IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Elias Dersahaguian
June 23, 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA with Data Visualization
  - EDA with SQL
  - Interactive Visual Analytics with Folium
  - Building a Dashboard with Plotly Dash
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Predictive Analytics result

Introduction

- Project background and context

SpaceX 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 SpaceX 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 you want to find answers
  - What features determine if a first stage will land successfully?
  - Can we use machine learning to predict the success of landing outcome of the first stage?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected through SpaceX API and through web scraping Wikipedia
- Perform data wrangling
  - Data cleaning of irrelevant attributes, complete missing data with the mean for some attributes, changing other attributes into one-hot encoding data fields
- 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 LR, SVM, DT and KNN classifiers and use GridSearchCV to find best parameters for each model

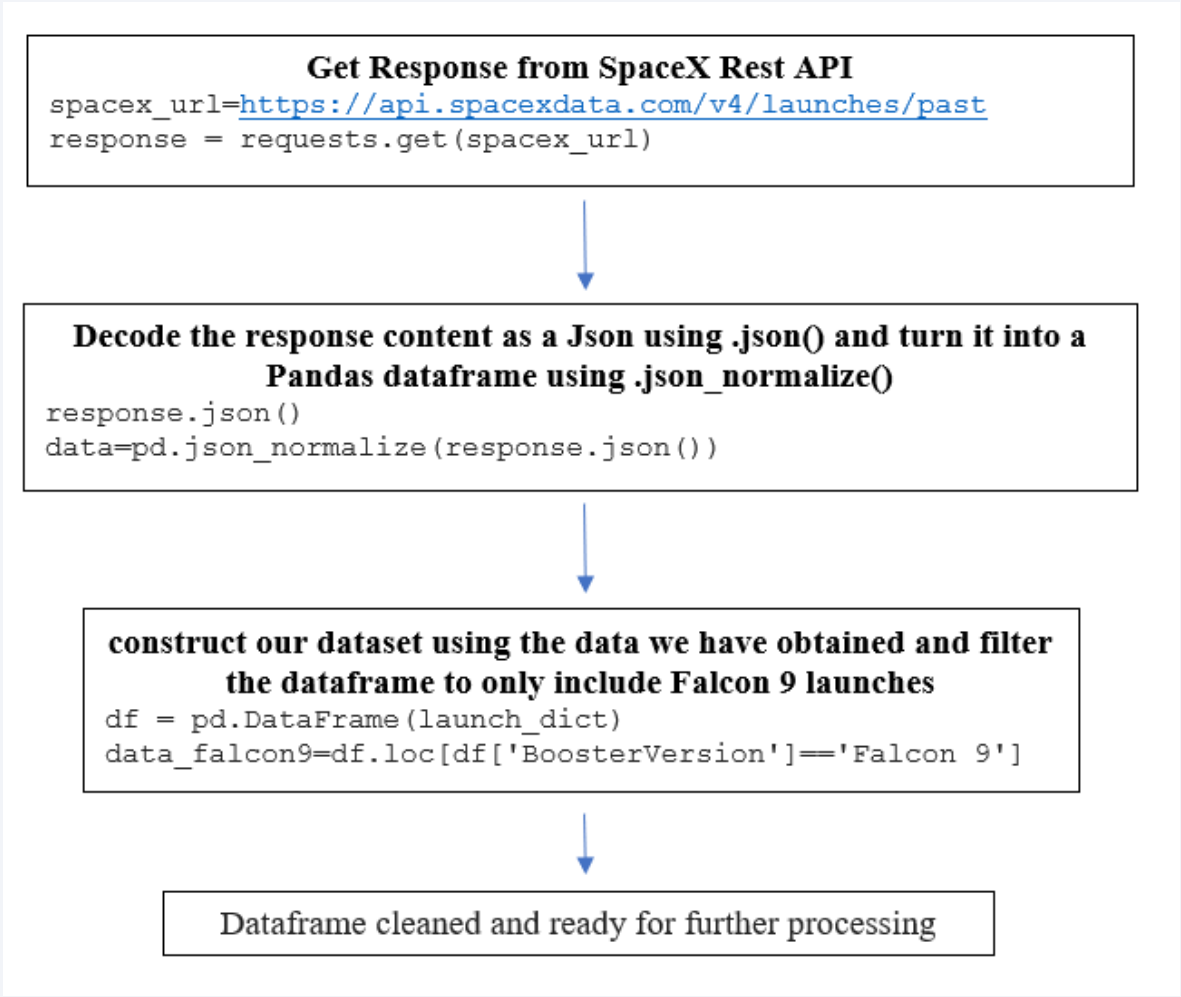
Data Collection

- Describe how data sets were collected.
  - Raw data was collected through the SpaceX API using GET requests
  - The response was decoded using .json and then was normalized into a dataframe using .json_normalize()
  - A new dataframe was created with information extracted from the original dataframe that is relevant for this study
  - The final dataframe was cleaned and processed for missing values, where the mean was used for 5 missing palyload masses. And only data about Falcon 9 booster version was included
  - Next, web scrapping with BeautifulSoup was used to parse HMTL tables from Wikipedia for further information about launch records including Falcon 9 booster versions for Falcon 9 launch records

Data Collection – SpaceX API

- Here is a flowchart explaining the SpaceX API data collection process

- And here is the link to the Jupyter Notebook of the process: https://github.com/eliasods/Applied_DS_Capstone/blob/main/Data%20Collection%20API.ipynb

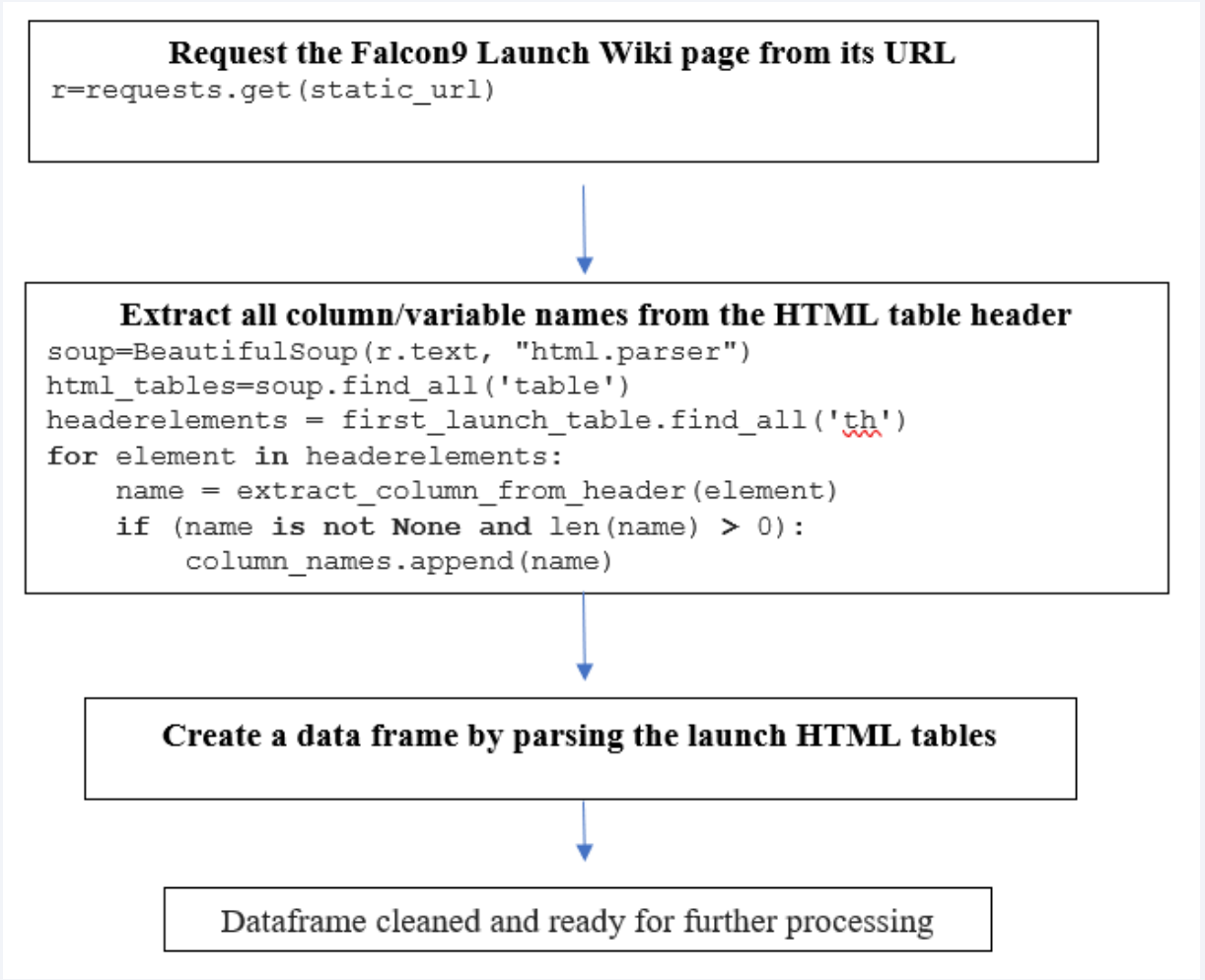


Data Collection - Scraping

- Here is a flowchart explaining the web scraping data collection process

- And here is the link to the Jupyter Notebook of the process:

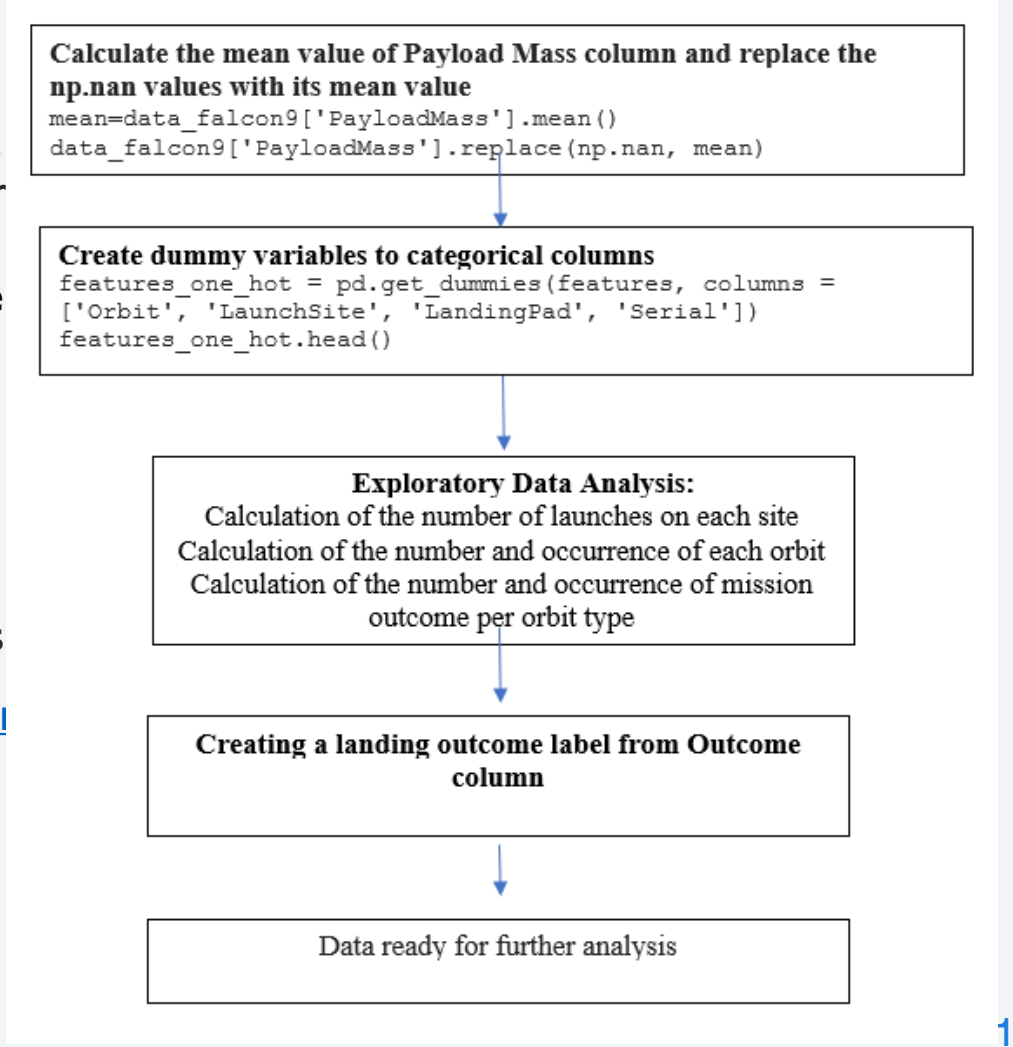
https://github.com/eliasods/Applied_DS_Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



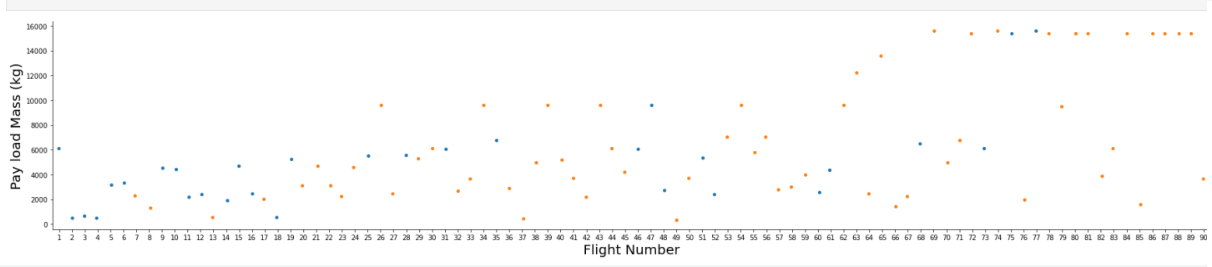
Data Wrangling

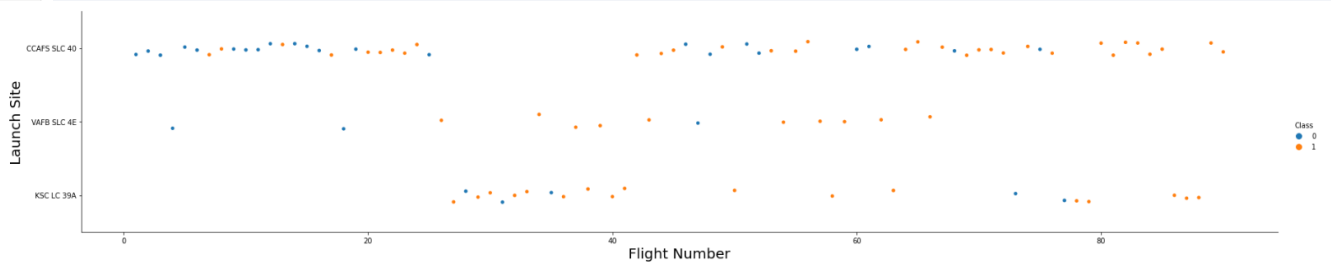
- Before the data was processed, some exploratory data analysis was performed including counting the number of launches on each site, the occurrence of different orbits, the number and occurrence of mission outcome per orbit type
- Then data was processed so that so that we landing outcome label was created from Outcome column
- here is the link to the Jupyter Notebook of the process

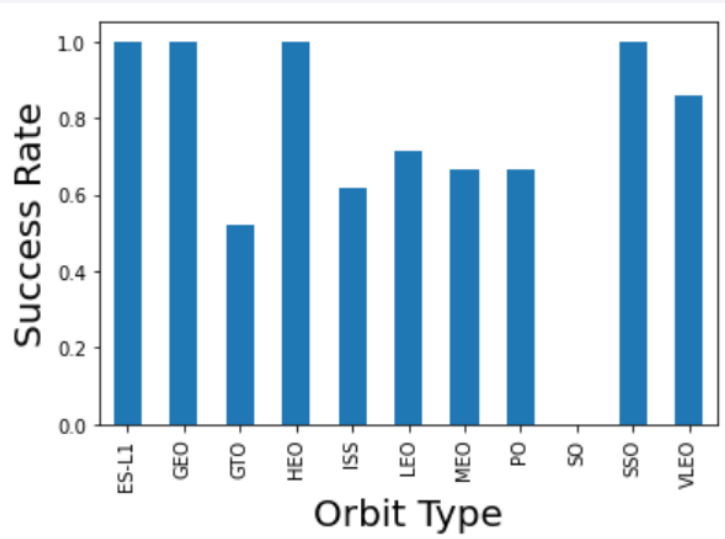
https://github.com/eliasods/Applied_DS_Capstone/blob/main/Data%20wrangling.ipynb

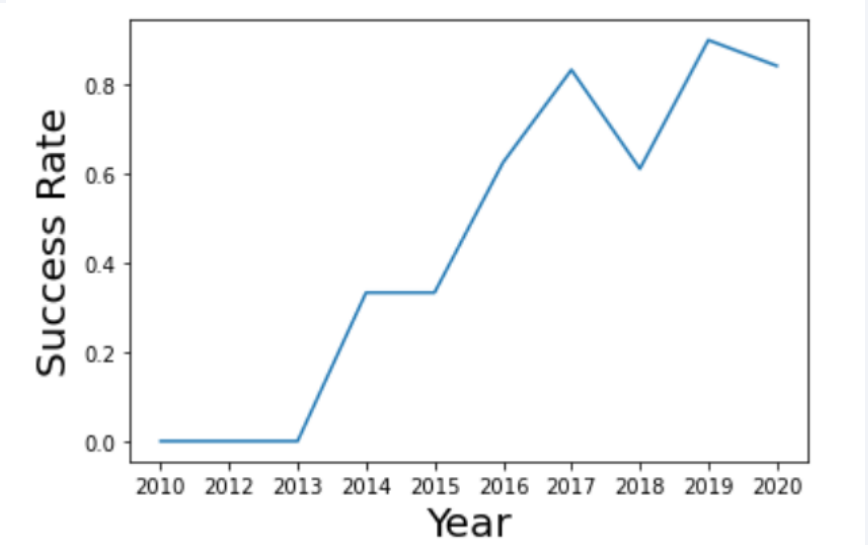


EDA with Data Visualization









- The graphs show vital information about success rates and different features
- We can visualize through the line graph the success rate progress over the years and the histogram the success rate as of each orbit type
- The scatter graphs show the relationship between payload mass, launch sites, and other features with each launch flight

https://github.com/eliasods/Applied_DS_Capstone/blob/main/EDA%20with%20dataviz.ipynb

