IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Garrett Isaacs
6/3/24

Outline

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

Executive Summary

- To achieve our results, data collection and data mining were employed for data building. Initial exploratory data analysis was used to try to find trends between booster landing results and collected features. Finally, as a classification problem, regression analysis, clustering, and decision making were utilized.
- With shocking consistency across all testing platforms, we can speak with certainty if the launch of a first stage booster will successfully land, but struggle with predicting if it will fail. Each model scored .833 repeating. An additional attempt with more curated data yielded a similar result. Predictions from these models can be taken in good confidence.

Introduction

- The cost of establishing a space fairing agency is exorbitant. SpaceX has managed to mitigate much of this expense by creating boosters with the capability of safely landing to be used once again on future missions. Its not fool proof, and incidents happen where the booster is lost.
- Knowing that saving the booster is a huge step in the direction of cost reduction for space travel, can we construct a model that will predict if a launched booster will successfully land?

Section 1

Methodology

Methodology

Executive Summary

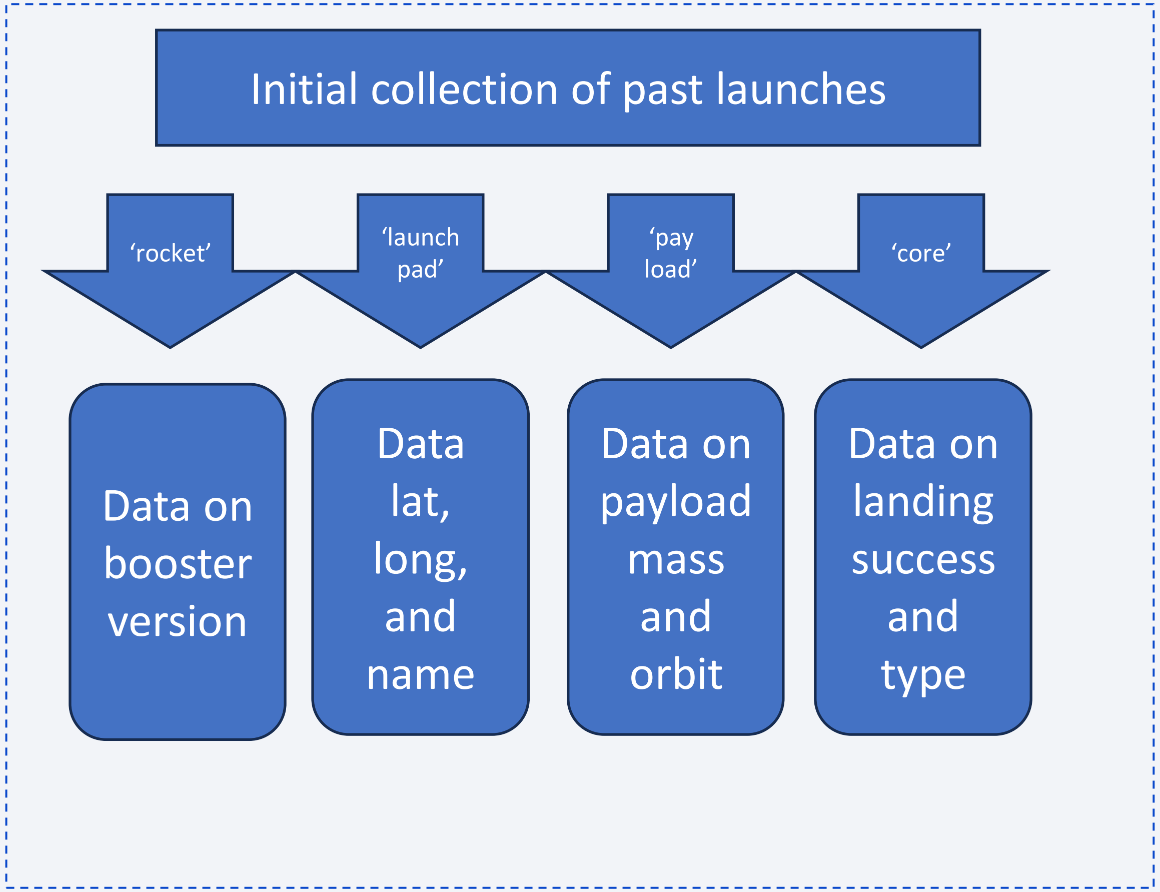
- Data collection methodology:
  - SpaceX records data of their launches for public use via their API’s, as well as publicly curated data on Wikipedia that was scrapable.
- Perform data wrangling
  - With a focus on landing success, information about booster version, launch site, landing type, payload weight, and many other rocket based features were organized.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Categorical classification models were employed and tested to find a model providing optimal results.

Data Collection

- The data was gathered from two publicly available sources.
- The primary source was SpaceX themselves.
  - Their data is stored and accessible across multiple API sources.
  - For making use of their data, it was necessary to make recurrent calls to each of their API's to build a full picture of the data set we required
- The secondary source was scraping charts on Wikipedia
  - Their charts are convenient for data collection, as they have all the data points we are looking for.

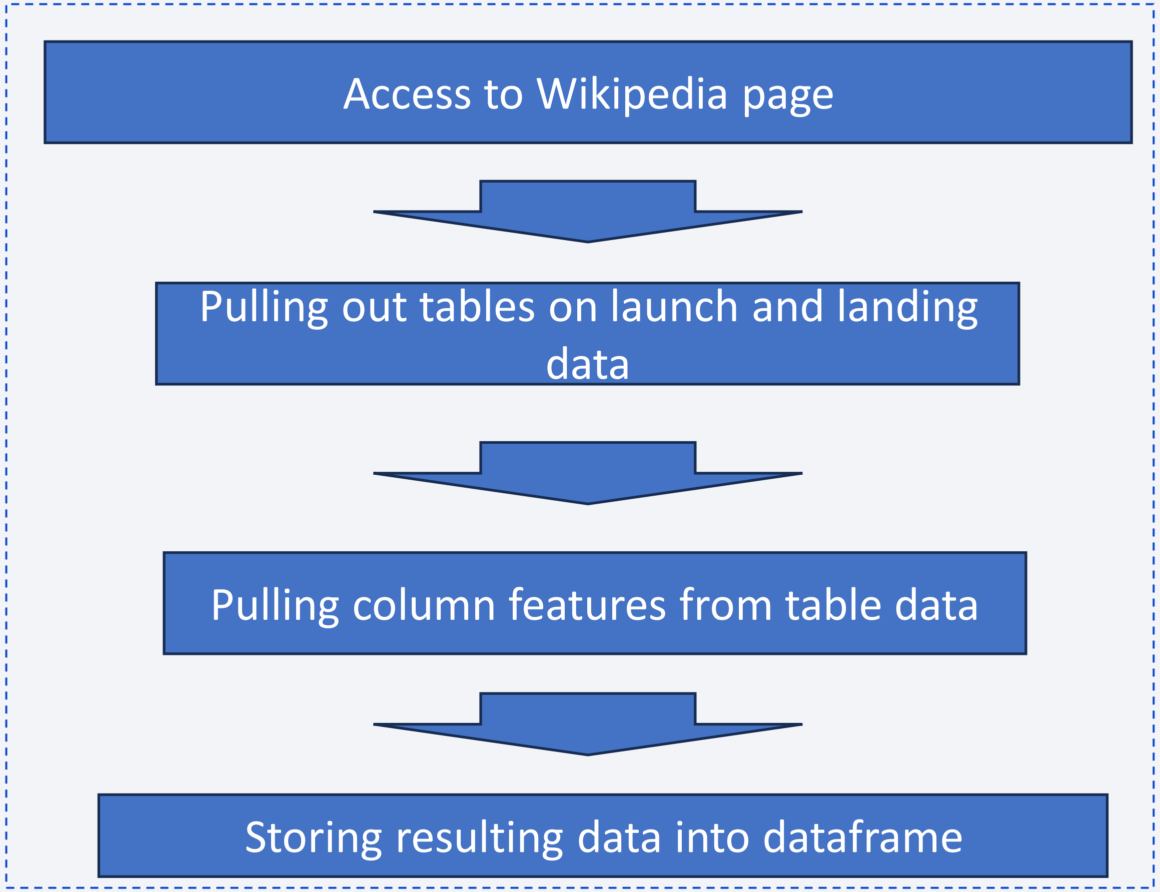
Data Collection – SpaceX API

- The initial data call was to /v4/launches/past. This resulted in primarily useful data such as rocket ID, mission success, date time, but not everything we needed.
  - v4/rockets/ Gave us access to information on the boosters
  - v4/launchpads/ took the launchpad id to grant us coordinates and site name
  - v4/payloads/ let us know the weight and intended trajectory of each payload
  - v4/cores/ gave use core serial numbers, reuse count, block data, the actual landing success and landing tyle, and info about the landing like legs, pad, and fins.
- With each successive API call we were able to build a more complete picture around the landing of the booster.
- Here is the notebook where the API data was collected.


Initial collection of past launches
'rocket'
'launch pad'
'pay load'
'core'
Data on booster version
Data lat, long, and name
Data on payload mass and orbit
Data on landing success and type

Data Collection - Scraping

- Much more straight forward, Wikipedia has their data in accessible tables we can make use of with web scrapping
  - Data including launch count, date, booster version, payload mass, launch orbit, and booster landing outcome.
  - Compared to the API approach, all the data we were looking for was this time in one place.
- Here is the notebook where web-scrapping was used for data collection.



Access to Wikipedia page

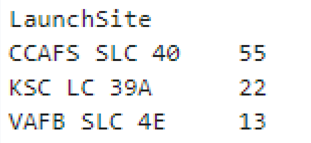
Pulling out tables on launch and landing data

Pulling column features from table data

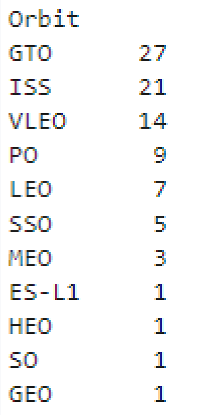
Storing resulting data into dataframe

Data Wrangling

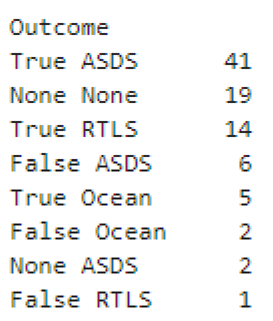
- With the data collected, we can now begin the initial findings step.
  - Firstly the objective marker was reduced from the different types of landing and success rating to just a Boolean of landing successful.
  - We’re able to see how many launches took place at each site
    - A significant majority launched from CCAFS SLC 40
  - What the objective of these launches was
    - Mostly geosynchronous orbit, ISS missions, and very low earth orbit
  - And what the landing occurrence of the booster module was
    - Most commonly a successful drone ship landing.



```
LaunchSite
CCAFS SLC 40      55
KSC LC 39A        22
VAFB SLC 4E       13
```



```
Orbit
GTO        27
ISS        21
VLEO       14
PO          9
LEO         7
SSO         5
MEO         3
ES-L1       1
HEO         1
SO          1
GEO         1
```



```
Outcome
True ASDS       41
None None       19
True RTLS       14
False ASDS       6
True Ocean       5
False Ocean      2
None ASDS        2
False RTLS       1
```

- From here we take the landing outcomes and simplify it into the categorical objective for our study: if the landing was successful or not.

