IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Carlos Eduardo Ugarteche Virhuez>
<20 December 2023>

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data Collection: Utilized SpaceX REST API and web scraping techniques.
  - Data Wrangling: Processed data to create a success/fail outcome variable.
  - Exploratory Data Analysis (EDA):
    - Explored data visually, considering payload, launch site, flight number, and yearly trends.
    - Employed SQL to calculate statistics such as total payload, payload range for successful launches, and total counts of successful and failed outcomes.
  - Launch Site Analysis:
    - Investigated launch site success rates and their proximity to geographical markers.
    - Visualized launch sites with the highest success rates and successful payload ranges.
  - Predictive Modeling:
    - Developed models using logistic regression, support vector machine (SVM), decision tree, and K-nearest neighbor (KNN) to predict landing outcomes.
- **Summary of all results**
  - Exploratory Data Analysis:
    - Over time, there has been a notable improvement in launch success.
  - Visualization/Analytics:
    - Most launch sites are located near the equator and in close proximity to the coast.
  - Predictive Analytics:
    - All models performed similarly on the test set, with the decision tree model slightly outperforming others in 2023.

Introduction

- Project background

SpaceX, a trailblazer in space exploration, has been dedicated to democratizing space travel, achieving milestones like spacecraft missions to the international space station and deploying a revolutionary satellite constellation for global internet access. The company's groundbreaking approach involves reusing the first stage of its Falcon 9 rocket, resulting in remarkably affordable launches at $62 million, in stark contrast to competitors' costs exceeding $165 million without reusability. Acknowledging the pivotal role of first stage landings in cost determination, a startup competitor aims to leverage data science and machine learning to predict landing outcomes. This strategic initiative is crucial for determining competitive bidding prices against SpaceX, ensuring viability and success in the dynamic space industry. As we embark on this project, inspired by SpaceX's innovation, we seek to navigate the challenges of the industry with a forward-thinking mindset, echoing the spirit of progress that defines space exploration.

- Problems to answer
  - Influence of Payload Mass, Launch Location, Flight Count, and Orbital Trajectories on the Success of First-Stage Landings.
  - Analysis of the Trend in Successful Landings Over Time.
  - Identification of the Optimal Predictive Model for First-Stage Landing Success.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX REST API and web scrapping.
- Perform data wrangling
  - Data was processed using one-hot encoding for categorical features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models.

Data Collection

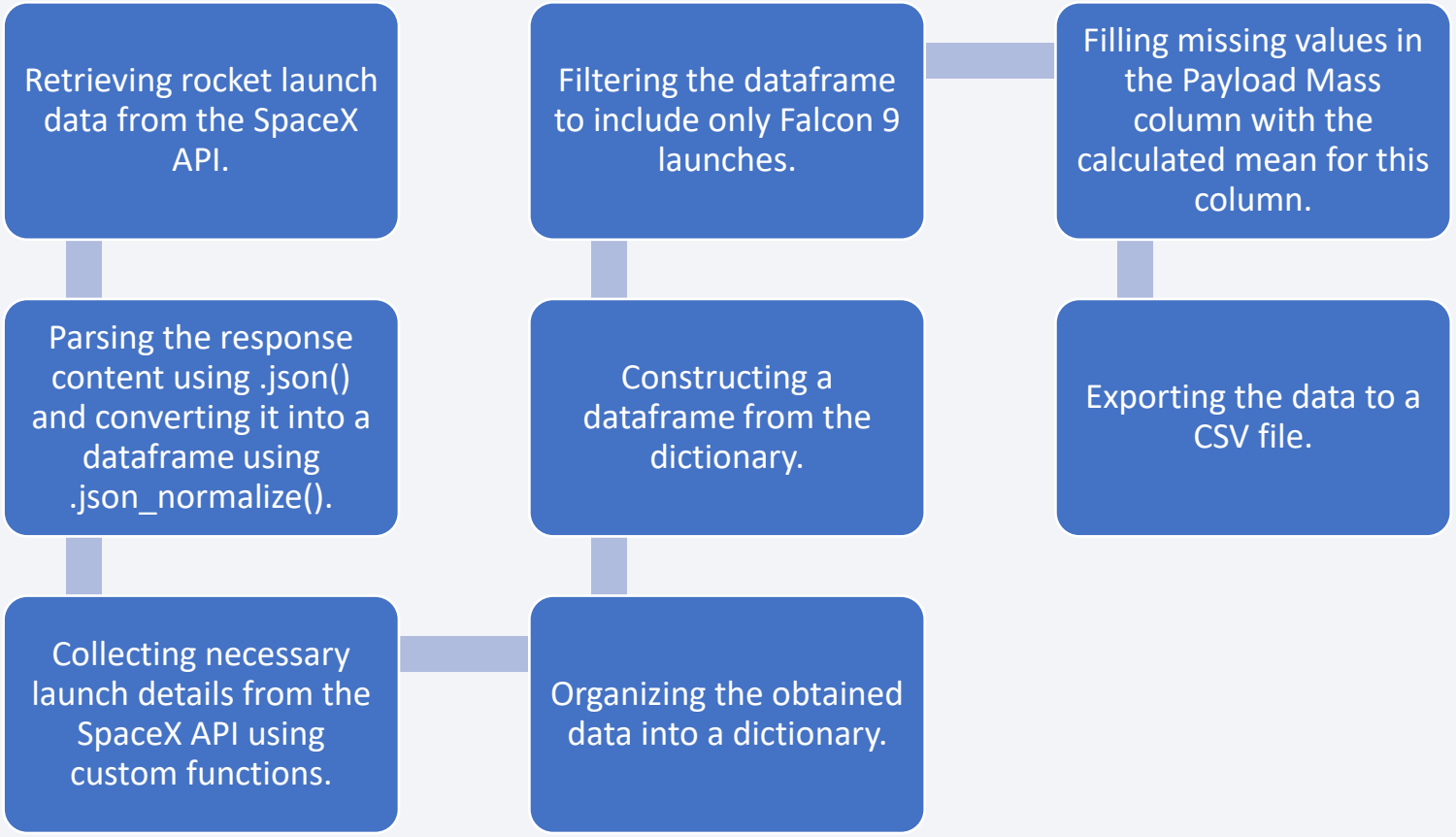
- This process involved a combination of API requests from SpaceX's REST API and web scraping data from a table in SpaceX's Wikipedia entry. Both methods were used to ensure comprehensive data collection for a detailed analysis of launches.

- By combining these two sources of data, we aimed to gather comprehensive information about launches for a more detailed and thorough analysis.

SpaceX's REST API	Wikipedia web scraping
FlightNumbers	Flight No.
Date	Launch site
BoosterVersion	Payload
Orbit	PayloadMass
PayloadMass	Orbit
Customer	Customer
LaunchSite	Launch outcome
Outcome	Version Booster
Flights	Booster landing
GridFins	Date
Reused	Time
Legs	
LandingPad	
Block	
ReusedCount	
Serial	
Longitude	
Latitude	

