IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jonathan Duncan
12/31/2024

Outline

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

Executive Summary

- The following methodologies were used to analyze data:
  - Data Collection using REST API from SpaceX databases along data scraping and normalization
  - Exploratory Data Analysis (EDA), including data wrangling data visualization and interactive visual analytics
  - Machine Learning Prediction.
- Summary of all results
  - Used EDA to identify which features are the best to predict success
  - Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using collected data.

Introduction

- SpaceX is the most successful company in the current commercial space age and this is mostly due to their ability to re-use the first stage of their launches
- The Goal of this project is to understand the factors that lead to the successful landing of their 1st stage in order to determine how best to replicate those successes

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - REST API calls to SpaceX databases
- Performed data wrangling
  - Data was normalized, filtered and collected into a pandas dataframe
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Built, tuned, evaluated classification models

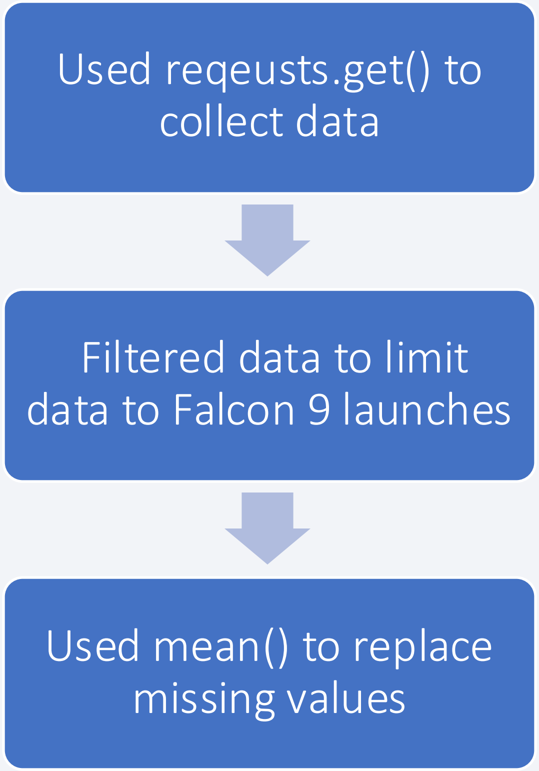
Data Collection

- Data sets were collected using REST API calls to the following sources using web scraping techniques
  - Space X API (https://api.spacexdata.com/v4/rockets/
  - Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

- Data was collected using REST API calls to publicly available SpaceX databases.

- GitHub link for the completed SpaceX API calls notebook



Used reqeusts.get() to collect data

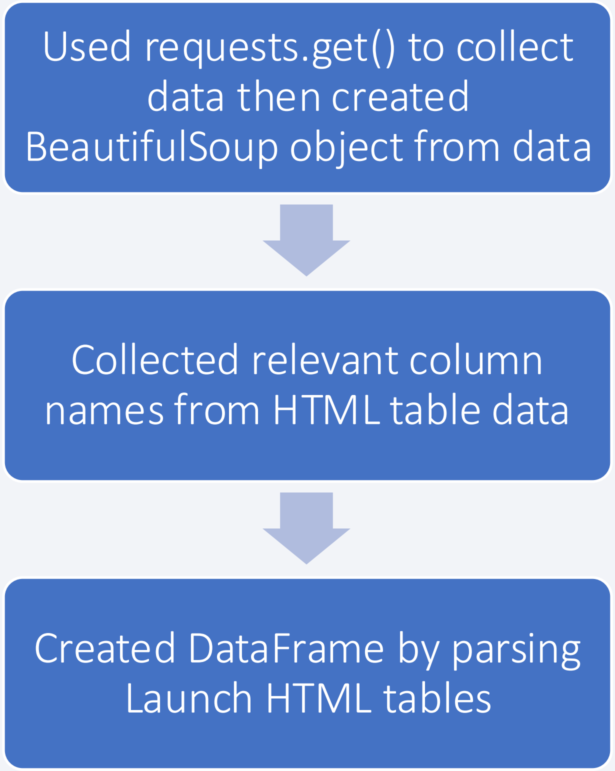
Filtered data to limit data to Falcon 9 launches

Used mean() to replace missing values

Data Collection - Scraping

- Collected launch data from Wikipedia using web-scraping techniques

- GitHub link for the completed web scraping notebook



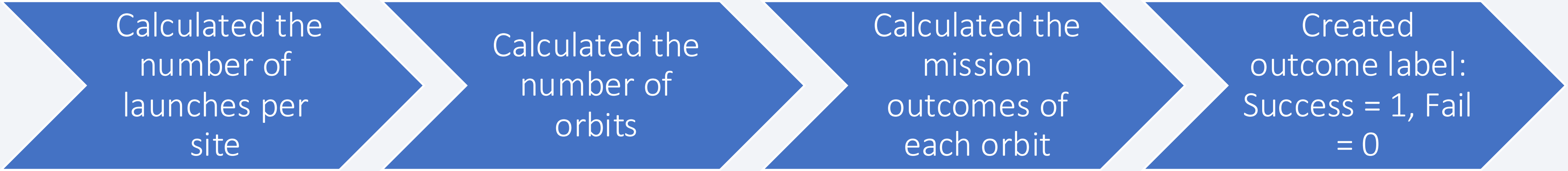
Used requests.get() to collect data then created BeautifulSoup object from data

Collected relevant column names from HTML table data

Created DataFrame by parsing Launch HTML tables

Data Wrangling

- Performed Exploratory Data Analysis (EDA) to find patterns in the data and determine the label for training supervised models


Calculated the number of launches per site

Calculated the number of orbits

Calculated the mission outcomes of each orbit

Created outcome label: Success = 1, Fail = 0

- GitHub link for the completed data wrangling notebook

EDA with Data Visualization

- Used scatter plots to show the relationships between the following
  - Flight Number and Launch Site
  - Payload Mass and Launch Site
  - FlightNumber and Orbit type
  - Payload Mass and Orbit type
- Used Bar chart to show success rate of each orbit type
- Used line chart to Visualize the launch success yearly trend
- GitHub link of completed EDA with data visualization notebook

EDA with SQL

- Summary of the SQL queries performed
  - Names of the unique launch sites in the space mission
  - Top 5 launch sites whose name begin with the string 'CCA’
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first successful landing outcome in ground pad was achieved
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
  - Total number of successful and failure mission outcomes
  - Names of the booster versions which have carried the maximum payload mass
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.
- GitHub link of completed EDA with SQL notebook

Build an Interactive Map with Folium

- Added Markers of all Launch Sites:
  - Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
  - Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.
- Added Colored Markers of the launch outcomes for each Launch Site:
  - Colored Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.
- Added Distances between a Launch Site to its proximities:
  - Colored Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- GitHub link of completed interactive map with Folium map

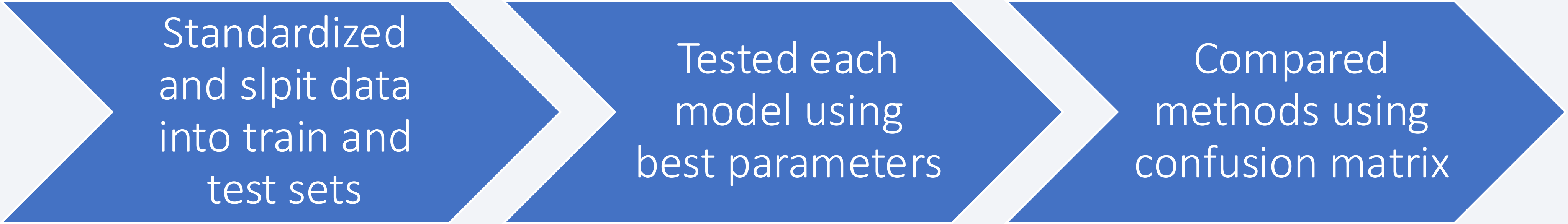
Build a Dashboard with Plotly Dash

- The following charts and plots were used to visualize the data
  - Pie chart to show percentages of launches by launch site
  - Scatter plot showing the payload range by launch site
- These charts and plots were chosen to better show the impact of launch site and payload range on the successful landing of the mission

- GitHub link of completed Plotly Dash lab

Predictive Analysis (Classification)

- Standardized the data, split it into training data and test data, found the best parameters and classification methods for predicting outcomes using the following models
  - SVM
  - Classification Trees
  - Logistic Regression
  - K Nearest Neighbors



Standardized and slpit data into train and test sets

Tested each model using best parameters

Compared methods using confusion matrix

GitHub link to completed predictive analysis lab

Results

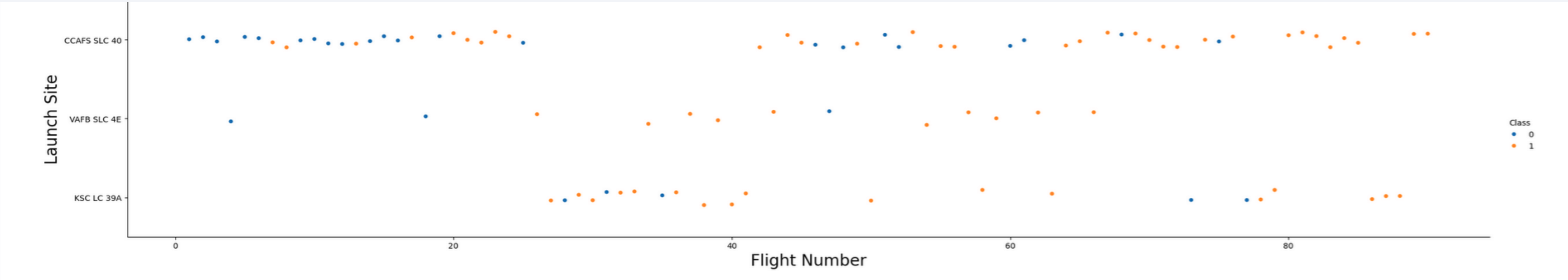
- Exploratory data analysis results
  - The KSC LC-39A site has the highest number of successful launches and highest success rate
  - The payload range with the highest launch success rate was 2000 – 4000 KG
  - The payload range with the lowest launch success rate was 6000 – 8000 KG
  - The FT Booster version had the highest launch success rate
  - The number of successful landing outcomes increased over time
- Link to screenshot 1 of interactive dashboard
- Link to screenshot 2 of interactive dashboard