- Here is the notebook where the initial data wrangling took place.

EDA with Data Visualization

- For initial exploratory analysis, scatter plots and catter plots were used to try to visually identify trends and patterns in the data.
  - A catter plot is a scatter plot with one or more axis being a categorical value. Dots have some shake to make it more visually clear the volume of a category
  - This was preformed on Flight Number vs Payload Mass, Flight Number vs Launch Site, Payload Mass vs Launch Site, Flight Number vs Orbit and Payload Mass vs Orbit. In each chart, dots were color coded to identify the success of failure of the booster landing.
  - In addition, bar and line chats were employed to visualize the average success of booster landing by Orbit, and the rate of booster landing over the years.
- Here is the notebook where the initial data visualization took place.

EDA with SQL

- SQL queries were used for more quantitative data analysis.
- With this method, we are able to easily compare the landing outcomes from specific sites, what their payload mass was, when the launch was, and their orbit objective
- We are able to deduce which customers are most common and the quantity of their payload launches.
- We can also see which boosters are responsible for what payload masses, and see trends from a booster to launch the heaviest payloads.

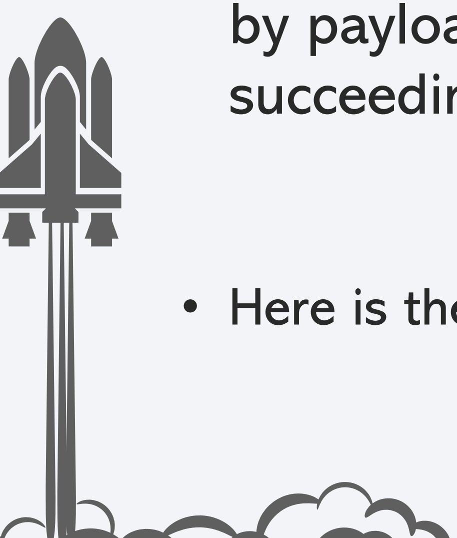
- Here is the notebook where the SQL data exploration took place.

Build an Interactive Map with Folium

- Folium is a tool that allows for the visualization of geographic data. The data we collected has launch site information and coordinates. As such we are able to visualize launch and landing information on the map.
- With Folium, markers were clustered and placed at each launch site. Color of red and green were used to denote if the landing of the booster was a failure or success respectively.
- In addition, we are able to denote geographic distances, and can plot markers to key points, like cities, freeways, and railroads to the launch site.
- Here is the notebook where the geographic data exploration took place.

Build a Dashboard with Plotly Dash

- Creating a Plotly Dashboard allows for the quick change of parameters to get new visualization in data.
- A plotly pie chart was set up to demonstrate the landing success of boosters from each individual site, and with a change show the pie chart of the success rate of booster landings from a particular site.
- A catter plot was set up to demonstrate the success rate of a booster landing by payload mass and booster. This way we can identify trends of booster succeeding to land given their payload mass.



- Here is the script where the dashboard can be run.

Predictive Analysis (Classification)

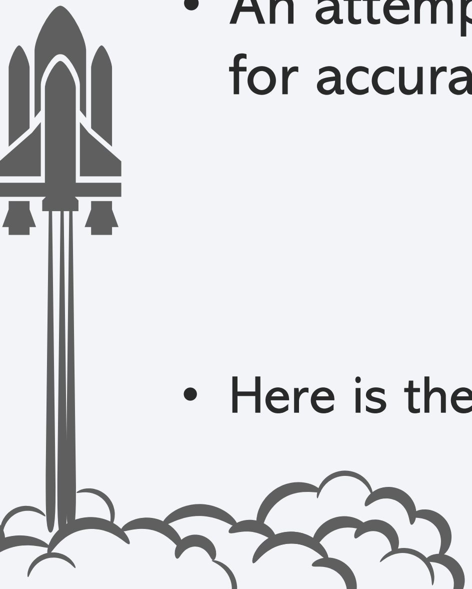
- To generate classification models for the data, pipelines were set up to test varying hyper parameters in an attempt to find the best tuned model per model type.
- Classification models employed were logistic regression, decision tree, support vector machines, and k-nearest neighbors.
- Each model was set in a grid search pipeline with 10 convulsions each along with varying parameter settings to test for best classification.
- In the end, all models preformed the same.

- Here is the notebook where the geographic data exploration took place.

Predictive Analysis (Classification) part 2

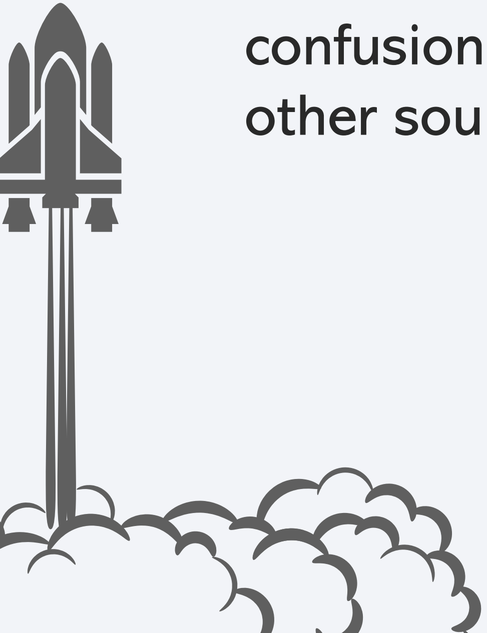
- As an additional attempt at finding better preforming models, half of the successful booster landing results were removed from the data set to train and test on a balanced class.
- This ended up preforming worse, but unlike the previous results the confusion matrices showed different classification rates.
- An attempt to average the result of the models for a final output was tested for accuracy, leading to an accuracy score matching the previous models.

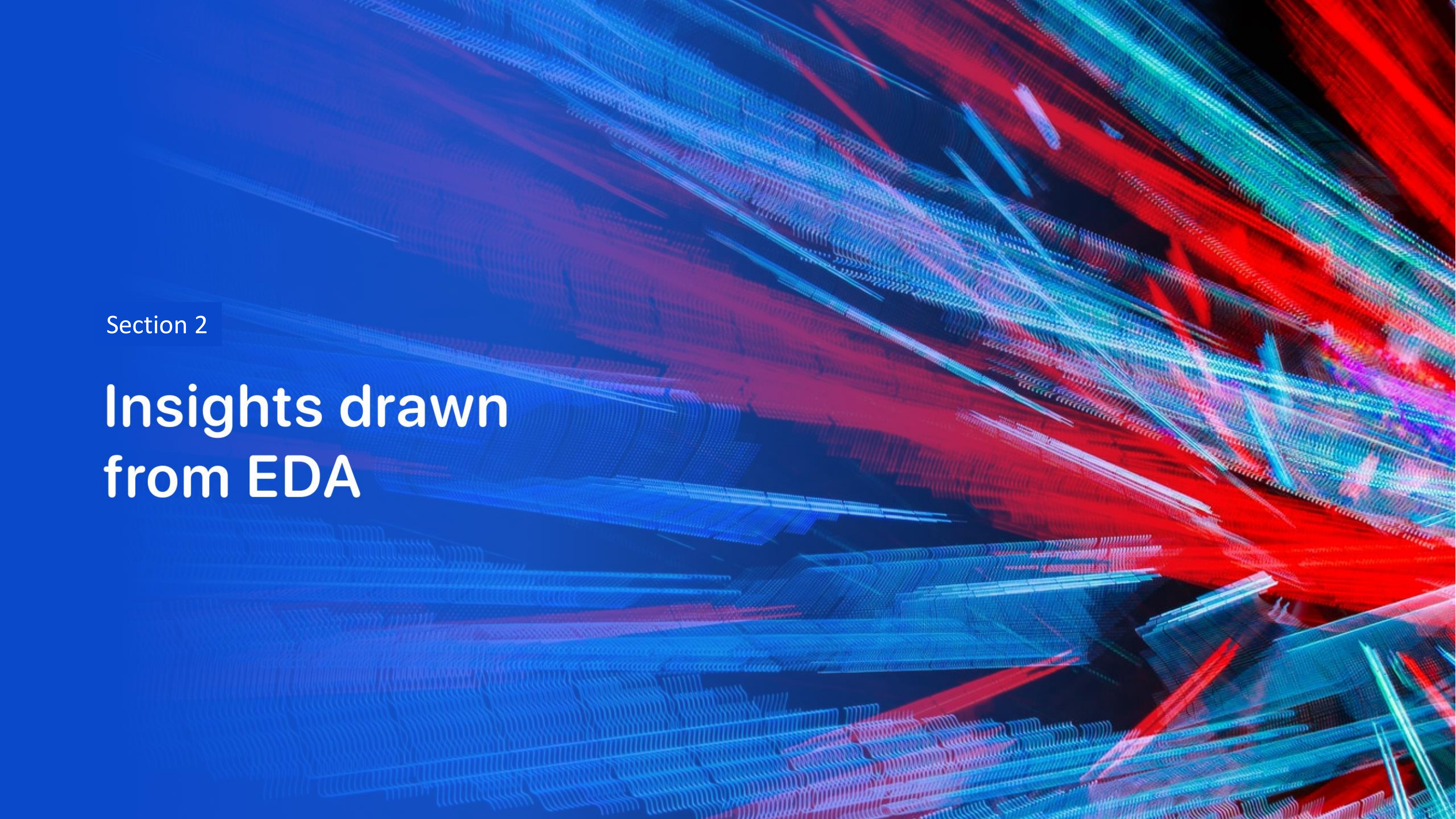
- Here is the [notebook]() where the geographic data exploration took place.



Results

- The initial trend from the visual analysis seemed to place significant weight on payload mass. Sections of success could be drawn in regions of the mass by orbit objective.
- The interactive analysis showed a trend of successful landings from the CC AFS site.
- With the models generated, a consistent score of .833 repeating and matching confusion matrices showed vary stable data, though a need for exploration into other sources if improvement is desired.



