IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Avvari Srisujitha

July 3, 2023

TABLE OF CONTENTS

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

Executive Summary

Summary: Choose the accurate and finest model for future prediction

- SpaceX promotes Falcon 9 rocket launches that cost 62 million dollars.
- Compared to other suppliers, SpaceX may save up to 165 million dollars every flight, in large part because the first stage can be recycled.
- The main goal of this study is to use the SpaceX landing data to apply a comprehensive data-driven analytical approach to anticipate a successful conclusion to winning the space race.
- This study involves data collect, data wrangling, web scraping, EDA by using SQL, EDA by using data visualization , building interactive map using folium, interactive dash board creation using plotly and predictive analysis using machine learning

Introduction

Business problem:

- SpaceX promotes 62 million USD Falcon 9 rocket flights. Because the first stage can be recycled, SpaceX could save up to 165 million dollars per flight in comparison to other suppliers.
- The first stage may occasionally be sacrificed by SpaceX depending on the orbit, payloads and the customer.
- The objective of this study is to accurately forecast the Possibility of a first-stage rocket landing successfully as a metric for the cost of launch

Problems to find answers:

- Will SpaceX's first-stage landing be successful?
- What are the factors that influence the SpaceX landing attempt's success?

Section 1

Methodology

Methodology

- Data collection methodology:
  - The data was acquired through the SpaceX API's get request
  - web scraping Wikipedia's records of the Falcon 9 and Falcon Heavy launches.
- Perform data wrangling
  - It involves handling missing values, adding new columns, deleting unwanted columns, and displaying data frames using Pandas.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - compared the performance of four classification models (logistic regression, tree, SVM, and KNN), and select the one with the highest accuracy model among them.

Data Collection

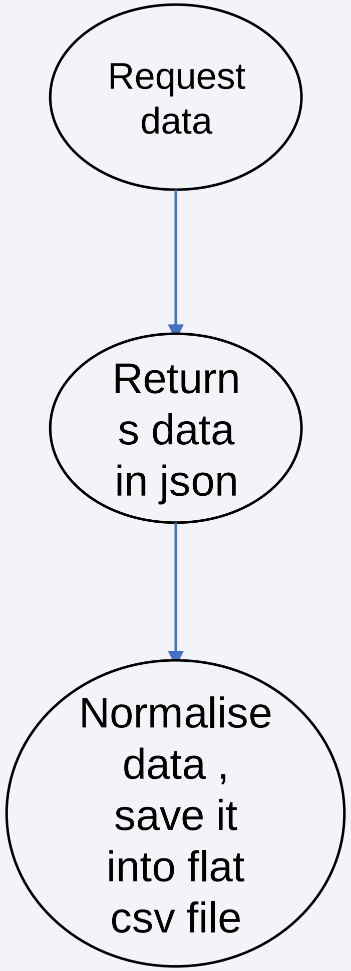
- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

Data collection:

API:

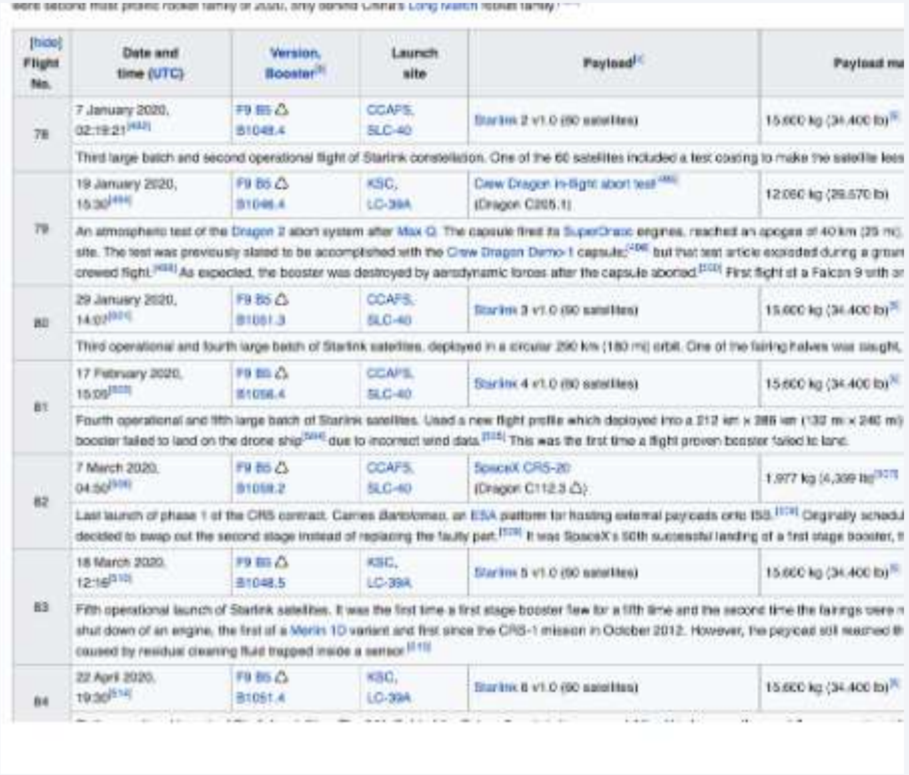
- obtain previous launch data for SpaceX using open source REST API
- GET request was used to get and interpret information on the SpaceX launch.
- Filter the dataframe to only contain launches of the Falcon 9
- Missing payload mass numbers from covert missions were replaced with mean.



Github url: https://github.com/Srisujitha6/AppliedDs/blob/main/jupyter-labs-spacex-data-collection-api.ipynbn

Data Collection - Scraping

- the Falcon9 Launch Wiki request by visiting its url.
- Take note of each column and variable names in the HTML table header.
- Parse the launch HTML tables to produce a data frame.



were second most prolific rocket family of 2020, only behind China's Long March rocket family.

[hide] Flight No.	Date and time (UTC)	Version, Booster[b]	Launch site	Payload[c]	Payload ma
78	7 January 2020, 02:19:21[482]	F9 B5 ♺ B1049.4	CCAFS, SLC-40	Starlink 2 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]
	Third large batch and second operational flight of Starlink constellation. One of the 60 satellites included a test coating to make the satellite less				
79	19 January 2020, 15:30[494]	F9 B5 ♺ B1046.4	KSC, LC-39A	Crew Dragon in-flight abort test[495] (Dragon C205.1)	12,050 kg (26,570 lb)
	An atmospheric test of the Dragon 2 abort system after Max Q. The capsule fired its SuperDraco engines, reached an apogee of 40 km (25 mi), site. The test was previously slated to be accomplished with the Crew Dragon Demo-1 capsule;[496] but that test article exploded during a groun crewed flight.[498] As expected, the booster was destroyed by aerodynamic forces after the capsule aborted.[500] First flight of a Falcon 9 with ar				
80	29 January 2020, 14:07[501]	F9 B5 ♺ B1051.3	CCAFS, SLC-40	Starlink 3 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]
	Third operational and fourth large batch of Starlink satellites, deployed in a circular 290 km (180 mi) orbit. One of the fairing halves was caught,				
81	17 February 2020, 15:05[503]	F9 B5 ♺ B1056.4	CCAFS, SLC-40	Starlink 4 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]
	Fourth operational and fifth large batch of Starlink satellites. Used a new flight profile which deployed into a 212 km × 386 km (132 mi × 240 mi) booster failed to land on the drone ship[504] due to incorrect wind data.[505] This was the first time a flight proven booster failed to land.				
82	7 March 2020, 04:50[506]	F9 B5 ♺ B1059.2	CCAFS, SLC-40	SpaceX CRS-20 (Dragon C112.3 ♺)	1,977 kg (4,359 lb)[507]
	Last launch of phase 1 of the CRS contract. Carries Bartolomeo, an ESA platform for hosting external payloads onto ISS.[508] Originally schedu decided to swap out the second stage instead of replacing the faulty part.[509] It was SpaceX's 50th successful landing of a first stage booster, t				
83	18 March 2020, 12:16[510]	F9 B5 ♺ B1048.5	KSC, LC-39A	Starlink 5 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]
	Fifth operational launch of Starlink satellites. It was the first time a first stage booster flew for a fifth time and the second time the fairings were r shut down of an engine, the first of a Merlin 1D variant and first since the CRS-1 mission in October 2012. However, the payload still reached th caused by residual cleaning fluid trapped inside a sensor.[511]				
84	22 April 2020, 19:30[514]	F9 B5 ♺ B1051.4	KSC, LC-39A	Starlink 6 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]

Github url:

https://github.com/Srisujitha6/AppliedDs/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Define the missing value.
- Estimate the amount of launches each site has had.
- Evaluate the quantity and frequency of each orbit.
- Compute the quantity and frequency of each orbit.
- From the Outcome column, create a landing outcome label.

Github url:

https://github.com/Srisujitha6/AppliedDs/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- Read the dataset using the pandas data frame
- Using scatter plot: visualize the relationship between the Fight Number vs. Payload Mass, Fight Number vs. Launch Site, Payload vs. Launch Site, Orbit vs. Fight Number, Payload vs. Orbit Type, Orbit vs. Payload Mass.
- Using line graph: visualize the relationship between the yearly trends.
- Using bar plot: visualize the relationship between the success rate of each orbit type.

Github url:

https://github.com/Srisujitha6/AppliedDs/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- 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
- List the records which will display the month names, failure landing outcomes in drone sh 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

Github url:

https://github.com/Srisujitha6/AppliedDs/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Map out each of the launch sites.
- Mark each site's both successful and failed launches on the map.
- Determine the separations between a launch location and its surroundings.
    - Whether or whether it is near the coast
    - Whether or whether it is near a railway
    - Whether or not it is near a highway
    - Whether or not it is near the city

Github url: https://github.com/Srisujitha6/AppliedDs/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

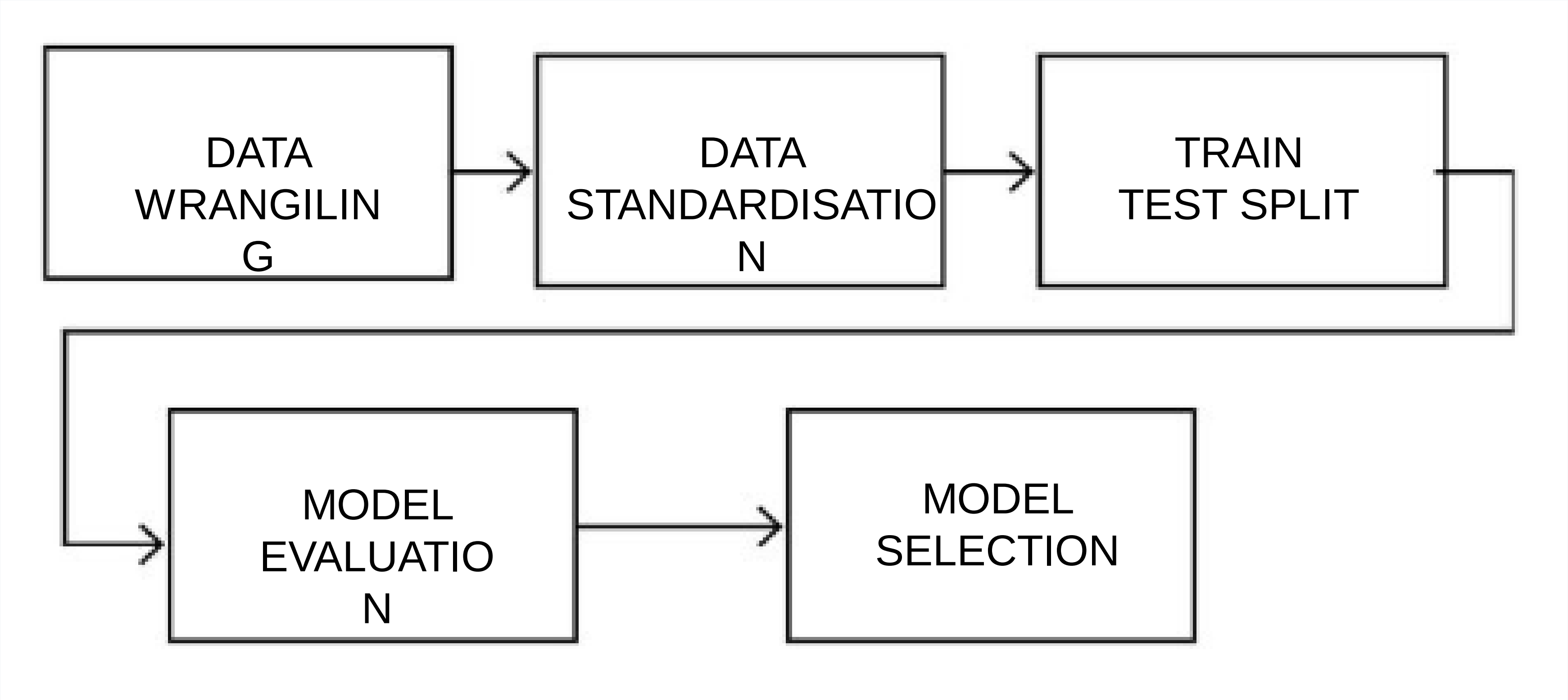
Build a Dashboard with Plotly Dash

- Dashboard Records Launch
- Plotly Dash, a Python interactive dashboarding module, was used to give stakeholders access to interactive, real-time data exploration and manipulation.
- Pie graph displaying the success rate
- based on launch place and colour Graph of payload mass versus landing success Booster version-specific colour coding with a range slider to restrict payload size
- Choose between all sites and individual launch sites using the drop-down menu.

