IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Abhinay Vardhan
August 2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA with Data Visualization using Matplotlib
  - EDA with SQL Querying
  - Interactive Map with Folium
  - Dashboarding with Plotly Dash
  - Predictive Analysis using Scikit-learn (Classification Problem)
- Summary of all results
  - EDA results
  - Analytics (Dashboard/Folium Markers/SQL results)
  - Predictive Analytics results – Comparing accuracy from different ML Models

Introduction

- Project background and context
  - SpaceX is the most successful company of the commercial space, targeting to make space travel affordable.
  - The company advertises Falcon 9 rocket launches on its website, with a cost of 69.75 million dollars per launch which is significantly cheaper as compared to other providers sur to the reusability of launchers.
  - Hence, if we can determine which factors favors the successful landing of the first stage launchers, we can significantly reduce failures which in turn will add to the savings of each launch. Based on public information and machine learning models, we are going to predict if SpaceX will be able to reuse its launcher.
- Answers to be found:
  - Which factors determine successful landing: Launch sites, payload, boosters, orbits etc.?
  - Has the reusability improved over the years, if yes which payloads, boosters and launch sites have been used for recent launches?
  - Best algorithm to use for such a machine learning problem

Section 1

Methodology

Methodology

Executive Summary

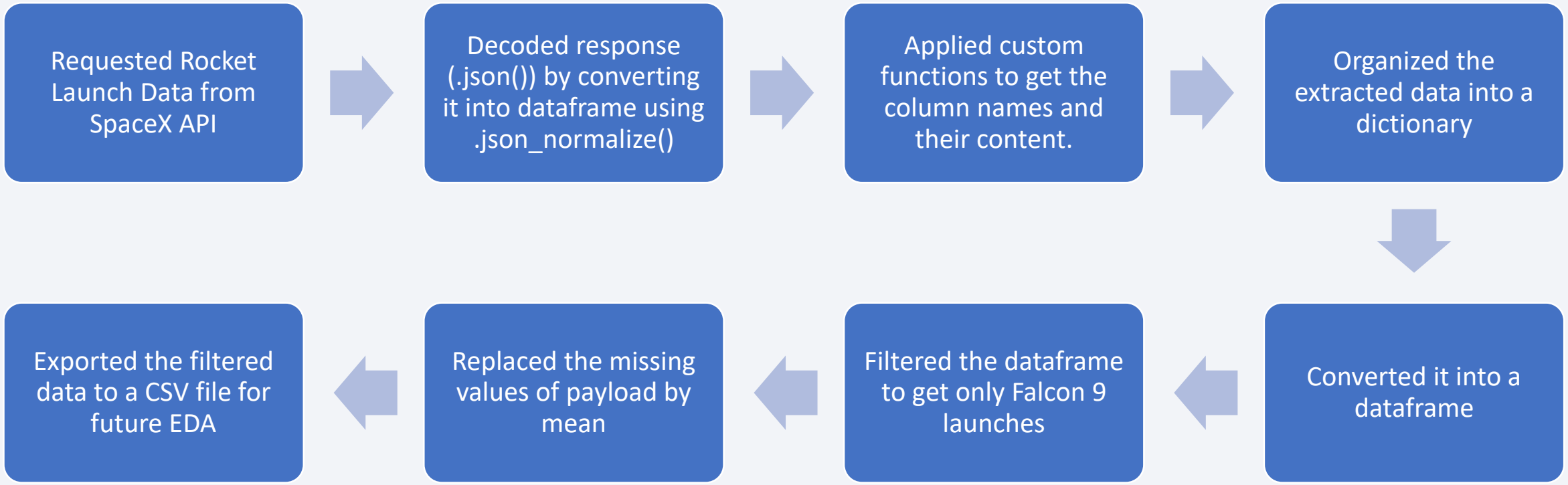
- Data collection methodology:
  - Using SpaceX Rest API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - Filtering Data
  - Dealing with Missing Values
  - Feature Classification using One Hot Encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models to get the best possible outcome

Data Collection

Data collection process involved API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

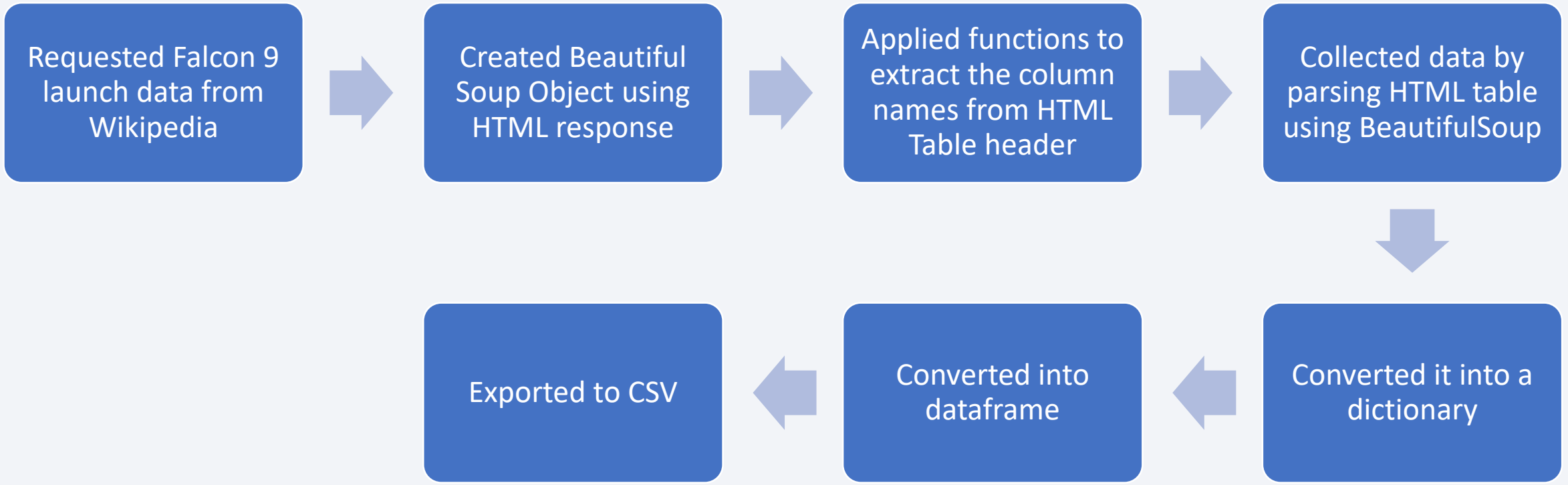
- We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.
- Later on, the data obtained from websites were parsed using BeautifulSoup and column names as well as the column content was obtained.
- Data Columns are obtained by using SpaceX REST API:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Data Columns are obtained by using Wikipedia Web Scraping:
  - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



Github: Data Collection using API

Data Collection - Scraping



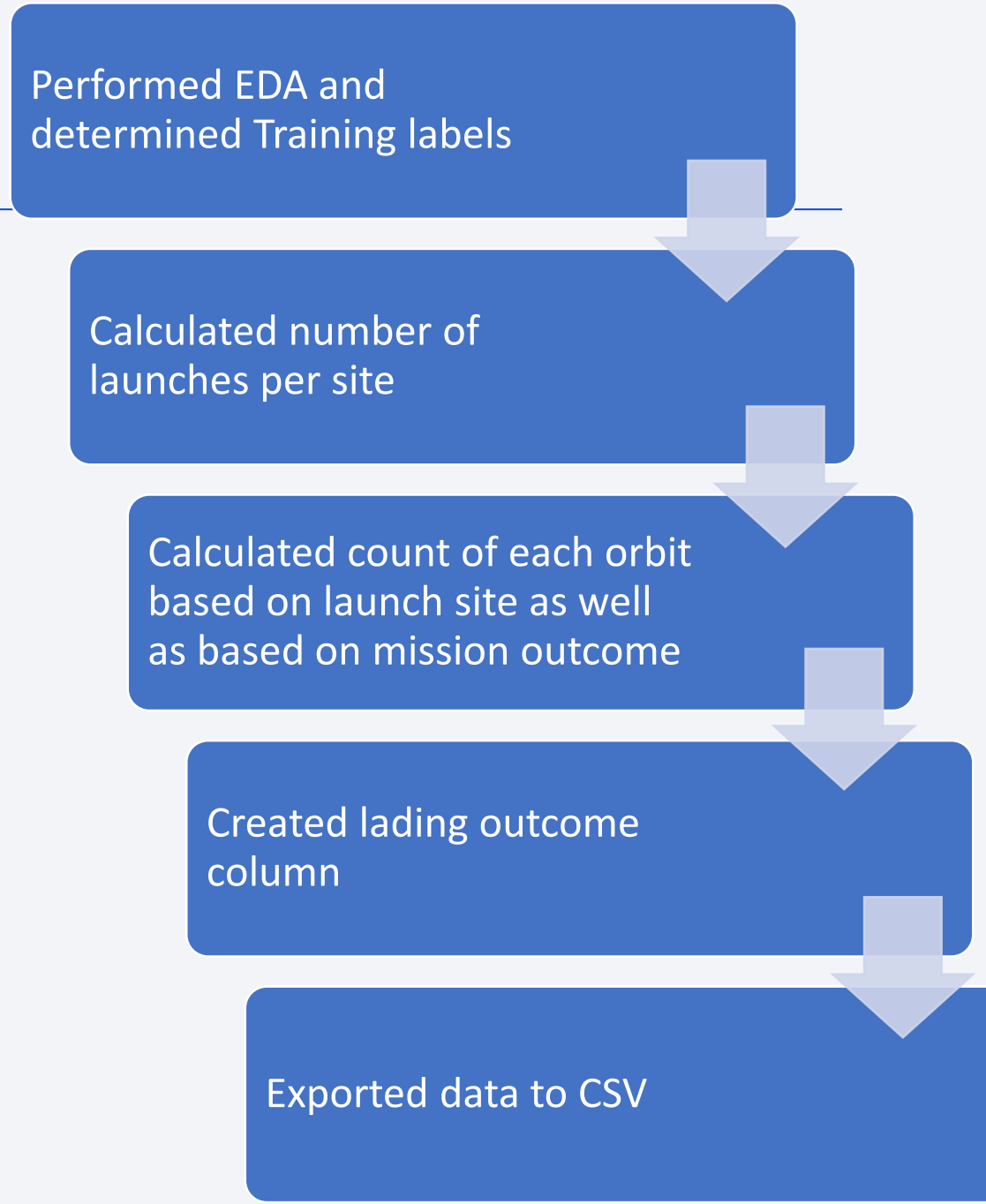
Github : Web Scraping

Data Wrangling

Feature Extraction:

- In the data set, there are several different cases when the booster did not land successfully.
- Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed on a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean.
- True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.
- True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.
- We mainly convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful.

