Winning Space Race with Data Science

Daniel Tibaquira
17/01/22

Outline

Executive Summary

Introduction

Methodology

Results

Conclusion

Appendix

Executive Summary

Summary of methodologies

- Data Collection using Web Scraping
- Data Wrangling
- Exploratory Data Analysis with Visualization
- Exploratory Data Analysis with SQL
- Interactive Visualization using Folium
- Dashboard using plotly

Summary of all results

- Results and Analyzing the Data
- Using Machine Learning to make Predictions

Introduction

Project background and context

It won't take long before we change our vacation plans, now we will start to take **spaceships** instead of airplanes. Even though it isn't precisely cheap to do it right now, there are companies aiming to achieve this while keeping it affordable.

One most successful companies in this industry is SpaceX. In their accomplishments we can find: Spacecraft to the ISS and Starlink, a satellite internet constellation providing satellite internet access. Also, we can explore BlueOrigin, a company that already made their first **commercial space travel** and has many people trying to book their flight.

One of the reason for these companies to have such success is because the manufacturing of the rockets are becoming relatively **less expensive**, and some parts are becoming **reusable** too!

This project aims find out the price for each rocket launch for company SpaceY, a competitor of SpaceX, so that we can generate **key insights** for decision making. Rockets costs are dependent on the reuse rate of the first stage rocket, that's why we will be studying parameters for the First Stage of rocket launches.

Project objective

we would like to find out how the parameters in the first stage affect on the **successful** landing rate of Falcon 9.

We can find out the **best parameters** in the first stage to ensure the successful landing rate of our First Stage rocket so that we will be able to **determine the best cost** for the Rocket Launch.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- Using the Space X REST API
- Web Scraping data for Falcon 9
- Scraping Falcon Launches from Wikipedia using BS4

Perform data wrangling

- Converting outcomes into Training Labels
  - 1 Being that the booster successfully landed
  - 0 if it was unsuccessful

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- Obtain the best Hyperparameters for SVM, ClassificationTrees and Logistic Regression.

Data Collection

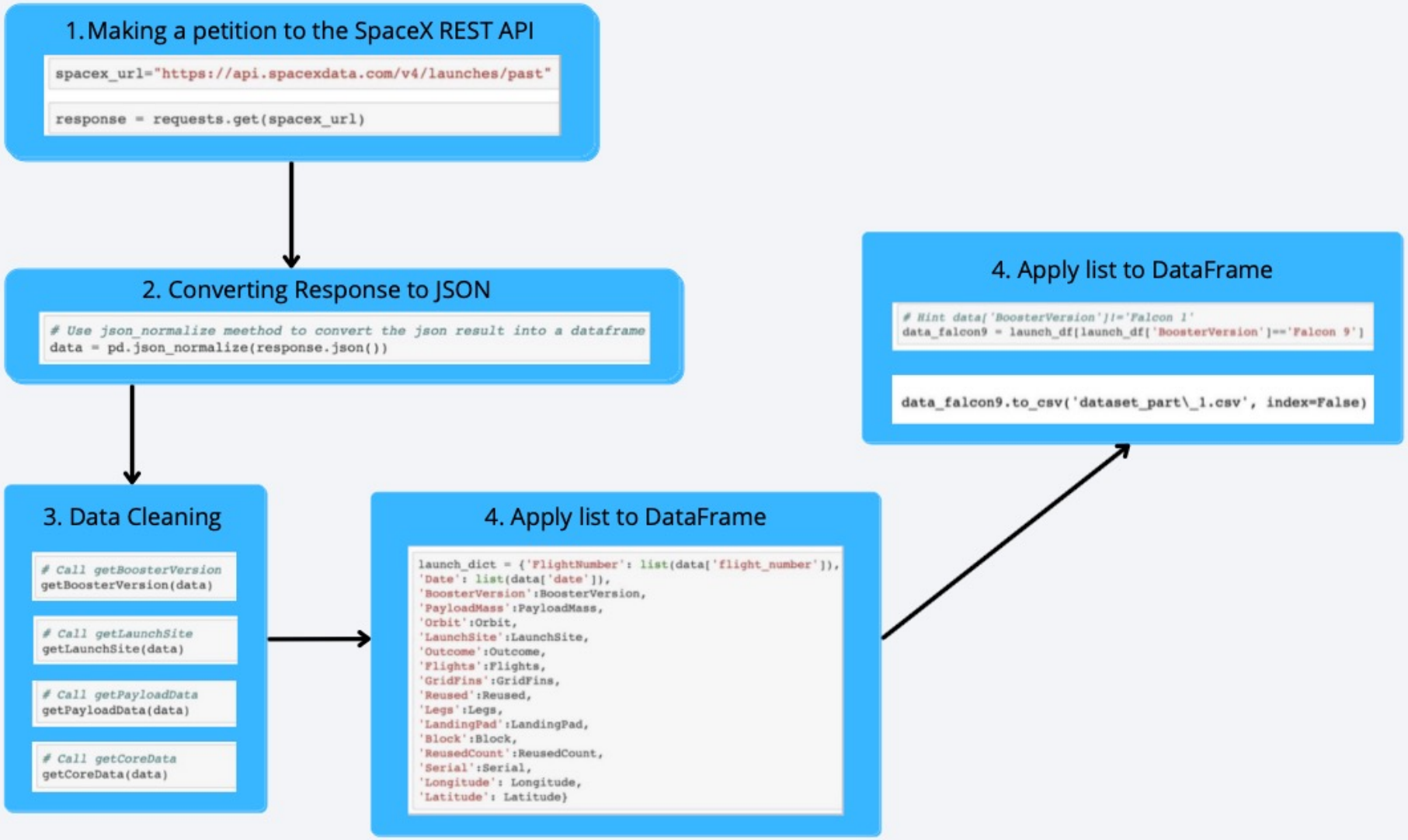
The Data collection process includes a combination of API request from SpaceX REST API

The API provide us information such as:

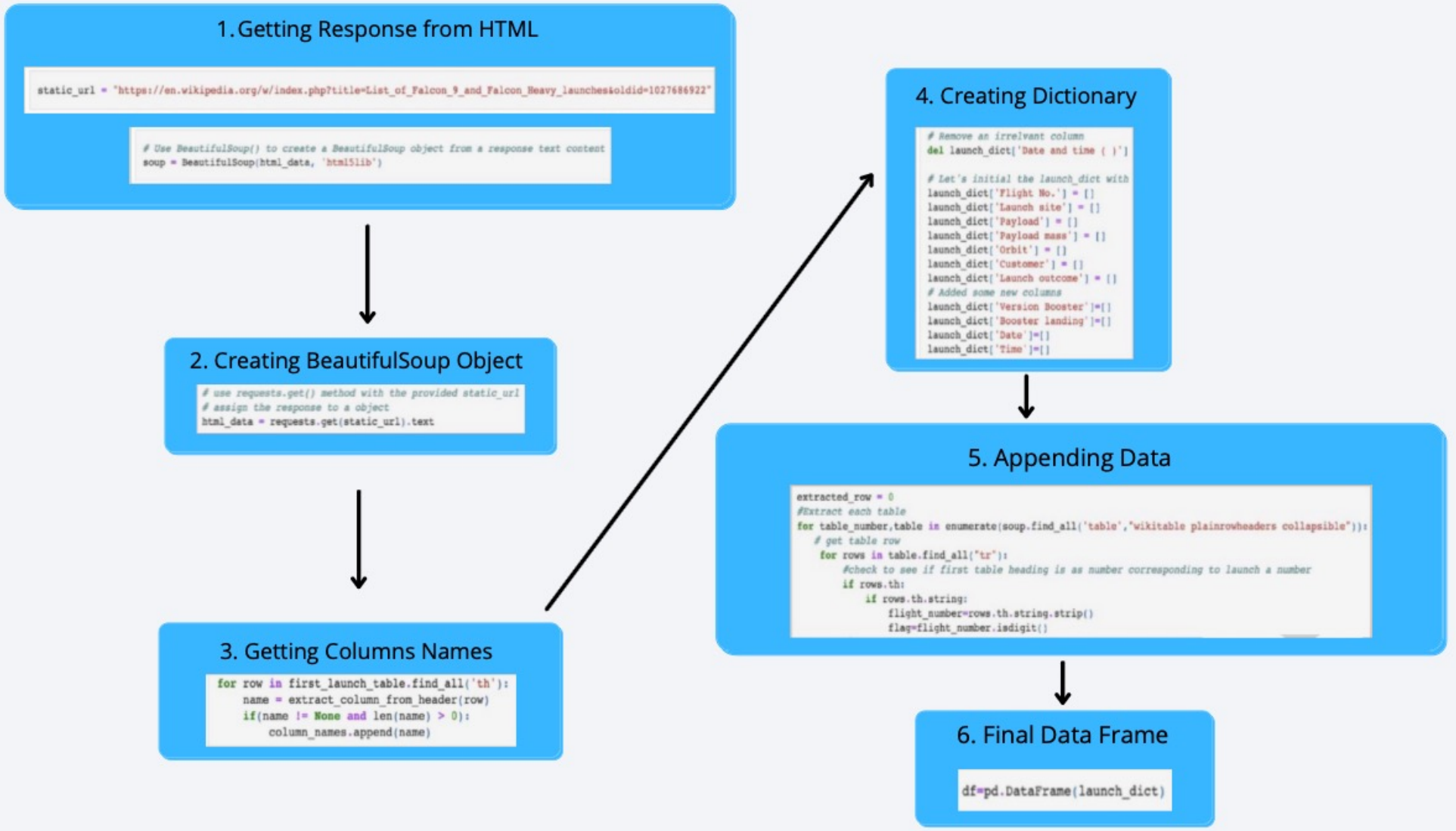
- Information of rocket
- payload delivered.
- Launch specification
- Landing specification
- Landing outcome
- location and etc.

