IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Aditya Raj
06-03-2024

Outline

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

Executive Summary

- The project focuses on analyzing the SpaceX Falcon 9 data and building a prediction model to predict if a particular rocket launch will successfully land or not.
- The entire presentation consists of the following sections:
  - Methodology
  - Insights from Exploratory Data Analysis
  - Launch Sites Proximities Analysis
  - Building a Dashboard with Plotly Dash
  - Predictive Analysis and Classification Model Building
- The outcome of each step were analyzed and stored in separate Jupyter notebooks, images, and CSV files, wherever applicable. All the files have been shared in the GitHub repository.

Introduction

- SpaceX is a leading private space agency that specializes in low-cost rocket launches. With its reusable first stage components of the Falcon 9 rockets, SpaceX is able to keep its operations costs almost one third of compared to other space agencies.
- This project aims to analyze how successful is the Falcon 9 first stage landing. It also tries to predict whether a particular rocket launch will result in a successful landing of the first stage of the Falcon 9 rockets.
- With different exploration tasks, the project also aims to analyze the effect of different factors such as launch site, payload mass, booster version, etc.

Section 1

Methodology

Methodology

Executive Summary

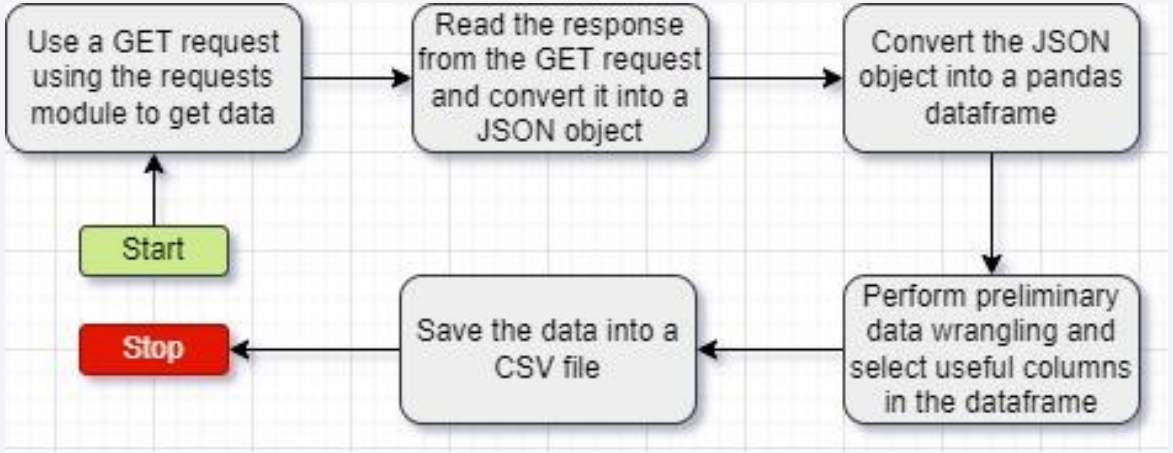
- Data collection methodology:
  - For the project, the data was collected using API calls to api.spacexdata.com using the requests module in Python. We also scraped data from Wikipedia page of Falcon 9 launches.
- Perform data wrangling
  - To process data, we used the pandas module in Python for analysis and feature engineering.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - To build classification models, we used the scikit-learn module in Python.

Data Collection

- The data was collected using two approaches.
  - The first approach uses the requests module in Python to fetch data from api.spacexdata.com.
  - The second approach uses the requests module and the BeautifulSoup module to fetch data from “List of Falcon 9 and Falcon Heavy launches” Wikipedia page.
- The data collection process has been presented in the next slides.

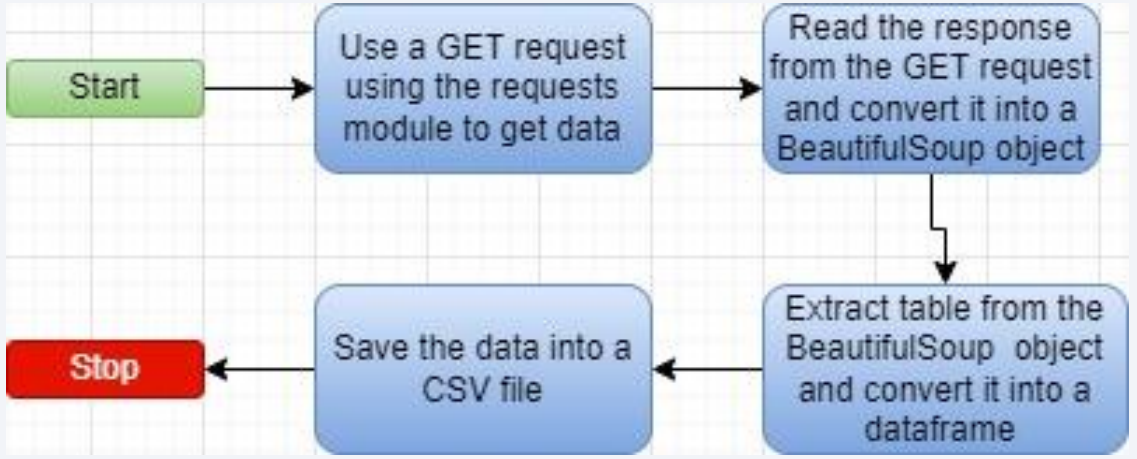
Data Collection – SpaceX API

- To collect data, we use the requests module in Python to make API calls to different endpoints at api.spacexdata.com.
- The Jupyter notebook for this task can be found at this link.



Data Collection - Scraping

- To get Falcon 9 data using web scraping, we scraped data from Wikipedia page "List of Falcon 9 and Falcon heavy launches". Then we extracted data from tables on the page into a CSV file.
- The Jupyter notebook for data collection with web scraping can be found at this link.



Data Wrangling

- After collecting the data, we performed data wrangling to get basic insights from the data and transform it into a usable format. For this, we performed the following tasks.
  - **Calculate the number of launches on each site.**
  - **Calculate the number and occurrence of each orbit**
  - **Calculate the number and occurrence of mission outcome of the orbits**
  - **Create a landing outcome label from Outcome column for building classification models.**

- The Jupyter notebook for this task can be found at this link.

EDA with Data Visualization

- After data wrangling, we proceeded with data visualization to find answers to our queries on how different factors affect the success of rocket landings.
  - **To visualize the relationship between success rate of each orbit type, we used a bar chart.**
  - **To Visualize the relationships between Payload and Orbit type, Flight Number and Orbit type, Flight Number and Launch Site, we used a scatter plot.**
  - **To visualize the launch success yearly trend, we used a line chart.**
- **The Jupyter Notebook for this task can be found at this link.**

EDA with SQL

- After data visualization, we used SQL to gain deeper insights into the data. For this, we used the following queries.
  - ***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 records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.***
  - ***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.***
- *The Jupyter Notebook for the task can be found at this link.*

