IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anca Zamfirescu
7th June 2024

Outline

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

Executive Summary

- The purpose of this project is to find a way to predict as accurately as possible whether Falcon9 first stage will land successfully after launches.
- To this end, data on previous Falcon 9 launches was collected from two sources: SpaceX, via REST API calls and Wikipedia, using webscraping.'
- Exploratory analysis, conducted using SQL analyses and visualization techniques, indicated several characteristics of the launches which can serve to predict the success or failure of the first stage landing.
- The analysis shows that success rates increased in time, thus booster version categories are good predictors of success. Also, launch sites, orbits and payload mass are somewhat correlated with each other and also good predictors. Other predictors are the physical characteristics of the core and the number of times it was reused.
- The success or failure of first stage landings of Falcon 9 launches can be predicted with reasonable accuracy (0.83), Successes can be quite accurately determined, however failures are harder to predict. The tested models yield 50% false positives.
- Three classification models proved to be equaly suitable for the prediction: Logistic Regression, SVM and K-Nearest Neighbours.

Introduction

- SpaceX offers Falcon 9 rocket launches at a significantly lower cost compared to its competitors: 62 million dollars versus as much as 165 million dollars asked by other companies for each launch. The smaller cost owes to the fact that SpaceX can reuse the first stage, if it has landed successfully.
- Therefore, Falcon 9 launch costs are highly dependent on the success of the first stage landing.
- The purpose of this project is to find a way to predict as accurately as possible whether Falcon9 first stage will land successfully.
- Key objectives:
  - Find correlations between features of the rockets and launches and the success/failure of the landing
  - Create a machine learning model that can predict landing success or failure based on the most relevant features

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from two sources: SpaceX, via REST API calls and Wikipedia, using webscraping
- Perform data wrangling
  - Dealt with missing values, checked data types and ran initial summaries to get familiarized with the data and determine suitable training labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - After pre-processing the data and created the training and test sets, I split the data for independent variables (X) and dependent (Y) into training and test data.
  - Several models were created and evaluated. Also, several parameters were tested for each model using GridSearchCV.

Data Collection

- Data was collected in two ways:

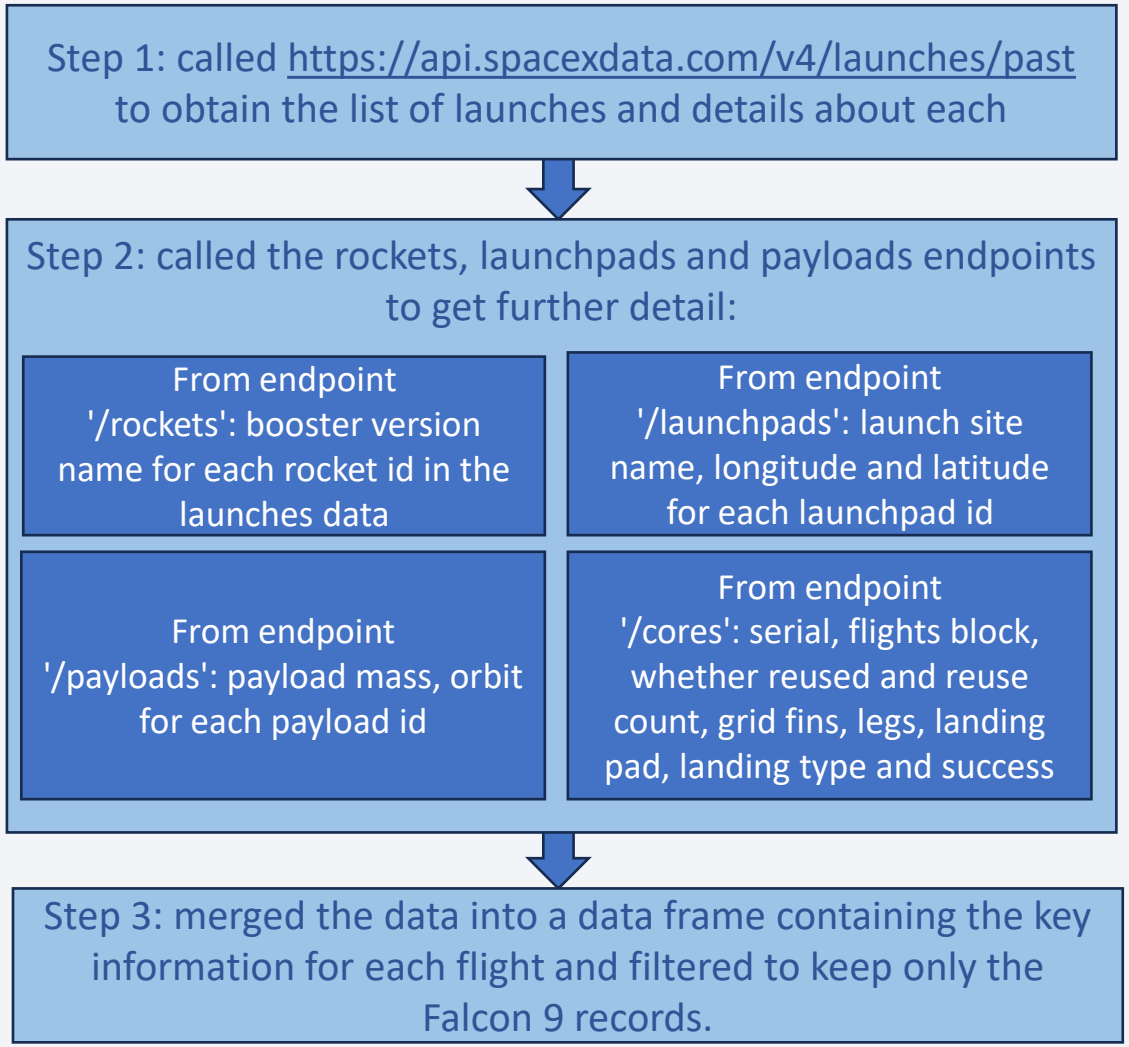
API call SpaceX API https://api.spacexdata.com/v4/	Webscraping Wikipedia, page *List of Falcon 9 and Falcon Heavy launches* https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- Key data collected:
  - information on the launch site (including longitude and latitude) and launchpad,
  - rocket, booster version, cores data (with or without legs, how many times it was reused, landing type and landing pad), existence of grid fins.
  - payload and orbit
  - success or failure of the landing

Data Collection – SpaceX API

- First data source is SpaceX API:
  - Base URL: https://api.spacexdata.com/v4/
  - Endpoints:
    - launches/past
    - rockets
    - launchpads
    - payloads
    - cores
- REST API calls notebook can be found here:

  https://github.com/ancaz/IBM-DS-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



