IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Sheba Rachel Varghese>
<16/03/2024>

Outline

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

Executive Summary

- This project follows a robust methodology encompassing various stages of data analysis and predictive modeling. Data collection was executed through the utilization of ***API*** and ***Web Scraping*** techniques, coupled with meticulous ***Data Wrangling***.
- The subsequent Exploratory Data Analysis (EDA) leveraged ***SQL, Pandas****, and* ***Matplotlib*** for a comprehensive understanding of the dataset.
- The findings were visually represented through ***Interactive Visual Analytics and Dashboards***, offering stakeholders an intuitive grasp of key insights.
- Additionally, the project incorporated Predictive Analysis using advanced ***Machine Learning models****,* enhancing the ability to make informed decisions based on data-driven predictions.
- The successful implementation of data collection from APIs and web scraping, coupled with an in-depth analysis of essential dataset features, underscores the project's commitment to robust data handling. Multiple machine learning models were trained, ensuring the highest predictive accuracy. ***Decision tree performed the best*** out of all models.

Introduction

In the dynamic landscape of space exploration, SpaceX has emerged as a pioneering force, achieving remarkable success in launching rockets at a comparatively low cost. This project is framed against the backdrop of SpaceX's achievements, aiming to apply insights from their launches to predict the launch costs of a new entrant, Space Y.

- **Project Background and Context:** SpaceX's proven success in cost-effective rocket launches serves as an invaluable foundation for this endeavor. The objective is to leverage historical data from SpaceX launches to construct a predictive model for estimating the cost associated with each launch conducted by Space Y.
- **Problems to Address:**

➢ ***Determine the Price of Each Launch:*** Establish a predictive framework to estimate the cost of individual launches by Space Y, drawing insights from SpaceX's cost-effective model.

➢ ***Identify Crucial Predictive Features and Success Rate:*** Uncover the most influential factors that contribute to predicting launch costs and success rates, offering a strategic advantage for Space Y's planning and decision-making.

➢ ***Explore Interrelations Between Factors:*** Investigate potential correlations and relationships between various factors, shedding light on the complex interplay influencing launch costs and success rates.

By addressing these challenges, this project aims to provide actionable insights that can empower Space Y in its pursuit of cost-effective and successful rocket launches.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology

- SpaceX API: Data extraction from the SpaceX API at https://api.spacexdata.com/v4/launches/past provided a foundational dataset comprising pertinent details about past launches.
- Web Scraping Wikipedia: Additional data enrichment was achieved through web scraping the Wikipedia page at https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches, ensuring a comprehensive dataset for analysis.

Perform data wrangling

- A dedicated column was introduced to represent the landing outcome, enhancing the dataset's completeness and relevance.

Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary (Contd.)

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models:

- The dataset was split into training and testing sets and subjected to evaluation through four classification models(Logistic Regression, SVM, Decision Tree and KNN). By employing GridSearchCV for parameter tuning, this methodology ensures that the classification models are fine-tuned to their optimal configurations, enhancing their ability to make accurate predictions on new and unseen data.
- The performance of each model was assessed using relevant evaluation metrics such as accuracy and confusion matrix. These metrics provided a comprehensive understanding of how well the models performed in classifying and predicting outcomes.

Data Collection

- **API Integration**

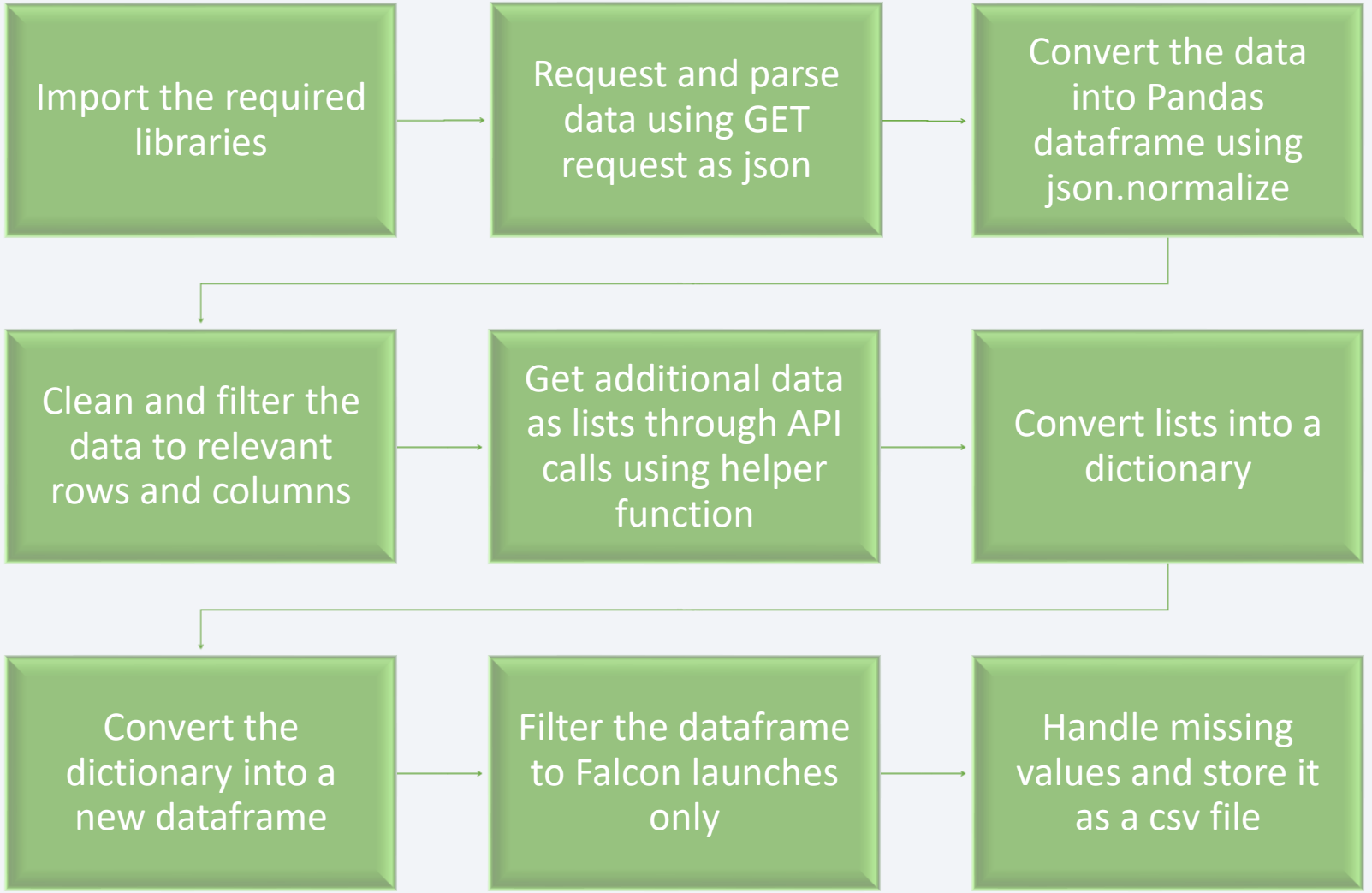
Utilized the SpaceX API (https://api.spacexdata.com/v4/launches/past) to access and retrieve information about past launches conducted by SpaceX. The API served as a structured and direct source of data, providing details such as launch dates, mission success, and other relevant parameters.

- **Web Scraping Wikipedia**

Employed web scraping techniques to extract additional information from the Wikipedia page (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches). The scraping process focused on gathering supplementary data related to Falcon 9 and Falcon Heavy launches, complementing the information obtained from the API.

