IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mikhail Nikitin
15/10/2024

Outline

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

Executive Summary

Summary of methodologies:

- ✓ Data collection: Web Scrapping and SpaceX API
- ✓ EDA: SQL, Data Visualization (Matplotlib, Seaborn, Plotly, Follium, Dash)
- ✓ ML prediction: Logistic Regression, SVM, Decision Tree, K Nearest Neighbors

Summary of all results:

- ✓ All necessary data was collected
- ✓ Exploratory Data analysis was conducted and insights about data were used for Machine Learning prediction of launch outcome
- ✓ Launch outcome was predicted using different ML techniques

Introduction

In the era of commercial space travel, reducing costs is crucial. SpaceX advertises its Falcon 9 rocket launches at a fraction of the cost of other providers, largely due to the reuse of the first stage. By predicting the success of these landings, companies can optimize their bids and potentially compete with SpaceX.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected through two main sources: SpaceX API and Wikipedia page webscrapping
- Perform data wrangling
  - Data was analyzed and described, empty values were filled and one-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four different models (LogReg, SVM, Decision Trees and k-NN) were applied to predict landing outcome

Data Collection

Space X API Data:

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

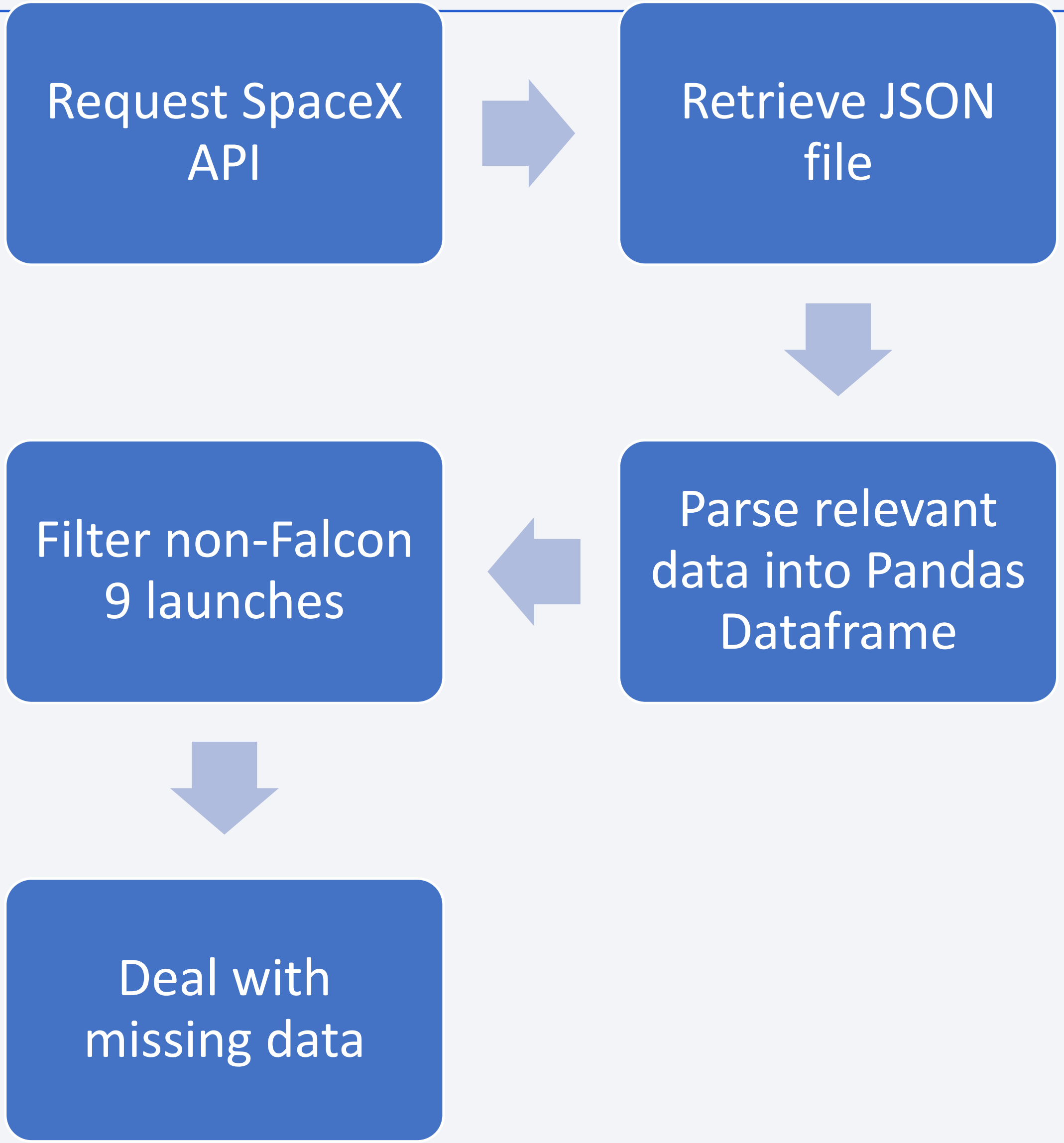
Wikipedia Webscrape Data:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date and Time.

Data Collection – SpaceX API

- SpaceX offers a public API which contains the data about all launches
- The data was obtained and used according to the flowchart

Notebook is here



Request SpaceX API

Retrieve JSON file

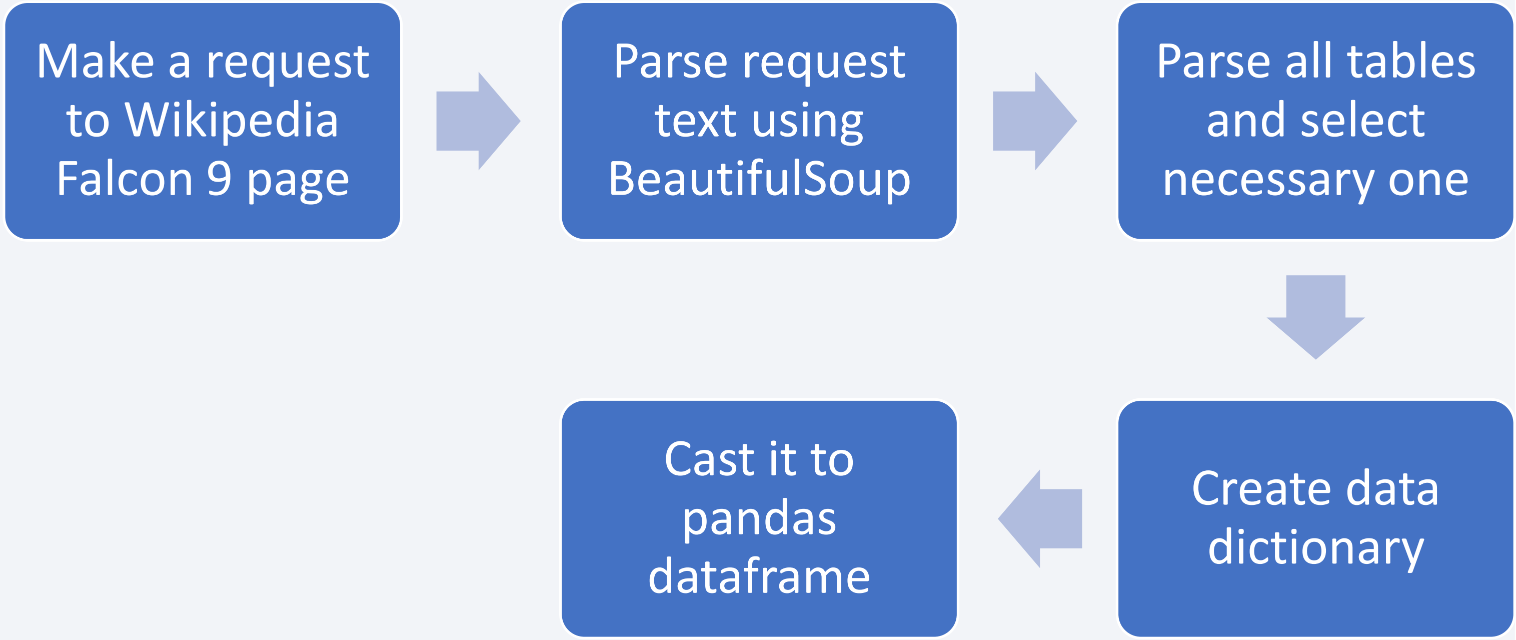
Parse relevant data into Pandas Dataframe

Filter non-Falcon 9 launches

Deal with missing data

Data Collection - Scraping

- Data from Wikipedia were parsed using BeautifulSoup and transformed into dataframe
- Notebook is here



Make a request to Wikipedia Falcon 9 page

Parse request text using BeautifulSoup

Parse all tables and select necessary one

Create data dictionary

Cast it to pandas dataframe

Data Wrangling

1. identify the data types of the columns.

2. Determine the number of values for each attribute.

3. Calculate the percentage of the missing values.

4. To determine the label, we apply zero/one hot encoding to the "Outcome" column to classify landing to either 1(Success) of 0 (Failure)

EDA with Data Visualization

- To explore data Scatter plots, Line charts, and Bar plots were used to compare relationships between variables to decide if any direct or indirect relationship exists.
- The variables: Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year were visualized graphically. Plots Used: Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend.
- Data Visualization notebook

EDA with SQL

Dataset was analyzed using sqlite3 and following queries were received:

- Names of the unique launch sites in the space mission.
- 5 records where launch sites begin with the string 'CCA’
- The total payload mass carried by boosters launched by NASA (CRS).
- Average payload mass carried by booster version F9 v1.1
- The date when the first successful landing outcome in ground pad was achieved.
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- The total number of successful and failure mission outcomes.
- The names of the booster versions which have carried the maximum payload mass.
- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

SQL EDA Notebook

Build an Interactive Map with Folium

Markers, circles, lines, and marker clusters were used with folium Maps.

- Markers indicate points like Launch Sites.
- Circles indicate highlighted areas around specific co-ordinates like NASA Johnson Space Centre.
- Lines speak of the distance between two co-ordinates.
- Marker Clusters indicate group of events in each co-ordinate for e.g. launches in a launch site.

