IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hrushikesh Sanap
19.05.2025

Outline

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

Executive Summary

Summary of methodologies

- Data collected by Web Scraping and SpaceX API
- Exploratory Data Analysis, Data Wrangling, Data Visualization and Machine Learning Predictive Analysis were done on data.

Summary of all results

- Valuable data was successfully gathered from public sources.
- Exploratory Data Analysis (EDA) helped identify the most relevant features for predicting launch success.
- Machine learning predictions revealed the most effective model for determining which characteristics best drive this opportunity, utilizing all the collected data.

Introduction

The objective is to assess the feasibility of the new company, Space Y, competing with SpaceX.

Key questions to address include:

- What is the most effective way to estimate the total cost of launches by predicting successful first-stage landings?
- Which launch sites are the most optimal for conducting launches?

Section 1

Methodology

Methodology

Data Collection Methodology:

Data related to SpaceX was gathered from two primary sources:

SpaceX API: https://api.spacexdata.com/v4/rockets/

Web scraping from Wikipedia: List of Falcon 9 and Falcon Heavy launches

Data Wrangling:

The collected data was cleaned and enhanced by generating a landing outcome label, derived from outcome data after summarizing and analyzing key features.

Exploratory Data Analysis (EDA):

EDA was performed using visualizations and SQL to uncover patterns and insights from the data.

Methodology

Interactive Visual Analytics:

Tools such as Folium and Plotly Dash were used to create interactive visualizations for deeper exploration of geographic and performance data.

Predictive Analysis:

Classification models were employed to perform predictive analysis.

The processed data was normalized and split into training and testing datasets.

Four different classification models were evaluated, with their accuracy assessed based on various parameter combinations.

Data Collection

Data related to SpaceX was gathered from two primary sources:

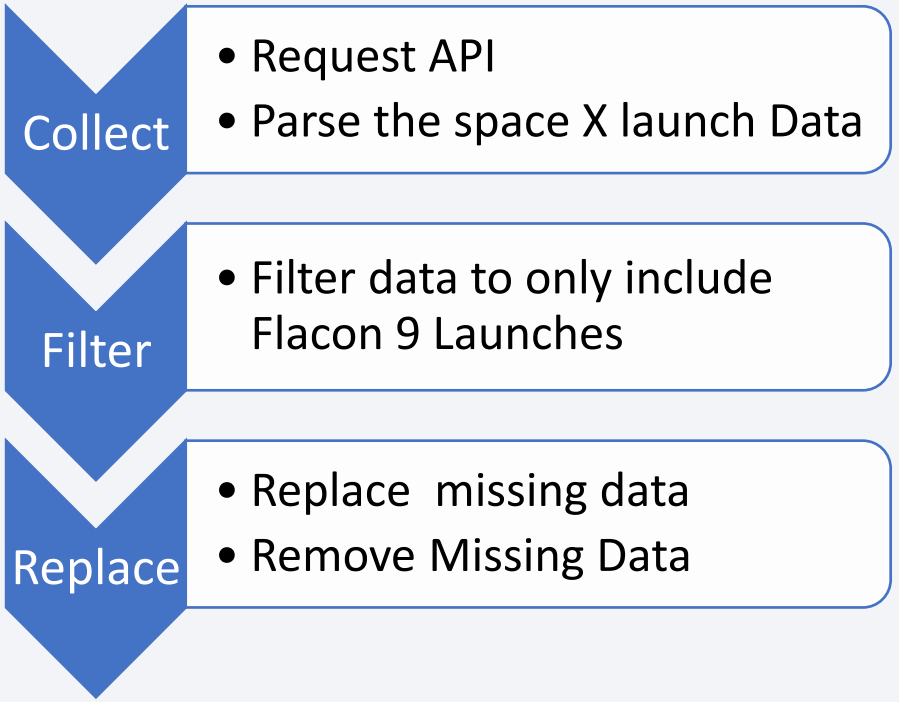
SpaceX API: https://api.spacexdata.com/v4/rockets/

Web scraping from Wikipedia: List of Falcon 9 and Falcon Heavy launches

Check following slides for details:

Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used
- This API was used according to the flowchart beside and then data is persisted.



- Notebook:

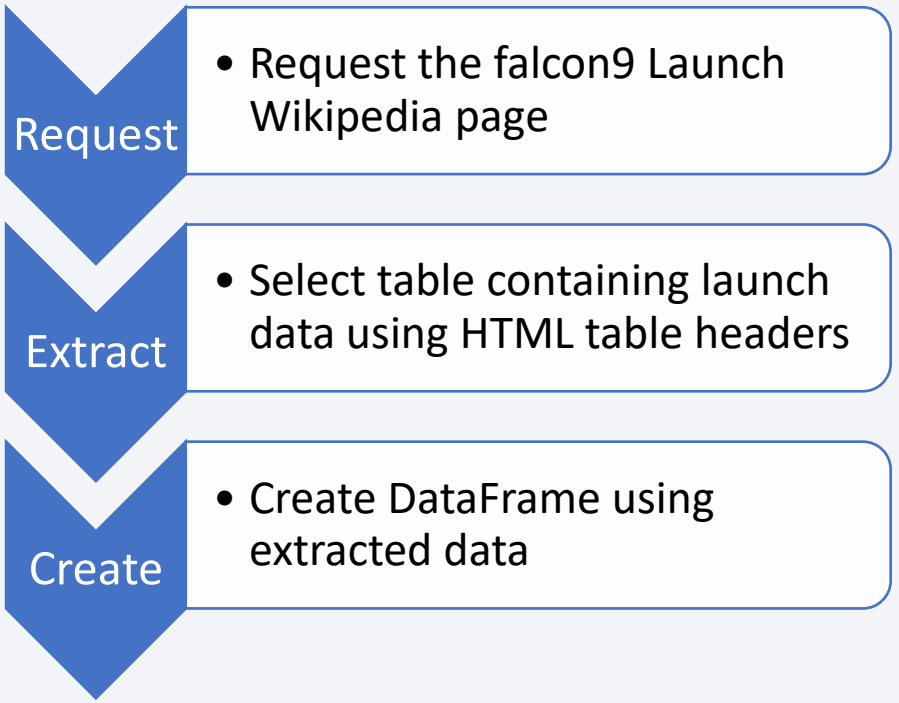
https://github.com/HrushiSanap/ds_specalization_capstone_falcon9_landing/blob/main/c10m1nb1_data_collection_api.ipynb

Data Collection – Web Scraping

- Data from Space X launches can also be obtained from Wikipedia
- Data are downloaded from Wikipedia according to the flowchart and then persisted.

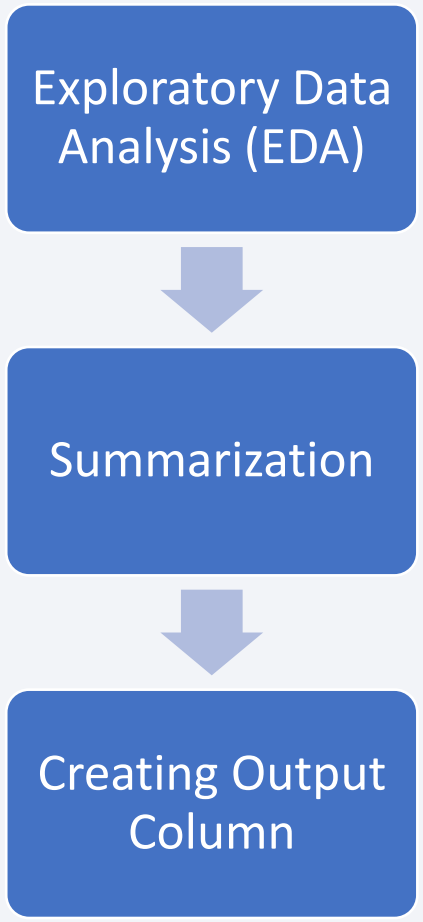
- Notebook:

https://github.com/HrushiSanap/ds_specalization_capstone_falcon9_landing/blob/main/c10m1nb2_webscraping.ipynb



Data Wrangling

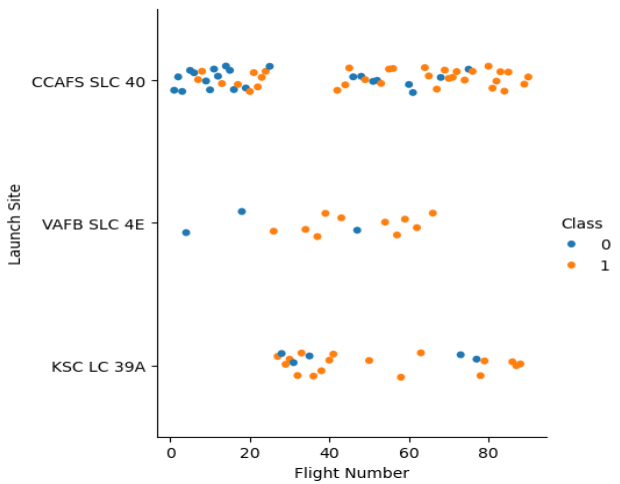
- Initial Exploratory Data Analysis (EDA) was conducted on the dataset.
- Summaries were generated for launches by site, frequency of each orbit type, and mission outcomes by orbit type.
- Finally, a landing outcome label was created based on the values in the Outcome column.

