IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Chandan Gowda NG

17th August 2025

Outline

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

Executive Summary

- **Summary of methodologies are:**
  - Collected data through the SpaceX API and web scraping techniques.
  - Performed data cleaning, wrangling, and preprocessing.
  - Conducted exploratory analysis using SQL magic commands and Python visualization libraries.
  - Created interactive maps and dashboards for data analytics.
  - Applied machine learning classification algorithms for predictive analysis

- Sum**mary of all results:**
  - Findings obtained from exploratory data analysis.
  - Insights derived from interactive dashboards and maps.
  - Evaluation and interpretation of model results.
  - Identification of the most effective classification model for predicting success.

Introduction

- SpaceX rockets are renowned for their reusable technology and consistent vertical landings.
- Falcon 9 rockets from SpaceX have demonstrated reliability and efficiency in launching payloads into space, though there have been occasional failures.
- This applied data science capstone project focuses on predicting the success of Falcon 9 first stage rocket landings by utilizing classification models.
- The main goal is to determine whether the first stage can be successfully reused.

Introduction

- Problems to find answers to:

- Is the rocket launch outcome influenced by the payload size?
- Does the launch site impact the success of a rocket launch?
- Does the orbital destination have an effect on the outcome of the rocket launch?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describes how data was collected using Space X API and web scraping from Wikipedia website
- Perform data wrangling
  - Describes how the data was processed by finding and replacing missing values and also converting columns with categorical values to numerical
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Describes how to build, tune, and evaluate classification models

Data Collection

- This process outlines the methods used to collect data on SpaceX Falcon 9 rockets through various techniques
  - **Data collection using Space X API involved:**
- Retrieved data from the SpaceX API URL using the request.get() method.
- Parsed the response content with the .json() function.
- Transformed the data into a Pandas DataFrame using .json_normalize().
- Extracted additional details such as booster name, payloads, launch site, landing outcome, flight number, and date via other SpaceX API endpoints.
- Filtered the DataFrame to include only Falcon 9 launches and imputed missing values with their respective mean values.

Data Collection

- **Data collection using Web scraping involved:**

- Imported the requests, BeautifulSoup, and pandas libraries.
- Fetched data from the Wikipedia URL using the requests.get() method.
- Created a BeautifulSoup object from the HTML response using the BeautifulSoup() constructor.
- Extracted all tables from the HTML using soup.find_all().
- Retrieved the third table and iterated through it to obtain the headers or column names.
- Built a DataFrame by iterating through each HTML table to extract information such as Flight Number, Launch Site, Payload, Payload Mass, Orbit, Customer, Launch Outcome, Version Booster, Booster Landing, Date, and Time.

Data Collection – SpaceX API

- Figure 1 illustrates the data collection process utilizing Python's requests library and the SpaceX API.
- The final data frame is saved as a CSV file using df.to_csv()
- The URL below is the GIT repository containing the Jupyter notebook

https://github.com/Chandan-39/Data-Science-Capstone-Project-IBM---SpaceX-Falcon9-landing-predictions/blob/main/Data%20collected%20by%20APIs/jupyter-labs-spacex-data-collection-api.ipynb

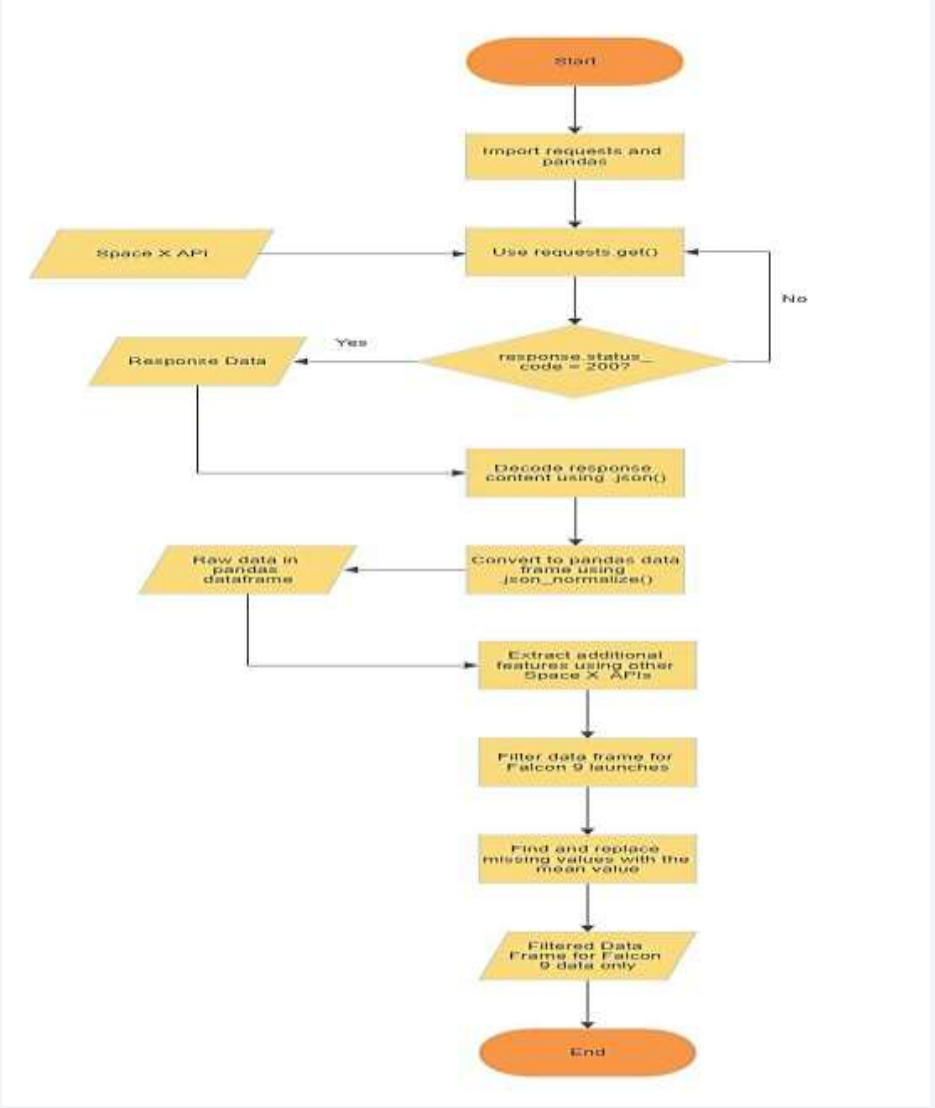


Figure 1. Flow chart of Data collection process using Space X API

Data Collection - Scraping

- Figure 2 illustrates the data collection process utilizing Python's requests and BeautifulSoup libraries.
- The final data frame is saved as a CSV file using data_falcon9.to_csv()
- The URL below is the GIT repository containing the Jupyter notebook

https://github.com/Chandan-39/Data-Science-Capstone-Project-IBM---SpaceX-Falcon9-landing-predictions/blob/main/Web%20Scraping/jupyter-labs-webscraping.ipynb

