Space racing with Data Science

Luiz Guardini

March 6th 2024

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusions

EXECUTIVE SUMMARY

- Objective: Determine the price of each launch of a commercial exploration launcher.
  - Can we reuse the first stage rocket?
  - Can we create a company to compete in the space exploration using SpaceX data to determine the reusability of the first stage launcher?
- By gathering information from SpaceX space program (methodologies).
  - Collect public data through APIs and Web Scraping.
  - Complete Data Wrangling
  - Perform Exploratory Data Analysis with SQL and Data Visualization
  - Implement Visual Analysis with Data Visualization and Folium
  - Execute Predictive Analysis with the aid of Machine Learning techniques, such as SVM, KNN, Decision Trees and Logistic Regression
- Summary of the results:
  - SpaceY has to be ready for a learning curve period where most likely many unsuccessful mission will occur, but with the correct investment, it is possible to create a solid space exploration company.
  - This analysis helped us to understand that the success rate of over 76% can be achieved by choosing the right location for a launch site in combination with Orbit and Payload values.
  - Detailed results will be provided though out the presentation.

Section 1

Methodology

METHODOLOGY

DATA COLLECTION AND DATA WRANGLING

Data Collection from public SpaceX data.

- api.spacexdata.com/v4/
- API has different endpoints:
  - /capsules
  - /cores
  - /past
  - /rockets
  - /launchpad
  - /payloads

Web scraping

- Gather data from SpaceX wiki page
- Use BeautifulSoup to handle data

Data Wrangling

- Convert variables to correct type
- Identify and correct missing values
- Remove irrelevant data

METHODOLOGY

DATA COLLECTION AND DATA WRANGLING – SPACE-X API

SpaceX REST API

Open Source REST API for launch, rocket, core, capsule, starlink, launchpad, and landing pad data.

- Data Collection from public SpaceX data.
  - api.spacexdata.com/v4/
  - API has different endpoints:
    - /capsules
    - /cores
    - /past
    - /rockets
    - /launchpad
    - /payloads

METHODOLOGY

DATA COLLECTION AND DATA WRANGLING – SPACE-X API

https://api.spacexdata.com/v4/

https://api.spacexdata.com/v4/capsules/


```
[{"reuse_count":0,"water_landings":1,
1.0","id":"5e9e2c5bf35918ed873b2664"}
1.0","id":"5e9e2c5bf3591882af3b2665"}
1.0","id":"5e9e2c5bf3591835983b2666"}
1.0","id":"5e9e2c5bf359189ef23b2667"}
1.1","id":"5e9e2c5bf3591859a63b2668"}
["5eb87ce7ffd86e000604b33b","5eb87d03
{"reuse_count":0,"water_landings":1,"
{"reuse_count":2,"water_landings":3,"
["5eb87cecffd86e000604b33f","5eb87d0e
{"reuse_count":0,"water_landings":1,"
```


https://api.spacexdata.com/v4/cores/


```
[{"block":null,"reuse_count":0,"rtls_
["5eb87cd9ffd86e000604b32a"],"serial"
burn and transition to second stage,
["5eb87cdaffd86e000604b32b"],"serial"
led to collision between stage 1 and
{"block":null,"reuse_count":0,"rtls_a
RazakSAT.","launches":["5eb87cdbffd86
{"block":null,"reuse_count":0,"rtls_a
{"block":1,"reuse_count":0,"rtls_atte
{"block":1,"reuse_count":0,"rtls_atte
["5eb87cdeffd86e000604b330"],"serial"
```


…

https://api.spacexdata.com/v4/payloads/


```
[{"dragon":{"capsule":null,"mass_retu
2","type":"Satellite","reused":false,
["SSTL"],"mass_kg":20,"mass_lbs":43,"
earth","longitude":null,"semi_major_a
n_anomaly":null,"id":"5eb0e4b5b6c3bb0
{"capsule":null,"mass_returned_kg":nu
rs":["DARPA"],"norad_ids":[],"nationa
earth","longitude":null,"semi_major_a
"mean_anomaly":null,"id":"5eb0e4b6b6c
{"capsule":null,"mass_returned_kg":nu
tomers":["NASA"],"norad_ids":[],"nati
```

METHODOLOGY

DATA COLLECTION AND DATA WRANGLING – WEB SCRAPING

Gather data:

https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

[hide] Flight No.	Date and time (UTC)	Version, Booster [b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
1	4 June 2010, 18:45	F9 v1.0[7] B0003.1[8]	CCAFS, SLC-40	Dragon Spacecraft Qualification Unit		LEO	SpaceX	Success	Failure[9][10] (parachute)
	First flight of Falcon 9 v1.0.[11] Used a boilerplate version of Dragon capsule which was not designed to separate from the second stage.(more details below) Attempted to recover the first stage by parachuting it into the ocean, but it burned up on reentry, before the parachutes even deployed.[12]								
2	8 December 2010, 15:43[13]	F9 v1.0[7] B0004.1[8]	CCAFS, SLC-40	Dragon demo flight C1 (Dragon C101)		LEO (ISS)	NASA (COTS) NRO	Success[9]	Failure[9][14] (parachute)
	Maiden flight of Dragon capsule, consisting of over 3 hours of testing thruster maneuvering and reentry.[15] Attempted to recover the first stage by parachuting it into the ocean, but it disintegrated upon reentry, before the parachutes were deployed.[12] (more details below) It also included two CubeSats,[16] and a wheel of Brouère cheese.								
3	22 May 2012, 07:44[17]	F9 v1.0[7] B0005.1[8]	CCAFS, SLC-40	Dragon demo flight C2+[18] (Dragon C102)	525 kg (1,157 lb)[19]	LEO (ISS)	NASA (COTS)	Success[20]	No attempt
	Dragon spacecraft demonstrated a series of tests before it was allowed to approach the International Space Station. Two days later, it became the first commercial spacecraft to board the ISS.[17] (more details below)								
4	8 October 2012, 00:35[21]	F9 v1.0[7] B0006.1[8]	CCAFS, SLC-40	SpaceX CRS-1[22] (Dragon C103)	4,700 kg (10,400 lb)	LEO (ISS)	NASA (CRS)	Success	No attempt
				Orbcomm-OG2[23]	172 kg (379 lb)[24]	LEO	Orbcomm	Partial failure[25]	

Create BeautifulSoup object

Find tables

```
<table class="wikitable plainrowheaders
<tbody><tr>
<th scope="col">Flight No.
</th>
<th scope="col">Date and<br/>time (<a h
</th>
<th scope="col"><a href="/wiki/List_of_
</th>
<th scope="col">Launch site
</th>
<th scope="col">Payload<sup class="refe
</th>
<th scope="col">Payload mass
```

Create DataFrame

METHODOLOGY

DATA COLLECTION AND DATA WRANGLING – DATA WRANGLING

1. Data analysis → Identify which columns are numerical and categorical.

```
df.dtypes

FlightNumber      int64
Date             object
BoosterVersion   object
PayloadMass     float64
Orbit            object
LaunchSite       object
Outcome          object
Flights           int64
GridFins           bool
Reused             bool
Legs               bool
LandingPad       object
```

