IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

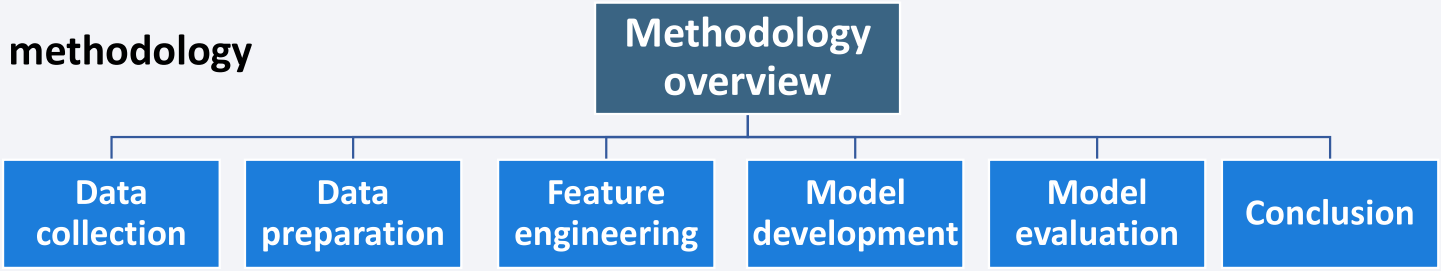
Nguyen Thi Tam Dan
01/04/2025

Outline

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

Executive Summary

Summary of methodology



Methodology overview

Data collection

Data preparation

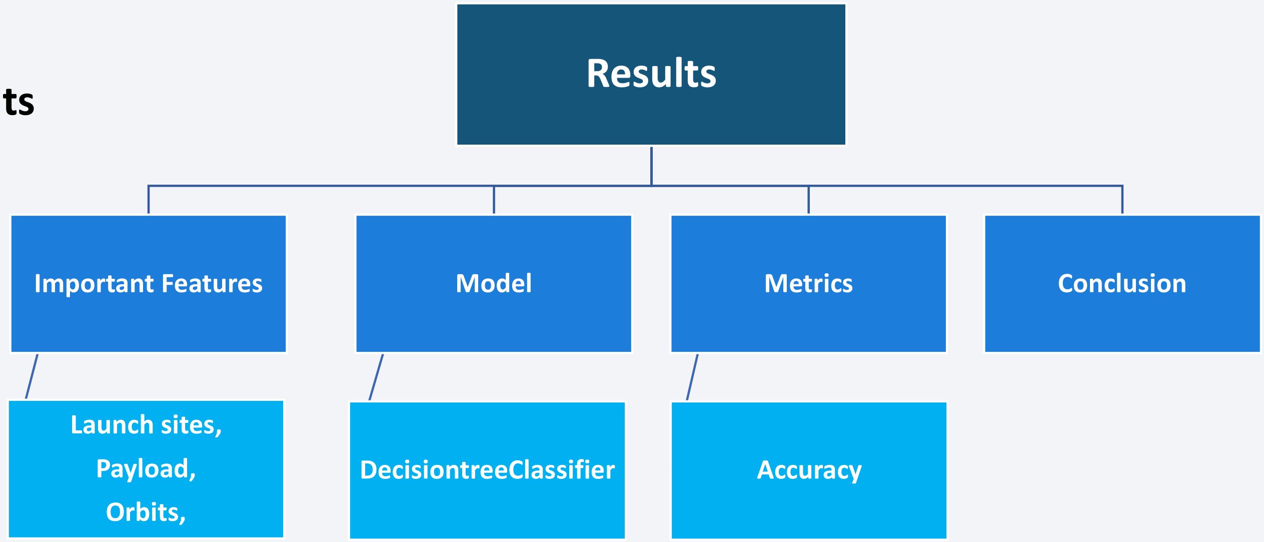
Feature engineering

Model development

Model evaluation

Conclusion

Summary of results



Results

Important Features

Launch sites,
Payload,
Orbits,

Model

DecisiontreeClassifier

Metrics

Accuracy

Conclusion

Introduction

Background

SpaceX is planning to launch rockets into orbit between 2025 and 2030 and is currently assessing potential locations for its launch sites. Selecting an optimal site is crucial for maximizing mission success and operational efficiency.

Objective

By analyzing historical rocket launch data from 2010 to the present, SpaceX seeks to identify the key factors influencing launch success and booster landing reliability. Gaining these insights will enable data-driven decisions to enhance future launch strategies and improve overall mission outcomes.

Section 1

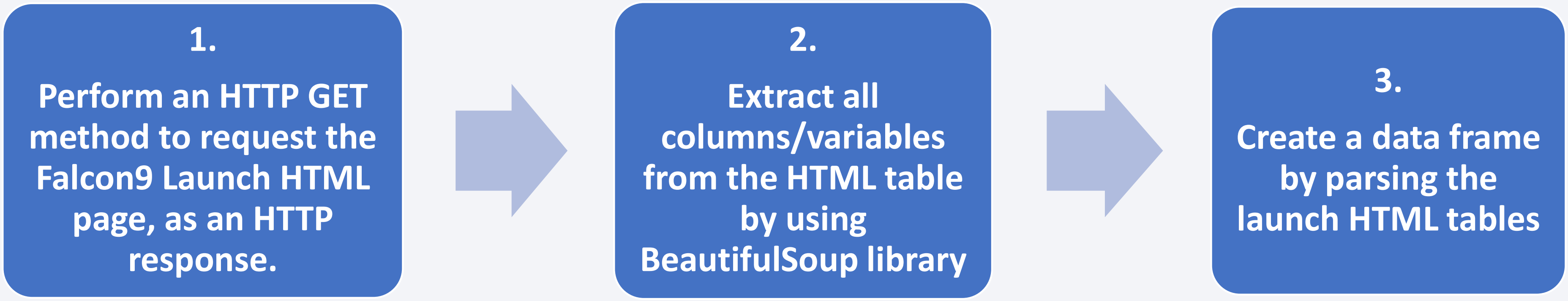
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Scraping data about Falcon 9 historical launch records from Wikipedia by Using Requests and BeautifulSoup libraries
- Perform data wrangling
  - Handled missing values data using Pandas
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built & fine-tuned Logistic Regression, Support Vector Machine, Decision Tree Classifier, and Random Forest models.
  - Evaluation Metrics: Accuracy score

Data Collection



1.

Perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.

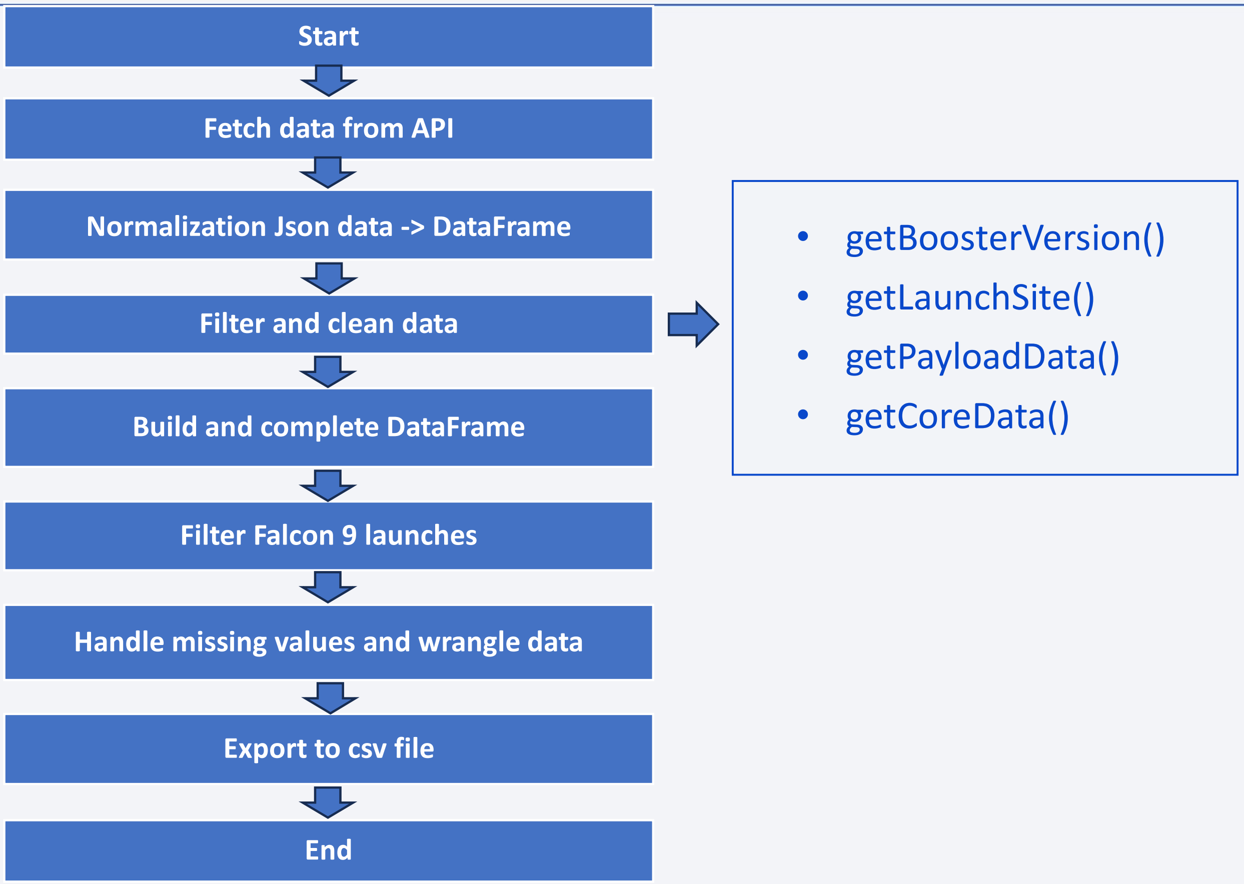
2.