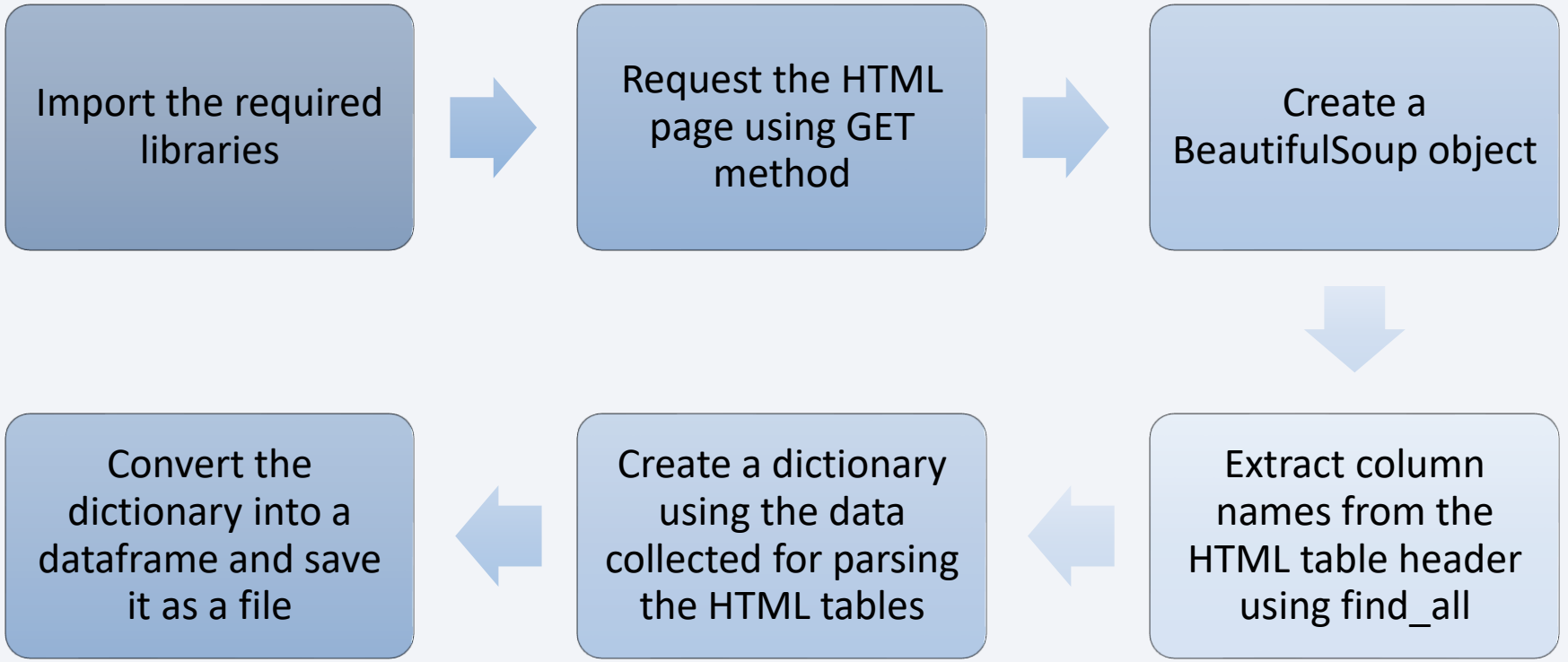
Data Collection – SpaceX API

- GitHub URL - SpaceX-API



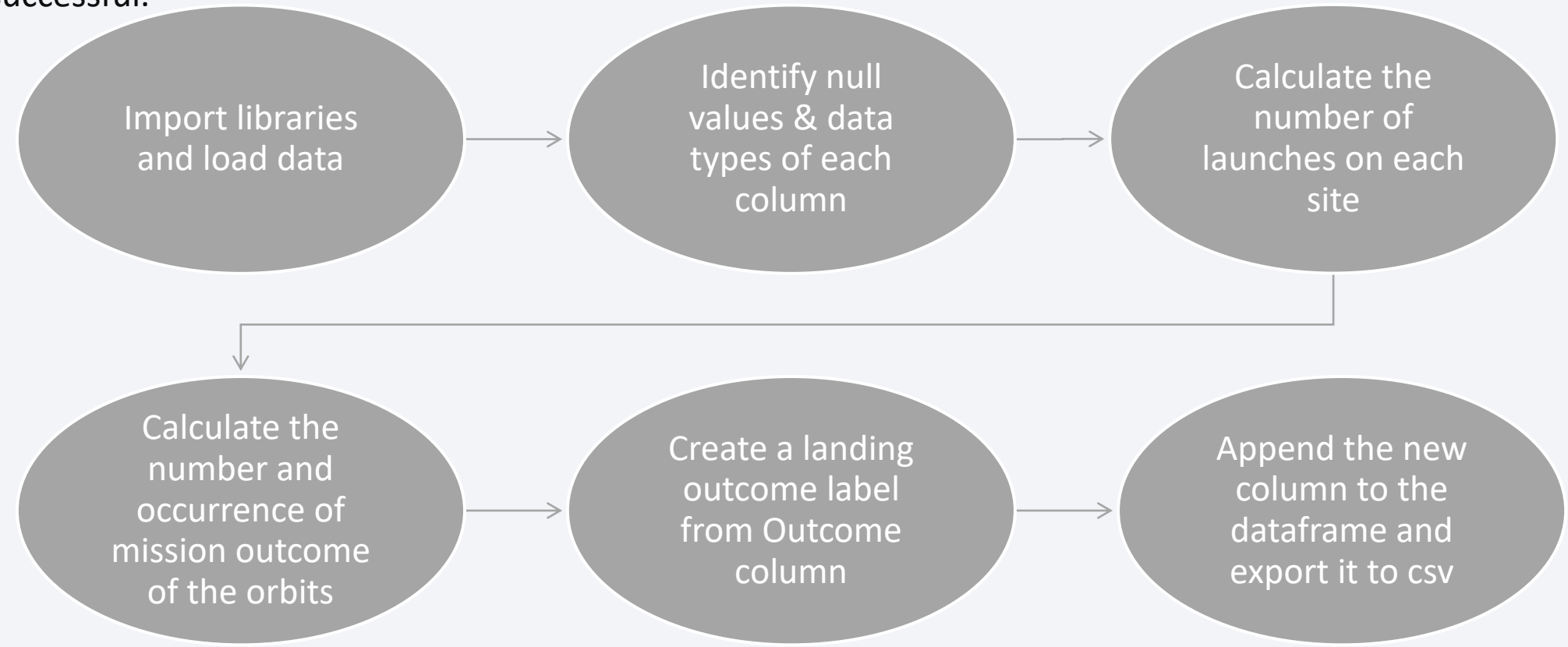
Data Collection - Scraping

- GitHub URL - Web Scraping



Data Wrangling

- Through the process of Exploratory Data Analysis (EDA), we identify patterns in the data. Subsequently, a new variable is incorporated, which serves as a training label by indicating whether the landing was successful or unsuccessful.



- GitHub URL : Data Wrangling

EDA with Data Visualization

- Scatter plots, bar charts, and line charts were used to gain a comprehensive understanding of the dataset. The plots mainly focused on the relationship between the variables : Class, Flight Number, Launch site, payload mass, success rate, orbit type, and also, yearly trends. This helps in determining how these features would impact the success rate.
- Each chart type was strategically chosen based on the nature of the data and the specific insights sought.
    - Scatter plots are ideal for identifying patterns and relationships between pairs of variables, offering insights into potential correlations.
    - Bar charts are effective for presenting categorical data and facilitating easy comparisons between different elements in the dataset.
    - Well-suited for displaying time series data, enabling a clear visualization of trends and patterns over successive time periods.
- GitHub URL - Data Visualization

EDA with SQL

The following SQL queries were performed on the data:

- Names of the unique launch sites in the space mission
- Five records where launch sites begin with the string 'CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved.
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass.
- Month, booster versions, launch site for failed outcomes in drone ship in the year 2015.
- The count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

- GitHub URL: SQL

Build an Interactive Map with Folium

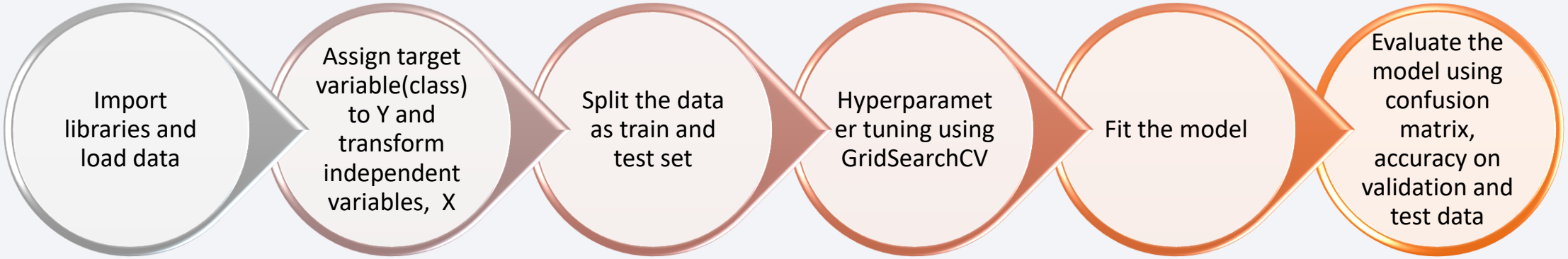
- The process of identifying an optimal location for a launch site involves analyzing existing launch site locations to uncover relevant factors. In this analysis,
- Circle and Marker symbols were utilized to denote the initial center location, the NASA Johnson Space Center in Houston, Texas, as well as four other launch sites.
- Marker clusters were employed to visualize launch outcomes for each site, allowing for the identification of high success rates.
- Additionally, MousePosition functionality was integrated to display the latitude and longitude of any point of interest on the map.
- Polylines were used to draw lines indicating distances between launch sites and nearby landmarks such as cities, railways, or highways.