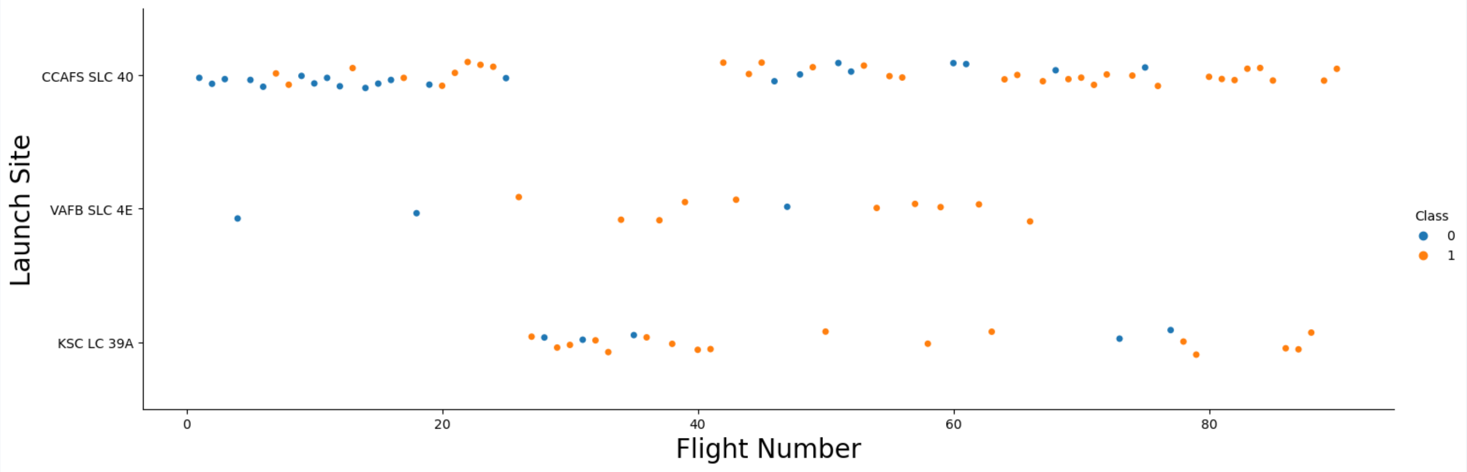


Section 2

Insights drawn from EDA

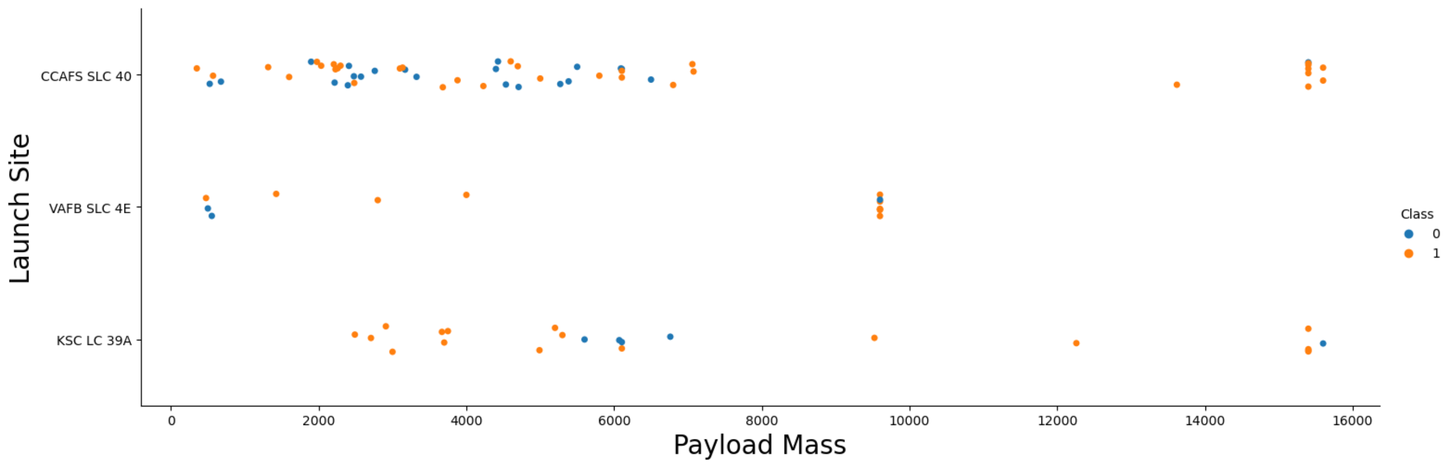
Flight Number vs. Launch Site

- This chart showcases how each launch over time has done from each launch site.
- For launch site, it appears KSC LC 39A has achieved the lowest rate of failure, but the journey definitely began at CCAFS SLC 40 with a bit of a struggle.



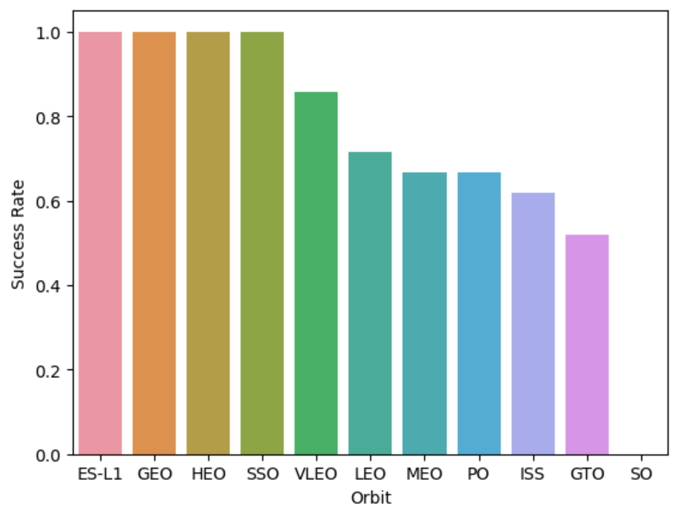
Payload vs. Launch Site

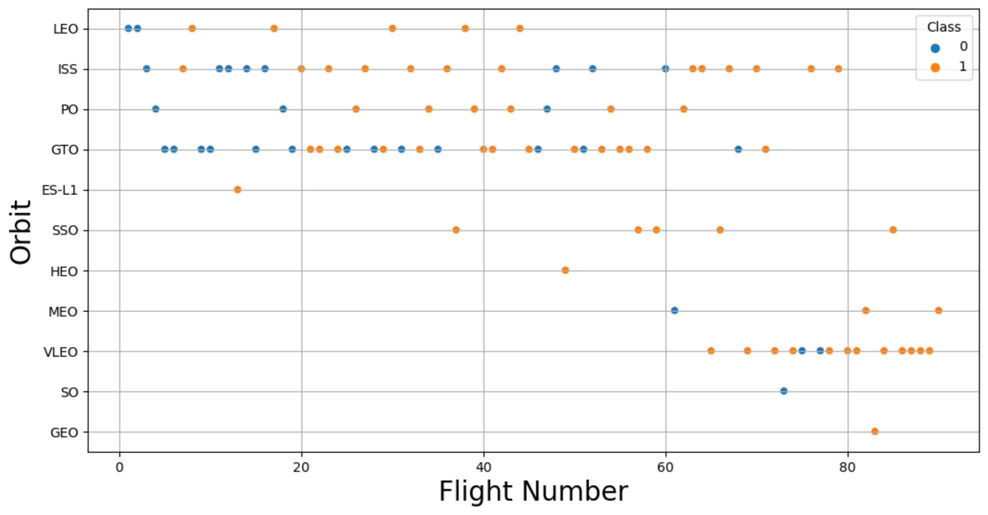
- This chart showcases the success of each launch by payload mass at each launch site.
- VAFB SLC 4E seems to be almost exclusively the site for mid range payload weights, and if it’s a light payload (5500 kg or under); KSC LC 39A shows almost a pure success rate.



Success Rate vs. Orbit Type and Flight Number vs. Orbit Type

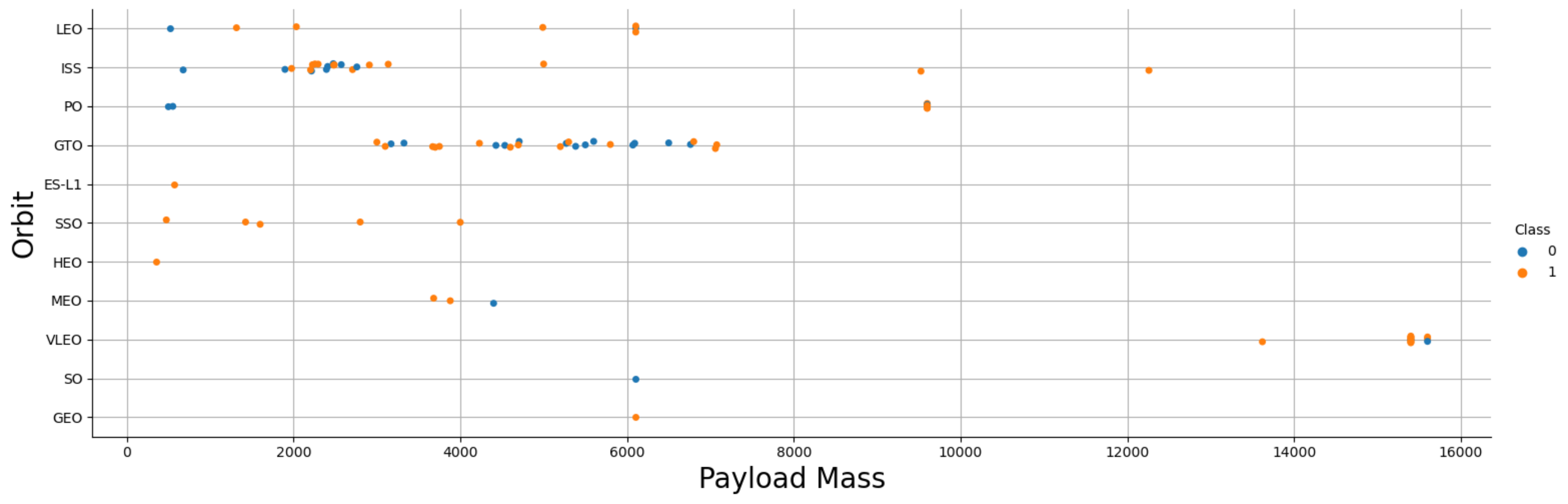
- I have elected to show both charts at the same time as I believe it completes a potentially misleading picture. Data must be viewed from multiple angles. The first chart shows the overall success rate of launches given their orbit objective, and the scatter plot shows the success of each orbit objective launch over time.
- We can see HEO, GEO and ES-L1 all have a guaranteed success rate until you notice only 1 mission has been launched for each. Though SSO is impressive on 5 for 5.
- Its also interesting to note one of their most successful types, VLEO, makes up majority of their overall launches after the 60th launch. This to a degree makes some sense as the payload doesn't have as far to travel perhaps.





