IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hong Zhao
2024/6/20

Outline

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

Executive Summary

- This analysis collected data from SpaceX API and Wikipedia and transformed into a dataframe. it was later visualized, queried, and classified
    - It follows normal EDA process, used visualization, SQL, and machine learning technology to proceed.
- Summary of all results
    - From the data analysis, we can find that some factors such as payload mass, orbit, and time are all influential to the success rate
    - Through folium map, we can find some common attributes from the launch sites
    - When we are trying to predict the probability of success, all the methods are doing well

Introduction

- Since 2002, SpaceX has become an undeniable force in the global aerospace industry with its innovative spirit and relentless pursuit. However, rocket launches, like other scientific explorations, are full of unknowns and uncertainties.
- The purpose of this study is to review the rocket launch records of Falcon 9 and analyze the key factors that affect the launch results in a data science manner. Hopefully, this analysis can start with data and provide some insights for the subsequent launch of SpaceX.

Section 1

Methodology

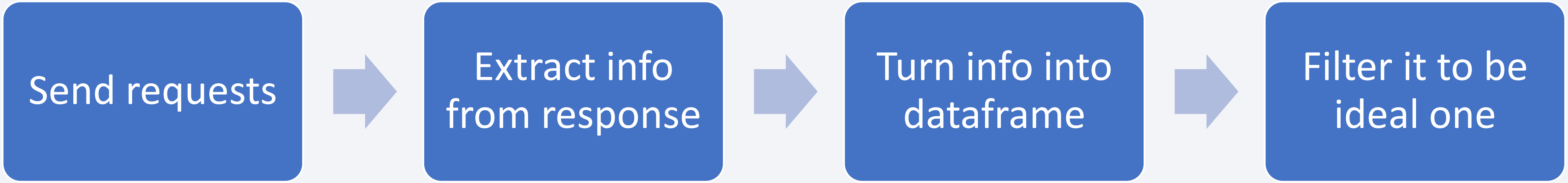
Methodology

Executive Summary

- Data collection methodology:
  - The information is sourced from the SpaceX API and Wikipedia. They were discovered from web pages and parsed from responses in Json and Html formats
- Perform data wrangling
  - Here all information was normalized, checked, and re-organized into a dataframe
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The dataset was separated into train and test set, they were further fitted by different models.

Data Collection

- The datasets were collected from SpaceX API and Wikipedia.
- The collection started with requests and getting response in different format, and the response were all transformed into dataframe which is also filtered later.



Send requests

Extract info from response

Turn info into dataframe

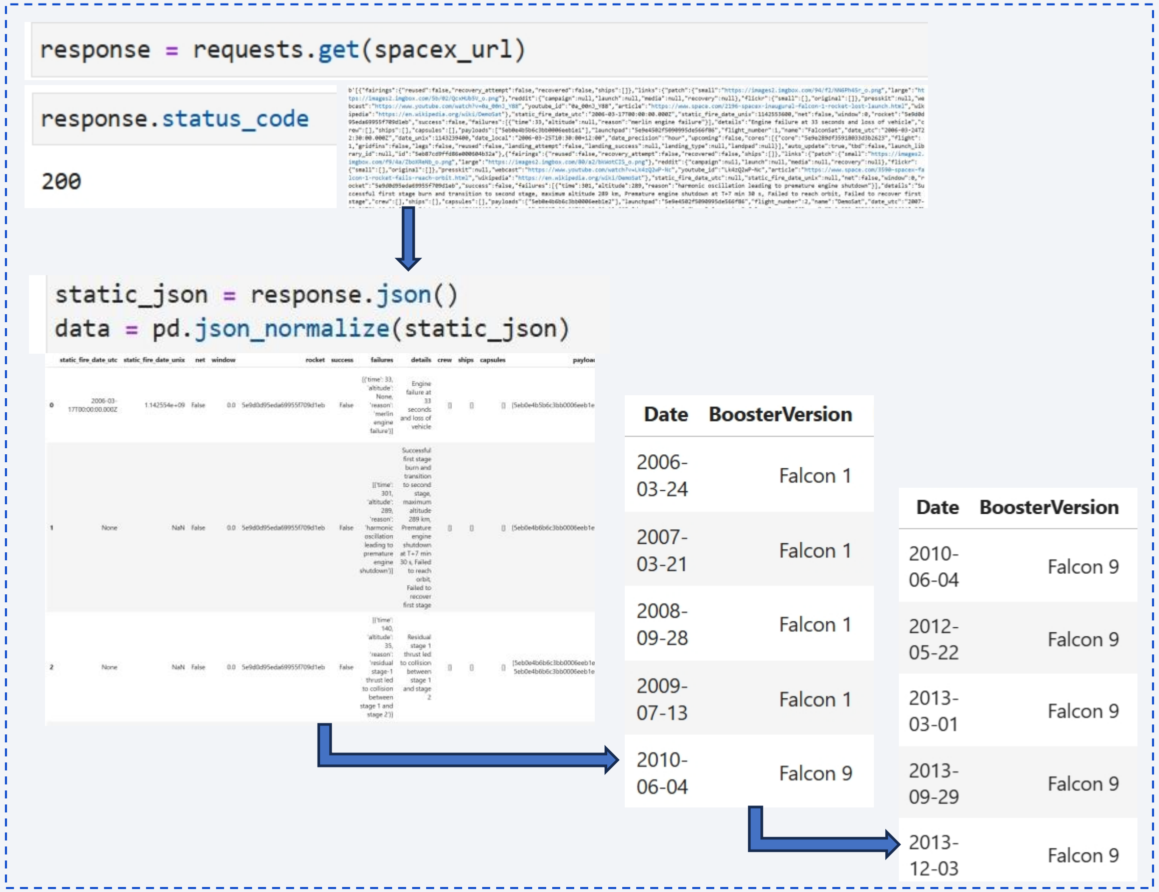
Filter it to be ideal one

Data Collection – SpaceX API

In this part:

- Requests of rocket launch data had been sent to SpaceX API with URL
- the Json result had been normalized and turned it into a dataframe
- The original dataframe was then filtered with a defined dictionary and further filtered to be a falcon 9 only dataframe

- For detail of the workflow, you may check the notebook here.



```
response = requests.get(spacex_url)
```

```
response.status_code
```

```
200
```

b'[{"fairings":{"reused":false,"recovery_attempt":false,"recovered":false,"ships":[]},"links":{"patch":{"small":"https://images2.imgbox.com/94/f2/NN6Ph45r_o.png","large":"https://images2.imgbox.com/5b/02/QcxHUb5V_o.png"},"reddit":{"campaign":null,"launch":null,"media":null,"recovery":null},"flickr":{"small":[],"original":[]},"presskit":null,"webcast":"https://www.youtube.com/watch?v=0a_00nJ_Y88","youtube_id":"0a_00nJ_Y88","article":"https://www.space.com/2196-spacex-inaugural-falcon-1-rocket-lost-launch.html","wikipedia":"https://en.wikipedia.org/wiki/DemoSat"},"static_fire_date_utc":"2006-03-17T00:00:00.000Z","static_fire_date_unix":1142553600,"net":false,"window":0,"rocket":"5e9d0d95eda69955f709d1eb","success":false,"failures":[{"time":33,"altitude":null,"reason":"merlin engine failure"}],"details":"Engine failure at 33 seconds and loss of vehicle","crew":[],"ships":[],"capsules":[],"payloads":["5eb0e4b5b6c3bb0006eeb1e1"],"launchpad":"5e9e4502f5090995de566f86","flight_number":1,"name":"FalconSat","date_utc":"2006-03-24T22:30:00.000Z","date_unix":1143239400,"date_local":"2006-03-25T10:30:00+12:00","date_precision":"hour","upcoming":false,"cores":[{"core":"5e9e289df35918033d3b2623","flight":1,"gridfins":false,"legs":false,"reused":false,"landing_attempt":false,"landing_success":null,"landing_type":null,"landpad":null}],"auto_update":true,"tbd":false,"launch_library_id":null,"id":"5eb87cd9ffd86e000604b32a"},{"fairings":{"reused":false,"recovery_attempt":false,"recovered":false,"ships":[]},"links":{"patch":{"small":"https://images2.imgbox.com/f9/4a/ZboXReNb_o.png","large":"https://images2.imgbox.com/80/a2/bkWotCIS_o.png"},"reddit":{"campaign":null,"launch":null,"media":null,"recovery":null},"flickr":{"small":[],"original":[]},"presskit":null,"webcast":"https://www.youtube.com/watch?v=Lk4zQ2wP-Nc","youtube_id":"Lk4zQ2wP-Nc","article":"https://www.space.com/3590-spacex-falcon-1-rocket-fails-reach-orbit.html","wikipedia":"https://en.wikipedia.org/wiki/DemoSat"},"static_fire_date_utc":null,"static_fire_date_unix":null,"net":false,"window":0,"rocket":"5e9d0d95eda69955f709d1eb","success":false,"failures":[{"time":301,"altitude":289,"reason":"harmonic oscillation leading to premature engine shutdown"}],"details":"Successful first stage burn and transition to second stage, maximum altitude 289 km, Premature engine shutdown at T+7 min 30 s, Failed to reach orbit, Failed to recover first stage","crew":[],"ships":[],"capsules":[],"payloads":["5eb0e4b6b6c3bb0006eeb1e2"],"launchpad":"5e9e4502f5090995de566f86","flight_number":2,"name":"DemoSat","date_utc":"2007-

```
static_json = response.json()
data = pd.json_normalize(static_json)
```

	static_fire_date_utc	static_fire_date_unix	net	window	rocket	success	failures	details	crew	ships	capsules	payloa
0	2006-03-17T00:00:00.000Z	1.142554e+09	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 33, 'altitude': None, 'reason': 'merlin engine failure'}]	Engine failure at 33 seconds and loss of vehicle	[]	[]	[]	[5eb0e4b5b6c3bb0006eeb1e
1	None	NaN	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 301, 'altitude': 289, 'reason': 'harmonic oscillation leading to premature engine shutdown'}]	Successful first stage burn and transition to second stage, maximum altitude 289 km, Premature engine shutdown at T+7 min 30 s, Failed to reach orbit, Failed to recover first stage	[]	[]	[]	[5eb0e4b6b6c3bb0006eeb1e
2	None	NaN	False	0.0	5e9d0d95eda69955f709d1eb	False	[{'time': 140, 'altitude': 35, 'reason': 'residual stage-1 thrust led to collision between stage 1 and stage 2'}]	Residual stage 1 thrust led to collision between stage 1 and stage 2	[]	[]	[]	[5eb0e4b6b6c3bb0006eeb1e 5eb0e4b6b6c3bb0006eeb1e

