IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Stephanie Edwards
18/01/2024

Outline

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

Executive Summary

SPACE Y want to compete in the rocket launch market. As has been learnt with SPACE X, significant savings can be made if the first stage of launch can be reused. Therefore, the aim of this project is to create a machine learning pipeline to predict if the first stage will land successfully.

Summary of Methodologies:

1. Data Collection from API and Web scraping
2. Data Wrangling
3. Exploratory Data Analysis (EDA) using visualisation and SQL
4. Interactive visual analytics using Folium and Plotly Dash
5. Predictive analysis using classification models

Summary of all results:

- Logistic regression, SVM, KNN and Tree classifier models all have an accuracy score of 0.83
- The above 4 models all have the same the confusion matrix. Whilst the models can distinguish between the different classes, the major problem is false positives.

Introduction

Project background and context:

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of $62 million; other providers cost upwards 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, like Space Y, wants to bid against SpaceX for a rocket launch.
- The aim of this project is to create a machine learning pipeline to predict if the first stage will land successfully.

Problems you want to find answers

- What factors determine if the rocket will land successfully?
- What is the interaction amongst various features that determines the success rate of a successful landing?
- What is the probability of a future SpaceX Falcon 9 rocket launch landing successfully?

Section 1

Methodology

Methodology

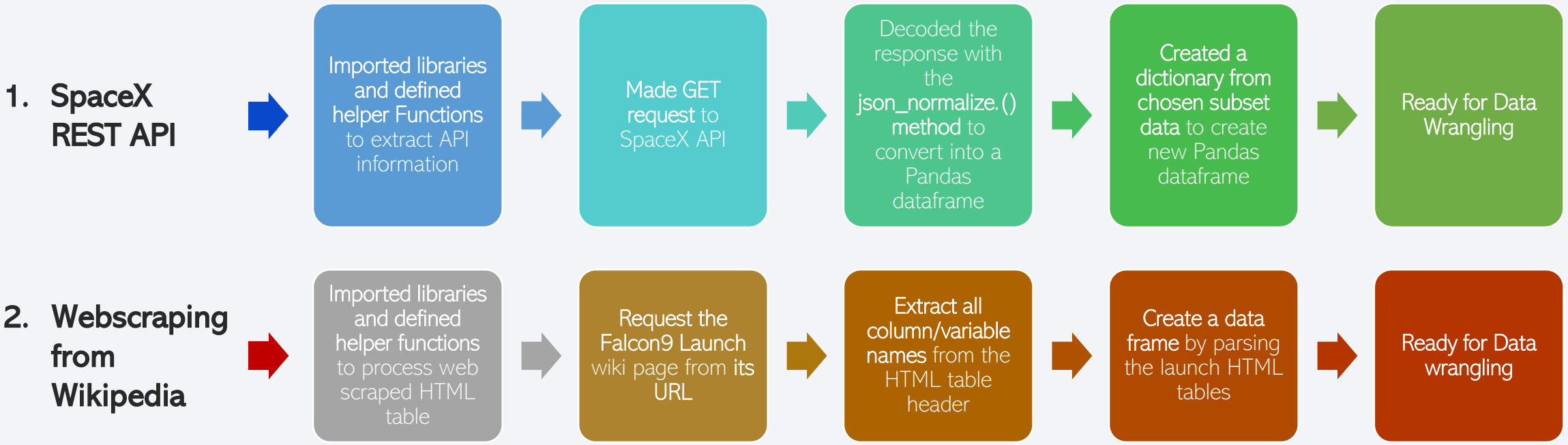
Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX REST API and web scraping from Wikipedia
- Perform data wrangling
  - One-hot encoding was applied to categorical features
  - Data cleaning of null values and irrelevant data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, KNN, SVM and Decision Tree models have been built and evaluated for the best classifier

Data Collection

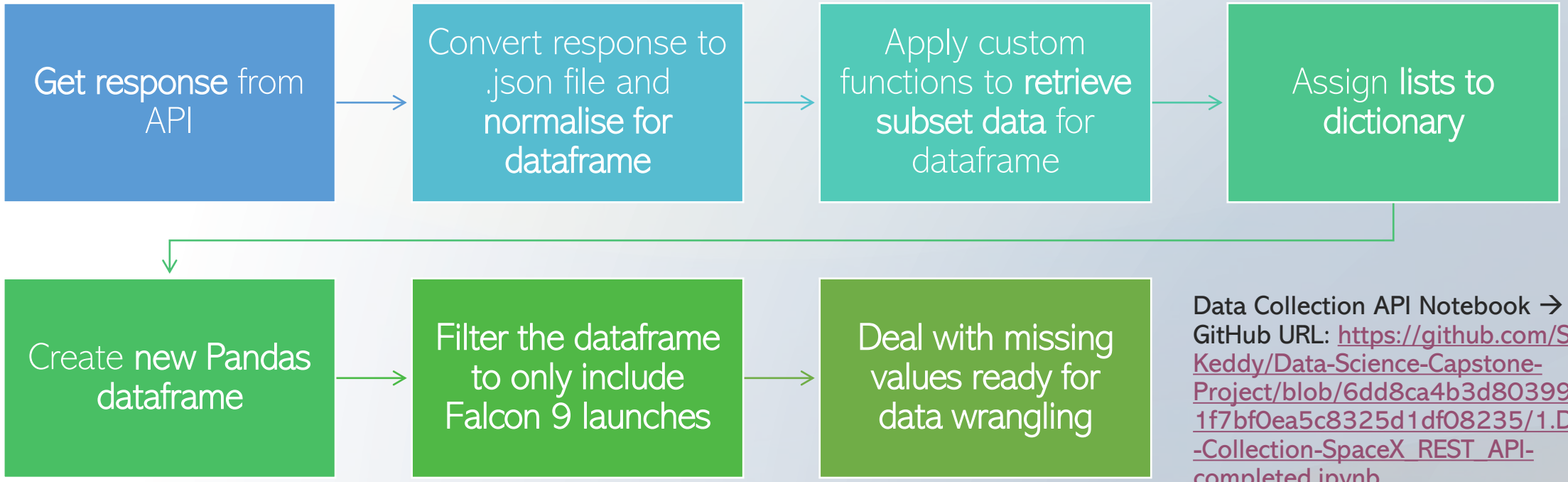
Data about the SpaceX Falcon9 was collected by using the SpaceX REST API and from Webscraping Wikipedia using BeautifulSoup.

Flowcharts of data collection process:



Data Collection – SpaceX API

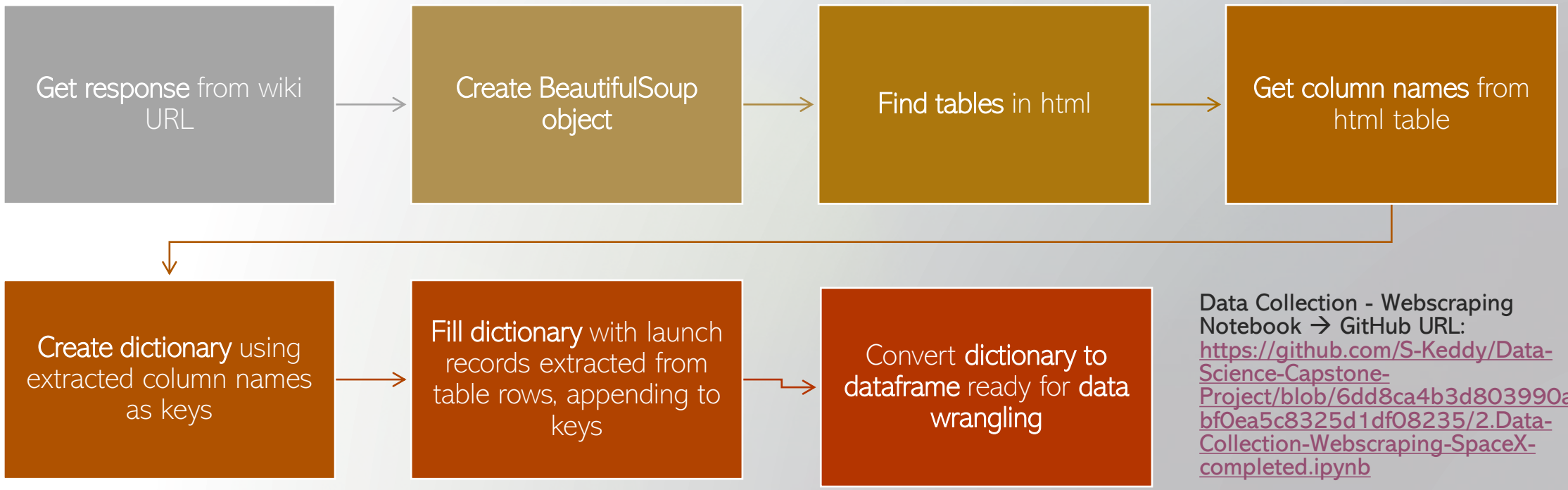
Use SpaceX REST API to collect Falcon9 historical launch data: *https://api.spacexdata.com/v4*



Data Collection API Notebook → GitHub URL: https://github.com/S-Keddy/Data-Science-Capstone-Project/blob/6dd8ca4b3d803990a1f7bf0ea5c8325d1df08235/1.Data-Collection-SpaceX_REST_API-completed.ipynb

