IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kai Schofield
09/28/2023

Outline

- Executive Summary – slide 3
- Introduction – slide 4
- Methodology – slide 5-15
- Results – slide 16-44
- Conclusion – slide 45
- Appendix – slide 46

Executive Summary

- Using python, the data from SpaceX was collected, filtered, and pulled into a usable Pandas format.
- Using mySQL and graphs, the data was explored before going into more detail with Folium, plotly Dash, and machine learning.
- Several factors are key to a successful landing- location, payload mass, booster version, orbit, and an organization's technical expertise gained over time.

Introduction

- Working for a new space launch company, what best practices could be used to compete with SpaceX?
- Which factors best predict a successful landing, which dramatically saves on cost?
- When creating our own rockets, what needs to be considered to maximize chances of our own landings?

Section 1

Methodology

Methodology

Executive Summary

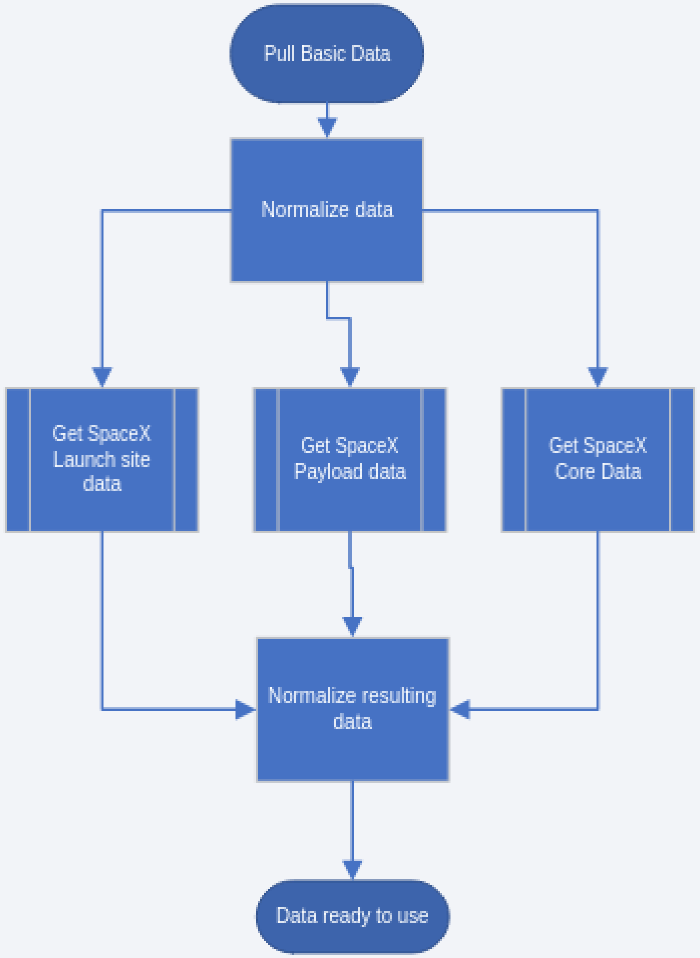
- Data collection methodology:
  - Data was pulled from the SpaceX website and formatted for usability and focus
- Perform data wrangling
  - Landing outcome was converted from categorical to Boolean for ease of use
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using sklearn, several models were fit, transformed, and tested. Hyperparameters were automatically tested to select the best values for each model.

Data Collection

- The requests library in python was used to connect with the SpaceX api and pull data from both Coursera and the SpaceX website
- The initial Coursera dataset was first normalized, and additional data on booster versions, launch sites, payload, and cores was added to the resulting pandas dataframe directly from the SpaceX website.
- This dataset was trimmed to focus on single core launches of the latest booster within a certain timeframe to limit focus and make insights easier to obtain.
- Finally, missing payload values were replaced with the mean payload value

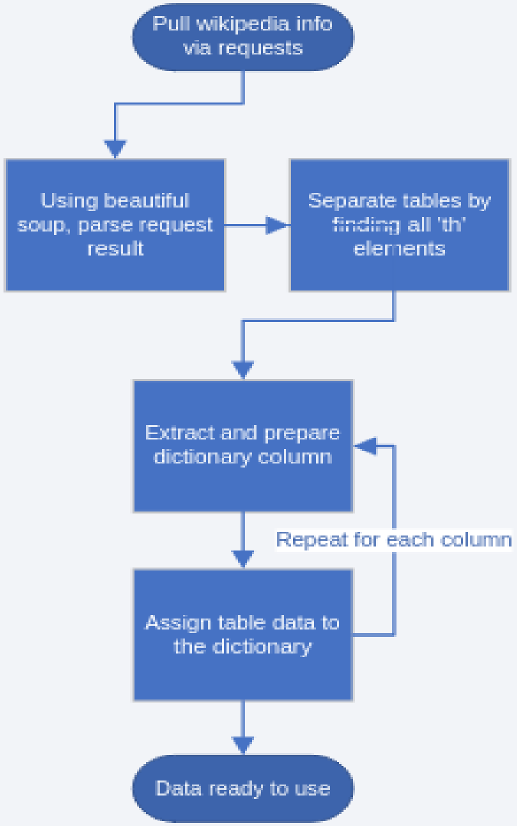
Data Collection – SpaceX API

- Once an initial dataset was assembled from previously scraped data, the SpaceX api was called several times to get auxiliary data from several of their webpages.
- The extra data collected includes booster names, launch site info, payload info, and secondary core characteristics
- The jupyter notebook for data collection can be found on GitHub here



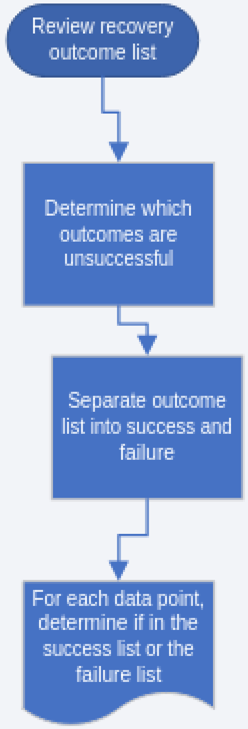
Data Collection - Scraping

- Using requests and beautiful soup, the Wikipedia was scraped for its information on Falcon 9 launches
- The resulting tables were parsed into a python dictionary for each type of data, with the results being collated into a pandas dataframe
- Webscraping notebook can be found on GitHub here


Pull wikipedia info via requests
Using beautiful soup, parse request result
Separate tables by finding all 'th' elements
Extract and prepare dictionary column
Repeat for each column
Assign table data to the dictionary
Data ready to use

Data Wrangling

- The spaceX dataset from the API was pre-processed for clarity
- The number of launches per launch site and orbit were noted for future reference- these factors could potentially influence launch success
- Landing outcomes were then separated into success and failure, so that future models may utilize a binary pass/fail metric.
- GitHub URL for data wrangling is here



Review recovery outcome list

Determine which outcomes are unsuccessful

Separate outcome list into success and failure

For each data point, determine if in the success list or the failure list

EDA with Data Visualization

- Plotted the relationship between payload and launch site vs flight number, to see how these characteristics changed with each launch.
- Separated launch sites to compare how the payloads are spread between launch sites and checked for orbit influences success rate.
- Compared payload mass and flight number to orbit, with color based on success status, checking to see if success rate depends on these factors when combined with orbit. Noted that Polar, LEO, and ISS had superior landing rates for heavy payloads.
- Finally, checked overall success rate over time- it has been increasing
- Link to EDA notebook on github

EDA with SQL

- The first several queries were highly exploratory and only really intended to get a grasp of the data represented in the table
- The next two queries the sum of payload mass for all NASA launches, and the average Falcon 9 v1.1 payload mass.
- Then queries found the first successful landing date and a list of Falcon 9 booster versions. This was followed with a count of successful and failed landings, and then a listing of versions which carried the maximum noted payload.
- The final two queries listed the landing outcomes for 2015, followed by a ranked count of outcomes between 2010 and 2017.
- The GitHub URL for EDA with SQL notebook is here

Build an Interactive Map with Folium

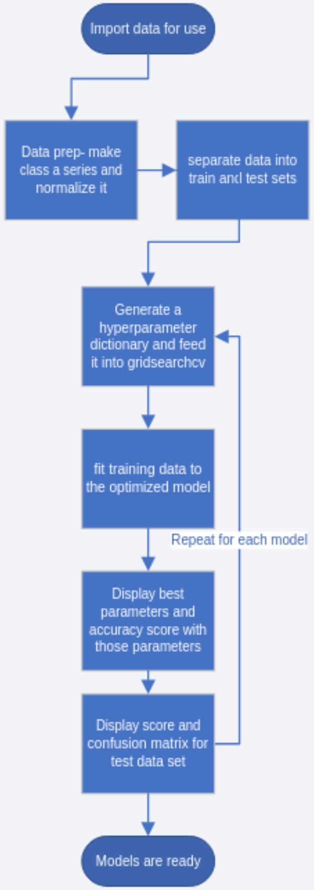
- First, all launch sites were marked on a map, giving a visual representation of where the SpaceX launches have been occuring.
- Then, the success and failures for each site were marked, showing where each outcome tends to occur in a visual manner
- Finally, the distance between a launch site and nearby supported infrastructure was noted as this ease of access may have an impact of success rate
- GitHub URL for the Folium map is here, but note that you may need to trust the notebook to have the folium maps show for you

Build a Dashboard with Plotly Dash

- The dashboard shows two charts- a pie chart of success count by site, and a scatter chart that gives payload mass vs launch site, with success class colored for each instance.
- The pie chart allows for easy comparison of success counts between launch sites and gives an idea of which sites have the most success.
- The second chart allows for zeroing in on success rates for each site as determined by payload carried. The slider allows for fine-tuning of the visualization to focus on desired payload range.
- When the two graphs are combined, a user can see how much of an overall site's success count is represented by a given payload range, and how the overall rate may change with differing payloads.
- The GitHub for the plotly dashboard is here. As this is not a notebook, some work may be needed to visually see the outcome

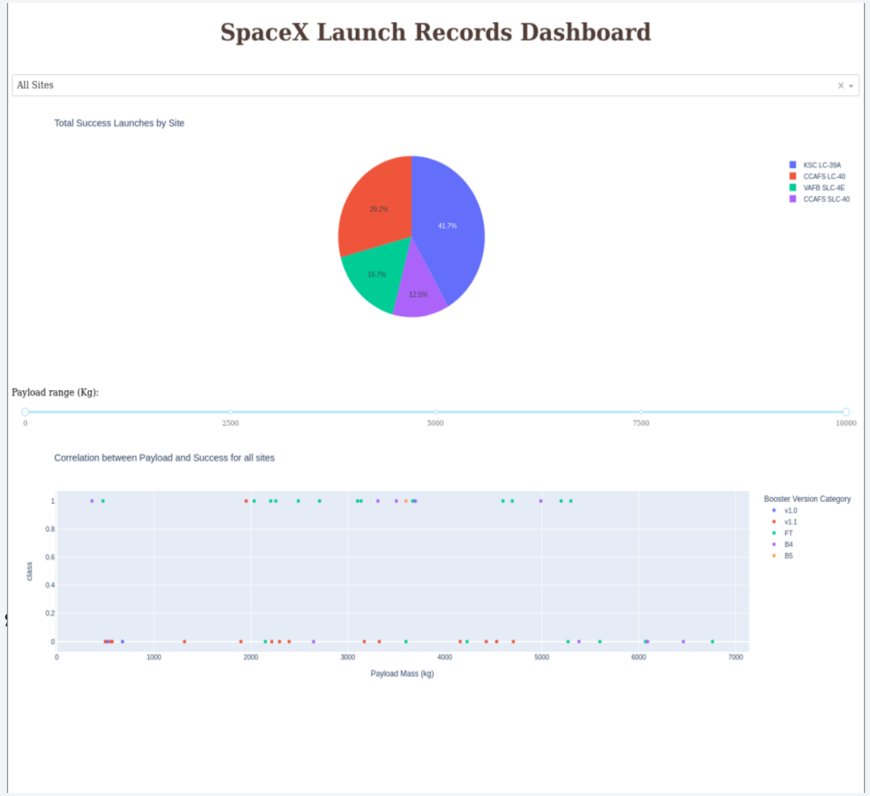
Predictive Analysis (Classification)

