Winning Space Race with Data Science

Name: An Nguyen
Date: 26/03/2022

Github link for the project

Outline

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

Executive Summary

- Summary of methodologies

Collected data from public SpaceX API and Wikipedia. -> Perform data Wrangling -> Perform Explanatory Data Analysis (EDA) -> Visualize Data -> Create interactive dashboard -> Perform Predictive Analysis with Machine Learning

- Summary of all results

Decision tree is the best model for predicting the outcome of Space X Falcon 9 First Stage Landing

Introduction

- Project background and context

We pretend to be Space Y Company. We already know that Space X Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

- Problems and Objectives

Create a machine-learning model to predict if the first stage will land

Section 1

Methodology

Methodology

Executive Summary

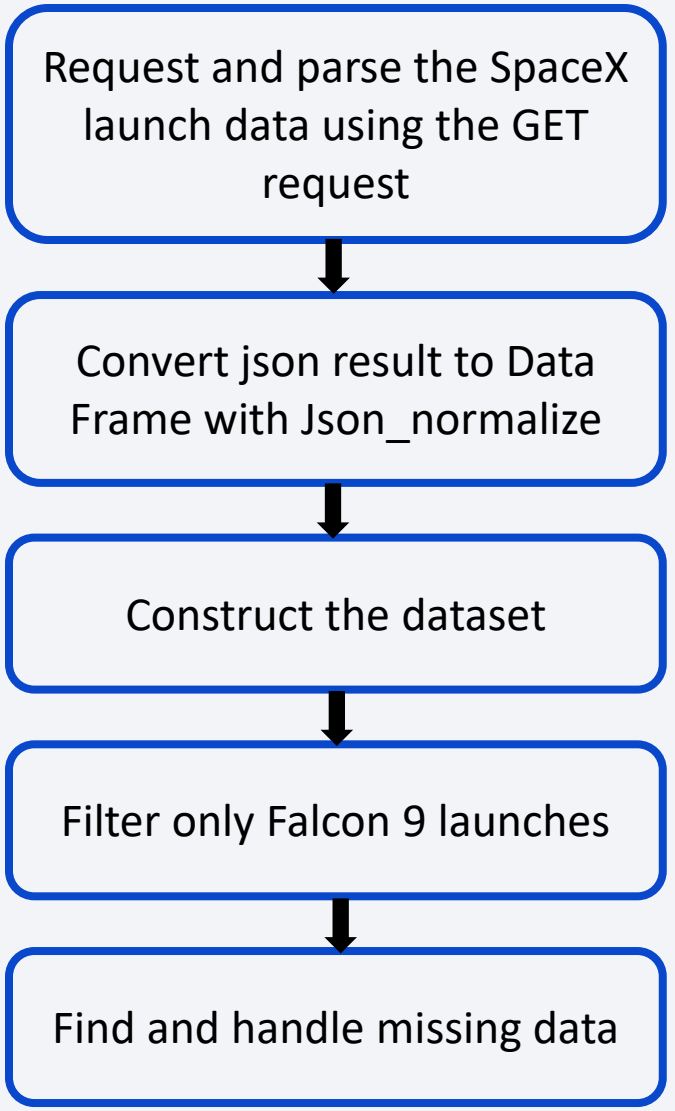
- Data collection methodology:
  - Data was collected from public SpaceX API and Wikipedia
- Perform data wrangling
  - Convert outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build and evaluate KNN, Decision Trees,…models

Data Collection

- Data was collected from public SpaceX API and SpaceX Wikipedia:
- The dataset From SpaceX API provides information about the launch date, time, outcomes,…
- The dataset from SpaceX Wikipedia gives us a list of Falcon 9 and Falcon Heavy launches, a few example of unsuccessful and successful landing,…

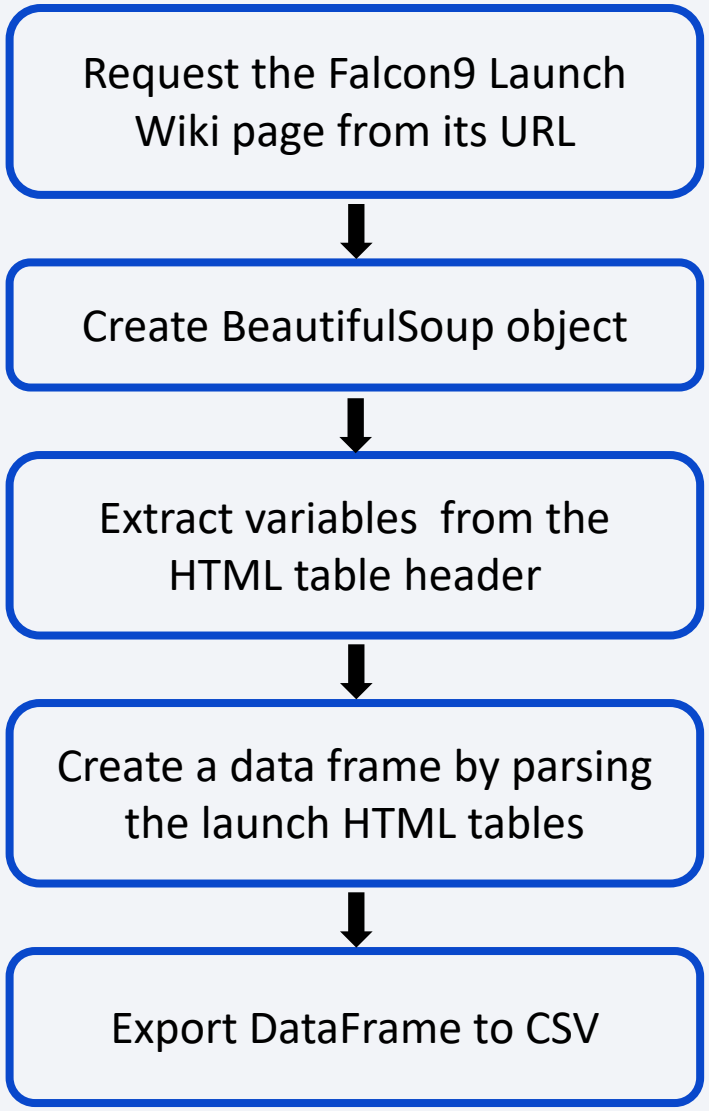
Data Collection – SpaceX API

- Data collection with SpaceX REST calls using key phrases and flowcharts
- GitHub URL of the completed SpaceX API calls notebook https://github.com/annguyenthanh96/SpaceX_Rocket_Lauch_Analysis/blob/master/Data%20Collection%20API.ipynb



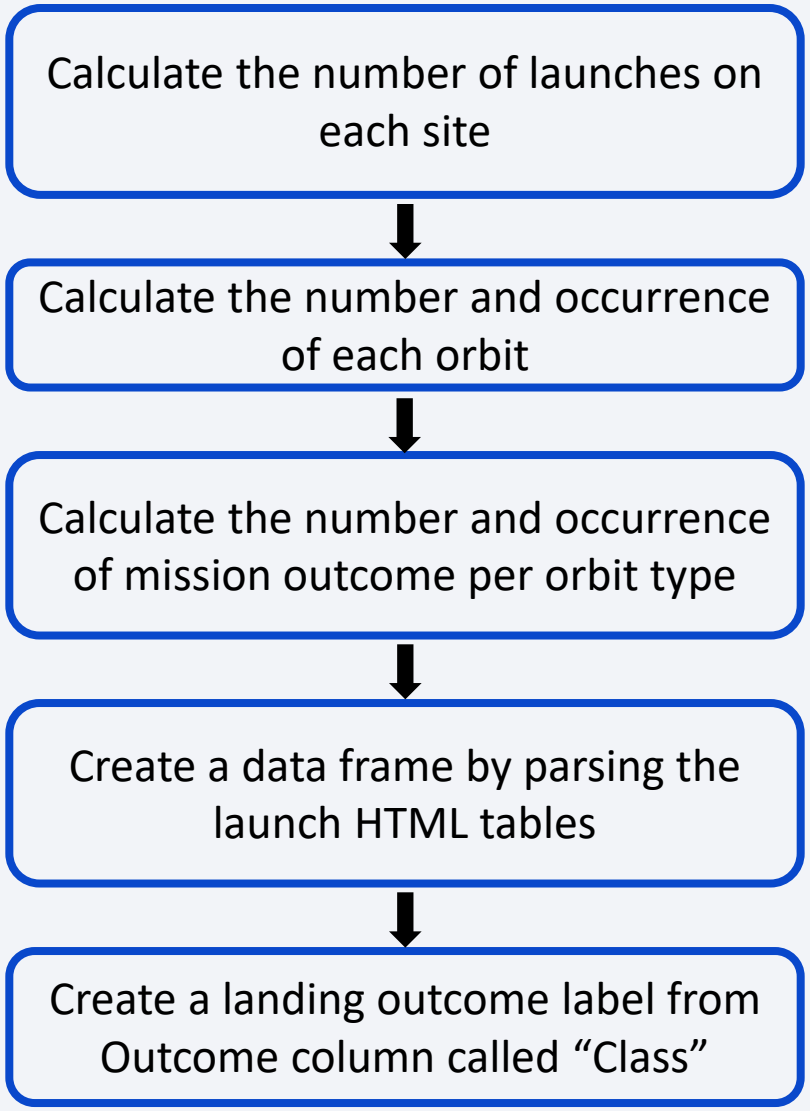
Data Collection - Scraping

- Web scraping process using key phrases and flowcharts
- GitHub URL of the completed web scraping notebook https://github.com/annguyenthanh96/SpaceX_Rocket_Lauch_Analysis/blob/master/Data%20Collection%20with%20Web%20Scraping%20lab.ipynb



