Winning Space Race with Data Science

Manasa
01/09/2025

Outline

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

Executive Summary

- First the required data of SpaceX is collected using APIs and then data wrangling is done after which EDA is performed.
- Using folium and ploty visualizations are created to analye the given data ,later a predictive ml models are tested to find the best fit
- It is clear that decision tree model is best fit for the prediction of launcher with an accuracy of over 94 percent.

Introduction

- Project background and context : SpaceX Falcon 9 first stage Landing Prediction
- The goal is to 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. In this lab, you will collect and make sure the data is in the correct format from an API. The following is an example of a successful and launch
- Problems you want to find answers : The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations. Finding best fit ML model is also important.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data is collected with web scraping using API
- Perform data wrangling
  - Data is first cleaned and outliers are handled and is normalized.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Various classification models are built and analysed for best fit.

Data Collection

- To collect the data first made a get request to the SpaceX API.
- Also basic data wrangling and formatting is done in here.

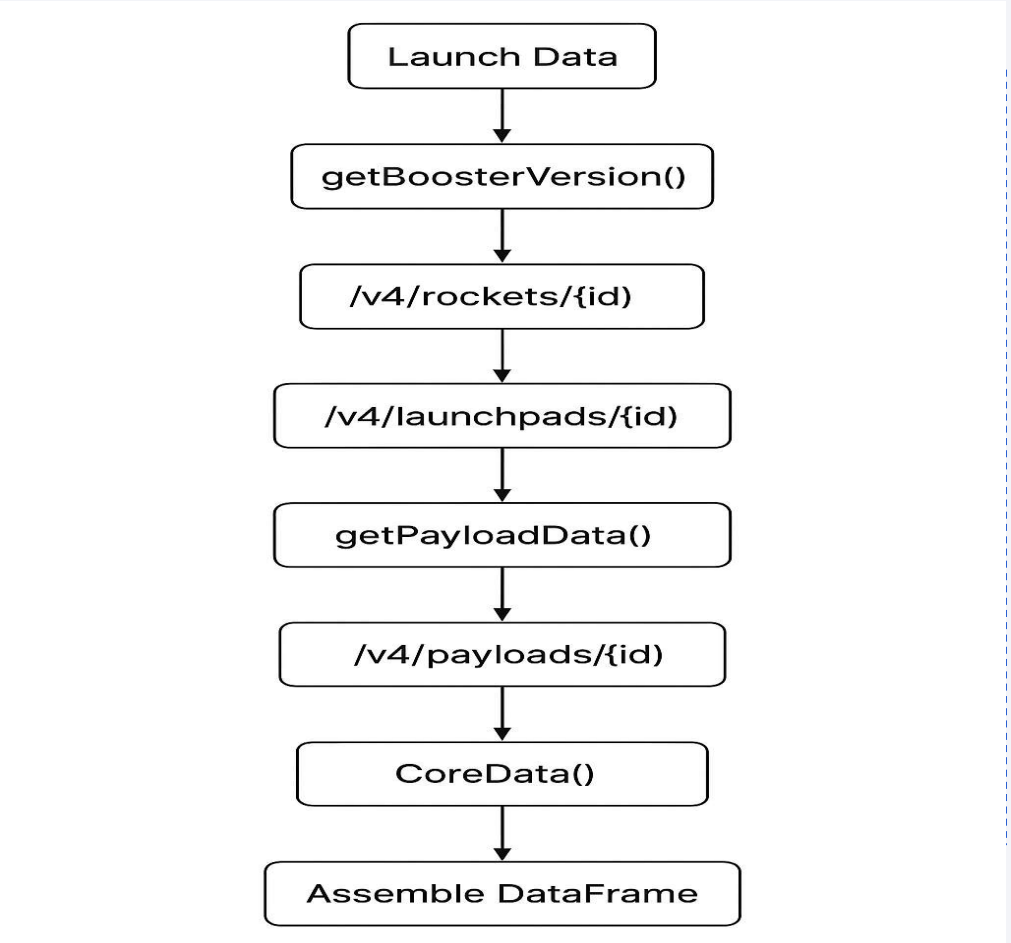
- Request to the SpaceX API

- Clean the requested data

- After the data is collected and cleaned it is then saved into an csv file for further use.

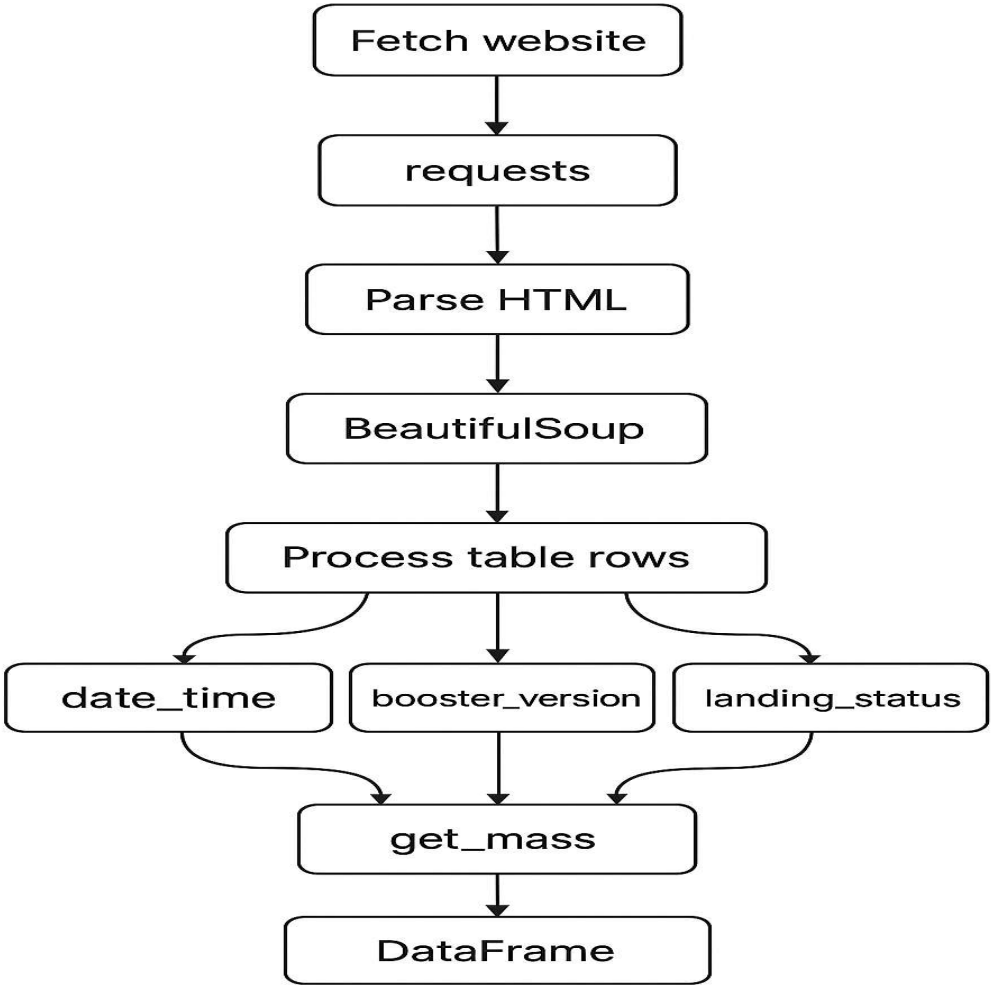
Data Collection – SpaceX API

- https://github.com/M27113/DS_SpaceX/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Web scrap Falcon 9 launch records with `BeautifulSoup`:
- - Extract a Falcon 9 launch records HTML table from Wikipedia
- - Parse the table and convert it into a Pandas data frame



Data Wrangling

- In this step I have performed web scraping 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
- https://github.com/M27113/DS_SpaceX/blob/main/jupyter-labs-webscraping.ipynb

EDA with Data Visualization

In this stage I have performed exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib

- Exploratory Data Analysis
- Preparing Data Feature Engineering
- I have generated scatter plots , bar graph and line plot to get better insights.
- https://github.com/M27113/DS_SpaceX/blob/main/edadataviz.ipynb

EDA with SQL

- This step EDA is performed to know the data better and analyze is further , to get clear idea.
- First understand the Spacex DataSet
- Loaded the dataset into the corresponding table in a Db2 database
- Executed SQL queries to analyze the data.
- https://github.com/M27113/DS_SpaceX/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- In this stage I have done following tasks:
- **TASK 1:** Marked all launch sites on a map
- **TASK 2:** Marked the success/failed launches for each site on the map
- **TASK 3:** Calculated the distances between a launch site to its proximities
- https://github.com/M27113/DS_SpaceX/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.

