IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anna Korponay
05/03/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using API and Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis using SQL
  - Exploratory Data Analysis using Data Visualization
  - Interactive Data Visualization with Folium
  - Interactive Dashboard with Plotly Dash
  - Predictive Analysis
- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive Data Visualization
  - 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. If it can be determined whether the first stage will land, the cost of a launch can be determined.
  - Machine learning models and public information will be used to predict if SpaceX will reuse the first stage.
- Problems you want to find answers
  - What are the characteristics that affect a successful landing?
  - What happened to the rate of successful landings over time?
  - Which model can best predict whether the first stage will land?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web Scraping using Wikipedia
- Perform data wrangling:
  - Cleaning the data
  - One hot encoding for success and failure
- 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 Collection

SpaceX REST API

This API will provide us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

SpaceX REST API URL:

api.spacexdata.com/v4/launches/past

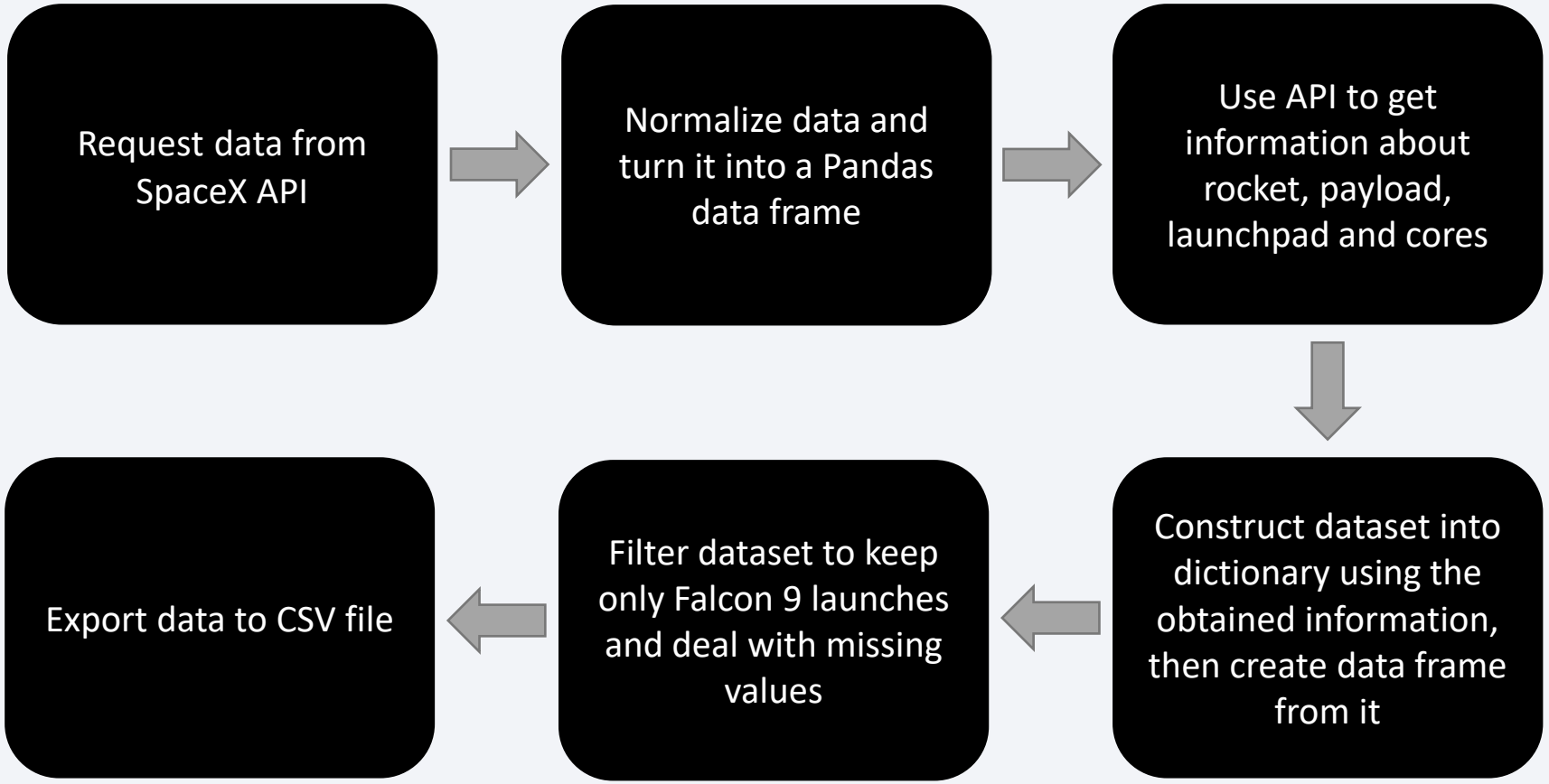
Web Scraping using Wikipedia

The web scraping will provide us data about launches, including the flight number, launch site, payload, orbit, customer, launch outcome, as well as the date and time of the launches.

URL for Wikipedia page used:

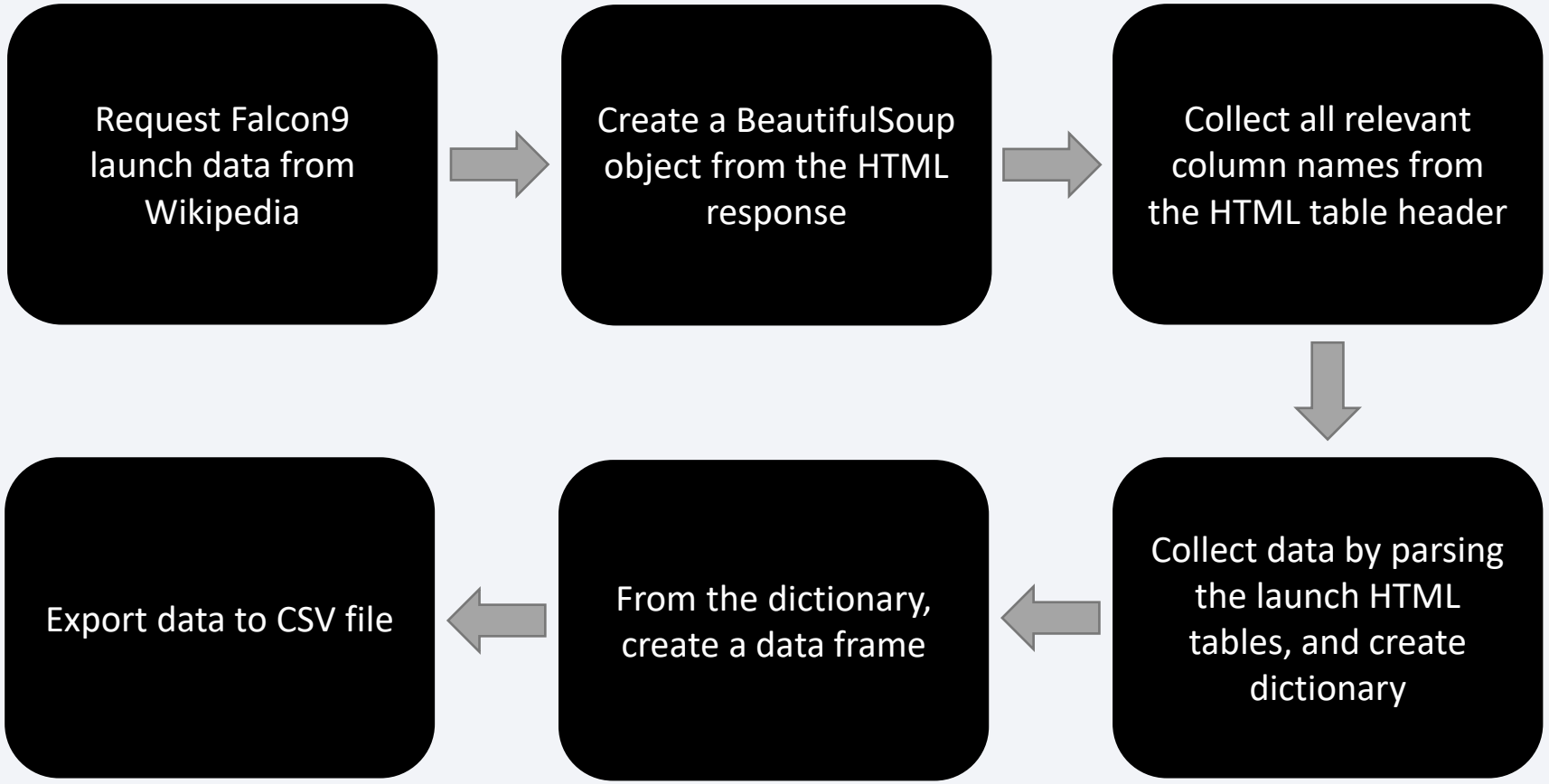
https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data Collection – SpaceX API



