IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Mohab Bahnassy>
<8/21/2024>

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA
  - Prediction
- Summary of all results
  - EDA Results
  - Predicition Results

Introduction

Space X advertises 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. This information can be used if an alternate company wants to bid against space X for a rocket launch.

The goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and from Wikipedia
- Perform data wrangling
  - One hot encoding and dropping rows
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Compared different classification models (ex. LR, SVM, Classification Trees)

Data Collection

- The datasets were collected using SpaceX API and Wikipedia
- Wikipedia
  - HTML Response ---> BeautifulSoup Object ---> Web Scraping Methods ----> Pandas Dataframe
- API
  - API Call ----> JSON File ----> Pandas Data Frame

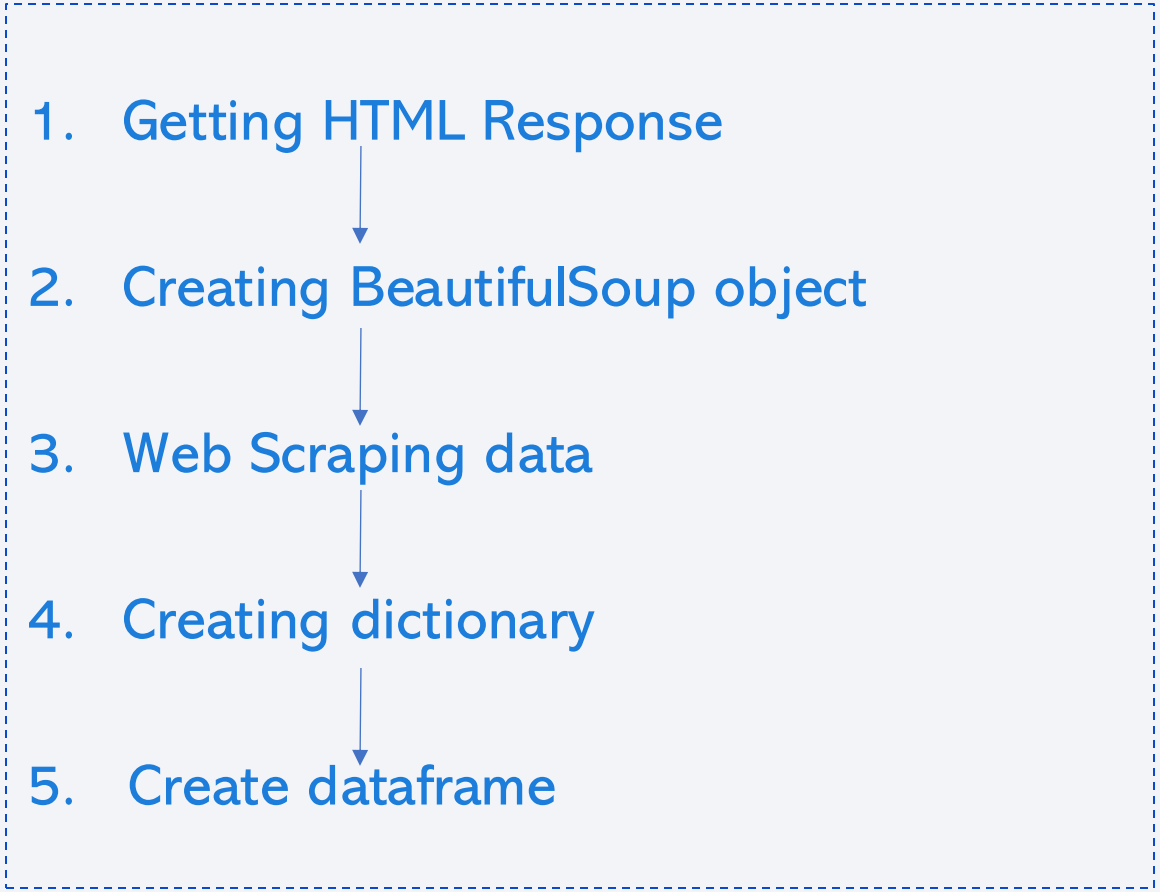
Data Collection – SpaceX API

- We used the SpaceX API and get request function to retrieve data, we then cleaned the requested data.

- https://github.com/mohabbahnassy/final_dsc/blob/main/Data%20Collection%20API.ipynb

Data Collection - Scraping

- Web scraping methods were used to scrape date from the Wikipedia page for the Falcon 9 page
- https://github.com/mohabbahnassy/final_dsc/blob/main/DataCollection_Webscraping.ipynb



Data Wrangling

- We created a result column based on the results in the original data set.
  - True Ocean, True RTLS, True ASDS means the mission has been successful.
  - False Ocean, False RTLS, False ASDS means the mission was a failure.

  We also calculated the number of launches at each site, and the number and occurrence of each orbits

  https://github.com/mohabbahnassy/final_dsc

EDA with Data Visualization

Scatter Graphs

- Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload vs. Launch Site, Orbit vs. Flight Number, Payload vs. Orbit Type, Orbit vs. Payload Mass

Bar Graph

- Success rate vs. Orbit

Line Graph

- Success rate vs. Year

https://github.com/mohabbahnassy/final_dsc/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- SQL Queries:
  - • Displaying the names of the unique lauunch 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 achieved. • 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 records which will display the month names, faiilure landing_ouutcomes in drone ship, booster versions, launch_site for the months in year 2015. • Rank the count of successful landiing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order

Build an Interactive Map with Folium

- Red circle at NASA Johnson Space Center's coordinate with label showing its name (folium.Circle, folium.map.Marker). • Red circles at each launch site coordinates with label showing launch site name (folium.Circle, folium.map.Marker, folium.features.DivIcon). • The grouping of points in a cluster to display multiple and different information for the same coordinates (folium.plugins.MarkerCluster). • Markers to show successful and unsuccessful landings. Green for successful landing and Red for unsuccessful landing. (folium.map.Marker, folium.Icon). • Markers to show distance between launch site to key locations (railway, highway, coastway, city) and plot a line between them. (folium.map.Marker, folium.PolyLine, folium.features.DivIcon
- https://github.com/mohabbahnassy/final_dsc/blob/main/LaunchSite_Location.ipynb

Build a Dashboard with Plotly Dash

Plotted pie charts showing the total launches by a certain sites

Plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster versions.