Follium notebook

Build a Dashboard with Plotly Dash

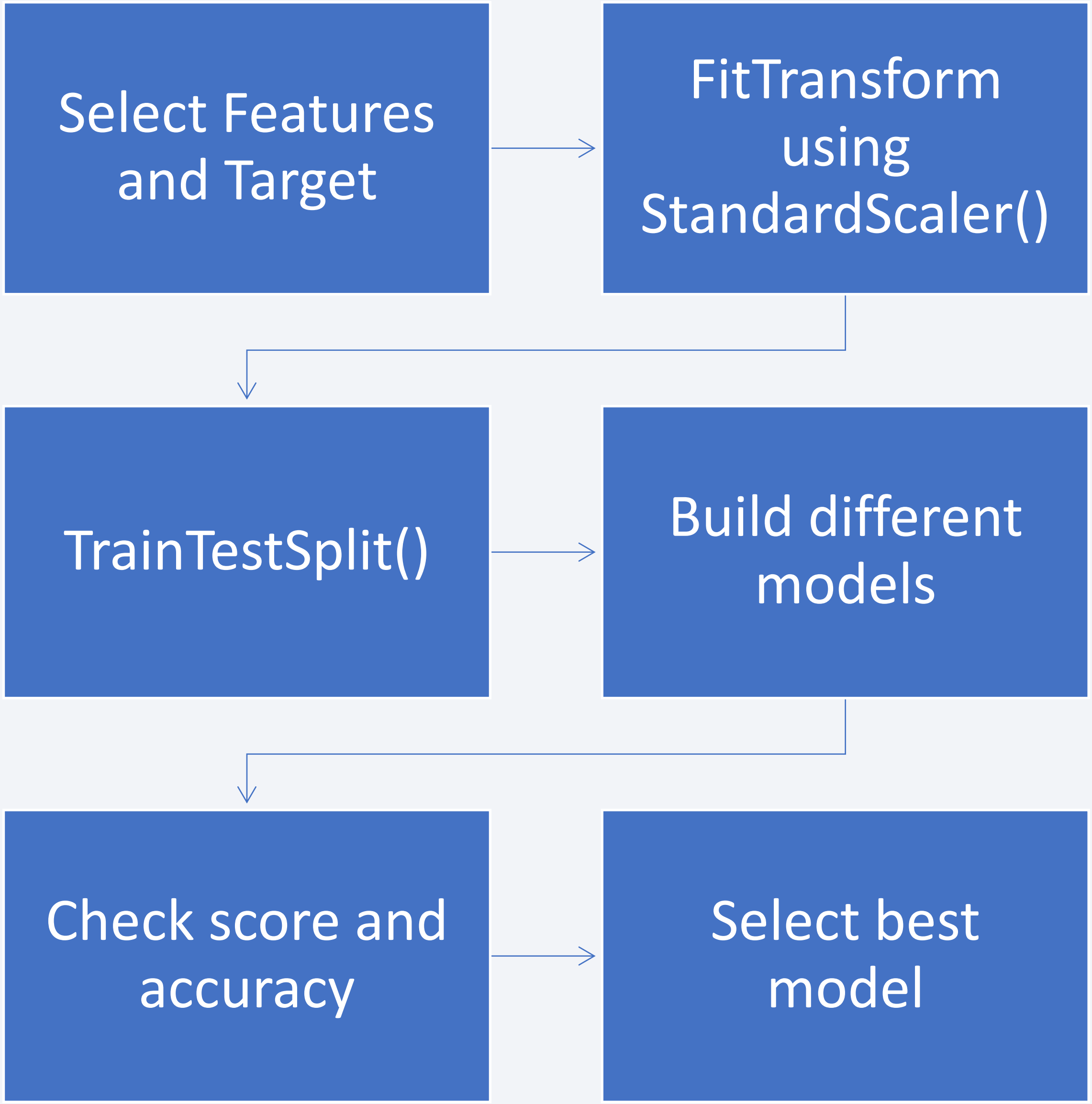
- We have used a pie chart and a scatter plot.
- • Pie chart is used to show the distribution of successful landings across all launch sites and can also be used to show individual launch site's success rates.
- • Scatter plot takes two inputs: All sites or any individual launch site and Payload mass on a slider between 0 and 10000 kg.

This combination allowed us to quickly analyze the relation between Payload and launch sites so that we can select best launch sites.

Dash Python code

Predictive Analysis (Classification)

- Four Machine Learning Models were compared- Logistic regression, Support Vector Machine, K Nearest Neighbor and Decision Trees
- Based on accuracy, Decision trees made the best result with 88.9%
- Source code



Select Features and Target

FitTransform using StandardScaler()

TrainTestSplit()

Build different models

Check score and accuracy

Select best model

Results

Exploratory data analysis results o SpaceX uses 4 different launch sites

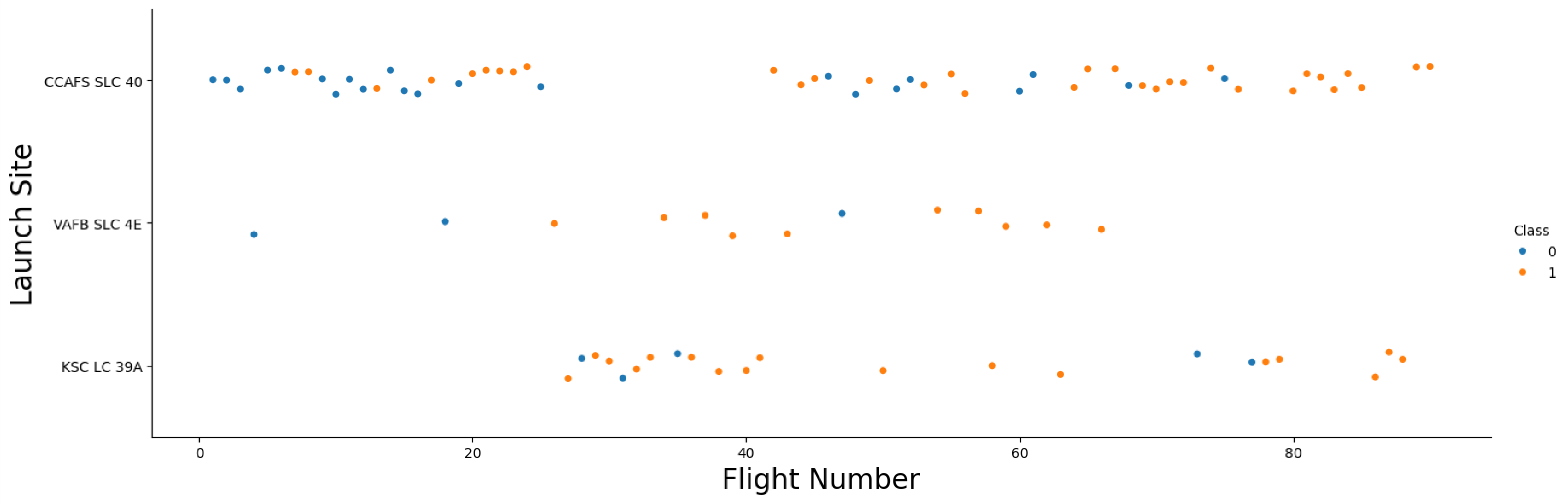
- The first launches were done to Space X itself and NASA
- The average payload of F9 v1.1 booster is 2,928 kg
- The first success landing outcome happened in 2015 five year after the first launch
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- Almost 100% of mission outcomes were successful;
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
- The number of landing outcomes became as better as years passed.

Section 2

Insights drawn from EDA

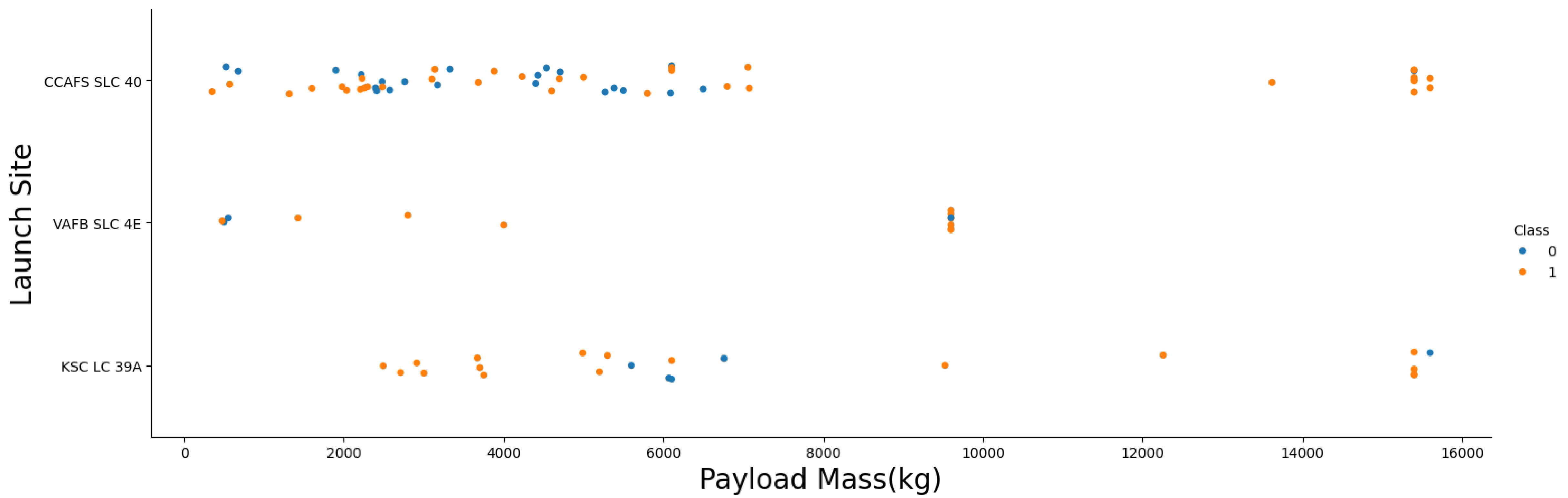
Flight Number vs. Launch Site

- In late flight numbers, VAFB SLC4E launch site was left
- Most likely, there was a huge development around flight 20 that greatly improved the success rate.
- Success rate was increasing with flight number for all sites, with no unsuccessful flight numbers after flight #77



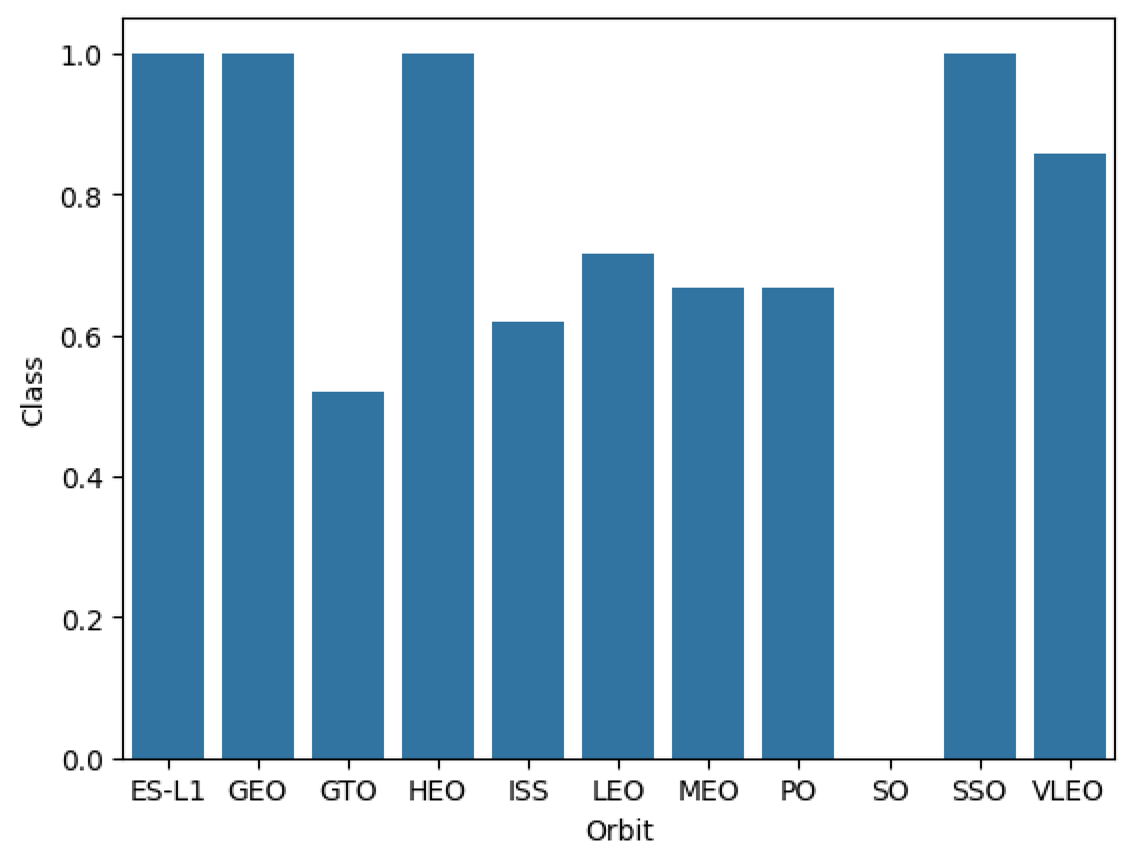
Payload vs. Launch Site

- All the flights with payload > 10,000 kg were made from CCAFS SCL 40 and KSC LC 39A sites
- There was only two flight with that heavy payload which were unsucessfull



