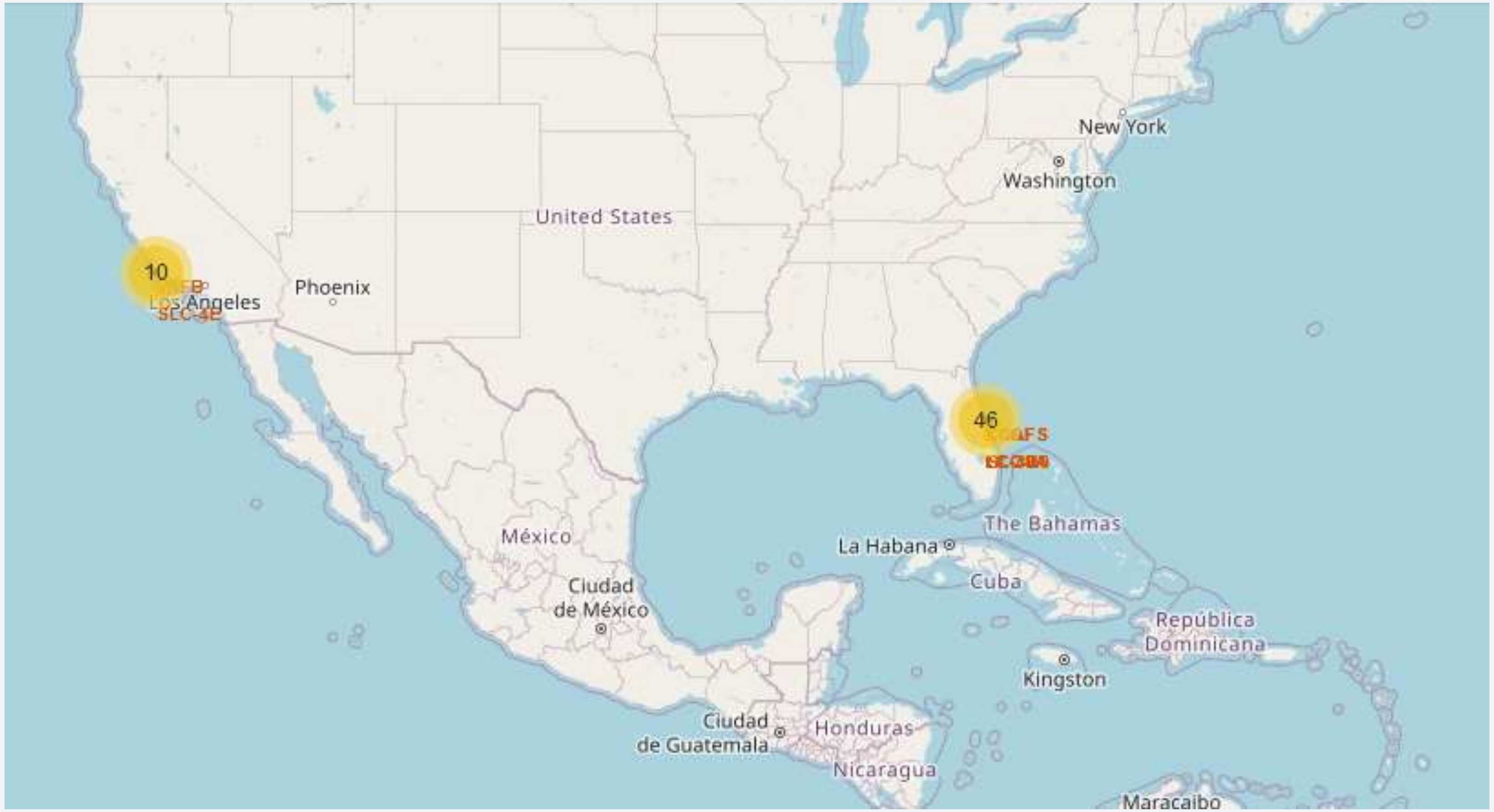
```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	count
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

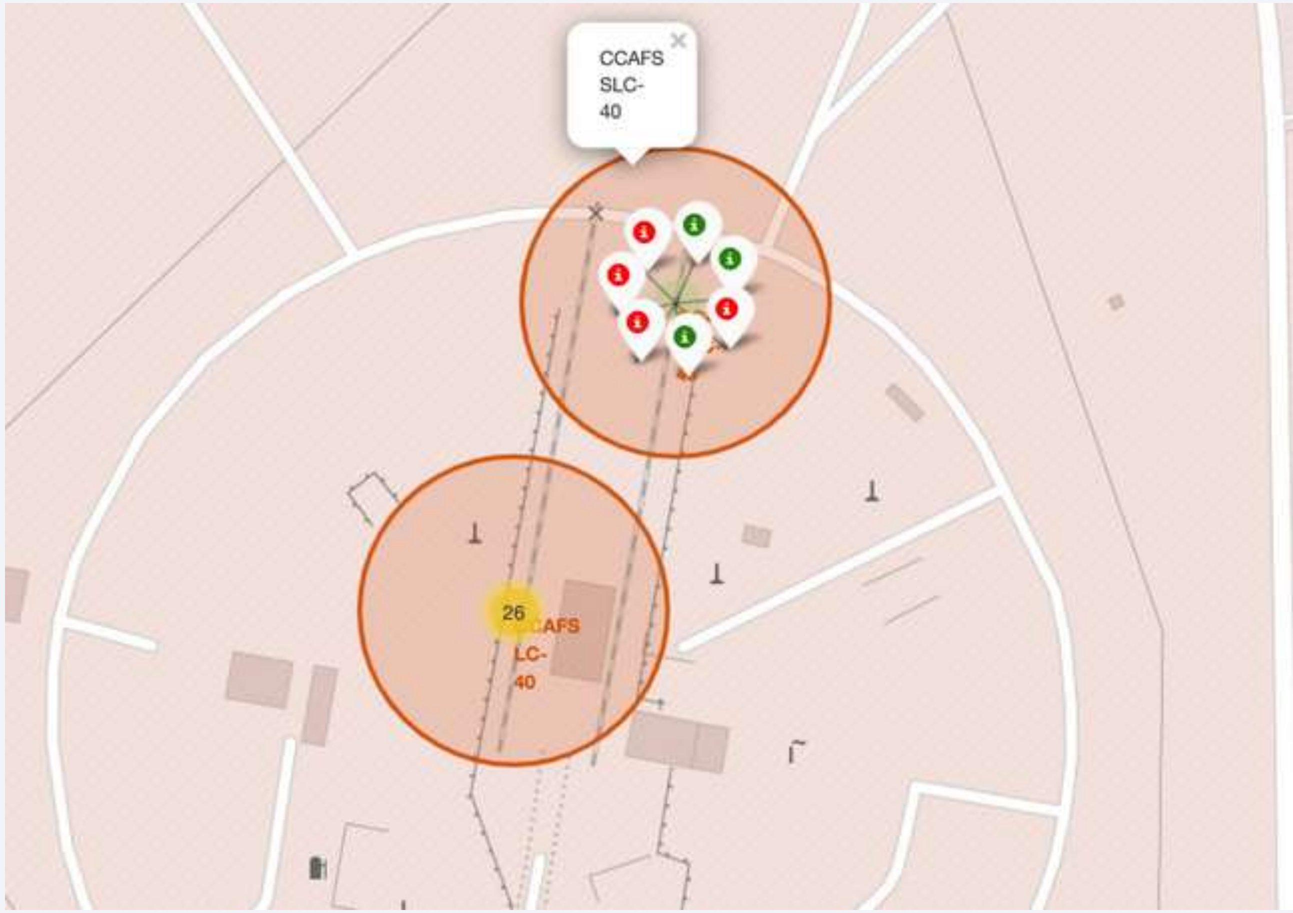