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

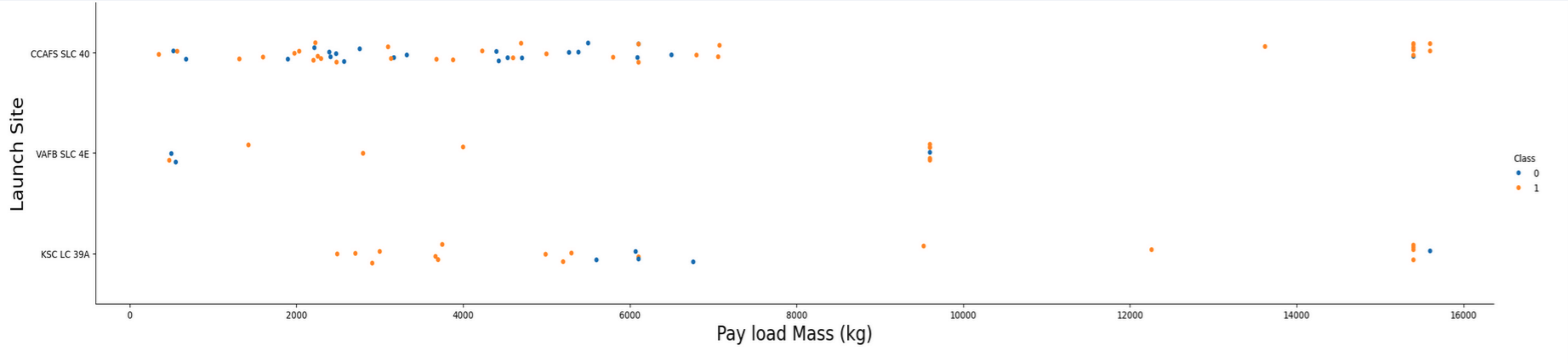
- Scatter plot of Flight Number vs. Launch Site



- This plot shows that the majority of launches occurred at the CCAF5 SLC 40 site and that the success rate improved as more launches occurred

Payload vs. Launch Site

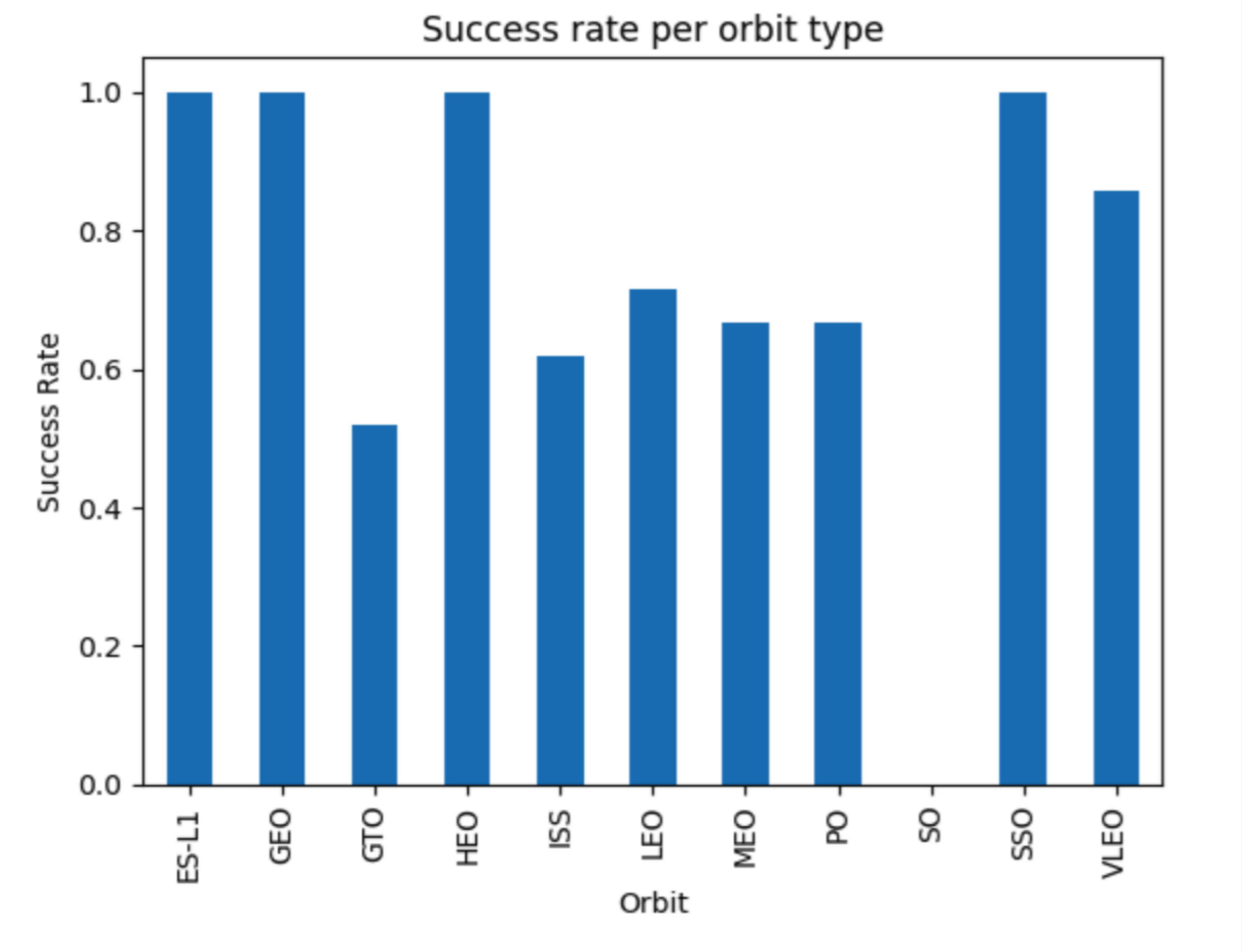
- Scatter plot of Payload vs. Launch Site



- Launches with payloads of over 9500 KG have a very high success rate
- The heaviest payload launches occurred at the CCAFS SLC 40 and KSC LC 39A sites

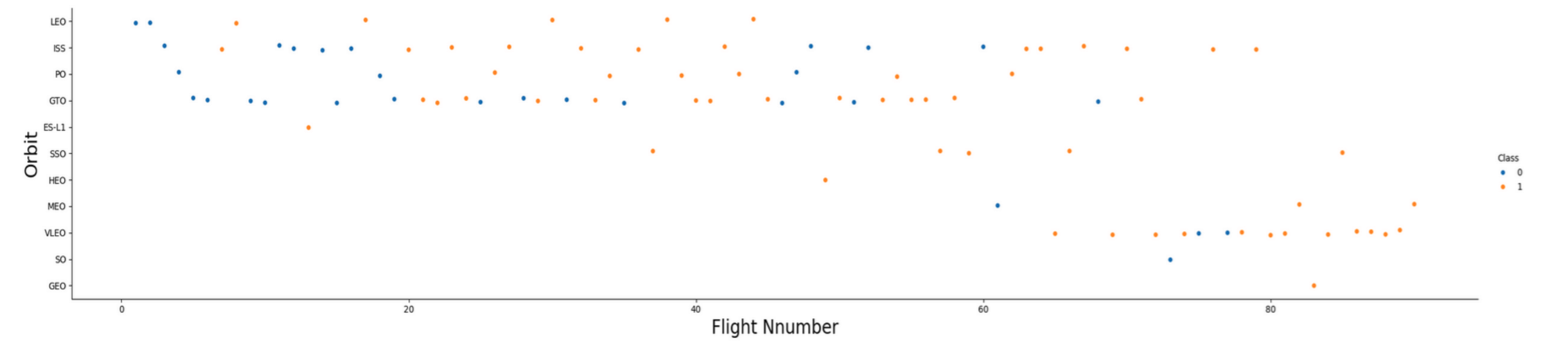
Success Rate vs. Orbit Type

- Bar chart showing the success rate of each orbit type
  - The ES-L1, GEO, HEO and SSO Orbits had the highest success rate.
  - The SO and GTO orbits had the lowest success rates



Flight Number vs. Orbit Type

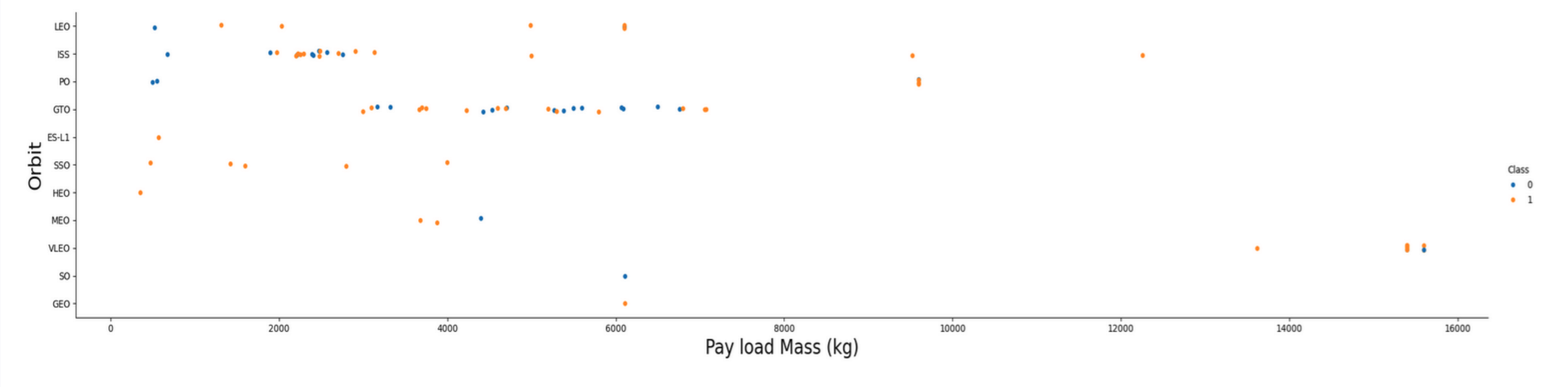
- Scatter point of Flight number vs. Orbit type



- For most orbits the success rate has improved as the flight number increased
- the VLEO orbit has become the most frequent orbit since the Flight Number has surpassed 60 flights

Payload vs. Orbit Type

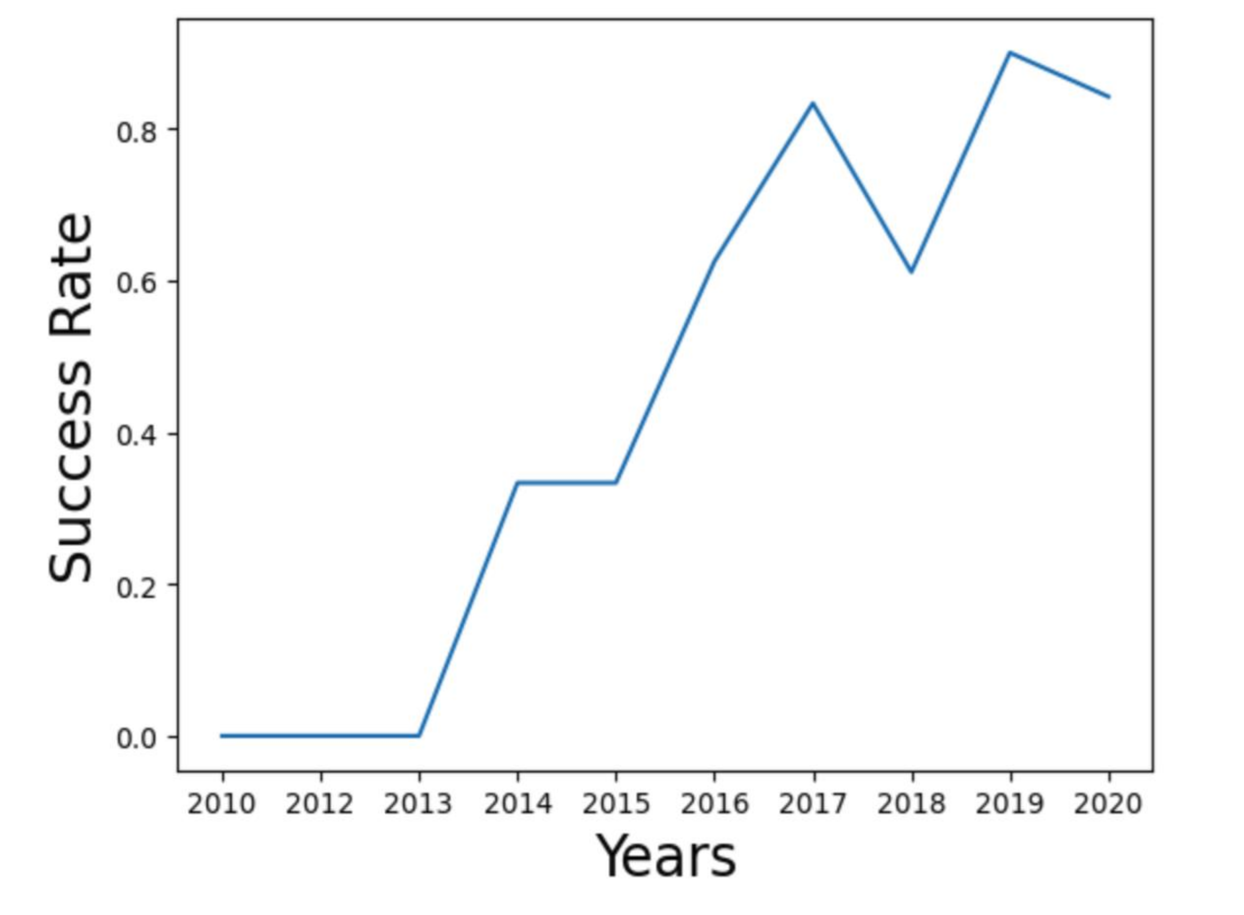
- Scatter point of payload vs. orbit type