Using BeautifulSoup for web scraping on Wikipedia (Falcon9 Launch data information)

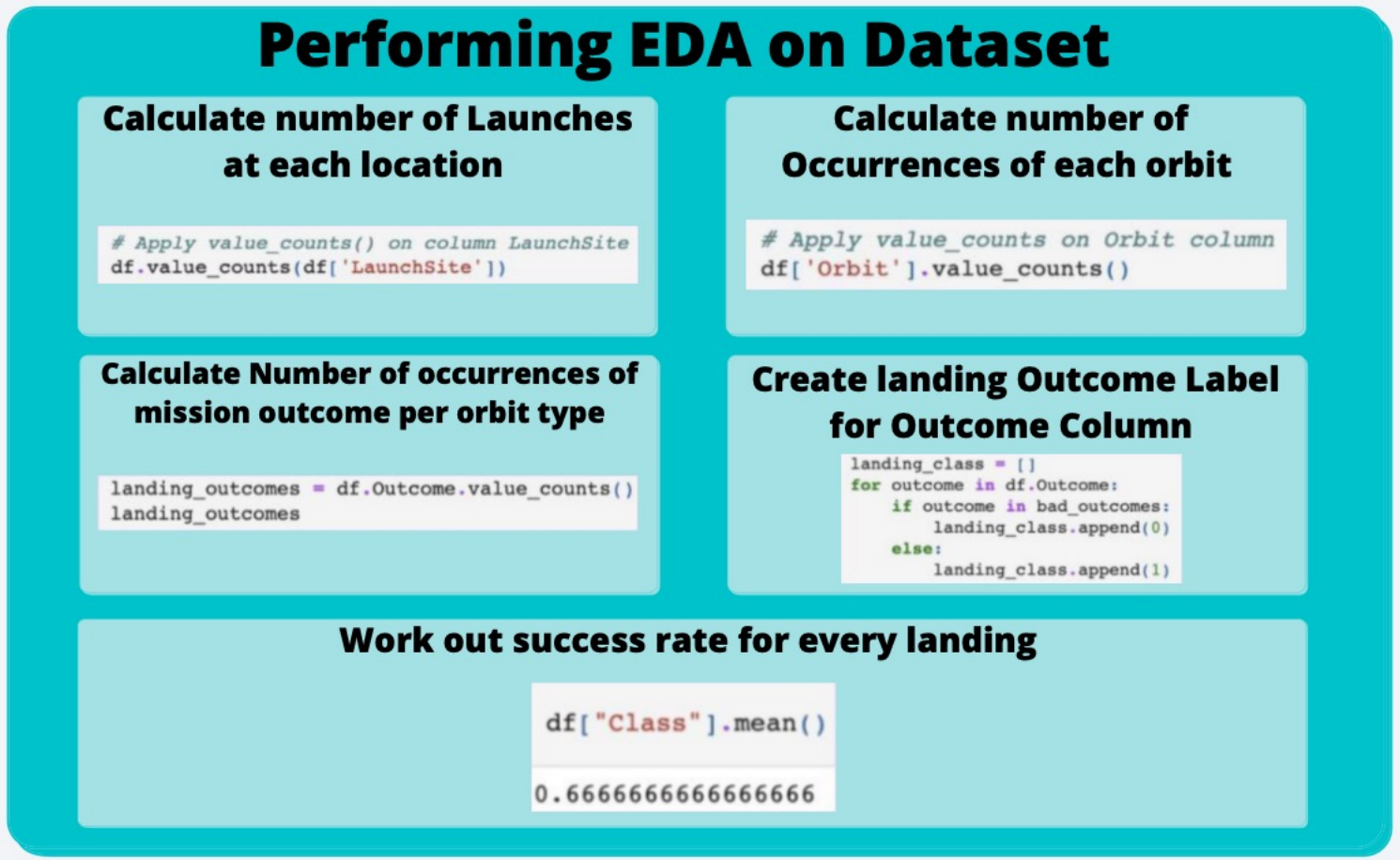
Data Collection – SpaceX API - Repo



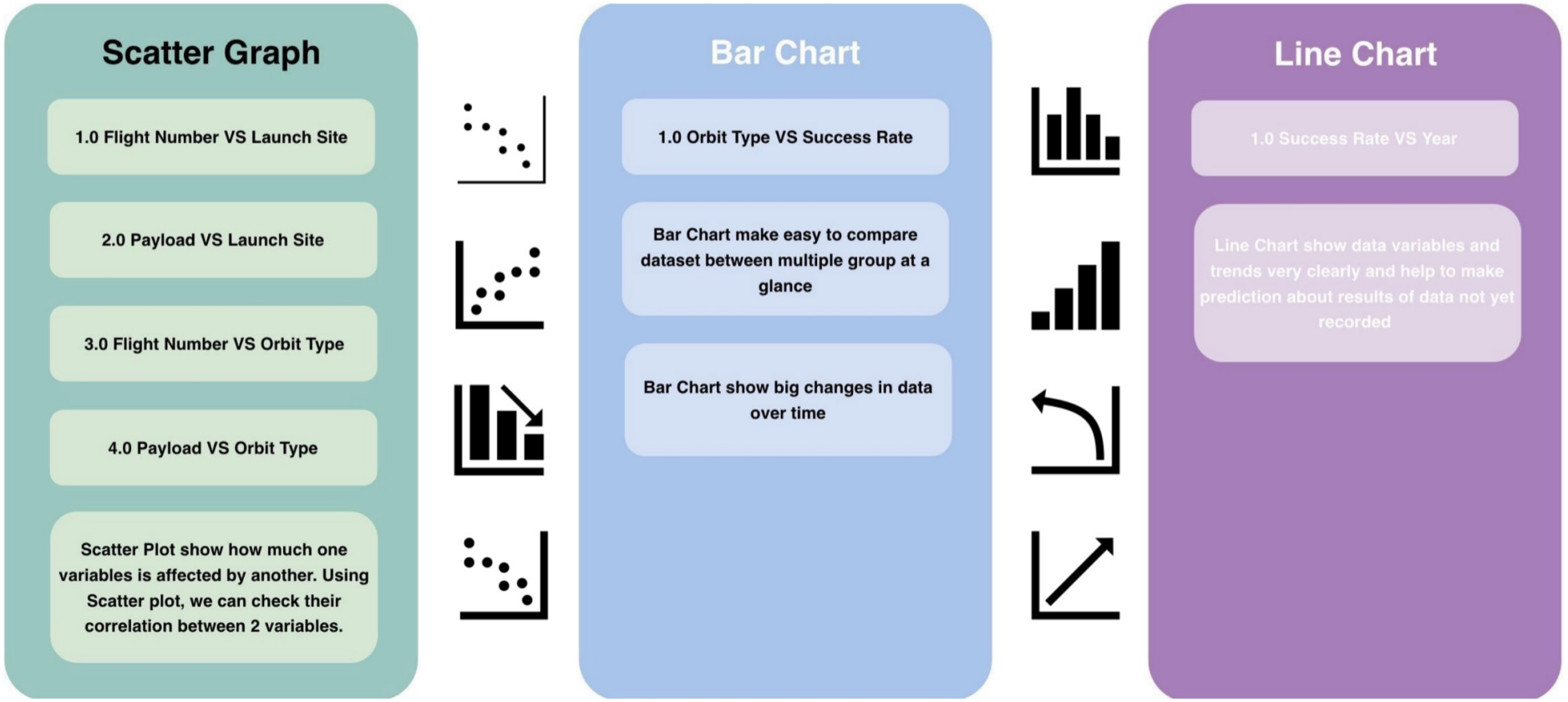
Data Collection – Scraping - Repo



Data Wrangling - Repo



- In the data set, there are several different cases where the booster did not land
- successfully. Sometimes a landing was attempted but failed due to an accident; for
- example,
  - •True Ocean = mission outcome was successfully landed to a specific region of
  - the ocean
  - •False Ocean = mission outcome was unsuccessfully landed to a specific region of
  - the ocean.
  - • True RTLS = mission outcome was successfully landed to a ground pad
  - • False RTLS = mission outcome was unsuccessfully landed to a ground pad.
  - • True ASDS = mission outcome was successfully landed on a drone ship
  - • False ASDS = mission outcome was unsuccessfully landed on a drone ship.
- We will mainly convert those outcomes into Training Labels with 1 means the
- booster successfully landed 0 means it was unsuccessful.



EDA with Data Visualization - Repo

EDA with SQL - Repo

- Performed SQL queries to gather information about the dataset.
- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'KSC’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date where the successful landing outcome in drone ship was achieved.
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass.
- Listing the records which will display the month names, successful landing outcomes in ground pad ,booster versions, launch site for the months in year 2017
- Ranking the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order

Build an Interactive Map with Folium - Repo

Object Created and Added to the Folium Map

Markers that show all launch sites on map

Markers that show the success/failed launches for each site on the map

Lines that show the distances between a launch site to its proximities

By adding these objects, following geographical patterns, about launch sites are found:

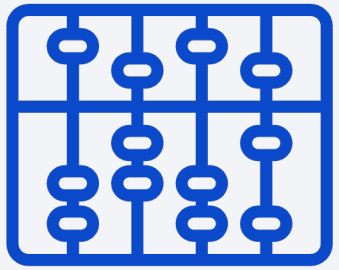
Are launch sites near railways? Yes

Are launches sites near highways? Yes

Are launch sites near coastline? Yes

Do launch sites keep certain distance away from cities? Yes

Build a Dashboard with Plotly Dash - Repo



The dashboard application contains a
pie chart and a scatter point chart.

- For showing total success launches by sites
-This chart can be selected to indicate a successful
landing distribution across all launch sites or to
indicate the success rate of individual launch sites.