Section 3

Launch Sites Proximities Analysis

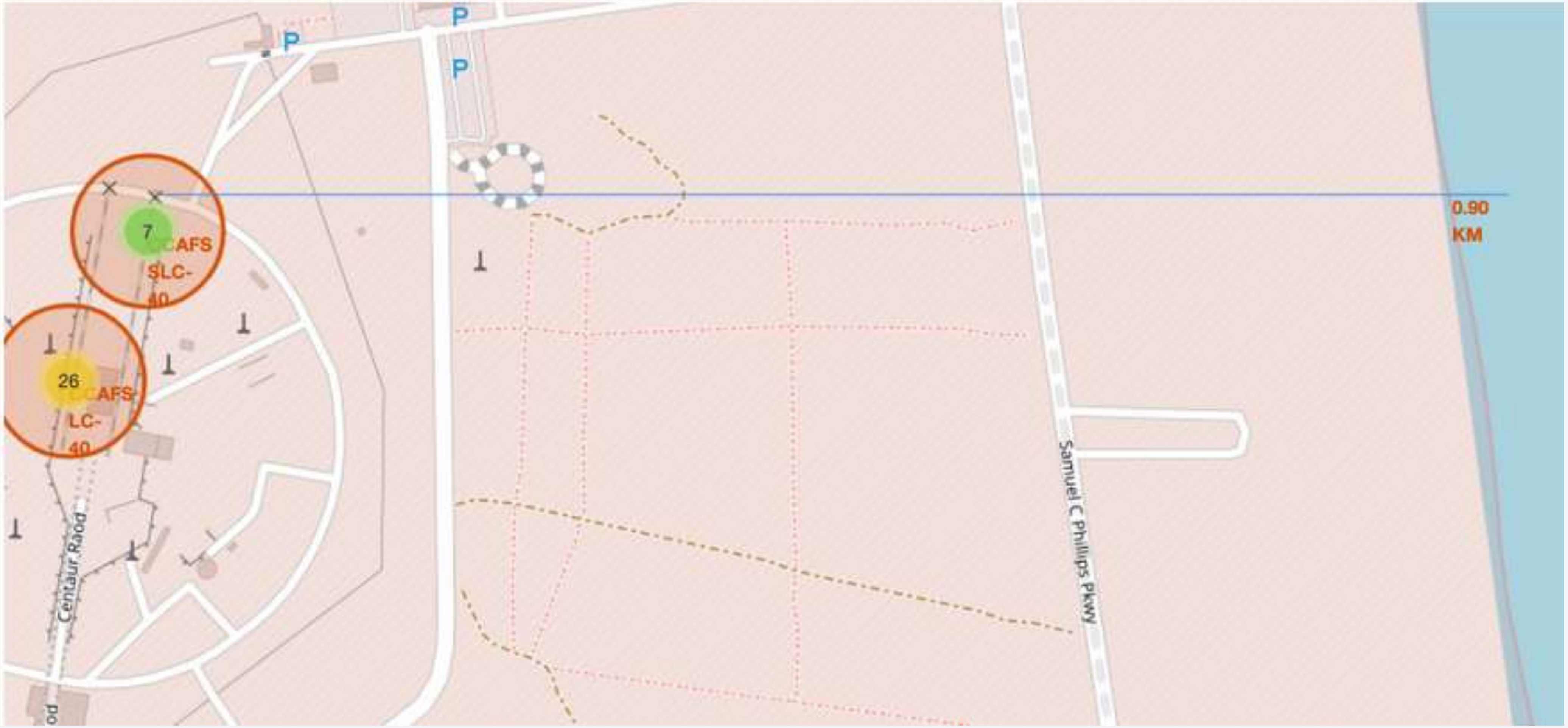
All launch sites global map markers

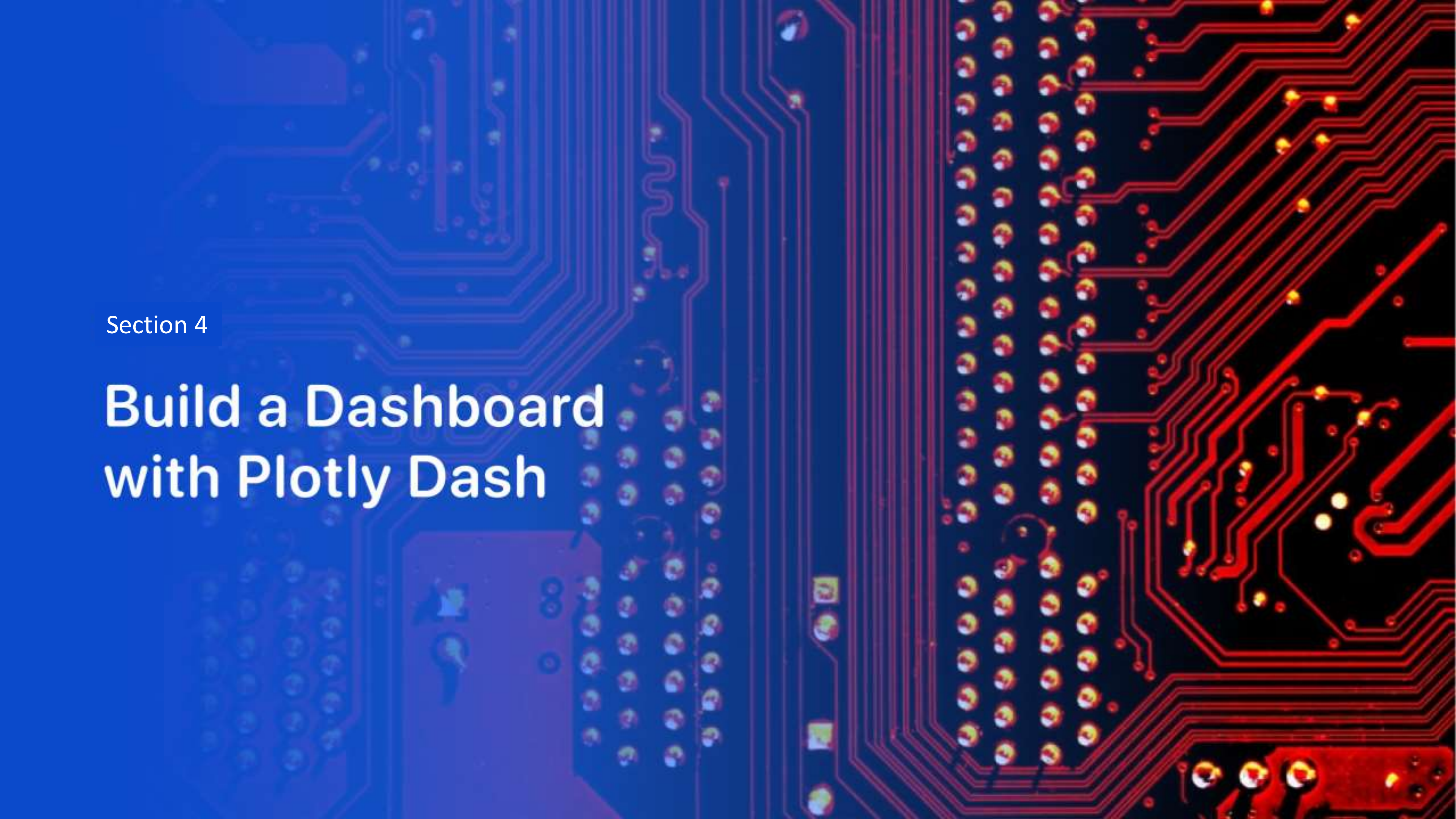


Markers showing launch sites with color labels