•Added a Launch Site Drop-down Input Component

•Added a callback function to render success-pie-chart based on selected site dropdown

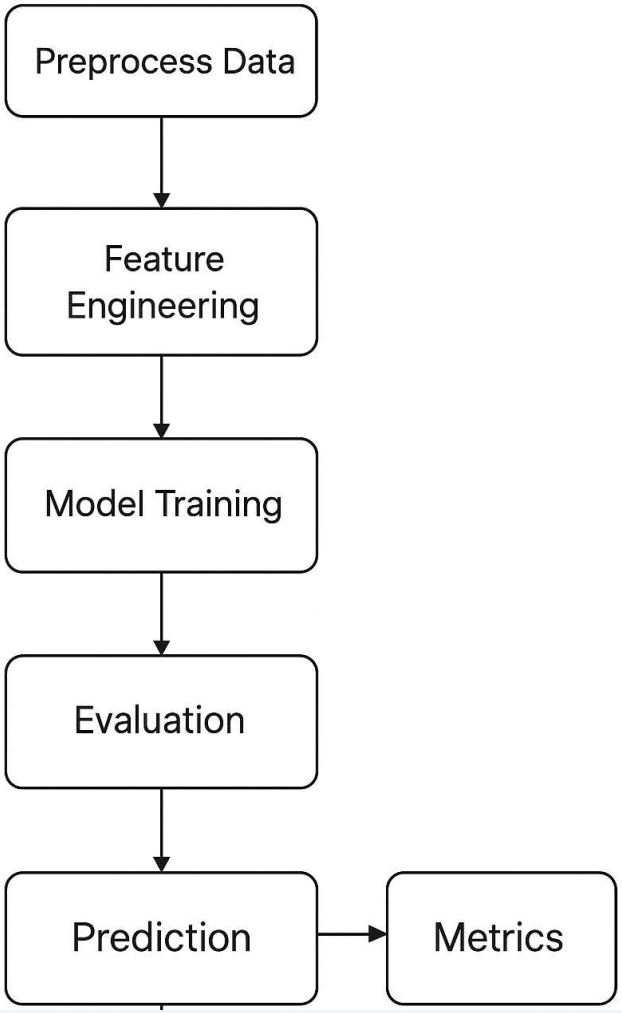
•Added a Range Slider to Select Payload

•Added a callback function to render the success-payload-scatter-chart scatter plot

- https://github.com/M27113/DS_SpaceX/blob/main/spacex-dash-app%20(1).py

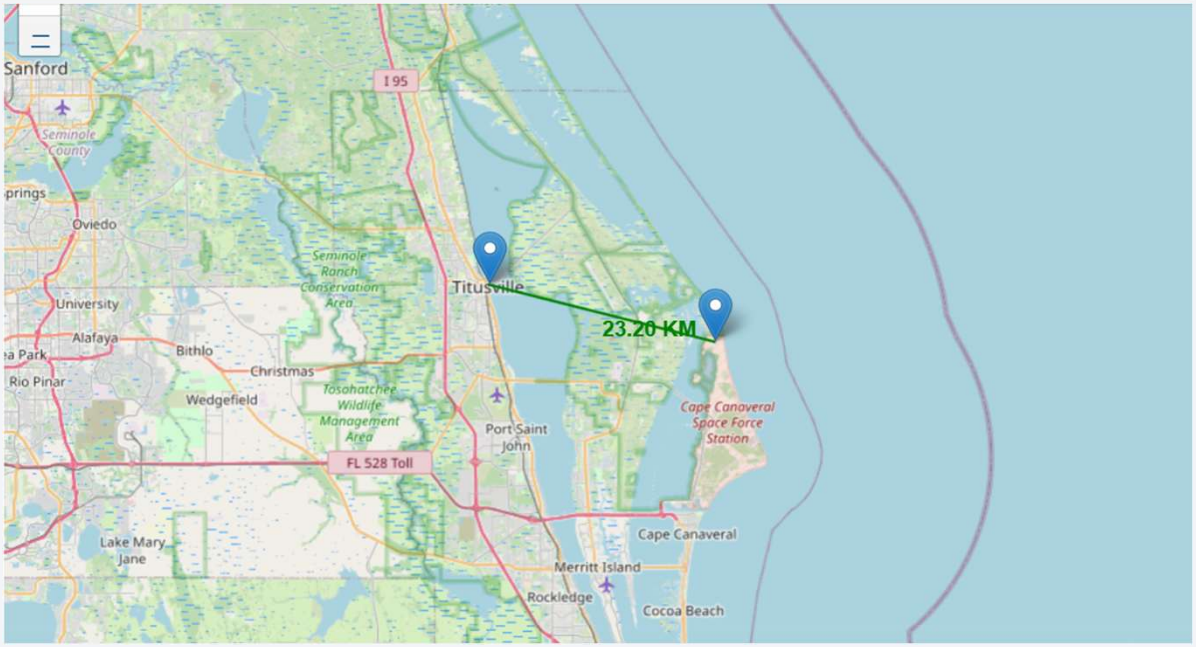
Predictive Analysis (Classification)

- Perform exploratory Data Analysis and determine Training Labels
- Created a column for the class
- Standardized the data
- Data is split into training data and test data
- Tested to find best Hyperparameter for SVM, Classification Trees and Logistic Regression and the method performs best using test data
- https://github.com/M27113/DS_SpaceX/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

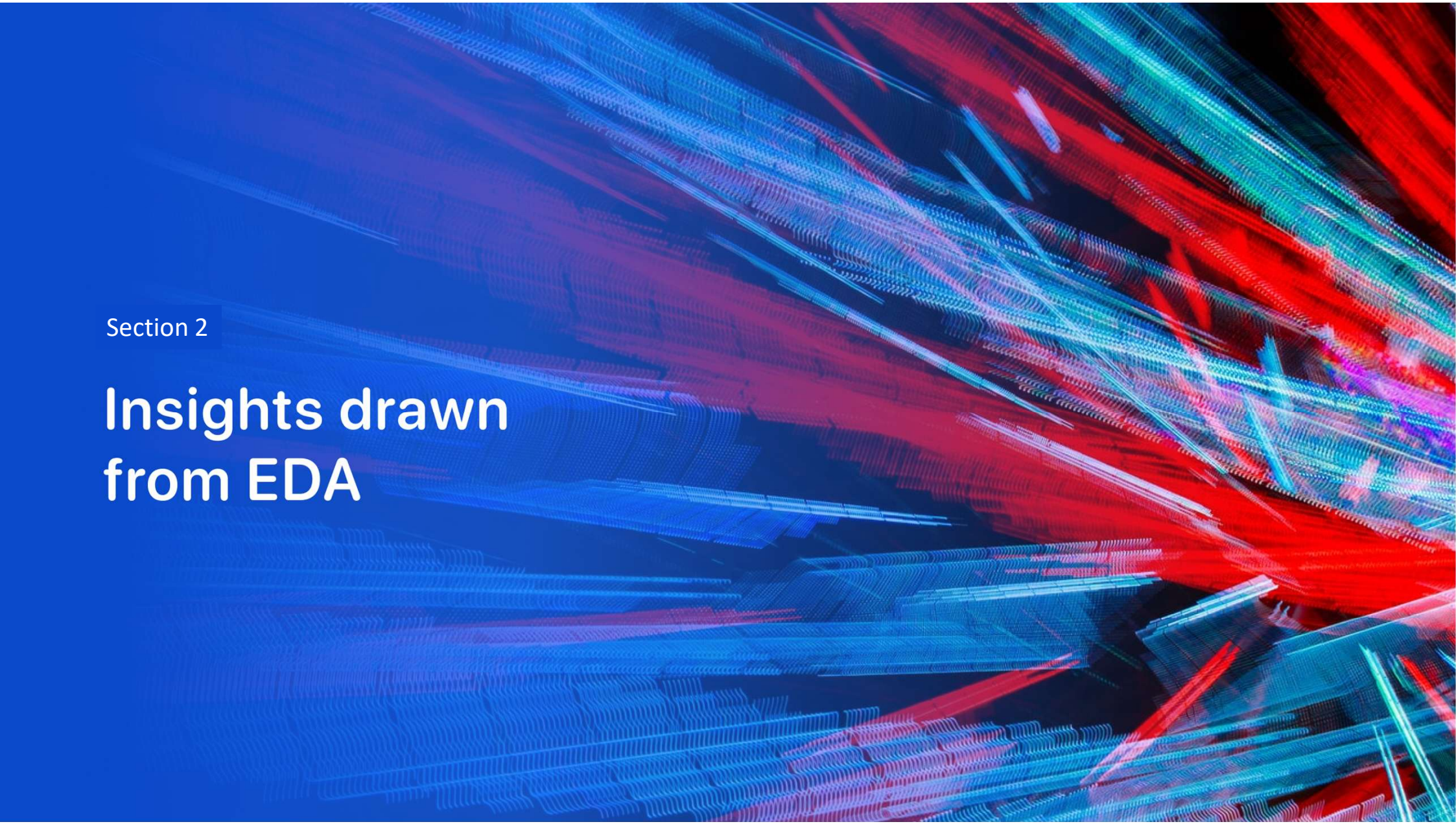


Results

Landing_Outcome	outcome_count
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