EDA with SQL

- SQL queries performed include:
  - **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 records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.**
  - **Ranking the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.**

https://github.com/eliasods/Applied_DS_Capstone/blob/main/EDA%20with%20SQL_sqllite.ipynb

Build an Interactive Map with Folium

- Using Folium map, markers and popups were introduced in order to:
  - Mark all launch sites on a map
  - Mark the success/failed launches for each site on the map
  - Calculate the distances between a launch site to its proximities
- These objects were added to answer questions such as:
  - Are all launch sites in very close proximity to the coast?
  - Are launch sites in close proximity to coastline?
  - Do launch sites keep certain distance away from cities?

https://github.com/eliasods/Applied_DS_Capstone/blob/main/Interactive%20Viz%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Using Plotly Dash a Dashboard was created to display the following interactive charts:
  - A pie chart to show the total successful launches count for all sites, and show the Success vs. Failed counts for a specific launch site if it was selected
  - scatter chart to show the correlation between payload and launch success with a selected payload mass
- These plots were added because they give insights about the success rates on the bases of launch sites and payload masses

https://github.com/eliasods/Applied_DS_Capstone/blob/main/Interactive%20Dashboard%20with%20Ploty%20Dash.py

Predictive Analysis (Classification)

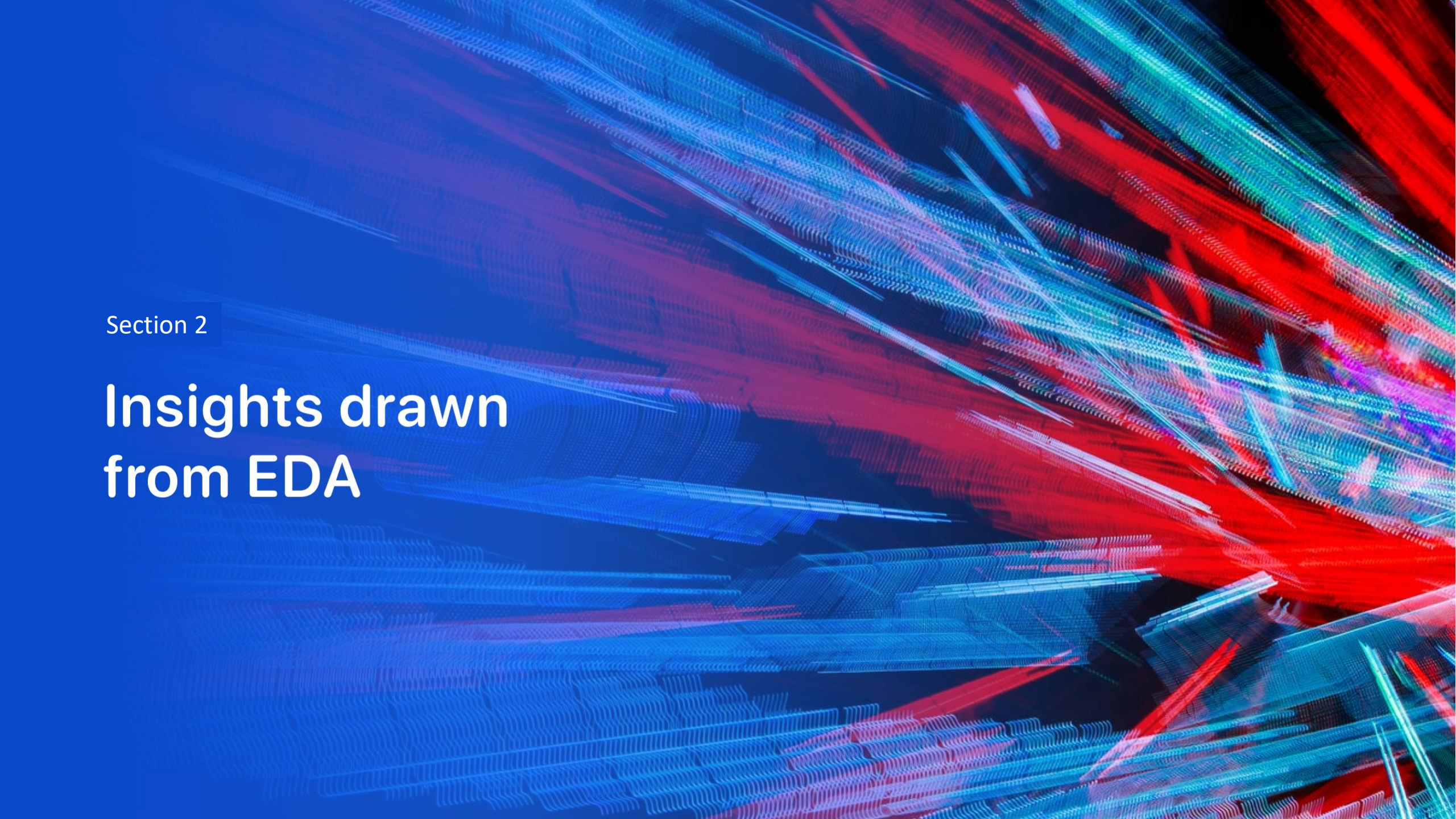
- In order to find the best classification model, we used logistic regression, support machine vector, decision tree, and k nearest neighbor.
- First, we split the dataset into training and testing sets
- Second, we used GridSearchCV to find the best parameters for each model via the training set
- Then we used the models with the best parameters to check the accuracy of each model on the testing set

https://github.com/eliasods/Applied_DS_Capstone/blob/main/Machine%20Learning%20Prediction.ipynb

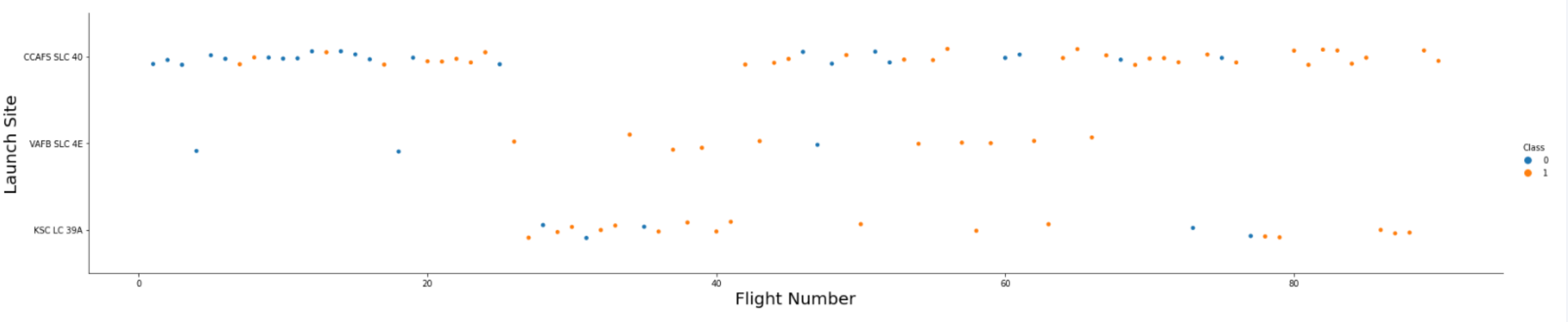
Results

The results are summarized in the following four sections:

- Insights drawn from Exploratory Data Analysis
- Launch sites proximities analysis
- Interactive dashboard with Plotly Dash
- Predictive analysis results