Build an Interactive Map with Folium

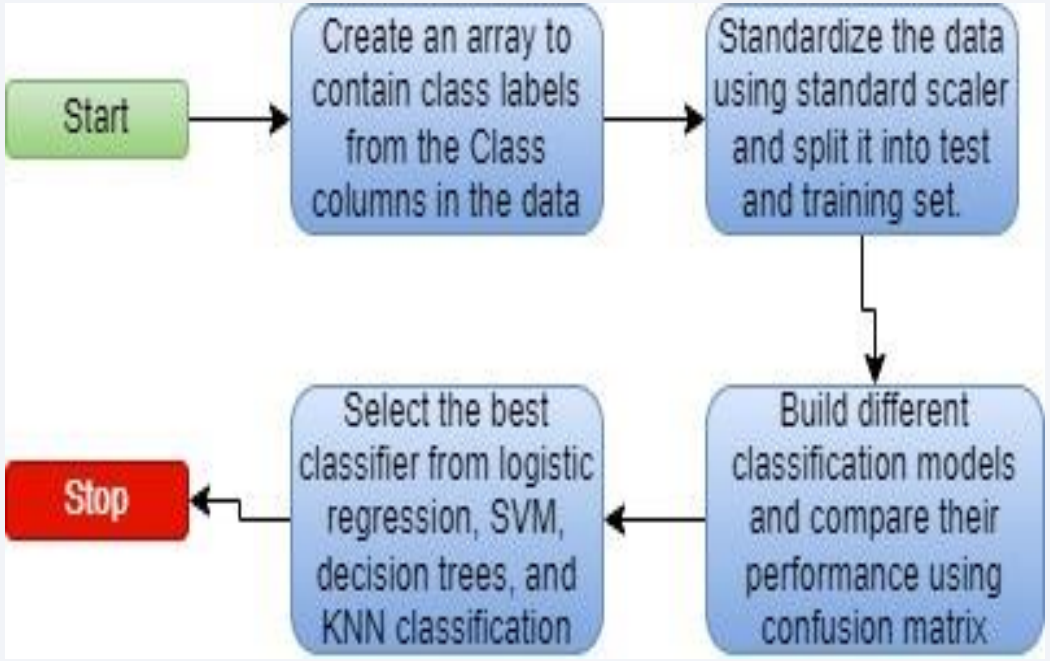
- To build interactive maps with Folium, I completed the following tasks.
  - Mark all launch sites on a map.
  - Mark the success/failed launches for each site on the map.
  - Calculate the distances between a launch site to its proximities
- For this, I used different folium objects such as Map to create the map object, Circle to add a highlighted circle area with a text label on a specific coordinate, Marker to mark a particular coordinate, and MarkerCluster to group multiple markers.
- The Jupyter notebook for this activity can be found at this link.

Build a Dashboard with Plotly Dash

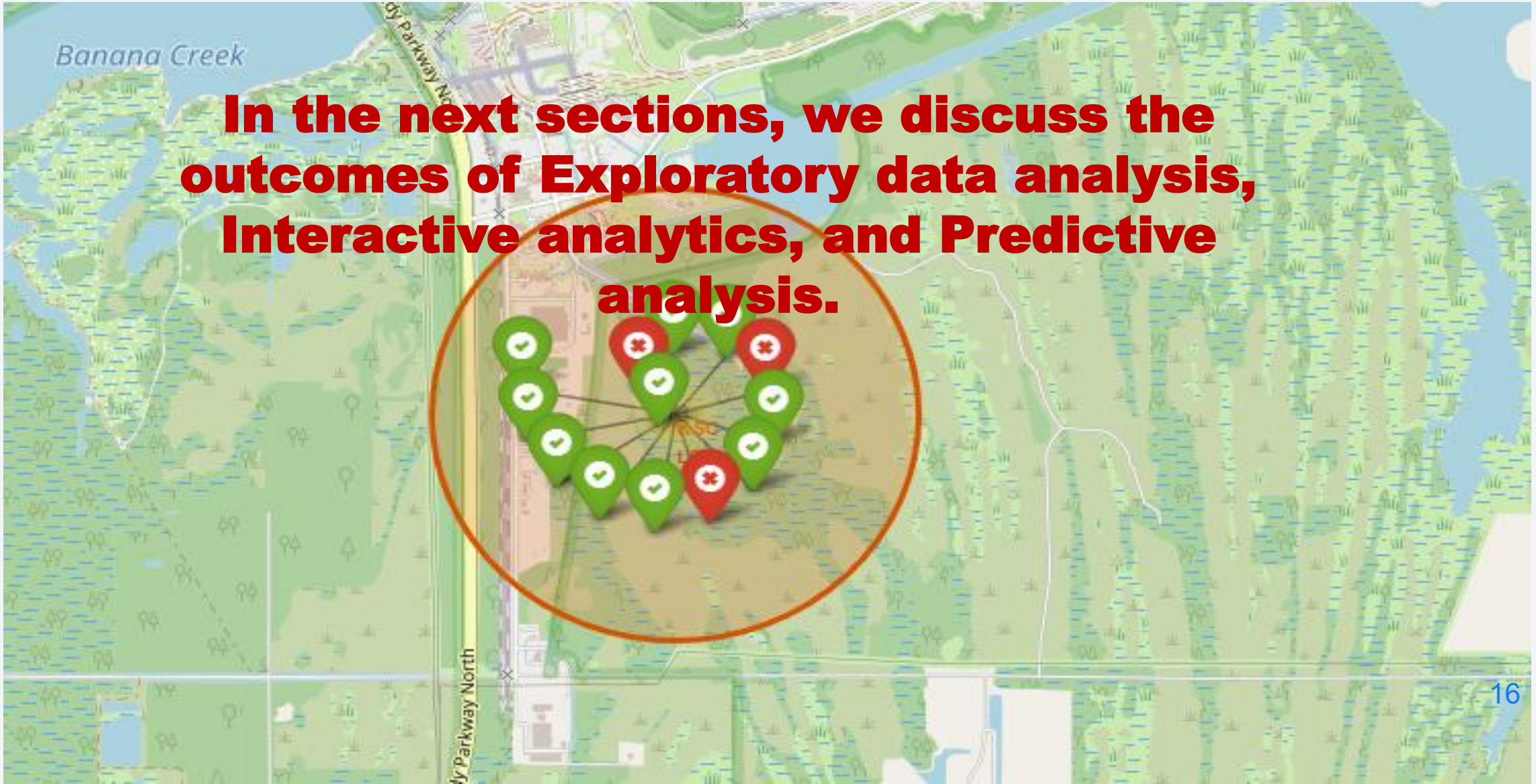
- To get better insights into the data, I created a dashboard to visualize the success vs failed launches for each SpaceX launch site. It also uses a pie chart to show the total successful launches count for all sites. I have also used a scatter chart to show the correlation between payload and launch success for each site.
- The Python file for this activity can be found at this link.

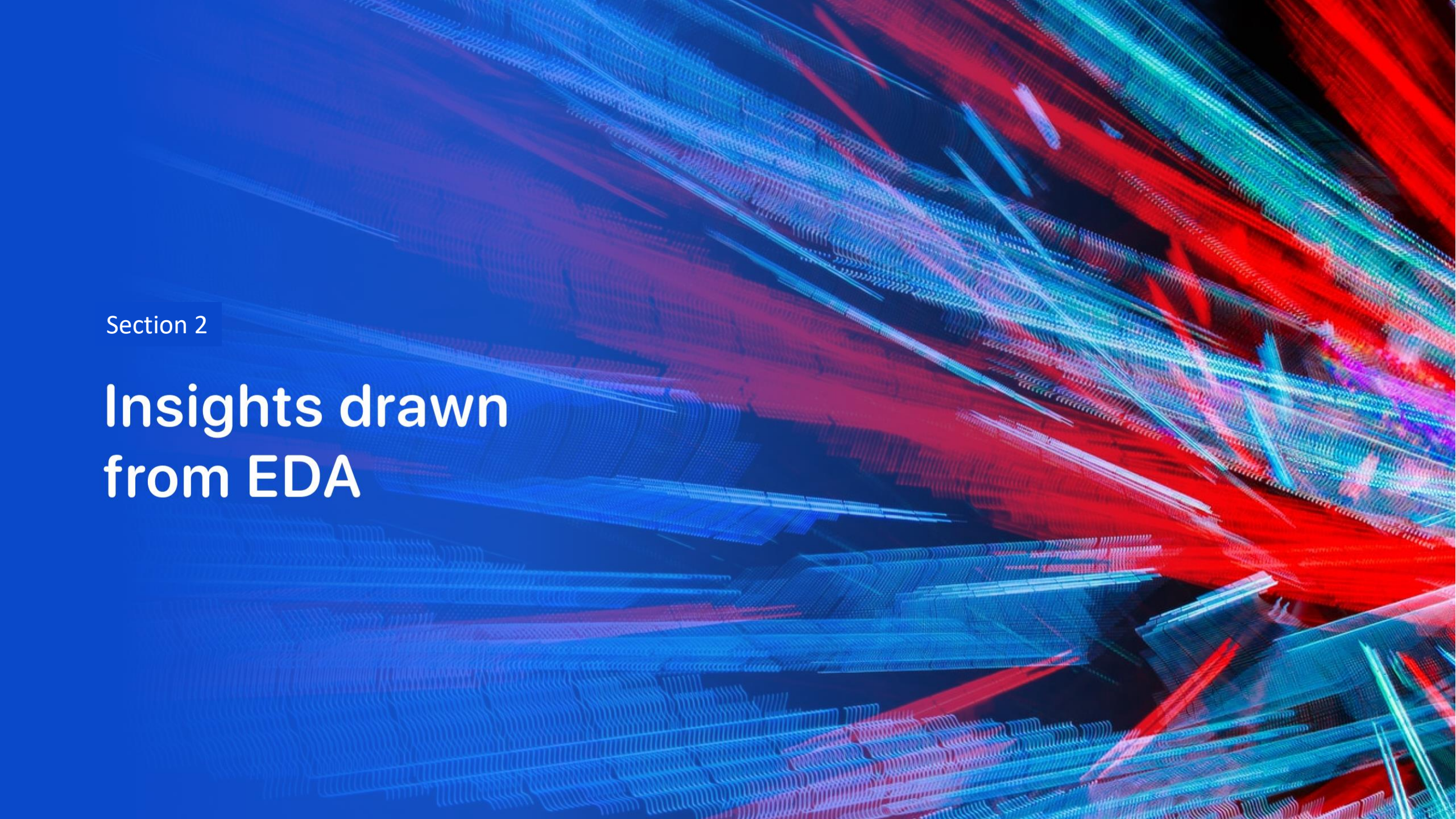
Predictive Analysis (Classification)