Github: Data Wrangling



EDA with Data Visualization

- Charts plotted:
  - Scatter plot of Flight Number vs. Payload Mass
  - Scatter plot of Flight Number vs. Launch Site
  - Scatter plot of Payload Mass vs. Launch Site
  - Bar plot Orbit Type vs. Success Rate
  - Scatter Plot of Flight Number vs. Orbit Type
  - Scatter plot of Payload Mass vs Orbit Type and
  - Bar plot of Success Rate Yearly Trend

Github: EDA with Data Visualization

EDA with SQL

List of SQL Queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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

Github: EDA with SQL

Build an Interactive Map with Folium

- Using markers and circles identified the launch sites on the global map
- Added total launch instances to these sites
- Added occurrence of success/failures on each site
- Calculated the minimum distances from railways, coastlines, highways and city centers
- This is to determine what things to keep in mind while considering an optimum location for launch sites

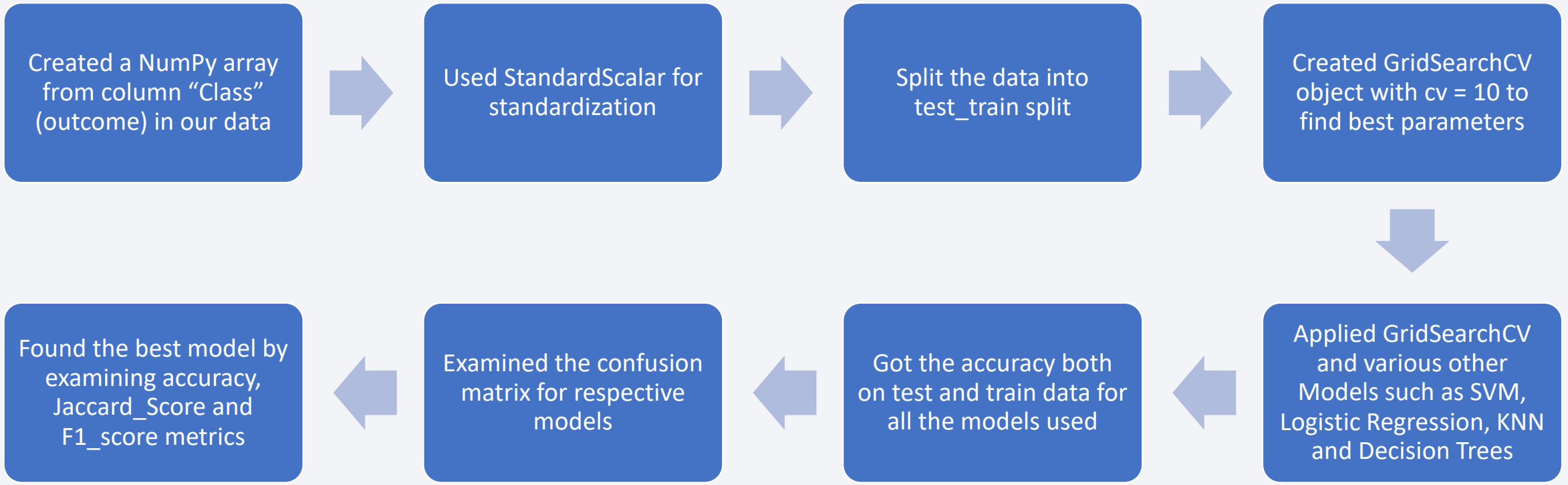
Github: Interactive Map with Folium

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List:
  - Added a dropdown list to enable Launch Site selection.
- Pie Chart showing Success Launches (All Sites/Certain Site):
  - Added a pie chart to show the total successful launches count for all sites and the
- Success vs. Failed counts for the site, if a specific Launch Site was selected.
  - Slider of Payload Mass Range
  - Added a slider to select Payload range.
- Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:
  - Added a scatter chart to show the correlation between Payload and Launch Success

Github: Plotly Dash

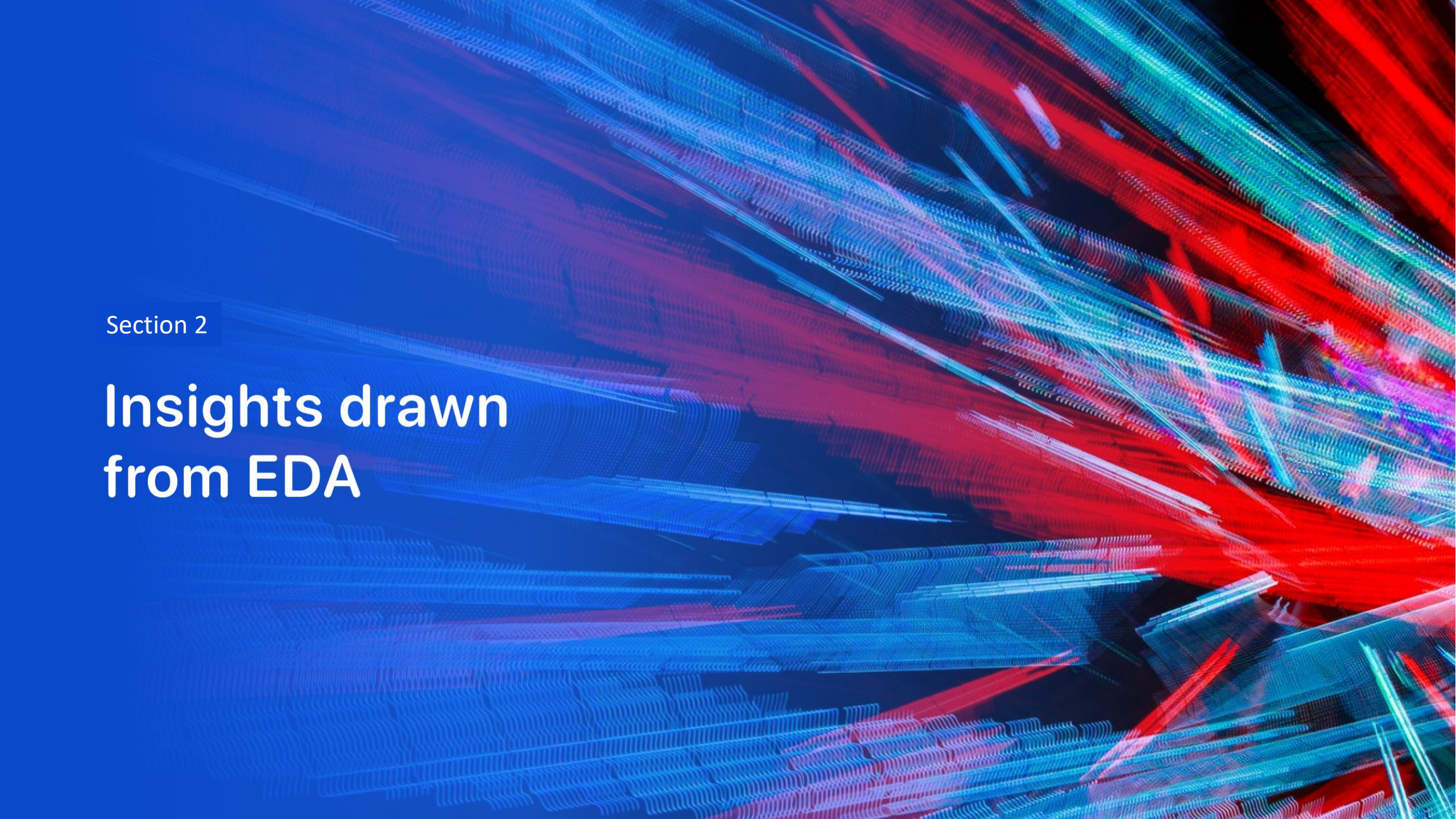
Predictive Analysis (Classification)



Github: ML Predictive Analysis

Results

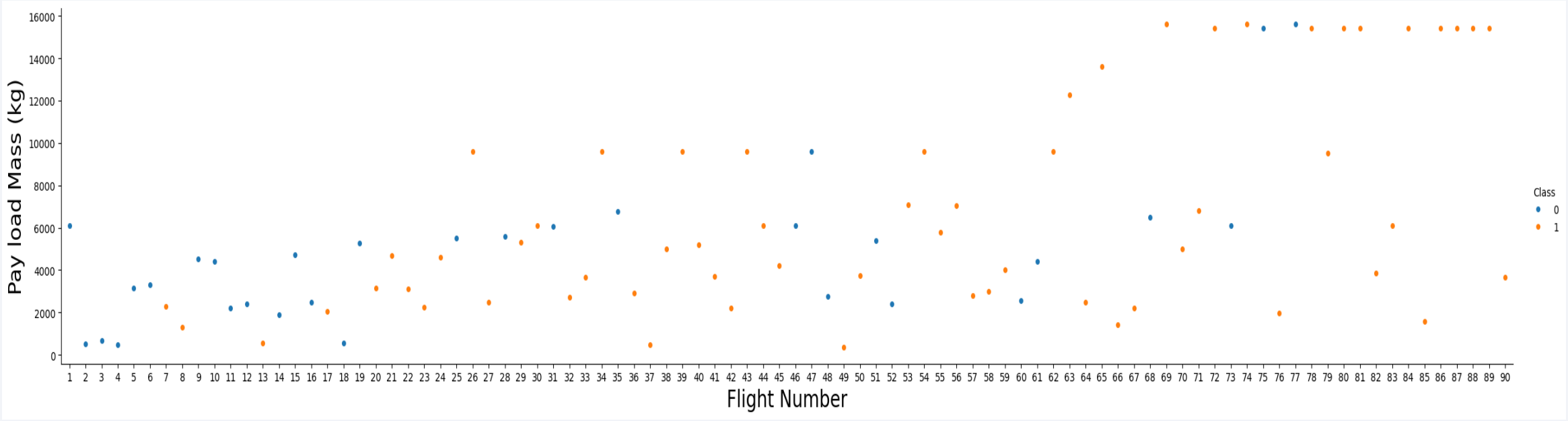
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

Section 2A **VISUALIZATIONS**

Insights drawn from EDA

Flight Number vs. Payload Mass (Kg)



