IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yash Rao
15th August 2023

Outline

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

Executive Summary

- The project deals with prediction of landing of Falcon 9 stage one component based on factors such as Payload Mass used in the rocket, booster version installed , launch pad used, etc. Data acquisition for the project is implemented using SpaceX API as well as scraping data from Wikipedia pages that list landings of Falcon rocket launches.
- Data collected was then modified to deal with missing values in extracted features and the target variable was encoded. Proper data visualizations using various charts and Folium maps was performed, **showing successful landings from KSC LC - 39A.**
- Data exploration was implemented as well using python and SQL to understand relationships in between features and target variable, showing an increase in successful landings post 2013. A dashboard was also created for a more enhanced view into relationships between successful landings, booster versions, launch pad and payload mass. **Exploring the data displayed successful landings attributed to launch years greater than 2017 and lighter payloads (in between 2000 – 4000 kgs).**
- Categorical features were encoded using one-hot encoding. Finally, a predictive analysis was performed between classifiers like KNN, Decision Trees, Logistic Regressors, etc. and performance metrics were evaluated. **Data was split into training and testing sets, with competitive accuracy of 83.33% on test sets.**

Introduction

SpaceX, an organization that manufactures and controls spacecraft machinery, has recently been in the news. Falcon series spacecrafts and rockets are one of its most prominently featured products, with Falcon 9 being an answer to the costly problem of space launches, costing only $62 million per launch. This is particularly due to recovery of its stage one component, leading to its reusage on other rockets, drastically reducing overall manufacturing costs.

A successful recovery looks something like this -

Introduction (Cont.)

However, not all recoveries are successful due to -

- Engine failure
- Incorrect coordinates for landing
- Defective recovery pads
- Fuel shortages, etc.

Hence, there is a need to understand factors that contribute to successful landings and predict the same based on these factors.

Introduction (Cont.)

The project aims to answer questions such as –

- What factors contribute to successful landings?
- What are the relationships in between these factors?
- Are there any modern algorithms that can predict successful landings based on these factors? If so, how accurate are they based on the landing history of Falcon 9?

These questions were successfully answered using multiple data science methodologies like Data Collection, Data Wrangling, Feature Engineering, Data Analysis, Data Visualization and Predictive Analysis using Machine Learning.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX API
  - Web Scraping from Wikipedia page documenting Falcon 9 landings.
- Perform data wrangling
  - Assessing data for null values and appropriately dealing with them
  - Categorizing target values for EDA
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Splitting data into training and testing sets
  - Training models (KNN, Decision Trees, SVM, Logistic Regression) on training set
  - Evaluating trained models using accuracy obtained on test sets.

Data Collection

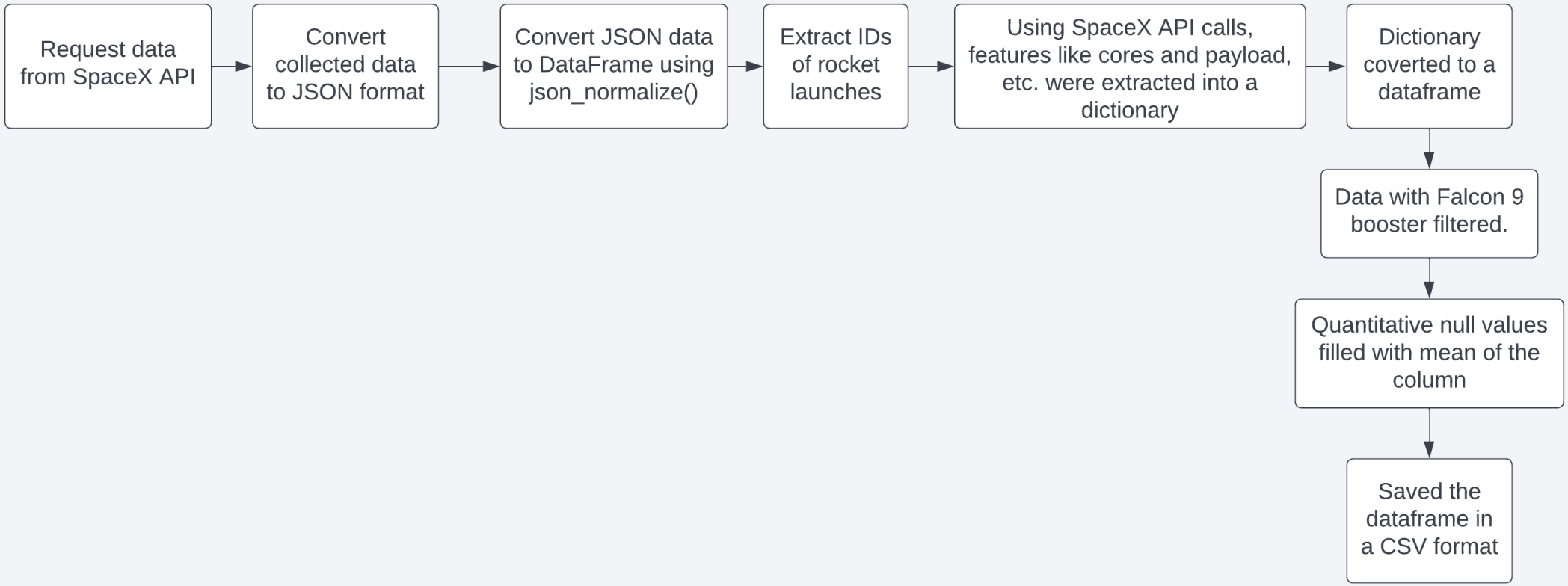
- Data was collected in the following manner –
  - SpaceX API –
    - Data collected by requesting data using SpaceX API.
    - Data was cleaned afterwards
  - Web Scraping –
    - Data scraped from Wikipedia page documenting Falcon 9 recoveries
    - Convert scraped data into format that is easily accessible for further processes

Data Collection – SpaceX API

- SpaceX API provides information on booster versions, rocket names, payloads used, dates of launch, details of launch, etc. Using the link https://api.spacexdata.com/v4/launches/past, data was collected and parsed.
- Necessary details like booster versions, cores, names of rockets, latitude, longitude, date and time of launch, etc. were collected. This was done using the request and response methods imported from “requests” library. The content recovered was parsed to a JSON (JavaScript Object Notation) file.
- The JSON file is then converted to a Dataframe using json_normalize(json_data) method. Features were collected and using API calls (https://api.spacexdata.com/v4/rockets/, https://api.spacexdata.com/v4/launchpads/, etc.) to IDs from the previously obtained Dataframe. The collection was stored as a dictionary.
- The dictionary obtained was finally converted to a Dataframe, out of which Falcon 9 boosters' data were selected. Quantitative Null values in the dataset were dealt with by replacing NaN values with the mean of that column. The cleaned data was saves as a csv file.

Data Collection – SpaceX API (Cont.)

The link to notebook is provided here. A flowchart is also provided below -

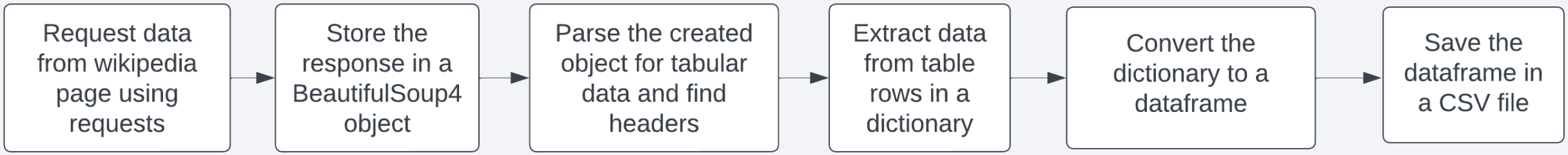


Data Collection – Scraping

- Another way for data collection was by using Web Scraping. The primary data used was extracted from a Wikipedia page (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922) documenting launches of Falcon 9 over the years.
- Data was extracted using BeautifulSoup4 object, which aided in converting the text response into a HTML object. Headers were extracted from the table itself.
- After parsing and iterating over all table rows, data found was stored in a dictionary and keys were the headers extracted from the table itself (with splitting date to date and time of the launch).
- This dictionary was converted to a Dataframe. Finally, the Dataframe was saved in a CSV file.

Data Collection – Scraping (Cont.)

The link for web scraping notebook is provided here. A flowchart is displayed as well -

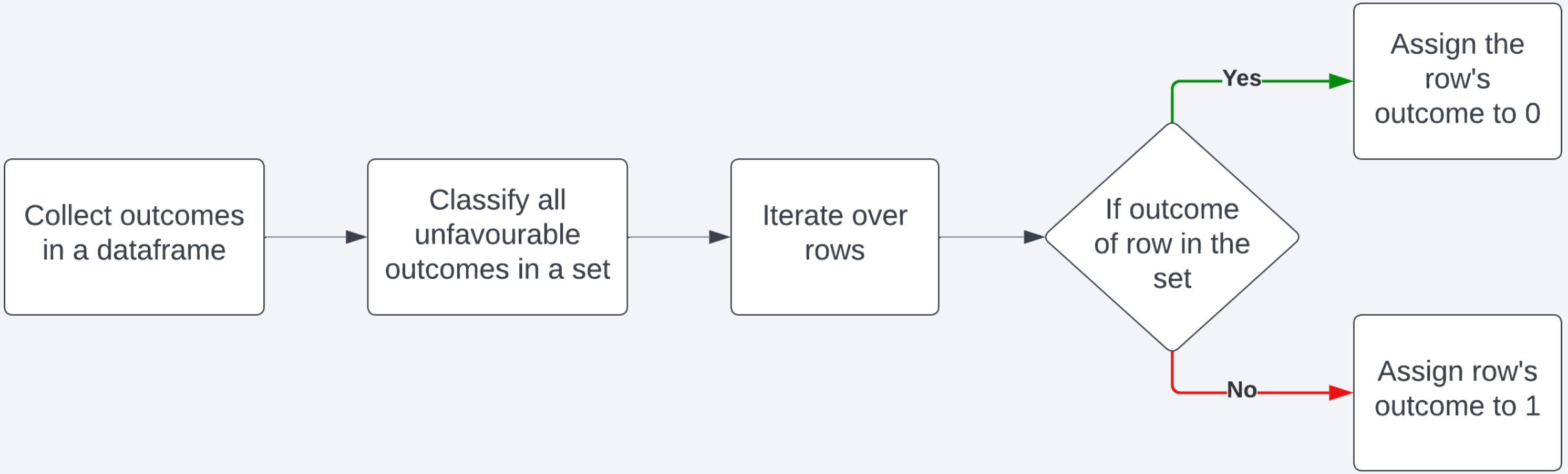


Data Wrangling

- After collecting data, the next step is to understand data collected and clean it for data analysis and further processes. Data regarding orbits was collected. In addition, outcomes for the data was divided into various classes. A description for the same is presented below (bold ones are considered good outcomes) –
  - **True Ocean – Recovery successful in ocean**
  - False Ocean – Recovery unsuccessful, landing compromised in ocean
  - **True RTLS – Recovery successful in ground pad**
  - False RTLS – Recovery unsuccessful, landing compromised in a ground pad
  - **True ASDS – Recovery successful in a drone ship**
  - False ASDS – Recovery unsuccessful, compromised in a drone ship
  - None ASDS; None None – Recovery unsuccessful due to undisclosed reasons
- In order to understand good outcomes and distinguish them from the unfavorable ones, the class variable was encoded in a way that –
  - Bold outcomes in the list were assigned a value of 1
  - Other outcomes were assigned a value of 0

Data Wrangling (Cont.)