Data Wrangling

- Create new labels (Column “Class”) to distinguish: if landing successful = 1 & failure = 0
- If the Outcome column says (True ASDS, True RTLS, & True Ocean), the label is 1
- If the Outcome label says (None None, False ASDS, None ASDS, False Ocean, False RTLS), the label is 0
- GitHub URL https://github.com/annguyenthanh96/SpaceX_Rocket_Lauch_Analysis/blob/master/Data%20Wrangling%20.ipynb



EDA with Data Visualization

- We used 3 types of plots

➢ Scatter plot: View relationship between variables Eg: FlightNumber vs. PayloadMass, FlightNumber vs LaunchSite, Payload and Launch Site, FlightNumber and Orbit type, Payload and Orbit type.

➢ Line chart: View trend by year. Eg: Visualize the launch success yearly trend

➢ Bar chart: Compare data of different groups. Eg: Visualize the relationship between success rate of each orbit type

- GitHub URL
  https://github.com/annguyenthanh96/SpaceX_Rocket_Lauch_Analysis/blob/master/EDA%20with%20Visualization.ipynb

EDA with SQL

- Summarize the SQL queries performed

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful 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 the names of the booster_versions which have carried the maximum payload mass
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for 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
  https://github.com/annguyenthanh96/SpaceX_Rocket_Lauch_Analysis/blob/master/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

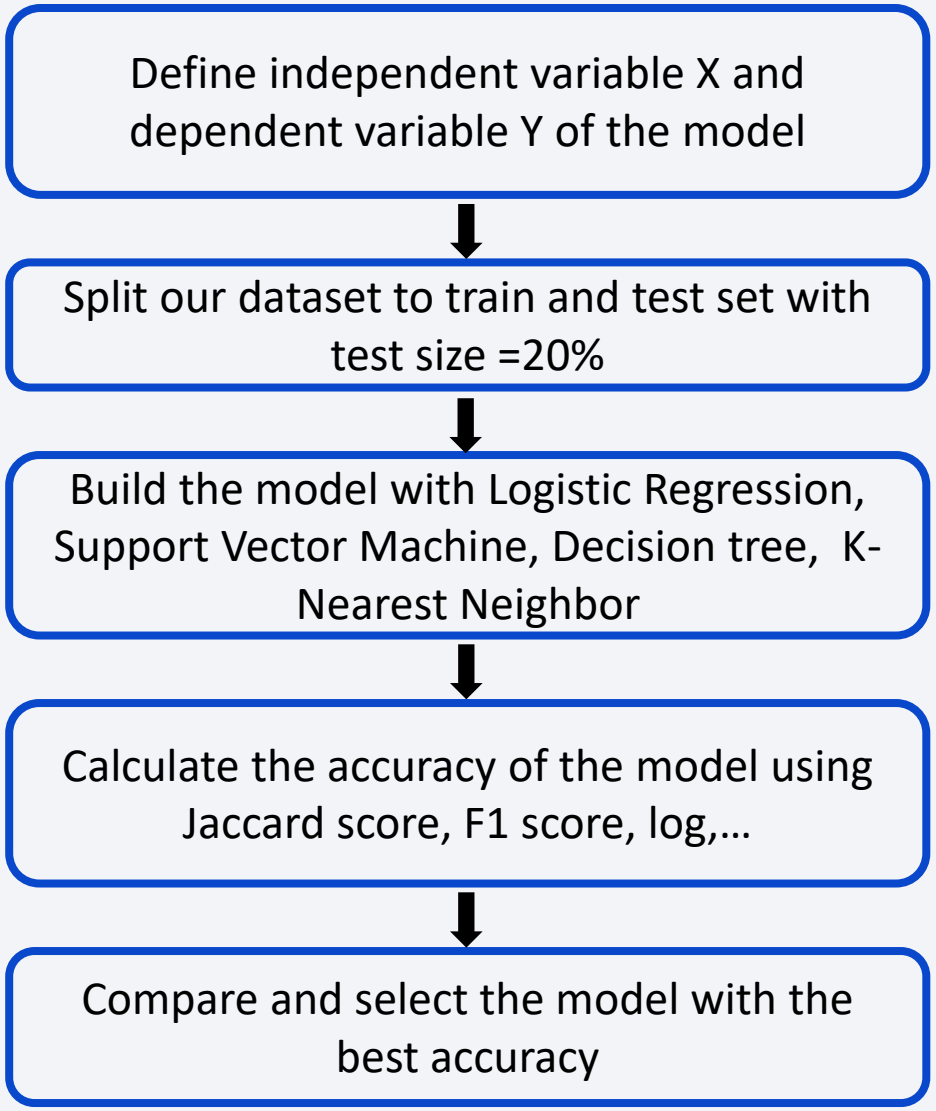
- First, we created a folium Map object, with an initial center location to be NASA Johnson Space Center at Houston, Texas
- We created folium.Circle object to add a highlighted circle area with a text label on a specific coordinate. For each launch site, add a Circle object based on its coordinate (Lat, Long) values.
- We created a MarkerCluster object and customize the Marker's icon property to indicate if this launch was successful or failed,
- We created a `folium.PolyLine` object using the coastline coordinates and launch site coordinate to measure the distance to coastline
- We created a marker with distance to measure distance to a closest city, railway, highway, etc.
- GitHub URL https://github.com/annguyenthanh96/SpaceX_Rocket_Lauch_Analysis/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- The pie chart helps determine the success rate for each launch site. We can select each launch site to see the result
- The scatter plot helps us observe how payload may be correlated with mission outcomes for selected site(s). We can select Payload range from 0 to 10,000 for each site
- GitHub URL https://github.com/annguyenthanh96/SpaceX_Rocket_Lauch_Analysis/blob/master/Build%20an%20Interactive%20Map%20with%20Folium.ipynb

Predictive Analysis (Classification)

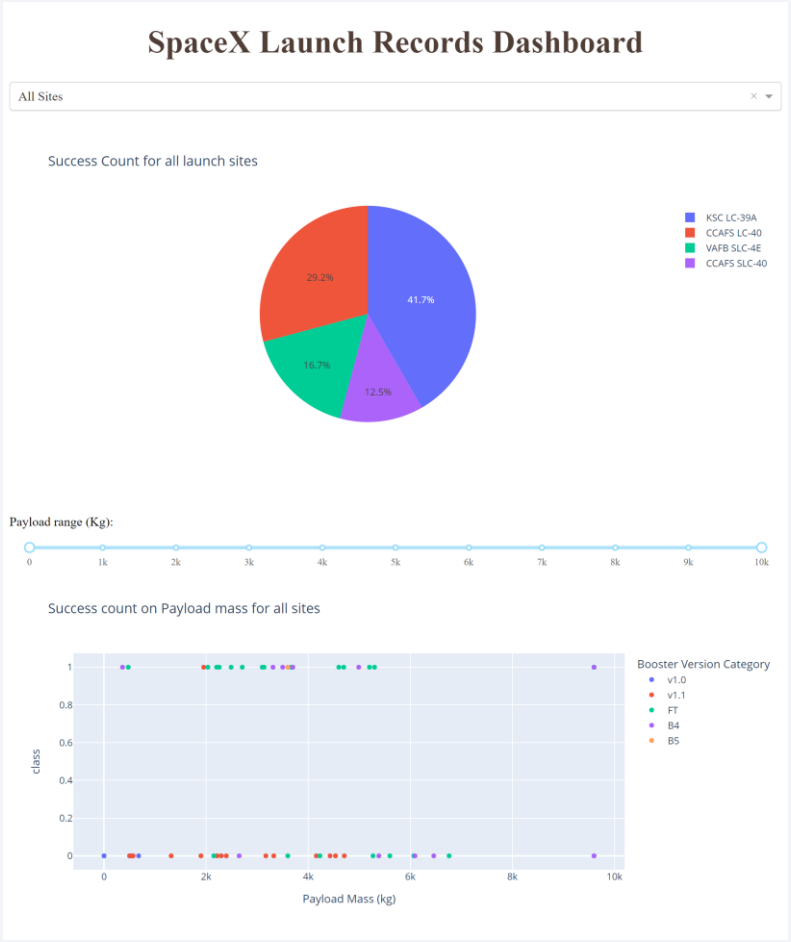
- We created serval machine learning models (Logistic Regression, Decision Trees, K- Near Neighbor)to predict the success of the the First Stage landing to determine what is the best prediction model
- GitHub URL https://github.com/annguyenthanh96/SpaceX_Rocket_Lauch_Analysis/blob/master/Machine%20Learning%20Prediction.ipynb