Launch Site distance to landmarks

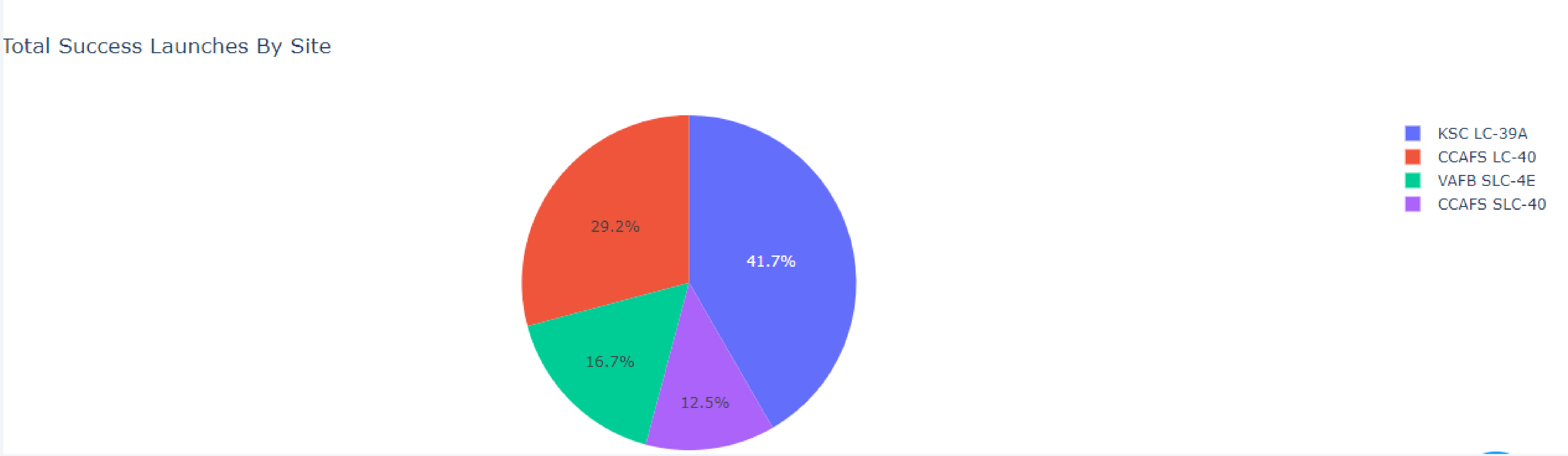




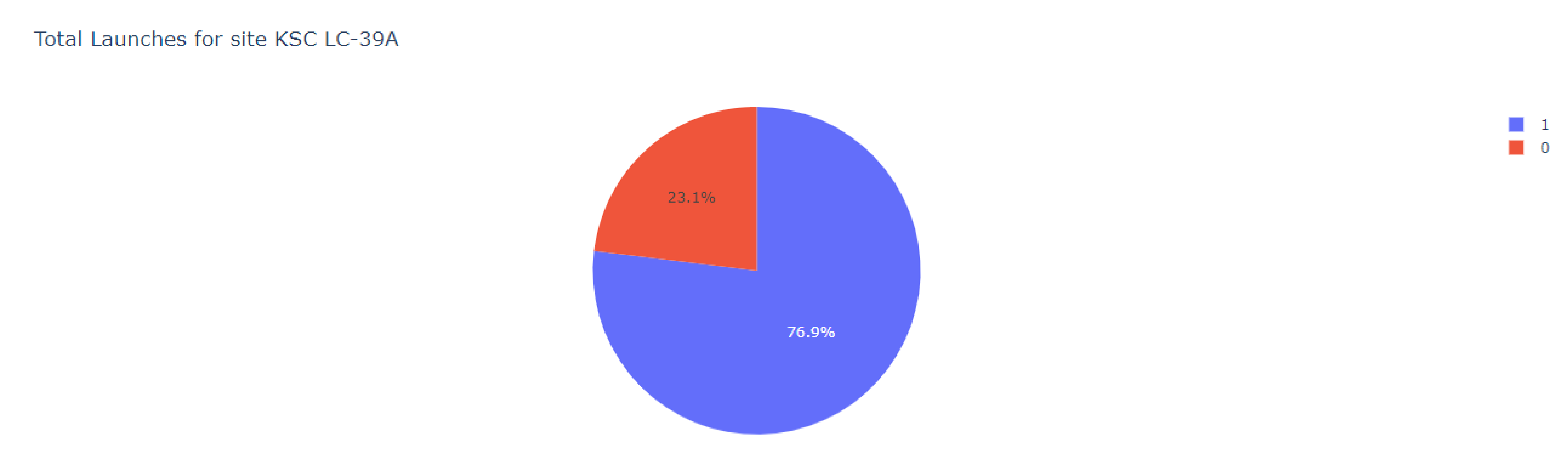
Section 4

Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site