Data Collection - Webscraping

Webscraping to collect Falcon 9 historical launch records from a Wikipedia page titled “List of Falcon 9 and Falcon Heavy launches” : *https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches*



Data Collection - Webscraping Notebook → GitHub URL: https://github.com/S-Keddy/Data-Science-Capstone-Project/blob/6dd8ca4b3d803990a1f7bf0ea5c8325d1df08235/2.Data-Collection-Webscraping-SpaceX-completed.ipynb

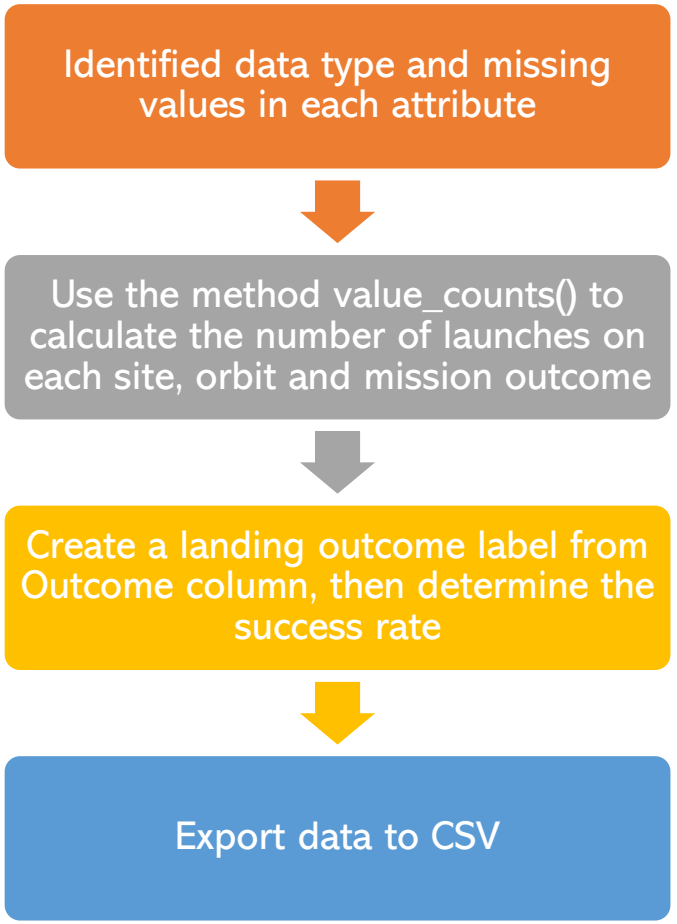
Data Wrangling

We performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.

In the data set, there are 6 different mission outcome labels. We need to convert those outcomes into training labels with 1 meaning the booster successfully landed and 0 meaning it was unsuccessful.

Data Wrangling Notebook → GitHub URL: https://github.com/S-Keddy/Data-Science-Capstone-Project/blob/6dd8ca4b3d803990a1f7bf0ea5c8325d1df08235/3.Data-Wrangling-SpaceX-completed.ipynb

Data Wrangling Process:



EDA with Data Visualization

Using the collected data, 7 charts were plotted as follows:

1.	**Scatter plot**: to see how the Flight Number (indicating the continuous launch attempts) and Payload variables would affect the launch outcome.*
2.	**Scatter plot**: to visualise the relationship between Flight Number and Launch Site
3.	**Scatter plot**: to visualise the relationship between Payload Mass and Launch Site
4.	**Scatter plot**: to visualise the relationship between Flight Number and Orbit Type
5.	**Scatter plot**: to visualise the relationship between Payload and Orbit Type
6.	**Bar chart**: to visualise the relationship between success rate of each Orbit Type
7.	**Line chart**: to visualise the launch success yearly trend

EDA with Data Visualisation Notebook → GitHub URL: https://github.com/S-Keddy/Data-Science-Capstone-Project/blob/0c1c9dc02a4388b0bdbd0377d4a70f863c34f6f8/5.EDA-data-visualisation-completed.ipynb

*See appendix for this scatter plot, slide 46

EDA with SQL

Summary of the SQL queries performed:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass
- List the failure landing_outcomes in drone ship, booster versions, launch site by month in year 2015
- In descending order, rank the count of landing outcomes between 2010-06-04 and 2017-03-20

EDA with SQL Notebook → GitHub URL: https://github.com/S-Keddy/Data-Science-Capstone-Project/blob/26b2cd38ea2e12ee25619f8915d8fb64da047b4e/4.EDA-SQL-completed.ipynb

Build an Interactive Map with Folium

Using Folium, a map was created with the following map objects:

- Circles and markers have been added as a Marker Cluster for each launch site.
- Each marker has been colour coordinated as either green or red to denote if the launch was a success or failure, respectively.
- MousePosition was added to the map to get the coordinate for a mouse over a point in the map.
- Distances between launch site and the nearest coastline, railway, highway and city have been calculated and marked on the map. A line has also been drawn from each of these distance markers to the launch site.

These object markers have been added so we can understand some geographical patterns about launch sites, such as proximity to the nearest coastline, railway, highway and city.

Interactive Map with Folium Notebook → GitHub URL: https://github.com/S-Keddy/Data-Science-Capstone-Project/blob/d27ca49e5dcfe96bb9ca82ff16c6e33ba051518b/6.Interactive-Analytics-FOLIUM-SpaceX-launch-sites-completed.ipynb

Build a Dashboard with Plotly Dash

A dashboard has been created with 4 components:

1. *Dropdown menu* for launch site: can select ALL sites, or individual site
2. *Range slider* for payload mass: can select specific payload range
3. *Pie chart* to visualise success rate of launch site(s)
    i. *Interacts with dropdown menu so can either view [ALL] 'Total Success Launches by Site' or 'Launch success vs failure for site [SELECTED]'*
4. *Scatter plot* to visualize relationship between launch site, payload and booster version
    i. *Interacts with launch site dropdown menu and payload mass range slider*

These plots and interactions have been added to analyze SpaceX launch data to understand:

- Which site has the largest successful launches?
- Which site has the highest launch success rate?
- Which payload range(s) has the highest launch success rate?
- Which payload range(s) has the lowest launch success rate?
- Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate?

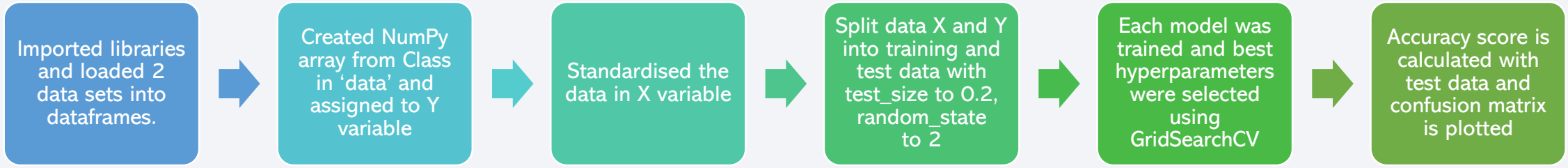
Build a Dashboard with Plotly Dash Notebook → GitHub URL: https://github.com/S-Keddy/Data-Science-Capstone-Project/blob/6dd8ca4b3d803990a1f7bf0ea5c8325d1df08235/7.Interactive-Dashboard-Plotly-Dash-app-SpaceX-completed.py

Predictive Analysis (Classification)

Built and trained 4 machine learning models to determine which would be the best choice to predict if the first stage will successfully land given data the data from preceding labs.

