IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Salvatore Baglieri
31/03/2023

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data Collection
  - Data Wrangling
  - EDA (Exploratory Data Analysis)
  - Interactive Visual Analytics
  - Predictive Analysis
- **Summary of all results**
  - EDA results
  - Geospatial analytics
  - Interactive dashboard
  - Predictive analysis of classification models

Introduction

Background

SpaceX launches Falcon 9 rockets at a cost of around $62m.

This is considerably cheaper than other providers (which usually cost upwards of $165m), and much of the savings are because SpaceX can land, and then re-use the first stage of the rocket.

If we can make predictions on whether the first stage will land, we can determine the cost of a launch, and use this information to assess whether an alternate company should bid and SpaceX for a rocket launch.

Problem

Train a machine learning model to predict successful Stage 1 recovery.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

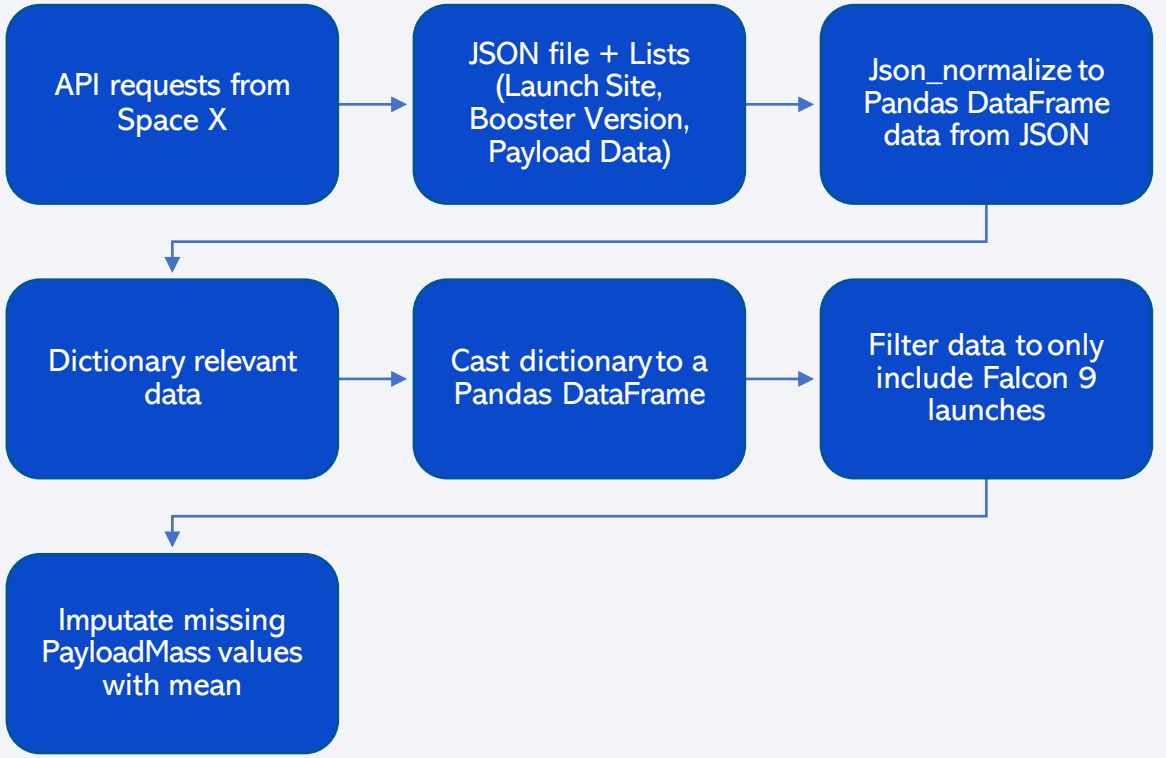
Data Collection

Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.

The next slide will show the flowchart of data collection from API and the one after will show the flowchart of data collection from webscraping.

Data Collection – SpaceX API

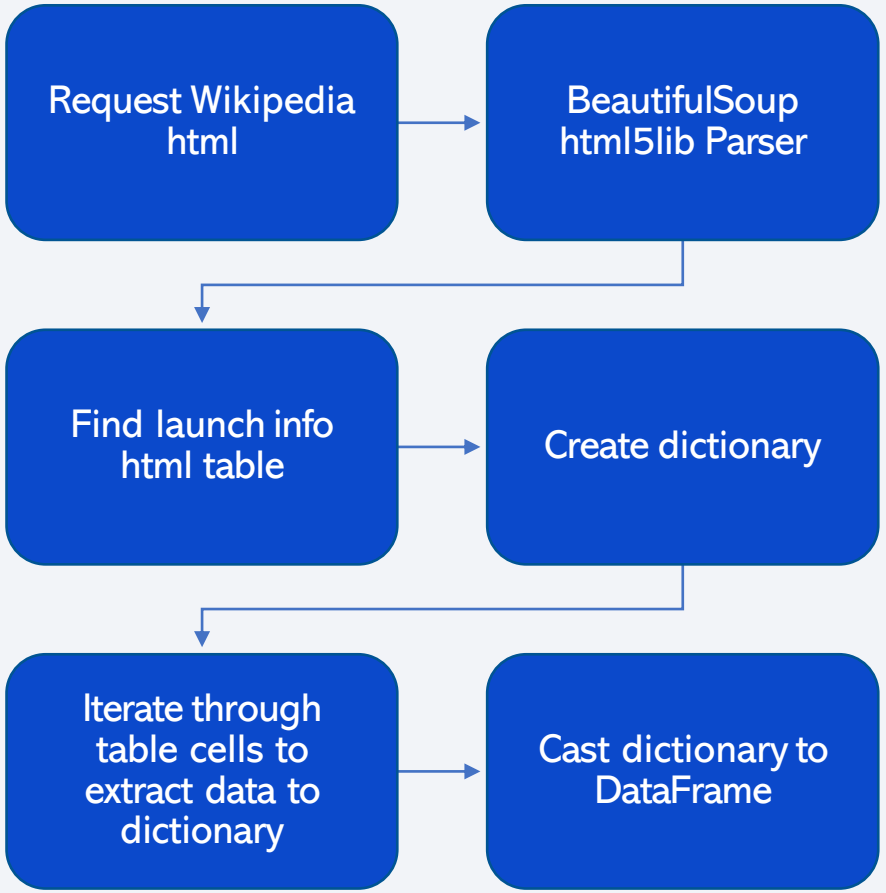
Using the SpaceX API to retrieve data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.



GitHub URL

Data Collection - Scraping

Using the SpaceX API to retrieve data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.