- https://github.com/mohabbahnassy/final_dsc/blob/main/plotly.py

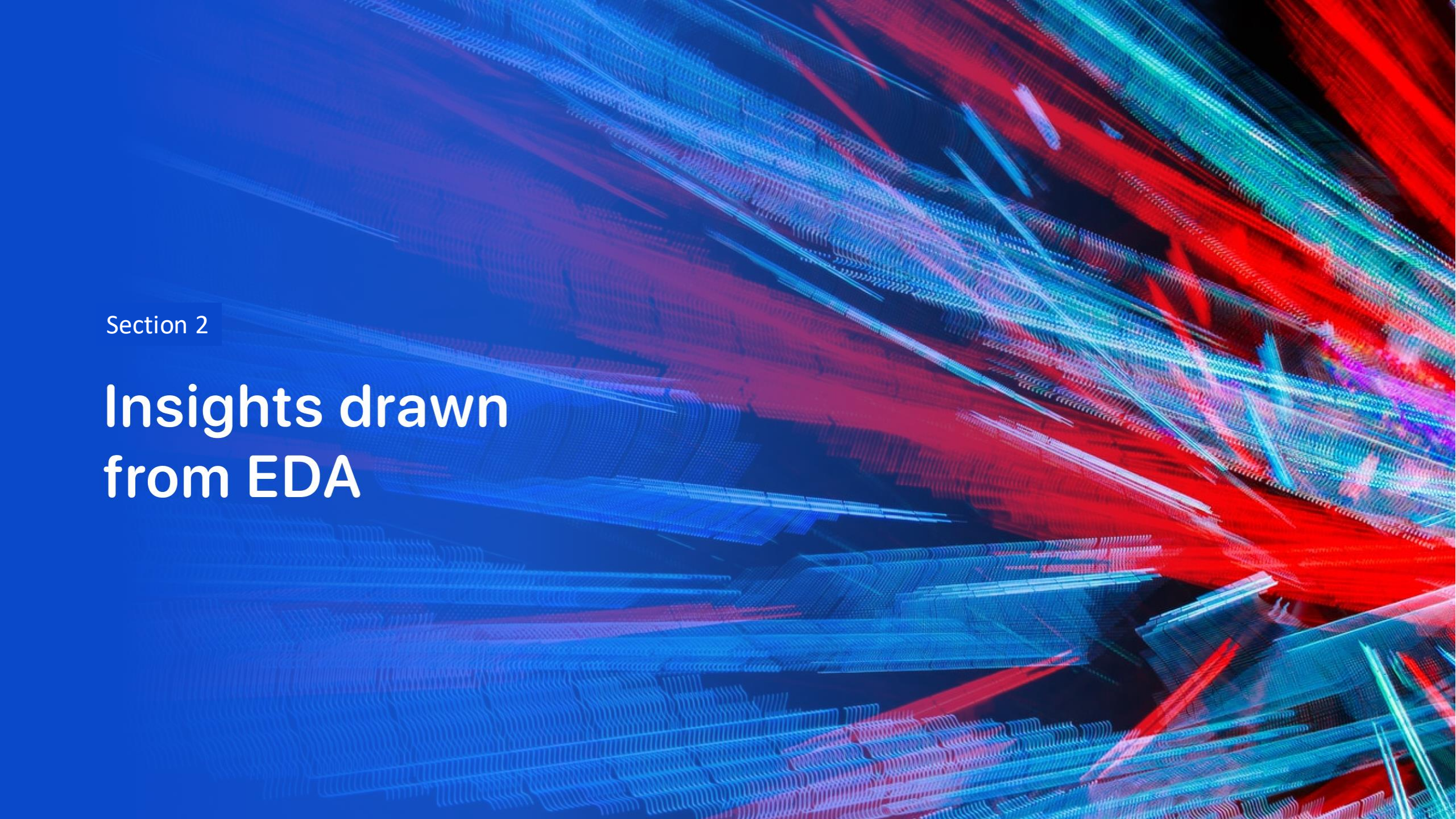
Predictive Analysis (Classification)

- Preparing the data (normalizing, splitting, etc.)
- Building different machine learning models and tune different hyperparameters using GridSearchCV.
- Evaluating and comparing between different classification models based on accuracy metrics.

- https://github.com/mohabbahnassy/final_dsc/blob/main/Machine%20Learning%20Prediction.ipynb

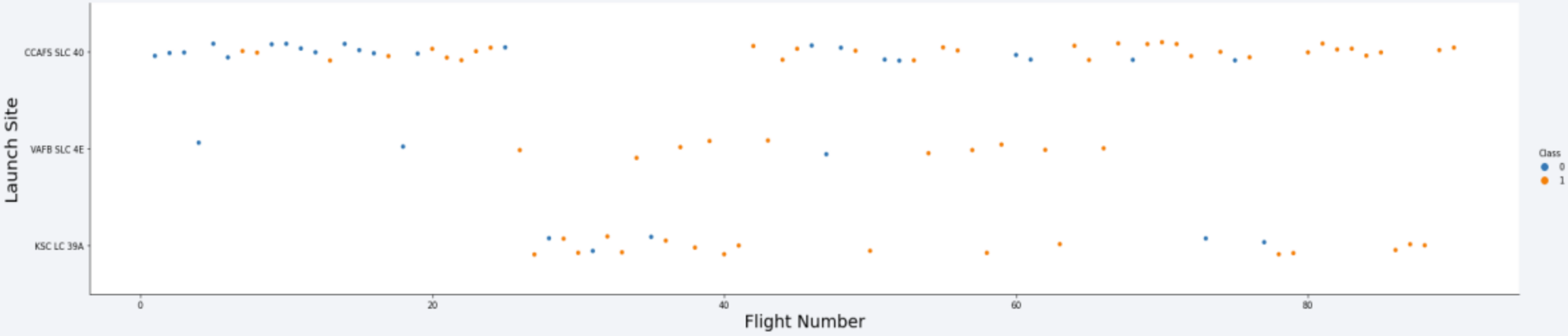
Results

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


Section 2
Insights drawn from EDA

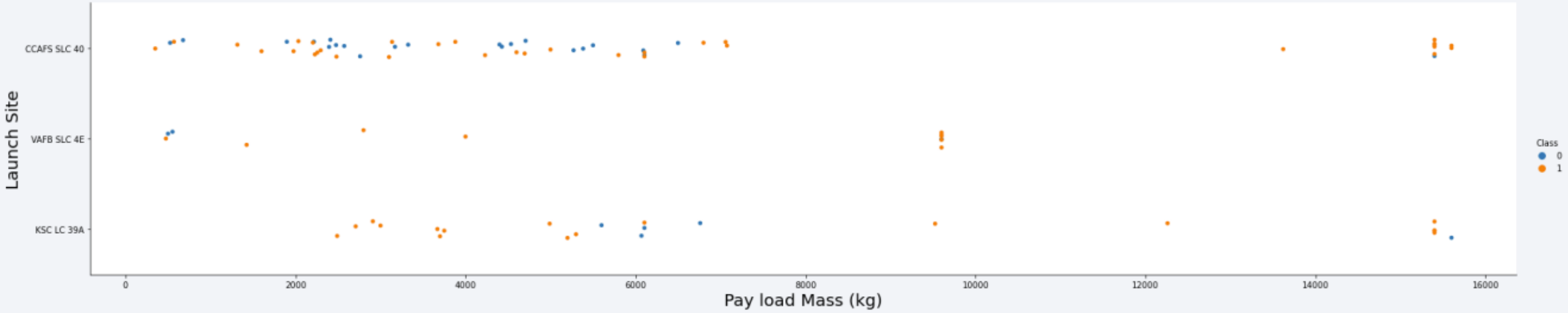
Flight Number vs. Launch Site

- Success Rate is increasing.

