Winning Space Race with Data Science

Yap Jin Hong
02.11.2022

Outline

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

Executive Summary

- Following methods are used to analyze the data
  - Data Collection using SpaceX API and Web scrapping on Wikipedia.
  - Exploratory Data Analysis (EDA) using Data Visualization and SQL as well as interactive map.
  - Building a Dashboard.
  - Machine Learning Prediction.
- Summary of all results
  - Possible to collect valuable results from the web.
  - Able to identify important features using EDA and Dashboard.
  - Able to identify which machine learning model and its parameter is best to predict the successful landing.

Introduction

- The objective is to evaluate the important element for the new company Space Y to compete with Space X.
- Questions to be answered
  - Which variable affects the success of first stage landing?
  - Where is the best place for landing?
  - Which algorithm is the best to predict the success of the first stage landing?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX REST API and Web Scrapping from Wikipedia.
- Perform data wrangling
  - Filter and dealing with missing data then create a landing outcome label.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data collected was processed and evaluted using four different classification mdoels and choose the best model.

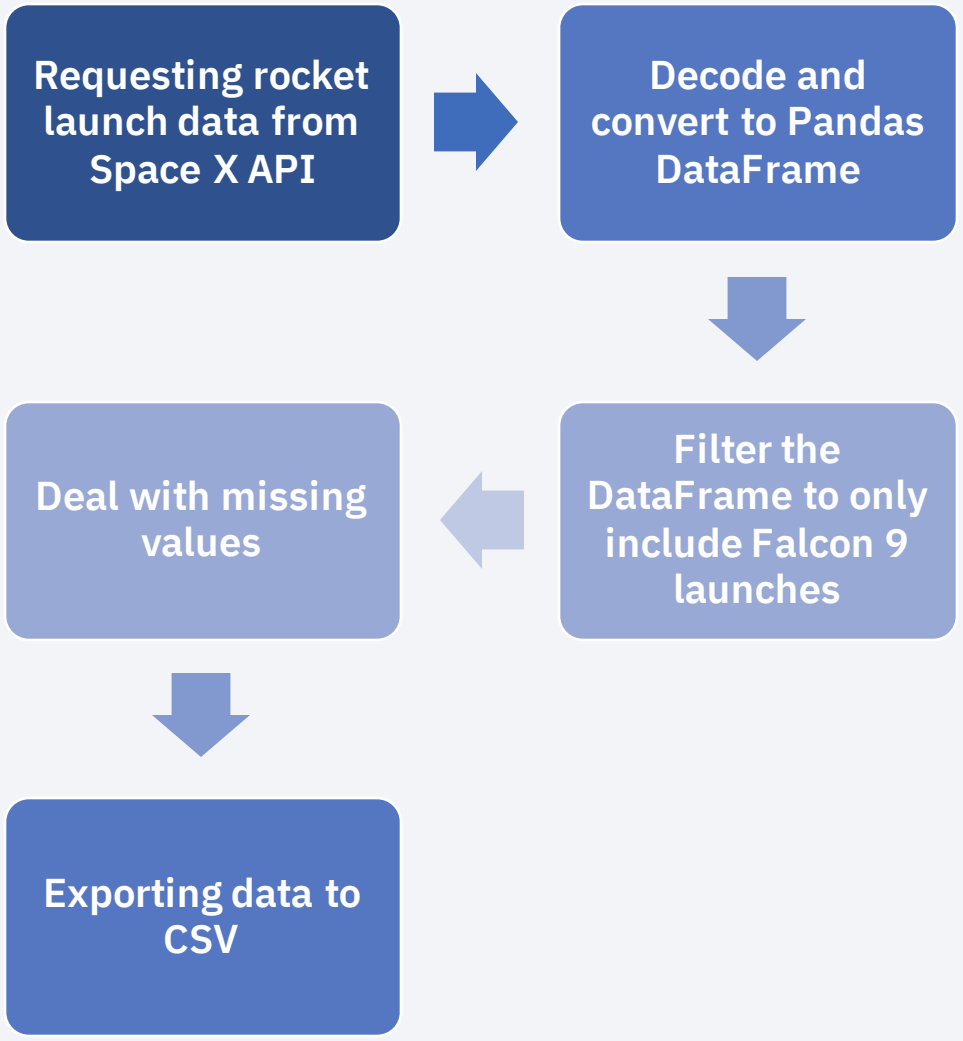
Data Collection

- The data of Space X can be obtained from:
  - Space X API (https://api.spacexdata.com/v4/rockets/)
  - Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Both sites are used to collect data so that we have more information about the launches for detailed analysis.

Data Collection – SpaceX API

- The data of Space X can be obtained from their public API, it can be downloaded and used.

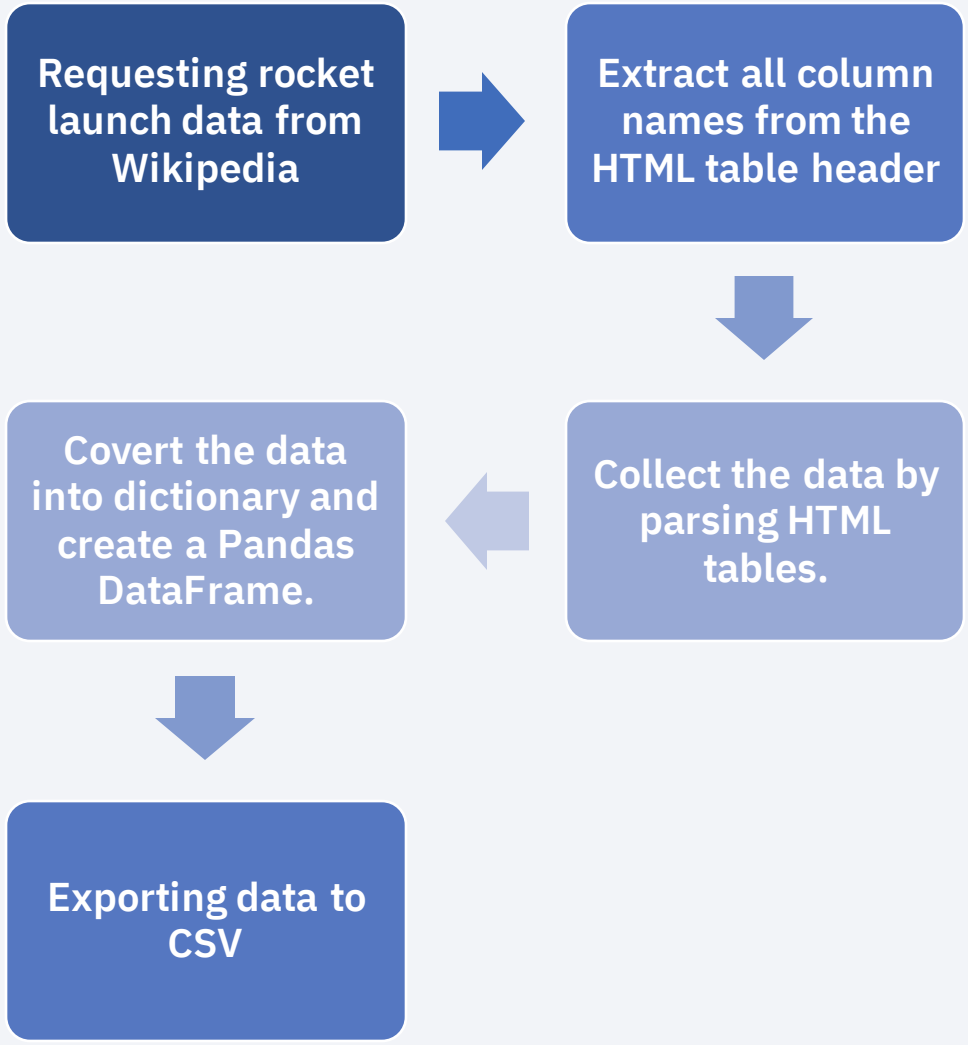
- GitHub URL : Data Collection using API



