IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vilinie Singh
25/02/2025

Outline

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

Executive Summary

- Summary of methodologies
  - ✓ Data was collected from the SpaceX Api and the SpaceX Wikipedia's page
  - ✓ Data Wrangling was done (categorical features were modified using one-hot encoding)
  - ✓ EDA were carried out using SQL, Data Visualization methods and Folium (maps, dashboards, graphs, charts)

- Summary of all results
  - ✓ Machine learning predictions were made by producing the following four: K Nearest Neighbours, Support Vector Machines, Decision Tree Classifier and Logisitic Regression; all of which produced similar results of 83 percent, all of which predicted successful landings.

Introduction

- Project background and context

As seen on the Space X website the Falcon 9 rocket launches with a cost of 62 million dollars. However, there exists others with costs of 165 million dollars. The drastic levels in cost is a result of Space X being able to recover and reuse parts of the rocket within Stage 1.

The purpose of this project is to determine whether the Stage 1 will launch successfully; this is determined through the means of a machine learning pipeline.

- Problems you want to find answers
  - ✓ What properties will help to determine a successful launch?
  - ✓ What conditions are ultimately necessary to ensure a successful launch?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected through the means of the SpaceX API and web scraping from the Wikipedia website.
- Perform data wrangling
  - Data was processed by applying one-hot encoding to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Models were tuned using GridSearchCV

Data Collection

- Data was collected using the API from the public SpaceX API and the web scraping data from the SpaceX Wikipedia page.

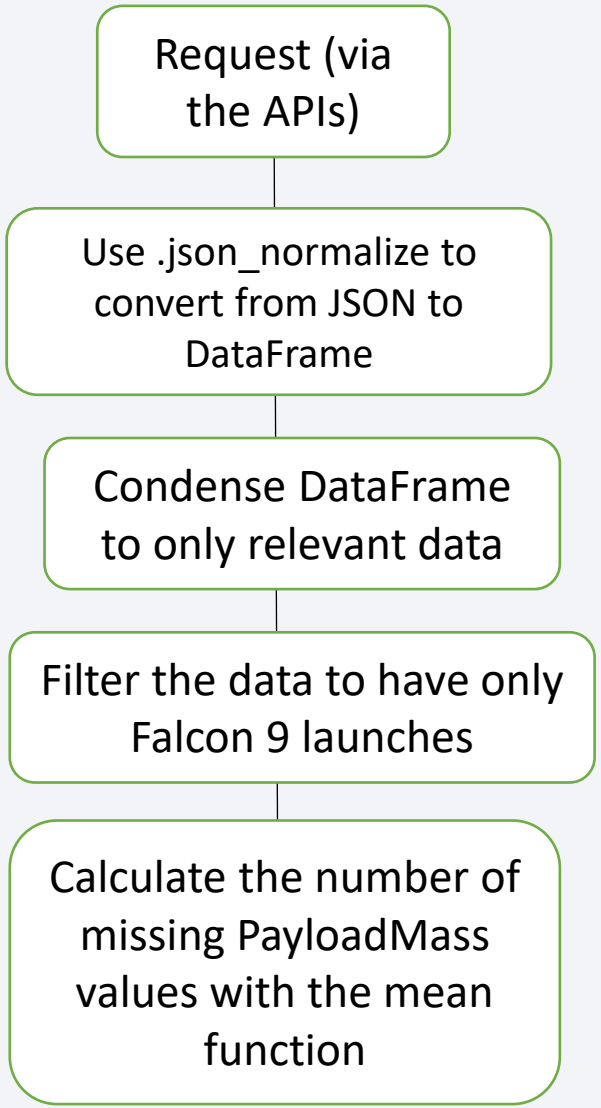
  ✓ The .json() function call was used to decode the response content information from the API and then was converted to a dataframe by using the Pandas library and the .json_normalize() function call.

  ✓Furthermore, the BeautifulSoup library was used to carry out the web scraping on the Wikipedia page. The launch records were collect and converted into a dataframe using the Pandas library for further analysis.

Data Collection – SpaceX API

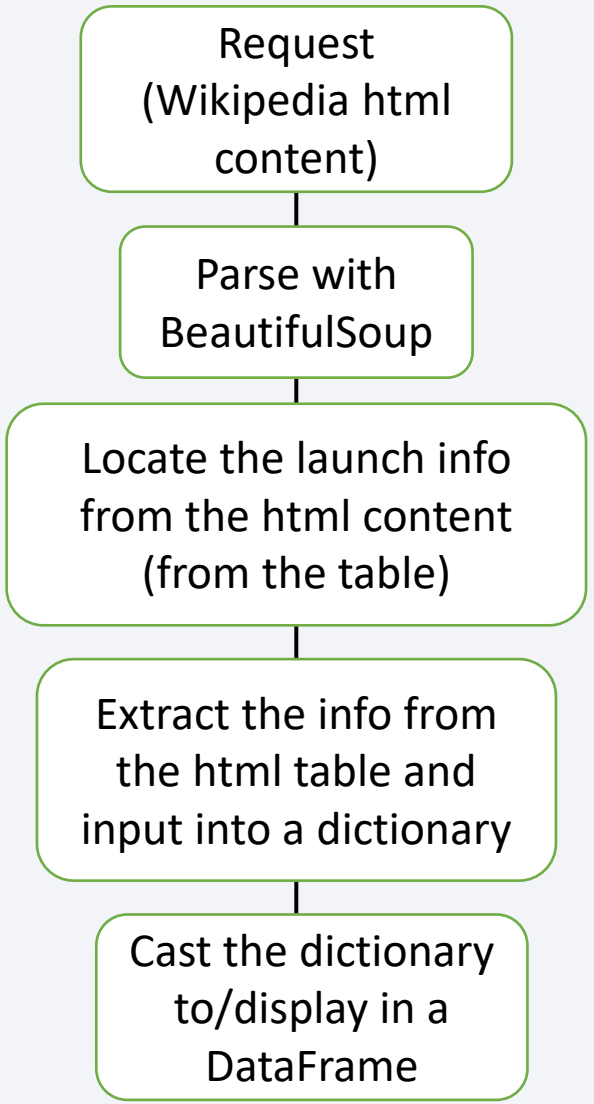
- The GET request method was used to collect data from the SpaceX API.

