IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gauthama S Nair
4-3-2024

Outline

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

Executive Summary

- Data was collected using both web scraping and api. Data was cleaned and sufficient field added. Data analysis was done using SQL and data visualization. Predictive models were built and compared.
- The results collected show that success rate began to increase for launches as flight number increased. ES 11, GEO, HEO and SSO were orbits with highest success rate. Success rate increased with payload mass.

Introduction

- In this project we will predict if the Falcon 9 first stage will land successfully or not. Much of the savings of Spacex is because of reusing first stage. If we are able to determine this, we will be able to predict the cost of launch, for an alternate company.
- Some of the questions that need to be answered include which launch site, orbit has more successes? How success is impacted by payload mass, booster version? How does flight number affect success?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using Spacex Api data of past launches were collected and relevant subsequent api calls were made. Data was cleaned and filtered for Falcon 9 launches.
  - Also web scraped Wiki page of the Falcon 9 launches and cleaned the data.
- Perform data wrangling
  - The various types of landing outcomes were classified into success or failure and added as a field to the dat
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

- Perform predictive analysis using classification models
  - LogisticRegression, SVCClassifier, DecisionTreeClassifer and KNeigborClassifier models were used to train and test the data. GridSearchCV was used for hyper-parameter tuning. All were performing equally well on test data.

Data Collection

- The data was collected using 2 methods
  1. Using APIs
  2. By Web Scraping Wikipedia page.

Data Collection – SpaceX API

Launches API	Relevant Fields	Additional Fields	Falcon 9 Launches	Missing Values
Get data for past launches using Spacex API and store it in data frame.	Retain only relevant fields namely rocket, payloads, launchpad, cores, flight_number and date_utc.	Using rocket, launchpad, payloads in subsequent API calls to get booster version, latitude, longitude, launch site, orbit and payload mass.	Filter the data for Falcon 9 launches.	Check for missing values in columns of data frame. For missing values of payload mass use the mean value of the column.

GitHub URL

Data Collection - Scraping

Request Falcon 9 Wiki page	Extract all column names from header table	Create data frame by parsing html tables
Request the Falcon 9 Wikipedia page and parse it and create a BeautifulSoup object.	The third table is the required table. From this get all the elements with tag ‘th’ and then try and extract the contents of this and check if it is not just numbers. This would be the set of columns.	Create a dictionary of lists. Now parse the html table row by row. Get the flight number and check if it is only digits, if so go through the cells. From the cells extract all the columns and add it to the dictionary of lists with the corresponding key. Now convert the dictionary to data frame.

Github URL

Data Wrangling

Get Launch Site numbers	Get Orbit numbers	Get Outcome numbers	Add variable Class
For each Launch Site get value counts to get the numbers.	For each Orbit get value counts to get the numbers.	For each Outcome get value counts to get the numbers. This also help us in understanding what are the various types of outcomes, and which can be classified as success and failure.	Add a variable class which if the outcome is present in failure outcomes is of value 0, else it is of value 1.

GitHub URL

EDA with Data Visualization

- Flight Number - Pay load Mass(kg)

  A cat plot for these two with Class as hue shows that as flight number increases, the chance of success increases. Similarly as pay load mass increases the chance of success increases.

- Flight Number - Launch Site

  A cat plot for these two with Class as hue shows that CCAFS LC-40 has the most launches. While initially there are lot of failures, as flight numbers increased, more success was observed. VAFB SLC 4E had initial couple of failures, followed by more successes. KSC LC 39 A has it’s first flight number later, hence starts out with success.

EDA with Data Visualization

- Pay load Mass(kg) - Launch Site

  A cat plot for these two with Class as hue shows that, for VAFB SLC 4E there have been no launches with more than 10000kg payload. Both the others have some, and they tend to be more successful.

- Orbit - Class

  A bar plot for these two shows that SO orbit has had no successes. ES L1, GEO, HEO and SSO have all successful launches. GTO has the most failure/success ratio.A bar plot for these two shows that SO orbit has had no launches. ES L1, GEO, HEO and SSO have all successful launches. GTO has the most failure/success ratio.

- Flight Number - Orbit

EDA with Data Visualization

A cat plot for these two with Class as hue shows that, for LEO the initial flight numbers were failures, followed by successes. For GTO there are failures in initial flight numbers and later on as well. VLEO orbit was used later on and SO has only 1 launch.

- Pay load Mass(kg) - Orbit

  A cat plot for these two with Class as hue shows that VLEO has most of the high payload launches. For GTO there is no relationship between Pay load Mass and success, since failure is present in both high and low payloads.

- Date - Class

  A line plot for these two(with mean of class) shows that there were no successes till 2013, but after that the mean has gradually increased to about 0.8.

GitHub URL

EDA with SQL

- Initially the top 5 rows of SPACEXTBL were fetched.
- Next top 5 rows whose Launch Site starts with CCA were fetched, this was done by matching Launch Site with ‘%CCA%’ using like.
- Next sum function was used on Payload Mass whose customer was NASA(CRA)
- Next avg function was used on Payload Mass whose Booster Version was F9 v1.1
- Next first success was checked for using min function in Date and checking for landing outcome like ‘%Success%’
- Next Booster Version was checked for rows with Landing Outcome

EDA with SQL

Success (drone ship) and Payload Mass between 4000 and 6000 kg.

- Next count function was checked for based on Mission Outcome
- Next using sub query and function max on Payload Mass the maximum was checked for and the Booster Version of these listed.
- Next using substr function on Date the month and year was taken and used for listing the month, Landing Outcome, Booster Version and Launch Site in the year 2015 for all landing outcome which was of type drone ship, and which was not Success (drone ship).
- Next between 2010-06-04 and 2017-03-20 Date, Landing Outcome was grouped and count function was applied and sorted in descending order of count.

GithHub URL

Build an Interactive Map with Folium

- Mark all launch sites on a map

  A data frame for Launch Site was created. This was iterated and for each row, a corresponding circle was added into the map, based on the latitude and longitude of the site.

- Mark all success/failed launches for each site on a map

  Each row of the data frame is iterated, and based on the success and failed color added and then the marker is added to the map.

- Calculate the distance between a launch site and its proximities

  Choose a launch site and coastal point, calculate the distance between them based on latitude and longitude. Add a distance marker on the coastal point. Also add a polyline from launch site to coastal point. Same way choose a city, calculate the distance between city and launch site using latitude and longitude. Add a distance marker on the city. Add the polyline from city to launch site.

GitHub URL

Build a Dashboard with Plotly Dash

- Success Pie Chart with Dropdown

  A success pie chart is added with dropdown. There are 2 scenarios to be shown on the chart. In case of all sites option, chosen on dropdown. The pie chart shows the percentage count of successes each launch site contributed. In case a particular site is chosen, the pie chart shows the success to failure in percentages.