Data Collection - Scraping

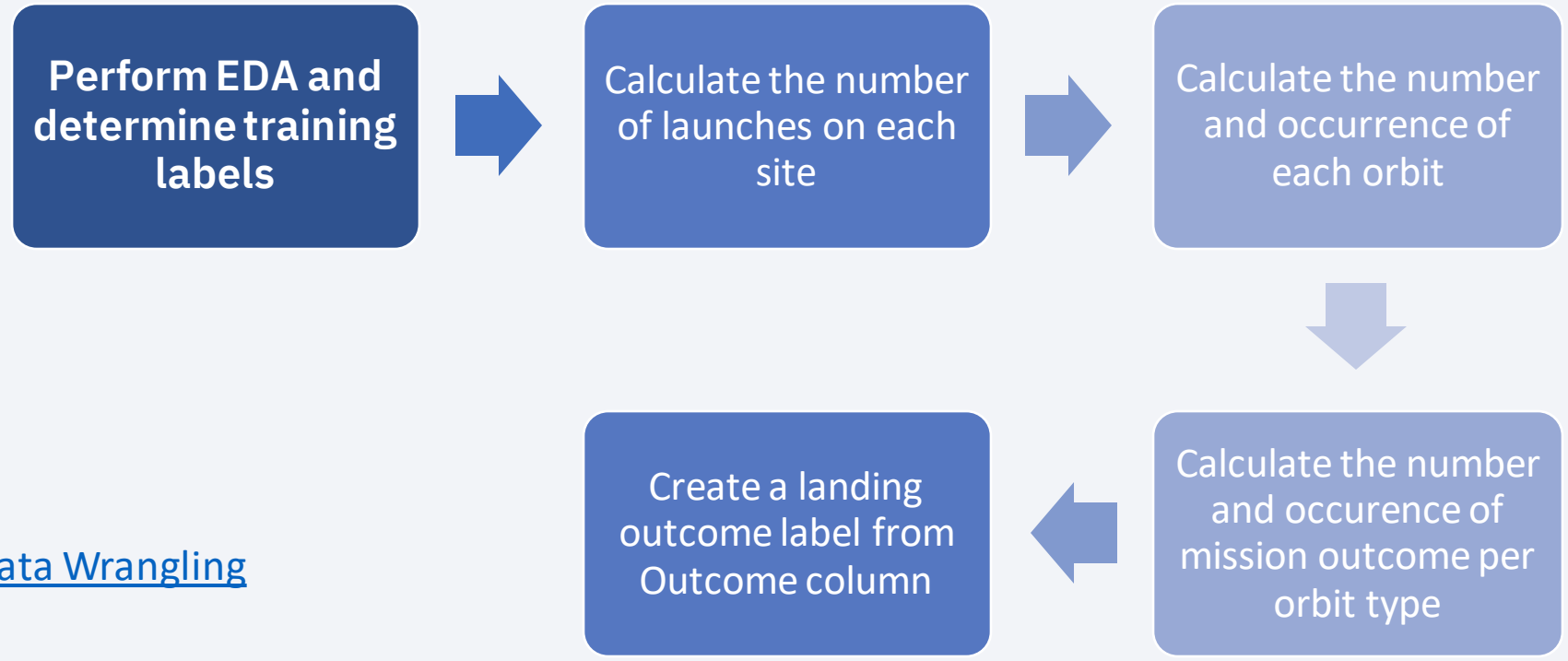
- The data of Space X can be obtained from Wikipedia, it can be downloaded and used.

- GitHub URL : Data Collections on Web Scrapping



Data Wrangling

- From the dataset, there are different outcome for both successful and unsuccessful landing.
- To further understand the dataset some features are calculated.
- The landing outcome label was created from Outcome column where '1' represent successful.



- GitHub URL : Data Wrangling

EDA with Data Visualization

- Charts are plotted to visualize and show the relationship between variables.
  - Payload Mass vs. Flight Number
  - Flight Number vs. Launch Sites
  - Payload Mass vs. Launch Sites
  - Success Rate of each Orbit
  - Flight Number vs. Orbit Type
  - Payload Mass vs. Orbit Type
  - Launch Success yearly trend

GitHub URL : EDA with Data Visualization

EDA with SQL

- **The following SQL queries are performed.**
  - The names of the unique launch sites in the space mission.
  - Top 5 records where launch sites begin with the string 'CCA'.
  - Total payload mass carried by boosters launched by NASA (CRS).
  - Average payload mass carried by booster version F9 v1.1.
  - Date when the first successful landing outcome in ground pad was achieved.
  - Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - Total number of successful and failure mission outcomes.
  - Names of the booster versions which have carried the maximum payload mass.
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.
  - The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

GitHub URL : EDA with SQL

Build an Interactive Map with Folium

- **Markers of all Lauch Sites**
  - Added Marker with Circle, Popup Label and Text Label on
    - NASA Johnson Space Center.
    - All Launch Sites.
- **Colored Markers of all launch outcomes for each launch site:**
  - Added colored marker of success (Green) and failed (Red) launches to identify which launch sites have relatively high success rates.
- **Distances between a Launch Site to its proximities:**
  - Added colored lines to show distances between launch site to its proximities like Railway, Highway, Coastline and Closest City.

GitHub URL : Interactive Visual Analytics with Folium.