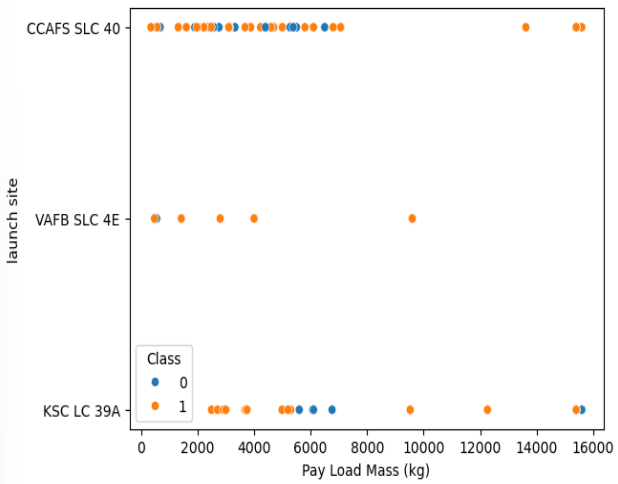


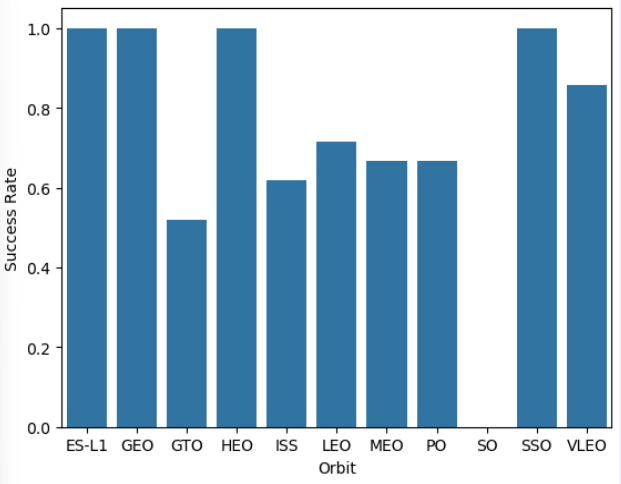
Notebook:
https://github.com/HrushiSanap/ds_specalization_capstone_falcon9_landing/blob/main/c10m1nb3_data_wrangling.ipynb

EDA with Data Visualization

- Scatter plot, bar plots, line plots, were generated to visualize relationship between various features.
- At the end feature engineering such as creating dummy variables, one-hot-encoding was done to create final data set.
- All plots and graphs are included in notebook mentioned below.







Notebook:
https://github.com/HrushiSanap/ds_specalization_capstone_falcon9_landing/blob/main/c10m2nb2_eda_data_visaualization.ipynb

EDA with SQL

Following SQL queries were performed:

- 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 succesful landing outcome in ground pad was acheived.
- 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 all the booster_versions that have carried the maximum payload mass. Use a subquery.
- 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.

Notebook:
https://github.com/HrushiSanap/ds_specalization_capstone_falcon9_landing/blob/main/c10m2nb1_eda_sql.ipynb

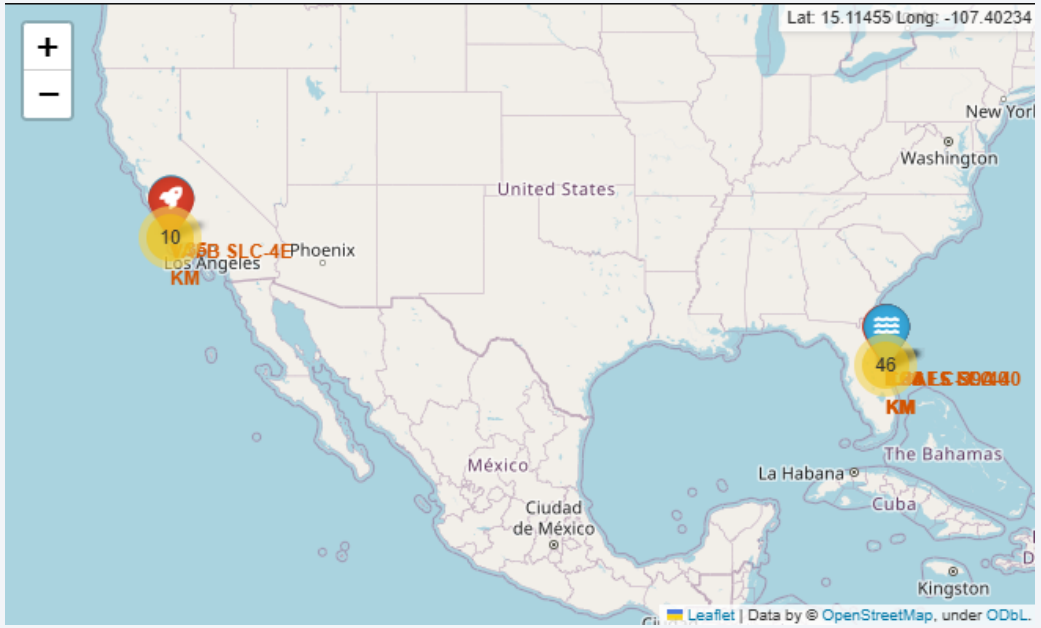
EDA: Dashboard with Plotly Dash

- The following graphs and plots were used to visualize the data:
    - **Percentage of launches by site**
    - **Payload range distribution**
- This combination enabled a quick analysis of the relationship between payload sizes and launch sites, helping to identify the most suitable locations for launches based on payload capacity.

Notebook:
https://github.com/HrushiSanap/ds_specalization_capstone_falcon9_landing/blob/main/c10m3nb2_plotly_dash.py

EDA: Interactive Map with Folium

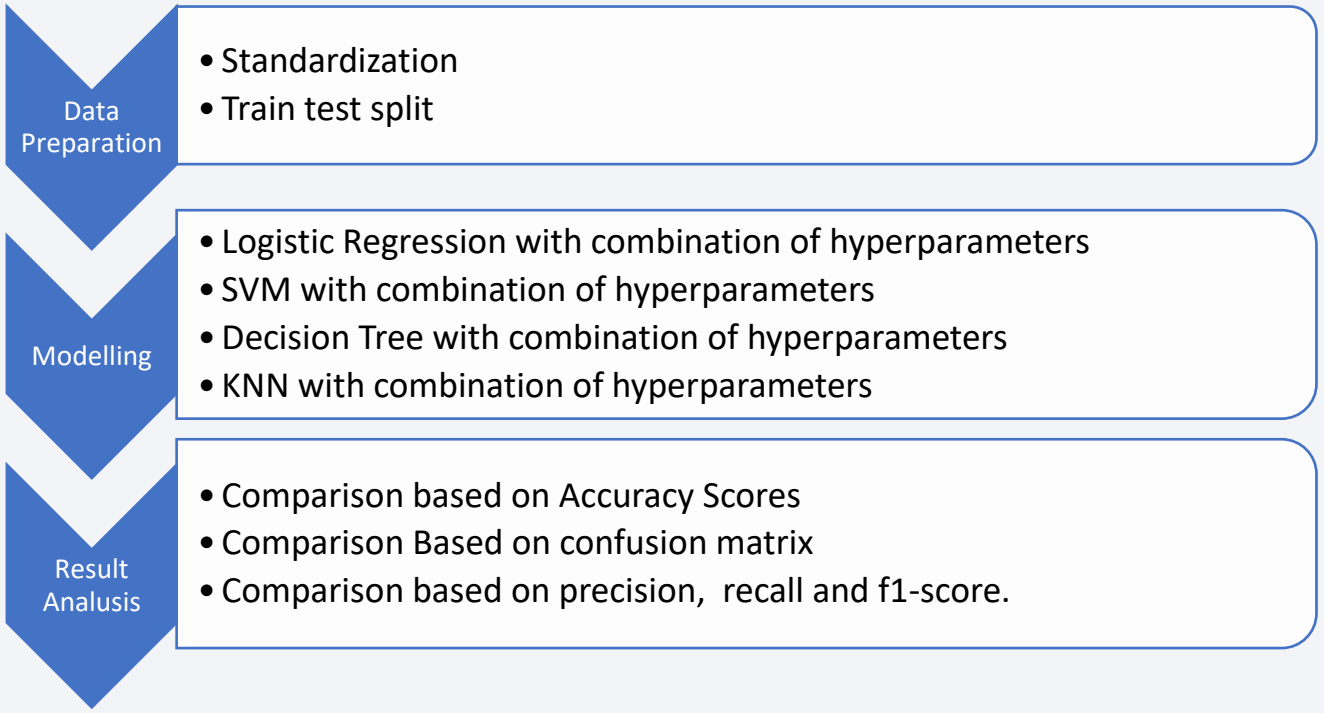
- Folium Maps were utilized with various elements, including markers, circles, lines, and marker clusters.
- Markers were used to represent specific locations, such as launch sites.
- Circles highlighted areas around particular coordinates, for example, the NASA Johnson Space Centre.
- Marker clusters grouped multiple events occurring at the same location, such as multiple launches from a single site.
- Lines were drawn to represent distances between two geographic coordinates.