- Scatter Chart with Dropdown and Range Slider

  A scatter chart is added which depends on dropdown as above as well as range slider. The scatter chart is for payload mass and class with booster version as hue. The range of payload mass is taken from range slider. If all sites are chosen all launches are considered, else only launch specific to site

GitHub URL

Predictive Analysis (Classification)

Preprocess Data	Logistic Regression	Support Vector Machine	Decision Tree Classifier	K Nearest Neigbors
Assign Class as the target variable and others as feature variables. Transform feature variables using Standard Scalar. Split the data into train and test data for both feature and target variable.	Using GridSearchCV, check for C parameter with 0.01, 0.1 and 1 in logistic regression on train data. The solver and penalty are kept constant at lbfgs and l2 respectively. C as 0.01 performs best, use this to predict test data and check accuracy and Confusion Matrix.	Using GridSearchCV, check for kernel with linear, rbf, poly and sigmoid, C and with values between -3 and 3(equally split into 5) for svm. Sigmoid kernel with C as 1.0, and gamma around 0, works best for train data. Use this to predict test data and check accuracy and Confusion Matrix.	Using GridSearchCV, check for criterion with entropy and gini, splitter with best and random, maxfeatures with log2 and sqrt, maxdepth, minsampleleaf and minsamplesplit. Train data works best for gini, random, log2. 1, 10, and 6 being leaf, depth and split, and check the results on test data.	Using GridSearchCV, check for algorithm with auto, balltree, kdtree and brute, neigbors with 1 to 10, p with 1 and 2. It works best for algorithm with auto, neighbors 10 and p as 1 for train data. Using this to predict test data and check accuracy and Confustion Matrix.

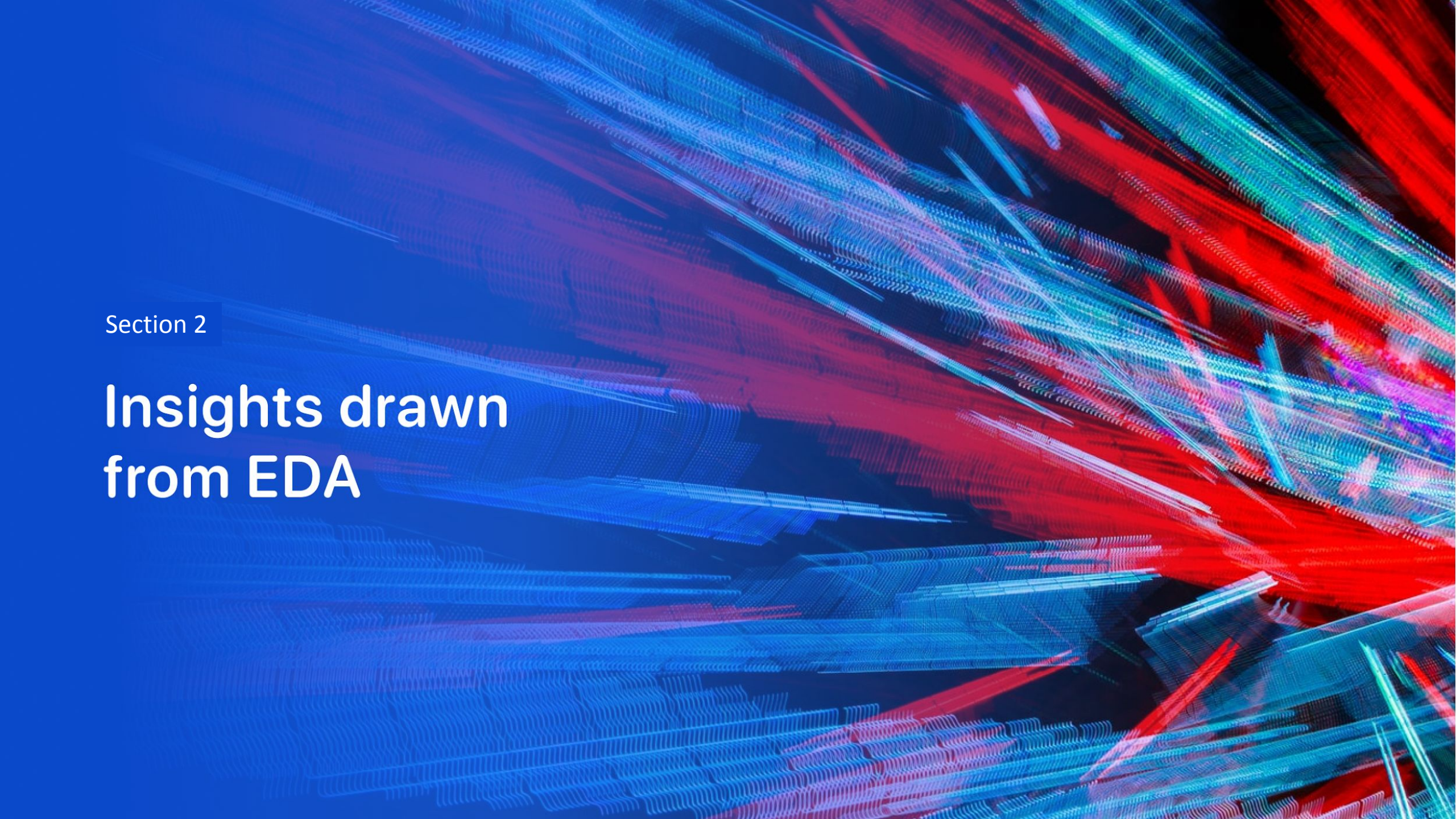
[GitHub URL]

Results

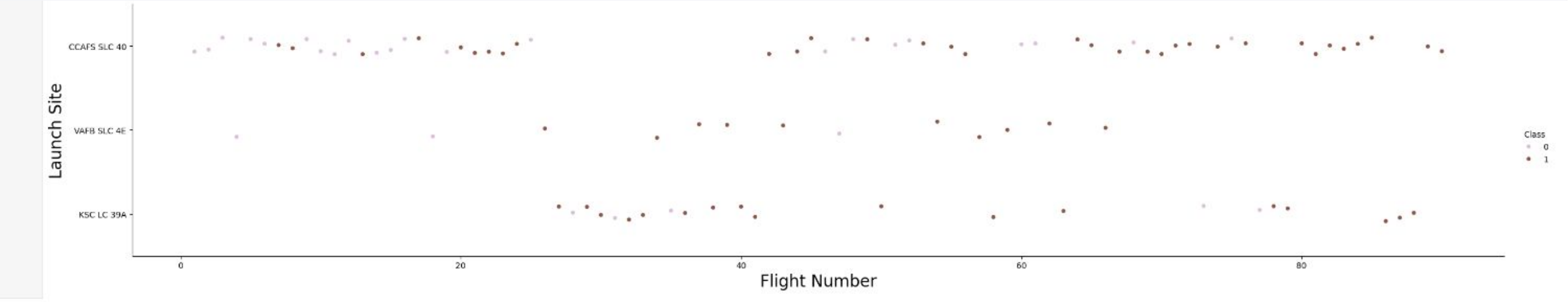
- Exploratory data analysis, emphasize the importance of flight number, orbits and payload mass.
- The interactive app helps us in understanding the launch results with parameters that can be varying.
- For the given set of parameters, the models seem to be working equally well.