-For showing the relationship between Outcomes and
Payload mass (kg) by different boosters.
-Has 2 inputs: All sites/ individual site & Payload mass
on a slider between 0 and 10000 kg
-This chart helps determine how success depends on
the launch point, payload mass, and booster version
categories

Predictive Analysis (Classification) - Repo

Building Model

Load our dataset into Numpy and Pandas

Transform Data

Split our data into training and test data sets

Check how many test samples
Decide on which type of machine learning algorithms to apply

Evaluating Model

Check accuracy for each model

Get tuned hyperparameters foreach type of algorithms

Improving Model

Feature engineering

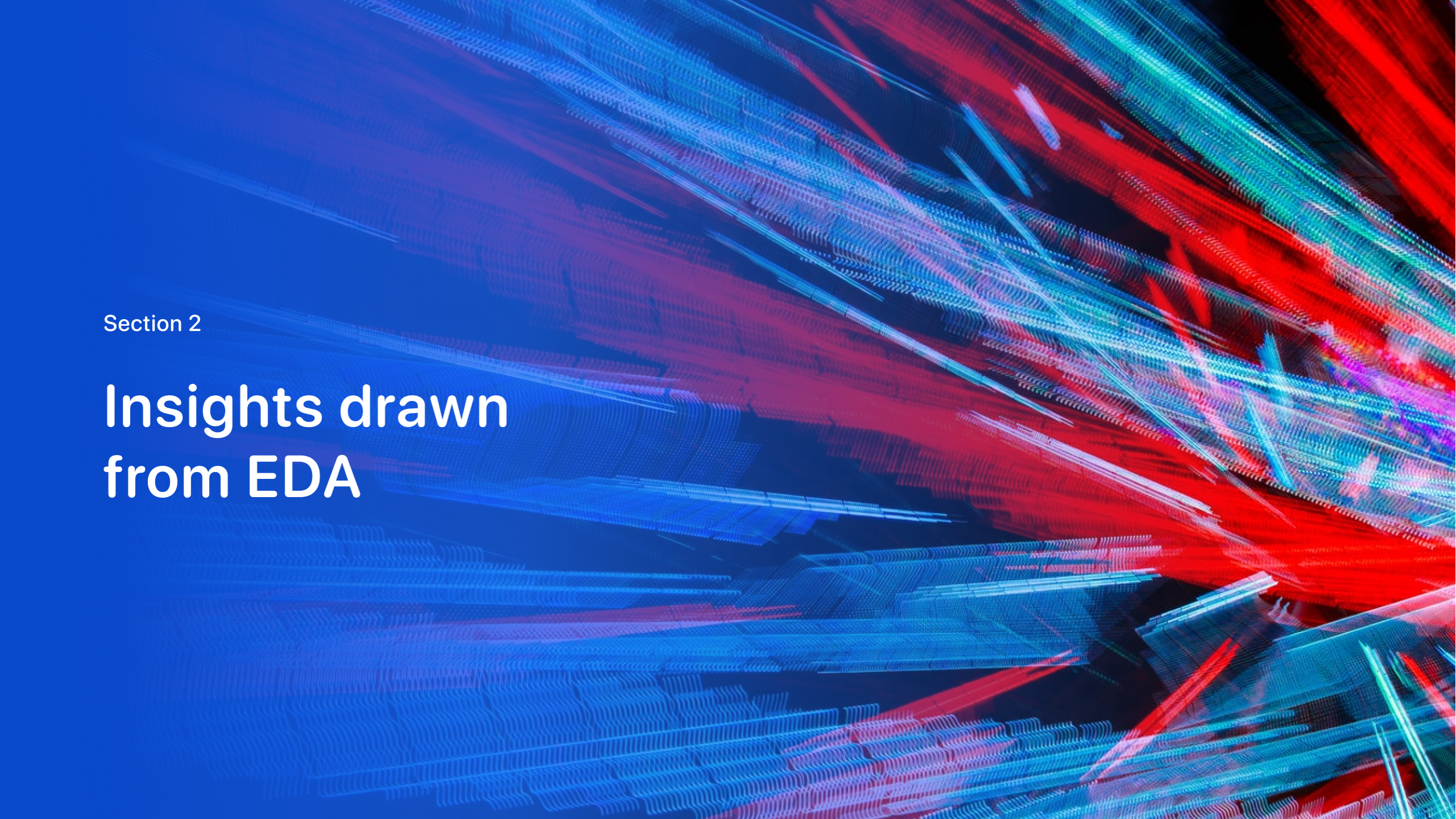
Algorithm Tuning

Finding best algorithm

The model with the best accuracy score
wins the best performing model

Results

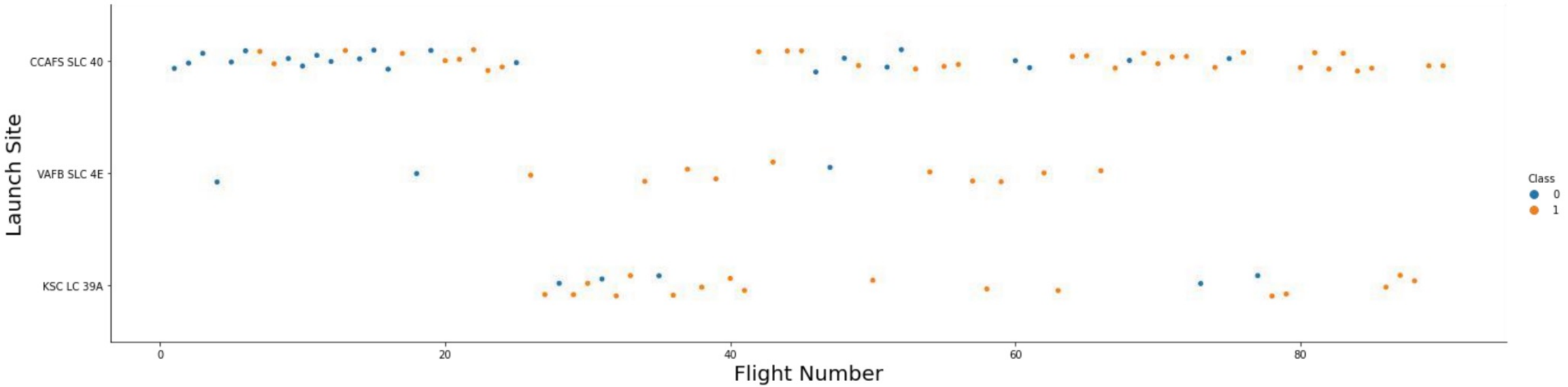
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

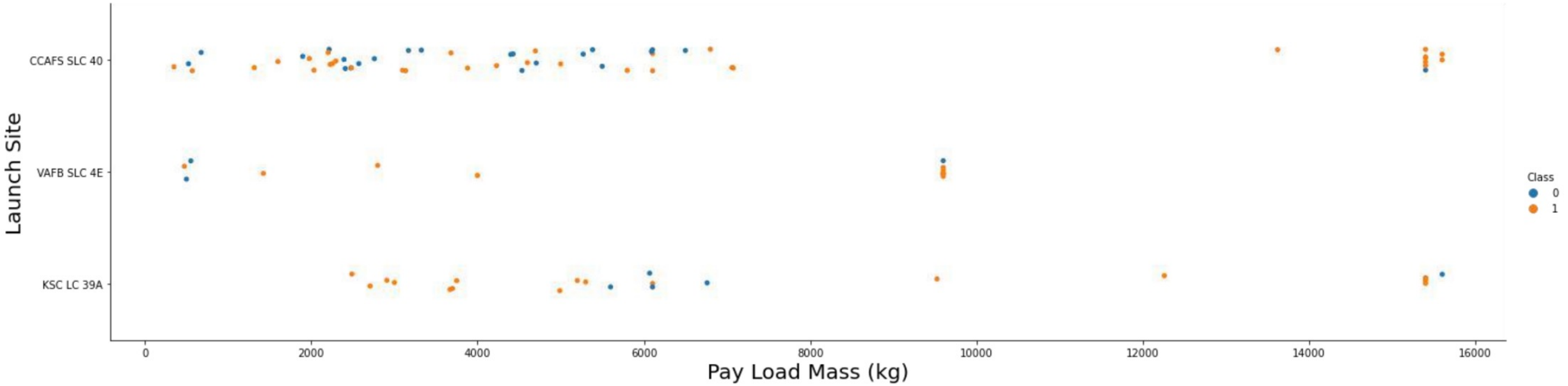
Insights drawn from EDA

Flight Number vs. Launch Site



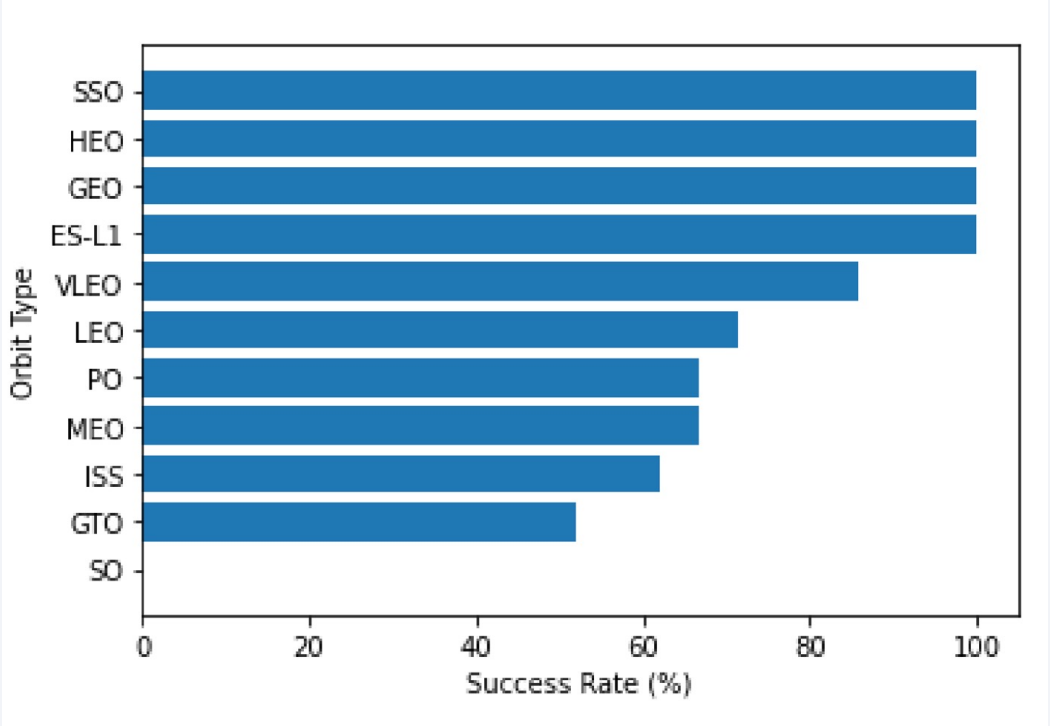
- Class 0 (Blue) represents unsuccessful launch, and Class 1 (orange) represents successful launch.
- This figure shows that the success rate increased as the number of flights increased
- As the success rate has increased considerably since the 20th flights. This point seems to be a big breakthrough