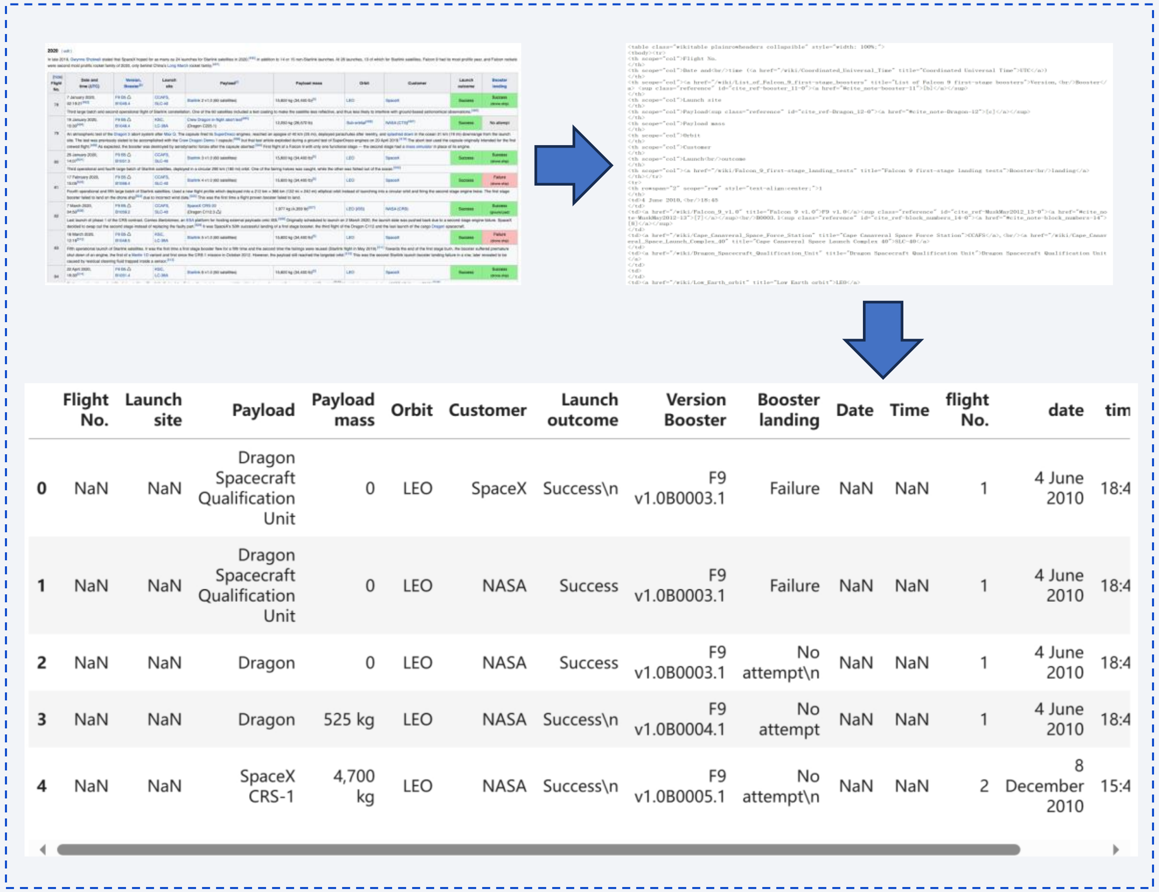
Date	BoosterVersion
2006-03-24	Falcon 1
2007-03-21	Falcon 1
2008-09-28	Falcon 1
2009-07-13	Falcon 1
2010-06-04	Falcon 9

Date	BoosterVersion
2010-06-04	Falcon 9
2012-05-22	Falcon 9
2013-03-01	Falcon 9
2013-09-29	Falcon 9
2013-12-03	Falcon 9

Data Collection - Scraping

In this part:

- Requests of rocket launch data had been sent to Wiki
- Target was parsed with beautiful soup
- A dictionary had been defined to reorganize the information and it was later turned into a dataframe
- For detail of the workflow, you may check the notebook here.



```
<table class="wikitable plainrowheaders collapsible" style="width: 100%;">
<tbody><tr>
<th scope="col">Flight No.
</th>
<th scope="col">Date and<br/>time (<a href="/wiki/Coordinated_Universal_Time" title="Coordinated Universal Time">UTC</a>)
</th>
<th scope="col"><a href="/wiki/List_of_Falcon_9_first-stage_boosters" title="List of Falcon 9 first-stage boosters">Version,<br/>Booster</
a> <sup class="reference" id="cite_ref-booster_11-0"><a href="#cite_note-booster-11">[b]</a></sup>
</th>
<th scope="col">Launch site
</th>
<th scope="col">Payload<sup class="reference" id="cite_ref-Dragon_12-0"><a href="#cite_note-Dragon-12">[c]</a></sup>
</th>
<th scope="col">Payload mass
</th>
<th scope="col">Orbit
</th>
<th scope="col">Customer
</th>
<th scope="col">Launch<br/>outcome
</th>
<th scope="col"><a href="/wiki/Falcon_9_first-stage_landing_tests" title="Falcon 9 first-stage landing tests">Booster<br/>landing</a>
</th></tr>
<tr>
<th rowspan="2" scope="row" style="text-align:center;">1
</th>
<td>4 June 2010,<br/>18:45
</td>
<td><a href="/wiki/Falcon_9_v1.0" title="Falcon 9 v1.0">F9 v1.0</a><sup class="reference" id="cite_ref-MuskMay2012_13-0"><a href="#cite_no
te-MuskMay2012-13">[7]</a></sup><br/>B0003.1<sup class="reference" id="cite_ref-block_numbers_14-0"><a href="#cite_note-block_numbers-14">
[8]</a></sup>
</td>
<td><a href="/wiki/Cape_Canaveral_Space_Force_Station" title="Cape Canaveral Space Force Station">CCAFS</a>,<br/><a href="/wiki/Cape_Canav
eral_Space_Launch_Complex_40" title="Cape Canaveral Space Launch Complex 40">SLC-40</a>
</td>
<td><a href="/wiki/Dragon_Spacecraft_Qualification_Unit" title="Dragon Spacecraft Qualification Unit">Dragon Spacecraft Qualification Unit
</a>
</td>
<td>
</td>
<td><a href="/wiki/Low_Earth_orbit" title="Low Earth orbit">LEO</a>
```

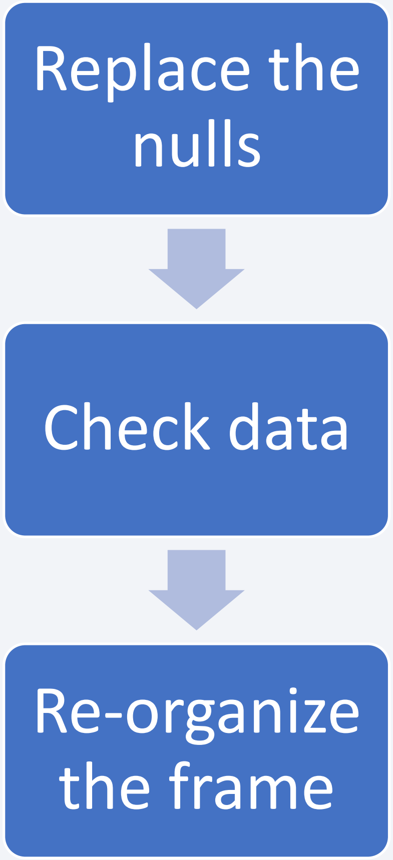
	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time	flight No.	date	tim
0	NaN	NaN	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	NaN	NaN	1	4 June 2010	18:4
1	NaN	NaN	Dragon Spacecraft Qualification Unit	0	LEO	NASA	Success	F9 v1.0B0003.1	Failure	NaN	NaN	1	4 June 2010	18:4
2	NaN	NaN	Dragon	0	LEO	NASA	Success	F9 v1.0B0003.1	No attempt\n	NaN	NaN	1	4 June 2010	18:4
3	NaN	NaN	Dragon	525 kg	LEO	NASA	Success\n	F9 v1.0B0004.1	No attempt	NaN	NaN	1	4 June 2010	18:4
4	NaN	NaN	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0005.1	No attempt\n	NaN	NaN	2	8 December 2010	15:4

Data Wrangling

During the process to get the ideal dataframe:

- All null values had been replaced by average number to ensure no abnormalities
- Types of data were checked
- Occurrence of orbits and outcomes were also checked
- The “class” column was created from outcomes, to comprehensively separate good and bad outcomes

- For detail of the workflow, you may check the notebook here.