Extract all columns/variables from the HTML table by using BeautifulSoup library

3.

Create a data frame by parsing the launch HTML tables

Data Collection – SpaceX API



Start

Fetch data from API

Normalization Json data -> DataFrame

Filter and clean data

Build and complete DataFrame

Filter Falcon 9 launches

Handle missing values and wrangle data

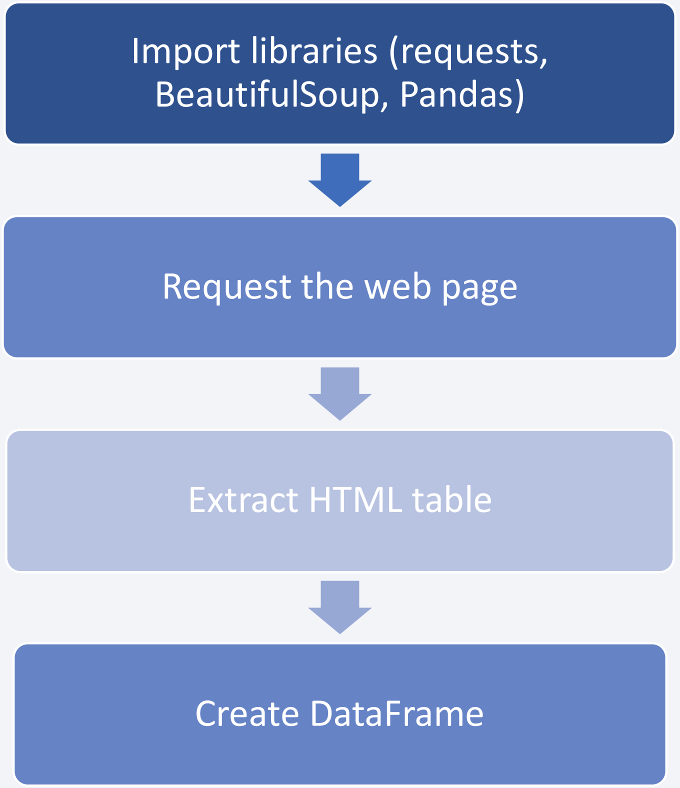
Export to csv file

End

- getBoosterVersion()
- getLaunchSite()
- getPayloadData()
- getCoreData()

Github URL:
https://github.com/Tamdannguyen/TamDan_SpaceX_data_project/blob/main/1.SpaceX%20API%20calls%20notebook.ipynb

Data Collection - Scraping



Import libraries (requests, BeautifulSoup, Pandas)

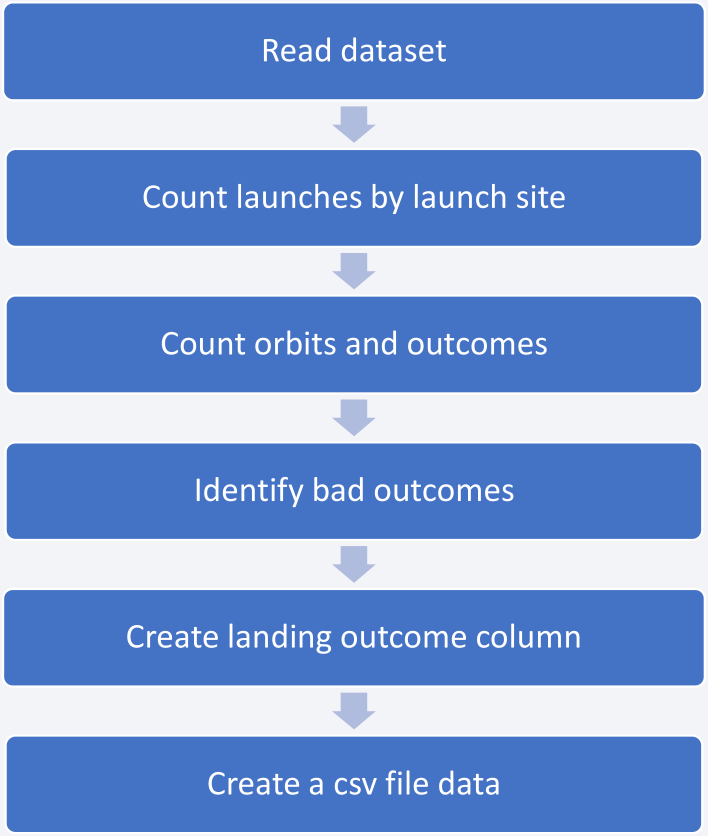
Request the web page

Extract HTML table

Create DataFrame

Github URL:
https://github.com/Tamdannguyen/TamDan_SpaceX_data_project/blob/main/2.webscraping.ipynb

Data Wrangling



Read dataset

Count launches by launch site

Count orbits and outcomes

Identify bad outcomes

Create landing outcome column

Create a csv file data

Github URL:
https://github.com/Tamdannguyen/TamDan_SpaceX_data_project/blob/main/3.Data%20wrangling.ipynb

EDA with Data Visualization

Plot kind	Relationships	Explanation
Scatter plot	FlightNumber vs. PayloadMass	The scatter plot displays the correlation between 2 variables
	FlightNumber vs LaunchSite	
	Launch sites vs Payload mass	
	FlightNumber vs Orbit type	
	Payload Mass vs Orbit type	
Bar plot	Success rate of each orbit type	The bar plot displays the comparison between the launch success rate of each orbit type
Line plot	Launch success yearly trend	The line plot displays the development of the launch success rate through years

EDA with SQL

- Display the names of the unique launch sites in the space mission.
- Display 5 records where launch sites begin with the string 'CCA'.
- Display the total payload mass carried by boosters launched by NASA (CRS).
- Display average payload mass carried by booster version F9 v1.1.
- List the date when the first successful landing outcome in the ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater. than 4000 but less than 6000.
- List the total number of successful and failed mission outcomes.
- List all the booster_versions that have carried the maximum payload mass.
- List the records that will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in the year 2015.
- Rank the count of landing outcomes between the dates 2010-06-04 and 2017-03-20, in descending order.

GitHub URL: *https://github.com/Tamdannguyen/TamDan_SpaceX_data_project/blob/main/4.EDA-SQL-coursera_sqllite.ipynb*

Build an Interactive Map with Folium

The following objects to the folium map:

- **Markers:** To mark the location of Launch sites on the map
- **Circles:** Show the number of launches at each launch site and classify each launch as successful or failed
- **Lines:** To draw the shortest path from the launch site to specific locations such as coastline, highway, train,….

GitHub URL:
https://github.com/Tamdannguyen/TamDan_SpaceX_data_project/blob/main/5.Launch_site_location_%20Folium.ipynb

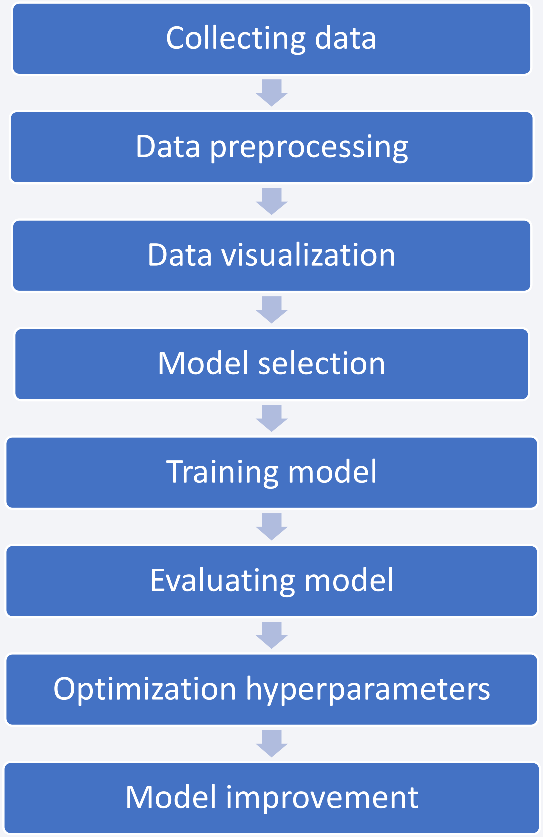
Build a Dashboard with Plotly Dash

The following plots to the Plotly dashboard:

- **Scatter plot:** displays the correlation between two variables Launch sites and Payload mass, detailed in each Booster Version.
- **Pie plot:** Show the success rate for each launch site in total launch sites.

