IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Matías Mödinger
13-09-2022

Outline

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

Executive Summary

- We started by collecting data regarding past launches of SpaceX through the SpaceX REST API and web scraping Wikipedia tables. Then performed data wrangling to get the data in a form that allows calculations and analysis. After this we did some analysis such as visualization and SQL queries to gather information and gain insights on the data. Finally we built several predictive models to classify if a landing will be successful or not using the features that were selected as relevant on the analysis phase.
- During analysis we found out that payload mass, orbit and launch site where the features that impacted the most a successful launching/landing. Finally we built 4 classification models, and when comparing accuracies the best model was a Tree classifier.

Introduction

- SpaceX is the most dominant space company with several successful rocket launches and landings, an important reason for this is that they're able to reuse the first stage of rockets when it's able to land. So the cost of a SpaceX rocket is about 62 million dollars while for other companies the cost can be up to 165 million dollars.
- When will the first stage land successfully? Is the main question to be answered, in order to do this we will build a model to predict the outcome of a landing based on previous launches. Another important question is if a rocket launch will be successful or not, for this we'll be performing data analysis on launch sites and past launch outcomes.

Section 1

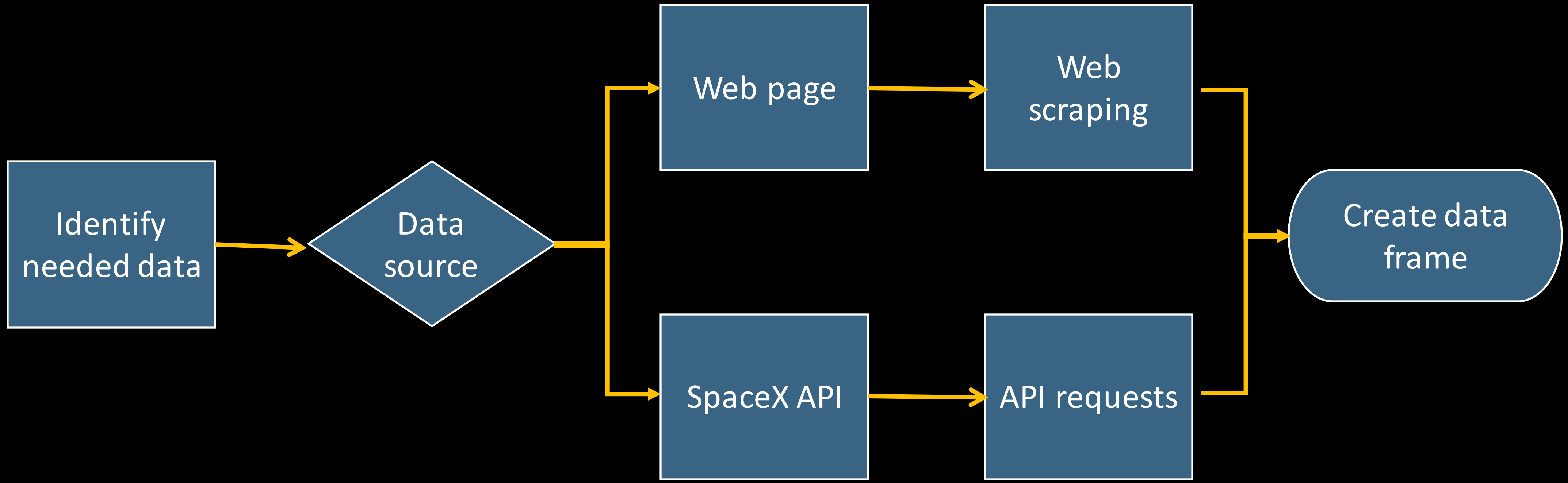
Methodology

Methodology

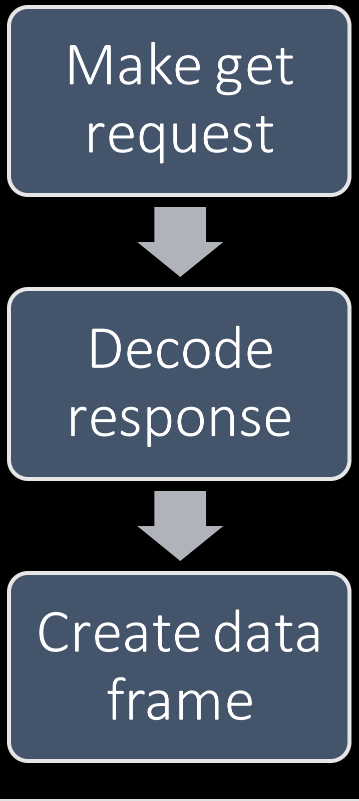
- Executive Summary
- Data collection methodology:
  - Using the SpaceX REST API and through web scraping in list falcon 9 and falcon heavy launches Wikipedia page
- Perform data wrangling:
  - Removed instances which had features that were not of interest, replaced some missing values, created some new features that were useful and preprocessed data before it was given to models.
- 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 classification models where built. Each model had its parameters optimized using a grid search. With these optimized parameters each model was evaluated on the test set to measure them.

Data Collection

- First we identified that data on launches, rockets, launch sites and landings was needed. There were two sources: the SpaceX API and a Wikipedia page, for the first source data was collected through API requests and for the second source web scraping was utilized. Then a data frame was created containing the collected data.



Data Collection – SpaceX API



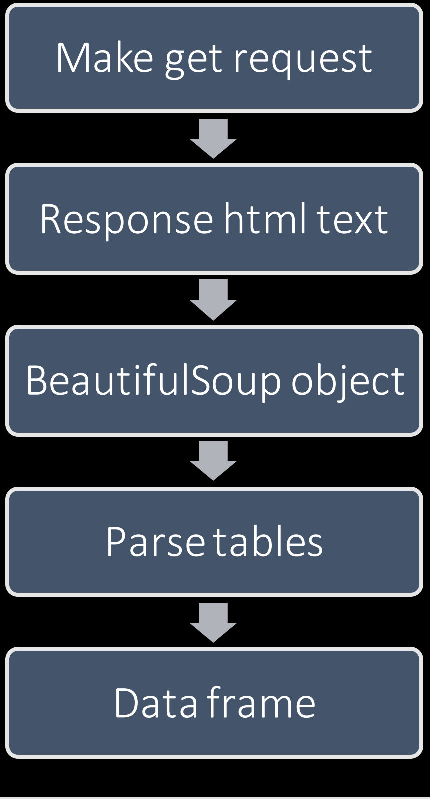
Make get request

Decode response

Create data frame

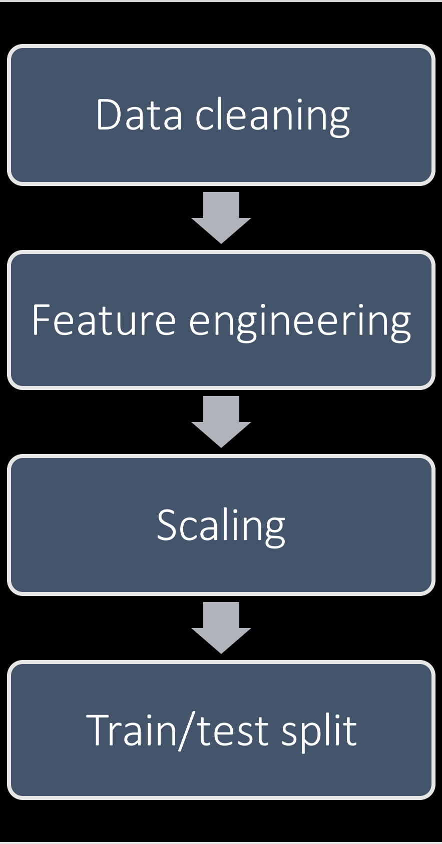
- We collect data from the SpaceX REST API by passing past launches, payloads, cores, booster version and launch sites URLs to the get() function in the requests library, then decode the response as Json using .json() method, to finally use json_normalize() pandas function to create a data frame with the data.
- Noteboook link: [API data collection]()

Data Collection - Scraping



- For web scraping we used list of falcon 9 and falcon heavy launches Wikipedia page to obtain data from the html tables regarding falcon 9 launches. To do this first we get a response from the requests.get() function, then pass the html code in the response to a BeautifulSoup object, and with this object we search for the table we are interested. And finally create a data frame using the rows and headers from this table.
- Github link: Web scraping

Data Wrangling



- Records with multiple cores or payloads where dropped
- Missing values in payload mass where replaced with the mean
- A class feature was created encoding the success (1) or fail (0) of a landing
- Dummy variables where created using one hot encoding for the features: orbit, launch site, landing pad and serial.
- Values for the features selected where standardized so that different scales in features don't affect models performance.
- Data was splitted into train and test data, so we can measure models capacity in classifying data that was not seen.
- Data wrangling related notebooks: API data collection , data wrangling , data visualization , predictive model

EDA with Data Visualization

- Scatter plot of launch sites vs flight number with the outcome as overlay, to see if there's any trend as the number of flight increases on each launch site.
- Scatter plot of launch sites vs payload mass with the outcome as overlay, to see in which sites where launched the heavier payloads.
- Bar chart to visualize the success rate of landing for each orbit.
- Scatter plot of orbits vs flight number with outcome as overlay, to visualize if there's any relation with the increase of flights on each orbit.
- Scatter plot of orbits vs payload mass with the outcome as overlay, to see if the success rate on each orbit has any relation with the payload being heavier or lighter.
- Line plot to visualize the trend of the success rate on each year.
- Notebook link: data visualization

EDA with SQL

Queries performed:

- Names of unique launch sites
- Five records where the launch site name begins with "CCA"
- Total payload mass carried by booster launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date of first successful landing outcome on a ground pad
- Names of boosters which have success in drone ship and have payload mass between 4000 and 6000
- Total number of successful and failure mission outcomes
- Name of booster versions which have carried the maximum payload mass
- Failed landing outcomes in drone ship, their booster versions and launch site names that occurred in 2015
- Rank the count of landing outcomes between 2010-06-04 and 2017-03-20
- Notebook link: EDA with SQL

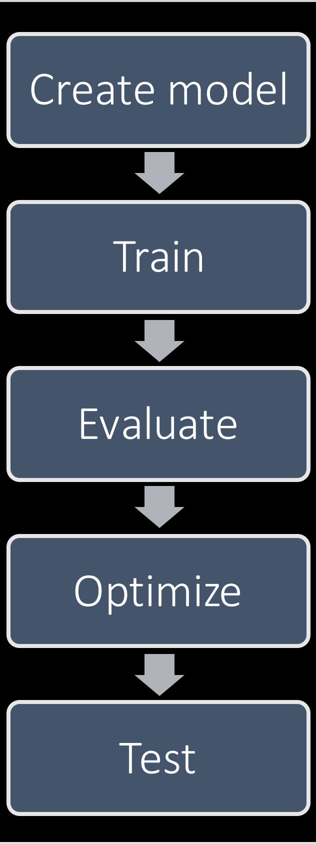
Build an Interactive Map with Folium

- Marker objects where made to label launch sites names, distances and launch outcomes in the map. Some markers included a DivIcon object to add text in the map
- Marker clusters to cluster the markers with the launch outcome information, because there where many with the same location.
- Circle objects to visualize locations such as launch sites or places near launch sites. Some circles included a Popup object to see the name of the location when clicking the circle.
- Lines object to join a launch site to near locations and visualize the distances.
- MousePosition to know the coordinates pointed by the mouse.
- Notebook link: [folium map]

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Dropdown to choose a launch site or all launch sites data for the graphs displayed.
- Pie charts; for all launch sites the pie chart contained the overall launch success for each site, and for a specific site the pie chart contained the launch success rate for this site. The purpose of these charts was to see the site with best success rate and how it compared to other sites.
- Scatter plot where the x-axis was payload mass and the y-axis the launch outcome (1 or 0), with an overlay (color) containing the booster version. The purpose was to see if there's any relation between payloads and launch outcomes and at the same time is there's any between the booster versions and the launch outcome.
- A range slider was made to select the range in kg displayed for the payload mass in the scatter plot.
- Script link: dashboard script

Predictive Analysis (Classification)



Create model

Train

Evaluate

Optimize

Test

- Four classification models where chosen: logistic regression, support vector machine, tree classifier and k-nearest neighbors.
- Data was split into train and test data. Train data was given to models to train them.
- Training was performed using grid search, which allows to train, evaluate and optimize at the same time. For each model class a GridSearchCV object was created and a list of parameters was given to it to optimize, this object did 10-fold cross validation for each combination of parameters, and kept the model with best score for each of the four classes.
- Finally the best model for each class was tested using the test data to see which one had better predictive capabilities.
- Notebook link: [predictive analysis]

