IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sergey Dayneko
February 12th 2023

Executive Summary

The project goal is to predict the likelihood of the first stage landing and estimate how probable the official Space X statement is about 62 million dollars of a launch cost. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

Data for analysis was collected using SpaceX Rest API and Web Scraping from Wikipedia. Launch sites locations were visualized on a map with Folium library. This allowed to evaluate the distances between launch sites to proximities. Dashboard built with Plotly Dash made it possible to interactively analyze the landing success rate statistics by launch sites. Predictive Analysis has been implemented as a Classification model. Evaluation metrics calculated for the full sample revealed that the best classifier is the Decision Tree model.

Below are the key conclusions built based on data analysis performed with application of Data Science methodology:

- The following factors influence first stage landing success rate: **Launch site, Payload**, and **Booster version**.
- The following pair of predictors has a correlation that determine the success rate of the first stage landing: Payload - Booster version.
- To ensure the best first stage landing success rate, the following operating conditions are recommended:
  - use KSC LC 39A as main launch site,
  - use Booster version F9 FT for payloads between 2000 and 5500 kg and Booster version F9 B5 for payloads over 7000 kg.
  - tend to use payloads from two ranges (First: between 2000 and 5500 kg; Second: between 9000 and 15500 kg) as well-tested and proven high landing success rates.

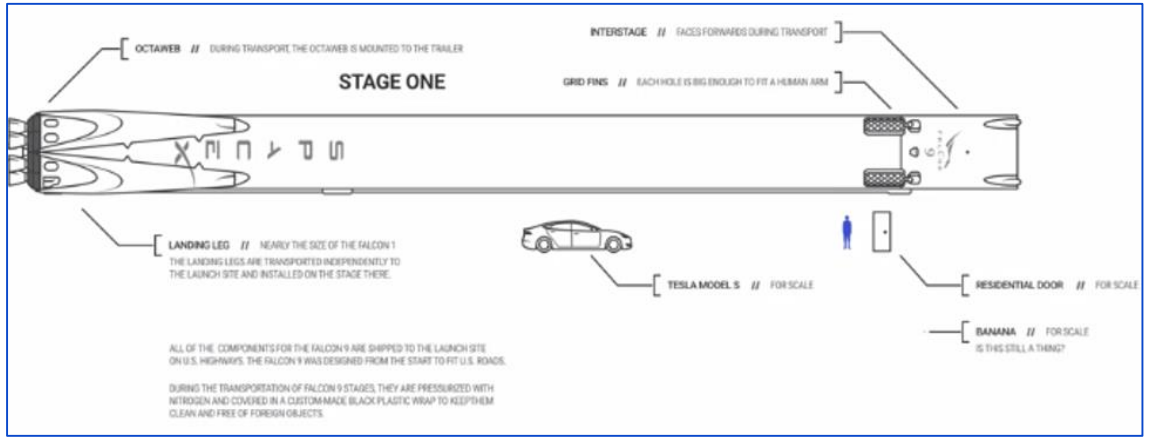
Outline

- Introduction
- Methodology
- Results
- Conclusion

Introduction

- Project background and context

  Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, we need to apply Data Science methodology to analyze publically available data on Falcon 9 rocket launches to predict the likelihood of the first stage landing and estimate how probable the official Space X statement is about 62 million dollars of a launch cost.



SpaceX's Falcon 9 First stage is shown above. This stage is quite large and expensive, and does most of the work of getting payloads into orbit.

Source: Forest Katsch, zlsadesign.com

- Questions to be answered
  - What factors influence first stage landing success rate?
  - Identify interaction amongst various predictors (independent variables) that determine the success rate of the first stage landing.
  - What operating conditions need to be in place to ensure the best first stage landing success rate?

Section 1

Methodology

Methodology

- Data collection methodology:
  - Data was collected using SpaceX Rest API and Web Scraping from Wikipedia
- Data wrangling:
  - Data preprocessing included filtering and formatting, handling missing values, normalization and one-hot encoding
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models with the following evaluation and tuning to obtain optimal results

Data Collection

In order to get all necessary data on the first stage landing statistics we utilized two sources: SpaceX REST API and SpaceX's Wikipedia.

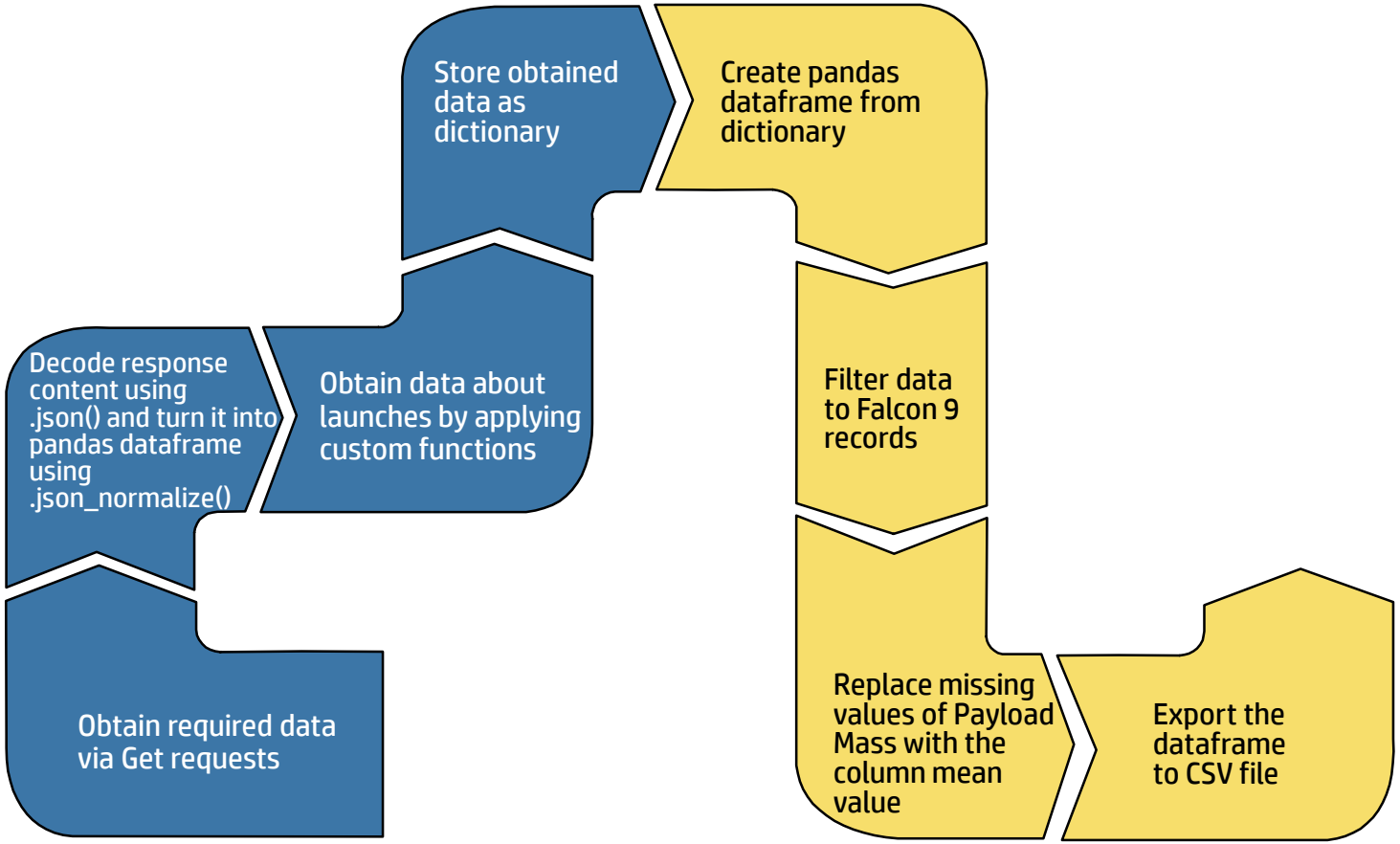
- **Data obtained via SpaceX REST API:**

Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude

- **Data obtained via Wikipedia Web Scraping:**

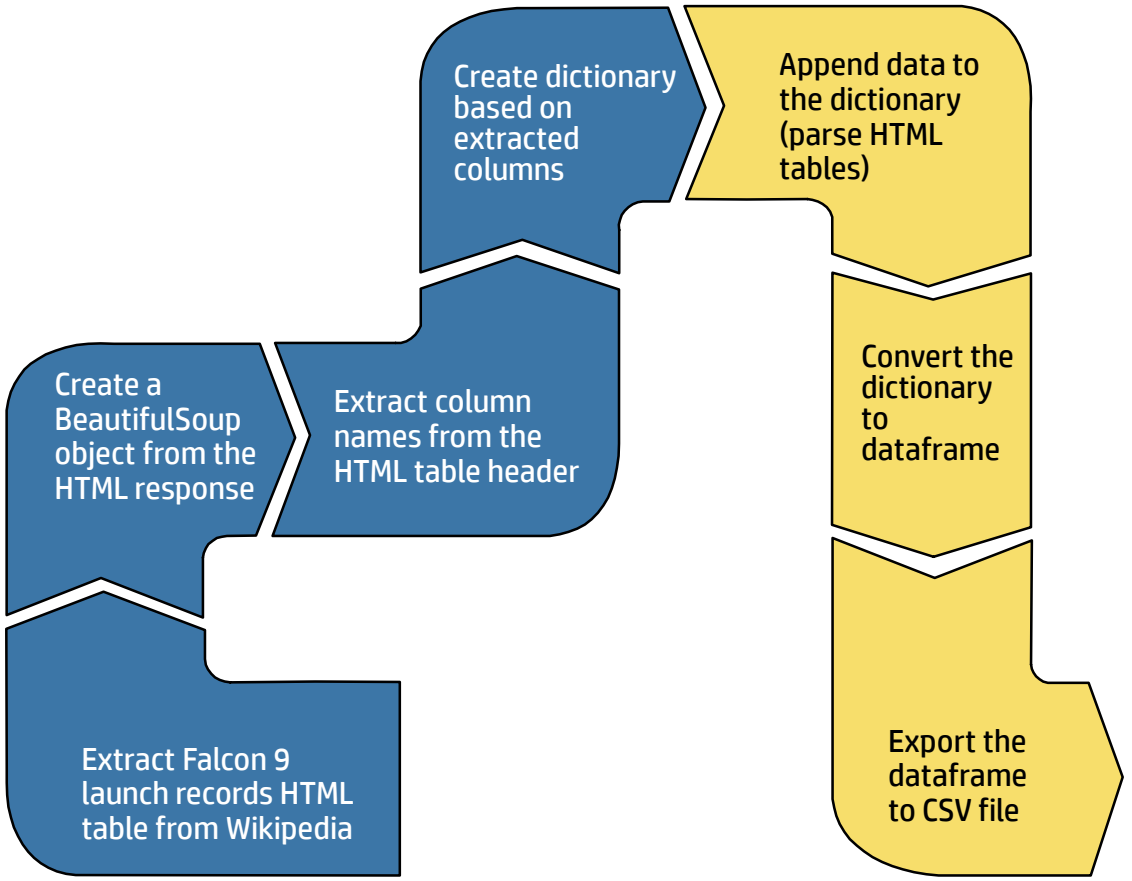
Flight No., Launch site, Payload, Payload Mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



GitHub URL

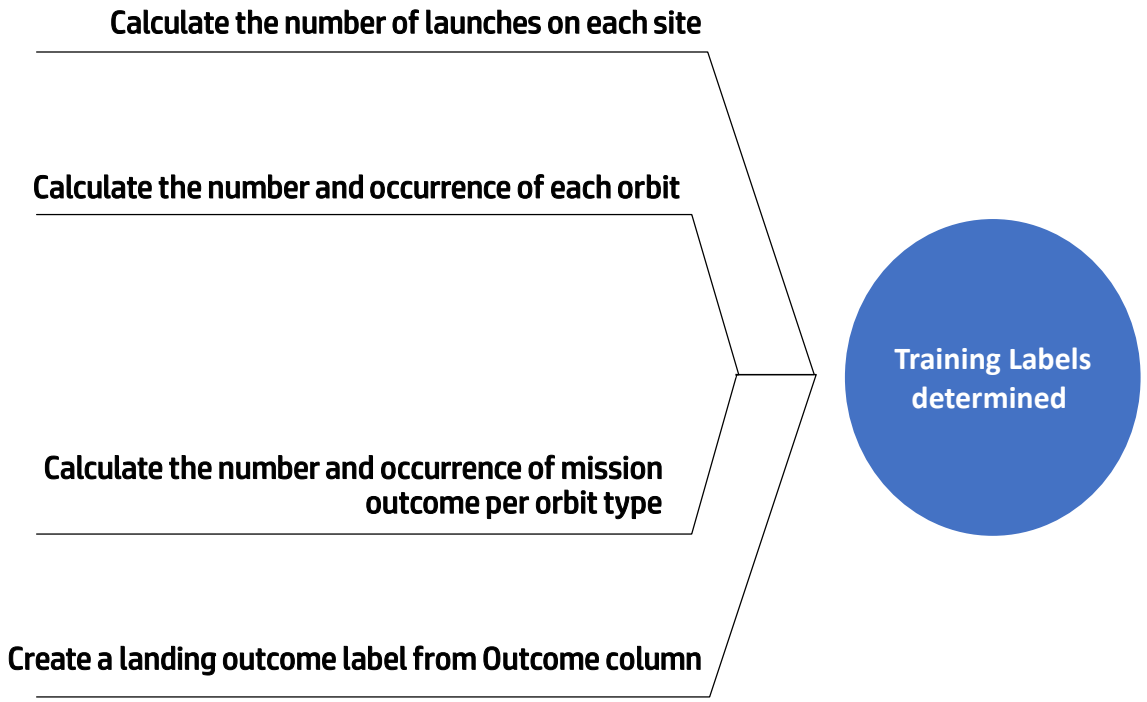
Data Collection – Web Scraping



GitHub URL

Data Wrangling

- We performed Exploratory Data Analysis to find some patterns in the data and determine what would be the label for training supervised models.
- Though in the data set there are several outcomes describing the status of the first stage landing, we converted those outcomes into Training Labels where **1** meant the booster successfully landed, and **0** meant unsuccessful land.



GitHub URL

EDA with Data Visualization

- In order to evaluate the relationship between the studied variables we built the following **Scatter plots**:
  - Payload Mass vs. Flight Number
  - Launch Site vs. Flight Number
  - Launch Site vs. Payload Mass
  - Orbit Type vs. Flight Number
  - Orbit Type vs. Payload Mass
- To demonstrate how orbit type influences success rate of the first stage landing we built corresponding **Bar chart**.
- We built **Line chart** to track launch success yearly trend.

GitHub URL

EDA with SQL

Following information were obtained via SQL queries:

- Names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string ‘CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranked count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order

GitHub URL

Build an Interactive Map with Folium

To visualize launch sites locations on the map:

- Using corresponding latitude and longitude coordinates we added markers with Circle, Popup Label and Text Label for all launch sites of our dataset.

To identify which sites have relatively high success rates:

- According to the ‘class’ column of our dataset we added green markers (for success) and red markers (for fails) using Marker Cluster to identify which launch sites have relatively high success rates.

To evaluate the distances between launch site (CCAFS SLC-40) to its proximities:

- We drew lines to show distances between the launch site and its proximities: railway, highway, coastline and city.

GitHub URL

Build a Dashboard with Plotly Dash

To analyze landing success rate statistics:

- We created a dropdown list to enable launch site selection
- Plotted a Pie chart representing a share of success launches per launch site

To analyze landing success rate per launch site:

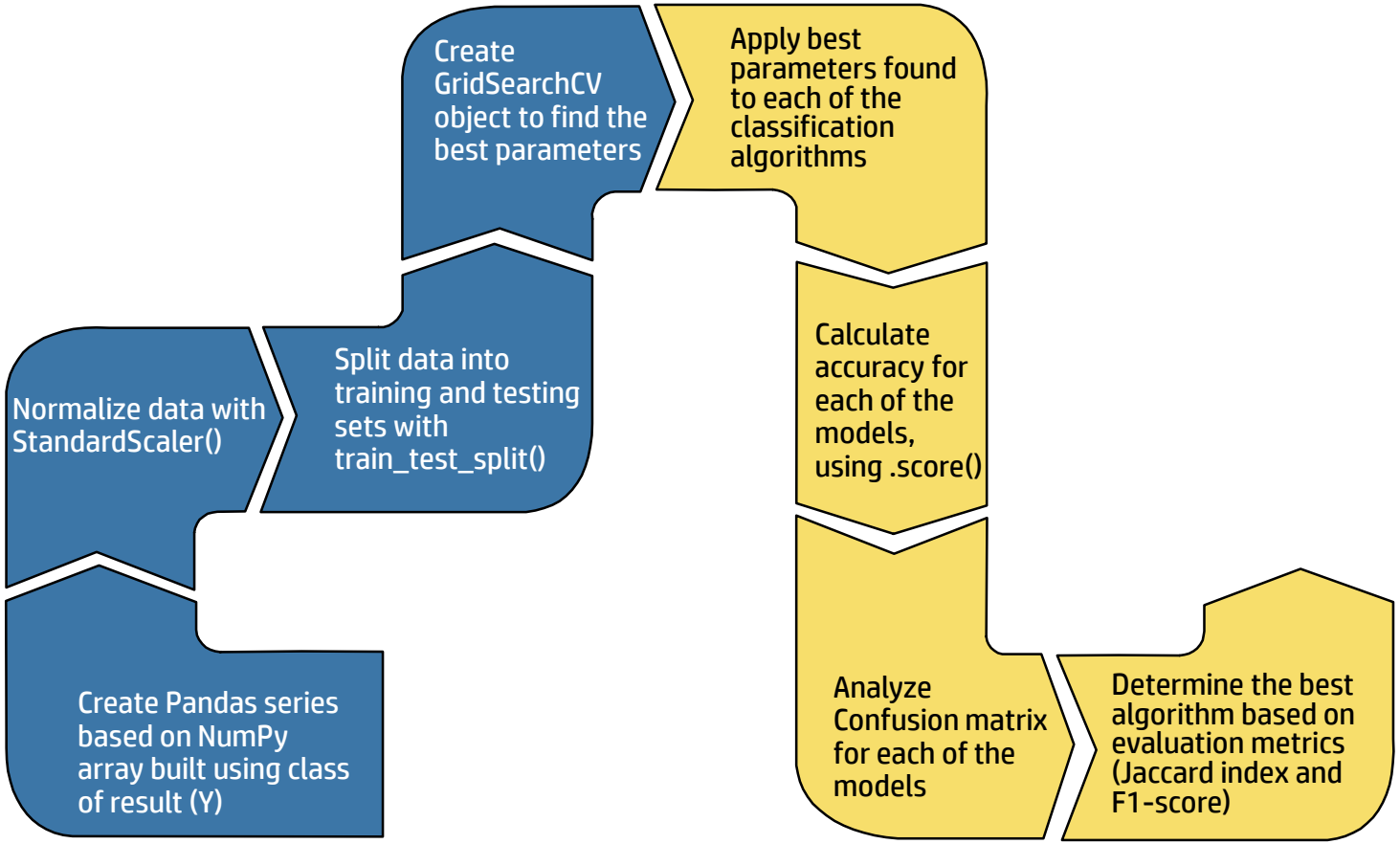
- We created Pie chart representing a share of success vs. failed launches for a selected launch site

