IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Raho Osman
March 12,2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusions

Executive Summary

- Summary of methodologies

1. SpaceX Data Collection using API and Web Scraping
2. SpaceX Data Wrangling
3. Exploratory Analysis using SQL, Pandas and Matplotlib
4. Exploratory Analysis via Data Visualization
5. Predictive Analysis using Machine Learning

- Summary of all results

1. EDA outcomes and results
2. Interactive Dashboards
3. Predictive Analysis Conclusions

Introduction

- As more companies are making their foray into space travel, we are taking a look into SpaceX's endeavors and how it will us at our company to learn from their launches.
- Our goal here is to find out the cost of each launch and if SpaceX will we reuse the first stage of their Falcon9 rocket. The goal of our findings is to help us compete with SpaceX

Section 1

Methodology

Methodology

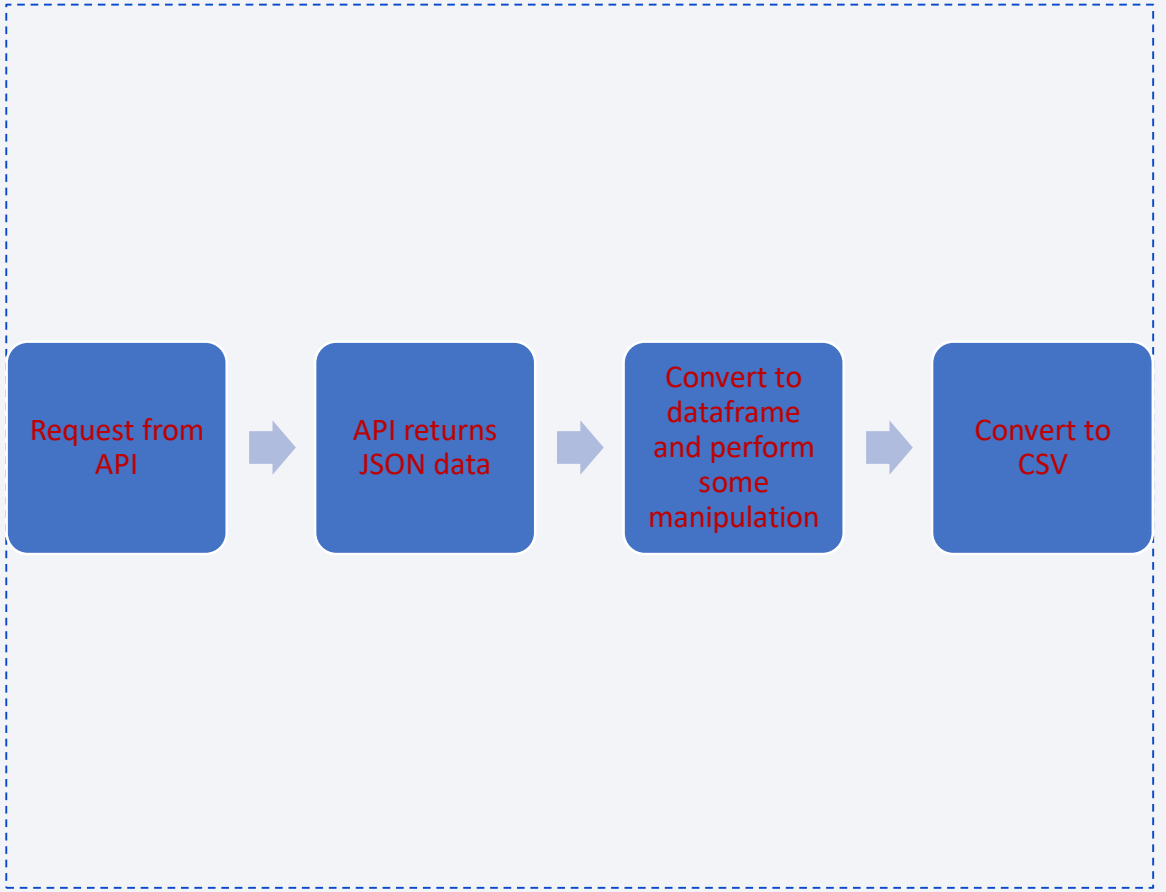
Executive Summary

- Data collection methodology:
  - Describing how data was collected
- Perform data wrangling
  - Describing how data was processed
- 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 – SpaceX API

To predict whether or not the first stage of the Falcon 9 will land successfully or fail, we gathered data from a few sources.

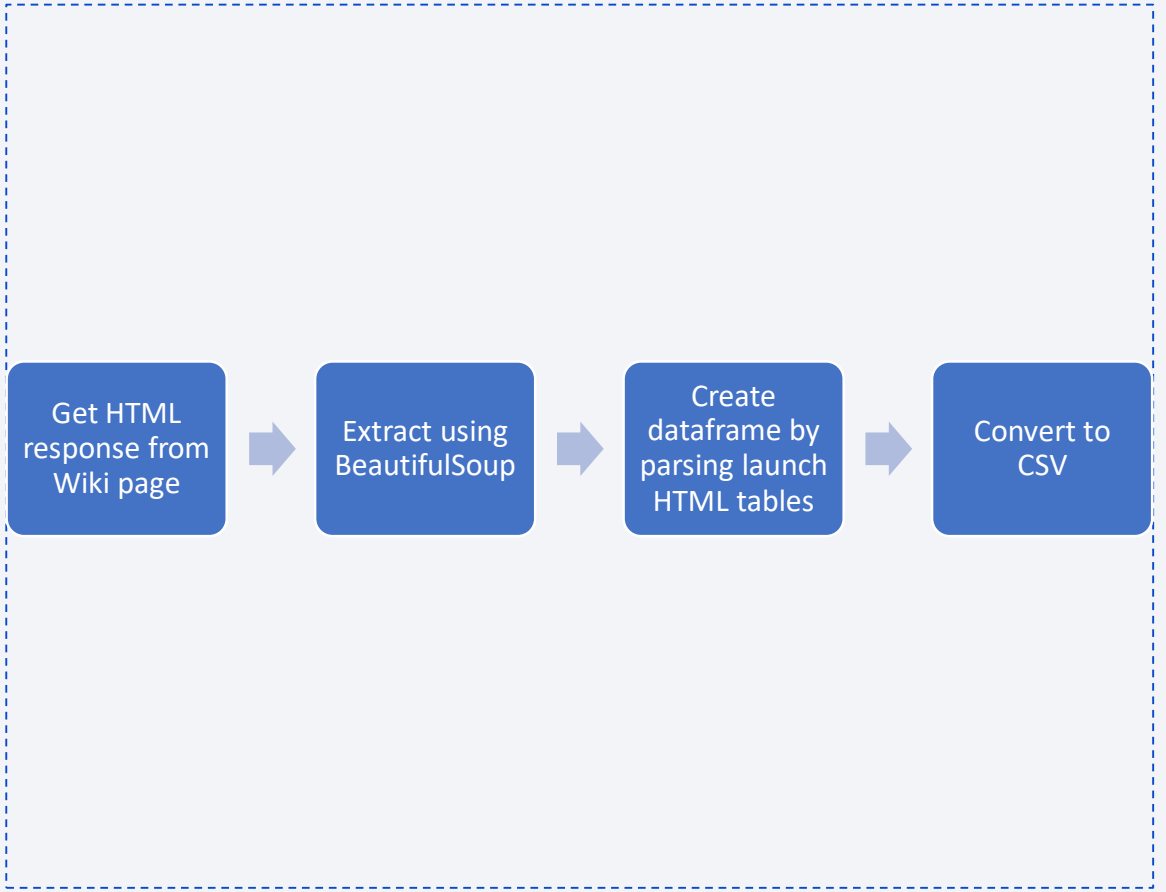
- Launch Data was collected via the SpaceX REST API
  - This provided us with historical data of launches, rocket use, payload, landing outcomes etc.
  - This was the URL used https://api.spacexdata.com/v4/launches/past
- https://github.com/raho93/SpaceXCapstoneproject2024/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