Section 2

Insights drawn from EDA

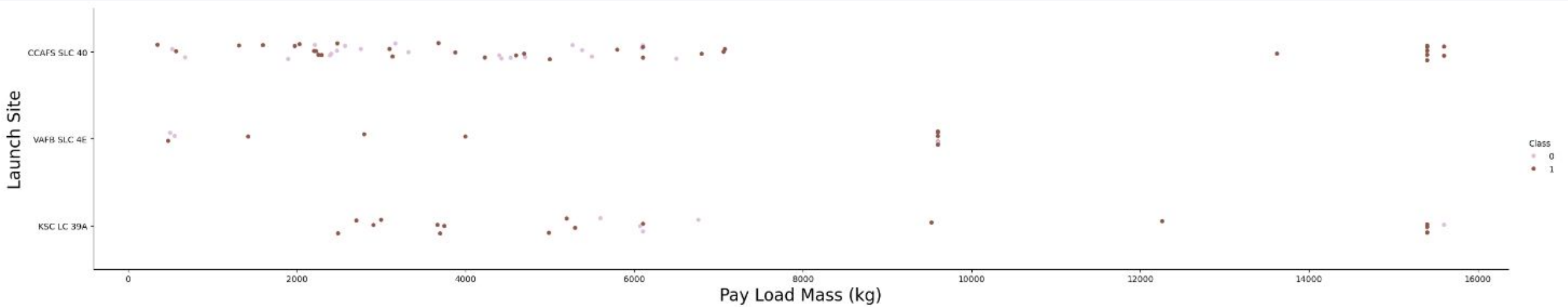


Flight Number vs. Launch Site



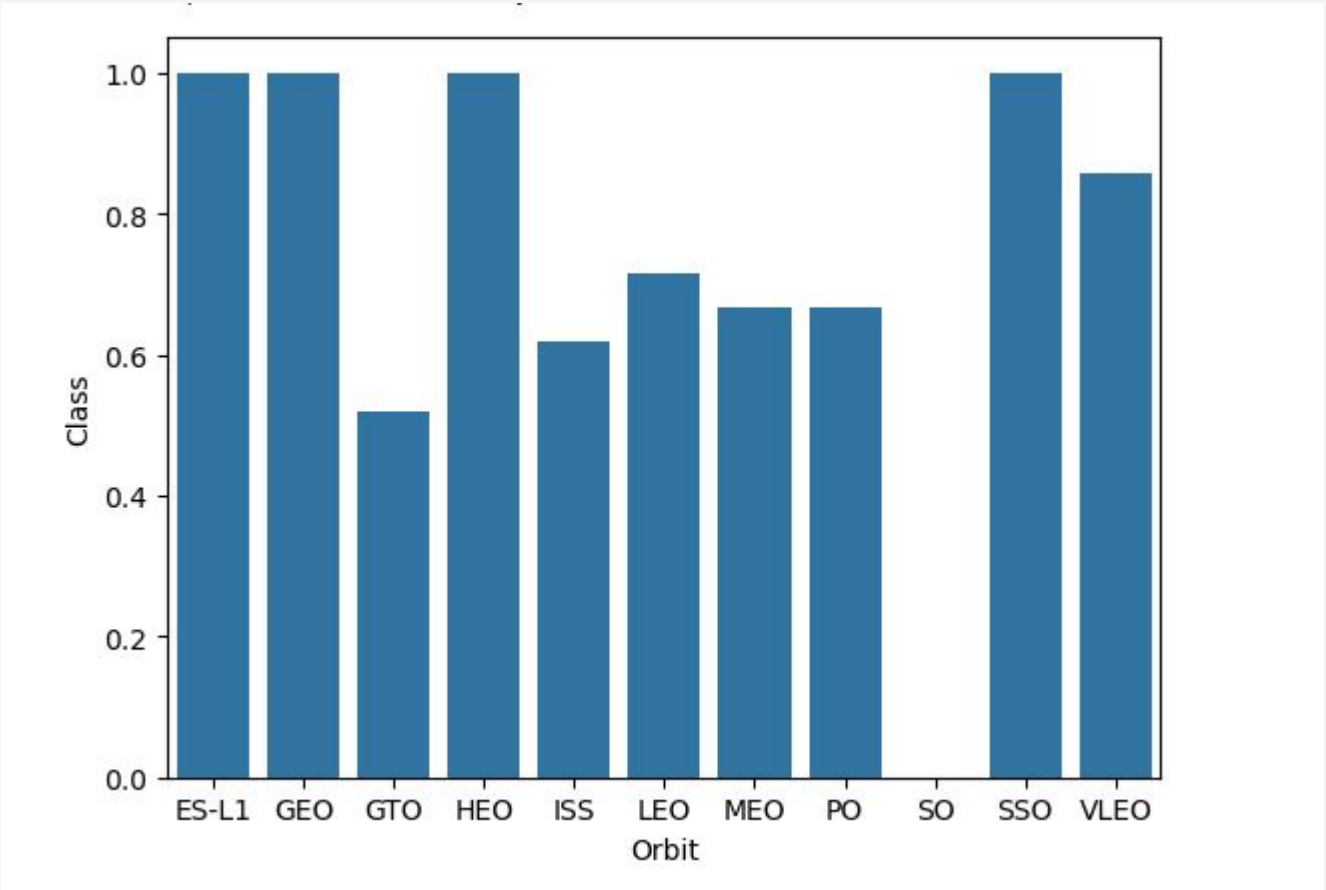
This cat plot between flight number and launch site has class as hue. We can see that initial flight numbers were failures. Most of them were done in CCAFS SLC 40 and some in VAFB SLC 40. KSC LC 39 A have their first flight numbers later on and hence starts off with success.

Payload vs. Launch Site



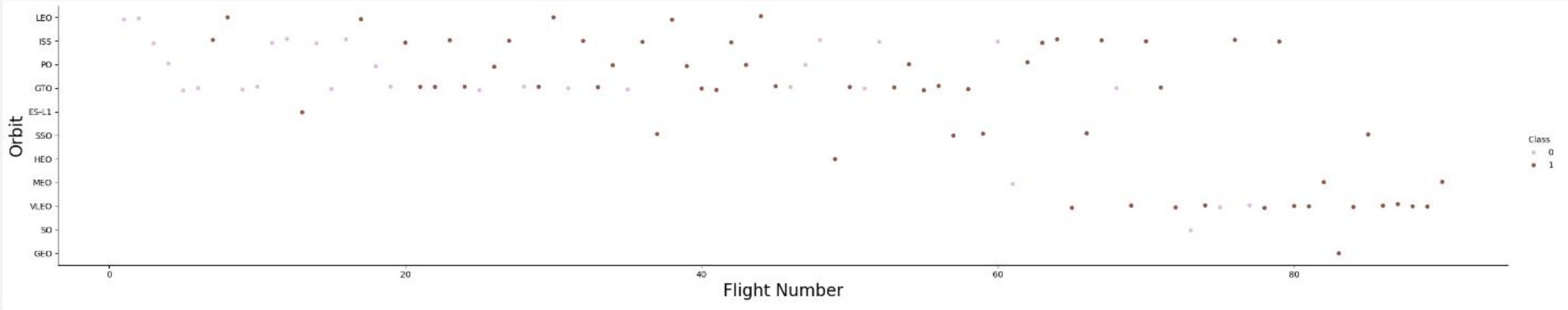
This cat plot between payload mass and launch site has class as hue. We can see that for VAFB SLC 4E there have been no launches with payload mass greater than 10000kg. For CCAFS SLC 40 and KSC LC 39 A there have been some, which were mostly successful. Unsuccessful launches have payloads in the low to mid range usually.

