IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ekaterina Elagina

13.04.2023

Outline

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

Executive Summary

- Data for the study was obtained in two ways: API access and web scraping. After the data was cleared, missing values were removed, the necessary metrics were calculated, and additional variables were created. Information was obtained with the help of EDA and an analysis of the geographical location of the launch sites was carried out.
- The results of the study showed that the most favorable location for launch sites is near the railway, highway and coastline for the most convenient delivery of components and safety buffer zone, and for safety, it is extremely important that the launch site is located far from the city. Near the equator line, which helps save fuel and reach orbit faster.

Introduction

- The aim of this research is to predict the successful landing of the Falcon 9 first stage, which is a crucial component of SpaceX's rocket launches. The cost of a launch is significantly reduced by reusing the first stage, making it a critical factor in determining the overall cost.
- Therefore, accurate predictions of the first stage landing can provide crucial information for alternate companies to determine the cost of their launch bids against SpaceX.
- This report will provide background information on the significance of the Falcon 9 first stage and the context in which the project is situated, as well as detailing the problem being addressed and the proposed solution.

Section 1

Methodology

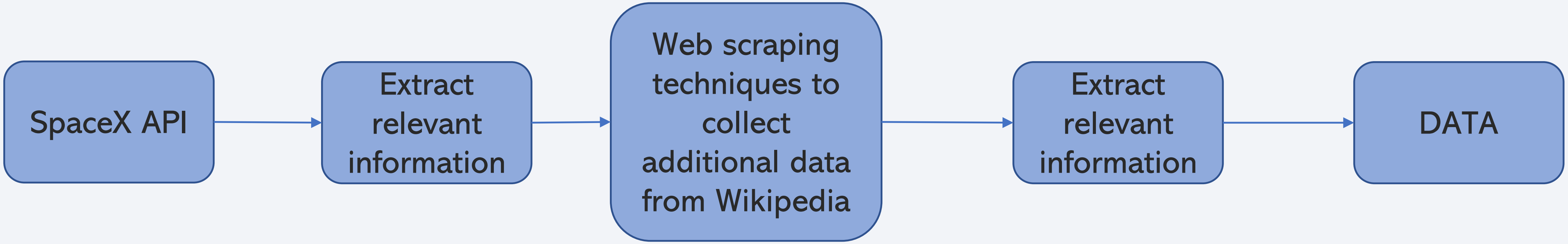
Methodology

Executive Summary

- Data collection methodology:
  - data for analysis was collected using the API and additional data using web scraping
- Data wrangling:
  - missing values were removed from the data, the data type was checked, the main metrics were calculated, the target variable was created
- Exploratory data analysis (EDA) using visualization and SQL was performed
- Interactive visual analytics using Folium and Plotly Dash were performed
- Predictive analysis using classification models is including:
  - building different models using libraries, choosing the best parameters, checking the model on test data

Data Collection

- The data for this research was collected through two primary methods: utilizing the SpaceX API and web scraping information from Wikipedia.



SpaceX API

Extract relevant information

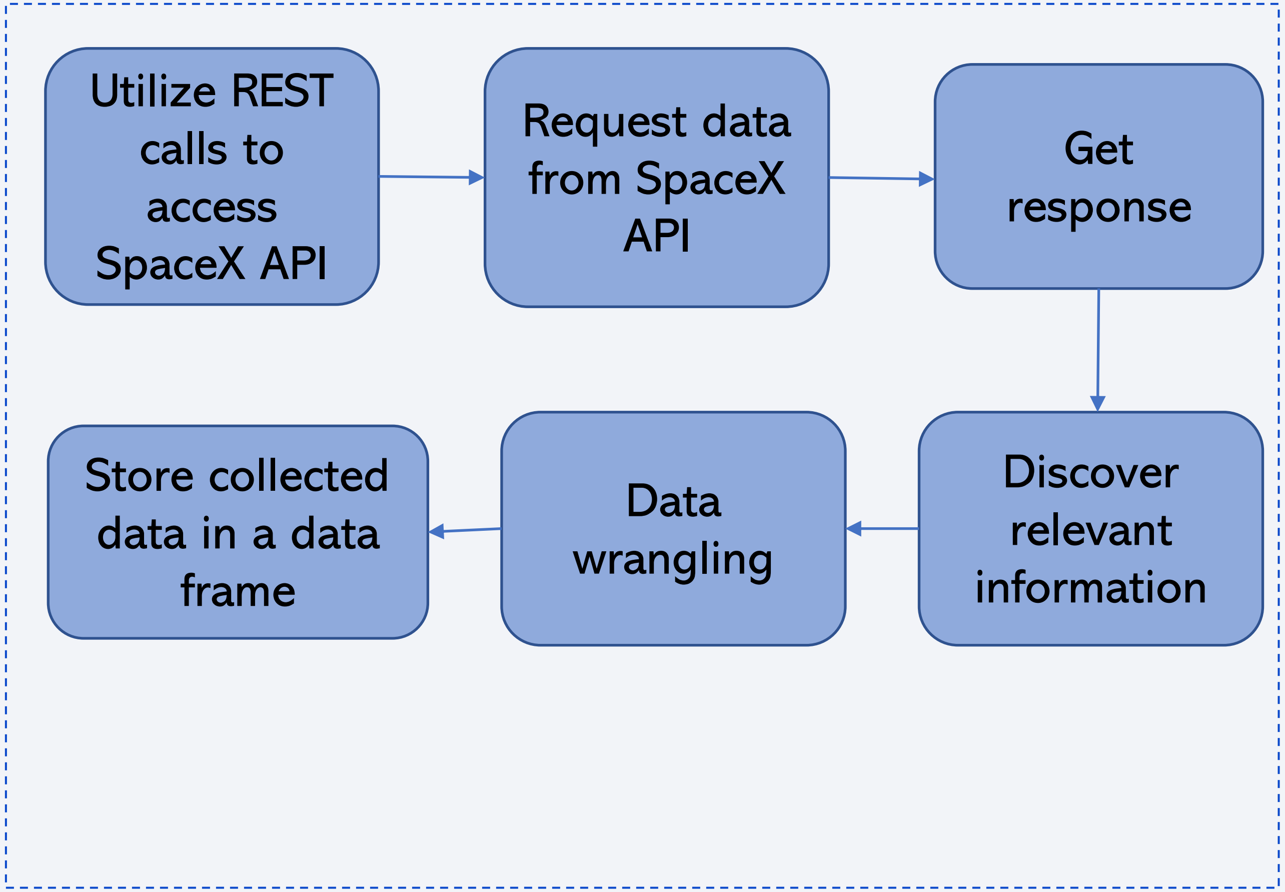
Web scraping techniques to collect additional data from Wikipedia

Extract relevant information

DATA

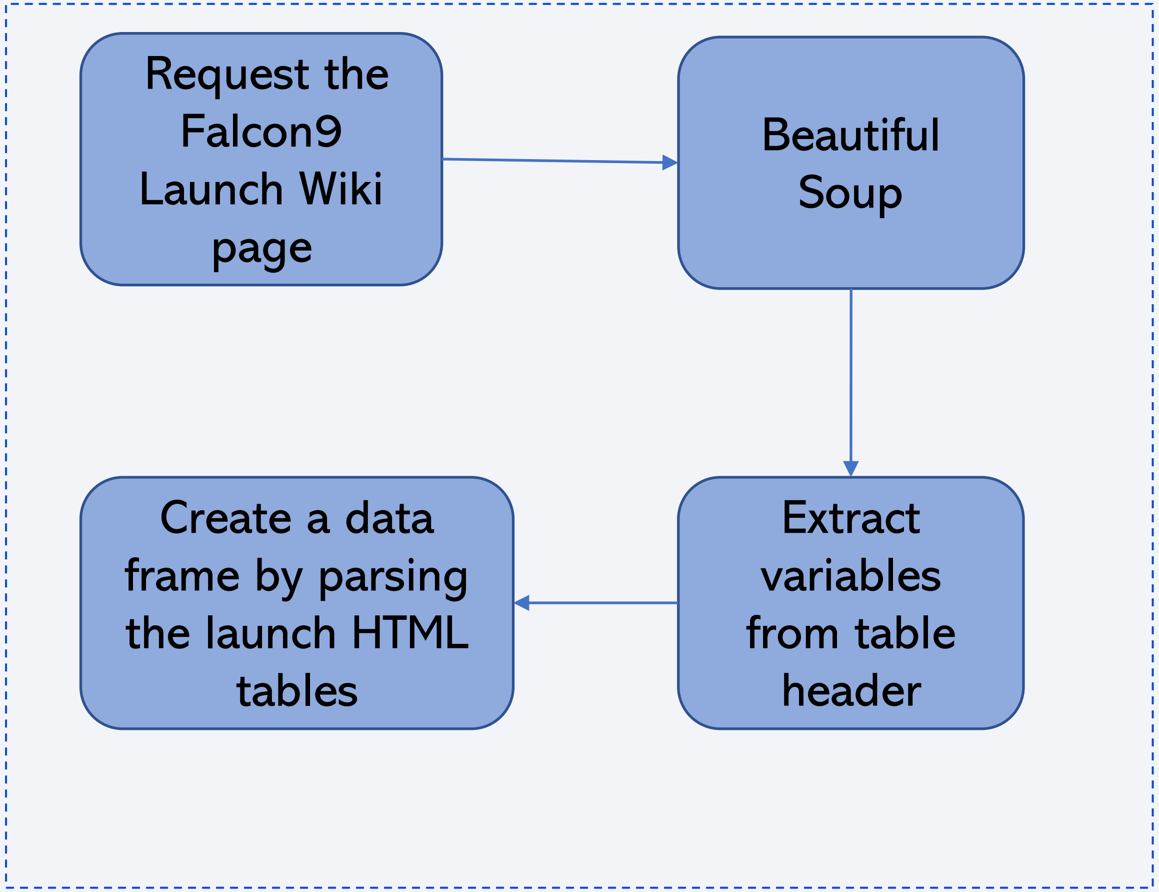
Data Collection – SpaceX API

- Utilize REST calls to access SpaceX API
- Request and parse the SpaceX launch data using the GET request
- Filter the data frame to only include Falcon 9 launches
- Data Wrangling
- GitHub


Utilize REST calls to access SpaceX API
Request data from SpaceX API
Get response
Discover relevant information
Data wrangling
Store collected data in a data frame

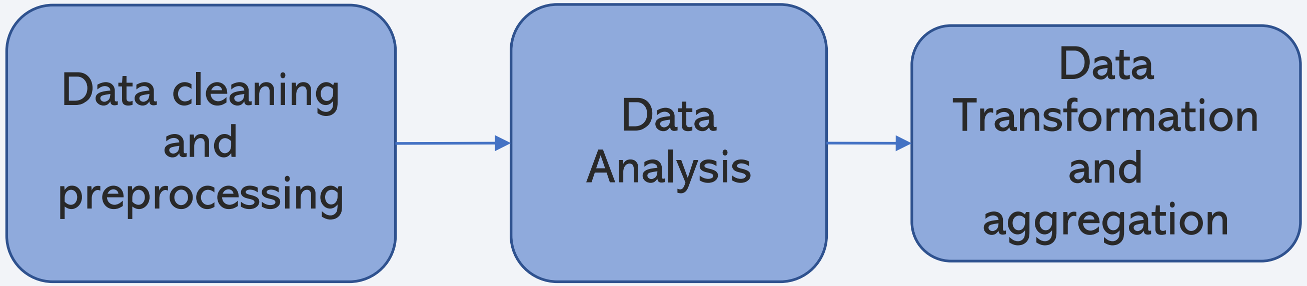
Data Collection - Scraping

- Utilizing web scraping techniques to access Wikipedia pages
- Create a BeautifulSoup object from the HTML response
- Extract all column names from the HTML table header
- Storing collected data in a data frame
- GitHub


Request the Falcon9 Launch Wiki page
Beautiful Soup
Extract variables from table header
Create a data frame by parsing the launch HTML tables

Data Wrangling

- Data cleaning and preprocessing
- Data analysis (missing values, data types)
- Data transformation and aggregation
- Calculation of key metrics (number of launches on each site, number and occurrence of each orbit, number and occurrence of mission outcome per orbit type)
- Created a new column called "Class" in the dataset, where the value 0 is assigned if the launch did not land successfully, and the value 1 is assigned if the launch landed successfully
- GitHub



Data cleaning and preprocessing

Data Analysis

Data Transformation and aggregation

EDA with Data Visualization

- During the research, we utilized several visualization techniques to explore the relationships between different variables:
  - scatter plots to identify any correlations between variables,
  - bar plot were used to visualize the relationship between the success rate of each orbit type.
  - line plot to visualize the yearly trend of the launch success rate.

  These visualizations enabled to gain insights into the data and draw meaningful conclusions.