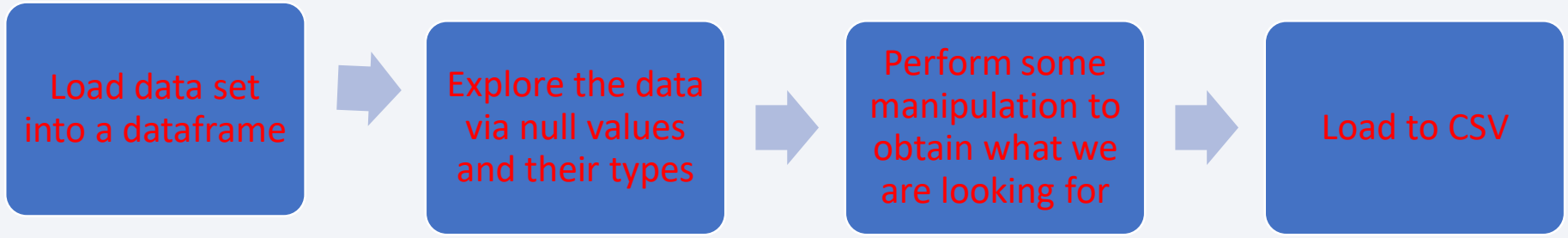
- More data was scrapped from the Wikipedia page of Falcon9's past launches. Then Beautiful soup was used to extract the data and put it into a dataframe.

- https://github.com/raho93/SpaceXCapstoneproject2024/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- After the data collection process, the dataframe that was formed was then used. Other rocket data was filtered out so that we were just left with Falcon9 data. Among some of the data that was calculated was number of launch sites, number and occurrences of orbits. Landing outcomes were also categorized into 0 and 1 to help later with predictive analysis.



- https://github.com/raho93/SpaceXCapstoneproject2024/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter plots were used to visualize certain relationships between variables. Matplotlib and pandas were used to get those visualizations
- The scatter plots included the relationship between Flight number and Launch Site, Flight number and Orbit type as well as Payload and Orbit type
- Another visualization that had been done is a bar chart of each orbit and its success
- A line plot was done to visualize the launch yearly success trend
- https://github.com/raho93/SpaceXCapstoneproject2024/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- The following queries were used:

To display unique launch sites

```
%sql SELECT DISTINCT (LAUNCH_SITE) FROM SPACEXTBL;
```

To display total payload mass carried by boosters launched by NASA

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

To display average payload mass carried by booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