Success Rate vs. Orbit Type



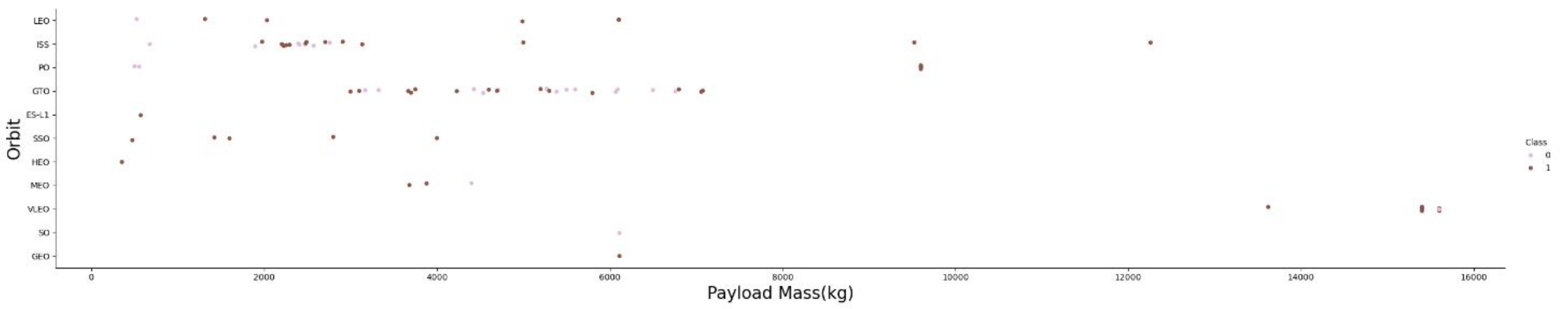
This bar chart is between Orbit and Class mean for each Orbit. Orbit type SO has no successful launches. ES L1, GEO, HEO and SSO have all launches as successful. GTO has a high failure/success ratio.

Flight Number vs. Orbit Type



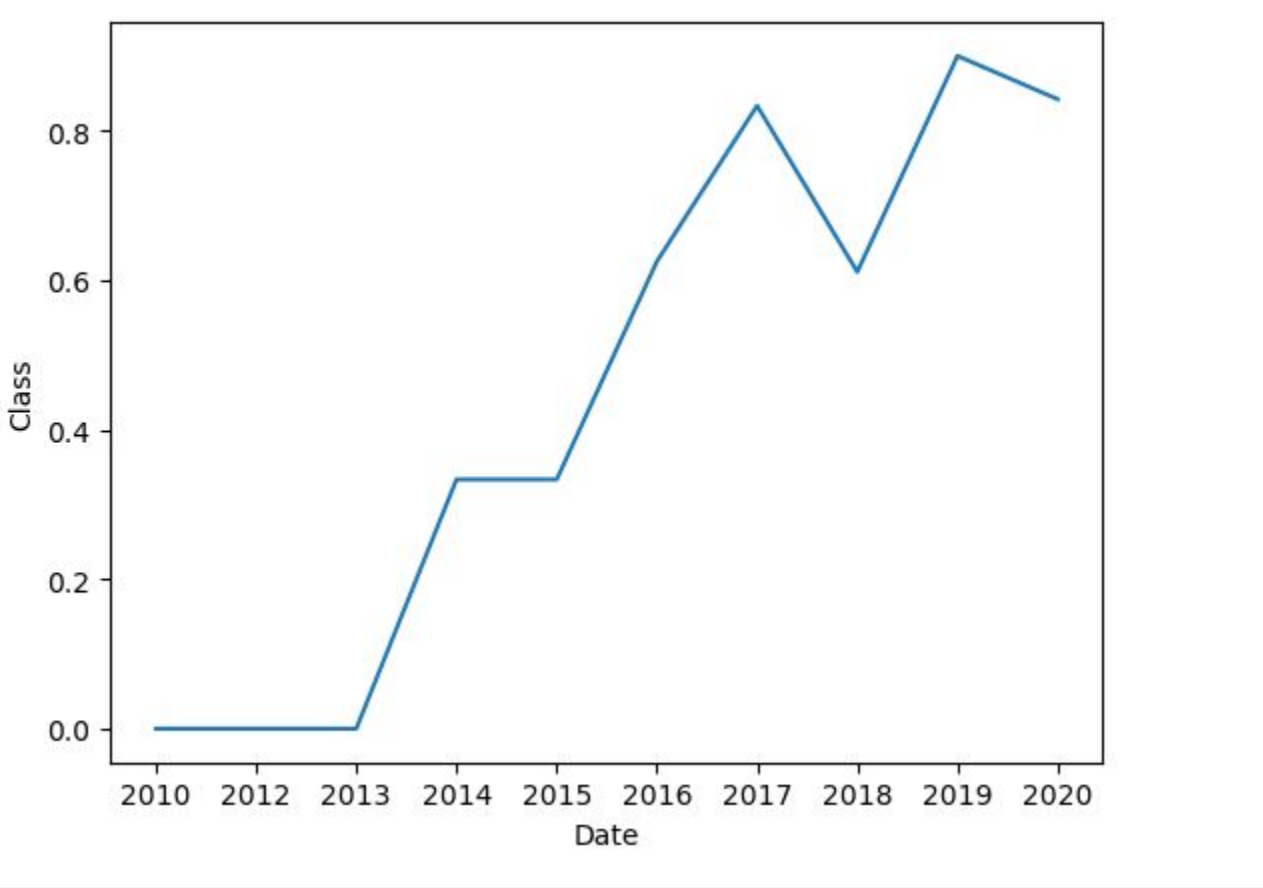
For GTO successes and failures appear even later on with flight numbers. For VLEO it comes up only on later launches, it has mostly successes. All Orbits with all successes have relatively less number of launches.

Payload vs. Orbit Type



The high payload launches seem to be mostly in VLEO orbit, and most of them seem to be successful. GTO has mid range launches with success and failures mixed. ES L1, SSO, and HEO have launches upto 4000 all successful and GEO has one around 6000, which is successful.

Launch Success Yearly Trend



Initial success seem to have come about 2014, and the relative ratio of success/failure has gradually increased, barring a dip in 2018 from 0.8 to 0.6.

All Launch Site Names

```
%sql select Launch_Site from SPACEXTBL group by Launch_Site
```

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

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

The launch sites were grouped to present distinct launch sites, which reflects 4 distinct launch sites as above.

Launch Site Names 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
2010-06-04	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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

The results are the top 5 results of launch sites with ‘%CCA%’ in it. As seen above, it is only CCAFS LC 40 launch site in these results. The mission outcomes were successful and landing outcomes were not. Booster versions have been of F9 v1.0 variety.

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

The above reflects the total payload mass with NASA(CRS) as customer. As the sum suggests multiple launches have been done by them.

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1'
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

The average payload mass by F9 V1.1 is tending towards the lower- middle range as the number suggests.

First Successful Ground Landing Date

```
%sql select min(date) from SPACEXTBL where Landing_Outcome like '%Success (ground pad)%'
```

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

min(date)
2015-12-22

The first successful landing outcome with success in ground pad is about 5 years after the initial launches.