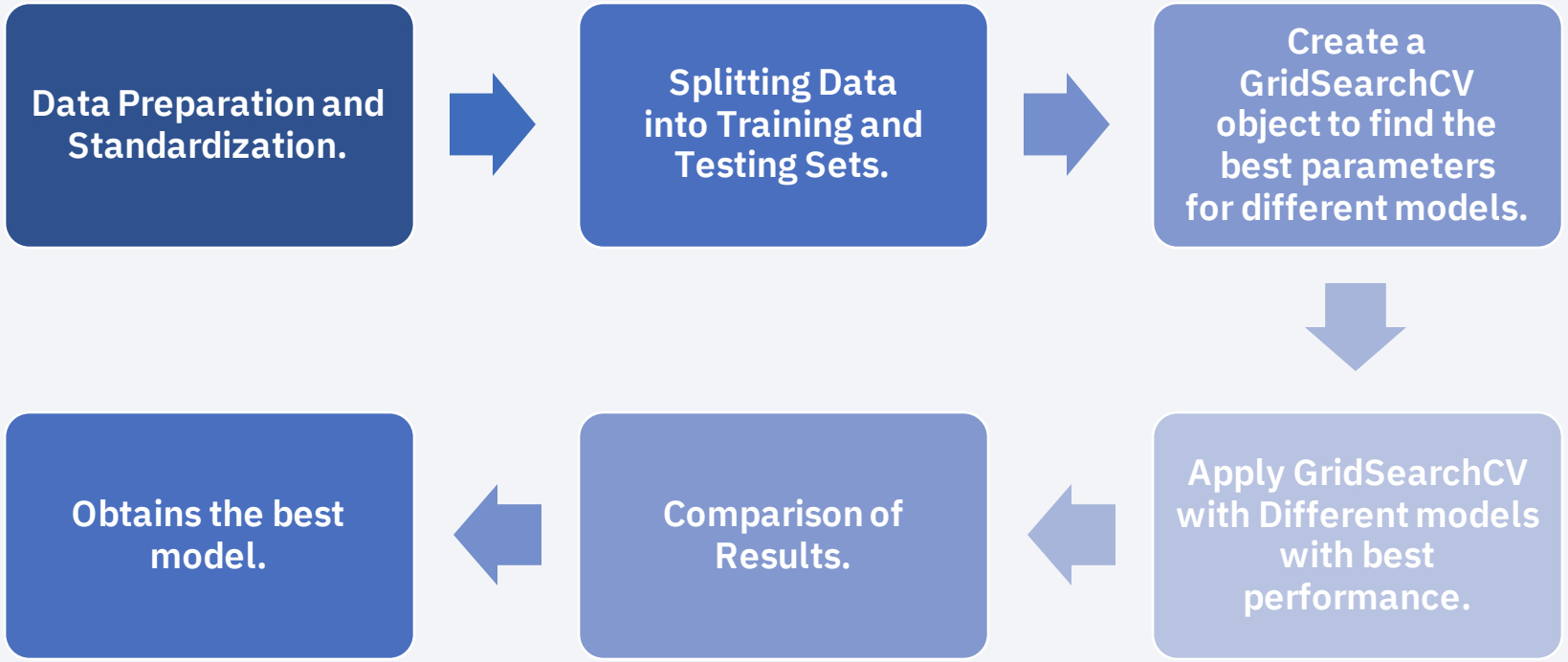
Build a Dashboard with Plotly Dash

- **Launch Sites Dropdown List:**
  - Added a dropdown list to enable Launch Site selection.
- **Pie Chart showing Success Launches (All sites / Certain Site):**
  - Added a pie chart to show the total successful launches for all sites and selected launch site.
- **Slider of Payload Mass Range:**
  - Added a slider to select Payload range.
- **Scatter Chart of Payload Mass vs. Success Rate for different Booster Version:**
  - Added a scatter chart to show the correlation between Payload Mass and Success Launches.

GitHub URL : SpaceX Dash App

Predictive Analysis (Classification)

- Four classification models were compared: Logistic Regression, Support Vector Machine, Decision Tree and K-Nearest Neighbors.



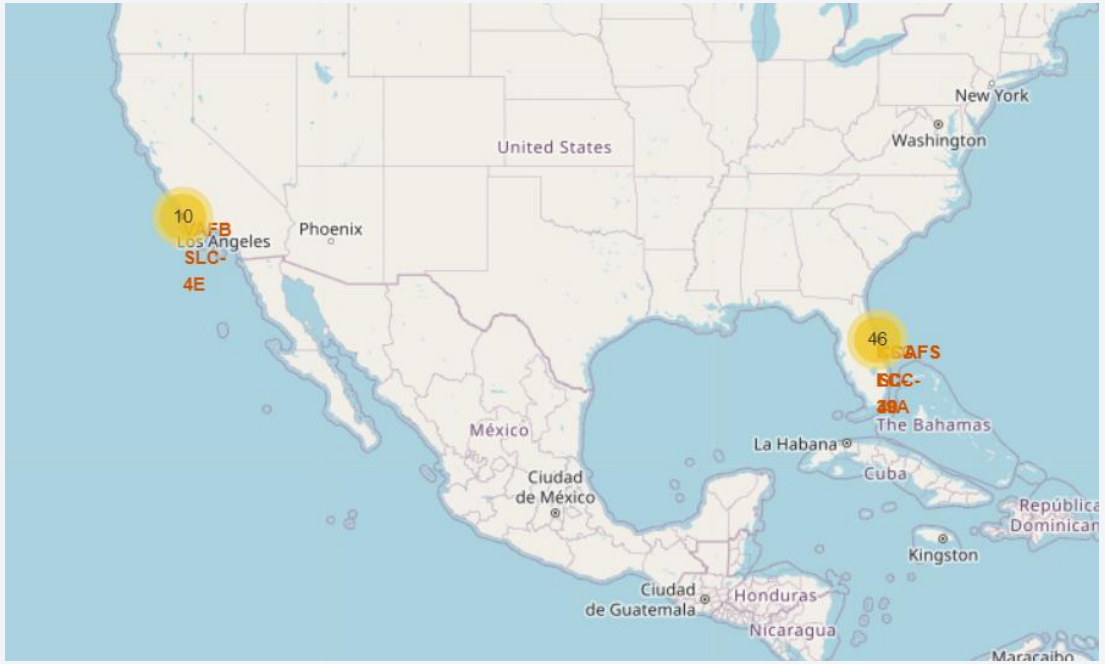
GitHub URL : Predictive Analysis/

Results

- Exploratory data analysis results:
  - Space X uses 4 different launch sites.
  - The first launches were done by Space X itself and NASA
  - The average payload of F9 v1.1 booster is 2,928 kg.
  - The first success landing outcome happened in 2015.
  - Many Falcon 9 booster versions were successful at landing in drone ships havibn payload above the average.
  - Two booster versions failed at landing in drone ships in 2015 is the F9 v1.1 B1012 and F9 v1.1 B1015.
  - The number of landing outcomes became better as years passed.

Results

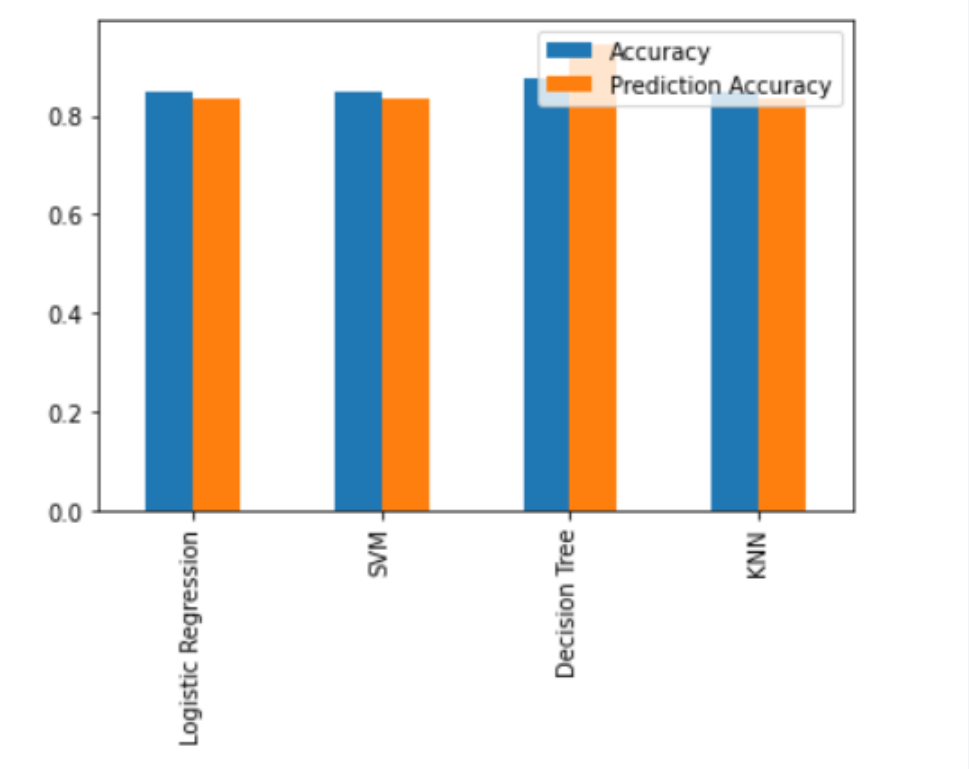
- Interactive analytics we able to identify that all the launch sites are located at safety places, near to the sea and have a good logistic infrastructure around.
- Most launches happens at east coast launch site.



