IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gary Lin

2022/07/18

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection – Rest API, Webscraping
  - Data Wrangling
  - EDA with Visualization
  - EDA with SQL
  - Build an interactive map with Folium
  - Build a dashboard with plotly dash
  - Predictive Analysis (Classification)
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- Project background and context

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. 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 SpaceX for a rocket launch.

- Problems you want to find answers
  - Find the relationship between each variable that affects the outcome of landing.
  - Observe the insights of each launch site, geography attributes, and similarity.
  - Predict outcome for further launches, and reduce cost to the minimum.

Section 1

Methodology

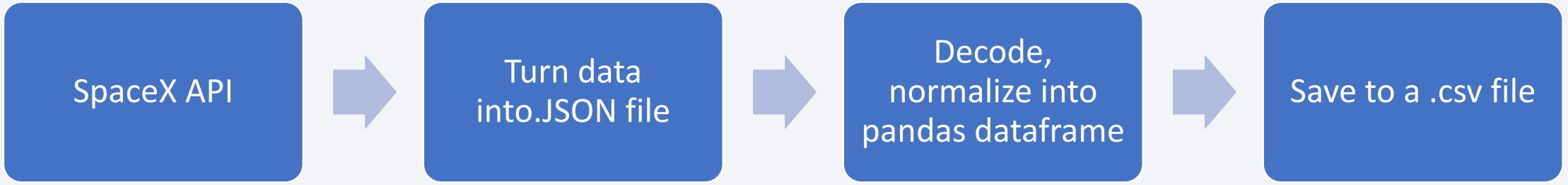
Methodology

Executive Summary

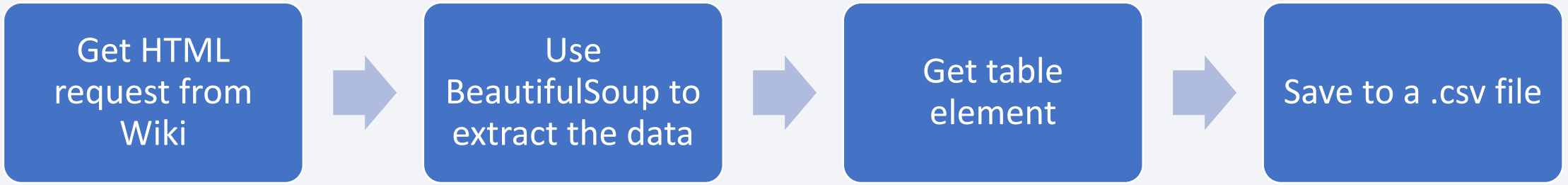
- Data collection methodology
  - SpaceX REST API
  - Webscraping from Wikipedia
- Perform data wrangling
  - Replace ‘PayloadMass’ with mean value
  - As for other columns with missing value, we delete those rows.
  - Use One Hot Coding for column ’Class’ for further classification work.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Use KNN, Logistic Regression, Decision Tree, SVM to predict outcome.