To display the date of the first successful landing outcome in ground pad

```
%sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

- https://github.com/raho93/SpaceXCapstoneproject2024/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Created folium map to marked all the launch sites. Also map objects were created such as markers, circles, lines to mark the success or failure of launches for each launch site.
- A launch set outcomes (failure=0 or success=1) were also created.
- These objects were created to explore whether there is a relationship between distance of nearby objects and the success rate
- https://github.com/raho93/SpaceXCapstoneproject2024/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

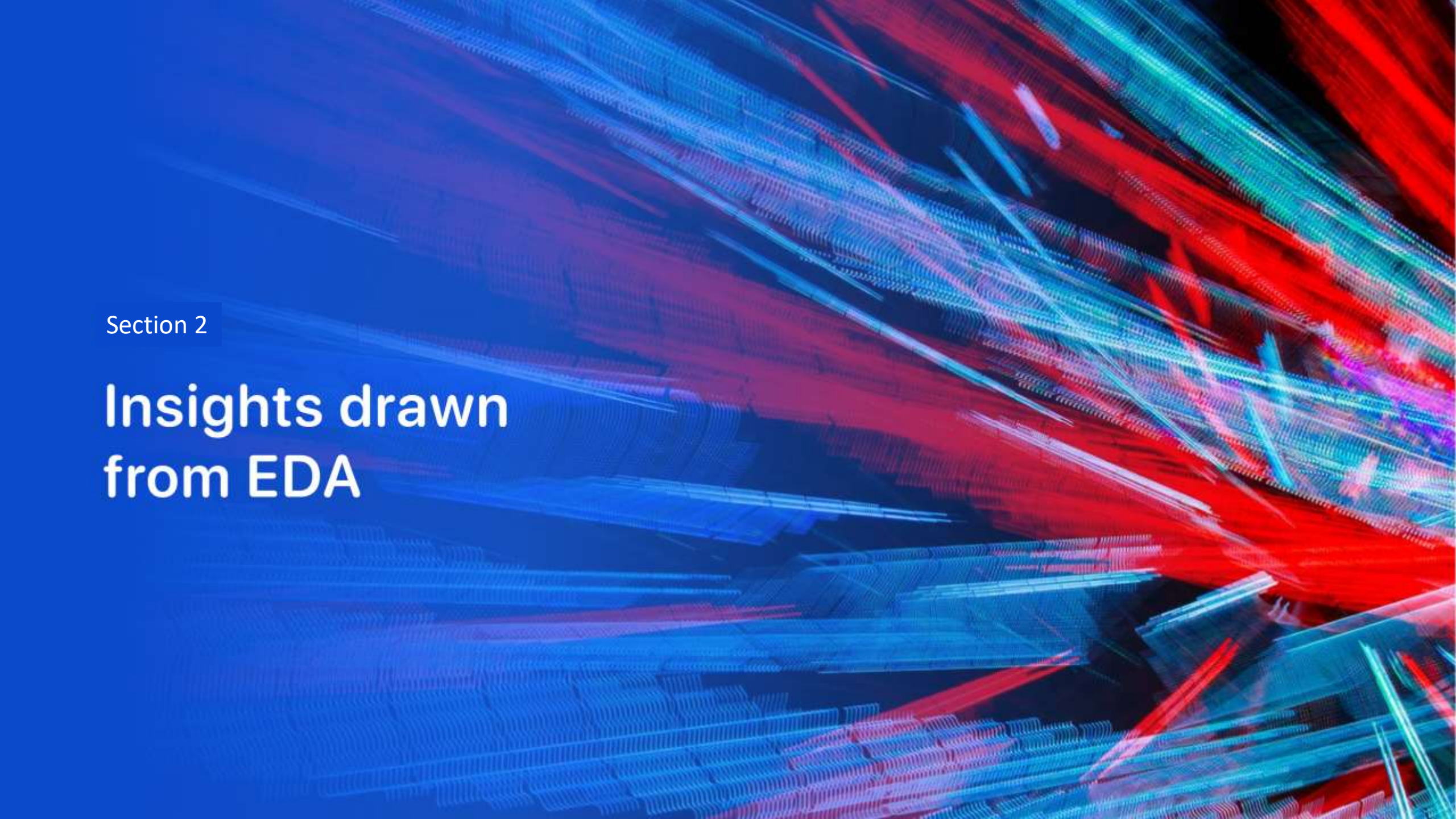
- In the dashboard, total success by all sites was added as well as Total success launches per site. Another addition to the dashboard is a scatter plot of Launch rate vs Payload