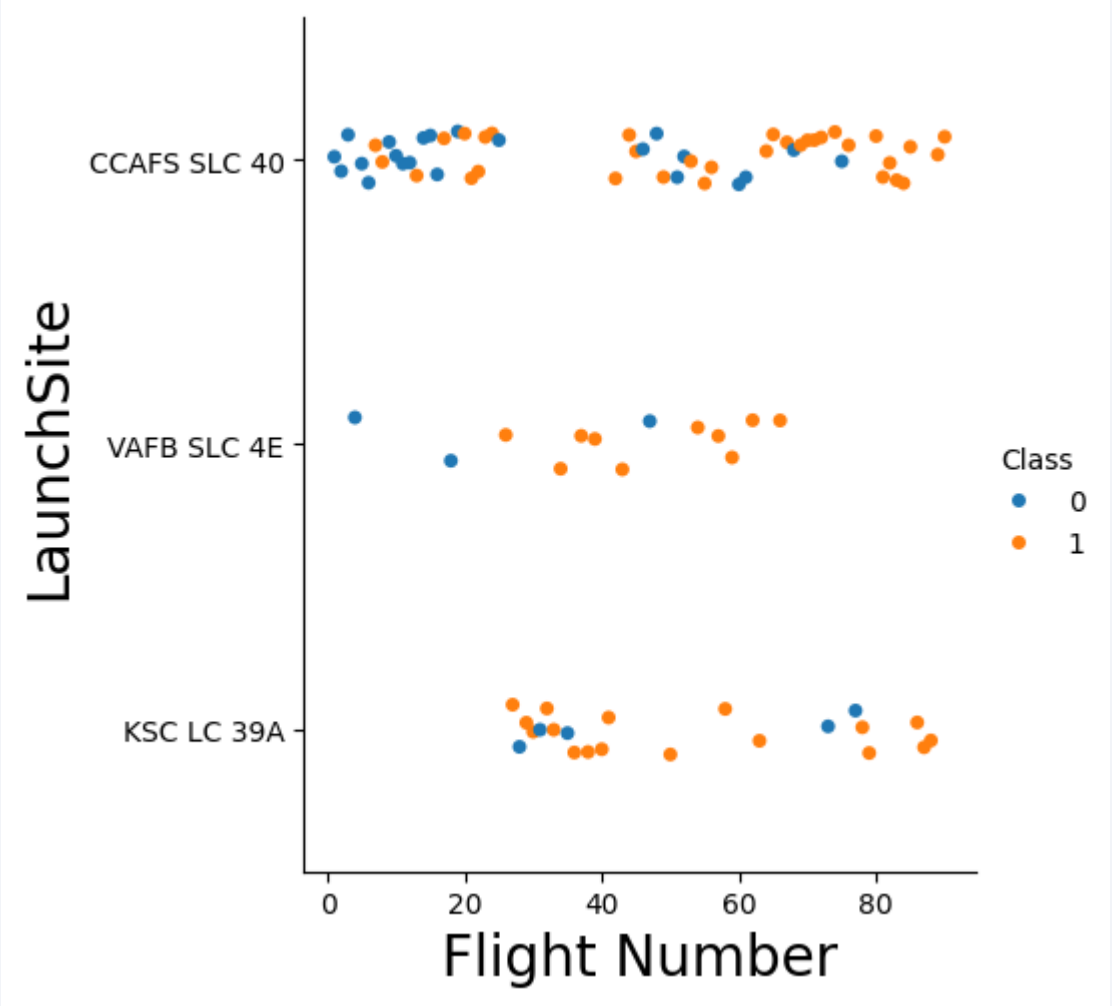
Scatter Plot of Flight Number vs Launch Sites:

- Early flights had lower payload (Most occurred failures)
- Most recent flights had higher payload (Most successful)

Flight Number vs. Launch Site

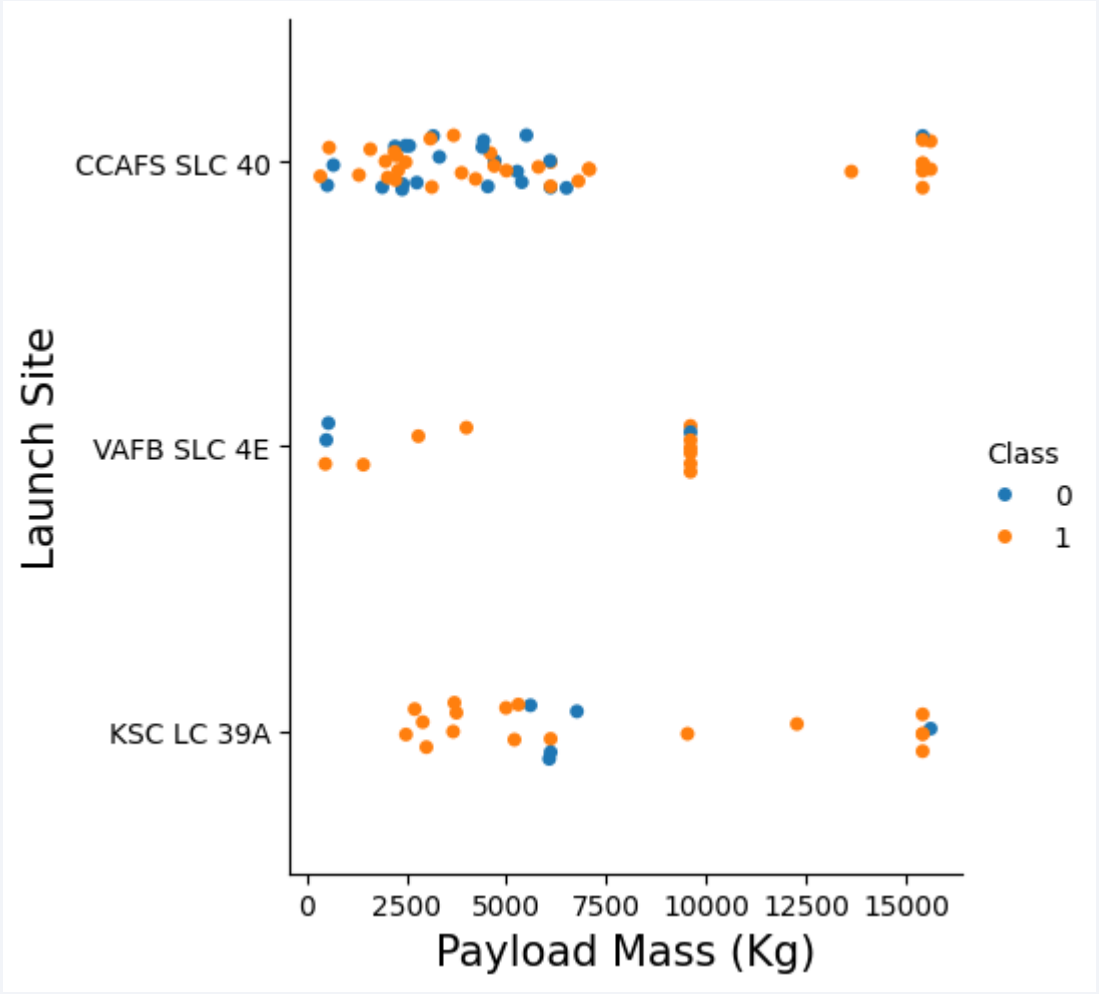
Scatter Plot of Flight Number vs Launch Sites:

- Early fights were from CCAFS SLC 40 or VAFB SLC 4E, most of them unsuccessful.
- Maximum flights launched from CCAFS SLC 40, recent flights all successful
- KSC LC 39A, maximum successful flights
- Recently, most sites are launched either from CCAFS SCL 40 or KSC LC 39A.
- Recent launches have higher success rate



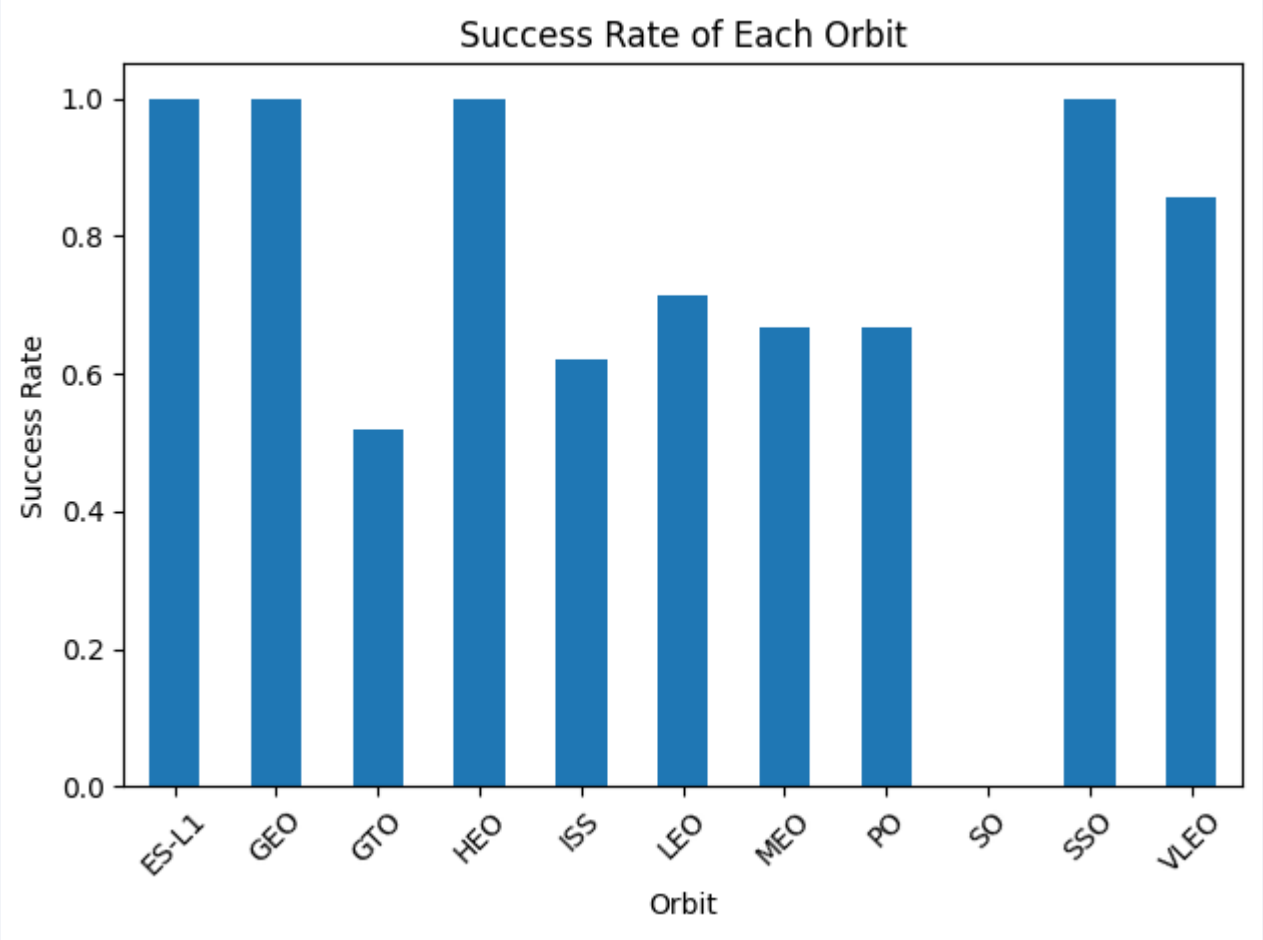
Payload vs. Launch Site

- If we combine insights from both flight number and payload mass as well as launch sites:
- Earlier flights with low payload, had high failure from all the sites.
- Once the technology improved, higher payloads were tried in later flights, observation: On each site higher the payload, higher success rate
- Later flights have high success rate, even with higher payload. Only a few instances of failure observed