GitHub URL

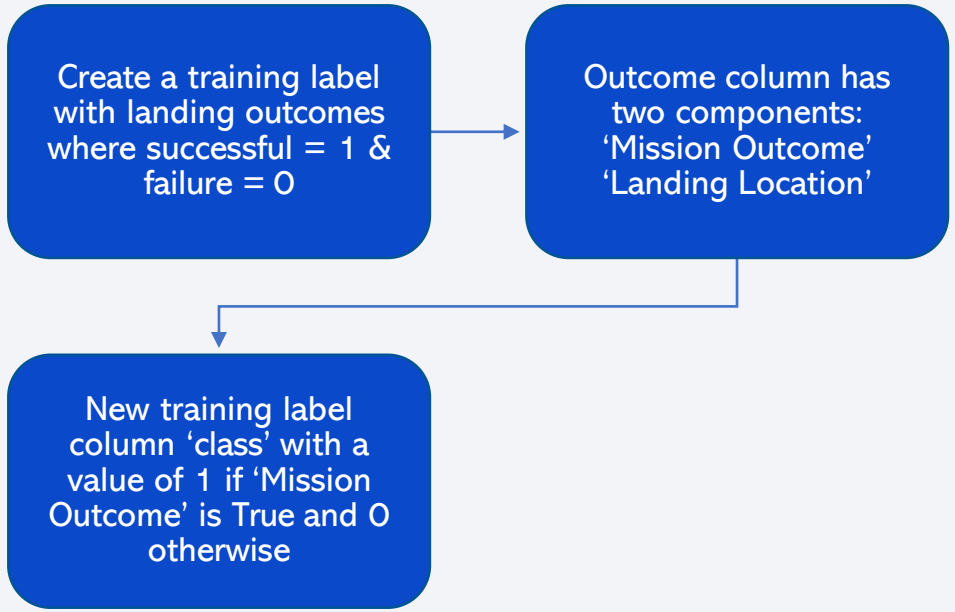
Data Wrangling

The SpaceX dataset contains several Space X launch facilities, and each location is in the *LaunchSite* column.

Each launch aims to a dedicated orbit, and some of the common orbit types are shown in the figure below. The orbit type is in the *Orbit* column.

The landing outcome is shown in the Outcome column:

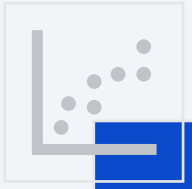
- True Ocean – the mission outcome was successfully landed to a specific region of the ocean
- False Ocean – the mission outcome was unsuccessfully landed to a specific region of the ocean.
- True RTLS – the mission outcome was successfully landed to a ground pad
- False RTLS – the mission outcome was unsuccessfully landed to a ground pad.
- True ASDS – the mission outcome was successfully landed to a drone ship
- False ASDS – the mission outcome was unsuccessfully landed to a drone ship.
- None ASDS and None None – these represent a failure to land.



GitHub URL

EDA with Data Visualization

Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

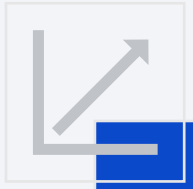


SCATTER PLOT

- Scatter charts were produced to visualizethe relationships between:
  - Flight Number and Launch Site
  - Payload and Launch Site
  - Orbit Type and Flight Number
  - Payload and Orbit Type
- Scatter charts are useful to observe relationships, or correlations, between two numeric variables.

BAR CHART

- A bar chart was produced to visualize the relationship between:
  - Success Rate and Orbit Type
- Bar charts are used to compare a numerical value to a categorical variable. Horizontal or vertical bar charts can be used, depending on the size of the data.



LINE CHART

- Line charts were produced to visualize the relationships between:
  - Success Rate and Year (i.e. the launch success yearly trend)
- Line charts contain numerical values on both axes, and are generally used t show the change of a variable over time.

GitHub URL

EDA with SQL

The SQL queries performed on the data set were used to:

- 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 the average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome on a ground pad was achieved
- List the names of the boosters which had success on a drone ship and a payload mass between 4000 and 6000 kg
- List the total number of successful and failed mission outcomes
- List the names of the booster versions which have carried the maximum payload mass
- List the failed landing outcomes on drone ships, their booster versions, and launch site names for 2015
- 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

GitHub URL

Build an Interactive Map with Folium

Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.

This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.

GitHub URL

Build a Dashboard with Plotly Dash

Dashboard includes a pie chart and a scatter plot.

Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.

Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.

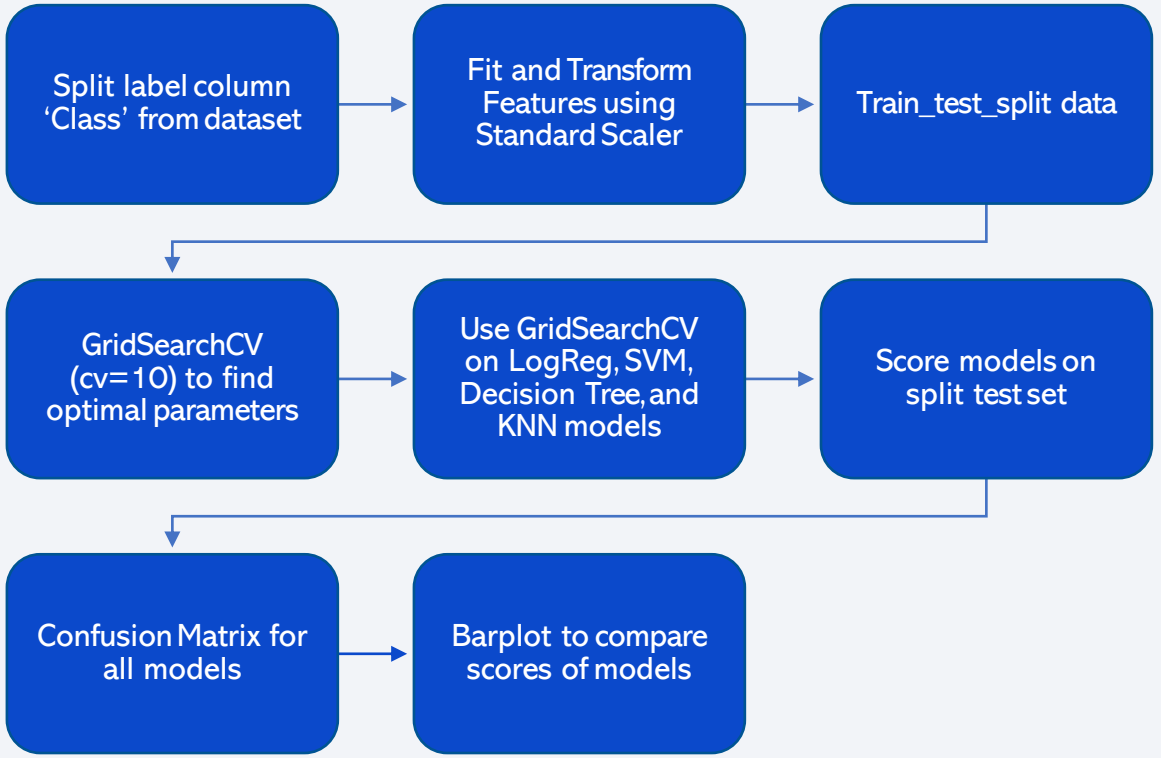
The pie chart is used to visualize launch site success rate.

The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category.

