IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ming-Chung Chan
2022-11-30

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- In this study, dataset of rocket landing was retrieved from SpaceX API and wiki web page.
- Some statistics are calculated by Pandas and SQL queries.
- By visualization, landing success rate is analyzed with different features like launch site, orbit type, payload mass, etc.
- Different machine learning models were trained. Logistic regression model was selected as the best performing model with accuracy 83.3%.

Introduction

Cost of rocket launching can be reduced by their reuse.

In this study, our aims are to:

1. Investigate the factors that lead to leading success or failure by analyzing the historical data from SpaceX
2. Train and select the best machine learning model for predicting landing success rate

Section 1

Methodology

Methodology

Executive Summary

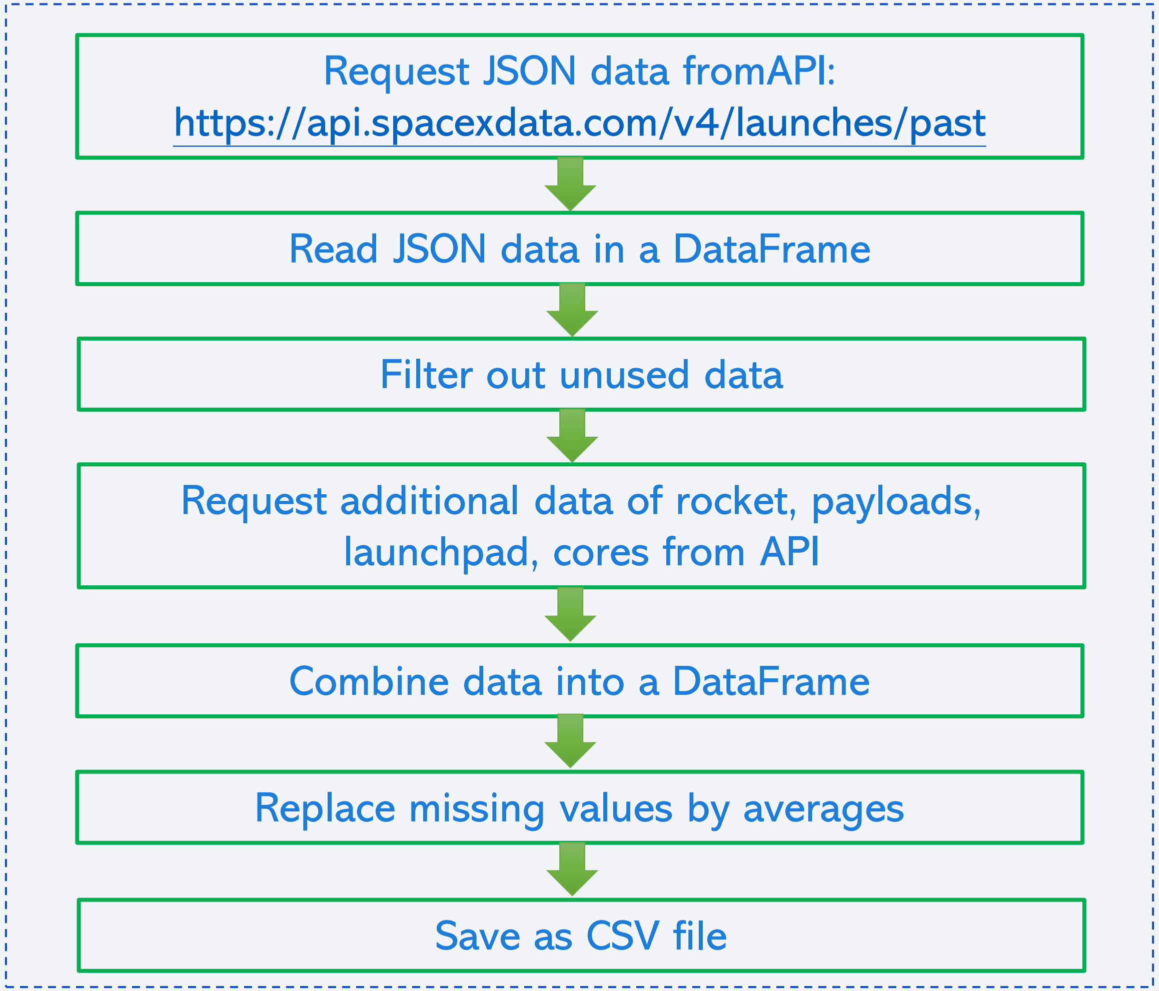
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection – SpaceX API

- What we did:

1. Request to the SpaceX REST API
2. Clean the requested data

- SpaceX API calls notebook: https://github.com/stoneagemcc/Data_Science_MC/blob/main/1_Data_Collection_API.ipynb



Request JSON data fromAPI: https://api.spacexdata.com/v4/launches/past

Read JSON data in a DataFrame

Filter out unused data

Request additional data of rocket, payloads, launchpad, cores from API

Combine data into a DataFrame

Replace missing values by averages

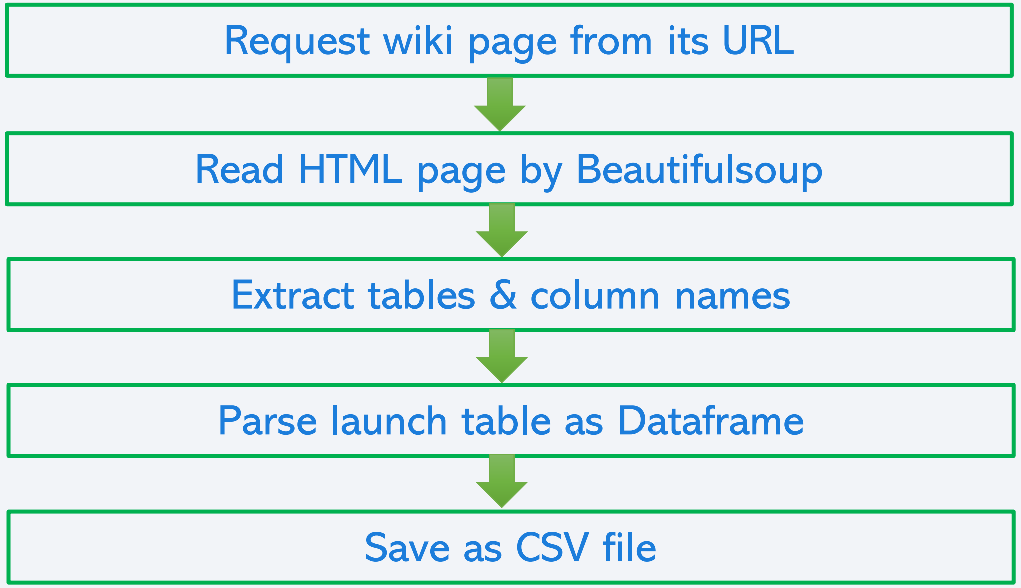
Save as CSV file

Data Collection - Scraping

- What we did:

1. Extract a Falcon 9 launch records HTML table from Wikipedia
2. Parse the table and convert it into a Pandas data frame

- Web scraping notebook: https://github.com/stoneagemcc/Data_Science_MC/blob/main/2_Data_Collection_Web_Scraping.ipynb



Request wiki page from its URL

Read HTML page by Beautifulsoup

Extract tables & column names

Parse launch table as Dataframe

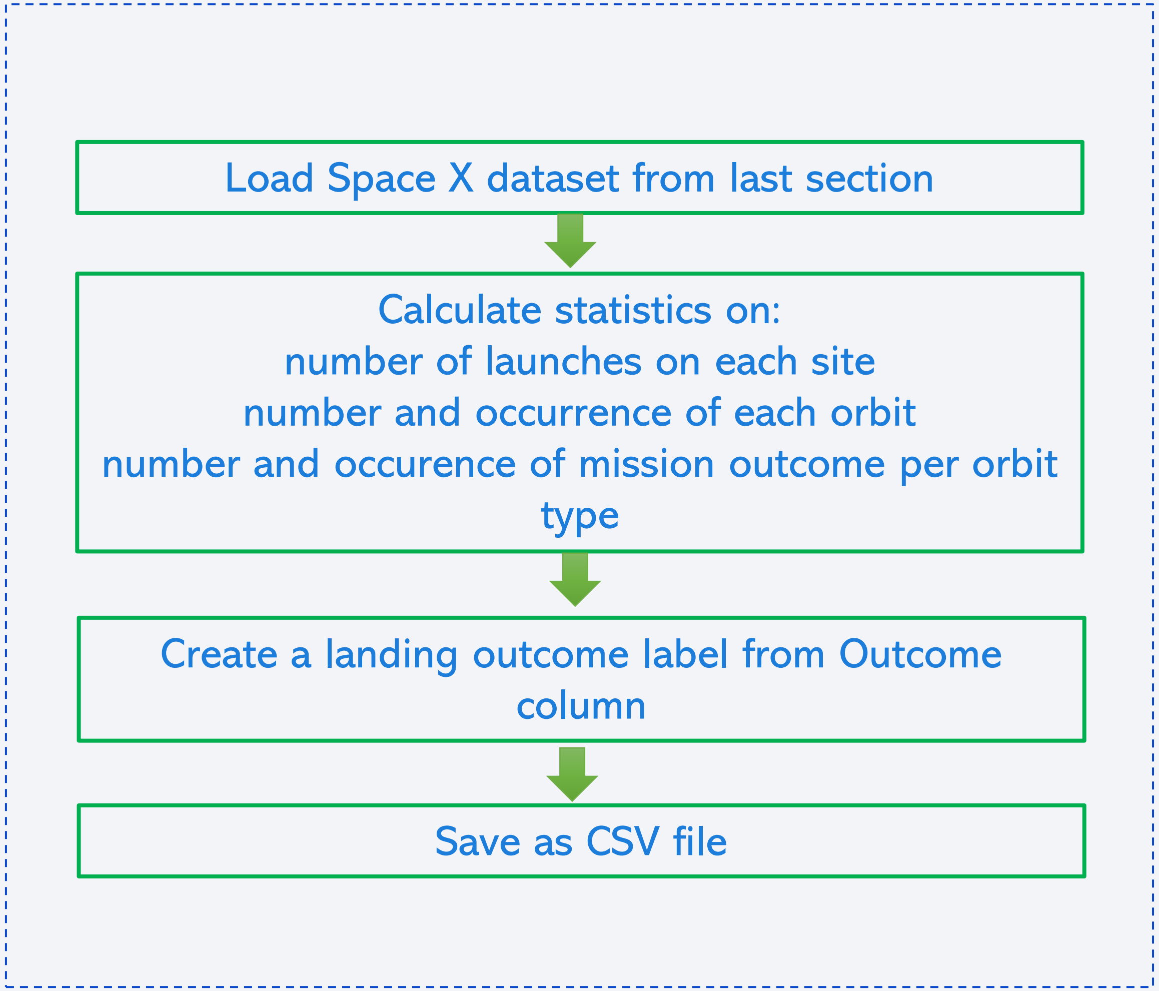
Save as CSV file

Data Wrangling

- What we did:

1. Understand some patterns by EDA
2. Determine Training Labels

- Data wrangling notebooks: https://github.com/stoneagemcc/Data_Science_MC/blob/main/3_Data_Wrangling.ipynb



Load Space X dataset from last section

Calculate statistics on:
number of launches on each site
number and occurrence of each orbit
number and occurence of mission outcome per orbit type

Create a landing outcome label from Outcome column

Save as CSV file

EDA with Data Visualization

- Objectives:

1. Exploratory Data Analysis
2. Preparing Data Feature Engineering

- Data are visualized in:

1. scatter plots to show the relationships on different combination of variables
2. bar charts to show the landing success rate on different categories
3. line charts to show the yearly trend of landing success rate

- EDA with data visualization notebook:
  https://github.com/stoneagemcc/Data_Science_MC/blob/main/5_EDA_with_Visualization.ipynb

EDA with SQL

- SQL queries performed to answer the following questions:

1. names of the unique launch sites in the space mission
2. 5 records where launch sites begin with the string 'CCA’
3. total payload mass carried by boosters launched by NASA (CRS)
4. average payload mass carried by booster version F9 v1.1
5. the date when the first succesful landing outcome in ground pad was achieved
6. boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. total number of successful and failure mission outcomes
8. booster_versions which have carried the maximum payload mass
9. records with month names, failure landing_outcomes in drone ship ,booster versions, launch site in year 2015
10. Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017

- EDA with SQL notebook:
  https://github.com/stoneagemcc/Data_Science_MC/blob/main/4_EDA_with_SQL.ipynb

Build an Interactive Map with Folium

- Folium maps were created to show information related to locations.