Notebook:
https://github.com/HrushiSanap/ds_specalization_capstone_falcon9_landing/blob/main/c10m3nb1_maps_geospatial_data.ipynb

Predictive Analysis (Classification)

Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbours.

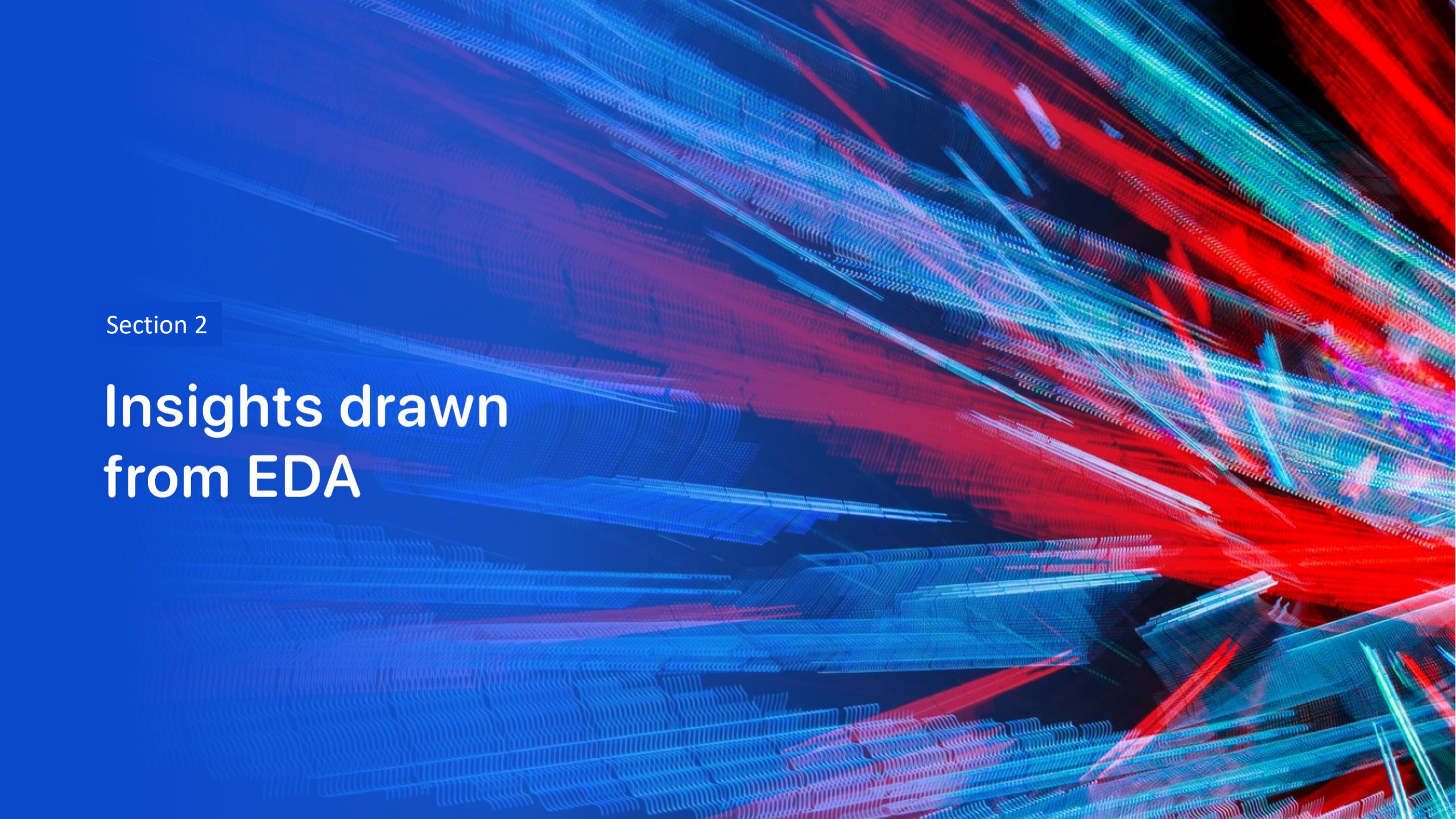


Notebook:
https://github.com/HrushiSanap/ds_specalization_capstone_falcon9_landing/blob/main/c10m4nb1_machine_learning_prediction.ipynb

Results

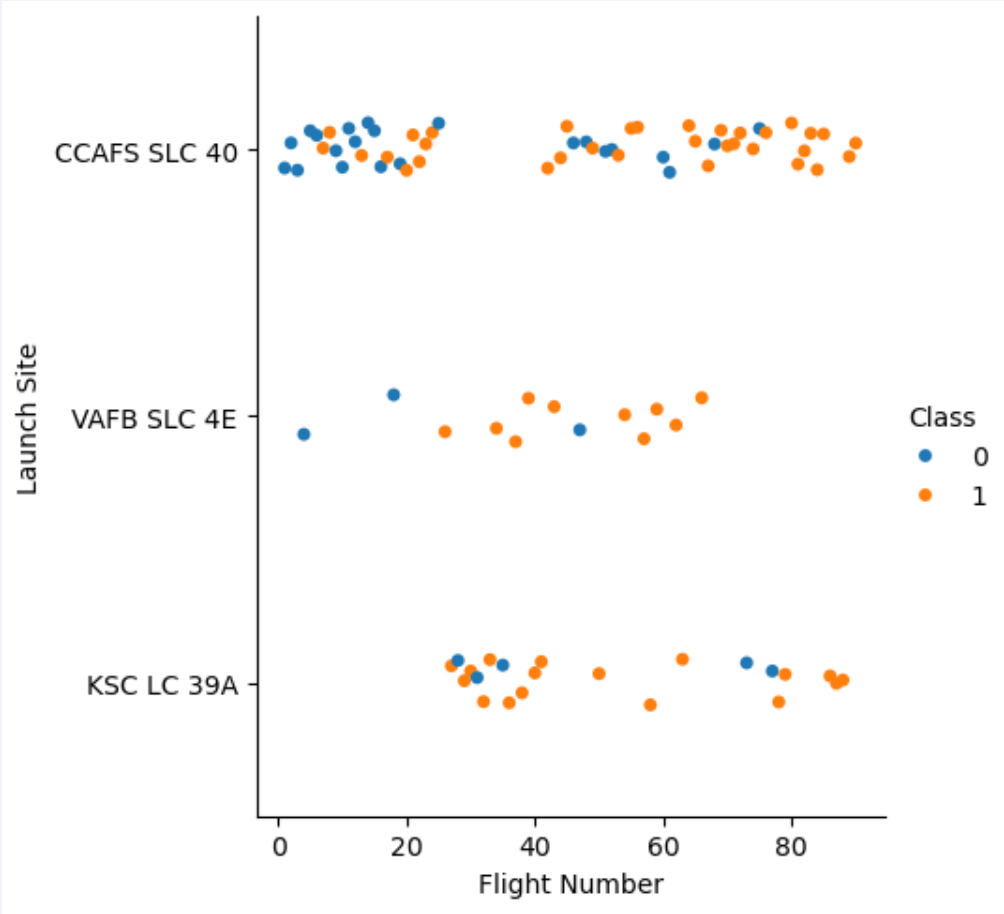
Exploratory Data Analysis Results:

- SpaceX has operated from four distinct launch sites.
- Initial launches were primarily conducted for SpaceX itself and NASA.
- The average payload carried by the Falcon 9 v1.1 booster is approximately 2,928 kg.
- The first successful landing occurred in 2015, five years after the initial launch.
- Several Falcon 9 booster versions successfully landed on drone ships while carrying payloads above the average.
- Nearly all mission outcomes have been successful.
- In 2015, two booster versions—F9 v1.1 B1012 and F9 v1.1 B1015—failed to land on drone ships.
- The frequency of successful landings has improved significantly over the years.
- It was noticed that most launches were carried out at east coast.
- Any of Logistic Regression, SVM, or KNN, with a slight preference for Logistic Regression due to its simplicity and interpretability.