Results

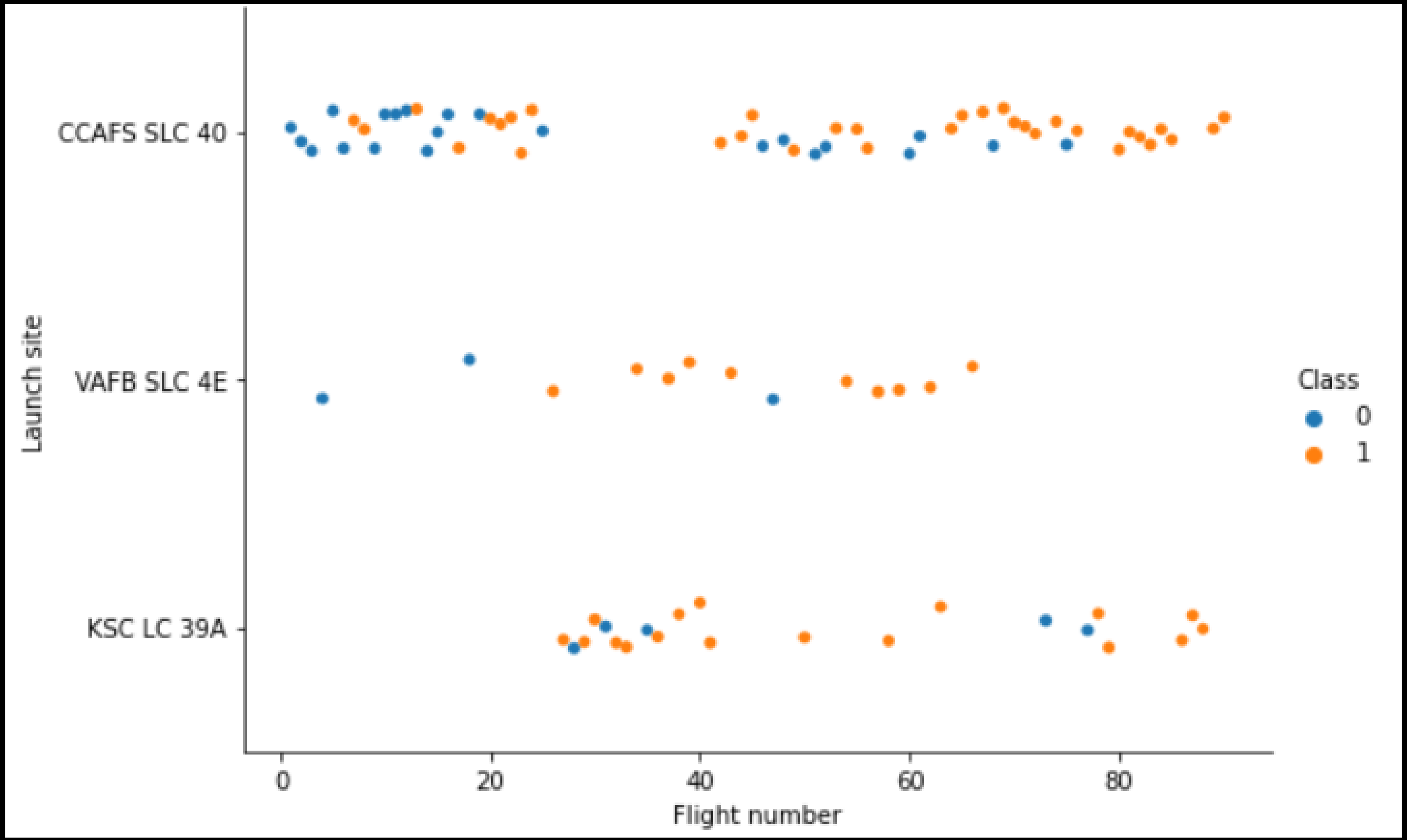
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

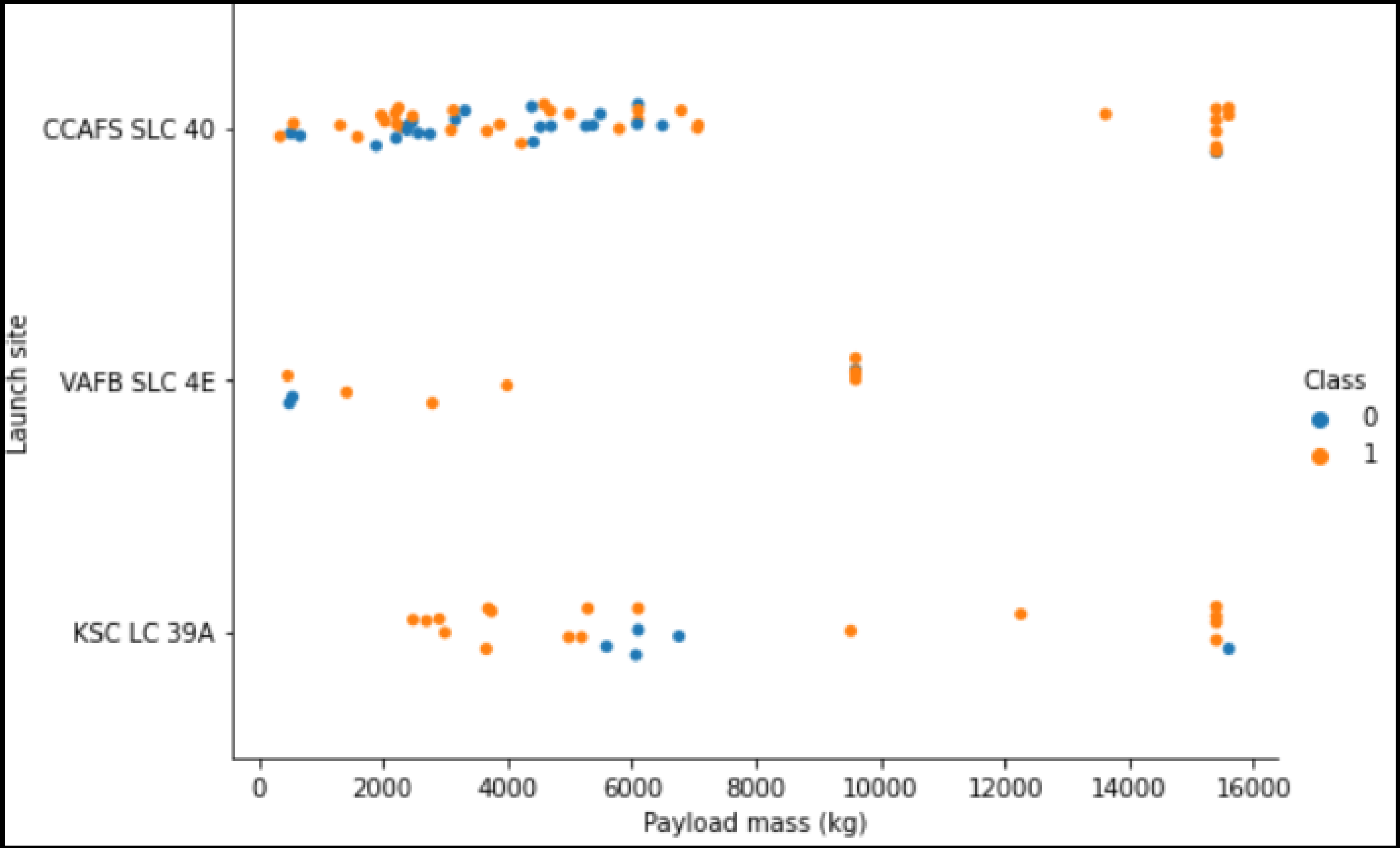
Flight Number vs. Launch Site

- We can see that as the number of flights increases the general trend is for more success on every launch site; on CCAFS SLC-40 and KSC LC 39A we have 100% success from 80 flights onwards.



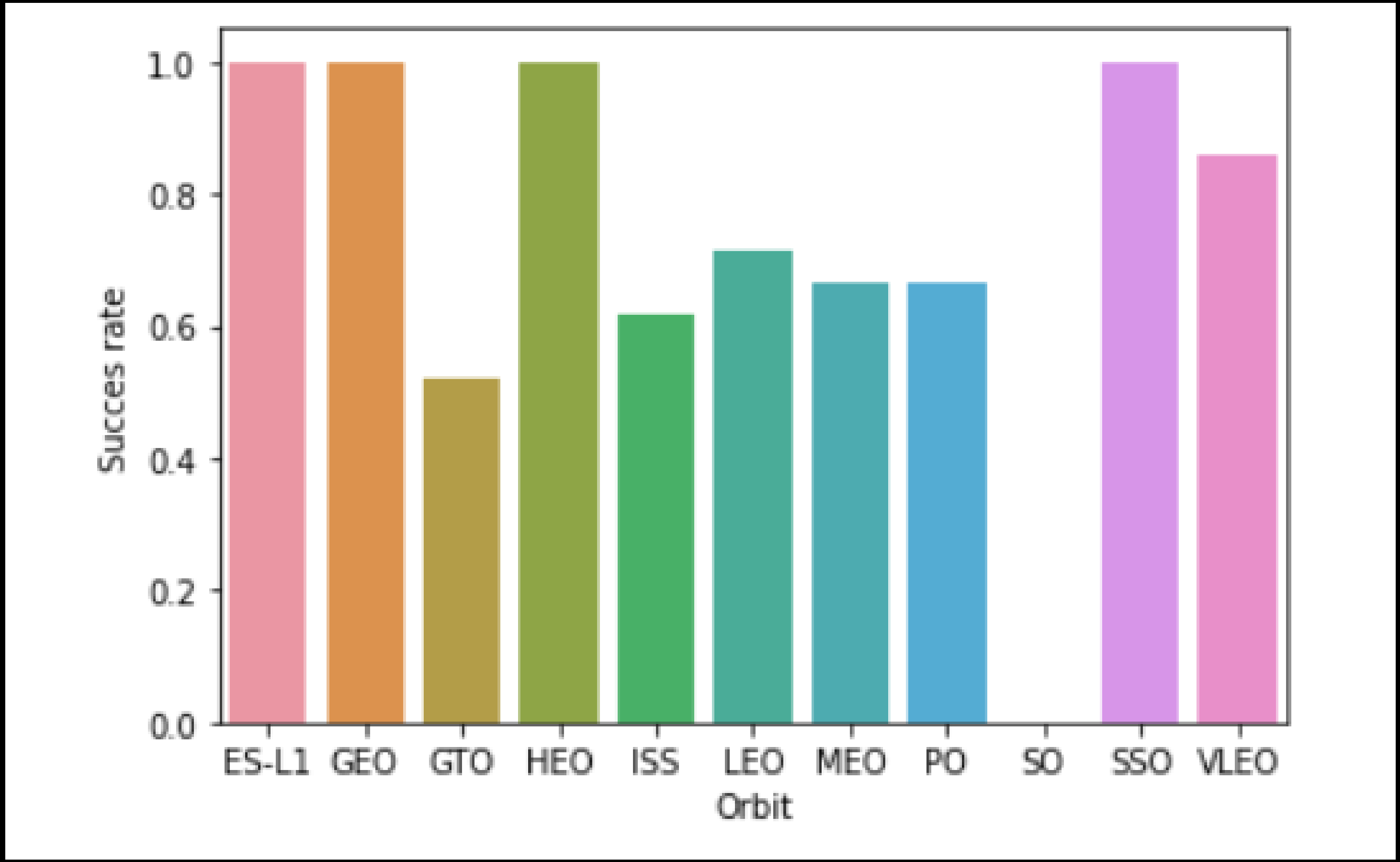
Payload vs. Launch Site

- We can see that heavier payload masses have greater success rate, from 8000 kg onwards there's just one failure on each launch site. This trend is more evident on CCAFS SLC 40 site; with payloads lighter than 8000 kg the success rate is near 50%.



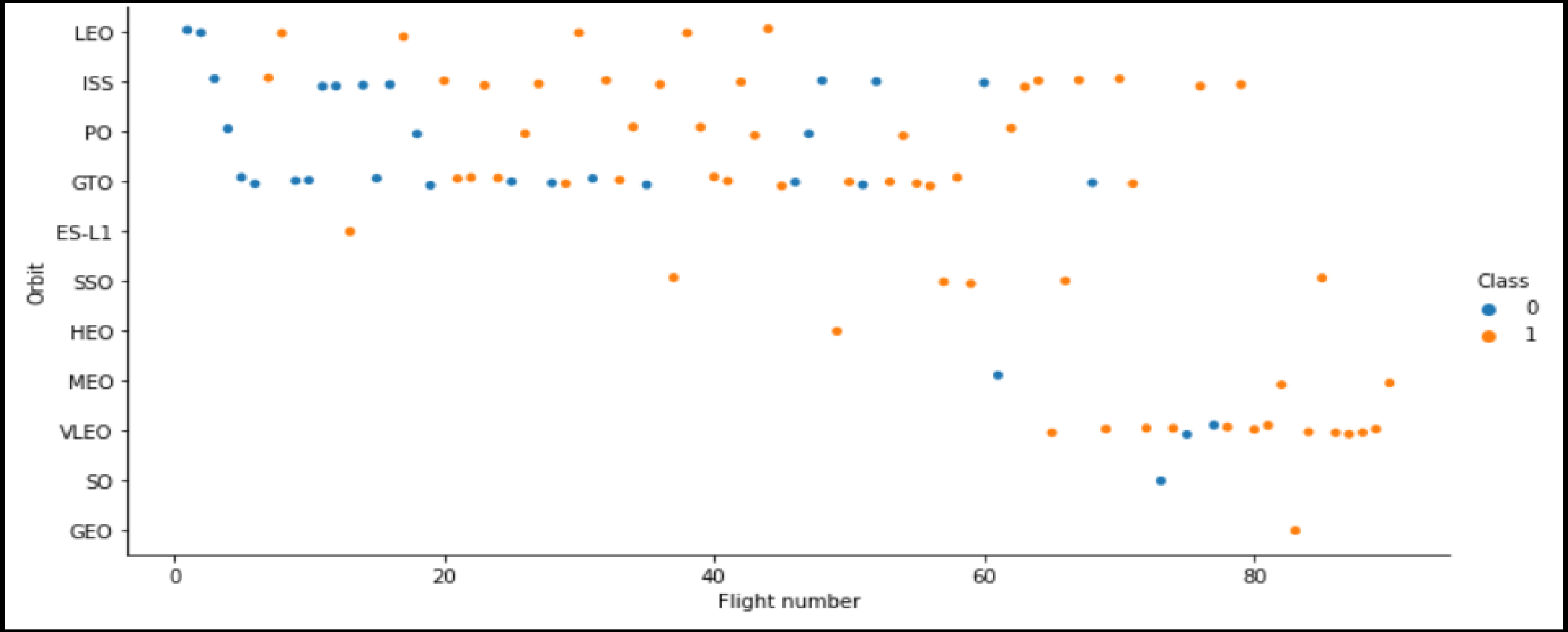
Success rate vs. Orbit type

- Rockets coming from ES-L1, GEO, HEO and SSO orbits have a 100% success rate, but each of this orbits have only 1 record except for SSO that has 5 records. GTO on the other hand is the one with lowest success rate but it's the most common type of orbit in the records.



