IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Stevanus Ong
12/01/2026

Outline

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

Executive Summary

- Methodologies
  - Data collection
  - Data wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Build an interactive visual analytic with Folium
  - Build an interactive dashboard with Plotly Dash
  - Predictive analysis (machine learning with classification task)
- Results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- Background and Context
    - The commercial space age is here, with companies making space travel affordable for everyone. One notable company, SpaceX, has achieved significant milestones, including sending spacecraft to the International Space Station, providing satellite internet, and conducting manned missions to space. The company advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because it reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. We are going to predict if SpaceX will reuse the first stage, based on publicly available information and machine learning.
- Question to be answered
    - How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
    - How does the rate of successful landings increase over the years?
    - What is the best machine learning that can be used for this case (binary classification)?

Section 1

Methodology

Methodology

Data collection methodology:	Using SpaceX Rest API
	Using Publicly available data from Wikipedia (Webscrapped)
Perform data wrangling	Filtering the necessary data
	Impute missing data
Perform exploratory data analysis (EDA) using visualization and SQL	
Perform interactive visual analytics using Folium and Plotly Dash	
Perform predictive analysis using classification models	One-Hot-Encoding and standardize the data
	Split the data to train-test with 80:20 ratio
	Build, tune and test several machine learning with GridSearchCV
	Compare the accuracy of the machine learning on train and test set

Data Collection

Data collection process involved a combination of API requests from SpaceX REST API and webscrapping data from SpaceX Wikipedia.

We had to use both of these data collection methods in order to get complete information about the historical launch for downstream analysis

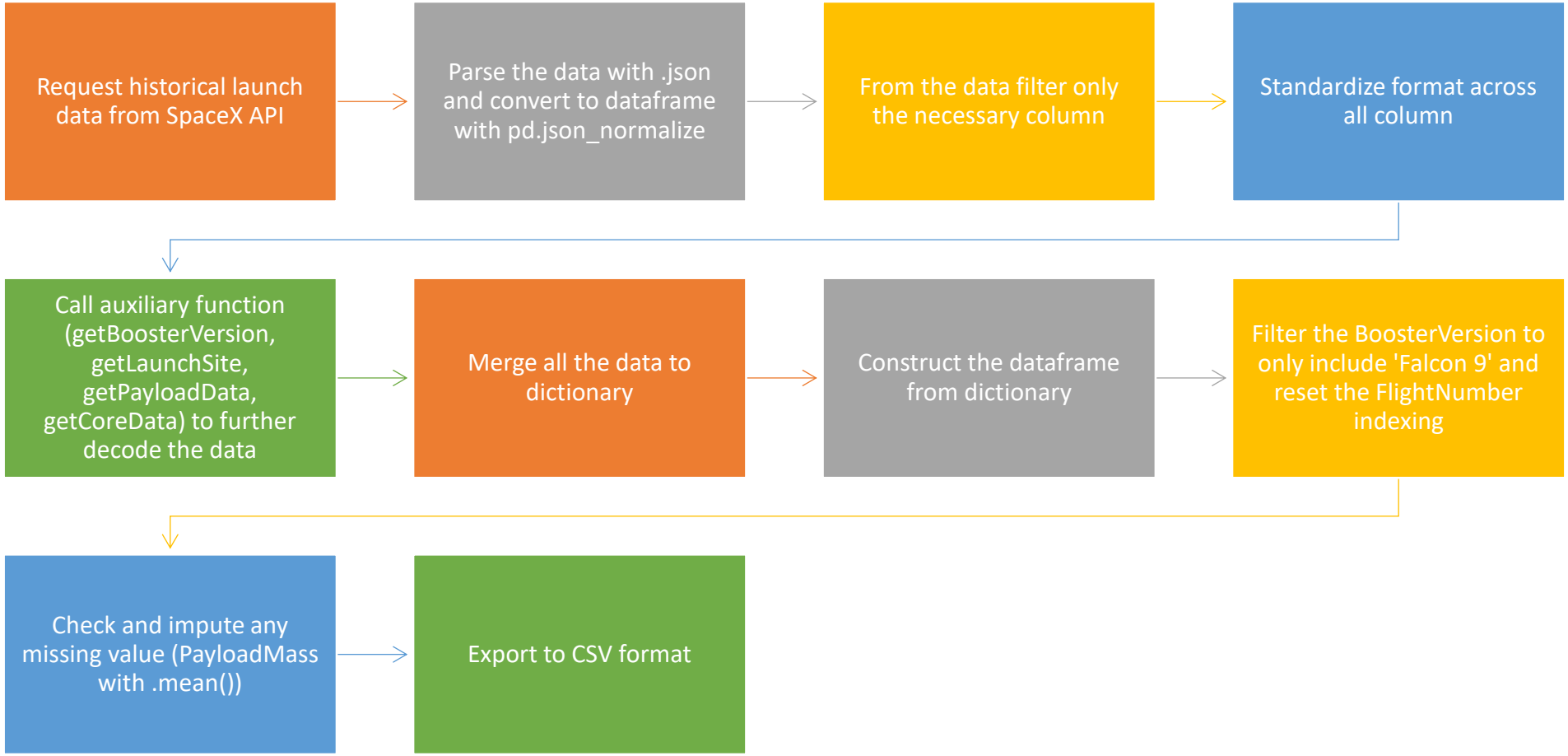
Data obtained from SpaceX API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Data obtained from SpaceX Wikipedia:

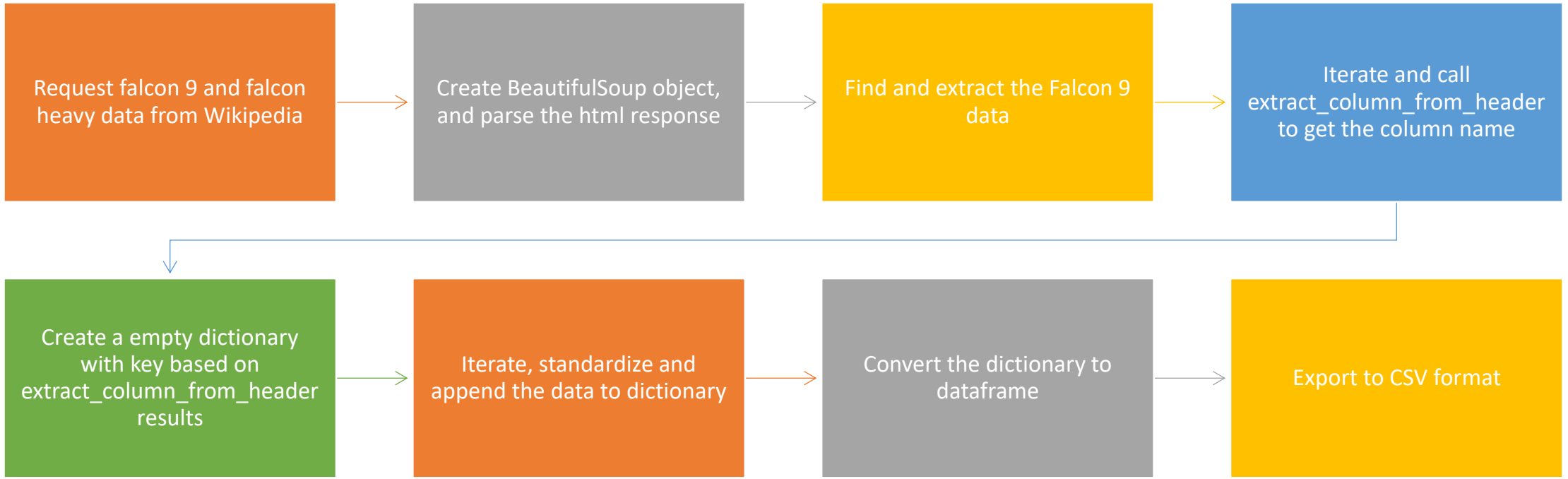
Flight No., Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX REST API



Github URL: [SpaceX Data Collection API]

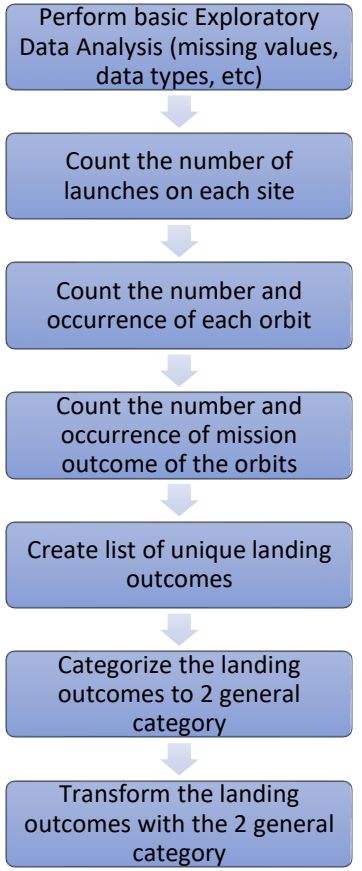
Data Collection – SpaceX Wikipedia



Github URL: [SpaceX Data Collection Data Collection with Web Scraping]

Data Wrangling

- In the dataset, there are several different cases where 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 to 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.
- In this case, we mainly narrow it down to two category that can be used for downstream analytic or machine learning building. We transform it to two “1” mean the booster successfully landed, “0” mean it was unsuccessful.