- The sklearn library was used to generate models for various types of classification using a train/test data split of 80%/20%
- The GridSearchCV method was used to determine the best hyperparameters for each model
- Using accuracy score on both testing and training data, as well as a confusion matrix, allowed for quick comparison between models.
- Ultimately, in what was probably user error each model had the same test score. However, the decision tree had the highest score for training data, and since is the one difference, decision tree is probably the best model.
- Here is the lab- feel free to leave a commment about this issue


Import data for use
Data prep- make class a series and normalize it
separate data into train and test sets
Generate a hyperparameter dictionary and feed it into gridsearchcv
fit training data to the optimized model
Repeat for each model
Display best parameters and accuracy score with those parameters
Display score and confusion matrix for test data set
Models are ready

Results

- Orbit, payload mass, and launch site are all predictors for a successful landing. The best orbits were ones where the orbit is highly circular and predictable, or ones where the satellite is easy to monitor (VLEO). The difference between PO and SSO, despite being ostensibly similar orbits, suggests regularity contributes to success rate. The flight number and time also corresponded to success, as experience with launches helps ensure a proper landing.
- In the SQL data analysis, ground landings were more successful than ocean landings, and most launches were on the lighter end. Most launches during the observed time were lighter, likely due to cost and complication. The heavier payloads did still land frequently, so some other metric is likely why they are fewer in number.
- To the right is a screenshot of the SpaceX launch records dashboard, which allows for easy visualization of total success launches by site, and correlation between payload and success for any or all site(s)
- All the models appeared to operate similarly on the test data, though this may be user error. Of the tested models, the decision tree worked best of the training data.



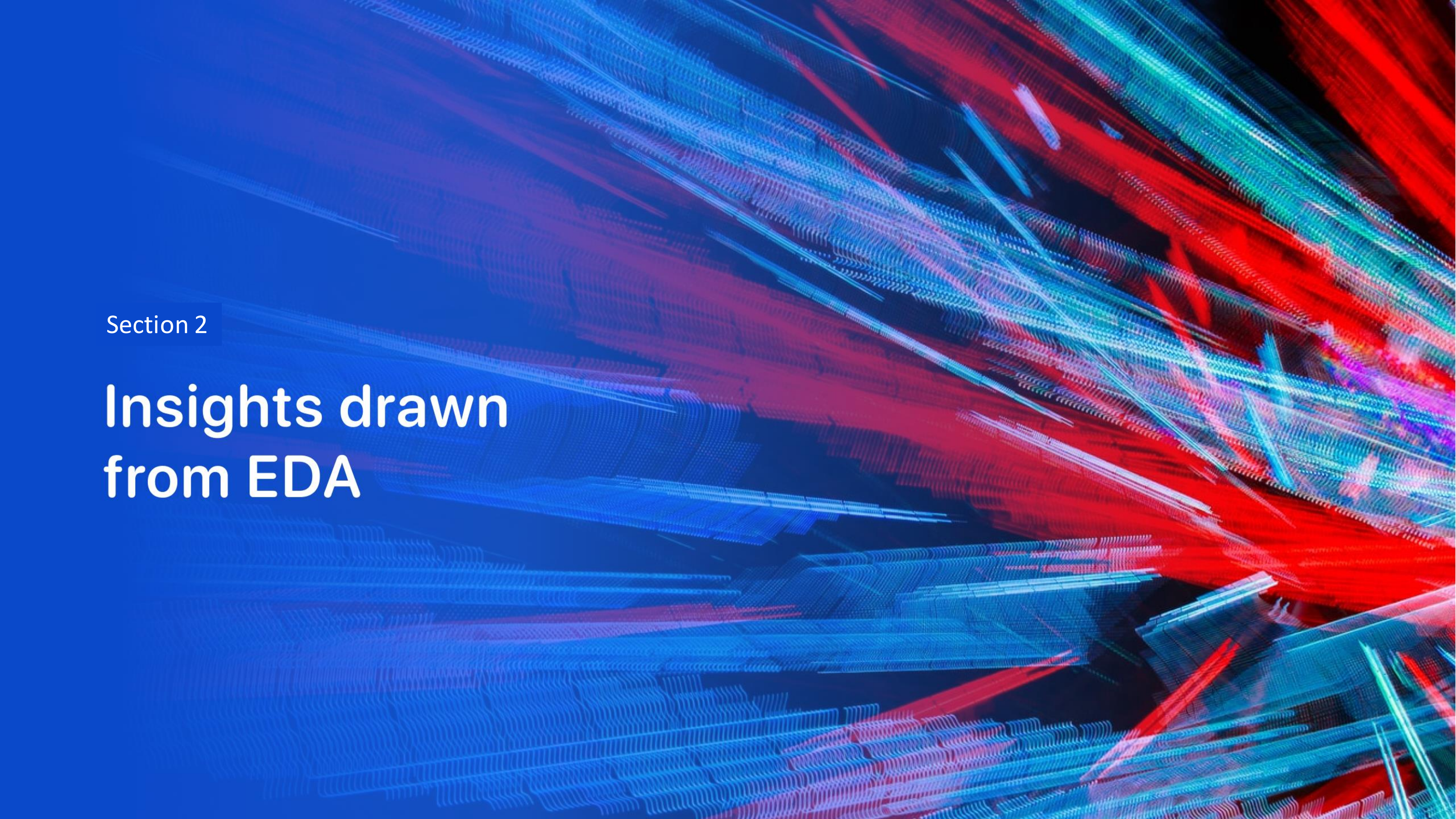
SpaceX Launch Records Dashboard

All Sites

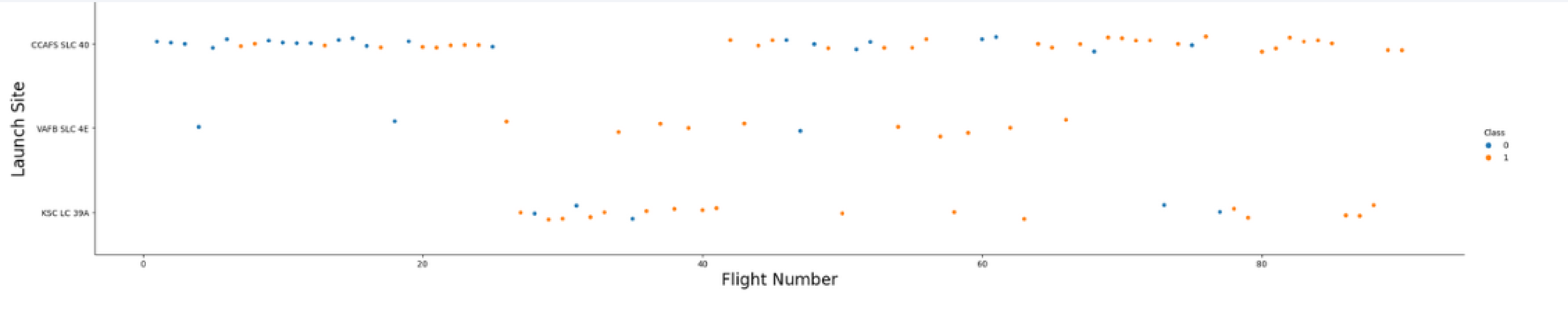
Payload range (Kg):

Section 2

Insights drawn from EDA

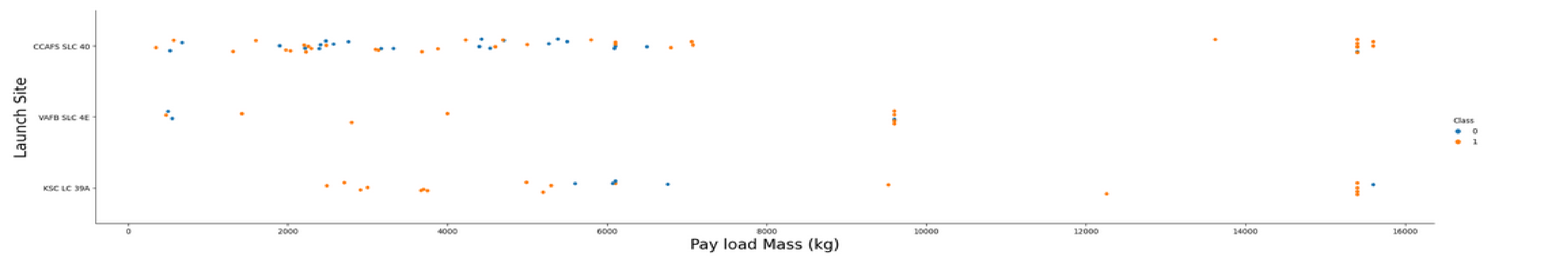


Flight Number vs. Launch Site



One launch site was largely used for early flights, before bringing up two more. The VAFB was no longer used after flight ~70, likely due to its location on the opposite side of the USA from the other two sites

