IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Marc Robert
Data Scientist
Space Y
November 29, 2023

Outline

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

Executive Summary

- Summary of methodologies

In order for Space Y to succeed in an ever-increasing competitive market for commercialized space travel, a thorough analysis was undertaken to review the launch data publicly available for SpaceX. The analysis will derive key information to support Space Y's bidding process;

- Collect SpaceX launch data from the SpaceX REST API and the Wiki page for List of Falcon 9 and Falcon Heavy launches
- Perform exploratory data analysis with SQL to analyze and investigate data, to uncover key insights, data patterns, and address any data quality issues.
- Determine relevant column as features and identified the hyperparameters that are most relevant to predicting successful landings.
- Build an interactive visual analytics in Folium and Plotly, to allow users to interact with key variables.
- Build Test Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors models to predict if the first stage of the Falcon 9 lands successfully

Introduction

Project background

The commercial space age is here, and companies are making space travel affordable for everyone. The competitive landscape is filling with entrants such as Virgin Galactic, Rocket Lab, Blue Origin and of course SpaceX. The cost of launching rockets continues to be the key factor in a sustainable business model. Currently, Space X has an ever-improving successful launch rate, while drastically reducing the overall mission costs significantly. The Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each. The SpaceX Falcon 9 rocket is designed to potentially reuse the first stage rocket, which is one of the major cost components to a rocket.

Project Objectives

Space Y that would like to compete with SpaceX. The ability for Space Y to determine the cost of a particular SpaceX launch (including the prediction if the SpaceX will likely successfully land the first stage rocket) will allow us to competitively bid against SpaceX for a rocket launch based on cost.

The objectives of the analysis will be to derive the following;

- Provide insight to estimate price of each SpaceX launch in order to better understand the SpaceX mission cost model
- Build and train a machine learning model and leverage public information to predict if SpaceX will reuse the first stage.
- Provide additional insights based on location, Rocket models to provide success rates

Section 1

Methodology

Methodology

Executive Summary

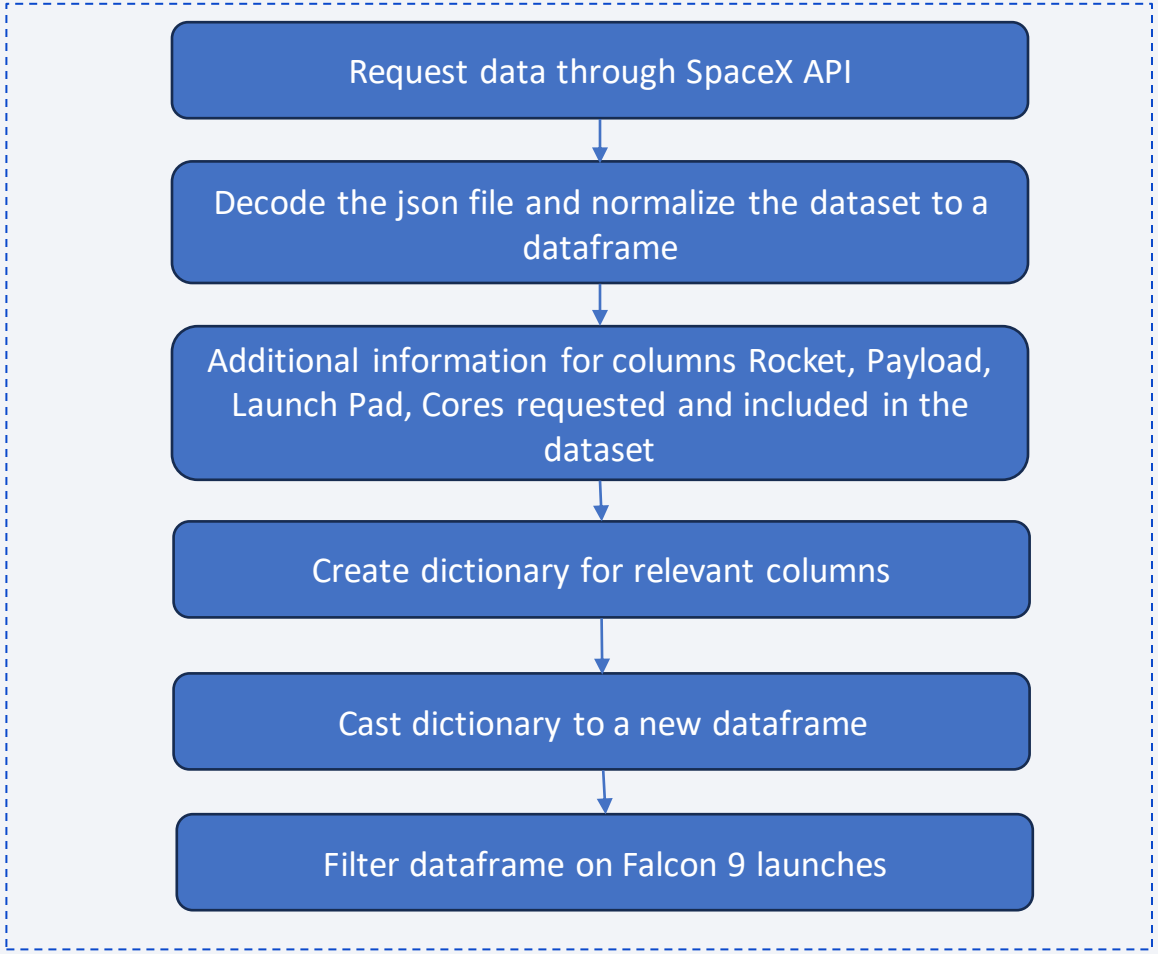
- Data collection methodology:
  - Data is collected from public sources available (SpaceX API, Wiki pages)
- Perform data wrangling
  - Reviewed data, perform cleansing of missing data, and transformed categorical data to a numeric outcome variable
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
- Build, tune, evaluate classification models
  - The use of GridSearchCV was used to tune all four models built

Data Collection

- Two primary data sets were collected to conduct this analysis. Both dataset were from the public domain and accessed through two key methods
  - API Connections - the SpaceX public website, with API call “api.spacexdata.com/v4/launches/past”
  - Webscrapping - the Wiki page containing a Falcon 9 launch data table was scrapped and loaded to dataset from the following website https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- The API request call extracts historical launch data from the SpaceXdata.com site.
- The data contains launch historical launch information for all SpaceX missions. There are 17 core data columns and 200 requests returned in the request call
- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