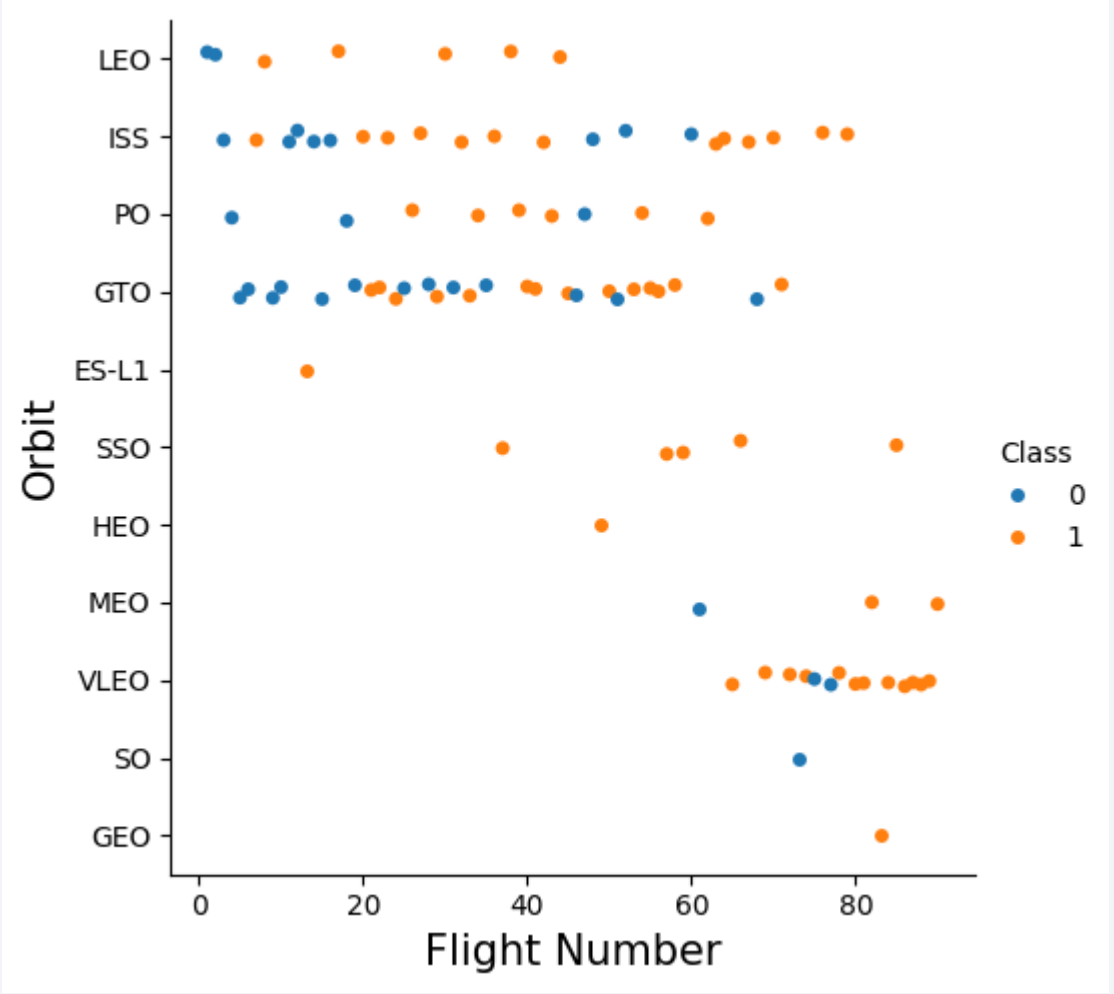
Success Rate vs. Orbit Type

- Highest success rate - 100% for : ES-L1, GEO, HEO, SSO
- Average success rate (50% - 70%) for GTO, ISS, LEO, MEO, PO
- Lowest Success rate for SO (0%)



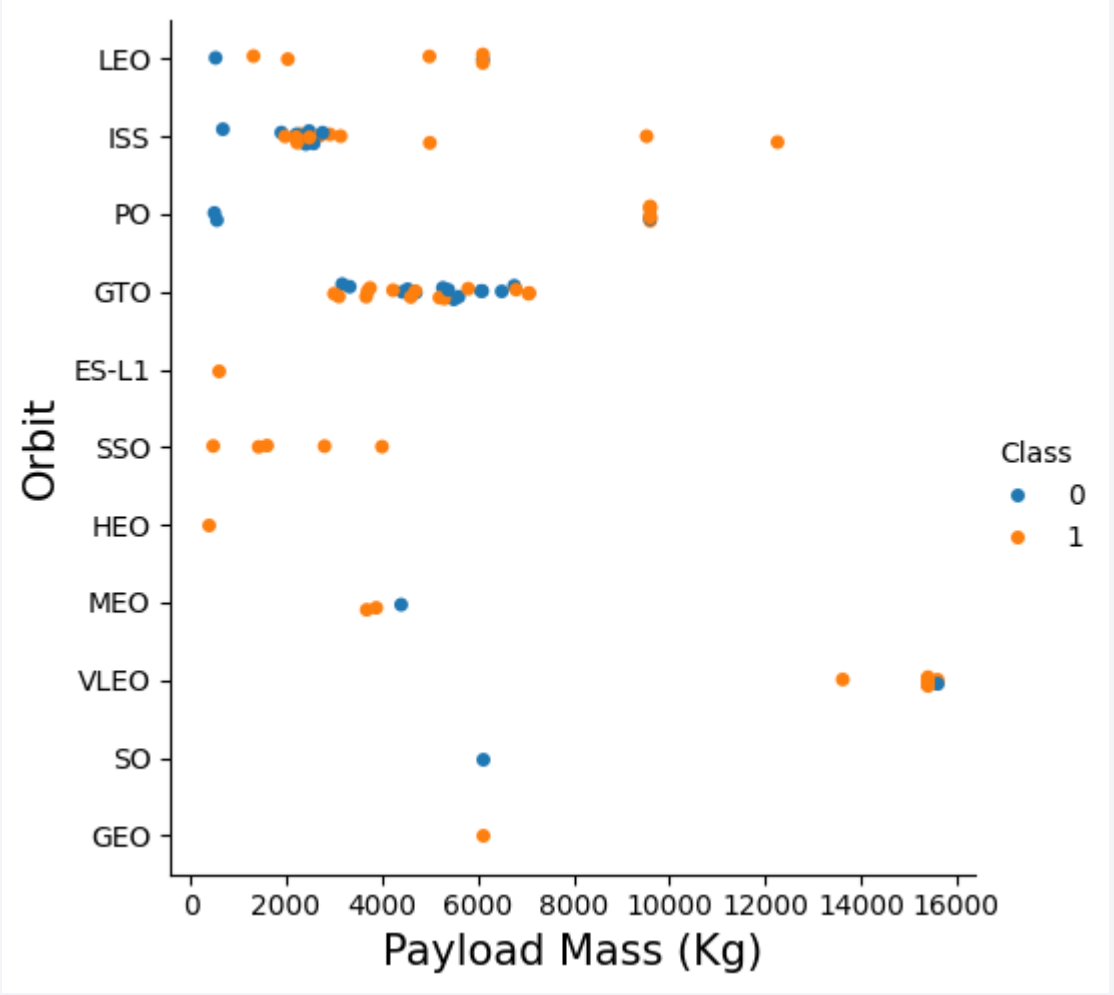
Flight Number vs. Orbit Type

- Early launches targeted LEO, ISS, PO, GTO orbits
- Later launches targeted SSO, MEOM, VLEO, SO and GEO orbits,
- Later flights have all been mostly successful especially ones targeting VLEO
- SSO orbit launches have never faced a failure



