IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Joanne Ho
March 16, 2024

Outline

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

Executive Summary

- Using the historical SpaceX launch data to conduct data analysis, find correlations, and build prediction models using Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbor.

- In summary, all 4 prediction models generate similar prediction with the accuracy rate of 0.833 on test data. All have ~ 17% false positive (predict successful landing, but it was not) which is a concern.

- Further model tuning is recommended. Considering to get more data and collect other external data attributes (such as weather esp wind speed and direction) to find additional correlations and build a more accurate model.

Other findings include:

- Site KSC LC-39A has the highest success ratio with 76.9% success rate
- Site VAFB SLC 4E did not launch any heavy rocket with payload Mass more than 10K, but its success rate is very high
- Orbits ES-L1, GEO, HEO, and SSO have the highest success rate

Introduction

Background

SpaceX 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 SpaceX can reuse the first stage.

Problems statement

Accurate prediction of successful landing the Falcon 9 first stage can determine the cost of rocket launch while maximizing the saving on launching SpaceX rockets in the long run.

Section 1

Methodology

Methodology

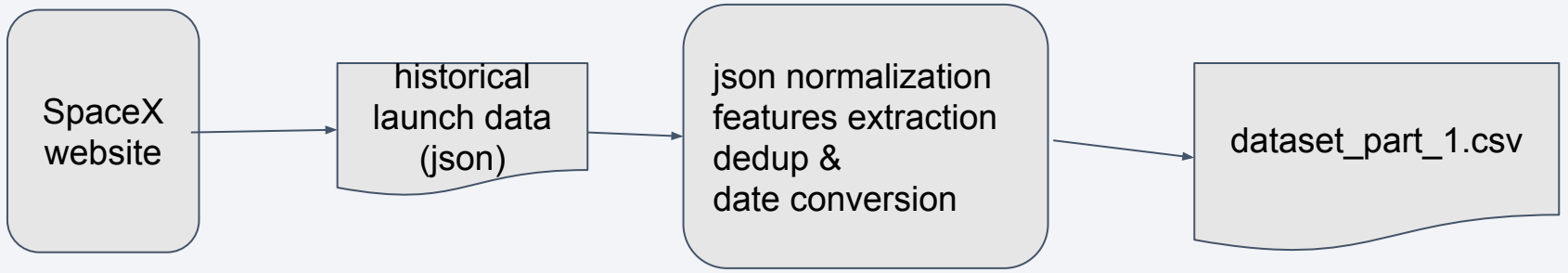
Executive Summary

- Data collection methodology
  - The historical SpaceX launch data was pulled from SpaceX API website by using requests.get(). The data is in json format.
- Perform data wrangling
  - Fixing null values and changing different data types.
- 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

The historical SpaceX launch data was pulled from SpaceX API website. The data is in json format.

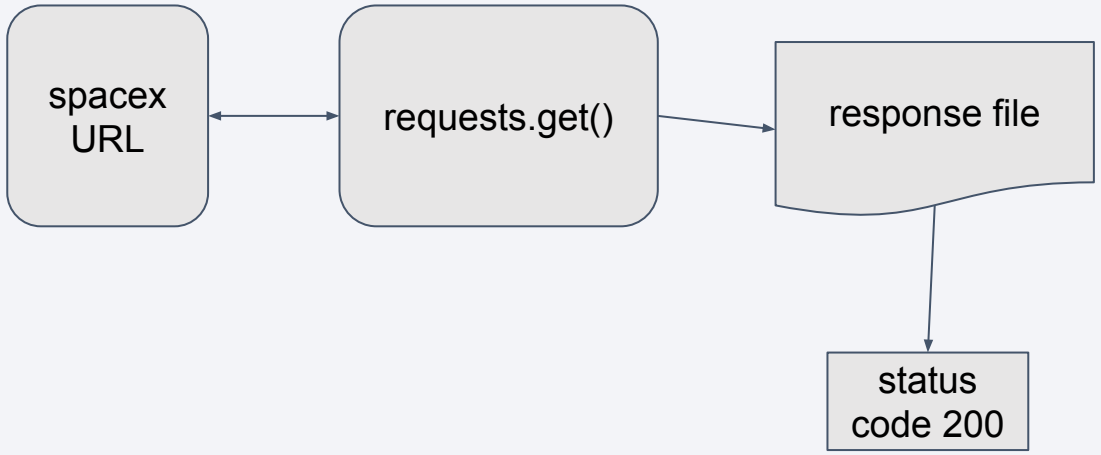
Here’s the key phrases and flowcharts of Data Collection



Data Collection – SpaceX API

- Here's the key phrases and flowcharts of pulling the data using REST API call and get command from SpaceX URL

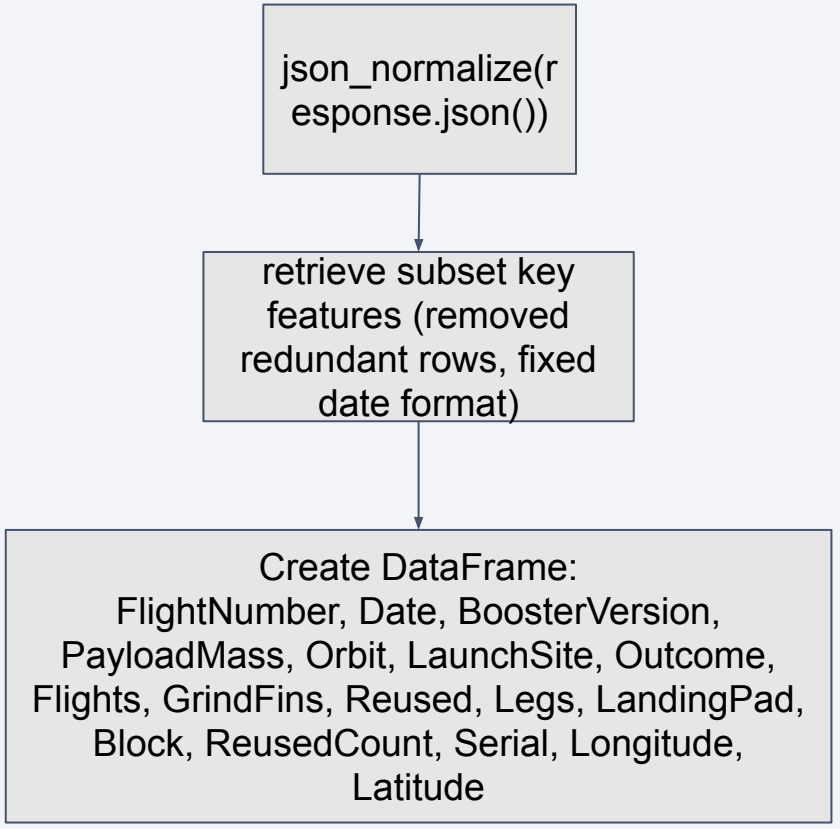
- GitHub URL of SpaceX API calls



Data Collection - Scraping

Here's the key phrases and flowcharts of Web Scraping to pull and prepare SpaceX launch data in HTML format

- GitHub URL of SpaceX web scraping

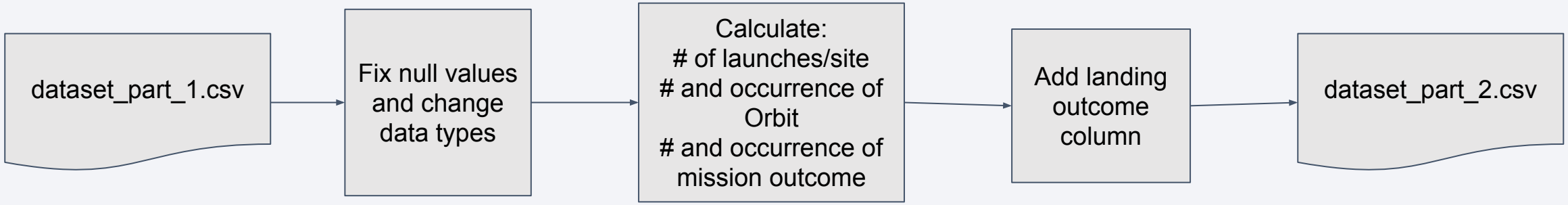


Data Wrangling

In data wrangling phase:

- Fixed null values and changing different data types
- Calculated the number of launches per launch site
- Calculated the number and occurrence of each Orbit
- Calculated the number and occurrence of mission outcome of the Orbits
- Created a landing outcome based on the Outcome column value

Here's the key phases and flow chart of Data Wrangling



GitHub URL of SpaceX data wrangling

EDA with Data Visualization

We used **catplot** as visualization to analyze the followings:

- Flight Number and Payload Mass - We observed that different launch sites have different success rates. `CCAFS LC-40`, has a success rate of 60 %, while `KSC LC-39A` and `VAFB SLC 4E` has a success rate of 77%.
- Launch Site and Flight Number - We noticed `CCAFS LC-40` launched the most rockets, but it also had the most failure launches

We used **scatterplot** as visualization to analyze the followings:

- Payload Mass and Launch Site. We observed VAFB SLC 4E did not launch any heavy rocket with payload Mass more than 10K, but the success rate was very high
- Flight Number and Orbit. We noticed LEO and SSO had the most success rate
- Payload Mass and Orbit. With heavy payloads the successful landing or positive landing rate were more for Polar,LEO and ISS. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) were both there and here.

