IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Viraj Koparkar
25-03-2024

Outline

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

Executive Summary

In this project we will try to predict if SpaceX Falcone 9 first stage will land successfully using several machine learning algorithms

- Following concepts and methods were used to collect and analyze data, build and evaluate machine learning models, and make predictions:
    - Collect data through API and Web scraping
    - Transform data through data wrangling
    - Conduct exploratory data analysis with SQL and data visuals
    - Build an interactive map with folium to analyze launch site proximity
    - Build a dashboard to analyze launch records interactively with Plotly Dash
    - Finally, build a predictive model to predict if the first stage of Falcon 9 will land successfully

- Summary of results:
    - As the numbers of flights increase, the first stage is more likely to land successfully
    - Launch Site ‘KSC LC-39A’ has the highest launch success rate
    - Orbits ES-L1, GEO, HEO, and SSO have the highest launch success rates

Introduction

- The commercial space age is here, companies are making space travel affordable for everyone. Perhaps the most successful is SpaceX. One reason SpaceX can do this is the rocket launches are relatively inexpensive. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Which can be used by competitors to compete against SpaceX
- Problems you want to find answers
  - Determine if the first stage of SpaceX Falcon 9 will land successfully
  - Impact of different parameters/variables on the landing outcomes (e.g., launch site, payload mass, booster version, etc.)
  - Correlations between launch sites and success rates

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web scrap Falcon 9 and Falcon Heavy launch records from Wikipedia (link)
- Perform data wrangling
  - Determined labels for training the supervised models by converting mission outcomes in to training labels (0-unsuccessful, 1-successful)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
- LR, KNN, SVM, Decision Tree, models were built using hyperparametrs to optimize accuracy

Data Collection

- The data was collected via SpaceX API and Web scrapping Wiki pages for relevant launch data.

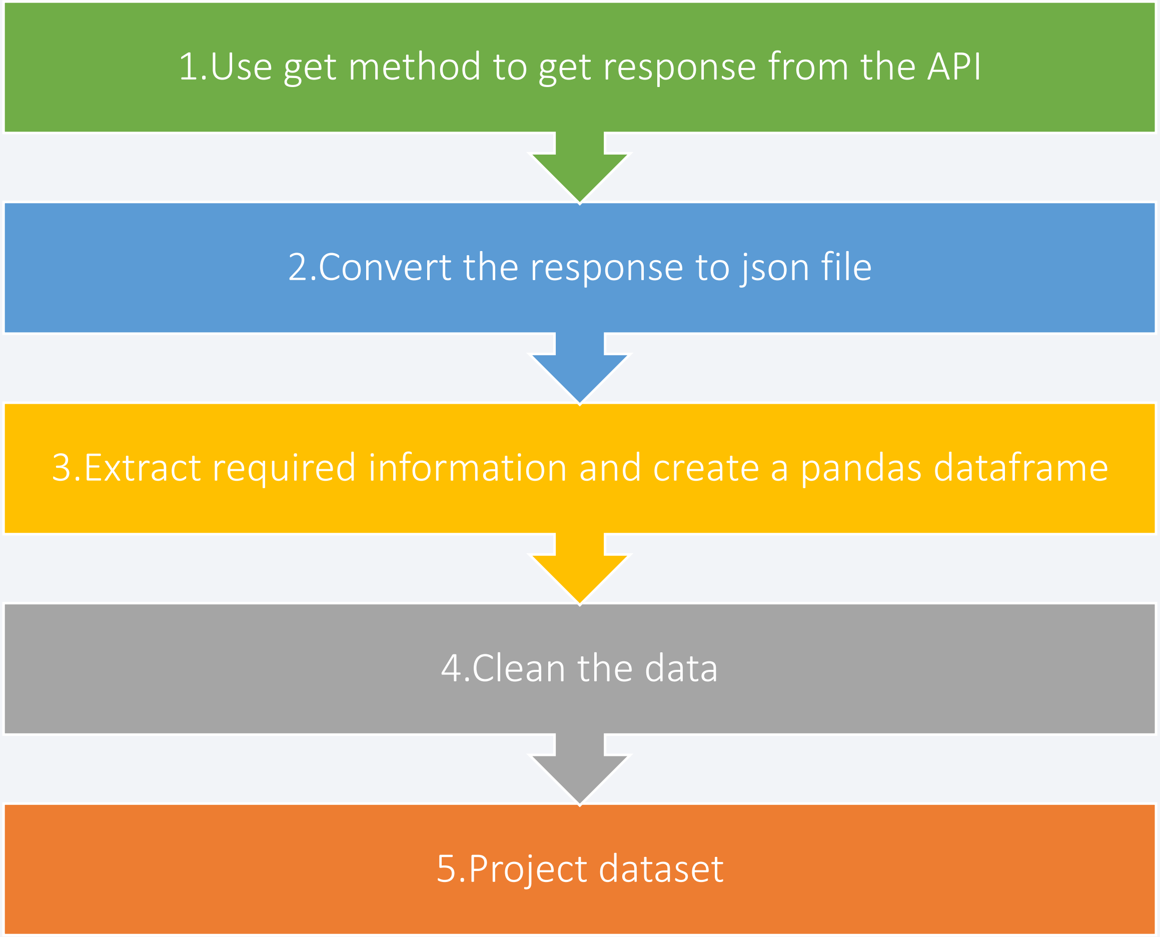
- The data includes:

FlightNumber
Date
BoosterVersion
PayloadMass
Orbit
LaunchSite
Outcome
Flights
GridFins
Reused
Legs
LandingPad
Block
ReusedCount
Serial
Longitude
Latitude
Class

Data Collection – SpaceX API

- Data collection with SpaceX REST calls:
  - The API used is https://api.spacexdata.com/v4/rockets
  - Missing values in the data are replaced by respective means.
  - The data is filtered to include only Falcone 9 launches

- Github Notebook



1.Use get method to get response from the API

2.Convert the response to json file

3.Extract required information and create a pandas dataframe

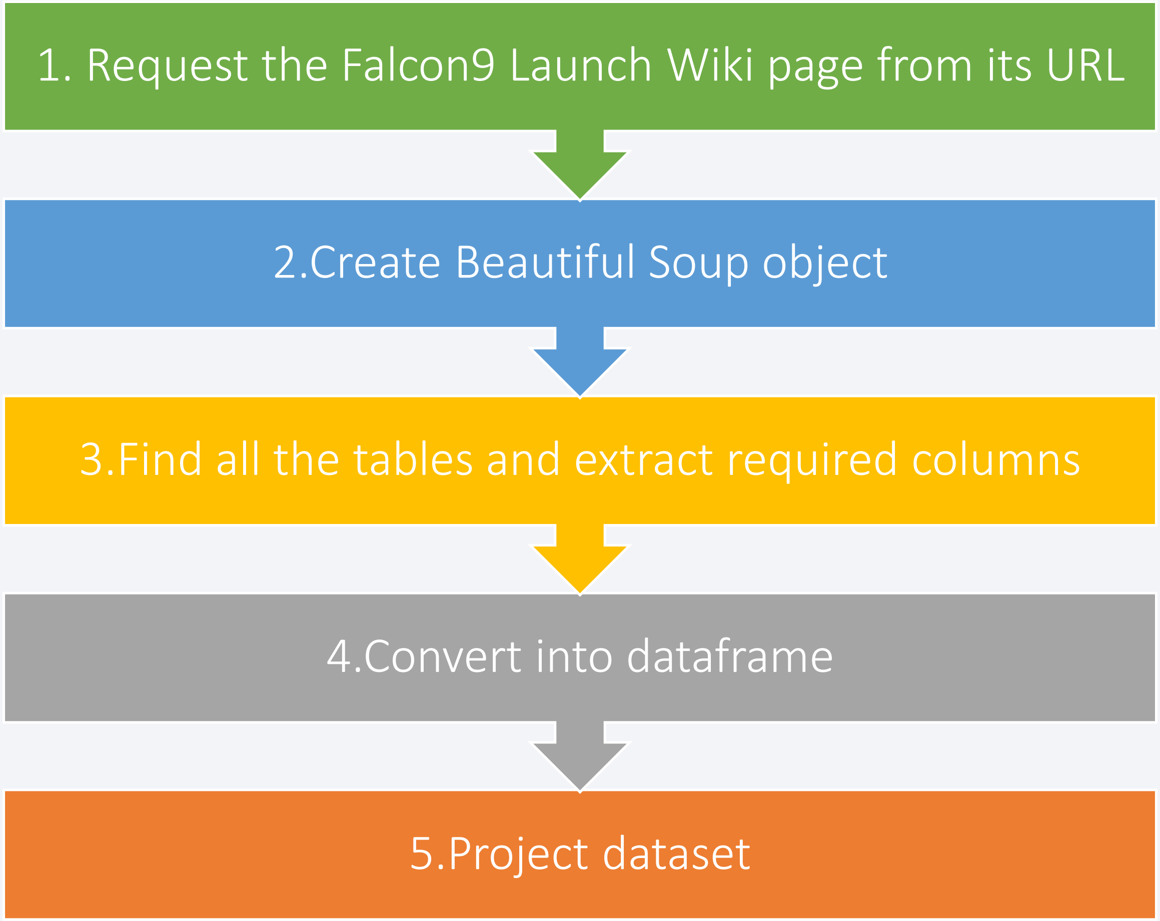
4.Clean the data

5.Project dataset

Data Collection - Scraping

- web scraping:
  - The data was scraped from https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
  - The data was scraped using beautiful Soup

- Github Notebook


1. Request the Falcon9 Launch Wiki page from its URL

2.Create Beautiful Soup object

3.Find all the tables and extract required columns

4.Convert into dataframe

5.Project dataset

Data Wrangling

- The data set contained missing values which were replaced with respective column means.
- The data set contained various mission outcomes that were converted into Training Labels with 1 meaning the booster successfully landed and 0 meaning booster was unsuccessful in landing.
- These labels will be used to perform EDA and model fitting in later stages

- Github Notebook

EDA with Data Visualization

- As part of the Exploratory Data Analysis (EDA), following charts were plotted to gain further insights into the dataset:

1. Scatter plot:
    - Shows relationship or correlation between two variables making patterns easy to observe
        - Relationship between Flight Number and Launch Site
        - Relationship between Payload and Launch Site
        - Relationship between Flight Number and Orbit Type
        - Relationship between Payload and Orbit Type
2. Bar Chart:
    - Bar charts makes it easy to see which groups are highest/common and how other groups compare against each other.
    - Plotted following Bar chart to visualize:
        - Relationship between success rate of each orbit type
3. Line Chart:
    - It helps depict trends over time.
    - Plotted following Line chart to observe:
        - Average launch success yearly trend

Github Notebook

EDA with SQL

- To better understand SpaceX data set, following SQL queries/operations were performed on an IBM DB2 cloud instance:
  1. Display the names of the unique launch sites in the space mission
  2. Display 5 records where launch sites begin with the string 'CCA’
  3. Display the total payload mass carried by boosters launched by NASA (CRS)
  4. Display average payload mass carried by booster version F9 v1.1
  5. List the date when the first successful landing outcome in ground pad was achieved.
  6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  7. List the total number of successful and failure mission outcomes
  8. List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
  9. List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  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

Build an Interactive Map with Folium

- Folium interactive map helps analyze geospatial data to perform more interactive visual analytics and better understand factors such location and proximity of launch sites that impact launch success rate.
- Following map object were created and added to the map:
  - Mark all launch sites on the map. This allowed to visually see the launch sites on the map.
  - Added ‘folium.circle’ and ‘folium.marker’ to highlight circle area with a text label over each launch site.
  - Added a ‘MarkerCluster()’ to show launch success (green) and failure (red) markers for each launch site.
- Calculated distances between a launch site to its proximities (e.g., coastline, railroad, highway, city)
  - Added ‘MousePosition() to get coordinate for a mouse position over a point on the map
  - Added ‘folium.Marker()’ to display distance (in KM) on the point on the map (e.g., coastline, railroad, highway, city)
  - Added ‘folium.Polyline()’ to draw a line between the point on the map and the launch site
  - Repeated steps above to add markers and draw lines between launch sites and proximities – coastline, railroad, highway, city)

Build an Interactive Map with Folium

- Building the Interactive Map with Folium helped answered following questions:
    1. Are launch sites in close proximity to railways?

       YES
    2. Are launch sites in close proximity to highways?

       YES
    3. Are launch sites in close proximity to coastline?

       YES
    4. Do launch sites keep certain distance away from cities?

       YES

- Github Notebook

Build a Dashboard with Plotly Dash

Built a Plotly Dash web application to perform interactive visual analytics on SpaceX launch data in real-time.

Added Launch Site Drop-down, Pie Chart, Payload range slide, and a Scatter chart to the Dashboard.

