IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vishal Gaware
16 Jan 2022

Outline

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

Executive Summary

- SpaceY is a commercial space company who looks to compete against the market leader SpaceX
- SpaceX advertises the cost of $62 mn for each rocket launch while other providers do the same for $165 mn making it more than $100 mn cheaper than other providers.
- Much of the cost reduced by SpaceX is due to their landing and reusability of first stage rocket.
- At SpaceY we will use the public data available of SpaceX and apply Data Science to predict if SpaceX is reusing their first stage rocket.

Introduction

- Applying Data Science to predict the probability if the Falcon 9 rocket will be reused.
- Sometimes SpaceX will sacrifice the first stage due to mission parameters such as payload, orbit, and customer.
- The report aims to accurately predict the likelihood of the first stage rocket landing.
- This will help in determining the launch cost of SpaceX next mission, and will help SpaceY to successfully bid against them to win the contract.

Falcon 9 vertical landing

Section 1

Methodology

Methodology

Executive Summary

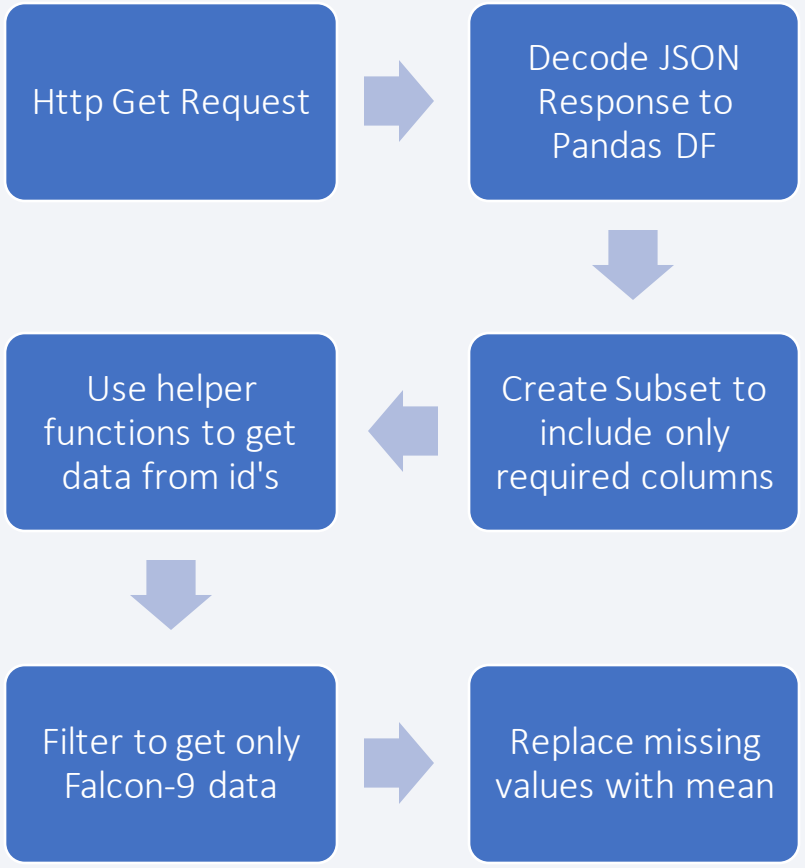
- Data collection methodology:
  - Using SpaceX API
  - Loading CSV file into IBM DB2 database
- Perform data wrangling
  - Webscraping the wikipedia page of SpaceX
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Finding the best hyperparameter for SVM, Classification Trees and Logistic regression

Data Collection

- SpaceX API
  - Historical launch data from Open Source REST API of SpaceX
  - Loading the data into the pandas dataframe, filtering it to only include Falcon 9 launches
  - Cleaning the data to replace missing values with mean
- Webscraping the wikipedia page
  - Wikipedia webpage used List of Falcon 9 and Falcon Heavy launches
  - Extracted data from HTML tables

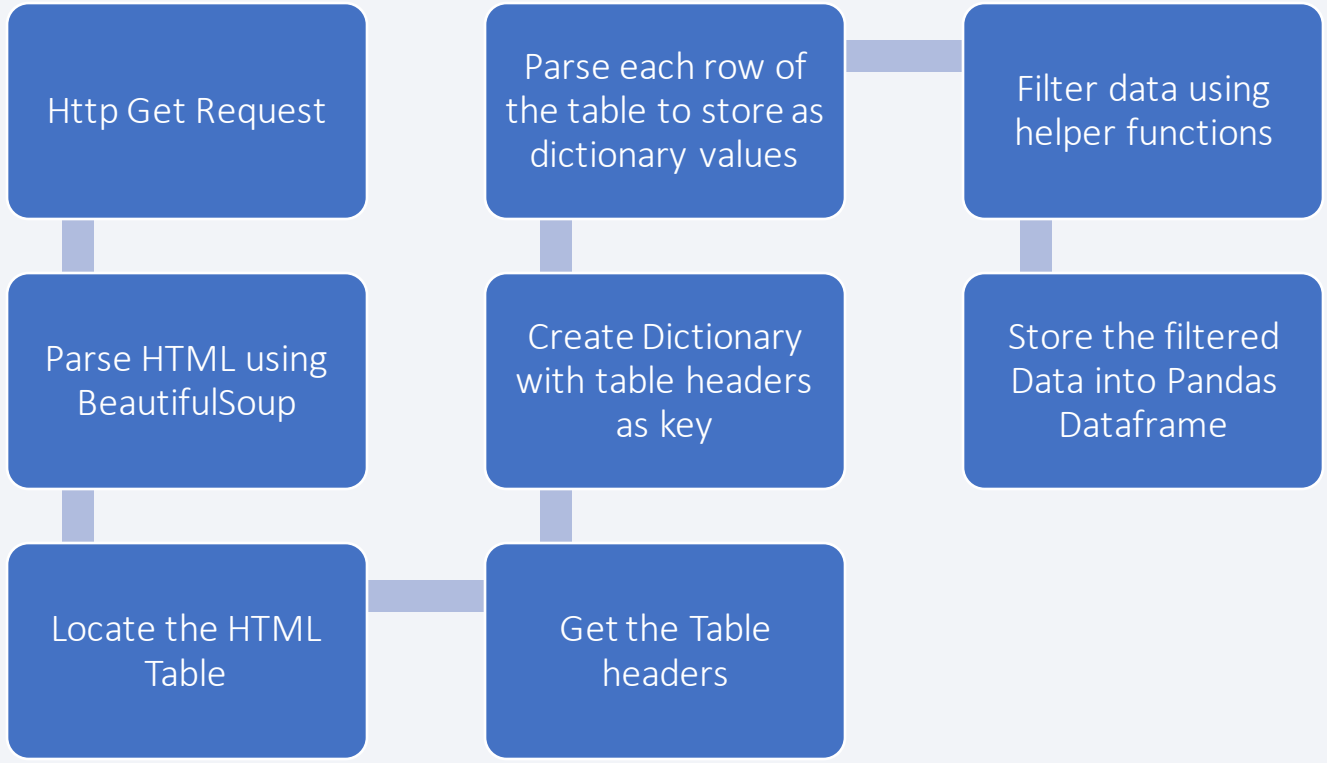
Data Collection – SpaceX API

- API
  URL: https://api.spacexdata.com
- Version: V4
- SpaceX Data Collection API Notebook



