IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

J-Philippe Caron
09-2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

On the SpaceX website, we can read the announcement of Falcon 9 rocket launch at the cost of 62 million dollars when competitors announce 165 million dollars for a similar service. Much of the savings is because SpaceX can reuse the first stage.

- Problems you want to find answers:
  - with what factors the rocket will land successfully?
  - Does the rate of successful landings increase over the years?
  - What is the best algorithm to be used for a binary classification

Section 1

Methodology

Methodology

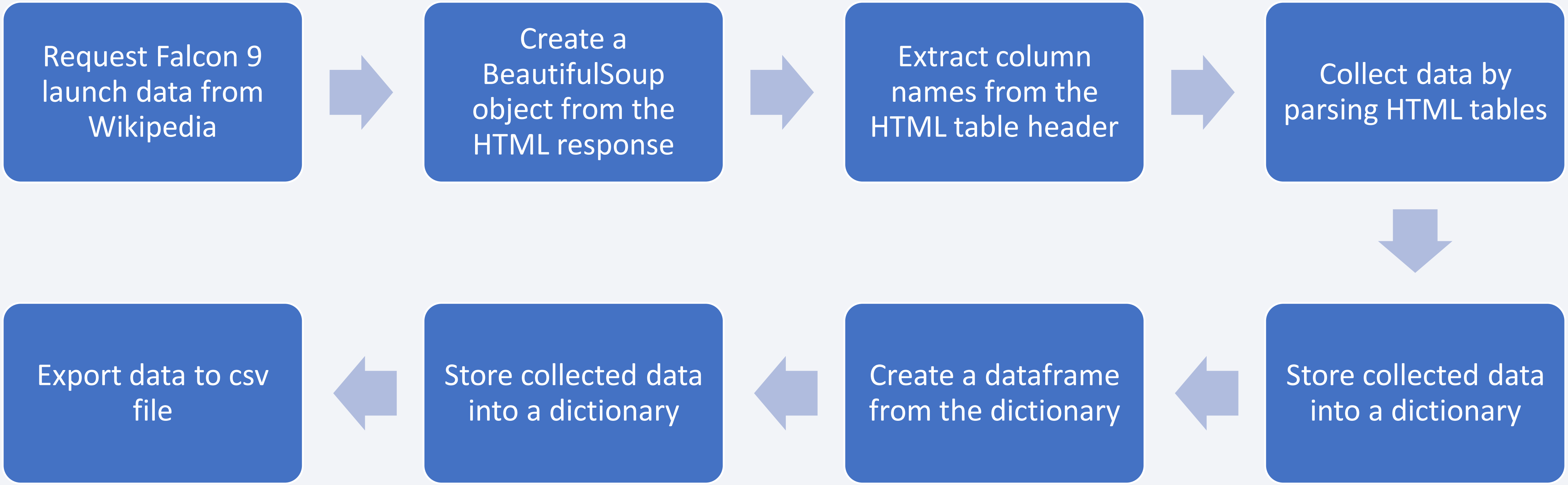
Executive Summary

- Data collection methodology:
  - Via SpaceX Rest API
  - Web scrapping from Wikipedia
- Perform data wrangling
  - Dropping irrelevant columns and using one hot encoding data field for machine learning
  - Dealing with missing values
  - Filtering the data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning, evaluating of classification models to ensure the results

Data Collection

- Data collection process is a combination of API requests from SpaceX Rest API and Web scraping data from a table in SpaceX's Wikipedia page.
- With these 2 data collections, we are able to get enough information about the launches to perform analysis.
- Data columns obtained from SpaceX Rest API are:
  - Flight number, Date, Booster Version, PayloadMass, Orbit, LaunchSite, Outcome, Flights, Gridfins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.
- Data columns obtained by using Wikipedia Web scraping:
  - FlightNo, LaunchSite, Payload, PayloadMass, Orbit, Customer, Launch Outcome, Version Booster, Booster Landing, Date, Time.
- Source code: https://github.com/j-ph/datasciencecoursera/blob/758a9d5b87d95e183d6483fbce73d9b367b99b56/SpaceX%20-%20Data%20Collection.ipynb

Data Collection – SpaceX API



Request Falcon 9 launch data from Wikipedia

Create a BeautifulSoup object from the HTML response

Extract column names from the HTML table header

Collect data by parsing HTML tables

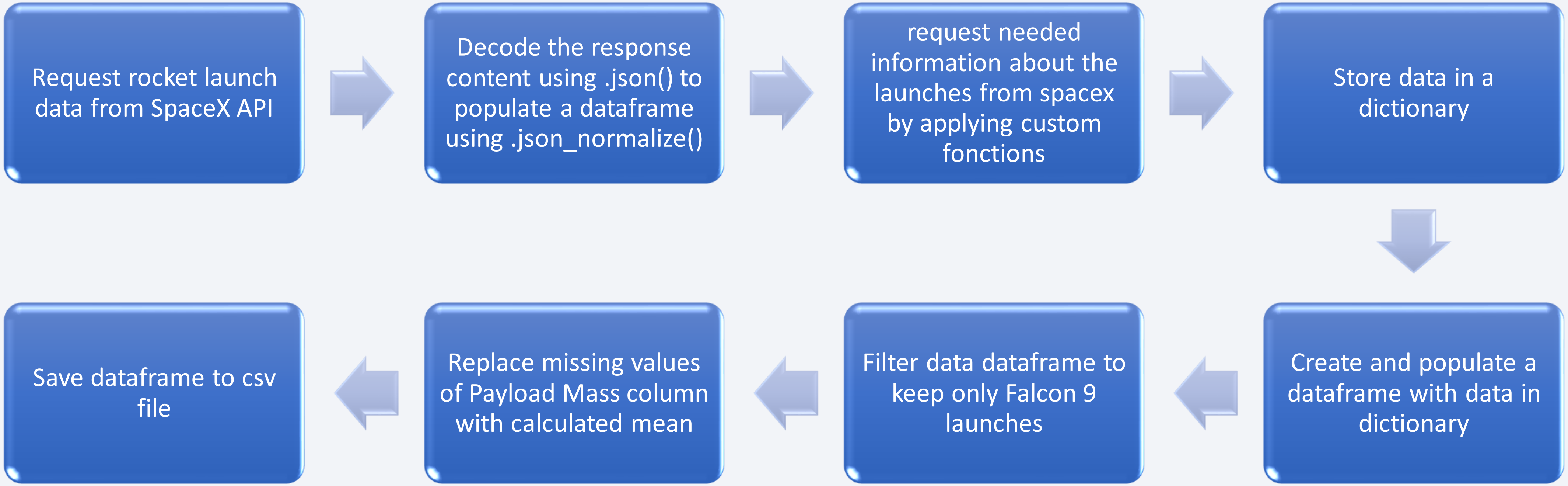
Store collected data into a dictionary

Create a dataframe from the dictionary

Store collected data into a dictionary

Export data to csv file

Data Collection - Scraping


Request rocket launch data from SpaceX API

Decode the response content using .json() to populate a dataframe using .json_normalize()

request needed information about the launches from spacex by applying custom fonctions

Store data in a dictionary

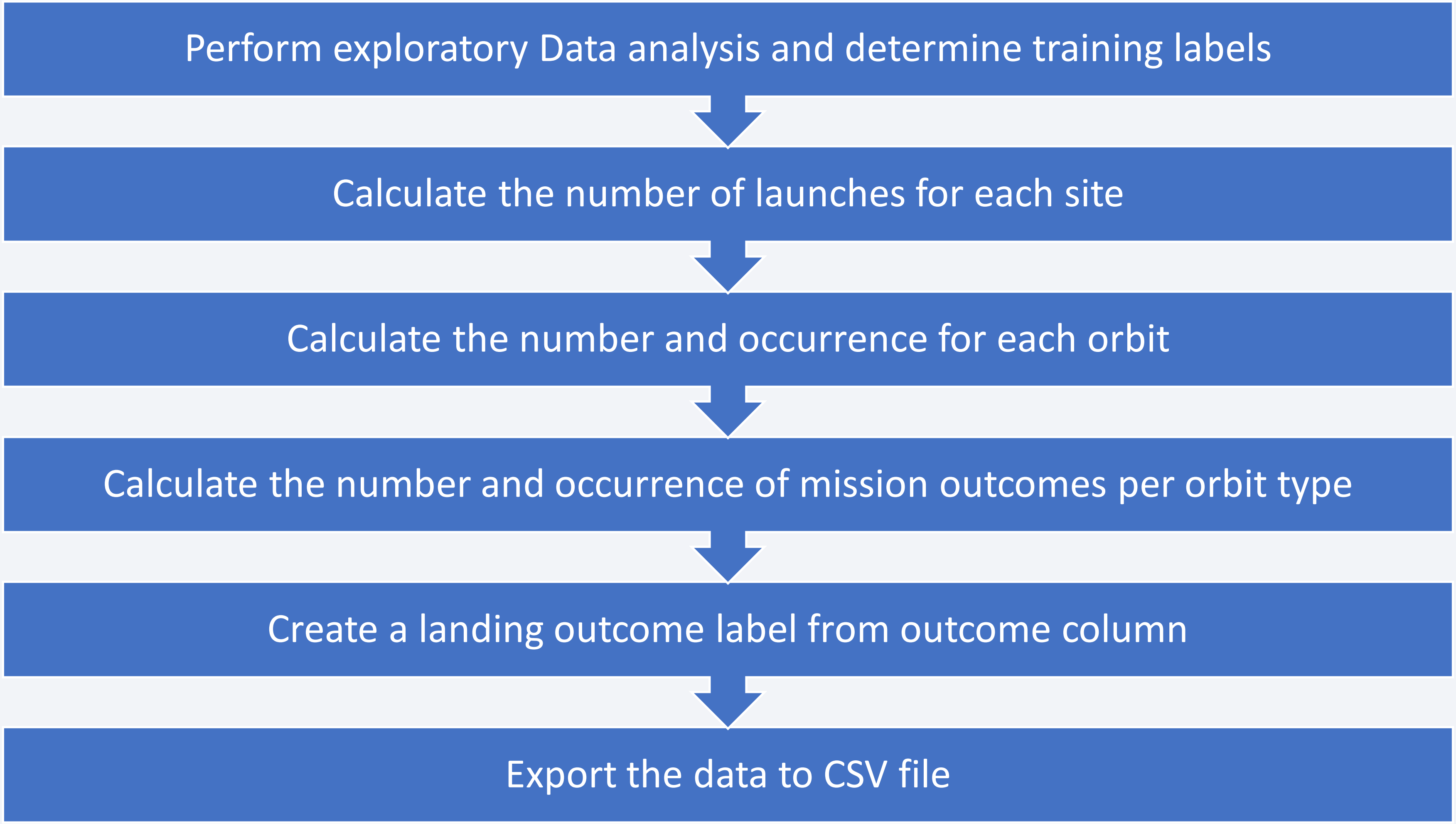
Create and populate a dataframe with data in dictionary

Filter data dataframe to keep only Falcon 9 launches

Replace missing values of Payload Mass column with calculated mean

Save dataframe to csv file

Data Wrangling



Perform exploratory Data analysis and determine training labels

Calculate the number of launches for each site

Calculate the number and occurrence for each orbit

Calculate the number and occurrence of mission outcomes per orbit type

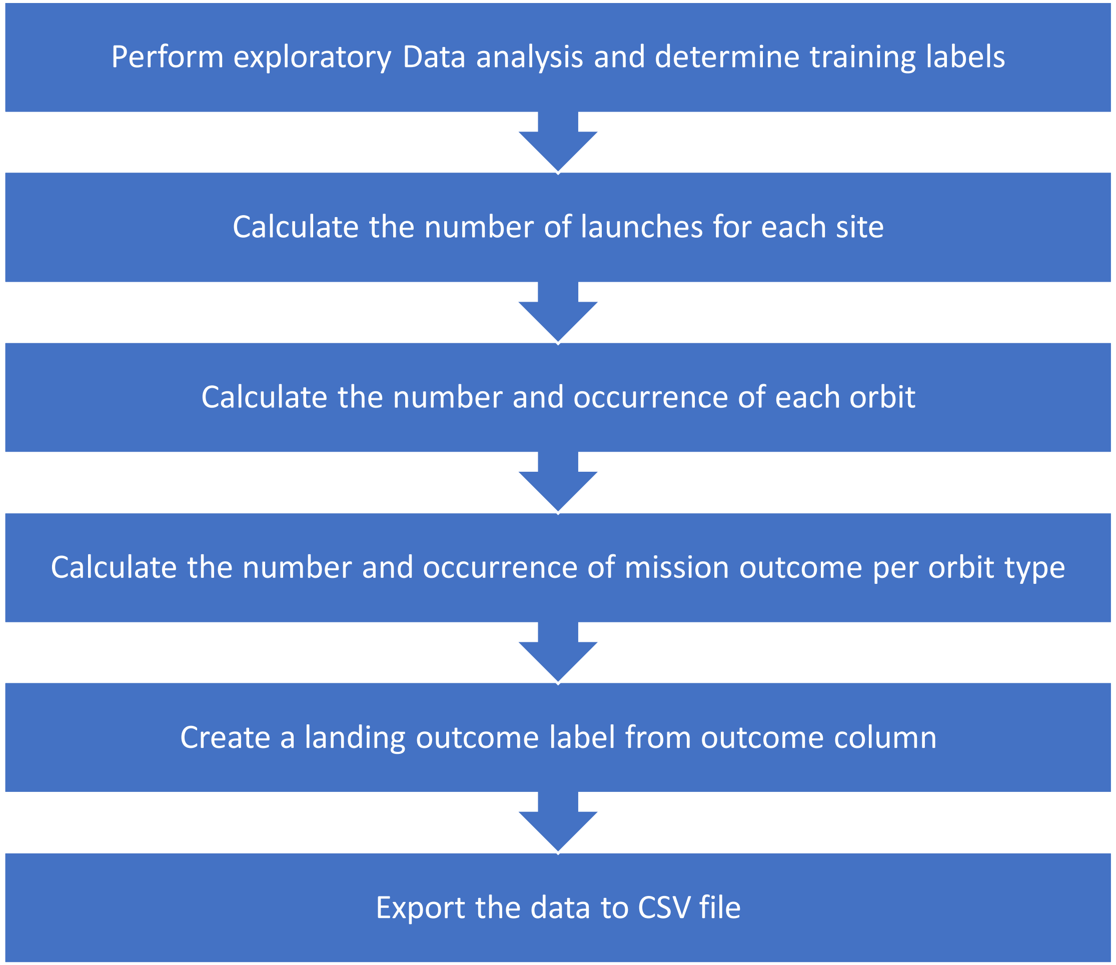
Create a landing outcome label from outcome column

Export the data to CSV file

Data wrangling

All these outcomes are converted into a binary indicator where ‘1’ means success and ‘0’ is a failure.