1. Added a Launch Site Drop-down Input component to the dashboard to provide an ability to filter Dashboard visual by all launch sites or a particular launch site
2. Added a Pie Chart to the Dashboard to show total success launches when ‘All Sites’ is selected and show success and failed counts when a particular site is selected
3. Added a Payload range slider to the Dashboard to easily select different payload ranges to identify visual patterns
4. Added a Scatter chart to observe how payload may be correlated with mission outcomes for selected site(s). The color-label Booster version on each scatter point provided missions outcomes with different boosters

Build a Dashboard with Plotly Dash

- Dashboard helped answer following questions:

1. Which site has the largest successful launches? KSC LC-39A with 10 2.
2. Which site has the highest launch success rate? KSC LC-39A with 76.9% success
3. Which payload range(s) has the highest launch success rate? 2000 – 5000 kg
4. Which payload range(s) has the lowest launch success rate? 0-2000 and 5500 - 7000
5. Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate? FT

Githib Notebook

Predictive Analysis (Classification)

- The Scikit-learn library is used to create machine learning models
- Building machine learning models includes the following steps:

  1. Standardize regression variables.
  2. Split the data into train, test dataset.
  3. Create machine learning models and fit best hyperpameters.
  4. Evaluate the models based on their accuracy score and confusion matrix.

- Github Notebook

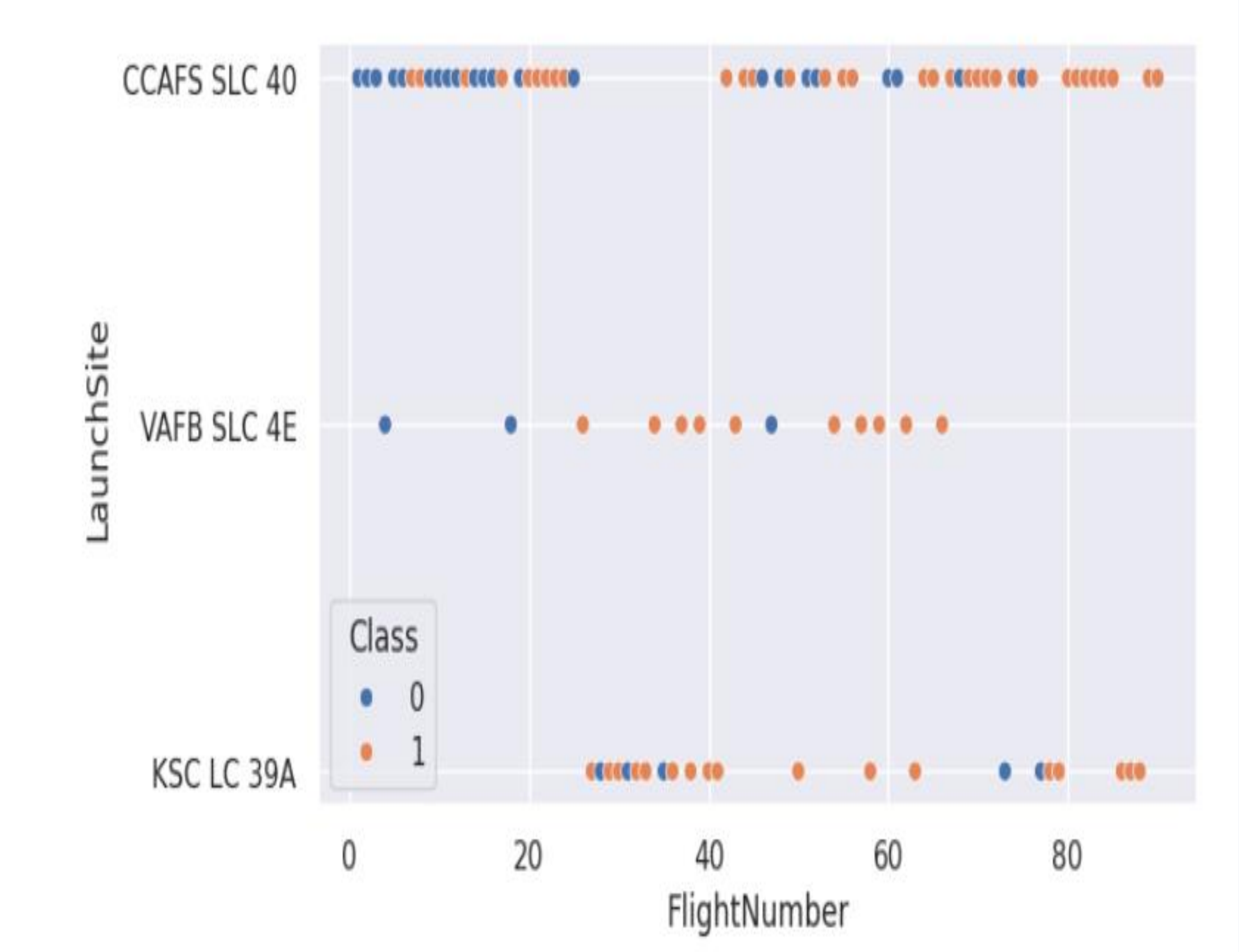
Results

- Exploratory data analysis results:
  - KSC LC-39A has the highest success rate.
  - Success rate seems to be increasing with flight number.
  - Orbits GEO, HEO, SSO, ES L1 have the highest success rate.
  - The success rate is increasing steadily over the years.
- Predictive analysis results
  - All the predictive models have almost same test set accuracy (about 0.83) except KNN which has better training accuracy but slightly worse test set accuracy.

Section 2

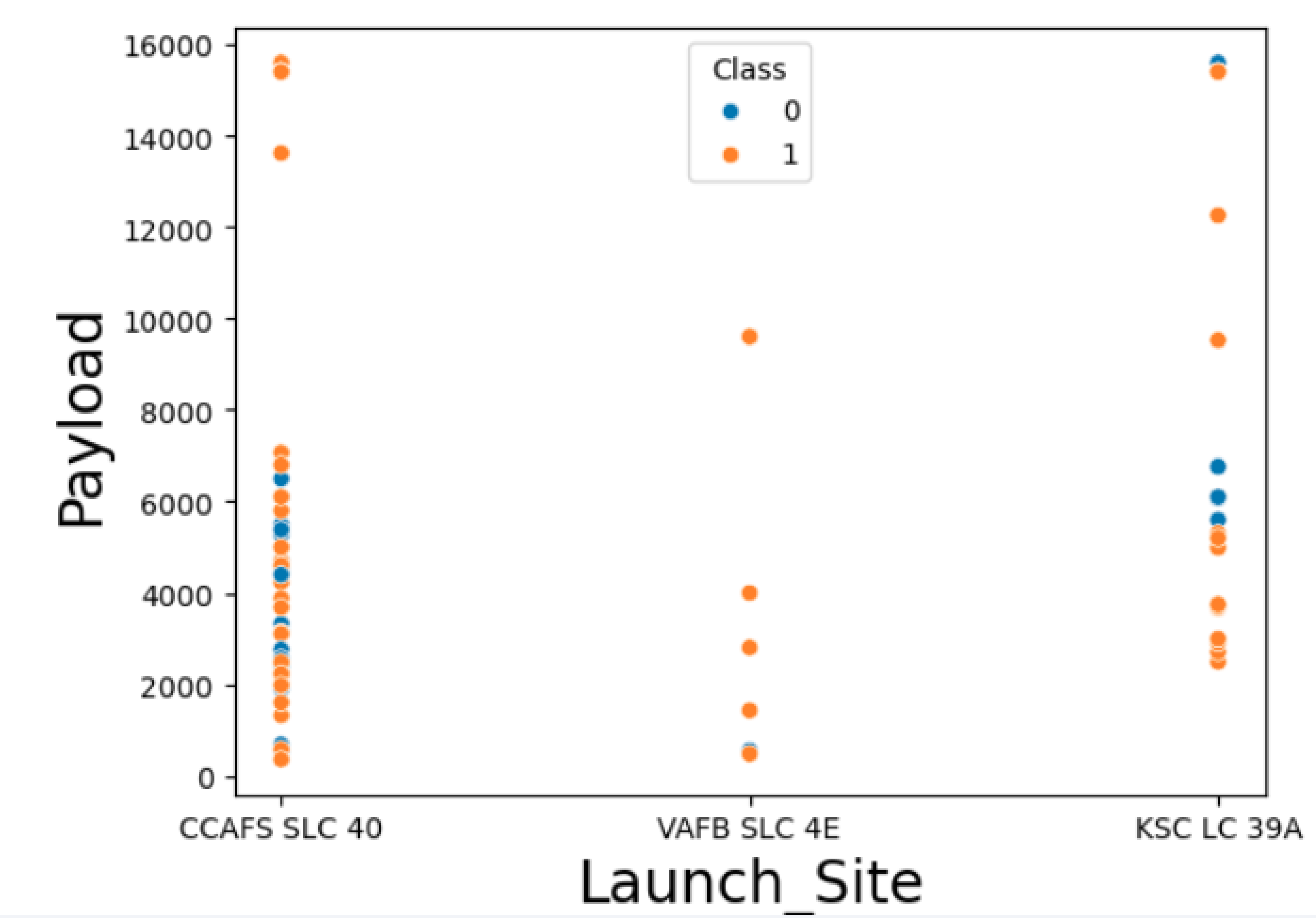
Insights drawn from EDA

Flight Number vs. Launch Site



- Success rates (Class=1) increases as the number of flights increase

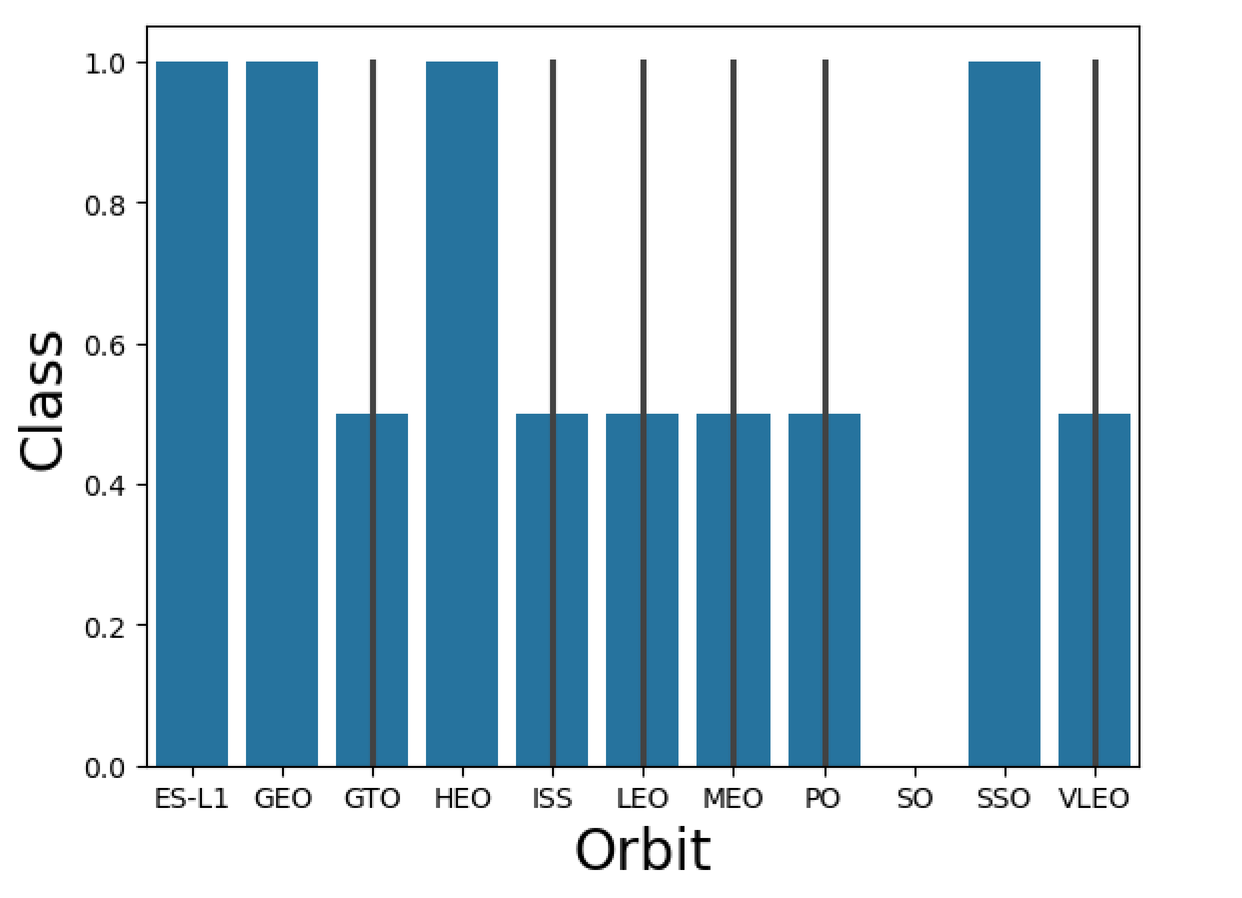
Payload vs. Launch Site



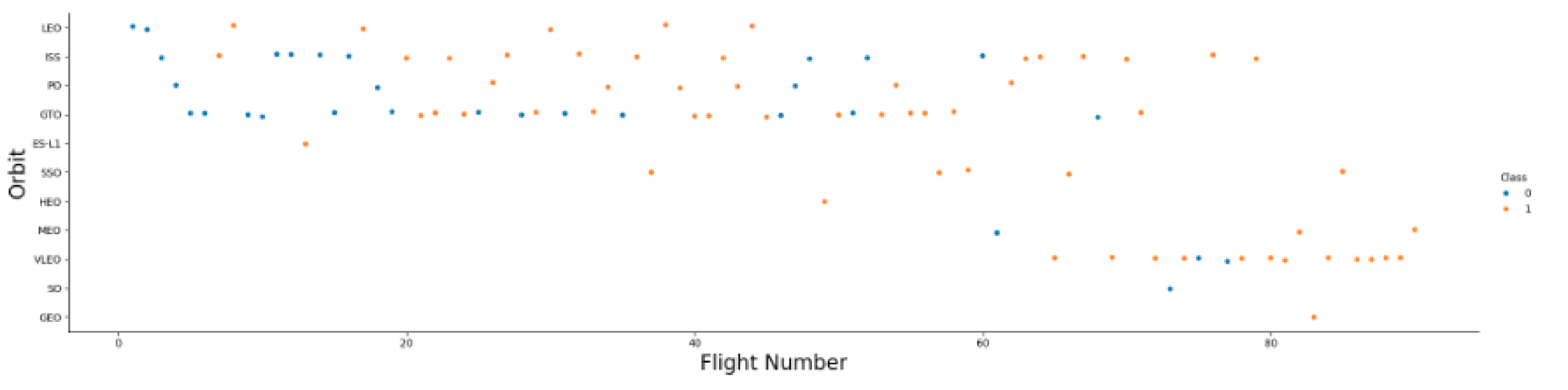
- For launch site ‘VAFB SLC 4E’, there are no rockets launched for payload greater than 10,000 kg
- Percentage of successful launch increases for launch site ‘VAFB SLC 4E’ as the payload mass increases
- There is no clear correlation or pattern between launch site and payload

