IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Randy Myrick>
<May 20, 2023>

Presentation Topics

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

Executive Summary

- In order to accurately price our launches, we need the ability to predict if our first stage rockets will land successfully. This report presents the results of our analysis and efforts o model and predict successful first stage launches.
- We followed a rigorous data science-based approach to build a model to predict successful first stage landings. We collected information about Space X launch results and created an interactive dashboard for the team to utilize to understand the underlying data. Further, we trained a machine learning model using public data to predict if SpaceX launches successfully land their first stage. We believe this model can be utilized to make similar predictions of first stage launch success
- Our analysis and modelling followed a standard methodology which included: data collection, data preparation, exploratory data analysis, interactive visual analytics and dashboard development, and finally predictive analysis and modelling. These topics will be discussed more fully in the balance of the report.
- Results of the analysis suggest an ability to identify successful first stage launch success with a mid-80% accuracy. The creation and accuracy determination of modelling will be discussed in more detail below.

Introduction

- Project background and context
  - The commercial space age is here, companies are making space travel affordable for everyone. Examples of this new market include Virgin Galactic providing suborbital spaceflights, Rocket Lab is a small satellite provider, Blue Origin manufacturing sub-orbital and orbital reusable rockets, and the most successful is SpaceX. SpaceX’s accomplishments include sending spacecraft to the International Space Station. Starlink, a satellite internet constellation providing satellite Internet access, and sending manned missions to Space.
  - One reason SpaceX can do this is the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.
  - First stage rockets are quite large and expensive. Unlike other rocket providers, SpaceX's Falcon 9 can recover the first stage. Sometimes the first stage does not land. Other times, Space X will sacrifice the first stage due to the mission parameters like payload and orbit.
  - As a new rocket company. Space Y needs to compete with SpaceX. To do so successfully, we need the ability to determine the price of each launch.
  - SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Problems you want to find answers
  - Simply stated, can we predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - How data was collected
- Perform data wrangling
    - 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

- Data sets collected.
  - A key source of SpaceX launch data was gathered from SpaceX REST API.
  - This API will gave us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
  - Our goal was to use this data to predict whether SpaceX will attempt to land a rocket or not.
  - The SpaceX REST API endpoint utilized for this analysis was api.spacexdata.com/v4/launches/past.
  - Another source of data on Falcon 9 Launches are wiki pages. We used the Python BeautifulSoup package to web scrape HTML tables that contain valuable Falcon 9 launch records.
  - Additional detail on the data collection process is provided below including the API calls utilized.

Data Collection – SpaceX API

- Key data collection thrue SpaceX REST calls

- GitHub URL of the SpaceX API-based data collection.

```
#Request launch data from SpaceX
spacex_url=https://api.spacexdata.com/v4/launches/past
# Use json_normalize meethod to convert the json result into a dataframe
data = pd.json_normalize(response.json())
```

Data Collection - Scraping

- Web scraping process to collect SpaceX launch data from Wikipedia

- GitHub URL of SpaceX web scraping of Wikipedia pages for launch data

Link to Wikipedia page on SpaceX launches

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"

web_data  = requests.get(static_url).text

soup = BeautifulSoup(web_data,"html.parser")