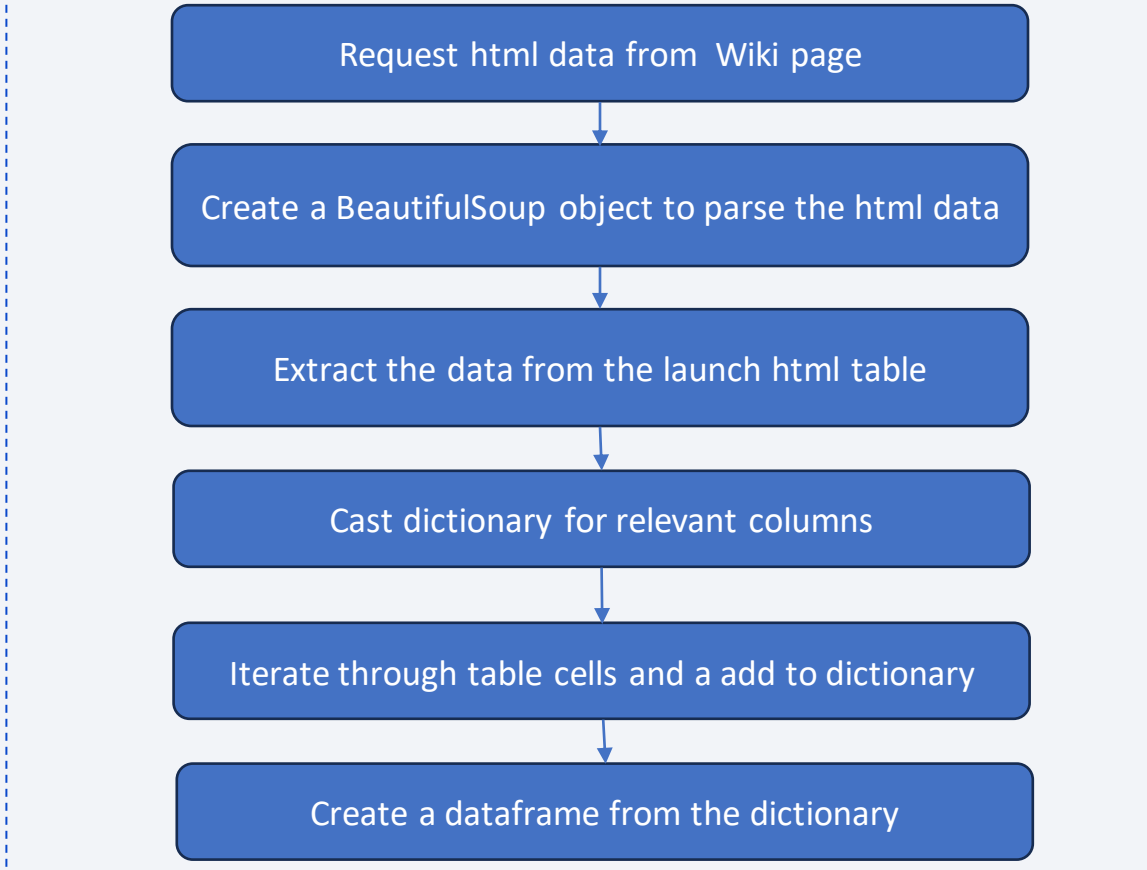


Data Collection - Scraping

- The historical launch data is located in a table on the following Wikipedia page

➢ https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose



Data Wrangling

- Data was reviewed for data completion. Two columns contained null values and required review.
  - 26 null values for Landing Pad was determined to be valid value
  - Payload Mass contained 5 null records, and was treated by replacing the null value with the mean of the Payload Mass to ensure continuity
- A training label called Class was added to the dataset to create a binary value for Landing outcomes (Successful (1) or Failure (0))
  - Class = 1 where Landing_Outcome in (True ASDS, True RTLS, True Ocean)
  - Class = 0 where Landing_Outcome in (None None, False ASDS, None ASDS, False Ocean, False RTLS)

- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

EDA with Data Visualization

- Scatterplots were used to visualize the relationship between various variables
  - Flight Number vs Payload
  - Flight Number vs Launch site
  - Flight Number vs Orbit
  - Payload vs Orbit
- Bar Chart was created to evaluate the relationship of Success Rate and Orbit
- Line Chart was created to understand the Average Success Rate by Year to better understand if the trend in success rate over time
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose

EDA with SQL

- Created connection to DB2 environment
- Created the SpaceX table and loaded data from a panda data frame
- Ran various queries using Python SQL integration, to better explore the data
    - Launch Sites
    - Mission Outcomes
    - Payloads sized
    - Booster versions
    - Landing Outcomes
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

Build an Interactive Map with Folium

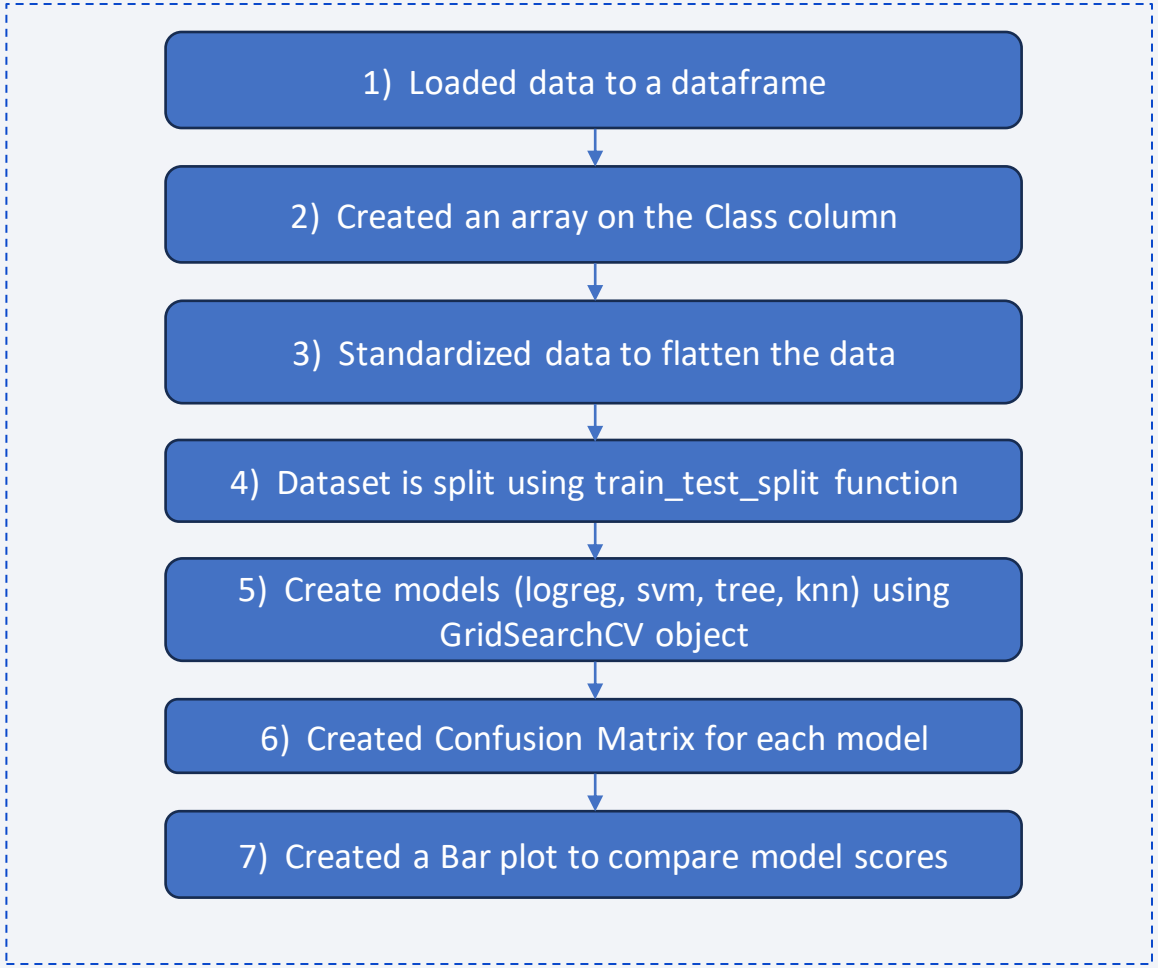
- On the visualization we added the following;
  - Created circle objects for each launch site, including popup labels
  - Create a marker cluster object to show success/failure of launches at launch sites
  - Create distance calculator between geo positions, allowing the distance to be added to polylines as labels
  - Added polylines between different geo positions on the map
- The addition of map objects, and visual tags (like the count of successful/unsuccessful launches at a site) allows users to quickly understand data and results. The ability to overlay data visually provide a depth of information that is extremely impactful
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