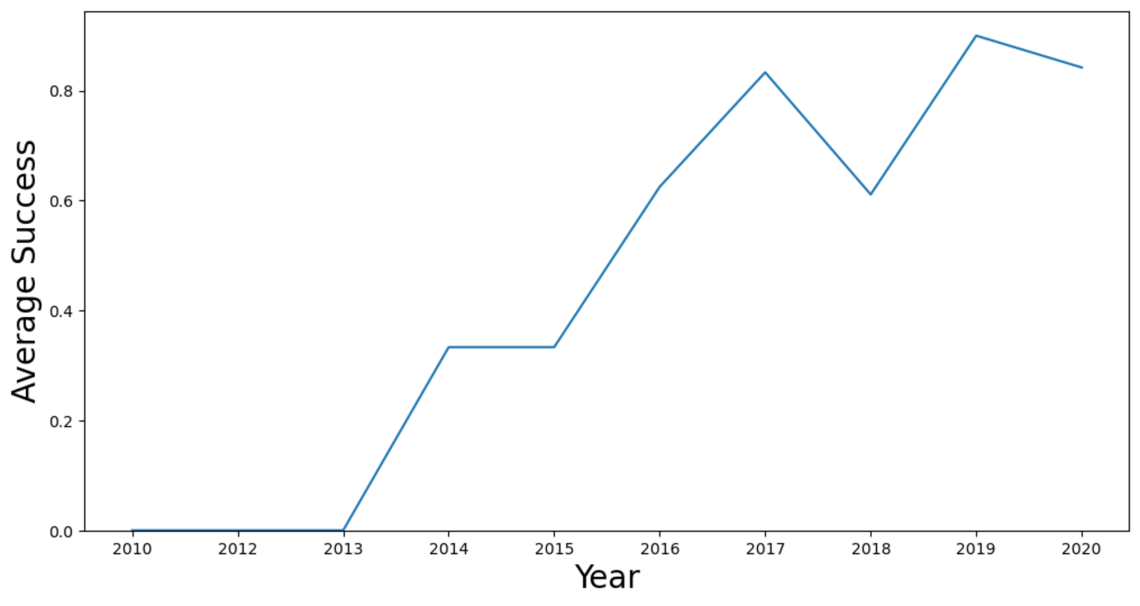
Payload vs. Orbit Type

- This chart showcases the success of launches given their payload mass and what their orbit type objective was.
- Heavier payloads seem to have a better overall success rate, but its interesting to learn almost all the heaviest payloads are VLEO. PO, ISS, and LEO each see significant success for their heaviest payloads.



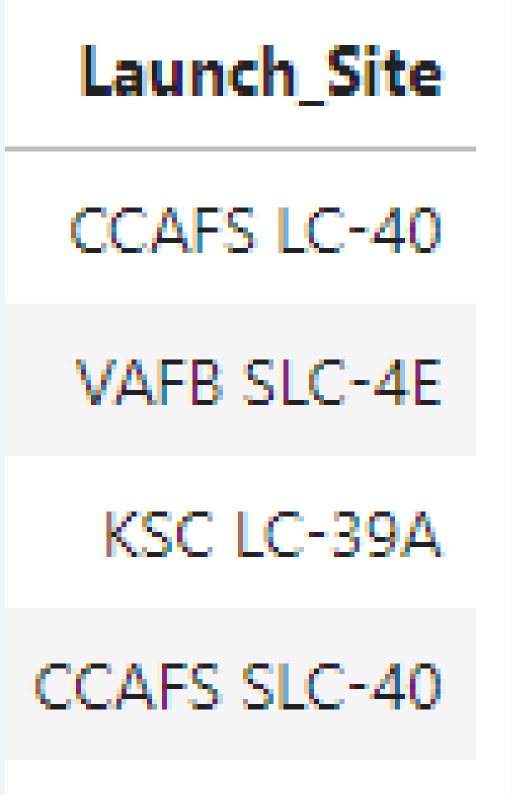
Launch Success Yearly Trend

- This chart showcases the average success of booster landings over the years.
- We can see for the first few years, the booster never successfully landed, but from 2014 and on, landings were finally achieved, spiking in 2017 and recovering in 2019.
- Still, not every booster lands.



All Launch Site Names

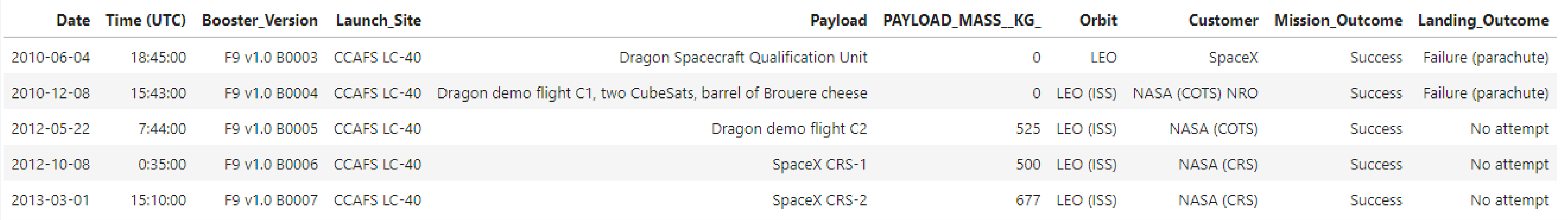
- One important thing to note. Previous graphs showcase 3 launch sites, but in the SQL data and the data moving forward there are 4.
- As will be clear on the map, CCAFS LC-40 and CCAFS SLC-40 are different launching platforms in the same facility. This will increase specificity moving forward.



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

Launch Site Names Begin with 'CCA'

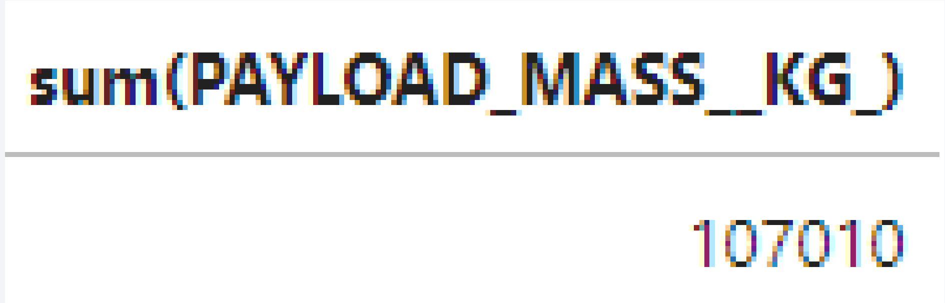
- Here is an example of the data showcasing all features. This data is specifically from CCAFS sites.
- We can see the features date, time, booster version, launch site, payload, payload mass, mission orbit, launch customer, the mission success, and the success or failure of the booster landing.



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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

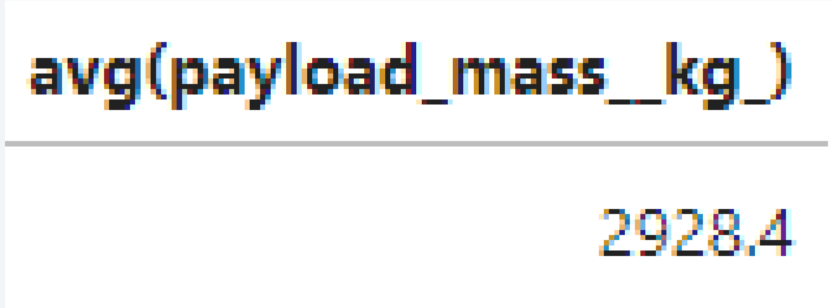
- NASA has been a significant customer. Launching a whopping 107,010 kgs of payload through SpaceX's service.



sum(PAYLOAD_MASS__KG_)
107010

Average Payload Mass by F9 v1.1

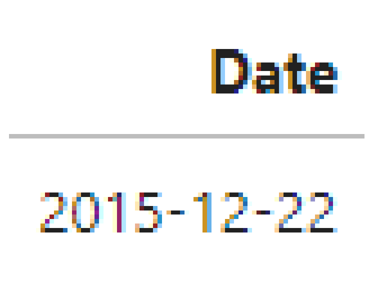
- The Falcon 9 has been SpaceX's more well know rocket. Here we can see that the v1.1 of the Falcon 9 launches on average roughly 2,900 kgs of payload.



avg(payload_mass__kg_)
2928.4

First Successful Ground Landing Date

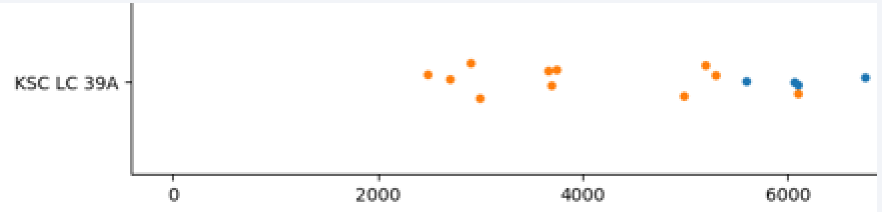
- While famous for their water landings on drone ships, their first successful ground landing of a booster was late December in 2015.

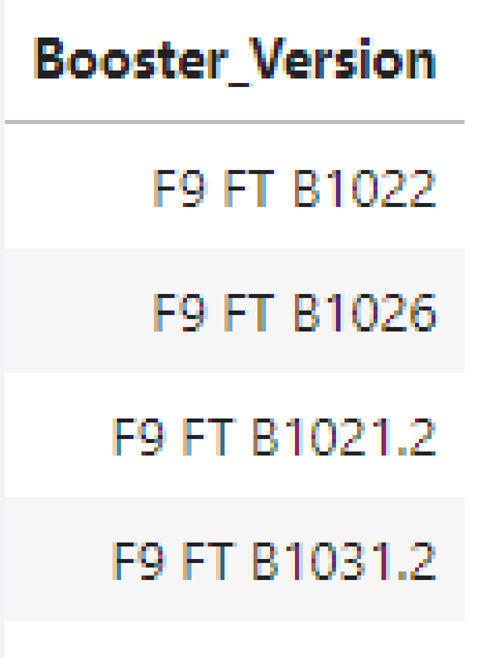


Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- With this query we can see all the boosters that successfully landed on a drone ship after carrying a payload between 4000 and 6000.
- Recall from the Payload vs Launch Site chart previously (and depicted below) that site KSC LC-39A had a high success rate in that range. So there is a possible correlation to explore with these boosters and that site.

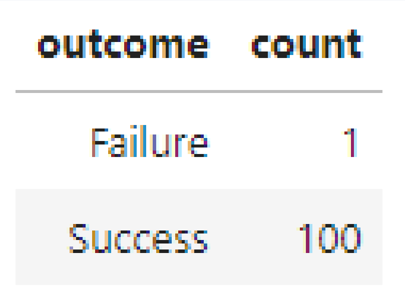




Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

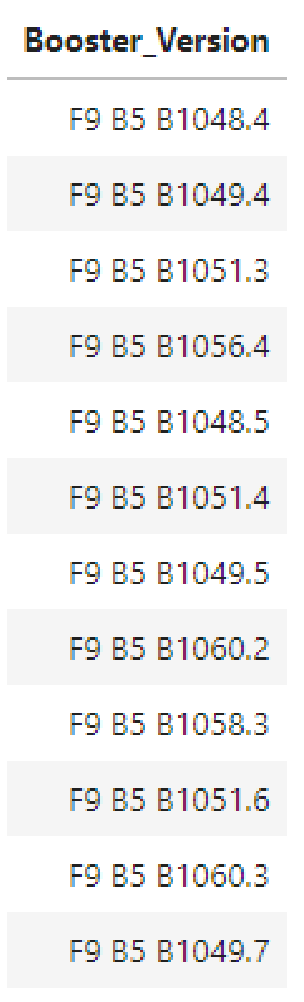
- Though the booster isn’t always recoverable, we are able to see that SpaceX has been very successful in their mission objectives, with only 1 recorded failure as of 2020.



outcome	count
Failure	1
Success	100

Boosters Carried Maximum Payload

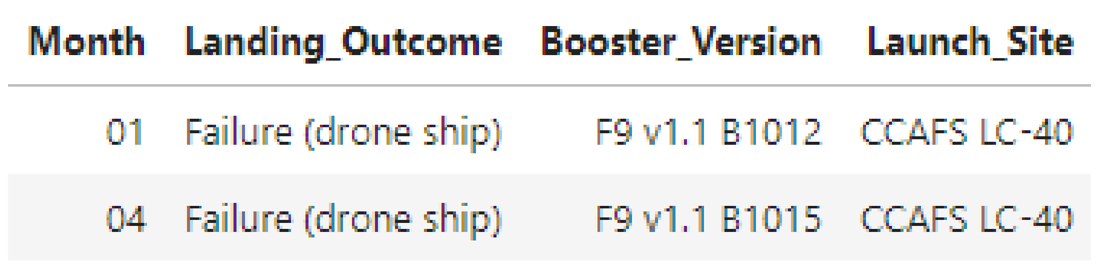
- Boosters carrying the maximum payload all seem to be of the B5 series.
- This makes sense, as up to the end of our data window, the B5 heavy series was their largest rocket in the ‘family’. So it would be responsible for the heaviest payloads.