- GitHub

EDA with SQL

Performed SQL queries:

- Display 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 the average payload mass carried by booster version F9 v1.1.
- List the date when the first successful landing outcome on a ground pad was achieved.
- List the names of the boosters which had success in drone ship and had 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 carried the maximum payload mass.
- List the records displaying the month names, failure landing_outcomes in drone ship, booster versions, and launch sites for the months in year 2015.
- Rank the count of successful landing outcomes between the dates 04-06-2010 and 20-03-2017 in descending order.

- GitHub

Build an Interactive Map with Folium

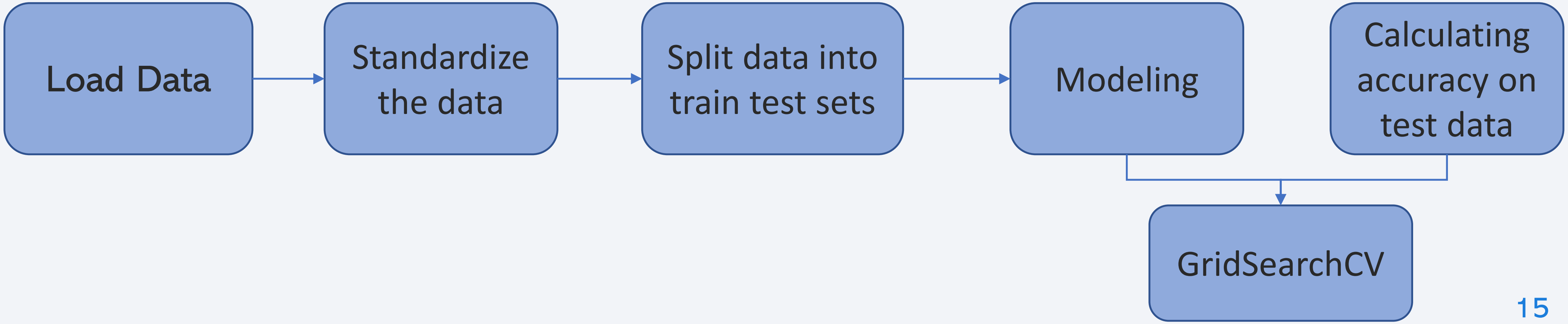
- In order to better understand the locations of launch sites and their success rates, we propose visualizing the launch sites on a map with markers indicating success or failure. Marker clusters can be used to simplify a map with many markers in the same coordinate. This will allow us to easily identify which sites have high success rates. Additionally, we can use lines to demonstrate the distance to important objects and analyze the proximity of the launch sites.
- GitHub

Build a Dashboard with Plotly Dash

- The dashboard includes a dropdown list to select the Launch Site, enabling users to filter data by site.
- Additionally, a pie chart will be added to visualize the total number of successful launches across all sites.
- A slider will also be included to filter data by Payload Range.
- Lastly, a scatter chart will be used to show the correlation between Payload and Launch Success, providing a clear understanding of how these variables are related.
- GitHub

Predictive Analysis (Classification)

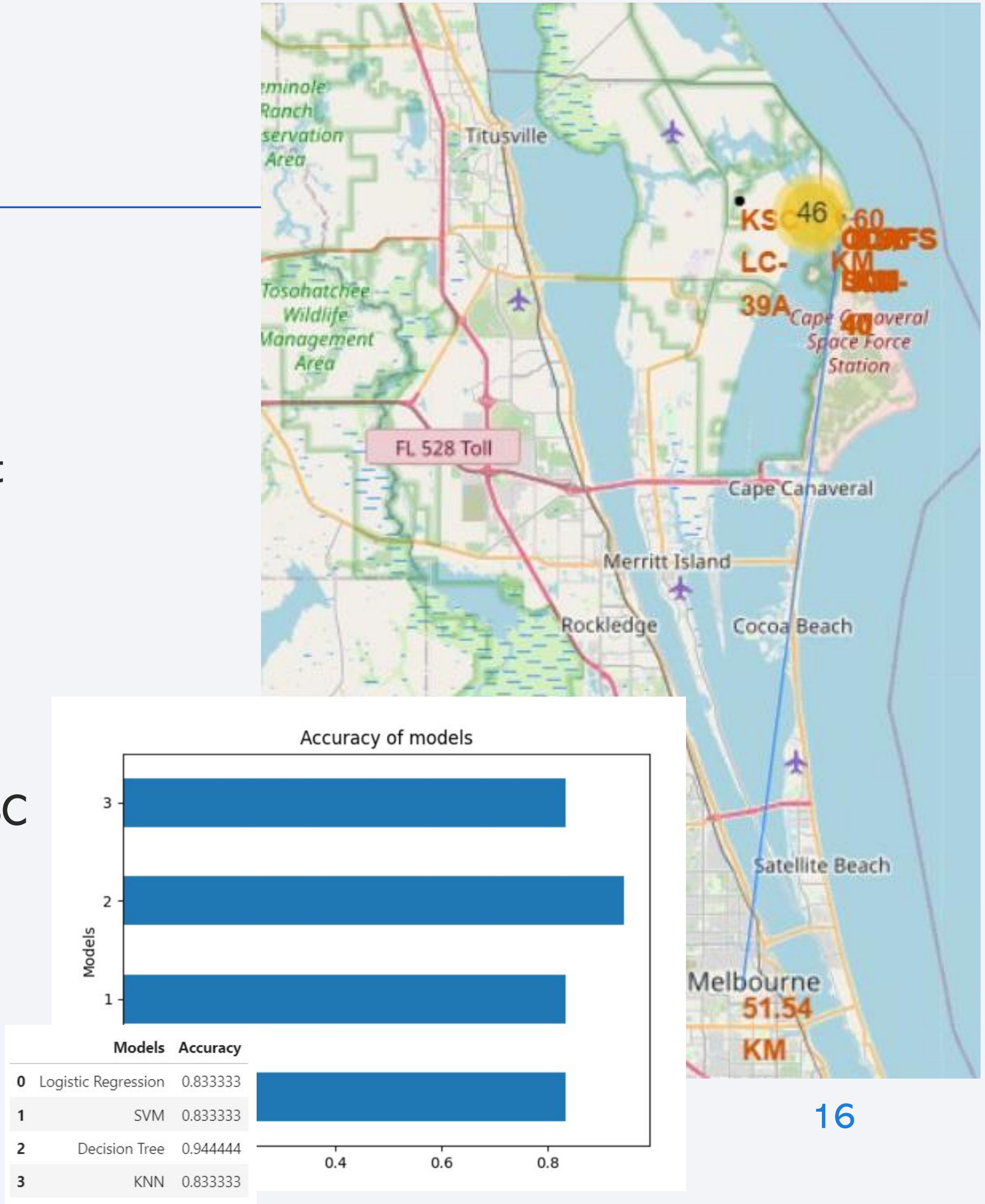
- In this slide presentation, we explore different classification models including Logistic Regression, SVM, Decision Tree, and KNN. We use GridSearchCV to find the best hyperparameters for SVM, Classification Trees, and Logistic Regression models. The results of this analysis can be used to determine which model performs the best for our classification task.
- GitHub


Load Data
Standardize the data
Split data into train test sets
Modeling
Calculating accuracy on test data
GridSearchCV

Results

The most favorable location for launch sites is:

- near the railway, highway and coastline
- far from the city;
- near the equator line, which helps save fuel and reach orbit faster;
- The best model for data is the decision tree;
- site with the largest successful launches is KSC LC-39A;
- site has the highest launch success rate is KSC LC-39A;
- payload ranges with the highest launch success rate are KSC LC-39A and CCAFS LC-40;
- payload ranges with the lowest launch success rate are CCAFS SLC-40 and VAFB SLC-4E;
- F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) with the highest launch success rate is FT



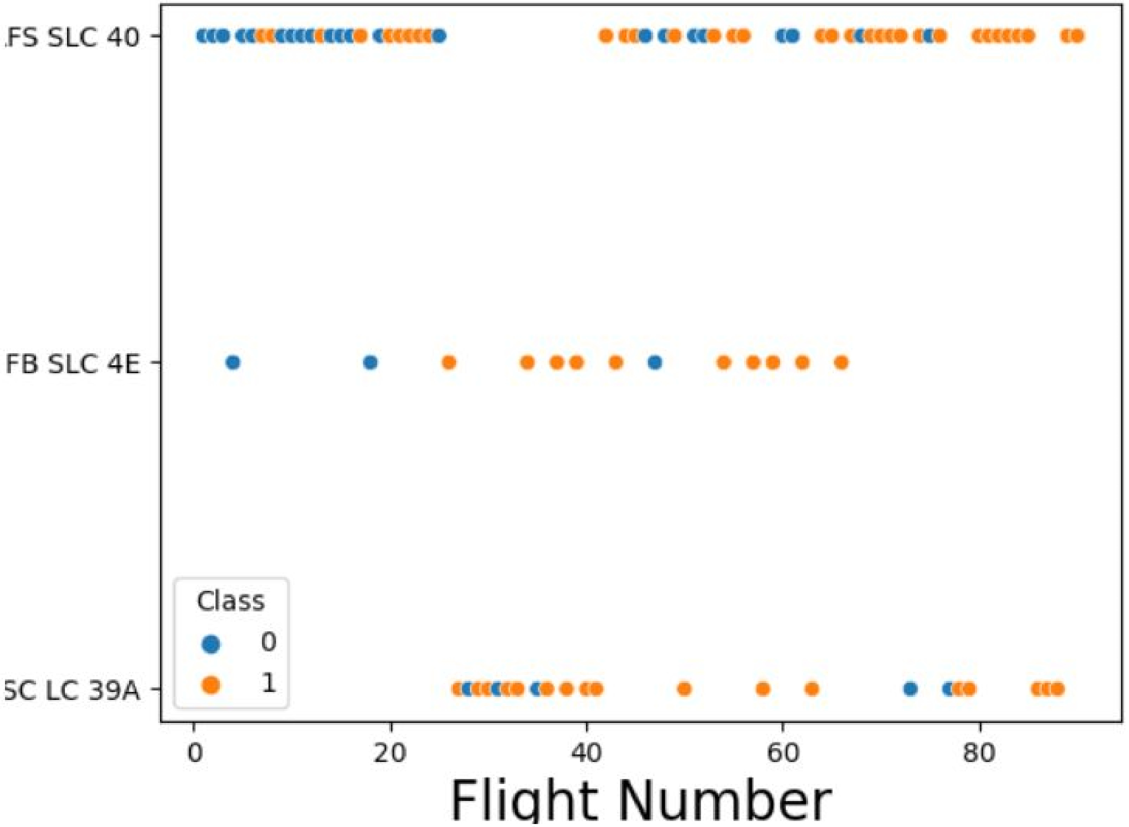
	Models	Accuracy
0	Logistic Regression	0.833333
1	SVM	0.833333
2	Decision Tree	0.944444
3	KNN	0.833333

Section 2

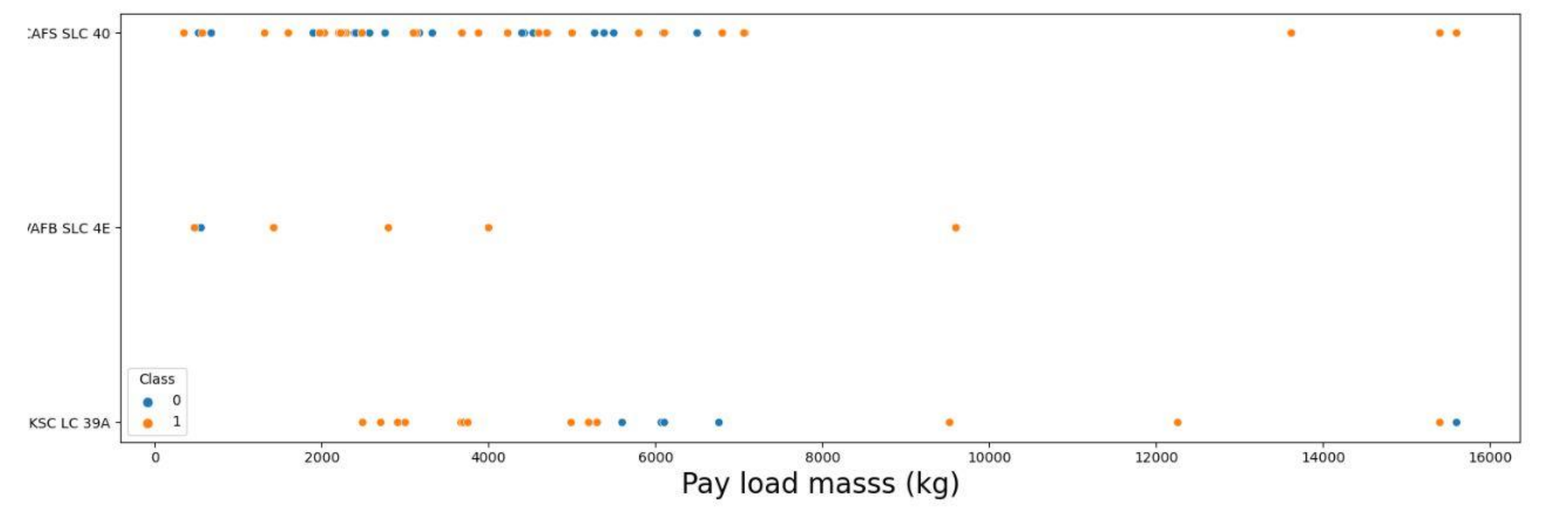
Insights drawn from EDA

Flight Number vs. Launch Site

- On the scatter plot, you can see that the higher the flight number, the more successful the launch on the launch site (FB SLC 4E, FS SLC 40)
- Also, FB SLC 4E has more successful launches then failed one.