Data Collection

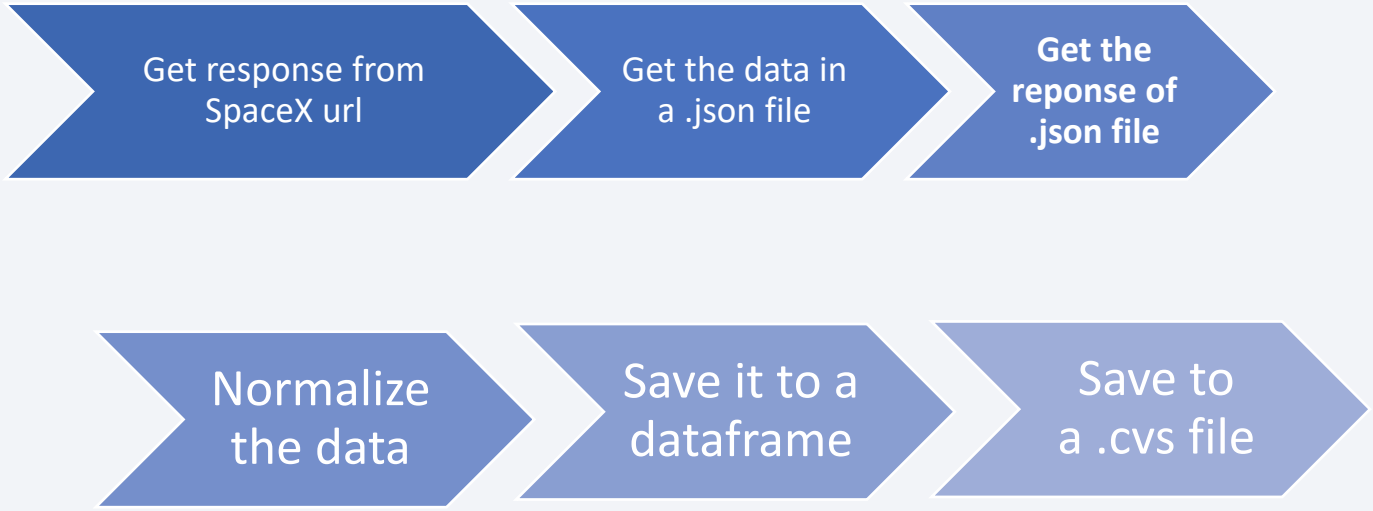
Using SpaceX REST API



Webscraping from Wikipedia

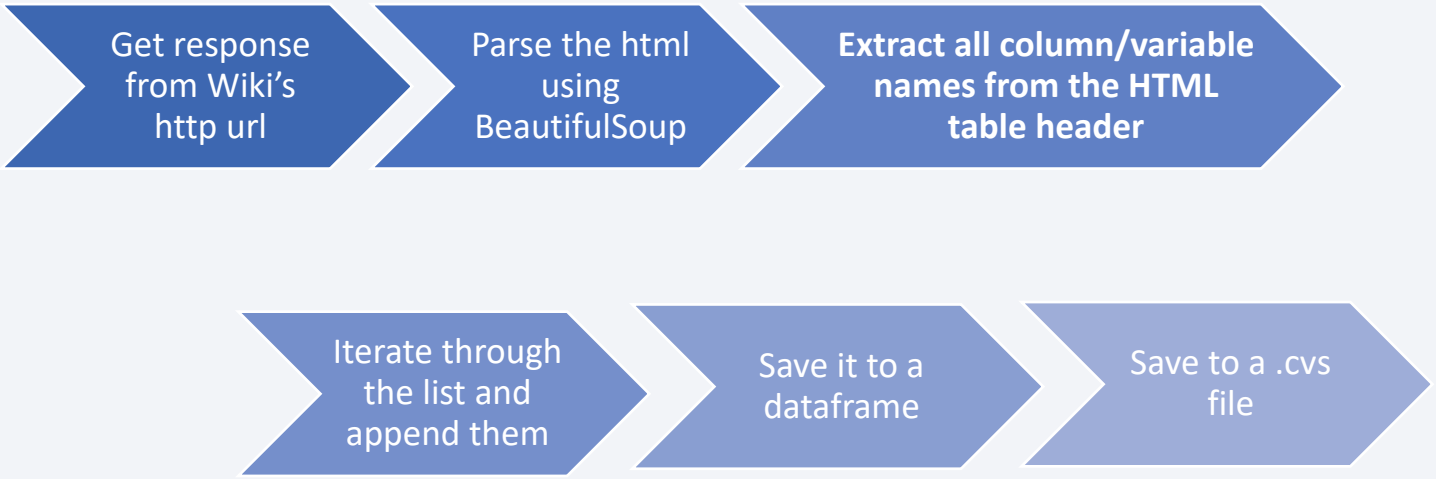


Data Collection – SpaceX API



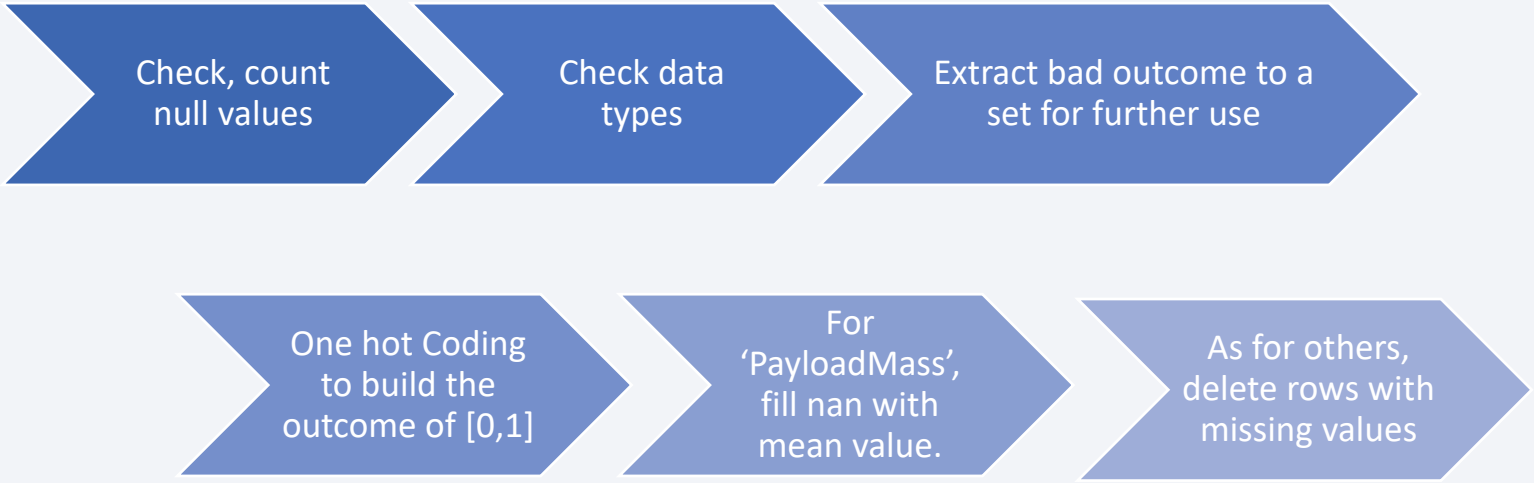
Data Collection_SpaceX API

Data Collection - Scraping



Data Collection with Webscraping

Data Wrangling

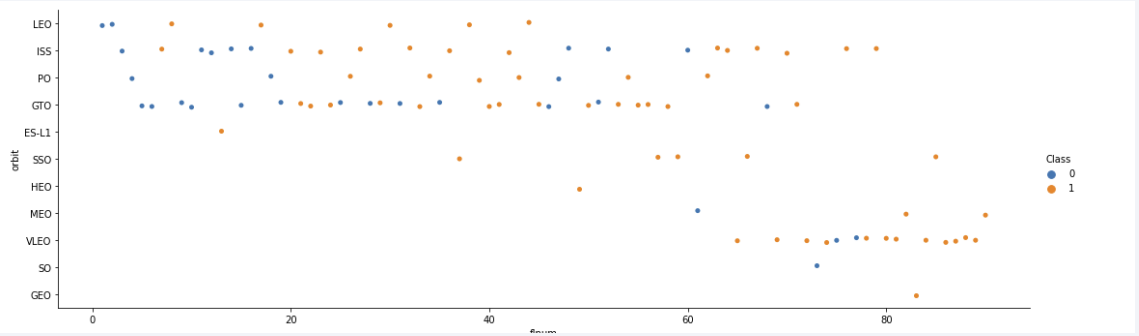


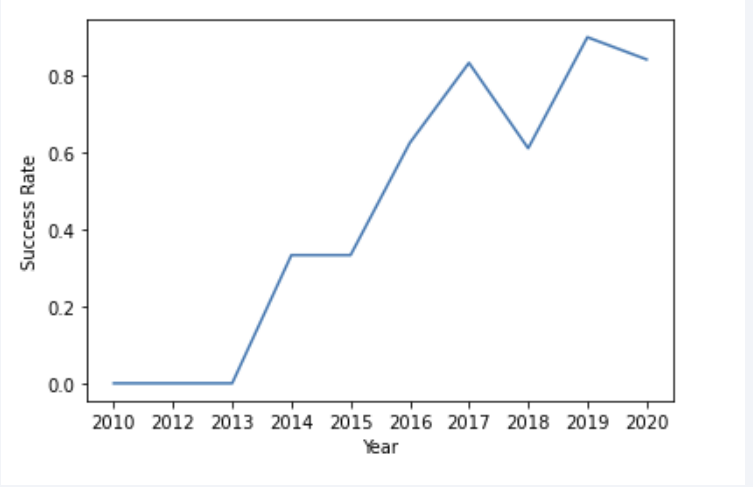
Data Wrangling

EDA with Data Visualization

We use visualization to see relationship between variables.

- Flight number vs Payload Mass
- Flight number vs Launch Site
- Payload Mass vs Launch Site
- Success rate for each orbit type
- Flight number vs orbit type
- Payload vs orbit type
- Yearly Launch trend





EDA with Visualization

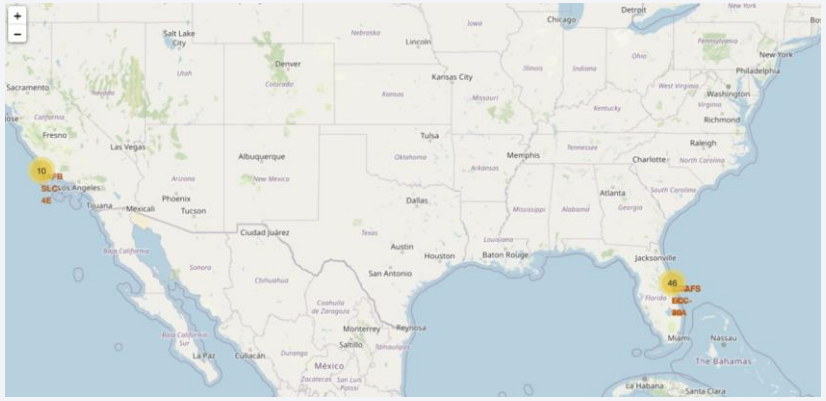
EDA with SQL

- Display the names of the unique launch 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 succesful landing outcome in ground pad was acheived.
- 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. Use a subquery
- 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.
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Build an Interactive Map with Folium

Stuffs I added :

- Add labels to each launch site
- Mark the cluster of launch for each launch site
- Use green, red label to mark the success and fail launch
- Calculate distance of launch site to its proximities.

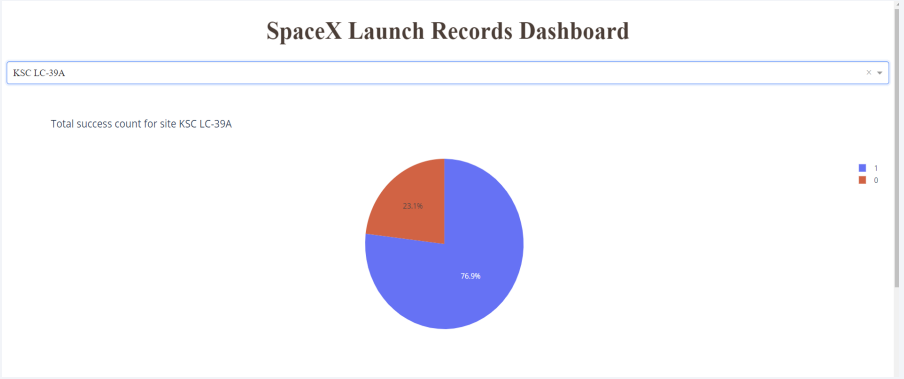


Interactive Map with Folium

Build a Dashboard with Plotly Dash

Stuffs I added :

- Add pie chart to see the success launch distribution for each launch site.
- Scatter plot established with a Payload range slider to see what range of payload leads to greater chance of successful launch



Dashboard with Plotly Dash

Predictive Analysis (Classification)