We used **bar chart** to analyze Orbit and Success Rate. We noticed orbits ES-L1, GEO, HEO and SSo orbits had the highest success rates.

Finally, we used **line chart** to analyze Year and Success Rate. Success Rate increase in the recent years. Our assume is due to lessons learned over the course of time, continuous on gaining experience and knowledge, and mature technology

GitHub URL of EDA with data visualization

EDA with SQL

We loaded the data into SQLite database to get the following data:

- Get Launch Site names
- Get 5 records where launch sites begin with the string ‘CCA’
- Get total payload mass of NASA (CRS)
- Get average payload mass carried by booster version F9 V1.1
- Get the date of the first successful landing outcome in ground pad was achieved
- Get the list of boosters that have success drone ship with payload mass between 4000 and 6000
- Get the total of success and failure mission outcome
- Get the list of booster version that carried the maximum payload mass
- Get the month when failure occured landing on drone ship
- Rank the count of landing outcomes occurred between 6/4/2010 and 3/20/2017

GitHub URL of EDA with SQL

Build an Interactive Map with Folium

We use Folium to build the interactive map to analyze the launch site and its distance at close proximity:

- Mark the success and failure launches on all Launch Sites using MarkerCluster()
- Add marker of the closest coastline point and draw Polyline to measure the distance between coastline point and the launch site. For comparison, add markers of nearly railroad, highway and city, and draw Polyline to measure the distance from the launch site

Putting the launch site on the map can get insights on location (close to coast and equator, etc). Setting markers to analyze if the launch site is at close proximity to railways, city, airport and coastline, which will impact the community (noise, dust, fire, etc)

GitHub URL of interactive map with Folium map

Build a Dashboard with Plotly Dash

Build an interactive application to display Dashboard based on user's selection of data points:

- Add a drop down menu to select all launch sites or a specific launch site to show the success launches rate in pie chart and scatter plot
- Add the range slider of Payload mass to see the success rate result in the pie chart and and scatter plot

Using pie chart and scatter plot can better illustrate the result. The interactive dashboard can immediately show the result based on user's selection

GitHub URL of your Plotly Dash lab

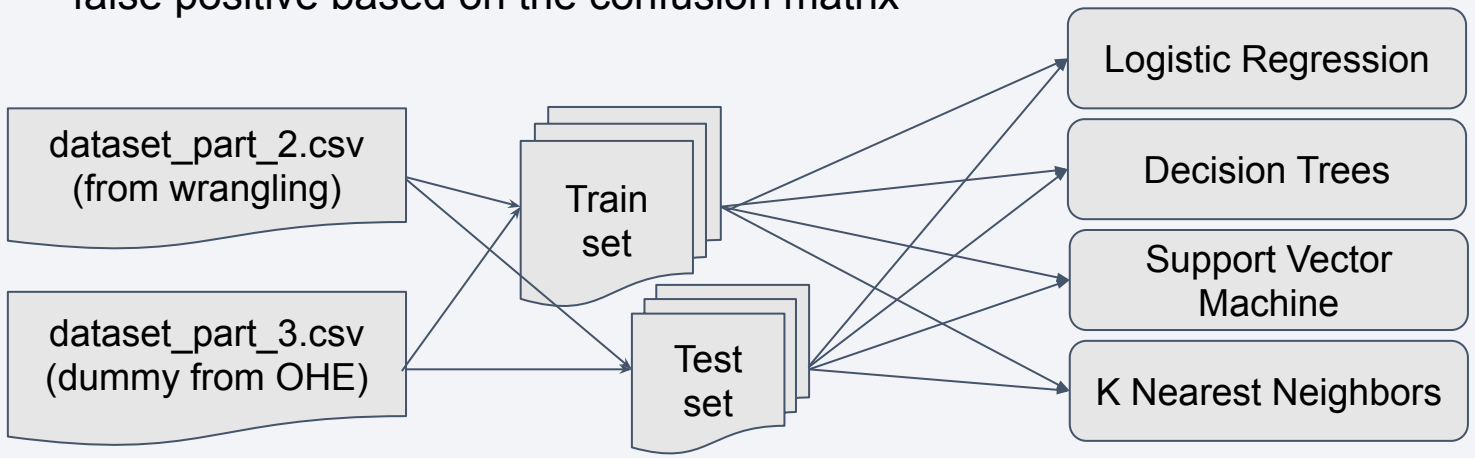
Predictive Analysis (Classification)

I used the output data of Data Wrangling and with dummy data after one hot encoding to conduct the predictive analysis. Splitting the data to train and test data sets with 8:2 ratios. (total of 90 records)

I trained the following models with GridSearchCV object:

- **Logistic Regression** using LBFGS solver with the train accuracy rate of 0.8196. Accuracy on testing data is 0.833 with 66% true positive and 17% false positive.
- **Support Vector Machine** object to train the model. Kernel **Sigmoid** generated the train model with 0.84 accuracy rate. Accuracy on testing data is 0.833 with 66% true positive and 17% false positive.
- **Decision Trees** to train the model with the train accuracy rate of 0.875. Accuracy on testing data is 0.833 with 66% true positive and 17% false positive.
- **K Nearest Neighbors** with auto algorithm and N_neighbors 3. The train accuracy rate is 0.664. Using the score method to calculate accuracy, it only has 0.611 accuracy rate. This model is the least accurate

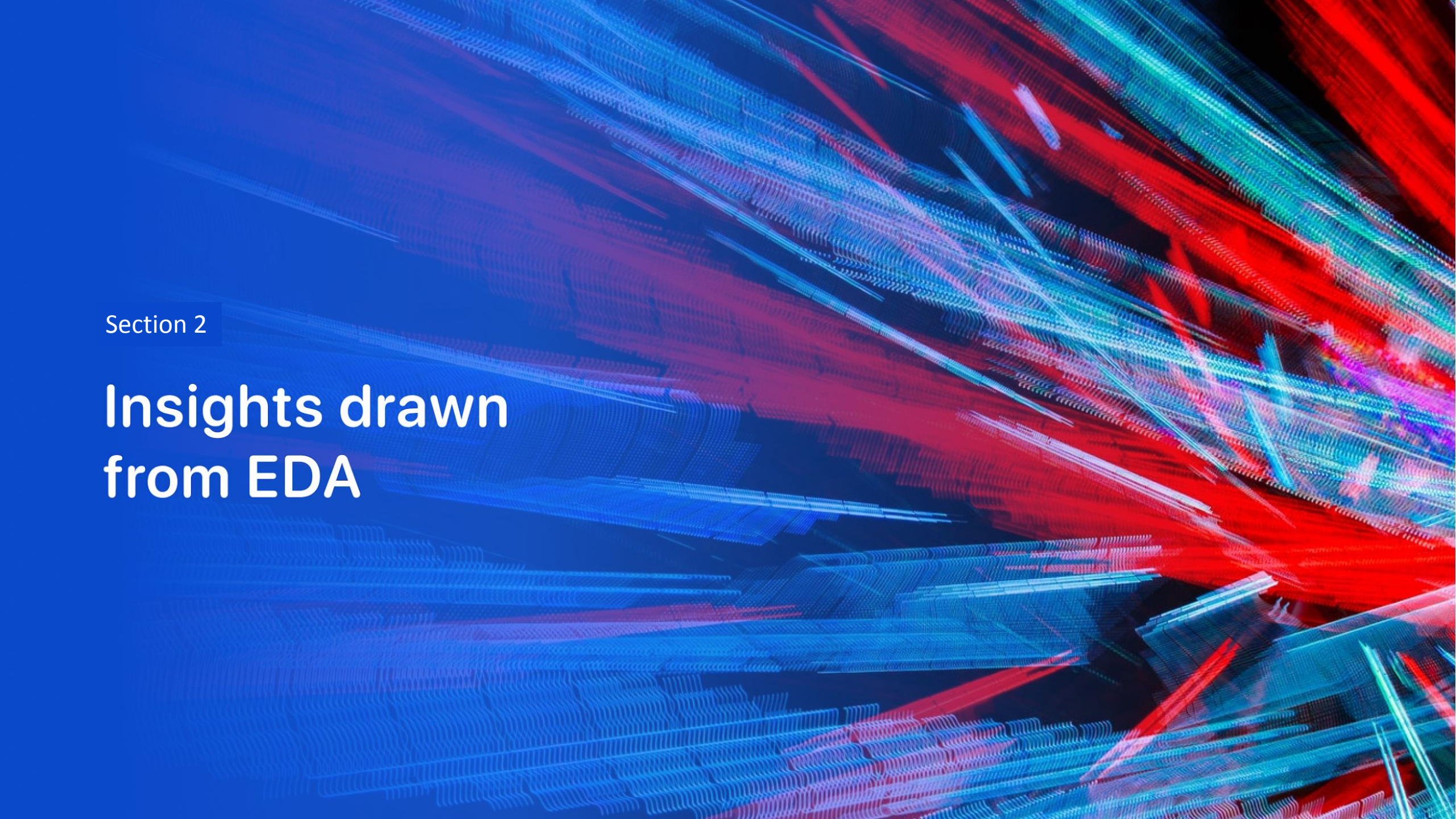
Based on the accuracy rate, t**he best model is Decision Trees** with 0.875 train accuracy rate. However, there is 17% of false positive based on the confusion matrix