- These plots and interactions are very important in determining the success of launches and how it will be used for future predictions.
- https://github.com/raho93/SpaceXCapstoneproject2024/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Summary:
  - Load CSV into a dataframe
  - Define what is X and Y
  - Preprocess and standardize the data in X
  - Split into training and test sets
  - Use the algorithms: Logistic Regression, SVM, Decision Trees and KNN
  - Train each model using GridSearchCV and find the best parameters
- https://github.com/raho93/SpaceXCapstoneproject2024/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

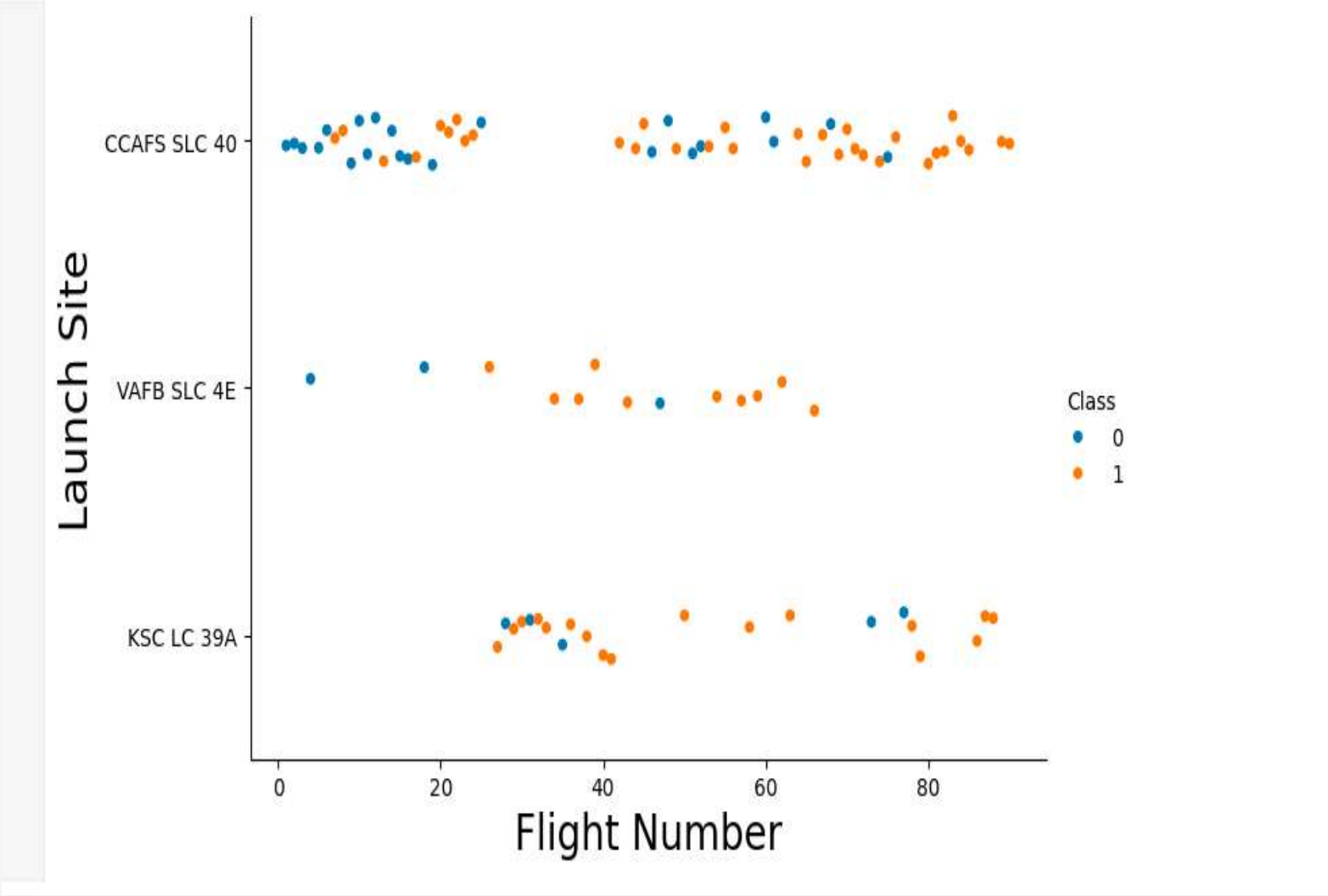
Results

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


Section 2
Insights drawn
from EDA

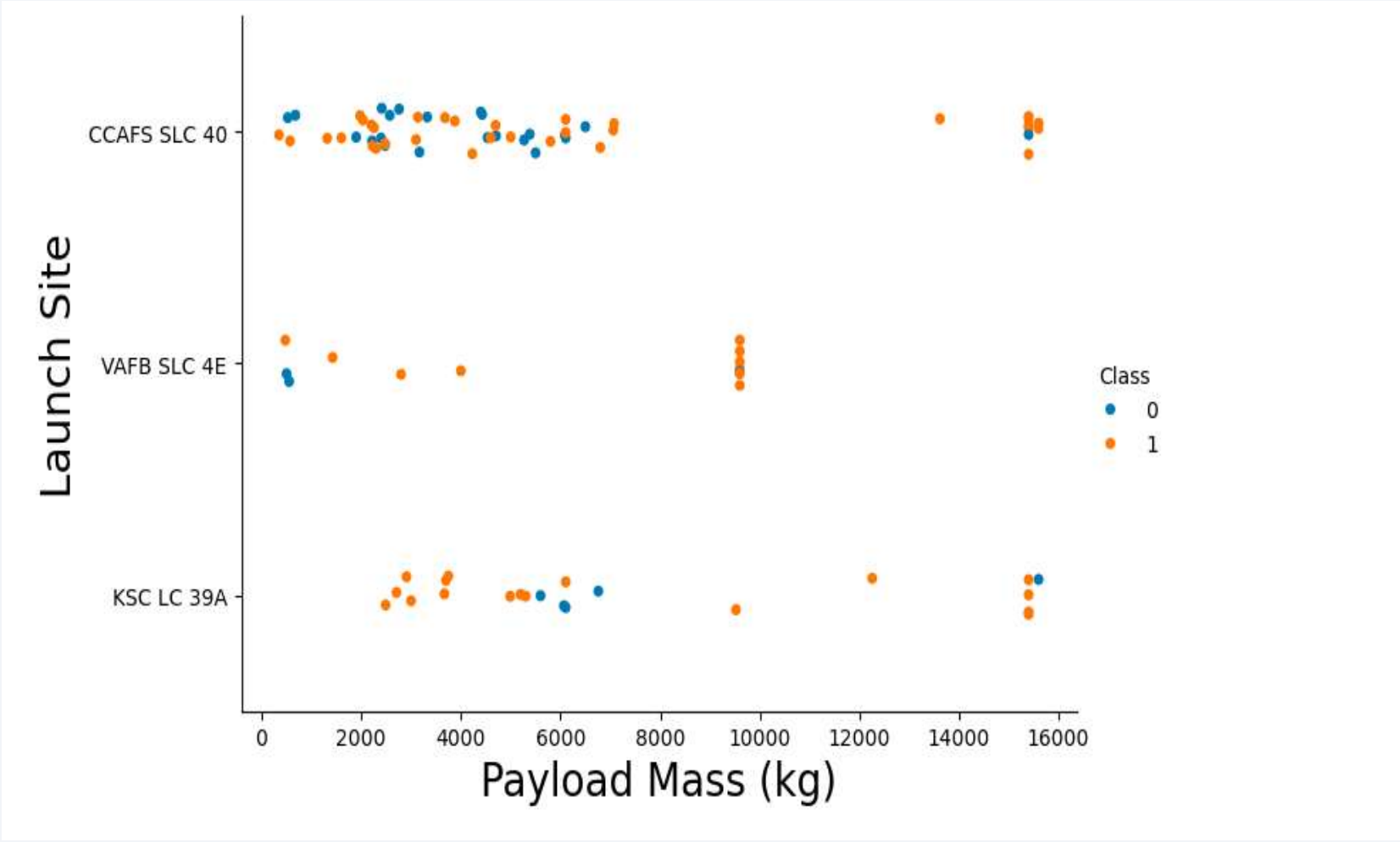
Flight Number vs. Launch Site

- Looking at the scatter plot we can deduce that as the flight number increases so does the success rate. This is applicable to all 3 launch sites



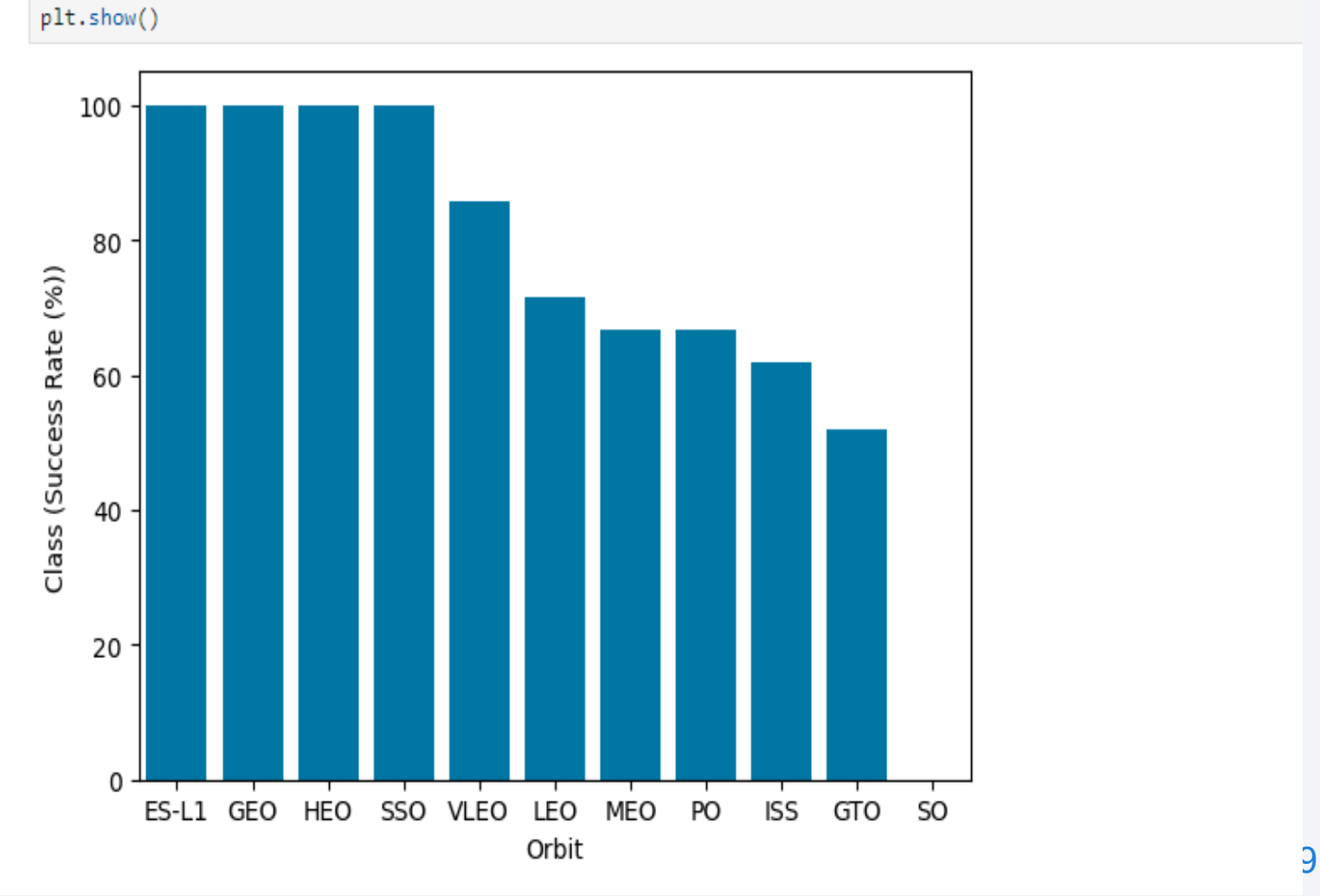
