IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shanmukhi
23/03/2022

Outline

Executive Summary

- **Summary of methodologies**
- Data Collection, SPACEX API and SPACEX Wikipedia page using Beautiful Soup Obj.
- EDA, via basic data summaries, SQL queries and visualizations
- Launch sites visual analytics using Folium for interactive maps
- Interactive Dashboard using Plotly using SpaceX launch data in real-time
- Finding Best hyperparameters using Grid Search and predictions using various Classification techniques
- **Summary of all results**
- The accuracy score on test data showed ~83.33% for all classification algorithms.

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 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.
- **Problems you want to find answers**
- Predict whether first stage will land or not
- Factors that influence for successful landing

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SPACEX REST API and SPACEX Wiki Page by Web scrapping
- Perform data wrangling
  - Irrelevant columns removed, filtered dataframe with only Falcon9 launches
  - Replaced Payload missing values with mean and One hot encoding for Outcome as 0/1
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardized and split the data, performed tuning and fit the model, predicted and evaluated models.

Data Collection

- SPACEX REST API

Requested and collected the launch data from SPACEX API from url https://api.spacexdata.com/v4/launches/past. json() results are normalized and turned into pandas dataframe. API is again used to get more information about launches and stored in list which is later used to create new dataframe that is used for further analysis

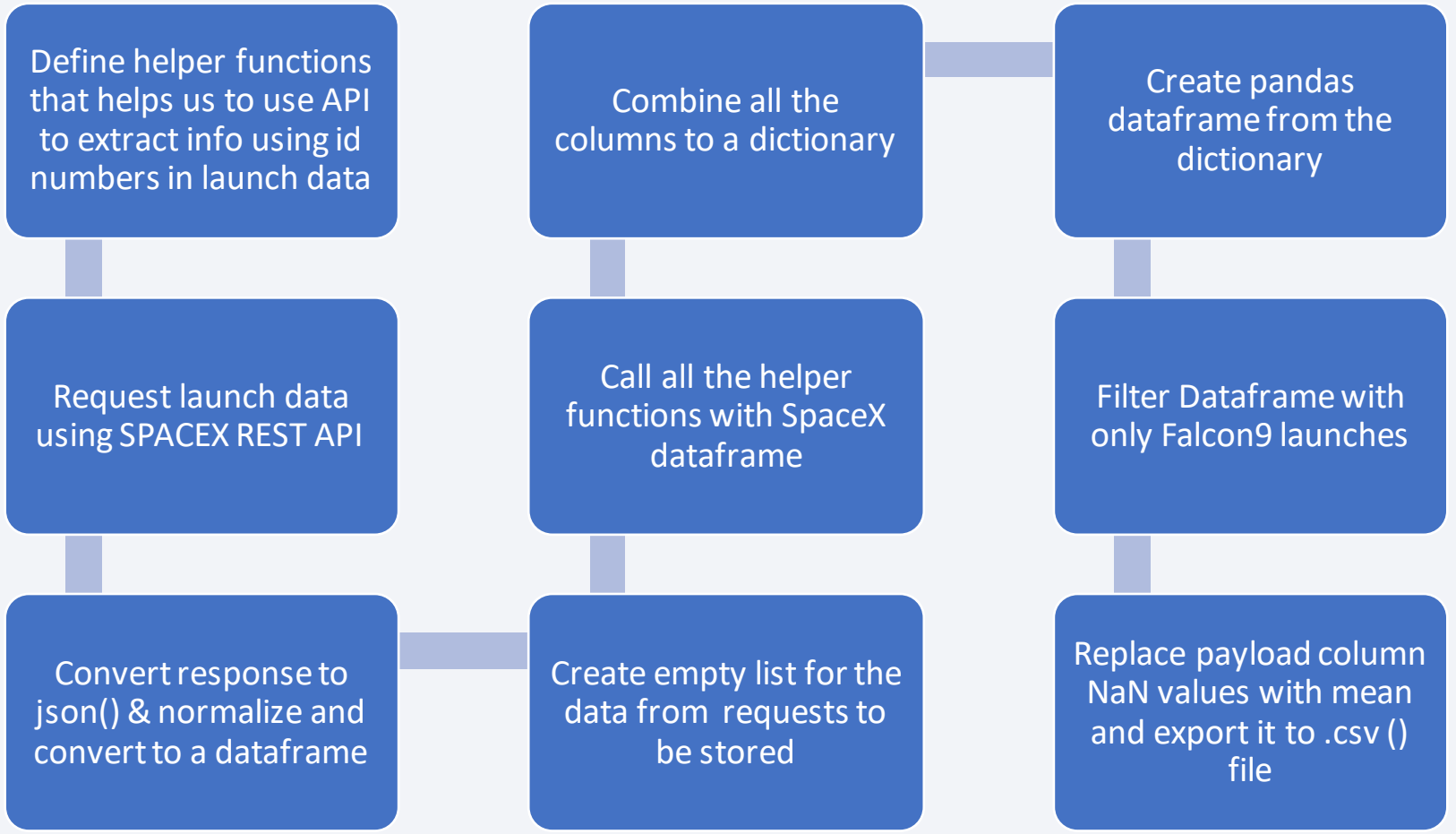
- WEBSCRAPPING

Data is collected from Wikipedia page titled List of Falcon9 and Falcon Heavy Launches

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches and web scrapped using beautiful soup object. Parsed the table and converted into pandas dataframe.

Data Collection – SpaceX API

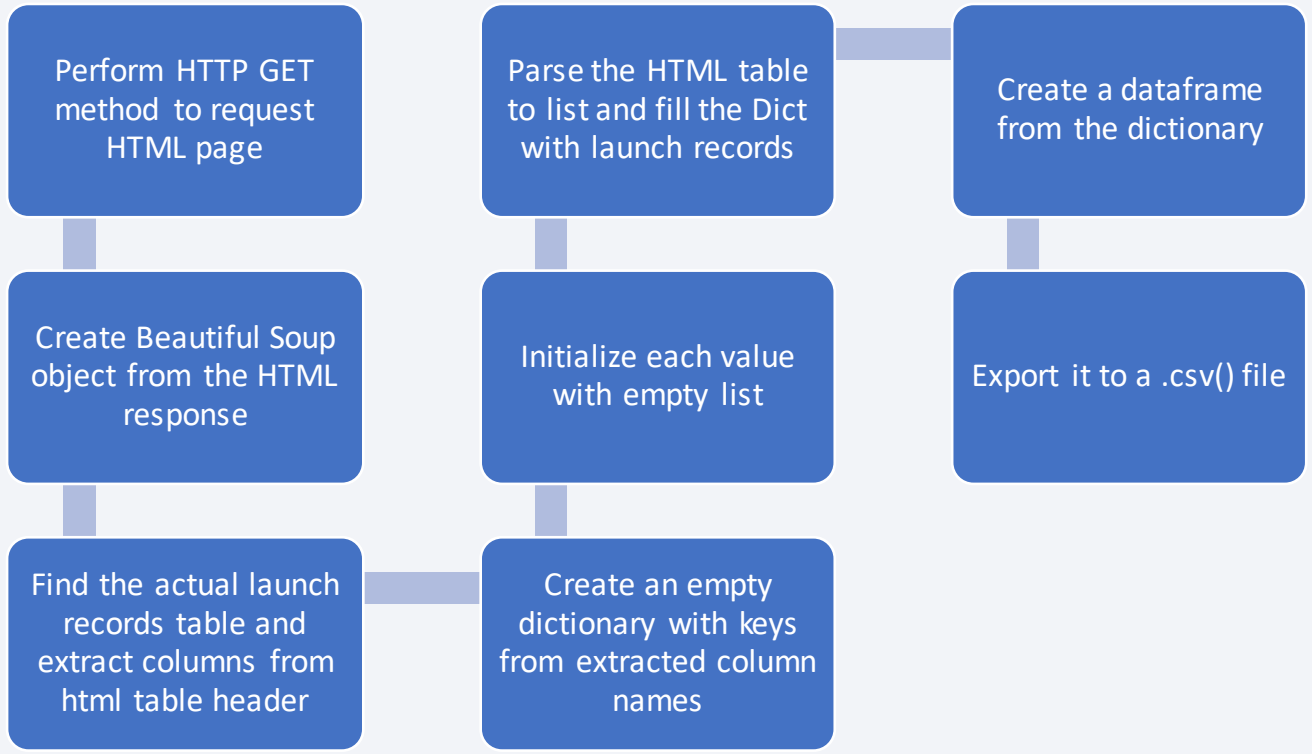
- Flow Chart



[Github url for notebook]