Github: Data Collection API

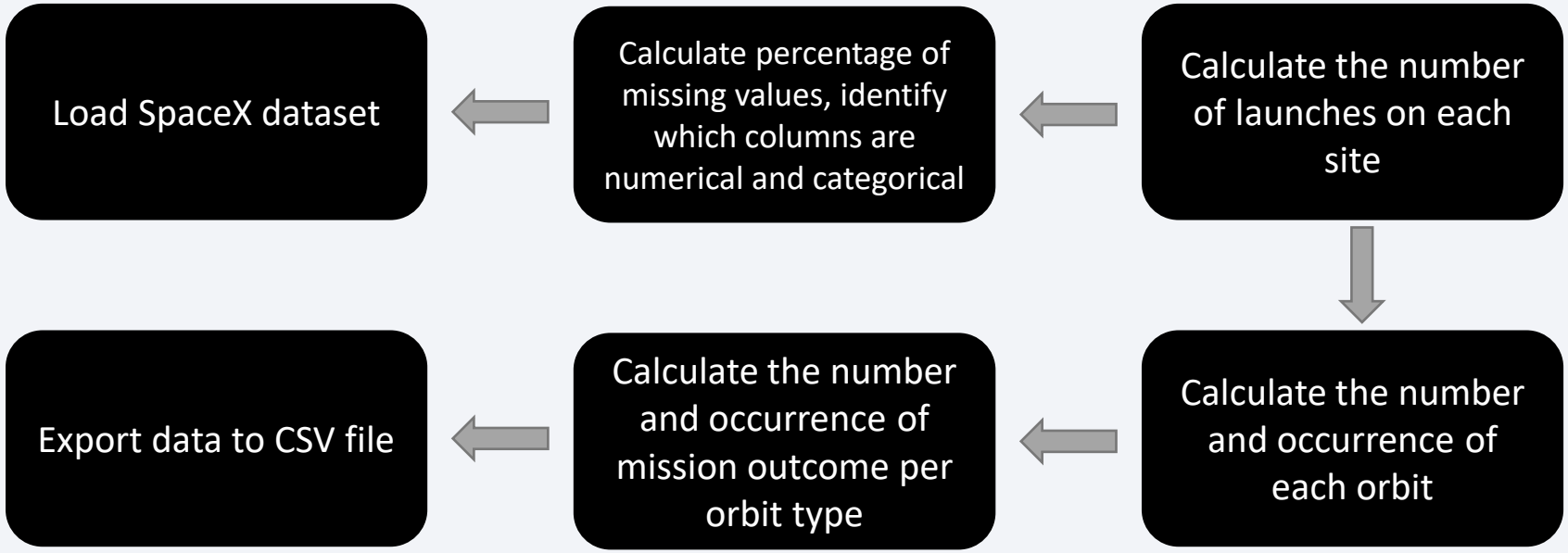
Data Collection – Scraping



Github: Data Collection Web Scraping

Data Wrangling

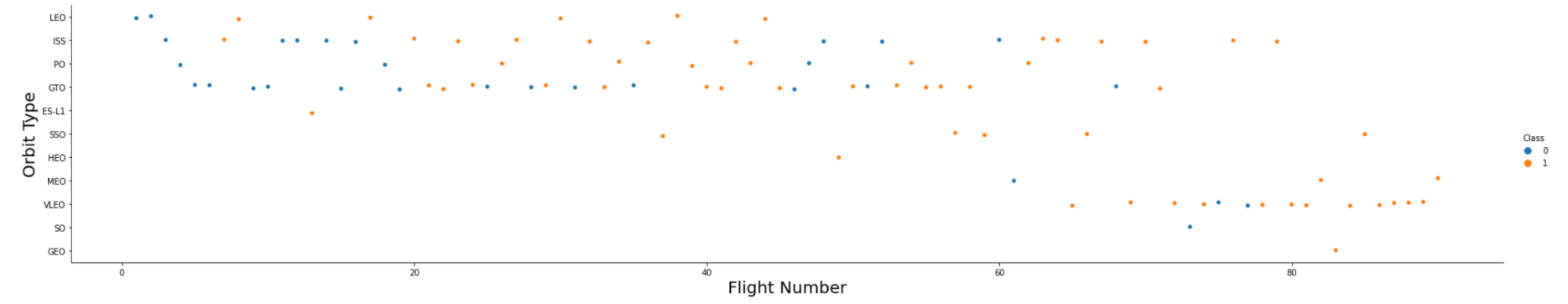
- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.



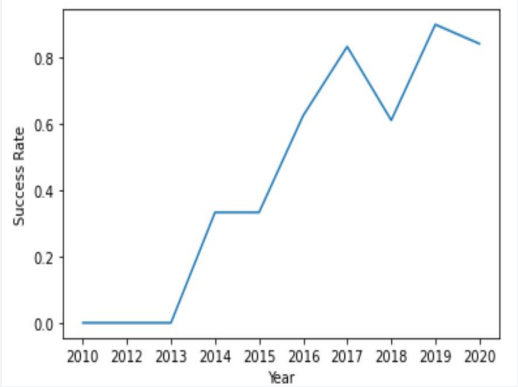
Github: Data Wrangling

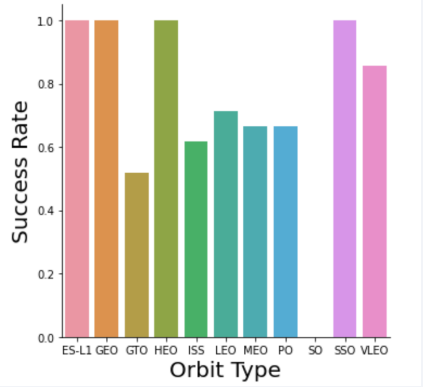
EDA with Data Visualization

- Scatter plots were used to show the relationship between flight number and payload mass, flight number and launch site, payload mass and launch site, orbit and flight number, orbit and payload mass.
- A bar chart was used to visually check whether there is any relationship between success rate and orbit type.
- A line chart was used to visualize the launch success yearly trend.
- Examples of charts:



Github: EDA Data Visualization





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 sites for the months in year 2015
    - Ranking the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Github: EDA with SQL

Build an Interactive Map with Folium

- Colored Markers
  - A green marker was used if the landing outcome was successful, otherwise the marker's color was red.
- Circles
  - A blue circle was created at NASA Johnson Space Center's coordinate with a popup label showing its name.
  - A circle was created for each launch site, with the launch site name as a popup label
- Lines
  - Lines were added to calculate the distance between a launch site to its proximities (railway, coastline, closes city)

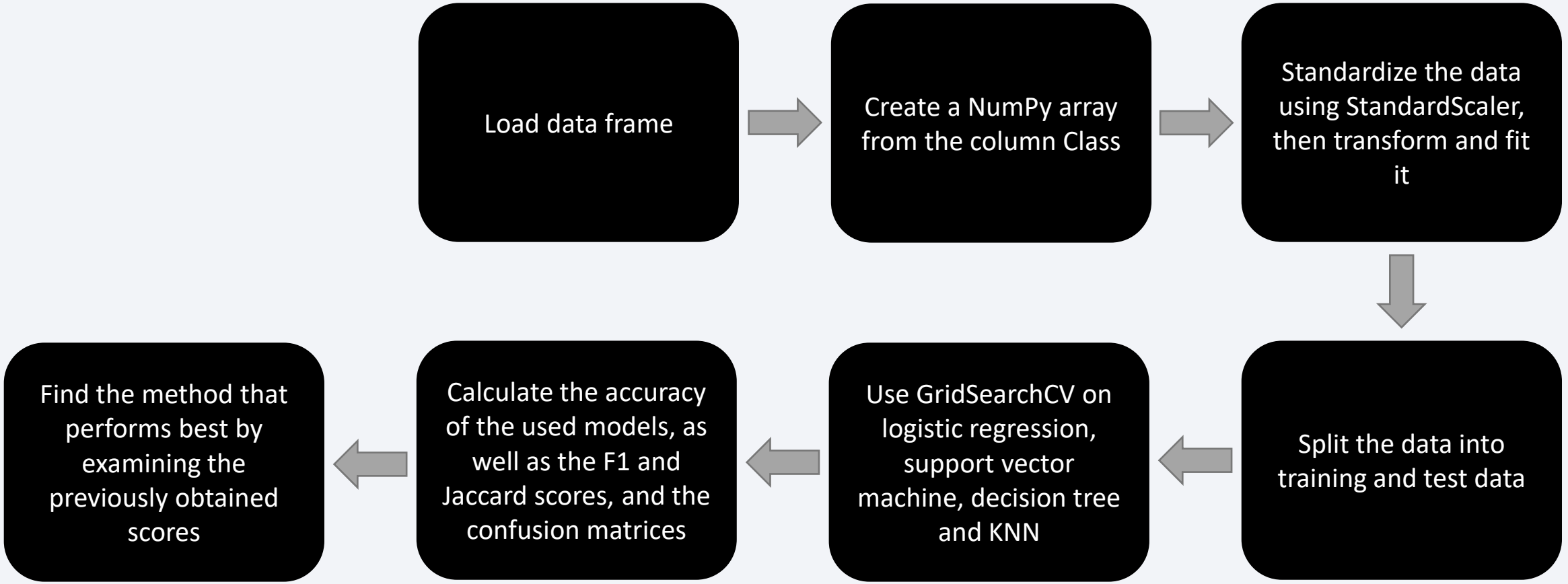
Github: Interactive Visual Analytics with Folium

Build a Dashboard with Plotly Dash

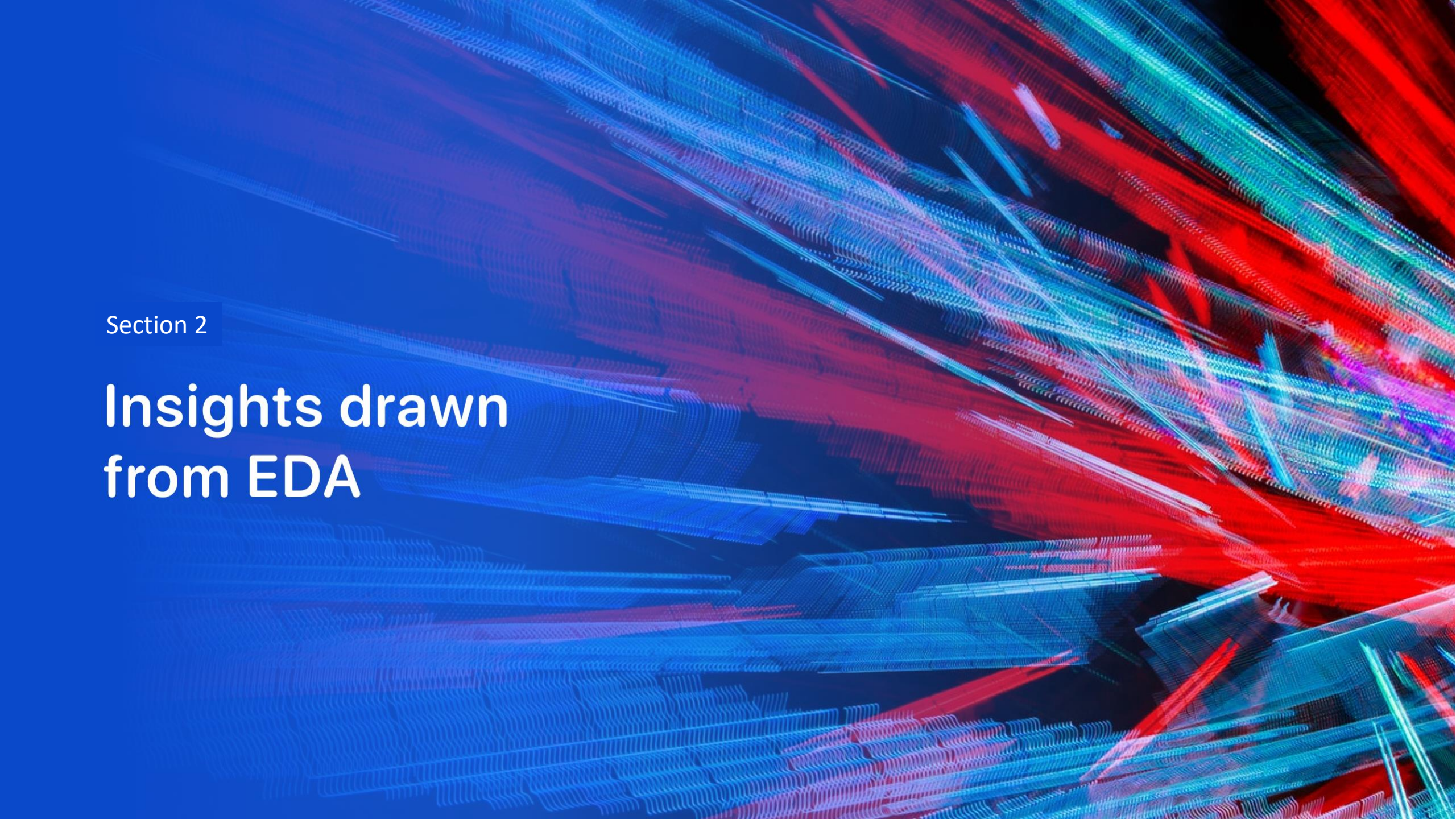
- The Plotly Dash has the following attributes:
  - Dropdown menu
    - This attribute helps in selecting the launch site of interest.
  - Pie chart
    - This attribute shows the successful launches for each site compared to failed ones, if a launch site is selected.
  - Range slider
    - This helps in selecting the payload mass, to check if this variable is correlated to the mission outcome.
  - Scatter chart
    - This chart shows the correlation between the payload mass and the launch site.