html_tables=soup.find_all('table’)
```

Once tables identified, the data table was targeted and values extracted on a row-by-row basis.

See GitHub URL for details on extraction.

Data Wrangling

- Multiple processing (i.e., data wrangling) steps were taken to prepare the data for analysis, including:
  - Replace missing data where appropriate
    - # Calculate the mean value of PayloadMass column
    - mean_payload = data_falcon9['PayloadMass'].mean()
    - # Replace the np.nan values with its mean value
    - data_falcon9['PayloadMass'].replace(np.nan, mean_payload, inplace=True)
  - Convert categorical first stage landing data into numeric
    - bad_outcomes=set(landing_outcomes.keys()[[1,3,5,6,7]])
    - landing_class = ~df['Outcome'].isin(bad_outcomes)
    - landing_class.replace({True: 1, False: 0}, inplace=True)
  - Multiple exploratory data analysis performed to understand data and look for patterns
    - Launches per site: df['LaunchSite'].value_counts()
    - Orbits targeted: df['Orbit'].value_counts()
    - Outcome frequency: landing_outcomes = df['Outcome'].value_counts()
- Additional detail can be found in the full data wrangling notebook

EDA with Data Visualization

- As part of the early data process an number of data visualizations were created to better understand the data and identify potential relationships
- The following visualizations were created:
  - Plot of payload mass versus flight number with first stage launch outcome identified
  - Plot of launch site versus flight number with first stage launch outcome identified
  - Plot of payload mass versus launch site with first stage launch outcome identified
  - Bar chart of first stage launch success for each orbit target
  - Plot of flight number versus orbit with first stage launch outcome identified
  - Plot of payload mass versus orbit with first stage launch outcome identified
  - Line graph of first stage launch success versus year
- GitHub URL for the full early exploration and visualization of the data

EDA with SQL

- Summary of SQL early data analysis queries

1. List SpaceX Launch Sites:
    1. v%sql select distinct "Launch_Site" from SPACEXTBL
2. Show records for CCA Launch Sites:
    1. %sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
3. Payload size launched by NASA:
    1. %sql select sum(PAYLOAD_MASS__KG_) FROM SPACEXTBL where "Customer" = 'NASA (CRS)’
4. Average payload for booster version F9 v1.1:
    1. %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version ='F9 v1.1’
5. Date of first successful first stage landing:
    1. %sql select min(Date) from SPACEXTBL where "Landing _Outcome" = 'Success (ground pad)’
6. Boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000:
    1. %sql select distinct Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ between 4000 and 6000 and "Landing _Outcome"='Success (drone ship)’
7. Summary of mission outcomes by type:
    1. %sql select "Mission_Outcome", count(*) from SPACEXTBL group by "Mission_Outcome“
8. Names of the booster_versions which have carried the maximum payload mass:
    1. %sql select distinct(Booster_Version) from SPACEXTBL where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS__KG_") from SPACEXTBL)
9. Landing failures in 2015 by month:
    1. %sql select substr(Date,4,2) as Month, substr(Date,7,4) as Year, Booster_Version, Launch_Site, "Landing _Outcome" from SPACEXTBL where Year = '2015' and "Landing _Outcome" like '%Failure%’
10. Count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017:
    1. %sql select "Landing _Outcome", count(*) as Count from SPACEXTBL where "Landing _Outcome" like '%Success%' and substr(Date,7,4)||substr(Date,4,2)||substr(Date,1,2) between '20100406' and '20170320’ group by "Landing _Outcome" order by Count desc

- GitHub URL for EDA with SQL notebook

Build an Interactive Map with Folium

Summary of map objects created and added to folium map analysis

Map Object	Rationale for Object
folium.Circle & Marker	Identify SpaceX lauch site locations
Marker Cluster	Manage multiple launch event markers at each site
Marker	For each launch event colored green for success and red for failure. Visualize launch performance.
MousePosition	Display to user the coordinates when mouse hovers over a location
Marker	Locate nearest water from launch site
Polyline	Display line between launch site and water marker

GitHub URL for completed interactive Folium-based map notebook.

Build a Dashboard with Plotly Dash

Summary of plots/graphs and within interactive dashboard created using Plotly. Dashbard provides interactive method to investigate SpaceX data and visualize relationships.

Plotly Object	Object Rationale
Dropdown	Allow selection of all launch sites of a single launch site. Selection makes corresponding change to dashboard graphs.