To build a classification model, we first created an array to contain class labels. Then, we used standard scaling algorithm to standardize the data. Next, we used different algorithms such as Logistic regression, Support vector machines, decision trees, and K-Nearest Neighbors algorithm to find the best classifier for predicting whether a rocket launch and landing will successful or not based on different factors. The Jupyter notebook for this activity can be found at this link.



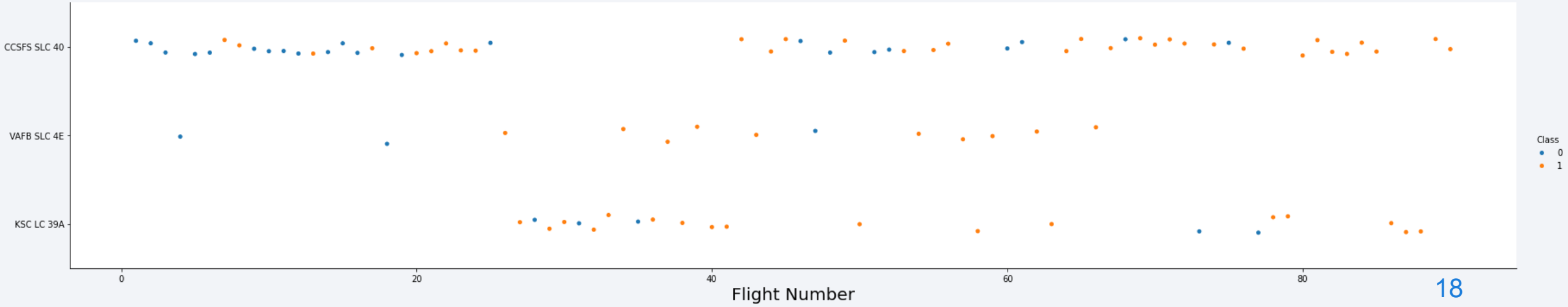
Results


Banana Creek
In the next sections, we discuss the outcomes of Exploratory data analysis, Interactive analytics, and Predictive analysis.


Section 2
Insights drawn from EDA

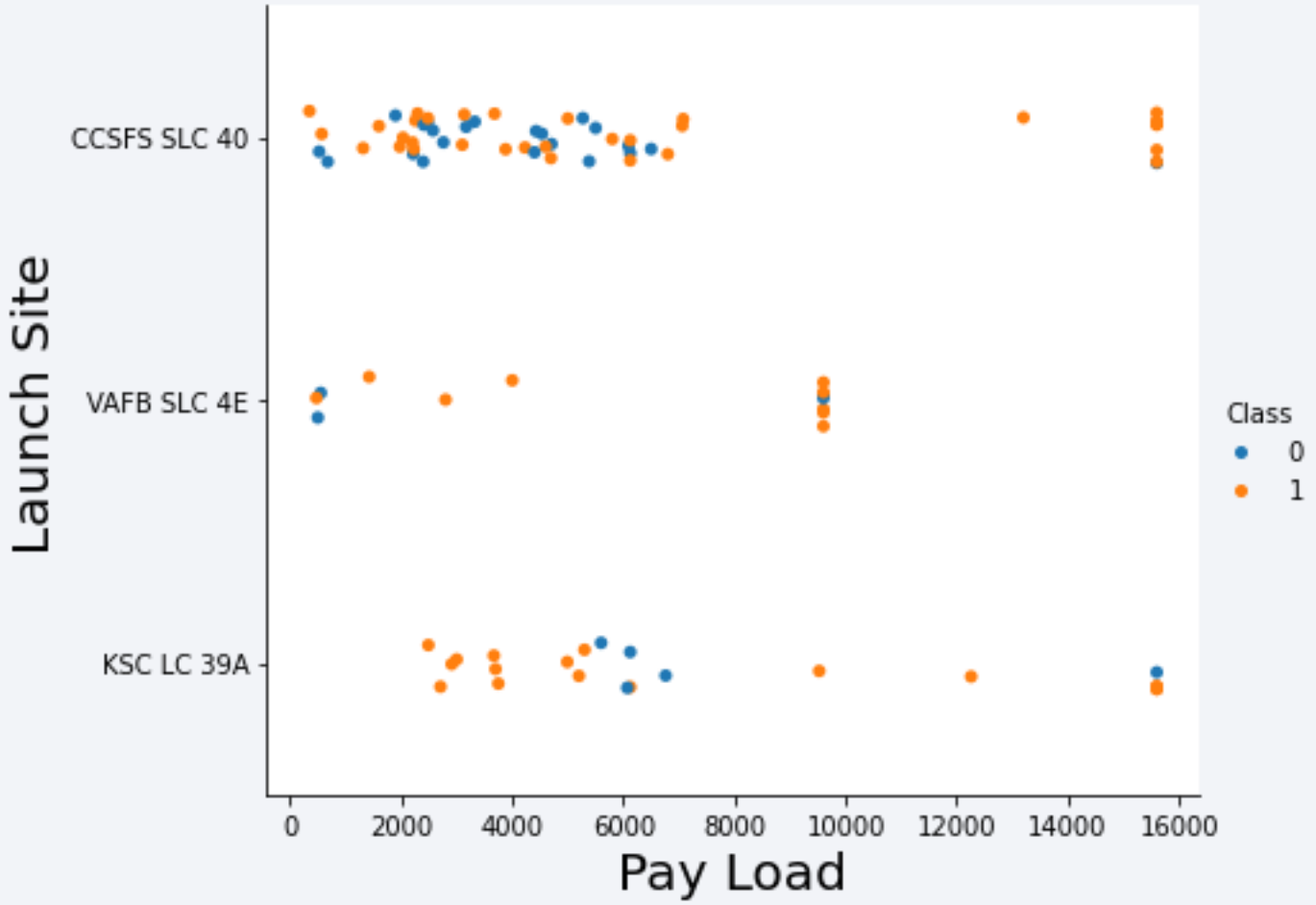
Flight Number vs. Launch Site

- The analysis on Flight Number vs Launch site shows that the site CCSFS SLC 40 was used to launch majority of rockets and VAFB SLC 4E was used to launch the least number of rockets as shown in the below scatter plot.