To analyze correlation between Payload mass and launch outcome for various Booster versions:

- We created a slider for Payload mass range selection
- Plotted a Scatter plot of launch outcome (class) vs. payload mass, where Booster versions were color coded

GitHub URL

Predictive Analysis (Classification)

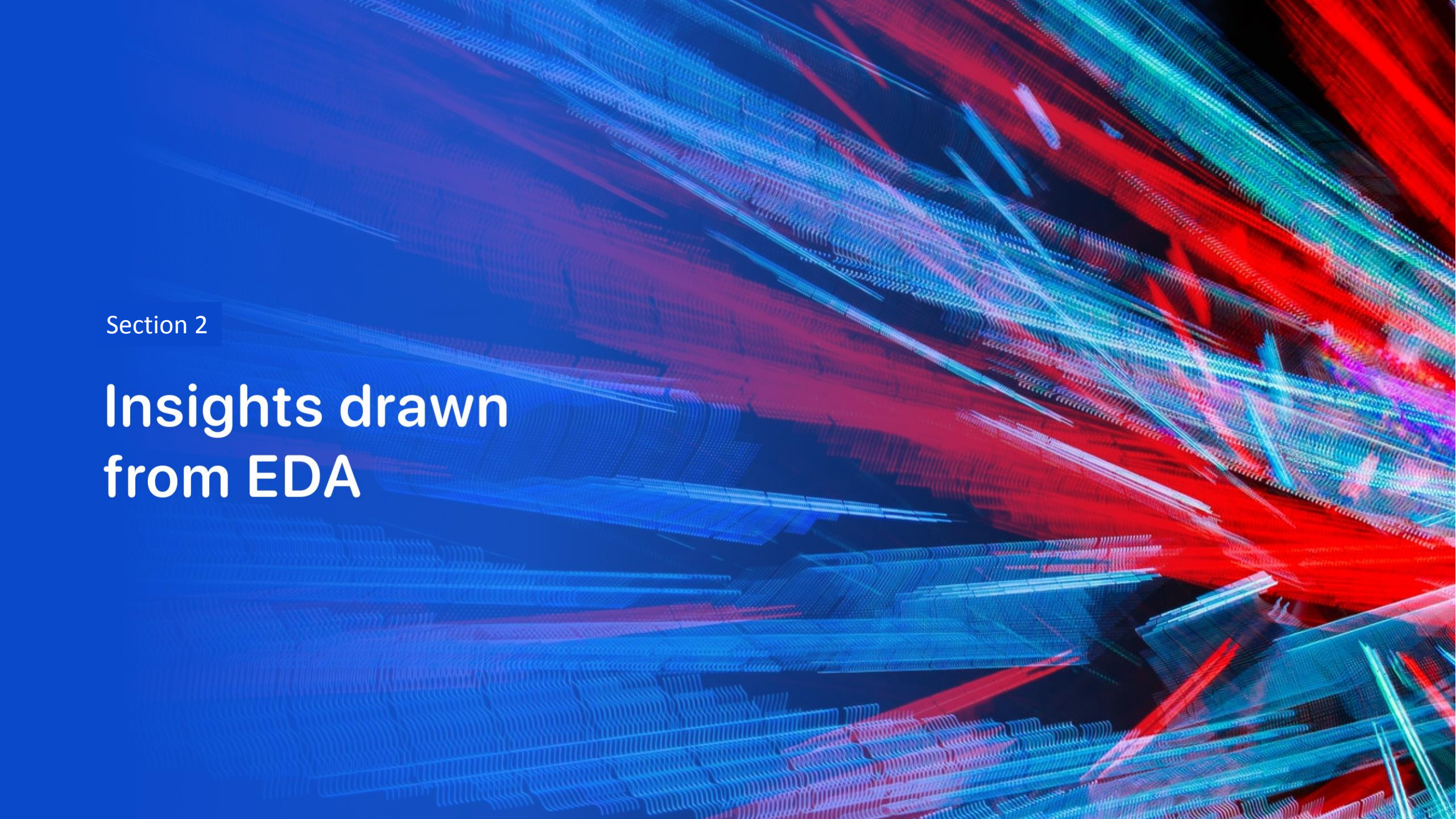


GitHub URL

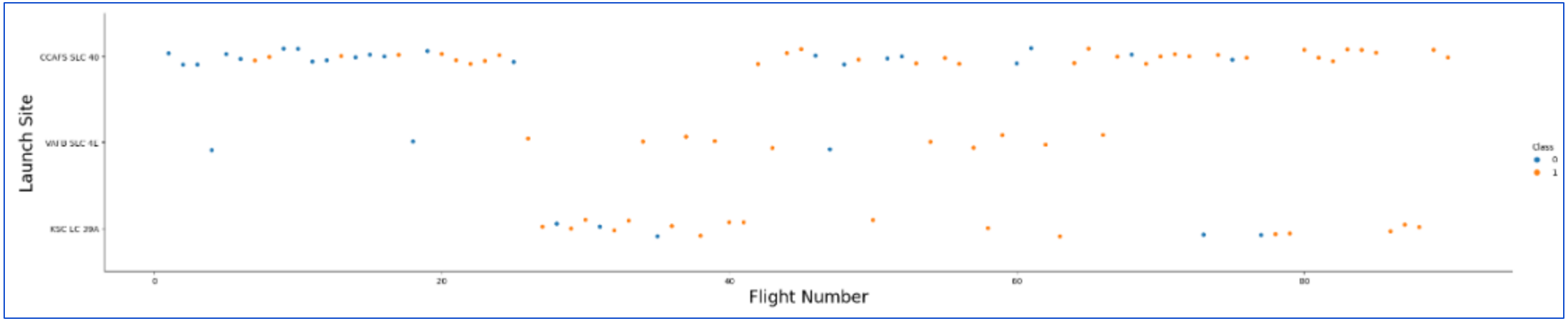
Results

The following sections contain results of the conducted analysis accompanied by charts and graphs with interpretations.

- Exploratory data analysis:
    - Data Visualization results
    - SQL results
- Interactive Map
- Dashboard
- Predictive Analysis


Section 2
Insights drawn from EDA

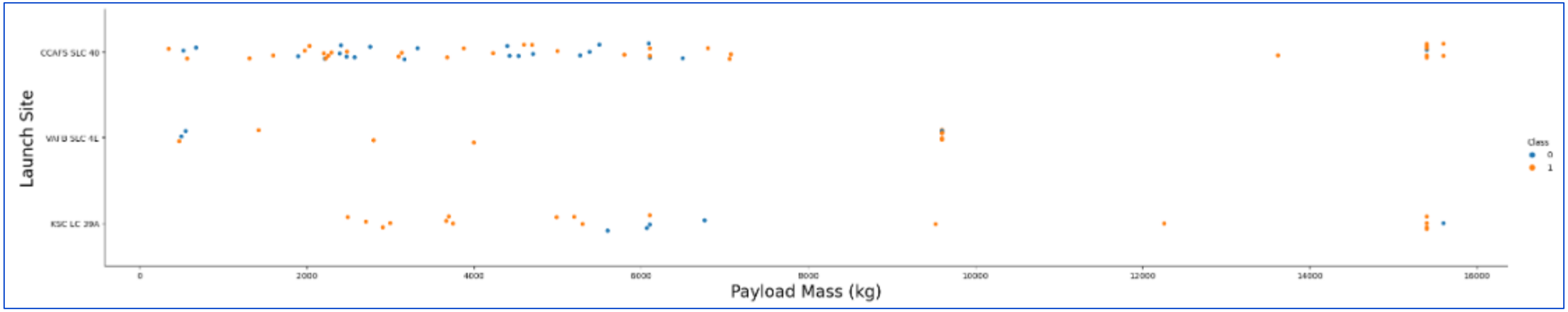
Flight Number vs. Launch Site



Interpretation:

1. General trend: landing success rate grow with the number of launches increase.
2. Launch site CCAFS SLC 40 is supposed to be the main since the vast majority of launches were made from it during the observation period.
3. Launch site KSC LC 39A was the last to put into operation. For some reason, all launches from Launch site CCAFS SLC 40 were transferred to KSC LC 39A between (approximately) 26 and 43 flights.
4. Among the three launch sites, VAFB SLC 4E considered as the least used. Starting from 66 flight (approximately), launches were stopped on it for some reason.