Section 2
Insights drawn from EDA

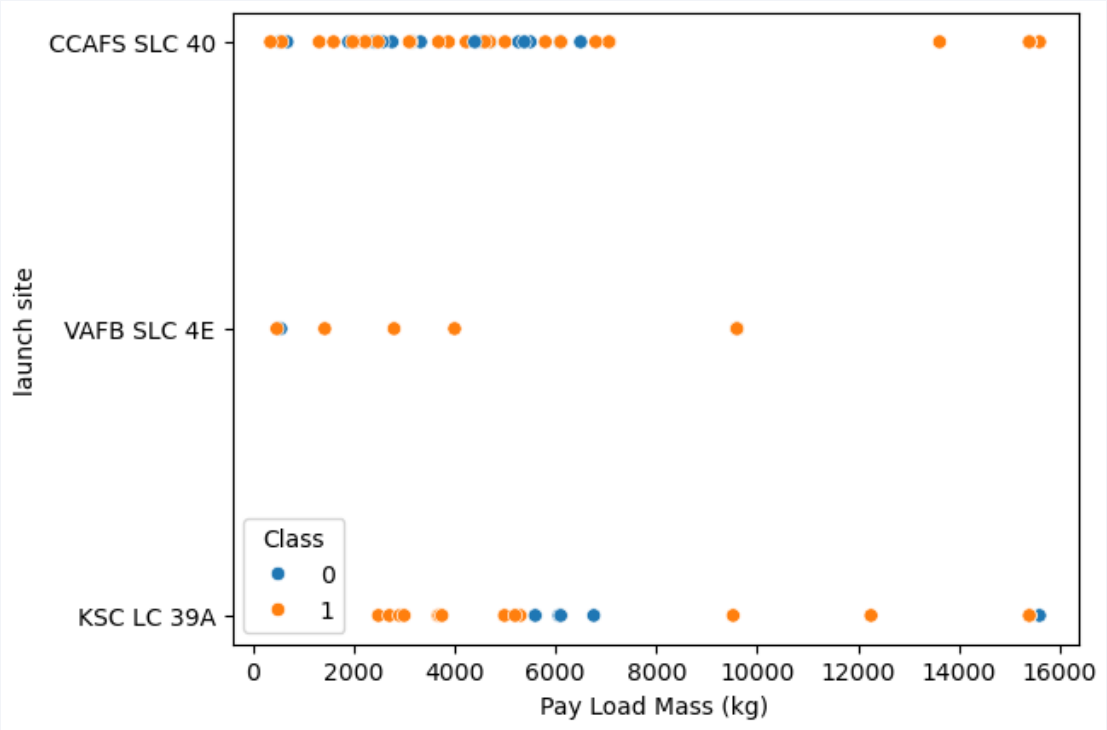
Flight Number vs. Launch Site

- Based on the analysis, the most reliable launch site currently is CCAFS SLC-40, which has recorded the highest number of successful recent launches.
- VAFB SLC-4E ranks second, followed by KSC LC-39A in third place.
- The data also indicates a steady improvement in the overall success rate of launches over time.



Payload vs. Launch Site

- Payloads over 9,000kg have excellent success rate
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites



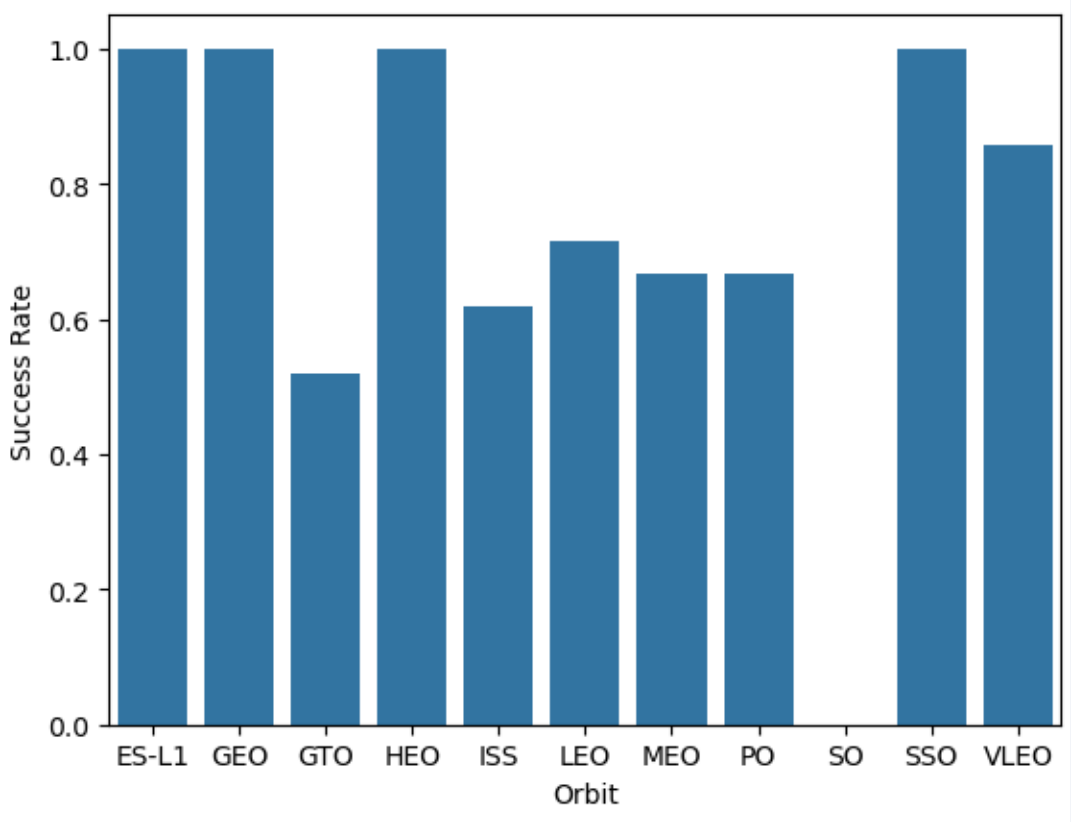
Success Rate vs. Orbit Type

The highest success rates were observed for missions targeting the following orbits:

- ES-L1
- GEO (Geostationary Orbit)
- HEO (Highly Elliptical Orbit)
- SSO (Sun-Synchronous Orbit)

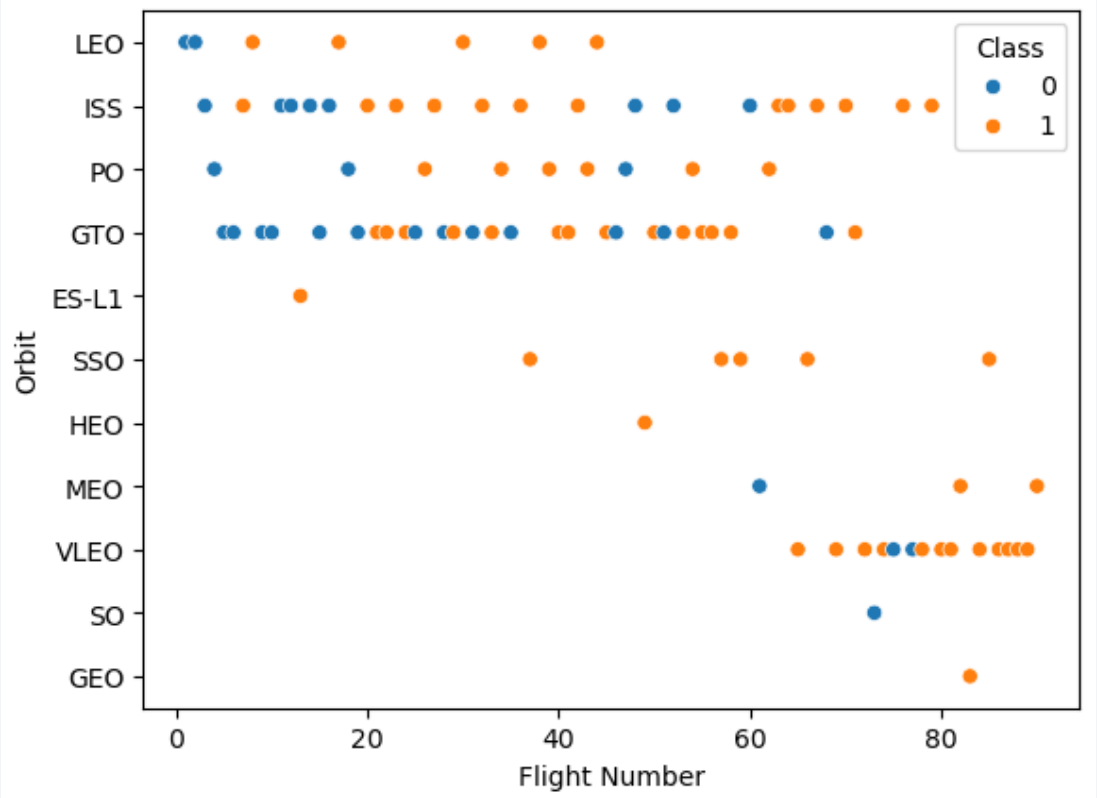
These were followed by:

- VLEO (Very Low Earth Orbit), with a success rate above 80%
- LFO (Low Earth Orbit), with a success rate above 70%



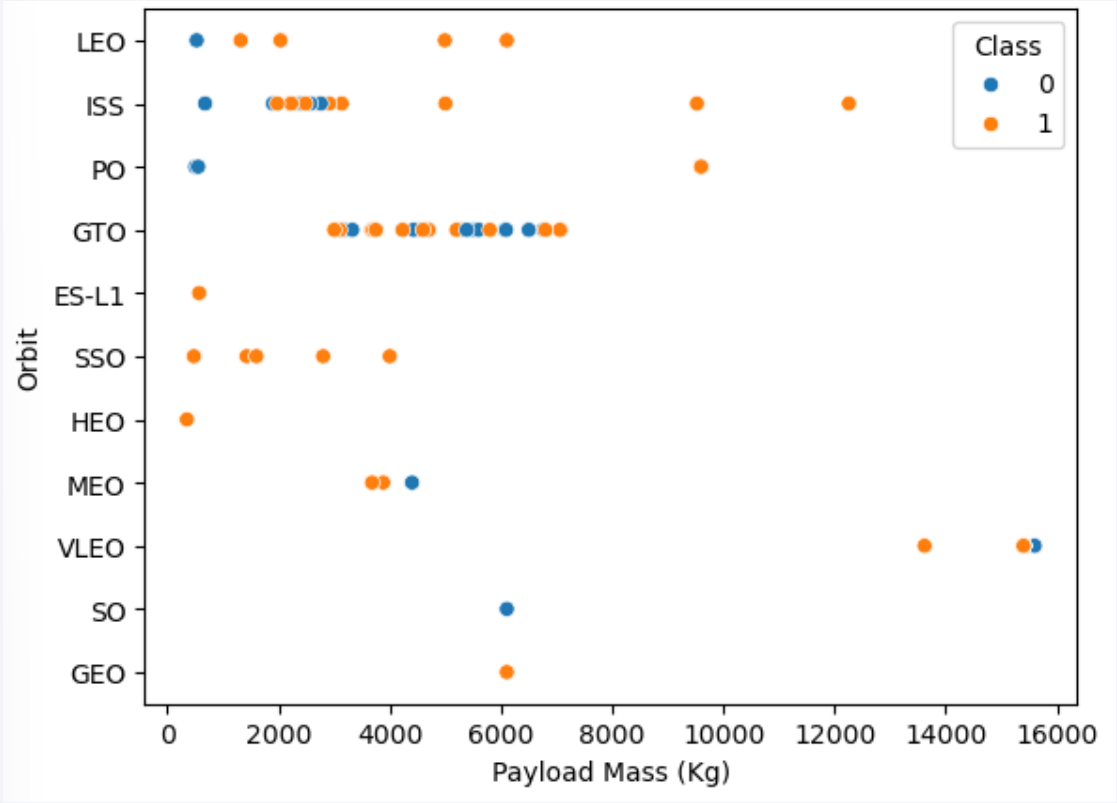
Flight Number vs. Orbit Type

- Success rate improved over time for all orbits
- VLEO orbit looks like a new business opportunity, because of recent increase in its frequency.



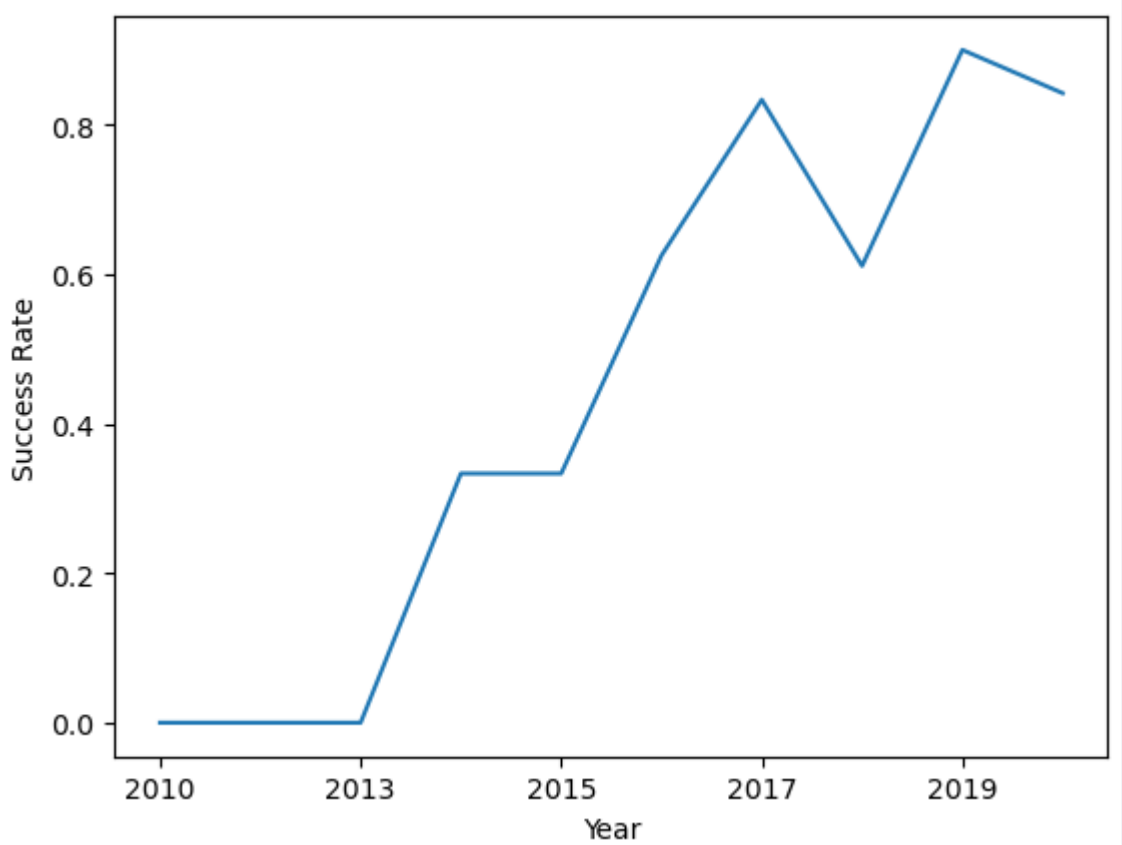
Payload vs. Orbit Type

- There appears to be no clear correlation between payload size and success rate for missions targeting the GTO (Geostationary Transfer Orbit).
- The ISS (International Space Station) orbit exhibits the widest payload range and maintains a high success rate.
- SO (Solar Orbit) and GEO (Geostationary Orbit) have had relatively few launches, making trend analysis limited for these orbits.



Launch Success Yearly Trend

- Success rate increased after 2013.
- Initial years were hard as there were not enough successful launches.



All Launch Site Names

- According to data, there are four launch sites.
- They are obtained by selecting unique occurrences of “launch_site” values from the dataset.

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

Launch Site Names Begin with 'CCA'

- Here we can see five samples of Cape Canaveral launches.