GitHub URL:
https://github.com/Tamdannguyen/TamDan_SpaceX_data_project/blob/main/6.Plotly%20Dashboard.py

Predictive Analysis (Classification)



Collecting data

Data preprocessing

Data visualization

Model selection

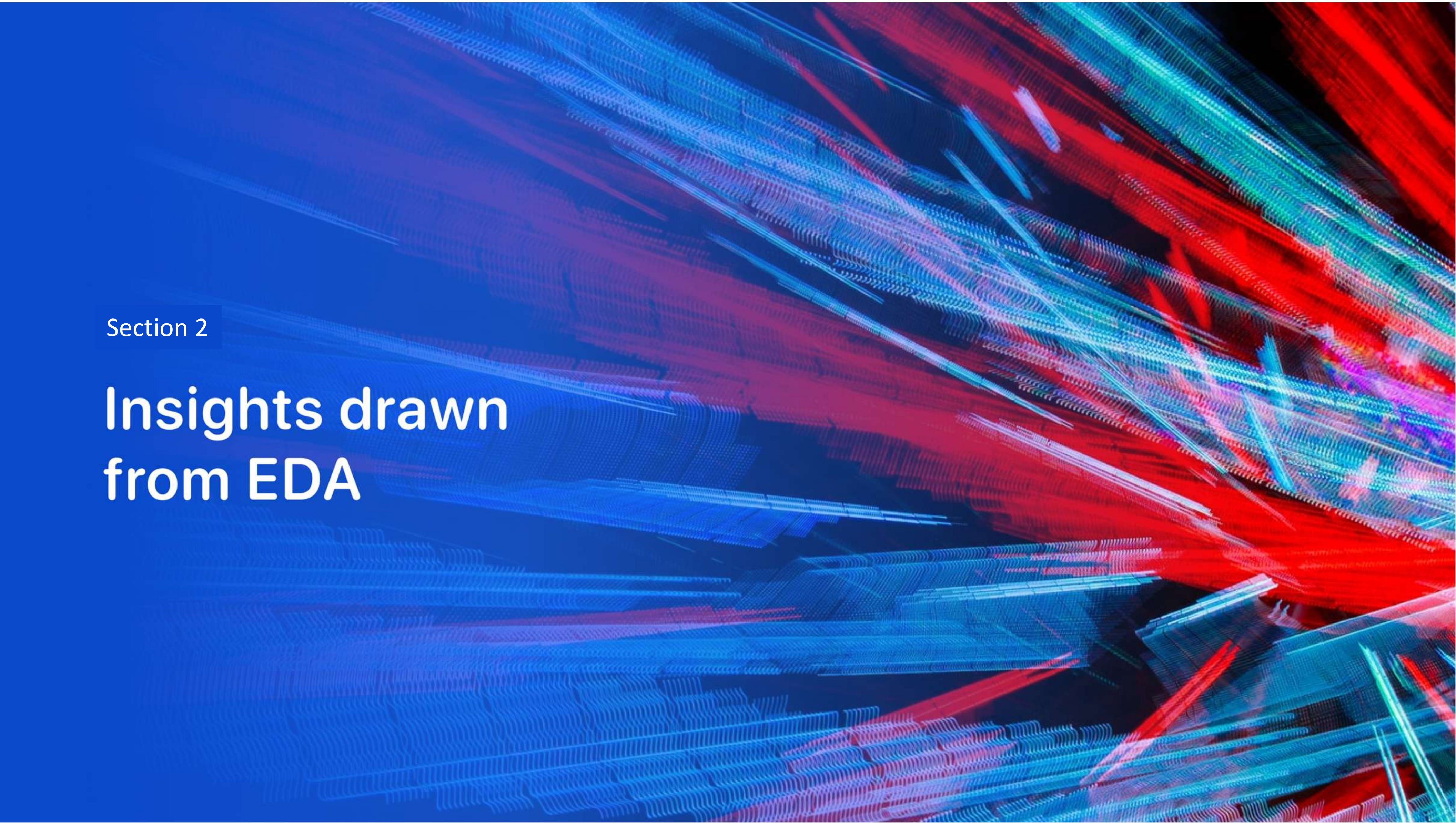
Training model

Evaluating model

Optimization hyperparameters

Model improvement

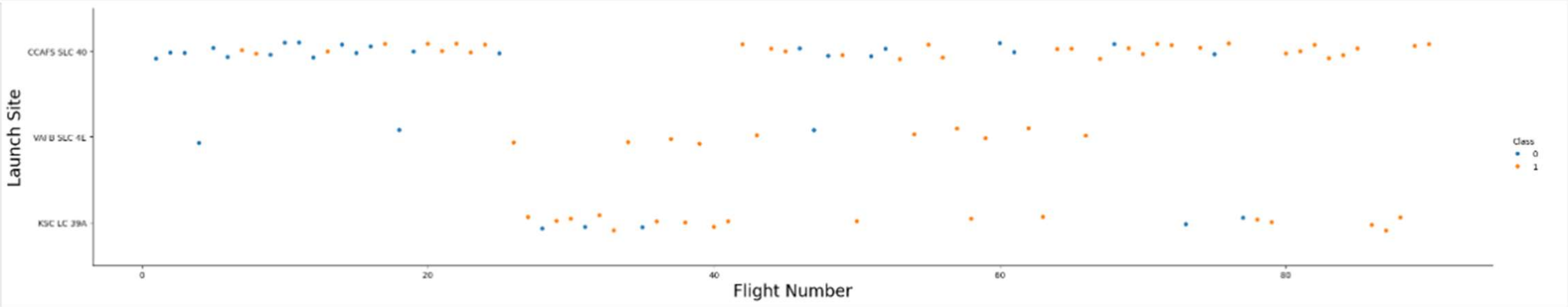
Github URL:



Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



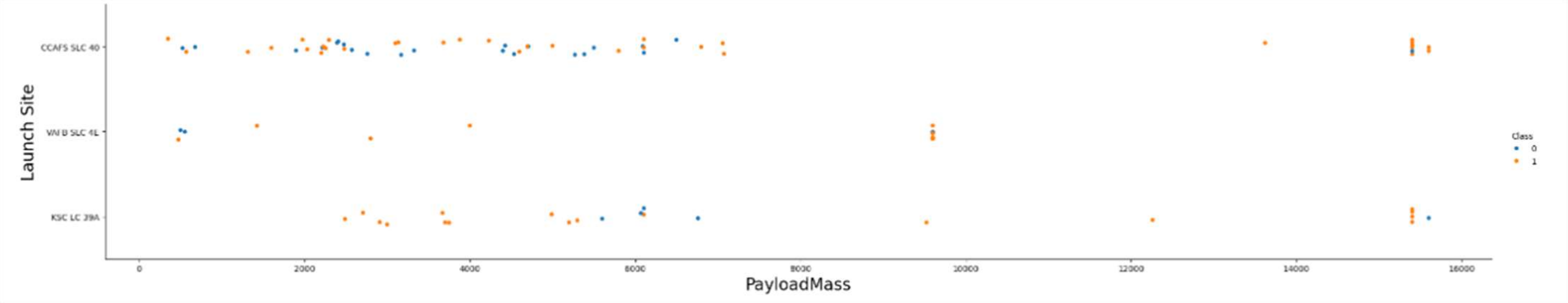
Data Distribution:

- It can be seen that the launches are spread out across all three launch sites.
- The number of failed launches (blue) appears to be more concentrated in the early launches.
- The number of successful launches (orange) increases as the number of launches increases. The CCAFS SLC 40 launch site appears to have the highest number of launches.

Trend:

Overall, the success rate appears to increase with the number of launches, suggesting improvements in launch technology or procedures.