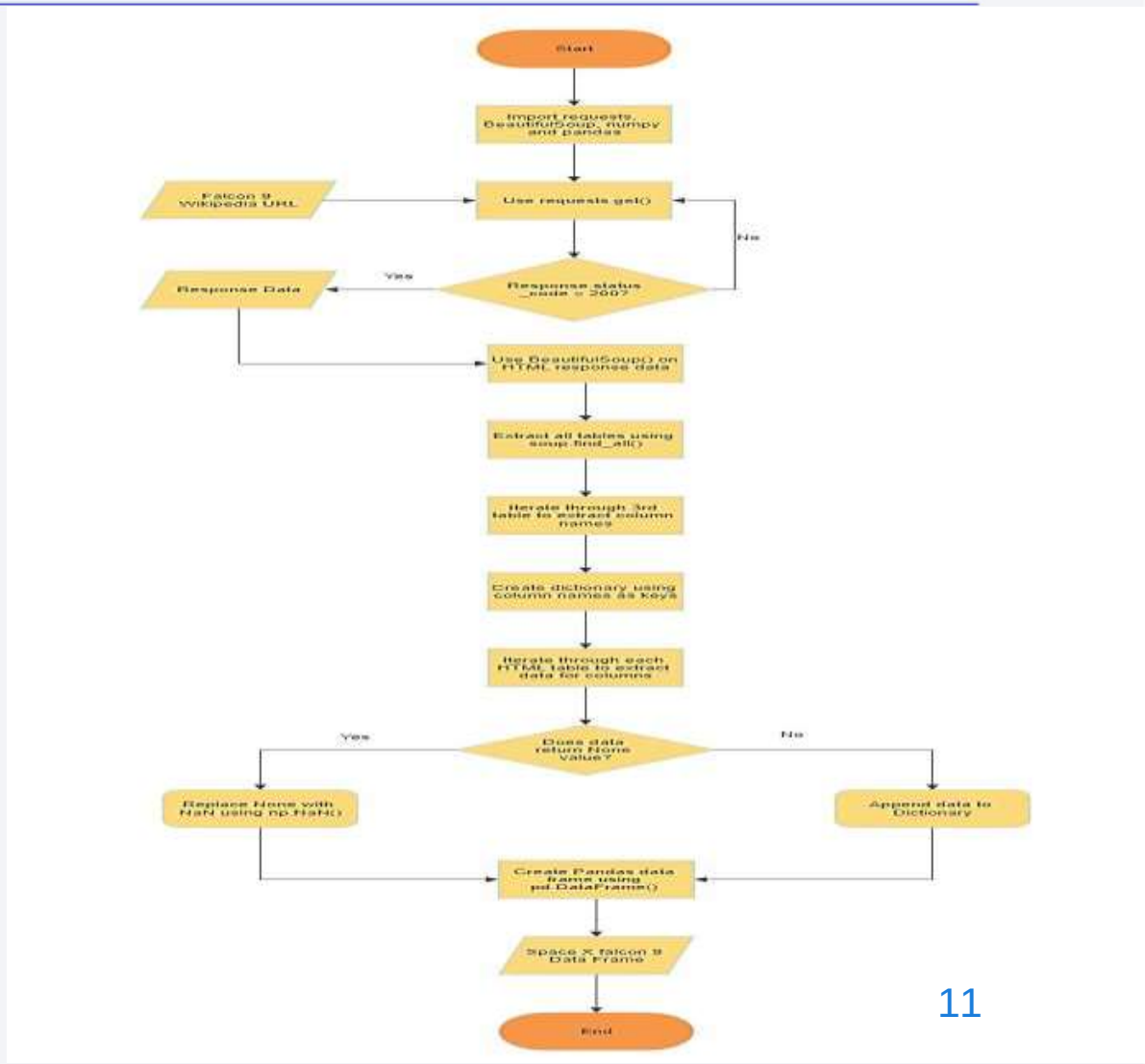


Figure 2. Flow chart of Data collection process using web

Data Wrangling

- This included performing exploratory data analysis on the dataset.
- Additionally, the outcomes or classes were transformed into training labels, where 1 indicates a successful landing and 0 represents all other results.
- Figure 3 presents a flow chart illustrating the data wrangling process.
- The URL below is the GIT repository containing the Jupyter notebook

https://github.com/Chandan-39/Data-Science-Capstone-Project-IBM---SpaceX-Falcon9-landing-predictions/blob/main/Data%20Wrangling/labs-jupyter-spacex-Data%20wrangling.ipynb

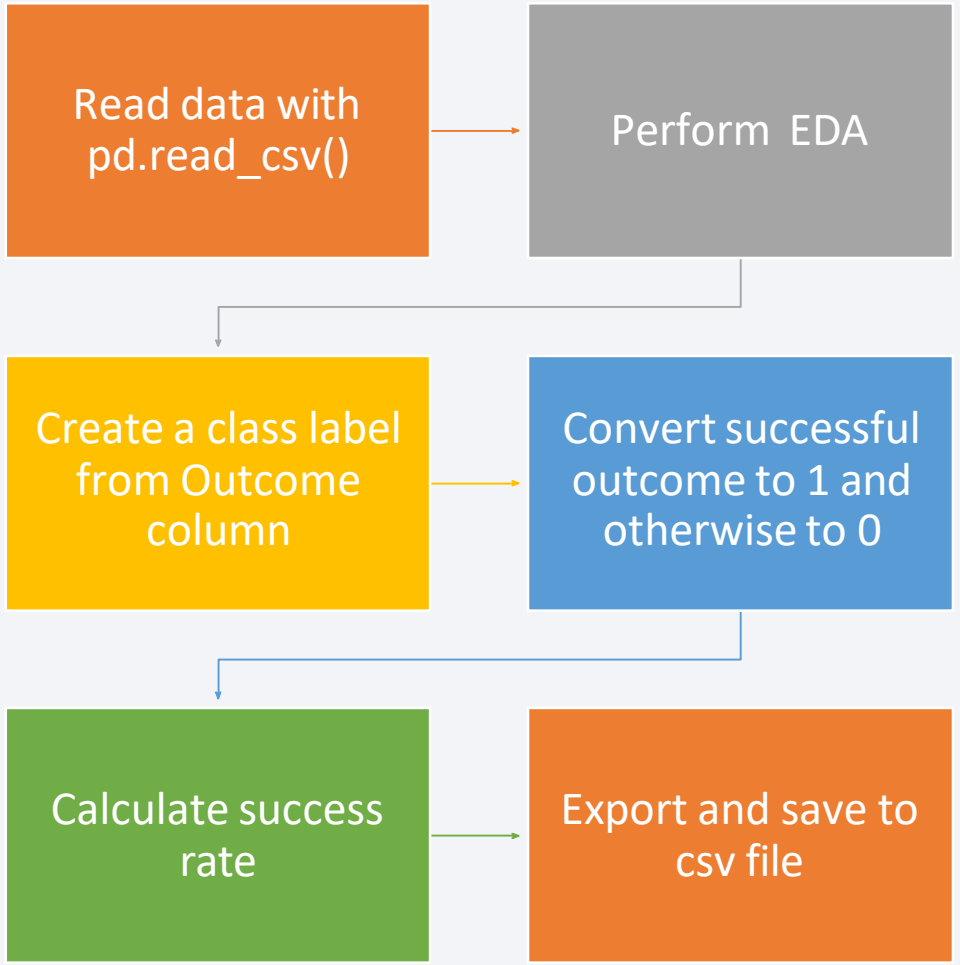


Figure 3. Flow chart of Data Wrangling process.

EDA with Data Visualization

- During exploratory data analysis, scatter plots were utilized to visualize the impact of various variables on launch outcomes, as well as the relationships among them, such as payload mass versus flight number, flight number versus launch site, and payload versus launch site.
- Bar charts were also employed to display the success rates across different orbit types, while scatter plots were used to examine the relationships between flight number and orbit type, as well as payload and orbit type.
- Finally, a line plot was utilized to depict the annual trend in launch successes
- The URL below is the GIT repository containing the Jupyter notebook

https://github.com/Chandan-39/Data-Science-Capstone-Project-IBM---SpaceX-Falcon9-landing-predictions/blob/main/EDA%20with%20Python/edadataviz.ipynb

EDA with SQL

- Exploratory data analysis conducted using SQL involved the following:
  - List the distinct launch sites used in space missions.
  - Identify launch sites that start with ‘CCA’.
  - Show the total payload mass carried by boosters launched for NASA (CRS).
  - Provide the average payload mass for the F9 v1.1 booster version.
  - Find the first successful landing outcome on a ground pad.
  - List boosters that successfully landed on drone ships and carried payloads between 4,000kg and 6,000kg.
  - Present the total count of successful and failed mission outcomes.

EDA with SQL

- Exploratory data analysis (cont'd):
  - List the booster version names that carried the highest payload mass.
  - Show the month names, failed drone ship landing outcomes, booster versions, and launch sites for launches occurring in 2015.
  - Rank the number of successful landing outcomes between 04-06-2010 and 20-03-2017, displaying the results in descending order.

The URL below is the GIT repository containing the Jupyter notebook for EDA with SQL

https://github.com/Chandan-39/Data-Science-Capstone-Project-IBM---SpaceX-Falcon9-landing-predictions/blob/main/EDA%20with%20SQL/jupyter-labs-eda-sql-coursera_sqllite%20(3).ipynb

Build an Interactive Map with Folium

- Certain folium map objects were used such as:
  - folium.Circle used to add a highlighted circle area of NASA JSC as an initial centre location
  - folium.map.Marker used to create a marker at a specific launch location on the map
  - MarkerCluster() used to create cluster markers of successful and failed launches for a particular site
  - MousePosition() provides a way to display the latitude and longitude coordinates of the mouse cursor’s position on a map. Used to calculate the distance of the launch sites to the coasts.
  - Folium.PolyLine() is used to create a series of connected line segments on the map to mark the distance of the launch sites to the coast, railways, highways, and major cities

Build an Interactive Map with Folium

- The URL below is the GIT repository containing the Jupyter Notebook for Folium Map

  https://github.com/Chandan-39/Data-Science-Capstone-Project-IBM---SpaceX-Falcon9-landing-predictions/blob/main/Folium%20Maps/lab_jupyter_launch_site_location.ipynb

- To view rendered Folium Map use the link below:

https://nbviewer.org/github/Chandan-39/Data-Science-Capstone-Project-IBM---SpaceX-Falcon9-landing-predictions/blob/main/Folium%20Maps/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The plots and graphs added to the dashboard include:
  - A drop-down menu featuring all available launch sites.
  - A pie chart displaying success rates according to the selected launch site.
  - A range slider for choosing the payload mass.
  - A scatter plot illustrating the relationship between payload mass and launch success for the chosen sites.

- The URL below is the GIT repository containing the Jupyter Notebook

https://github.com/Chandan-39/Data-Science-Capstone-Project-IBM---SpaceX-Falcon9-landing-predictions/blob/main/Plotly%20Dashboard/spacex-dash-app.py

Predictive Analysis (Classification)

- The following are the steps taken in building and evaluating the classification models used.
    - The dataset was loaded and separated into feature and target variables.
    - Feature columns were normalized, and the target column was converted into a NumPy array.
    - Data was divided into training and testing sets.
    - GridSearchCV was applied to all classification algorithms to identify optimal parameters and highest scores, using .best_params_ and .best_score_.
    - The test set accuracy was evaluated using the .score() method.
    - Finally, a confusion matrix was plotted to visualize the classification results.