Github url:

https://github.com/Srisujitha6/AppliedDs/blob/main/Dash%20board%20spacex%20launch

Predictive Analysis (Classification)



DATA WRANGILING

DATA STANDARDISATION

TRAIN TEST SPLIT

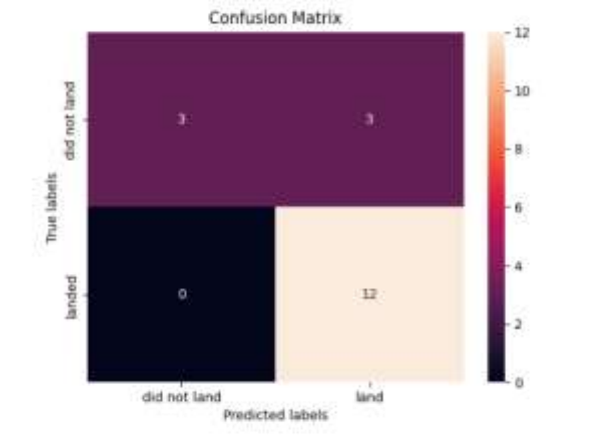
MODEL EVALUATION

MODEL SELECTION

Github url:
https://github.com/Srisujitha6/AppliedDs/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

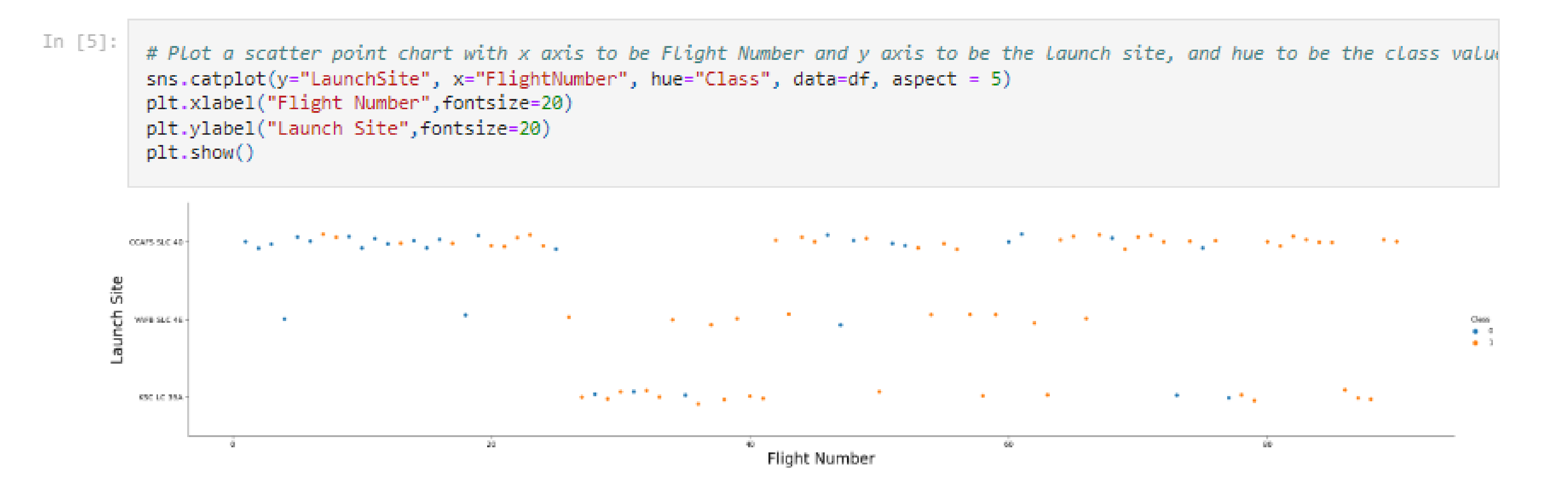
- For this data set the SVM, KNN, and Logistic Regression models had the highest prediction accuracy.
- Low weight payloads perform better than heavier payloads, and SpaceX launch success rates are strongly correlated with the number of years it will take them to perfect their missions.
- From all the sites, KSC LC 39A had the most prosperous launches.
- The best success rate is in Orbit GEO HEO, SSO, and ES L1.



Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

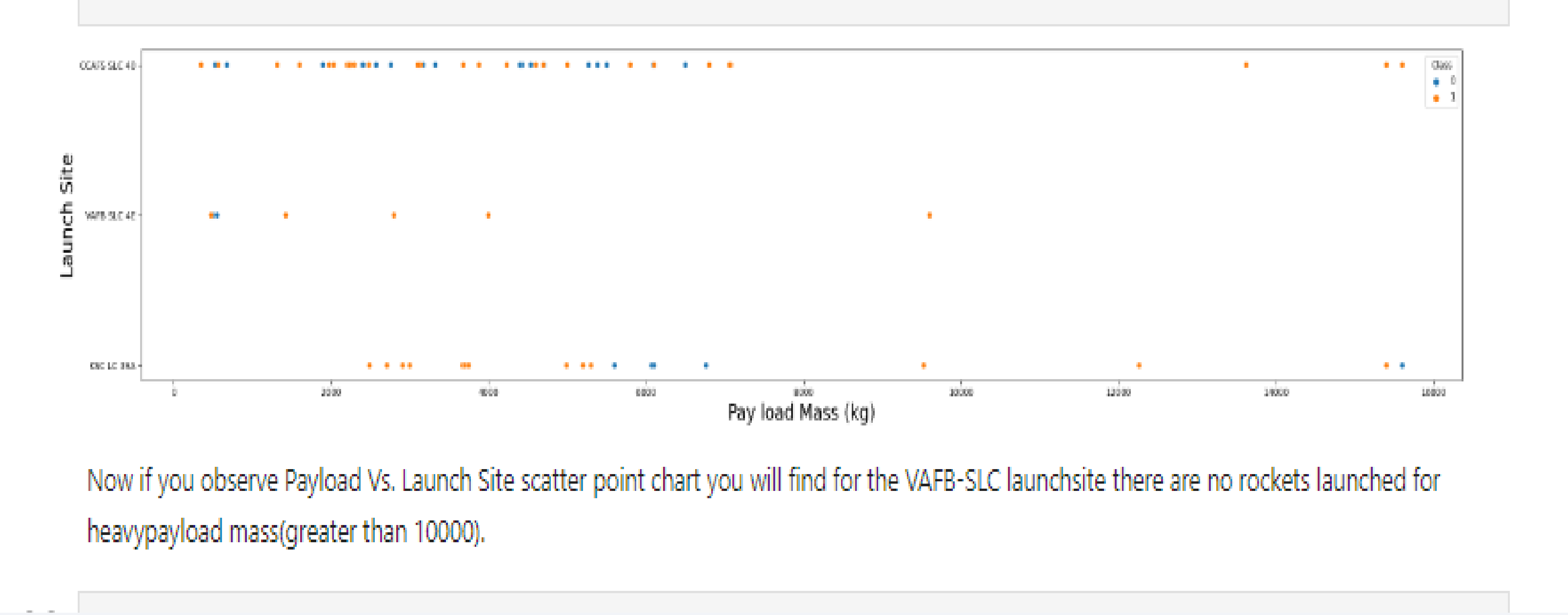


In [5]:

```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class valu
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

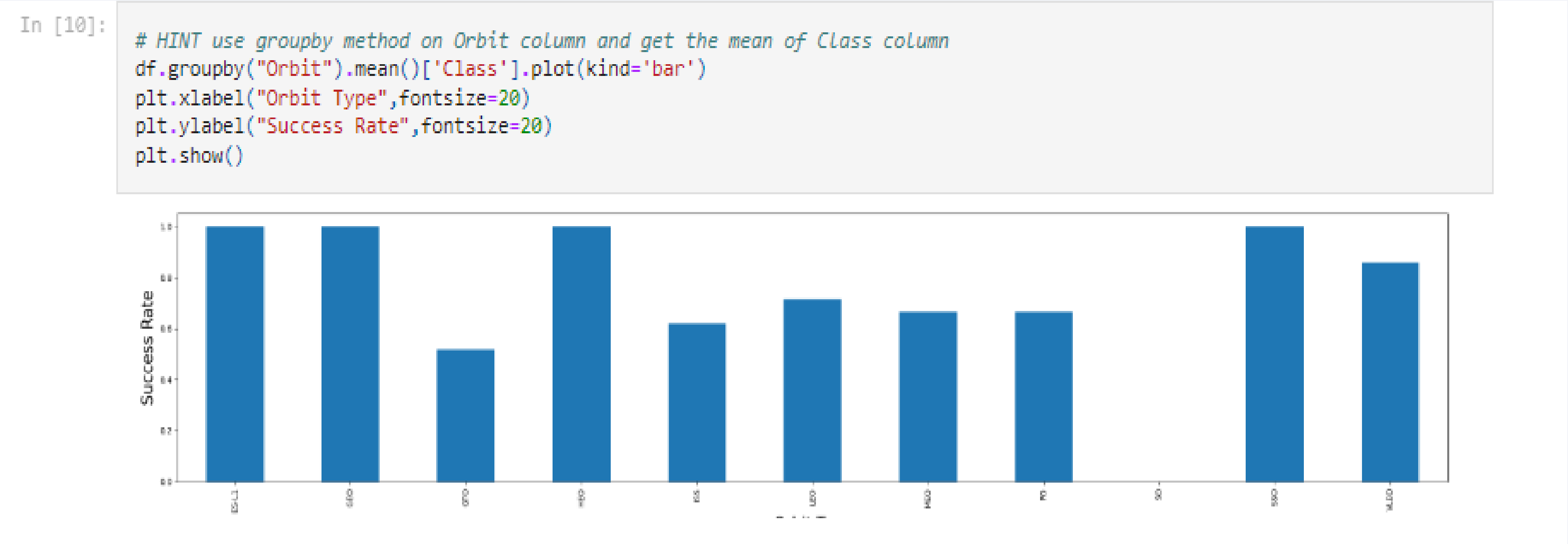
- CCAFS SLC 40 appears to have been where most of the early 1st stage landing failures took place

Payload vs. Launch Site



Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

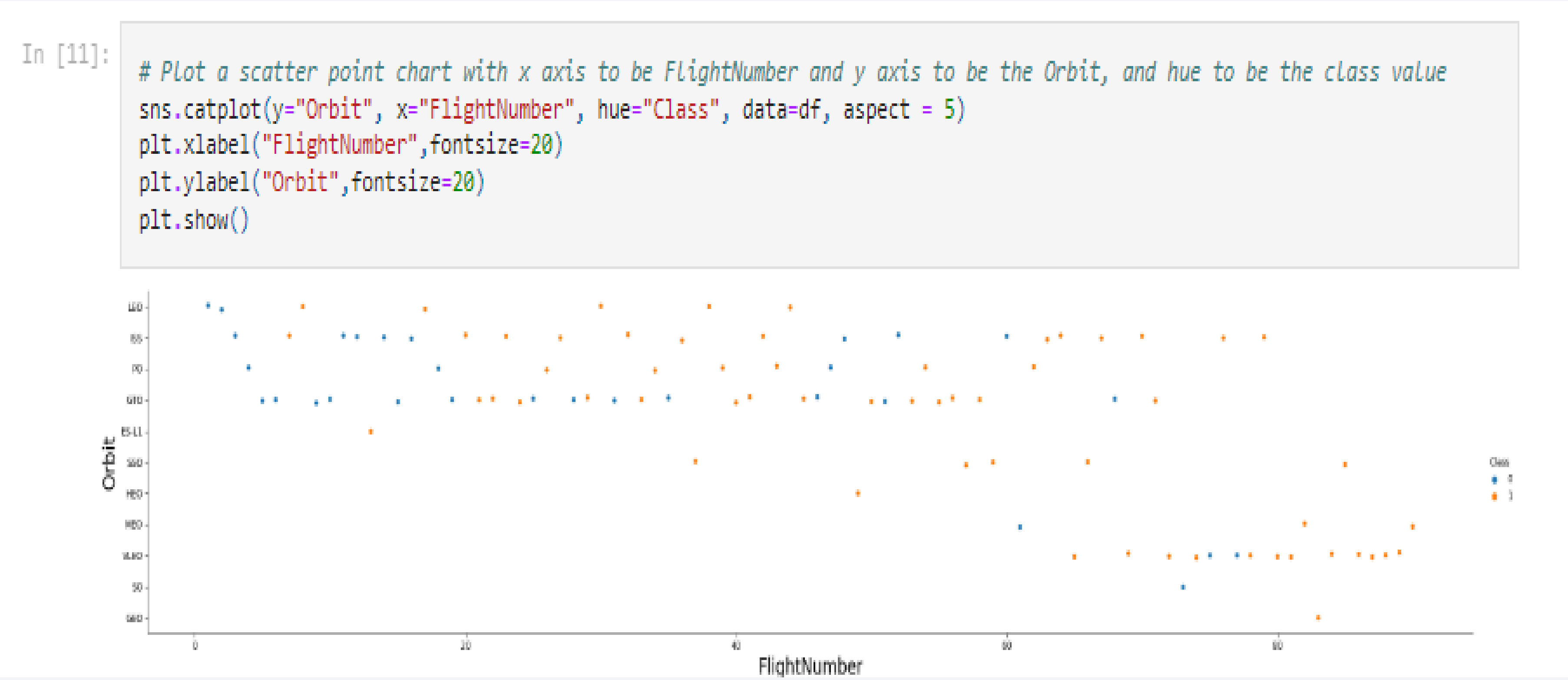
Success Rate vs. Orbit Type



In [10]:

```
# HINT use groupby method on Orbit column and get the mean of Class column
df.groupby("Orbit").mean()['Class'].plot(kind='bar')
plt.xlabel("Orbit Type",fontsize=20)
plt.ylabel("Success Rate",fontsize=20)
plt.show()
```

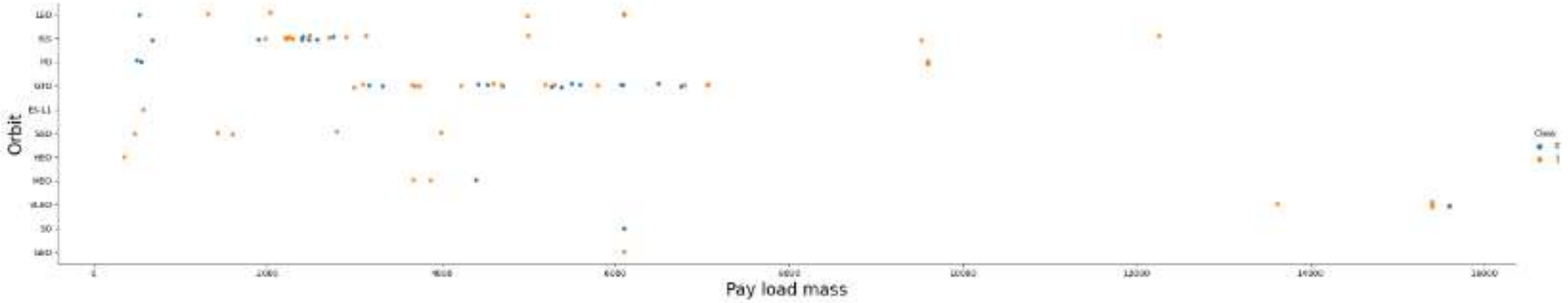
Flight Number vs. Orbit Type



In [11]:

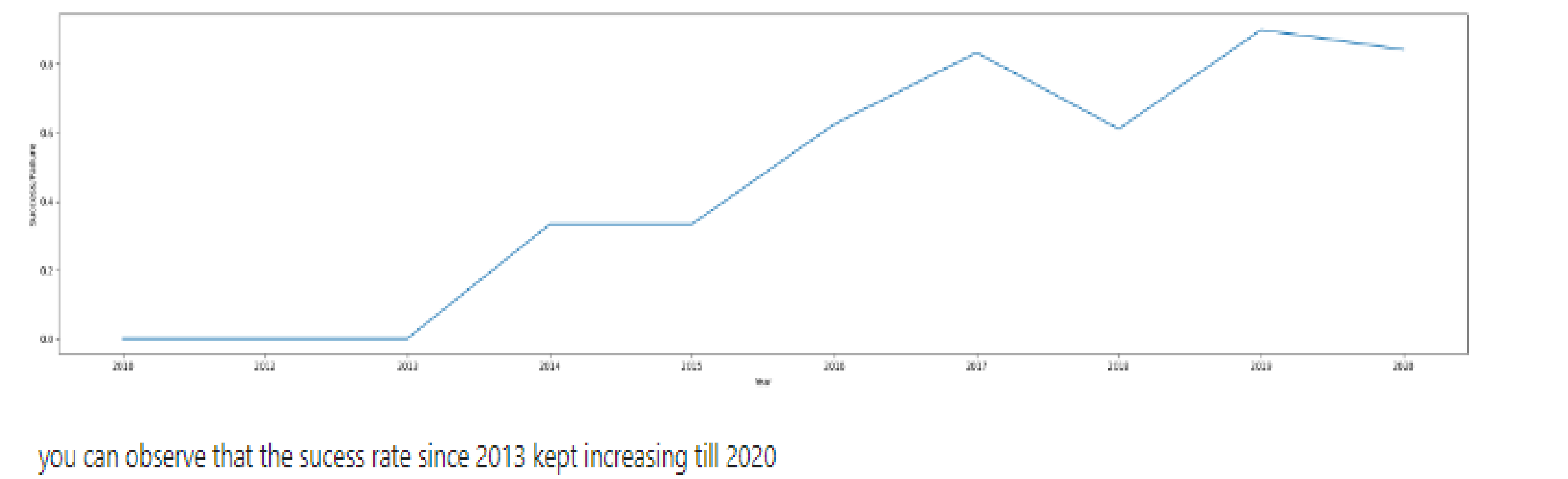
```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("FlightNumber",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

Payload vs. Orbit Type



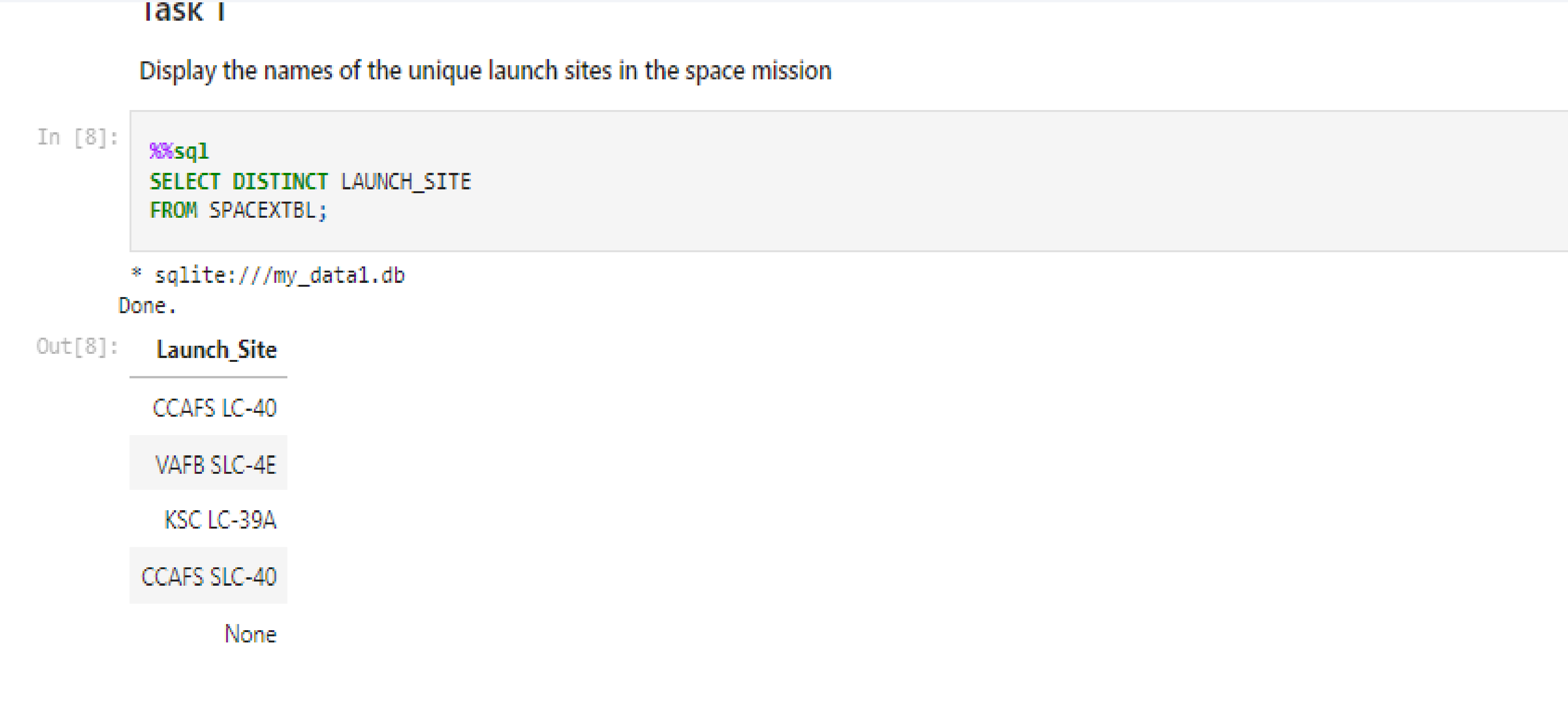
With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

Launch Success Yearly Trend



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names



Task 1

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

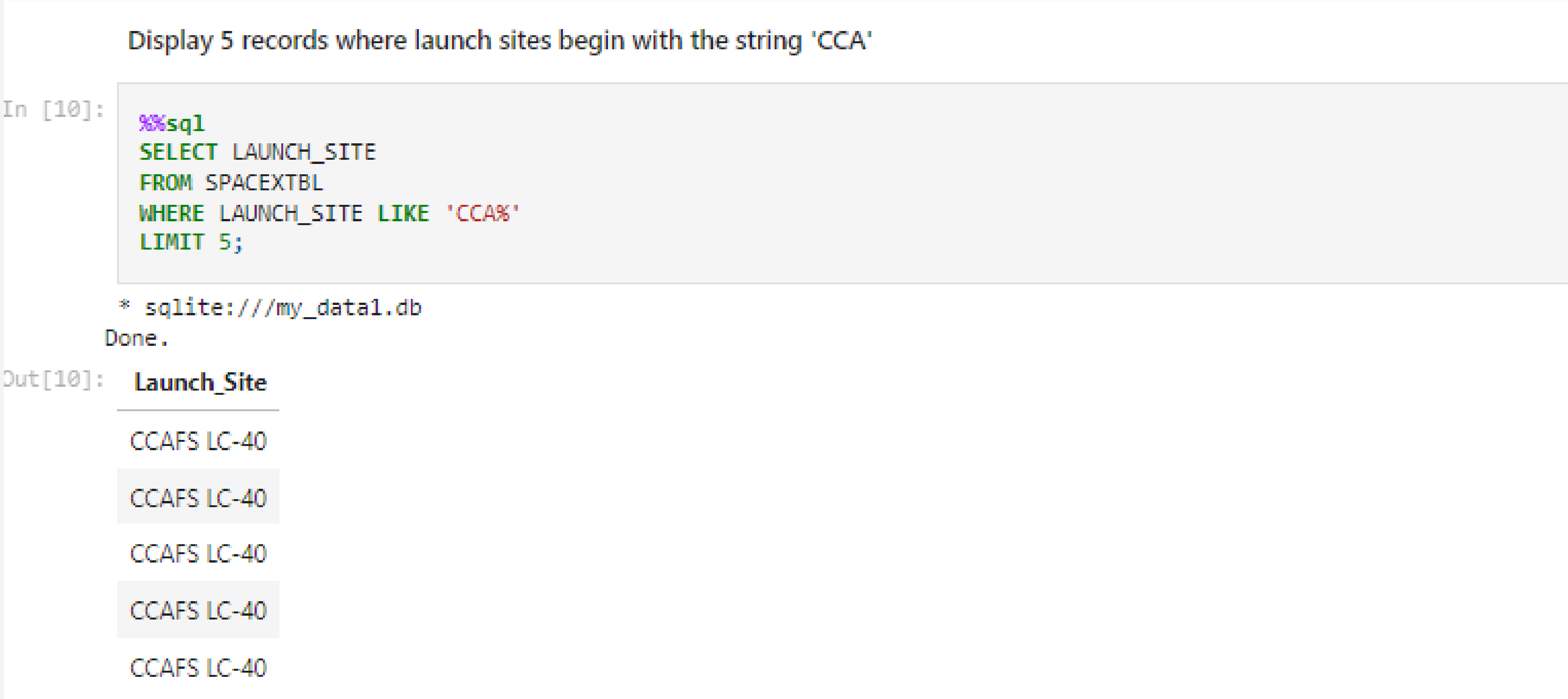
```
In [8]: %%sql
        SELECT DISTINCT LAUNCH_SITE
        FROM SPACEXTBL;
```