Preprocessing

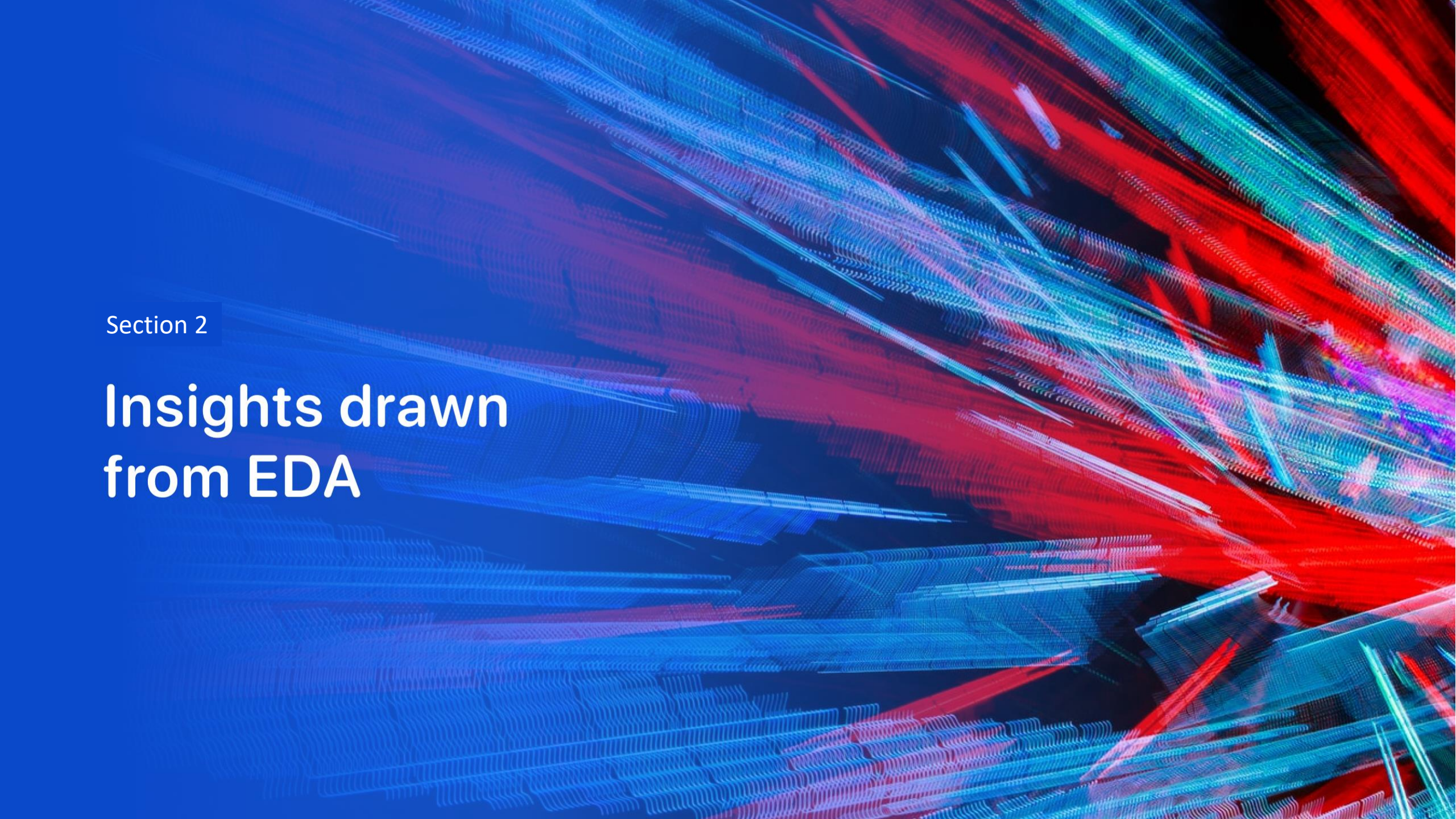
- Convert ‘class’ to numpy (Y)
- Standardize the data (X)
- Split X and Y into training, testing dataset with test size =0.2

Modeling (Using GridSearch CV to find the best parameter)

- Logistic Regression
- SVM
- Decision Tree
- KNN

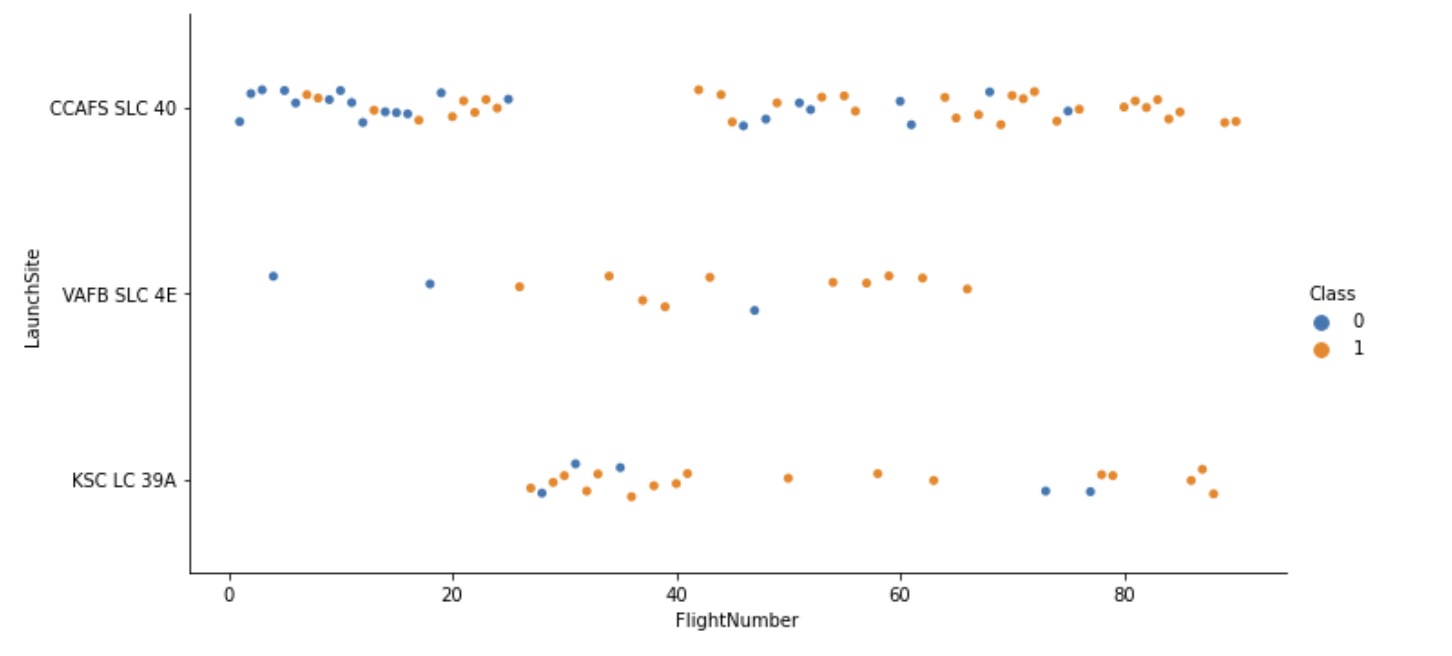
Results

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


Section 2
Insights drawn from EDA

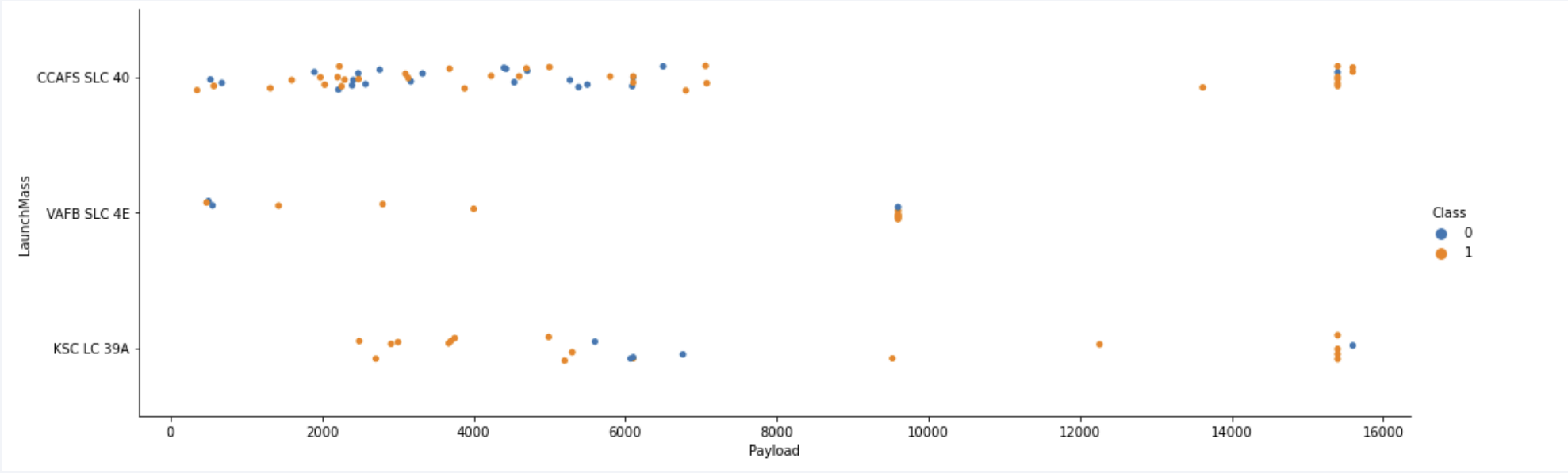
Flight Number vs. Launch Site

- For all launch sites, we can see that **successful launches increase with the increase of FlightNumber**



Payload vs. Launch Site

- Once **Payload >7000kg**, the chance of successful launch **highly increases**.
- For **VAFB SLC-4E**, there is **no launch** with **payload greater than 10000**.