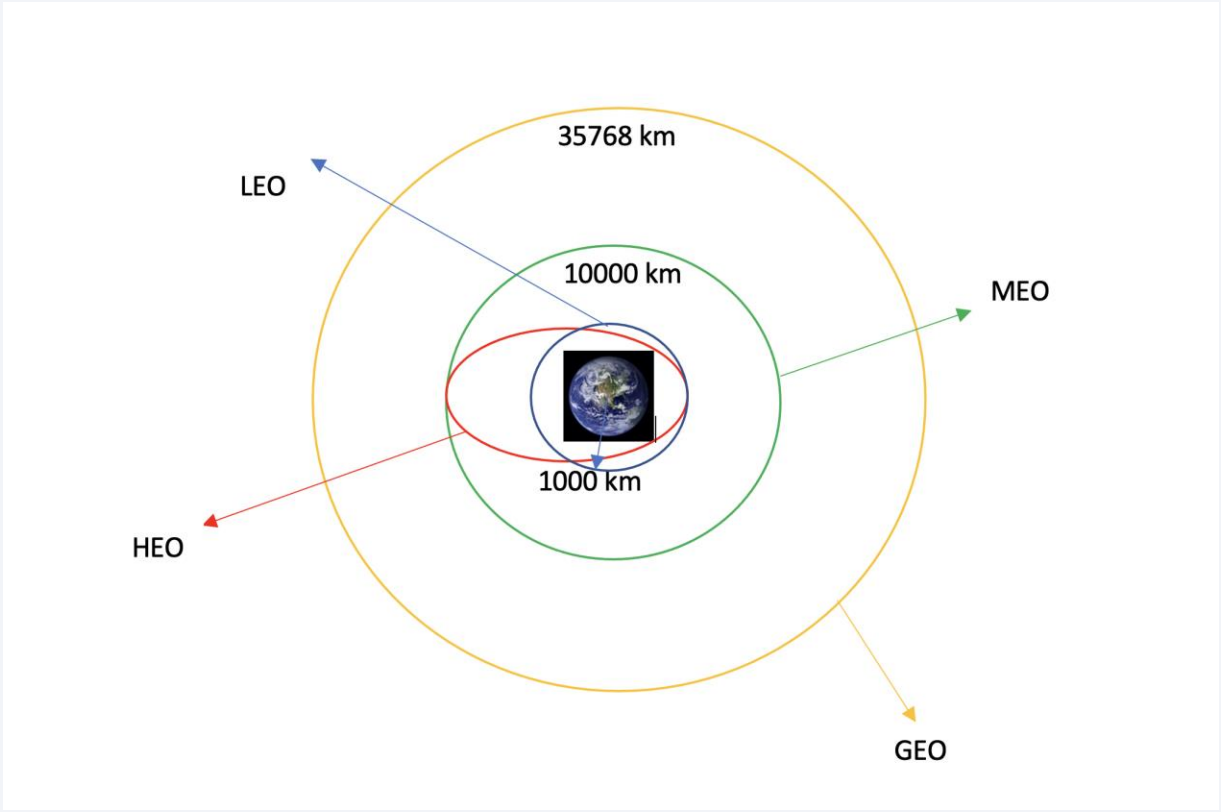
Data Collection - Scraping

- Wiki Page: List of Falcon 9 and Falcon Heavy launches
- Library: BeautifulSoup, Requests
- Web Scraping Notebook



Data Wrangling

- Perform Exploratory Data Analysis and determine Training Labels
- Exploratory Data Analysis
  - Calculate the number of launches on each site
  - Calculate number and occurrence of each orbit
- Training Labels
  - Calculate number and occurrence of mission outcome per orbit type
  - Create landing outcome label from outcome column



- Data Wrangling Notebook

EDA with Data Visualization

- Perform Exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib
- **Exploratory Data Analysis**
  - Payload Mass vs Flight Number
  - Launch Site vs Flight Number
  - Launch Site vs Payload Mass
  - Success Rate vs Orbit Type
  - Flight Number vs Orbit Type
- EDA with Data Visualization Notebook

EDA with Data Visualization

- **Exploratory Data Analysis**
  - Payload vs Orbit Type
  - Yearly Success trend

- **Feature Engineering**
  - Features used for success prediction:
    - Flight Number, Payload Mass, Orbit, Launch Site, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial
- EDA with Data Visualization Notebook

EDA with SQL

- **Information obtained through SQL queries**
  - There are 4 launch sites namely: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E
  - Total payload mass carried by boosters launched by NASA (CRS) is **45596 kg**
  - Average payload mass carried by booster version F9 v1.1 is **2928 kg**
  - First successful landing outcome in ground pad was achieved on **22nd December 2015**
  - Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 are **F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2**
- EDA with SQL Notebook

EDA with SQL

- **Information obtained through SQL queries**
  - The total number of successful and failure mission outcomes

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

  - Names of the booster_versions which have carried the maximum payload mass.

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

EDA with SQL

- **Information obtained through SQL queries**
  - The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 (Fig a)

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

Fig a

  - Ranking 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 (Fig b)

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

Fig b

Build an Interactive Map with Folium

- Taking Location data (Latitude and Longitude) we added a Circle Marker around each launch site with a label of the launch site name
- The column *launch_outcomes* contains information of the result of the launch (success or failure) using this information two markers Green and Red on the map in MarkerCluster represent success and failure respectively.
- Using **MousePosition** to calculate distances between launch site to its proximities such as railway, highway, coastline etc.
- Interactive Map with Folium Notebook

Build a Dashboard with Plotly Dash

- **Pie Chart:** shows success rate color coded by launch site
- **Scatter Chart:** shows payload mass vs landing outcome, color coded by booster version. **Range Slider** is used to select payload range
- **Drop Down:** For selecting individual launch site
- Plotly Lab Github

Predictive Analysis (Classification)

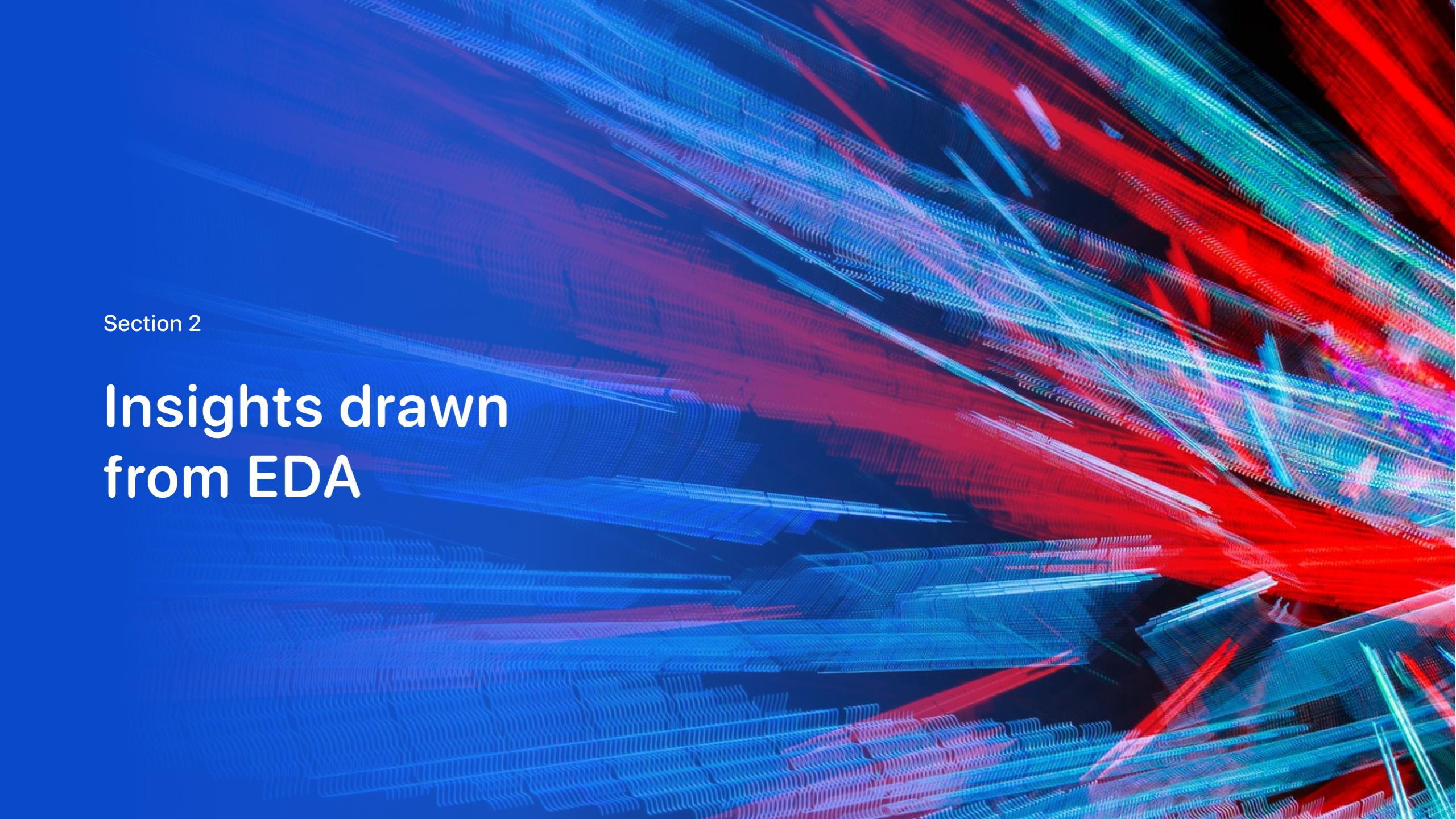
- **Model Building**
  - Loading data into Numpy and Pandas
  - Standardizing and Transforming the data
  - Splitting data into train and test dataset (80% train, 20% test dataset)
  - Algorithms used for training
    - Logistic Regression
    - SVM
    - Decision Tree Classifier
    - KNN
- Machine Learning Prediction Notebook