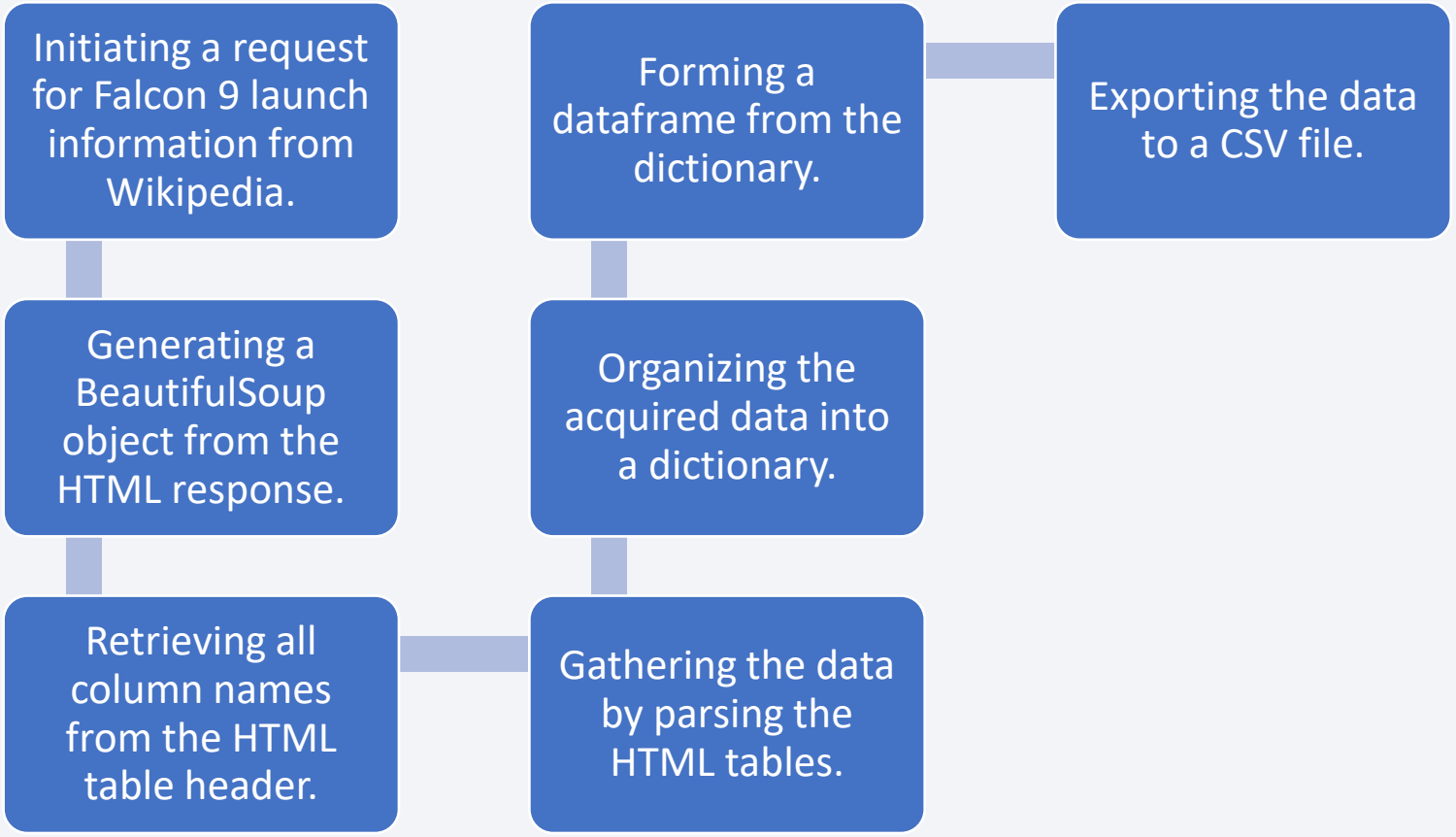
Data Collection – SpaceX API

- Reference: Github



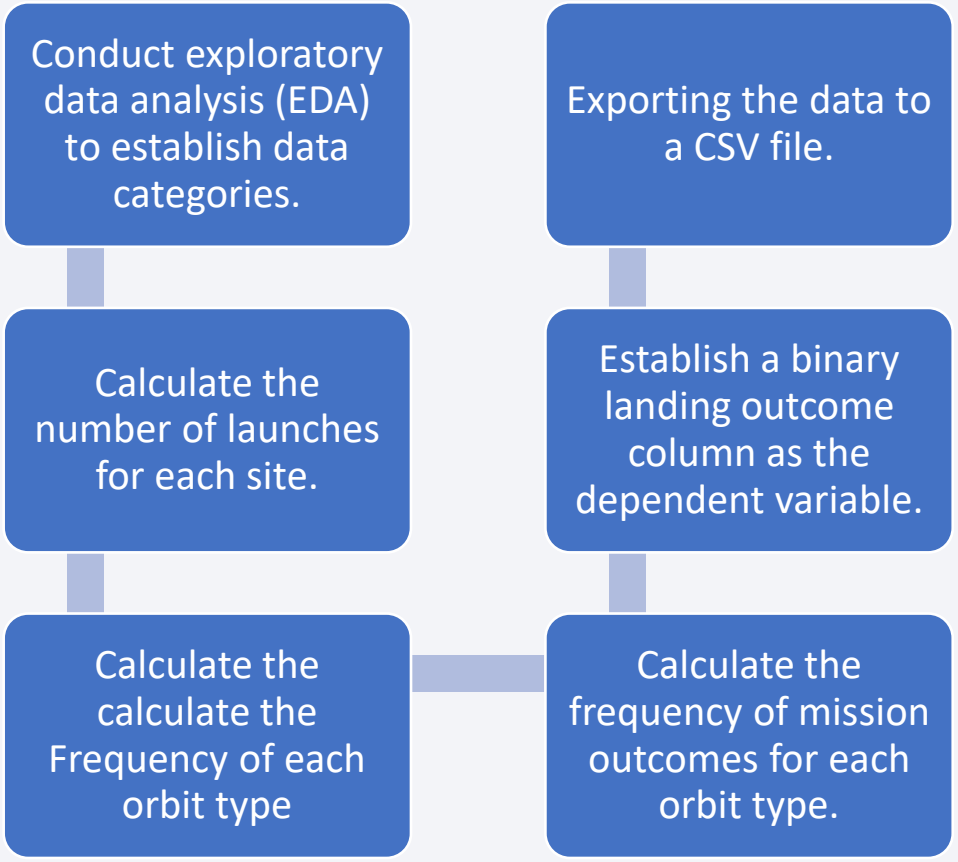
Data Collection - Scraping

- Reference: Github



Data Wrangling

- Reference: Github



EDA with Data Visualization

Scatter plots visually represent the relationship between different attributes. Once patterns are discerned from these graphs, it becomes easier to identify the factors that have the most impact on the success of landing outcomes.

We used the scatter plots to represent the relationship between:

- Flight Number and Launch Site
- Payload and Launch Site
- Flight Number and Orbit type
- Payload and Orbit

Plus we used a bar chart to visualize the relationship between success rate of each orbit type.

Reference: Github

EDA with SQL

Executed SQL queries:

1. Displaying the names of the unique launch sites in the space mission.
2. Displaying 5 records where launch sites begin with the string ‘CCA’.
3. Displaying the total payload mass carried by boosters launched by NASA (CRS).
4. Displaying the average payload mass carried by the booster version F9 v1.1.
5. Listing the date when the first successful landing outcome on a ground pad was achieved.
6. Listing the names of the boosters which have successfully landed on a drone ship and have a payload mass greater than 4000 but less than 6000.
7. Listing the total number of successful and failed mission outcomes.
8. Listing the names of the booster versions which have carried the maximum payload mass.
9. Listing the failed landing outcomes on a drone ship, along with their booster versions and launch site names, for the months in the year 2015.
10. Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20 in descending order.

Reference: Github

Build an Interactive Map with Folium

Generated markers for all launch sites.

- Ranked the count of landing outcomes, Failure or Success between June 4, 2010, and March 20, 2017, in descending order.

Created markers with colors to represent the launch outcomes for each launch site.

- Green markers signify successful launches, while red markers indicate failures. Implemented Marker Cluster to visually identify launch sites with notably high success rates.

Illustrated distances between a launch site and its proximities,.

- Included railways, highways, coastlines, and the closest city. Utilized colored lines to display these distances, using KSC LC39A as an example.

Reference: Github

Build a Dashboard with Plotly Dash

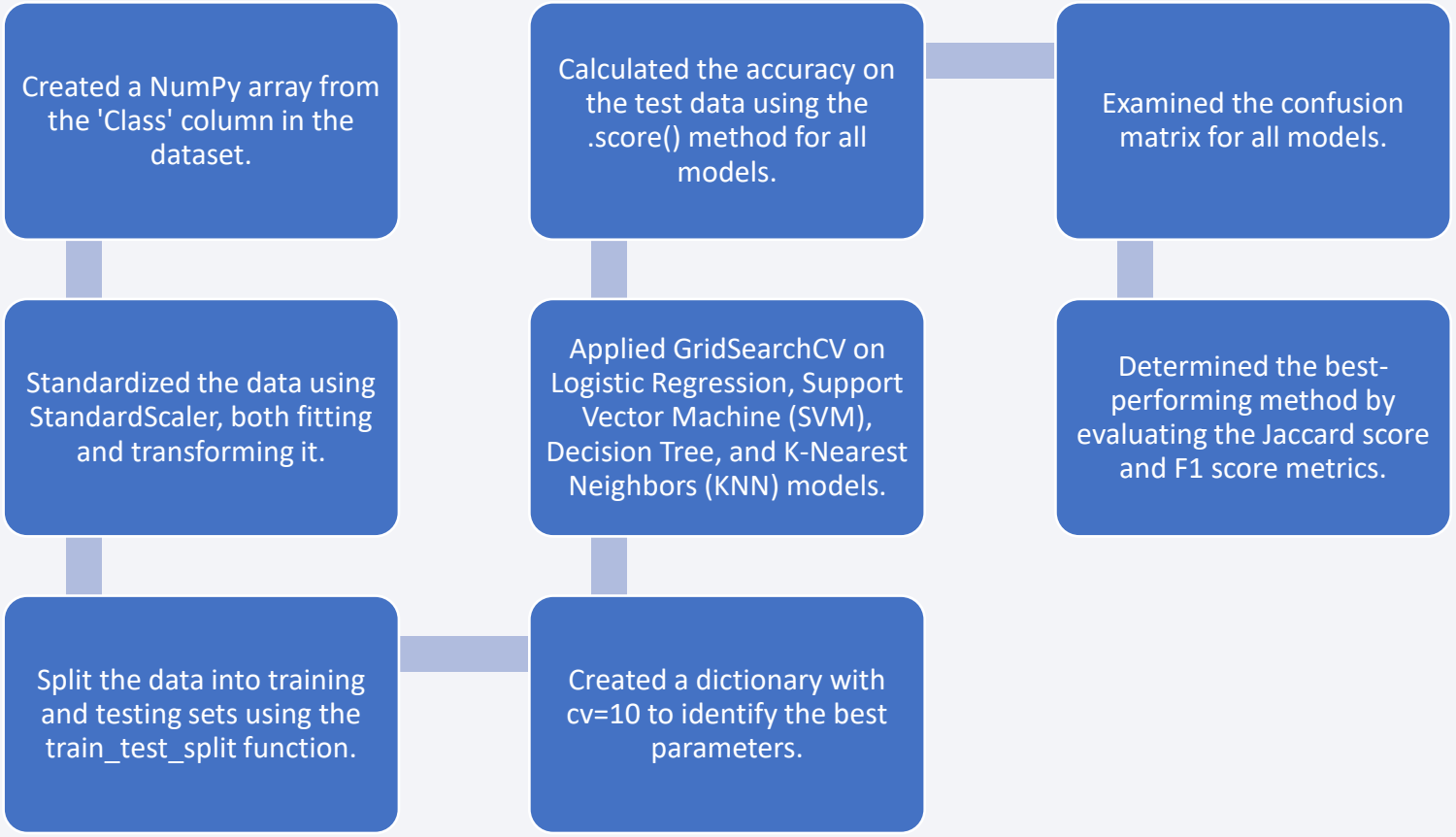
I've included the following Charts:

1. Dropdown List with Launch Sites
    - Enables users to choose all launch sites or a specific one.
2. Pie Chart Showing Successful Launches
    - Presents successful and unsuccessful launches as percentages of the total.
3. Slider for Payload Mass Range
    - Allows users to select a range of payload masses.
4. Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version
    - Displays the correlation between payload mass and launch success, categorized by booster version.