Github URL: [SpaceX Data Wrangling]

EDA with Data Visualization

In this dataset we used 3 type of charts to do visualization:

Scatter Plot

This plot visualizes the relationship between two numerical variables to identify correlations, clusters, and outliers.

- FlightNumber vs. PayloadMass, FlightNumber vs. Launch Site, PayloadMass vs. LaunchSite, FlightNumber vs. OrbitType, PayloadMass vs. OrbitType

Bar Chart

This plot compares categorical data (like OrbitType) against a numerical feature (like the mean of SuccessRate) to show differences between groups.

- SuccessRate vs. OrbitType

Line Chart

This plot displays trends and continuous changes in data over a chronological period.

- Yearly Launch Success Trend

Github URL: EDA with Visualization

EDA with SQL

- Performed SQL querying:
  - 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, using a subquery with a suitable aggregate function.
  - 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.

Github URL: EDA with SQL (SQLite)

Build an Interactive Map with Folium

Marker of all Launch site:

Added marker with circle and text label to highlight NASA Johnson Space Center on the map.

Added marker with circle and text label to highlight all launch sites on the map.

Marker and cluster of all launch attempt on each sites:

Added marker and cluster all the launch attempt on each site, green marker if the launch attempt successful, red marker if failed.

Display the distances between launch site to its proximities

Calculate the distance and added line and text label to show the distance.

Github URL: [Interactive Visual Analytics with Folium]

Build a Dashboard with Plotly Dash

Add dropdown selection and pie chart

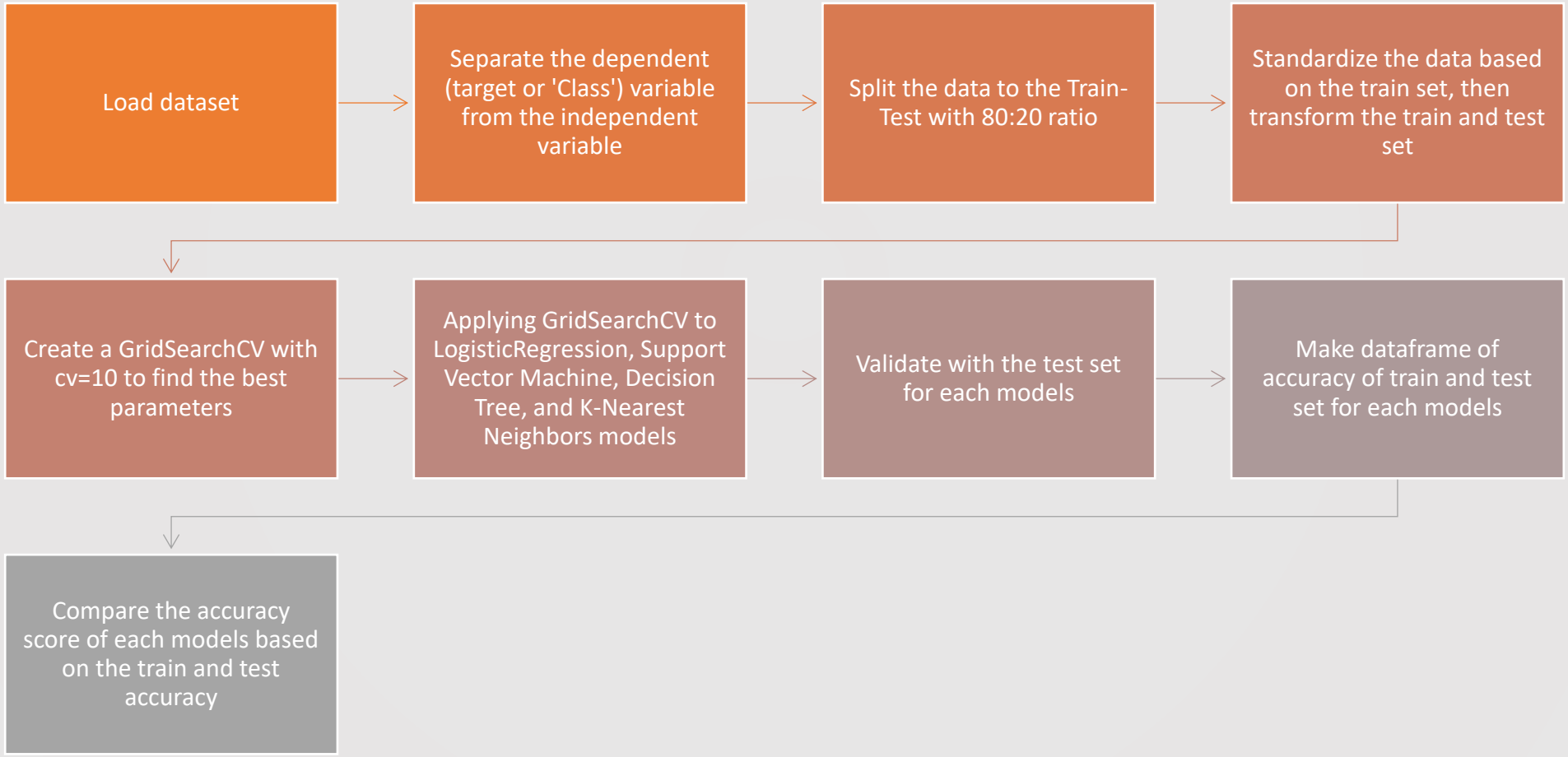
- Add dropdown launch site selection, enable user to include or exclude launch site for visualization
- Add pie chart that will show count of all successful launch for all sites
- Add callback that will let pie chart filter based on the dropdown selection

Add a slider and scatter plot

- Add slider of payload mass with range of (0-10000), enable user to filter the payload mass
- Add scatter plot that will show the correlation between payload and launch success
- Add callback that will let scatter plot filter based on dropdown selection of launch sites and slider of payload mass.

Github URL: [Interactive Dashboard with Plotly Dash]

Predictive Analysis (Classification)