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

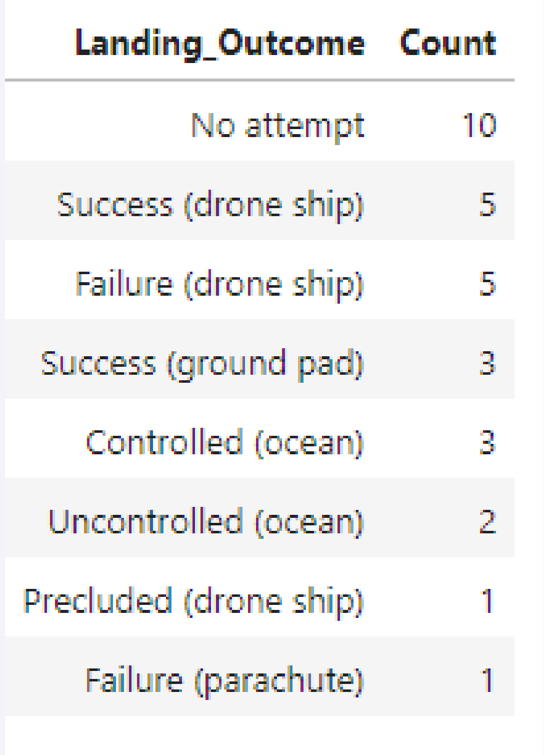
- Since our analysis is on booster landing, we can see how they fair by time. This small chart showcases the failures to land on a drone ship in 2015. It is fortunately a short list, and interestingly all from site CCAFS LC-40. Since this is the southern Florida location, it makes sense.



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

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

- Recall in a previous chart that booster landing success was climbing and peaked in 2017. So we can take a look in that window.
- Initially it makes sense, no attempt to land was made as the technology was still new.
- We can see drone ship is responsible for the next main series of landing attempts, as landing in the water has historically been the preferred and safer method of rocket landing.



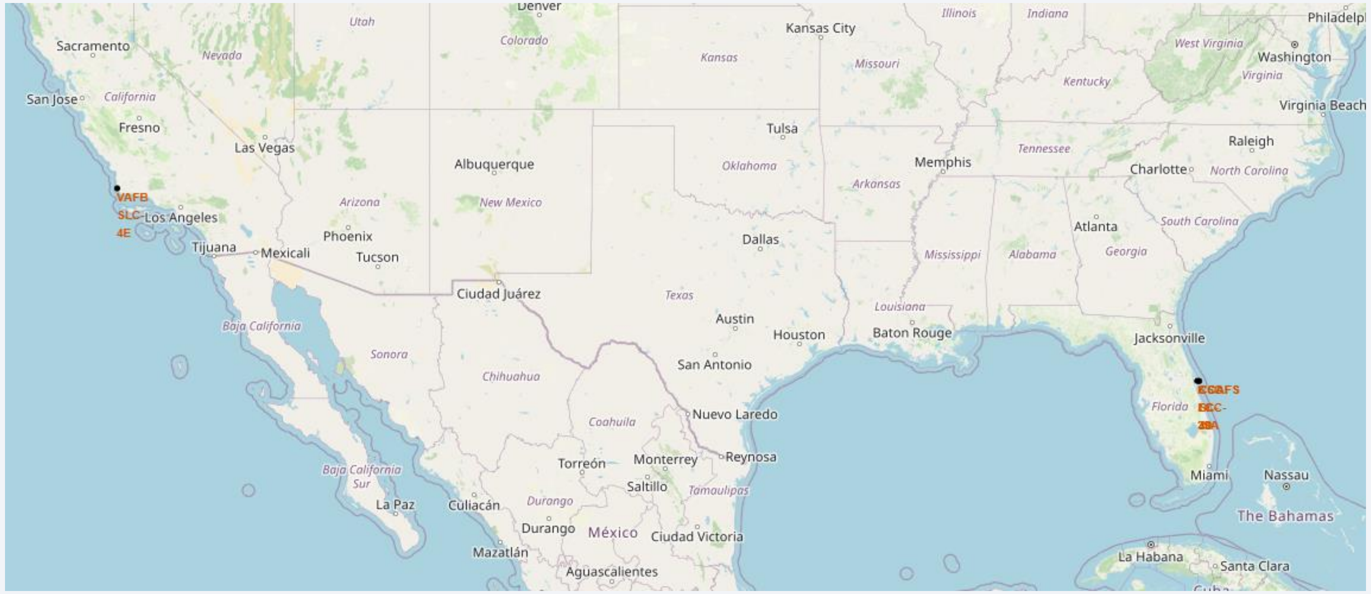
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

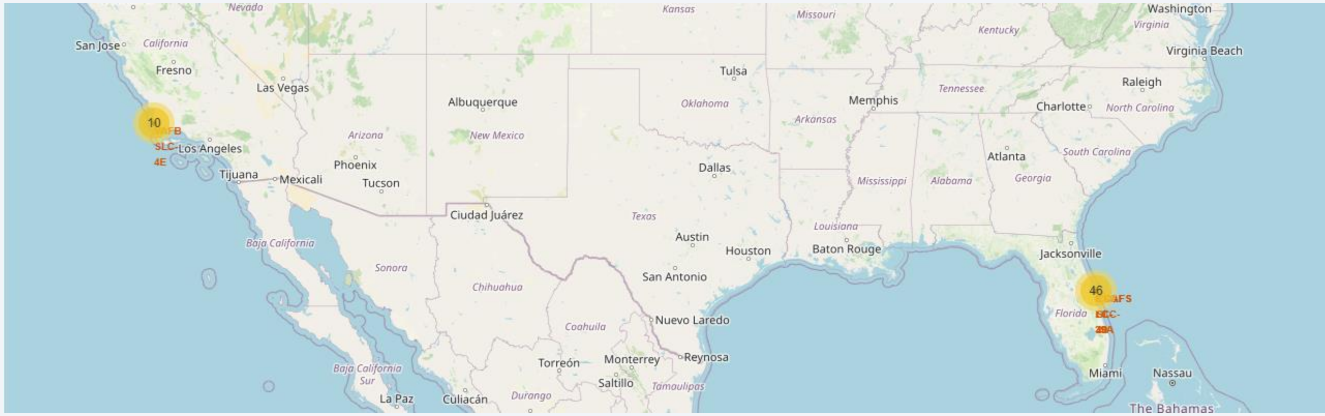
Folium Map1 - Location Sites

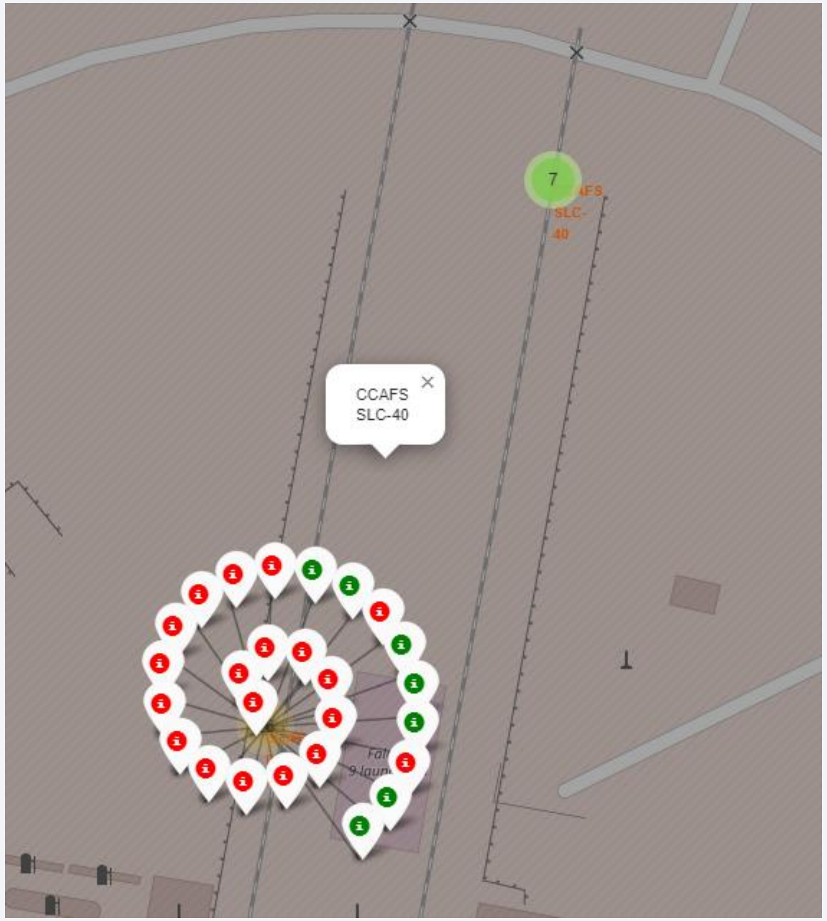
- With Folium, we’re able to make interactive maps showcasing geographical data. Walking through what was achieved, we first plot the launch locations. Of the 4 launch sites, 3 are in Florida, and 1 (VAFB SLC-4E) is located in southern California.