GitHub URL

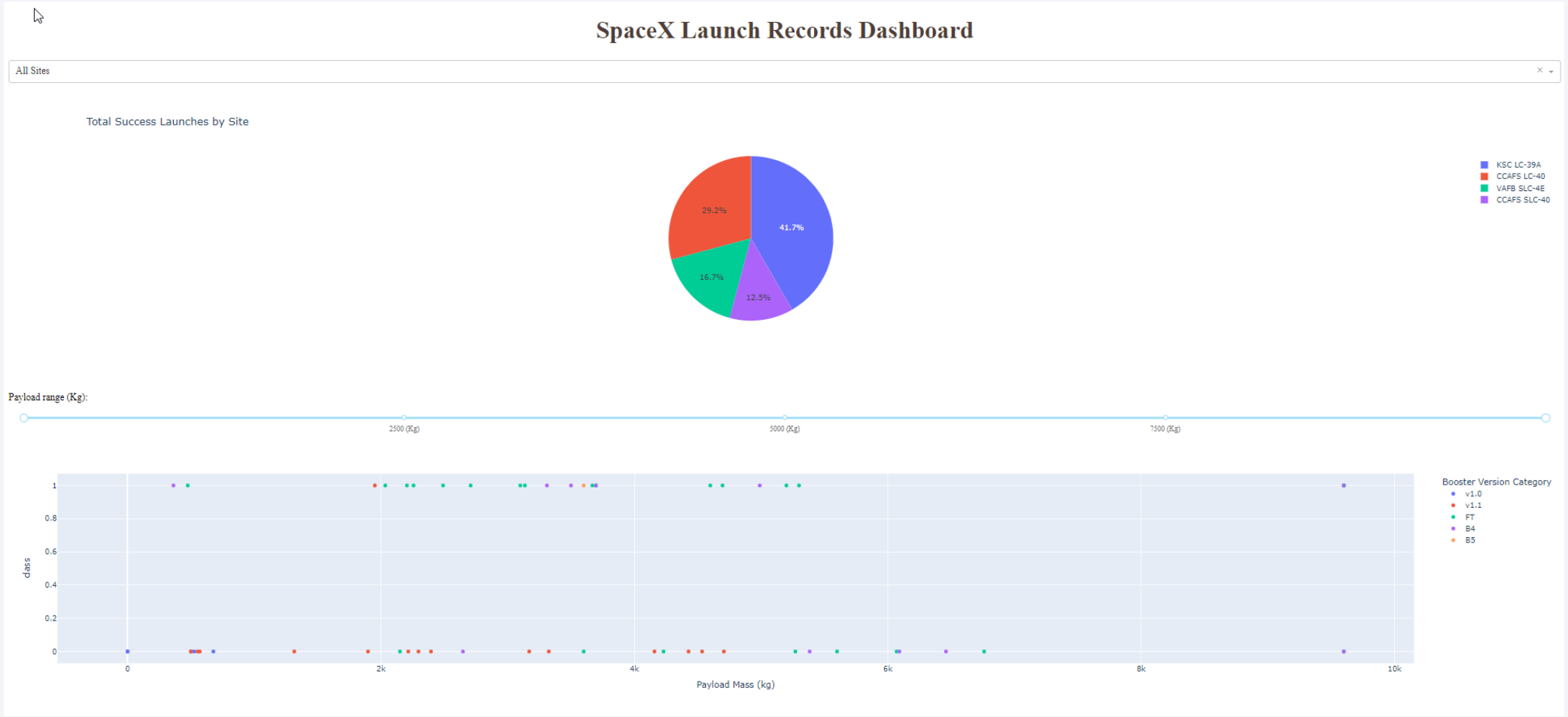
Predictive Analysis (Classification)

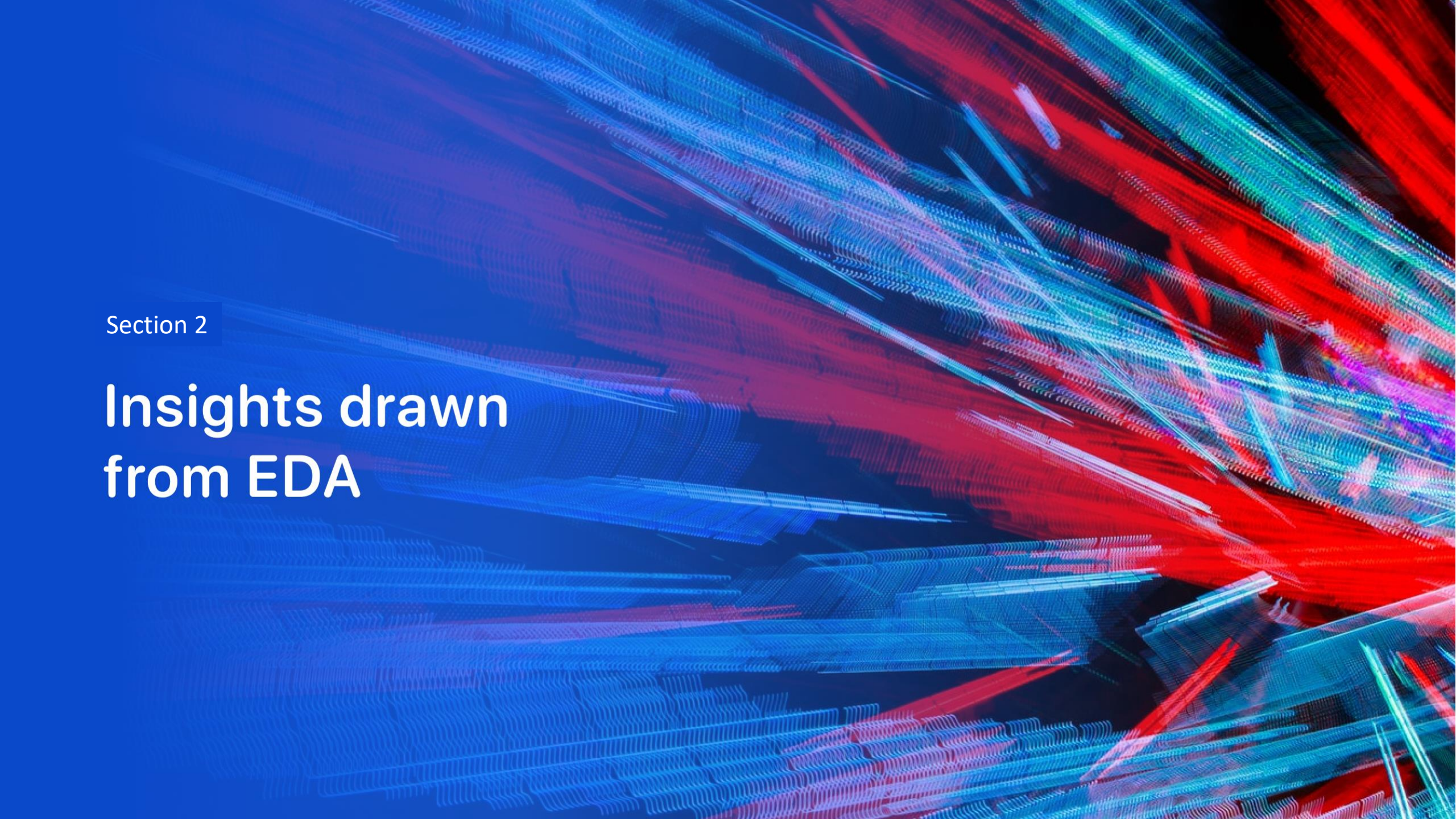
The following steps were taking to develop, evaluate, and find the best performing classification model