GitHub URL: Folium maps

Build a Dashboard with Plotly Dash

- An interactive dashboard was developed using Plotly Dash, featuring two charts:
  - A pie chart displaying the total successful launches per launch site.
  - A scatter plot illustrating the relationship between the launch outcome and the payload mass so that we can visually observe how payload may be correlated with mission outcomes for selected site(s).
- Additionally, the dashboard includes interactive elements such as:
  - A dropdown menu provides access to all launch sites, enabling data filtering in both charts. This facilitates the comparison of success rates across all launch sites and individual sites, aiding in the identification of the highest and lowest success rates.
  - A range slider allows control over the range of payload mass displayed in the scatter plot. This feature aims to explore the correlation between variable payload and mission outcome, providing insights into potential relationships.
- GitHub URL: Dash

Predictive Analysis (Classification)



- The data underwent transformation and was divided into training(80%) and testing datasets(20%)
- Four distinct models were employed: Support Vector Machine, Decision Tree, Logistic Regression, and K-Nearest Neighbor.
- Each model underwent hyperparameter tuning procedures using GridSearchCV to optimize performance.
- Evaluation of each model was based on selected metrics: F1-Score, Accuracy, and confusion matrix
- GitHub URL : Machine Learning

Results

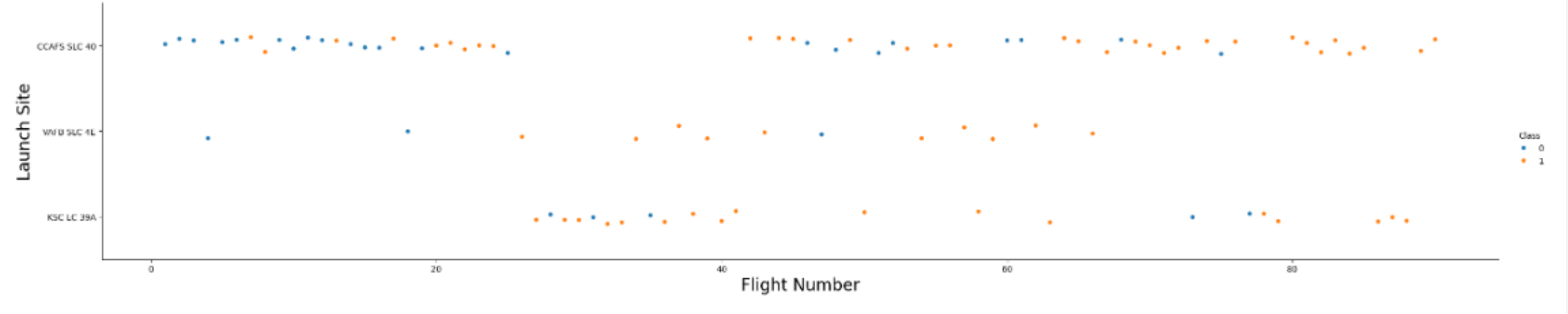
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

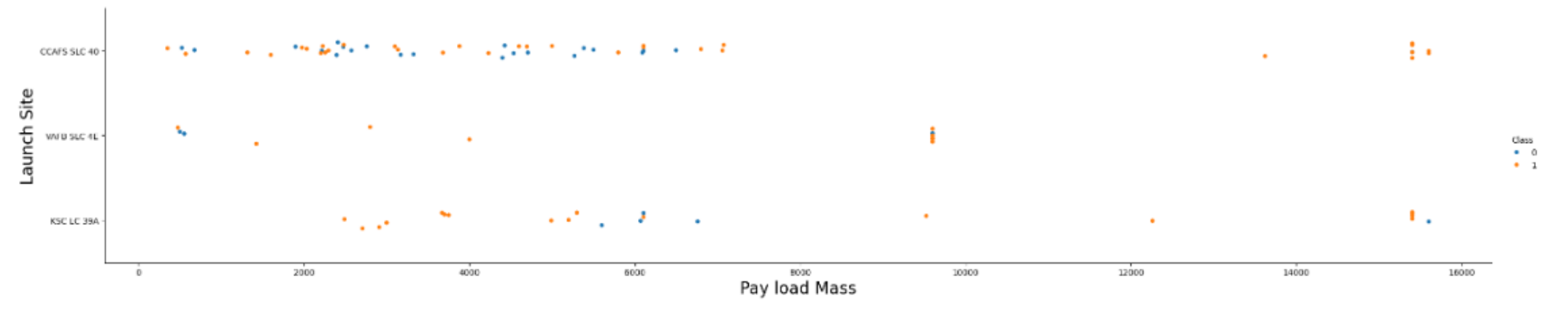
- A scatter plot of Flight Number vs. Launch Site



- **CCAFS SLC 40 has the most number of launches** while VAFB SLC 4E has the least.
- Compared to CCAFS SLC 40**, VAFB SLC 4E and KSC LC 39A has fewer unsuccessful landing**.
- As the flight number is increasing success rate is increasing as well.
- There is 100% success rate after flight number 80.

Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site

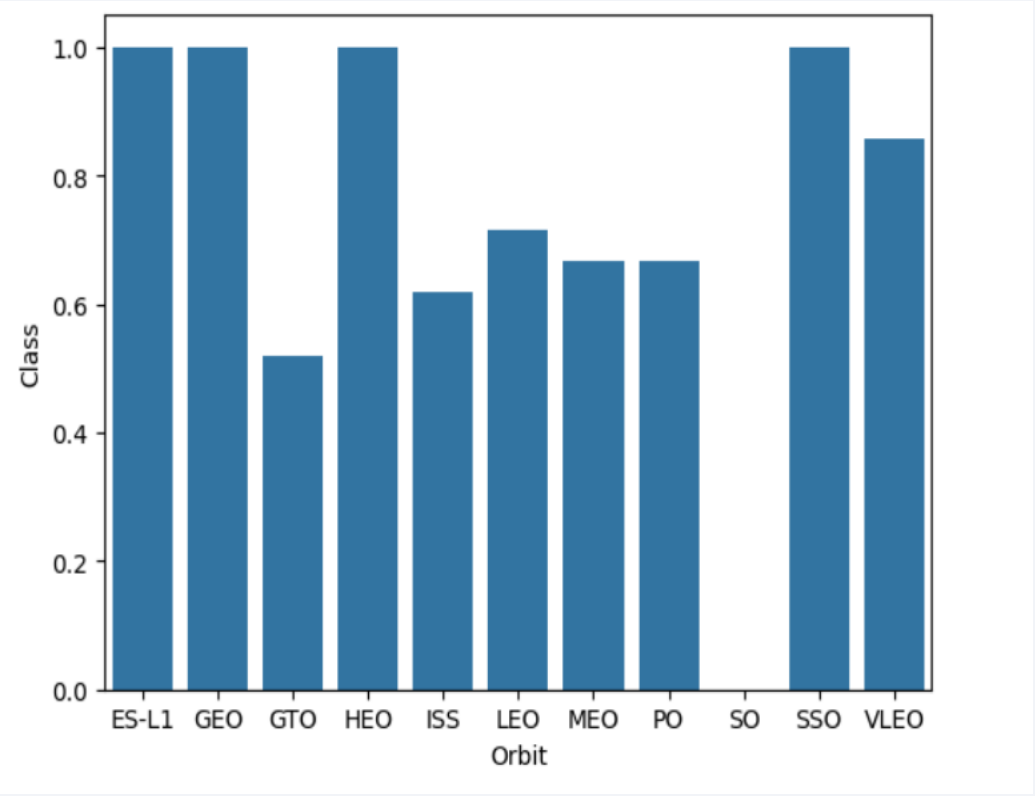


- There seems to be a positive correlation between higher payload mass and success rate. **Missions with payloads greater than 10000 kg seems to be more successful** even though there are fewer missions.
- More missions range between a payload mass of 2000kg and 6000kg. There are **no missions for VAFB SLC 4E above 10000kg**. This observation could imply limitations in the capabilities of this launch site for handling heavier payloads or a strategic decision to prioritize lighter payloads for missions launched from this site.
- **KSC LC 39A achieved a 100% success rate for payloads under 5000kg**. This observation suggests that this launch site may be particularly well-suited for handling missions with lighter payloads, possibly due to infrastructure, technology, or operational factors that contribute to higher success rates in this payload range.

Success Rate vs. Orbit Type

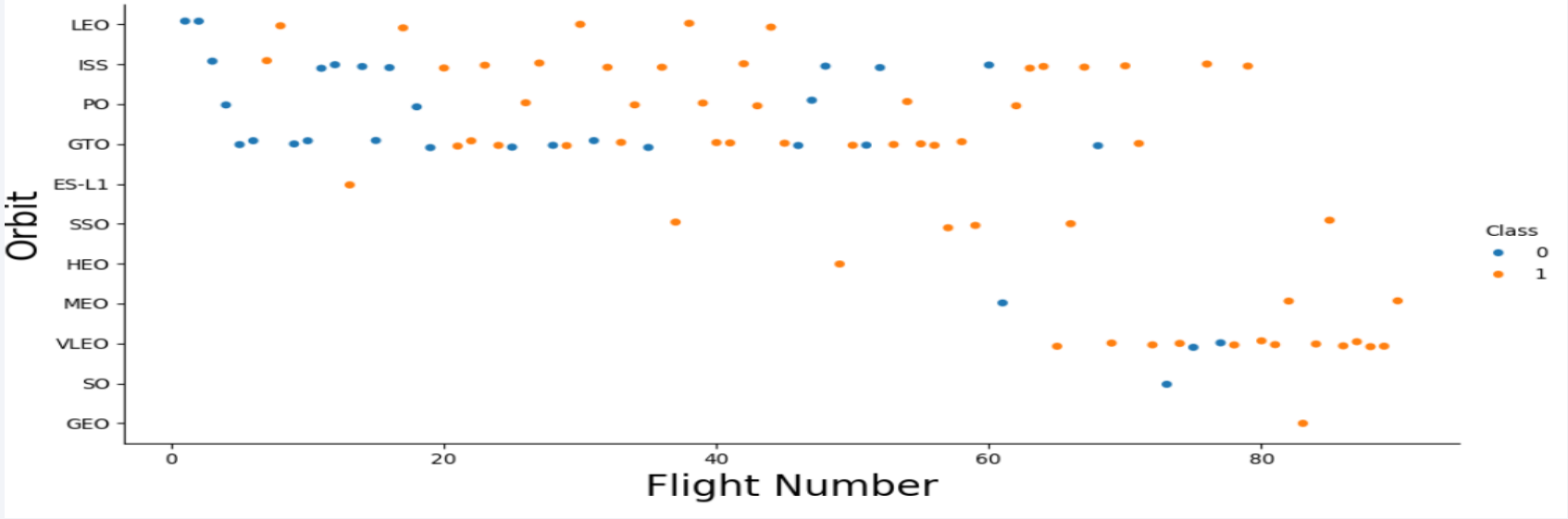
Bar chart for the success rate of each orbit type

- **Orbits ES-L1, GEO,HEO, and SSO has 100% success rate** suggesting a high level of reliability and effectiveness in launching missions to these destinations.
- **Orbit SO has 0 success rate**. indicating that missions aimed at this orbital destination did not achieve their objectives. This could be due to various factors such as technical challenges, operational issues, or unforeseen complications specific to missions targeting this orbit.