Payload vs. Launch Site

- For the VAFB SLC 4E launch site, a payload no rockets are launched that are more than 10000 kg. For the other two launch sites as the payload mass increases there are more successes than there are failures



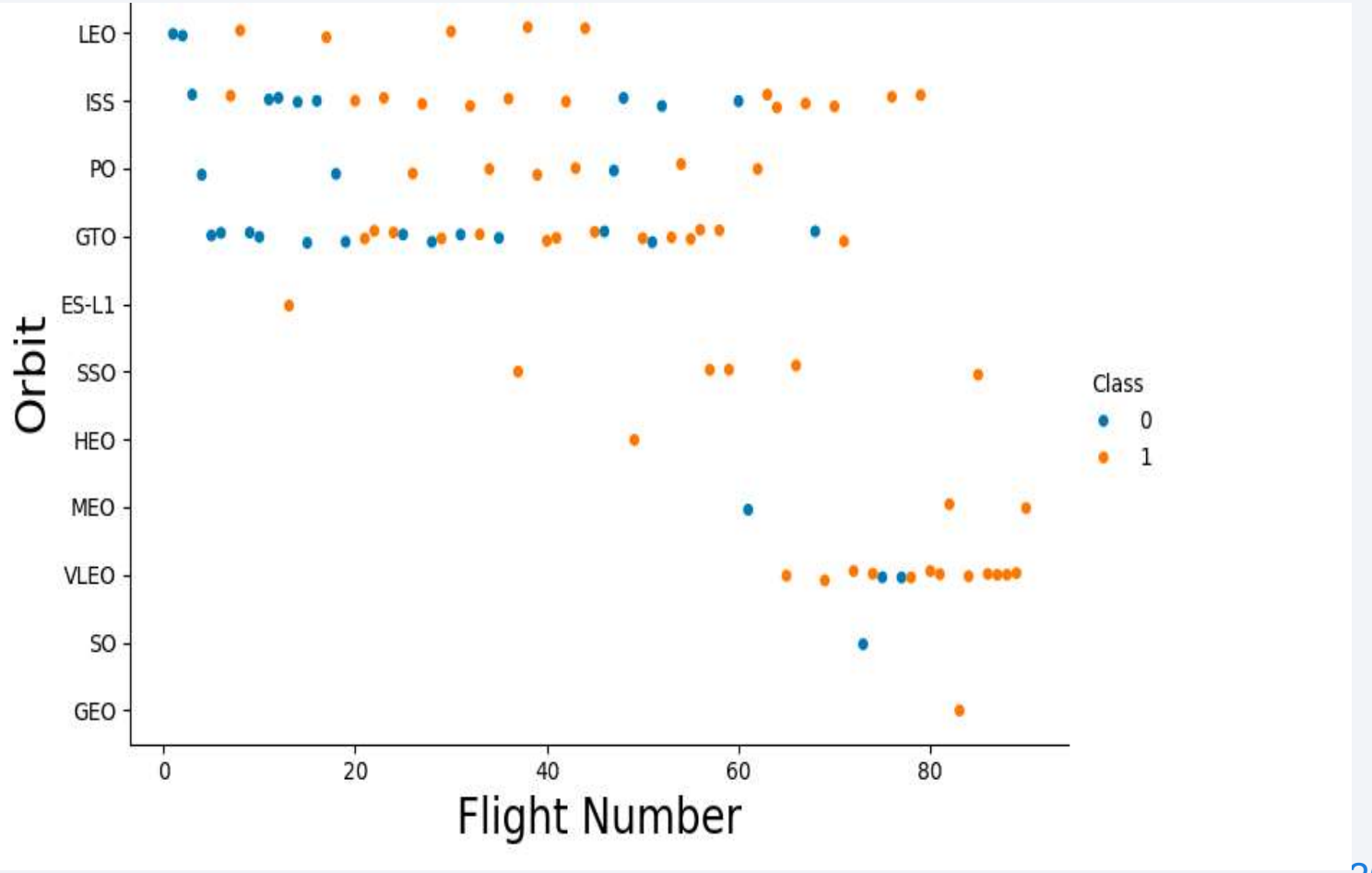
Success Rate vs. Orbit Type

Orbit SO has 0% success while Orbits ES-L1, GEO,HEO,SSO have success rates of 100%



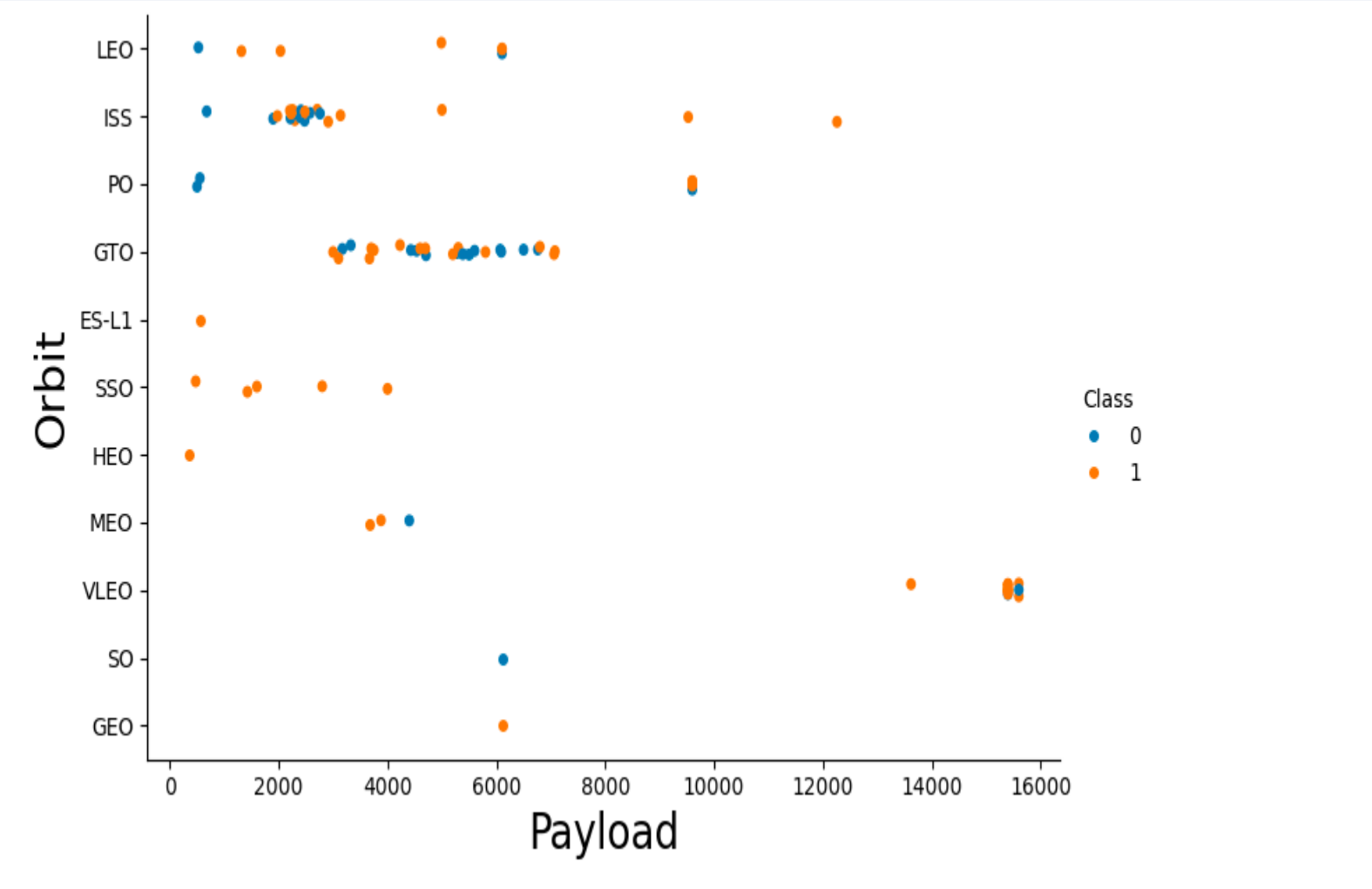
Flight Number vs. Orbit Type

There seems to be no relationship between flight when GTO is in orbit. However for orbit LEO the more flights the more success.



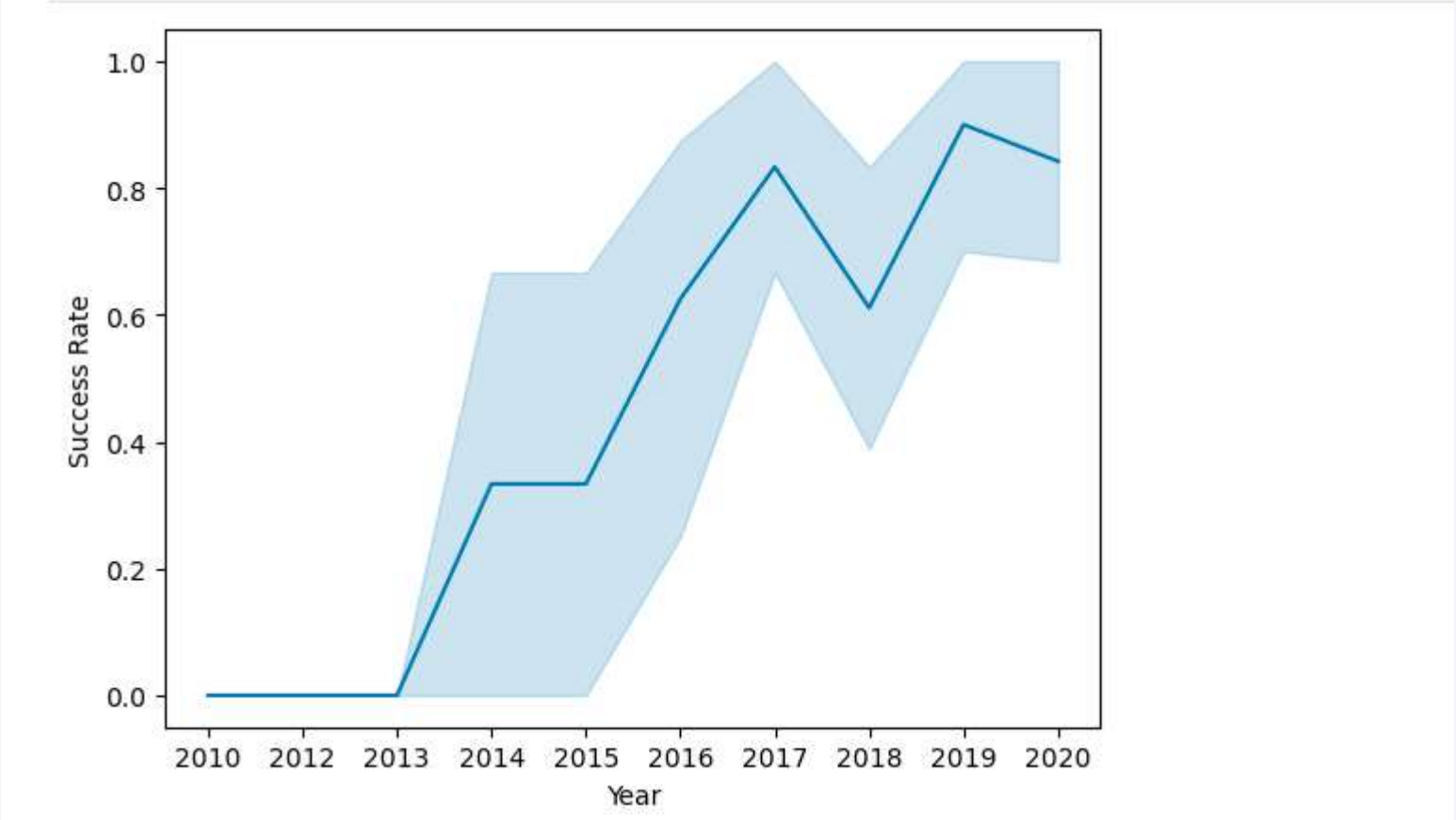
Payload vs. Orbit Type

For orbits PO,LEO and ISS there seems to be more success with higher payloads. With orbit GTO it is difficult to discern a relationship between payload and orbit in terms of success or failure.



Launch Success Yearly Trend

Between 2013 and 2020 there seems to be an increase in success rate.



All Launch Site Names