Data Collection - Scraping

- Flow Chart



[Github url for notebook]

Data Wrangling

- Have to convert the outcomes into Training labels with 1 for booster successful landing and 0 for unsuccessful.
- Created a set of outcomes where second stage didn’t land as bad_outcomes i.e., {'False ASDS', 'False Ocean', 'False RTLS', 'None ASDS', 'None None’}
- Created a landing outcome label from outcome column and then mapped to the dataframe

One-hot encoding:

- If outcome belongs to set of bad outcome → Class is labeled as 0
- If outcome doesn’t belong to bad outcome → Class is labeled as 1

Github url to notebook

EDA with Data Visualization

Main aspect of visualization is to find out the relationship between the features Payload Mass, Flight Number, Orbit, Launch Site, Class, Year

- Plots:

1. Scatter Plot: Flight Number vs Payload Mass, Flight Number vs Launch Site, Payload Mass vs Launch Site, Flight Number vs Orbit, Payload Mass vs Orbit
2. Bar Plot: Success rate vs Orbit
3. Line Plot: Launch success year trend

Github Url

EDA with SQL

- In order to Understand the dataset we have loaded it into IBM DB2 database and executed SQL queries
- Queries:

1. Display the names of the unique launch sites in the space mission.
2. Display 5 records where launch sites begin with the string 'CCA'.
3. Display the total payload mass carried by boosters launched by NASA (CRS).
4. Display average payload mass carried by booster version F9 v1.1.
5. List the date when the first successful landing outcome in ground pad was achieved.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
7. List the total number of successful and failure mission outcomes.
8. List the names of the booster_versions which have carried the maximum payload mass. Use a subquery.
9. List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.
10. 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

Github Url

Build an Interactive Map with Folium

❖Marking launch Sites on Map

- Circle() object is used to create circle at specified location and Popup() to add a label popup
- Marker() is used to add text label to the circle object

❖Mark the success/failure launch at each site

- MarkerCluster() is used to mark numerous failed and successful outcomes at specified site with distinguished colors

❖ Calculated Distance from launch site to its proximities and Polyline() is used to draw the line between them

Github Url

Build a Dashboard with Plotly Dash

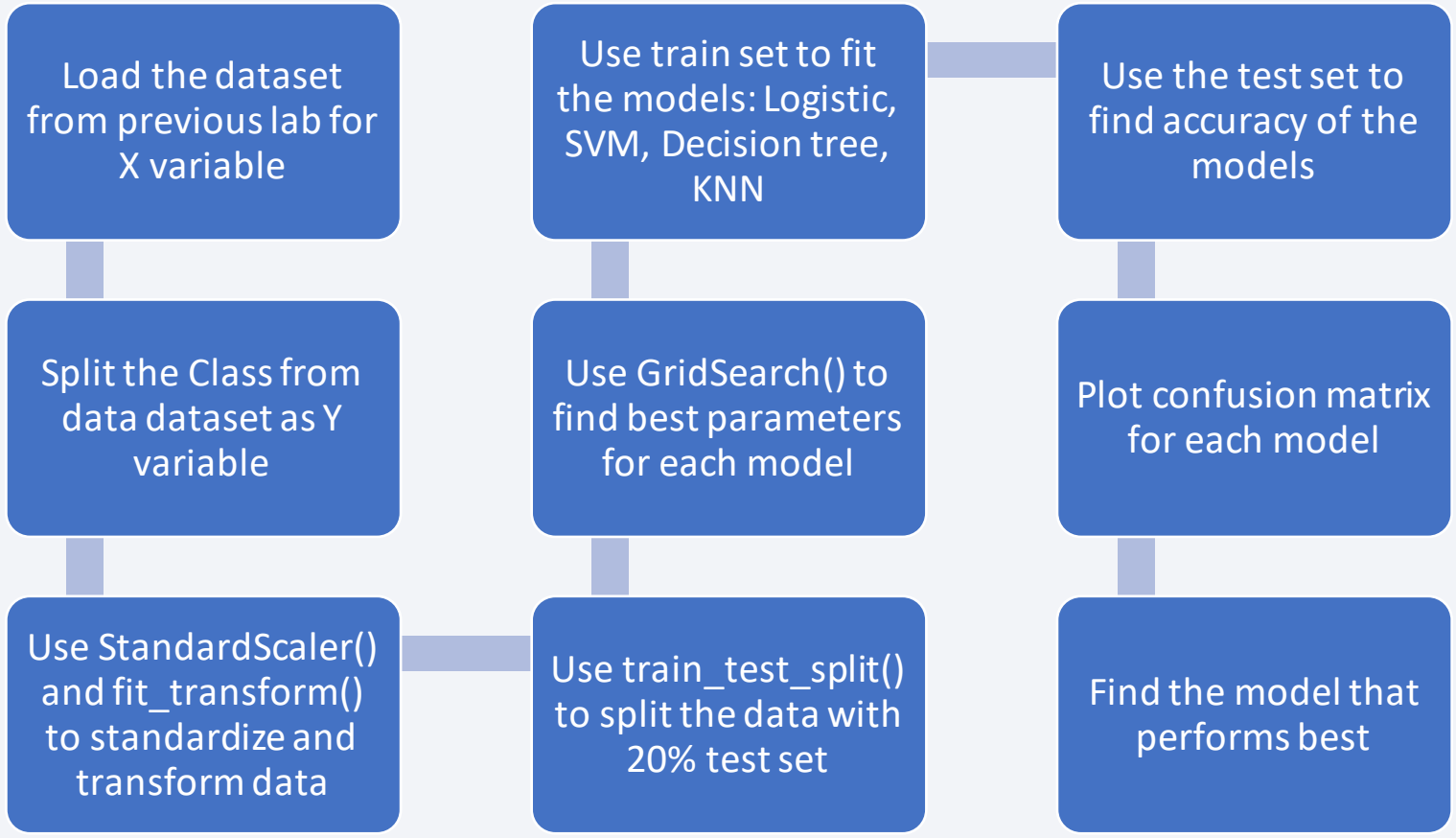
- An Interactive Dashboard has been created with Plotly Dash library. It consists of

1. Pie Chart: Showing total success launches for all sites and success/failure ratio for respective sites
2. Scatter Plot: Based on the site selected and range of Payload mass selected it shows the scatter plot between mass and launch outcome varied by booster version

Github Url

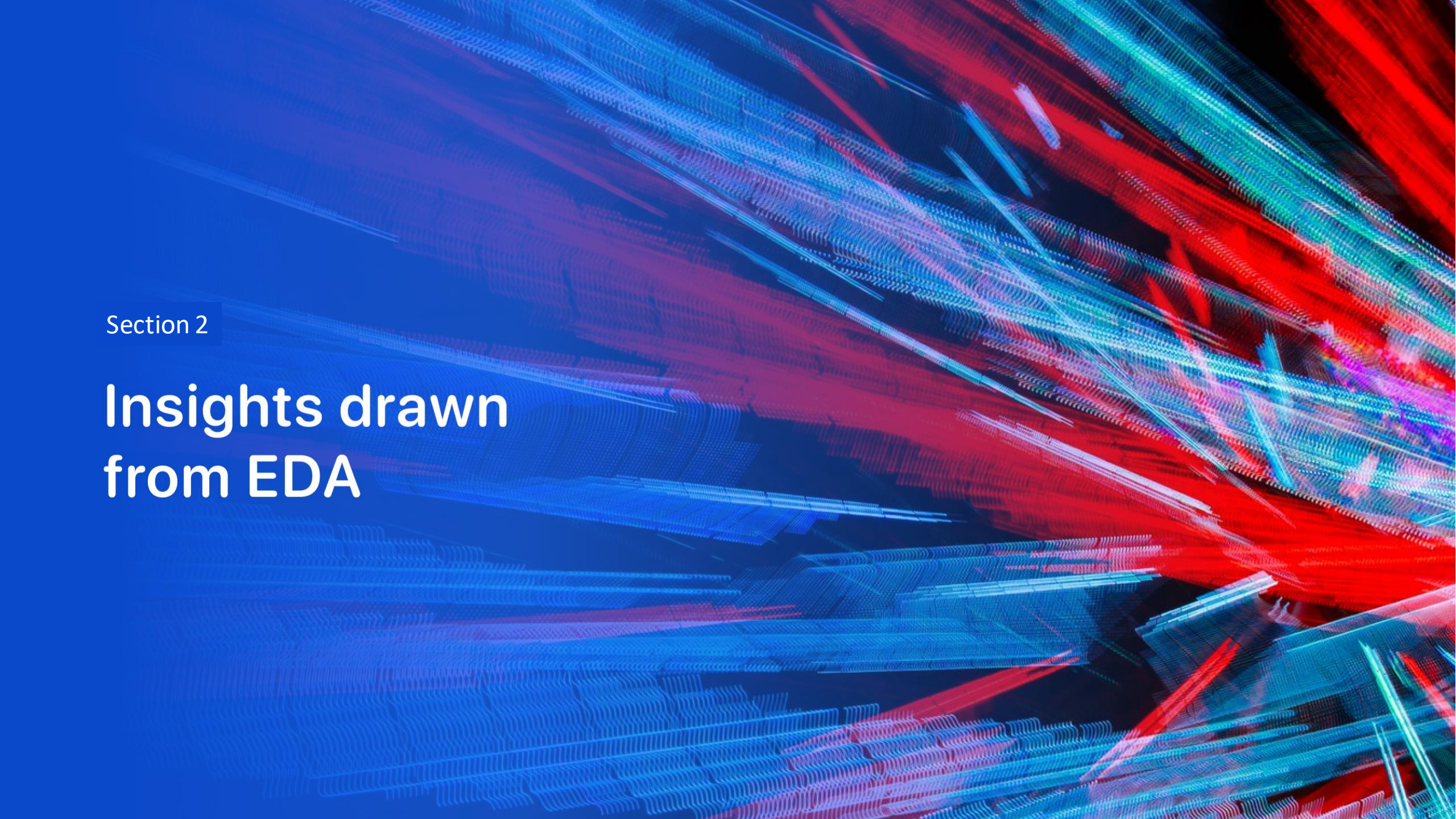
Predictive Analysis (Classification)

