IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alireza Samimi
05.02.2025

Outline

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

Executive Summary

- Summary of methodologies

✓ Data Collection through API

✓ Data Collection with Web Scraping

✓ Data Wrangling

✓ Exploratory Data Analysis with SQL

✓ Exploratory Data Analysis with Data Visualization

✓ Interactive Visual Analytics with Folium

✓ Machine Learning Prediction

- Summary of all results

✓ Exploratory Data Analysis result

✓ Interactive analytics in screenshots

✓ Predictive Analytics result

Introduction

- Project background and context

Space X 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 Space X 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 space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- Problems you want to find answers
- What factors determine if the rocket will land successfully?
- The interaction amongst various features that determine the success rate of a successful landing
- What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology

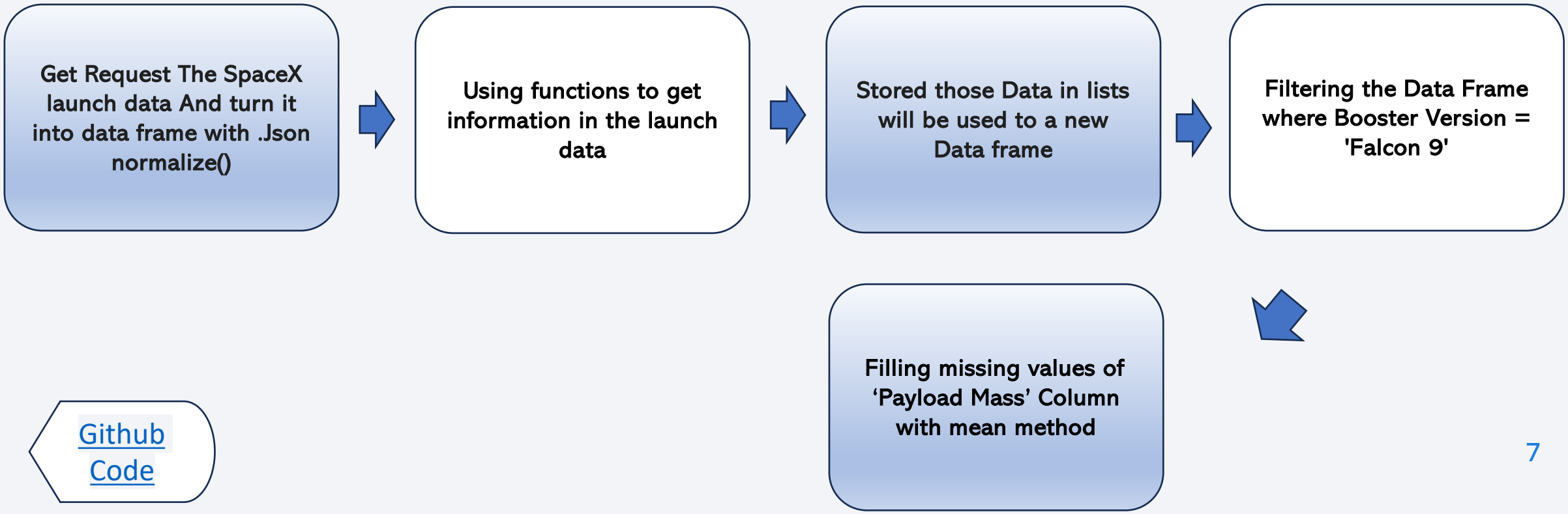
Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX APl and web scraping from Wikipedia.
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - - Building, tuning and evaluation of classification models to get the best results

Data Collection SpaceX API

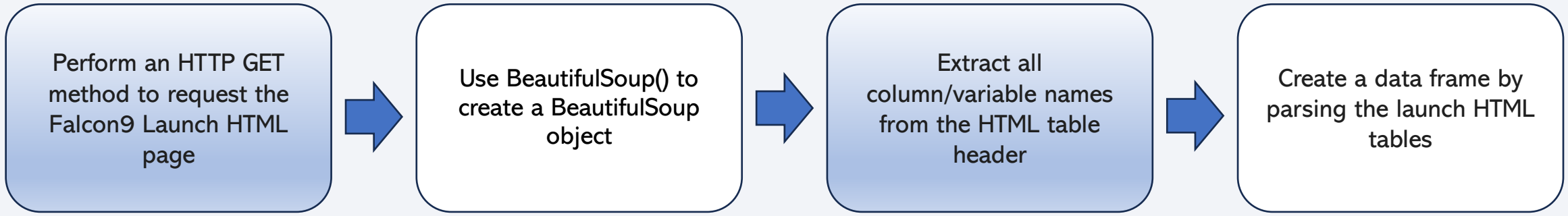
- **The key difference between scraping and API**

API: You're using an official interface designed for programmatic access
Scraping: You're extracting data that wasn't necessarily meant to be accessed programmatically



Github Code

Data Collection - Scraping



Github Code

Data Wrangling

- We need to check each categorical column to ensure not having the value-in-consistency
- There are some categories in landing outcome column so that need to collapsing them to 2 categories :1 represented did land successfully landed and 0 represented did not land successfully.
- For figure outing those cases as well as determining success rate, we should do these steps below:

1. Calculate the number of launches on each site
2. Calculate the number and occurrence of each orbit
3. Calculate the number and occurrence of mission outcome of the orbits
4. Collapsing categories of landing outcome column to 2 categories (0 and 1)
5. Assign the result of before step to new column named class
6. Use mean() method on class column to determine success rate

Github Code

EDA with Data Visualization

- We used seaborn package to visualize relationship between columns so that figure it out any pattern or relationship.
- Three kind of plot have plotted include: scatterplot, bar plot and line plot.
- **Scatter plot**
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Orbit vs. Flight Number
  - Payload vs. Orbit Type
  - Orbit vs. Payload Mass
- **Bar plot**
  - Success rate vs. Orbit
- **Line plot**
  - Success rate vs. Year

Github Code

EDA with SQL

- **Performed SQL queries:**
  - Displaying the names of the unique launch sites in the space mission
  - Displaying 5 records where launch sites begin with the string 'CCA'
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying average payload mass carried by booster version F9 v1.1
  - Listing the date when the first successful landing outcome in ground pad was achieved
  - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster versions which have carried the maximum payload mass
  - Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
  - Ranking 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

Github Code

Build an Interactive Map with Folium

- Folium map object is a map centered on NASA Johnson Space Center at Houston, Texas
- Red circle at NASA Johnson Space Center's coordinate with label showing its name (folium.Circle, folium.map.Marker)
- Red circles at each launch site coordinates with label showing launch site name (folium.Circle, folium.map. Marker, folium.features.Div/con)
- The grouping of points in a cluster to display multiple and different information for the same coordinates (folium.plugins.MarkerCluster).
- Markers to show successful and unsuccessful landings. Green for successful landing and Red for unsuccessful landing. (folium.map Marker, folium.Icon).
- Markers to show distance between launch site to key locations (railway, highway, coastway, city) and plot a line between them. (folium.map.Marker, folium.PolyLine, folium.features.Divicon)
- These objects are created in order to understand better the problem and the data. We can show easily all launch sites, their surroundings and the number of successful and unsuccessful landings.

Github Code

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 count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.
- **Slider of Payload Mass Range:**
  - - Added a slider to select Payload range.
- **Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:**
  - - Added a scatter chart to show the correlation between Payload and Launch Success.