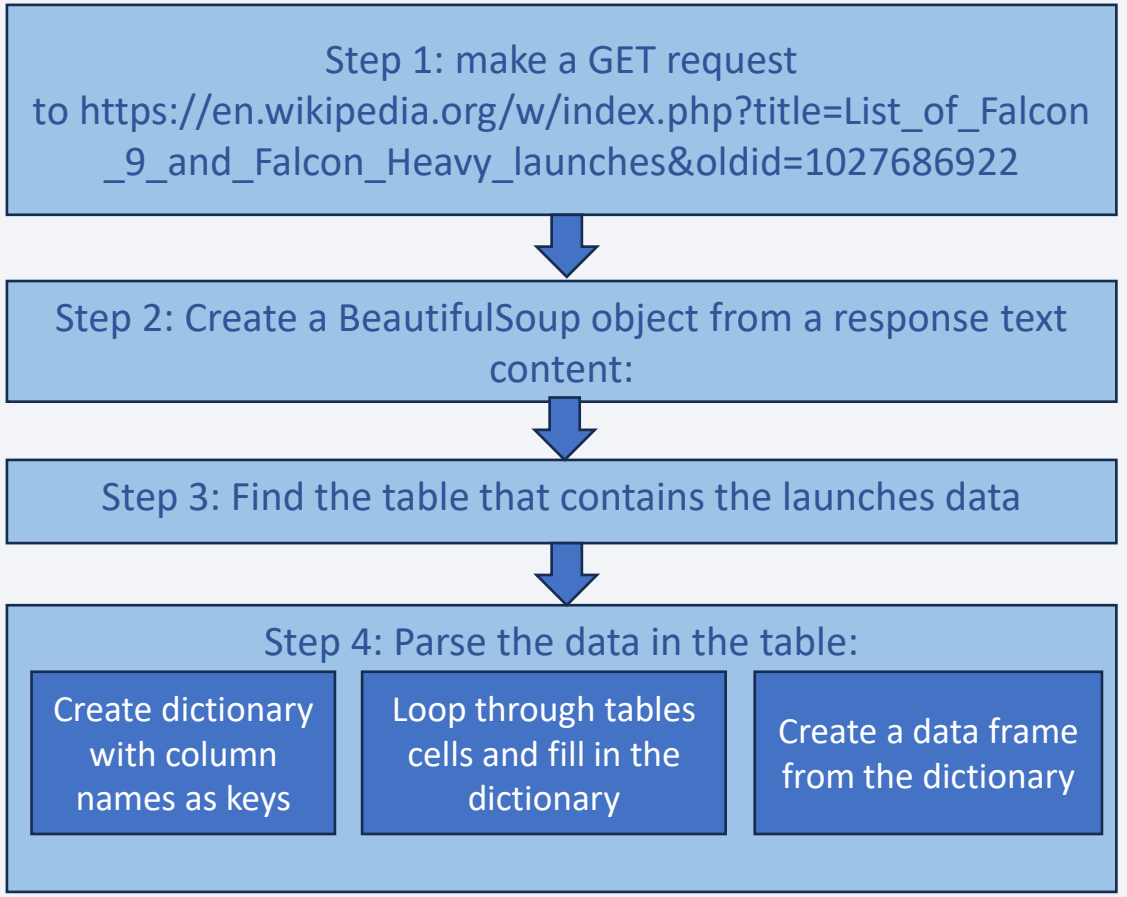
Data Collection - Scraping

- Second data source is Wikipedia page List of Falcon 9 and Falcon Heavy launches
- I extracted the Falcon 9 launch records HTML table from

  https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

  using BeautifulSoup
- Then parsed the table and convert it into a Pandas data frame
- Web scraping notebook can be found here:

  https://github.com/ancaz/IBM-DS-Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Dealt with missing values
  - We have 5 missing values for the payload mass in the dataset, which I replaced with the mean of the existing values
  - There are 26 missing values for landing pad, where landing pads were not used, and it makes sense to keep them as they are.
- Checked data types
- Ran initial summaries to get familiarized with the data and determine suitable training labels:
  - Calculated the number of launches on each site, number and occurrence of each orbit, number and occurrence of mission outcome of the orbits
  - Created a landing outcome label from Outcome column
- The data wrangling notebook can be found here:
  - https://github.com/ancaz/IBM-DS-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Charts created in order to identify relationships between features and the success/ insuccess of launches:
  - Chart 1: scatter plot of flight number vs. payload mass, overlaying the outcome of the launch – in order to gauge the connection between flight number (indicating that they are earlier or later flights) and success, also in relation to payload mass.
  - Charts 2 and 3: scatter plots of flight number vs. launch site, and payload mass vs. launch site, respectively, overlaying the outcome of the launch – in order to gauge whether particular launch sites seem to have different success rates, in time and in relation to the payload mass.
  - Chart 4: bar chart of success rates by orbit – to identify any differentiation
  - Charts 5 and 6: scatter plot of flight number vs. orbit, payload mass vs. orbit, respectively, overlayed with the launch outcome – in order to check how success for each orbit has improved with subsequent flights and in the context of different payload mass.
  - Chart 7: line chart showing success over time
- Selected features based on the analysis: flight number, payload mass, orbit, launch site, number of flights, presence of grid fins and legs, whether reused, number of reuses, landing pad, block, serial
- EDA with data visualization notebook can be found here:
  - https://github.com/ancaz/IBM-DS-Capstone/blob/main/edadataviz.ipynb

EDA with SQL

- SQL queries performed:
  - Names of the unique launch sites
  - 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
  - Rotal number of successful and failure mission outcomes
  - Names of the booster versions which have carried the maximum payload mass
  - Records for the months in year 2015 displaying month names, failure landing outcomes in drone ship , booster versions, launch site
  - Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.
- Notebook of completed EDA with SQL can be found here:
  - https://github.com/ancaz/IBM-DS-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

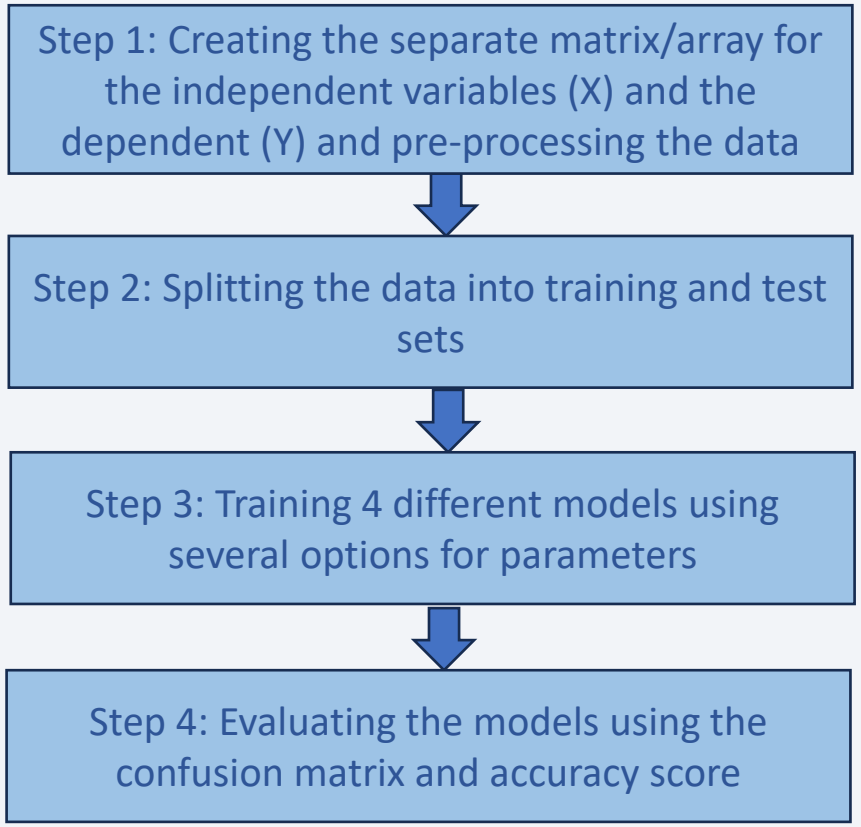
- To have a clearer image of the launch sites, all were represented on a map using Folium
  - Sites were represented using markers on their coordinates and 1000 radius circles around them
  - Successes and failures were marked for each site using marker clusters
  - Markers and poly lines were used to show distances from a launch site to its proximities. This way, we can gauge how close launch sites are to railways, highways, coastlines and cities
- The notebook with the completed interactive map with Folium map can be found here. **The file needs to be 'trusted' so that the interactive maps can be shown.**
  - https://github.com/ancaz/IBM-DS-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Built a dash app using Plotly Dash, that shows:
  - A pie chart showing the success of launches, controlled by a dropdown where the launch site can be chosen. When the all sites options is chose, the chart shows the structure of successes by site. When a site is chosen, the pie shows the percentage of successes and failures.
  - A scatter plot representing the relationship between payload mass and successes, in the context of the booster version category. The plot is controlled by the dropdown of launch sites and a slider where payload mass intervals can be chosen.
- Analyzing the data in this dashboard shows which launch sited had the largest number of successful launches and best success rate, which payload ranges have the highest and lowest success rates, and also which booster version categories have the highest success rate.
- The python script that creates the dashboard can be found here:
  - https://github.com/ancaz/IBM-DS-Capstone/blob/main/spacex_dash_app.py