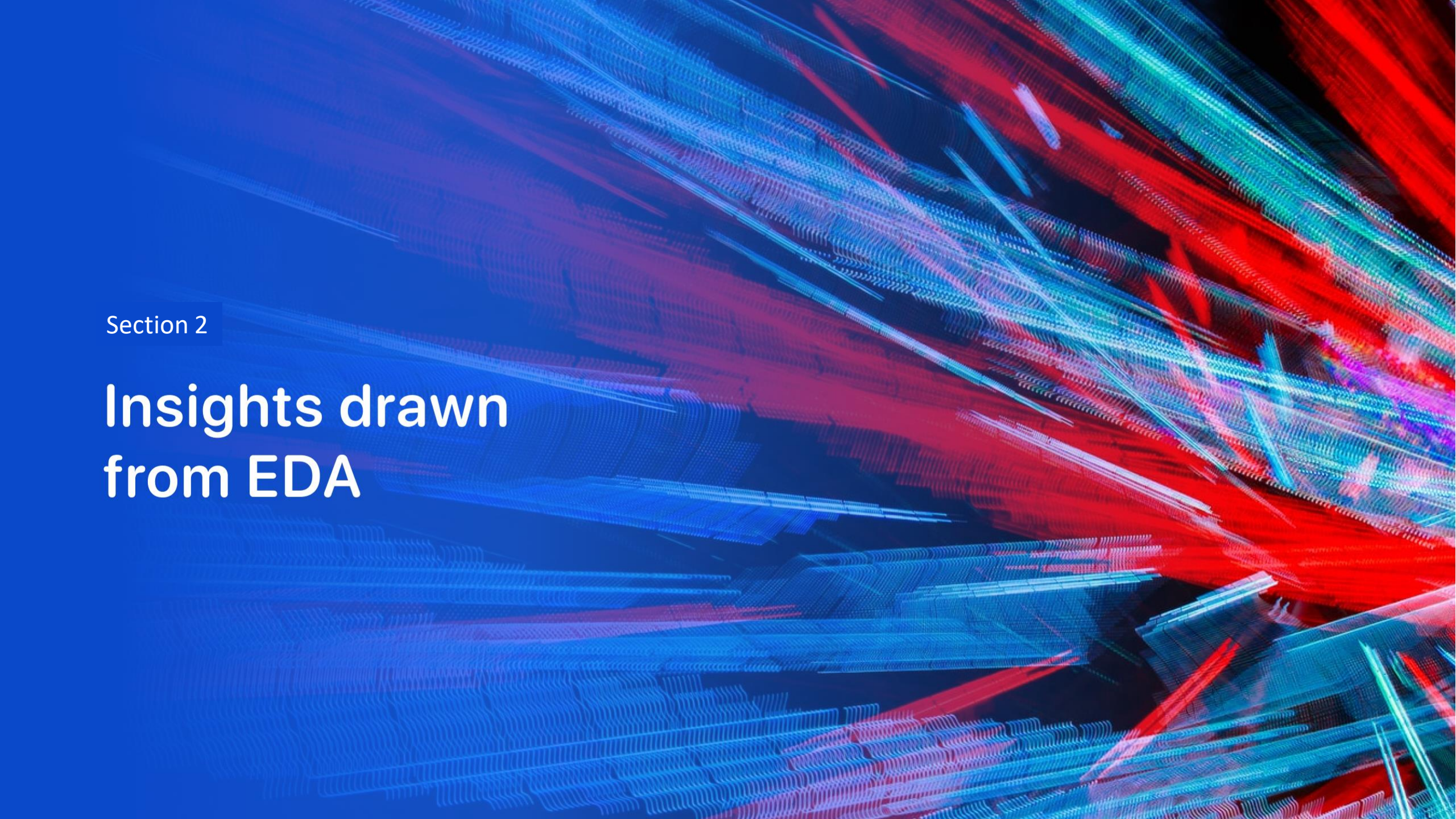
Build a Dashboard with Plotly Dash

- The dashboard allows the user to interact with the data in a dynamic setting. The user can explore the relationships and focus their queries to explore unique characteristics.
  - Dropdown list to allow users to select the launch site to display (default is all).
  - A pie Chart to display the success rate of the select site(s).
  - A slider object to filter results the payload range.
  - A scatterplot showing the correlation between payload and launch success.
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

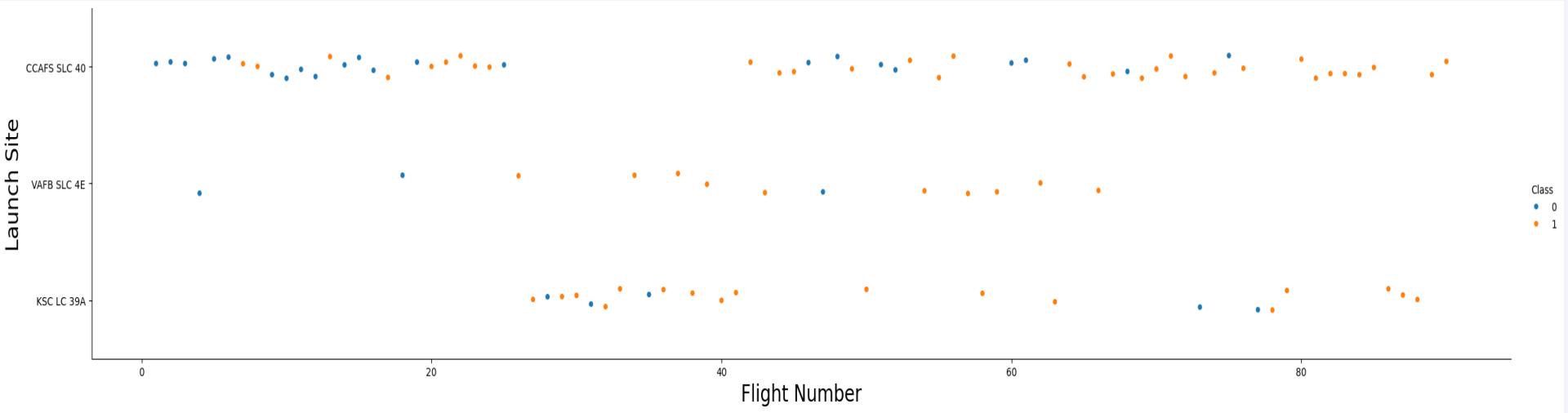
- The models built share the initial 4 steps in the flow.
- Steps 5 and 6 are repeated to build each model.
- Step 7 will compare the 4 model scores to determine is there is a clear winner
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose




Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- CCAFS SLA 40 is by far the most used launch site.
- There has been success and failure at each of the launch sites
- Early Flight Numbers had a high failure rate, while later Flight Numbers have a higher rate of success.

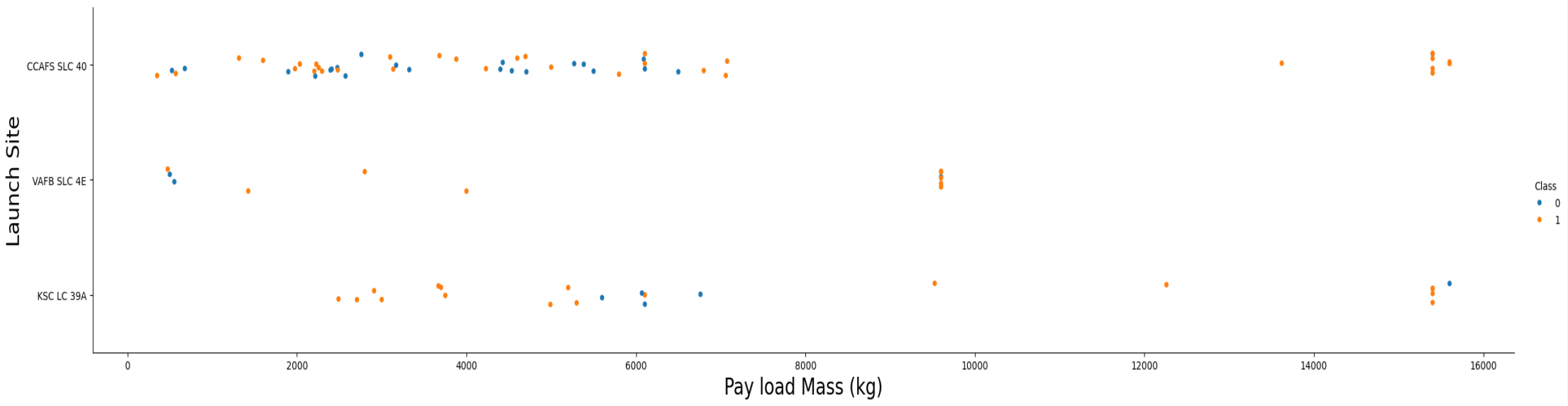


Success = 1 (Orange)

Failure = 0 (Blue)