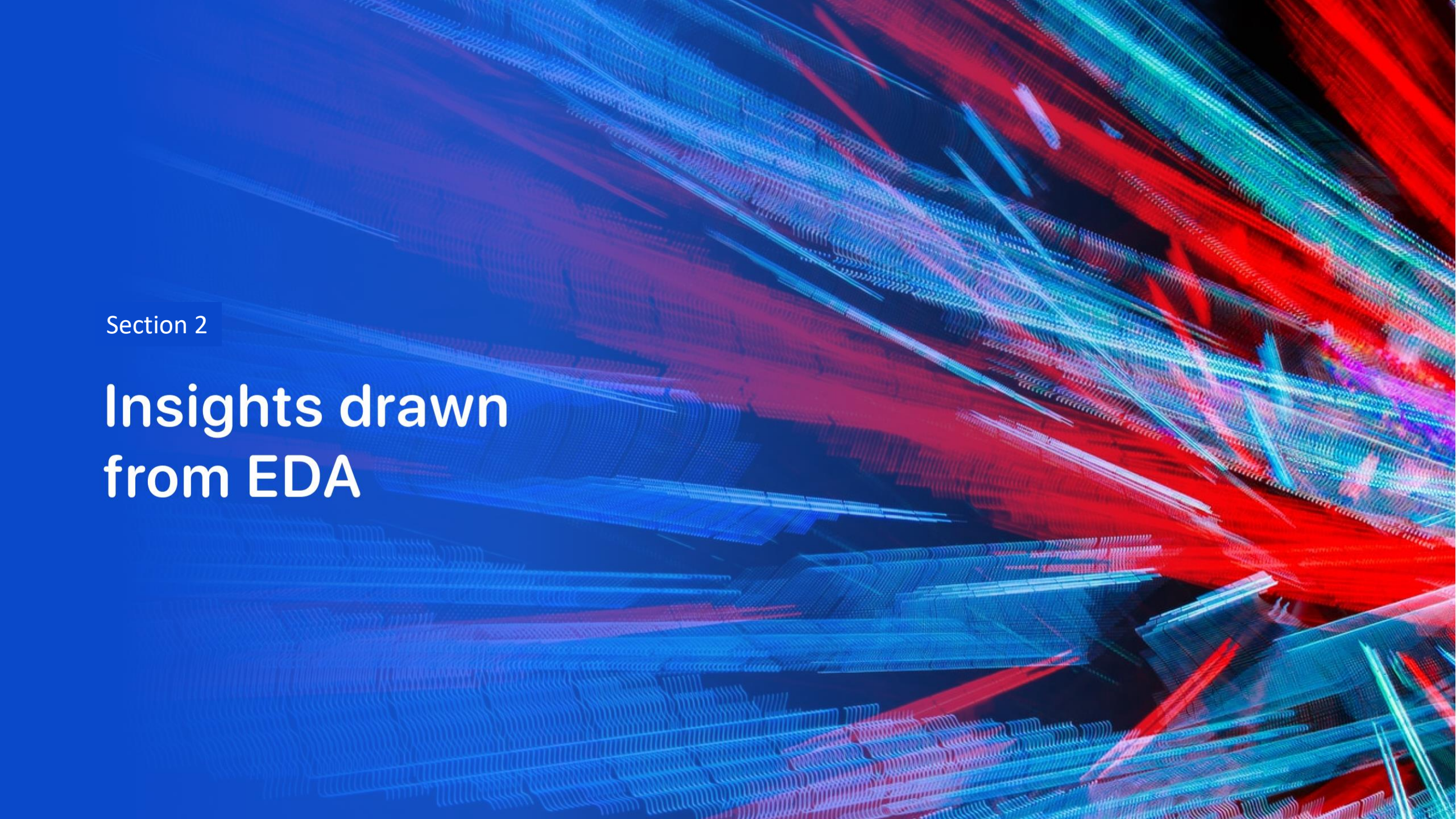


Results

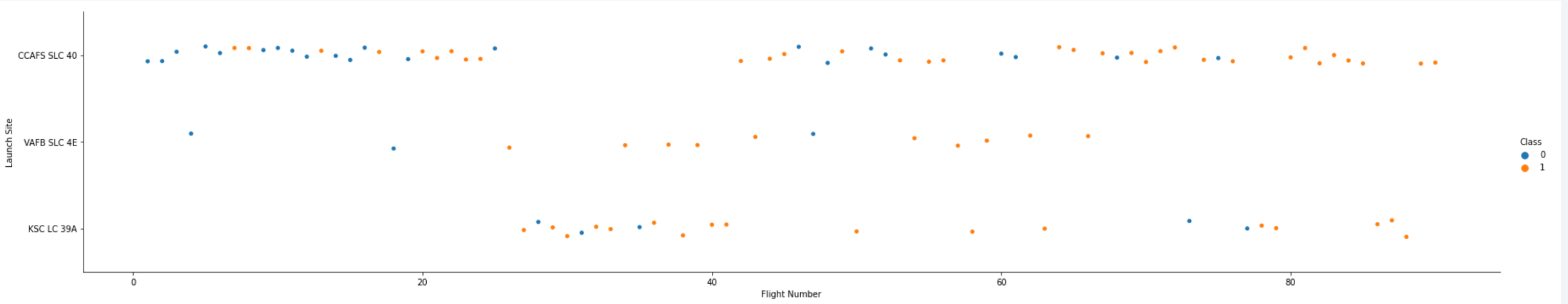
We found that Decision Tree is the most promising model because it produces the highest accuracy (94% accuracy compared 83% accuracy of other models)

Interactive analytics demo screenshots



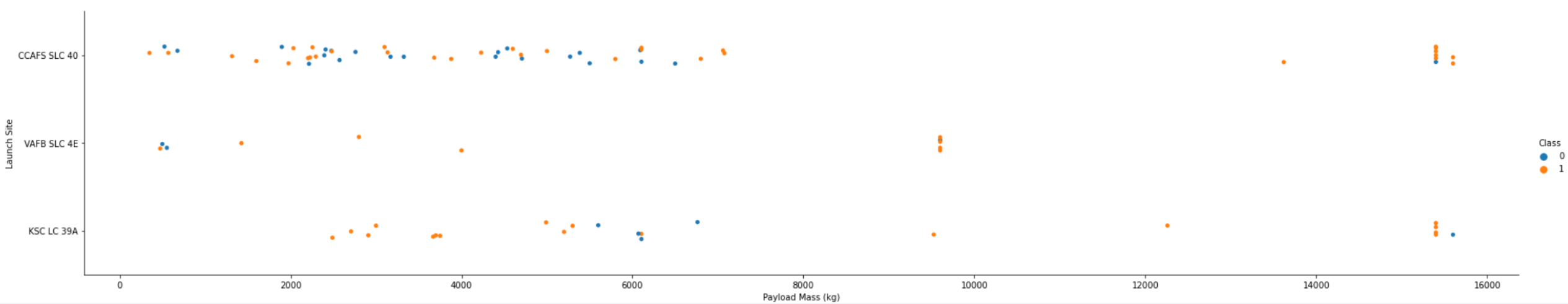

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- In this graph, Blue means a unsuccessful launch, Orange means a successful launch
- CCAFS SLC 40 is has the most launches

