IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kelly Queen
1.24.2024

Outline

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

Executive Summary

- **Summary of Methodologies**
  - Data collection via API calls
  - Data collection via webscraping techniques
  - Data preparation
  - SQL analysis
  - Data visualization analysis
  - Geographical analysis with Folium
  - Dashboard generation with Dash
  - Machine Learning method testing and selection

- **Results**
  - Specific orbital type launches show most success
  - Payload mass shows direct correlation to successful launch
  - Launch site location selection is crucial
  - Later launches show higher success rates

Introduction

- SpaceX has operated without a direct competitor for years. SpaceY will step into the space race and bring competition to the affordable rocket launch market.
- To do this we must be able to keep the cost of launches down and determine the reusability of rocket stages. SpaceX reuses the first stage of their launches, and this is the primary driver of their significantly lower launch prices. SpaceX can operate at an average cost of 62m per launch vs other providers quoting 165m per launch.
- SpaceX publishes their data publicly; and their data set will be used for analysis.
- Analysis seeks to determine the cost of each launch by predicting whether or not the first stage of SpaceX launches will land successfully based on previous launch data.
  - What features determine whether stage 1 will land successfully?
  - What conditions are required for a successful landing?
  - Success rate of landing based on specific factors, such as payloadmass and orbit type.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was extracted from SpaceX via API and webscraping Wikipedia launch tables
- Perform data wrangling
  - Categorical data regarding launch success converted to continuous values and appended to the dataframe
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

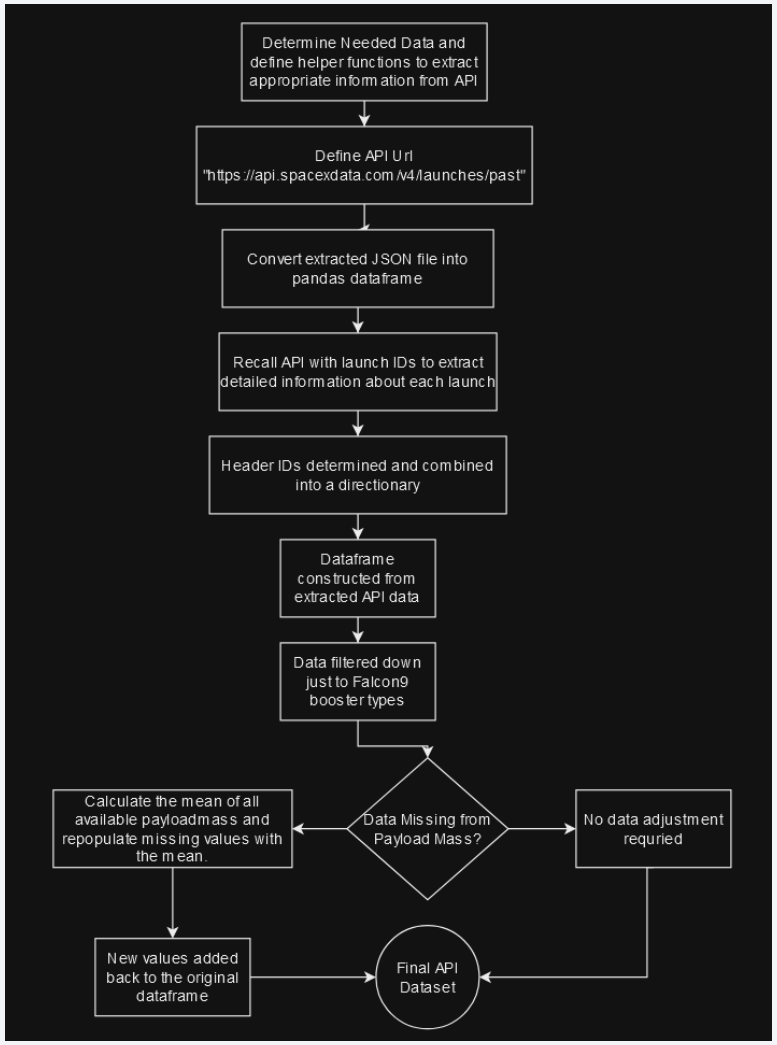
Data Collection

- Data was collected from various sources
  - A GET request was used to extract directly from the SpaceX API
  - The received JSON file was converted into a pandas dataframe for analysis
  - Data was cleaned with a focus on removing and appending missing data
  - Webscraping techniques were utilized to extract table data from the Falcon9 launch records Wikipedia entry, and then decoded using BeautifulSoup
  - Cleaned API and Webscraping data was combined into a singular dataframe for further analysis

Data Collection – SpaceX API

- To determine past data for rocket, payload, launchpad, and core features – an API call was made directly to the SpaceX datastore.

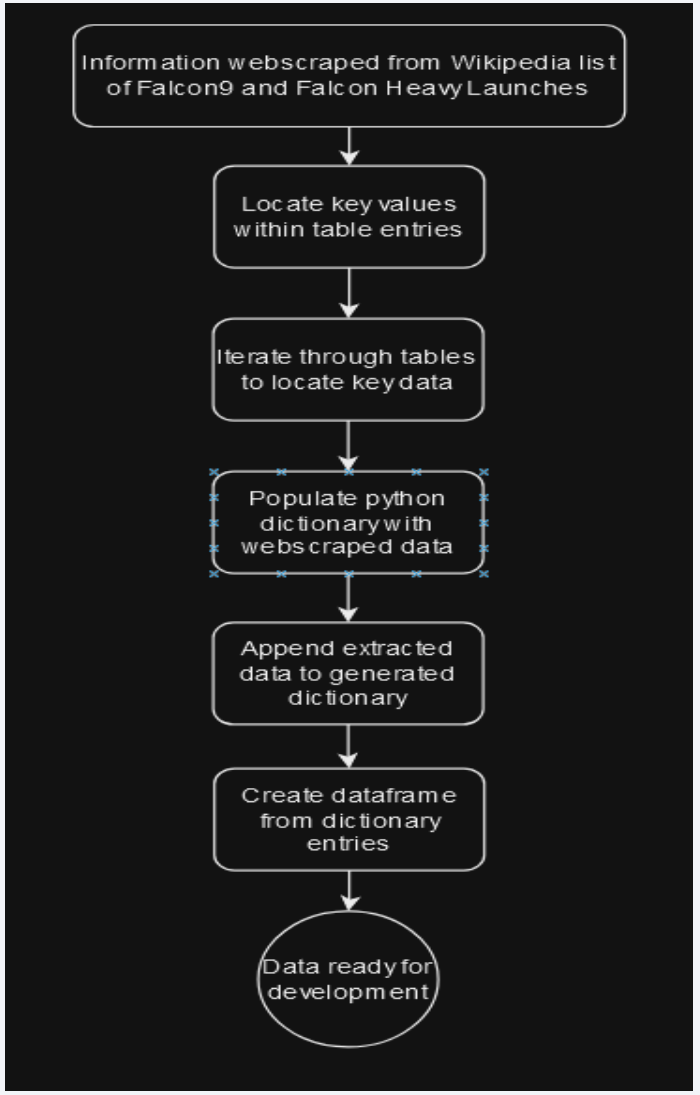
- Github URL
  - https://github.com/sublimetiger/spacey/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- To determine success/failure mode of the SpaceX launches, data was scraped from the Falcon9 and Falcon Heavy Launches Wikipedia HTML chart
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

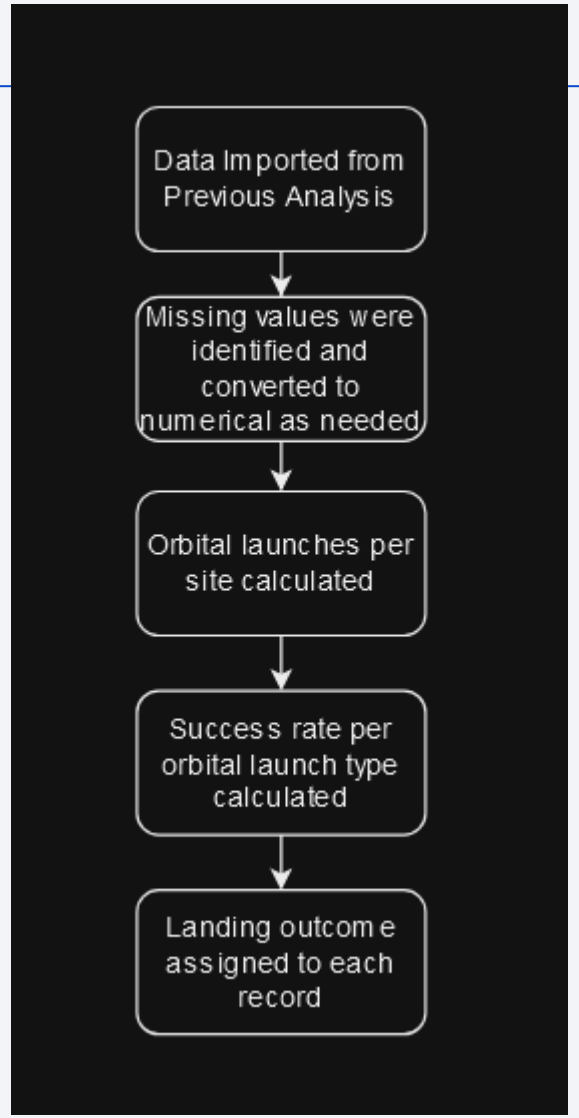
- Github URL
  - https://github.com/sublimetiger/spacey/blob/main/jupyter-labs-webscraping.ipynb