Payload vs. Launch Site



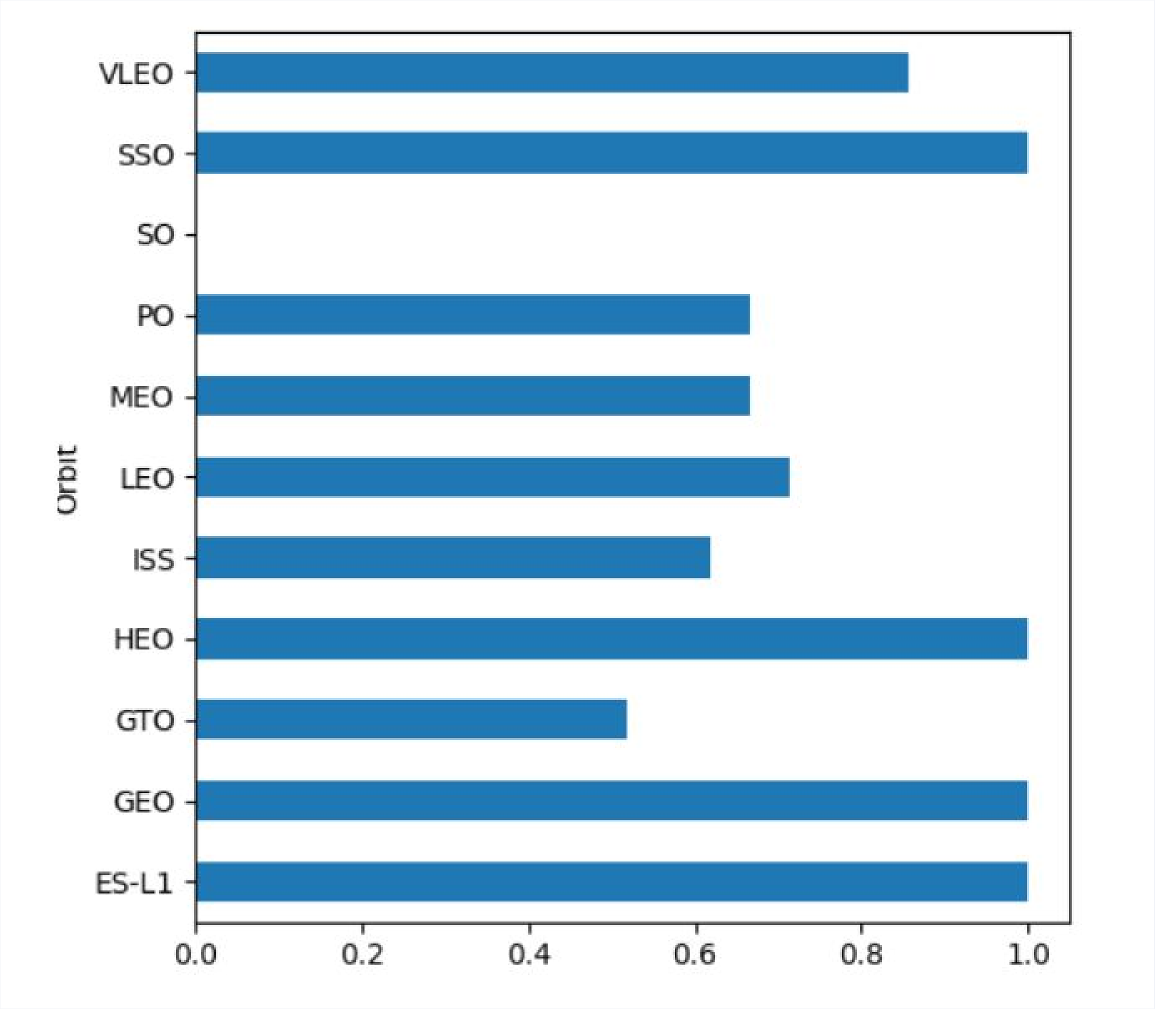
The most common pay load mass is in the range from 1000 to 6500.

For KSC LS 39A, unsuccessful launches were with a pay load mass in the range from 5500 kg to 6500 kg.

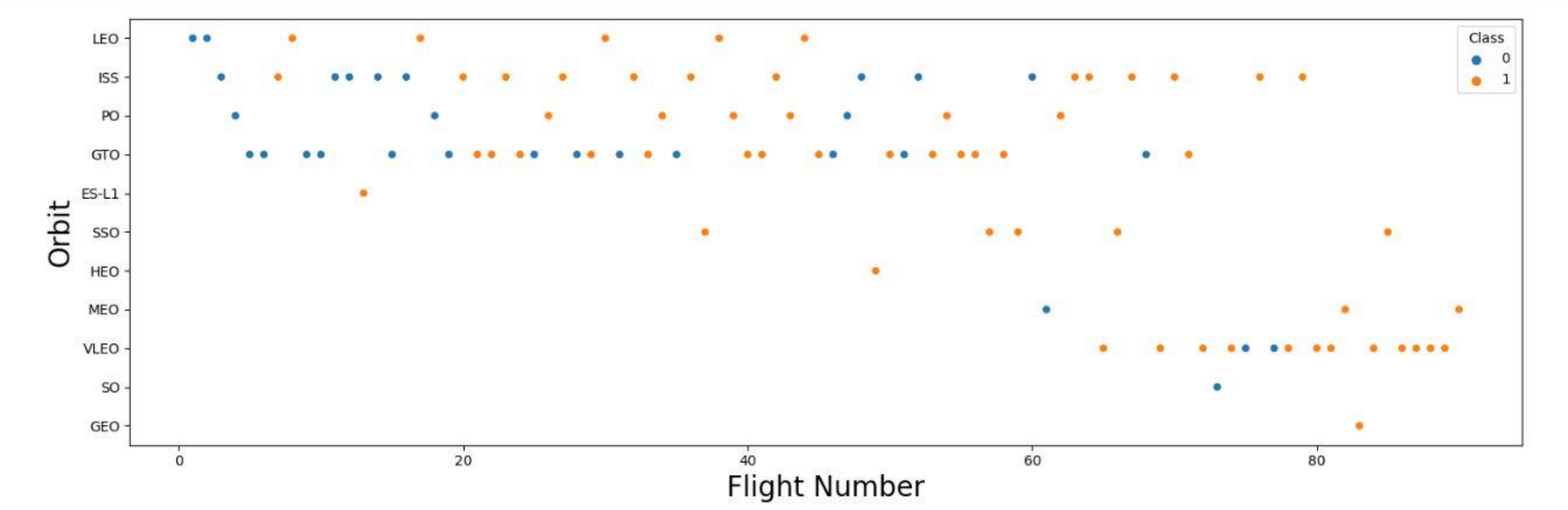
Conclusion: the higher the pay load mass, the more successful the launch

Success Rate vs. Orbit Type

- The slide shows a bar plot of the probability of a successful launch in each of the Orbit type.
- As we can see only four of all Orbit type have probability equal 1 and five with probability more than 60%.

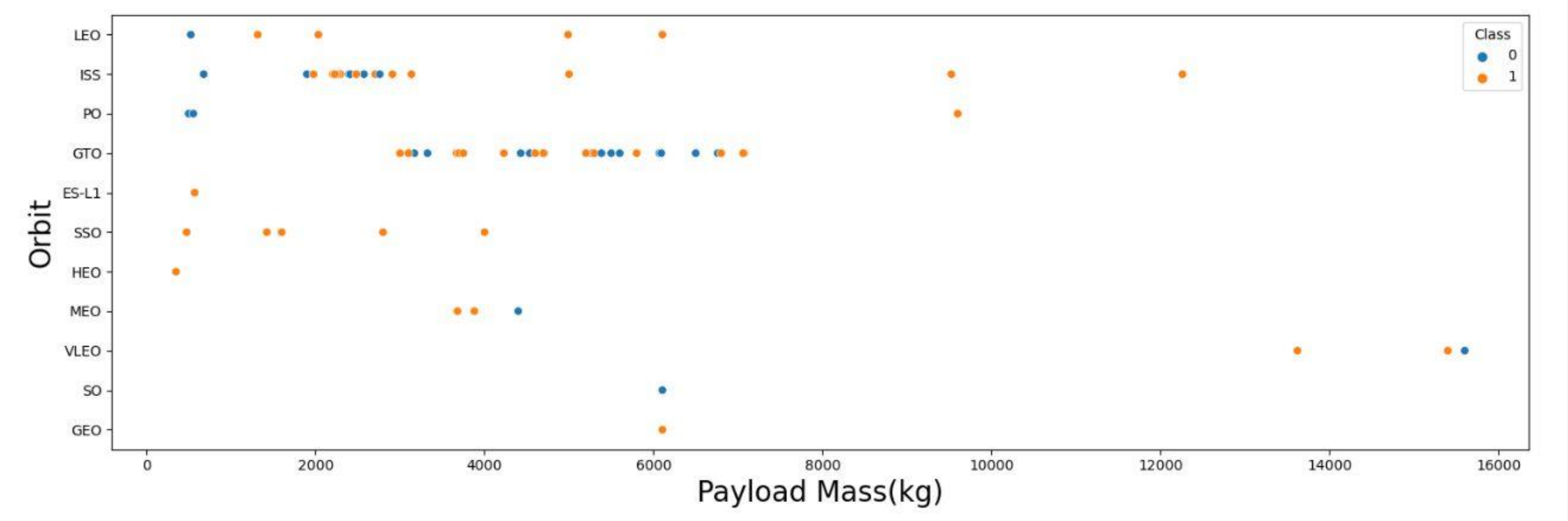


Flight Number vs. Orbit Type



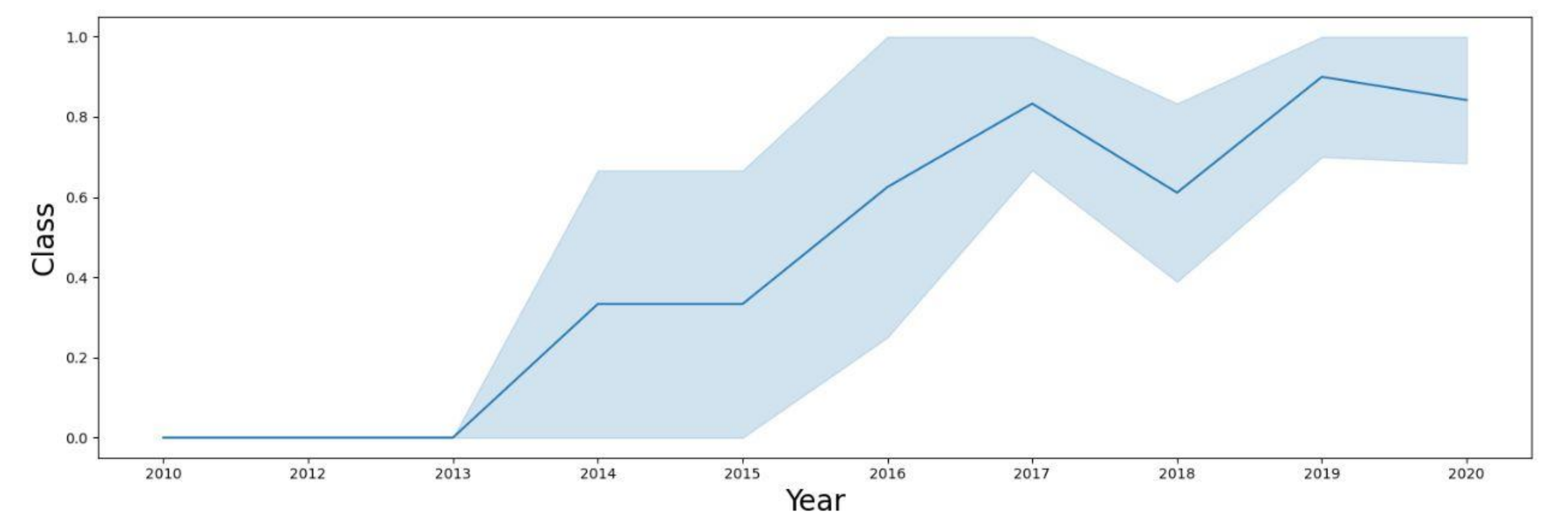
- We can see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here