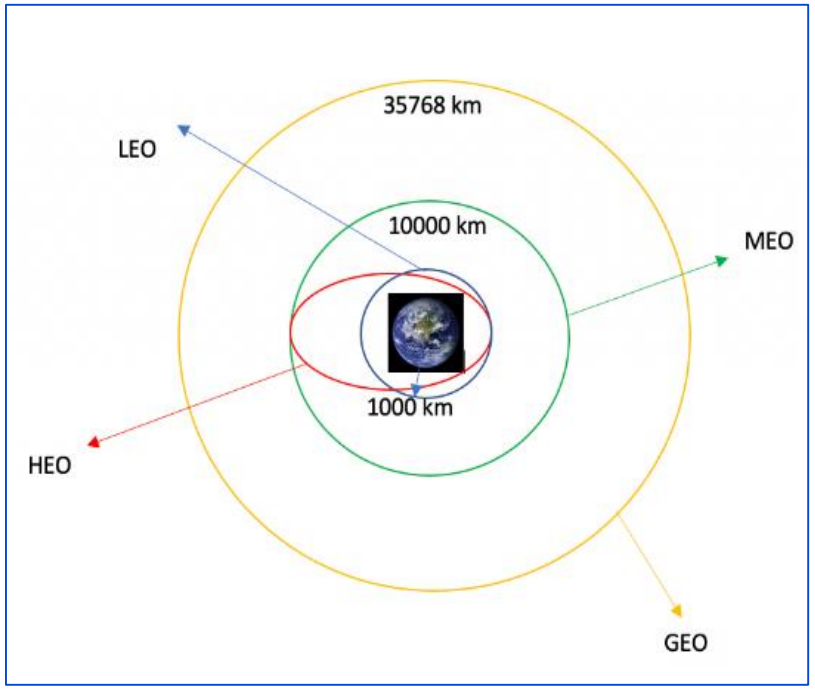
Payload vs. Launch Site

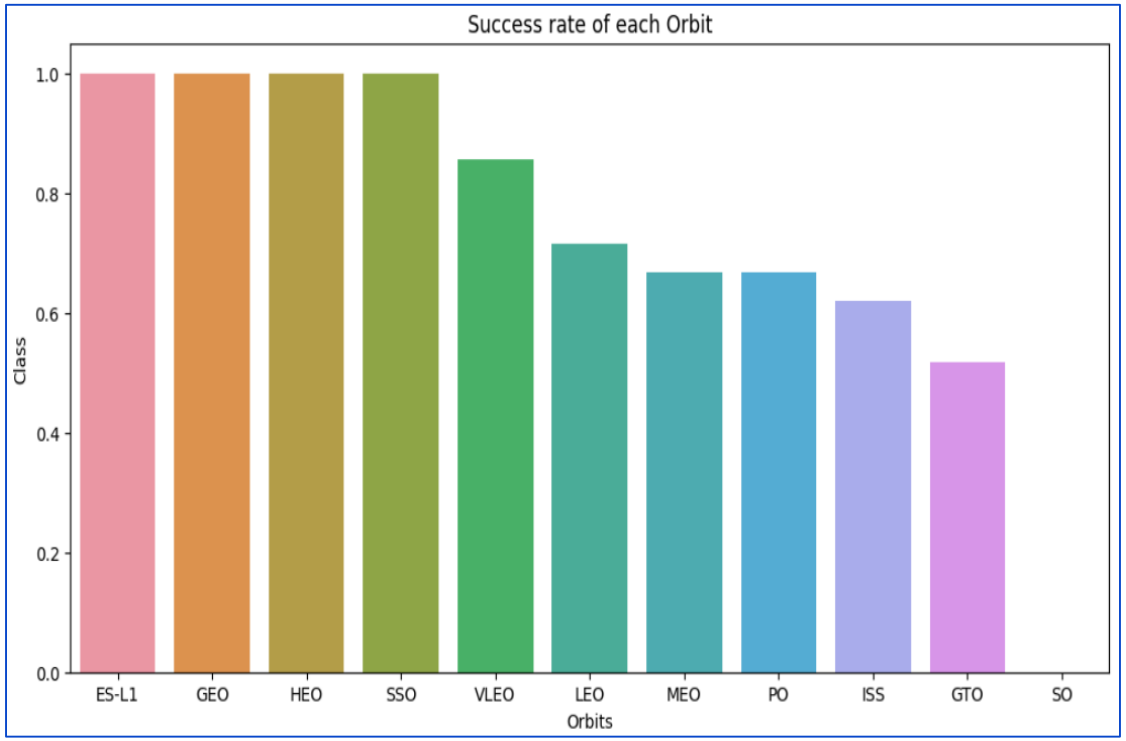


Interpretation:

1. General trend: landing success rate is higher for launches with Payload over 7000 kg.
2. Launch site KSC LC 39A has 100% landing success rate for payload mass under 5500 kg and between 9000 and 15500 kg.
3. For CCAFS SLC 40 and KSC LC 39A launch sites the most frequent Payload mass is in the ranges: 1) under 7000 kg 2) between 15000 and 16000 kg. No launches with Payload over 10,000 kg were made at VAFB SLC 4E launch site.

Success Rate vs. Orbit Type





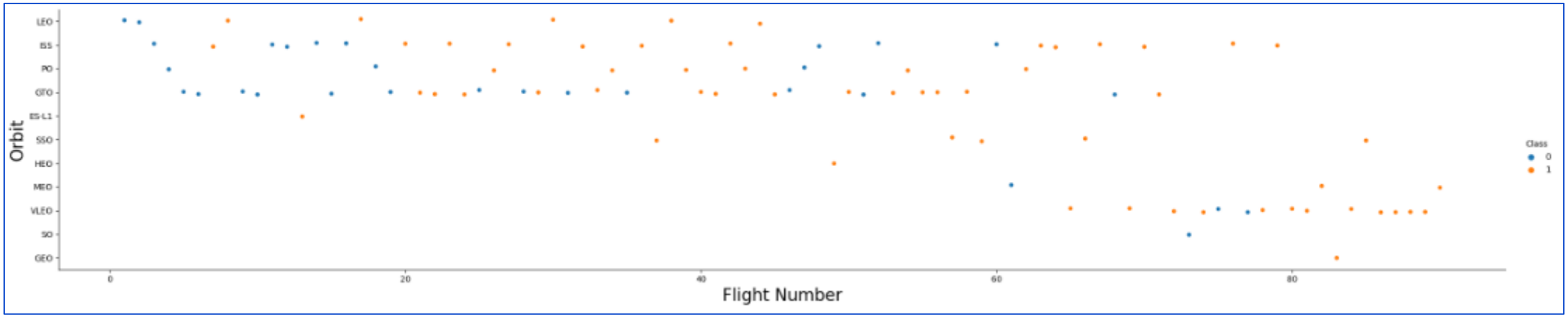
Interpretation:

1. Orbits with 100% landing success rate: **ES-L1, GEO, HEO, SSO**

2. Orbits with landing success rate between 85% and 50%: **VLEO, LEO, MEO, PO, ISS, GTO**

3. Orbit with no landing success rate: **SO**

Flight Number vs. Orbit Type

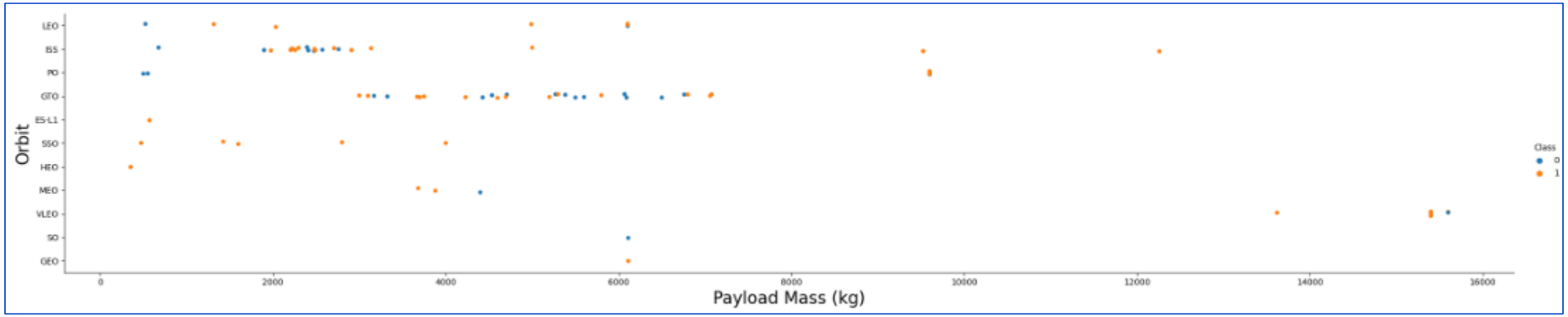


Interpretation:

1. As the number of launches increased, spectrum of orbits in use began to increase, starting from the 55th flight (approximately).