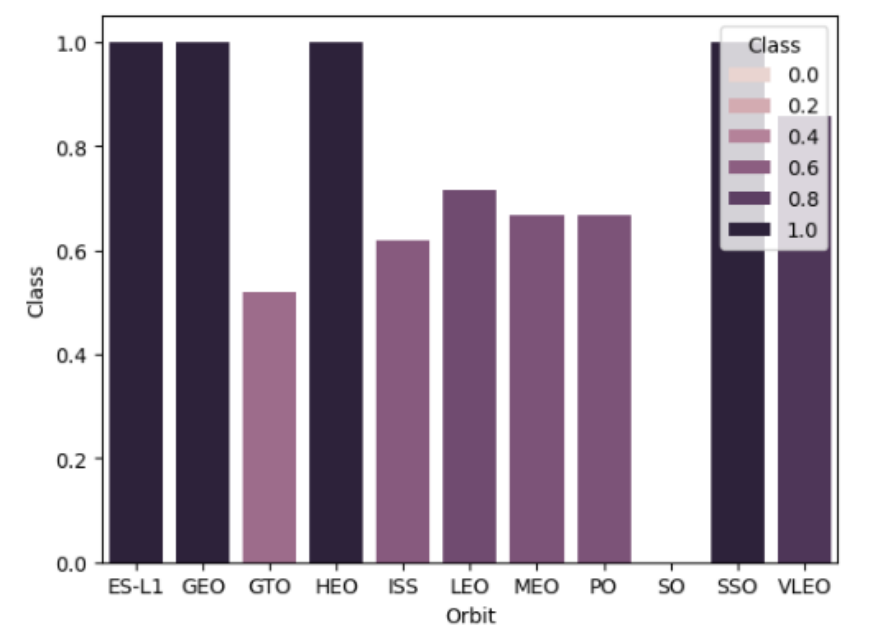
Data Wrangling

- Data was categorized to determine categorical or continuous and converted as necessary
- The number of launches and successful launches per site was calculated
- Outcome labels were determined for each record were the first stage landed successfully

- Github Link https://github.com/sublimetiger/spacey/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

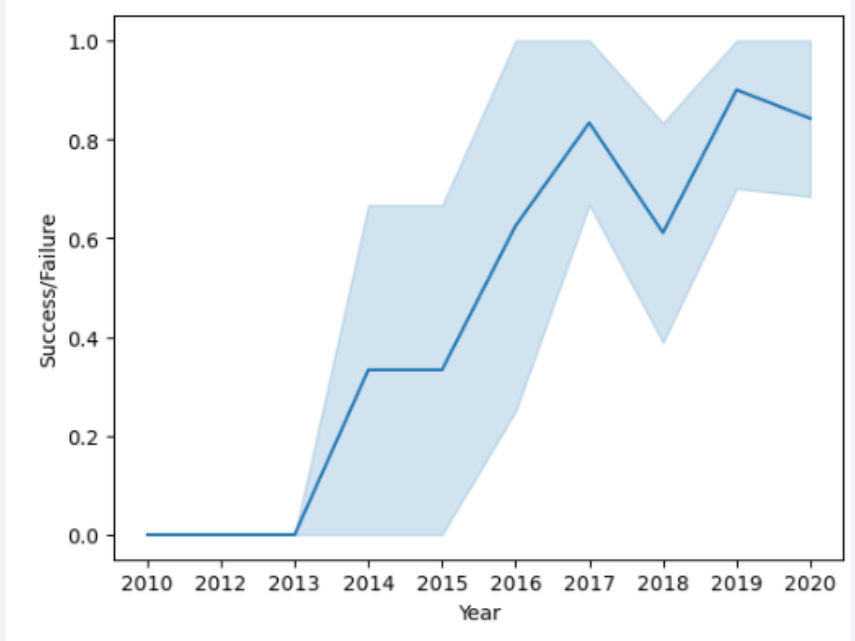


EDA with Data Visualization



Data was explored through viewing the visual relationship between orbit type and success rate, payloadmass and launch site, flightnumber and orbit.

Left chart indicates the success rate by orbit type and right visualizes success rate by year.



- Github URL
  - https://github.com/sublimetiger/spacey/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Analysis of the data continued with SQL
  - Unique launch sites were determined
  - Launch sites narrowed to CCAFS sites
  - Total payload volume for NASA launches calculated
  - Average payload mass carried by the F9 v1.1 booster determined
  - First successful landing in a ground pad date found
  - Drone shipments that were successful with midrange payloads
  - Total number of success and failure missions
  - Boosters that have carried the max mass
  - Ranking of the landing outcome types by count

- Github Notebook:
  - https://github.com/sublimetiger/spacey/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- All launch sites were marked, and notations to mark the success or failure of each launch indicated with:
  - Markers
  - Circles
  - Lines
- Class 1 was assigned to Success, and Class 0 Assigned to Failure
- Launch sites with high success rates were denoted with color-labeled marker clusters
- Distances between launch sites and proximal entities were calculated
  - Nearest Coastline
  - Nearest Railway
  - Nearest Highway
  - Nearest City
- The distances to nearest proximal entities answered questions such as are launch sites always near mass transportation routes for large items, and how close launch sites are to cities.

- Github URL
- https://github.com/sublimetiger/spacey/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

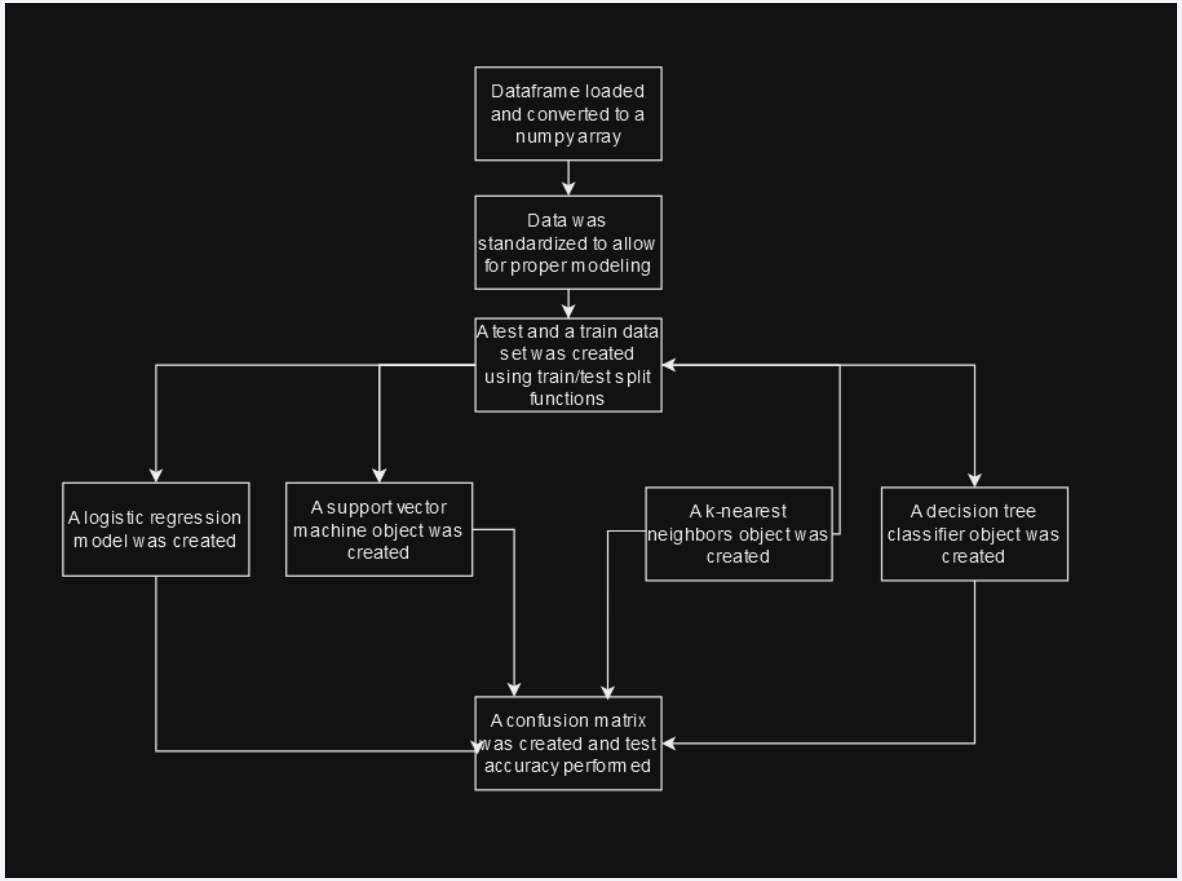
- An interactive dashboard was constructed in Plotly Dash featuring
  - A pie chart indicating the success count for all launch sites
  - Success count on payload mass for all sites with reference to booster type
  - A dropdown to narrow search by launch site
  - A slider to narrow view to a specific size of payload mass