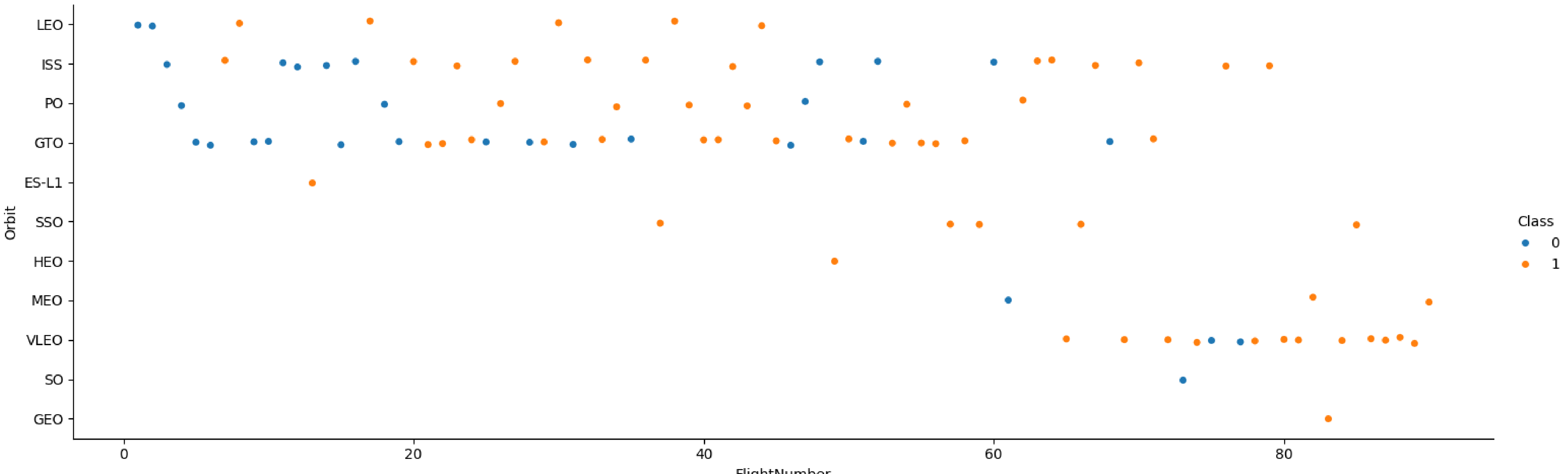
Success Rate vs. Orbit Type

- For ES-L1, GEO and SSO orbits launches were always successful
- SO, GTO and ISS orbits were the least successful, with only one launch on SO orbit failed



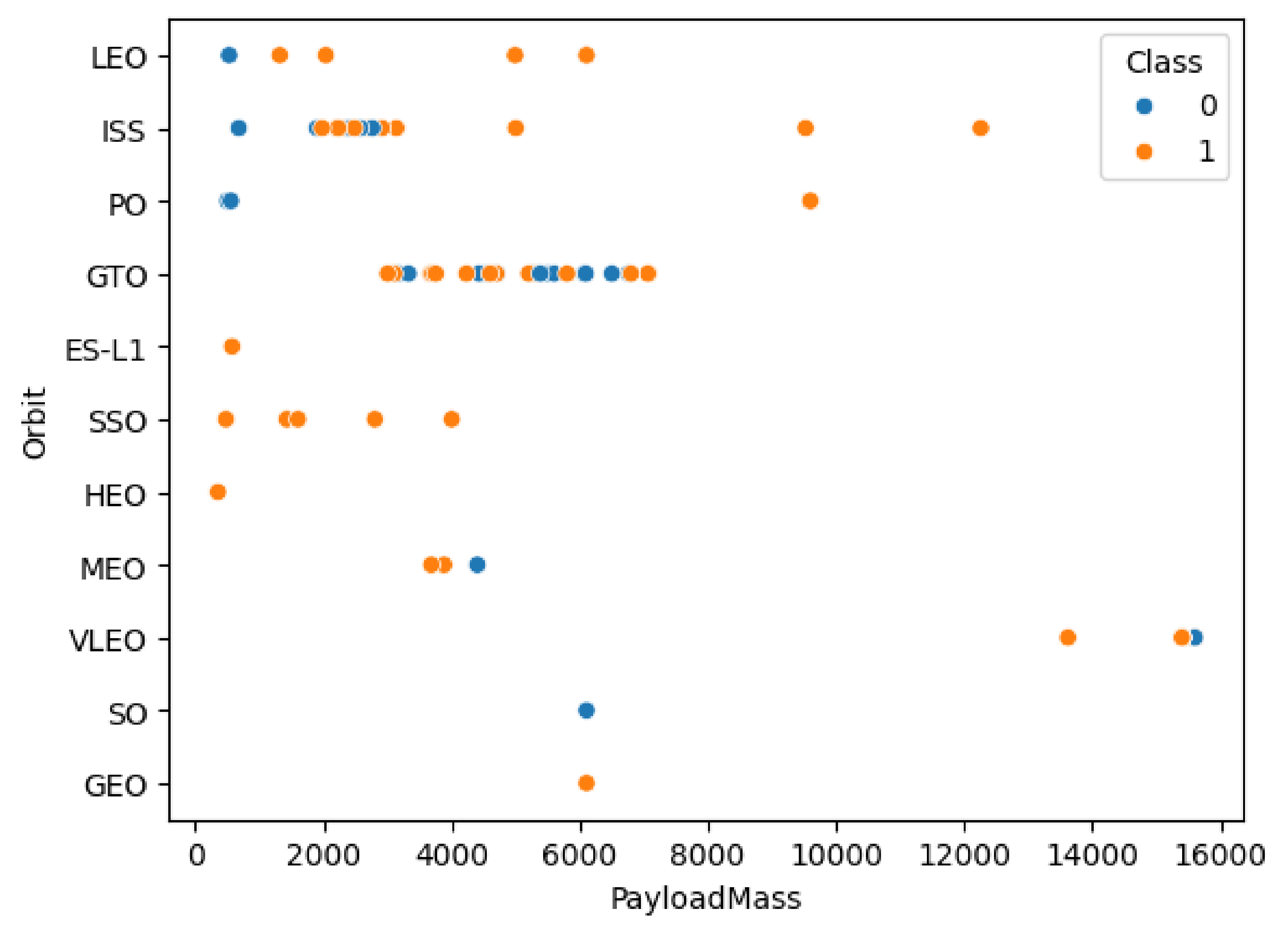
Flight Number vs. Orbit Type

- 24 • Launch Orbit preferences changed over Flight Number.
- • Launch Outcome seems to correlate with this preference.
- • SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches
- • SpaceX appears to perform better in lower orbits or Sun-synchronous orbits



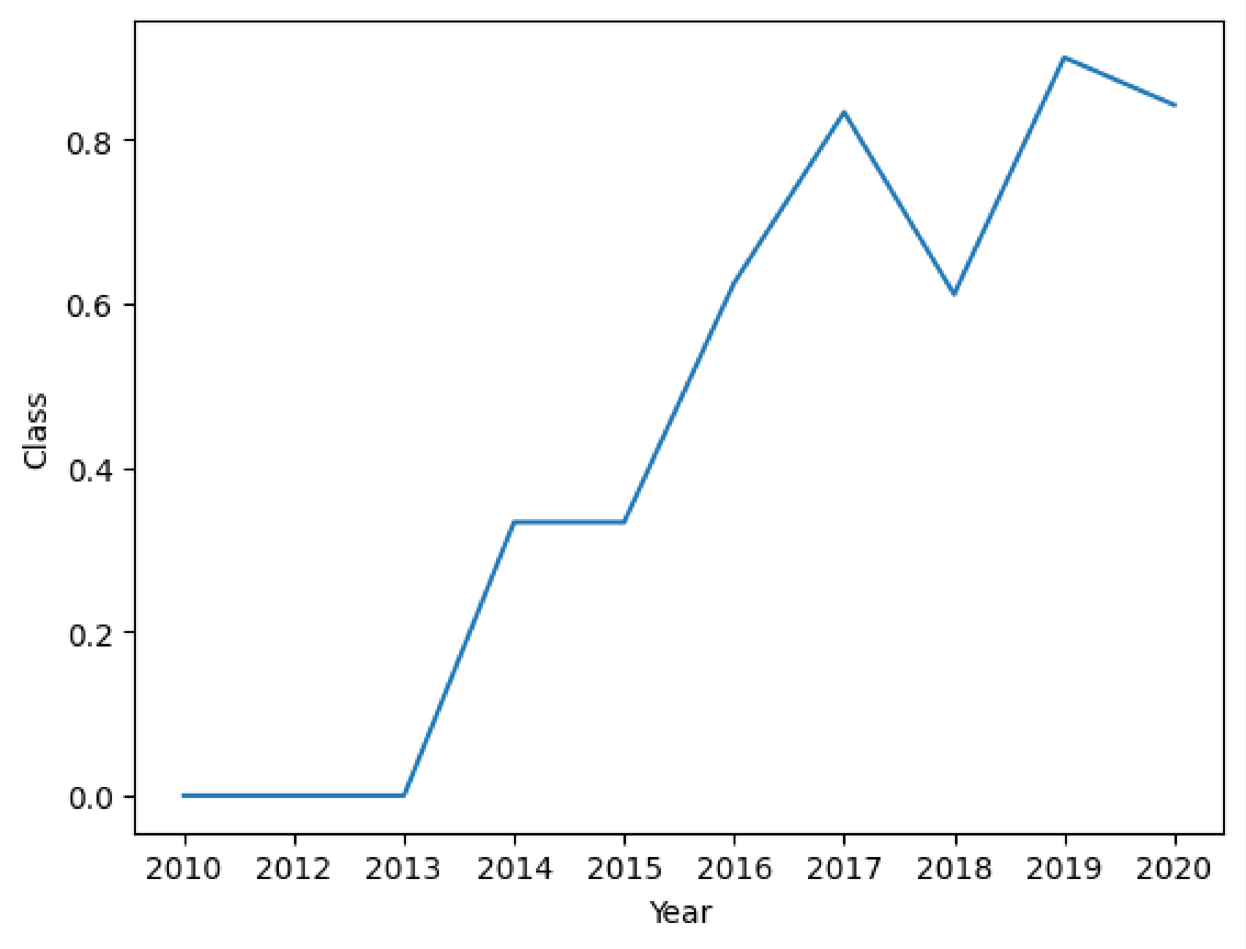
Payload vs. Orbit Type

- Different orbits are used for different payloads
- VLEO orbit was used for heavy payload
- Launches with high payload mass was more successful in average



Launch Success Yearly Trend

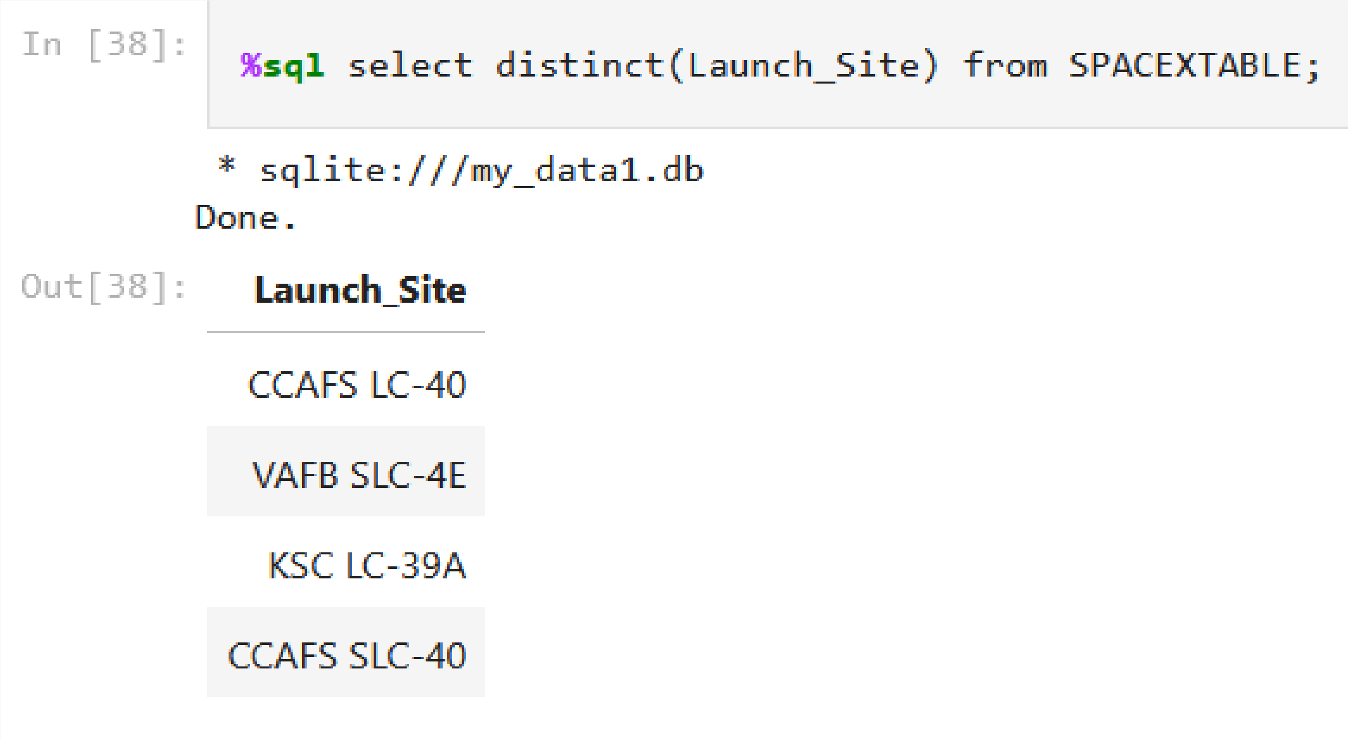
- Launch success rate were growing over time with a little delay in 2014-15 and little decrease in 2018
- After 2013, technologies were most likely modified, which led to a increasing performance