Source code: https://github.com/j-ph/datasciencecoursera/blob/758a9d5b87d95e183d6483fbce73d9b367b99b56/SpaceX%20-%20Data%20wrangling.ipynb



Perform exploratory Data analysis and determine training labels

Calculate the number of launches for each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from outcome column

Export the data to CSV file

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose
- Source code: https://github.com/j-ph/datasciencecoursera/blob/758a9d5b87d95e183d6483fbce73d9b367b99b56/SpaceX%20-%20Data%20Visualization.ipynb

EDA with SQL

SQL queries are used to find the following answers:

- Display unique names of launch sites
- Display 5 records where the launch site name begins with ‘CCA’
- Calculate the total payload mass carried by boosters launched by NASA (CRS)
- Calculate the average payload mass carried by booster version F9 v1.1
- Give the date when the first successful landing outcome in the ground pad was achieved
- List the names of the boosters which have success in drone ships and have payload mass greater than 4000 kg but less than 6000 kg

EDA with SQL

- List the total number of successful and failed mission outcome
- List the name of booster versions which have carried the maximum payload mass
- List the failed landing outcomes in drone ship and their version and launch site name for the month in year 2015
- Rank the count of landing outcomes failed and successful for the period between 2010—6-04 and 2017-03-20 in descending order

- Source code: https://github.com/j-ph/datasciencecoursera/blob/758a9d5b87d95e183d6483fbce73d9b367b99b56/Spacex%20-%20SQL%20%20EDA.ipynb

Build an Interactive Map with Folium

- We marked all launch sites and added map objects such as markers, circles, and lines to mark the success or failure of launches for each site on the folium map.
- We assign the feature launch outcome (0 or 1: failure or success)
- On the map launch outcome is visualized with color marker clusters, we can appreciate ===
- We calculate the distances between a launch site to its proximities to be able to answer some questions like:
  - What is the distance between the launch site and the cities around it?
  - Are launch sites close to railways, highways, and coastlines?
- Source code: https://github.com/j-ph/datasciencecoursera/blob/758a9d5b87d95e183d6483fbce73d9b367b99b56/SpaceX%20-%20Interactive%20Map.ipynb

Build an interactive map with Folium

- Folium markers show the SpaceX launch sites and their nearest important landmarks like railways, highways, towns, and coastlines.

- Polylines are used to connect launch sites to their nearest landmarks.

- Folium circles are used to circle areas of launch sites.

- To distinguish successes from failures for each site, marker clusters are displayed on the map. Red means failure while green means success.

Build a Dashboard with Plotly Dash

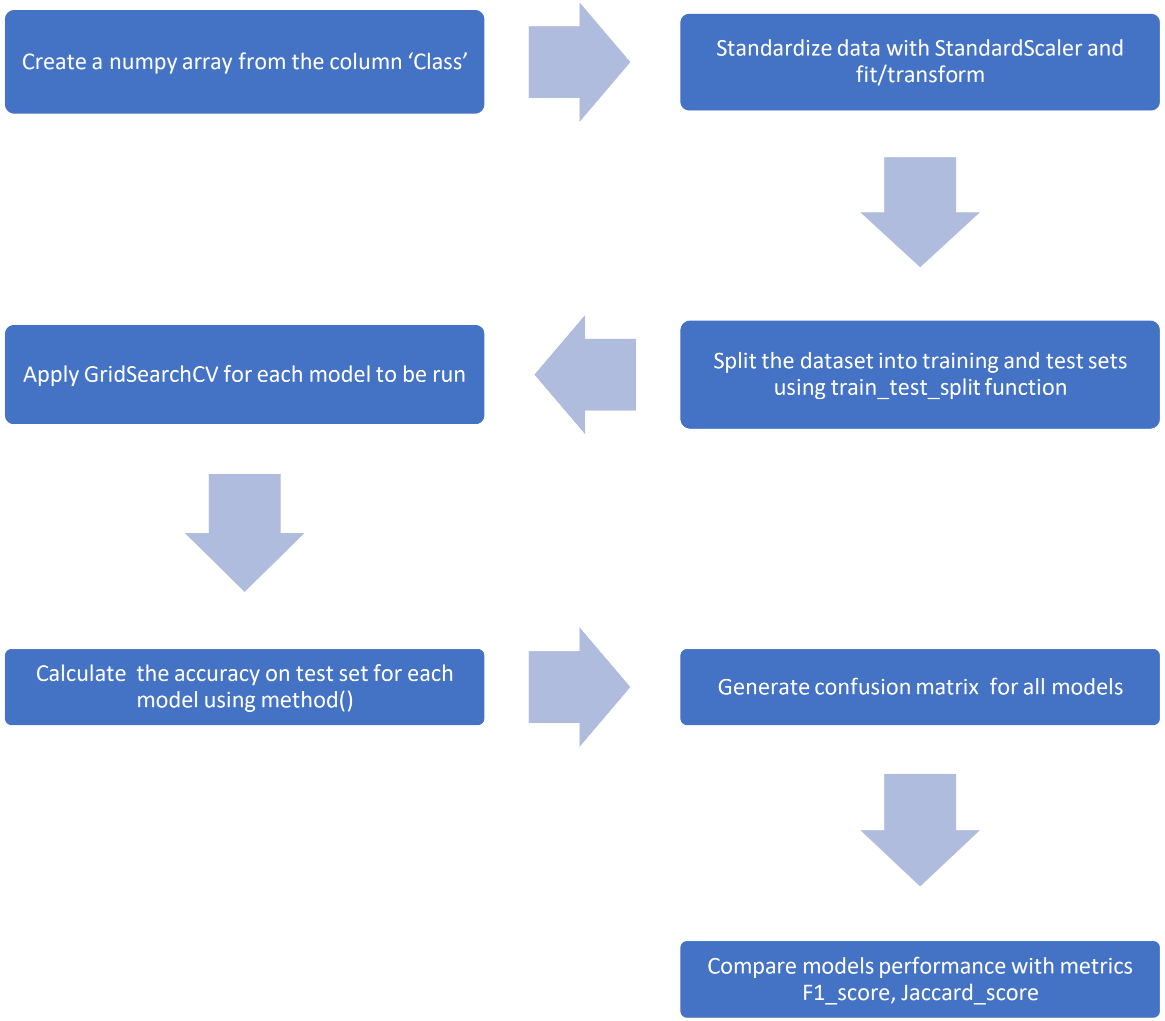
- Launch sites dropdown list:
  - Added a dropdown list to enable launch site selection
- Pie chart showing successful launches:
  - Added a pie chart to show the total successful launches count for all sites and also the success vs failed counts for the site if a specific launch site is selected.
- Slider of payload mass range:
  - Added a slider to select the payload range
- Scatter chart of payload mass vs success rate for the different booster versions:
  - Added a scatter chart to show the correlation between Payload and Launch success.
- Source code: https://github.com/j-ph/datasciencecoursera/blob/6d2f00ec28ac11d0dc8b1207f59e57864493934d/SpaceX%20-%20Dashboard%20(1).py

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Source code: https://github.com/j-ph/datasciencecoursera/blob/758a9d5b87d95e183d6483fbce73d9b367b99b56/SpaceX%20-%20Machine%20Learning%20Prediction.ipynb

Predictive analysis (Classification)

- Sickit-Learn is the library used for predictive analysis
- Split the dataset into a training set and test set
- Use a GridSearchCV to improve results
- Compare the predictions



Create a numpy array from the column 'Class'

Standardize data with StandardScaler and fit/transform

Split the dataset into training and test sets using train_test_split function

Apply GridSearchCV for each model to be run

Calculate the accuracy on test set for each model using method()

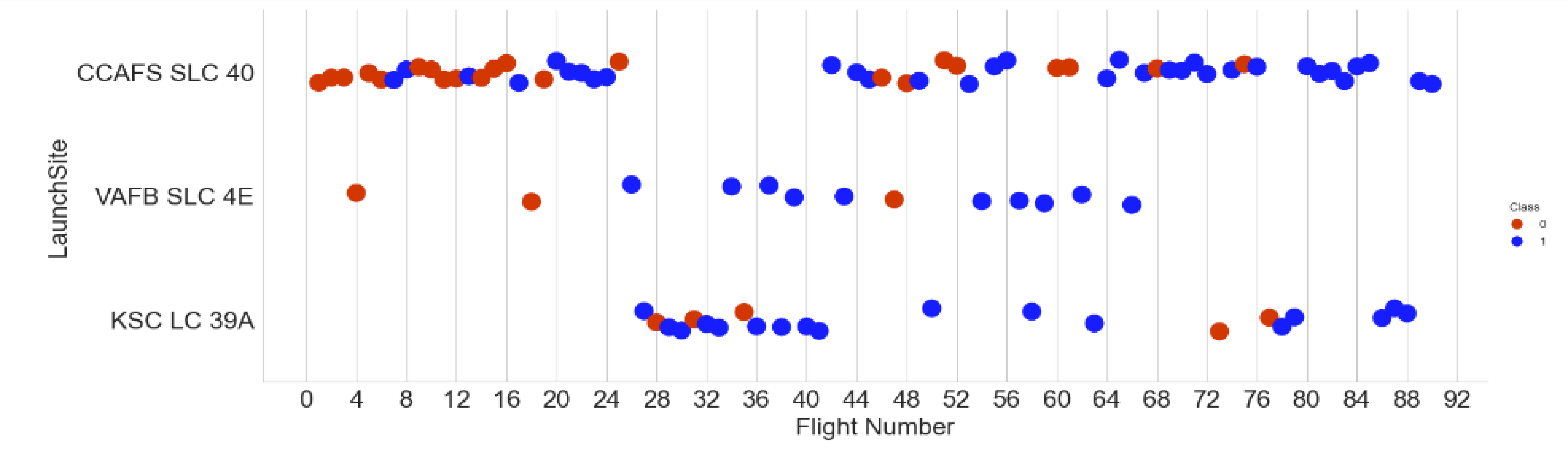
Generate confusion matrix for all models