Payload vs. Launch Site



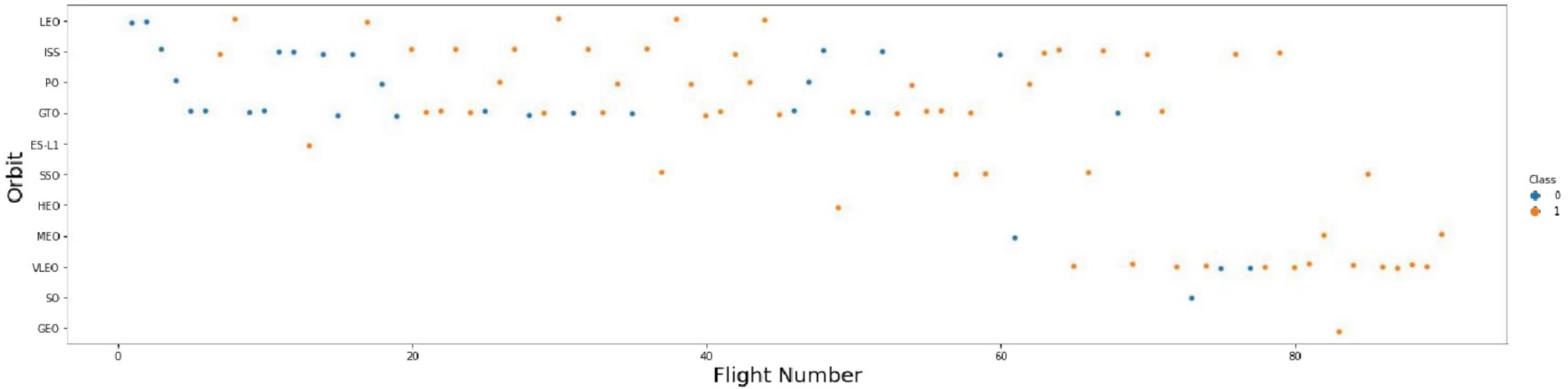
- Class 0 (Blue) represents unsuccessful launch, and Class 1 (orange)represents successful launch.
- At first glance, the larger pay load mass, the higher the rocket's success rate, but it seems difficult to make decisions based on this figure because no clear pattern can be found between successful launch Pay Load Mass

Success Rate vs. Orbit Type



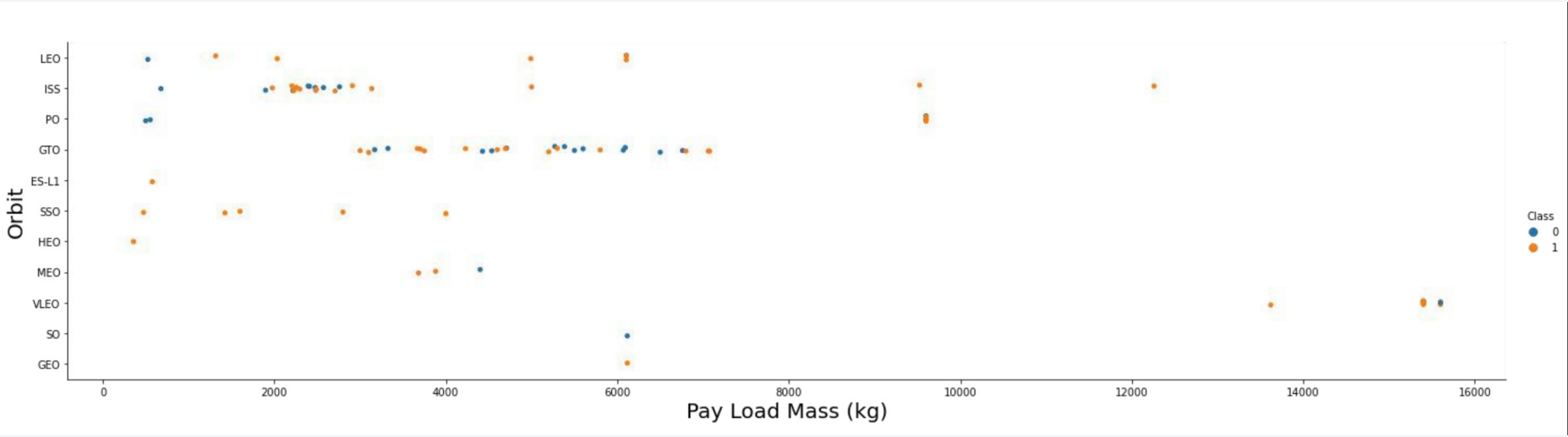
- Orbit types SSO, HEO, GEO, and ES-L1 have the highest success rates (100%)
- On the other hand, the success rate of orbit type GTO is only 50%, and it's the lowest except for type SO, which recorded failure in a single attempt.

Flight Number vs. Orbit Type



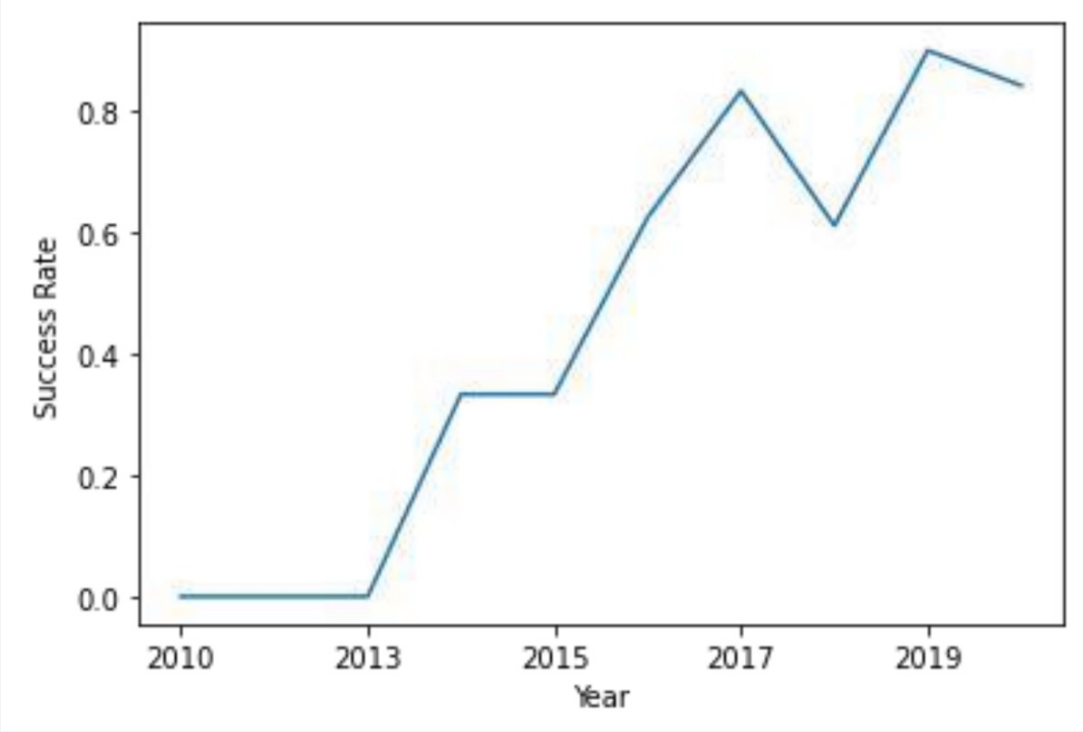
- Class 0 (Blue) represents unsuccessful launch, and Class 1 (orange)represents successful launch.
- In most cases, the launch outcome seems to be correlated with the flight number.
- On the other hand, in GTO orbit there seems to be no relationship between flight numbers and success rate.
- SpaceX starts with LEO with a moderate success rate, and it seems that VLEO, which has a high success rate, is used the most in recent launches

Payload vs. Orbit Type



- Class 0 (Blue) represents unsuccessful launch, and Class 1 (orange)represents successful launch.
- With heavy payloads, the successful landing rate are higher for LEO and ISS.
- However, for GTO case, it is hard to distinguish between the positive landing rate and the negative landing because they are all
gathered.