Predictive Analysis (Classification)

- **Model Evaluation**
  - Checking accuracy for each model
  - Plotting Confusion Matrix
- **Model Improvement**
  - Getting tuned hyperparameters to get the best accuracy
- Best Performing Model
  - All the models gave the same accuracy while tested on test data

Results

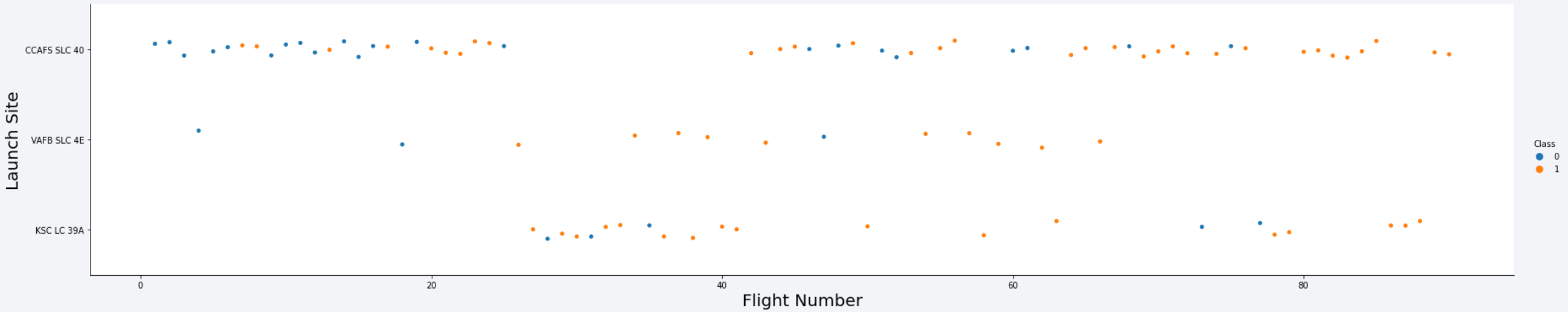
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

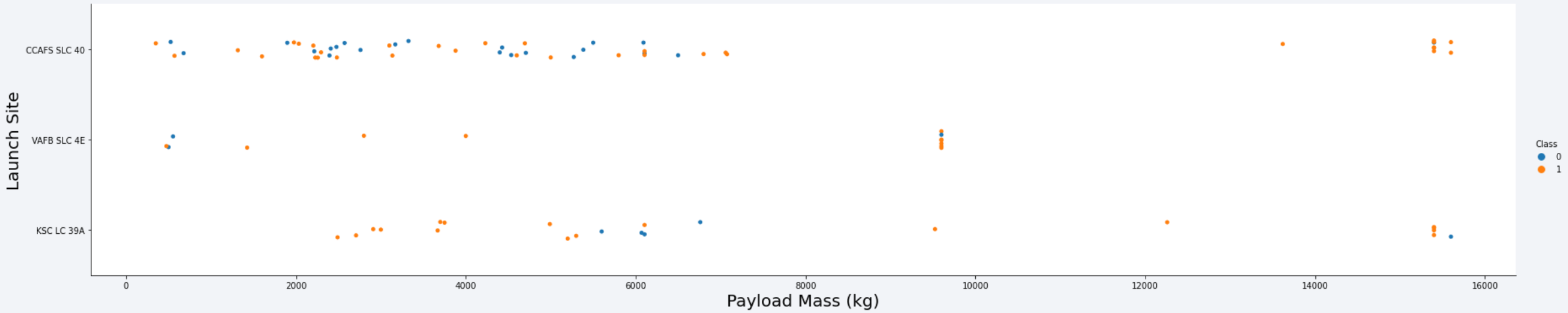
Insights drawn from EDA

Flight Number vs. Launch Site



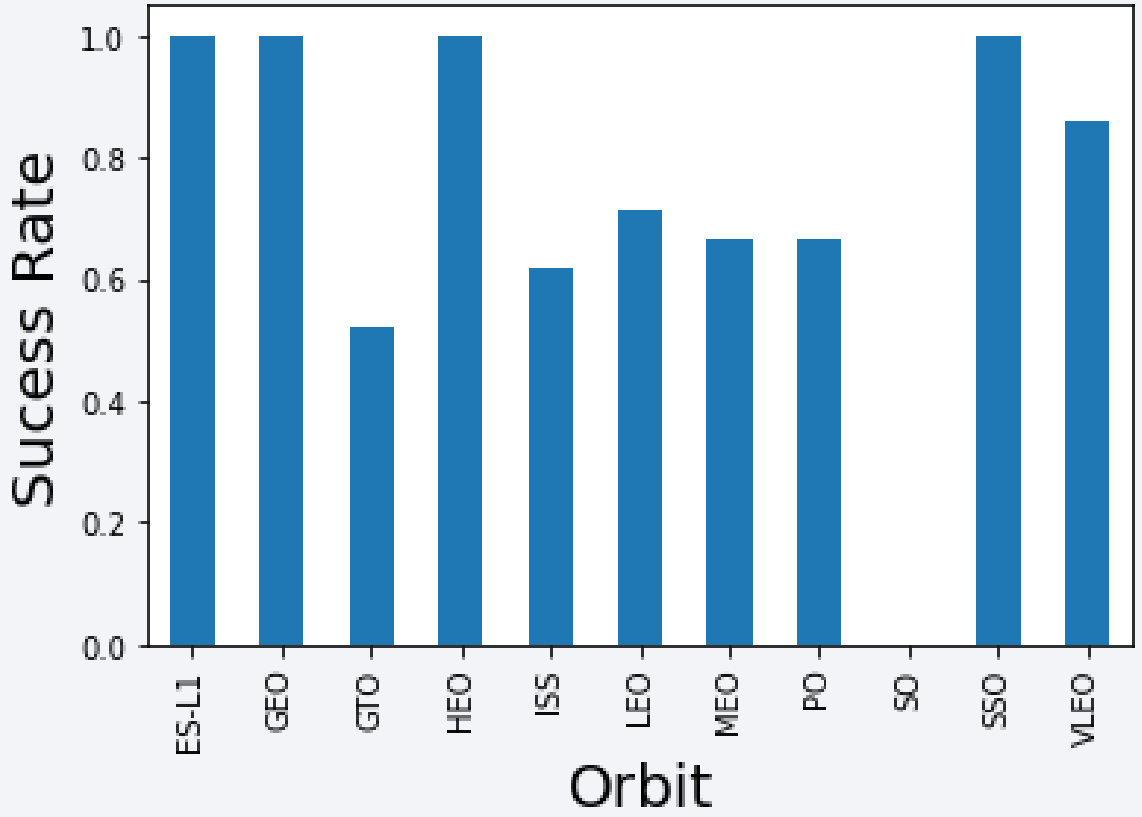
- For CCAFS SLC 40 and VAFB SLC 4E as the Flight number increases there is less probability for failed mission

Payload vs. Launch Site



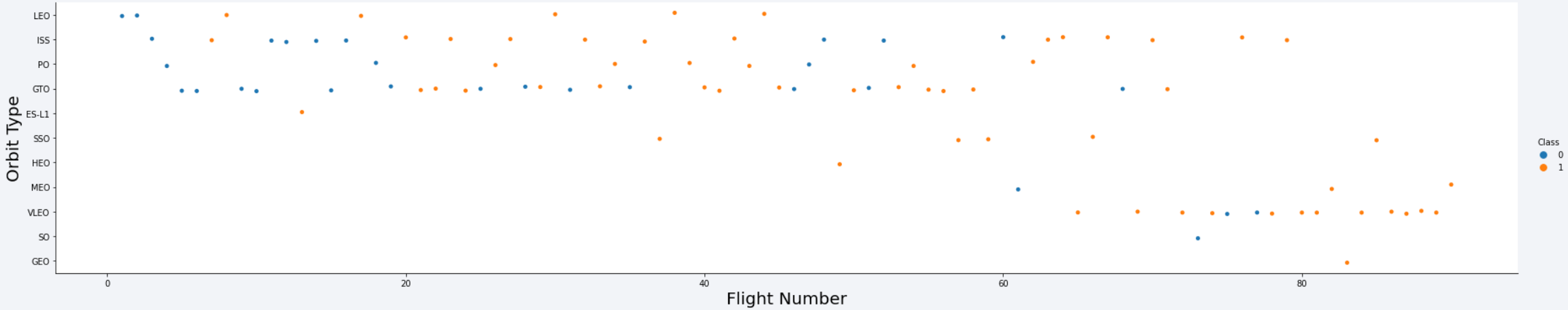
- For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type