Range Slider	Specify range of payloads to be included in scatter plot. Changes dynamically make corresponding change to scatter plot.
Pie Chart	Display launch success by launch site. Show success/failure split if a single launch site is selected
Scatter chart	Displays correlation between payload mass and launch success for all sites or an individual one.

GitHub URL for Plotly Dash code and screenshot.

Predictive Analysis (Classification)

The following provides a summary of the steps to build and evaluate a variety of machine learning classification models to predict successful recovery of the first stage of the Falcon 9 rocket.

Step	Activity
Preprocessing	Convert categorical data value to numerical and normalize data to facilitate machine learning analysis. Utilized sklearn standard scalar.
Train/Test Split	Randomly split data into train and test buckets. Split data 80% for training and 20% for testing via sklearn train_test_split function.
Train Algorithms	Trained 4 machine learning classification algorithms. Utilized sklearn grid search to identify optimum hyper parameters for each algorithm. Genereted best score for each.
Test Algorithms	Determined algorithm accuracy against test data.
Confusion Matrix	Created confusion matrix based on test data for each optimized algorithm depicting number of correct and incorrect predictions versus actual outcomes.

GitHub URL for completed predictive analysis notebook

Results

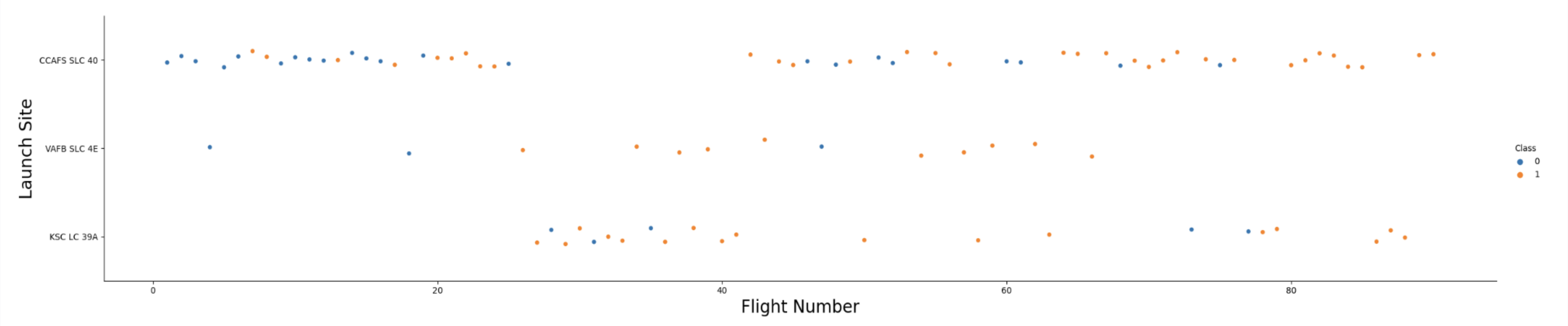
- Exploratory data analysis results
  - As flight number increases first stage recovery is more likely. The more massive the payload the less likely the first stage will land successfuly.
  - Success rates are dependent on launch site. CCAFS LC-40 has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
  - Success improves as number of flights increases.
  - Orbit target is related to successful stage 1 recovery
  - Success rate steadily improving since 2013
- Interactive analytics demo in screenshots
  - CCAFS LC-40 site has most launches (26) of 4 sites
  - CCAFS LC-40 and CCAFS SLC-40 are located very close together
  - KSC LC-39A has high success rate with 10 of 13 stage 1 rockets being recovered
  - VAFB SLC-4e has low success rate with only 4 of 10 stage 1 rockets being recovered
- Predictive analysis results
  - All 4 classification algorithms yielded meaningful accuracy scores in the mid 80% range
  - All 4 classification algorithms were better at predicting successful landings than failed landings
  - The Tree classification algorithm was slightly better at predicting failed landings and as a result was identified as the best predictor of the 4 algorithms.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

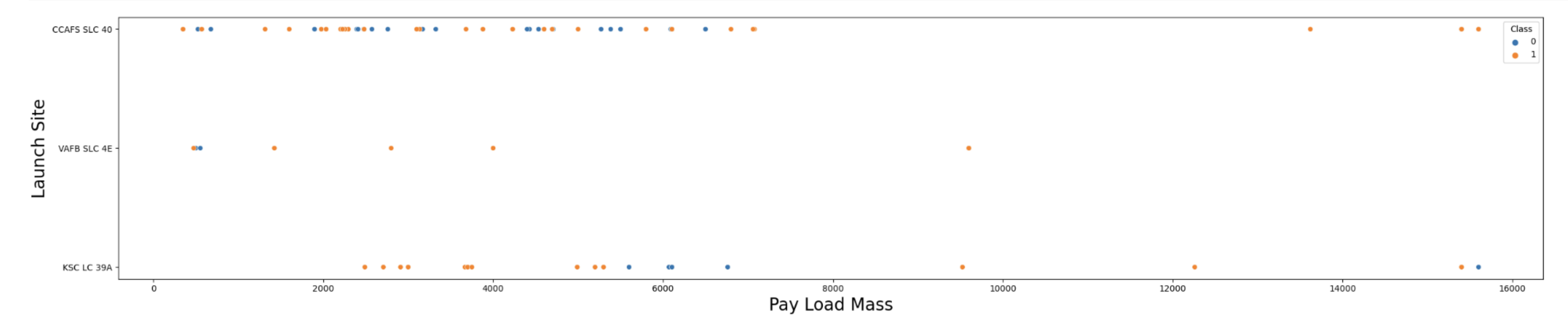
Scatter plot of Flight Number vs. Launch Site



Points are colored red for successful landing and blue for failure. Plot indicates higher level of launches and higher failure rates. Shows relatively high success rate of VAFB site. Also demonstrates improved success rate in the latest launches.

Payload vs. Launch Site

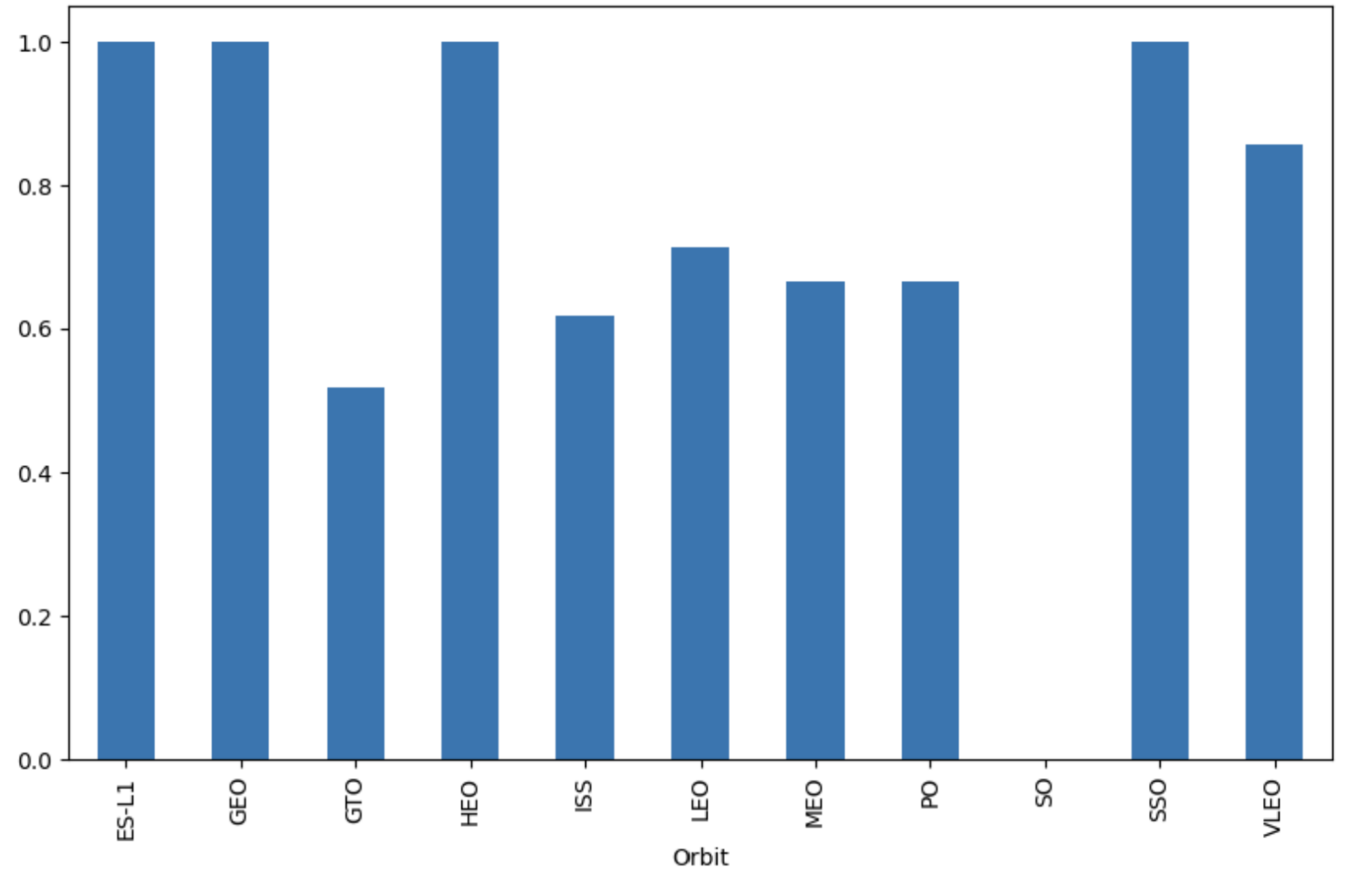
Scatter plot of Payload vs. Launch Site



Points are colored red for successful landing and blue for failure. At KSC site, higher payloads have been less successful. Overall, higher payloads result in lower success.

Success Rate vs. Orbit Type

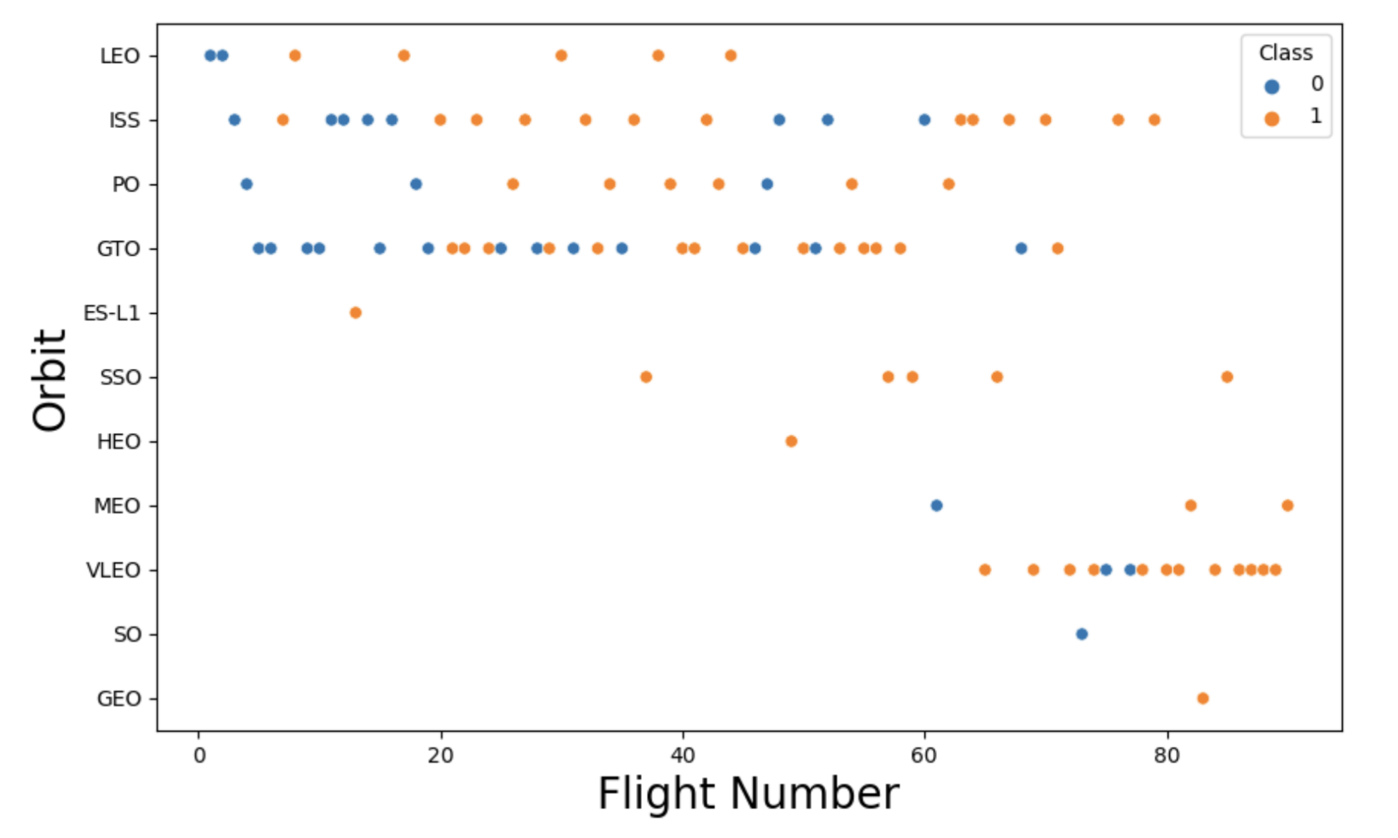
Bar chart for the success rate of each orbit type



Y-axis is mean success rate with 1 representing success and 0 representing failure. Each bar shows the mean success rate for all launches to the orbit indicated on the x-axis. Some orbits have a higher success rate than others.

Flight Number vs. Orbit Type

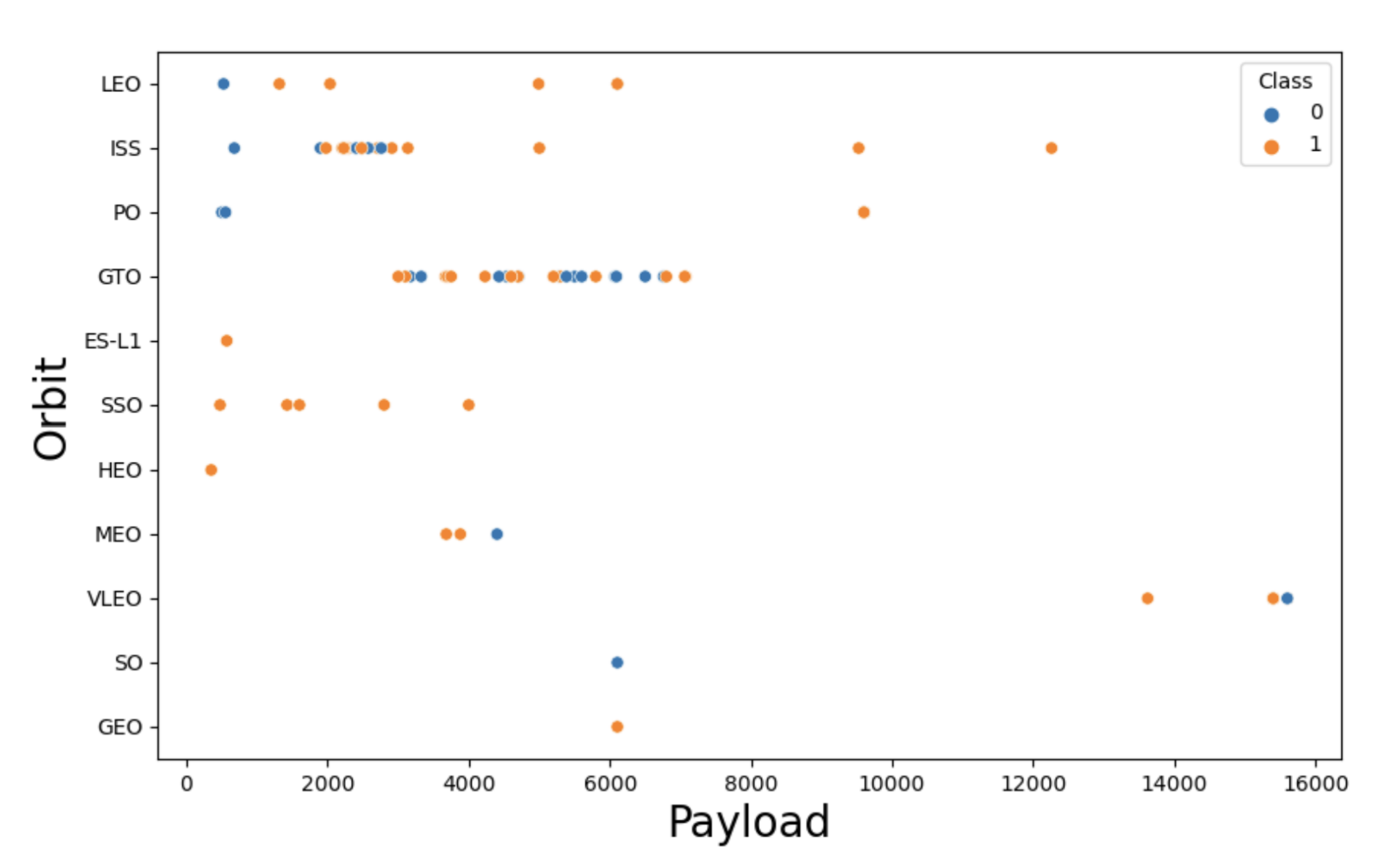
Scatter point of Flight number vs. Orbit type



With the red point representing successful landing, its evident in this plot that success increases as repeated launches occur. This pattern appears to hold true for all orbits targeted.

Payload vs. Orbit Type

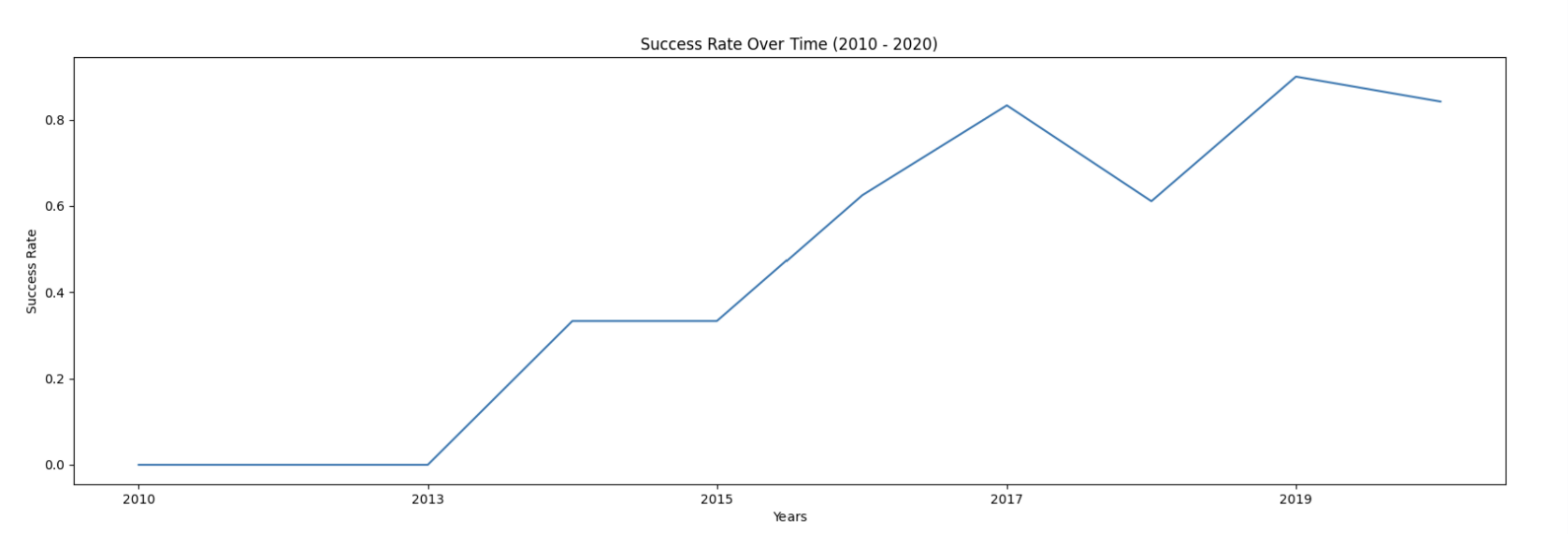
Scatter point of payload vs. orbit type



With the red point representing successful landing, its evident in this plot that orbit and payload are related for most orbits. The GTO orbit looks to be less predictable.

Launch Success Yearly Trend

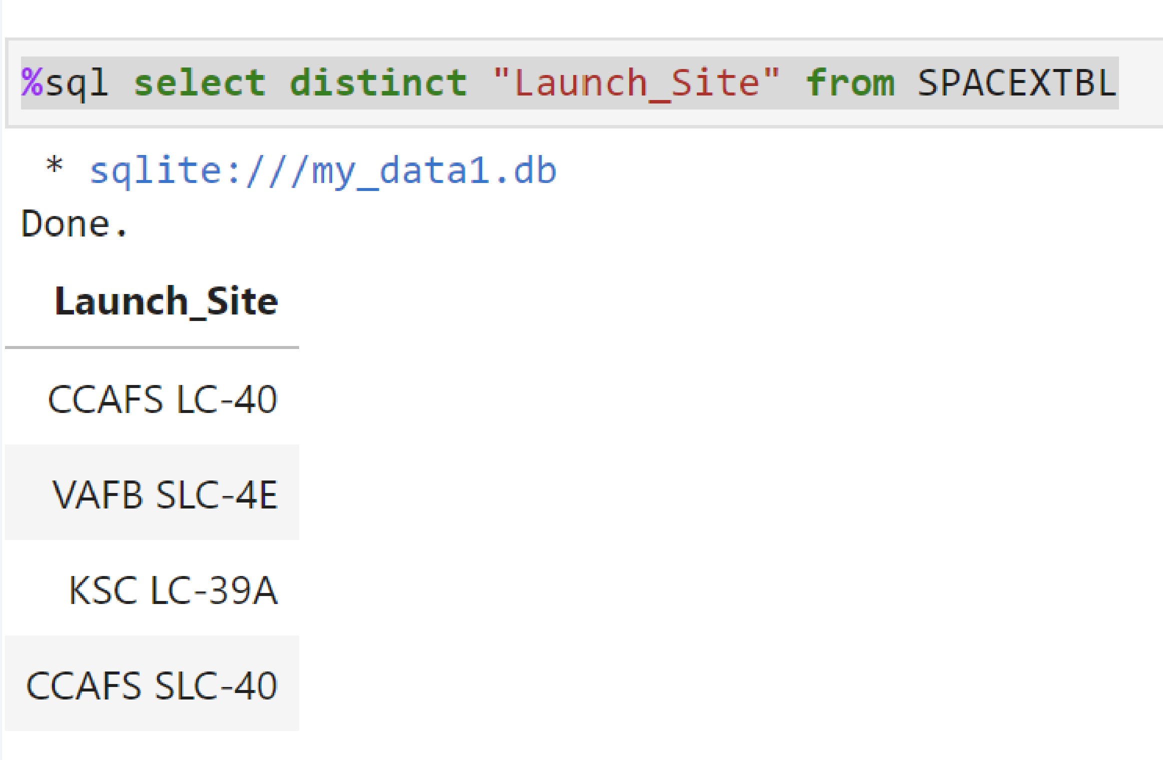
Line chart of yearly average success rate



Steady improvement since 2013. Currently ~80% of first stage rockets are landing successfully.

All Launch Site Names

Find the names of the unique launch sites



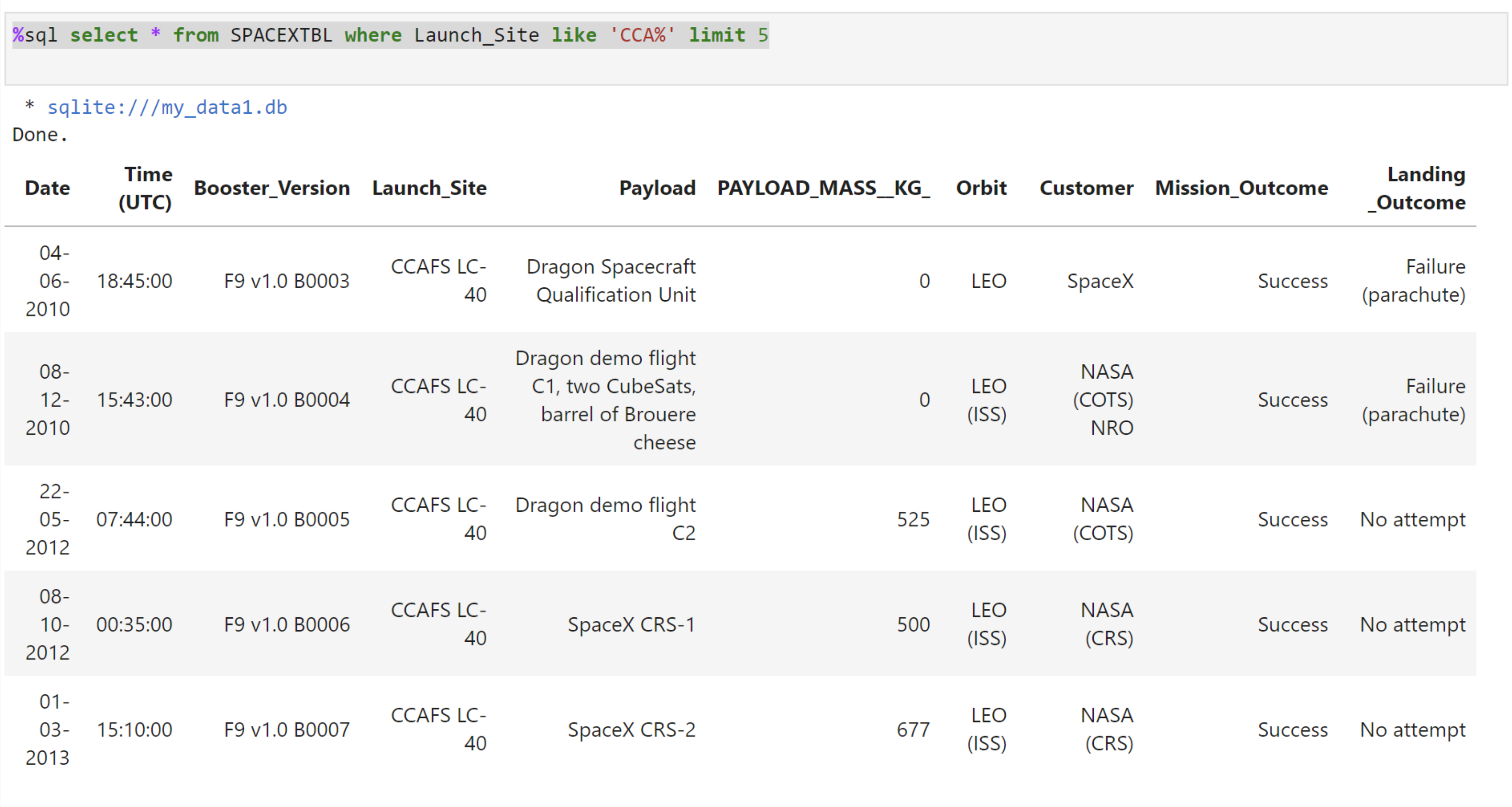
```
%sql select distinct "Launch_Site" from SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