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_ > 4000 and Payload_Mass__KG_ < 6000
```

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

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

The boosters for success in drone ship landing outcome and between payload range 4000 and 6000 are all of F9 FT variety.

Total Number of Successful and Failure Mission Outcomes

```
%sql select Mission_Outcome, Count(*) from SPACEXTBL group by Mission_Outcome
```

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

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

There has only been one failure, which was in flight for mission outcomes. There is one success with payload status unclear.

Boosters Carried Maximum Payload

```
%sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_ in (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

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

Boosters carrying maximum payload have all been of F9 B5 variety.

2015 Launch Records

```
%sql select substr(Date, 6, 2) as month , Landing_Outcome, Booster_Version, Launch_Site From SPACEXTBL where Landing_Outcome <> 'Success (drone ship)' \
and Landing_Outcome like '%drone ship%' and substr(Date,0,5) = '2015'
```

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

month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40

Failure drone ship for 2015 have all been from launch site CCAFS LC-40 and have been of F9 v1.1 variety.

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

```
%sql select Landing_Outcome, count(*) from SPACEXTBL where Date > '2010-06-04' and Date < '2017-03-20' group by Landing_Outcome order by count(*) DESC
```

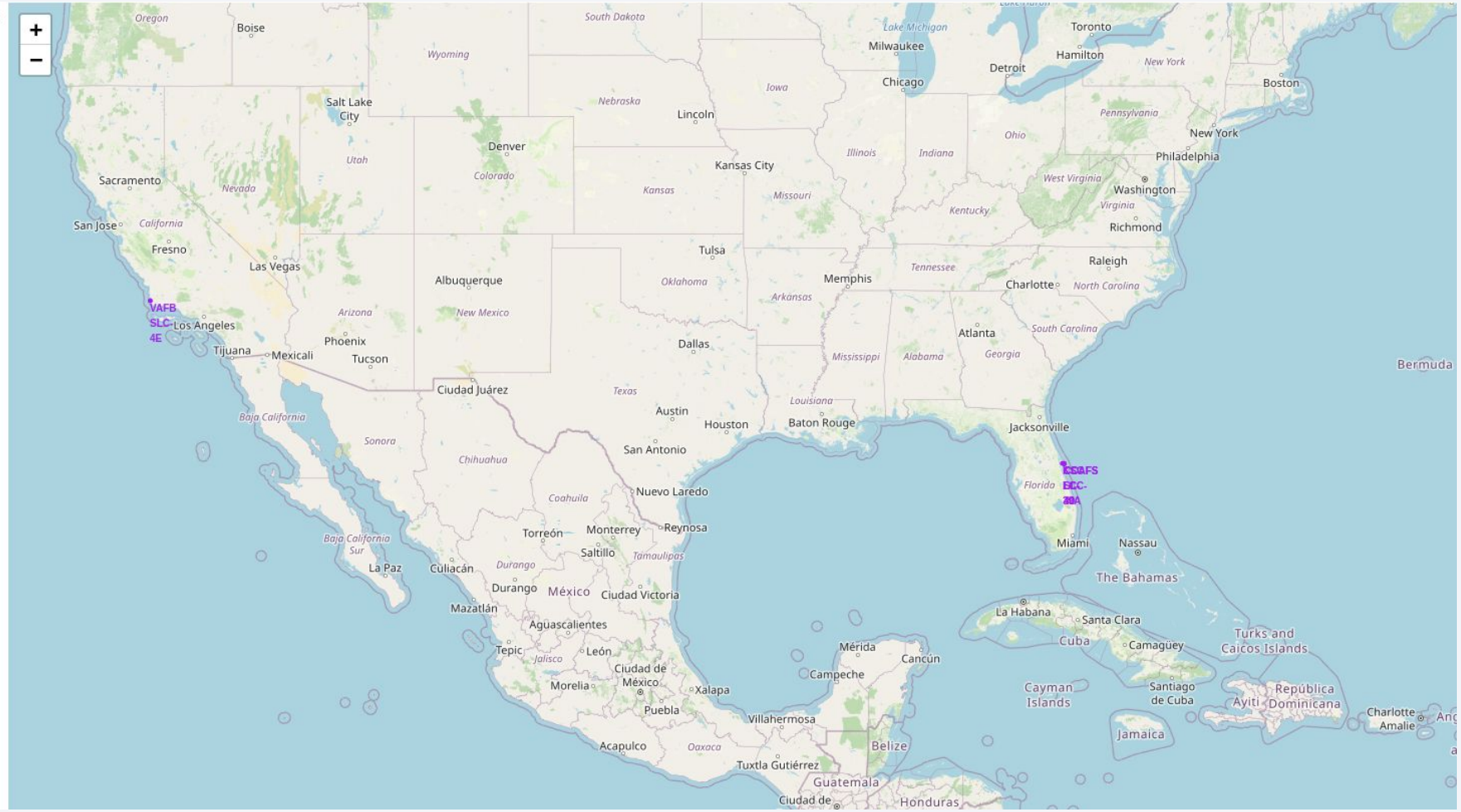
```
 * sqlite:///my_data1.db
Done.
```

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

The maximum outcome for the period is in which no attempt was made to land. Drone ship has equal number of success and failure landings. All ground pad landings were successes.