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

Out[8]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'



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

In [10]:

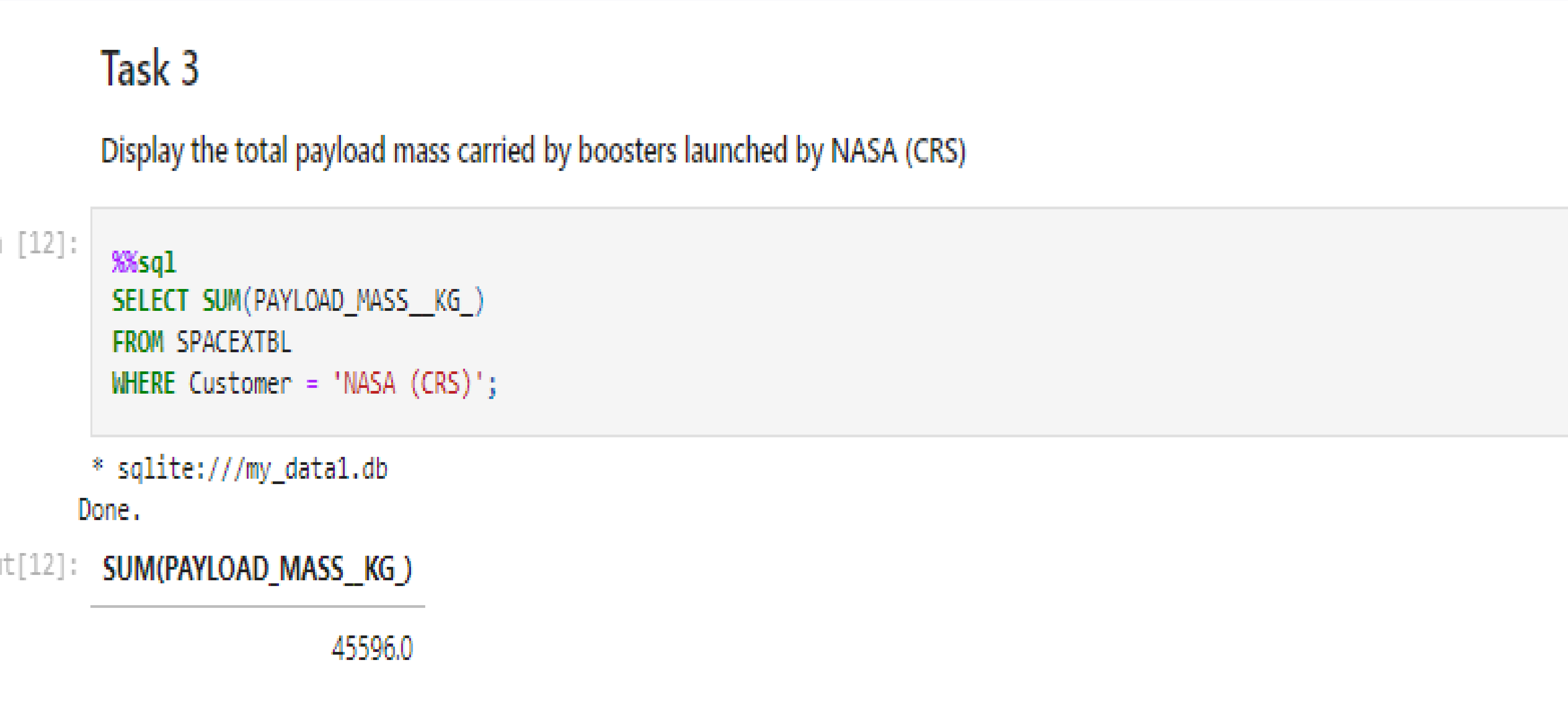
```
%%sql
SELECT LAUNCH_SITE
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

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

Out[10]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass



Task 3

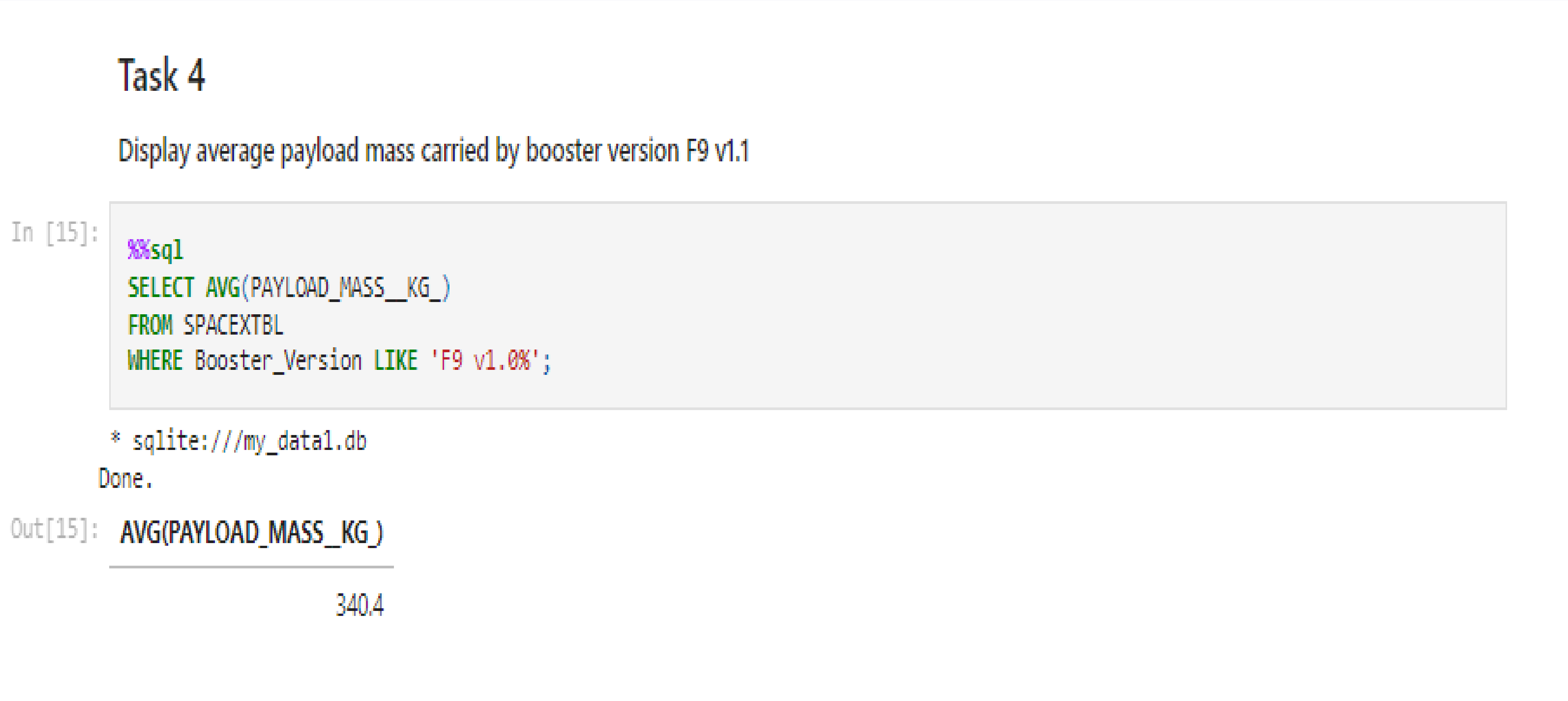
Display the total payload mass carried by boosters launched by NASA (CRS)

```
[12]:

%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

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

```
t[12]:  SUM(PAYLOAD_MASS__KG_)

                       45596.0
```

Average Payload Mass by F9 v1.1



Task 4

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

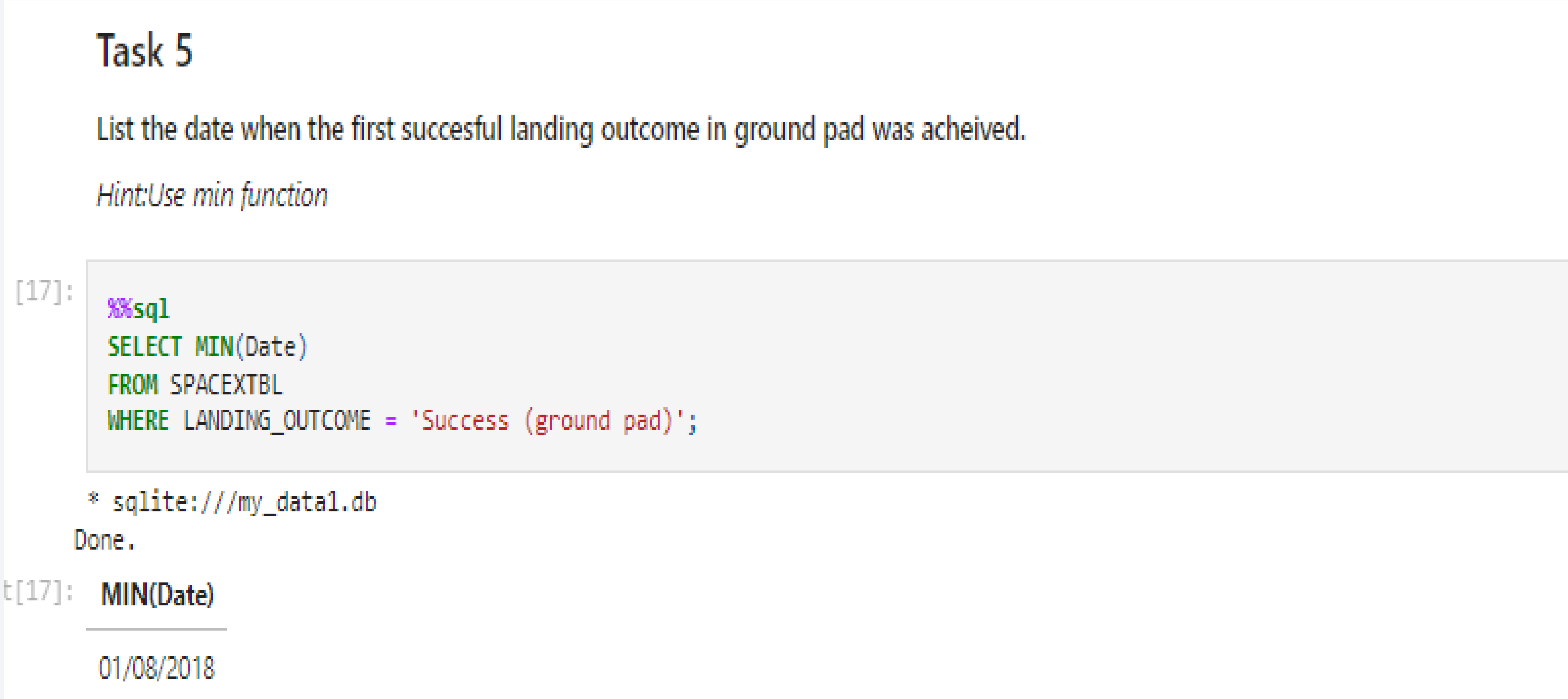
```
In [15]:

%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.0%';
```

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

Out[15]:

