Winning Space Race with Data Science

Jojo Tan
30 Oct 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

If we can predict whether the first stage of SpaceX's launches will land successfully, we can better determine whether it will reuse the first stage and subsequently the price of each launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Data had been collected for this purpose through REST calls and web-scraping and after which wrangled before exploratory data analysis (visualisation and SQL) could be performed on it. Interactive analyses such as Folium and Plotly Dash were also executed alongside predictive analyses done using classification models.

Introduction

Background

SpaceX 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 SpaceX 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 SpaceX for a rocket launch.

Section 1

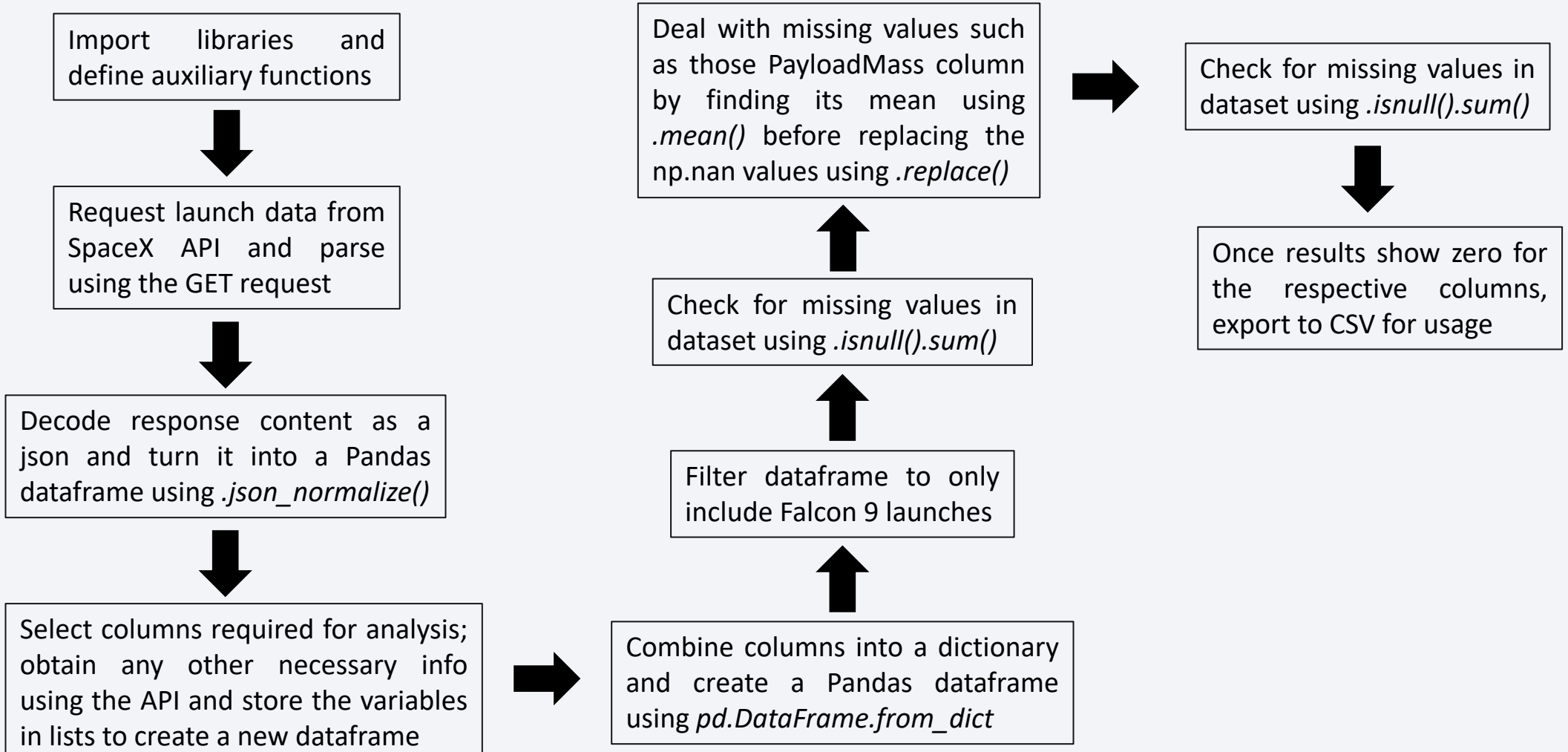
Methodology

Methodology

Executive Summary

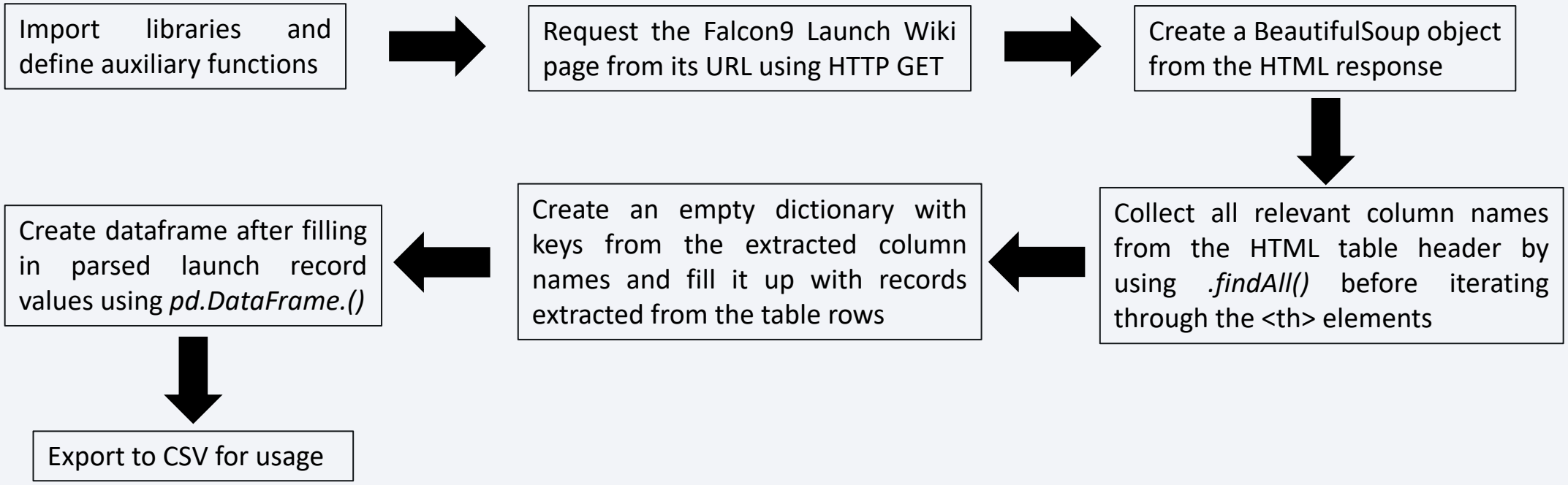
- Data collection methodology:
  - Done via SpaceX REST calls and webscraping using BeautifulSoup
- Perform data wrangling
  - Explored and cleaned data through filling in missing values and creating necessary columns derived from existing ones for further analysis
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build models using best parameter settings and train them on training sets before administrating them on testing sets; compare accuracy scores for respective models to determine best one to be used

Data Collection – SpaceX API



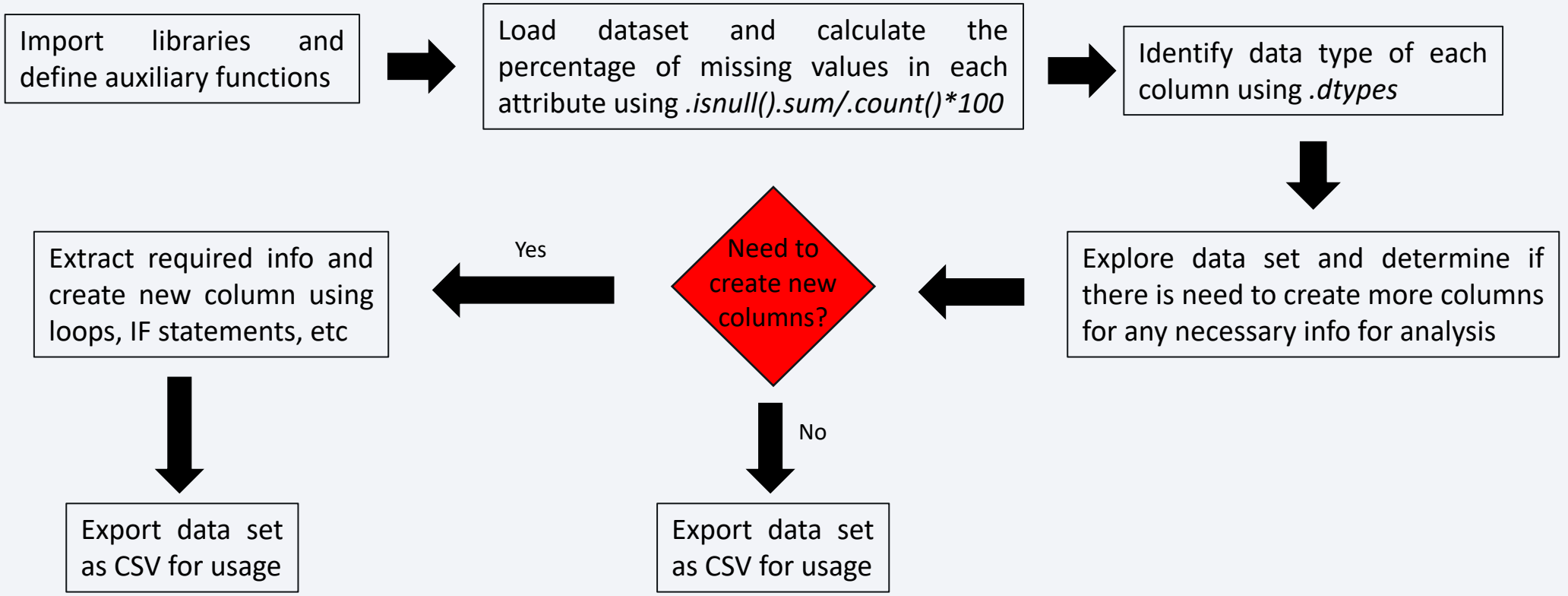
SpaceX API Calls Notebook

Data Collection - Scraping



[Completed Web Scraping Notebook]

