IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Juan Camilo Jaramillo Tascón
2023-01-17

Outline

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

Executive Summary

- Data collection was carried out using an API and doing web scraping, cleaning and reordering of the data was managed to later carry out an exploratory data analysis with visualization and with database queries. Finally, four classification models were used to determine which of the models is more accurate when predicting whether or not a rocket launch will be successful.
- Three of four models have the same score of 83.3%.
- All the four models have the same confusion matrix.

Introduction

- 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 if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

Section 1

Methodology

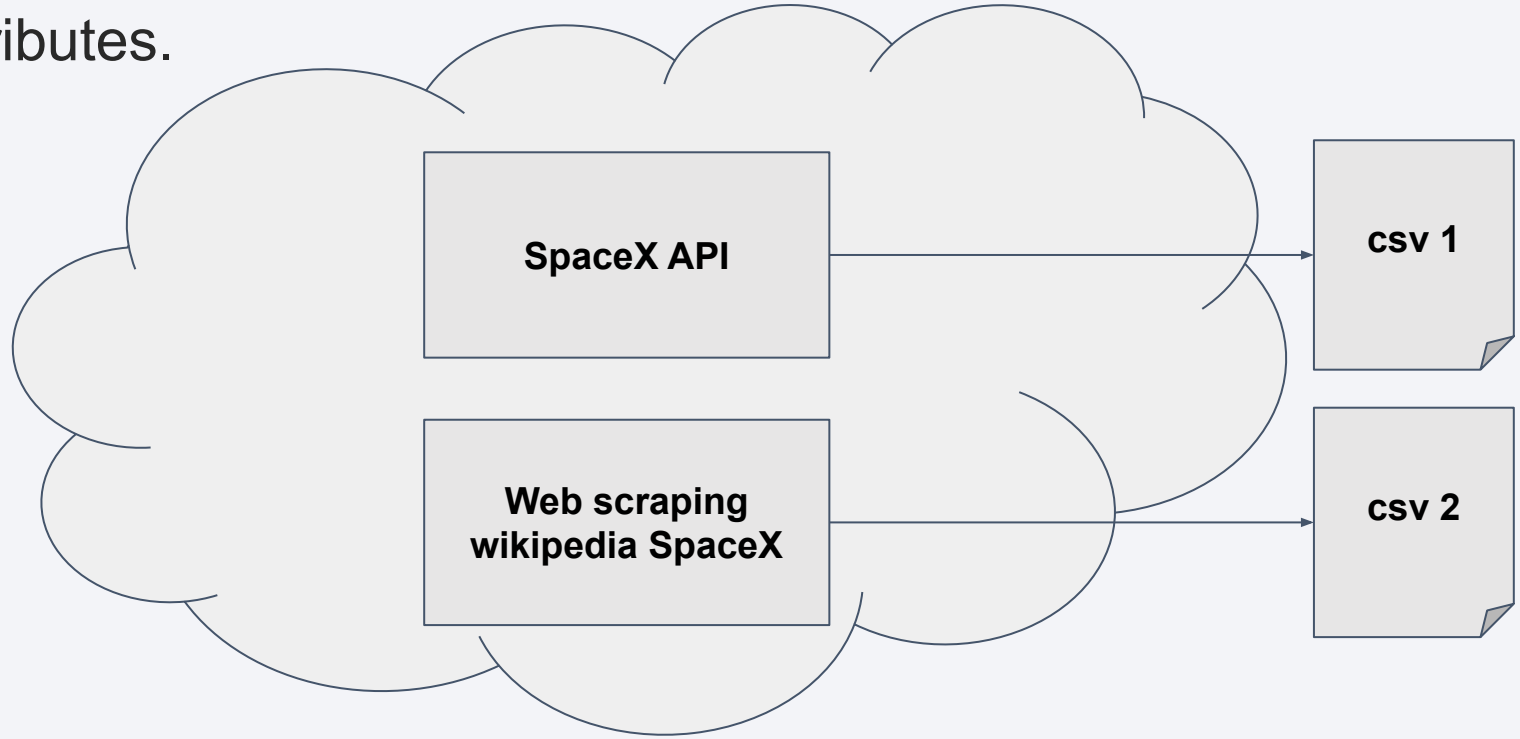
Methodology

Executive Summary

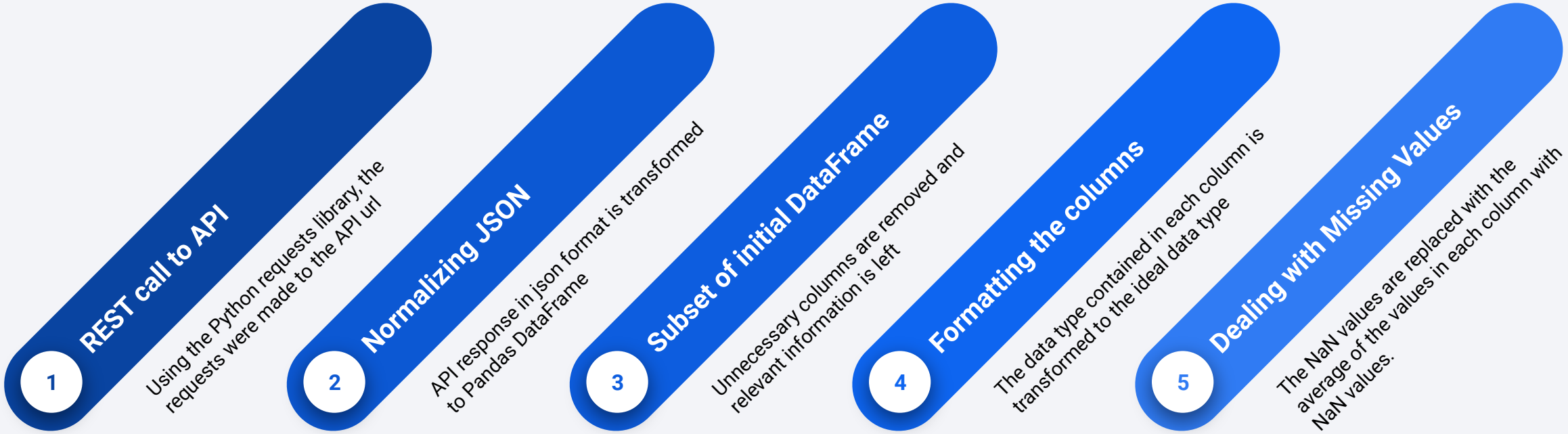
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The data was collected from two information sources, the first using the **SpaceX API** with the Python requests library, and the second performing **Web Scraping**, with the Python BeautifulSoup library, to a Wikipedia page that contains the data in the form of Table of the launches that have been made into space, with some attributes.



Data Collection – SpaceX API



Github Link

Data Collection - Scraping

1. Request the Falcon9 Launch Wiki page from its URL
2. Extract all column/variable names from the HTML table header
3. Create a data frame by parsing the launch HTML tables

Github Link

Data Wrangling

Calculate the number of launches on each site

Use the method **value_counts()** on the column **LaunchSite** to determine the number of launches on each site.

Calculate the number and occurrence of each orbit

Use the method **value_counts()** to determine the number and occurrence of each orbit in the column **Orbit**.

Calculate the number and occurence of mission outcome per orbit type

Use the method **value_counts()** on the column **Outcome** to determine the number of **landing_outcomes**. We create a set of outcomes where the second stage did not land successfully.

Create a landing outcome label from Outcome column

Using the **Outcome**, create a list where the element is zero if the corresponding row in **Outcome** is in the set **bad_outcome**; otherwise, it's one. Then assign it to the variable **landing_class**.

Github Link

EDA with Data Visualization

Github Link

01	**Relationship between Flight Number and Launch Site**	• This chart was needed because we wanted to determine if Launch Site and Launch Flight were related to successful launches.
02	**Relationship between Payload and Launch Site**	• This plot was necessary as we wanted to determine if Launch Site and Payload mass had any relationship to successful launches.
03	**Relationship between success rate of each orbit type**	• This chart was needed as we want to determine if the Orbit type was related to successful launches.
04	**Relationship between FlightNumber and Orbit type**	• This chart was needed as we wanted to determine if Orbit type and Launch Flight had any relationship to successful launches.
05	**Relationship between Payload and Orbit type**	• This plot was necessary as we wanted to determine if Orbit type and Payload mass had any relationship to successful launches.
06	**Launch success yearly trend**	• This graph was necessary as we wanted to determine if the launch year had any relationship to successful launches.