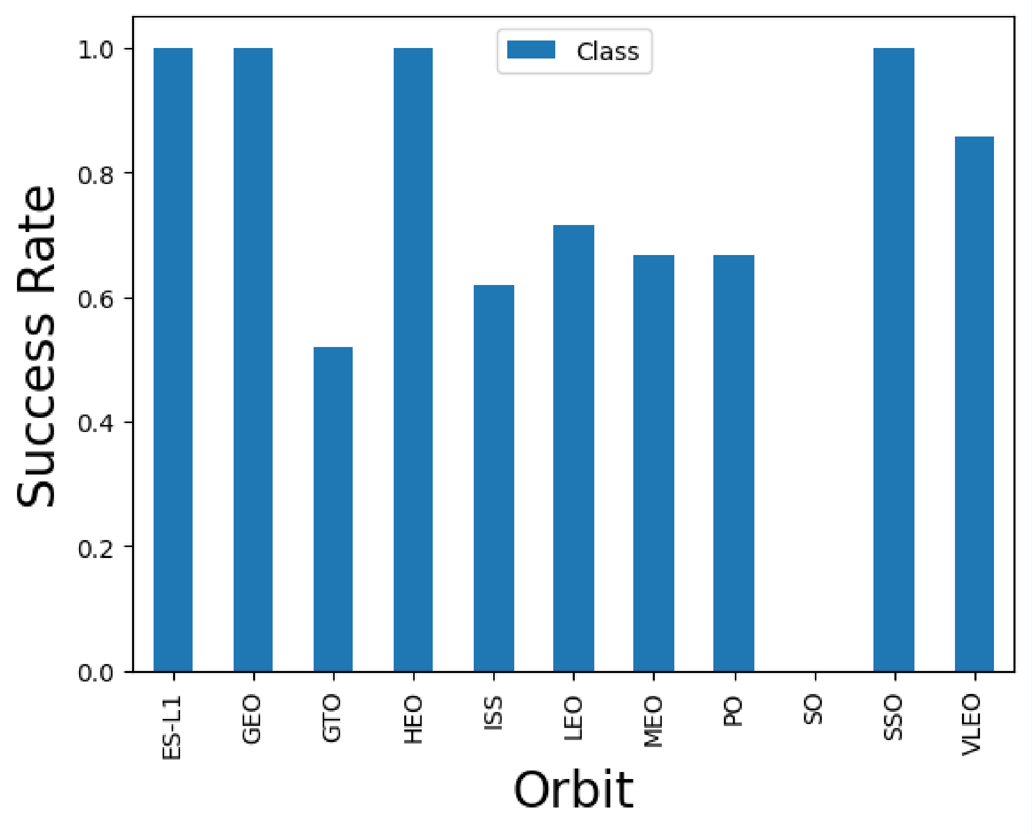
Payload vs. Launch Site



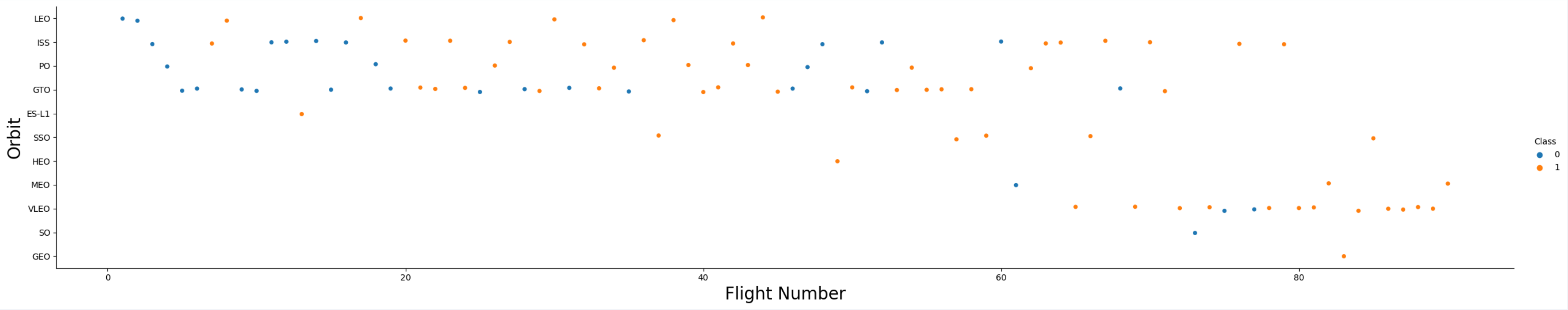
- Certain launch sites have higher success rates, with high payloads showing a higher overall success rate.
- Low to medium payloads have varying success rate, with every launch site holding a moderate success rate in this range.

Success Rate vs. Orbit Type

- Certain orbits are more elliptic or intermediate steps to a more stable, circular orbit.
- The more circular and stable an orbit, the better the landing success rate. Many of the highest success rate orbits are synchronous.
- Aside from these, lower orbits had higher success rates with LEO and VLEO beating the remaining orbits

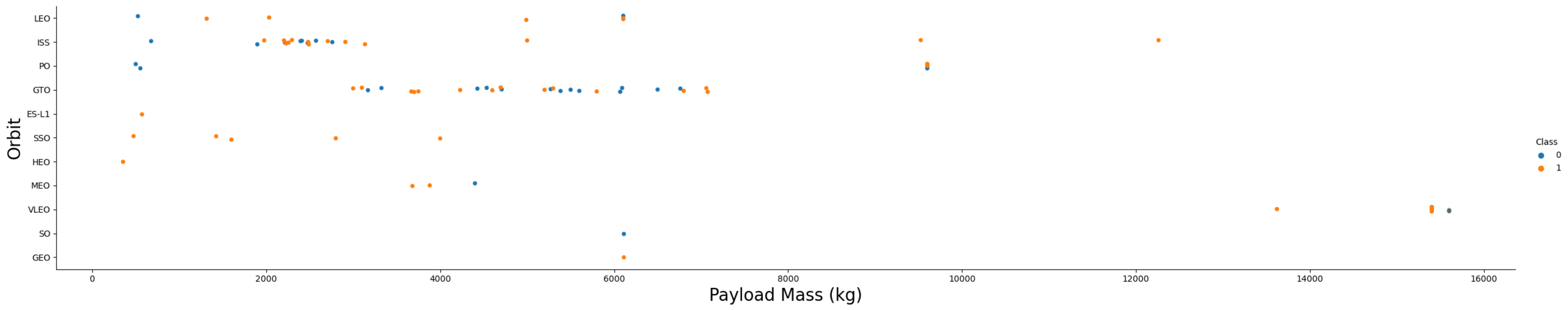


Flight Number vs. Orbit Type



- Early orbits were among common satellite orbits and showed low success rate as initially landings were rare.
- The highest success rates are among orbits that were not attempted until later in the program, and were attempted relatively few times.
- Orbits attempted early on had higher failure rates, with failures decreasing over time.

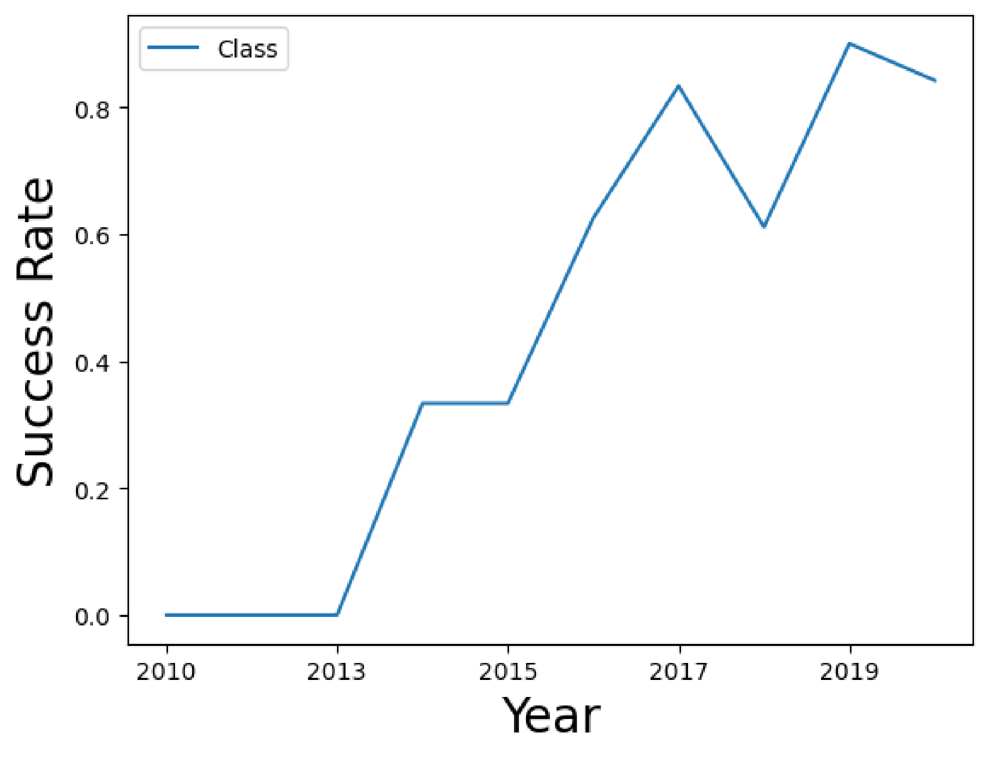
Payload vs. Orbit Type



- Only certain orbits saw the highest payloads, namely in ISS and VLEO- both of which are low orbits and need less fuel to complete for a given payload.
- Most orbits have a cluster of launches in a certain payload range that seems most common for that orbit.
- Orbits have similar payloads show both success and failure, suggesting that payload isn't the strongest predictor for success.

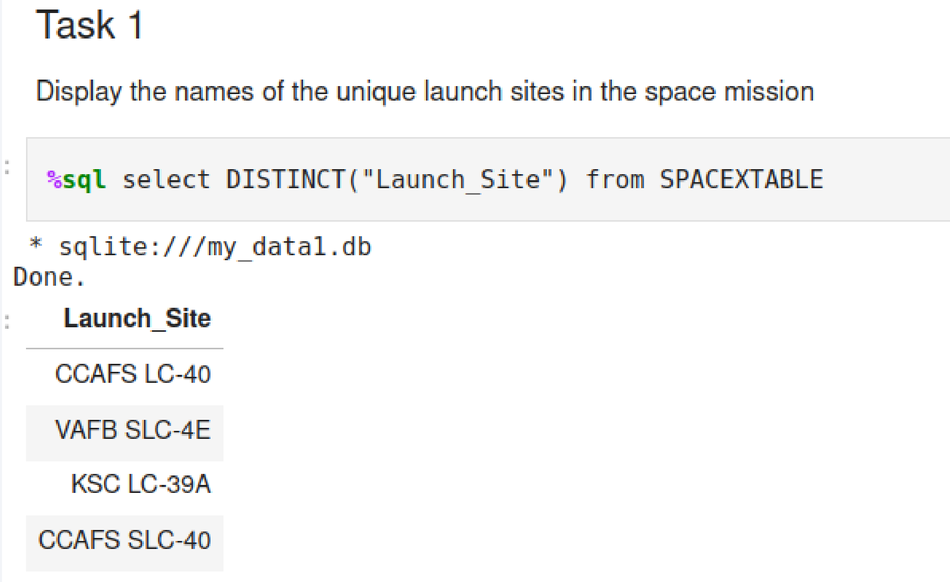
Launch Success Yearly Trend

- Over time the success rate of landings has increased as a general trend.
- The time periods of 2014-2015 and 2017-2018 are notable exceptions. Other factors temporarily disrupted this trend.



All Launch Site Names

- The unique launch site names are CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.
- By selecting distinct launch sites we can quickly grab all unique names from the database.