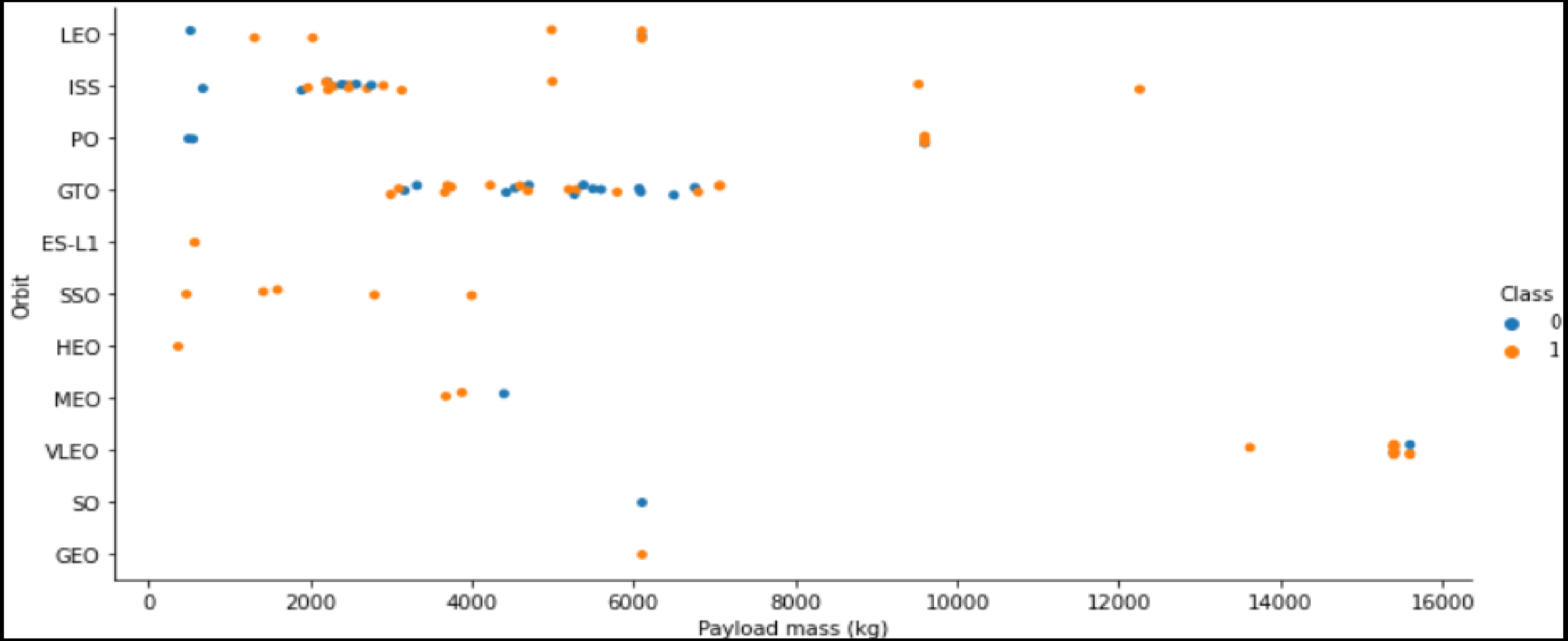
Flight number vs. Orbit type

- As the number of flights increases the general trend on the launch sites is to have more success rate, except on GTO type orbits which have failed outcomes spread out.



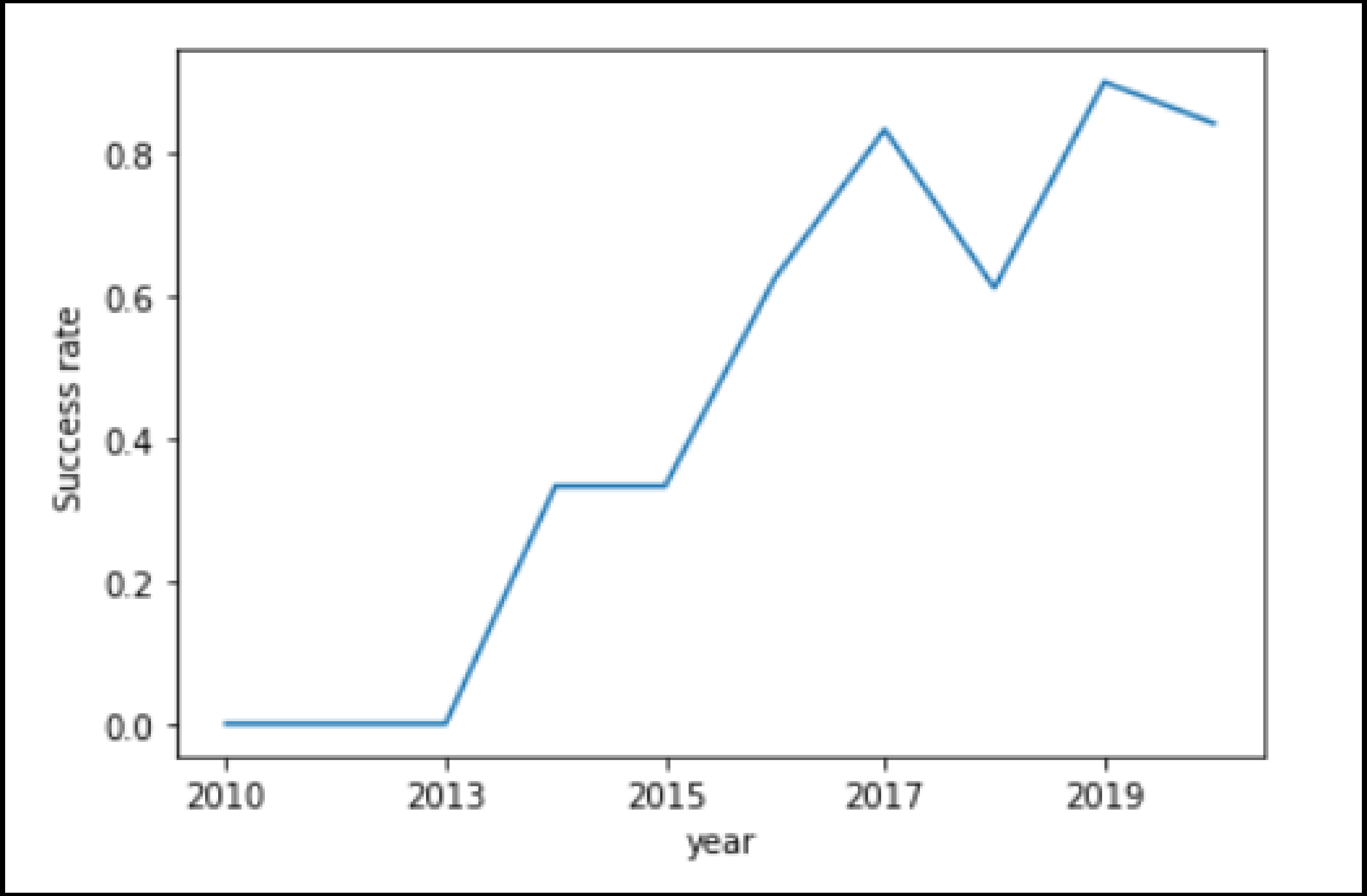
Payload vs. Orbit type

- For LEO, ISS and PO orbit types there seems to be a benefit as the payloads get heavier, whereas for GTO type orbits the payload mass doesn't seem to affect the success or failure



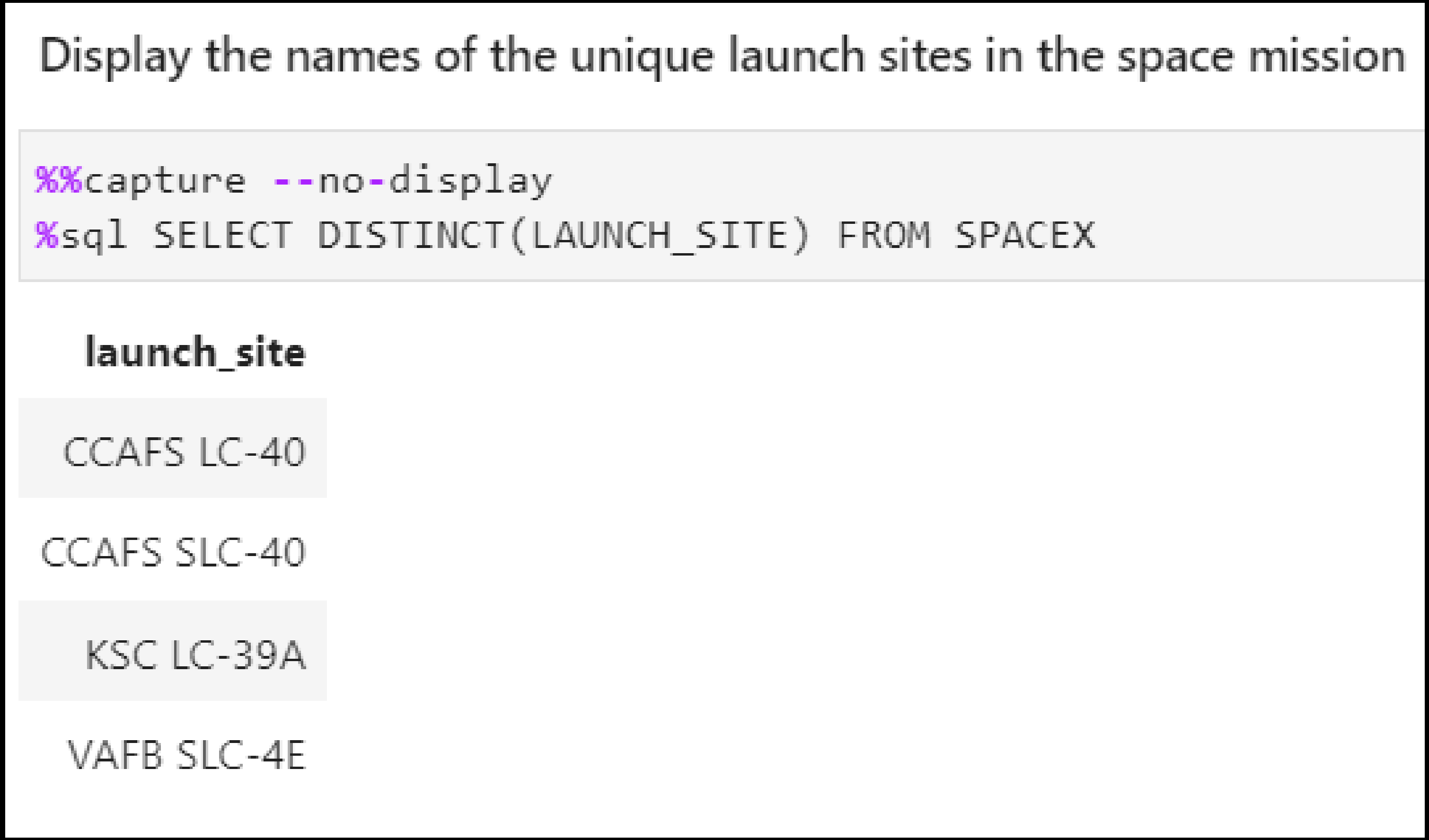
Launch success yearly trend

- Since 2013 there has been a solid increase on the success rate throughout the years, with a fall in 2018 but leading into a greater increase on 2019.



All launch site names

There are four launch sites.