Find 5 records where launch sites begin with `CCA`



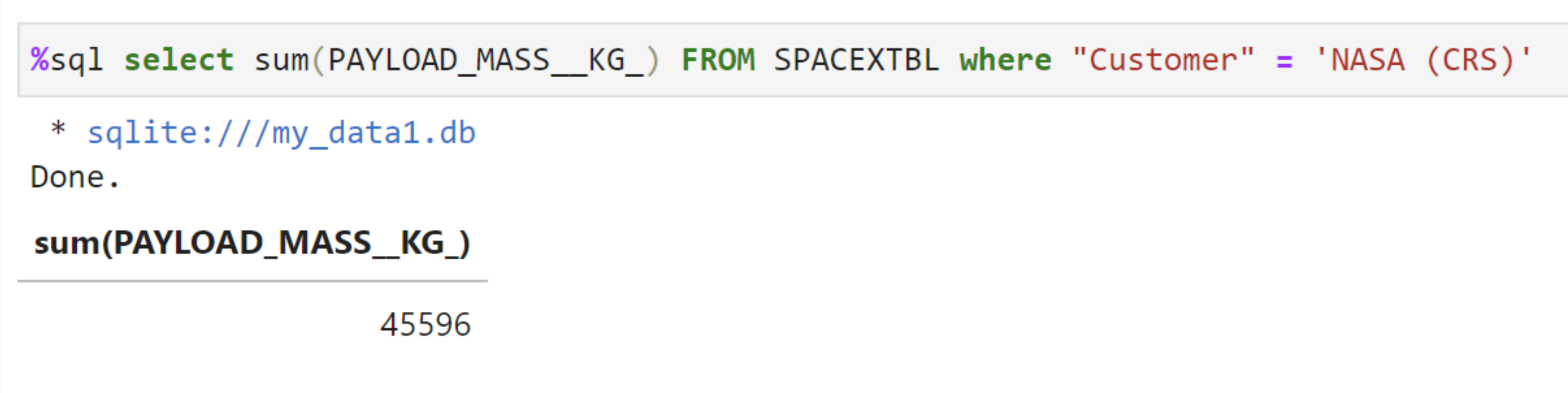
```
%sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Calculate the total payload carried by boosters from NASA



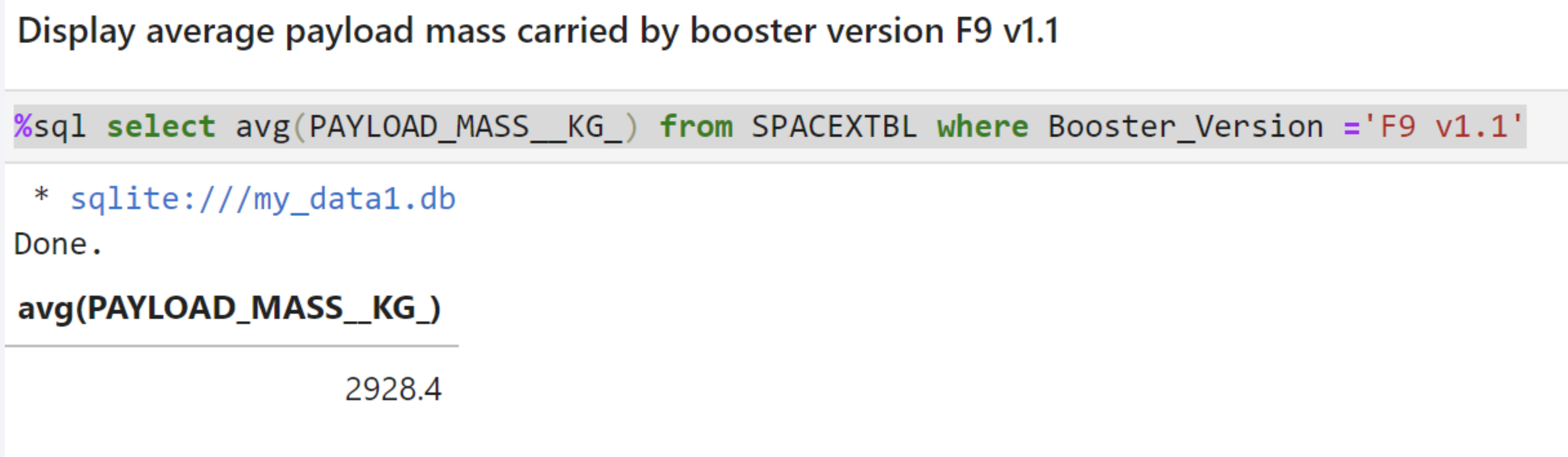
```
%sql select sum(PAYLOAD_MASS__KG_) FROM SPACEXTBL where "Customer" = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Calculate the average payload mass carried by booster version F9 v1.1



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

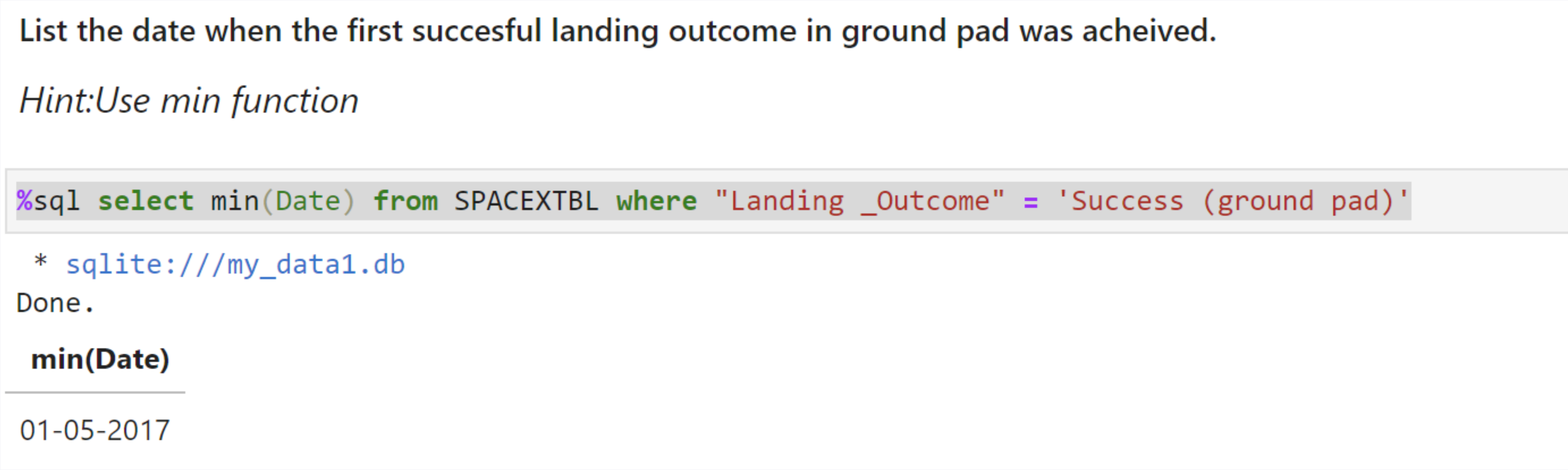
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version ='F9 v1.1'
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

Find the dates of the first successful landing outcome on ground pad



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

Hint:Use min function

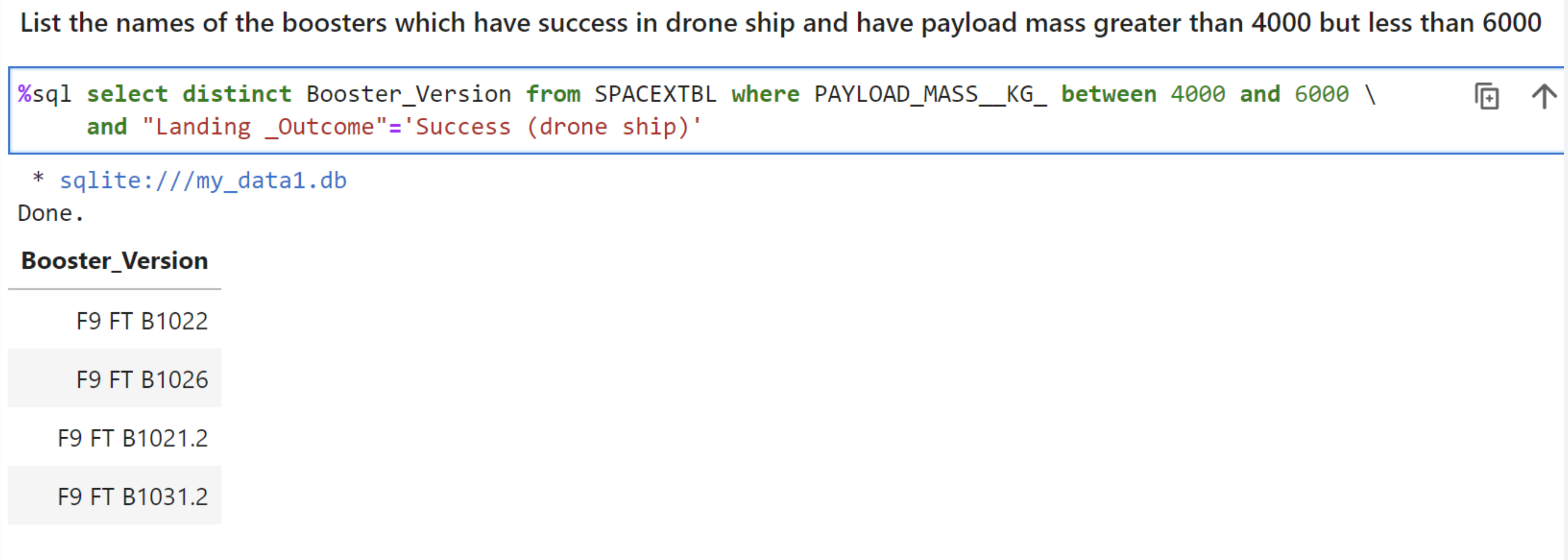
```
%sql select min(Date) from SPACEXTBL where "Landing _Outcome" = 'Success (ground pad)'
```