2. Deal with missing values → Replace missing Payload mass by its mean value

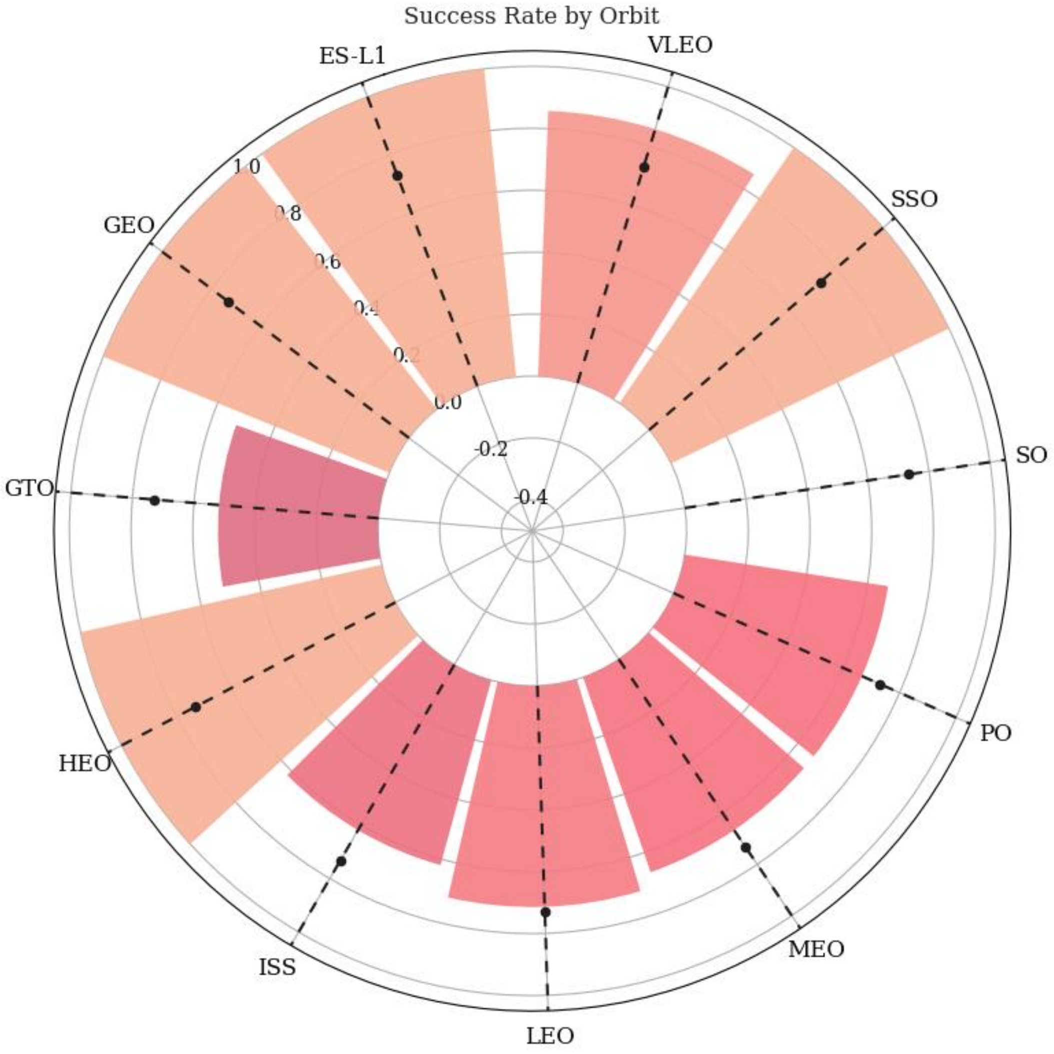
```
# Calculate the mean value of PayloadMass column
mean_payloadmass = data_falcon9['PayloadMass'].mean()

# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'].replace(np.NaN, mean_payloadmass, inplace=True)
```

3. Correct Mission labels → Create dummy variables to categorize successful/unsuccessful missions

	Class
0	0
1	0
2	0
3	0
4	0
5	0
6	1
7	1

METHODOLOGY

EDA with DATA VISUALIZATION



- Use Matplotlib and Seaborn to create relevant graphs to grasp information on the following indicators:
  - Evolution of successful missions.
  - Launch Success Yearly Trend.
  - Launch Site success rate.
  - Relationship between Payload and Launch Site.
  - Success launcher landing per Orbit.
  - Relationship between Flight Number and Orbit Type.

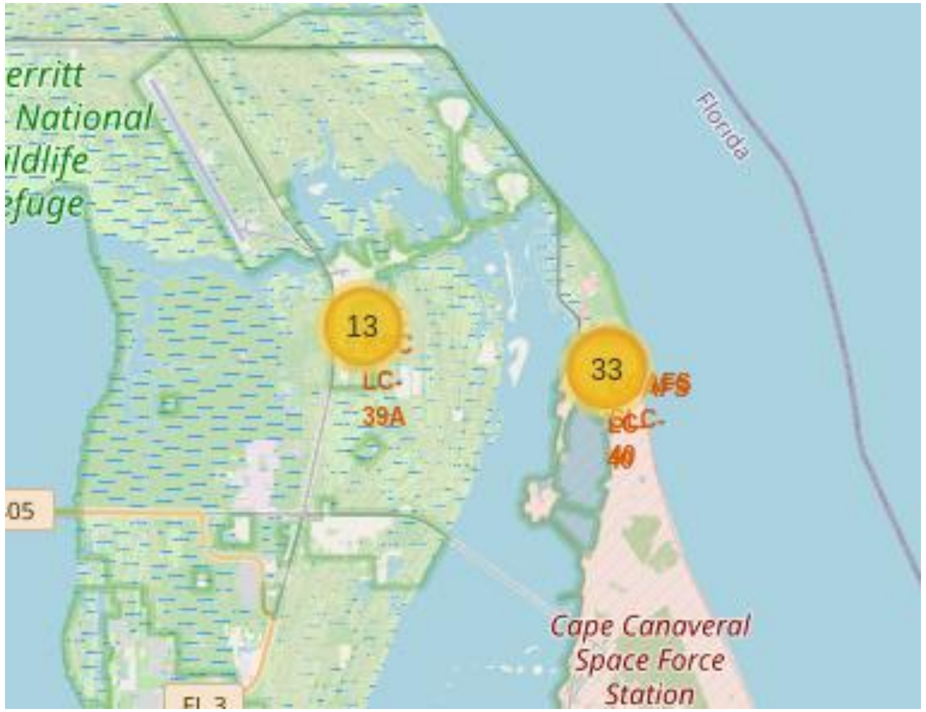
METHODOLOGY

EDA with SQL

- Load SpaceX dataset into PostgreSQL database to work within Jupyter Notebook.
- Get insights from the data for the following instances:
  1. Unique launch sites
  2. Records where launch sites begin with string 'CCA'
  3. Total Payload mass carried.
  4. Average payload mass carried by booster F9 v1.1.
  5. Date of the first successful landing.
  6. Boosters with success in drone ship with specific payload range
  7. Number of successful and failure mission outcomes
  8. List boost versions
  9. Failures for a corresponding year
  10. Count landing outcomes

METHODOLOGY

EDA with INTERACTIVE VISUAL ANALYTICS - Folium



- Mark all launch sites
  - Add map objects such as markers, circles and lines.
  - Mark success or failure of launches for each site
- Assign feature launch outcomes according to feature class (0 for failure, 1 for success)
- Use color-labeled marker clusters to identify site rate of success.
- Calculate the distance between the site and a point of interest.