- Github URL
  - https://github.com/sublimetiger/spacey/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

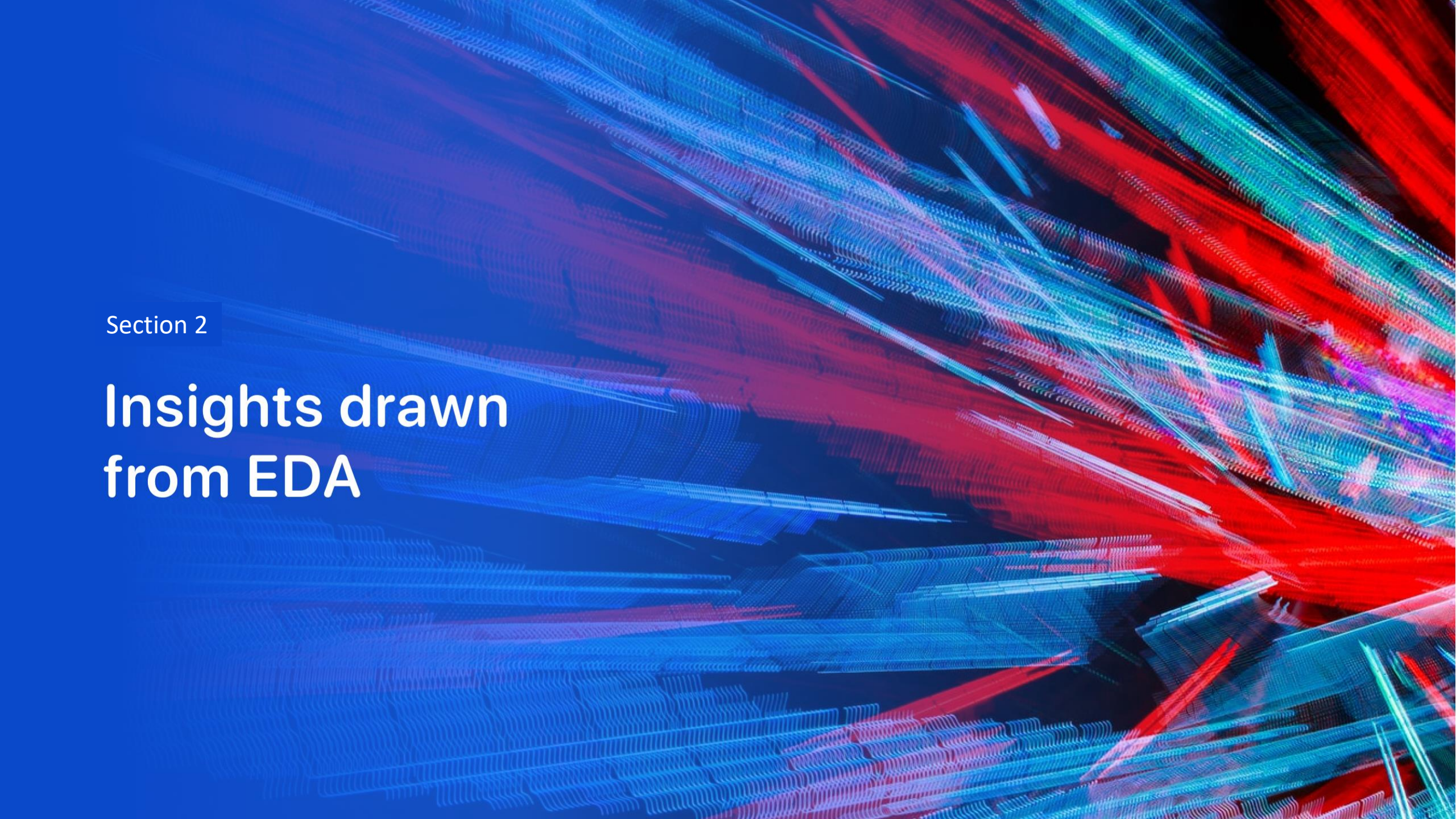
- After transforming the data into a numpy array, the data was split into train and test sets
- Different machine learning models were build and tuned to different hyperparameters
- Accuracy for each model was tested
- A confusion matrix was developed for each model
- Using feature engineering and algorithm tuning, the best performing model was found

- Github URL
  - https://github.com/sublimetiger/spacey/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



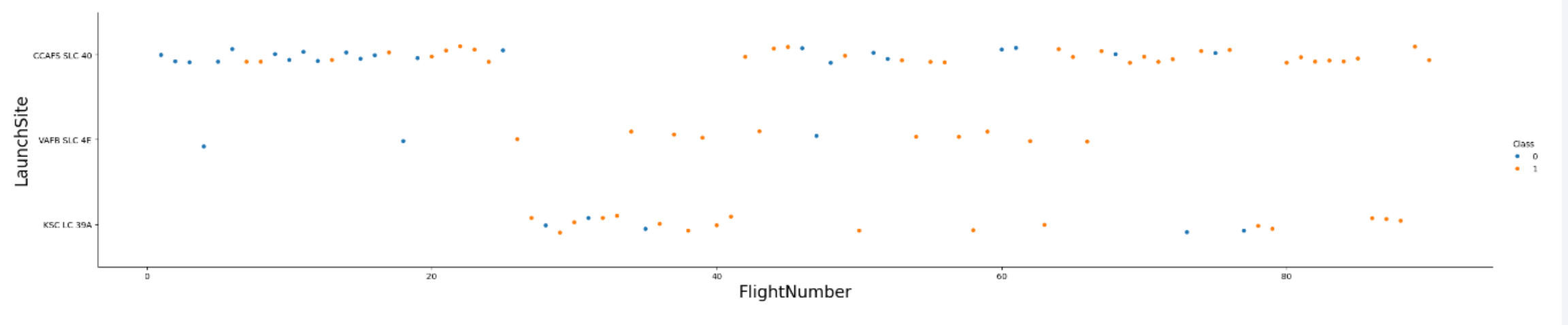
Results

- Payloadmass and Orbit type are primary drivers of success rates
- All machine learning models operated within a narrow margin, with false positives showing as the “biggest problem” with prediction


Section 2
Insights drawn from EDA

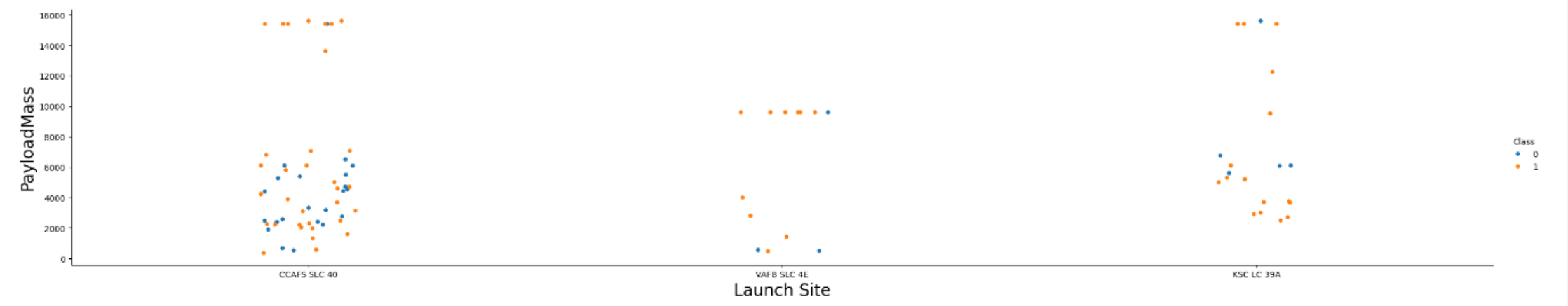
Flight Number vs. Launch Site

- From the scatter we can extrapolate that the higher the flight number at the launch site, the greater likelihood of success