Launch Success Yearly Trend



- Since 2013, the success rate has continued to increase until 2017
- The rate decreased slightly in 2018
- Recently, it has shown a success rate of about 80%

All Launch Site Names

```
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL
```

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

Query Explaination

Using the word DISTINCT in the query means that it will only show Unique values in the Launch_Site column from SpaceX Table

There are four unique launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT * FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5
```

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 of the SpaceX table were displayed using LIMIT 5 clause in the query
- Using the LIKE operator and the percent sign (%) together, the LAUNCH_SITE name starting with CAA will be called

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS_KG
FROM SPACEXTBL
WHERE CUSTOMER = 'NASA (CRS)'
```

total_payload_mass_kg
45596

- Using the SUM() function to calculate the sum of column PAYLOAD_MASS_KG
- The WHERE Clause filter the dataset to only perform calculations on CUSTOMER NASA (CRS)

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 v1.1'
```

avg_payload_mass_kg
2928

- Using the AVG() function to calculate the average value of column PAYLOAD_MASS_KG
- The WHERE clause filters the dataset to only perform calculation on Booster_version = F9v1.1

First Successful Ground Landing Date

```
SELECT MIN(DATE) AS FIST_SUCCESFUL_LANDING_DATE
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

fist_succesful_landing_date
2015-12-22

- Using the MIN() function to find out the earliest date in column DATE
- The WHERE clause filters the dataset to only perform filtration on LANDING_OUTCOME

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 6000)
```

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

- Selecting only BOOSTER_VERSION
- The WHERE clause filters the dataset to LANDING_OUTCOME = Success (drone ship)
- The AND and BETWEEN clause specifies additional filter condition PAYLOAD_MASS_KG BETWEEN 4000 AND 6000

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT MISSION_OUTCOME, COUNT(*) AS total_number
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME
```

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

- Using the COUNT() function to filter the total number of columns
- Using GROUP BY function to group rows that have same values into summary rows to find the total number in each MISSION_OUTCOME
- SpaceX successfully completed nearly 99% of its mission based on the dataset

Boosters Carried Maximum Payload

```
%%sql
SELECT DISTINCT BOOSTER_VERSION, PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL);
```

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

- Using a subquery, find the maximum value of the payload using MAX() function, and then filter the dataset to perform search IF PAYLOAD_MASS_KG_ is the maximum value
- From the result, F9 B5 B10xx.x boosters carried the maximum payload

2015 Launch Records

```
%%sql
SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND YEAR(DATE) = '2015'
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- In the WHERE clause, filter the dataset to perform a search if Landing_Outcome is Failure(drone ship)
- Use AND operator to display a record if YEAR is 2015
- There were two landing failure son drone ships in 2015

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