Results

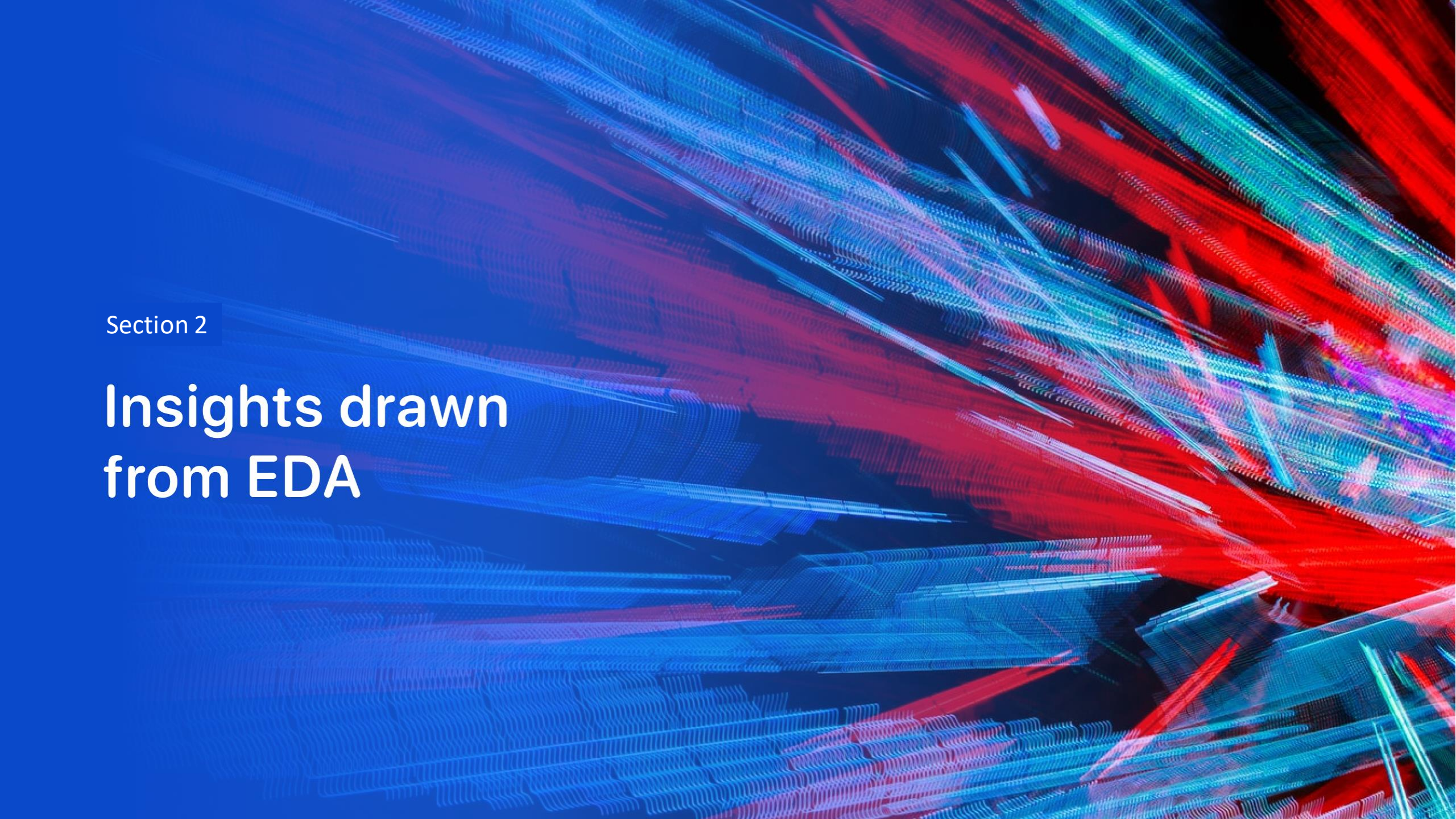
- Predictive analysis shows that the Decision Tree Classifier is the best model to predict successful landing having train accuracy of 87.5% and test accuracy of 94.4%.

	Accuracy	Prediction Accuracy
Logistic Regression	0.846429	0.833333
SVM	0.848214	0.833333
Decision Tree	0.875000	0.944444
KNN	0.848214	0.833333

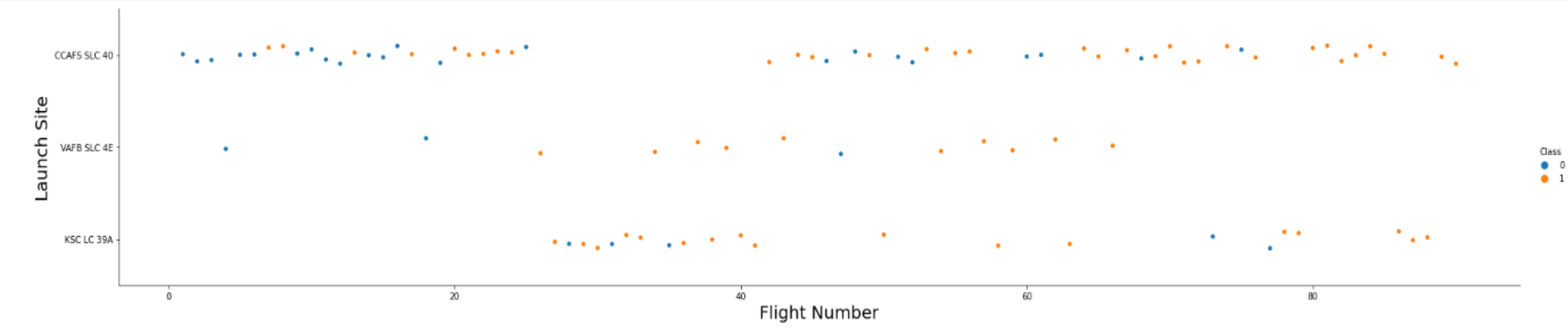


Section 2

Insights drawn from EDA

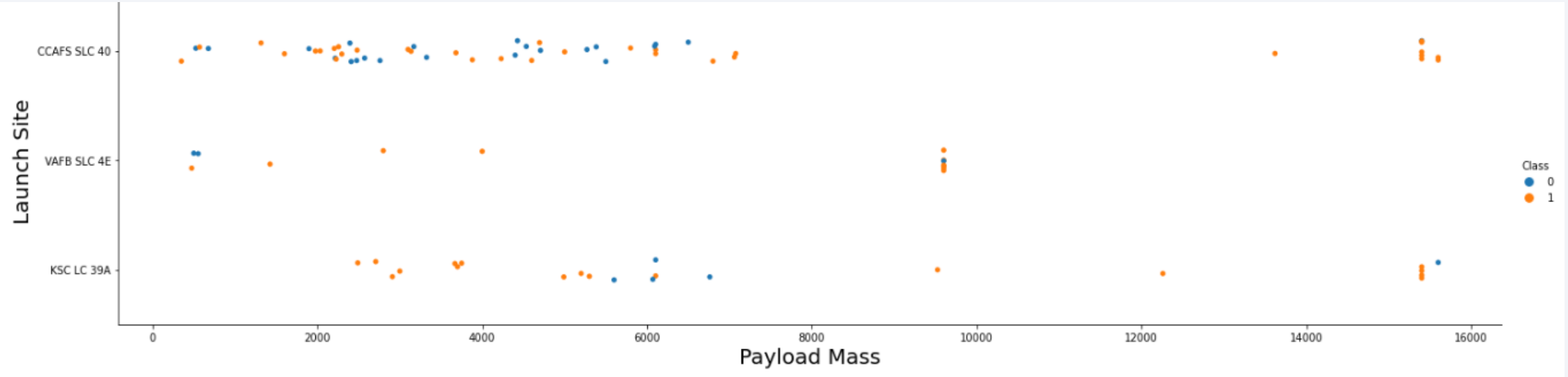


Flight Number vs. Launch Site



- The CCAFS SLC 40 launch site has about half of all launches.
- KSC LC-39A and VAFB SLC 4E have higher success rate.
- It can be assumed that the new launches has a higher success rate.