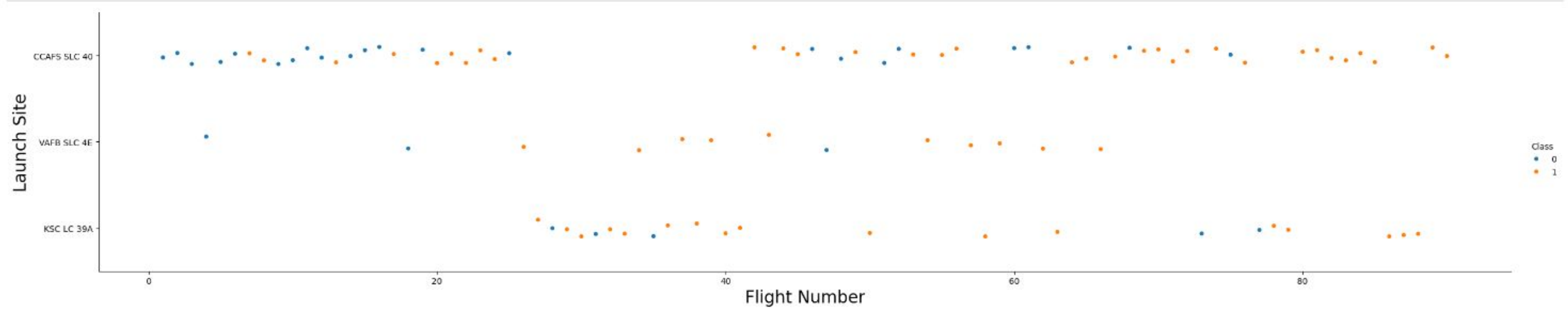
GitHub URL of predictive analysis lab

Section 2

Insights drawn from EDA



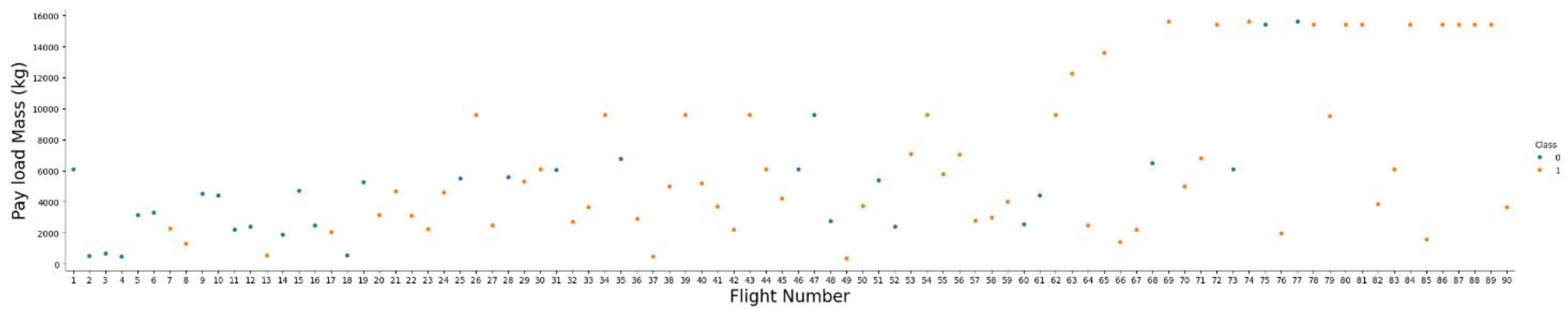
Flight Number vs. Launch Site



Key Takeaways:

1. Launch Site CCAF SLC-40 landed the most number of flights, but it also had the most failures than the other two sites with 60% success rate.
2. The other SLC site VAFB SLC 4E landed the least number of flights, but it has 76.9% success rate.
3. KSC LC 39A have the highest success rate of landing at 77%

Payload vs. Launch Site



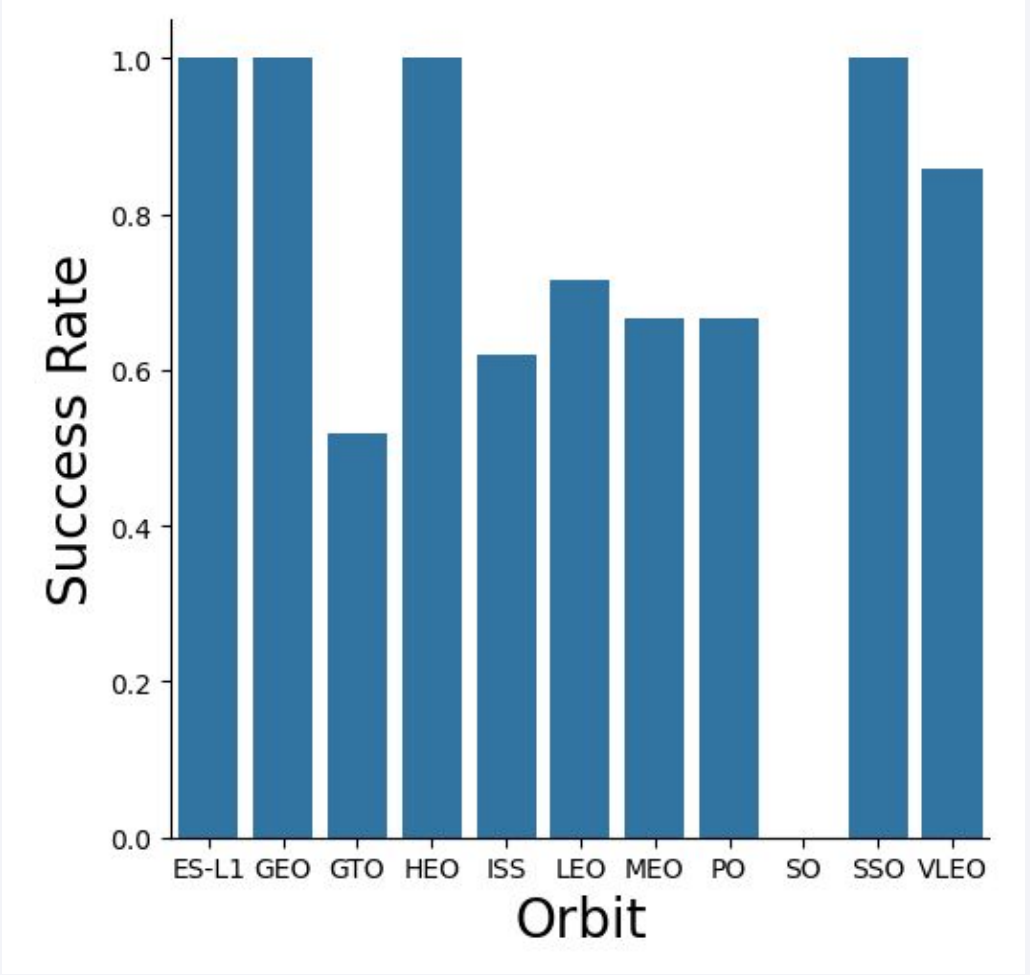
Key Takeaways:

1. There is no clear correlation between heavier payload and success rate.
2. Lighter payload (less than 10K payload mass) tended to cause more failure, especially in the early days
3. The success rates increase in the later days with larger flight numbers.