- The following map objects were added to the map:

1. Circles: to encircle the highlighted regions
2. Icons: to show the site names
3. Makers with popups: to show the information at specific locations
4. Maker cluster: to simplify a map such that confusion of too many markers is avoided
5. Line: to show the closest distance between two locations

- Interactive map with Folium map notebook:
  https://github.com/stoneagemcc/Data_Science_MC/blob/main/6_Interactive_Visual_Analytics_with_Folium.ipynb

Build a Dashboard with Plotly Dash

- Interactive dashboard was built to perform interactive visual analytics

- The following interactive plots or components were added:

1. Launch Site Drop-down Input: to allow user to choose which launch site to analyze
2. Bar chart: to show the landing success statistics for selected site
3. Range Slider: to allow user to choose the range of payload mass to analyze
4. Scatter Plot: to visualize the correlation between payload and landing success for selected site and payload mass range

- Plotly Dash notebook: https://github.com/stoneagemcc/Data_Science_MC/blob/main/7_Interactive_Dashboard_with_Ploty_Dash.py

Predictive Analysis (Classification)

- Each classification model for landing success/failure was built, evaluated by:

1. Split the dataset into training and testing sets.
2. Select range of hyperparameters for grid search.
3. Train the model on training dataset with cross-validation to choose the best hyperparameter set
4. Evaluate the model accuracy on testing dataset
5. Evaluate the model confusion matric on testing dataset

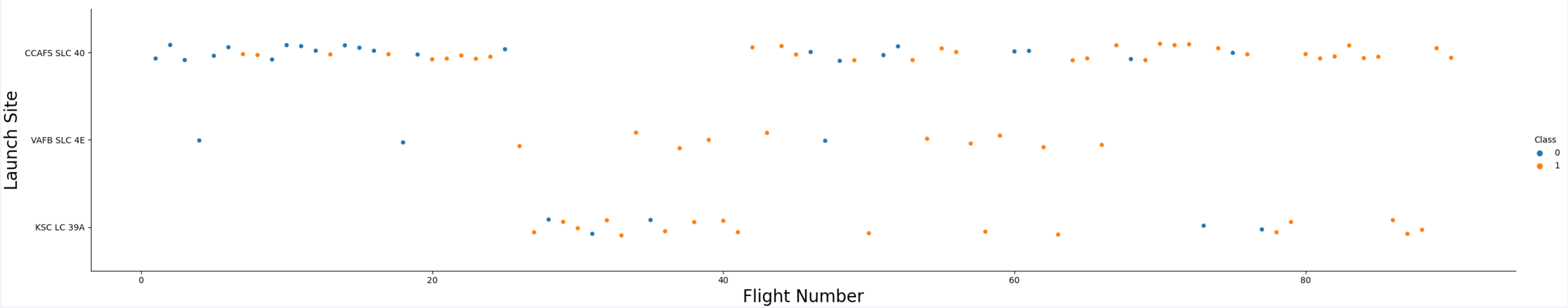
- Logistic regression, SVM, KNN and decision tree models were trained and the best one is selected based on accuracy on testing dataset

- Predictive analysis by machine learning notebook: https://github.com/stoneagemcc/Data_Science_MC/blob/main/8_Machine_Learning_Prediction.ipynb

Section 2

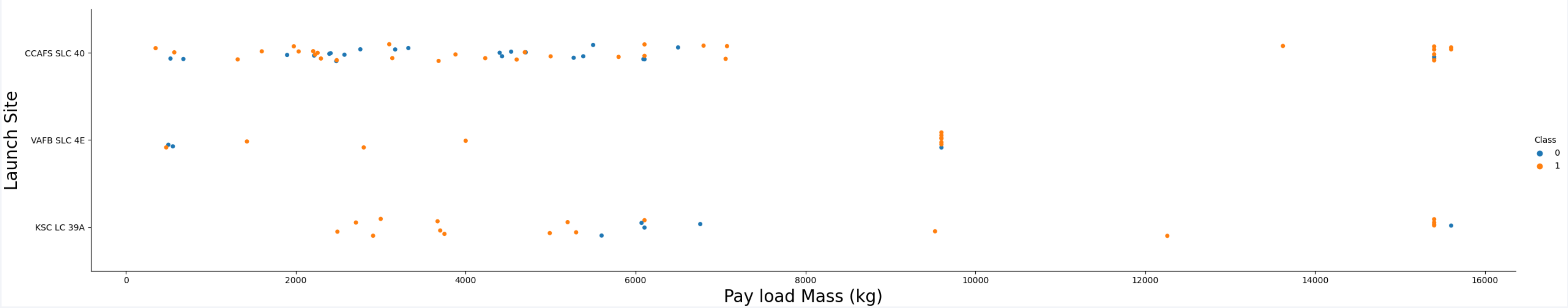
Insights drawn from EDA

Flight Number vs. Launch Site



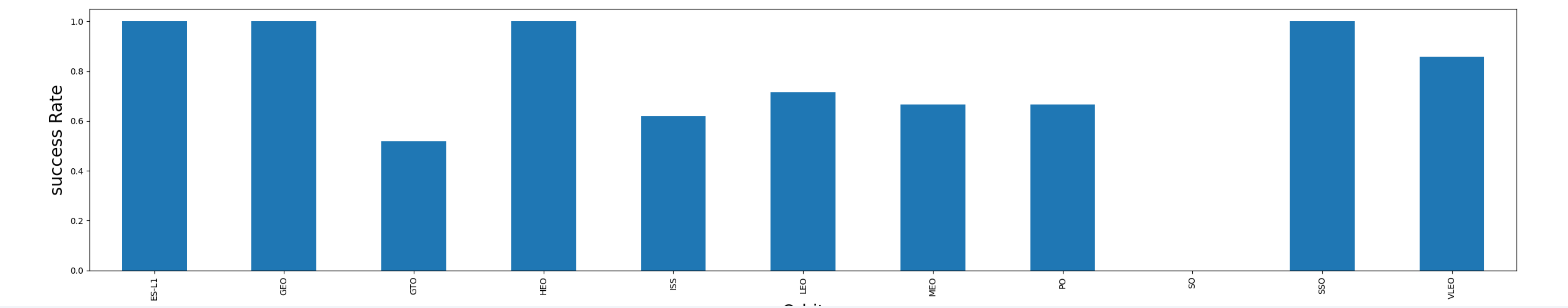
Success rates of sites VAFB SLC and KSC LC 39A are higher than success rate of CCAFS SLC 40.

Payload vs. Launch Site



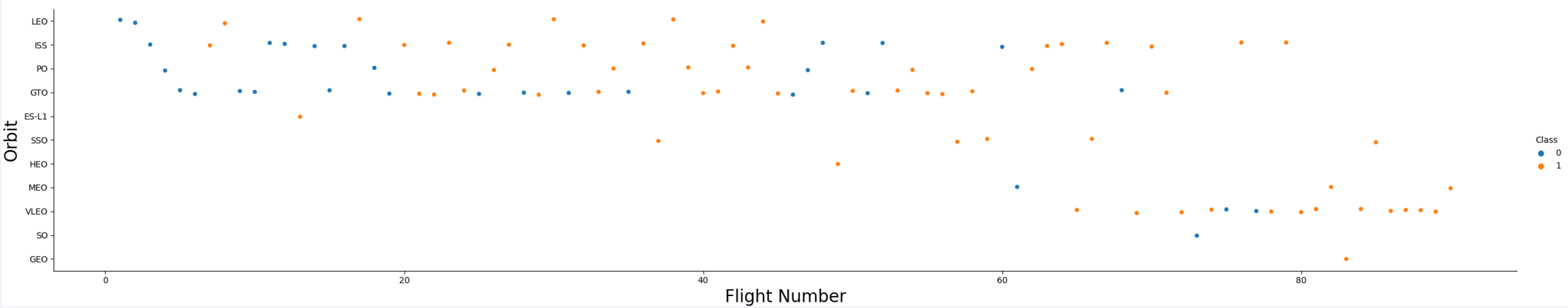
There are no rockets launched for heavy payload mass (greater than 10000) for the launch site VAFB-SLC.

Success Rate vs. Orbit Type



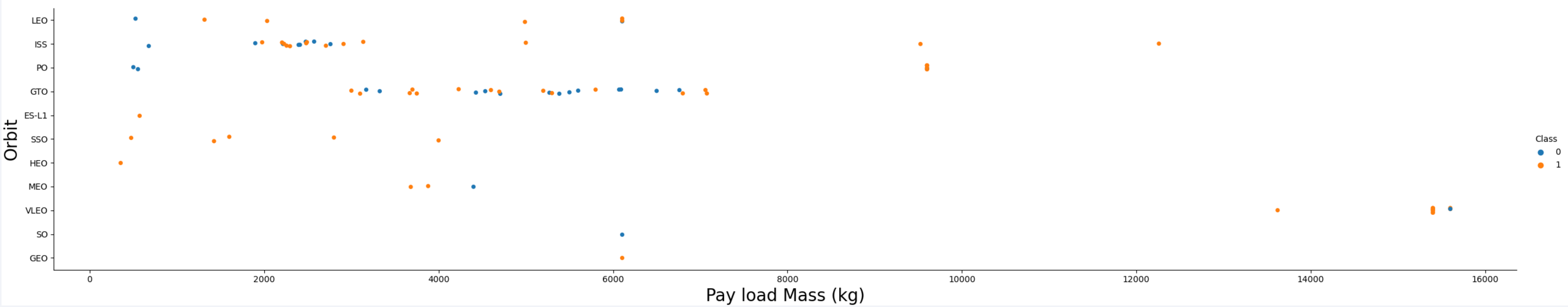
ES-L1, GEO, HEO and SSO are the orbit types with 100% success rate while So is the orbit type with 0% success rate.

Flight Number vs. Orbit Type



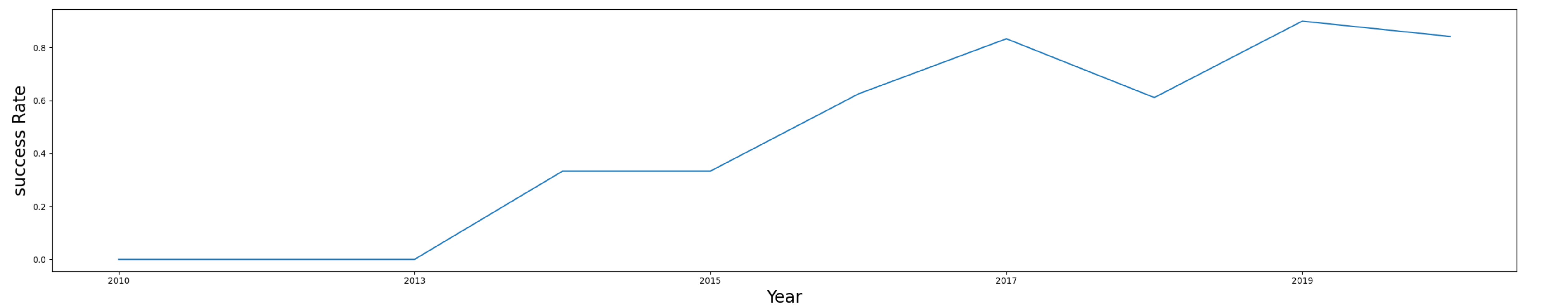
LEO orbit the Success appears related to the number of flights.
There seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