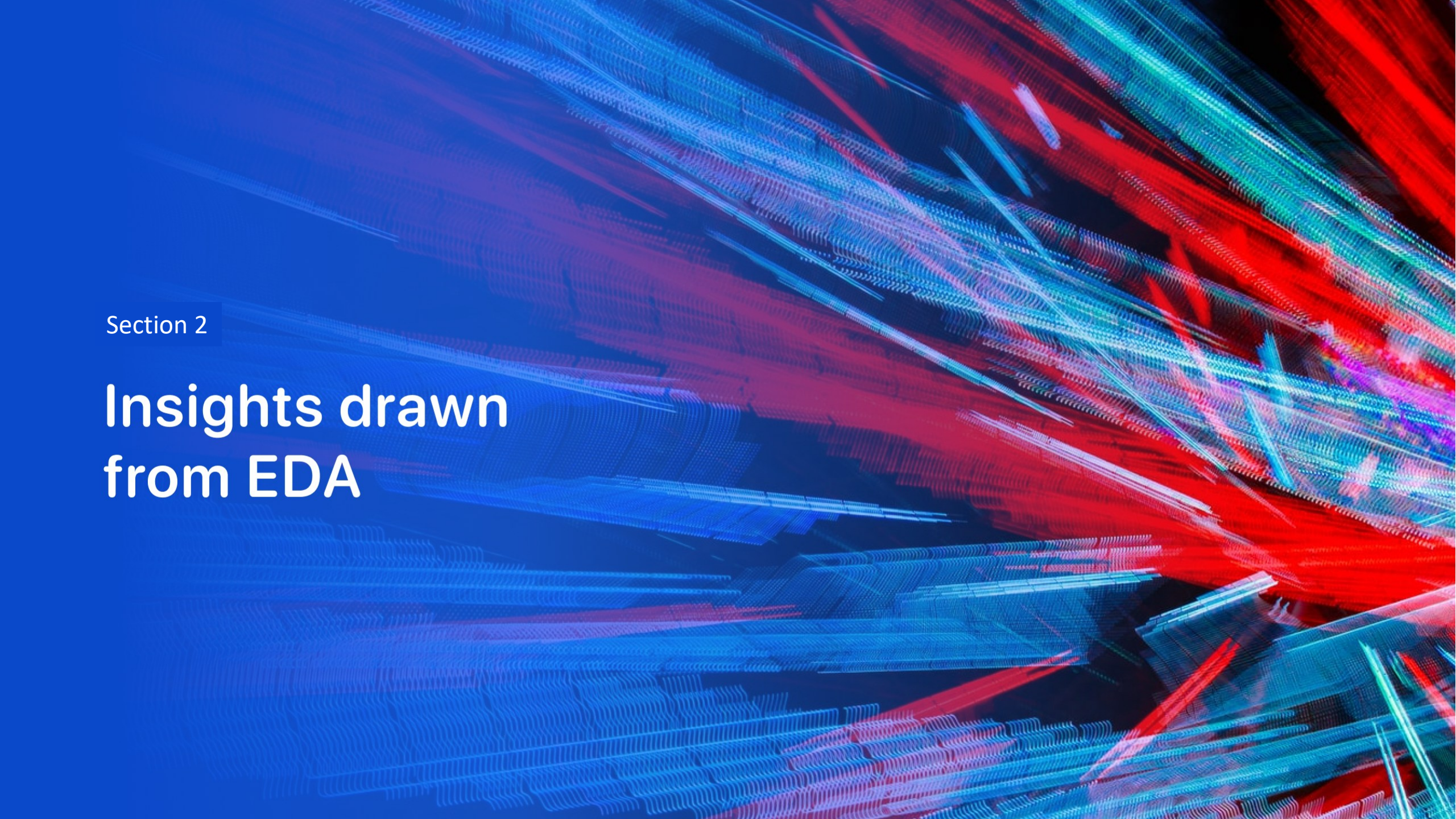
Github URL: [SpaceX Machine Learning Prediction]