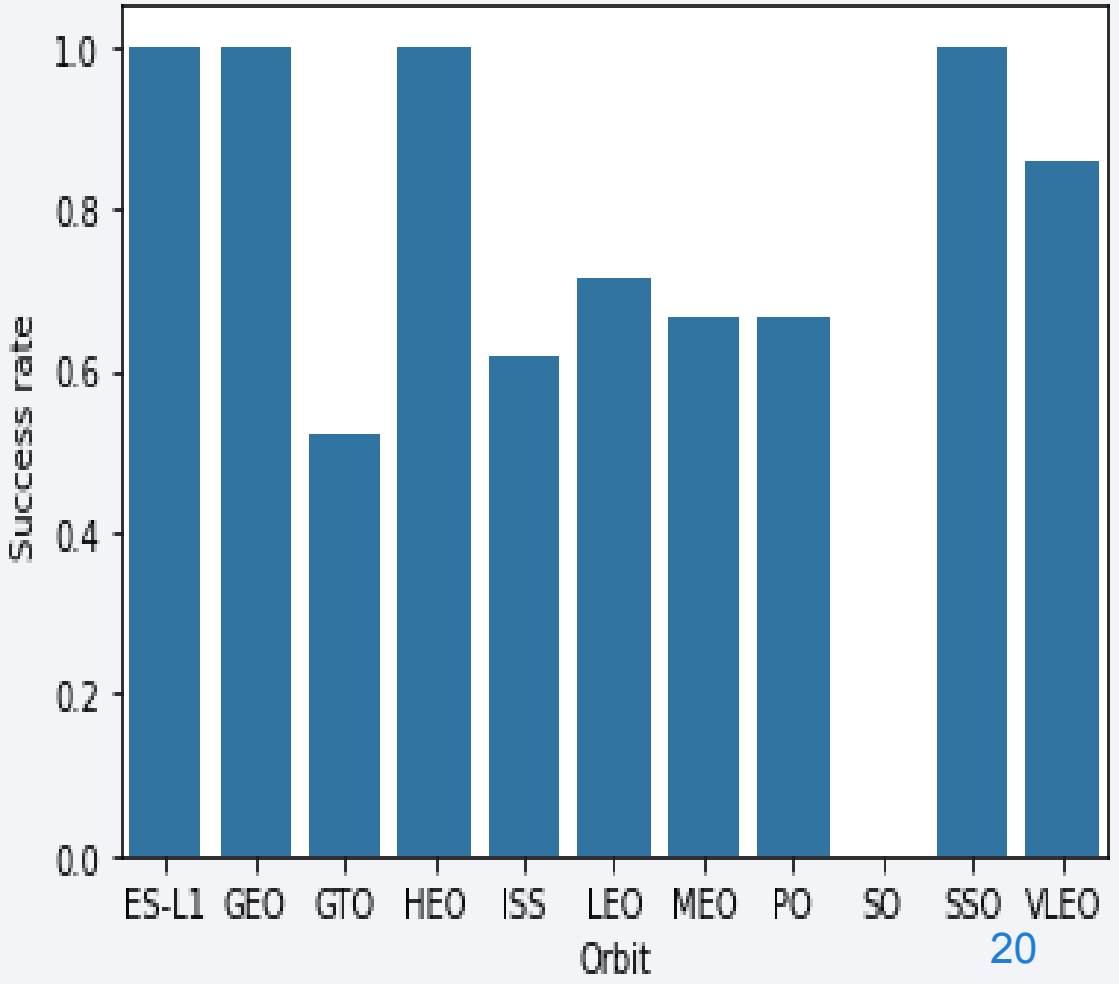
Payload vs. Launch Site

- The scatter plot on payload vs Launch site shows that the site CCSFS SLC 40 launched most of the rockets with payload less than 8000 Kgs. The site KSC LC 39A launched most of the rockets with payloads between 2000 to 7000 Kgs.



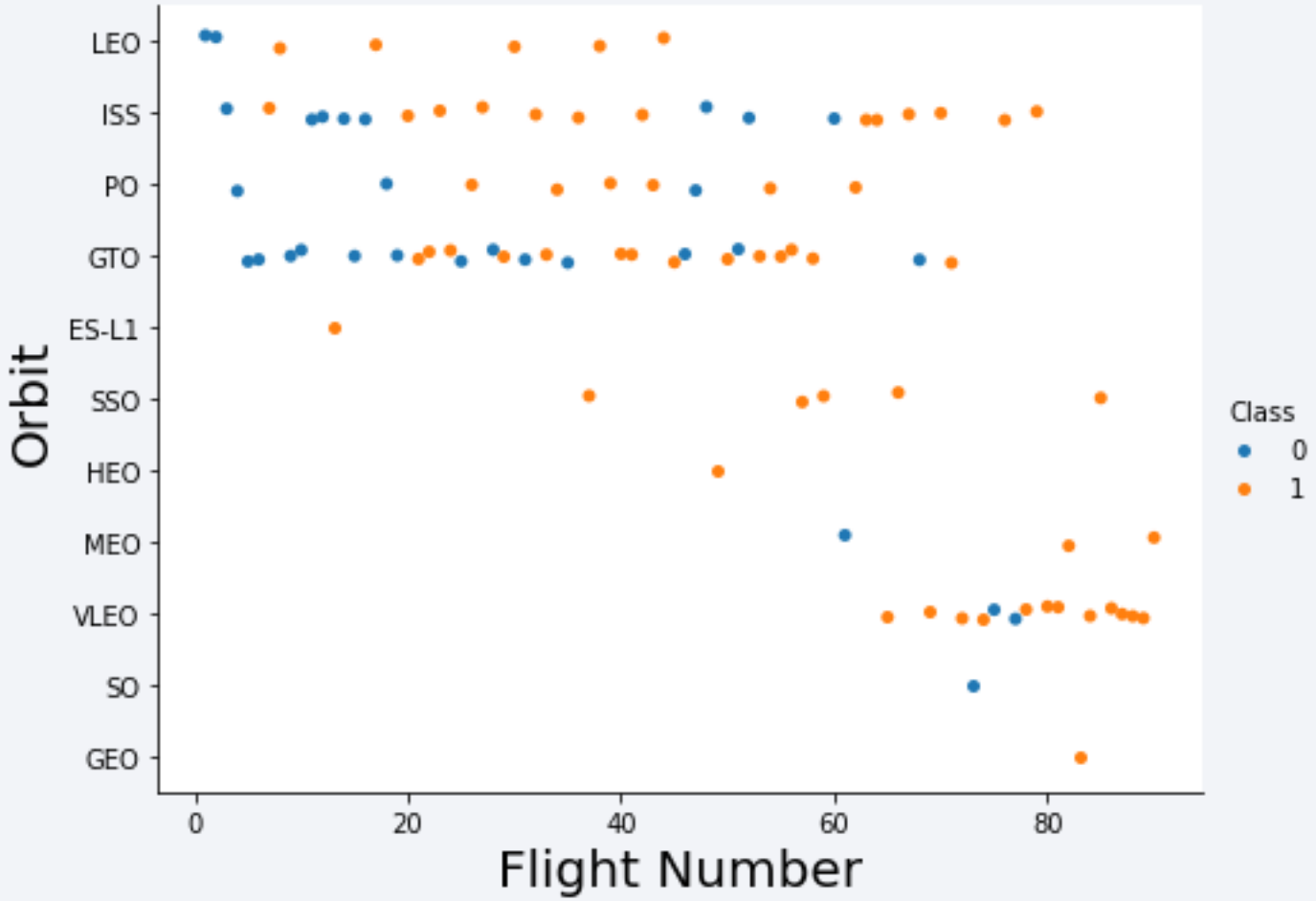
Success Rate vs. Orbit Type

- The bar chart on success rate vs orbit type shows that launches to 4 orbits namely ES-L1, GEO, HEO, and SSO have 100 percent success rate whereas launches to SO have 0 percent success rate.