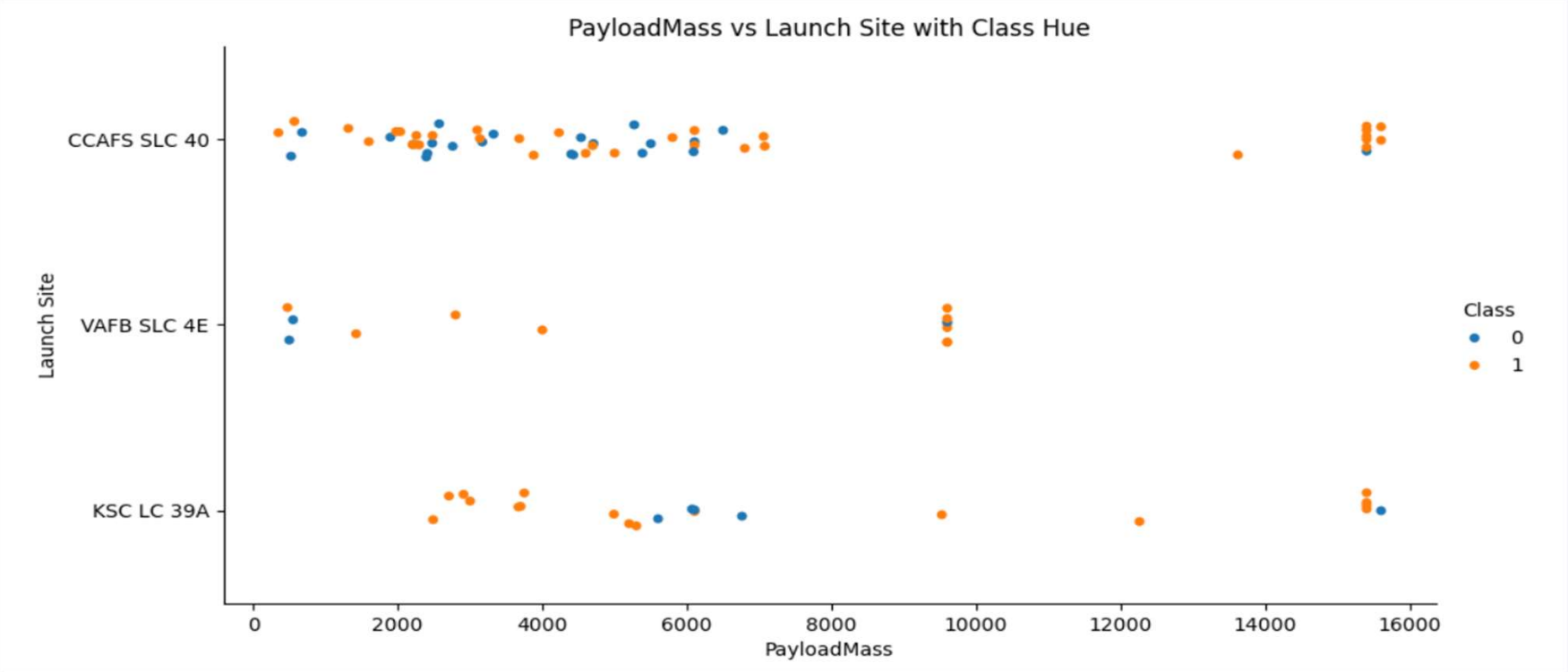
- Above images show results for EDA sql and distance from launch site results.
- Predictive analysis results shows that Decision tree is best classification model


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

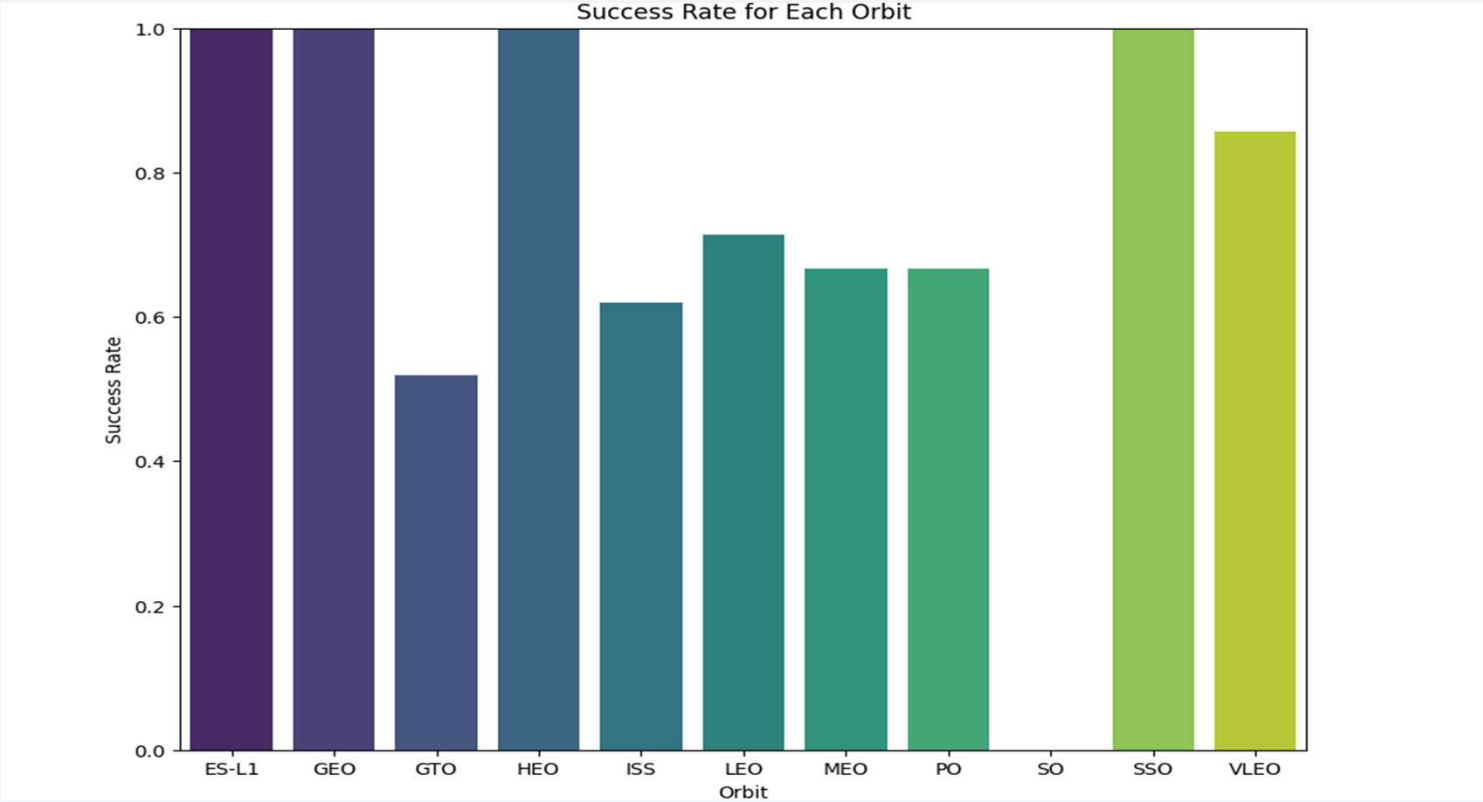
- The image shows a scatter plot of Flight Number vs. Launch Site

Payload vs. Launch Site



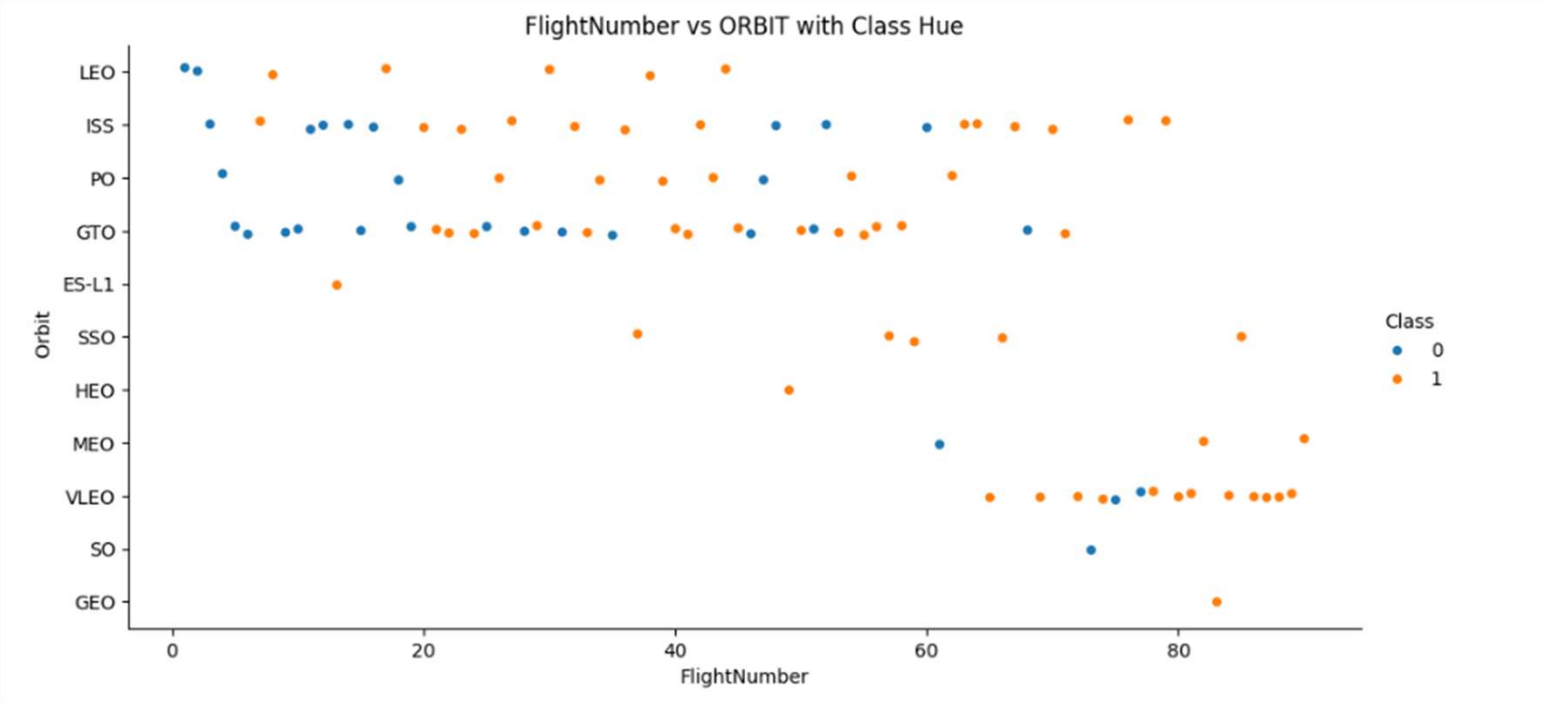
- In the Payload Mass Vs. Launch Site scatter point chart the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)

