IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

IHENACHO UGOCHI K.

17TH NOVERMBER 2023

Outline

Executive Summary

Introduction

Methodology

Results

Conclusion

Appendix

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- EDA with data visualization
- EDA with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

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

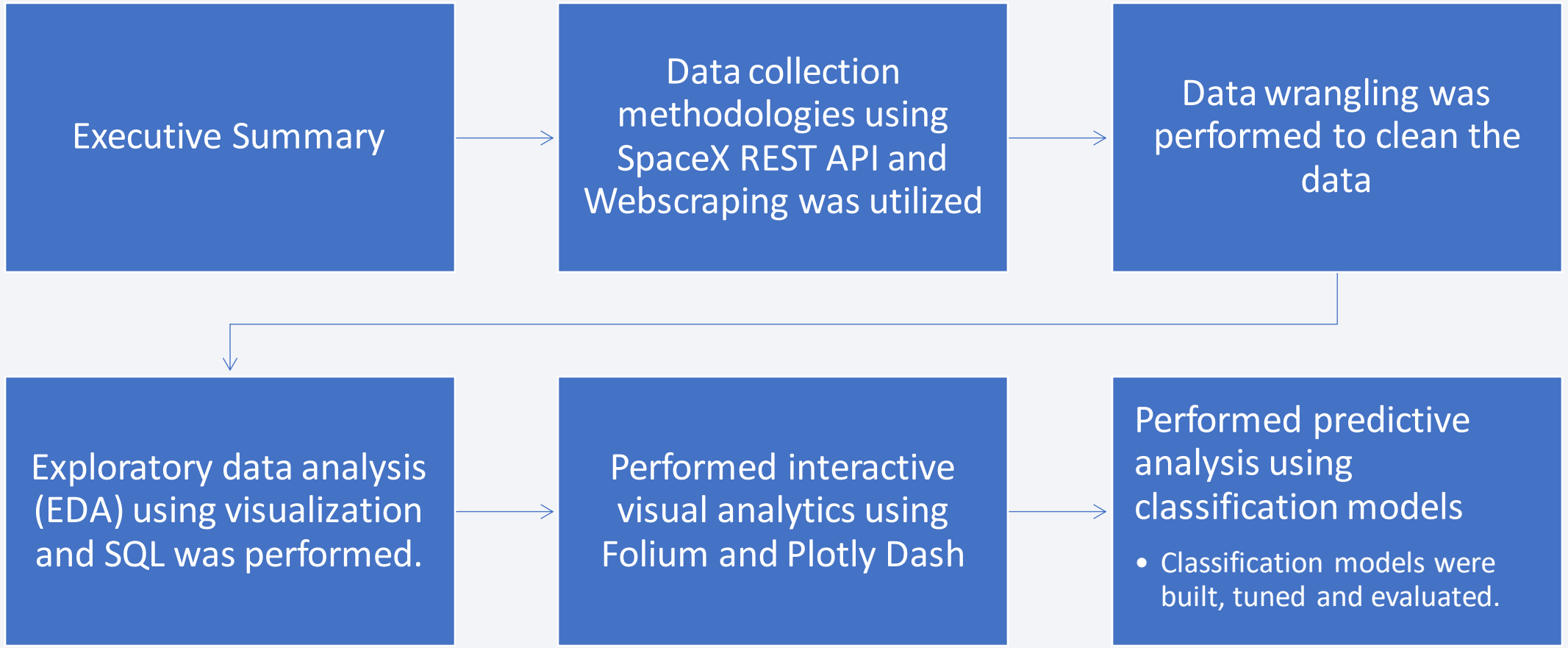
Introduction

- SpaceX is perhaps the most successful company that is into the manufacture of reusable rockets.
- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars. They have much of the savings is because they can reuse the first stage.
- This presentation will cover the prediction if the first stage will land.
- By determining this, the cost of a launch can be determined thereafter.
- Past data records from SpaceX will be utilized for this data analysis and prediction.
- With known features such as - rockets from different sites with different payloads and different orbit types, flight numbers we can determine the landing and mission outcome of the launch.

Section 1

Methodology

Methodology



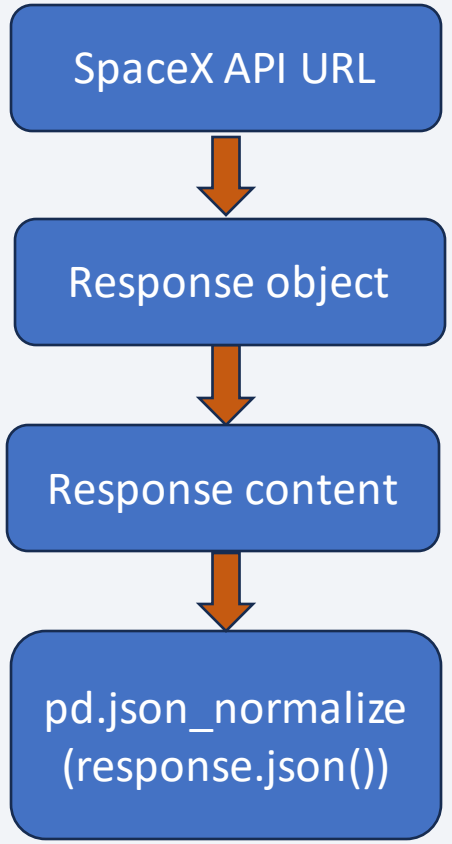
Data Collection

- The first data was collected using the SpaceX REST API.
- Information based on past launches was gathered from this API.
- The required columns: rockets used, payloads delivered in mass(kg), Cores, launch specifications, landing specifications and landing outcome was extracted.
- A get request was performed on the url to return a response object.
- The content of the response object was viewed using response.content.
- The response content was decoded using .json() which was further converted to a dataframe using pd.json_normalize(response.json()).
- The data was then wrangled and missing values were handled being replaced with the column mean.

Data Collection – SpaceX REST API

This is simple process flow summary using process using key phrases and flowcharts

This is the [SpaceX Data Collection API GitHub URL]



requests.get(SpaceX API URL) is used to get a response object.

response.content is used to get a response content.

Transform content to JSON Extract data by converting to a pandas dataframe.

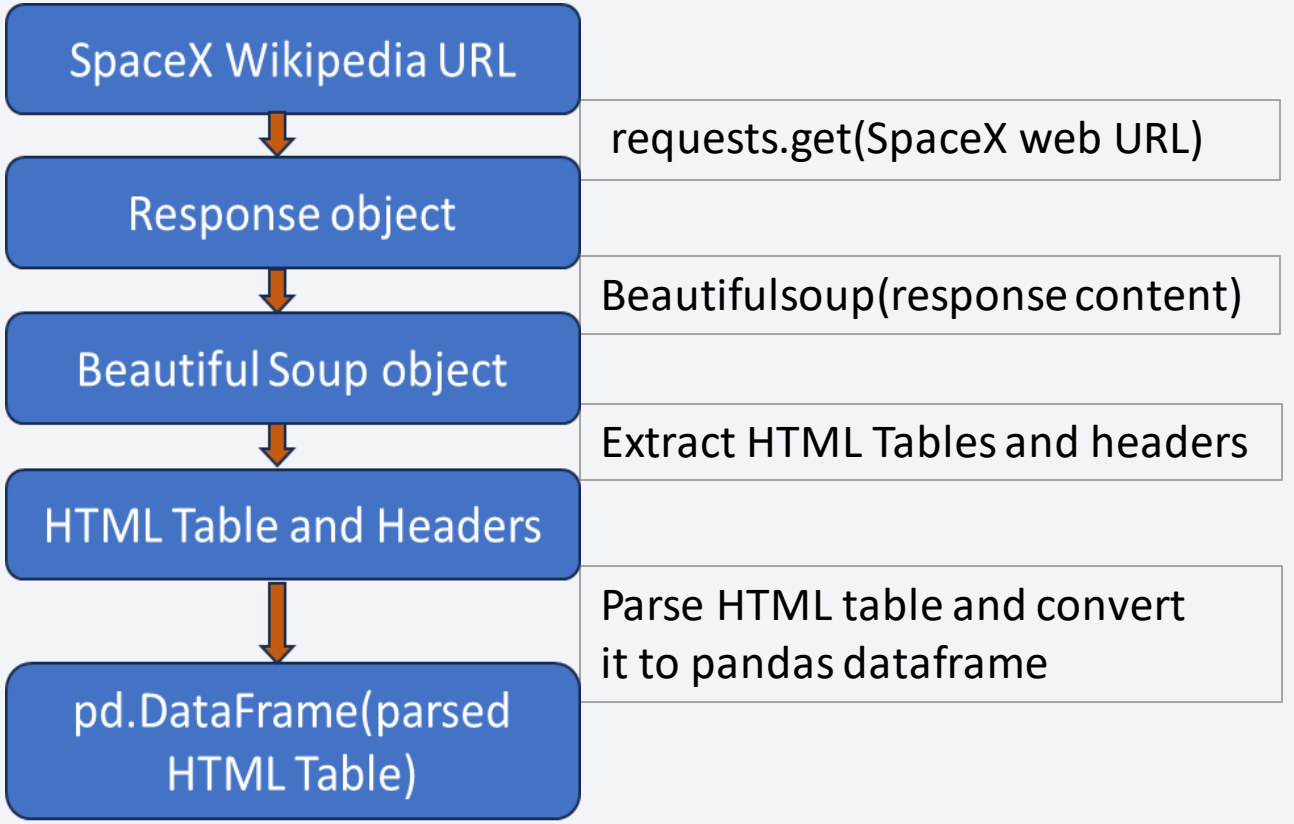
Data Collection – Web scraping

- The second data was gathered from 9 historical launch records from a Wikipedia page titled "List of Falcon 9 and Falcon Heavy launches"
- The method requests.get() was used to get a HTML response object called data,
- Python BeautifulSoup() object was created from the HTML response.
- All the tables on the page was extracted using soup.find_all() and assigning it to the HTML table header.
- The required table was extracted from the table header using list index.
- The table columns and variable names from the required table was extracted.
- The data was then parsed from these tables and converted to pandas dataframe for further visualization and analysis.