Task 1

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

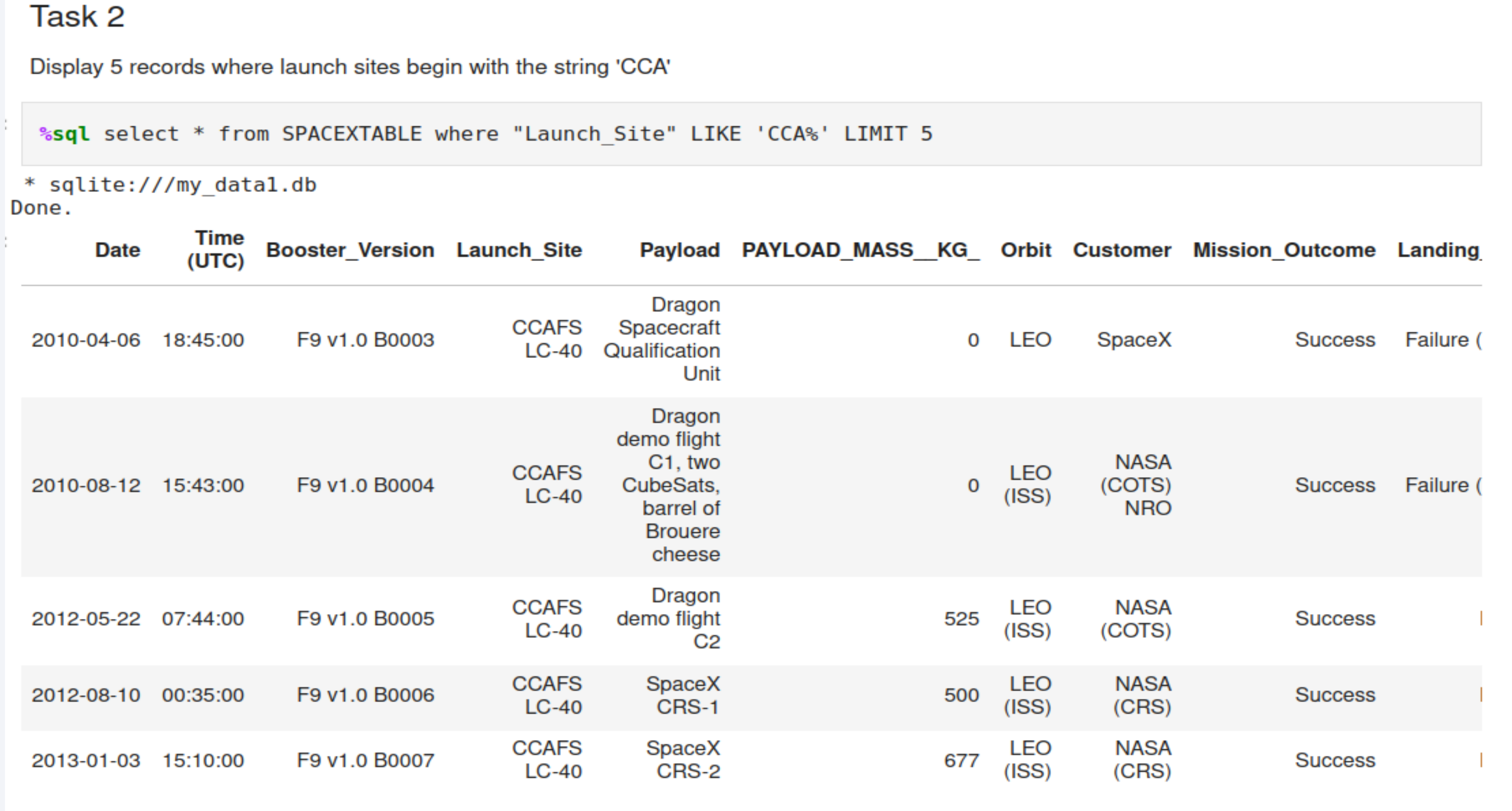
```
%sql select DISTINCT("Launch_Site") from SPACEXTABLE
```

```
 * 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'

- Selecting flights where the launch site is like CCA% asks for any launch site beginning with CCA, but allowing for anything to come after.
- This query would show the results of both CCAFS LC-40 and CCAFS SLC-40, but for ease of presentation only the first 5 results are shown here.



Task 2

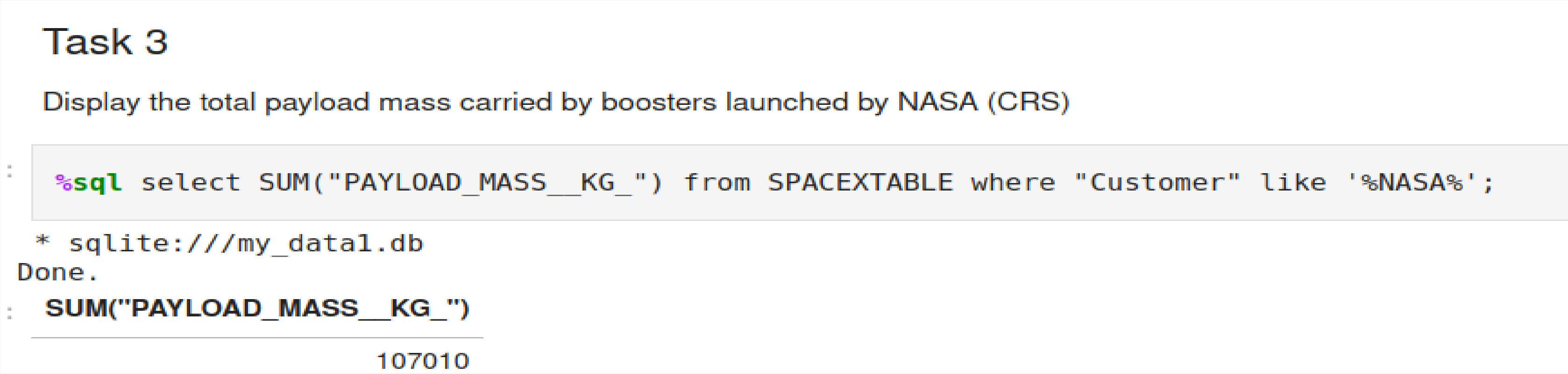
Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * from SPACEXTABLE 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_
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (
2010-08-12	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 (
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass



Task 3

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

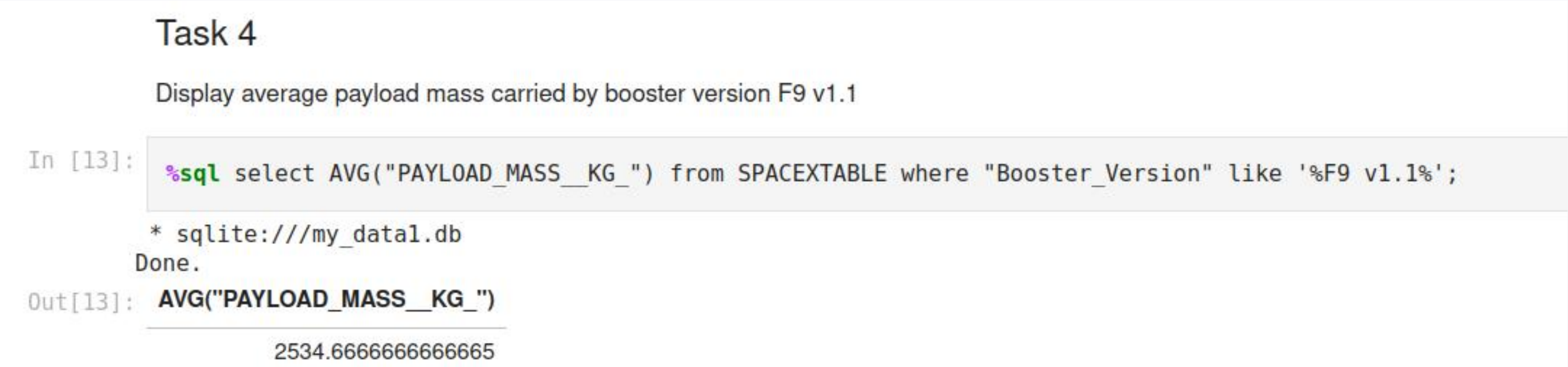
```
%sql select SUM("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Customer" like '%NASA%';
```

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

SUM("PAYLOAD_MASS__KG_")
107010

- The total payload carried by boosters on behalf of NASA was 107,010 kg.
- By filtering the table to only take customers with NASA in the name, all launches on behalf of NASA were queried.
- The result of the where clause had all payload masses summed to obtain this result

Average Payload Mass by F9 v1.1



Task 4

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

```
In [13]:  %sql select AVG("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Booster_Version" like '%F9 v1.1%';
```

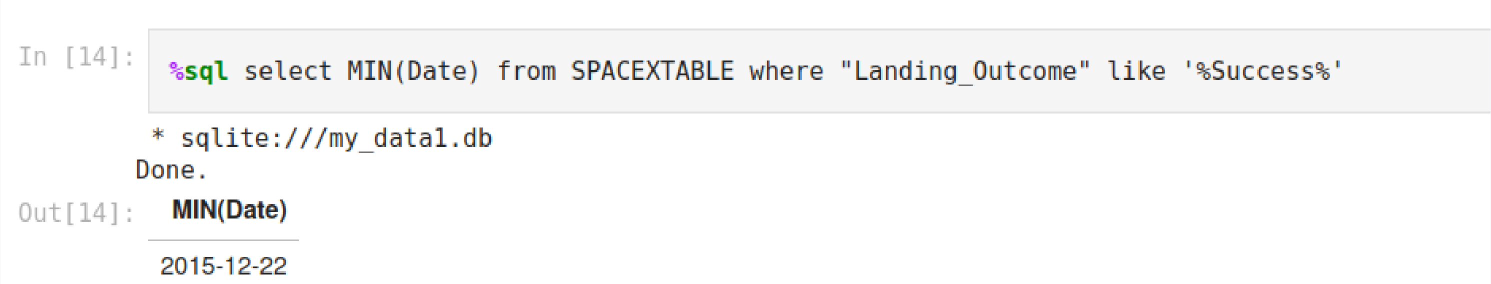
```
 * sqlite:///my_data1.db
Done.
```

Out[13]:

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

- The average payload mass for F9 v1.1 boosters was 2534.667 kg.
- Using a where clause, we limit booster versions to any containing the phrase F9 v1.1
- We then average the payload mass kg column using the AVG function.

First Successful Ground Landing Date



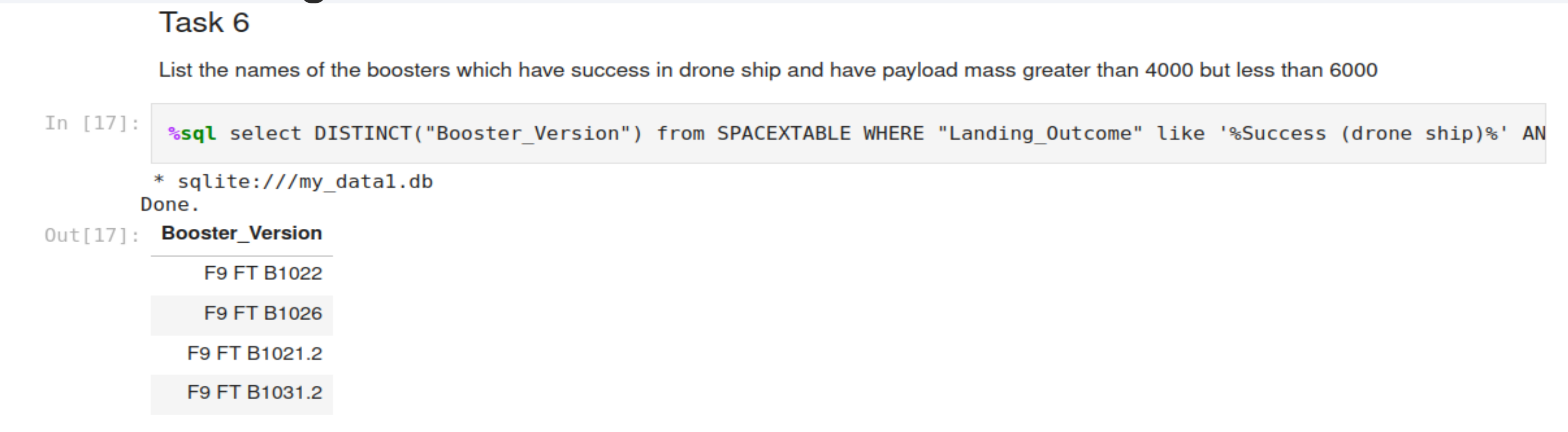
```
In [14]:  %sql select MIN(Date) from SPACEXTABLE where "Landing_Outcome" like '%Success%'

