IBM Developer
SKILLS NETWORK

Space X Falcon 9 First Stage Landing Prediction

Winning Space Race with Data Science

Wahyu Muhammad

May 13th, 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data was collected from SpaceX API and Wikipedia through a web-scraping process. Data wrangling was performed to identify patterns and explore all variables for supervised machine learning.
  - SQL language was used in exploratory data analysis (EDA) along with static and interactive dashboard data visualization in order to get insights.
  - Analytical predictions were performed with supervised algorithms such as Logistic Regression, SVM, Decision Tree, and KNN. To get the optimal model, these algorithms performed hyperparameter grid search optimization.

- Summary of all results
  - Several variables play an important role in a rocket launch such as booster version, payload mass, orbit, etc.
  - The selected algorithms have the same result in predicting test data.

Introduction

- Project background and context

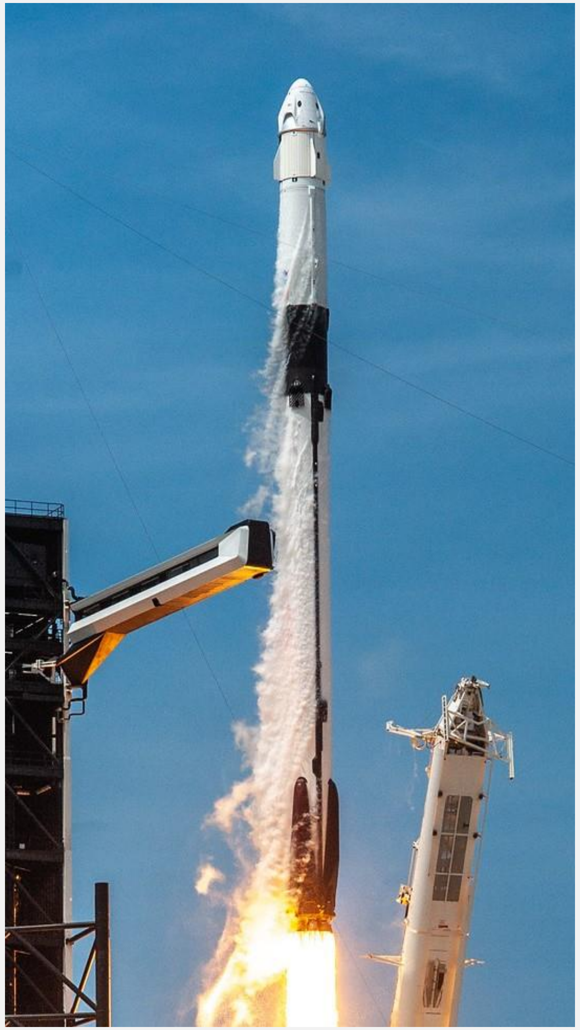
  SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. If we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- Problems

  - What are the most important factors predicting the success of a Falcon 9 rocket launch?

  - How does the performance of Falcon 9 rocket launches change over time? Are there any trends or patterns?

  - What machine learning algorithms can accurately predict the outcomes of Falcon 9 rocket launches?



Section 1

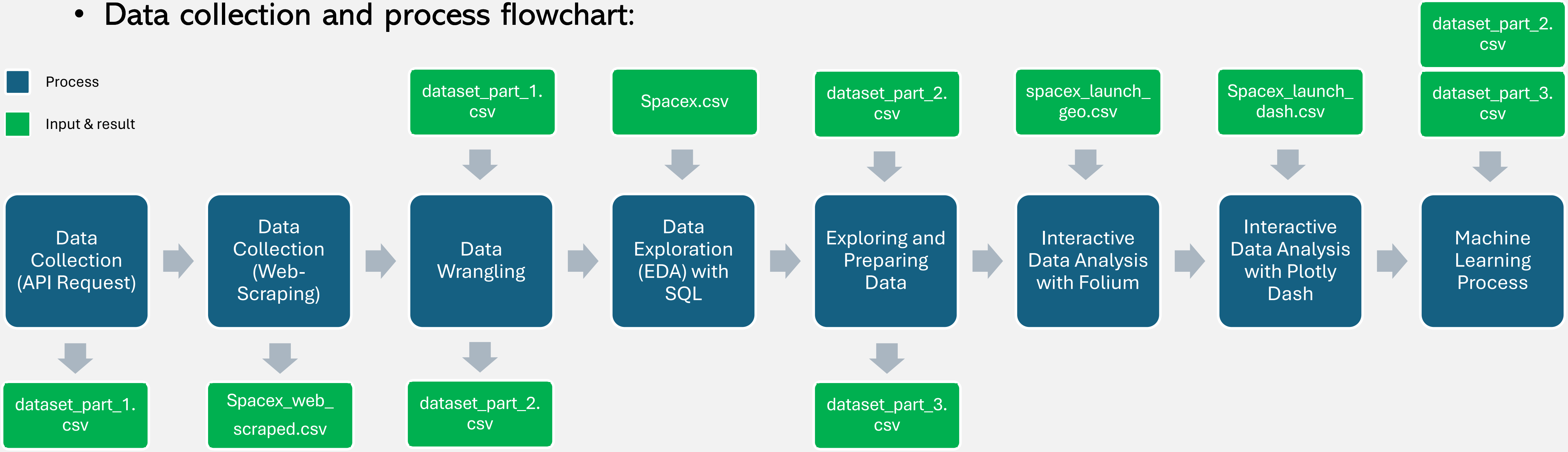
Methodology

Methodology

- Data collection methodology
  - Data was collected from the SpaceX API and Wikipedia through web-scraping process.
- Perform data wrangling
  - Data was processed by first analyzing all variables especially the launch outcomes and the variable Y was created to represent the classification variable.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Select a classification algorithm, select optimal parameters by using hyperparameter tuning, and evaluate which model gets highest accuracy on the dataset.

Data Collection

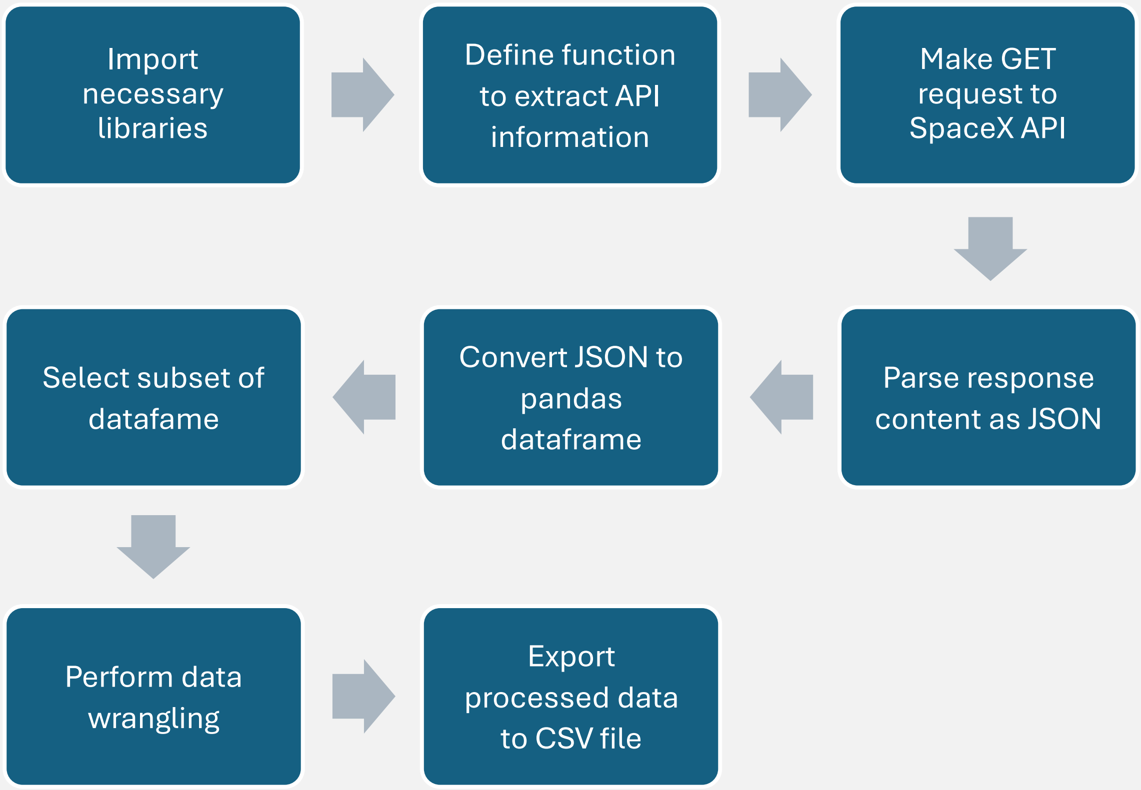
- Launch dataset were collected from:
    - SpaceX API (https://api.spacexdata.com/v4) by request and parse the launch data from JSON into pandas datafame
    - Converted HTML table from static Wikipedia webpage (date: 9th June 2021) (URL) through web-scraping process
    - IBM developer web data storage
- Data collection and process flowchart:



Data Collection – SpaceX API

- Access data through GET request to the SpaceX API using a predefined function to collect multiple dataset. Convert the obtained JSON data into a pandas dataframe.
- Perform necessary data wrangling tasks to clean and preprocess the data. Export the processed data for further analysis.
- GitHub URL = Data Collection API Lab