Data Collection – Web Scraping

This is simple process flow summary using process using key phrases and flowcharts.

This is the SpaceX Data Collection via Webscraping GitHub link



Data Wrangling

- The dataset was filtered to remove Falcon 1 launches leaving only the required Falcon 9.
- Null values (NaN) in the payload mass column were replaced with average values of all the payload masses.
- The number of launches on each site, the number and occurrence of each orbit was calculated using value counts.
- The number and occurrence of mission outcome of the orbits was also calculated.
- The landing outcome label of 0 and 1 where 0 is failure and 1 is success of landing was created and assigned to the column Class.
- Landing success rate was determined from the class variable as 0.67.
- This is the SpaceX Data Wrangling GitHub link

EDA with Data Visualization

- Here the dataset was analyzed to find trends, patterns and relationships that will result in a successful landing outcome.
- Relationships between key features was visualized. Some features includes but not limited to, Flight number vs Launch site, Payload mass vs Launch site using scatterplots and so on.
- Launch success yearly trends was visualized.
- Variables useful for further analysis was created using one hot encoding from specific features: orbits, launch site, landing pad and serial
- All numeric variables were casted to float64 data types.
- This is the SpaceX EDA with Data viualization GitHub link

EDA with SQL

Data was analyzed utilizing functions such as distinct, min, like, substr, group by, etc:

- The names of the unique launch sited in the space mission was found. There are 4 unique launch sites.
- 5 records where launch sites begin with the string 'CCA' were displayed using the "Distinct" function.
- The total payload mass carried by boosters launched by NASA (CRS) calculate to be 45596
- The average payload mass carried by booster version F9 v1.1 was calculated to be 2928.4
- For the landing outcomes, the date when launch was first achieved in ground pad was found as 2015-12-22.
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 was listed.
- There was a total number of 100 successful mission and1 failed mission.
- The names of the booster versions which have carried the maximum payload mass were found.
- The months in 2015 where there was failure in drone ship was found to be in October and April..
- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 was analyzed with 10 being the max and 1 the min.

This is the SpaceX EDA with SQL GitHub link

Build an Interactive Map with Folium

- The coordinates (latitude and longitude) of launch sites, coordinates of closest coastline, marker color, circle object was added to a map object that was created using Folium map.
- The coordinates of launch sites was extracted to be able to estimate the distance of the various launch sites location to the coasts, equator line and highly populated areas.
- The coordinate of the closest coastline was picked to estimate the distance of launch sites to the coast and to give an understanding of the reason behind the close proximities of launch sites to coastlines and far proximity towards densely populated areas.
- The circle object was used to label the location of the 4 unique launch sites.
- The marker color was added to distinguish between each vessels landing outcome, where red represents failed outcome and green represents successful outcome.
- This is the SpaceX Visualization with Folium GitHub link

Build a Dashboard with Plotly Dash

- Pie charts, scatter plots and range sliders were utilized in this task.
- A pie chart rendered by the first callback function - used to show the total percentage success rates for each launch sites. A site dropdown function to navigate the chart to view each sites mission's percentage success and failure outcomes was added.
- A scatter plot rendered by a callback function - added to show the amount of successful and failed landing outcomes of the payload mass with respect to the booster versions category.
- A range slider for payload with a min value of 0 and a max value of 10000 was also added to visualize the relationship between the payload and the mission outcome.
- Plots and interactions was added to enable users to perform interactive visual analytics on the SpaceX launch data in real-time.
- This is the SpaceX Interactive Dasboard with Plotly GitHub link

Predictive Analysis (Classification)

- The required features were assigned to the variables, X, and Y which was standardized thereafter.
- The train test split function was then on these new transformed datasets to get training sets and testing dets of data for or data modelling and prediction.
- Each of the models: logistic regression, KNN, SVM and Decision Tree models were uniquely built using the GridSearchCV object.
- A scoring parameter of "accuracy", a cv of 10, the model object and their parameters were then passed to the GridSeachCV object.
- The accuracy score was determined using the method .score() on the test data.
- The prediction was then carried out on the test data to be able to plot a confusion matrix. This was plotted thereafter.
- The best model was selected based on the comparison of the accuracy scores of all the models.
- This is the SpaceX Model Building and Prediction GitHub Link

Results

1

Exploratory data analysis results

- There are 4 unique sites from which launches took place.
- Although many of the launches was recorded to be from site CCAFS SLC 40, site KSC LC-39A had the highest mission success rate of 41.7% and CCAFS SLC 40 had the lowest success rate of 12.5%

2

Interactive analytics demo in screenshots

- Launch sites are usually situated at close proximities to the coastlines.