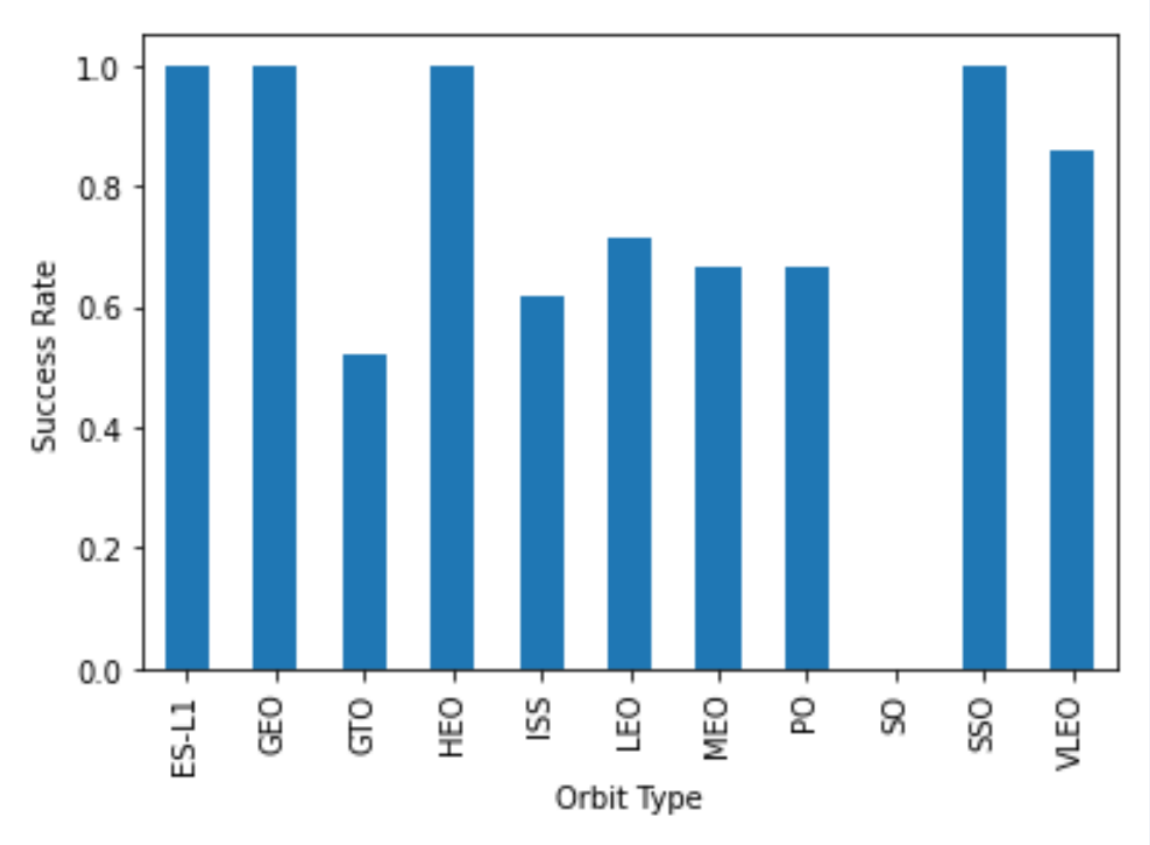
Payload vs. Launch Site



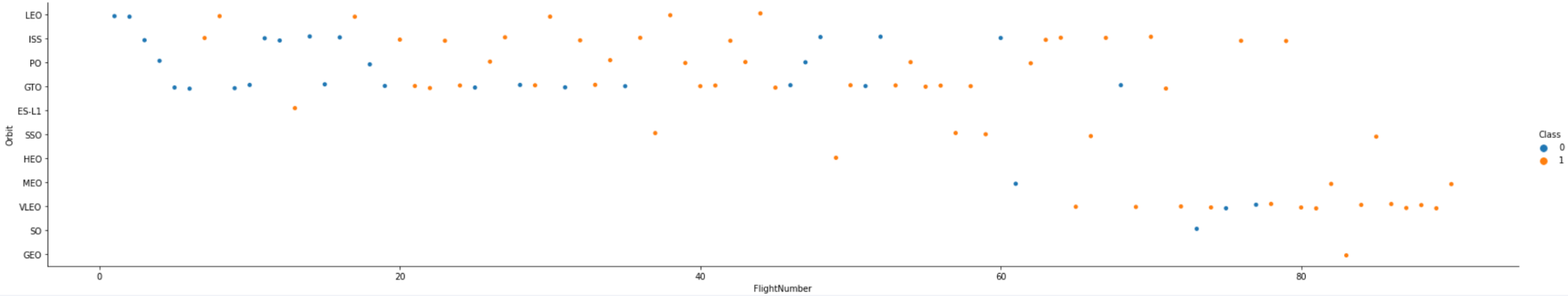
- In this graph, Blue means a unsuccessful launch, Orange means a successful launch
- Payload is centered around 1000-7000 kg

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO are the most successful Orbit Type with 100% success rate

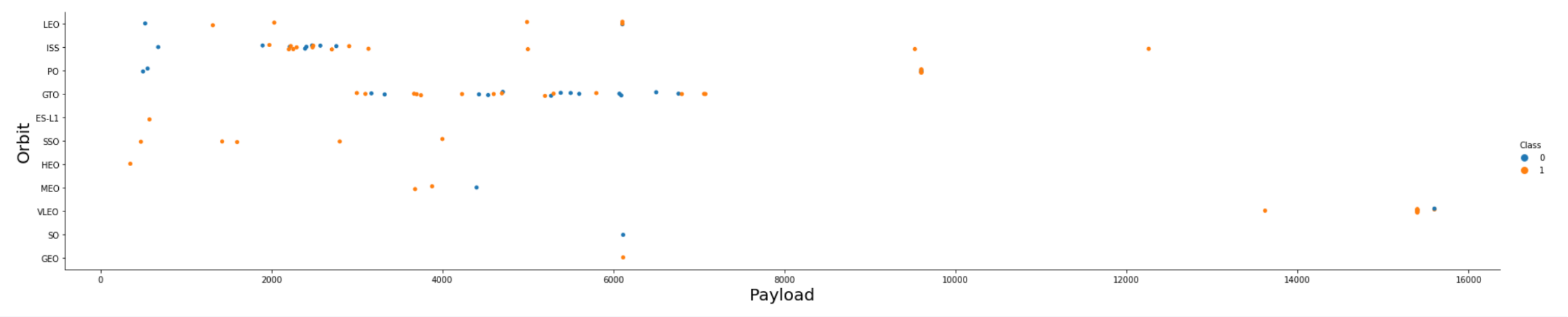


Flight Number vs. Orbit Type



- In this graph, Blue means a unsuccessful launch, Orange means a successful launch
- Orbit preference changes over number of flight