```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS total_number
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY total_number DESC
```

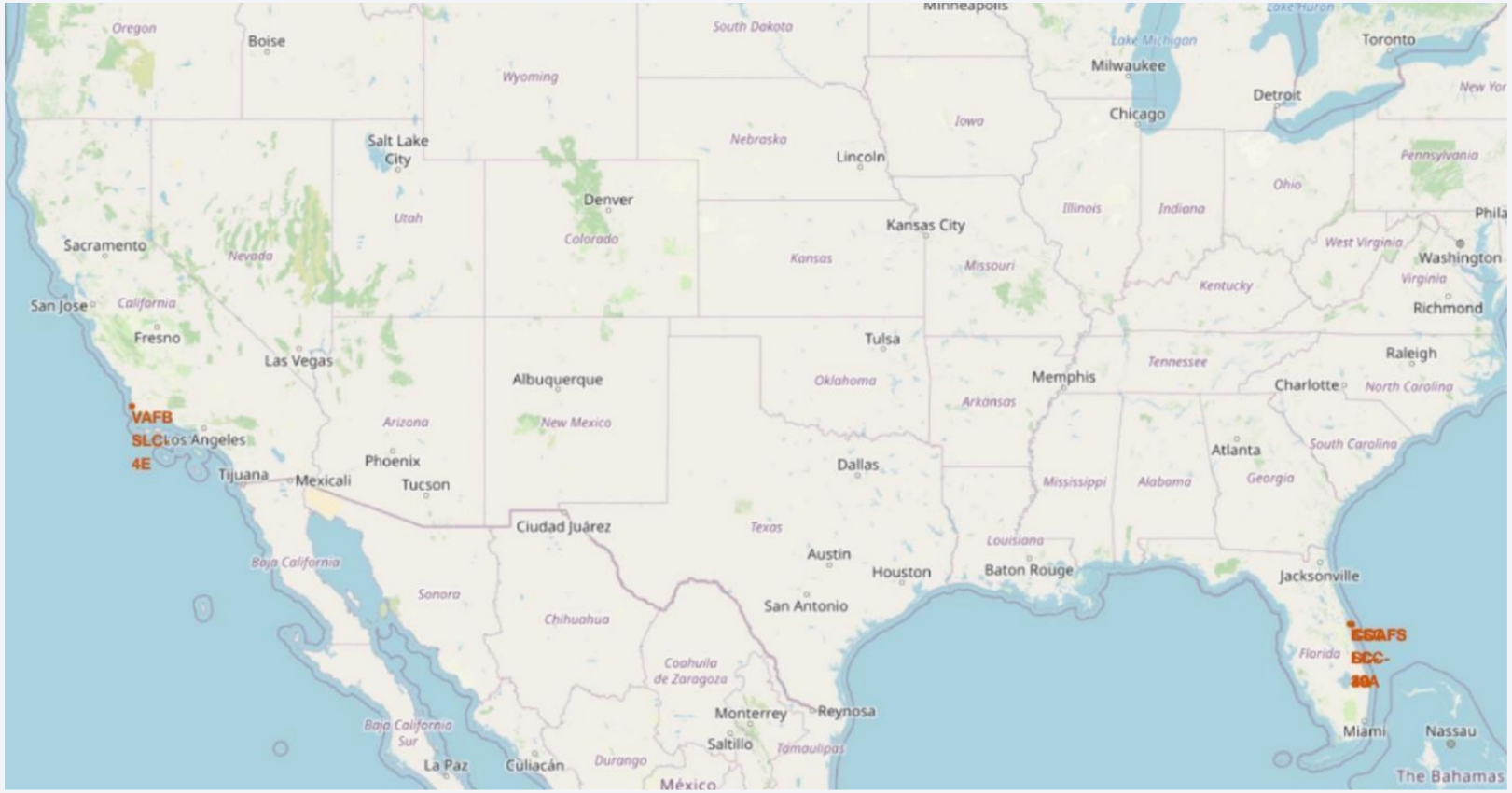
landing__outcome	total_number
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- In the WHERE clause, filter the dataset to perform search for DATE between 2010-06-04 and 2017-03-20
- Using ORDER clause to sort the records by total number of landing and DESC clause to sort the records in descending order.

Section 4

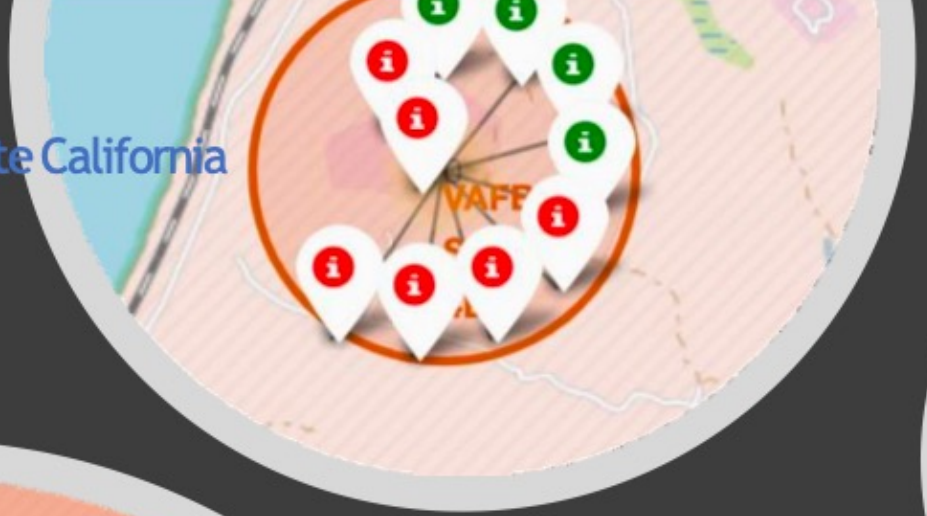
Launch Sites Proximities Analysis

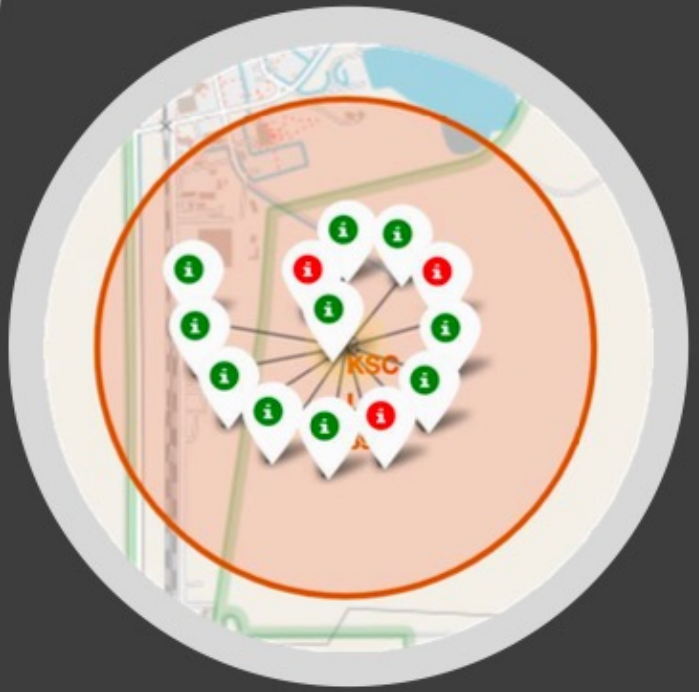
All Launch Site Location

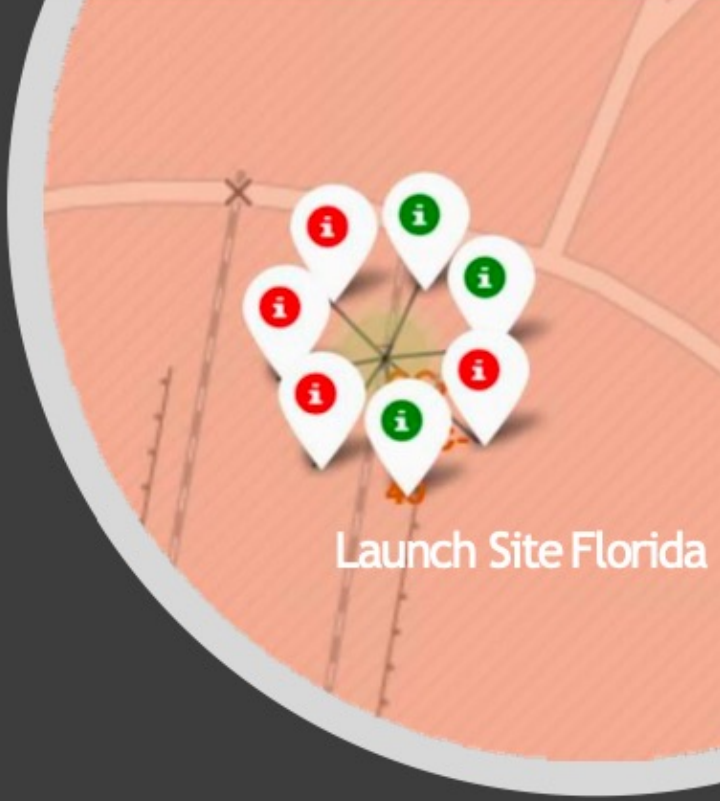


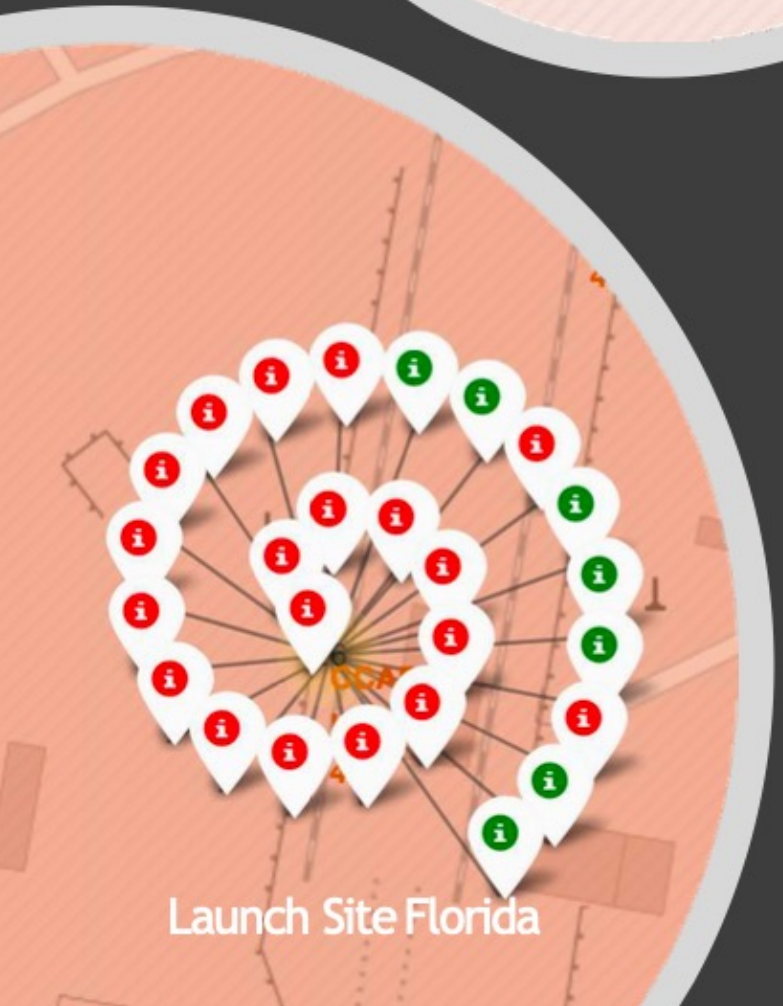
Most of the SpaceX launch site are in US coast area

Florida and California

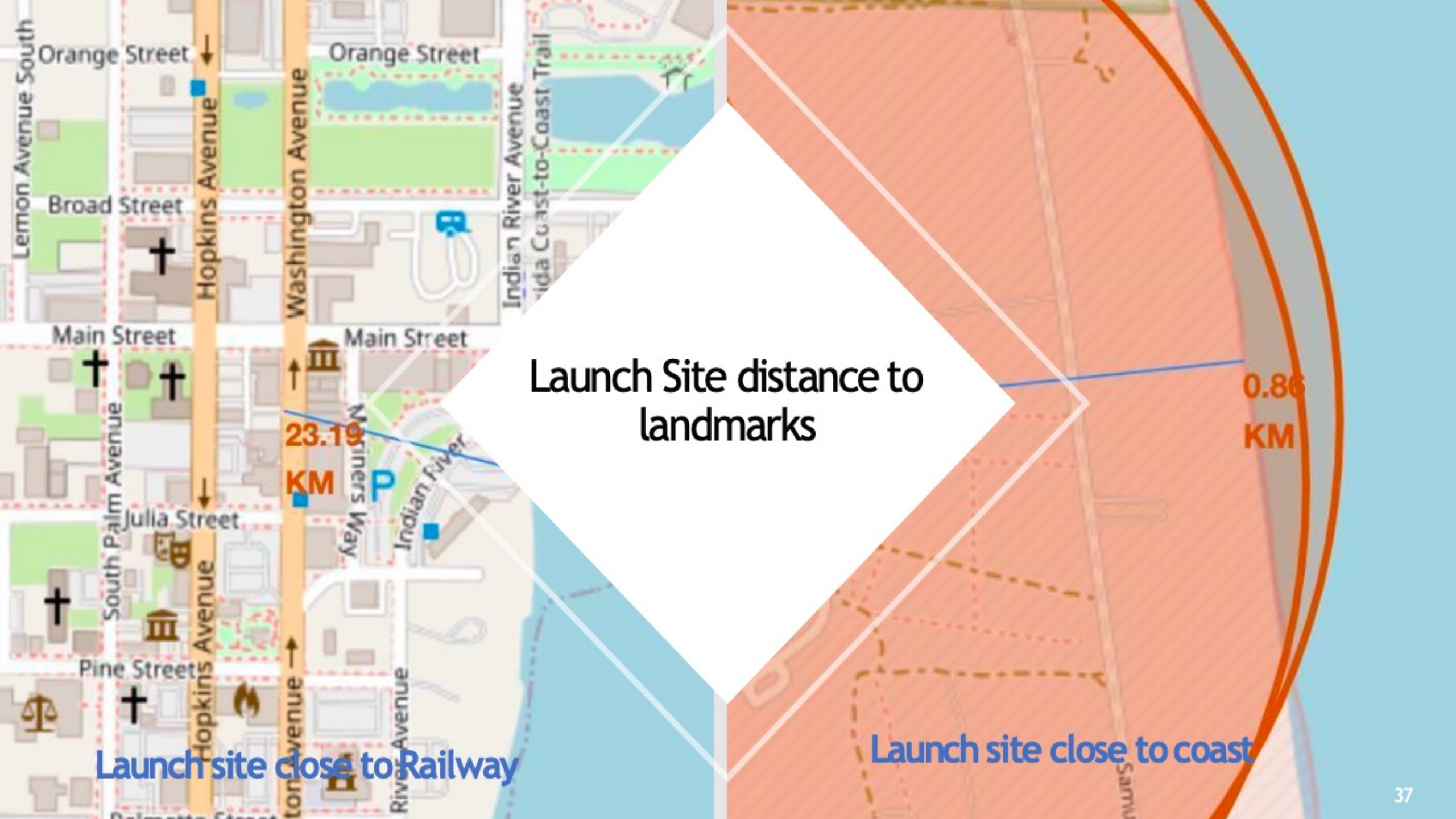




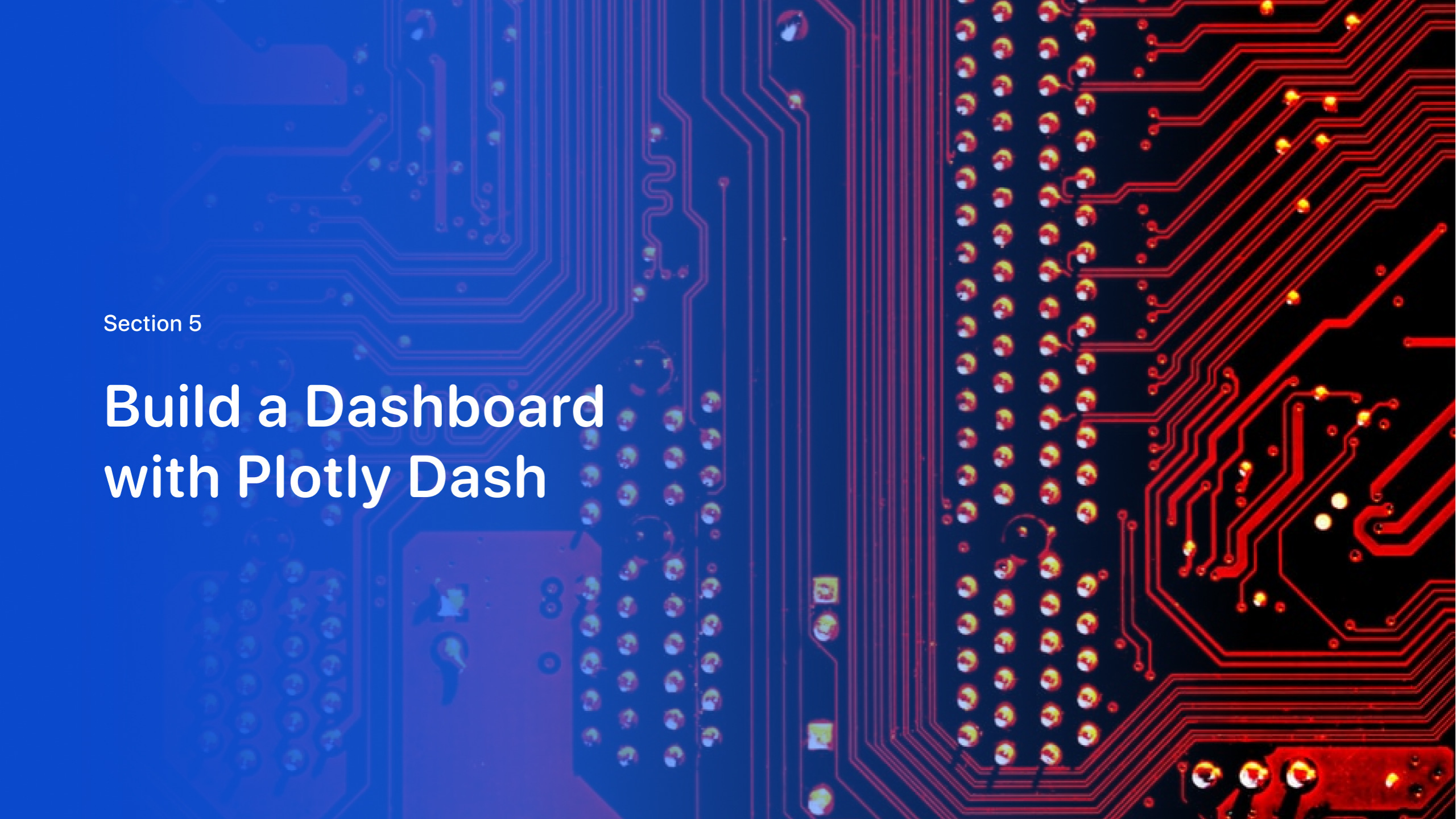