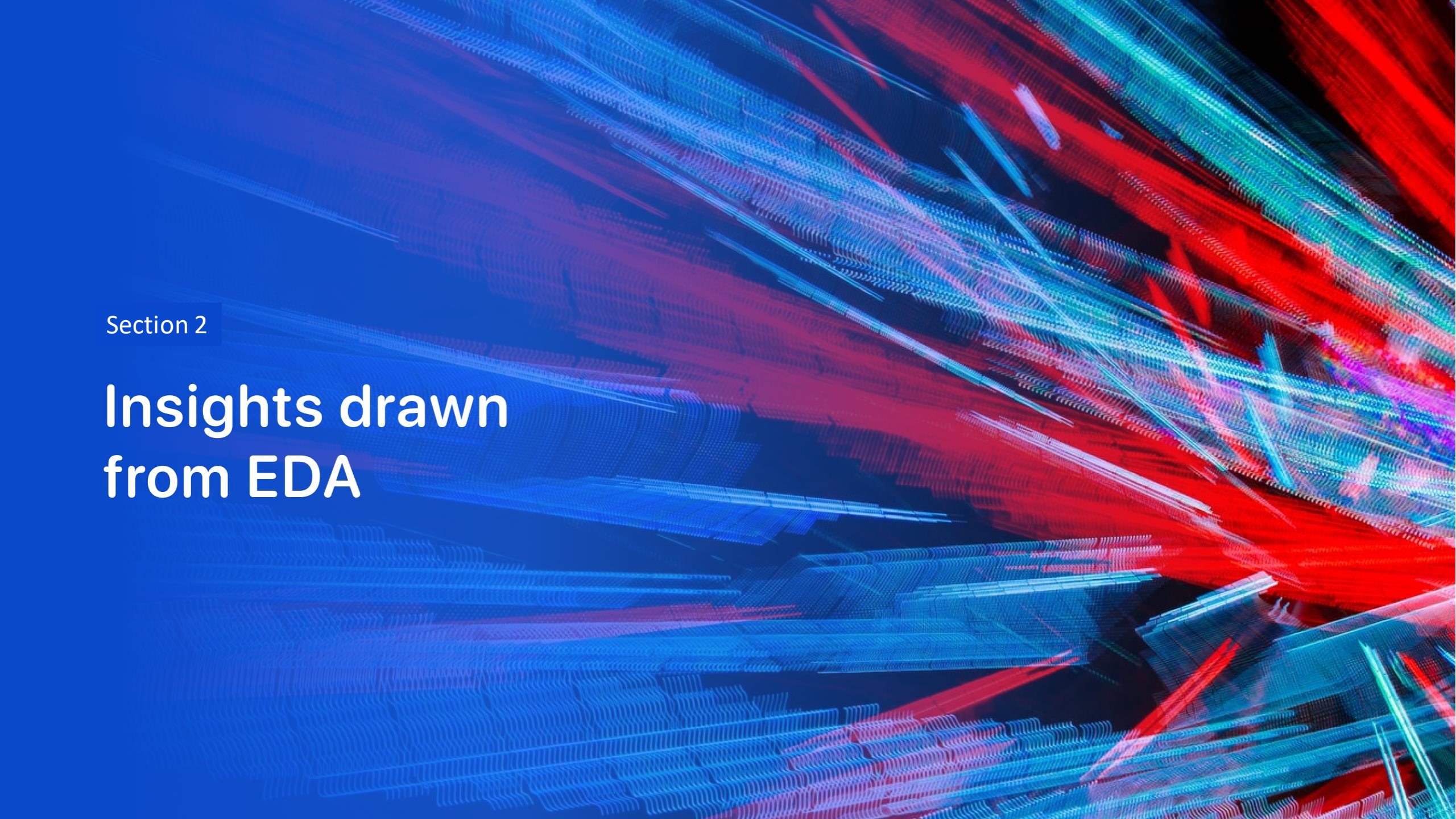
3

Predictive analysis results

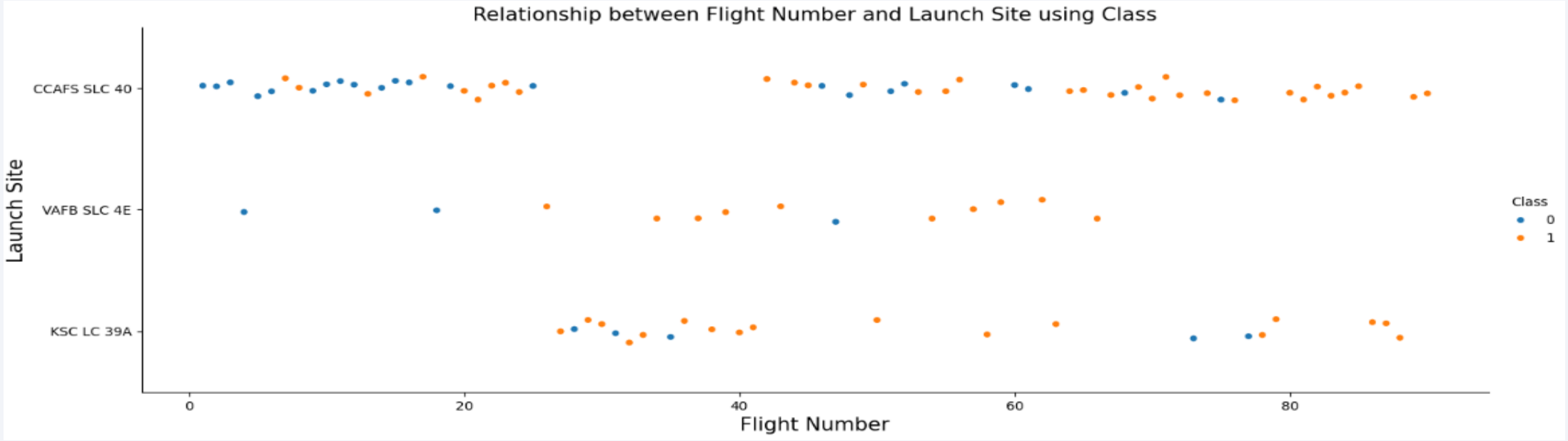
- All models gave the same accuracy score. Meaning the accuracy score of 0.834 was the best score

Section 2

Insights drawn from EDA

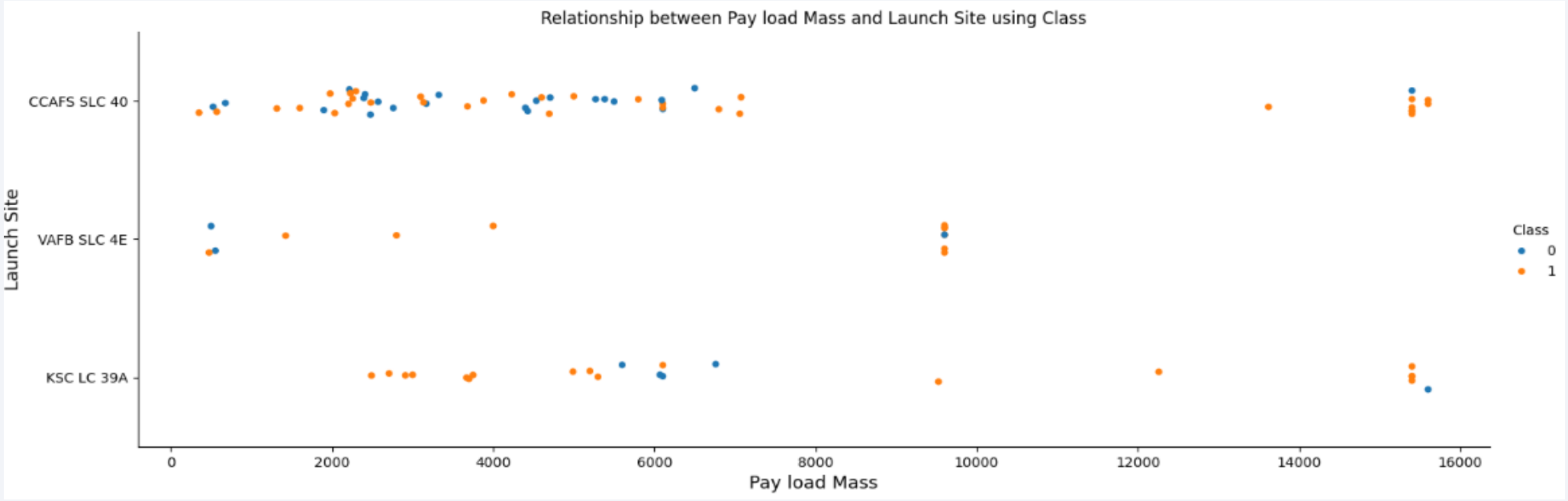


Flight Number vs. Launch Site



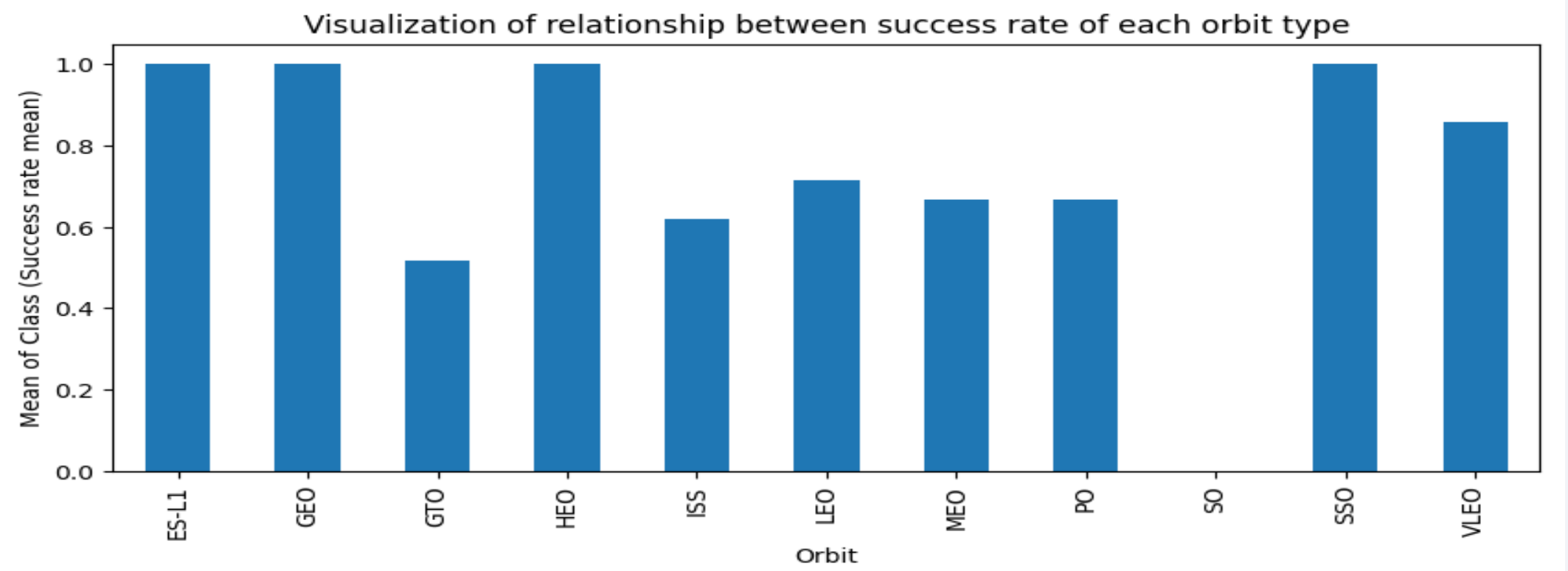
- Most of the missions with higher flight numbers that successfully landed are from site CCAFS SLC 40.
- It was observed that KSC LC 39A had no records for flight number below 27.
- VAFB SLC 4E had no records of flight number after 66 and CCAFS SLC 40 had no record between flight numbers 26 to 41.
- A majority of the launches with landing outcome occurred from site CCAFS SLC 40.