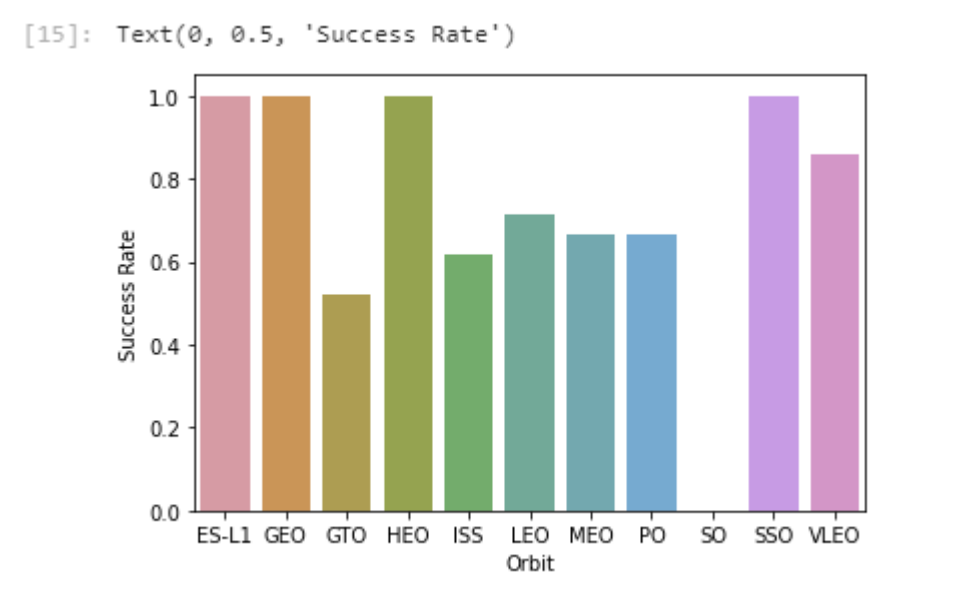
Success Rate vs. Orbit Type

Orbit type with Highest success rate:

- ES-L1
- GEO
- HEO
- SSO

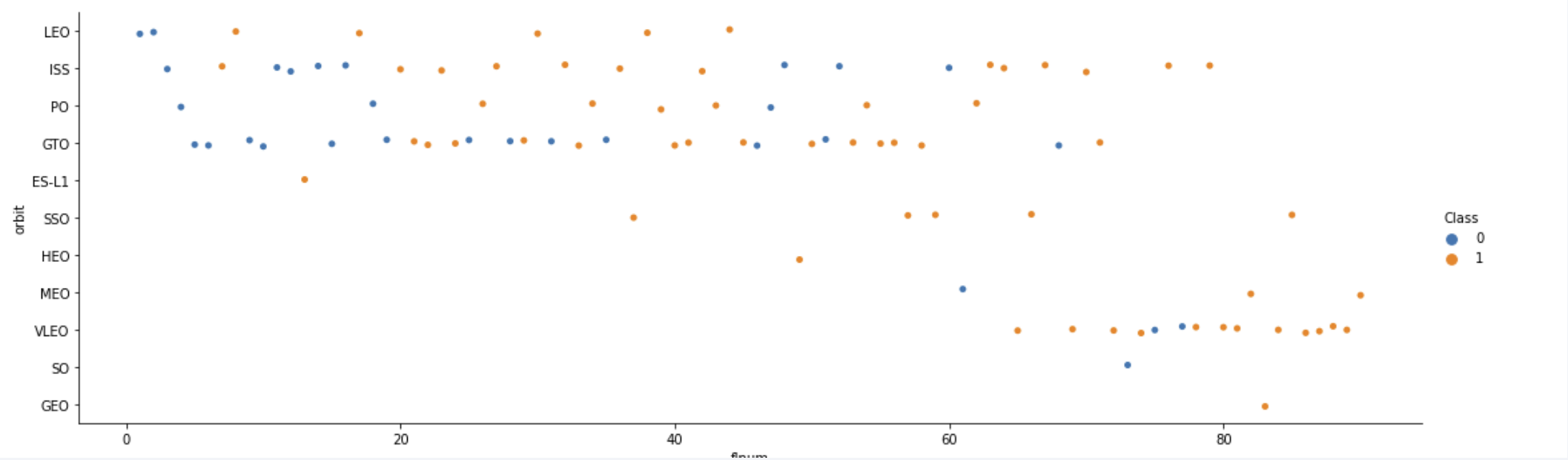
Orbit type with Lowest success rate:

- SO



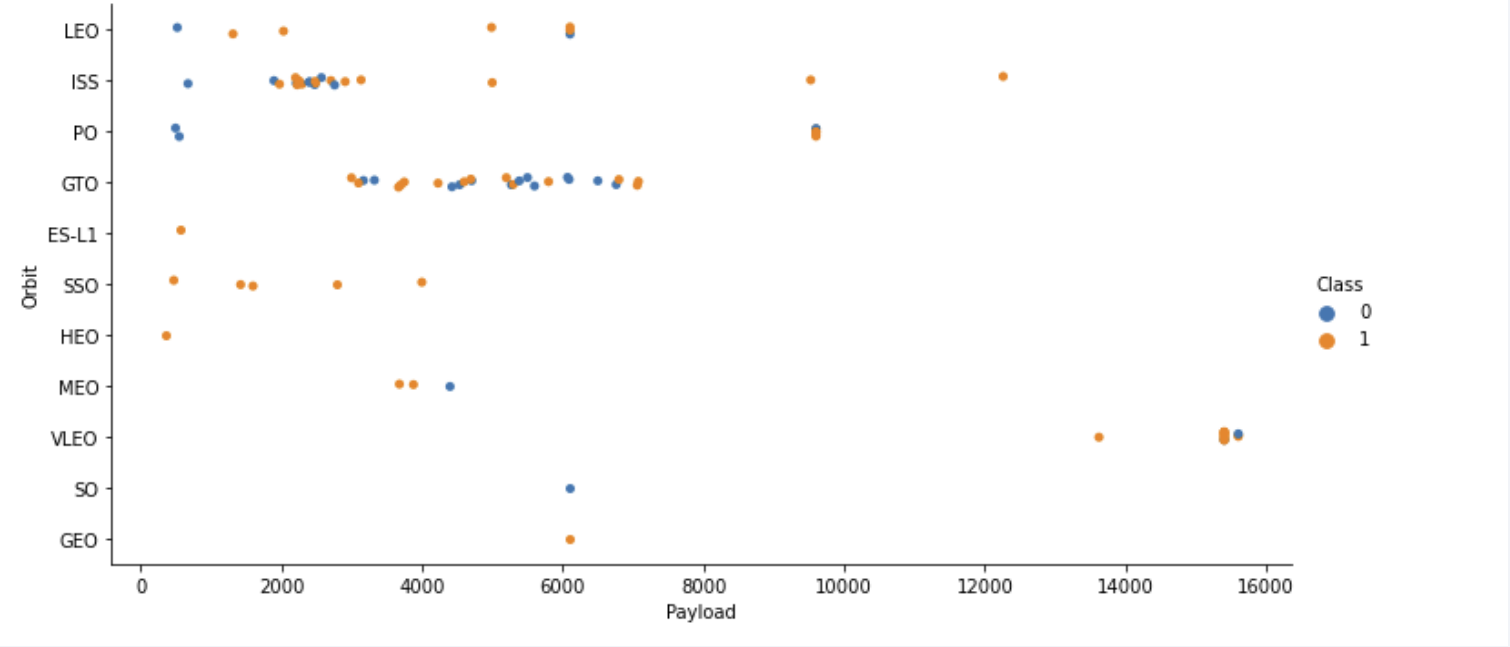
Flight Number vs. Orbit Type

- Generally, the **success rate increase as the FlightNumber increase.**



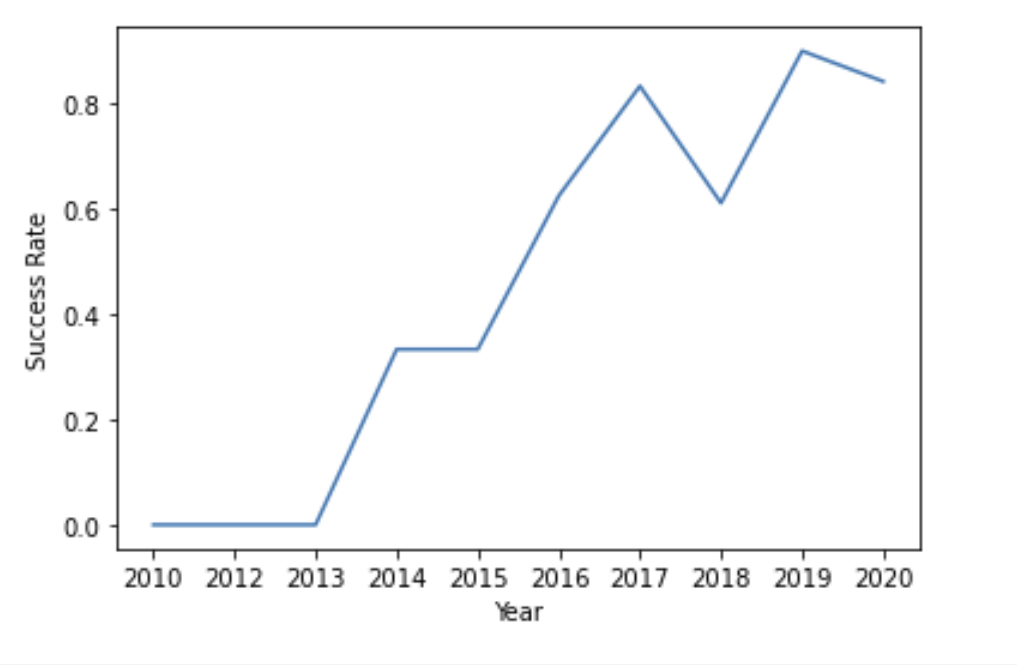
Payload vs. Orbit Type

- For LEO, ISS, PO, VLEO, the **success rate increase** as the **payload increase.**
- For GTO, there's no clear evidence that show the relationship **between payload vs Orbit type**.