AVG(PAYLOAD_MASS_KG_)
340.4

First Successful Ground Landing Date



Task 5

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

Hint:Use min function

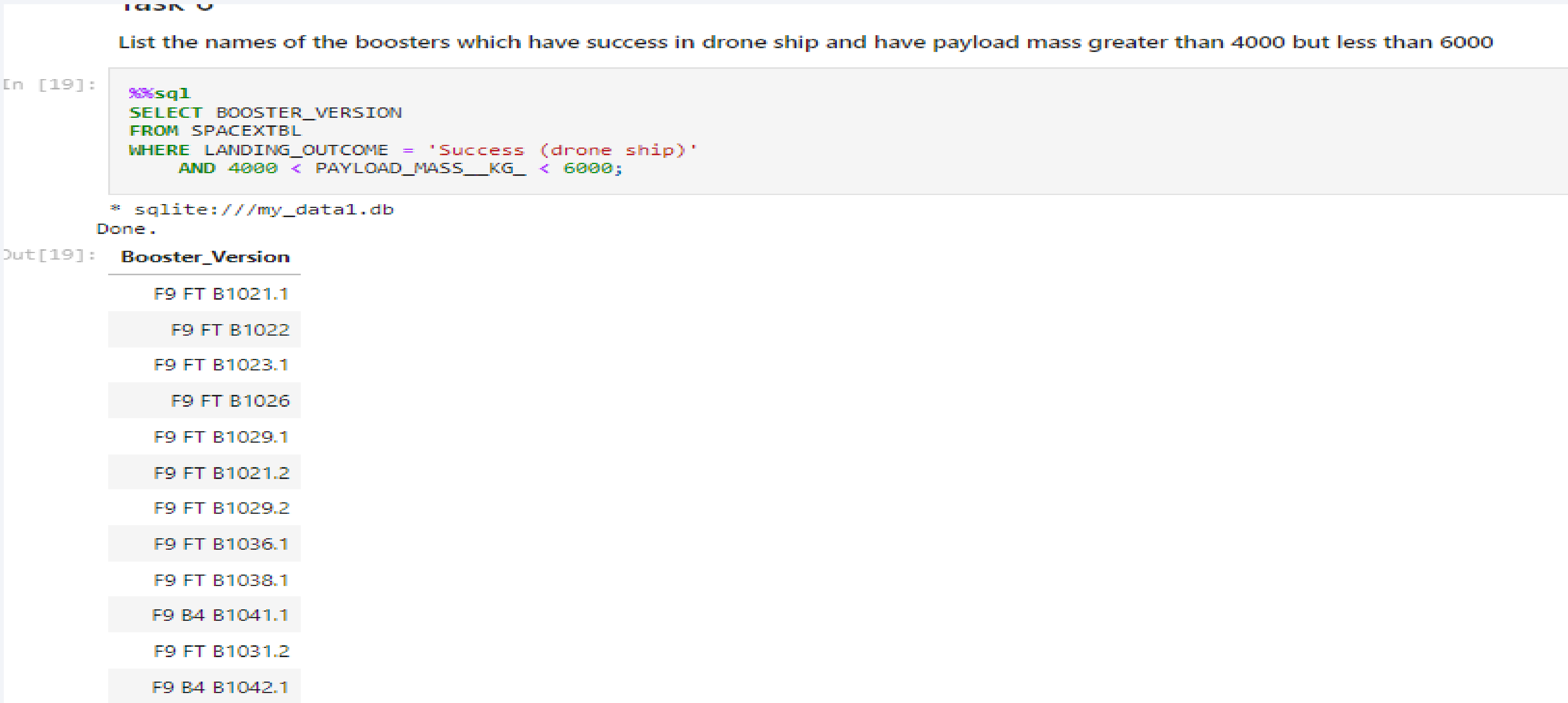
```
[17]:
%%sql
SELECT MIN(Date)
FROM SPACEXTBL
WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

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

t[17]:

MIN(Date)
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000



Task 6

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

In [19]:

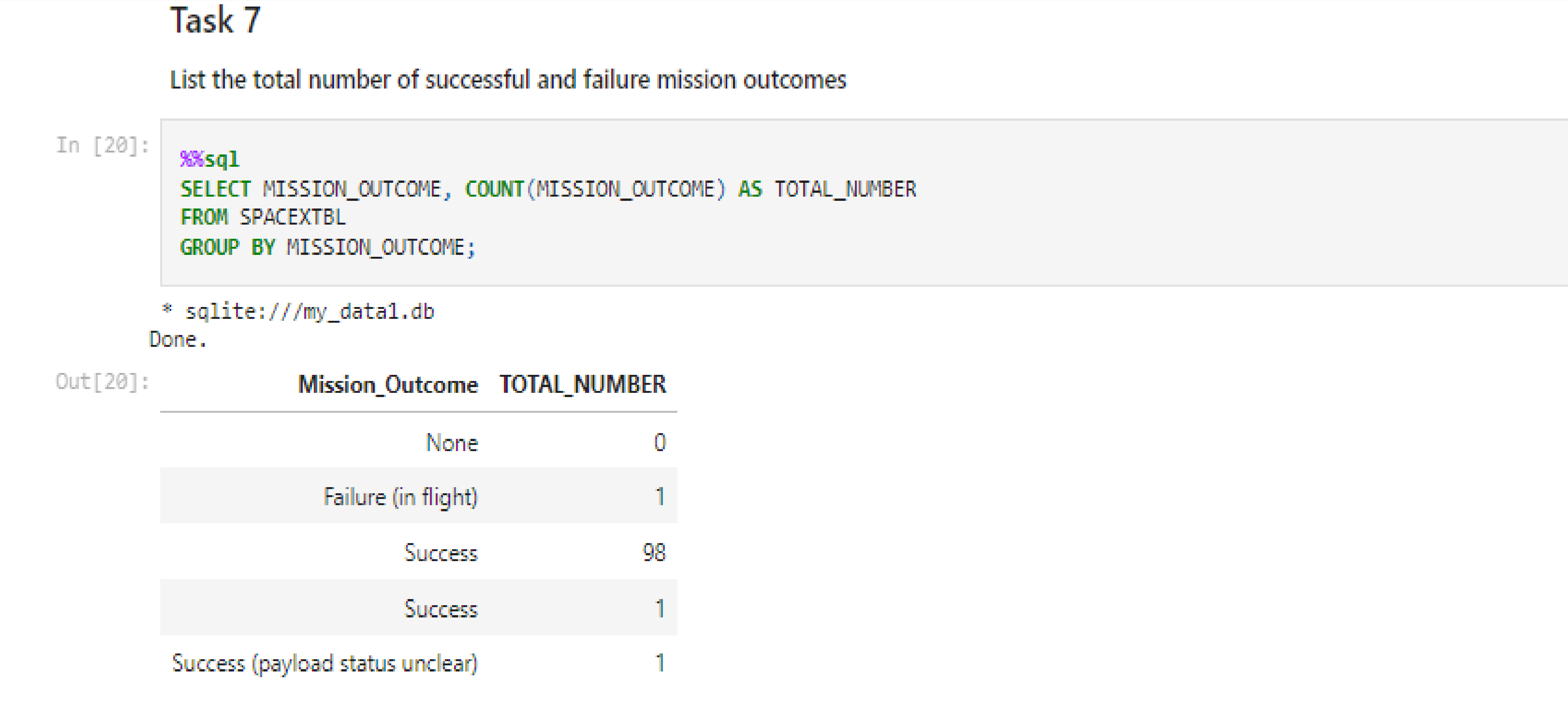
```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING_OUTCOME = 'Success (drone ship)'
    AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

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

Out[19]:

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1

Total Number of Successful and Failure Mission Outcomes



Task 7

List the total number of successful and failure mission outcomes

In [20]:

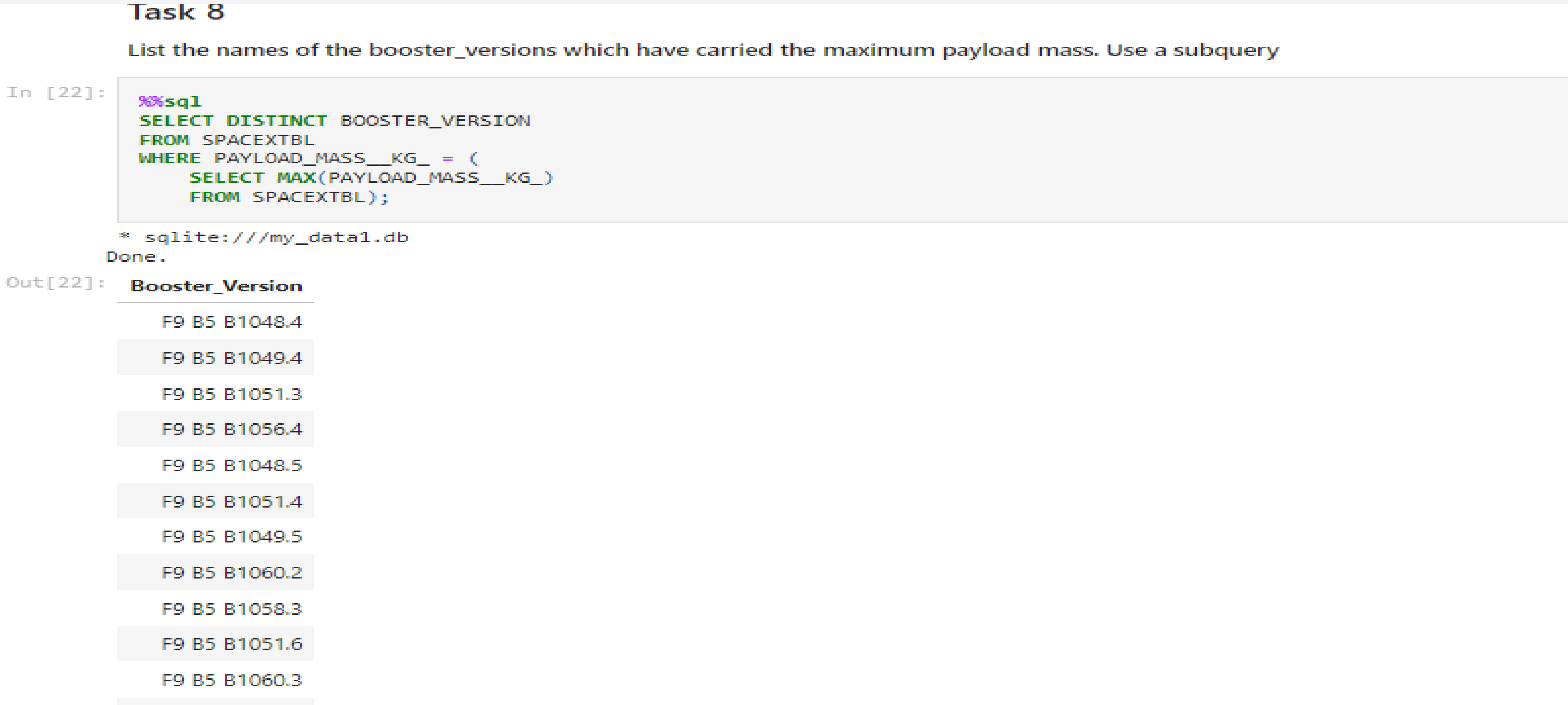
```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

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

Out[20]:

Mission_Outcome	TOTAL_NUMBER
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload



Task 8

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

In [22]:

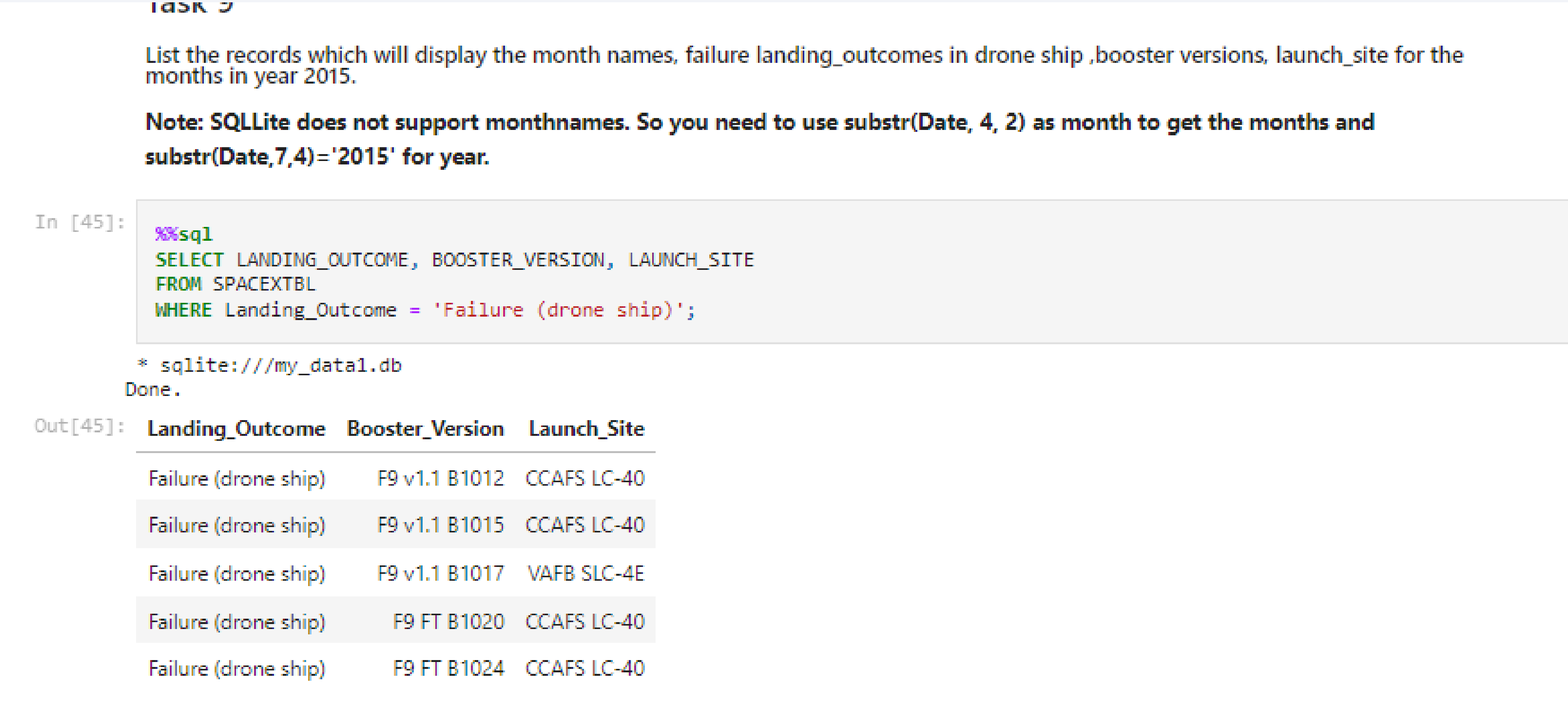
```
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL);
```

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

Out[22]:

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

2015 Launch Records



Task 9

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.

In [45]:

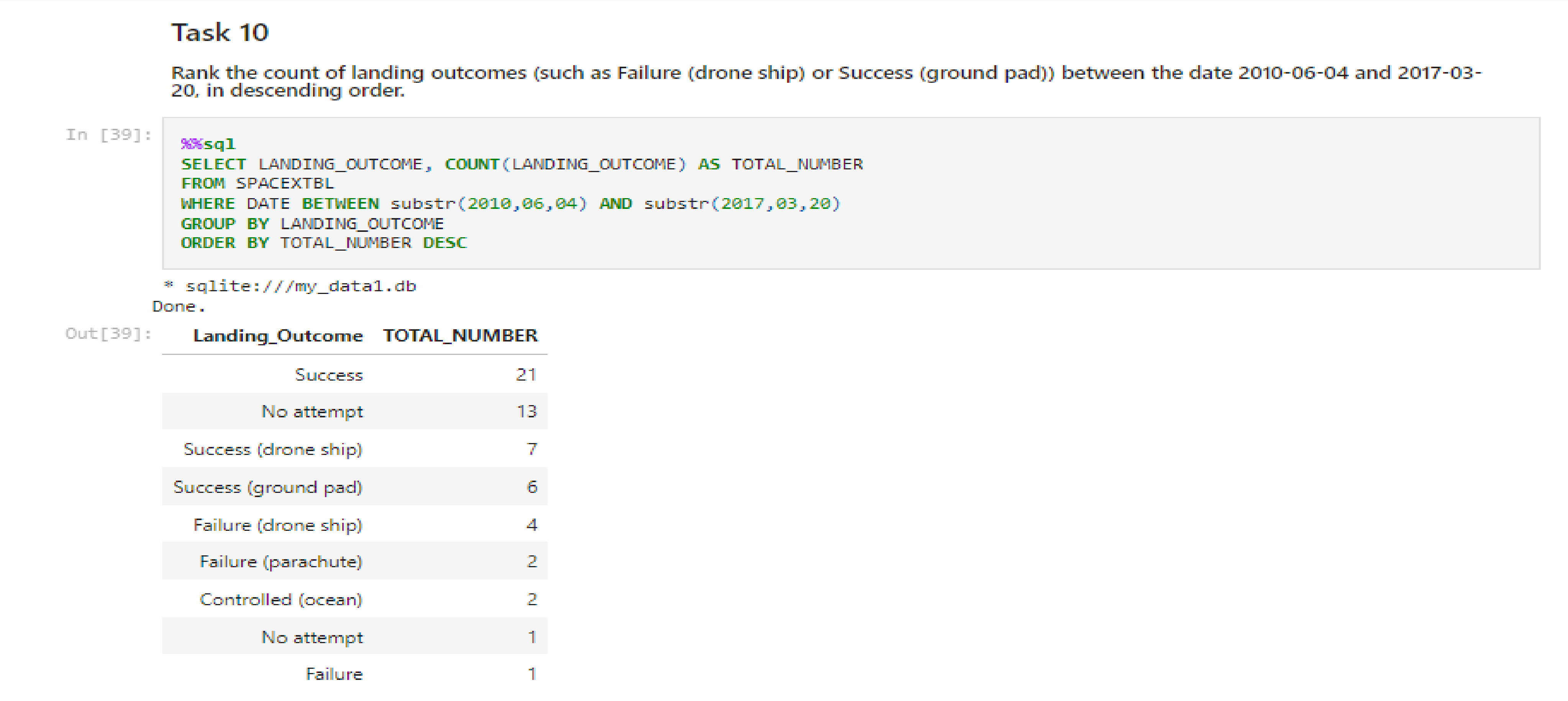
```
%%sql
SELECT LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)';
```

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

Out[45]:

Landing_Outcome	Booster_Version	Launch_Site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E
Failure (drone ship)	F9 FT B1020	CCAFS LC-40
Failure (drone ship)	F9 FT B1024	CCAFS LC-40

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



Task 10

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.

In [39]:

```
%%sql
SELECT LANDING_OUTCOME, COUNT(LANDING_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE DATE BETWEEN substr(2010,06,04) AND substr(2017,03,20)
GROUP BY LANDING_OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

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

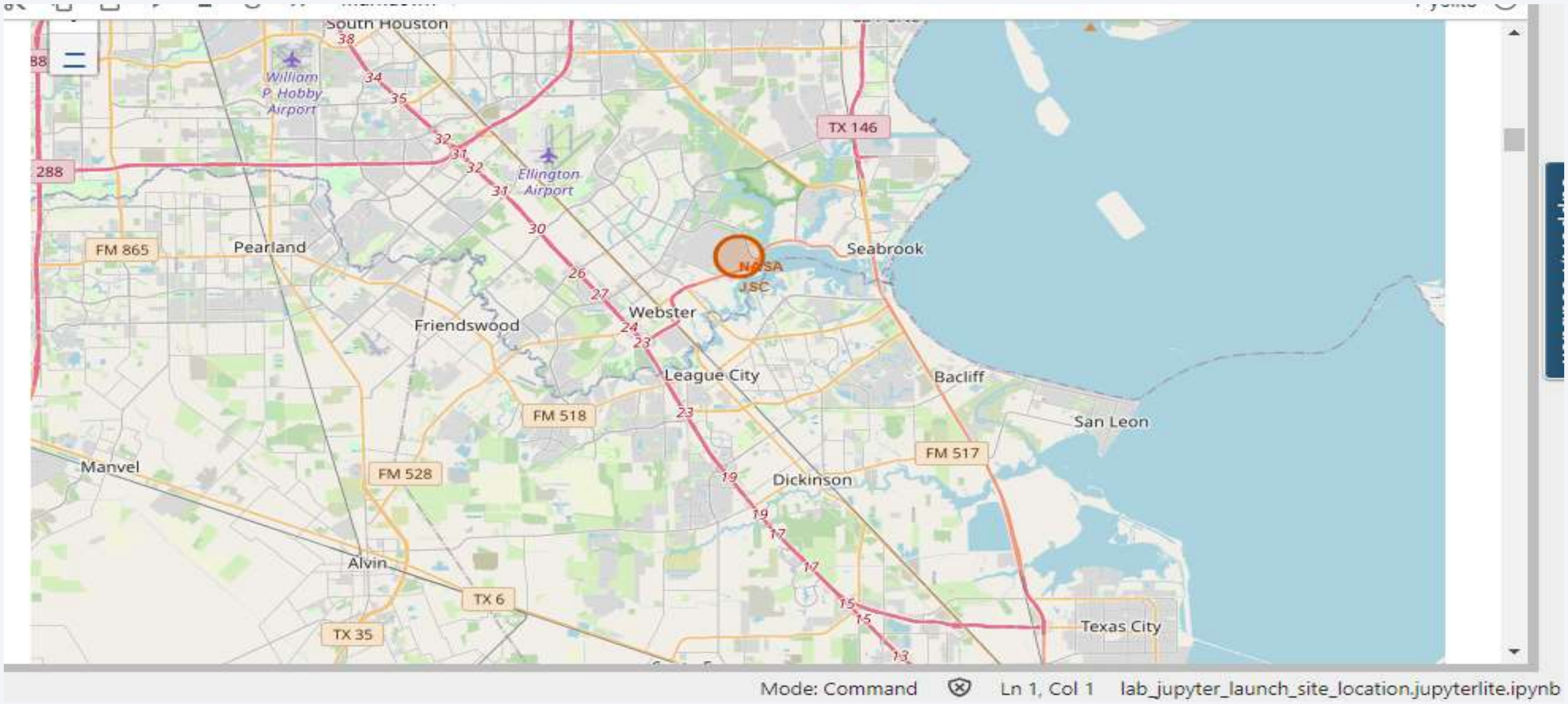
Out[39]:

Landing_Outcome	TOTAL_NUMBER
Success	21
No attempt	13
Success (drone ship)	7
Success (ground pad)	6
Failure (drone ship)	4
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1
Failure	1

Section 3

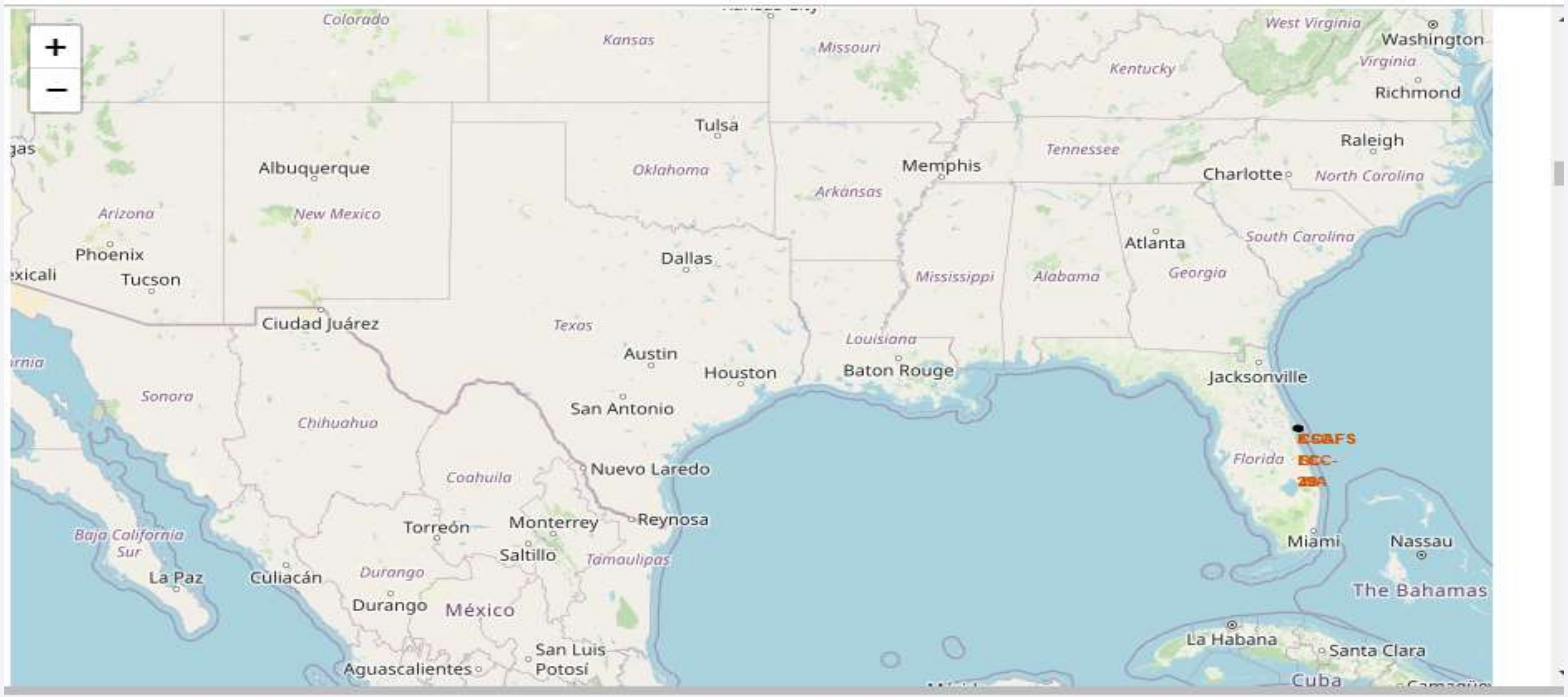
Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