Launch Success Yearly Trend

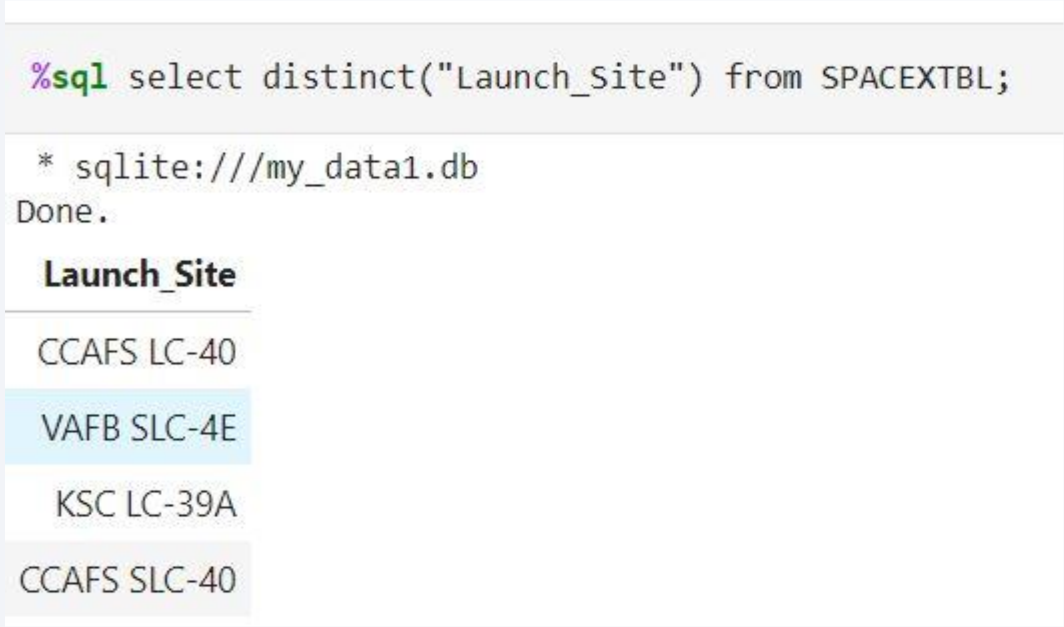


You can observe that the success rate since 2013 kept increasing till 2020

All Launch Site Names

This slide represents the unique launch sites:

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

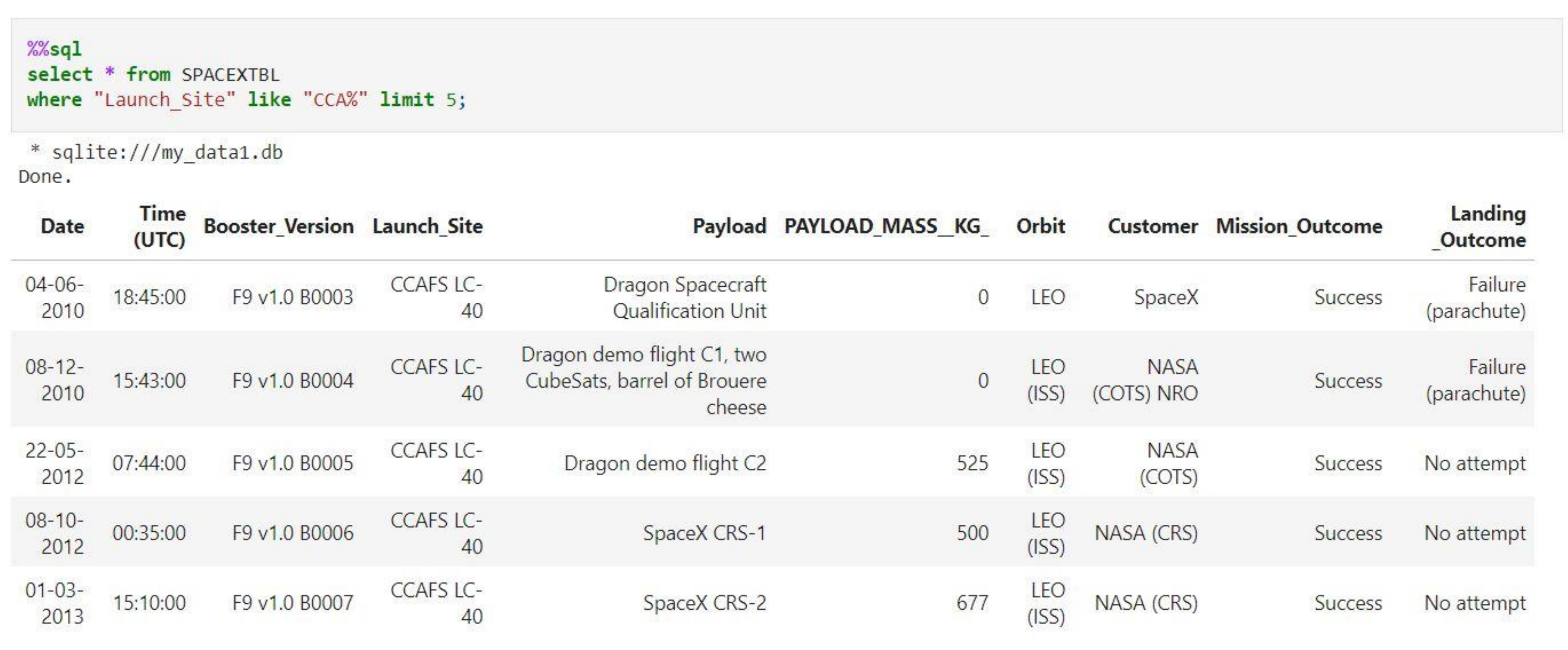


```
%sql select distinct("Launch_Site") from SPACEXTBL;
```

```
 * 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'



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

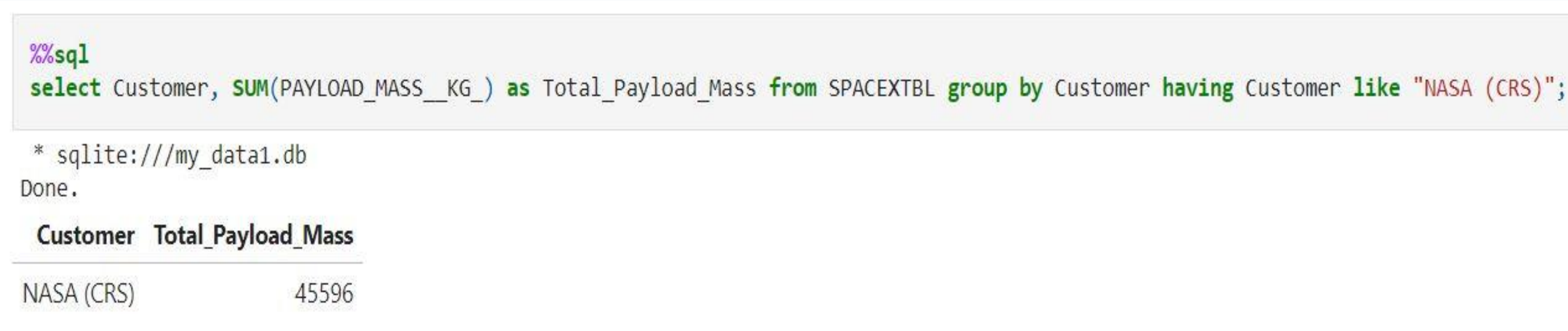
```
 * sqlite:///my_data1.db
Done.
```

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

- The slide shows the result of the query, which displays only the first 5 records, where the name of the site starts with ‘CCA’

Total Payload Mass

- This query result shows a total payload mass:



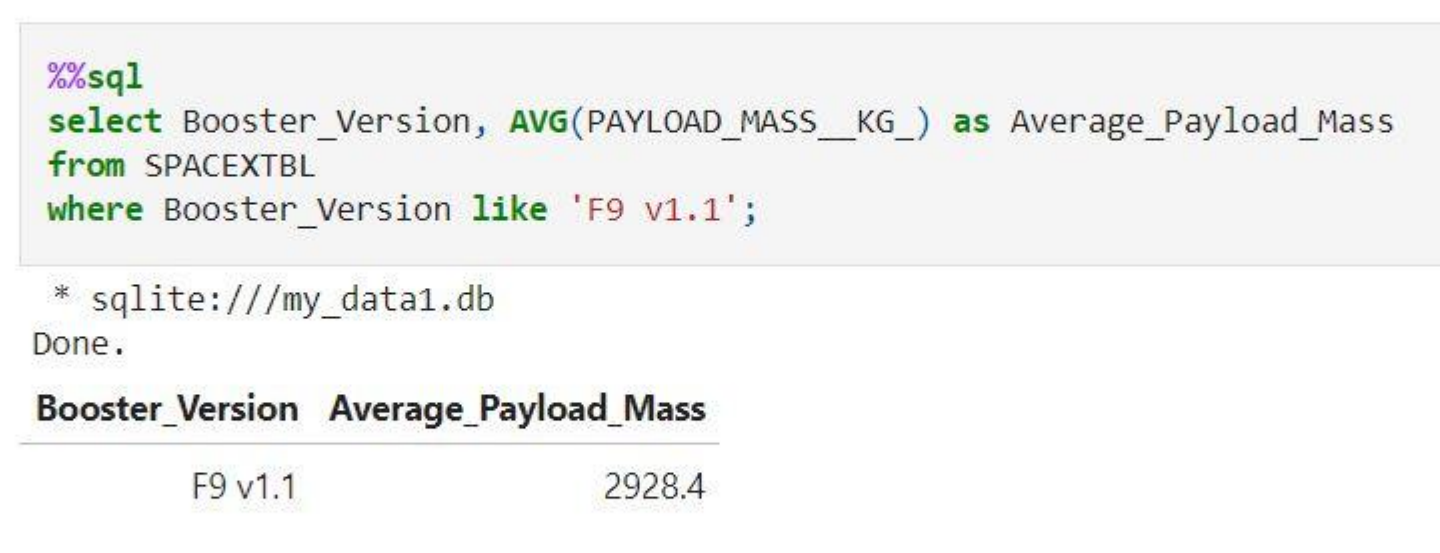
```
%%sql
select Customer, SUM(PAYLOAD_MASS__KG_) as Total_Payload_Mass from SPACEXTBL group by Customer having Customer like "NASA (CRS)";
```

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

Customer	Total_Payload_Mass
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- Query result present the average payload mass where version of booster is F9 v1.1:



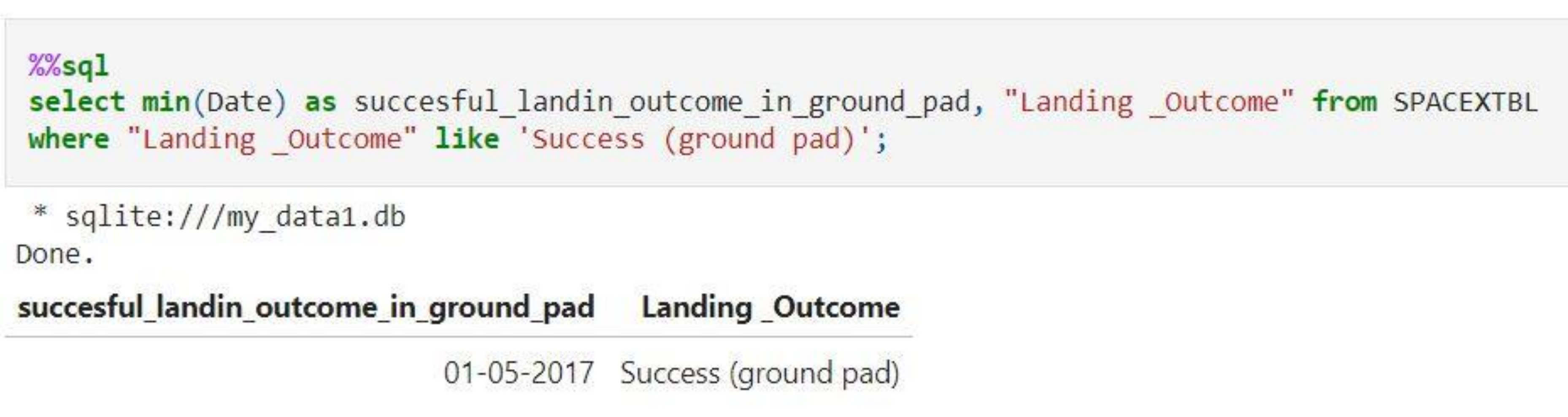
```
%%sql
select Booster_Version, AVG(PAYLOAD_MASS__KG_) as Average_Payload_Mass
from SPACEXTBL
where Booster_Version like 'F9 v1.1';
```

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

Booster_Version	Average_Payload_Mass
F9 v1.1	2928.4

First Successful Ground Landing Date

- Query result that present the successful landing on ground pad:



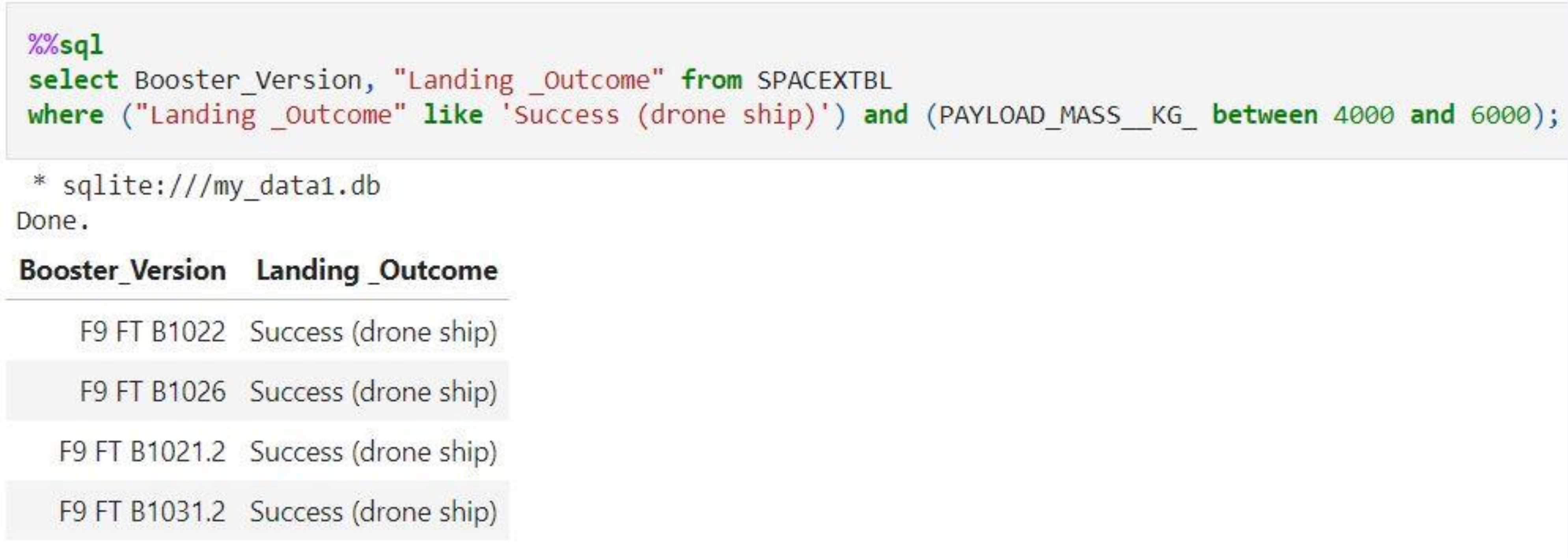
```
%%sql
select min(Date) as succesful_landin_outcome_in_ground_pad, "Landing _Outcome" from SPACEXTBL