Success Rate vs. Orbit Type

- Orbits ES-LI, GEO, HEO, and SSO have the highest success rates
- SO orbit has the lowest success rate



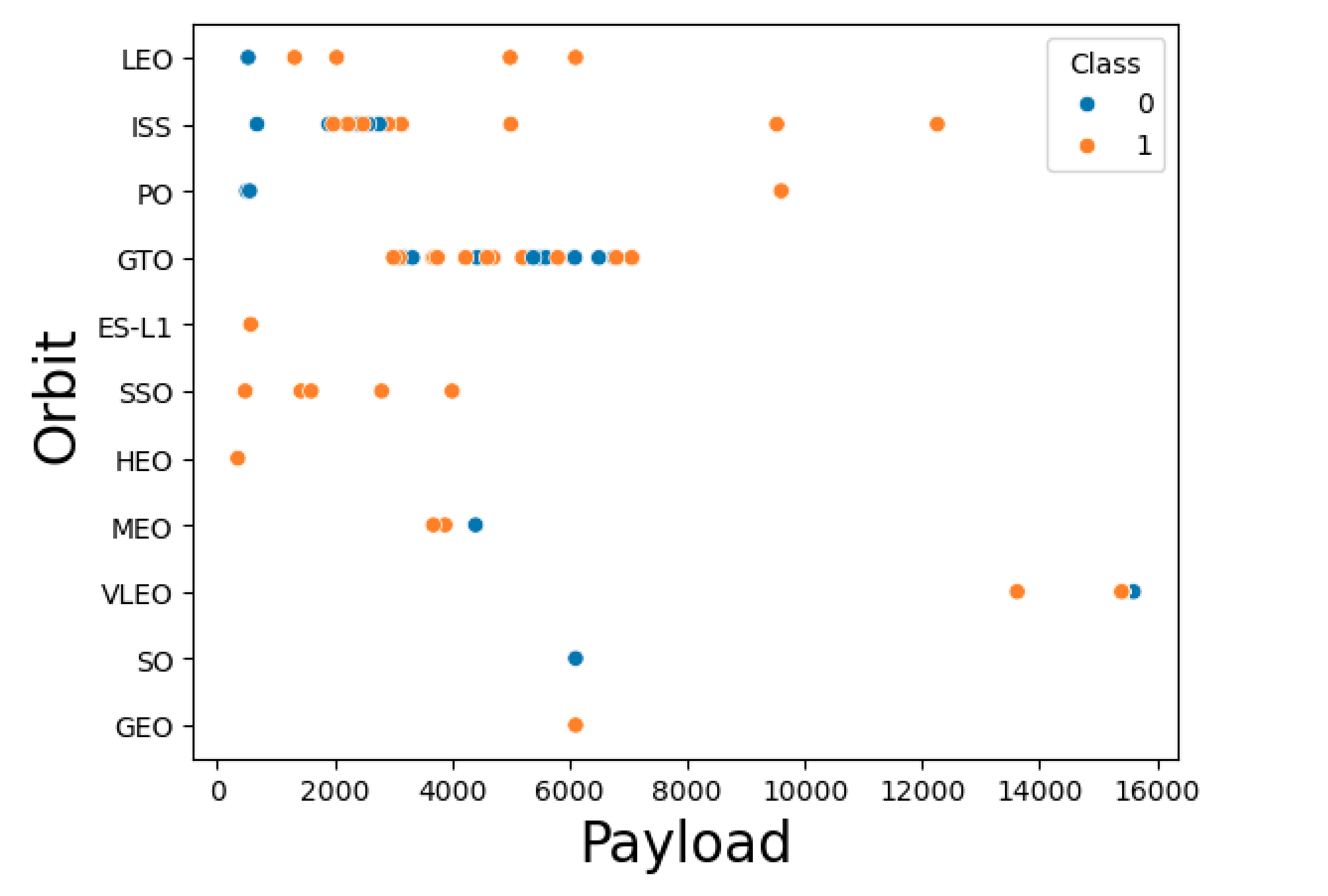
Flight Number vs. Orbit Type



- For most orbits (LEO, ISS, PO, SSO, MEO, VLEO) successful landing rates appear to increase with flight numbers.
- There is no relationship between flight number and orbit for GTO

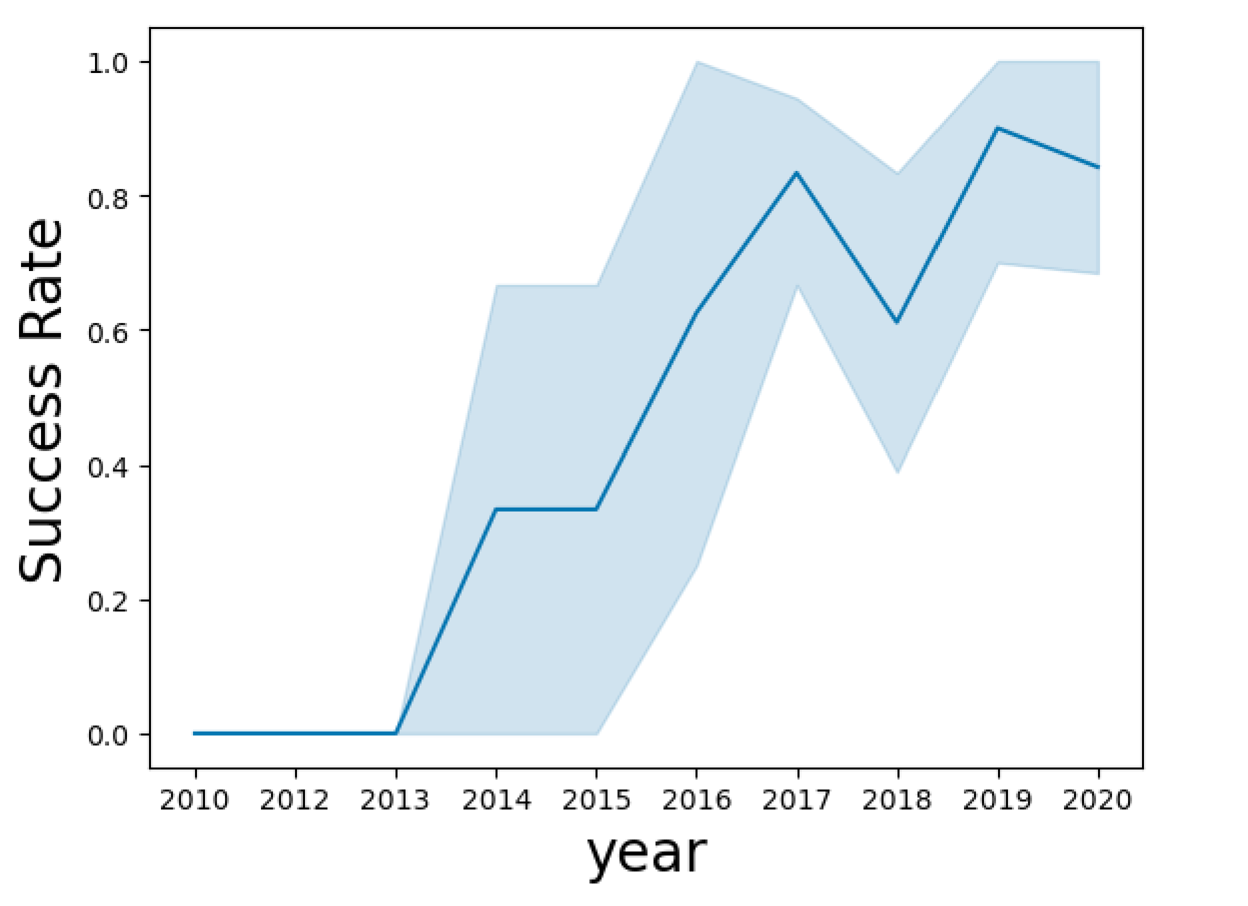
Payload vs. Orbit Type

- Successful landing rates appear to increase with pay load for orbits LEO, ISS, PO, and SSO
- For GEO orbit, there is not clear pattern between payload and orbit for successful or unsuccessful landing



Launch Success Yearly Trend

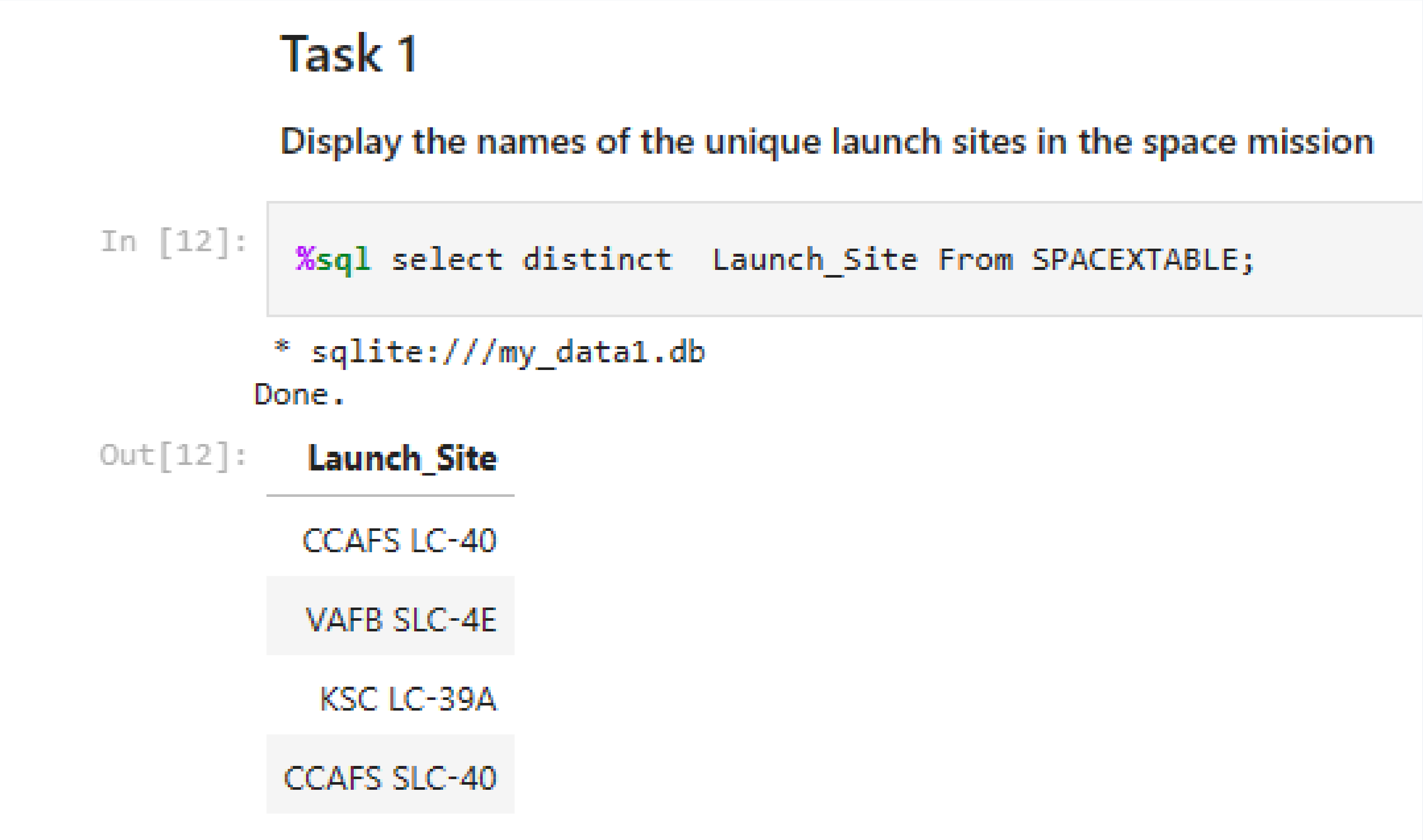
- Success rate increased by about 80% between 2013 and 2020
- Success rates remained the same between 2010 and 2013 and between 2014 and 2015
- Success rates takes a huge dip in 2018 which may be due to the fact that it was the year with most launches. Further investigation is required.