Reference: Github

Predictive Analysis (Classification)

- Reference: Github



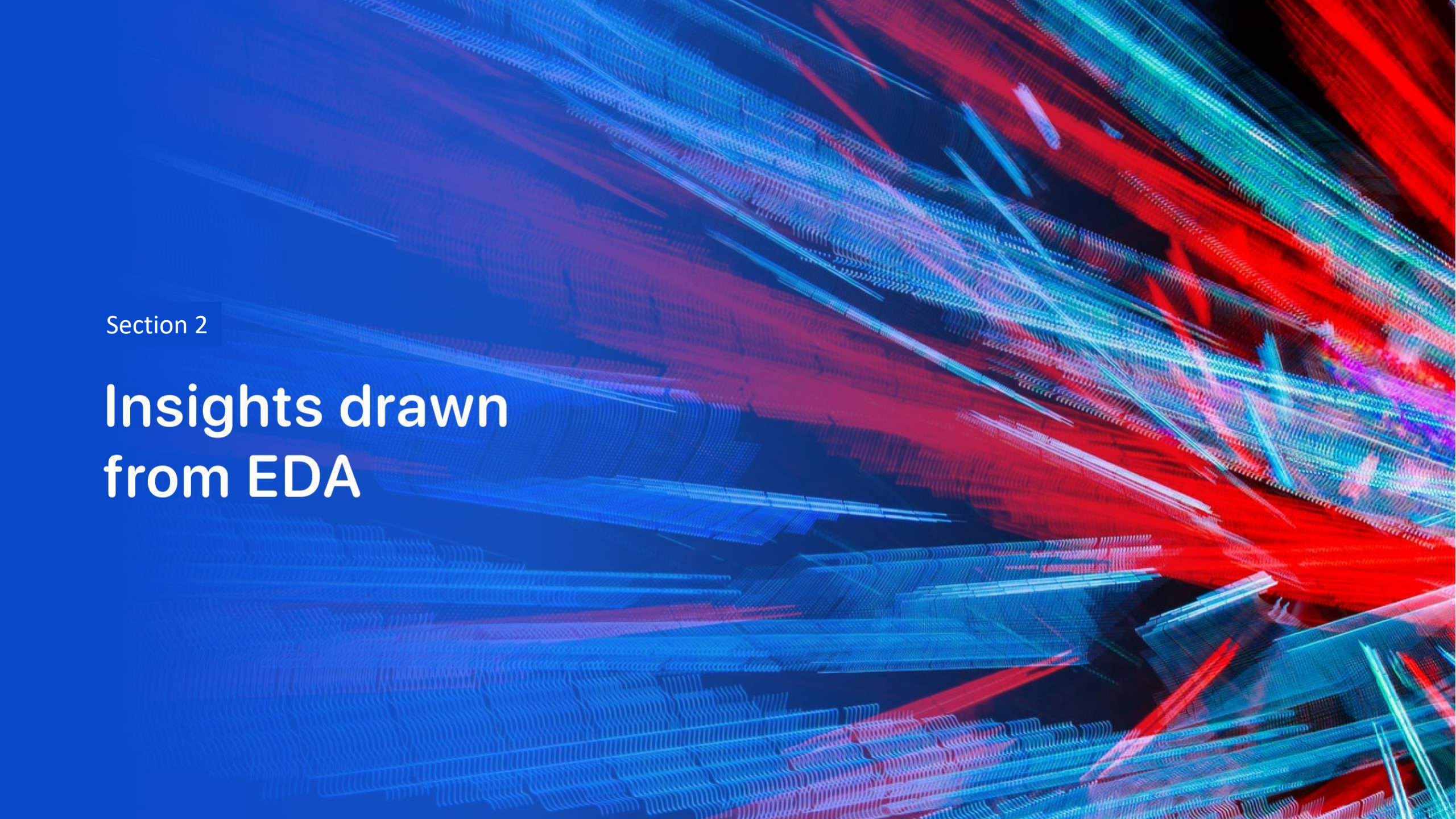
Results

On Section 2 and 3 we will explore the results of ours:

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

Section 2

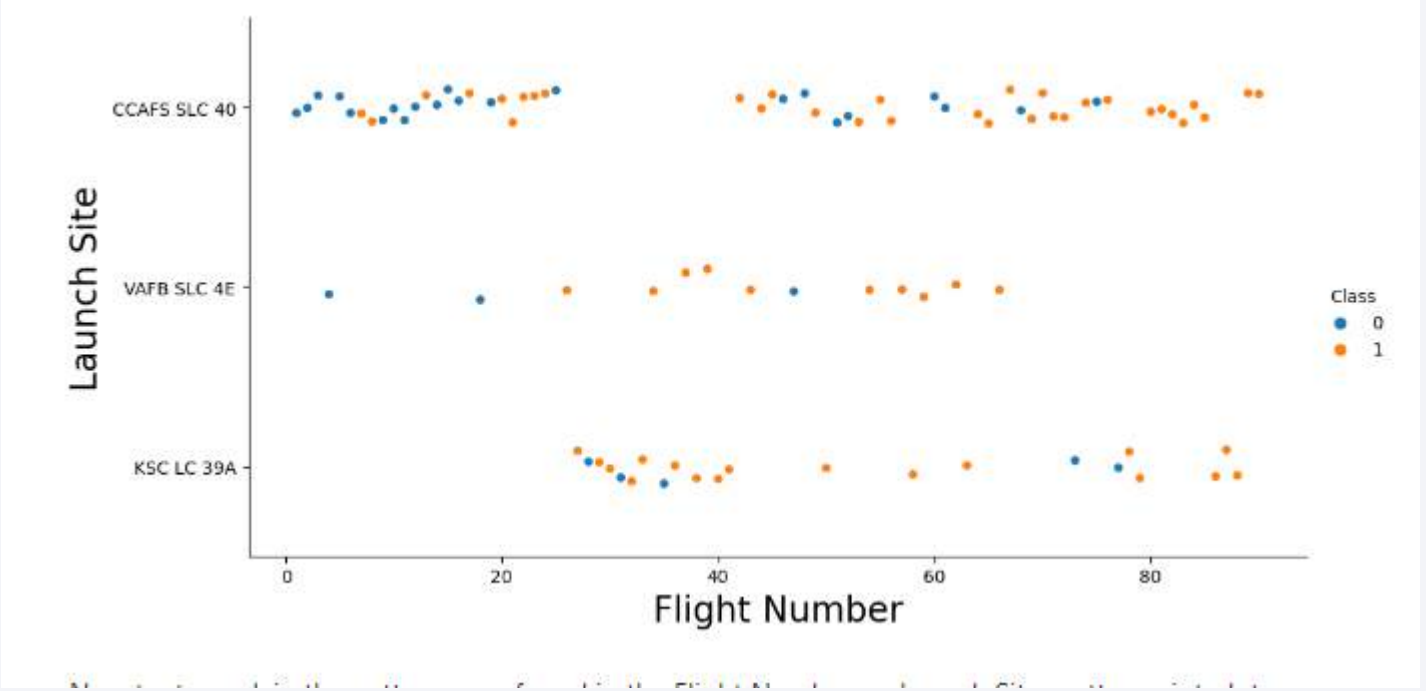
Insights drawn from EDA



Flight Number vs. Launch Site

Observations where blue represents failed launches and orange represents success:

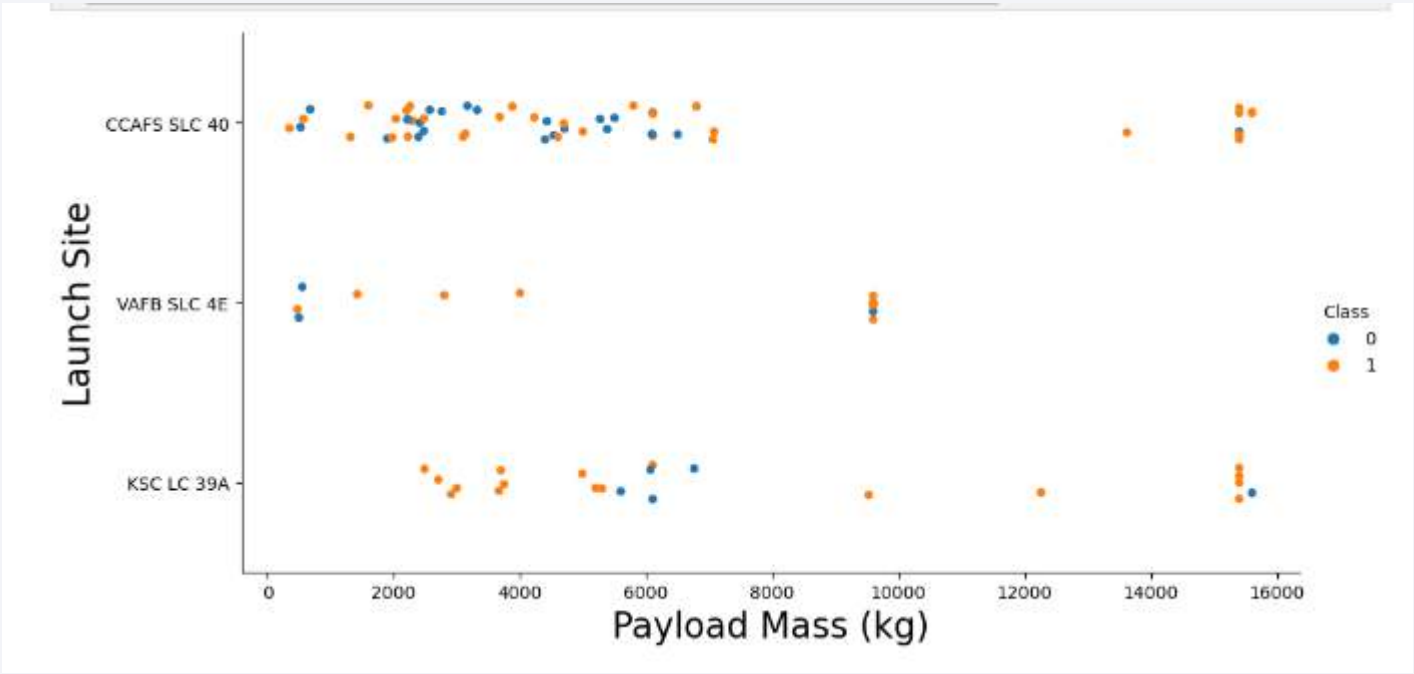
- The CCAFS SLC 40 launch site accounts for approximately half of all launches.
- The earliest flights experienced failure, while the latest flights have all succeeded.
- VAFB SLC 4E and KSC LC 39A have notably higher success rates.
- It can be inferred that each new launch has a higher likelihood of success.



Payload vs. Launch Site

Observations where blue represents failed launches and orange represents success:

- For every launch site, there is a positive correlation between payload mass and success rate, indicating that higher payload masses tend to result in higher success rates.
- For KSC LC 39A, there is a 100% success rate for payload masses under 5500 kg.
- Additionally, a majority of launches with payload masses above 7000 kg were successful.



Success Rate vs. Orbit Type

Success rate from Orbits:

0% Success rate:

- SO

1% to 49% Success rate:

- NONE

50% to 75% Success rate:

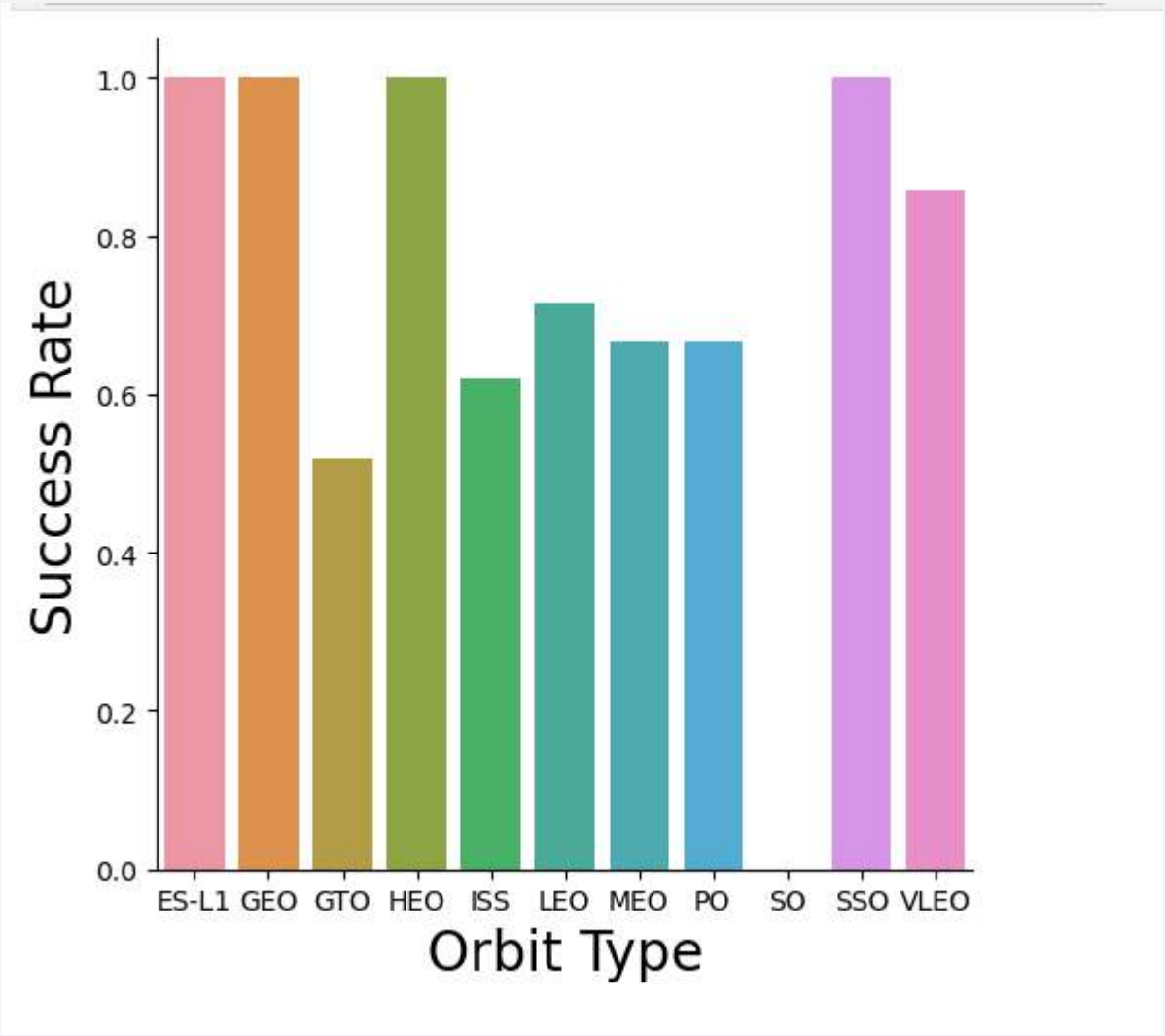
- GTO, ISS, LEO, MEO and PO

76% to 99% Success rate:

- VLEO

100% Success rate:

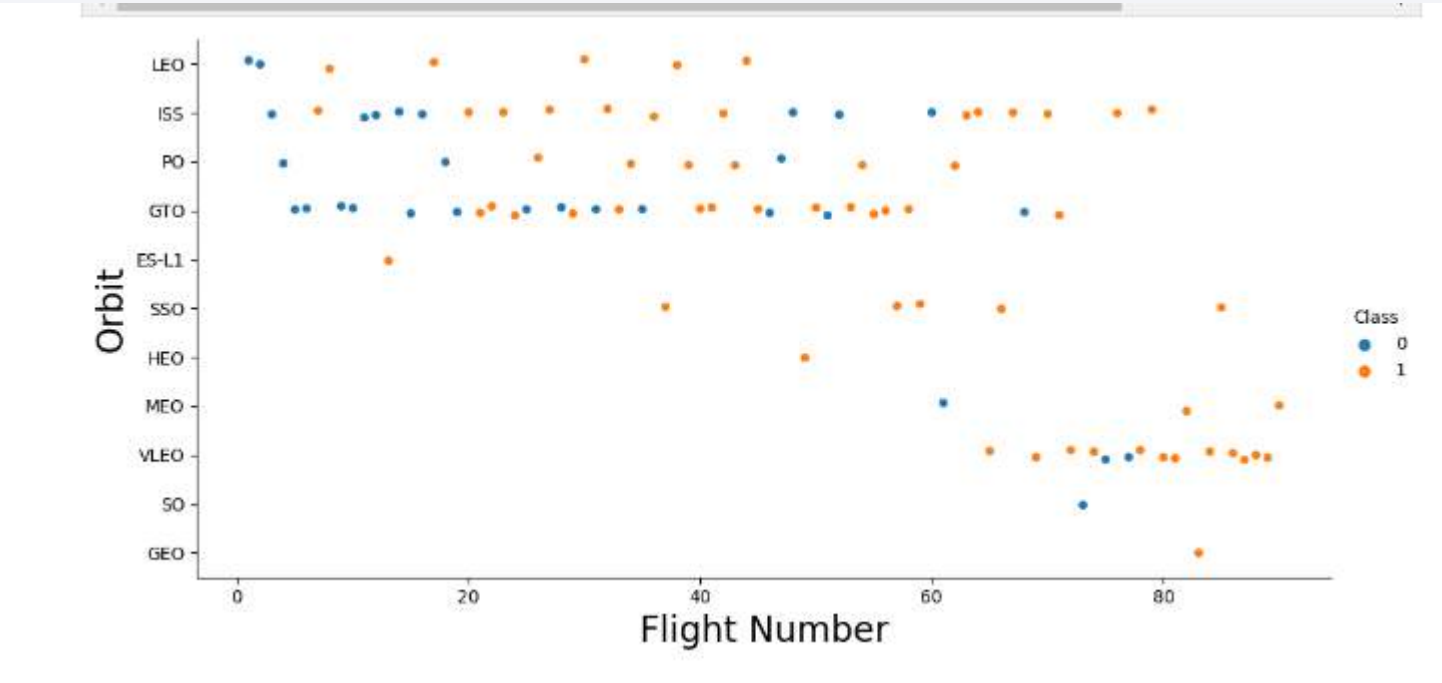
- ES-L1, GEO, HEO and SSO



Flight Number vs. Orbit Type

Observations where blue represents failed launches and orange represents success:

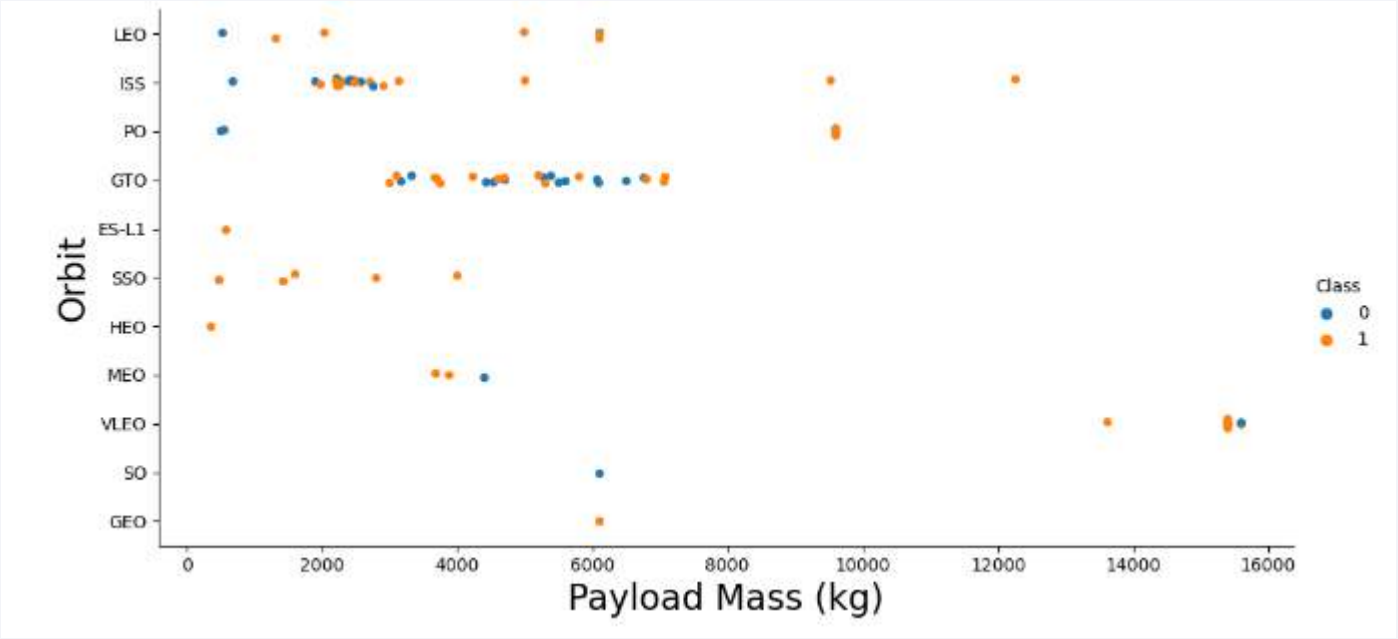
In Low Earth Orbit (LEO), the success rate seems to be influenced by the number of flights, suggesting that more flights in LEO orbit are associated with higher success rates. However, in Geostationary Transfer Orbit (GTO), there appears to be no discernible relationship between the flight number and the success rate.



Payload vs. Orbit Type

Observations where blue represents failed launches and orange represents success:

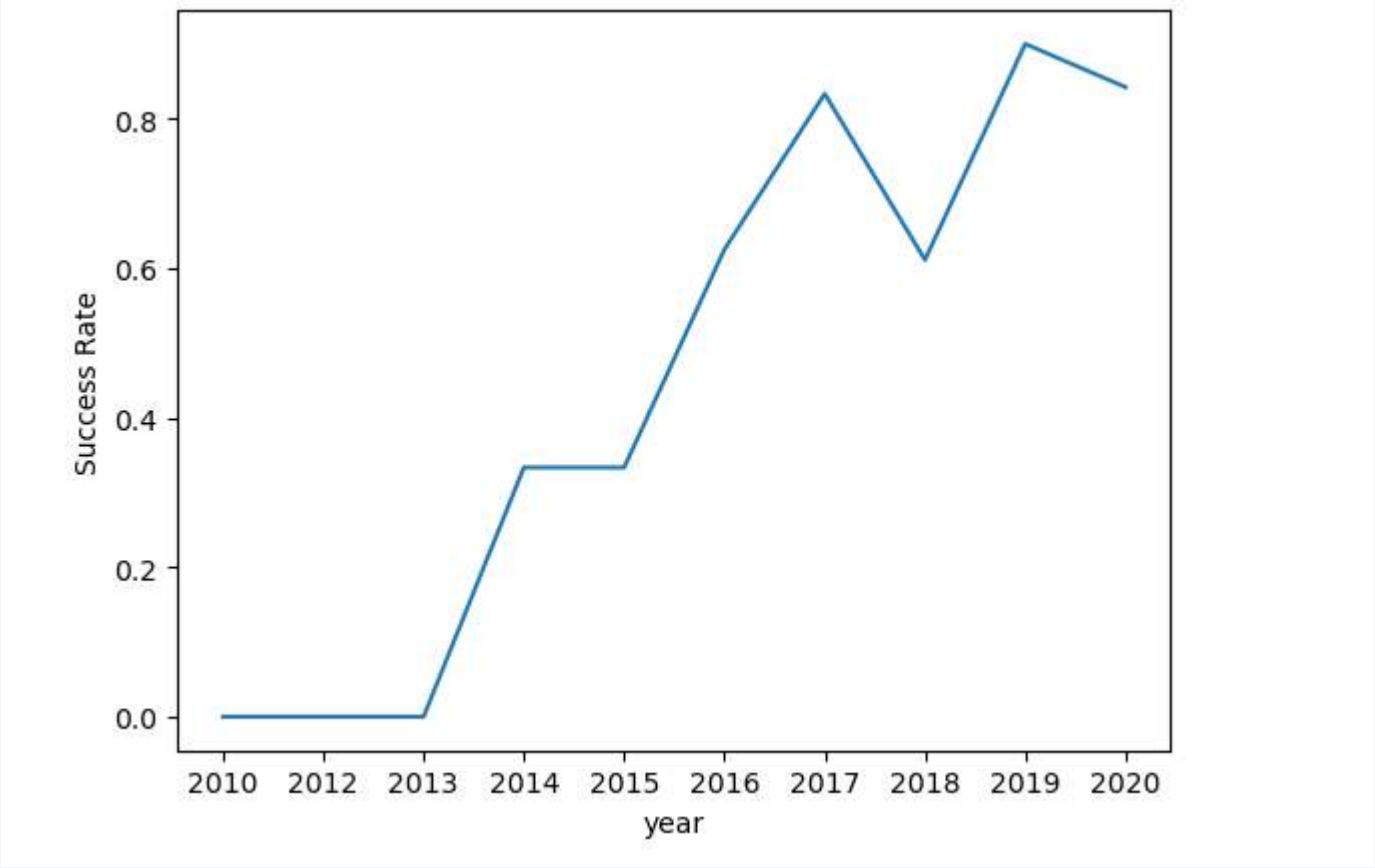
- Heavy payloads demonstrate better success rates in Low Earth Orbit (LEO), International Space Station (ISS), and Polar Orbit (PO).
- In contrast, the Geostationary Transfer Orbit (GTO) shows mixed success rates with heavier payloads, indicating less consistency in successful launches for this orbit with heavier payloads.



Launch Success Yearly Trend

Observations:

- The success rate showed improvement from 2013-2017 and 2018-2019.
- However, there was a decrease in the success rate from 2017-2018 and again from 2019-2020.
- Despite these fluctuations, the overall trend indicates an improvement in the success rate since 2013.