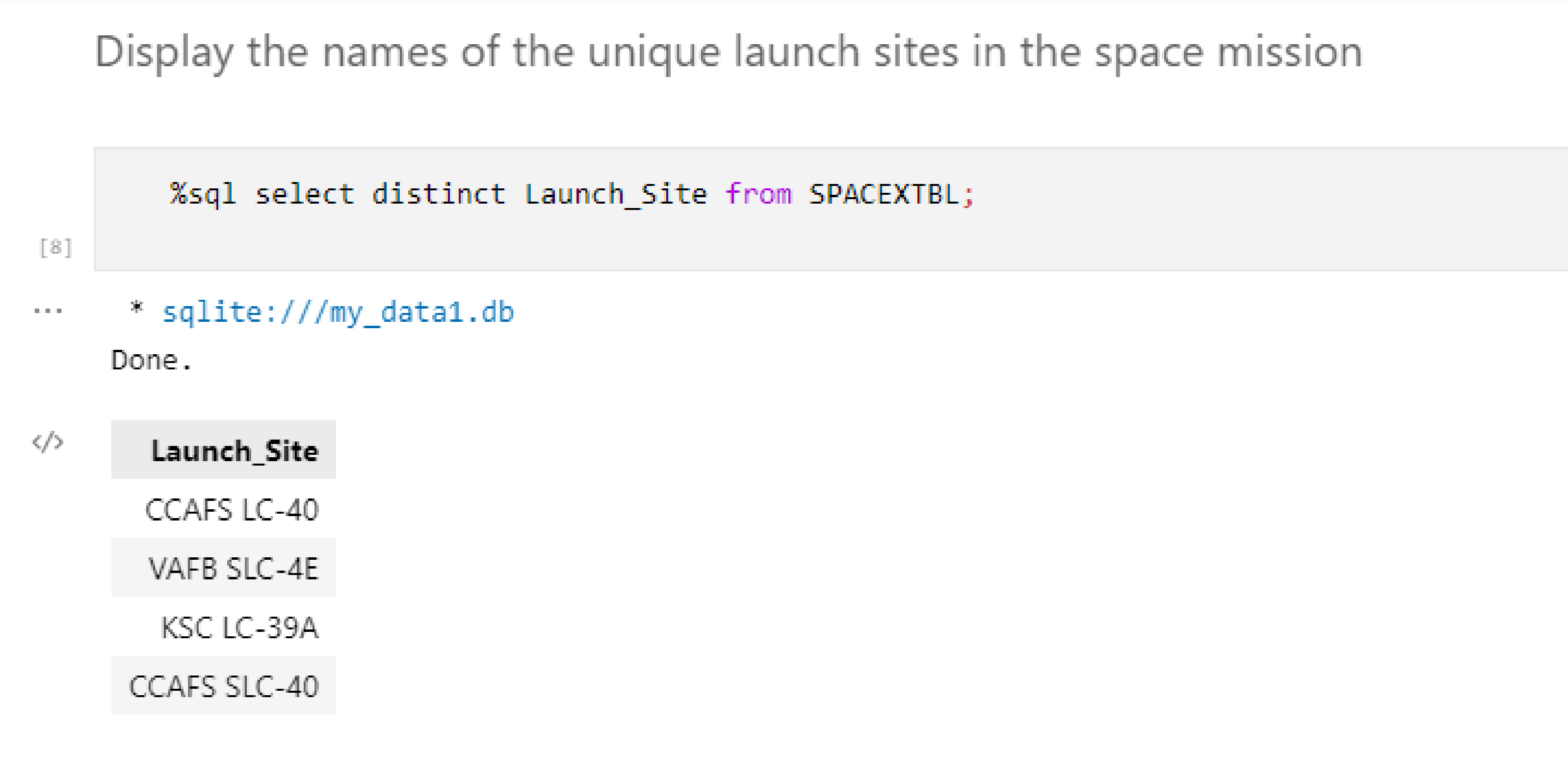
With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



The success rate since 2013 kept increasing till 2020.

All Launch Site Names



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

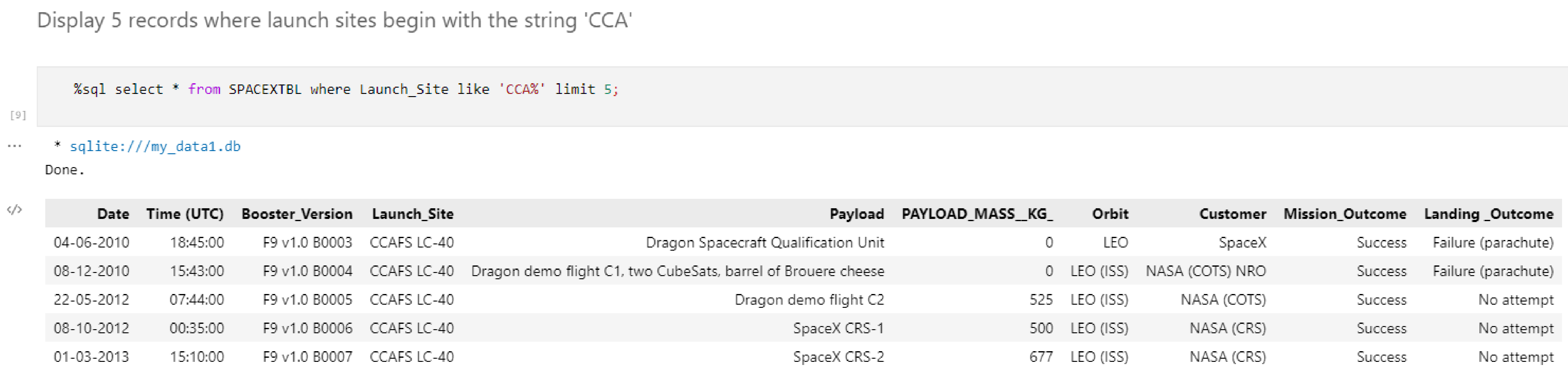
```
[8]    %sql select distinct Launch_Site from SPACEXTBL;
```

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

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

Launch site includes 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 the string 'CCA'

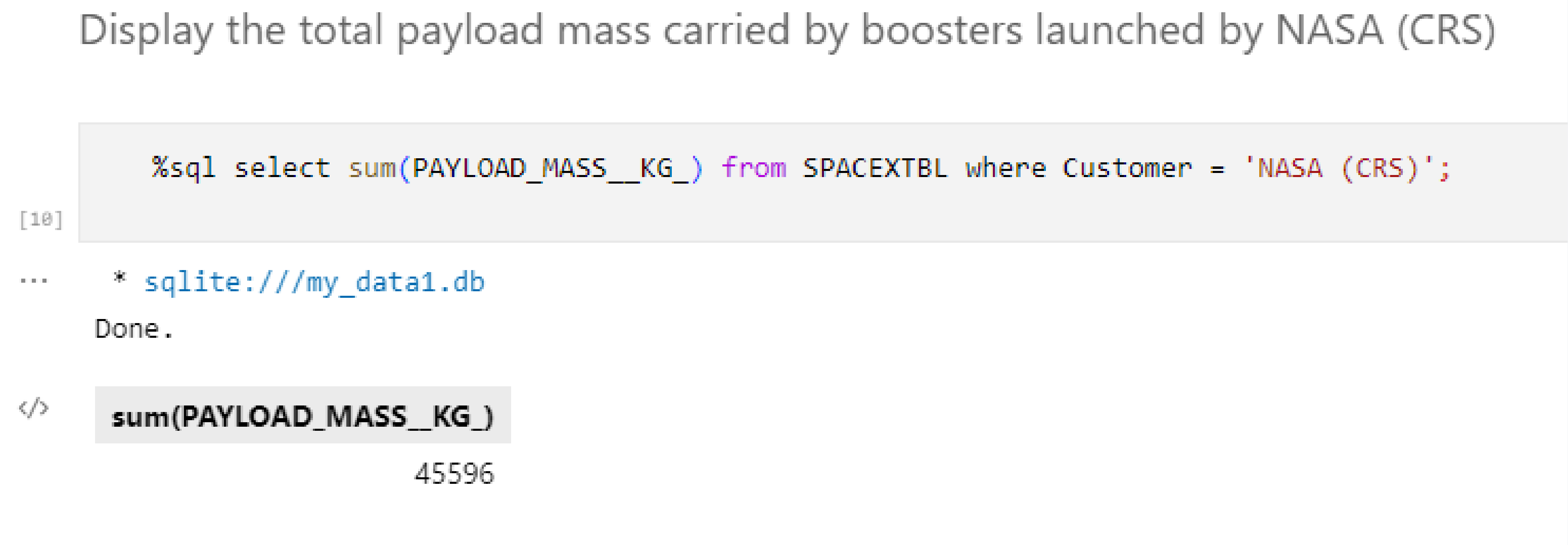
```
%sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5;
```

```
 * 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 table shows the 5 records where names of launch sites begin with ‘CCA’.

Total Payload Mass



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

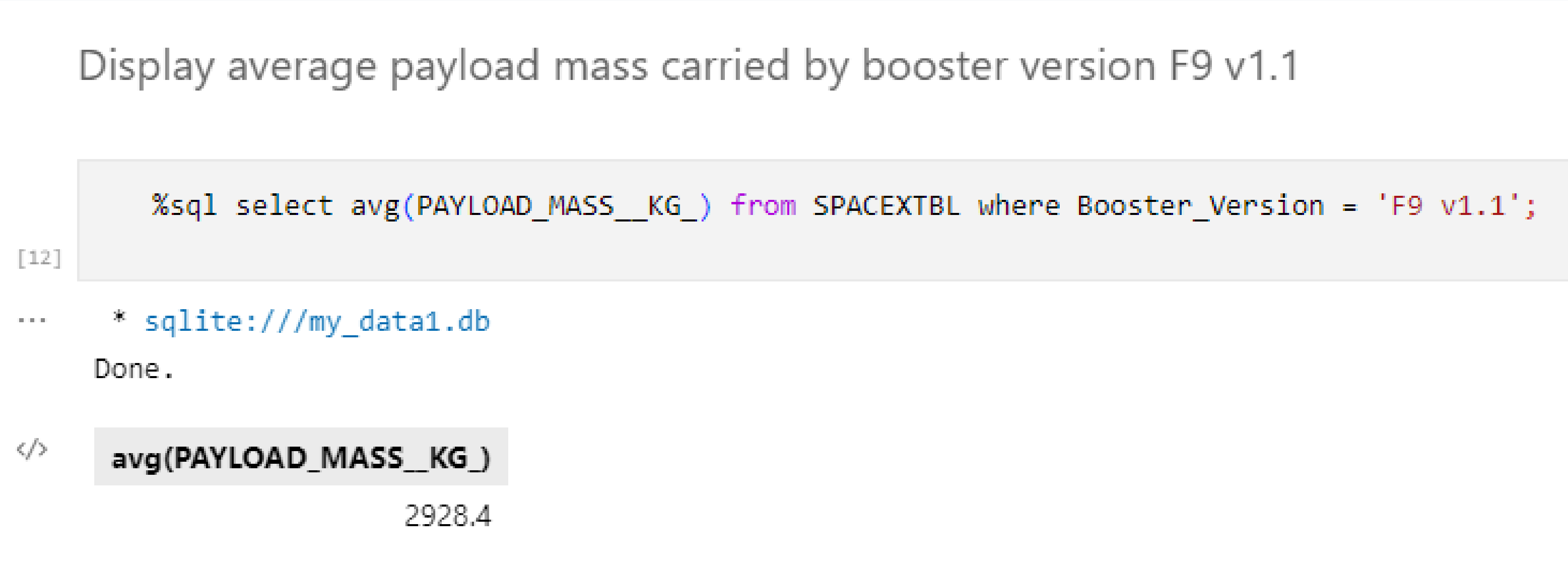
```
[10]
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)';
```

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

sum(PAYLOAD_MASS__KG_)
                 45596
```

Total payload mass launched by NASA (CRS) is 45596 kg.