where "Landing _Outcome" like 'Success (ground pad)';
```

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

succesful_landin_outcome_in_ground_pad	Landing _Outcome
01-05-2017	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

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



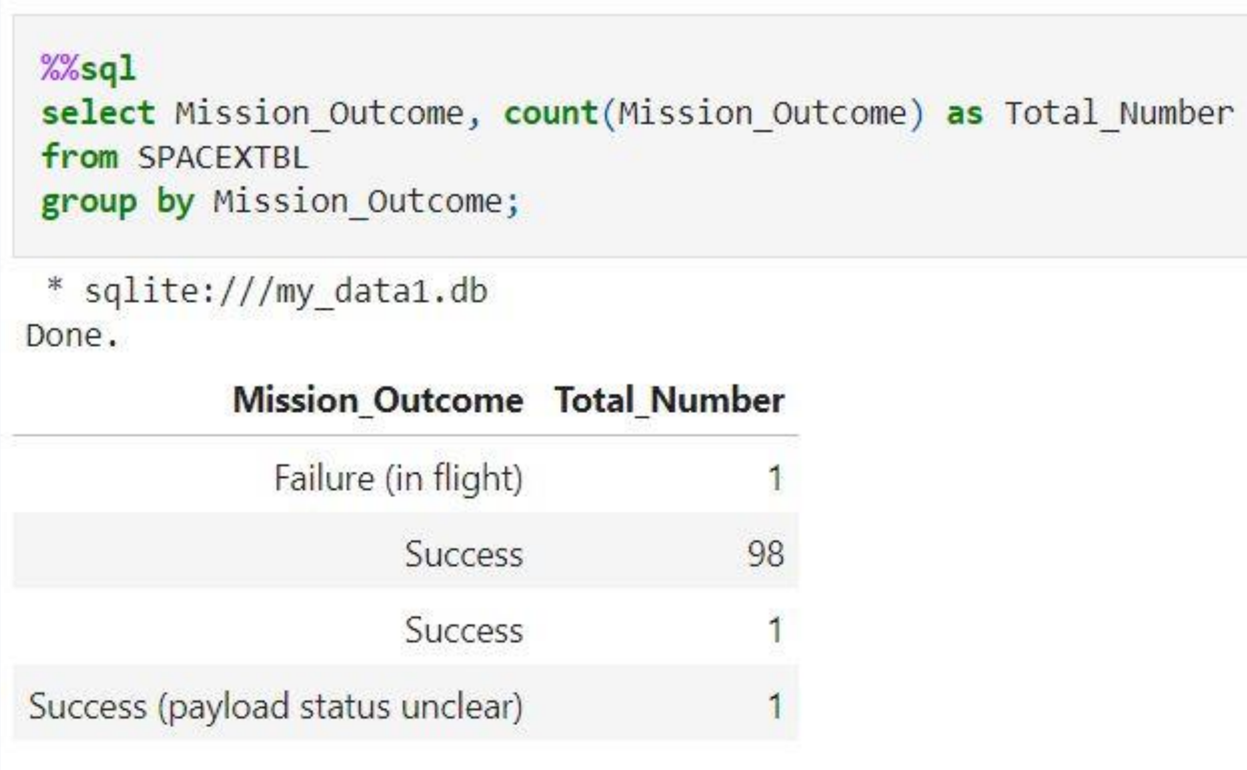
```
%%sql
select Booster_Version, "Landing _Outcome" from SPACEXTBL
where ("Landing _Outcome" like 'Success (drone ship)') and (PAYLOAD_MASS__KG_ between 4000 and 6000);
```

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

Booster_Version	Landing _Outcome
F9 FT B1022	Success (drone ship)
F9 FT B1026	Success (drone ship)
F9 FT B1021.2	Success (drone ship)
F9 FT B1031.2	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes, as we can see only one is failure:



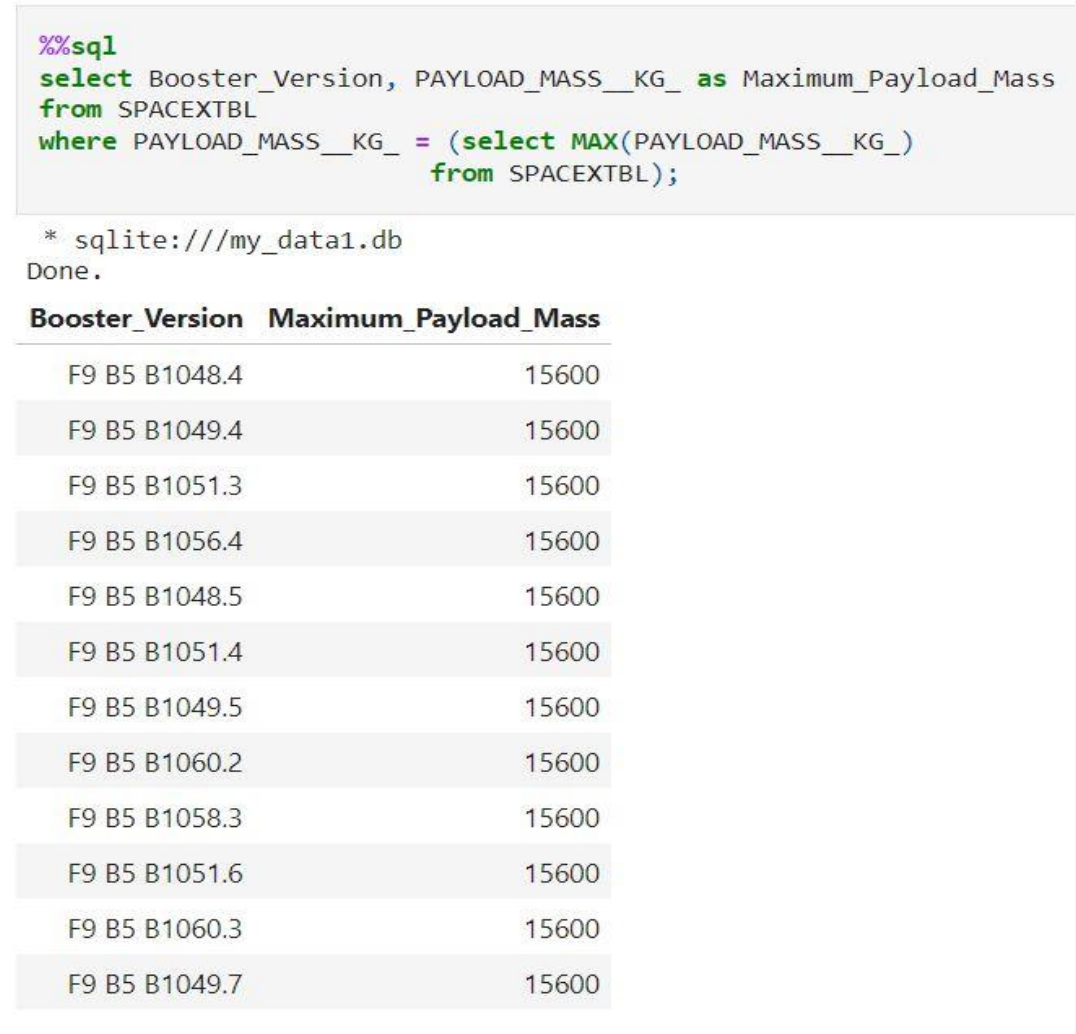
```
%%sql
select Mission_Outcome, count(Mission_Outcome) as Total_Number
from SPACEXTBL
group by Mission_Outcome;
```

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

Mission_Outcome	Total_Number
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:



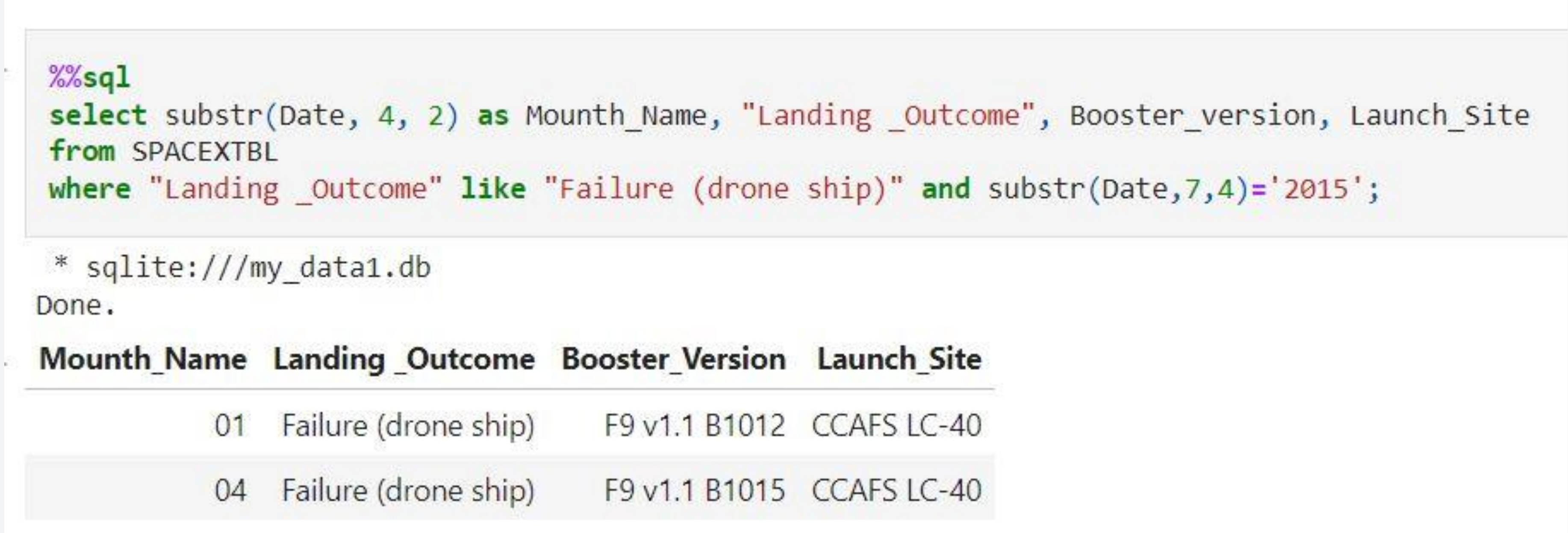
```
%%sql
select Booster_Version, PAYLOAD_MASS__KG_ as Maximum_Payload_Mass
from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_)
                          from SPACEXTBL);
```

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

Booster_Version	Maximum_Payload_Mass
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:

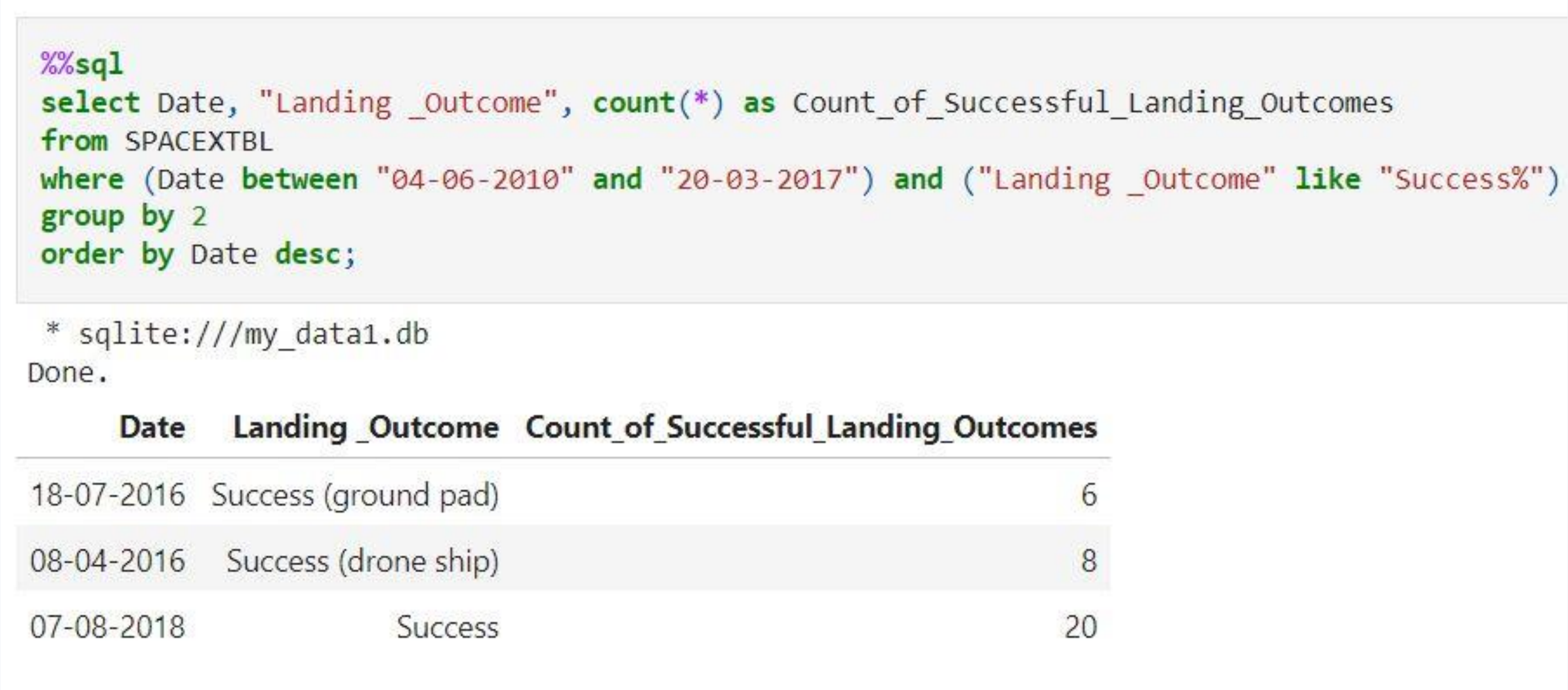