```
%sql SELECT DISTINCT (LAUNCH_SITE) FROM SPACEXTBL;
```

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

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

Used Select Distinct to obtain the all launch sites

Launch Site Names Begin with 'CCA'

- Used like ‘CCA%’ to get the launch sites starting with CCA and also used the limit 5 to only get the top 5 rows of the table with launch sites starting with 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
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

- Used the sum function to get the total of payload mass column for boosters launched by NASA

```
1]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
1]: sum(PAYLOAD_MASS__KG_)
                      45596
```

Average Payload Mass by F9 v1.1

Used AVG function to obtain the average payload mass carried by booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Used the MIN function on the date column and used WHERE to specify that a successful landing is being looked for.

```
%sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

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

min(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Used WHERE as well as AND to filter out the rows where the drone landing was successful and the payload was 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 Booster_Version from SPACEXTBL WHERE Landing_Outcome = '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

- Used GROUP BY and COUNT to display the total number of successful and failed mission outcomes

```
%sql select Mission_Outcome,  count(Mission_Outcome) from SPACEXTBL 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

- Used Subquery to first get the max payload then in the main query selected the Booster that had that maximum payload mass

```
%sql select 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

- Used Substr to get the month in 2015 where the landing outcome as a failure

```
%sql SELECT substr(Date,0,5) as year, substr(Date, 6, 2) as month ,Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTBL
WHERE Landing_Outcome LIKE 'Failure%drone%' AND SUBSTR(Date,0,5) = '2015'
```

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

year	month	Booster_Version	Launch_Site	Landing_Outcome
2015	01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015	04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Used WHERE and GROUP BY to filter the landing between those dates. Then used ORDER BY to order by descnding

```
%sql SELECT Landing_Outcome, COUNT(*) AS CountOfLaunches FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome ORDER BY CountOfLaunches DESC;
```

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

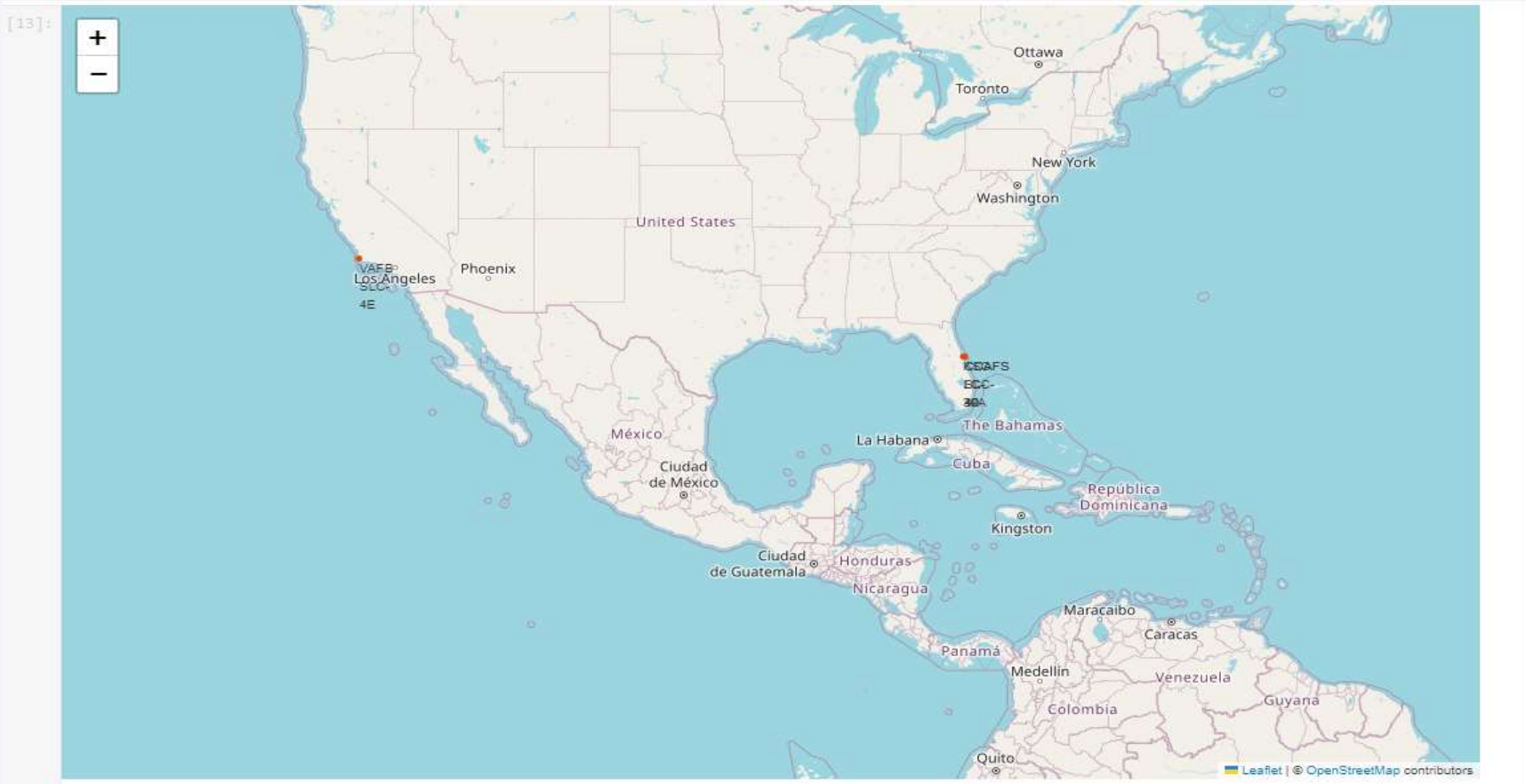
Landing_Outcome	CountOfLaunches
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Launch Sites on a global map

- Observing the location of the launch sites, the conclusion is that they are more towards the south of the US and close to the equator. Another observation is they are in proximity to coastlines



Florida Launch sites with markers

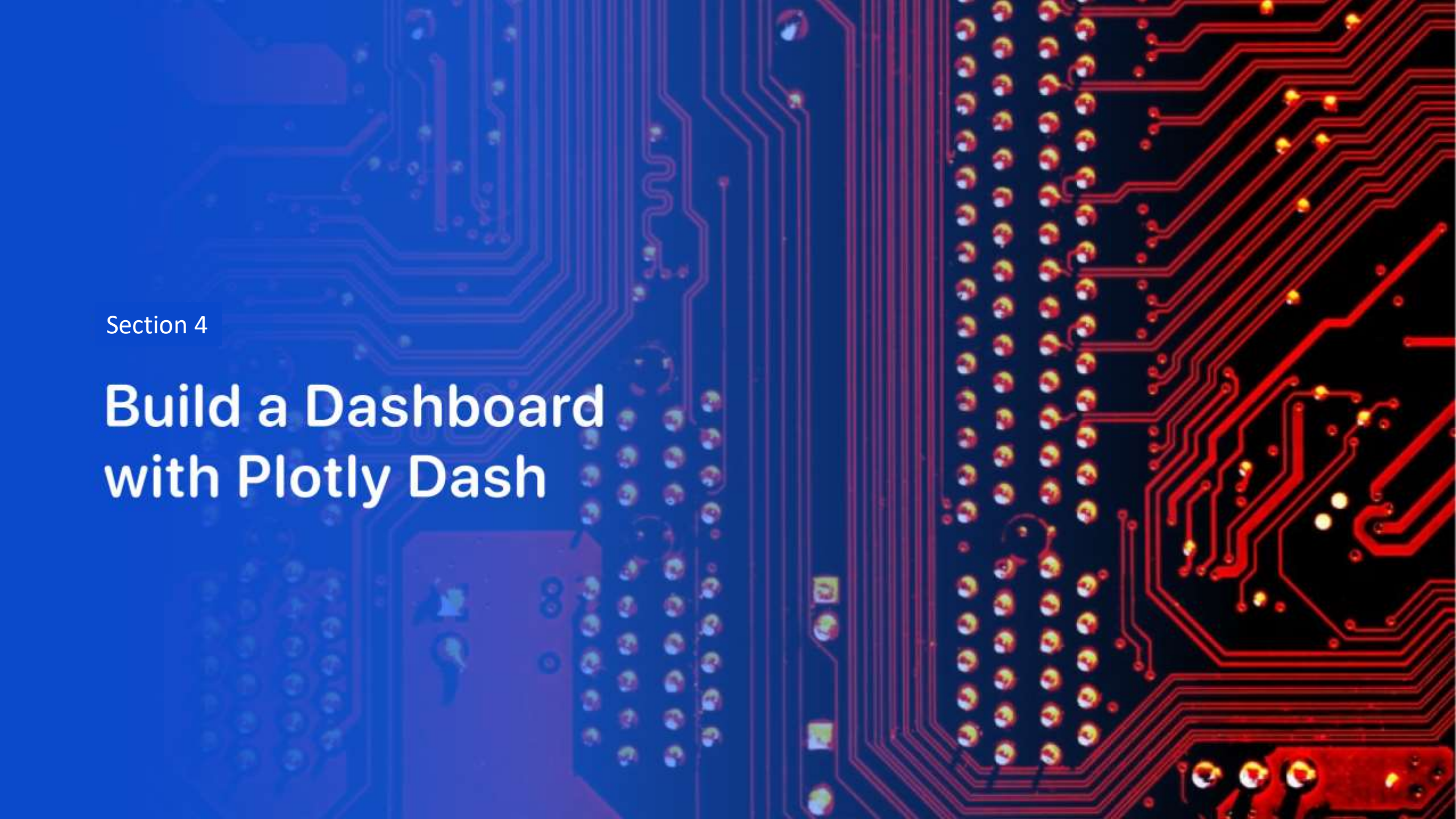
- For the Florida launch sites, KSC LC-40 has the highest success compared to the other two sites