2. Based on this plot, we cannot conclude that orbit is a factor that influences the first stage landing success rate.

Payload vs. Orbit Type

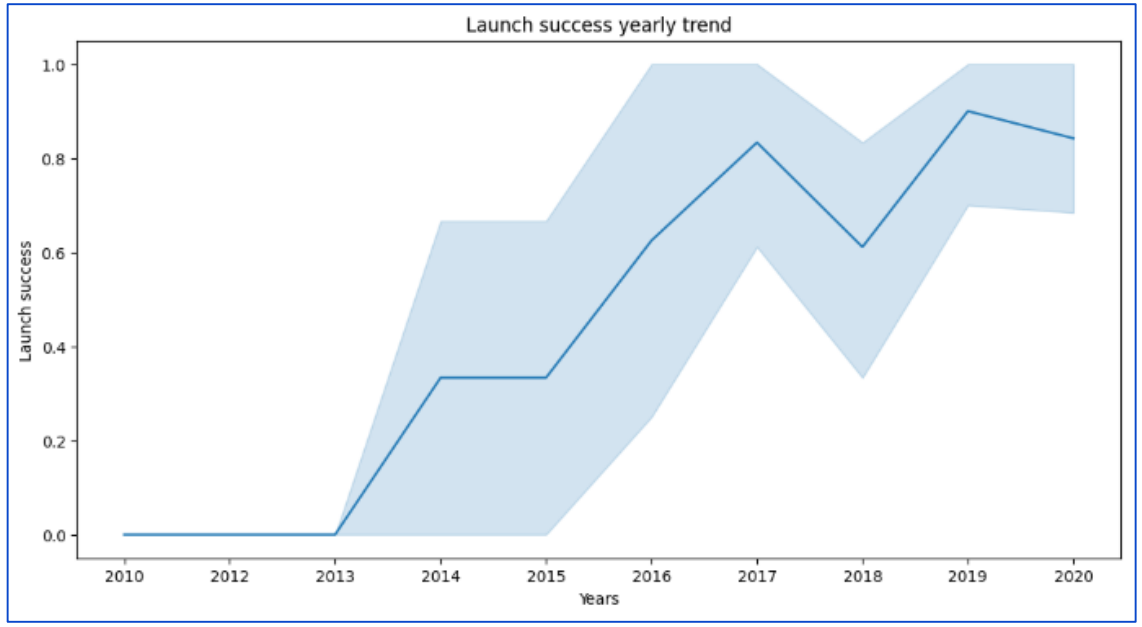


Interpretation:

1. Sufficient amount of data points for analysis is available for only two orbits: ISS for Payload between 2000 and 3000 kg, and GTO for Payload between 3000 and 7000 kg.

2. Based on this plot, landing success / failure rates are about the same for ISS and GTO orbits Payload ranges specified above.

Launch Success Yearly Trend



Interpretation:

1. Since 2013 there has been a growing trend in landing success rates, with a possible plateau in the region of 0.9 after 2020.

2. The local minimum in 2018 requires a more detailed study to discover the root causes.

All Launch Site Names

```
%sql select distinct launch_site from SPACEXTBL;
```

```
 * ibm_db_sa://pqy06220:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

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

Names of the unique launch sites in the space mission

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://pqy06220:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

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

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTBL where customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://pqy06220:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

total_payload_mass
45596

Total payload mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTBL where booster_version like '%F9 v1.1%';
```

```
 * ibm_db_sa://pqy06220:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

average_payload_mass
2534

Average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
%sql select min(date) as first_successful_landing from SPACEXTBL where landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://pqy06220:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

first_successful_landing
2015-12-22

Date when the first successful landing outcome in ground pad was achieved

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version from SPACEXTBL where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and
```

```
 * ibm_db_sa://pqy06220:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

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

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

Total Number of Successful and Failure Mission Outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXTBL group by mission_outcome;
```

```
 * ibm_db_sa://pqy06220:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

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

Total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
%sql select booster_version from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL);
```

```
 * ibm_db_sa://pqy06220:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

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

Names of the booster versions which have carried the maximum payload mass

2015 Launch Records

```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXTBL
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://pqy06220:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXTBL
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

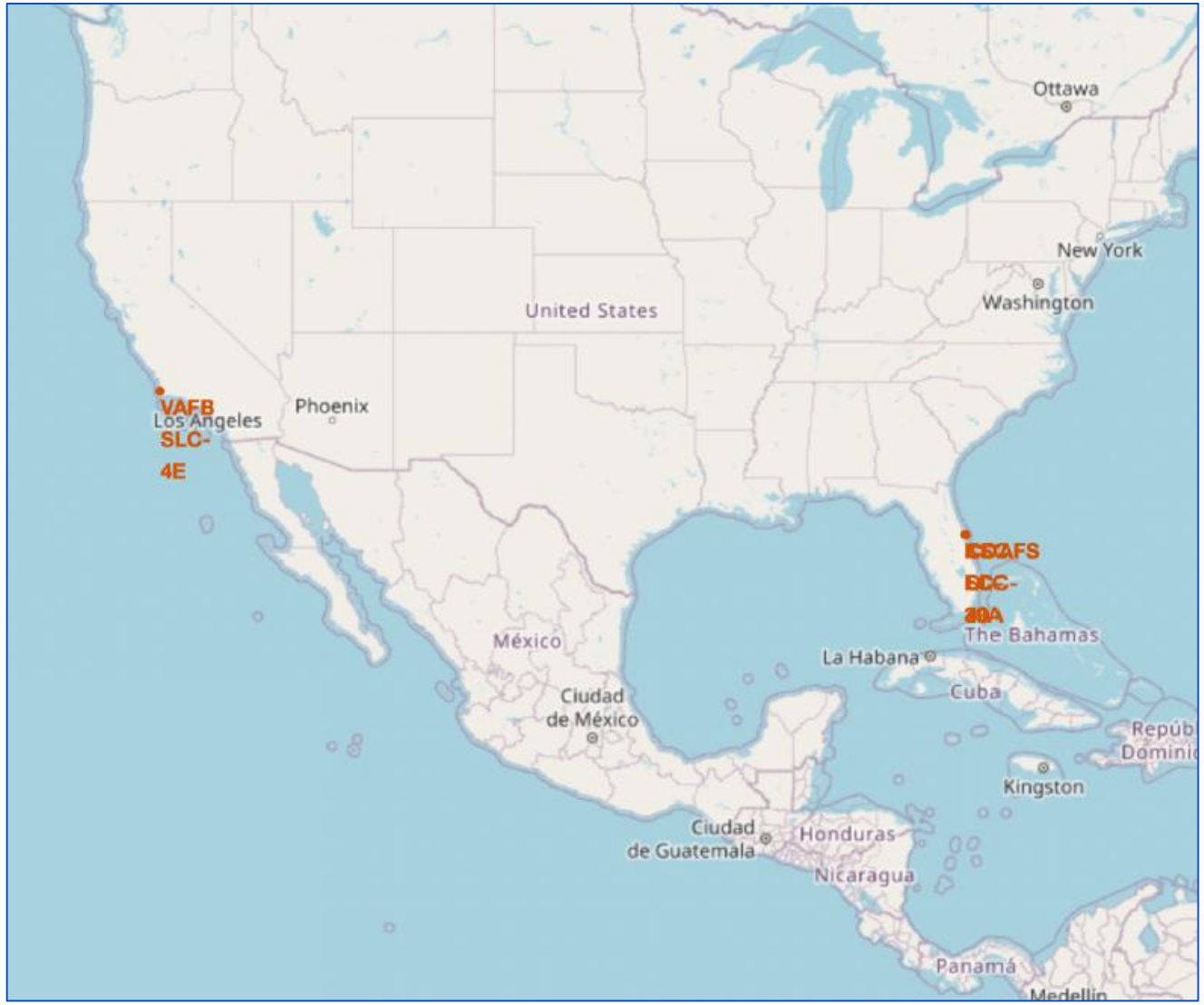
```
 * ibm_db_sa://pqy06220:***@b70af05b-76e4-4bca-a1f5-23dbb4c6a74e.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32716/BLUDB
Done.
```