Predictive Analysis (Classification)

- Figure 4 presents a flow chart outlining the process of developing and evaluating classification models.
- The URL below is the GIT repository containing the Jupyter Notebook

https://github.com/Chandan-39/Data-Science-Capstone-Project-IBM---SpaceX-Falcon9-landing-predictions/blob/main/Predictive%20Analysis%20using%20ML/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

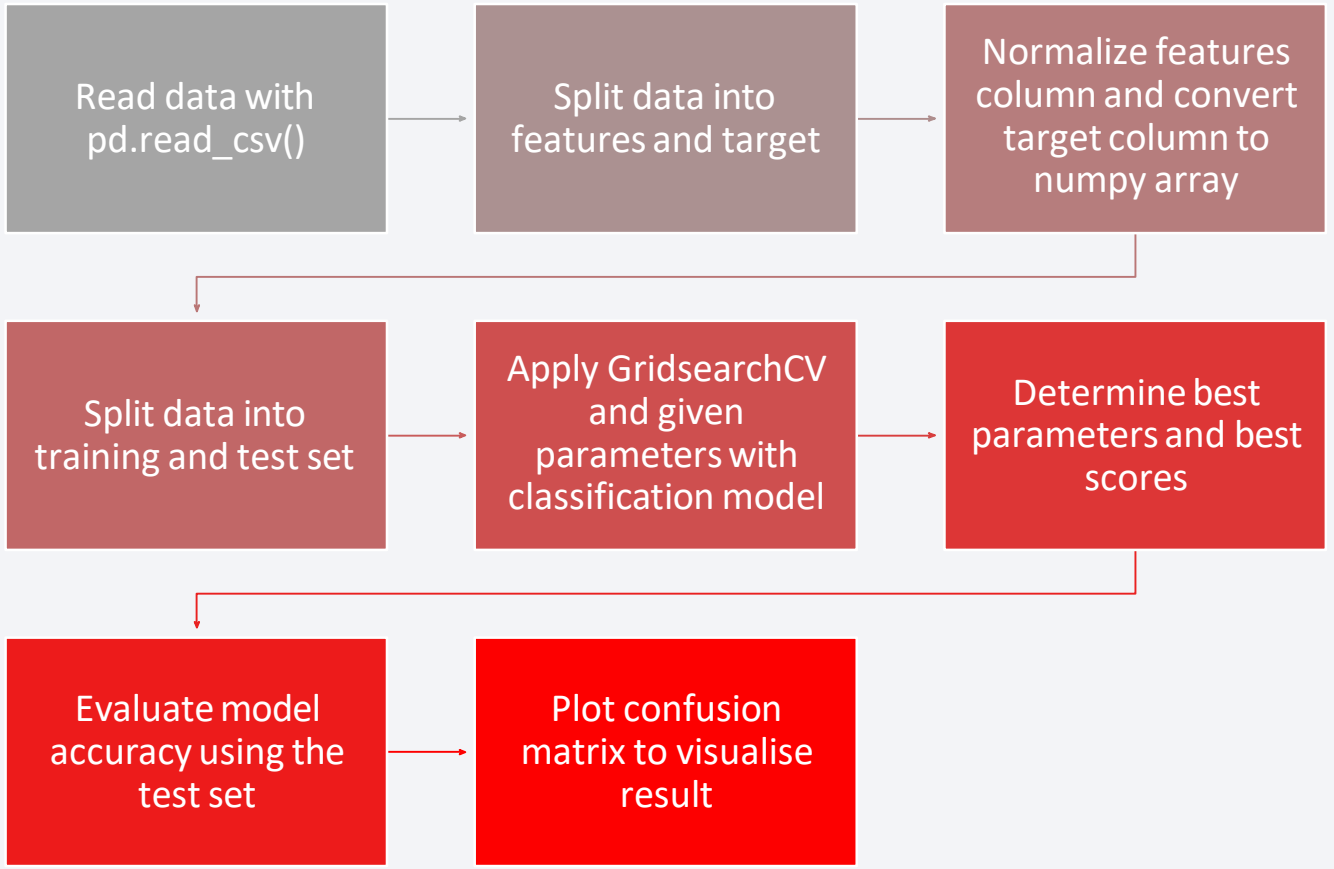
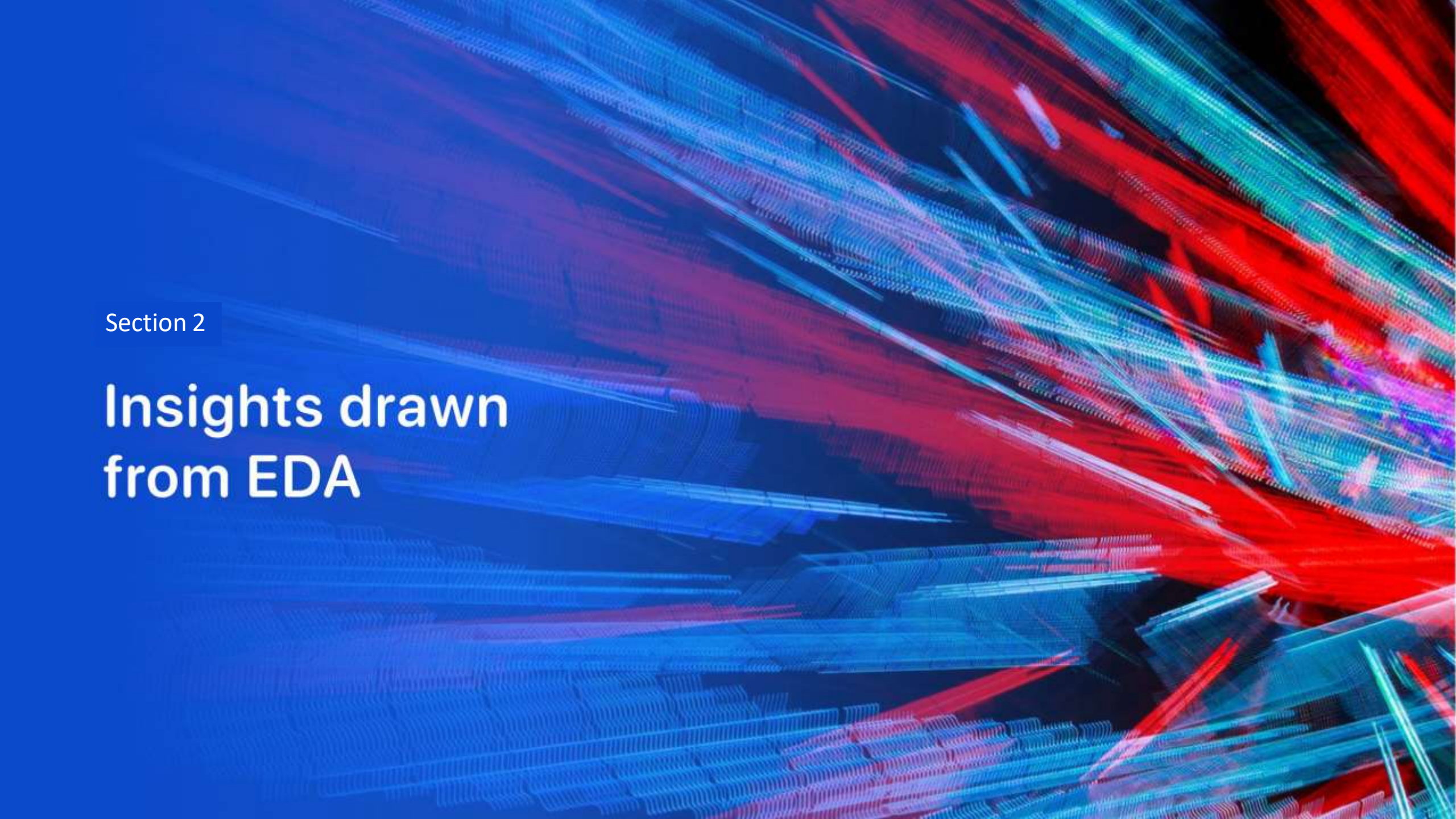


Figure 4. Flow chart of Model development and Evaluation process

Results

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


Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site

- The figure below shows that the launch site CCAFS SLC 40 has launched more rockets than any of the other sites.
- It is also shown that the later flights from the launch sites VAFB SLC 4E and KSC LC 39A had a higher success rate than the earlier flights.