Compare models performance with metrics F1_score, Jaccard_score

Section 2

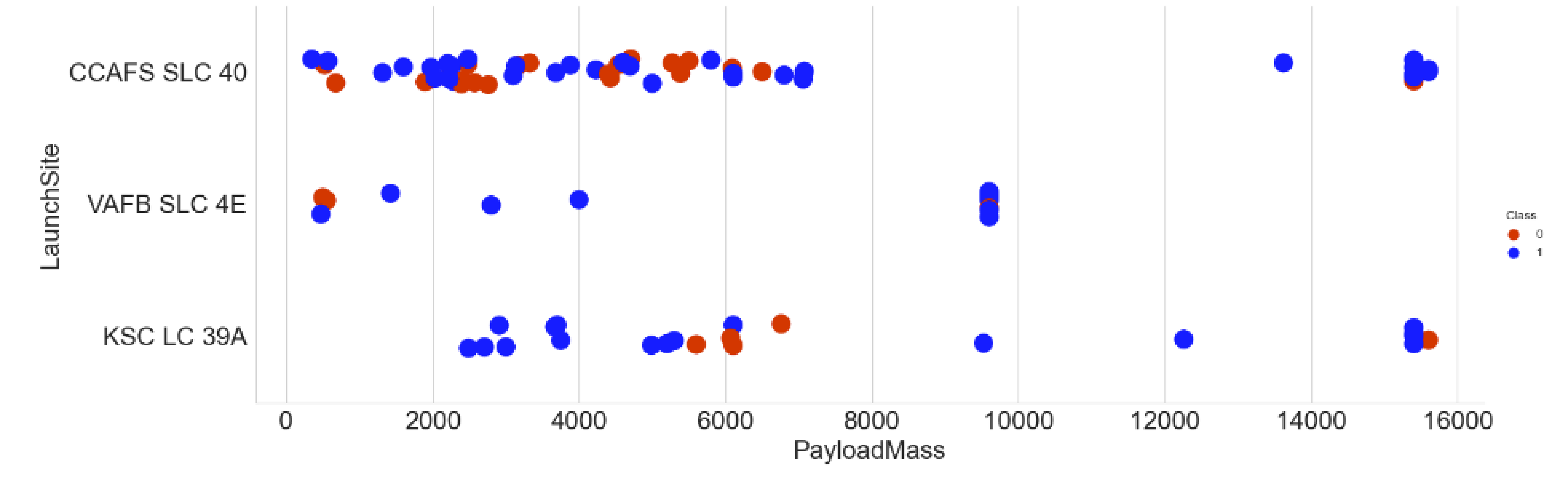
Insights drawn from EDA

Flight Number vs. Launch Site



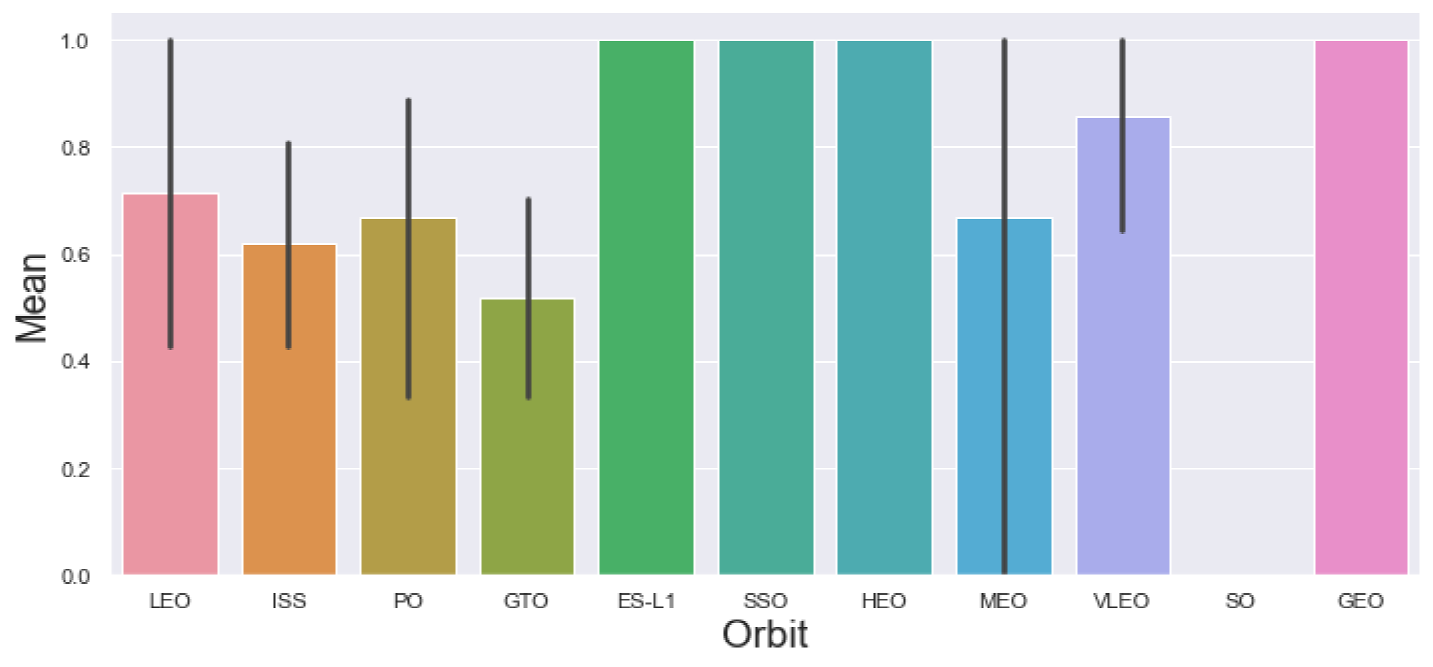
- The earliest flights have failed while latest flights succeeded
- The CCFAS SLC 40 launch site has about the half of all launches
- VAFB SLC 4E and KSC LC 39A sites have higher success rates
- New launches have a higher success rate

Payload vs. Launch Site



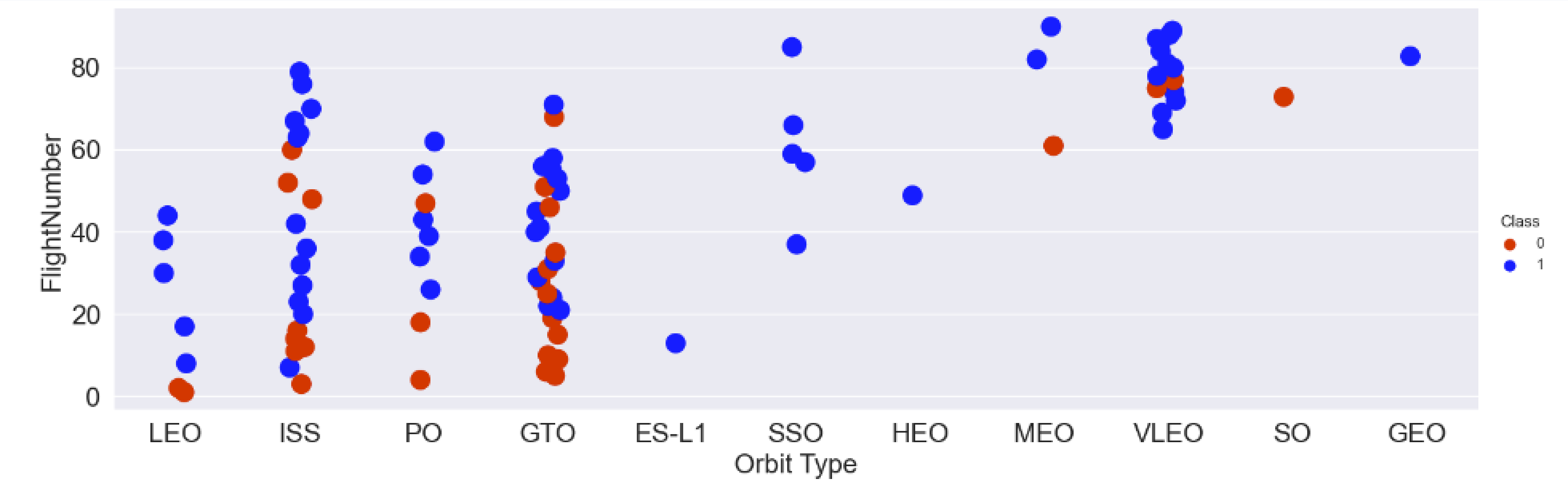
- for every launch site the higher the payload mass, the higher success rate is.
- Most of the payload mass above 7000 kg are successful.
- KSC LS 39A launch site has a 100% success rate with payload mass under 5500 kg.

Success Rate vs. Orbit Type



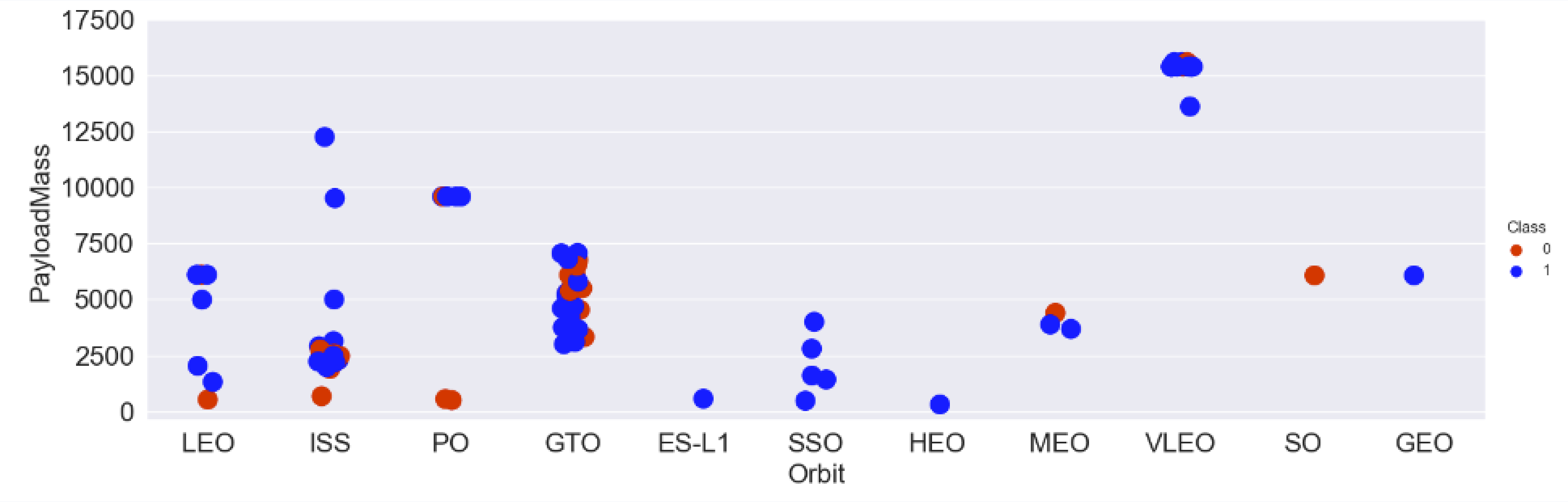
- Orbits with 100% success rate are: ES-L1, HEO, LEO, SSO.
- Conversely orbit with a 0% success rate is SO.
- Orbits with a success rate between 50% and 85% are GTO, ISS, LEO, MEO and PO.

Flight Number vs. Orbit Type



- For LEO orbit, success rate increase with number of flights. For orbits like GTO, there is no relation between number of flights and success rate.
- For SSO orbit, we can suppose it benefits from the experience of previous flights.

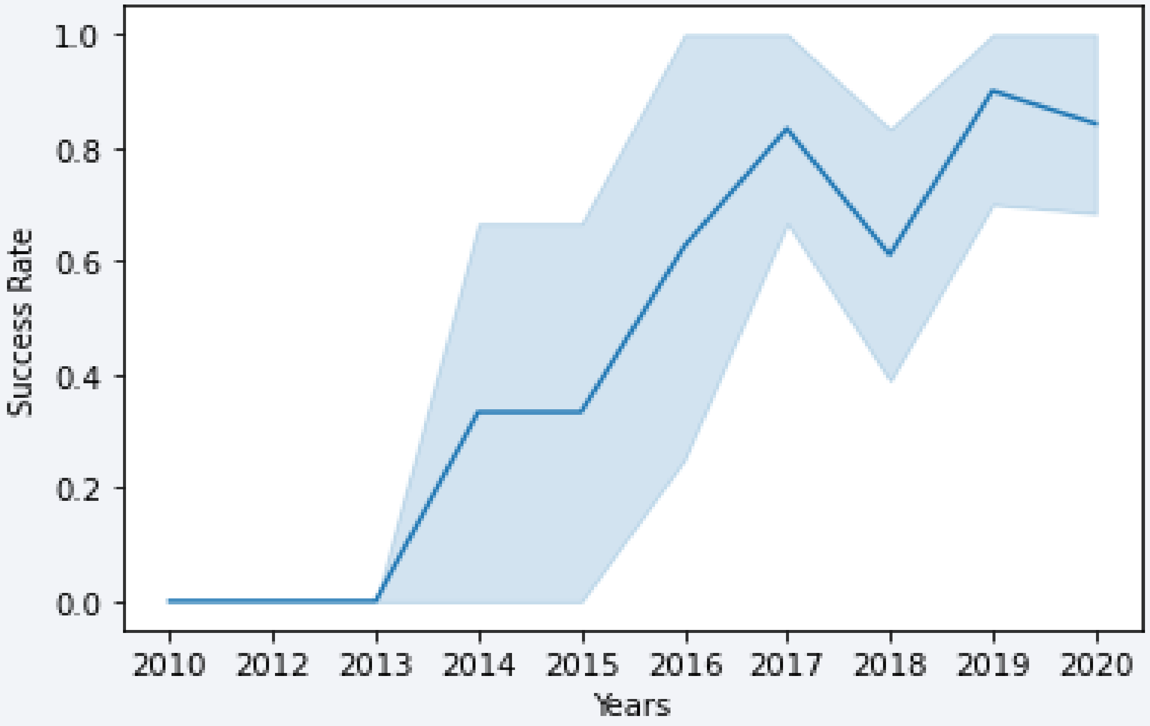
Payload vs. Orbit Type



- Heavy payloads have negative influence on GTO orbits and a positive one on Polar LEO (ISS) and GEO.

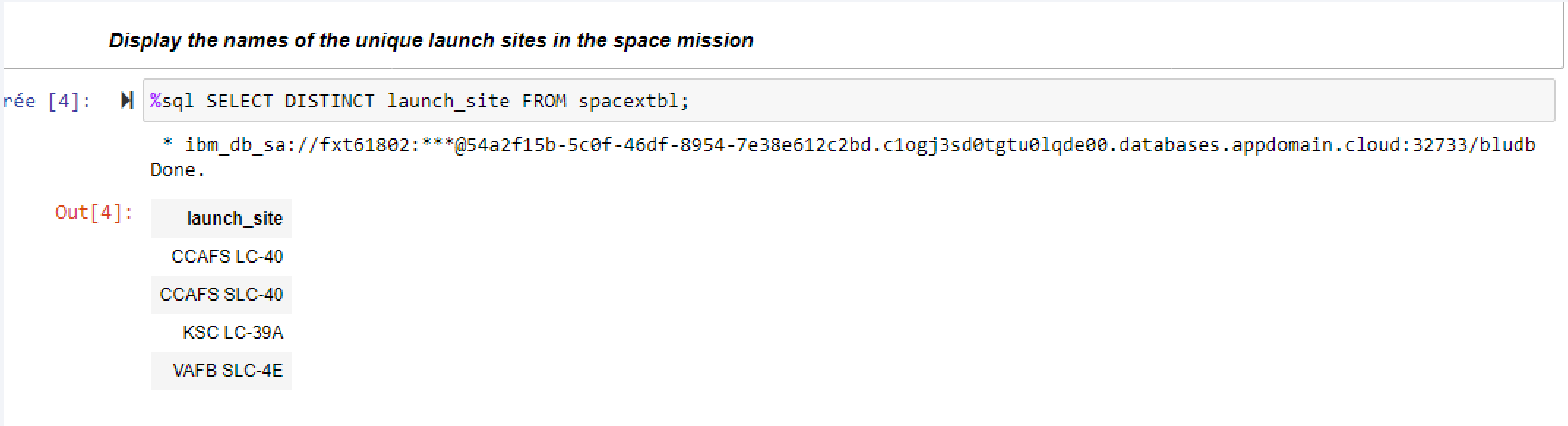
Launch Success Yearly Trend

- Success rate kept increasing since 2013 until 2020



All Launch Site Names

- The list display unique names of launch sites, all duplicates are discarded by query with “distinct” clause on column LaunchSite.