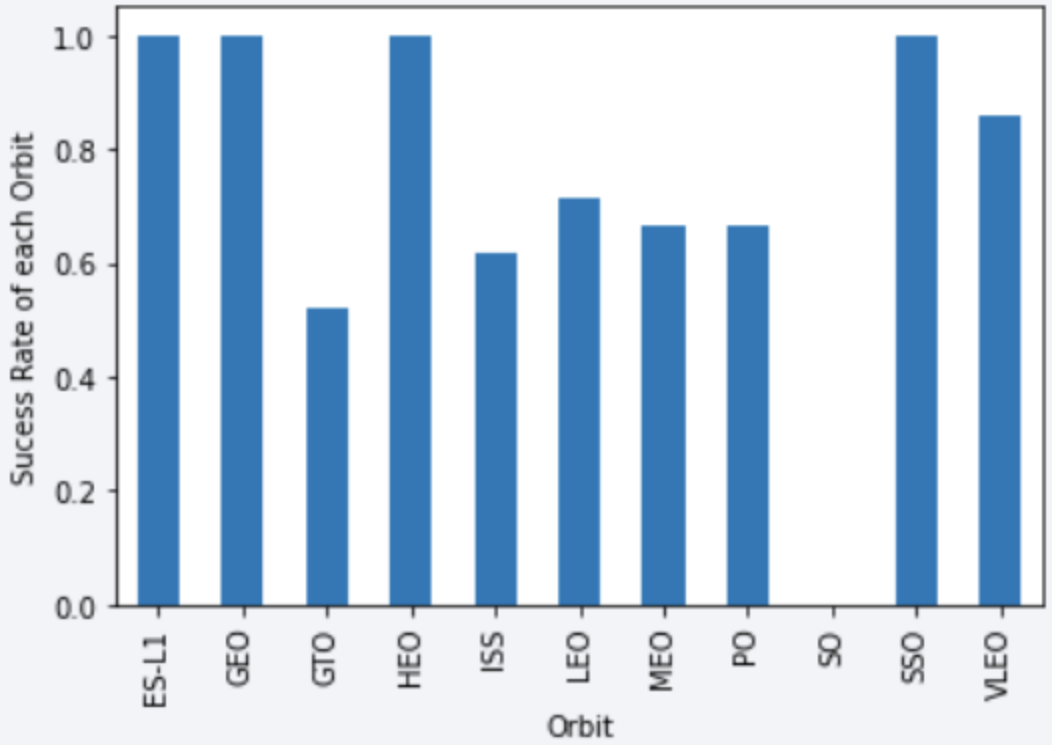


Payload vs. Launch Site

- Generally, higher payload means higher success rate.

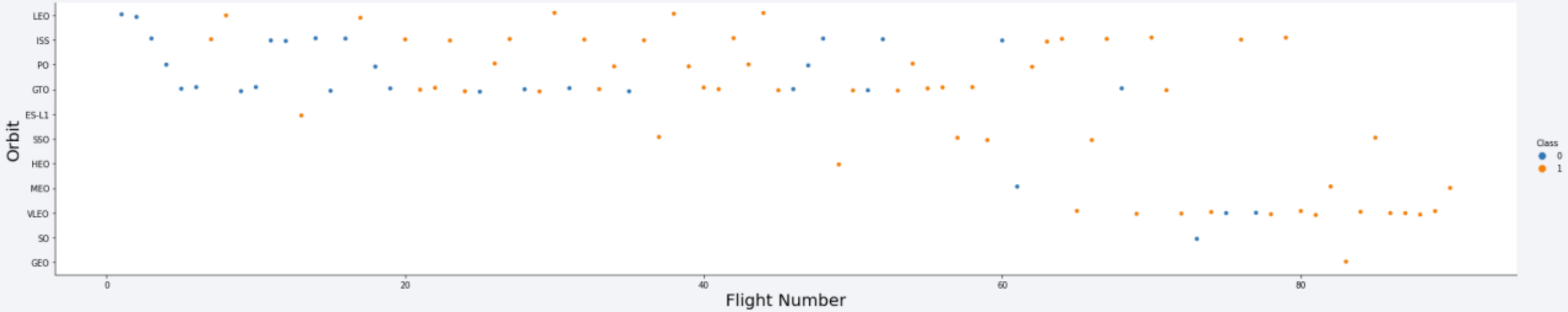


Success Rate vs. Orbit Type



Flight Number vs. Orbit Type

- In the LEO orbit, success increases with number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.