Github Code

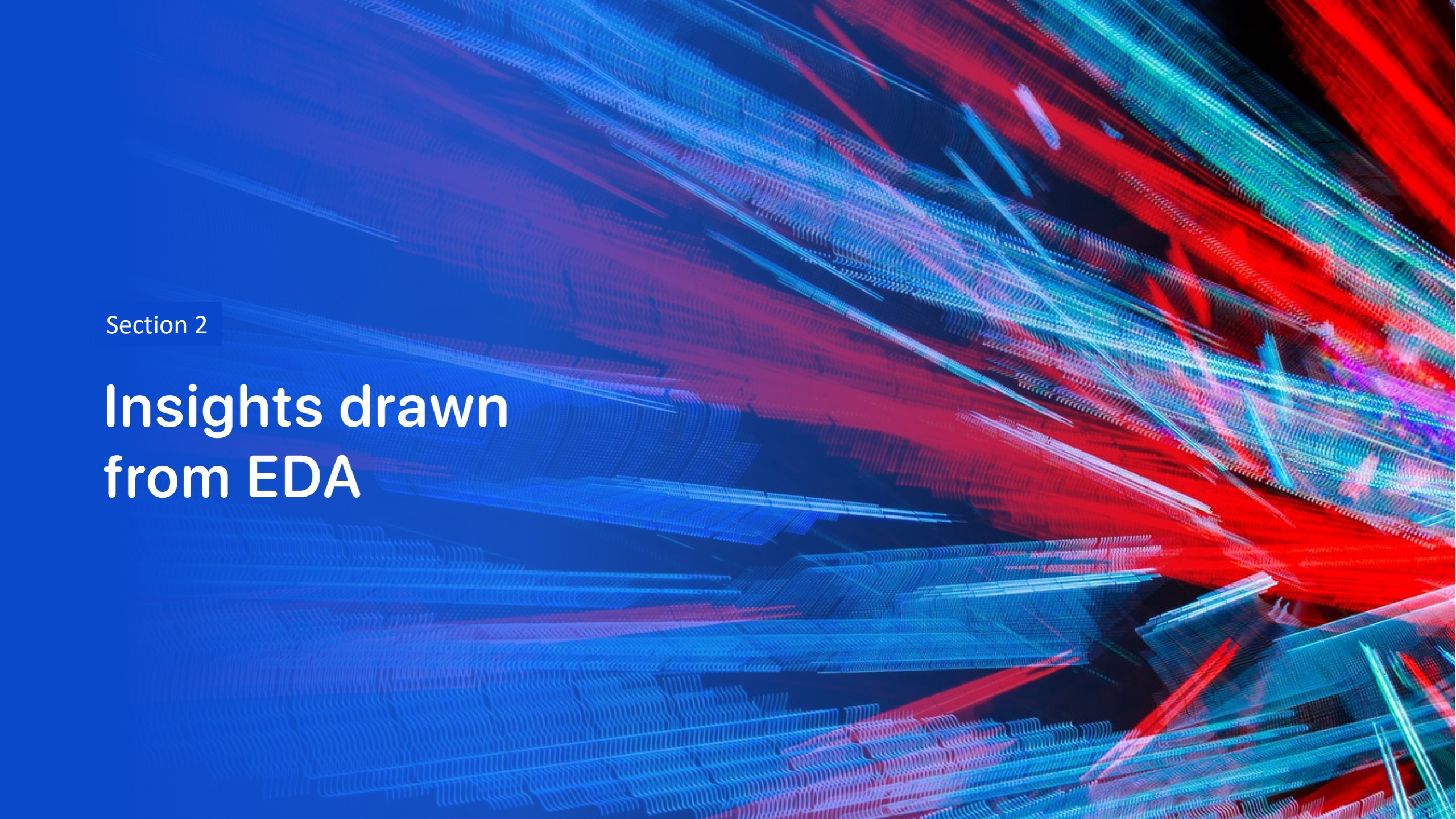
Predictive Analysis (Classification)

- Create a NumPy array from the column Class in data set, Assign to Y
- Standardizing the data in X then reassign it to the variable X
- Splitting data into training and test sets.
- Selection of ML algorithms
- Set parameters for each algorithm to GridSearchCV
- Training GridSearchModel models with training dataset
- Get best hyperparameters for each type of model
- Compute accuracy for each model with test dataset include confusion Matrix, then the best accuracy will be reportet.