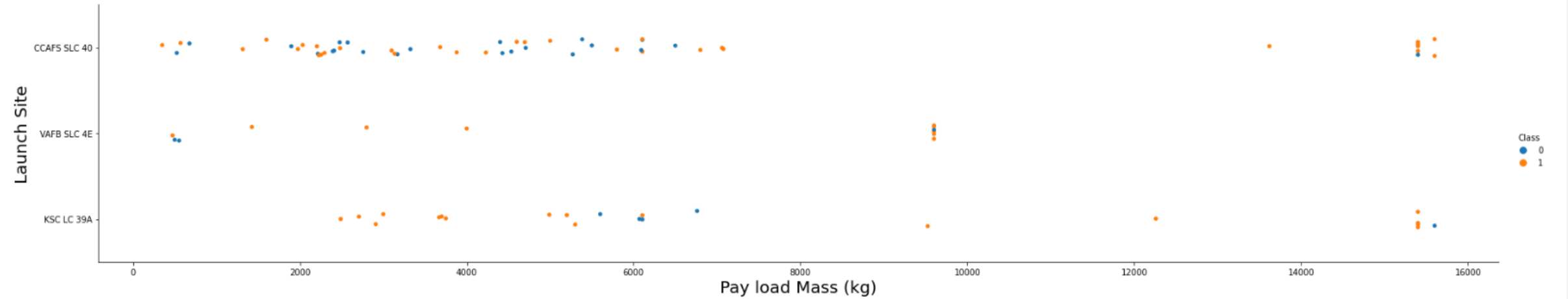

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



We notice that the success rate of all launch sites were higher in the second half of the launch record than the first half, meaning that all launch sites have improved their success rate with flight number

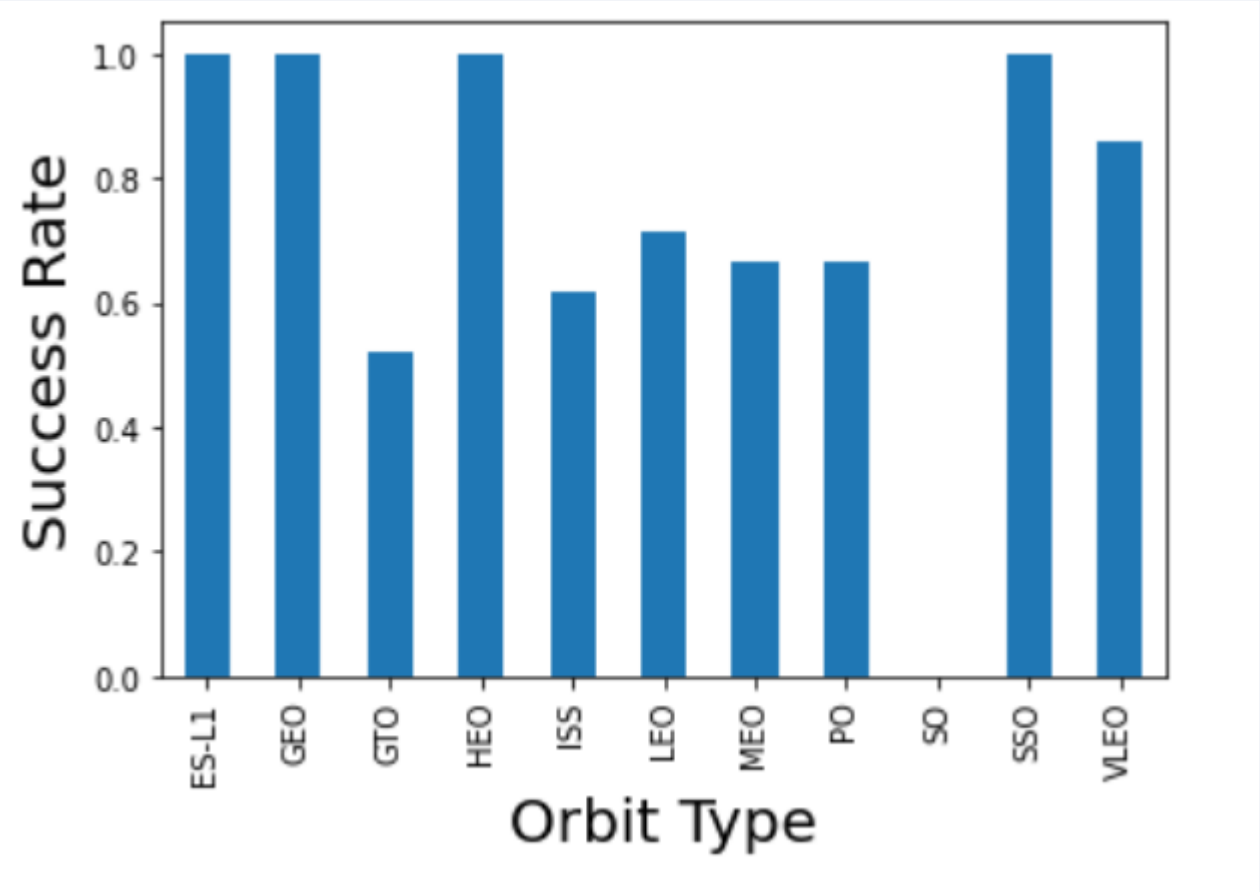
Payload vs. Launch Site



While CCAFS SLC 40 success rate is higher for heavier payloads(>15,000), the success rate for KSC LC 39A is high both at the lower mass and higher mass payloads (<5,000 and > 15,000)

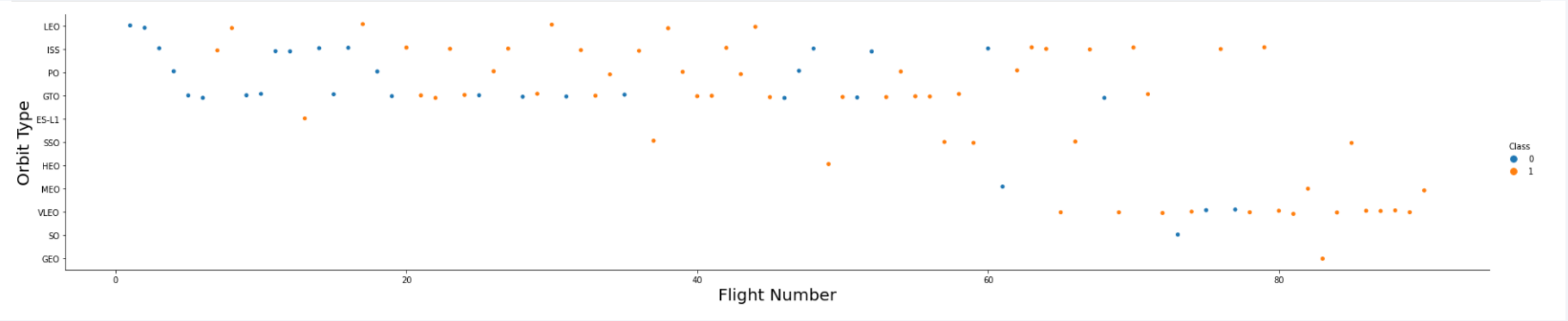
VAFB SLC 4E success rates are high within the range of payload mass it carries, it was never used for higher payloads (>10,000)

Success Rate vs. Orbit Type



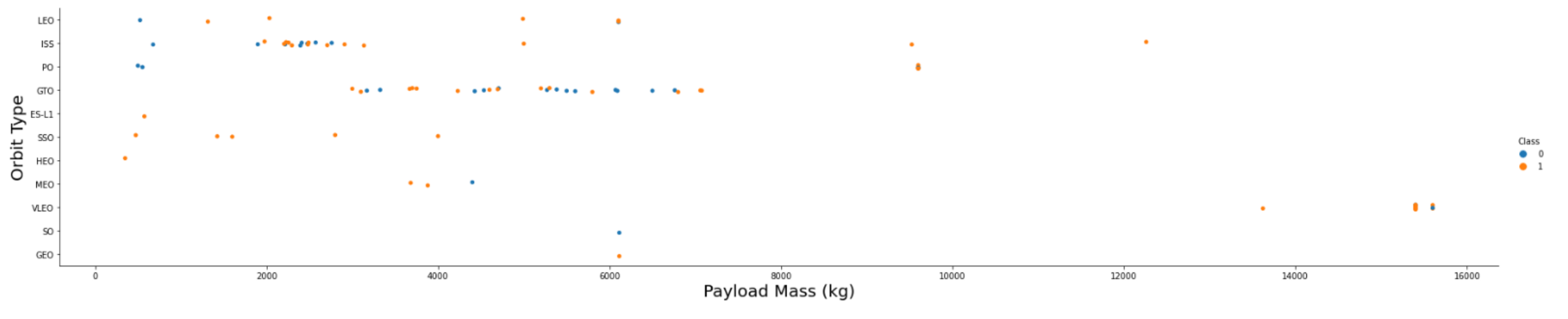
- Success rate for ES-L1 GEO, HEO, SSO, were 100%
- While the poorest outcome was when the orbit was GTO (50% success rate)