Replace the nulls

Check data

Re-organize the frame

EDA with Data Visualization

- Scatter was used here to estimate the relationship between different factors
  - FlightNumber vs. Launch Site: relationship between Flight Number and Launch Site
  - Payload Mass vs. Launch Site: relationship between Payload Mass and Launch Site
  - Flight Number vs. Orbit Type: relationship between Flight Number and Orbit type
  - Payload vs. Orbit Type: relationship between Payload and Orbit type
- Success Rate vs. Orbit Type: a bar chart to compare the success rates of each orbit
- A Line was also plotted to show the Launch Success Yearly Trend:

- Further detail, please check the notebook here.

EDA with SQL

- SQL queries performed to this data set
  - select distinct Launch_Site from SPACEXTABLE
  - select * from SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5
  - select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE Customer='NASA (CRS)'
  - select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE Booster_Version='F9 v1.1'
  - select min(Date) from SPACEXTABLE where Landing_Outcome='Success (ground pad)'
  - select * from SPACEXTABLE where Landing_Outcome='Success (drone ship)' AND PAYLOAD_MASS__KG_>4000 AND PAYLOAD_MASS__KG_<6000
  - select Mission_Outcome, count(*) from SPACEXTABLE group by Mission_Outcome
  - select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
  - select substr(Date, 6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where Landing_Outcome='Failure (drone ship)' and substr(Date,0,5)='2015'
  - select Landing_Outcome, COUNT(*) AS Outcome_No from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by Outcome_No DESC
- Result of queries, please check here

Build an Interactive Map with Folium

- Several objects have been created and added to a folium map, including:
  - Circles and markers for all launch sites on the map
  - Markers for the success/failed launches for each site on the map
  - Line for the closest coast line
- This is for you to:
  - Notice the location of all sites, and it will enable you to select the site and read its description
  - Find a cluster of outcomes from different sites, which you may click to see every outcomes happened
  - Understand the selection of those locations for safety concern of a bad outcome

- Further detail, please check here

Build a Dashboard with Plotly Dash

- Here are 5 pies and 5 scatters in this dashboard:
  - If you select all sites(default), it will be the share of launches from 4 sites, and the scatter down below will contain all samples
  - If you select a launch site, it will return you a the success share of this site, and the scatter down below will contain the samples from this site
  - You can also use the slide bar to select the range of payload

- For the code, please check here

Predictive Analysis (Classification)

- In order to predict the probability of a good or bad outcome, column “class” was set to Y, other columns were set to X
- 20% of the data set was set as test group
- 4 models have been applied to this dataset
    - Logic regression
    - SVM
    - Decision tree
    - KNN
- Score function and confusion matrix were applied to evaluate the models
- Eventually, all models returned the same result in test
- For the full process, please check here

Results

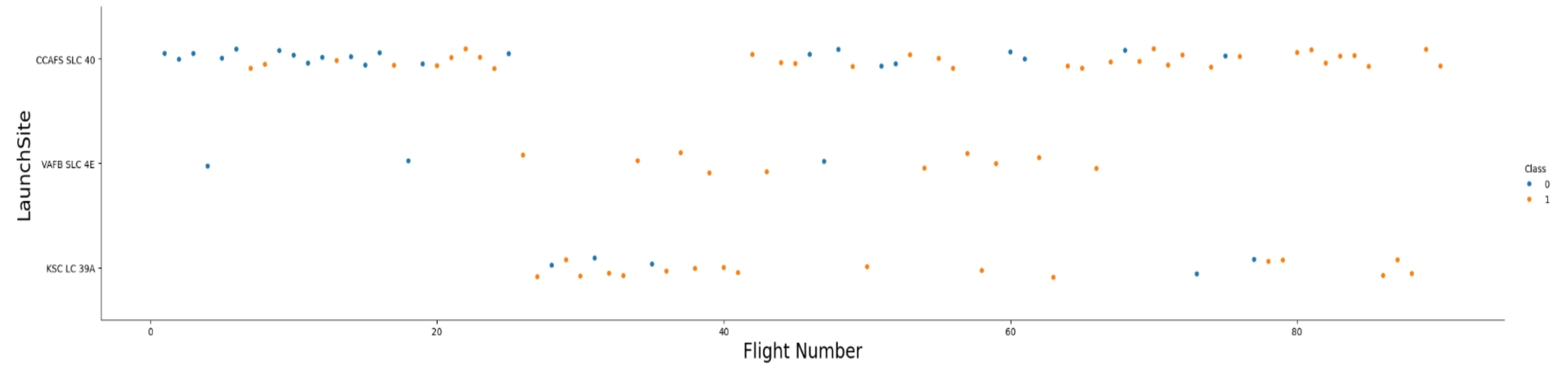
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

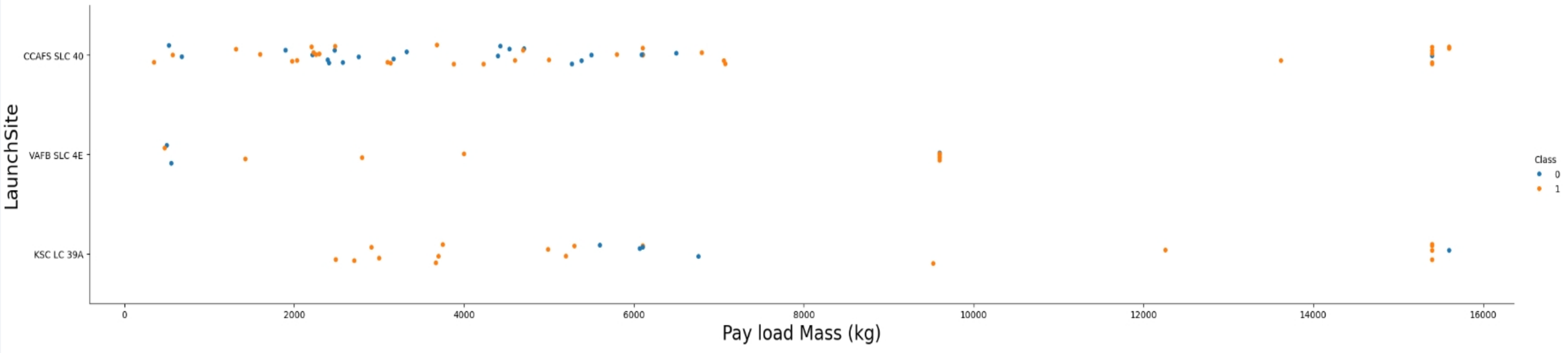
Flight Number vs. Launch Site

- VAFB SLC 4E has the best success rate
- However, it may because of its small sample
- As a result, we cannot say the outcome is relevant with launch site yet