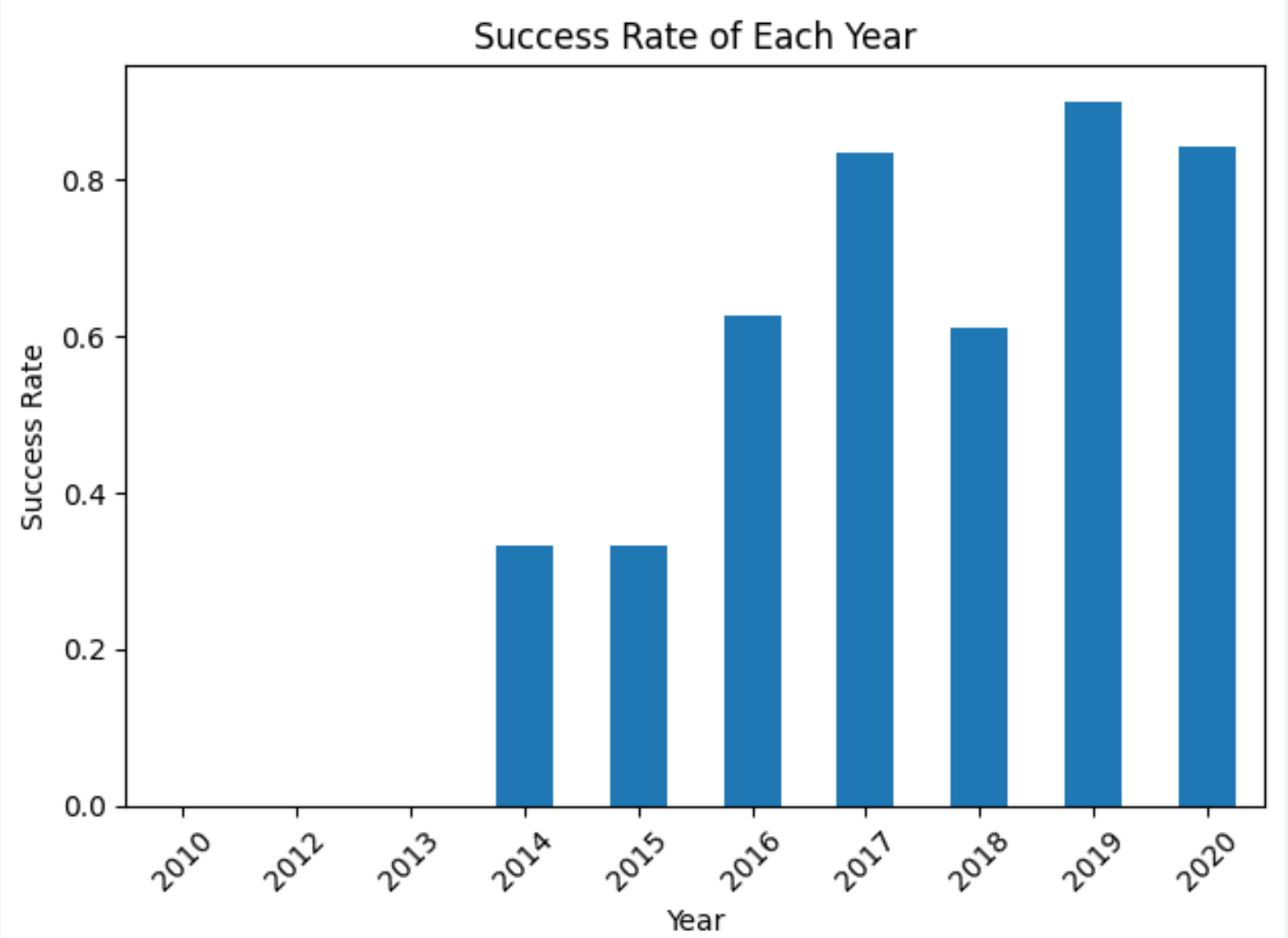
Payload vs. Orbit Type

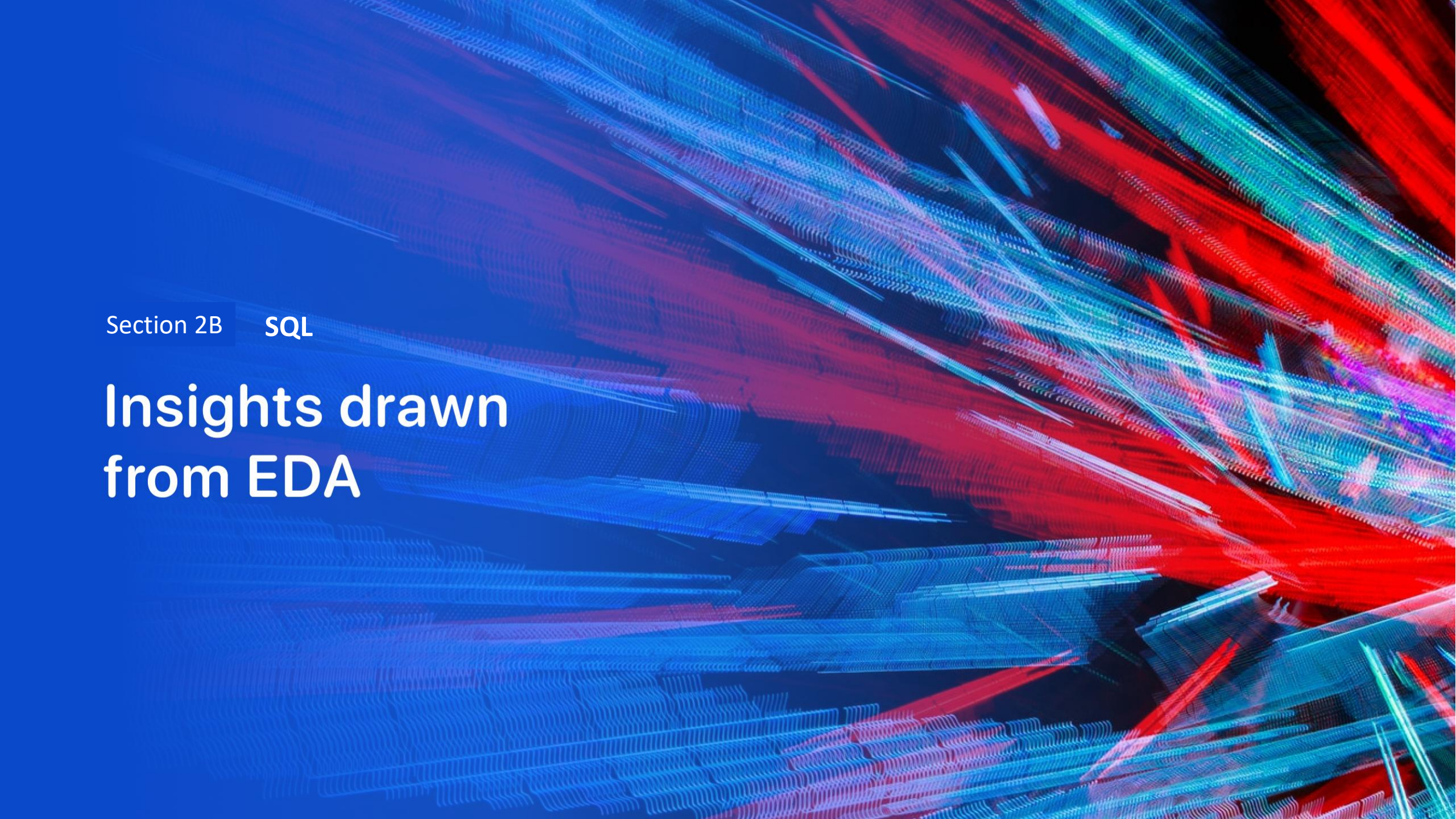
- Heavy payloads have been successful for ISS, PO and VLEO orbits
- GTO seems to have no correlation with payload
- Most of the payload for SSO have been lighter and successful.



Launch Success Yearly Trend

- Success rate improved gradually from 2014 onwards.




Section 2B
SQL
Insights drawn
from EDA

All Launch Site Names

- Four distinct Launch sites:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

```
%%sql
SELECT DISTINCT Launch_Site FROM SPACEXTBL LIMIT 5;
```

[9]

```
 * 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'

```
[10]
%%sql
SELECT Launch_Site FROM SPACEXTBL WHERE Launch_Site like 'CCA%' LIMIT 5;
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Find 5 records where launch sites begin with `CCA`

Total Payload Mass

```
[11]
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';

 * sqlite:///my_data1.db
Done.

SUM(PAYLOAD_MASS__KG_)
                 45596
```

- NASA boosters are named as ‘NASA (CRS)’, aggregating over the sum of payloads,
- Total mass of payloads = 45,596Kg

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version like 'F9 v1.1';
```

[13]

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

AVG(PAYLOAD_MASS__KG_)
2928.4

- Average payload mass carried by booster version F9 v1.1 by filtering using 'LIKE' keyword in SQL
- Average Payload Mass = 2928.4 Kg

First Successful Ground Landing Date

```
[14]
%%sql
SELECT MIN(DATE) FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)'
```

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

MIN(DATE)
2015-12-22

- Queried the minimum date of launch when landing outcome was successful on ground pad
- Date = 22nd Dec, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

```
[15]
%%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

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- 4 booster versions found

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT Mission_Outcome, COUNT(*) FROM SPACEXTBL GROUP BY Mission_Outcome;
```

[16]

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

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

- Most mission outcomes were successful except one failure (in flight)

Boosters Carried Maximum Payload

```
%%sql
SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

[17]

```
 * 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

- Queried booster versions using subquery where payload was equal to the maximum payload
- All of them are F9 B5 versions

2015 Launch Records

```
%%sql
SELECT SUBSTR(Date,6,2) AS month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)' AND SUBSTR(Date,0,5)='2015';
```

[20]

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

month	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

- Listed the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- One failure from the month of January, another from April. Both from launch sites CCAFS LC – 40, Both versions F9 v1.1

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