All Launch Site Names

Query:

Results:



Task 1

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

In [12]:

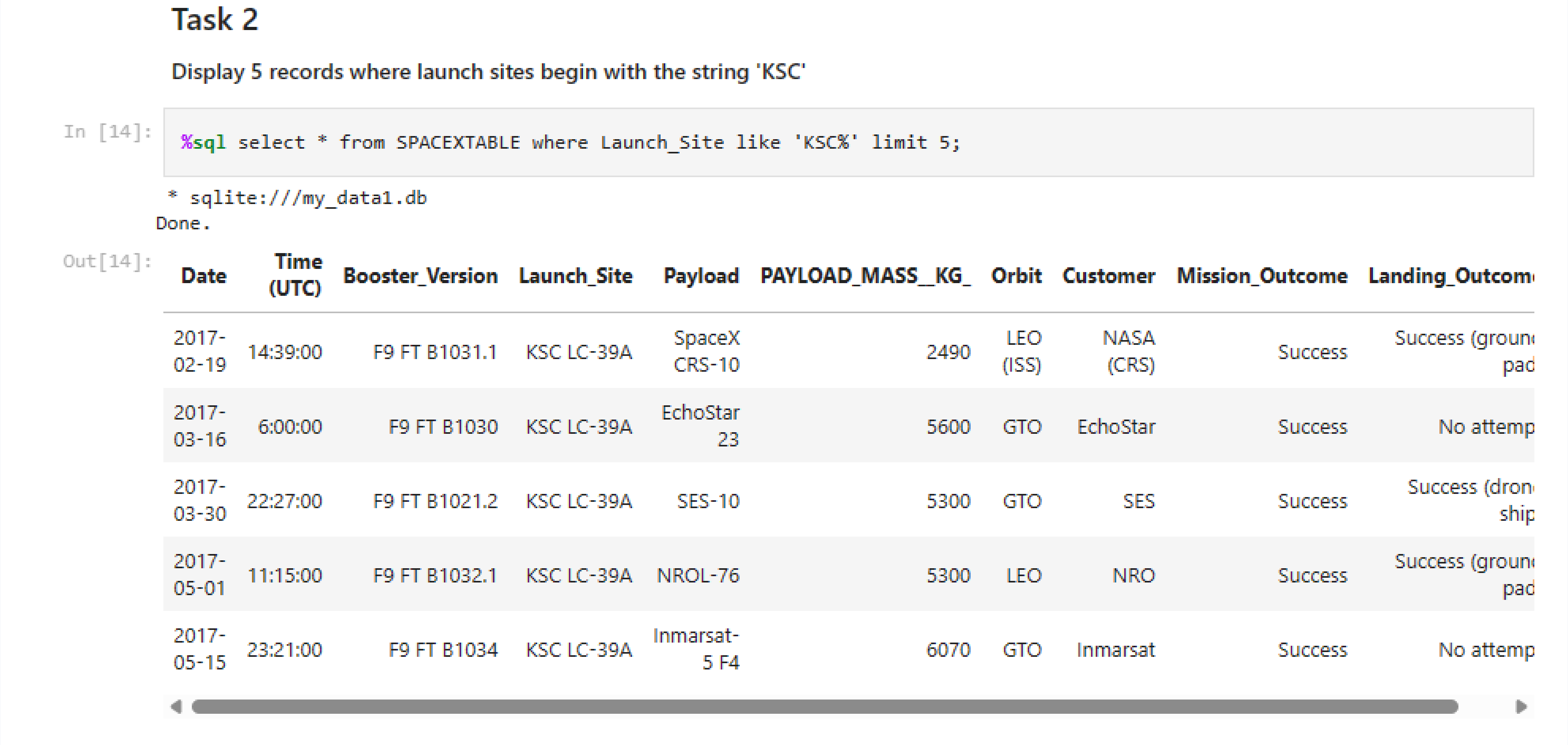
```
%sql select distinct  Launch_Site From SPACEXTABLE;
```

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

Out[12]:

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

Launch Site Names Begin with 'KSC'



Task 2

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

In [14]:

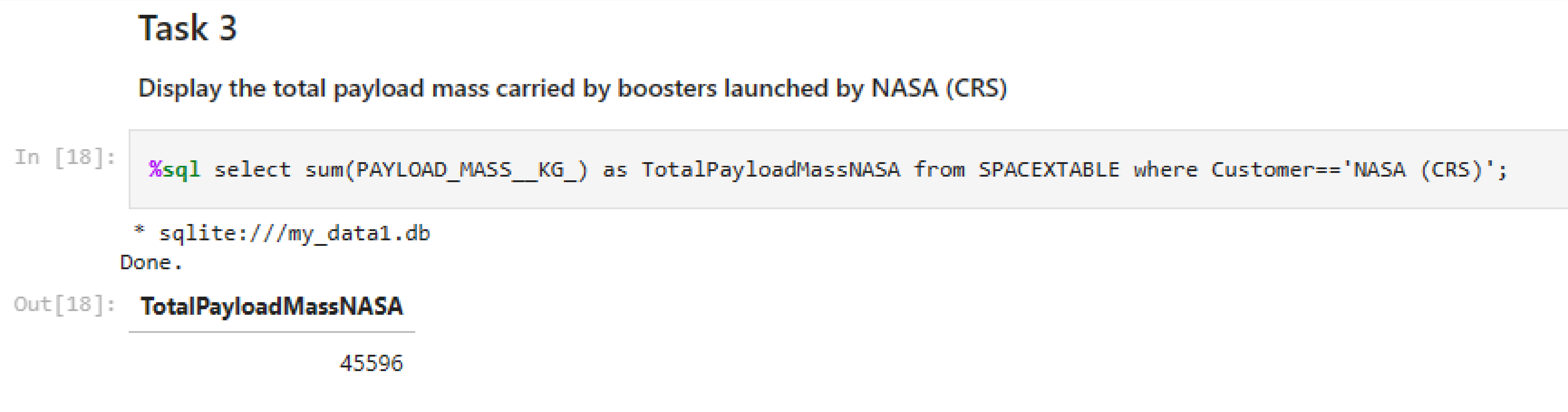
```
%sql select * from SPACEXTABLE where Launch_Site like 'KSC%' limit 5;
```

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

Out[14]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcom
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (groun pad
2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attemp
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (dron ship
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (groun pad
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attemp

Total Payload Mass



Task 3

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

In [18]:

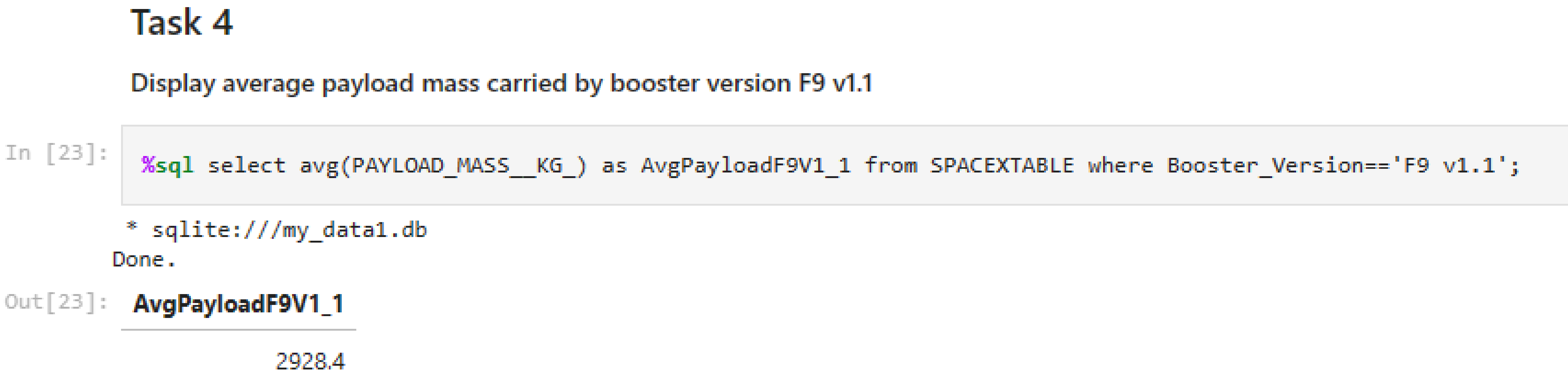
```
%sql select sum(PAYLOAD_MASS__KG_) as TotalPayloadMassNASA from SPACEXTABLE where Customer=='NASA (CRS)';
```

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

Out[18]:

TotalPayloadMassNASA
45596

Average Payload Mass by F9 v1.1



Task 4

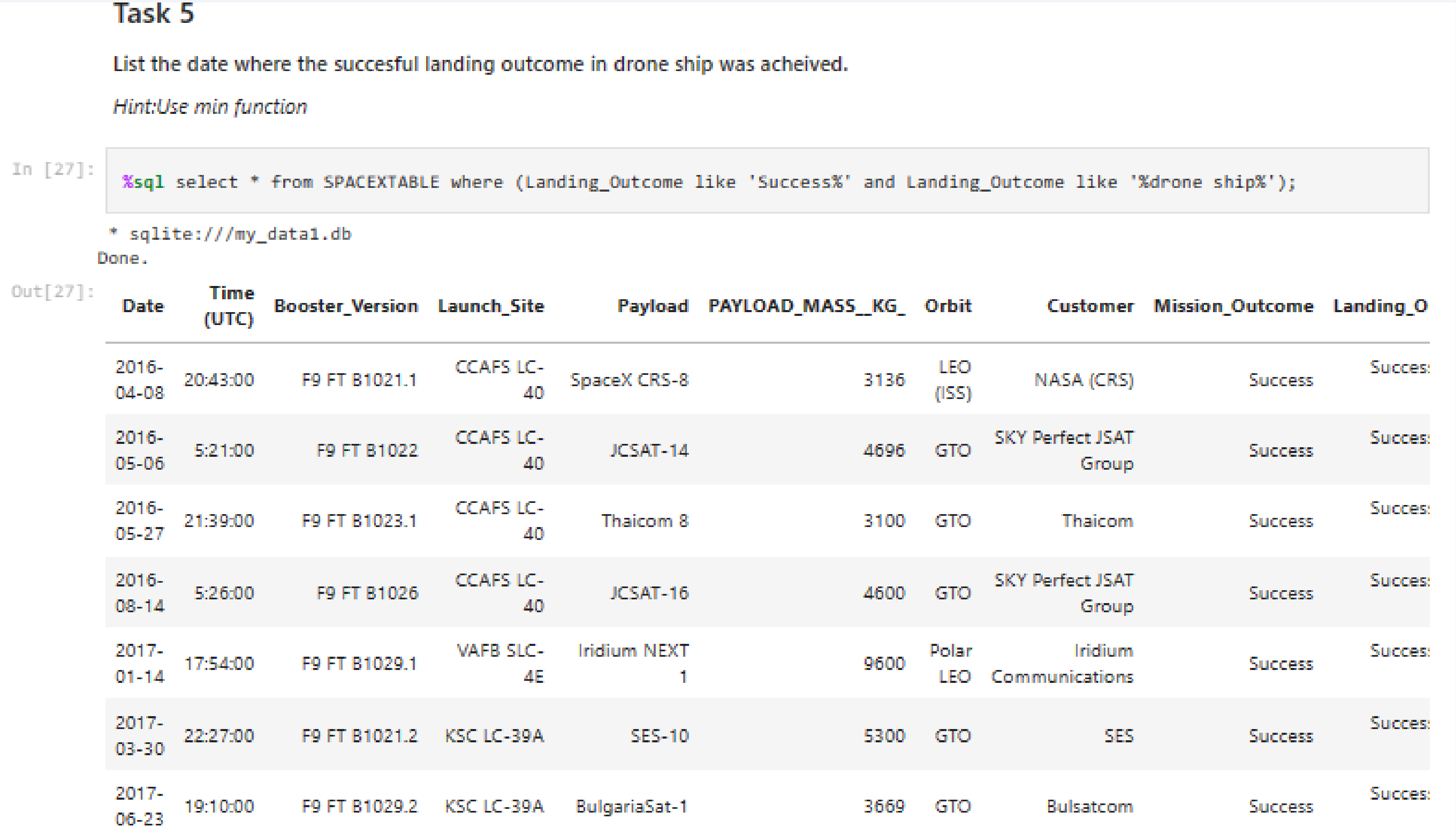
Display average payload mass carried by booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) as AvgPayloadF9V1_1 from SPACEXTABLE where Booster_Version=='F9 v1.1';
```

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

AvgPayloadF9V1_1
2928.4

First Successful Ground Landing Date



Task 5

List the date where the succesful landing outcome in drone ship was acheived.

Hint:Use min function

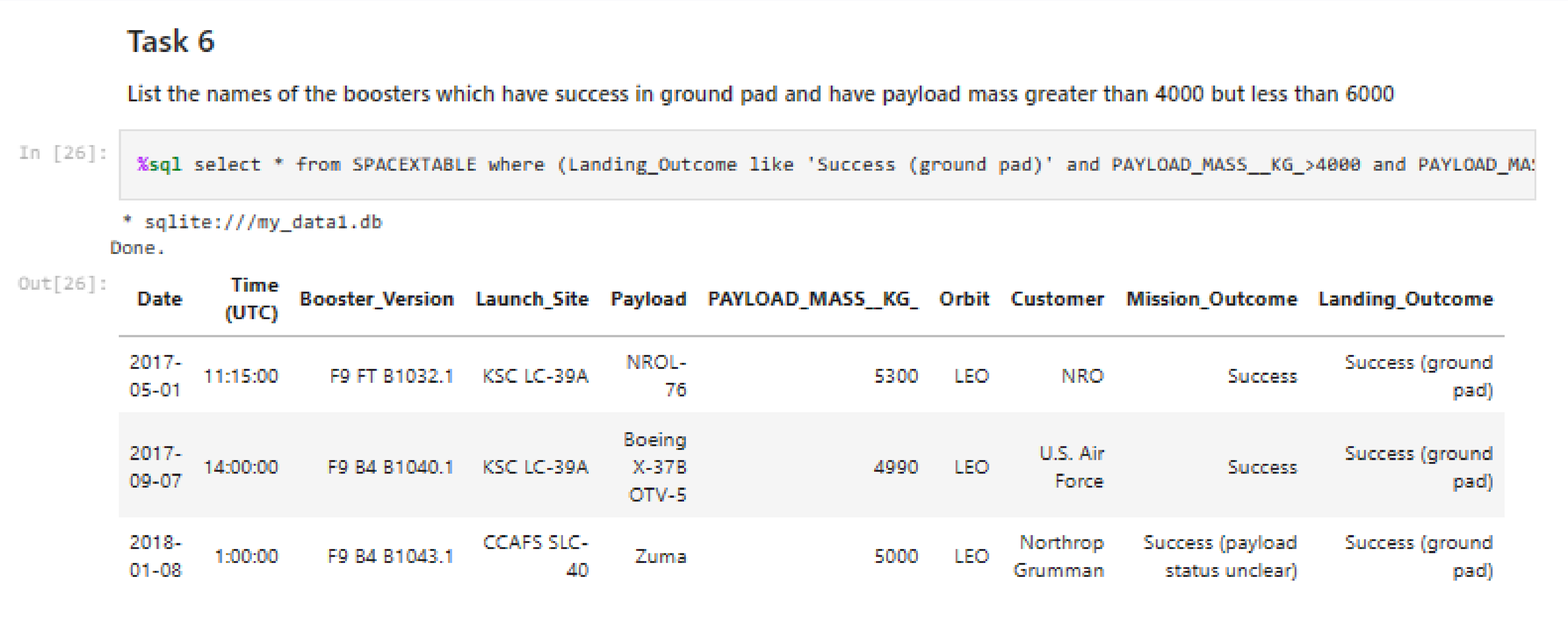
In [27]:
```
%sql select * from SPACEXTABLE where (Landing_Outcome like 'Success%' and Landing_Outcome like '%drone ship%');
```

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

Out[27]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_O
2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Succes
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Succes
2016-05-27	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Succes
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Succes
2017-01-14	17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	Succes
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Succes
2017-06-23	19:10:00	F9 FT B1029.2	KSC LC-39A	BulgariaSat-1	3669	GTO	Bulsatcom	Success	Succes

Successful Drone Ship Landing with Payload between 4000 and 6000



Task 6

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

In [26]:

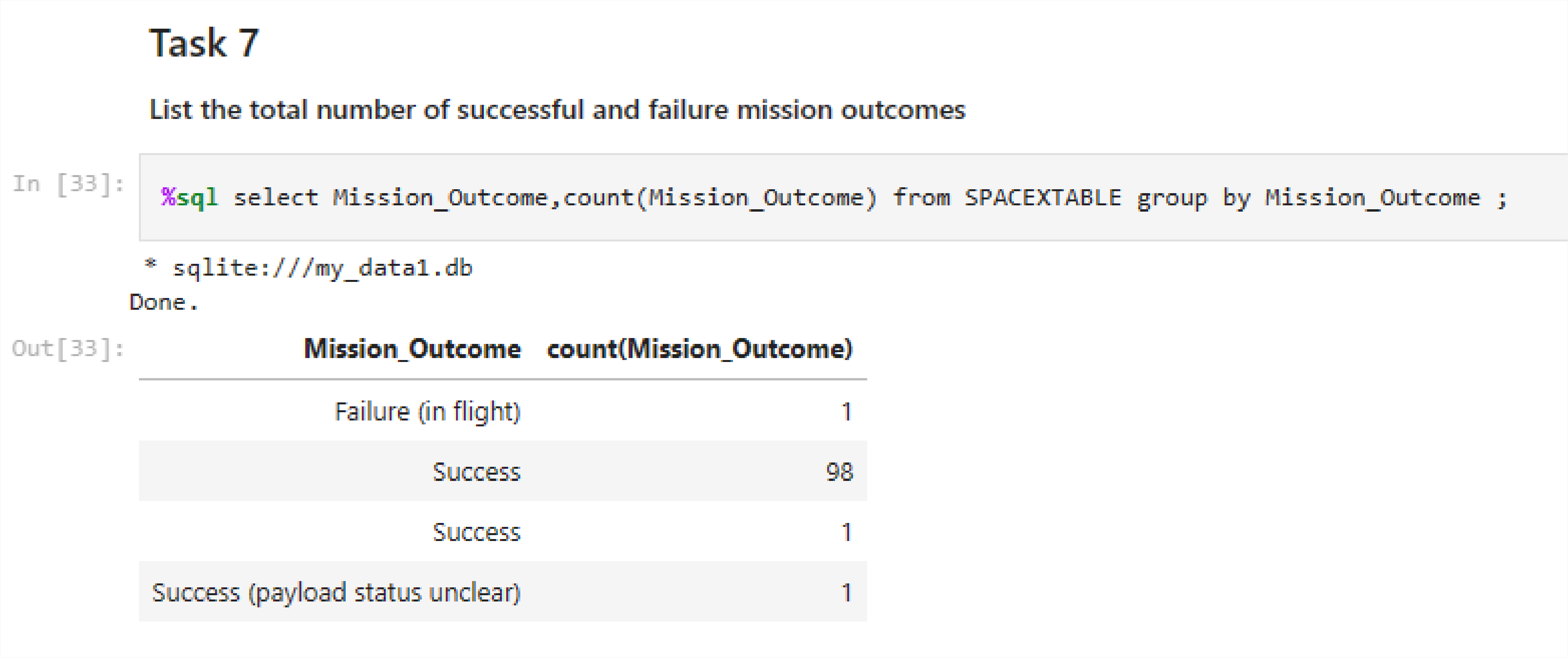
```
%sql select * from SPACEXTABLE where (Landing_Outcome like 'Success (ground pad)' and PAYLOAD_MASS__KG_>4000 and PAYLOAD_MA
```

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

Out[26]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-09-07	14:00:00	F9 B4 B1040.1	KSC LC-39A	Boeing X-37B OTV-5	4990	LEO	U.S. Air Force	Success	Success (ground pad)