- Flow Chart

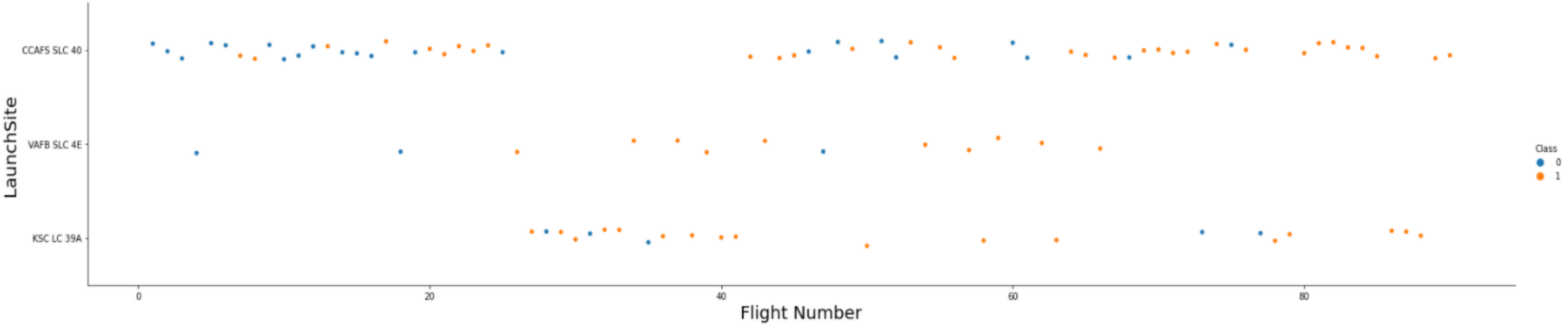


Results

- Results of the following are shown in the upcoming slides
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

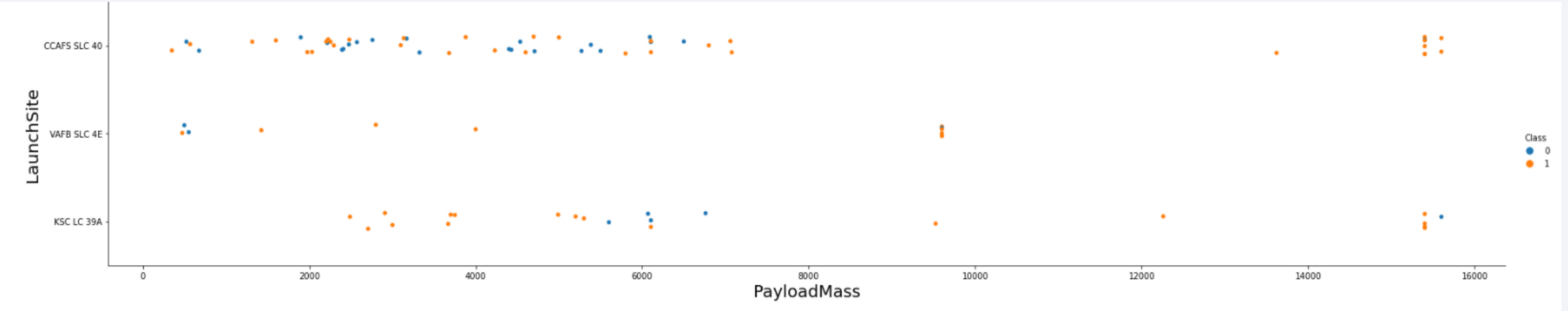

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



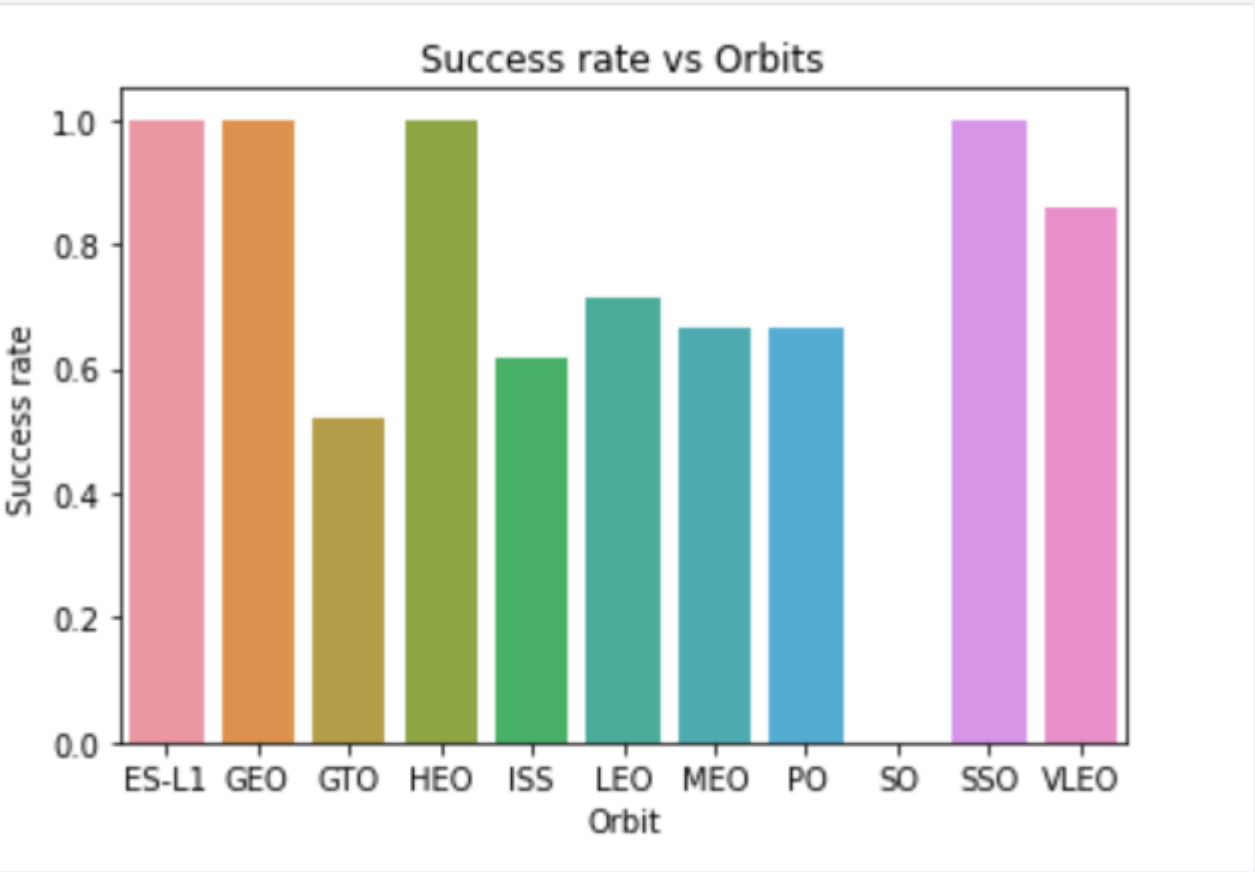
- Orange indicates the successful launches, Blue indicates unsuccessful launches
- Unsuccessful launches were more frequent in early flight numbers and success improved with recent flights