- Github Link: https://github.com/veeweenee/IBM-DS-Capstone-Course/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



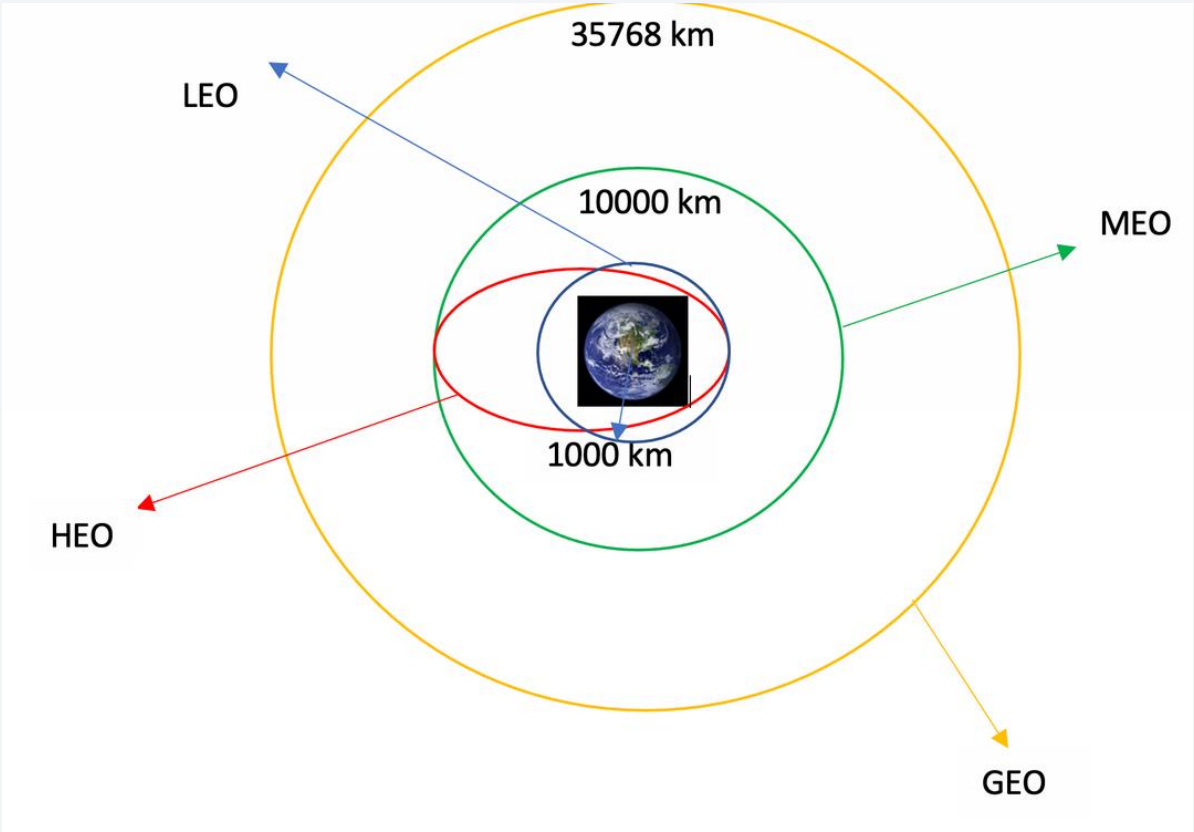
Data Collection - Scraping

- Data was scraped from the Wikipedia page by using the BeautifulSoup library.
- Findings of the launch information were then casted to a DataFrame for further exploration.
- GitHub URL: https://github.com/veeweene/IBM-DS-Capstone-Course/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- An exploratory data analysis was done to determine the training labels where the landing outcomes were successful (1) or a failure (0)
- Launches at each site, sightings at all orbits were calculated.
- Github URL: https://github.com/veeweenee/IBM-DS-Capstone-Course/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Several plots were used to compare the different relationships between the variables. The purpose of this is to test to see viable relationships could be used to train the machine. This type of analysis were performed using the following variables:
    1. FlightNumber
    2. PayloadMass
    3. LaunchSite
    4. Class
    5. Orbit
    6. Year
- Github URL: https://github.com/veeweenee/IBM-DS-Capstone-Course/blob/main/edadataviz.ipynb

EDA with SQL

- An exploratory data analysis was carried out but by using SQL.
- Mainly, queries were made to fully understand the data within the dataset. Queries were done based on landing outcomes, booster versions, payload sizes, etc.
- Github URL: https://github.com/veeweenee/IBM-DS-Capstone-Course/blob/main/jupyter-labs-eda-sql-coursera_sqllite(1).ipynb

Build an Interactive Map with Folium

- An interactive map was made using the Folium library to mark landings (successful and failed ones), launch sites and specific locations (such as railways, cities).

- This was done to fully grasp why launch sites have been based within a specific location.

- Github URL: https://github.com/veeweenee/IBM-DS-Capstone-Course/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

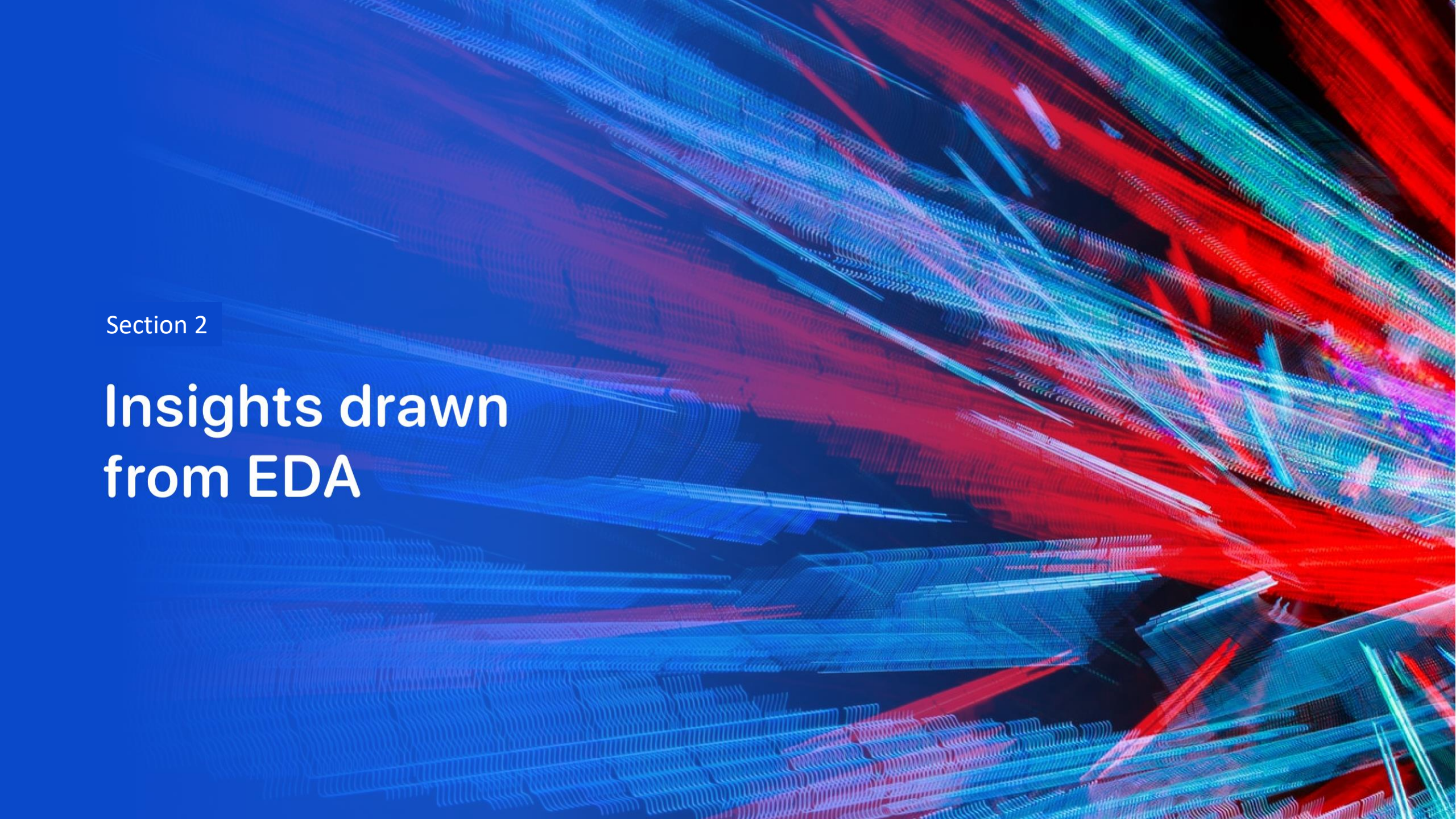
- Added to the dashboard are plots such as a scatterplot and pie chart.
- The scatterplot was added to display the variation when it comes to success rates in relation to booster versions, payload masses, etc.
- The pie chart was added as a display of the various success rates of launch sites.
- GitHub URL:https://github.com/veeweenee/IBM-DS-Capstone-Course/blob/main/DV0101EN-Final-Assign-Part-2-Questionss.py