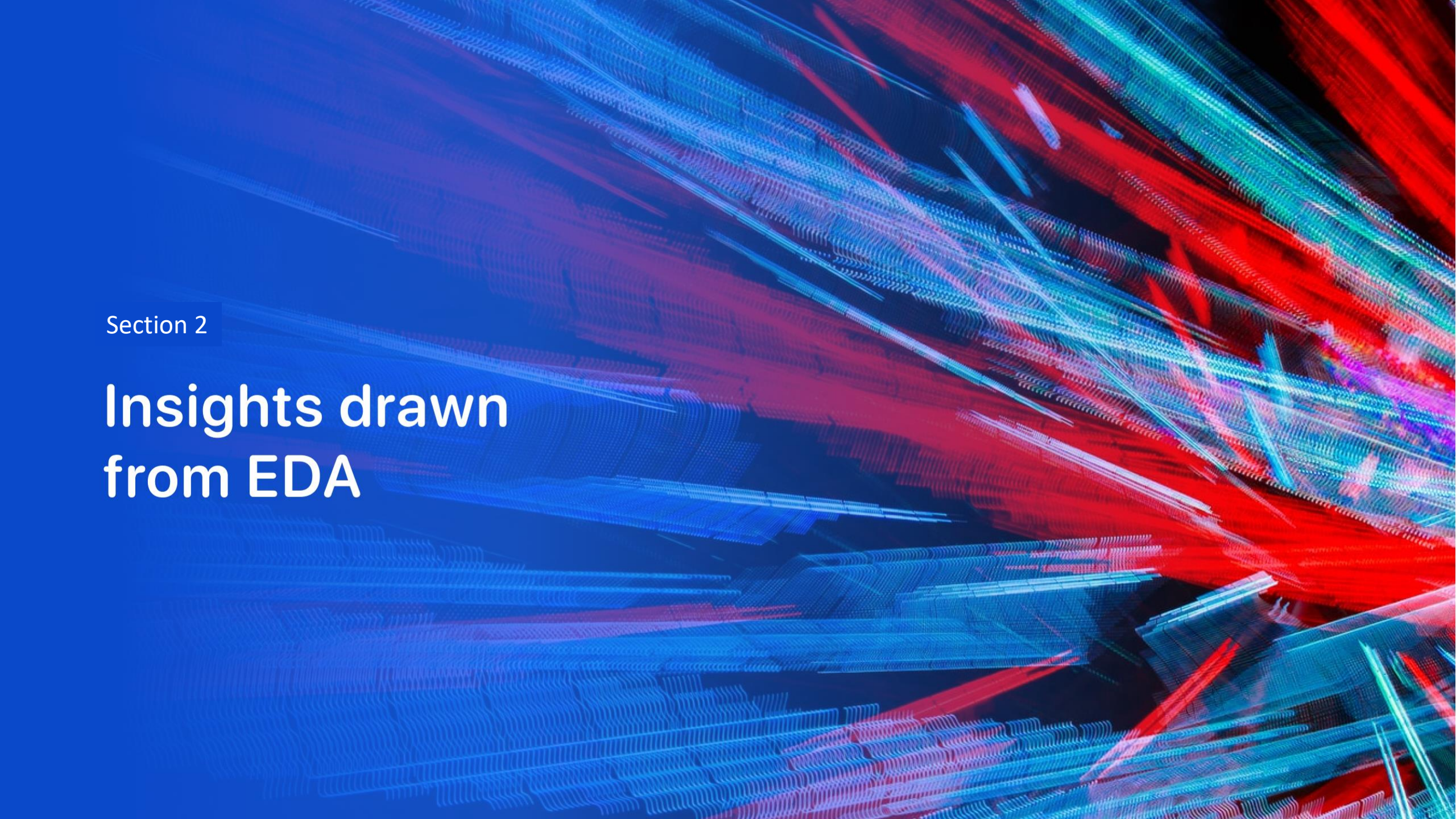
Predictive Analysis (Classification)

- Before starting to build the classification models, I pre-processed the data and created the training and test sets:
  - Created a NumPy array from the column Class in data (containing the success/ failure information, representing our dependent variable).
  - Standardized the data in the independent variables set using the StandardScaler function
  - Split the data for independent variables (X) and dependent (Y) into training and test data. Test data represents 20% of available data.
- Several models were created and evaluated. Also, several parameters were tested for each model using GridSearchCV.
  - Logistic regression
  - Support Vector Machine
  - Decision tree
  - K Nearest Neighbours
- The completed predictive analysis lab can be found here:
  - https://github.com/ancaz/IBM-DS-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