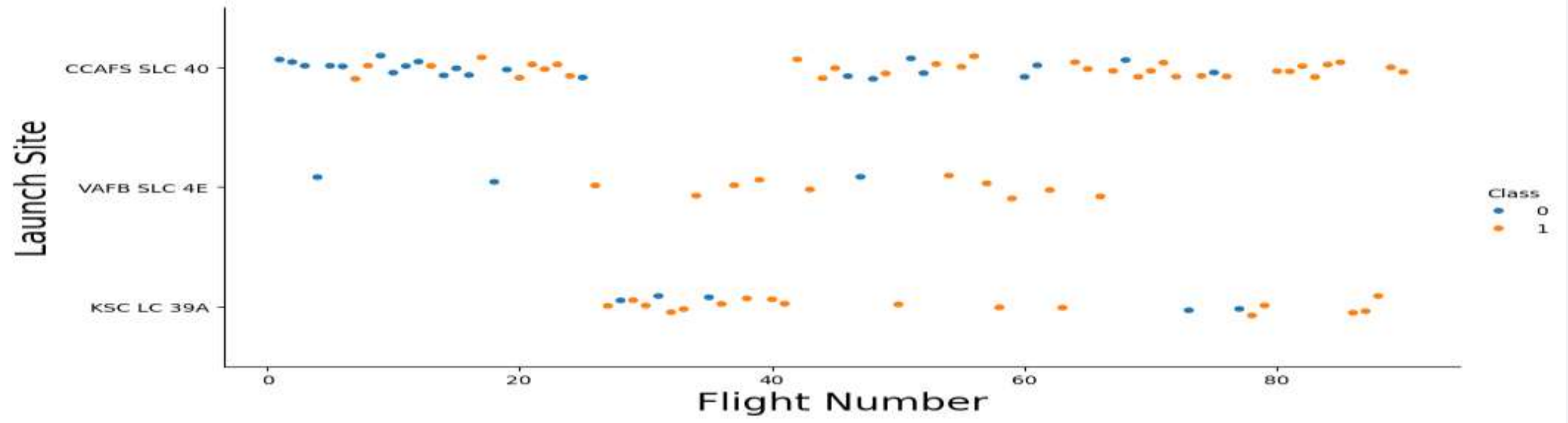


Figure 5. Scatter plot of Flight Number Vs. Launch Site

Payload vs. Launch Site

- The figure below shows that the VAFB-SLC 4E launch site has not launched any rockets with a payload mass exceeding 10,000 kg.
- It is also observed that the majority of rockets launched from all sites carry payloads under 9,000 kg.
- While both VAFB-SLC 4E and KSC LC 39A have handled heavy payloads, the CCAFS SLC 40 site shows a higher success rate for launches involving payloads of 14,000 kg and 16,000 kg.

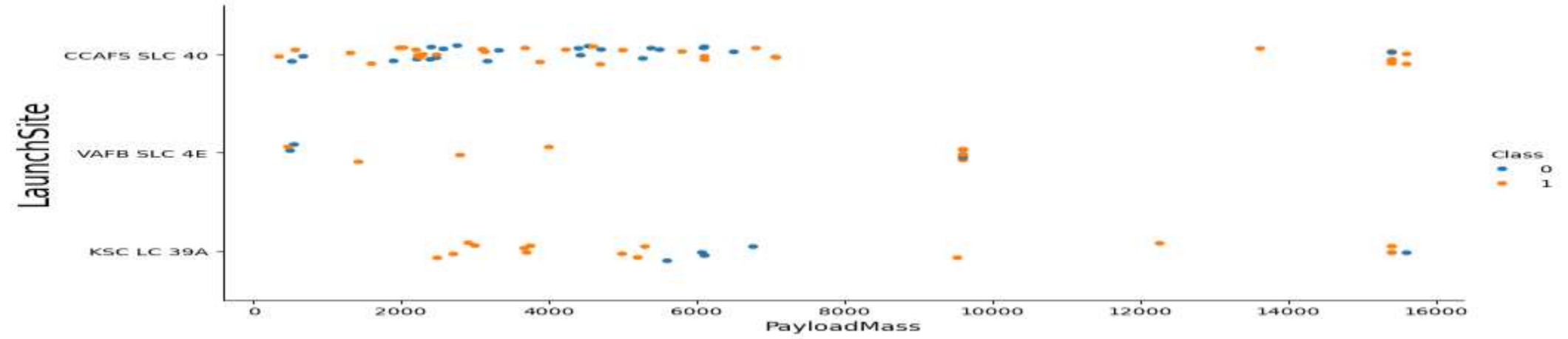


Figure 6. Scatter plot of Payload Vs. Launch Site

Success Rate vs. Orbit Type

- **ES-L1, GEO, HEO, SSO** all show a perfect success rate of **1.000**, indicating flawless landing outcomes for missions targeting these orbits.
- These orbits may involve fewer launches but demonstrate **high reliability**, possibly due to mature mission profiles or optimized vehicle configurations.

Figure 7. Bar plot of Success rate Vs. Orbit Type

Payload vs. Orbit Type

- There is a higher success rate for rockets with lighter payloads launched in ES-L1, HEO, and SO orbits.
- Rockets launched in GEO and VLEO orbits show mixed landing outcomes across all payload ranges.
- Rockets launched in GTO orbit display both successful and failed landings across a wide range of payload masses.

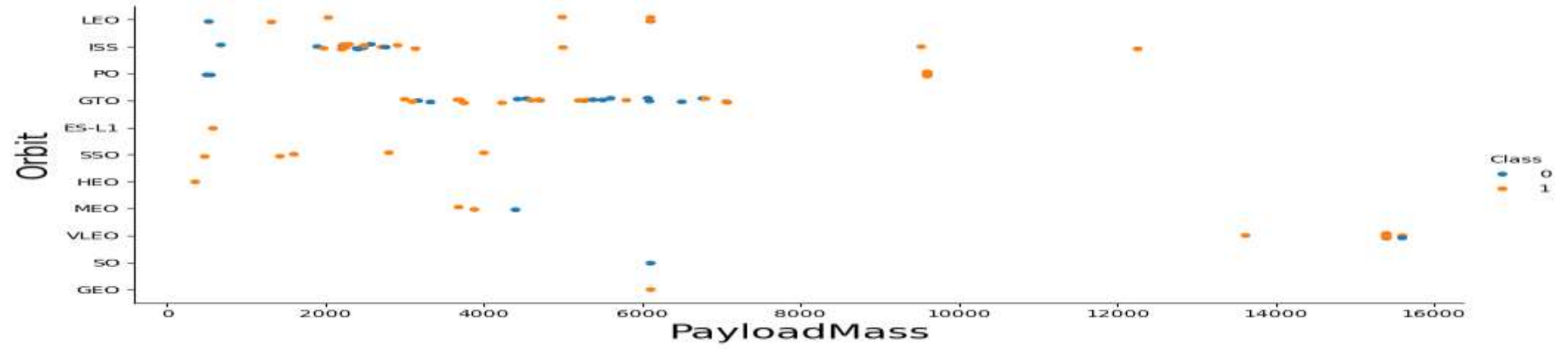


Figure 9. Scatter plot of Payload Vs. Orbit type

Launch Success Yearly Trend

- Figure 10 shows a clear upward trend in launch success rates from 2013 to 2020.