Github Code

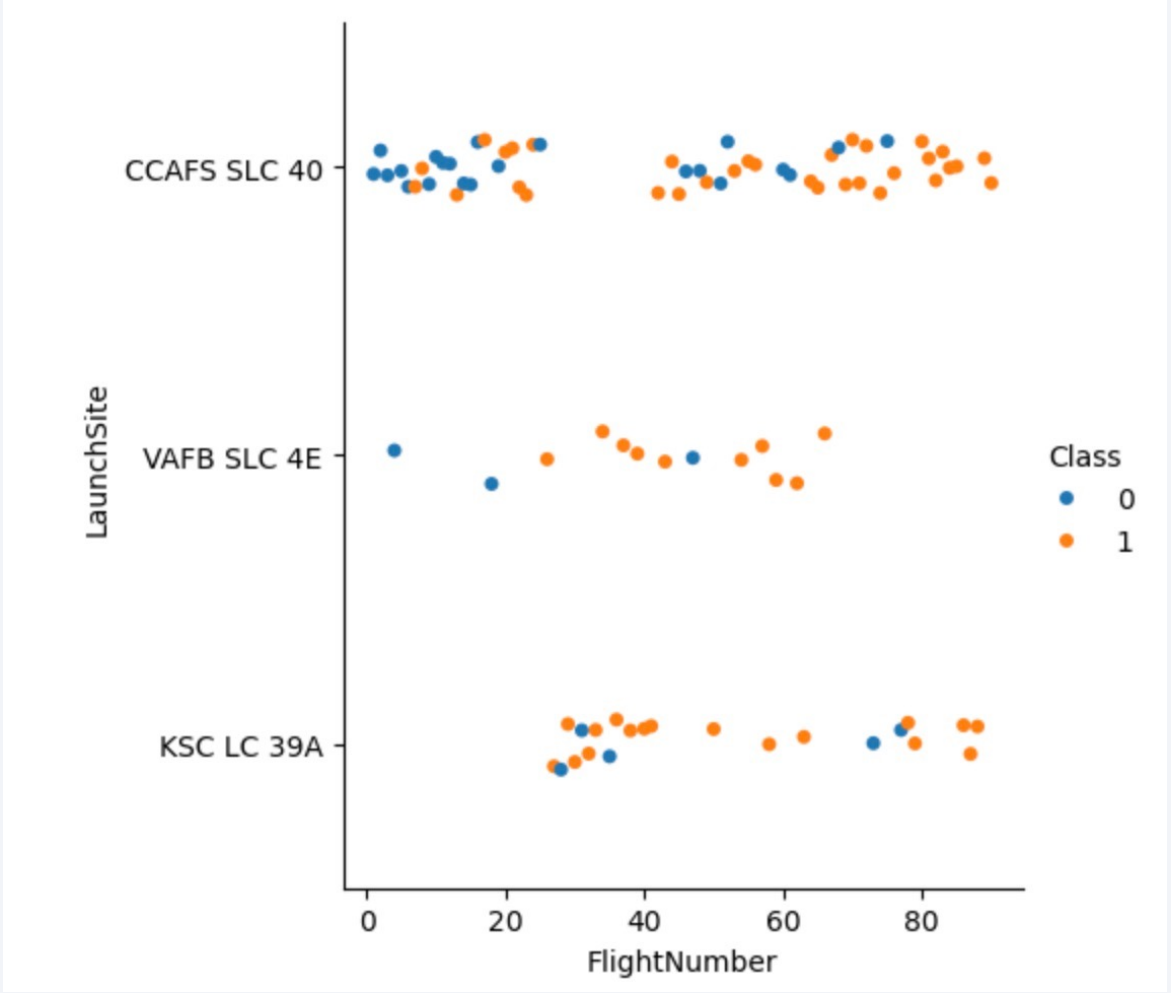
Results

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


Section 2
Insights drawn from EDA

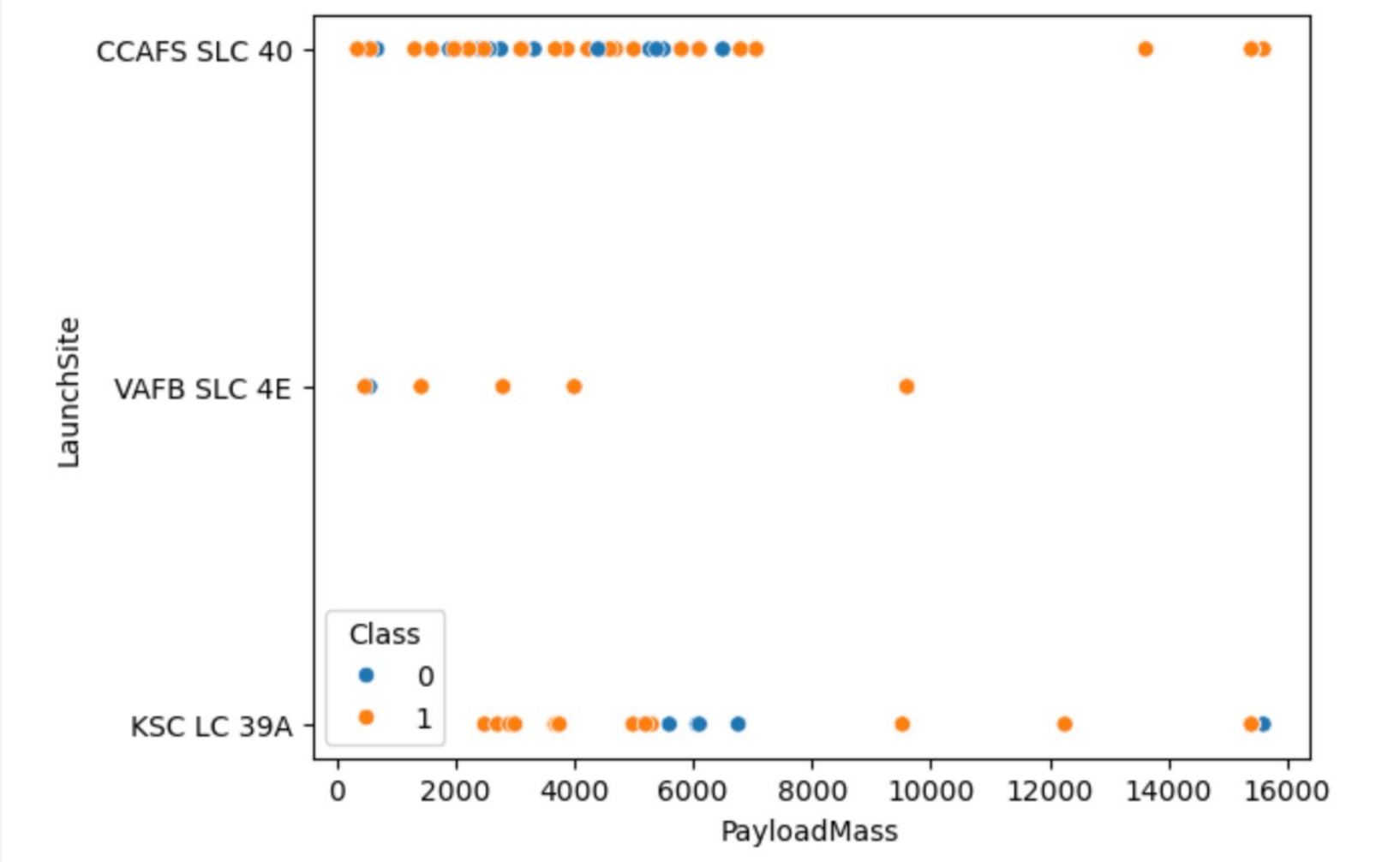
Flight Number vs. Launch Site

- Its seems to be KSC LC 39A and VAFB SLC 4E have higher success rates.
- The Success rate is increasing by larger the Flight number.



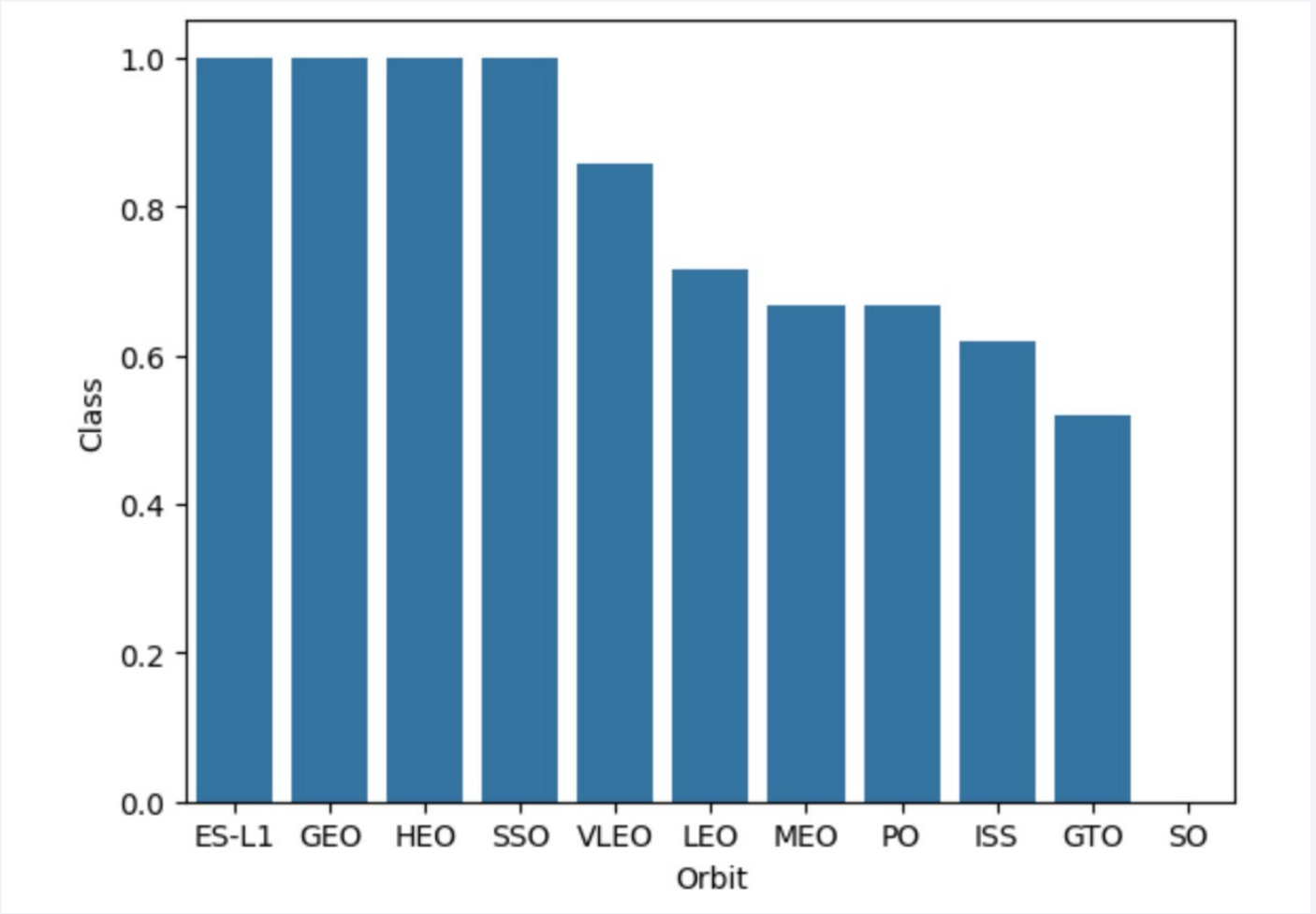
Payload vs. Launch Site

- For every launch site the higher payload mass, the higher success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.