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

- Total payload carried by boosters was 45596 Kg

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:

Average_Payload_Mass
2928.4

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

First Successful Ground Landing Date

- First successful landing outcome on ground pad:

First_Successful_Landing
2015-12-22

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it’s possible to identify the first occurrence, that happened on 12/22/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

- Selecting distinct booster versions according to the filters above, these 4 are the result.

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

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

- Grouping mission outcomes and counting records for each group led us to the summary above.

Boosters Carried Maximum Payload

- These are the boosters which have carried the maximum payload mass registered in the dataset.

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

- List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Month_Name	Booster_Version	Launch_Site	Landing_Outcome
January	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- The above list has only two rows.

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

- 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

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

- This view of data alerts us that “No attempt” must be taken in account.

Section 3

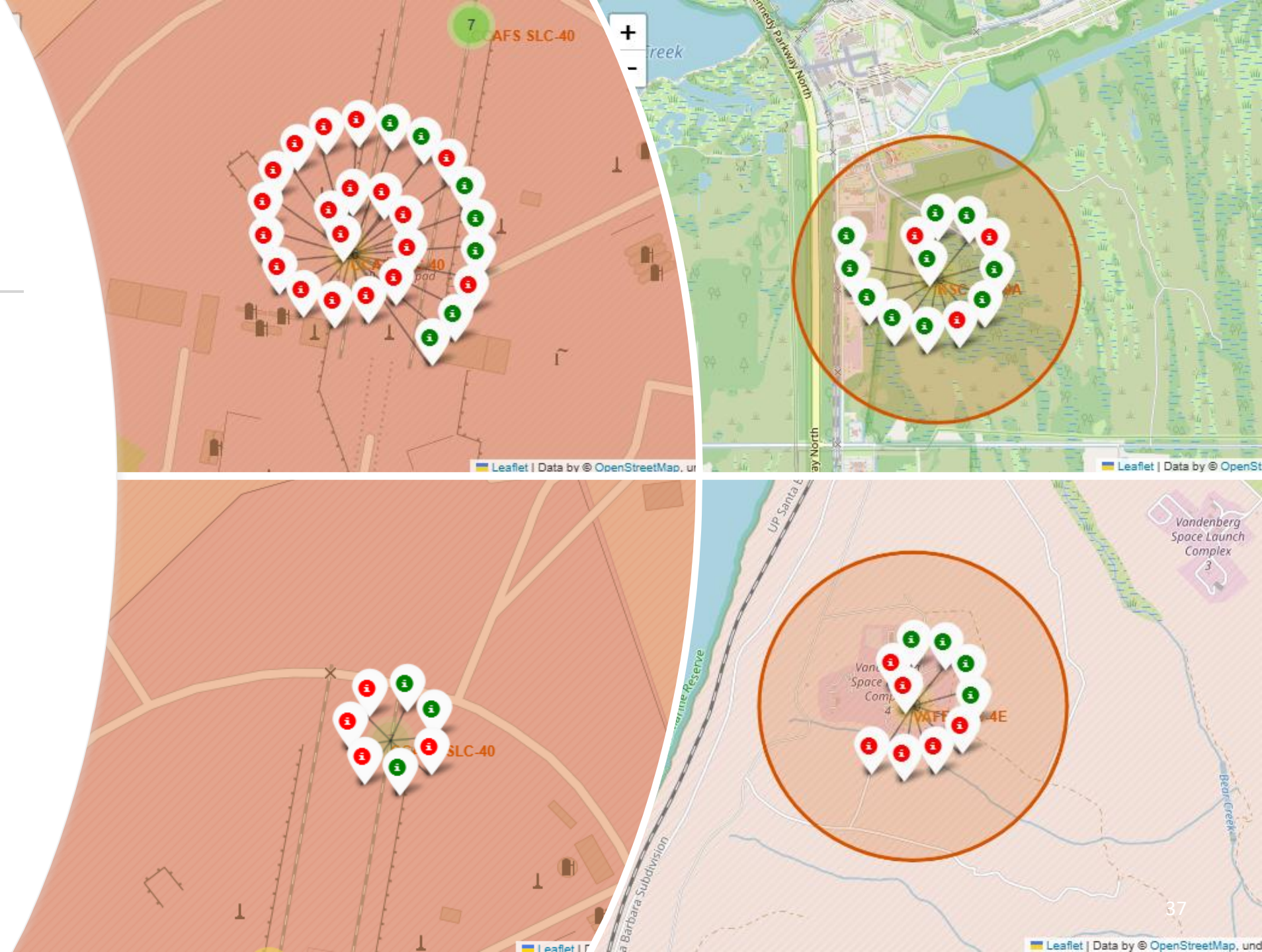
Launch Sites Proximities Analysis

Launch Sites

- Launch sites are near sea, probably by safety, but not too far from roads and railroads.