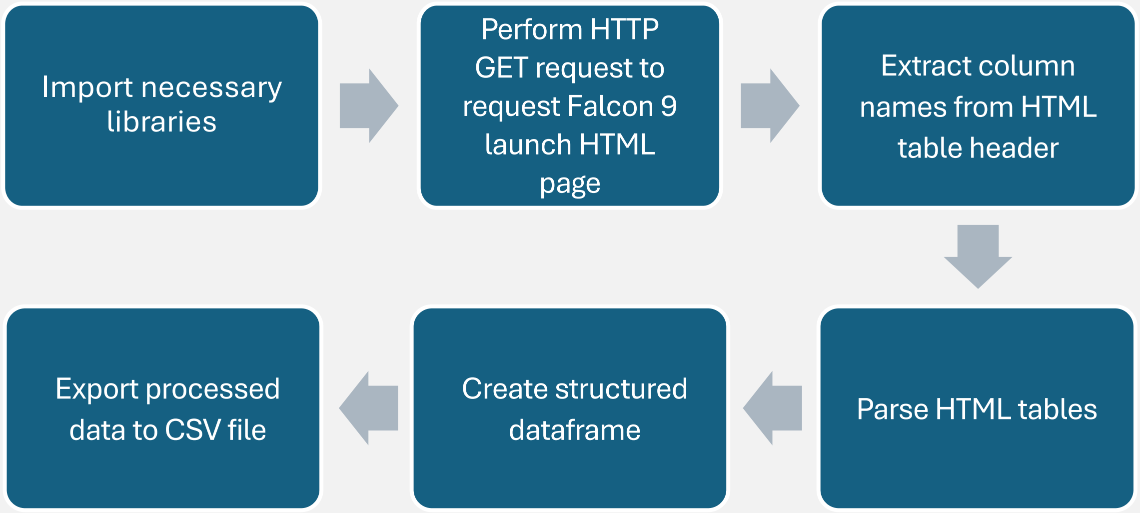
SpaceX API data collection flowchart



Data Collection – Scraping

- Perform an HTTP GET method to request the Falcon9 Launch HTML at Wikipedia. Extract all column or variable names from the HTML table header.
- Create a DataFrame by parsing launch HTML tables. Export the processed data for further analysis.
- GitHub URL = [Data Collection with Web Scraping Lab]

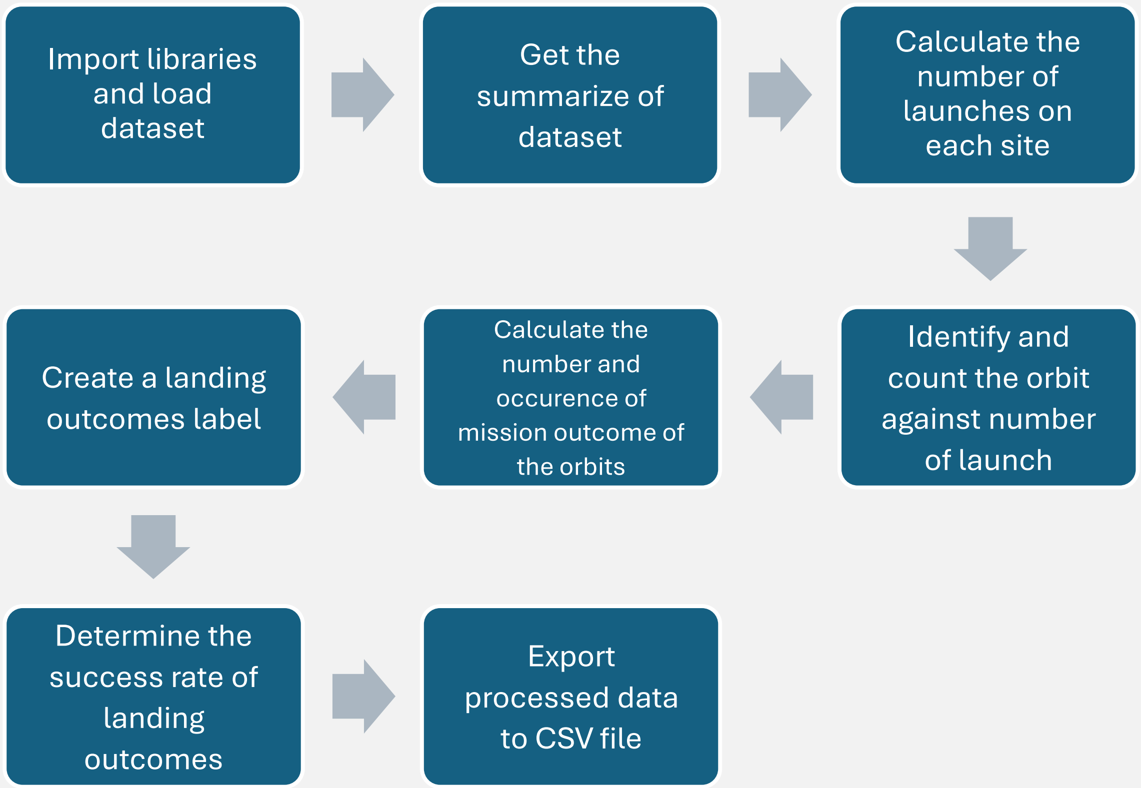
SpaceX data collection through web scraping flowchart



Data Wrangling

- Perform exploratory data analysis to get insights from dataset. Summarize the statistics of dataset and examine of specific features or columns.
- Create the landing outcomes ‘Class’ column from ’Outcome’ represented by binary values (0 = ‘unsuccessfully landed’ and 1 = ‘successfully landed’). Export the processed data for further analysis.
- GitHub URL = [Data Wrangling]

Data wrangling flowchart



EDA with Data Visualization

- Scatter, categorical, bar, and line charts are used in Exploratory Data Analysis (EDA).
- The following table shows the types of charts, descriptions and charts that are in the EDA process.

Chart Type	Description	Chart Title
Scatter	Visualize the relationship between two variables	Flight Number vs. Payload Mass
Categorical	Examine the distribution of categorical variables and identify patterns	Flight Number vs. Launch Site Payload Mass vs. Launch Site Flight Number vs. Orbit Type Payload Mass vs. Orbit Type
Bar	Compare the value or frequency between categorical variables	Success Rate by Orbit Type
Line	Identify trends or patterns in over short and long periods of time	Success Rate by Year

- GitHub URL = [EDA with Visualization Lab]

EDA with SQL

- By using SQL language, there are some output and specific information about:
  - Launch sites
  - Payload mass
  - Customers
  - Booster versions
  - Mission and landing outcomes
  - Dates
- GitHub URL = [Complete the EDA with SQL]

Build an Interactive Map with Folium

- Object list at Folium map:
  - Circles represented the NASA Johnson Space Center and launch site locations.
  - Colored markers were used at launch sites to indicate the count of launch results.
  - Lines and markers were used to illustrate the distance between the CCAFS SLC-40 launch site and the nearest railway, highway, and city.
- GitHub URL = Interactive Visual Analytics with Folium lab

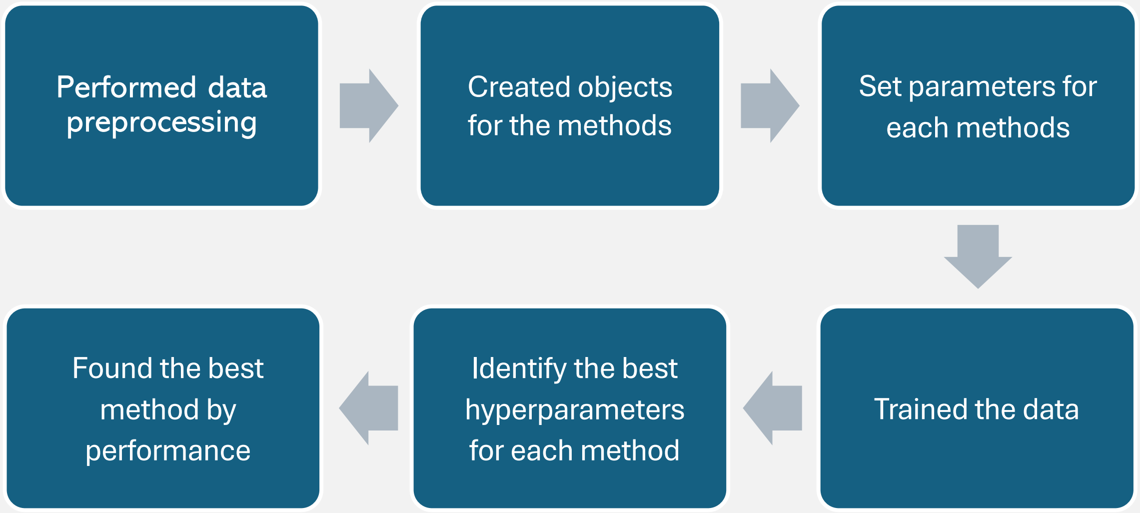
Build a Dashboard with Plotly Dash

- The Plotly Dash interactive dashboard elements:
  - Input dropdown featured all launch sites with specific options to display both a pie chart and scatter plot.
  - Pie chart showed the total successful launches across all sites or for the selected launch site.
  - Input slider to filter payload mass values within a specified range.
  - Scatter plot visualized the correlation between payload mass and launch results categorized by booster version, filtered by the slider.
- GitHub URL = Build an Interactive Dashboard with Ploty Dash