Payload vs. Launch Site



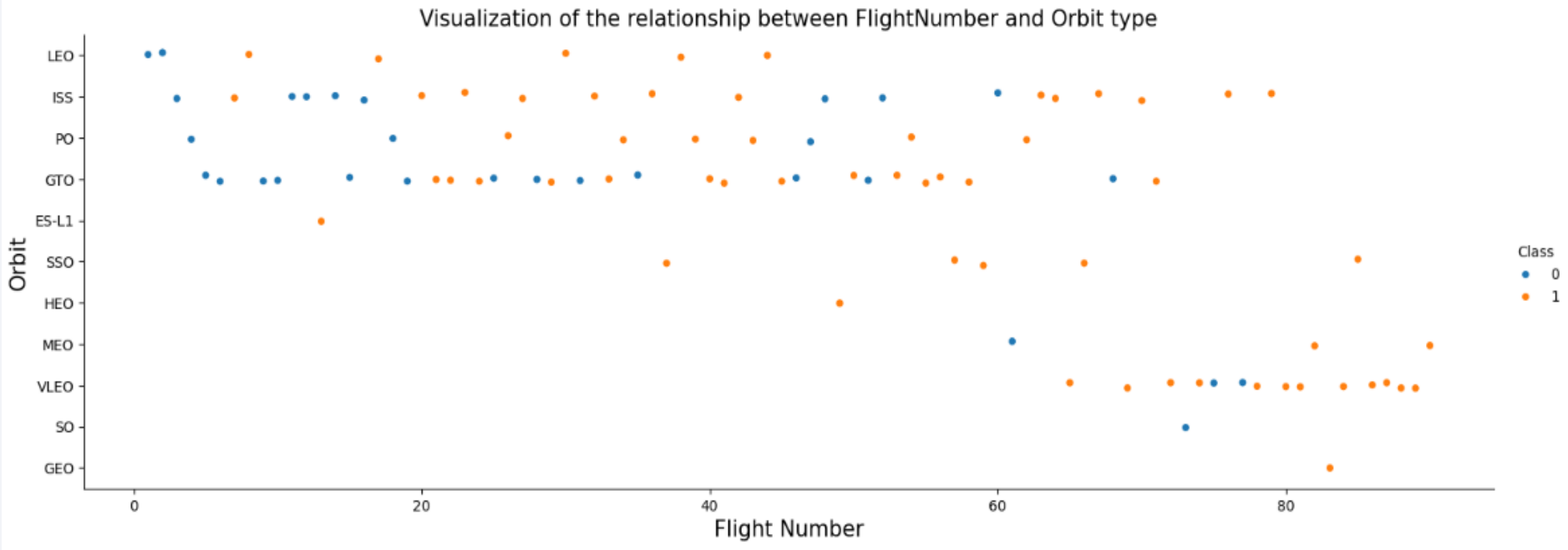
- Majority of the vessels from all sites had a payload mass range of 500 to 7000.
- For the VAFB-SLC launch site there are no records of rockets launched nor their outcome with heavy payload mass (greater than 10000).
- Majority of rockets launched with payload mass greater than 10000, are from site CCAFS SLC 40 and had a higher success rate than launches from site KSC LC 39A.

Success Rate vs. Orbit Type



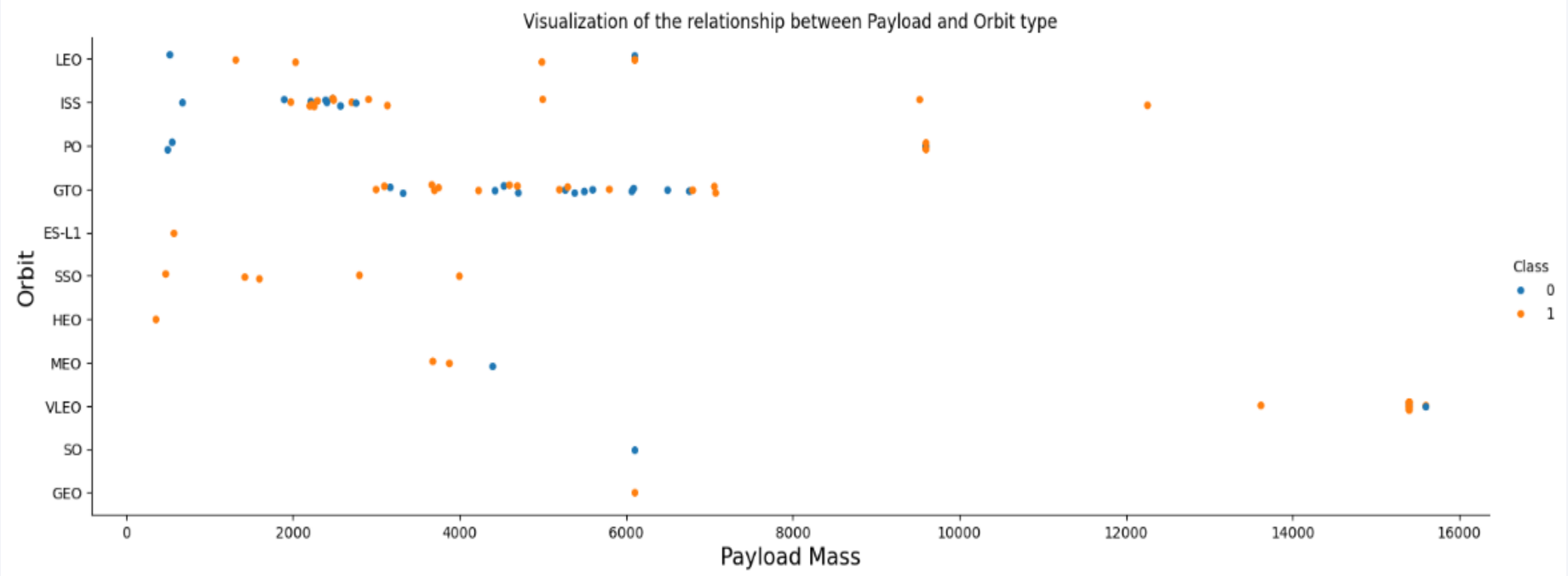
- It was observed that the orbits type with the highest average success rate of 1 are; ES-L1, GEO, HEO and SSO.
- SO has the lowest average success rate.

Flight Number vs. Orbit Type



- It was observed that in the LEO orbit, the success appears related to the number of flights.
- The orbit with the greatest number of success for higher flight numbers (from 80 and greater) is VLEO.
- GTO had the highest amount of launches. But its success rate is not related to the increase or decrease in flight number.

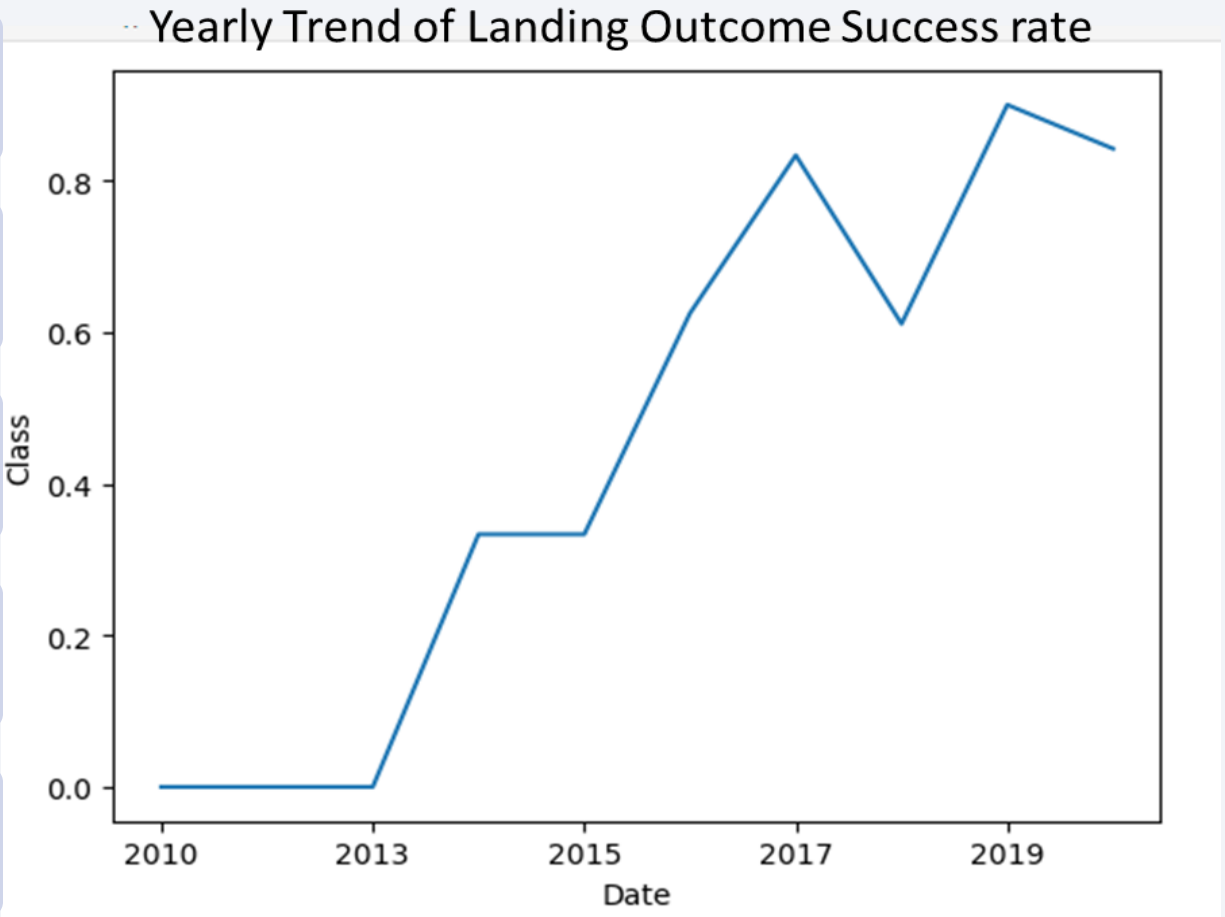
Payload vs. Orbit Type



- It was observed that orbit GTO and ISS had the most flights with different payloads.
- ISS and PO success rate seemed to have a relationship with increased payload mass.
- SSO is a low payload mass orbit.
- SSO had all successful outcomes.