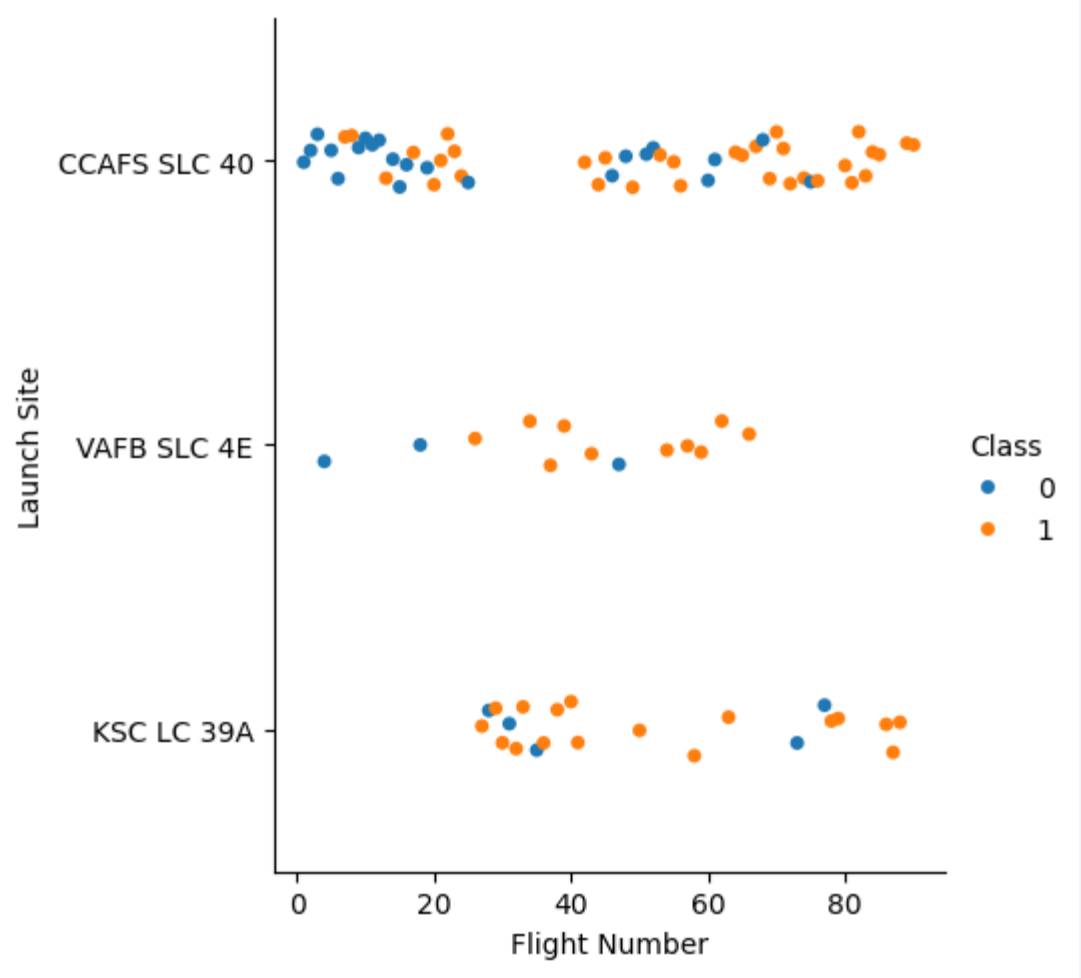
Results

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


Section 2
Insights drawn from EDA

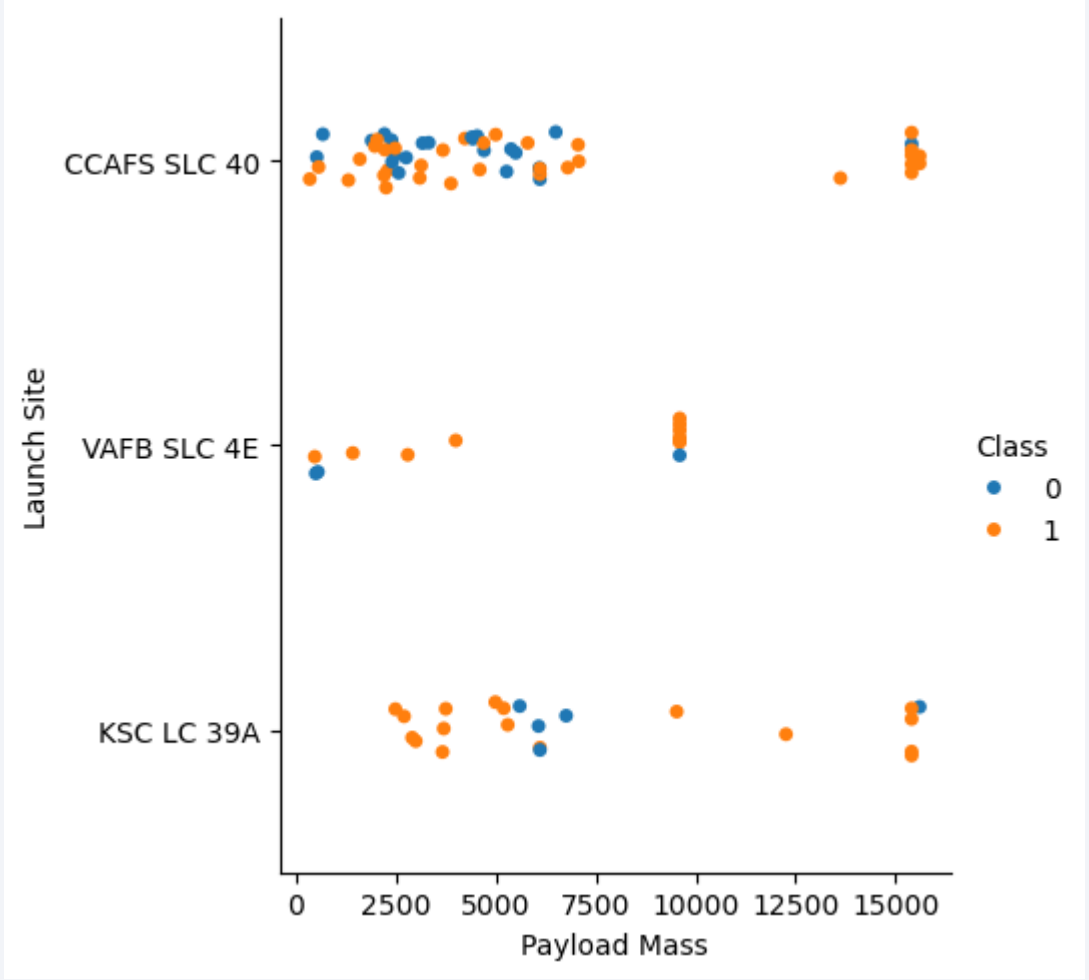
Flight Number vs. Launch Site

- The number of successes vs. failures increases with the flight number
- There are more successes compared to failure for launches from sites VAFB SLC 4E and KSC LC 39A, but most flights were launched from CCAFS SLC 40