Success Rate vs. Orbit Type



- The above figure displays bar chart for the success rate of each orbit type

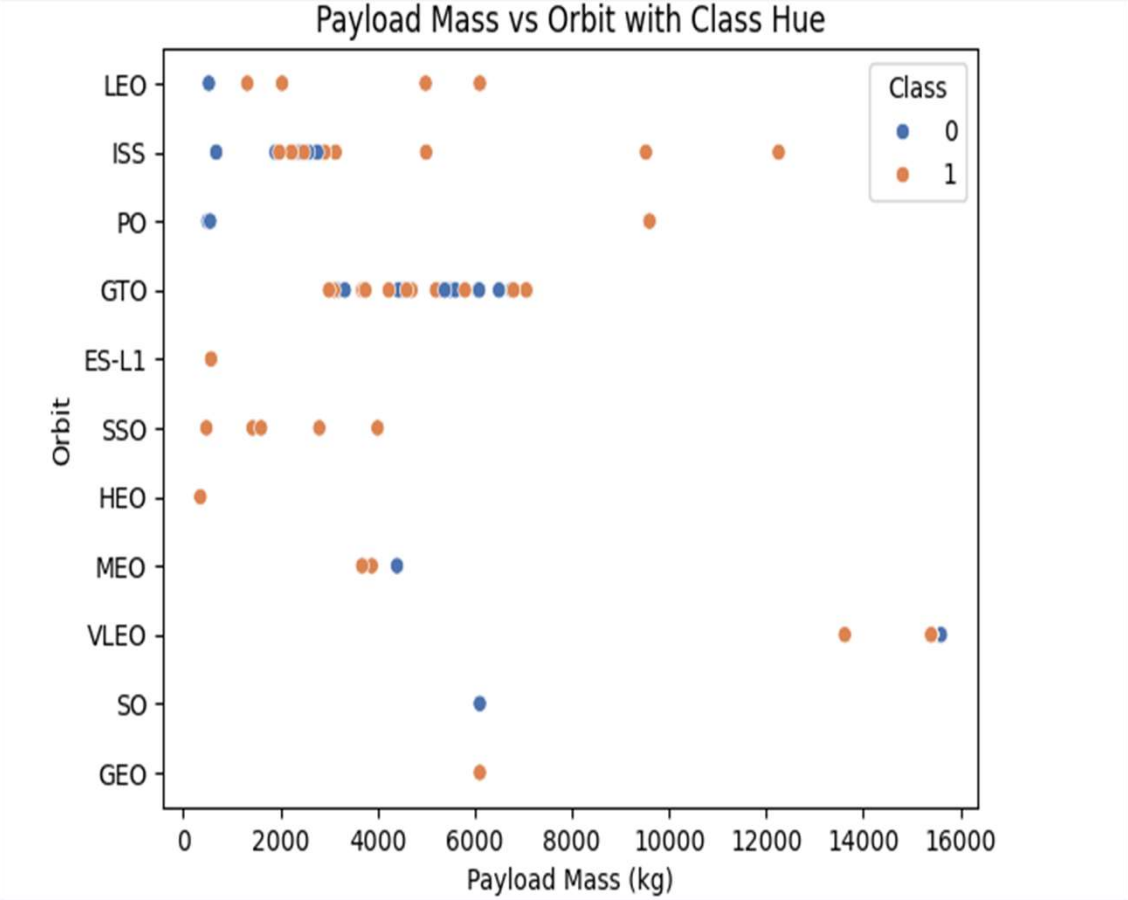
Flight Number vs. Orbit Type



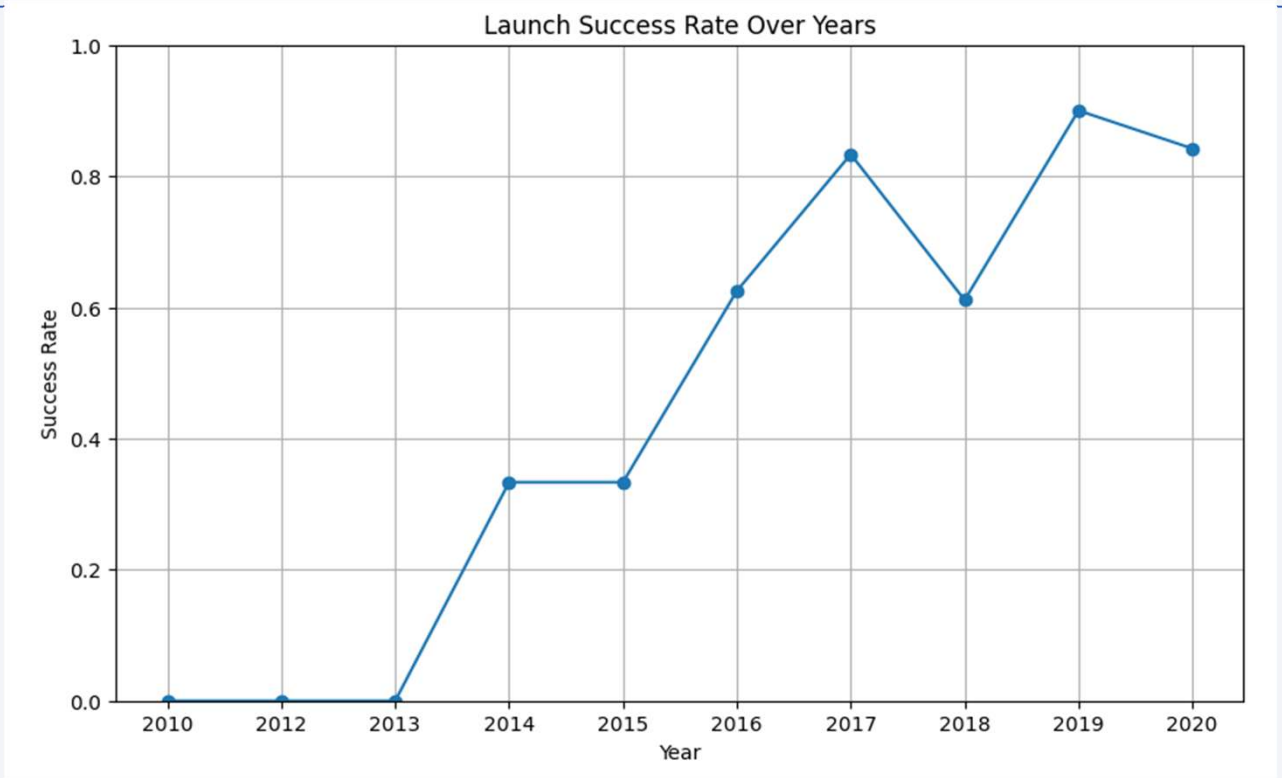
- The above chart in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success

Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



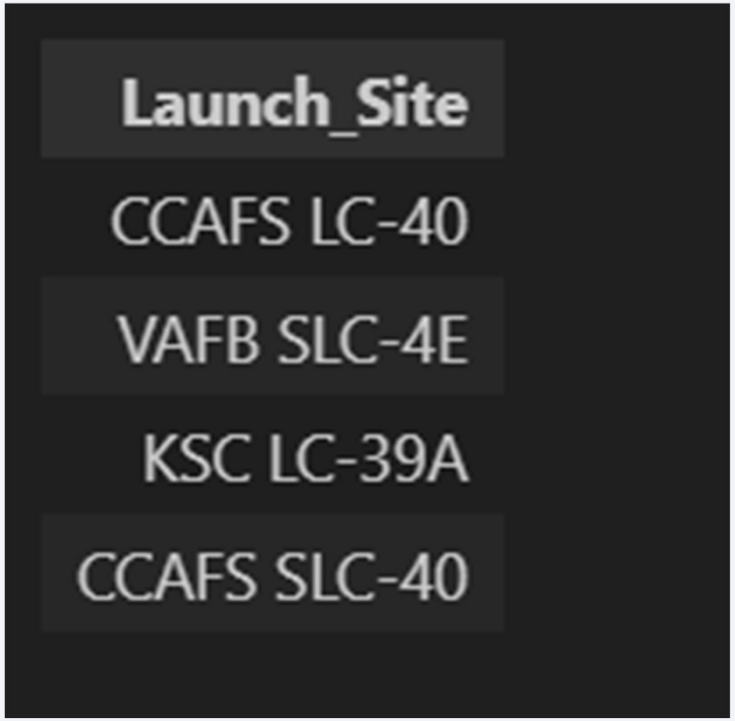
Launch Success Yearly Trend



- Show a line chart of yearly average success rate. I observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- The sql query used to find the distint launch site names is as follows:
- %sql SELECT DISTINCT launch_site FROM SPACEXTABLE
- It shows that there are 4 distint launch sites as shown in the figure.



Launch Site Names Begin with 'CCA'

- The query for 5 records where launch sites begin with `CCA` is :
- %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
2010-06-04	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

Total Payload Mass

- The total payload mass is 45596 kgs and the query used and results are as shown in figure below

```
%sql SELECT SUM(payload_mass__kg_) AS total_payload_mass FROM SPACEXTABLE WHERE customer = 'NASA (CRS)';
```

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

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1is 2928.4 kgs.
- The query used can be seen in below figure.

```
%sql SELECT AVG(payload_mass__kg_) AS average_payload_mass FROM SPACEXTABLE WHERE Booster_Version = "F9 v1.1";
```

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

average_payload_mass
2928.4

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad IS 22-12-2015
- The query and result is shown in figure below.