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

min(Date)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

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



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

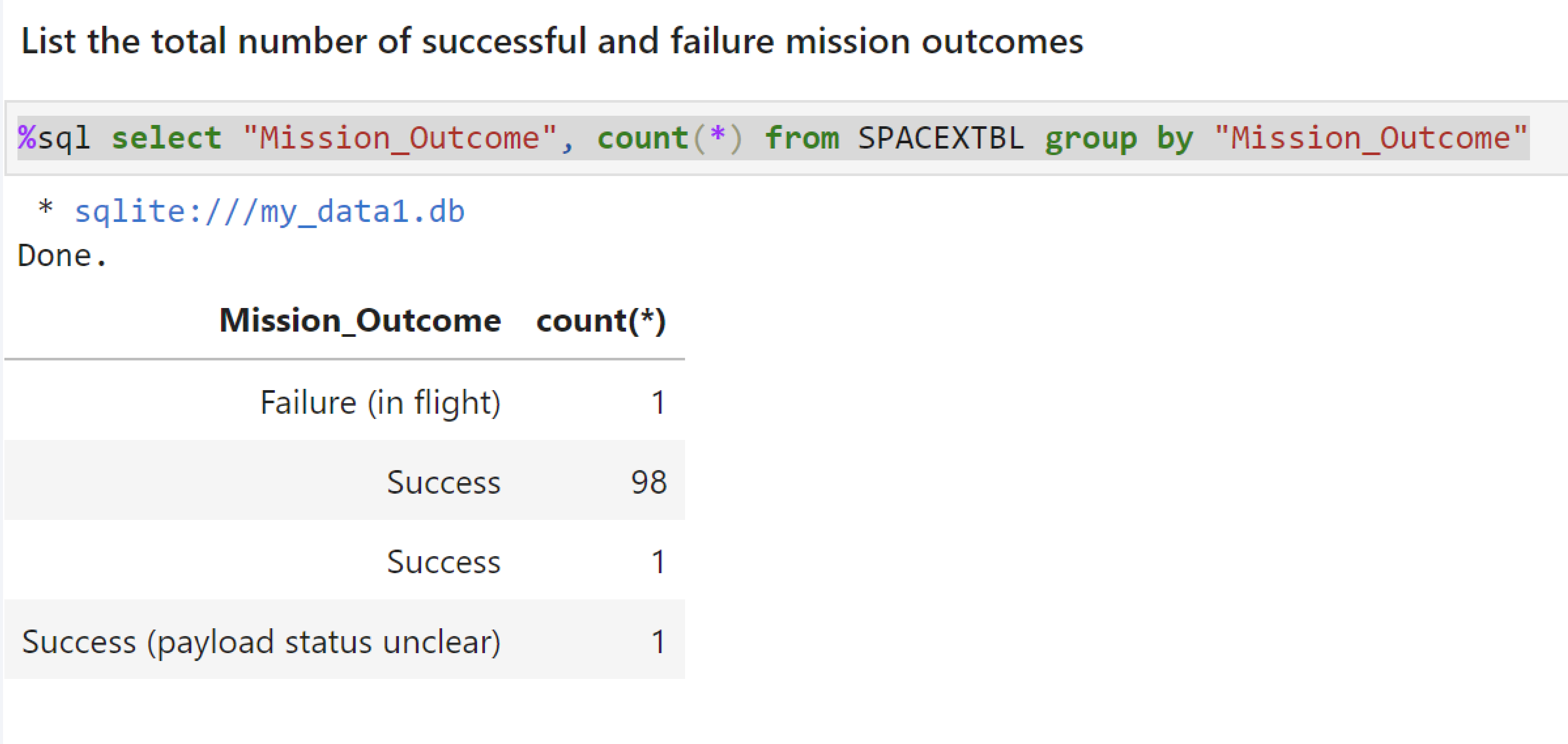
```
%sql select distinct Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ between 4000 and 6000 \
    and "Landing _Outcome"='Success (drone ship)'
```

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

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

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes



List the total number of successful and failure mission outcomes

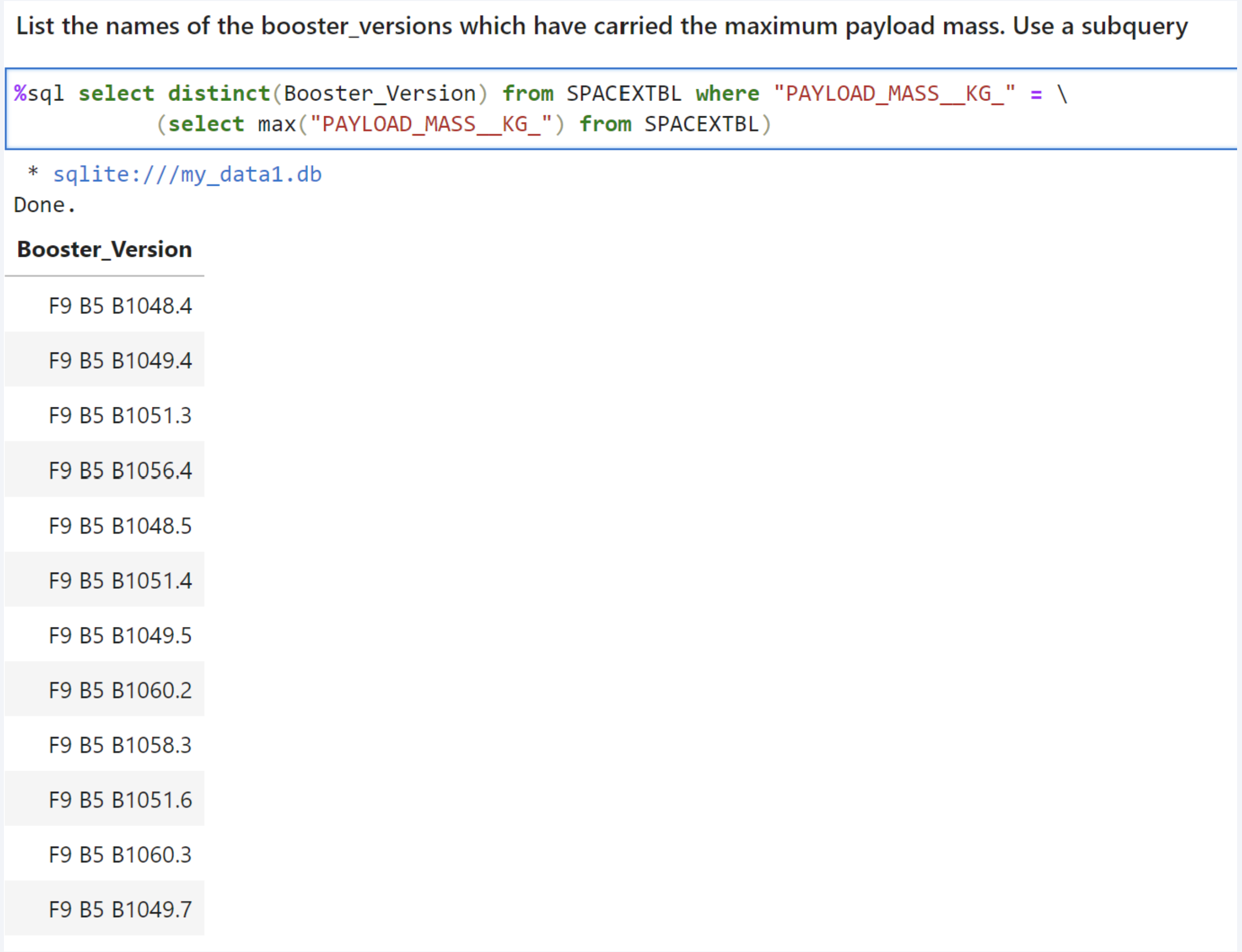
```
%sql select "Mission_Outcome", count(*) from SPACEXTBL group by "Mission_Outcome"
```

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

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

Boosters Carried Maximum Payload

List the names of the booster which have carried the maximum payload mass



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

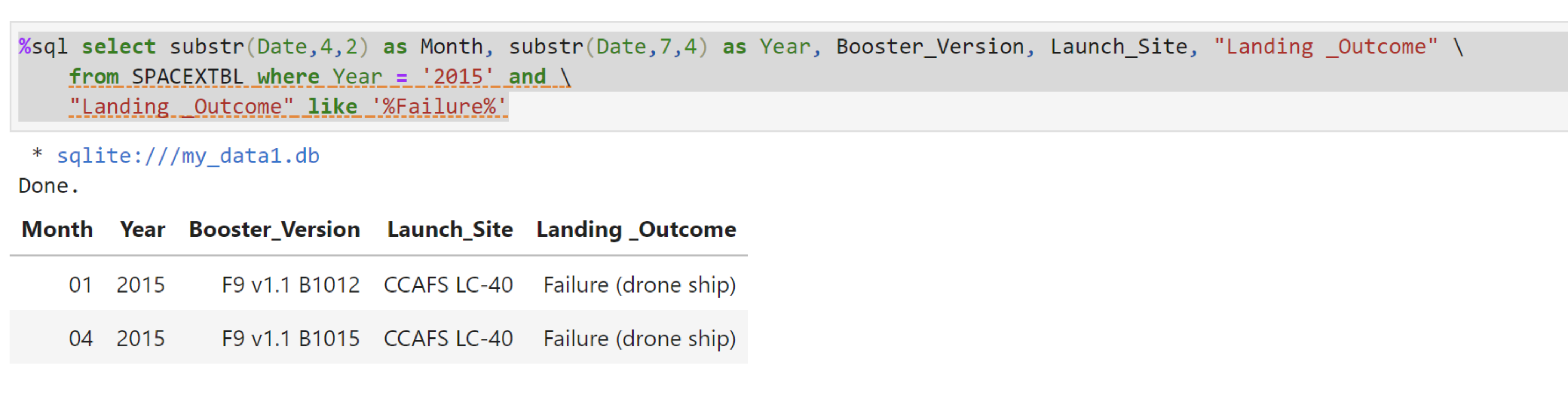
```
%sql select distinct(Booster_Version) from SPACEXTBL where "PAYLOAD_MASS__KG_" = \
          (select max("PAYLOAD_MASS__KG_") from SPACEXTBL)
```

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

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

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015



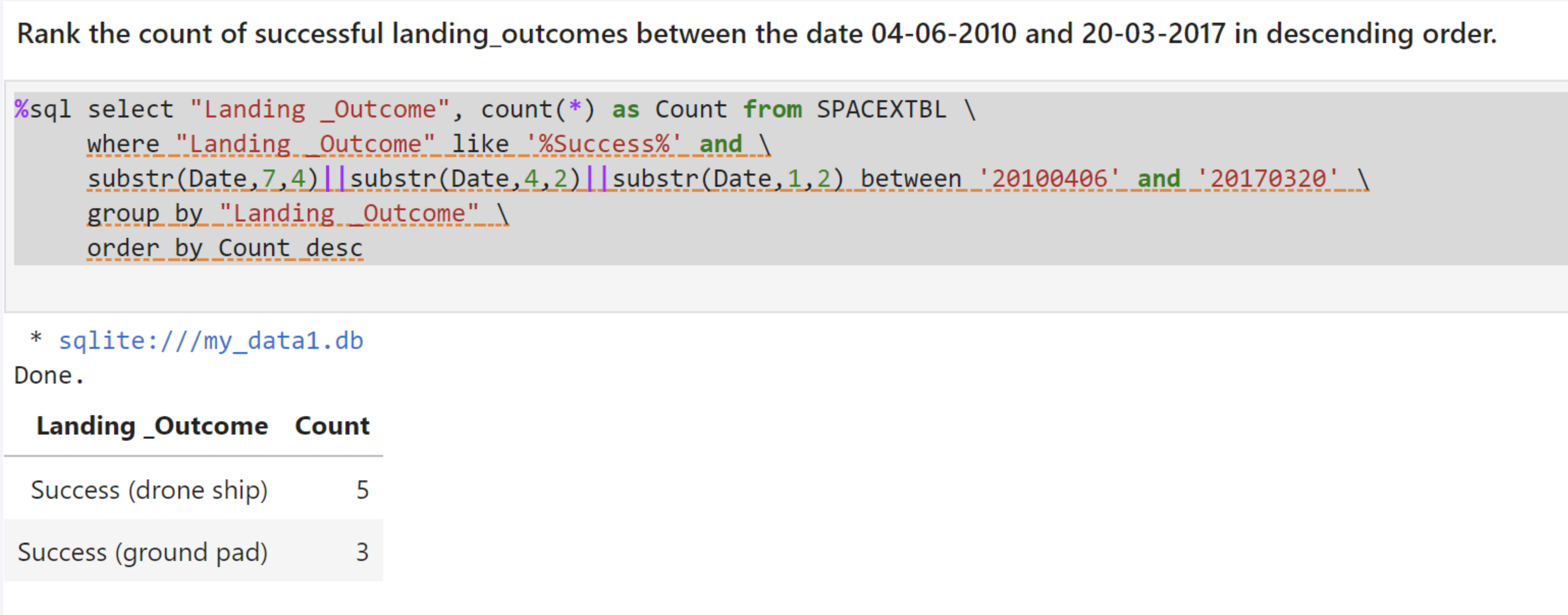
```
%sql select substr(Date,4,2) as Month, substr(Date,7,4) as Year, Booster_Version, Launch_Site, "Landing _Outcome" \
    from SPACEXTBL where Year = '2015' and \
    "Landing _Outcome" like '%Failure%'
```