Launch Success Yearly Trend

- The success rate is increasing **Year by Year**, and **reaches its peak in 2019.**



All Launch Site Names

All Launch Site Names :

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

```
%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'

- Use **LIMIT, %** to determine the **5 records begin 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
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

- Use **SUM** to determine **total payload mass**.

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL;
```

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

sum(PAYLOAD_MASS__KG_)
619967

Average Payload Mass by F9 v1.1

- Use **AVG** to determine **average payload mass**.

```
%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

- Use **MIN(date)** to determine **the First successful ground landing date.**

```
select min(date) from SPACEXTBL  where LANDING__OUTCOME = 'Success (ground pad)';
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- Use **Between and** to determine the records, and it shows **4 results** here.

```
select Booster_Version from SPACEXTBL
where LANDING__OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4001 and 5999;
```

歷程 | Results

結果集 1 | 詳細資料

Filter table 總計:4

BOOSTER_VERSION
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

Use **Count()**to determine the records, which has :

- 99 Success
- 1 Failure
- 1 Success with unknown payload status.

```
select Mission_Outcome, count(Mission_Outcome)  from SPACEXTBL group by Mission_Outcome ;
```

歷程 | Results

結果集 1 | 詳細資料

Filter table — 總計:3

MISSION_OUTCOME	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Use **subquery** to determine the records, along with 12 results here.

```
select Booster_Version from SPACEXTBL
where  PAYLOAD_MASS__KG_ =  (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

2015 Launch Records

- Use subquery, YEAR(Date)=2015 to determine the records, along with 2 results here.

```
select Launch_Site, Booster_Version, Date, Landing__Outcome from SPACEXTBL
where Landing__Outcome = 'Failure (drone ship)' and YEAR(Date) = 2015;
```

歷程 | Results

結果集 1 | 詳細資料

Filter table 總計:2

LAUNCH_SITE	BOOSTER_VERSION	DATE	LANDING__OUTCOME
CCAFS LC-40	F9 v1.1 B1012	2015-01-10	Failure (drone ship)
CCAFS LC-40	F9 v1.1 B1015	2015-04-14	Failure (drone ship)

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

- Use **group by, order by** to determine outcome between 2010-06-04 and 2017-03-20