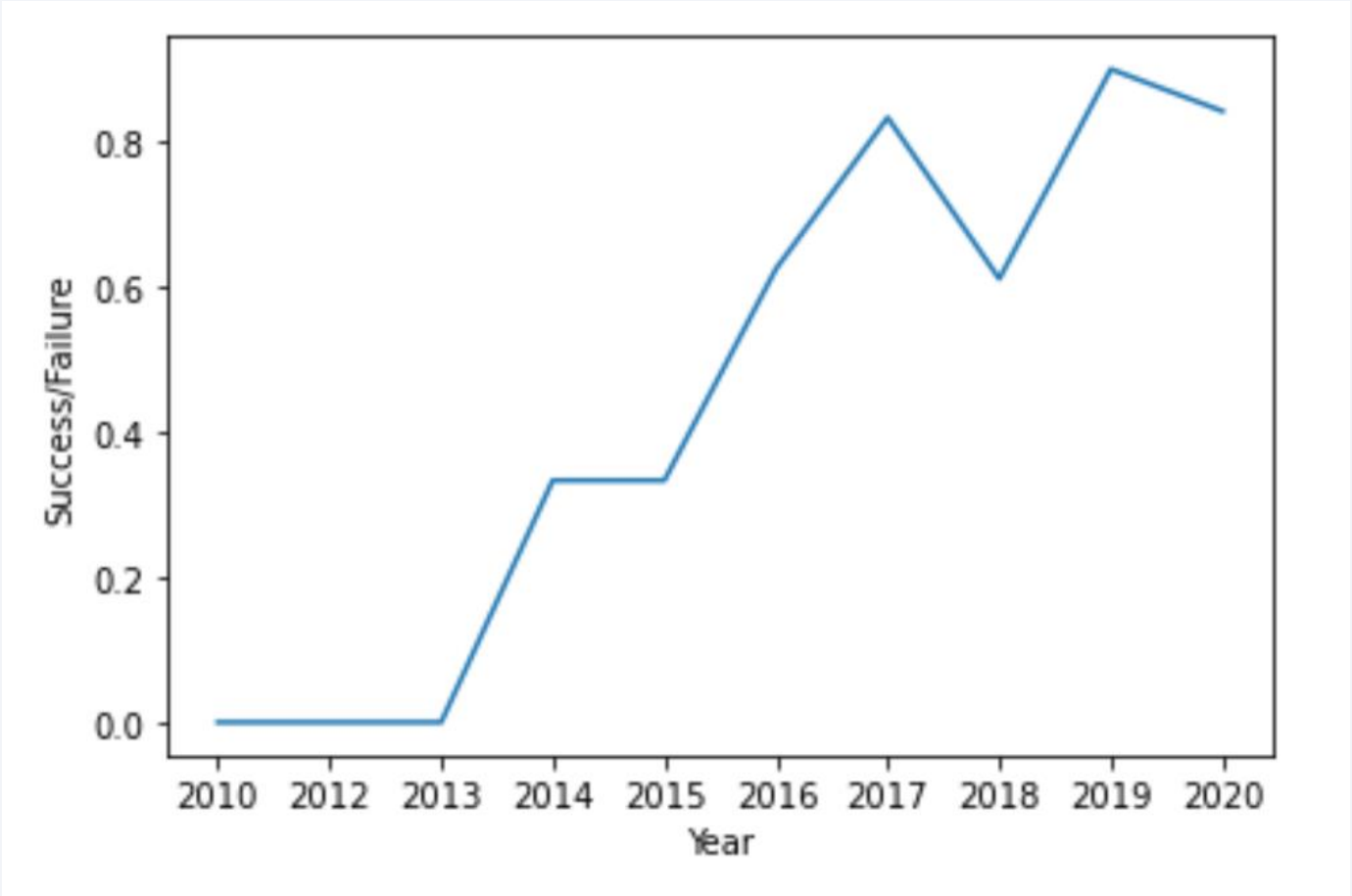
Payload vs. Orbit Type



- In this graph, Blue means a unsuccessful launch, Orange means a successful launch
- SSO seems to have the highest success rate with low to medium payload mass

Launch Success Yearly Trend

The success rate has an increasing trend since 2013 although there was a strong decrease in 2018



All Launch Site Names

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

Query: SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL;

There are 4 launch sites although there are confusing between CCAFS LC-40 and CCAFS SLC-40

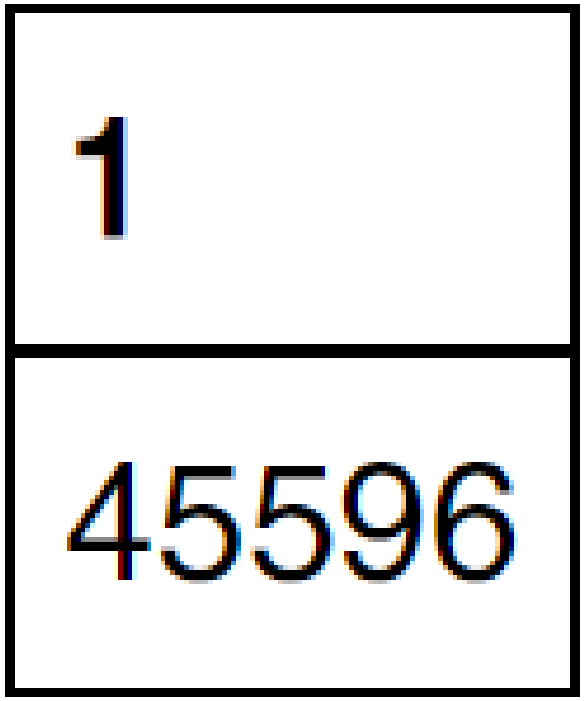
Launch Site Names Begin with 'CCA'

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Query: SELECT LAUNCH_SITE
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;

Retrieved 5 entries in database with Launch Site name beginning with CCA

Total Payload Mass



Query: SELECT

SUM(PAYLOAD_MASS__KG_)

FROM SPACEXTBL

WHERE Customer = 'NASA (CRS)';

The total Payload Mass is 45596

Average Payload Mass by F9 v1.1

1
340

Query: SELECT

AVG(PAYLOAD_MASS__KG_)

FROM SPACEXTBL

WHERE Booster_Version LIKE 'F9 v1.0%';

The average Payload Mass is 340

First Successful Ground Landing Date

1
2015-12-22

Query: SELECT MIN(Date)
FROM SPACEXTBL
WHERE Landing__Outcome = 'Success (ground pad)';

The first successful launch date was 22-12-2015

Successful Drone Ship Landing with Payload between 4000 and 6000

booster_version
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1029.2
F9 FT B1038.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Query: SELECT BOOSTER_VERSION

FROM SPACEXTBL

WHERE LANDING__OUTCOME = 'Success (drone ship)'

AND 4000 < PAYLOAD_MASS__KG_ < 6000;

There are 7 booster versions

Total Number of Successful and Failure Mission Outcomes

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Query: SELECT MISSION_OUTCOME,
COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;

99% of outcome is successful

Boosters Carried Maximum Payload

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

Query: SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
SELECT MAX(PAYLOAD_MASS__KG_)
FROM SPACEXTBL);

There are 12 boosters that carried maximum payload of 15600

2015 Launch Records

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Query: SELECT LANDING__OUTCOME,
BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE Landing__Outcome = 'Failure (drone ship)'
AND YEAR(DATE) = 2015;

There were failed 2 Drone Ship Landing Records

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

landing__outcome	total_number
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

Query: SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS

TOTAL_NUMBER

FROM SPACEXTBL

WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'

GROUP BY LANDING__OUTCOME

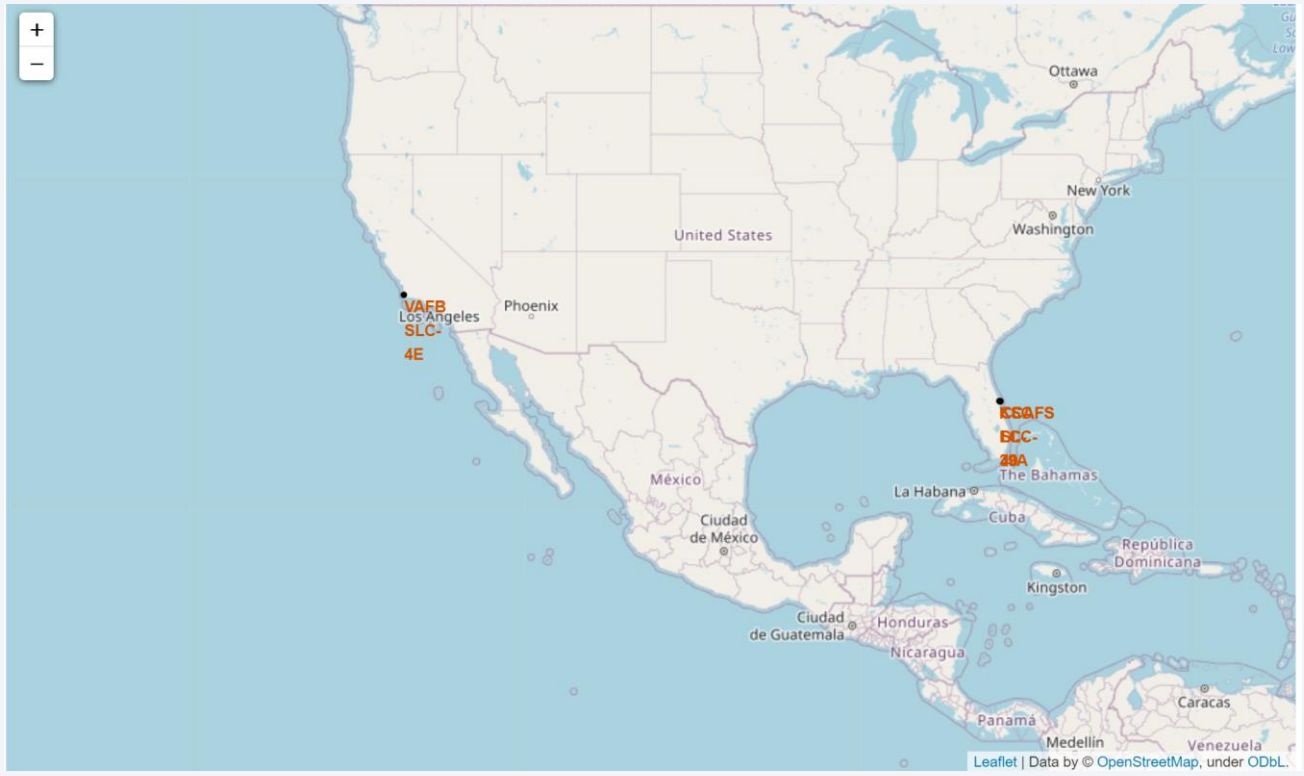
ORDER BY TOTAL_NUMBER DESC

Between the date 2010-06-04 and 2017-03-20, the Success and Failure of Drone Ship was equal