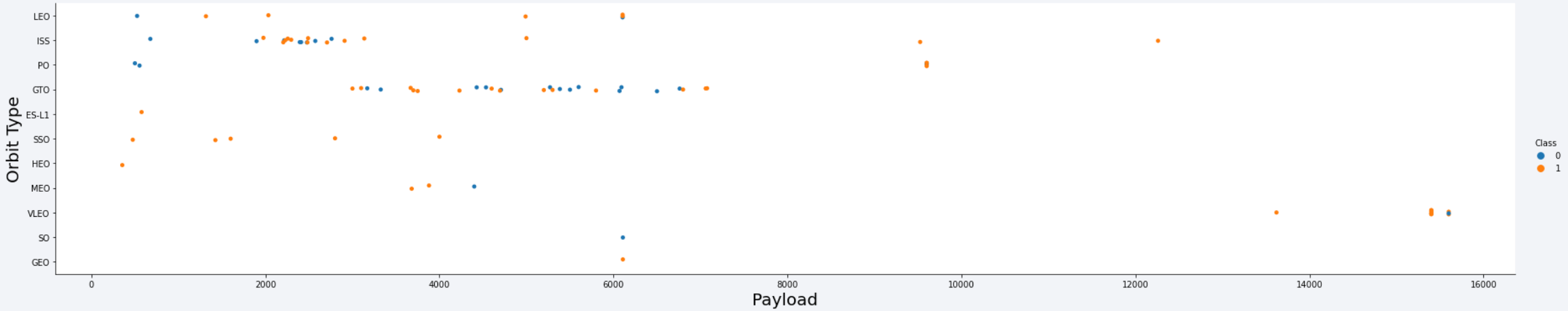
- **Orbits with high success rates**
  - ES – L1
  - GEO
  - HEO
  - SSO
  - VLEO

Flight Number vs. Orbit Type



- LEO orbit the Success appears related to the number of flights.
- on the other hand, there seems to be no relationship between flight number when in GTO orbit.

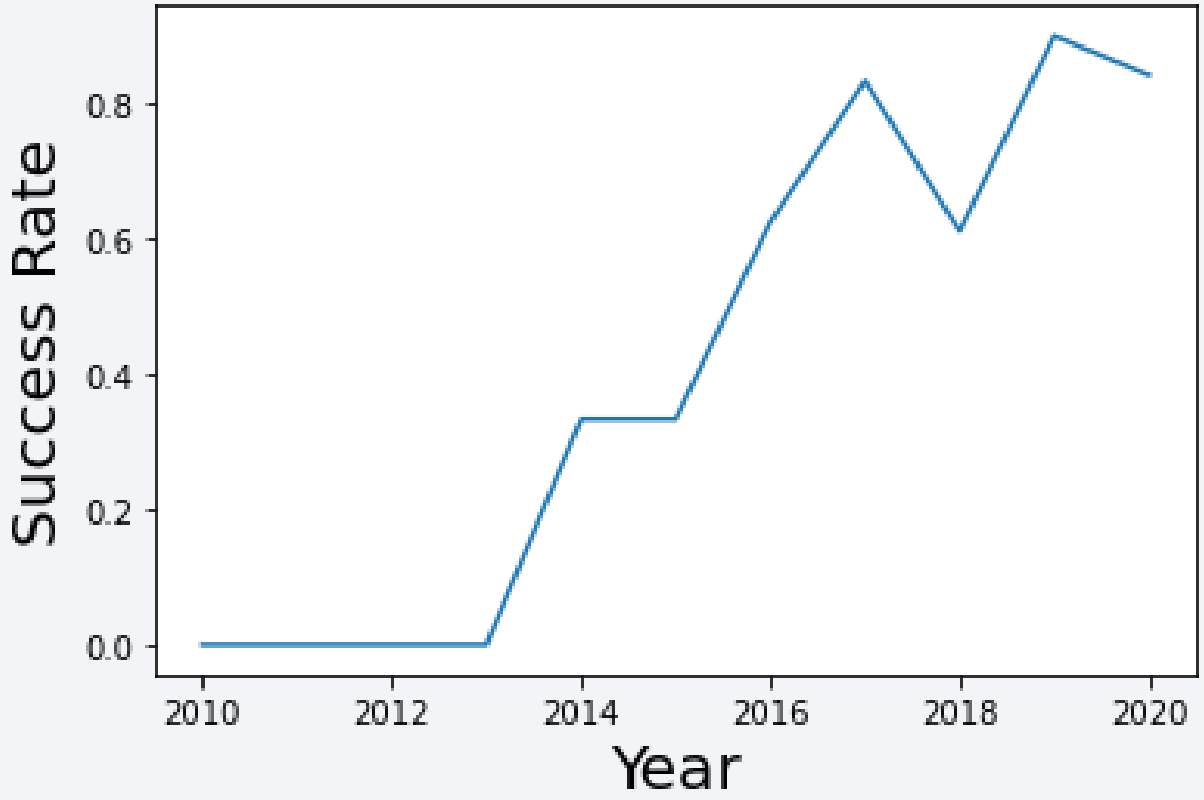
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

The sucess rate since 2013 kept increasing till 2020



All Launch Site Names

Query

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

The ***distinct*** keyword eliminates duplication, hence the site names are not repeated and we get all the launch site names.

Result

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

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXDATASET 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

- As seen in the previous query, there are two sites that begin with CCA, i.e CCAFS LC-40 and CCA SLC – 40 respectively
- Since we selected only first 5 rows the results shows only CCAFS LC – 40
- The ***'%'*** wildcard used after the ***like*** keyword in the query is used for text search. It is used to match the pattern of the string, when used after CCA ('CCA%'), it ignores rest of the string and checks only first 3 characters to match if the string begins with 'CCA'

Total Payload Mass

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

total_payload
45596

- The ***sum*** is an inbuilt function which adds all the values for the column *payload_mass__kg_*
- Additionally we specified ***where*** clause to include only NASA customers

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as avg_payload from spacexdataset where booster_version='F9 v1.1';
```

avg_payload
2928

- The ***avg*** function here calculates the average payload load mass from the *payload_mass__kg_* column.
- By using the ***where*** clause, we specify to include only those payload mass whose booster version is **F9 v1.1**

First Successful Ground Landing Date

```
%sql select date from spacexdataset where landing__outcome='Success (ground pad)' order by date asc limit 1;
```

DATE
2015-12-22

- This is a complex query.
- First we find all the results which have the *landing__outcome* as **"Success (ground pad)"**
- Next we order the results in descending order w.r.t date so that oldest date comes first and the latest date comes last
- Finally we limit the result to just one row, hence we get the first successful Ground Landing Date.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
select booster_version from spacexdataset
where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ > 4000 and payload_mass__kg_ < 6000;
```

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

- In this query we give 3 conditions
- In first condition we select the *landing__outcome* to **"Success (drone ship)"**. This selects all the rows who has successful drone ship landing
- The second condition specifies *payload_mass__kg_* to be greater than 4000, this selects all the rows which has successful drone ship landing and payload mass greater than 4000 kg
- This final condition specifies the *payload_mass__kg_* to be less than 6000, hence now only those successful drone ship landing are selected that have payload mass greater than 4000 kg but less than 6000 kg.

Total Number of Successful and Failure Mission Outcomes

```
%%sql
select mission_outcome, count(mission_outcome) as total
from spacexdataset group by mission_outcome;
```

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

- The main clause here is the ***group by*** clause.
- The ***group by*** is used to arrange identical data into groups. In our query we use ***group by*** to group all the rows w.r.t *mission_outcome.*
- We then count all the mission outcomes for a particular mission outcome value. Here we have 3 mission outcomes and their counts are represented in ***total*** column in the above figure.