Average Payload Mass by F9 v1.1



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

[12]

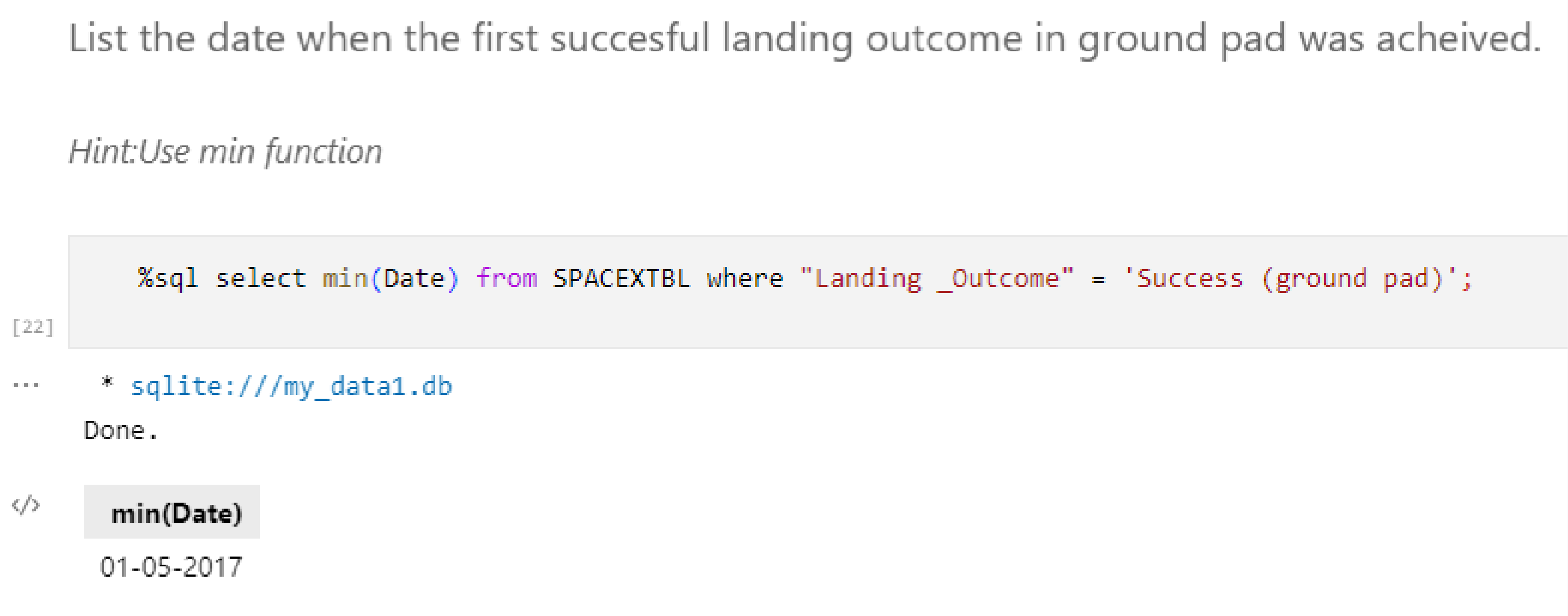
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1';
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

Average payload mass carried by F9 v1.1 is 2928.4 kg.

First Successful Ground Landing Date



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

Hint:Use min function

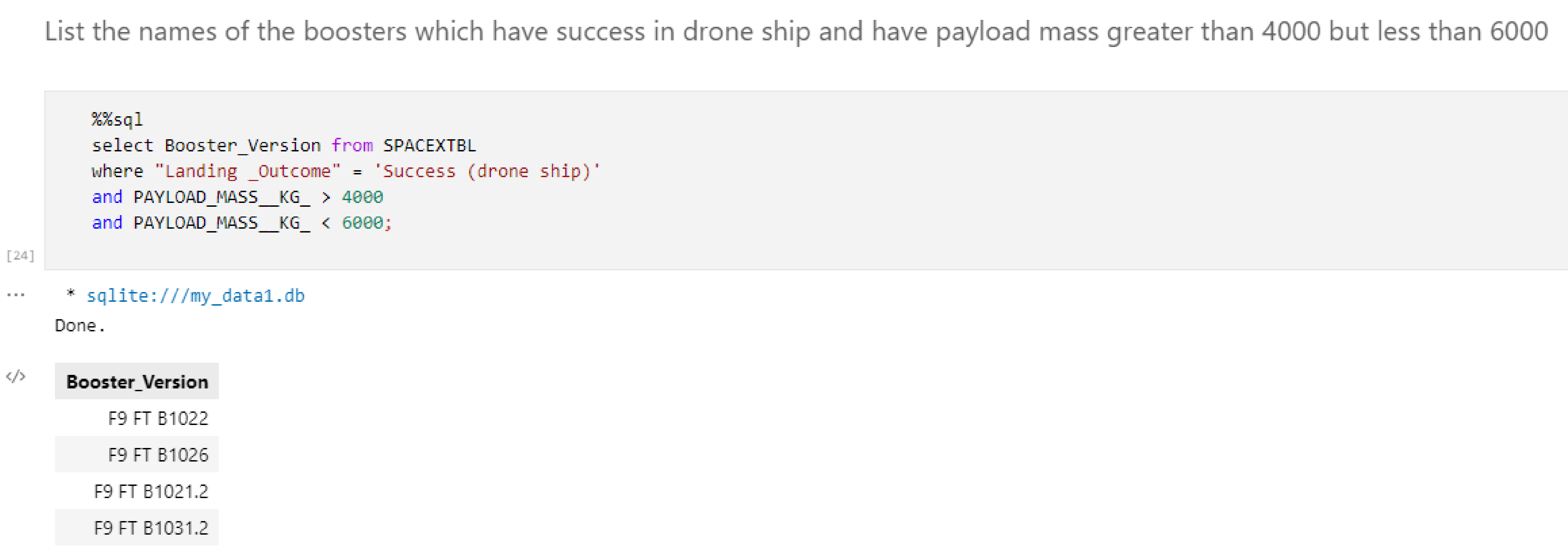
```
%sql select min(Date) from SPACEXTBL where "Landing _Outcome" = 'Success (ground pad)';
```

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

min(Date)
01-05-2017

The first successful landing in ground pad is on 01-05-2017.

Successful Drone Ship Landing with Payload between 4000 and 6000



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

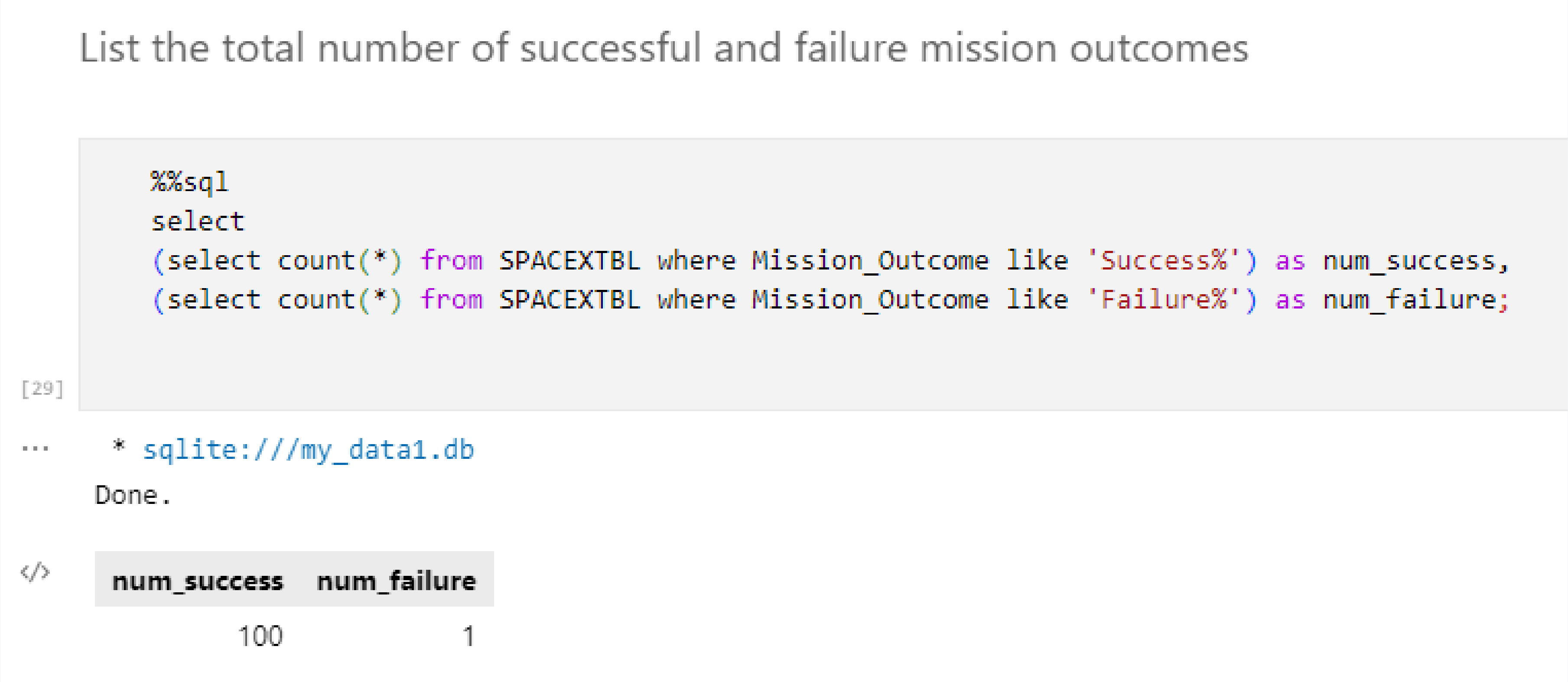
```
%%sql
select Booster_Version from SPACEXTBL
where "Landing _Outcome" = 'Success (drone ship)'
and PAYLOAD_MASS__KG_ > 4000
and PAYLOAD_MASS__KG_ < 6000;
```

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

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

Boosters with successful landing in drop ship having payload between 4000 kg & 6000 kg includes B1002, B1026, B1021.2 & B2031.2.

Total Number of Successful and Failure Mission Outcomes



List the total number of successful and failure mission outcomes

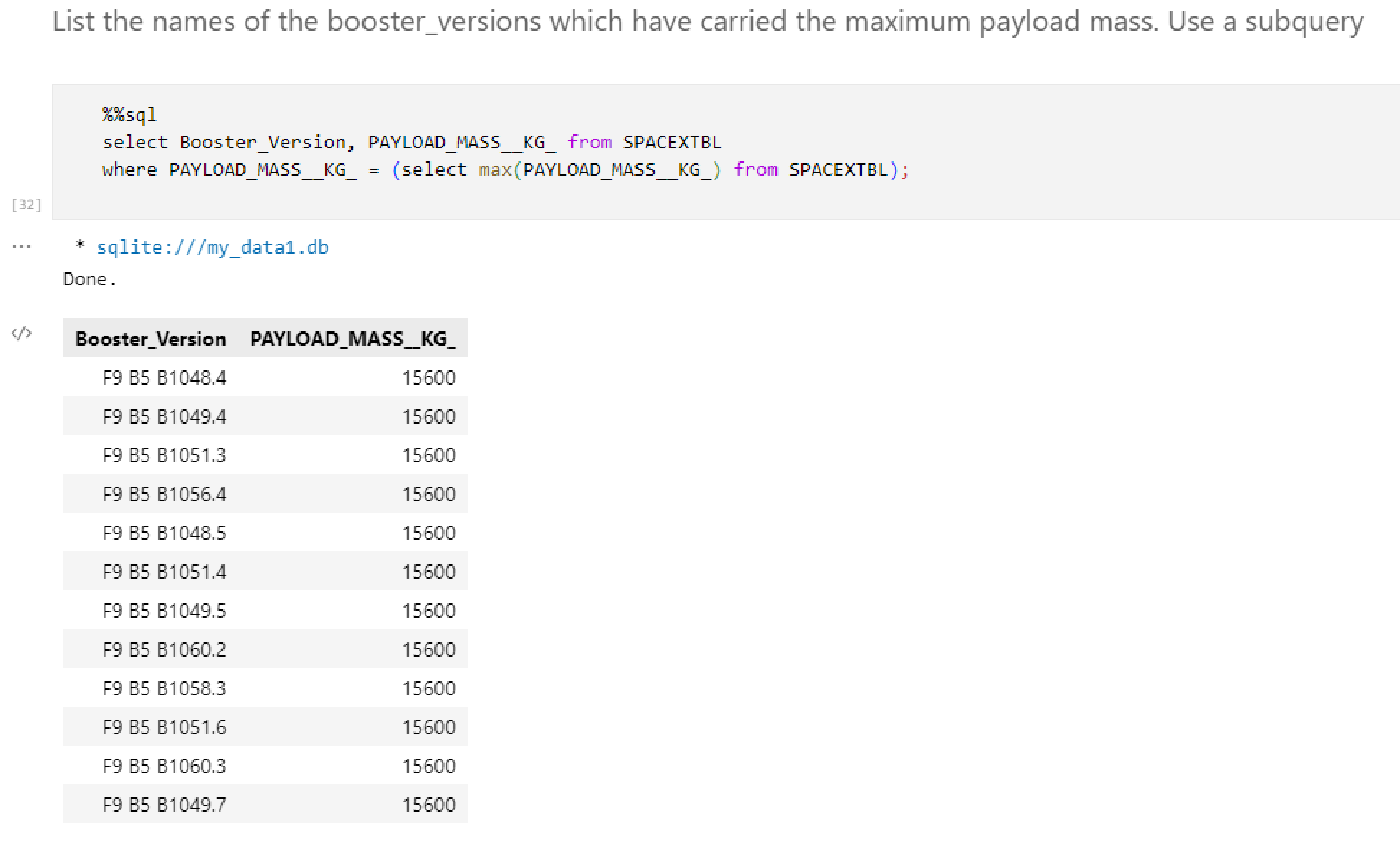
```
%%sql
select
(select count(*) from SPACEXTBL where Mission_Outcome like 'Success%') as num_success,
(select count(*) from SPACEXTBL where Mission_Outcome like 'Failure%') as num_failure;
```