Success Rate vs. Orbit Type

Key Takeways:

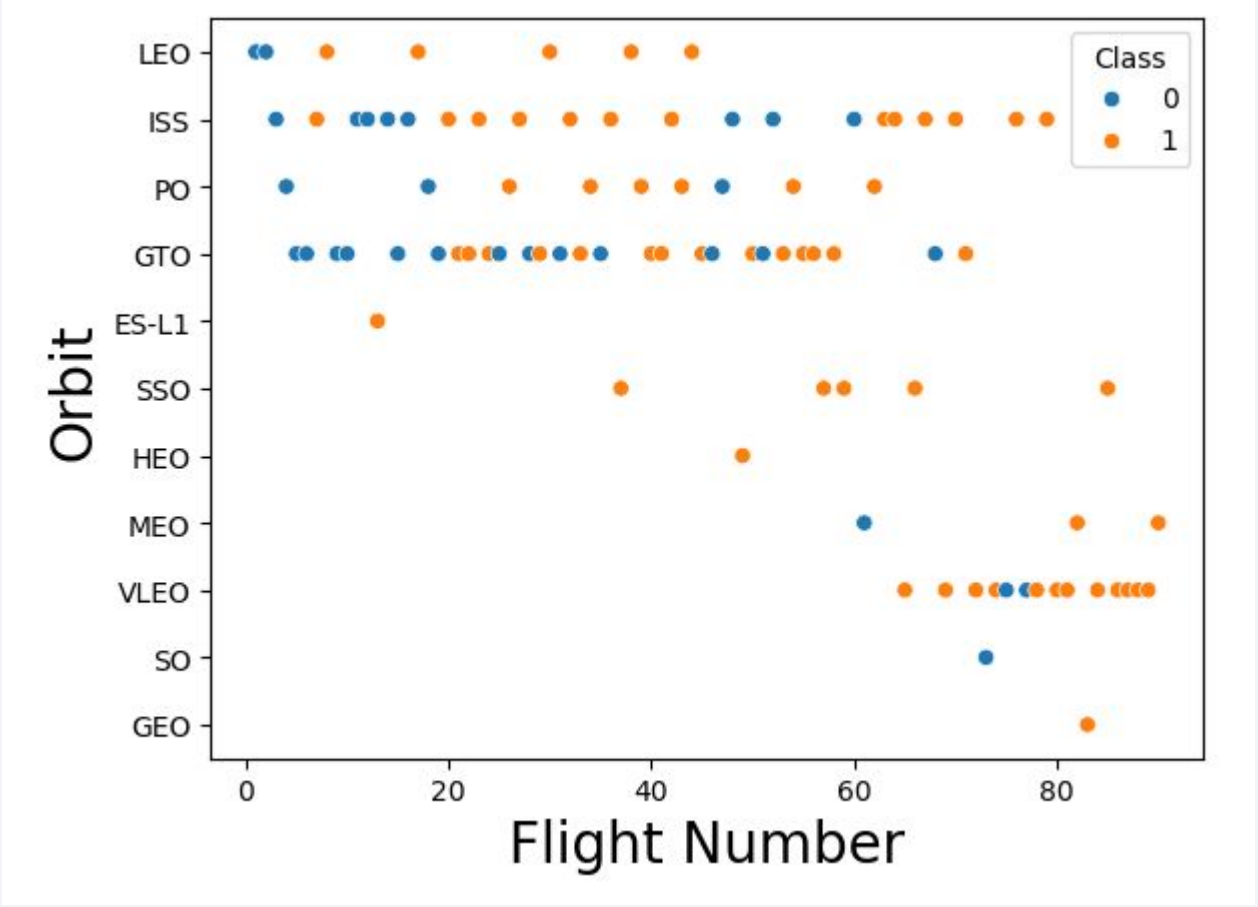
- Orbits ES-L1, GEO, HEO, and SSO have the highest success rate.
- Orbits GTO and ISS have the lowest success rate.



Flight Number vs. Orbit Type

Key Takeaways:

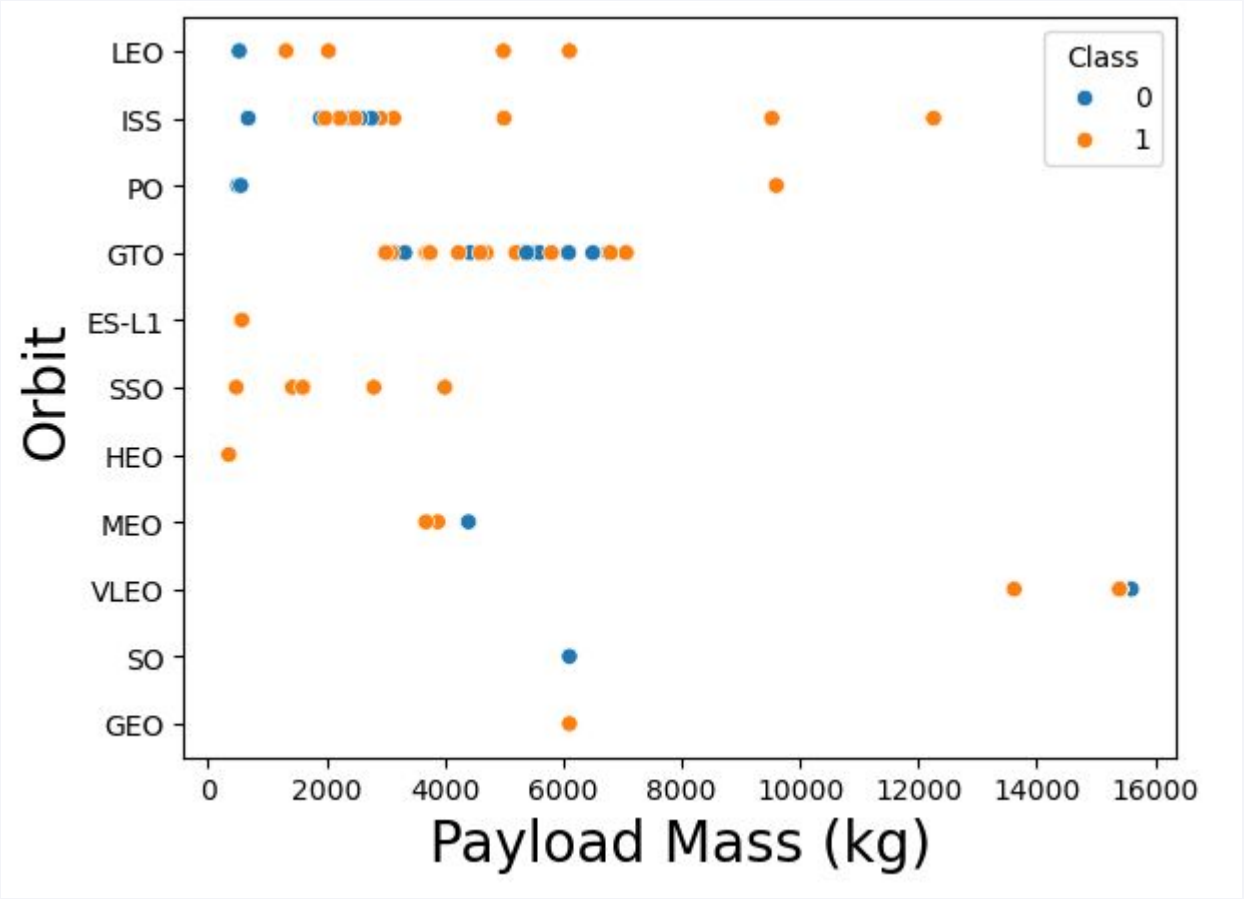
- LEO orbit the Success appears related to the number of flights.
- There seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

Key Takeaways:

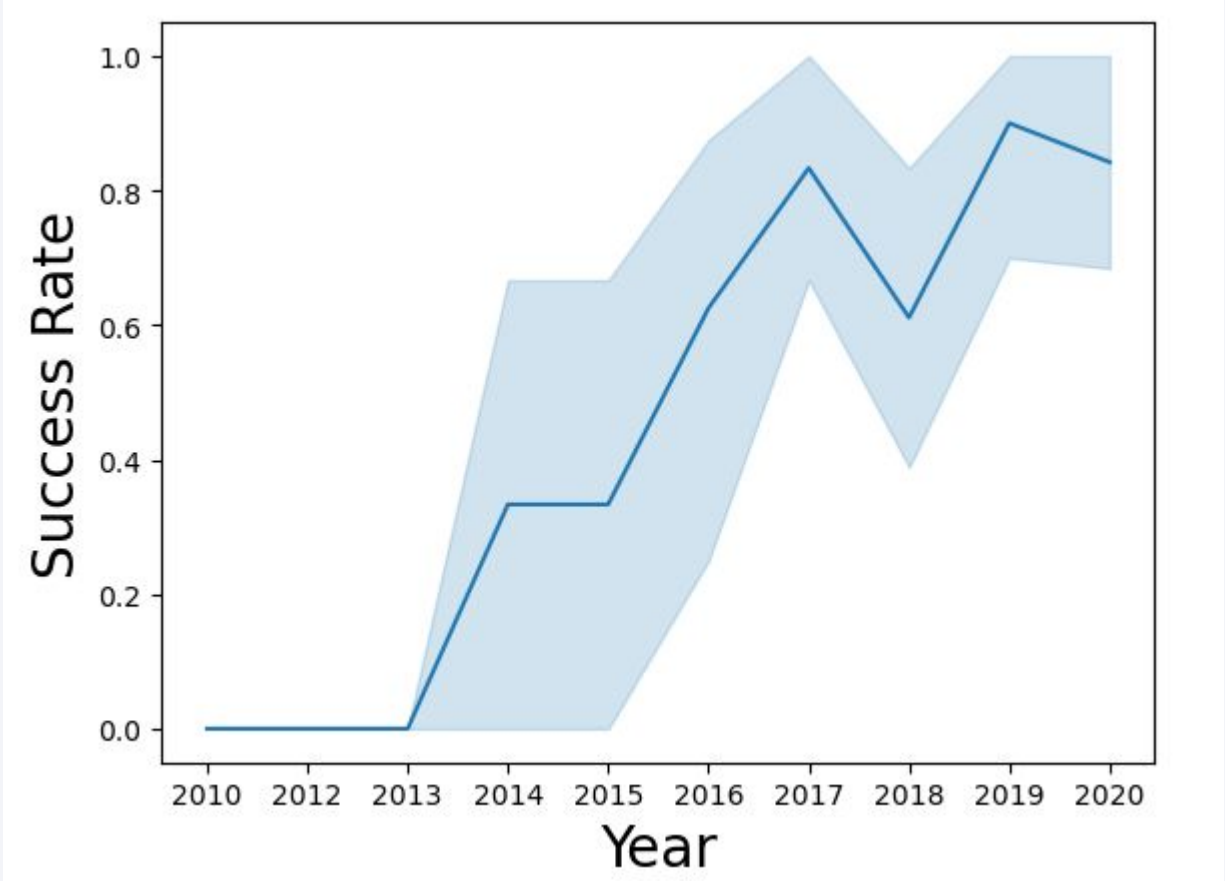
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- SSO, ES-L and HEO had all success landing with lighter payload (less than 4kg)
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here



Launch Success Yearly Trend

Key Takeways:

- Success rate has increased YoY since 2013. However, there was dip on success rate in 2018. The success rate went up again in 2019.



All Launch Site Names

There were 4 launch Sites for SpaceX

- CCAFS LC-40, CAFB SLC-4E, KSC LC-39A and CCAFS SLC-40

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

```
]: %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'

Here’s the sample of 5 launch records at Launch Site name starting with ‘CCA’

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	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

So far SpaceX boosters carried 45,596 kg of total Payload mass launched by NASA (CRS)

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

```
%sql SELECT sum(PAYLOAD_MASS__KG_) as "Total Payload Mass" from SPACEXTABLE where Customer = 'NASA (CRS)'
```

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

Total Payload Mass
45596

Average Payload Mass by F9 v1.1

The average Payload Mass carried by booster version F9 v1.1 is 2,928.4 kg

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

```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

2015/12/22 was the dates of the first successful landing outcome on ground pad

```
%sql Select min(date) from SPACEXTABLE where Landing_outcome = 'Success (ground pad)'
```

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

min(date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Here’s the list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 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, Payload from SPACEXTABLE where Landing_outcome = 'Success (drone ship)' and ( PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000)
```

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

Booster_Version	Payload
F9 FT B1022	JCSAT-14
F9 FT B1026	JCSAT-16
F9 FT B1021.2	SES-10
F9 FT B1031.2	SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

The total number of successful mission outcomes: 100
The total number of failure mission outcomes: 1

List the total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(*) from SPACEXTABLE group by 1
```

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

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

Boosters Carried Maximum Payload

List the names of the booster which have carried the maximum payload mass

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_ in (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

Here’s the list of failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql Select substr(Date, 6,2) as month, landing_outcome, Booster_Version, Launch_Site
from SPACEXTABLE
where substr(Date,0,5)='2015' and landing_outcome = 'Failure (drone ship)'
```

```
 * 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

Here the ranking of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

```
%sql select landing_outcome, count(*) from SPACEXTABLE
where Date between '2010-06-04' and '2017-03-20'
group by landing_outcome order by count(*) desc
```