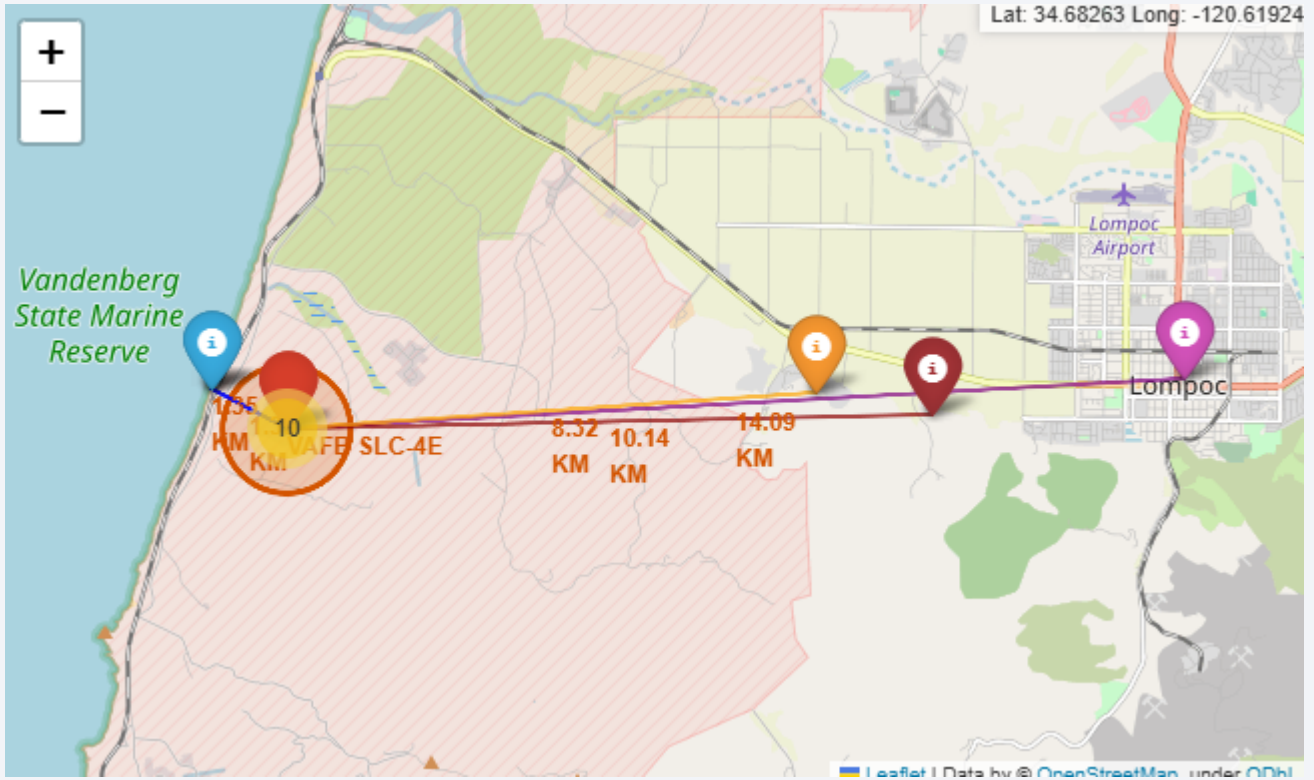
Launch Outcomes for Individual Site

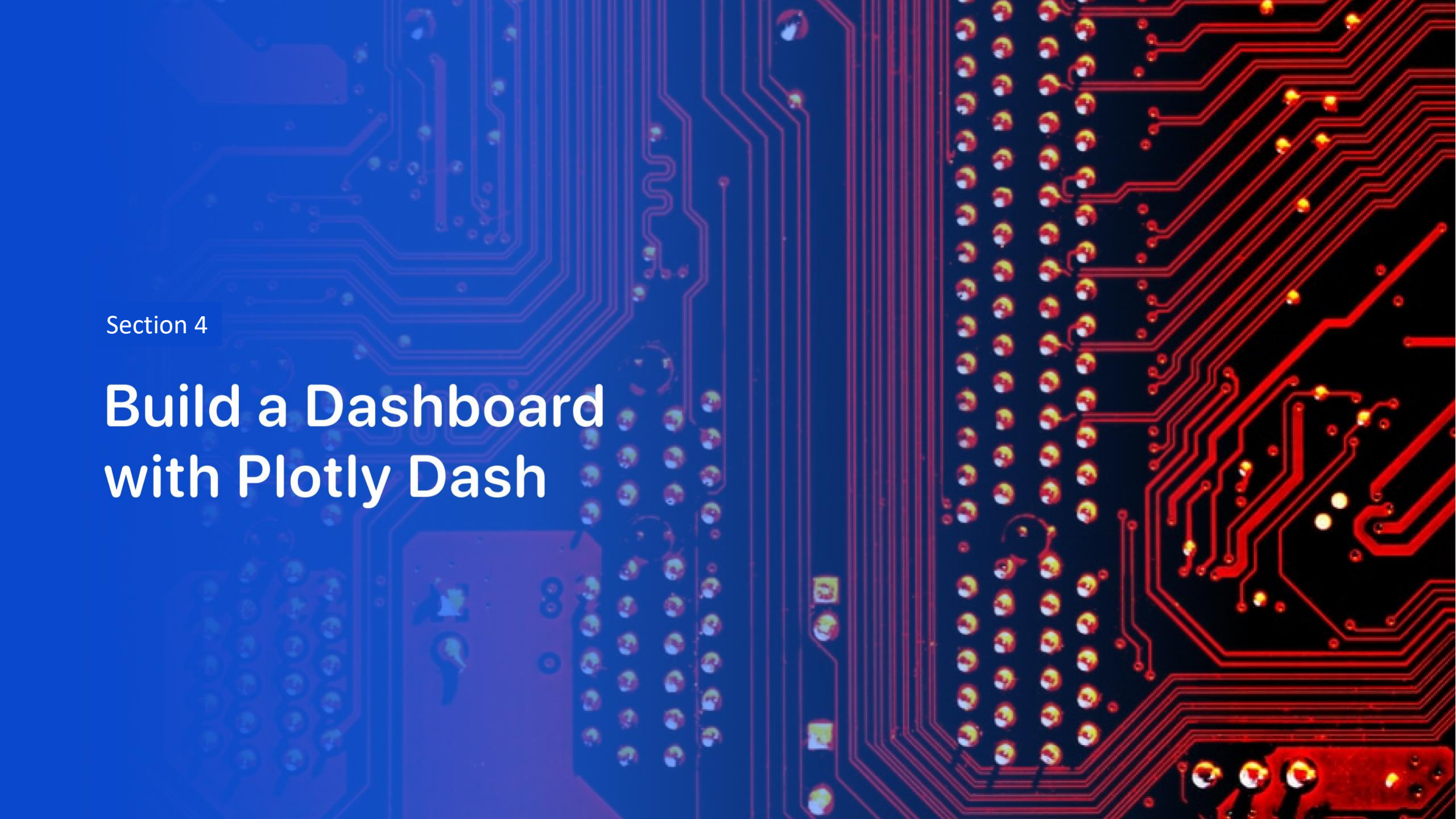
- Green markers indicate successful and red ones indicate failure.



Safety and Logistics

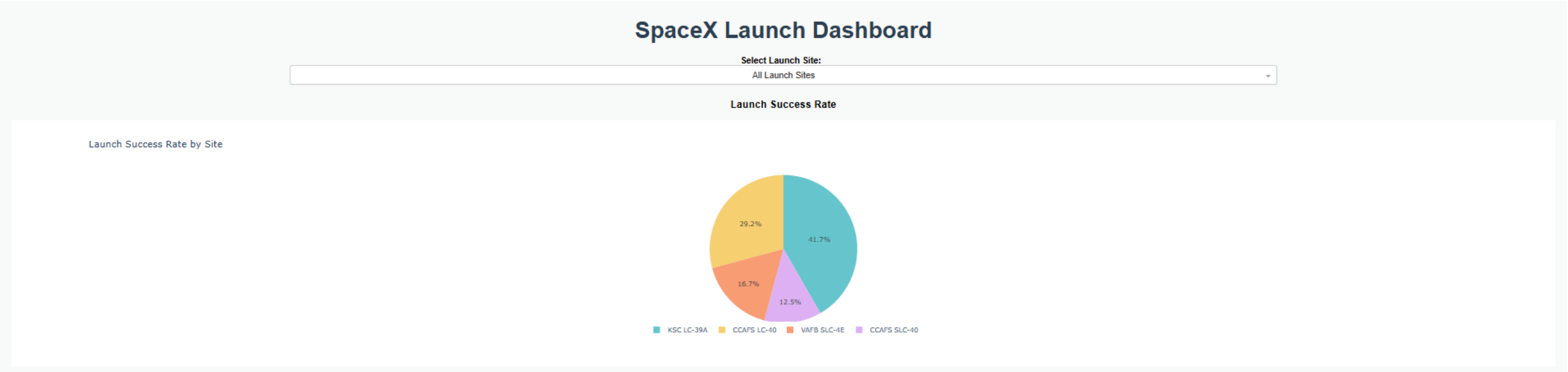
- All Sites are at convenient location when compared on basis of how close they are to major rail roads and inhabited areas. One of the example is as follows.




Section 4
Build a Dashboard with Plotly Dash

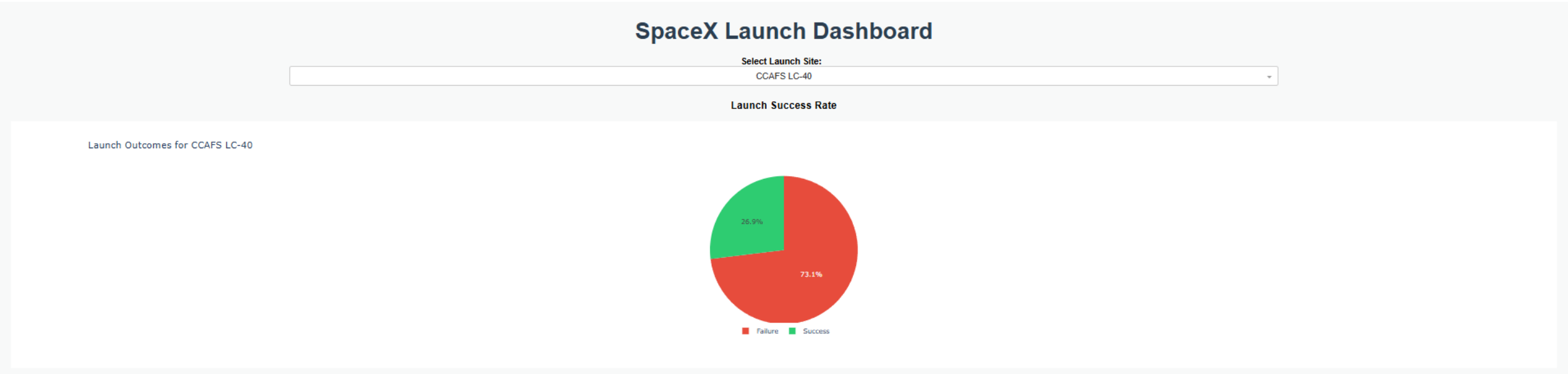
Successful Launches by Site

- The place from where launches are done seems to be a very important factor of success of missions.



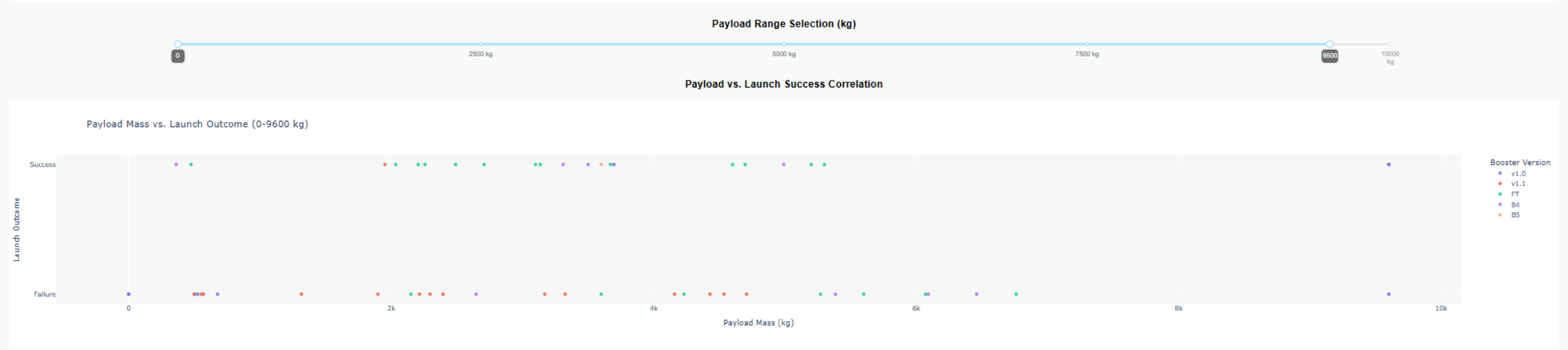
Launch Success Ratio for CCAFS LC-40

- 73.1% of launches are successful in this site.