All Launch Site Names

There are four unique launch sites as per data:

- CCAFS LC-40(Old name for CCAFS SLC-40)
- CCAFS SLC-40 (Cape Canaveral Space Launch Complex 40) (Brevard County, Florida)
- KSC LC-39A (Kennedy Space Center Launch Complex 39) (Merritt Island, Florida)
- VAFB SLC-4E(Vandenberg Space Launch Complex 4) (Vandenberg Space Force Base, California)



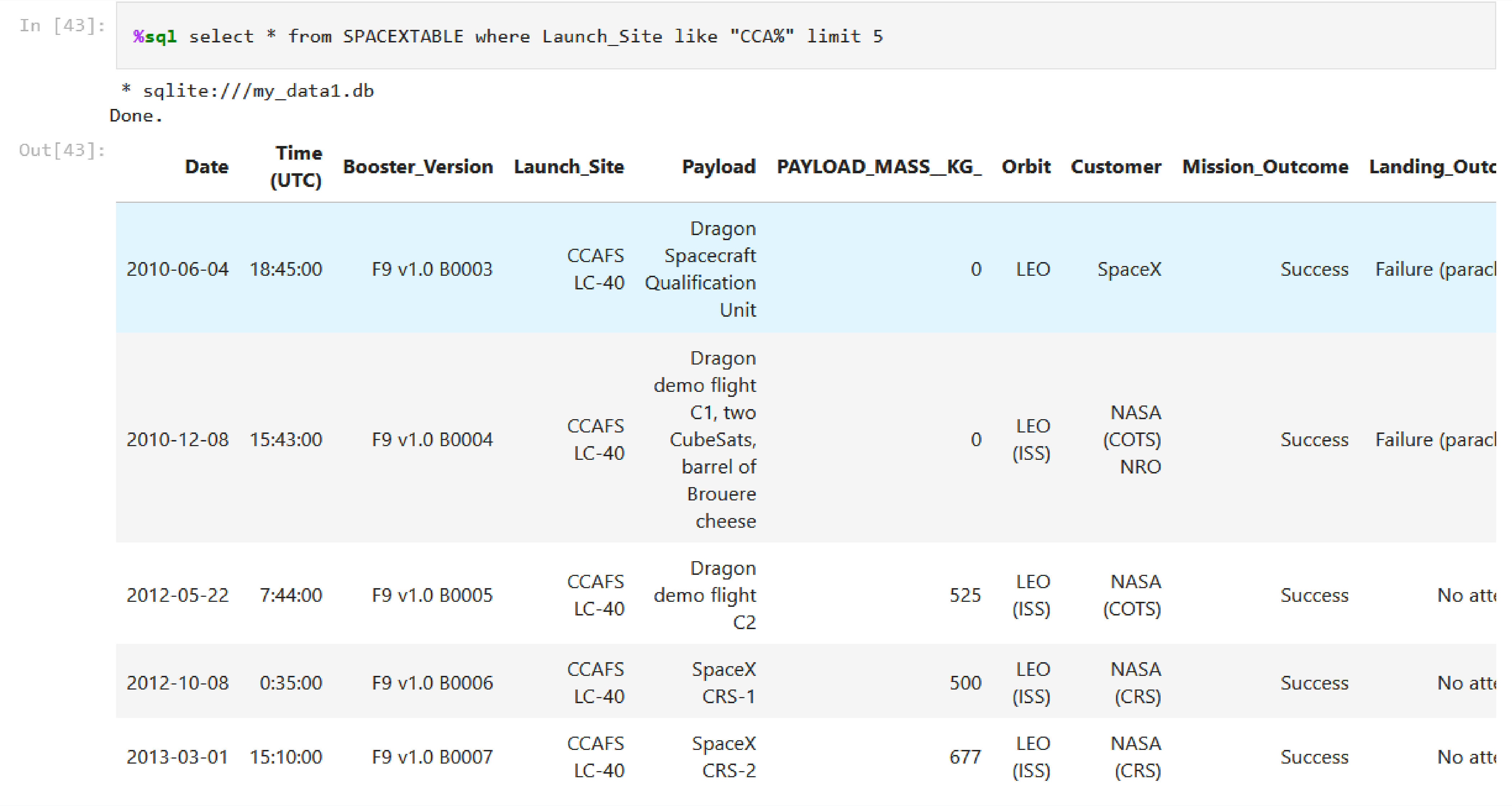
```
In [38]:  %sql select distinct(Launch_Site) from SPACEXTABLE;

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

Out[38]:

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

Launch Site Names Begin with 'CCA'



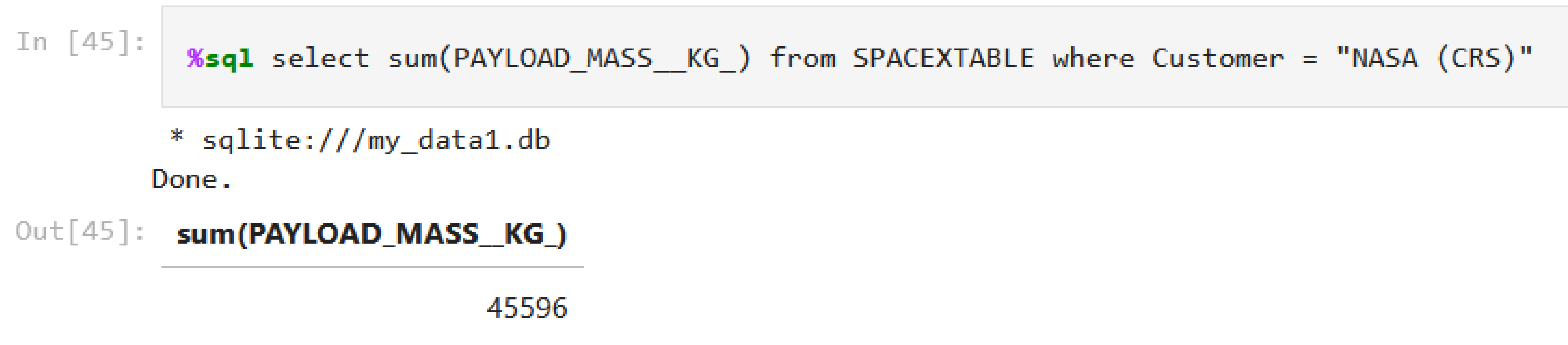
```
In [43]:  %sql select * from SPACEXTABLE where Launch_Site like "CCA%" limit 5
```

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

Out[43]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (paracl
2010-12-08	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 (paracl
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No atte
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No atte
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No atte

Total Payload Mass



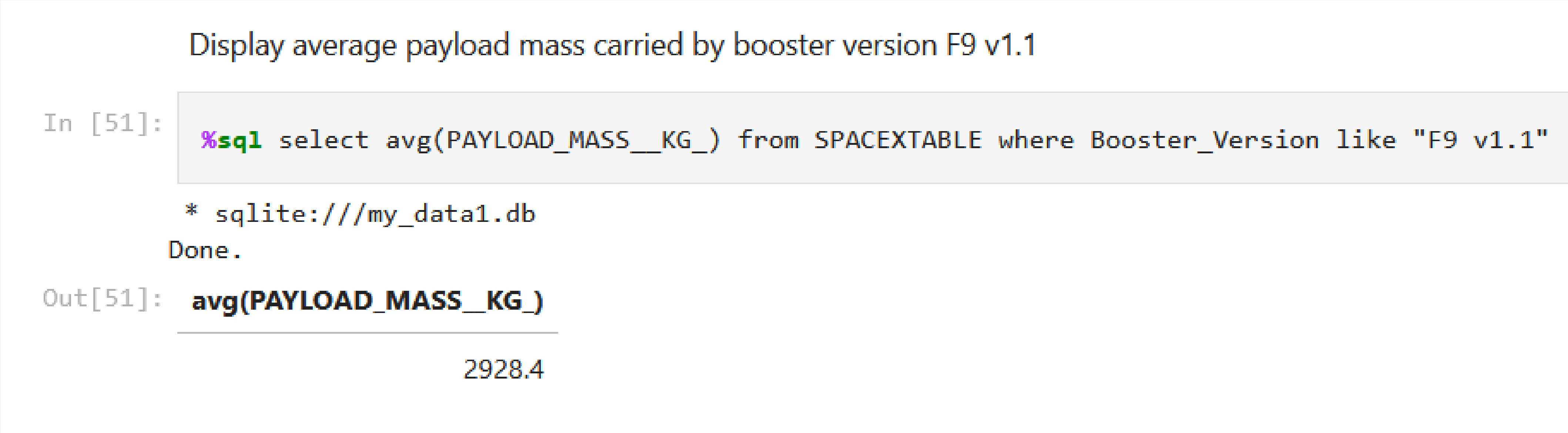
```
In [45]:  %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = "NASA (CRS)"

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

Out[45]:

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1



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

In [51]:

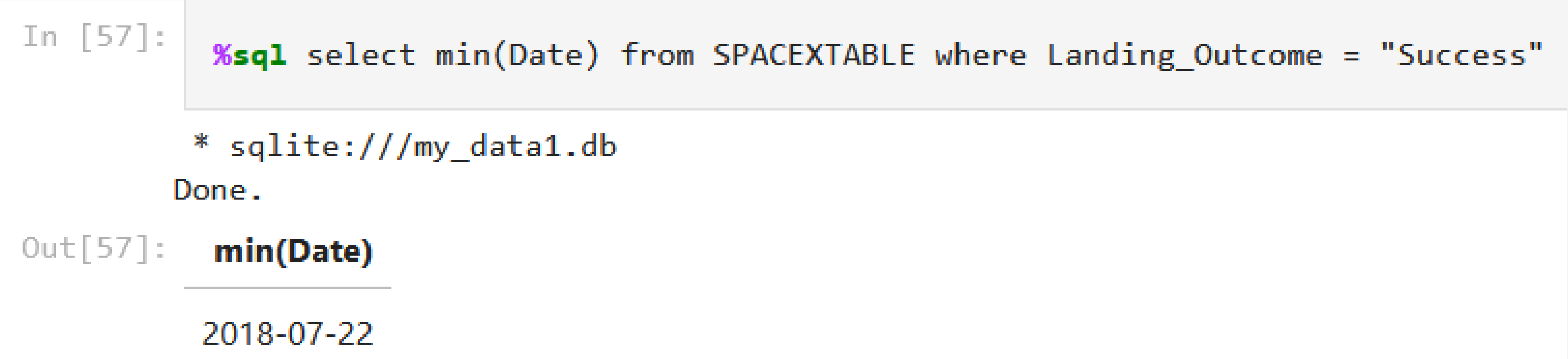
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like "F9 v1.1"
```

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

Out[51]:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date



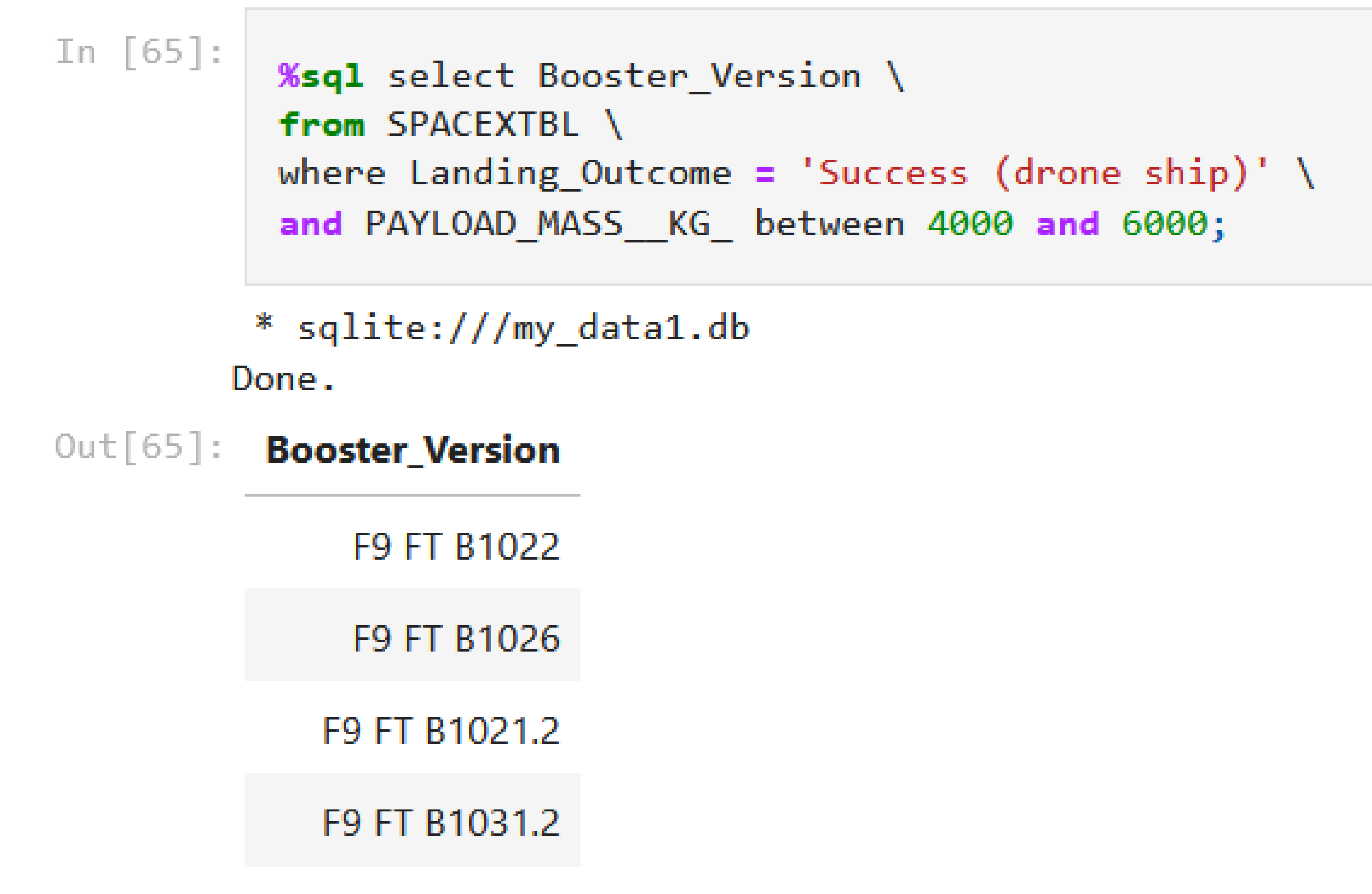
```
In [57]: %sql select min(Date) from SPACEXTABLE where Landing_Outcome = "Success"

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

Out[57]:

min(Date)
2018-07-22

Successful Drone Ship Landing with Payload between 4000 and 6000



In [65]:

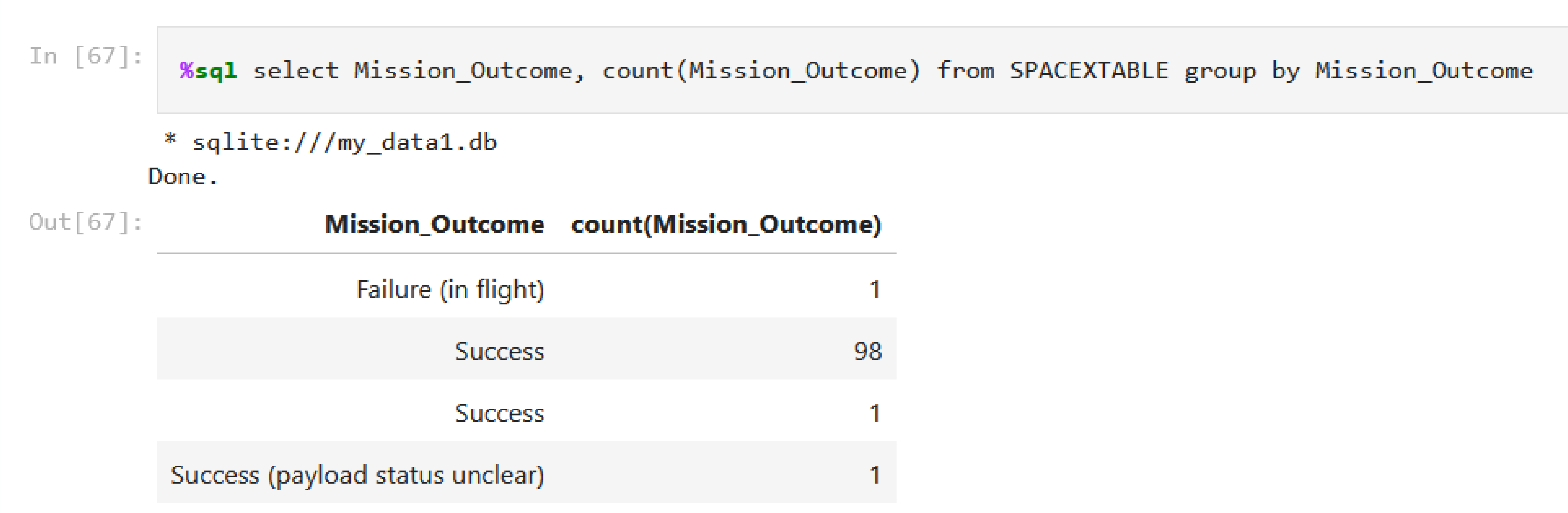
```
%sql select Booster_Version \
from SPACEXTBL \
where Landing_Outcome = 'Success (drone ship)' \
and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

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

Out[65]:

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

Total Number of Successful and Failure Mission Outcomes



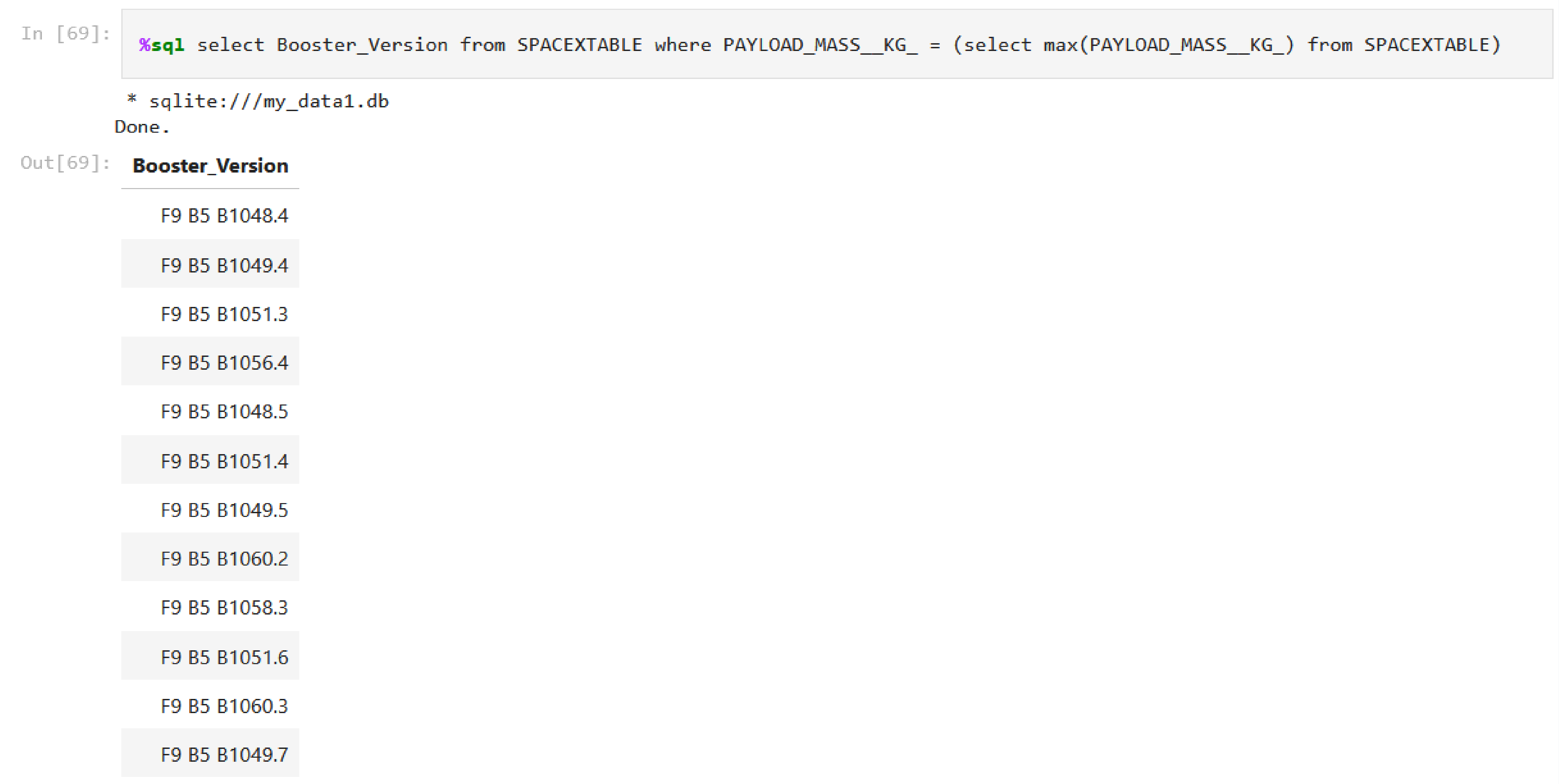
```
In [67]: %sql select Mission_Outcome, count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome
```

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

Out[67]:

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

Boosters Carried Maximum Payload



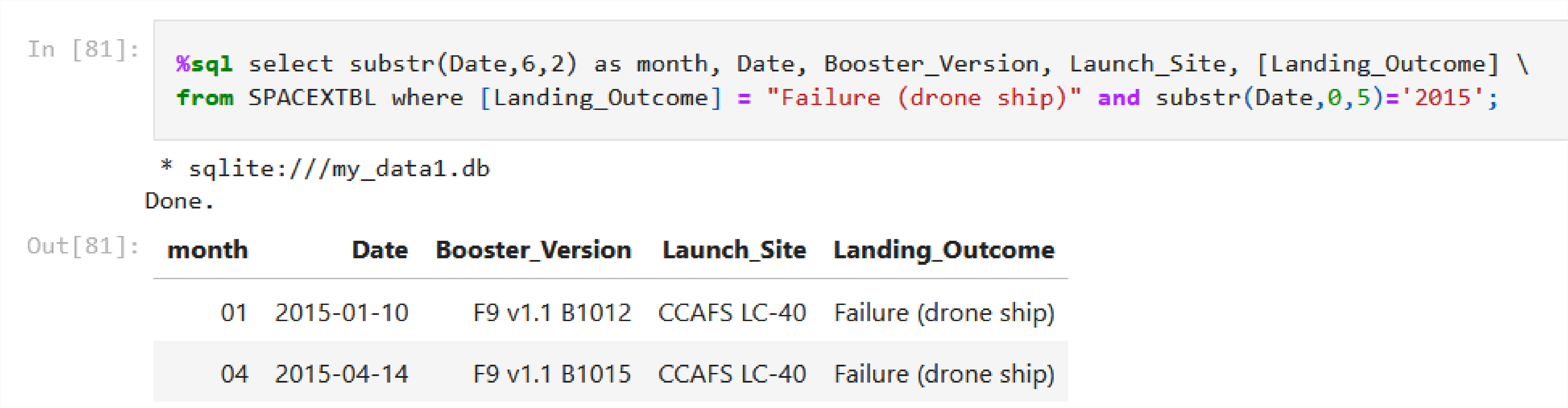
```
In [69]: %sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)