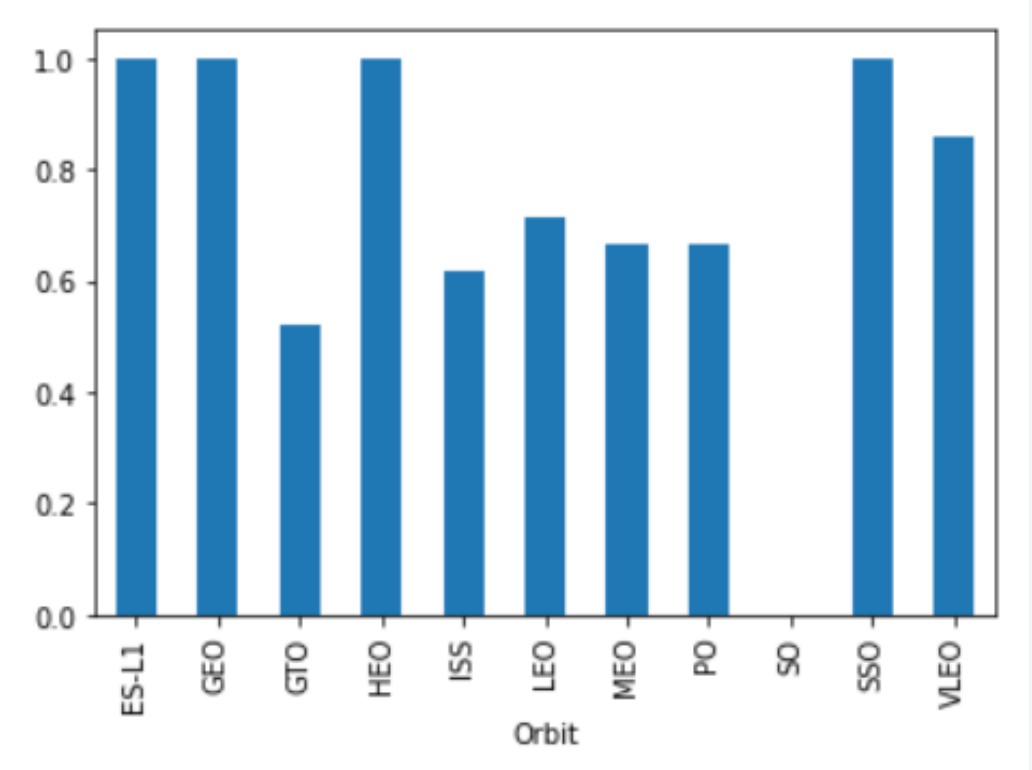
Payload vs. Launch Site



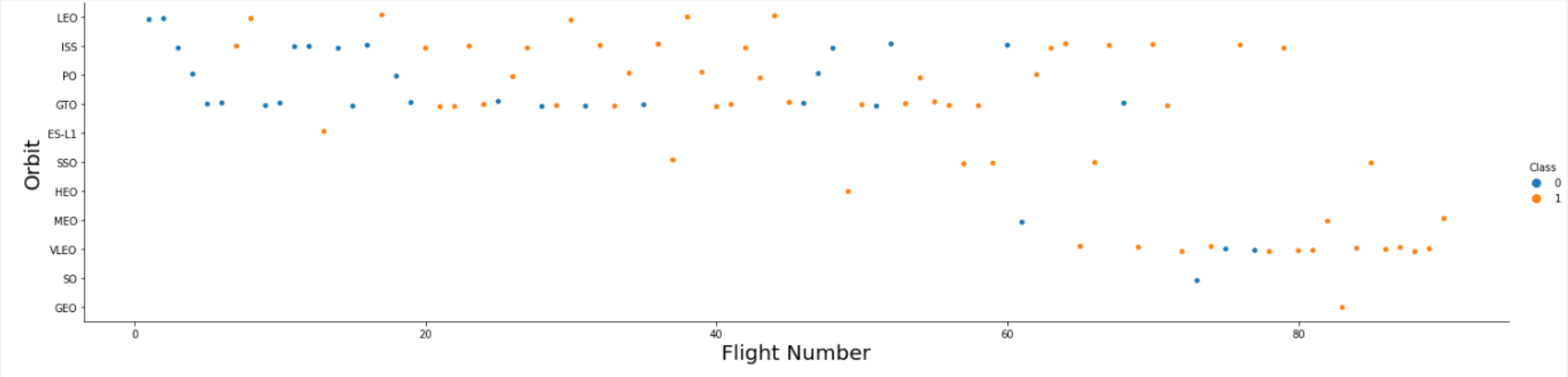
- Payload Mass over 9000kg have higher success rate.
- For every launch site, the higher the payload mass, the higher the success rate.
- KSC LC-39A have 100% success rate for payload mass less than 5000kg.

Success Rate vs. Orbit Type

- Orbit with highest success rate:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Orbit SO has a 0% success rate. Show the screenshot of the scatter plot with explanations
- VLEO have the success rate of 90% and LEO have 70%.

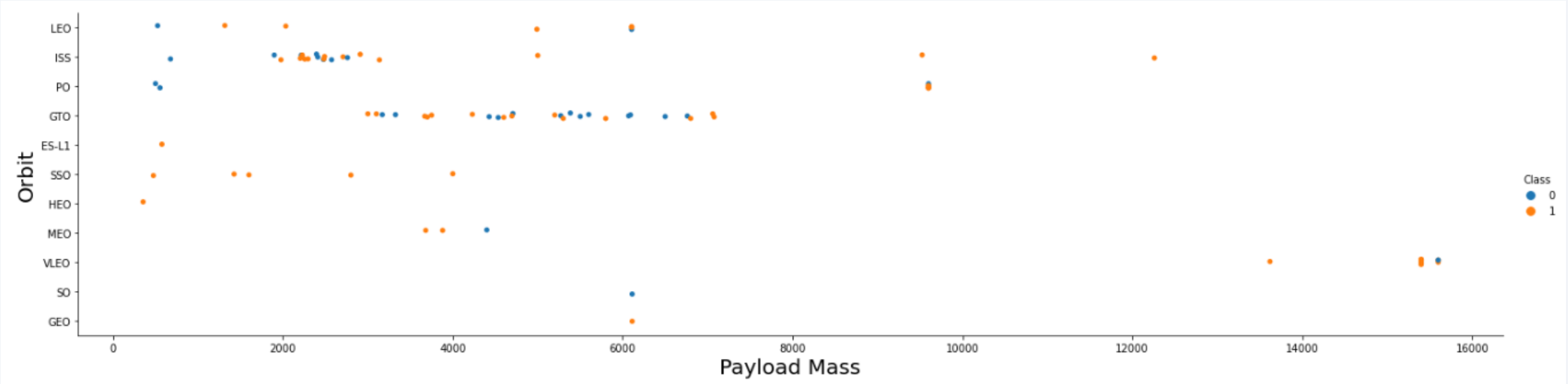


Flight Number vs. Orbit Type



- VLEO orbit seems to have more focus recently due to increase its frequency.
- Overall, according to the plot, there's no relationship between flight number when in GTO orbit.