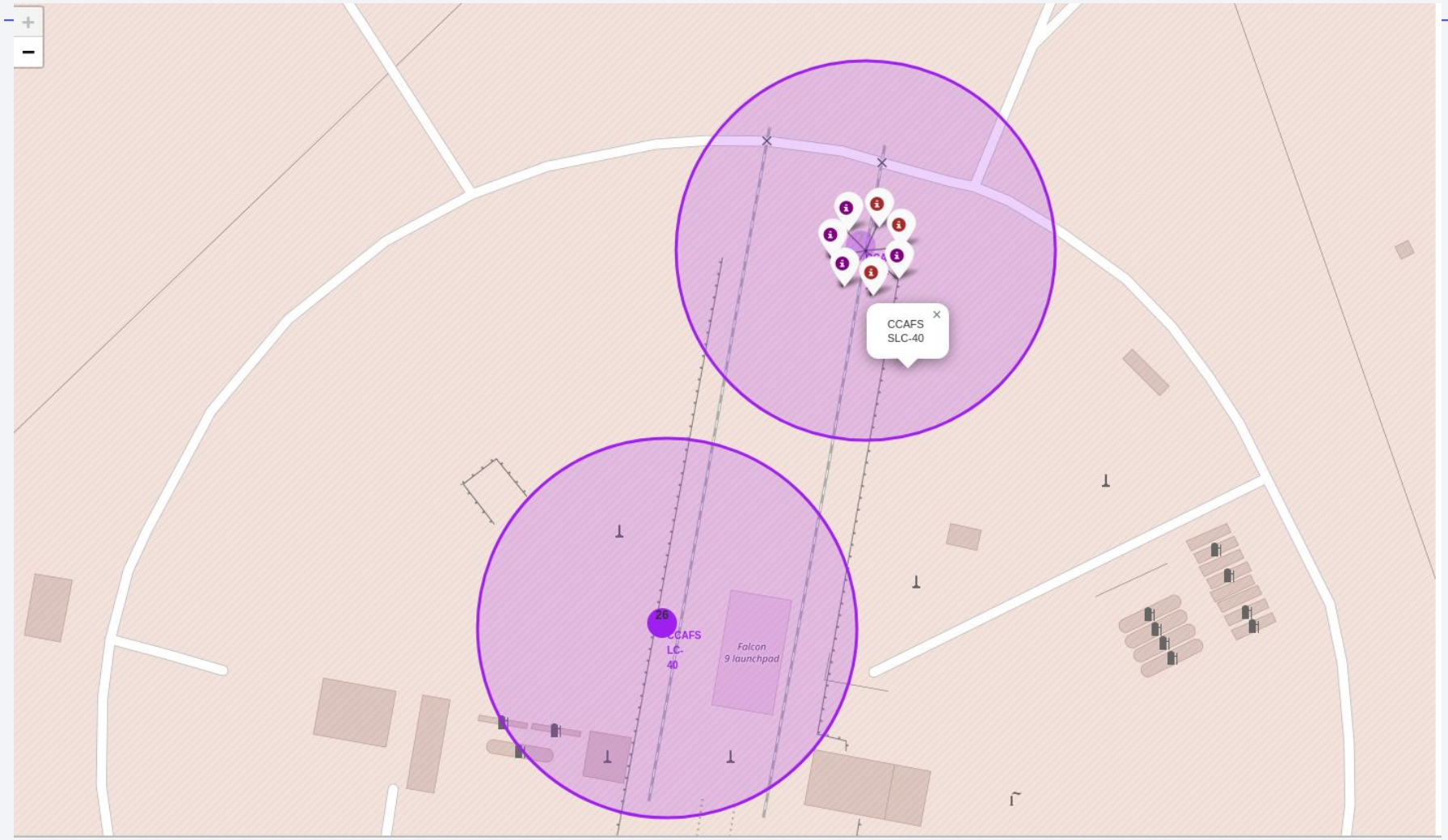

Section 3
Launch Sites
Proximities Analysis

Launch Site Locations



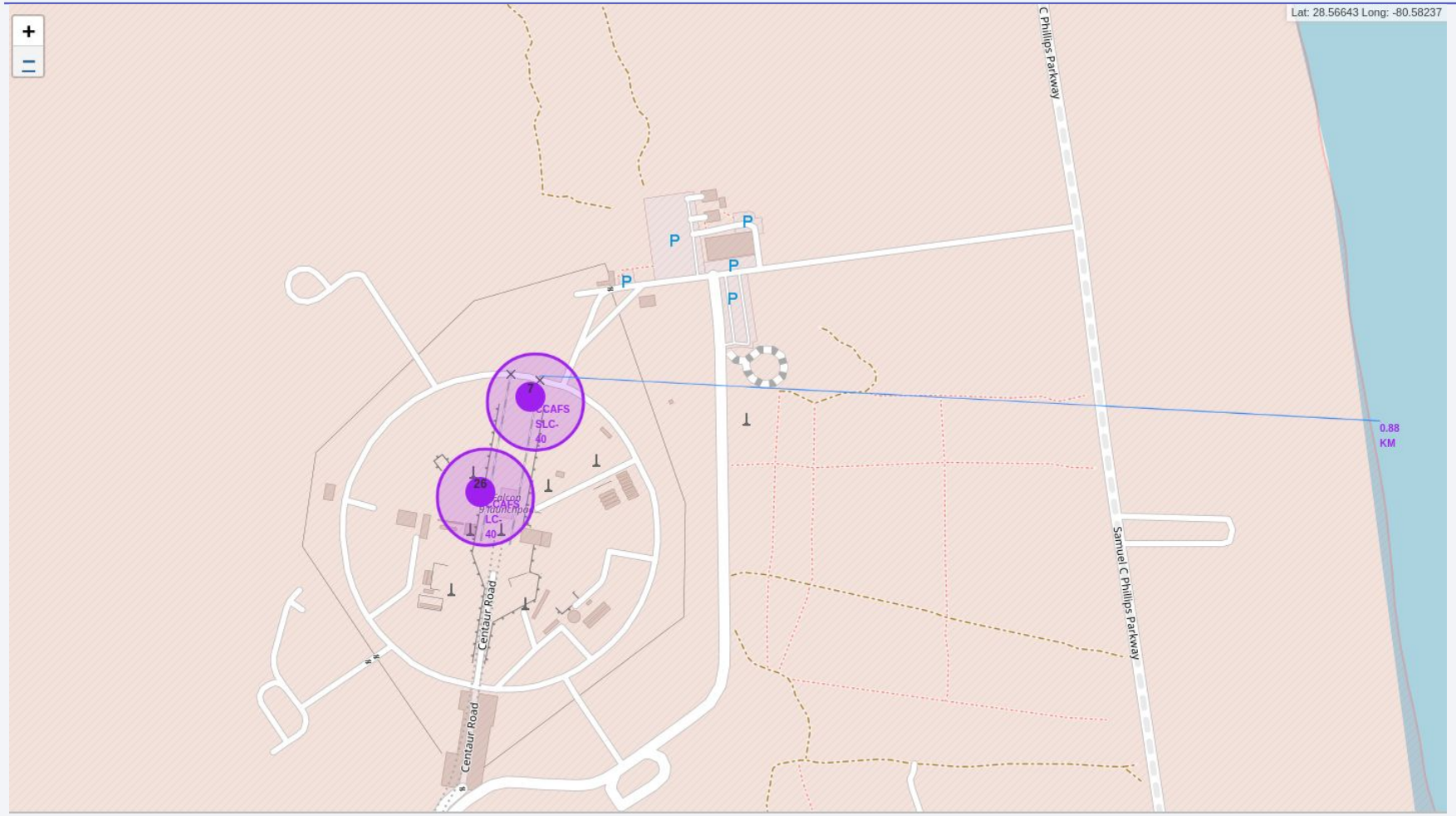
Launch sites are located on east and west side. Of the 4, 3 on east.