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

Count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between 2010-06-04 and 2017-03-20, in descending order


Section 3
Launch Sites
Proximities Analysis

Launch sites visualization on a map



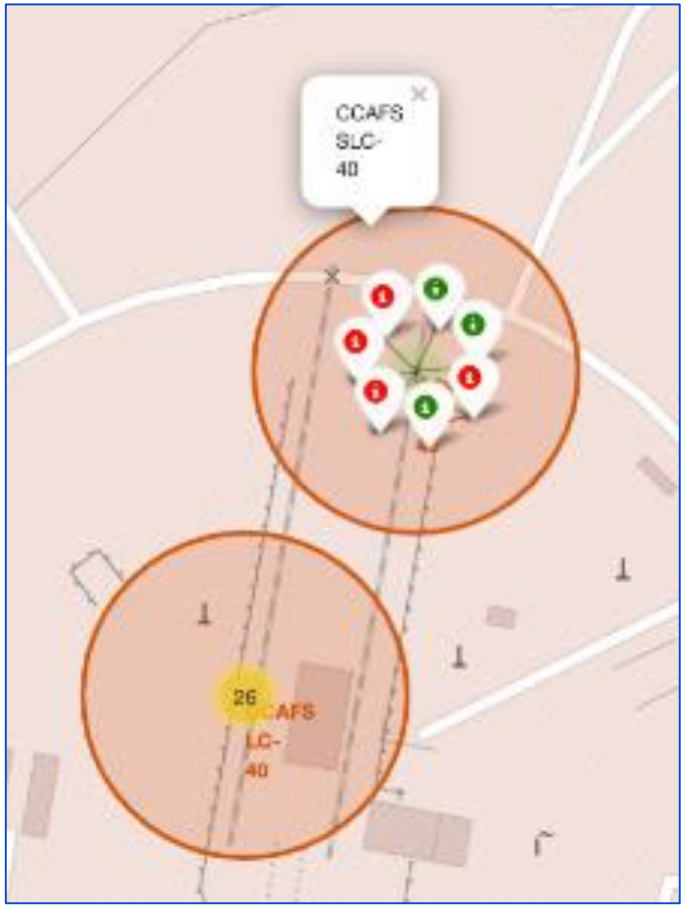
SpaceX launch sites locations mapped based on latitude and longitude coordinates

Launch sites with color markers

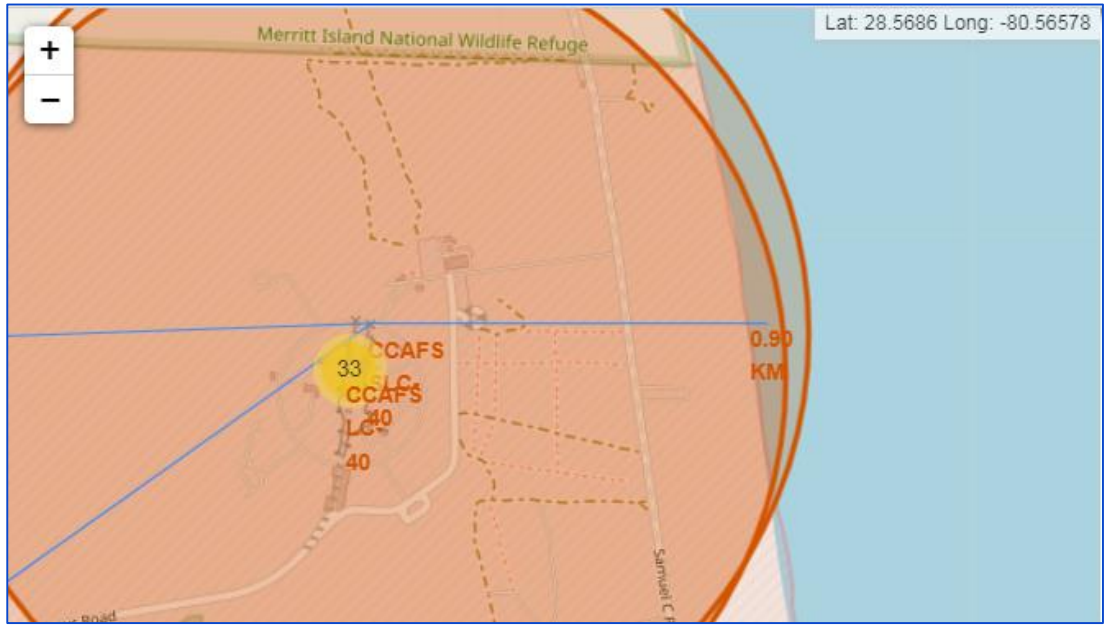
Based on data in ‘class’ column indicating if a launch was successful or not, we created Green and Red markers:

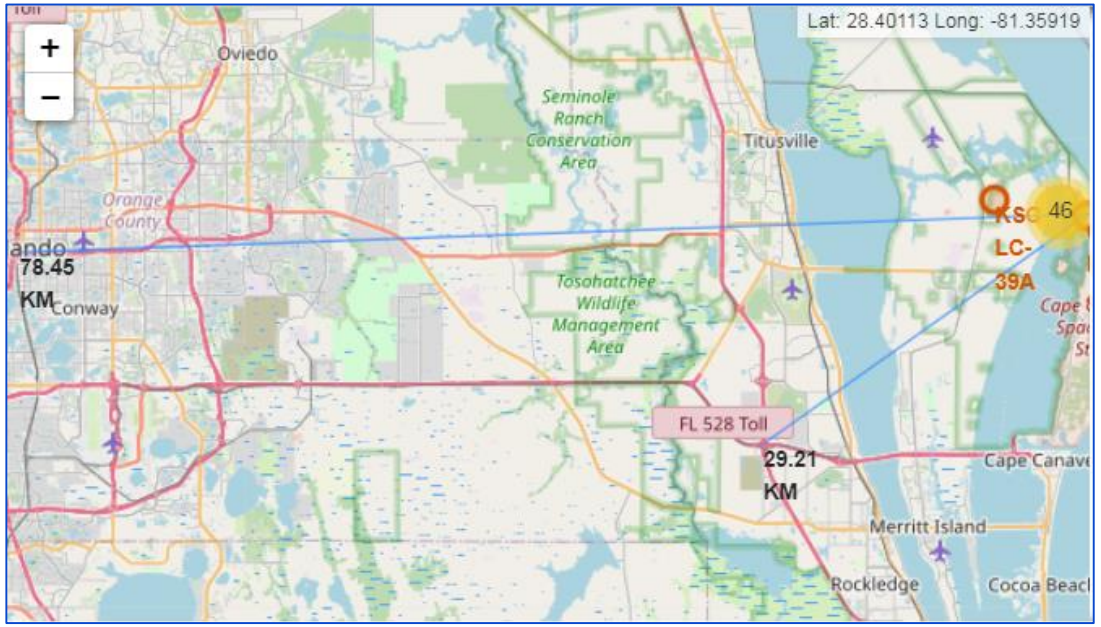
- Green marker – high success rate
- Red marker – low success rate

With the color-labeled markers it is easy to identify which launch sites have relatively high success rates.