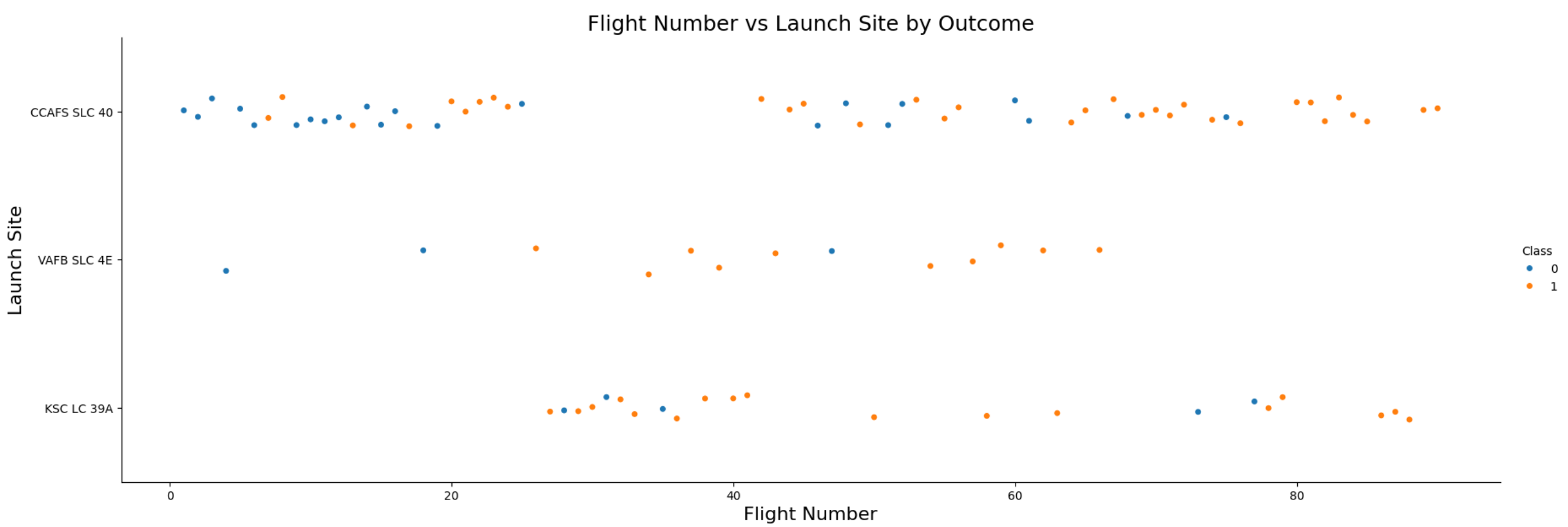
Results

EXPLORATORY DATA ANALYSIS RESULTS

INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS

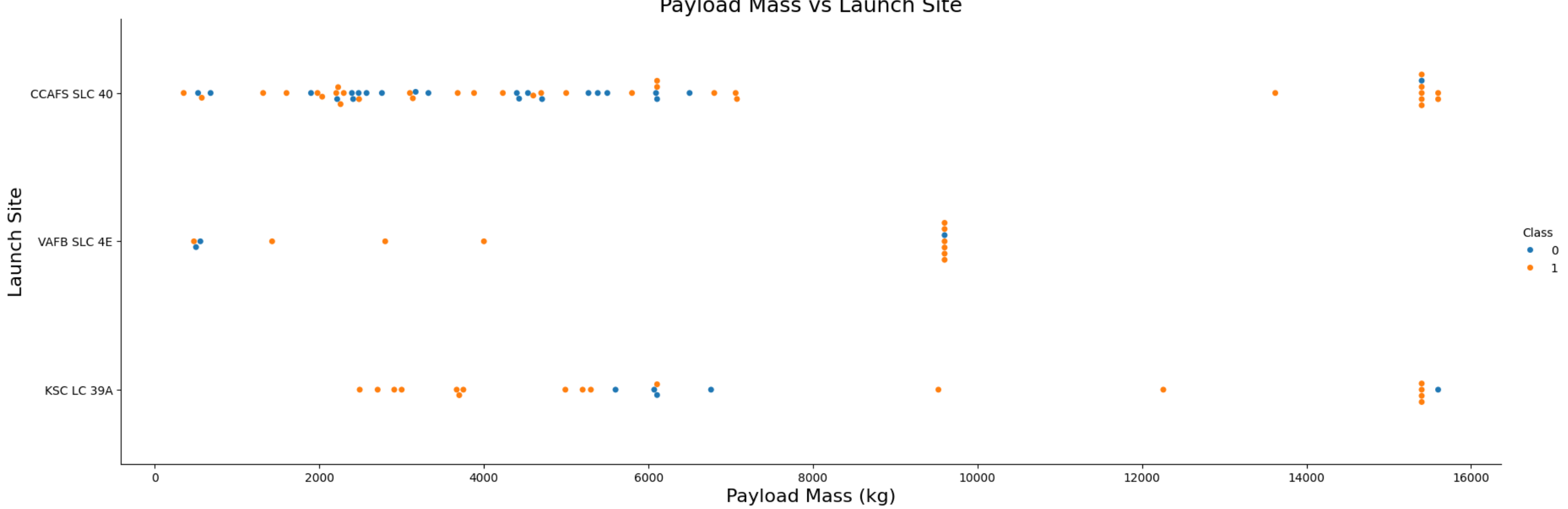
PREDICTIVE ANALYSIS RESULTS


Section 2
Insights drawn from EDA



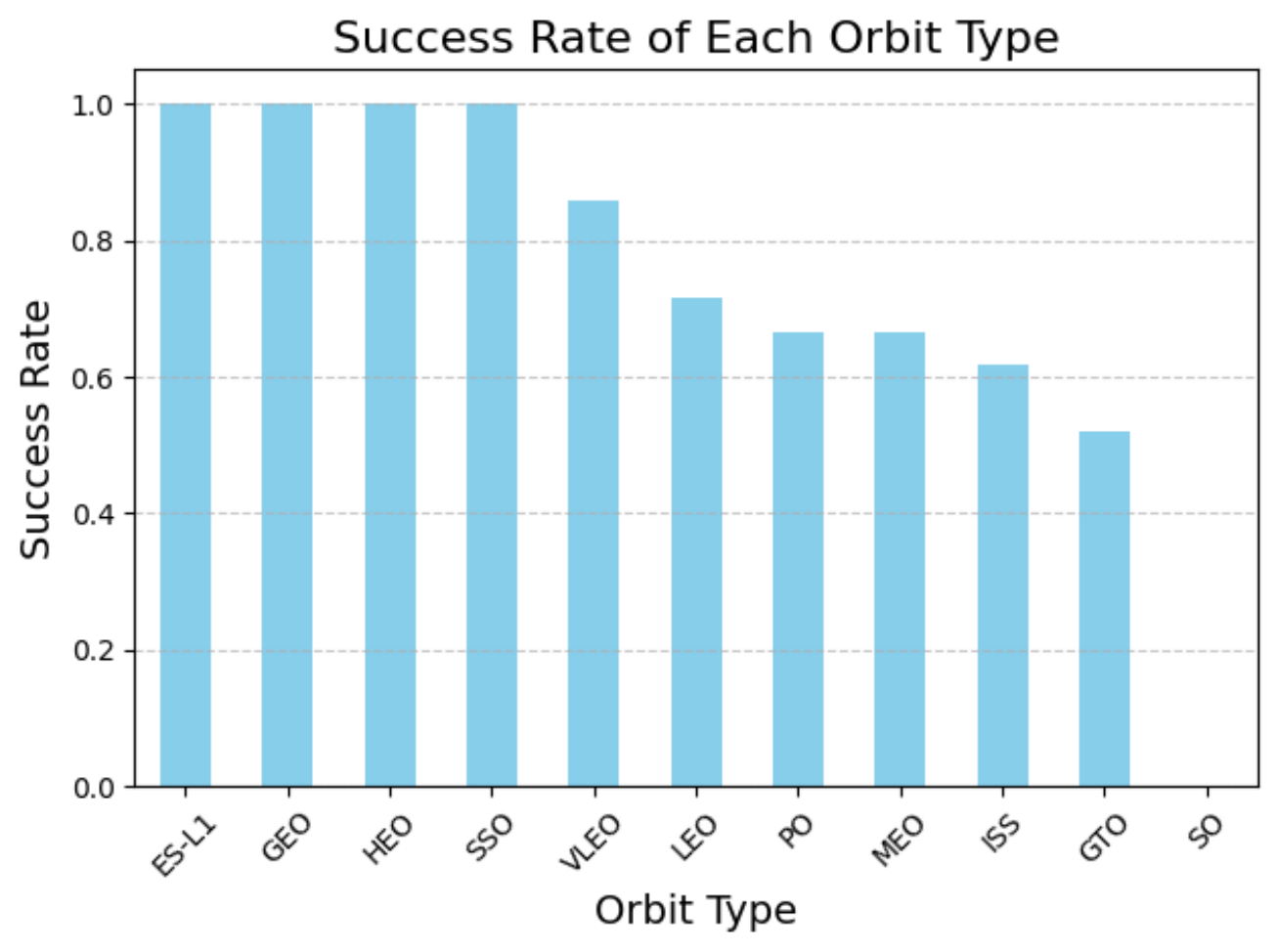
Flight Number vs. Launch Site

- The earliest launch result more on failed, while latest launch result in more success
- Most of launch happen in CCAFS SLC-40
- VAFB SLC-4E and KSC LC-39A have higher success rate
- As FlightNumber increase, we can see that the launch success rate increasing



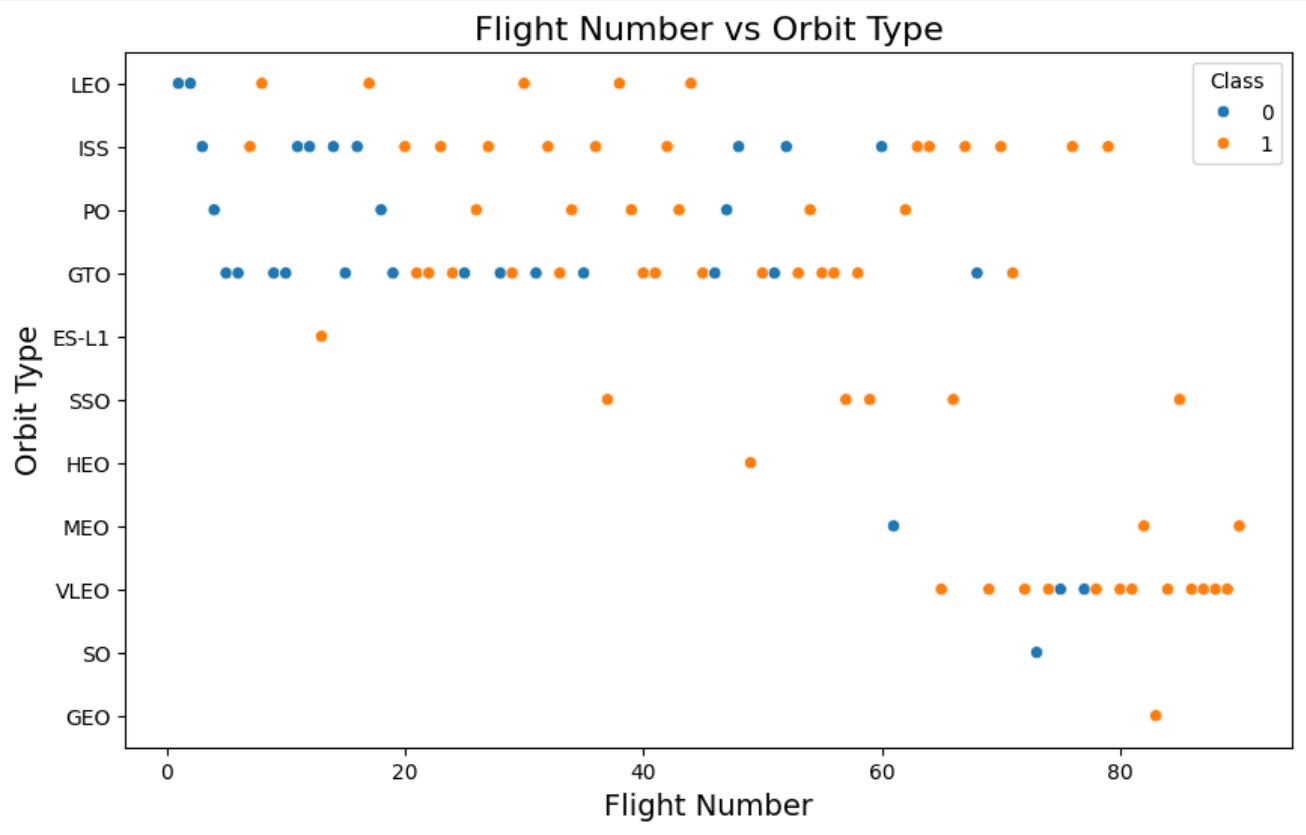
Payload vs. Launch Site

- KSC LC-39A and CCAFS SLC-40 have higher chance of success with higher payload mass
- VAFB SLC-4E specialized and payload limited to lighter to mid-range payload (generally below 10000 kg)