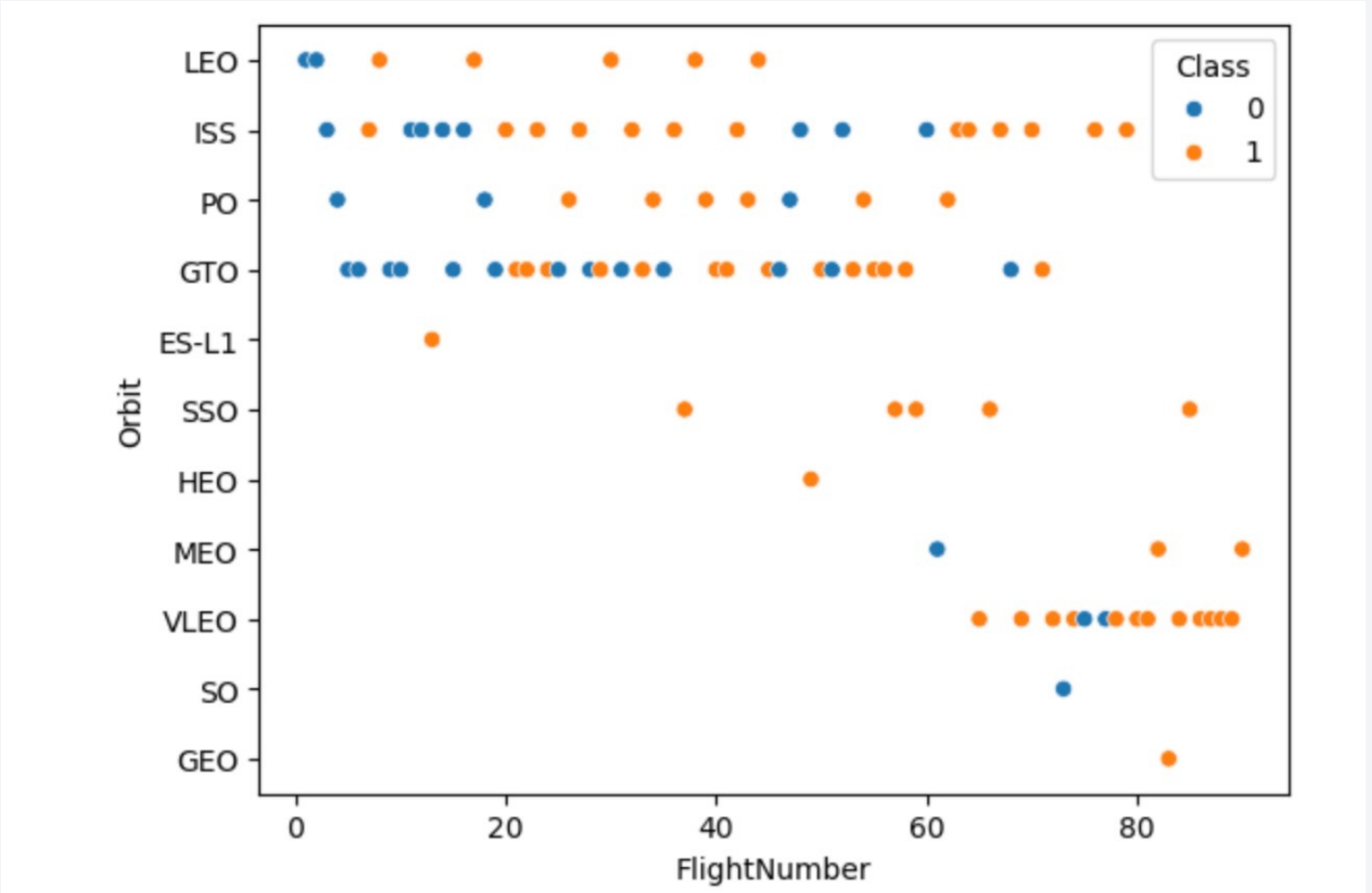
Success Rate vs. Orbit Type

- Orbits with 100% success rates are - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate is - SO



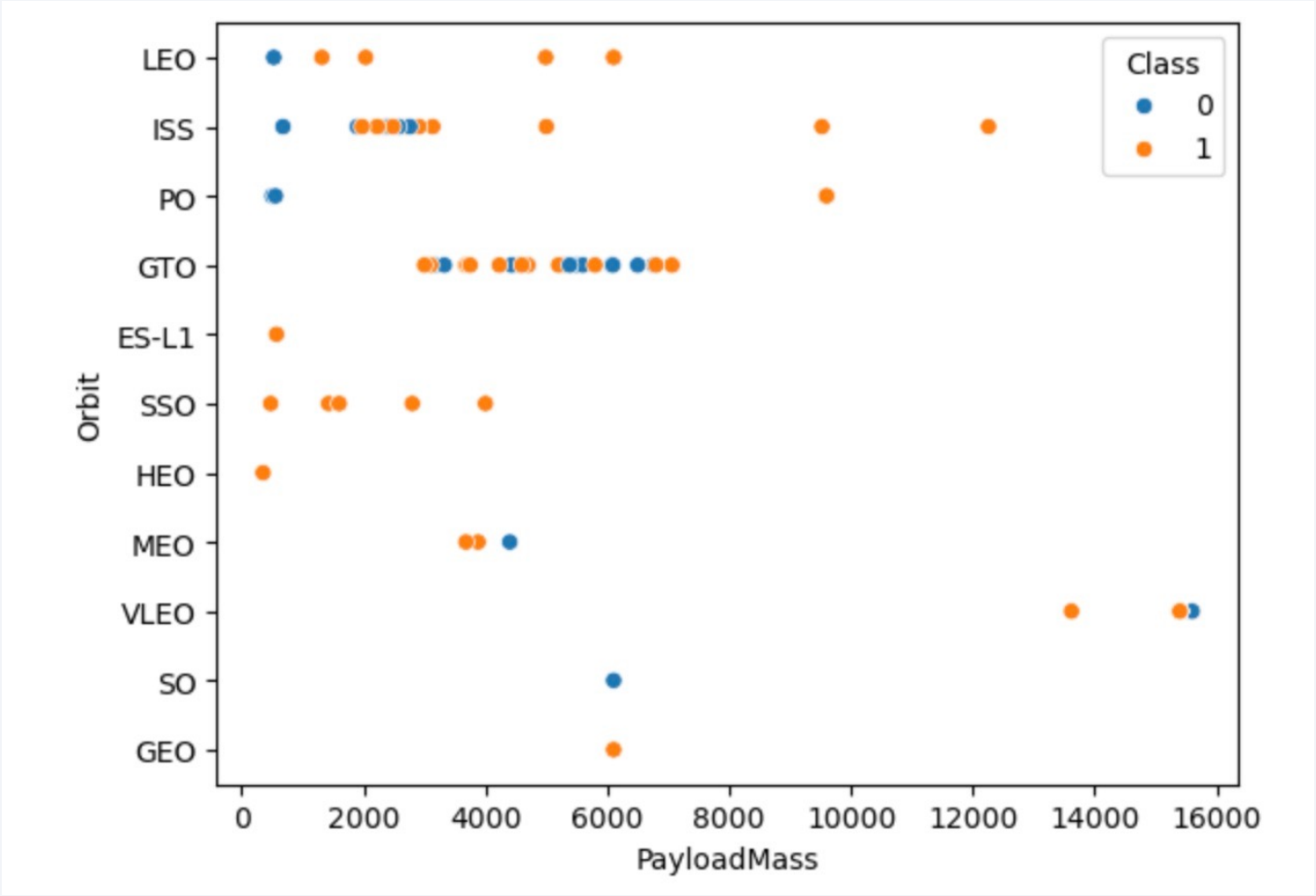
Flight Number vs. Orbit Type

- the LEO orbit, success is related to the number of flights, however in the GTO orbit, there is no relationship between flight number and the success rate.