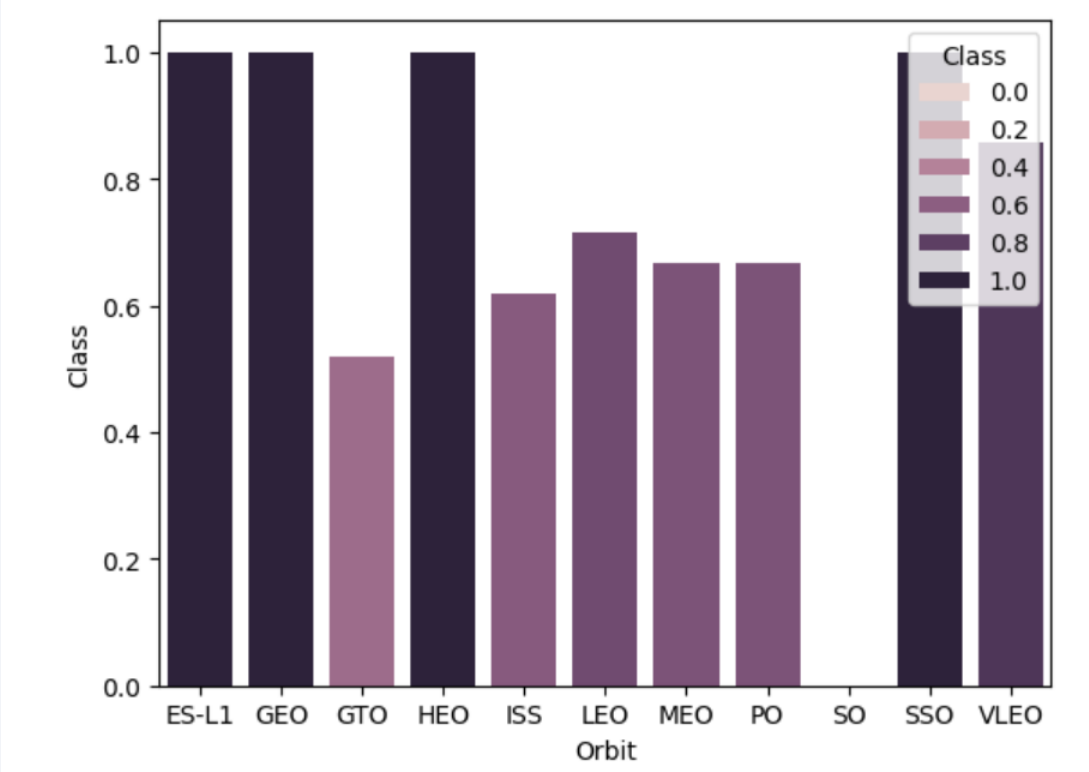
Payload vs. Launch Site

- VAFB SLC does not launch payloads higher than 10k
- The higher the payload, the higher the average for success in all sites
- Payloads have a tendency to run below 8k, or higher than 15k with few payloads between that range



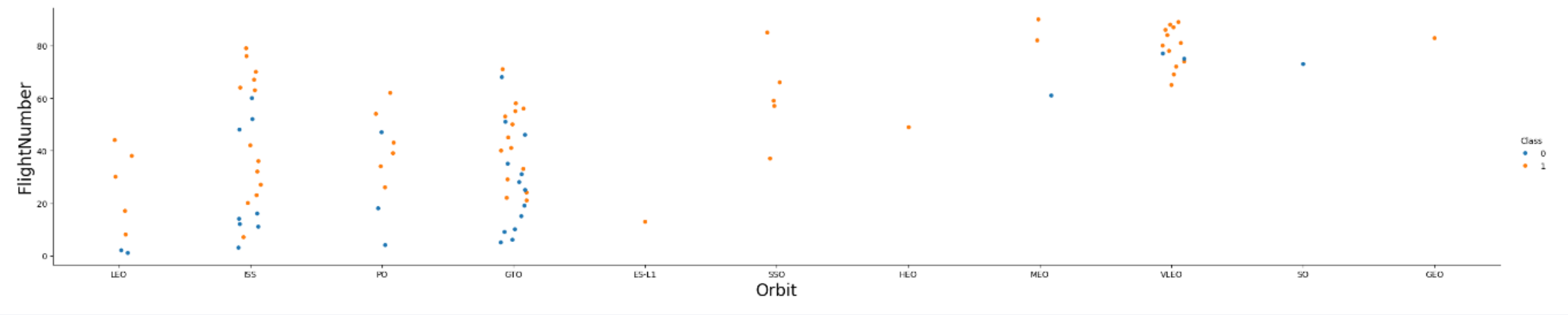
Success Rate vs. Orbit Type

- ES-L1 (language point L1,) GEO (circular geosynchronous,) HEO (geocentric,) and SSO (sun-synchronous) have the highest rates of success by orbit type
- SO (heliosynchronous) did not demonstrate any success



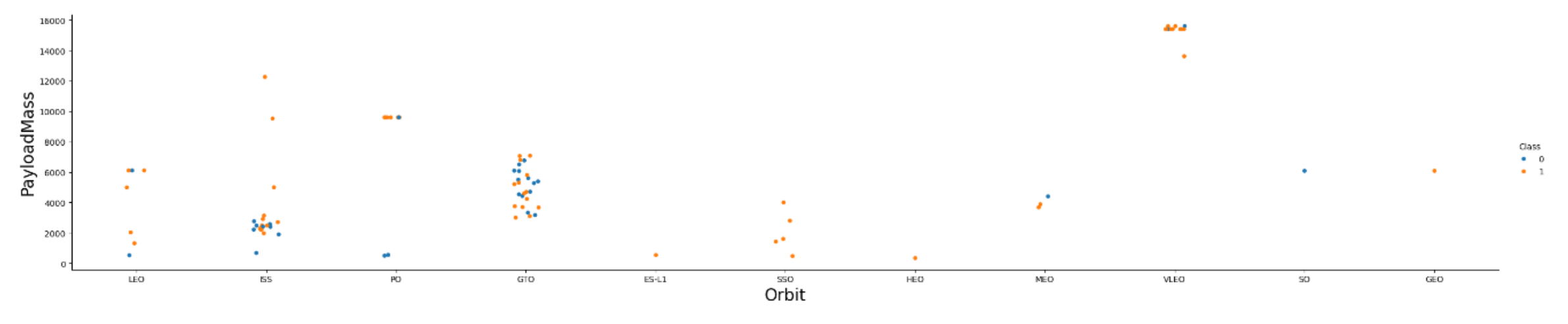
Flight Number vs. Orbit Type

- VLEO demonstrates a relationship between a higher flight number and orbit type, suggesting this orbit type was not attempted until later launches
- ISS, PO, and GTO do not show a strong relationship between flight number and orbit type, suggesting these are common orbit types