Predictive Analysis (Classification)

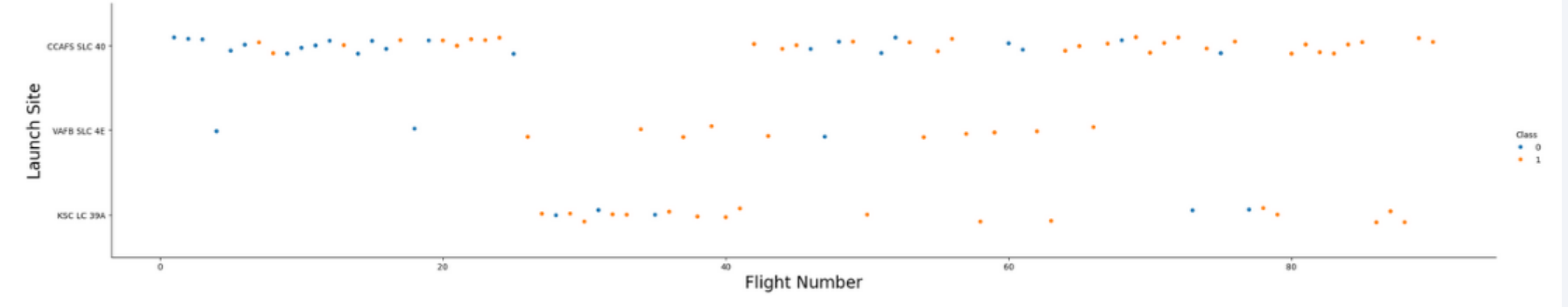
- Data was split into two sets (training and testing)
- Various machine models were built using GridSearchCV (Logistic Regression, Support Vector Machine, Decision Tree Classifier and K Nearest Neighbours)
- Accuracy was used as a metric for the models and the best performing model was found.
- Github URL: https://github.com/veeweenee/IBM-DS-Capstone-Course/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

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

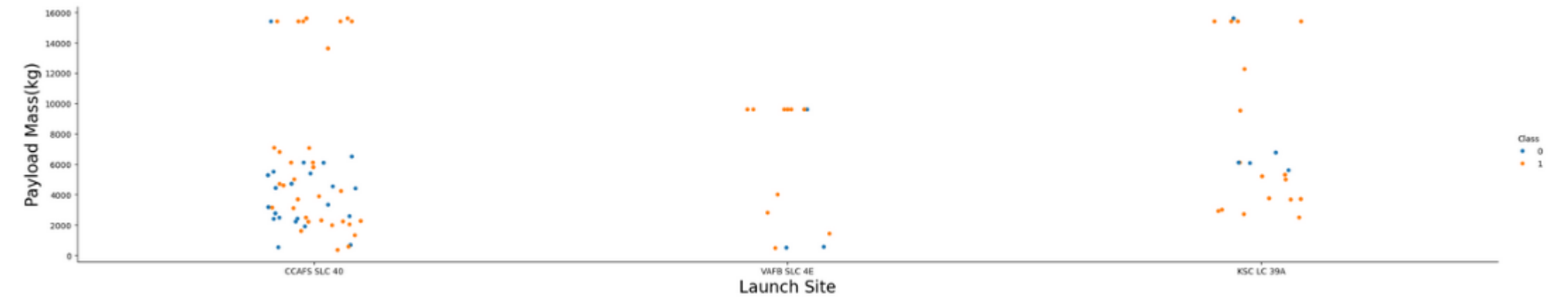

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- It can be seen from the plot above that the number of flights present at a launch site, the higher is the success rate.

Payload vs. Launch Site

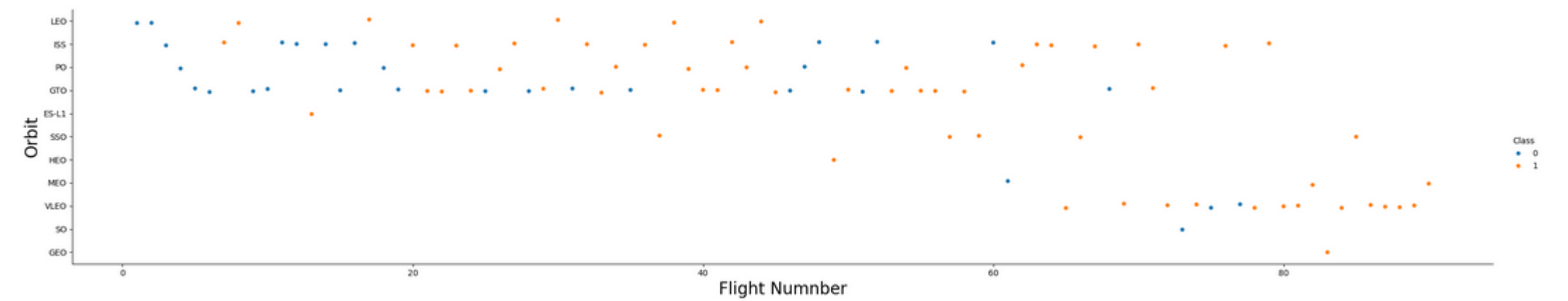


- It can be seen from the plot above that the payload mass is at the CCAFS SLC 40 launch site, the higher is the success rate for the Falcon 9 rocket