Models built and trained: Logistic Regression, SVM, Decision Tree and KNN

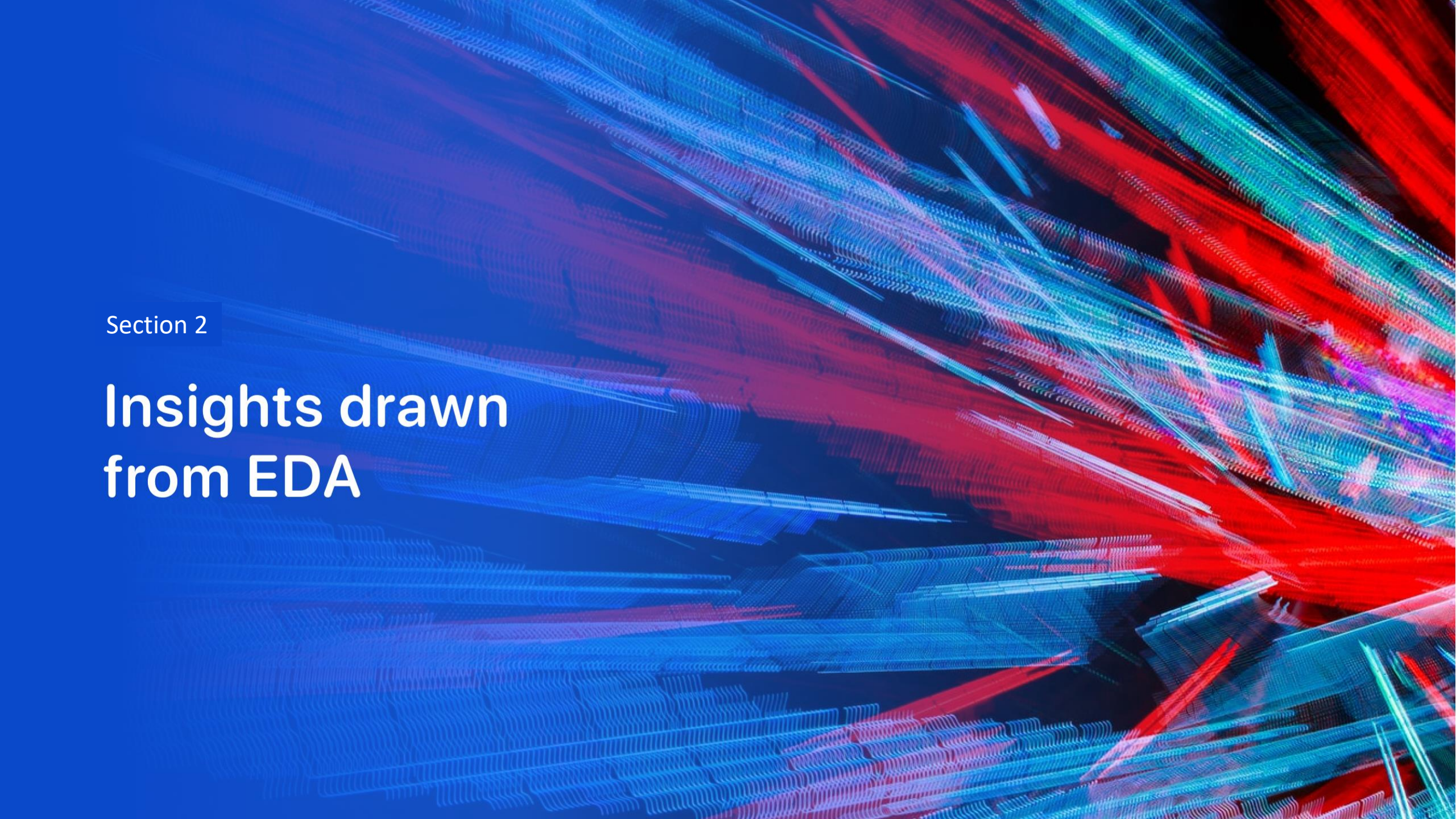
Flow chart of process:



Predictive Analysis – Machine Learning Notebook → GitHub URL: https://github.com/S-Keddy/Data-Science-Capstone-Project/blob/6dd8ca4b3d803990a1f7bf0ea5c8325d1df08235/8.Predictive-Analysis-Classification-Models-Machine-Learning-completed.ipynb