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

num_success	num_failure
100	1

Total num. of successful mission = 100
Total num. of failure mission = 1

Boosters Carried Maximum Payload



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

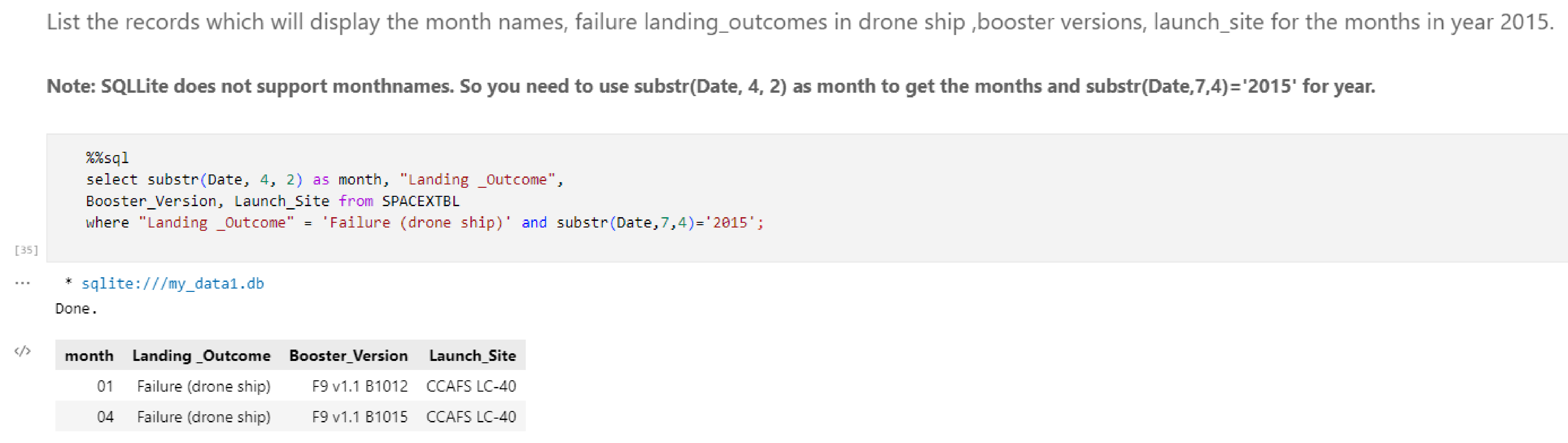
```
%%sql
select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

```
 * 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

The table shows the booster versions that carried the maximum payload mass.

2015 Launch Records



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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

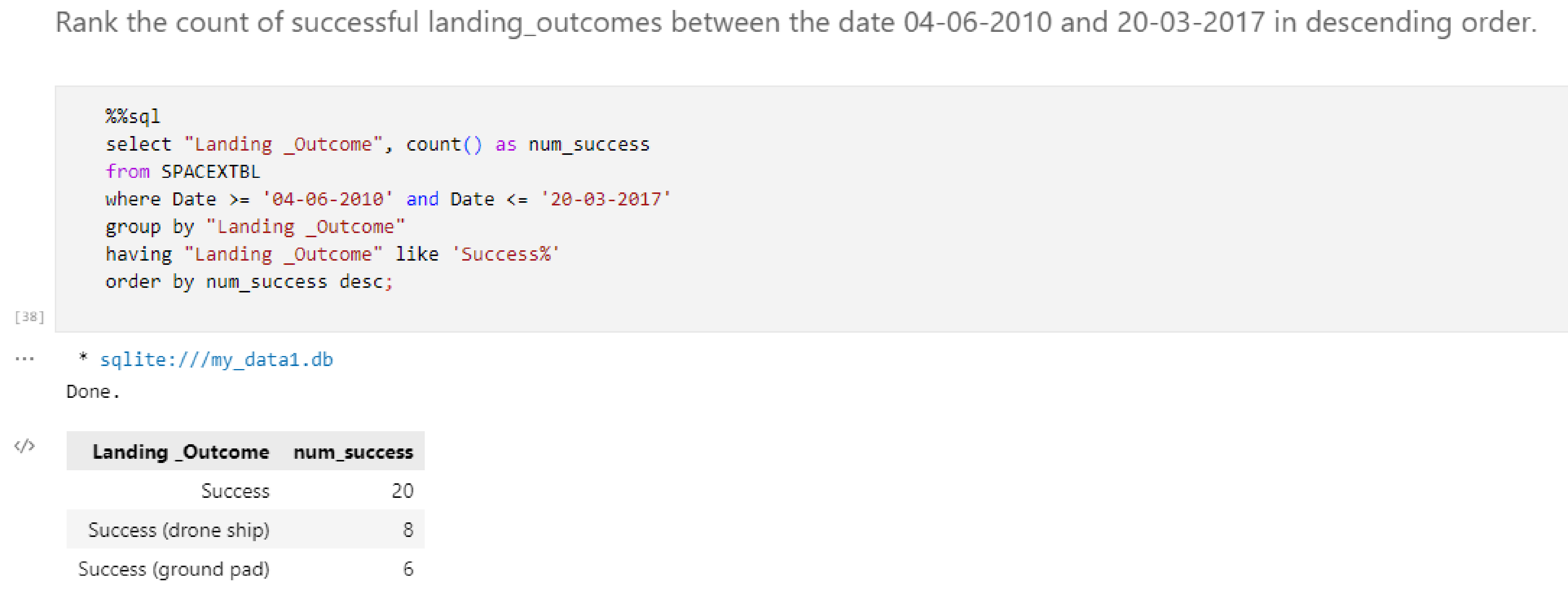
```
%%sql
select substr(Date, 4, 2) as month, "Landing _Outcome",
Booster_Version, Launch_Site from SPACEXTBL
where "Landing _Outcome" = 'Failure (drone ship)' and substr(Date,7,4)='2015';
```

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

month	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

In year 2015, two records of failure landing outcomes in drop ship were found.