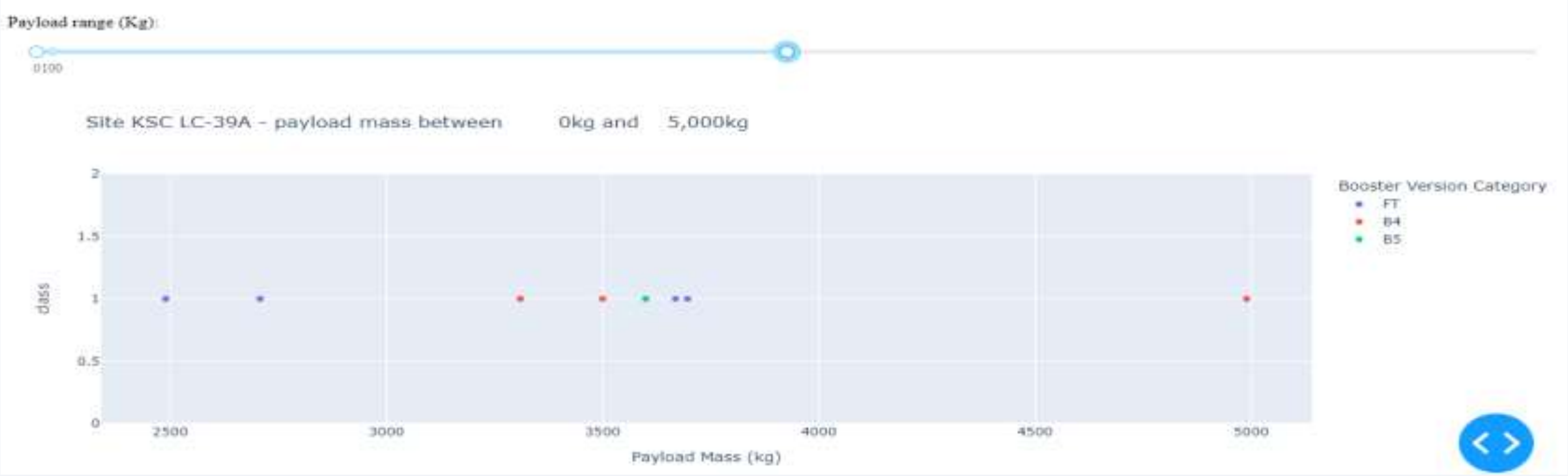


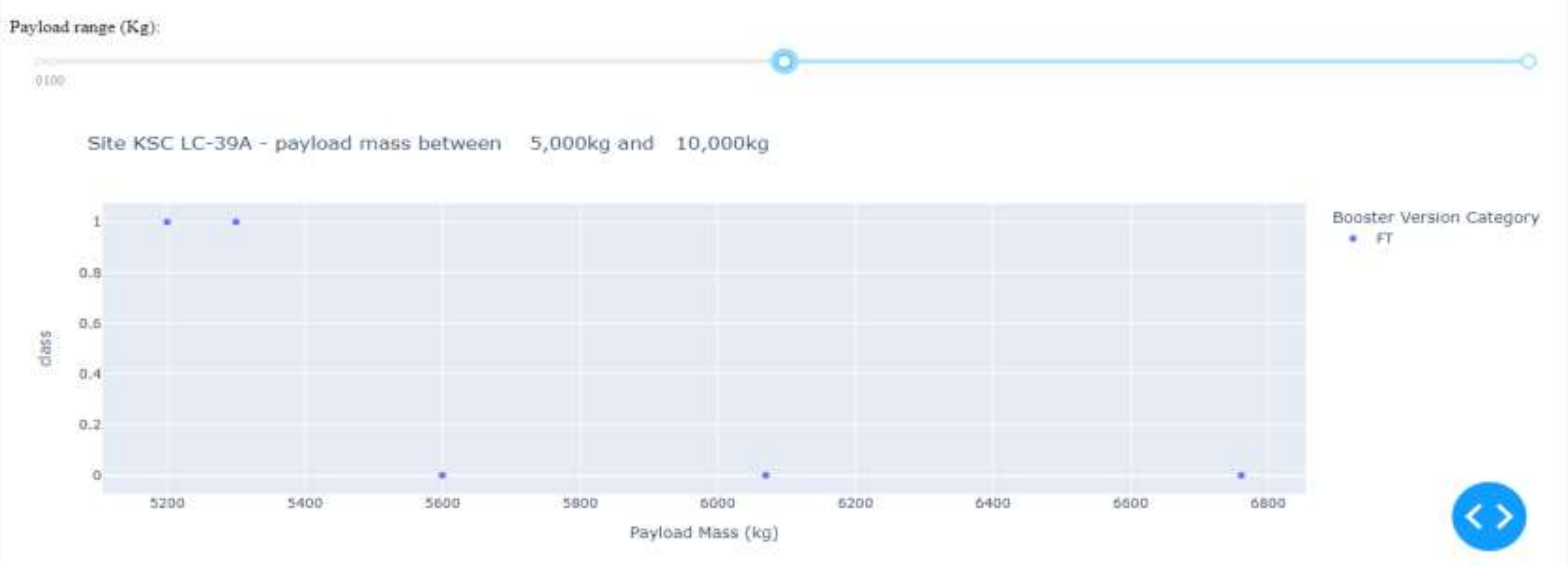
Pie chart showing the Launch site with the highest launch success ratio



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider



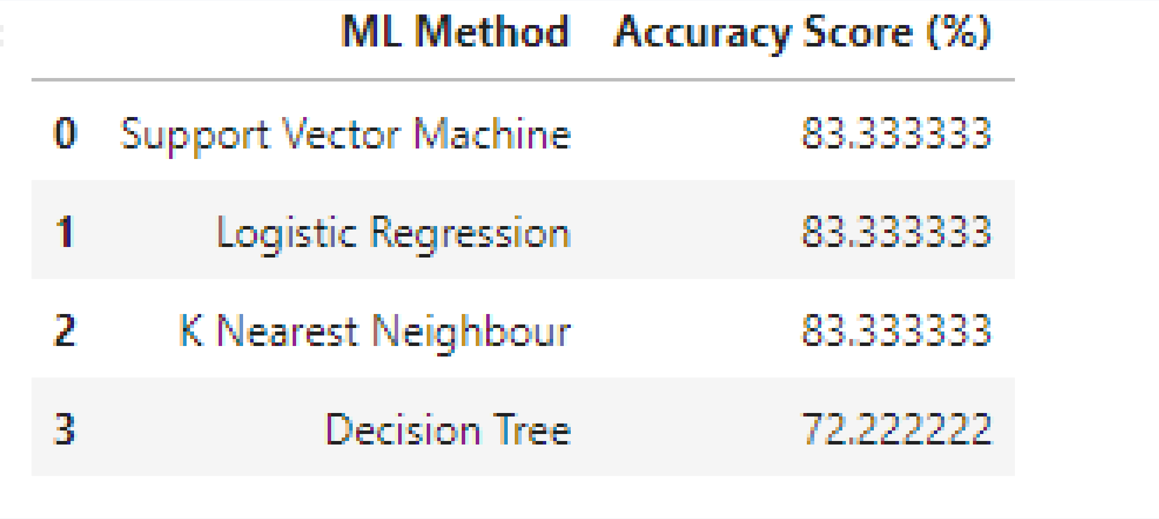


Section 5

Predictive Analysis (Classification)

Classification Accuracy

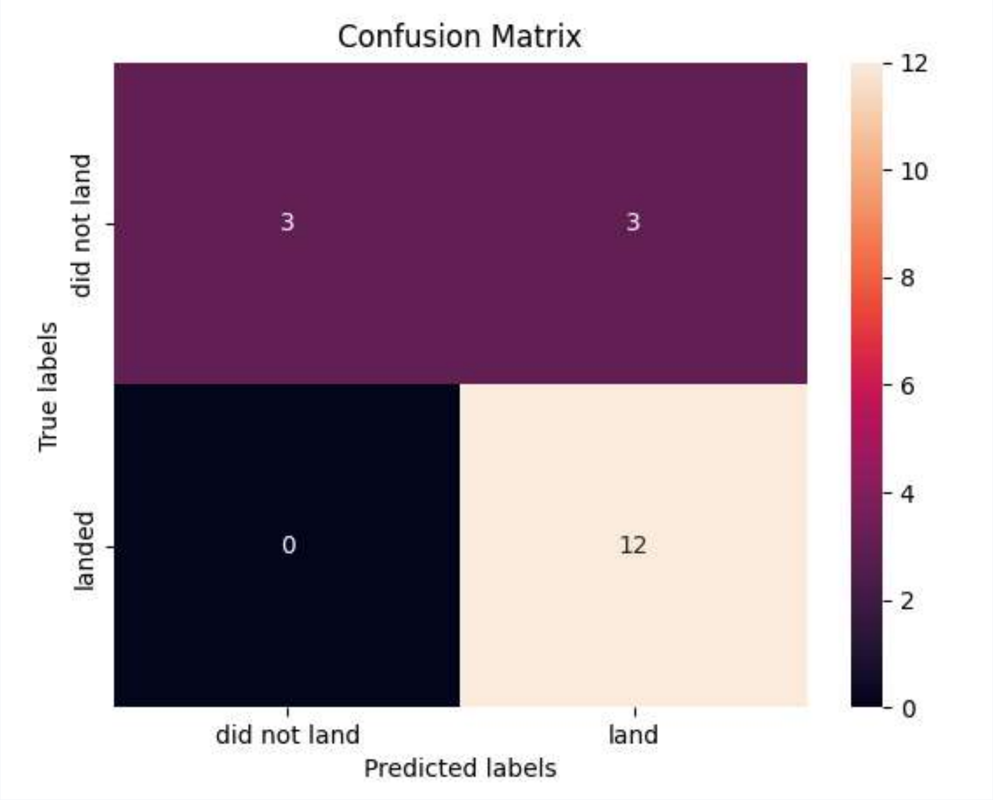
- Logistic Regression, SVM and KNN models are the best in terms of prediction accuracy (83%) for this dataset.



	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	83.333333
3	Decision Tree	72.222222

Confusion Matrix

- The confusion matrix for the SVM shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.



Conclusions

We can conclude that:

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the best success rate.
- KSC LC-39A had the most successful launches from all of sites.
- Logistic Regression, SVM and KNN models are the best in terms of prediction accuracy (83%) for this dataset.

Thank you!