Launch Outcomes

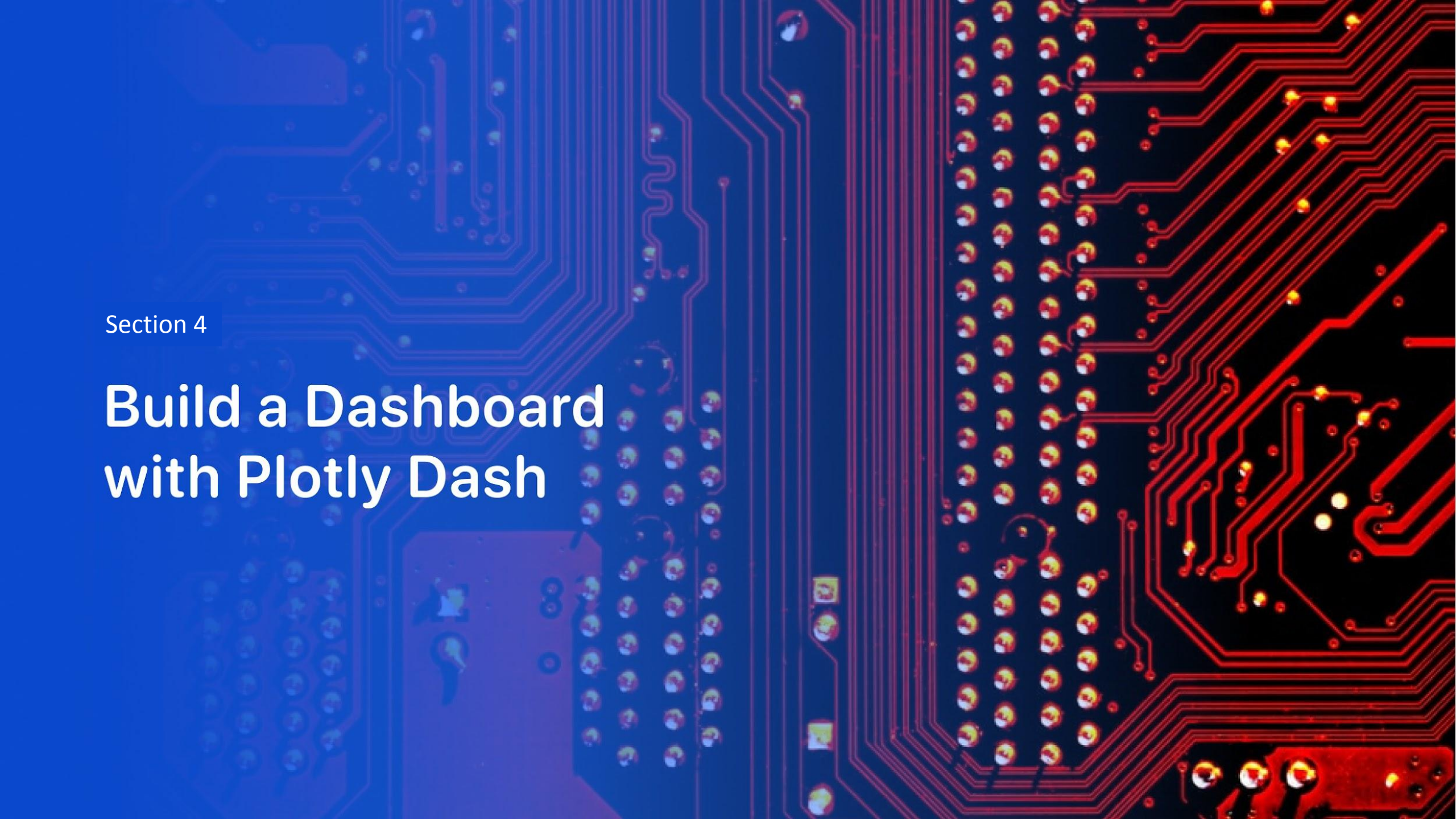


CCAFS SLC 40 and CCAFS LC 40 launch sites are nearby. For CCAFS SLC 40 there are 4 failure outcomes and 3 successes. There are 26 launches for CCAFS LC 40.

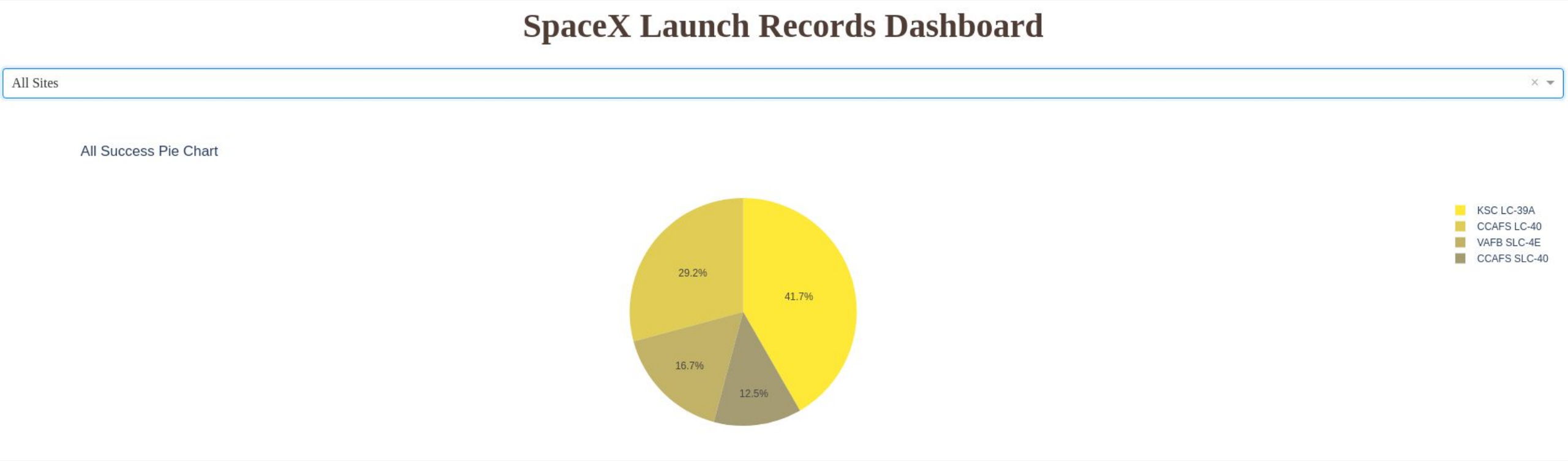
Launch Site Proximity



CCAFS SLC 40 is 0.88 km from coastline.