Predictive Analysis (Classification)

- Performed data preprocessing by creating a target/dependent variable column, standardizing the data, and splitting it into training and test datasets.
- Created objects for the methods - Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-nearest Neighbors (KNN) - set the parameters, and trained the data.
- Identified the best hyperparameters for the methods using *best_params_*.
- Found the method that performed best by accuracy score and made conclusions from it.
- GitHub URL = [Complete the Machine Learning Prediction lab]

Predictive analysis (classification) flowchart

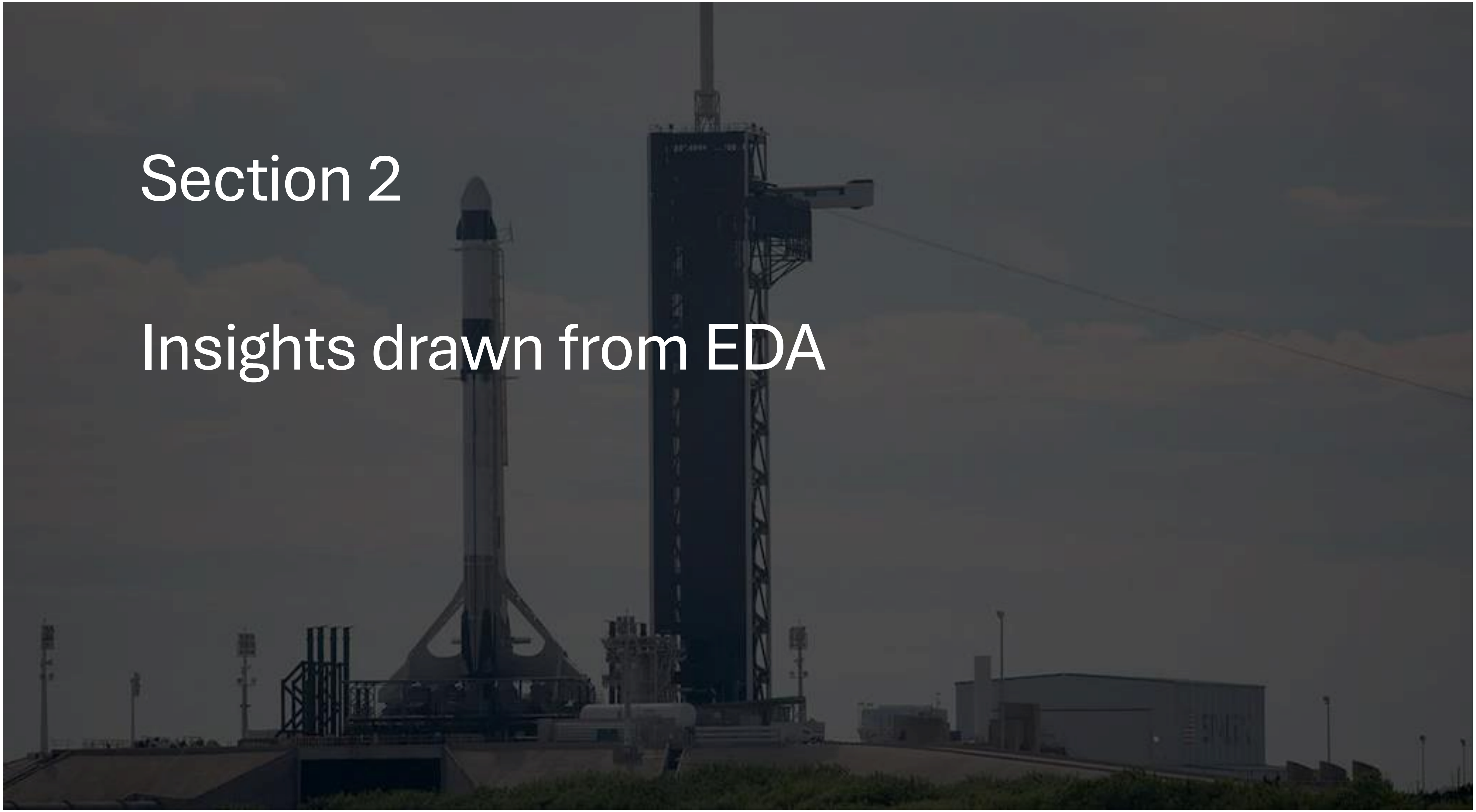


Results

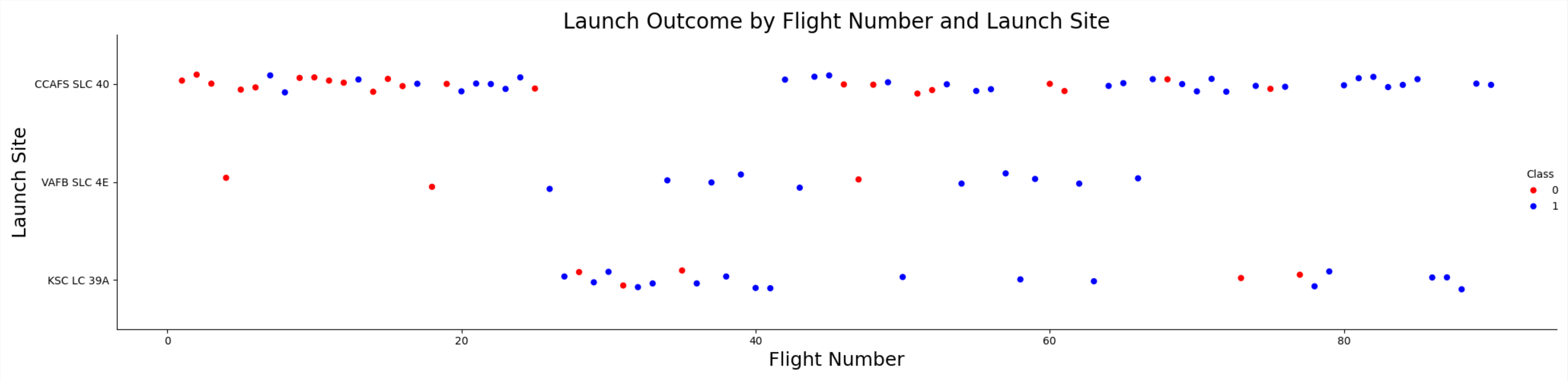
- Insights from exploratory data analysis
- Interactive and static data visualization dashboards
- Predictive analysis results

Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

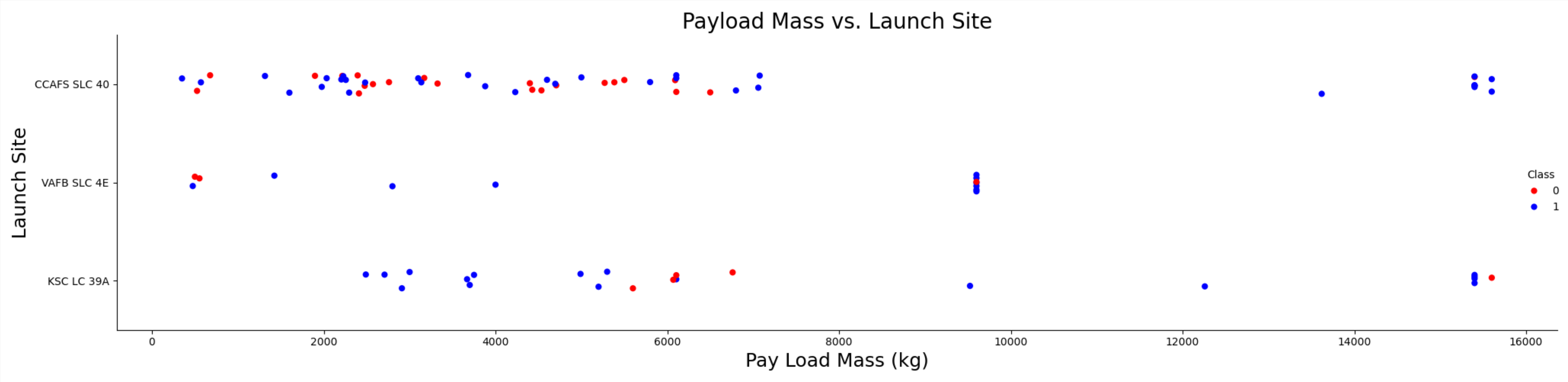


Class 0 : failed launch outcome
Class 1 : successful launch outcome

- Insight:

The mission frequency at the VAFB SLC 4E was lower compared to other launch sites. With lower success rate, CCAFS SLC 40 was used in several early 20s missions before moving to KSC LC 39A. There is a trend of increasing mission success as the increased flight number from all launch sites.