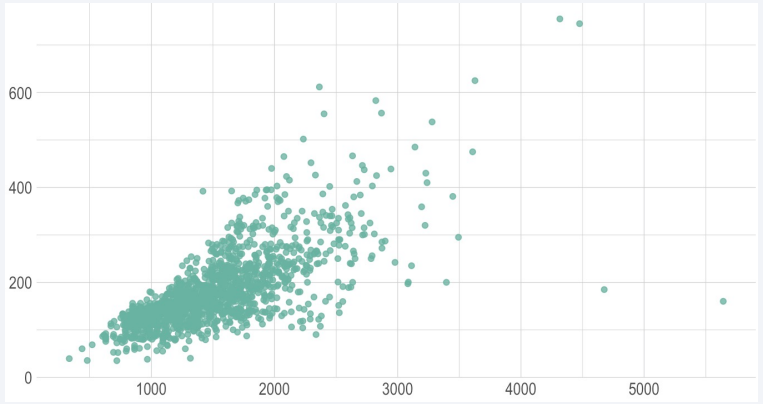
Data Wrangling

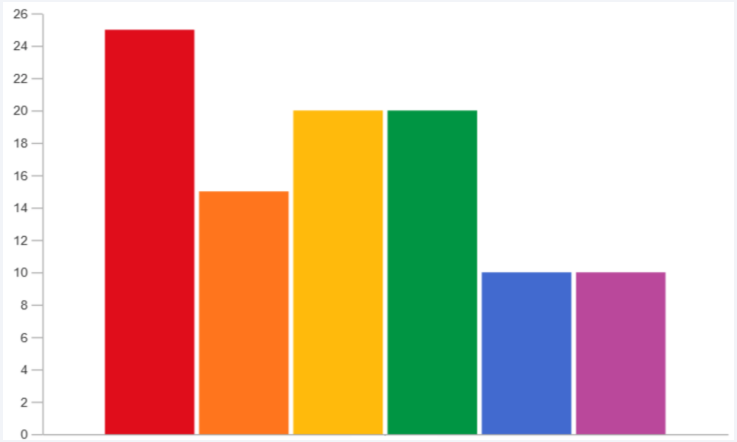


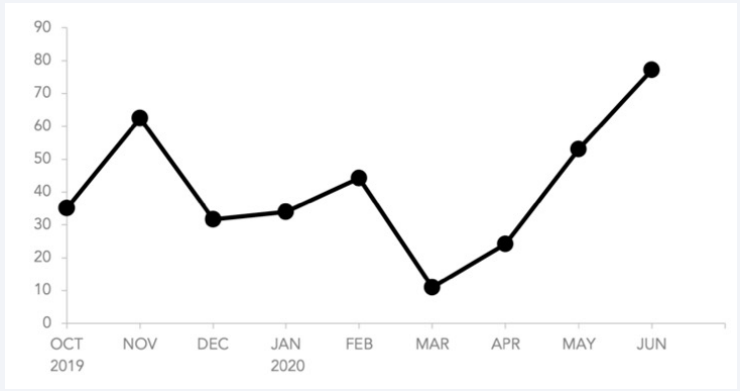
[Completed Data Wrangling Notebook]

EDA with Data Visualization

- Charts plotted: scatter plot, bar and line graphs







- Graphs used were selected based on their efficiency and ease of usage in interpreting the data selected and the trends formed (e.g. line graph to view pattern over time)

EDA Data Visualisation Notebook

EDA with SQL

Queries performed:

- Finding the various launch sites' names
- The total payload mass carried by a particular customer
- The average payload mass carried by a particular booster
- Finding the first successful ground landing date
- The total number of successful and failed outcomes
- The type of booster(s) that carried the max payload
- Filtering through the data for launch records in a particular year
- The count of the different outcomes between a time period

EDA with SQL Notebook

Build an Interactive Map with Folium

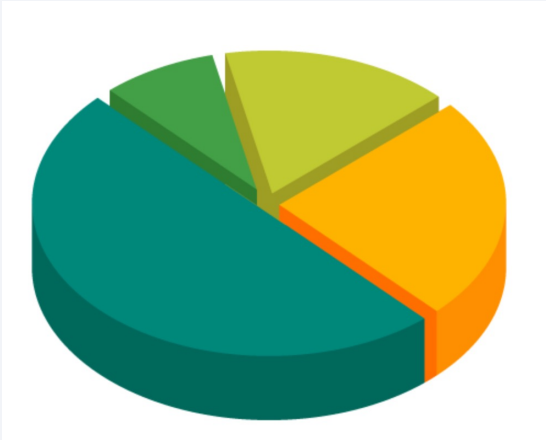
- Created and added markers and circles to a folium map to pin point the locations of the various launch sites
- Then added marker clusters to the launch sites to indicate the number of successful and failed launches they each had upon zooming into each one of them
- Calculated and added polylines to show the distance between launch site and its proximities such as railway, highway, coastline and city

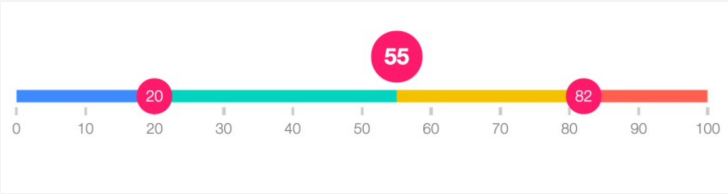
Completed Interactive Map

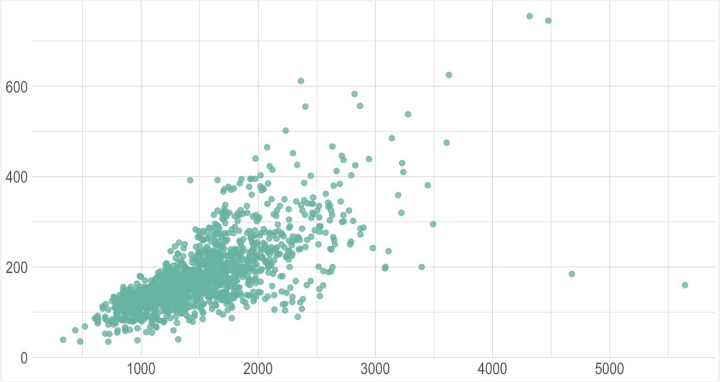
Pls view it using https://nbviewer.org/ instead as github is unable to display the folium maps, thanks!

Build a Dashboard with Plotly Dash

- Items added to dashboard: pie chart, slider, scatter plot chart





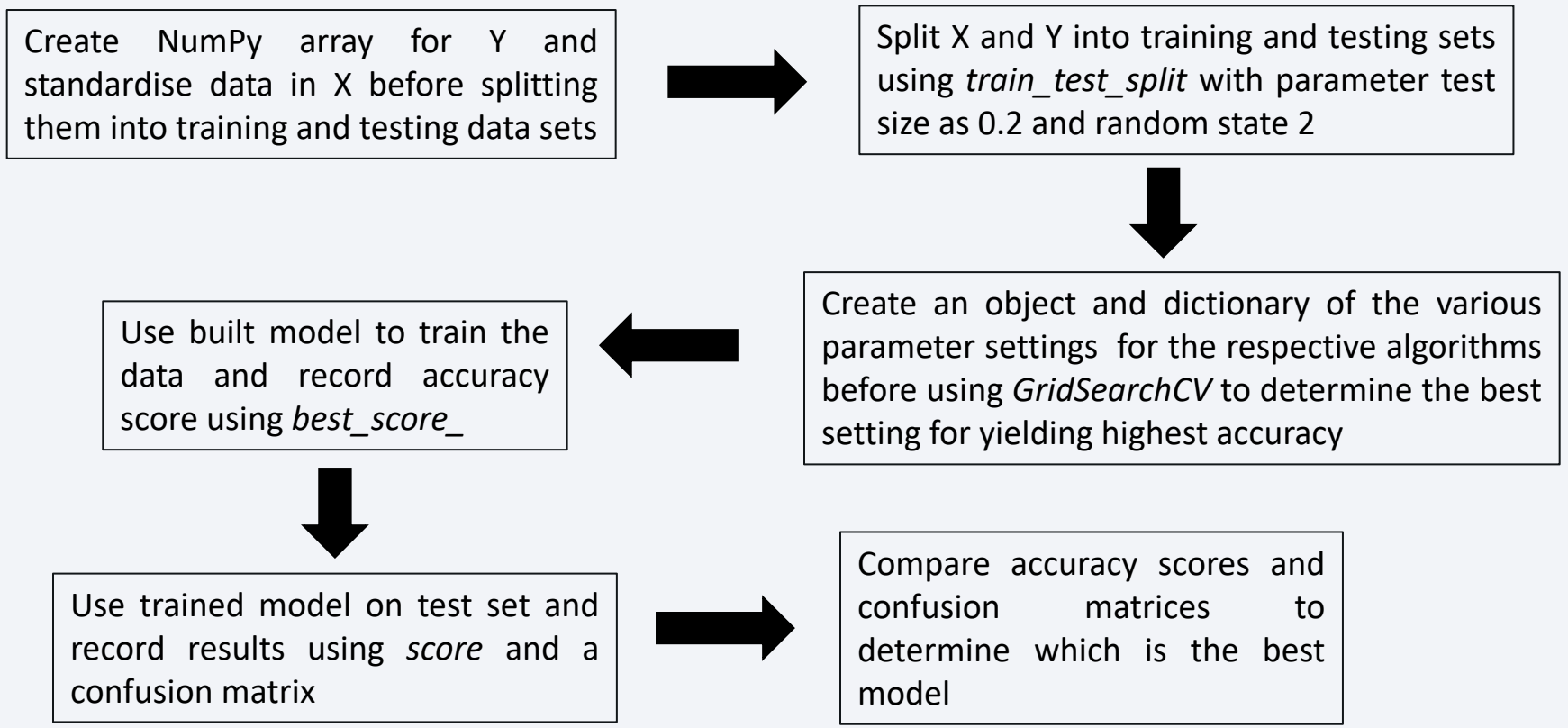


- Items used were selected based on their efficiency and ease of usage to interpret the data selected (e.g. slider to conveniently select the criteria user is interested in knowing about)