```
select count(*), LANDING__OUTCOME from SPACEXTBL
where Date between '04-06-2010' and '20-03-2017' group by LANDING__OUTCOME order by count(*) DESC;
```

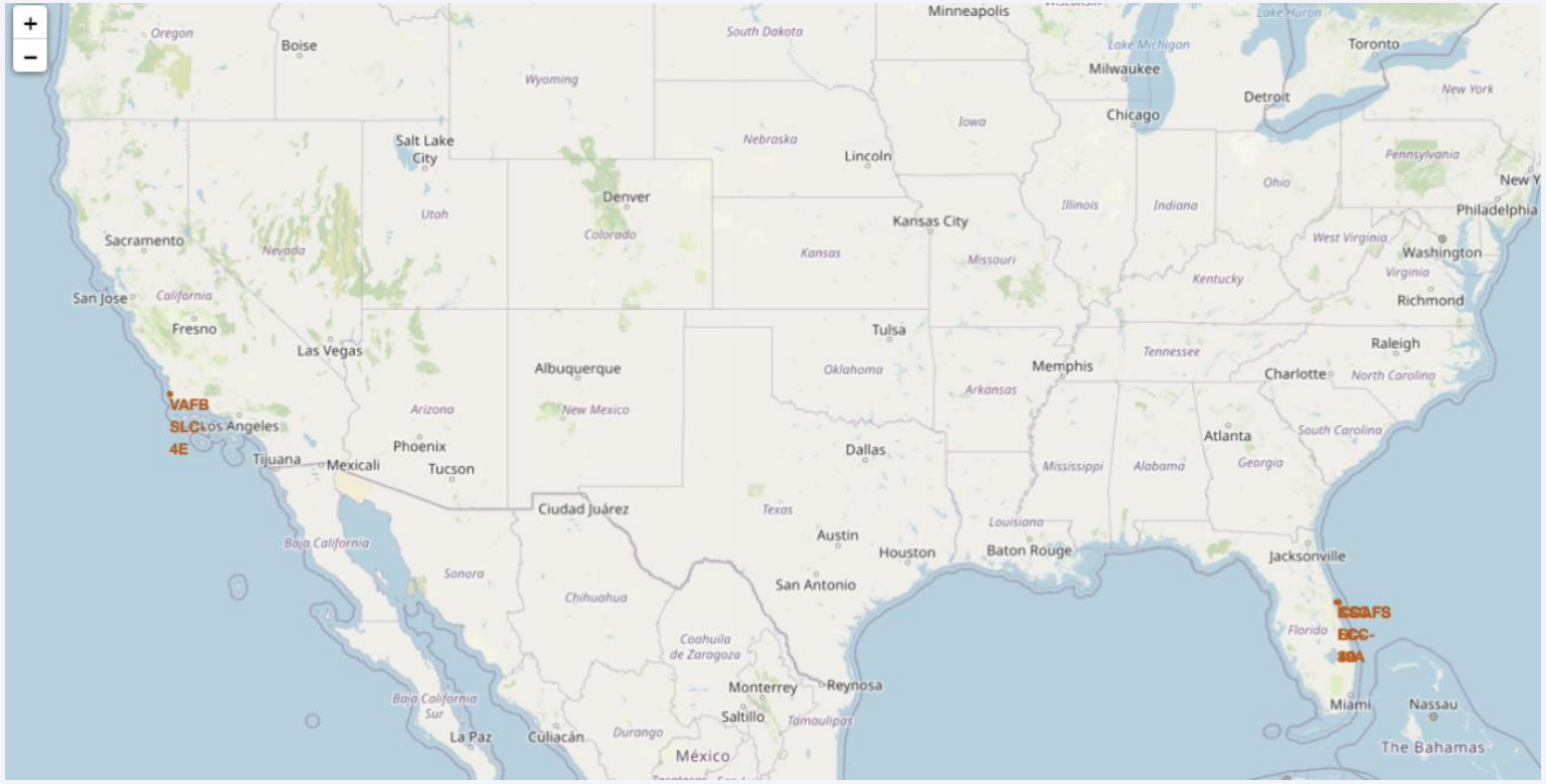
Filter table 總計:8

LANDING__OUTCOME	COUNT_LAUNCHES
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

Locations Of Launch Sites

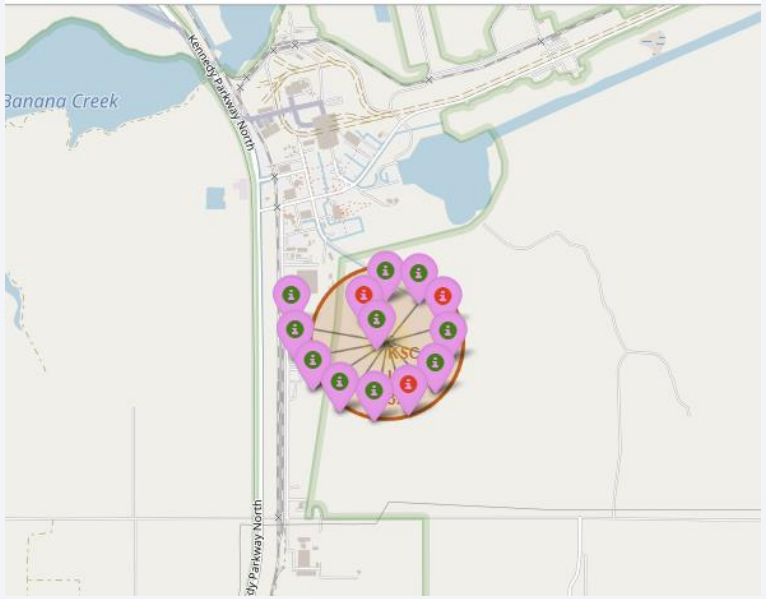
- We can observe that Launch Sites are **all near coastline**, and between **0-30 latitude.**



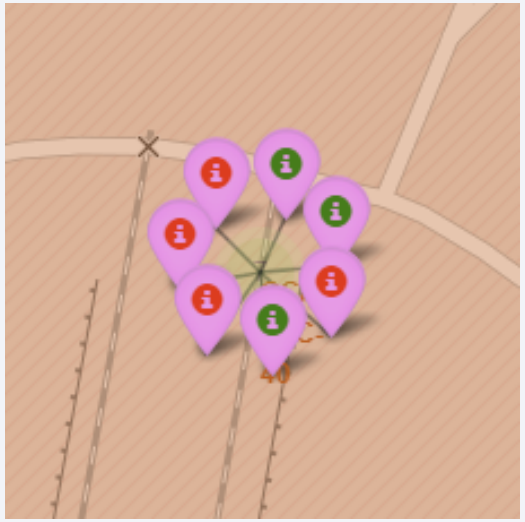
Launch Outcome with Color Labels

- Green Label : Success Launch
- Red Label : Fail Launch

✓ KSC LC-39A : 10 Success, 3 Fail.



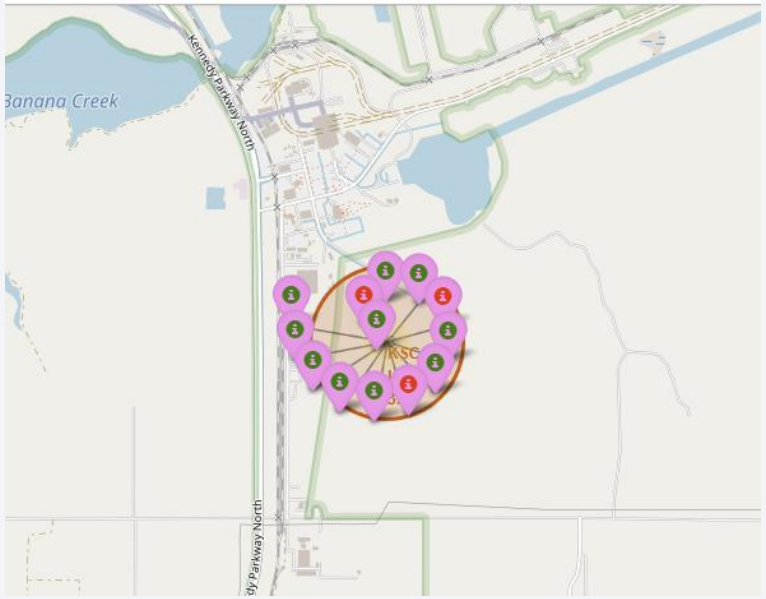
✓ CCAFS SLC-40: 3 Success, 4 Fail.



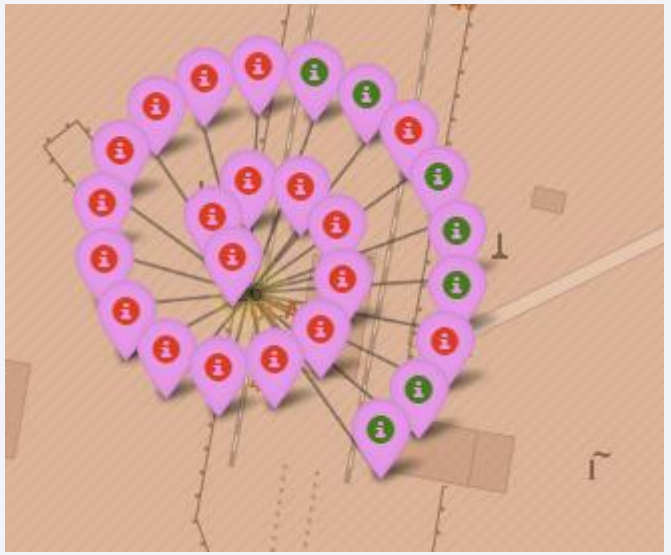
Launch Outcome with Color Labels

- Green Label : Success Launch
- Red Label : Fail Launch

✓ VAFB SLC-4E: 4 Success, 6 Fail.