Flight Number vs. Orbit Type

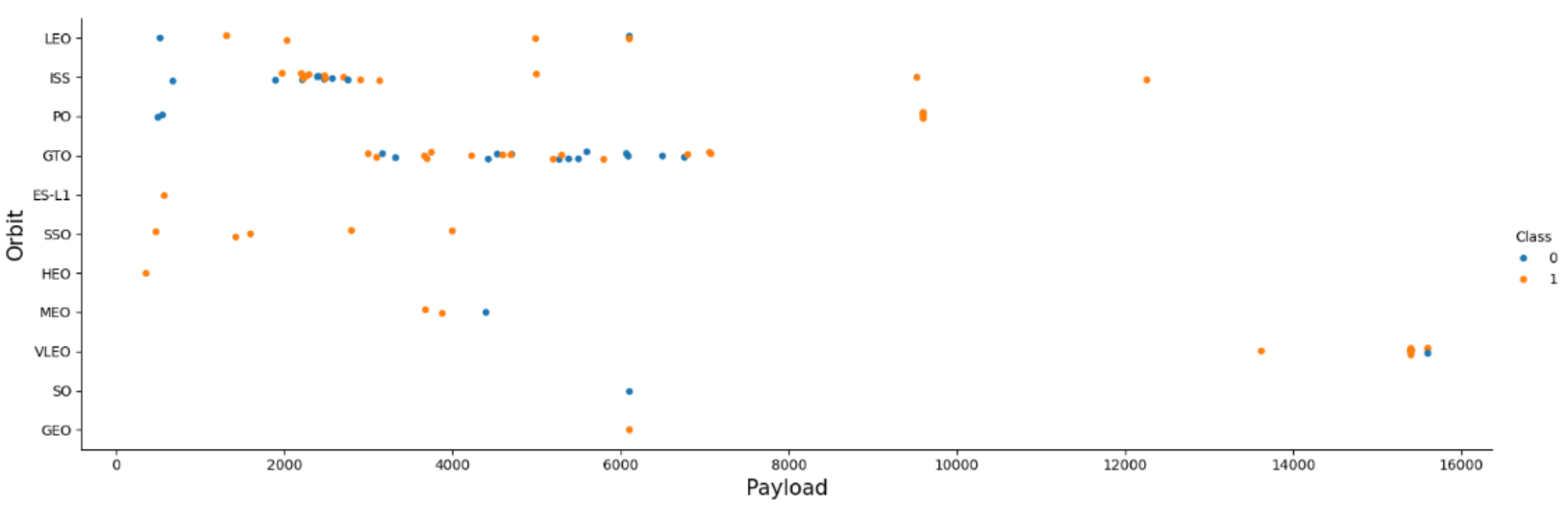
- Scatter point of Flight number vs. Orbit type



- There may be a **higher frequency of launches targeting the orbits ISS and GTO**. Potential reasons for this relationship could include the demand for satellite deployment missions. SSO has all successful mission.
- **LEO and PO had reasonable amount of successful missions** compared to the number of missions conducted. Rest of the orbits had only one launch each, making it impossible to draw any conclusion.
- **There is a notable trend towards targeting VLEO** after flight number 60, with a significant number of successful missions observed in this orbit. **It highlights the viability and effectiveness of targeting this orbit for future missions**.

Payload vs. Orbit Type

- Scatter point of payload vs. orbit type



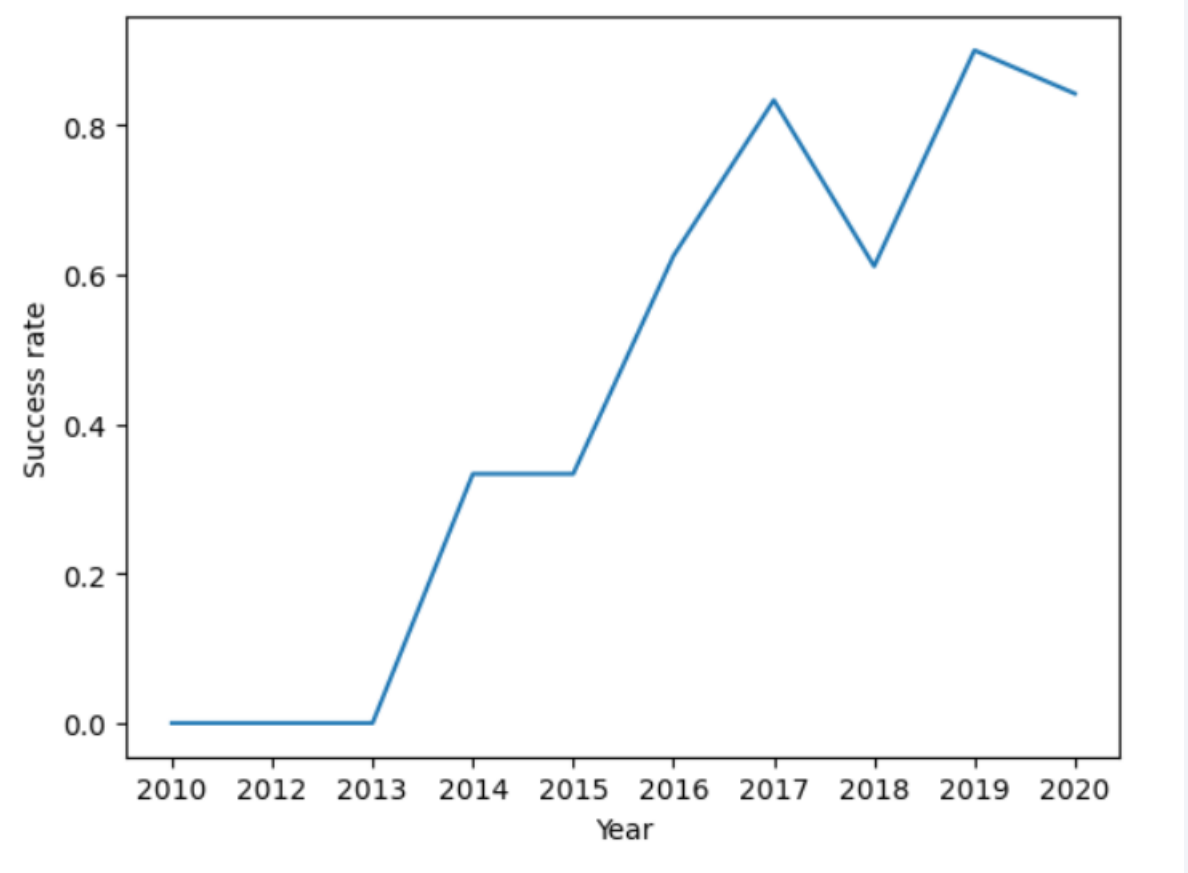
- Majority of the orbits has missions with payload that range between 2000 and 8000kg while VLEO have all the missions with payload mass greater than 13000kg, which are mostly successful.
- GTO has no record of a mission greater than 7000kg. All the missions in SSO, which are successful as well, had payload mass less than 4000kg.
- In general, **majority of the orbits are targeted of missions with payload less than 7000kg**. Possible reasons for this preference may include cost considerations, technological limitations, and operational constraints associated with launching heavier payloads.

Launch Success Yearly Trend

A line chart of yearly average success rate

- Initially, the **first three periods show a 0% success rate**. Subsequently, there is a notable increase in success rates until 2017 followed by **fluctuations in the succeeding years.**

- Despite these fluctuations, **there is an overall increase in success rates over the entire period**.



All Launch Site Names

Names of **four different launch sites** were obtained from the data using DISTINCT statement in SQL which are CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40.

In [9]:
```
%sql select DISTINCT(LAUNCH_SITE) from SPACEXTBL
```

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

Out[9]:

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

Launch Site Names Begin with 'CCA'

The following 5 records from spacex table were displayed using '**LIKE**' statement in SQL where the launch sites begin with `CCA`

```
%sql select * from SPACEXTBL 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

Using **sum()** and **WHERE** condition the total payload carried by boosters from NASA is found to be 45596kg

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

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

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Using **avg()** and the **WHERE** condition the average payload mass carried by booster version F9 v1.1 is found to be 2928.4kg.

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

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

The date of the first successful landing outcome on ground pad is 22nd December, 2015. **Min()** and **WHERE** condition were used to obtain this.

```
%sql select MIN(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)';
```

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

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of four boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are displayed here using **WHERE** condition.

```
%sql select Booster_Version from SPACEXTBL where (Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000);
```

```
 * 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 total number of successful and failed mission outcomes are displayed here using **COUNT()** and **GROUP BY** statements.