Plotly Dash Lab

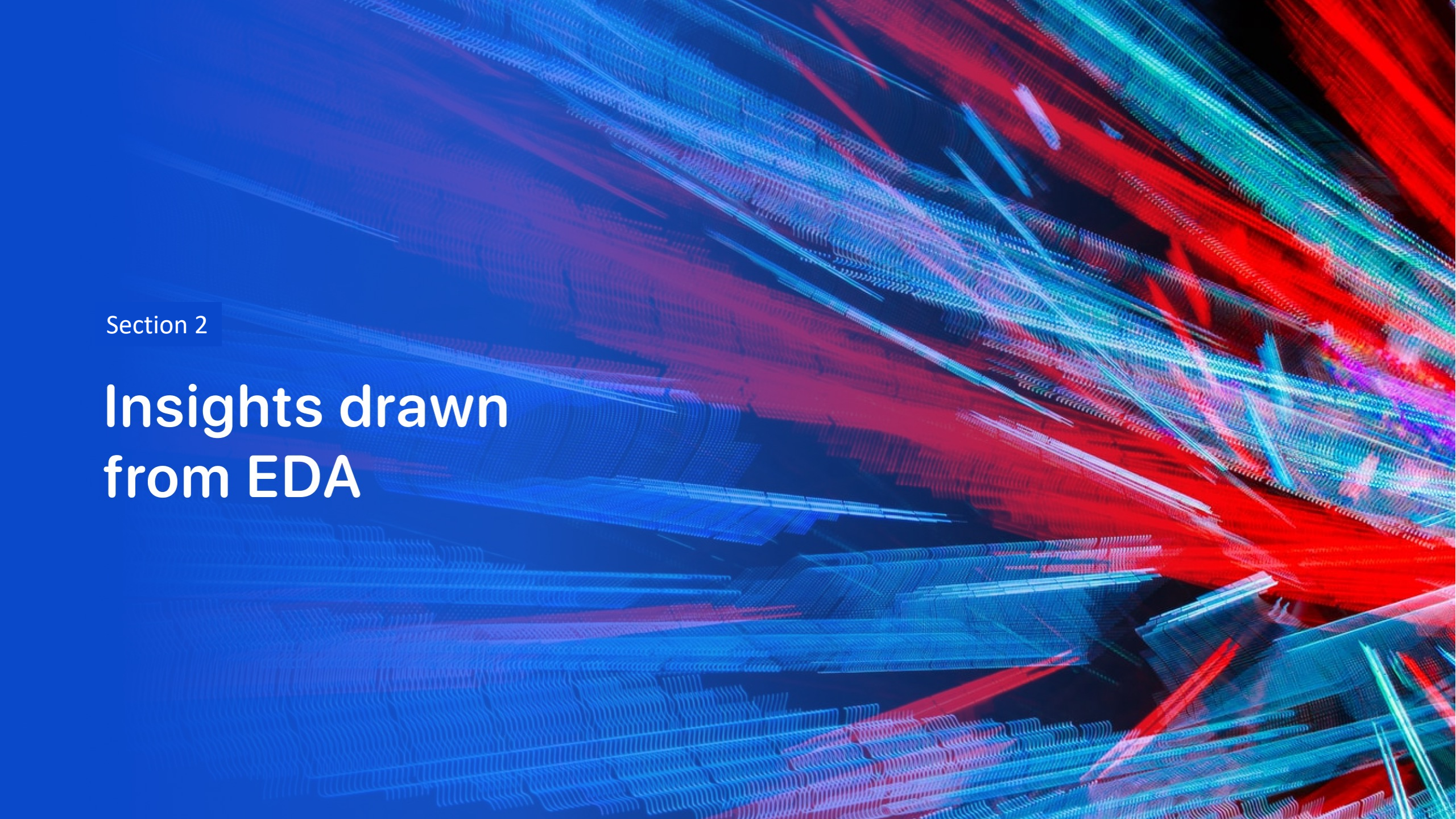
Predictive Analysis (Classification)

- Models tested: Logistic Regression, Support Vector Machine, Decision Tree and K Nearest Neighbours

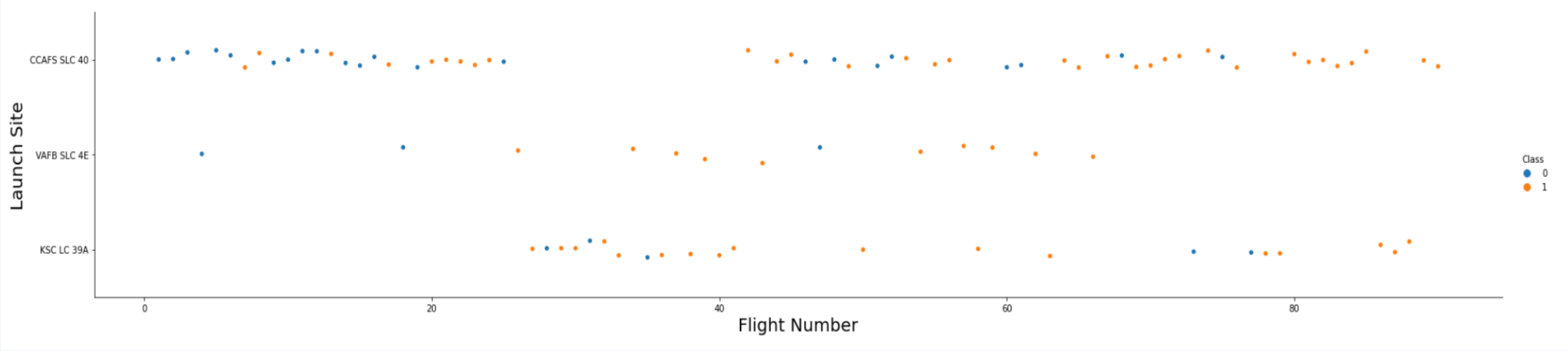


Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

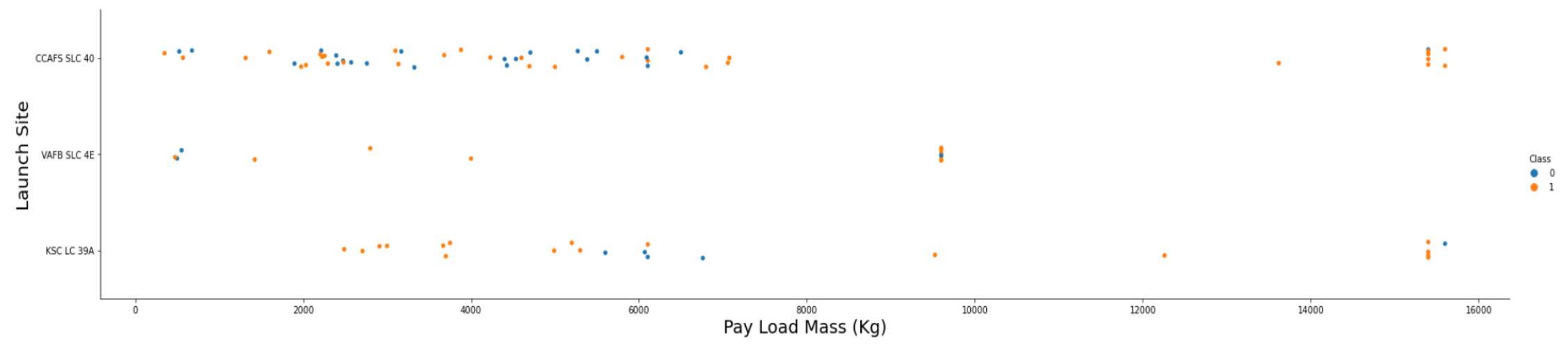


- Majority of the flights were done at CCFAS SLC-40 and the least at VAFB SLC-4E

- Different success rates for different launch sites:
  CCFAS SLC-40: 60%
  VAFB SLC-4E: 77%
  KSC LC-39A: 77%