EDA with SQL

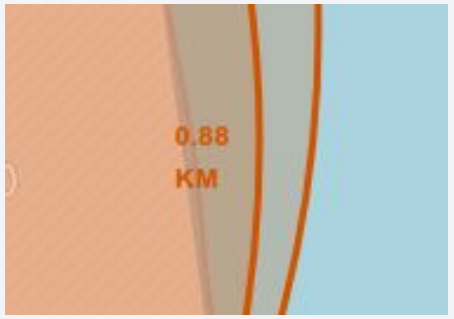
- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Github Link

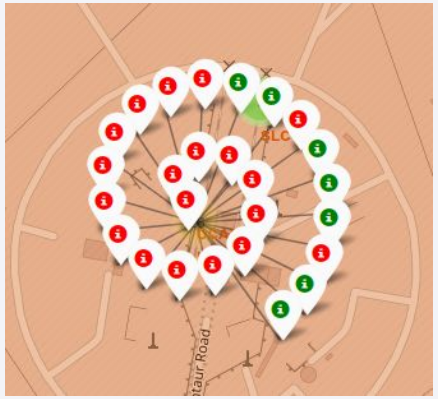
Build an Interactive Map with Folium

Github Link

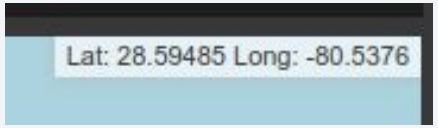
folium.Circle:
Used to mark an area of a place



folium.Marker:
Used to give information about a place or area



MarkerCluster:
Used to display different elements in the same location

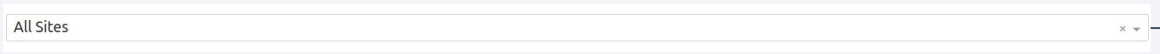


MousePosition:
Used to display the coordinates of the mouse pointer location

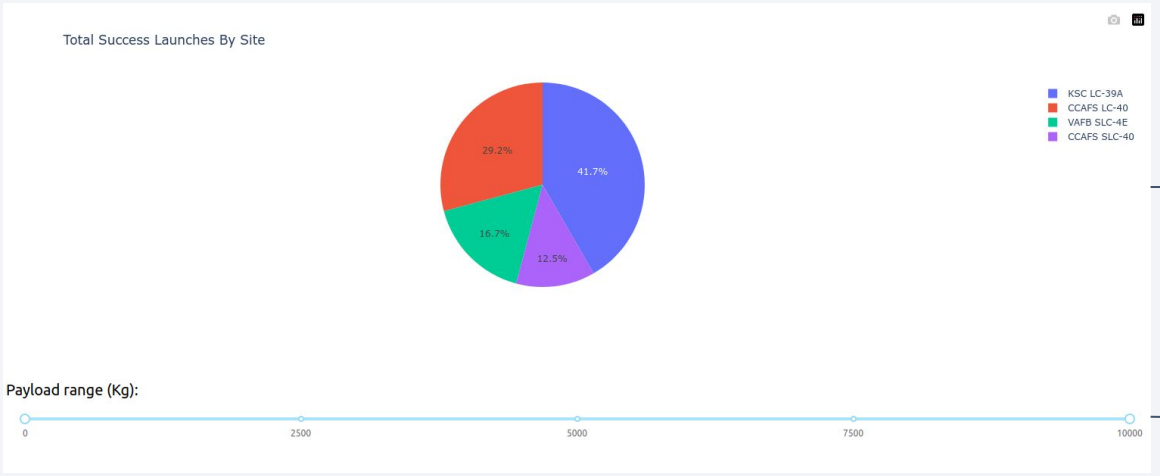
folium.PolyLine:
Used to join two geographic locations

Build a Dashboard with Plotly Dash

Github Link

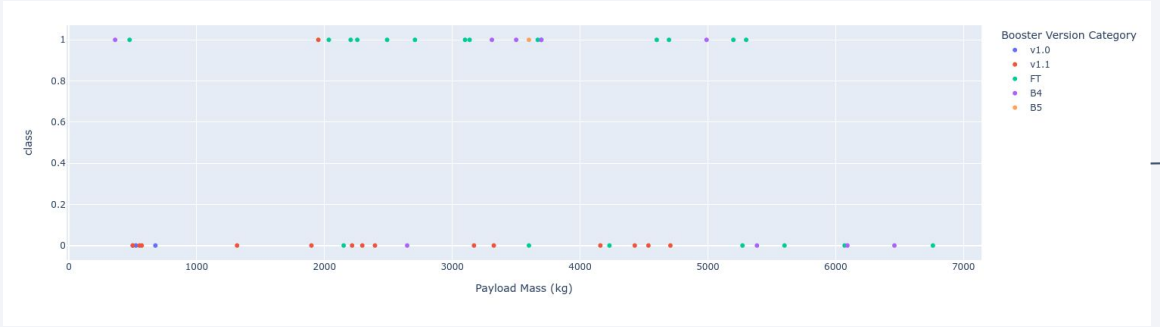


Dropdown: used for filter the launch sites



Pie chart: used for show the comparison of successful launches

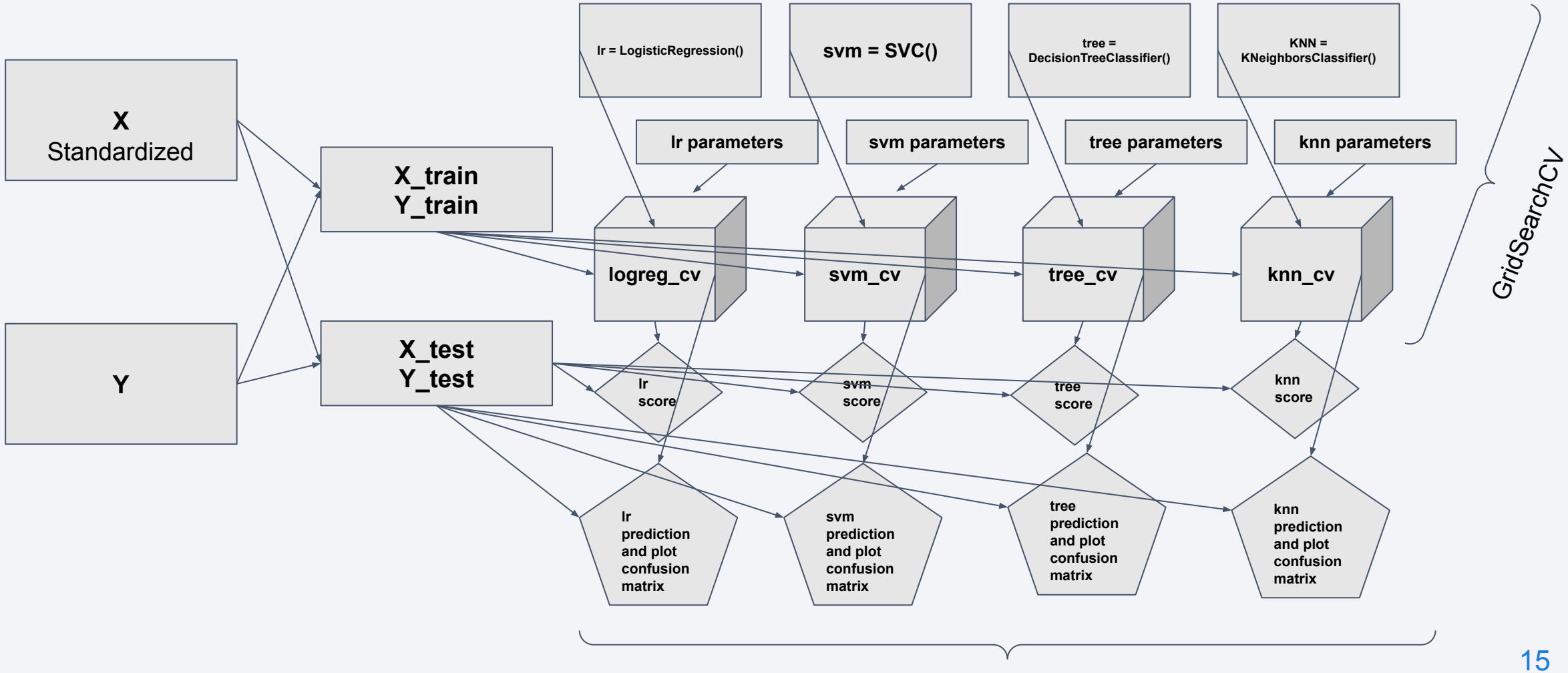
Range slider: used for filter the payload in kg of the launches