```
%%sql
select substr(Date, 4, 2) as Mounth_Name, "Landing _Outcome", Booster_version, Launch_Site
from SPACEXTBL
where "Landing _Outcome" like "Failure (drone ship)" and substr(Date,7,4)='2015';
```

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

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

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



```
%%sql
select Date, "Landing _Outcome", count(*) as Count_of_Successful_Landing_Outcomes
from SPACEXTBL
where (Date between "04-06-2010" and "20-03-2017") and ("Landing _Outcome" like "Success%")
group by 2
order by Date desc;
```

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

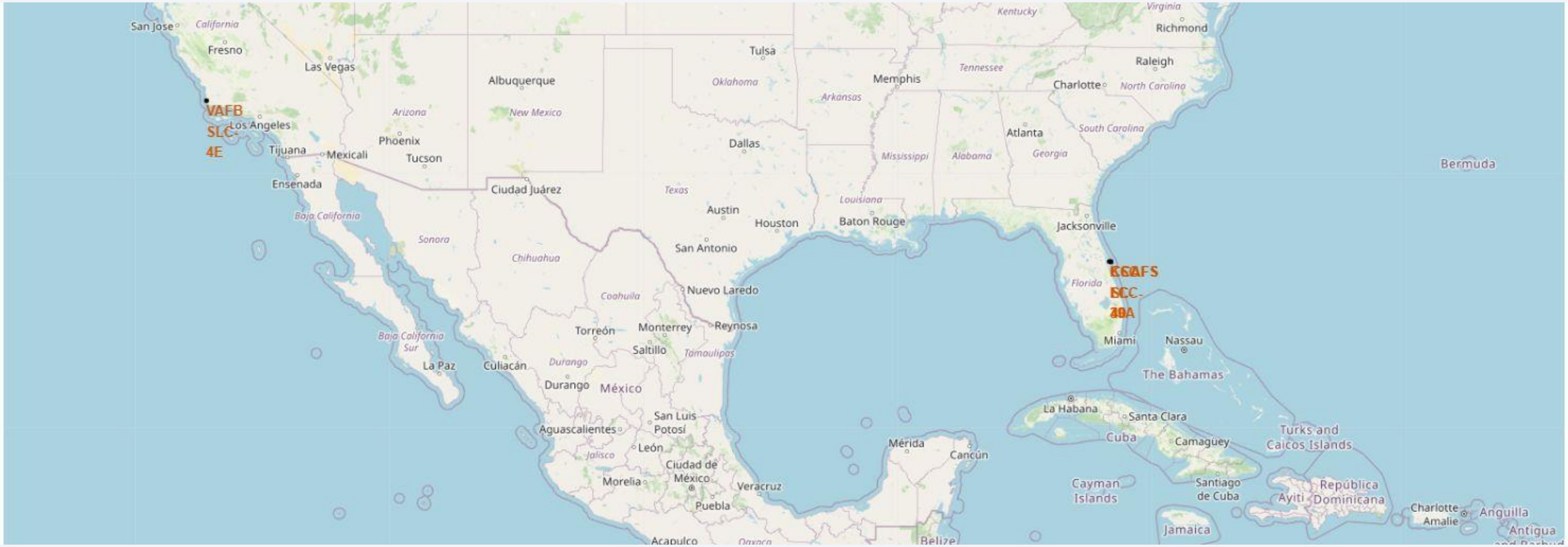
Date	Landing _Outcome	Count_of_Successful_Landing_Outcomes
18-07-2016	Success (ground pad)	6
08-04-2016	Success (drone ship)	8
07-08-2018	Success	20

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

Section 3

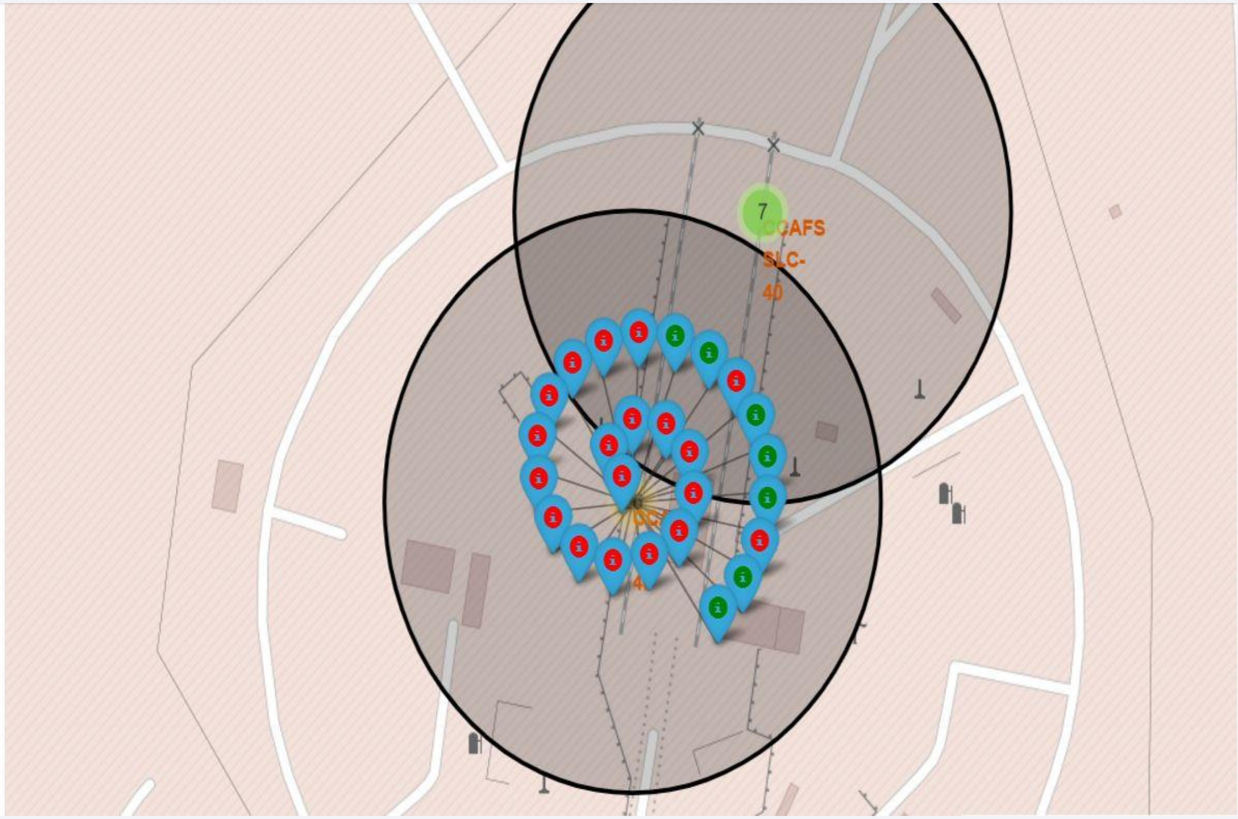
Launch Sites Proximities Analysis

All launch sites on a map



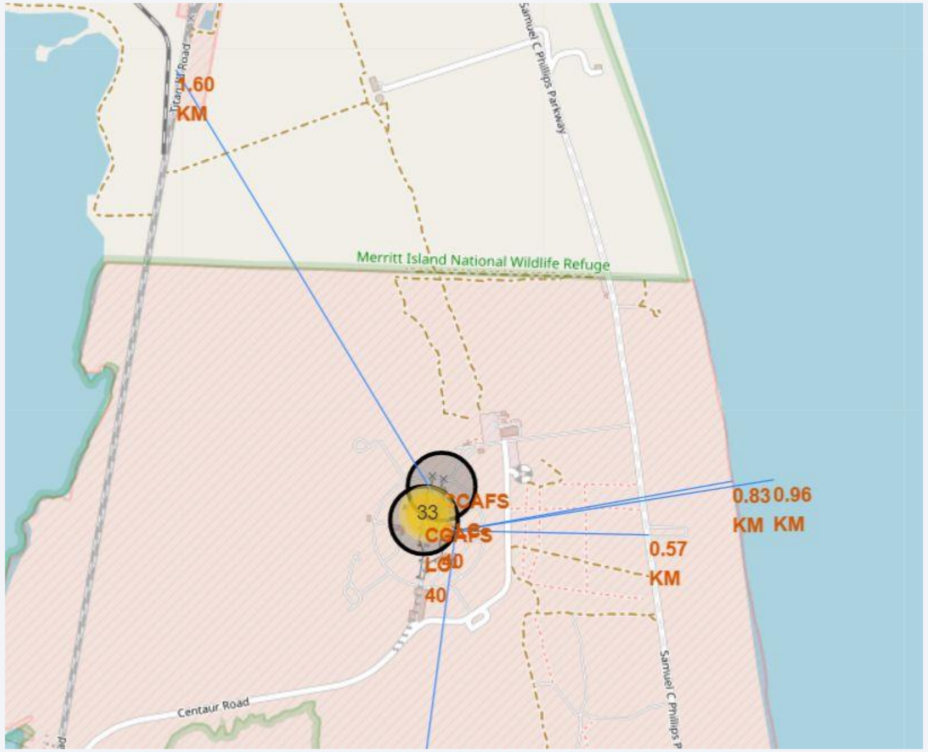
- Launch sites on the map are located near the coastline and not far from the equator line.

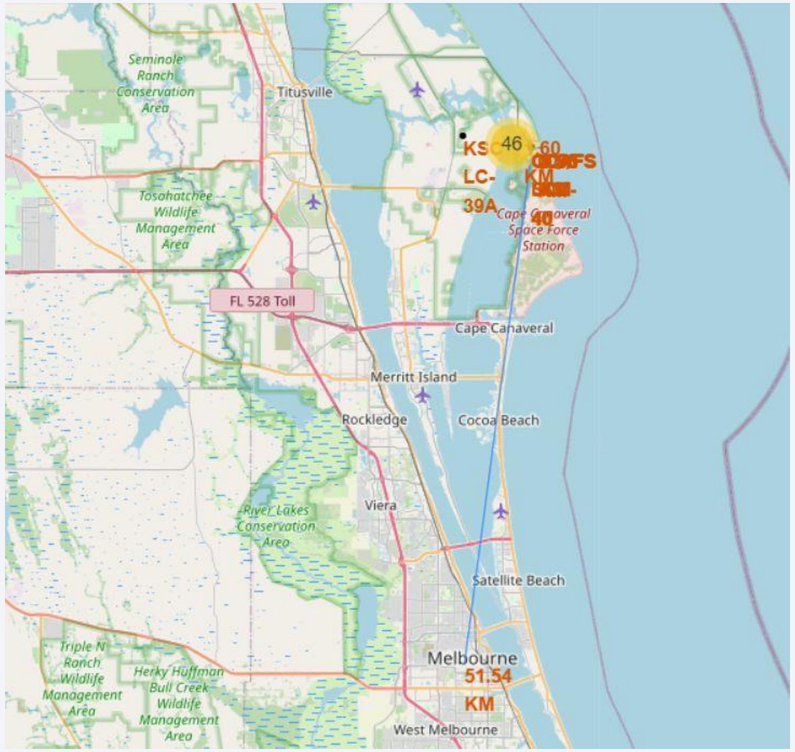
The success/failed launches for each site on the map



- In the figure on the left, you can see that on the selected launch pad, there are more unsuccessful launches than successful ones

Distances between a launch site to its proximities




Seminole Ranch Conservation Area
Titusville
Tosohatchee Wildlife Management Area
46
KSC LC-39A
CCAFS
KM
Cape Canaveral Space Force Station
FL 528 Toll
Cape Canaveral
Merritt Island
Rockledge
Cocoa Beach
Viera
River Lakes Conservation Area
Satellite Beach
Melbourne
51.54 KM
West Melbourne
Triple N Ranch Wildlife Management Area
Herky Huffman Bull Creek Wildlife Management Area

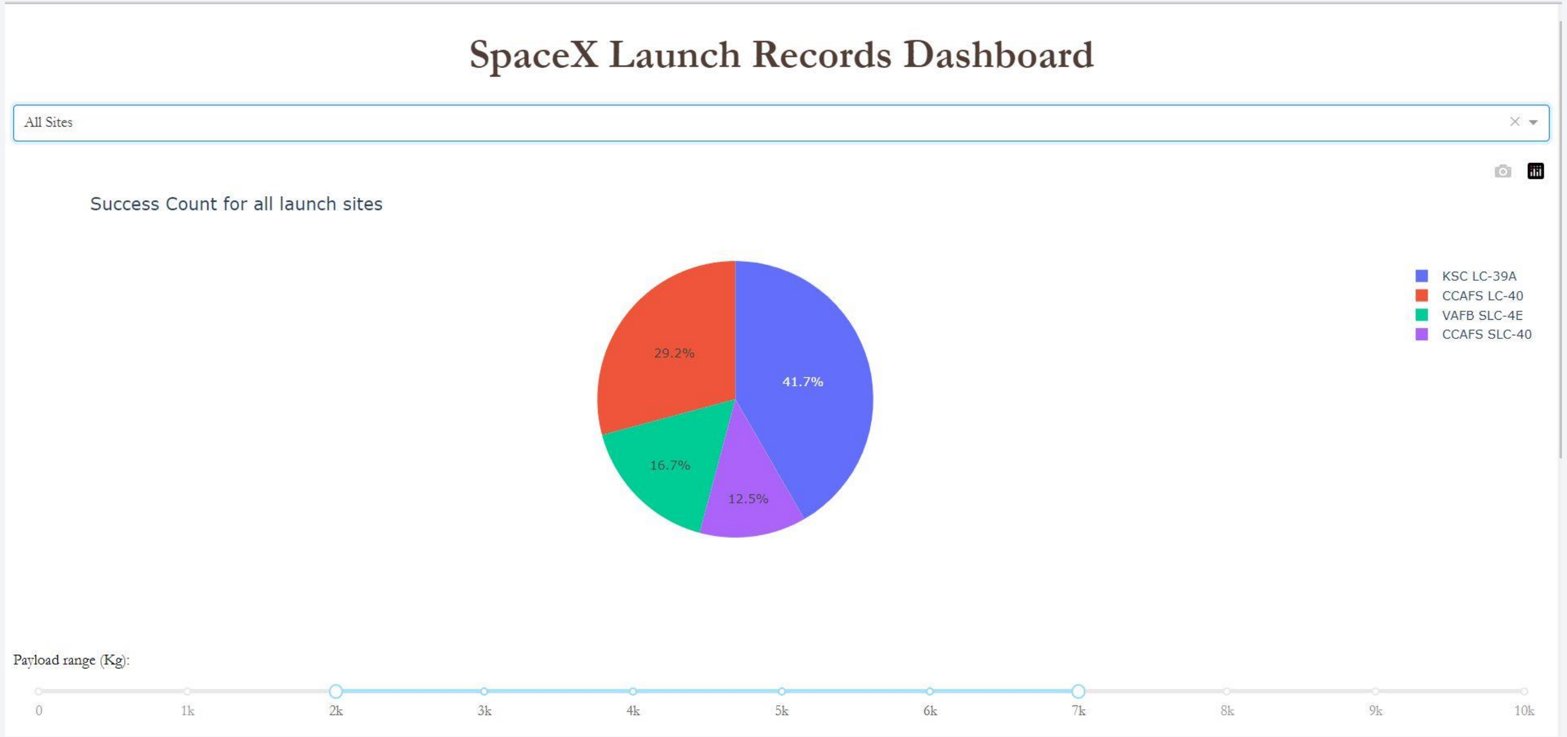
Closest railway:
1.5974725636445368 km
Closest highway:
0.5727253692913614 km
Closest city:
51.535890774647925 km

- Launch sites are located near the coastline (safety buffer zone) and the equator line (which allows the rocket to go into orbit easier and with less fuel). I also want to note that it is important to deliver some equipment located near the railway or highway.