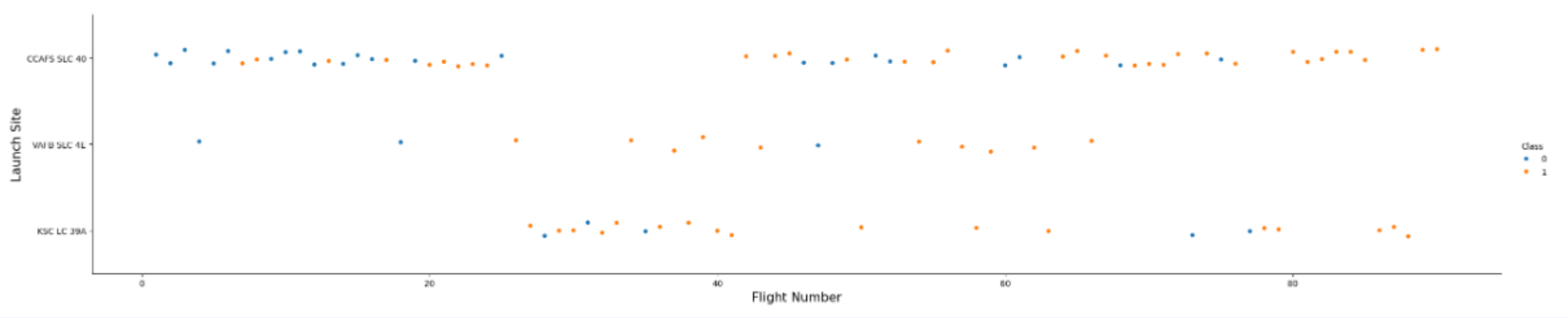
Results

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


Section 2
Insights drawn from EDA

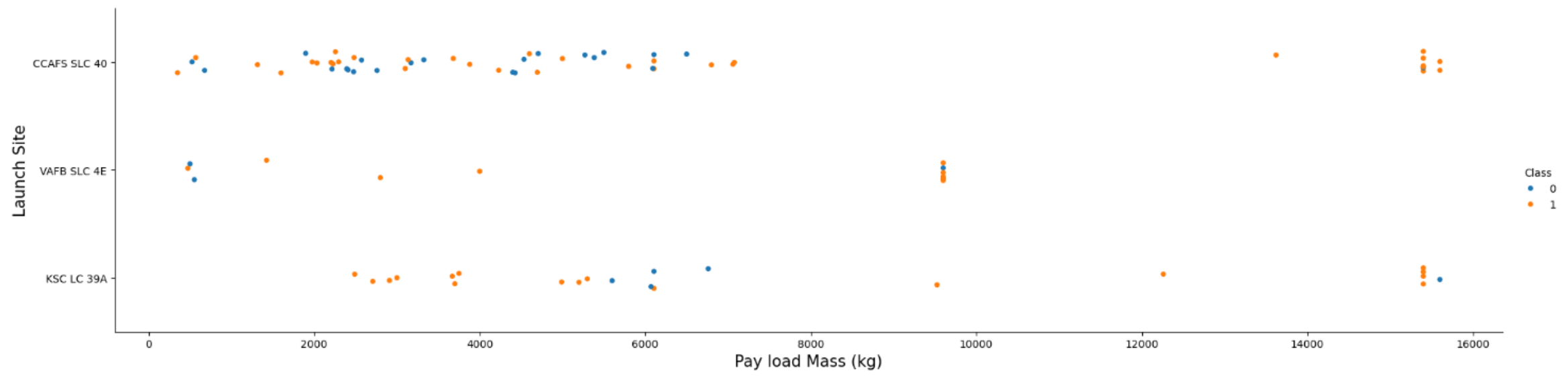
Flight Number vs. Launch Site

Scatter plot of Flight Number vs. Launch Site



Site CCAFS SLC 40 has the highest number of launches than any other site

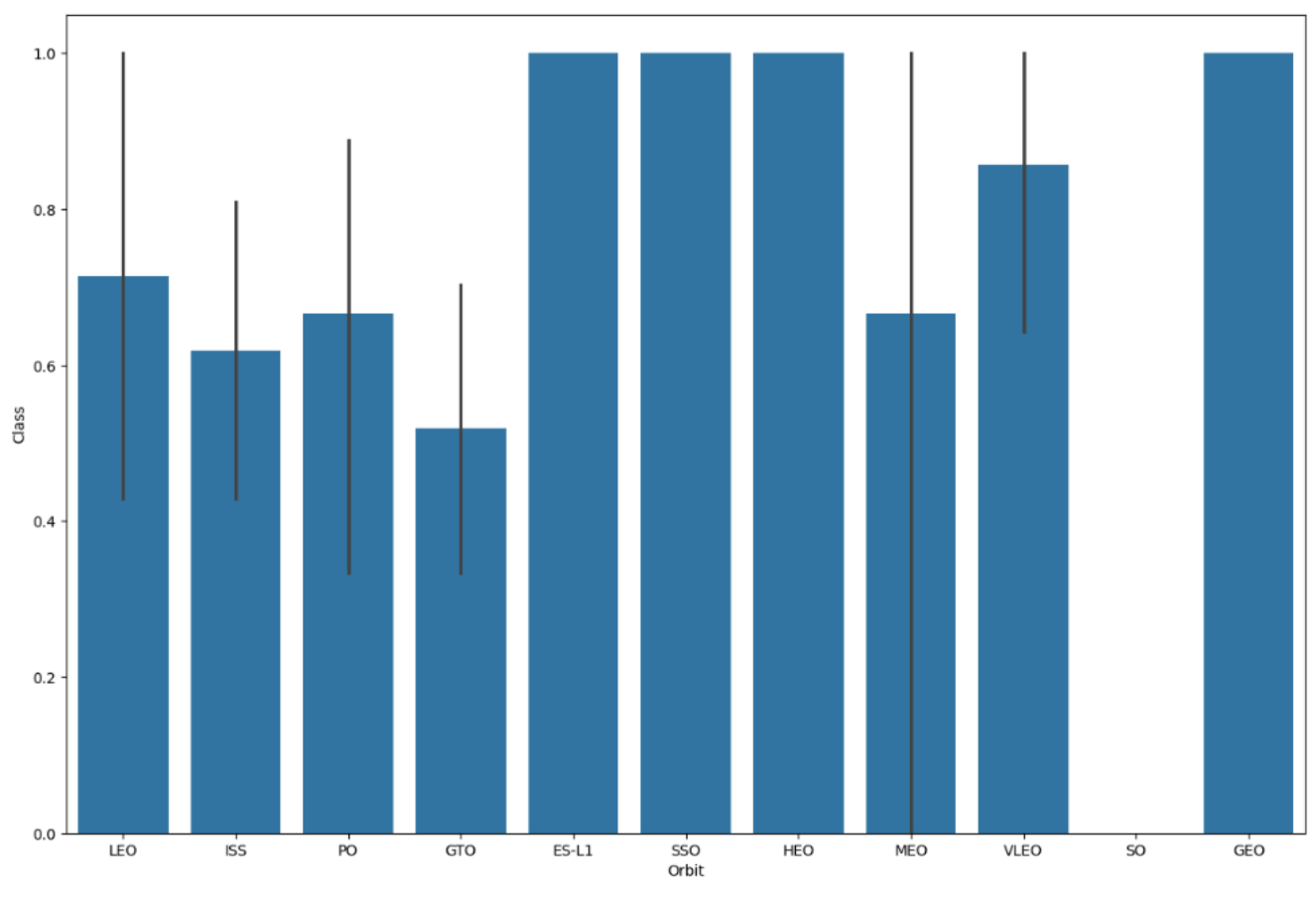
Payload vs. Launch Site



Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

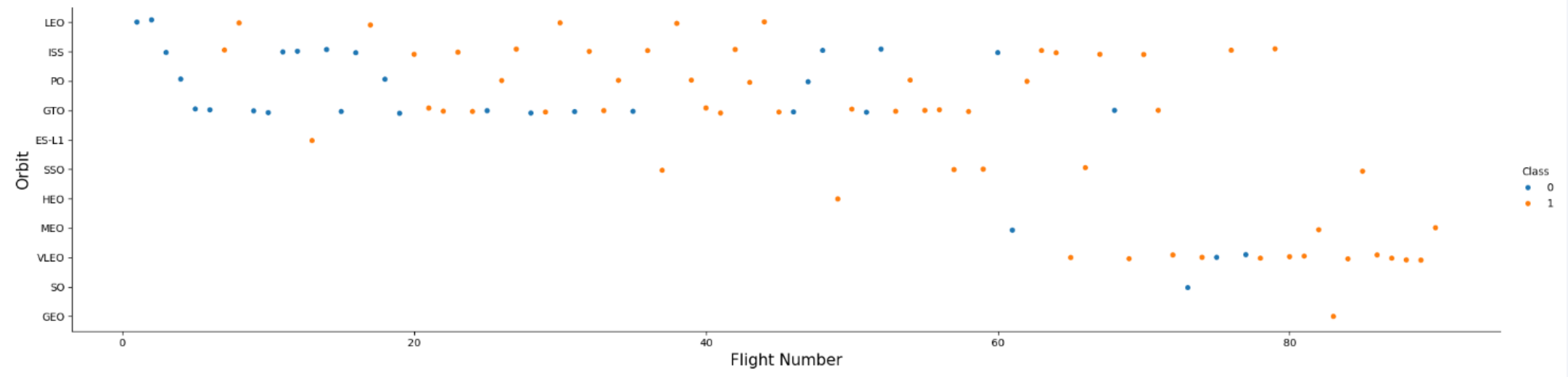
Bar chart for the success rate of each orbit type



Orbit ES-L1 SSO, HEO and GEO have the best success rate

Flight Number vs. Orbit Type

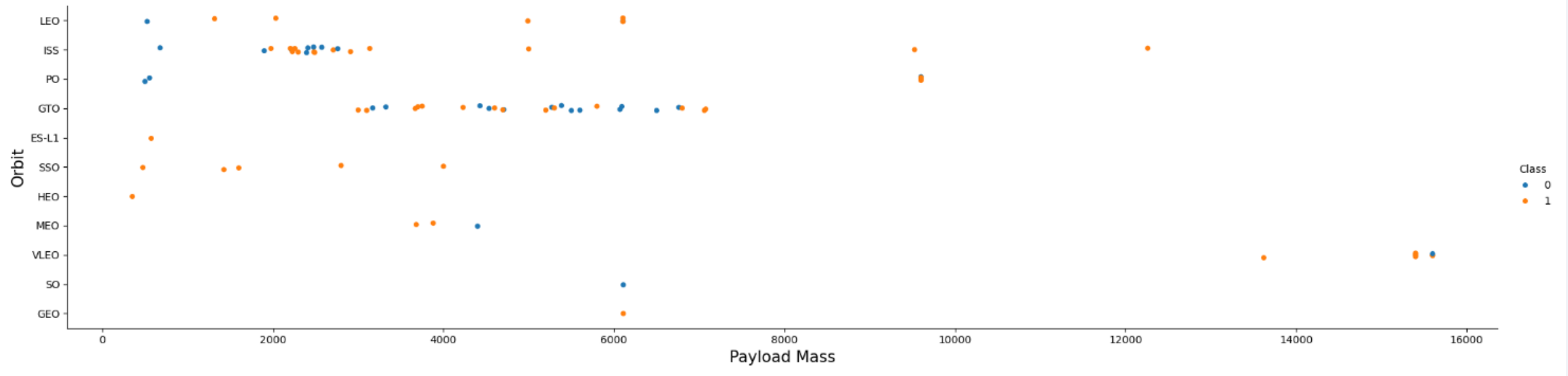
Scatter plot of Flight number vs. Orbit type



You should see that in the 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

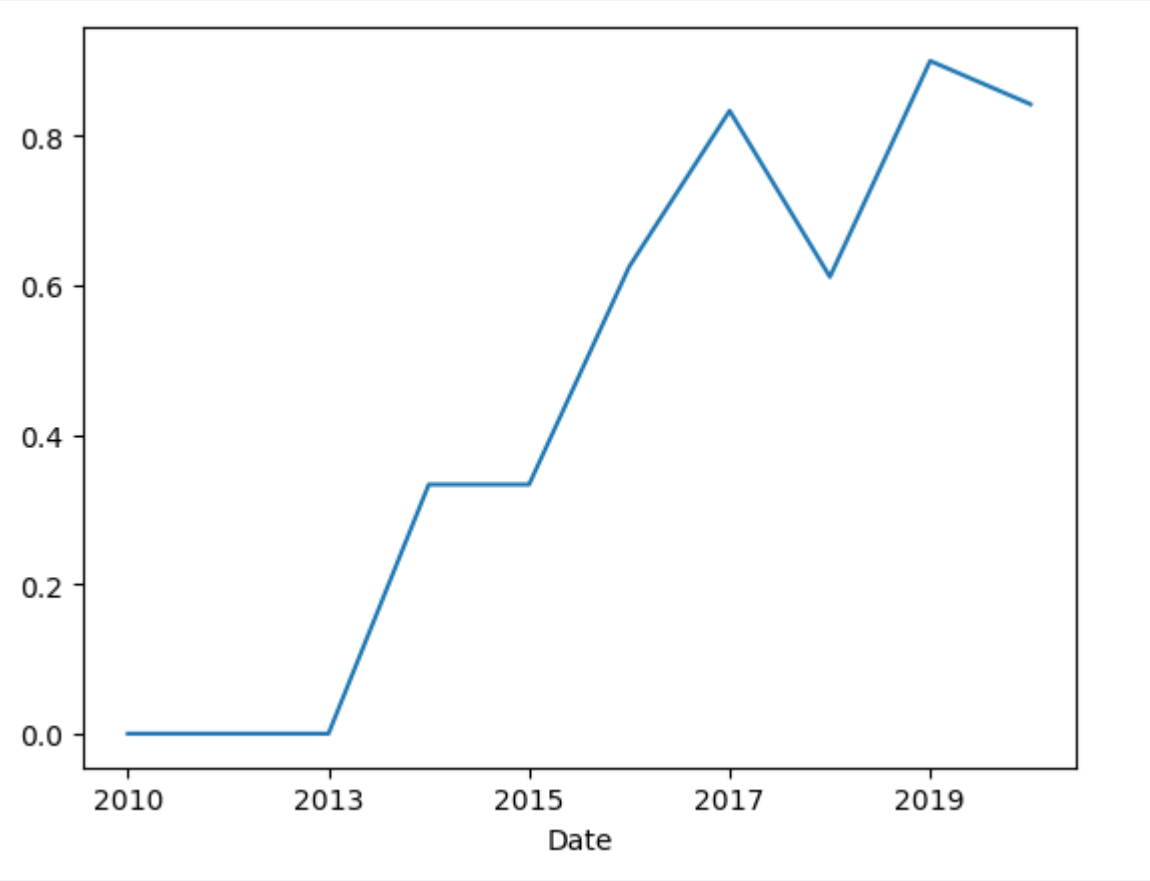
Scatter point of payload vs. orbit type



We can observe that with heavy payloads, the successful landing rate are more for PO, LEO and ISS

Launch Success Yearly Trend

Line chart of yearly average success rate



We can observe that the success rate since 2013 kept increasing until 2020

All Launch Site Names

Names of the unique launch sites:

```
In [10]: %sql SELECT distinct("Launch_site") FROM SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

Out[10]:

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

We used the **DISTINCT** statement to show only unique launch sites from the SpaceX data

As the table shows, there are 4 unique launch sites.

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * FROM SPACEXTABLE 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 table shows 5 records where launch sites begin with ‘CCA’.

The query uses WHERE, LIKE and LIMIT statements.

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT Customer, SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS From SPACEXTABLE  WHERE Customer = "NASA (CRS)"
```

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

Customer	TOTAL_PAYLOAD_MASS
NASA (CRS)	45596

The total payload mass carried by boosters launched by NASA (CRS) is 45,596 KG

This query has been created using SUM function and WHERE clause.

Average Payload Mass by F9 v1.1

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

Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT Booster_Version, AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS FROM SPACEXTABLE WHERE Booster_Version = "F9 v1.1"
```

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

Booster_Version	AVG_PAYLOAD_MASS
F9 v1.1	2928.4

This query has used the AVG function, and WHERE clause

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT MIN("Date"), Landing_Outcome FROM SPACEXTABLE WHERE Landing_Outcome = "Success (ground pad)"
```

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

MIN("Date")	Landing_Outcome
2015-12-22	Success (ground pad)

The first successful landing outcome in ground pad was on 2015-12-22.