- For most obits the success rate improved as the Payload Mass increased except for the GTO Orbit that doesn't appear to have a relationship between success rate and payload.

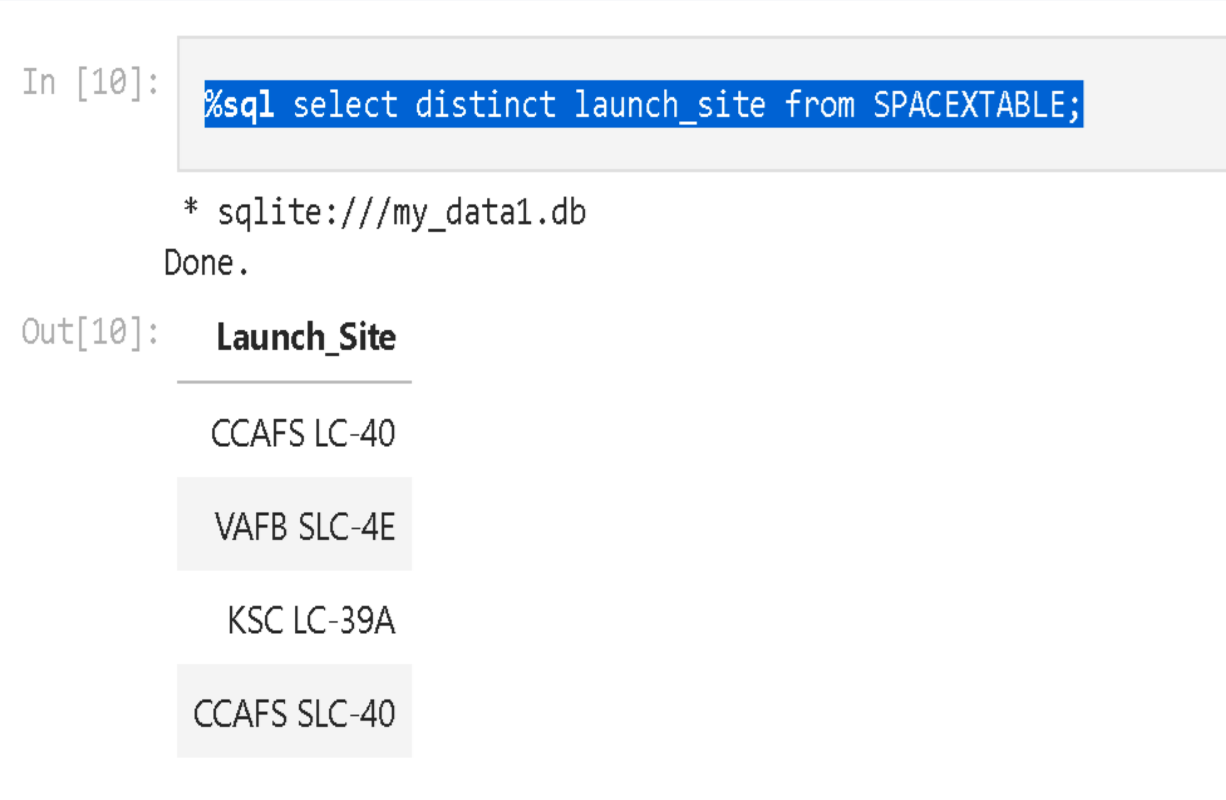
Launch Success Yearly Trend

- Line chart showing yearly average success rate
- The chart shows that the success rate has improved over the years.



All Launch Site Names

- The names of the unique launch sites are shown in image
- SQL query used to collect this information - `%sql select distinct launch_site from SPACEXTABLE;`



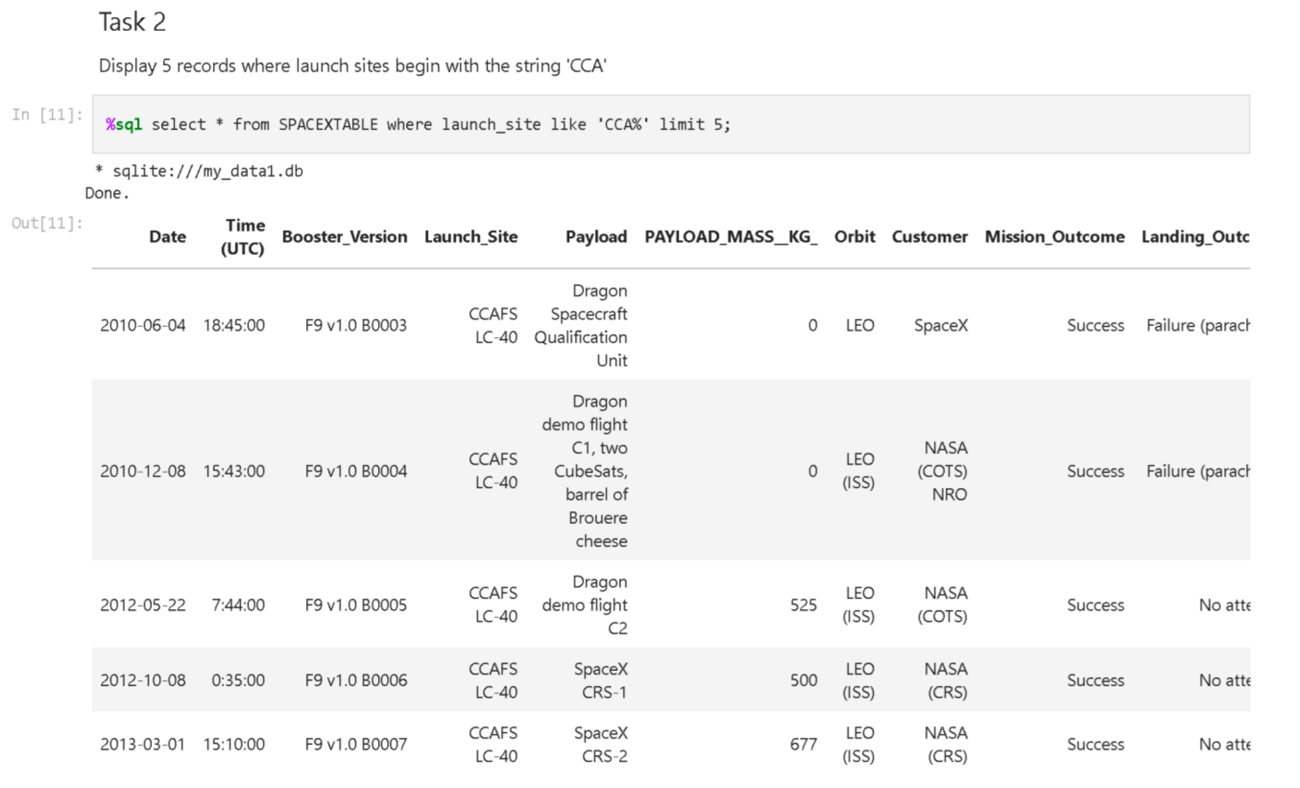
```
In [10]: %sql select distinct launch_site from SPACEXTABLE;

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

Out[10]:

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`
- SQL Query used to collect this data - `%sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;`



Task 2

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

```
In [11]: %sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
```

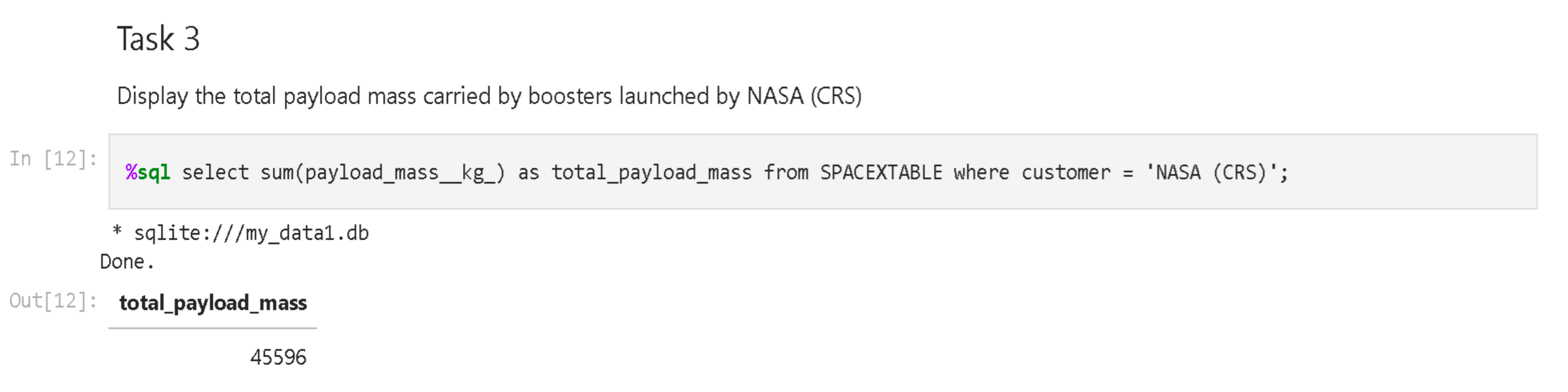
```
 * sqlite:///my_data1.db
Done.
```

Out[11]:

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 (parach
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 (parach
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

- The total payload carried by boosters from NASA



Task 3

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

```
In [12]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';