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



Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%%sql
select "Landing _Outcome", count() as num_success
from SPACEXTBL
where Date >= '04-06-2010' and Date <= '20-03-2017'
group by "Landing _Outcome"
having "Landing _Outcome" like 'Success%'
order by num_success desc;
```

[38]

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

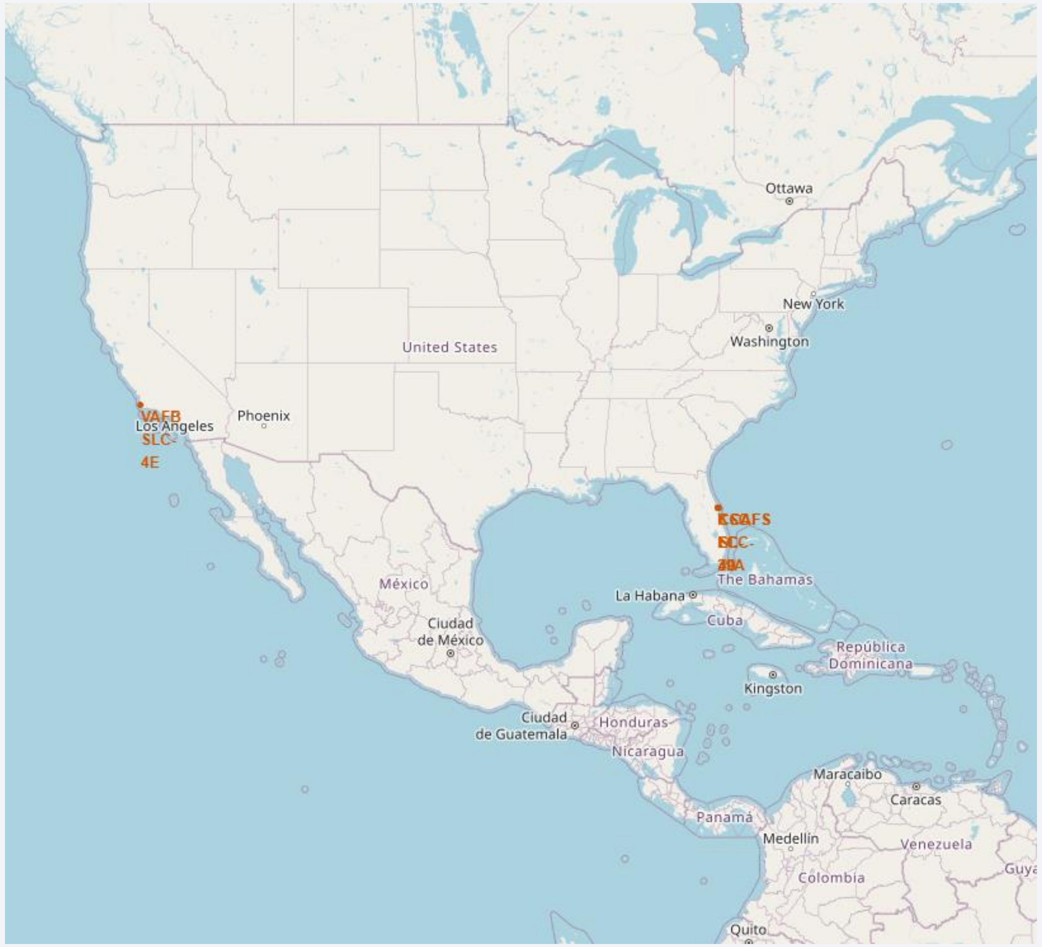
Landing _Outcome	num_success
Success	20
Success (drone ship)	8
Success (ground pad)	6

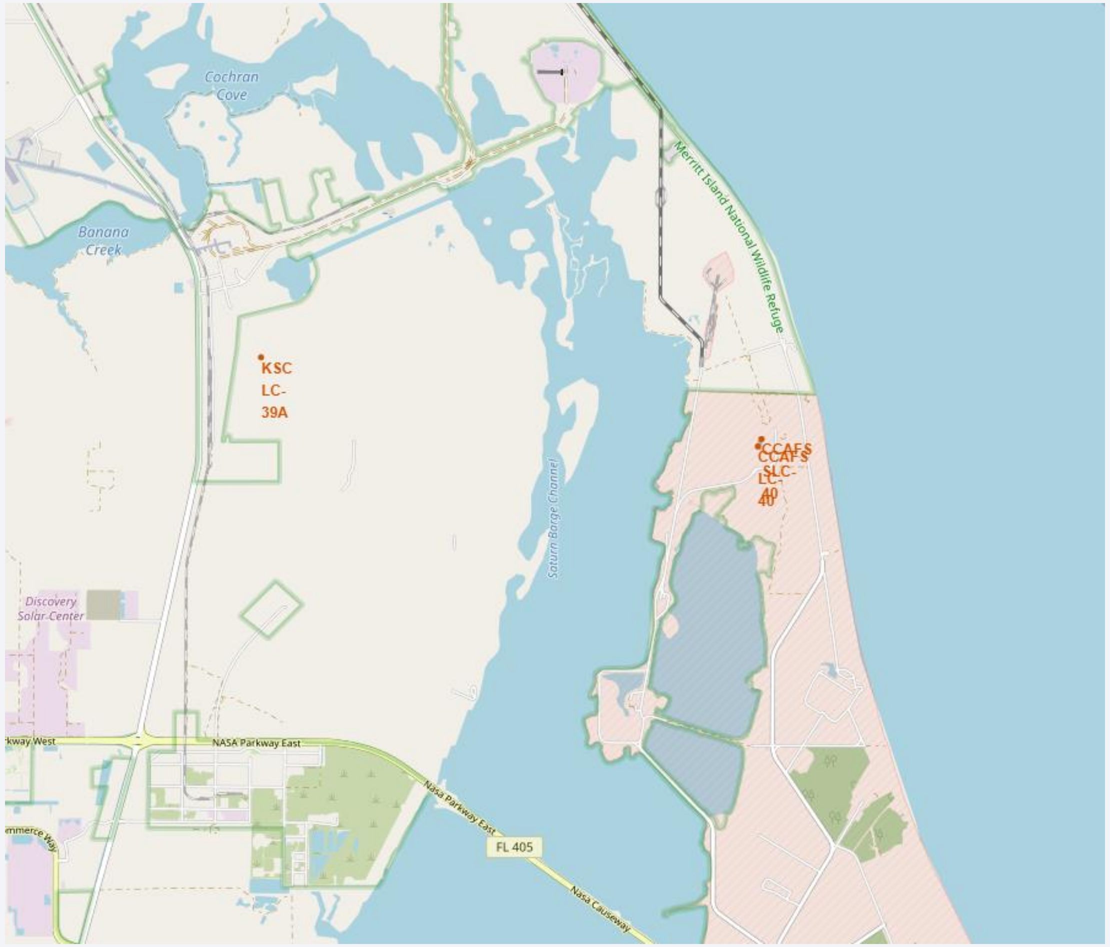
The rank of successful landing outcomes by number of success is shown above.

Section 3

Launch Sites Proximities Analysis

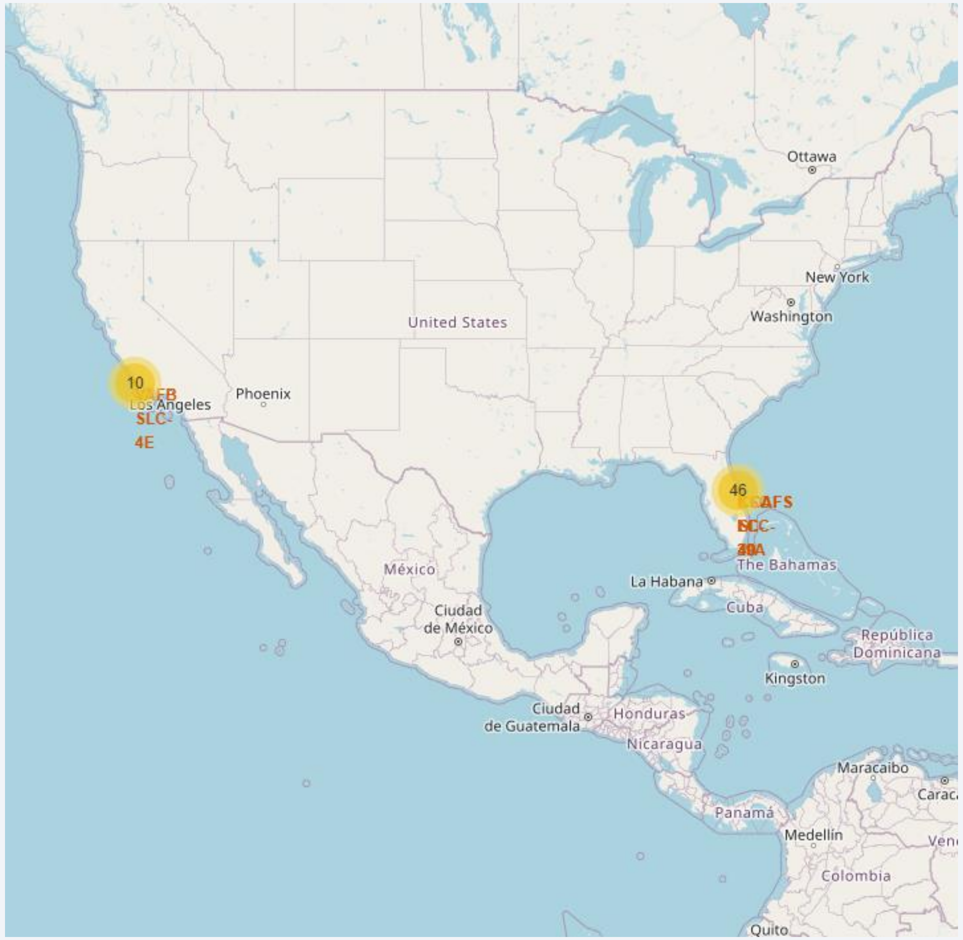
All Launch Sites on a Map

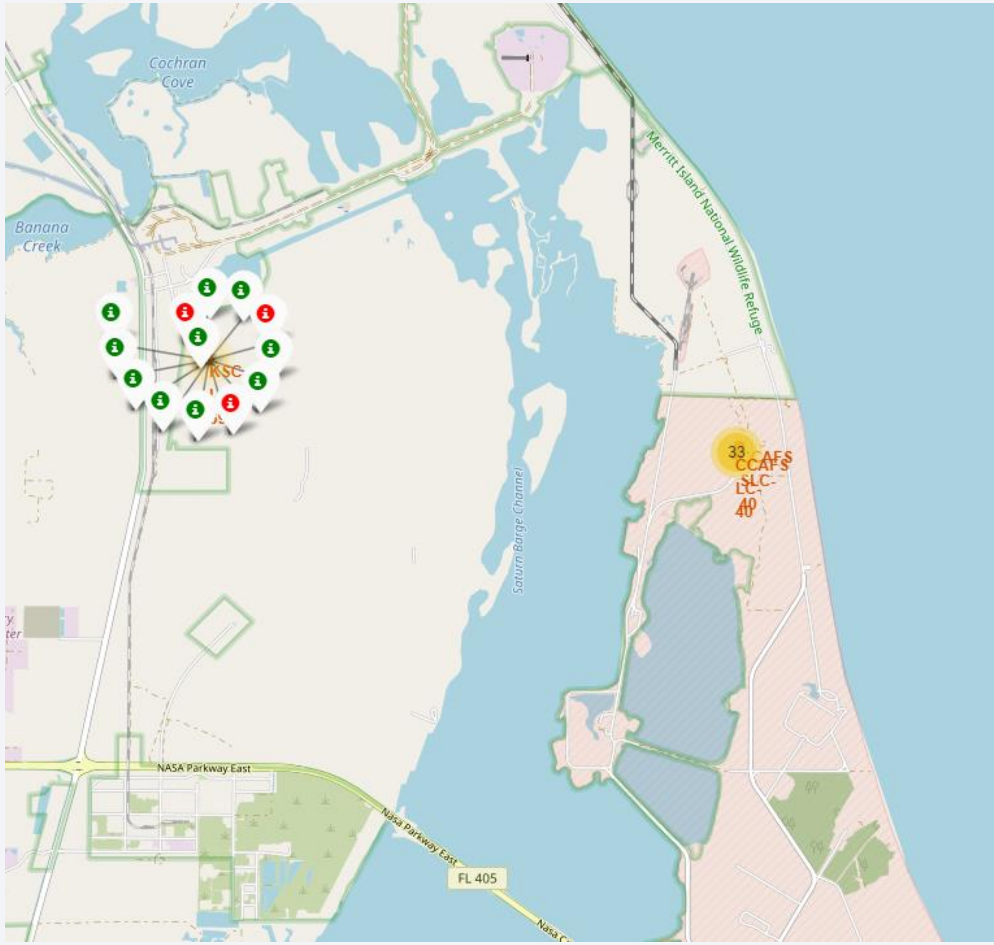




All launch sites are in very close proximity to the coast.

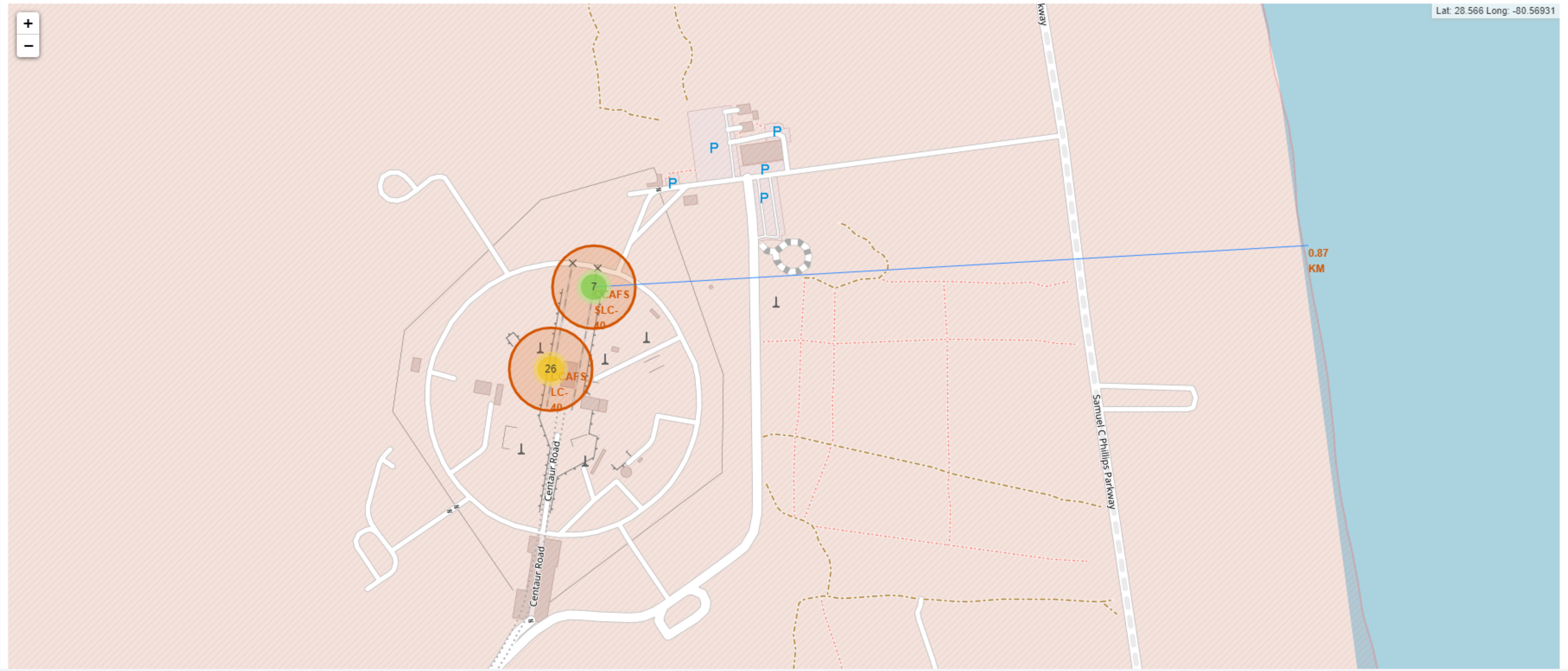
Success/Failed Launches for Each Site on a Map