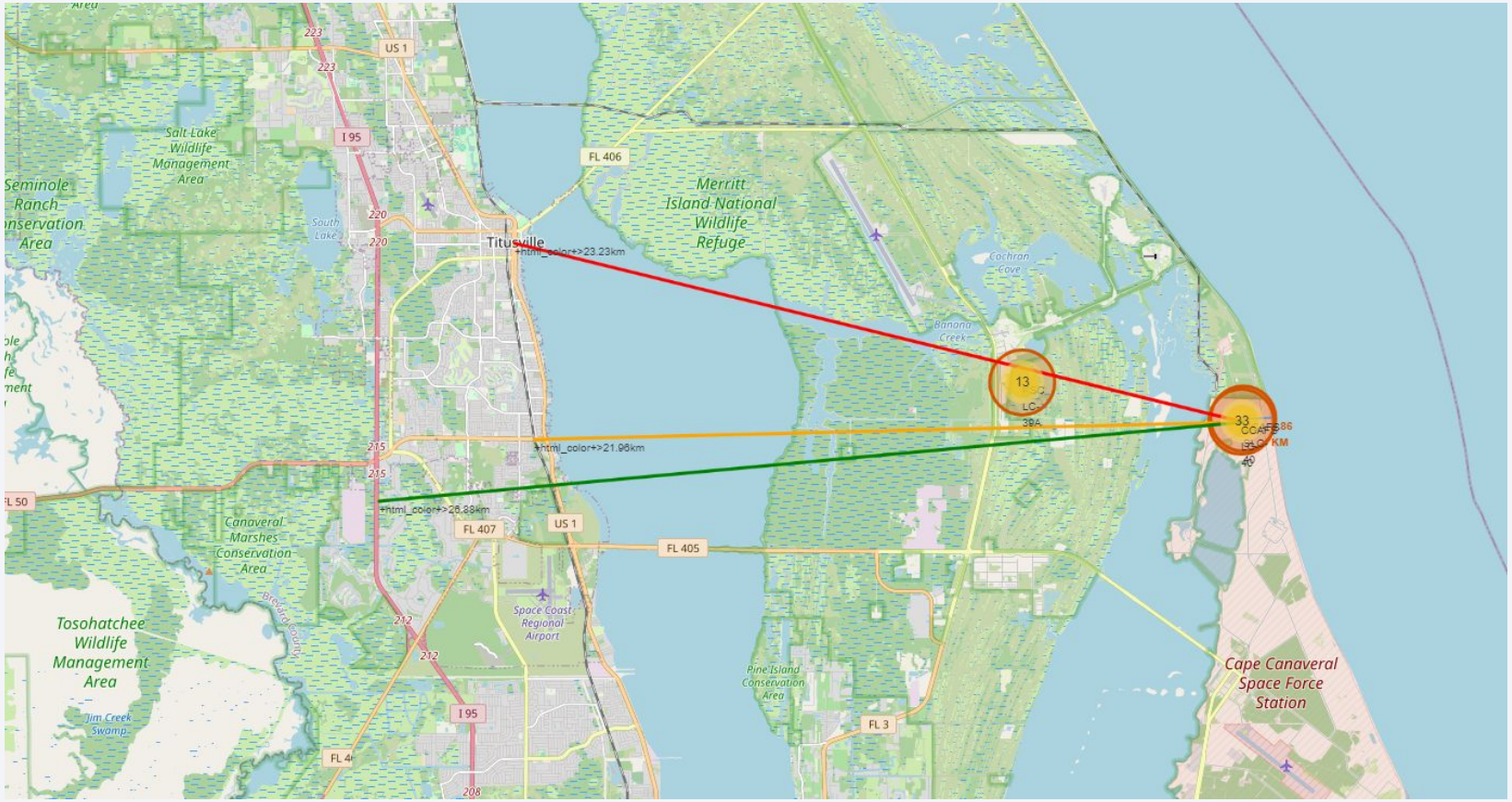
```
 * sqlite:///my_data1.db
Done.
```

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

Distances between launch sites to its proximities

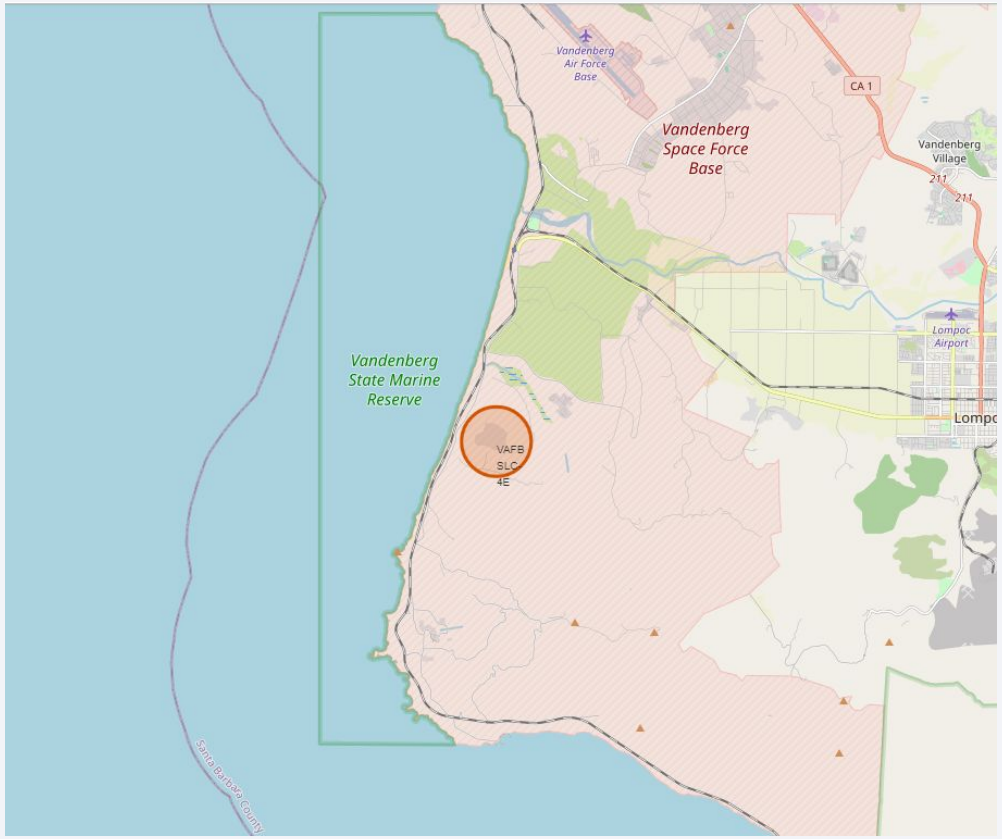
Launch sites were located far from railroad, highway, and cities in order to avoid impact (noise, pollution, fire, damage, etc)

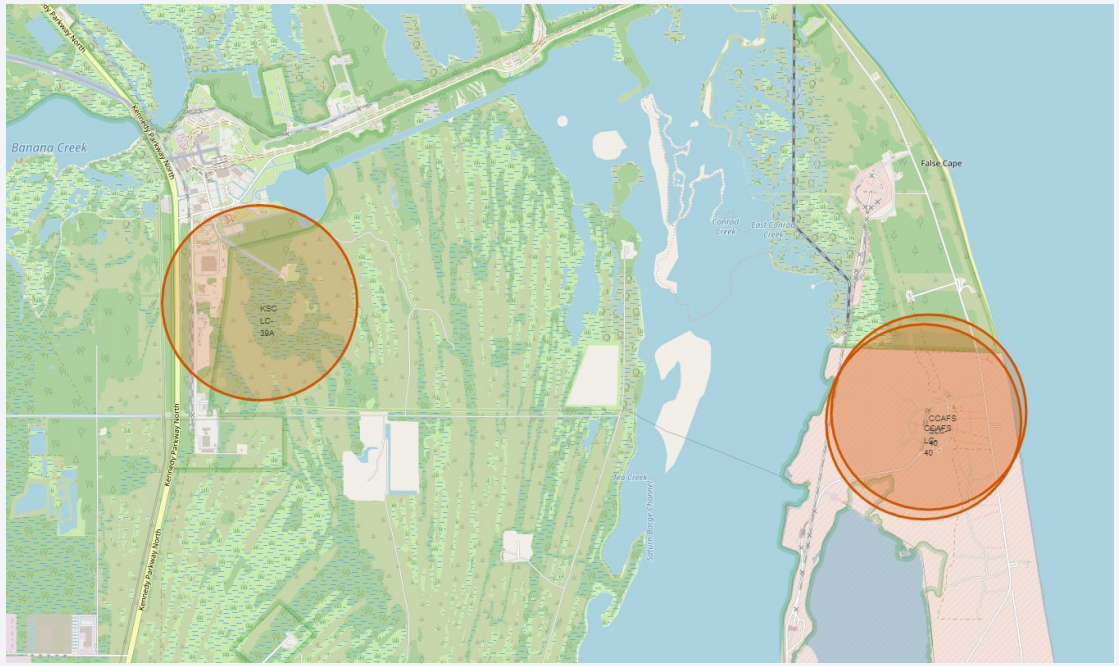


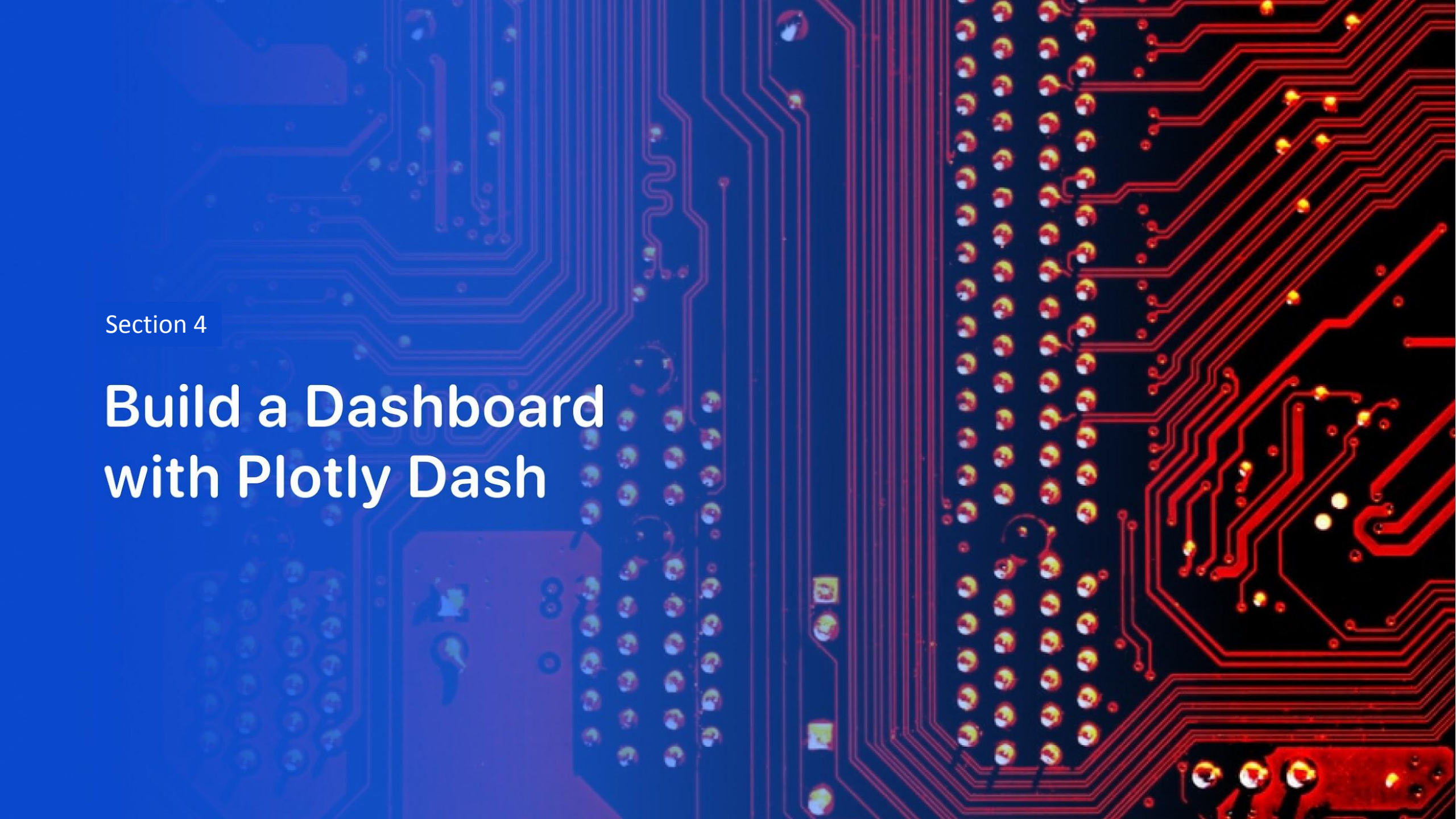
Mark all launch sites on a map

There were 4 launch sites. The screenshot at the left shows CAFB SLC-4E in California. The screenshot at the right shows 3 launch sites CCAFS LC-40, KSC LC-39A and CCAFS SLC-40 in Florida.