Scatter plot: used for show the comparison between success of a launch, payload and Booster version

Predictive Analysis (Classification)

Github Link



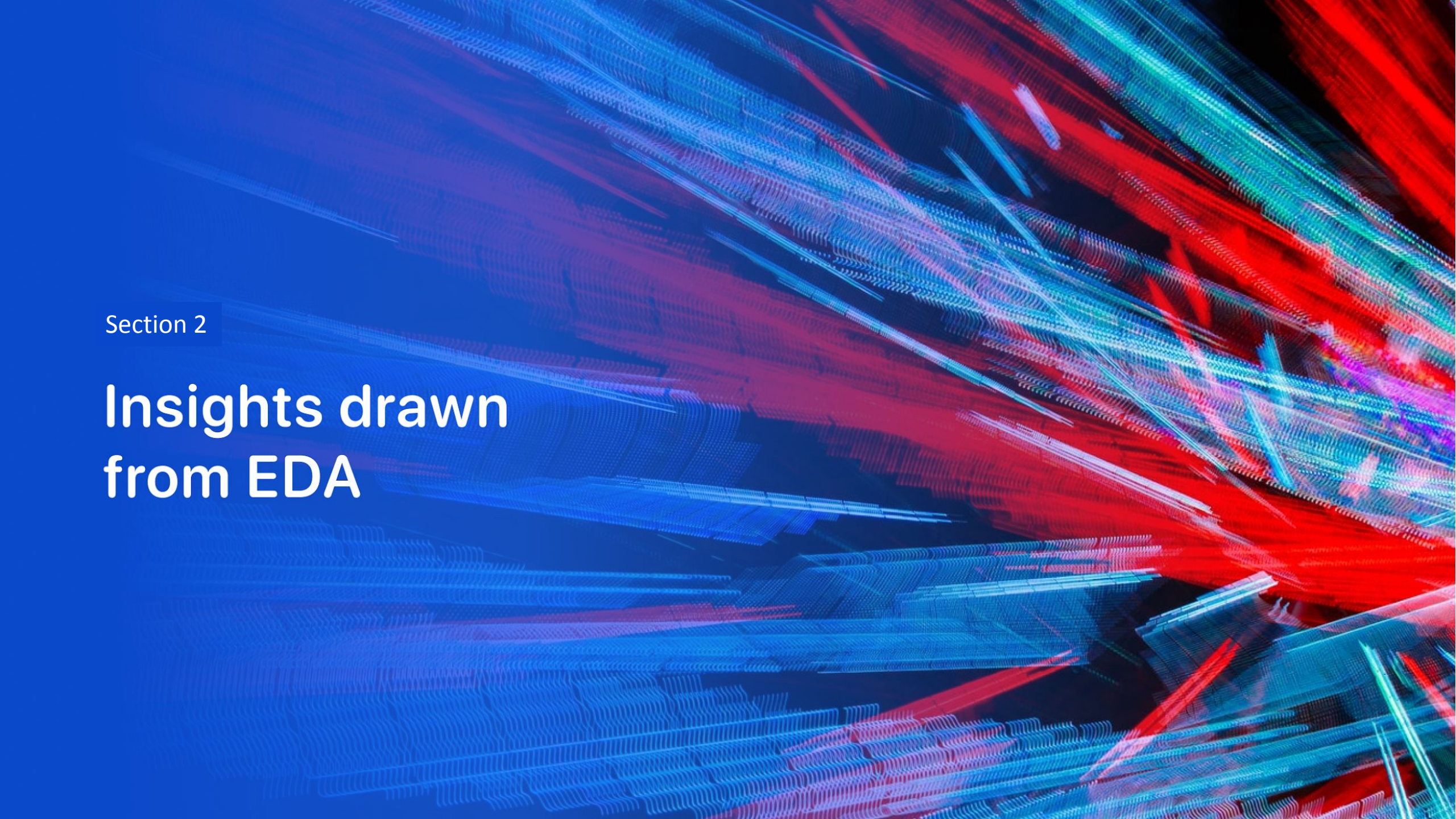
Choose the best score, prediction and confusion matrix

Results

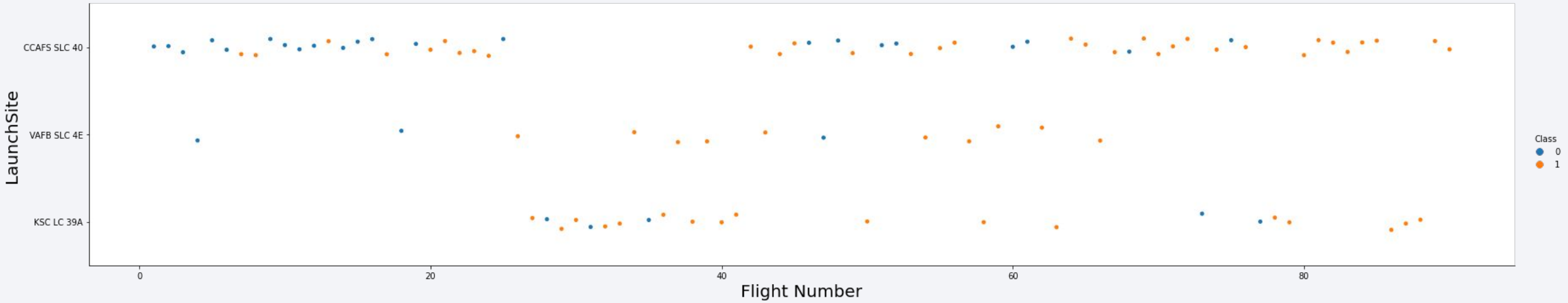
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

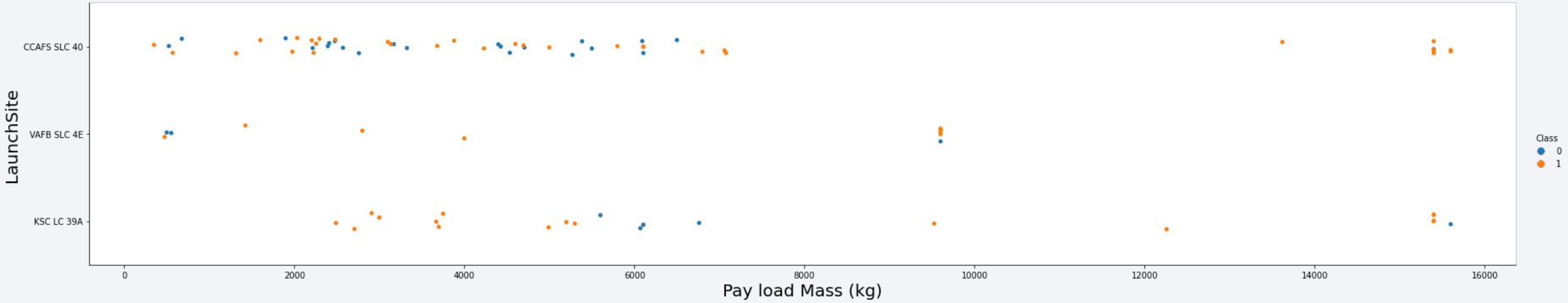


Flight Number vs. Launch Site



- It is observed that from flight number 25 the number of successful flights increased for each of the flight stations.
- The first 6 flights were not successful.
- The station with the highest percentage of successful launches is VAFB SLC 4E.