Launch Success Yearly Trend

- An increase in landing outcome success rate was first recorded in 2013.
- The second increase was recorded in 2015 after which it experienced a decline.
- The first decline in success rate was recorded in 2017.
- The second decline was experienced in 2019.
- From 2010 to 2019 the success rate has generally increased.



All Launch Site Names

- The distinct function was used to retrieve the unique launch sites.
- There are 4 unique launch sites and can be seen on the screenshot on the right.

```
%%sql
SELECT DISTINCT Launch_Site FROM SPACEXTABLE
```

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

Launch Site Names Begin with 'CCA'

- The query is shown below.
- The result is shown on the right
- The wildcard "LIKE" used with the % sign, is used to filter records containing the word in the string.

```
%%sql
SELECT * FROM SPACEXTABLE
WHERE Launch_Site LIKE "CCA%"
LIMIT 5
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- The LIMIT function is used to filter the amount of records to be retrieved.
- It was observed that the 5 records that were displayed are from site CCAFS LC 40.

Total Payload Mass

- The query utilizes the "WHERE" clause for filtering

```
%%sql
SELECT SUM("PAYLOAD_MASS__KG_") AS "Total Payload Mass For NASA (CRS)"
FROM SPACEXTABLE
WHERE Customer == "NASA (CRS)"
```

- The total payload mass the total payload mass carried by boosters launched by NASA (CRS) is 45596kg

Total Payload Mass For NASA (CRS)
45596

Average Payload Mass by F9 v1.1

- The query utilizes the "WHERE" clause for filtering and AVG to find the mean.

```
%%sql
SELECT AVG("PAYLOAD_MASS__KG_") AS "Average Payload Mass by Booster_Version F9 v1.1"
FROM SPACEXTABLE
WHERE Booster_Version == "F9 v1.1"
```

- The average payload mass carried by booster version F9 v1.1 is 2928.4kg.

Average Payload Mass by Booster_Version F9 v1.1
2928.4

First Successful Ground Landing Date

- The query utilizes the "WHERE" clause for filtering and MIN to find the minimum record.

```
%%sql
SELECT MIN(Date) as "First Successful Landing Date", "Time (UTC)", "Landing_Outcome"
FROM SPACEXTABLE
WHERE "Landing_Outcome" == "Success (ground pad)"
```

- The date when the first successful landing outcome in ground pad was achieved is 22nd December 2015.

First Successful Landing Date	Time (UTC)	Landing_Outcome
2015-12-22	01:29:00	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are seen below on the right.
- It is observed that out of these 4, F9 FT B1021.2 had the highest payload of 5300.
- And F9 FT B1026 had the lowest of 4600 out of all 4 booster versions.

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

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)

Total Number of Successful and Failure Mission Outcomes

- The query makes use of the sum of cases when a particular condition is met.

```
%%sql
SELECT
SUM(CASE WHEN Mission_Outcome LIKE '%Success%' THEN 1 ELSE 0 END) AS "Total Successful Mission",
SUM(CASE WHEN Mission_Outcome LIKE '%Failure%' THEN 1 ELSE 0 END) AS "Total Failed Mission"
FROM SPACEXTABLE;
```

- It was observed that there was a total of 100 successful missions
- And there was a total of 1 failed mission.

Total Successful Mission	Total Failed Mission
100	1

Boosters Carried Maximum Payload

- This query makes use of the subquery and the "GROUP BY" function.
- The names of the booster which have carried the maximum payload mass are displayed below.
- There are 12 unique booster versions.

```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_
== (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE )
GROUP BY Booster_Version
```

Query Results

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

2015 Launch Records

- This query makes use case and substr() to extract the date with a given format.
- The 2015 failed landing outcomes in drone ship occurred in October and April.
- They both occurred at the launch site CCAFS LC 40.
- The 2 booster versions are seen below.

Month_Name	Month_Number	Landing_Outcome	Booster_Version	Launch_Site
October	10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Query Code

```
%%sql
SELECT
    CASE SUBSTR(Date, 6, 2)
        WHEN '01' THEN 'January'
        WHEN '02' THEN 'February'
        WHEN '03' THEN 'March'
        WHEN '04' THEN 'April'
        WHEN '05' THEN 'May'
        WHEN '06' THEN 'June'
        WHEN '07' THEN 'July'
        WHEN '08' THEN 'August'
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
    END as Month_Name,
    SUBSTR(Date, 6, 2) as Month_Number,
    Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Failure (drone ship)'
AND SUBSTR(Date, 1, 4) = '2015';
```

Boosters Carried Maximum Payload

- This query makes use of the subquery and the "GROUP BY" function.
- The names of the booster which have carried the maximum payload mass are displayed below.
- There are 12 unique booster versions.

```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_
== (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE )
GROUP BY Booster_Version
```

Query Results

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

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

- The query is seen below:

```
%%sql
SELECT Date, Landing_Outcome, COUNT(Landing_Outcome)
AS "Total Outcome"
FROM SPACEXTABLE
WHERE DATE(SUBSTR(Date, 1, 4) || '-'
           || SUBSTR(Date, 6, 2) || '-'
           || SUBSTR(Date, 9, 2))
BETWEEN DATE('2010-06-04') AND DATE('2017-03-20')
GROUP BY Landing_Outcome
ORDER BY "Total Outcome" DESC;
```

- The highest outcome of 10 was seen to occur on 22nd May 2015 where no attempt was made.
- The are 8 unique lading outcomes for the stated period.

- Query results

Date	Landing_Outcome	Total Outcome
2012-05-22	No attempt	10
2015-12-22	Success (ground pad)	5
2016-08-04	Success (drone ship)	5
2015-10-01	Failure (drone ship)	5
2014-04-18	Controlled (ocean)	3
2013-09-29	Uncontrolled (ocean)	2
2015-06-28	Precluded (drone ship)	1
2010-08-12	Failure (parachute)	1

- Between this period, the first outcome which was on 12-08-2010 was a failure.


Section 3
Launch Sites
Proximities Analysis

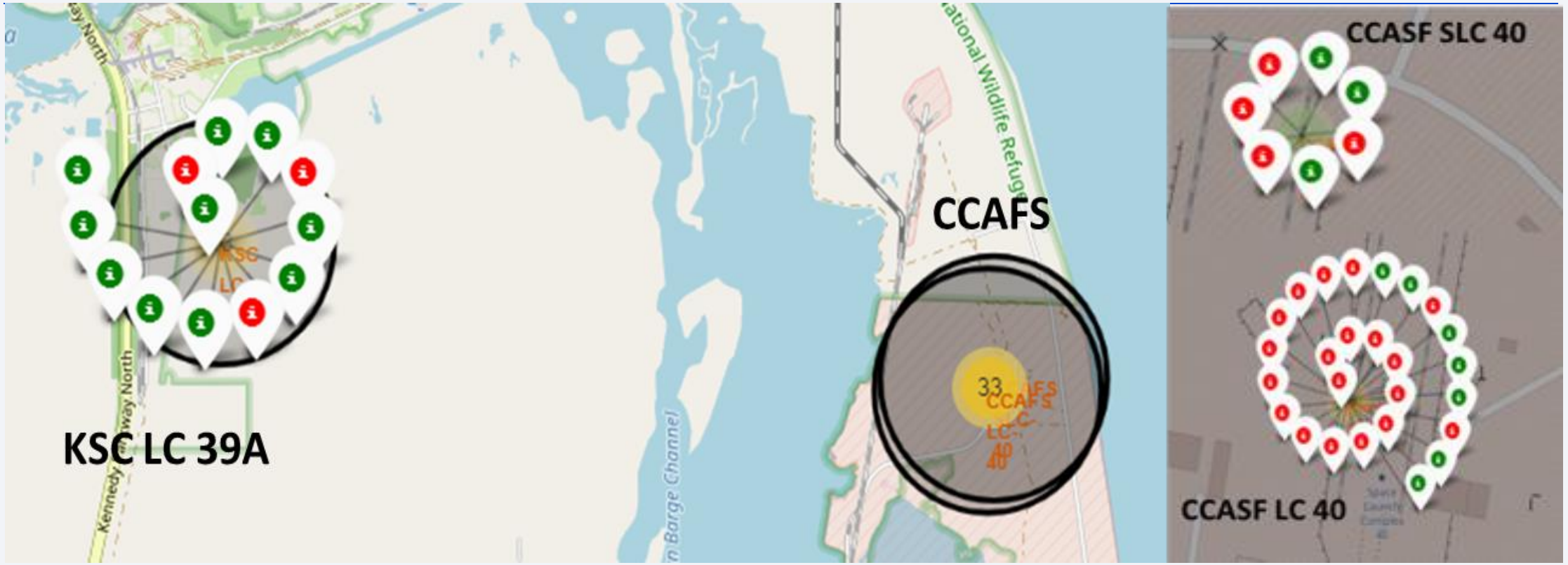
Folium Map – Launch Sites Location Markers

- All 4 launch sites are located on the southern part of USA.
- 3 launch sites are located southeast region and only VAFB SLC 4E is located at the southwest region.
- All launch sites are located at very close proximities to the coast (or coastline).

Folium Map – Launch Outcomes for Southwest site

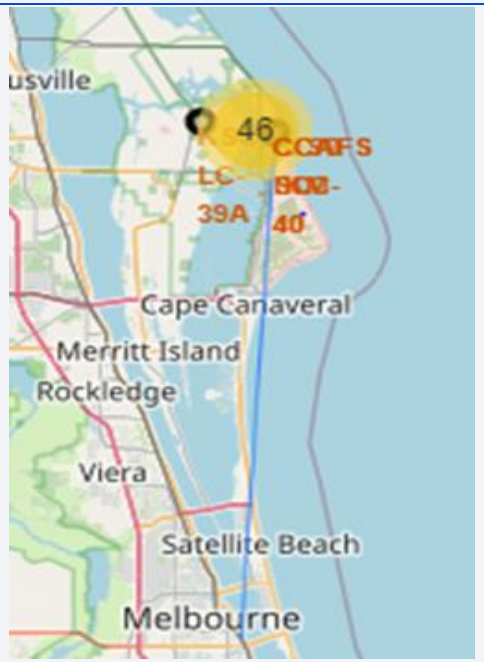
- There was a total of 56 landing outcomes for southeast and southwest regions, with site KSC LC 39A having the highest success rate out of the 4 sites.
- The southwest site region – VAFB SLC 4E had a total of 10 landing outcomes.
- There was 4 successful landing outcomes and 6 failed landing outcomes.

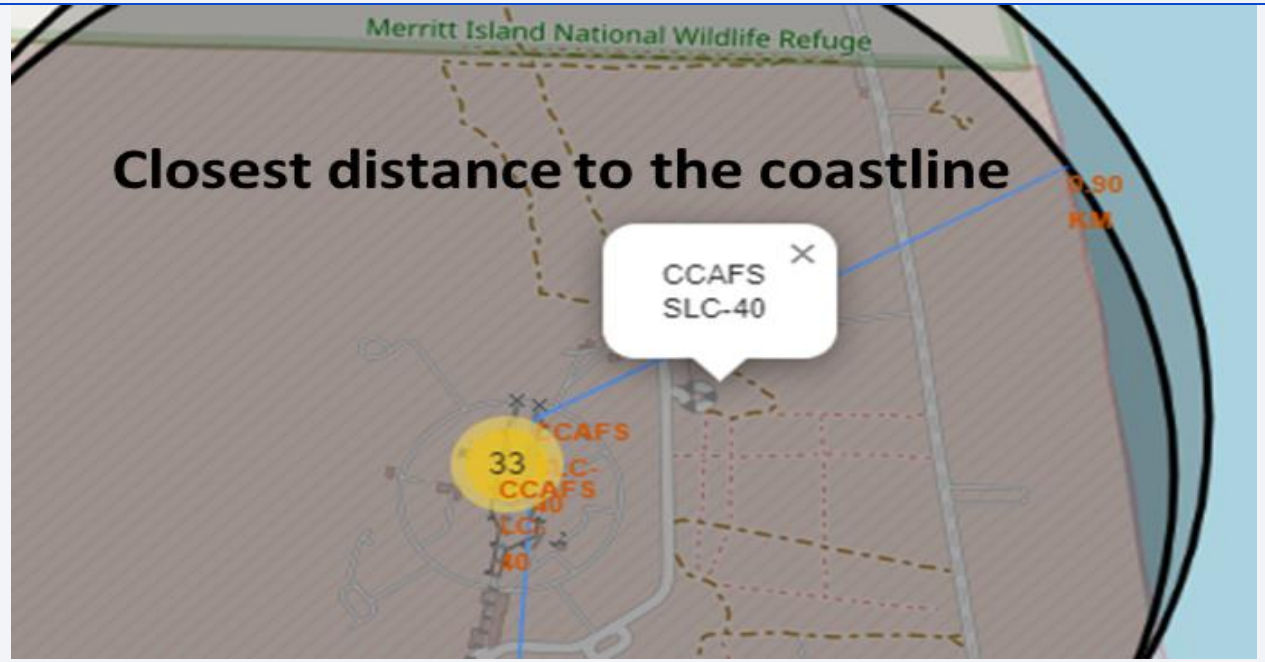
Folium Map – Launch Outcomes for Southeast sites



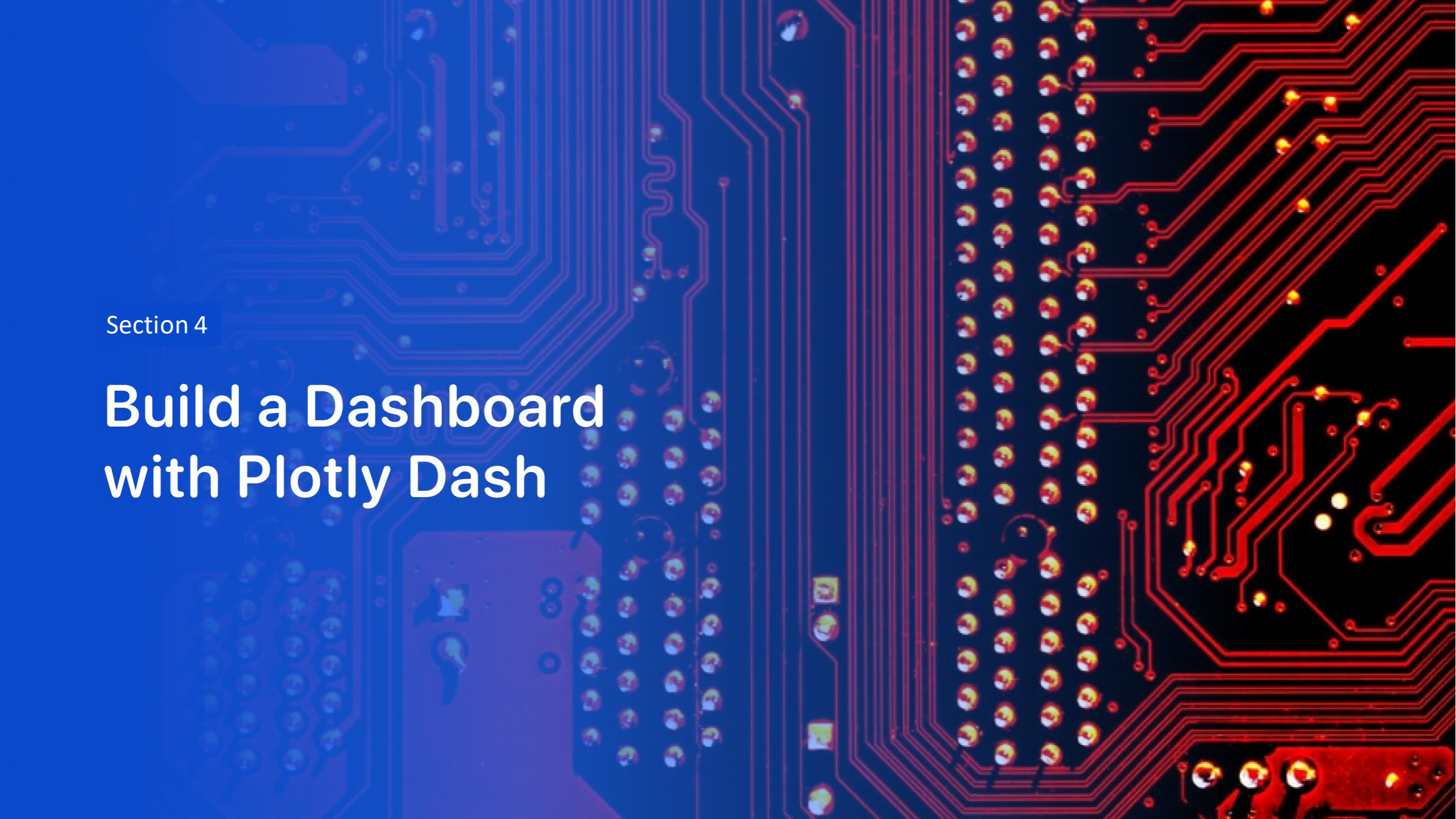
- The site KSC LC 39A had a total of 13 outcomes with 10 successes and 3 failures.