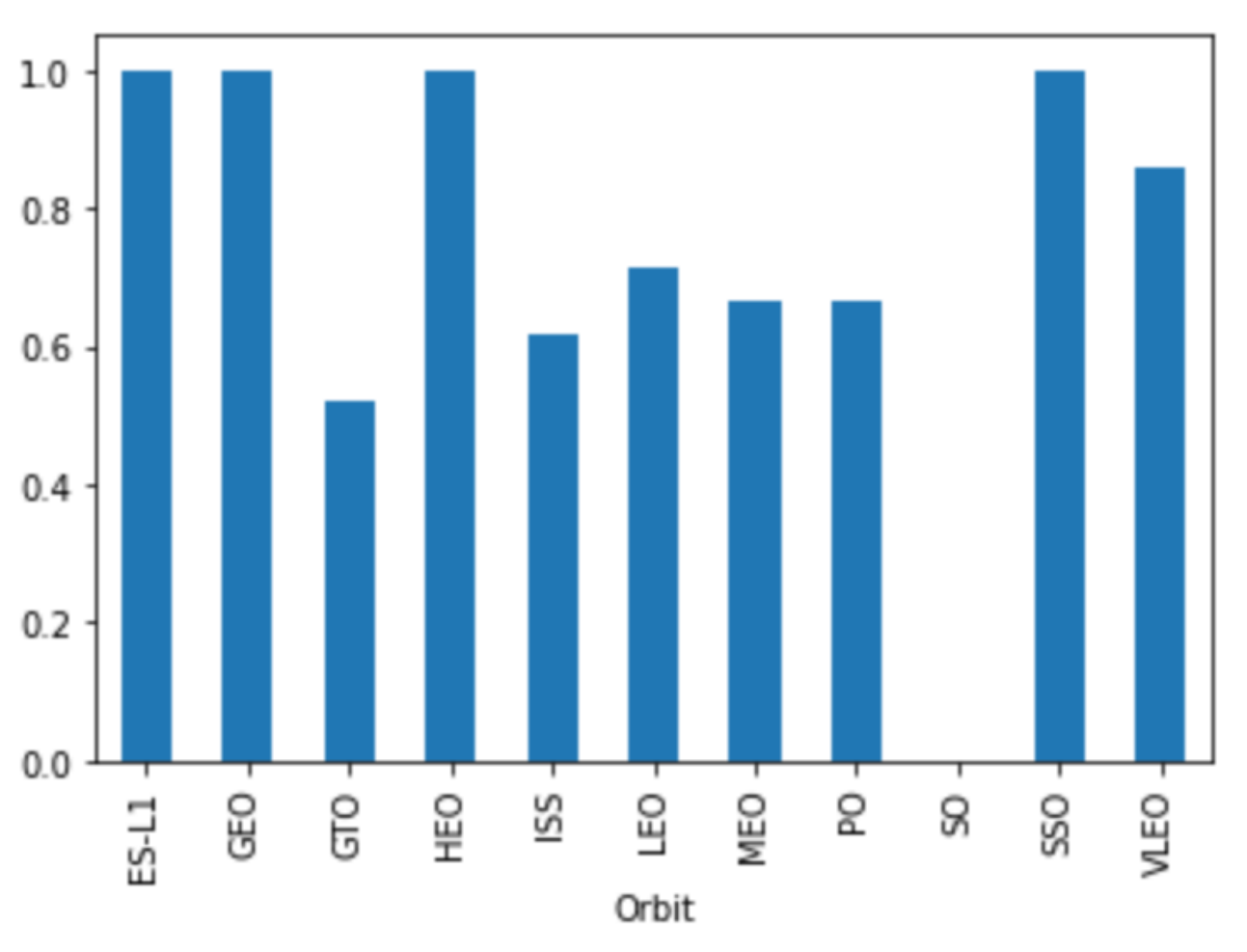
- Increasing success over time with the increase in flight number

Payload vs. Launch Site



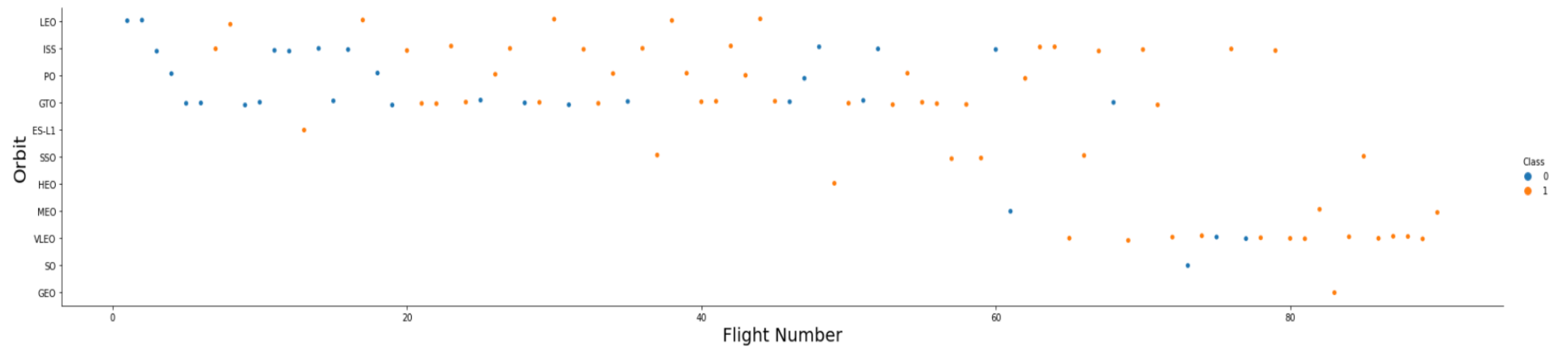
- Most rockets launched are under 8,000kg
- No rockets greater than 10,000kg launched for VAFB SLC-4E
- Lower failure rate for launched rockets greater than 10,000kg

Success Rate vs. Orbit Type



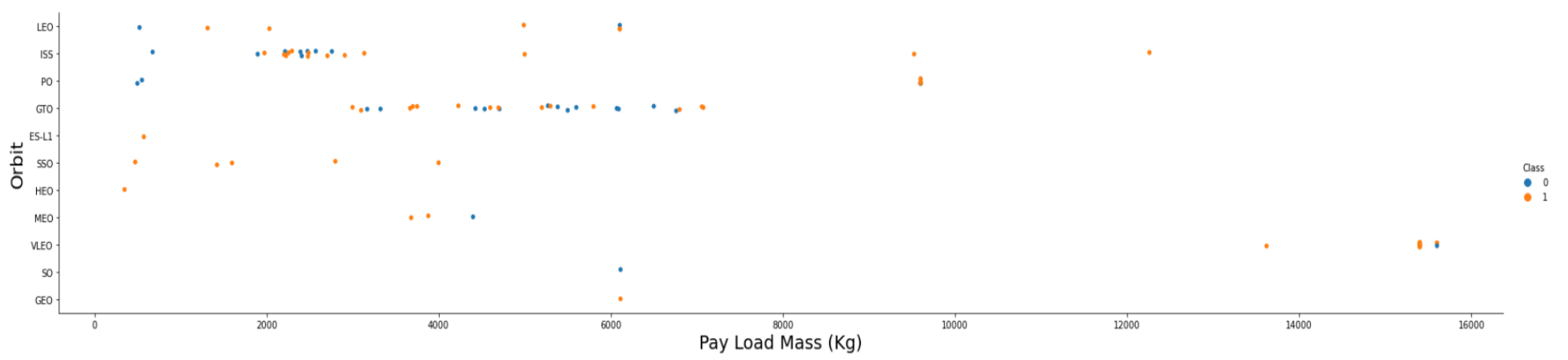
- High success rates for ES-L1, GEO, HEO and SSO orbits
- Lowest for SO orbits

Flight Number vs. Orbit Type



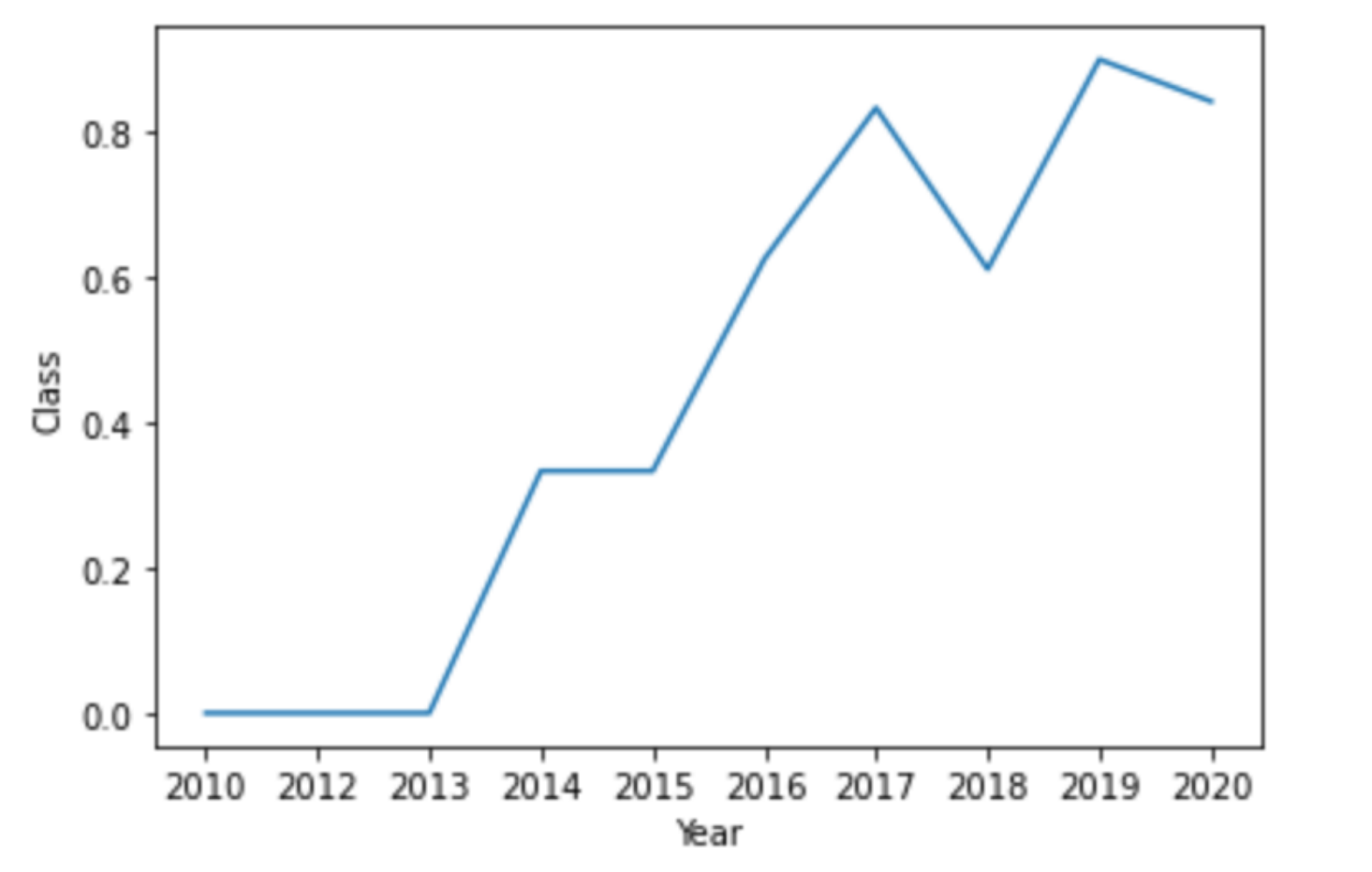
- Most number of flights taken in GTO orbit with no apparent trend of success / failure rate over time
- In comparison, flights taken in LEO orbit become successful over time / more flights
- Orbits ES-L1, HEO and GEO have only taken one flight each, which all turned out successful – putting into question whether their individual success rate is reliable since only one flight has been taken