Github: Spacex Dash App

Predictive Analysis (Classification)

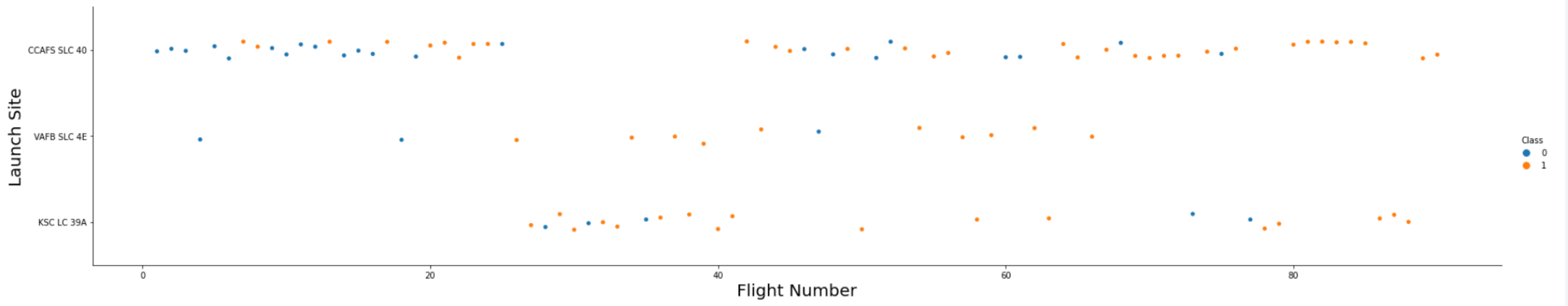


Github: Machine Learning Prediction


Section 2
Insights drawn from EDA

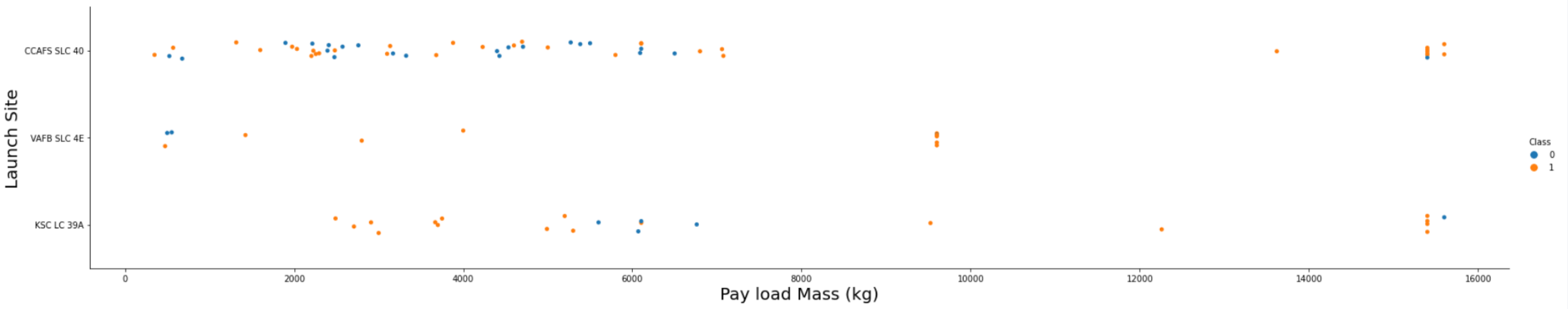
Flight Number vs. Launch Site

- The scatter plot shows the relationship between the flight number and launch site.
- The earliest flights were likely to fail, while the later ones all succeeded.
- The CCAFS SLC 40 launch site has the most launches, however, the other two launch sites have higher success rates.



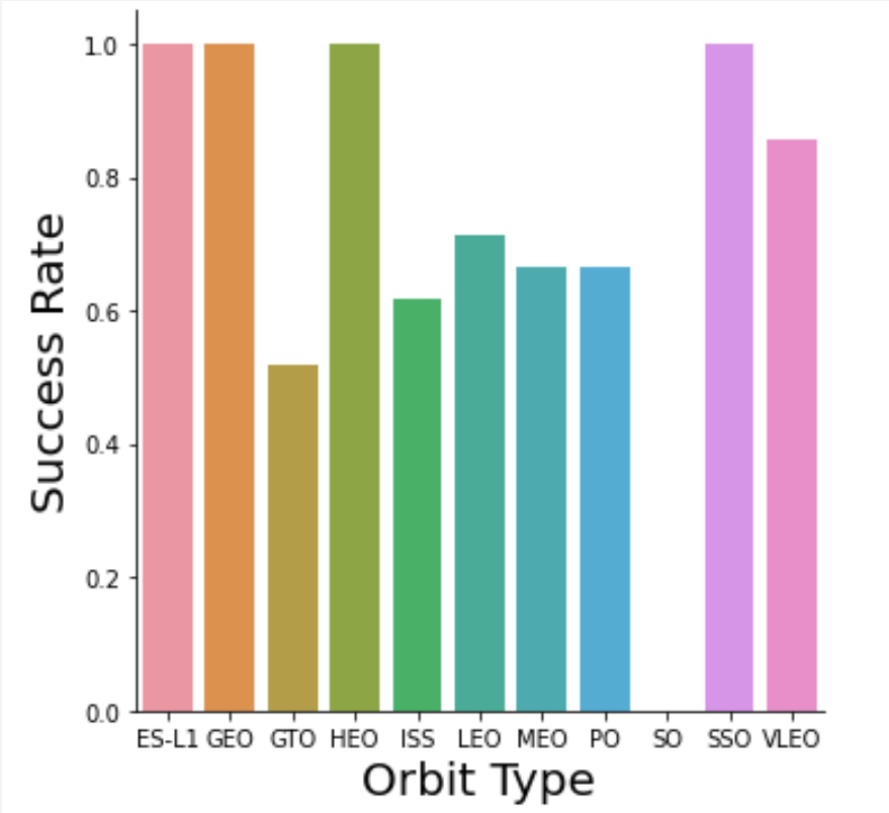
Payload vs. Launch Site

- The scatter plot shows the relationship between the payload and launch site.
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000). For the other launch sites there are rockets launched for various payload masses – heavy ones, as well as not so heavy ones.