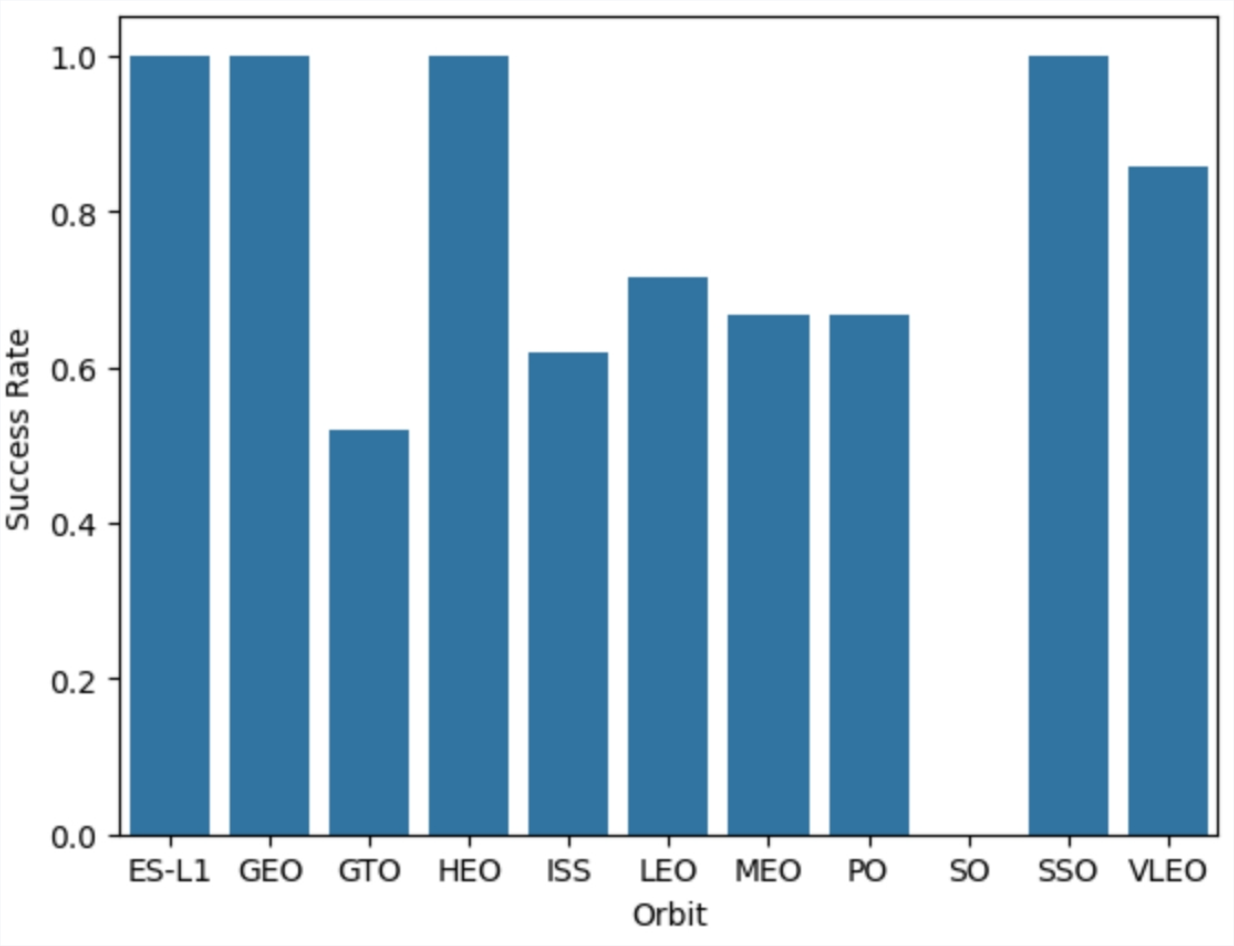
Payload vs. Launch Site

- Replacing the flight number with payload can significantly bring some change
- Obviously, VAFB SLC 4E has several outcomes from the same payload which is in a high success rate area
- Compare with the launch site, the payload seems to be a bigger factor
- However, though the sample size is still small, KSC LC 39A and VAFB 4E showed better potential than CCAFS SLC40 when payload is between 1500 and 5500



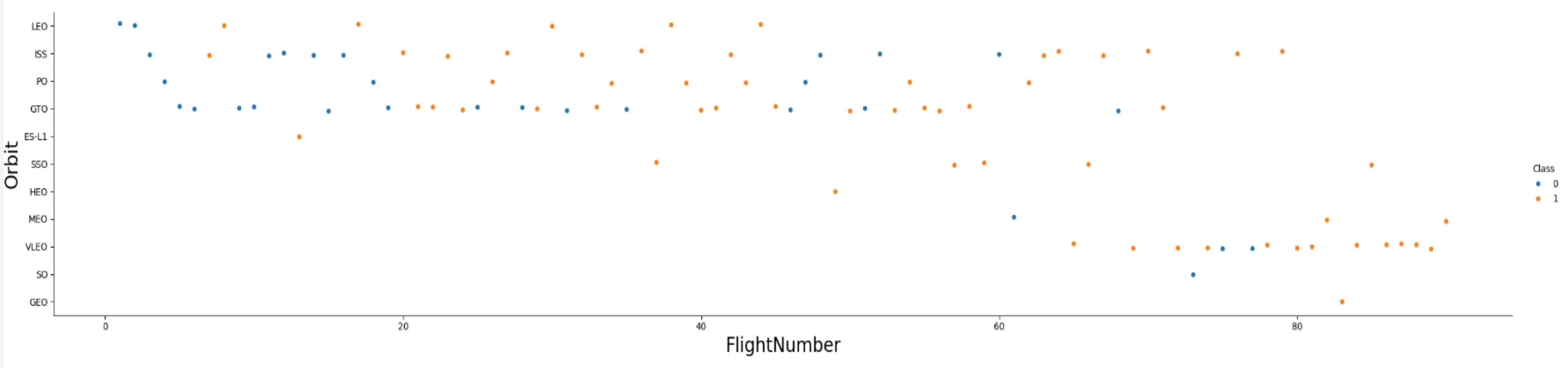
Success Rate vs. Orbit Type

- Similar process can been repeated in Orbit types
- A bar chart can pick the Orbits with better success rate immediately



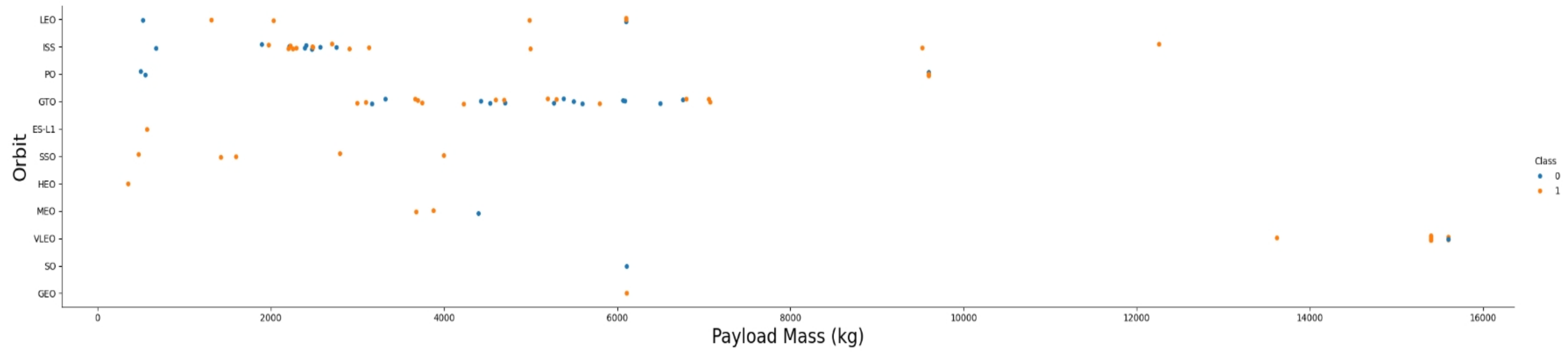
Flight Number vs. Orbit Type

- The flight number shows that several Orbits have significant less samples
- Most 100% success Orbits has only one outcome
- SSO is the only 100% success Orbit with more than one outcome, but still a small subset



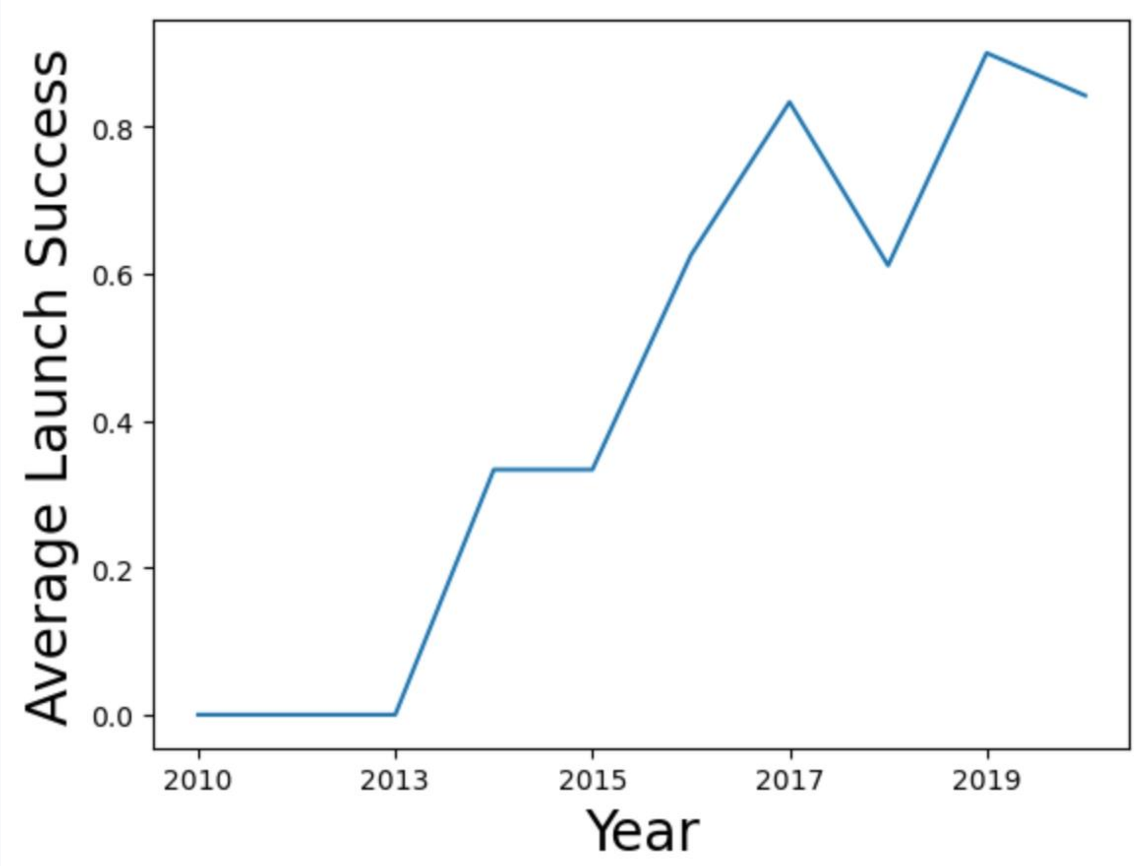
Payload vs. Orbit Type

- All SSO samples are from the payload area 0-6000, where you may still remember a lot of failure from CCAFS site
- LEO is not too far from SSO with only twice bad outcome in the same payload range
- Still, payload > 6000 range, not many bad outcomes



Launch Success Yearly Trend

- In flight number scatter, a noticeable fact is: bigger number works better
- The flight number is relevant to the time and experience, and this guess is verified by the trend line right side



All Launch Site Names

- Through SQL query, here you may find the unique members of Launch_Site family. In another word, we got 4 Launch Sites in total and here they are.