Flight Number vs. Orbit Type



LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

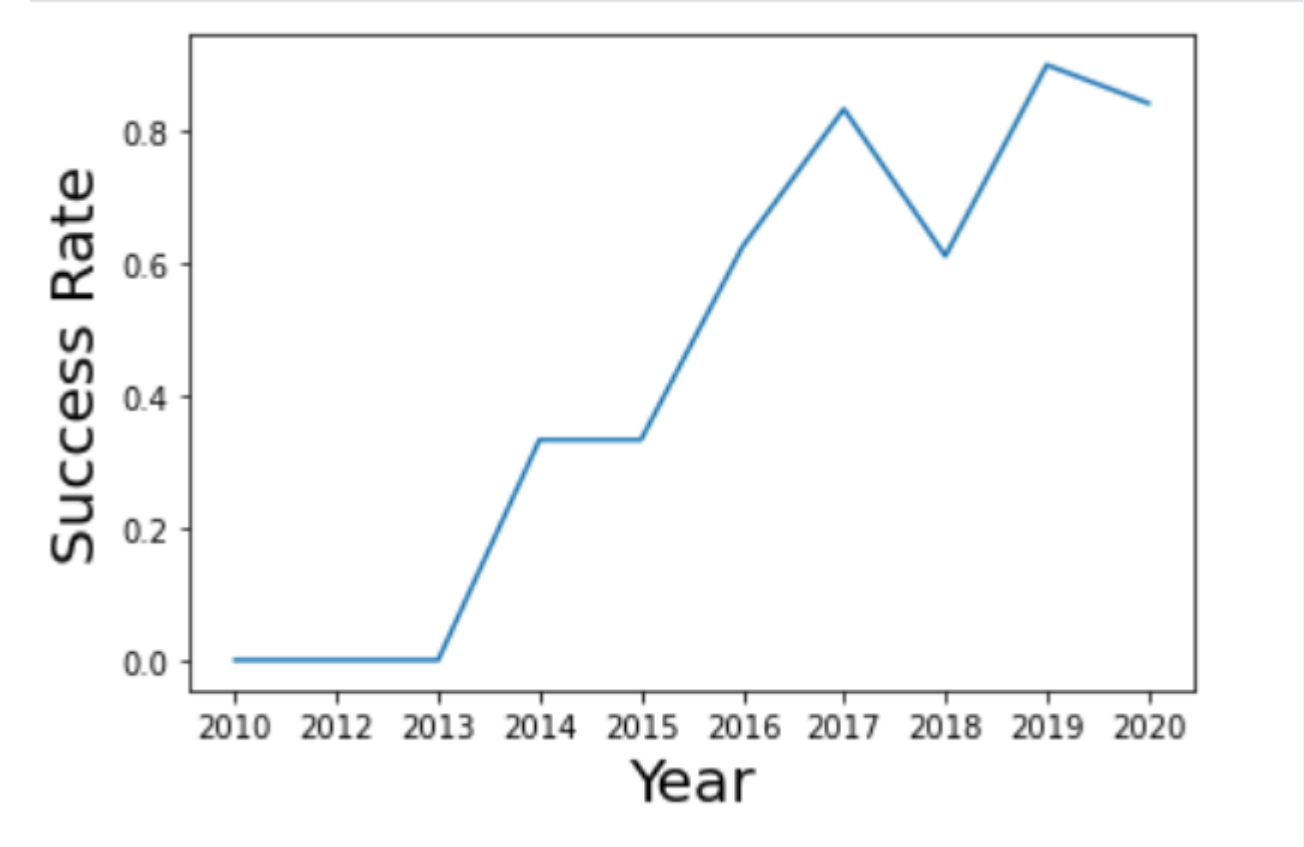
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here

Launch Success Yearly Trend

The success rate passed 0.6 the first time in 2016, and fluctuated between 0.6 and 0.9 between 2016 and 2020



All Launch Site Names

Display the names of the unique launch sites in the space mission

```
%%sql
select DISTINCT Launch_site from SPACEXTBL
```

```
 * 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'

Display 5 records where launch sites begin with the string 'CCA'

```
%%sql
select * from Spacextbl Where Launch_site like 'CCA%' Limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	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 carried by boosters launched by NASA (CRS)

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

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

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

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%%sql
SELECT Date FROM SpacexTBL WHERE "Landing _Outcome" = 'Success (ground pad)' limit 1;
```

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

Date
22-12-2015

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%%sql
Select Distinct Booster_Version from spacextbl where ("Landing _Outcome" = 'Success (drone ship)') and (PAYLOAD_MASS__KG_ >4000) and (PAYLOAD_MASS__
```

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

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

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

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

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

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%sql
Select Distinct Booster_Version from spacextbl where PAYLOAD_MASS__KG_ = (Select Max(PAYLOAD_MASS__KG_) from spacextbl)
```

```
 * 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

2015 Launch Records

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%%sql
SELECT substr(Date, 4, 2) as Month, "LANDING _OUTCOME", BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL
WHERE "Landing _Outcome" LIKE 'Failure%' AND substr(Date,7,4)='2015'
```

```
 * 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

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

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%%sql
SELECT "LANDING _OUTCOME", COUNT("LANDING _OUTCOME") AS TOTAL_NUMBER FROM SPACEXTBL
WHERE DATE BETWEEN '04-06-2010' AND '20-03-2017'
GROUP BY "LANDING _OUTCOME"
ORDER BY TOTAL_NUMBER DESC
```

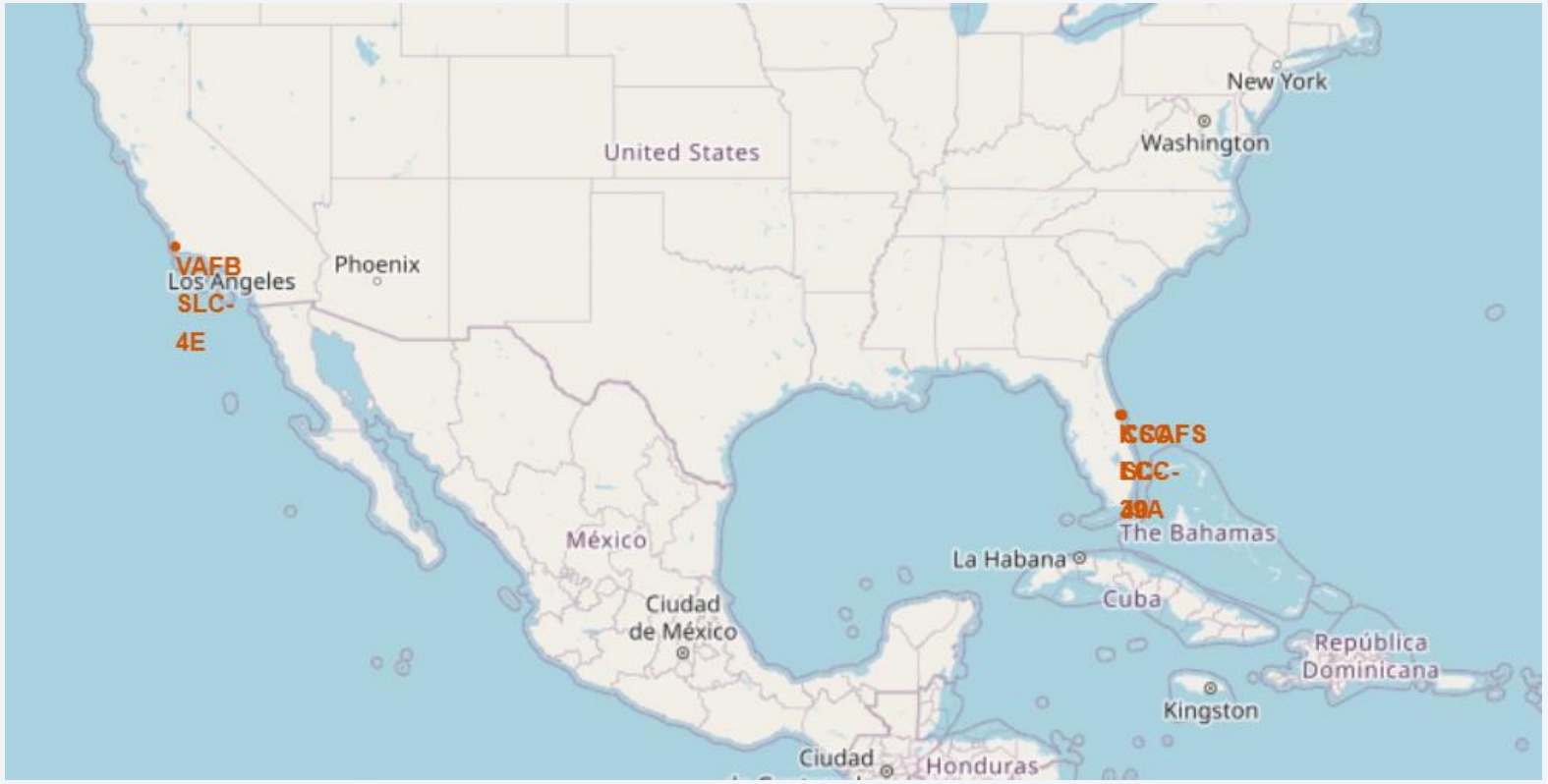
```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	TOTAL_NUMBER
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