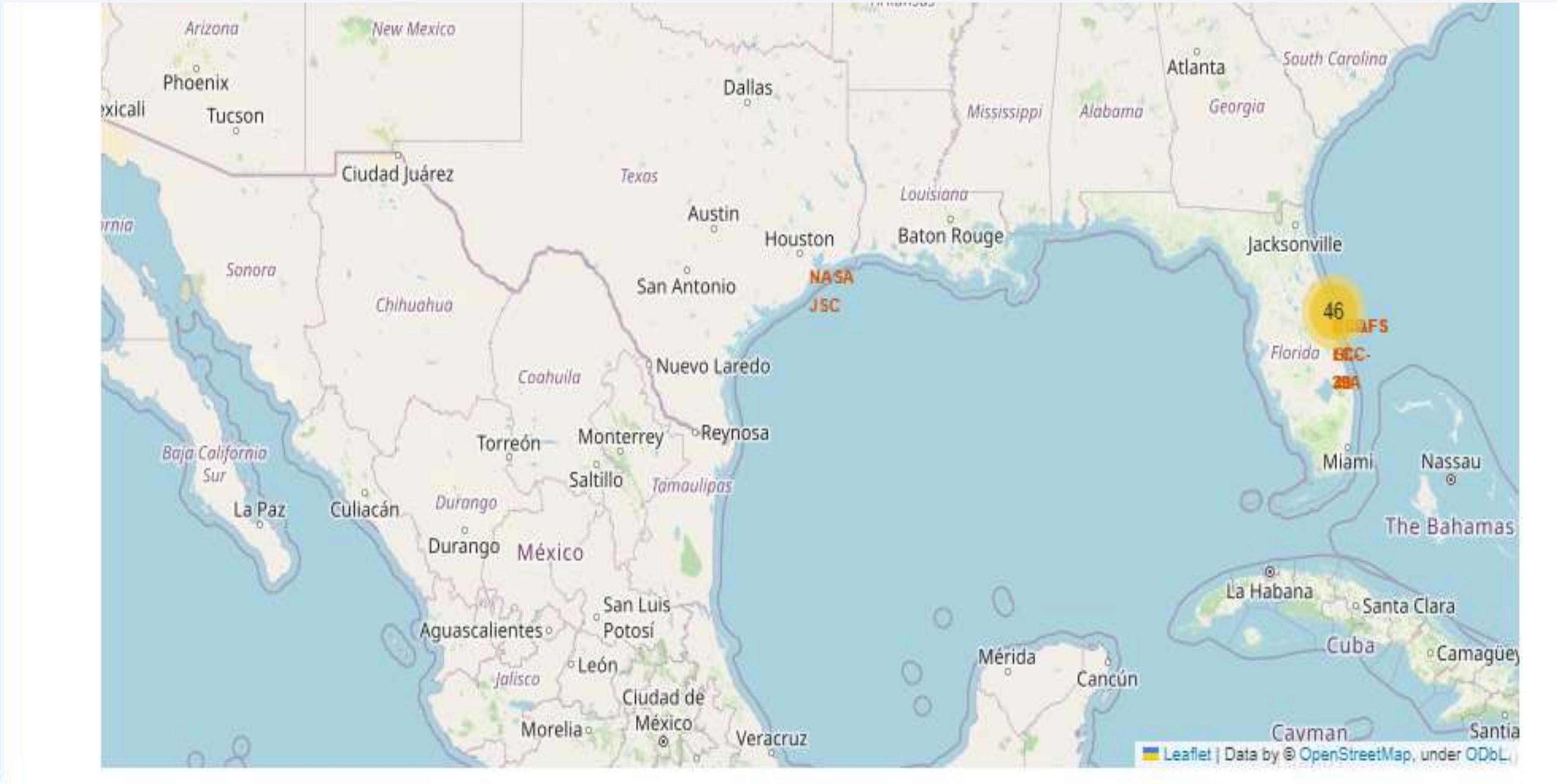


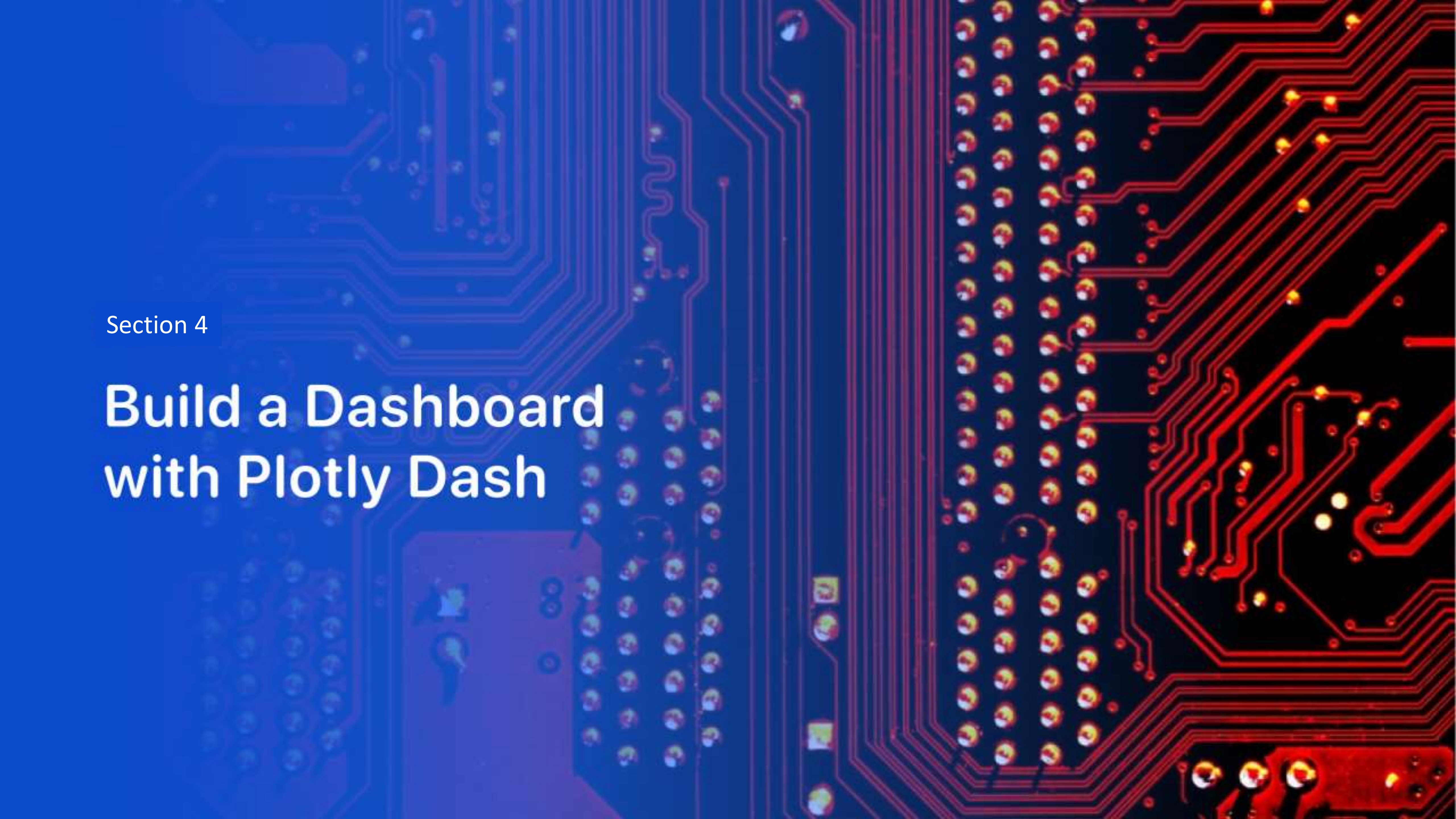
Mode: Command Ln 1, Col 1 lab_jupyter_launch_site_location.jupyterlite.ipynb

<Folium Map Screenshot 2>


+
−
Colorado
Kansas
Missouri
West Virginia
Washington
Virginia
Kentucky
Richmond
Tulsa
Raleigh
Tennessee
Albuquerque
Oklahoma
Memphis
Charlotte
North Carolina
Arkansas
Arizona
New Mexico
South Carolina
Atlanta
Phoenix
Dallas
Mississippi
Alabama
Georgia
Tucson
Ciudad Juárez
Texas
Louisiana
Austin
Houston
Baton Rouge
Jacksonville
Sonora
San Antonio
Chihuahua
CCAFS
Florida
KSC-
39A
Coahuila
Nuevo Laredo
Reynosa
Torreón
Monterrey
Baja California Sur
Saltillo
Tamaulipas
Miami
Nassau
La Paz
Culiacán
Durango
The Bahamas
Durango
México
La Habana
San Luis Potosí
Santa Clara
Aguascalientes
Cuba

<Folium Map Screenshot 3>

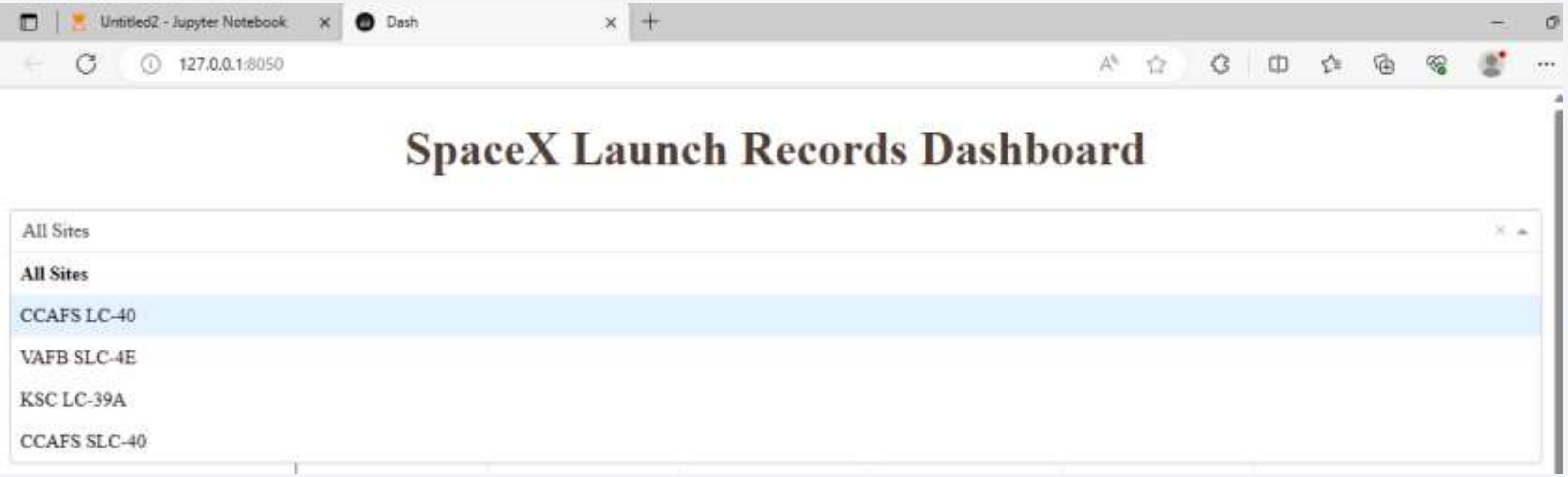

Arizona
New Mexico
Phoenix
Tucson
Dallas
Texas
Ciudad Juárez
Austin
Houston
San Antonio
NASA
JSC
Mississippi
Alabama
Louisiana
Baton Rouge
Atlanta
Georgia
South Carolina
Jacksonville
46
AFS
Florida
CC-
39A
Miami
Nassau
The Bahamas
Sonora
Chihuahua
Coahuila
Nuevo Laredo
Monterrey
Reynosa
Torreón
Saltillo
Tamaulipas
Baja California Sur
La Paz
Culiacán
Durango
Durango
México
San Luis Potosí
Aguascalientes
León
Jalisco
Ciudad de México
Morelia
Veracruz
Mérida
Cancún
La Habana
Santa Clara
Cuba
Camagüey
Cayman
Santia
Leaflet | Data by © OpenStreetMap, under ODbL.