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

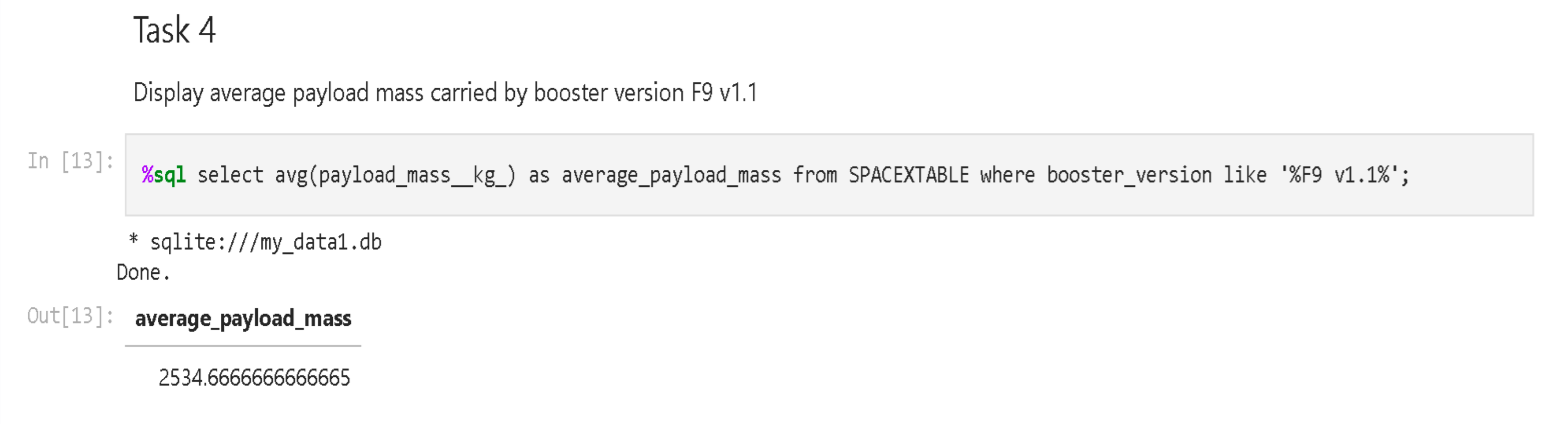
Out[12]:

total_payload_mass
45596

- SQL Query used to collect this information - `%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';`

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1



Task 4

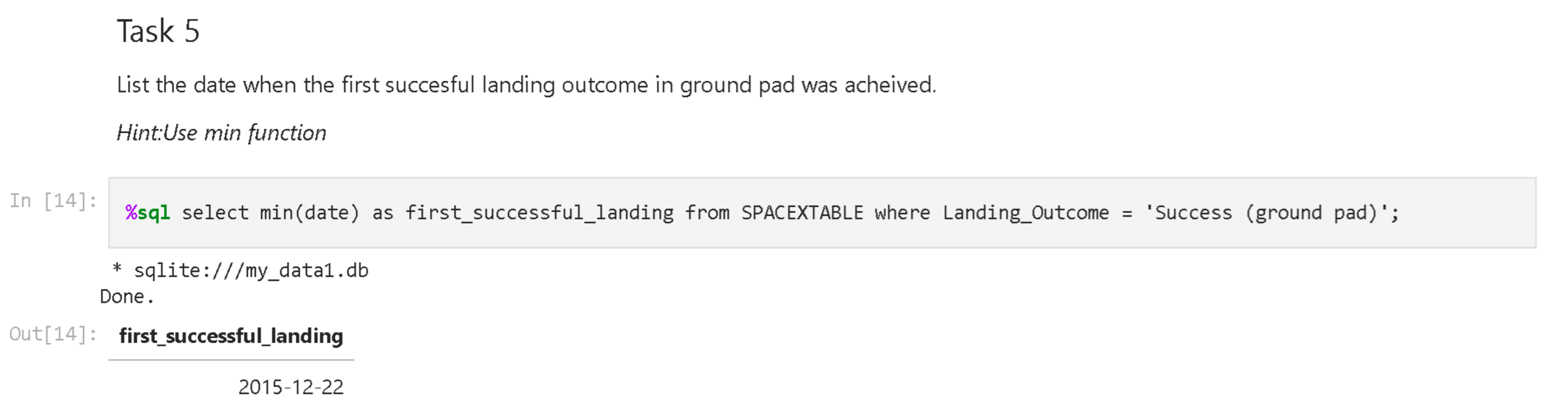
Display average payload mass carried by booster version F9 v1.1

```
In [13]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';

 * sqlite:///my_data1.db
Done.

Out[13]: average_payload_mass
         2534.6666666666665
```

- SQL Query used to collect this information - `%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';`

First Successful Ground Landing Date

- The Date of the first successful landing outcome on ground pad



Task 5

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

Hint:Use min function

```
In [14]:  %sql select min(date) as first_successful_landing from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)';
```

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

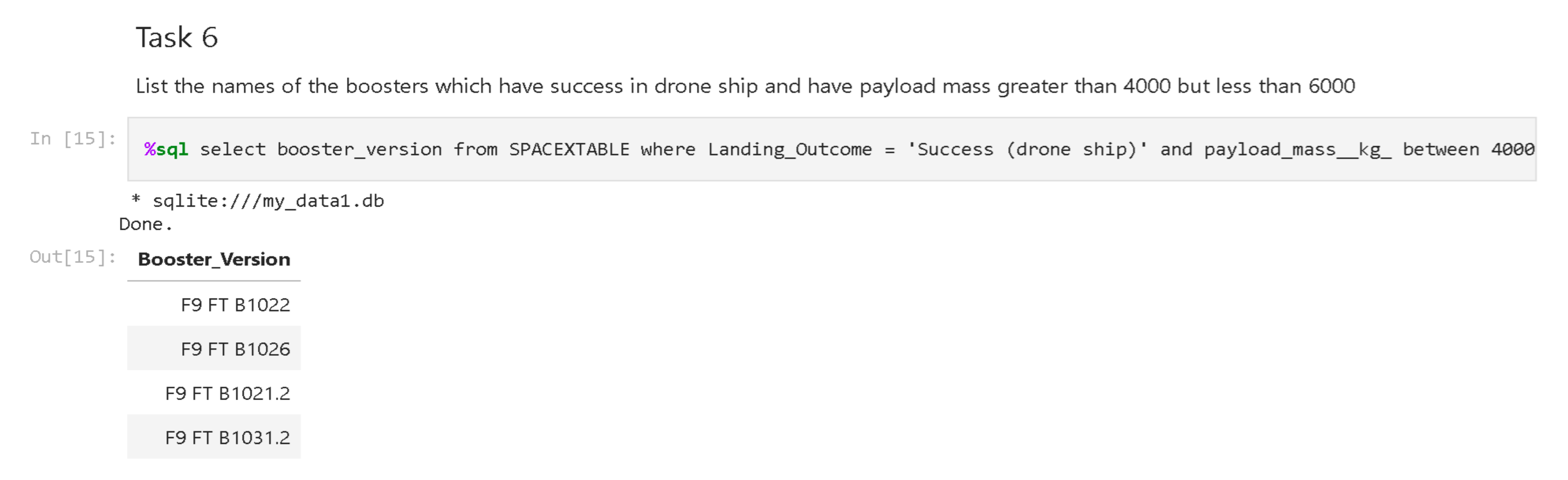
Out[14]:

first_successful_landing
2015-12-22

- SQL Query used to collect this information - `%sql select min(date) as first_successful_landing from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)';`

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 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 [15]: %sql select booster_version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000
```

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

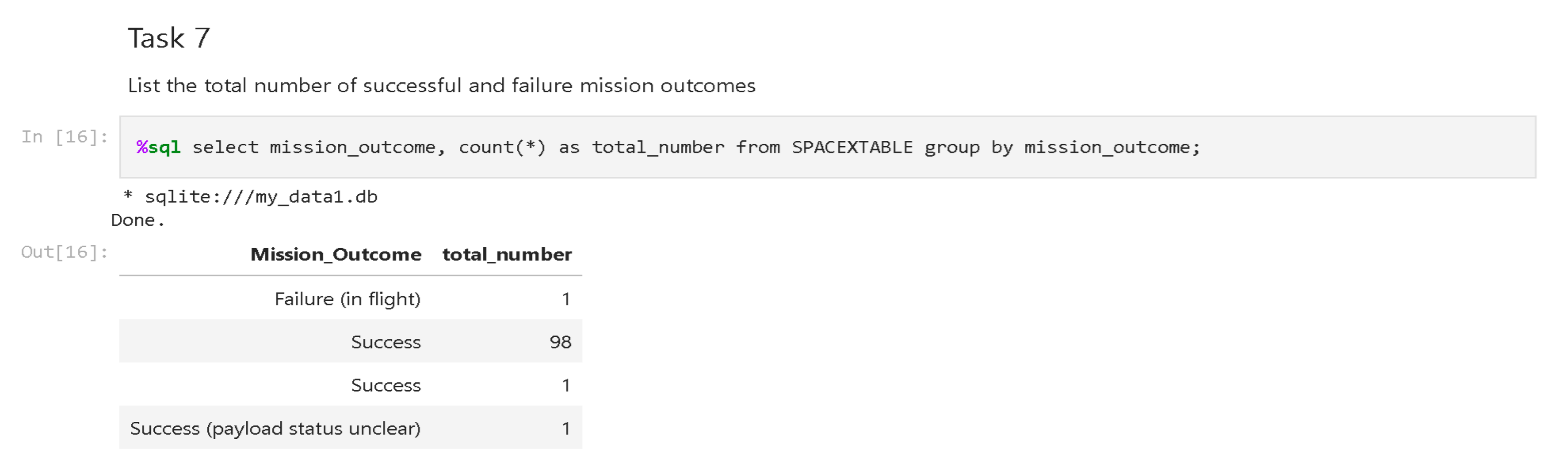
Out[15]:

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

- SQL Query to collect this information - `%sql select booster_version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;`

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes



Task 7

List the total number of successful and failure mission outcomes

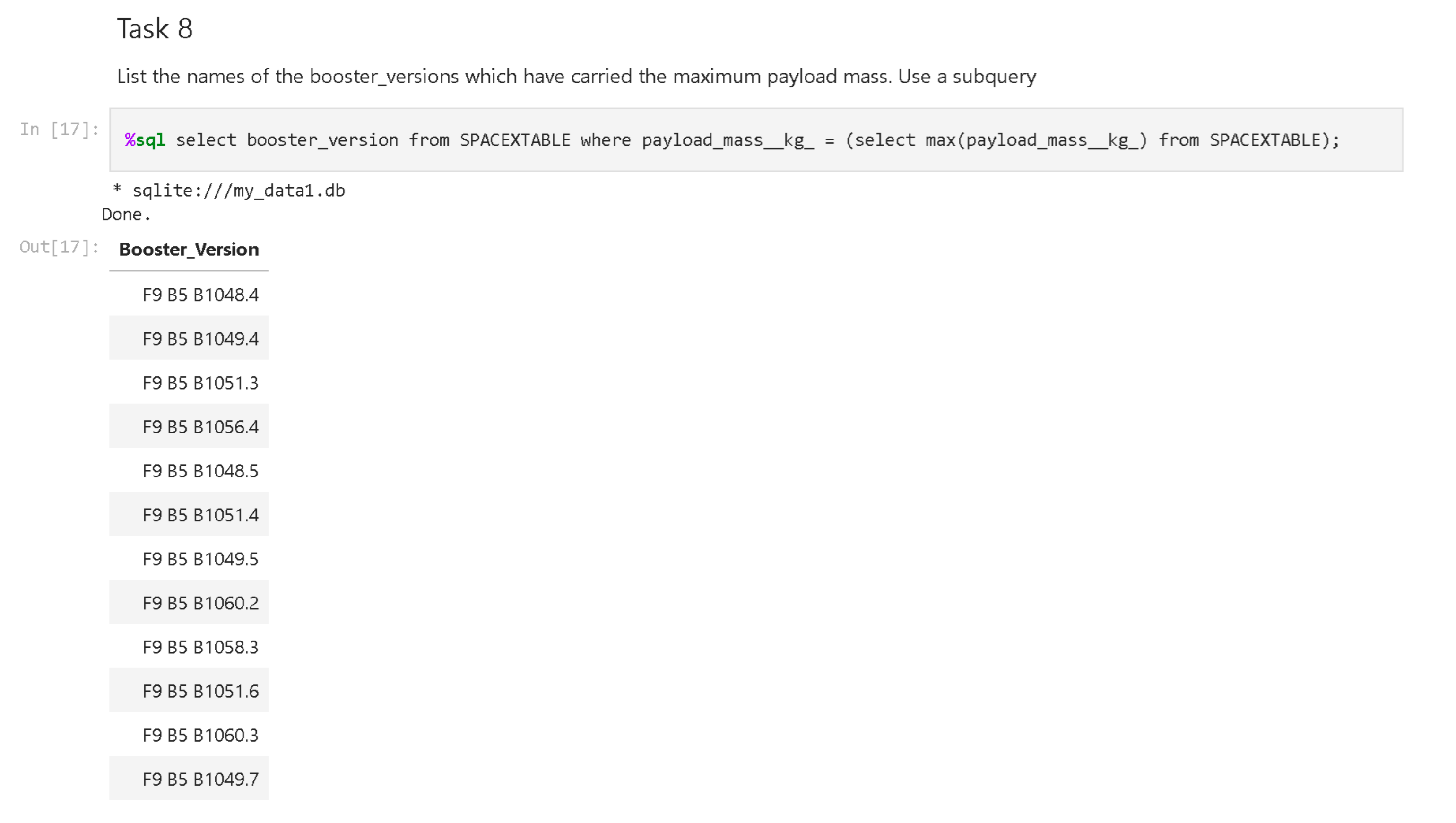
```
In [16]:  %sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;