List the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome, Count(Mission_Outcome) as Total_Number from SPACEXTBL GROUP BY Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

The names of the 12 boosters which have carried the maximum payload mass using **WHERE** and **MAX()**.

```
%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

```
 * 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 the year 2015 is displayed here using **WHERE**. Month is extracted from the given data in the data using **substr()**.

```
%sql select substr(Date,6,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, Landing_Outcome FROM SPACEXTBL \
where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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 is shown here using **WHERE**, **COUNT()**, **GROUP BY** and **ORDER BY.**

```
%sql select Date,Landing_Outcome, count(*) as outcome_count FROM SPACEXTBL WHERE DATE between '2010-06-04' and '2017-03-20' \
group by Landing_Outcome order by outcome_count DESC;
```

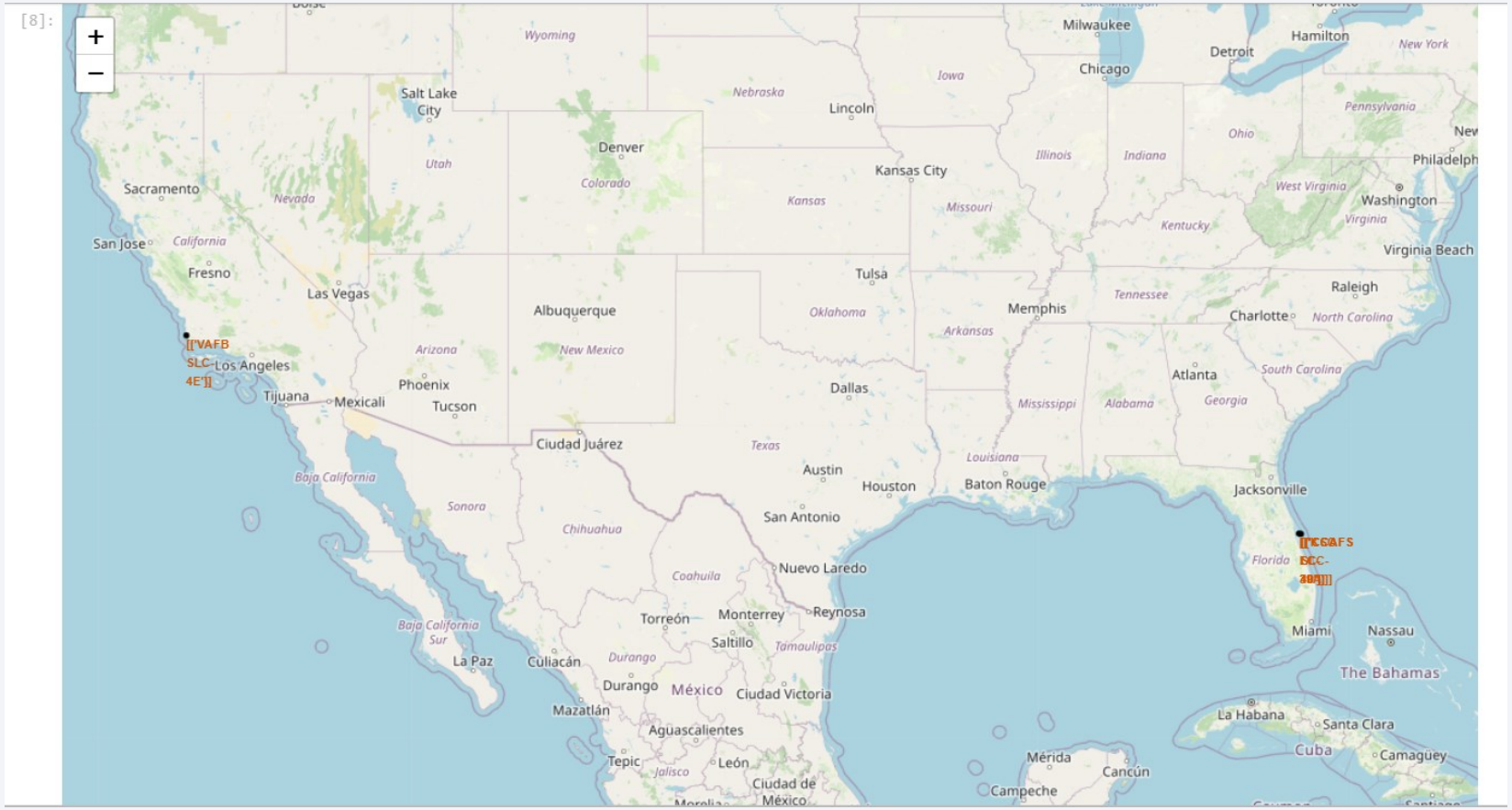
```
 * sqlite:///my_data1.db
Done.
```

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

Section 3

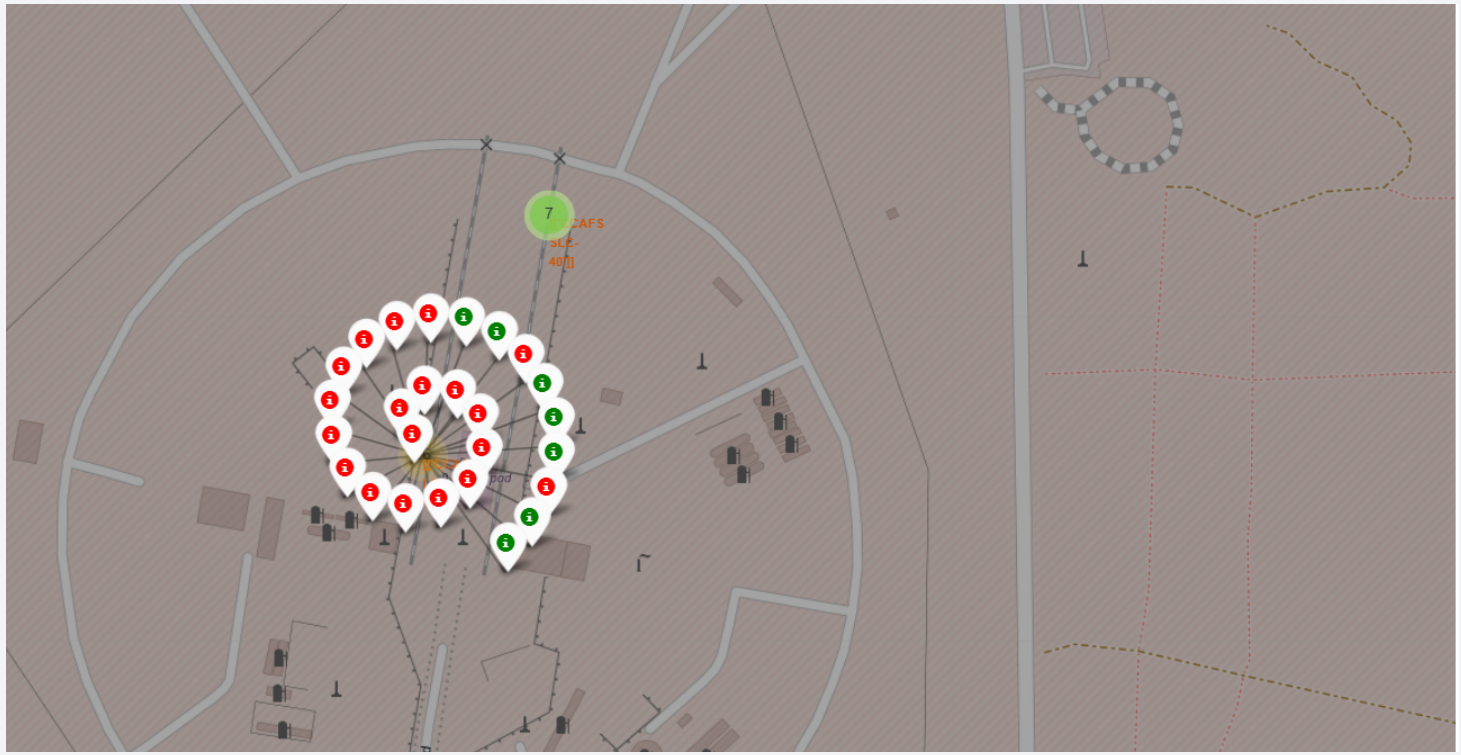
Launch Sites Proximities Analysis

Folium Map for all launch sites



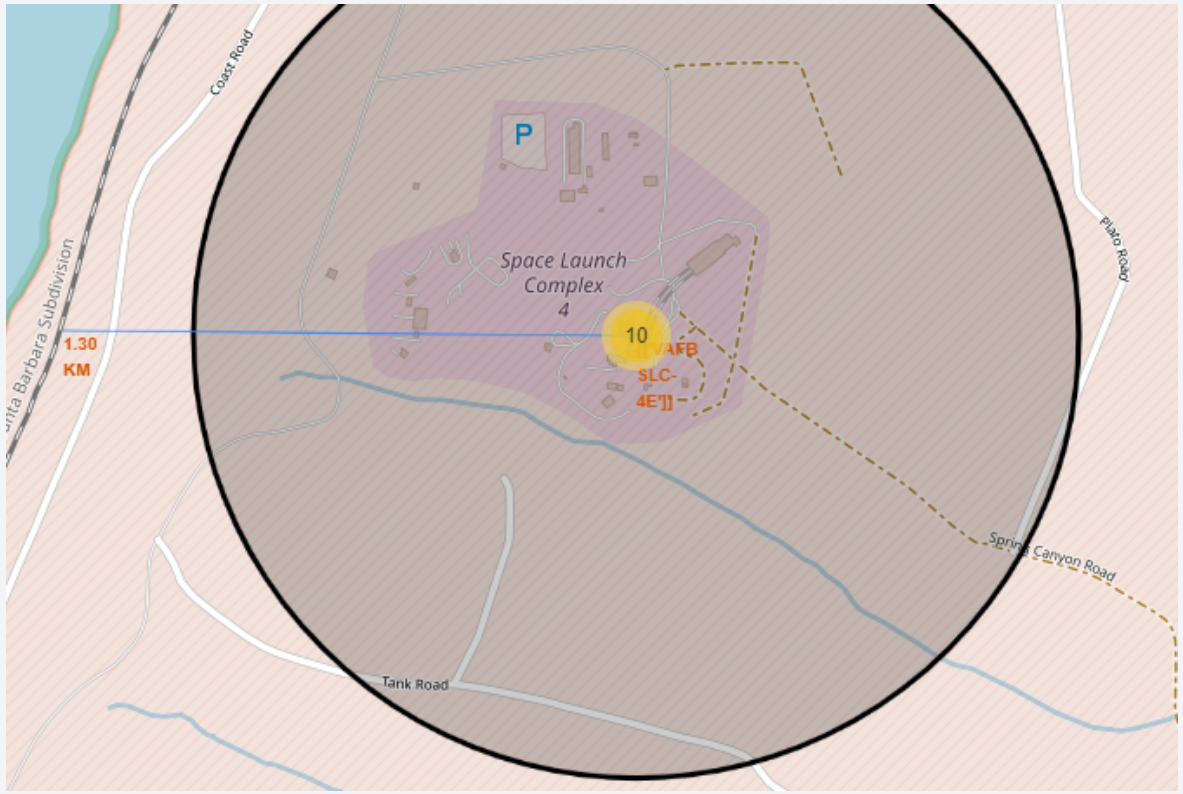