Success Rate vs. Orbit Type

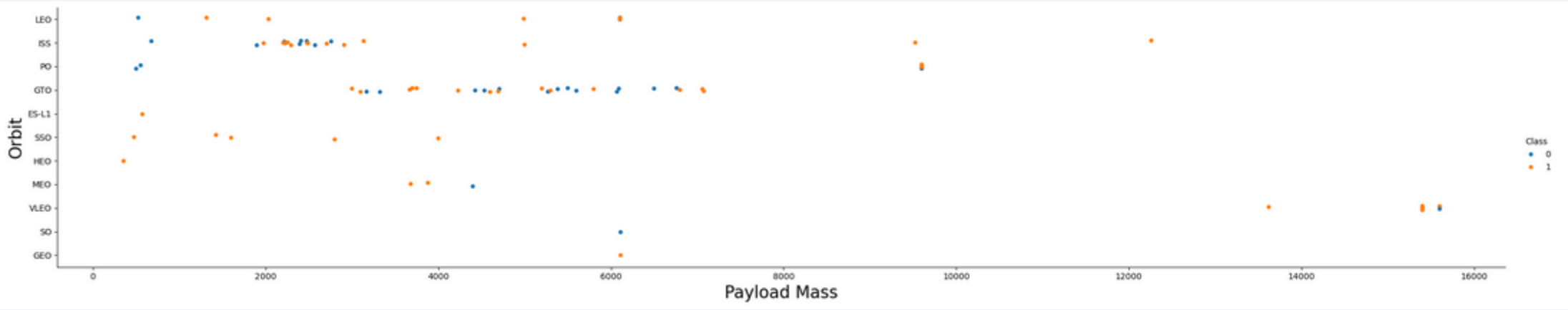
From the bar plot it can be seen the four of the Orbits have higher success rate than the others.

Flight Number vs. Orbit Type



- It can be seen from the plot there are two notable findings:
  - There is a relationship between the success of the number of flights when analyzing the LEO orbit.
  - There exists no relationship between the number of flights when analyzing the GTO orbit.

Payload vs. Orbit Type



- As seen in the scatterplot above, successful landings occur more for PO, ISS and LEO orbits when considering heavy payloads.

Launch Success Yearly Trend

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block
0	1	2010	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0
1	2	2012	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0
2	3	2013	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0
3	4	2013	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0
4	5	2013	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0

- Using the information above a line chart was plotted which displayed that the success rate kept a steady increase from 2013 until 2010

All Launch Site Names

- Using the DISTINCT keyword, four unique launch sites were found, namely:

1. KSC LC-39A
2. CCAFLS LC-40
3. VAF8 SLC-4E
4. CCAFLS SLC-40

Display the names of the unique launch sites in the space missi

```
[10]: %sql select DISTINCT launch_site from SPACEXTABLE;
```

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

[10]:

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

Launch Site Names Begin with 'CCA'

- The following query was used to find the names beginning with CCA:

```
%sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5
```

]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAY
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	

Total Payload Mass

- Using the following query the total payload mass was calculated:

```
%sql select sum(payload_mass__kg_) as sum from SPACEXtable where customer like 'NASA (CRS)'
```

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

sum
45596

Average Payload Mass by F9 v1.1

- Using the following query, the average payload mass by F9 c1.1 was calculated to 2534.66

```
%sql select avg(payload_mass__kg_) as sum from SPACEXtable where booster_version like 'F9 v1.1%'
```

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

sum
2534.6666666666665

First Successful Ground Landing Date

- Using the following query, the first successful ground landing date was calculated to 2010-06-04

```
%sql select min(date) as Date from SPACEXtable where mission_outcome like 'Success'
```

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

Date
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

- Using the query below, the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are as follows:

```
%sql select booster_version from SPACEXtable where mission_outcome like 'Success' AND (payload_ma
```

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

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

Total Number of Successful and Failure Mission Outcomes

- The calculation of the total number of successful and failure mission outcomes were done. It is seen that there was a failure (in flight), 99 successes and a single success where the payload status was unclear.

```
%sql SELECT mission_outcome, count(*) as Count from SPACEXTABLE GROUP by mission_outcome ORDER BY mission_outcome
```

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

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

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass were calculated by using the query (seen below) and produced the following results:

```
maxm = %sql select_max(payload_mass__kg_) from SPACEXTABLE
%sql select booster_version from SPACEXTABLE where payload_mass__kg_= (select max(payload_mass__kg_) from SPACEXTABLE)
```

```
 * sqlite:///my_data1.db
(sqlite3.OperationalError) near "select_max": syntax error
[SQL: select_max(payload_mass__kg_) from SPACEXTABLE]
(Background on this error at: https://sqlalche.me/e/20/e3q8)
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 are listed below:

```
]: %sql select MONTHNAME(DATE) as Month, landing_outcome, booster_version, launch_site from SPACEXTABLE where DATE like '2015%'
 * sqlite:///my_data1.db
(sqlite3.OperationalError) no such function: MONTHNAME
[SQL: select MONTHNAME(DATE) as Month, landing_outcome, booster_version, launch_site from SPACEXTABLE where DATE like '2015%'
AND landing_outcome like 'Failure (drone ship)']
(Background on this error at: https://sqlalche.me/e/20/e3q8)
```

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

- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order are ranked below:

```
[31]: %sql selecting

 * sqlite:///my_data1.db
(sqlite3.OperationalError) near "selecting": syntax error
[SQL: selecting]
(Background on this error at: https://sqlalche.me/e/20/e3q8)
```

Section 3

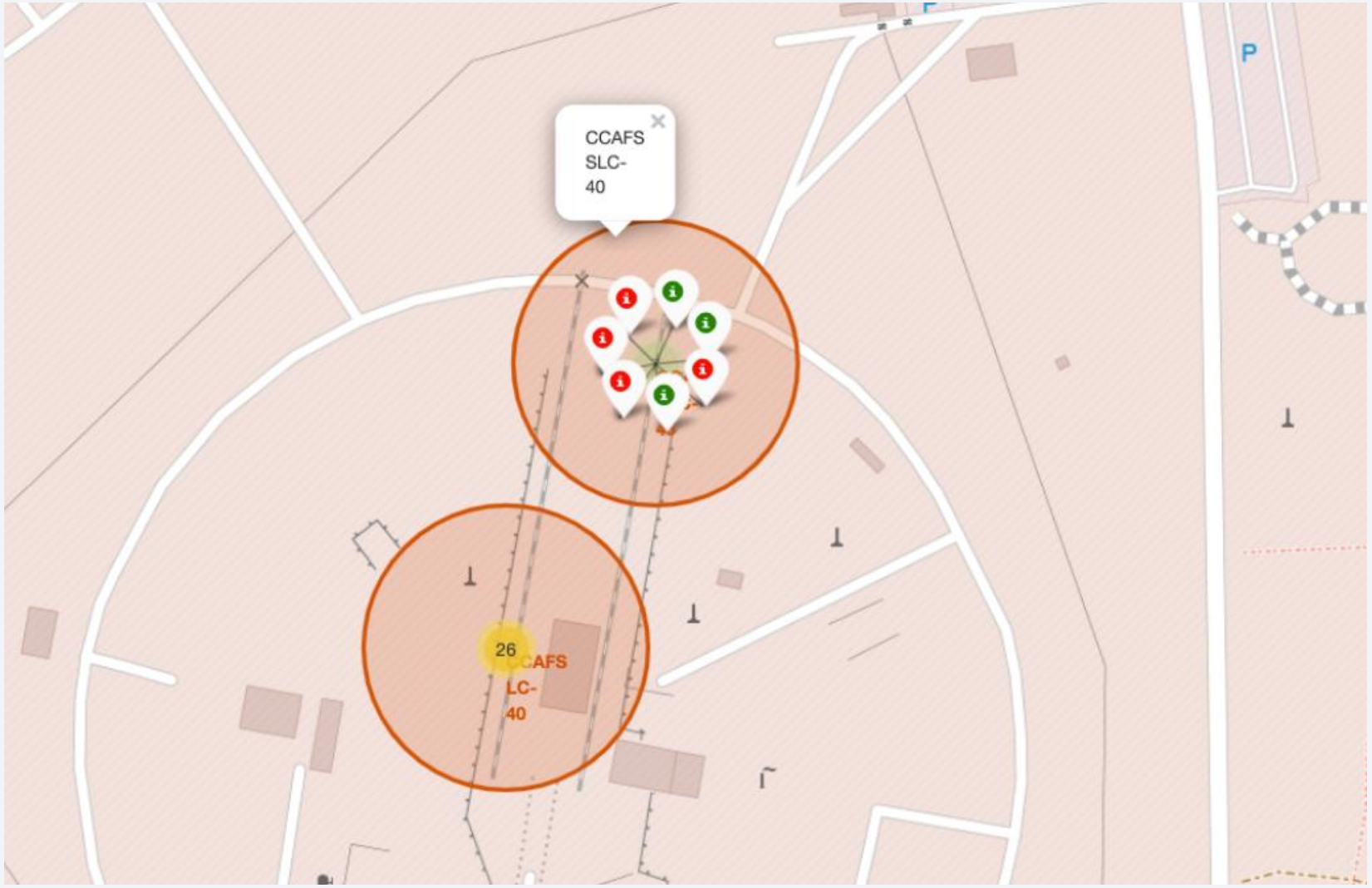
Launch Sites Proximities Analysis

Launch Site locations

It can be seen that the launch site is in Florida and near the coast.

- Here we can see both launch sites are located in Florida and in California, by the coasts.

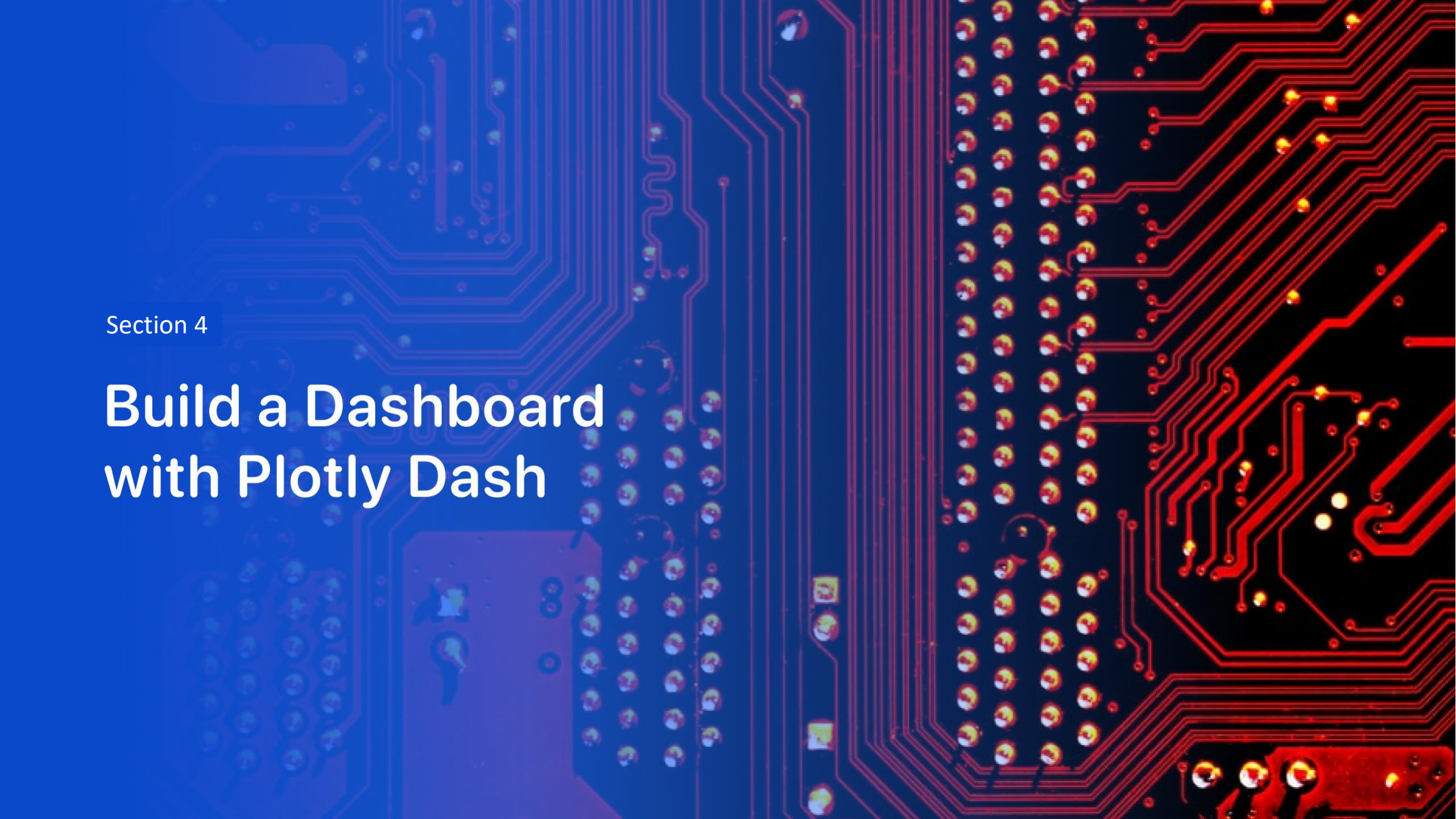
Color-coded markers showing launch sites



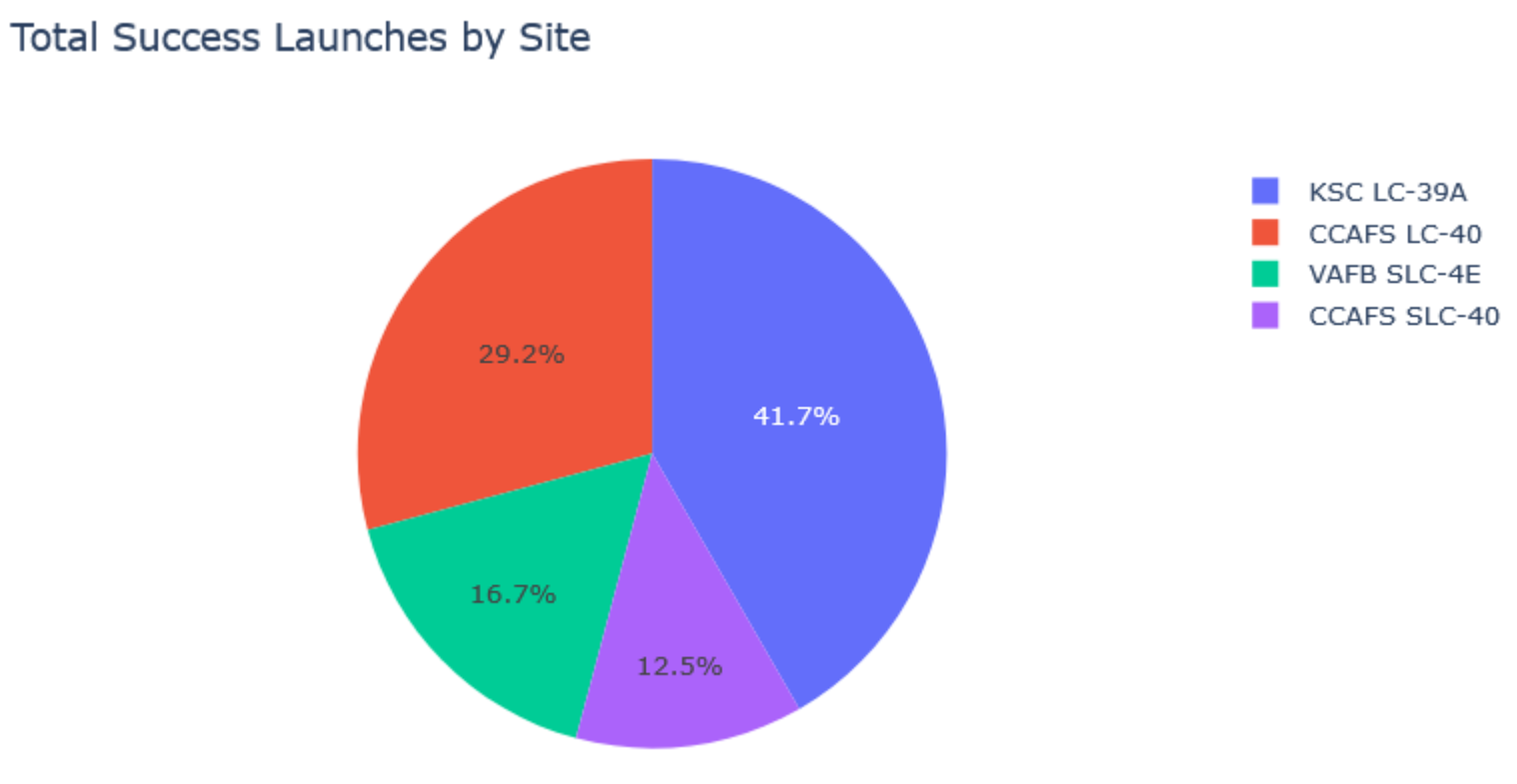
- Red markers show failed outcomes of launching; green show successful outcomes of the CCAFS SLC-40

Key Location Proximities to Launch Sites

- According to the map, it is seen that launch sites are stationed towards the coastlines, away from the city, more towards railways for the main purpose of transportation for workers and materials needed.