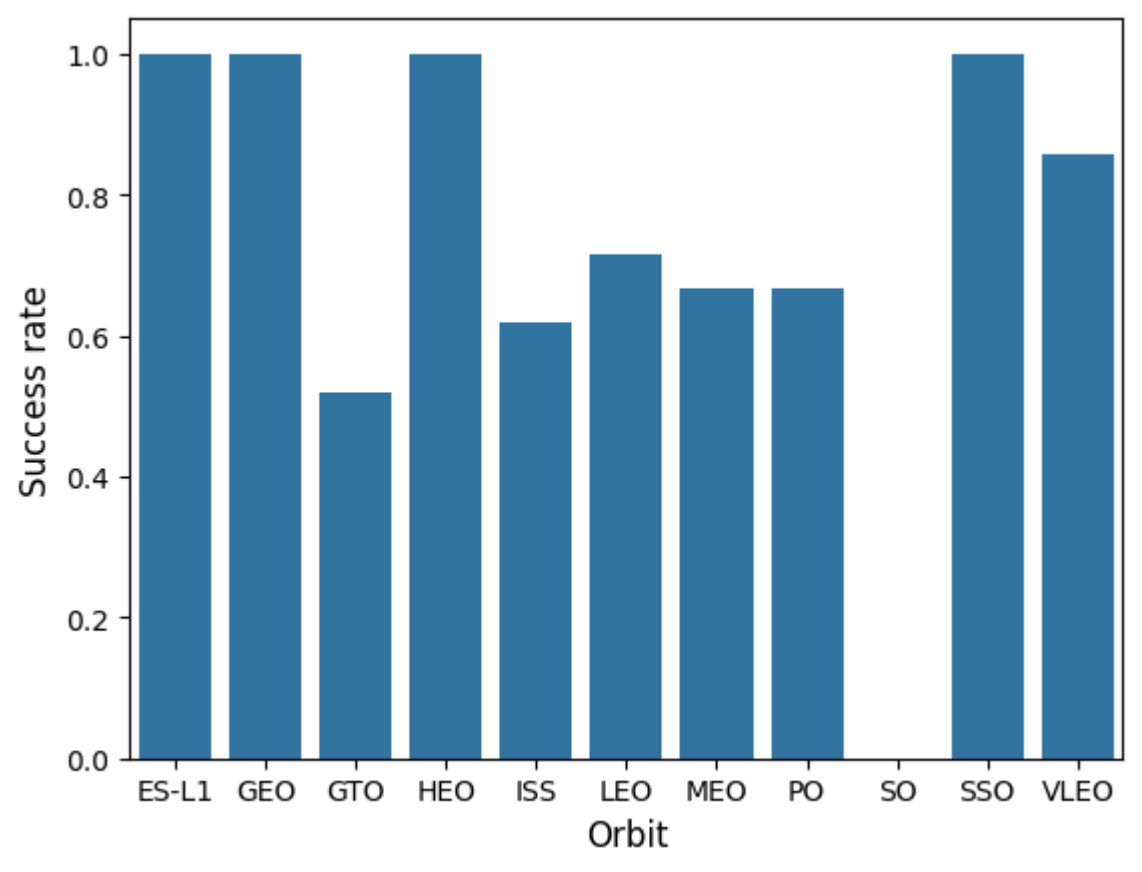
Payload vs. Launch Site

- Flights from site CCAFS SLC 40 have the best rates of success when they fly heavy payloads (15000 kg).
- In general, flights with payloads over 7500kg have higher chances of landing the first stage successfully.
- From site VAFB SLC 4E, mostly small and medium payloads are launched.



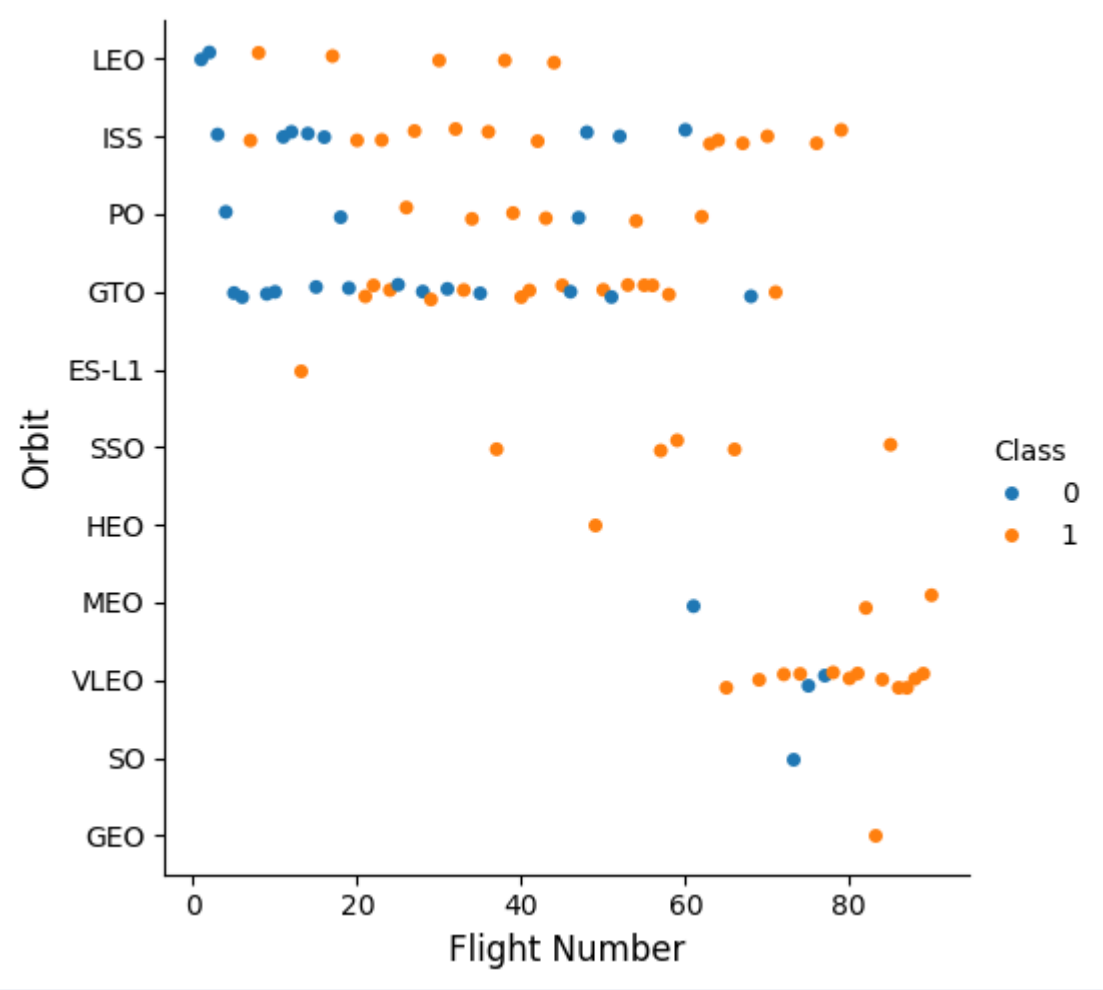
Success Rate vs. Orbit Type

- Only one flight each was launched for orbits ES-L1 (L1 Lagrange point), GEO (geosynchronous orbit) and HEO (highly elliptical orbit), and all were successes.
- Combining SSO and SO (both indicating the same Sun-synchronous orbit), out of 6 flights, 5 landed the first stage successfully.
- On the other hand, the smallest success rates was obtained for launches to GTO (the orbit of satellites) and the ISS, while these two were the destination of the highest number of launches (27 and 21, respectively).