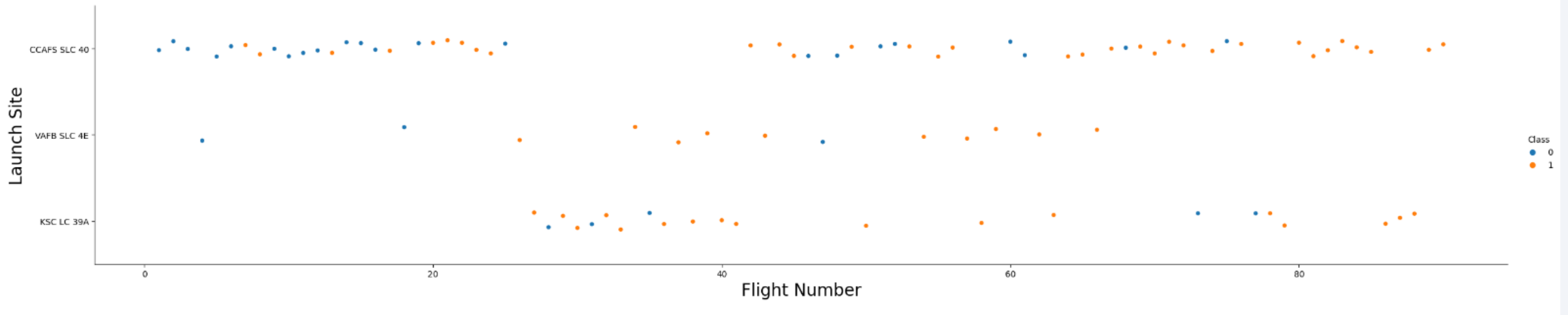
GitHub URL

Results


SpaceX Launch Records Dashboard
All Sites
Total Success Launches by Site
29.2%
41.7%
16.7%
12.5%
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
Payload range (Kg):
2500 (Kg)
5000 (Kg)
7500 (Kg)
class
Payload Mass (kg)
Booster Version Category
v1.0
v1.1
FT
B4
B5

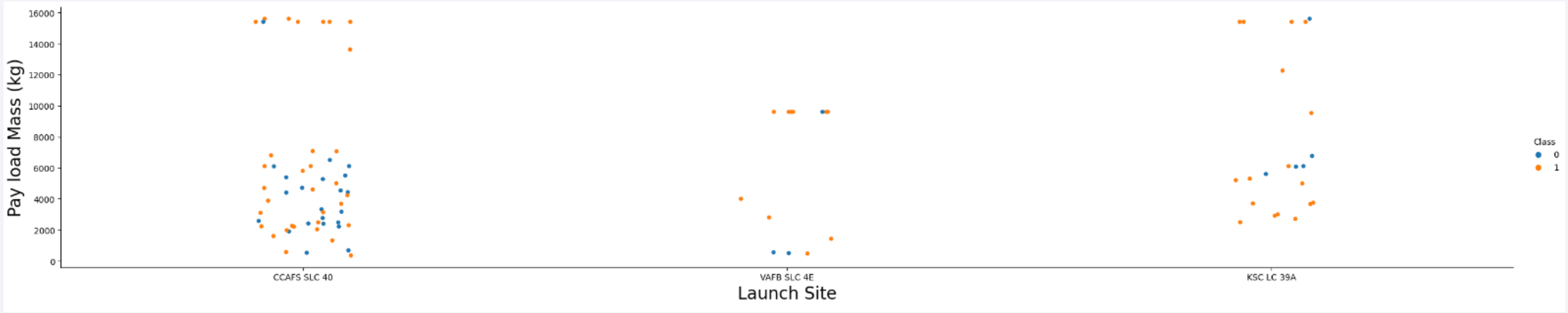

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



Graphic suggests an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appears to be the main launch site as it has the most volume.

Payload vs. Launch Site



Payload mass appears to fall mostly between 0-6000 kg. Different launch sites also seem to use different payload mass.

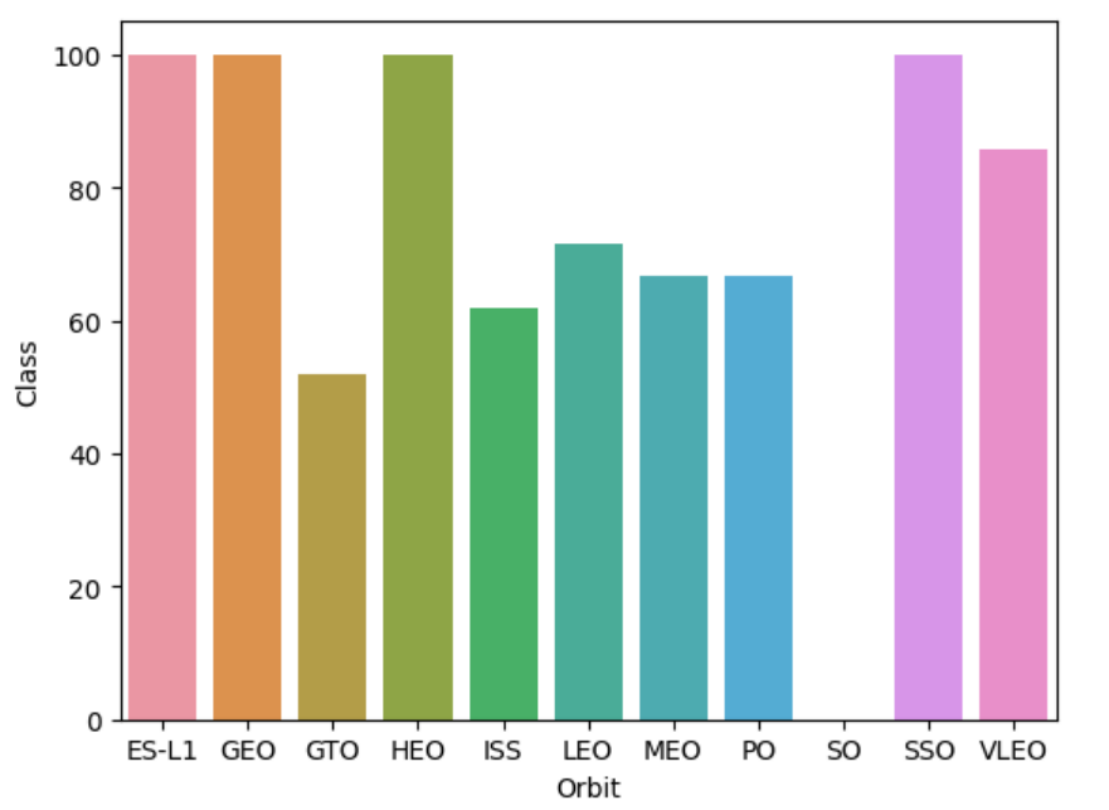
Success Rate vs. Orbit Type

The bar chart of Success Rate vs. Orbit Type shows that the following orbits have the highest (100%) success rate:

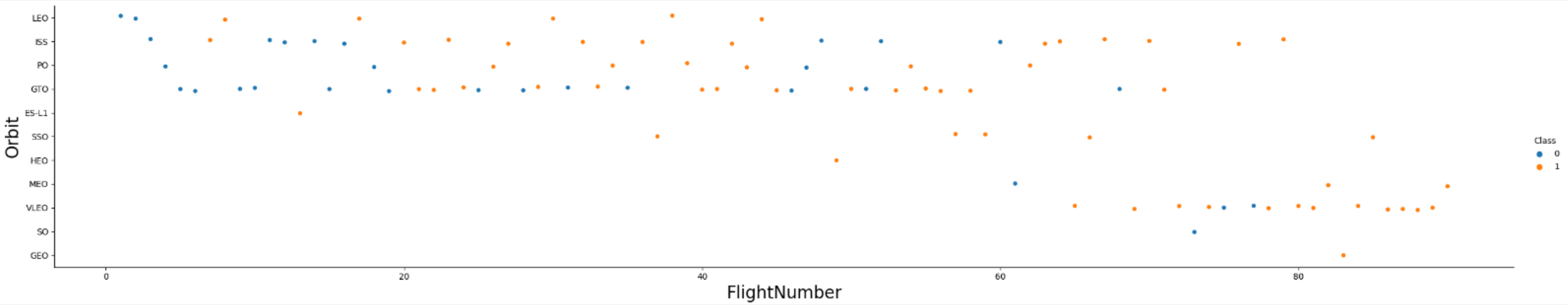
- ES-L1 (Earth-Sun First Lagrangian Point)
- GEO (Geostationary Orbit)
- HEO (High Earth Orbit)
- SSO (Sun-synchronous Orbit)