Payload vs. Launch Site



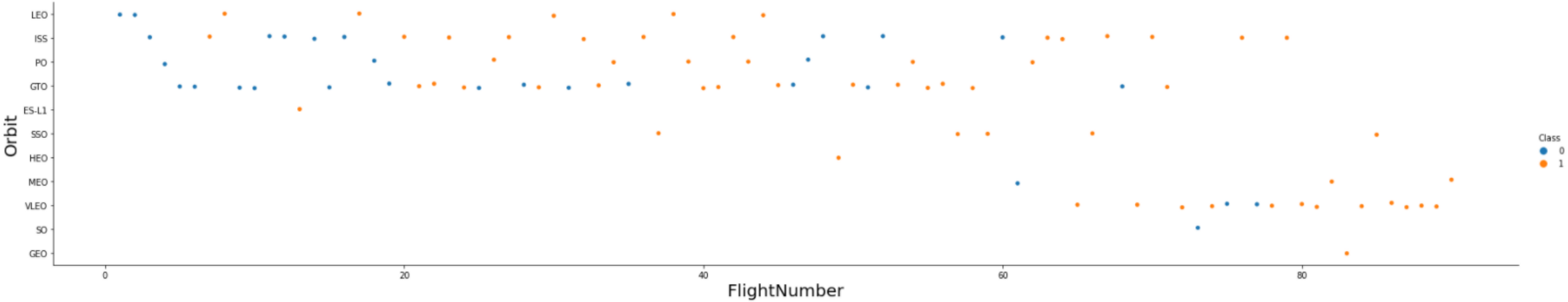
- Orange indicates the successful launches, Blue indicates unsuccessful launches
- VAFB SLC 4E has no launches with payload mass greater than 10000
- CCAFS SLC 40 has more number and VAFB SLC 4E has least number of launches

Success Rate vs. Orbit Type



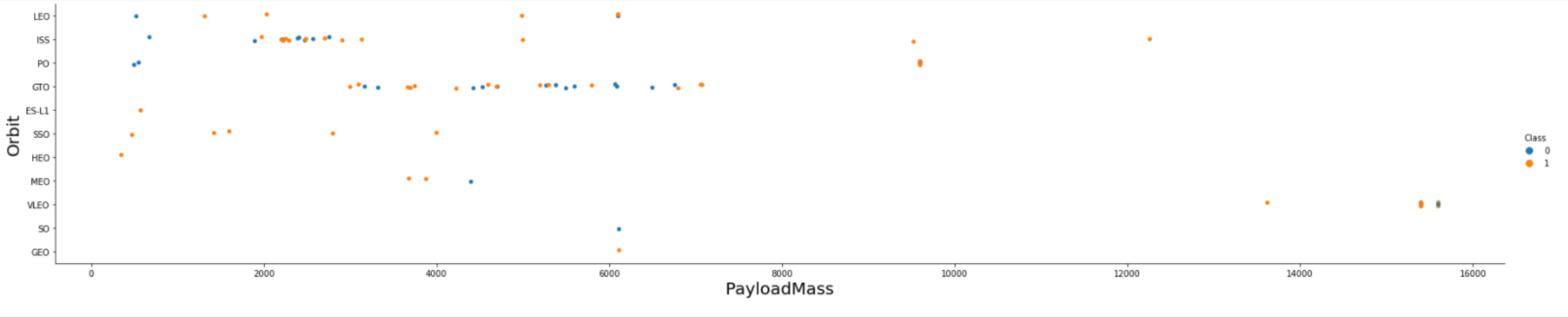
- Launch into ES-L1, GEO, HEO, SSO Orbits have 100% success rate where as SO orbit has 0% success rate
- Launch into GTO orbit has less than 60 % of success rate
- ISS, LEO, MEO, PO & VLEO have more than 60% success rate

Flight Number vs. Orbit Type



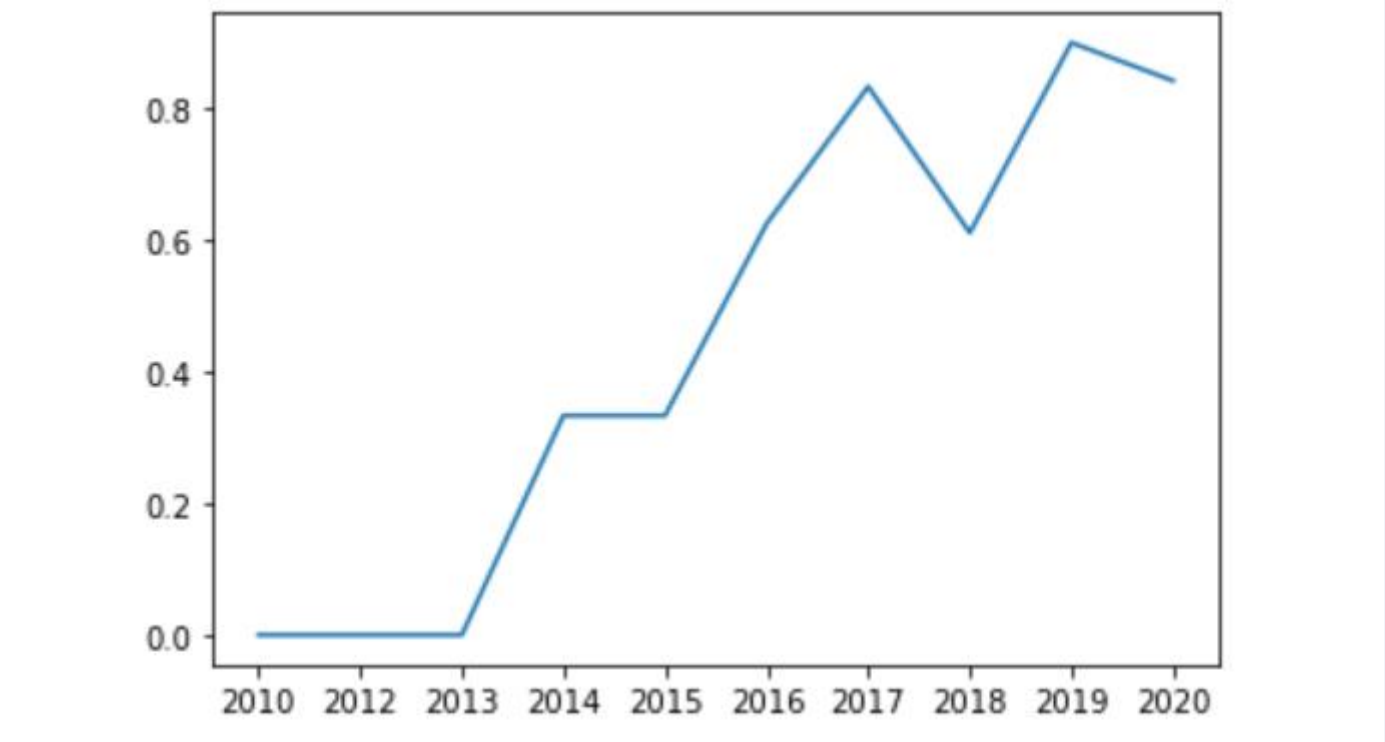
- Previously LEO, ISS, PO, GTO Orbits had major of the launches
- Recent time most of the launches were into VLEO orbit with maximum success rates.

Payload vs. Orbit Type



- Heavy Payload mass launches with success rate were into VLEO Orbit.
- GTO orbit has 50-50% of success and failure launches with all of the payload mass less than 8000.
- ISS Orbit has only one heavy Payload mass of 12000 launch which has been successful.

Launch Success Yearly Trend



- Since 2016, the launch success rate has been more than 50%
- The success rate has been increasing since 2010 except a decrement to 60 % in the year 2018
- In the recent years the success rate has reached to 80%

All Launch Site Names

```
%sql SELECT DISTINCT(LAUNCH_SITE)FROM SPACEX;
```

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

- Returns the names of all the launch sites
- We can see the first 2 observations are the same but a small mistake while entering them.