Payload vs. Orbit Type



- Generally positive landing rates for most orbits with payloads at 4,000kg and below
- Successful landing rates are higher for orbits LEO, ISS and PO with heavier payloads
- This however does not seem to be the case for GTO which has a mix of success and failure landings even with similar payloads

Launch Success Yearly Trend



A general upward trend from 2013 – sans for a plateau between 2014 and 2015 – up till 2017 where success rate began to dip before picking up again in 2018 and peaked in 2019 before dipping again

All Launch Site Names

- Four launch sites in total:
  1. CCAFS LC-40
  2. VAFB SLC-4E
  3. KSC LC-39A
  4. CCAFS SLC-40

- Query drawn by selecting distinct values of Launch_Site column of database

```
1 %sql SELECT DISTINCT Launch_Site \
2     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'

```
%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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Query drawn using the 'WHERE' clause with a wildcard ('%') on the Launch_Site column to filter out entries with launch site names beginning with 'CCA' before using the 'LIMIT' clause to obtain the top five records
- Judging from the query result, the table might have been sorted in ascending order either by the Date or Booster_Version column

Total Payload Mass

```
%sql SELECT SUM (PAYLOAD_MASS__KG_) FROM SPACEXTBL \
    WHERE Customer LIKE 'NASA (CRS)';
```

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

SUM (PAYLOAD_MASS__KG_)
45596

- Total payload carried by boosters from NASA: 45,596kg
- Result calculated using the SUM function on the PAYLOAD_MASS__KG_ column and filtering with the WHERE clause on the Customer column

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL \
    WHERE Booster_Version LIKE 'F9 v1.1';
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

- Average payload mass carried by F9 v1.1: 2,928.4kg
- Result derived from using AVG function on the PAYLOAD_MASS__KG_ column and filtering with the WHERE clause on the Booster_Version column

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) AS DATE_OF_FIRST_SUCCESSFUL_LANDING, [Landing _Outcome]\
    FROM SPACEXTBL \
    WHERE "Landing _Outcome" LIKE "Success (ground pad)";
```

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

DATE_OF_FIRST_SUCCESSFUL_LANDING	Landing _Outcome
01-05-2017	Success (ground pad)

- First Successful Ground Landing Date: 1 May 2017
- Result derived from using MIN function on the Date column and filtering with the WHERE clause on the Landing _Outcome column

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_, [Landing _Outcome] \
    FROM SPACEXTBL \
    WHERE "Landing _Outcome" LIKE "Success (drone ship)" AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

Booster_Version	PAYLOAD_MASS__KG_	Landing _Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

- Four successful drone ship landing with payload between 4,000 and 6,000kg
- Lightest at 4,696kg and heaviest at 5,200kg
- Result derived from filtering the Landing _Outcome column using the WHERE clause and the BETWEEN operator on the PAYLOAD_MASS__KG column, setting the lower and upper limits as 4,000 and 6,000

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT COUNT(Mission_Outcome) AS TOTAL_NUMBER_OF_OUTCOMES, Mission_Outcome \
    FROM SPACEXTBL \
    GROUP BY Mission_Outcome;
```

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

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

- Total Number of Successful Outcomes: 100
- Total Number of Failure Outcomes: 1
- Numbers derived from using COUNT function on the Mission_Outcome column before grouping result by Mission_Outcome
- A possible reason for there being two rows of ‘Success’ outcome is the recording of the data (e.g. a space after the word vs no space)

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT Booster_Version \
    FROM SPACEXTBL \
    WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) \
    ORDER BY Booster_Version;
```

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

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

- Total of 12 boosters which had carried the maximum payload
- List is derived from using a nested query to execute the MAX function on the PAYLOAD_MASS__KG column as a filter option with the WHERE clause
- Results show that the main booster that carried the maximum payload is the Falcon 9 Block 5

2015 Launch Records

```
%sql SELECT SUBSTR(Date, 4, 2) as Month, "Landing _Outcome", Booster_Version, Launch_Site \
    FROM SPACEXTBL \
    WHERE "Landing _Outcome" = 'Failure (drone ship)' AND SUBSTR(Date,7,4) = '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

- Results depicts two failed landing outcomes in drone ship in the months of January and April of 2015 at CCAFS LC-40
- The booster used in both instances was the Falcon 9 v1.1, the second version of SpaceX's Falcon 9 orbital launch vehicle
- Records are derived using the SUBSTR function to extract the year from the Date column to add on as a filter criteria in the WHERE clause on top of the failed outcome in drone ship under the Landing _Outcome column

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

```
%sql SELECT "Landing _Outcome", COUNT (*) as Number_of_Outcomes\
    FROM SPACEXTBL \
    WHERE Date BETWEEN '04-06-2010' AND '20-03-2017' \
    GROUP BY "Landing _Outcome" \
    ORDER BY Number_of_Outcomes DESC;
```