The orbit with the lowest (0%) success rate is:

- SO (Heliocentric Orbit)



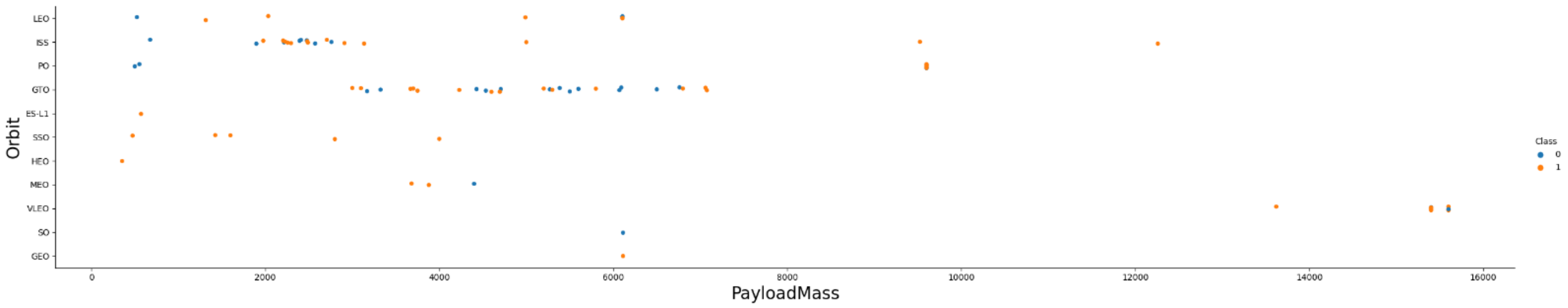
Flight Number vs. Orbit Type



Launch Orbit preferences changed over Flight Number. Launch Outcome seems to correlate with this preference.

SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

Payload vs. Orbit Type



Payload mass seems to correlate with orbit

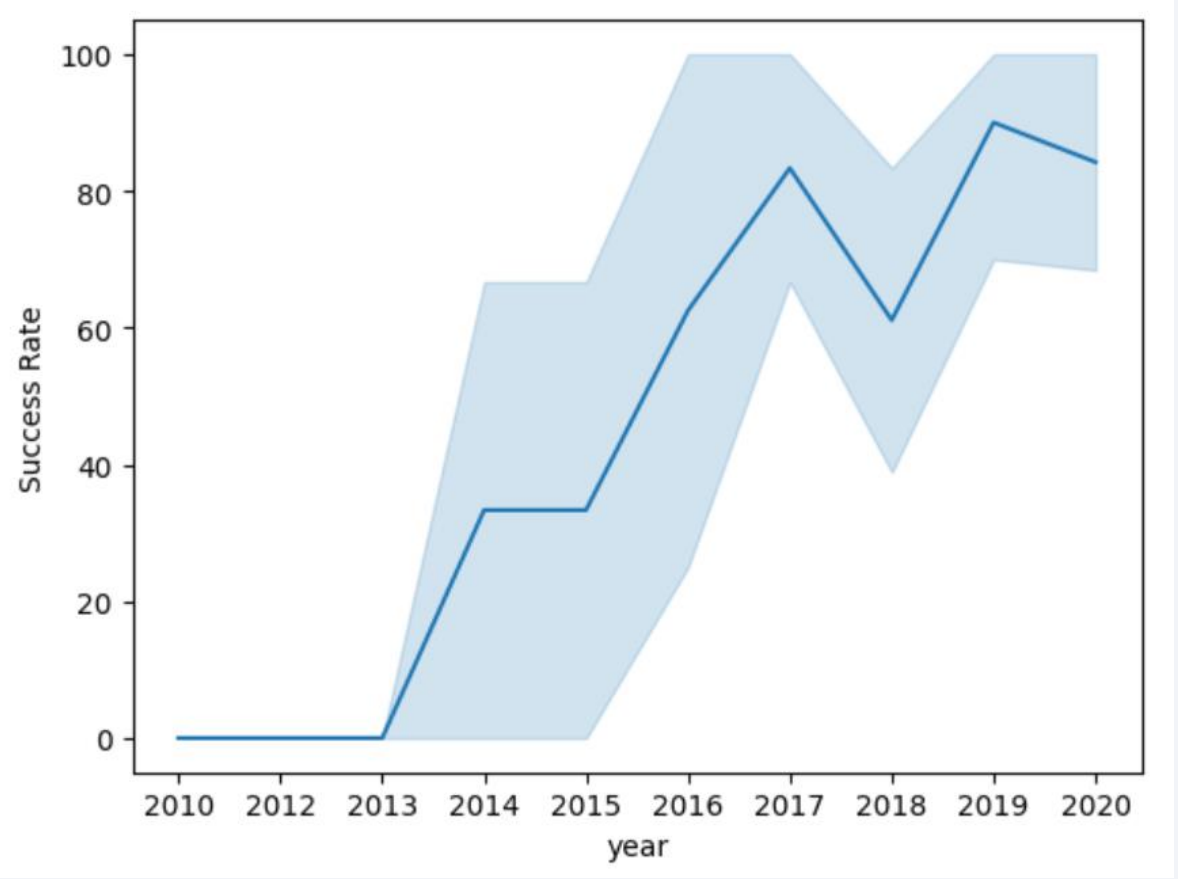
LEO and SSO seem to have relatively low payload mass

The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch Success Yearly Trend

The line chart of yearly average success rate shows that:

- Between 2010 and 2013, all landings were unsuccessful (as the success rate is 0).
- After 2013, the success rate generally increased, despite small dips in 2018 and 2020.
- After 2016, there was always a greater than 50% chance of success.



All Launch Site Names

The word UNIQUE returns only unique values from the LAUNCH_SITE column of the SPACEXTBL table.

```
%sql select DISTINCT LAUNCH_SITE from SPACEXTBL
```


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

Launch Site Names Begin with 'CCA'

LIMIT 5 fetches only 5 records, and the LIKE keyword is used with the wild card ‘CCA%’ to retrieve string values beginning with ‘CCA’.

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

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 SUM keyword is used to calculate the total of the LAUNCH column, and the SUM keyword (and the associated condition) filters the results to only boosters from NASA (CRS).

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


total_payload_mass
45596

Average Payload Mass by F9 v1.1

The AVG keyword is used to calculate the average of the PAYLOAD_MASS__KG column, and the WHERE keyword (and the associated condition) filters the results to only the F9 v1.1 booster version.

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

Average
2534.6666666666665

First Successful Ground Landing Date

The MIN keyword is used to calculate the minimum of the DATE column, i.e. the first date, and the WHERE keyword (and the associated condition) filters the results to only the successful ground pad landings.

```
%sql select min(date) as First_successful_ground_landing from SPACEXTBL where `landing _outcome` = 'Success (ground pad)'
```


First_successful_ground_landing
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

The WHERE keyword is used to filter the results to include only those that satisfy both conditions in the brackets (as the AND keyword is also used). The BETWEEN keyword allows for 4000 < x < 6000 values to be selected.