- Notebook displaying data wrangling is provided here. A flowchart for encoding outcomes is also provided below –



EDA with Data Visualization

- Multiple charts were plotted in order to understand relationships between features present in the data –
  - Flight Number VS Launch Site
  - Flight Number VS Orbit
  - Orbit Success Rate
  - Flight Number VS Payload
  - Launch Site VS Payload
  - Orbit VS Payload
  - Success rate per year
- Link for the notebook is presented here.
- A link to folder containing all visualizations is present here.

EDA with SQL

- EDA using SQL queries provides users with an easy way to understand data. Various SQL queries were implemented –
  - Displaying names of launch sites
  - Displaying sits with start with “CCA”
  - Calculating total payload used by NASA CRS
  - Calculating average payload carried by booster “F9 v1.1”
  - Displaying the date when first successful landing outcome in a drone ship occurred
  - Understanding boosters successfully landing in drone ships with payloads in between 4000 and 6000
  - Counting total successful and failed missions
  - Displaying booster versions that have carried maximum payloads
  - Understanding data for failures for missions occurring in 2015
  - Understating total count of each landing outcome in between 4th June 2010 and 20th March 2017
- The link for the notebook is presented here.

Build an Interactive Map with Folium

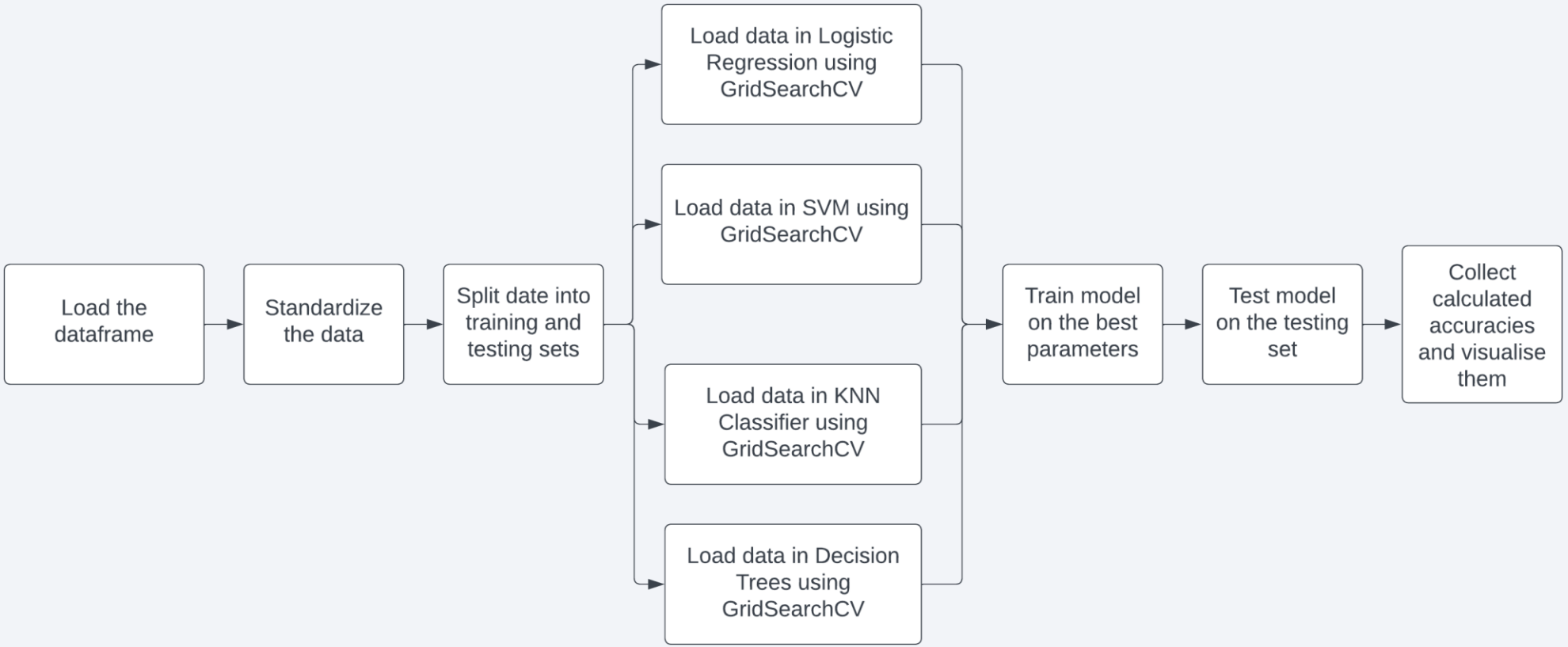
- This project was further able to visualize locations wherein launch sites were present and was able to derive important insights. This was performed using Folium library -
  - Launch sites on the map were marked by a red semi-transparent circle
  - Markers aided in understanding successful and unsuccessful launches from a particular site. Red markers indicated unsuccessful launches, while Green marked the successful launches.
  - Distances and lines helped visualize distance of these sites from coastlines, highways, railway lines and cities
- The link for the same is provided here (Folium maps are not supported by GitHub)

Build a Dashboard with Plotly Dash

- A dashboard was prepared using Plotly and Dash. Various charts were prepared in order to understand relationships between booster versions, payloads and landing outcomes –
  - Pie charts were created to understand relationships between booster versions and successful launches (which was updated depending on the booster version selected)
  - A scatter plot was created to understand relationship in between payload mass and landing outcome (the chart updated based on the payload mass range selected).
- The link is presented here.

Predictive Analysis (Classification)

- A flowchart is presented below, summarizing the process for predictive analysis –



- A link for the notebook is presented here.

Results (EDA)

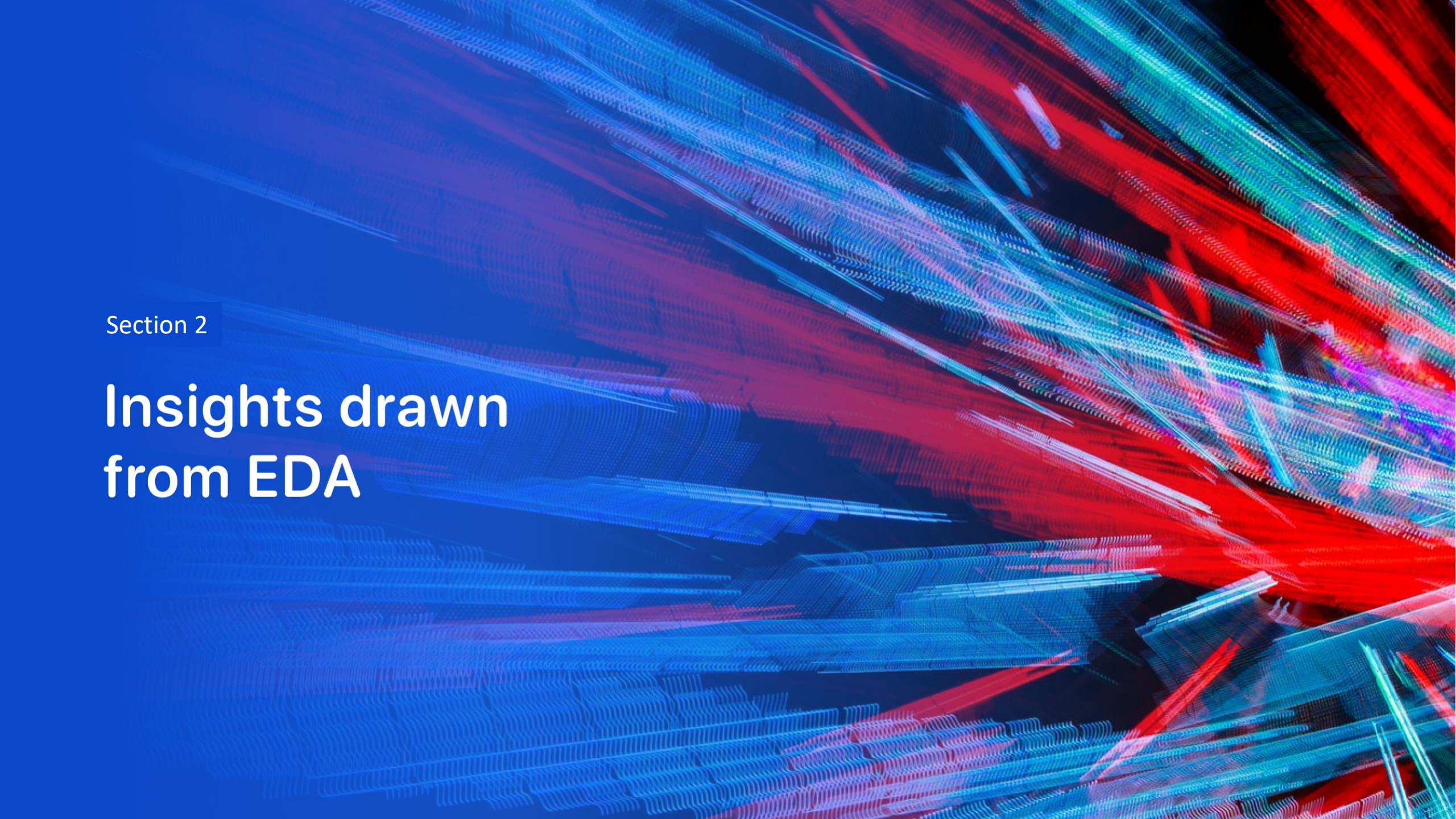
- Results from EDA speak –
  - Successful landings for Falcon 9 has been in a rise from 2013, with a depression in between 2017 – 2018 period, with recovery of the same being faster in further years
  - Light payload operations have a high chance of recovery (particularly payload masses in between 2000 – 4000 kgs)
  - Launch Site KSC LC-39A has the highest success rate for successful landings (around 75%)
  - Recovery methods involving drone ships have the highest success rate among all the successful landings
- Visualizations are presented in the next section.