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

```
rée [4]:  %sql SELECT DISTINCT launch_site FROM spacextbl;
```

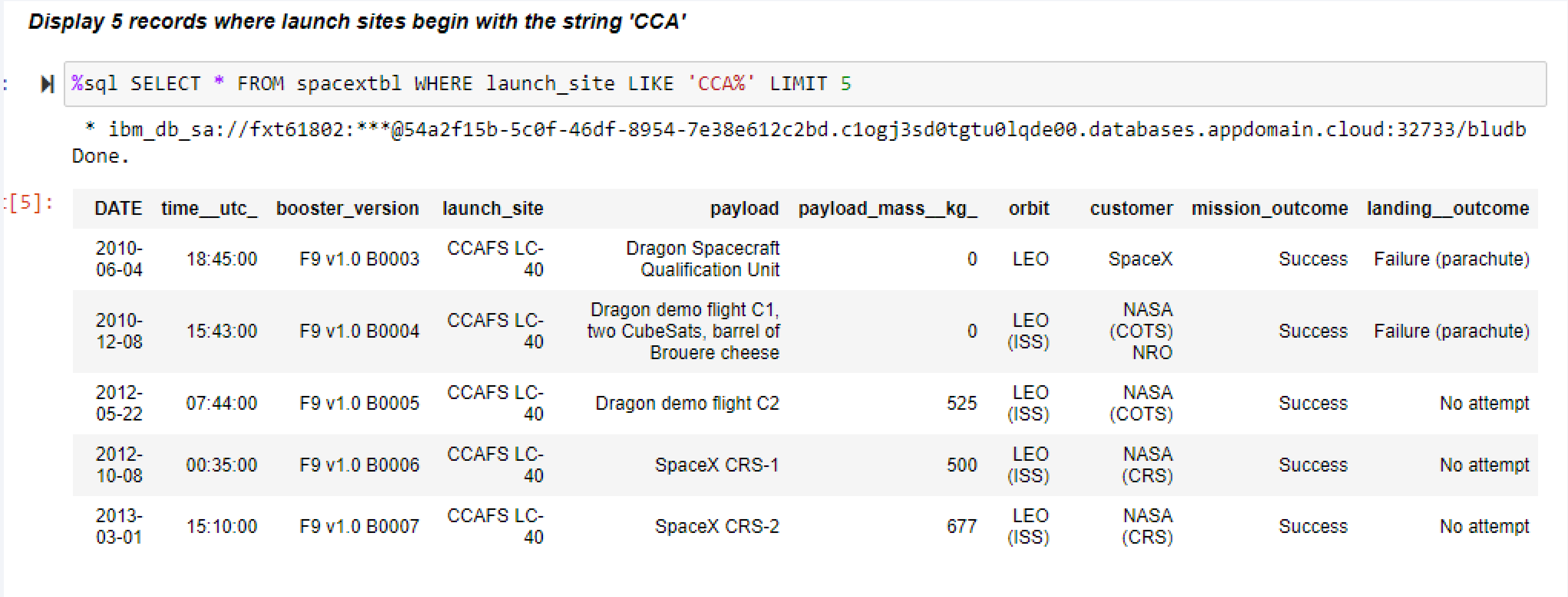
```
 * ibm_db_sa://fxt61802:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[4]:

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

Launch Site Names Begin with 'CCA'

- This query returns only 5 rows with a launch site name starting with “CCA”



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

```
%sql SELECT * FROM spacextbl WHERE launch_site LIKE 'CCA%' LIMIT 5
```

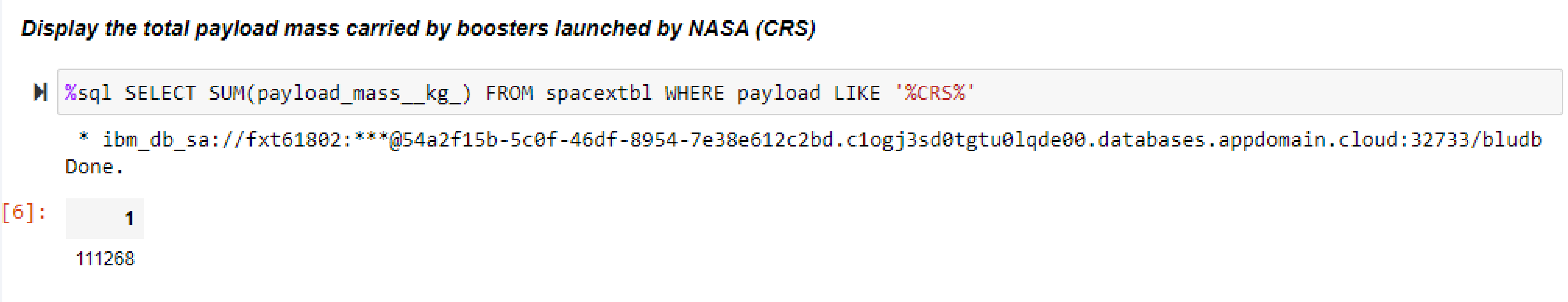
```
 * ibm_db_sa://fxt61802:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

[5]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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 (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- We calculate the total payload mass launched by NASA.



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

```
%sql SELECT SUM(payload_mass__kg_) FROM spacextbl WHERE payload LIKE '%CRS%'
```

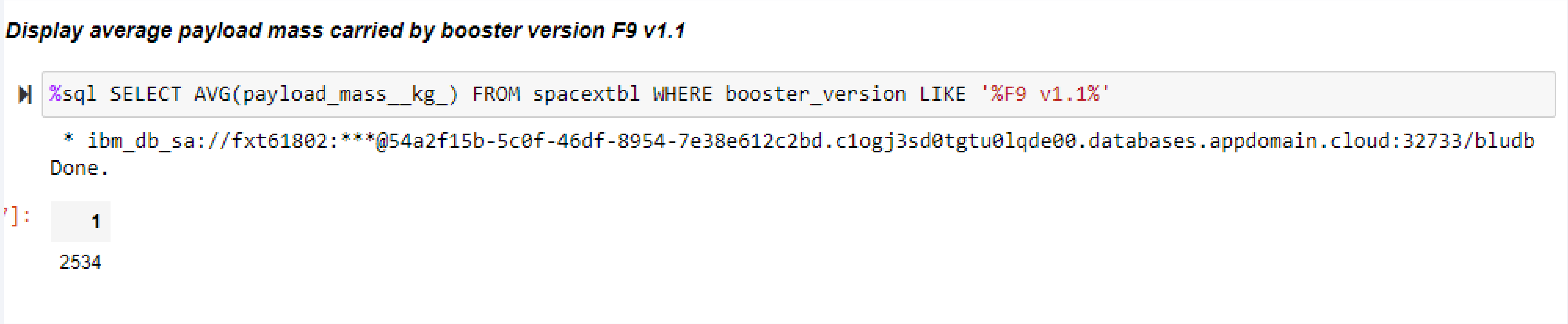
```
 * ibm_db_sa://fxt61802:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

[6]:

1
111268

Average Payload Mass by F9 v1.1

- This query displays the average payload mass carried by booster version F9 v1.1



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

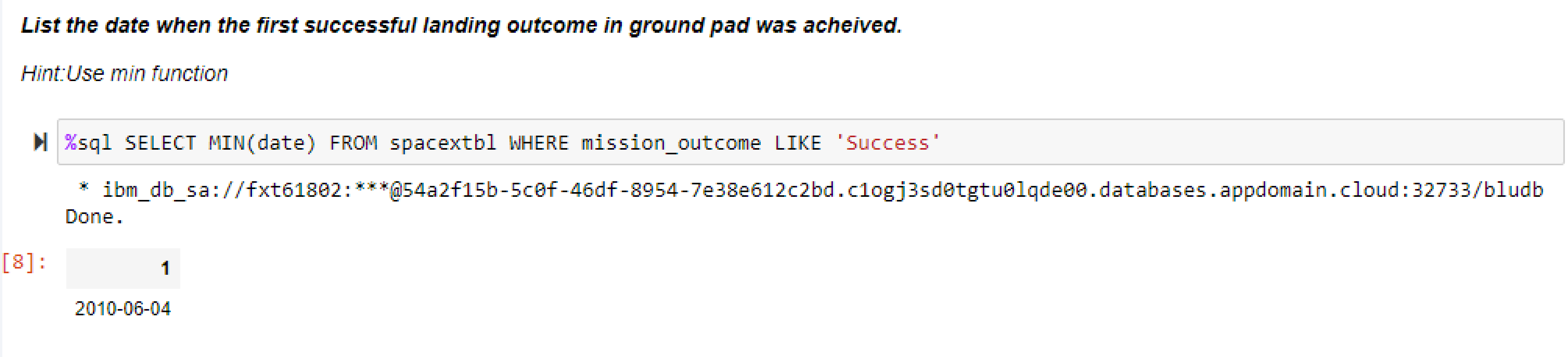
```
%sql SELECT AVG(payload_mass__kg_) FROM spacextbl WHERE booster_version LIKE '%F9 v1.1%'
```

```
 * ibm_db_sa://fxt61802:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

1
2534

First Successful Ground Landing Date

- **Find the dates** of the first successful landing outcome on a ground pad
- We find the earliest successful landing outcome on a ground pad



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

Hint:Use min function

```
%sql SELECT MIN(date) FROM spacextbl WHERE mission_outcome LIKE 'Success'
```

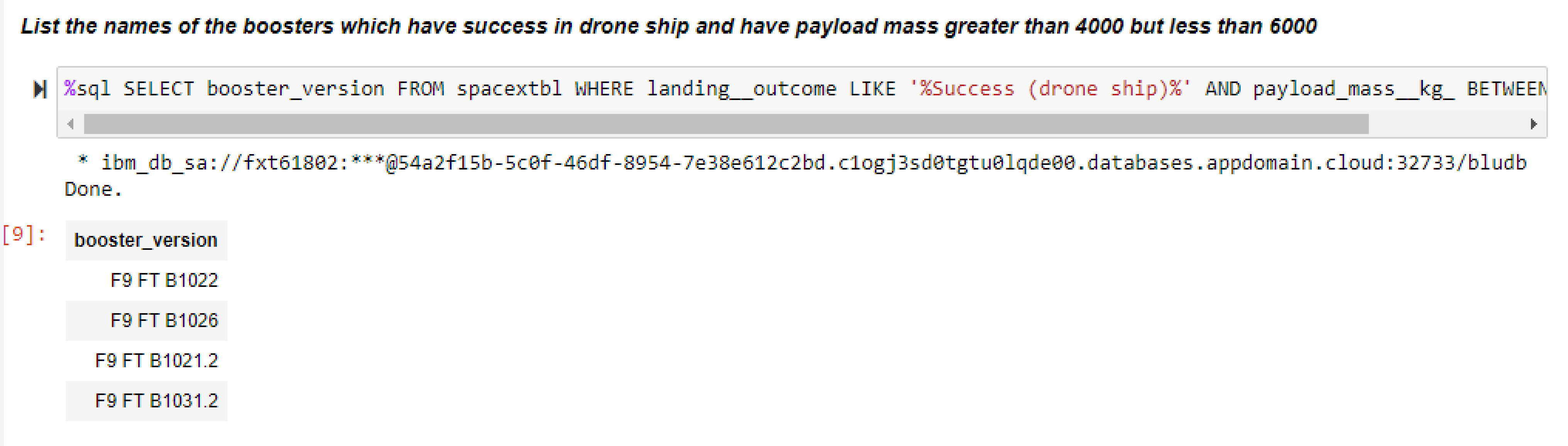
```
 * ibm_db_sa://fxt61802:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

[8]:

1
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query 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 booster_version FROM spacextbl WHERE landing__outcome LIKE '%Success (drone ship)%' AND payload_mass__kg_ BETWEEN
```

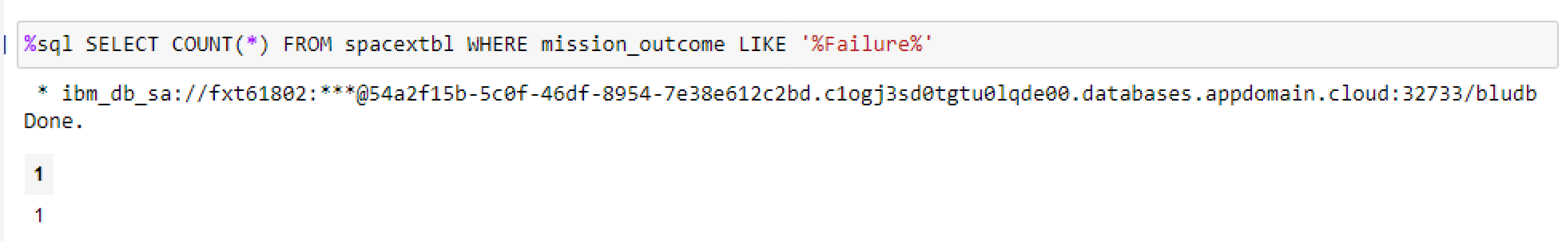
```
 * ibm_db_sa://fxt61802:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

[9]:

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

Total Number of Successful and Failure Mission Outcomes

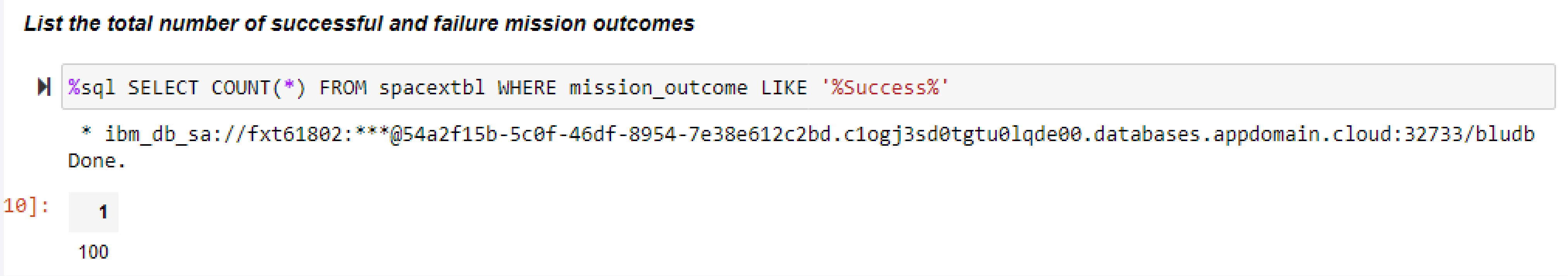
- Calculate the total number of successful and failure mission outcomes. It is done in 2 queries. First one for failure and second one for success.