Boosters Carried Maximum Payload

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

- To get the desired result we have used the nested query here.
- The exact maximum payload weight is not known, also if it is known earlier there are chances that it gets replaced by new maximum payload and the earlier known maximum payload gets outdated.
- Therefore we use the ***max*** function to get our current maximum payload.
- The result of ***max*** is used as input value in the ***where*** clause of the above query.

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

2015 Launch Records

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

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

- There are two occasions where the booster failed to land on the drone ship in the year 2015.
- Since we have to find all the failed drone ship launch record in the year 2015, we specify two conditions in the where clause.
- In the first condition we get all the failed drone ship landing outcomes, next we filter the result to include only those landing outcomes of the year 2015, we use ***year*** function to get the results for 2015.

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

```
%%sql
select landing__outcome, count(landing__outcome) as total
from spacexdataset
where date > '2010-06-04' and date < '2017-03-20'
group by landing__outcome order by total desc;
```

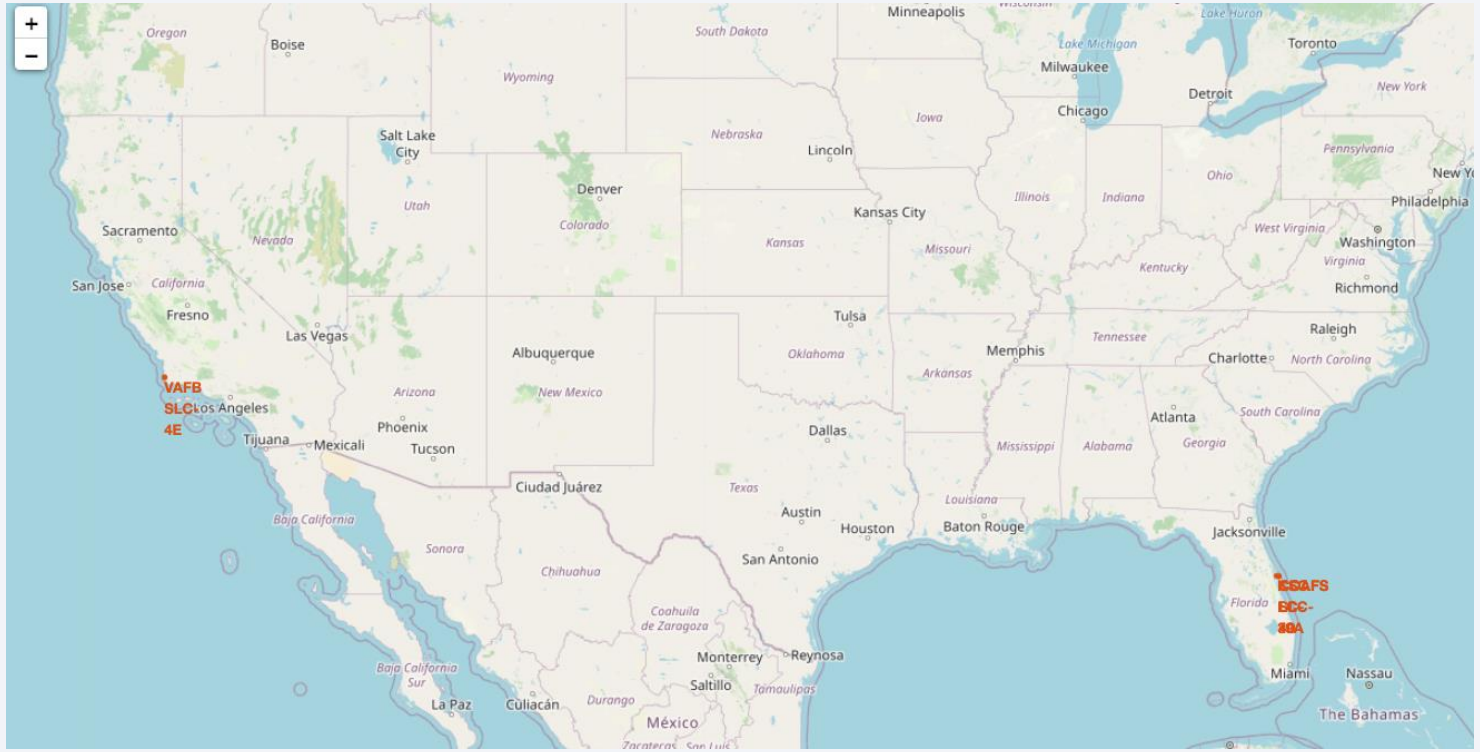
- We group the values for *landing__outcome* column
- Specify the date to be between 2010-06-04 and 2017-03-20
- Finally we order the results in descending order so that our results begin with the highest to the lowes w.r.t *total* column