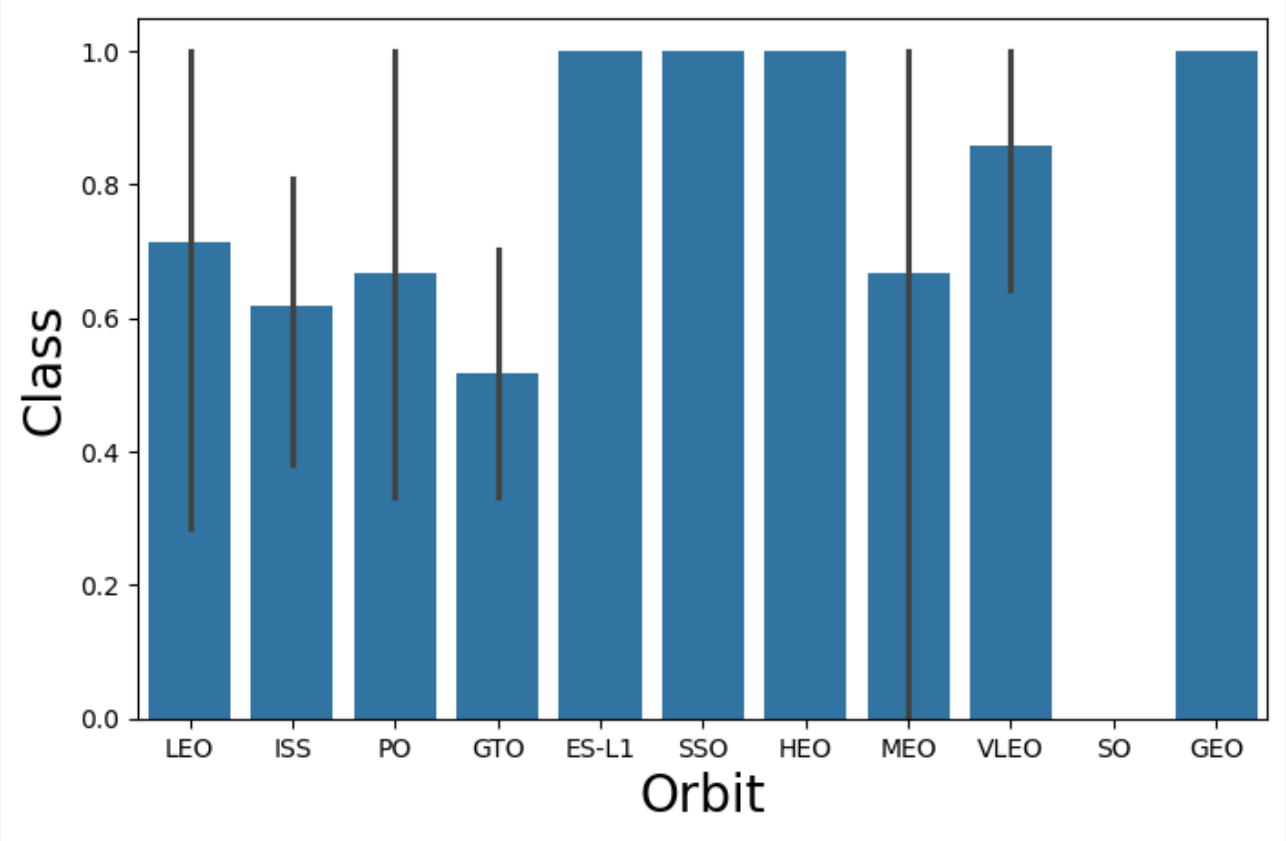
Payload vs. Launch Site

- Most payloads are between 2000 – 6000 Kg
- There has a high success rate of missions with Payload >8000 kg
- VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000)



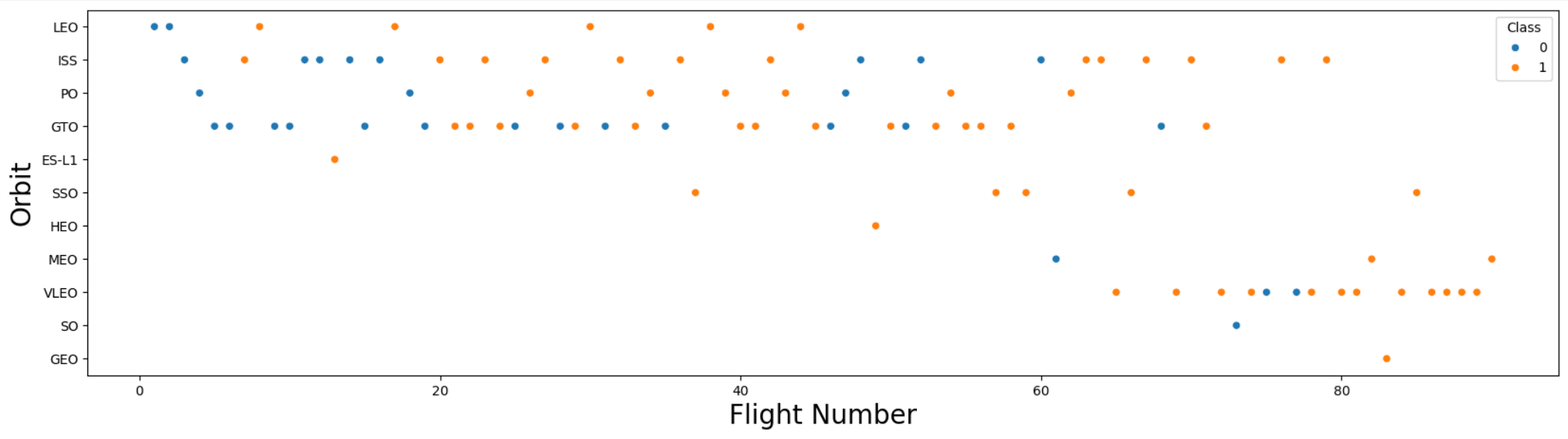
Success Rate vs. Orbit Type

- GTO is the most common orbit, and is at a 50% success rate
- VLEO has an 84% success rate and is the second most common orbit
- ES-L1, GEO, HEO have a 100% success rate on one mission
- SSO is also at 100% success rate, but has had 5 missions



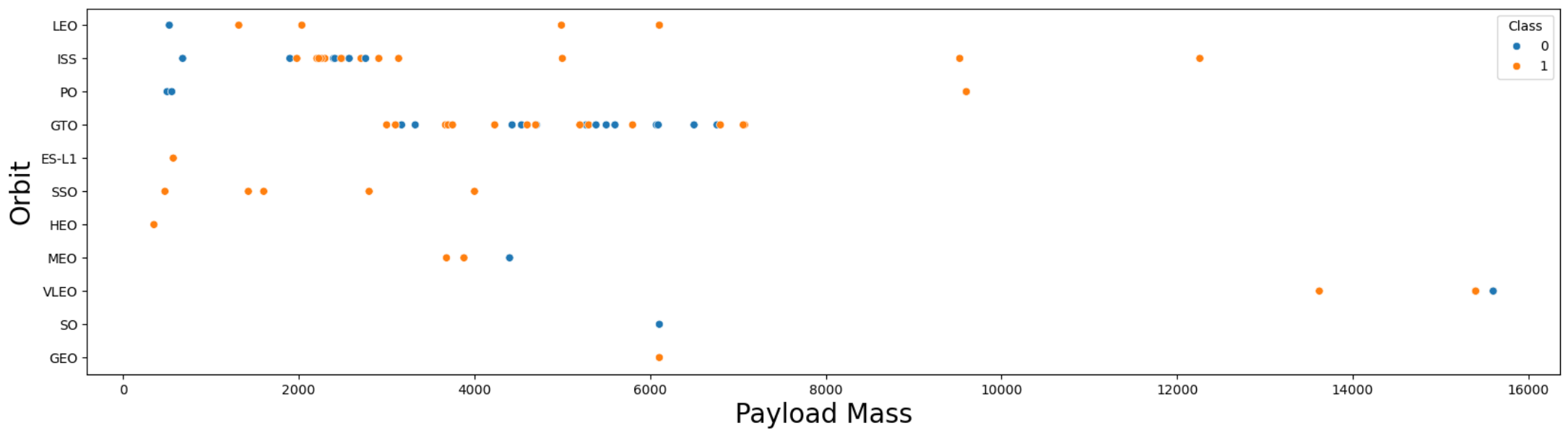
Flight Number vs. Orbit Type

- The VLEO orbit has been replaced the GTO orbit as the most frequent orbit since Flight Number 60.
- Success rate has significantly increased since Flight Number 55, indicating that it is more likely to be successful regardless of the orbit.