```
%sql SELECT COUNT(*) FROM spacextbl WHERE mission_outcome LIKE '%Failure%'

 * ibm_db_sa://fxt61802:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.

1
1
```



List the total number of successful and failure mission outcomes

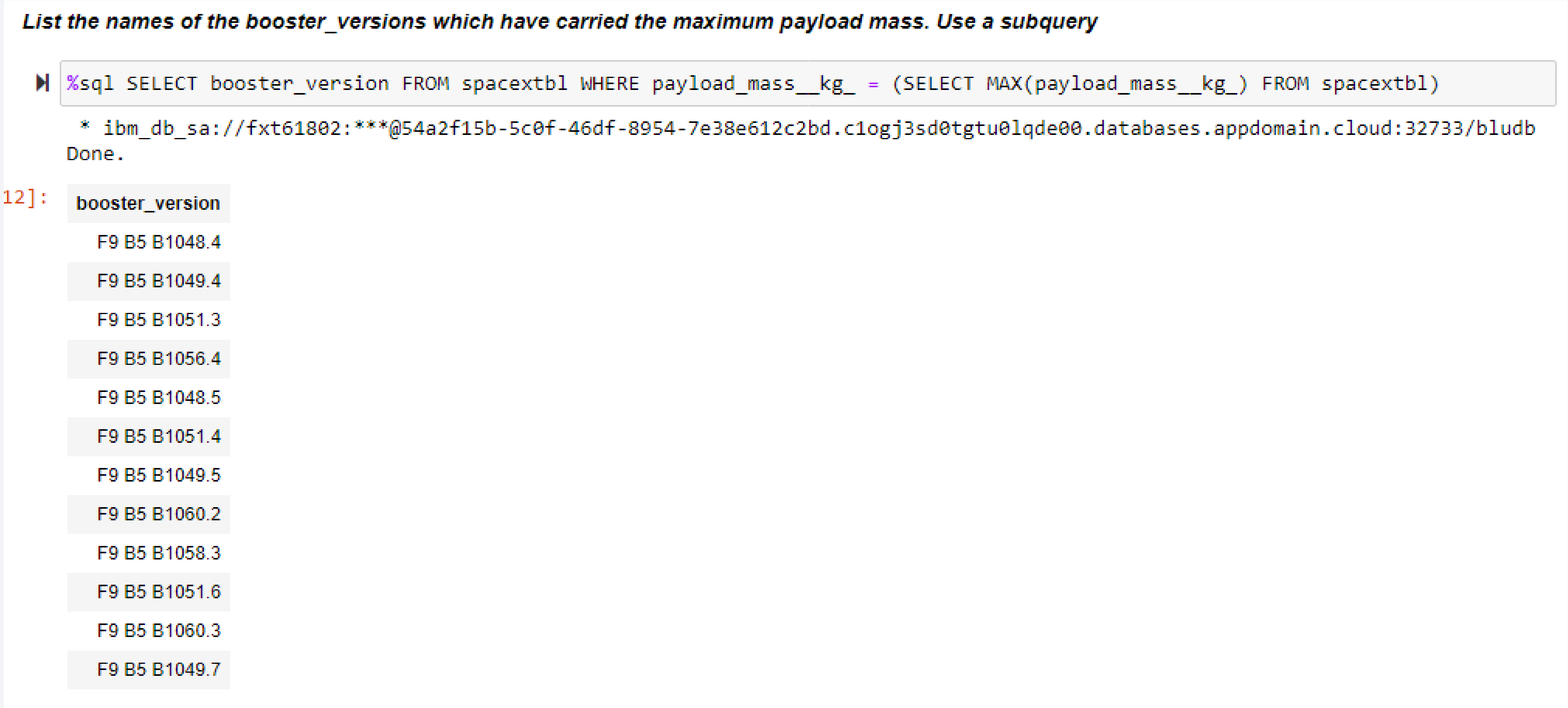
```
%sql SELECT COUNT(*) FROM spacextbl WHERE mission_outcome LIKE '%Success%'

 * ibm_db_sa://fxt61802:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.

10]:   1
     100
```

Boosters Carried Maximum Payload

- we determined the booster that have carried the maximum payload mass by using a subquery that first retrieve the maximum Payload Mass and then retrieve all Booster Version matching with that number.



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

```
%sql SELECT booster_version FROM spacextbl WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM spacextbl)
```

```
 * ibm_db_sa://fxt61802:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

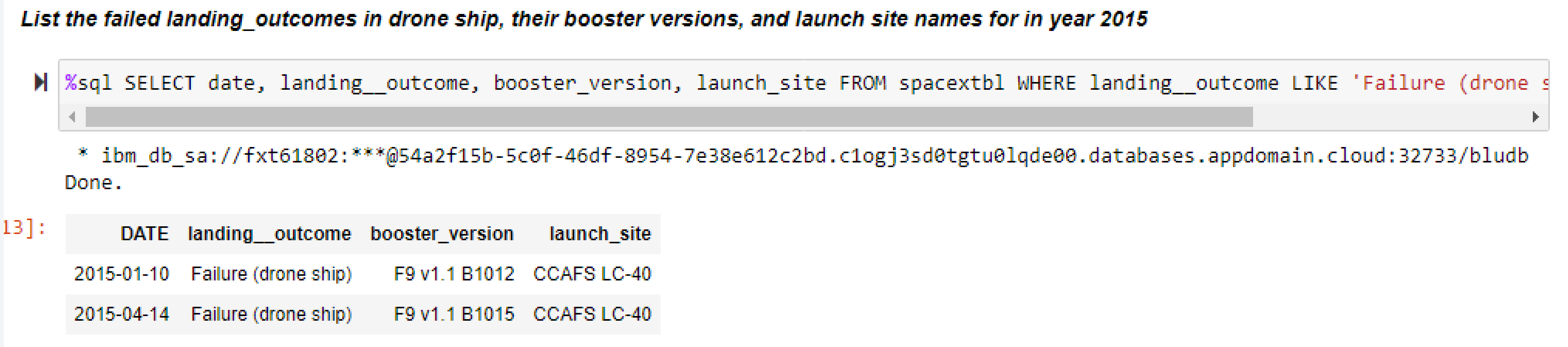
12]:

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

- Present your query result with a short explanation here



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

```
%sql SELECT date, landing__outcome, booster_version, launch_site FROM spacextbl WHERE landing__outcome LIKE 'Failure (drone s
```