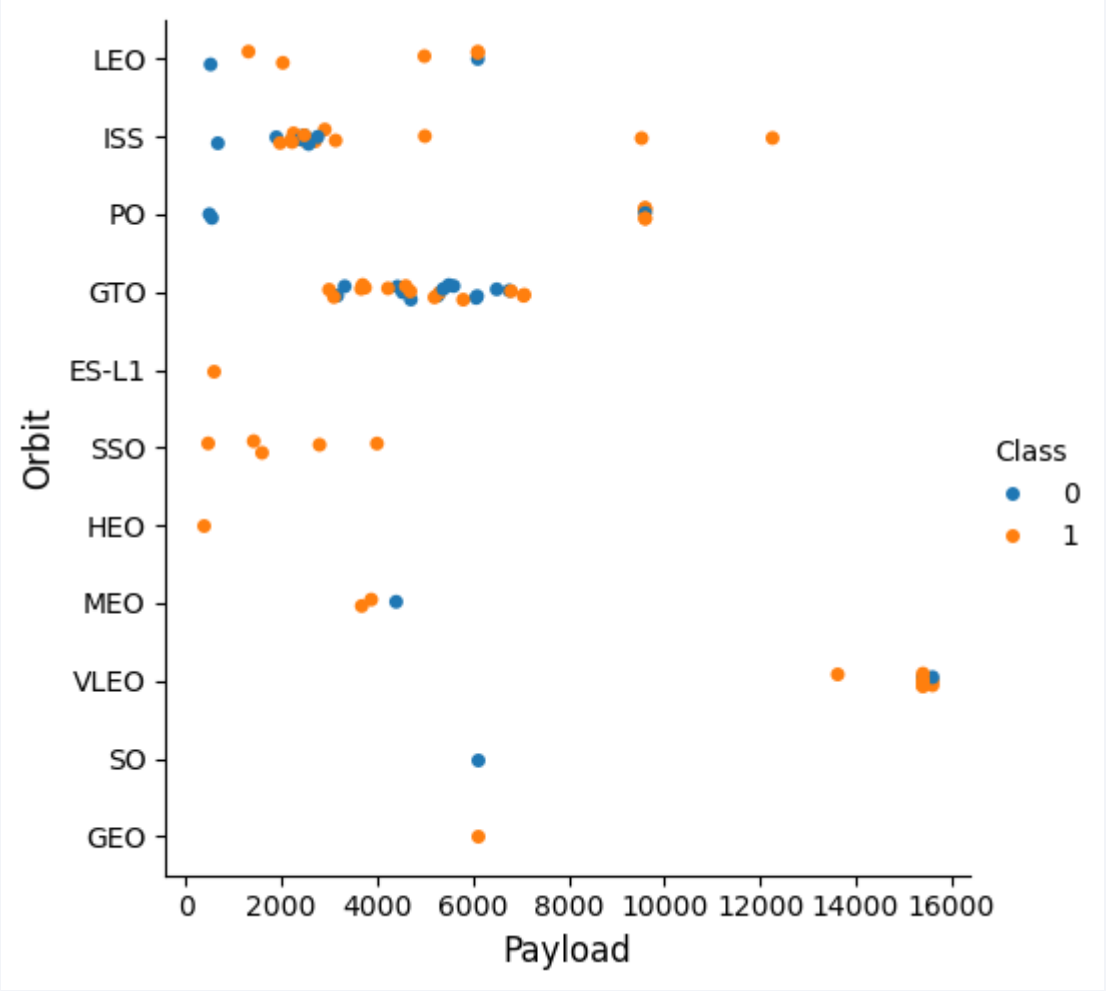
Flight Number vs. Orbit Type

- Flights to LEO orbit, while in small number, managed to land successfully after the first two failures, while flights to SSO always had successful first stage landings.
- For other orbits where a large number of flights were sent, success rates improved with subsequent flights, but results remain mixed.



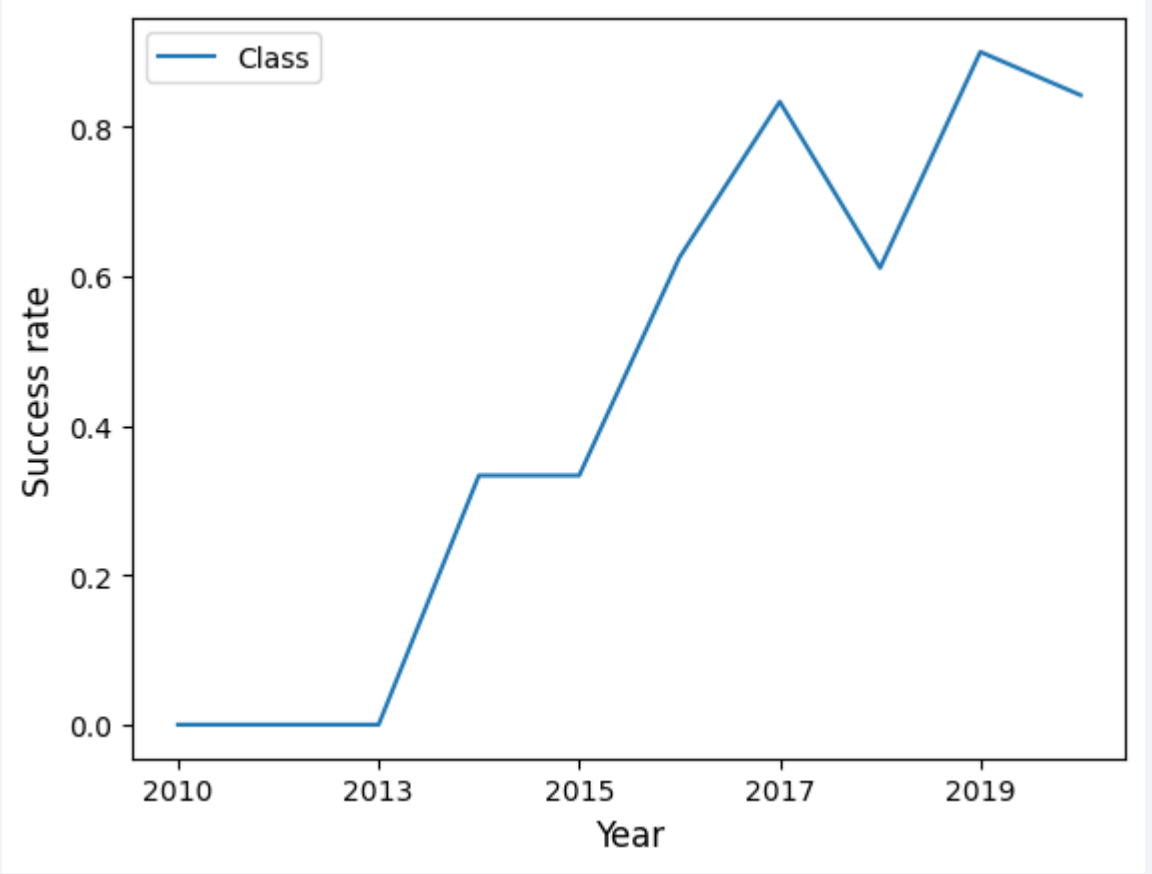
Payload vs. Orbit Type

- Landing results were mixed for most payload masses across orbits. However, launches to ISS had successful first stage landings for payloads over 5000kg.
- SSO/SO launches had successful landings for the smaller payloads.



Launch Success Yearly Trend

- Looking at success rate over time, it started to show significant and continuous improvements starting from 2014, with only a dip in 2018.



All Launch Site Names

- Space X flights were launched from four different launch sites

```
%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'

- Find 5 records where launch sites begin with `CCA` are all from CCAFS LC-40 site and start from June 2010, all either failing to land or making no attempt

```
%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

- Total payload carried by boosters from NASA is almost 100,000 kg

```
%sql select sum(PAYLOAD_MASS__KG_) FROM SPACEXTABLE where Customer like 'NASA%'
```

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

sum(PAYLOAD_MASS__KG_)
99980

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 is 2535kg

```
%sql select avg(PAYLOAD_MASS__KG_) FROM SPACEXTABLE where Booster_Version like 'F9 v1.1%'
```

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- The first successful landing outcome on ground pad came on 22nd December 2015

```
%sql select MIN(`Date`) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
```

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

MIN(`Date`)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Four different boosters carrying a payload mass between 4000 and 6000 have successfully landed on drone ship

```
%sql select distinct Booster_Version
from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)'
and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 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

- Almost all mission outcomes were successful (according to mission objective) and only one failed

```
%sql select Mission_Outcome, count(*) from SPACEXTABLE group 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

- 12 boosters have carried the maximum payload mass of 15600 kg.

```
%%sql
select Booster_Version, PAYLOAD_MASS__KG_
from SPACEXTABLE
where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- Two flights failed their landing in drone ship in 2015, both launched from CCAFS LC-40, both with booster versions F9 v1.1