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

Out[69]:

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



In [81]:

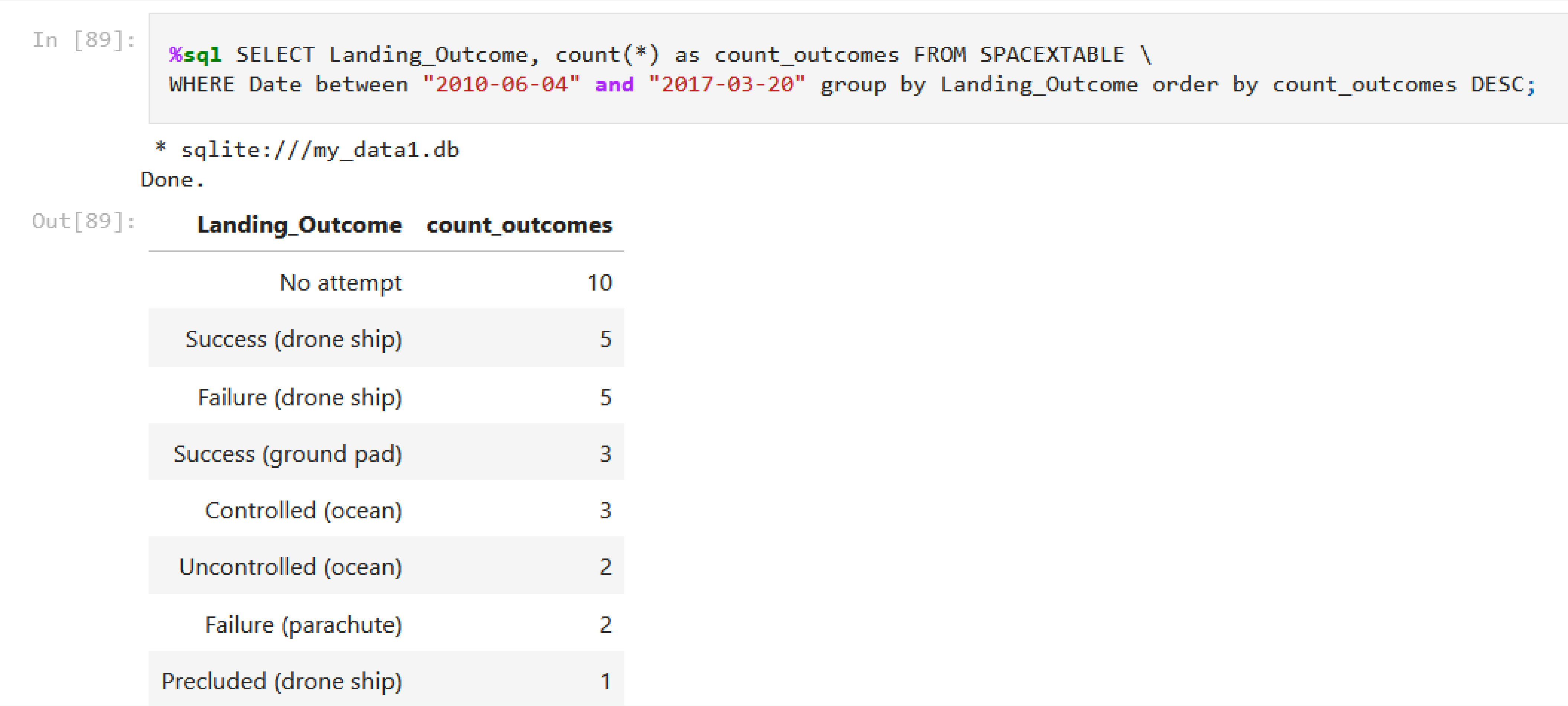
```
%sql select substr(Date,6,2) as month, Date, Booster_Version, Launch_Site, [Landing_Outcome] \
from SPACEXTBL where [Landing_Outcome] = "Failure (drone ship)" and substr(Date,0,5)='2015';
```

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

Out[81]:

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

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



In [89]:

```
%sql SELECT Landing_Outcome, count(*) as count_outcomes FROM SPACEXTABLE \
WHERE Date between "2010-06-04" and "2017-03-20" group by Landing_Outcome order by count_outcomes DESC;
```

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

Out[89]:

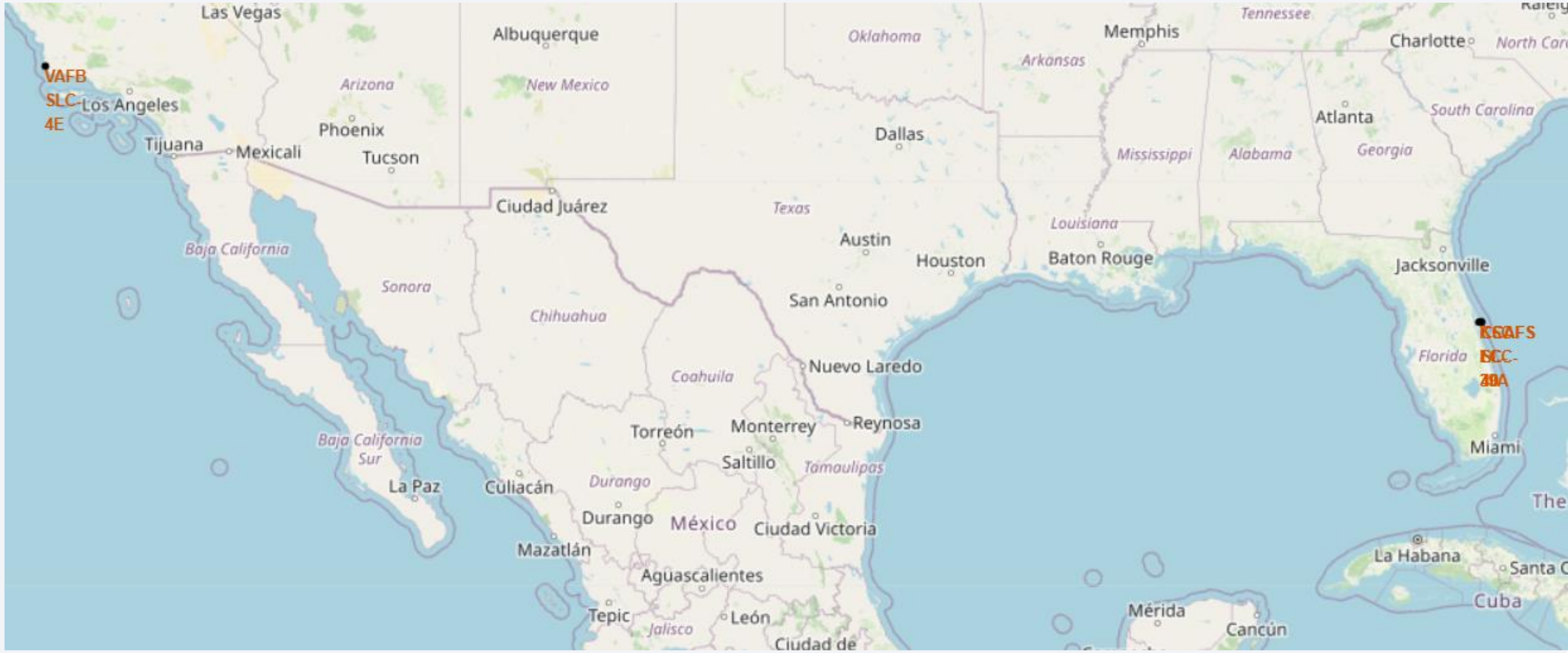
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

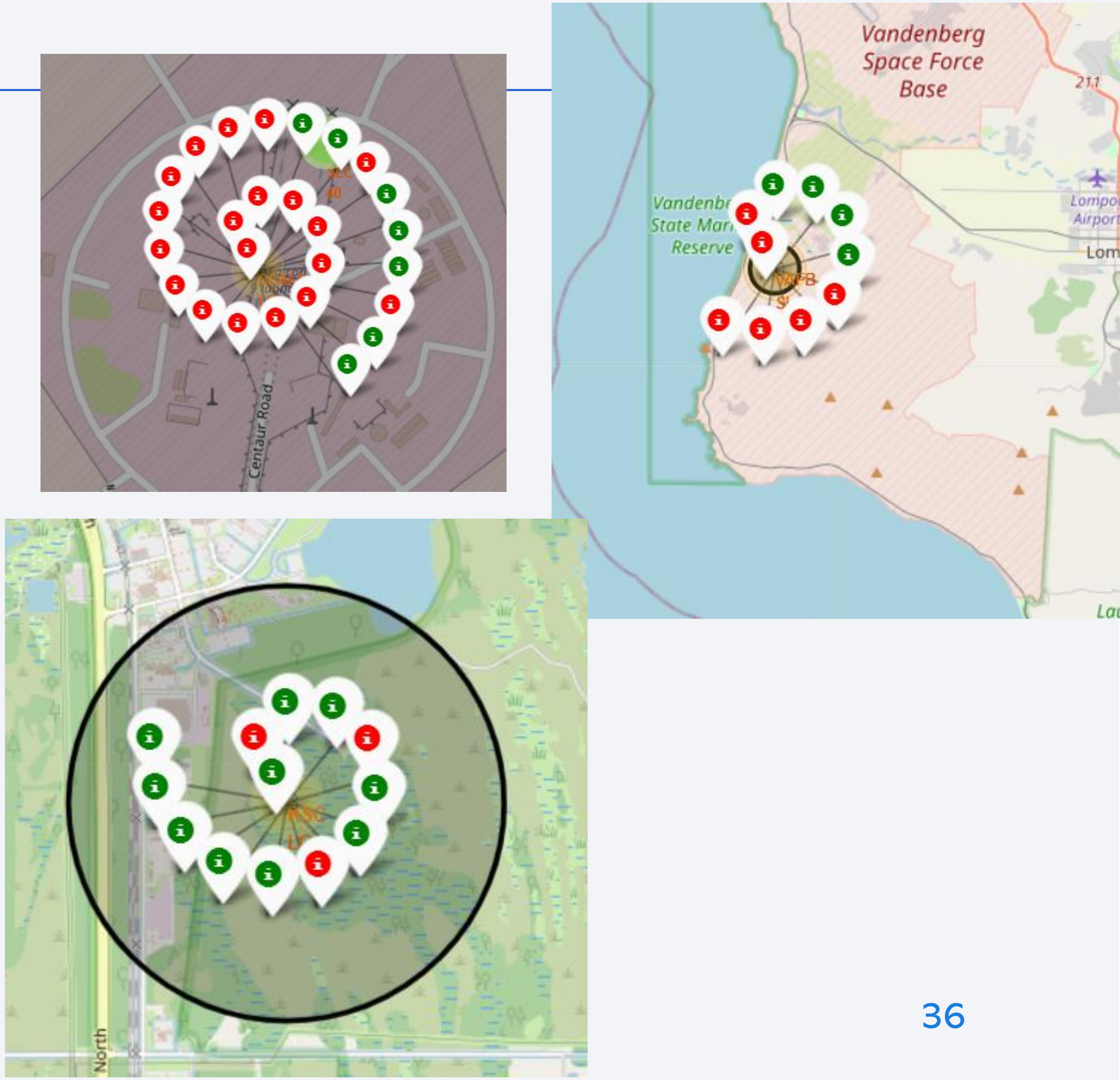
Launch site locations

- There are two distinct launch sites in USA, both located near the equator and both having ocean near it
- Two of them are located on East coast, and one – on West coast