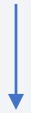
```
%sql select * from SPACEXTBL where (mission_outcome like 'Success') AND (payload_mass__kg_ BETWEEN 4000 AND 6000) AND (`landing _outcome` like 'Success (drone ship)')
```

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 COUNT keyword is used to calculate the total number of mission outcomes, and the GROUPBY keyword is also used to group these results by the type of mission outcome.

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



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

Boosters Carried Maximum Payload

A subquery is used here. The SELECT statement within the brackets finds the maximum payload, and this value is used in the WHERE condition. The DISTINCT keyword is then used to retrieve only distinct /unique booster versions.

```
maxm = %sql select max(payload_mass__kg_) from SPACEXTBL
maxv = maxm[0][0]
%sql select booster_version from SPACEXTBL where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEXTBL)
```

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 WHERE keyword is used to filter the results for only failed landing outcomes, AND only for the year of 2015.

```
%sql select substr(Date, 4, 2) as Month, `landing _outcome`, booster_version, launch_site from SPACEXTBL where DATE like '%2015' AND `landing _outcome` like 'Failure (drone ship)'
```


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

The WHERE keyword is used with the BETWEEN keyword to filter the results to dates only within those specified. The results are then grouped and ordered, using the keywords GROUP BY and ORDER BY, respectively, where DESC is used to specify the descending order.

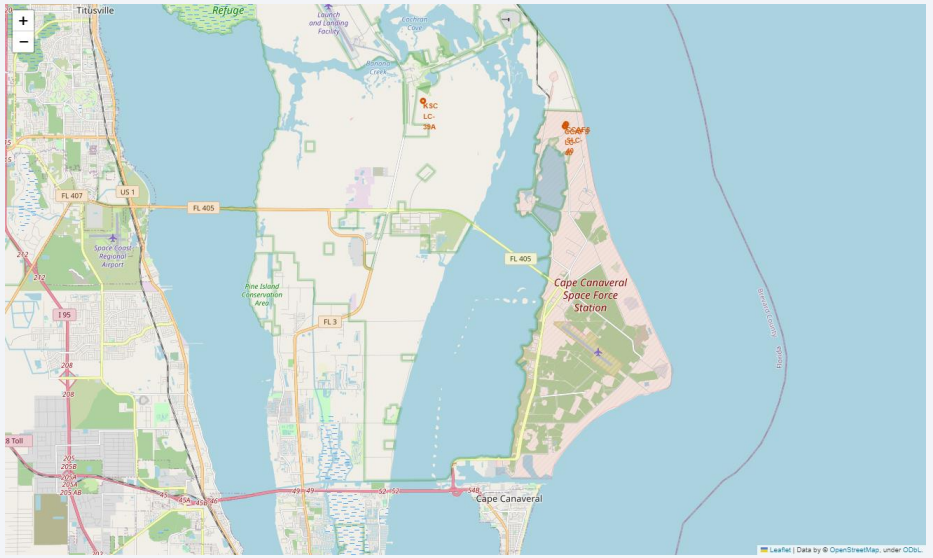
```
%sql select `landing _outcome`, count(*) as count from SPACEXTBL where Date >= '04-06-2010' AND Date <= '20-03-2017' GROUP by `landing _outcome` ORDER BY count Desc
```

Landing _Outcome	count
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

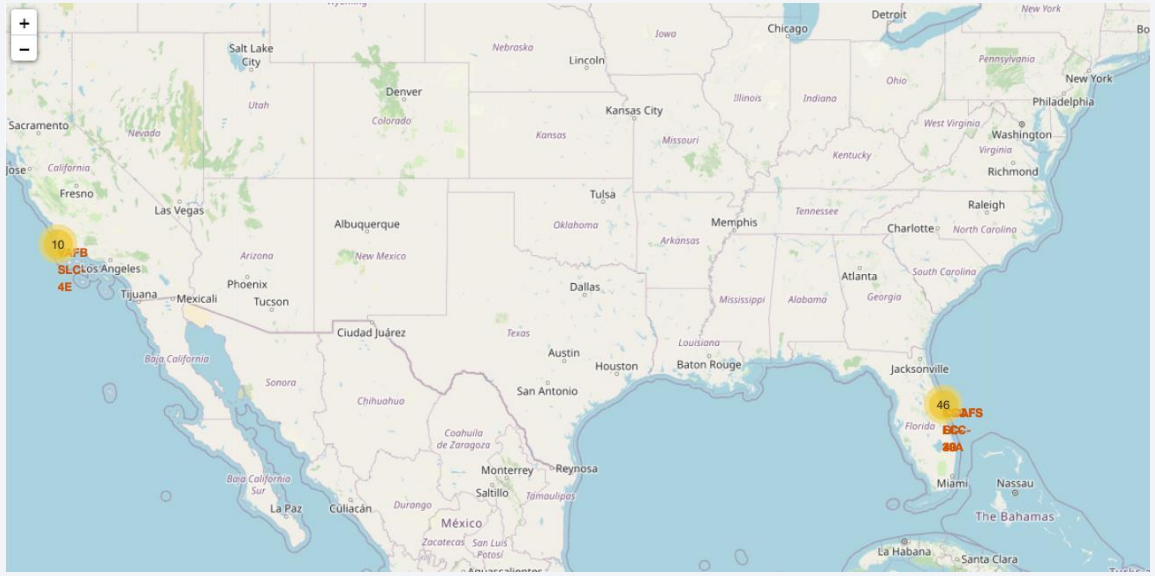
Launch Sites Proximities Analysis

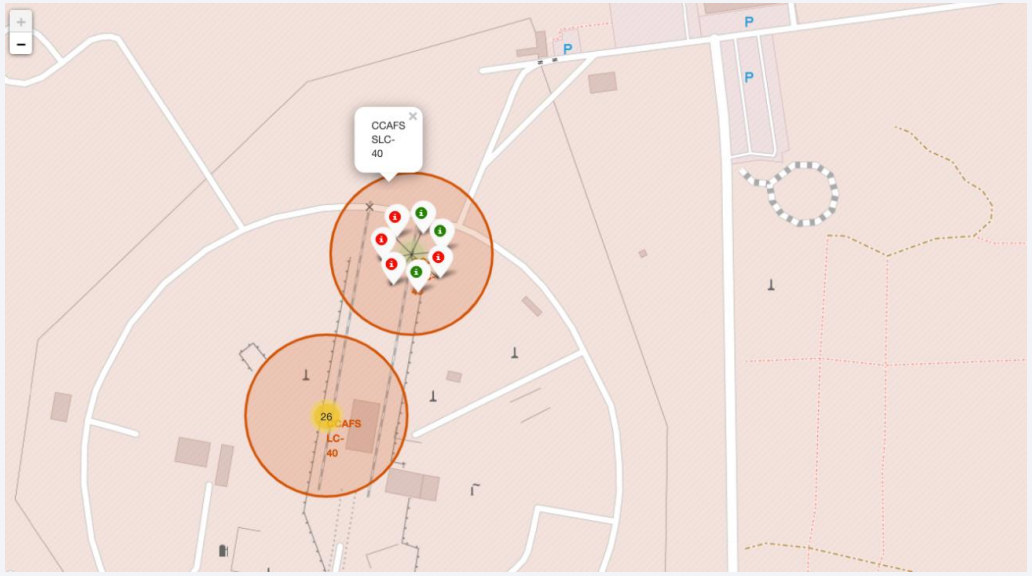
All launch sites on map



All SpaceX launch sites are on coasts of the United States of America, specifically Florida and California.