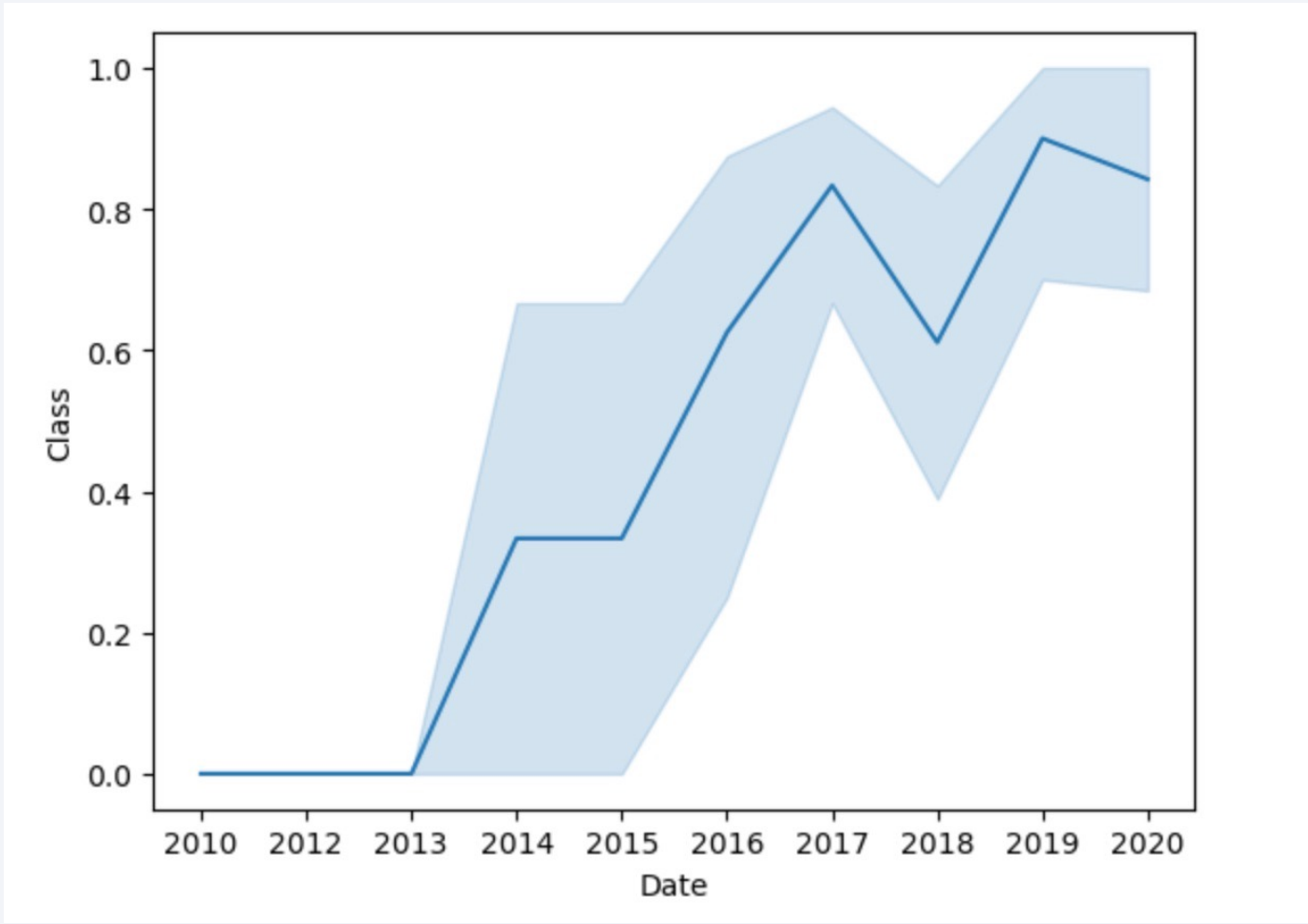
Payload vs. Orbit Type

- heavier payloads improve the success rate for the LEO orbit.
- For GTO, ES-L1, SSO,HEO orbits The Success rate, are independent from PayloadMass



Launch Success Yearly Trend

- The success rate since 2013 kept on increasing.



All Launch Site Names

- We used the key word DISTINCT to show only unique launch sites from the SpaceX data.

```
%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'

- Shows 5 records with launch sites that 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

- the total payload carried by boosters from NASA

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

```
%sql select sum(payload_mass__kg_) as sum from spacextable where customer like 'NASA (CRS)'
```

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

sum
45596

Average Payload Mass by F9 v1.1

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

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

```
%sql select avg(payload_mass__kg_) as Average from SPACEXTABLE where booster_version like 'F9 v1.1%'
```

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

Average
2534.6666666666665

First Successful Ground Landing Date

- dates of the first successful landing outcome on ground pad

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql select min(date) as Date from SPACEXTABLE where landing_outcome like 'Success (ground pad)'
```

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

Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- 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 from SPACEXTABLE \
where (landing_outcome like 'Success (drone ship)')AND (payload_mass__kg_ between 4000 and 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

- the total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXTABLE GROUP by mission_outcome ORDER BY mission_outcome
```

```
 * 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

- 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_=(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

- 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, substr(Date,0,5) as year,
    Landing_outcome, booster_version, launch_site from SPACEXTABLE
    where Landing_outcome like 'Failure (drone ship)%' AND substr(Date,0,5) ='2015'
```

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

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

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

- 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(*) as count from SPACEXTABLE
where Date >= '2010-06-04' AND Date <= '2017-03-20'
GROUP by landing_outcome ORDER BY count Desc
```