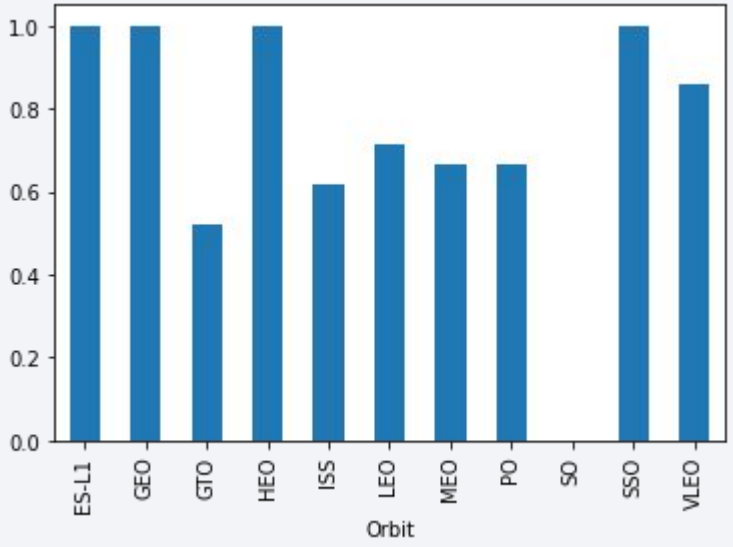
Payload vs. Launch Site



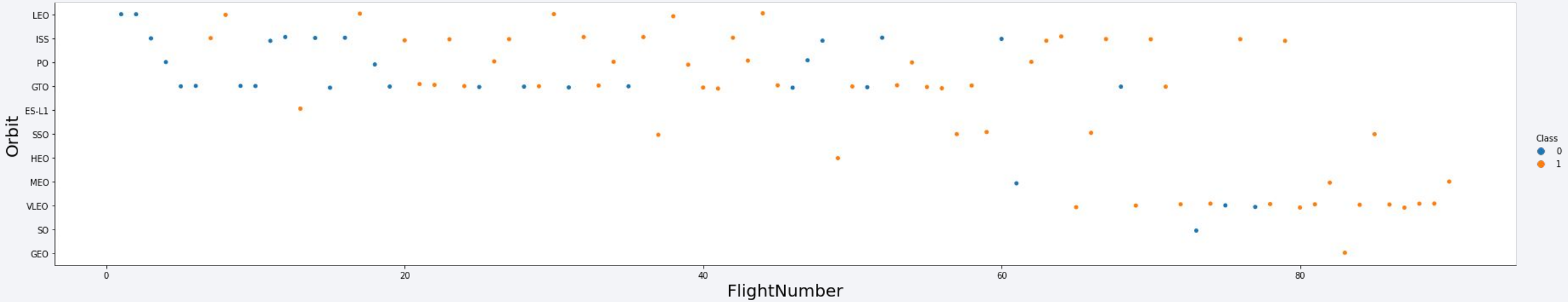
- For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000)
- It cannot be determined by eye that payload affects successful flights from station CCAFS SLC 40.

Success Rate vs. Orbit Type

- The orbits that have had 100% successful flights are ES-L1, GEO, HEO and SSO, while for orbit SO it has not had any successful flights.

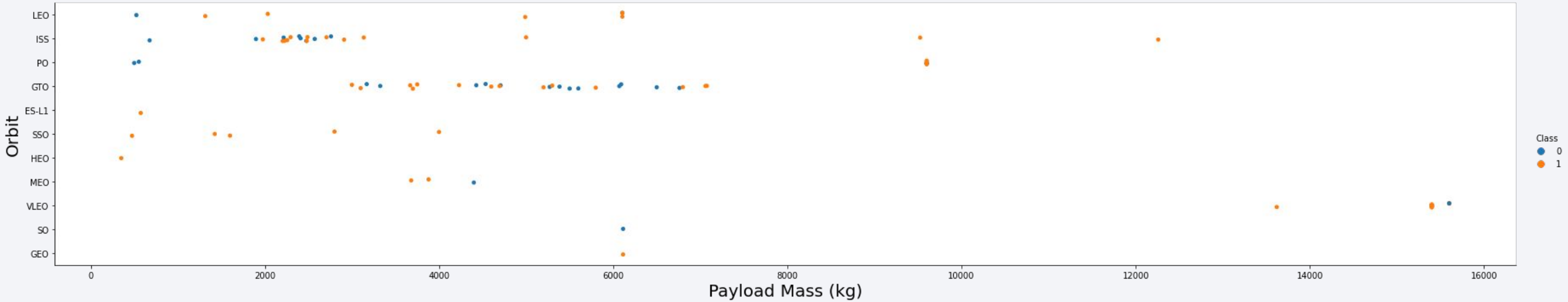


Flight Number vs. Orbit Type



- In the LEO orbit the Success appears related to the number of flights
- There seems to be no relationship between flight number when in GTO orbit.

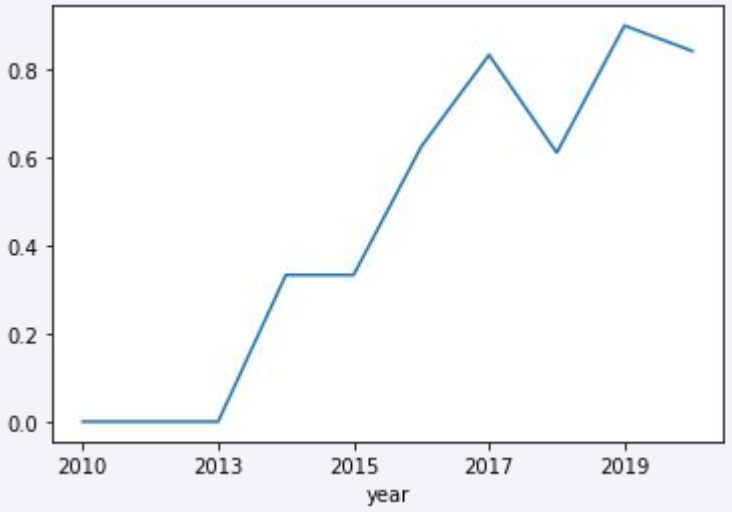
Payload vs. Orbit Type



- In the LEO orbit the Success appears related to the Payload mass

Launch Success Yearly Trend

- Observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

- The different launch sites were selected from the table

```
[ ]  %sql SELECT DISTINCT(Launch_Site) FROM SPACEXTBL;

 * sqlite:///my_data1.db
Done.
 Launch_Site
 CCAFS LC-40
 VAFB SLC-4E
 KSC LC-39A
 CCAFS SLC-40
```

Launch Site Names Begin with 'CCA'

- All columns in the table where the launch site began with the letters CCA are selected, and your search was limited to 5

```
[ ] %sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	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 sum of the PAYLOAD_MASS__KG column of the table was selected, where the client had the letters NASA (CRS)

```
[ ] %sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass_carried_by_boosters_launched_by_NASA_CRS FROM SPACEXTBL WHERE Customer LIKE '%NASA (CRS)%';
 * sqlite:///my_data1.db
Done.
total_payload_mass_carried_by_boosters_launched_by_NASA_CRS
48213
```

Average Payload Mass by F9 v1.1

- The sum of the PAYLOAD MASS KG column of the table was selected, where the Booster version had the letters F9 v1.1

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

```
[ ] %sql SELECT AVG(PAYLOAD_MASS__KG_) AS average_payload_mass_carried_by_booster_version_F9_v1_1 FROM SPACEXTBL WHERE Booster_Version LIKE '%F9 v1.1%';
```

```
 * sqlite:///my_data1.db
Done.
average_payload_mass_carried_by_booster_version_F9_v1_1
2534.6666666666665
```

First Successful Ground Landing Date

- The date was selected from the table where the mission was successful and the landing_outcome had the letters "ground pad" ordered by date and choosing the first row.