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

Launch Site Names Begin with 'CCA'

- There are 2 launch sites starts with ‘CCA’
- They are filtered out

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

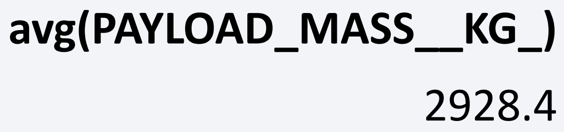
Total Payload Mass

- the total payload carried by boosters from NASA is 45596kg

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.4kg



avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad is 2015/12/22

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- 4 records have been found landed on a drone ship with payload 4000-6000 kg
- 2 from CCAFS LC-40 with payload around 4650, 2 from KSC LC-39A with payload around 5250
- All for GTO

Date	Time (UTC)	Booster_ Version	Launch_Site	Payload	PAYLOAD_ MASS__KG_	Orbit	Customer	Mission_ Outcome	Landing_ Outcome
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)

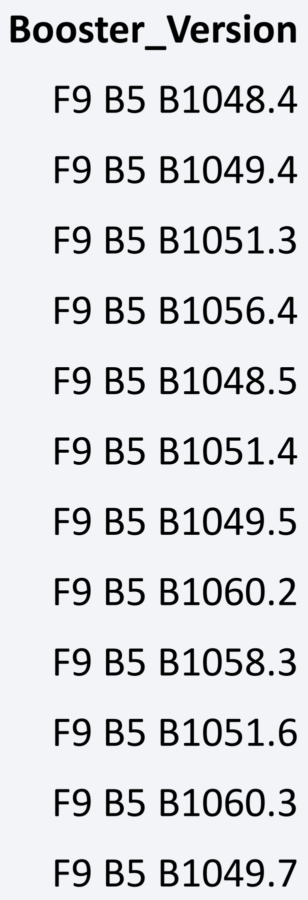
Total Number of Successful and Failure Mission Outcomes

- Almost all the boosters recorded as success in mission outcome category
- However mission success doesn’t mean successfully landed

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

Boosters Carried Maximum Payload

- Here we got the booster carried the maximum payload mass



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

- Here is the ones failed to land in drone ship in 2015
- Both of them are from 1H and Booster Version is reasonably close

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

- From 2010-06-04 to 2017-03-20, the most landing outcome is “no attempt” with 10 times
- For drone ship, its landing outcome is evenly distributed as 5
- Boosters successfully landed on the ground pad for 3 times

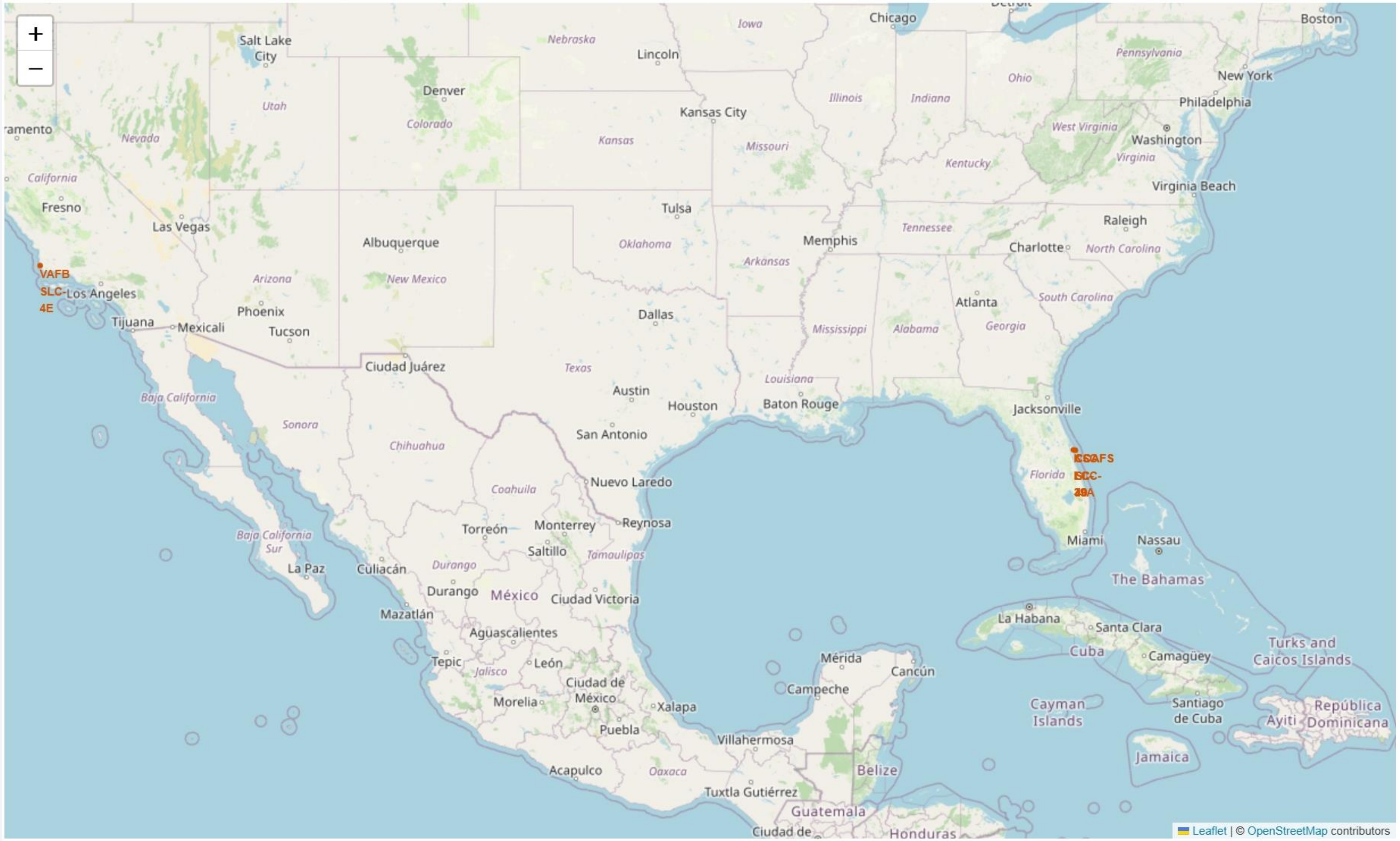
Landing_Outcome	**Outcome_No**
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

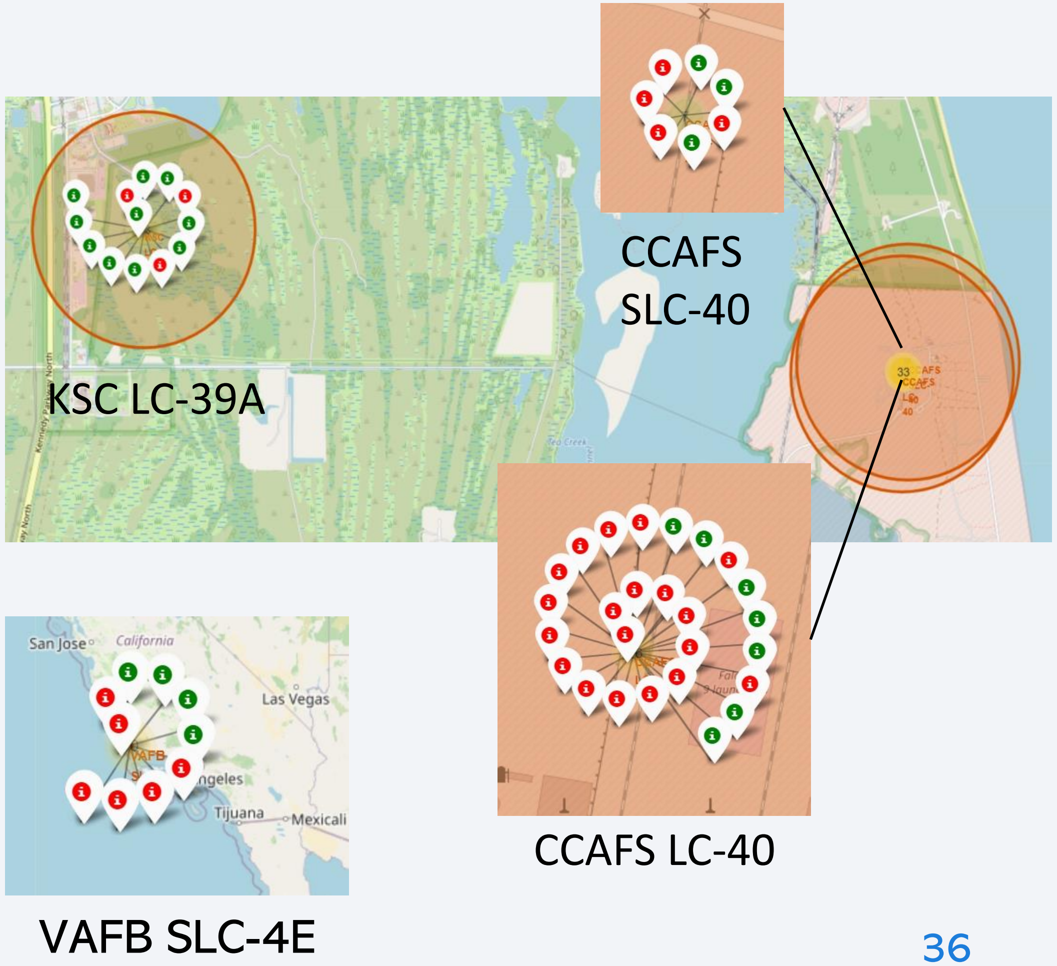
Location of Launch sites

- Through this map, we can easily find some fact:
    - Only 1 site from west coast
    - 3 sites located very closely in Florida
    - All the sites are beside coasts