Payload vs. Launch Site



Data Distribution:

- It can be seen that the launches are spread out across all three launch sites.
- The payload mass is fairly widely distributed across the launch sites. It appears that successful launches (orange) tend to be concentrated at higher payload mass.
- The CCAFS SLC 40 launch site appears to have a higher number of launches with larger payload mass.

Trend:

There appears to be a relationship between payload mass and launch results. Successful launches tend to have larger payload mass. However, there are still some unsuccessful launches with larger payload mass.

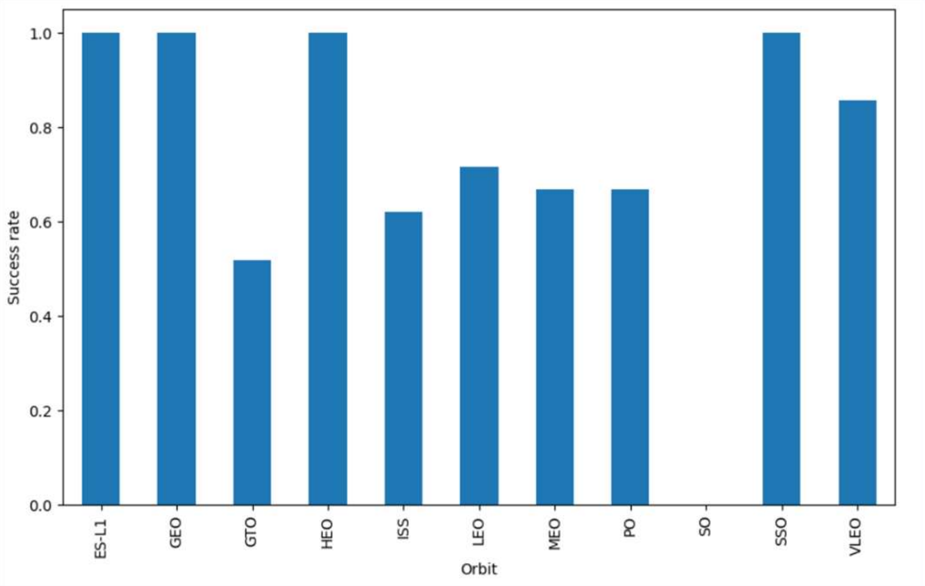
Success Rate vs. Orbit Type

Data:

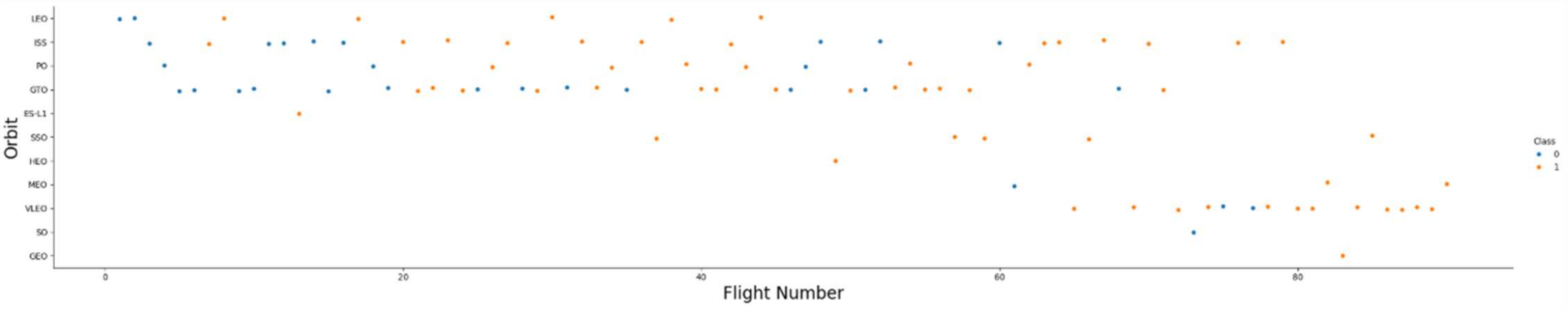
- ES-L1, GEO, HEO, SSO orbits have the highest success rates (1.0 or close to 100%). This indicates that launches into these orbits are generally very successful.
- GTO orbits have the lowest success rates (around 0.52). This indicates that launches into these orbits have a higher failure rate.
- ISS, LEO, MEO, PO, SO, VLEO orbits have intermediate success rates (ranging from 0.6 to 0.85).

Remarks:

- The chart shows significant differences in success rates between orbit types.
- Some orbits have near-perfect success rates, while others have significantly lower success rates.
- This can be explained by a number of factors, such as the complexity of the orbit, the technology used, or the purpose of the launch.



Flight Number vs. Orbit Type



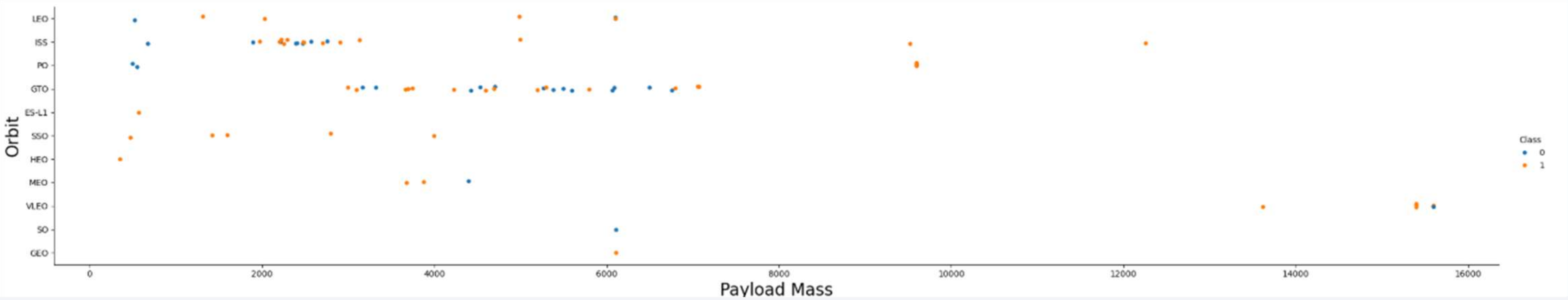
Data Distribution:

- It can be seen that the launches are spread out over different orbit types.
- Some orbit types, such as GEO, HEO, and SSO, appear to have higher success rates, with more orange (class 1) points.
- In contrast, some orbit types, such as GTO, appear to have higher failure rates, with more blue (class 0) points.
- The uneven distribution of data can be seen between orbit types, suggesting that there may be differences in complexity or risk between them.

Trend:

There is no clear trend in the success rate by number of flights. It appears that the launch results are not highly dependent on the number of flights.

Payload vs. Orbit Type



Data Distribution:

- It can be seen that launches are spread out over different orbit types and with different payload masses.
- Some orbit types, such as GEO, HEO, and SSO, appear to have higher success rates, with more orange (1) points.
- In contrast, some orbit types, such as GTO, appear to have higher failure rates, with more blue (0) points.
- The uneven distribution of data can be seen between orbit types and payload masses, suggesting that there may be differences in complexity or risk between them.

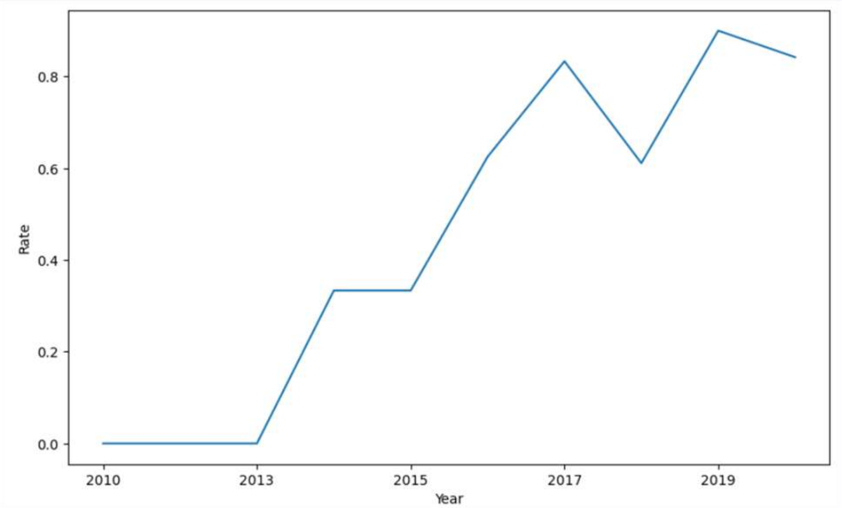
Trend:

There is no clear trend in success rates by payload mass. It appears that launch outcomes are not highly dependent on payload mass.