Payload vs. Launch Site



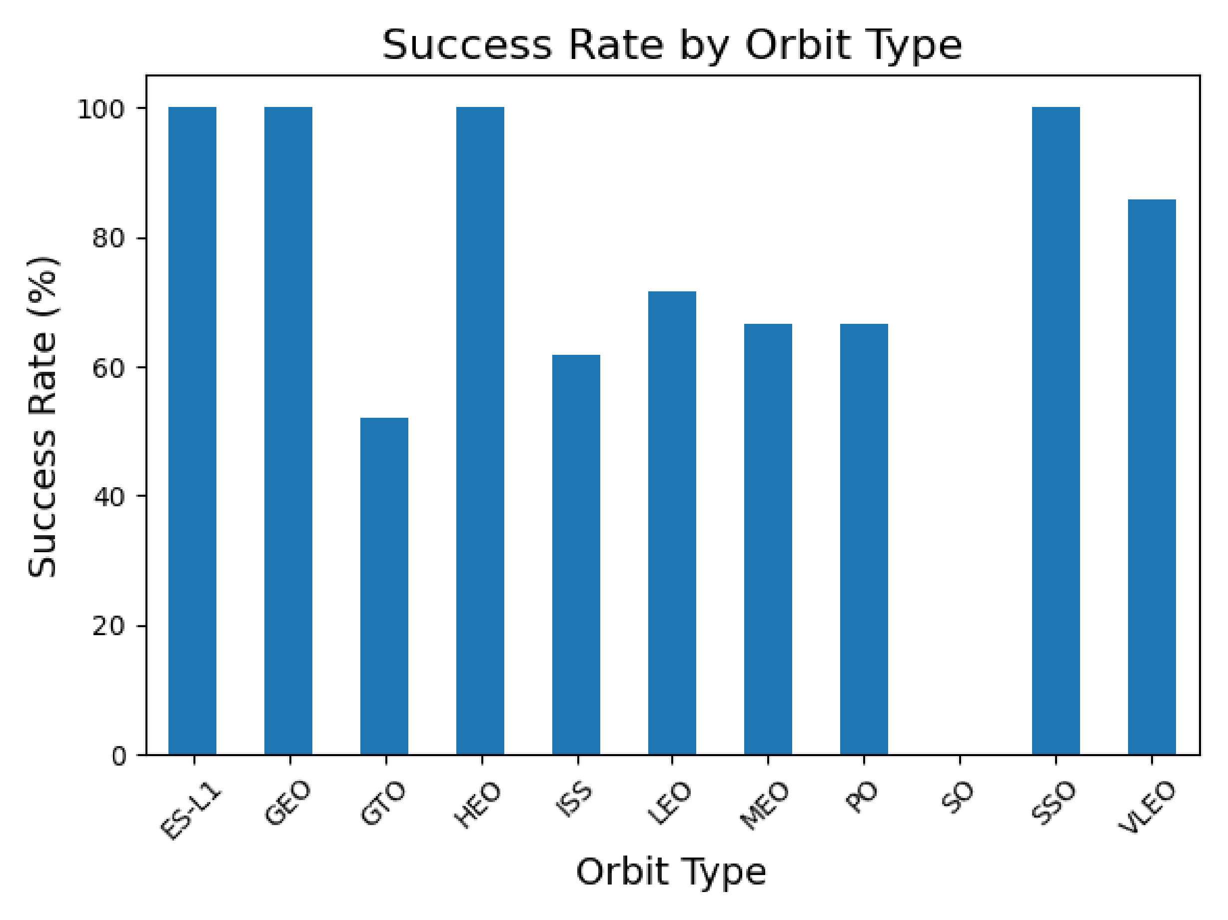
Class 0 : failed launch outcome
Class 1 : successful launch outcome

- Insights:
- There are no rockets launched for heavy payload mass (greater than 10,000 kg) at the VAFB-SLC launch site.
- For rocket launches with heavy payloads, they are at other launch sites with only two failed launch outcomes from 15 total launches.

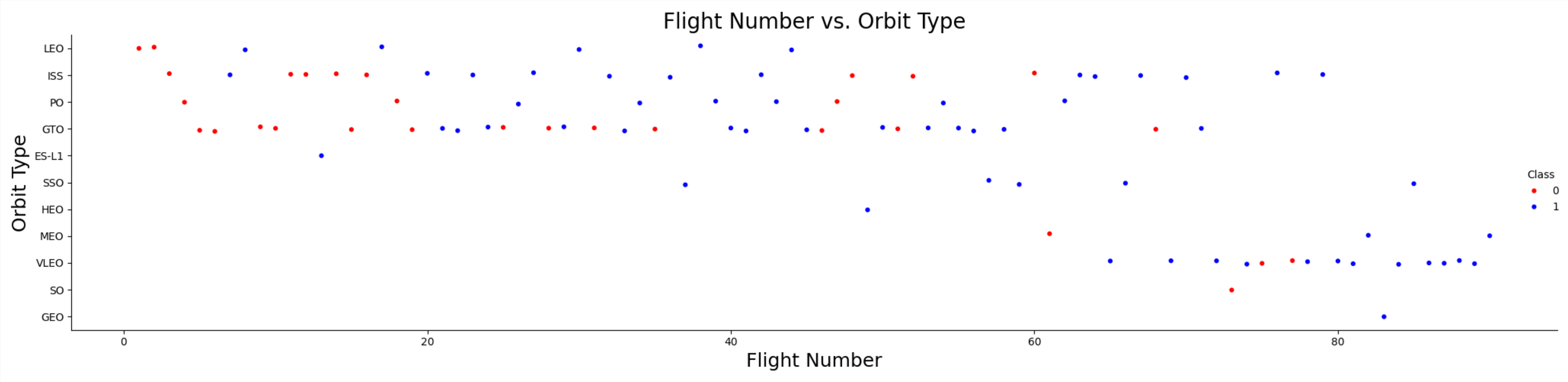
Success Rate vs. Orbit Type

- Insight:

The missions to ES-L1, GEO, and SSO orbits achieved a success rate of 100%, followed by VLEO and LEO orbit with lower success rate.



Flight Number vs. Orbit Type

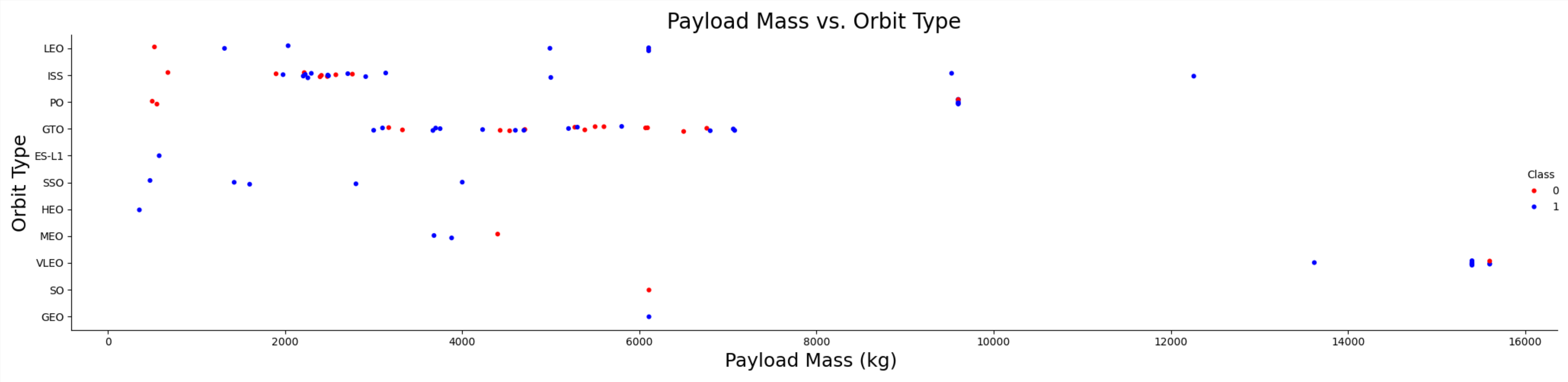


Class 0 : failed launch outcome
Class 1 : successful launch outcome

- Insight:

In the LEO orbit, there appears to be a relationship between successful missions and flight number. However, in the GTO orbit, there appears to be no relationship between the flight number and the mission success rate.