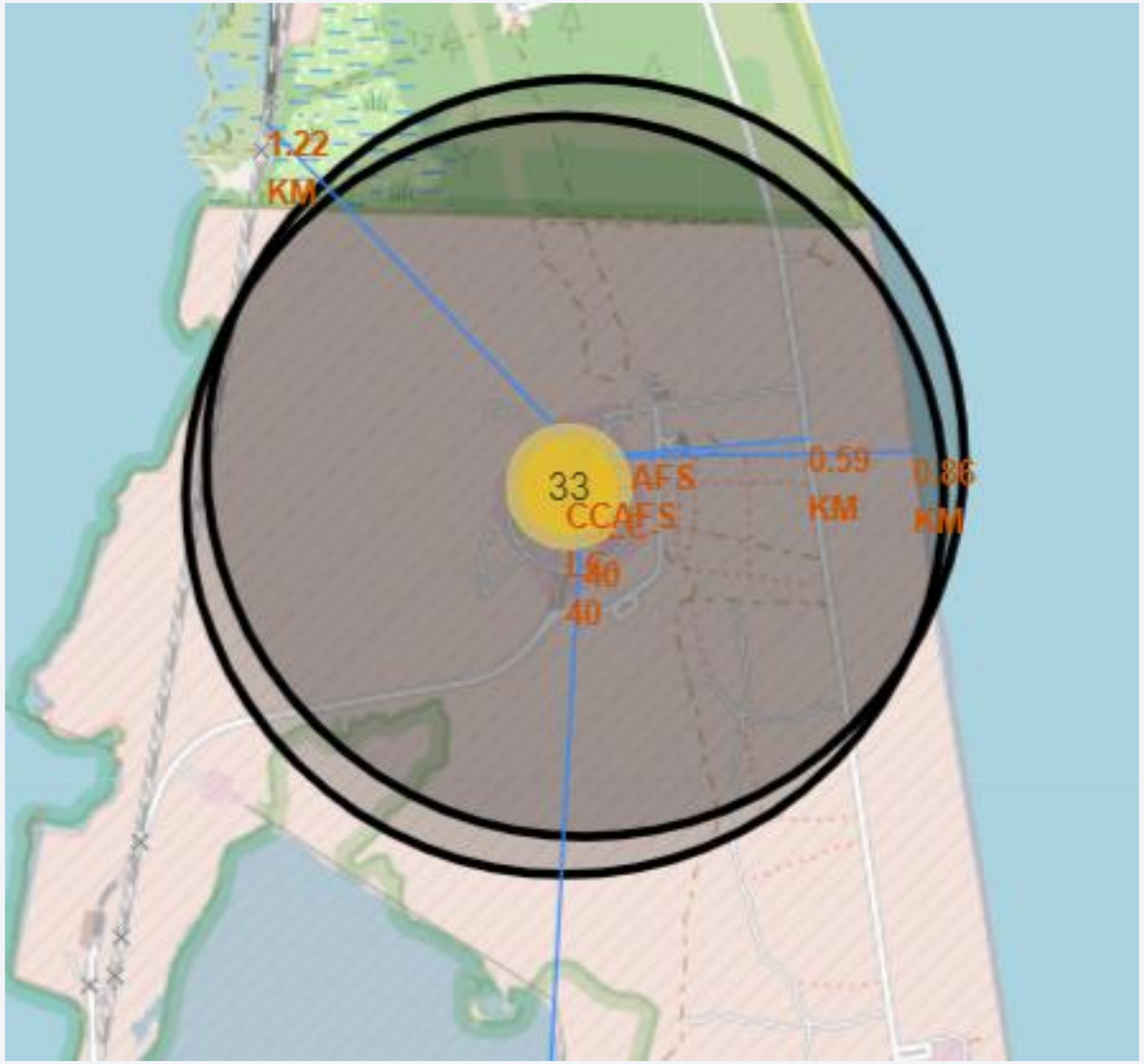
Launch success markers

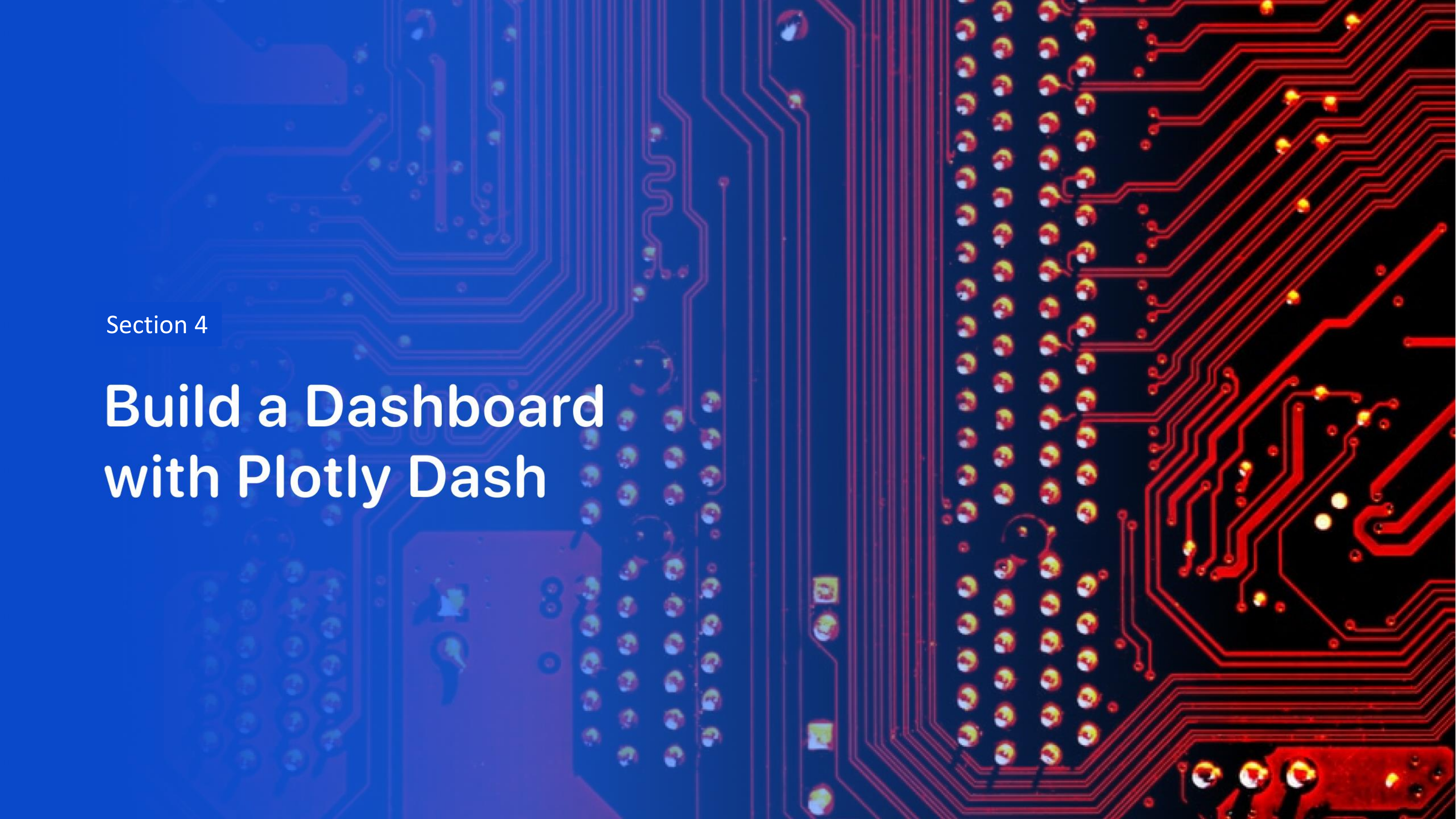
- Each launch site can be described by the number of launch and their successfulness
- Markers on the map can be convenient to describe how useful particular launch site can be
- Based on the data, we can say that KSC-LC 39A is the most successful launch site for Falcon 9



Visualizing distance between map objects

- Using Follium map, we can visualize how far are different objects from launch sites
- Launch sites are close to highways for human and supply transport.
- Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.



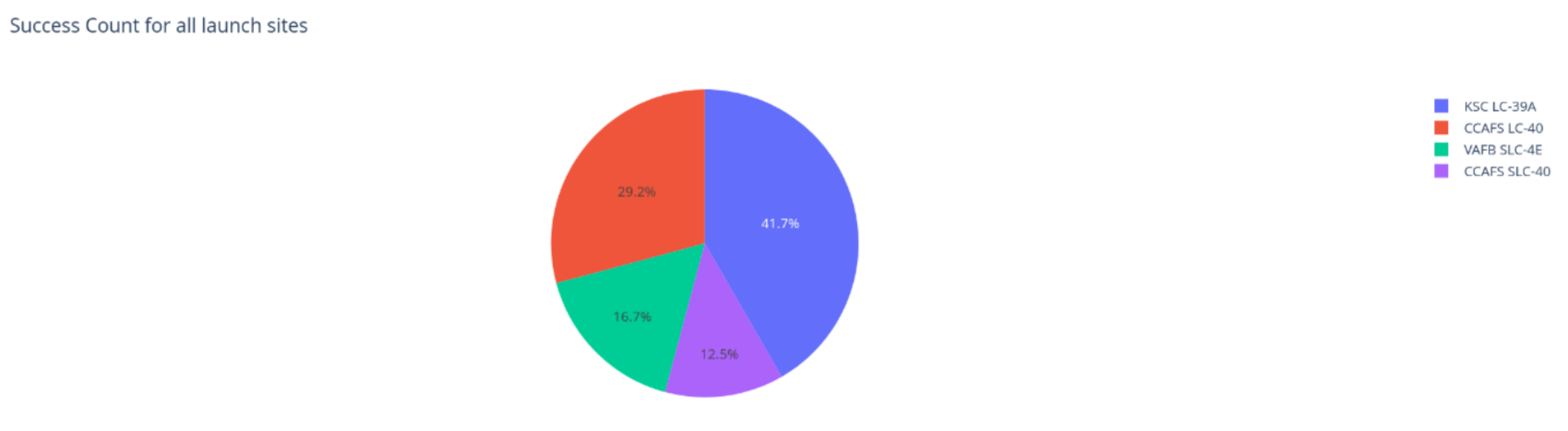


Section 4

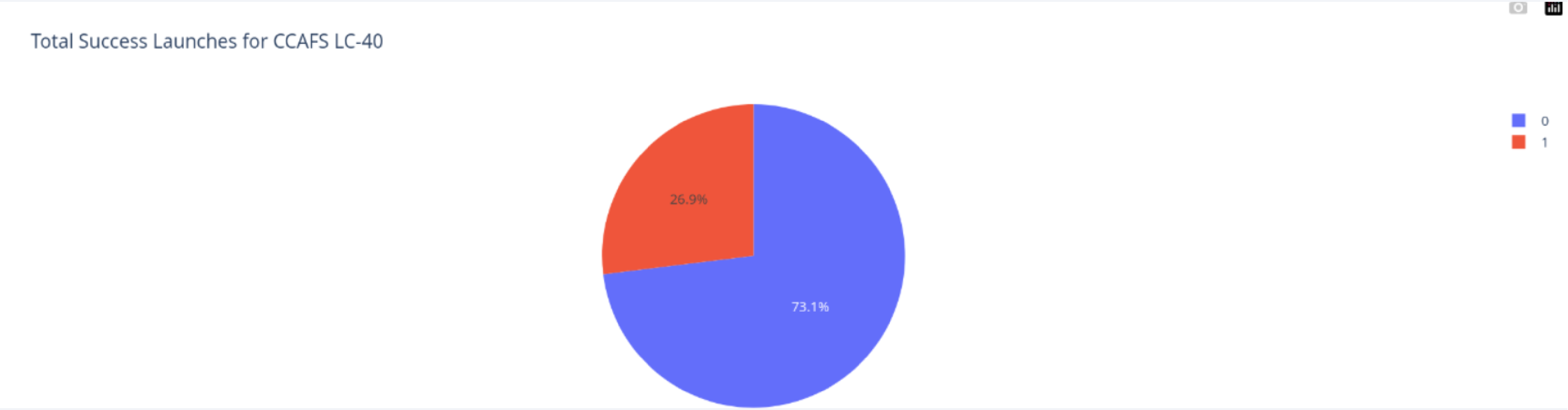
Build a Dashboard with Plotly Dash

Successful landing pie chart

- CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC LC 39-A have the same amount of successful landings, but a majority of the successful landings were achieved before the name change.
- VAFB SLC-4E has the smallest share of successful landings.