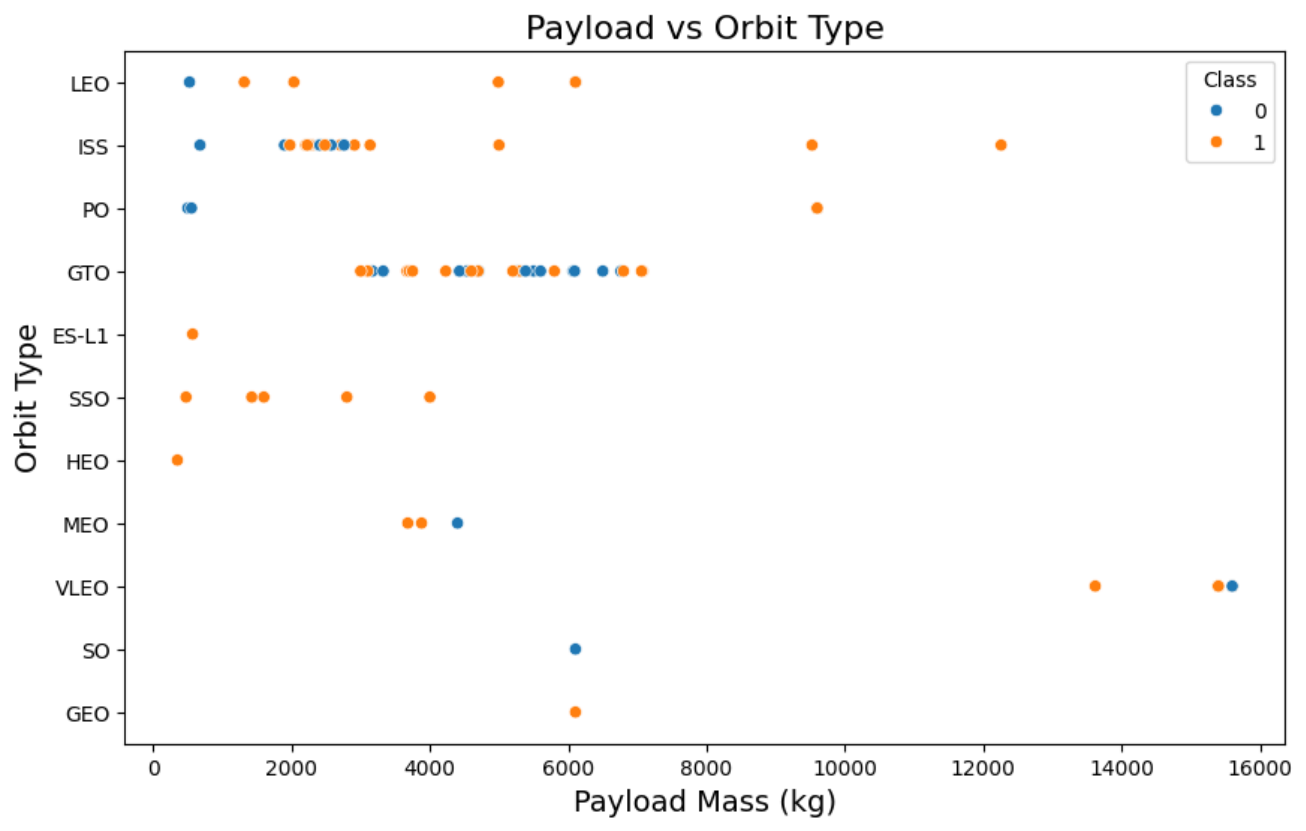
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO have a 100% success rate
- SO have 0% success rate
- VLEO, LEO, PO, MEO, ISS, and GTO have 85-50% success rate



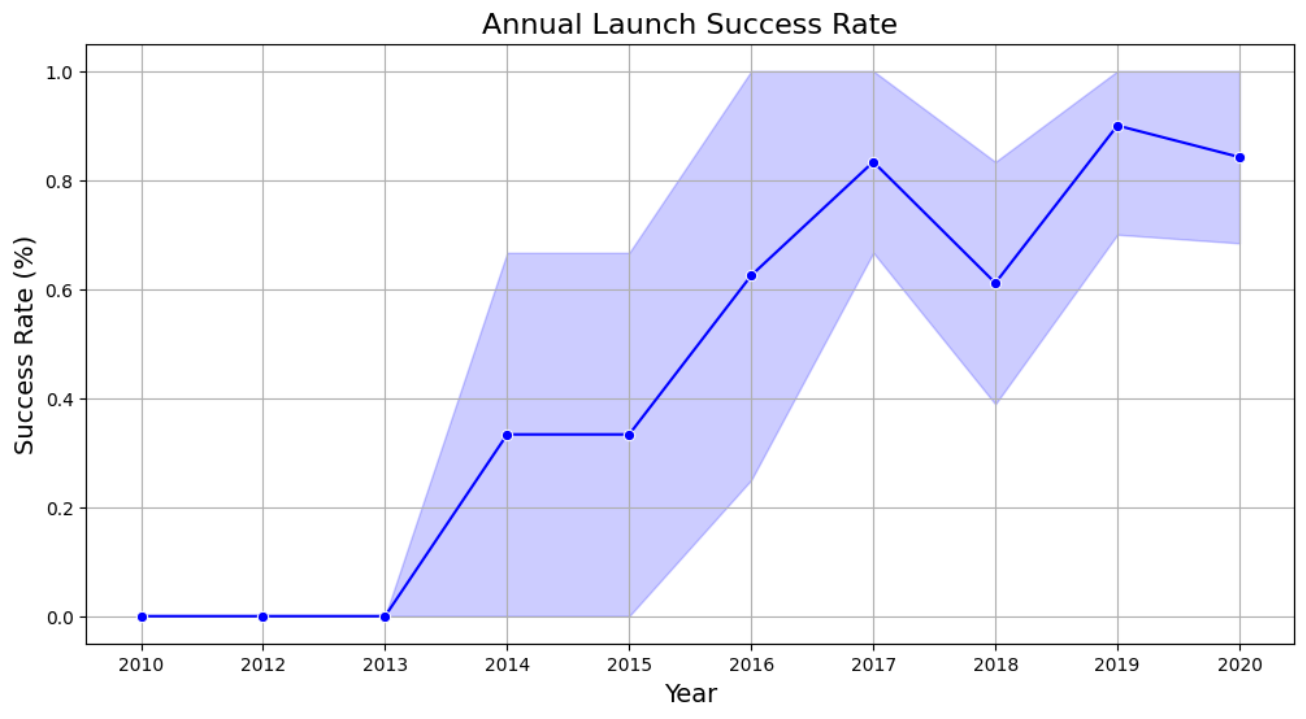
Flight Number vs. Orbit Type

- Early flight show a mix of success across multiple orbit types, while later flight have higher chance of success
- Most failed occur in LEO, ISS, PO and GTO during early flight, while other orbit have higher chance of success
- It can be assumed mission success improves significantly with more flight attempt
- Advanced or less frequent orbit types are introduced later and show high success rate



Payload vs. Orbit Type

- VLEO has the heaviest payload (13000-16000 kg)
- LEO and ISS generally carry light payload (< 7000 kg)
- GTO mission cluster around 3000-7000 kg with mix of success and failed
- ES-L1, SSO and HEO have 100% success rate



Launch Success Yearly Trend

- While early years (2014–2016) showed promise, the data was volatile due to low volume. However, the recent trend (2019–2020) shows both improving performance and increased stability.

All Launch Site Names

- Displaying all unique launch site SpaceX used

```
%%sql
SELECT DISTINCT "Launch_Site"
FROM SPACEXTBL
```

[9] ✓ 0.0s

```
 * 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'

- Displaying all record of launch that happen in launch site that has name start with ‘CCA’

```
%%sql
SELECT *
FROM SPACEXTBL
WHERE Launch_Site LIKE "CCA%"
LIMIT 5
```

[10] ✓ 0.2s

```
 * 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

- Display the Total Payload Mass has been launched by ‘NASA (CRS)’ with SpaceX service

```
[9]  1  %%sql
     2  SELECT Customer, SUM("PAYLOAD_MASS__KG_") AS "Total_Payload_Mass_KG"
     3  FROM SPACEXTBL
     4  WHERE Customer = "NASA (CRS)"
     5  GROUP BY Customer
     ✓  0.0s

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

...
```

Customer	Total_Payload_Mass_KG
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- Display the average payload mass carried by booster F9 v1.1 and all its derivative variants

```
%%sql
SELECT ROUND(AVG("PAYLOAD_MASS__KG_"), 2) AS Average_Payload_Mass_KG
FROM SPACEXTBL
WHERE Booster_Version LIKE "%F9 v1.1%"
```

[13] ✓ 0.0s

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

Average_Payload_Mass_KG
2534.67

```
%%sql
SELECT MIN(Date) AS First_Success_Date
FROM SPACEXTBL
WHERE Landing_Outcome = "Success (ground pad)"
ORDER BY Date ASC
LIMIT 1
```

[31] ✓ 0.0s

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

First_Success_Date
2015-12-22

First Successful Ground Landing Date

Display the date of the first successful ground landing achieved

```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE Landing_Outcome = "Success (drone ship)" AND ("PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000)
```

[32] ✓ 0.0s

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

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

Successful Drone Ship Landing with Payload between 4000 and 6000

- Display the booster name that successfully landed in drone ship with payload mass greater than 4000 but less than 6000

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

[53] ✓ 0.0s

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

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

Total Number of Successful and Failure Mission Outcomes

- Display the number of successful and failure mission outcomes

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
%%sql
SELECT DISTINCT Booster_Version, PAYLOAD_MASS__KG_ AS Max_Payload_Mass
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL
)
ORDER BY Booster_Version ASC
```

[55] ✓ 0.0s

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

Booster_Version	Max_Payload_Mass_KG
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

- Display the failed landing attempt that happen in drone ship in year = 2015