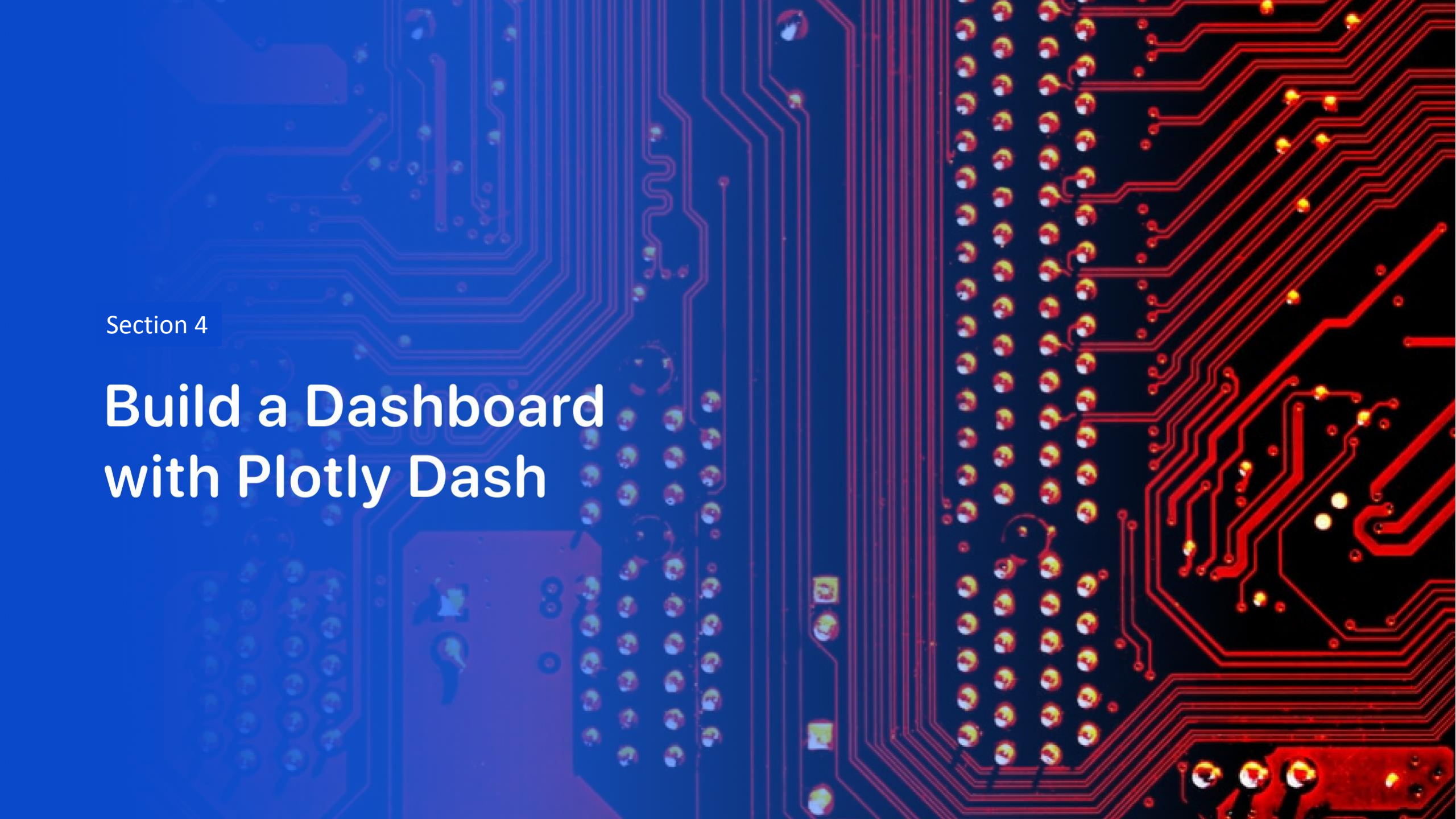
Distance from the launch site CCAFS SLC 40 to its proximities



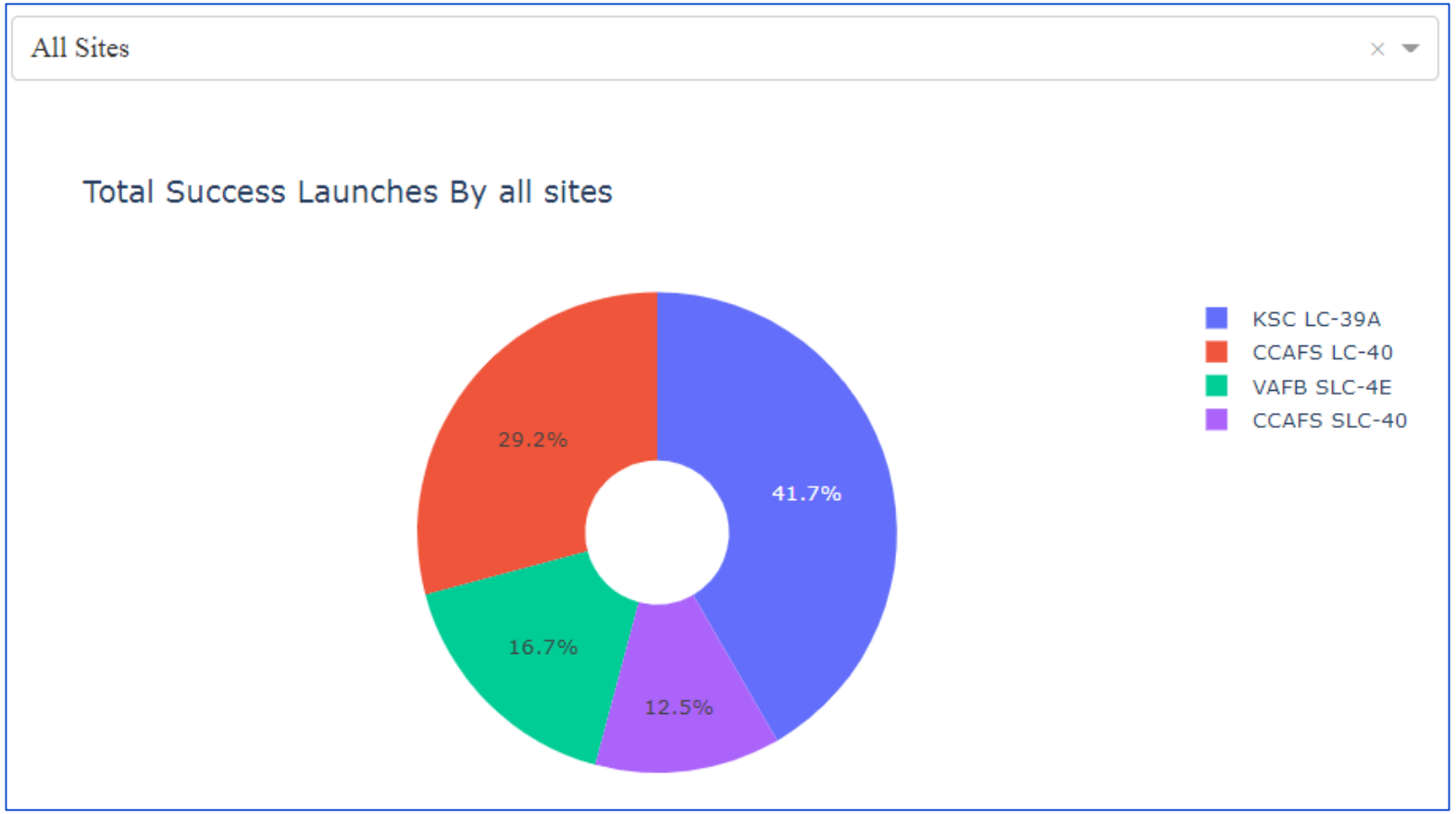


Interpretation:

- The launch site is not in close proximity to railways.
- The launch site is not in close proximity to highways.
- The launch site is not in close proximity to railways coastline.
- The launch site keeps certain distance away from cities.


Section 4
Build a Dashboard with Plotly Dash

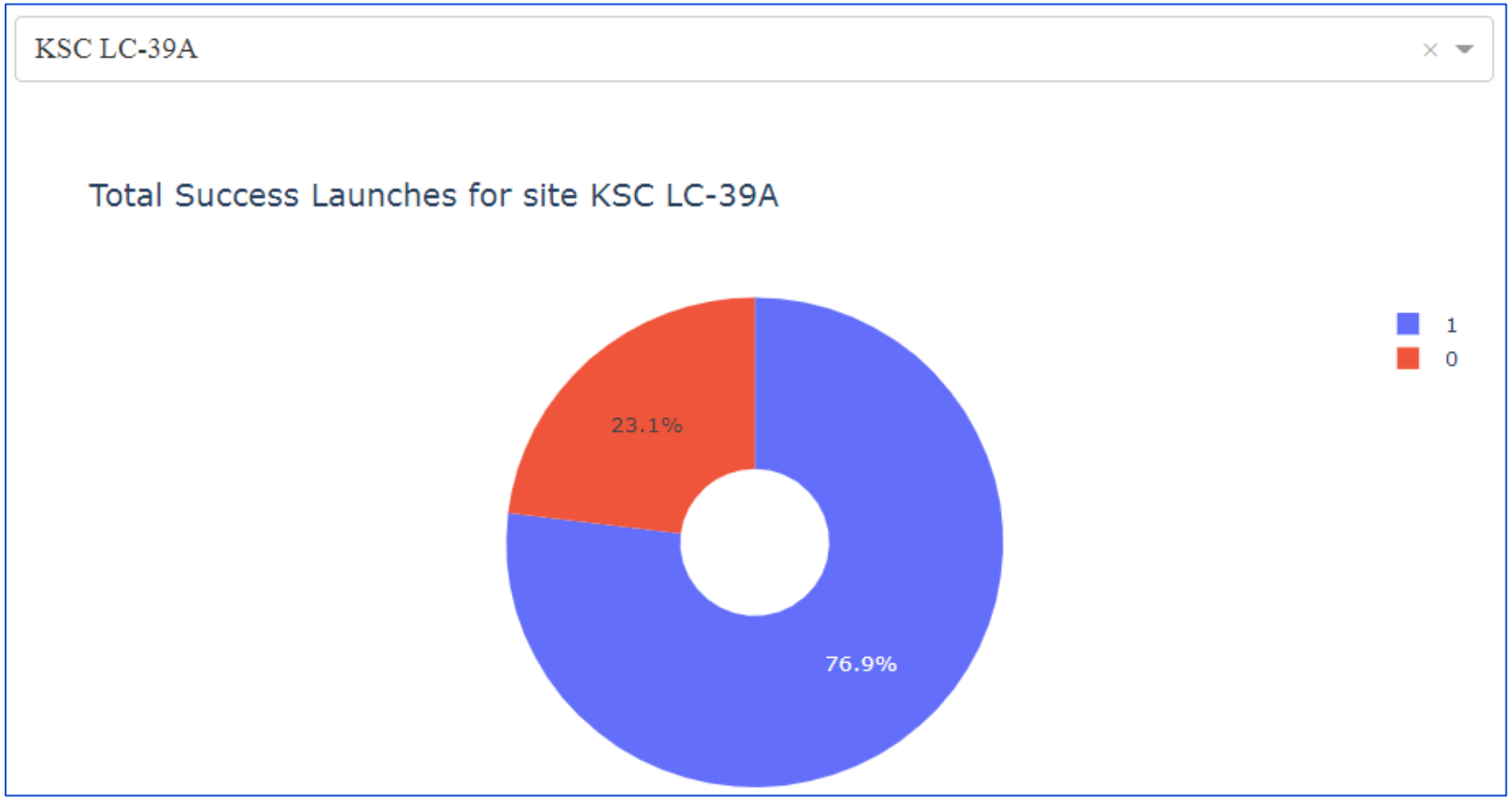
Landing success rate for all launch sites