Payload vs. Orbit Type



Class 0 : failed launch outcome
Class 1 : successful launch outcome

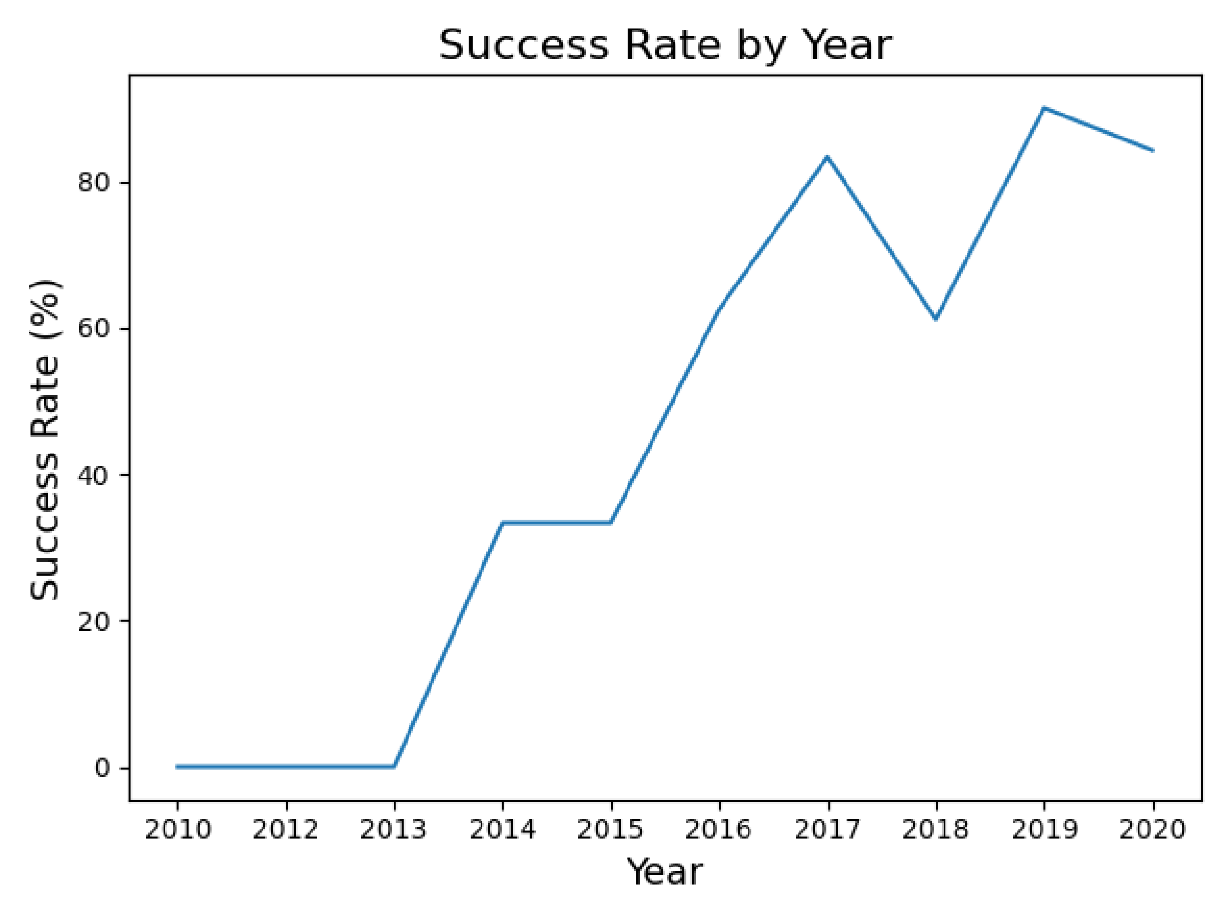
- Insight:

With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- Insight:

The success rate grows significantly each year. The highest success rate was achieved in 2019.



All Launch Site Name

```
[10]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE

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

[10]:

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

Explanation:

There are 4 launch sites for Falcon 9 rocket. They are CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.

Launch Site Name Begin with ‘CCA’

```
[11]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5
 * sqlite:///my_data1.db
Done.
```

[11]:

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

Explanation:

Displayed 5 records where launch sites begin with the string 'CCA’, sorted by launch date.

Total Payload Mass

```
[12]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS NASA_CRS_total_payload_mass FROM SPACEXTABLE where Customer = 'NASA (CRS)'

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

[12]:

NASA_CRS_total_payload_mass
45596

Explanation:

The total payload mass carried by boosters launched by NASA (CRS) is 45,596 kg.

Average Payload Mass by F9 v1.1

```
[13]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS average_payload_mass FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'

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

[13]:

average_payload_mass
2928.4

Explanation:

The average payload mass carried by booster version F9 v1.1 is 2,928.4 kg.

First Successful Ground Landing Date

```
[14]: %sql SELECT MIN(Date) AS first_succesful_landing FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'

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

[14]:

first_succesful_landing
2015-12-22

Explanation:

The first succesful landing outcome in ground pad was achieved at December 22th, 2015.

Successful Drone Ship Landing with Payload between 4,000 and 6,000

```
[15]: %sql SELECT Booster_Version FROM SPACEXTABLE\
      WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ between 4000 AND 6000

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

[15]:

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

Explanation:

The names of the boosters which have success in drone ship and have payload mass greater than 4,000 but less than 6,000.

Total Number of Successful and Failure Mission Outcomes

```
[16]: %sql SELECT DISTINCT Mission_Outcome, COUNT(Mission_Outcome) AS number_of_outcomes\
      FROM SPACEXTABLE GROUP BY Mission_Outcome

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

[16]:

Mission_Outcome	number_of_outcomes
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Explanation:

Displayed the total number of successful and failure mission outcomes. It shows mission outcome is dominated by successful outcome.

Booster Carried Maximum Payload

```
[17]: %sql SELECT Booster_Version FROM SPACEXTABLE\
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)

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

[17]:

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

Explanation:

Displayed the names of the booster versions which have carried the maximum payload mass. Falcon 9 Block 5 booster version dominates this list.

2015 Launch Records

```
[18]: %sql SELECT SUBSTR(Date,6,2) AS month_names, Landing_Outcome, Booster_Version, Launch_Site\
FROM SPACEXTABLE WHERE Landing_Outcome = 'Failure (drone ship)' AND SUBSTR(Date,0,5)='2015'

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

[18]:

month_names	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

Explanation:

There are 2 rows data in the failed landing outcome in drone ship at 2015. There were happened at January and April launched from CCAFS LC-40.

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

[19]:
```
%sql SELECT DISTINCT Landing_Outcome, COUNT(Landing_Outcome)\
FROM SPACEXTABLE WHERE Date > '2010-06-04' AND Date < '2017-03-20'\
GROUP BY Landing_Outcome ORDER BY COUNT(Landing_Outcome) DESC
```

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

[19]:

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

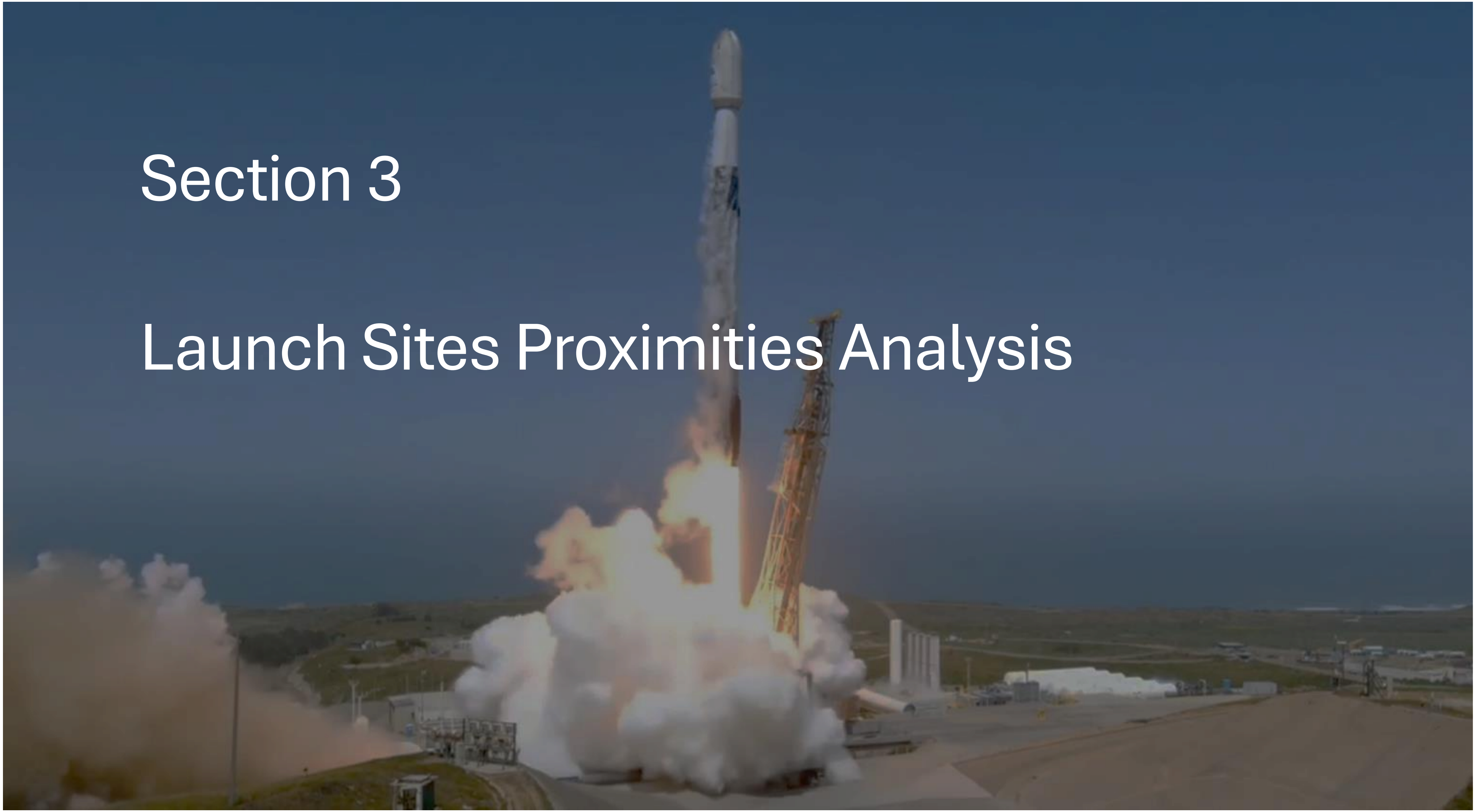
Explanation:

Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

No attempt landing outcome was the highest number, while successful and failed landing outcomes in drone ship has in equal number.