```
24  SELECT
25      CASE SUBSTR(Date, 6, 2)
26          WHEN '01' THEN 'January'
27          WHEN '02' THEN 'February'
28          WHEN '03' THEN 'March'
29          WHEN '04' THEN 'April'
30          WHEN '05' THEN 'May'
31          WHEN '06' THEN 'June'
32          WHEN '07' THEN 'July'
33          WHEN '08' THEN 'August'
34          WHEN '09' THEN 'September'
35          WHEN '10' THEN 'October'
36          WHEN '11' THEN 'November'
37          WHEN '12' THEN 'December'
38      END AS Month_Name,
39      Landing_Outcome,
40      Booster_Version,
41      Launch_Site
42  FROM SPACEXTBL
43  WHERE Landing_Outcome = 'Failure (drone ship)' AND SUBSTR(Date,
```

[56] ✓ 0.0s

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

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

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

- Rank the Landing Outcomes between 2010-06-04 and 2017-03-20 in descending order

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

[58] ✓ 0.0s

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

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

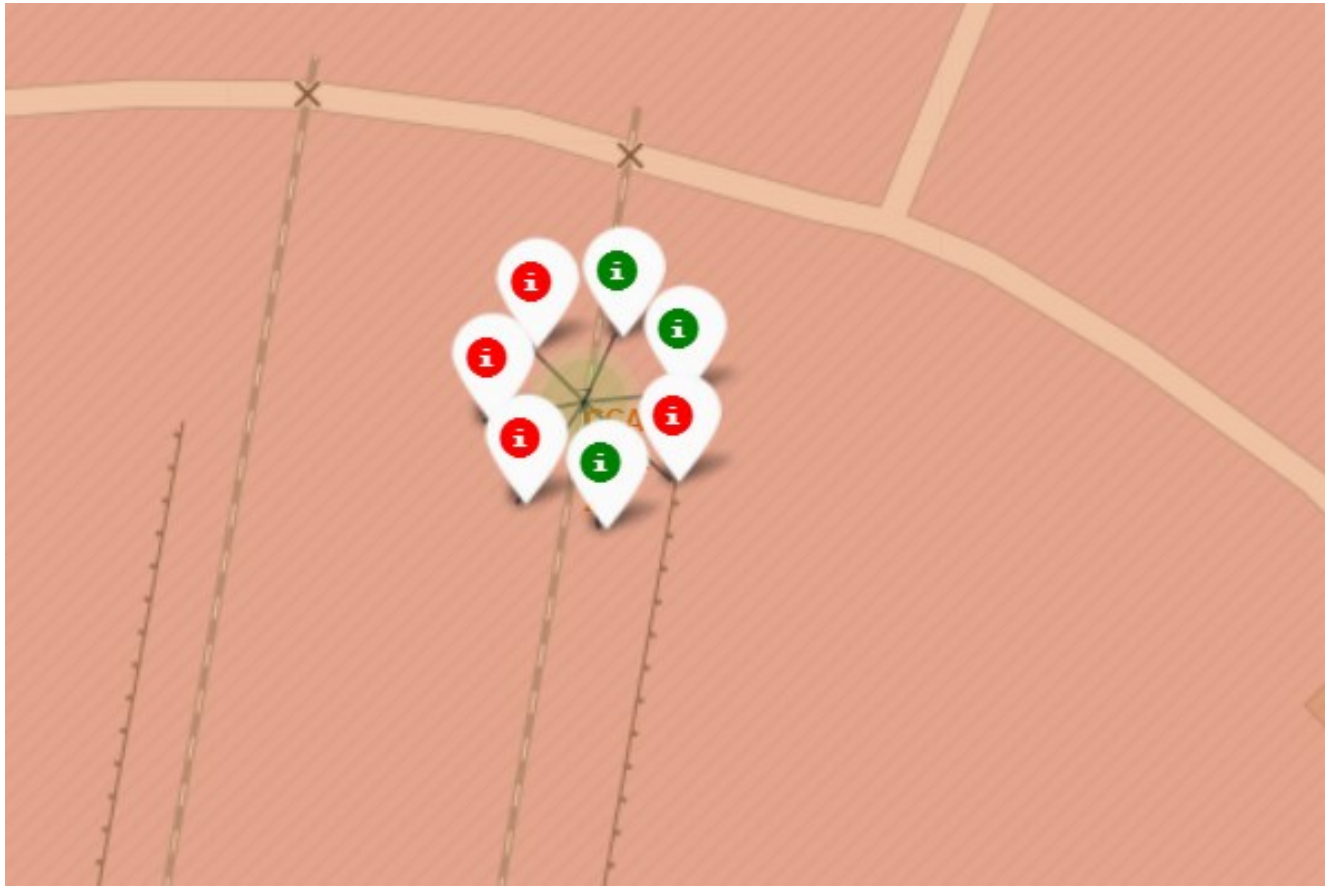

Section 3
Launch Sites
Proximities Analysis

All launch site's location markers

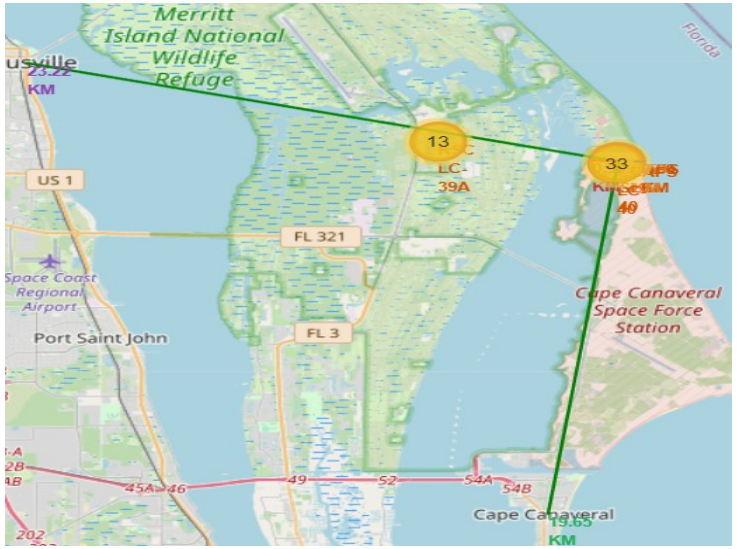
- Most space launch complexes are strategically located near the equator and large bodies of water to maximize physics and safety. Launching closer to the equator allows rockets to harness the Earth’s maximum rotational speed, acting like a natural 'slingshot' which significantly reduces the fuel required to reach orbit. Furthermore, being adjacent to the ocean provides a vast, unpopulated 'drop zone' where spent rocket stages or malfunctioning vehicles can safely fall without endangering human lives or property.