Section 4

Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>



SpaceX Launch Records Dashboard

All Sites

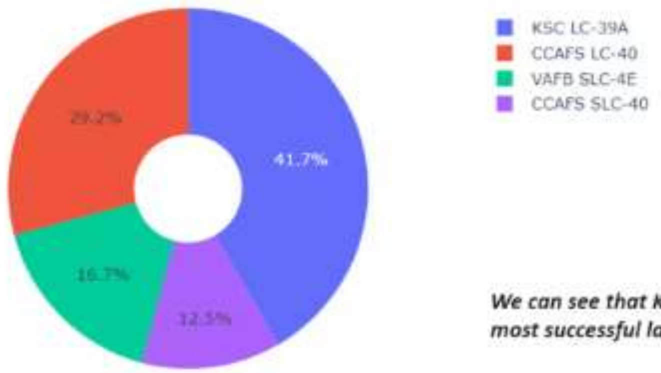
All Sites

CCAFS LC-40

VAFB SLC-4E

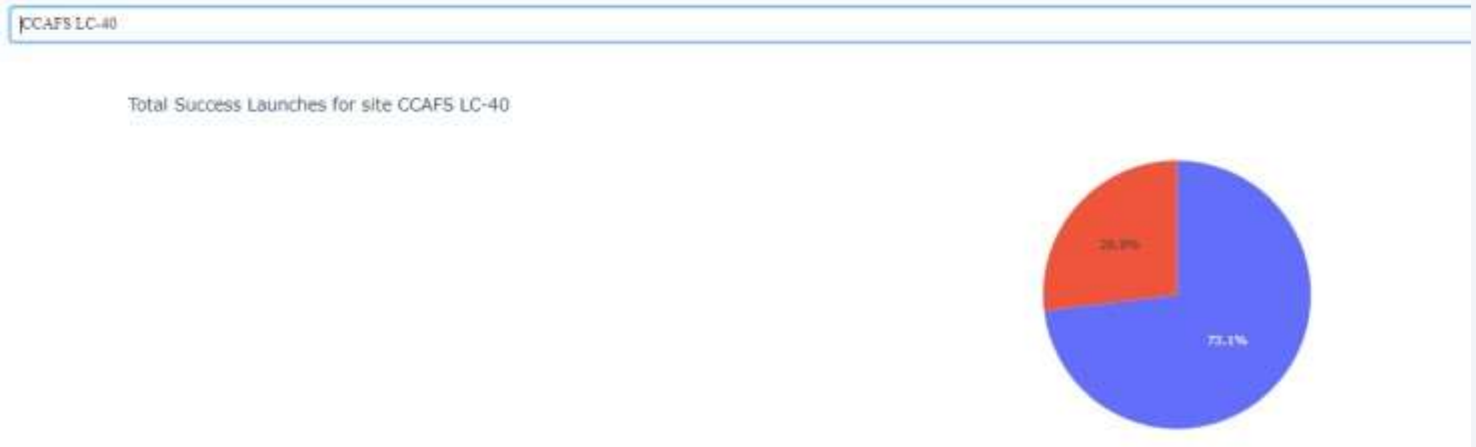
KSC LC-39A

CCAFS SLC-40



We can see that KSC LC-39A had the most successful launches from all the sites

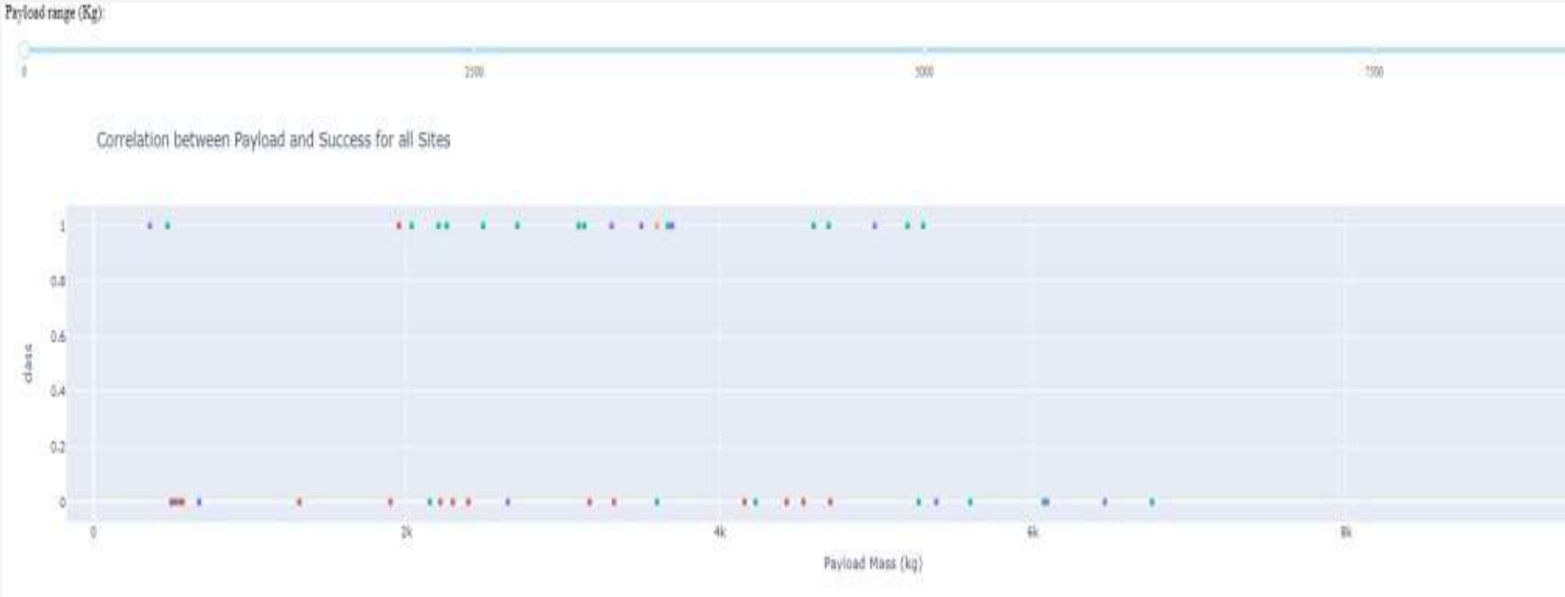
<Dashboard Screenshot 2>



CCAFS LC-40

Total Success Launches for site CCAFS LC-40

<Dashboard Screenshot 3>

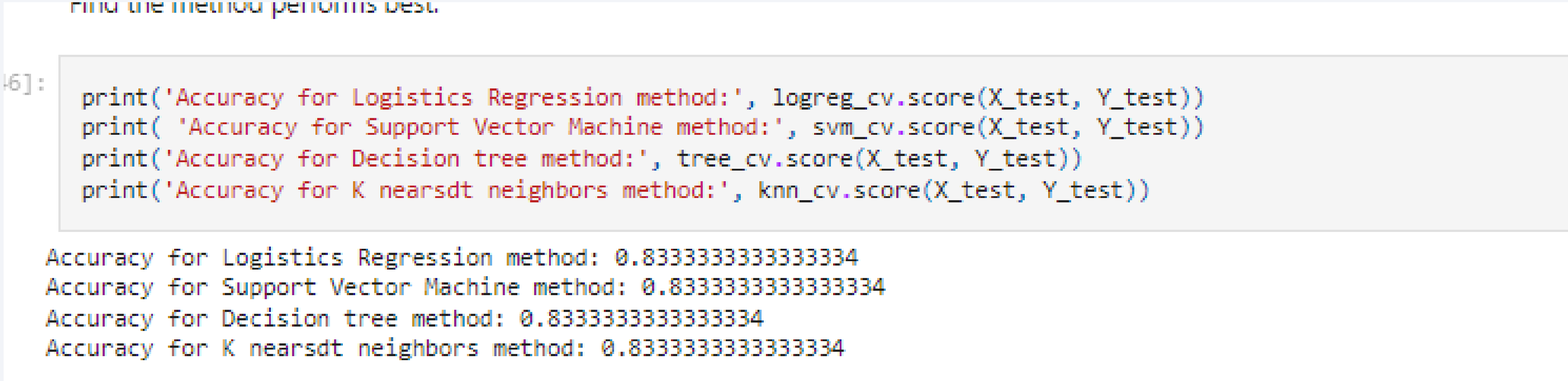


Payload range (Kg):
0
2500
5000
7500
Correlation between Payload and Success for all Sites
class
1
0.8
0.6
0.4
0.2
0
0
2k
4k
6k
8k
Payload Mass (kg)

Section 5

Predictive Analysis (Classification)

Classification Accuracy



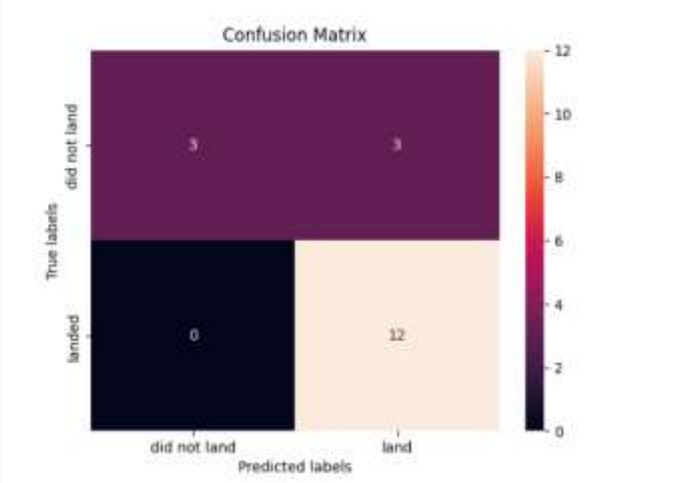
Find the method performs best.

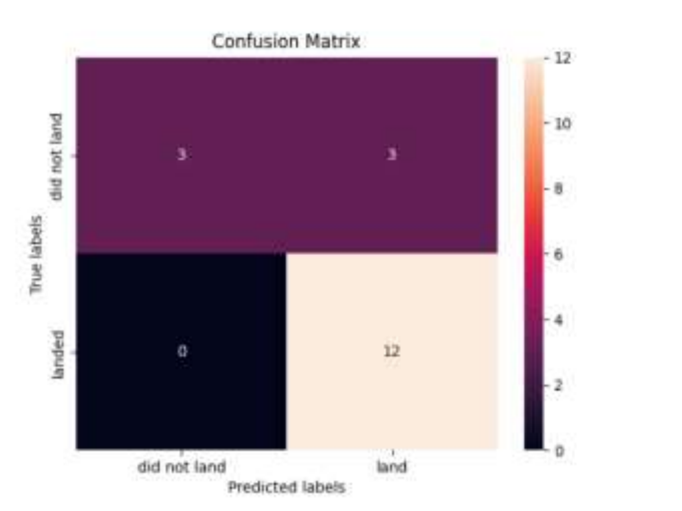
46]:

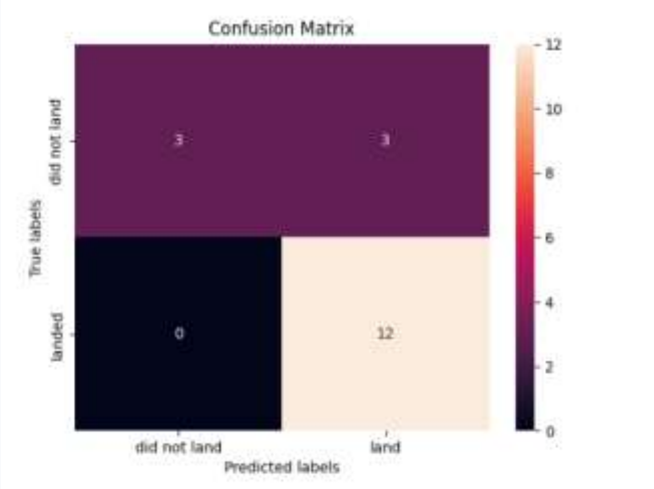
```
print('Accuracy for Logistics Regression method:', logreg_cv.score(X_test, Y_test))
print( 'Accuracy for Support Vector Machine method:', svm_cv.score(X_test, Y_test))
print('Accuracy for Decision tree method:', tree_cv.score(X_test, Y_test))
print('Accuracy for K nearsdt neighbors method:', knn_cv.score(X_test, Y_test))
```

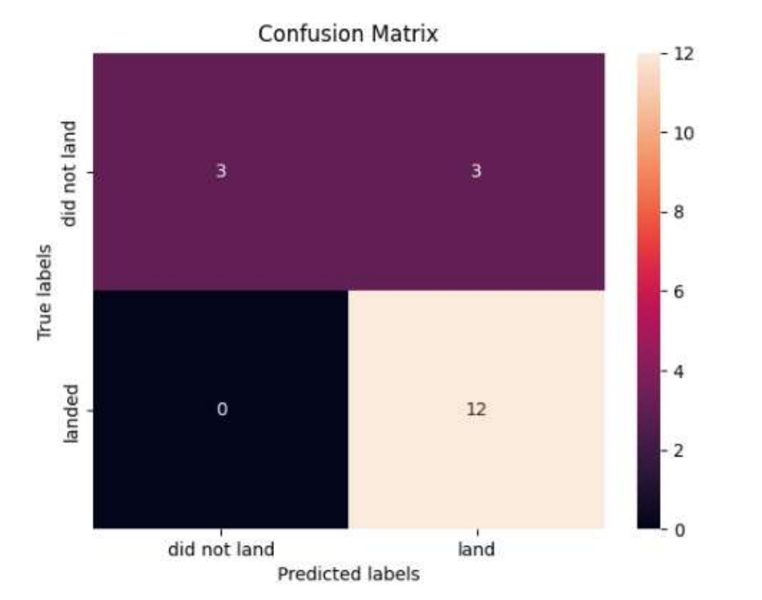
```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.8333333333333334
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```

Confusion Matrix









Conclusions

- For this data set the SVM, KNN, and Logistic Regression models had the highest prediction accuracy.
- Low weight payloads perform better than heavier payloads, and SpaceX launch success rates are strongly correlated with the number of years it will take them to perfect their missions.
- From all the sites, KSC LC 39A had the most prosperous launches.
- The best success rate is in Orbit GEO HEO, SSO, and ES L1.

Thank you!