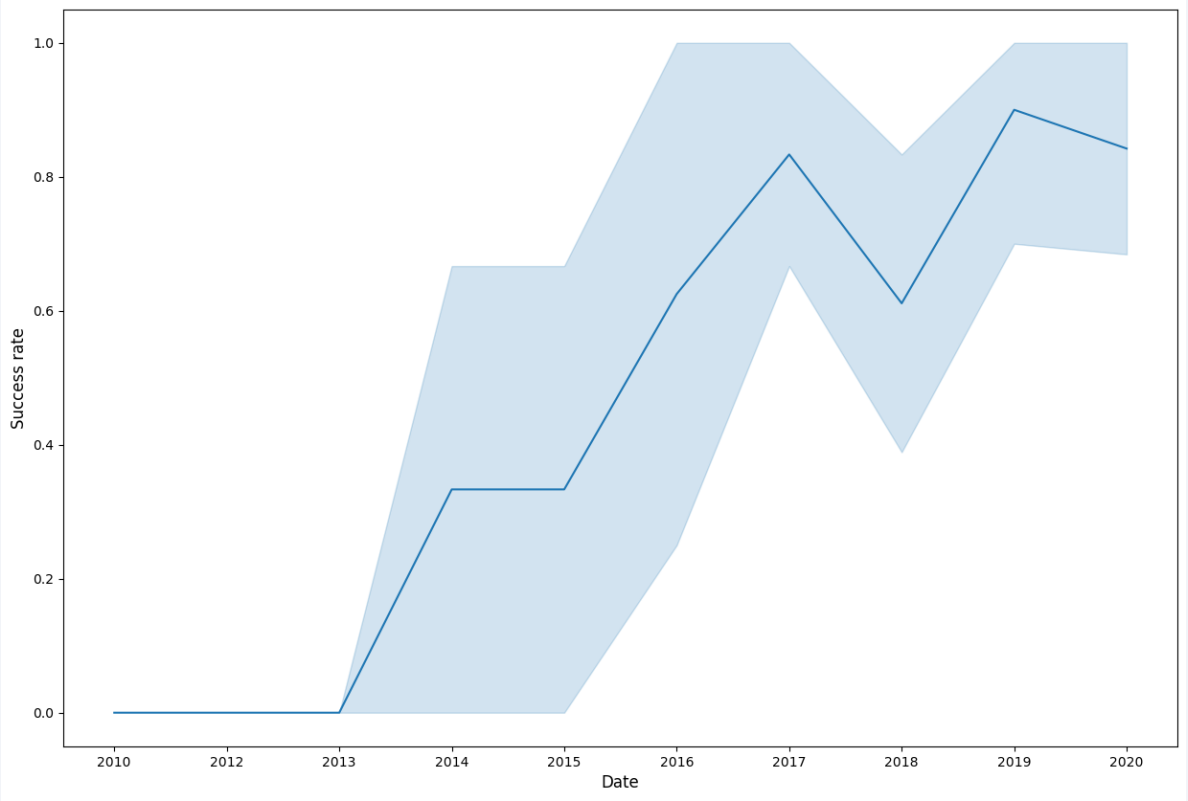
Payload vs. Orbit Type

- Payload mass seems to correlate with orbit
- LEO and SSO seem to have relatively low payload mass
- The largest payload masses have been sent to the VLEO orbit



Launch Success Yearly Trend

- There has been a markable increase in average success rate since 2013.
- The average success since 2019 is within the low 80%
- There appear to have been challenges in 2018, which would require further analysis to determine underlying issues



All Launch Site Names

- There are 4 unique launch sites listed in the SpaceX dataset, however there are only 3 actual launch sites
    - VAFB SLC-4E - Vandenberg Space Force Base Space Launch Complex 4E
    - KSC LC-39A - Kennedy Space Center Launch Complex 39A
    - CCAFS SLC-40 - Cape Canaveral Space Launch Complex 40
    - The CCAFS LC-40 was decommissioned and renamed as CCAFS SLC-40. For the purpose of the analysis, we will retain CCAFS LC-40 as a unique site
- The unique set is return by the following sql on the SpaceX table

%sql select distinct(Launch_Site) from SPACEXTABLE;

Launch Site Names Begin with 'CCA'

- Here are 5 sample records where launch sites begin with `CCA`

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

Total Payload Mass

- The payload carried by boosters from NASA is 45,596 KG. This payload is coded as CRS, indicating it as destined to the International Space Station.

Customer	Total
NASA (CRS)	45,596

- The query below will return the total Payload being carried by the boosters from NASA.

```
%sql select Customer,  sum(PAYLOAD_MASS__KG_)
as Total from SPACEXTABLE where Customer =
'NASA (CRS)';
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is located in following table

Customer	Booster_Version	AVG_PAYLOAD
NASA (CRS)	F9 v1.1	2296.0

- The query below will return the Average Payload being carried by the F9 v1.1 Booster from NASA.

```
%sql select Customer, Booster_Version, avg(PAYLOAD_MASS__KG_) as AVG_PAYLOAD
from SPACEXTABLE where Customer = 'NASA (CRS)' and Booster_Version = 'F9
v1.1';
```

First Successful Ground Landing Date

- The first successful landing outcome on a ground pad was achieved July 22, 2018

- The query to retrieve this date is below

```
%sql select min(Date) as First_Succesful_Landing_Date from SPACEXTABLE where
Landing_Outcome  = 'Success';
```

First_Succesful_Landing_Date
2018-07-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of the boosters which have successfully landed on drone ship with payload mass greater than 4000 but less than 6000 are located in the table below.

- The query to retrieve this data is below

```
%sql select Booster_Version, Landing_Outcome
from SPACEXTABLE where Landing_Outcome =
"Success (drone ship)" and (PAYLOAD_MASS__KG_ >
4000 and PAYLOAD_MASS__KG_< 6000);
```

Booster_Version	Landing_Outcome
F9 FT B1022	Success (drone ship)
F9 FT B1026	Success (drone ship)
F9 FT B1021.2	Success (drone ship)
F9 FT B1031.2	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- The Mission Outcomes to have been successfully met are reported to be 99%. The table identifies that the objectives of the mission have been successfully me 100 out of 101 missions. It is important to understand that some mission outcomes may be expected failures.
- The query to retrieve this data is below

Mission_Outcomes	Total
Failure	1
Successful	100

```
%sql Select Mission_Outcomes, count(*) as Total from \
(SELECT CASE WHEN Mission_Outcome LIKE 'Success%' THEN 'Successful' WHEN
Mission_Outcome LIKE 'Failure%' THEN 'Failure' END AS Mission_Outcomes from
SPACEXTABLE)\
Group by Mission_Outcomes;
```

Boosters Carried Maximum Payload

- The following list contains the names of the boosters which have carried the maximum payload mass
- The max payload is 15,600 KG

- The query to retrieve this data is below

```
%sql select Booster_Version from SPACEXTBL where
PAYLOAD_MASS__KG_ = (select
max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

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

- There were two failed landing_outcomes in drone ship in 2015. The table provides the month, landing outcome, booster version and launch site.

Month	Landing_Outcome	Booster_Version	Launch_Site
Jan	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Apr	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The query to retrieve this data is below

```
%sql select CASE substr(Date, 6, 2) WHEN '01' THEN 'Jan' WHEN '02' THEN 'Feb'
WHEN '03' THEN 'Mar' WHEN '04' THEN 'Apr' WHEN '05' THEN 'May'\
    WHEN '06' THEN 'June' WHEN '07' THEN 'July' WHEN '08' THEN 'August'    WHEN
'09' THEN 'September' WHEN '10' THEN 'October'    WHEN '11' THEN 'November'\
    WHEN '12' THEN 'December' ELSE 'Unknown' END AS Month, Landing_Outcome,
Booster_Version, Launch_Site from SPACEXTBL where substr(Date,0,5)='2015' \
and Landing_Outcome = 'Failure (drone ship)';
```

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

- There were 8 Landing Outcomes recorded between the date 2010-06-04 and 2017-03-20. There were 5 Failures (drone ship) and 3 Success (ground pad) outcomes.