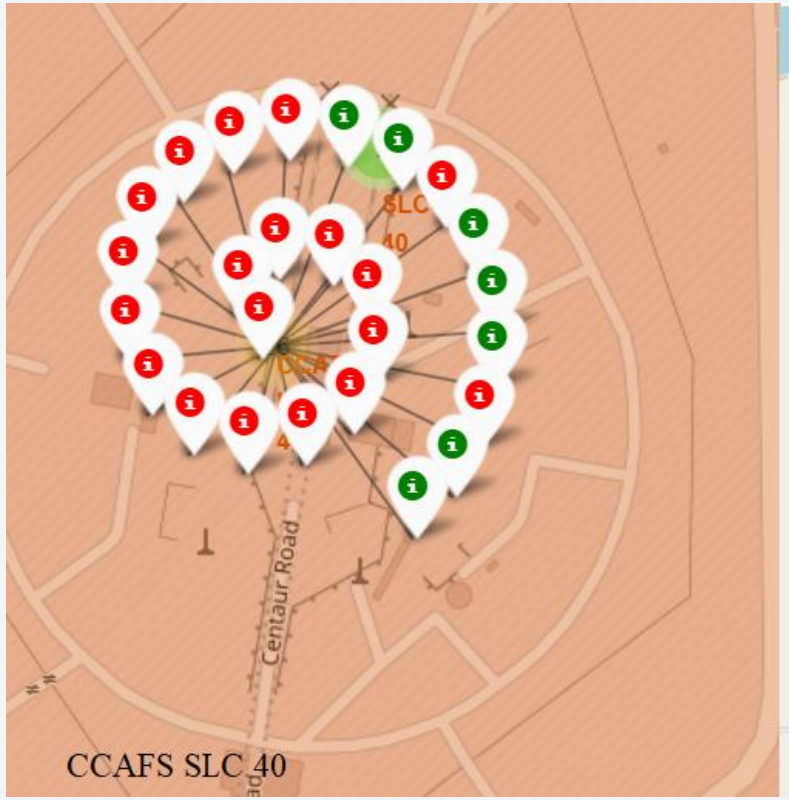
Launch Sites Proximities Analysis

All Launch Sites

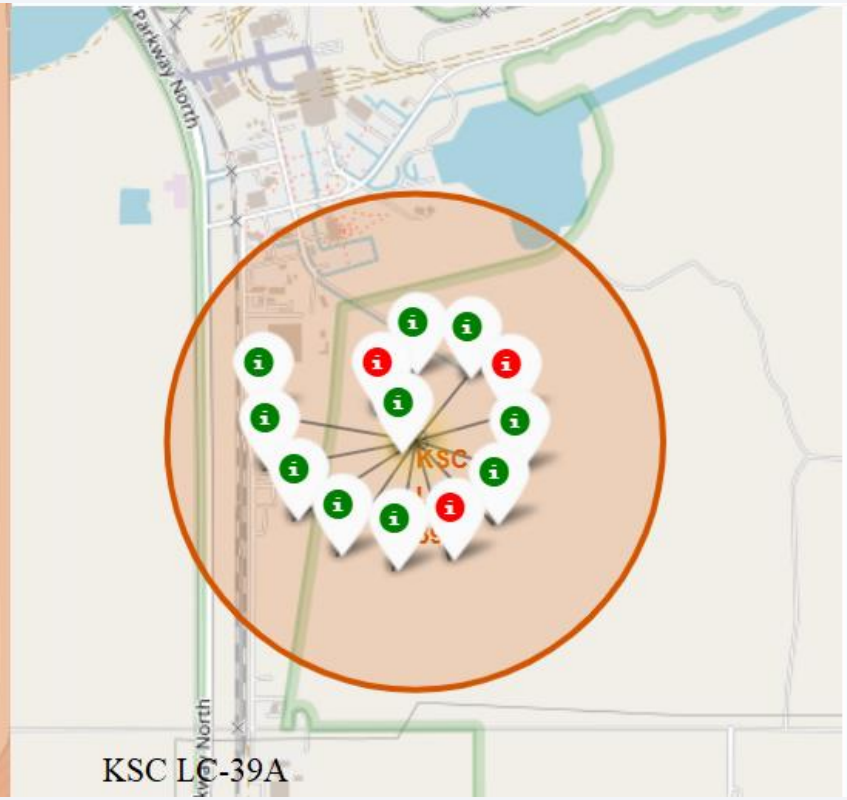


All launch sites are within the US territory. There are two main site areas: Florida and California, however both areas are close to the coastline

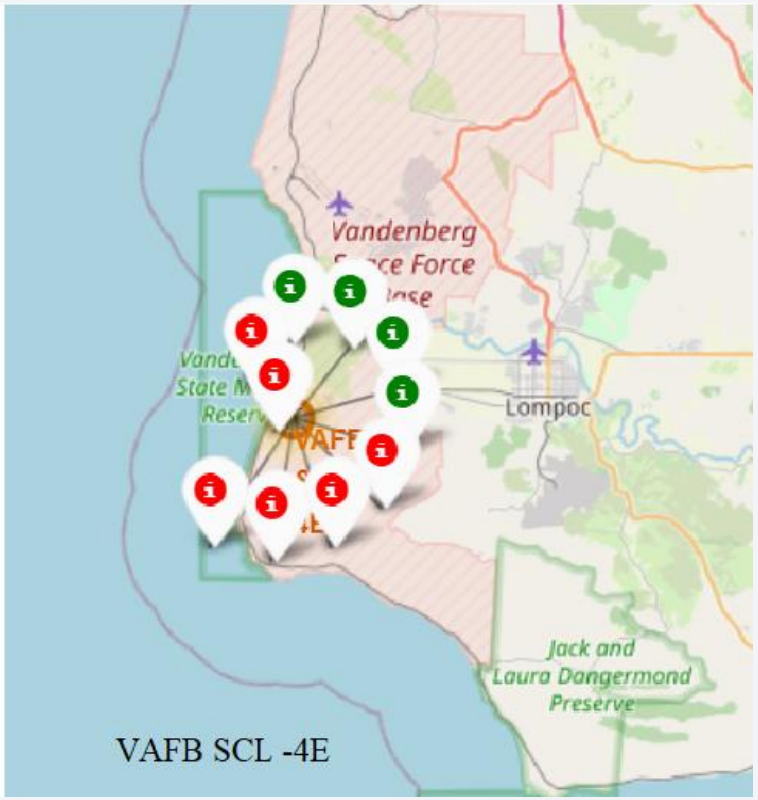
The Success/Failed Launches for Each Site