Folium Map 2 – Site Outcomes

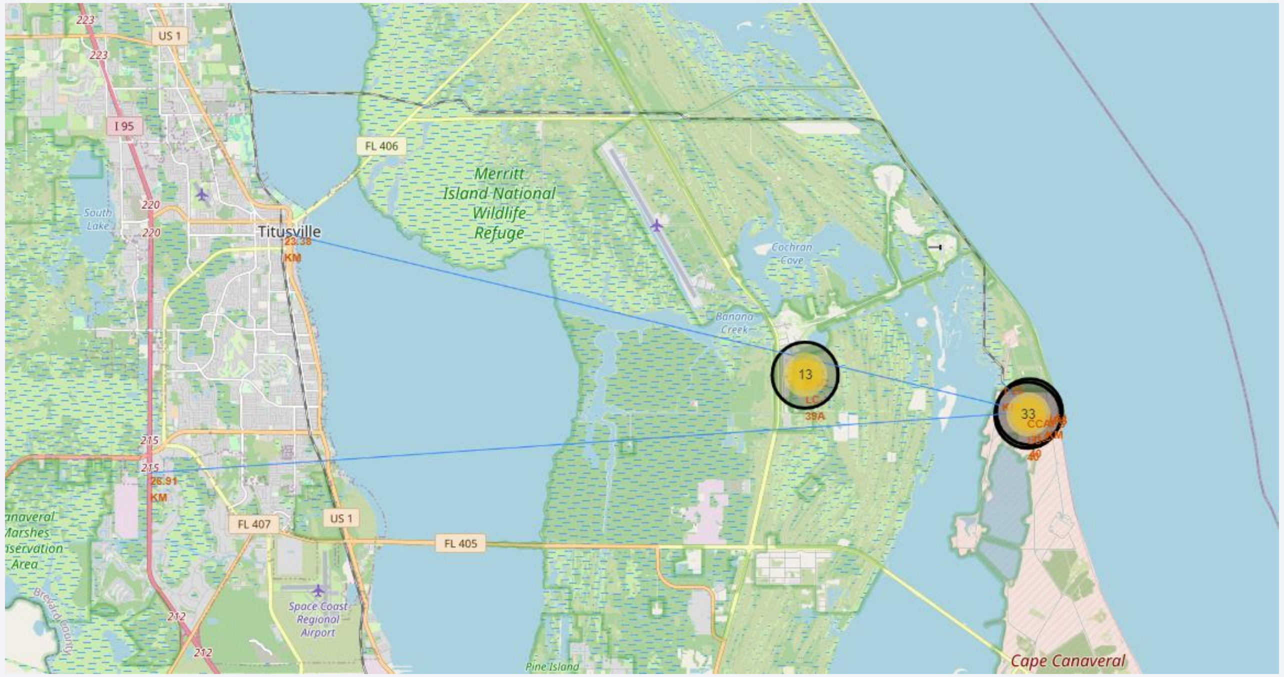
- On top of these sites, we can cluster the launch success and failures.
- Significantly more launches have taken place in Florida.
- We can see the results of each site, like CCAFS SLC-40





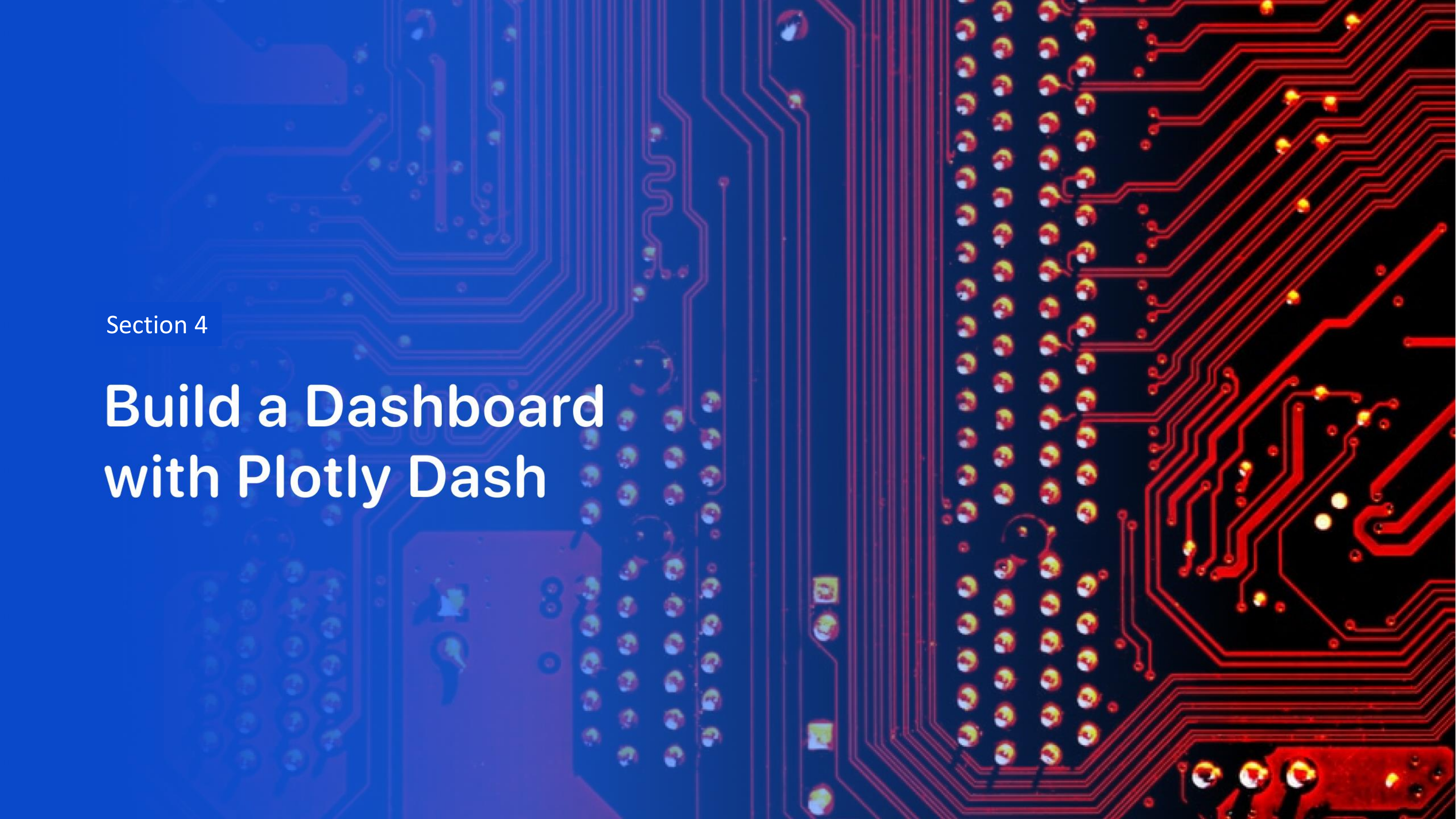
Folium Map 3 – Nearby Sites

- Near these sites is the city of Titusville. 23 km from the launch with the nearest major highway being I95. A railroad track comes directly to the site, presumable for ease of material transport.



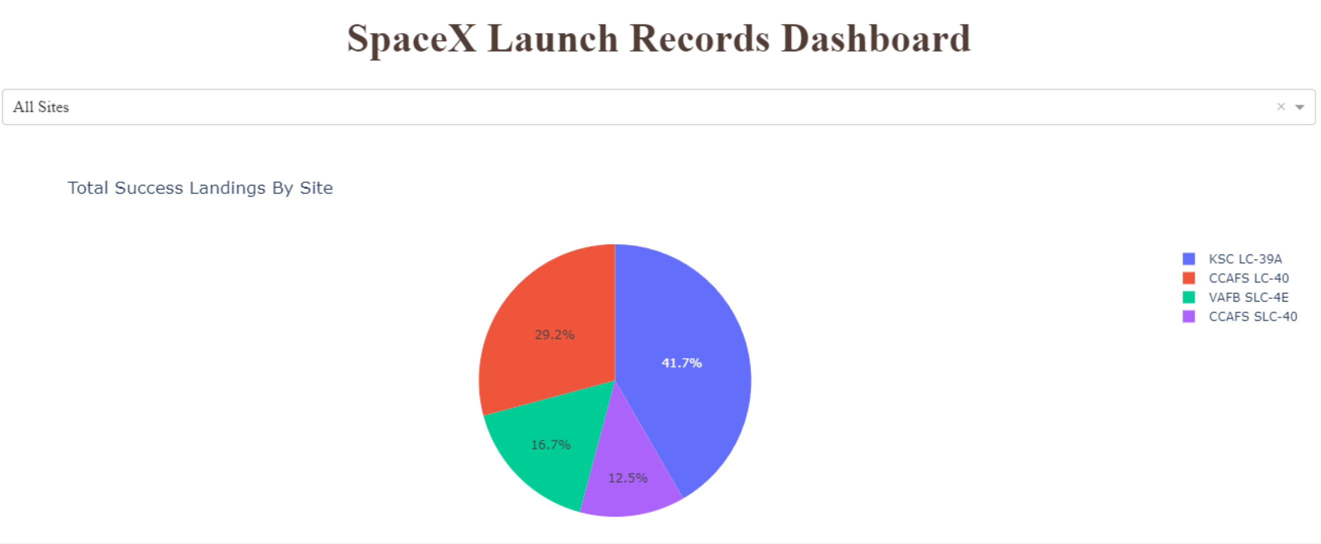
Section 4

Build a Dashboard with Plotly Dash



Dashboard 1-The success of all sites

- Here we are able to see that in terms of total successful launches, KSC LC-39A makes up almost half of all successful launches. However, combining the values of CCAFS brings us just above that result.
- This confirms what we saw on the map where a significant margin of launches were coming from the Florida sites as opposed to the California Site (VAFB)

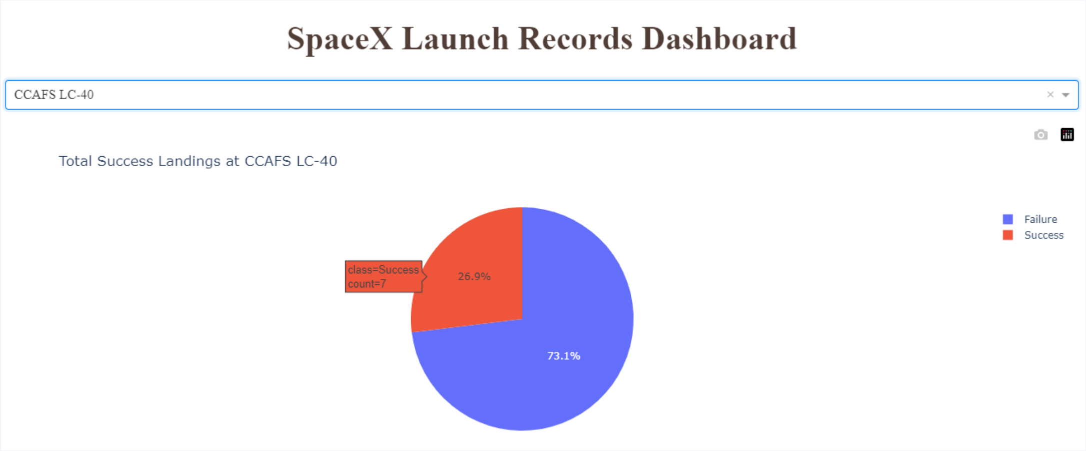


SpaceX Launch Records Dashboard

All Sites

Dashboard 2- Successful Landings by Site

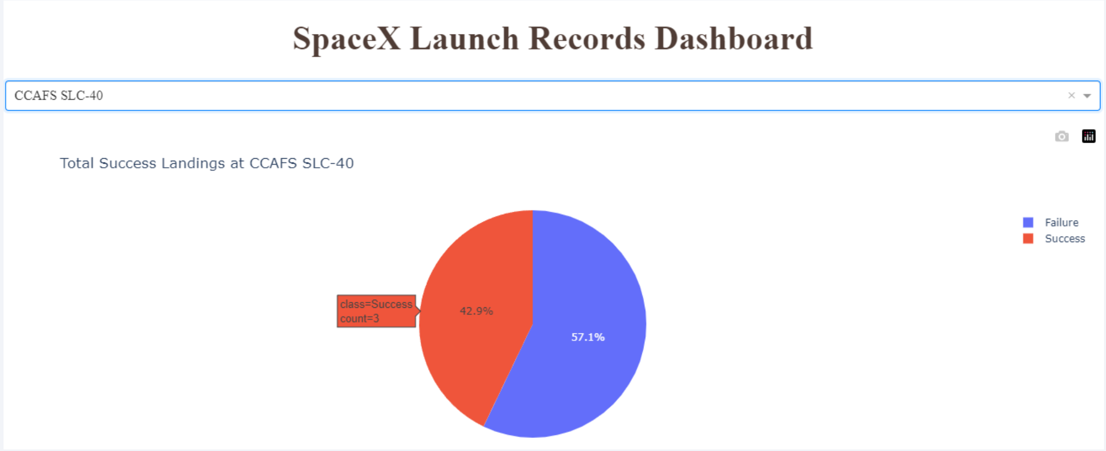
- Here we can see that the different CCAFS sites have achieved a maximum in two separate regards.
- CCAFS LC-40 is responsible for the most booster landings from its launches at 7, but with an overall landing success of 26.9%
- CCAFS SLC-40 is responsible for a rate of landing success of 42.9% but with only 3 landings.