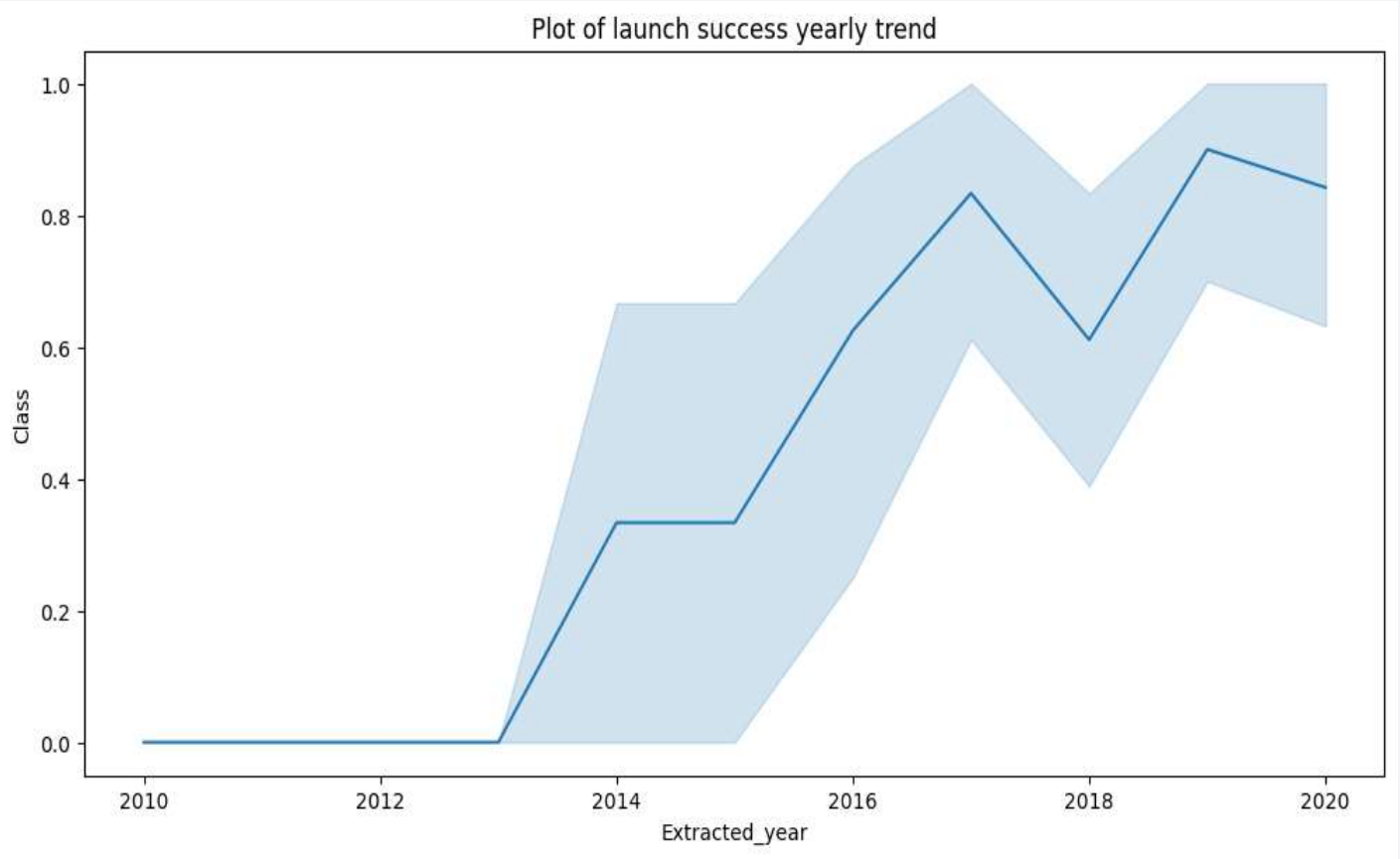


Figure 10. Line plot of Launch success Yearly Trend

All Launch Site Names

- An SQL table named SPACEXTBL was created using the existing data frame.
- To identify the unique launch sites, the DISTINCT keyword was applied to the relevant column.

```
]: %sql select distinct(LAUNCH_SITE) from SPACEXTBL

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

]:

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

Figure 11. SQL query for unique launch site names

Launch Site Names Begin with 'CCA'

- Keyword LIKE `CCA%` was used to get launch site names beginning with `CCA`.
- LIMIT 5 keyword was used to display only 5 records

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

Figure 12. SQL query for Launch site names beginning with CCA

Total Payload Mass

- SUM function was used to calculate the total payload mass of customers with the name ‘NASA (CRS)’.

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

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

sum(PAYLOAD_MASS__KG_)
45596

Figure 13. SQL query for Total payload Mass for NASA (CRS)

Average Payload Mass by F9 v1.1

The AVG function was used to calculate the average payload mass carried by the booster version F9 v1.1.

```
%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

Figure 14. SQL query for Average payload Mass for F9 V1.1 3

First Successful Ground Landing Date

- An SQL query was executed to find the first successful landing on the ground pad.
- The result indicates that the first successful ground landing occurred on December 22, 2015.

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

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

min(DATE)
2015-12-22

Figure 15. SQL query for First successful ground landing date

Successful Drone Ship Landing with Payload between 4000 and 6000

- Using the keywords BETWEEN and AND, the names of boosters that successfully landed on a drone ship and had payload mass greater than 4,000kg but less than 6,000kg were displayed.
- The result includes four rockets.

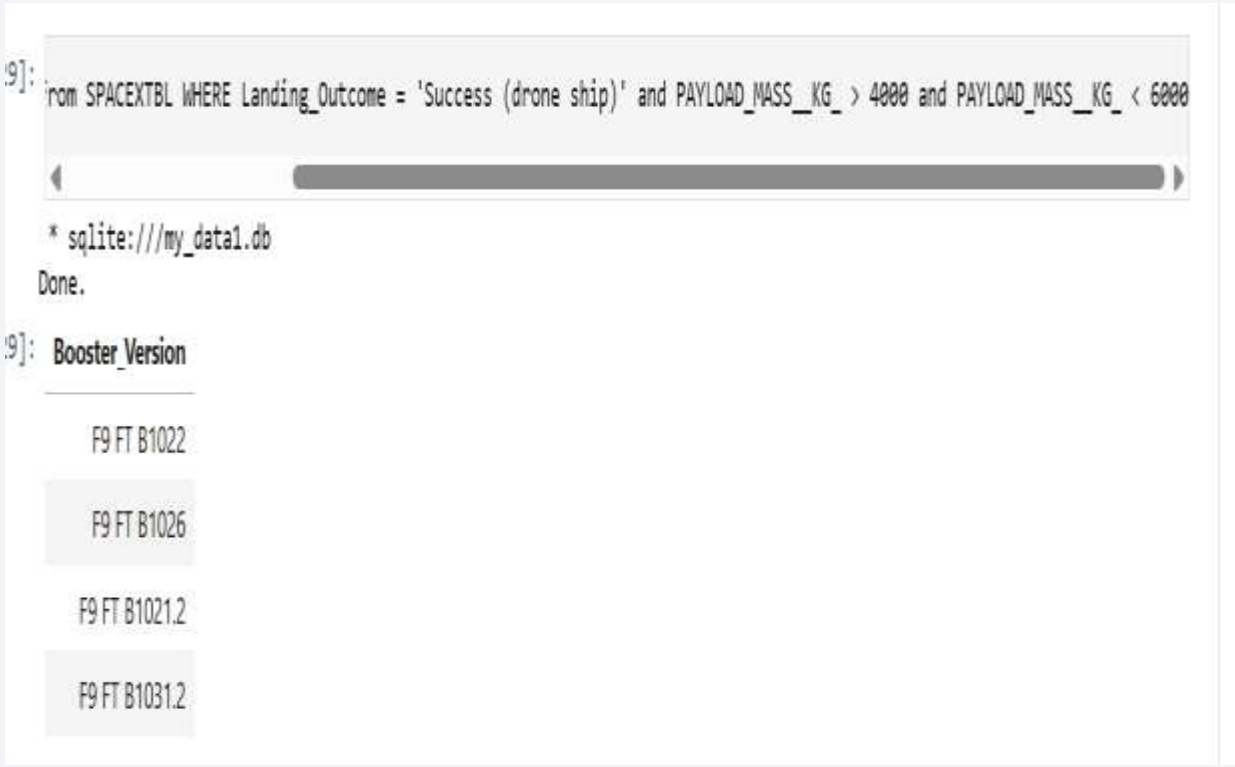


Figure 16. SQL query for successful drone ship landing between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

- The COUNT function was used to count the total number of successful and failed missions.
- The results show that there were 100 successful missions and 1 failed mission.

```
%%sql
SELECT Mission_Outcome, COUNT(*) AS Total
FROM SPACEXTBL
WHERE Mission_Outcome IN ('Success', 'Failure (in flight)')
GROUP BY Mission_Outcome;
```

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

Mission_Outcome	Total
Failure (in flight)	1
Success	98

Figure 17. SQL query for successful drone ship landing between 4000 and 6000.

Boosters Carried Maximum Payload

- A subquery using the MAX function was employed to identify the boosters that carried the maximum payload.
- The results indicate a total of 12 such boosters

```
%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

Figure 18. SQL query for boosters that carried the maximum payload

2015 Launch Records

- The substr() function was used to extract the month and year from the Date column.
- The WHERE and AND keywords were applied to retrieve launch records of failed drone ship landings in 2015.
- The results show that the failed landings occurred in April (04) and January (01).

```
%sql select substr(Date,6,2) as month, date, 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	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)

Figure 19. SQL query for failed drone ship landings in 2015

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

- Keywords such as GROUP BY, ORDER BY, and DESC, along with functions like substr() and COUNT(), were used to rank the count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order.
- The results show a high number of occurrences for no attempt (10), successful landings on the drone ship (5), and successful ground landings (3).
- Additionally, there was one instance of failure due to parachute deployment.

```
%sql select landing_Outcome, count(*) as count_outcomes \
from SPACEXTABLE \
where date between '2010-06-04' and '2017-03-20' \
group by landing_Outcome \
order by count_outcomes desc;
```

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

Landing_Outcome	count_outcomes
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

Figure 20. SQL query ranking landing outcomes between 2010-06-04 and 2017-03-20

Section 3

Launch Sites Proximities Analysis

Launch Site Locations for Space X Falcon 9

- All of the launch sites shown in the figure are situated in coastal cities within the United States of America.