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

Section 4

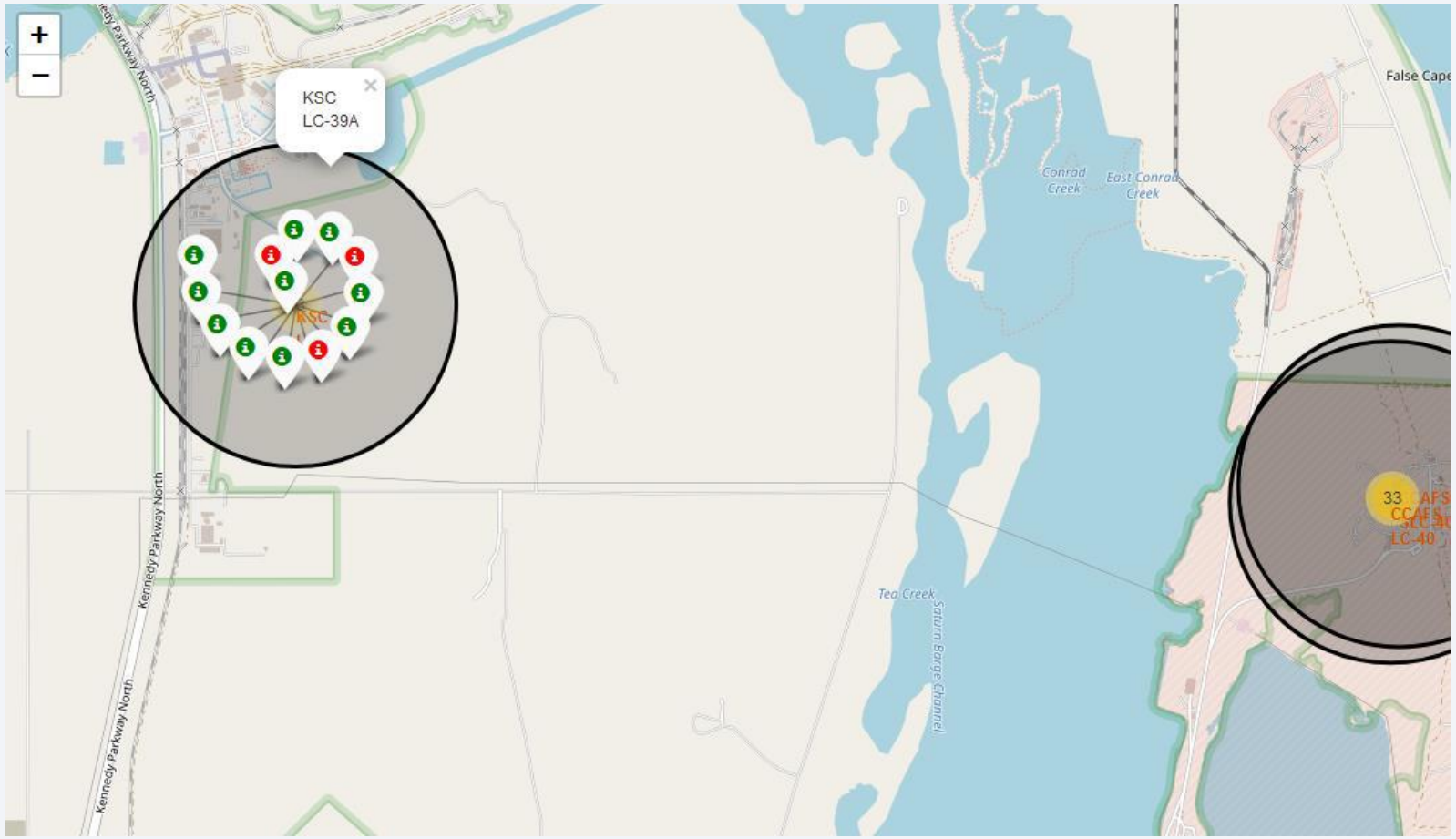
Launch Sites Proximities Analysis

Launch Site Markers



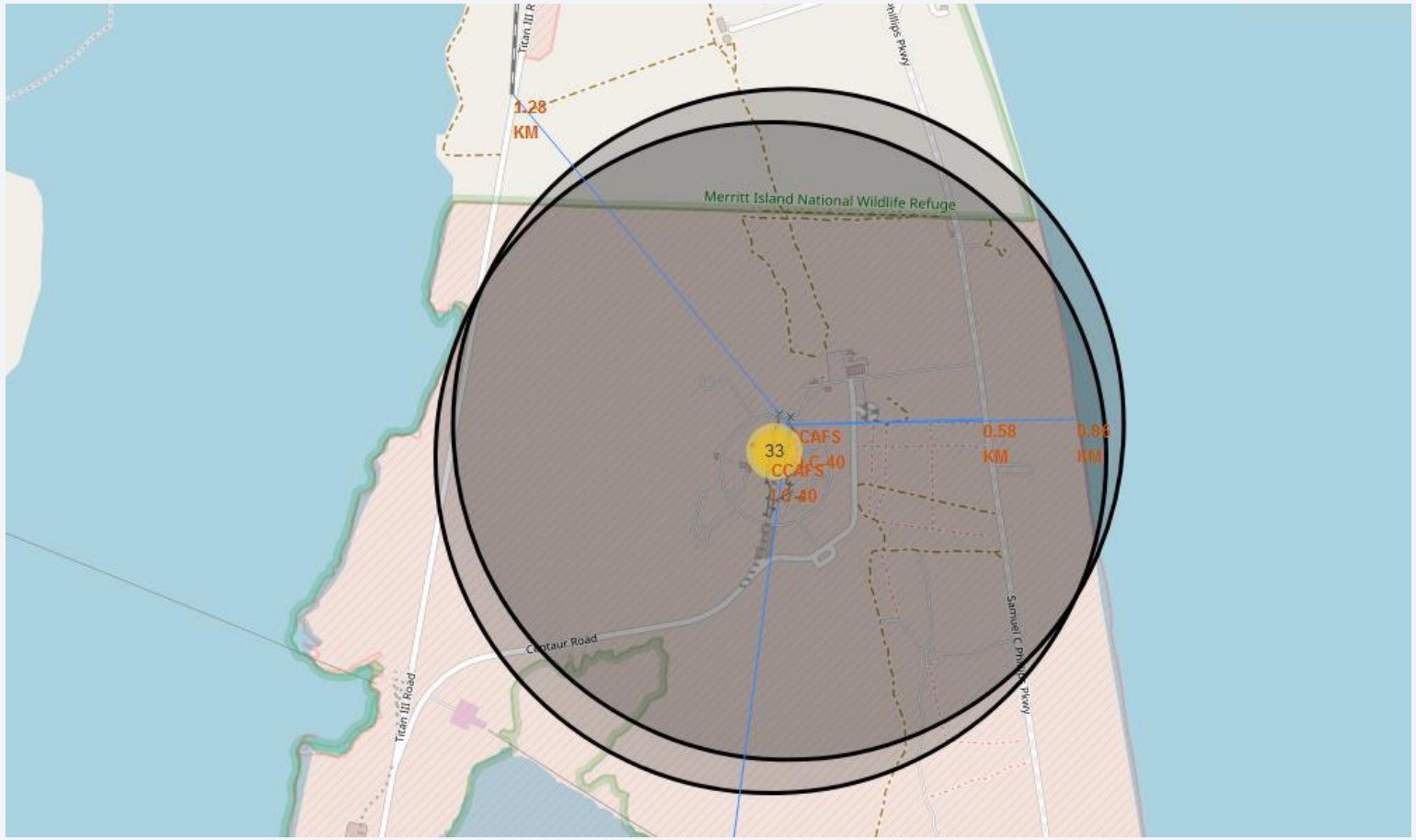
- Here we see two markers in red.
- One at the east coast and one at the west coast
- All the Launch Sites are near the coast and also nearby the equator.

Launch Site Analysis for site KSC LC-39A

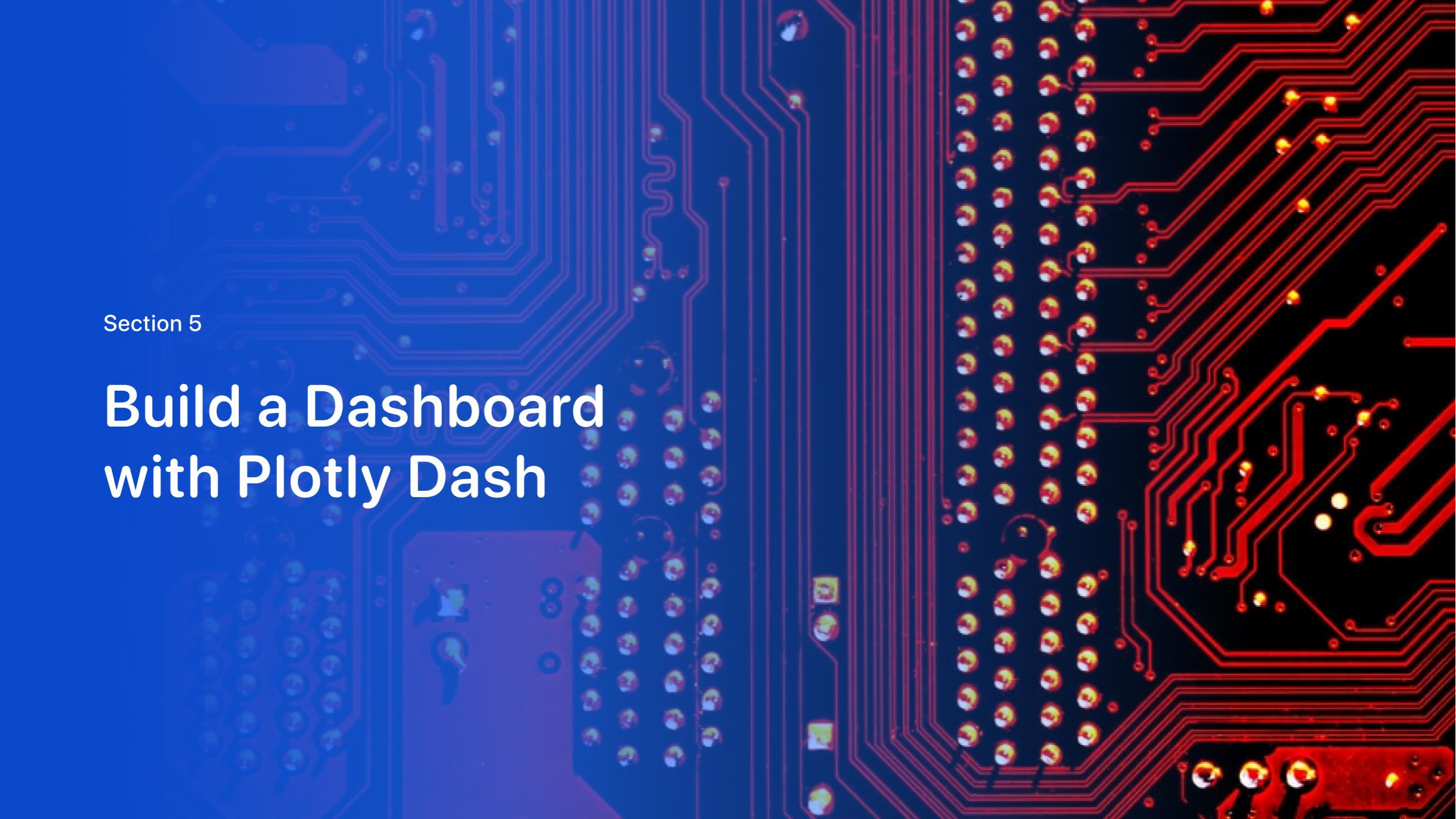


- This launch site has relatively higher success rate
- Markers are color coded to represent success and failure
- Green marker represents Success while Red marker represents Failure