All Launch Site Names

- We utilized the DISTINCT function to display only the unique launch sites from the SpaceX data.

```
%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

Launch Site Names Begin with 'CCA'

- We used the following query to display 5 records where launch sites begin with 'CCA':

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- We calculated the total payload carried by boosters from NASA to be 45596 using the following query:

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTBL where customer = 'NASA (CRS)';
```

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

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- We calculated the average payload mass carried by booster version F9 v1.1 as 2928.4

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTBL where booster_version like '%F9 v1.1%';
```

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

average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

- List the date when the first successful landing outcome on a ground pad was achieved:

```
%sql select min(date) as first_successful_landing from SPACEXTBL where Landing_Outcome = 'Success (ground pad)';
```

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

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of the boosters that have succeeded in landing on a drone ship and have a payload mass greater than 4000 but less than 6000:

```
%sql select booster_version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

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

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

Total Number of Successful and Failure Mission Outcomes

- List of the number of successful and failure mission outcomes:

```
%sql select mission_outcome, count(*) as total_number from SPACEXTBL group by mission_outcome;
```

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

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

Boosters Carried Maximum Payload

- We determined the booster that have carried the maximumpayload using a subquery in the WHERE clause and the MAX() function.

```
%sql select booster_version from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL);
```

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

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- Used a combination of the WHERE clause, LIKE, AND, and BETWEEN conditions to filter for failed landing outcomes on a drone ship, including their booster versions and launch site names, for the year 2015.:

```
%sql SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, Landing_Outcome FROM SPACEXTBL where Landing_Outcome = 'Failure (drone ship)' and DATE LIKE '2015-%';
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
5-	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