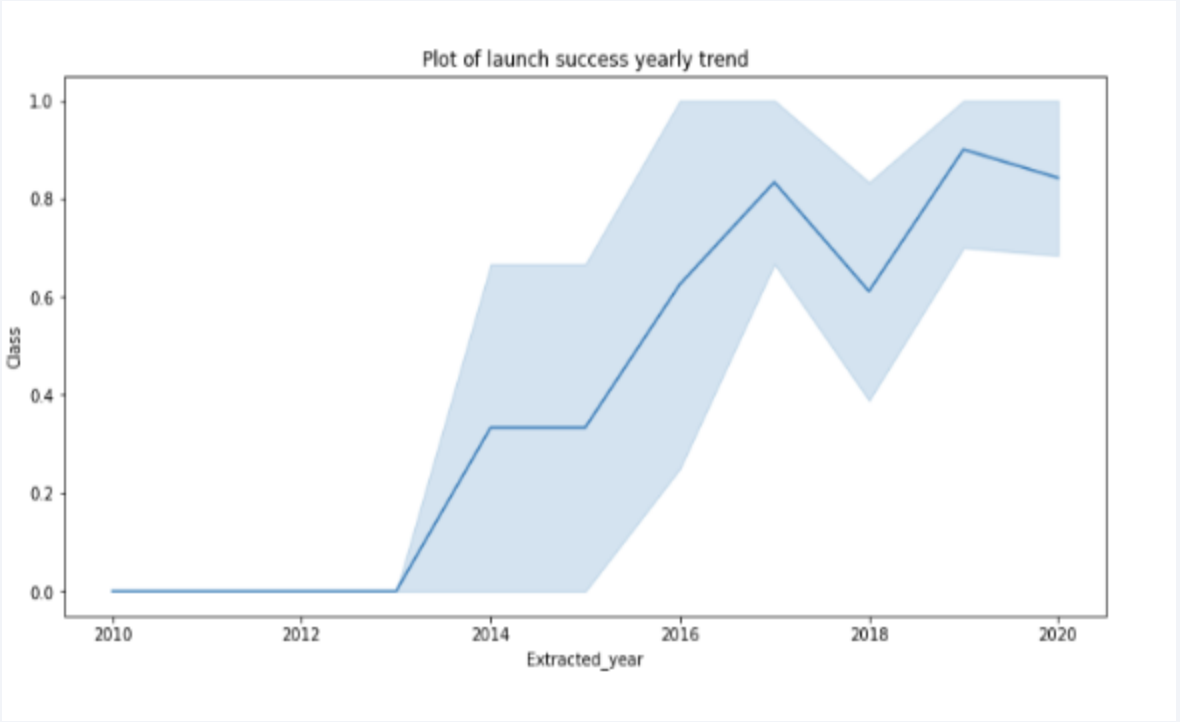


Payload vs. Orbit Type

- We can observe that with heavy payloads, the successful landing are more for select orbits.

Launch Success Yearly Trend

- Success rate increases from 2013 onwards.



All Launch Site Names

```
SELECT DISTINCT "LAUNCH_SITE" FROM SPACEXTBL
```

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

Launch Site Names Begin with 'CCA'

```
SELECT * FROM SPACEXTBL WHERE "LAUNCH_SITE" LIKE '%CCA%' LIMIT 5
```

We use the like clause and perctanage signs to extract launch sites that include CCA.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
08-12-2010	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
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

Total Payload Mass

```
SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "CUSTOMER" = 'NASA (CRS)'
```

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

- We use the like clause since we only need F9 v1.1

```
SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "BOOSTER_VERSION" LIKE '%F9 v1.1%'
```

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

- We use MIN to get the least date. (earliest)

```
SELECT MIN("DATE") FROM SPACEXTBL WHERE "Landing _Outcome" LIKE '%Success%'
```

MIN("DATE")
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql SELECT "BOOSTER_VERSION" FROM SPACEXTBL WHERE "LANDING _OUTCOME" = 'Success (drone ship)' \
AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;
```

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql SELECT (SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Success%') AS SUCCESS, \
(SELECT COUNT("MISSION_OUTCOME") FROM SPACEXTBL WHERE "MISSION_OUTCOME" LIKE '%Failure%') AS FAILURE
```

SUCCESS	FAILURE
100	1

Boosters Carried Maximum Payload

- Here we use the subquery to select only records where the payload mass is maximum.

```
%sql SELECT DISTINCT "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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql SELECT substr("DATE", 4, 2) AS MONTH, "BOOSTER_VERSION", "LAUNCH_SITE" FROM SPACEXTBL\
WHERE "LANDING _OUTCOME" = 'Failure (drone ship)' and substr("DATE",7,4) = '2015'
```

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

```
%sql SELECT "LANDING _OUTCOME", COUNT("LANDING _OUTCOME") FROM SPACEXTBL\
WHERE "DATE" >= '04-06-2010' and "DATE" <= '20-03-2017' and "LANDING _OUTCOME" LIKE '%Success%'\
GROUP BY "LANDING _OUTCOME" \
ORDER BY COUNT("LANDING _OUTCOME") DESC ;
```

Landing _Outcome	COUNT("LANDING _OUTCOME")
Success	20
Success (drone ship)	8
Success (ground pad)	6

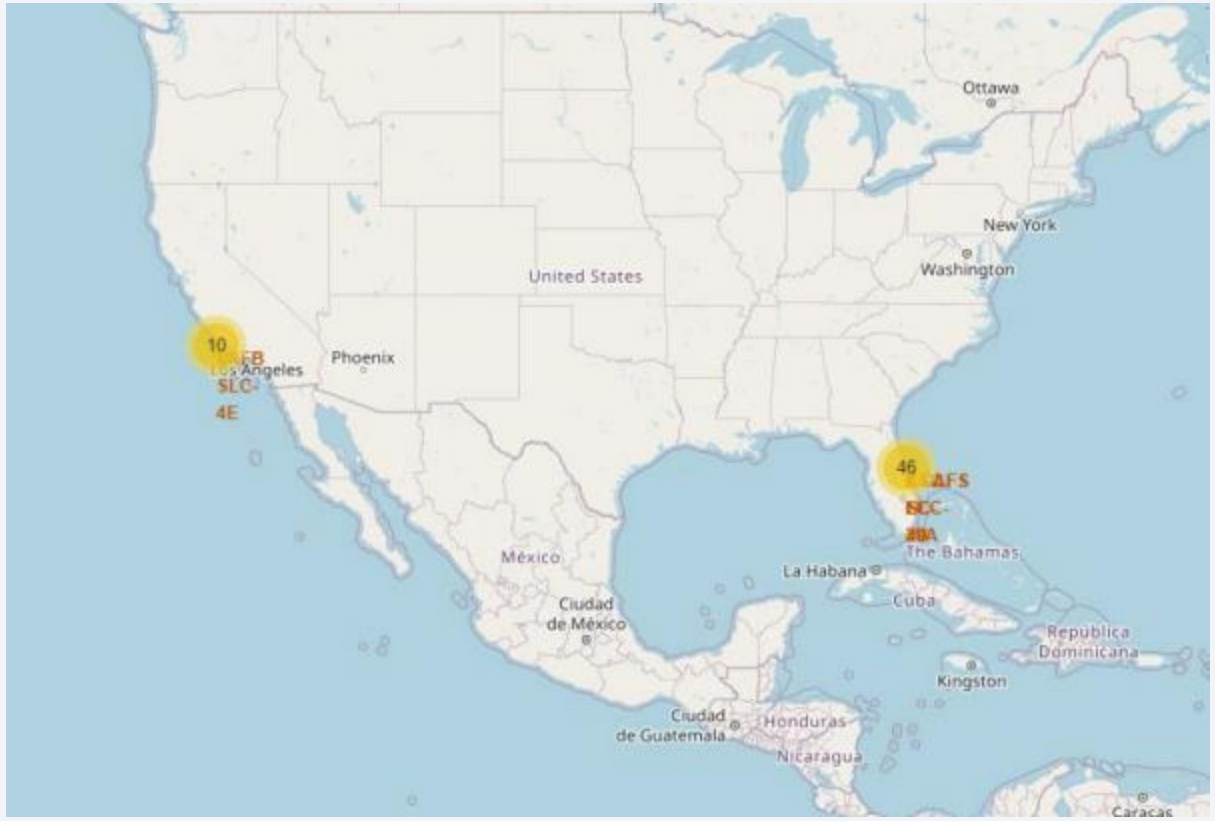
- Here we use the GROUP BY clause since we want to rank according to outcomes.