 * sqlite:///my_data1.db
Done.
Out[16]:
```

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

- SQL Query used to collect this information - `%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;`

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass
- SQL Query used to collect this information - `%sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);`



Task 8

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

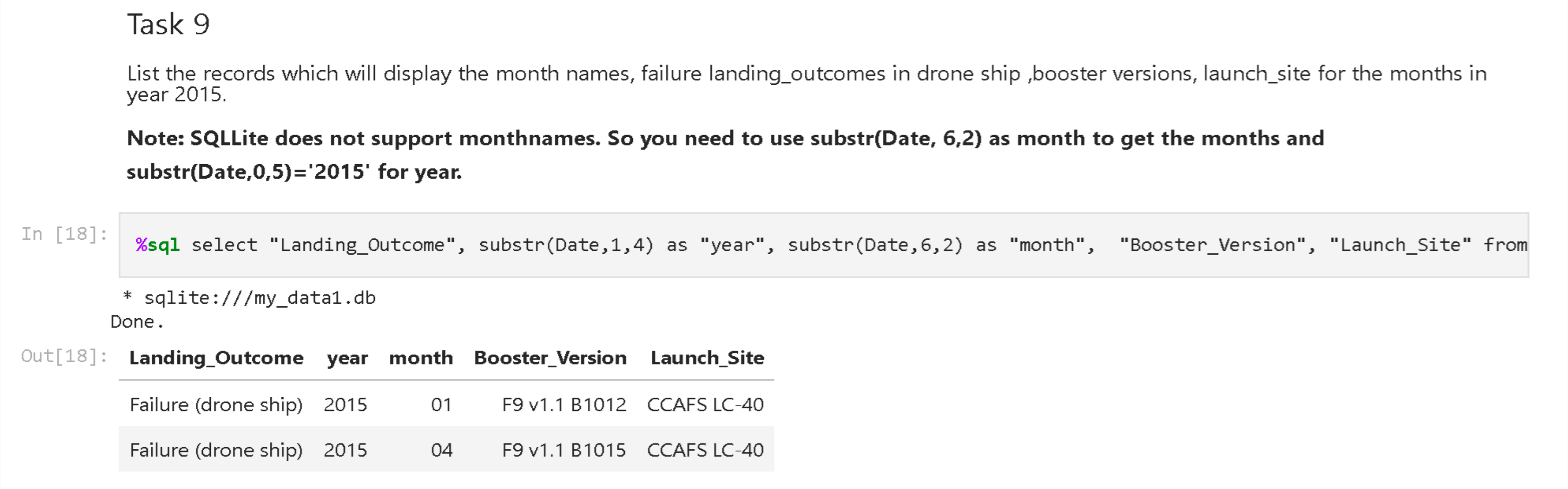
```
In [17]: %sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);

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

Out[17]:

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

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



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, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

In [18]:

```
%sql select "Landing_Outcome", substr(Date,1,4) as "year", substr(Date,6,2) as "month",  "Booster_Version", "Launch_Site" from
```

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

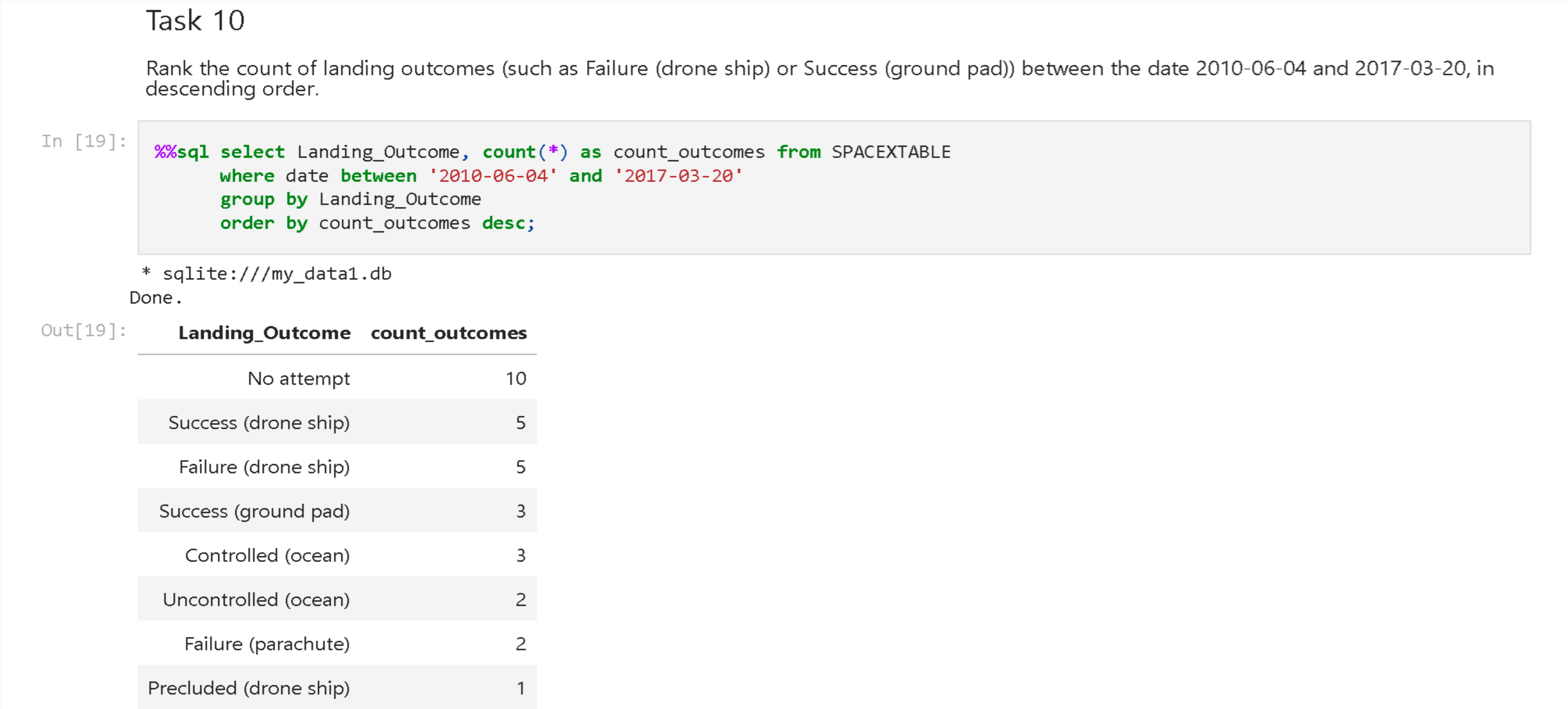
Out[18]:

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

- SQL Query used to collect this data - `%sql select "Landing_Outcome", substr(Date,1,4) as "year", substr(Date,6,2) as "month", "Booster_Version", "Launch_Site" from SPACE`