SpaceX Launch Records Dashboard

CCAFS LC-40

Total Success Landings at CCAFS LC-40



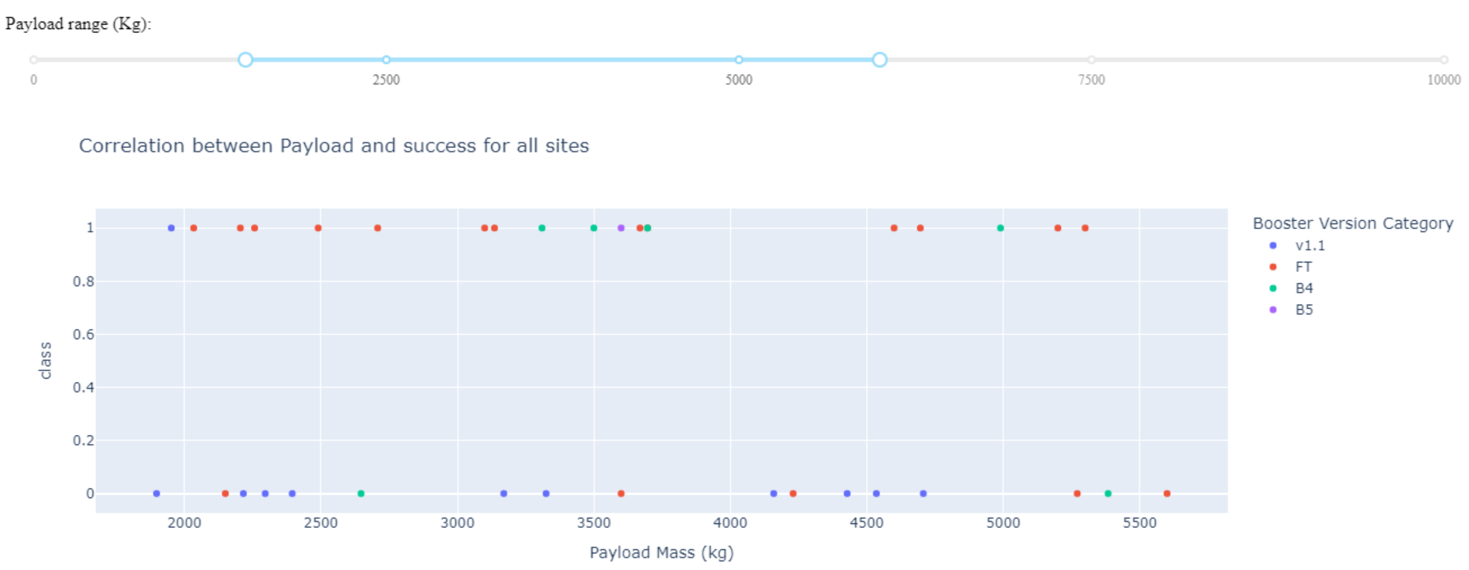
SpaceX Launch Records Dashboard

CCAFS SLC-40

Total Success Landings at CCAFS SLC-40

Dashboard 3- Success of booster landing by payload weight

- We can see the success and failure of launches with a payload in the range of 1500 to 6000 kgs. The booster v1.1 launching these payloads almost never saw a success, however FT and especially B4 and B5 saw majority if not complete booster landing success in this weight range.

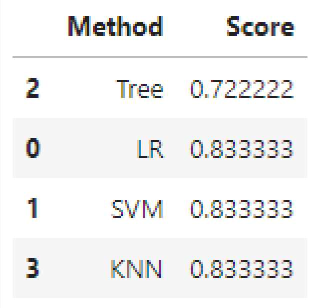


Section 5

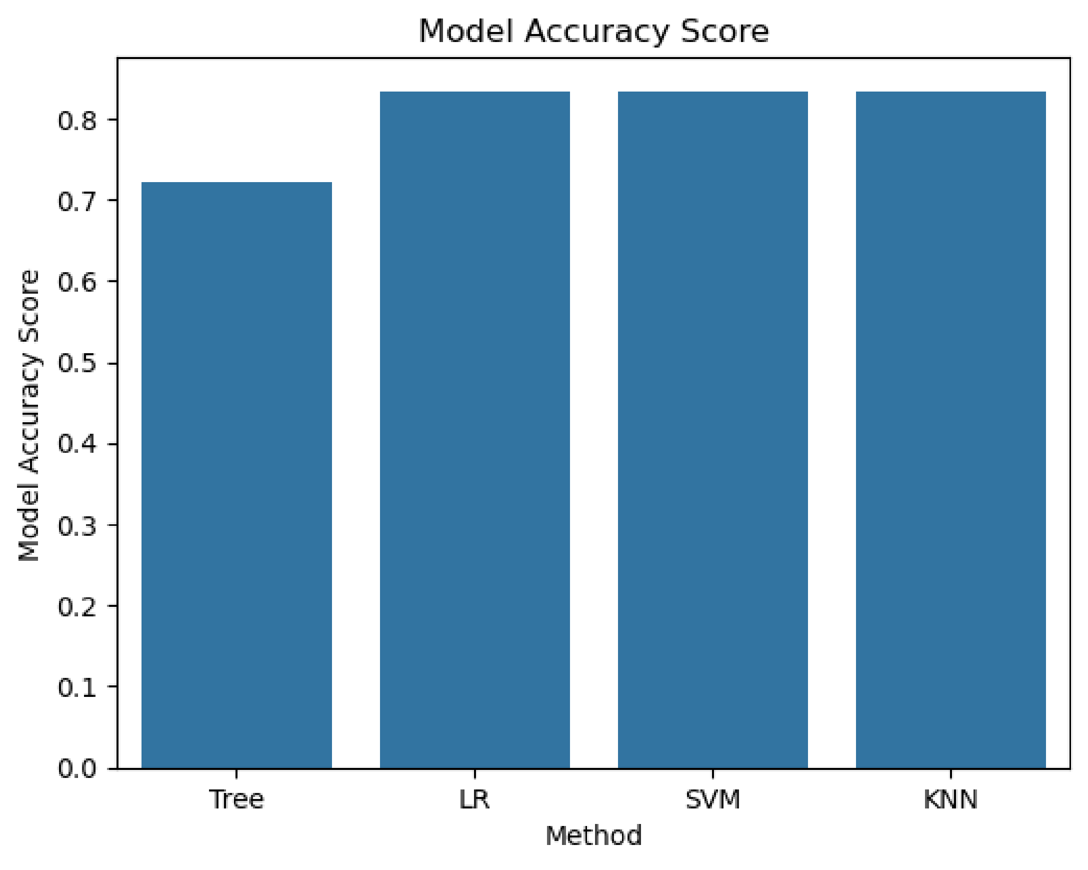
Predictive Analysis (Classification)

Classification Accuracy

- Of the 4 models tested to find optimal hyper parameters, with the exception of Decision Trees, each model scored the same confidence of .83 repeating.
- While this shows consistency, it also speaks to a need to test more approaches.

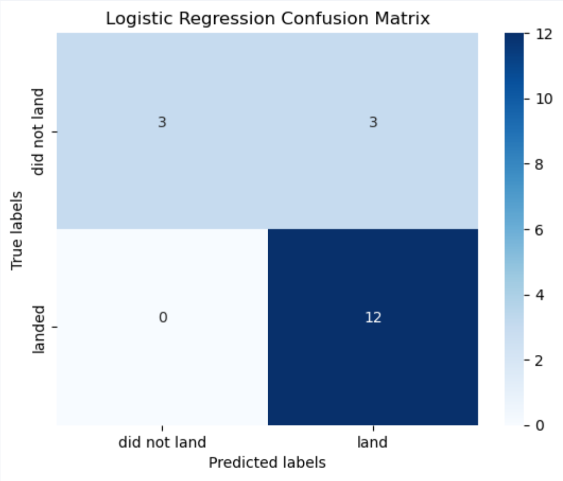


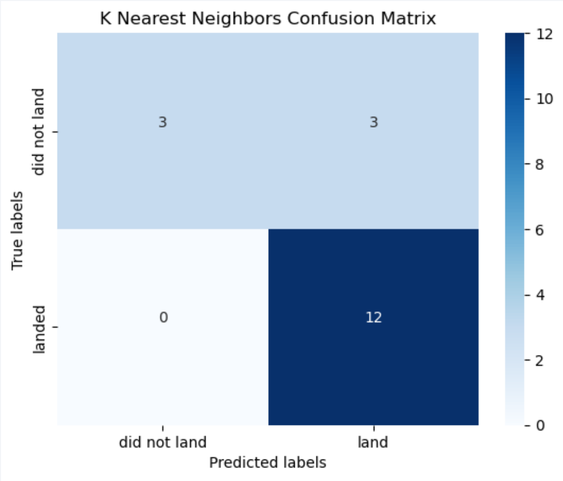
	Method	Score
2	Tree	0.722222
0	LR	0.833333
1	SVM	0.833333
3	KNN	0.833333

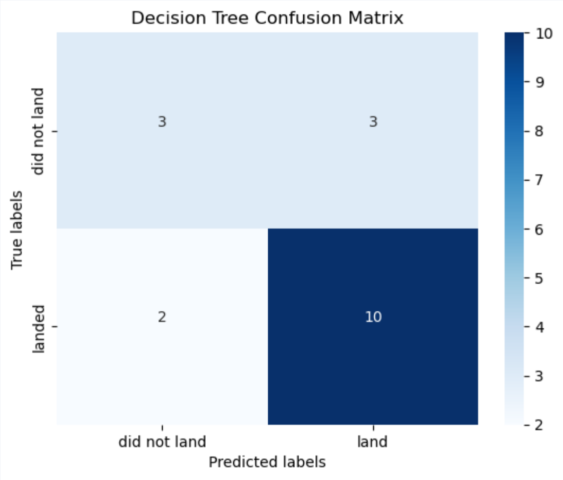


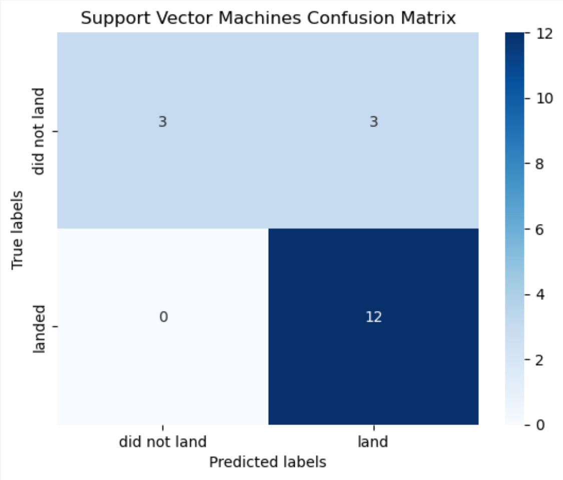
Confusion Matrix

- What is interesting is that each model was incredibly accurate in its classification of a landing. The error comes primarily from false positives on did not land.
- What this could speak to is a need for more cases where the booster failed to land. Our test case had 30 failed counts and 60 success counts.