2018-01-08	1:00:00	F9 B4 B1043.1	CCAFS SLC-40	Zuma	5000	LEO	Northrop Grumman	Success (payload status unclear)	Success (ground pad)

Total Number of Successful and Failure Mission Outcomes



Task 7

List the total number of successful and failure mission outcomes

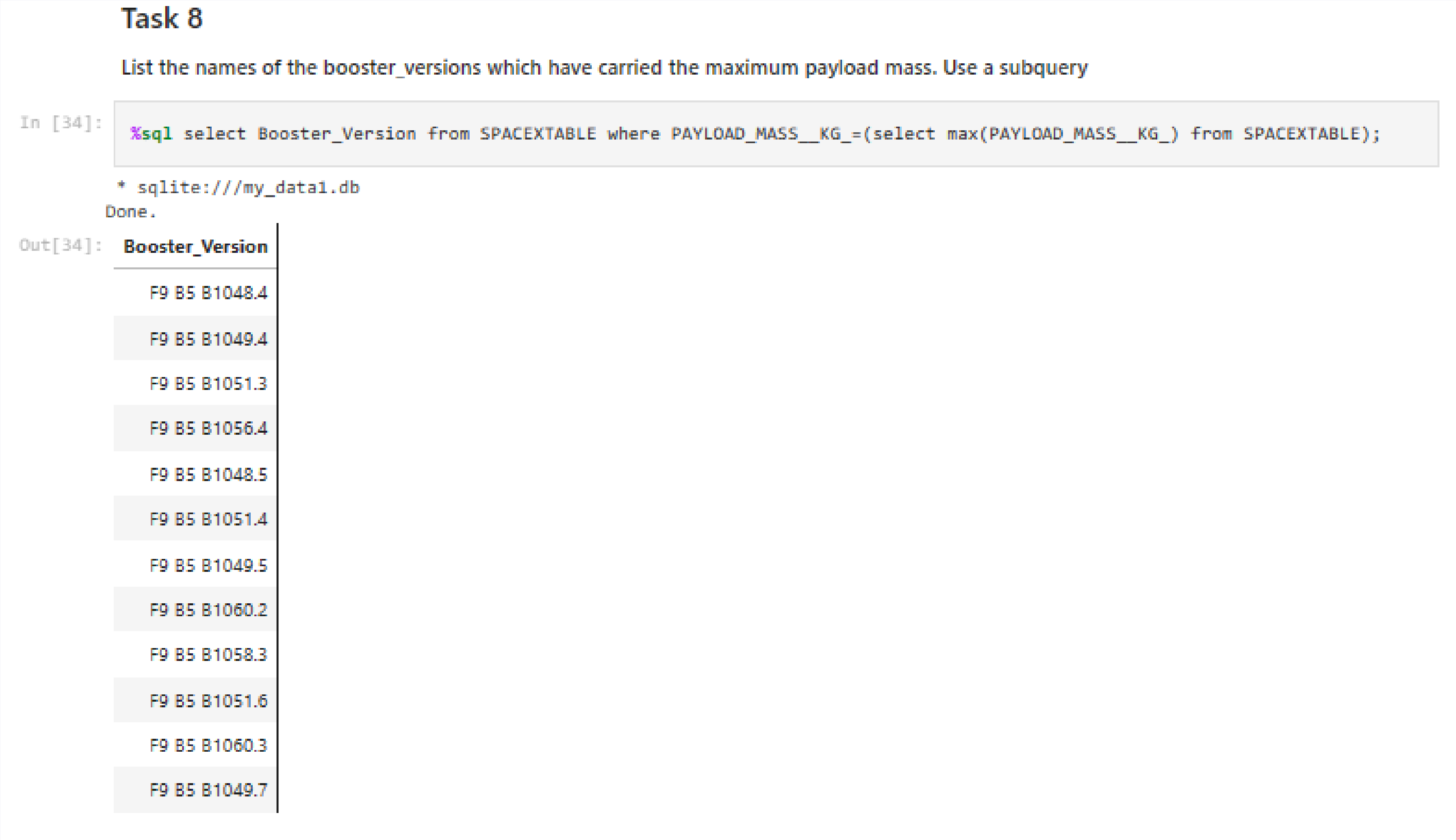
```
In [33]: %sql select Mission_Outcome,count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome ;
```

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

Out[33]:

Mission_Outcome	count(Mission_Outcome)
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

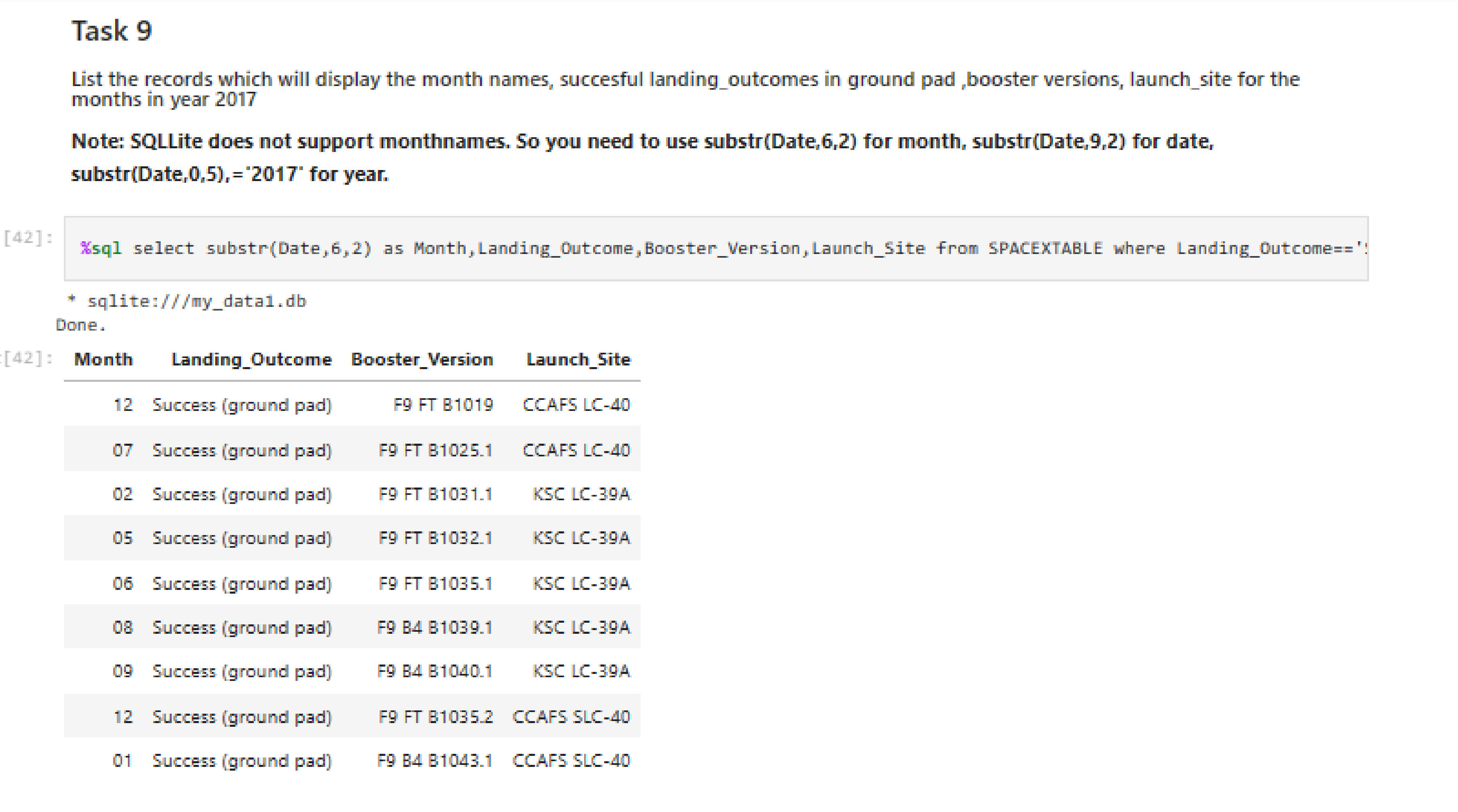
```
In [34]: %sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);
```

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

Out[34]:

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

2015 Launch Records



Task 9

List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

Note: SQLLite does not support monthnames. So you need to use substr(Date,6,2) for month, substr(Date,9,2) for date, substr(Date,0,5),='2017' for year.

[42]:

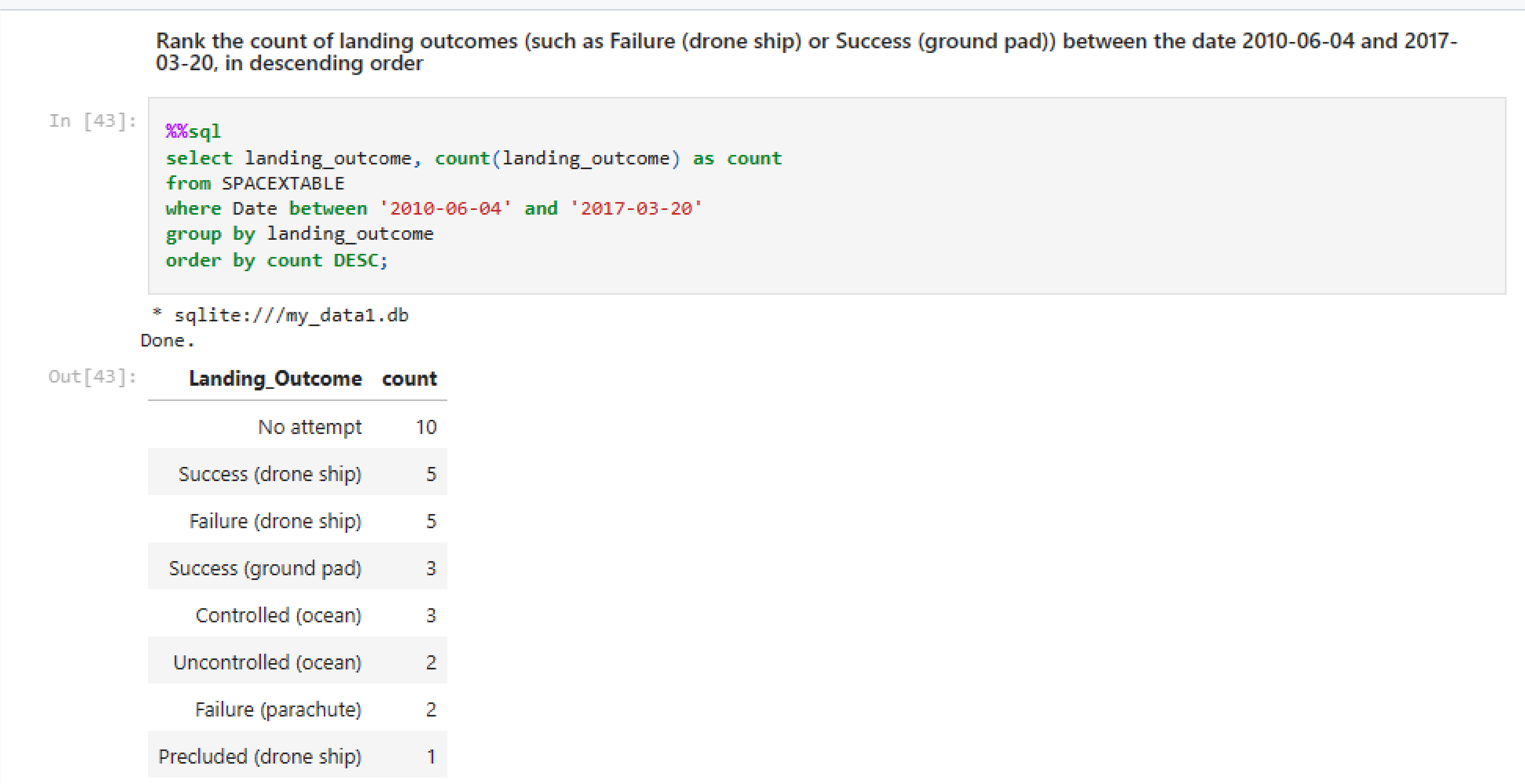
```
%sql select substr(Date,6,2) as Month,Landing_Outcome,Booster_Version,Launch_Site from SPACEXTABLE where Landing_Outcome=='
```

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

[42]:

Month	Landing_Outcome	Booster_Version	Launch_Site
12	Success (ground pad)	F9 FT B1019	CCAFS LC-40
07	Success (ground pad)	F9 FT B1025.1	CCAFS LC-40
02	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
05	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
06	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
08	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
09	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
12	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40
01	Success (ground pad)	F9 B4 B1043.1	CCAFS SLC-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

In [43]:

```
%%sql
select landing_outcome, count(landing_outcome) as count
from SPACEXTABLE
where Date between '2010-06-04' and '2017-03-20'
group by landing_outcome
order by count DESC;
```

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

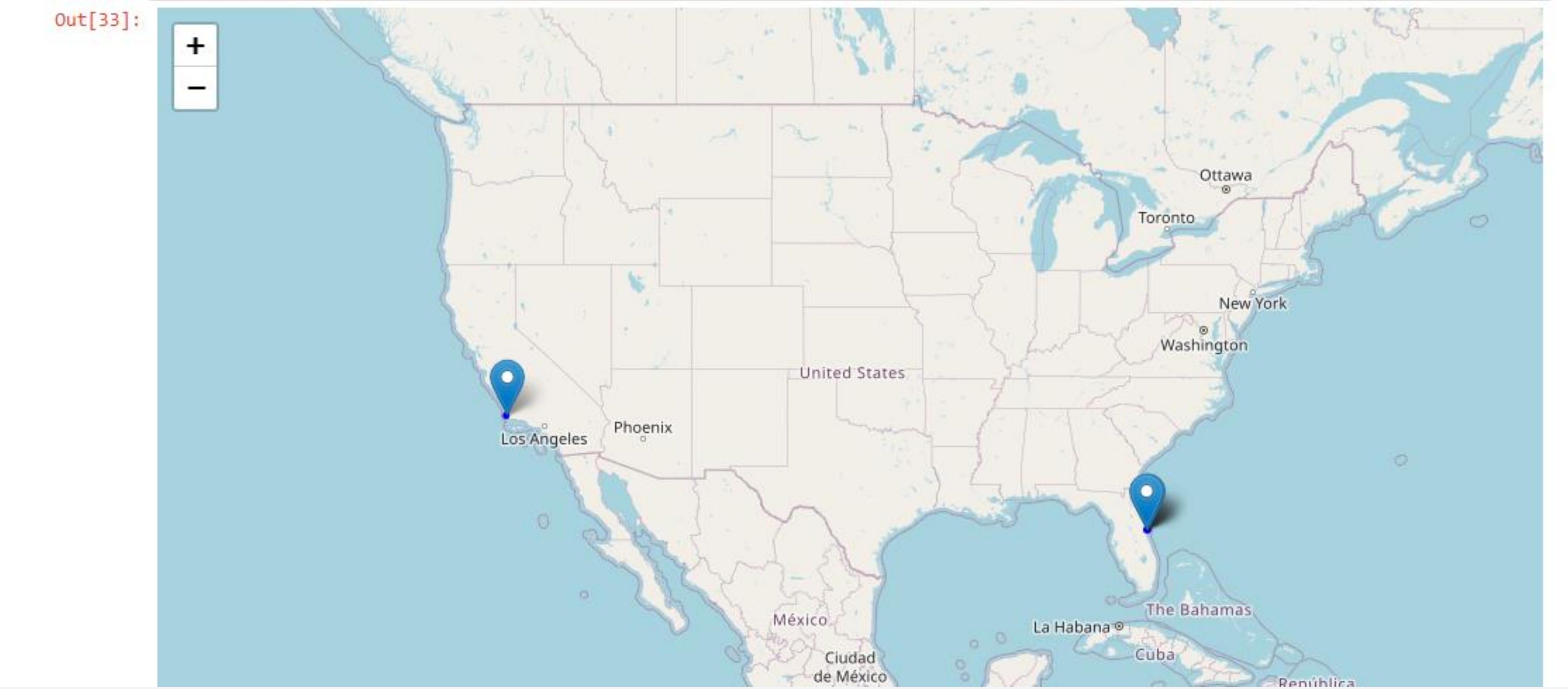
Out[43]:

Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

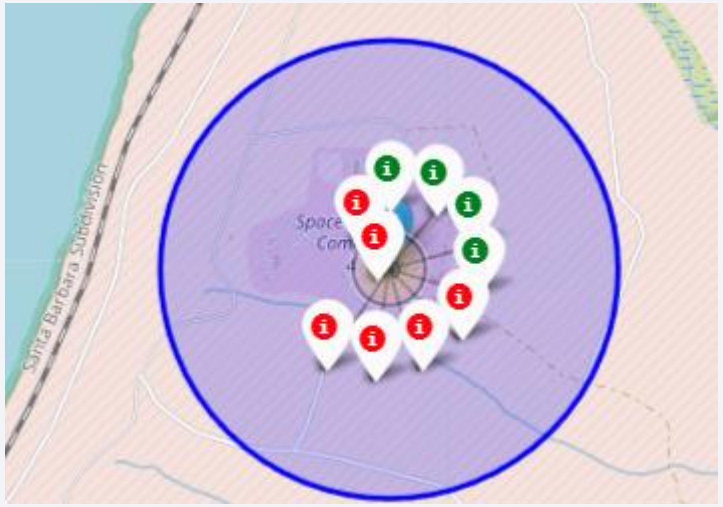
Section 3

Launch Sites Proximities Analysis

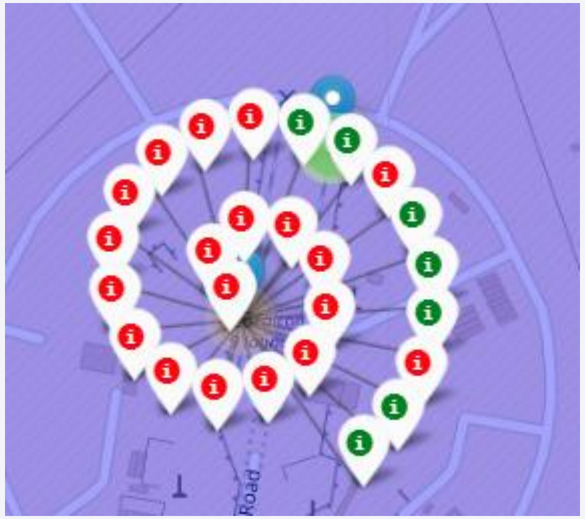
SpaceX Falcon9 - Launch Sites Map



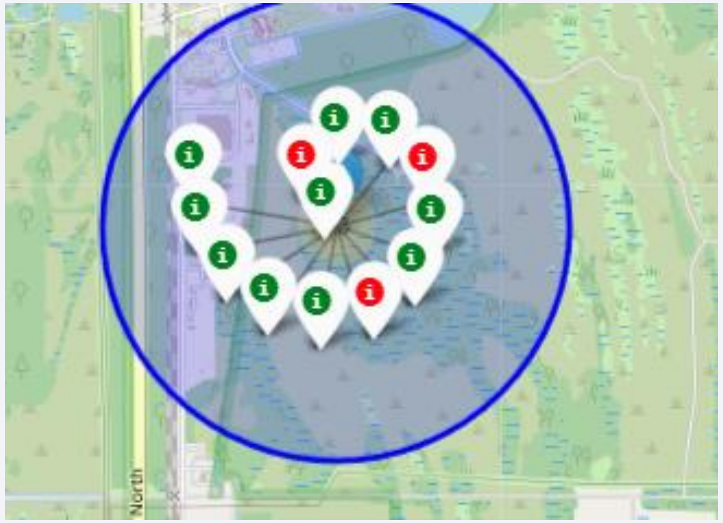
SpaceX Falon9 – Success/Failed Launch Map for all Launch Sites



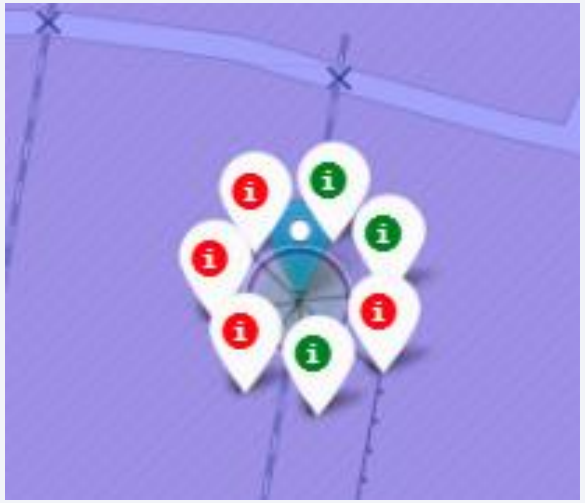
VAFB SLC 4E



CCAFS SLC-40



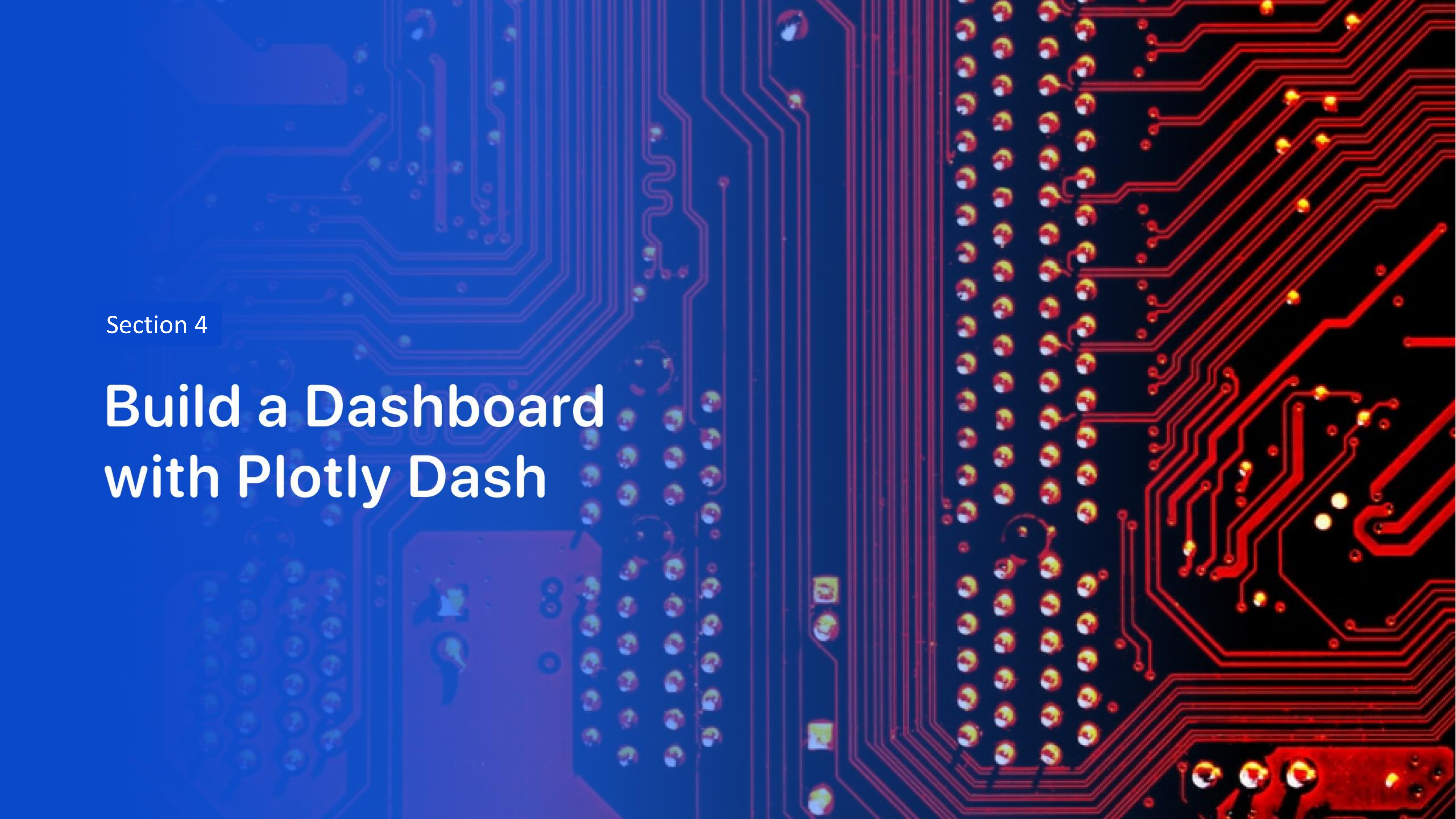
KSC LC-39A



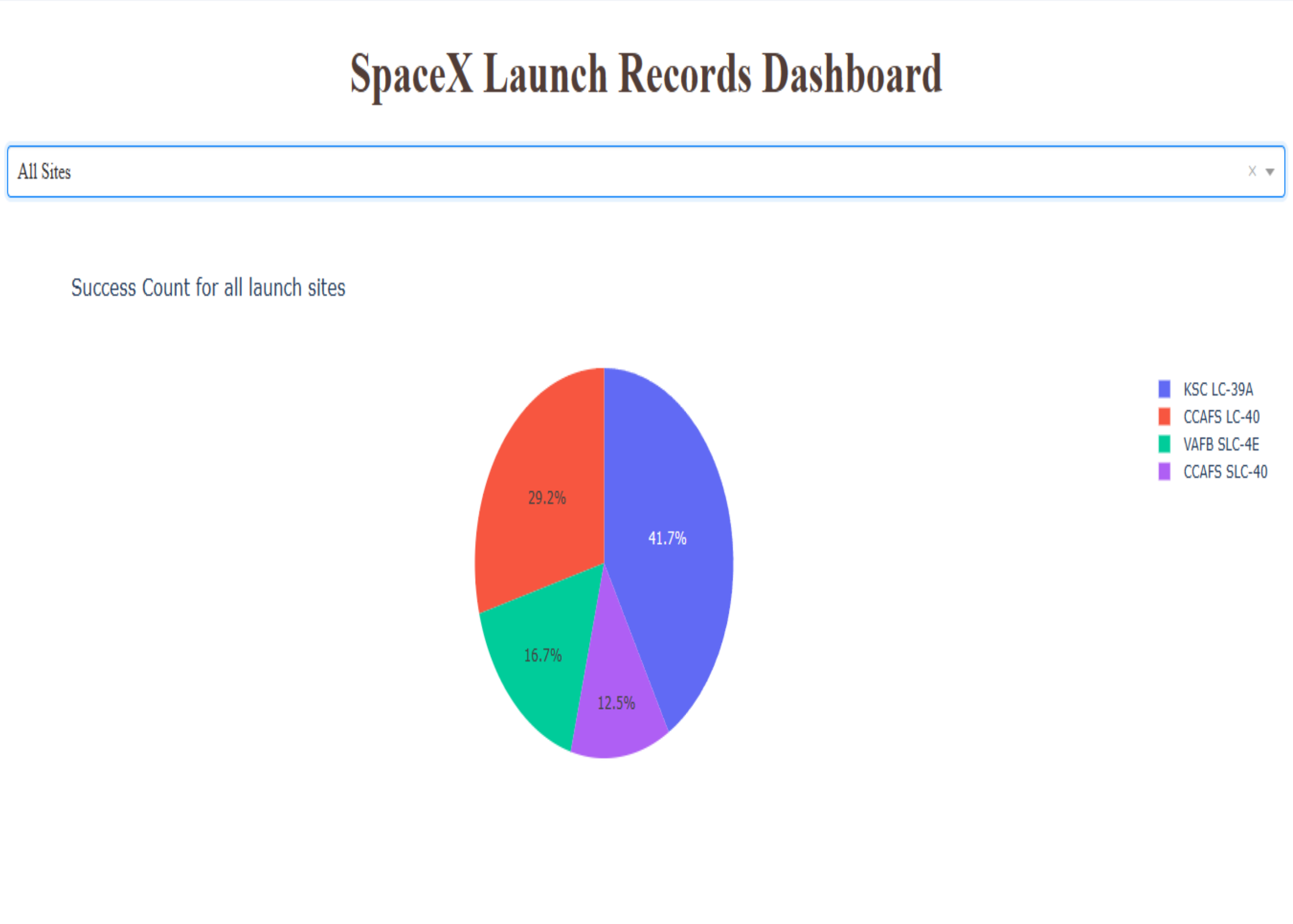
CCAFS SLC-40

Section 4

Build a Dashboard with Plotly Dash



Launch Success Counts For All Sites



SpaceX Launch Records Dashboard

All Sites

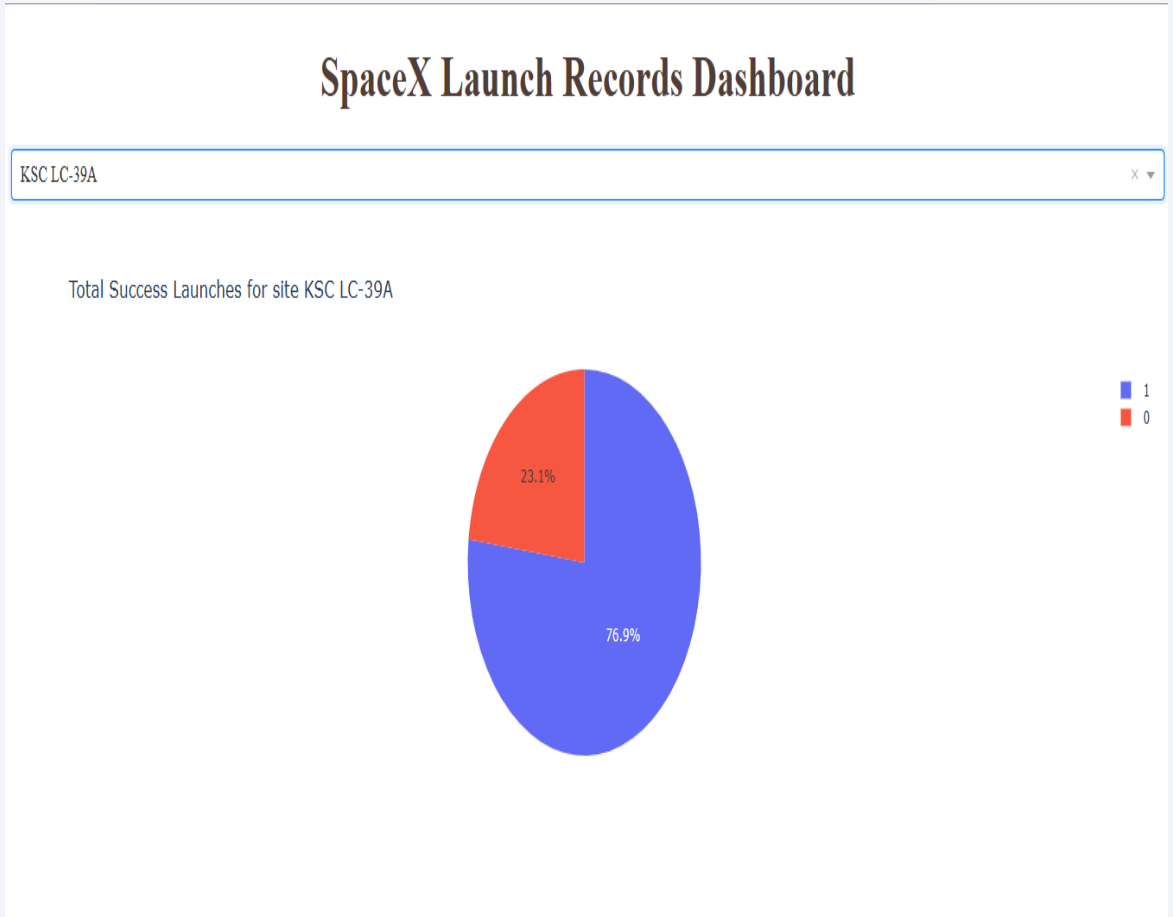
Success Count for all launch sites

- Launch Site ‘KSC LC-39A’ has the highest launch success rate

- Launch Site ‘CCAFS SLC-40’ has the lowest launch success rate

Launch Success Counts For KSC LC-39A

- KSC LC-39A Launch Site has the highest launch success rate and count
- Launch success rate is 76.9%
- Launch success failure rate is 23.1%

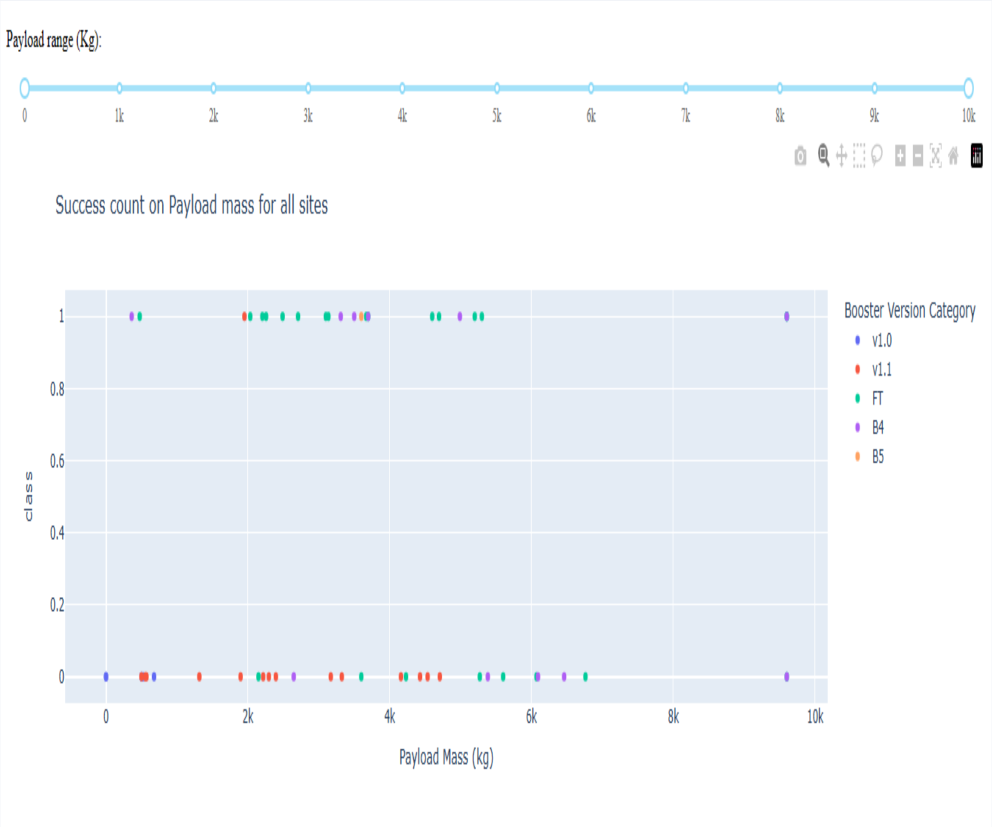


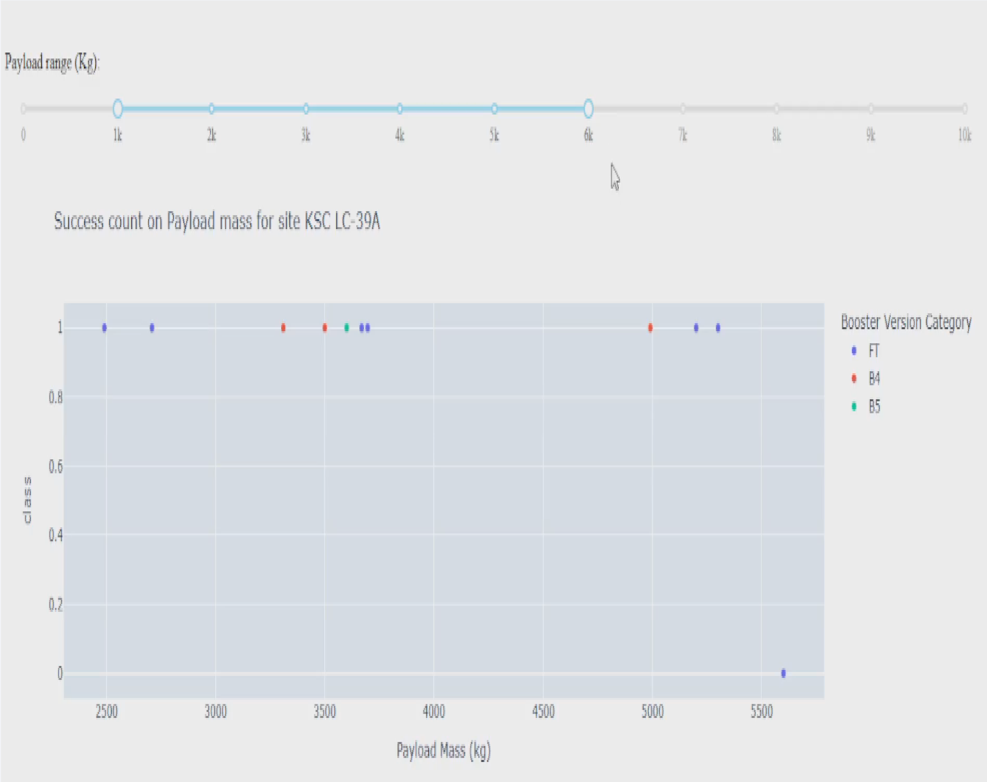
SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for site KSC LC-39A

Payload vs. Launch Outcome





- Most successful launches are in the payload range from 2000 to about 5500

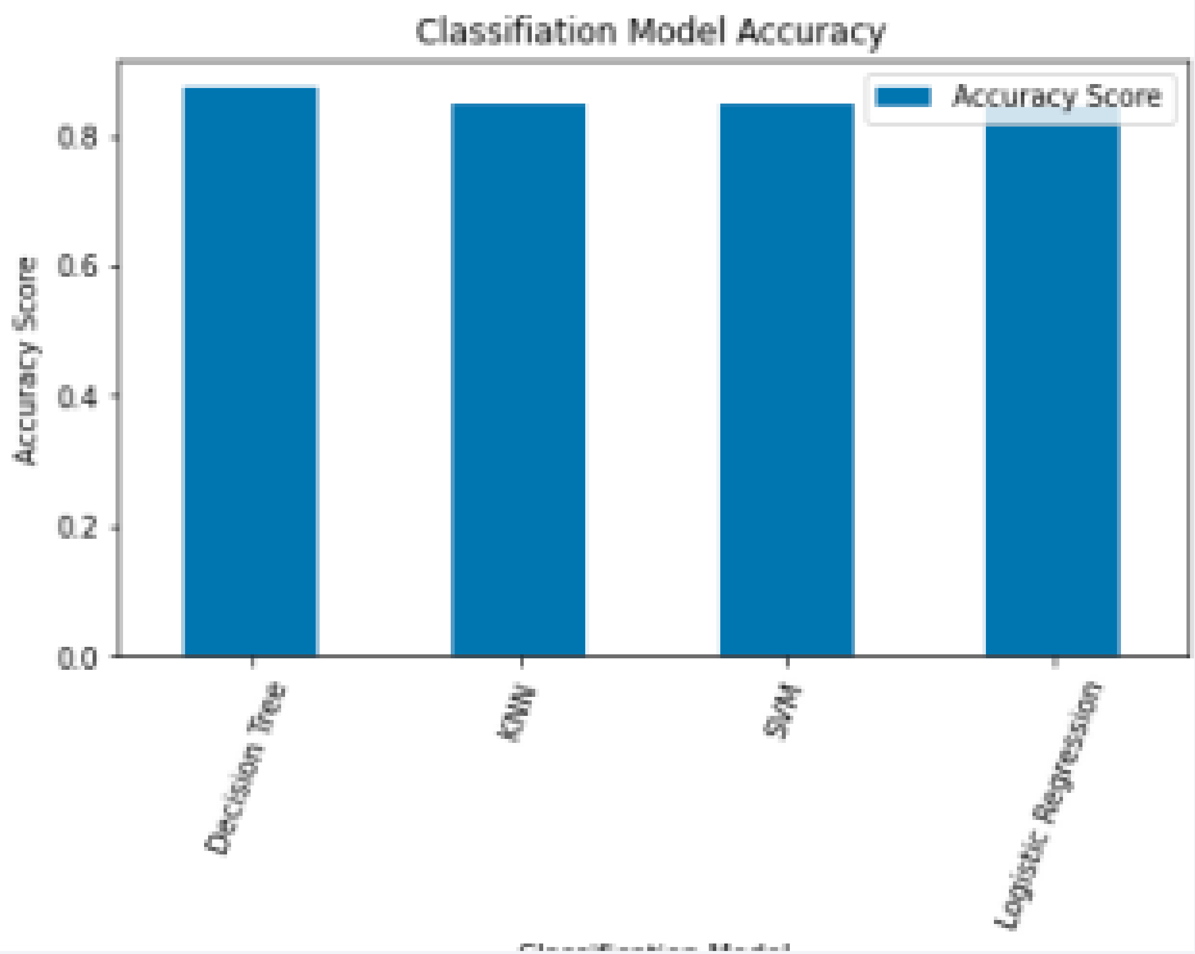
- The success rate for KSC LC 39A is almost perfect between 2000-6000 Kg