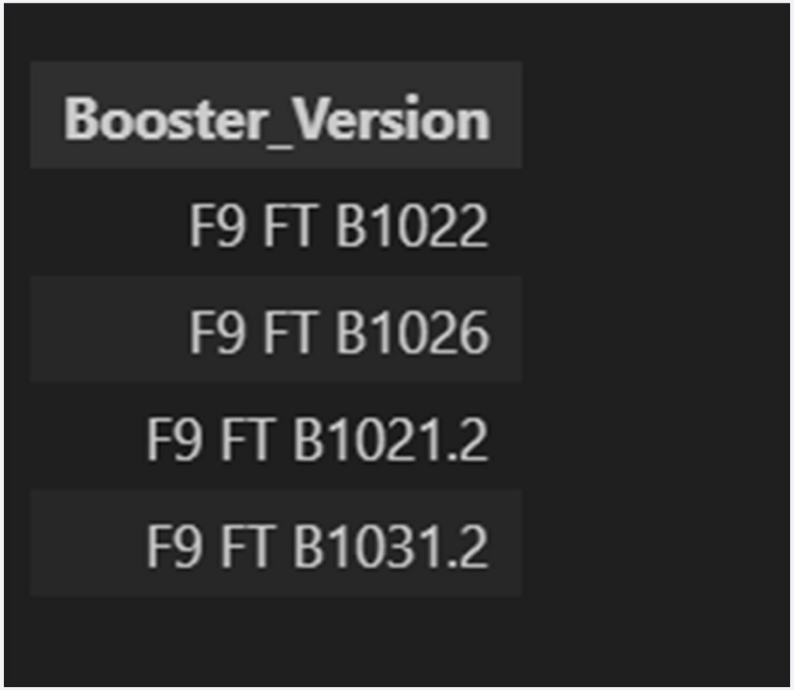
```
%sql SELECT MIN(date) AS first_successful_landing_date FROM SPACEXTABLE WHERE landing_outcome = 'Success (ground pad)';
```

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

first_successful_landing_date
2015-12-22

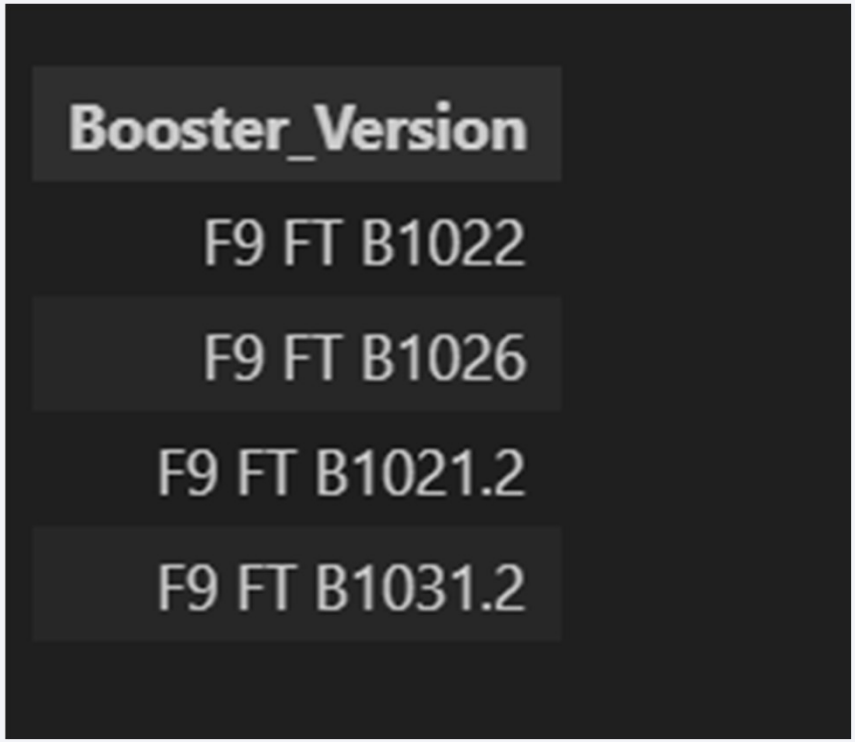
Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are shown in figure.
- %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE landing_outcome = 'Success (drone ship)' AND payload_mass__kg_ > 4000 AND payload_mass__kg_ < 6000;



Total Number of Successful and Failure Mission Outcomes

- Ttotal number of successful and failure mission outcomes is as shown is figure
- The sql query used is as follows:
- %sql SELECT mission_outcome, COUNT(*) AS total_missions FROM SPACEXTABLE GROUP BY mission_outcome;



Boosters Carried Maximum Payload

- List of names of the booster which have carried the maximum payload mass is shown in fugure.
- %sql SELECT booster_version FROM SPACEXTABLE WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTABLE);

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

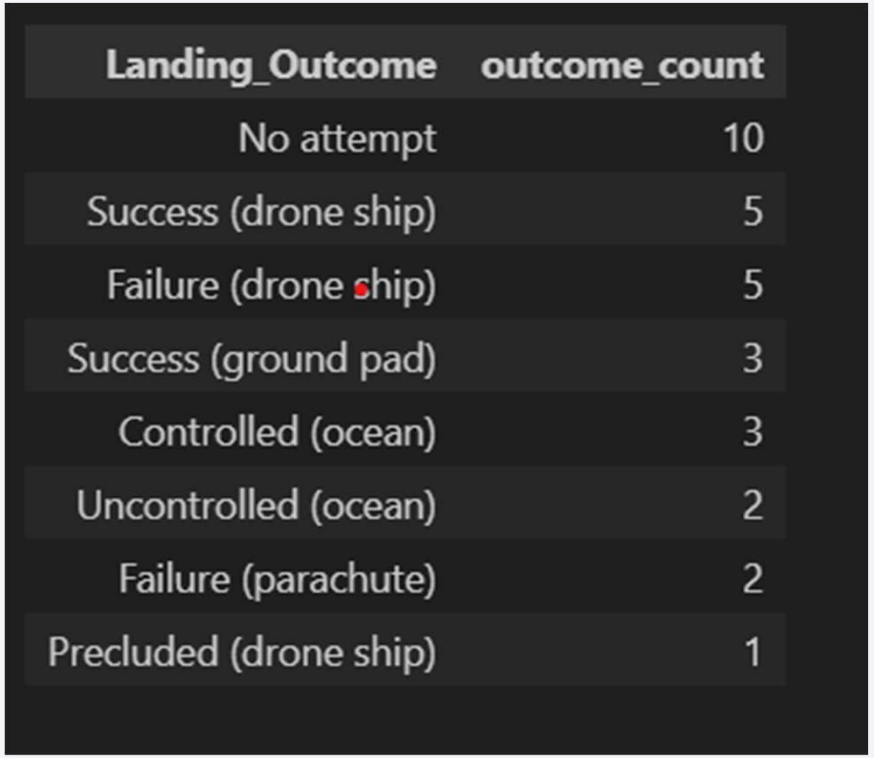
2015 Launch Records

- The query to list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 is:
- %sql SELECT substr(date, 6, 2) AS month,landing_outcome,booster_version,launch_site FROM SPACEXTABLE WHERE substr(date, 0, 5) = '2015' AND landing_outcome = 'Failure (drone ship)';

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

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

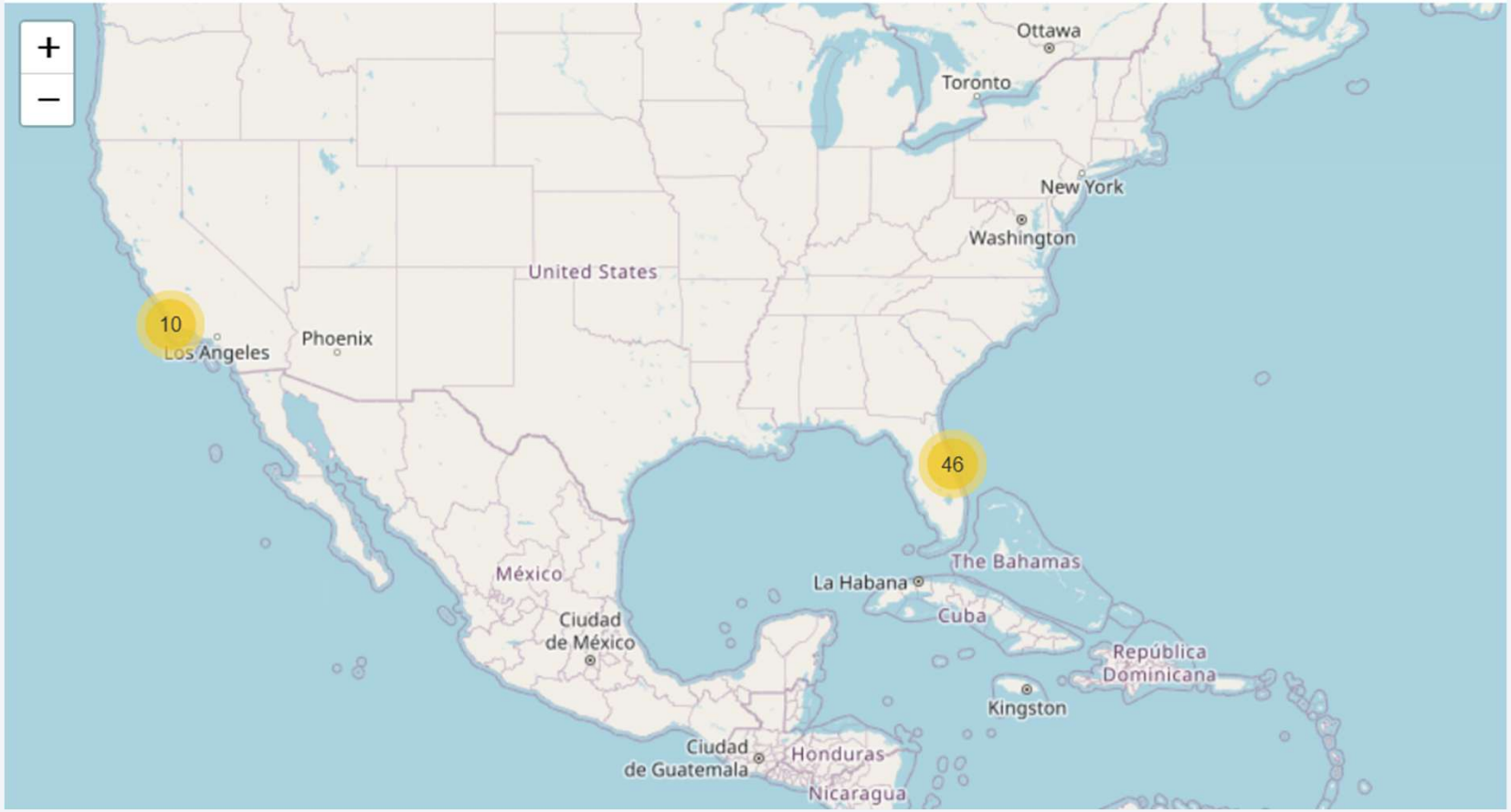
- The query to rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order is:
- %sql SELECT landing_outcome, COUNT(*) AS outcome_count FROM SPACEXTABLE WHERE date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY landing_outcome ORDER BY outcome_count DESC;