```
%%sql
select substr(Date, 6, 2) as Month, Landing_Outcome, Booster_Version, Launch_Site
from SPACEXTABLE
where Landing_Outcome = 'Failure (drone ship)' and substr(Date, 0, 5) = '2015'
```

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

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

- 10 missions done between 4th June 2010 and 20th March 2017 didn't attempt landings

```
%%sql
select Landing_Outcome, count(*) as Landing_Outcome_Counts
from SPACEXTABLE
where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by Landing_Outcome_Counts desc
```

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

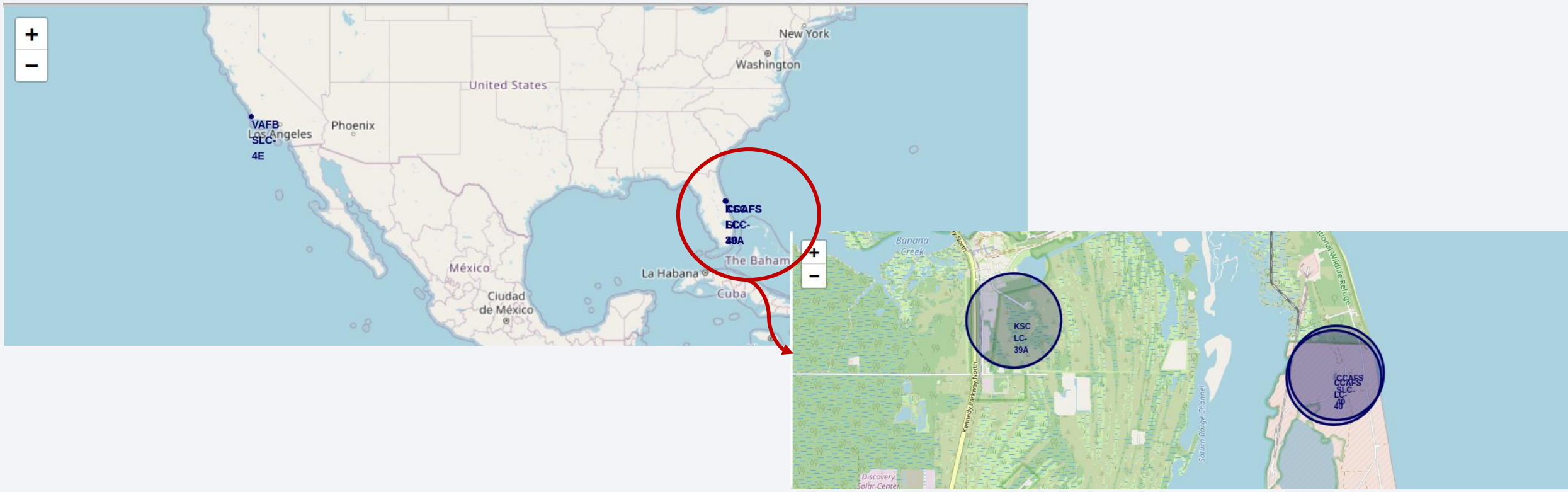
Landing_Outcome	Landing_Outcome_Counts
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

General Map of Launch Sites

- Out of the four different launch sites used by SpaceX to launch Falcon 9 rockets, three are on the East coast, in Florida.