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEX 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

- Returns the first five records where launch site name begins with ‘CCA’

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS "TOTAL_PAYLOAD_MASS" FROM SPACEX WHERE CUSTOMER = 'NASA (CRS)';
```

total_payload_mass
45596

- Returns the sum of payload mass in Kg of all records where Customer is NASA (CRS)

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS "AVERAGE_PAYLOAD_MASS" FROM SPACEX WHERE BOOSTER_VERSION = 'F9 v1.1';
```

average_payload_mass
2928

- Returns the average of payload mass in Kg of all records whose Booster version is F9 v1.1

First Successful Ground Landing Date

```
%sql SELECT MIN(DATE)AS DATE_OF_FIRST_SUCCESFUL_LANDING FROM SPACEX WHERE LANDING__OUTCOME = 'Success (ground pad)' ;
```

date_of_first_succesful_landing
2015-12-22

- Returns the first successful launch date where landing outcome was on Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT BOOSTER_VERSION FROM SPACEX 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

- Returns the names of the booster versions with successful landing on drone ship and having payload mass between 4000 and 6000 kg

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT MISSION_OUTCOME,COUNT(*) as COUNT FROM SPACEX GROUP BY MISSION_OUTCOME;
```

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

- Returns the count of both Successful and Failure landing mission outcomes.

Boosters Carried Maximum Payload

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

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

- Returns the names of the boosters which carried maximum payload.

2015 Launch Records

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

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- Returns booster version and their respective launch site whose landing was failed on drone ship in the year 2015.

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

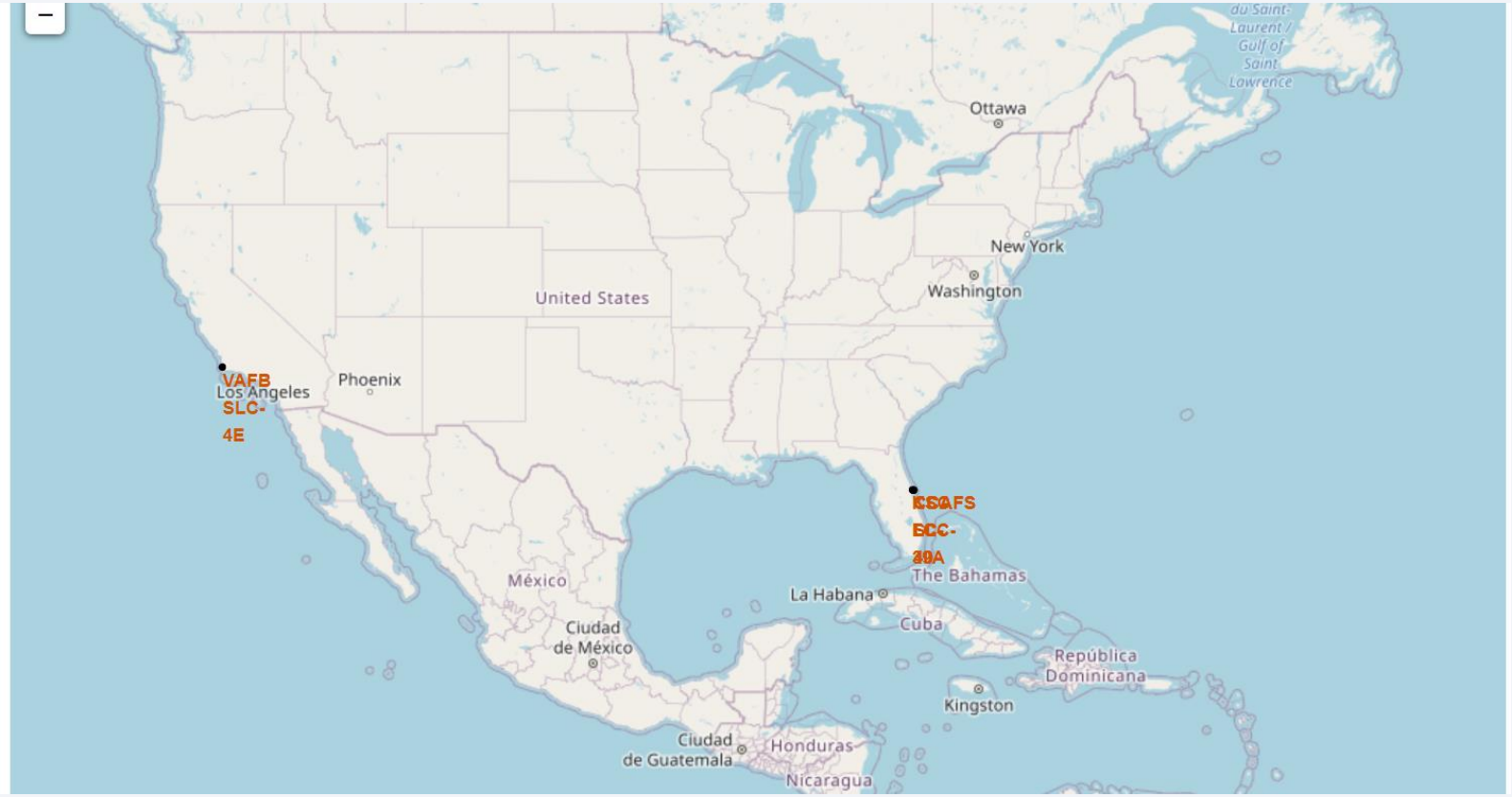
```
%sql SELECT LANDING__OUTCOME,COUNT(*) AS COUNT FROM SPACEX WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY COUNT
```

landing__outcome	COUNT
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

Returns the count of all of the landing outcomes between the dates 2010-06-04 and 2017-03-20 in descending order

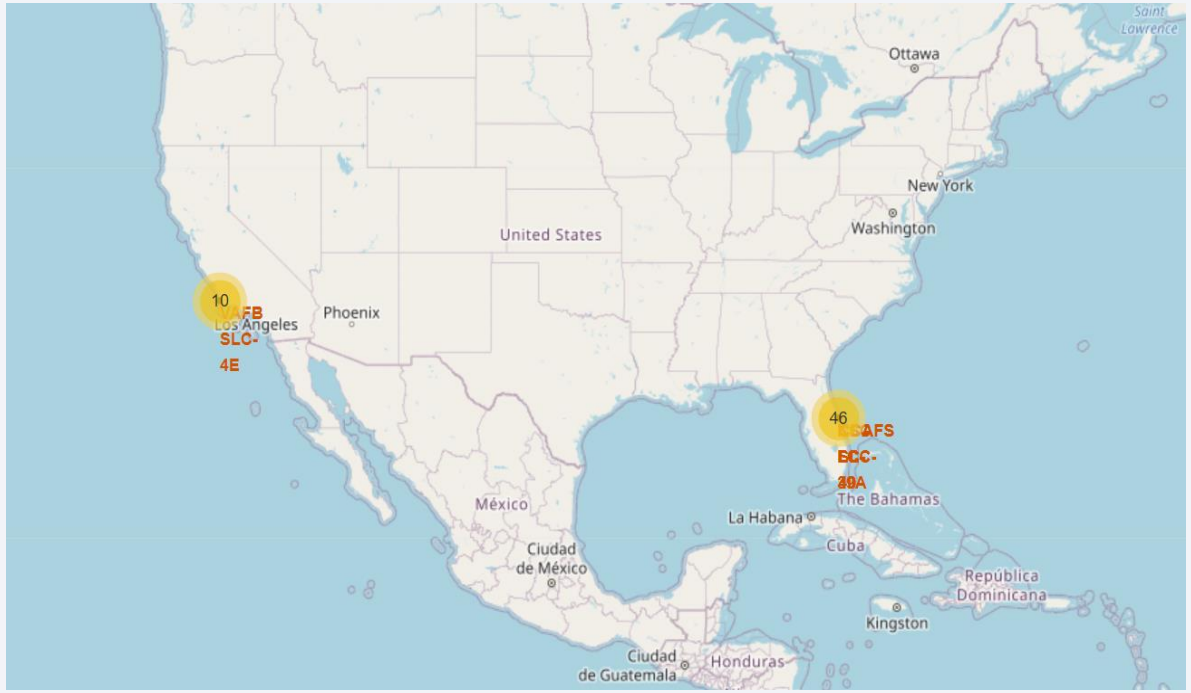

Section 3
Launch Sites
Proximities Analysis

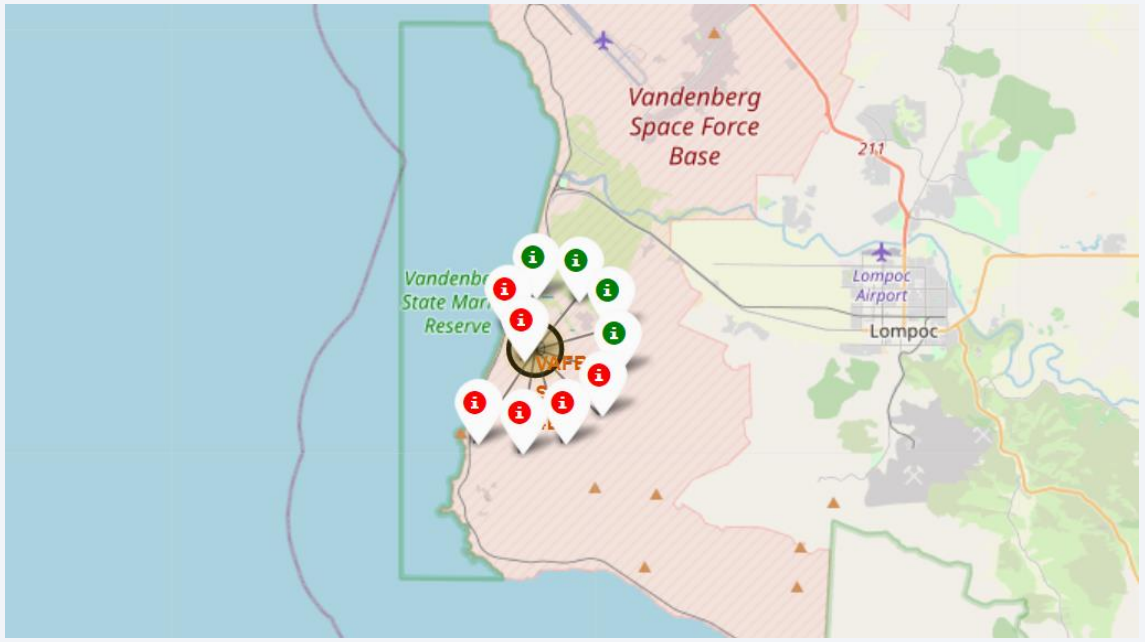
All Launch Sites on Map



- We can observe all launch sites are near to coastal regions.