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

Month	Year	Booster_Version	Launch_Site	Landing _Outcome
01	2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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



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 Count from SPACEXTBL \
     where "Landing _Outcome" like '%Success%' and \
     substr(Date,7,4)||substr(Date,4,2)||substr(Date,1,2) between '20100406' and '20170320' \
     group by "Landing _Outcome" \
     order by Count desc
```

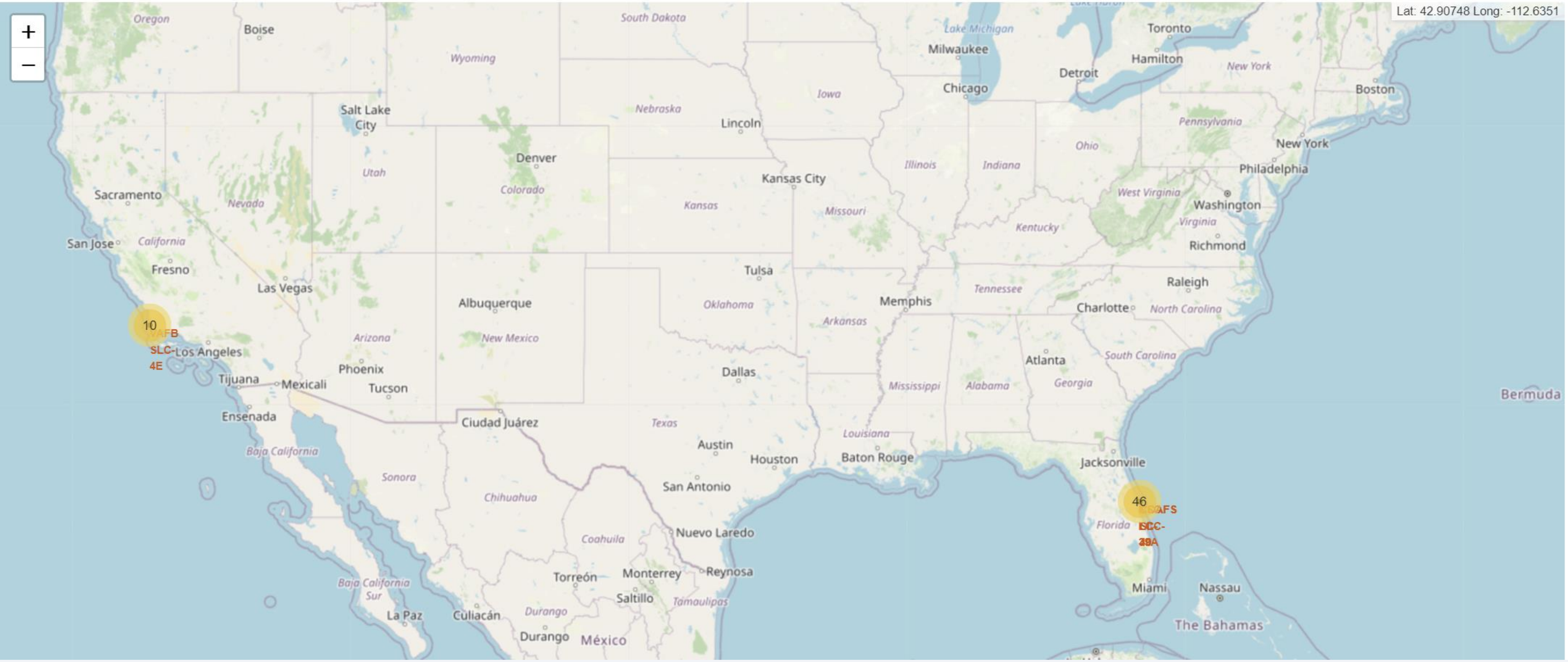
```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	Count
Success (drone ship)	5
Success (ground pad)	3

Section 3

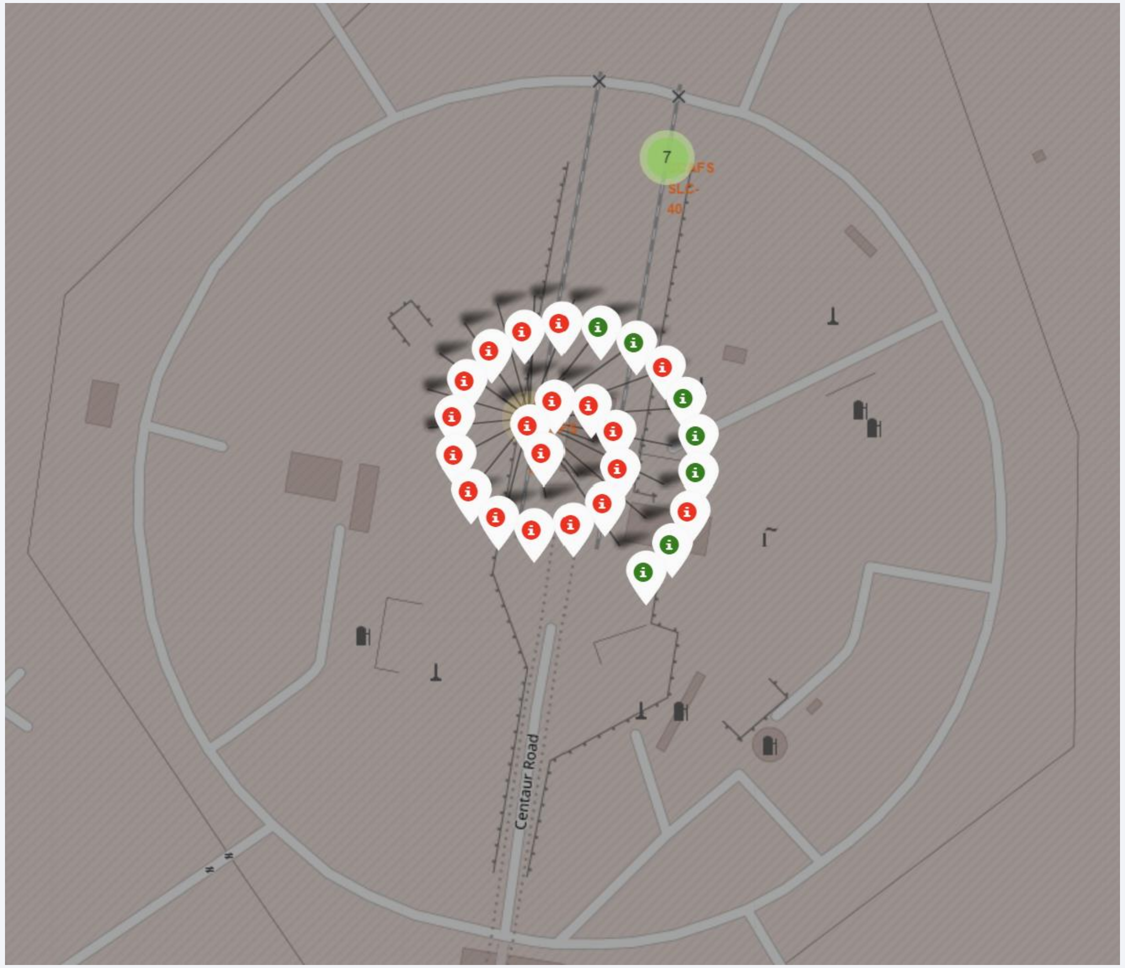
Launch Sites Proximities Analysis

SpaceX Launch Sites – National View



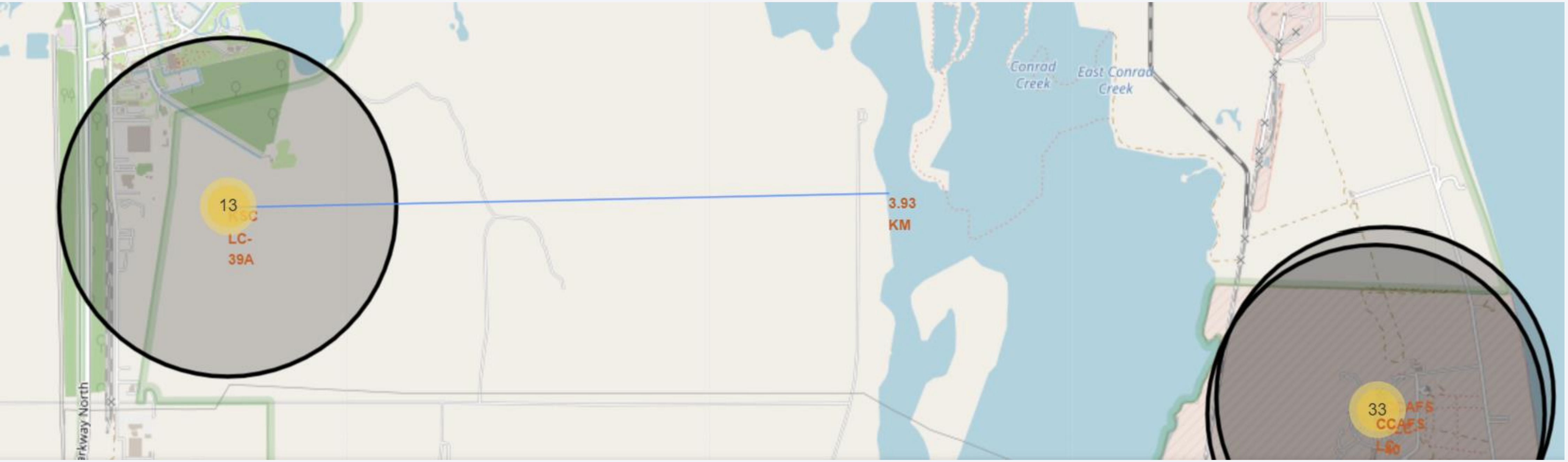
Screenshot of interactive map displaying SpaceX launch sites. The site in Florida is actually 3 sites where the majority of SpaceX launches occur. In the upper right corner lat and long coordinates are displayed based upon mouse hover location. Clicking on the yellow launch site circles zooms in on that location and provides additional detail as described in following slides.

SpaceX Launch Sites – Site and Launch View



Zooming in to the launch site level displays each launch radiating out from the center over time. Successful first stage landings are colored green and failures are red. Above the spiral is another nearby launch site which has had 7 launches. Similar launch data is rendered for that site by clicking on the green circle containing the number 7.