Launch Success Yearly Trend

Data:

- From 2010 to 2013: The ratio is 0.0, indicating little or no change during this period.
- 2013: The ratio begins to increase, indicating a significant improvement or change.
- 2013 to 2017: The ratio continues to increase gradually, indicating a continuous growth trend.
- 2017: The ratio peaks, indicating the highest level during the period under consideration.
- 2017 to 2018: The ratio drops significantly, indicating a decline or negative movement.
- 2018 to 2019: The ratio increases again, but does not reach the high level of 2017.

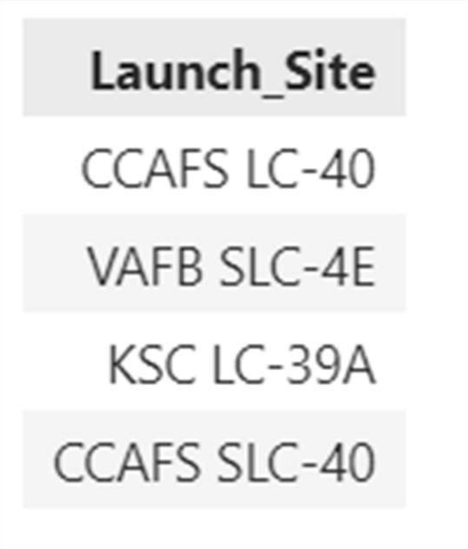


Remarks:

- The chart shows significant fluctuations in the ratio over time.
- There was a period of strong growth from 2013 to 2017, followed by a decline and partial recovery.

All Launch Site Names

SQL query: `%sql select distinct(Launch_Site) from SPACEXTABLE`

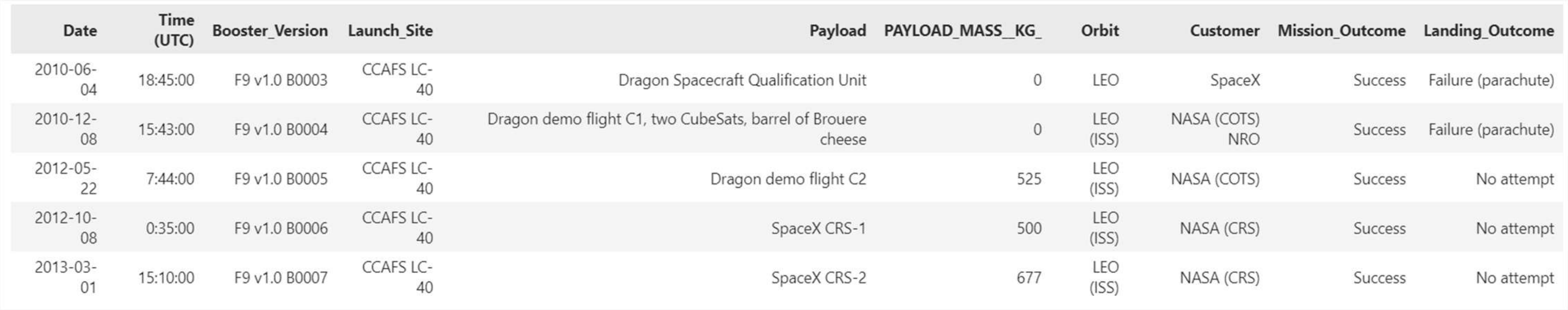


Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

There are 4 launch sites:

- 2 launch sites (CCAFS LC-40 and SLC-40) at Cape Canaveral Space Force Station (CCSFS) is an installation of the United States Space Force's Space Launch Delta 45, located on Cape Canaveral in Brevard County, Florida.
- Launch site KSC LC-39A at The John F. Kennedy Space Center, located on Merritt Island, Florida, is one of the National Aeronautics and Space Administration's (NASA) ten field centers.
- Launch site VAFB SLC-4E at Vandenberg Space Force Base, previously Vandenberg Air Force Base, is a United States Space Force Base in Santa Barbara County, California.

Launch Site Names Begin with 'CCA'



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

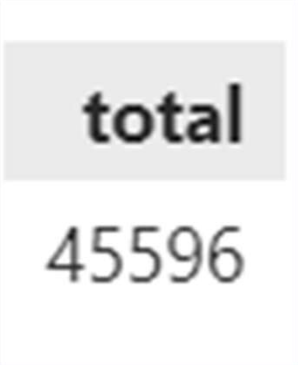
SQL query: `%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5`

Observation:

- All 5 launches were successful
- Early launches primarily did not succeed in recovering the booster
- Every launch in the dataset occurred at CCAFS LC-40
- Most payloads were relatively light (0 - 677 kg), likely because these missions were either test flights or light cargo deliveries to the ISS.
- All missions were aimed at Low Earth Orbit (LEO), especially the ISS (LEO (ISS)), SpaceX initially focused on supporting NASA’s ISS resupply missions.
- Most missions were carried out under contracts with NASA. In early years, SpaceX primarily operated under NASA contracts before expanding to commercial customers.

Total Payload Mass

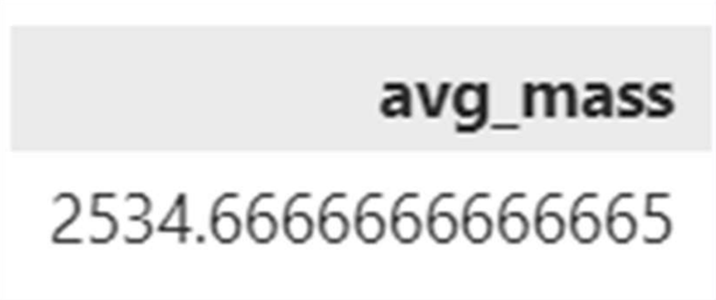
SQL query: %sql select sum(PAYLOAD_MASS__KG_) as total from SPACEXTABLE where Customer = 'NASA (CRS)’



total
45596

Average Payload Mass by F9 v1.1

```
SQL query: %sql select avg(PAYLOAD_MASS__KG_)
as avg_mass from SPACEXTABLE where
Booster_Version like 'F9 v1.1%'
```

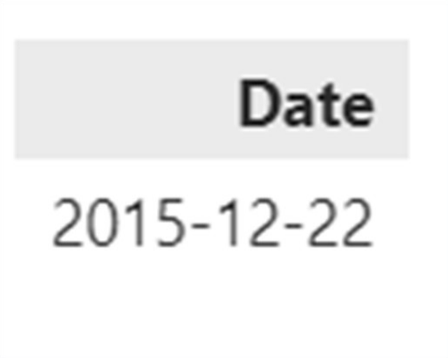


avg_mass
2534.6666666666665

First Successful Ground Landing Date

```
SQL query: %sql select min(Date) as Date from SPACEXTABLE
  where Landing_Outcome == 'Success (ground pad)’
```

On December 22, 2015, SpaceX achieved a major milestone in its history by successfully landing the first stage of its Falcon 9 rocket back on the ground. This was a major step forward in the development of reusable rockets, which would significantly reduce the cost of launching rockets and pave the way for more frequent space missions.

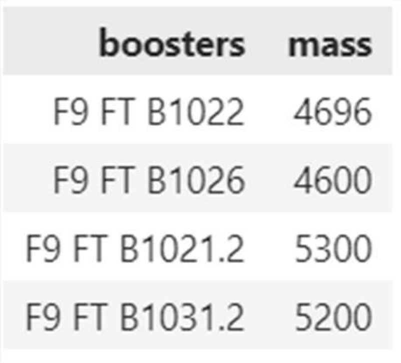


Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL query:

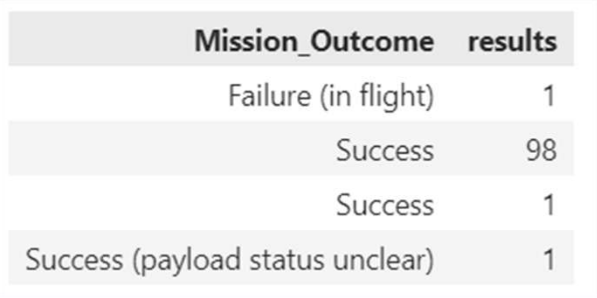
```
%sql select Booster_Version as boosters,
PAYLOAD_MASS__KG_ as mass from SPACEXTABLE where
Landing_Outcome == 'Success (drone ship)' and
(PAYLOAD_MASS__KG_ between 4000 and 6000)
```



boosters	mass
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

SQL query:
```
%sql select Mission_Outcome,
count(Mission_Outcome) as results from SPACEXTABLE group
by Mission_Outcome
```