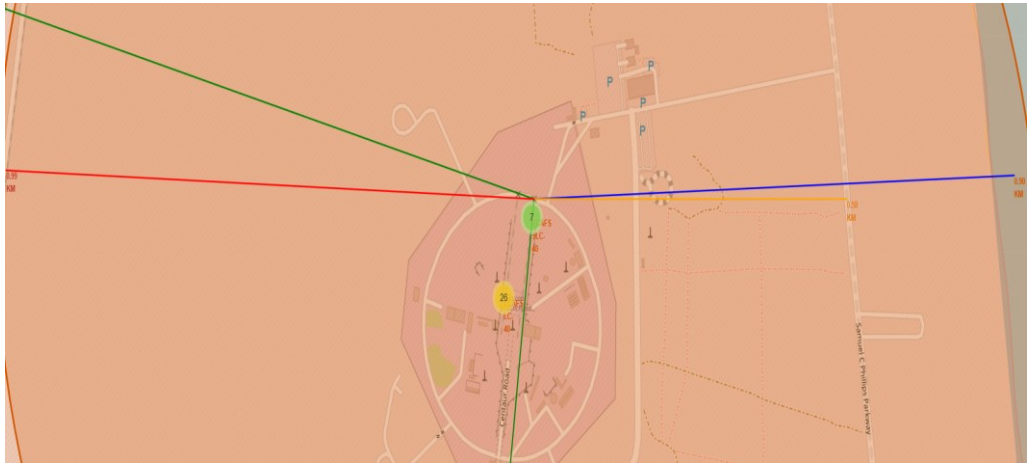
Colour-labeled launch outcome on the map

- Added colour labeled marker that will make it easily identify is the launch at launch site successful or failed.
    - Green colour marker: successful launch
    - Red colour marker: failed launch
- CCAFS SLC-40 have relatively more failed launch

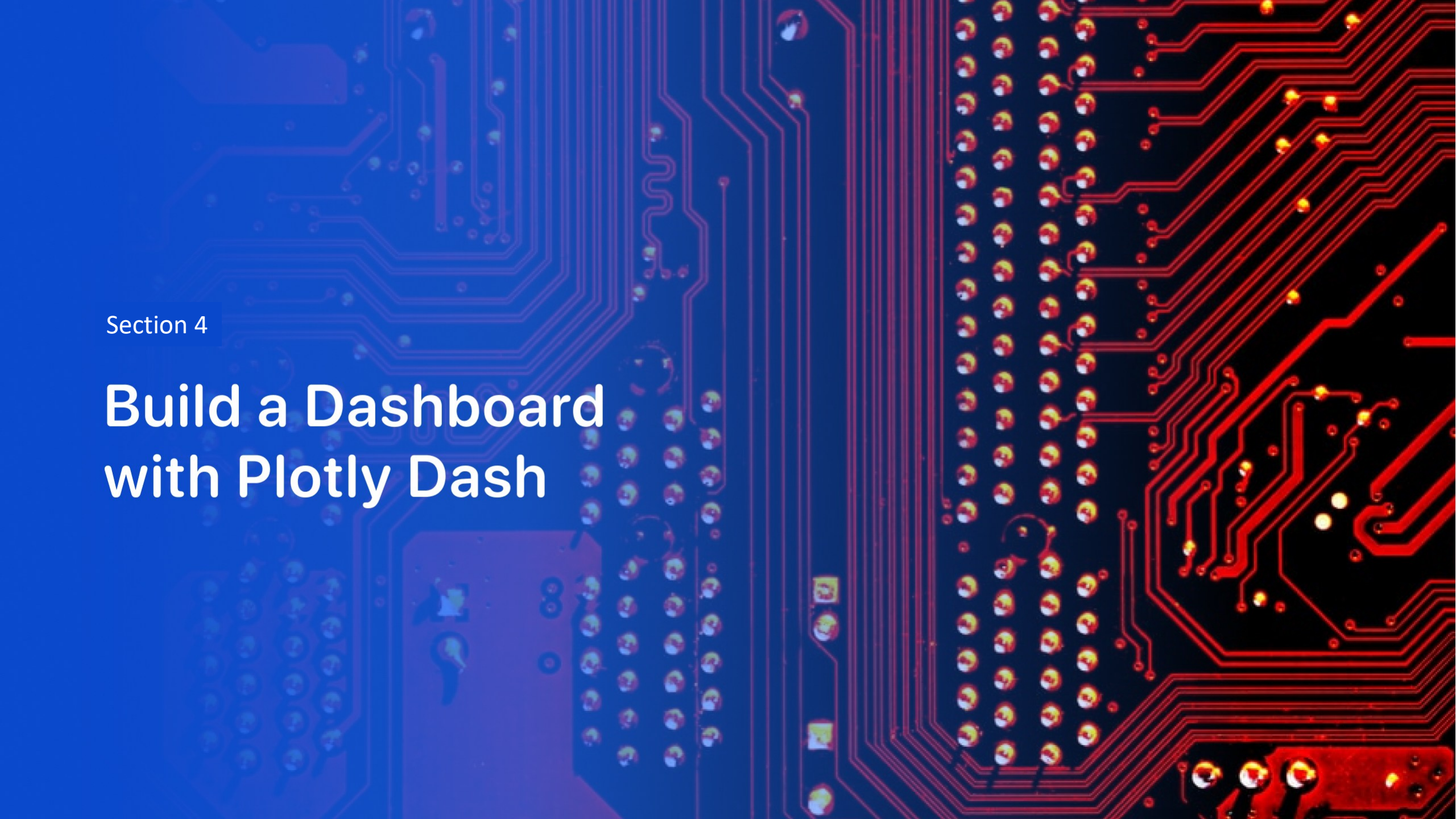


Distance from the launch site CCAFS SLC-40 to its proximities



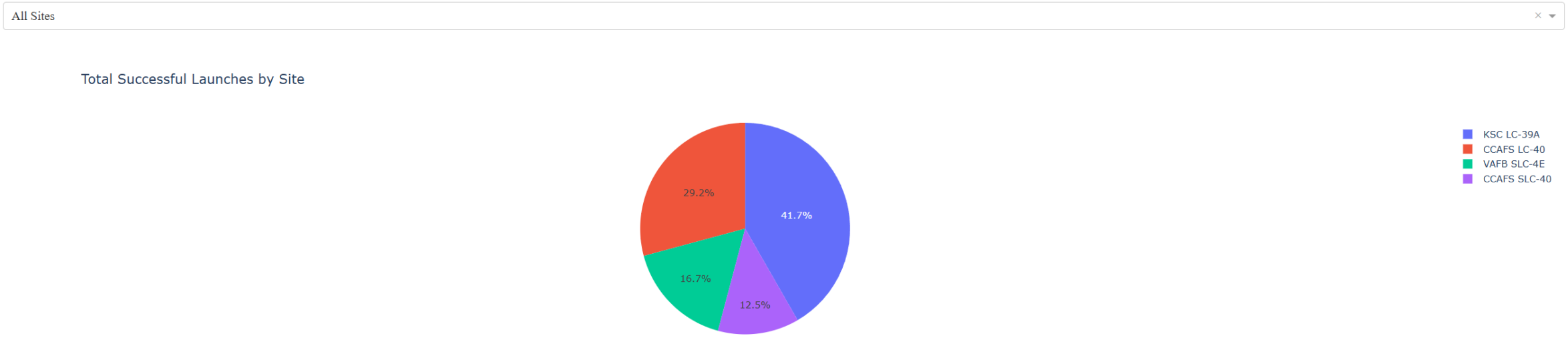


- From the visual, CCAFS SLC-40 we can see that the proximities it is:
  - launch site to closest coastline (blue line) : 0.89 km
  - launch site to Cape Canaveral City (green line) : 19.64 km
  - launch site to Titusville City (green line) : 23.22 km
  - launch site to closest highway (yellow line) : 0.58 km
  - launch site to closest railroad (red line) : 28.56 km
- Failed or unexpected event that happen during first stage may result in space debris around launch site (15-20km) or another unknown coordinate.


Section 4
Build a Dashboard with Plotly Dash

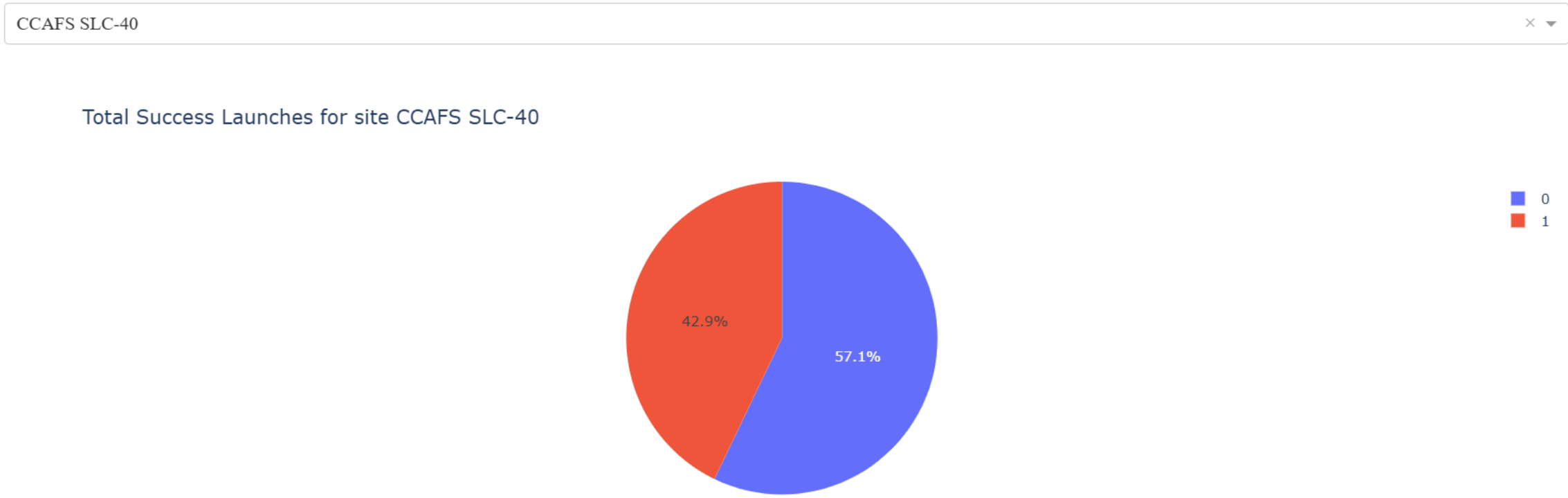
Launch Success for all sites

- The chart show all launch success for each site, and from that we can see that KSC LC-39A has the highest successful launch.



Launch site with the highest launch success ratio

- From the chart, CCAFS SLC-40 has the highest launch success ratio, with 3 (Class=1) successful launch and 7 failed (Class=0)



Payload vs. Launch Outcome for all sites

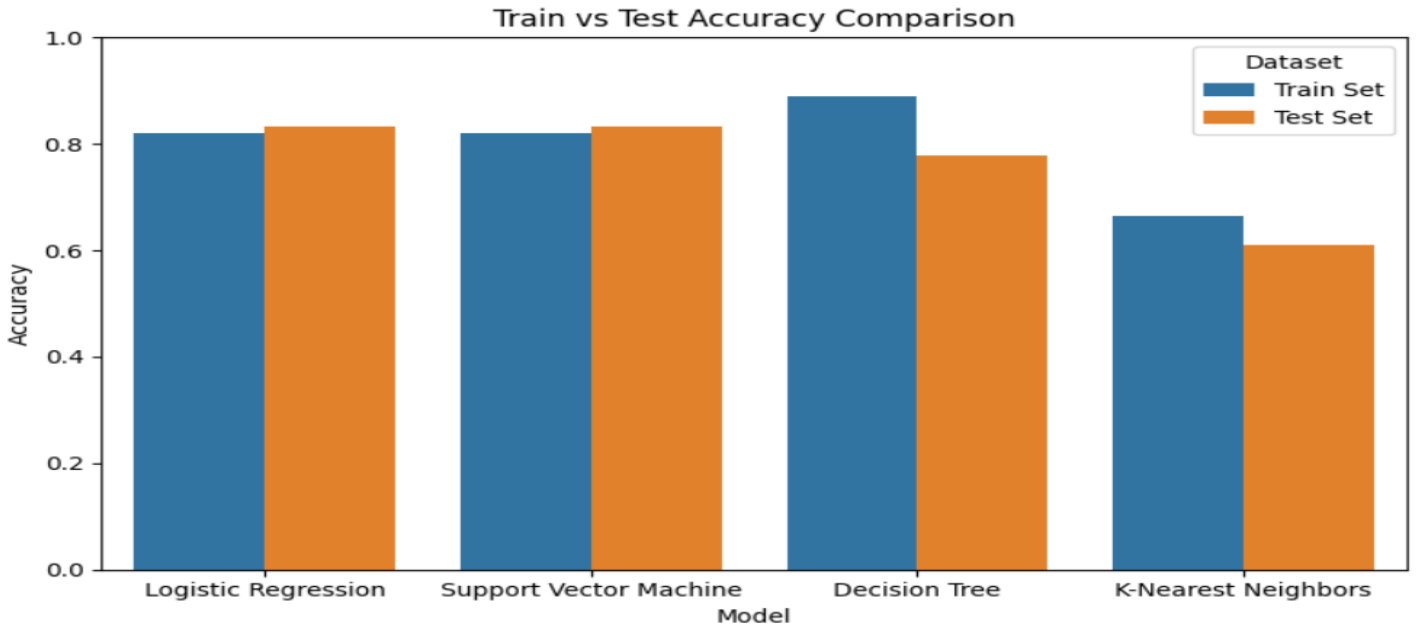
- From the chart, we can see most of success launch cluster around payload mass of 1900-3700 kg and 4500-5500 kg

Section 5

Predictive Analysis (Classification)

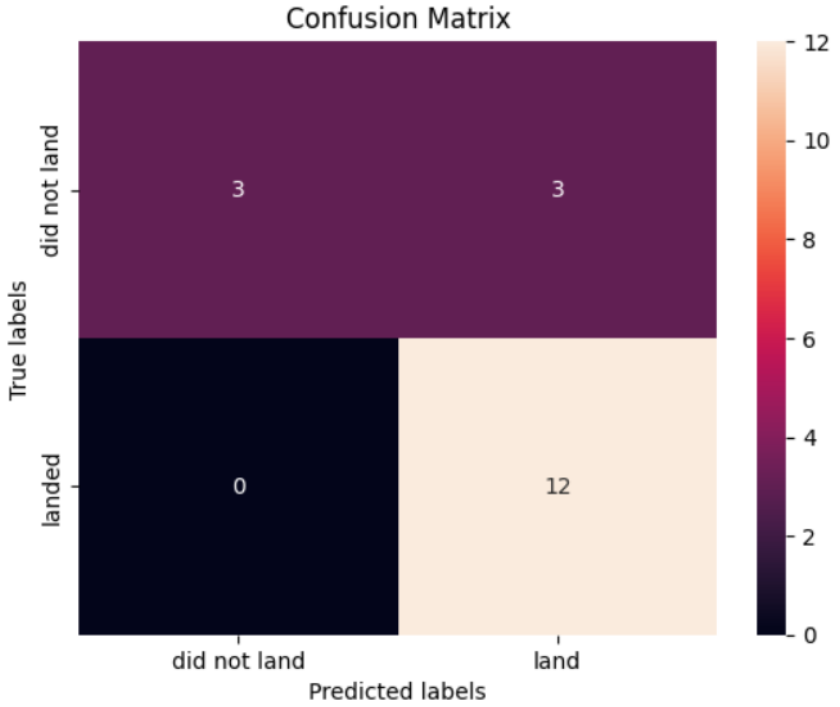
Classification Accuracy