- There are four launch sites marked in the map, one near California and other three near Florida.
- All the sites are situated **near to the coastal area, away from cities** which helps to avoids any risks to populated area.

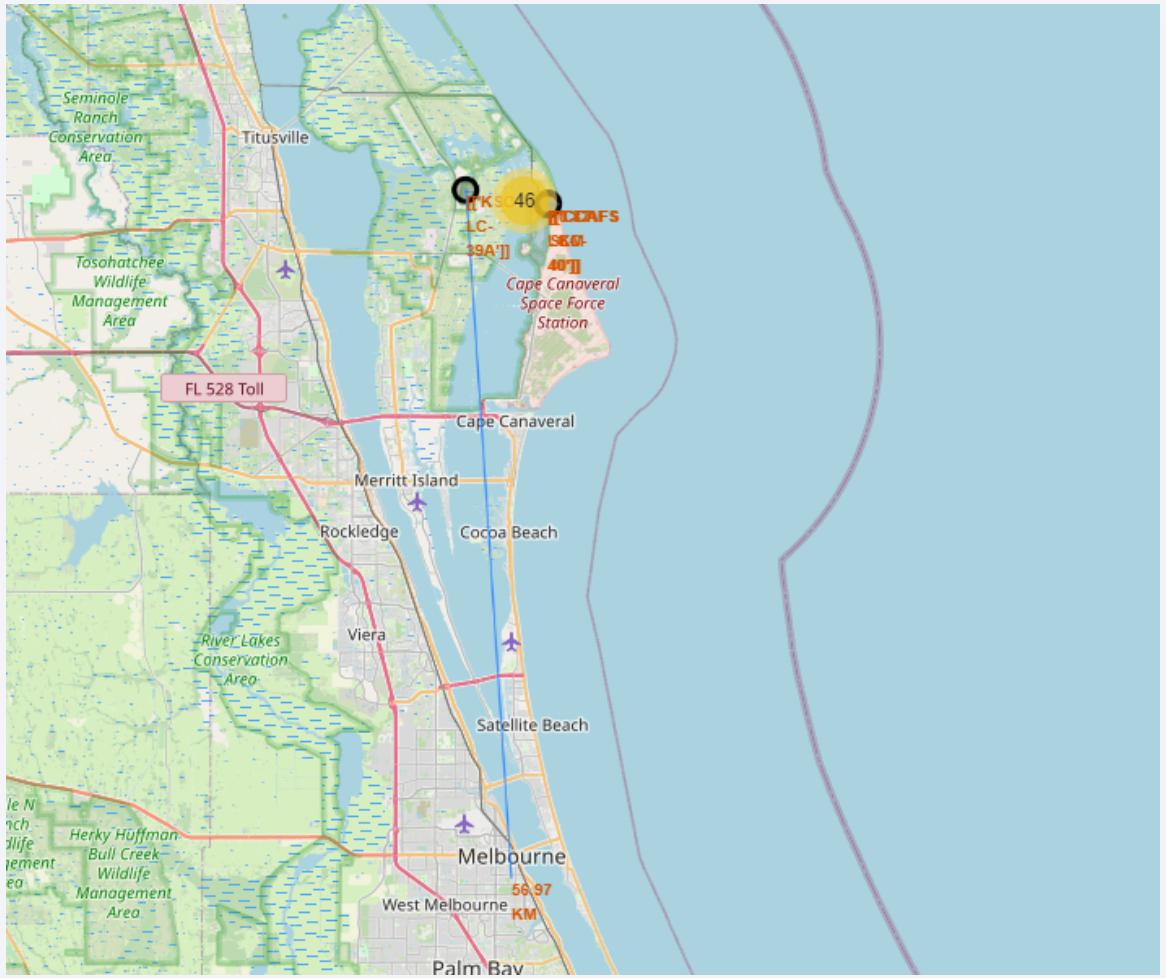
<Colour labeled map for Launch Outcomes>



- The map displays color-labeled outcomes for launches from the launch site **CCAFS LC-40**.
- Each launch site is assigned a color label indicating the outcome of its missions. **Red markers represent failures, while green markers indicate successful landings.**
- Specifically for **CCAFS LC-40, there are more unsuccessful landings than successful ones**, as indicated by the predominance of red markers on the map.

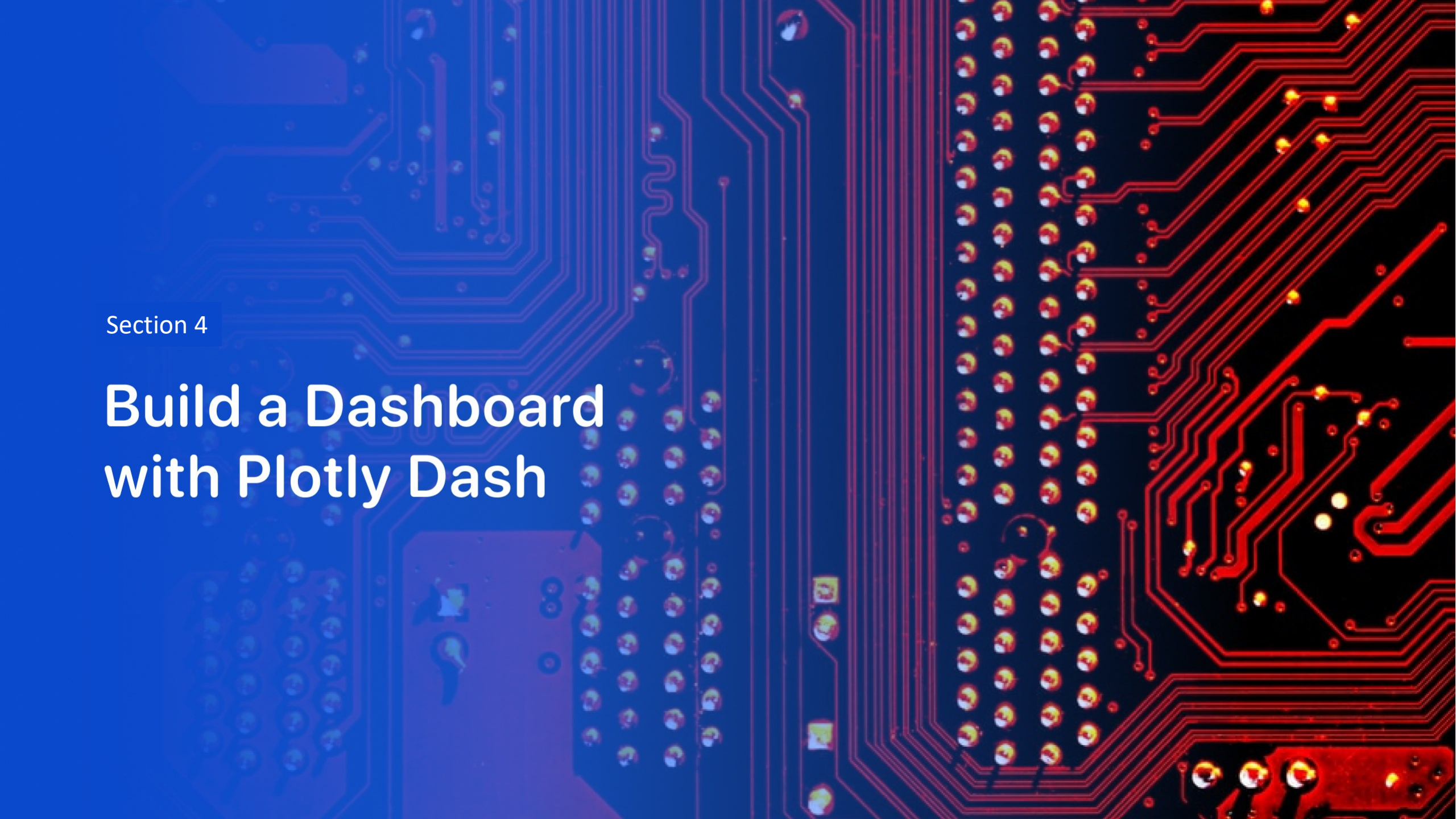
Proximity of Launch Site





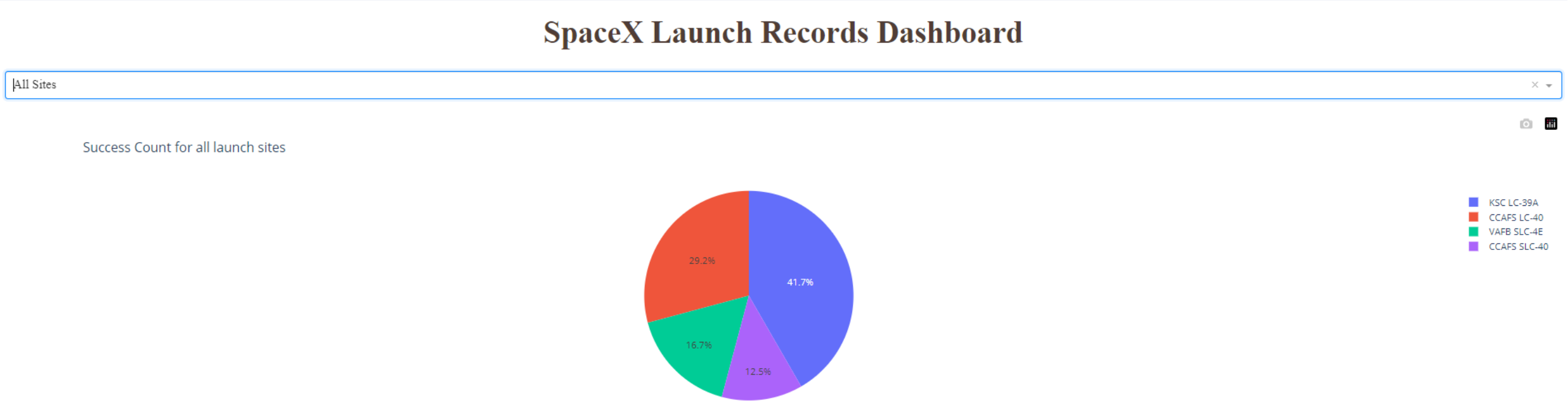
Proximity of Launch Site(Contd.)

- **Railway, Highway and Coastline are relatively in close proximity to the launch site** suggesting potential transportation and logistical advantages for accessing and transporting equipment to and from the launch site.
- The **far distance from the city** may indicate a strategic decision to locate the launch site away from densely populated areas for safety and operational reasons.


Section 4
Build a Dashboard with Plotly Dash

Success count for all launch sites