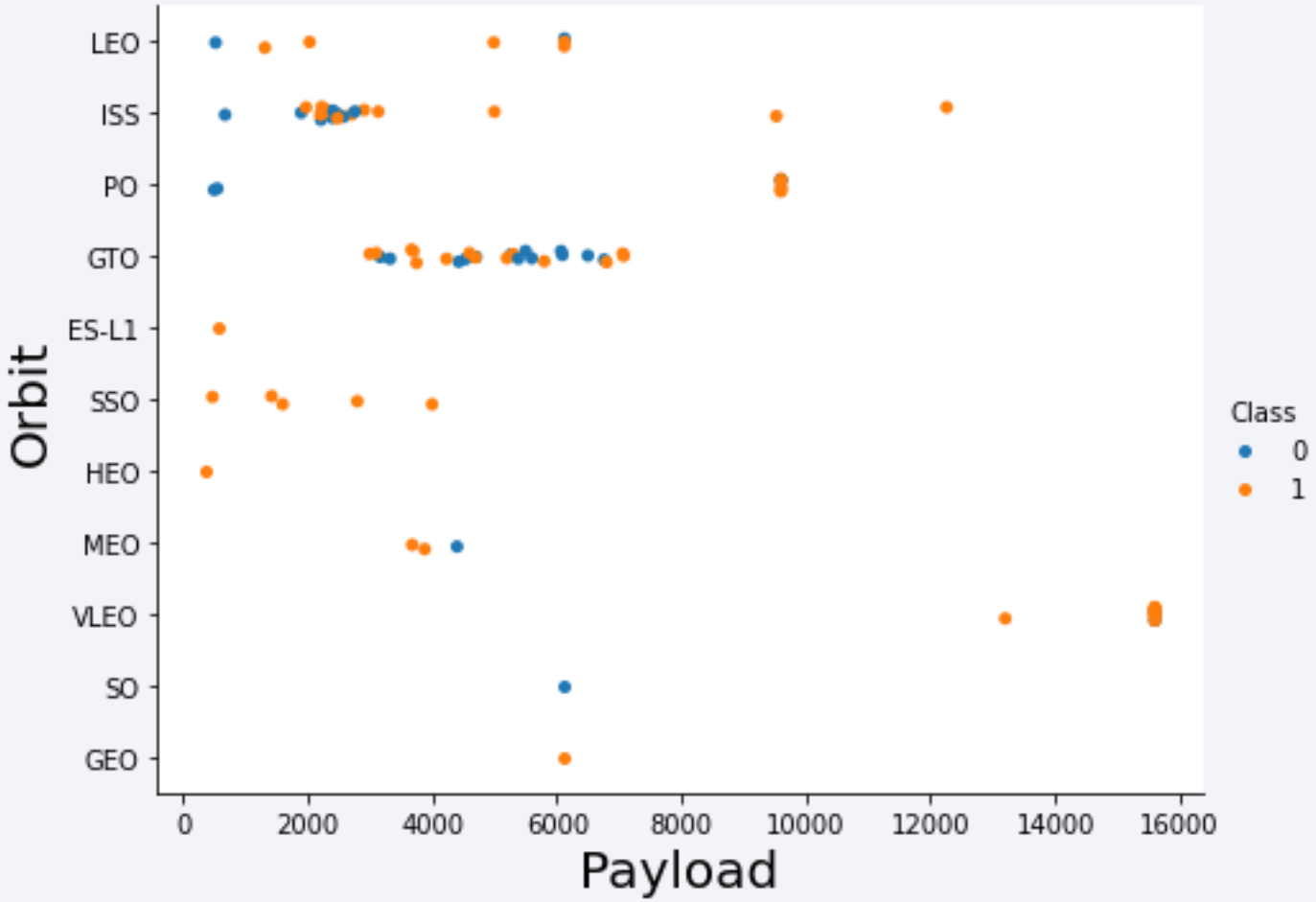
Flight Number vs. Orbit Type

- The scatter on Flight number vs Orbit type shows tat most of the launches are for LEO, ISS, PO, AND GTO. On the other hand, orbits ES-L1, HEO, SO, and GEO have only one launch.



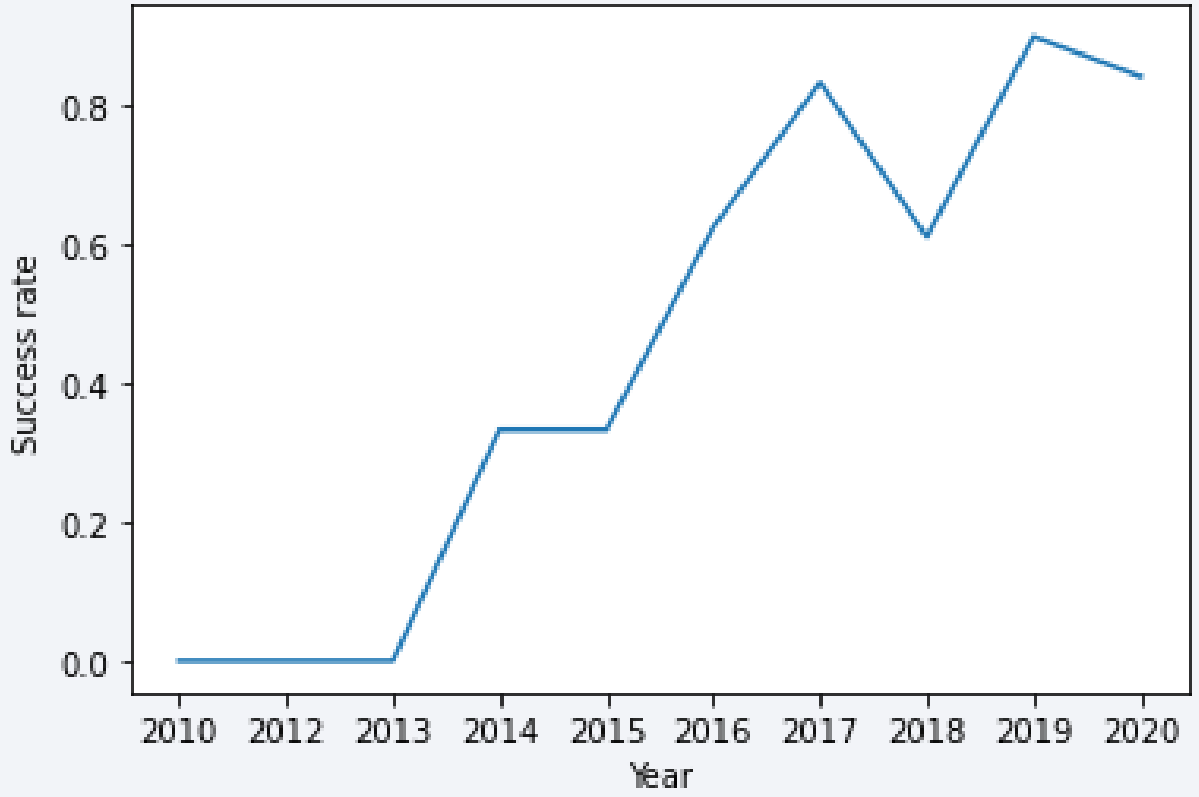
Payload vs. Orbit Type

- The scatter plot on payload vs orbit type shows that most of the launches are of less than 10000 kg payload.
- The ISS orbit has most launches between 3000 to 5000 kg whereas the GTO orbit has payloads in the 3000 to 8000 kg range.
- The VLEO orbit has two flights with 16000 kg payload.



Launch Success Yearly Trend

- The line chart for launch success yearly trend shows that SpaceX had no success in landing the first stage of rockets till 2013. After that, it has significantly increased its success rate each year, with a slight dip in the year 2018.



All Launch Site Names

- To get the all launch site names, we have used distinct keyword with the select statement. The results are as follows.

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

```
]: %sql select distinct Launch_Site from SPACEXTABLE;
```

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

]:

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

Launch Site Names Begin with 'CCA'

- To find 5 records where launch sites begin with `CCA`, I have used the like() operator and the limit keyword to get only 5 rows.