```
%%sql
SELECT
Landing_Outcome,
COUNT(Landing_Outcome) AS Total_Count
FROM SPACEXTBL
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Total_Count DESC;
```

[22]

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

Landing_Outcome	Total_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

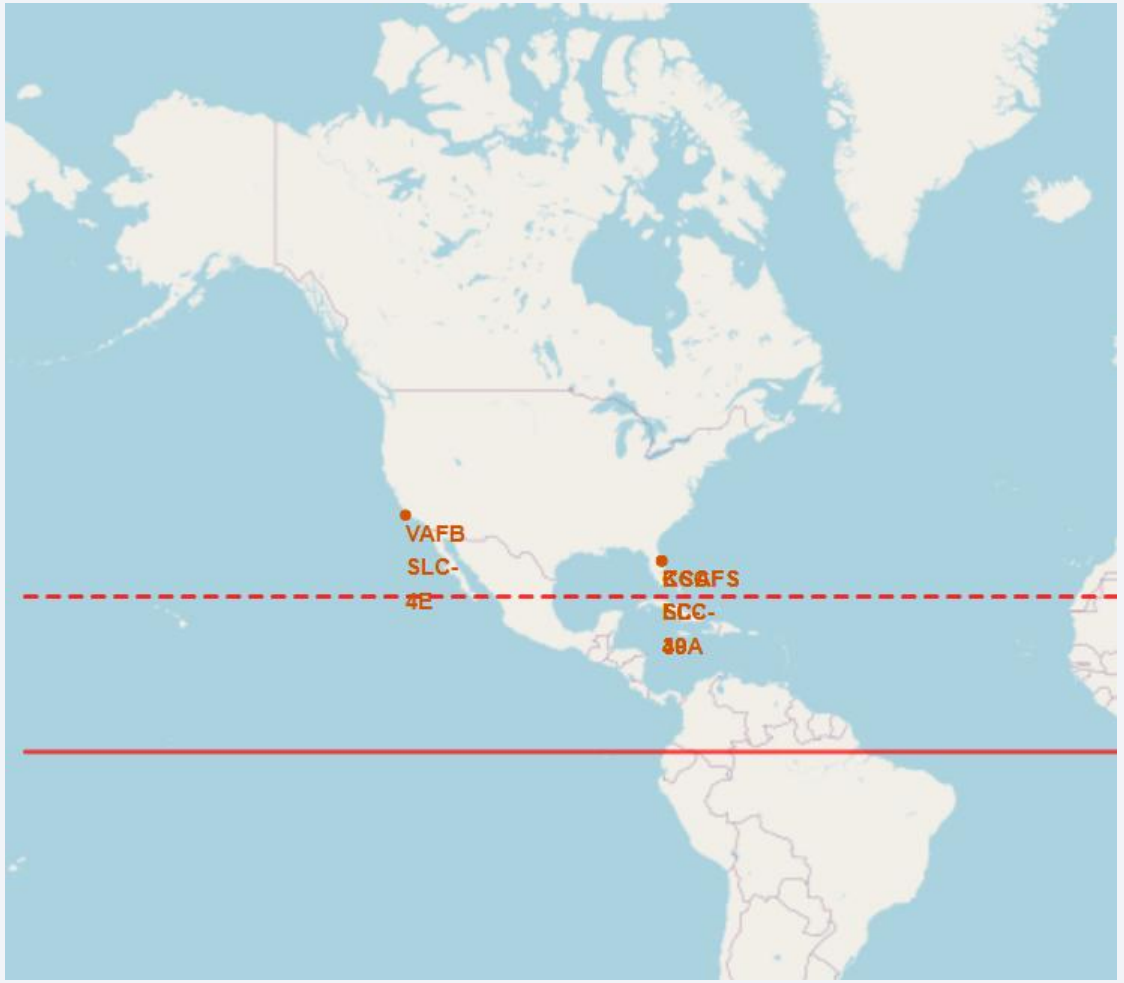
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- Maximum failures and successes were on a drone ship (5)

Section 3

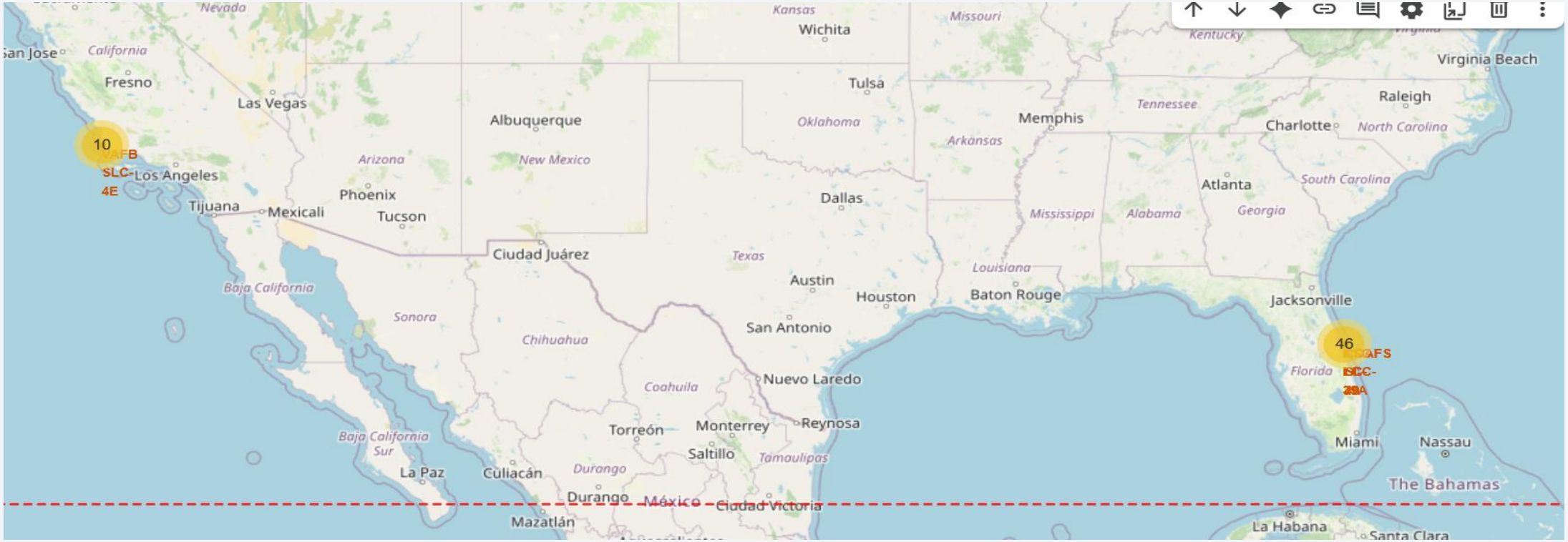
Launch Sites Proximities Analysis

Launch Sites Locations

- Most Launch sited are closer to the Equator (Equator-solid red line doesn't pass through US)
- They have selected the closest possible locations within US which are closer to the equator
- This is because the ESCAPE VELOCITY (velocity needed to leave the earth's orbit) required is slightly lower at equator as compared to poles
- The Earth's rotation provides a "free" speed boost at the equator, especially when launching eastward, which helps reduce the total velocity needed to escape its gravity

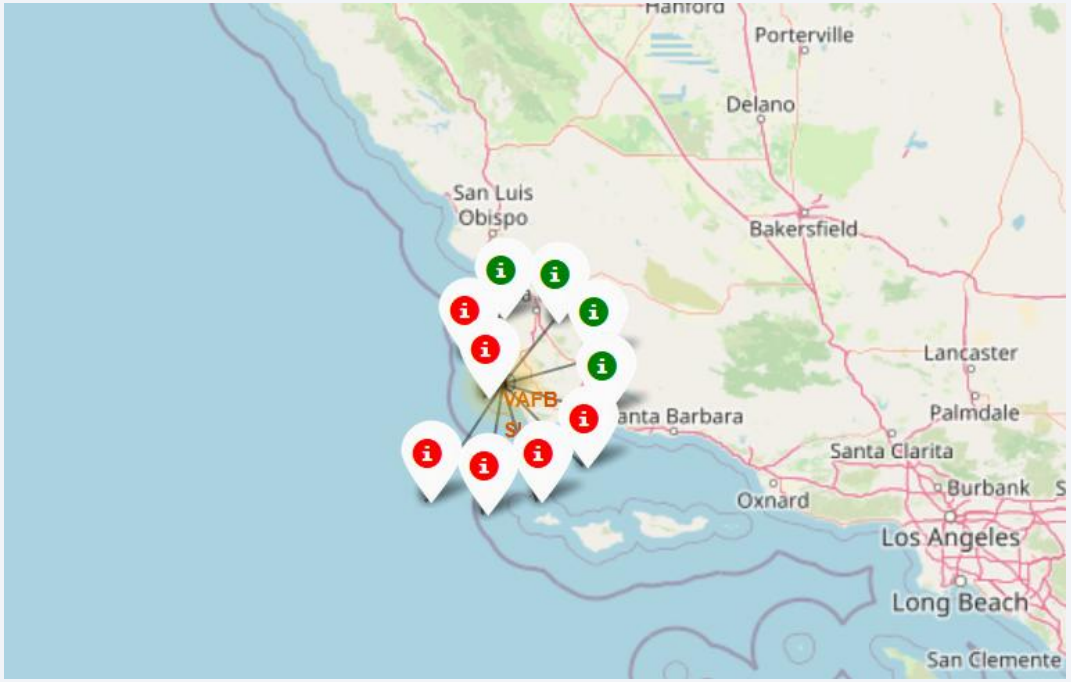


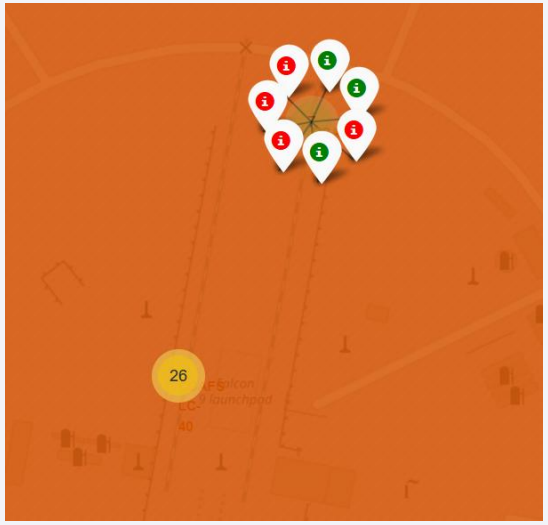
Distribution of Launch Sites

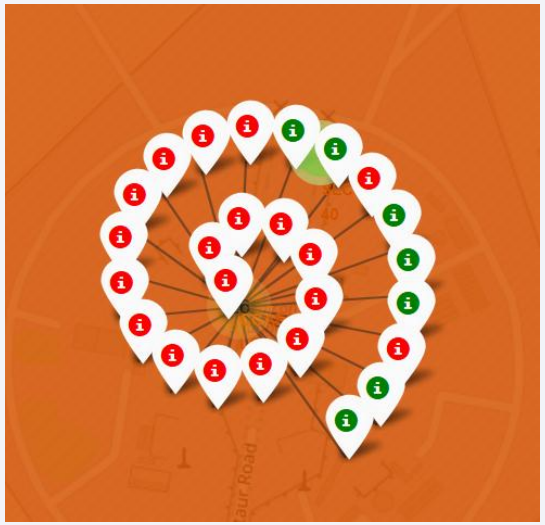


- Launch sites are either situated near the East/West Coasts
- They are also selected strategically to be away from densely populated cities
- Launch Sites on East Coast: 10
- Launch Sites on West Coast: 46

Color labeled Success/Failure instances on Map



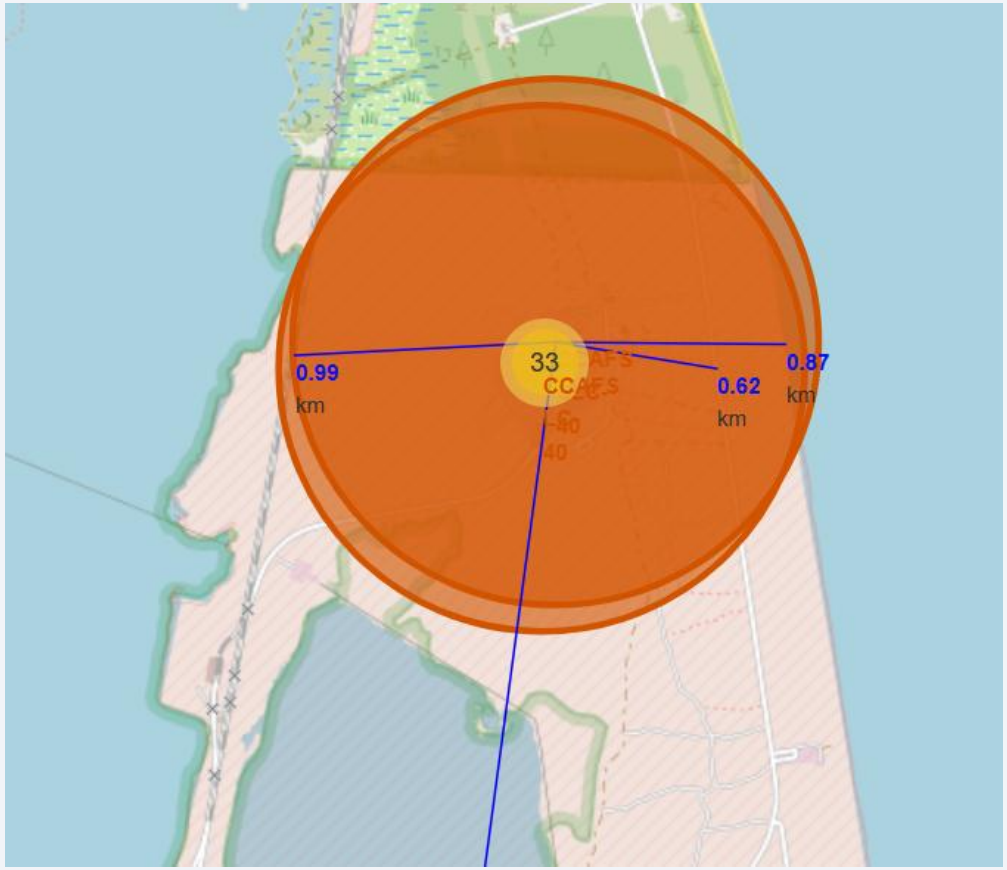


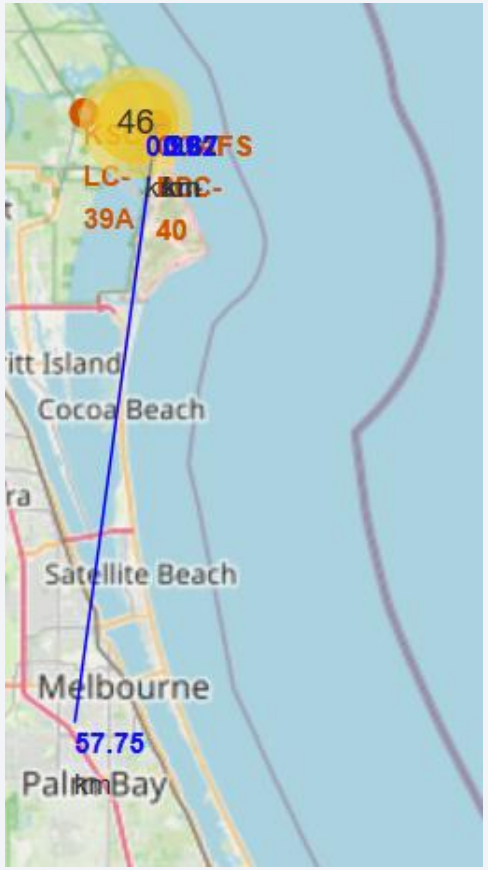