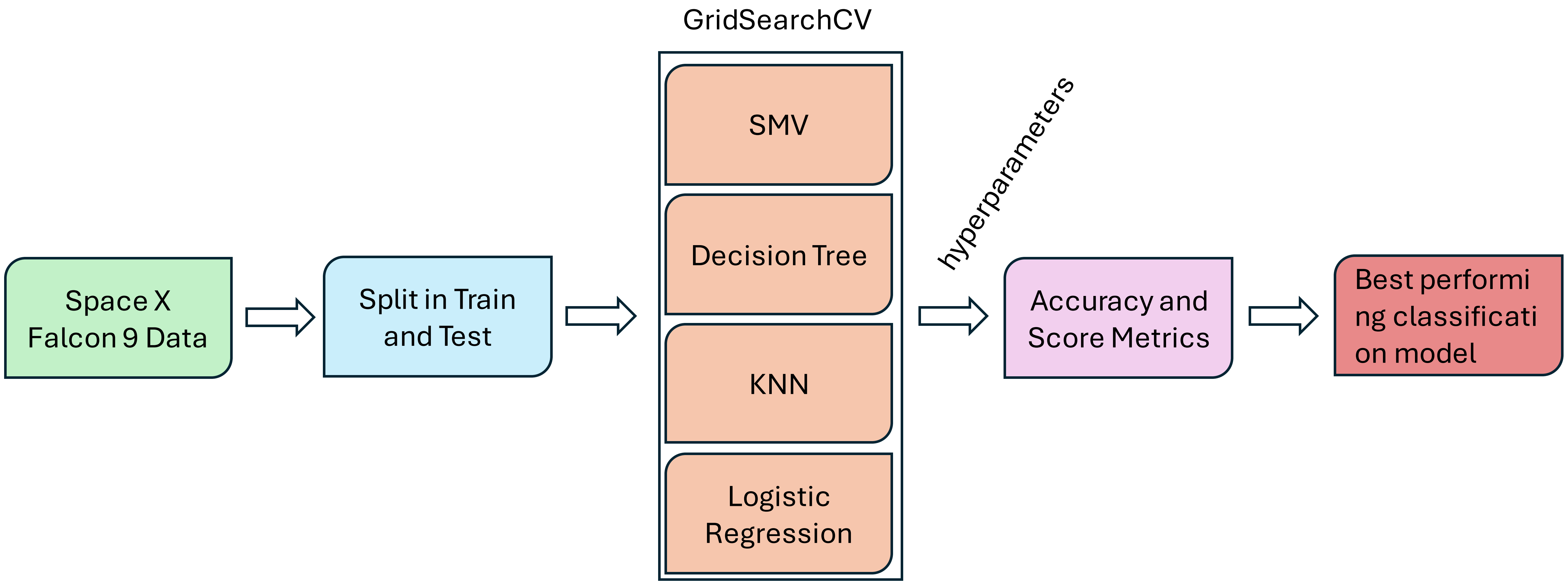
METHODOLOGY

EDA with INTERACTIVE VISUAL ANALYTICS - Dashboard

- Build an interactive dashboard with Ploty Dash
- Create a dropdown menu to select 'All sites' or a specific launch site
- Create a slide range to select payload mass
- Plot a pie chart to depict the launches per site
- Plot scatter graph to show relationship between Successful Outcome and Payload Mass [Kg] for different booster versions.

METHODOLOGY

PREDICTIVE ANALYSIS (CLASSIFICATION)

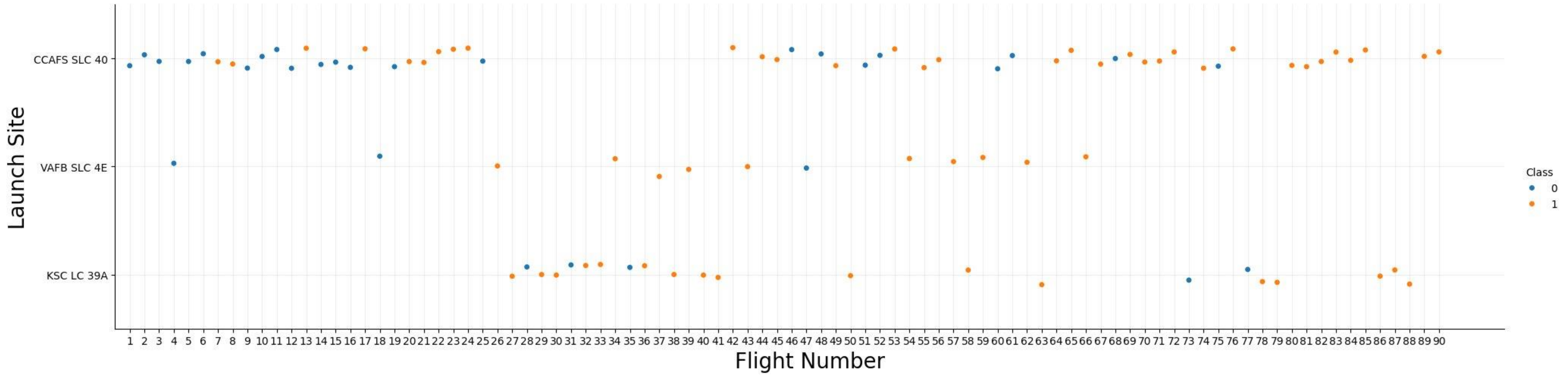


Section 2 - Results

Insights from visual EDA

RESULTS – EXPLORATORY DATA ANALYSIS

Flight Number vs. Launch Site

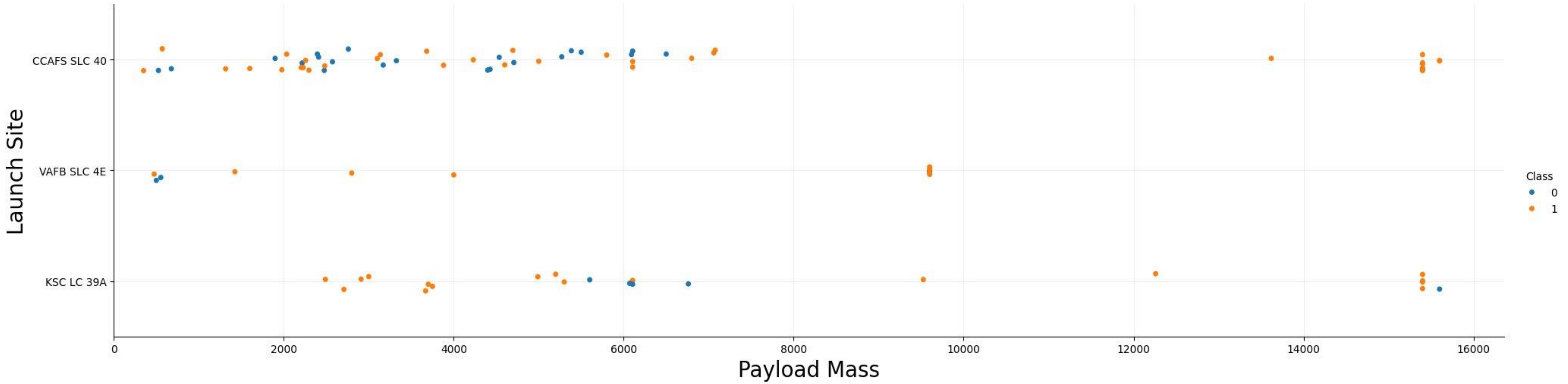


- Different launch sites have different success rates:
  - CCAFS LC-40, has a success rate of 60 %,
  - KSC LC-39A and VAFB SLC 4E have a success rate of 77%

	LaunchSite	Class
0	CCAFS SLC 40	0.600000
1	KSC LC 39A	0.772727
2	VAFB SLC 4E	0.769231

RESULTS – EXPLORATORY DATA ANALYSIS

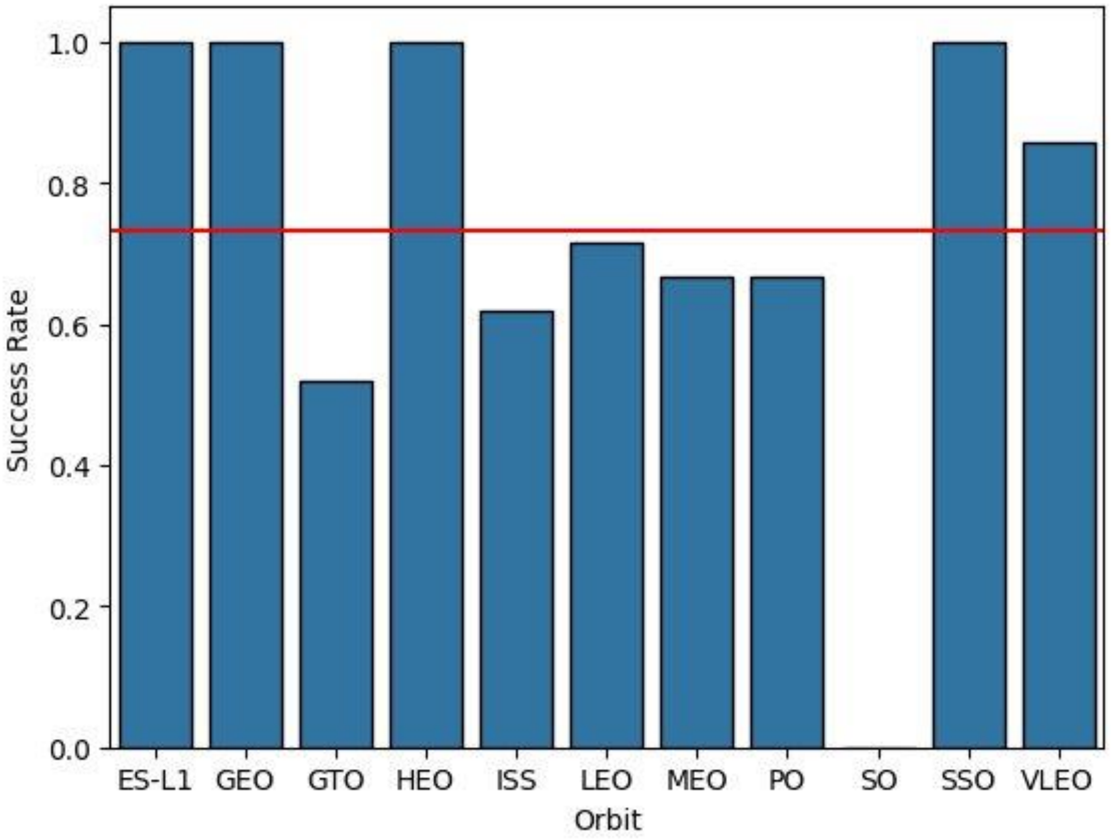
Payload vs. Launch Site



- VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).
- The higher the payload, the higher the success rate.