Launch Site proximity Analysis



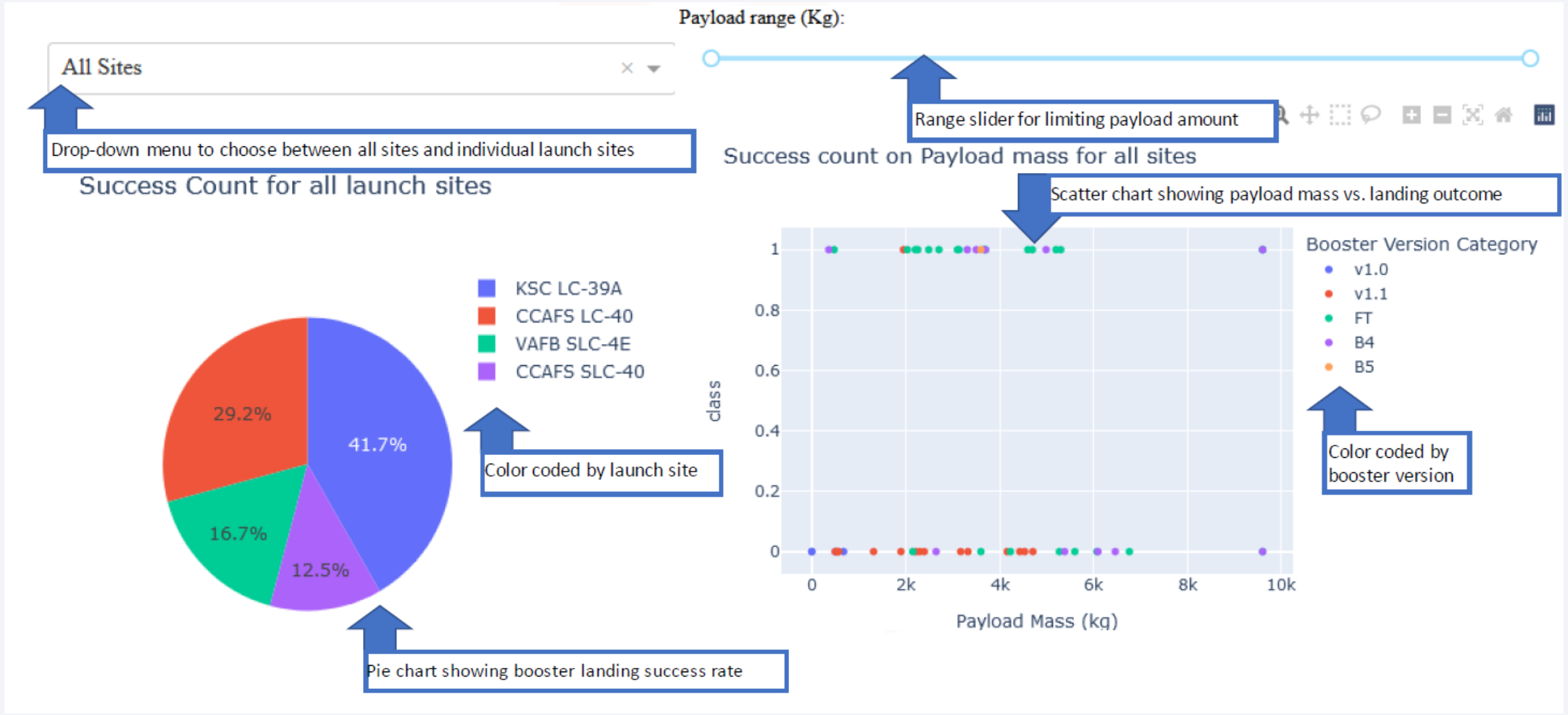
- The site chosen for proximity analysis is CCAFS SLC-40
- By proximity analysis we find that the site is 1.28 km away from the nearest railway (for carrying heavy cargo)
- The nearest highway is 0.58 km away
- Distance from the coastline to Launch Site is 0.86 km
- Distance from the city is 51.43 km



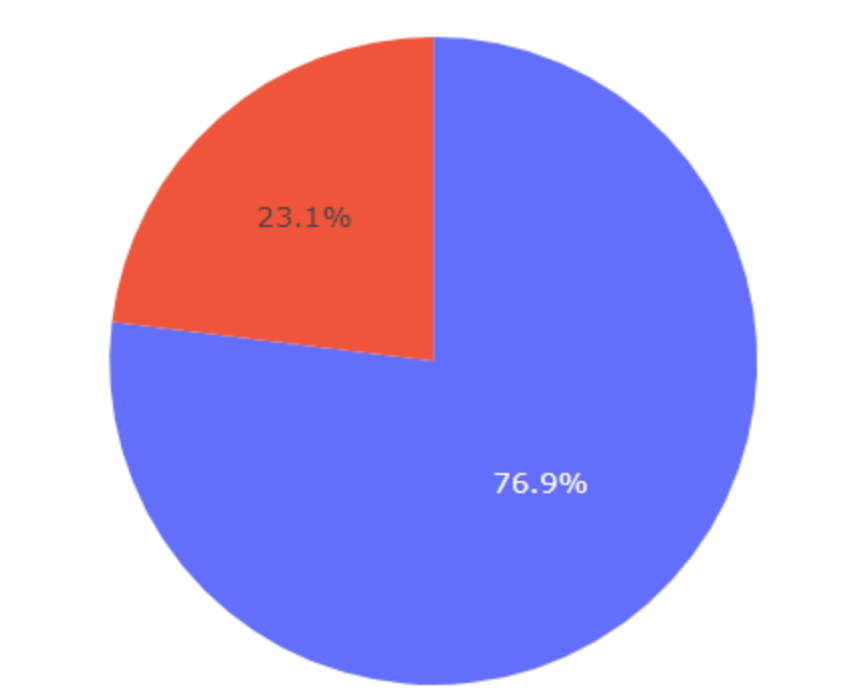
Section 5

Build a Dashboard with Plotly Dash

Plotly Dashboard for Launch Site and Payload



Success pie chart for site KSC LC - 39A



- The blue portion of the pie chart shows success, whereas the red portion shows the failure.
- The Launch Site **KSC LC - 39A** has the highest success ratio with **76.9%** success and **23.1%** failure

Scatter Plot: Low vs High payload success rate

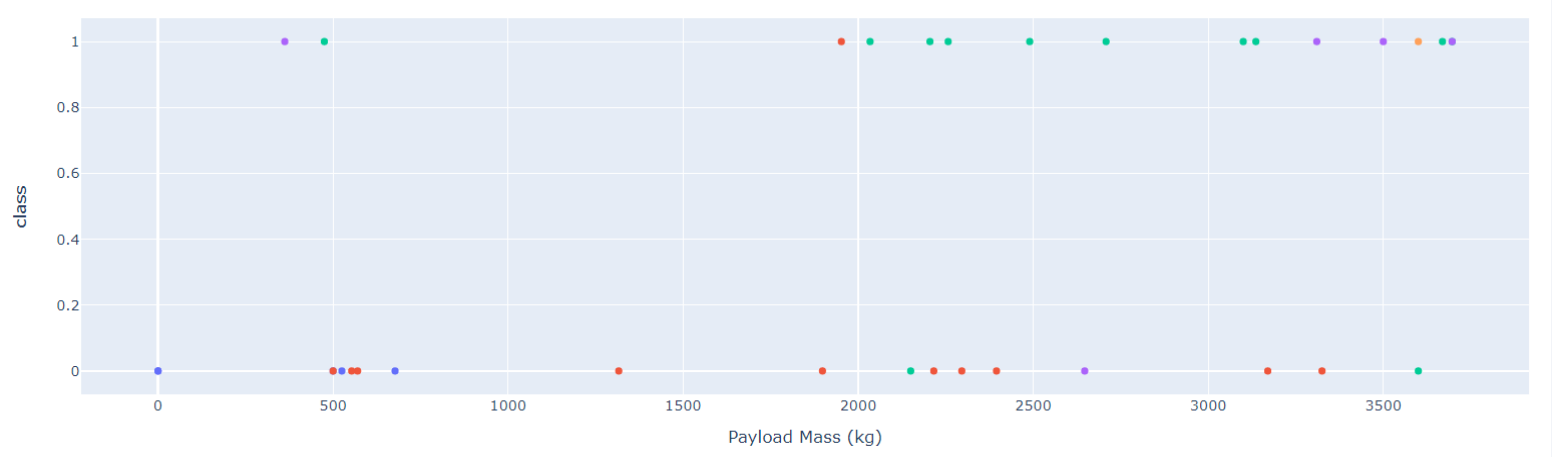


Fig a: Payload range 0-4000 kg (Low weight)