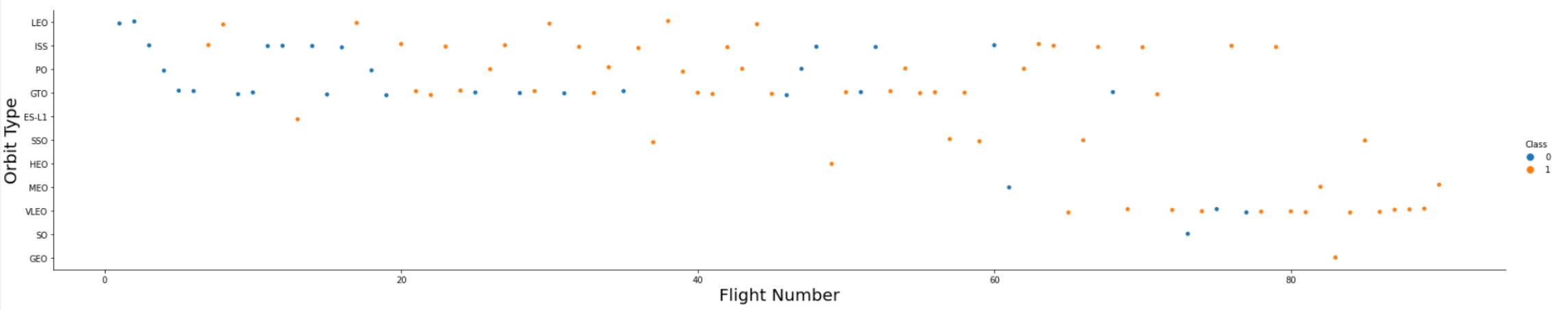
Success Rate vs. Orbit Type

- The bar chart shows the success rate of each orbit type.
- There was a 100% success rate for orbit types:
  - ES-L1
  - GEO
  - HEO
  - SSO



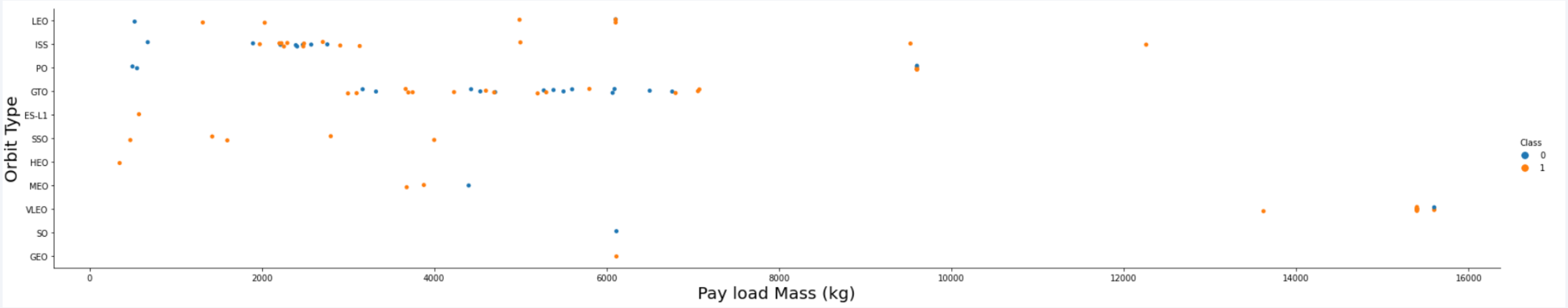
Flight Number vs. Orbit Type

- The scatter plot shows the relationship between the flight numbers and orbit type.
- LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



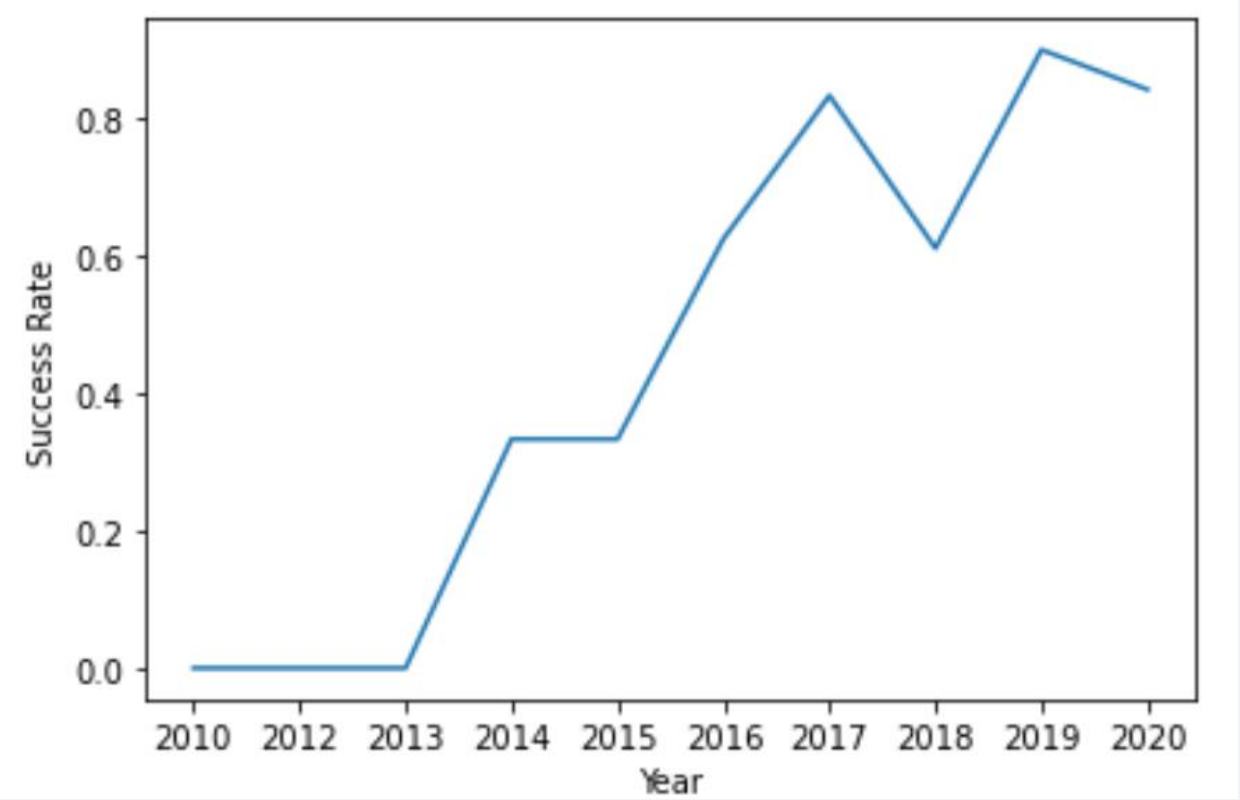
Payload vs. Orbit Type

- The scatter plot shows the relationship between the payload and 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

- The line chart shows the yearly average success rate.
- The success rate increased over time, with a drop in 2018.



All Launch Site Names

The query below displays the unique launch sites in the space mission.

```
%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'

The query below displays 5 records where the launch site name begins 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

The query below displays the total payload mass carried by boosters launched by NASA.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL WHERE Payload LIKE '%CRS%';
```

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

TOTAL_PAYLOAD_MASS
111268

Average Payload Mass by F9 v1.1

- The query below displays the average payload mass carried by booster version F9 v1.1.

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

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

AVG_PAYLOAD_MASS
2928.4

First Successful Ground Landing Date

- The query below displays the average payload mass carried by booster version F9 v1.1.

```
sql SELECT MIN(substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2)) AS FIRST_SUCCESS FROM SPACEXTBL WHERE "LANDING _OUTCOME" = 'Success (ground p
```

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

FIRST_SUCCESS
20151222

Successful Drone Ship Landing with Payload between 4000 and 6000

- The query below displays the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND "LANDING _OUTCOME" = 'Success (drone ship)';
```

```
 * 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 query below displays the total number of successful and failure mission outcomes.

```
%sql SELECT Mission_Outcome, COUNT(*) AS TOTAL FROM SPACEXTBL GROUP BY Mission_Outcome ORDER BY Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

- The query below displays the names of the booster which have carried the 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

- The query below displays failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.