Results (Dashboard)

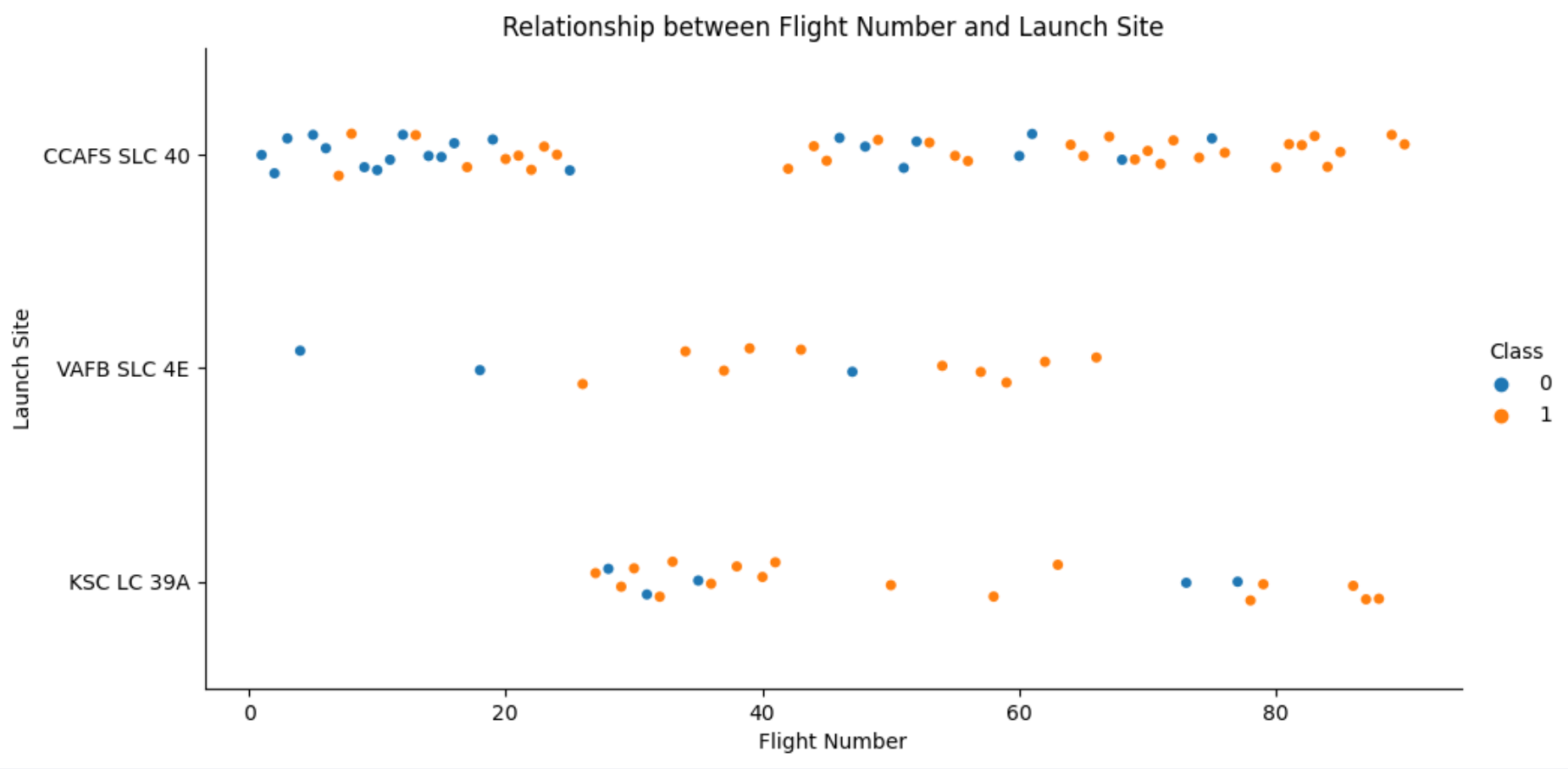
- Dashboard results are provided in section 4, with the link provided in slide 19.

Results (Predictive Analysis)

- Predictive Analysis shows best performing models represent 83.33% accuracy.
- Confusion Matrix shows high numbers of false positives

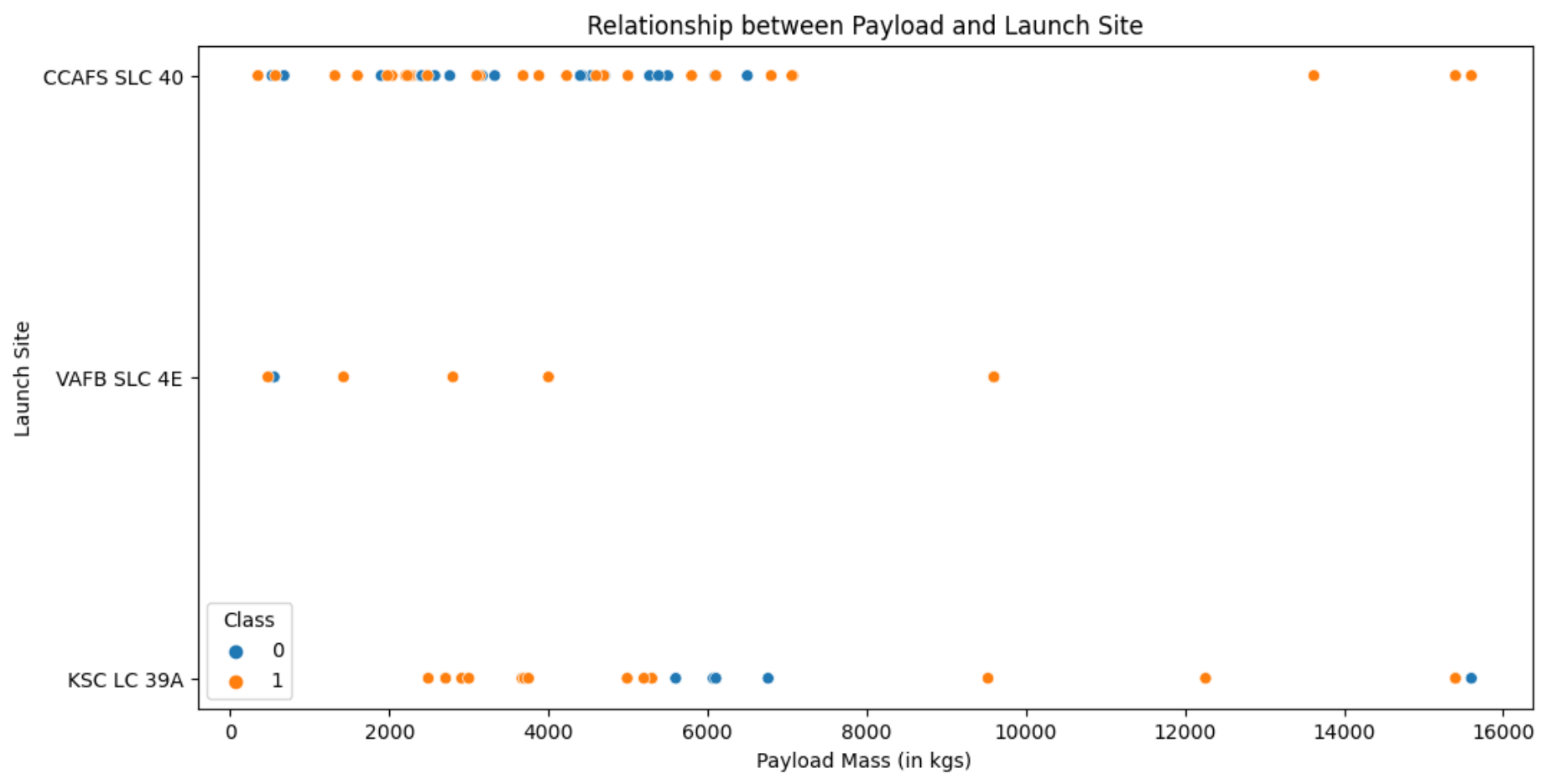

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



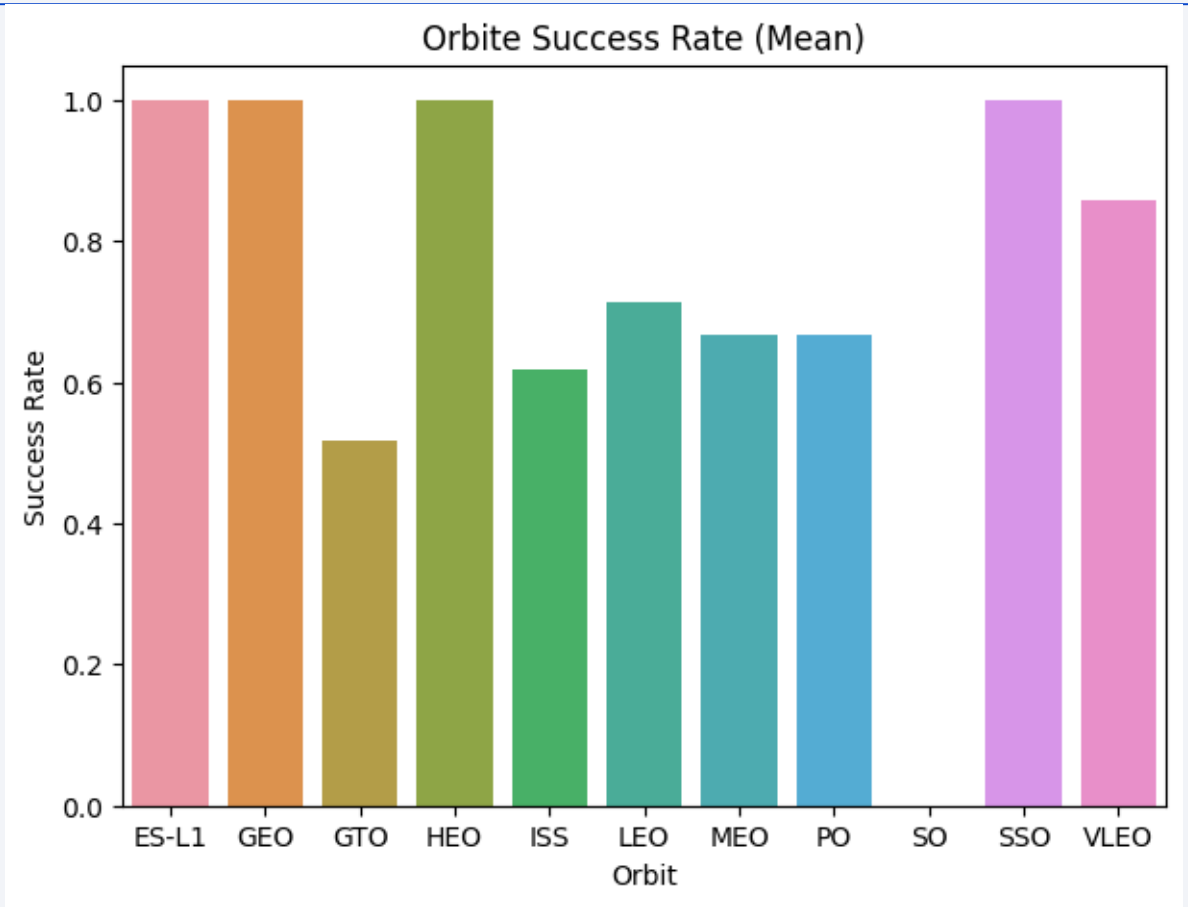
- KSC LC 39A shows the highest rate of successful landings, with it being the most consistent since it has appreciable data points with high success