- Based on train set, we can see that Decision Tree followed by Logistic Regression and SVM is the best at this classification task
- Based on test set both Logistic Regression and SVM is the best at test set with the same accuracy score, but decision tree is falling behind

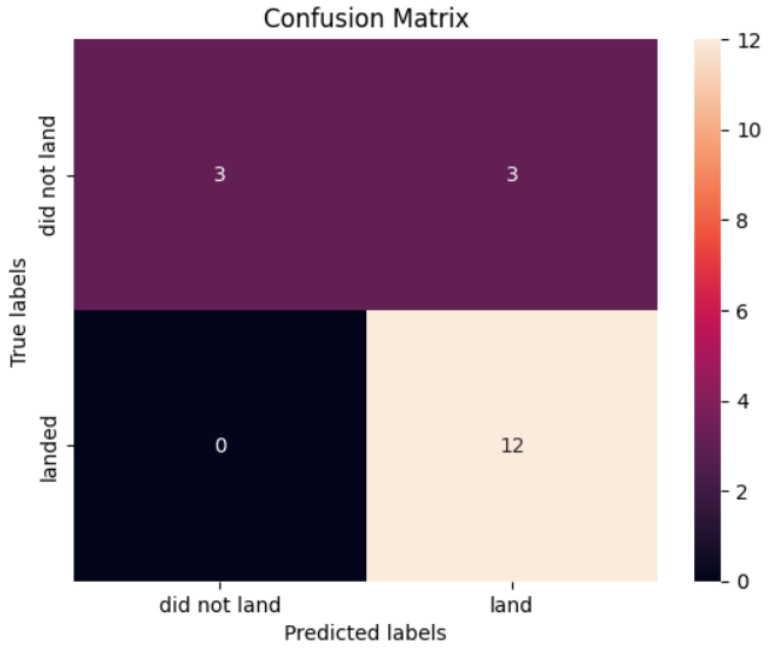


Model	Accuracy Score	Set
Decision Tree	0.889286	Train Set
Logistic Regression	0.819643	Train Set
Support Vector Machine	0.819643	Train Set
K-Nearest Neighbors	0.664286	Train Set

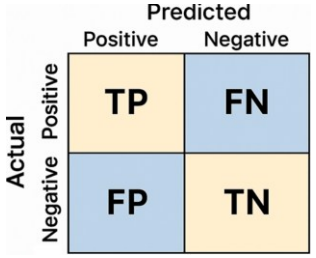
Model	Accuracy Score	Set
Logistic Regression	0.833333	Test Set
Support Vector Machine	0.833333	Test Set
Decision Tree	0.777778	Test Set
K-Nearest Neighbors	0.611111	Test Set



Confusion matrix for Logistic Regression on Test Set



Confusion matrix for SVM on Test Set



Confusion Matrix

- From the previous Bar chart, we know that Logistic Regression and SVM perform well both on train and test set, we also can see that both have the same confusion matrix result, both fall behind on False Negative resulting on error on 3 prediction.

Conclusions

Both Logistic Regression and SVM is the best algorithm for this dataset and task.

Most failed occur in LEO, ISS, PO and GTO during early flight, while other orbit have higher chance of success

The success rate of launches increase over year, shows both improving performance and increased stability.

Advanced or less frequent orbit types are introduced later and show high success rate

CCAFS LSC-40 launch site has the highest success rate of launches compare to all the other sites.

Most of launch sites are in proximity to the Equator line and all the sites are in proximity to the coastline

Appendix

- Special Thanks to
  - Instructor
  - IBM
  - Coursera

Thank you!