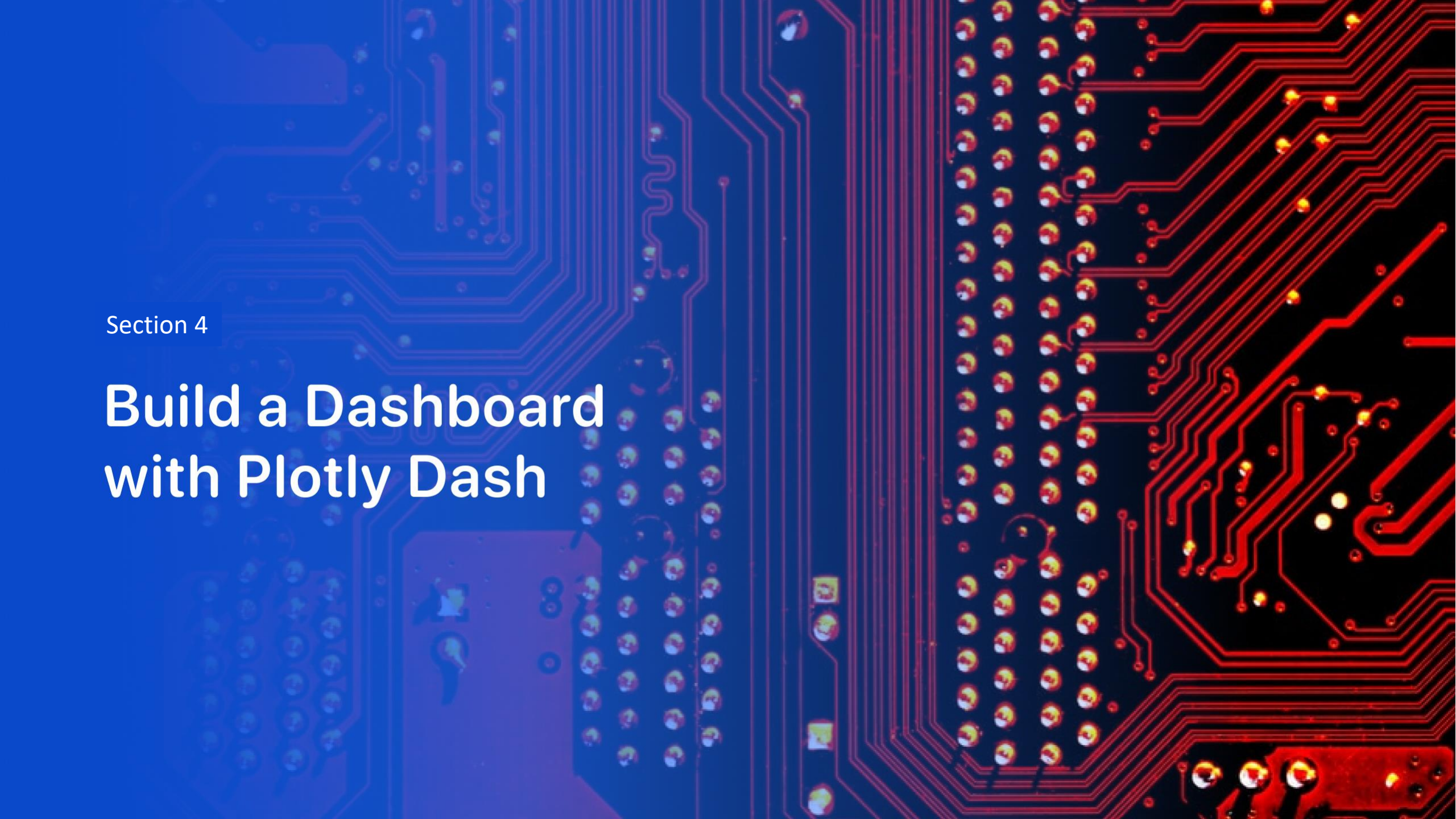


✓ CCAFS LC-40: 7 Success, 19 Fail.



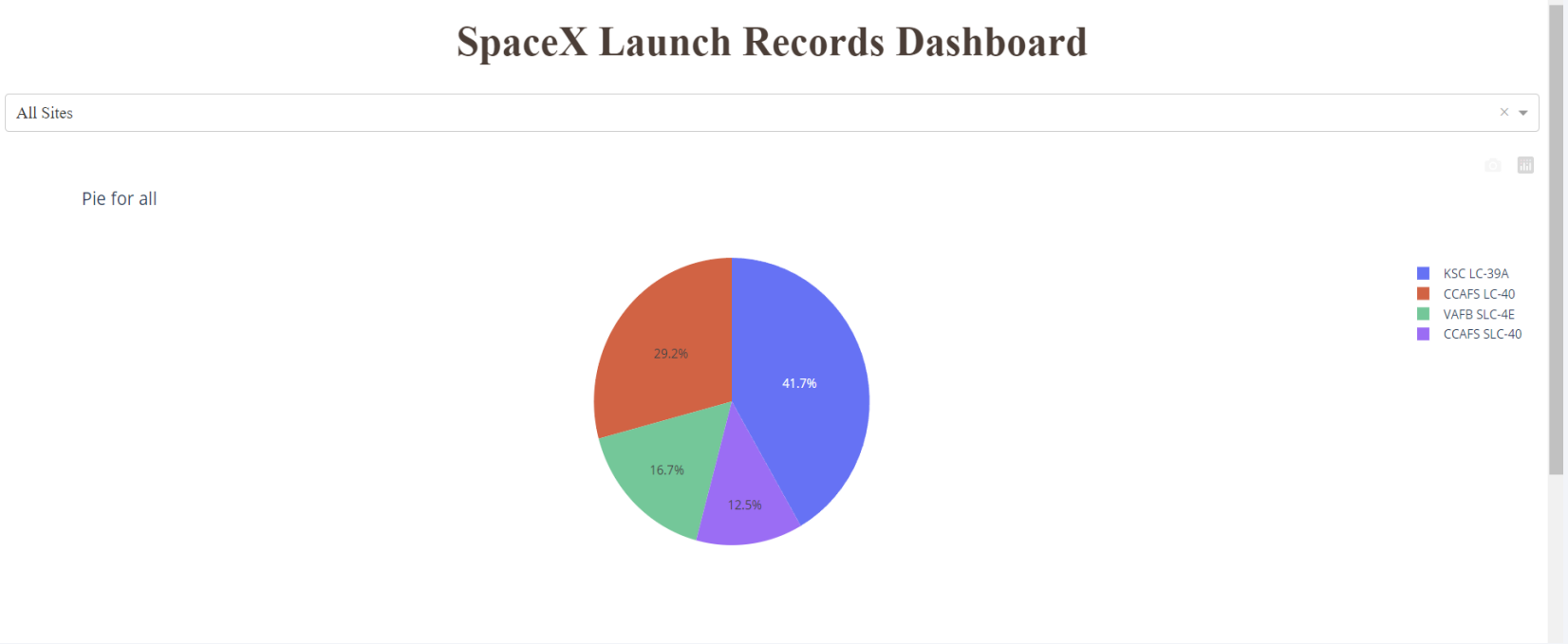
Launch Sites Proximities with Landmarks

- The distance of eastcoast Launch Site to its **nearest coastline** is **0.86KM.**


Section 4
Build a Dashboard with Plotly Dash

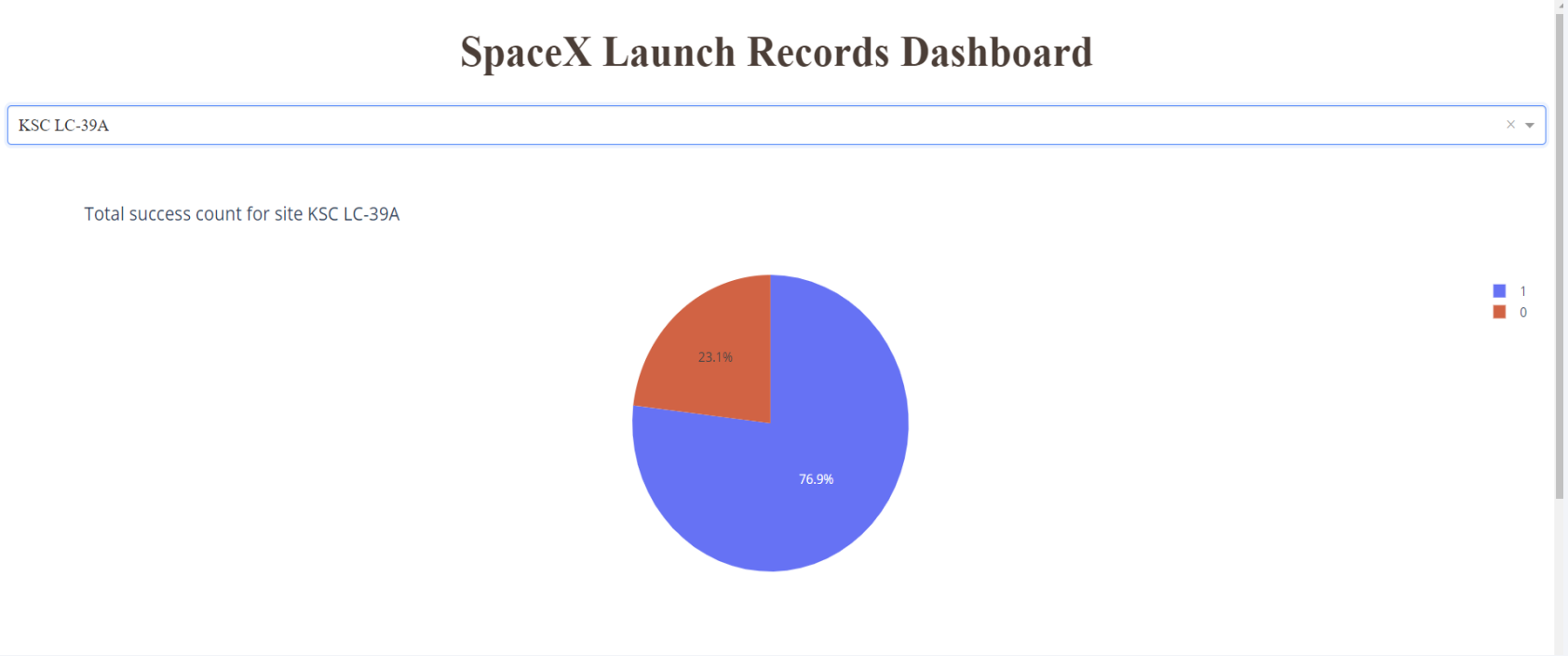
Total Success Count For All Launch Sites

- Launch site with Highest success count —— **KSC LC-39A**
- Launch site with Lowest success count —— **CCAFS SLC-40**



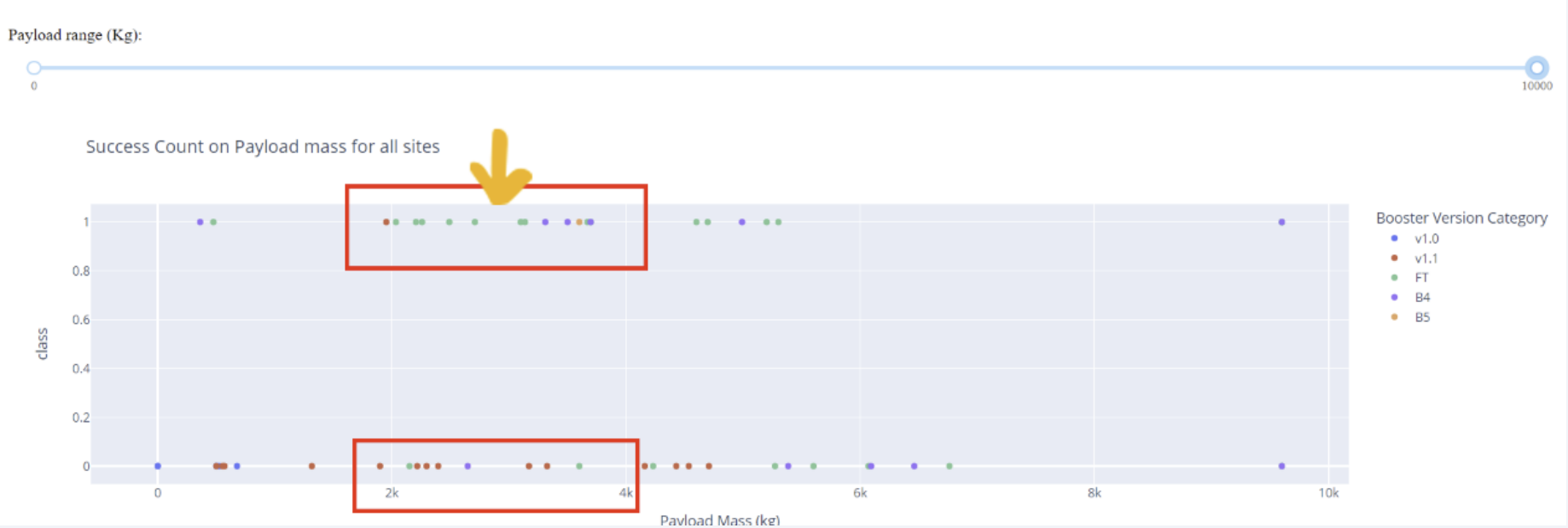
Launch Site with Highest Success Ratio

- KSC LC-39A has the Highest success rate of **76.9%**.



Payload vs Launch Outcome (All Launch Site)

- In Payload range **2K-4K,** we have the Highest success rate.



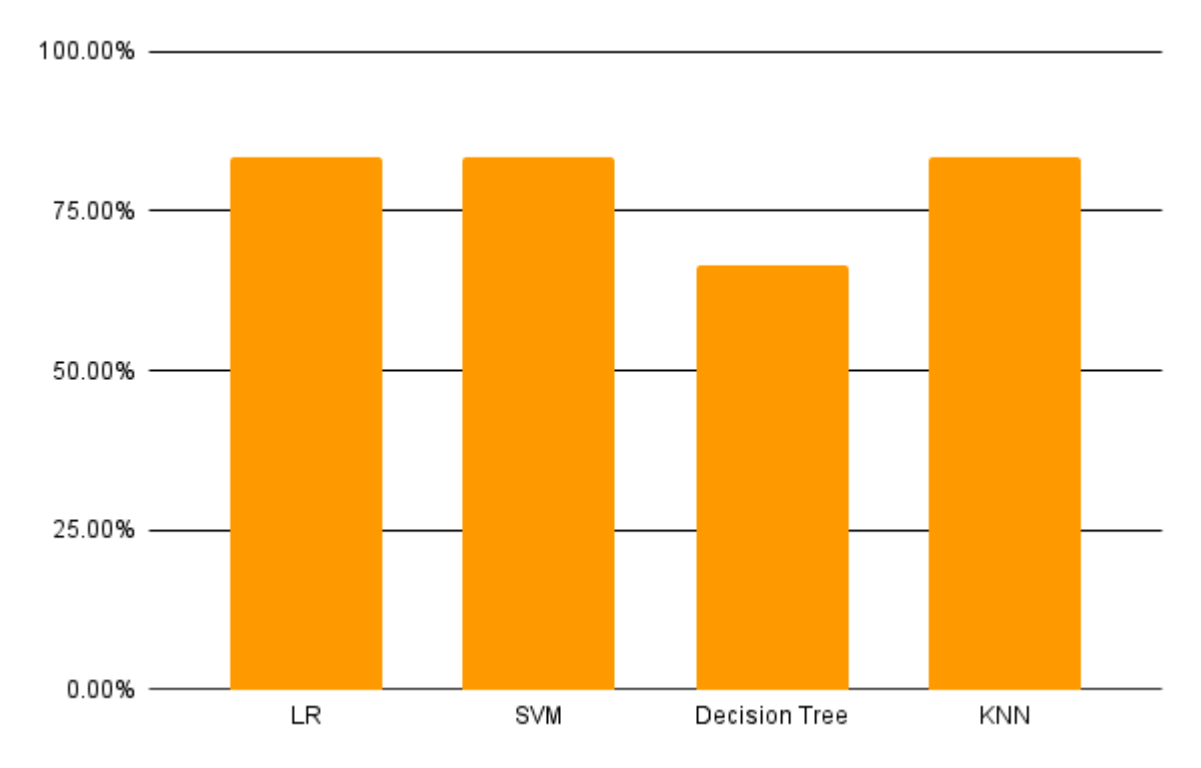
Payload vs Launch Outcome (All Launch Site)

- In **Higher Payload range**, the success rate decrease.