Payload vs. Launch Site



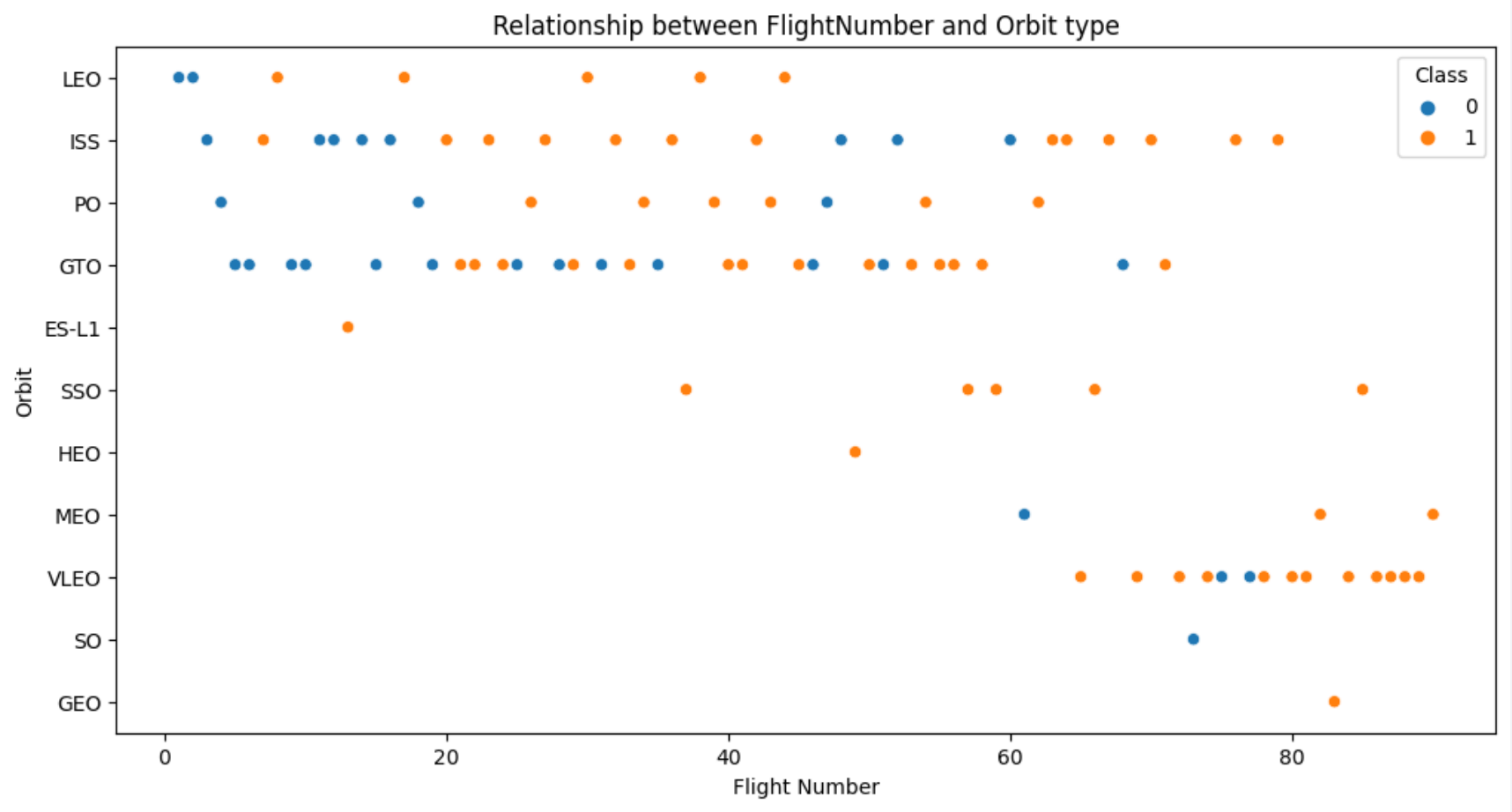
- Payloads less than 4000 kgs have the highest successful landings, with most of it being in KSC LC 39A site

Success Rate vs. Orbit Type



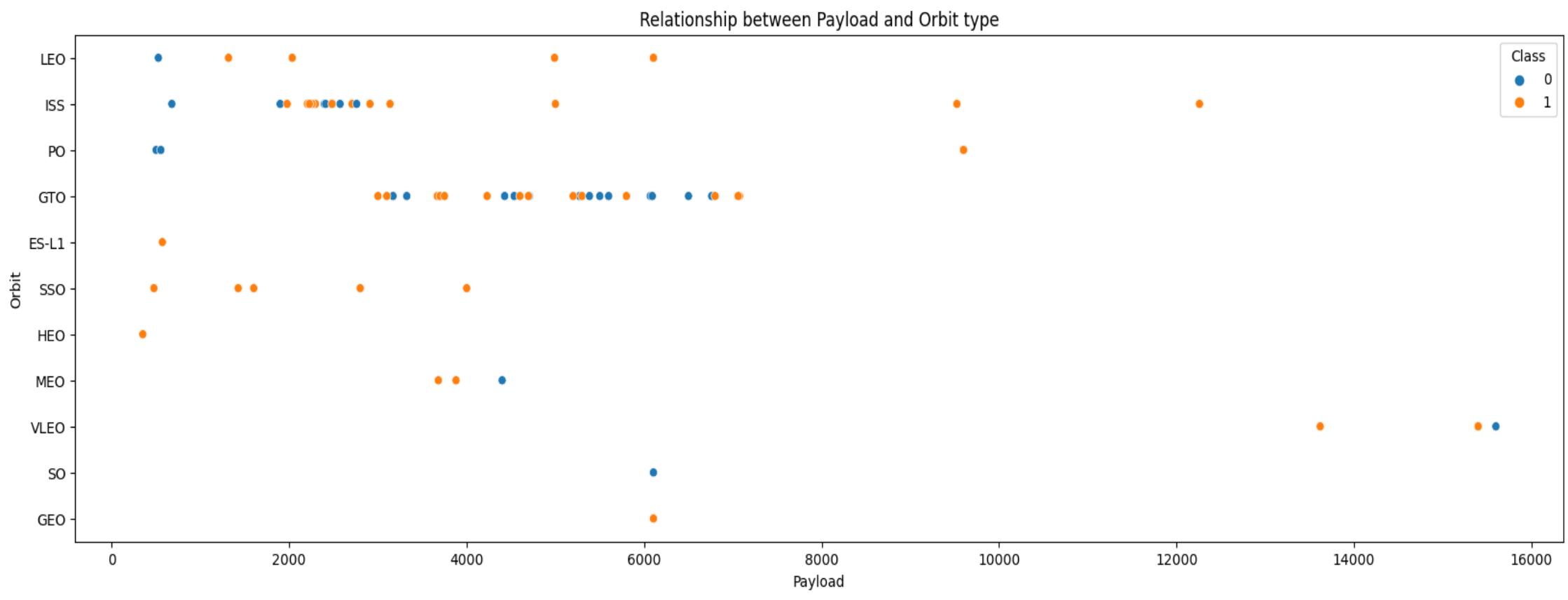
- Orbits ES-L1, GEO, HEO and SSO have the highest success rate (almost 100%).
- Orbit SO has the no success in stage one recoveries

Flight Number vs. Orbit Type



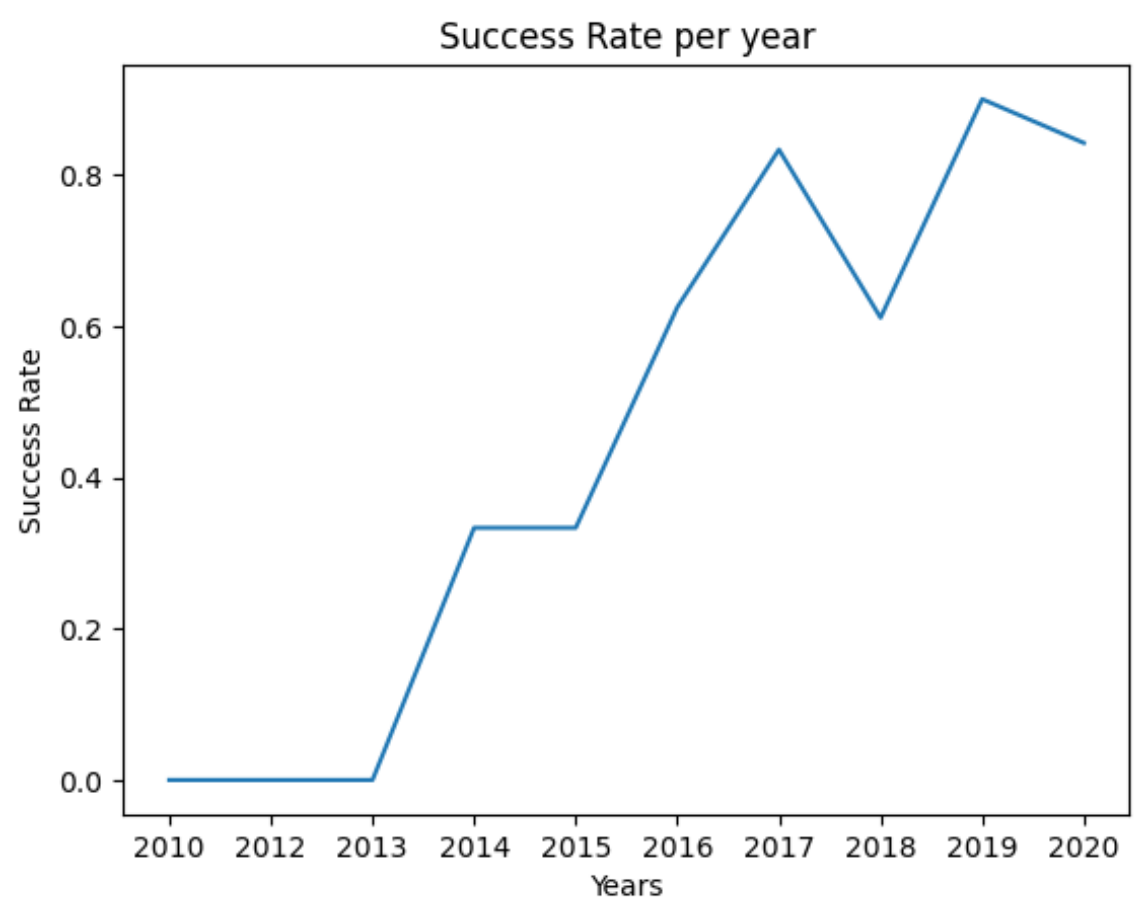
- Looking at the graph, the most consistent highest success rate orbit is SSO with appreciable data points in it.
- Next to it, LEO and VLEO seem to be the most consistent.

Payload vs. Orbit Type



- SSO, LEO and VLEO have less payloads, attributing to their successful landings.