           * sqlite:///my_data1.db
          Done.
Out[14]:   MIN(Date)
          2015-12-22
```

- By querying the minimum (earliest) date where landing outcome contained the word success, the first landing was found to occur on December 12th, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- The list of boosters which have landed on a drone ship with a payload above 4000 and less than 6000 is F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2.
- This was found by querying for distinct booster versions where the landing outcome was successful on a droneship and the payload mass was in the correct range.



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

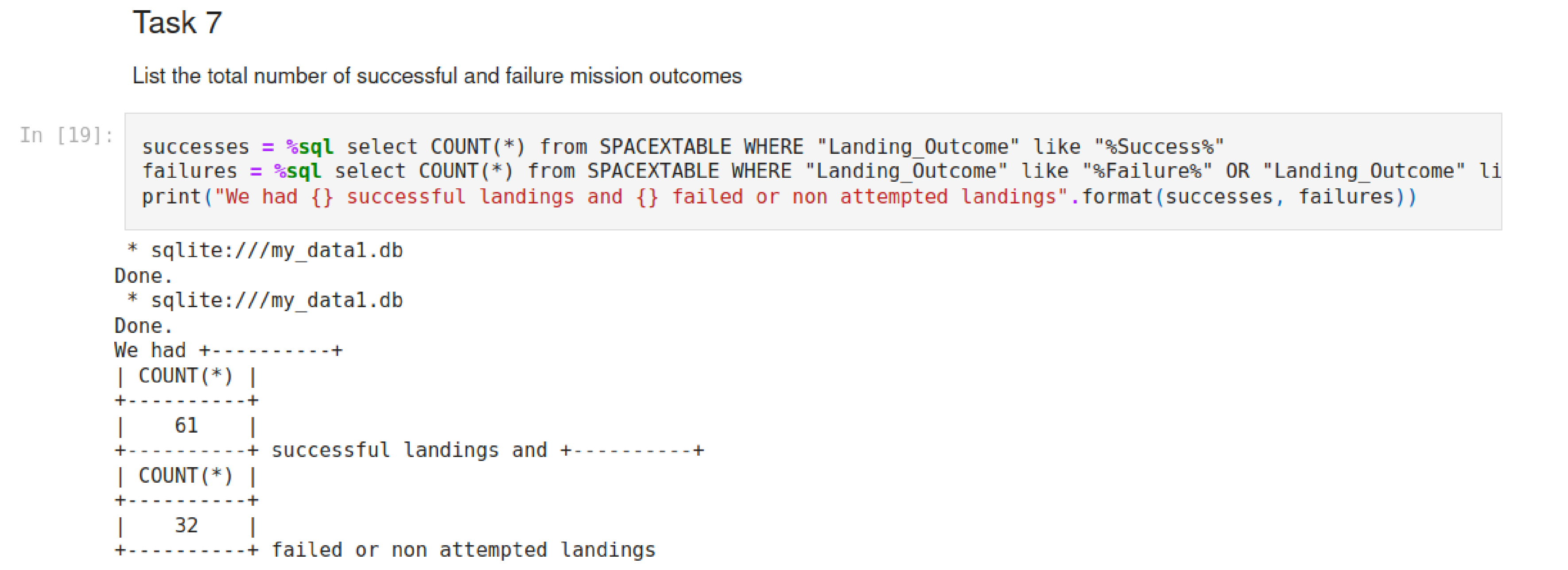
```
In [17]:  %sql select DISTINCT("Booster_Version") from SPACEXTABLE WHERE "Landing_Outcome" like '%Success (drone ship)%' AN
```

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

Out[17]:

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

Total Number of Successful and Failure Mission Outcomes



Task 7

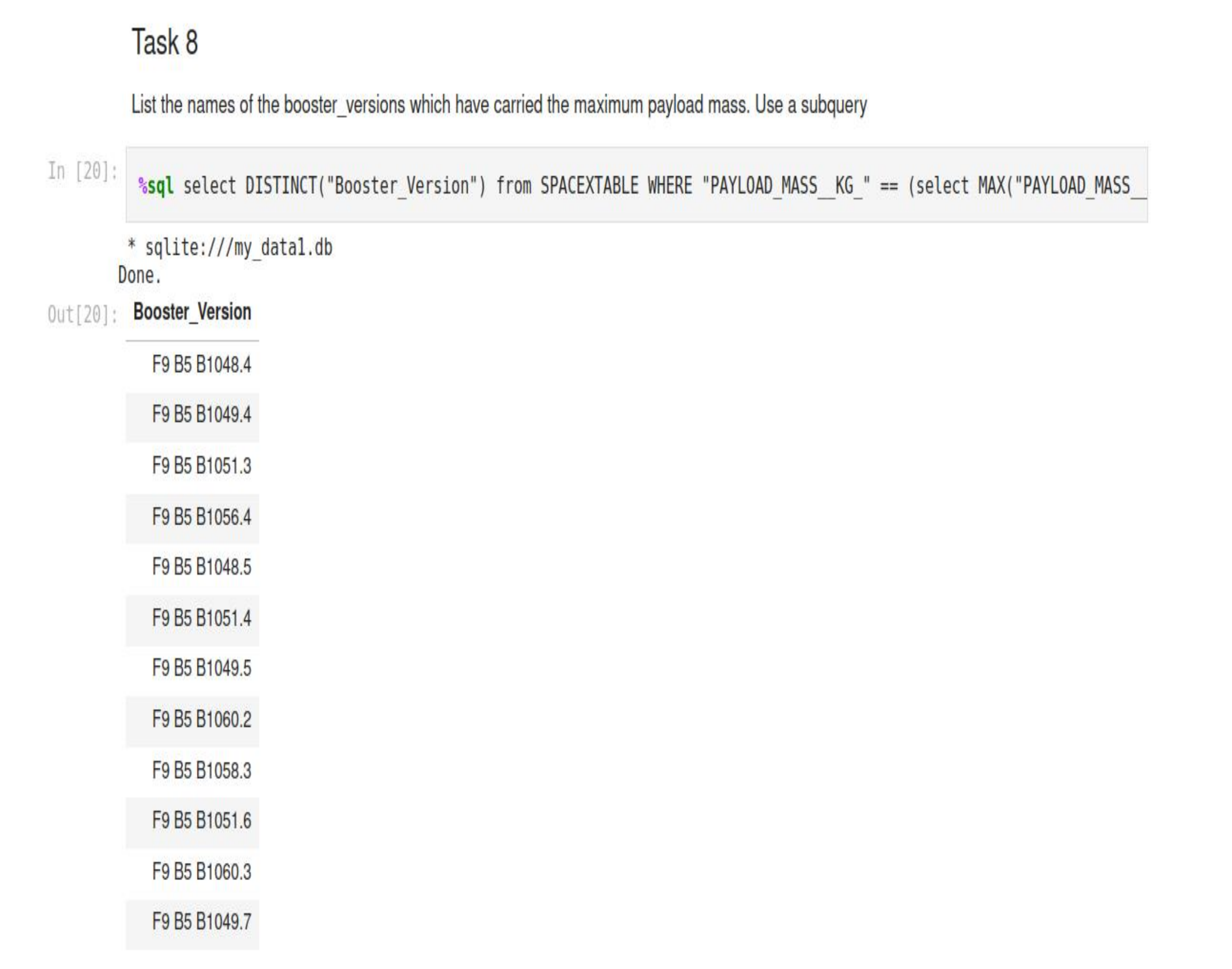
List the total number of successful and failure mission outcomes

```
In [19]: successes = %sql select COUNT(*) from SPACEXTABLE WHERE "Landing_Outcome" like "%Success%"
         failures = %sql select COUNT(*) from SPACEXTABLE WHERE "Landing_Outcome" like "%Failure%" OR "Landing_Outcome" li
         print("We had {} successful landings and {} failed or non attempted landings".format(successes, failures))
```

```
 * sqlite:///my_data1.db
Done.
 * sqlite:///my_data1.db
Done.
We had +----------+
| COUNT(*) |
+----------+
|    61    |
+----------+ successful landings and +----------+
| COUNT(*) |
+----------+
|    32    |
+----------+ failed or non attempted landings
```

- There were 61 successful landings and 32 failures. Using SQL, successful landings were outcomes that contained the word success. Failures were defined as failure or no attempt, since both instances lead to a non landing event.

Boosters Carried Maximum Payload



Task 8

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

```
In [20]: %sql select DISTINCT("Booster_Version") from SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" == (select MAX("PAYLOAD_MASS__
```

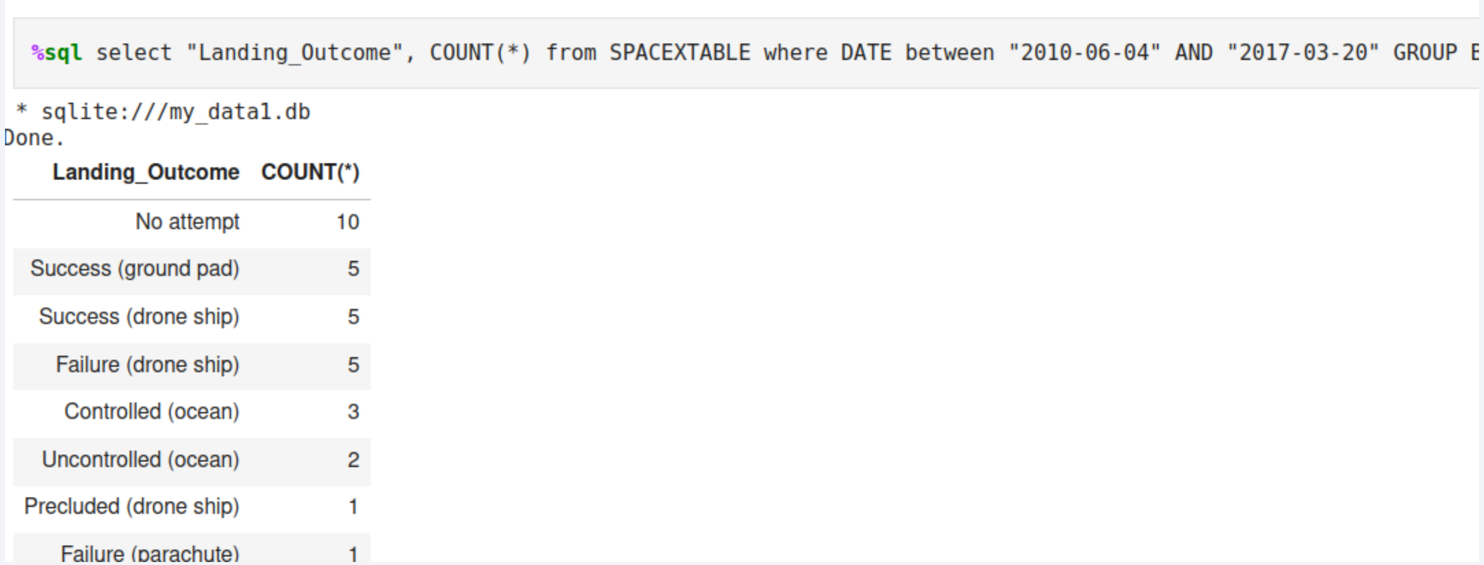
```
 * sqlite:///my_data1.db
Done.
```

Out[20]:

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

- Twelve separate boosters have carried the maximum payload mass, as visible on the left. All of them are F9 B9 rockets of varying versions.
- The query to obtain this list selected boosters used in launches where their mass was equal to the mass payload, as found via subquery.

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



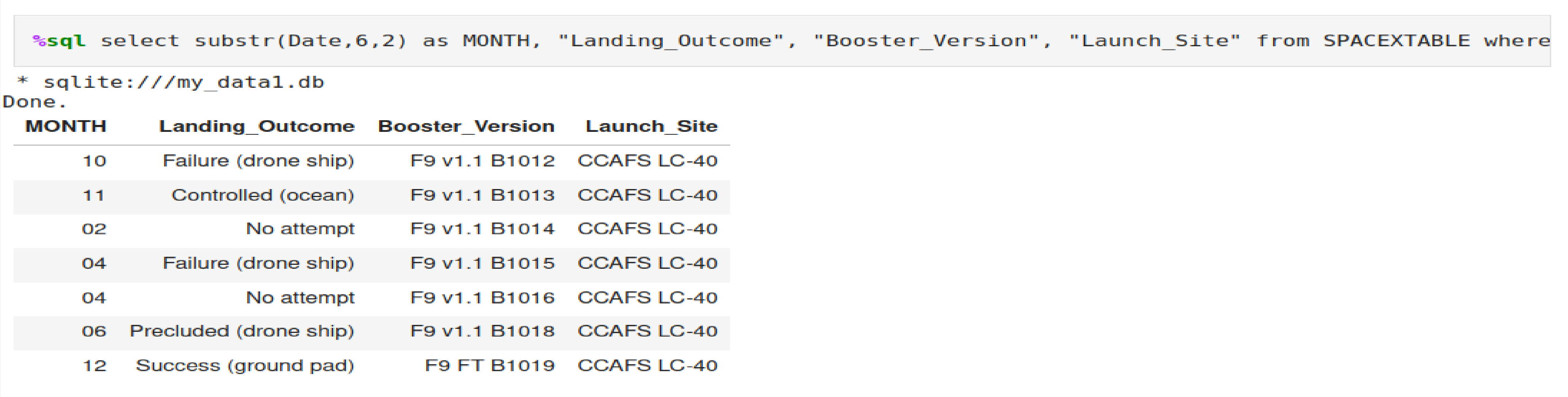
```
%sql select "Landing_Outcome", COUNT(*) from SPACEXTABLE where DATE between "2010-06-04" AND "2017-03-20" GROUP B
```

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

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

- The landing outcomes in the selected time range are presented in descending order above, from the most common outcome of No attempt to the least common of failure (parachute).
- The query is cut off above, but involves selecting landing outcome and count between two dates and grouping by landing outcome. The results are ordered by descending count.

2015 Launch Records