Landing_Outcome	Total
Failure (drone ship)	5
Success (ground pad)	3

- The query to retrieve this data is below

```
%sql SELECT Landing_Outcome , count(*) as Total FROM SPACEXTBL WHERE  (Date
>'2010-06-04' AND Date < '2017-03-20') \
AND Landing_Outcome in ('Failure (drone ship)','Success (ground pad)') Group by
Landing_Outcome;
```

Section 3

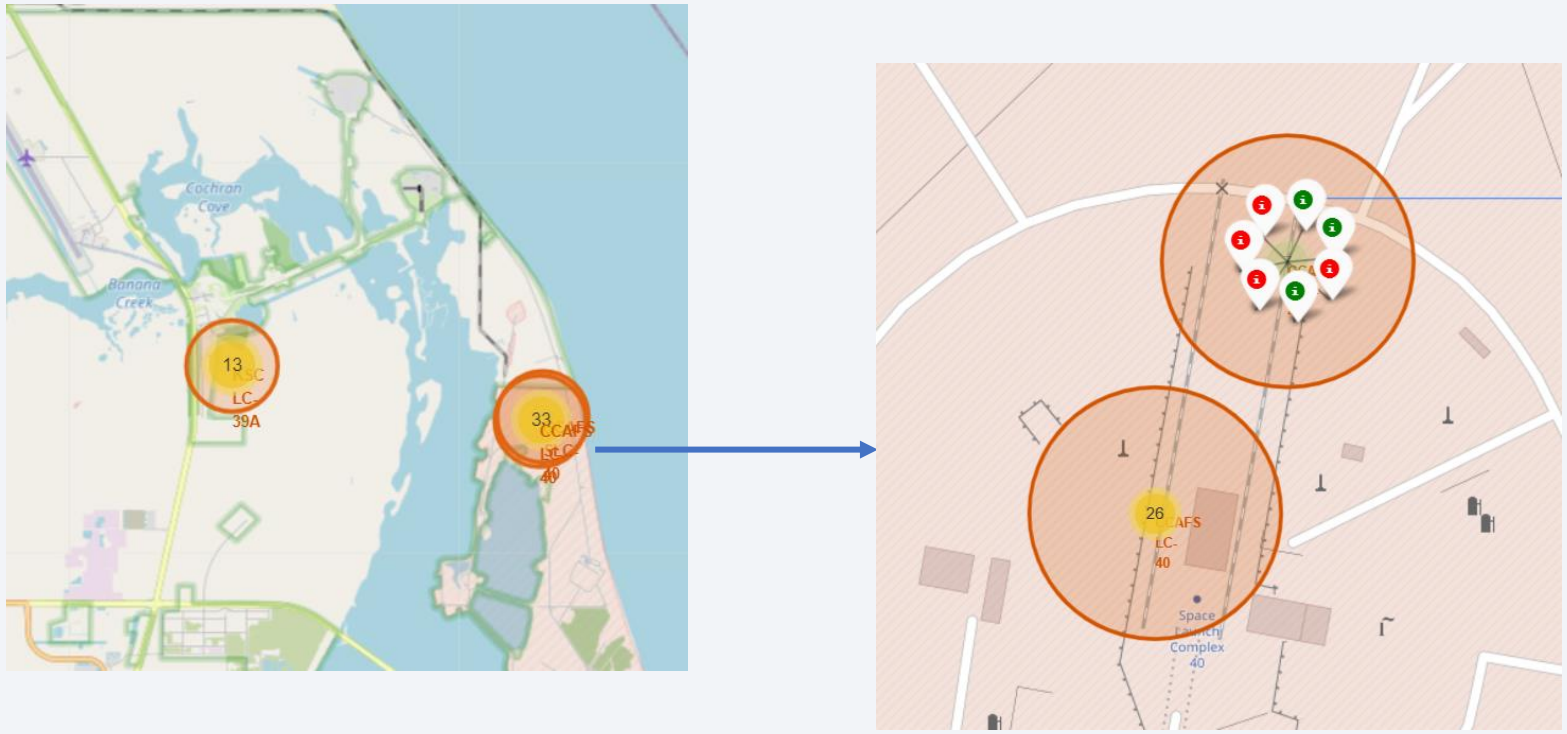
Launch Sites Proximities Analysis

Map of SpaceX Launch Sites

- There are launch sites on both the Eastern and Western United States.
- All launch sites located in proximity to the ocean

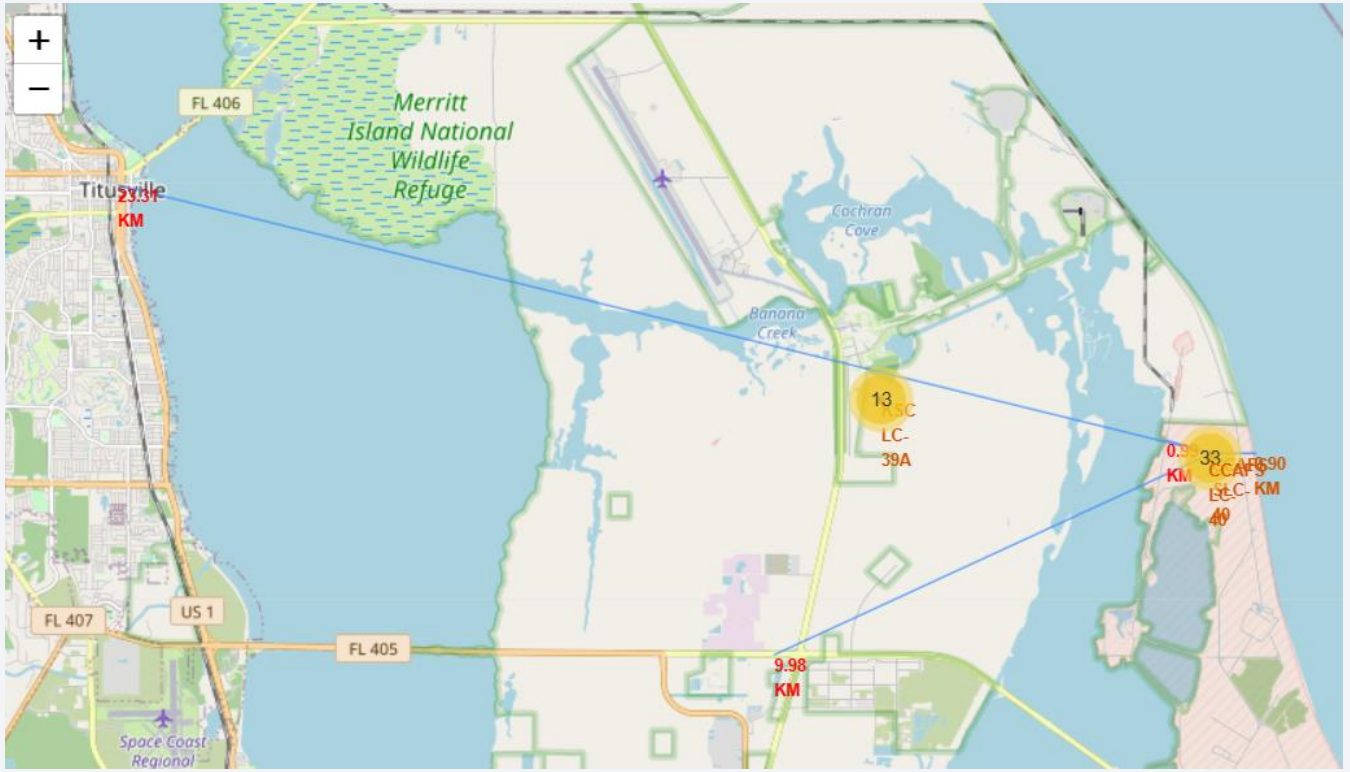
Launch Summary with Outcomes for CCAFS LC-40