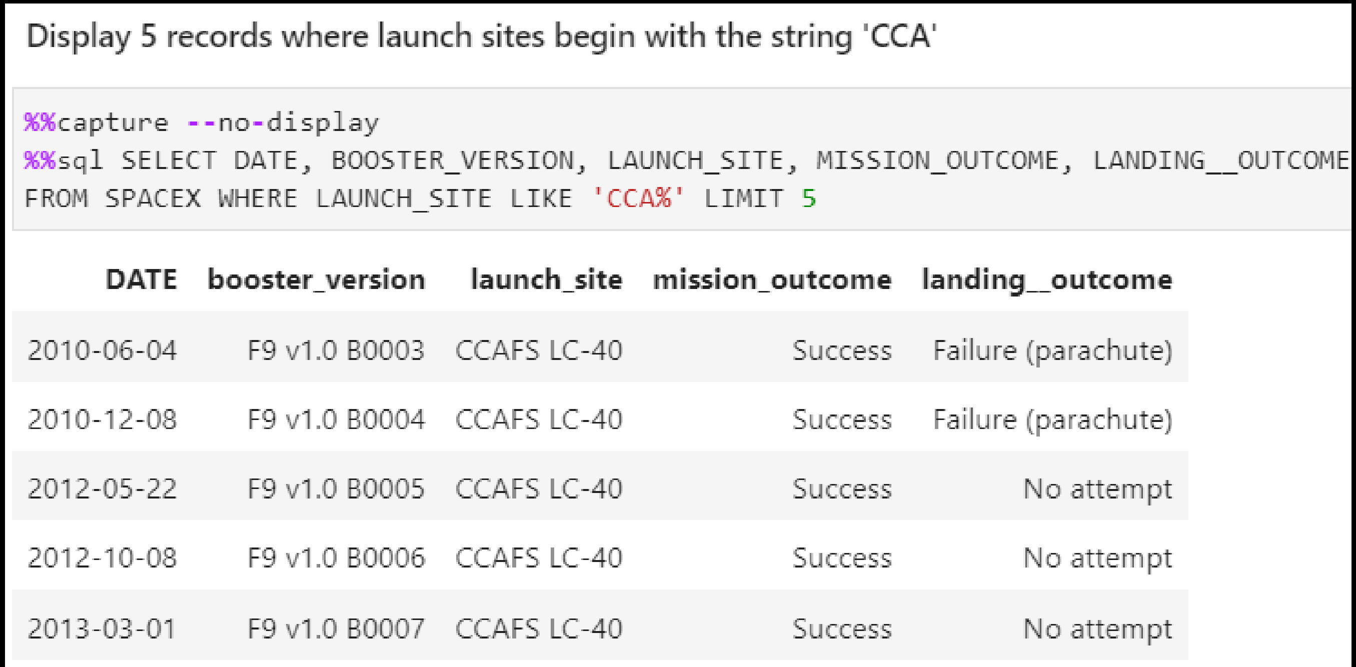


Display the names of the unique launch sites in the space mission

```
%%capture --no-display
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEX
```

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

Launch site names begin with 'CCA'



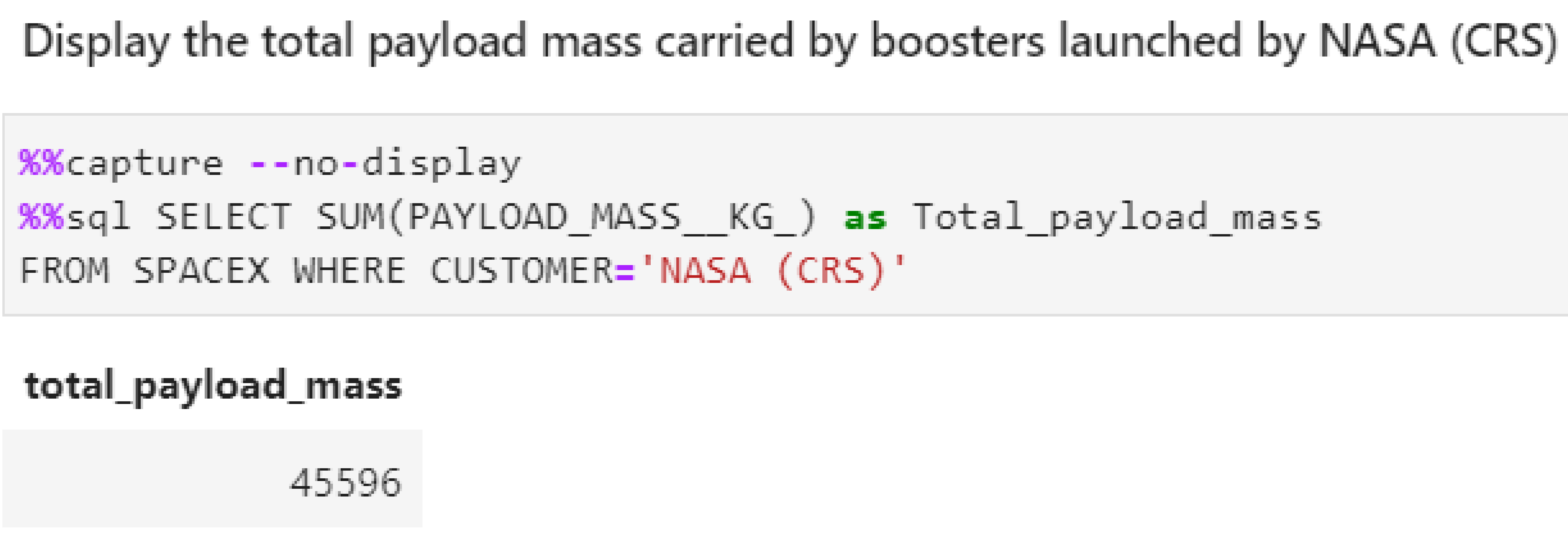
Display 5 records where launch sites begin with the string 'CCA'

```
%%capture --no-display
%%sql SELECT DATE, BOOSTER_VERSION, LAUNCH_SITE, MISSION_OUTCOME, LANDING__OUTCOME
FROM SPACEX WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

DATE	booster_version	launch_site	mission_outcome	landing__outcome
2010-06-04	F9 v1.0 B0003	CCAFS LC-40	Success	Failure (parachute)
2010-12-08	F9 v1.0 B0004	CCAFS LC-40	Success	Failure (parachute)
2012-05-22	F9 v1.0 B0005	CCAFS LC-40	Success	No attempt
2012-10-08	F9 v1.0 B0006	CCAFS LC-40	Success	No attempt
2013-03-01	F9 v1.0 B0007	CCAFS LC-40	Success	No attempt

Total payload mass carried by boosters from NASA (CRS)

Boosters launched by NASA (CRS) carried 45596 kg which is 7.3% of the total payload mass. See Appendix

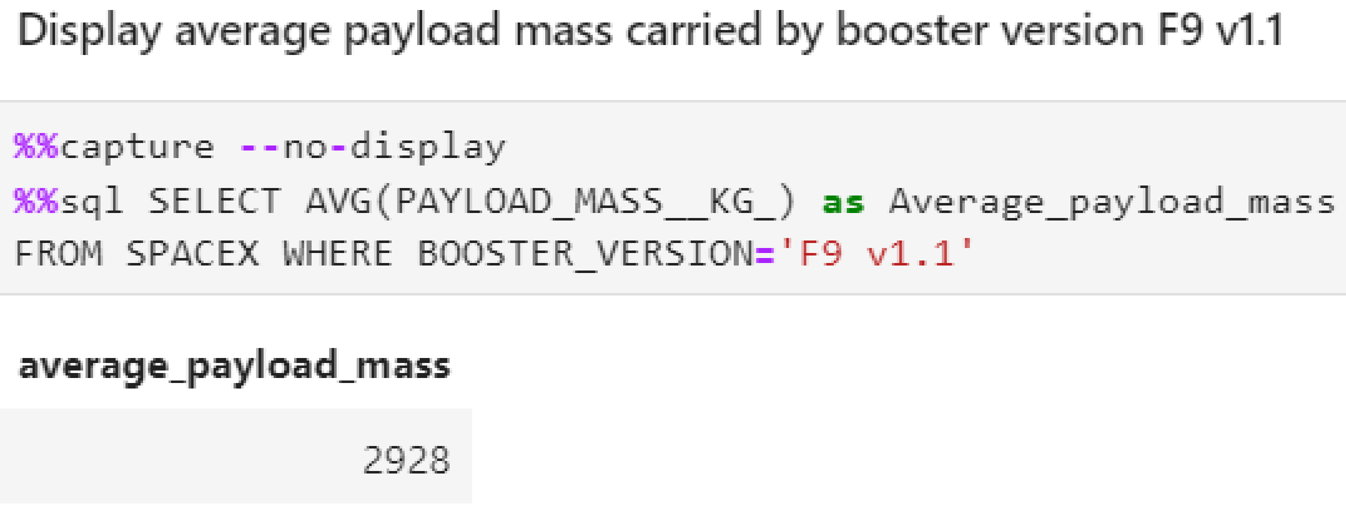


Display the total payload mass carried by boosters launched by NASA (CRS)

```
%%capture --no-display
%%sql SELECT SUM(PAYLOAD_MASS__KG_) as Total_payload_mass
FROM SPACEX WHERE CUSTOMER='NASA (CRS)'
```

total_payload_mass
45596

Average payload mass by F9 v1.1



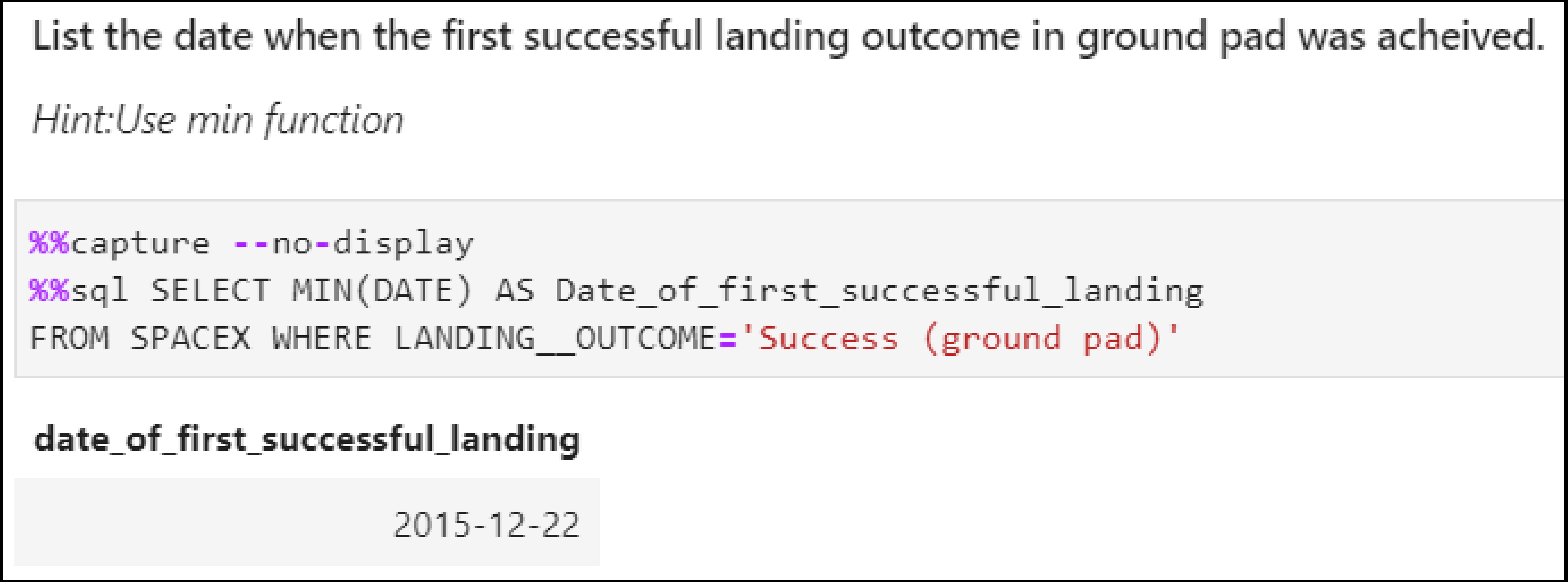
Display average payload mass carried by booster version F9 v1.1

```
%%capture --no-display
%%sql SELECT AVG(PAYLOAD_MASS__KG_) as Average_payload_mass
FROM SPACEX WHERE BOOSTER_VERSION='F9 v1.1'
```

average_payload_mass
2928

First successful ground landing date

On late 2015 SpaceX attempted the first ground pad landing and it was successful, actually they've never had a failed ground pad landing (see appendix).



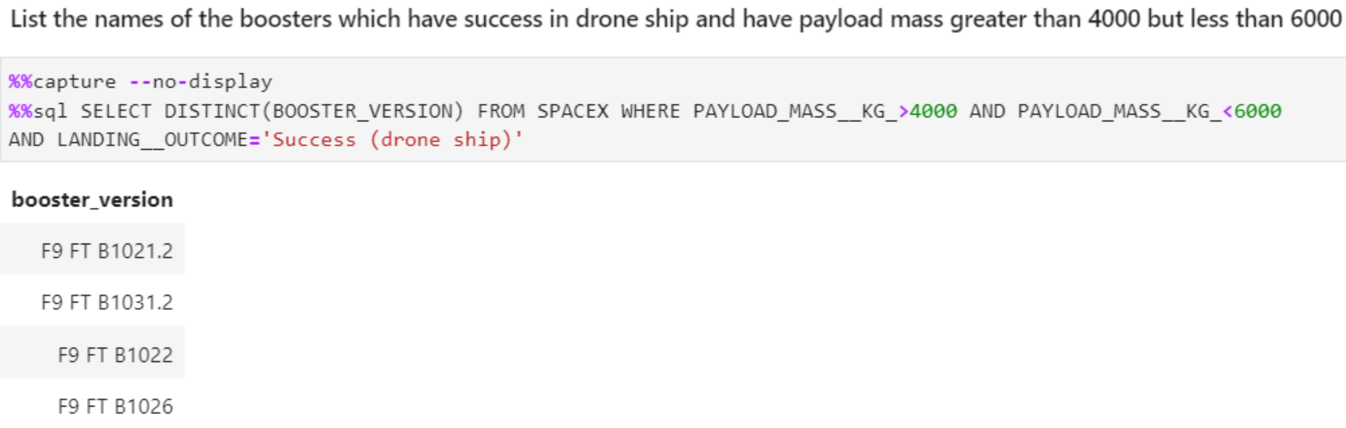
List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