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

Total Payload Mass

- To calculate the total payload carried by boosters from NASA, I have used the where keyword and the sum() function.

Task 3

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

```
%sql select sum(PAYLOAD_MASS__KG_) as total_mass from SPACEXTABLE where Customer='NASA (CRS)';
```

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

total_mass
45596

Average Payload Mass by F9 v1.1

- To calculate the average payload mass carried by booster version F9 v1.1, I have used the avg() function and where operation with the select statement.

Task 4

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

```
%sql select avg(PAYLOAD_MASS__KG_) as avg_mass from SPACEXTABLE where Booster_Version='F9 v1.1';
```

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

avg_mass
2928.4

First Successful Ground Landing Date

- To find the date of the first successful landing outcome on ground pad, I have used the min() function and the where condition.

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

Hint:Use min function

```
%sql select min(Date) as date from SPACEXTABLE where Mission_Outcome='Success';
```

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

date
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

- To list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000, I have used the and operator in the where condition to apply required conditions.

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql select Booster_Version 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
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- To calculate the total number of successful and failure mission outcomes, I have used the groupby() function with the count() function.

List the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome,count(*) as number from SPACEXTABLE group by Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

- To list the names of the booster which have carried the maximum payload mass, I have used a subquery to first get the maximum payload. Then, I used the where condition with distinct operation to get the desired results.

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
8]: %sql select distinct Booster_Version as Booster_Versions from SPACEXTABLE\
    where PAYLOAD_MASS__KG_ in (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);
```

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

8]:

Booster_Versions
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

- To list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015, I used the substr() function to get the launch year and used it in the where condition and the select statement.

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql select substr(Date, 6,2) as Month, Mission_Outcome, Booster_Version,Launch_Site, Landing_Outcome from SPACEXTABLE\
where Landing_Outcome='Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

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

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

- 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, I have used order by operation and the count() function.

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.