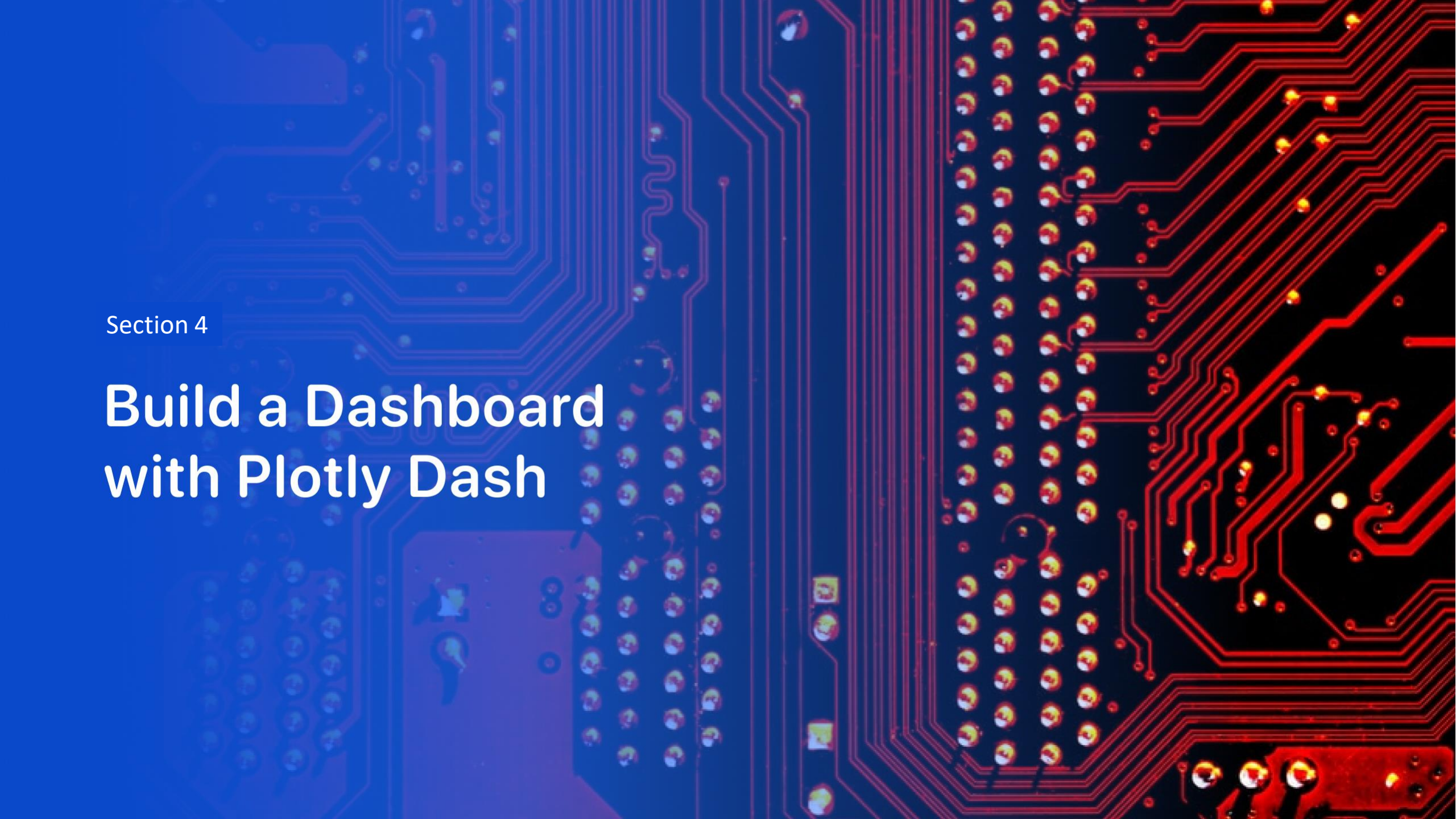
- Zooming into the CCAFS SLC-40 launch site, we see that there have been 3 Successful launches, and 4 Failed launches.
- The map allows for key information to be explored by the users, and launch outcomes to be shared easily



Launch Site Proximity to local objects

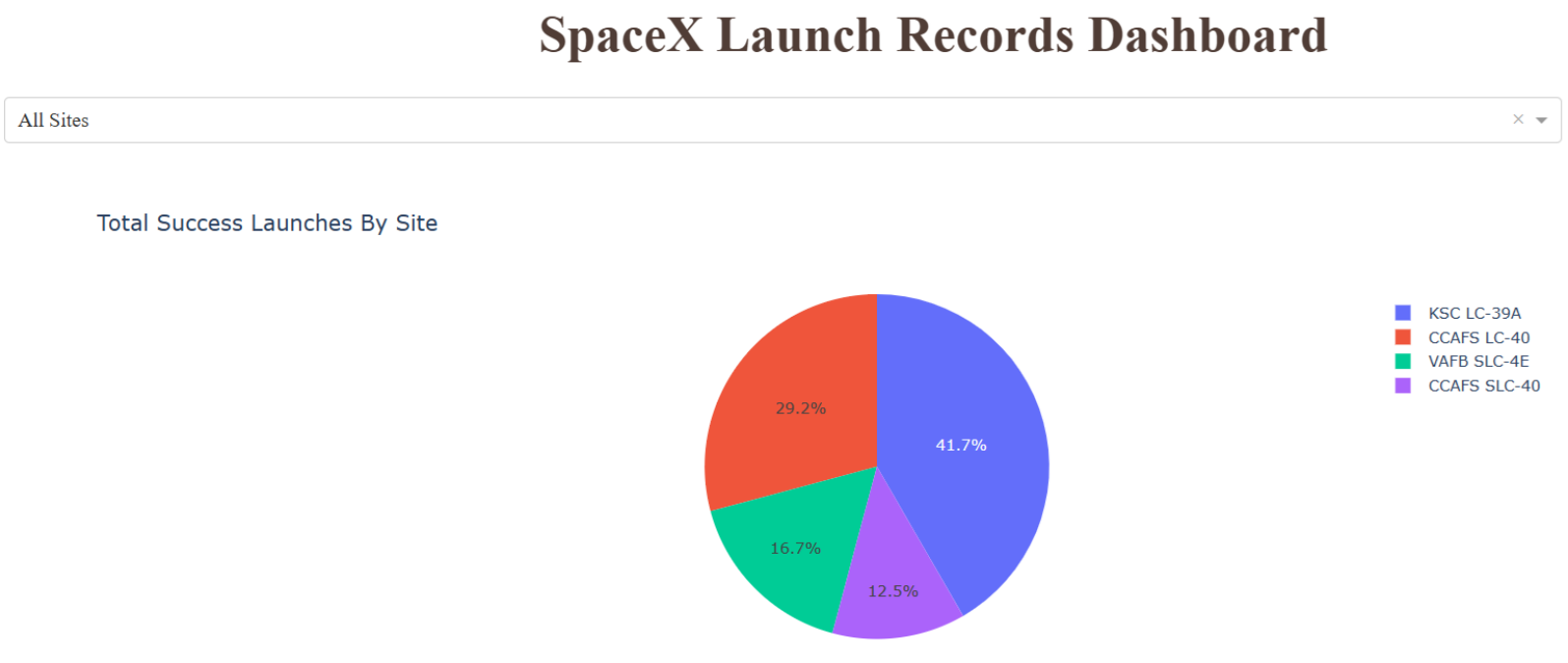
- Polylines are added to the visualization to provide distances to points of interest (railways, nearby cities, highways, etc)
- The following points of interest for CCAFS SLC-40 are listed below
- Highway FL 405 is 9.9 km away
- Rail line is 0.9 km away
- Titusville is 23.1 km