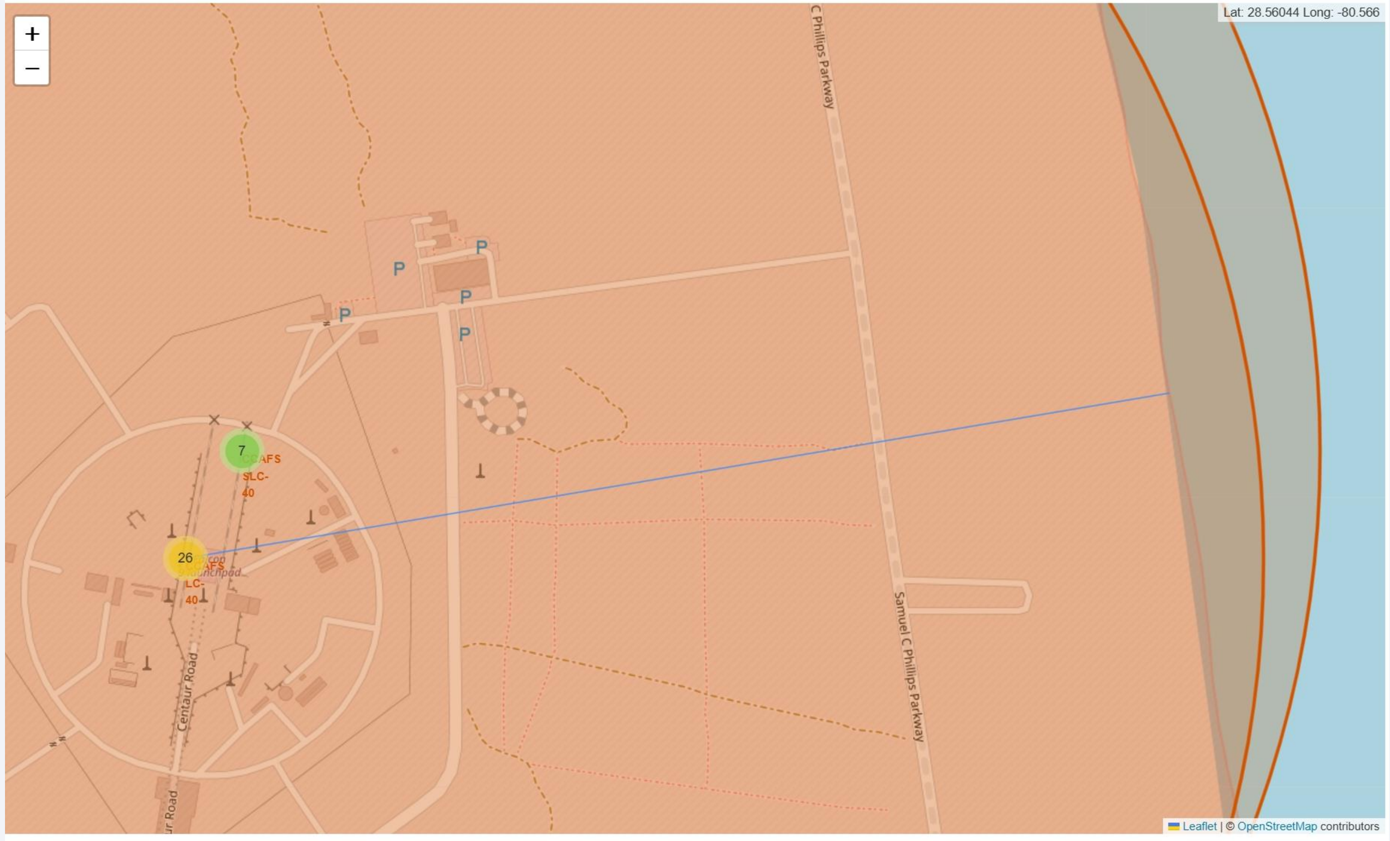
Outcomes from each Launch Sites

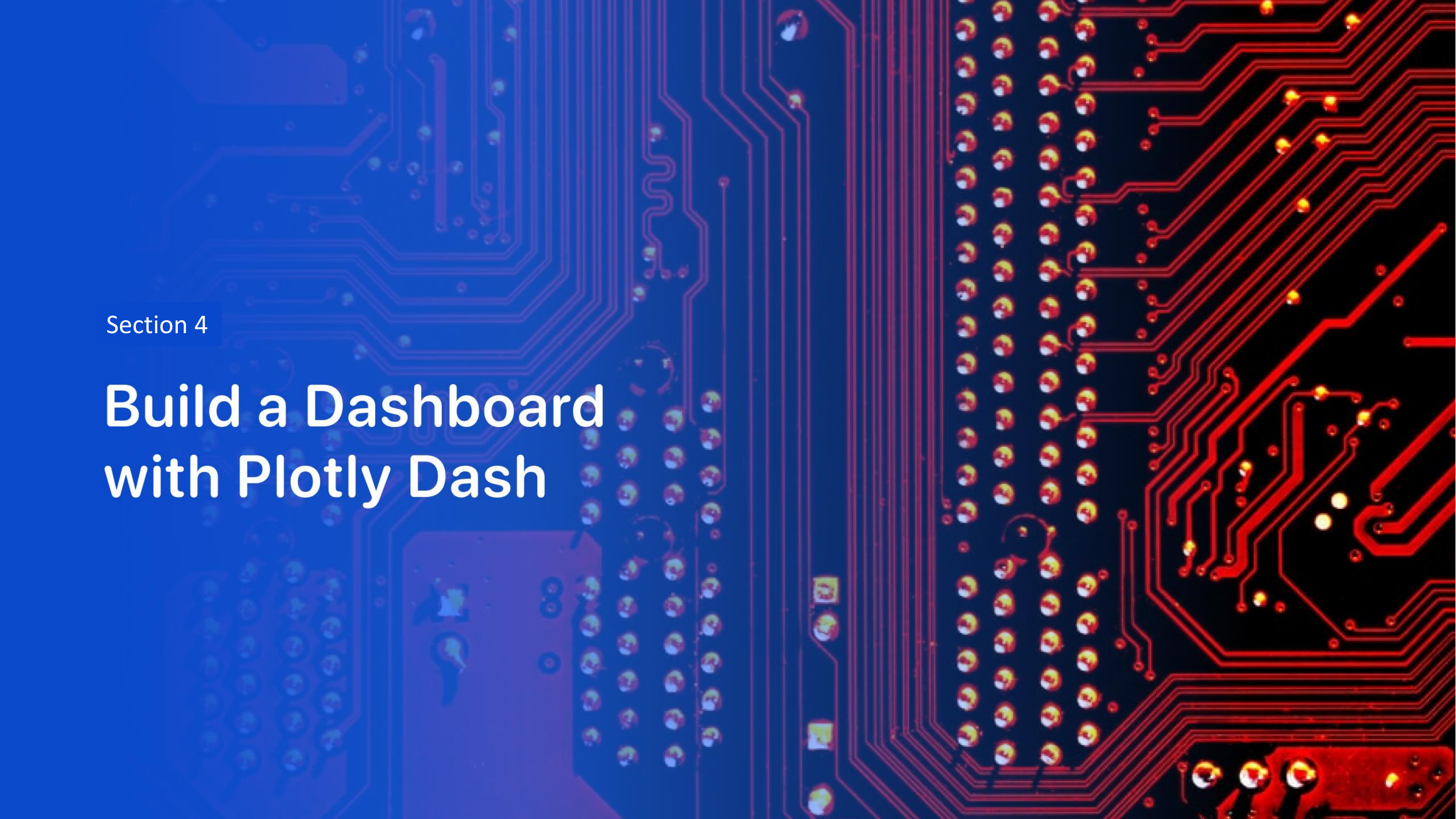
- KSC LC-39A has the best record
- All other sites have bad outcomes more than good outcomes
- CCAFS SLC-40 located at almost the same place of CCAFS LC-40 but recorded 3 good outcomes and 4 bad outcomes, which is closer to 50%



CCAFS sites are very close to the coast

- Different from KSC LC-39A, both CCAFS sites are located less than 1km from the coast line.
- As previously mentioned, CCAFS sites has the most launch record, also the failed ones.
- Such location offered better chance to turn bad outcome into controlled or uncontrolled (ocean)