This query was created using the MIN function and WHERE clause

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT
    Booster_Version,
    PAYLOAD_MASS__KG_,
    Landing_Outcome
FROM
    SPACEXTABLE
WHERE
    PAYLOAD_MASS__KG_ >4000 AND PAYLOAD_MASS__KG_<6000 AND Landing_Outcome ="Success (drone ship)"
```

```
 * 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)

This query was created with a WHERE clause and AND operator.

There were 4 successful drone ship landings with payload between 4000 and 6000, as per the table on the left.

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT
    'Success' AS Outcome,
    SUM(CASE WHEN Mission_Outcome LIKE '%Success%' THEN 1 ELSE 0 END) AS Count
FROM
    SPACEXTABLE
GROUP BY
    Outcome

UNION

SELECT
    'Failure' AS Outcome,
    SUM(CASE WHEN Mission_Outcome LIKE '%Failure%' THEN 1 ELSE 0 END) AS Count
FROM
    SPACEXTABLE
GROUP BY
    Outcome;
```

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

Outcome	Count
Failure	1
Success	100

- The total number of successful and failure mission outcomes was 100 and 1, respectively.

- To count the sum of all the ‘success’ outcomes, used a CASE statement with LIKE operator and grouped by ‘Outcome’. Then used a UNION operator and repeated the process for finding ‘success’ outcomes, but for ‘failure’ outcomes instead.

Boosters Carried Maximum Payload

```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE WHERE PAYLOAD_MASS__KG_= (select MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
 * 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

← Here is a list of names of boosters that carried the maximum payload mass of 15600KG

- A subquery was used in the WHERE clause to find the maximum payload using the MAX function

2015 Launch Records *[where landing outcome = Failure (drone ship)]*

```
%%sql
SELECT
    (SELECT SUBSTR(Date, 6,2)) AS Month,
    (SELECT substr(Date,0,5)) AS Year,
    Booster_Version, Launch_Site, Landing_Outcome
    FROM
        SPACEXTABLE
    WHERE Landing_Outcome = "Failure (drone ship)" AND "Date" Like "2015%"
```

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

Month	Year	Booster_Version	Launch_Site	Landing_Outcome
01	2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

← Here is a list of the failed landing outcomes in drone ship, also showing their respective booster version and launch site name in the year 2015.

- A Nested select was used to find Month and Year
- A WHERE clause was then used to filter the nested select by the year 2015 and Landing_Outcome to show only “Failure (drone ship)”.

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

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

←HERE is a table showing the rank of landing outcomes between 2010-06-04 and 2017-03-20.

←In total, there were 8 category types of landing outcome, with 'No attempt' ranking the highest, and 'precluded (drone ship)' as the lowest.

```
%%sql

SELECT
    Landing_Outcome,
COUNT
    (Landing_Outcome) As Count_of_Landing_Outcomes
FROM
    SPACEXTABLE
WHERE
    Date between "2010-06-04" and "2017-03-20"
GROUP BY
    Landing_Outcome
ORDER BY
    Count_of_Landing_Outcomes DESC
```

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

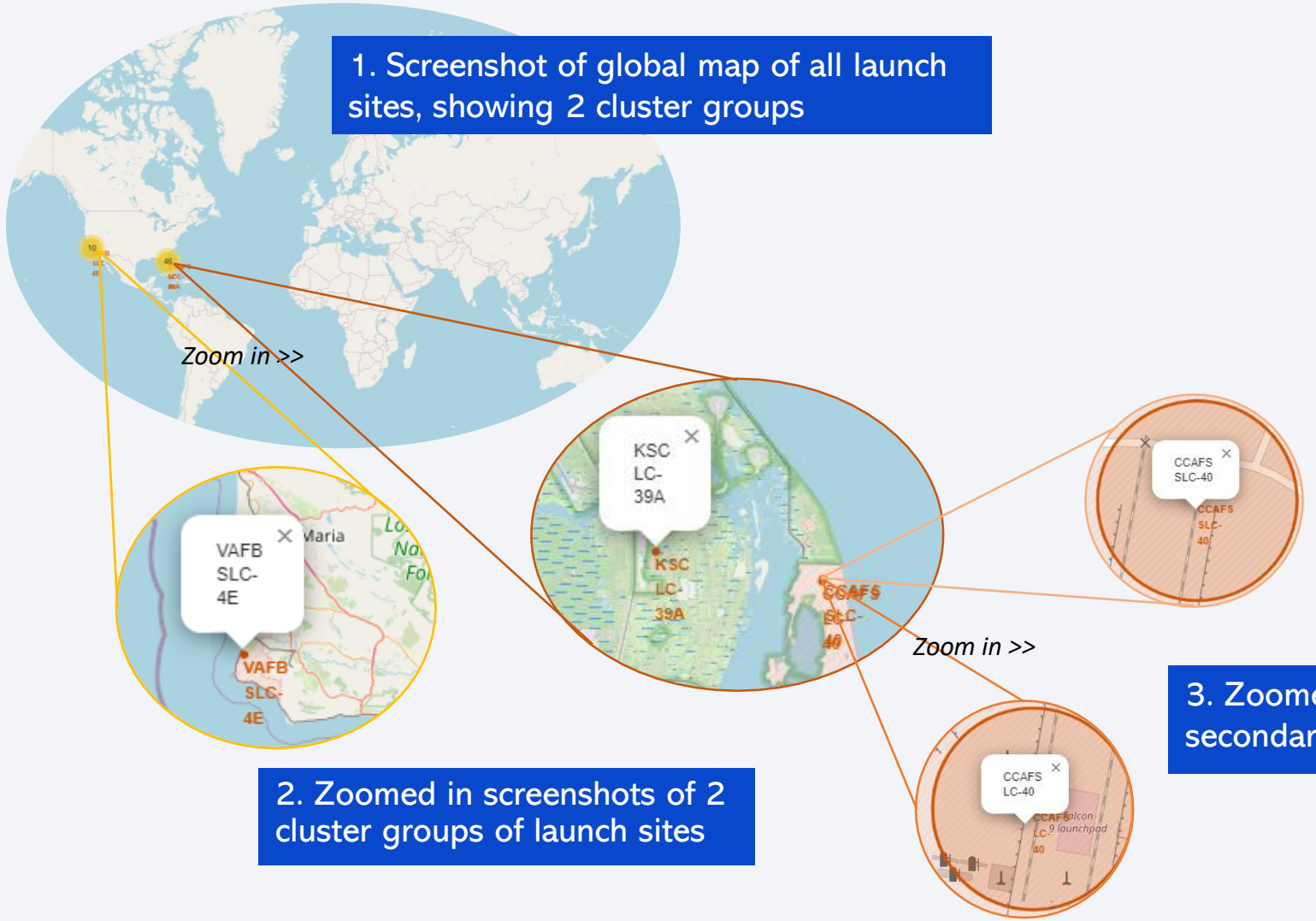
To create this rank of outcomes between 2010-06-04 and 2017-03-20, a query was created as follows:

- *Selected landing outcomes, and the COUNT of landing outcomes*
- *Then filtered by the desired date range in a WHERE clause*
- *Then grouped by landing outcome, and finally ordered by the count of landing outcomes in descending order.*

Section 3

Launch Sites Proximities Analysis

Global Map of All Launch Sites

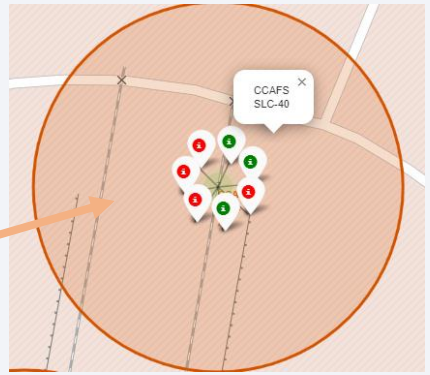


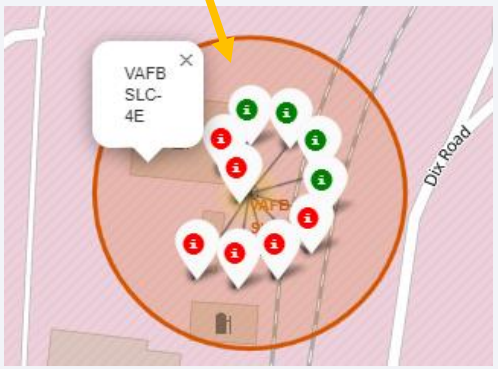
Findings:

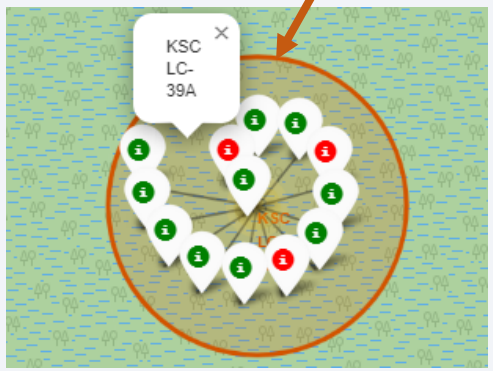
- All launch sites are in very close proximity to the coast
- All launch sites are in proximity to the equator
- All launch sites are in restricted areas

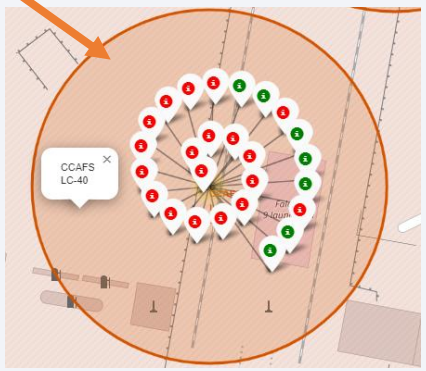
Launch site locations: Success vs Failure

Map showing outcomes of launches at each launch site location, with green markers representing ‘successful’ outcome and red markers representing ‘failure’ outcome.