- KSC LC-39A and CCAFS SLC- 40 are near to each other.

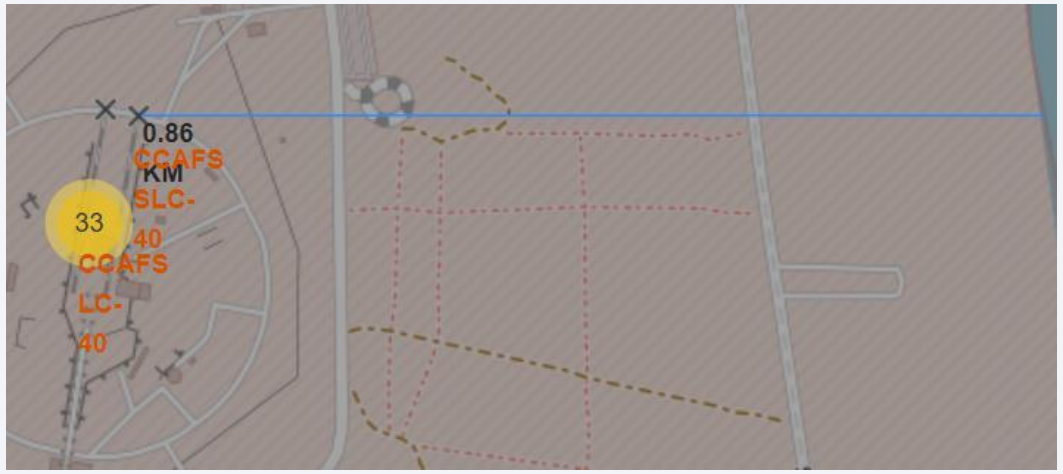
Launch Outcomes of each Site



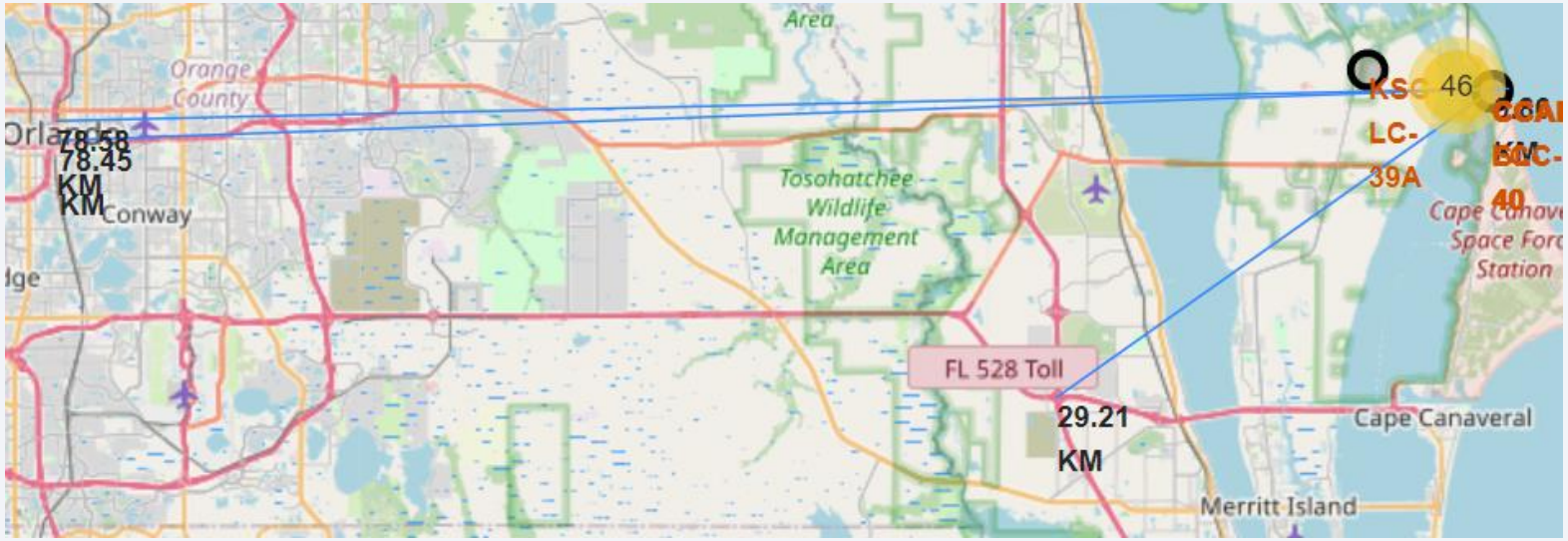


We can observe for each launch site the number of outcomes are displayed and upon zooming we can see the number of success/failure outcomes with color coded where red indicates failure and green indicates success

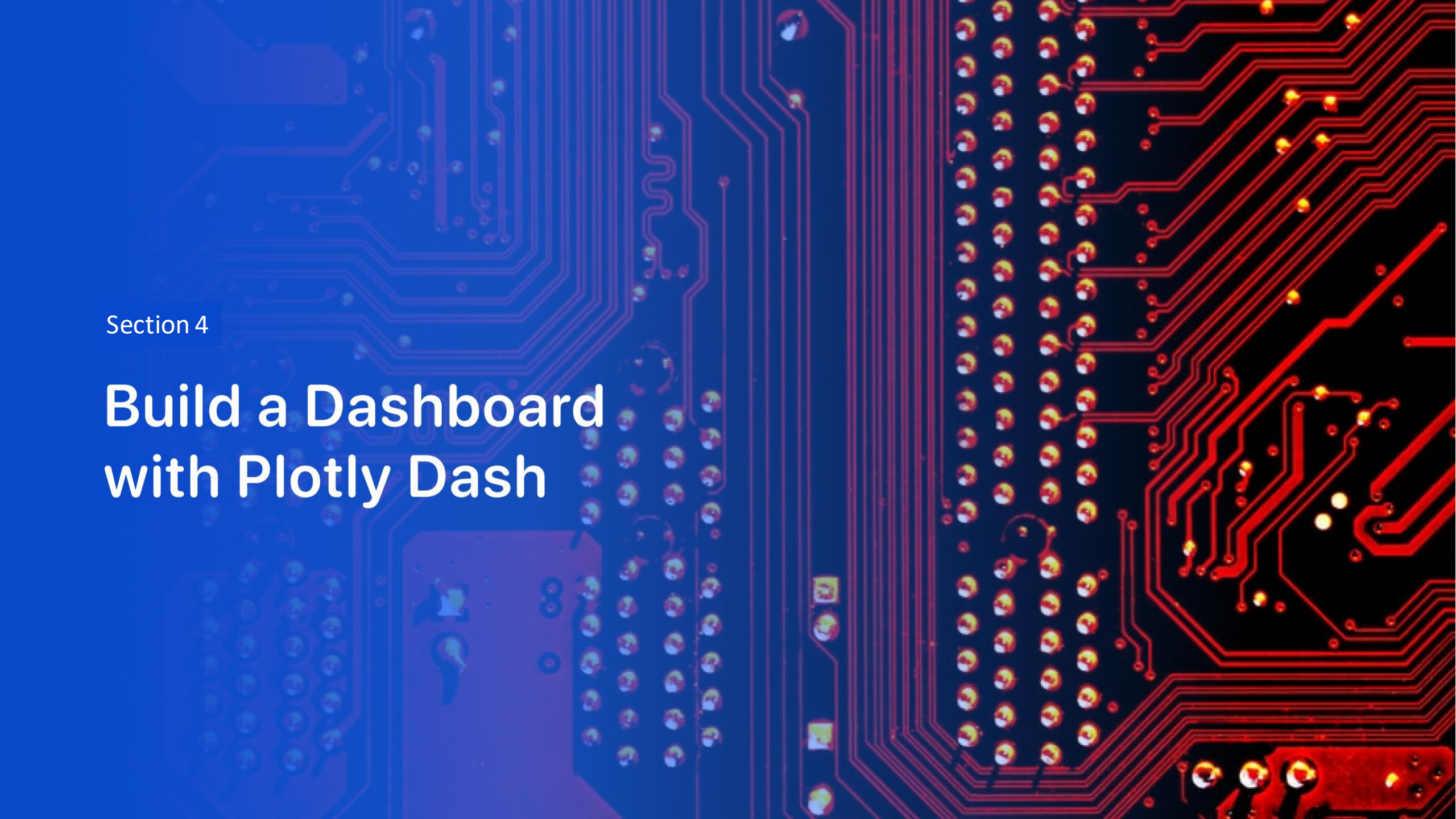
Launch Site distance to its proximities



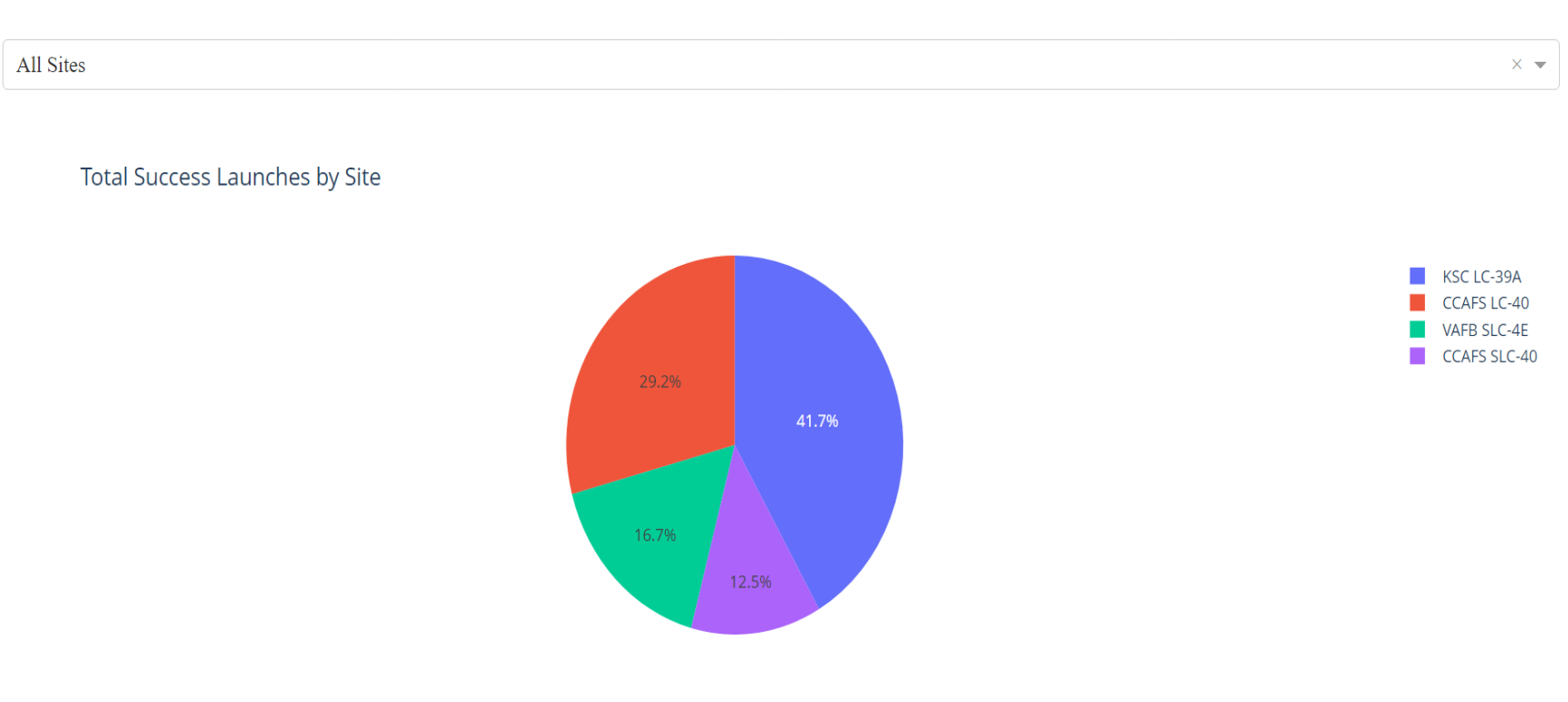
The distance from the launch site CCAFS SLC-40 and the coastal line is 0.86 km



1. The distance to nearest city Florida from site is 78.45 km
2. The distance to nearby railway line is 78.58 km
3. The distance to highway is 29.21 km

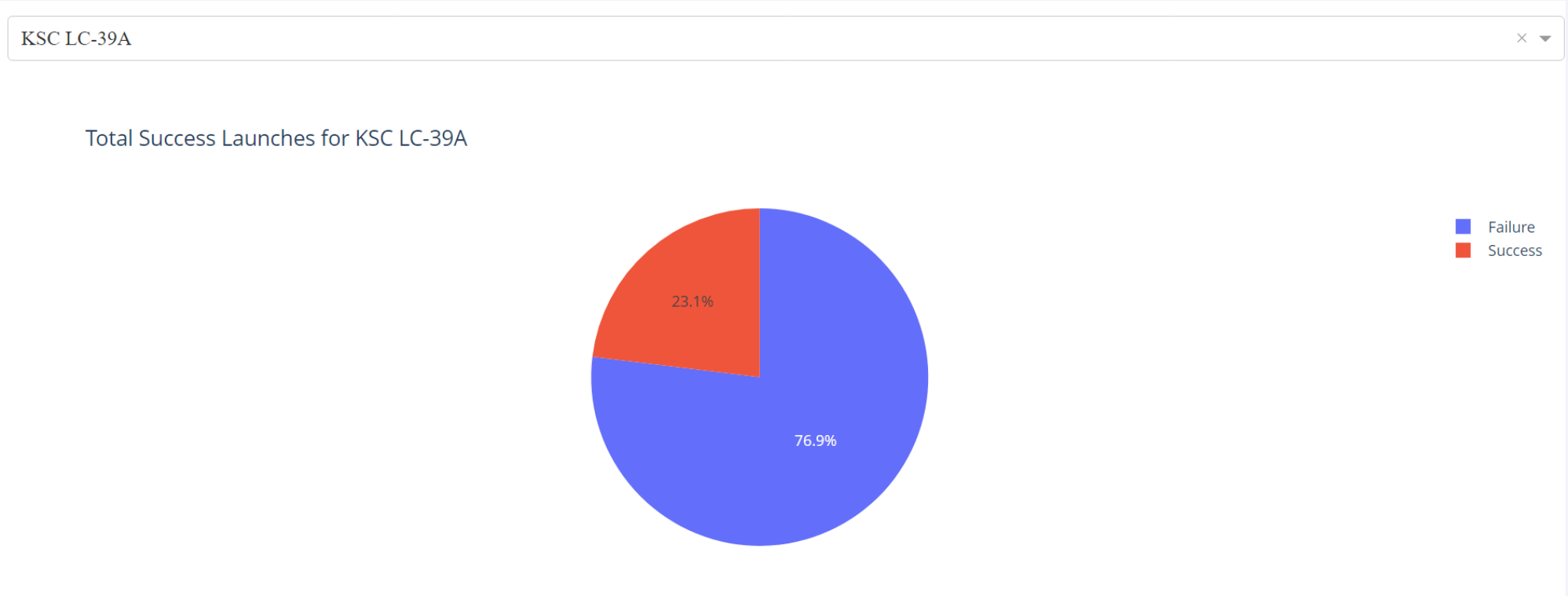

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by Site



We can see that KSC LC-39A Site has major success rate and CCAFS SLC-40 has least success rate

Launch Success Ratio for KSC LC-39A



77% of the launches were successful from KSC LC-39A launch site which is the highest among the others.

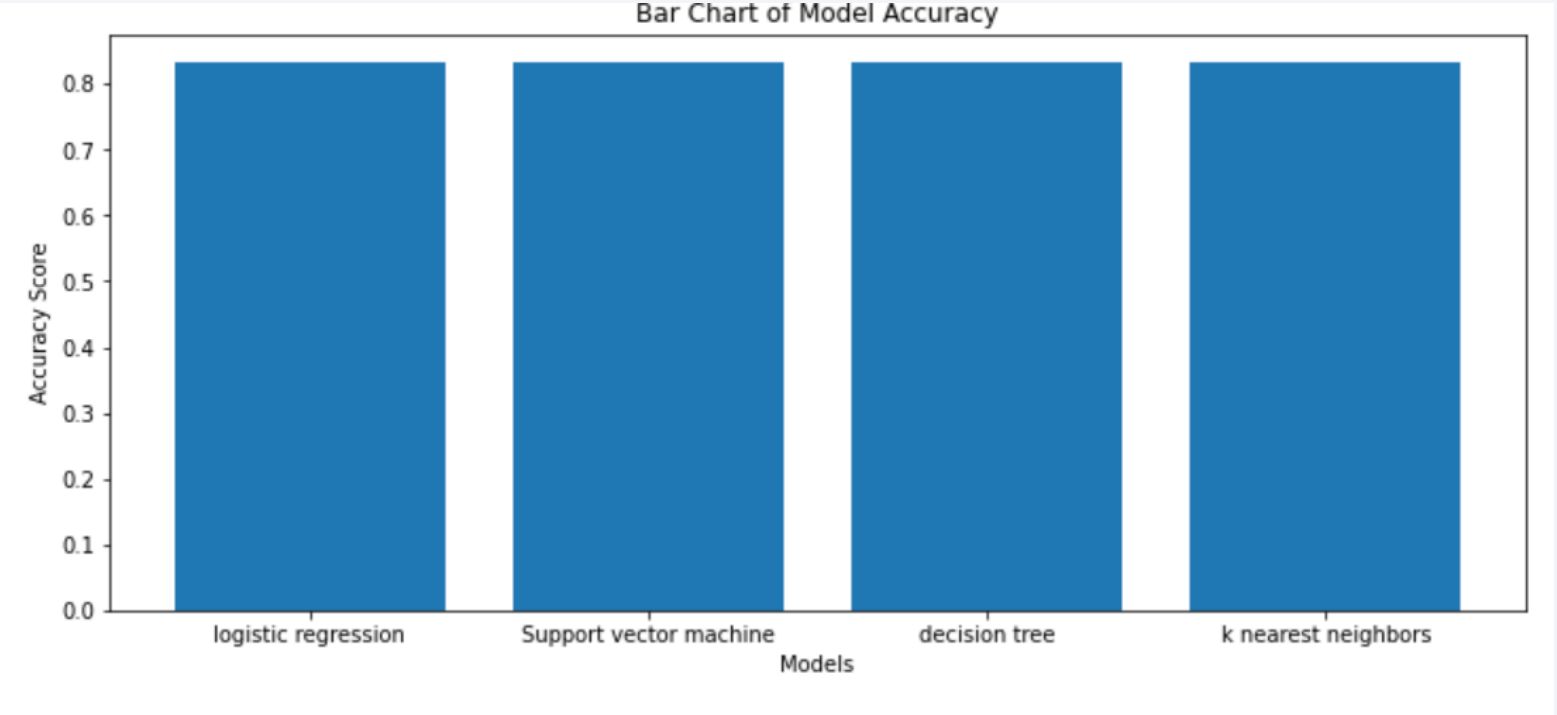
Payload vs Launch Outcome for All Sites

There are more failure rates as compared to successful launches for all sites with a payload capacity between 2500 kg and 7500kg