```
%%capture --no-display
%%sql SELECT MIN(DATE) AS Date_of_first_successful_landing
FROM SPACEX WHERE LANDING__OUTCOME='Success (ground pad)'
```

date_of_first_successful_landing
2015-12-22

Successful drone ship landing with payload between 4000 and 6000

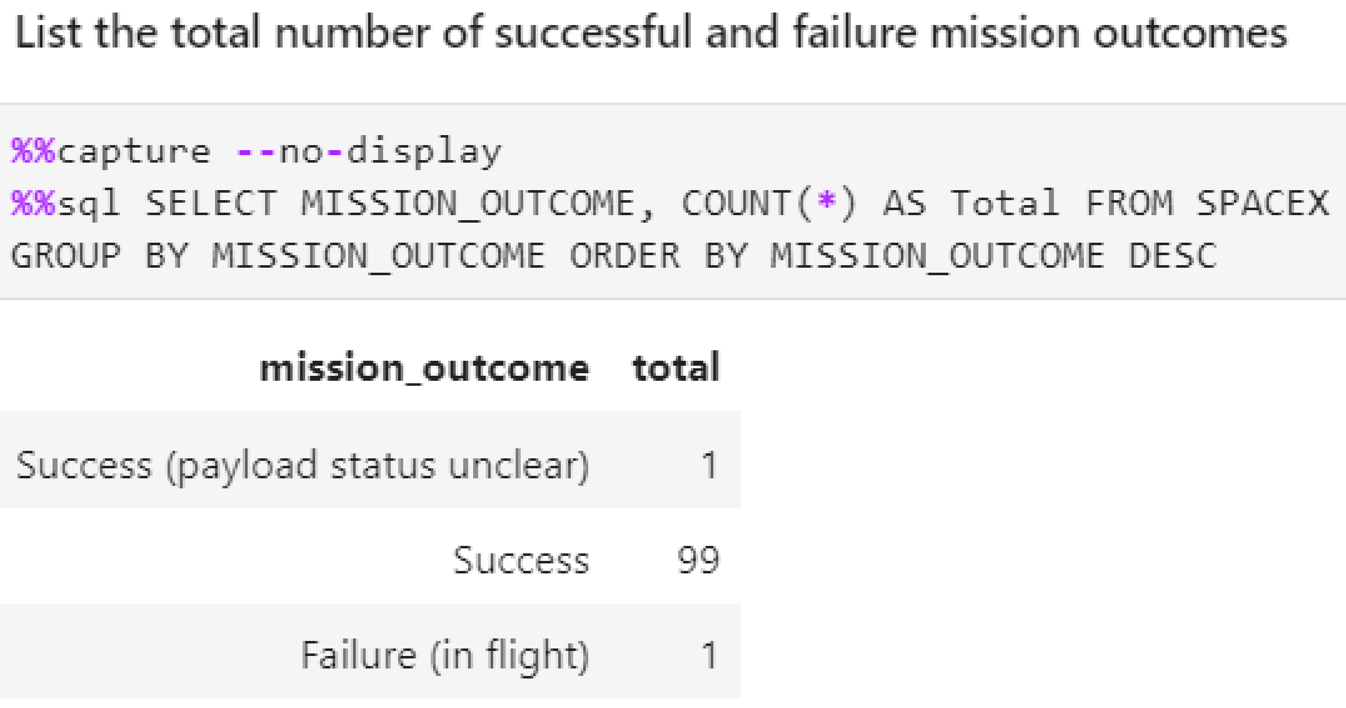


List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%%capture --no-display
%%sql SELECT DISTINCT(BOOSTER_VERSION) FROM SPACEX WHERE PAYLOAD_MASS__KG_>4000 AND PAYLOAD_MASS__KG_<6000
AND LANDING__OUTCOME='Success (drone ship)'
```

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

Total number of successful and failure mission outcomes

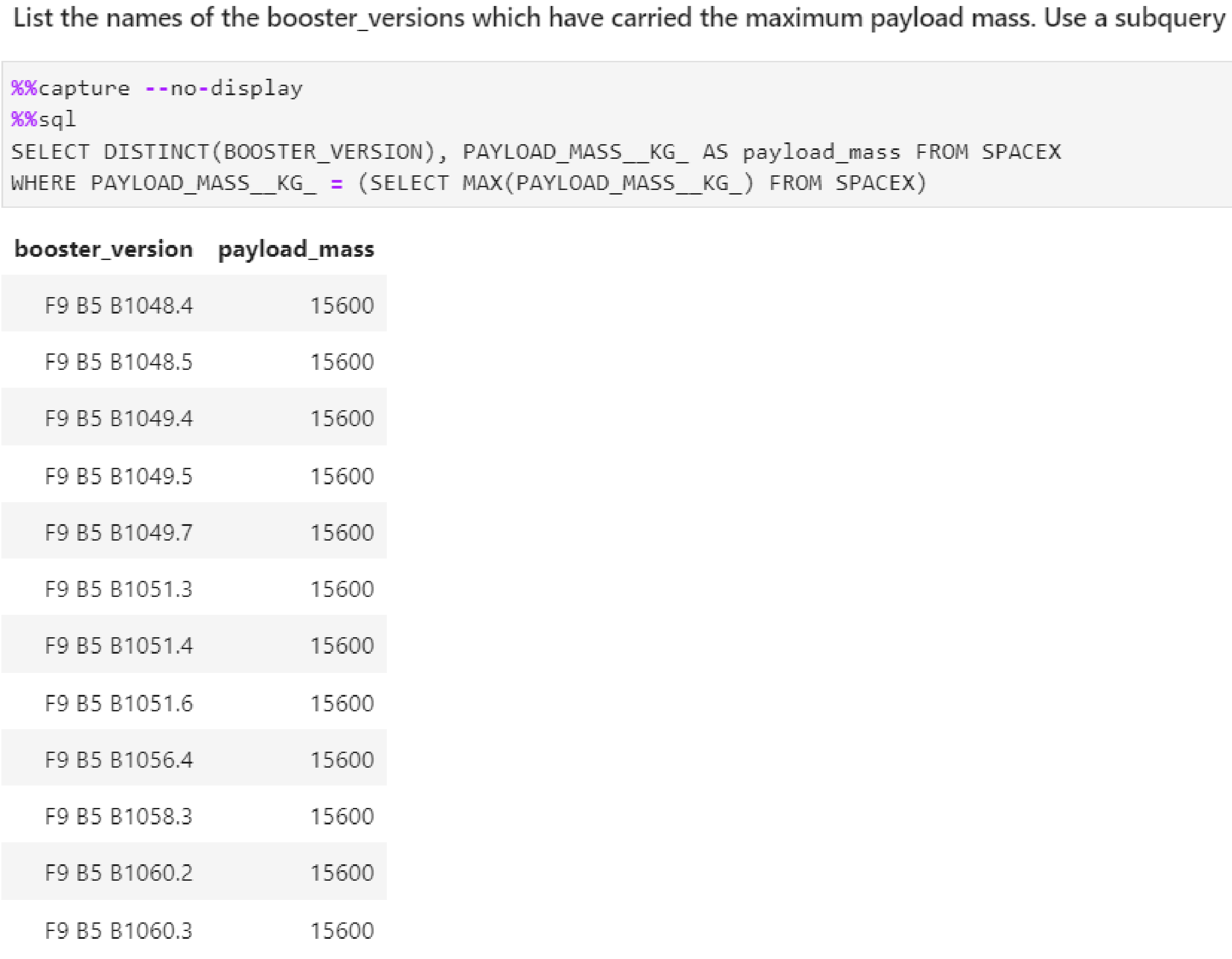


List the total number of successful and failure mission outcomes

```
%%capture --no-display
%%sql SELECT MISSION_OUTCOME, COUNT(*) AS Total FROM SPACEX
GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME DESC
```

mission_outcome	total
Success (payload status unclear)	1
Success	99
Failure (in flight)	1

Boosters that carried the maximum payload



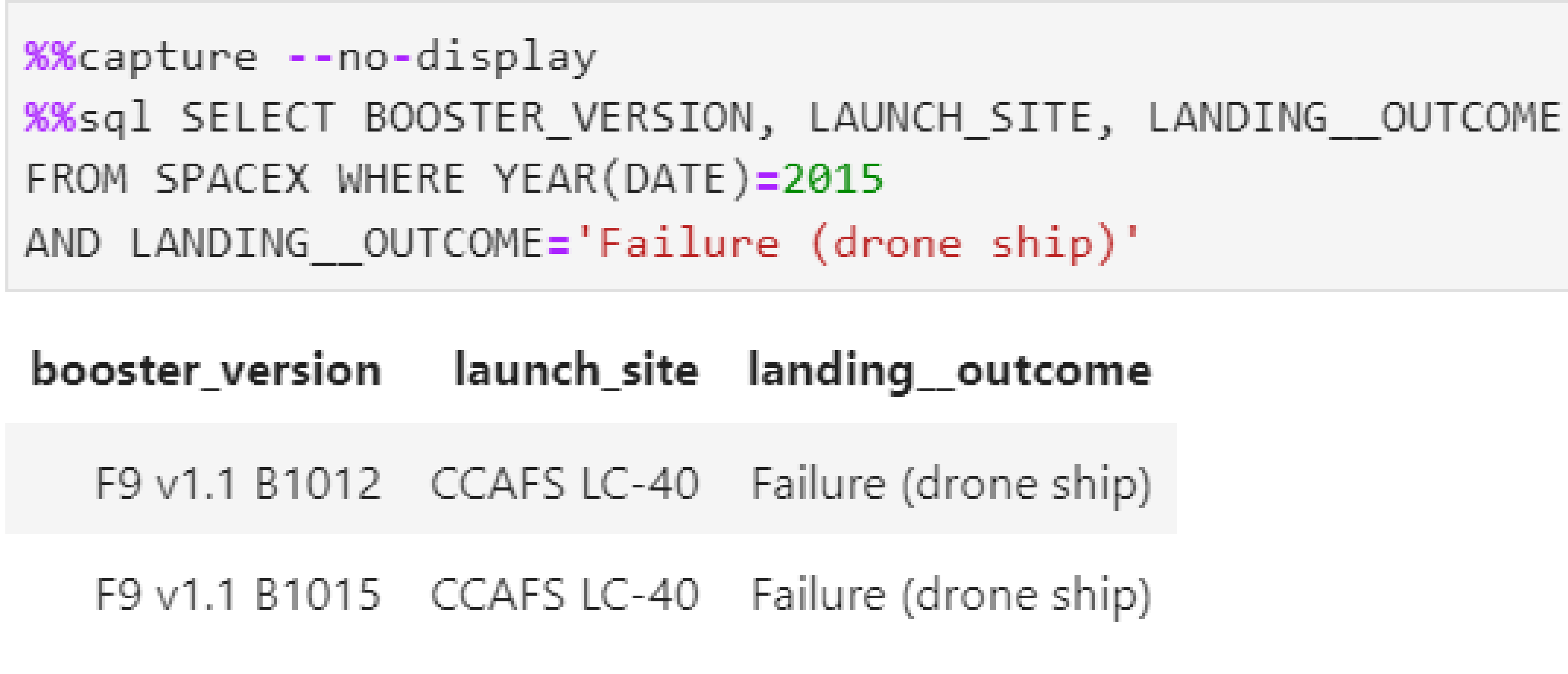
List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%capture --no-display
%%sql
SELECT DISTINCT(BOOSTER_VERSION), PAYLOAD_MASS__KG_ AS payload_mass FROM SPACEX
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX)
```

booster_version	payload_mass
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 launch records

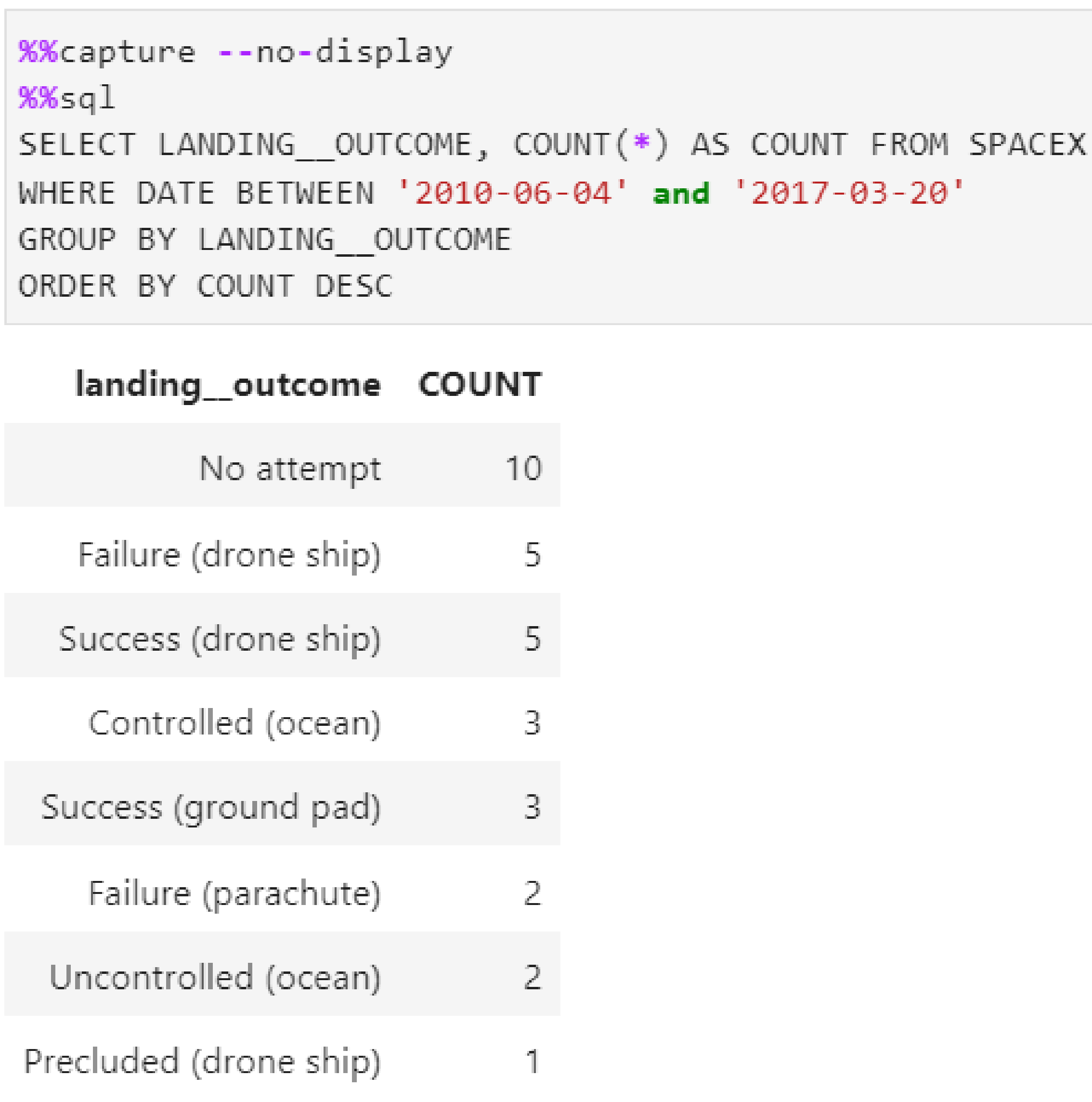
There were two failed drone ship landings in 2015 and both of them occurred in the same launch site.