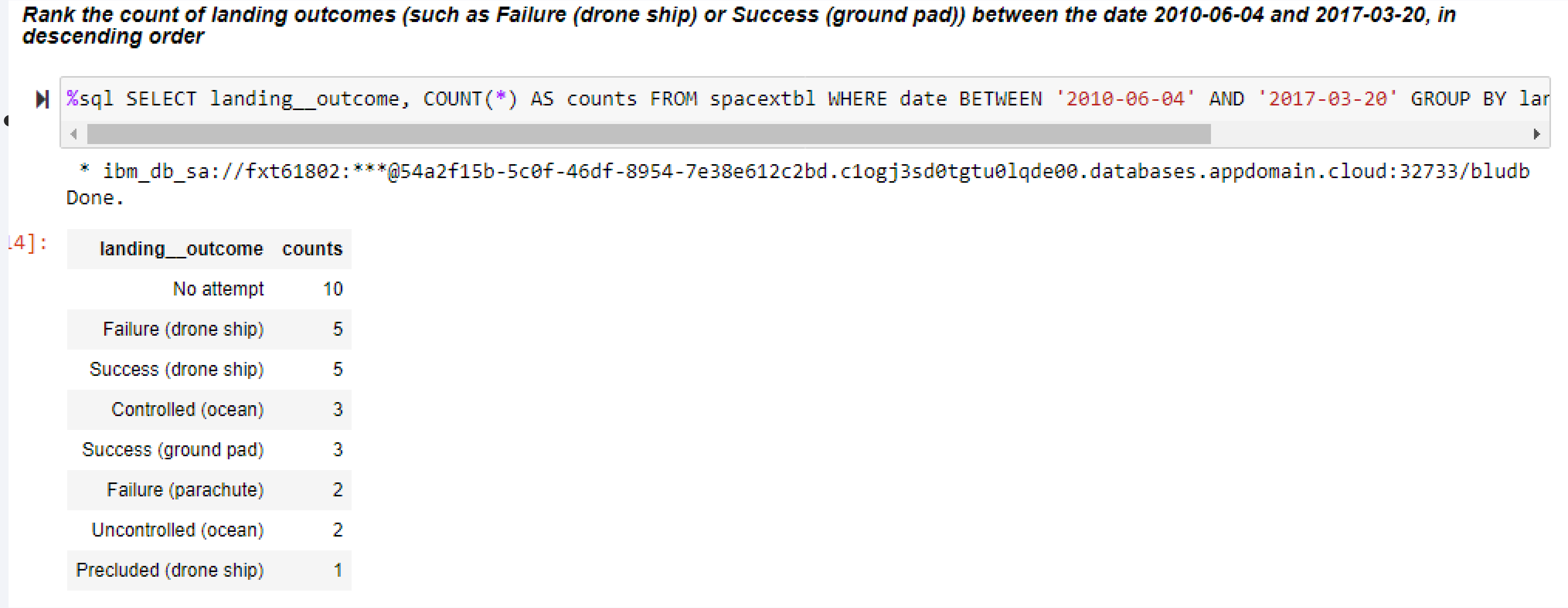
```
 * ibm_db_sa://fxt61802:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

13]:

DATE	landing__outcome	booster_version	launch_site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- For each type of Landing Outcome, the query counts the 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 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 SELECT landing__outcome, COUNT(*) AS counts FROM spacextbl WHERE date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY lar
```

```
 * ibm_db_sa://fxt61802:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

14]:

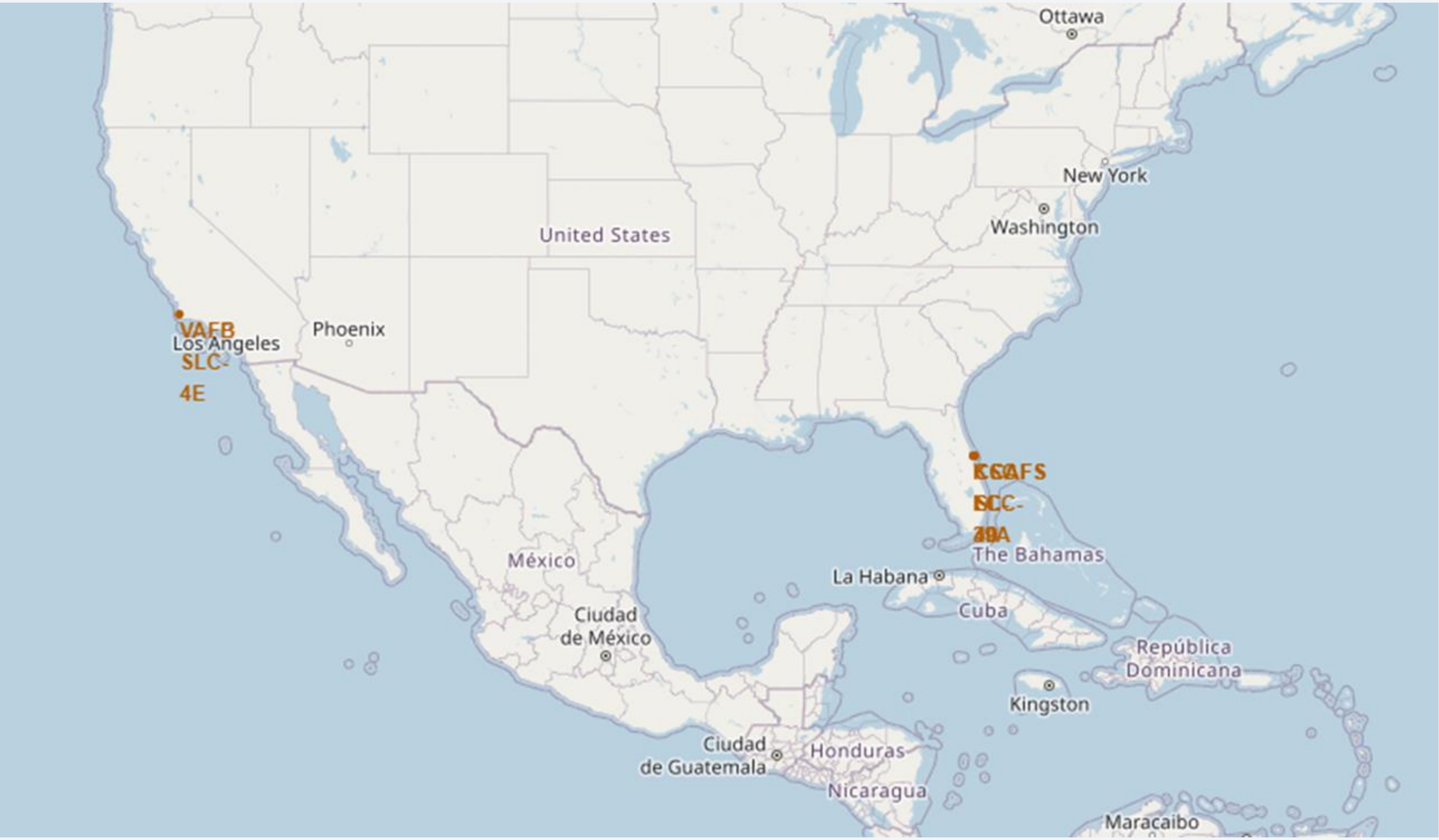
landing__outcome	counts
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

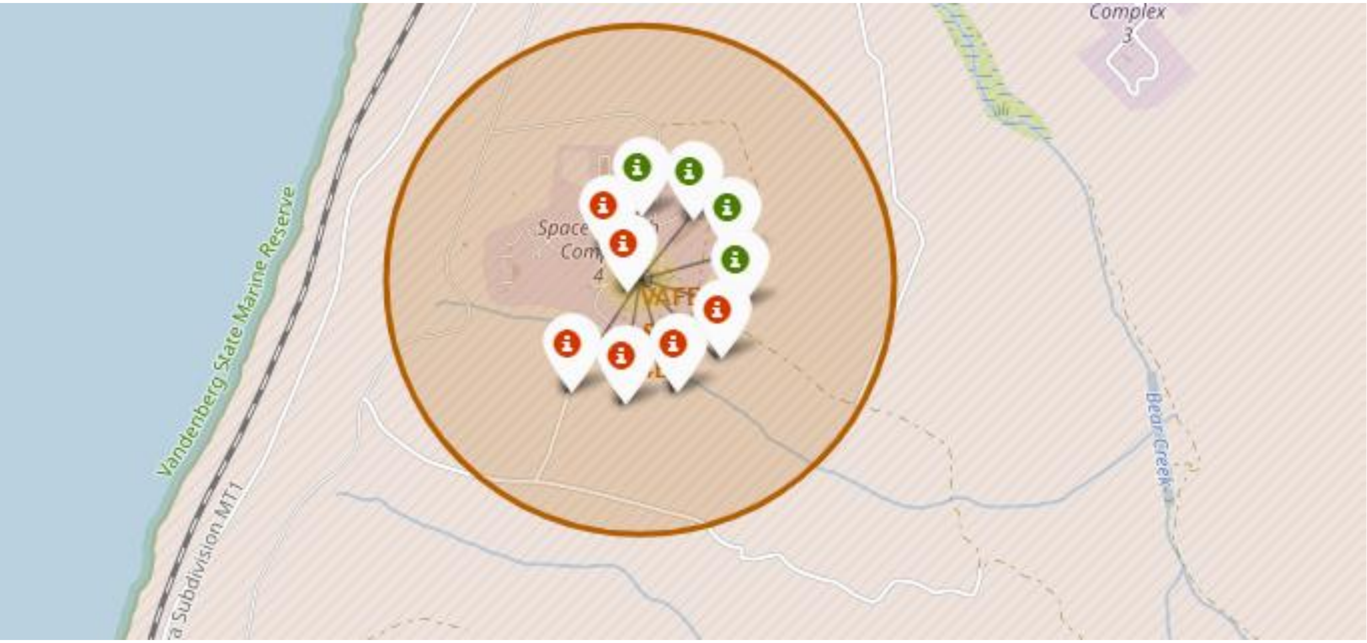
Launch Sites locations

- SpaceX uses a launch site on the Atlantic coast of Florida
- Another site is used on the Pacific coast of southern California.
- Both launch sites are near the sea. For safety reason, rocket launches and booster landings are done toward the ocean.



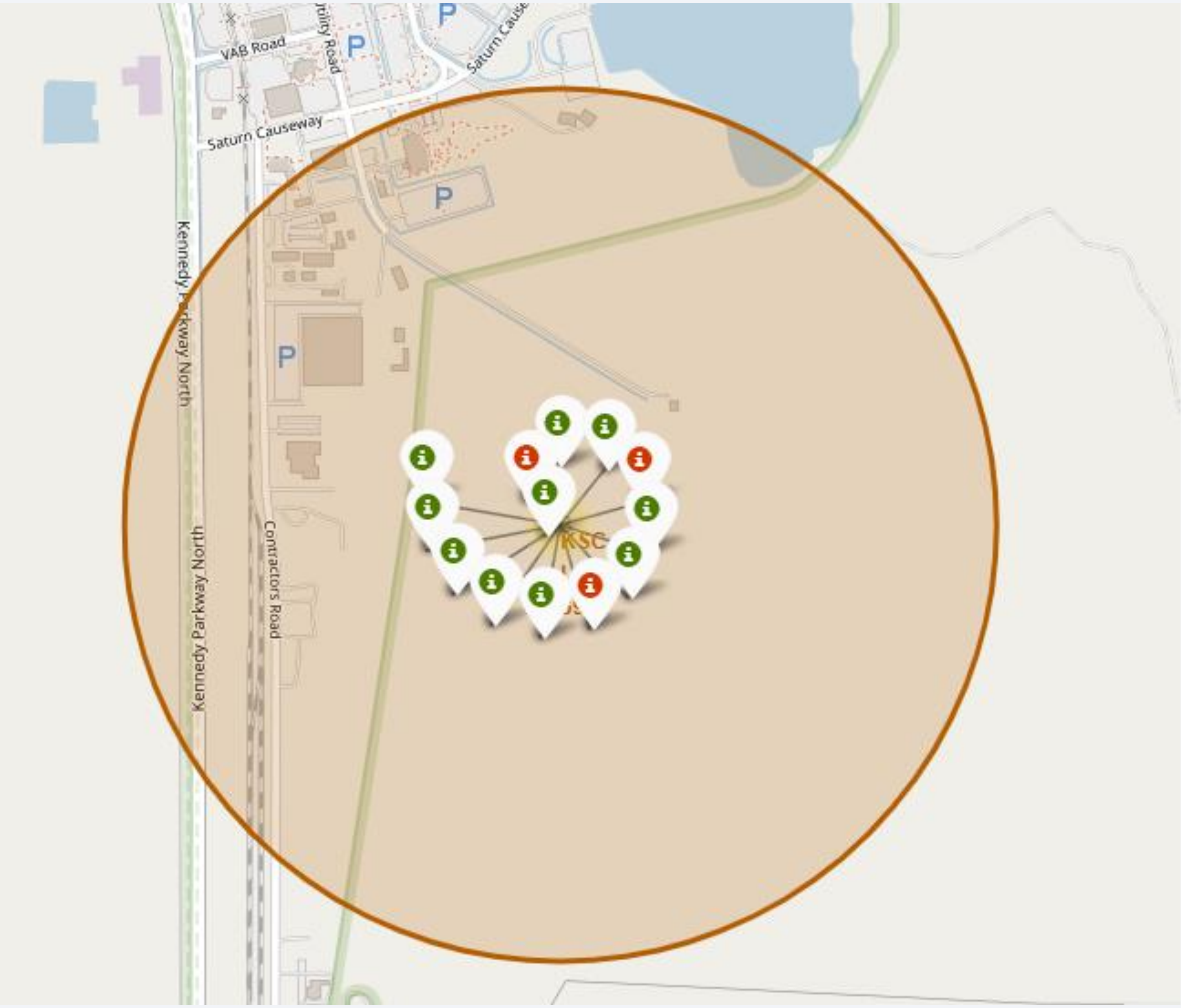
VAFB SLC-4E Recovery outcomes

- Red flag - Failure