Launch Success Yearly Trend



- Years further 2013 seem to have increased successful landings (exception being the 2017 – 2018 period)

All Launch Site Names

- Find the names of the unique launch sites

```
%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'

- Find 5 records where launch sites begin with `CCA`

```
%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
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA (CRS)

```
%sql select sum("PAYLOAD_MASS__KG_") as "Toal Payload Mass by NASA (CRS)" from SPACEXTBL where "Customer" = "NASA (CRS)";
```

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

Toal Payload Mass by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

```
%sql select avg("PAYLOAD_MASS__KG_") as "Average Payload Mass by F9 v1.1" from SPACEXTBL where "Booster_Version" = "F9 v1.1";
```

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

Average Payload Mass by F9 v1.1
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%sql select min("Date") as "First Successful Landing" from SPACEXTBL \
where "Landing_Outcome" = "Success (ground pad)";
```

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

First Successful Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql select distinct("Booster_Version") from SPACEXTBL \
where "Landing_Outcome" = "Success (drone ship)" and \
"PAYLOAD_MASS__KG_" between 4001 and 5999
```

```
 * 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

- Calculate the total number of successful and failure mission outcomes

```
%sql select count(*) as "Count (Successful)" from SPACEXTBL where "Landing_Outcome" like "Success%";
```

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

Count (Successful)
61

```
%sql select count(*) as "Count (Failure)" from SPACEXTBL where "Landing_Outcome" like "Failure%";
```

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

Count (Failure)
10

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
%sql select distinct(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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql select substr(Date, 4, 1) as "Month", Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL \
where substr(Date, 1, 4) = "2015" and Landing_Outcome like "Failure (drone ship)";
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
5	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
5	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