Section 3

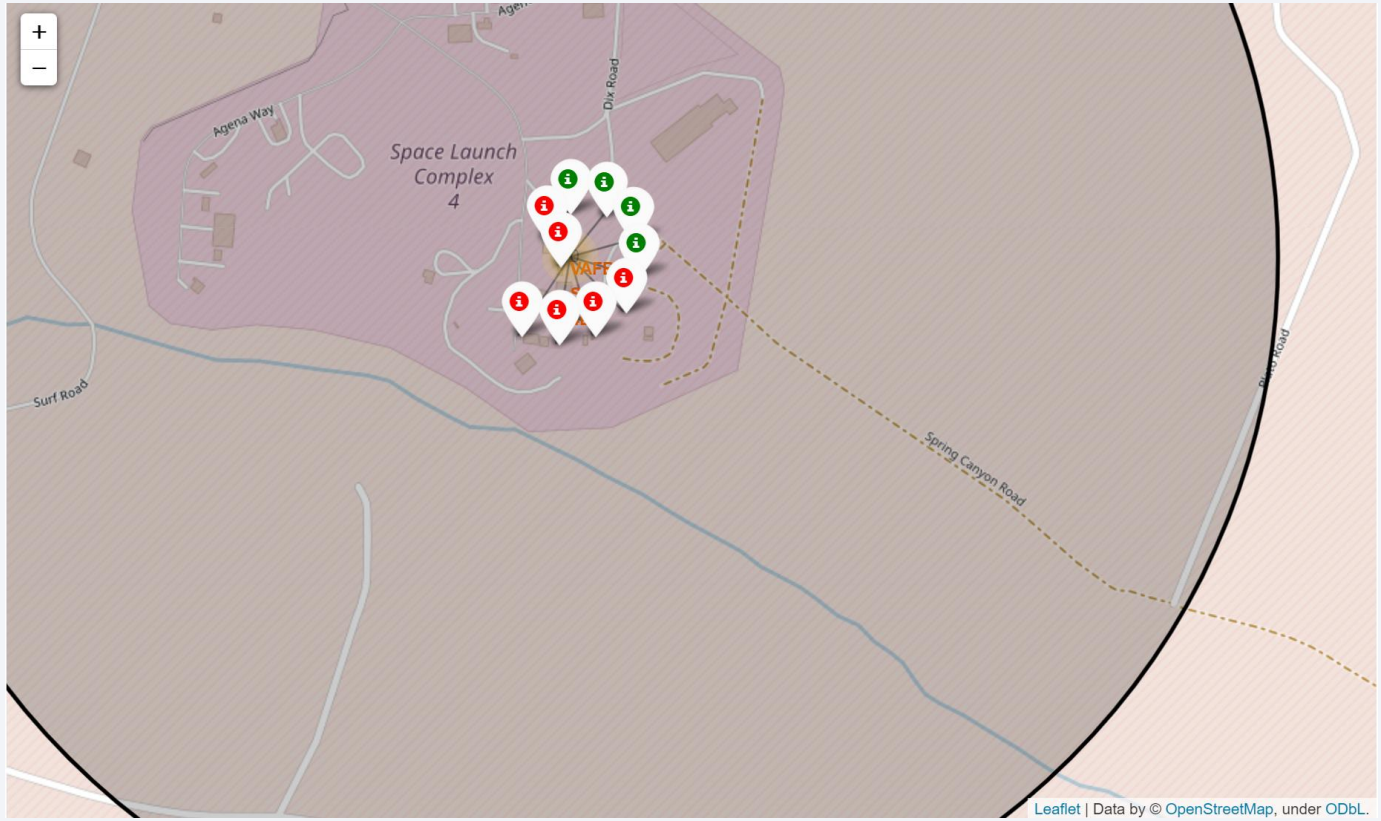
Launch Sites Proximities Analysis

Record All The Launch Sites



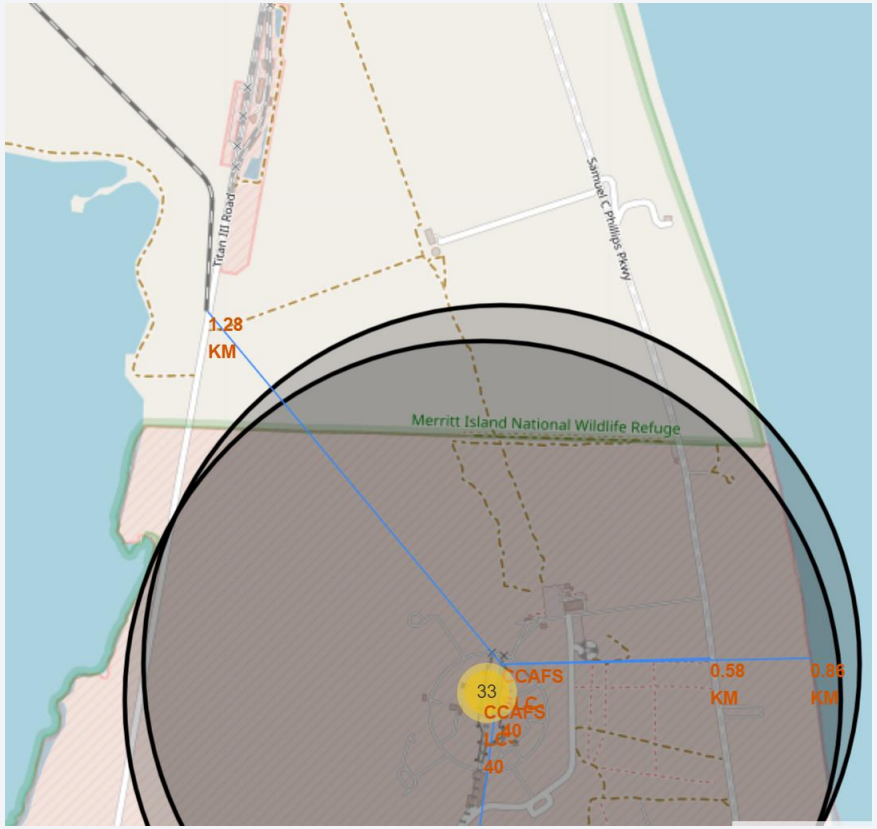
Launch sites are all in the US and close to the coast