Success/Failed launches for each site

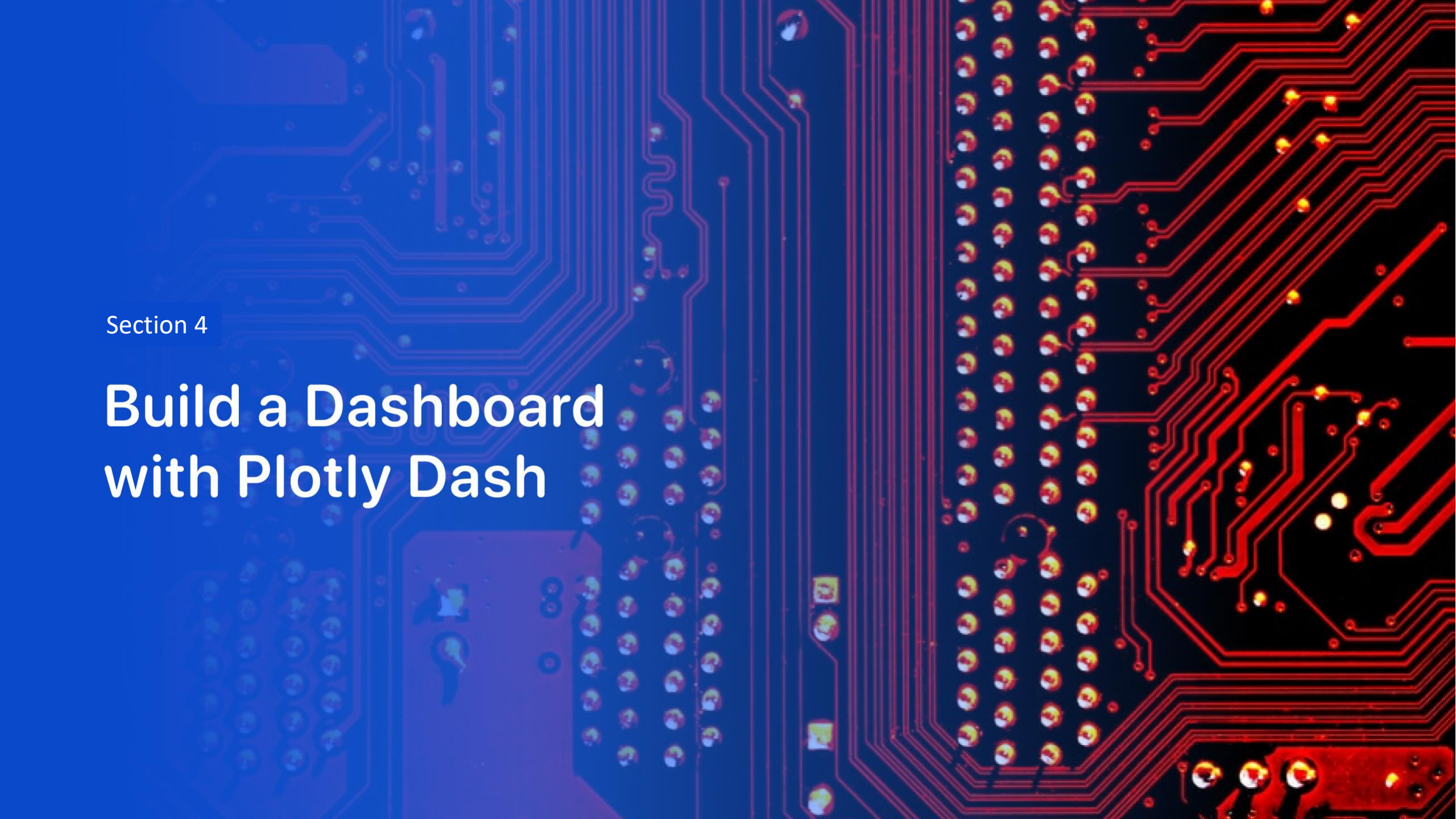




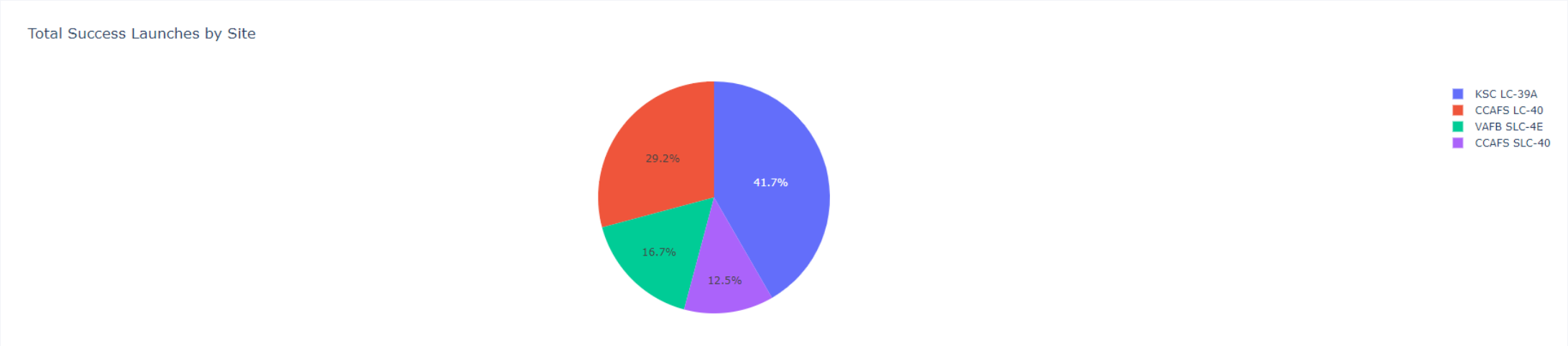
Launches have been grouped into clusters, and annotated with green icons for successful launches, and red icons for failed launches.

Proximity of launch sites to other points of interest

Using the CCAFS SLC-40 launch site as an example site, we can understand more about the placement of launch sites.

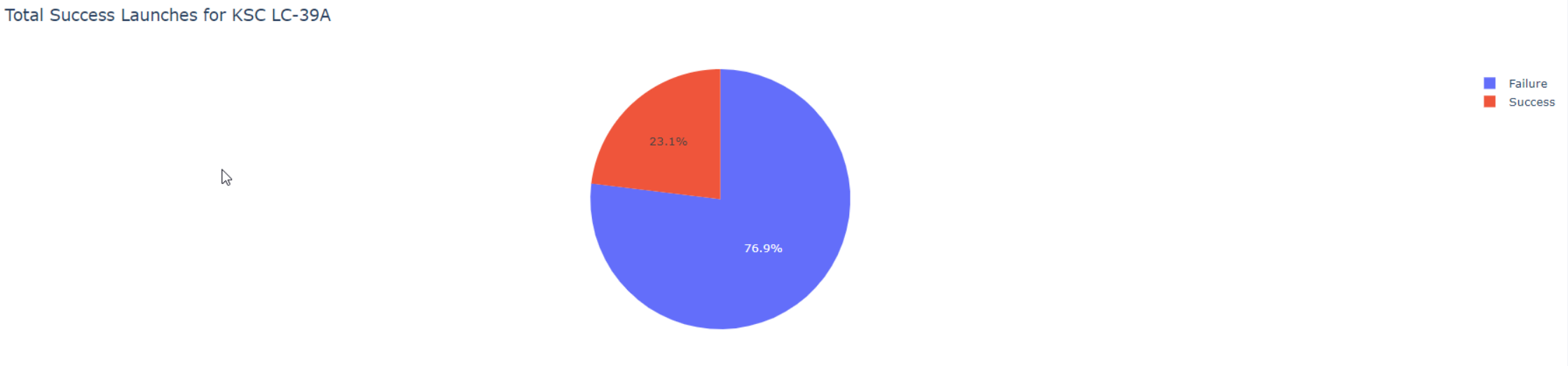

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



The launch site KSC LC-39 A had the most successful launches, with 41.7% of the total successful launches.