- The CCAFS SLC 40 site had a total of 7 outcomes with 3 successes and 4 failures.
- Site CCAFS LC 40 had a total of 26 outcomes with 7 successes and 19 failures

Proximities from launch sites to coast and cities

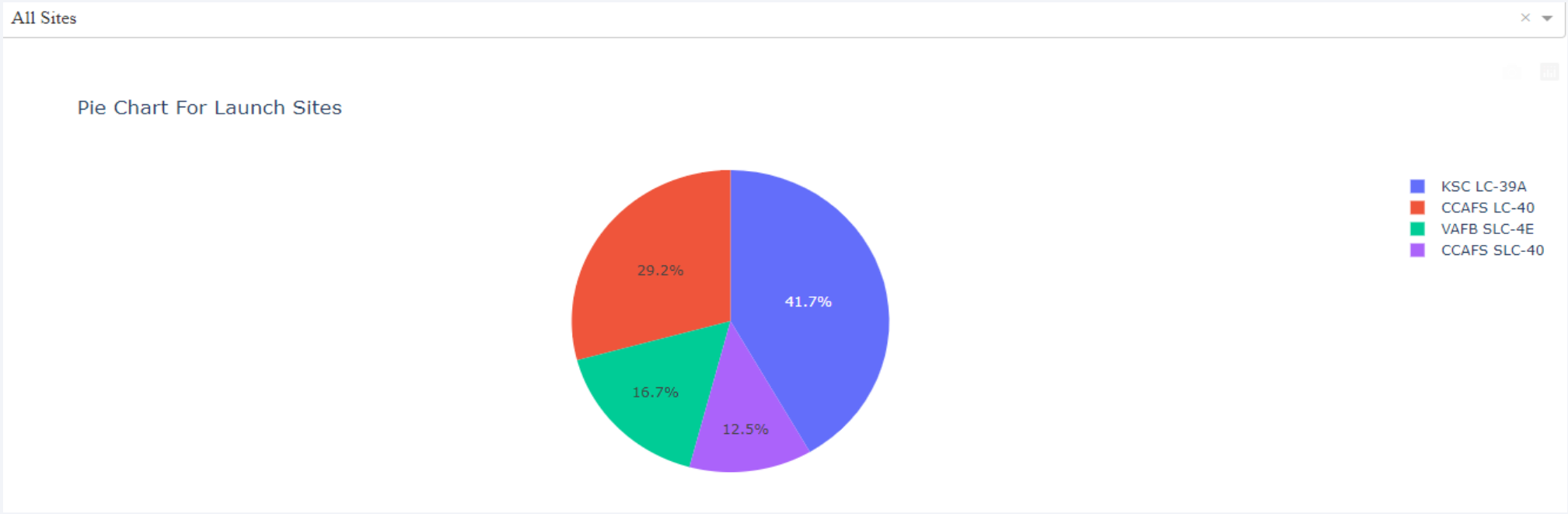




- Site with the closest distance to southeast coastline was CCAFS SLC 40 with 0.9km
- The city closest to the CCAFS SLC 40 site is Melbourne with a distance of approximately 53km
- Launch sites are situated far away from densely populated areas. But are in close proximity to the coastline because the vessel would have the advantage of flying over the ocean, minimizing the risk of having any particles exploding near humans.

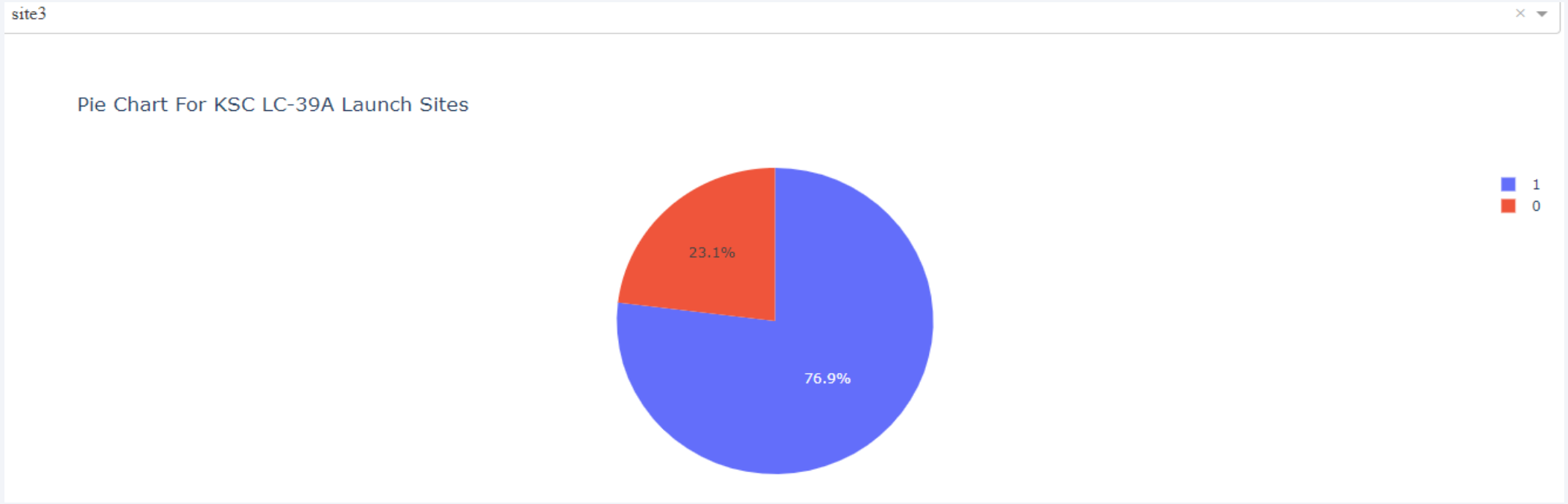

Section 4
Build a Dashboard with Plotly Dash

Interactive Visualization with Plotly - Total Success Rate



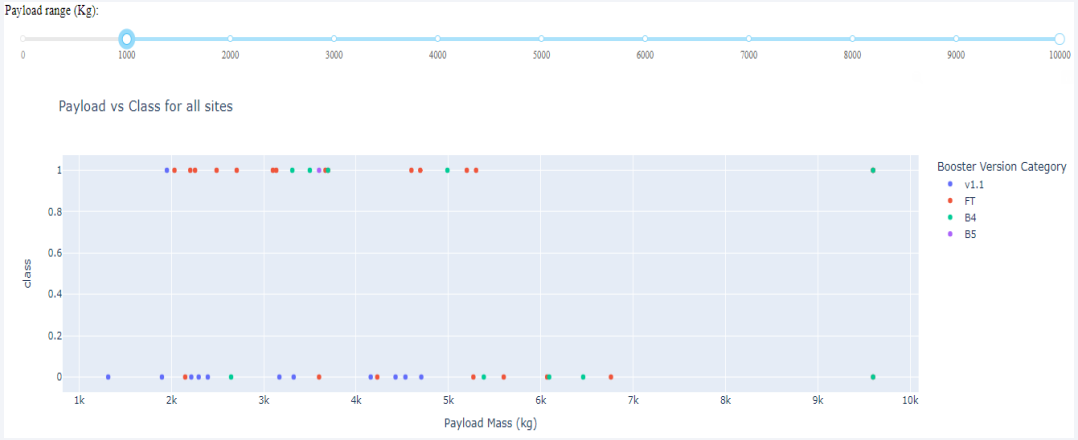
- This shows the total success count for all the sites with each site uniquely identified by a color on the color key.
- It was observed that KSC LC-39A had the highest mission success rate of 41.7% and CCAFS SLC 40 had the lowest success rate of 12.5%

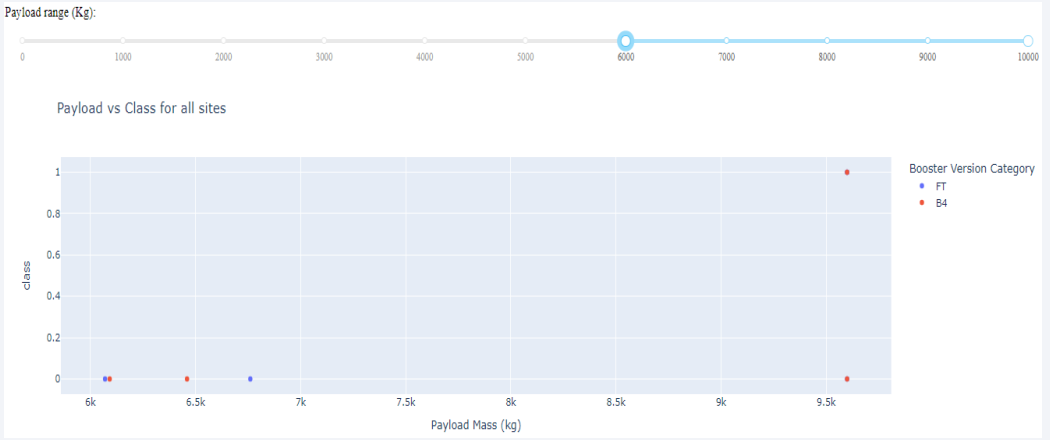
Mission Outcome Distribution for KSC LC-39A site



- In the color key, 1 represents successful mission while 0 represents failed mission.
- This site had a mission success rate of 76.9% while the remaining 23.1% represents the failed mission.

Payload and Class relationship using Range Slider





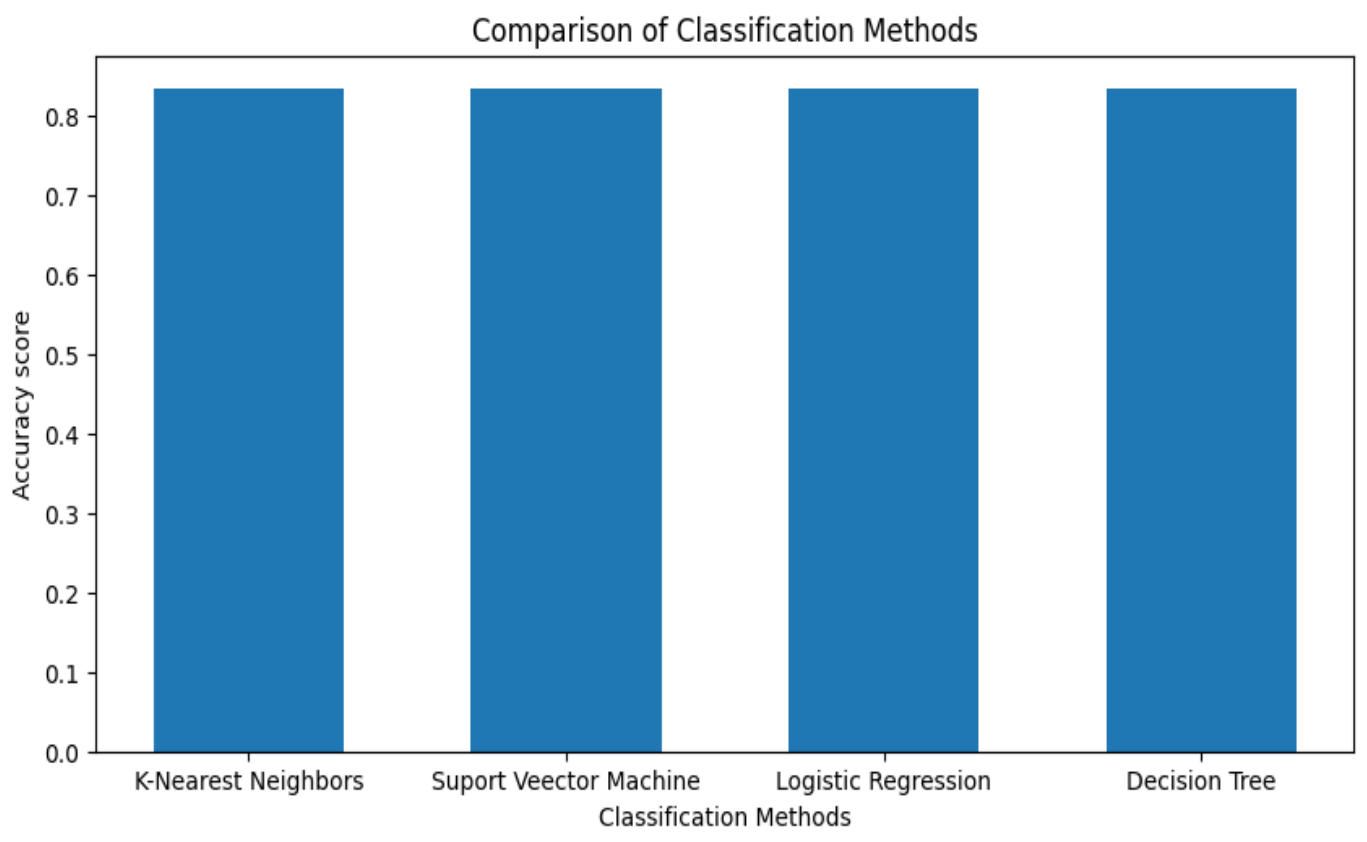
- The figure on the left is for 1kg (1000) payload mass while the figure on the right is for 6kg (6000) of payload mass.
- There was no outcome for booster v1.0 at a payload range of 1000. And only booster FT and B4 were present utilizing a payload range of 6000
- Booster version B5 had only 1 mission outcome which was a success.
- Booster FT had the highest mission success rate while v1.1 had the lowest for both payload range instances.