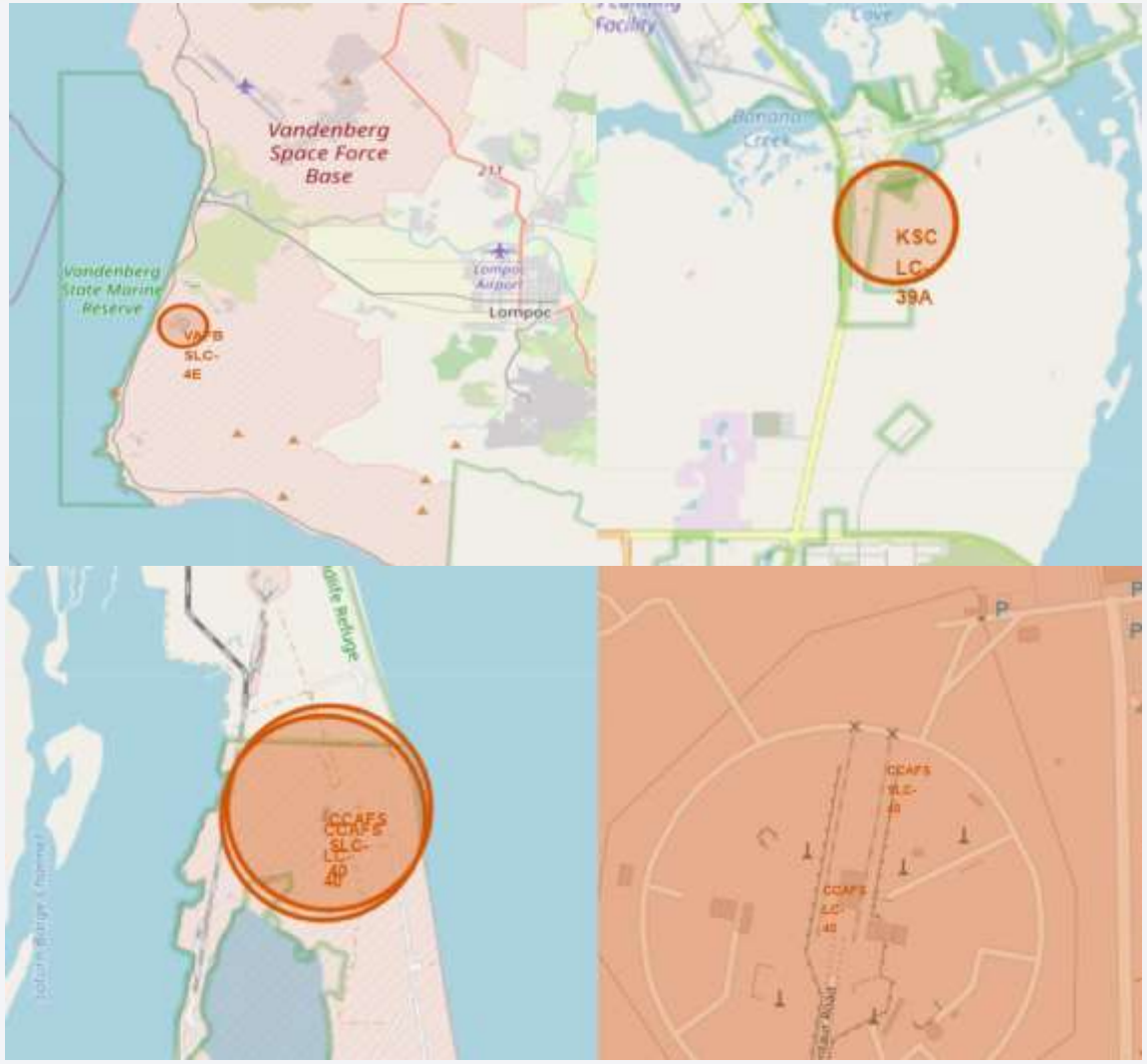


Figure 21. Folium Map showing locations of Space X Falcon 9 launch sites

Launch Outcomes for Space X Falcon 9

- The figure shows the launch outcomes for various launch sites;
  - Top left: VAFB SLC-4E
  - Top right: KSC LC-39A
  - Bottom left: CCAFS SLC-40
  - Bottom Right: CCAFS LC-40
- Red icons indicate the failed outcomes and the green icons indicate successful outcomes.

Figure 22. Folium Map showing Launch outcomes in their various launch sites

Launch site distance from coastline, cities, railways and highways.

- From the figure it is shown that launch sites are located very close to the coast i.e. 0.95km from CCAFS SLC 40 and 1.52km from VAFB SLC 4E
- The same can not be said for some railways and highways
- It is also evident that launch sites are located far from major cities, i.e. VAFB SLC 4E is 38.16km away from its closest city Santa Maria and CCAFS SLC 40 is 56.04km away from Melbourne

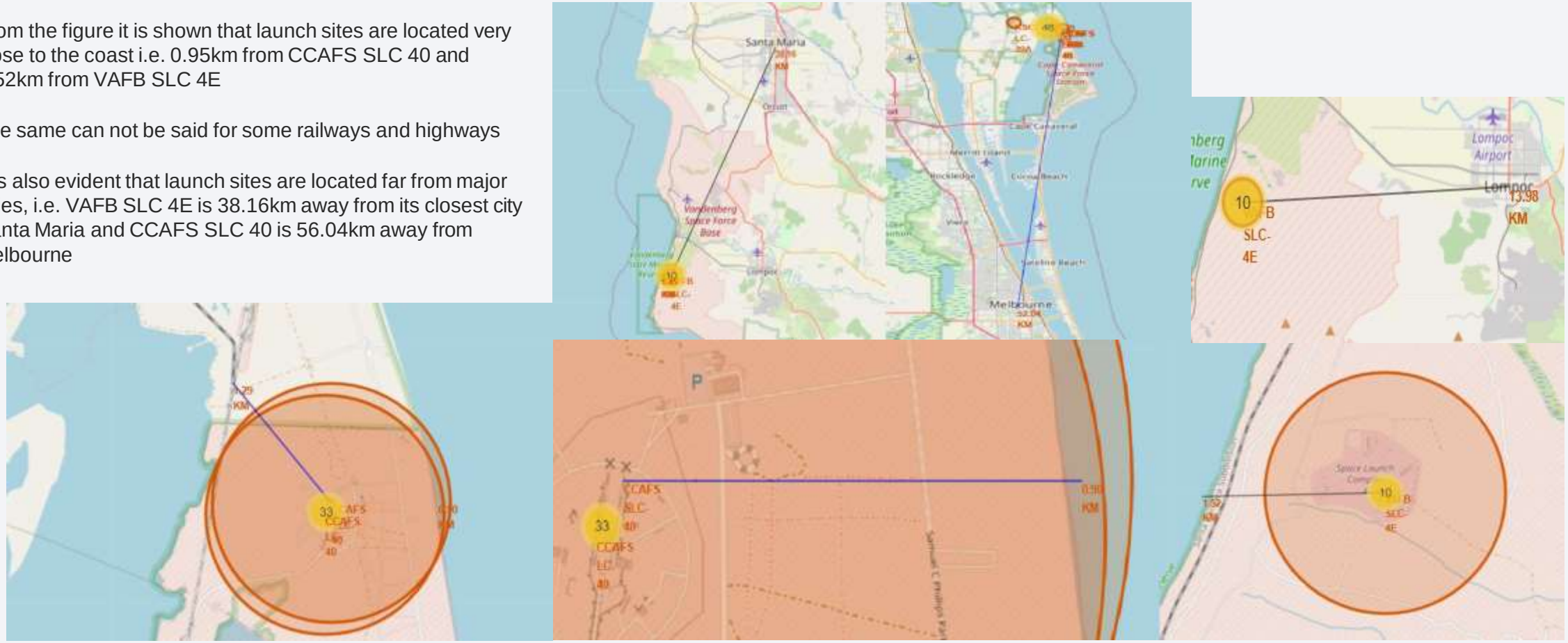
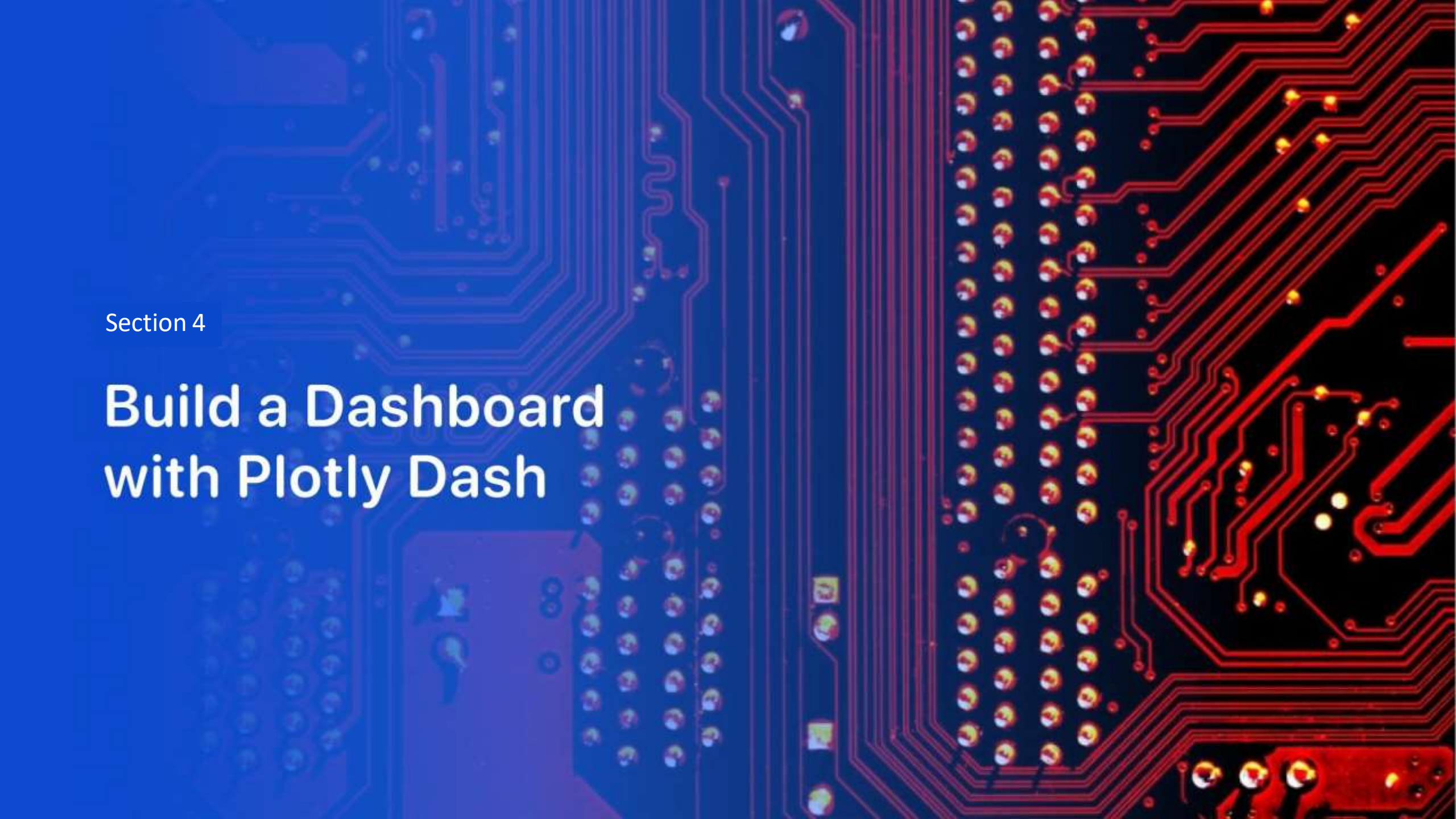


Figure 23. Folium Map showing Launch site distances from coastlines, cities, railways and highway

Section 4

Build a Dashboard with Plotly Dash



Pie Chart of Launch Success for all Sites

- From Figure 24, it is shown that KSC LC-39A has the largest success rate with about 41.2% of the total success ratio with other sites.