Colour Labelled Markers



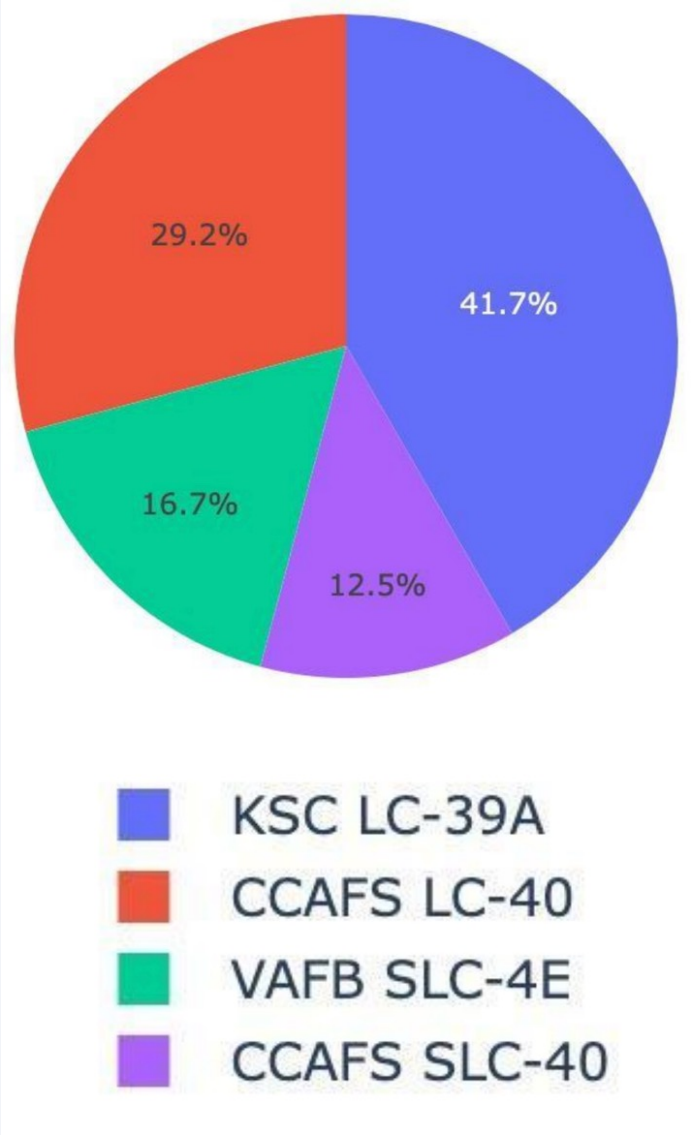
Launch Site distance to landmarks



Section 5

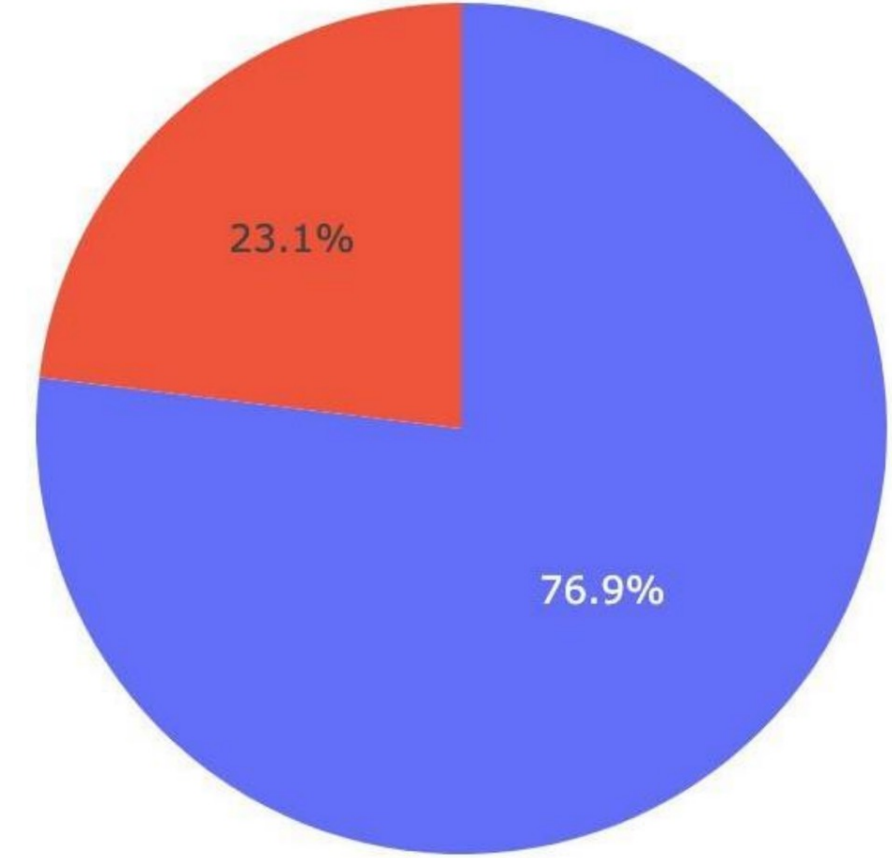
Build a Dashboard with Plotly Dash

Total Success Launches by allSites



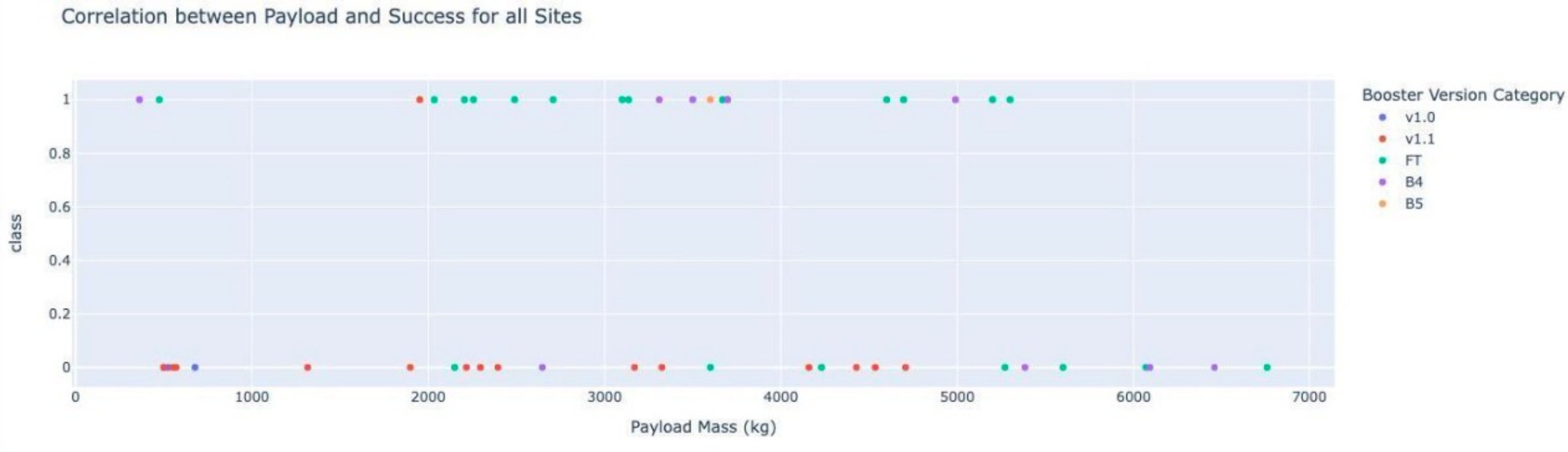
- KSLC– 39A records the most launch success among all sites.
- VAFB SLC-4E has the lowest success launch

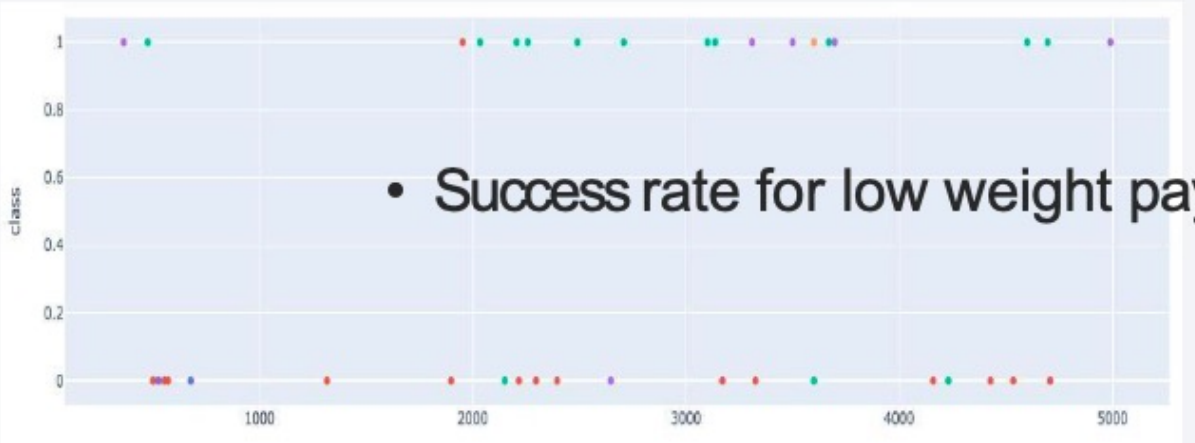
Launch Site with Highest launch Success Ratio

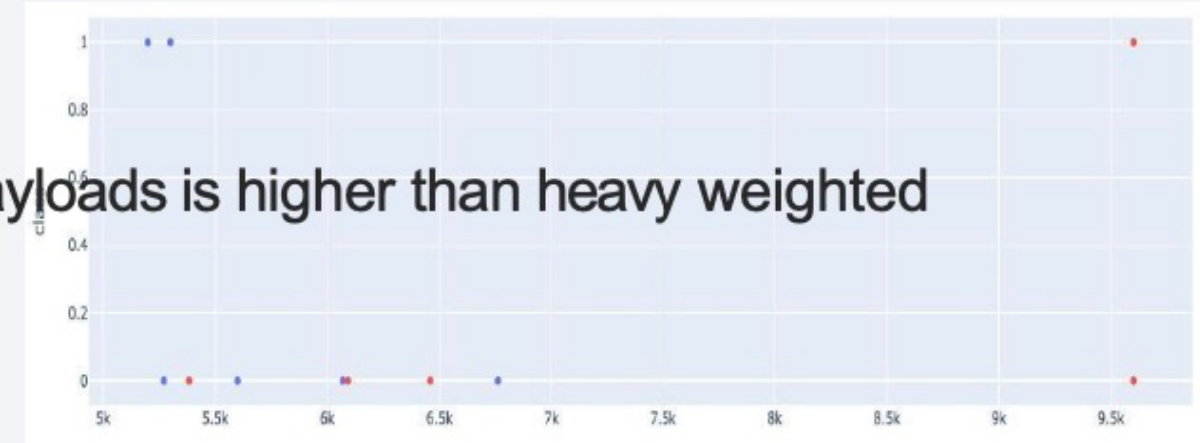


- KSC-LC-39A achieved a 76.9% success rate with total of 13 landing

Payload VS Launch Outcome Scatter Plot for all Sites





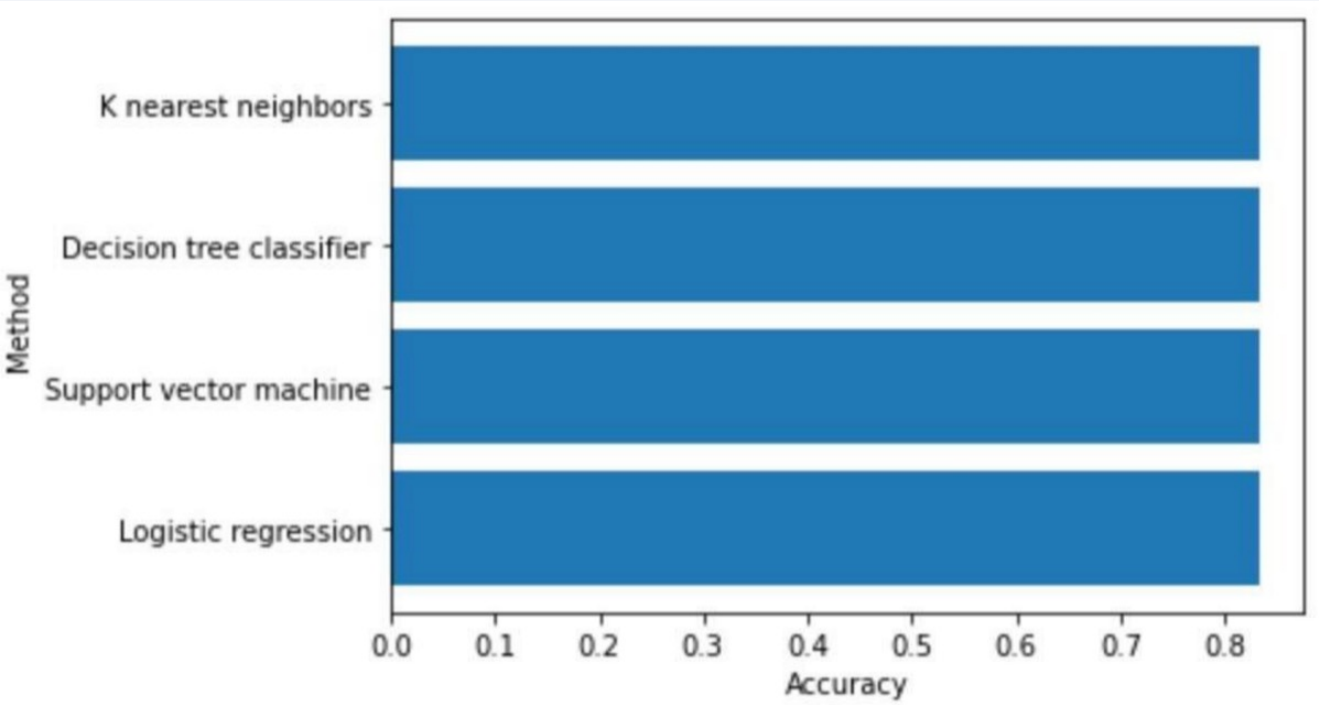


- Success rate for low weight payloads is higher than heavy weighted

Section 6

Predictive Analysis (Classification)

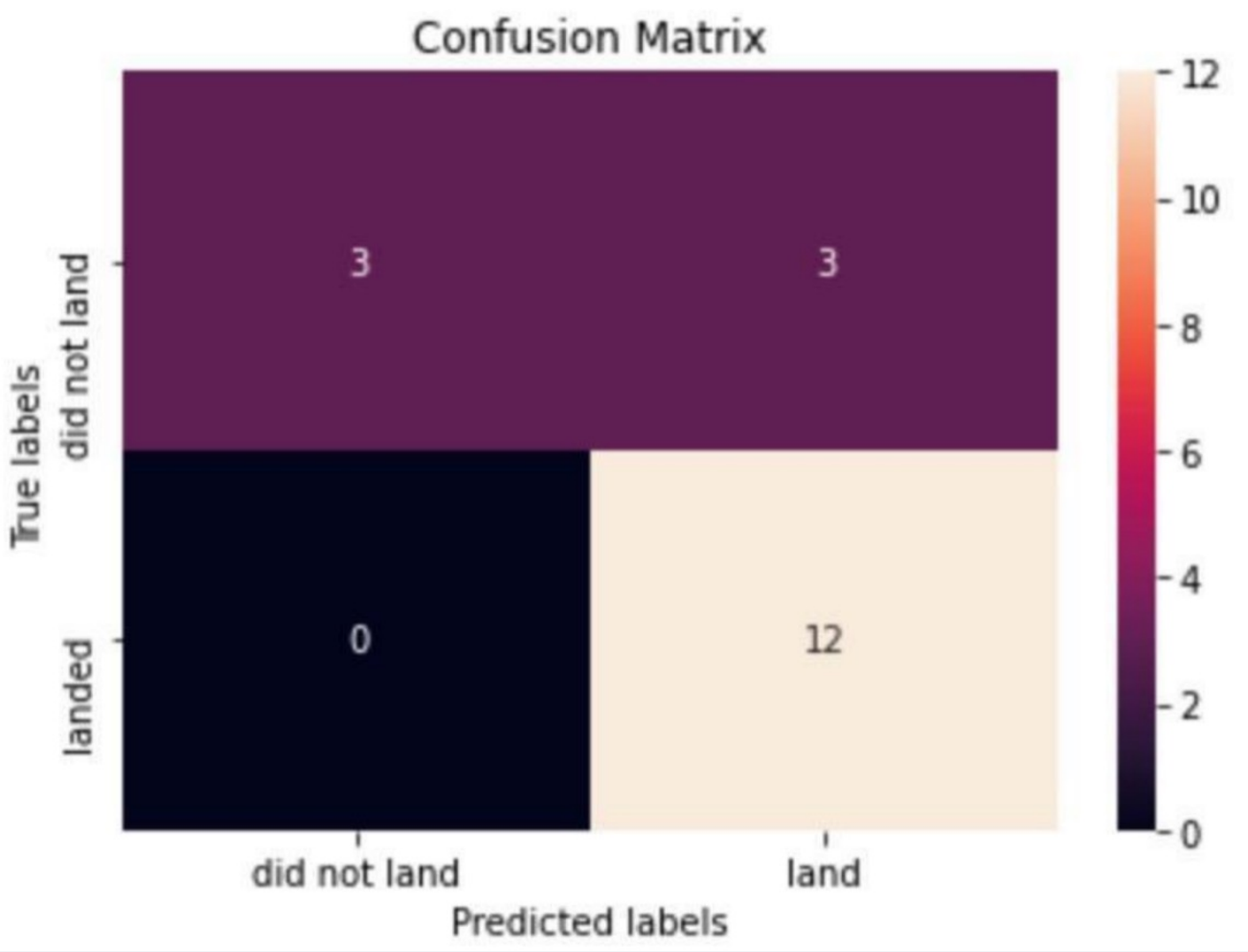
Classification Accuracy



- In the test set, the accuracy of all models was virtually the same at 83.33%

- More data is needed to improve the model

Confusion Matrix



- The confusion matrix is the same for all models
- The models predicted 12 successful landing when the true label was successful.
- The model predict 3 failed landing while the actual label was not successful.
- Overall, the model predict quite good at successful landings.

Conclusions

ORBITAL TYPES SSO, HEO,GEO, AND ES-L1 HAVE THE HIGHEST SUCCESS RATE(100%).

KSLC-39A HAS THE HIGHEST NUMBER OF LAUNCH SUCCESSES AND THE HIGHEST SUCCESS RATE AMONG ALL SITES.

LOW WEIGHTED PAYLOADS PERFORM BETTER THAN THE HEAVIER PAYLOADS.

IN THIS DATASET, ALL MODELS HAVE THE SAME ACCURACY (83.33%), BUT IT SEEMS THAT MORE DATA IS NEEDED TO DETERMINE THE OPTIMAL MODEL DUE TO THE SMALL DATA SIZE.


10916.69
10847.17
Buy
OMXRGI
OMX RIGA GI
6025.9680
5993.7030
599.40
28289.
OMX18
OMX ICEL

Appendix

Thank you!