Section 3

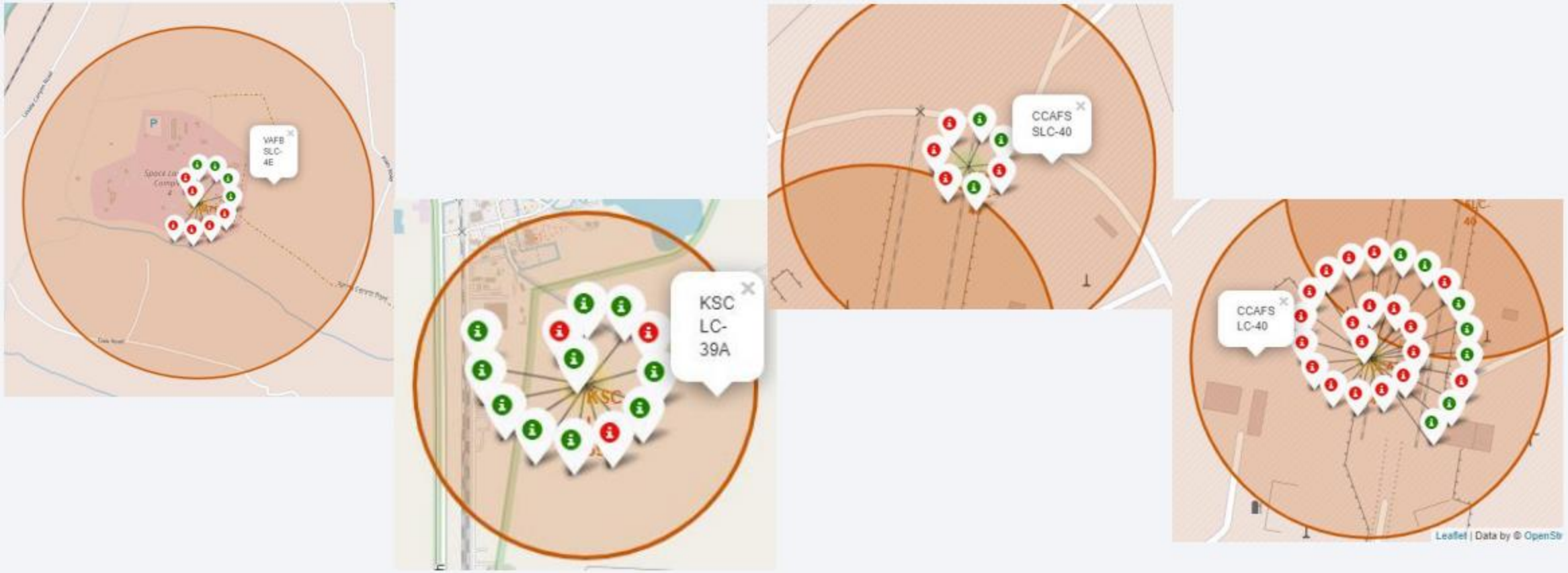
Launch Sites Proximities Analysis

Folium map – SpaceX Launch sites

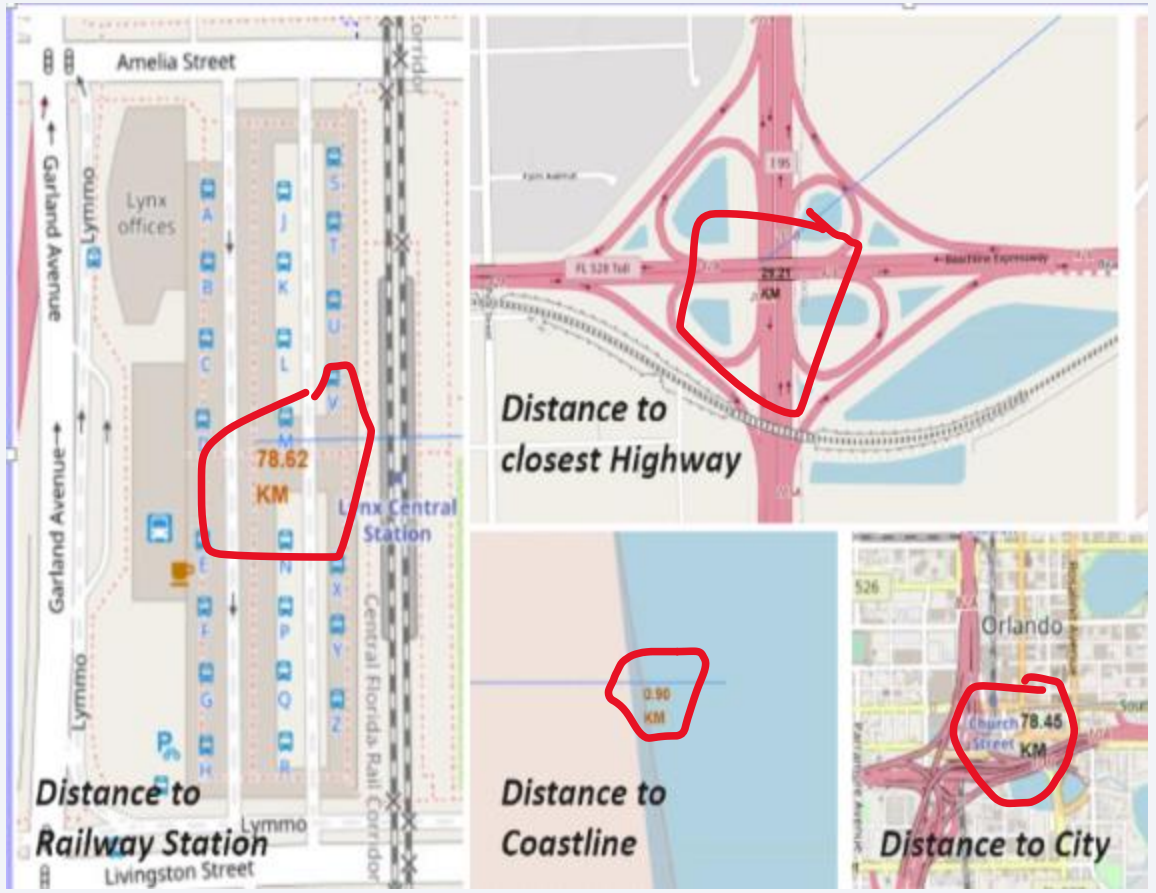
All of the sites are near US coastlines.