CCAFS SLC 40



KSC LC-39A



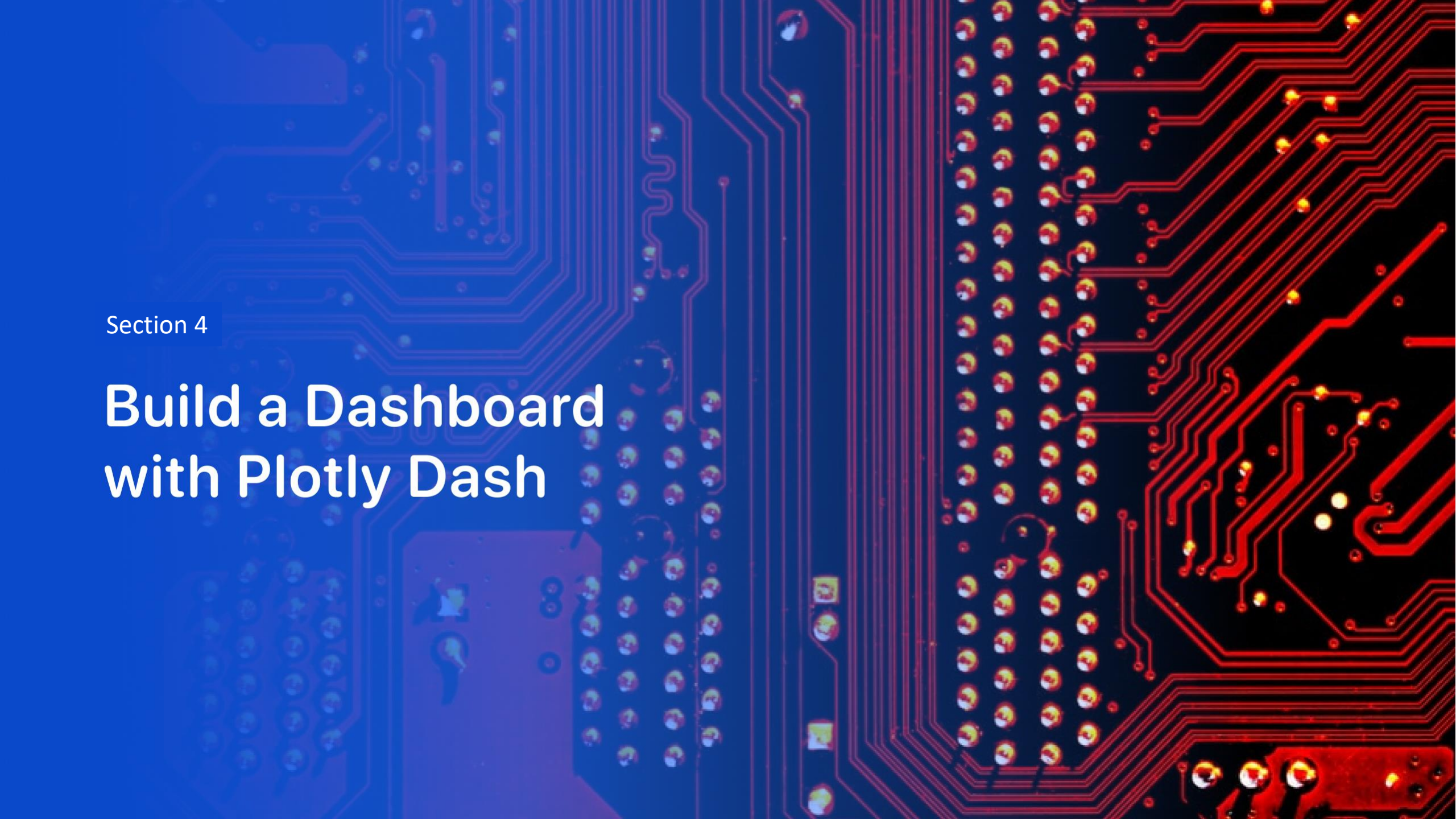
VAFB SCL -4E

We can notice easily through this visualization that the launching site KSC LC-39A has the highest success rate of all launch sites

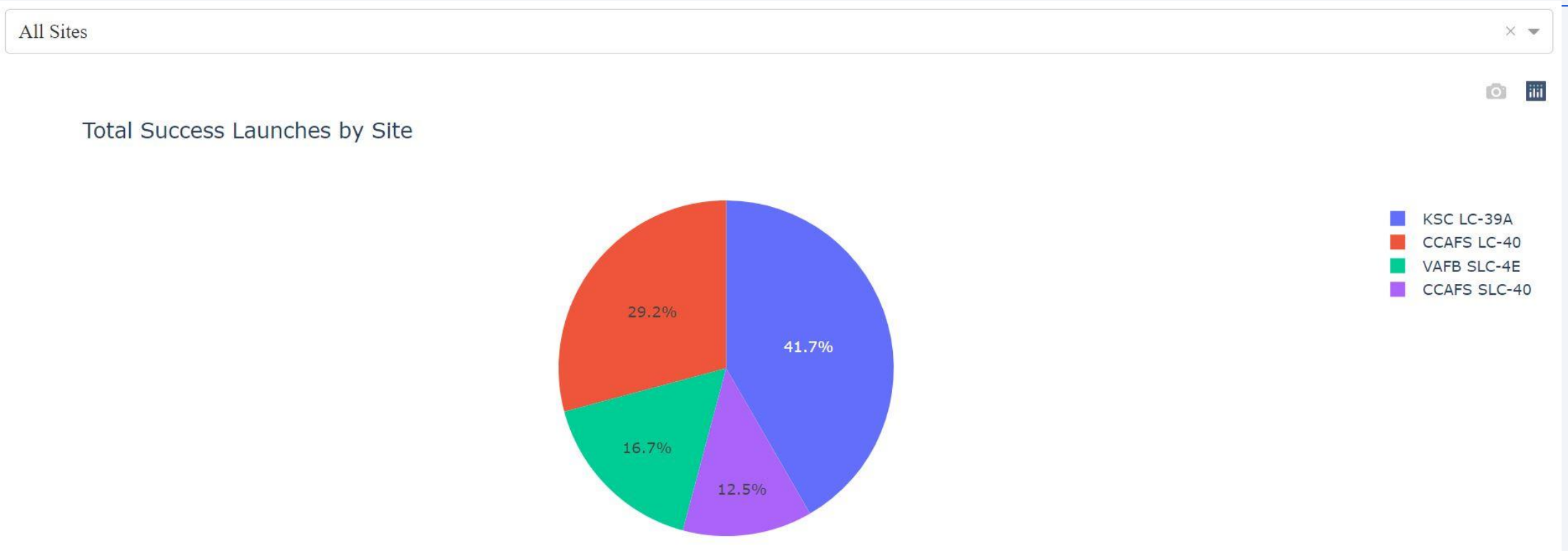
The Distancex between a launch site to its proximities

We notice that all launch sites are very close to the coastline. And that they are relatively close to cities (CCAFS SLC-40 within 25km to Titusville and similar distance of VAFB SCL-4E to Santa Maria)

Here we show two of several distance measurements that can be taken.