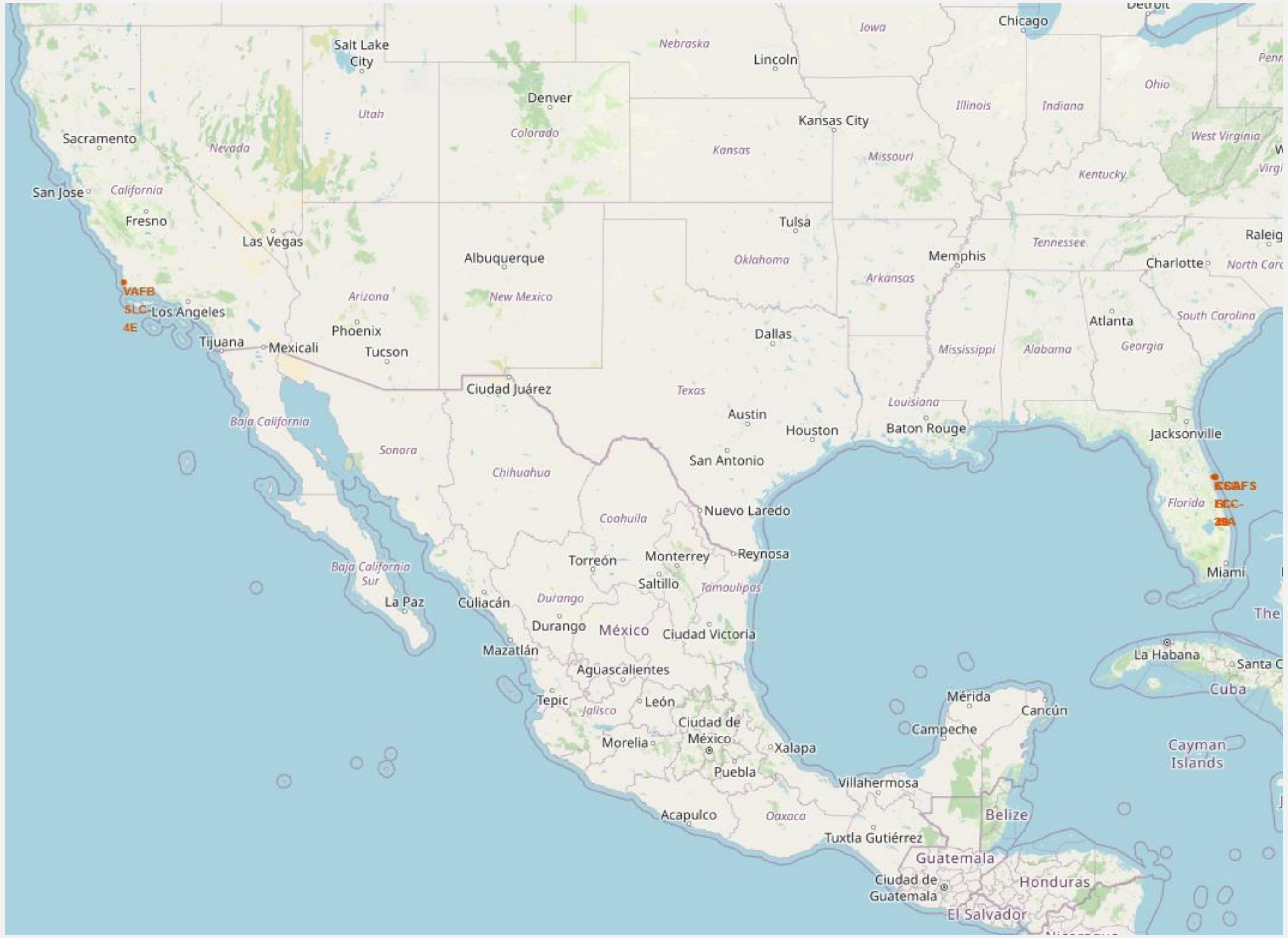
Section 3

Launch Sites Proximities Analysis



All Launch Sites on Map

- It is clear that all launch sites are close to the coast and the equator line. The VAFB SLC-4E launch site is on the west coast, while the other 3 launch sites are on the east.
- Launch sites near the coast allow quick access to transportation and provide a safe location for rocket debris to fall.
- To take advantage on the Earth's rotating speed, launch sites usually located close to the Equator. Rockets can reach orbit more efficiently because of the faster rotational speed near the equator, which enables them to achieve higher speeds while using less fuel.



Label Outcomes on Map

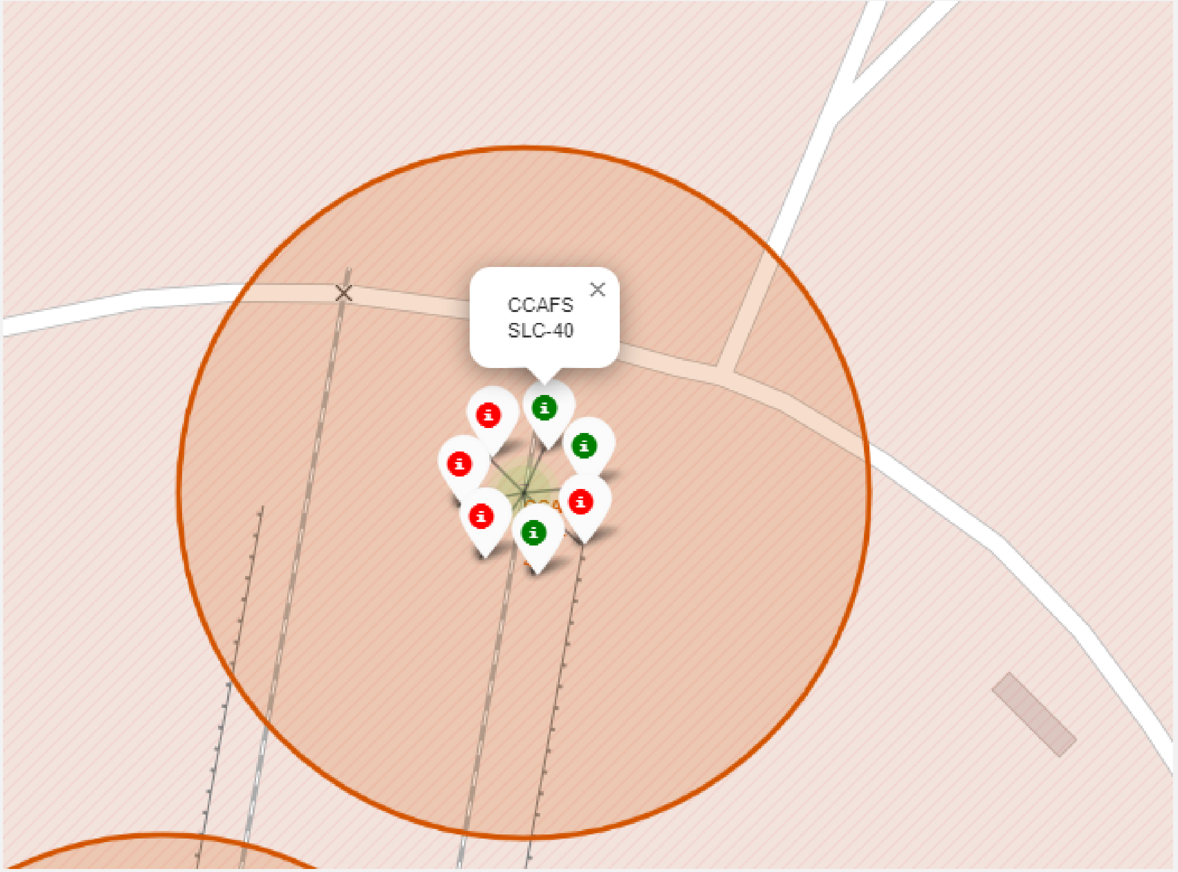
- The colored markers on each launch site can display each launch outcome.

  Green marker : successful launch outcome

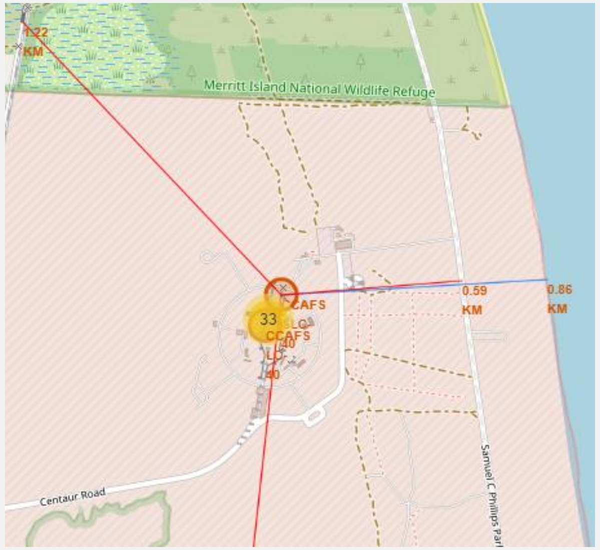
  Red marker : failed launch outcome

- For example:

  CCAFS SLC-40 launch has 42.86% success rate launch outcome from 3 successful launch and 4 failed launch.