```
%sql select Landing_Outcome, Count(*) as number from SPACEXTABLE\
where Date between '2010-06-04' AND '2017-03-20' group by Landing_Outcome order by number desc;
```

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

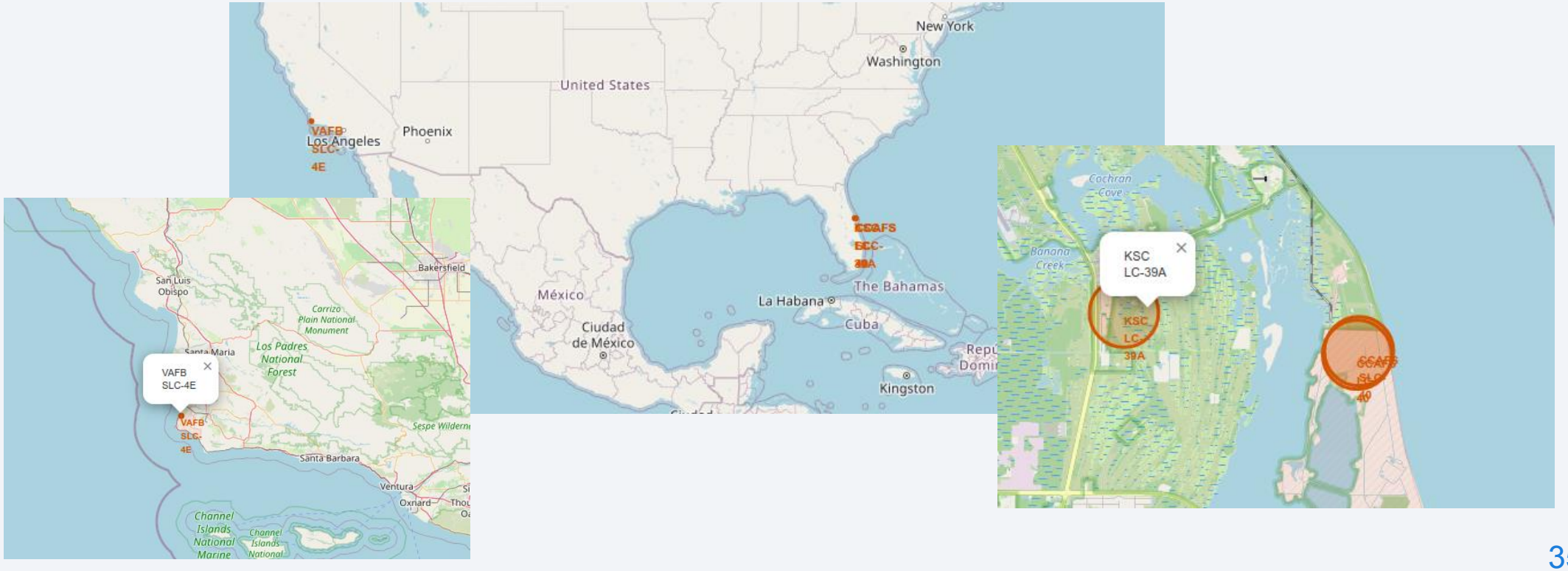
Landing_Outcome	number
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

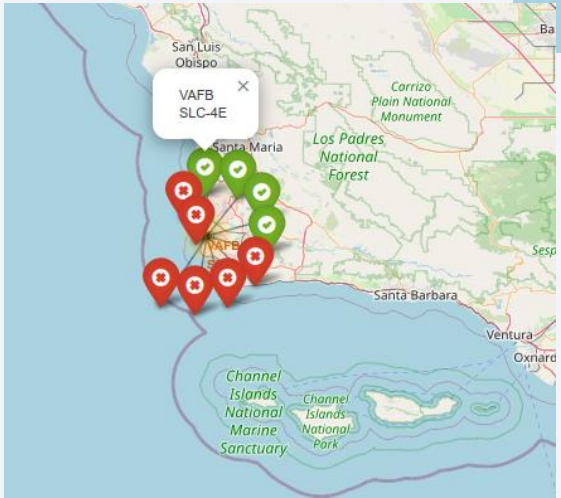
SpaceX Launch Sites

- As we can see, all the launch sites are in coastal regions close to the equator. This helps SpaceX launch rockets efficiently and land the first stage of rockets into water bodies.



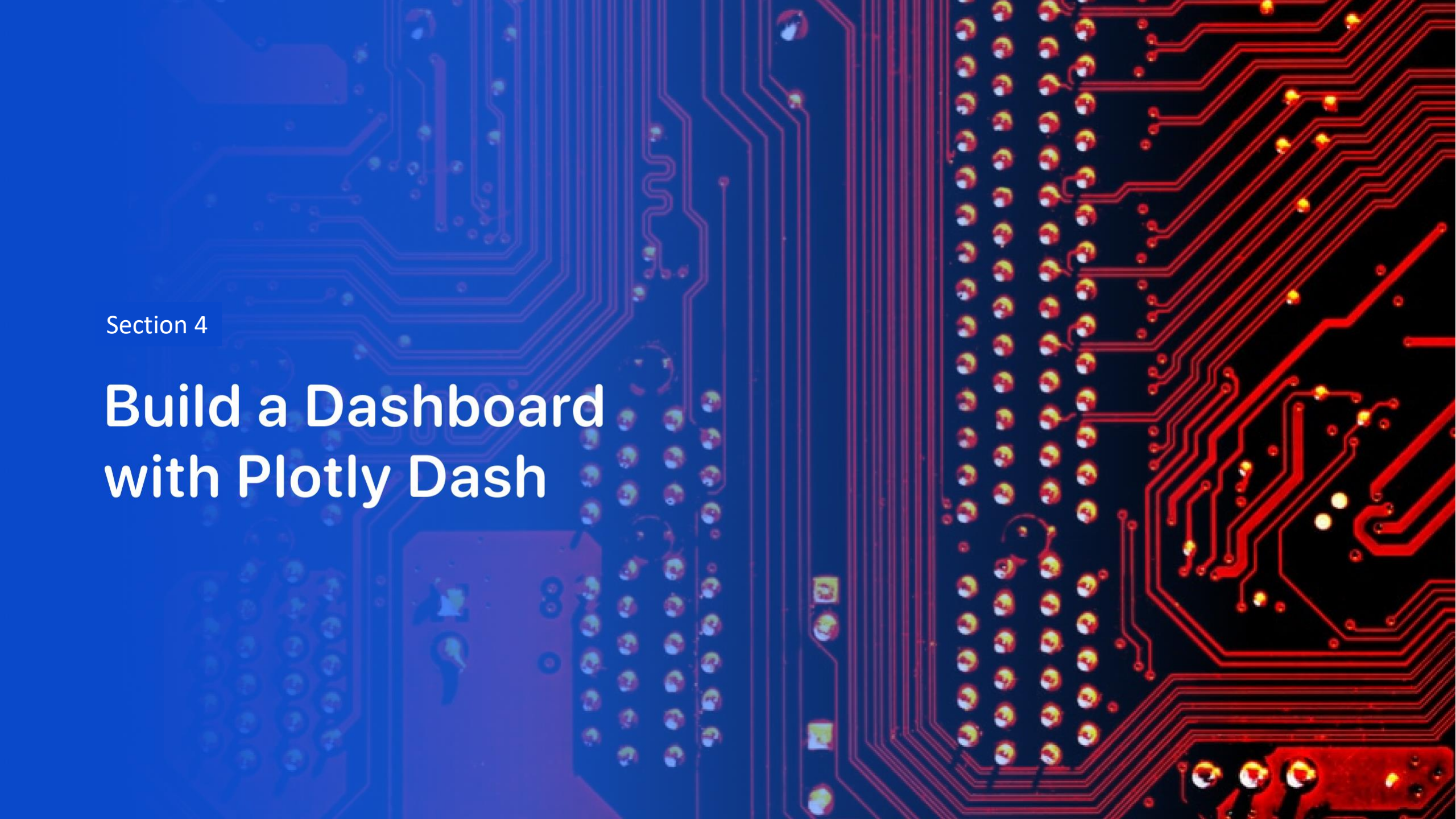
Success vs Failed Launches by SpaceX

- In the output map, we can observe that the site CCAFS LC-40 has a low success rate whereas the site KSC LC-39A has a high success rate.



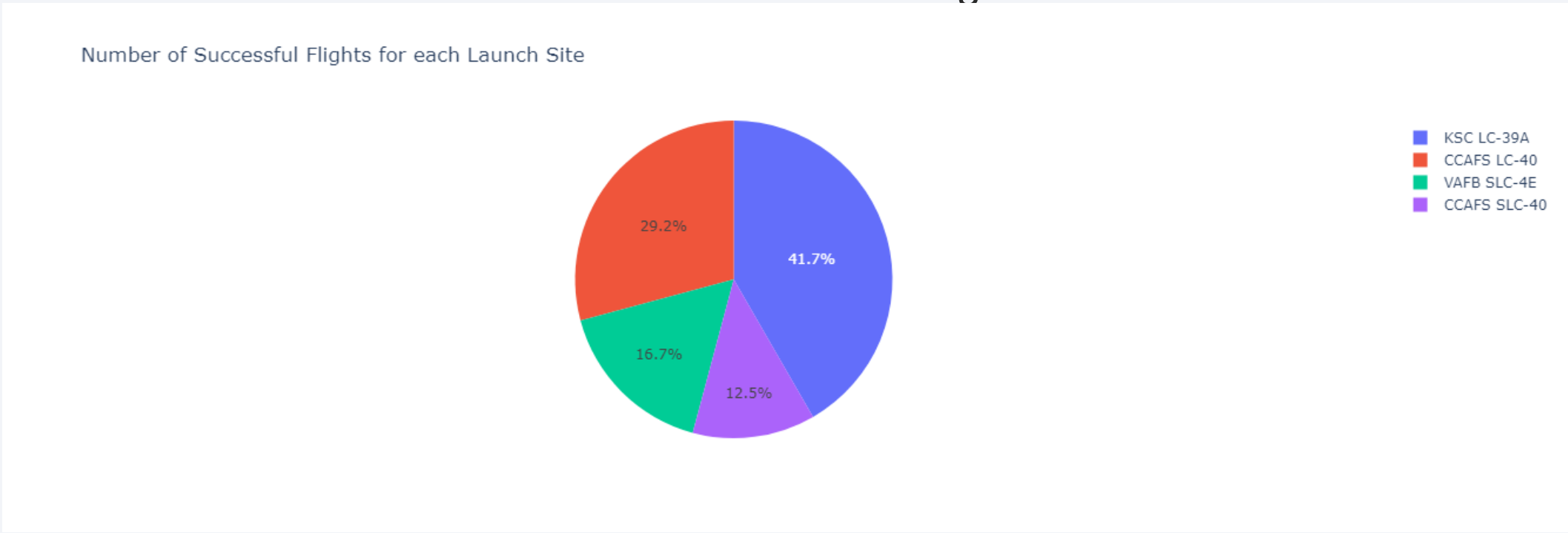
Nearest Coastline to VAFB SLC-4E

- In the output map, we can see that the nearest coastline to the Launchpad VAFB SLC-4E is only 1.3 KMs away.


Section 4
Build a Dashboard with Plotly Dash

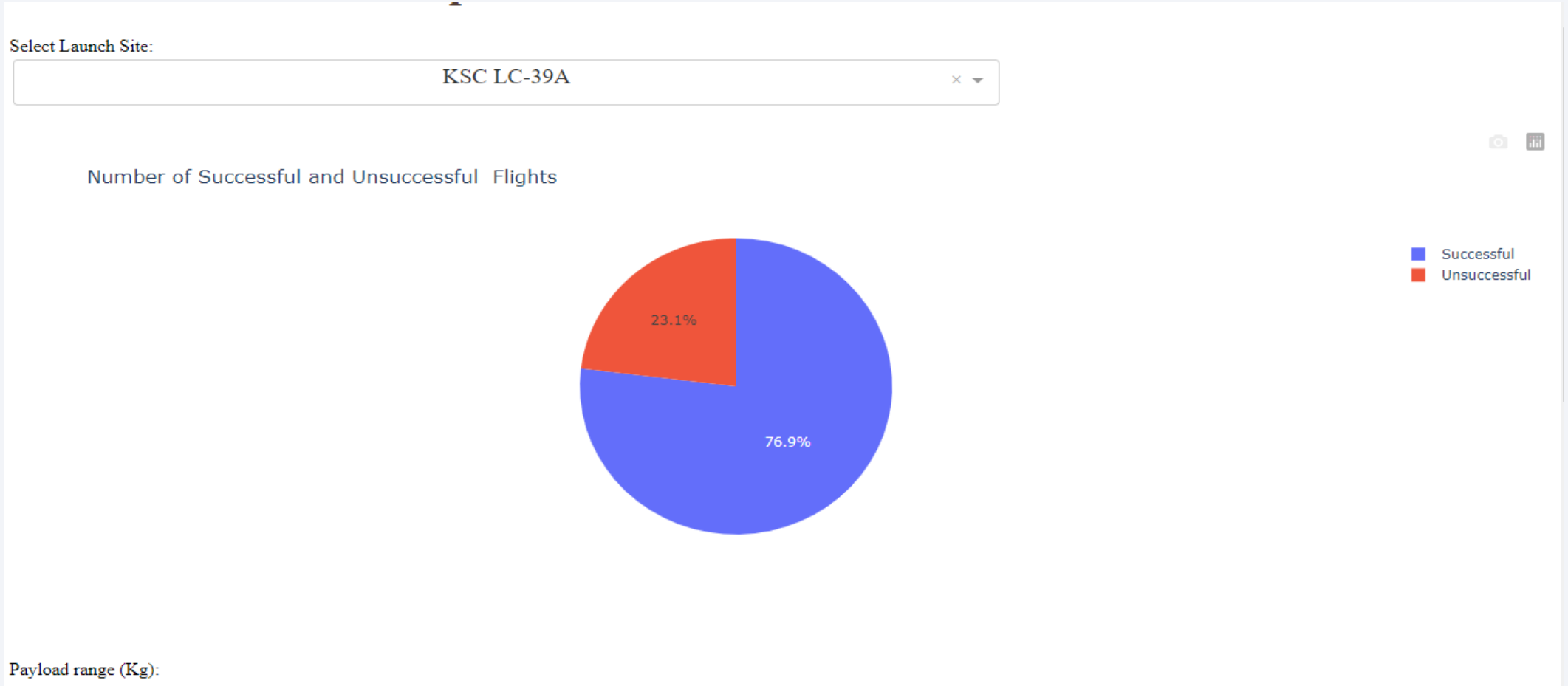