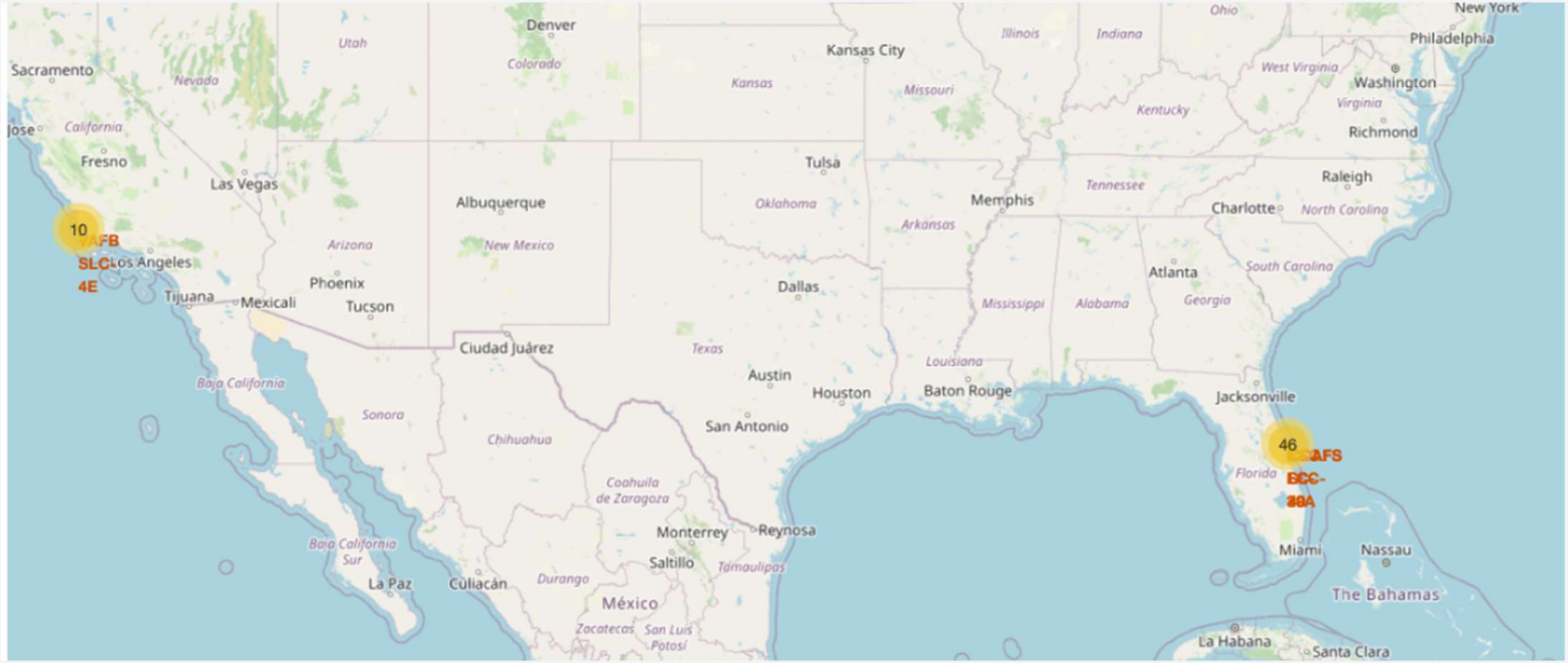
Landing_Outcome	outcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