Section 4
Build a Dashboard with Plotly Dash

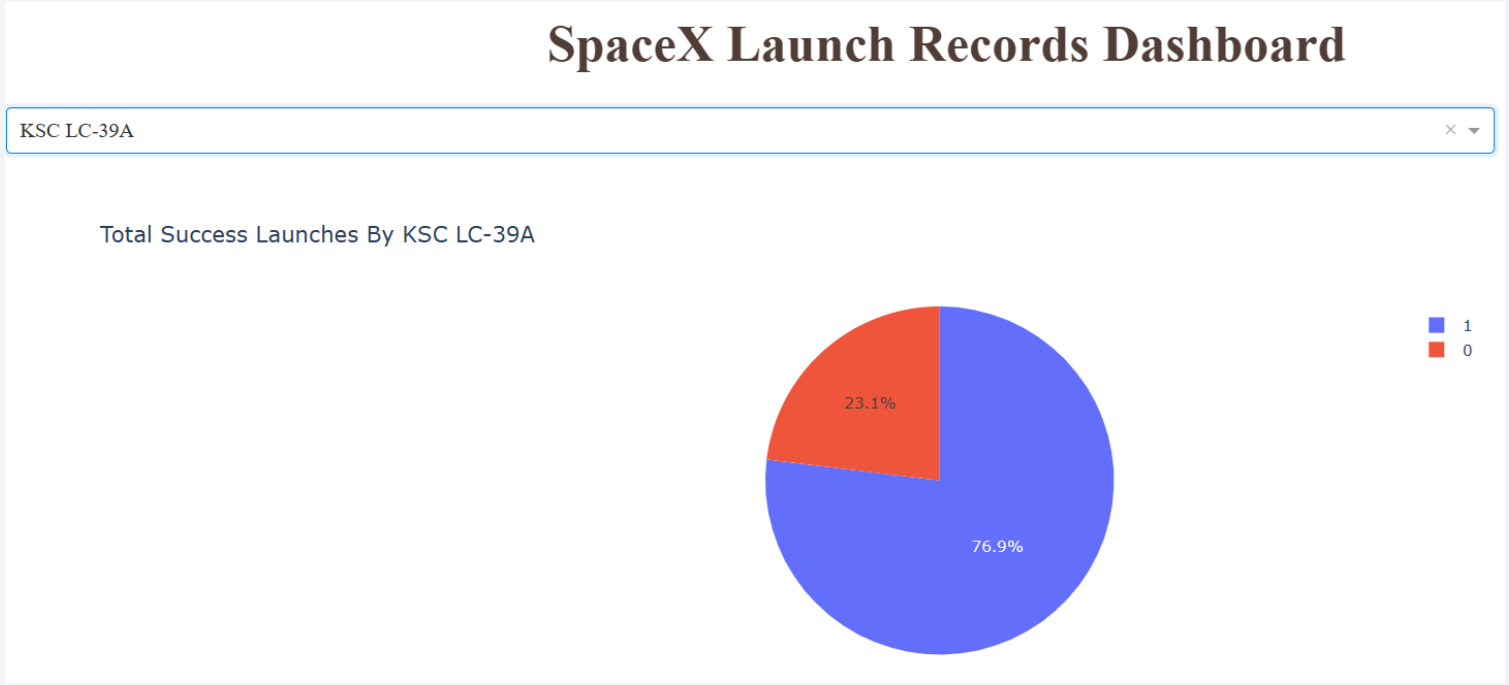
SpaceX Launch Dashboard - Success Rate

- The success rate for all of the SpaceX launch sites is presented below.
- It is evident that KSC LC-39A compared to all launch data has had the highest success rate overall at 41.7%
- The CCAFS LC-40 has seen a 29.2% success rate, however the same location – now renamed as CCAFS SLC-40 is only at 12.5% success rate



KSC LC-39A Launch Site – Launch Performance

- The success rate of the KSL LC-39A launch site is 76.9%



SpaceX Launch Dashboard – Payload vs Success

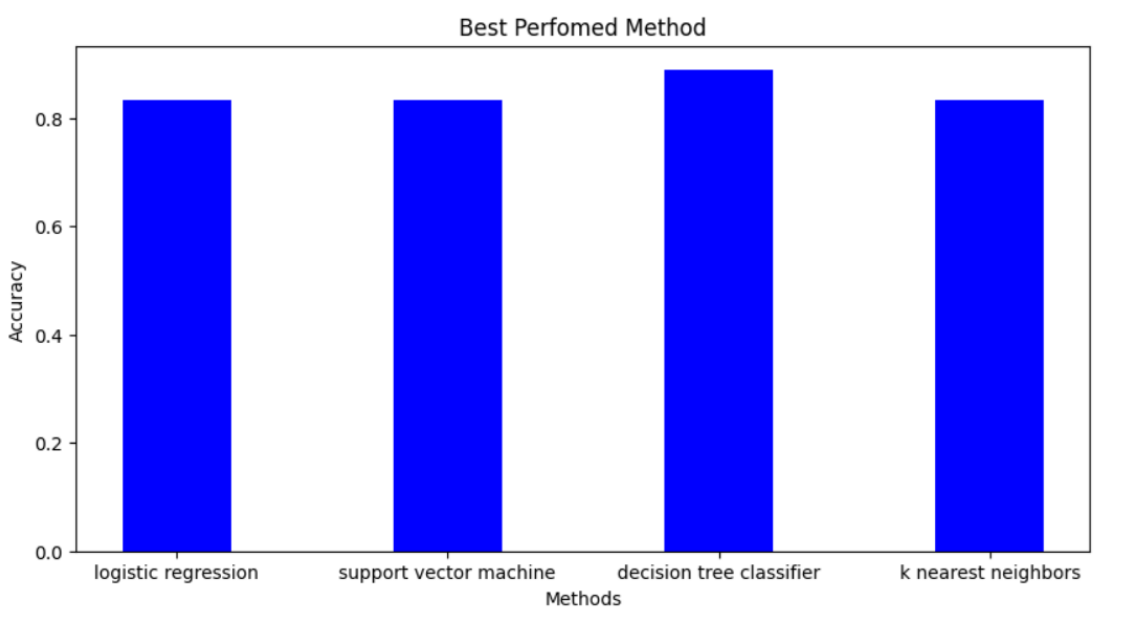
- The scatter plot demonstrates the correlation between Payload and outcome of launches
- The Class axis is represented by 1 = Successful, 0 = Failure
- The Booster versions help to identify which Booster versions are most successful. In this grid, booster version FT has seen the most success

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The accuracy of the Logistic regression, SVM, and KNN models are all equal at 83%
- The Decision Tree classifier model returned the highest accuracy at 88%

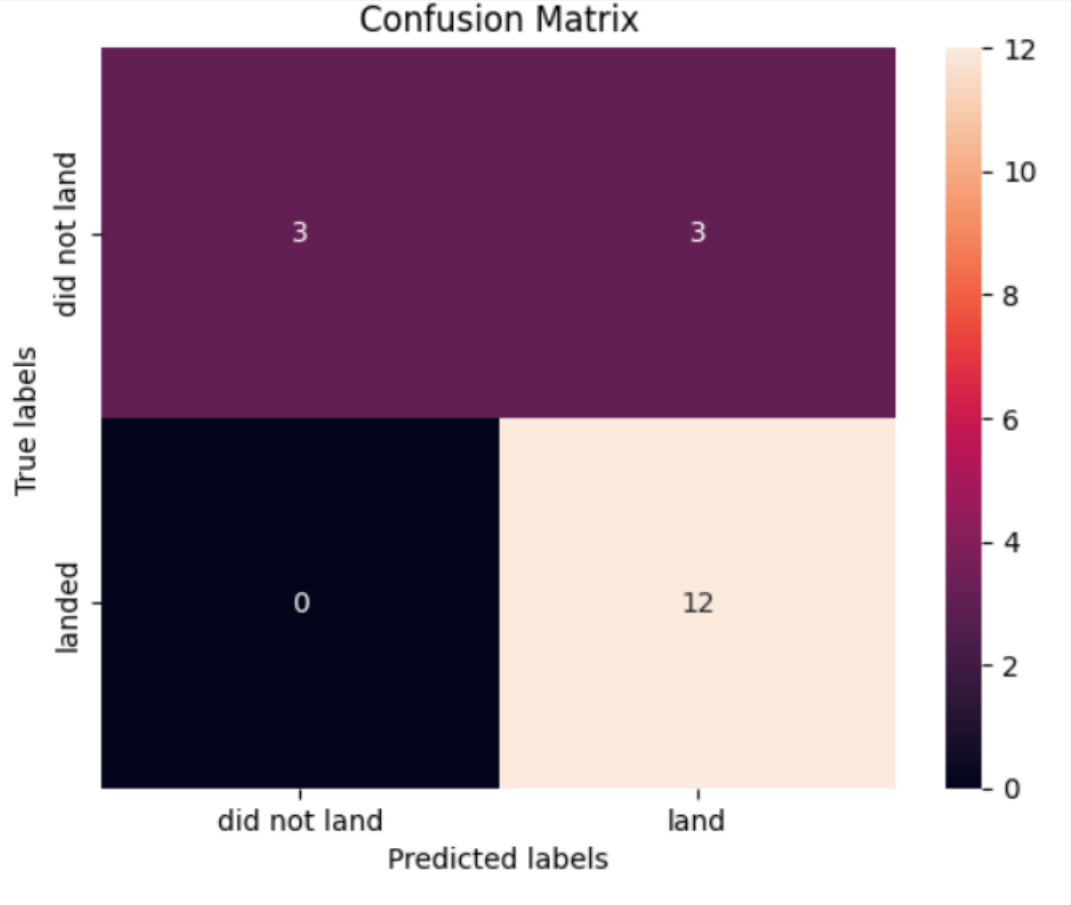


Confusion Matrix

- The confusion matrix was the same across all the three models; Decision Tree, Logistic regression, SVM, and KNN model.

- True Negative (TN) = 3
- True Positive (TP) = 12
- False Negative (FN) = 0
- False Positive (FP) = 3

- Precision = TP/(TP+FP) :. 12/15 = 80%
- Recall = TP/TP+FN :. 12/12 = 100%



Conclusions

- The success rate of SpaceX launches have improved since 2013. The rate of success has been fairly consistent in most recent 20 launches, demonstrating maturing stability
- SpaceX launches have been focusing on VLEO orbit, as a large portion of the last 20 launches have been destined to this orbit
- KSC LC-39 is the most frequently used launch site for SpaceX, and it has the highest success rate comparatively
- The Decision Tree Classifier model is the lead algorithm for predicting successful landing outcomes. As more launches occur, the modelling algorithms should be rerun to challenge accuracy of predictions.
- With the ability to predict the landing success of the SpaceX launch, Space Y can now use the predictions to build our competing bids based on SpaceX's launch success.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!