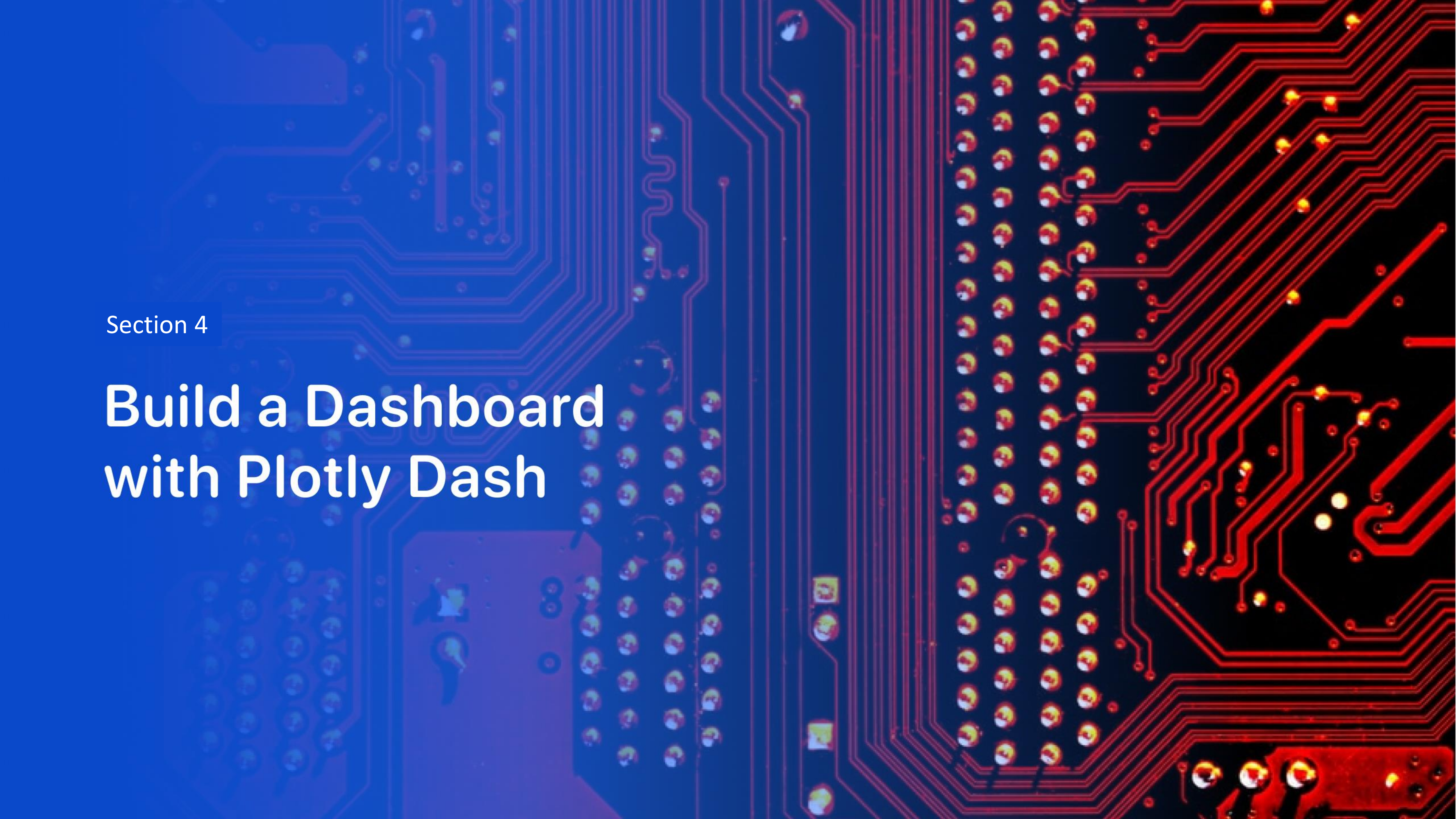
Proximity to coastline



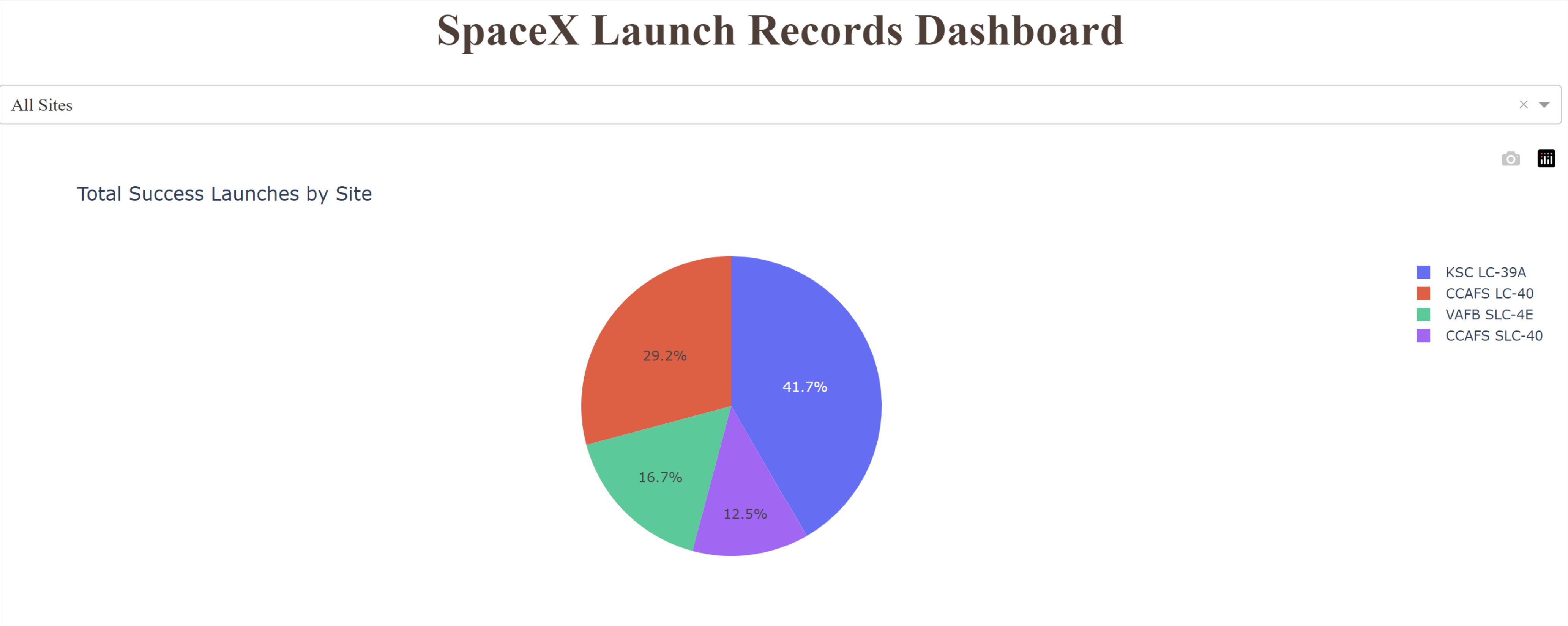
The yellow circle containing the number 13 represents one of three of SpaceX's Florida launch sites. The blue line represents the distance to the nearest coastline, measured at 3.93 km.

Section 4

Build a Dashboard with Plotly Dash



SpaceX Dashboard – Overall Launch Success

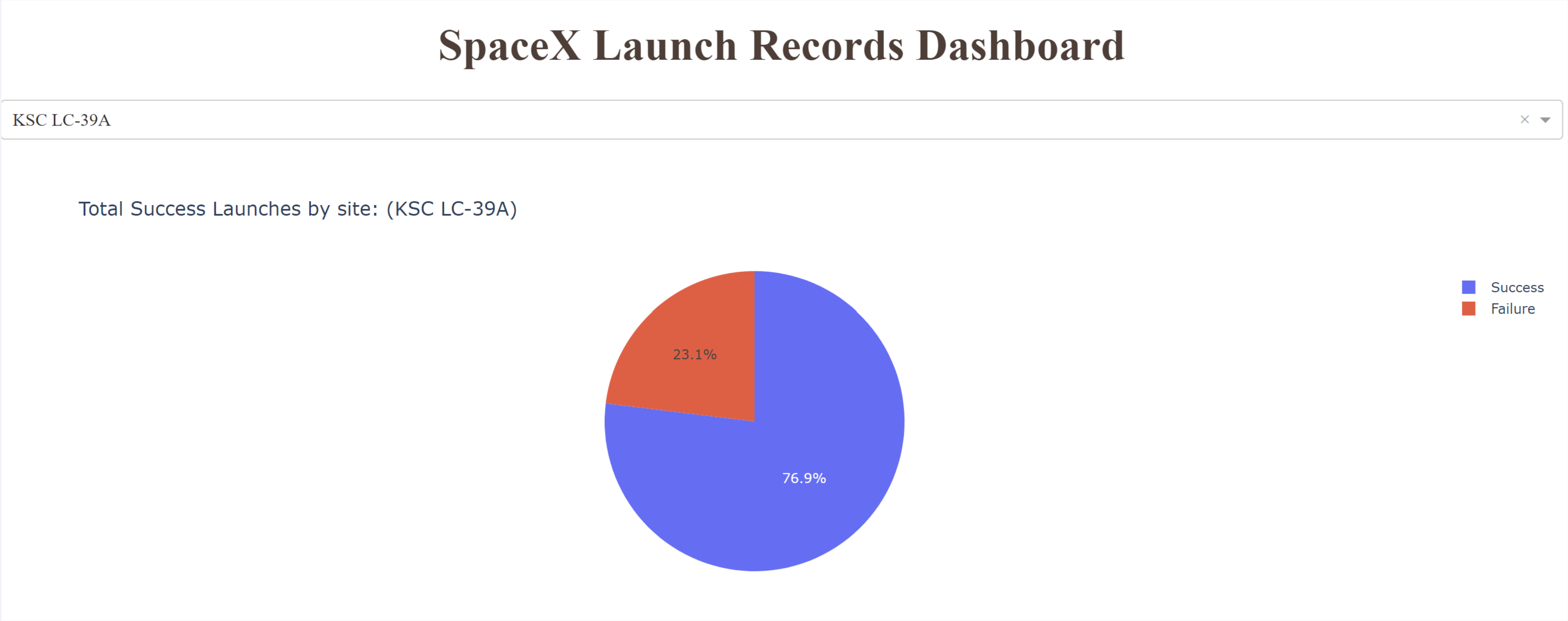


SpaceX Launch Records Dashboard

All Sites

Breakdown of SpaceX successful landings of stage 1 rockets by launch site. For example, the largest number of successful landings (41.7%) occurred at launch site KSC LC-39A. Individual site data can be selected through the dropdown menu beneath the dashboard title.

SpaceX Dashboard – Site-Specific Launch Success



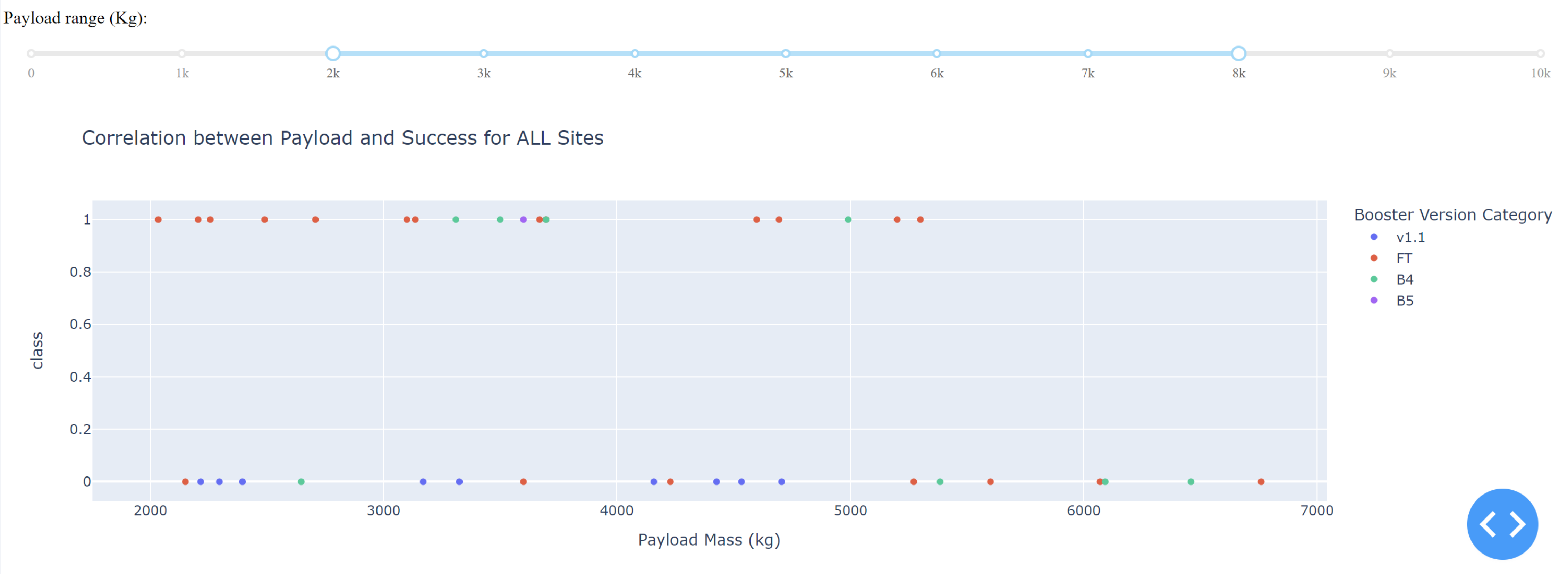
SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches by site: (KSC LC-39A)