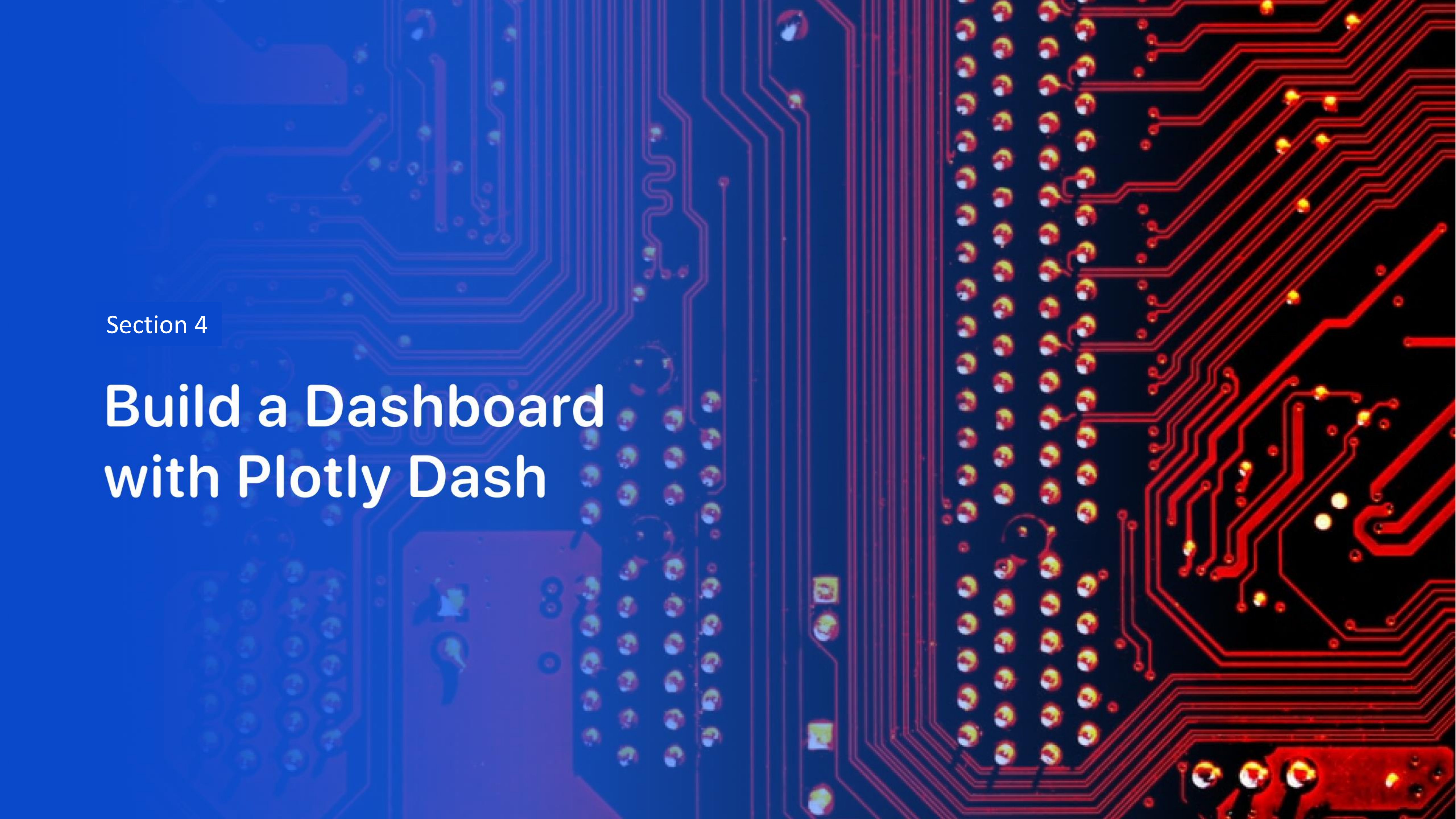


Launch site KSC LC-93A has a highest success rate.

Closest Distance Between a Launch Site and Coastline



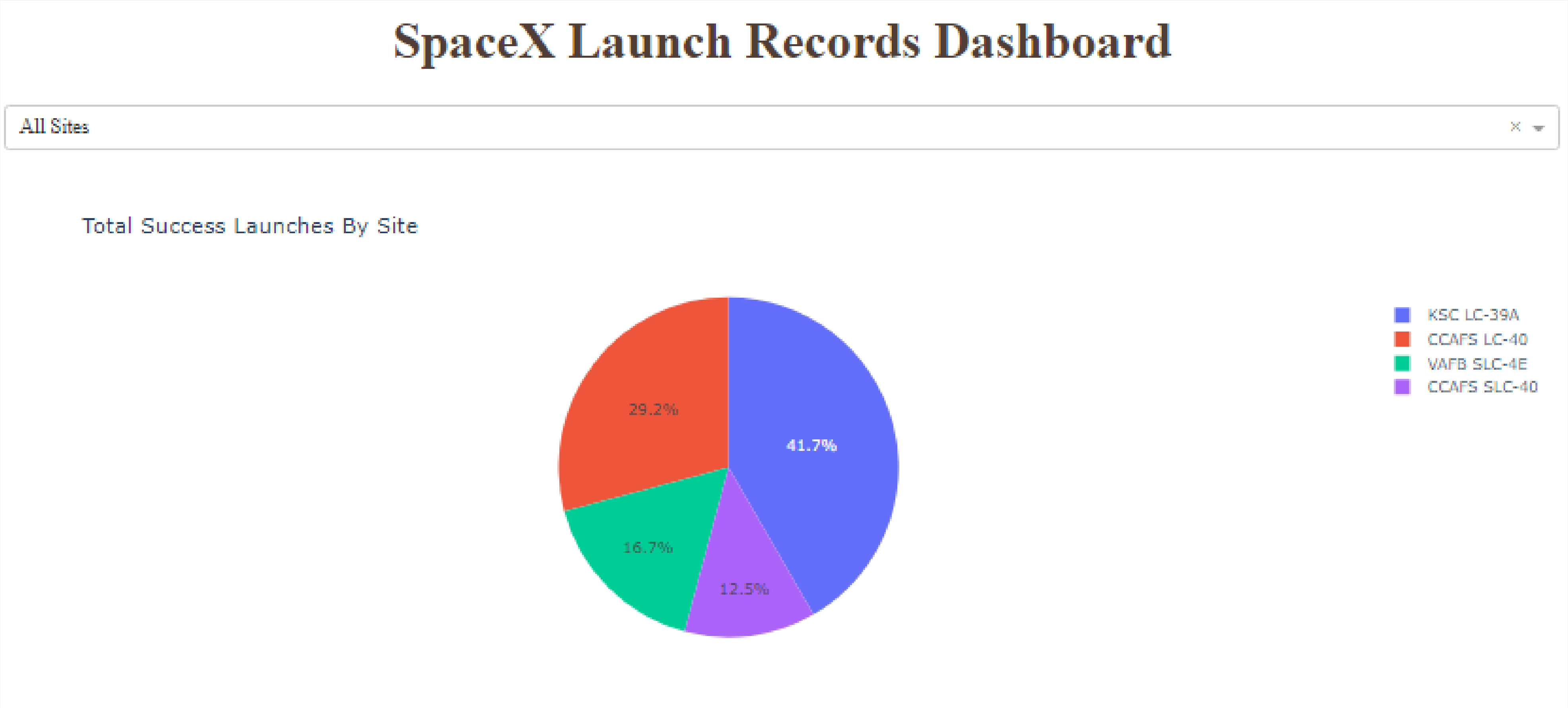
The closest distance between a launch site and coastline is 0.87 km, which is from the site CCAFS SLC-40.



Section 4

Build a Dashboard with Plotly Dash

Launch Success Count for All Sites

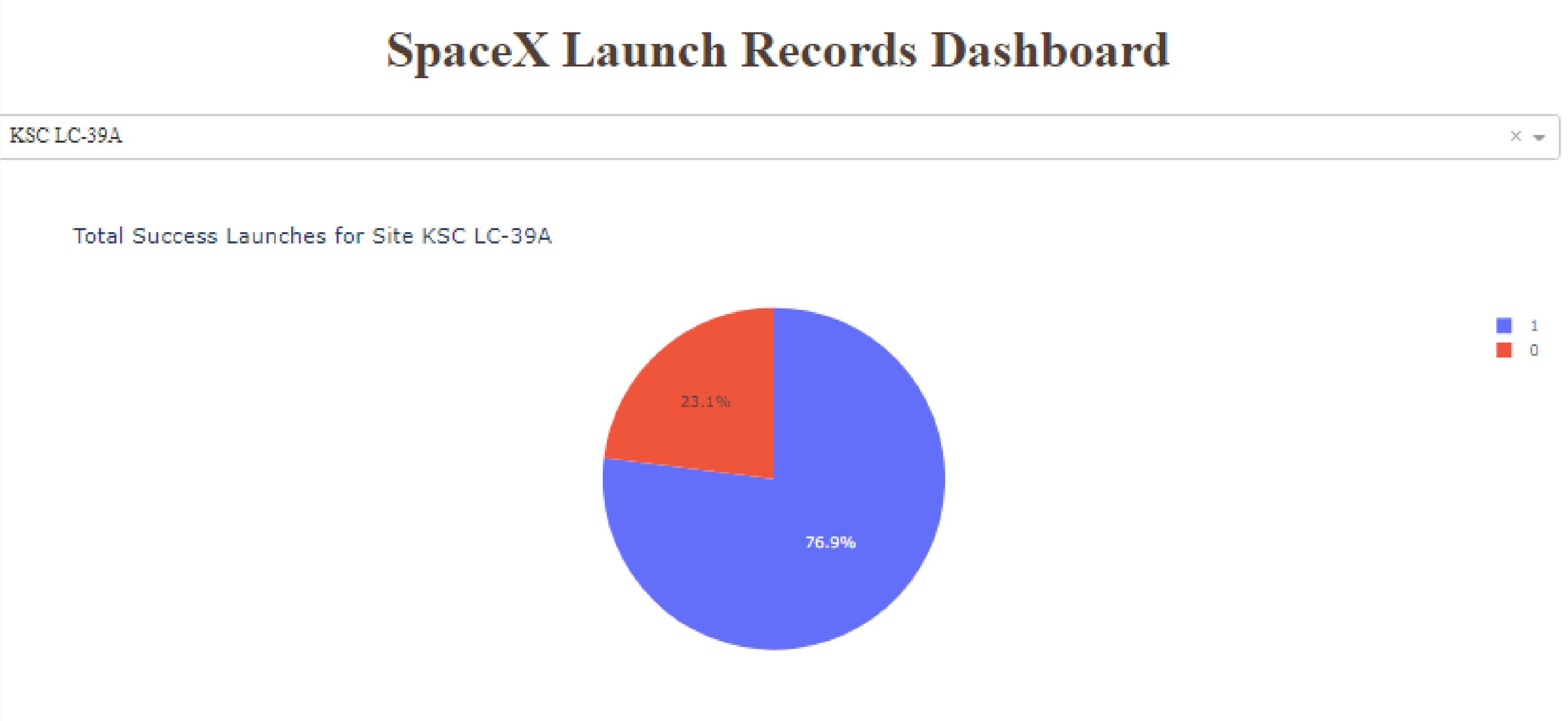


SpaceX Launch Records Dashboard

All Sites

The site KSC LC-39A has the most success launches.

Launch Site with Highest Launch Success Ratio



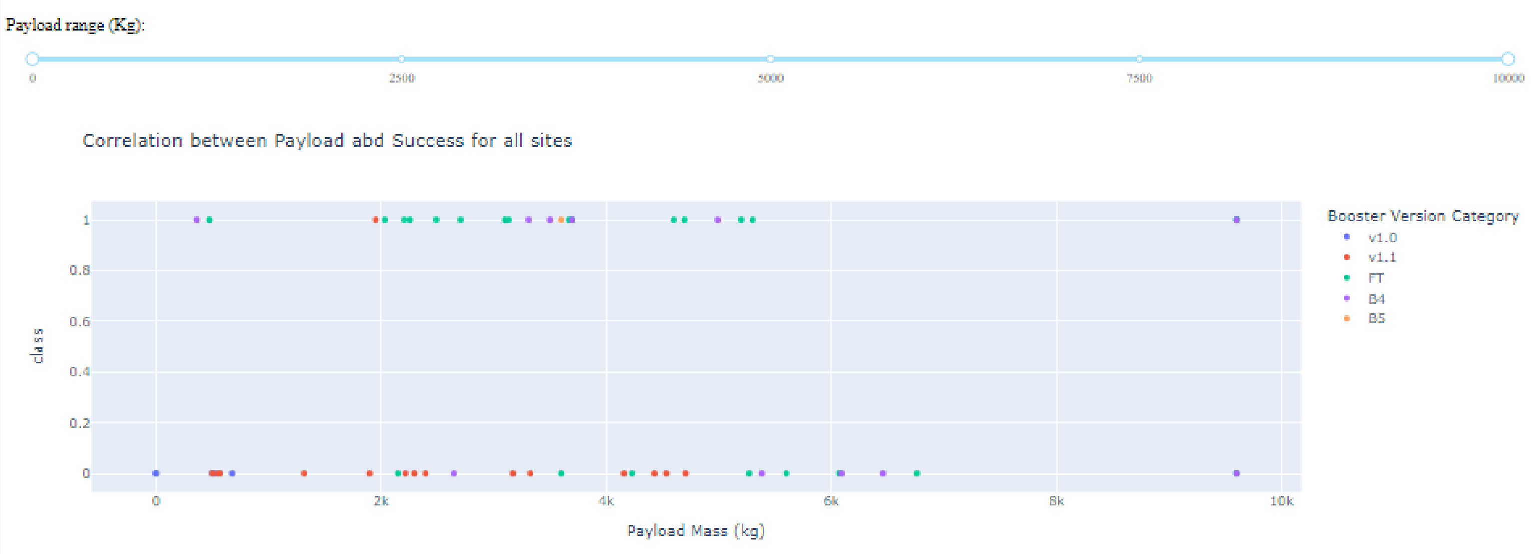
SpaceX Launch Records Dashboard

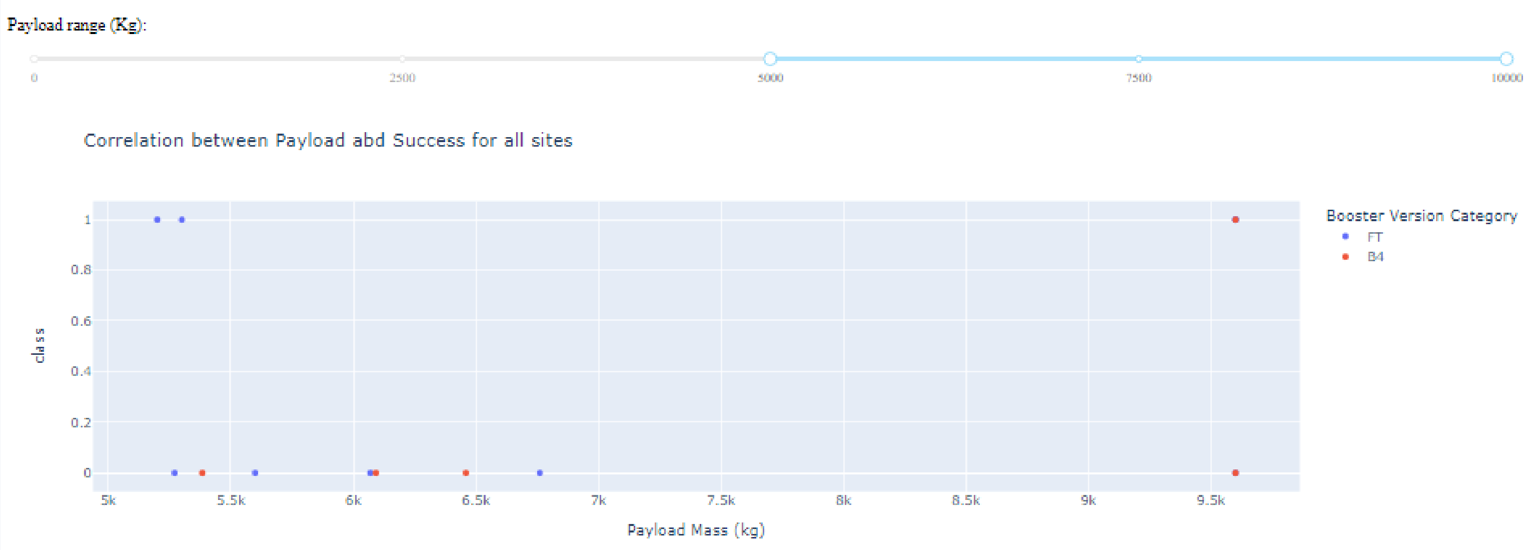
KSC LC-39A

Total Success Launches for Site KSC LC-39A

The site KSC LC-39A has the highest launch success ratio, which is 76.9%.

Payload vs. Launch Outcome for All Sites



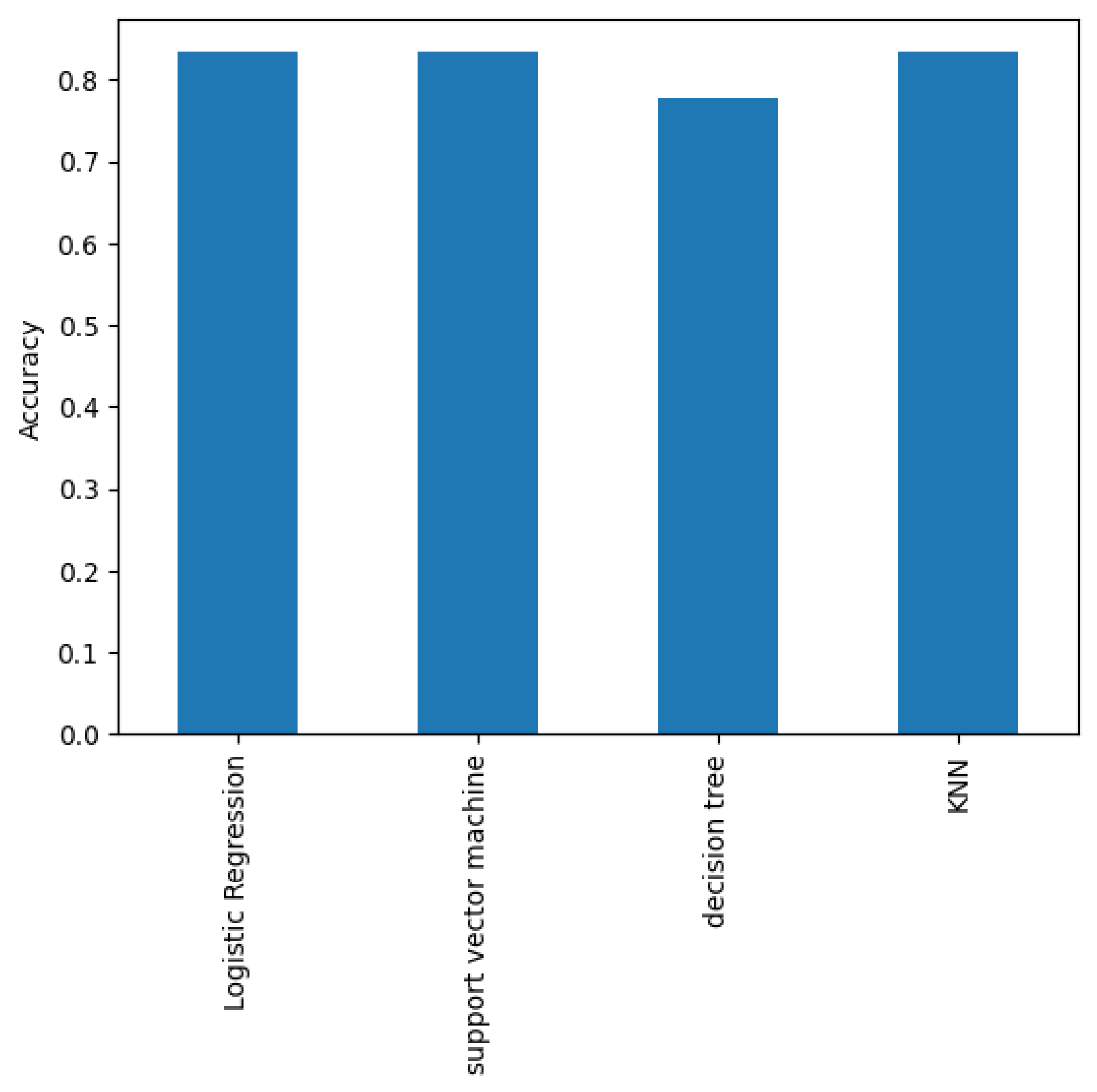


Most success cases are for payload mass between 2000 and 5000 kg.