Section 5

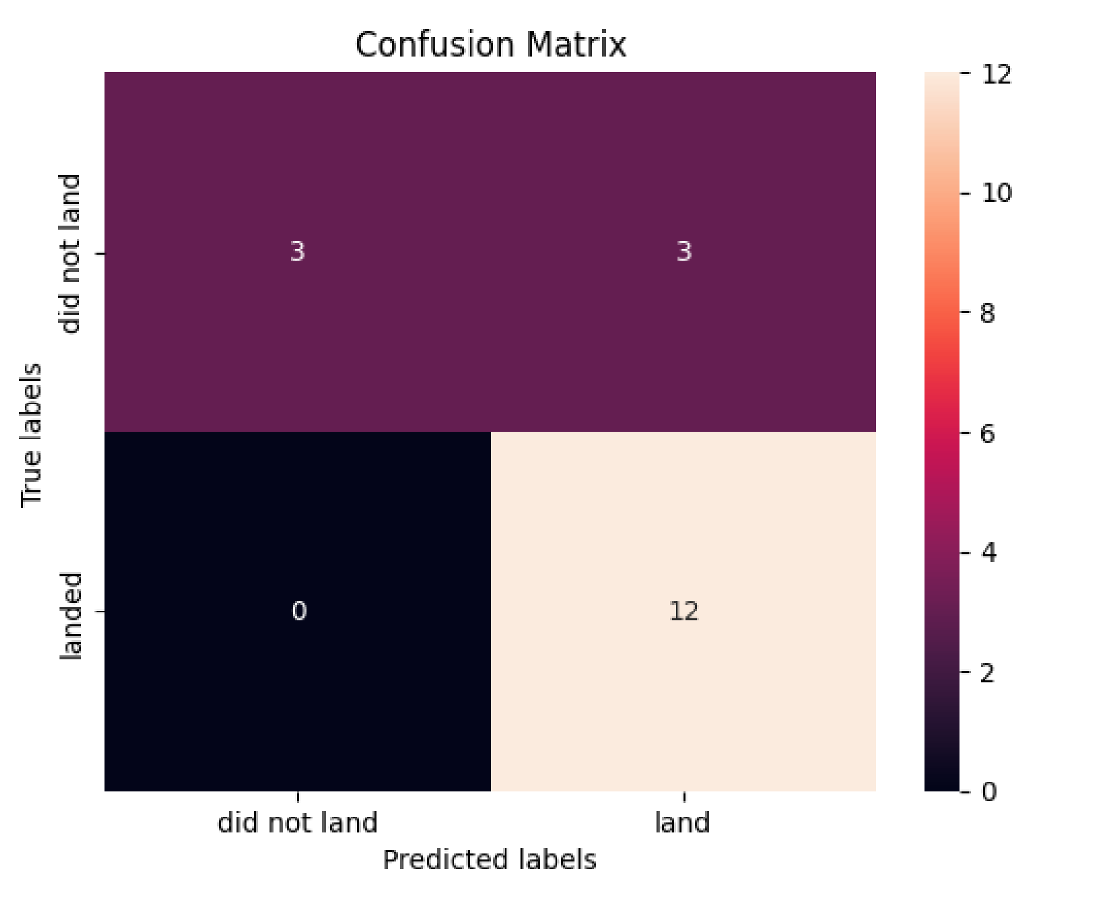
Predictive Analysis (Classification)

Classification Accuracy

- Based on the Training Accuracy, Decision Tree algorithm has the highest classification score with a value of .8750
- Accuracy Score on the test data is the same for all the classification algorithms based on the data set with a value of .8333



Confusion Matrix



- The confusion matrix is same for the models LR, SVM, KNN
- Per the confusion matrix, the classifier made 18 predictions
- 12 scenarios were predicted Yes for landing, and they did land successfully (True positive)
- 3 scenarios (top left) were predicted No for landing, and they did not land (True negative)
- 3 scenarios (top right) were predicted Yes for landing, but they did not land successfully (False positive)
- Overall, the classifier is correct about 83% of the time ((TP + TN) / Total) with a misclassification or error rate ((FP + FN) / Total) of about 16.5%

Conclusions

- As the numbers of flights increase, the first stage is more likely to land successfully
- Launch success rate increased by about 80% from 2013 to 2020
- Launch Site ‘KSC LC-39A’ has the highest launch success rate and Launch Site ‘CCAFS SLC-40’ has the lowest launch success rate
- Orbits ES-L1, GEO, HEO, and SSO have the highest launch success rates and orbit GTO the lowest
- Launch sites are located strategically away from the cities and closer to coastline, railroads, and highways
- Accuracy Score on the test data is the same for all the classification algorithms based on the data set with a value of .8333. The decision tree algorithm has slightly better train set accuracy.

Thank you!