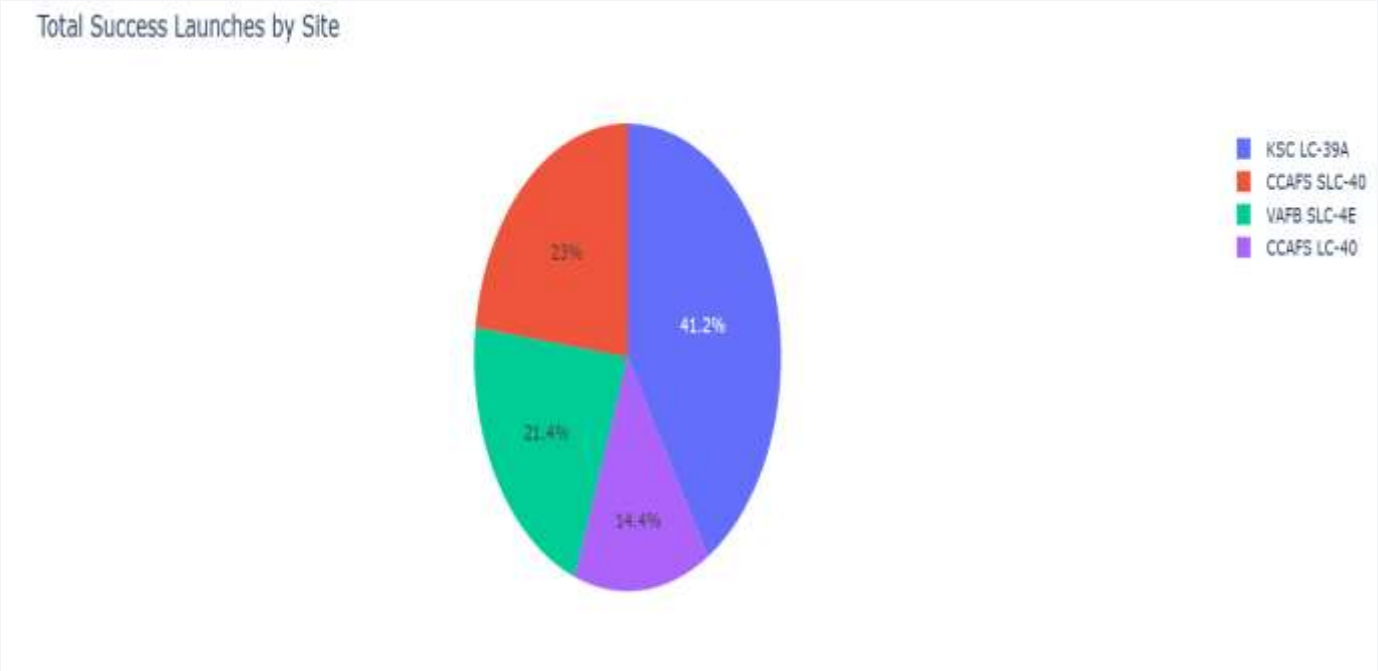


Figure 24. Pie showing the Success rate of all Launch sites

Pie chart of Launch site with highest success ratio

- Figure 25 also clearly shows that KSC LC-39A has the highest success rate, approximately 76.9%, compared to the other sites
  - 73.1% for CCAFS LC-40
  - 60% for VAFB SLC-4E
  - 57.1% for CCAFS SLC-40

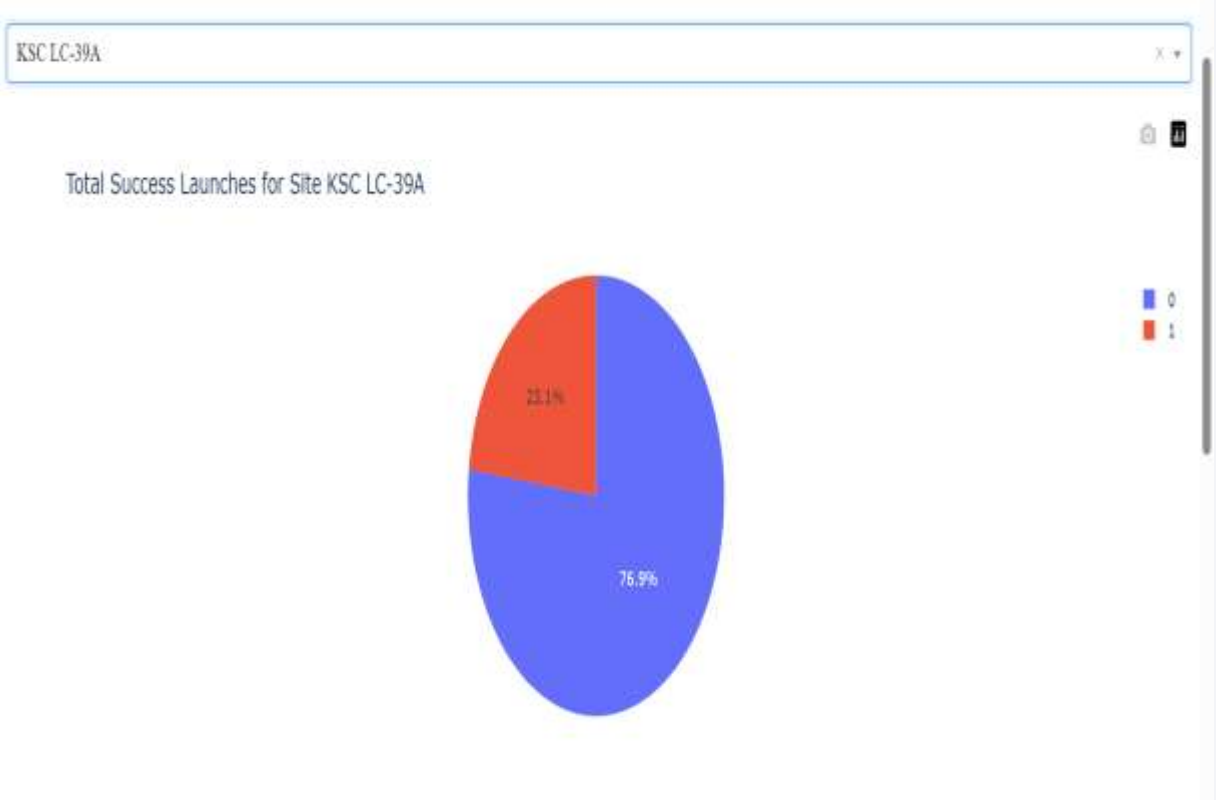


Figure 25. Pie showing the launch site with the highest success ratio

Payload vs Launch outcome for all sites

- The figures below indicate that Booster version FT has the highest success rate with payload masses ranging from approximately 700 kg to 5,500 kg.
- It is also evident that rockets carrying payloads above 5,500 kg have a lower success rate, suggesting that heavier payloads reduce the likelihood of a successful outcome.

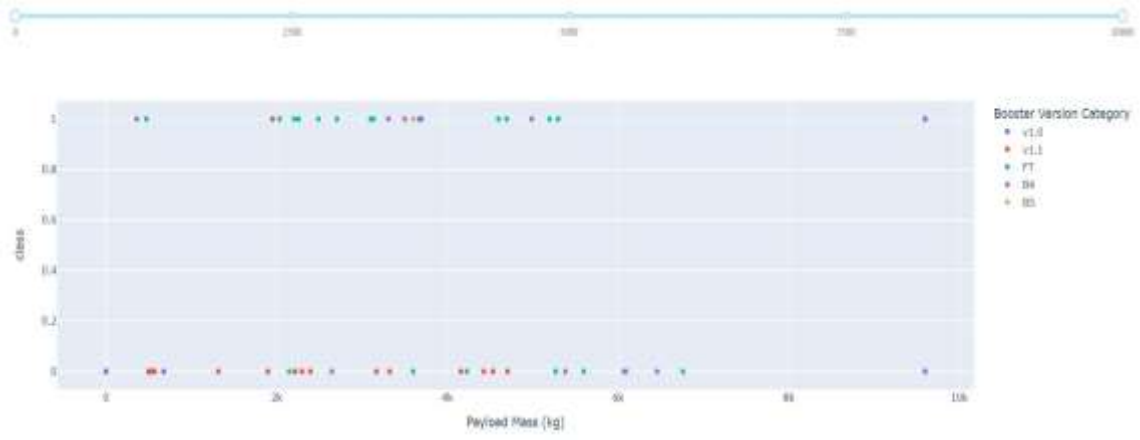


Figure 26. Scatter plot showing the booster versions with different payload mass for all launch sites