```
%sql select substr(Date,6,2) as MONTH, "Landing_Outcome", "Booster_Version", "Launch_Site" from SPACEXTABLE where
```

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

MONTH	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
11	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
02	No attempt	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	No attempt	F9 v1.1 B1016	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
12	Success (ground pad)	F9 FT B1019	CCAFS LC-40

- The record of months, landing outcomes, booster versions, and launch sites for 2015 are presented above. Three of these outcomes involved a drone ship- the first, fourth, and sixth in the above image

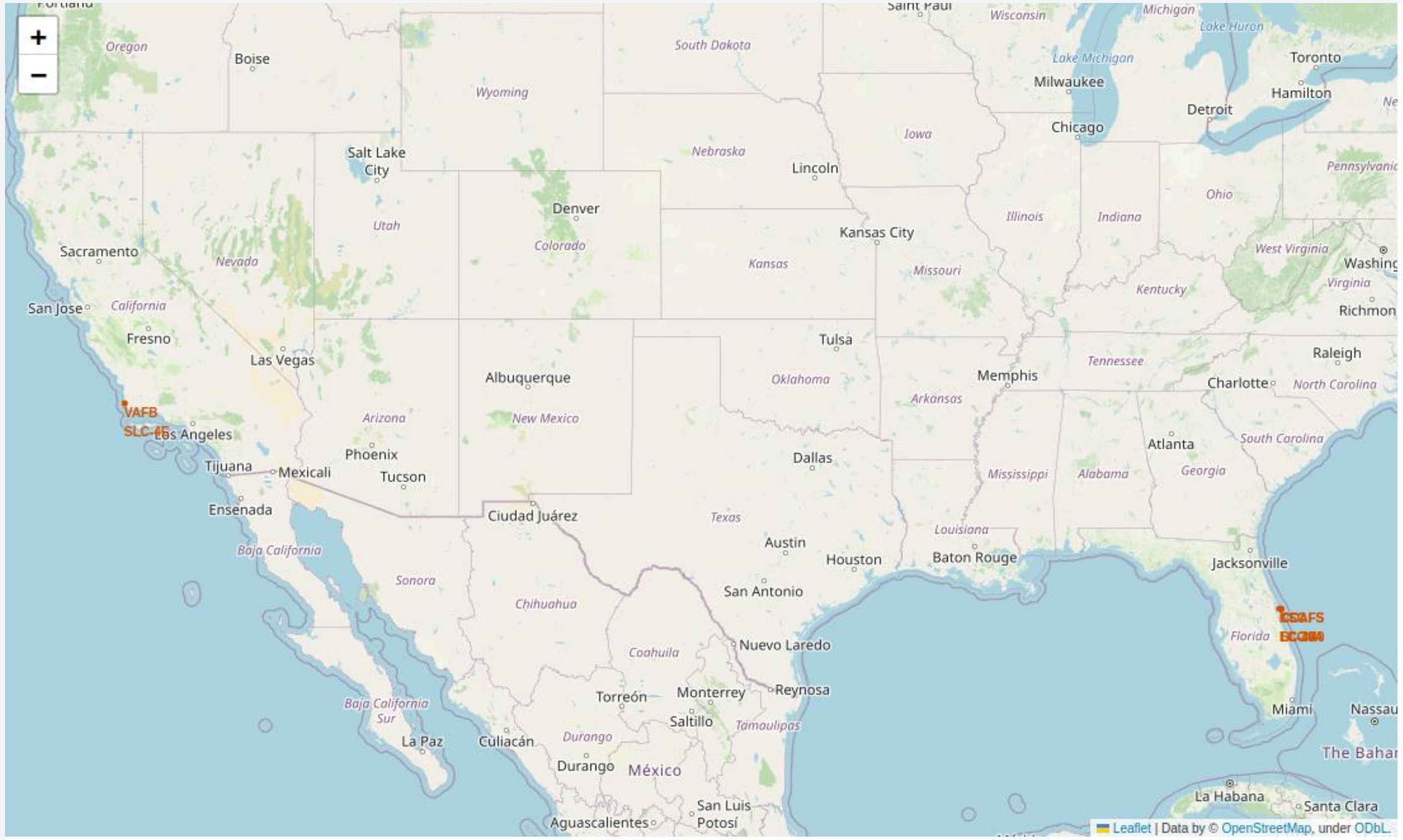
- The SQL query selects desired data from the SPACEXTABLE where the year is 2015.

Section 3

Launch Sites Proximities Analysis

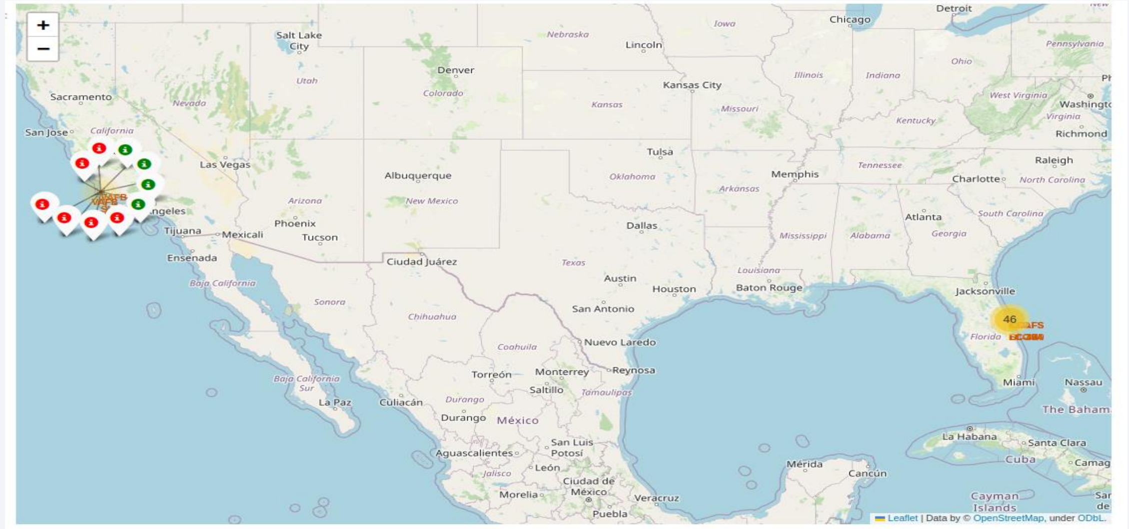
SpaceX Launch Site Map

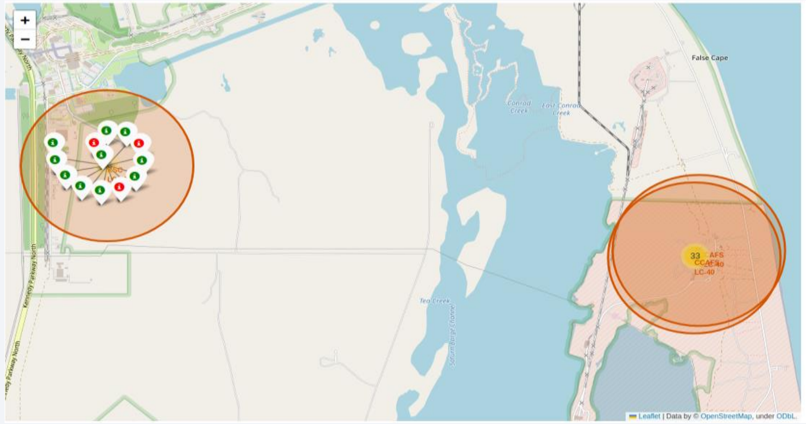
- The image to the right is a folium generated map that shows all SpaceX launch sites .
- There are four launch sites present, with three on the east coast of the USA in Florida, and the other in California. All sites are very close to the water as well as several modes of access



SpaceX Launch Site Outcome Map

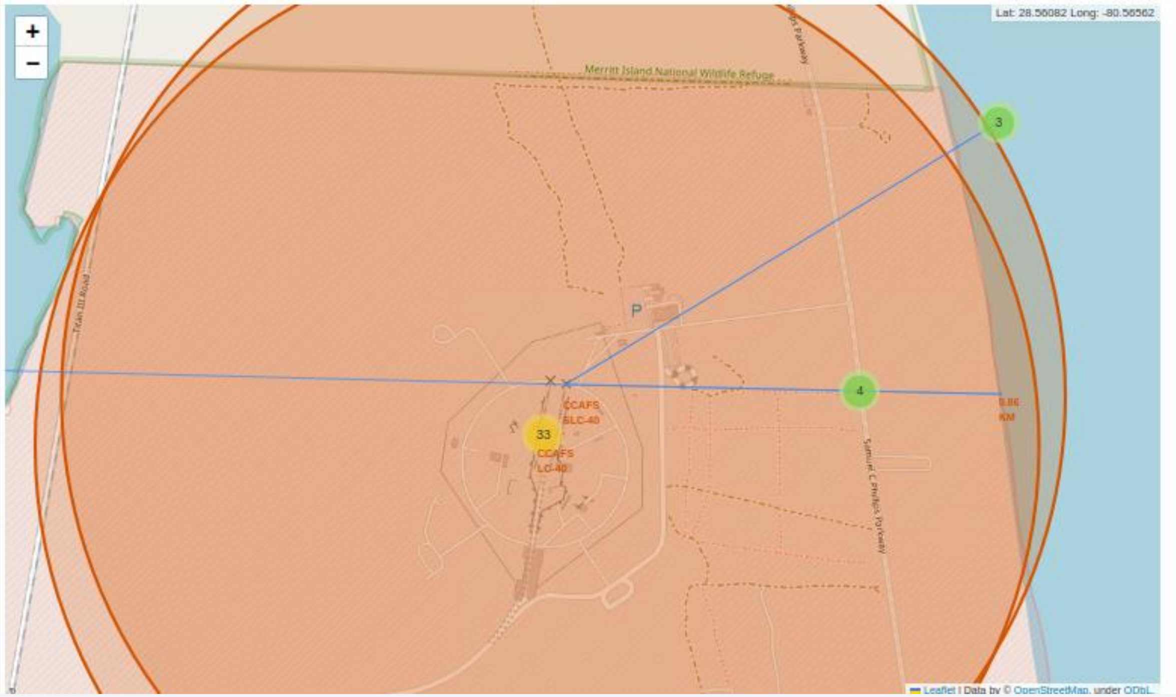
- In the first map on the right the same launch areas are visible once more, only this time the launch outcomes associated with each site are visible and color coded.
- The launch site in California has been clicked on to show color coded outcomes (green for success, red for failure).
- The three launch sites in Florida Are clustered closely, but clicking on the right launch sites could zoom in and show the launch clusters for each site. The second map shows one such view.

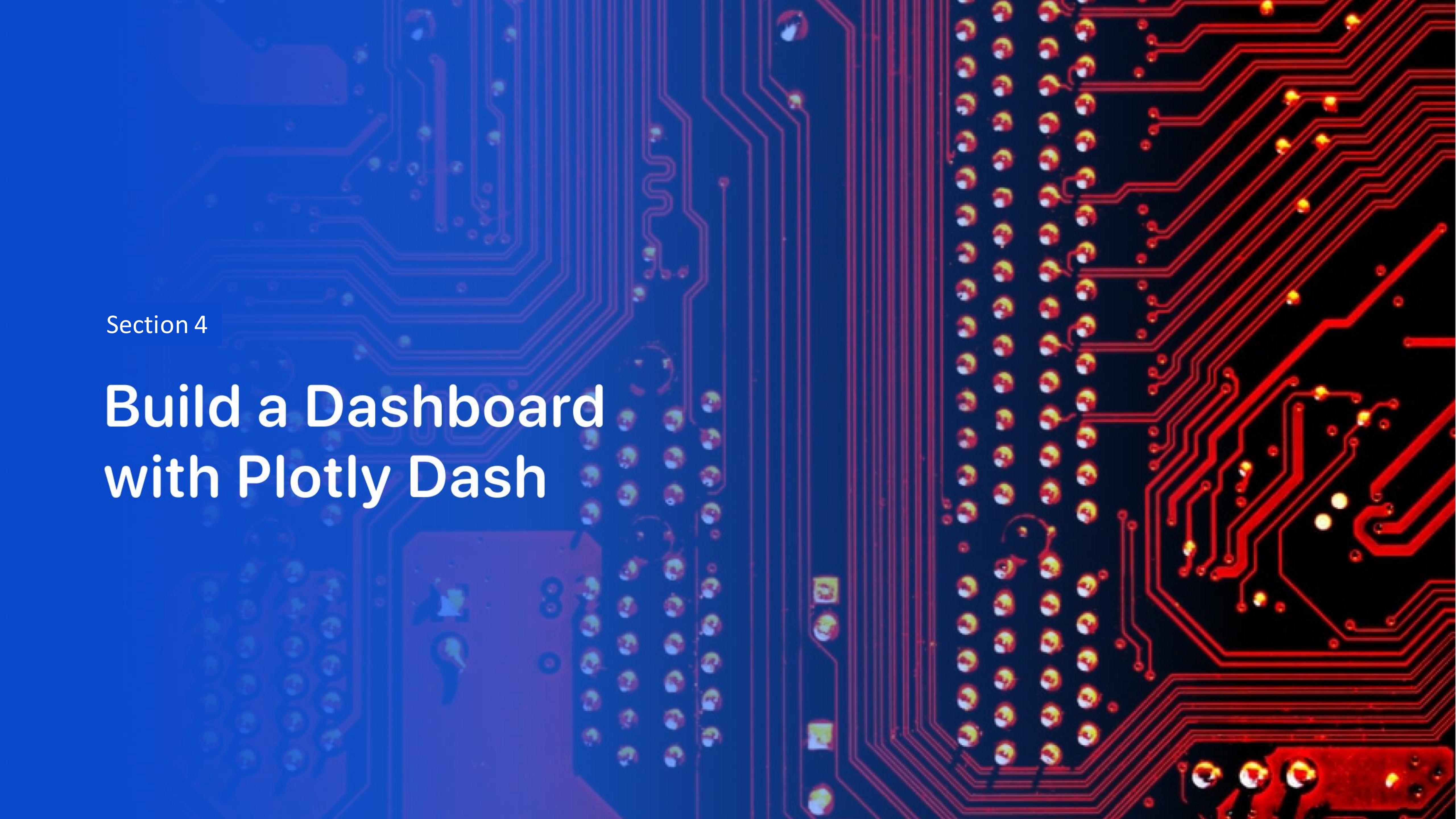




SpaceX Launch Site Proximity Map

- The map to the right highlights one example of a launch site's proximity to nearby access methods, coastline, and population centers.
- This particular launch site is ~0.86km from the coast, ~0.59km from the nearest highway, ~1.27km from the nearest railway, and ~21.65km from the nearest population center.