Breakdown of SpaceX successful landings of stage 1 rockets by a specific launch site. This example is for the most successful SpaceX launch site KSC LC-39A

SpaceX Dashboard – Launch Success by Payload



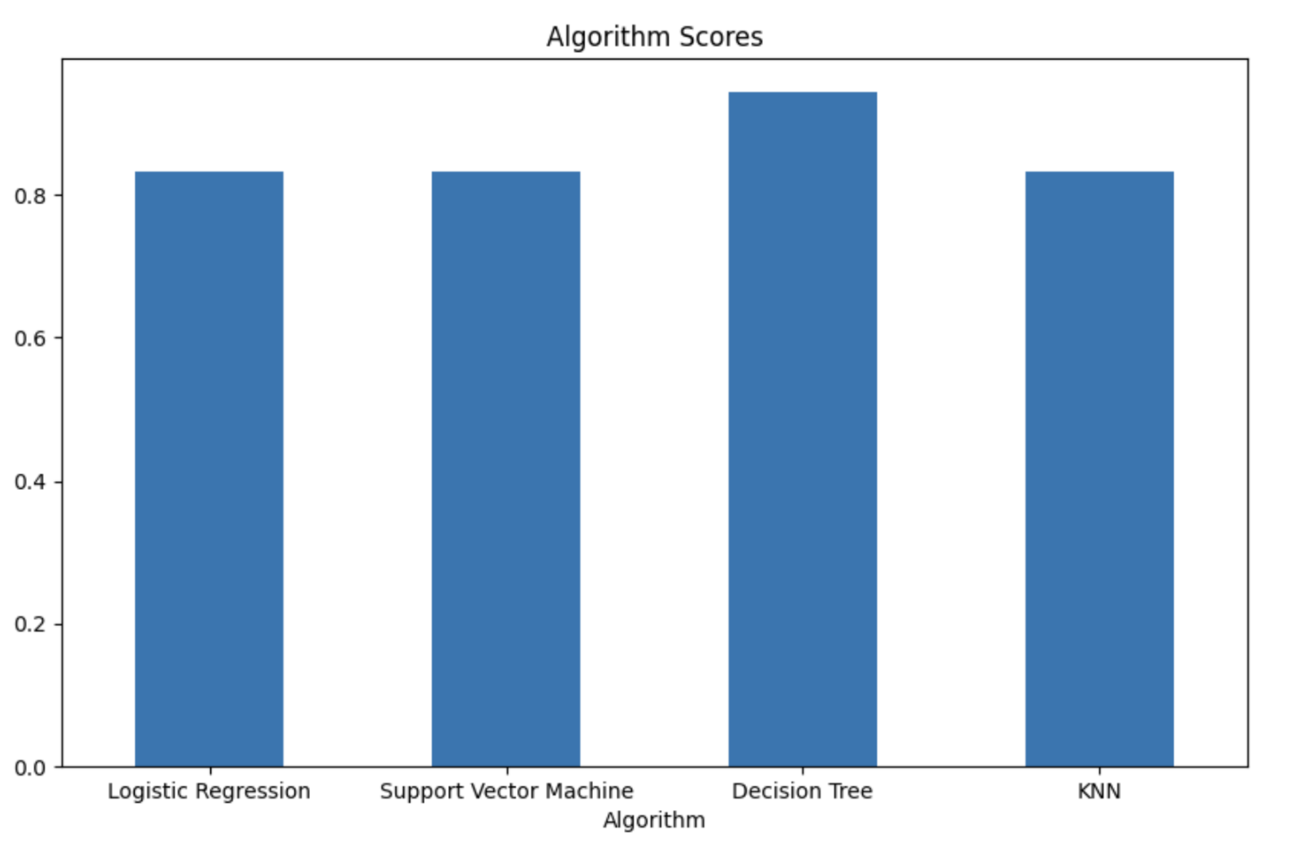
Breakdown of SpaceX successful landings of stage 1 rockets by all launch sites constrained by payload range selected in the payload range slider control. In this example we are seeing launch success for payload range from 2Kg to 8Kg. Data points are colored by the booster version utilized in the launch. Within this payload range, the FT booster version is shown to be the most successful.

Section 5

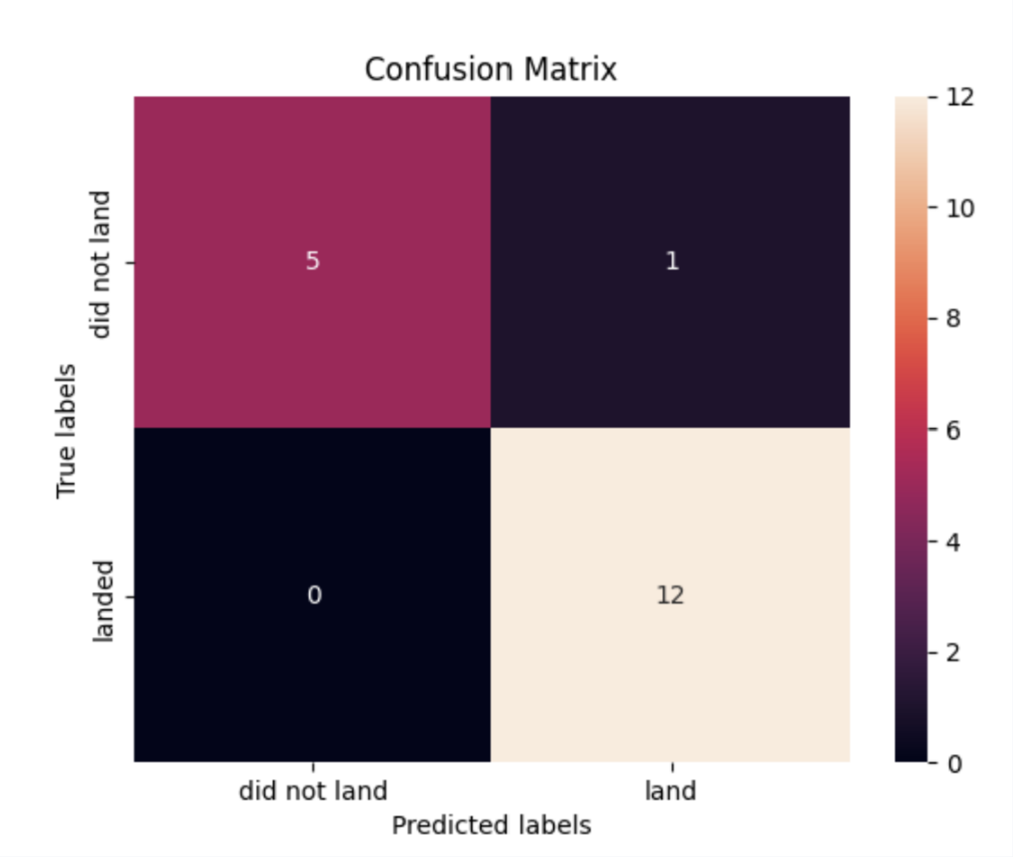
Predictive Analysis (Classification)

Classification Accuracy

Algorithm accuracy scores based on test data and optimum hyper parameters for each algorithm. The Decision Tree algorithm showed a marginally higher accuracy score of .94%



Decision Tree Confusion Matrix



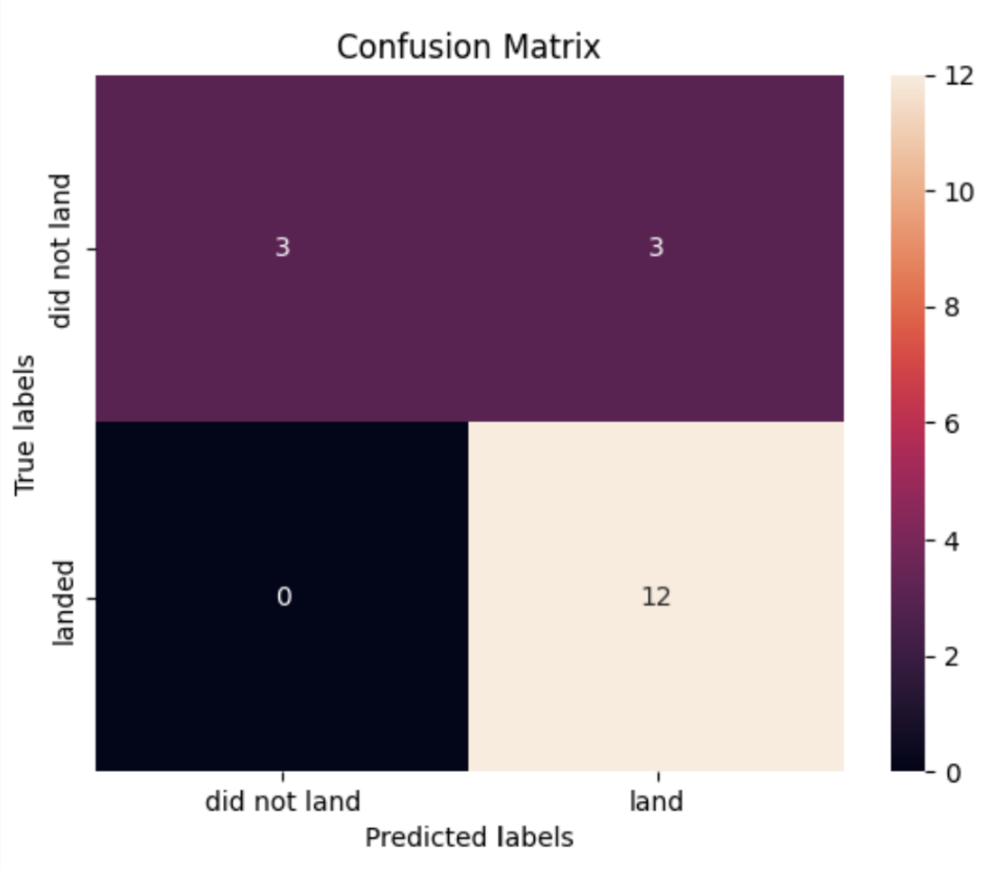
Based upon the test data set the decision tree confusion matrix above indicates the algorithm correctly predicted 12 of 12 successful landings and 5 of 6 unsuccessful landings.

Conclusions

1. All of the algorithms trained game meaningful predictions; particularly with respect to predicting successful landings.
2. The decision tree model gave the highest accuracy score (94% vs 83%) based upon the test data. The higher score is related to this algorithms improved ability to predict failed launches as well as successful ones.
3. The test sample was relatively small at only 18 launches. Additional launches will be important to include as they occur.
4. We should be able to make improvements to our launch pricing models based upon the predictions rendered by the decision tree algorithm trained as part of this analysis.

Appendix

- Included below for comparison is the confusion matrix generated for logistic regression, support vector machine, and KNN algorithms. The accuracy scoring for each of these algorithms came out the same.



Thank you!