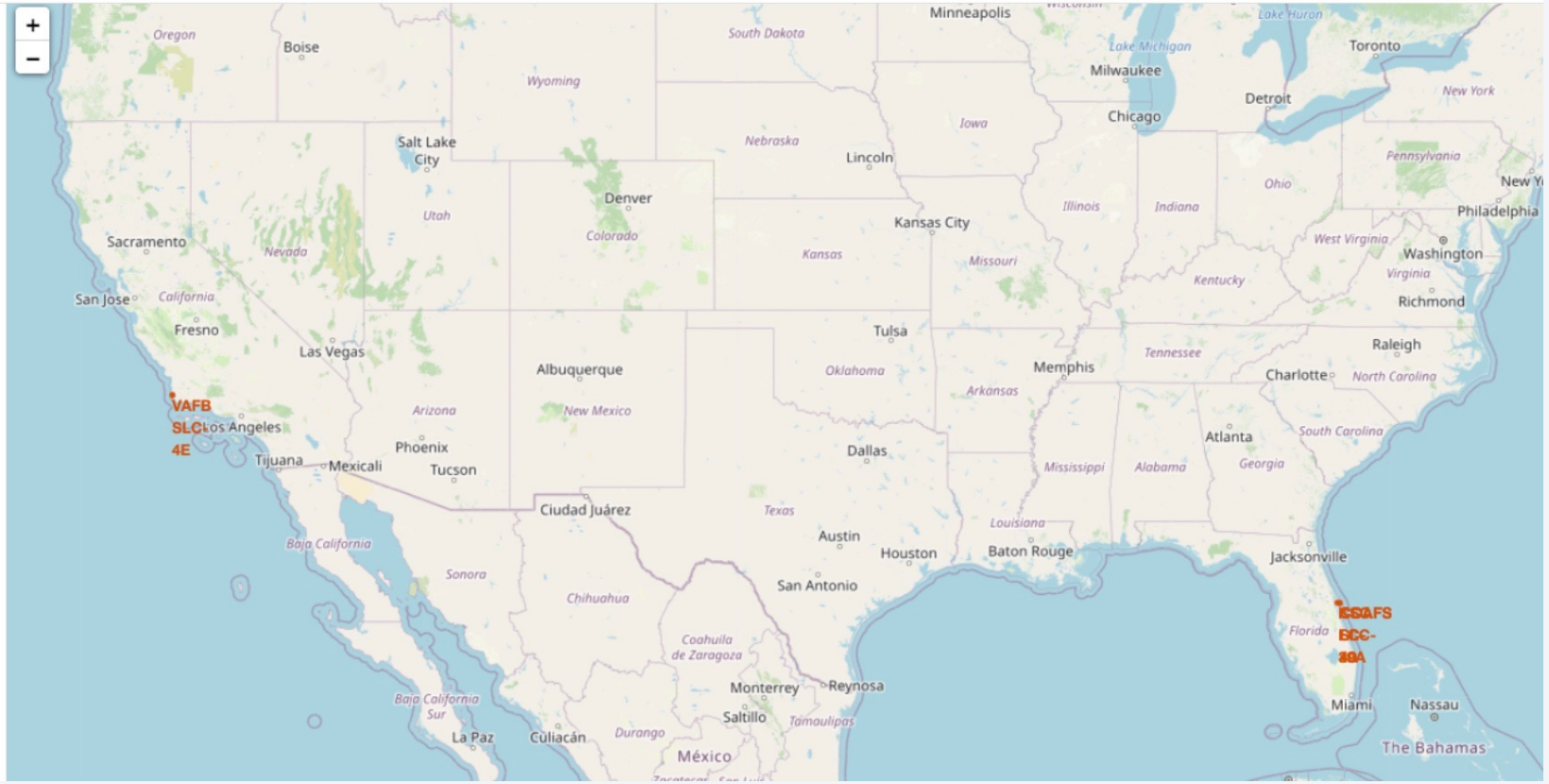
```
 * sqlite:///my_data1.db
Done.
```

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

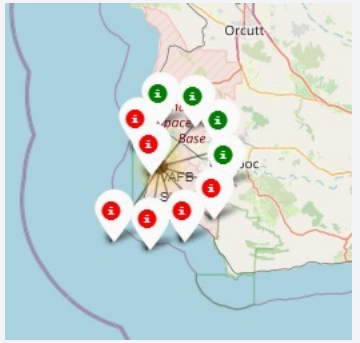
All launch sites’ locations

- We see that Space X launch sites are located on the coast of the United States



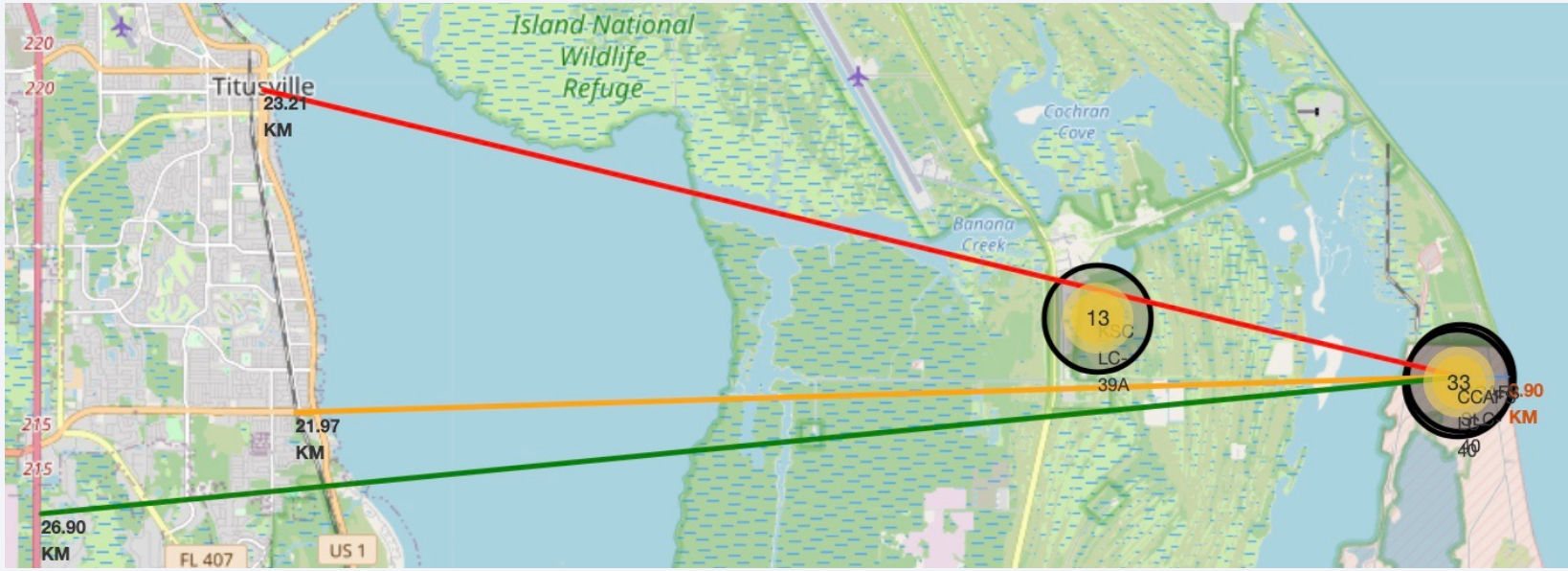
Color-labeled launch records on the map

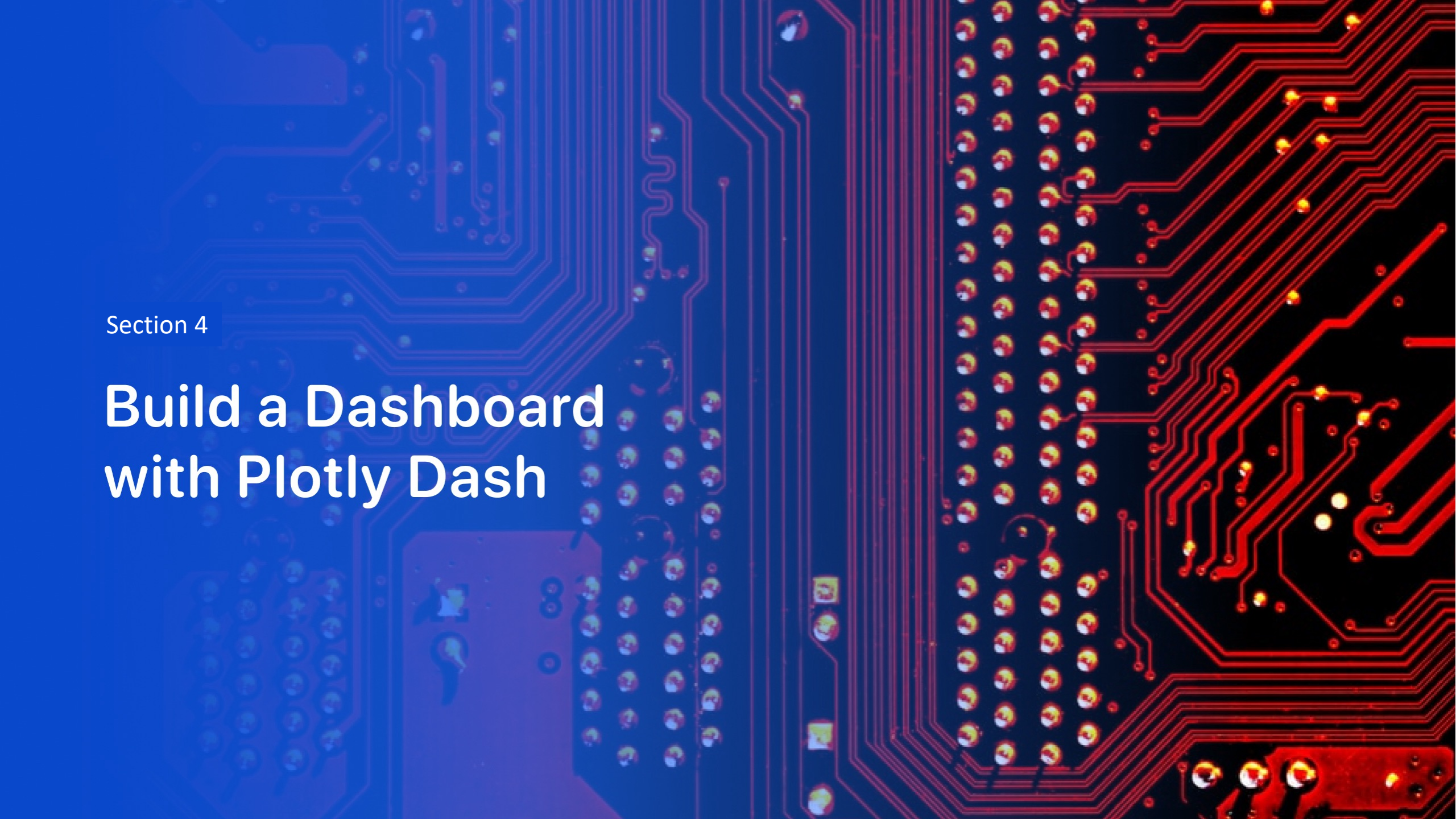
- From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates. - Green Marker = Successful Launch - Red Marker = Failed Launch
- Launch Site KSC LC-39A(Top right fig) has a very high Success Rate.