5-	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Extracted the landing outcomes and their respective counts from the dataset, applying a filter with the WHERE clause to include only those occurring between June 4, 2010, and March 20, 2010. We then used the GROUP BY clause to group the landing outcomes and the ORDER BY clause to arrange them in descending order based on their counts.

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

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

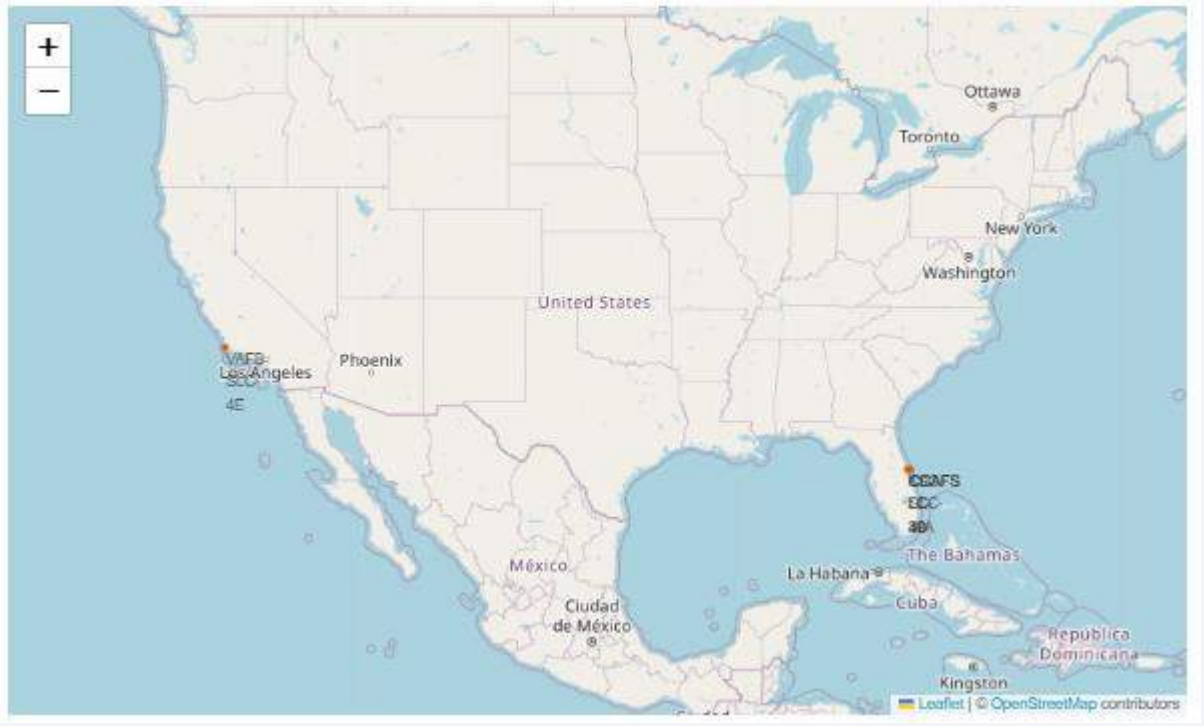
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

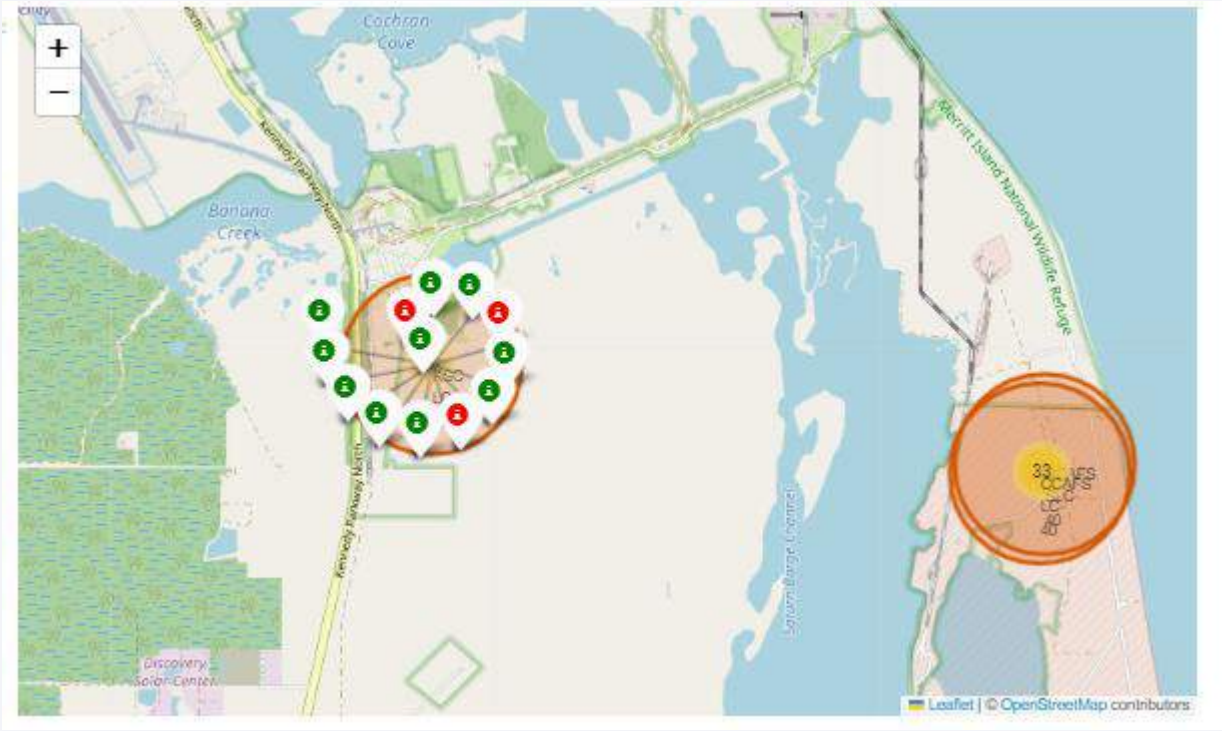
Launch Sites

- The majority of launch sites are located near the Equator, where the Earth's rotation provides a significant speed advantage. Objects at the Equator are already moving at approximately 1670 km/hr due to the Earth's rotation, and launching from there allows a spacecraft to inherit this velocity, aiding in achieving and maintaining orbit.
- Additionally, launching over the ocean minimizes the risk of debris falling in populated areas, making coastal locations preferable for launch sites.



Outcome of launches with markets

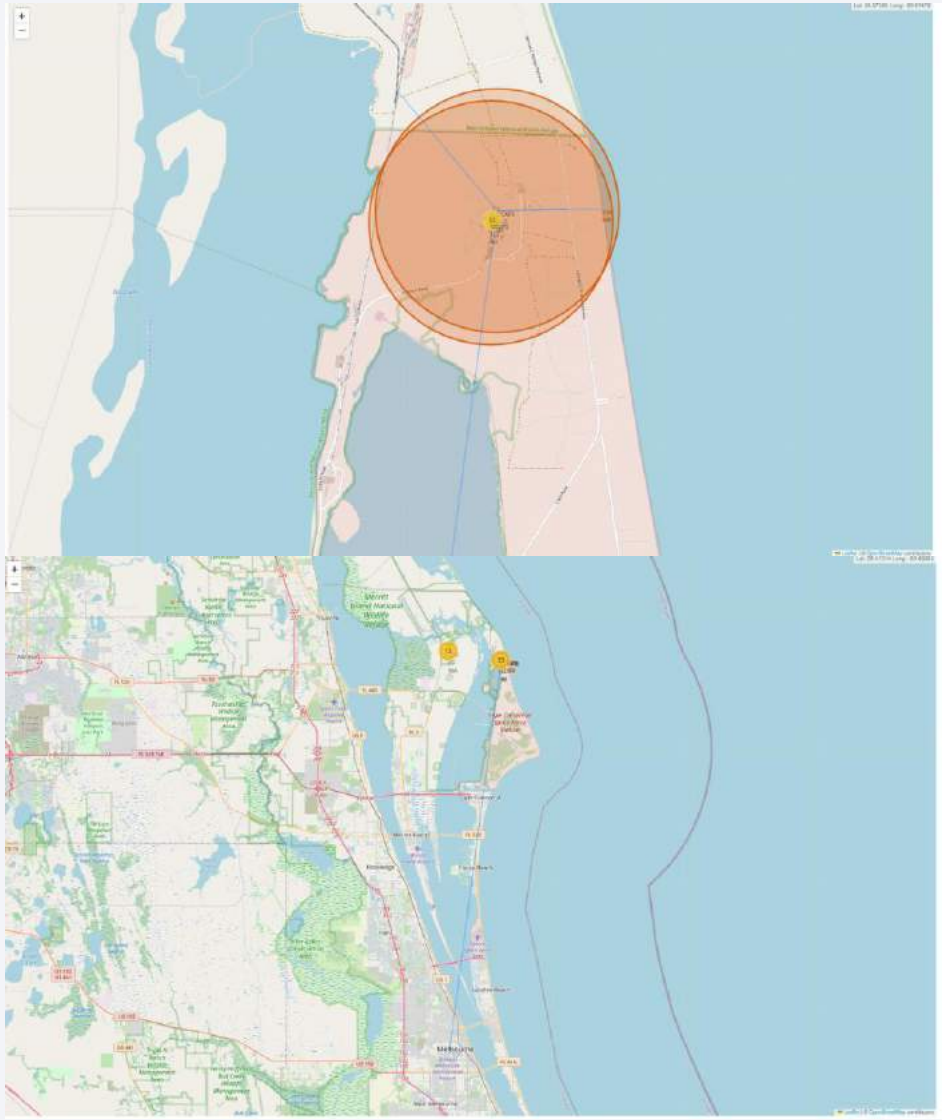
- We are be able to identify the launch sites with high success rates by using visual markers.
- Green markers can indicate successful launches, while red markers can indicate failed launches.
- Launch site KSC LC-39A has a notably high success rate.



Distances between a launch site to its proximities

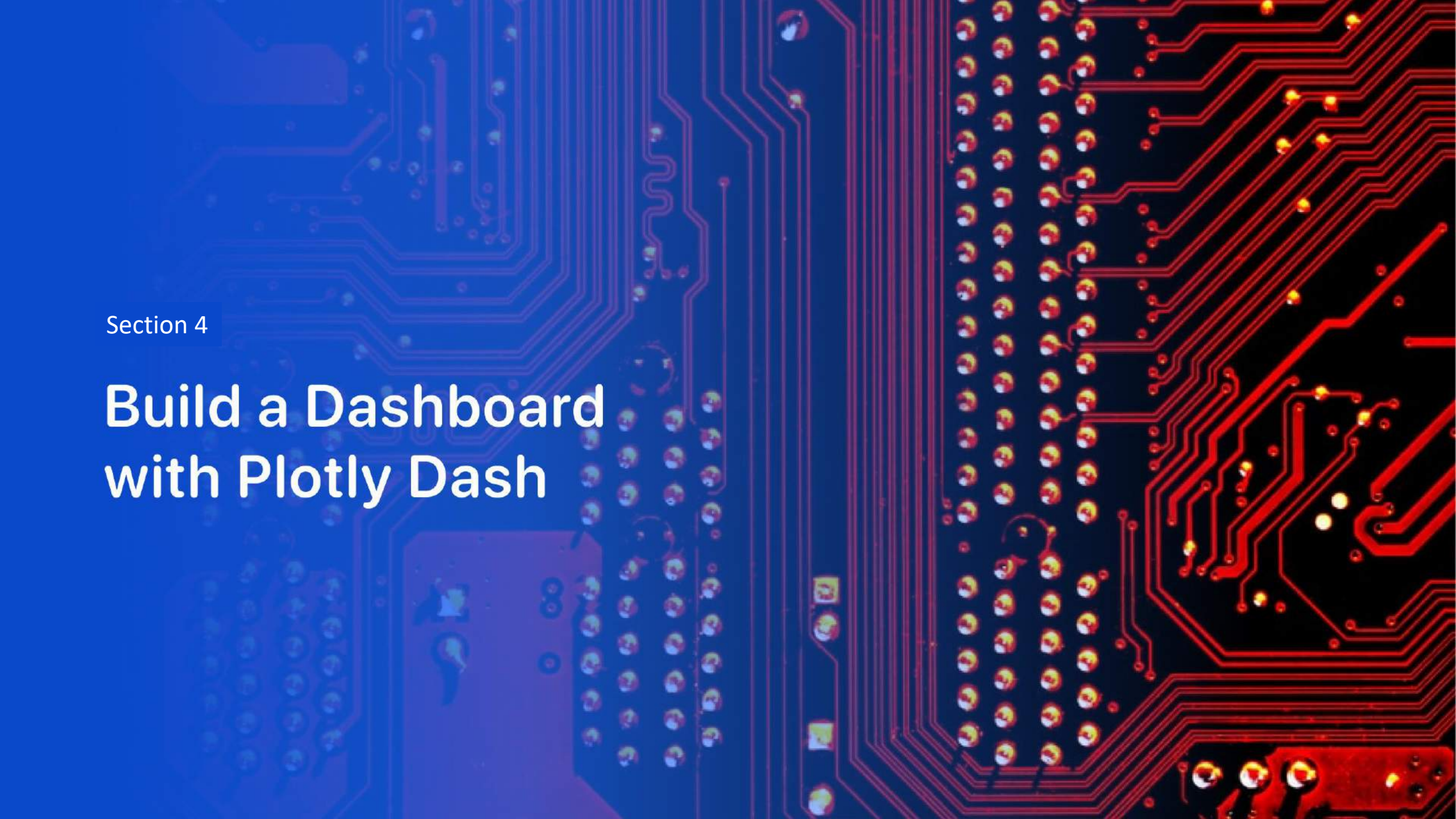
- We can deduct that launchsites need to be away from any urban location and any essential infrastructure as railways and highways are. Maintaining relative proximity to facilitate transportaition from materail and staff.

distance_city
51.43416999517233
distance_railroad
1.2845344718142522
distance_highway
0.5834695366934144



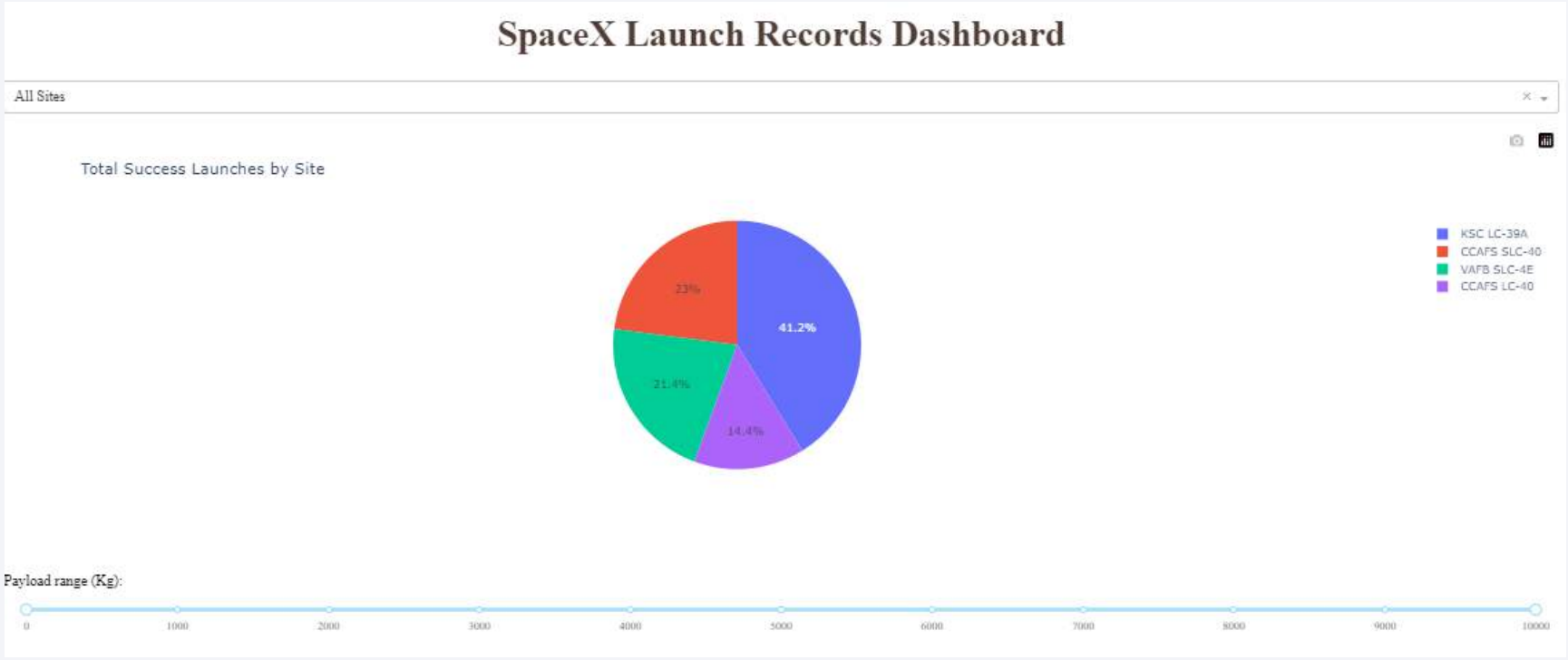
Section 4

Build a Dashboard with Plotly Dash



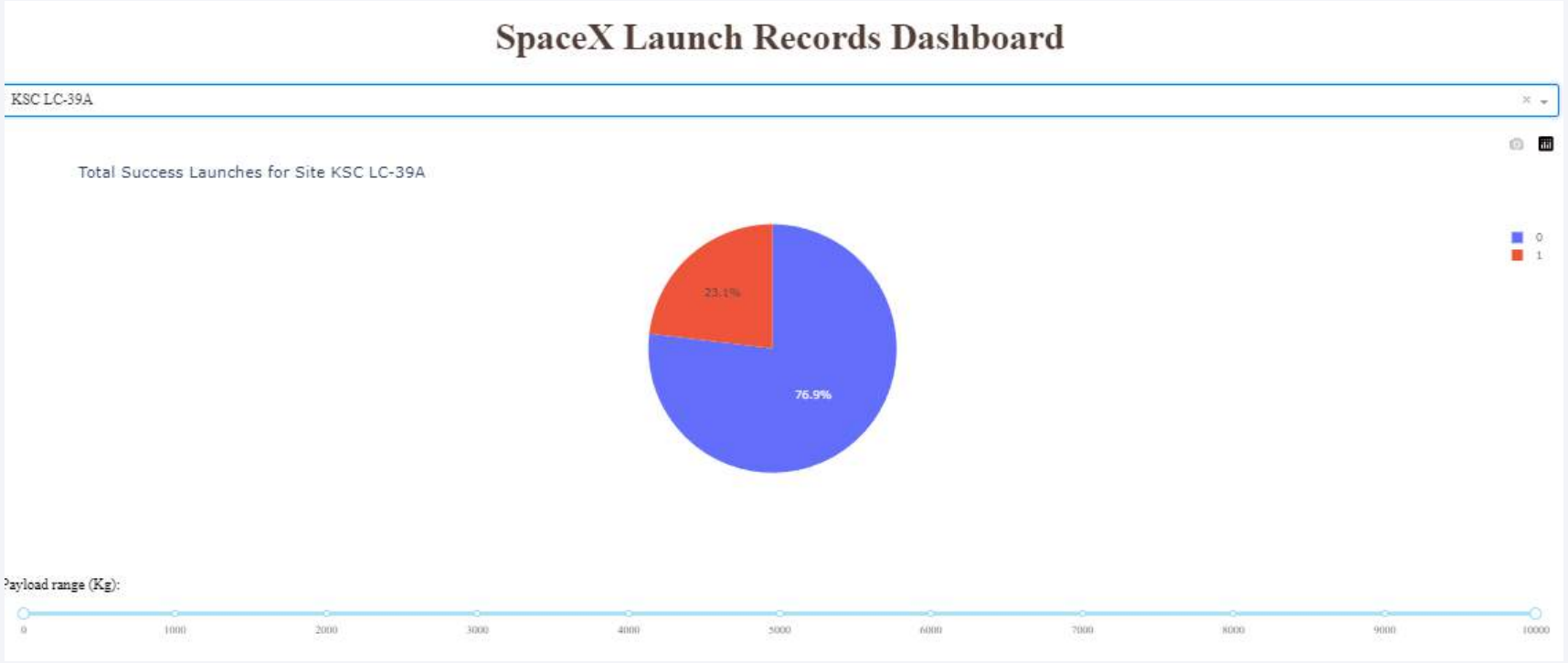
Launch Success

- KSC LC-39A has a 41.2% of successful launches among all sites.



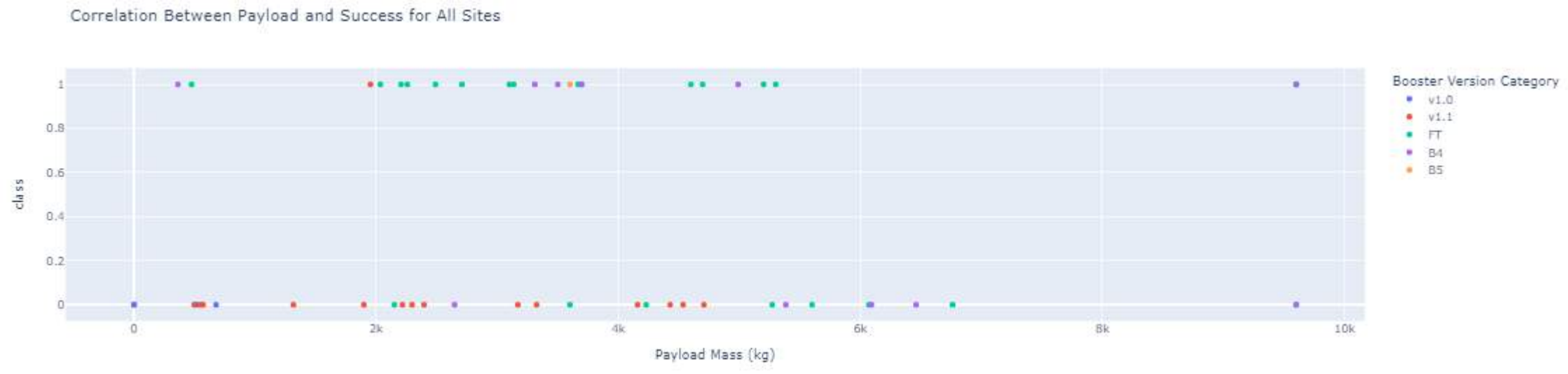
Launch Success KSC LC-39A

- KSC LC-39A has a 76.9% launch success rate with 10 successful and only 3 failed landing, making it the launch site with the highest launch success rate



Payload Mass vs Launch Outcome

- Payloads between 2,000 kg and 5,000 kg have the highest success rate.
- 1 indicating successful outcome and 0 indicating an unsuccessful outcome.



Section 5

Predictive Analysis (Classification)

Classification Accuracy

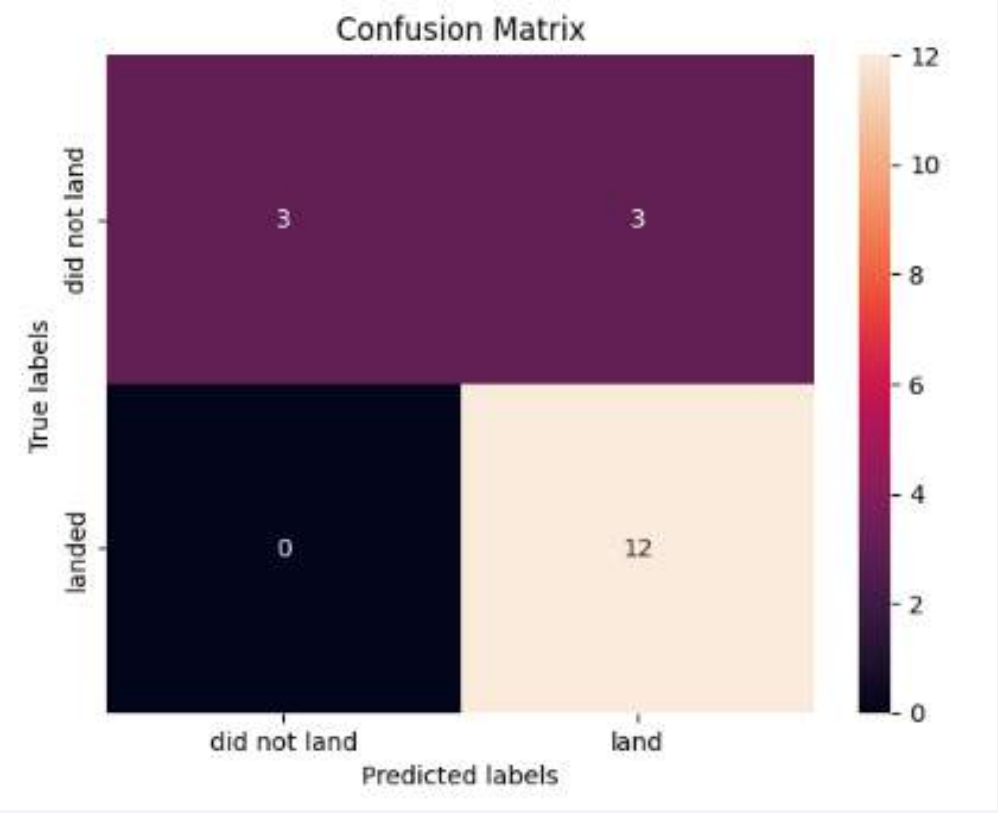
- All the models performed similarly and showed nearly identical scores and accuracy levels. This similarity in performance is likely attributed to the small size of the dataset.
- However, when considering the .best_score_ , the Decision Tree model marginally outperformed the others.