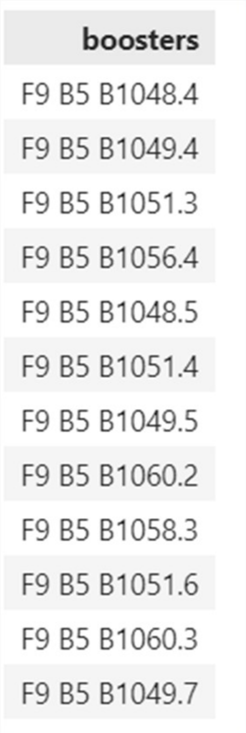


Mission_Outcome	results
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

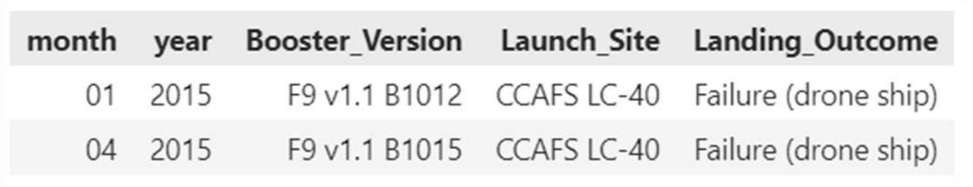
SQL query:

```
%sql select Booster_Version as boosters from
SPACEXTABLE where PAYLOAD_MASS__KG_ = (select
max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```



boosters
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



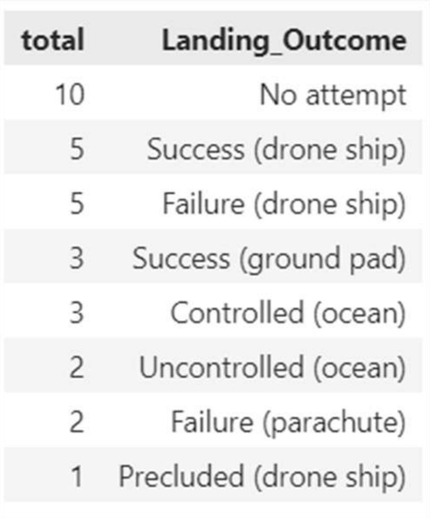
month	year	Booster_Version	Launch_Site	Landing_Outcome
01	2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

SQL query:

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

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

SQL query:
```
%sql select count(*) as total, Landing_Outcome from
SPACEXTABLE where (Date between '2010-06-04' and '2017-03-20')
group by Landing_Outcome order by total desc
```



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

Section 3

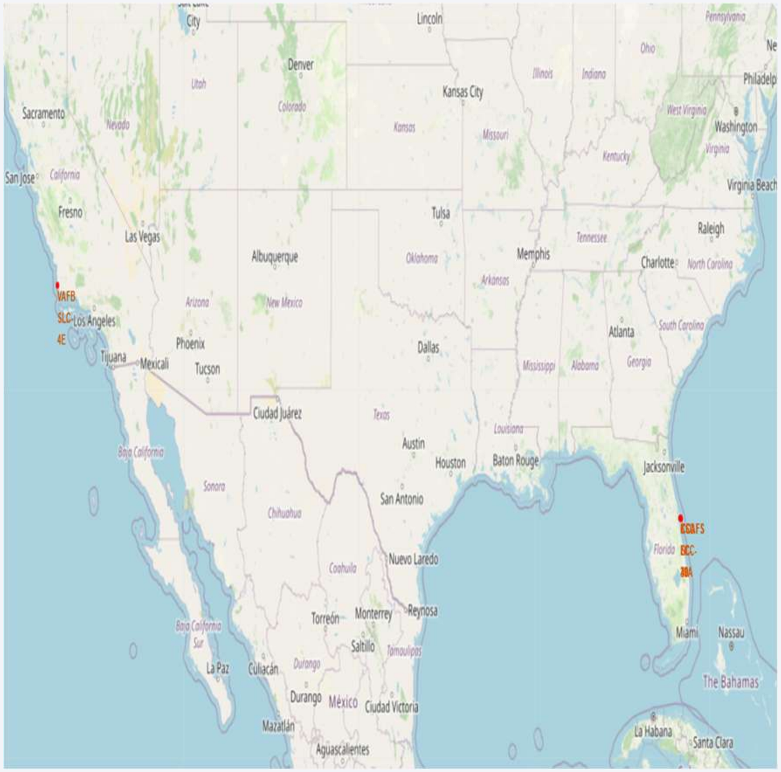
Launch Sites Proximities Analysis

All launch sites' location

Remarks:

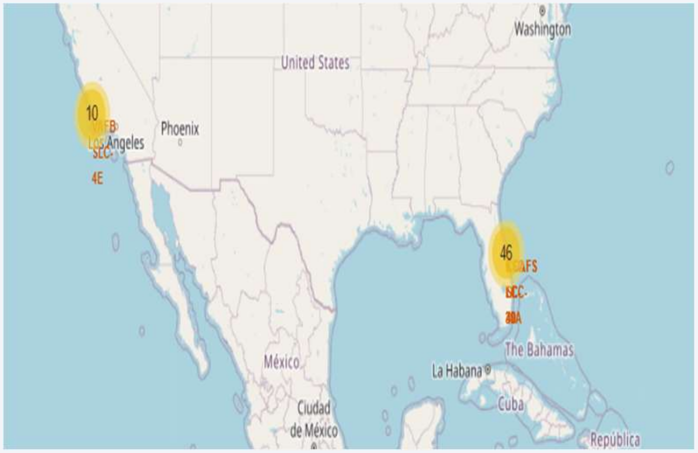
- 3 launch sites in Florida (CCAFS LC-40, CCAFS SLC-40 and KSC LC-39A) and 1 in California (VAFB SLC-4E).
- Geographic distribution: Launch sites are distributed on both coasts of the United States, allowing SpaceX to access a variety of orbits.
- Strategic Location: The launch sites are all located near the ocean, minimizing the risk to residential areas in the event of a launch failure.

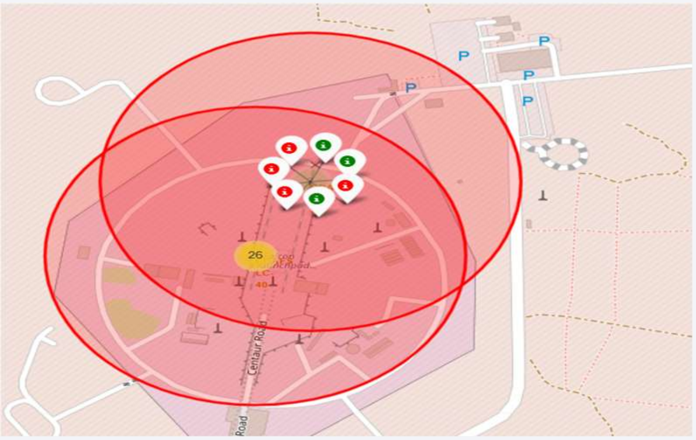
SpaceX has the ability to launch rockets from multiple locations, increasing its flexibility and reducing its dependence on a single location.



Success/failed launches for each site on the map

Area	Launch site	Number of launches	Success rate
Eastern	KSC LC-39A	13	76.92%
	CCAFS SLC-40	7	42.82%
	CCAFS LC-40	26	26.92%
Western	VAFB SLC-4E	10	40%





Launches from 2010-2018:

There does not appear to be a clear correlation between the number of launches and the success rate.

KSC LC-39A has the highest success rate with 13 launches.

CCAFS LC-40 has the highest number of launches with 26 but a low success rate. CCAFS SLC-40 has nearly same conditions for launch but it has higher success rate.

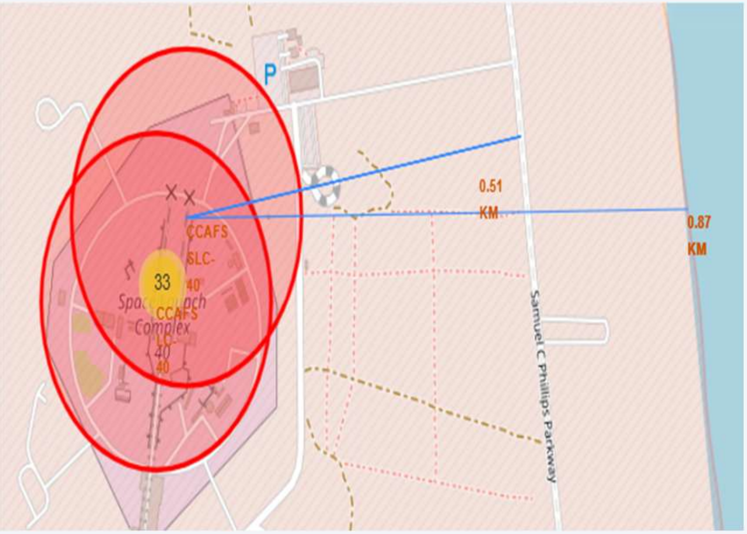
VAFB SLC-4E has 10 launches with an average success rate.

Distances between a launch site to its proximities

Launch sites located near the coastline, this helps to minimize the risk to residential areas in the event of a launch accident.

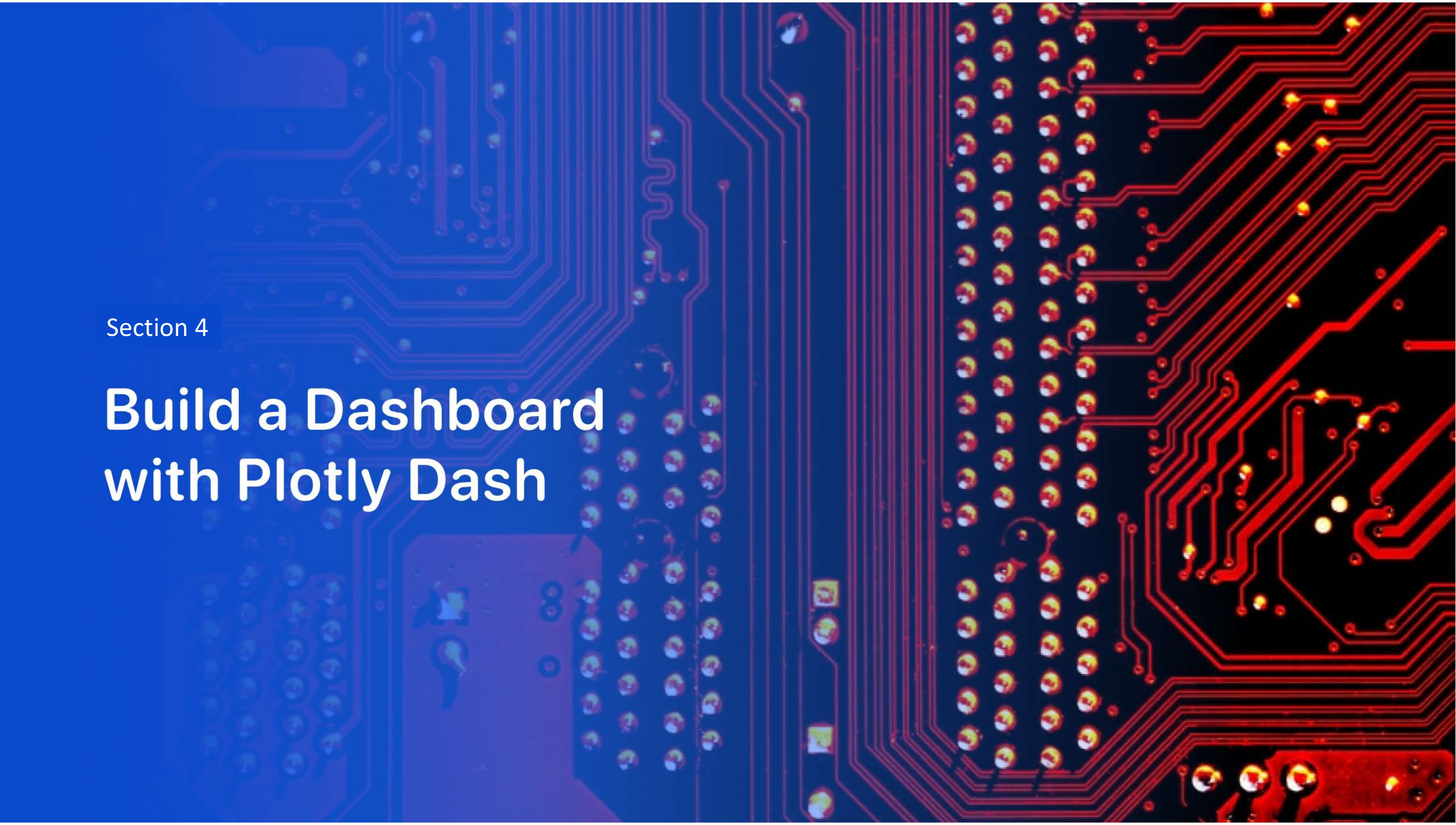
Space ships can be launched in the direction of the sea, avoiding densely populated areas.

Launch sites near transport facilities such as highways, railways, located in open areas and far from residential areas.



Section 4