Launch site distances from key points

- As shown in the top picture the two launch sites are 0.86 km from the coast.
- In the bottom picture, it is observed that CCAFS SLC-40 is 23.19 km away from the nearest highway

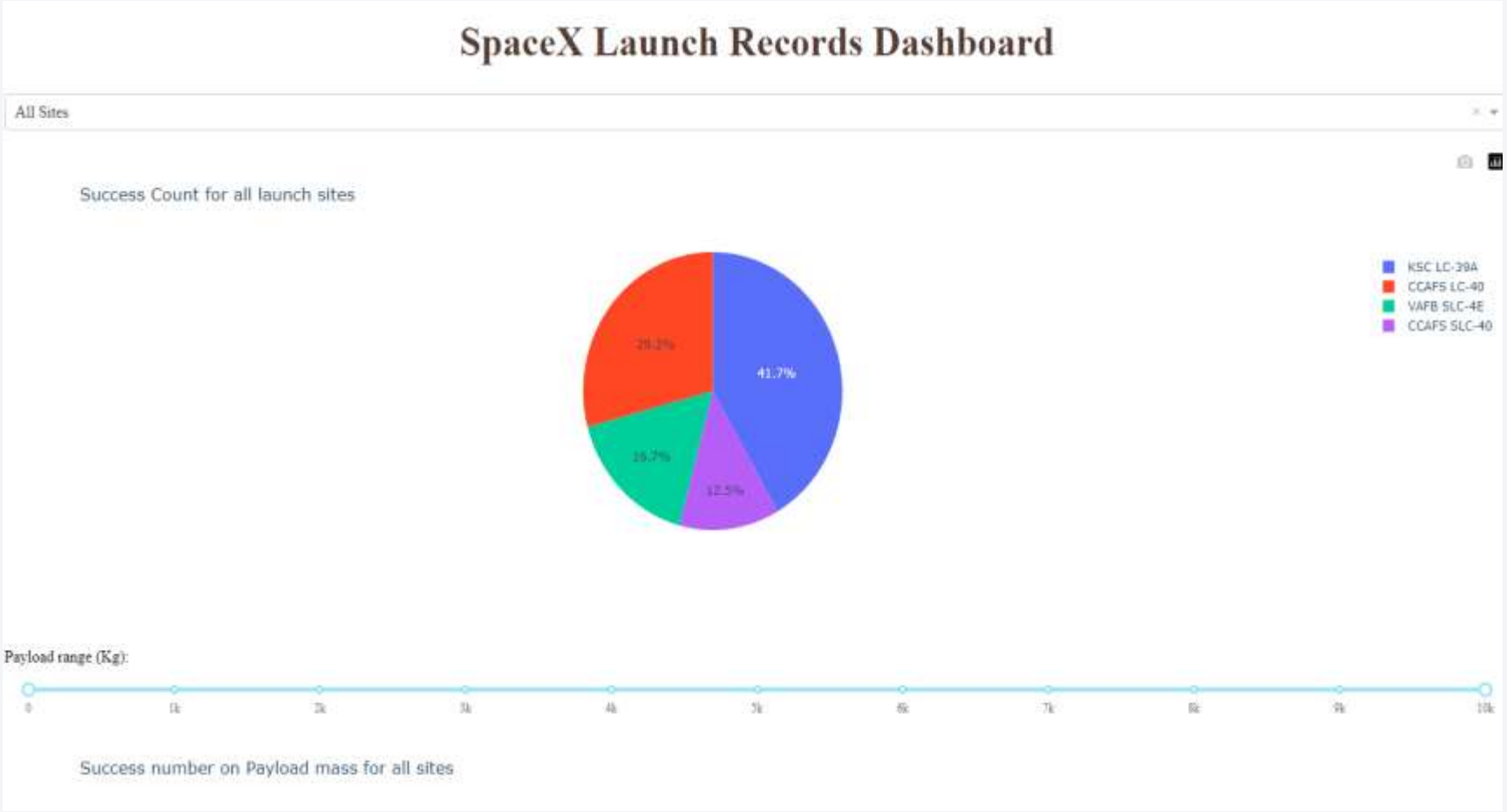
Section 4

Build a Dashboard with Plotly Dash



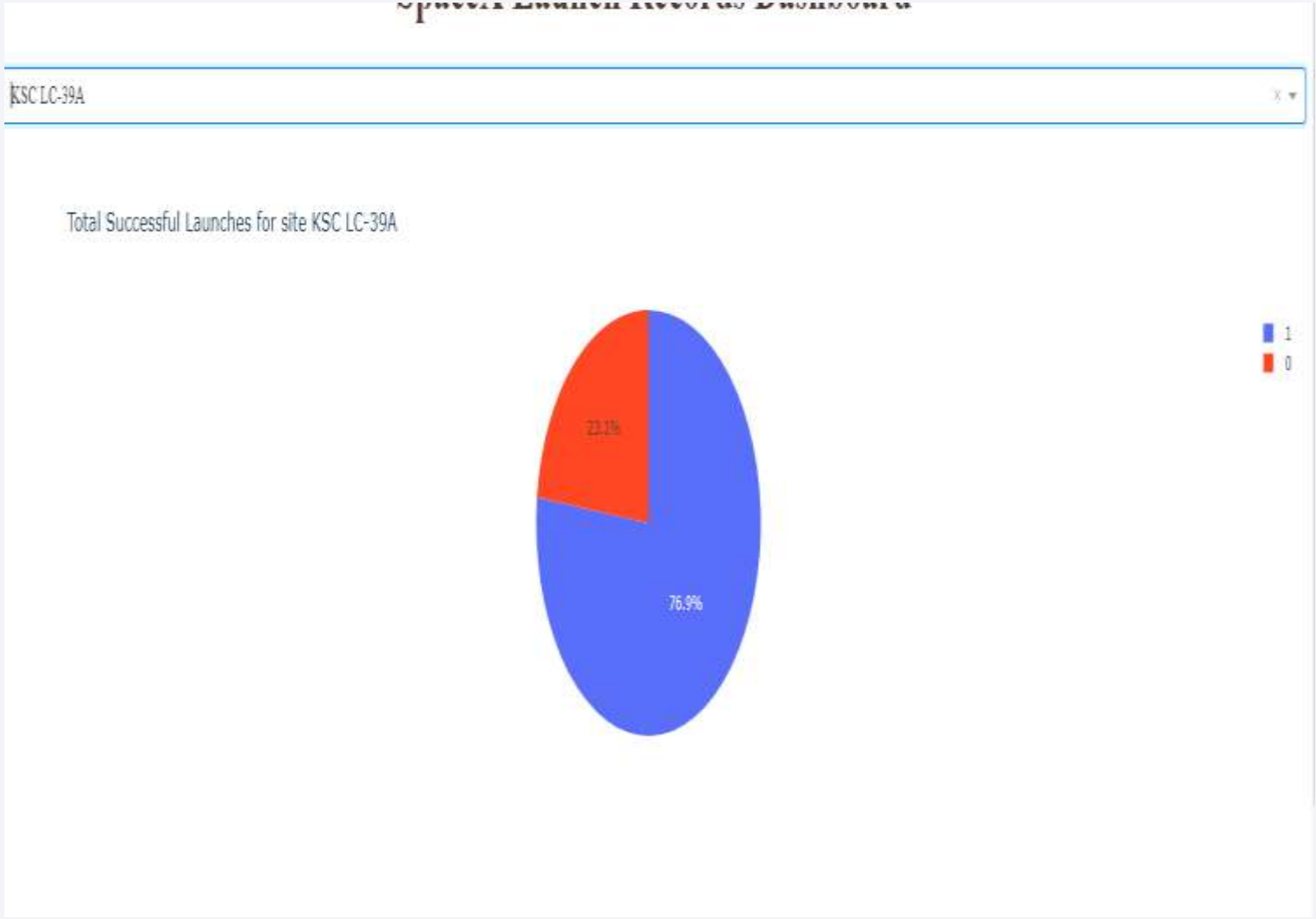
Success Count for all launch sites

Launch site KSCLC-39A has the highest success rate at 41.7% and CCAFS SLC-40 is the least successful at 12.5%