Section 5

Predictive Analysis (Classification)

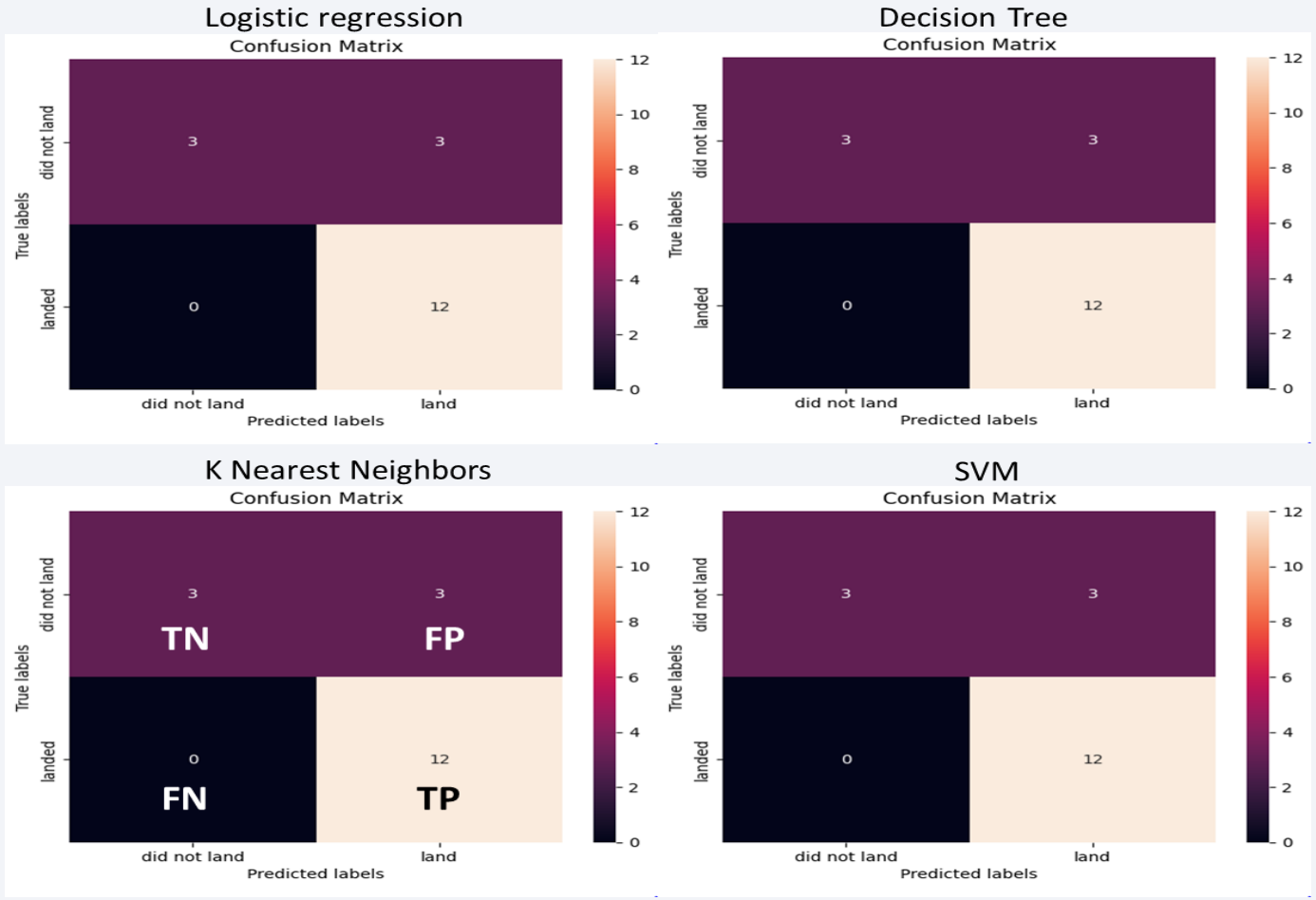
Classification Accuracy

- It was observed that practically all these algorithms gave the same accuracy score as seen on the graph on the right.
- The accuracy score was 0.8334 (83.34%) for each of the models.



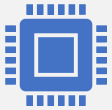
Confusion Matrix

- From the merged diagram on the right, practically all these algorithms gave the same result.
- This is the best result that can be gotten because the model accurately predicted all 12 of the test samples to have landed which landed.
- This is the True positive.
- The model predicted accurately 3 samples out of 6 to not land.



Conclusions

Features such as; launch sites, orbit type, booster versions and payload mass can influence the landing outcome.

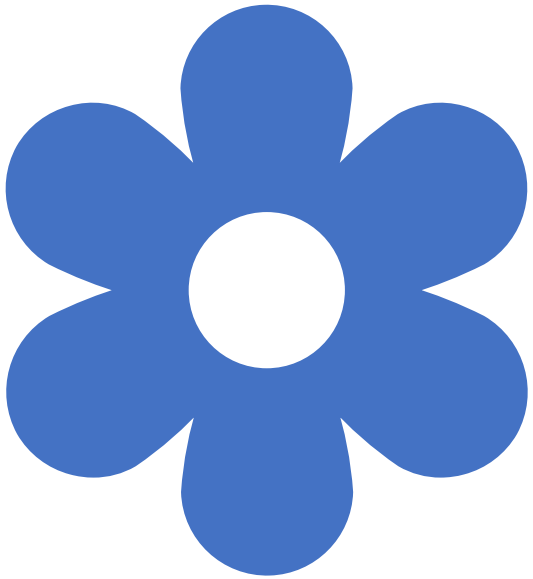


The increase in landing outcome success over the years clearly proves that technology is constantly improving. With data science and advancement of engineering and technology there is a high possibility of increased success rate in landing outcome over the next coming years.

The accuracy score of 83.34% for all the models shows a high percentage of prediction in the outcome of the mission.

This project is therefore a viable project and hence the cost estimation of launches can be proceed.

Appendix



- To ensure the same sequence of random numbers in the DecisionTreeClassifier model was produced, numpy.random.seed(0) was used because the value for the hyperparameter "splitter" of the model was ['best', 'random'].
- After the train test split, the data produced 18 test samples. Our predicted was tested on these 18 samples.

- URL for data collection API - SpaceX API URL
- URL for data collection with Web scraping - SpaceX Wikipedia URL
- URL to full project on GitHub - SpaceX Capstone Project

Thank you!