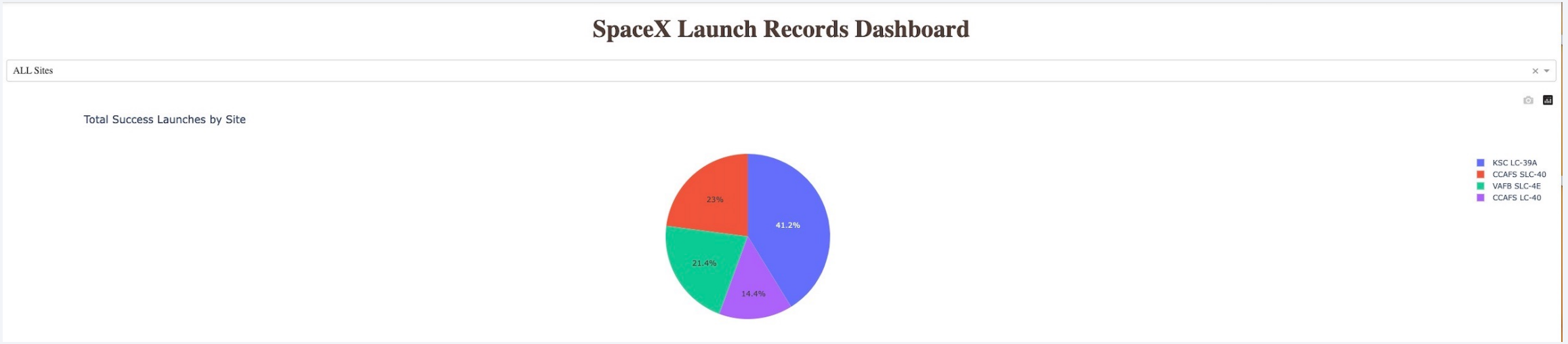
Distance from the launch site KSC LC-39A to its proximities

- From the visual analysis of the launch site CCAFS SLC-40we can clearly see that it is:
- relative close to railway (21.97 km)
- relative close to highway (26.90 km)
- relative close to city (23.21 km)




Section 4
Build a Dashboard with Plotly Dash

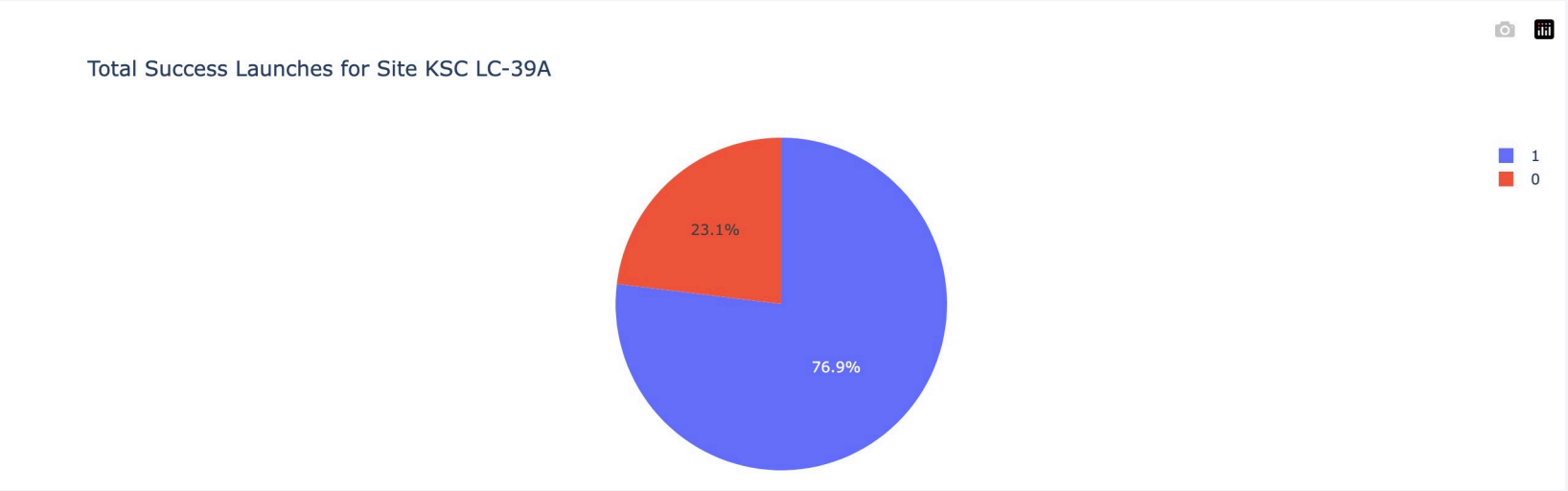
Launch success count for all sites



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

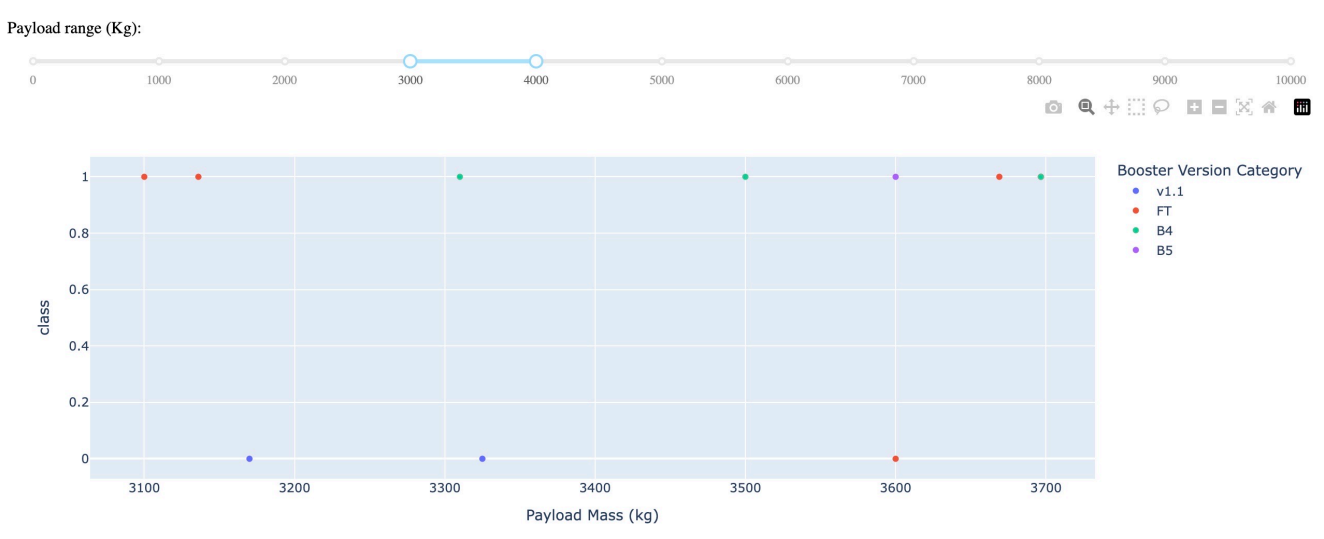
Launch site with highest launch success ratio

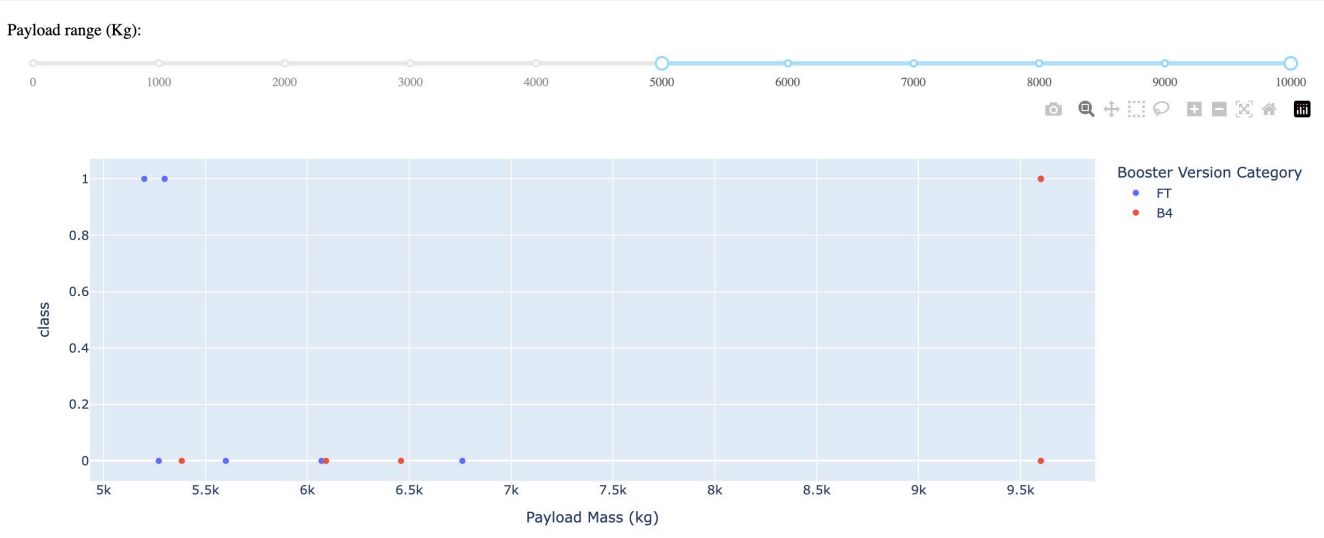
- We see KSC LC-39A launch site has the highest launch success rate (76.9%).