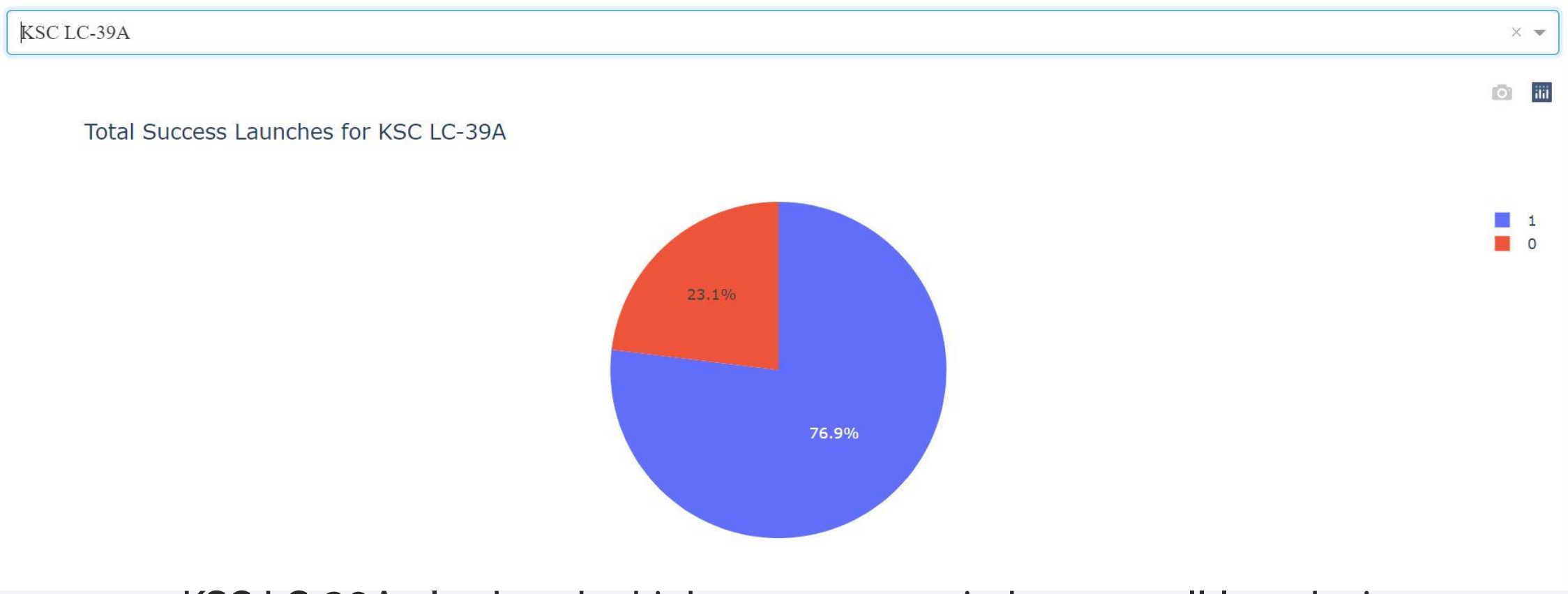

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for all Sites



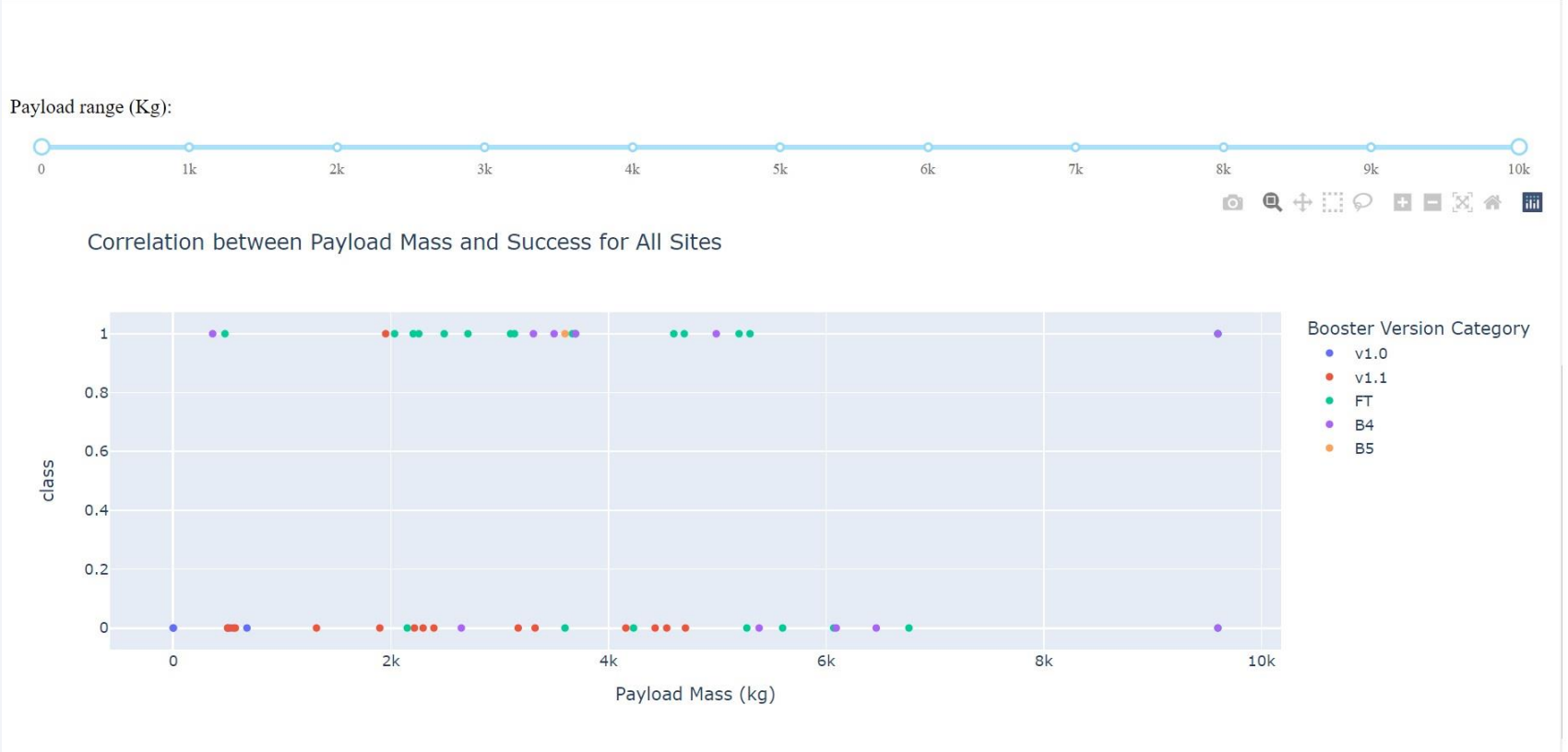
KSC LC-39A has the highest success rate between all launch sites.

Site with highest success ratio



KSC LC-39A also has the highest success ratio between all launch sites

Payload vs. Launch Outcome for all Sites

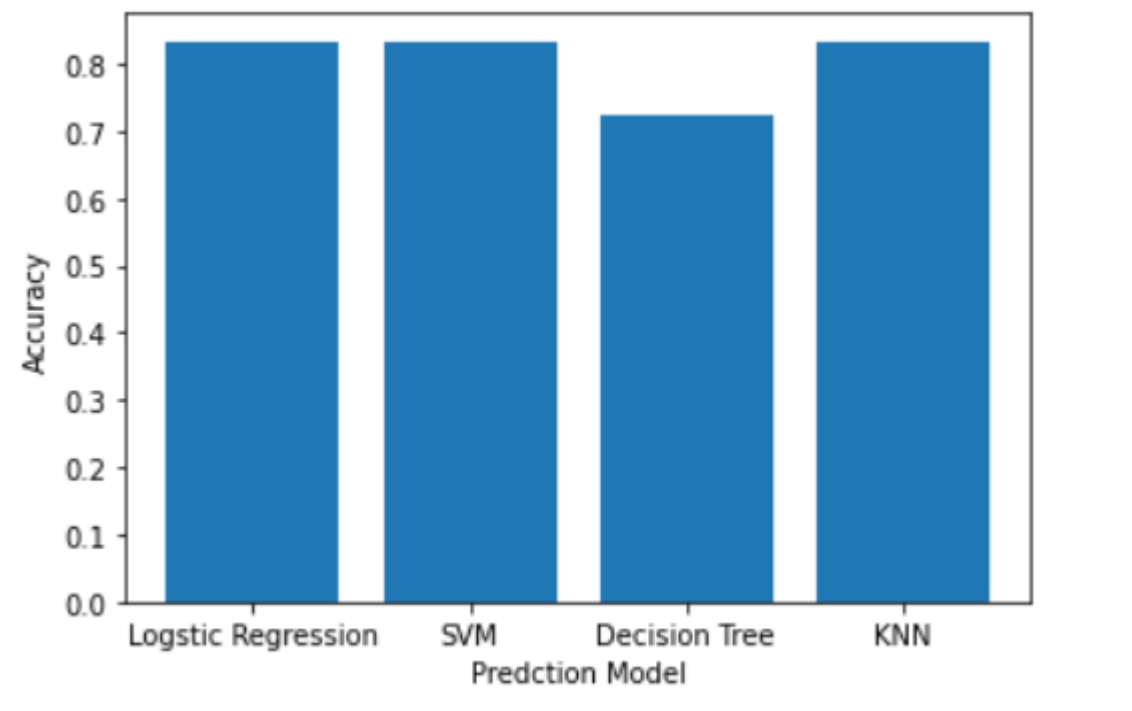


We notice that the most successful Booster Version is FT, and that it is most successful for payload masses between 2000 and 40000 kg

Section 5

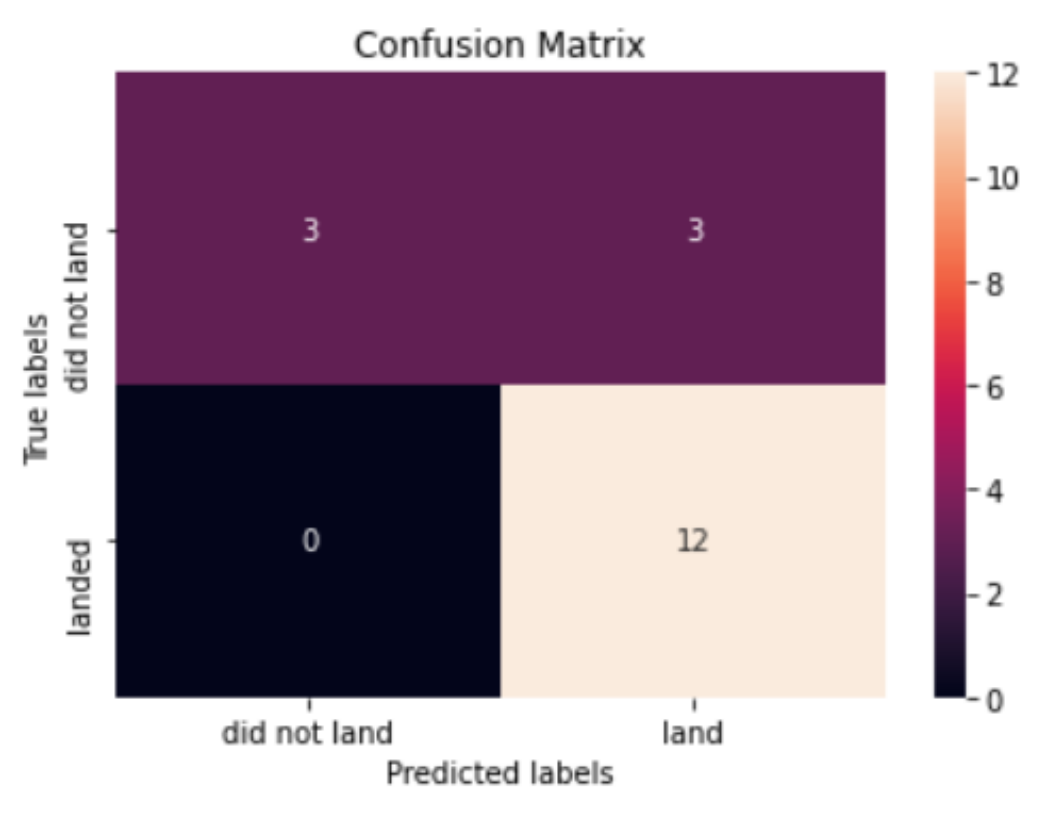
Predictive Analysis (Classification)

Classification Accuracy



- Logistic Regression, SVM and KNN models had equal accuracy of 83.33%
- While the Decision tree gave a lower accuracy score of 72.22%

Confusion Matrix



All three models including logistic regression, SVM, and KNN showed equal accuracy of 83.33% and a similar confusion matrix displayed here

Conclusions

We can conclude that:

- Success of the first stage landing increased with flight number, with a success rate above 0.6 starting in 2016 and reached its peak in 2020
- Orbits ES-L1, GEO, HEO, and SSO had the most success rate
- KSC LC-39A had the most successful launches of any sites
- The Logistic Regression, SVM and KNN models where the most successful models in predicting the outcome of a successful first stage landing

Thank you!