```
[ ] %sql SELECT Date FROM SPACEXTBL WHERE Mission_Outcome='Success' AND Landing_Outcome LIKE '%ground pad%' ORDER BY Date DESC LIMIT 1;
```

```
 * sqlite:///my_data1.db
Done.
   Date
22-12-2015
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- Selected the Booster_version from the table where the Landing_outcome was "Success (drone ship) and the Payload_mass was between 4000 and 6000 kg

```
[ ] %sql SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome='Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

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

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

Total Number of Successful and Failure Mission Outcomes

- Selected the Mission_outcome and the Mission_outcome account from the table, grouped by Mission_outcome

```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTBL GROUP BY Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

- A subquery was used to calculate the maximum payload mass and used as a search parameter to determine which Booster_version had releases with that mass.

```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

2015 Launch Records

- Selected the month, Landing_outcome, Booster_version, and Launch_site, from the table where the year was 2015 and the Landing_outcome had the letters "Failure (drone ship).

```
[ ] %sql SELECT substr(Date, 4, 2) AS month_name, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE substr(Date,7,4)='2015' AND Landing_Outcome='Failure (drone ship)';

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

month_name	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

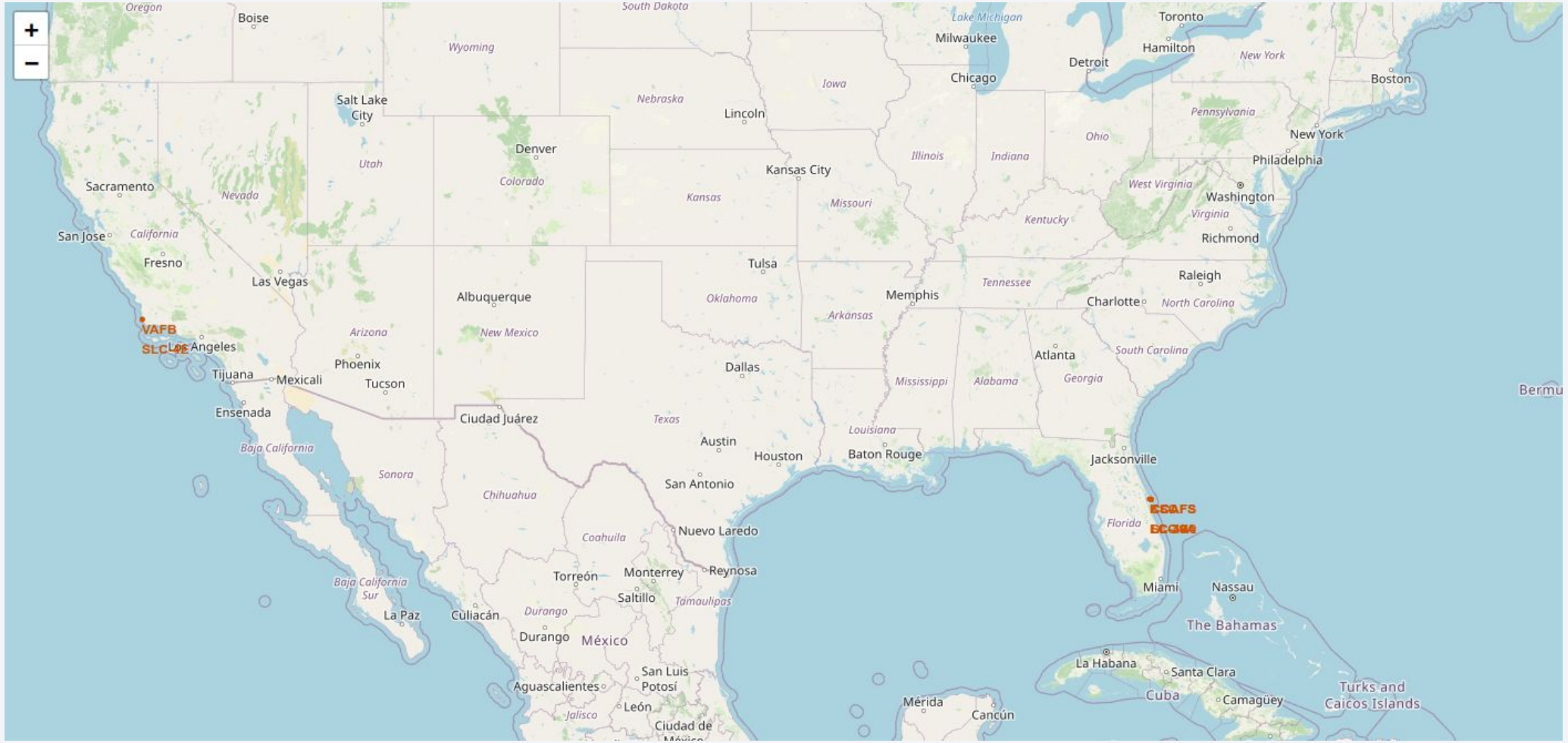
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Rank the 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

- Present your query result with a short explanation here

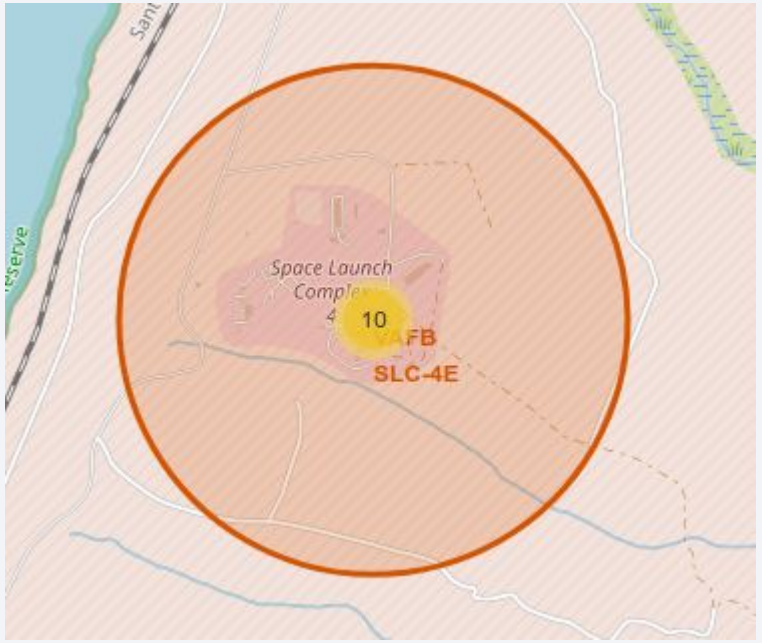

Section 3
Launch Sites
Proximities Analysis

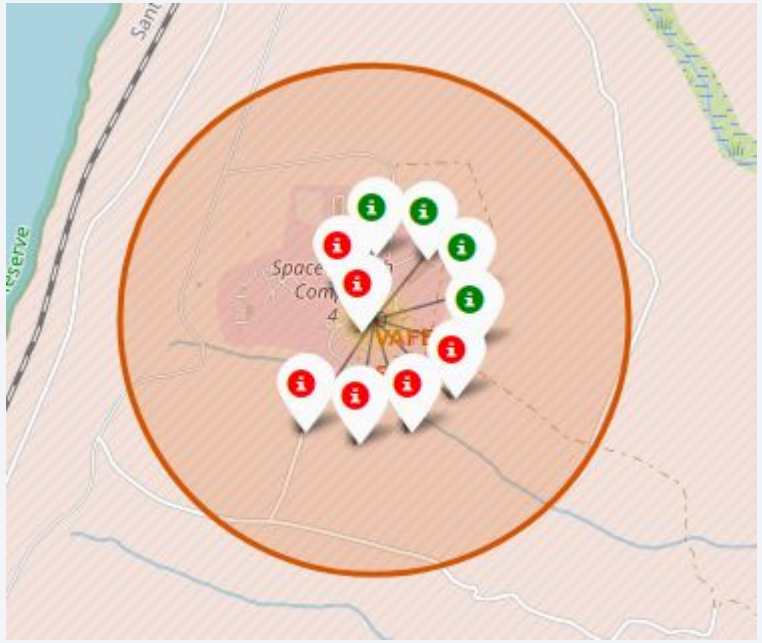
Launch sites location markers on a global map



- All the Launch Sites are near to the sea, one in the Pacific Ocean, and the others in the Atlantic Ocean.