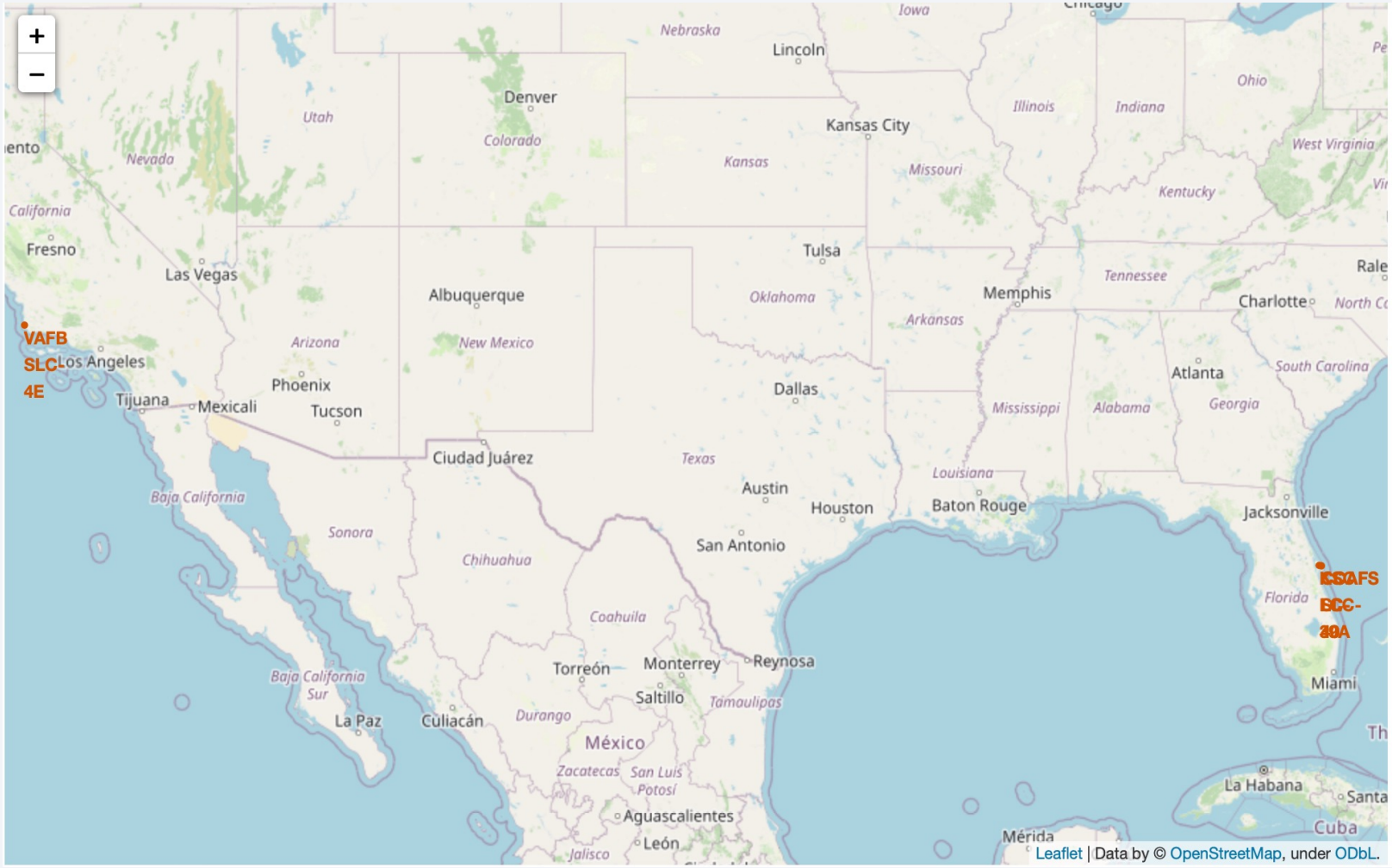
```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	Number_of_Outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

- Results show top landing outcome between 4 Jun 2010 and 20 Mar 2017 as ‘Success’ at a total of 20, followed by ‘No attempt’ at 10 and ‘No attempt’ as the last at 1
- This could be due to the way the record was being input in the database (e.g. a space after ‘attempt’ vs no space)
- The list is derived from using the COUNT function on the Landing _Outcome column and filtering the date using the WHERE clause and the BETWEEN operator before grouping the former and lastly sorting the result column in a descending order

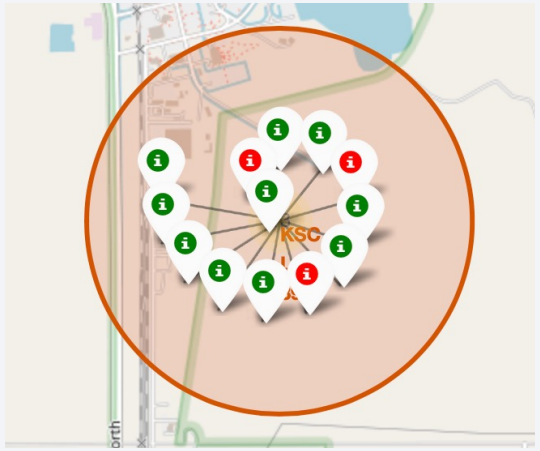

Section 3
Launch Sites
Proximities Analysis

Locations of All SpaceX Launch Sites

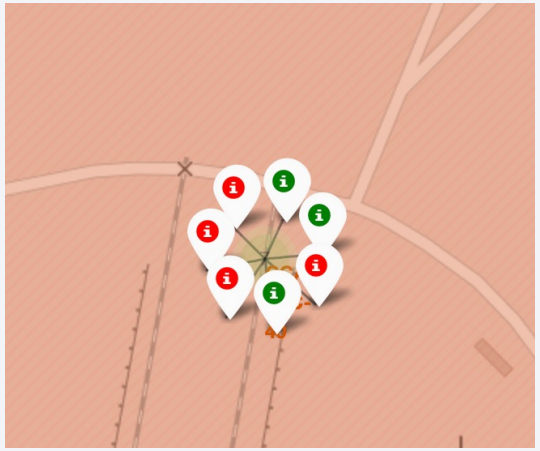


- 4 launch sites in total – 1 in the state of California (VAFB SLC-4E) and 3 in the state of Florida (CCAFS LC-40, CCAFS SLC-40, KSC LC-39A)
- All located near coasts to minimize risk of falling rockets damaging populated areas and for ease of debris retrieval
- Majority of the sites are located in Florida due to its proximity to the equator and the velocity the rocket will get from Earth's eastward rotation to help boost it into the atmosphere, saving fuel on launching and instead using it to increase speed and punch through the atmosphere

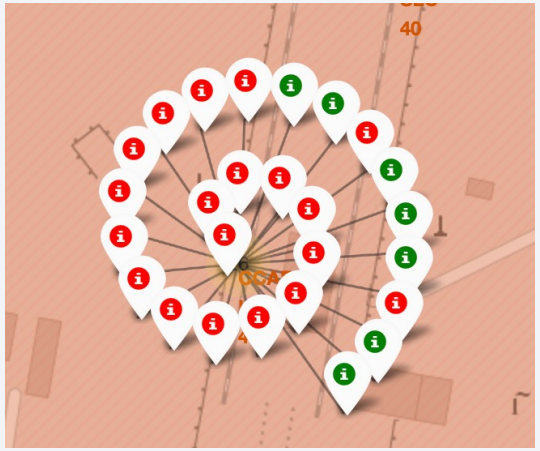
Successful and Failed Launches of Each Site



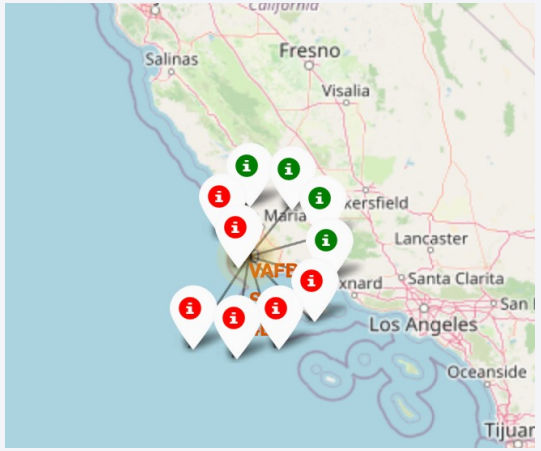
KSC LC-39A
State: Florida
Latitude: 28.573255
Longitude: -80.646895
Number of Launches: 13
Success Rate: 76.92%



CCAFS SLC-40
State: Florida
Latitude: 28.563197