- **Plotly Dash** was used to create this pie chart. It has a dropdown to select the launch site to display the respective success count as well.
- **KSC LC-39A had the highest count of success**(41.7%) followed by CCAFS LC-40(29.2%)
- **CCAFS SLC-40(12.5%) had the least success** compared to all the sites.



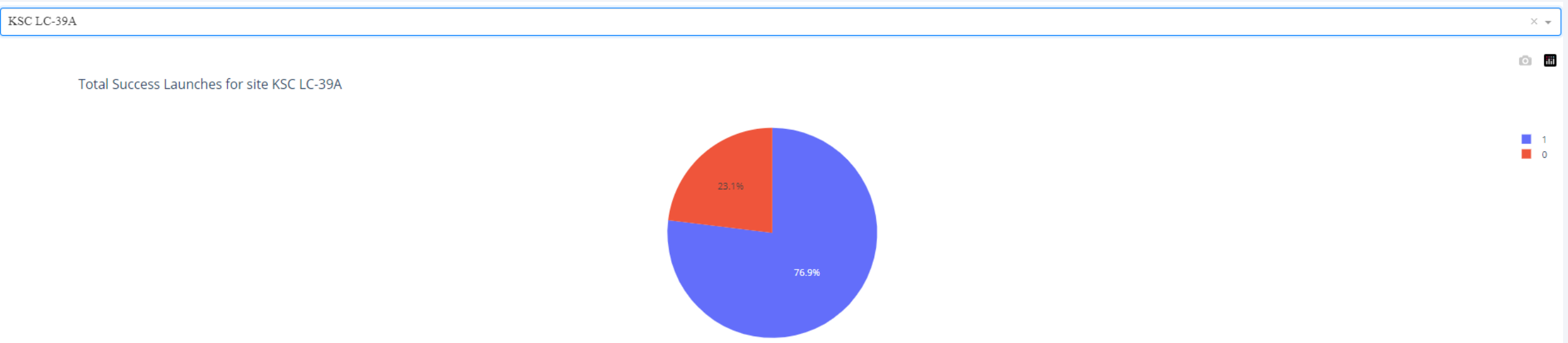
Site with Highest Launch Success

- **KSC LC-39A has the highest success launch rate** with 76.9% . Out of 13 missions 10 were successful.
- While the others sites have the following success rates:

  CCAFS LC-40: 73.1%

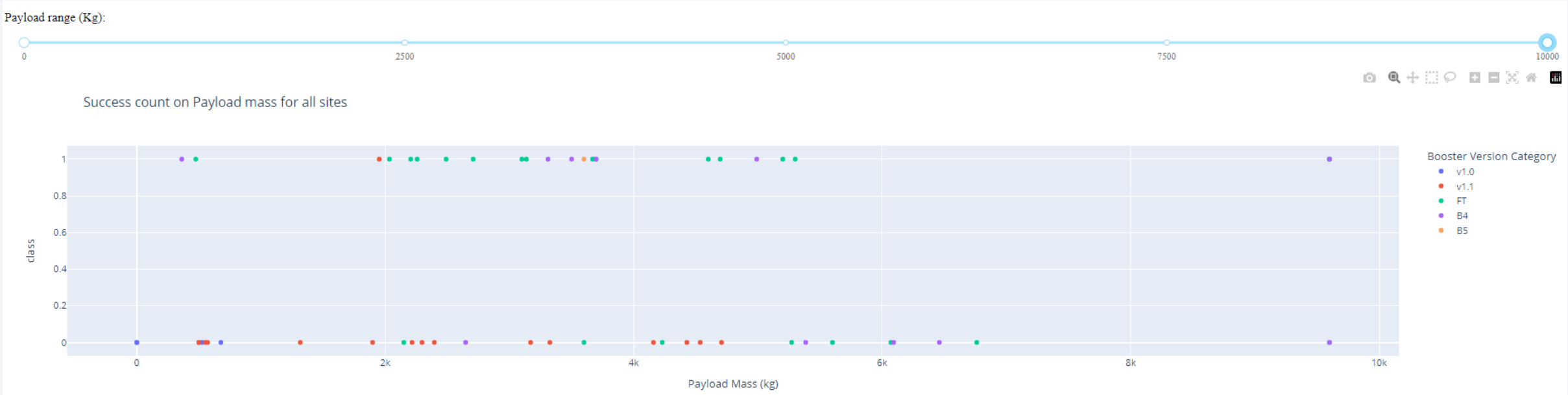
  VAFB SLC-4E: 60%

  CCAFS SLC-40: 57.1%



Booster Version vs Launch Outcome

Booster FT has the largest success rate compared to others boosters mostly between the payload mass 2000kg and 5500kg.



Payload vs Launch Outcome

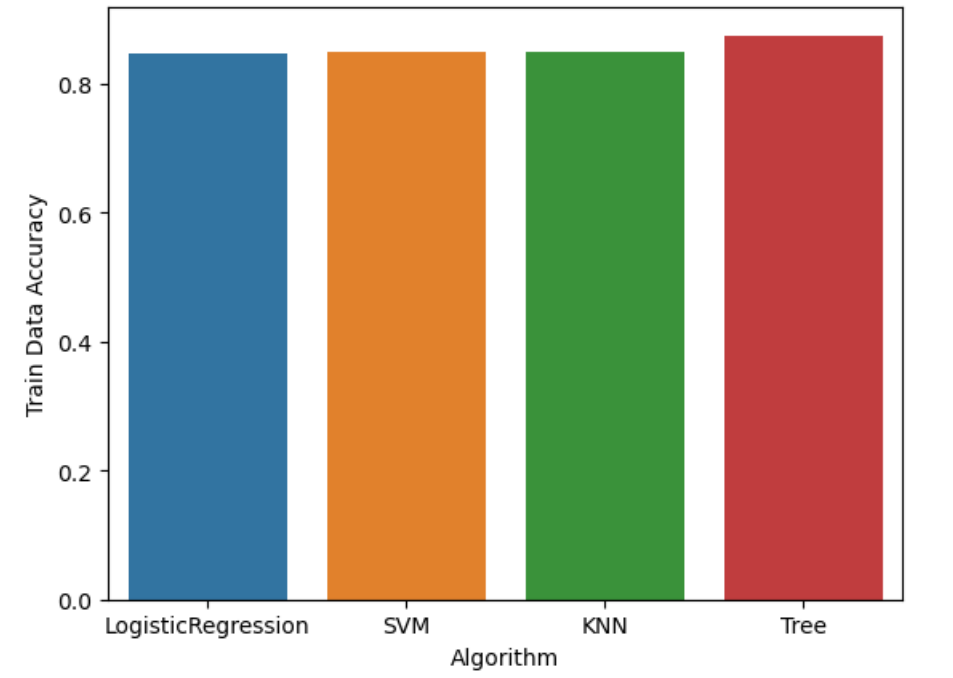
From the plot, **payload range between 2000 kg and 5500kg has the highest success rate**.

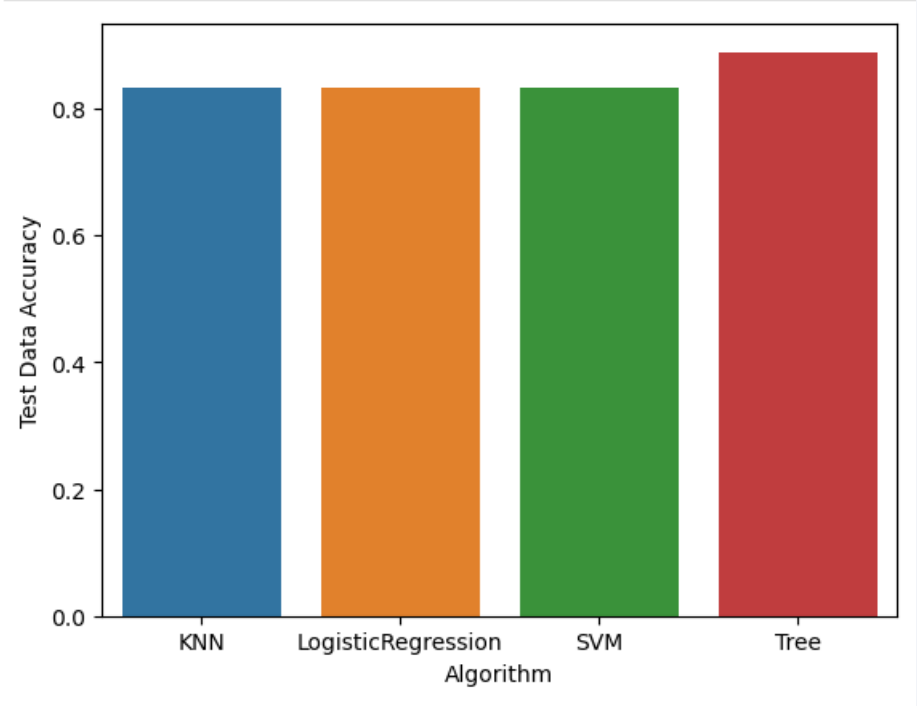
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The below bar charts represent accuracy of the four different models on train Data and Test Data. **All the models perform almost similar. However, decision tree has a slightly improved accuracy.**





Classification Accuracy

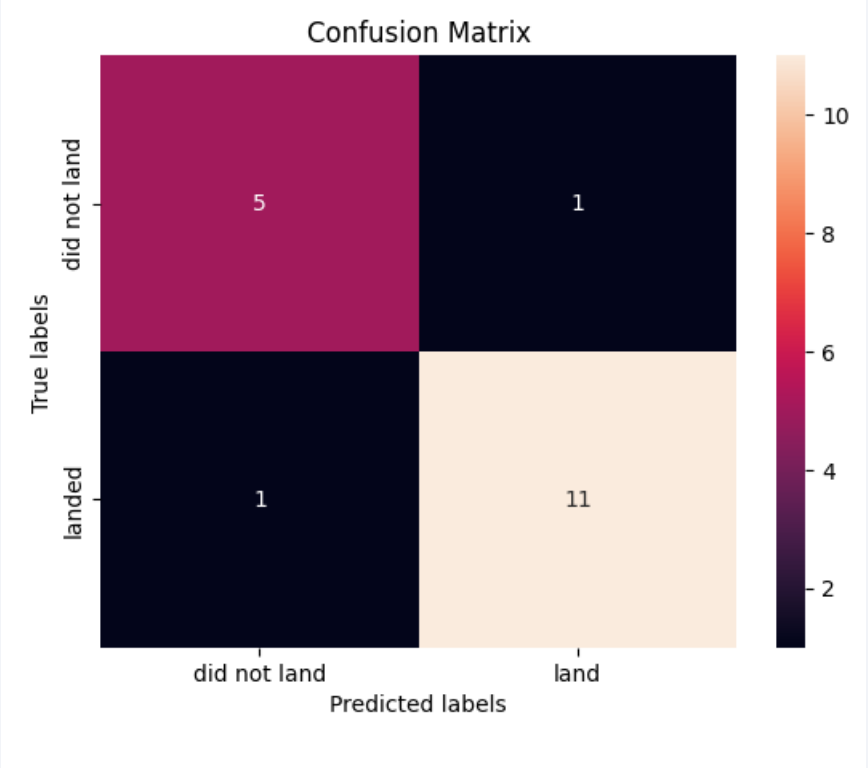