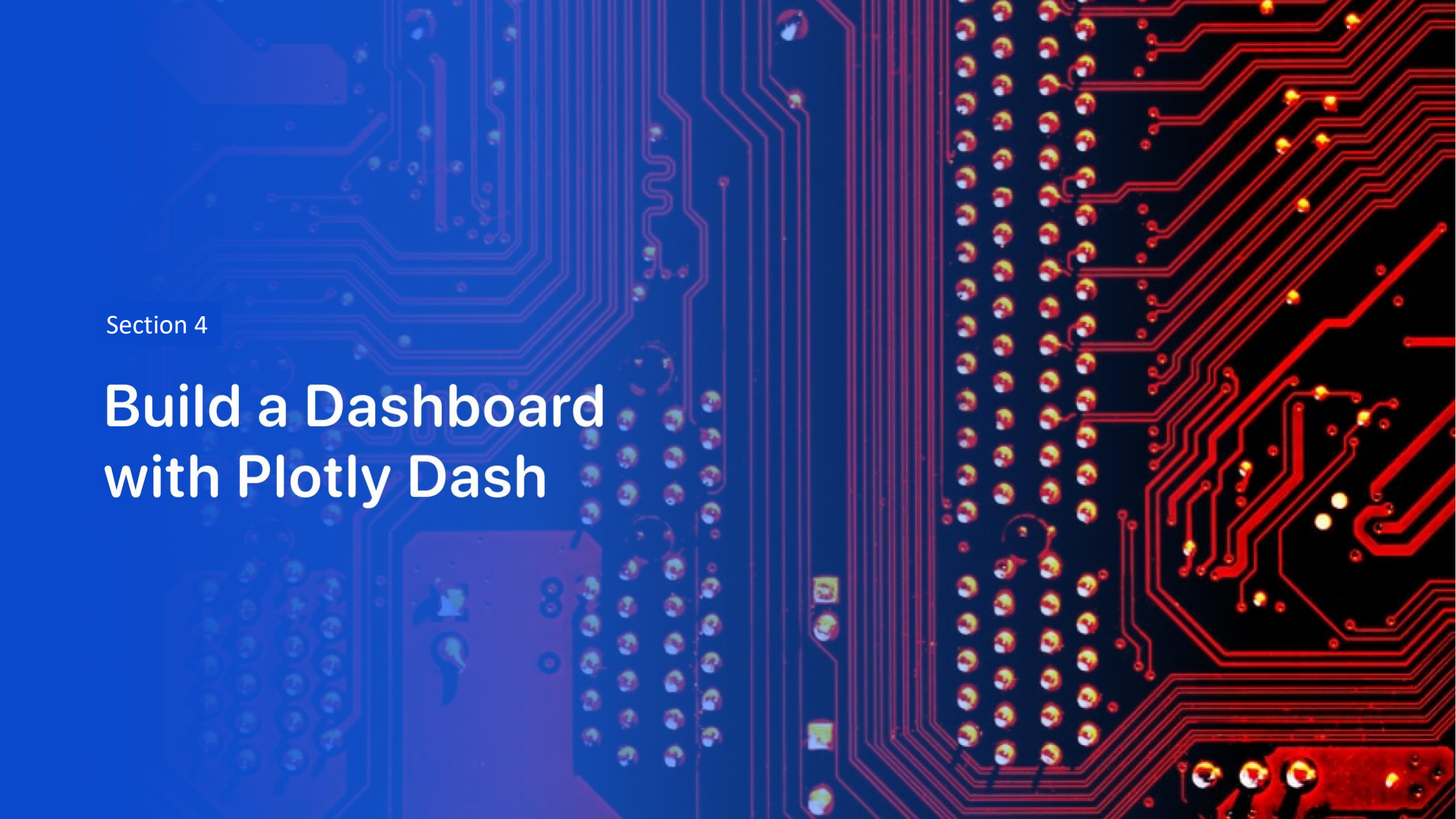
- Color Coding: Green(Success), Red (Failure)
- Total launches near West Coast: 10 (Success: 4; Failures:6), Success Rate: 40%
- Total Launches near East Coast: 46 (Success: 10, Failures: 36), Success Rate: 22%
- Since the sample size has a huge difference (10 vs 46), we cannot say for certainty that the west coast has better success rate. If the number of instances were comparable, this inference could definitely be explored in depth.

<Folium Map Screenshot 3>

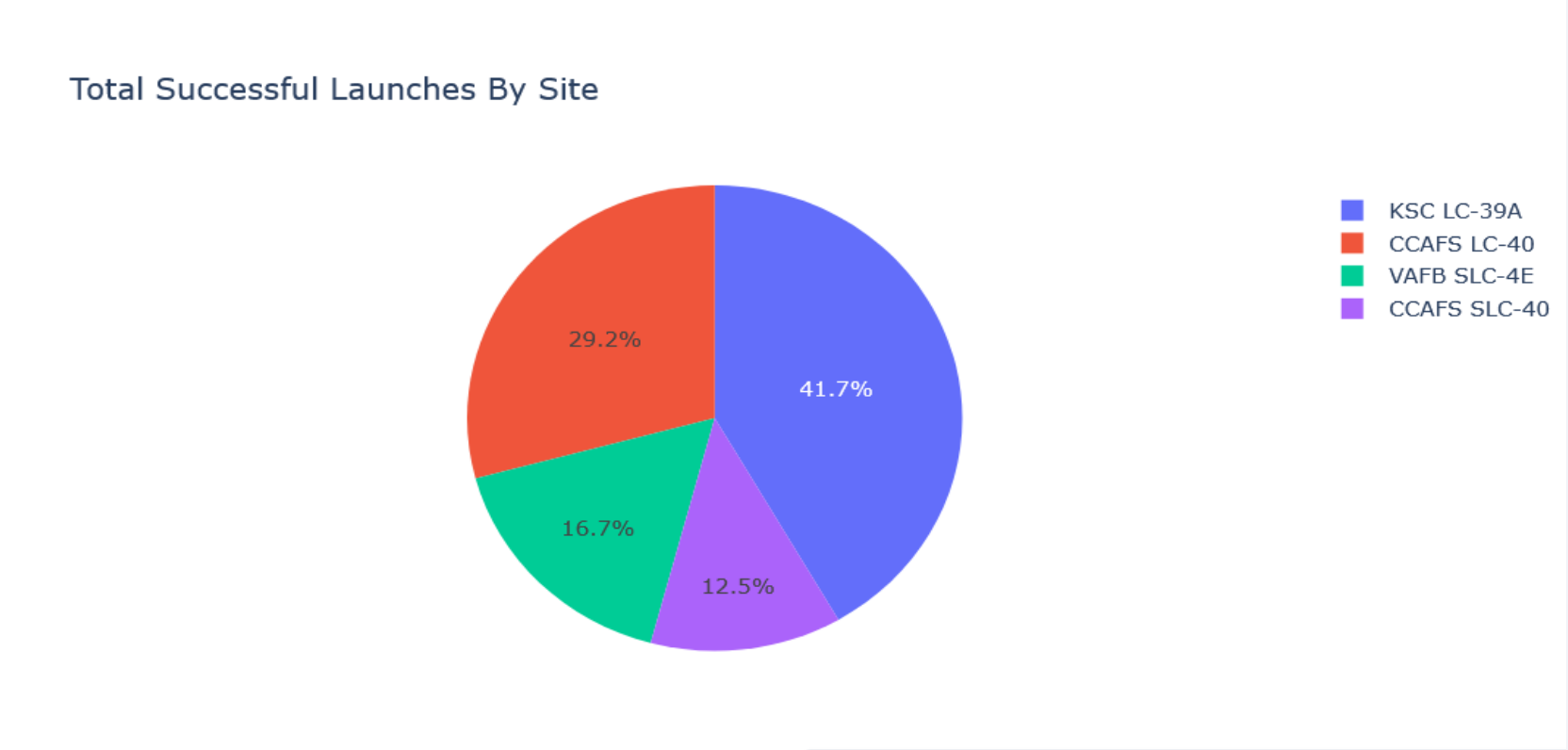
- The closest launch site when compared to the distance from coast, railway and highway line is CCAFS-SLC 40
  - Distance from Coast – 0.87km
  - Railway Line: 0.99 km
  - Highway: 0.62 km
  - Closest City: Melbourne (57.75km)
- The site is closer to railway and highways for greater connectivity to procure equipment for research and manufacturing, but it is far away from city centers to avoid problems occurring from explosions, such as air pollution, harm from falling debris and general obstruction to city life.






Section 4
Build a Dashboard with Plotly Dash

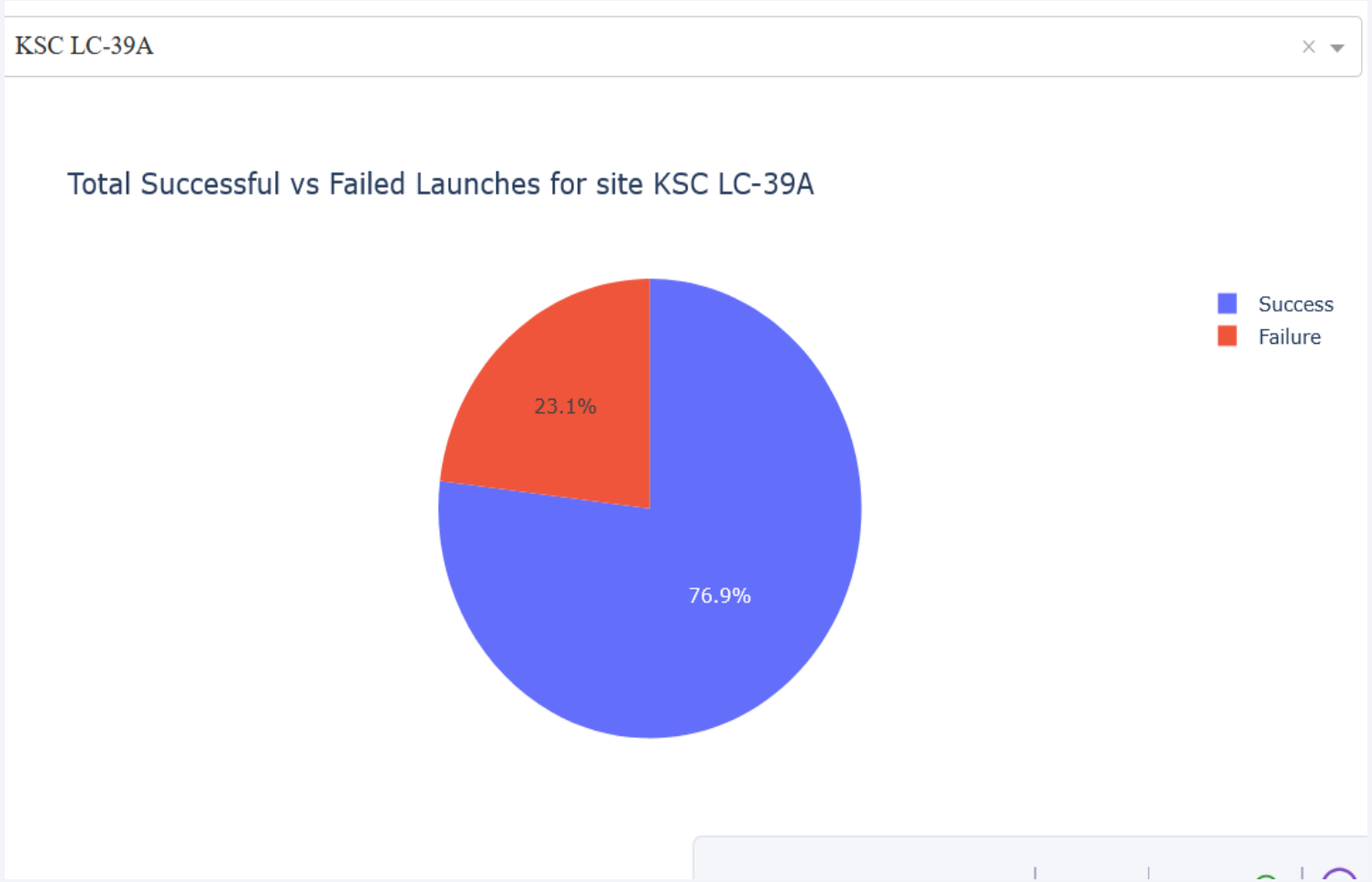
Success Rate by Launch Site



- Most successful Lunch Site is KSC LC – 39A (41.7%)
- Least Successful – CCAFS SLC 40 (12.5%)

Most successful Launch Site

- KSC LC - 39A is the most successful launch site
  - Success Rate: 76.9%
  - Success: 10
  - Failure: 3



Correlation between booster type, payload and success rate

- FT booster version shows the best performance in terms of success rate
- V1.1 and B4 shows the highest number of failures