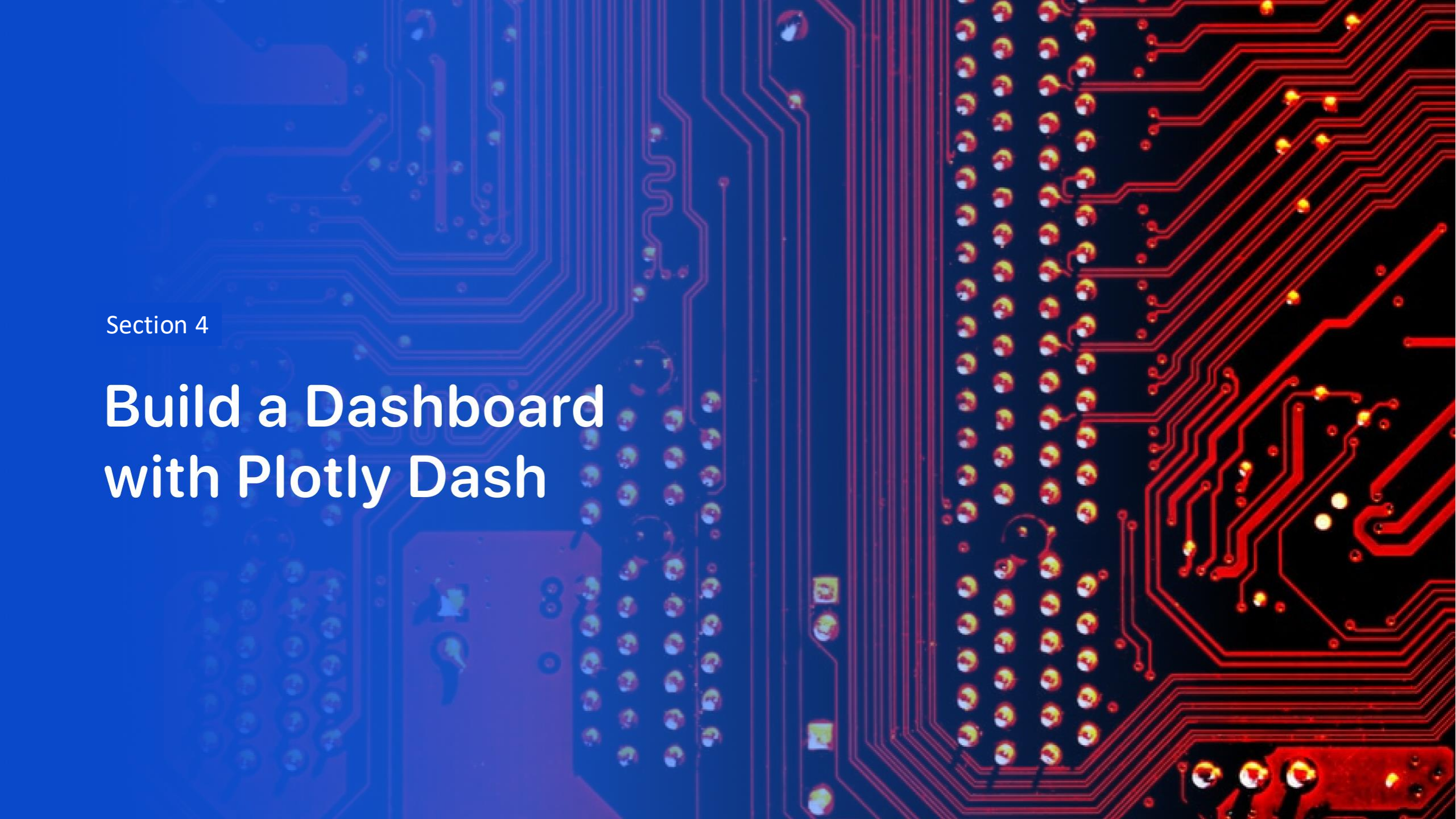


Labeled markers, representing success or failure

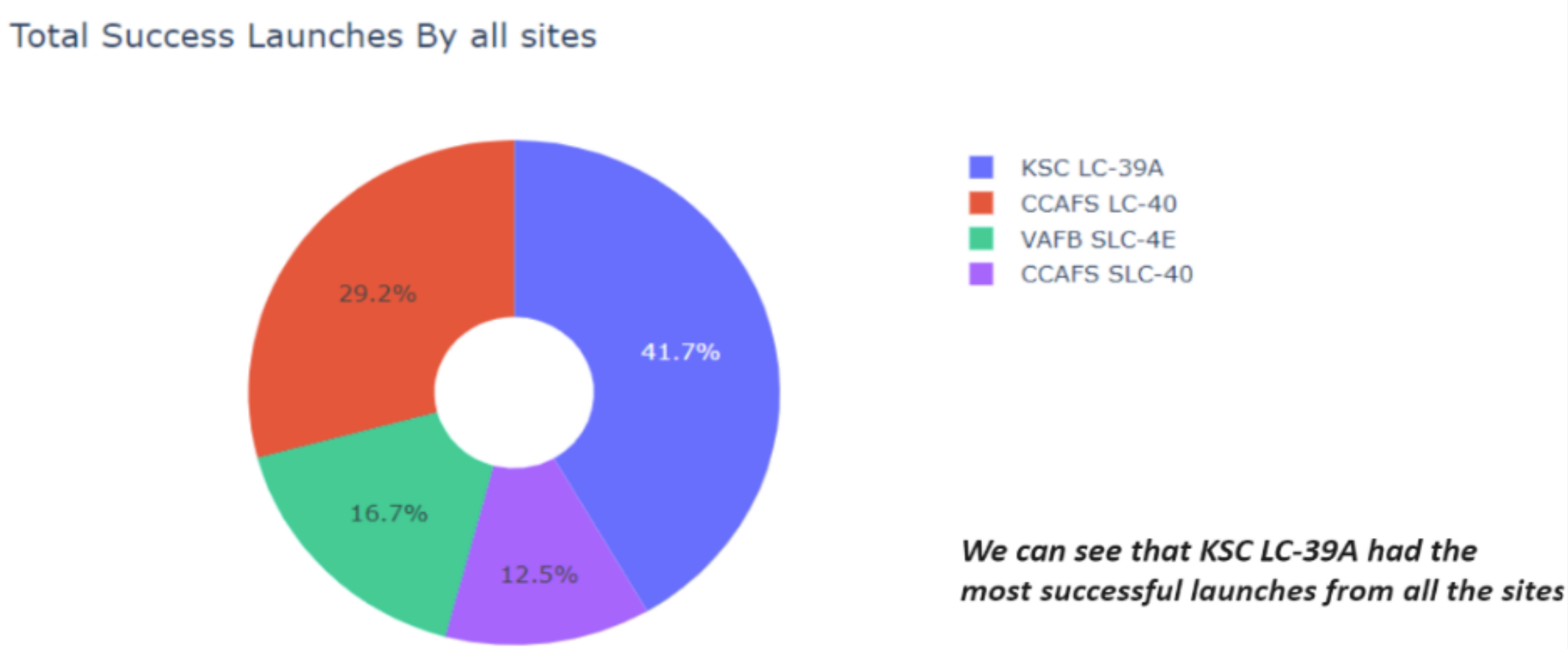


Launch site distance to landmarks

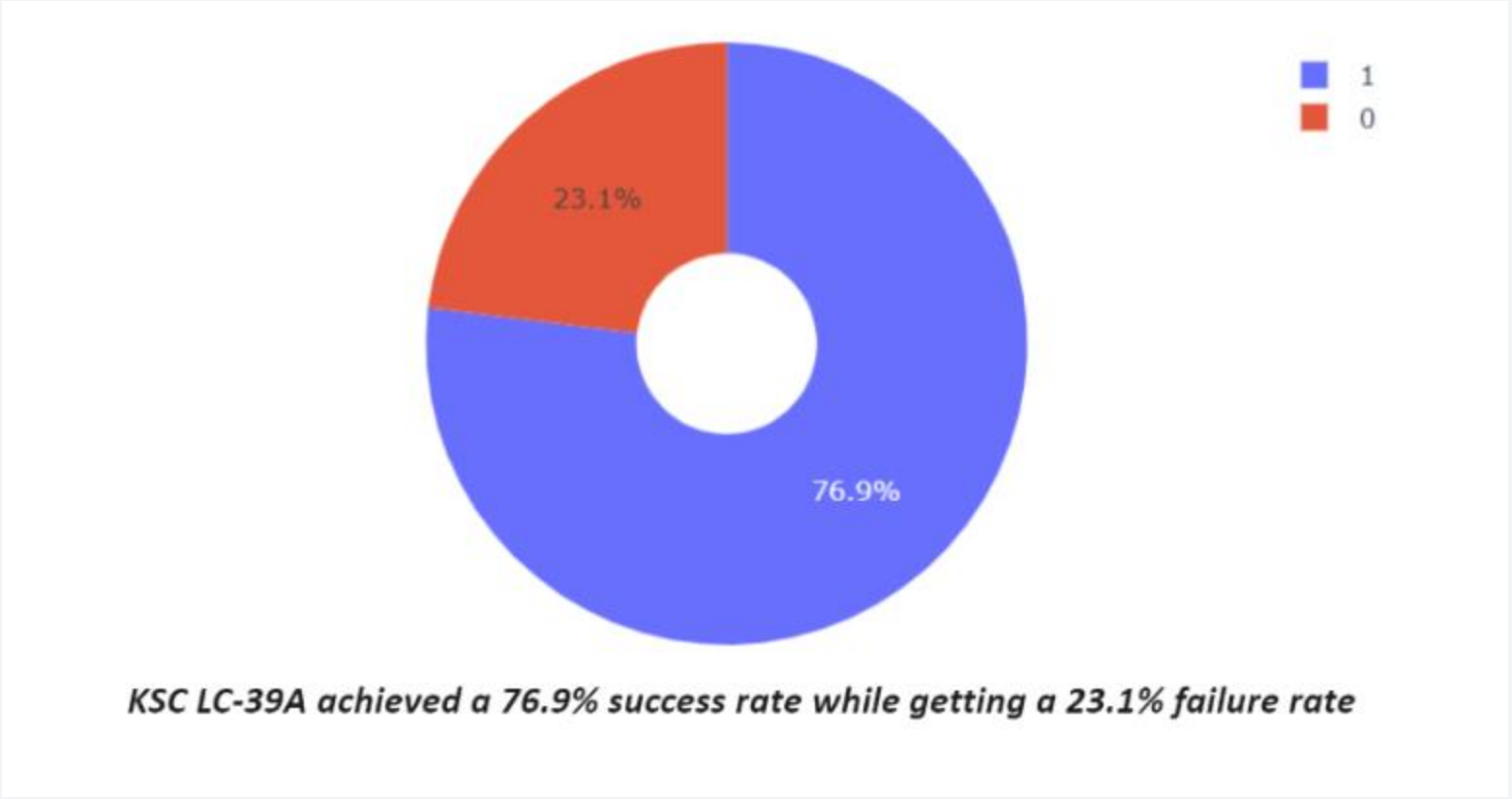



Section 4