Section 5

Predictive Analysis (Classification)

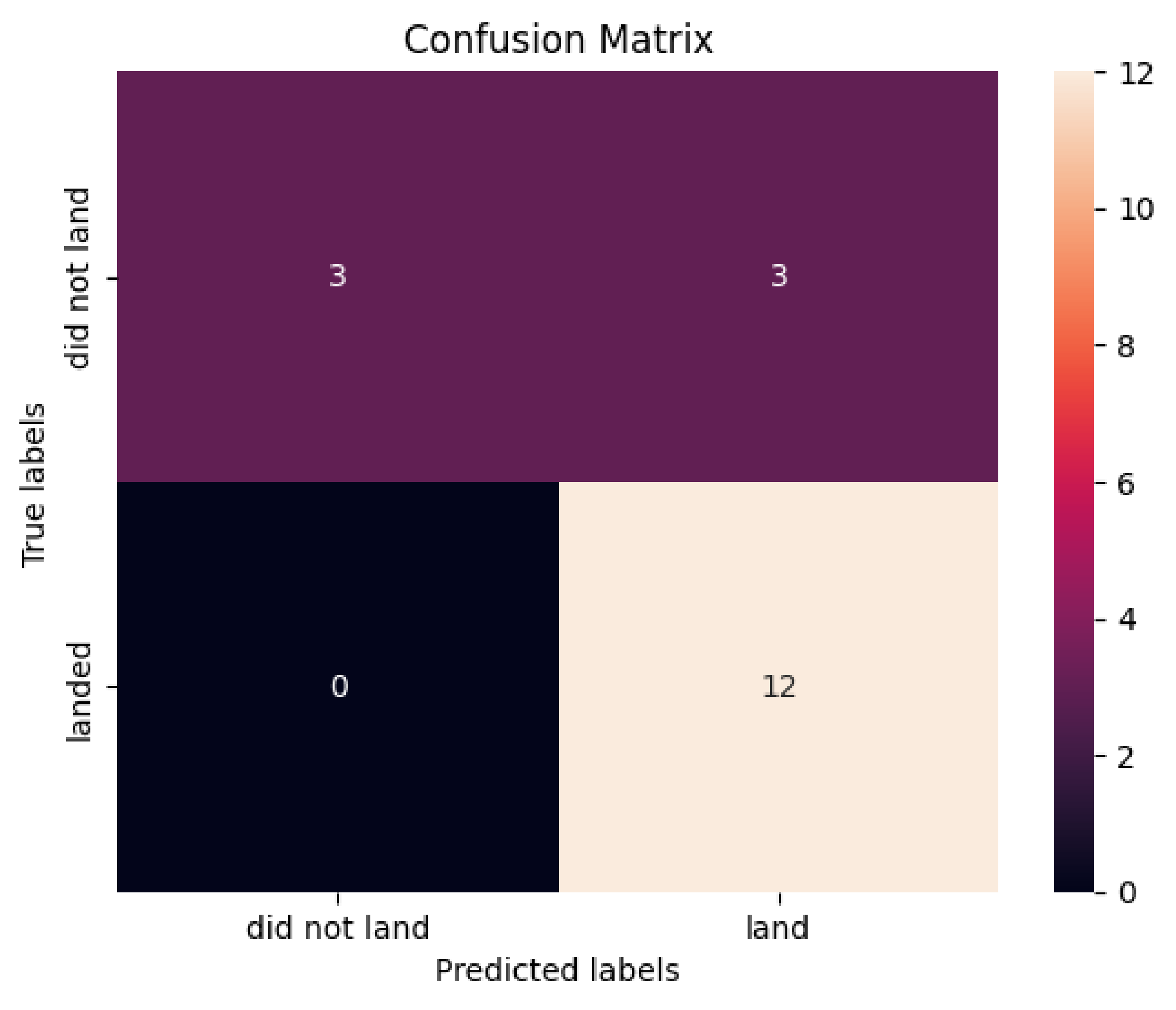
Classification Accuracy



Logistic regression, SVM & KNN shows the same accuracy of testing dataset.

Logistic regression model is selected as the best performing model since it can output the predicted probability as well.

Confusion Matrix



If the landing is in fact successful, the predictions are all correct.

If the landing is in fact failed, the predictions are 50% correct.

Conclusions

- ES-L1, GEO, HEO and SSO are the orbit types with 100% success rate while So is the orbit type with 0% success rate.
- The success rate since 2013 kept increasing till 2020.
- All launch sites are in very close proximity to the coast.
- Launch site KSC LC-93A has a highest success rate.
- Most success cases are for payload mass between 2000 and 5000 kg.
- Trained logistic regression model was selected as the best performing model with predictive accuracy of 83.3%.

Thank you!