Build a Dashboard with Plotly Dash



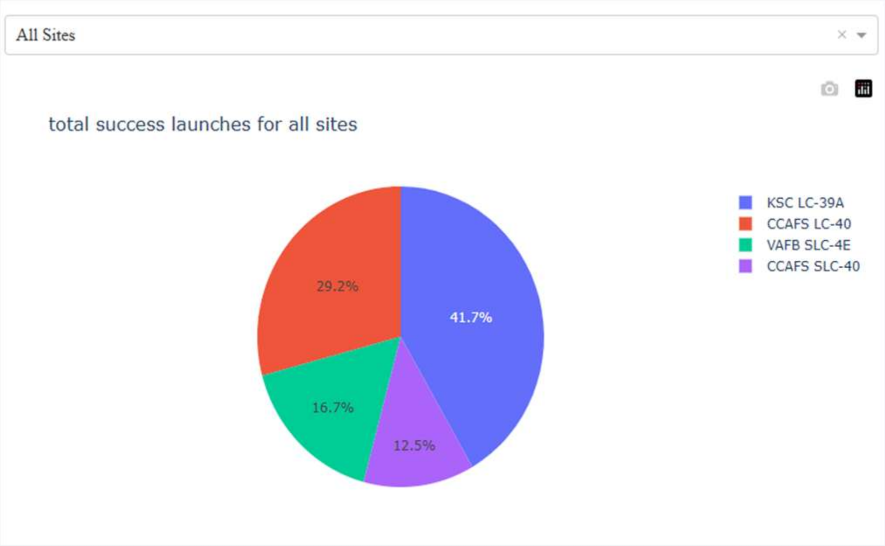
Total success launches for all sites

Analysis:

- KSC LC-39A: This is the launch site with the highest success rate, indicating that it is the most efficient launch site among the sites listed. This may be due to favorable geographical location, modern infrastructure, or the experience of the operating team.
- CCAFS LC-40: This site has the second-highest success rate, however, this rate is significantly lower than KSC LC-39A.
- VAFB SLC-4E: This site has the third lowest success rate. There may be unique challenges or limitations at this site, such as weather conditions or the type of rocket used.
- CCAFS SLC-40: This site has the lowest success rate. This suggests that there may be factors that negatively affect the success rate at this site.

Conclusion:

- The graph shows significant differences in success rates between launch sites.
- KSC LC-39A is the most efficient launch site, while CCAFS SLC-40 has the lowest success rate.
- Factors that influence success rates at each site need to be considered to make better decisions about launch site selection.



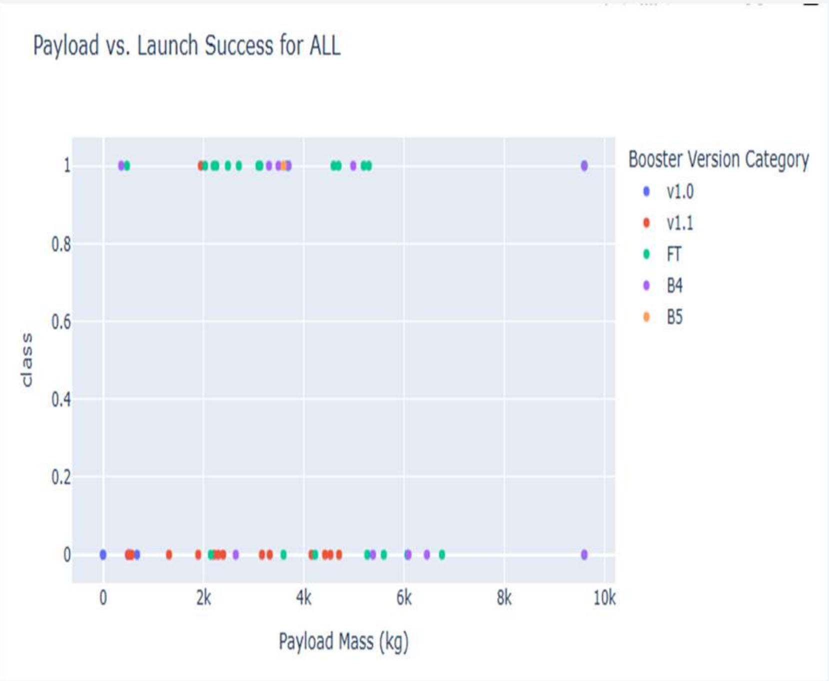
Payload vs. Launch Success for ALL

Data Distribution:

- It can be observed that successful launches (**class = 1**) are concentrated at lower payload masses (below **6000 kg**).
- Unsuccessful launches (**class = 0**) are scattered across the entire payload mass range.
- Different booster versions are used for launches with varying payload masses.
- **Booster versions v1.0 and v1.1** were mainly used for launches with lower payload masses.
- **Falcon 9 Full Thrust (FT) and B4 boosters** were used for launches with medium payload masses.
- **B5 boosters** were used for both low and high-payload mass launches.

Trends:

- There appears to be a relationship between payload mass and launch success. Successful launches tend to have lower payload masses.
- Booster type may also impact launch outcomes, as different boosters are used for launches with different payload masses.



Payload vs. Launch Success for each launch site

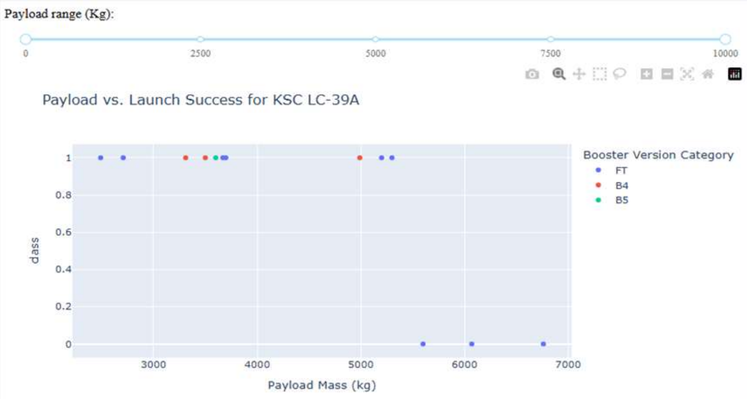
The period from 2010 to 2018:
At the KSC LC-39A launch site:

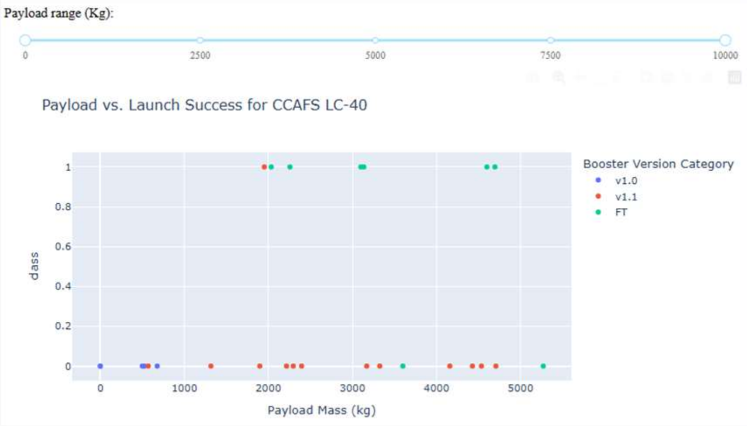
- Launches carrying lower payloads tended to have a higher success rate.
- The FT booster version was the most frequently used; however, it primarily carried low payloads (below 5,500 kg).

At the CCAFS LC-40 launch site:

- The FT booster also exhibited a high success rate with lower payloads, while the v1.1 booster version mostly failed, even with very low payloads.

Conclusion:
Payload mass appears to be a significant factor influencing success rates. However, no clear correlation between these two variables has been identified yet, requiring further evaluation with additional datasets.