+
−
Lat: 28.56044 Long: -80.566
C Phillips Parkway
Samuel C Phillips Parkway
Centaur Road
P
P
P
P
P
7
CCAFS SLC-40
26
CCAFS LC-40
Launchpad
Leaflet | © OpenStreetMap contributors

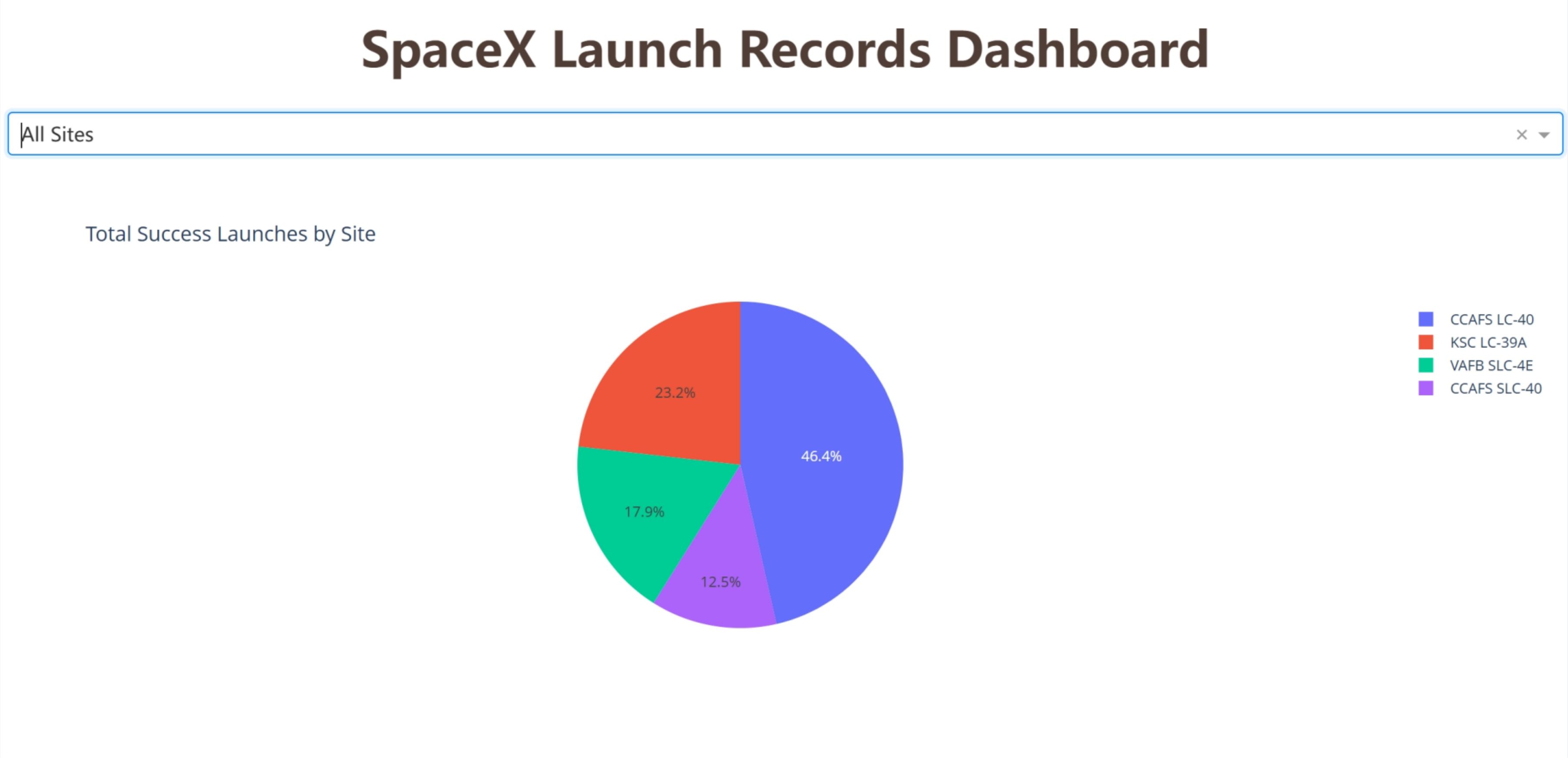


Section 4

Build a Dashboard with Plotly Dash

This is the SpaceX Dashboard

- Here you can see a dropdown option at the top to select different sites. It will change the pie chart below to your selection.

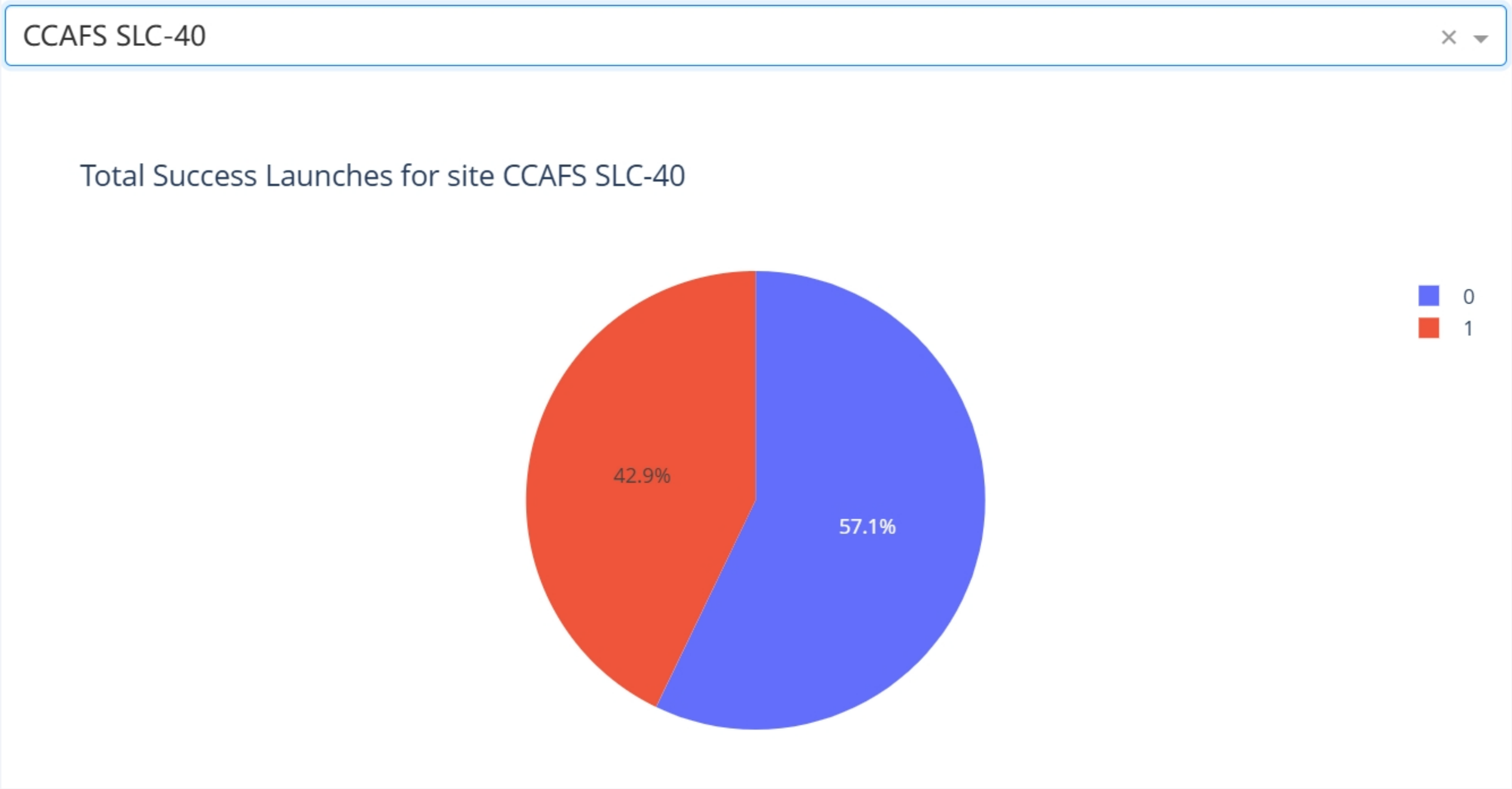


SpaceX Launch Records Dashboard

All Sites

The Launch Site with Highest Success Ratio

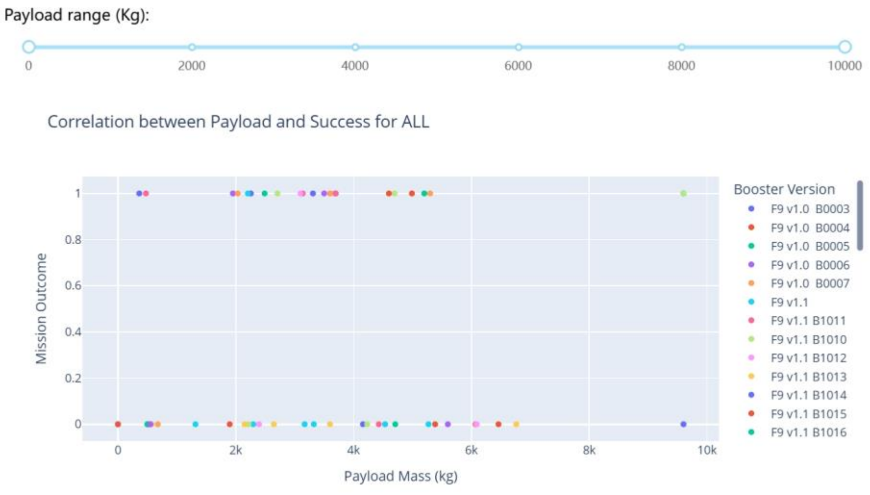
- In the preprocessed data, CCAFS SLC-40 has the highest success ratio
- Here, 0 represents bad outcome and 1 is the good outcome, so the success ratio of CCAFS SLC-40 is 42.9%