Color-labeled launch outcomes on the map

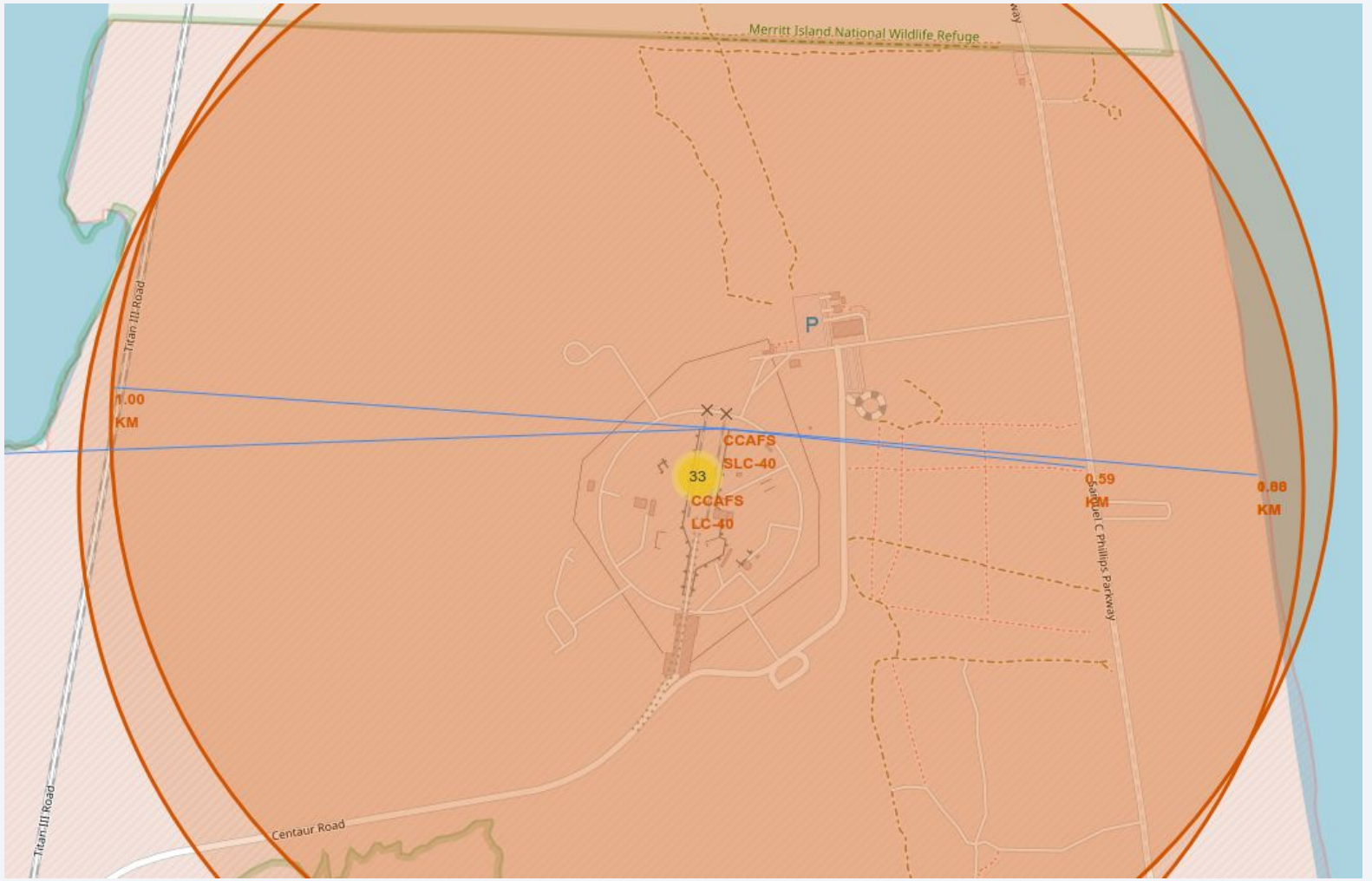


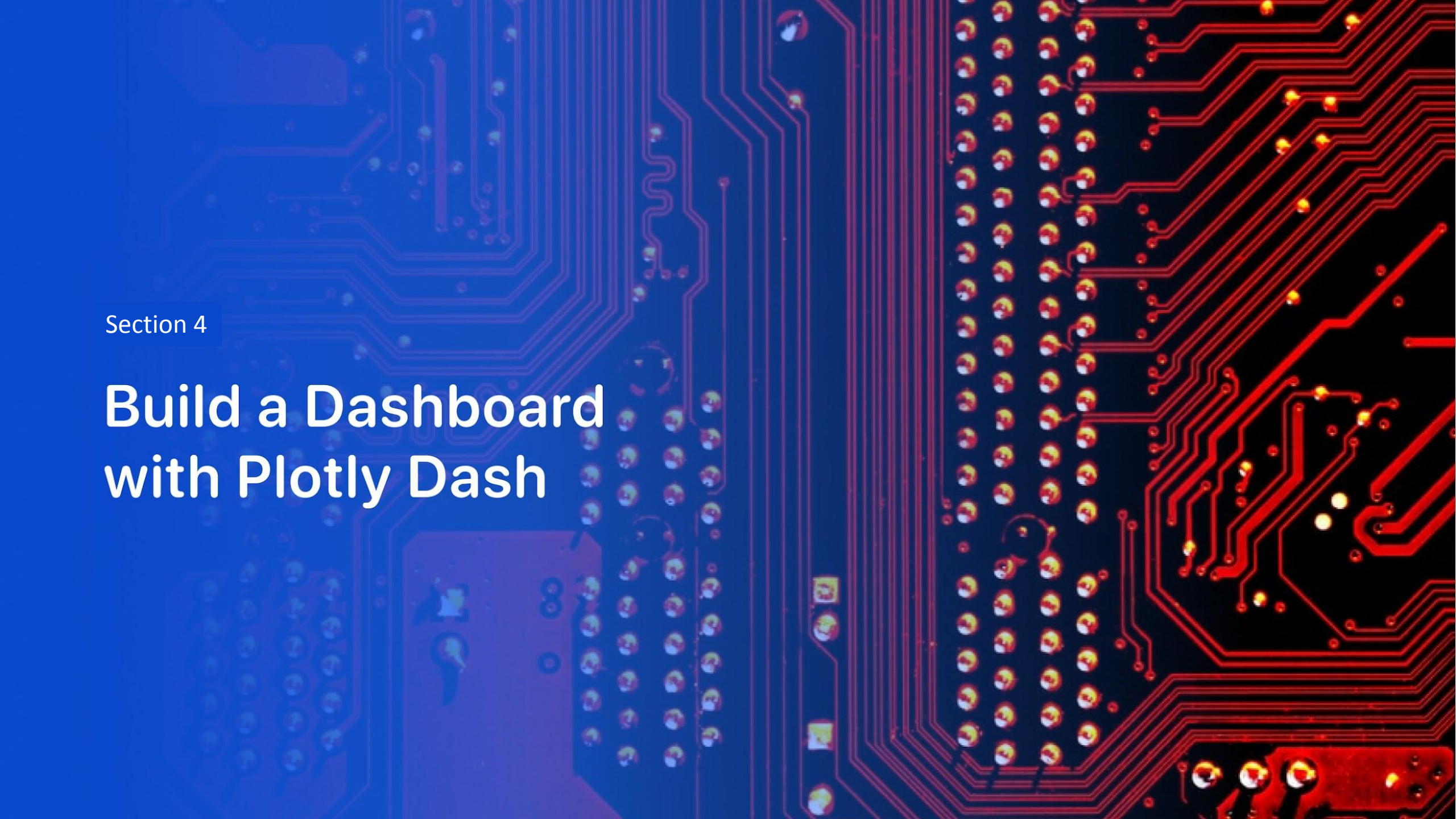


- The green markers are success launches, and the red are unsuccess, all are in the same Launch Site.