- Below are the **accuracy obtained for the validation data and testing data**. Decision Tree has slightly more accuracy than the other models.

	Algorithm	Train Data Accuracy
0	LogisticRegression	0.846429
1	SVM	0.848214
2	KNN	0.848214
3	Tree	0.875000

	Algorithm	Test Data Accuracy
0	KNN	0.833333
1	LogisticRegression	0.833333
2	SVM	0.833333
3	Tree	0.888889

Confusion Matrix for Decision Tree

- Decision Tree has **only one false positive and one false negative**. Rest of them are predicted correctly which indicates a good model.

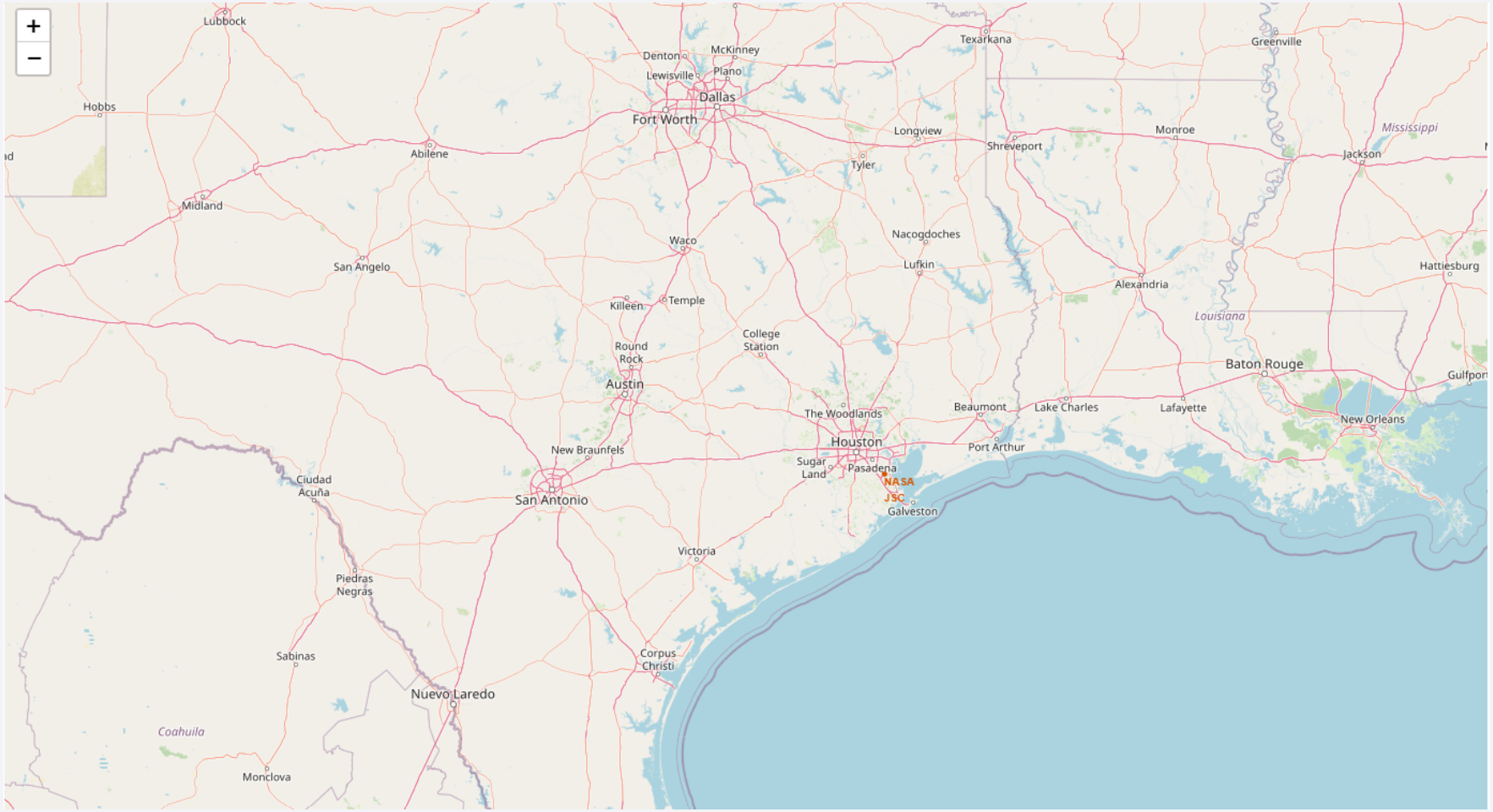


Conclusions

- CCAFS SLC 40 has the most number of launches while VAFB SLC 4E has the least.
- Missions with payloads greater than 10000 kg seems to be more successful even though there are fewer missions.
- Orbits ES-L1, GEO,HEO, and SSO had 100% success rate while Orbit SO has 0 success rate
- There is an overall increase in success rates over the entire period.
- KSC LC 39A achieved a 100% success rate for payloads under 5000kg.
- Decision tree is the model with highest accuracy even though all the models performed almost similar.

Appendix

Github URL - [Github]



Initial center location, NASA Johnson Space Center at Houston, Texas.

Appendix

Initial center location, NASA Johnson Space Center at Houston, Texas.

Thank you!