Build a Dashboard with Plotly Dash

Pie chart of success percentage achieved for each launch site

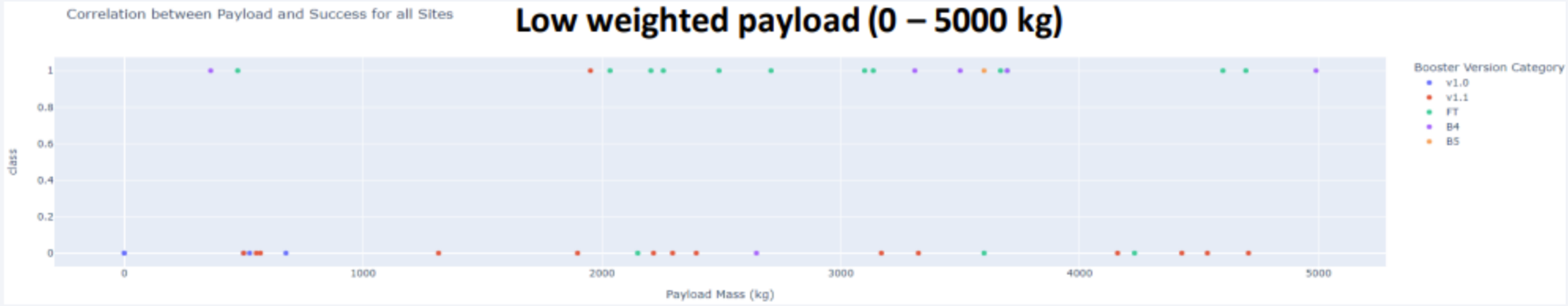


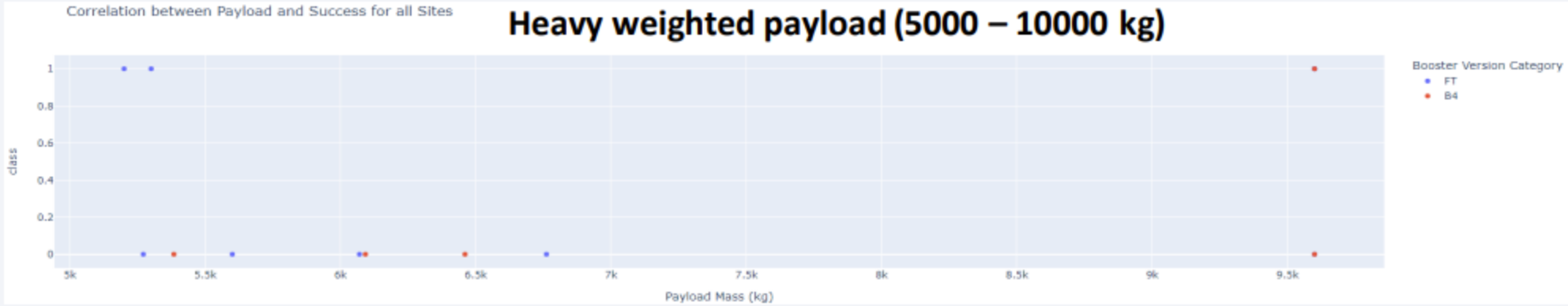
Pie chart showing the Launch site with the highest launch success ratio