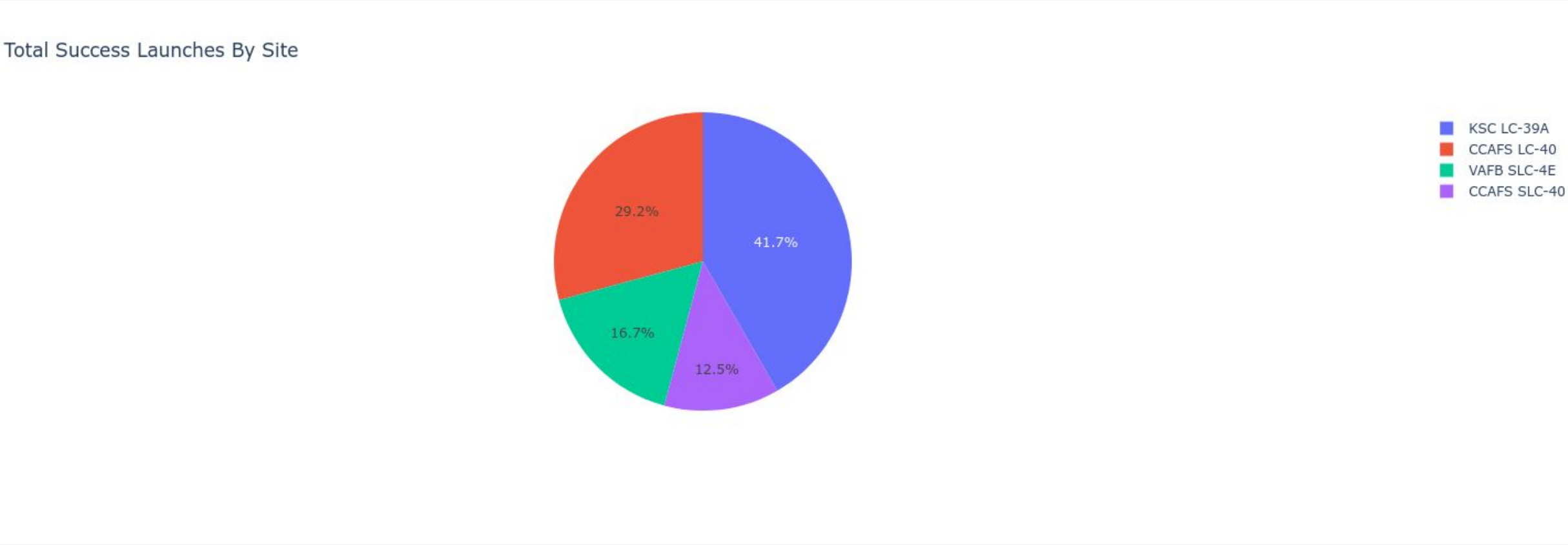
Launch site to its proximities

- The CCAFS SLC-40 Launch Site is 0.88 km near to the coastline, 0.59 km near to the highway, and 1 km near to the railway.



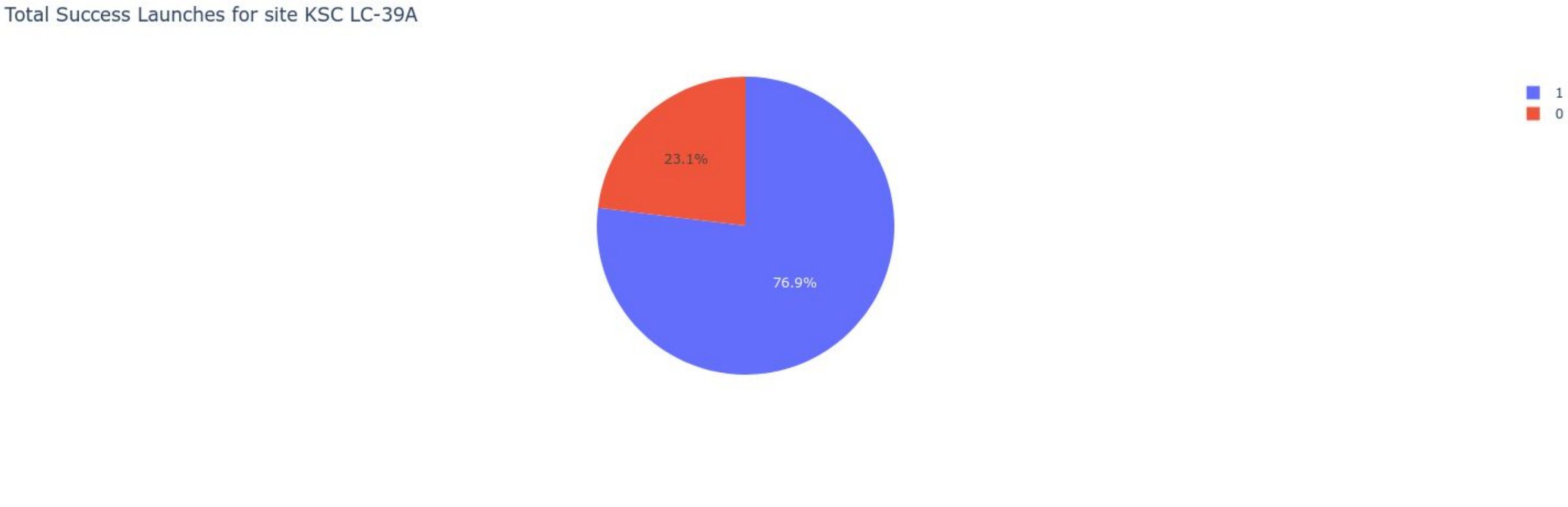

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



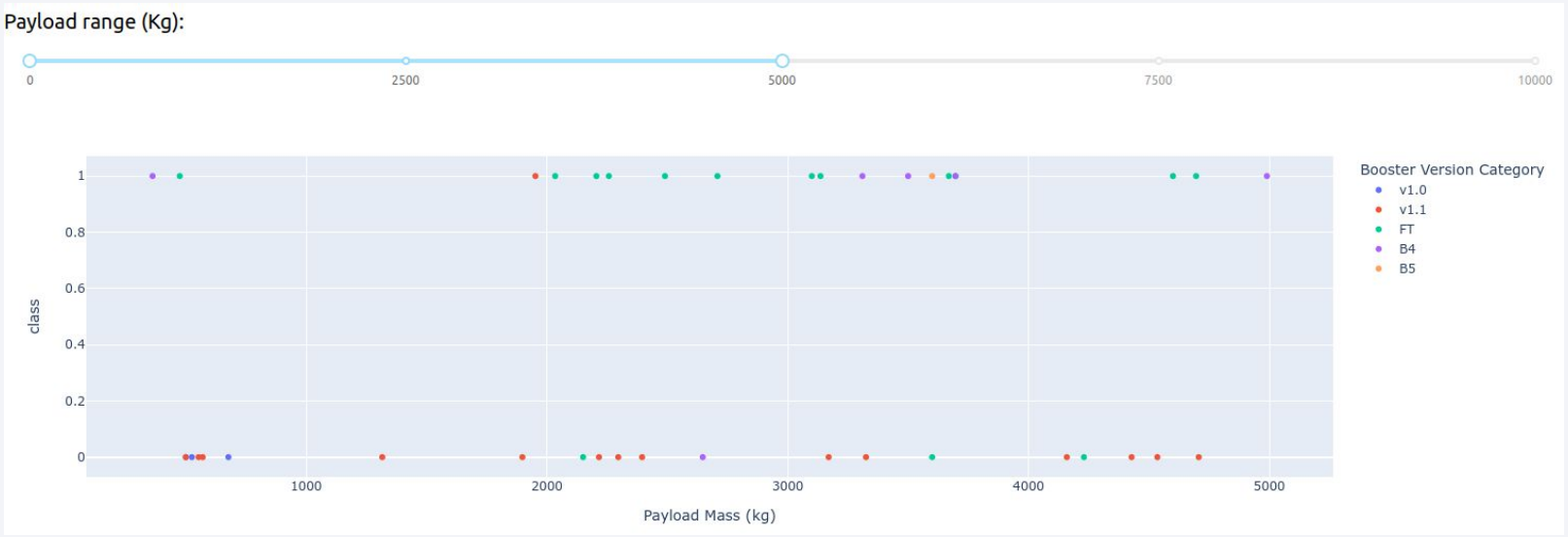
- The most successful Launch Site is KSC LC-39A