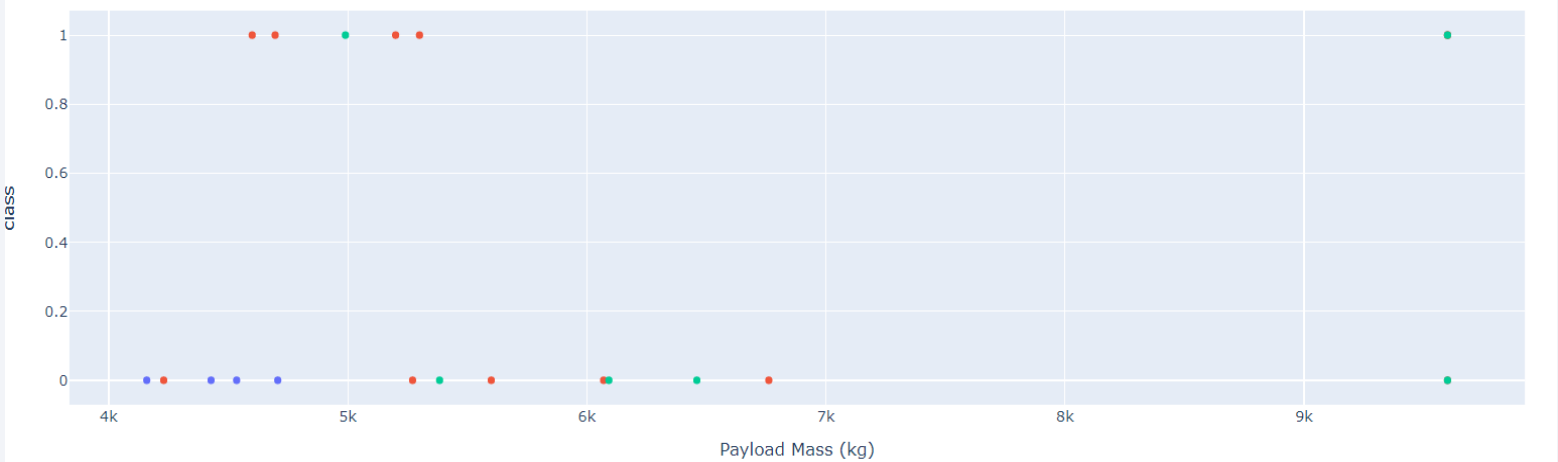


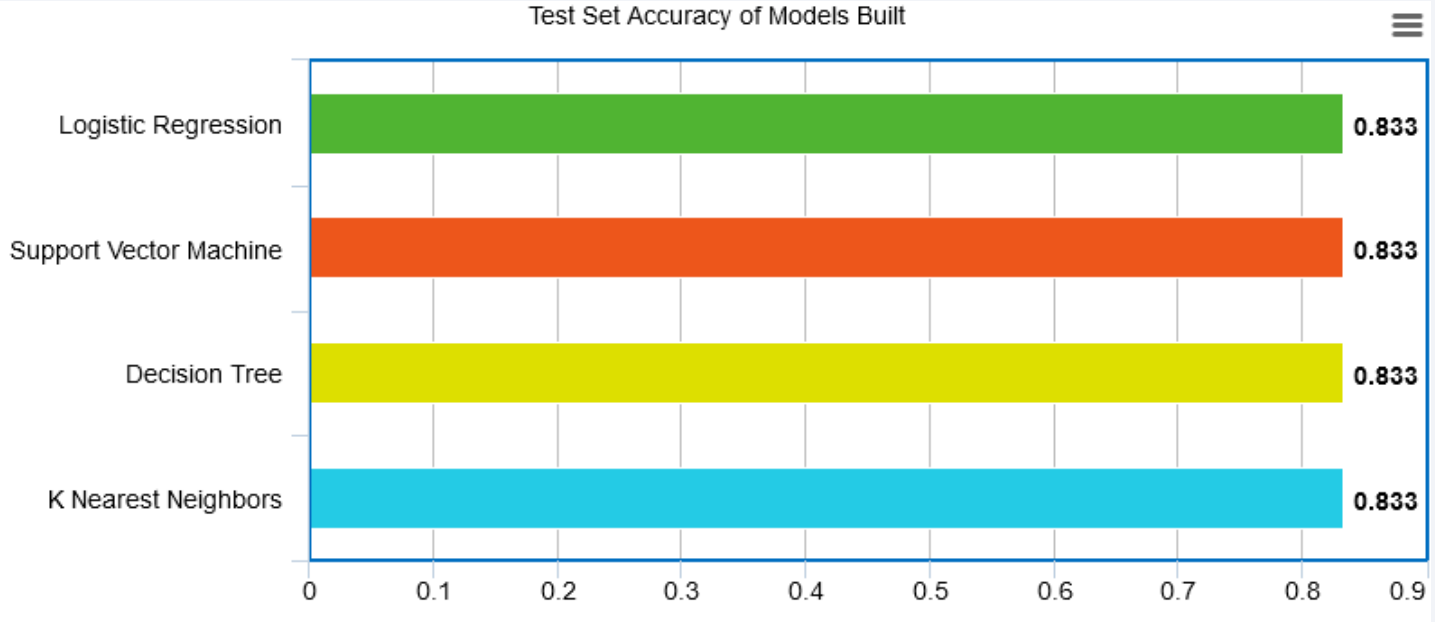
Fig b: Payload range 4000 - 10000 kg (High weight)

From the two scatter plots we can say that launches with low payload have higher success rates compared to higher payload

Section 6

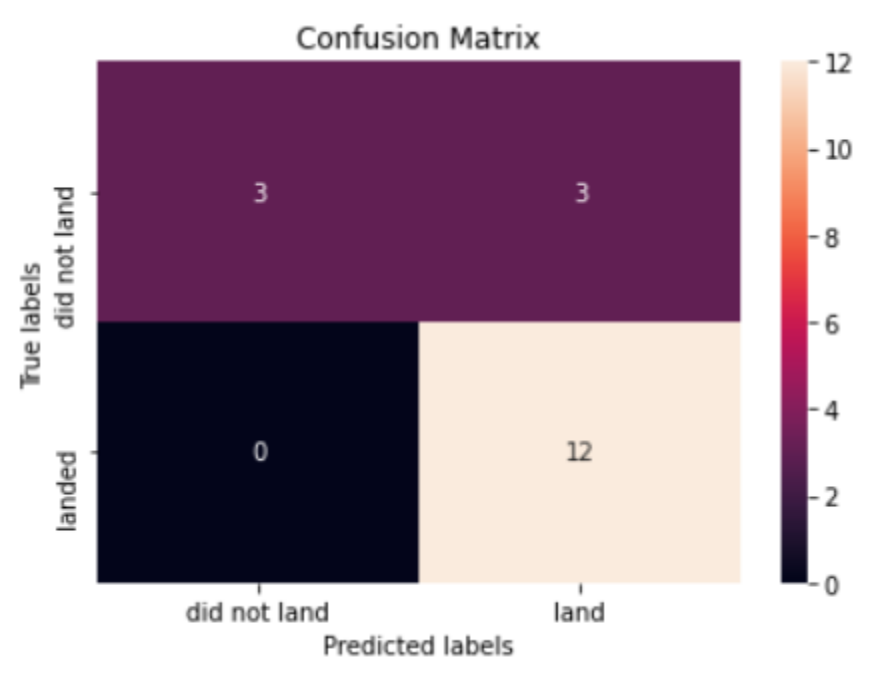
Predictive Analysis (Classification)

Classification Accuracy



When the accuracy was calculated for all the four algorithms, we found out that all of them had the same accuracy of **83%**

Confusion Matrix



- The confusion matrices of the best performing models (4 way tie) are the same
- The major problem is false positives as evidenced by the models incorrectly predicting the 1st stage booster to land in 3 out of 18 samples in the test set

Conclusions

- All the prediction algorithms had the accuracy of **83%**
- Orbits **GEO, HEO, SSO, ES-L1** had better success rates
- Launch Site **KSC LC-39A** had best success ratio of **77%**
- Success rate for SpaceX kept on getting better with coming flights
- Launches with low payload had better success rate compared to launches with higher payload

Appendix

- **Notebooks**
  - SpaceX Data Collection API
  - Data Collection with Web Scraping
  - Data Wrangling
  - EDA with Data Visualization
  - EDA with SQL
  - Interactive Visual Analytics with Folium lab
  - Machine Learning Prediction

- **Github Link**
  - Applied Data Science Capstone

Thank you!