```
%%capture --no-display
%%sql SELECT BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME
FROM SPACEX WHERE YEAR(DATE)=2015
AND LANDING__OUTCOME='Failure (drone ship)'
```

booster_version	launch_site	landing__outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Ranked landing outcomes between 2010-06-04 and 2017-03-20



```
%%capture --no-display
%%sql
SELECT LANDING__OUTCOME, COUNT(*) AS COUNT FROM SPACEX
WHERE DATE BETWEEN '2010-06-04' and '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY COUNT DESC
```

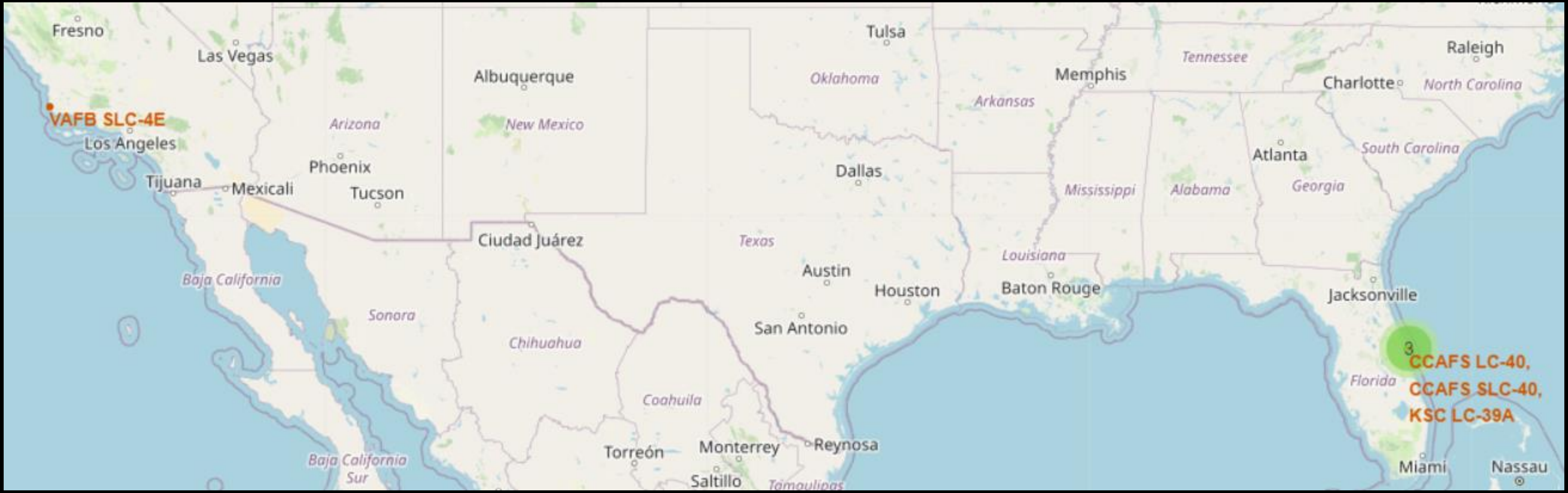
landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

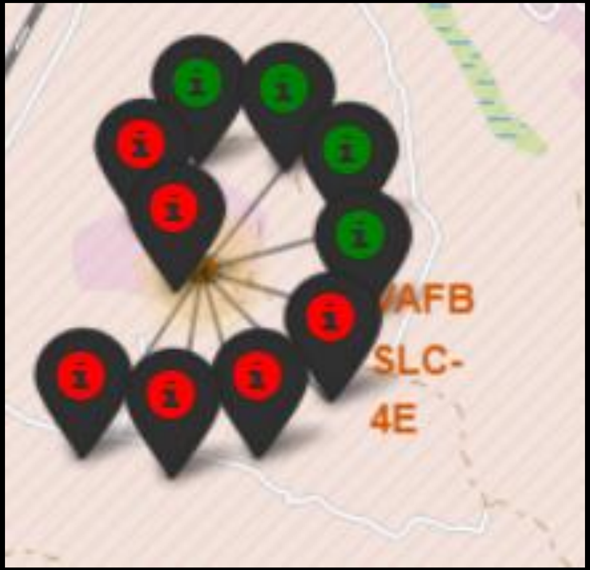
Launch sites locations

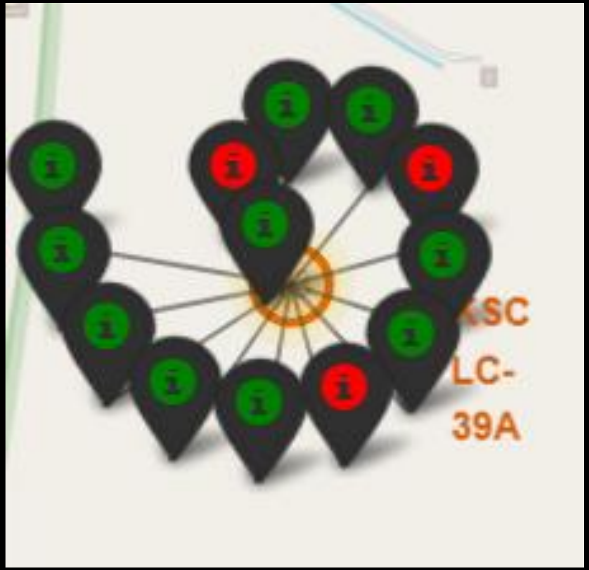
- Notice that all 4 launch site locations are near the coastline, three of which are very close in Florida and the other one is in California

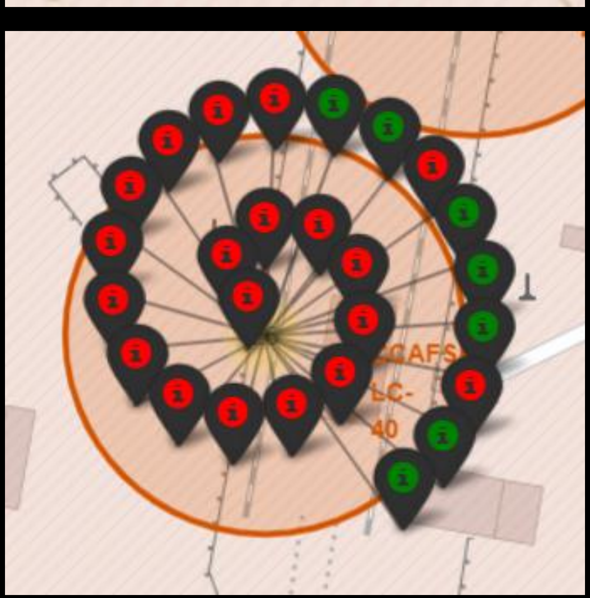


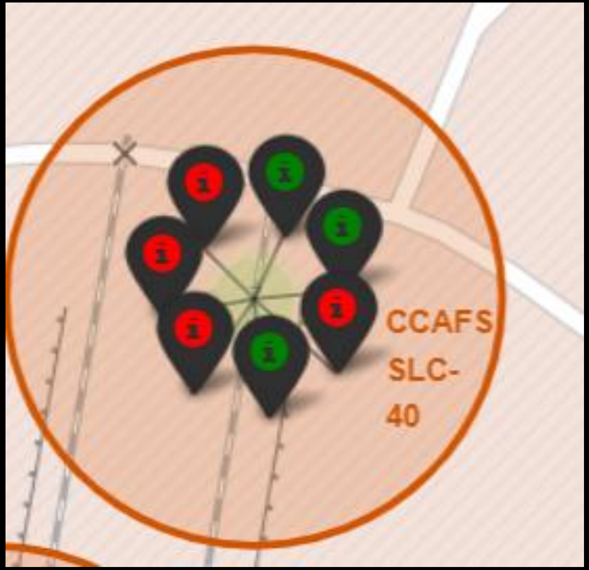
Success rate of launches

- Green markers denote a successful launch and red a failed one.
- It is evident that the launch site with the worst rate of success is CCAFS LC-40 with just 7 successful launches over a total of 26.
- On the other hand; KSC LC-39 is the launch site with greatest rate, having 10 out of 13 accomplished launches.



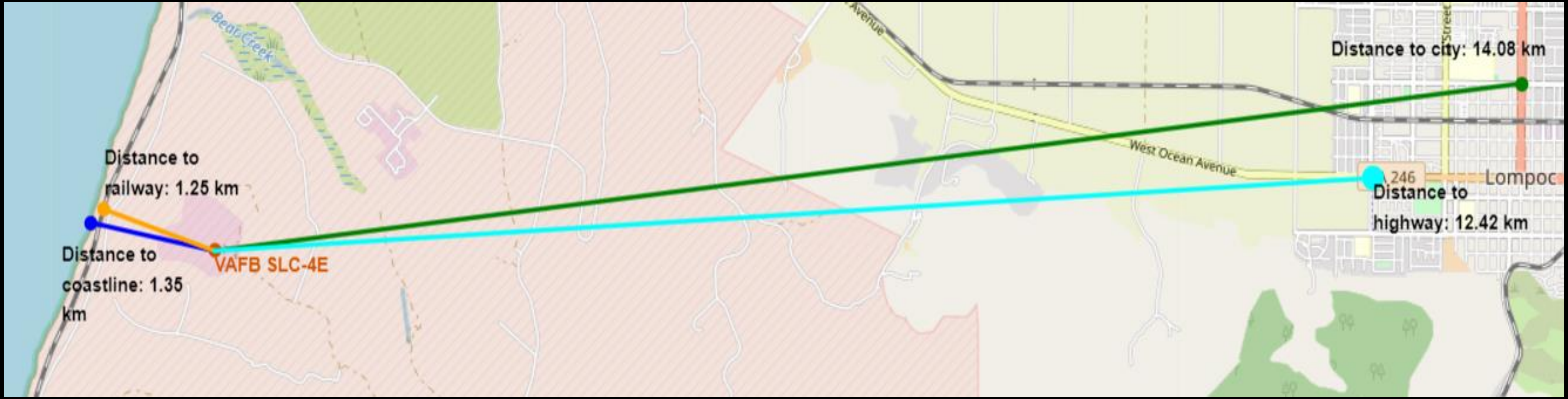






Close locations to VAFB SLC-4E

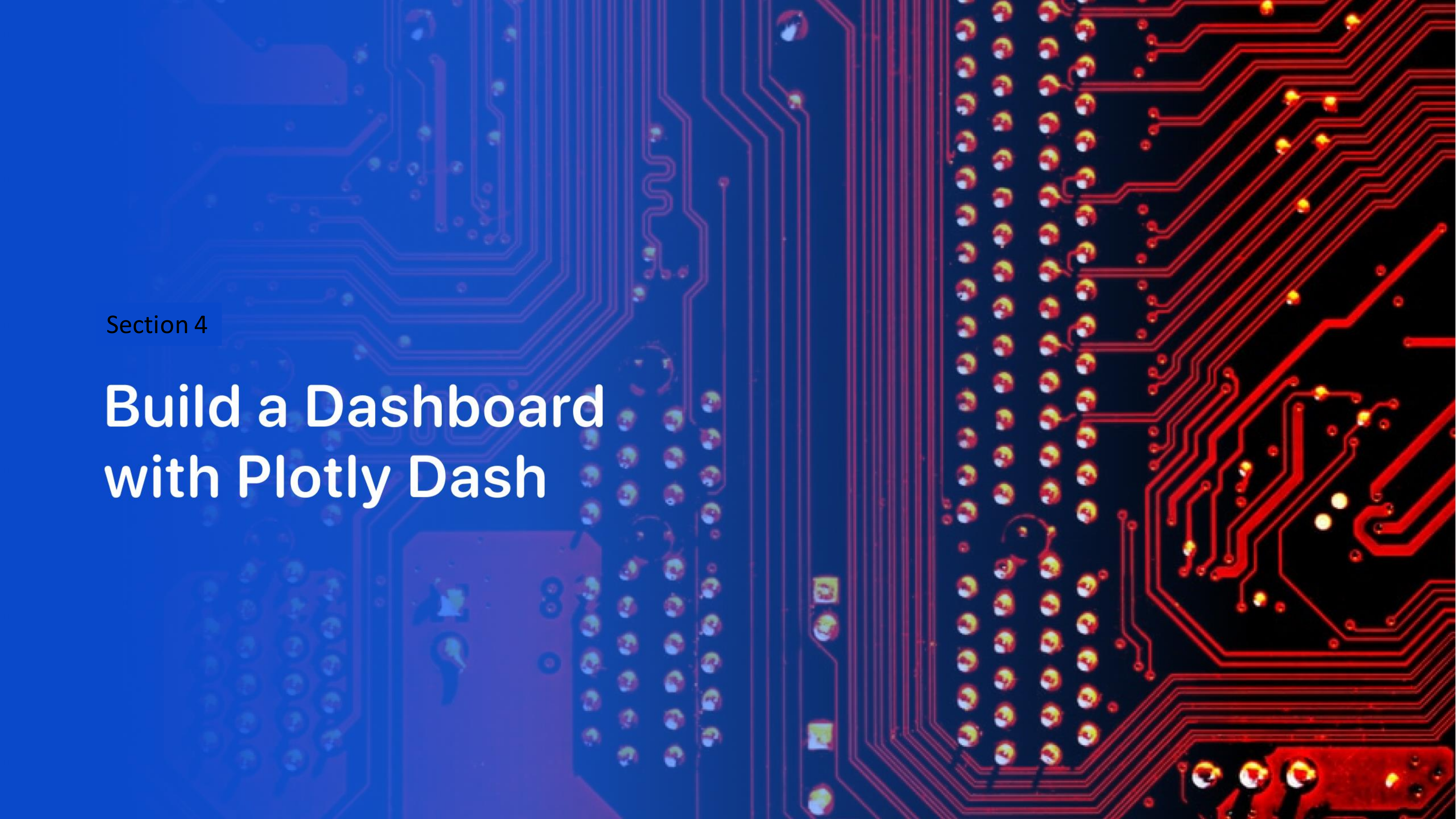
- The launch site is near the coastline which is important for ocean landing, the closest city is 14km and the nearest highway is at 12.42 km; so the base has a pretty safe distance to both of these locations, and it has a railway at just 1.25 km which could be important for supplies.



Distance to city: 14.08 km