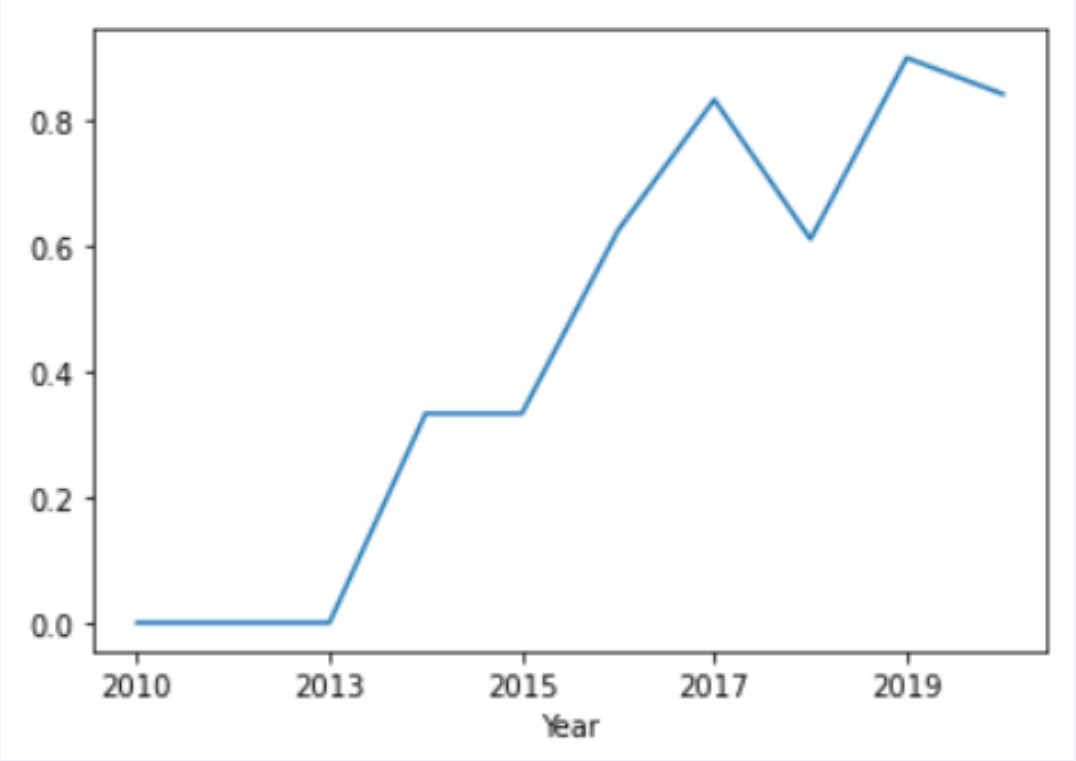
Payload vs. Orbit Type



- 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

- Success rate started to increase from year 2013 to year 2020 maybe to breakthrough of technologies.



All Launch Site Names

- According to data queries from SQL, the names of the launch sites is:

```
%sql SELECT DISTINCT launch_site FROM SPACEXTBL
```

```
 * ibm_db_sa://gyk98721:***@2d46b6b4-cbf6-40eb-bb
Done.
```

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

- They are obtained by selecting the distinct occurrences of unique "launch_site" value.

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE launch_site like 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://gyk98721:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- Displaying 5 records where launch sites begin with `CCA`

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE CUSTOMER like '%CRS%';
```

```
 * ibm_db_sa://gyk98721:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.
Done.
```

total_payload
48213

- Displaying the total payload carried by boosters from NASA

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://gyk98721:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.
Done.
```

average_payload
2928

- Displaying the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
%sql SELECT MIN(DATE) AS FIRST_SUCCESS FROM SPACEXTBL WHERE Landing_Outcome ='Success (ground pad)';
```

```
 * ibm_db_sa://gyk98721:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdom
Done.
```

first_success
2015-12-22

- Display the dates of the first successful landing outcome on ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE payload_mass__kg_ between 4000 and 6000 AND LANDING_OUTCOME ='Success (drone ship)';
```

```
 * ibm_db_sa://gyk98721:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

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

- Display the list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(*) FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://gyk98721:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1
Done.
```

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

- Display the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTBL)
```

```
 * ibm_db_sa://gyk98721:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- Display the list the names of the booster which have carried the maximum payload mass

2015 Launch Records

```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Failure (drone ship)' AND DATE_PART('YEAR', DATE) = 2015
```

```
 * ibm_db_sa://gyk98721:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- Display the list of failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

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