- Only 3 success instances of payload of over 5000+ kg, lower payload shows higher success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Metrics on Test Data

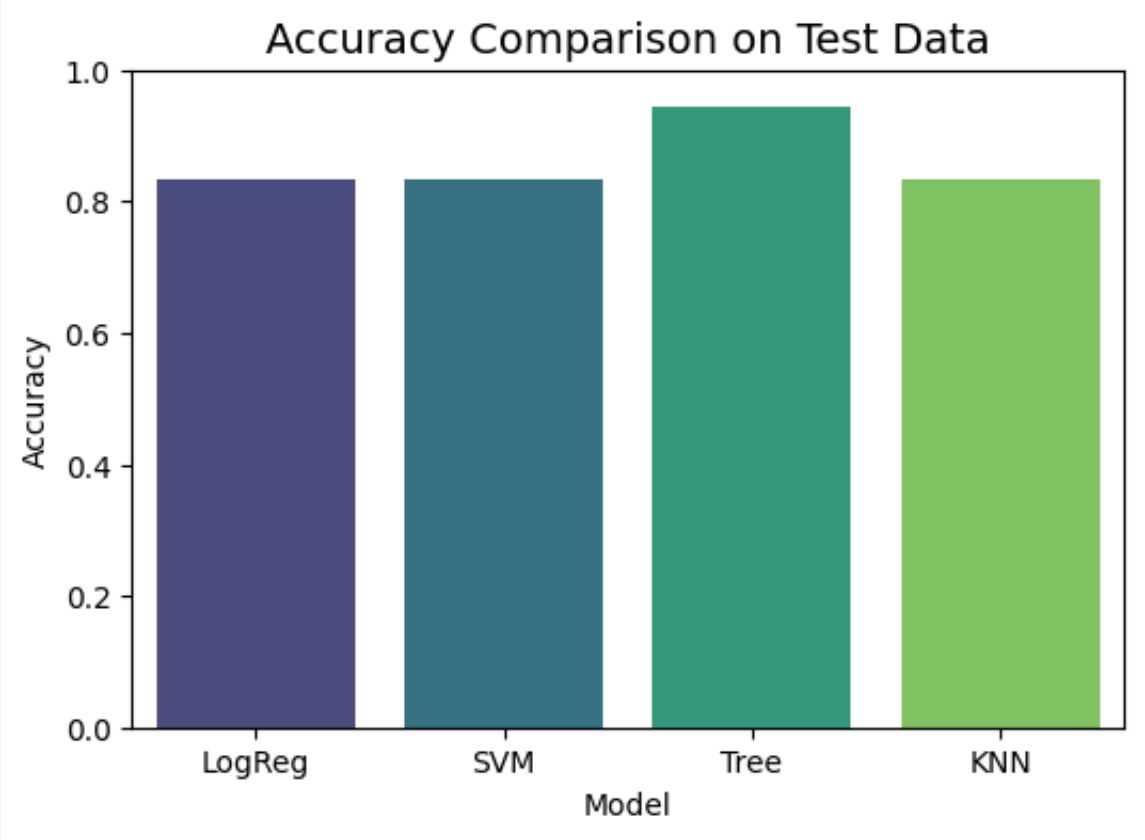
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.944444	0.833333

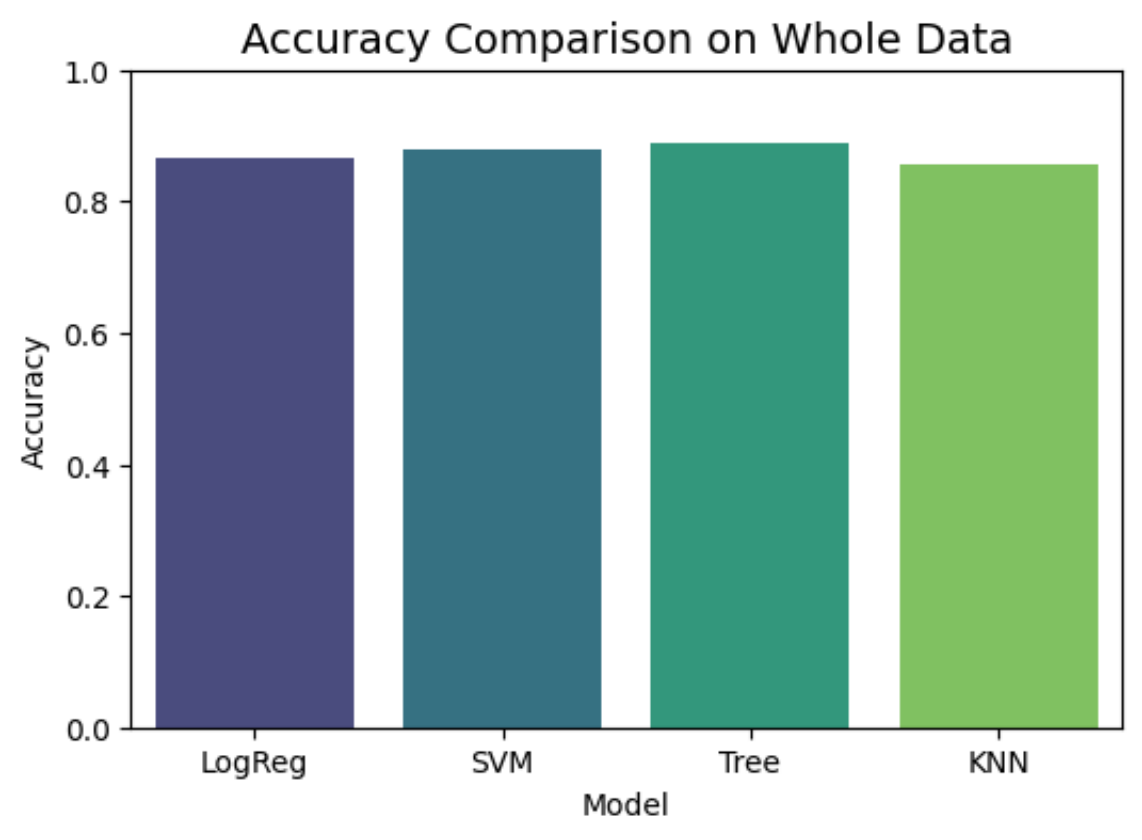
Metrics on Whole Data

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

- Model Accuracy for both the train and test data is observed by using: **"DECISION TREES"**
- Accuracy on **Whole data: 0.91, test data: 0.94**

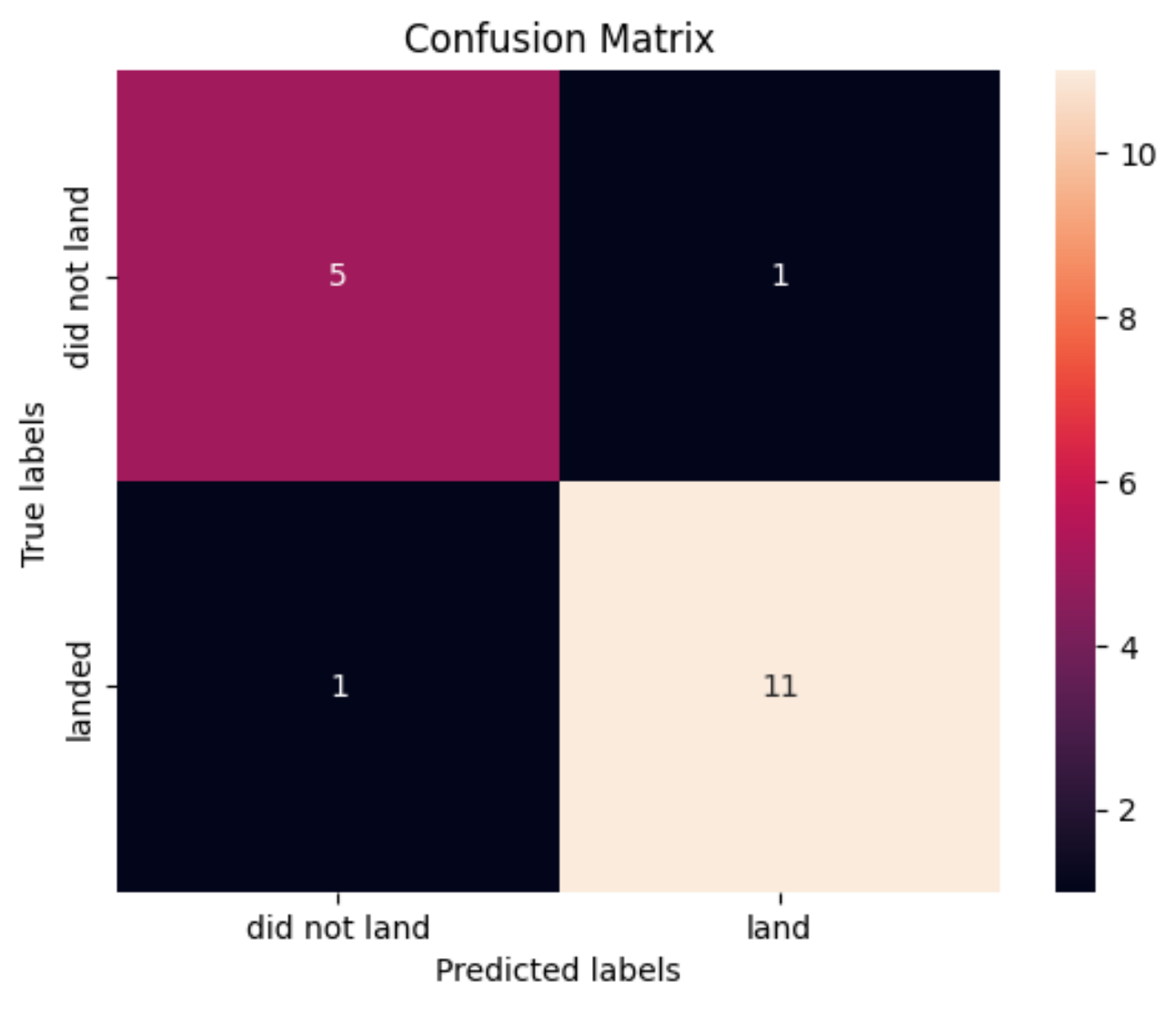
Classification Accuracy





For both testing and training data, accuracy for Decision Trees is the Highest

Confusion Matrix



Confusion matrix also shows only two errors, False Positive (1) and False Negative (1) by the Decision Tree Algorithm for the test data

In conclusion, this is the best performing model for the problem.

Conclusions

- Best algorithm for this classification problem was found to be “Decision Trees” with an F1 score of Point 0.92
- Later flights had higher success rate, this could be dur to improvement in technology and learnings from previous failures. In conclusion, success has improved over the years
- In general, higher payload had a better success rate.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate., though these do not have a high number of launches.
- Maximum number of flights were targeted at the GTO orbit but observed mixed success.

Acknowledgement

- Special Thanks to instructors at Coursera and IBM

Thank you!