Launch Outcomes

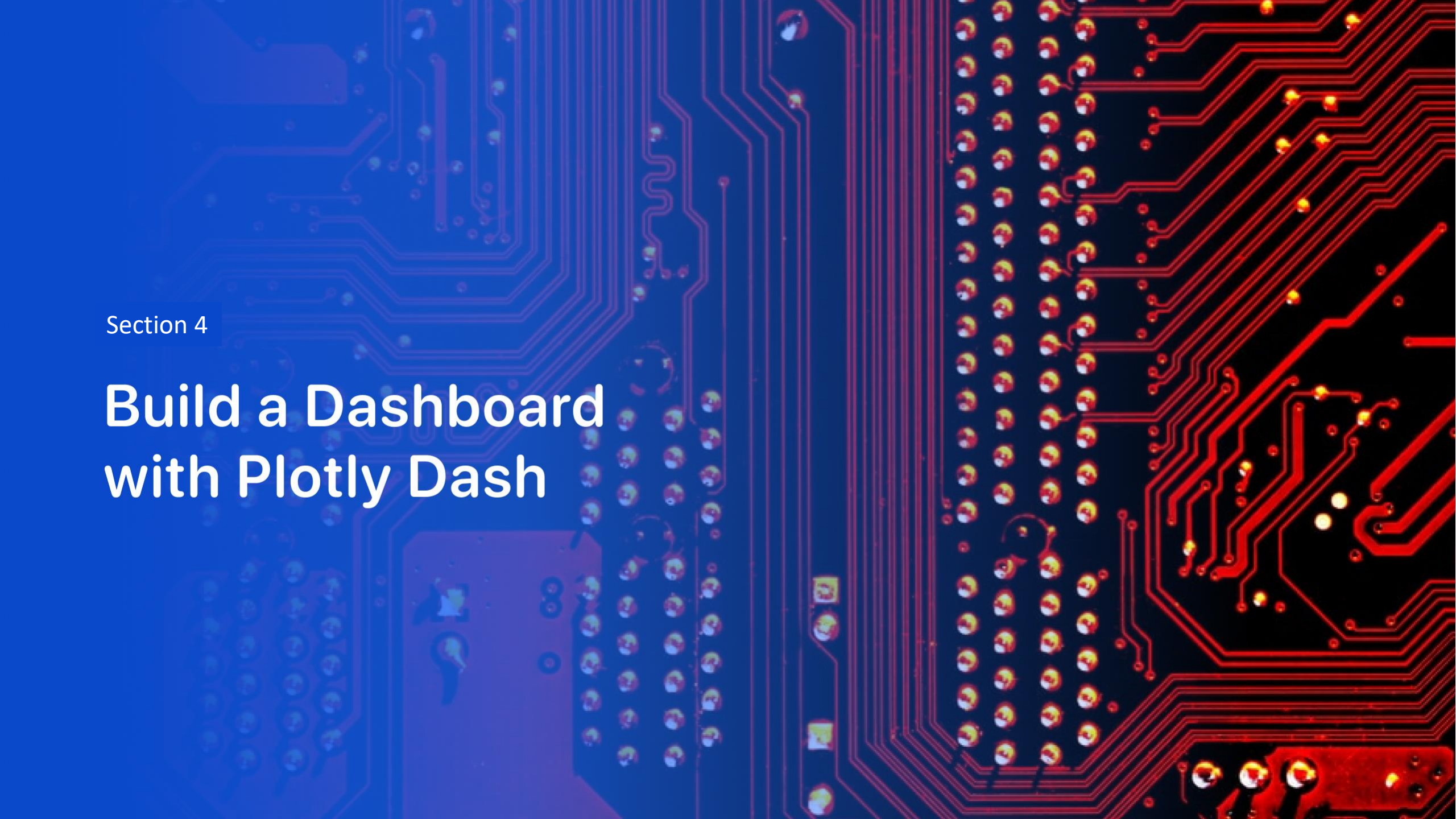


- In this graph, successful landing is labeled green icon and failed landing is labeled with red icon.
- From the graph, we can see that VAFB SLC-4E has 4 successful landings and 6 failed landings

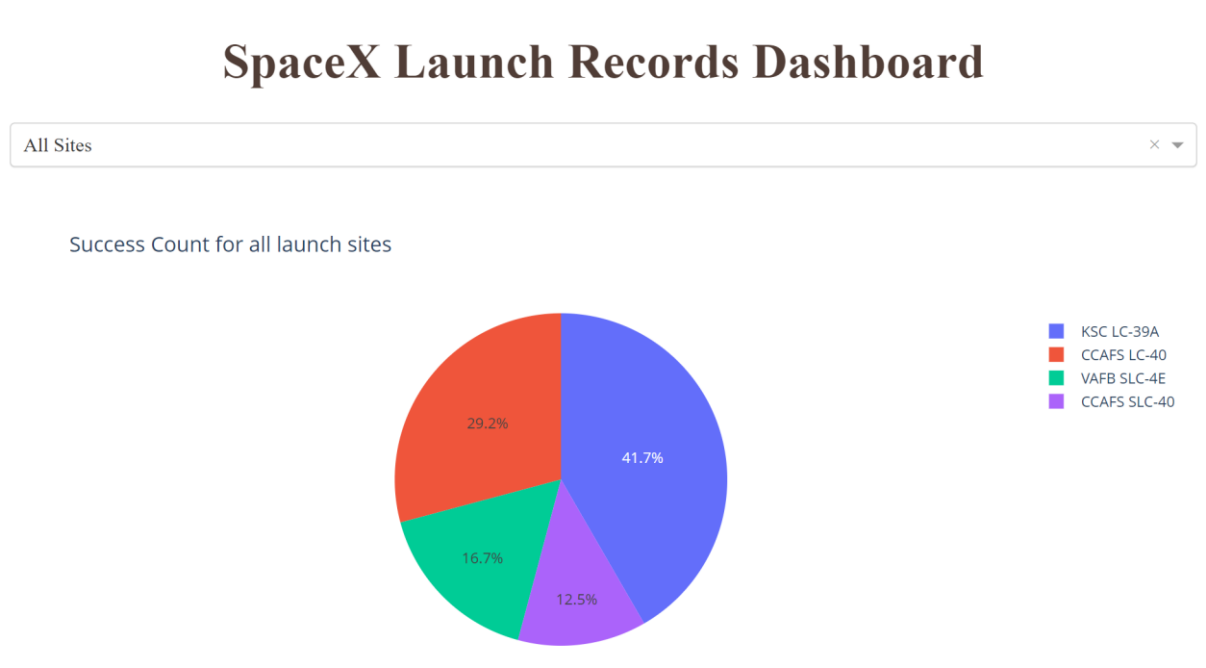
Calculate Distance From Launch Site



- Launch sites are not close to highways
- Launch sites are not close to railway
- Launch sites are close to the coast line
- Launch sites keep distance from the cities