- It is to be noted that launch sites are located away from the cities and more towards the sea area for the reason that if there are any failures in launch, rocket debris will not cause harm upon anyone/ anything within those areas and will fall in the sea or water areas.


Section 4
Build a Dashboard with Plotly Dash

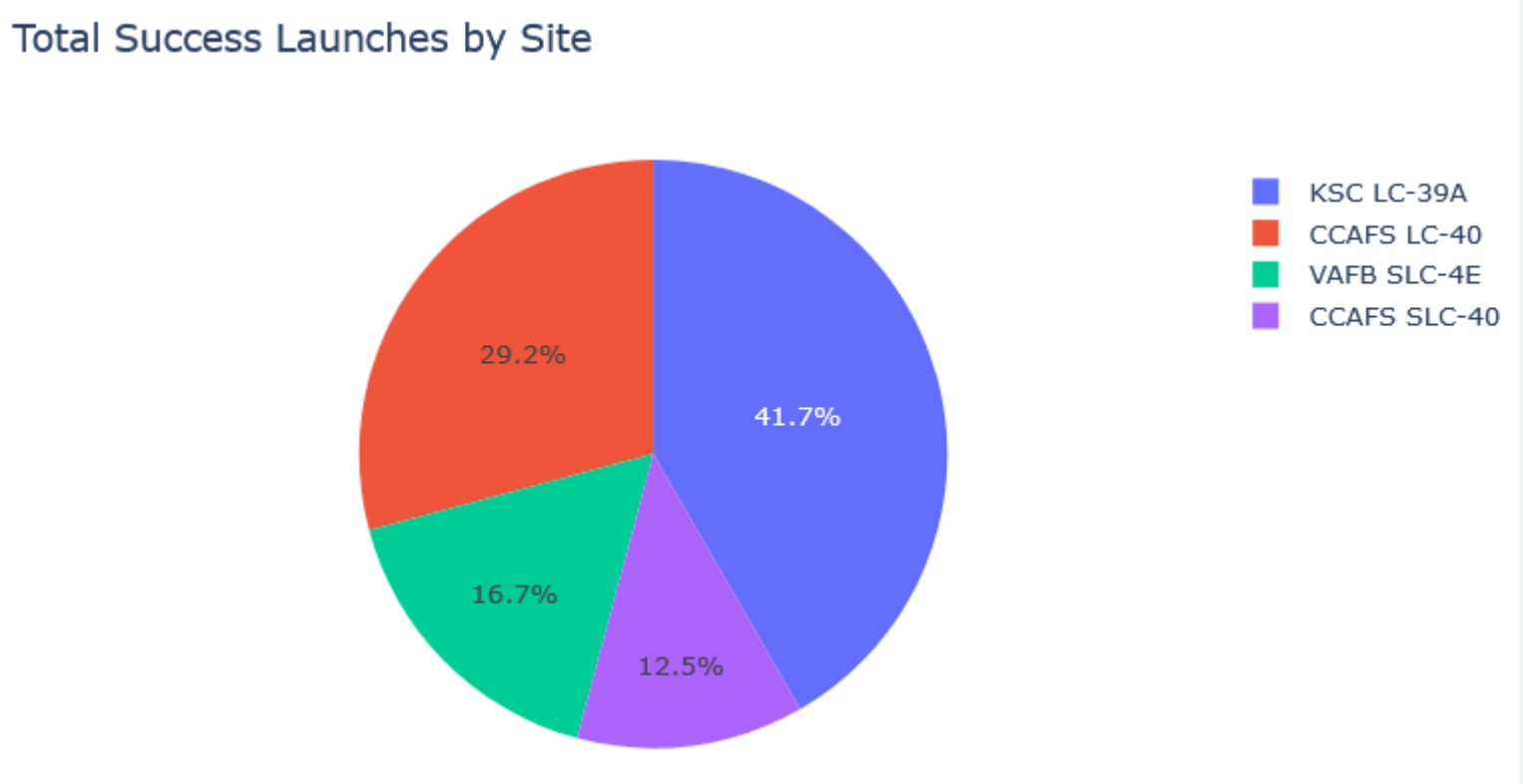
Success rate of each launch site by percentage



- Amongst the four launch sites, it is seen that the most popular one is the KSC LC-39A site.

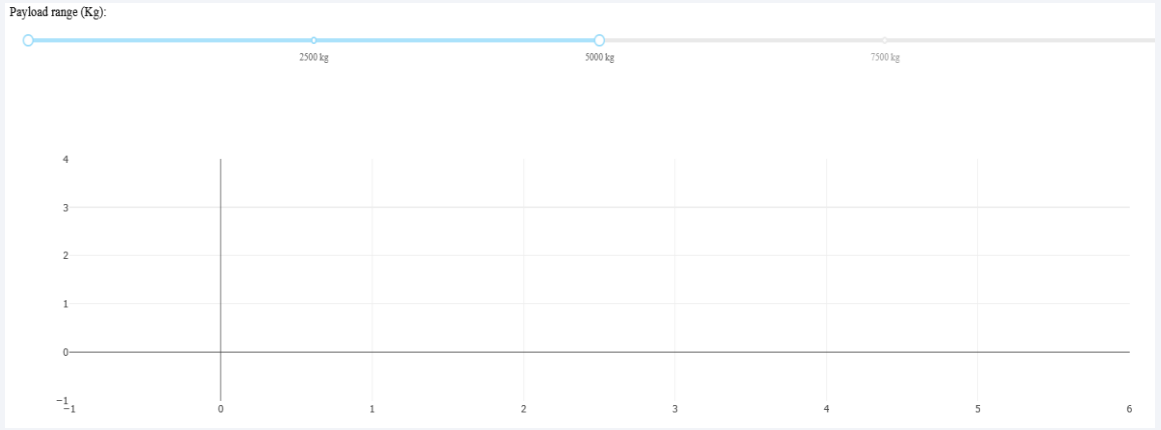
Launch Site with the highest success rate

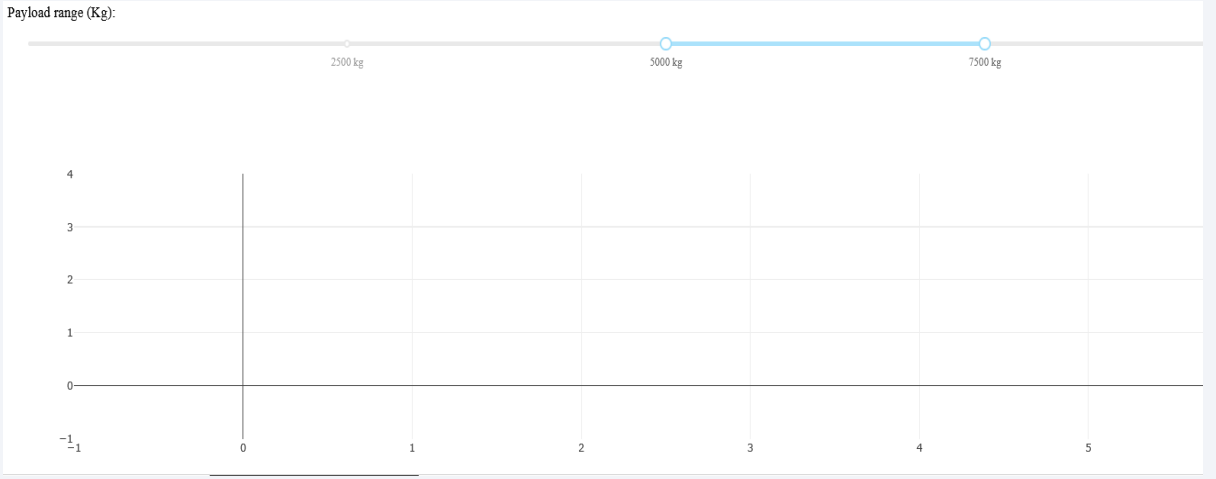
- As indicated by the pie chart, the KSC LC-39A has the highest success rate. With more than seventy percent of successful landings and under thirty percent of failed ones.