RESULTS – EXPLORATORY DATA ANALYSIS

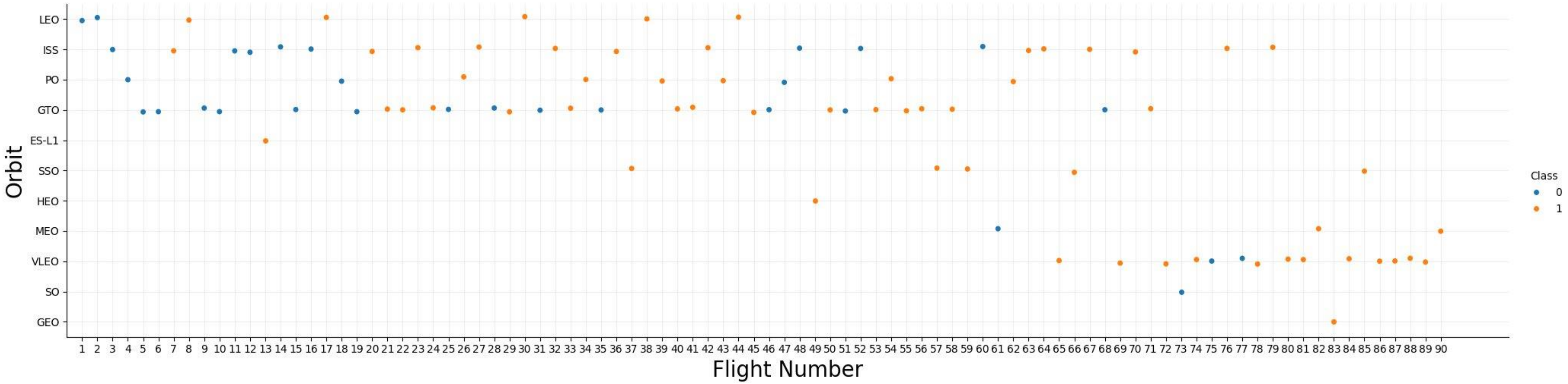
Success Rate vs. Orbit Type



- ES-L1, GEO, HEO, SSO and VLEO orbits have above the average success rate in stage 1 landing.
- ES-L1, GEO, HEO and SSO have 100% success rate in stage 1 landing.
- SO has no successful stage 1 landing

RESULTS – EXPLORATORY DATA ANALYSIS

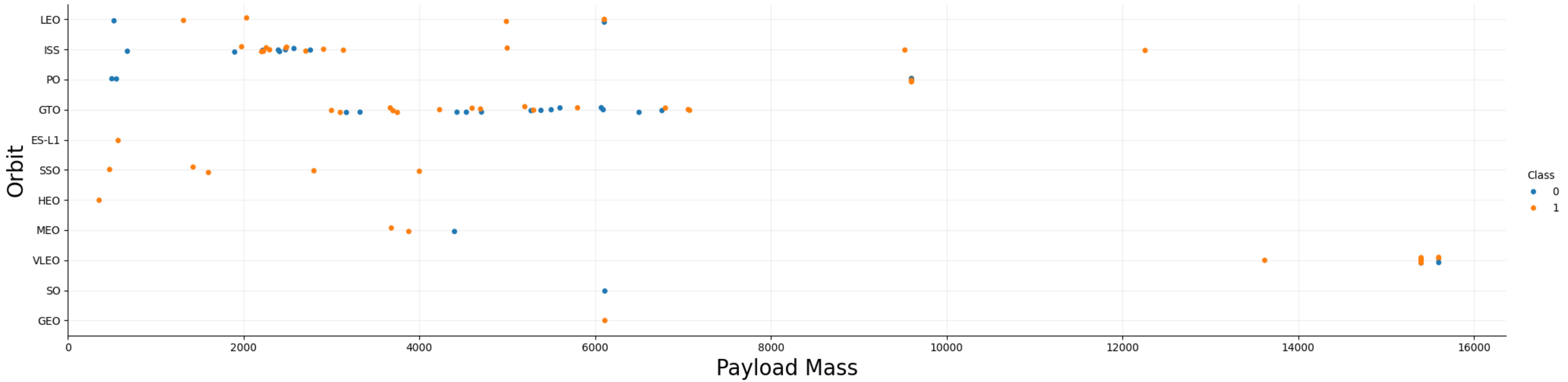
Flight Number vs. Orbit Type



- LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
- GEO has 100% success rate, but only one launch.
- SO has 100% failure rate, but only one launch.

RESULTS – EXPLORATORY DATA ANALYSIS

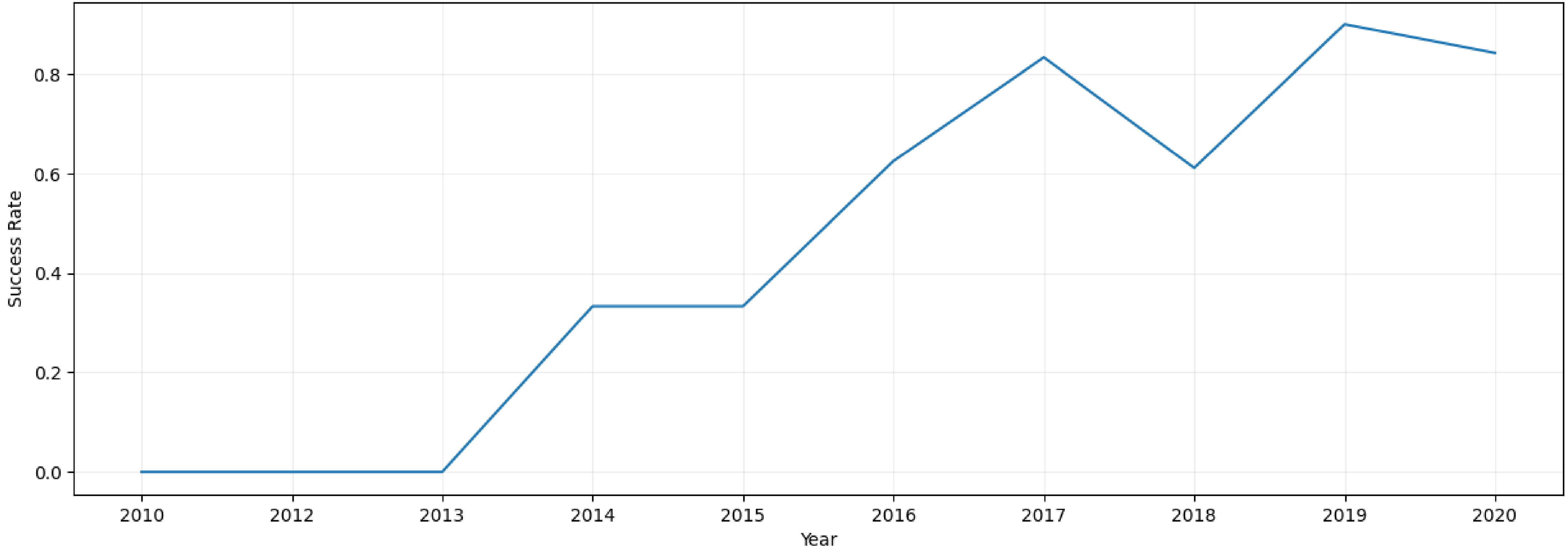
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- For GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