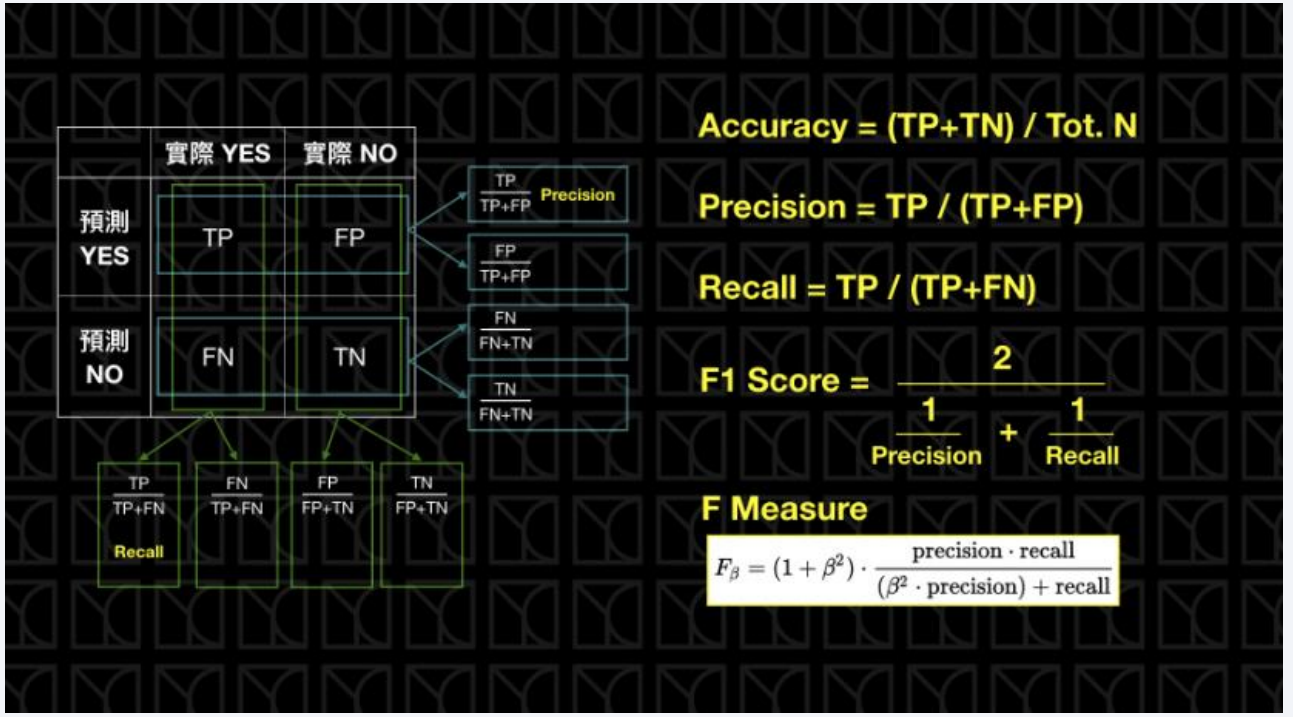
Section 5

Predictive Analysis (Classification)

Classification Accuracy



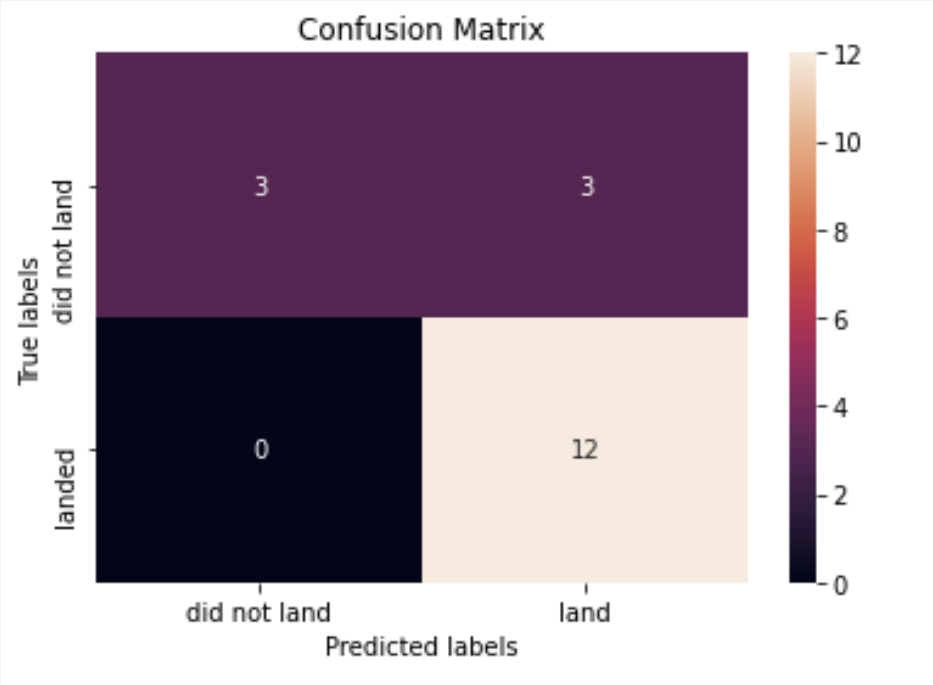
Confusion Matrix Info



Confusion Matrix

The confusion Matrix for SVM, Logistic Regression, KNN shows that :

- Recall = TP / (TP+FN) = 100%
- Precision = TP / (TP+FP) = 50%
- F1-score = 0.666



Conclusions

- In Payload range **2K-4K,** we have the Highest success rate, which implies that lower payload range performs better than higher range
- The success rate increase as the Flight Number increase
- Orbit type with Highest success rate : ES-L1, GEO, HEO, SSO
- Launch Sites are **all near coastline**, and between 0-30 latitude
- KSC LC-39A had the most successful launches among all sites
- The SVM, Logistic Regression, KNN are the best models for prediction with 83.3% accuracy

Appendix

- https://ithelp.ithome.com.tw/articles/10220716

Thank you!