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



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 [19]:  %%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.
```

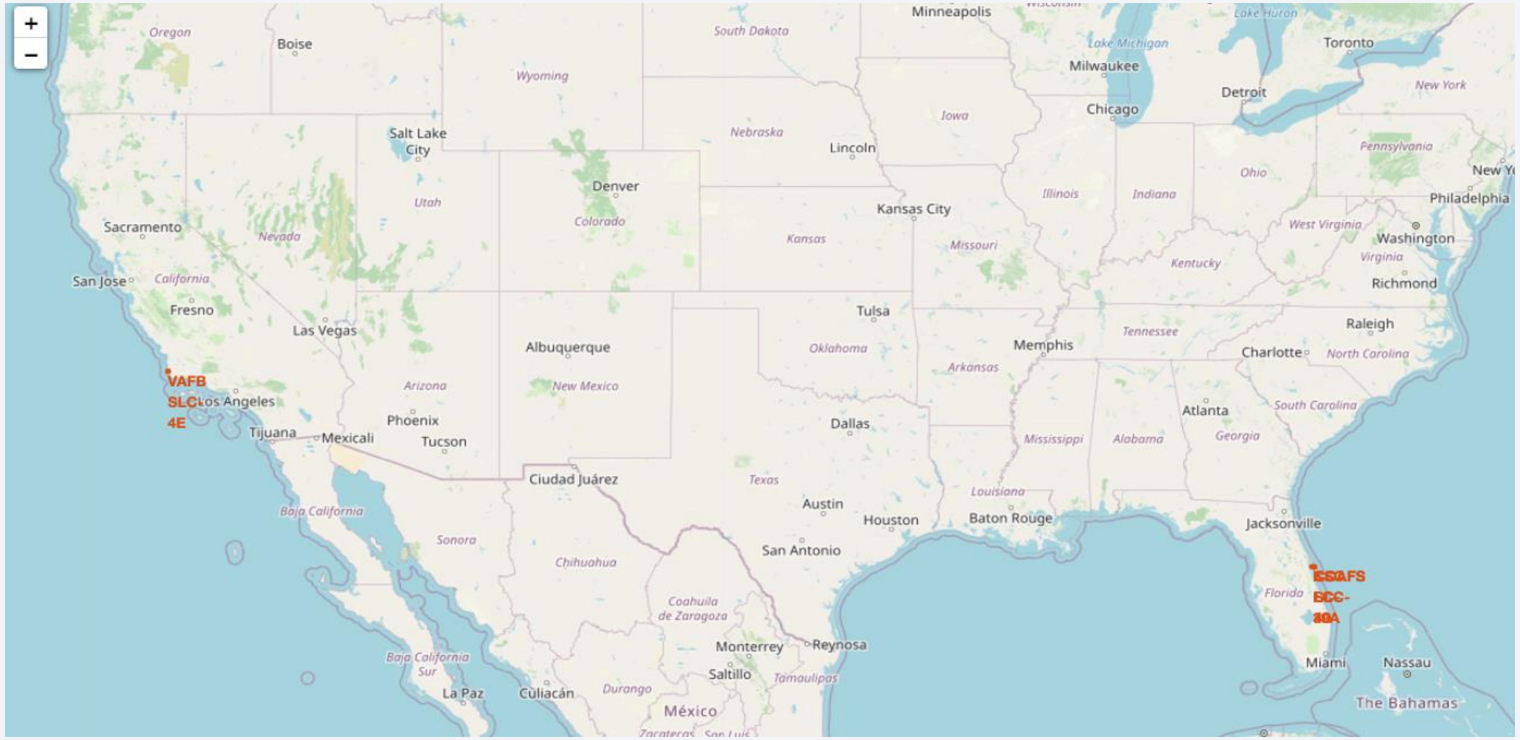
Out[19]:

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

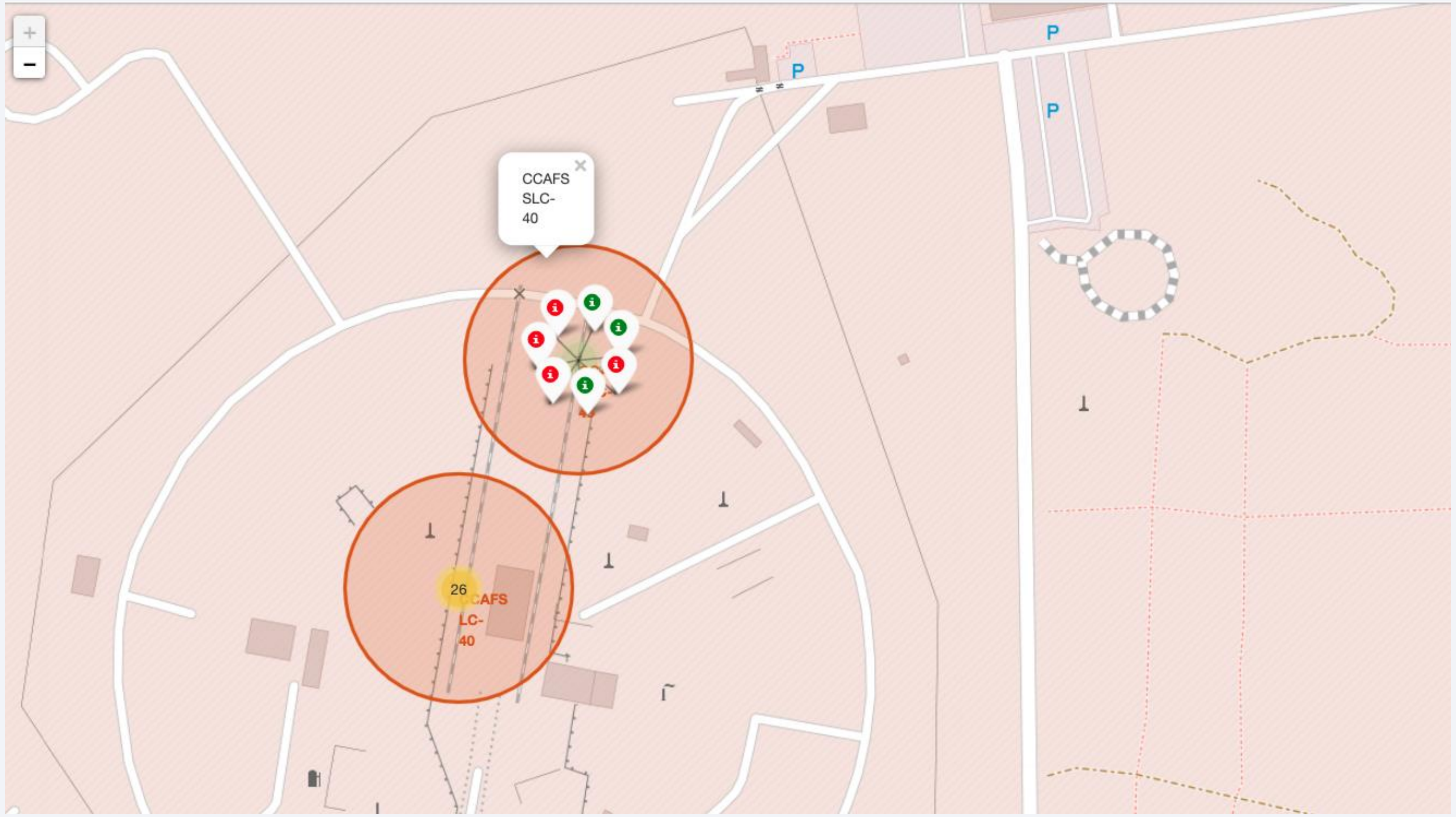
Launch Sites Proximities Analysis

Launch Sites



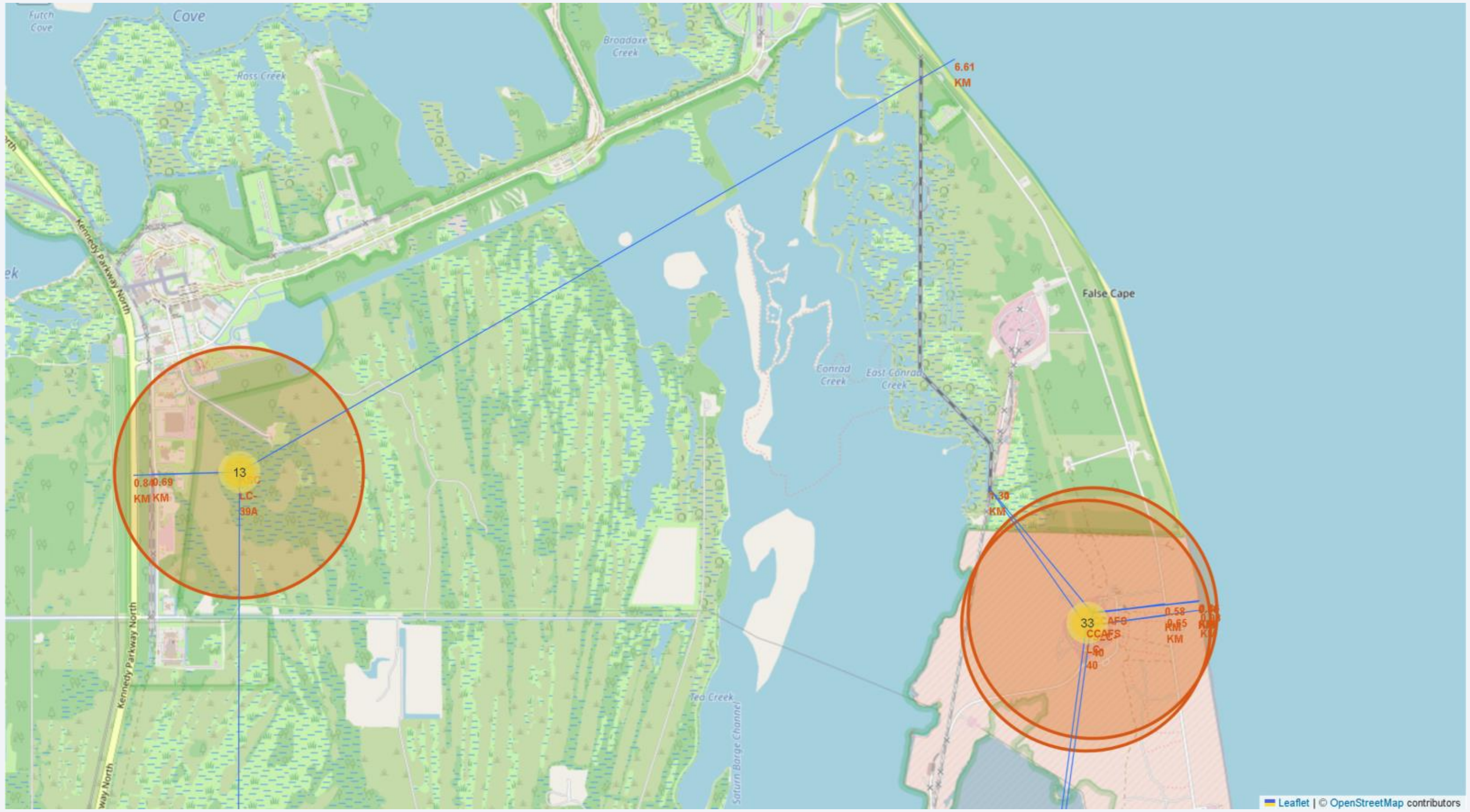
- Launch Sites located on each coast within close proximity to roads and railroads.

Launch Outcome at CCAFS SLC-40 site

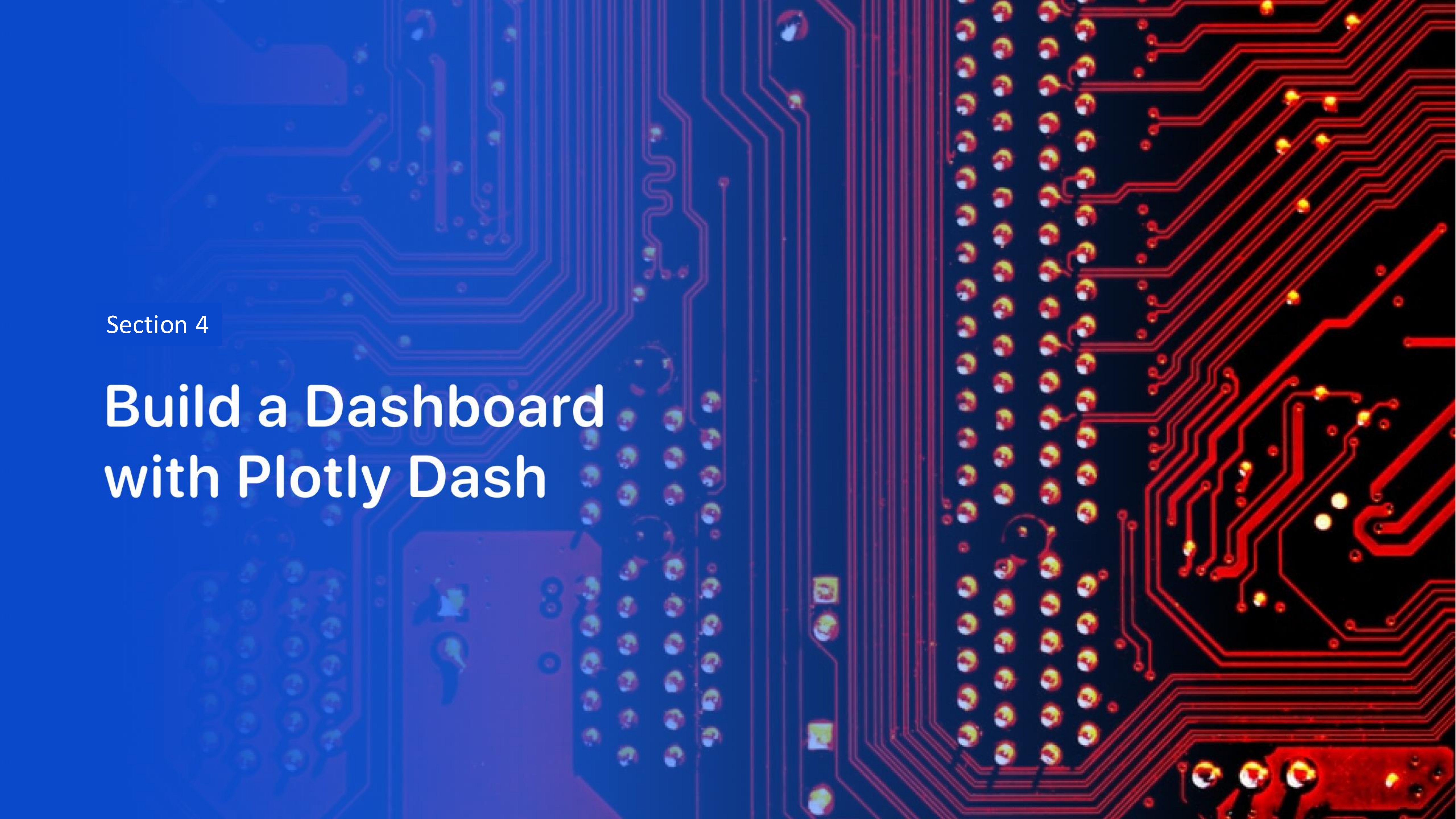


- Launch site outcomes at CCAFS SLC-40 site. Red markers indicate failures and green markers indicate successes

Launch Site Proximities



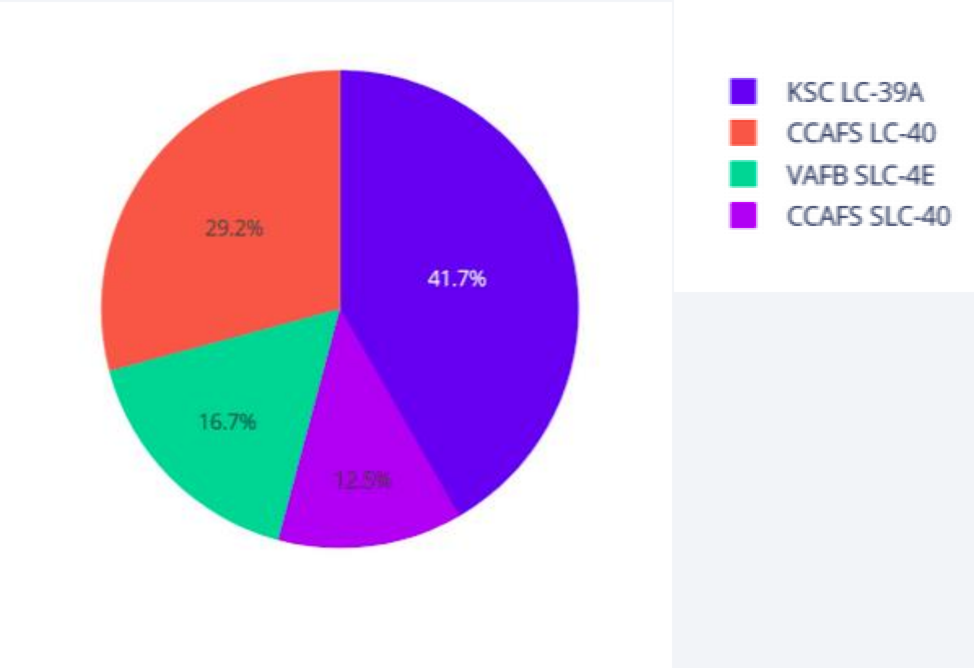
- The launch sites on the east coast are close to the coastlines, railroads and roads



Section 4