- These distances imply that there may be some benefit to a launch site being close to the coast, highway, and railway, with city centers kept relatively close but not so close as to be in danger of stray rockets.
- In this case, the railway being further away is interesting- the connections are such that deliveries to the site need to be carried by another delivery method that final 1.27km, while the highways come much closer to the actual launch pad.

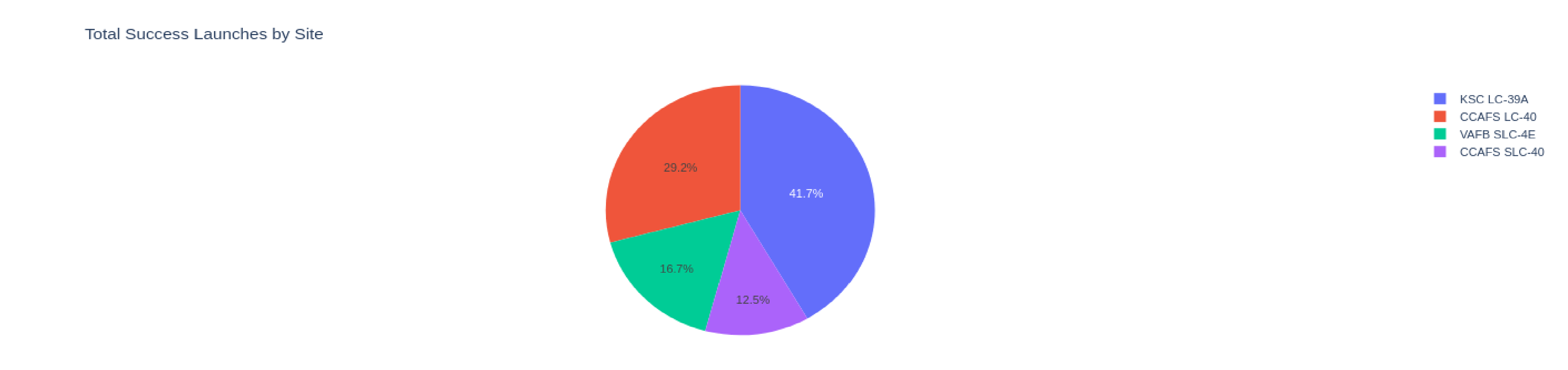

+
−
Lat: 28.56082 Long: -80.56562
Merritt Island National Wildlife Refuge
Titan III Road
3
P
4
0.86
KM
CCAFS
SLC-40
33
CCAFS
LC-40
Samuel C Phillips Parkway
Leaflet | Data by © OpenStreetMap, under ODbL.



Section 4

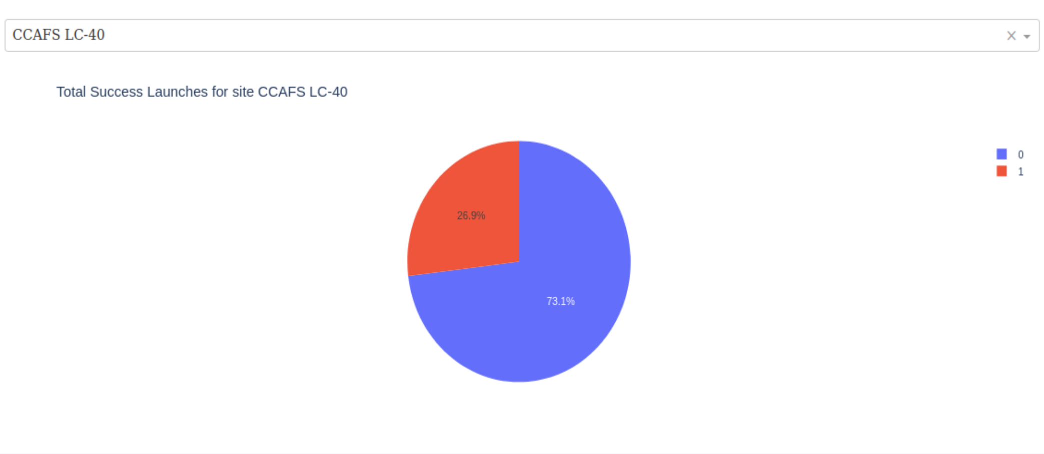
Build a Dashboard with Plotly Dash

SpaceX Launches by Site

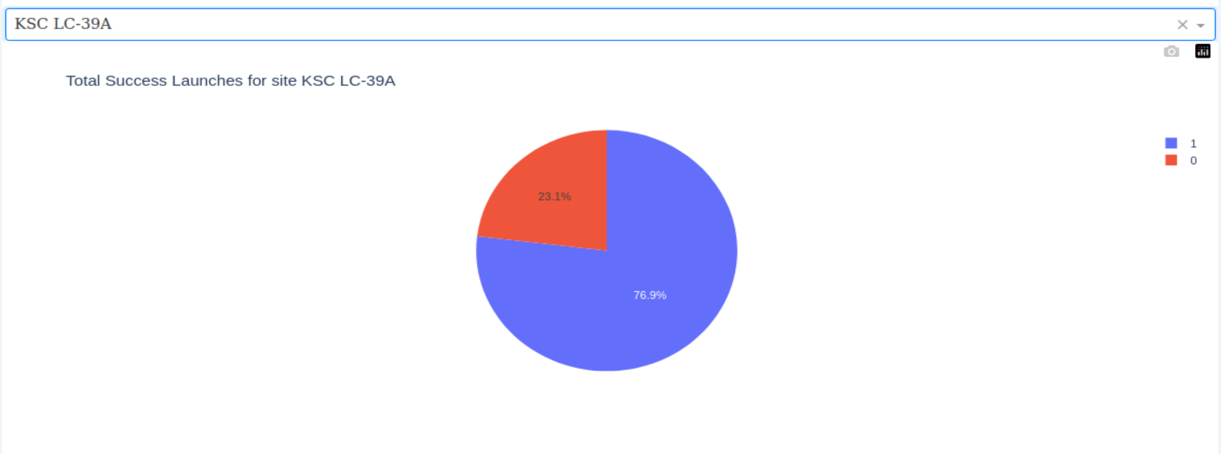


- Location matters for a successful landing- only 16.7% occurred on the west coast of the USA- that's four landings. Rockets are typically launched to the east, so a recovery that puts the stage over water is often safer than one that puts it over the continental US.
- Of the Florida based launch sites, 41.7%, or 10, of them are from the KSC. This was leased from NASA and benefits from their existing infrastructure and expertise.

Landing Site Comparison- Best and Worst



CCAFS LC-40

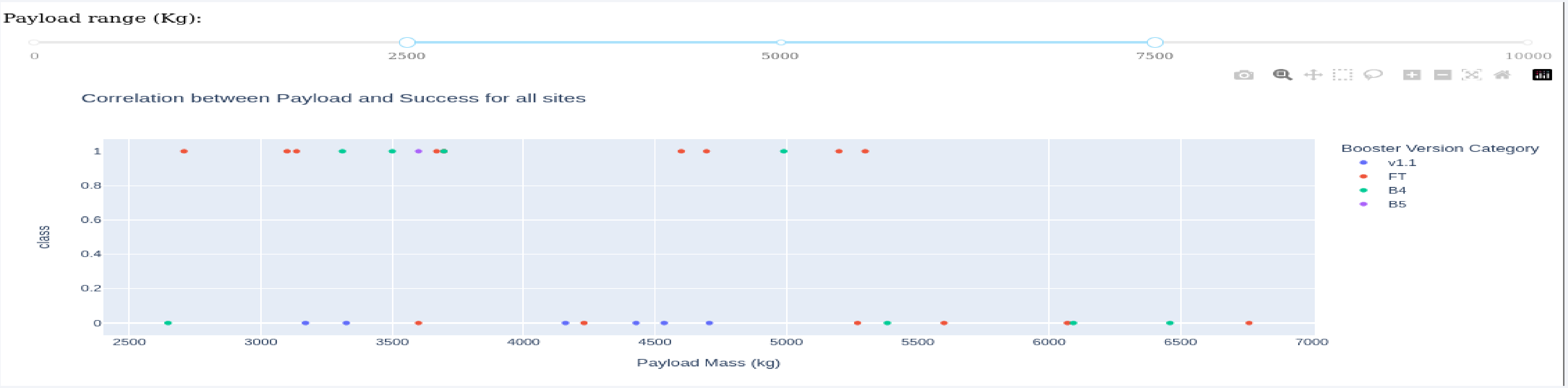


KSC LC-39A

- As noted on the prior slide, Kennedy Space Center has the best launch site for a successful landing, with 10/13, or 76.9% of the launches leading to a successful landing. While the NASA missions in Florida having a high success rate makes sense, more interestingly the worst success rate site is not the west coast.

- CCAFS LC-40 has a lower success rate than VAFB SLC-4E- while it is in Florida, location isn't everything and its 26.9% success rate is substantially worse than the 40% of the west coast site.

Payload and Landing Success



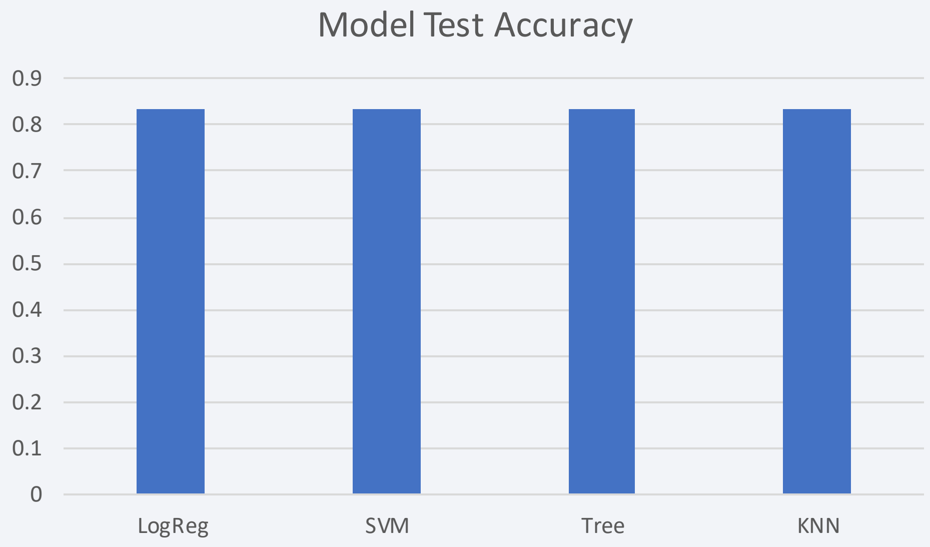
- When excluding the lightest and heaviest payloads from the data range, some rocket versions visibly perform better than others.
- The v1.1 rocket has no successful landings in this range, while the B5 has a single launch, a success. B4 is more mixed, with good reliability in lower payloads but no successes past 5000kg.