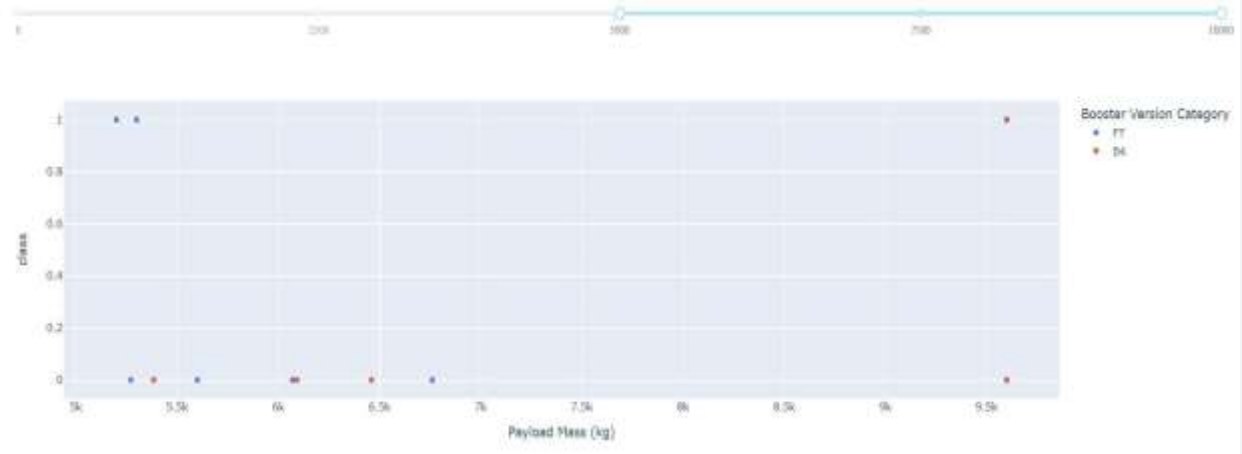


Figure 27. Scatter plot showing the booster versions of different payload mass greater than 5,500kg.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- From the bar chart in Figure 28, the Decision Tree classifier achieved the highest performance with an accuracy score of approximately 0.887, or about 88%

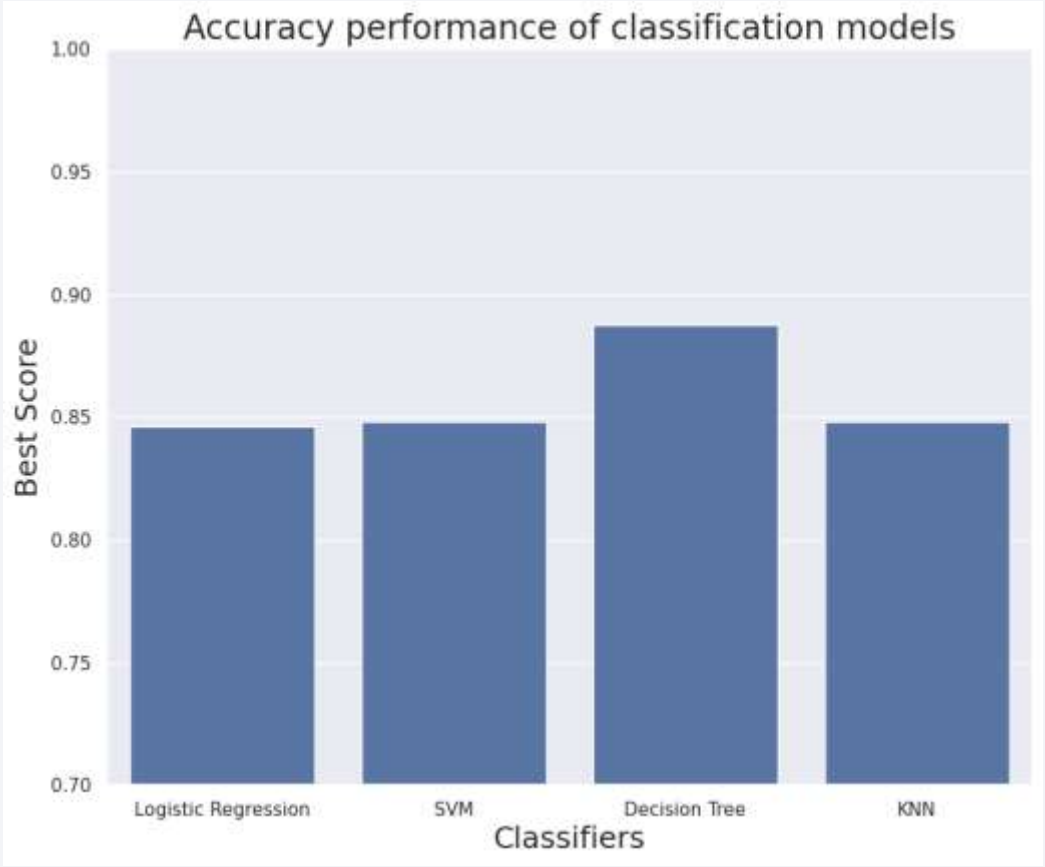


Figure 28. Bar plot showing different classifiers and their accuracies.

Confusion Matrix

- After splitting the dataset into training and test sets, there were only 18 samples in the test set.
- From these, the decision tree classifier correctly predicted 12 landings (12 True Positives) and 3 non-landings (3 True Negatives).
- The classifier had 0 False Negatives, meaning it did not mistakenly predict any successful landings as failures.
- However, it had 3 False Positives, where it incorrectly predicted a successful outcome for 3 observations that were not successful.

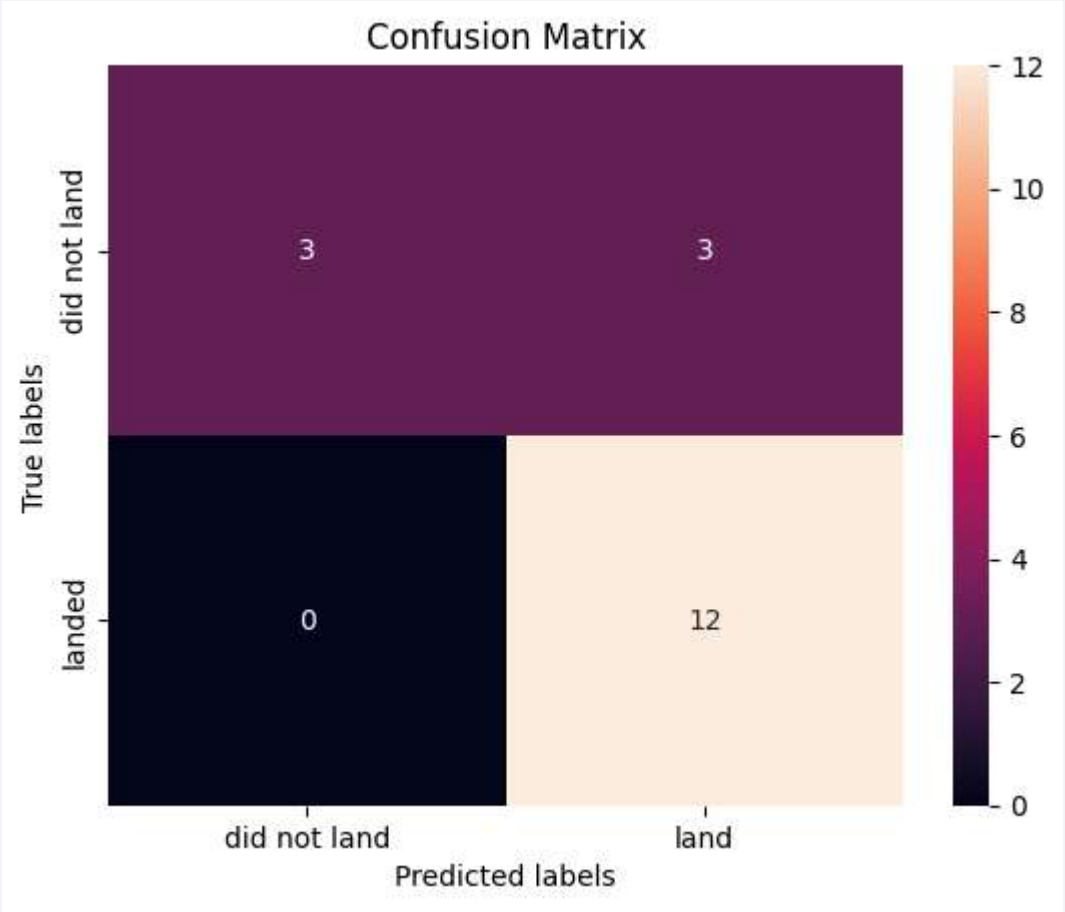


Figure 29. Confusion matrix of decision tree classifier.

Conclusions

- Payload mass is an important factor for mission success, as rockets with smaller payloads tend to have higher success rates.
- Orbit type also influences success, with rockets launched to orbits such as VLEO, ES-L1, GEO, HEO, and SSO showing higher success rates compared to other orbits.
- Launch sites are strategically located in coastal cities to facilitate easy retrieval and recovery, and are positioned away from busy areas like major highways and cities to reduce the risk of casualties in case of failure.
- Recent years have seen improved outcomes, with later flight launches achieving higher success rates.
- Among classification algorithms, decision tree classifiers performed best, achieving approximately 88% accuracy, making them a strong model choice for predicting landing outcomes.

Thank you!