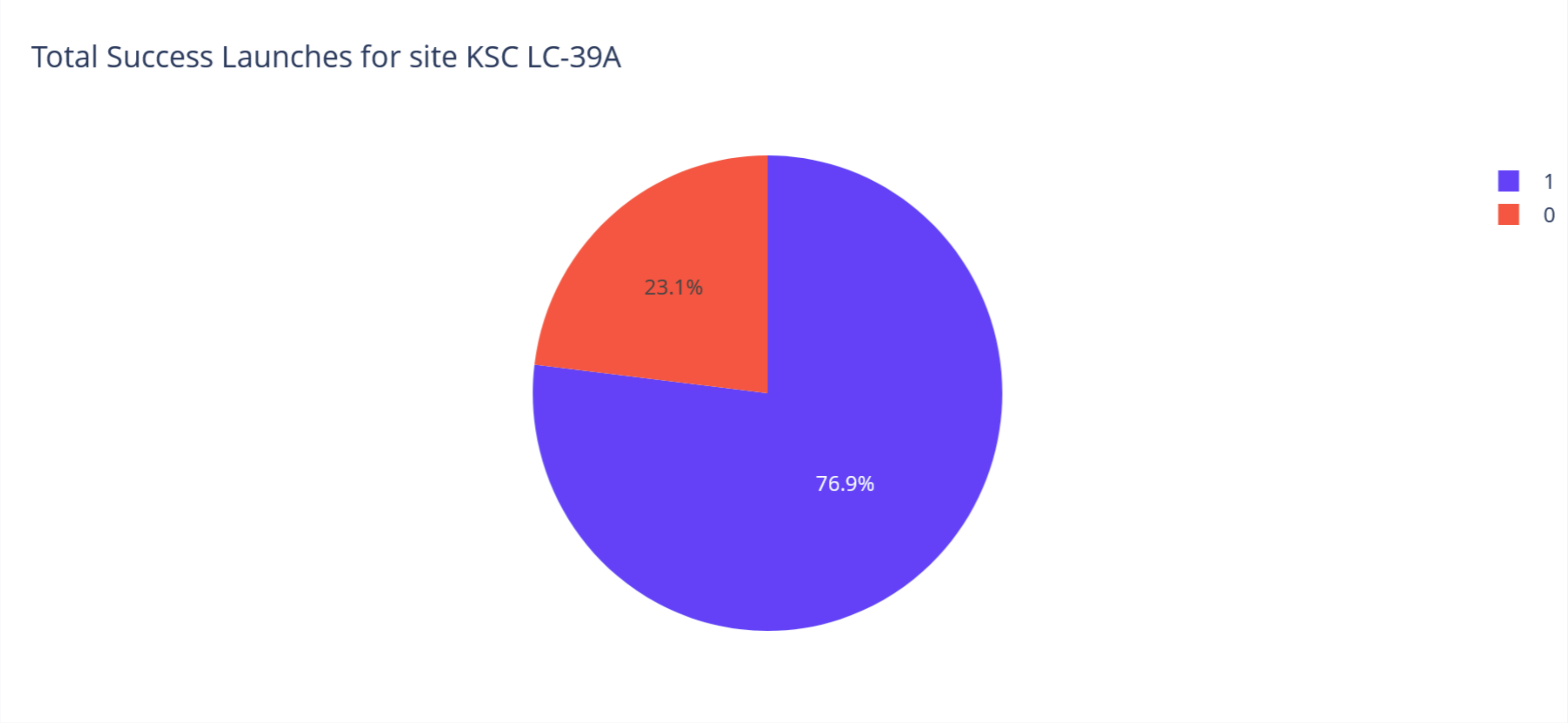
Build a Dashboard with Plotly Dash

Launch Success rate by Site



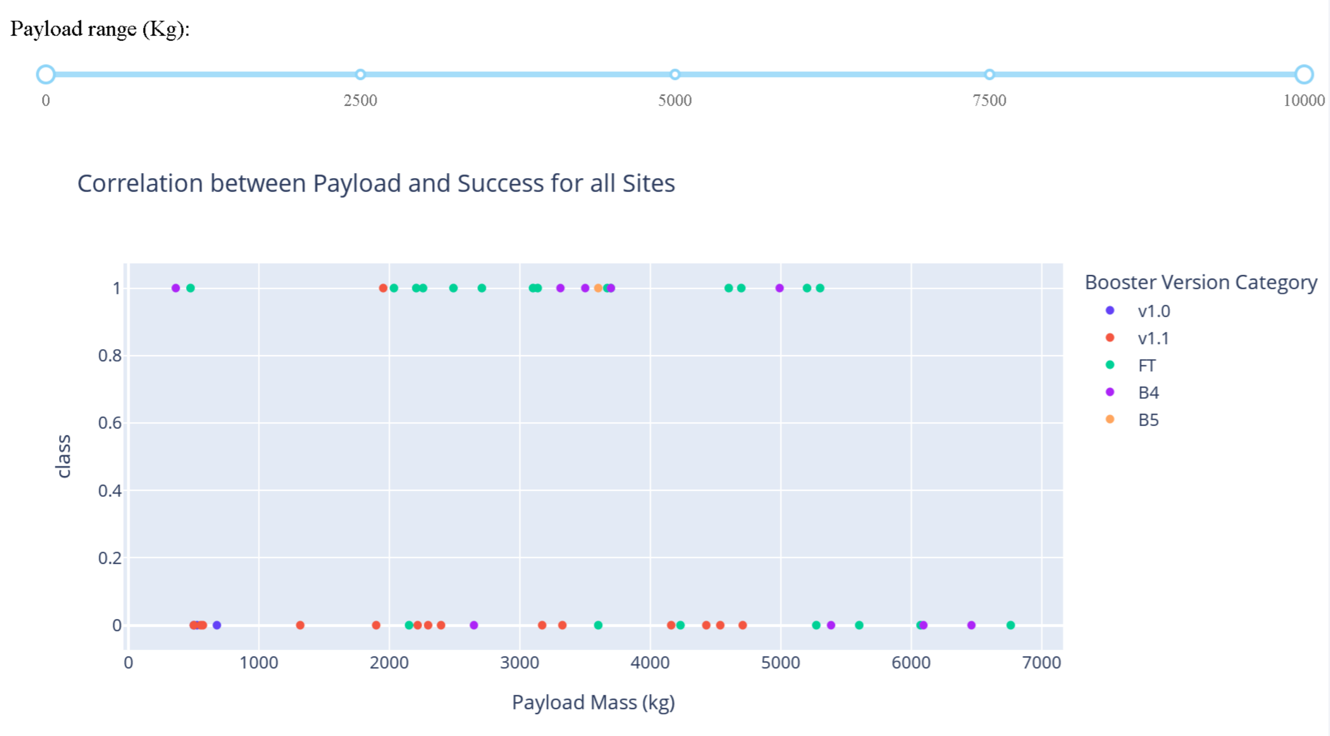
- The KSC LC-39A site has the highest success rate with the CCAFS LC-40 site having the second highest rate

Launch Success at the KSC LC-39A Site



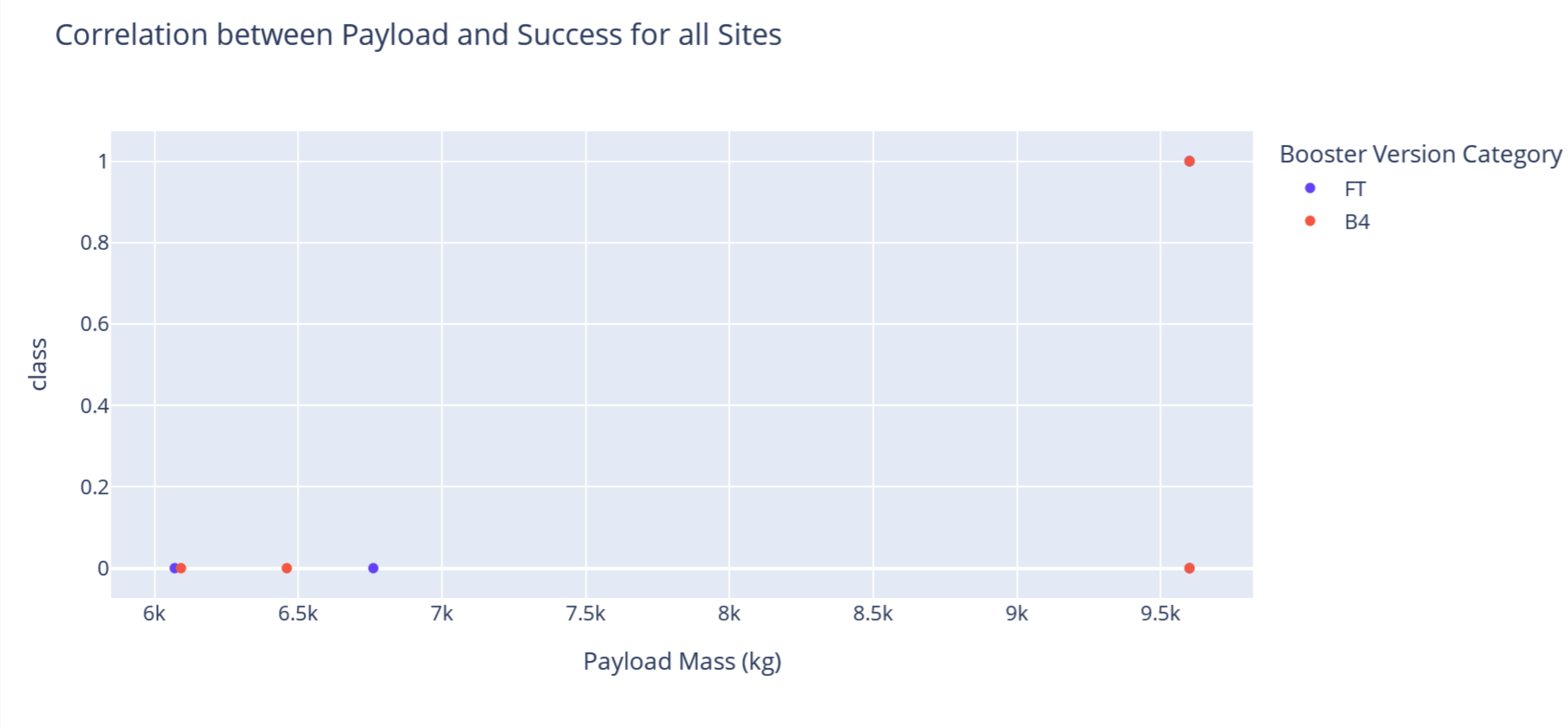
- The KSC LC-39A site has the highest succes rate at 76.9%

Payload vs Success rate for all Booster Versions



- Payloads of less than 6000 KG and the FT Booster has been the most successful combination

Payload vs Success rate for all Booster Versions



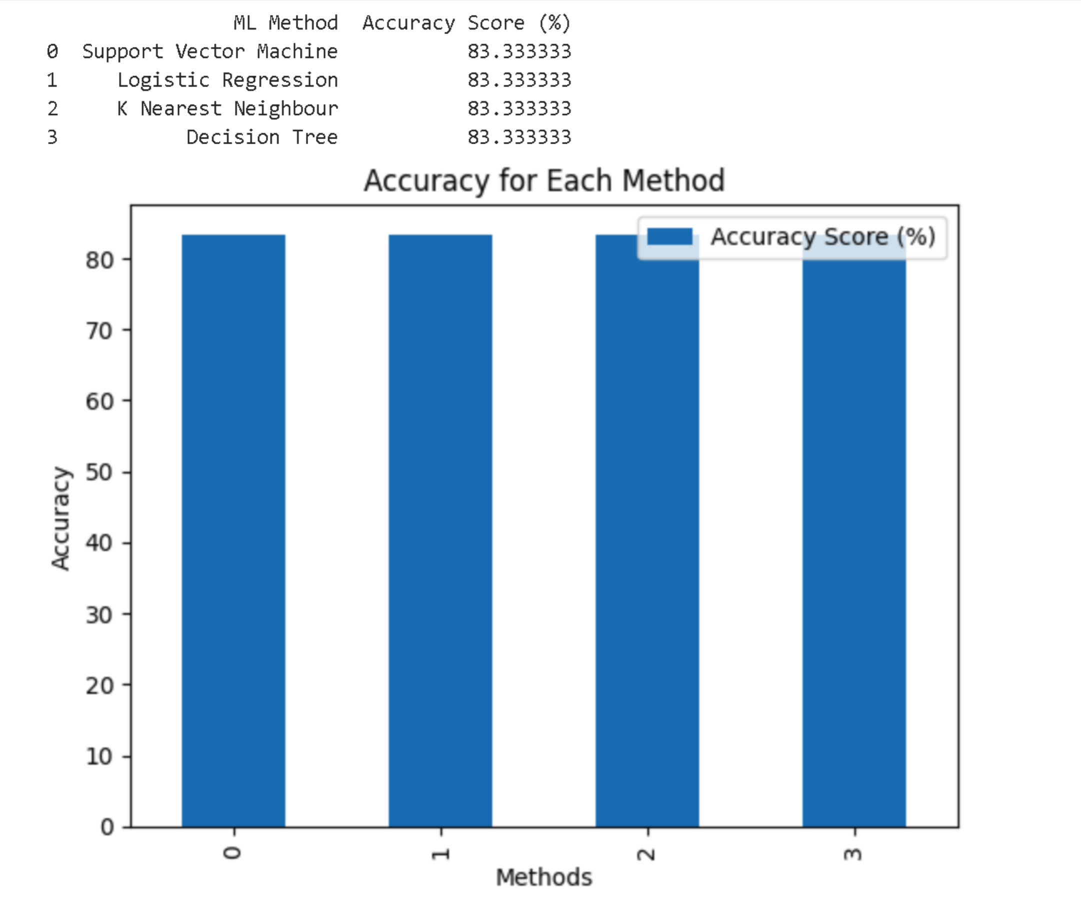
- Payloads of over 6000KG have been the least successful missions

Section 5

Predictive Analysis (Classification)

Classification Accuracy

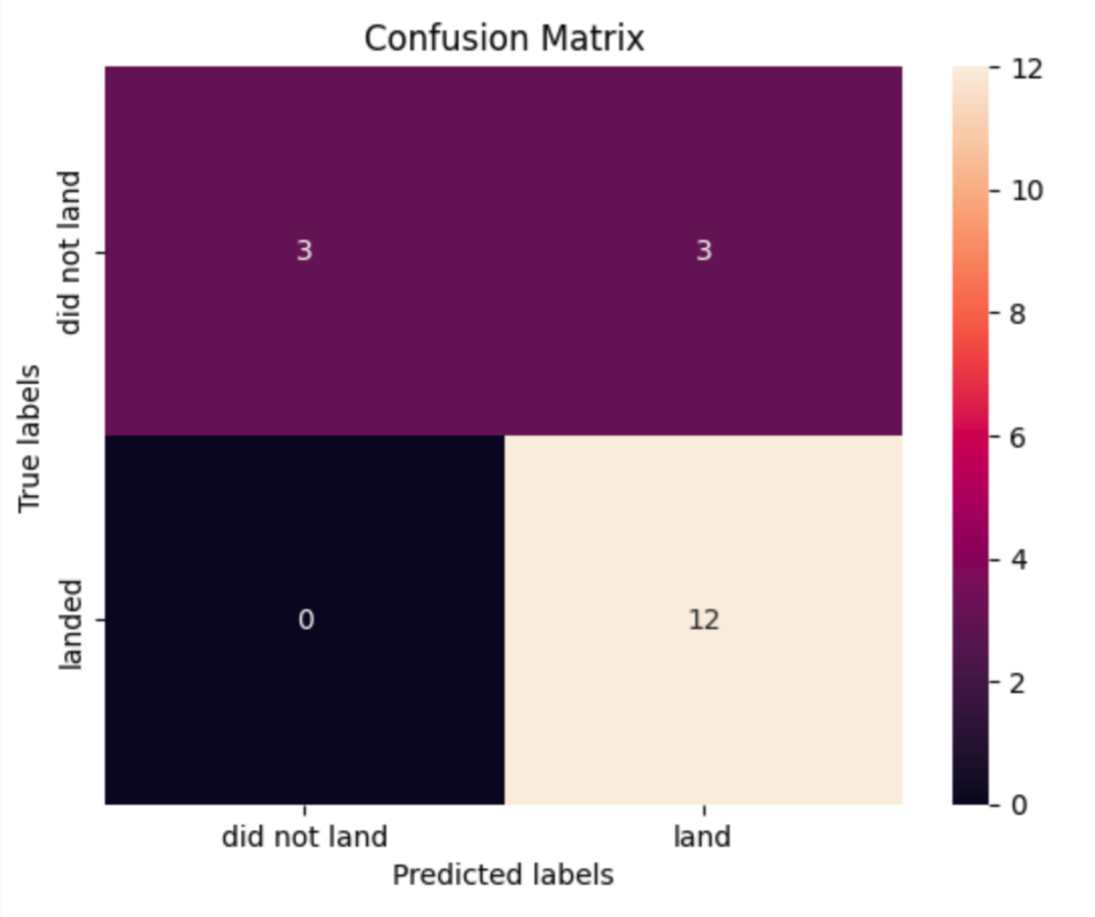
- This Bar Chart shows the built model accuracy for all built classification models
- All models were found to be 82% accurate.



```
                 ML Method  Accuracy Score (%)
0  Support Vector Machine           83.333333
1      Logistic Regression          83.333333
2      K Nearest Neighbour          83.333333
3            Decision Tree          83.333333
```

Confusion Matrix

- Each model was found to have the same Confusion Matrix as shown below



Conclusions

- The KSC LC-39A has the highest success rate of all the different sites.
- Missions with payloads less than 6000KG have a higher rate of success
- Success rate of launches has increased over time.
- Any of the Mchine Learning methods tried can predict the success outcome with an 83% accuracy.

Appendix

- GitHub link to all code created for this project - https://github.com/jduncangh/Applied_Data_Science_Capstone.git

Thank you!