Longitude: -80.576820
Number of Launches: 7
Success Rate: 42.86%



CCAFS LC-40
State: Florida
Latitude: 28.562302
Longitude: -80.577356
Number of Launches: 26
Success Rate: 26.92%

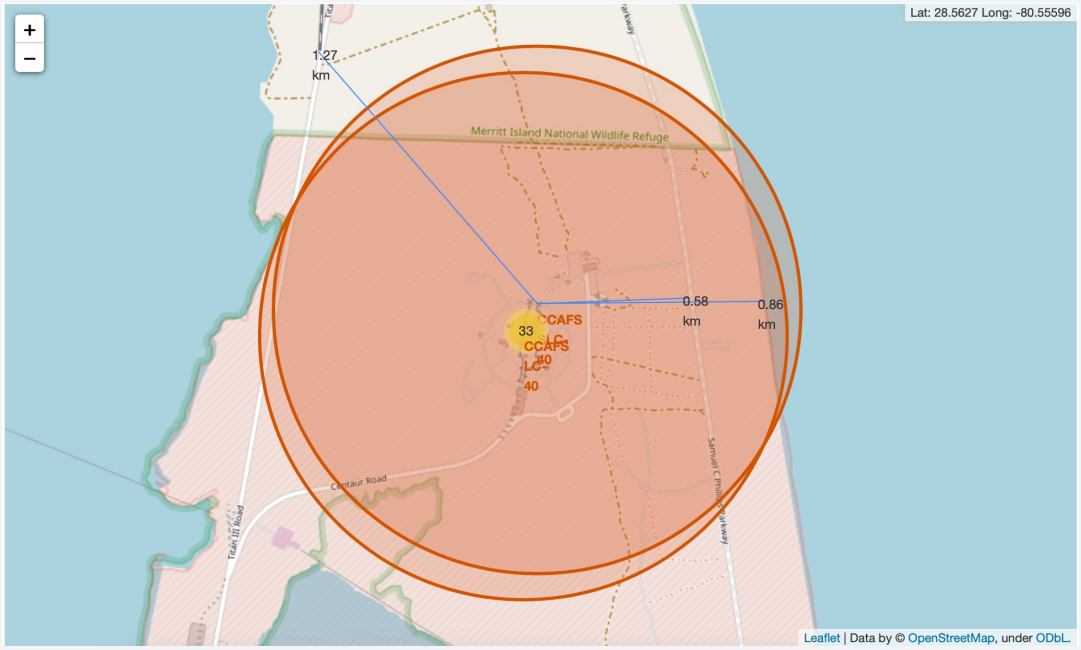


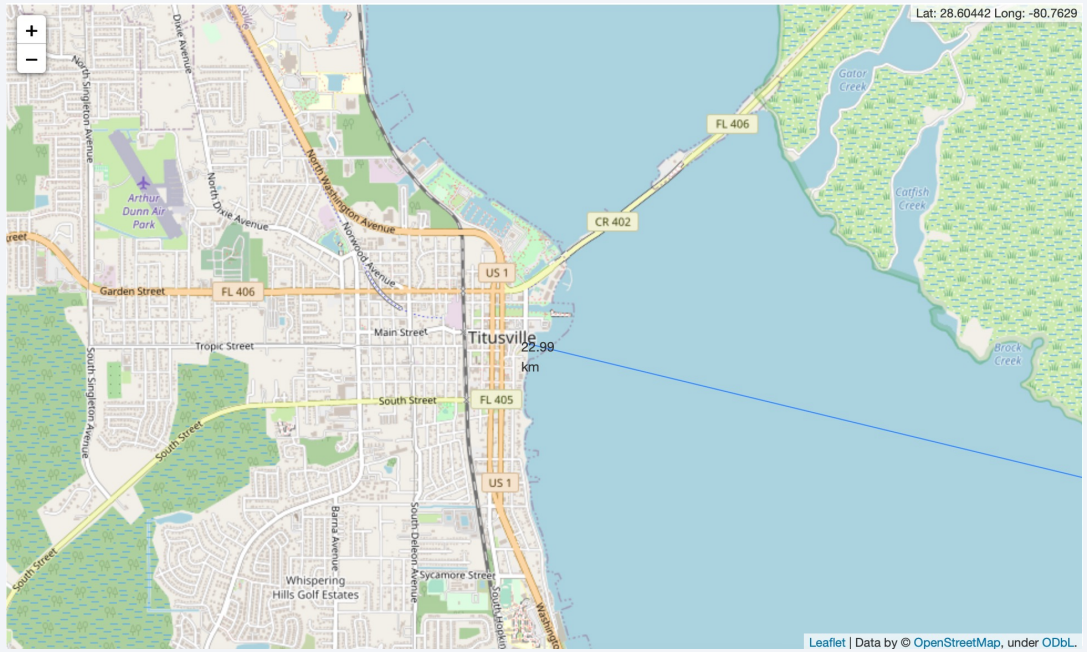
VAFB SLC-4E
State: California
Latitude: 34.632834
Longitude: -120.610746
Number of Launches: 10
Success Rate: 40%

- Each marker represents a launch with red indicating a failure and green a success
- Top two highest success rates are from sites located in the east – which adheres to the theory behind using the velocity from Earth's rotation speed to help boost the craft into the atmosphere
- A need to investigate why a drastic difference in success rate between CCAFS SLC-40 and CCAFS LC-40 despite their very close proximity to each other

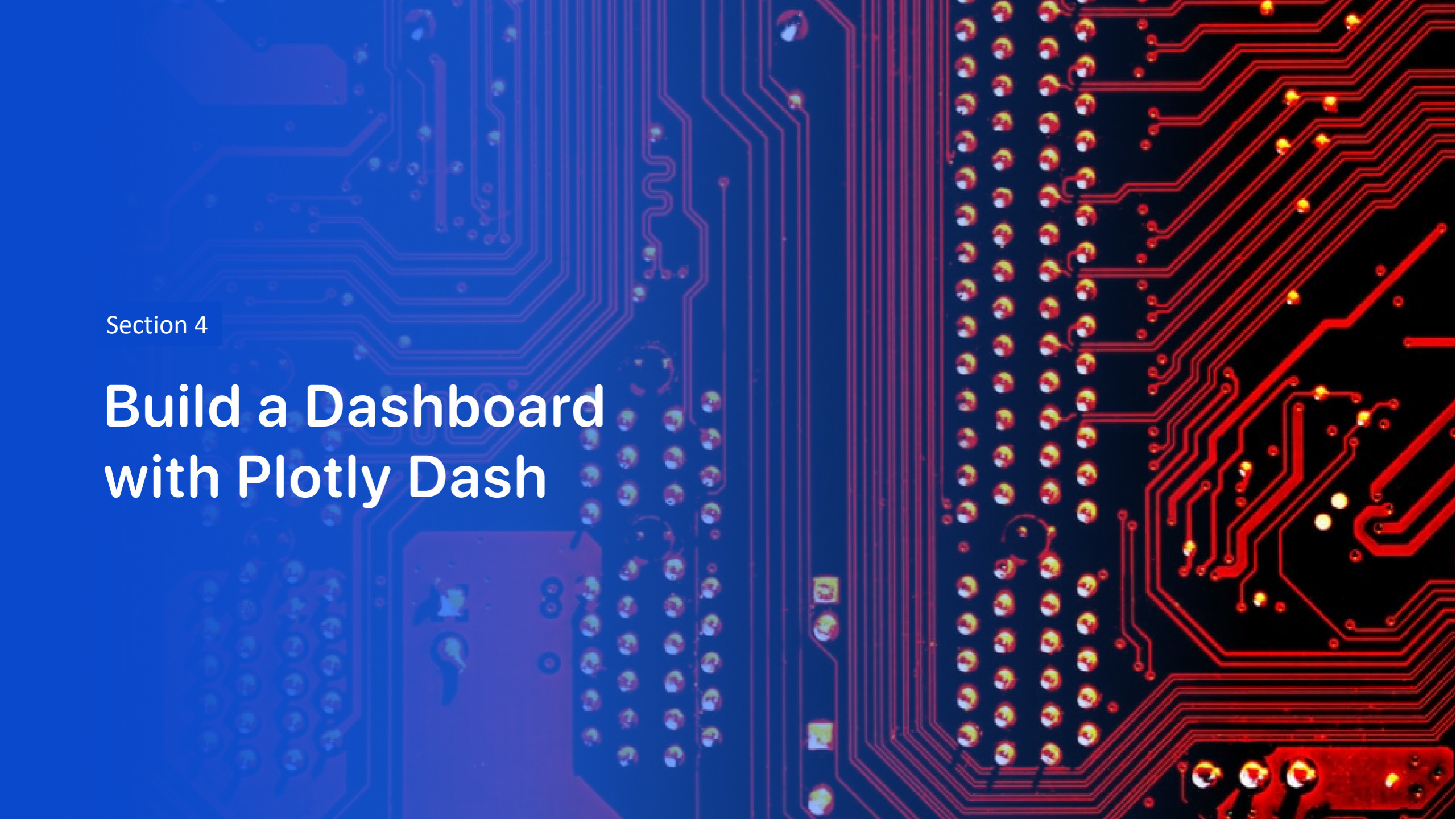
Exploring a Launch Site's Proximities

Launch Site Selected: CCAFS SLC-40

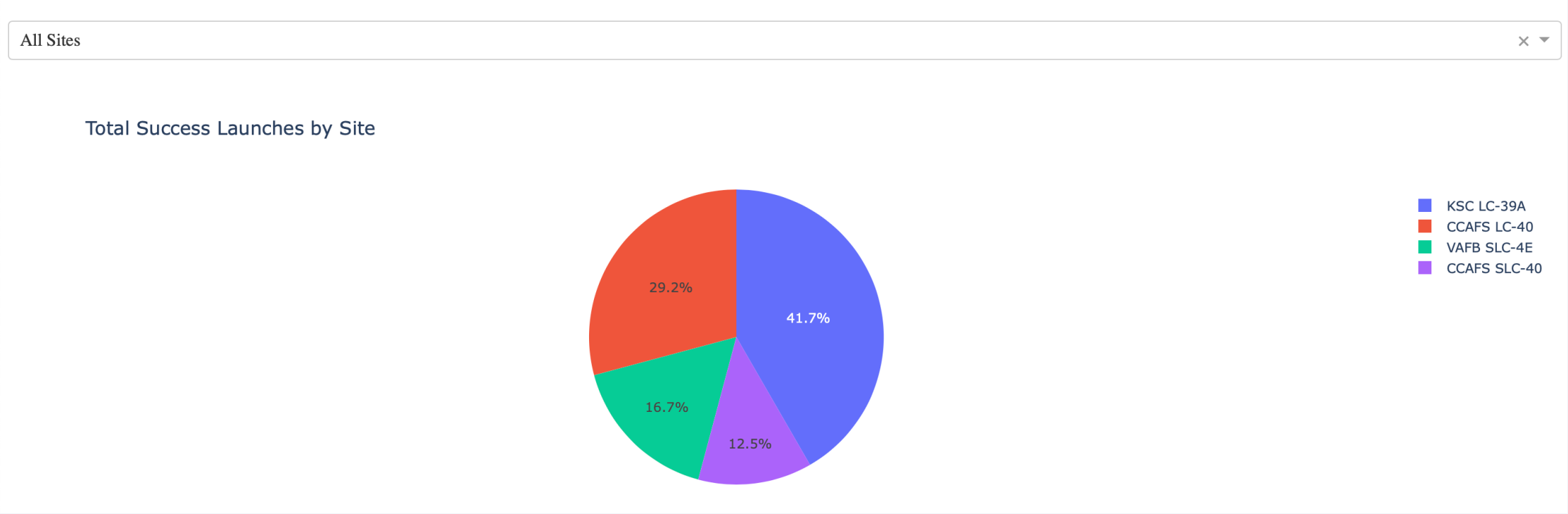




- Short distance to both railway (1.27km) and highway (0.58km) for ease of transportation of rocket and rocket parts
- Short distance to coastline (0.86km) and a substantial distance to nearest city, Titusville (22.99km), to minimize risk of falling rockets damaging populated areas and for ease of debris retrieval

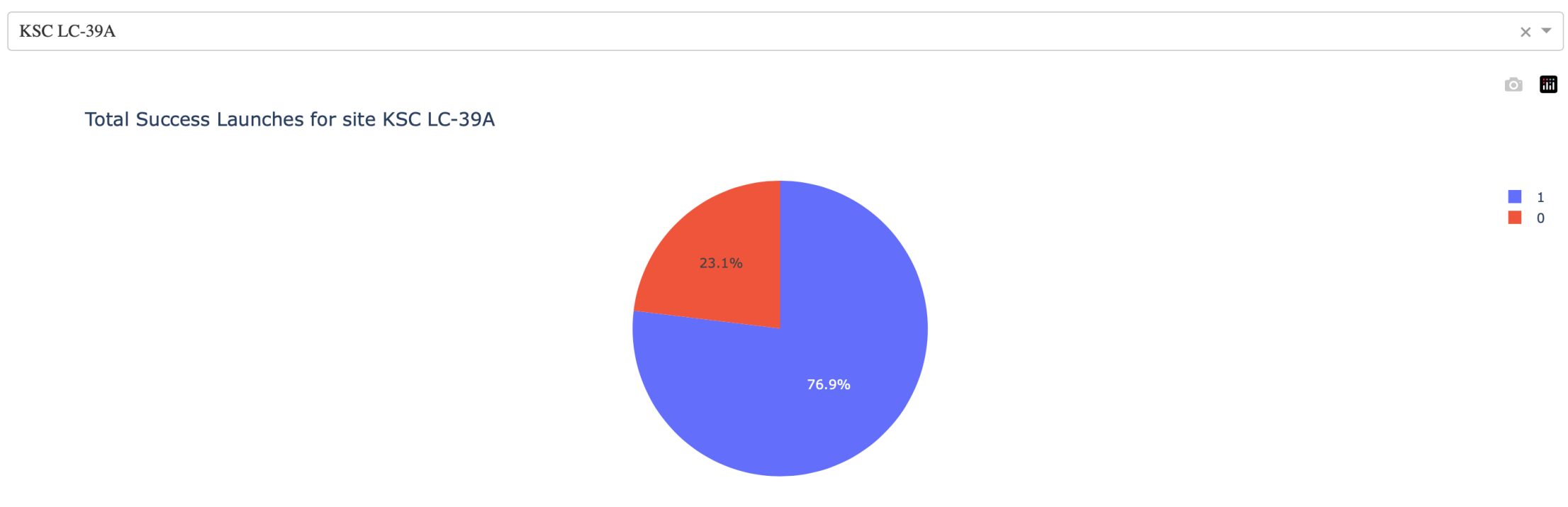

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches for All Sites



Pie chart depicts that site KSC LC-39A has the highest success rate at 41.7%, followed by CCAFS LC-40, VAFB SLC-4E and lastly CCAFS SLC-40 at 12.5%

Success Launches for Most Successful Site



- Launch site with highest launch success ratio based on previous slide is KSC LC-39A