Highest success rate launch sites



The launch site KSC LC-39 A also had the highest rate of successful launches, with a 76.9% success rate.

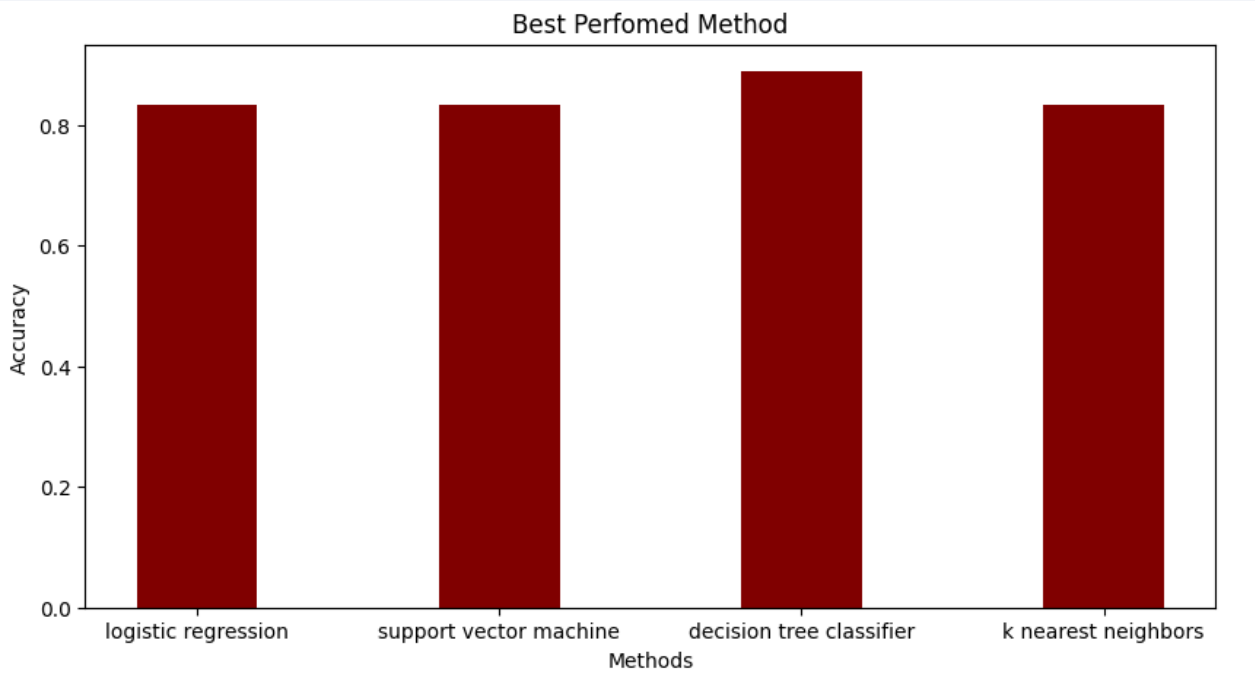
Launch outcome scatter plot for all sites

Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-6000, interestingly there are two failed landings with payloads of zero kg.

Section 5

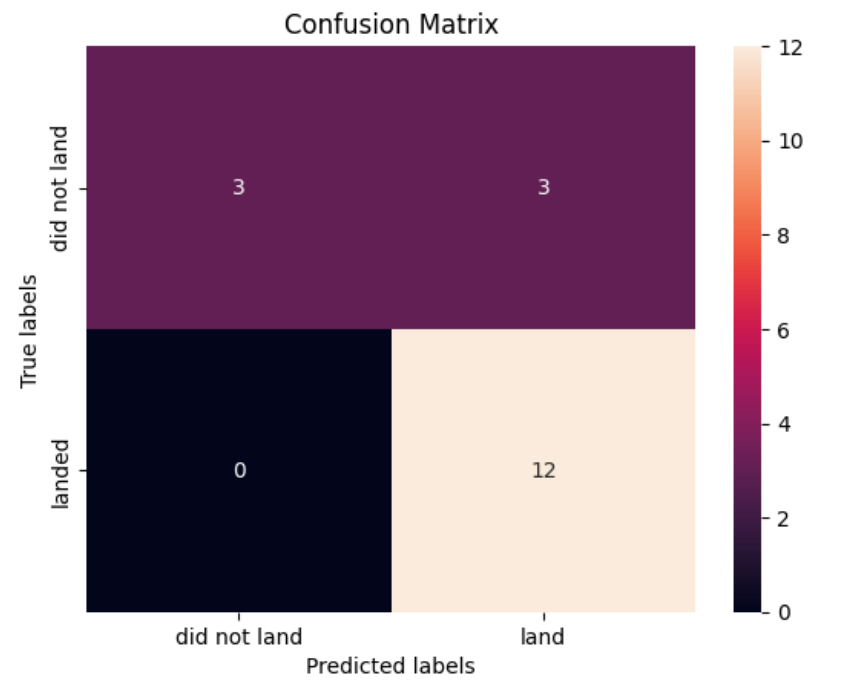
Predictive Analysis (Classification)

Classification Accuracy



Plotting the Accuracy Score and Best Score for each classification algorithm produces the following result: The Decision Tree model has the highest classification accuracy

Confusion Matrix



As shown previously, best performing classification model is the Decision Tree.

This is explained by the confusion matrix, which shows only 1 out of 18 total results classified incorrectly (a false positive, shown in the top-right corner).

The other 17 results are correctly classified (5 did not land, 12 did land).

Conclusions

- As the number of flights increases, the rate of success at a launch site increases, with most early flights being unsuccessful. I.e. with more experience, the success rate increases.
  - Between 2010 and 2013, all landings were unsuccessful (as the success rate is 0).
  - After 2013, the success rate generally increased, despite small dips in 2018 and 2020.
  - After 2016, there was always a greater than 50% chance of success.
- Orbit types ES-L1, GEO, HEO, and SSO, have the highest (100%) success rate.
  - The 100% success rate of GEO, HEO, and ES-L1 orbits can be explained by only having 1 flight into the respective orbits.
  - The 100% success rate in SSO is more impressive, with 5 successful flights.
  - The orbit types PO, ISS, and LEO, have more success with heavy payloads:
  - VLEO (Very Low Earth Orbit) launches are associated with heavier payloads, which makes intuitive sense.
- The launch site KSC LC-39 Ahad the most successful launches, with 41.7% of the total successful launches, and also the highest rate of successful launches, with a 76.9% success rate.
- The success for massive payloads (over 4000kg) is lower than that for low payloads.
- The best performing classification model is the Decision Tree model.

Appendix

GitHub repository URL

Thank you!