```
 * ibm_db_sa://gyk98721:***@2d46b6b4-cbf6-40eb-bbce-6251e6ba0300.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32328/bludb
Done.
```

landing_outcome	qty
No attempt	10
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

- Rank 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


Section 3
Launch Sites
Proximities Analysis

All Launch Site

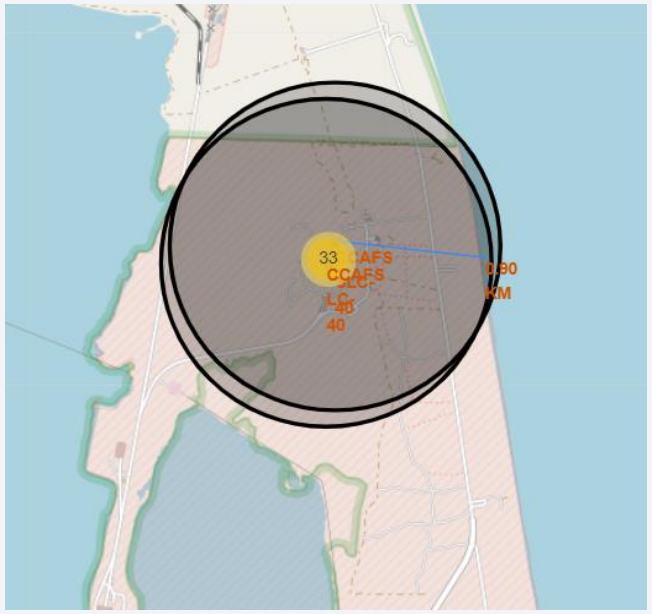
- Launches sites are near to the coast, while launching rockets it minimizes the risk of having explosion or any debris dropping on people.

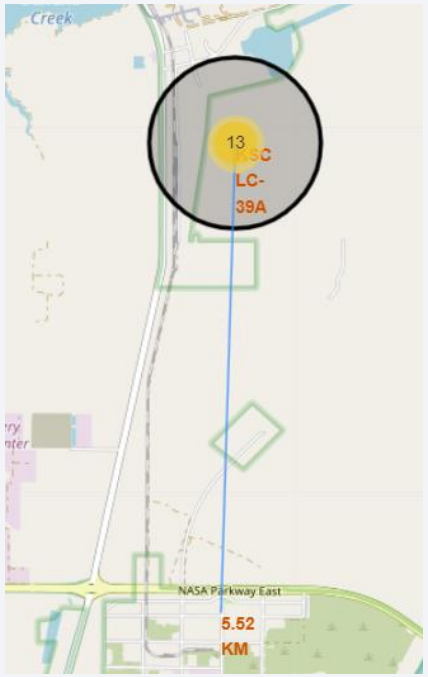
Lauch Outcome by Sites

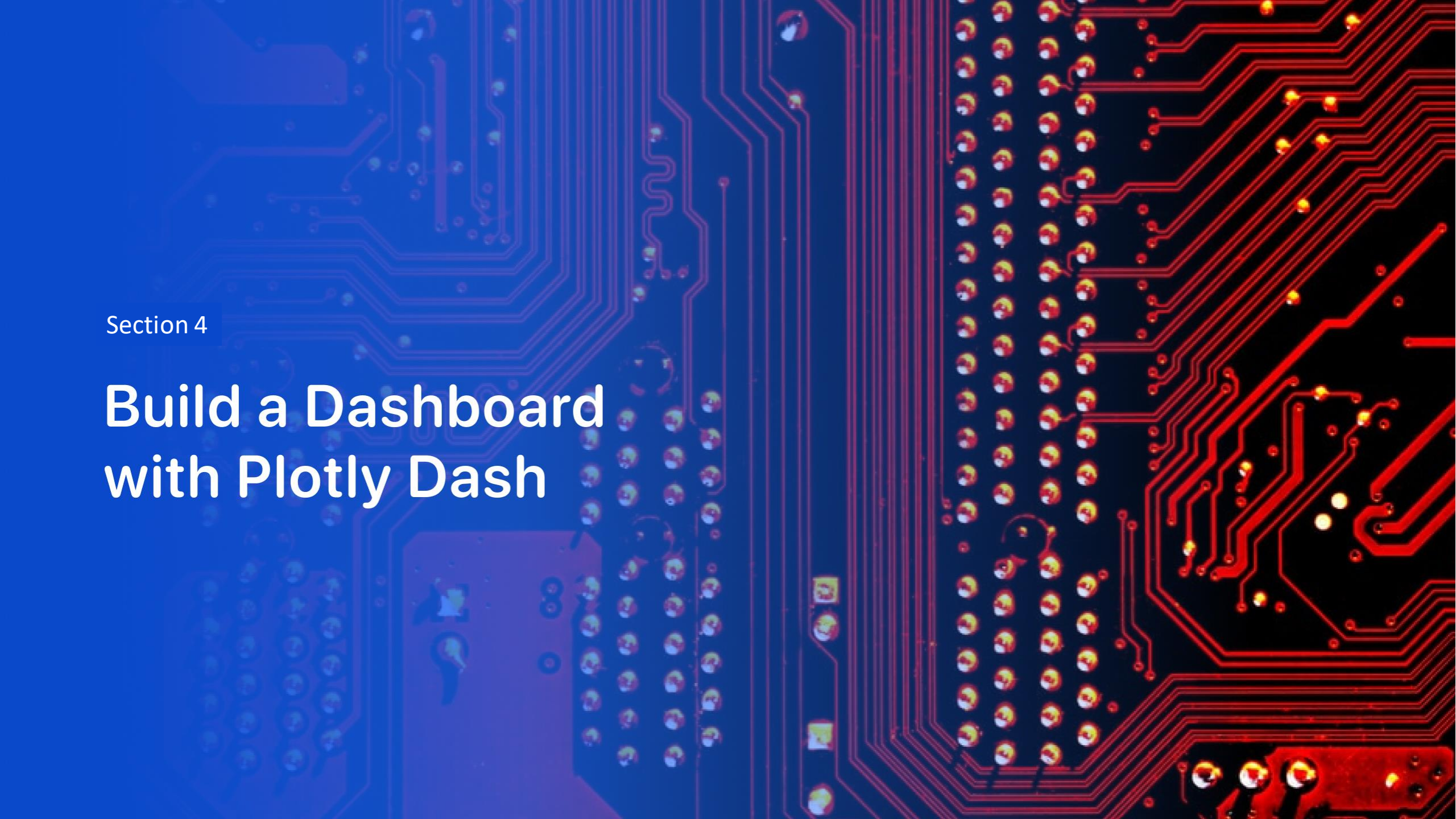
- Example of KSC-LC-39A launch site outcome.
- Green marker indicates successful launch while Red marker indicates failed launch.

Safety and Logistic

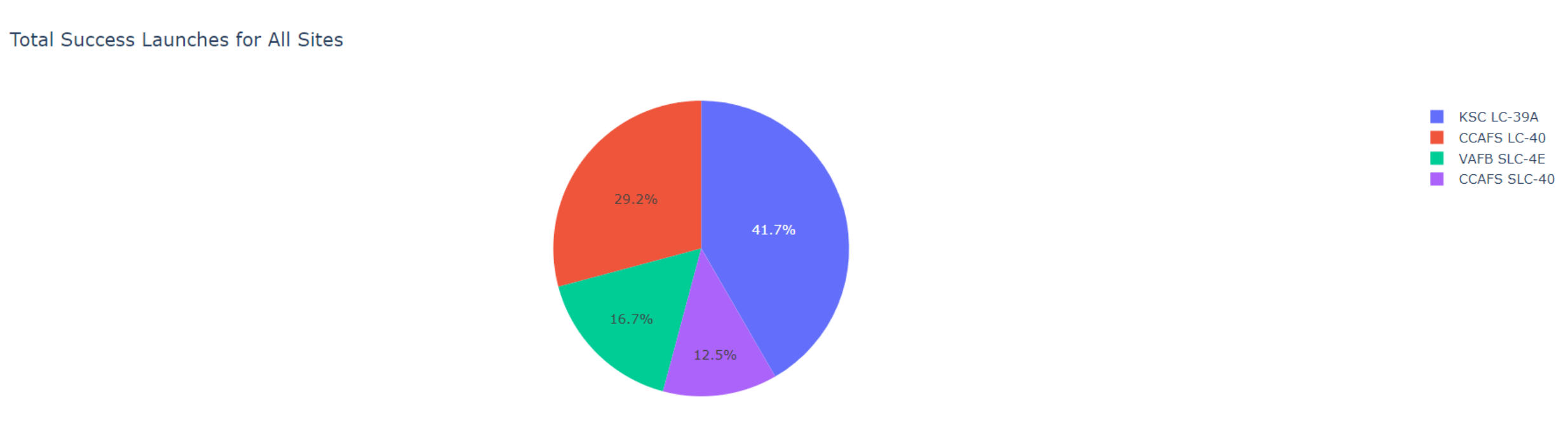
- From the left, we can see that CCAFS SLC-40 launch site is only 0.90km from the coasts. And it's far from any town.
- While on the right, KSC-LC-39A launch site is 5.52km from its nearest town and it have railway and roads.





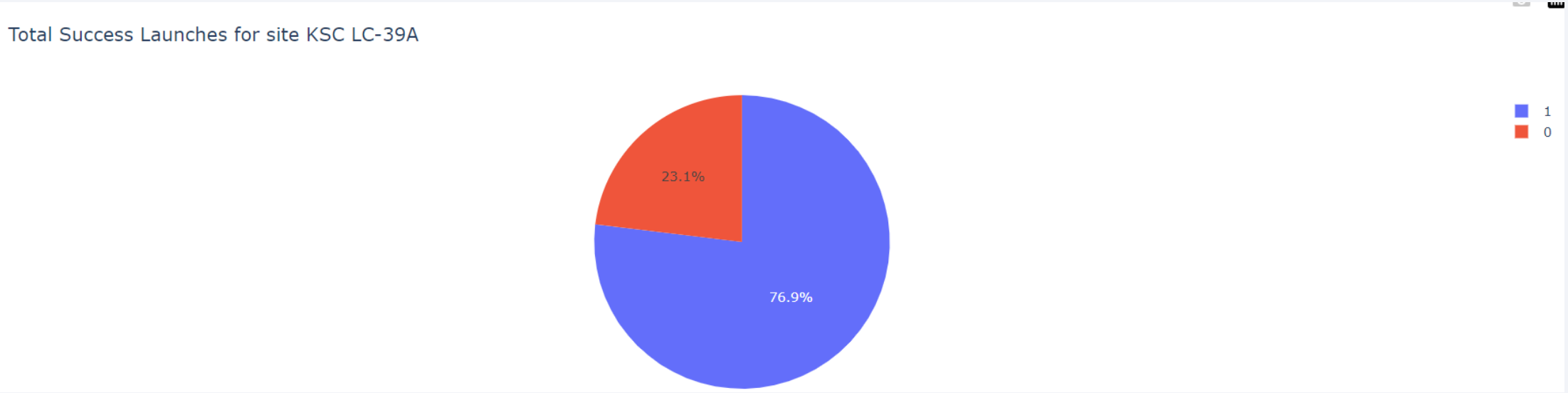

Section 4
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



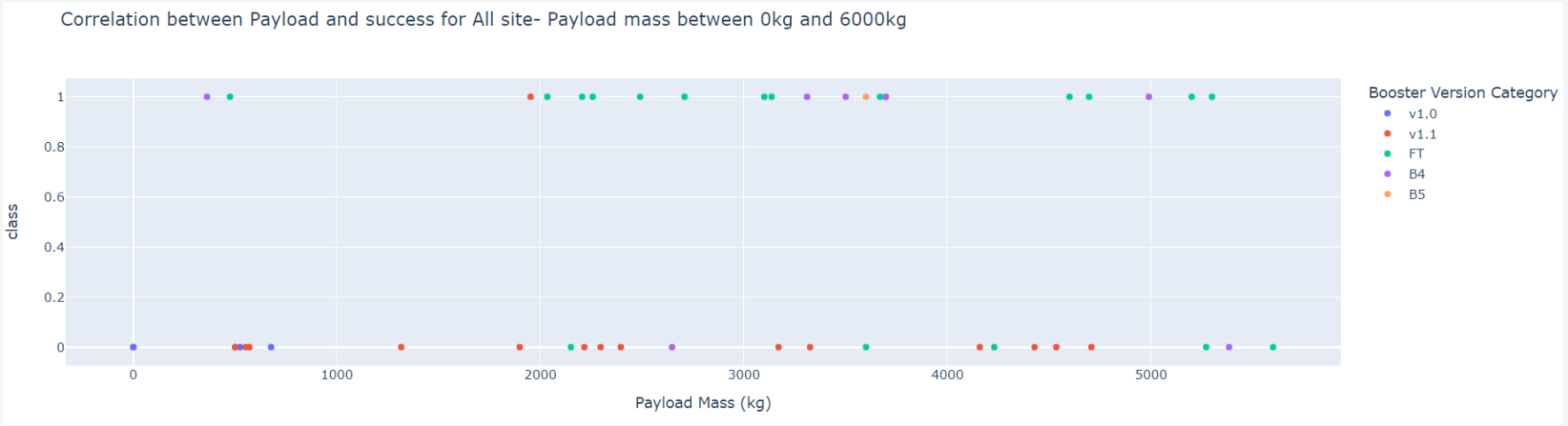