```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE "LANDING _OUTCOME" = 'Failure (drone ship)' AND substr(Date,7,4)='2015';
```

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

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- The query below ranks the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

```
%%sql
SELECT "Landing _Outcome",count("Landing _Outcome")as LANDING_OUTCOME_COUNT,DATE
from SPACEXTBL where substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2) between '20100604'
and '20170320'and "Landing _Outcome" like "Success%"
group by "Landing _Outcome" order by count("Landing _Outcome") desc
```

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

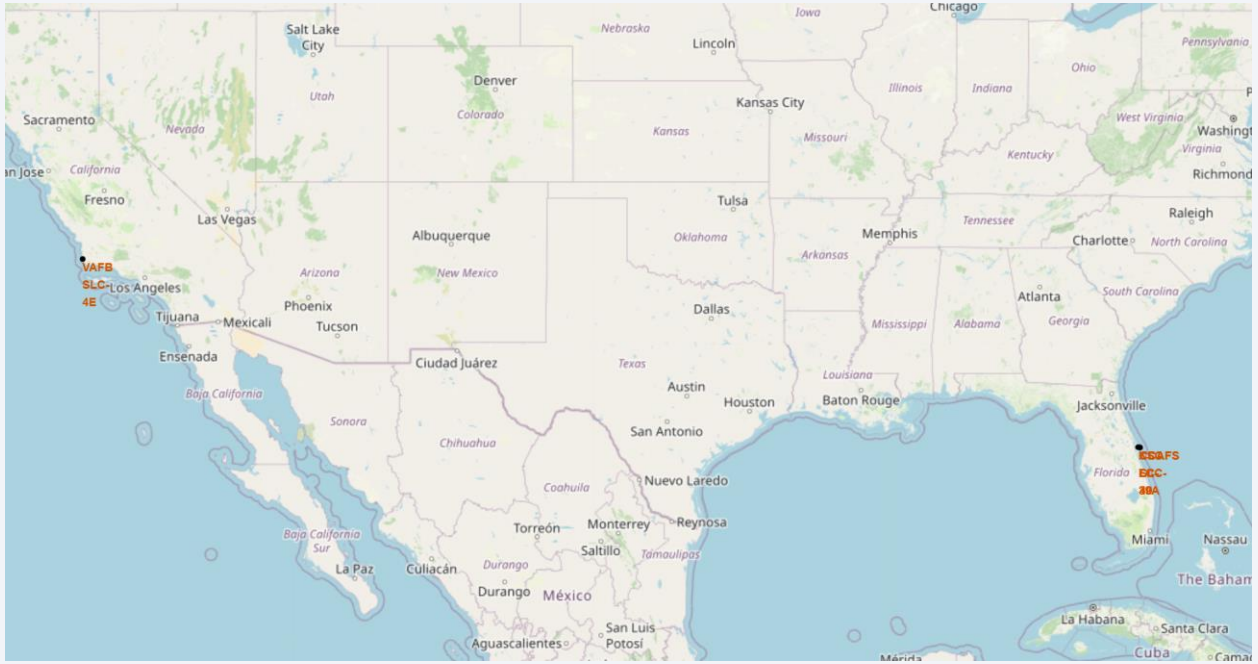
Landing _Outcome	LANDING_OUTCOME_COUNT	Date
Success (drone ship)	5	08-04-2016
Success (ground pad)	3	22-12-2015

Section 3

Launch Sites Proximities Analysis

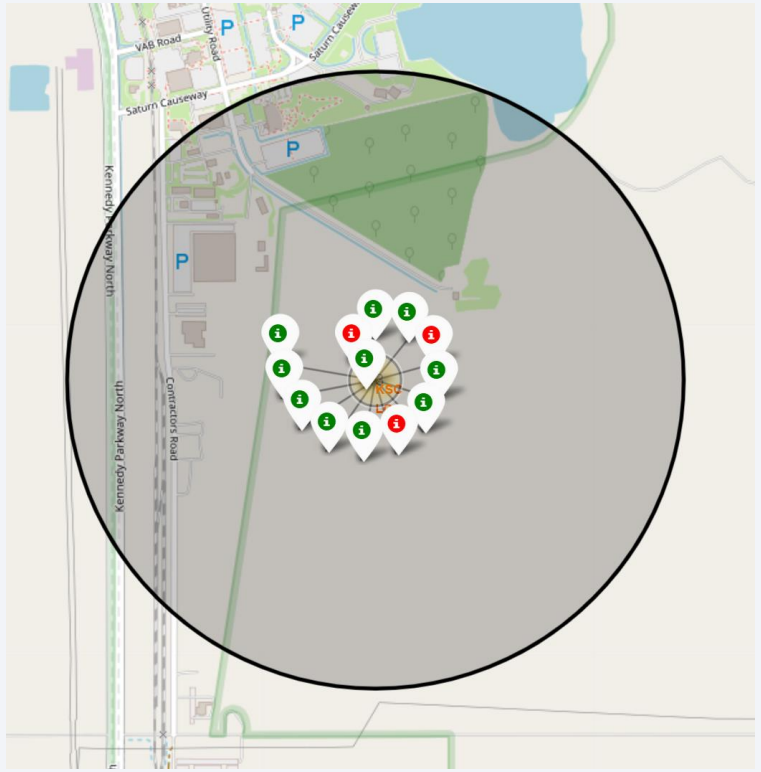
Launch site location markers on the map

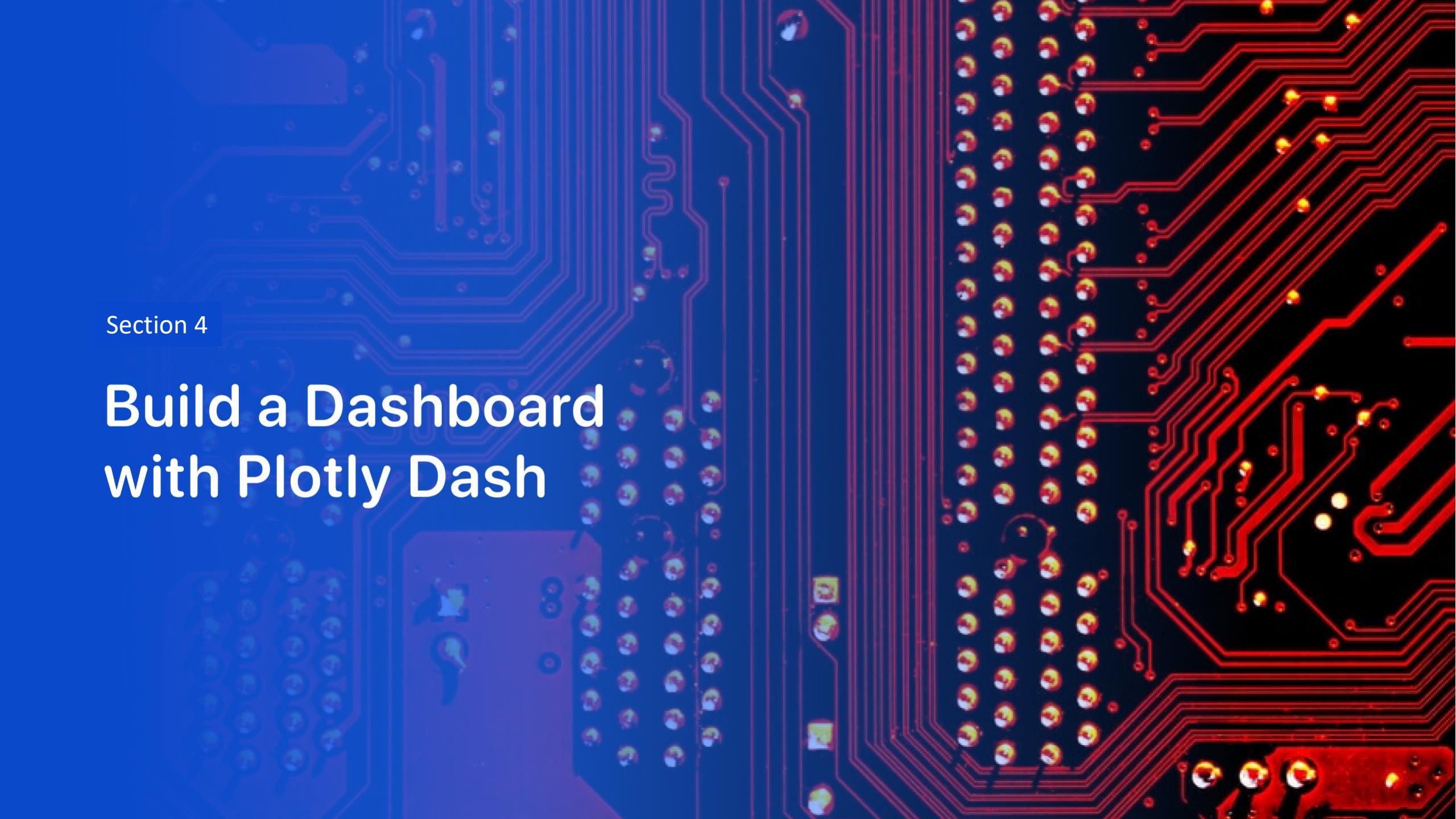
- From the picture it is visible that the launch sites are on the coast lines of the United States, in California and in Florida.