```
%sql select Landing_Outcome, count(*) as "Count" from SPACEXTBL \
where Date between "2010-06-04" and "2017-03-20" \
group by Landing_Outcome \
order by "Count" desc;
```

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

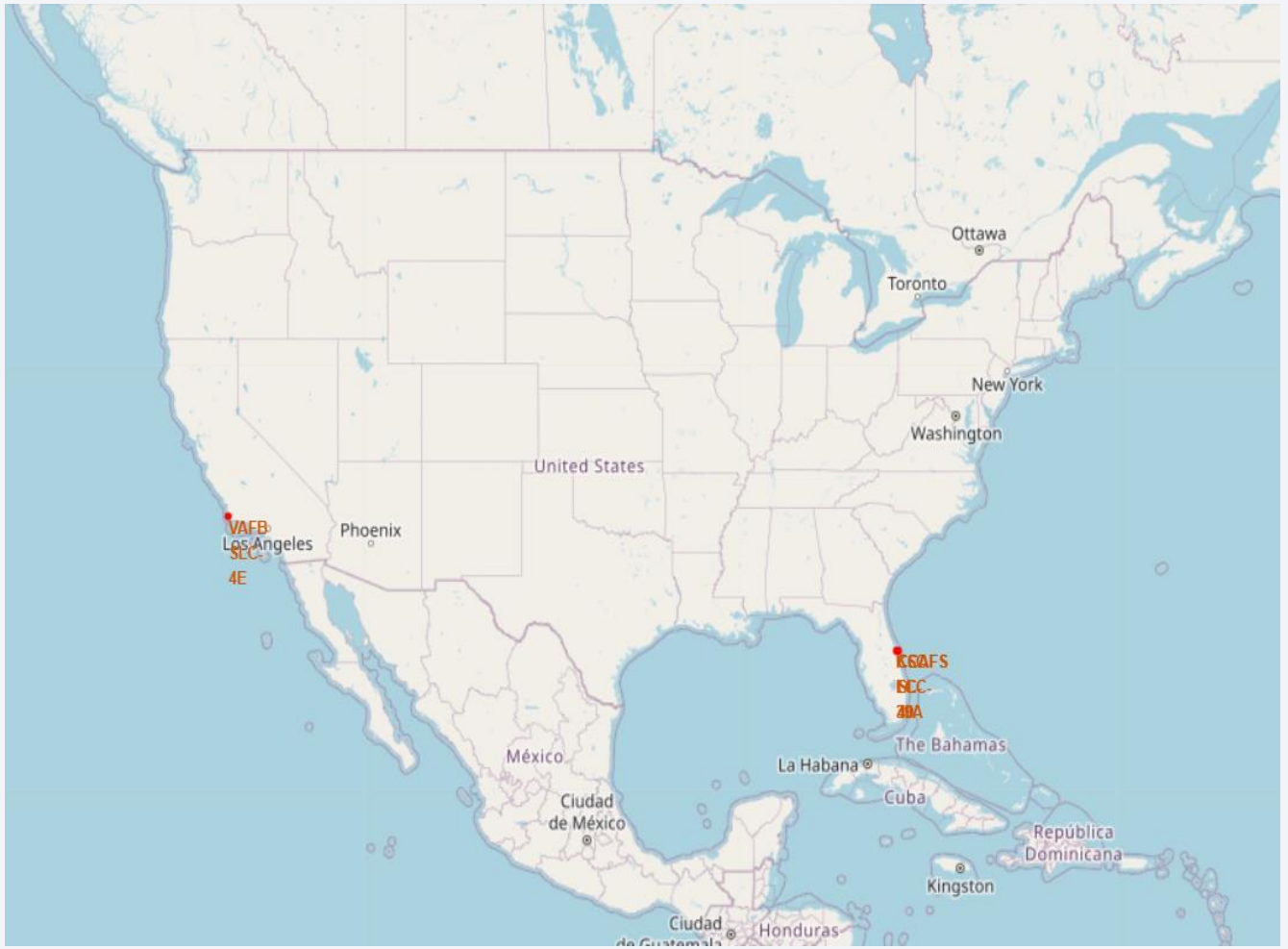
Landing_Outcome	Count
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

Launch Sites Location

- One launch site is in the California, whereas other sites in Florida.
- All the sites are near coasts

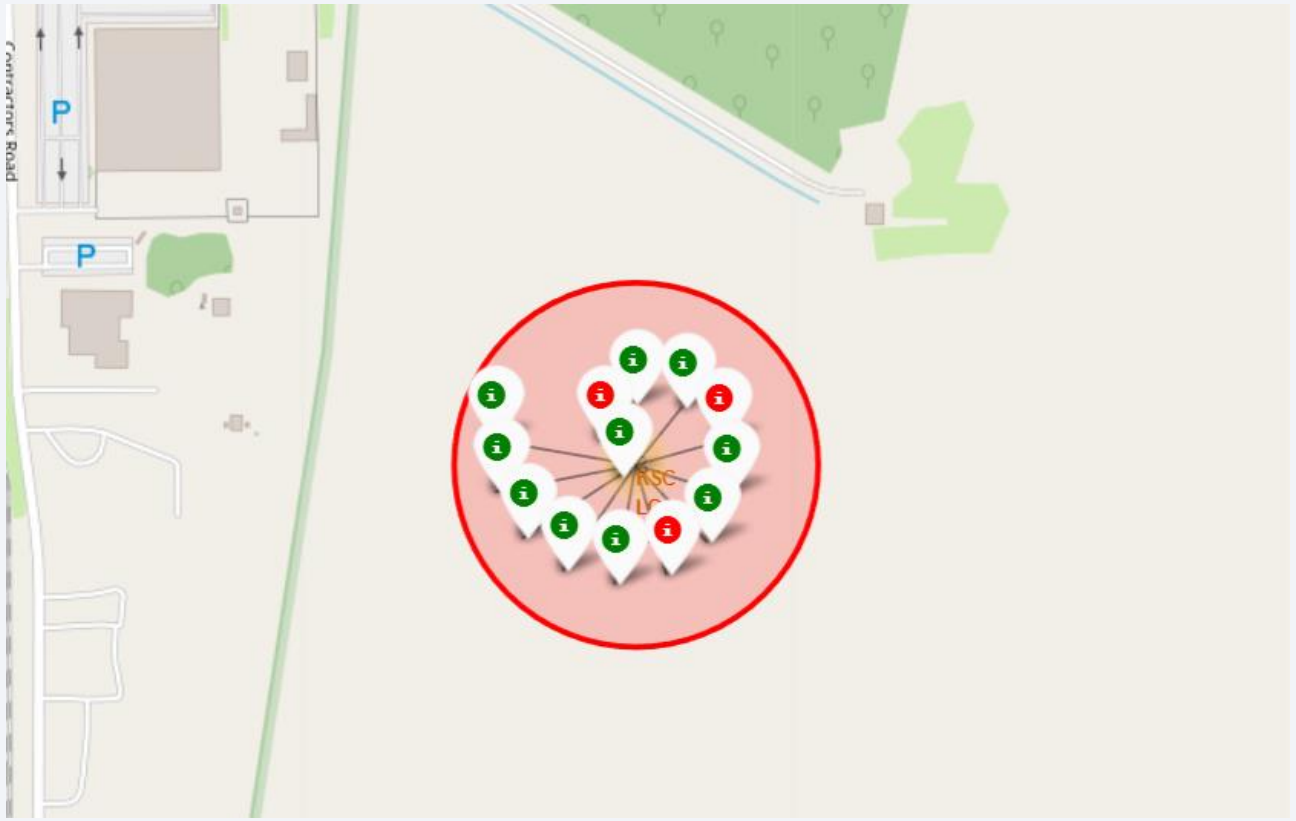


VAFB SLC - 4E Launch Outcomes

- VAFB SLC-4E outcomes show higher number of unsuccessful landings.

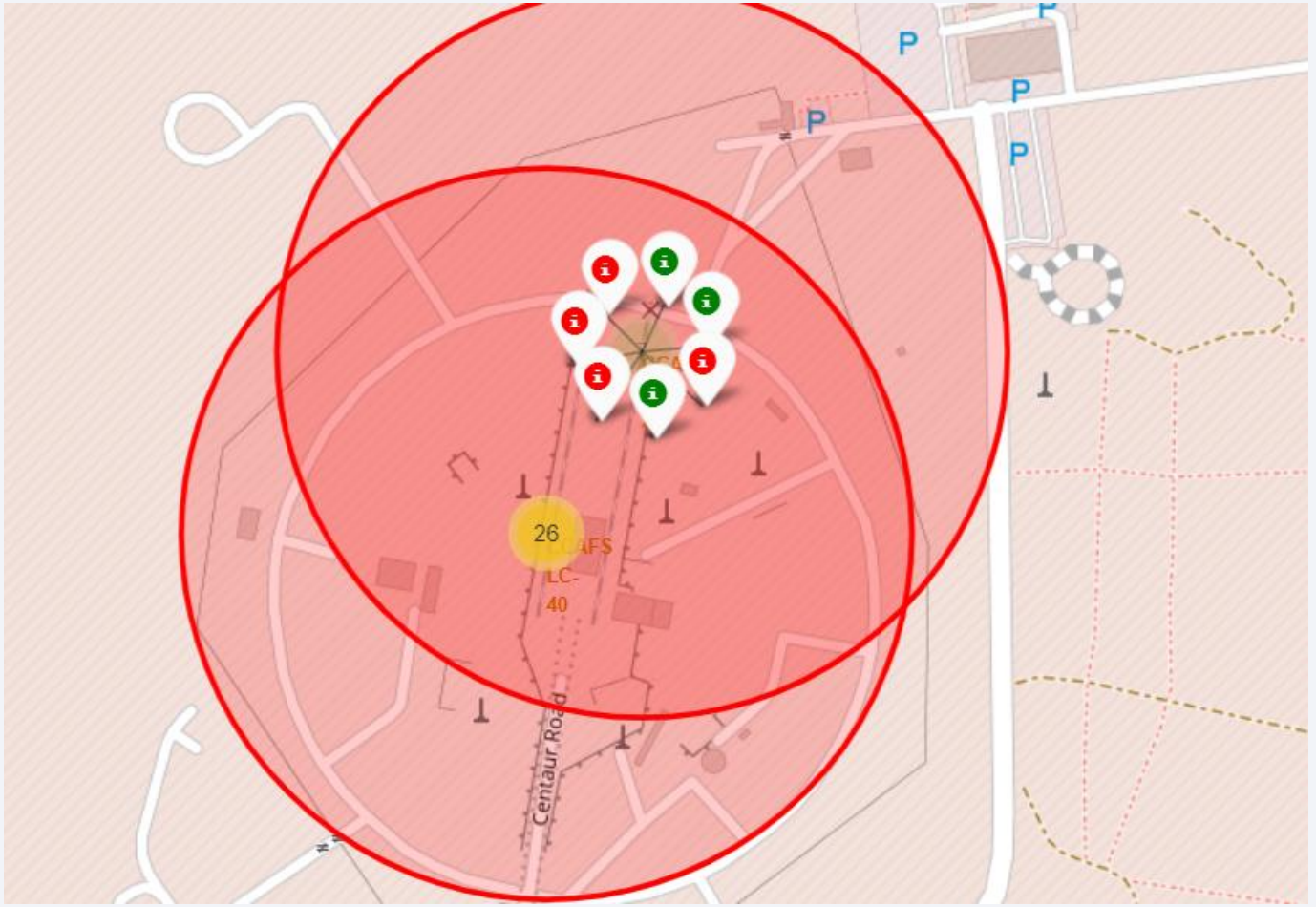
KSC LC - 39A Launch Outcomes

- KSC LC – 39A outcomes show higher number of successful landings. This can be helpful in deciding a launch site due to its high rate of successful landings



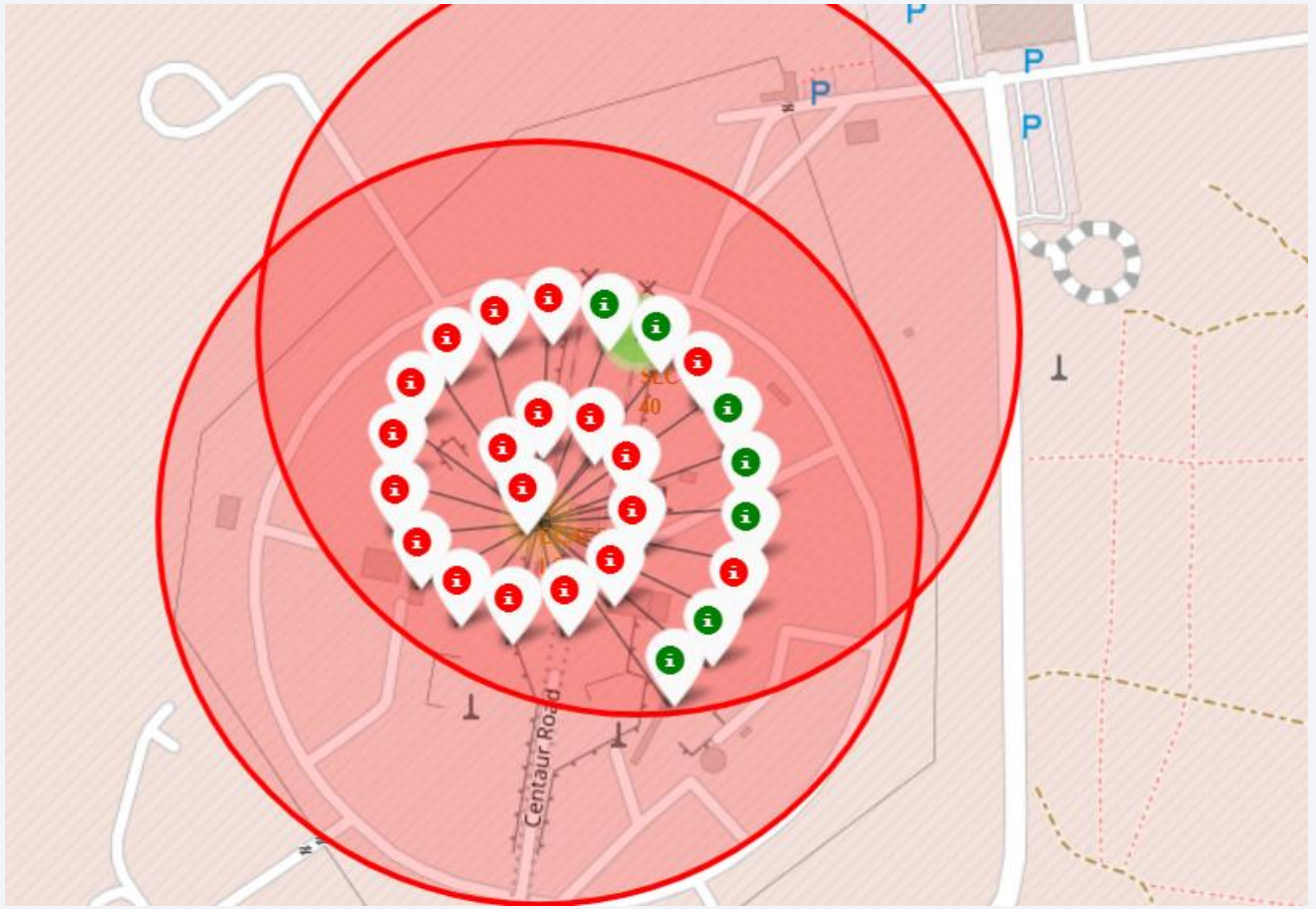
CCAFS SLC – 40 Launch Outcomes

- CCAFS SLC – 40 also shows a smaller number of successful landings, however since the total number of landings itself are smaller, no impactful inference can be made.

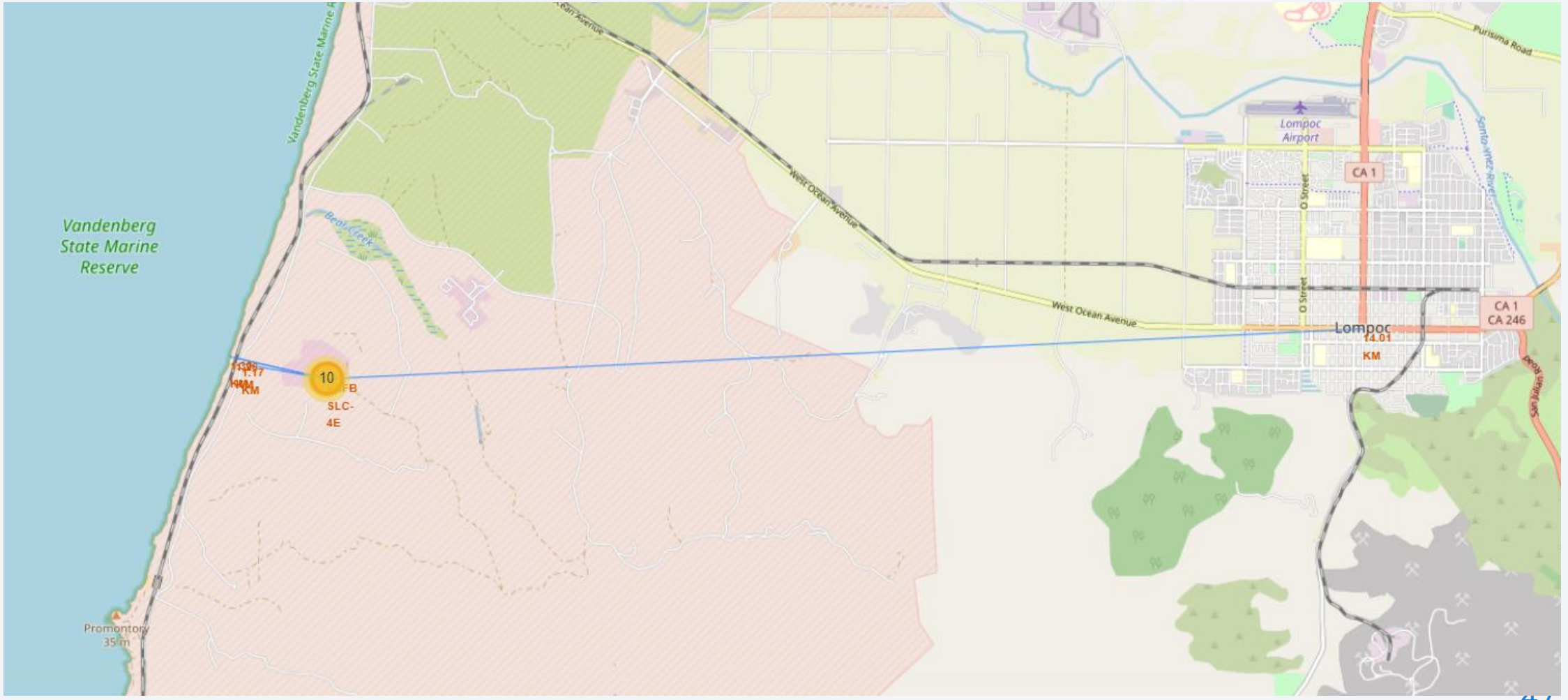


CCAFS LC – 40 Launch Outcomes

- CCAFS LC – 40 outcomes show higher number of unsuccessful landings as well. Since there are a high number of landings, we can say that this site is generally bad for successful recoveries.



VAFB SLC - 4E Proximities



VAFB SLC - 4E Proximities (Cont.)

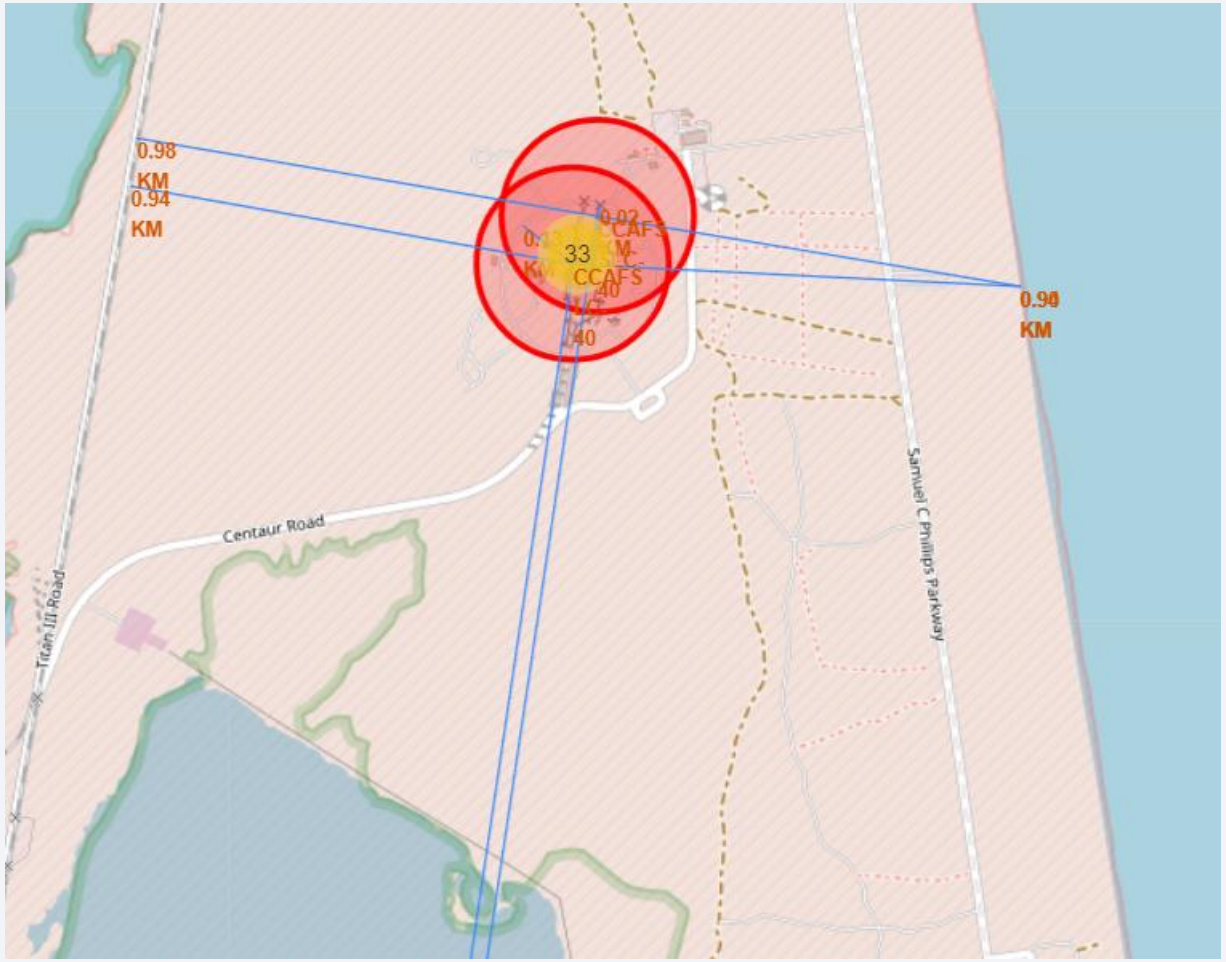
- Distances –
  - Coastline – 1.3 km
  - Railway – 1.26 km