Payload Mass vs. Launch Outcome scatter plot for all sites

- Low payloads mass have a better success rate than the heavy payloads mass specially in range 3000 – 4000 Kg.



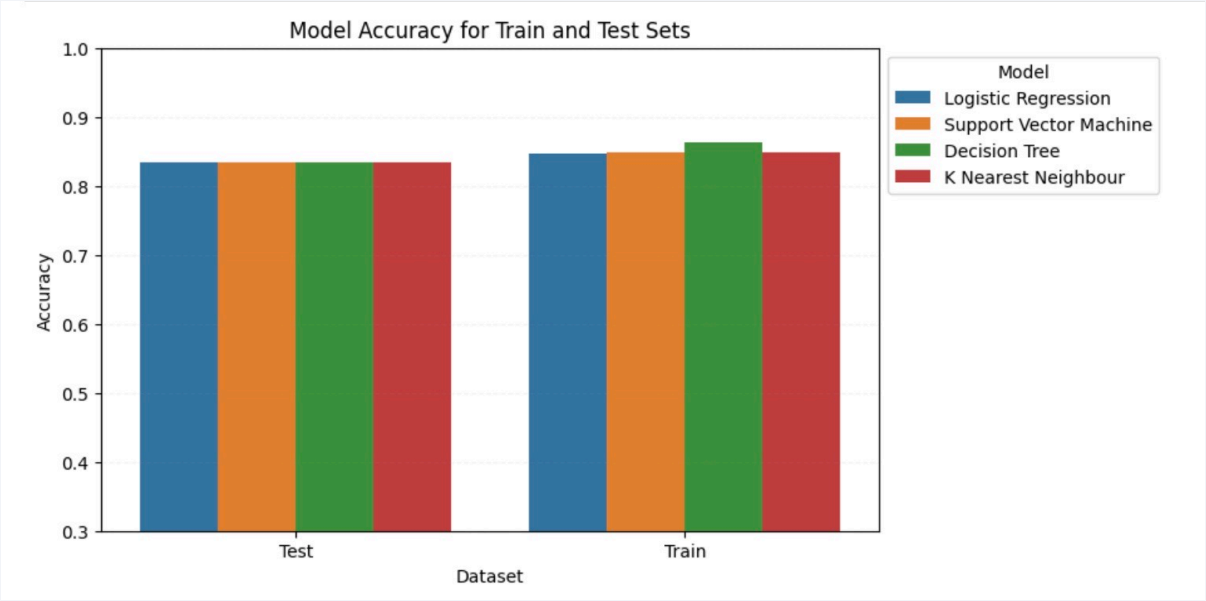


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- We see that all models have the same accuracy in the test dataset; however, the Decision Tree model has the best accuracy in the train dataset. The same test set accuracy may be due to the small test sample size



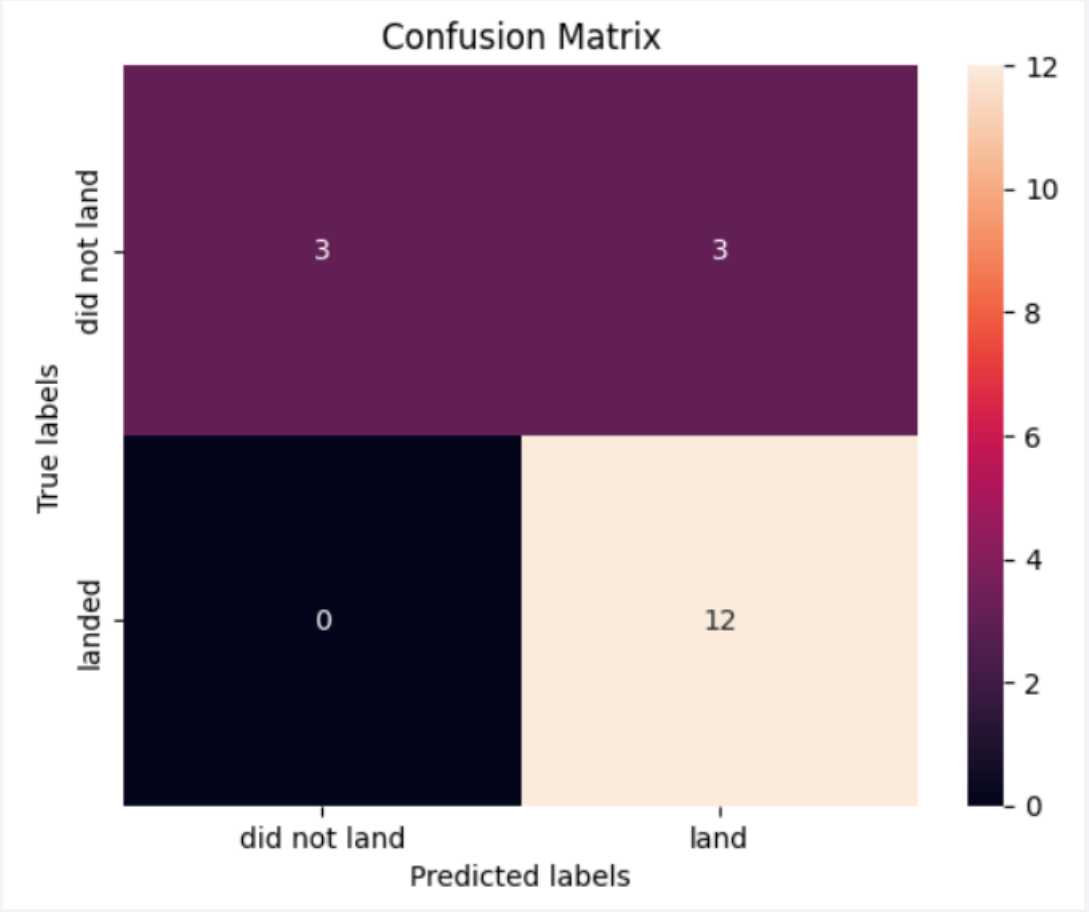
- The best parameters of best model are:

```
Best Model: Decision Tree
Best Parameters:
criterion: gini
max_depth: 8
max_features: sqrt
min_samples_leaf: 2
min_samples_split: 10
splitter: random
```

	Model	Accuracy_Test	Accuracy_Train_Cv
0	Logistic Regression	0.833333	0.846429
1	Support Vector Machine	0.833333	0.848214
2	Decision Tree	0.833333	0.862500
3	K Nearest Neighbour	0.833333	0.848214

Confusion Matrix

- As the test accuracy of all models are equal, the confusion matrices are also identical .The main problem of these models are false positives.



Conclusions

- The success rate of launches increases over the years.
- Orbits ES-L1, GEO, HEO and SSO had the most success rate.
- KSC LC-39A has the highest success.
- Depending on the orbits, the payload mass can be impact to the success of a mission. Some orbits require a light or heavy payload mass.
- In overall, Low payloads mass have a better success rate than the heavy payloads mass.
- Finally, we used some machine learning algorithms to learn the pattern to predict whether a mission will be successful or not based on the given features. Decision Tree Model is the best algorithm due to has the best train set accuracy.

Thank you!