Section 5

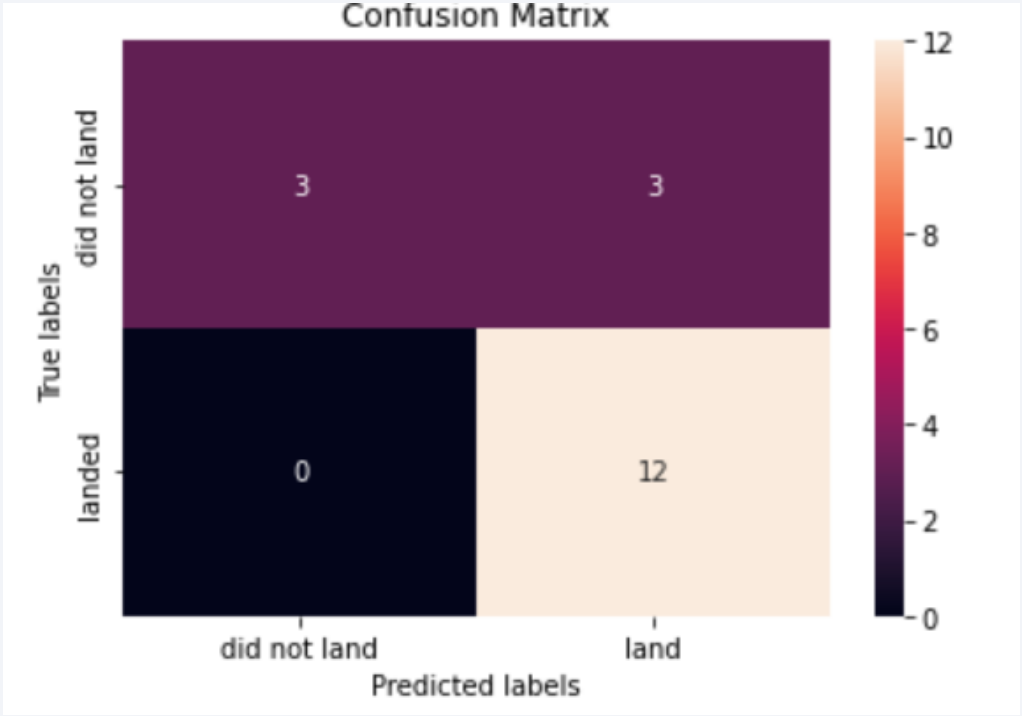
Predictive Analysis (Classification)

Classification Accuracy



- All models produced the same accuracy against the test data set (~83.33%).
- This is likely due to the limited data used, training and testing the models on larger data sets may produce more varied results.

Confusion Matrix



- All of the models produced the same confusion matrix
- The model correctly predicted all of the successful landings. The model predicted half of the unsuccessful landings correctly ad over predicted the landing outcomes which led to false positives

Conclusions

- Our main goal is to develop a model to predict the landing. By this Capstone Project we have developed different classification models with best parameters using GridSearch() technique and obtained an accuracy of 83% on test set.
- Insights :

1.KSC LC-39A has the highest amount of successful landings

2.Launch Sites are near to the coastlines

3. ES-L1, GEO, HEO, SSO Orbits have 100% success rate

4. Over the times the success rate has been improved

5. Most of the launches in the recent times have been mostly from CCAFS SLC-40

Appendix

- Github url of repository

- Refer the above link for Github repository of the project

Thank you!