- Green flag - Success



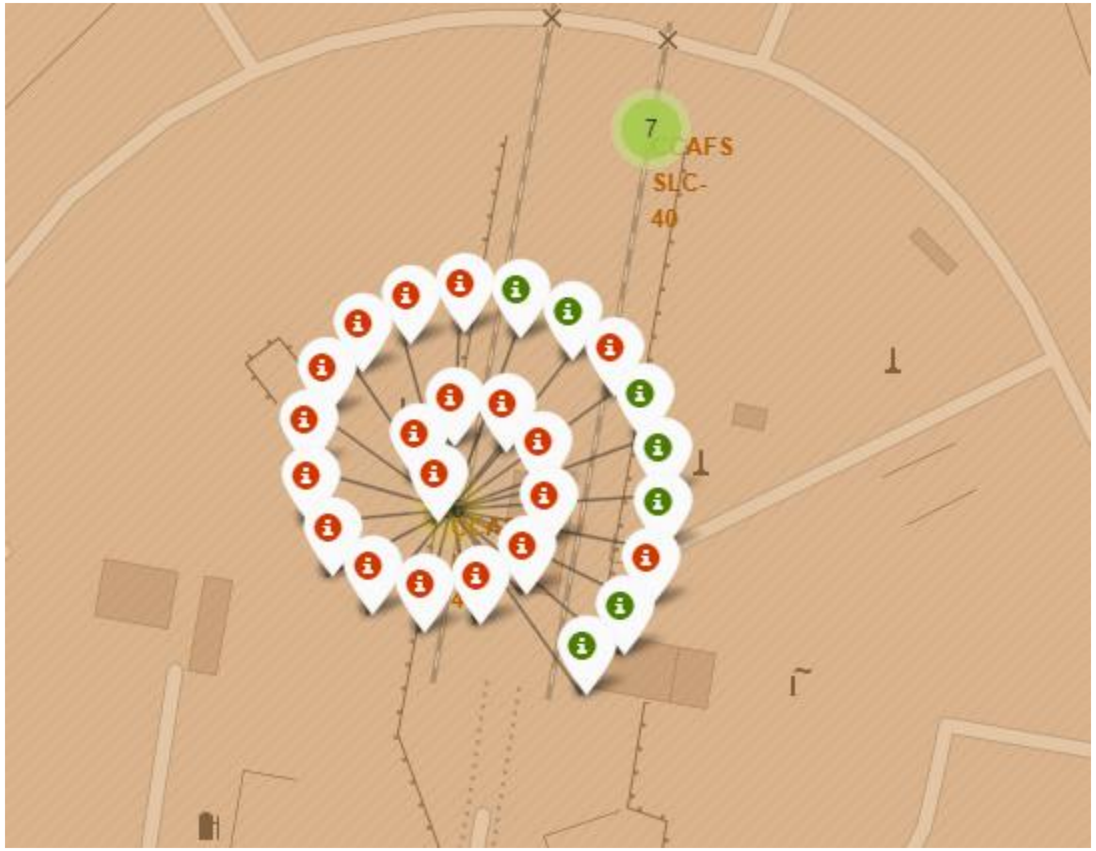
KSC LC-39A Recovery outcomes

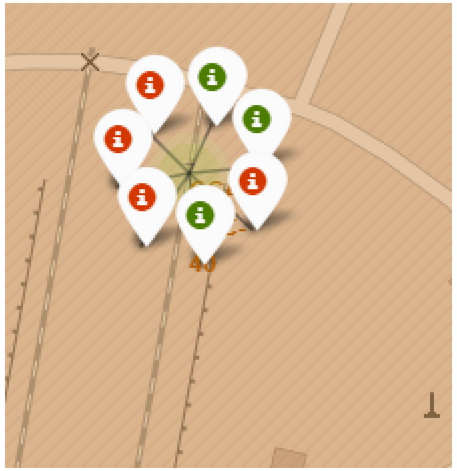
- KSC LC-39A is the launch site with the highest rate of success



CCAFS LC-40 and SLC-40 Recovery outcomes

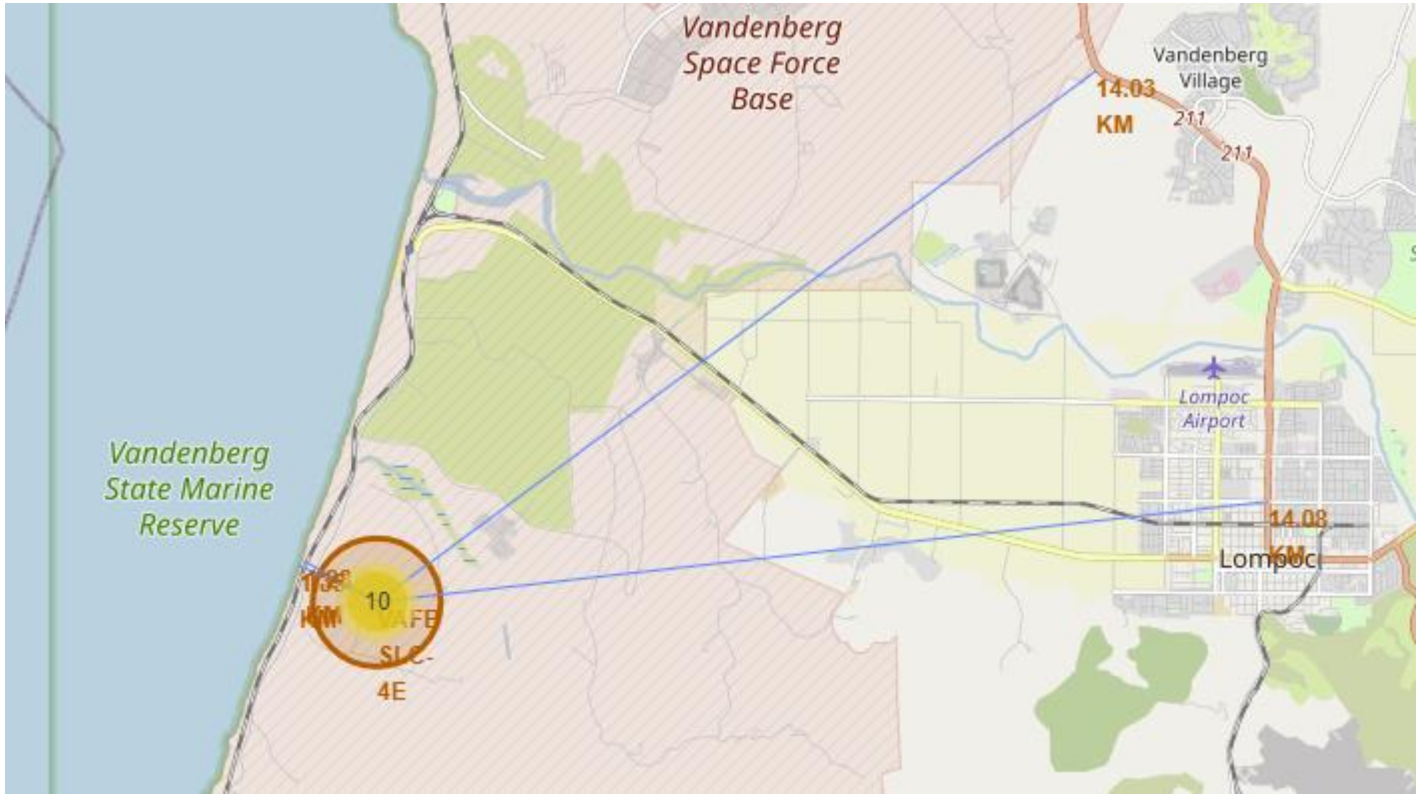
- Site CCAFS SLC-40 is not much used with 7 launches in contrast to site CCAFS LC-40 with 26 launches.
- Success rate for CCAFS SLC-40 is low.





VAFB SLC-4E Nearby Locations

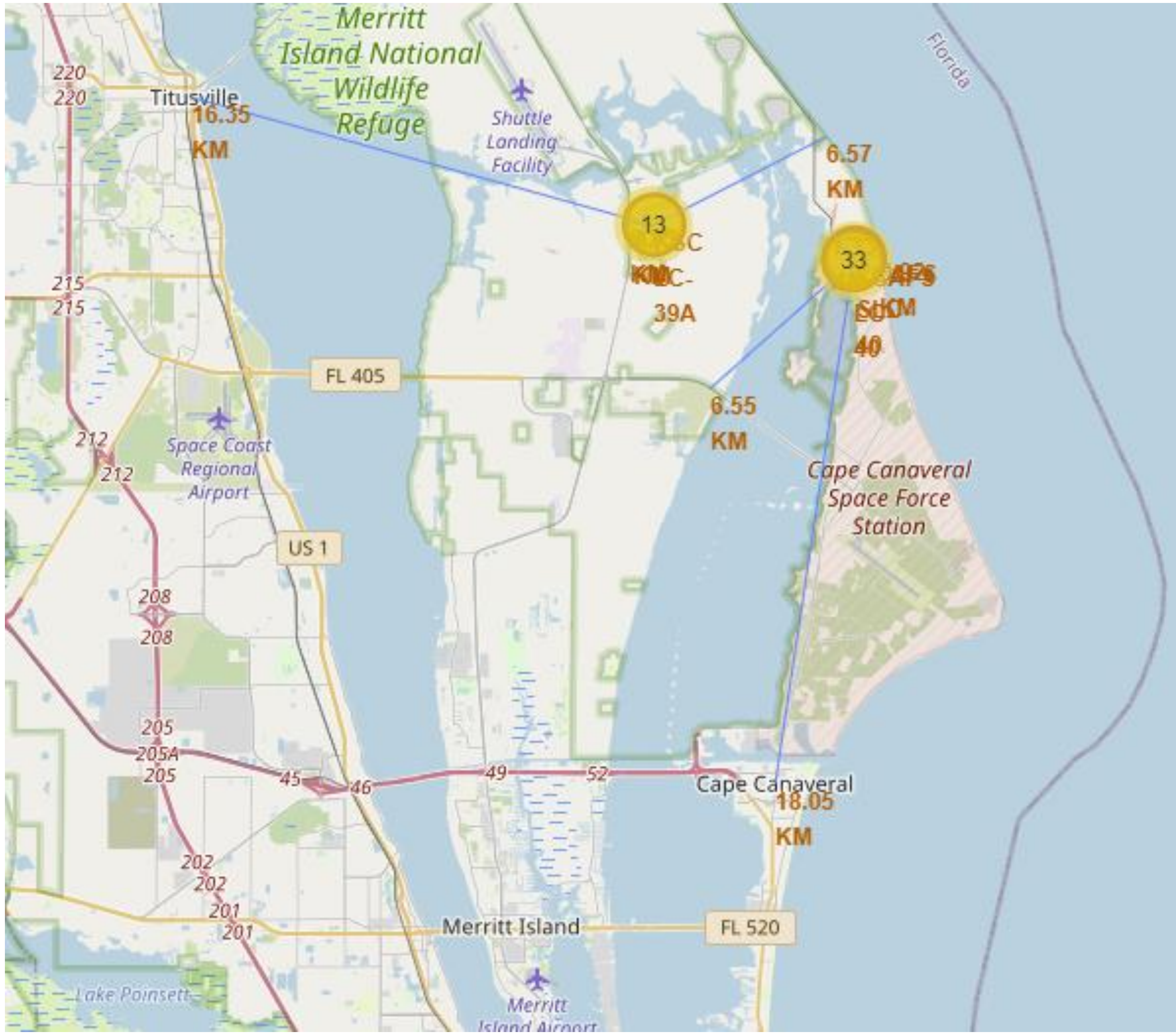
- Launch site is at a reasonable distance from the airport and the highway.

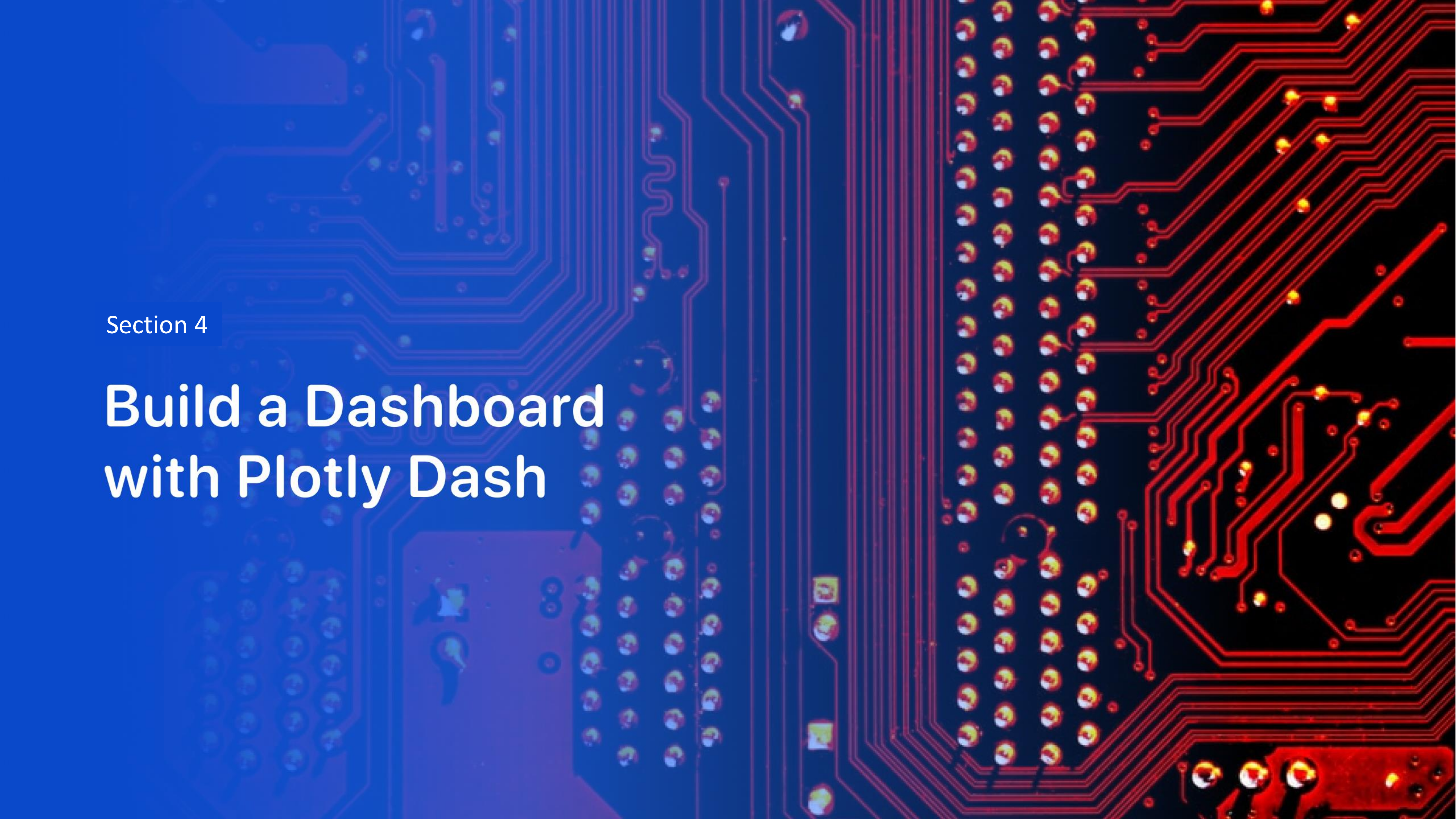


East coast sites nearby locations

- Launch sites are far enough from the nearest city (>15km)

- Launch sites are close to highways and railways(<7km)



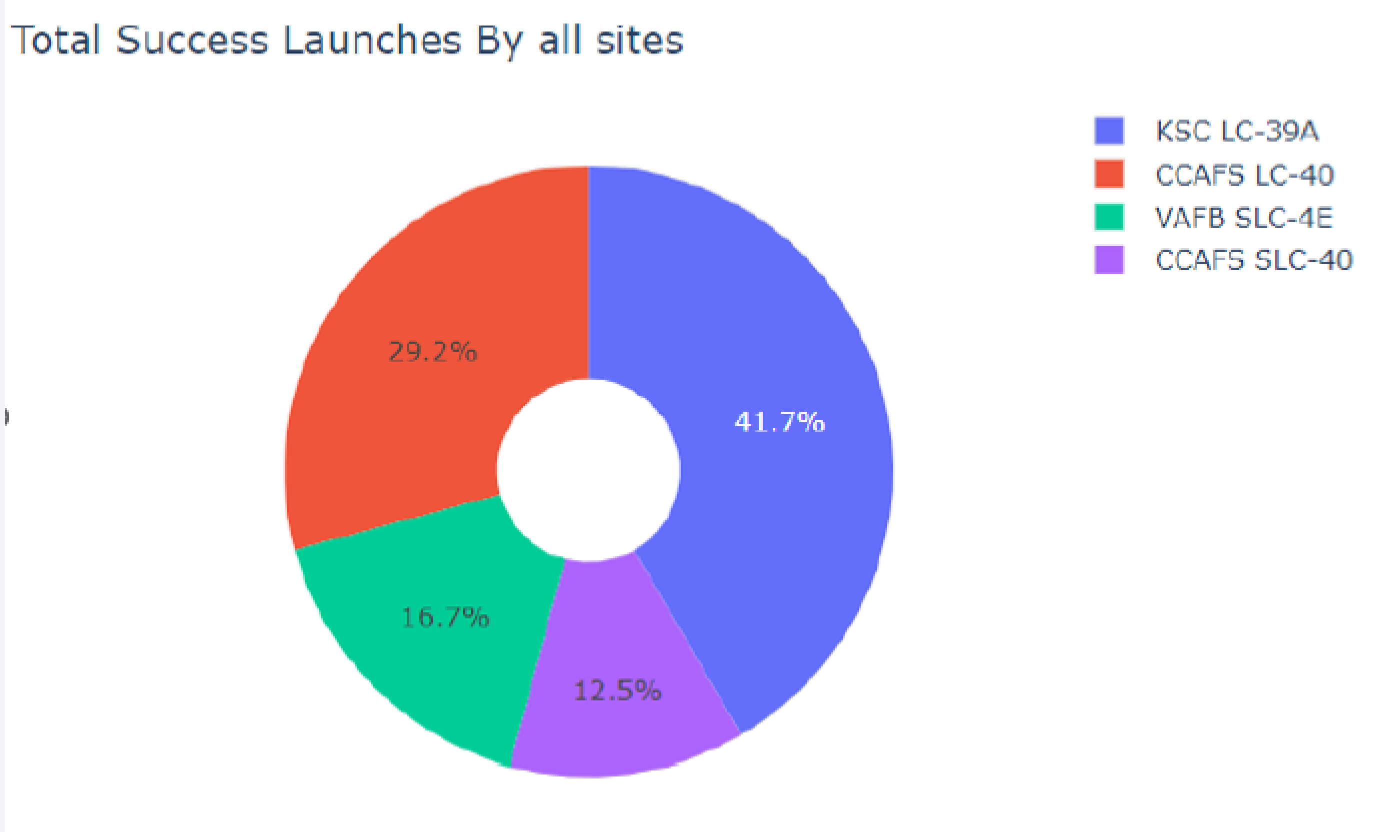


Section 4

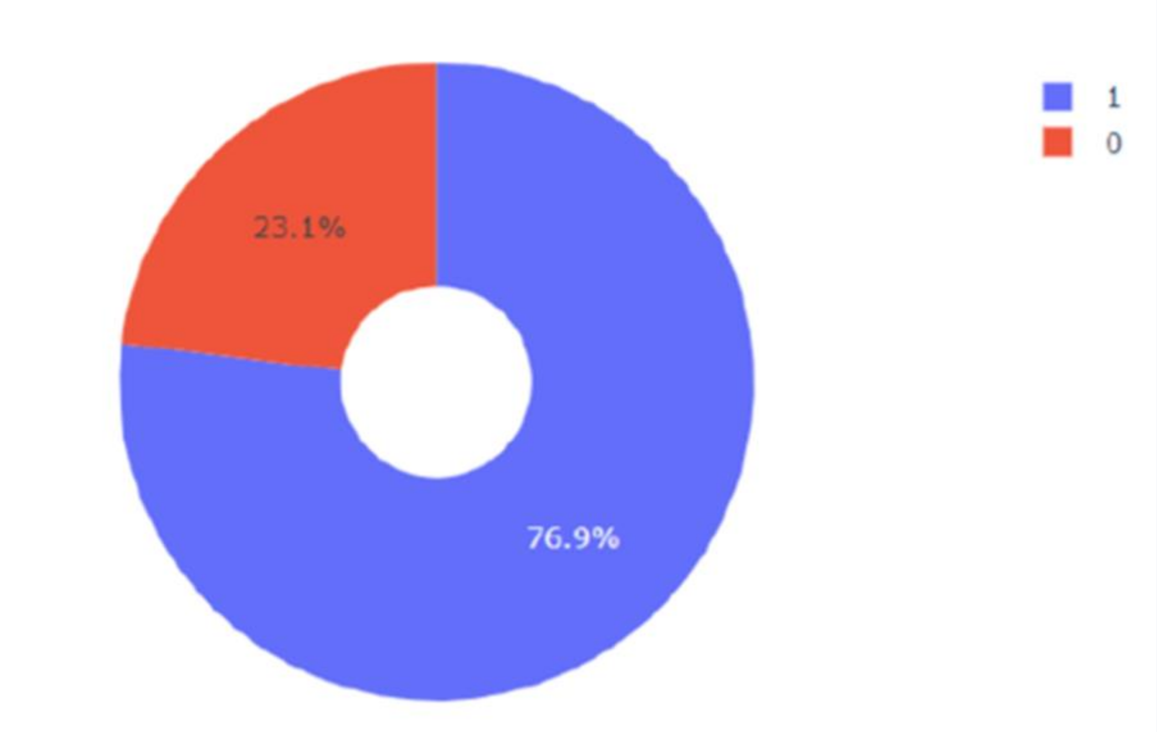
Build a Dashboard with Plotly Dash

Successful launches by site

- One site KSC LC-39A is clearly ahead from others.

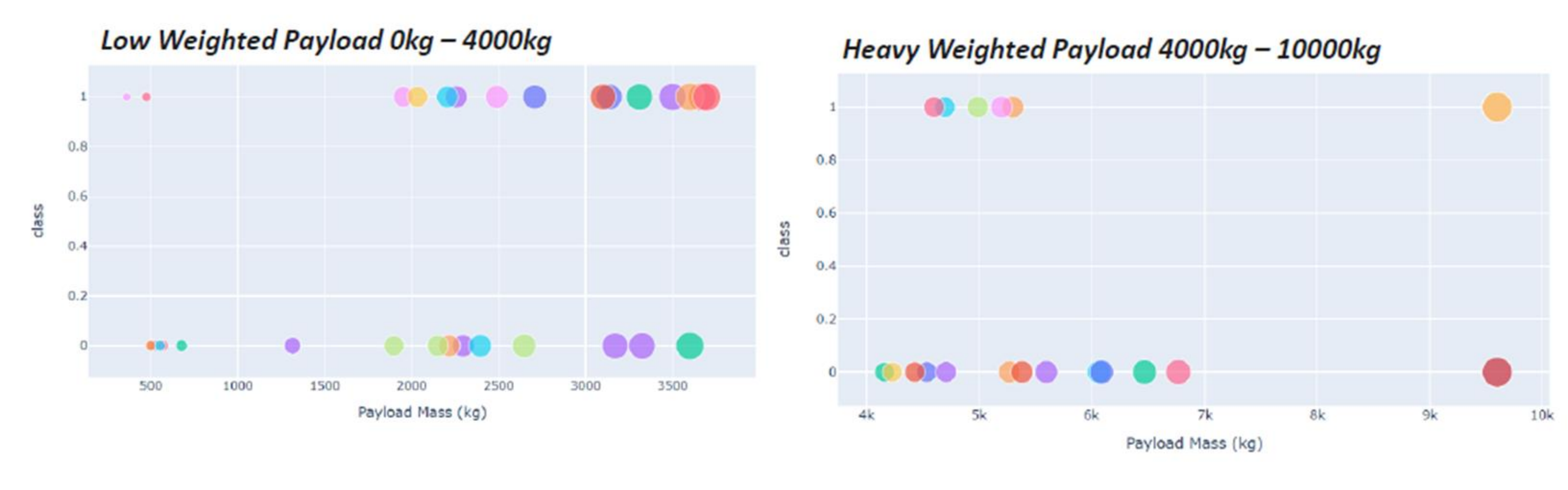


Launch success ratio for site KSC LC-39A



- Site KSC LC-39A as success ratio of 76.9%