Findings:

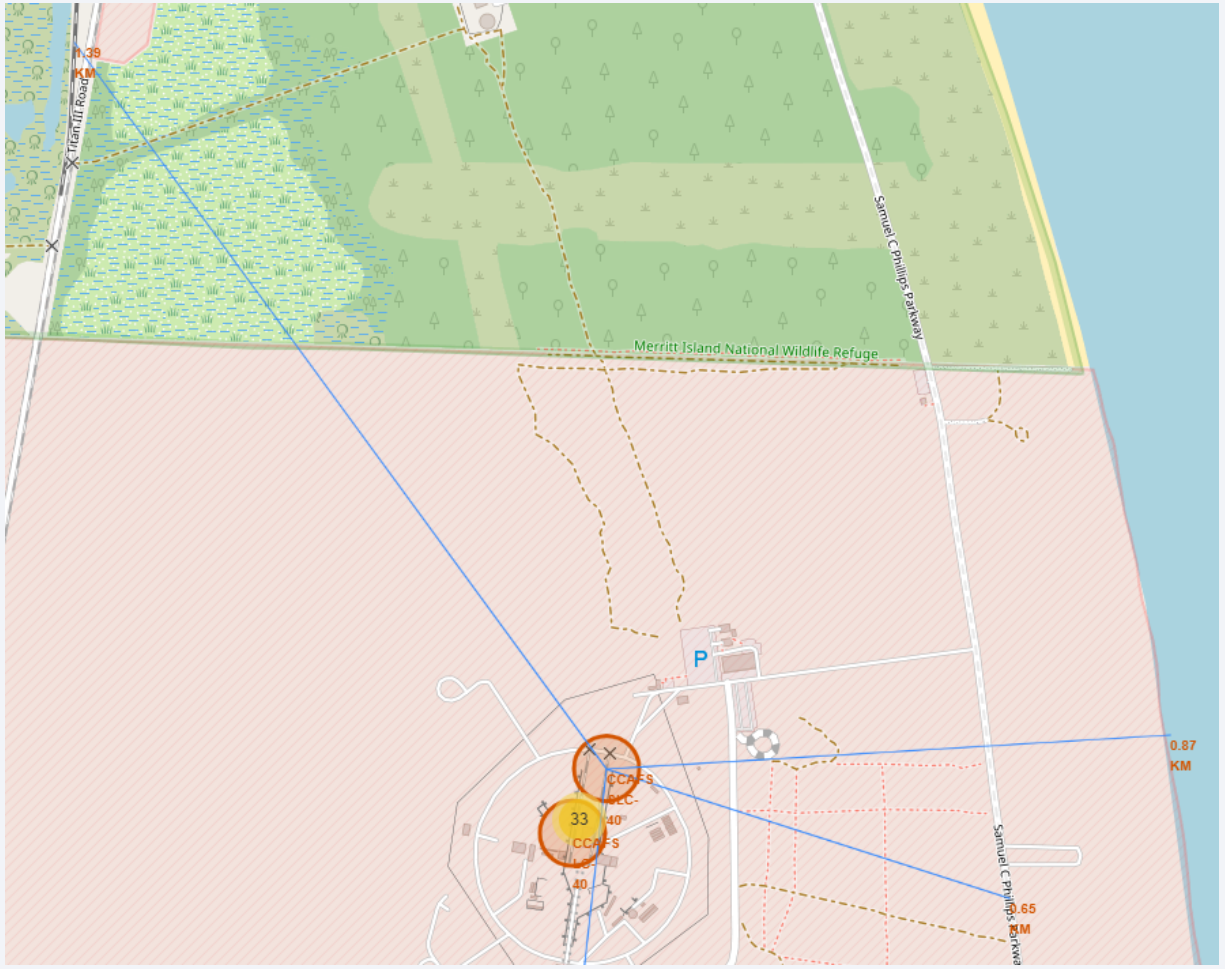
- Launch site KSC LC-39A has the highest proportion of successful outcomes

Key:

Success outcome

Failure outcome

Distance between Launch Site and its Proximities



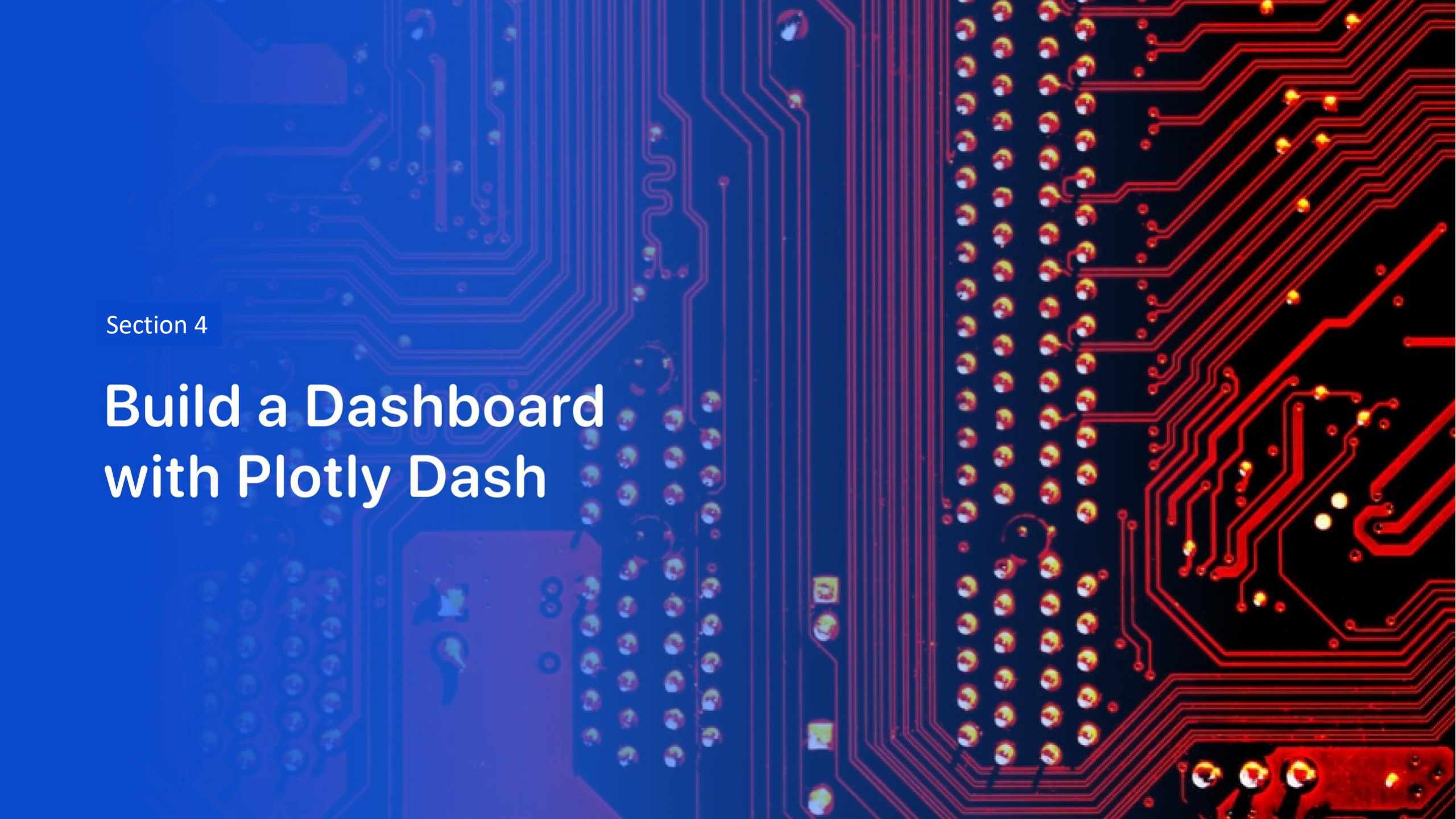
Selected launch site: **CCAFS SLC 40**

Distance to proximities:

- Coastline: 0.87KM
- Railway *(NASA Railroad)* : 1.39KM
- Highway *(Samuel Phillips Parkway)* : 0.65KM
- City *(Melbourne):* 51.95KM