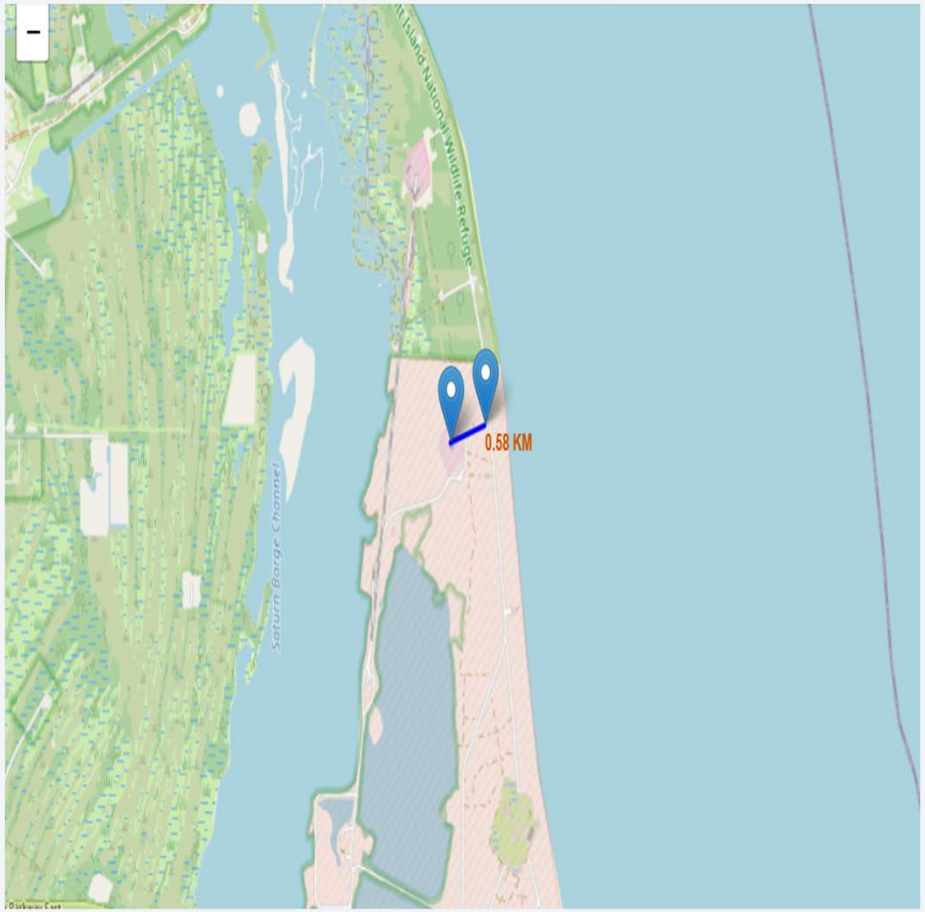
<Folium Map Screenshot 1>

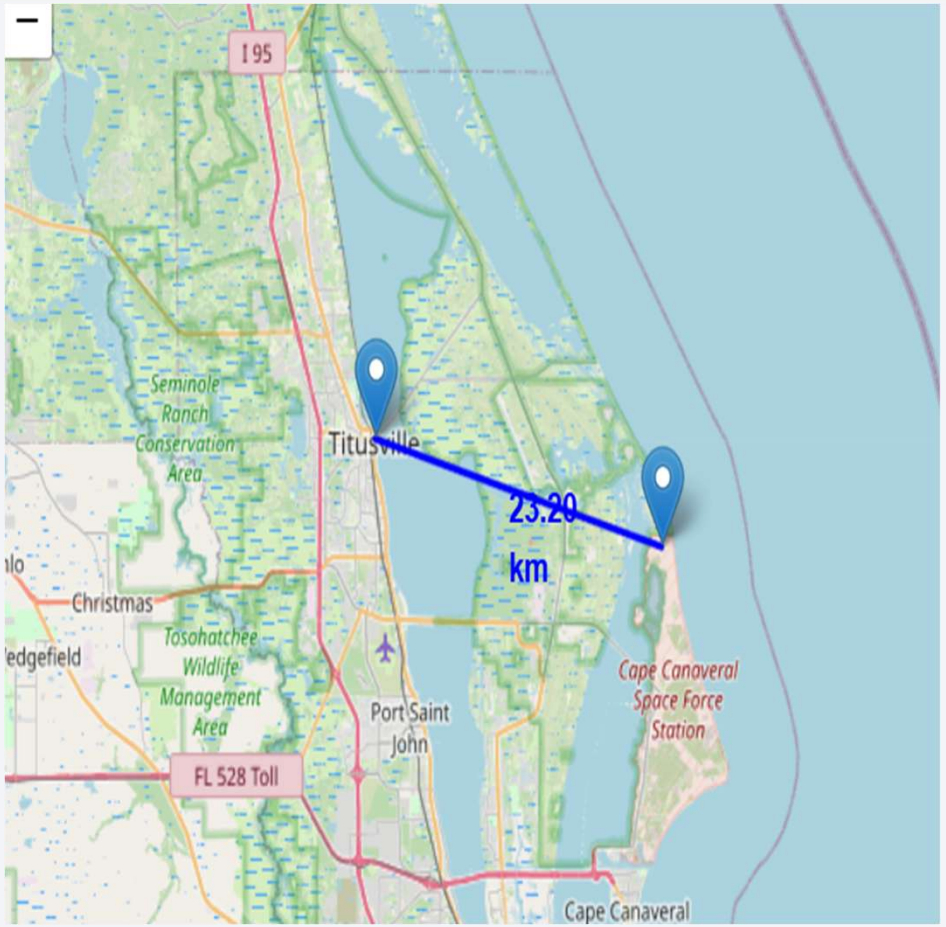


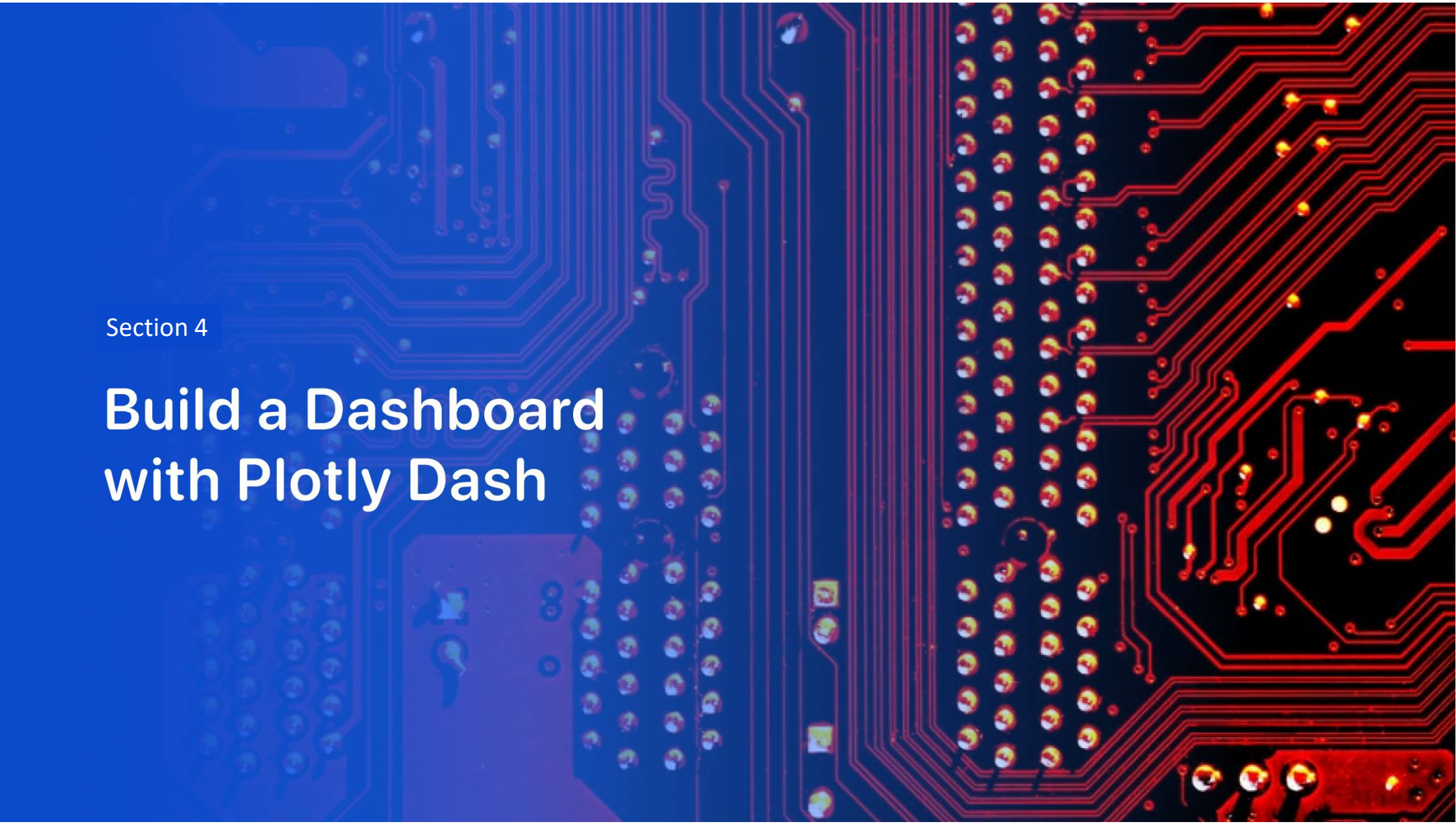
<Folium Map Screenshot 2>



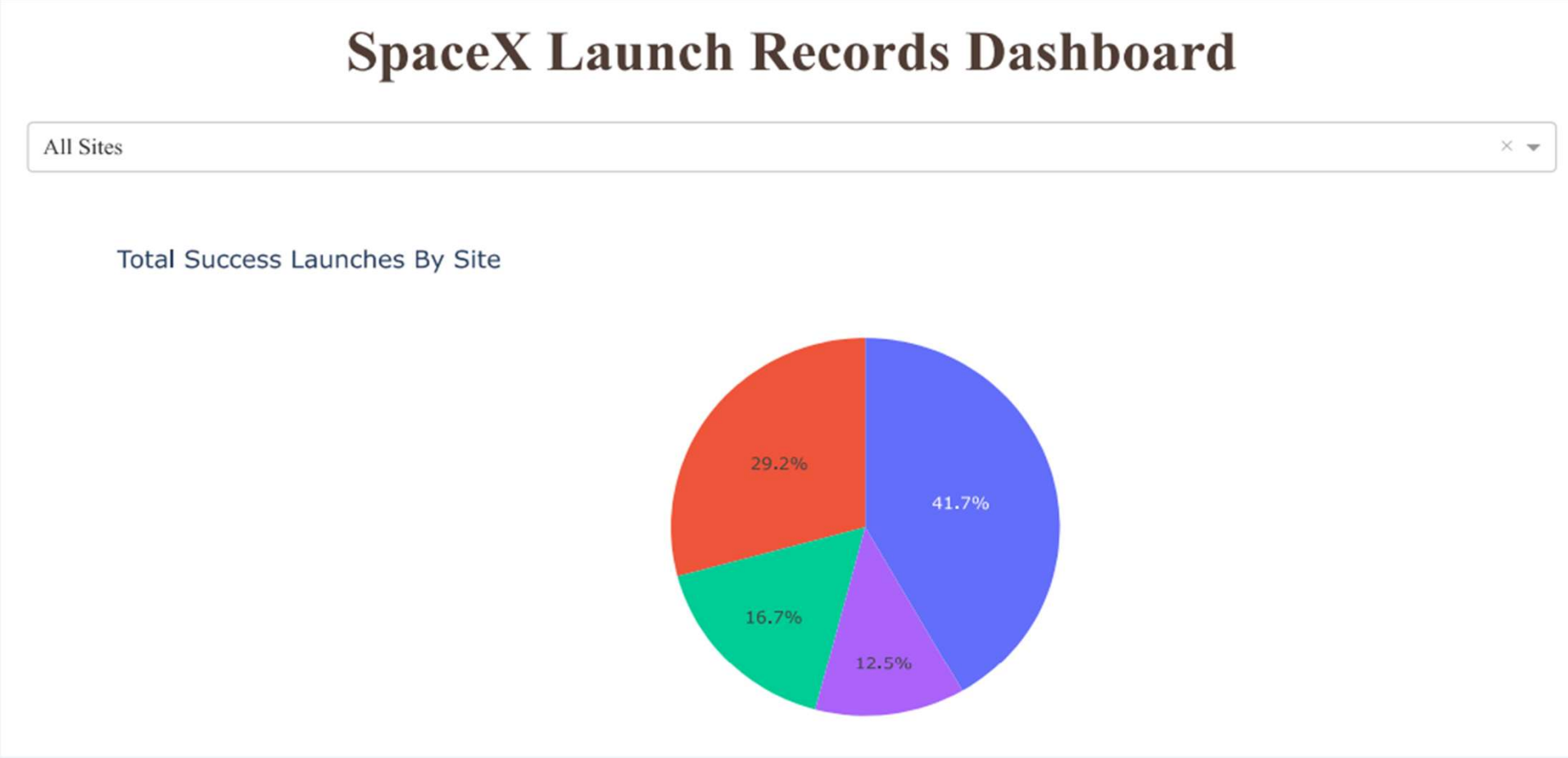
<Folium Map Screenshot 3>



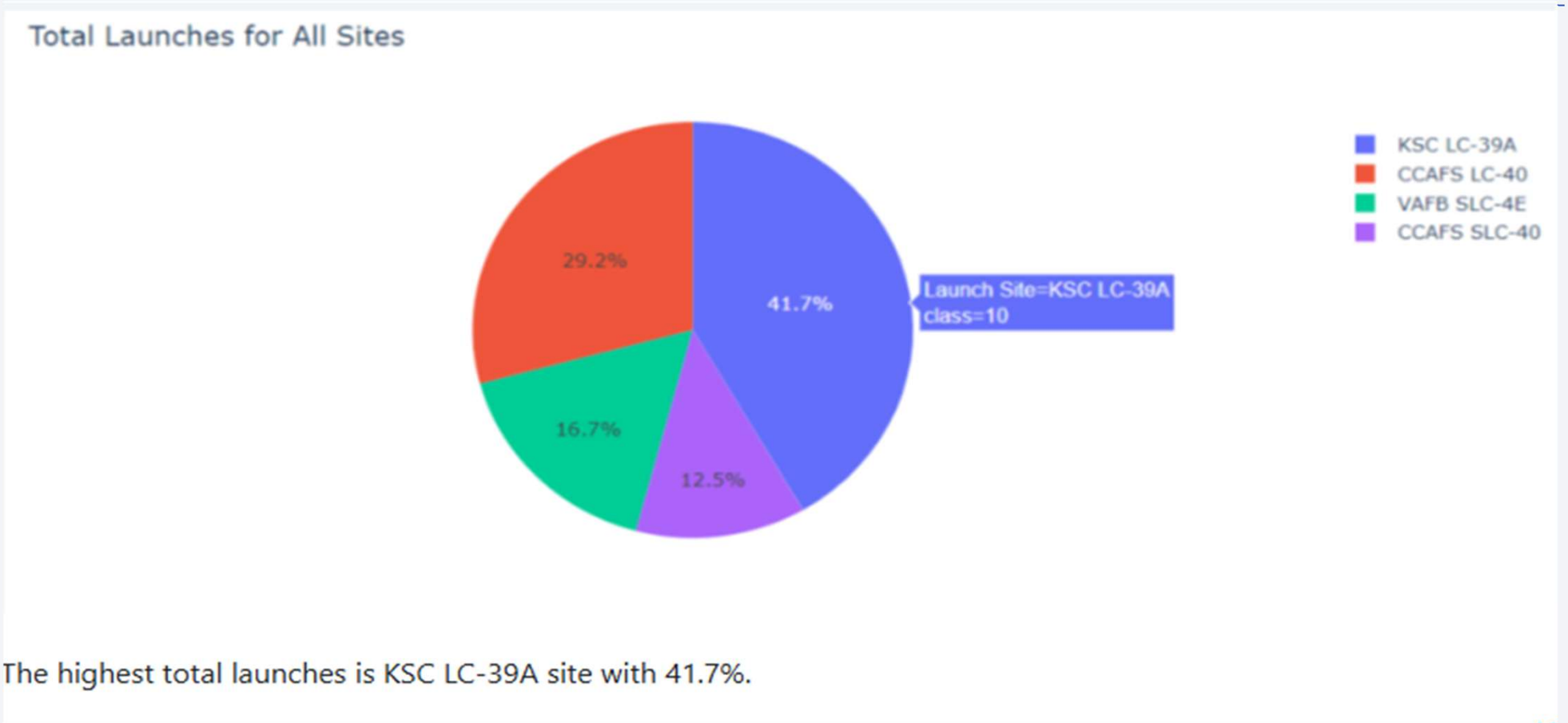



Section 4
Build a Dashboard with Plotly Dash

The pie chart of total success launch by site



Piechart for The Highest Launch Success Site



The highest total launches is KSC LC-39A site with 41.7%.

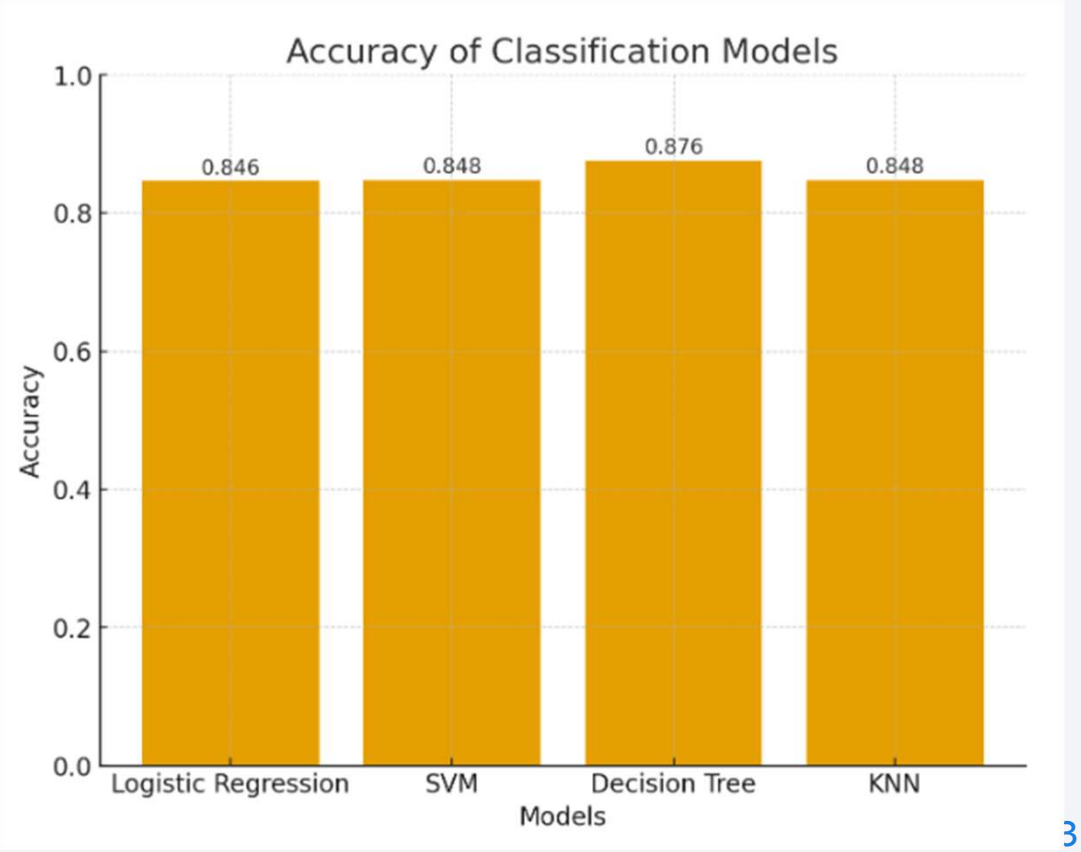
The scatter plot of payload and success class

Section 5

Predictive Analysis (Classification)

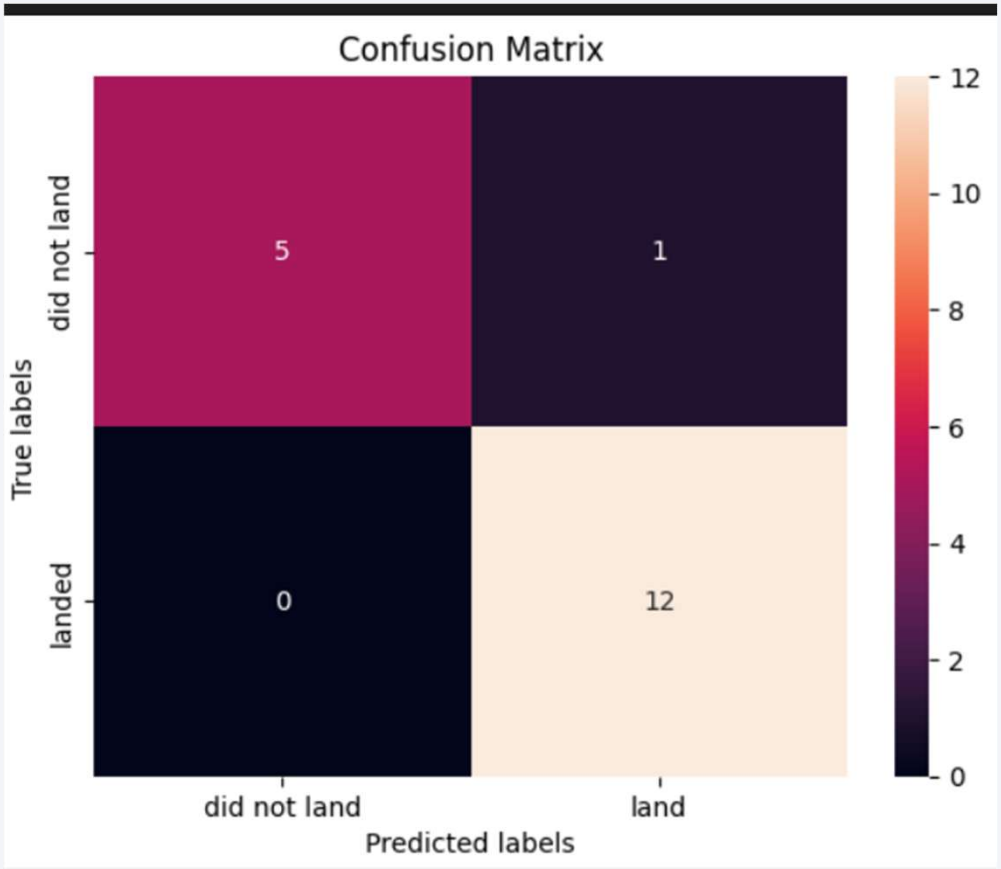
Classification Accuracy

- Bar graph for the built model accuracy for all built classification models is as shown in the figure.
- The decision tree model has highest accuracy.



Confusion Matrix

- This is the confusion matrix of the best performing model.
- The best performing model is Decision tree with model accuracy of 87 percent and accuracy on test as 94 percent



Conclusions

- While collecting data it is important to clean it first and perform data wrangling for further analysis.
- By visualizing the dataset we can get better outlook over the dataset.
- SQL queries gives us better scope to explore datasets in comparison with traditional EDA methods . Decision tree model is the most reliable since it has the highest out of sample accuracy of 87% and f1-score of 94%

Appendix

```
from js import fetch
import io

URL1 = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_
resp1 = await fetch(URL1)
text1 = io.BytesIO((await resp1.arrayBuffer()).to_py())
data = pd.read_csv(text1)
```

```
data.head()
```

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0

Thank you!