Logistic Regression Confusion Matrix

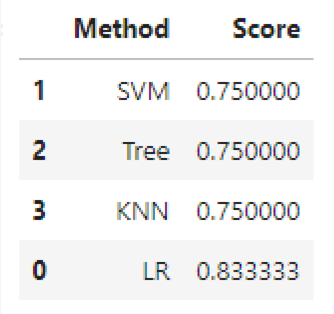

K Nearest Neighbors Confusion Matrix


Decision Tree Confusion Matrix

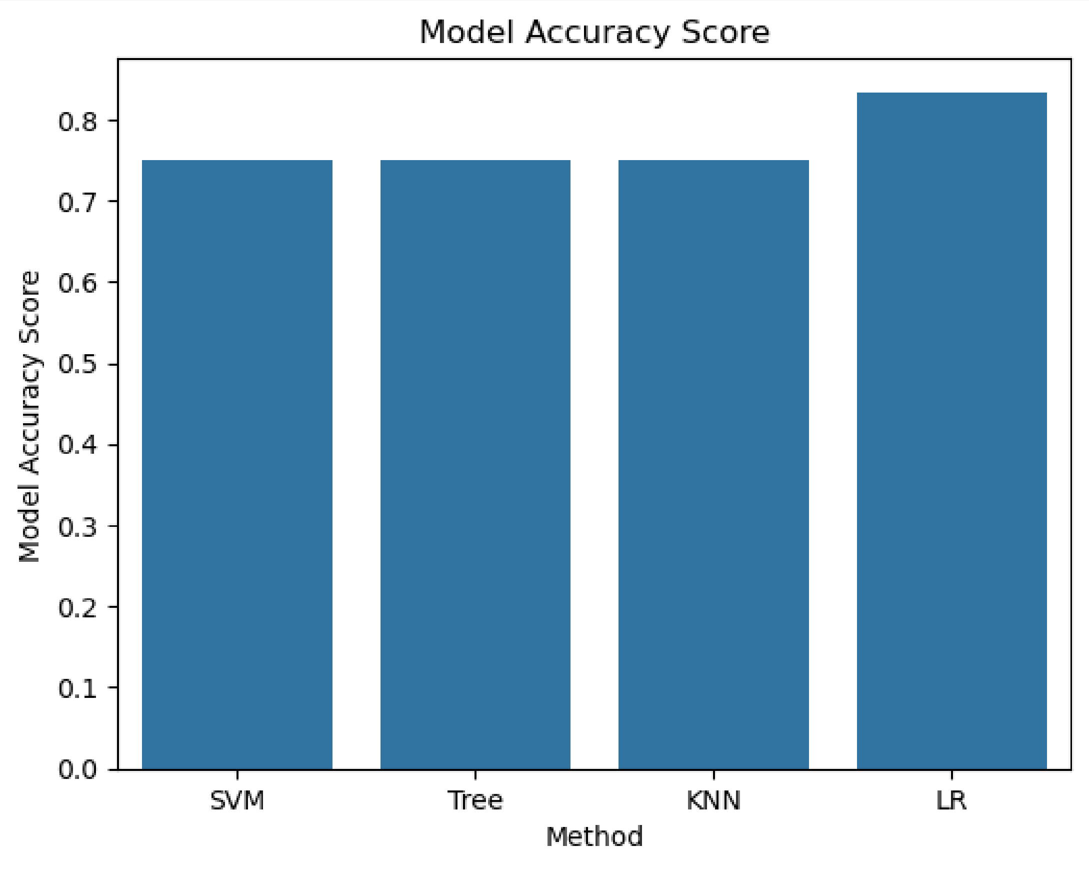

Support Vector Machines Confusion Matrix

Classification Accuracy with Less Success Instances

- For an additional test based on my own curiosity of having too many success results, I set about removing half of the success values and testing the models again.
- To my surprise, though there was a balance of 30/30 success results to failures, the models preformed worse over all.

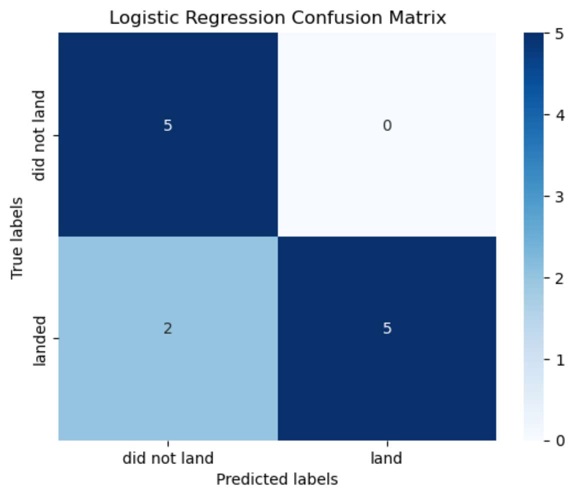


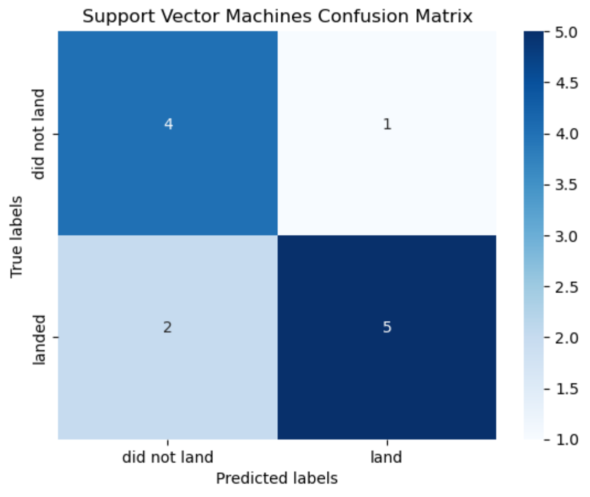
	Method	Score
1	SVM	0.750000
2	Tree	0.750000
3	KNN	0.750000
0	LR	0.833333

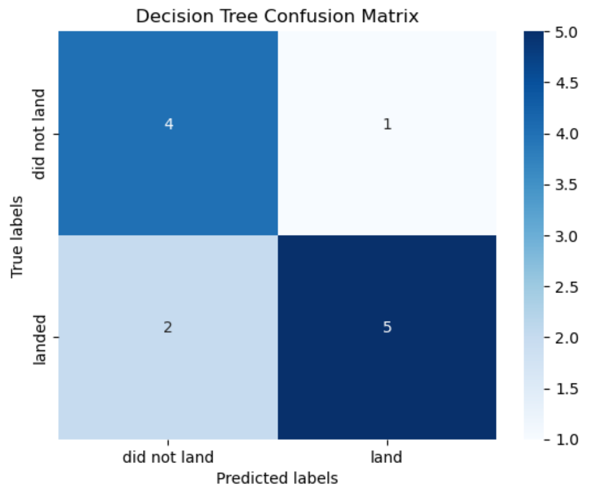


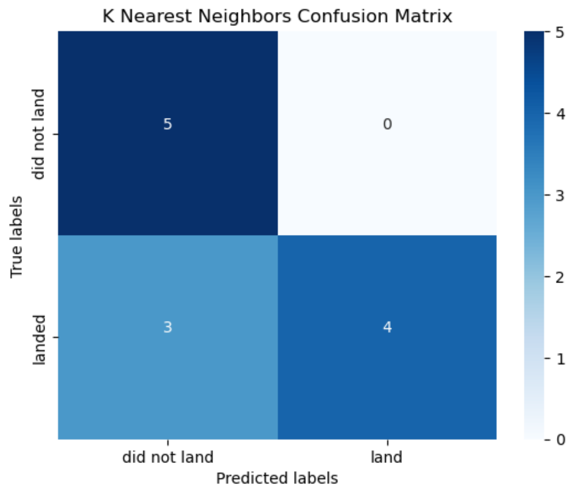
Confusion Matrix

- What is interesting is that each model actually preformed differently in terms of correct classification, unlike the previous with high priority of true positives over true negatives. Here the results seemed rather split.


Logistic Regression Confusion Matrix

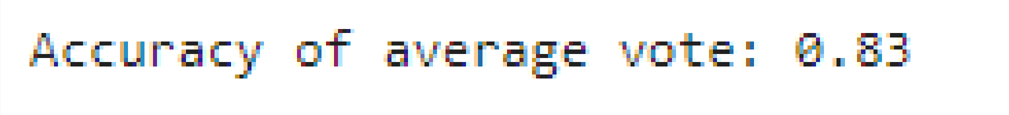

Support Vector Machines Confusion Matrix


Decision Tree Confusion Matrix


K Nearest Neighbors Confusion Matrix

Average Prediction Results

- Knowing that unlike last time, the classification results were not as consistent, I was curious if all 4 models together in vote would be stronger than their individual parts. Unfortunately the classification rate matched the success rate of Logistic Regression.



```
Accuracy of average vote: 0.83
```

Conclusions

- The models have high recall and perfect sensitivity. Each of the three models correctly identified every case of true positive (The booster Landed)
- The models have decent precision. 20% of instances labeled as Landed were actually ‘Did not Land’.
- With these in mind, we can be confident when the model labels a result as ‘Will not land’ as no false Positives occurred in the three preforming models. If the predicted result is labeled as ‘Will land’ then we can be pretty confident in the result.
- More avenues of data should be explored. All that was examined were features in company control. Features such as weather at landing site could yield more insight into why some booster landings, especially more recent ones, failed.
- For the sake of common answers, we also limited ourselves to not using more recent data. Landing results may have improved further in the early 2020’s.

Appendix

- Access to GitHub with all notebooks, images, and scripts.
- Rocket image from Freeimages.com

Thank you!