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Payload mass vs Outcome for all sites

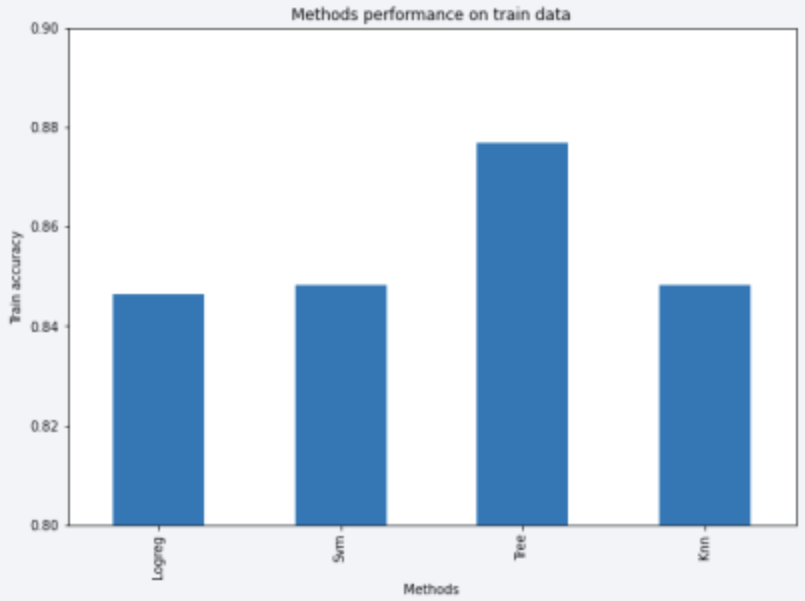


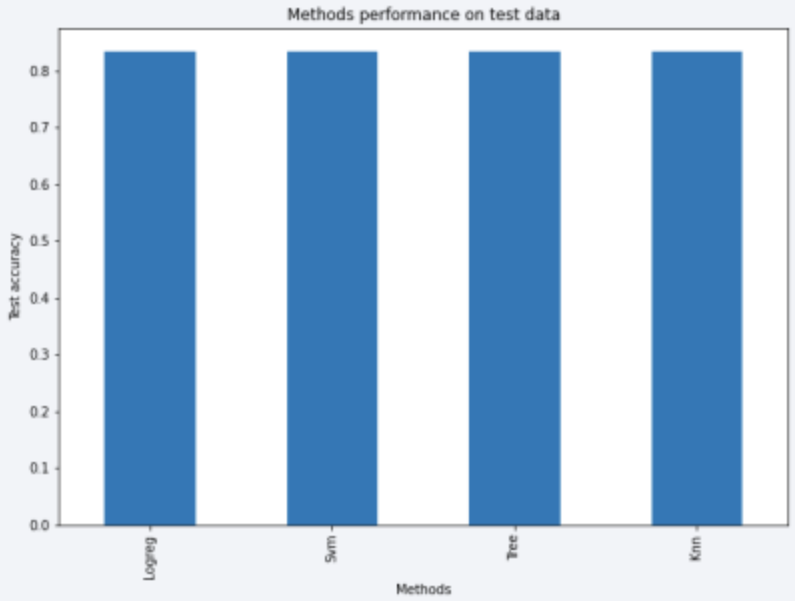


Section 5

Predictive Analysis (Classification)

Classification Accuracy

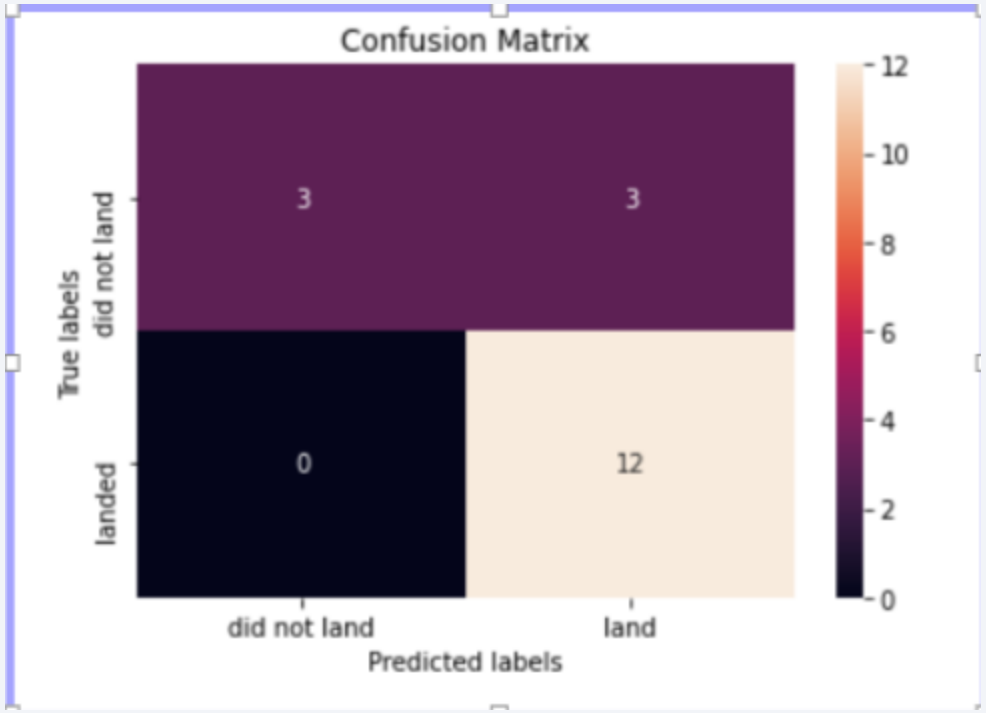




- Highest accuracy is tree.

Confusion Matrix

- The confusion matrix for the decision tree classifier mostly fails in false positives



Conclusions

- As flight amount increases, the success rate increases.
- Payload mass can not be used as a definitive criterion for success rate.
- The Decision tree classifier is the best machine learning algorithm for our task of classification.

Thank you!