Distance to railway: 1.25 km
Distance to coastline: 1.35 km
VAFB SLC-4E
Distance to highway: 12.42 km
West Ocean Avenue
Avenue
Bear Creek
246
Lompoc
Street

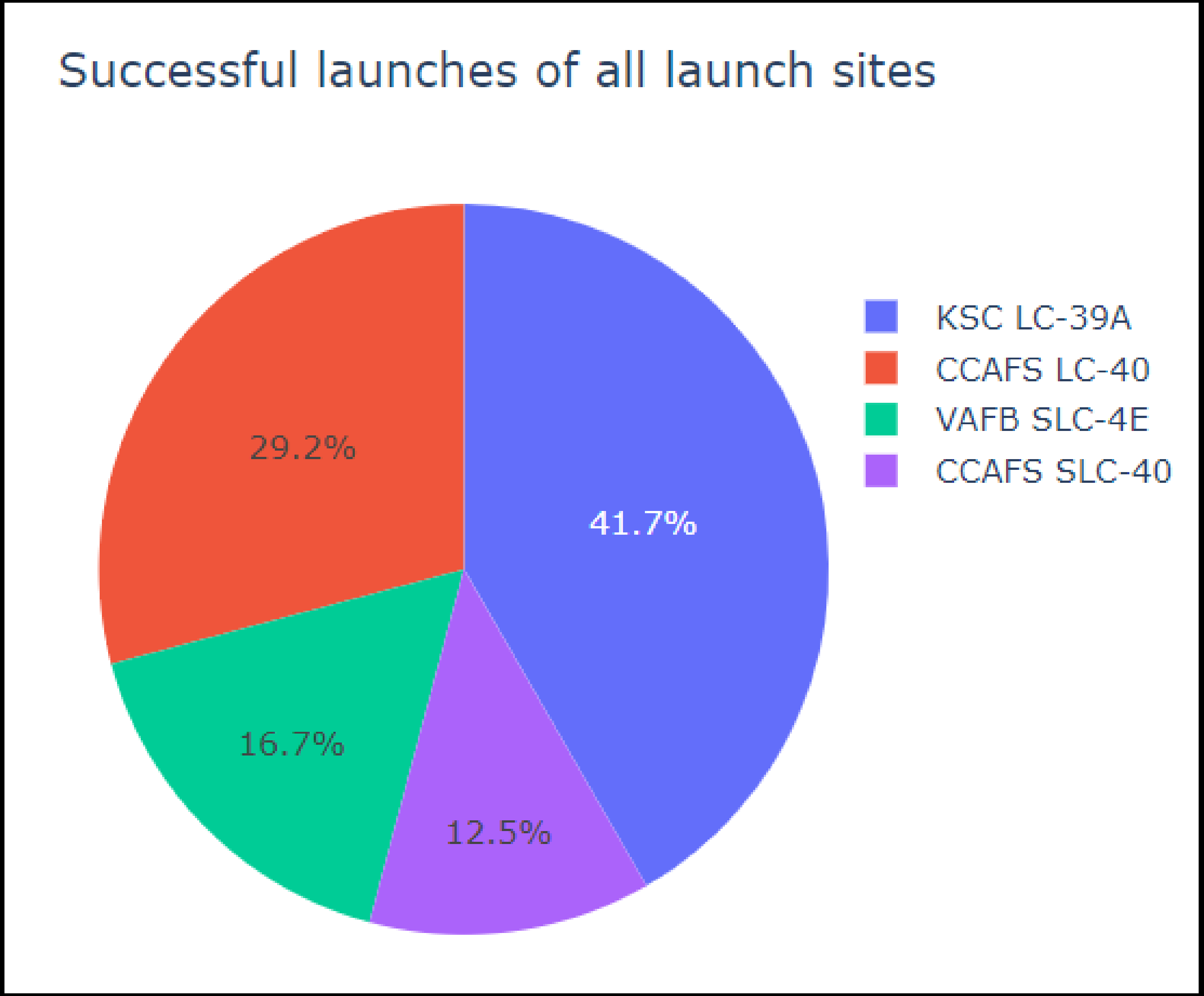
Section 4

Build a Dashboard with Plotly Dash



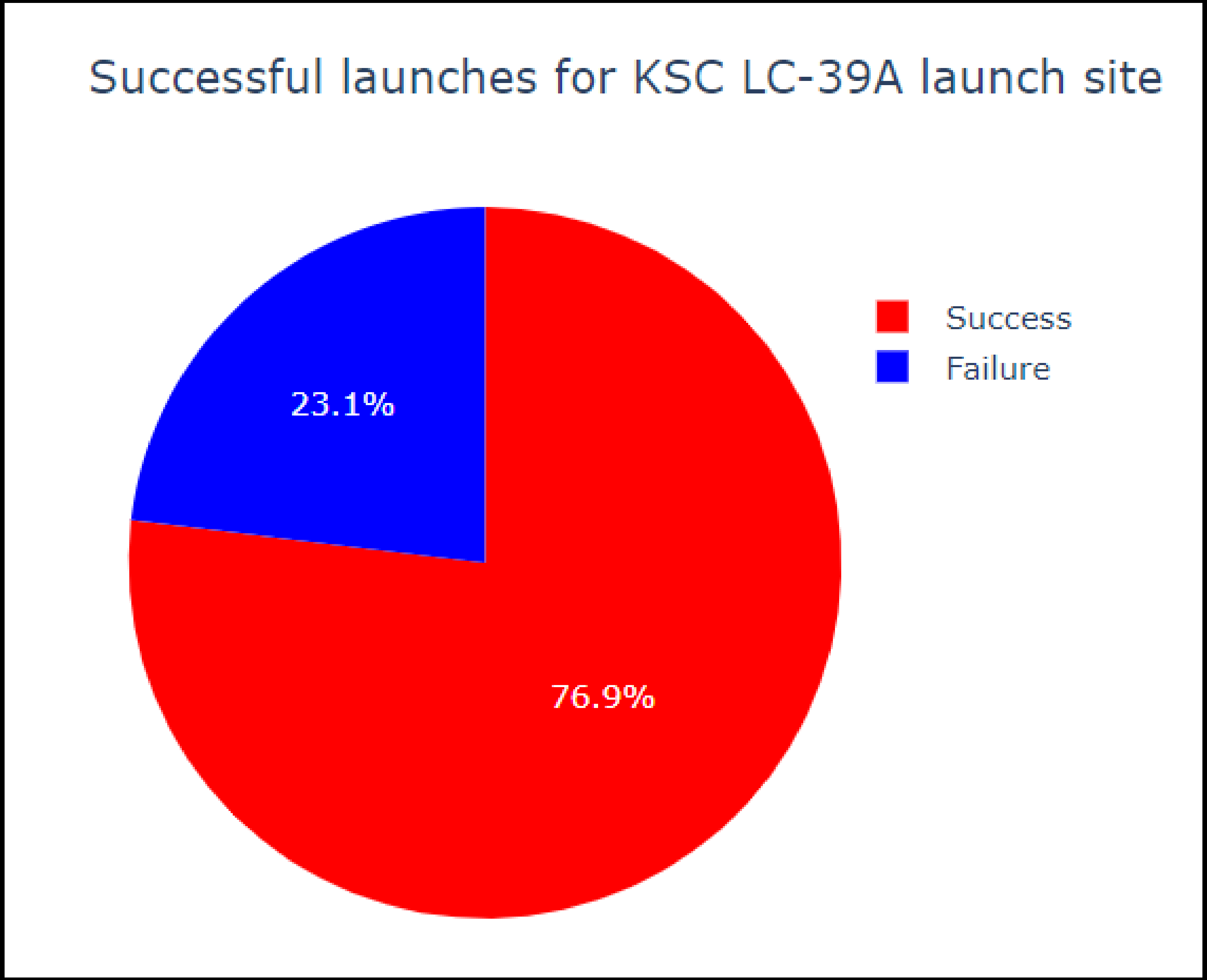
Launch success count for all sites

- KSC LC-39A is the site with most successful launches and CCAFS SLC-40 is the with the least amount.



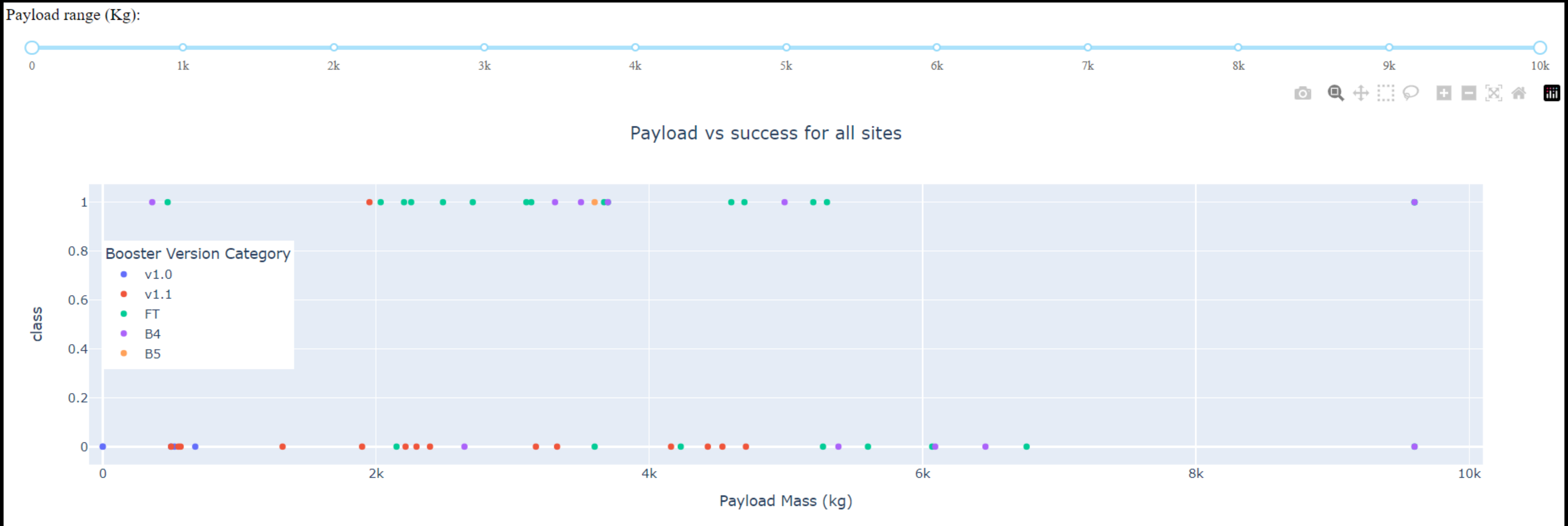
Launch site with highest launch success

- As seen in the previous pie chart, KSC LC-39A is the site with most successful launches and it also has the best success rate of the launch sites with almost 77% rate.



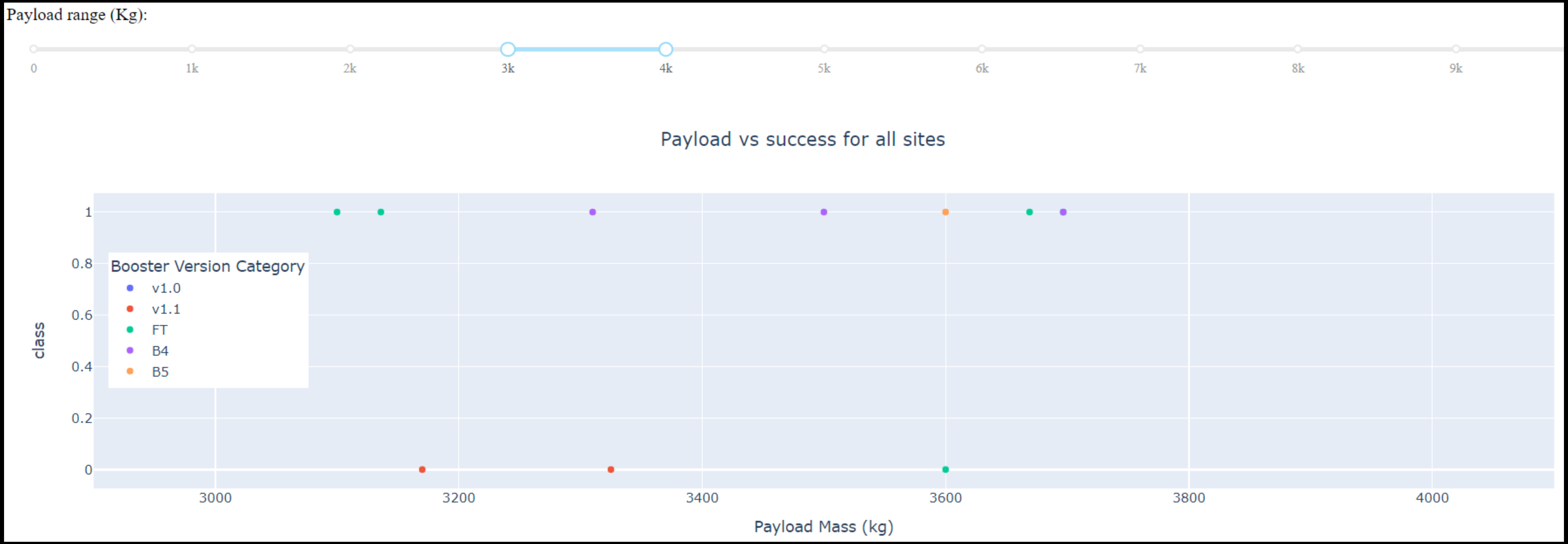
Payload vs launch outcome

- We can see that booster version v1.1 has a pretty bad success rate with only 1 successful launch, on the other hand version FT has the best success rate on launches.
- The range with most successful launches is between 2k and 6k, lighter and heavier payloads tend to do much worse.
- We find the best success rate between payloads of 3k kg and 4k kg.



Payload vs launch outcome (3k-4k kg)

- Most successful launches in this range come from the latter booster versions; B4 and B5.
- Booster version v1.1 contributes 2 of the 3 failed launches, so in this range keeps the trend found out in the general range.

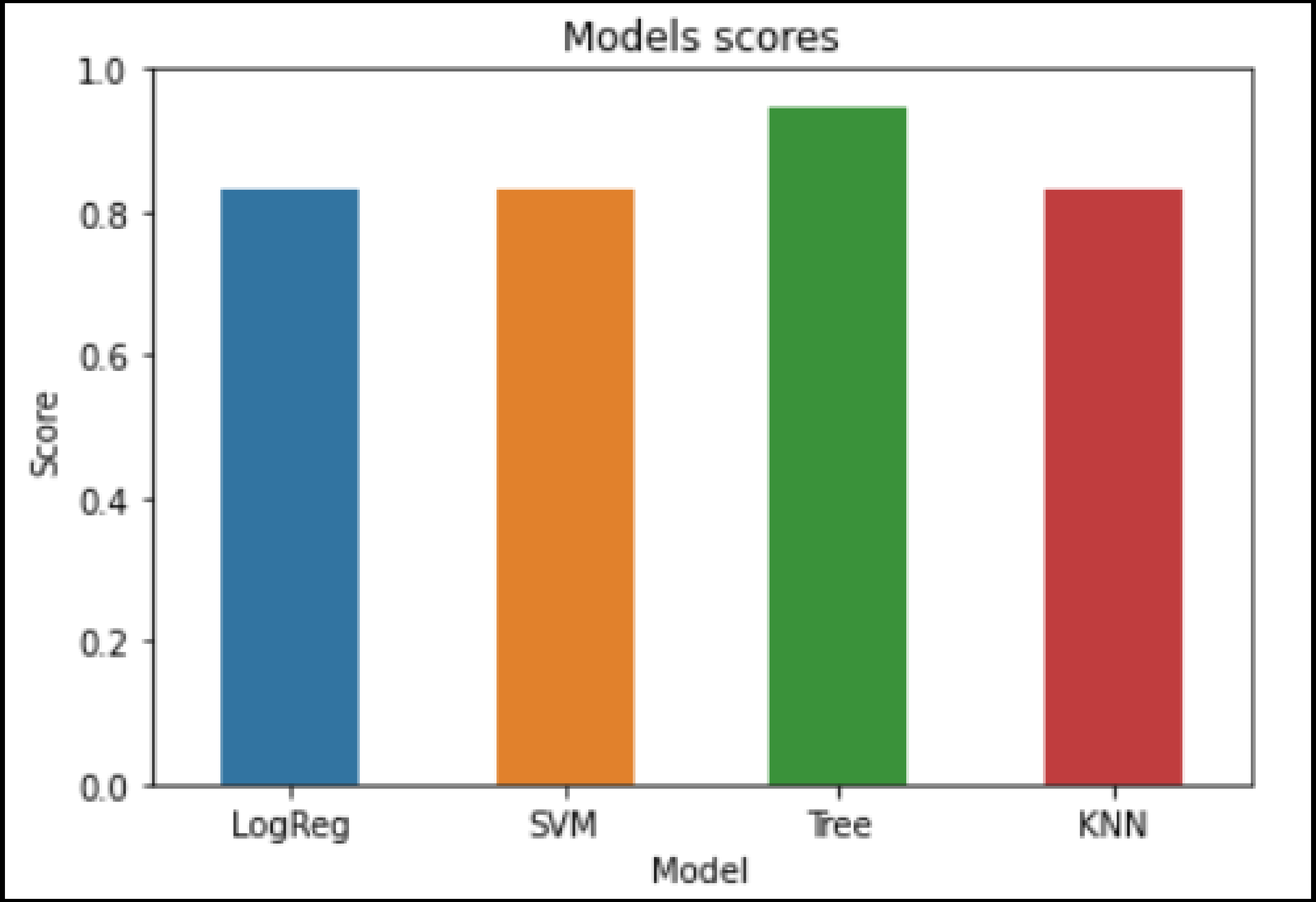


Section 5

Predictive Analysis (Classification)

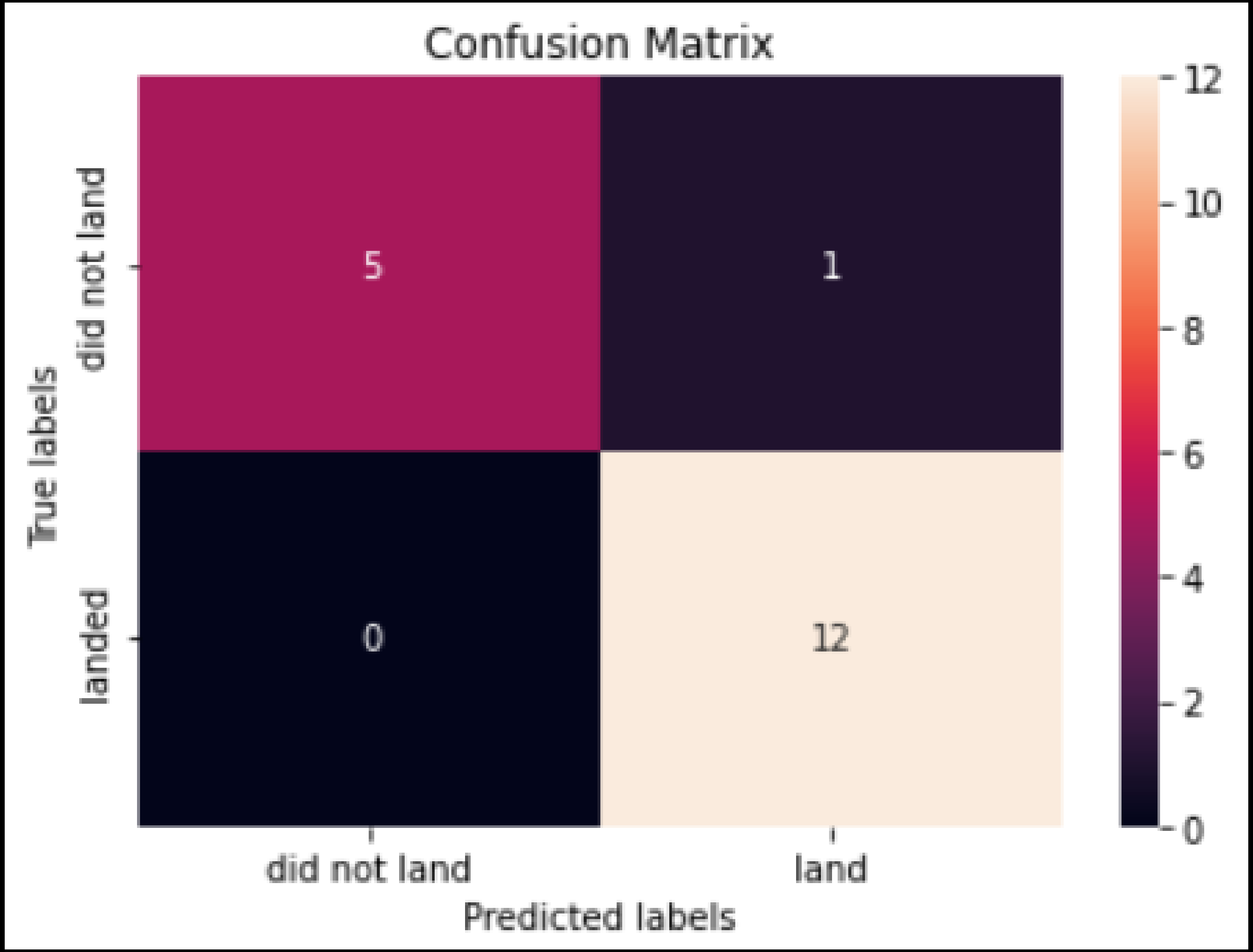
Classification Accuracy

- The score is calculated as the rate of correctly classified tests instances over the total test instances.
- Logistic regression, support vector machine and k-nearest neighbors models achieved around 83% accuracy, while Tree classifier model reached 94%. So our selected model will be Tree classifier.