'Number of Successful Flights for each Launch Site'

- The pie chart shows that the launch site KSC LC-39-A has highest percentage of successful flight. On the contrary, the launch site CCAFS SLC-40 has the lowest number of successful flights.



Percentage of Successful and Unsuccessful flights at KSC LC-39A

- The launch site KSC LC-39A has the highest success percentage at 6.9 percent.



Payload and Success Chart

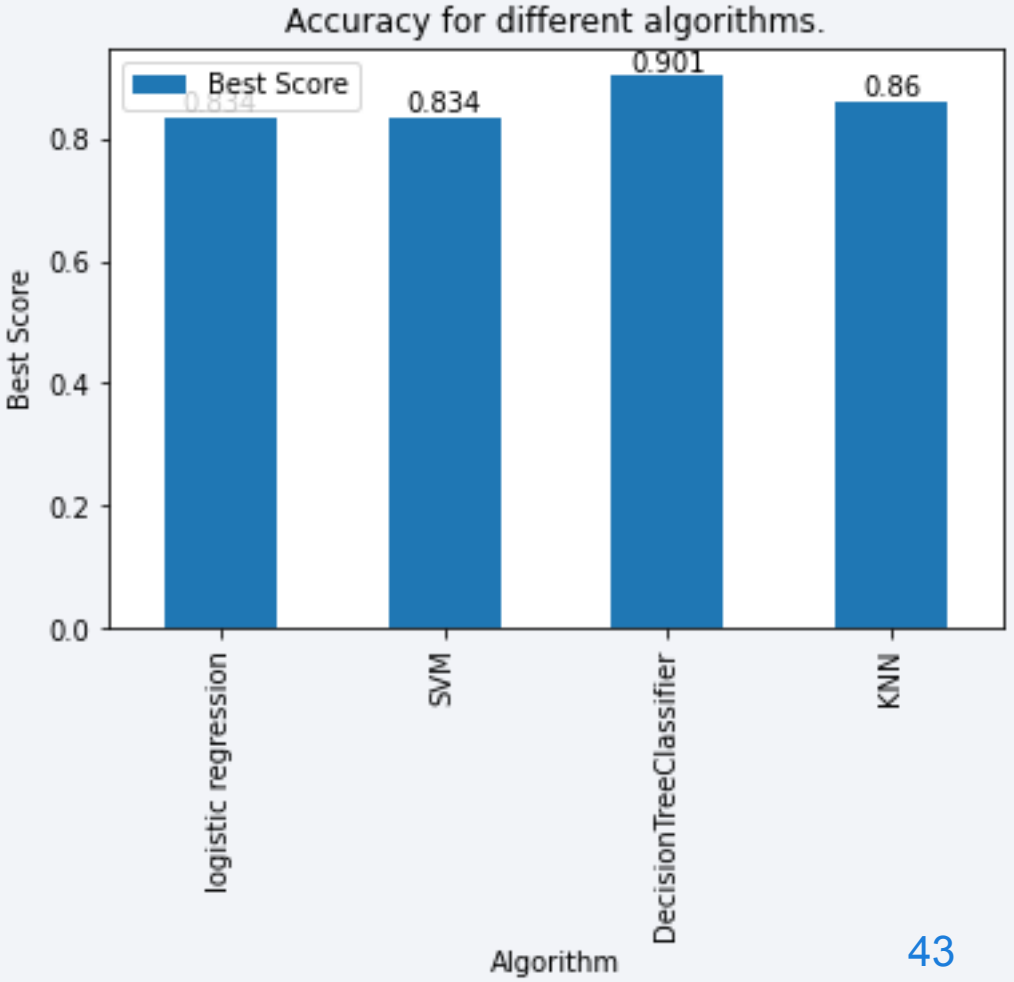
- In the output plot, we can observe that the booster version FT has the highest overall success rate.

Section 5

Predictive Analysis (Classification)

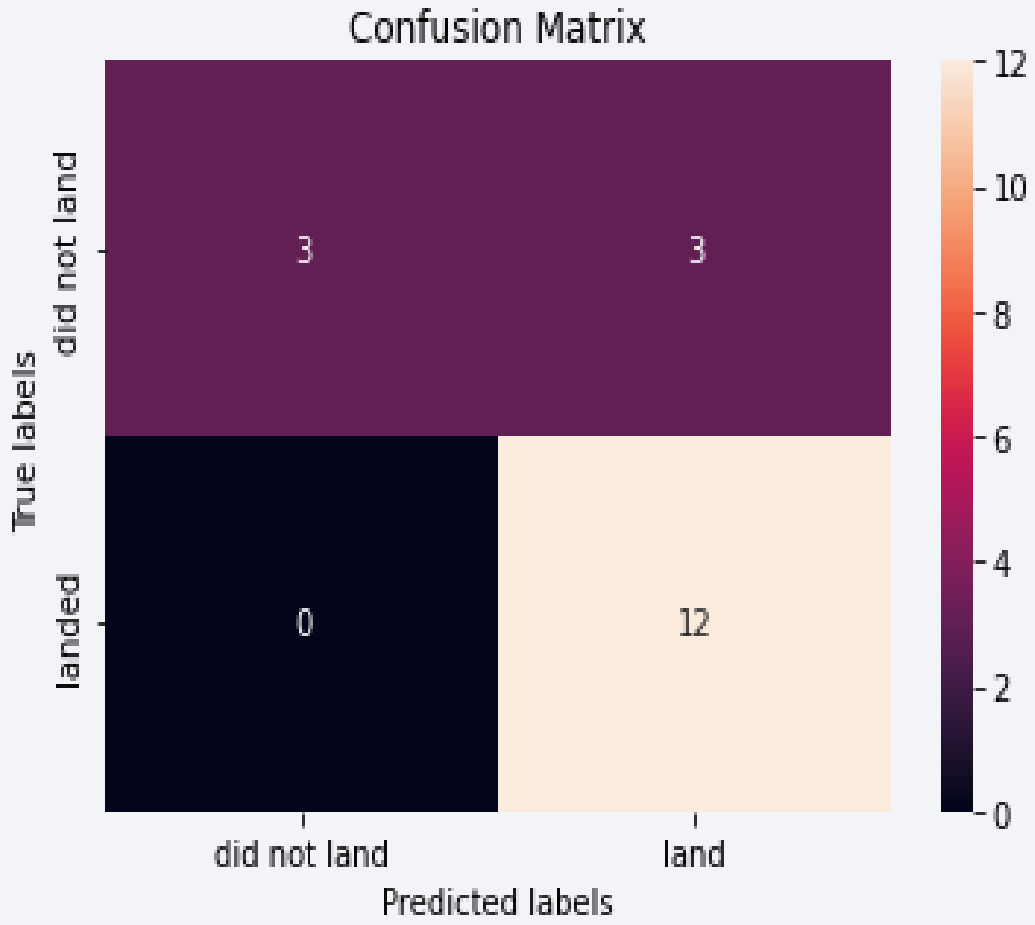
Classification Accuracy

- By looking at the bar chart, we can observe that the decision tree classifier has the highest accuracy on training data whereas the logistic regression and the SVM classifier have the lowest accuracy.
- On the test data, All the models showed the same accuracy.



Confusion Matrix

- On the test data, all the models showed similar results as captured in the confusion matrix.



Thank you!