  - Highway – 1.17 km
  - Nearest City - Lompoc (14.01 km)

KSC LC – 39A Proximities

KSC LC – 39A Proximities (Cont.)

- Distances –
  - Coastline – 3.92 km
  - Railway – 0.66 km
  - Highway – 0.69 km
  - Nearest City - Titusville (16.28 km)

CCAFS SLC – 40 and CCAFS LC - 40 Proximities

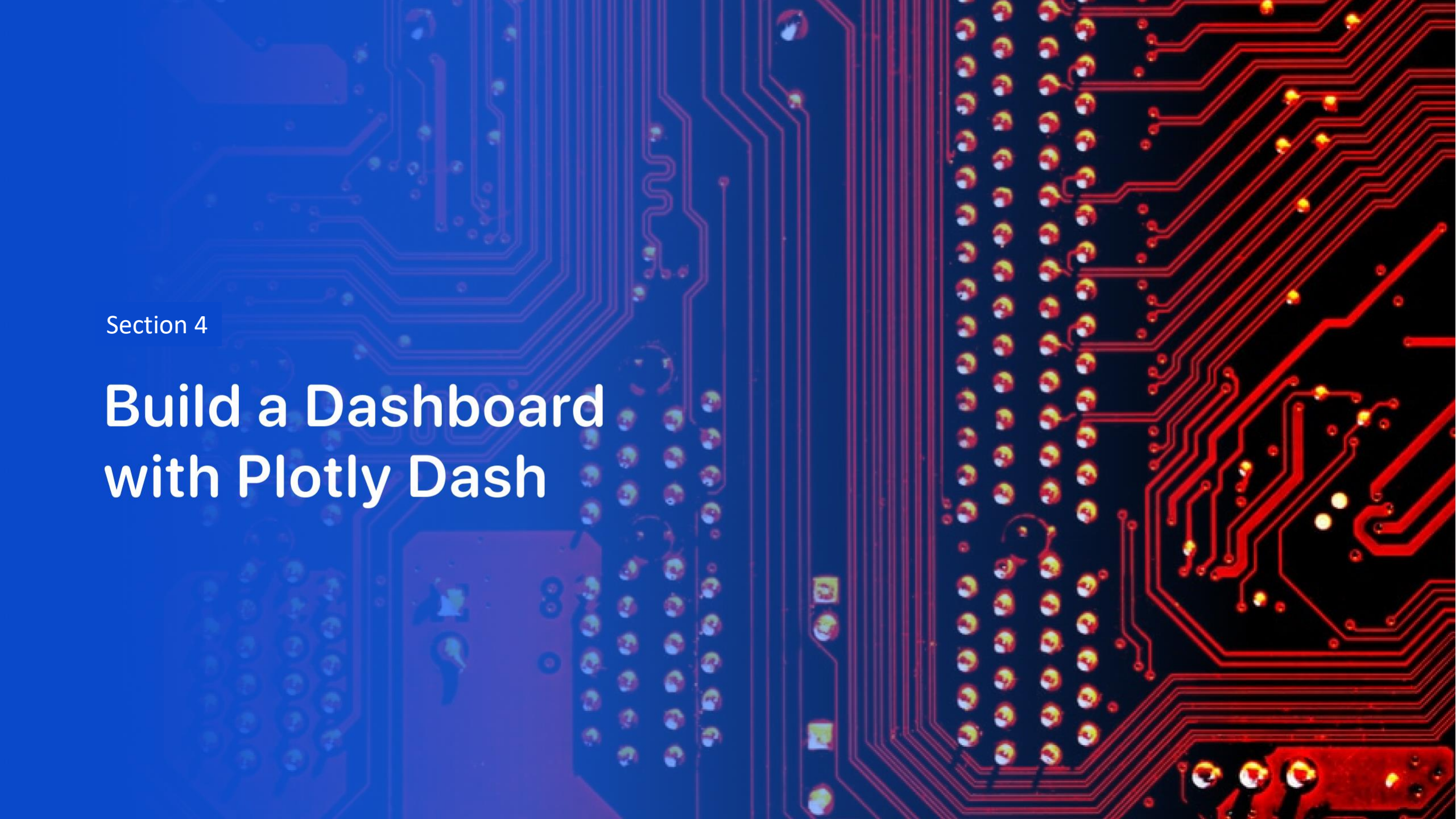


CCAFS SLC – 40 and CCAFS LC - 40 Proximities (Cont.)

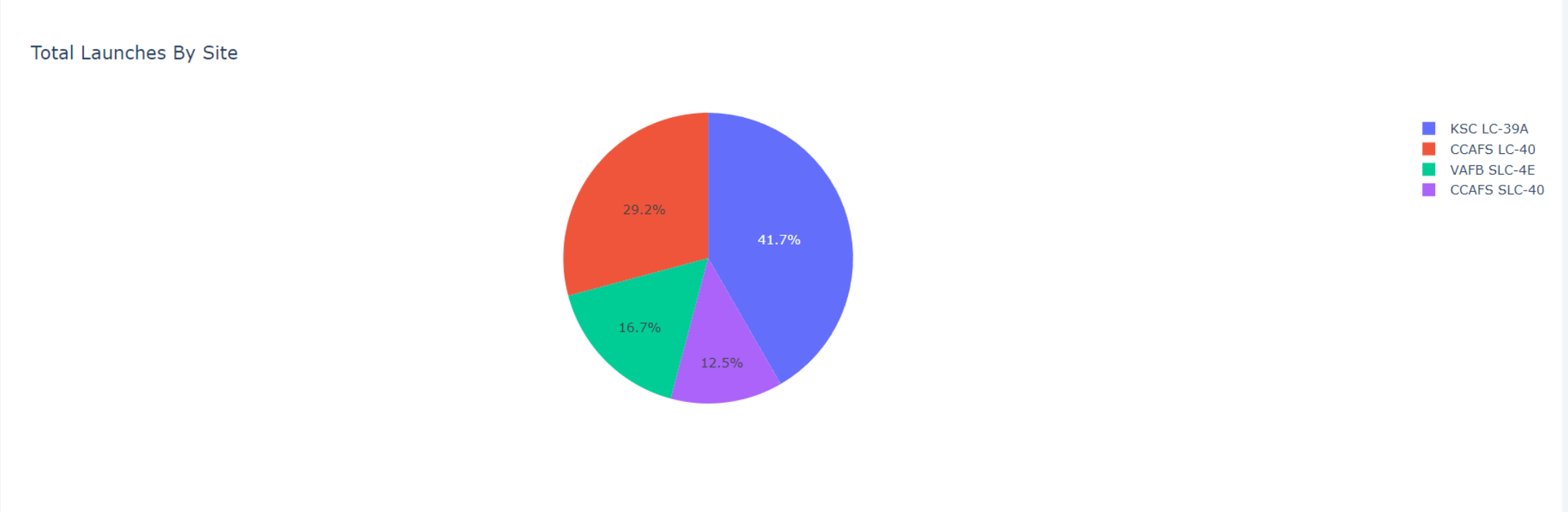
- Distances –
  - Coastline – 0.90 km
  - Railway – 0.98 km (CCAFS SLC - 40); 0.94 km (CCAFS LC - 40)
  - Highway – 0.02 km (CCAFS SLC - 40); 0.13 km (CCAFS LC - 40)
  - Nearest City - Cape Canaveral (18.07 km)

Inferences from proximities

- Distances from cities are far, with the highest being around 18 km.
- However, all launch sites and pads are at close proximities from railways, highways and roads and coastline. This can be the case due to -
  - Ease transport of materials between launch sites and space stations and vice versa in case of successful landing
  - Lessen costs on commutes from launch sites and work areas
  - Reduce risk to damage to materials due to transportation faults to and from launch sites

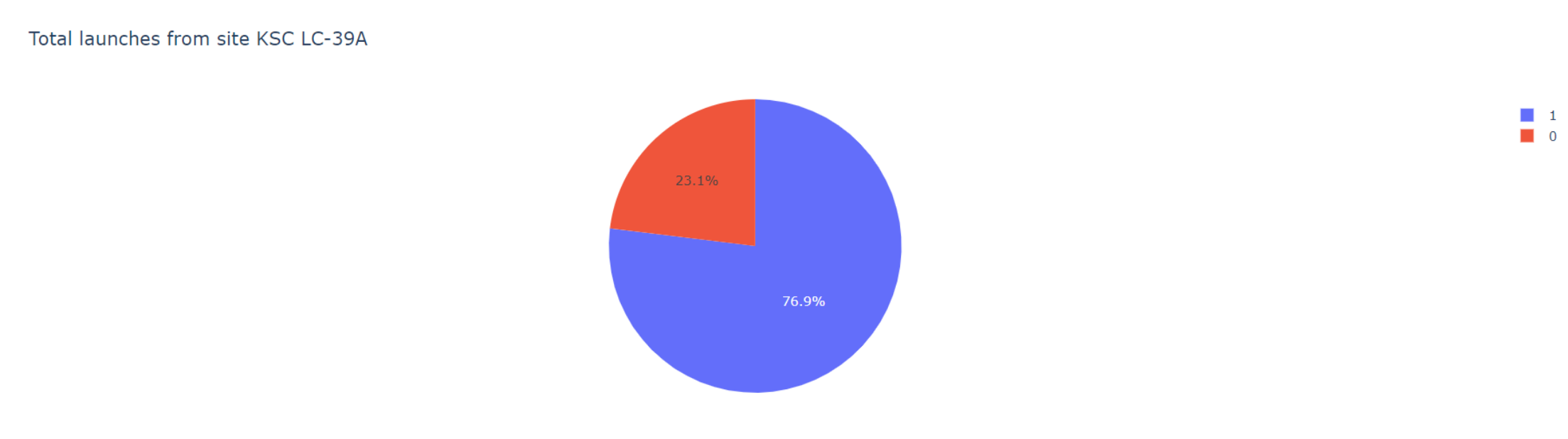

Section 4
Build a Dashboard with Plotly Dash

Launch success for all sites



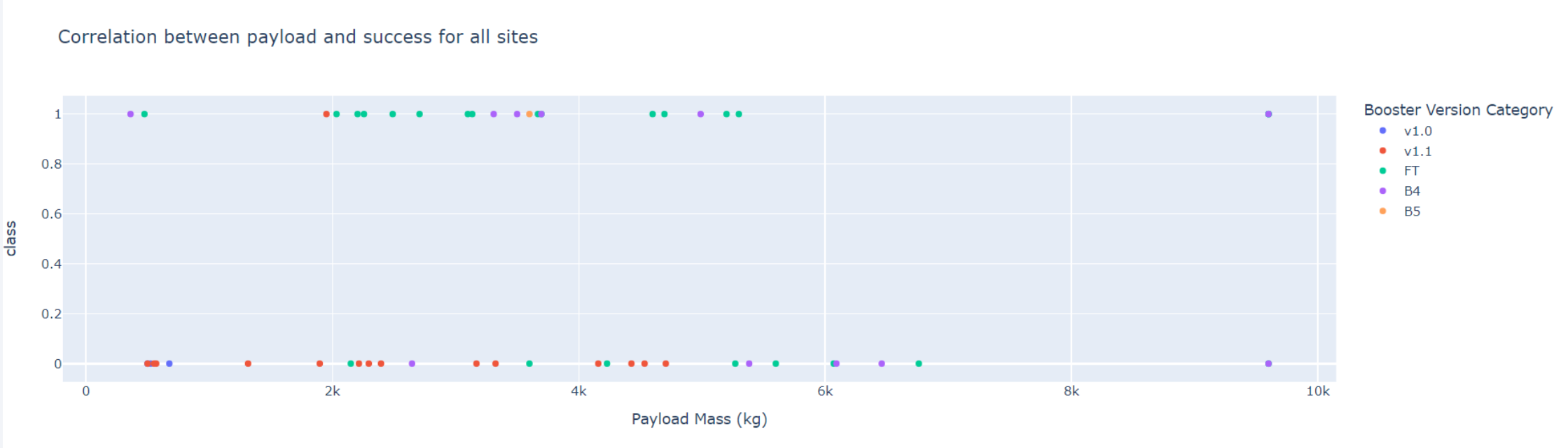
- Each color distinguishes sites from each other, while the percentages help understand the distribution of successful launches in between sites
- KSC LC – 39A has the highest number of successful launches (10 in number)