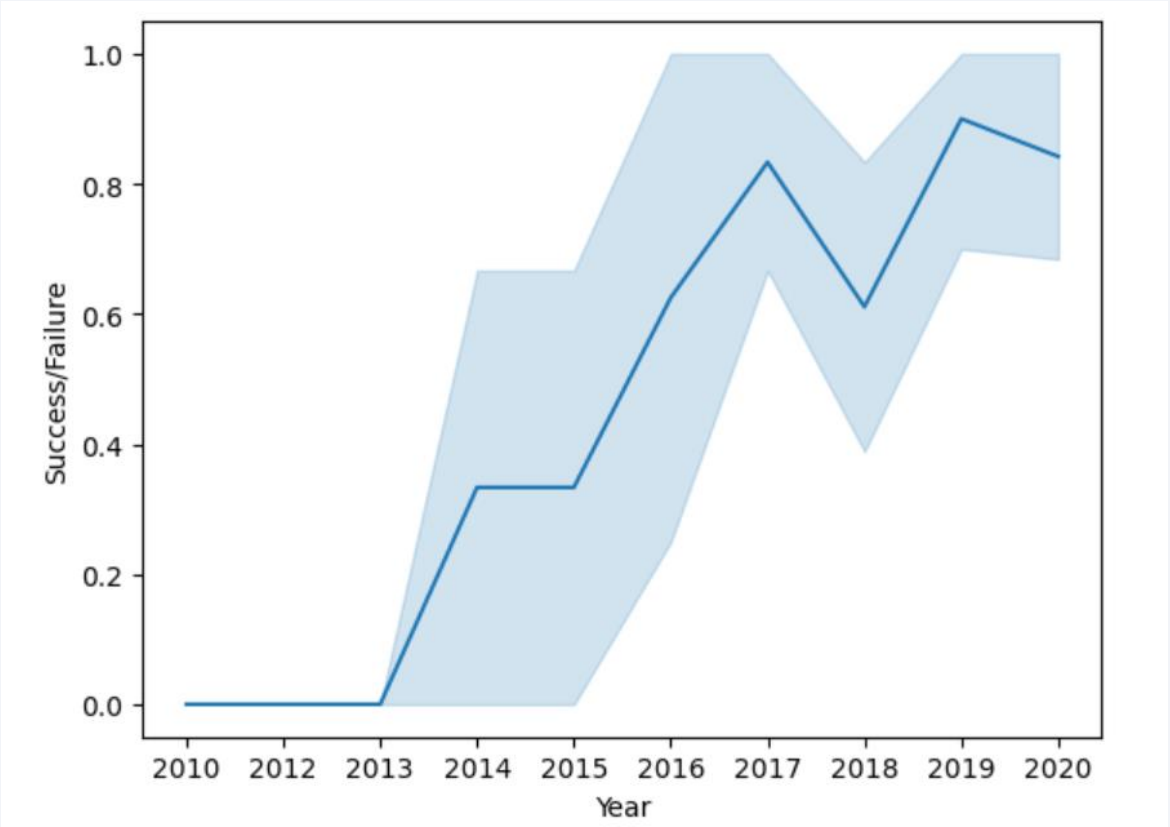
Payload vs. Orbit Type

- VLEO demonstrates a strong correlation between orbit type and high payload mass
  - GTO observations suggest payloadmasses between 3kkg and 9kkg



Launch Success Yearly Trend

- Observations indicate a strong linear relationship between years and success rates, with 2018 demonstrating a marked dip with two failures over the previous year



All Launch Site Names

- SQL (DISTINCT) was leveraged to determine a list of all unique launch sites

```
%%sql
Select
Distinct "Launch_Site"
From
SPACEXTABLE
```

```
 * 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'

- Launch site records were narrowed to the CCAFS LC-40 utilizing a limiter function

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

```
%%sql
Select *
From
SPACEXTABLE
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
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

- Total payload was calculated
- An error occurred in the SQL where statement to narrow results down to just NASA boosters

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%%sql
Select
Sum("PAYLOAD_MASS__KG_")
From
SPACEXTABLE
```

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

Sum("PAYLOAD_MASS__KG_")
619967

Average Payload Mass by F9 v1.1

- The average payload mass was calculated based on the payload mass (kg) feature, and is shown to be just below 3,000 kg

```
%%sql
Select
AVG("PAYLOAD_MASS__KG_")
From
SPACEXTABLE
Where
"Booster_Version" in ('F9 v1.1')
```

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

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

- The first date of a successful launch was found to be on June 4th, 2010

```
%%sql
Select
min("Date")
From
SPACEXTABLE
Where
"Mission_Outcome" = "Success"
```

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

min("Date")
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

- A WHERE clause was utilized to narrow the landing outcome to just successes of payloadmass totals between 4000kg and 6000kg

```
%%sql
Select
"Booster_Version"
From
SPACEXTABLE
Where
"Landing_Outcome" in ("Success (drone ship)")
and
"PAYLOAD_MASS__KG_" > 4000 and "PAYLOAD_MASS__KG_" < 6000
```

```
 * 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

- A count function and a groupby function were leveraged to calculate the total number of successes and failures based on the mission outcome variable

```
%%sql
Select
"Mission_Outcome",
Count("Mission_Outcome")
From
SPACEXTABLE
Group by "Mission_Outcome"
```

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

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

Boosters Carried Maximum Payload

- A WHERE clause utilizing a nested query was leveraged to determine what booster versions carried the max payloadmass

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

```
%%sql
Select
"Booster_Version"
From
SPACEXTABLE
Where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS__KG_") from SPACEXTABLE)
```

```
 * 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

- A WHERE clause was utilized to narrow the results to the months in which droneship failures occurred, showing to be focused on January and April

```
%%sql
Select
substr(Date, 6,2) as "Month",
"Booster_Version",
"Launch_Site"
From
SPACEXTABLE
Where
substr(Date,0,5)='2015'
and
"Landing_Outcome" = "Failure (drone ship)"
```

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

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

- Several unique query functions were leveraged, such as count, group by, and order by to observe the count of landing outcomes in a series