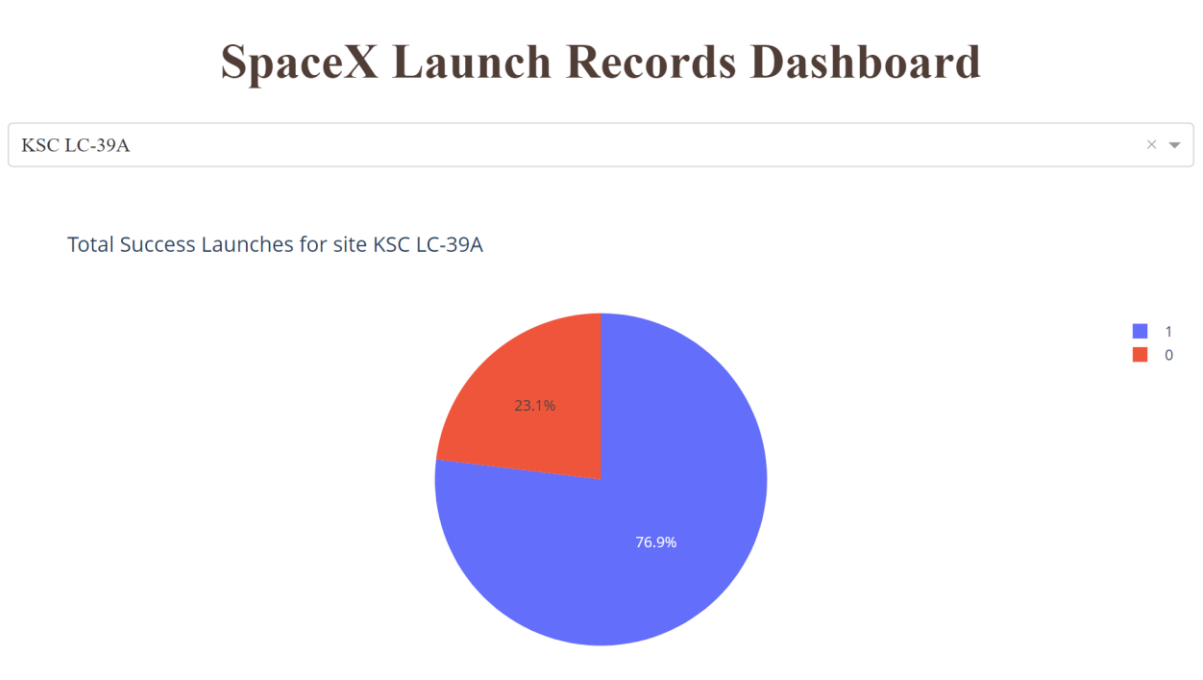

Section 4
Build a Dashboard with Plotly Dash

Successful Launch by All Sites



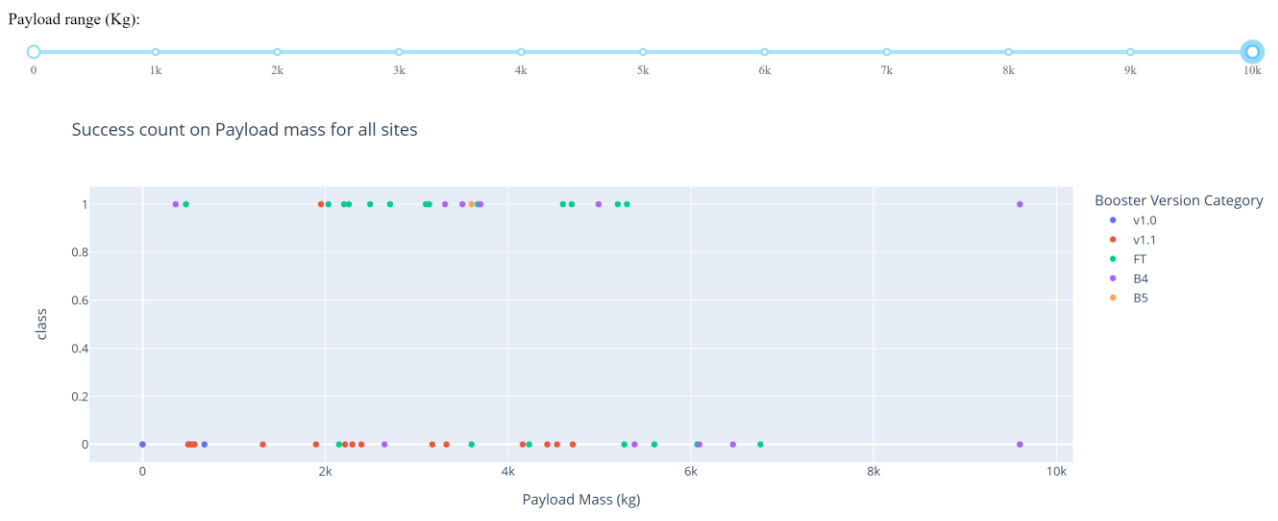
KSC LC-39A had the most successful launches from all the sites

Launch Site with the Highest Launch Success Ratio



- The best site is KSC LC-39A
- Success rate is 76.9% while getting a
- Failure rate is 23.1%

Payload vs. Launch Outcome for All Sites



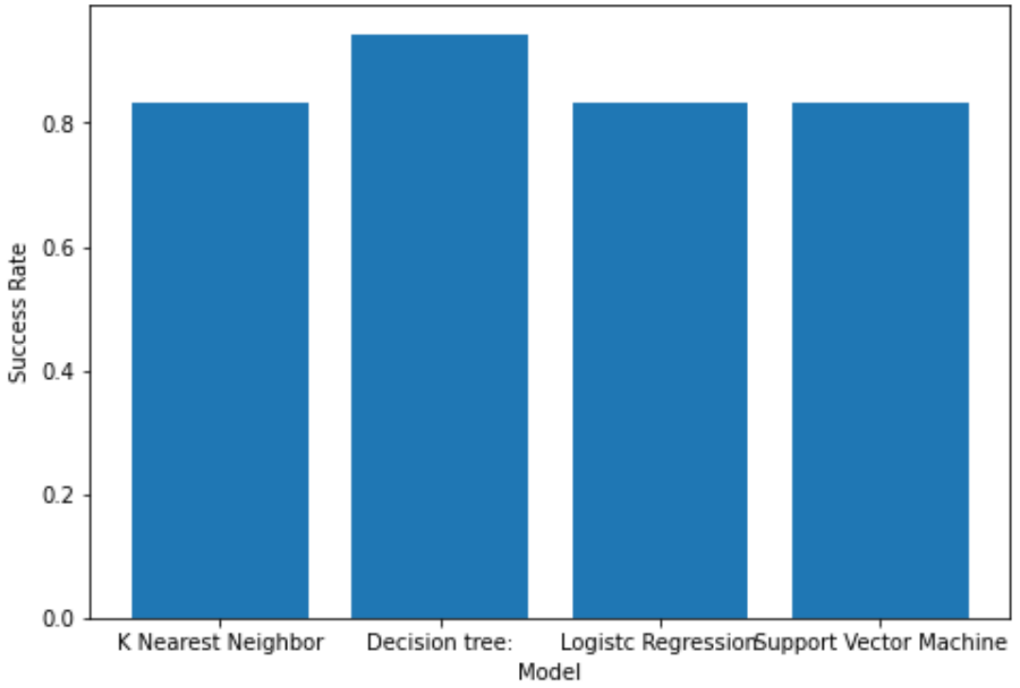
- Class indicates 1 for successful landing and 0 for failure
- Most of our data in the low to medium range payload
- The number of unsuccessful launch is slightly than successful launch in selected range

Section 5

Predictive Analysis (Classification)

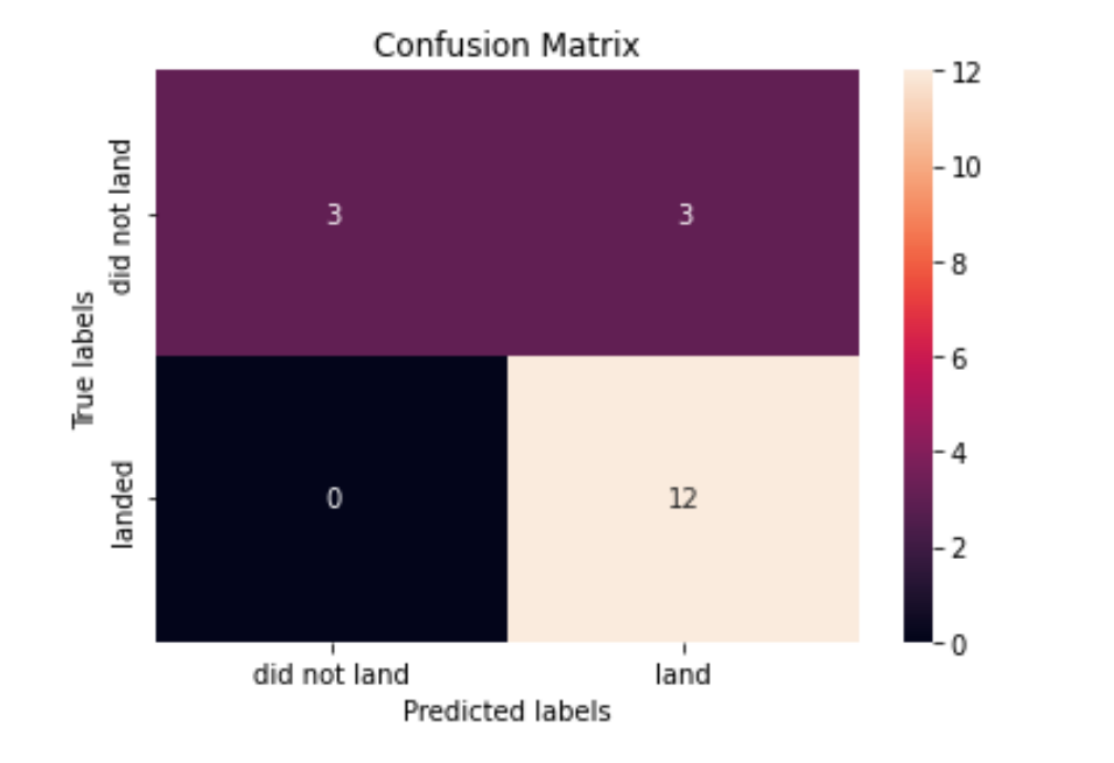
Classification Accuracy

- Decision Tree is the best model with the accuracy score of 0.94
- Other models including K Nearest Neighbor, Logistic Regression, Support Vector Machine has the same accuracy score of 0.83




```
Compare test accuracy
K Nearest Neighbor: 0.8333333333333334
Decision tree: 0.9444444444444444
Logistc Regression: 0.8333333333333334
Support Vector Machine: 0.8333333333333334
```

Confusion Matrix



- The models predicted 12 successful landings
- The models predicted 3 unsuccessful landings
- The models predicted 3 successful landings when the label was unsuccessful landings (false positives)

Conclusions

- We can see that KSC LC-39A is the most successful launch site
- GEO,HEO,SSO,ES-L1 are the best orbits to launch
- Decision tree is the best Machine Learning model that our company, Space Y can use for prediction

Thank you!