All 4 launch sites are close to the coastline and near equator

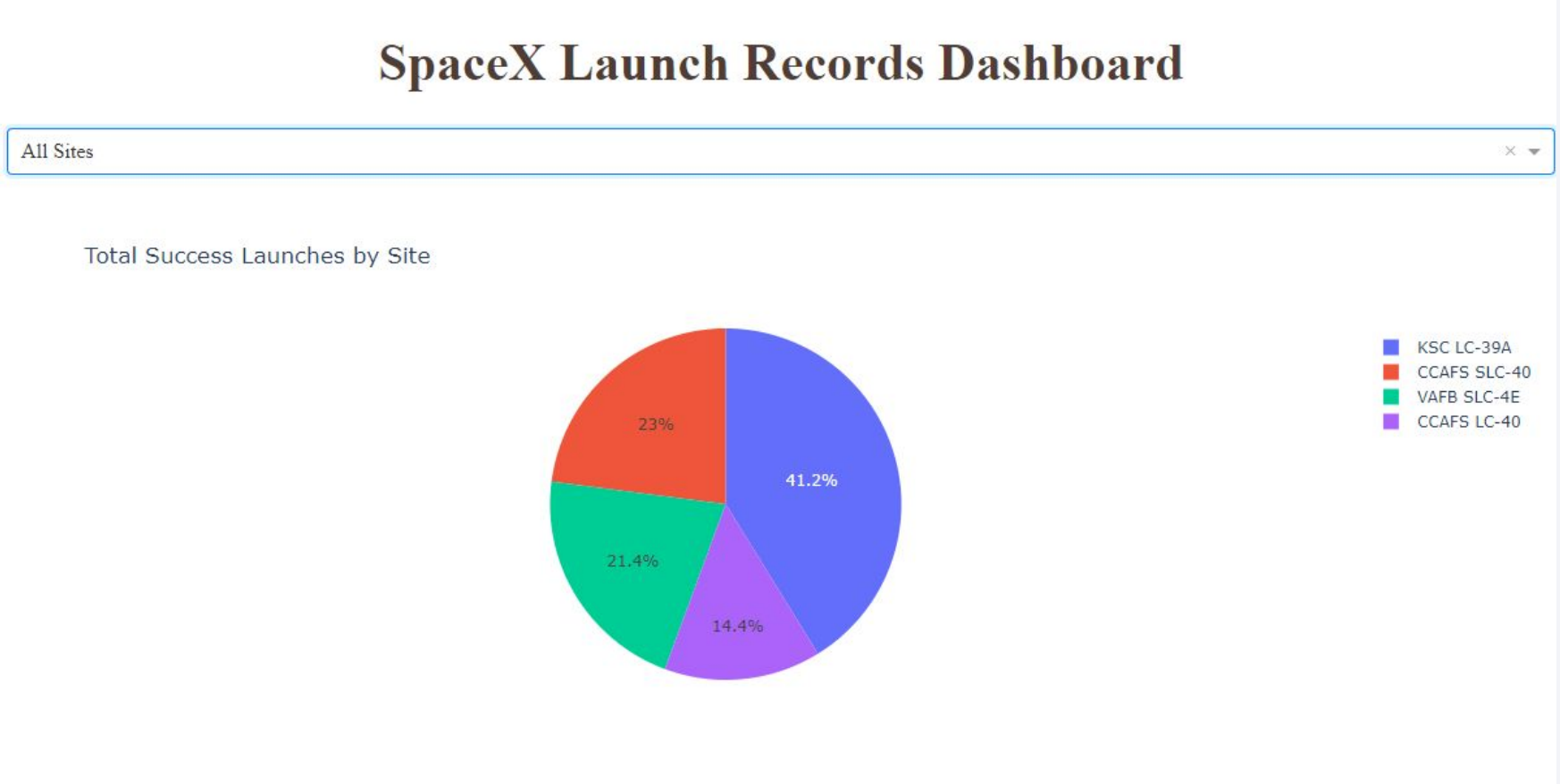





Section 4
Build a Dashboard with Plotly Dash

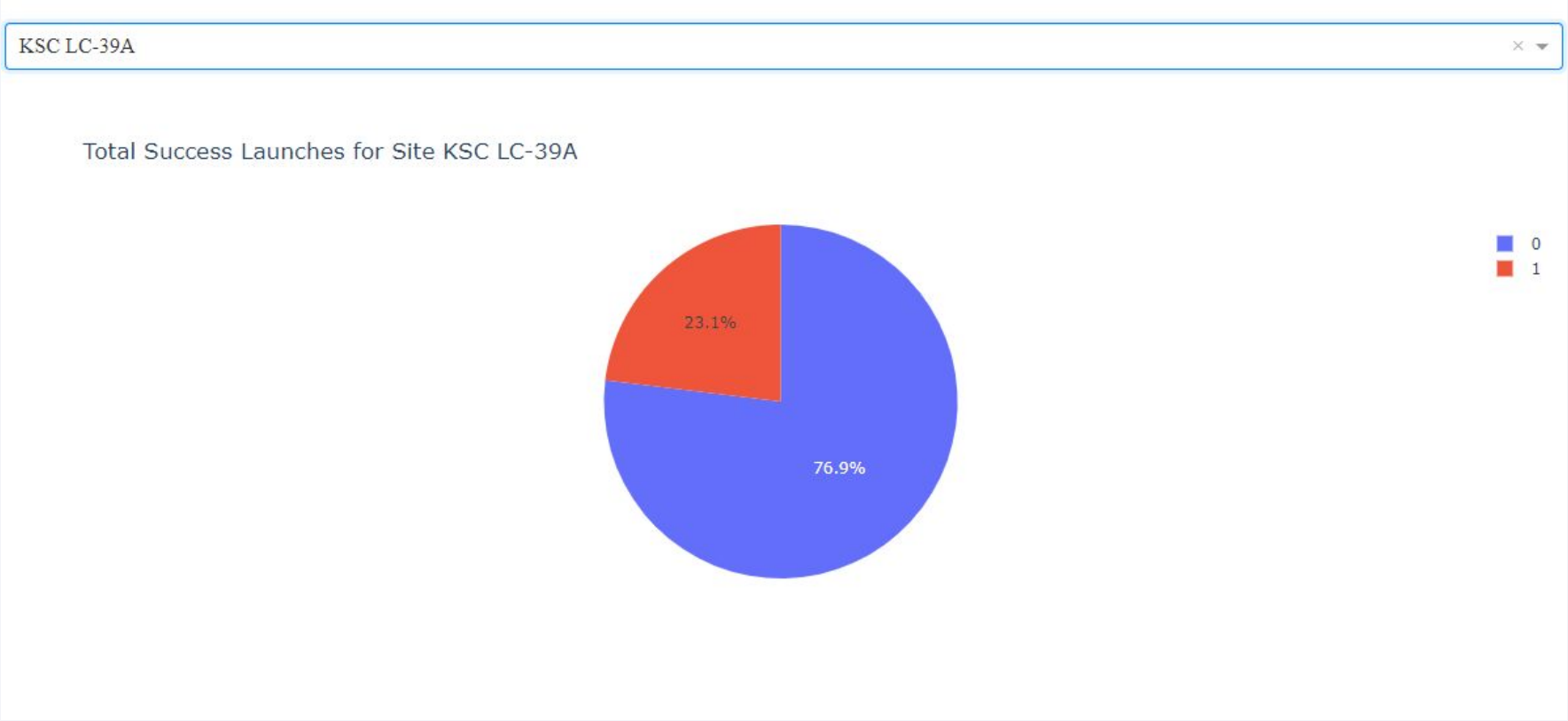
Total Success Launches for all sites

The following pie chart show the success launches % by site. KSC LC-39A has the highest success launches rate of 41.2%. CCAFS SLC-40 and VAFB SLC-4E have success launches rate of 23% and 21.4% respectively. CCAFS LC-40 has the lowest success rate of 14.4%