Launch site with highest launch success ratio

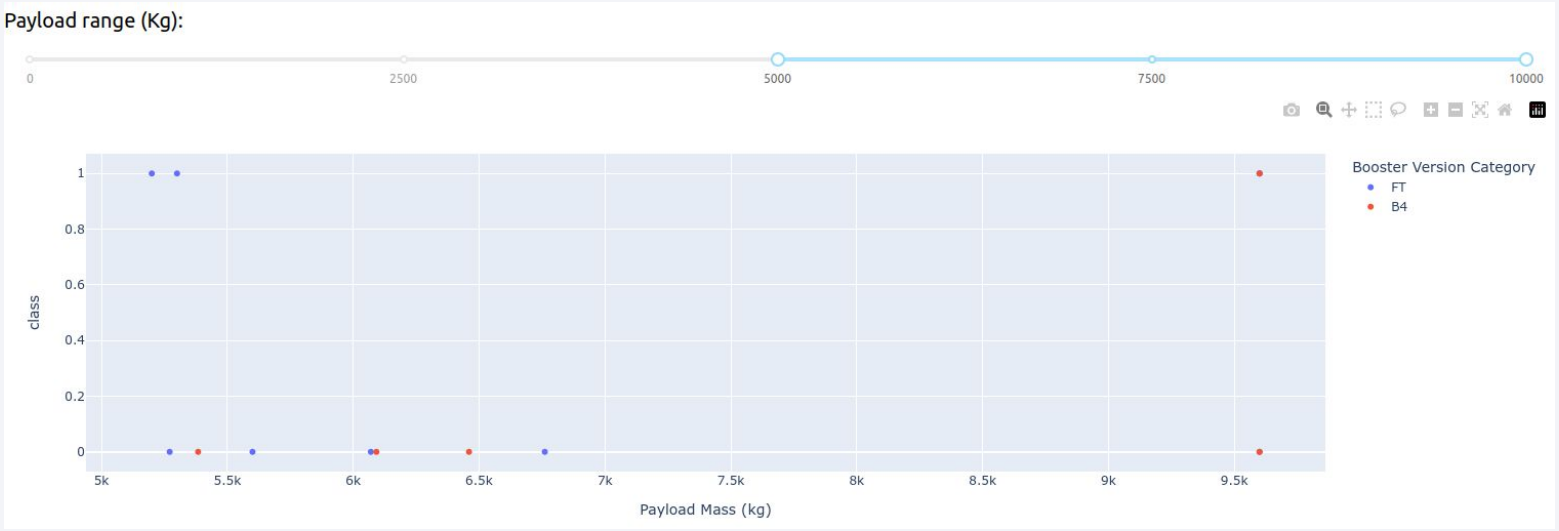


- 76.9 % of the launches of Launch Site KSC LC-39 A are success

Payload vs. Launch Outcome scatter plot for all sites



- The best version of Booster with the highest number of successful launches for payloads less than 5000 kg has been the FT.



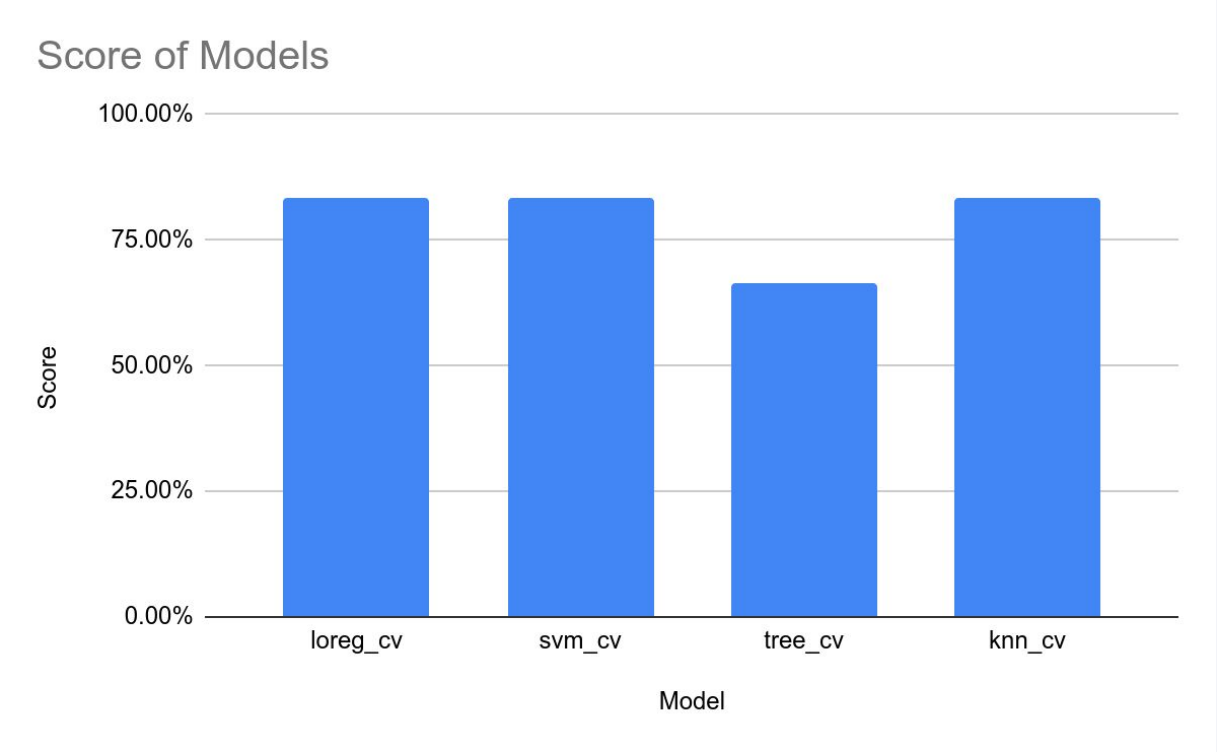
- There have been few successful launches with a payload greater than 5,000 kg.

Section 5

Predictive Analysis (Classification)

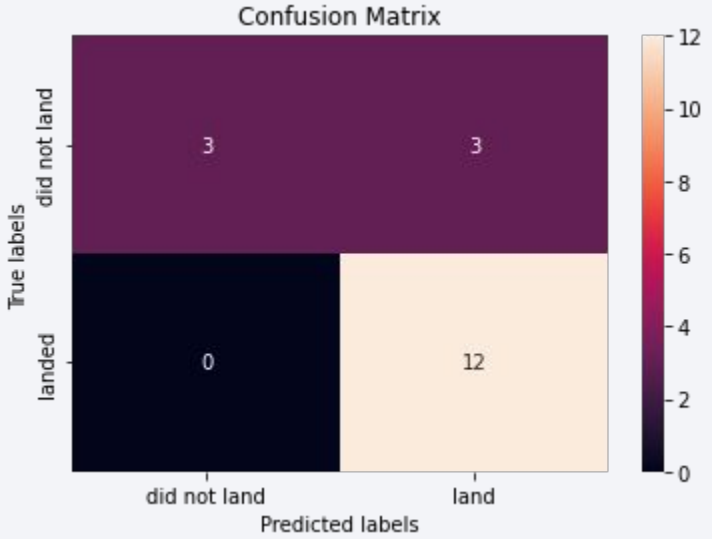
Classification Accuracy

- Three of four models have the same score of 83.3%



Confusion Matrix

- All the four models have the same confusion matrix.



Conclusions

- The most successful Launch Site is KSC LC-39A
- 76.9 % of the launches of Launch Site KSC LC-39 A are success
- The best version of Booster with the highest number of successful launches for payloads less than 5000 kg has been the FT.
- There have been few successful launches with a payload greater than 5,000 kg.
- Three of four models have the same score of 83.3%.
- All the four models have the same confusion matrix.

Appendix

- Github repository: https://github.com/jucajata/IBMDataScienceCapstone
- Autor: Juan Camilo Jaramillo Tascón - Data Scientist & Electrical Engineer

Thank you!