Payload Mass vs. launch outcome for all sites



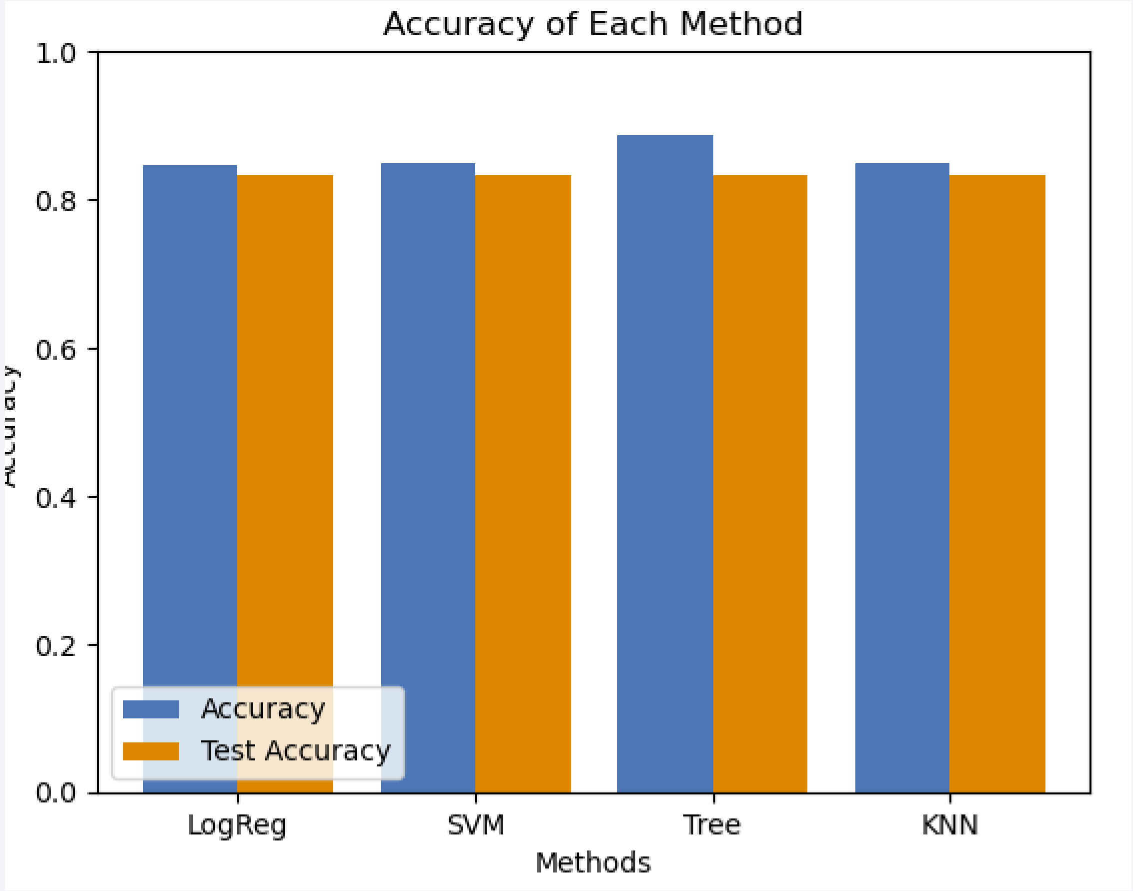
- Launches with a payload mass under 6000 kg and FT boosters are the most successful combination.
- Data above 7000 kg are missing for a better conclusion.

Section 5

Predictive Analysis (Classification)

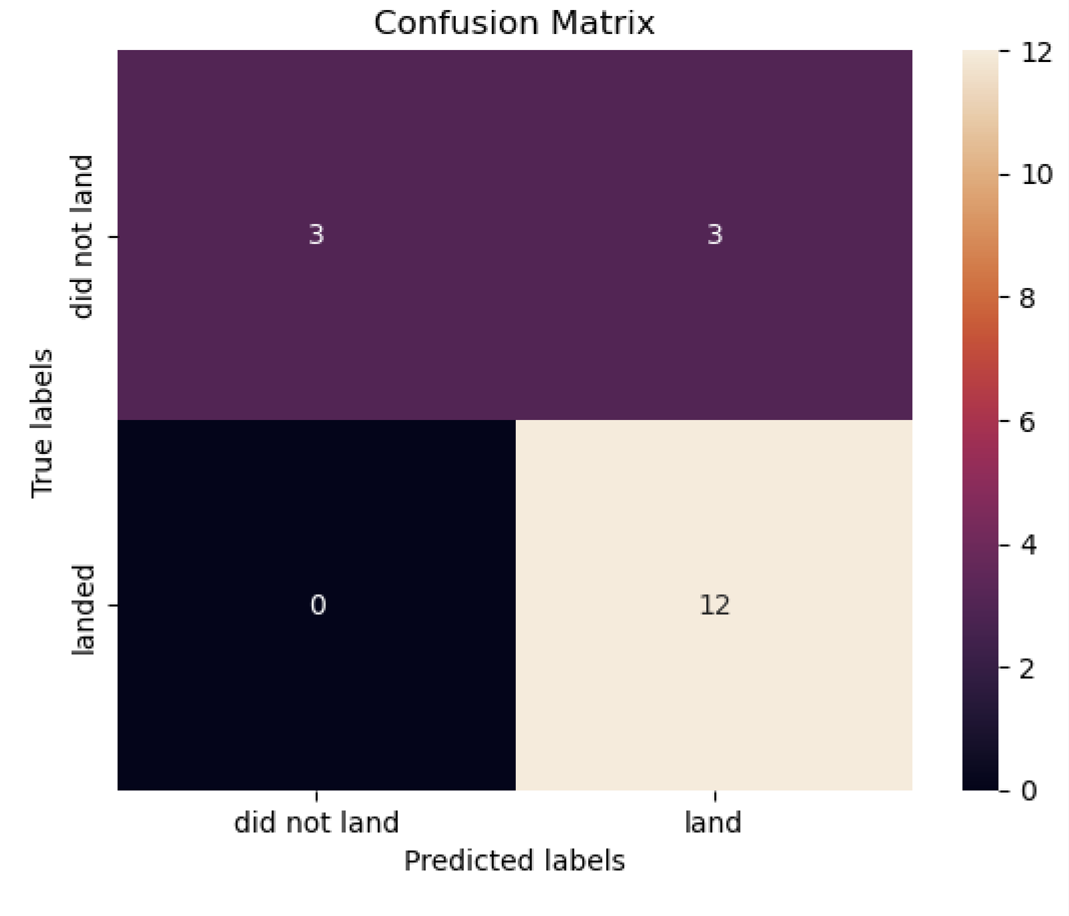
Classification Accuracy

- All models have an accuracy of 84% with slight better one for Decision Tree.



Confusion Matrix

- Models are rights except for 3 labels flagged as false negative.



Conclusions

- All launch sites are located on the coast and away from nearby cities. This avoids damage to the population in case of failure.
- Site KSC LC-39A has the highest rate of success out of all launch sites.
- Since 2015, the success rate of rocket landings has increased.
- Accuracy is the maximum for the Decision Tree Classifier.
- Launches with a payload over 7000kg are less risky.
- More observations would improve predictions and maybe make an algorithm appear more accurate than others.
- Model developed can predict the outcome of a given recovery with reasonable accuracy of 84%.

Thank you!