Launch Site with the highest launch success ratio

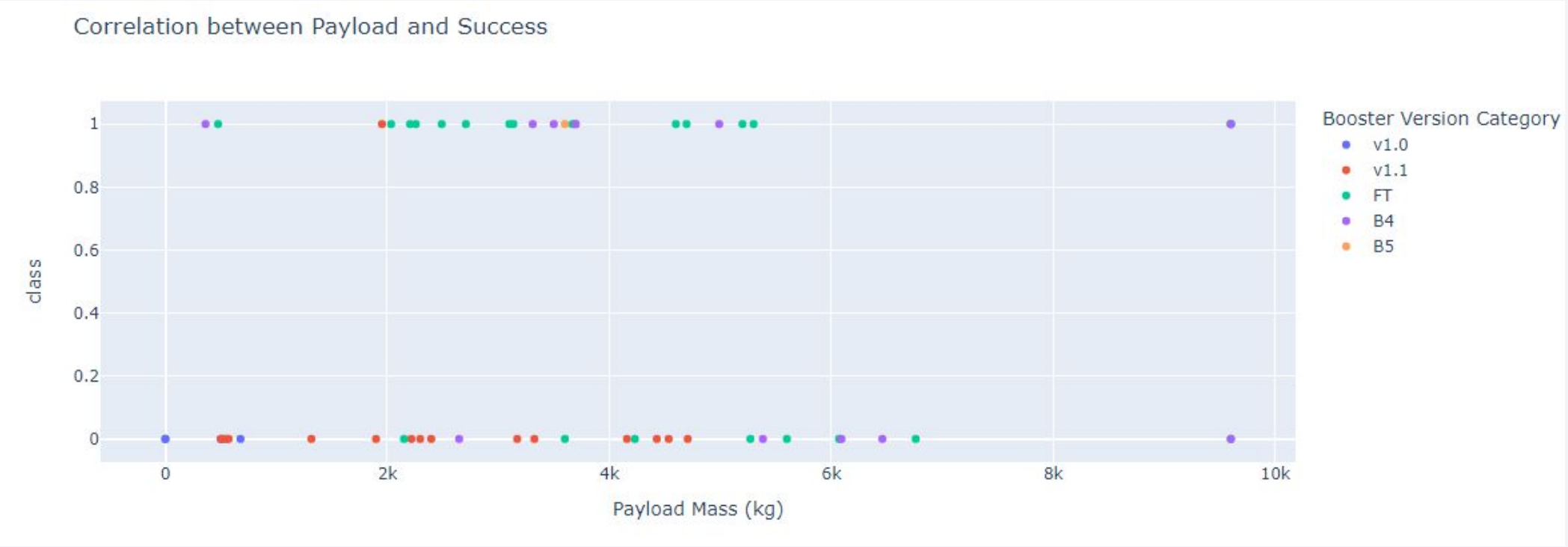
Site KSC LC-39A has the highest launch success ratio with 76.9% success rate



Correlation between Payload and Success Rate

Booster version v1.1 have the lowest success rate with the payload in range of 0.5K to 5K, with only 1 success landing with 2K payload.

In contrast, Booster version FT has the highest success rate with the payload in range of 2K and 5.5K. However, it failed landing with heavier payload over 5.5K



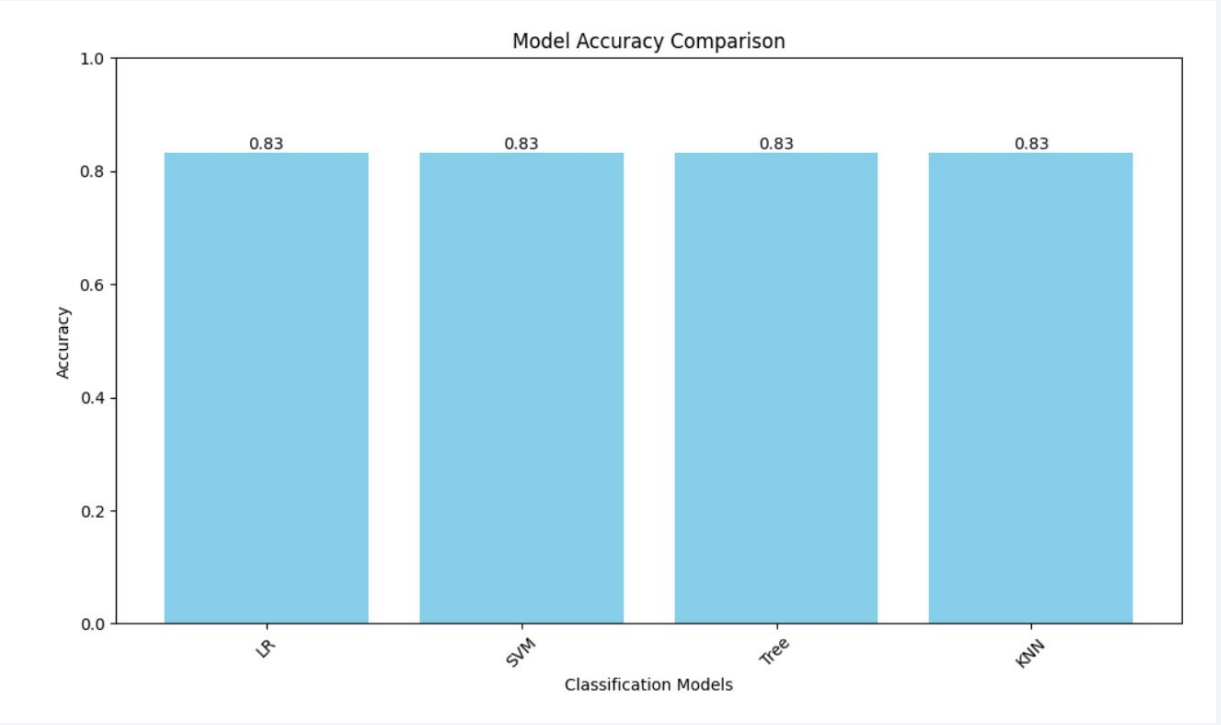
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- All 4 prediction models generate the similar accuracy of 0.83 on test data.
- However, all models generated false positive in the confusion matrices. KNN generated the most false positive (predict successful landing, but it was not)

All models need further fine-tuning to increase it accuracy.

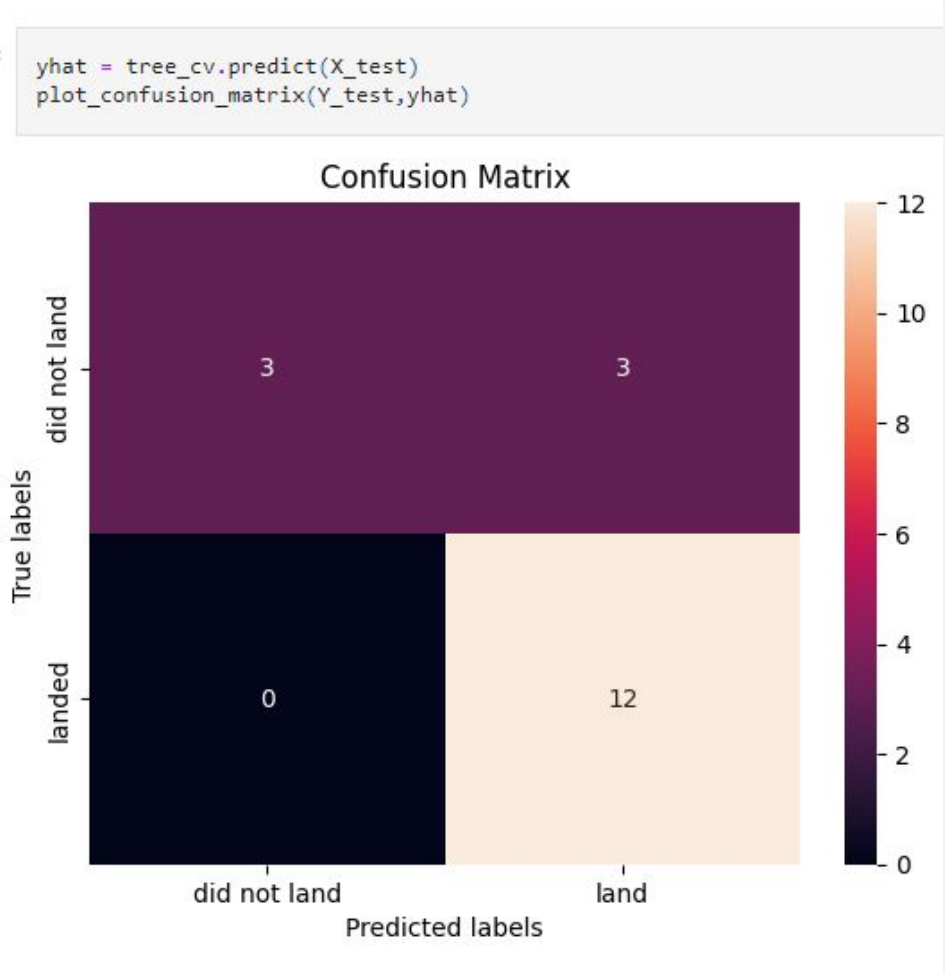


Confusion Matrix

Decision Tree, Logistic Regression and SVM all have similar confusion matrix. But Decision Tree model has higher train model accuracy of 0.875. And score accuracy of 0.833 on test data.

Out of 18 test data, 15 of it was predicted correctly (12 true positive and 3 true negative).

However, it predicted 3 with false positive, and 3 with false negative.



Conclusions

- All 4 prediction models generate similar prediction. All have false positive (predict successful landing, but it was not).
- Further model tuning is recommended. Considering to get more data and collect other attributes (such as weather esp wind speed and direction) to find correlations
  - Site KSC LC-39A has the highest success ratio with 76.9% success rate
  - Site VAFB SLC 4E did not launch any heavy rocket with payload Mass more than 10K, but its success rate is very high
  - Orbits ES-L1, GEO, HEO, and SSO have the highest success rate.

Thank you!