RESULTS – EXPLORATORY DATA ANALYSIS

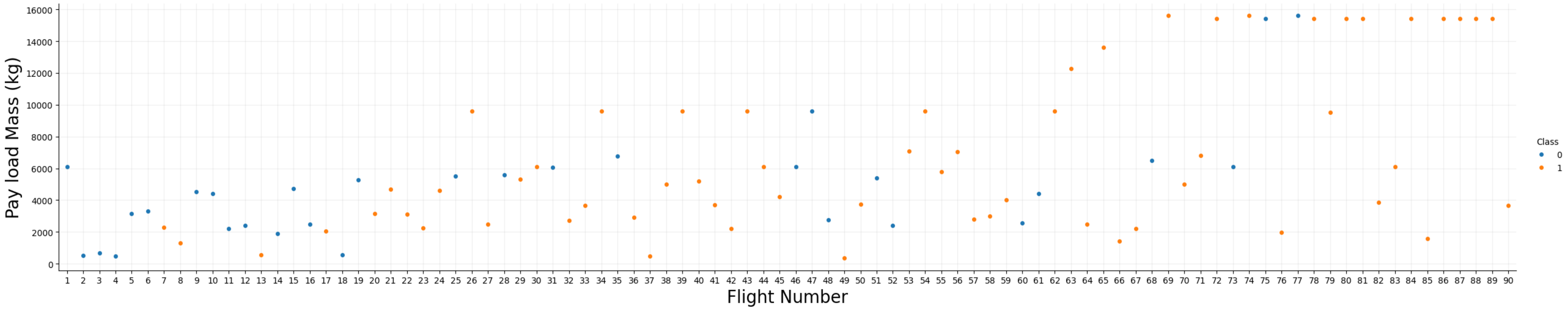
Launch success yearly trend



- Success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.

RESULTS – EXPLORATORY DATA ANALYSIS

Payload vs. Flight Number



- As the flight number increases, the first stage is more likely to land successfully (red dots).
- The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.

Section 3 - Results

Insights from SQL EDA

RESULTS – EDA - SQL

All launch site names:

```
%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

RESULTS – EDA - SQL

Launch site names begin with `CCA`:

```
%%sql SELECT * FROM SPACEXTABLE
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5
```

Python

* sqlite:///my_data1.db
Done.

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

RESULTS – EDA - SQL

Total payload mass:

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS 'TOTAL PAYLOAD MASS (KG)' FROM SPACEXTABLE WHERE Customer == 'NASA (CRS)'
✓ 0.0s

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

TOTAL PAYLOAD MASS (KG)
45596

RESULTS – EDA - SQL

Average payload mass by F9 v1.1:

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS 'AVERAGE PAYLOAD MASS (KG)' FROM SPACEXTABLE WHERE Booster_Version LIKE '%%F9 V1.1%'
```

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

AVERAGE PAYLOAD MASS (KG)
2534.6666666666665

RESULTS – EDA - SQL

First successful ground landing date:

```
%sql SELECT Date AS 'FIRST SUCCESSFUL LANDING DATE' FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success (ground pad)%' ORDER BY Date ASC LIMIT 1
```

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

FIRST SUCCESSFUL LANDING DATE
2015-12-22

RESULTS – EDA - SQL

Successful drone ship landing with payload between 4000 and 6000:

```
%%sql SELECT DISTINCT(Booster_Version), PAYLOAD_MASS__KG_ AS 'PAYLOAD MASS (KG)'
FROM SPACEXTABLE SPACEXTABLE WHERE Landing_Outcome LIKE 'Success (drone ship)%'
AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

✓ 0.0s

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

Booster_Version	PAYLOAD MASS (KG)
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

RESULTS – EDA - SQL

Total number of successful and failure mission outcomes:

```
%sql SELECT DISTINCT(Mission_Outcome), COUNT(Mission_Outcome) FROM SPACEXTABLE GROUP BY Mission_Outcome
✓ 0.0s

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

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

RESULTS – EDA - SQL

Boosters carried maximum payload:

```
%%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ AS 'MAX PAYLOAD MASS (KG)'
FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

✓ 0.0s

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

Booster_Version	MAX PAYLOAD MASS (KG)
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

RESULTS – EDA - SQL

2015 launch records:

```
%%sql SELECT strftime('%m', Date) AS Month, strftime('%Y', Date) AS Year, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE
WHERE Landing_Outcome LIKE 'Failure%' AND strftime('%Y', Date) == '2015'
```

✓ 0.0s

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

Month	Year	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

RESULTS – EDA - SQL

Rank success count between 2010-06-04 and 2017-03-20:

```
%%sql SELECT Date, Landing_Outcome AS 'COUNT LANDING OUTCOMES BETWEEN 2010/06/04 AND 2017/03/20' FROM SPACEXTABLE
WHERE Landing_Outcome == 'Failure (drone ship)' OR Landing_Outcome == 'Success (ground pad)'
ORDER BY 'Landing_Outcome' DESC
```

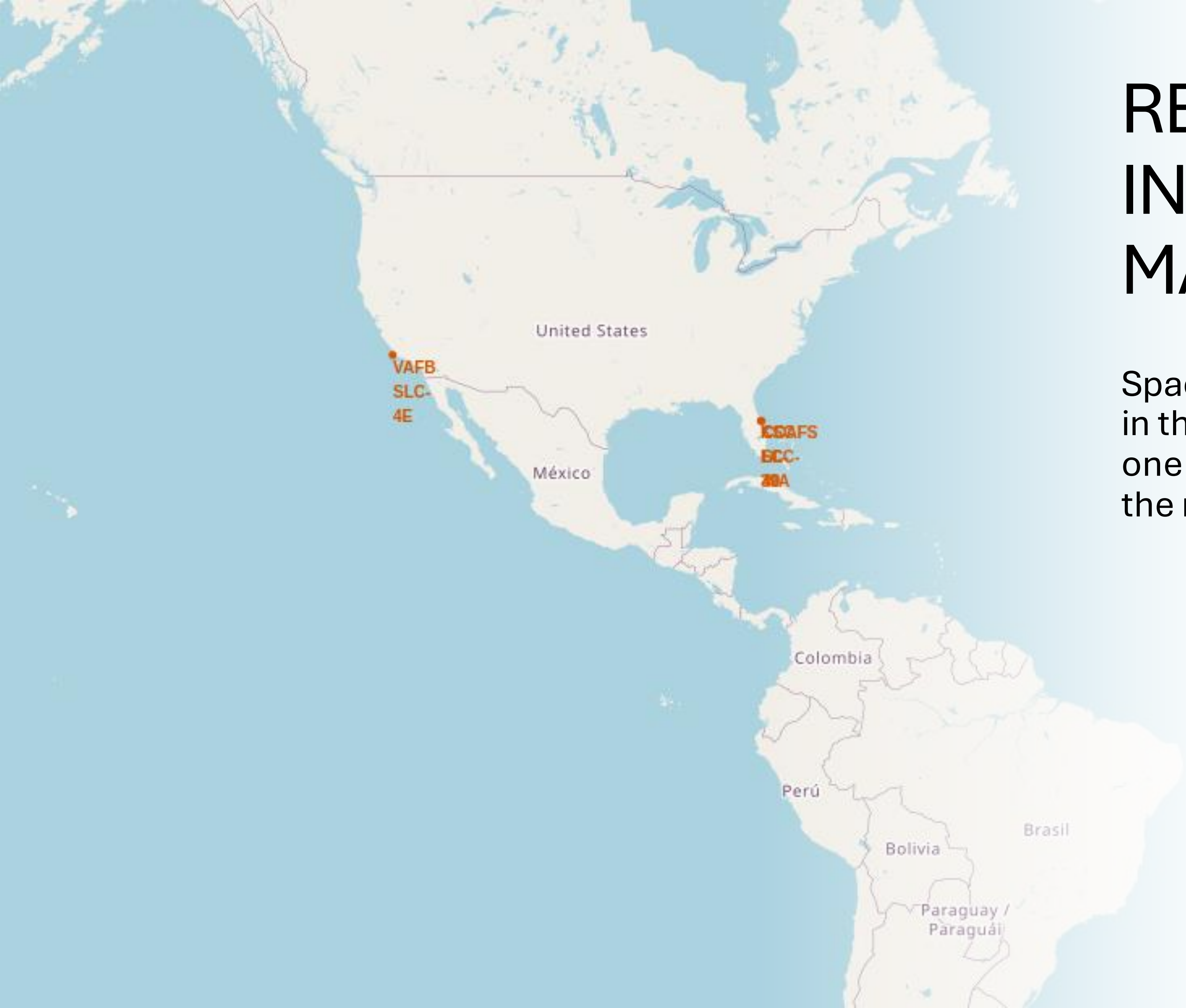
✓ 0.0s

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

Date	COUNT LANDING OUTCOMES BETWEEN 2010/06/04 AND 2017/03/20
2015-01-10	Failure (drone ship)
2015-04-14	Failure (drone ship)
2015-12-22	Success (ground pad)
2016-01-17	Failure (drone ship)
2016-03-04	Failure (drone ship)
2016-06-15	Failure (drone ship)
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)

Section 4 - Results

Launch Sites visual analysis

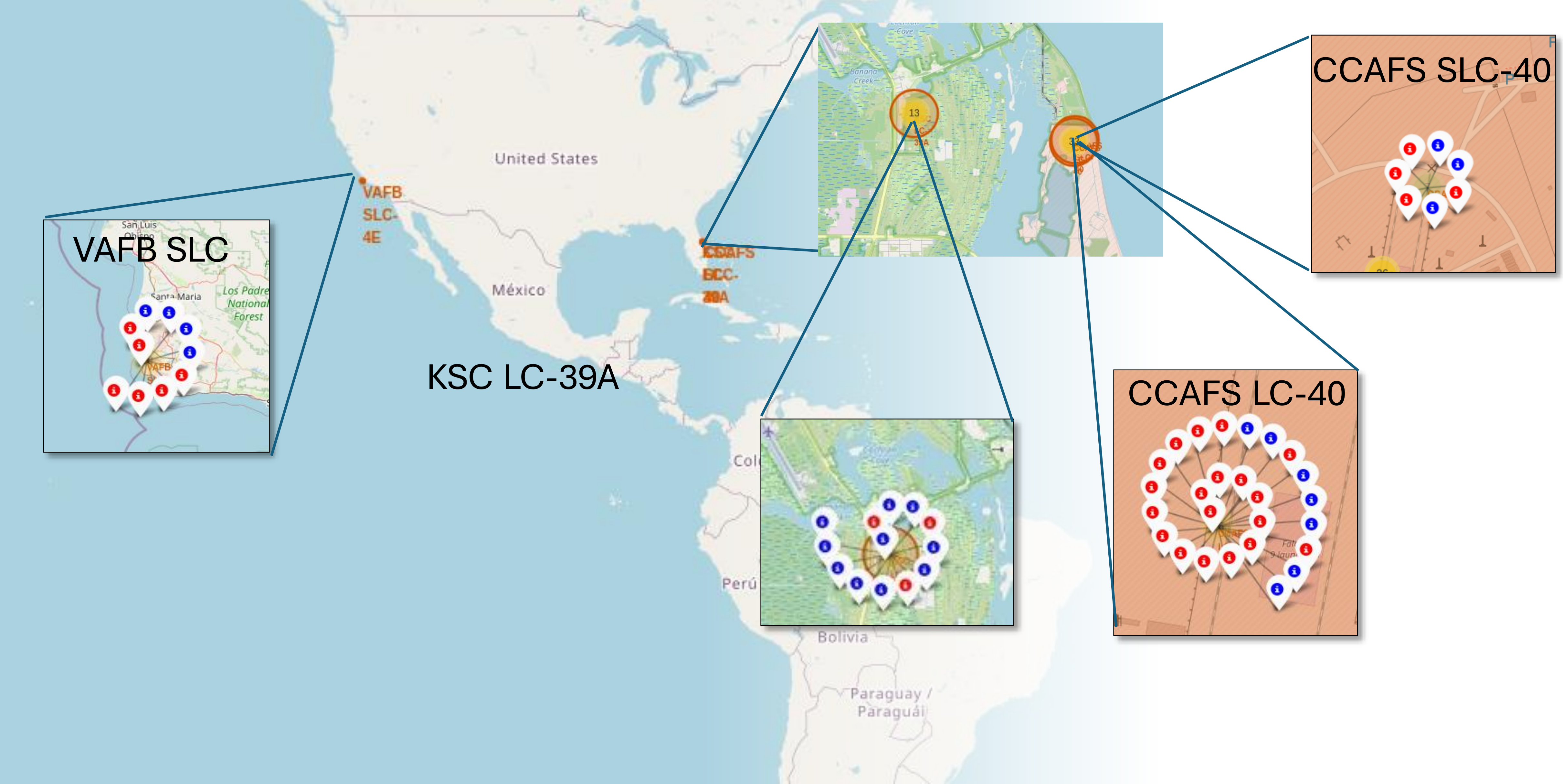


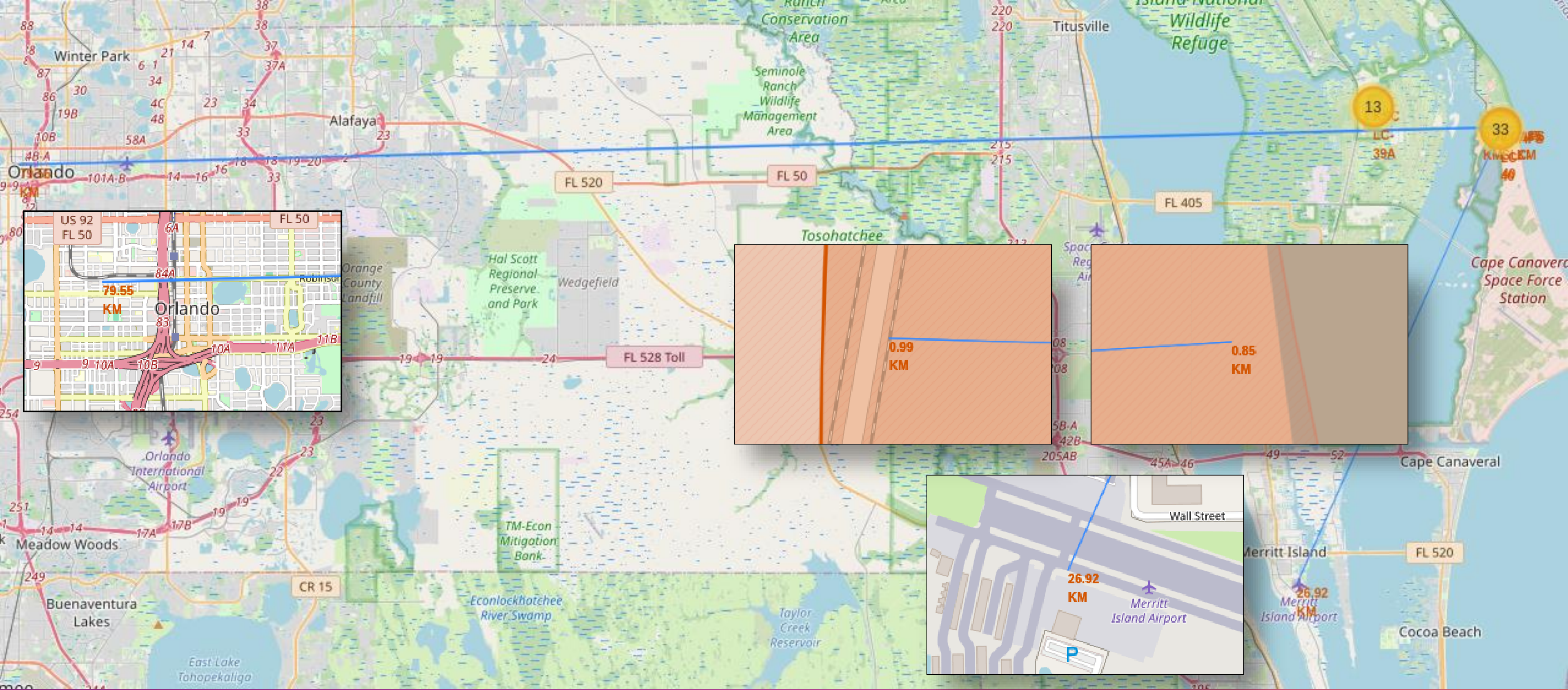
RESULTS – INTERACTIVE MAP

SpaceX launch sites are located in the United States of America, one in the coast of California and the rest near the coast of Florida

RESULTS – INTERACTIVE MAP

All Launch Records Per Site On The Map





RESULTS – INTERACTIVE MAP

CCAFS SLC-40 proximity analysis

Distance from Orlando: 79 Km

Distance from Airport: 27 Km

Distance from railroad: 0.99 Km

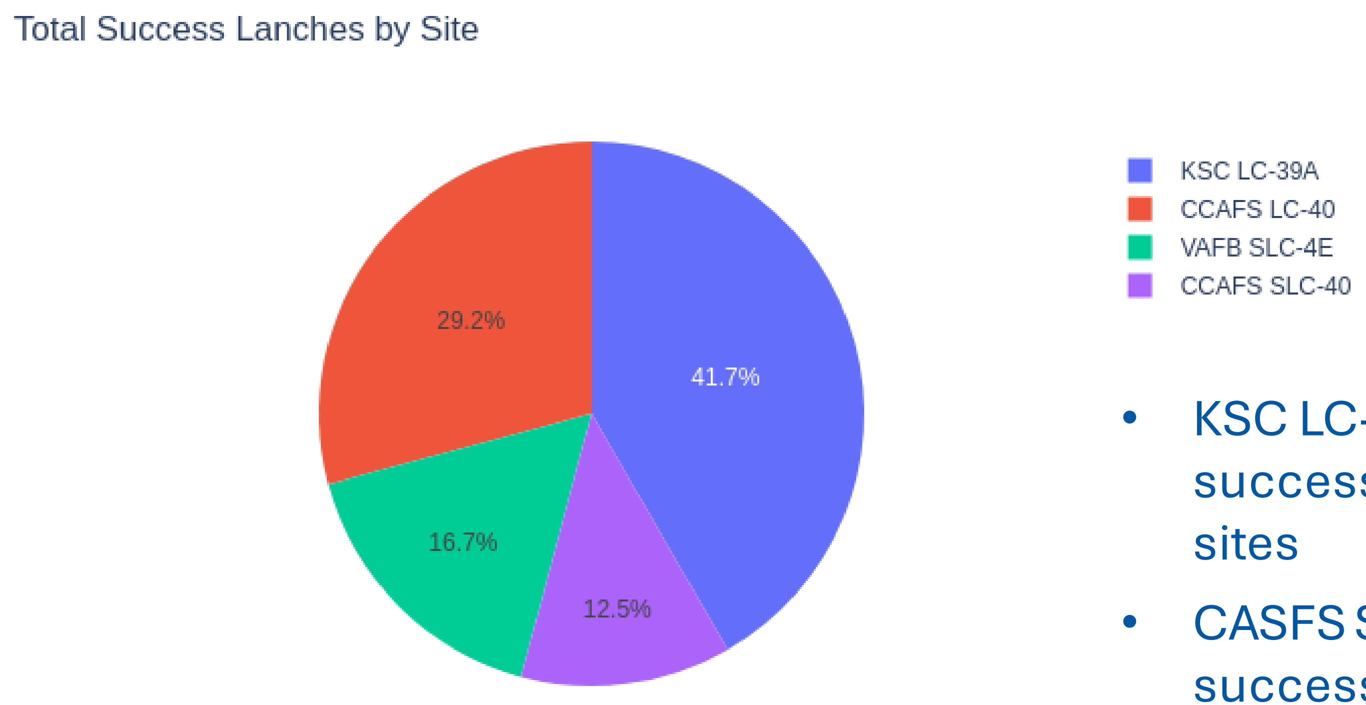