```
algorithms = {'KNN':knn_cv.best_score_,'Tree':tree_cv.best_score_,'LogisticRegression':logreg_cv.best_score_}
bestalgorithm = max(algorithms, key=algorithms.get)
print('Best Algorithm is',bestalgorithm,'with a score of',algorithms[bestalgorithm])
if bestalgorithm == 'Tree':
    print('Best Params is :',tree_cv.best_params_)
if bestalgorithm == 'KNN':
    print('Best Params is :',knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best Params is :',logreg_cv.best_params_)
```

```
Best Algorithm is Tree with a score of 0.875
Best Params is : {'criterion': 'gini', 'max_depth': 16, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5, 'splitter': 'best'}
```

Confusion Matrix

- Upon examining the confusion matrix, we observe that logistic regression is capable of distinguishing between different classes. However, a notable issue is the occurrence of false positives.



Conclusions

Based on our analysis, we can conclude the following:

- The Sun-Synchronous Orbit (SSO) has the highest success rate of 100% and has occurred more than once, indicating a high level of reliability for launches into this orbit.
- Lighter payloads (defined as 4000kg and below) exhibit better performance compared to heavier payloads.
- Starting from 2013, the success rate for SpaceX launches has shown a consistent increase, indicating a positive trend over time. This trend suggests a potential for further improvement in the future.
- KSC LC-39A has the highest success rate among all launch sites, at 76.9%.
- The Tree Classifier Algorithm is the most suitable machine learning approach for this dataset.

Thank you!