```
%%sql
Select
"Landing_Outcome",
Count(*) as "Count"
From
SPACEXTABLE
Where
Date between '2010-06-04' and '2017-03-20'
GROUP BY "Landing_Outcome"
Order by "Count" Desc;
```

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

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

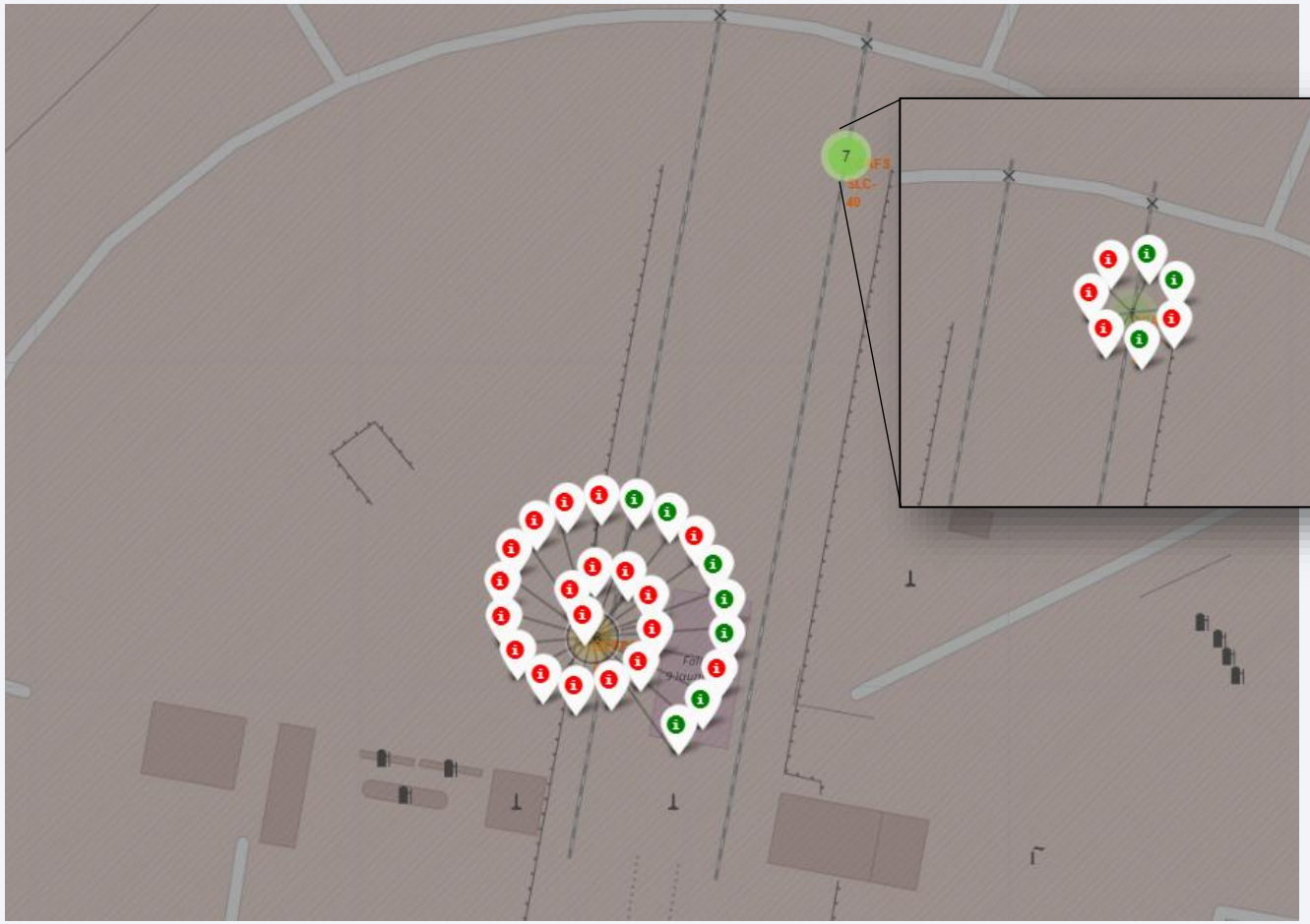
Section 3

Launch Sites Proximities Analysis

Global Launch Sites

- We can observe that all launch sites are within the United States

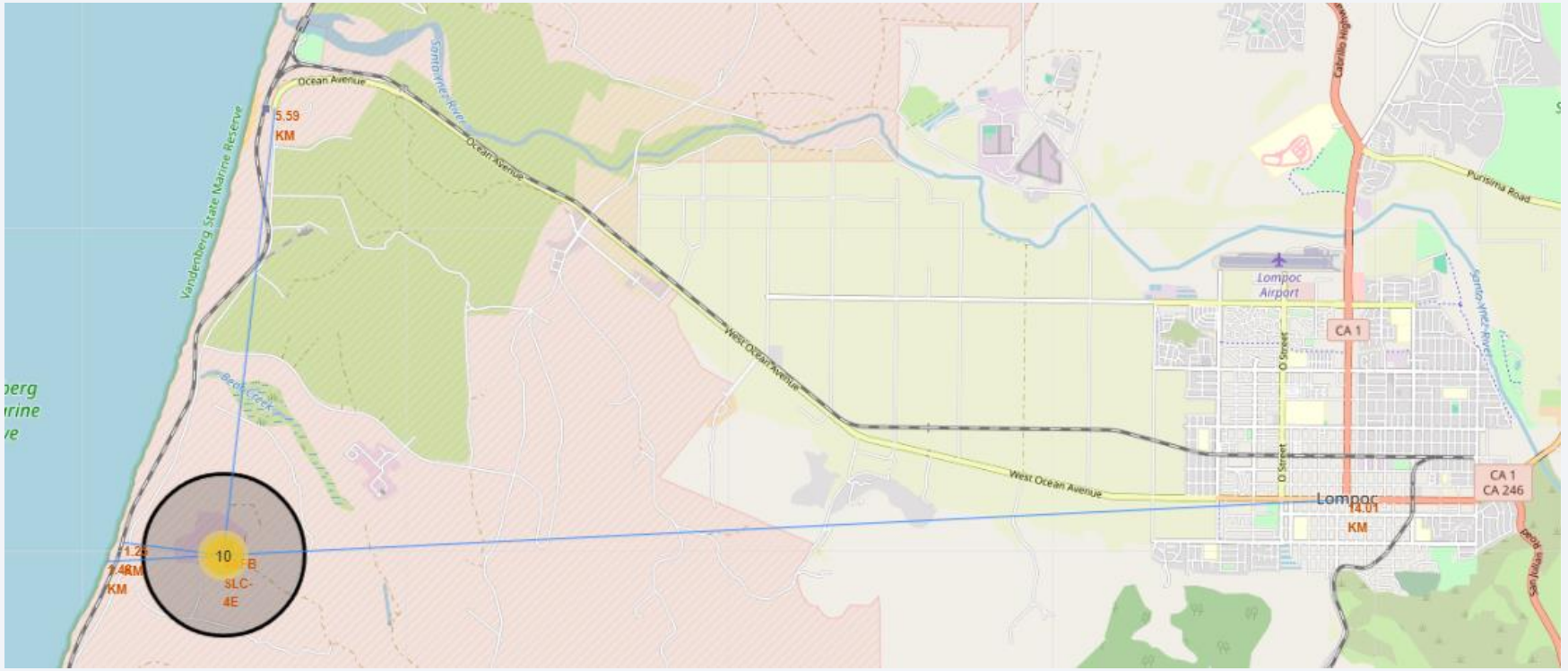
Success/Failure Marker Map Insets

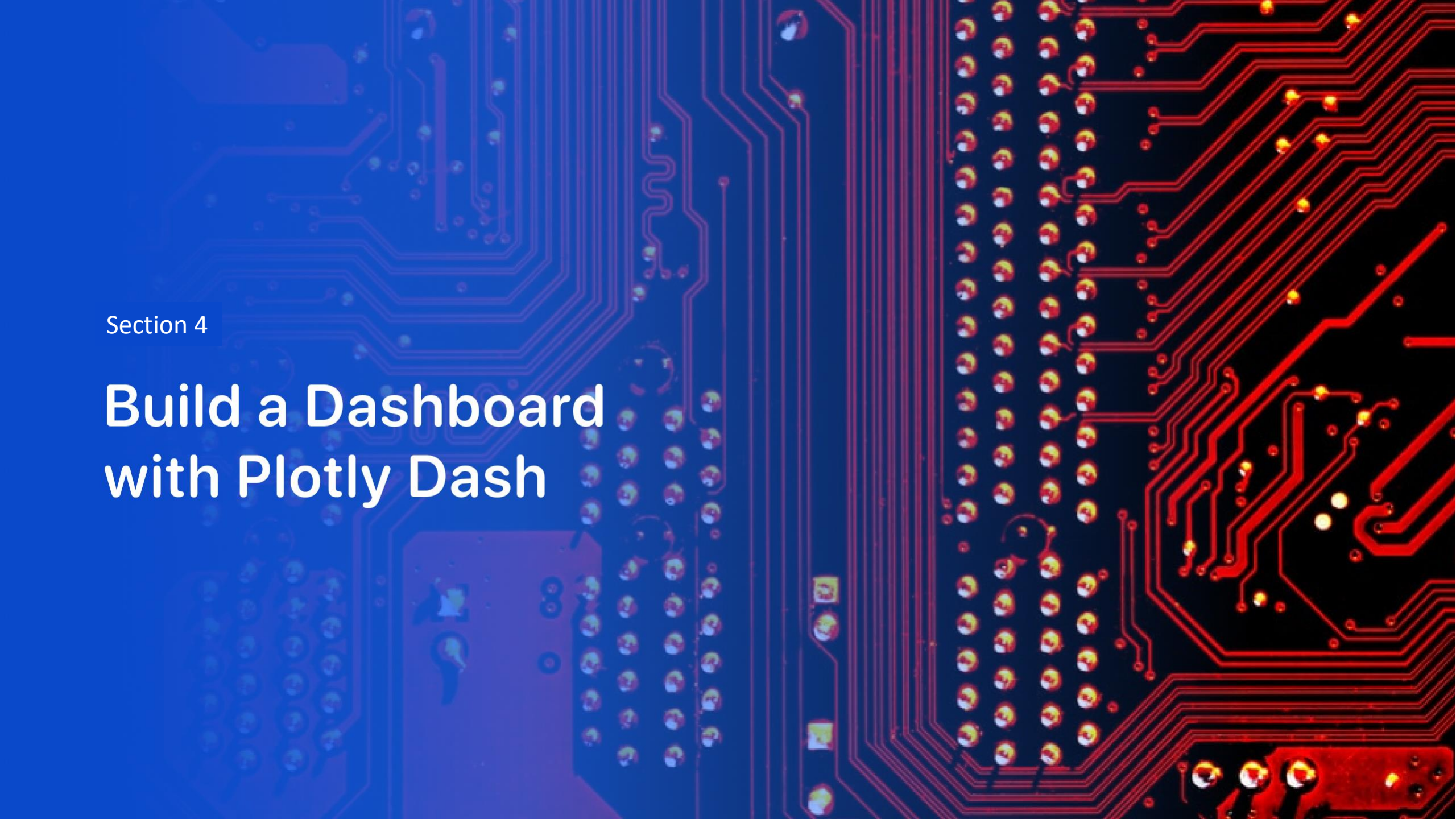


A drill down into site locations with corresponding colors to indicate success and failures at these locations

Observation of Launch Site Distance to Transporation

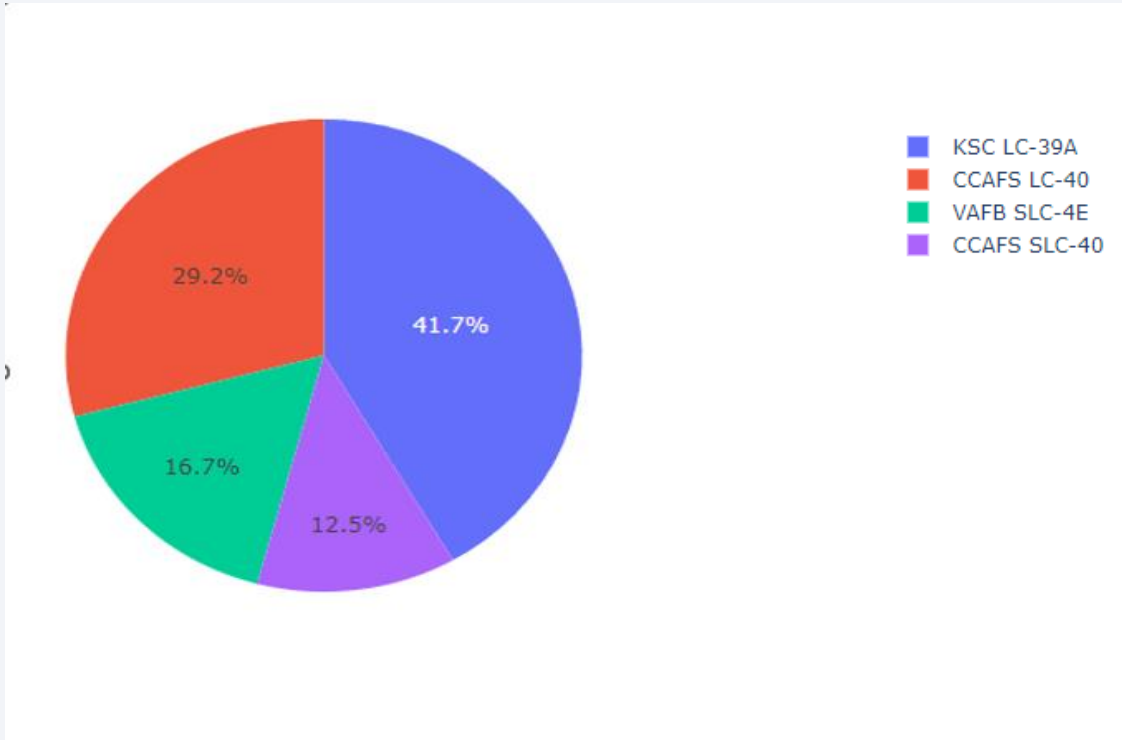
- From this observation we can see that location sites favor proximity to coastlines and railways
- City proximity is quite far comparatively