- The chart shows that from all the sites, KSC LC-39A has the most successful launches.

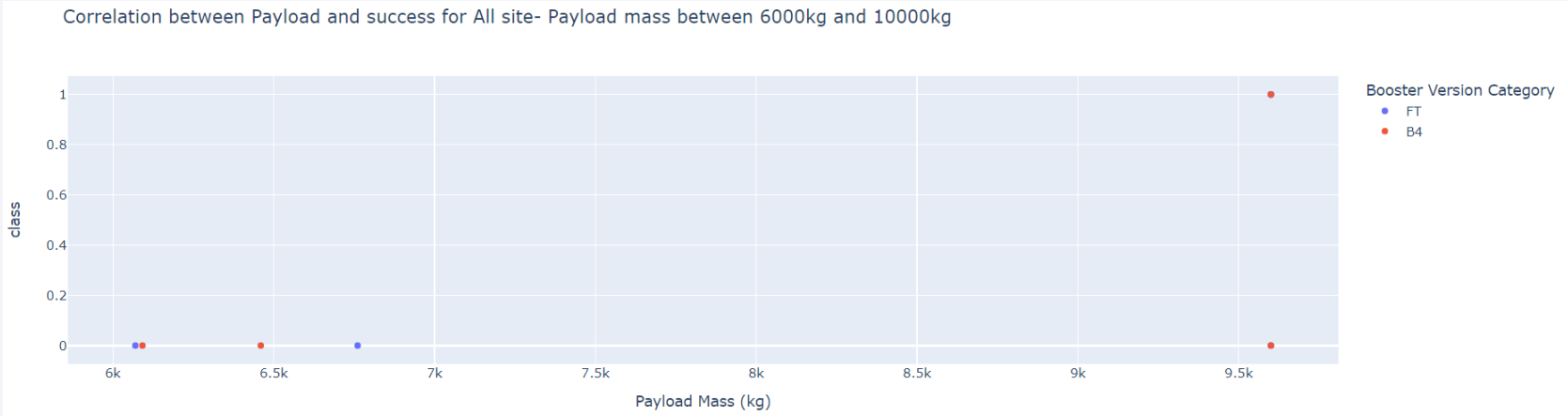
Launch Site with Highest Launch Success Ratio



- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful launch and only 3 failed launch.

Payload vs. Launch Outcome





- Payload between 2000kg to 6000kg have highest success launch.

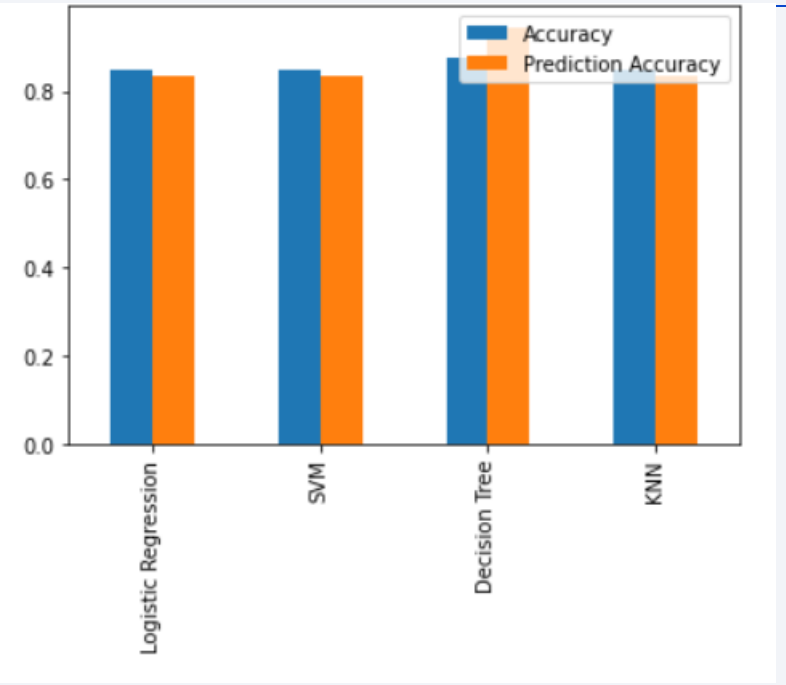
Section 5

Predictive Analysis (Classification)

Classification Accuracy

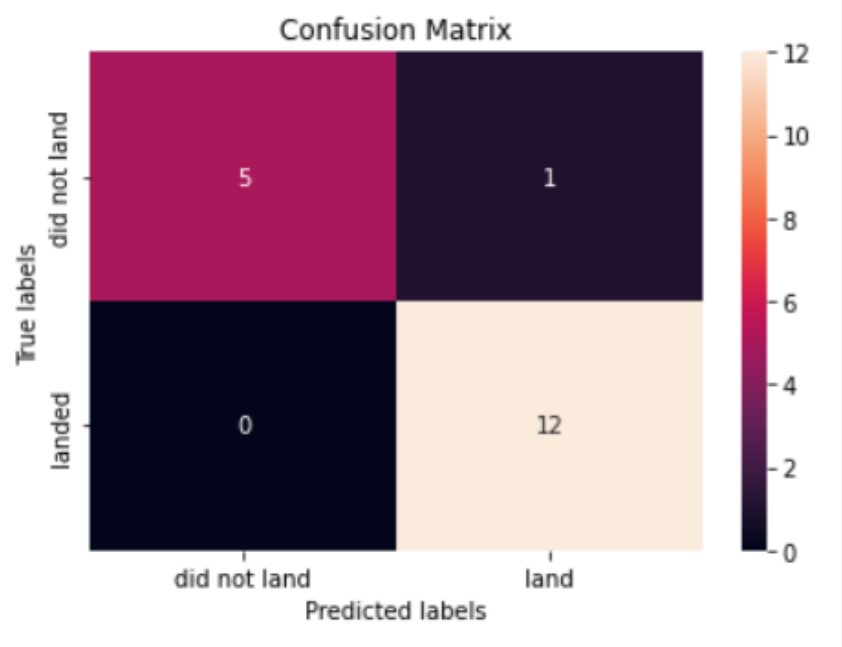
- Four classification model are tested, and their accuracy results are plotted.

- Decision Tree model has the highest classification accuracy with accuracy over 87%.



	Accuracy	Prediction Accuracy
Logistic Regression	0.846429	0.833333
SVM	0.848214	0.833333
Decision Tree	0.875000	0.944444
KNN	0.848214	0.833333

Confusion Matrix



- The confusion matrix of the Decision Tree shows high values True Positive and True Negative.

Conclusions

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results compare to larger payload mass.
- The success rate of launches increases over the years.
- KSC KC-39A has the highest success rate of the launches of all site.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Appendix

Thank you!