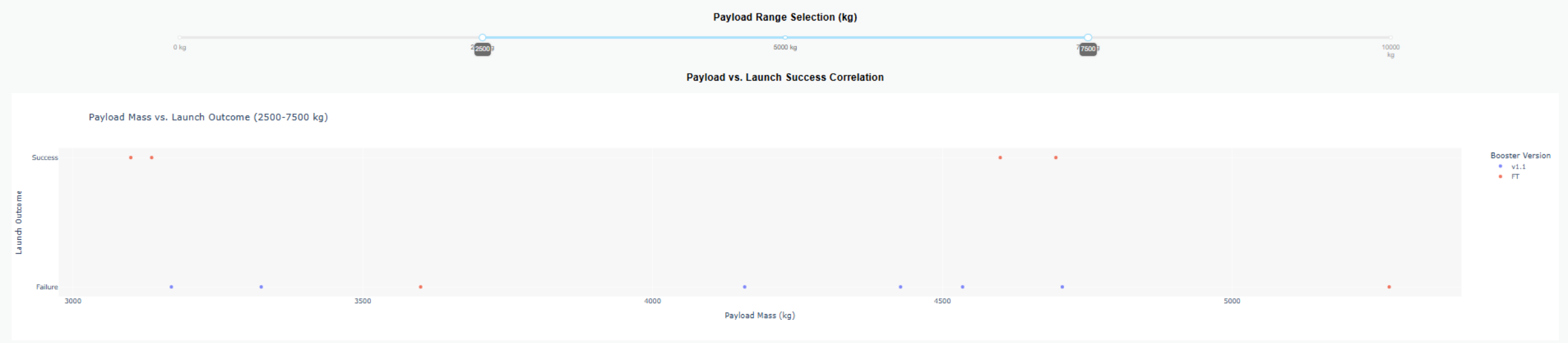
Payload vs. Launch Outcome

- • Payloads under 6,000kg and FT boosters are the most successful combination.



Payload vs. Launch Outcome

- There’s not enough data to estimate risk of launches over 7,000kg

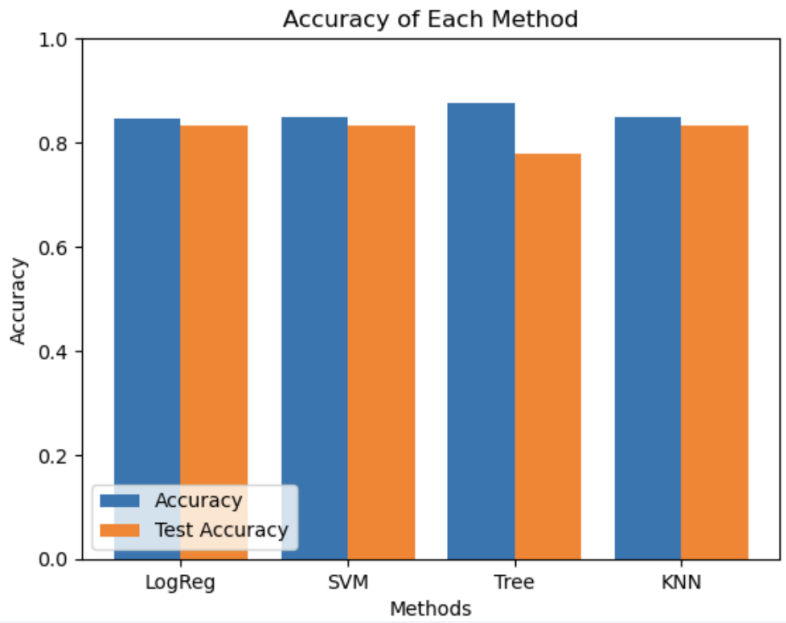


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.



Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.875	0.77778
KNN	0.84821	0.83333

Classification Accuracy

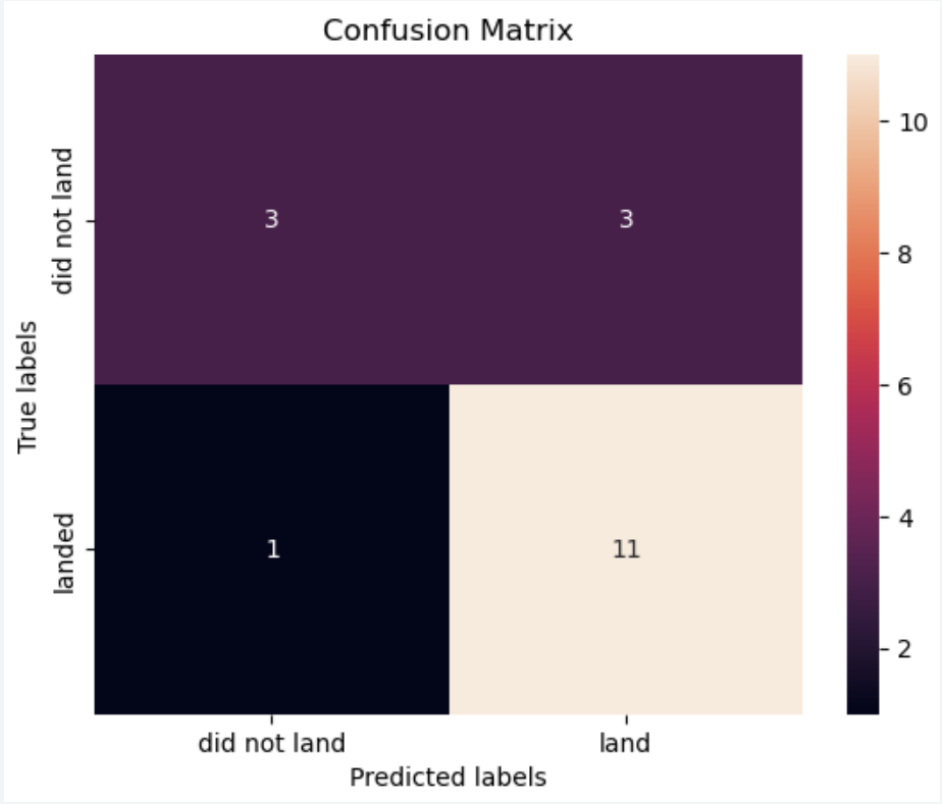
To determine the best algorithm among Logistic Regression, SVM, Decision Tree, and KNN, following things were considered.

- Test Accuracy – generalization performance on unseen data.
- Precision, Recall, and F1-score – especially for the more important class (usually class 1 if it's the positive class).
- Consistency between training and testing scores to avoid overfitting.

Model	Train Acc	Test Acc	Class 1 F1	Overfitting Risk	Notes
Logistic Regression	0.846	0.833	**0.89**	Low	Consistent, well-balanced
SVM	0.848	0.833	**0.89**	Low	Same as Logistic
Decision Tree	**0.860**	0.778	0.85	**High**	Overfitting likely
KNN	0.848	0.833	**0.89**	Low	Same as Logistic and SVM

Confusion Matrix

- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.



Conclusions

- Multiple data sources were analyzed, allowing for continuous refinement of insights throughout the process.
- The analysis identified KSC LC-39A as the most optimal launch site.
- Launches with payloads exceeding 7,000 kg appear to carry lower risk.
- While mission success rates are generally high, landing success rates have shown progressive improvement over time, reflecting advancements in technology and operational processes.
- Since Logistic Regression, SVM, and KNN perform equally well on all key metrics and Decision Tree shows signs of overfitting, the best algorithm is Any of Logistic Regression, SVM, or KNN, with a slight preference for Logistic Regression due to its simplicity and interpretability.

Appendix

- Please refer to the following repository for the source code.

ds_specalization_capstone_falcon9_landing

Thank you!