Distance from coast line: 0.85 Km

Section 5 - Results

Dashboard with Plotly Dash

RESULTS – INTERACTIVE MAP

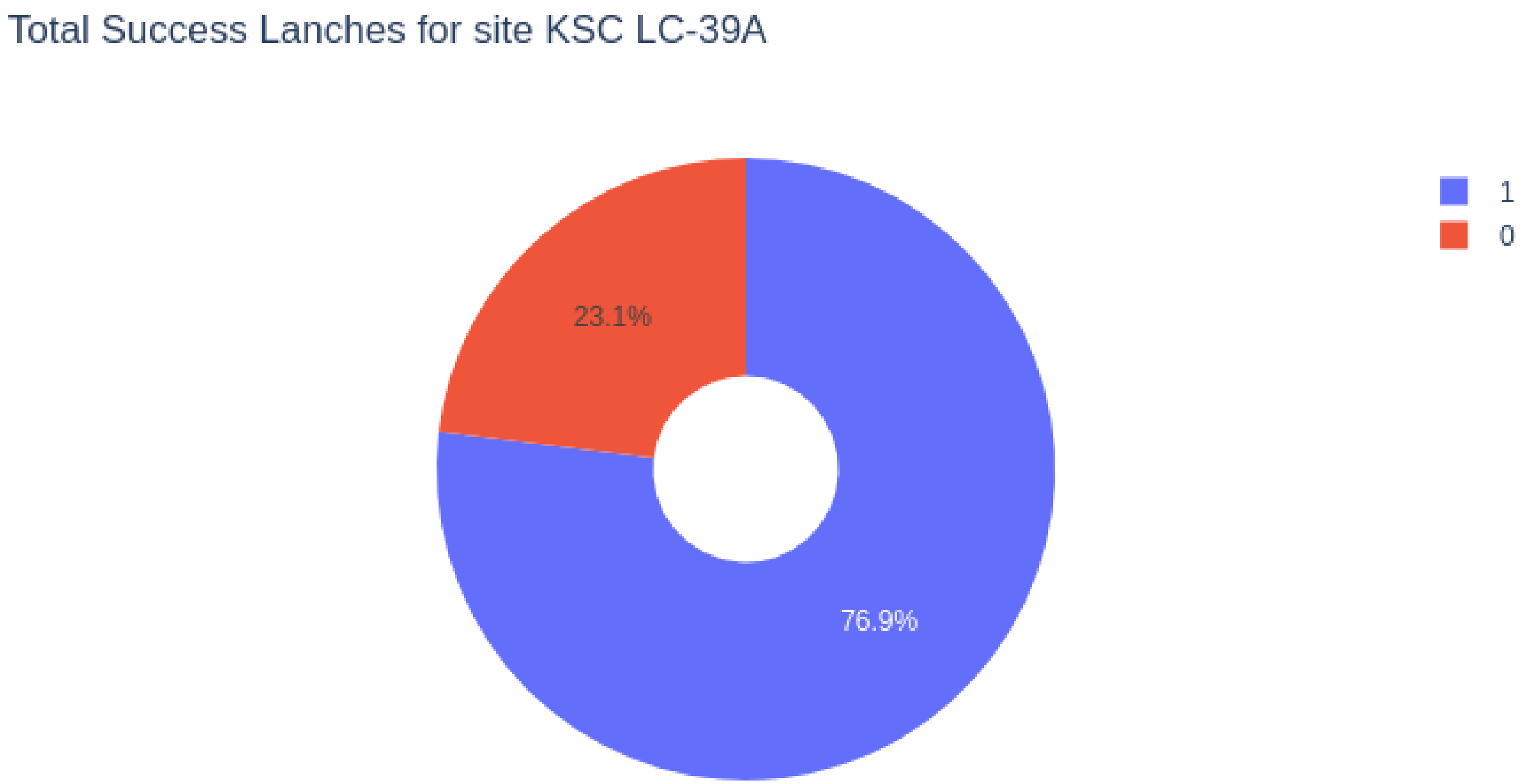
Launch success count for all sites



- KSC LC-39A holds the most successful launches from all sites
- CASFS SLC-40 has the least successful launch rate from all sites

RESULTS – INTERACTIVE MAP

Launch site with the highest launch success ratio

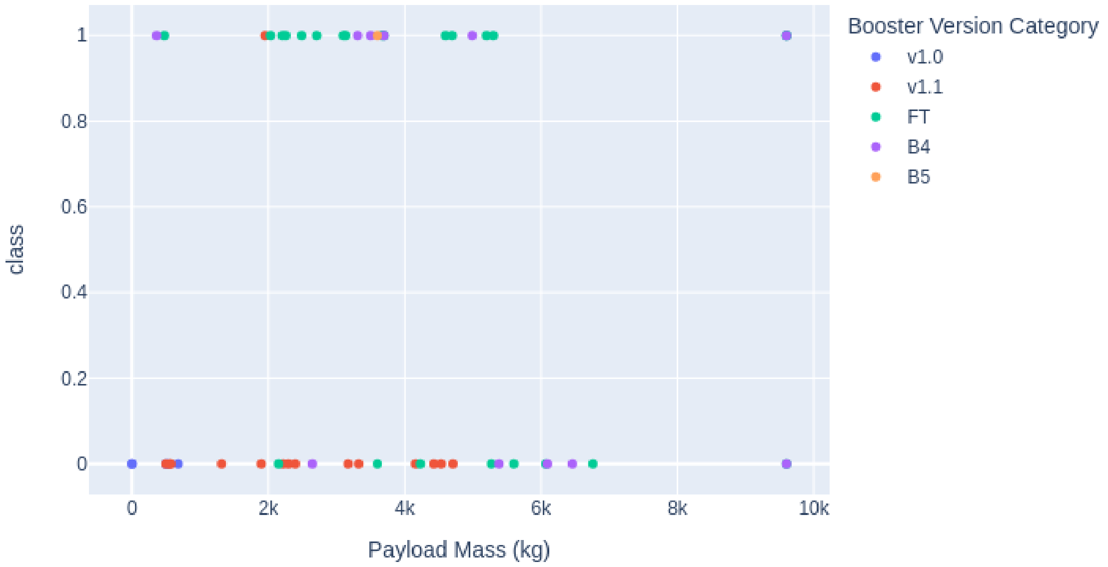


KSC LC-39A has a 76.9% success rate

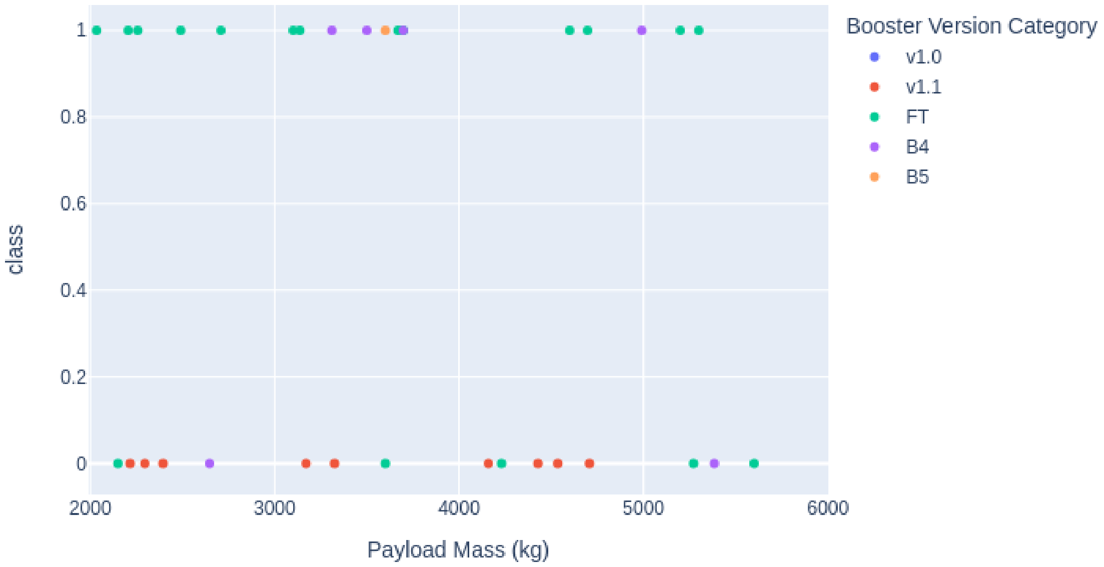
RESULTS – INTERACTIVE MAP

Payload vs. Launch outcome scatter plot for all sites

Payload from 0 to 10k



Payload from 2k to 6k



- FT booster version has a high success rate for loads up to 4000 Kg
- Booster v1.1 has a high failure rate for all payload range

Section 6

Predictive Analysis (Classification)

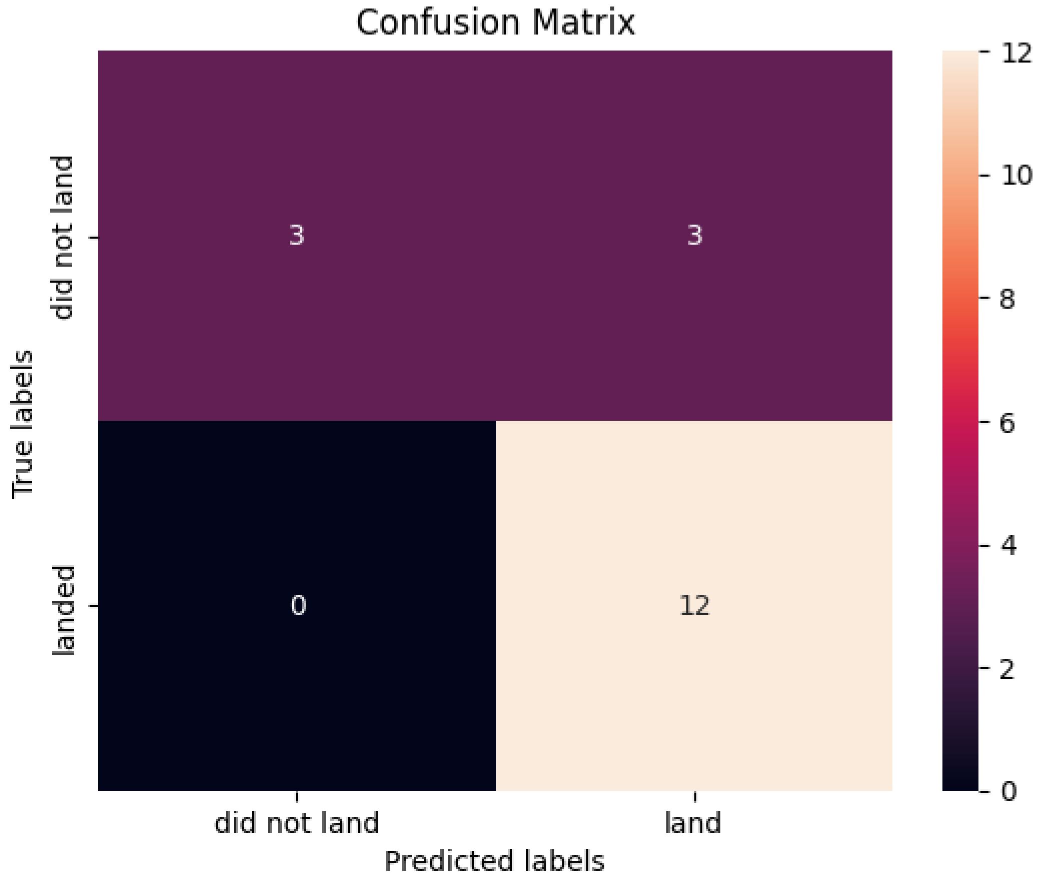
RESULTS – PREDICTIVE ANALYSIS

	Method	Score	Accuracy
0	LogReg	0.8333	0.8464
1	SVM	0.8333	0.8482
2	Decision Tree	0.8333	0.8893
3	KNN	0.8333	0.8482

- All methods scores equally in the mean accuracy on the given test data and labels.
- Decision Tree method has a better Accuracy for the given test data and labels.

RESULTS – PREDICTIVE ANALYSIS

Confusion Matrix is a specific table layout that allows visualization of the performance of an algorithm. The fields allows us to compute accuracy of the model



In this case, we have 100% of true positive and 0% false negative for 'did' not land'.
We have 20% (3 out of 15) false negative and 80% (12 out of 15) true positive for 'land'

CONCLUSION

- From EDA with data visualization one can observe:
  - The correct combination of Lauch site, Orbit and Payload result can lead in a high success rate.
  - With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- From EDA with SQL
  - There was a hiatus of 5 years, (2010 – 2015) until the fully successful mission.
- From interactive visual analytics
  - KSC LC-39A holds the most successful launches from all sites
  - CASFS SLC-40 has the least successful launch rate from all sites
  - FT booster version has a high success rate for loads up to 4000 Kg
  - Booster v1.1 has a high failure rate for all payload range
- From classification model results
  - All methods scores equally in the mean accuracy on the given test data and labels.
  - Decision Tree method has a better Accuracy for the given test data and labels, being the best ML model for analysis using the corresponding parameters and criteria