- CCAFS SLC – 40 has the least successes (3)

Site with the highest launch success ratio (KSC LC - 39A)



- Colors in the pie chart explains successful (denoted as 1) and unsuccessful (denote as 0) launches.
- KSC LC – 39A site has the highest launch success ratio (around 77% success rate)

Payload vs. Launch Outcomes



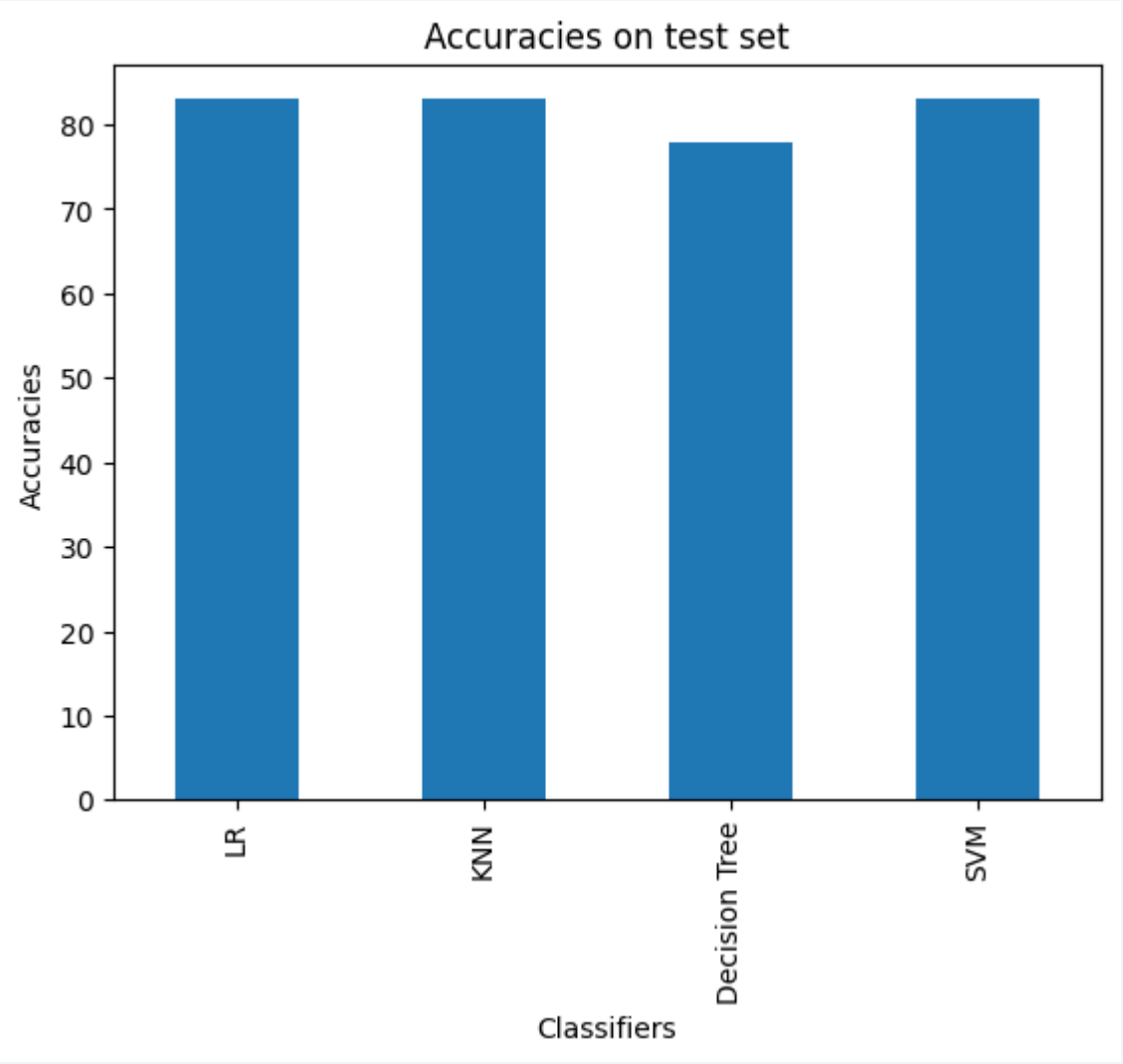
- Colors in the scatter plot help distinguish booster versions, with successful launches (denoted as 1) and unsuccessful ones (denoted as 0).
- From the chart, it is evident that payloads in range of 2000 to 4000 kgs are the most successful.
- Booster Version FT has the highest number of successful launches.

Section 5

Predictive Analysis (Classification)

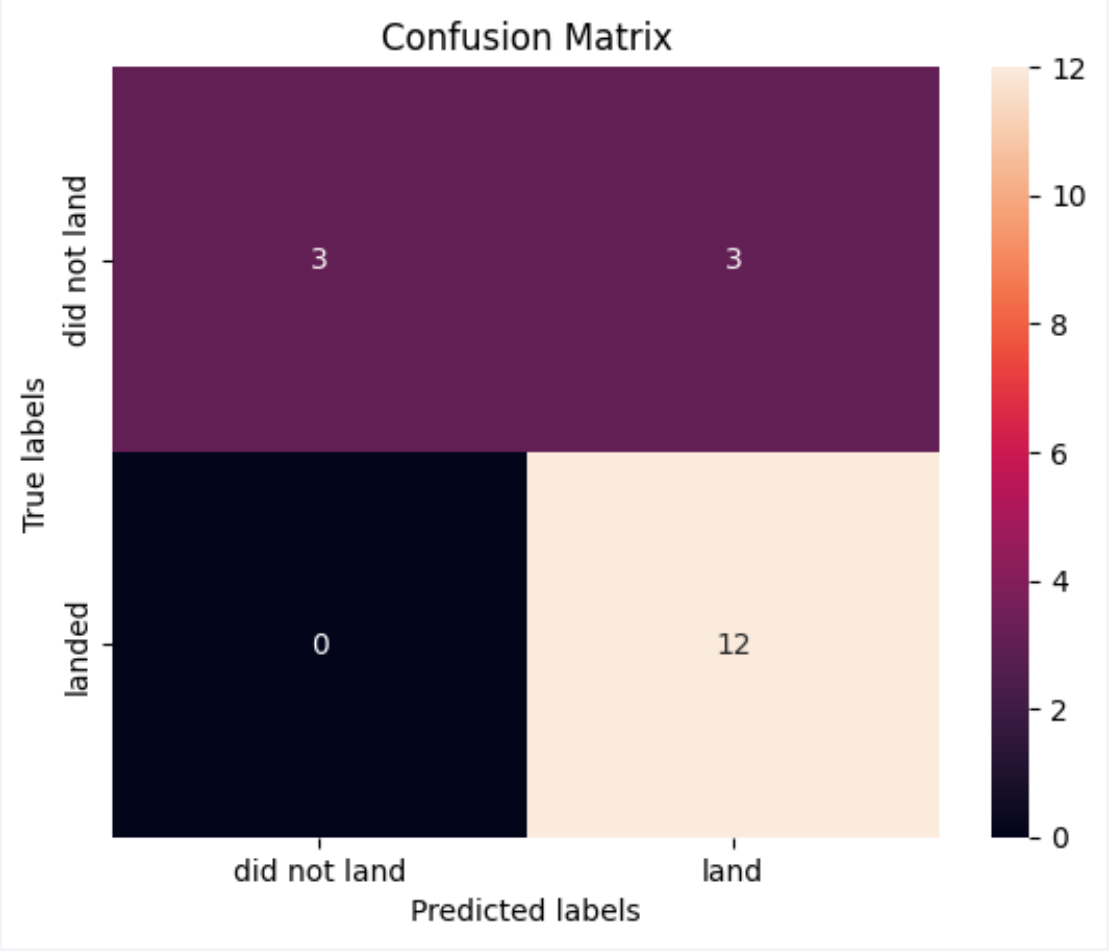
Classification Accuracy

- Chart displays accuracies of each classifier used on test set.
- Logistic Regression, KNN and SVM have high accuracies (83%)



Confusion Matrix

- The best confusion matrix is displayed, showing 3 false positives, causing the accuracy to go down to 83%



Conclusions

- Falcon 9's stage recovery successes attribute to –
  - Lighter payloads, particularly in between 2000 and 4000 kgs
  - Rockets launched from launch site KSC LC – 39A
  - Using Booster Version FT
  - Travelling SSO, LEO and VLEO orbits
- Predictive Analysis shows classifiers Logistic Regressor, Support Vector Machine and KNN classifier display 83.33% accuracy, whereas Decision Tree shows less accuracy (78%).

Thank you!