- Being close to the ocean, it is safer to launch the rockets.



Color Labeled Markers

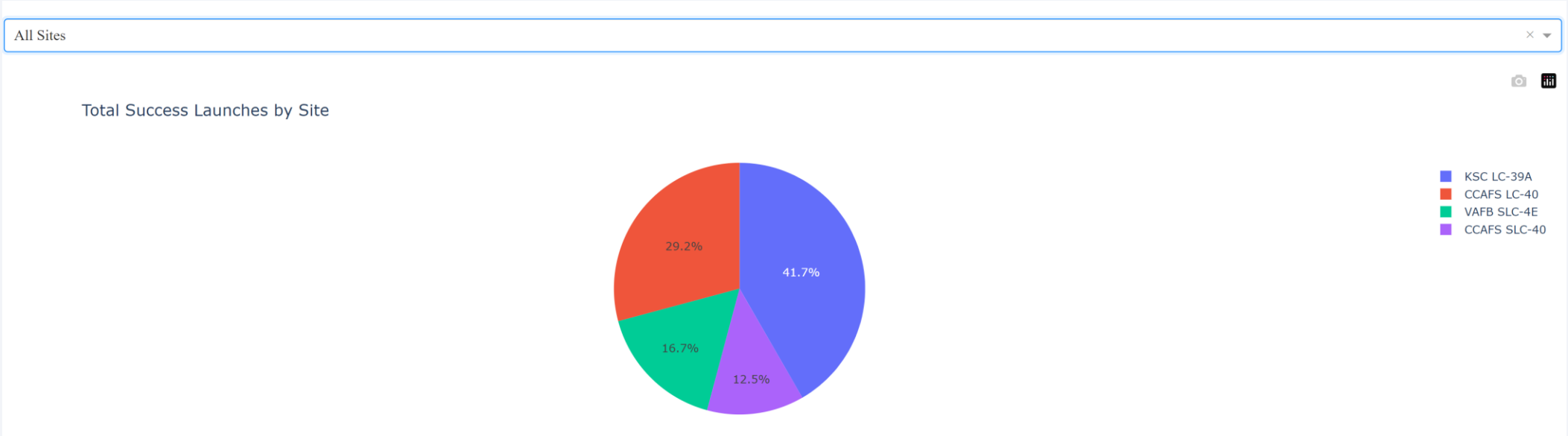
- The color labeled markers show us whether the launches were successful or if they failed.
- Green color represents success, while red represents failure.
- In the case of the launch site showed on the picture, we can conclude that it has a high success rate, as most markers are green.




Section 4
Build a Dashboard with Plotly Dash

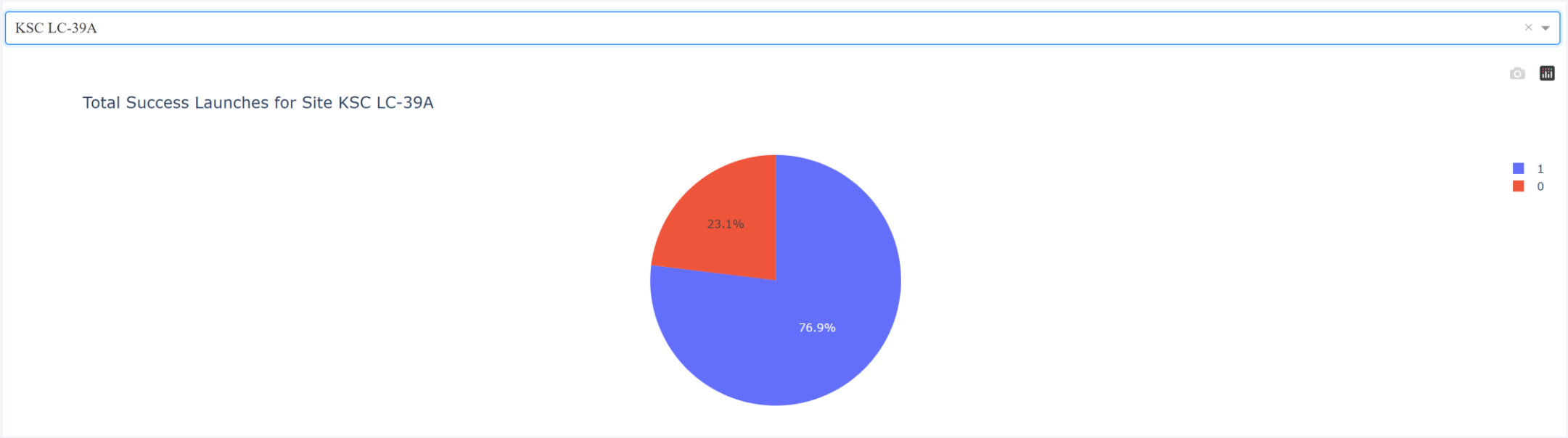
Launch success for all sites

- The pie chart below shows the total success launches by site.
- From all the launch sites, KSC LC-39A has the most successful launches.



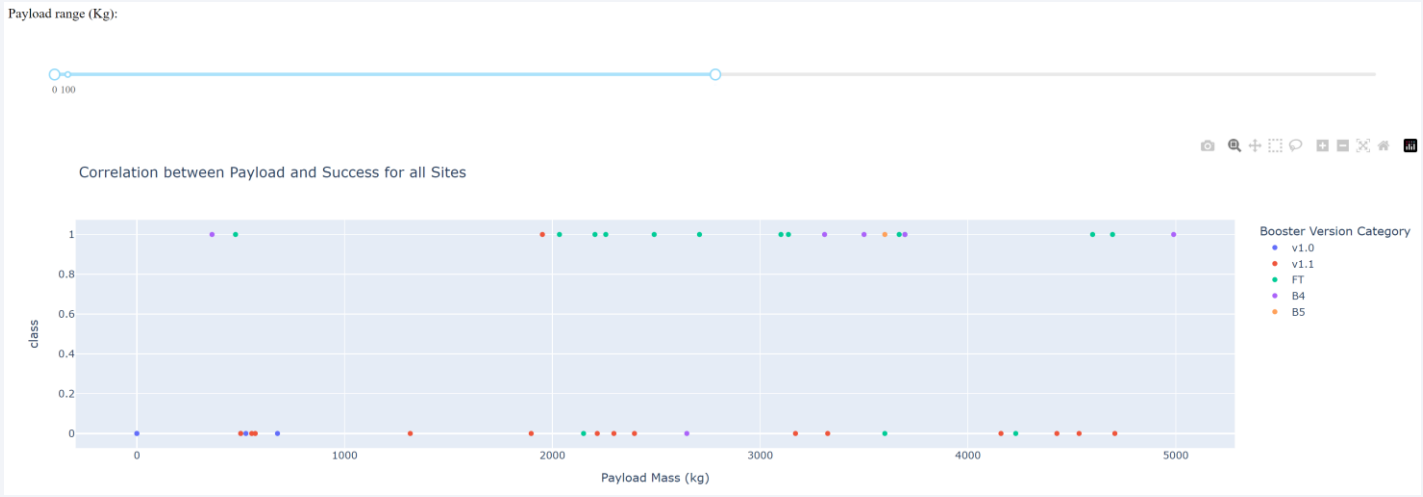
Launch site with highest success ratio

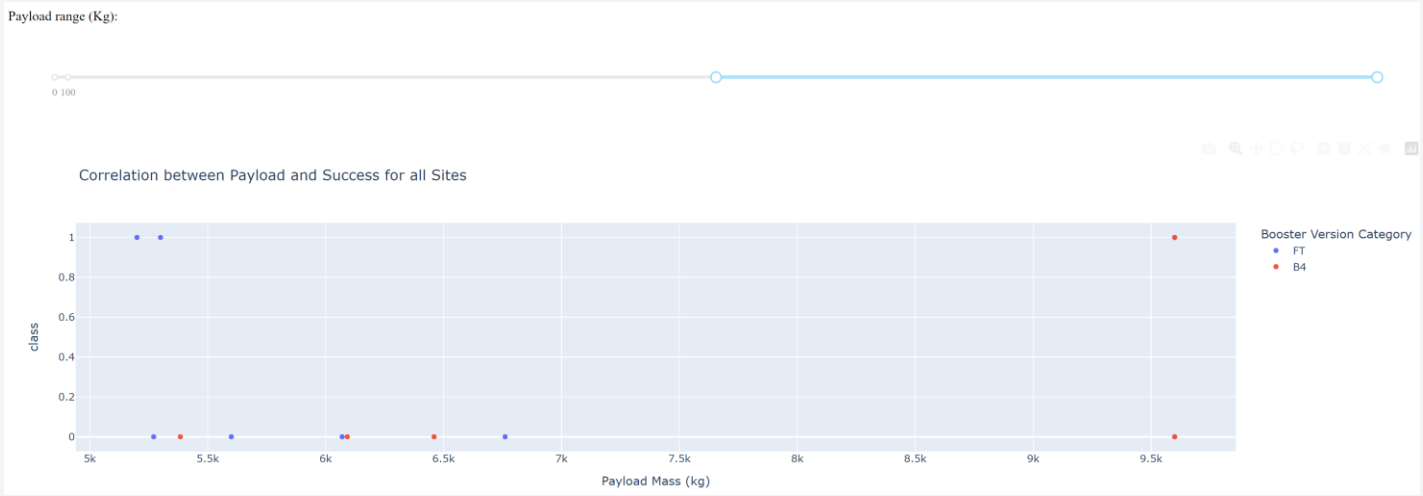
- As mentioned before, launch site KSC LC-39A has the most successful launches, with 76.9% successful landings, and only 23.1% failed landings.



Payload vs Launch Outcome

- The charts show that the lower ranged payloads have a higher success rate compared to heavier payloads.





Section 5

Predictive Analysis (Classification)

Classification Accuracy

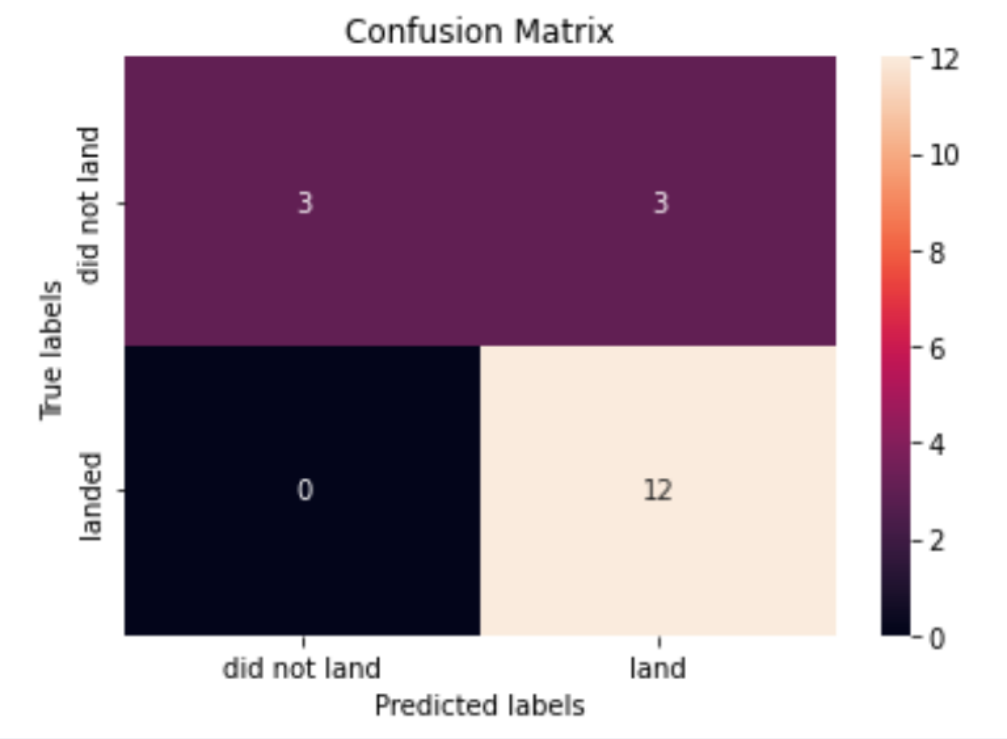
- Based on the scores we cannot predict which model performs best.
- For this reason, the scores were calculated for the whole dataset, and in this case, it is confirmed that the best model for the prediction is the decision tree model.

Scores and accuracy for the test set:

	LogReg	SVM	Tree	KNN
Jaccard Score	0.800000	0.800000	0.846154	0.800000
F1 Score	0.888889	0.888889	0.916667	0.888889
Accuracy	0.866667	0.877778	0.877778	0.855556

Confusion Matrix

- The confusion matrices are identical for the used models.
- The main problem with these models is the detection of false positives.



Conclusions

- The earlier flight were more likely to fail, while the flights later in time were more likely to succeed.
- Some launch sites have a higher number of flights; however, others have a higher percentage of successful landings (KSC LC-39A).
- There are orbit types that have a success rate (close to) 100%, like ES-L1, GEO, HEO, SSO. Some orbits require a low payload mass for a successful landing, while others can handle both capacities.
- Most of the launching sites are next to the coastline in the US.
- The success rate of the landing outcome improved steadily over the years.

Thank you!