Interpretation:

KSC LC-39A launch site has the largest share of landing success rate: 41.7%

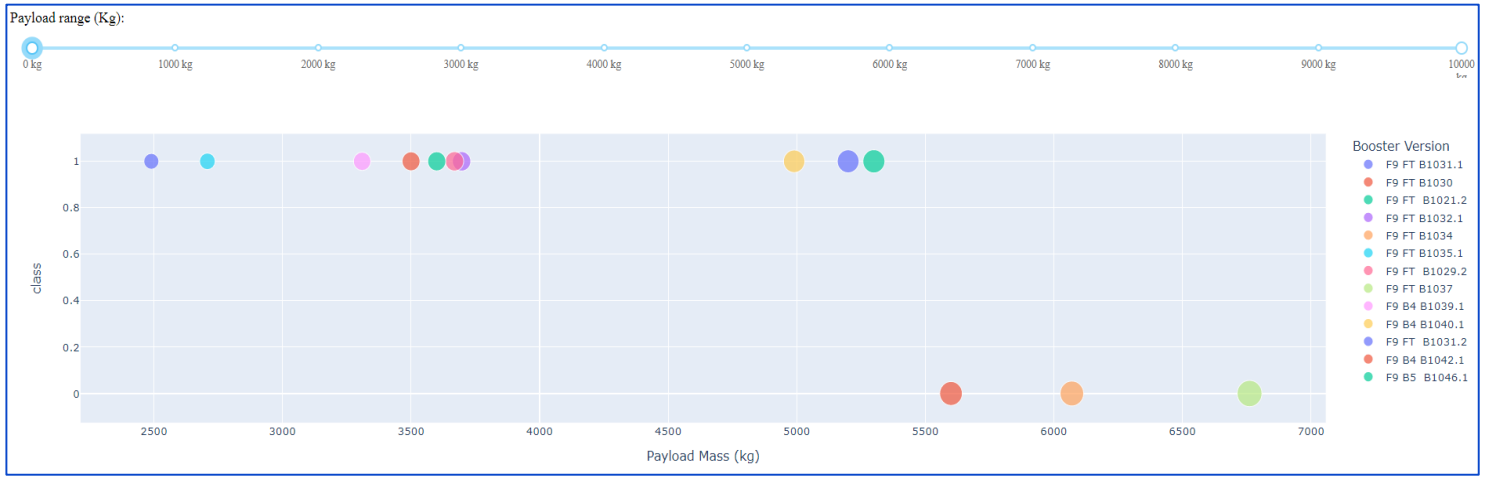
Launch site with the highest landing success rate ratio

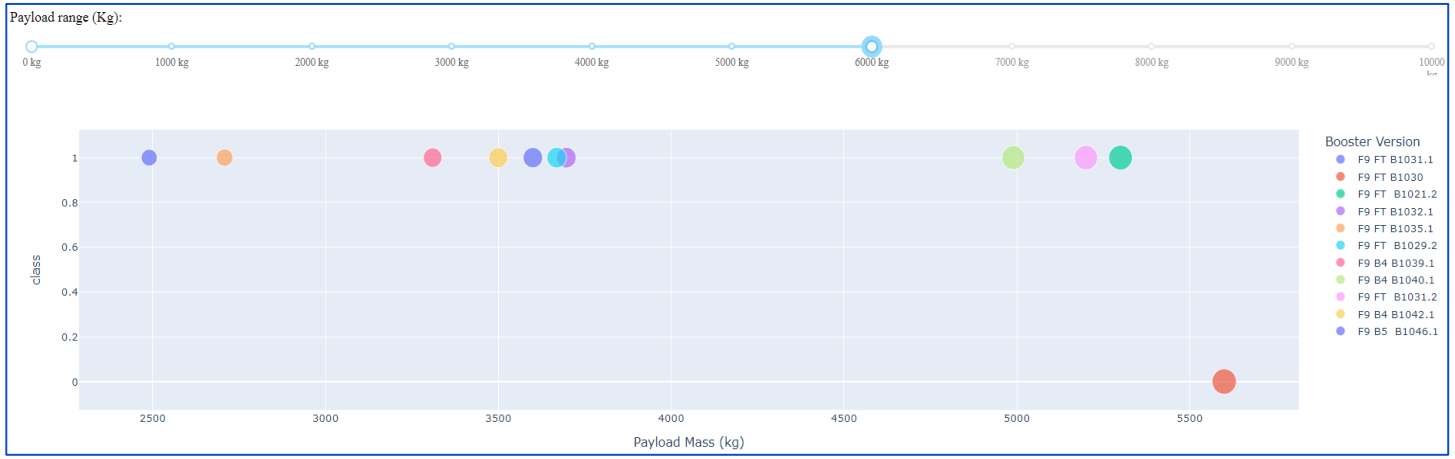


Interpretation:

KSC LC-39A has the highest landing success rate: 76.9%

Landing success rate vs. Payload





Interpretation:

1. Payloads between 2000 and 5500 kg have the highest Landing success rate.
2. F9 FT Booster version has the highest Landing success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Interpretation:

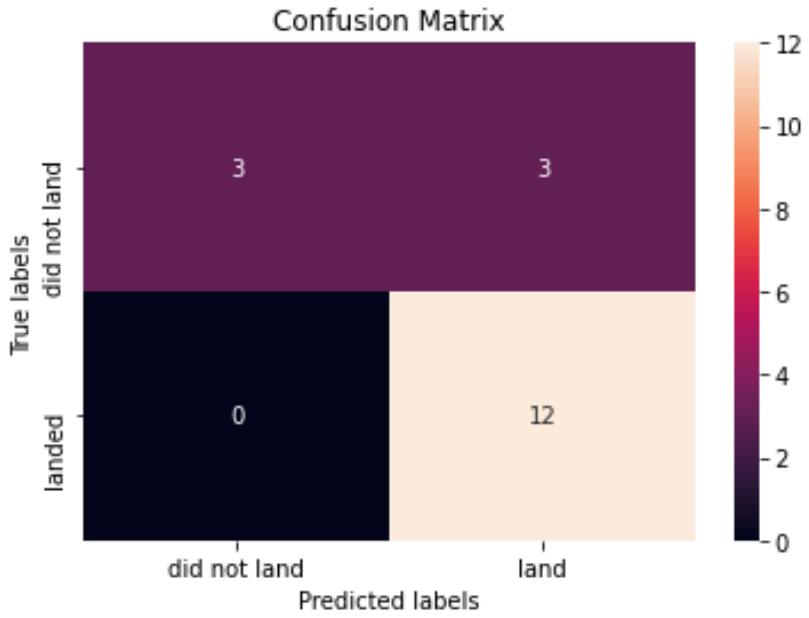
Evaluation metrics calculated for the full sample revealed that the best classifier is the Decision Tree model

	LogReg	SVM	DecisionTree	KNN
Jaccard_Score	0.833333	0.845070	0.848485	0.819444
F1_Score	0.909091	0.916031	0.918033	0.900763
Accuracy	0.866667	0.877778	0.888889	0.855556

Confusion Matrix

Interpretation:

Confusion matrix built for Decision Tree demonstrates that the classifier is able distinguish between the different classes. The remaining problem is False positives - when failed landings marked as successful



Conclusions

Based on the data analysis performed, we can draw the following conclusions in the context of the questions stated at the beginning of the study:

- The following factors influence first stage landing success rate: Launch site, Payload, and Booster version.
- The following pair of predictors has a correlation that determine the success rate of the first stage landing: Payload - Booster version.
- To ensure the best first stage landing success rate, the following operating conditions are recommended:
  - use KSC LC 39A as main launch site,
  - use Booster version F9 FT for payloads between 2000 and 5500 kg and Booster version F9 B5 for payloads over 7000 kg.
  - tend to use payloads from two ranges (First: between 2000 and 5500 kg; Second: between 9000 and 15500 kg) as well-tested and proven high landing success rates.

Thank you!