Scatterplot (Payload versus the Launch Outcomes for all of the sites)

- A comparison between the different payloads using the range slider can be made. With the slider ranging from 2500 to 7500 kg, a low weighted payload 2500 to 5000kg vs a higher weighted payload of 5000 to 700kg produced a significant finding:
    - The lower weighted payload produced more success rates than the heavy ones.



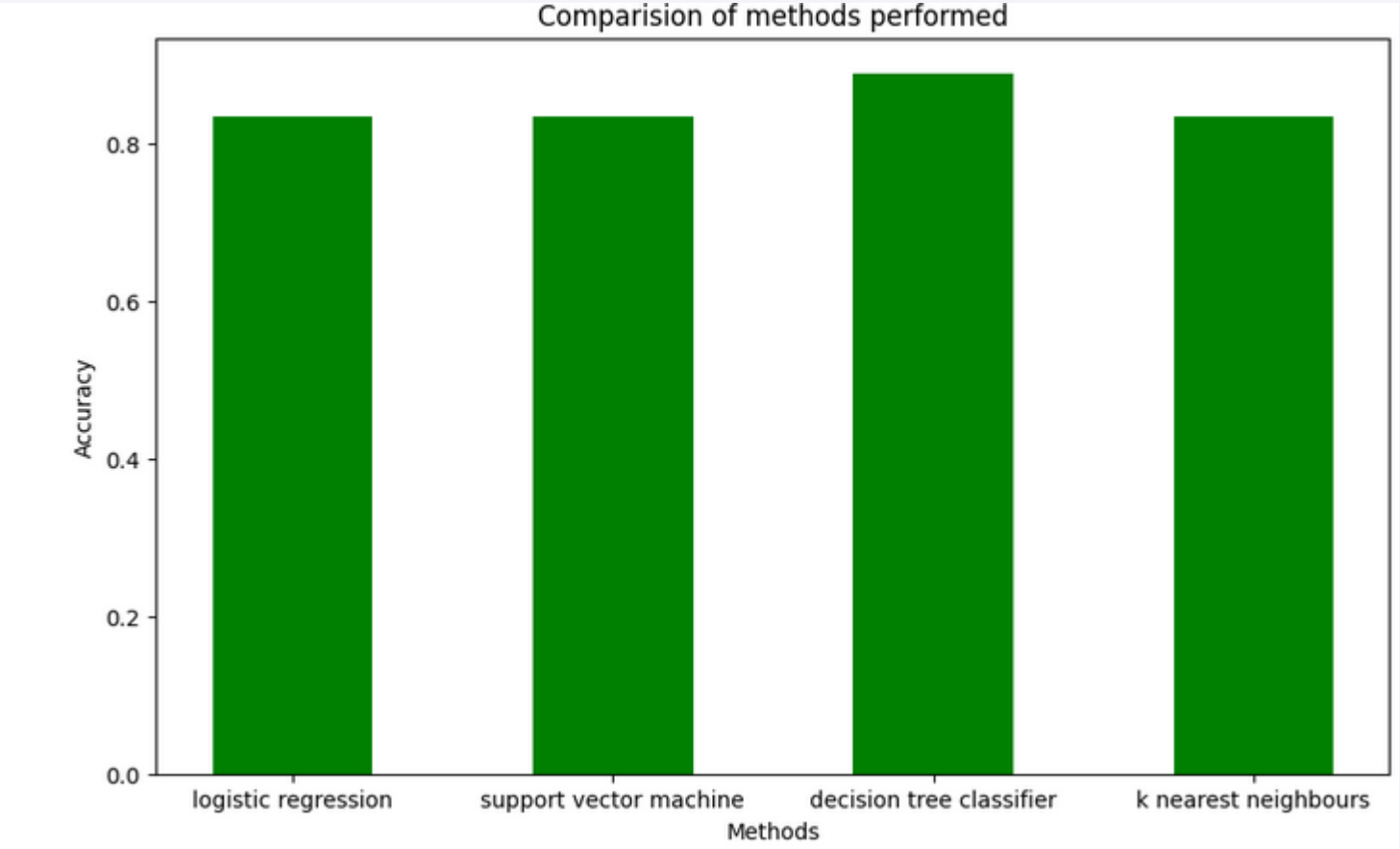


Section 5

Predictive Analysis (Classification)

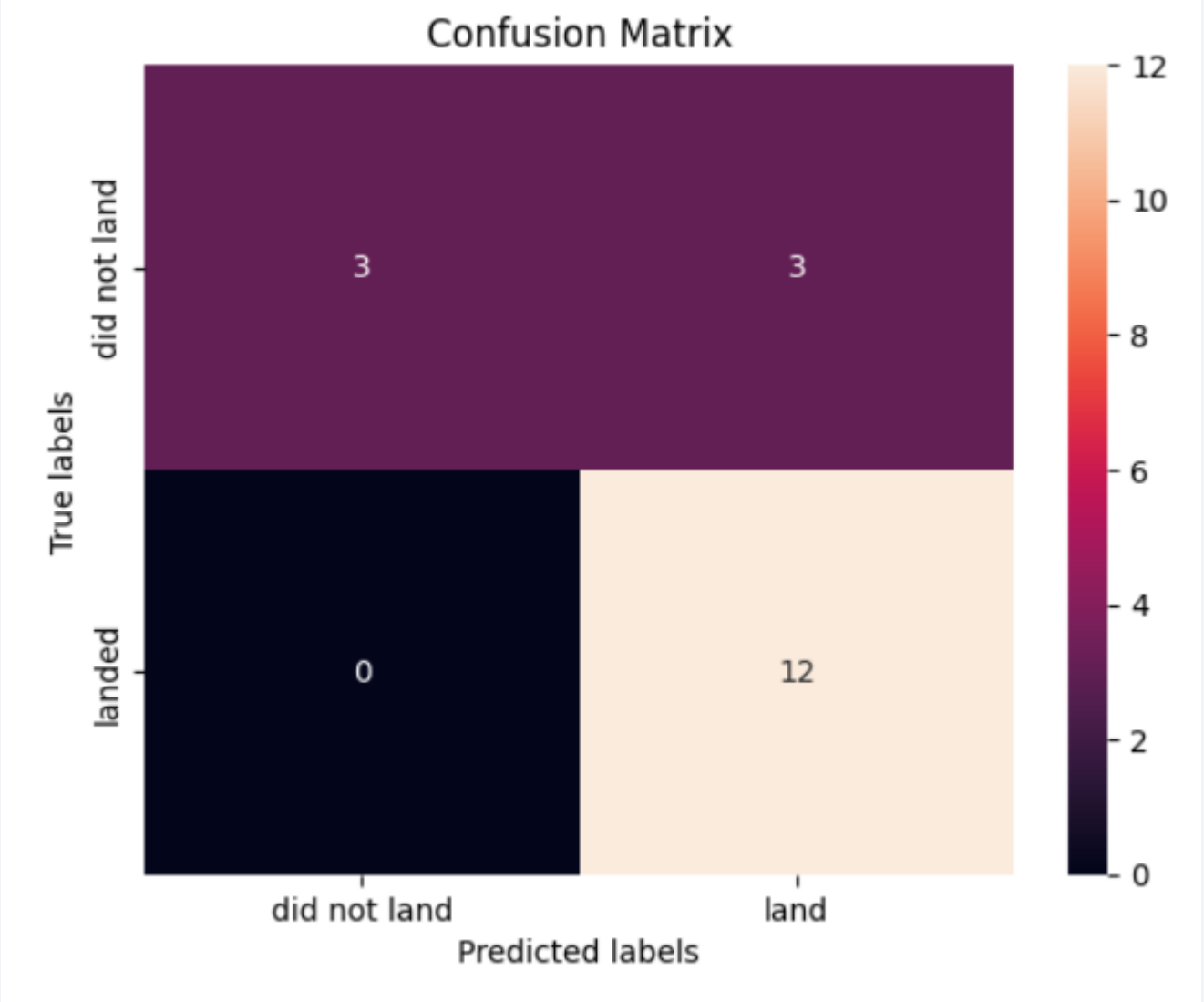
Classification Accuracy

- Almost all of the models have the same accuracy, except that of the decision tree classifier, which exceeds the 83.3 percent of the other models.



Confusion Matrix

- As seen in the confusion matrix (on the right), for the decision tree classifier model.
- There is an issue with false positives (true labels shows unsuccessful landings when the model indicates that 3 landings were successful)
- The classifier tends to over predict.



Conclusions

- Based on the project carried out, it can be concluded that:
  - Data was collected using the API from the public SpaceX API and the web scraping data from the SpaceX Wikipedia page.
  - The first successful ground landing date was calculated to 2010-06-04
  - The calculation of the total number of successful and failure mission outcomes were done: 1 failure (in flight), 99 successes and 1 single success where the payload status was unclear.
  - The Decision Tree Classifier is indicated to be the best choice to carrying out the Predictive Analysis stage of the project.
  - However, the model tends to over predict and has cases of false positives (as seen in the confusion matrix)

Thank you!