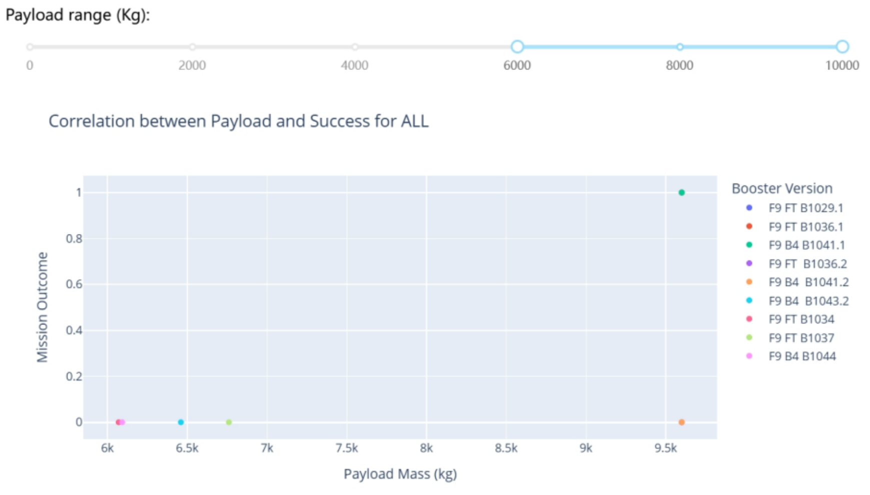


CCAFS SLC-40

Total Success Launches for site CCAFS SLC-40

Payload is a bigger factor





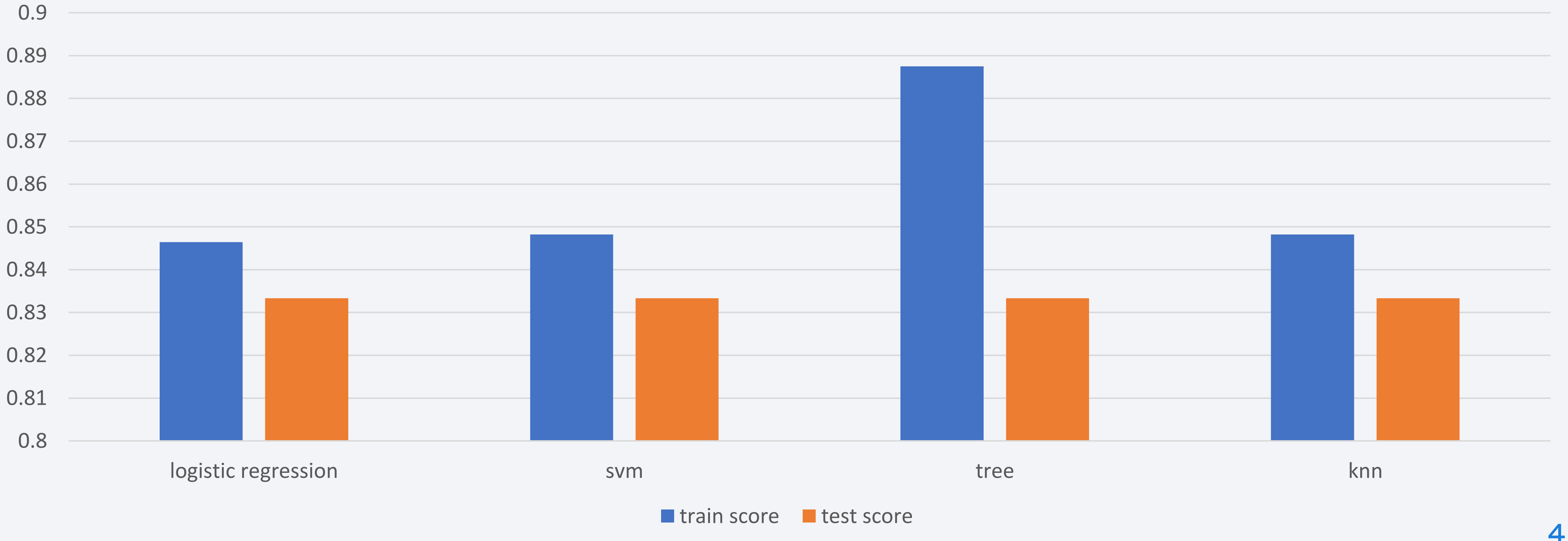
- From the scatter, we can notice that most success outcomes are from payload mass < 6000 kg range
- When payload > 6000kg, there are only 6 samples and only 1 success

Section 5

Predictive Analysis (Classification)

Classification Accuracy

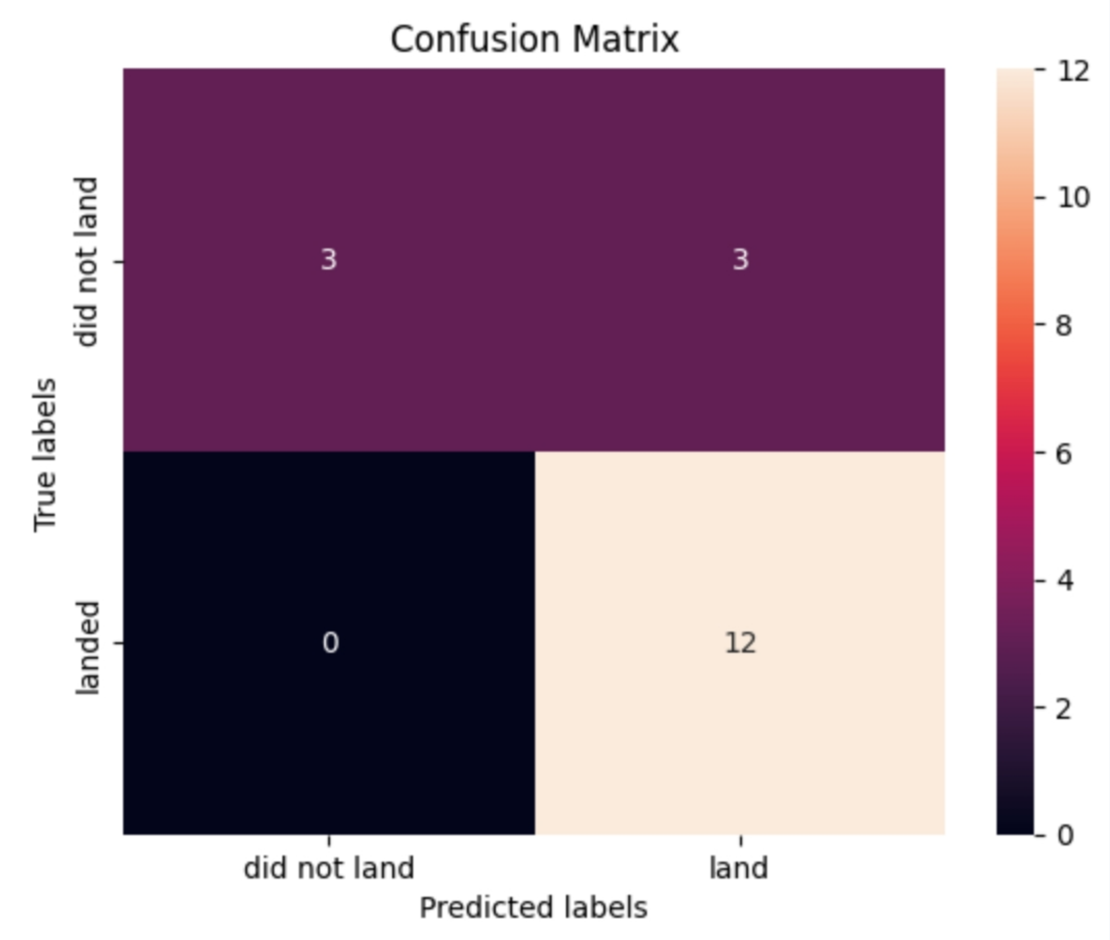
- The tree model has the highest classification accuracy for train set
- But all the models returned exactly same for test set, which means they works similar



Confusion Matrix

- All 4 models returned the same matrix it is shown on the right

- This result works perfectly on landed group but false land still happens which may get people over confident



Conclusions

- Through the data, we noticed that some factors such as payload mass, orbit, and time are all influential to the outcome
- With data science analysis, we can easily collect massive amount of record and turn it into one meaningful dataframe
- Through folium map, we can find some common attributes from the launch sites, for example they all located not far from the coast
- When we are trying to predict the outcome or class, all the method including logistic regression, SVM, decision tree, KNN, are doing well
- The analysis shows good potential to predict subsequent launch of SpaceX with iteration

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!