- But for safety, it is extremely important that the launch site is located far from the city.

Section 4

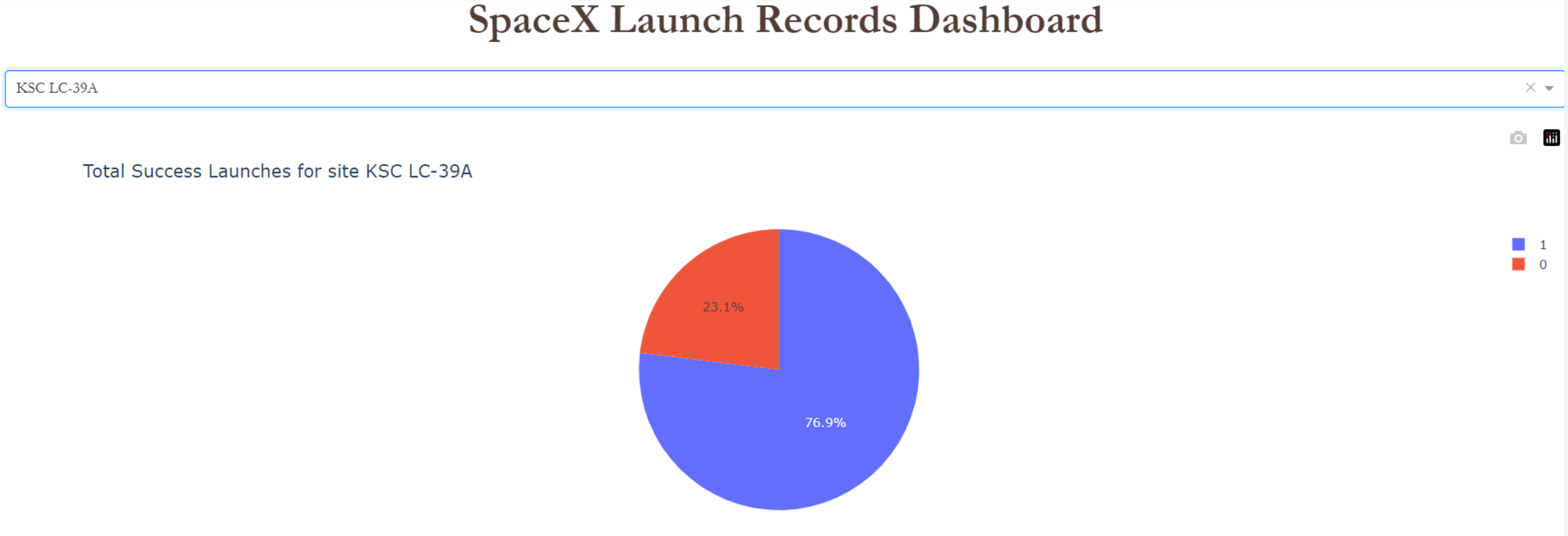
Build a Dashboard with Plotly Dash

Count of Success for all launch sites



- The graph shows that the largest number of successful landings at the launch pads: KSC LC-39A and CCAFS LC-40.

The launch site with highest launch success ratio



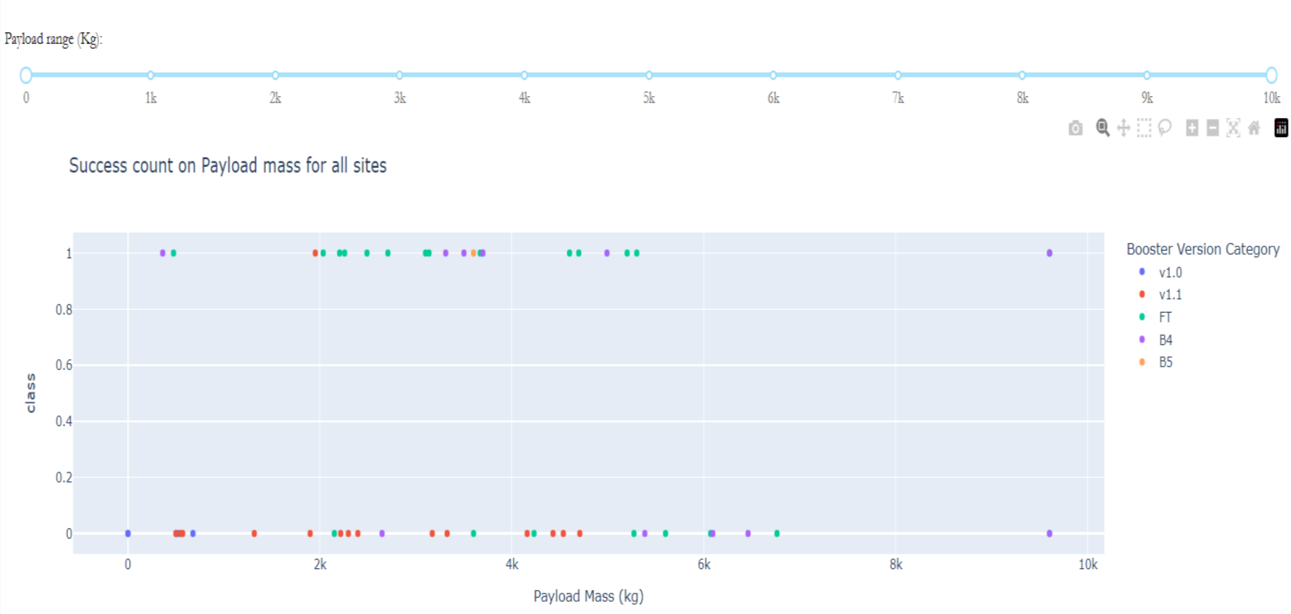
SpaceX Launch Records Dashboard

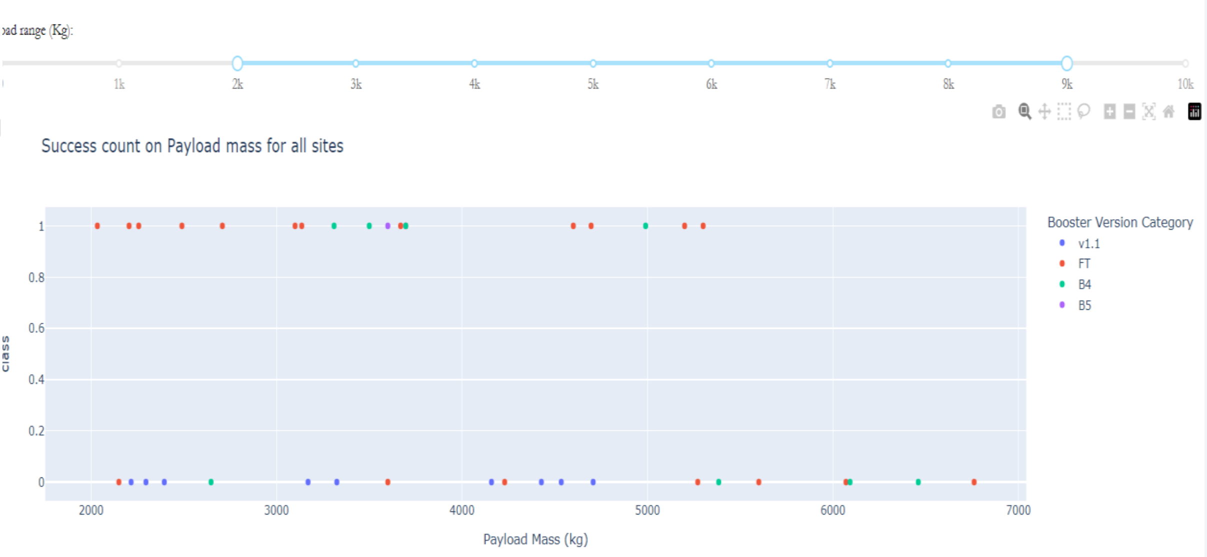
KSC LC-39A

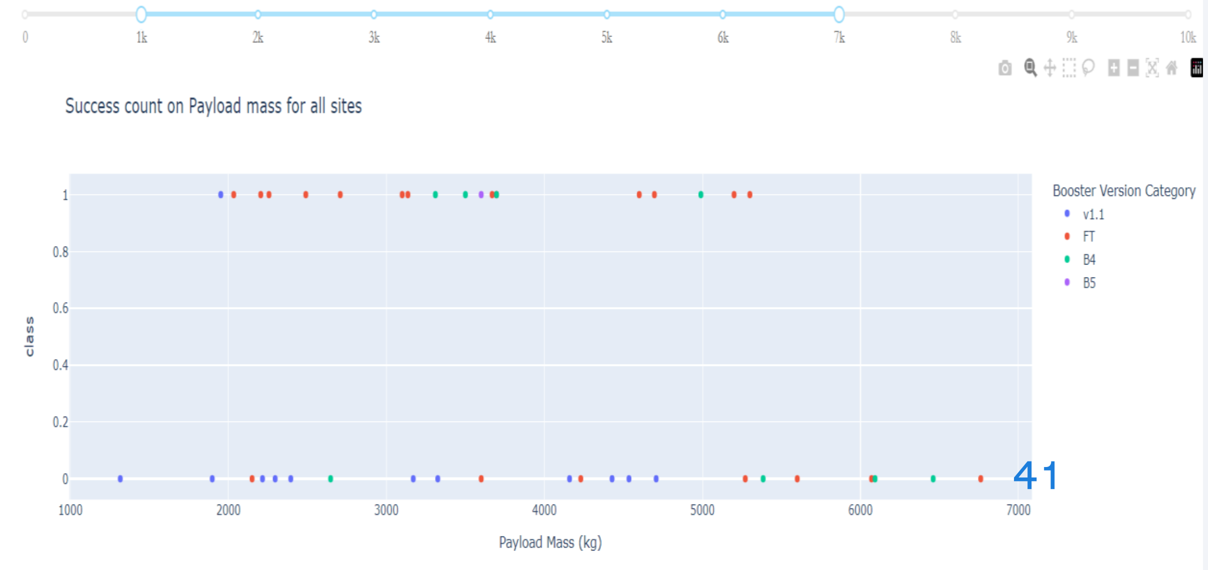
Total Success Launches for site KSC LC-39A

- the pie chart for the launch site with highest launch success ratio is KSC LC-39A

Payload vs. Launch Outcome scatter plot for all sites





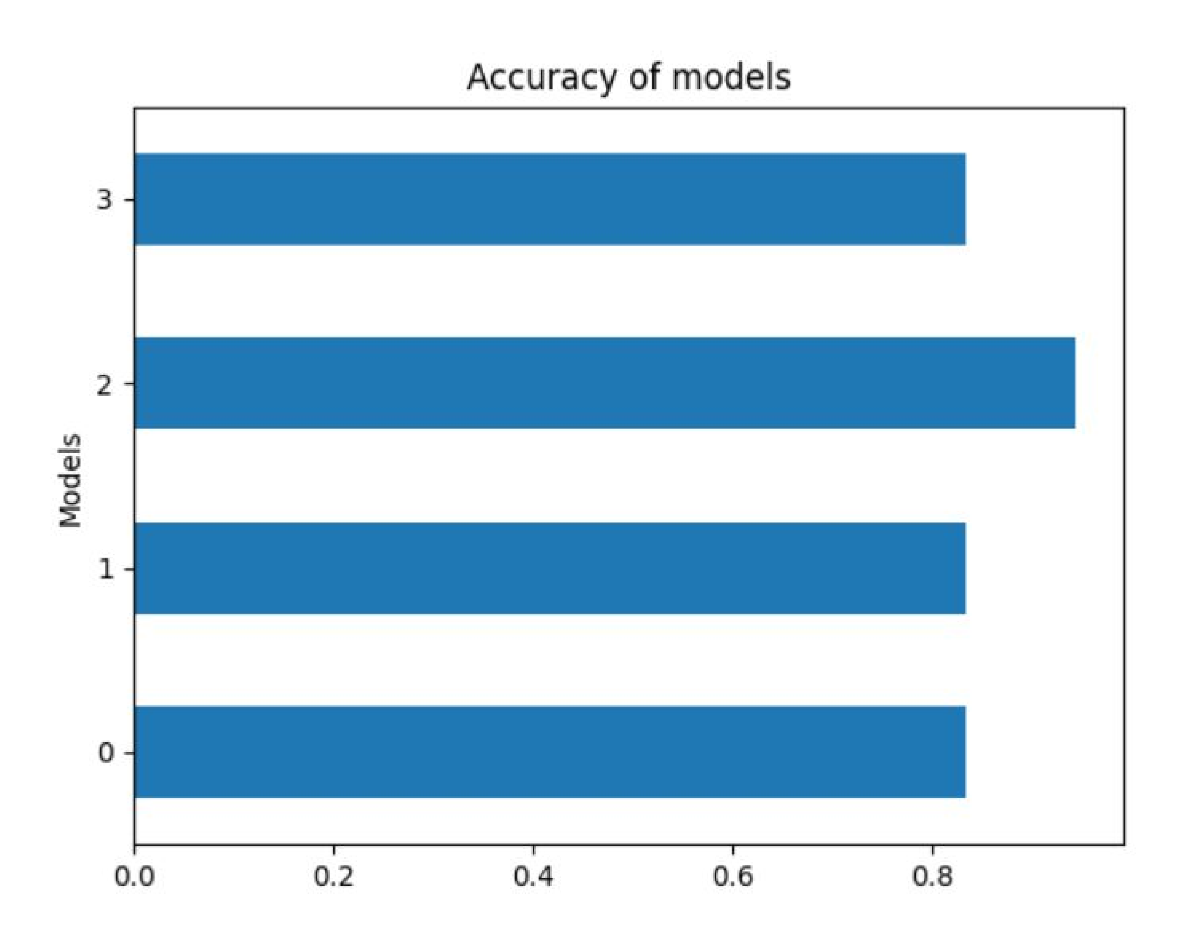


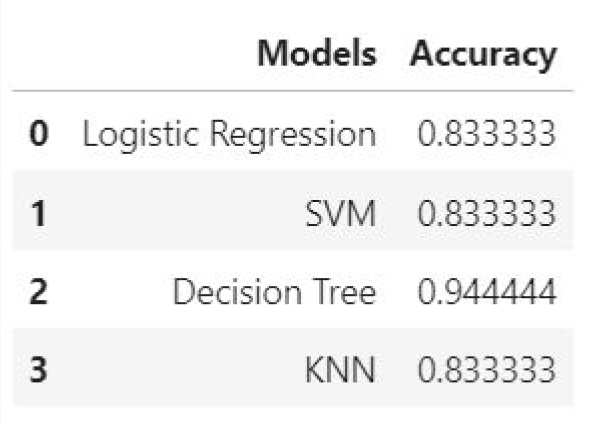
- The success depends on payload mass and F9 Booster version with the highest launch success rate is FT.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

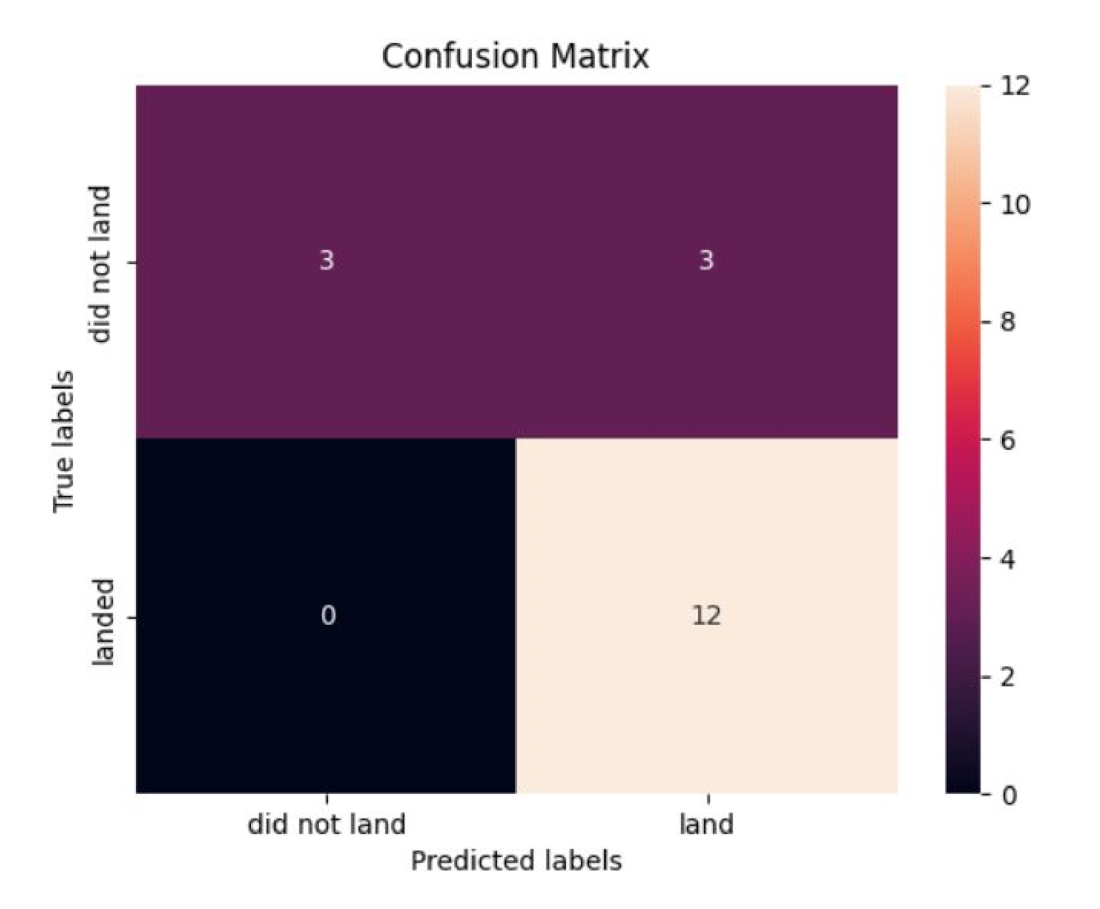




	Models	Accuracy
0	Logistic Regression	0.833333
1	SVM	0.833333
2	Decision Tree	0.944444
3	KNN	0.833333

As can be seen from the table and bar **plot**, the model based on decision trees shows the highest accuracy.

Confusion Matrix



The confusion matrix of the best performing model with an explanation:

- it can be seen that the model accurately determines the landing of the first stage and erroneously defines three as landing (false positive)

Conclusions

The most favorable location for launch sites is:

- near the railway, highway and coastline
- far from the city;
- near the equator line, which helps save fuel and reach orbit faster;

- site with the largest successful launches is KSC LC-39A;
- site has the highest launch success rate is KSC LC-39A;
- payload ranges with the highest launch success rate are KSC LC-39A and CCAFS LC-40;
- payload ranges with the lowest launch success rate are CCAFS SLC-40 and VAFB SLC-4E;
- F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) with the highest launch success rate is FT.

Thank you!