CCAFS SLC-40 Launch Site with Several nearby Places



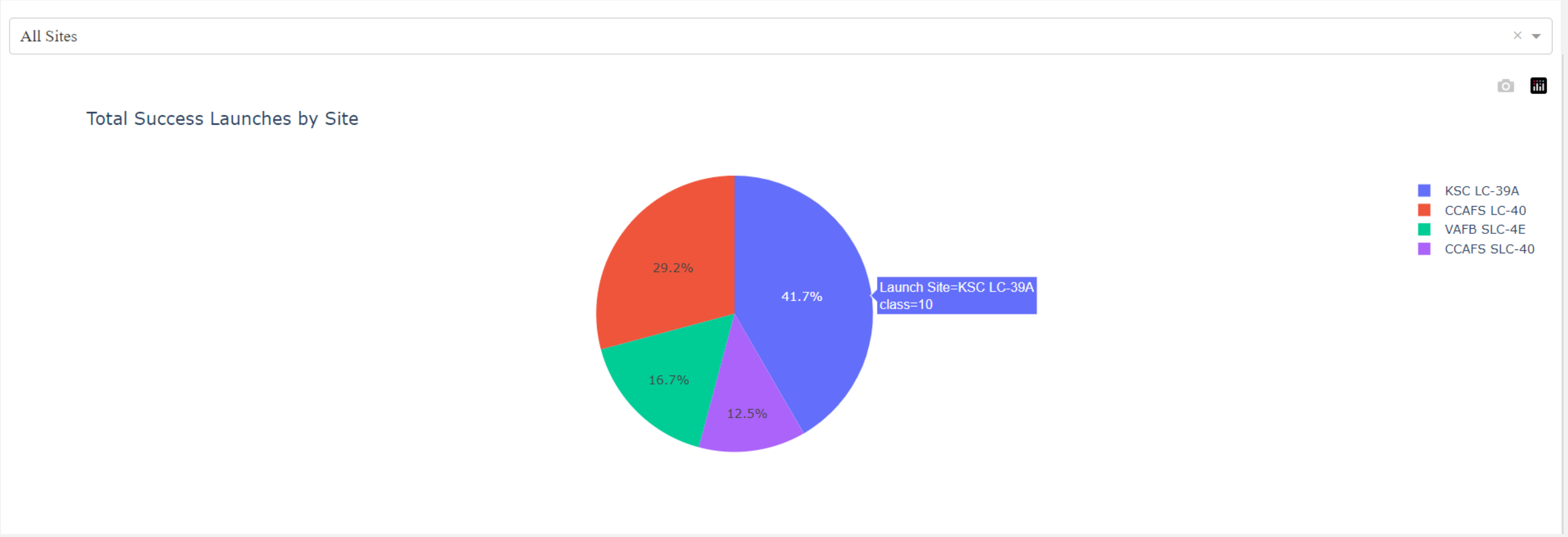
Places	Distance (km)
Coastline	0.86
Active railway	1.22
Active highway	0.59
City (Cape Canaveral)	17.64

- Launch sites are located close to railways and highways for transportation and logistical purposes. Being close to transportation infrastructure allows for quick evacuation procedures and emergency response capabilities if a rocket launch incident happens.
- Launch sites are located far away from cities to ensure public safety and reduce possible risks to populated areas in the case of launch-related incidents or failures.

Section 4

Build a Dashboard with Plotly Dash

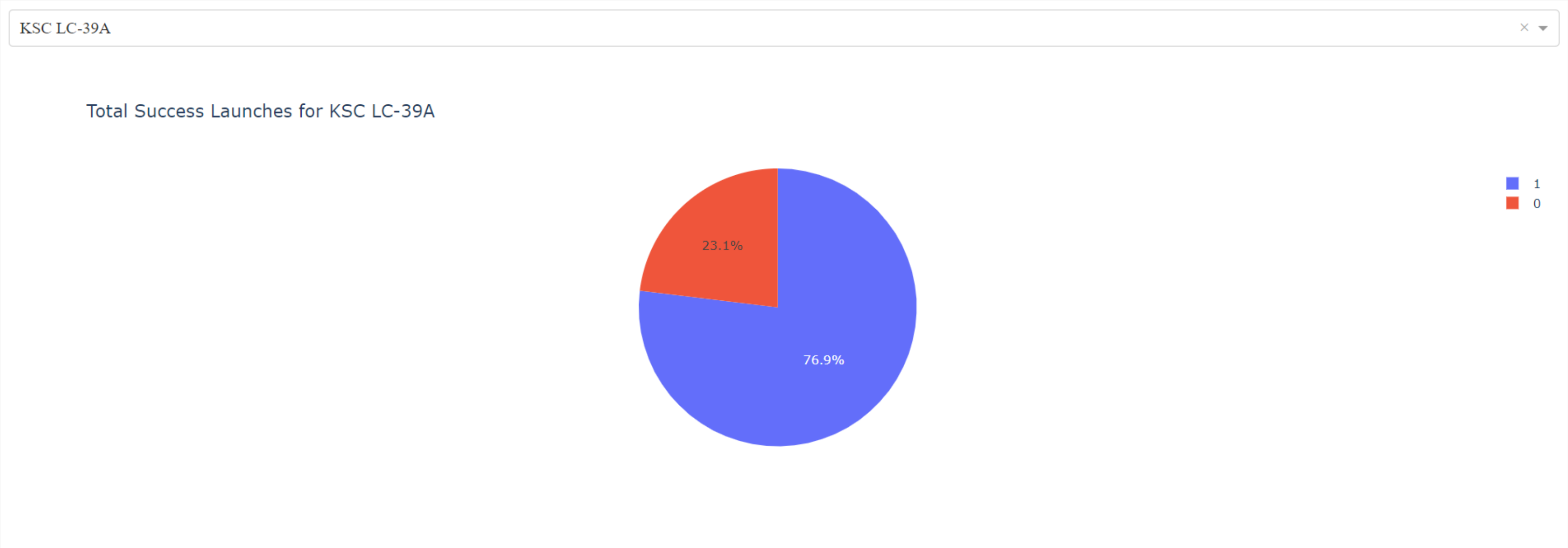
Largest Successful Launches by Launch Site



Insight:

With10 successful launch (41.7%), KSC LC-39A becomes the largest successful launches from all launch sites.

Launch Site with Highest Success Rate



Insight:

KSC LC-39A has the highest launch success rate, with a success rate of 76.9%.

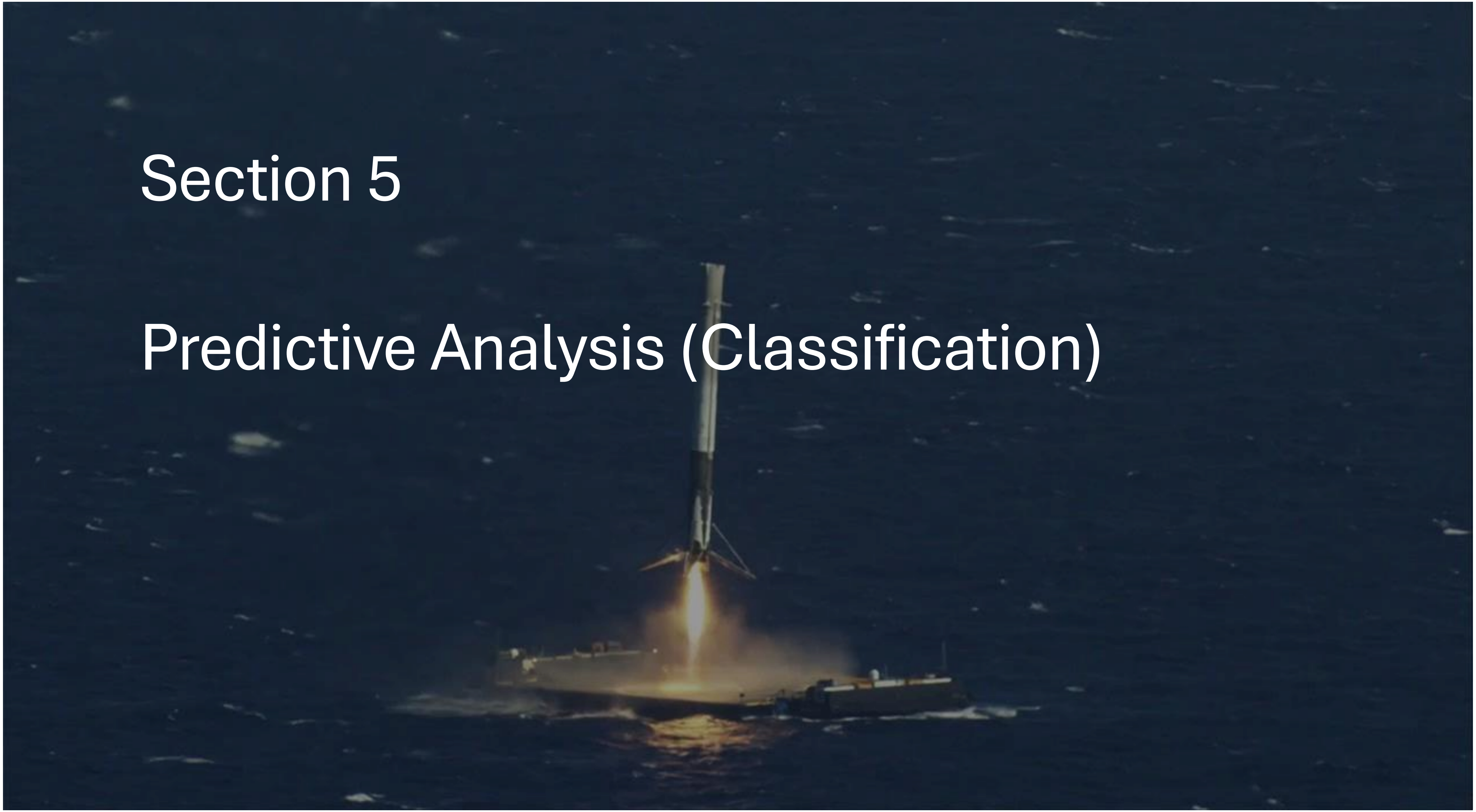
Payload Mass vs. Launch Outcome

Insight:

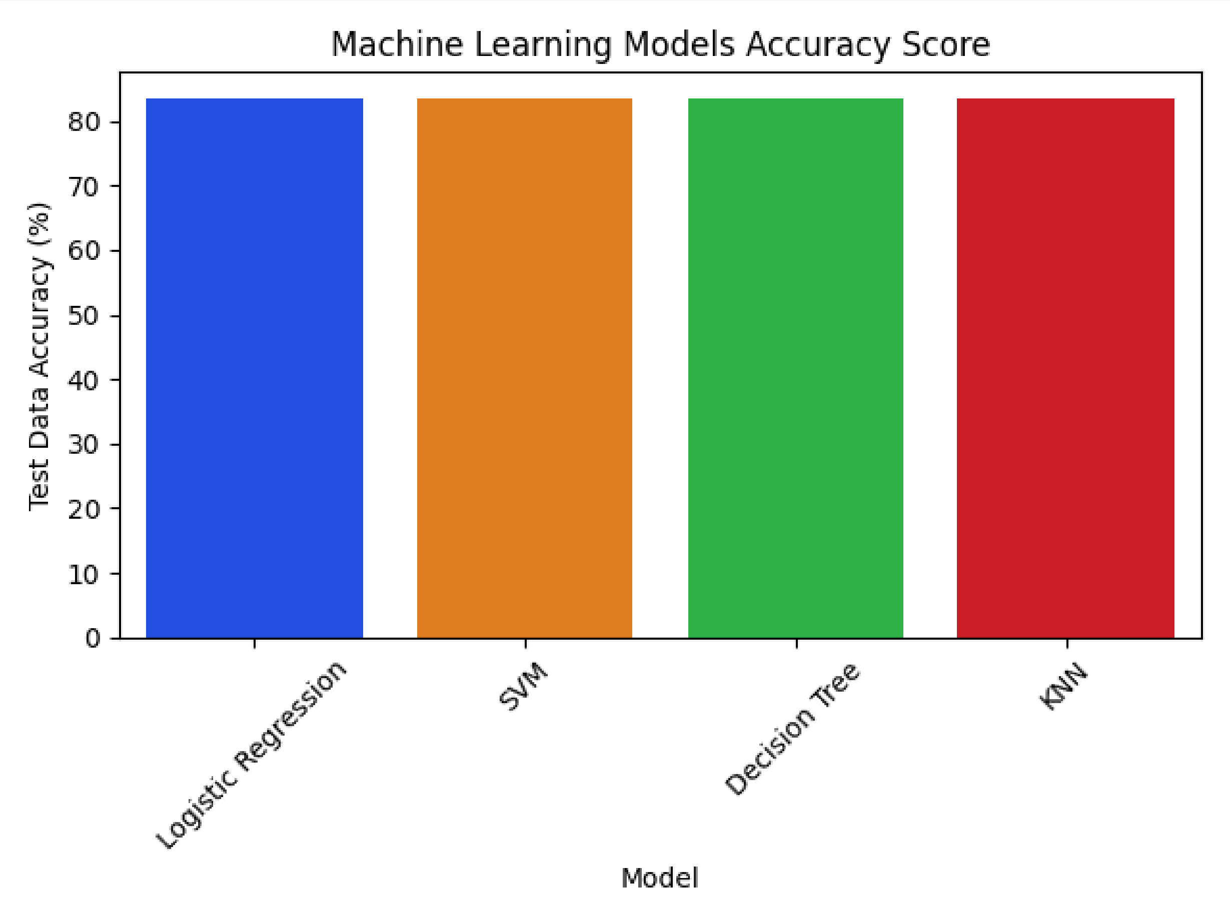
- Payloads with mass range 2,500 – 10,000 kg has the highest launch success rate with 48.57%
- Payload with range 5,000 – 7,500 kg has the lowest launch success rate with 22.23%
- Booster version B5 has the highest launch success rate 100% from 1 flight, followed by FT with 16 successful launch from 24 flights.

Section 5

Predictive Analysis (Classification)



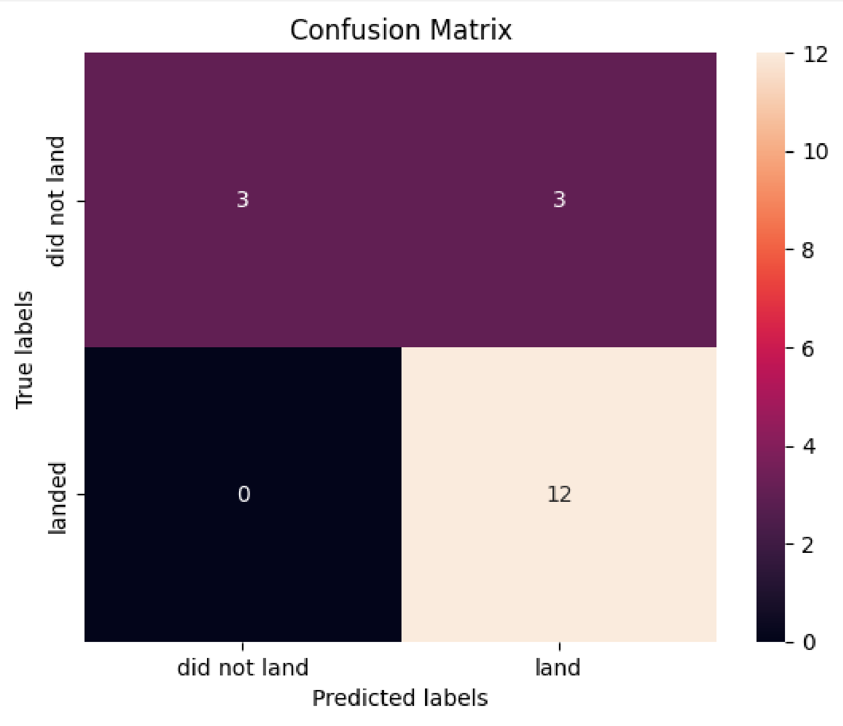
Classification Accuracy

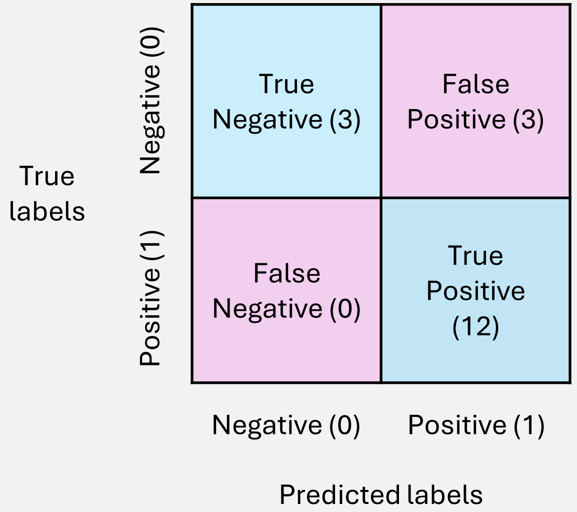


Insight:

All four models achieved the same accuracy of 83.33% for test data, indicating there is no significant difference in performance between them.

Confusion Matrix





Insights:

- All models have same confusion matrix result.
- The main problem from this classification model is false positives (predicted value by model labeled as positive, but true value is negative).
- The performance metrics (precision, recall, F1 score, etc.) show that they have performed equally well across all aspects of classification evaluation.

Conclusion

The conclusions of this report are:

- The success rate of Falcon 9 launch outcome grows significantly each year. Make it more reliable and trusted for the next few periods.
- Several variables such as orbit type and payload mass greatly influence the launch outcome. Other variables need to be studied further.
- All classification models have accuracy test data and confusion matrix with the same results. An evaluation model such as feature selection or cross-validation is needed.

Appendix

- Dataset URL:
  - dataset_part_1.csv
  - spacex_web_scraped.csv
  - dataset_part_2.csv
  - dataset_part_3.csv
  - Spacex.csv
  - spacex_launch_geo.csv
  - spacex_launch_dash.csv
- Images credit:
  - Space Exploration Technologies (SpaceX)
  - National Aeronautics and Space Administration (NASA)
  - European Space Agency (ESA)


Thank you!