Tree classifier confusion matrix

- We can that this model correctly classified 17 out of 18 instances in the test data, confirming what was stated before.
- At the same time the model achieved a perfect recall and almost perfect precision with only 1 instance being misclassified as a landing.

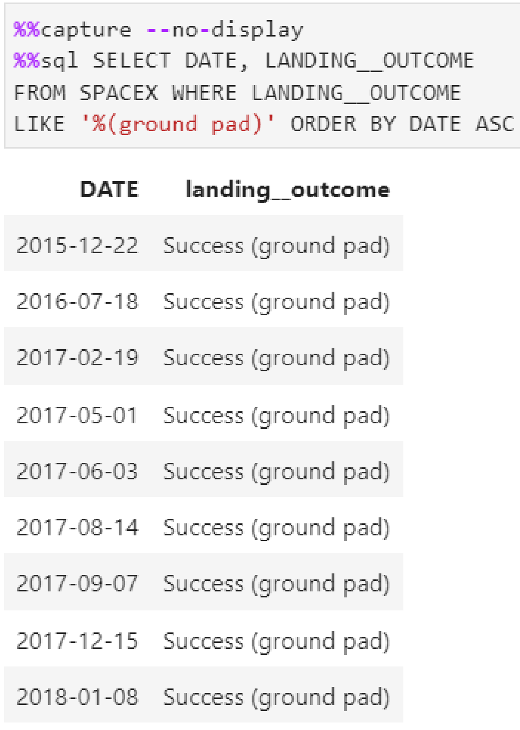


Conclusions

- Most rockets are launched into GTO orbits, and only about 60% perform a successful landing.
- Latter flights have all had a successful landing.
- Launchings done in KSC LC-39A site are probably going to be successful, for other sites the success rate is below 50%.
- We have a model with a 94% accuracy to be able to predict if a future landing will be successful or not, which was our principal question to be answered.

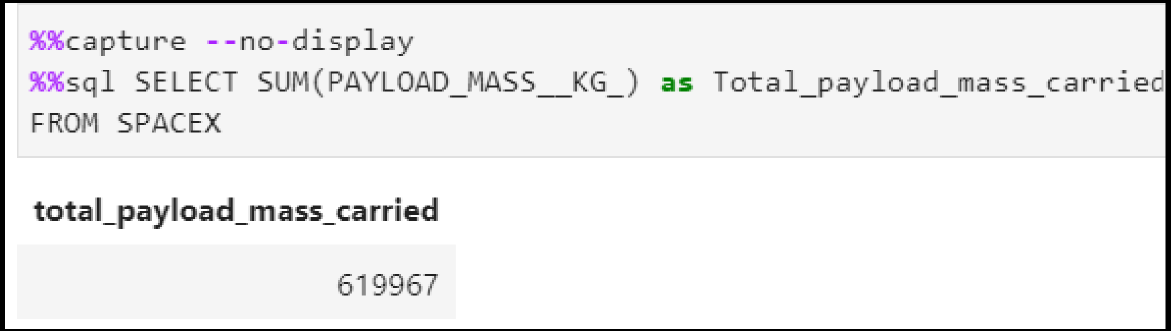
Appendix

- Here we can see all ground pad landings attempted, we notice that all of them were successful.
- And the total payload mass carried by falcon 9 rockets to compare with the total payload mass carried by boosters from NASA (CRS).



```
%%capture --no-display
%%sql SELECT DATE, LANDING__OUTCOME
FROM SPACEX WHERE LANDING__OUTCOME
LIKE '%(ground pad)' ORDER BY DATE ASC
```

DATE	landing__outcome
2015-12-22	Success (ground pad)
2016-07-18	Success (ground pad)
2017-02-19	Success (ground pad)
2017-05-01	Success (ground pad)
2017-06-03	Success (ground pad)
2017-08-14	Success (ground pad)
2017-09-07	Success (ground pad)
2017-12-15	Success (ground pad)
2018-01-08	Success (ground pad)



```
%%capture --no-display
%%sql SELECT SUM(PAYLOAD_MASS__KG_) as Total_payload_mass_carried
FROM SPACEX
```

total_payload_mass_carried
619967

Thank you!