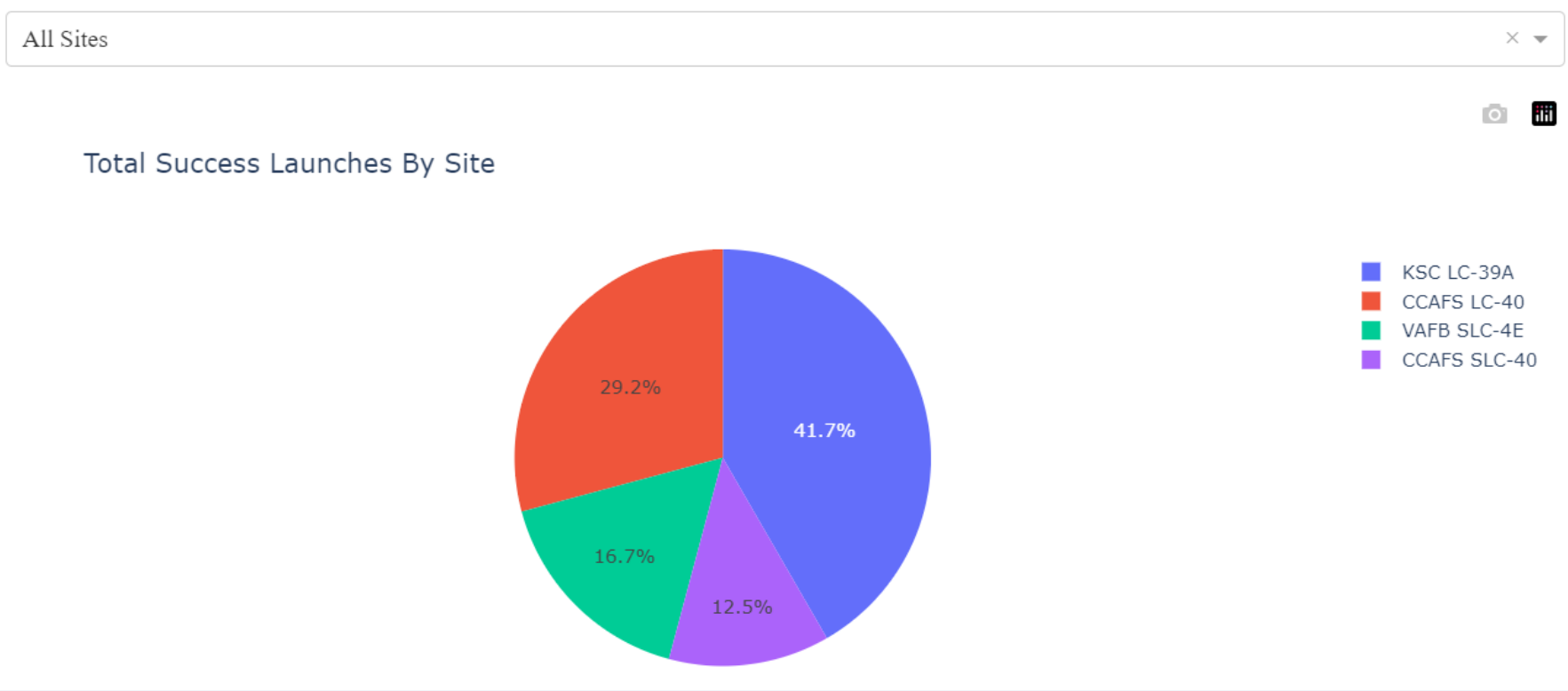
Findings:

- Launch sites are sizeable distance from cities
- Launch sites are close to the coastline


Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites

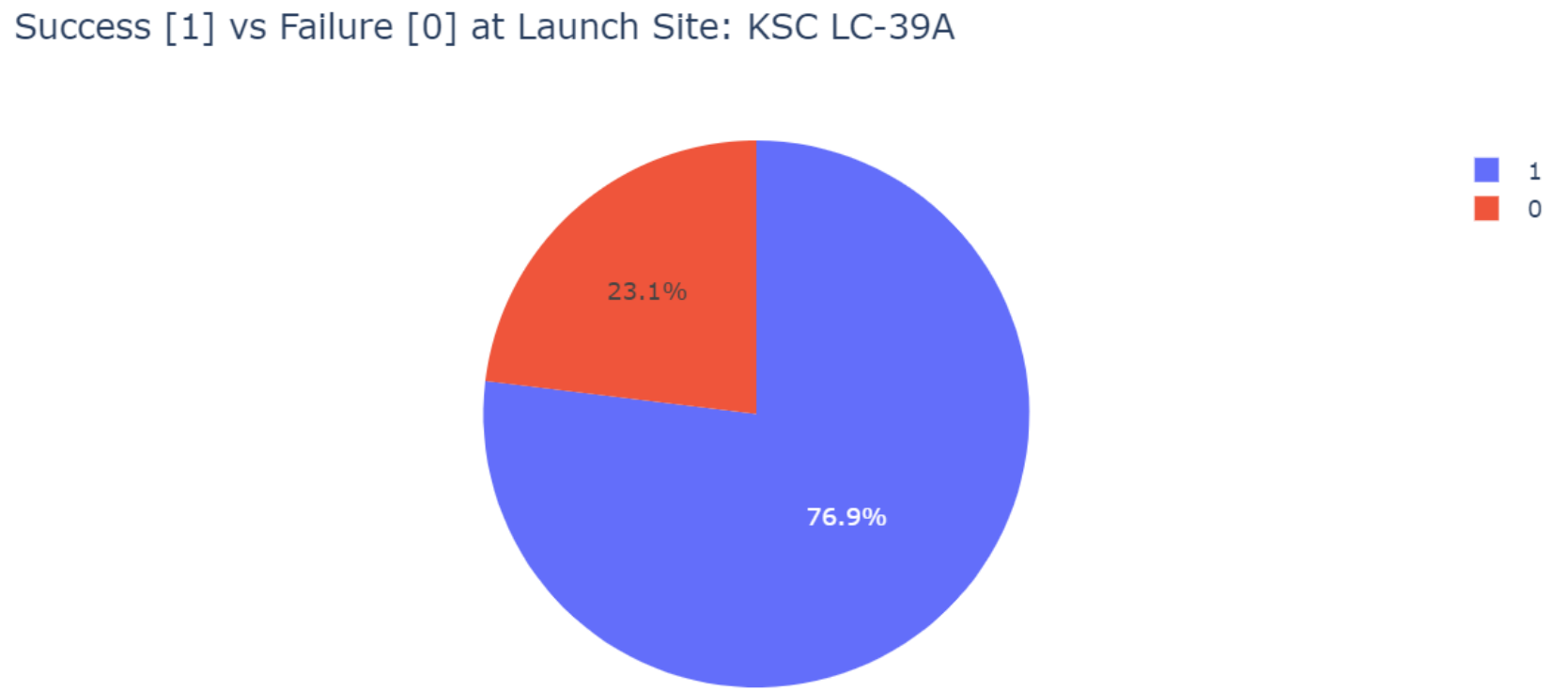
Pie chart showing the success percentage achieved by each launch site



Findings:

- Launch site KSC LC-39A had the highest percentage of successful launches
- Launch site CCAFS SLC-40 had the lowest percentage of successful launches *(i.e. highest no. of failed launches)*

Site KSC LC-39A – Launch Success Ratio



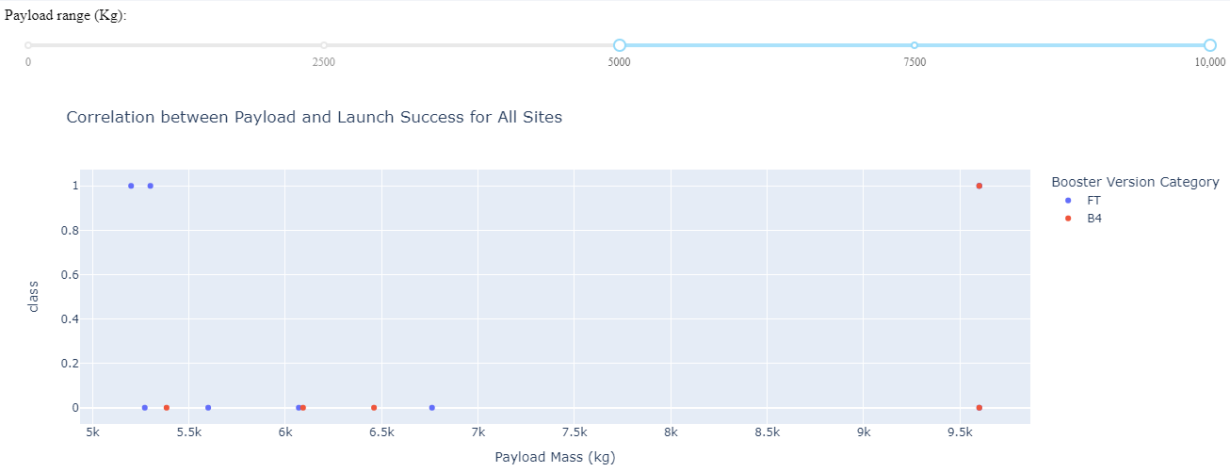
KSC LC-39A achieved:

- Success rate of 76.9%
- Failure rate of 23.1%

Payload vs Launch Outcome for All Sites

Payload range: 0KG-5000KG

Payload range: 5000KG-10,000KG



Findings:

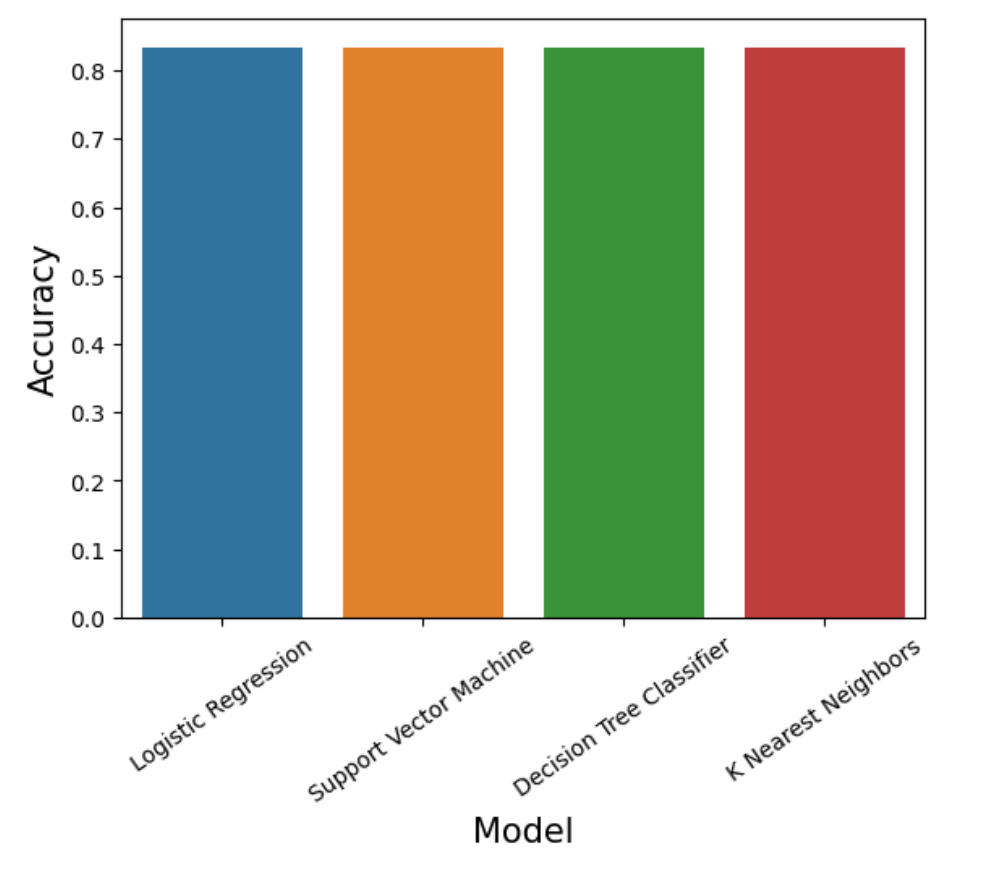
- Success rate for lower weighted payloads is higher than the heavier weighted payloads
- Only Booster Versions FT and B4 had a payload of over 5000KG
- Booster version FT has the highest success rate in pay load range 2000-5000KG

Section 5

Predictive Analysis (Classification)

Classification Accuracy

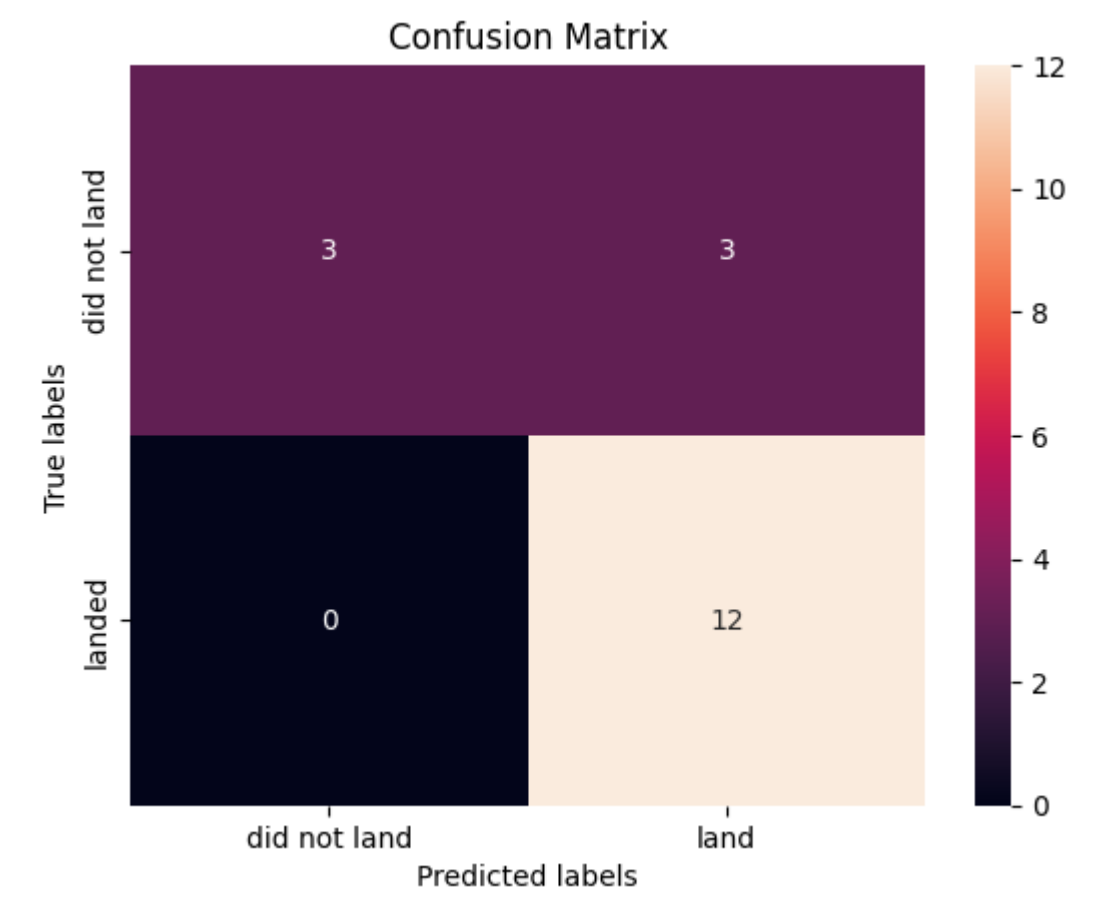
Bar chart showing the built model accuracy for all built classification models



Findings:

- All models have the same classification accuracy

Confusion Matrix



- All models have the same confusion matrix.
- Whilst all the models can distinguish between the different classes, the major problem is false positives.

Conclusions

EDA with SQL and Visualization

- Orbit ES-L1 SSO, HEO and GEO have the best success rate
- We can observe that with heavy payloads, the successful landing rate are more for PO, LEO and ISS Orbit types
- We can observe that the success rate since 2013 kept increasing until 2020
- In total, there are 12 booster versions which have carried the maximum payload mass of 15600KG

Folium Map & Dashboard

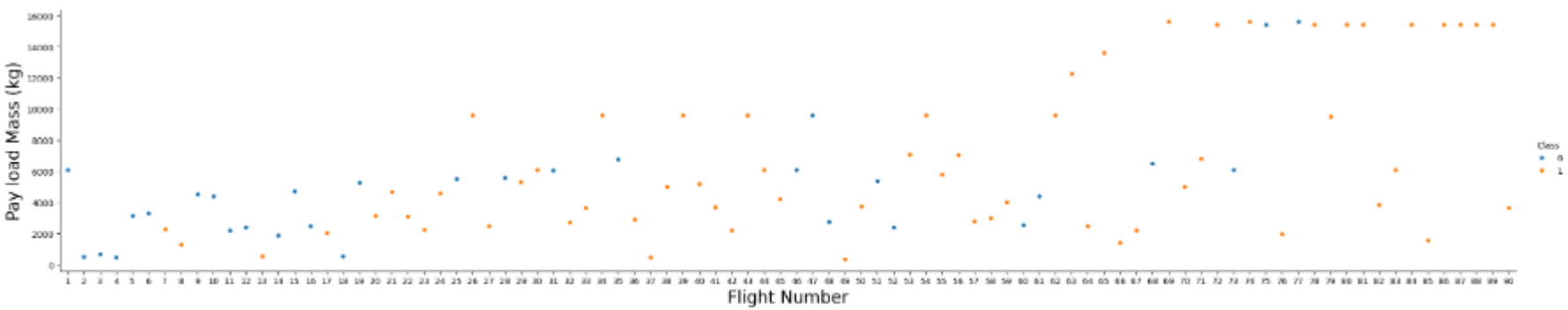
- All launch sites are in very close proximity to the coast
- Launch site KSC LC-39A has the highest percentage of successful launches
- Success rate for lower weighted payloads is higher than the heavier weighted payloads
- Booster version FT has the highest success rate in pay load range 2000-5000KG

Predictive Analysis

- Logistic Regression, SVM, KNN and Tree classifier models all have an accuracy score of 0.83
- The above 4 models all have the same the confusion matrix. Whilst the models can distinguish between the different classes, the major problem is false positives.

Appendix

- NB: On slide 11, the following scatter plot: **to see how the Flight Number (indicating the continuous launch attempts) and Payload variables would affect the launch outcome** has not been included in main body of presentation (as per template). Here is the missing seventh scatter plot and insight:



We see that different launch sites have different success rates. `CCAFS LC-40`, has a success rate of 60 %, while `KSC LC-39A` and `VAFB SLC 4E` has a success rate of 77%.

- GitHub URL for all Notebook outputs and data sets can be found here: https://github.com/S-Keddy/Data-Science-Capstone-Project

Thank you!