- Diving in, we can see that the site achieved 76.9% successful launches individually

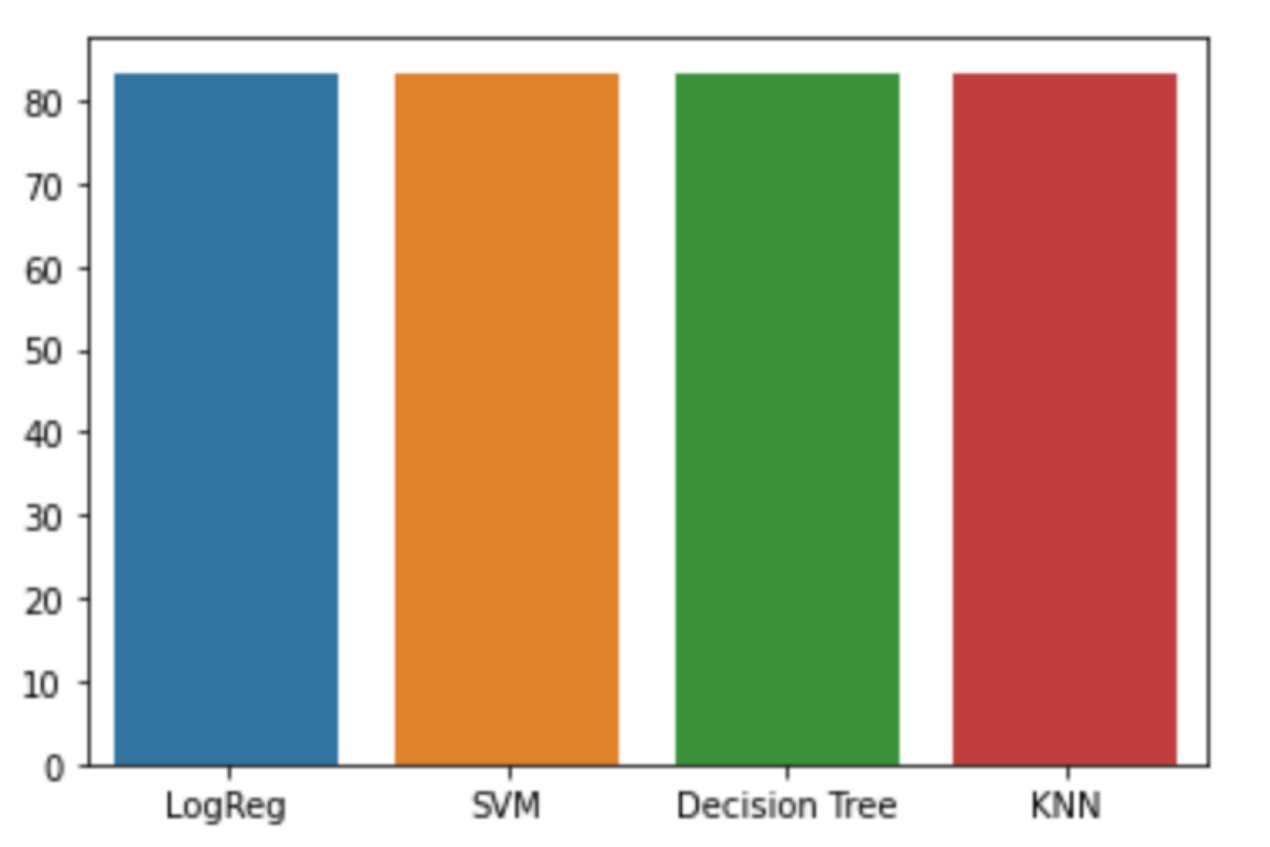
Payload vs Launch Outcome

- Payload range(s) that yields the highest success rate: 3,000 – 4,000kg
- Payload range(s) that yields the highest failure rate: 500kg - 1,000kg and 2,000kg – 3,000kg
- Booster version(s) with the highest success rate: FT
- Booster version(s) with the highest failure rate: v1.1

Section 5

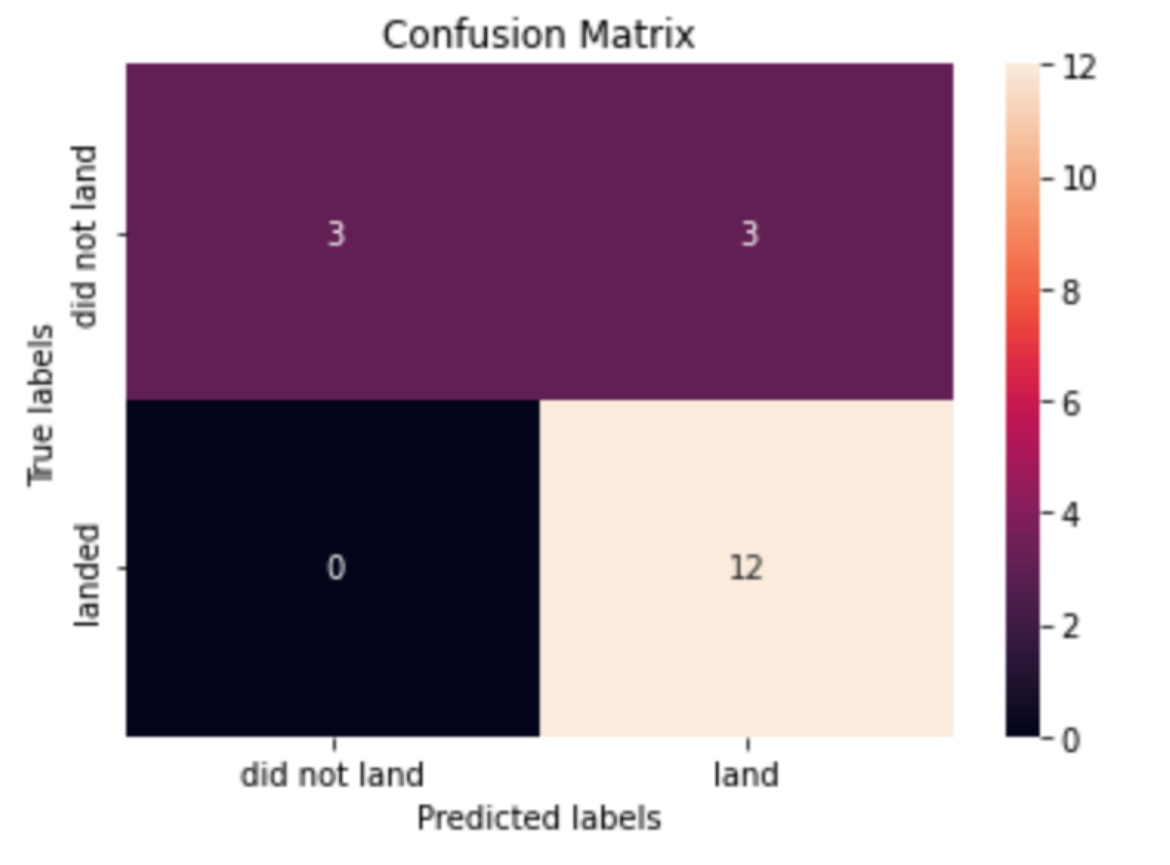
Predictive Analysis (Classification)

Classification Accuracy



All four algorithms produced the same accuracy at 83.33%

Confusion Matrix



Confusion matrix plotted for all four models is the same as they all displayed the same results and level of accuracy

Conclusions

If we can predict whether the first stage will land successfully, we can better determine whether SpaceX will reuse the first stage and subsequently the price of each launch.

We can do so through machine learning algorithms such as Logistic Regression, Support Vector Machine, Decision Tree and K Nearest Neighbours which have all yielded the same accuracy score

We can also help to facilitate a successful first landing by creating a favourable environment and setting through the analyses that we've done such as:

- The payload range that yields the highest success rate is between 3,000 to 4,000kg
- If we are launching a rocket at max payload, the booster that should be used is Falcon 9 Block 5
- Otherwise, we should use FT and avoid using v1.1
- Launches should be conducted at KSC LC-39A as it has been shown through the various analyses that its success rate is the highest
- The orbit that should be used is LEO

Thank you!