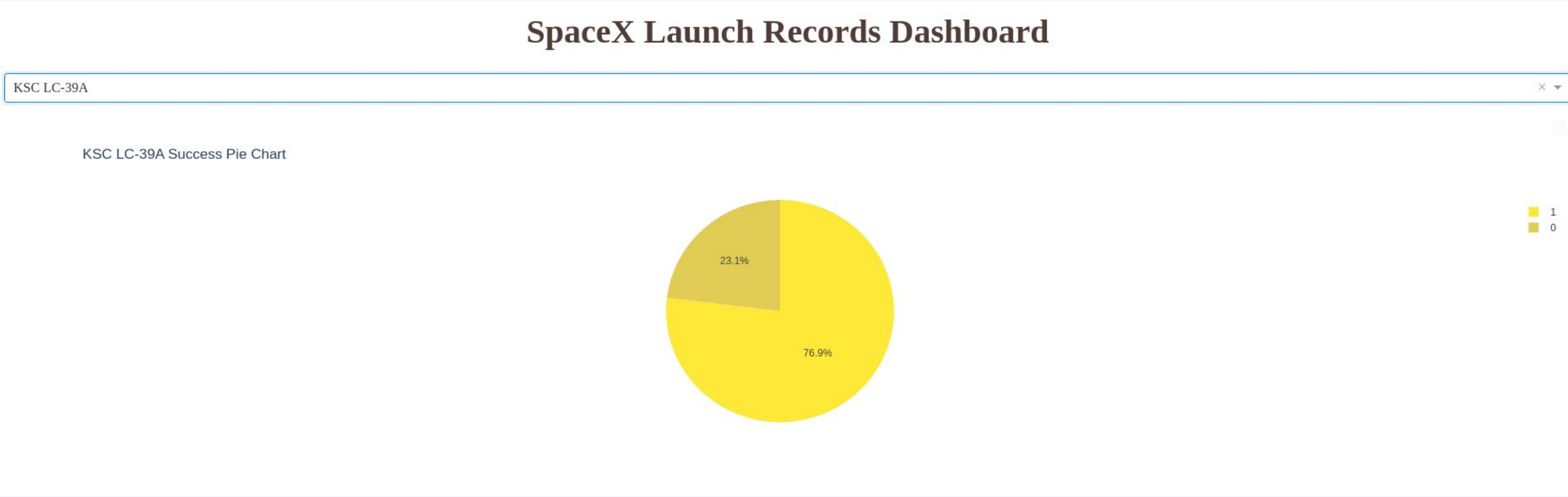

Section 4
Build a Dashboard with Plotly Dash

All Sites Success Pie Chart



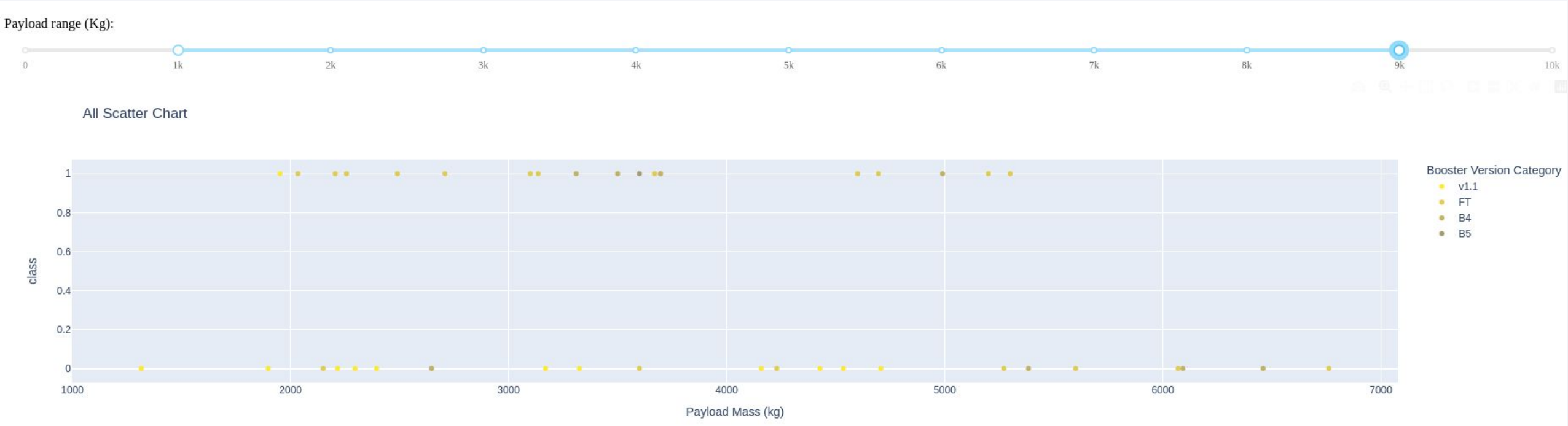
We see that KSC LC - 39 A has the highest amount of successes. And CCAFS SLC-40 has the least.

Highest Success Ratio Pie Chart



We see that KSC LC - 39 A has the highest success ratio as well. We see that inspite of having relatively lower number of launches, it has contributed to highest number of successes.

Selected Payload Range Scatter Chart

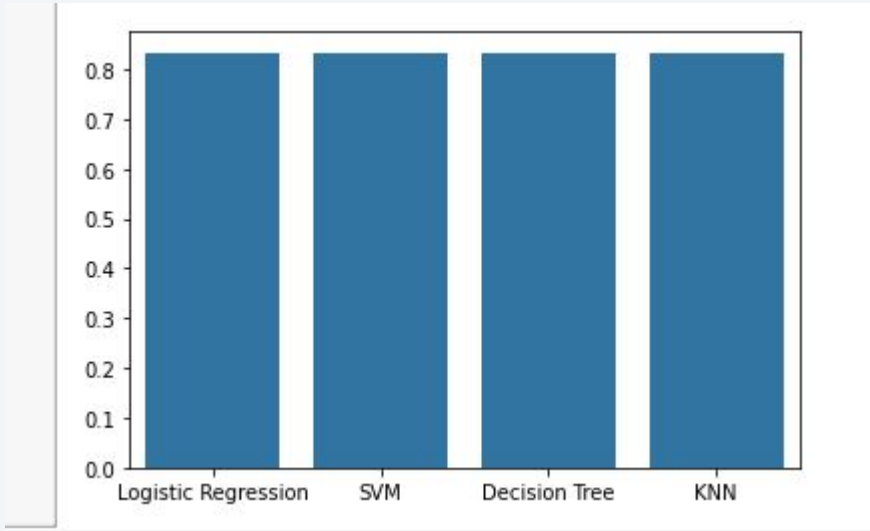


We see that V1.1 has lot of failures and few successes. Same with B4. B5 has single success. And FT has more successes than failures.

Section 5

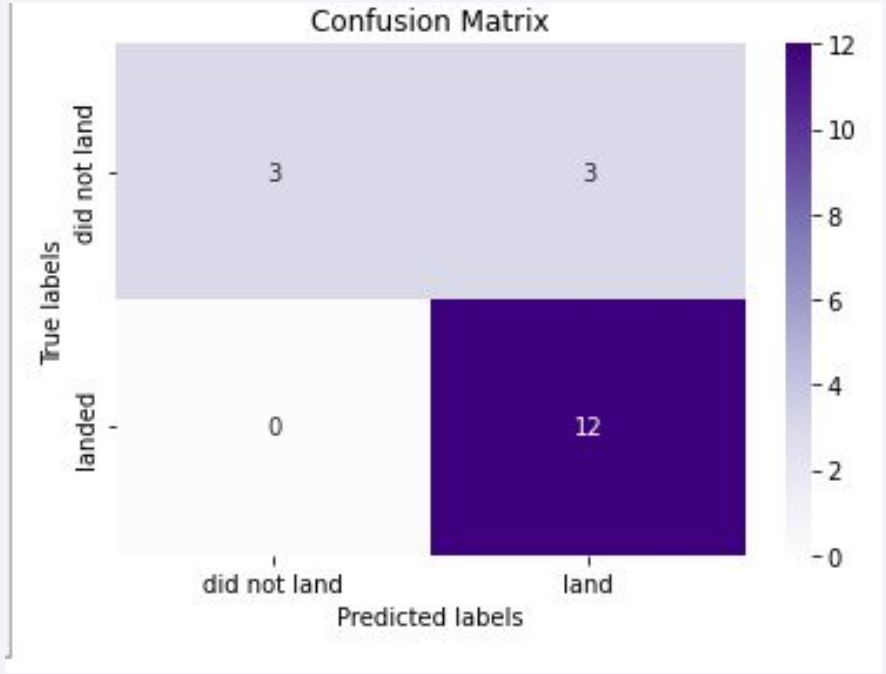
Predictive Analysis (Classification)

Classification Accuracy



The models are having the same level of accuracy of around 0.83.

Confusion Matrix



The results for the Confusion Matrix seem to be the same for all models.

Conclusions

- Launch successes have increased, as flight number increases, which may be due to learnings in earlier launches.
- High payloads (above 10000kg) have high success/failure ratio
- ES L1, HEO, GEO and SSO orbits have all successes. The number of launches in them are not high, and payloads are below 6000 kg.
- The launch sites tend to be near coastline and modes of transportation and far away from city.
- Booster version FT has more success/failure ratio, in the 1000-9000 kg payload mass range, with some number of launches.
- B5 booster version is used for high payloads.

Appendix

- The repository here has the data that was created as part of the assignment, data_set_part1.csv, data_set_part2.csv, data_set_part3.csv and mydb1.db

Thank you!