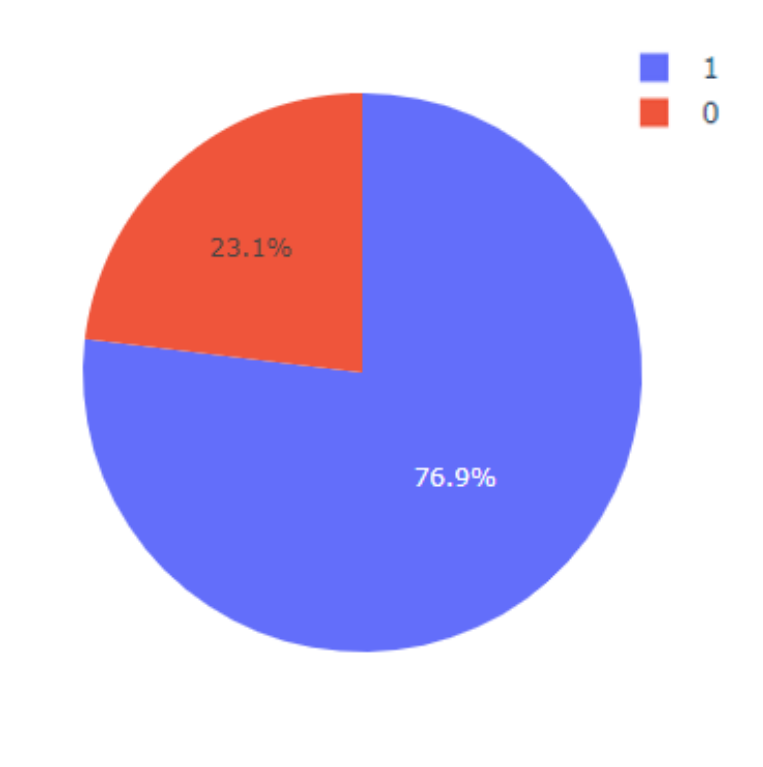

Section 4
Build a Dashboard with Plotly Dash

Success Rate by Launch Site

- We can see that KSC LC-39A is the primary launch site experiencing success
- CCAFS SLC-40 is the least successful

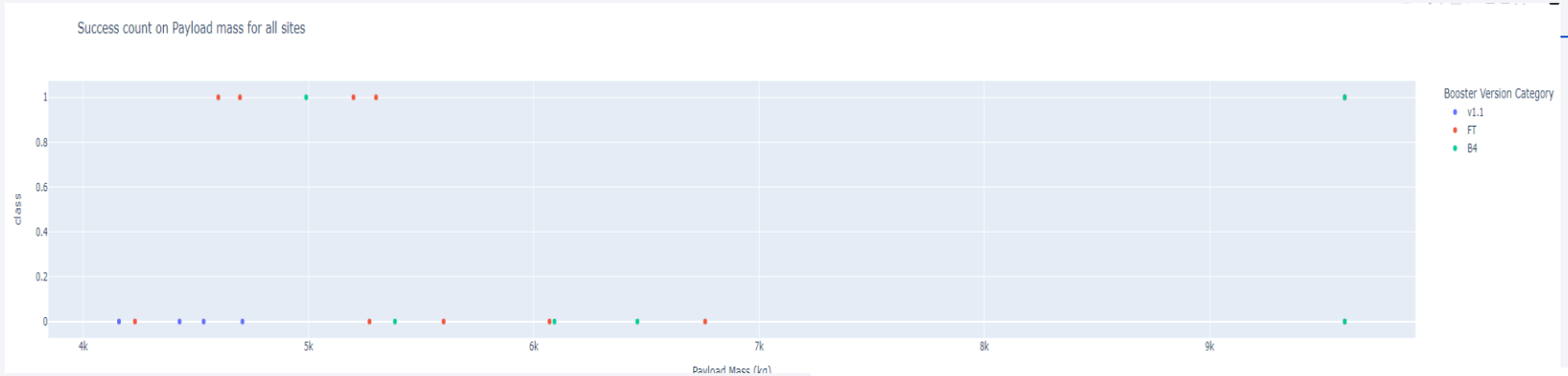


KSC LC-39A Success Statistics

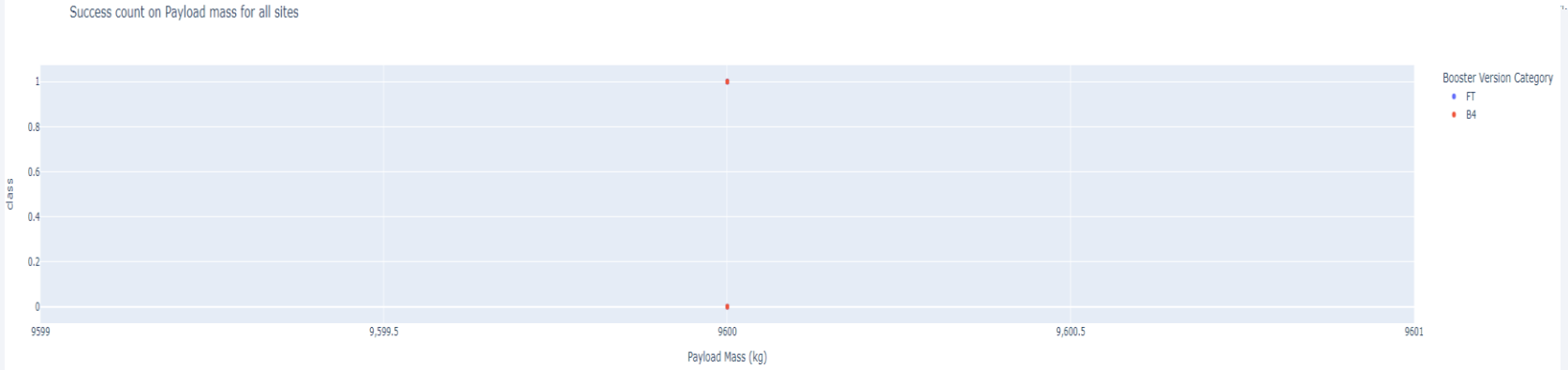


- Drilling down into the KSC LC-39A launch site, we can see that launches achieved a 76.9% success rate

Payload vs Launch Outcome



Success rates for low rated payloads is higher than heavy weighted payloads

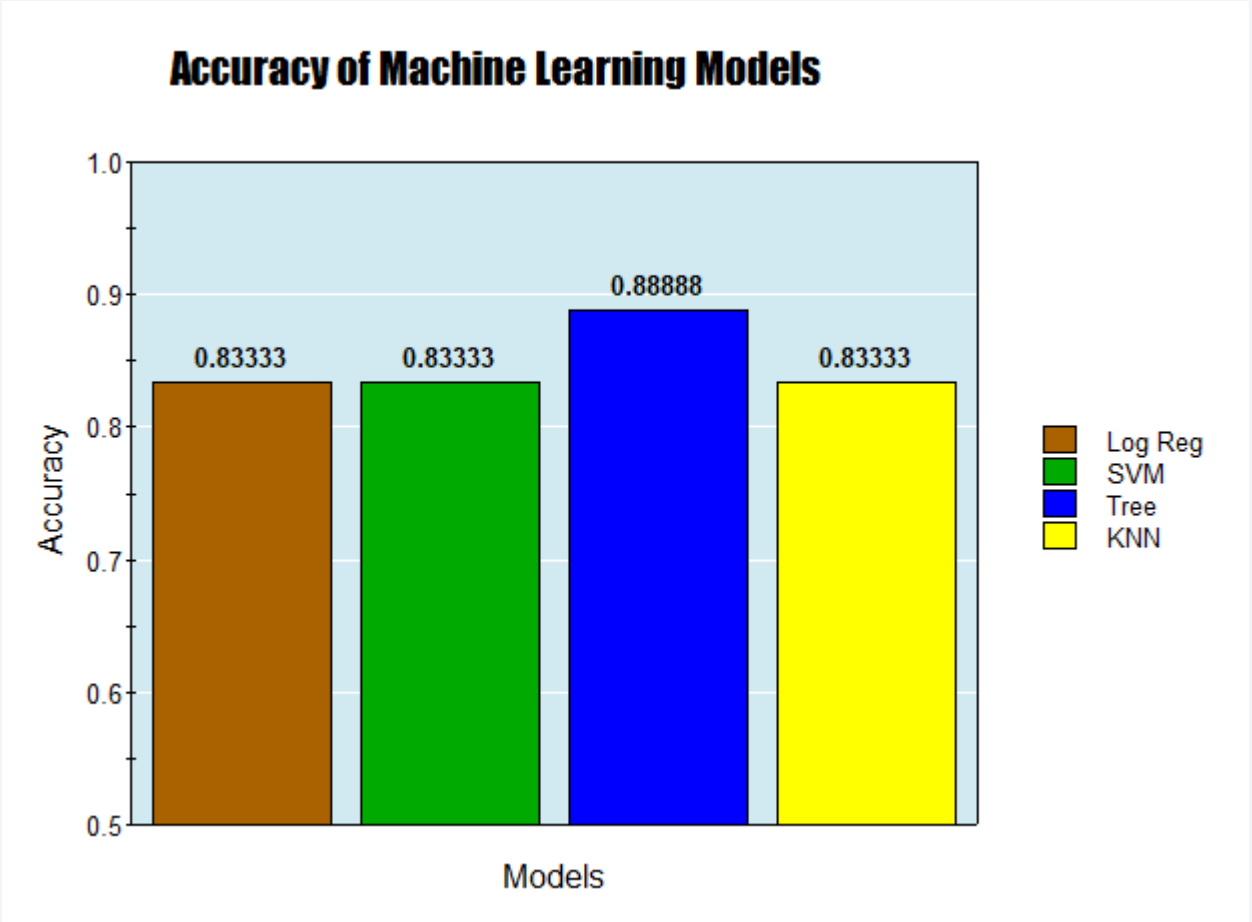


Section 5

Predictive Analysis (Classification)

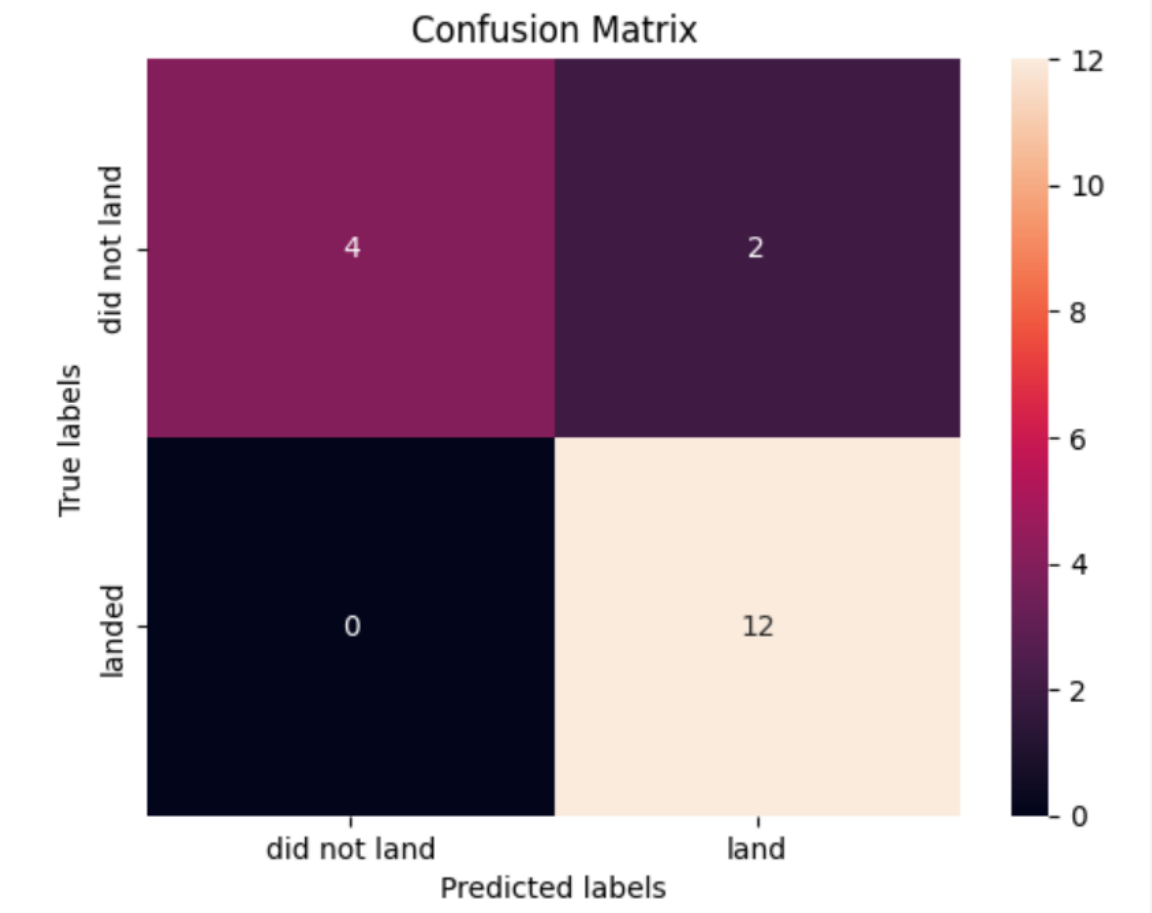
Classification Accuracy

- Model observation indicates a narrow scope of accuracy with a Decision Tree model being the most accurate by a narrow margin



Confusion Matrix

- In this confusion matrix based on accuracy results from the Decision Tree (most accurate model) we can see false positives (unsuccessful landings marked as successful landings) showed the greatest error rate



Conclusions

- Based on all observations and analysis, we can conclude that
  - The higher the flight number (most recent,) the greater likelihood of success
  - Launch success was first achieved in 2010, and has demonstrated yearly increases through 2020
  - Orbits VLEO, SSO, HEO, ES-L1, and GEO experienced the highest success rate
  - A decision tree classifier is the best machine learning model to classify future launches based on payload data and orbit type
  - KSC LC-39A is the location with the highest success rate percentage
  - An ideal launch site location should be close to coast lines and rail ways, but far from cities

Appendix

- All code and observations available at Github URL
  - https://github.com/sublimetiger/spacey
- Draw.io was utilized for flow chart creation

Thank you!