- The FT version is more mixed in its results, but generally has higher success rates for lower payloads.

Section 5

Predictive Analysis (Classification)

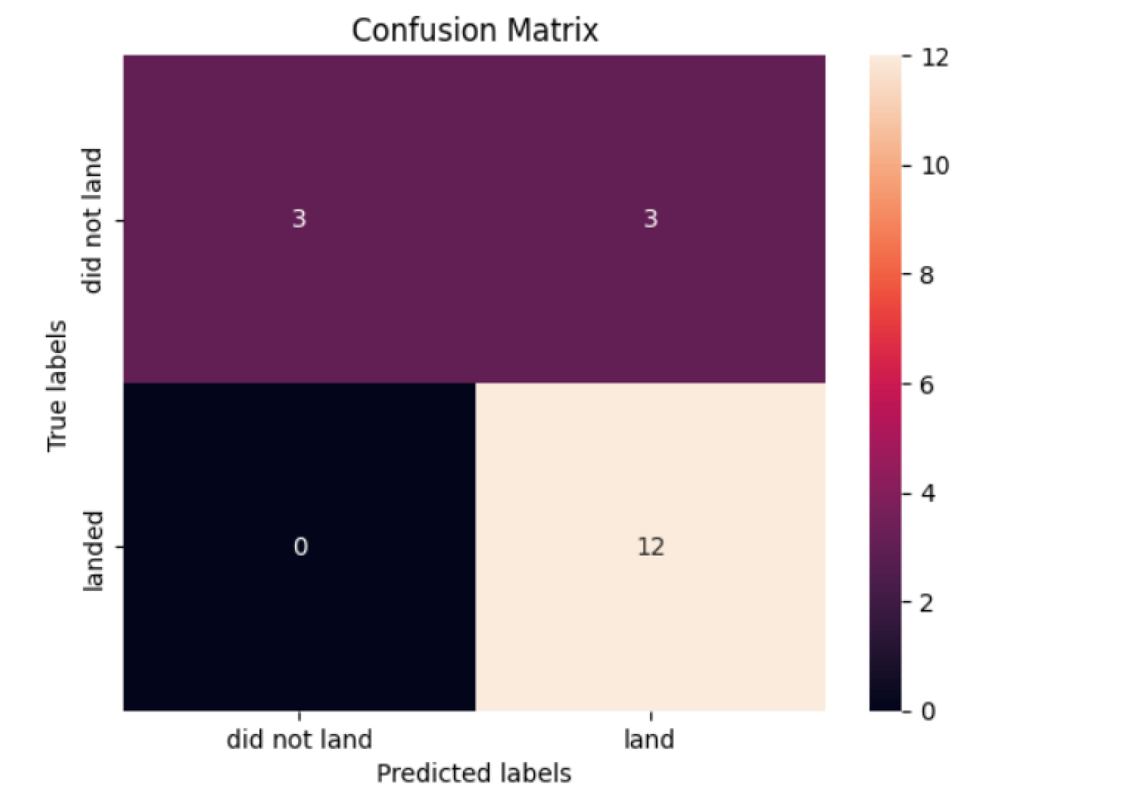
Classification Accuracy

- They all had the same accuracy in my tests. Not sure what went wrong here, but the .score method returned 0.8333 for all models.
- Maybe the tree is better? Higher in-sample accuracy.
- Possibly a bug with the method used, as using .best_estimator_.score(X_test,Y_test) showed Tree as far worse, but the other methods unchanged.



Confusion Matrix

- This is the confusion matrix for the "best-performing model", decision tree. Like accuracy they were all the same. Three failed landings were correctly predicted, and three were falsely predicted to be landings. Of the landing predictions, twelve were true and none were false.



Conclusions

- Expertise in launching rockets is a massive factor in successful landings- practice really does make a difference (from ~30% during early years of attempts to ~80% later).
- Orbit and payload mass also play into success rate- easier to obtain orbits mean smaller rockets, which are generally easier to land. This is carried through with the payload mass, which shows higher success rates with low payloads. The exception to this is low orbits- the two factors can compensate for each other.
- Location is key- good infrastructure and proximity to water makes recovery far simpler than trying to land it in a specific spot on land.
- Choosing the correct model, and not messing up the score with test data, is key to making accurate predictions.

Appendix

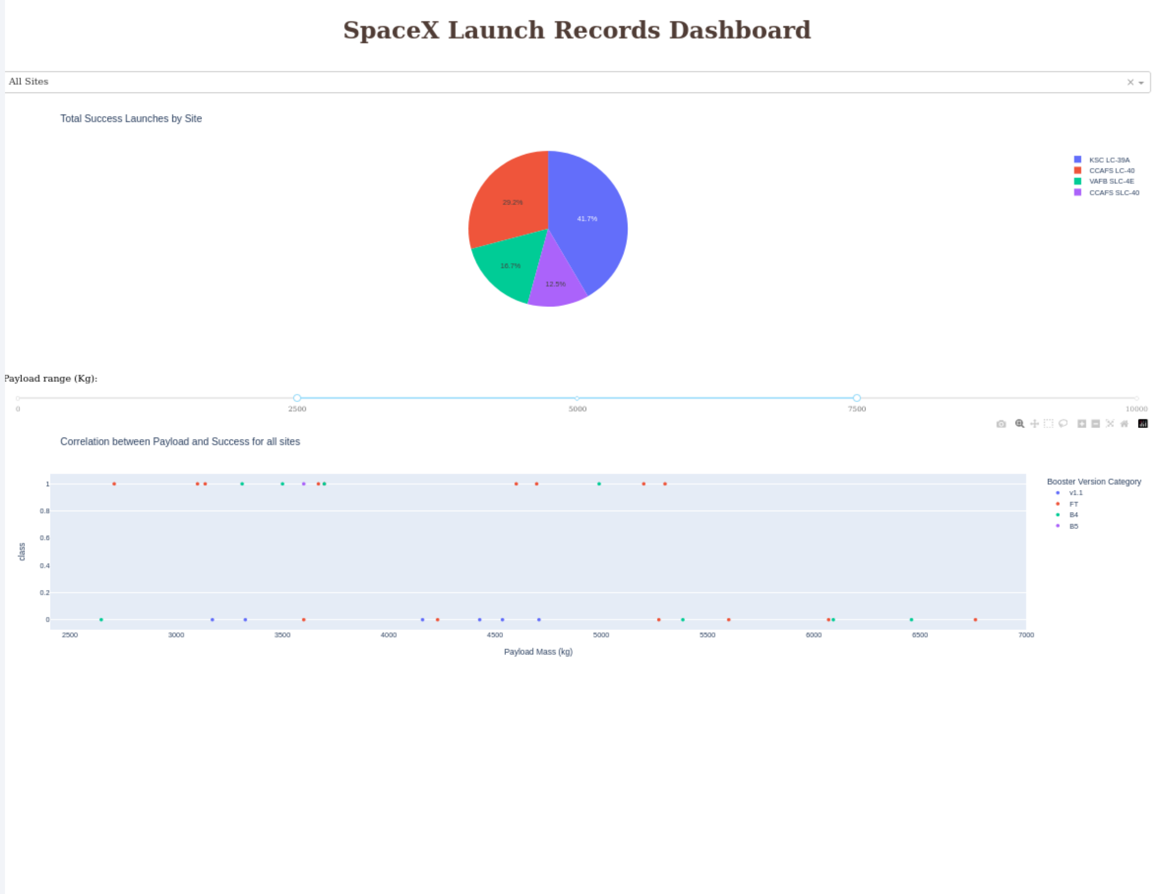
- In case it was missed, the link to the github database for this presentation is below:

https://github.com/kschofield65/Coursera
Capstone

- Below is the link to the data used for the launches, for your reference:

https://github.com/kschofield65/Coursera
Capstone/blob/main/spacex_launch_dash.
csv

- To the right is the whole dashboard:



SpaceX Launch Records Dashboard

All Sites

Thank you!

Thank you for the review! Hope your day has been wonderful!