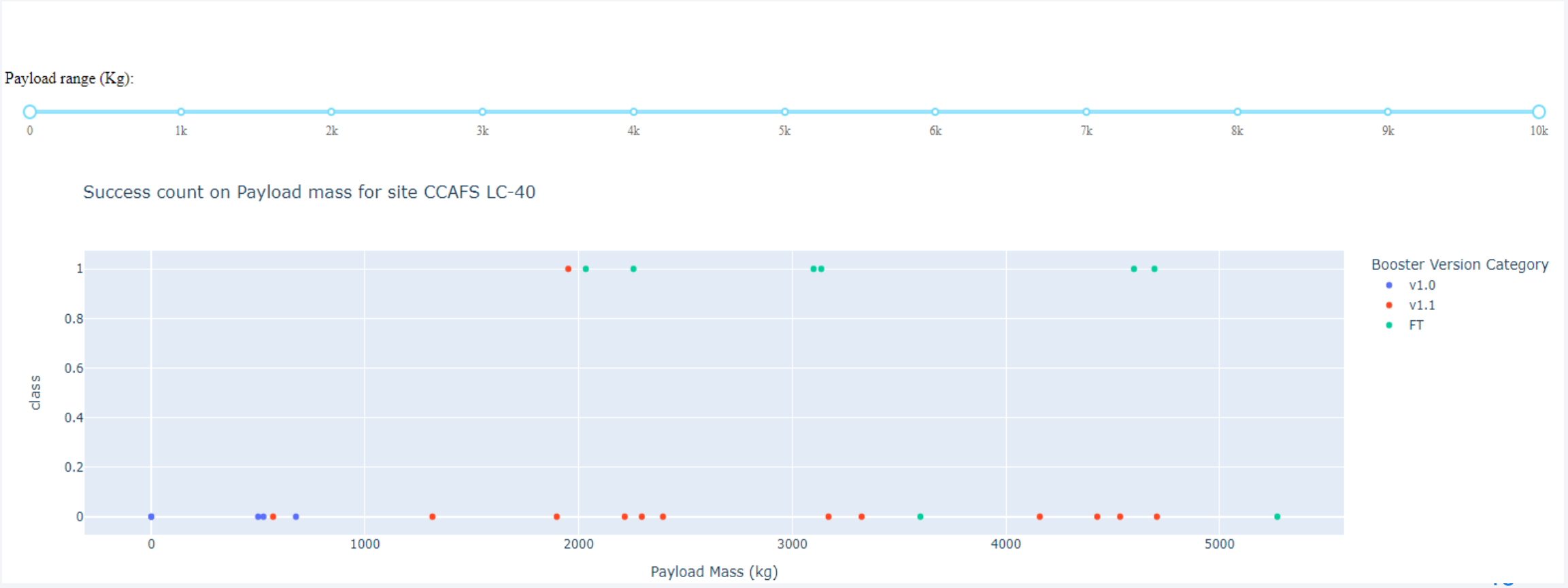
Total successful launches for KSC LC-39A

- Site KSC LC-39A has a success rate of 76.9% for all its launches and a failure rate of 23.1%



Success count on payload mass for site CCAFS LC-40

- Booster FT has the most success for Payload masses greater than 2000 kg



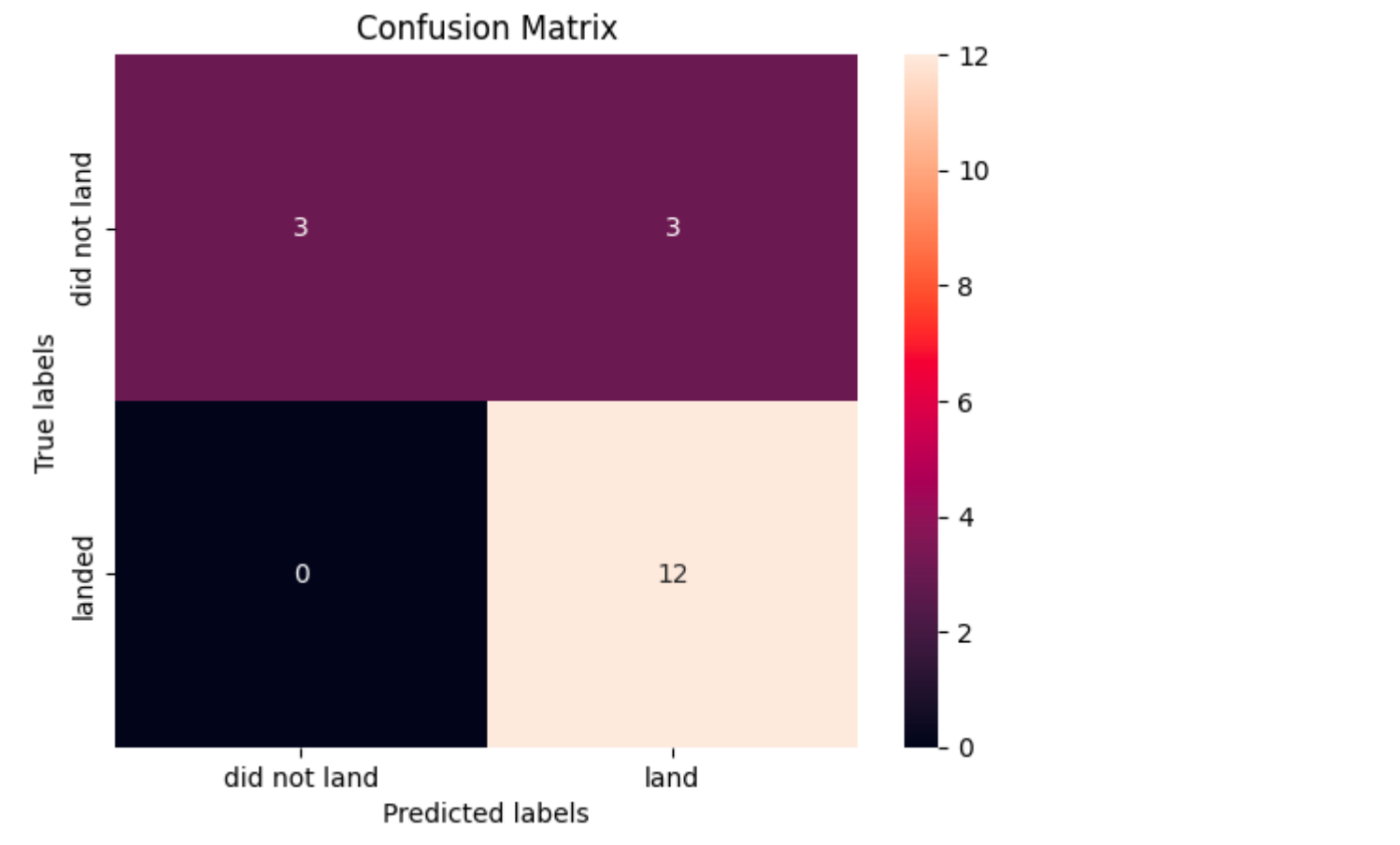
Section 5

Predictive Analysis (Classification)

Confusion Matrix

- All 4 models have the same confusion matrix and they all have the same accuracy

	0
Method	Test Data Accuracy
Logistic_Reg	0.833333
SVM	0.833333
Decision Tree	0.833333
KNN	0.833333



Conclusions

- Each launch site has a varied success rate. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 76.9%.
- Another observation is that as flight number increased, the success of the launch increased as well.
- For all launch sites except VAFB SLC 4E as payload mass increased the success rate also increased
- All orbits have a decent success rate except orbit SO that has a success rate of 0%
- The launch sites seem to all be close to the coast.
- All algorithmic models used have the same accuracy of 0.83333……

Thank you!