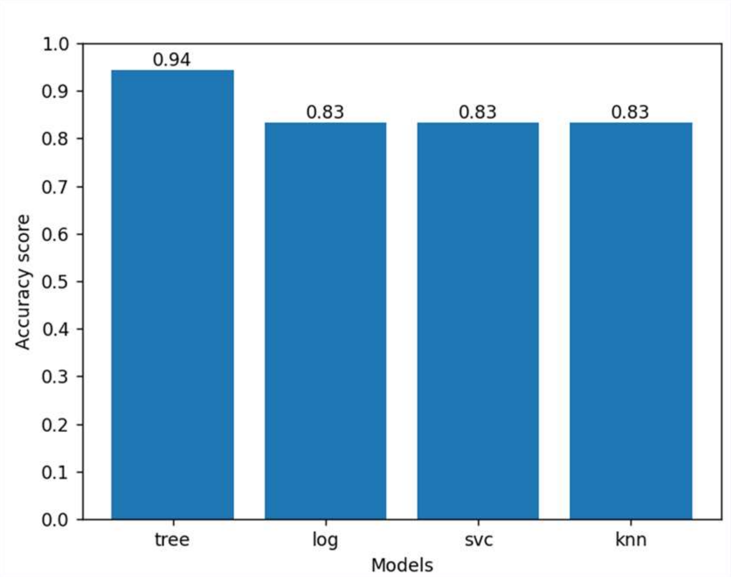
Section 5

Predictive Analysis (Classification)

Classification Accuracy

Remarks:

- The decision tree model achieved the highest accuracy score (0.94), indicating that it performed best in classifying the data.
- The other three models (LogisticRegression, Support Vector Machine, and K-nearest Neighbor) all had the same accuracy score (0.83), lower than the decision tree model.
- The difference in accuracy scores between the decision tree model and the other three models suggests that the decision tree model is better at capturing patterns in the data.



Conclusion:

Based on this diagram, the decision tree model is the best choice for classifying the data in this case.

model_tree = DecisionTreeClassifier(criterion = 'entropy', max_depth = 4, max_features = 'sqrt', min_samples_leaf = 1, min_samples_split = 10, splitter = 'random', random_state=42)

Confusion Matrix

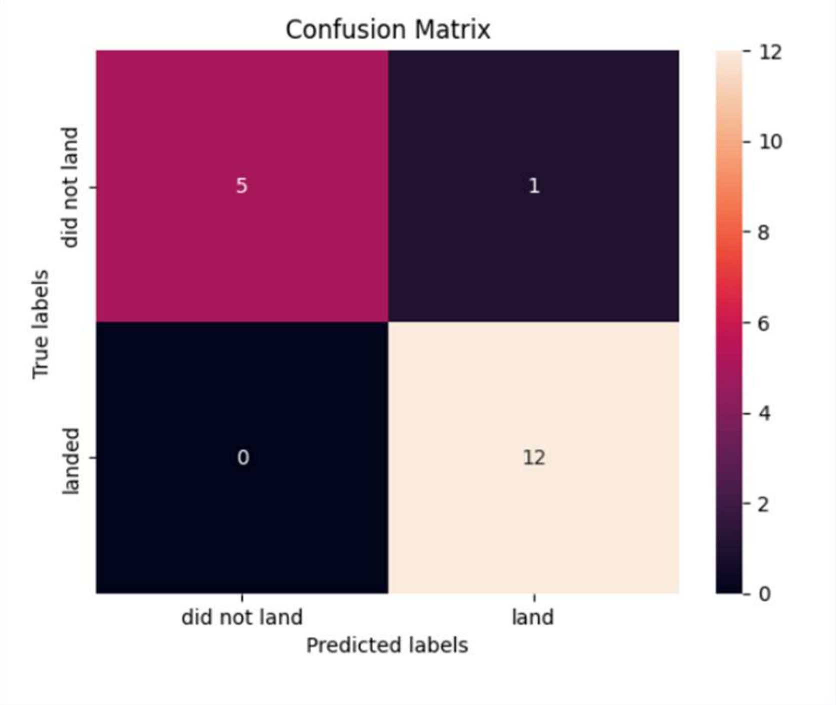
True Positive (TP): The model correctly predicts a sample belonging to the positive class (landed). In this case, TP = 12.

True Negative (TN): The model correctly predicts a sample belonging to the negative class (did not land). In this case, TN = 5.

False Positive (FP): The model incorrectly predicts a sample belonging to the negative class as positive (predicts landed but it actually did not land). In this case, FP = 1. This is a Type I error.

False Negative (FN): The model incorrectly predicts a sample belonging to the positive class as negative (predicts did not land but it actually landed). In this case, FN = 0. This is a Type II error.

The model gives a high accuracy rate of 94%, indicating that its prediction performance on the total number of predictions is effective and well-rated. However, to more specifically evaluate the model's performance in practice, it's necessary to consider the Precision, Recall, and F1-score metrics.



Conclusions

- The churn prediction success rate for Falcon9 launches provides a powerful tool for mitigating launch failure. By accurately identifying important features, the model enables proactive, targeted interventions that can significantly improve success rates. Insights from the feature important offer actionable areas for improvement in space missions.

- Implementing the recommendations will help SpaceX not only retain the prediction success rate but also optimize costs for launches. This data-driven approach lays the foundation for sustained growth for SpaceX.

Appendix

- Dataset: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Jupiter Notebook lab on Skills Network Labs
- Cloud IDE on Skills Network Labs

Thank you!