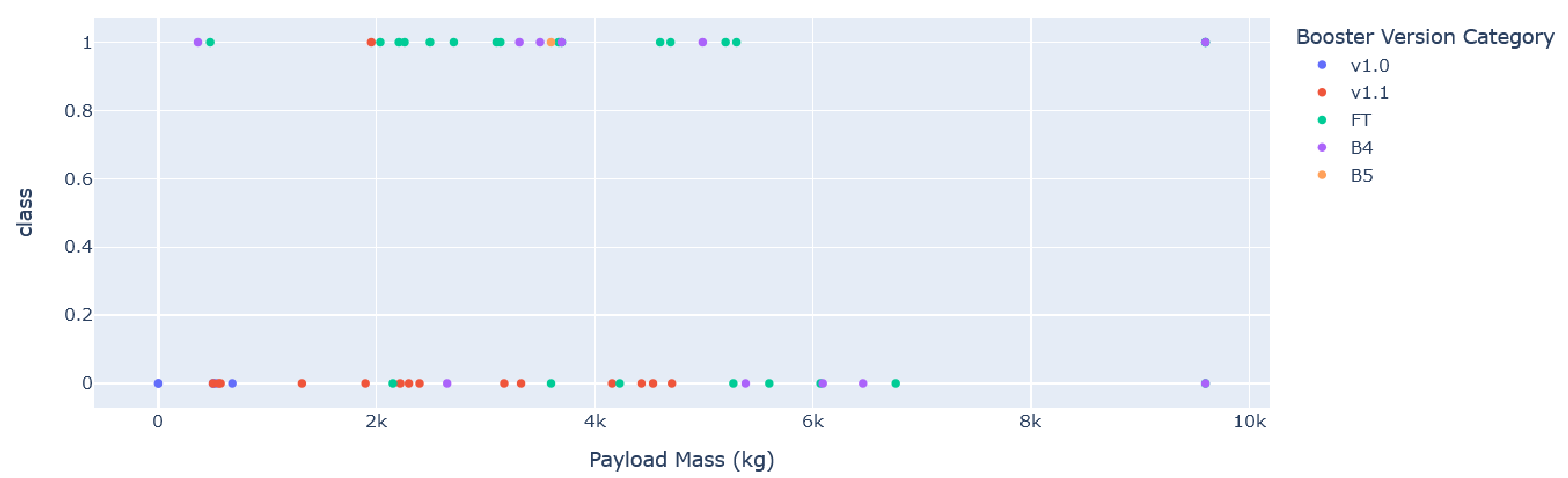


Successful landings pie chart for CCAFS LC-40



Payload Vs Launch Outcome

- Class indicates 1 for successful landing and 0 for unsuccessful landing
- Payloads under 6000 kgs and FT booster version combination has high success rates
- There is less data to estimate risk of launches over Payload Mass 7000 kgs.

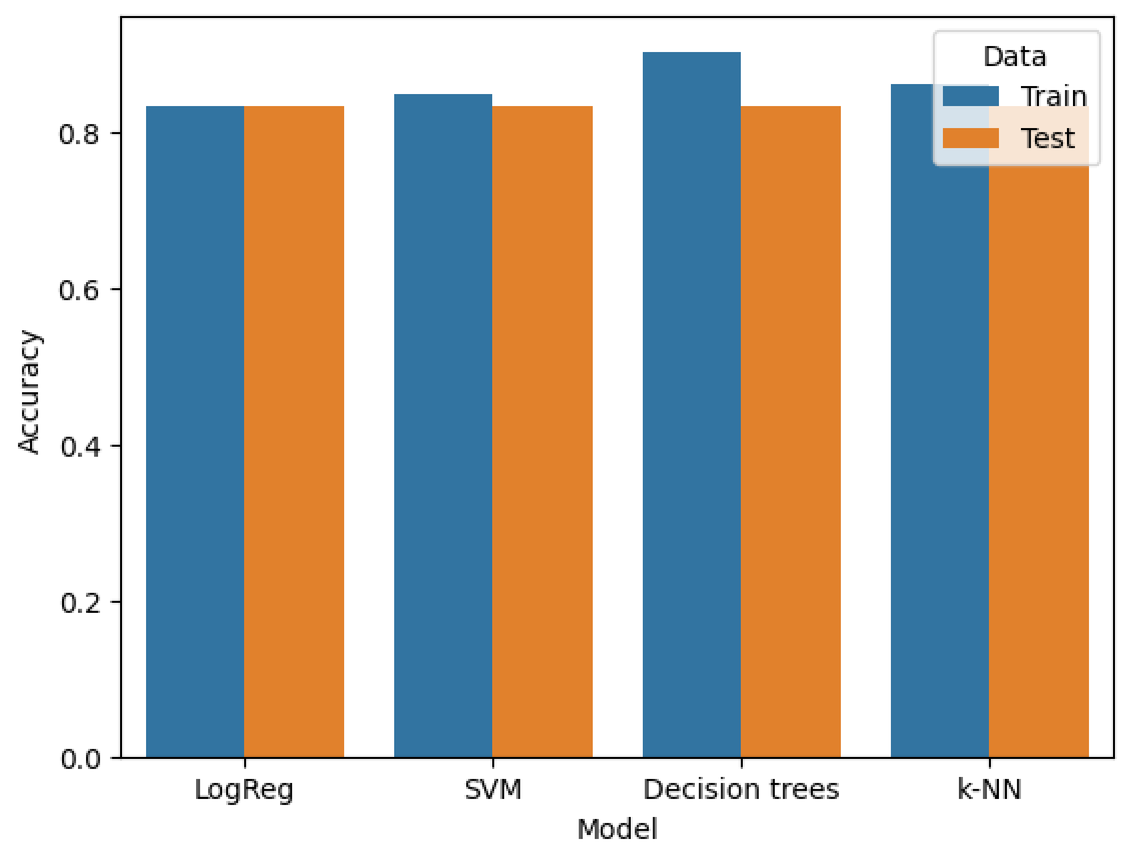


Section 5

Predictive Analysis (Classification)

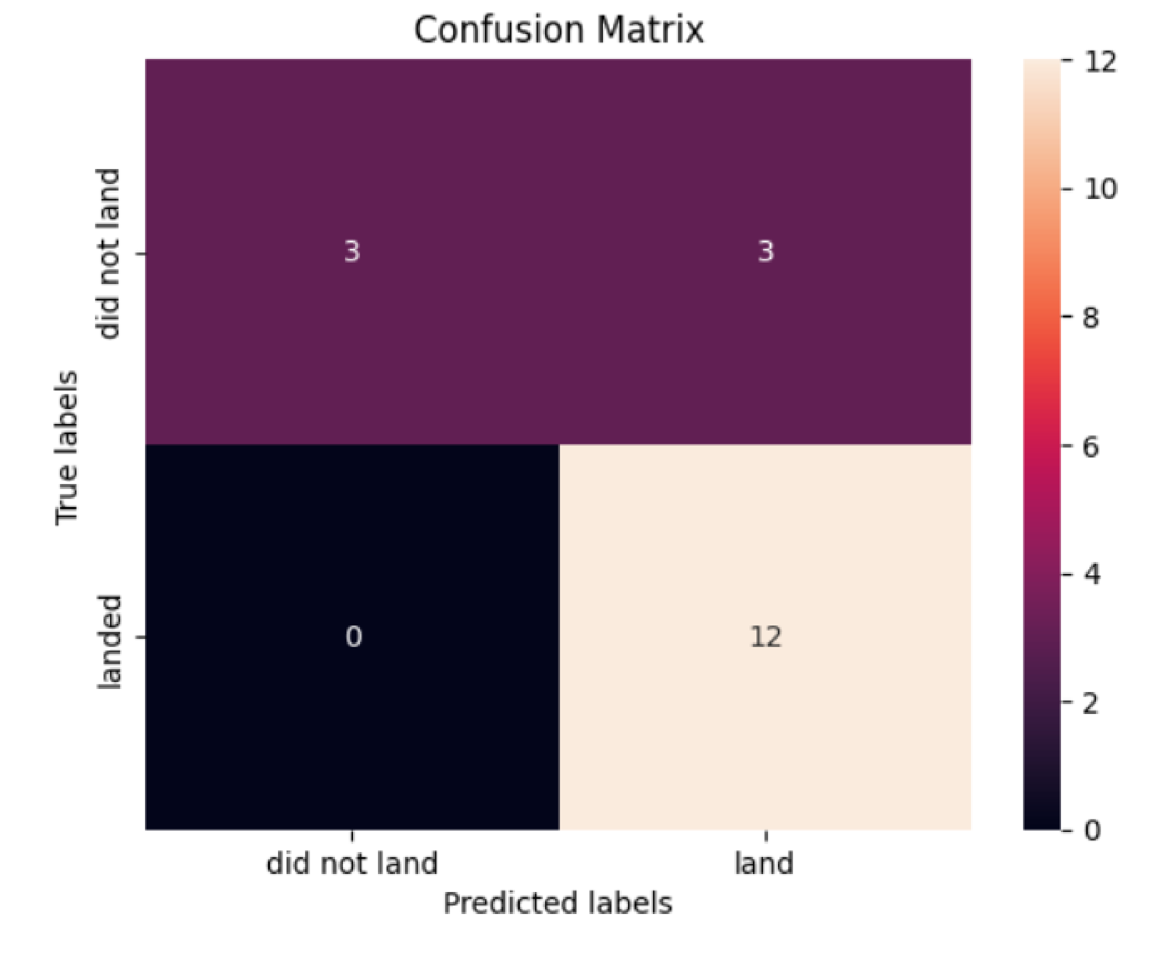
Classification Accuracy

- Four Models were tested: Logistic Regression, SVM, Decision Trees and KNN and their accuracy can be shown beside
- The best model which has highest classification accuracy is Decision Tree model with an accuracy of 88%
- All other models performed equally and have accuracy around 83%
- It should be noted that test size is small and is only contain 18 samples. This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.



Confusion Matrix

- The models predicted 12 successful landings when the true label was successful landing
- The models predicted 0 unsuccessful landings when the true label was landed
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landings
- The model predicted 3 successful landings when true label was unsuccessful landing
- Our models over predict successful landings.



Conclusions

- Problem: To develop a machine learning model for Space Y who wants to compete against SpaceX
- The goal of model is to predict whether Stage 1 will successfully land or fail to land
- Different data sources (API and Wiki Page) were analyzed. • The best launch site is KSC LC 39-A
- Launches above payloads 7000kg are more successful.
- Decision Tree Classifier can be used to predict successful landings and increase profits.
- However, If possible more data should be collected to determine the best machine learning model and improve accuracy.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!