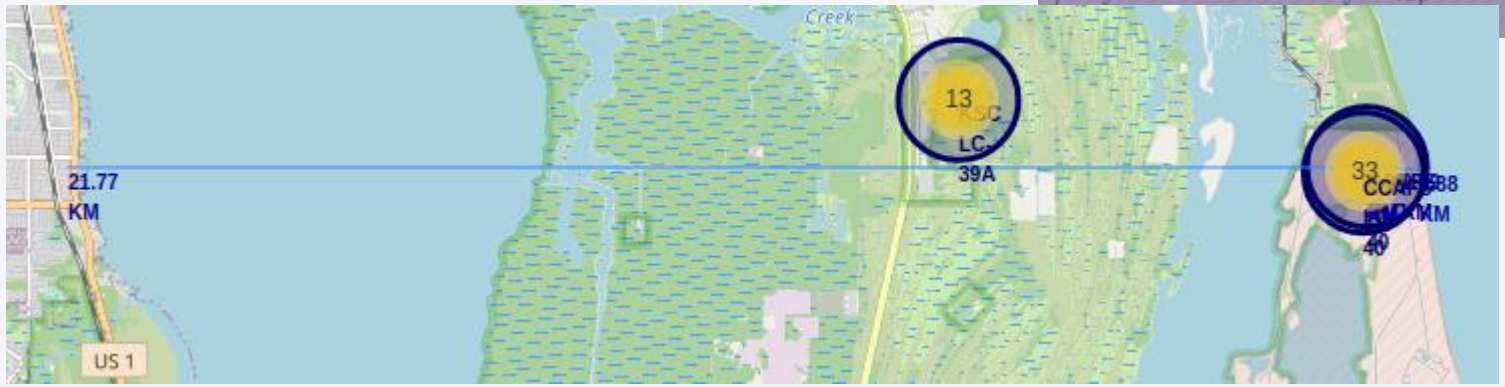


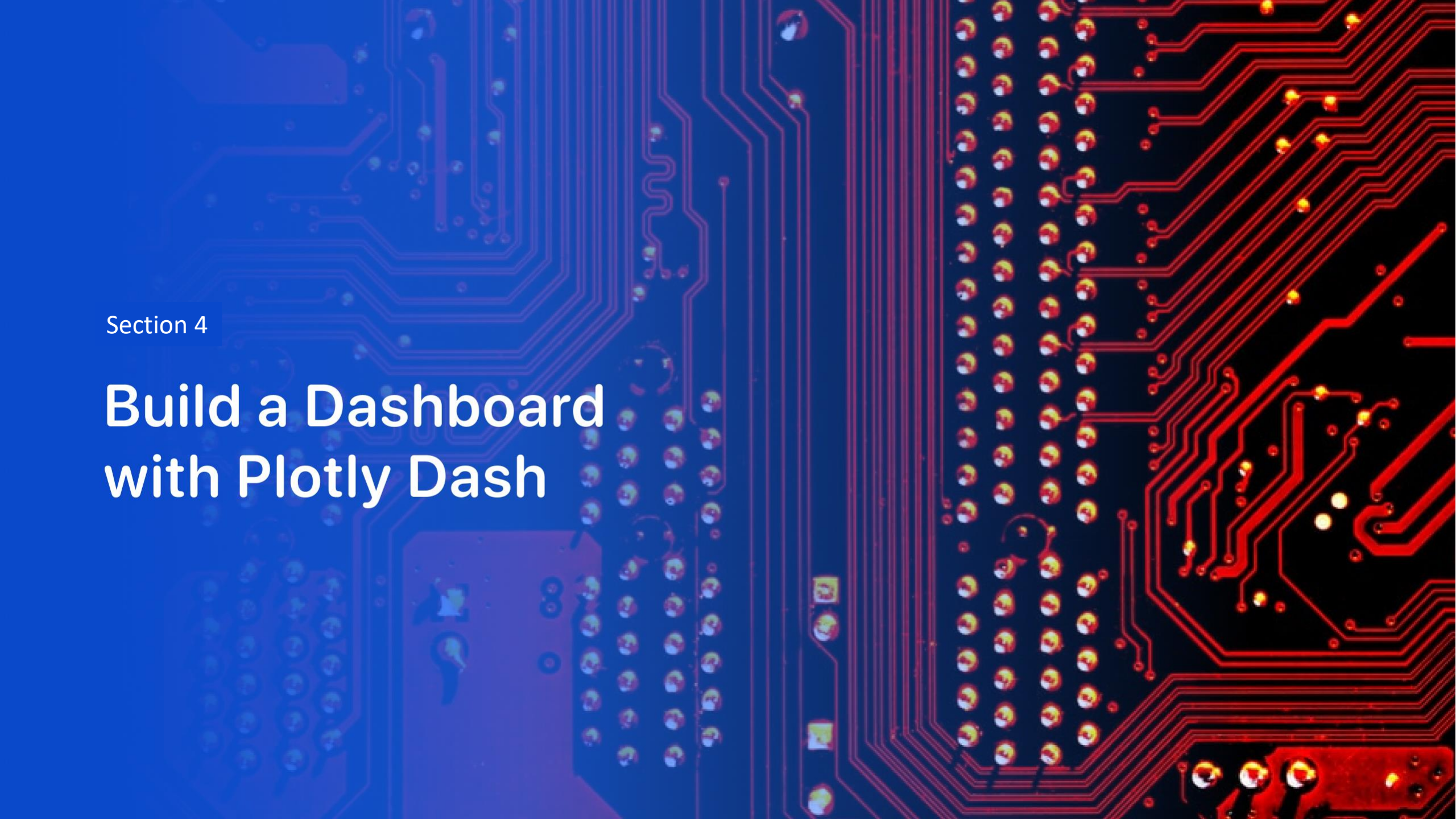
Successful and Faild Landings from KSC LC-39A Launch Site

- Markers represent each launch from site KSC LC-39A and whether it was successful (green) or has failed (red).
- Out of 13 launches from this site, 10 had successful landing of first stage

Proximities of Launch Site CCAPS LC-40

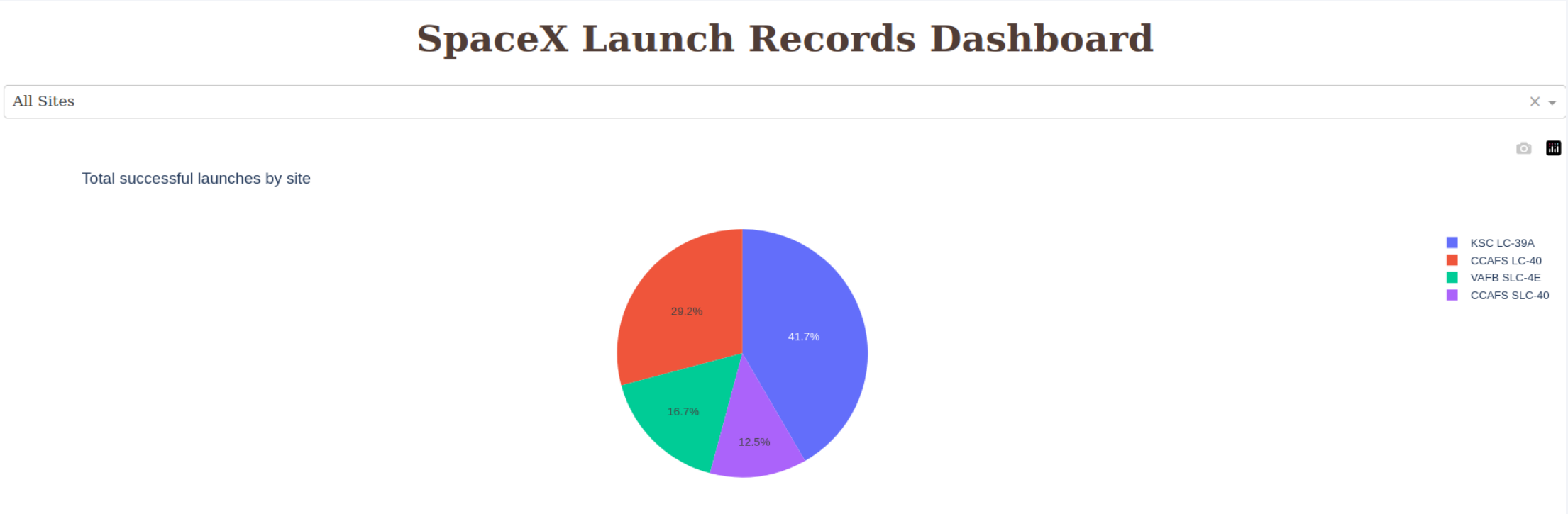
- Launch Site CCAPS LC-40 is quite close to the coastline, a highway and railroad less than 1km) and almost 22km from the nearest city




Section 4
Build a Dashboard with Plotly Dash

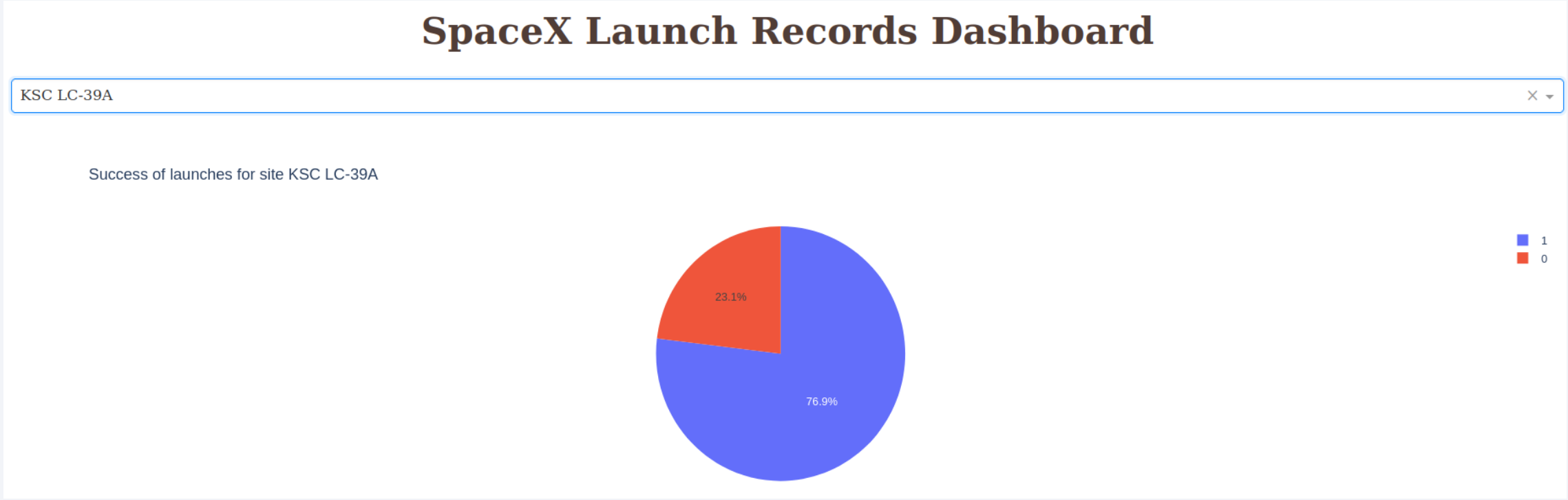
Total Successful Launches by Site

- Launches from KSC LC-39A had the highest number of successes, 41.7% of all successful landings of the first stage.



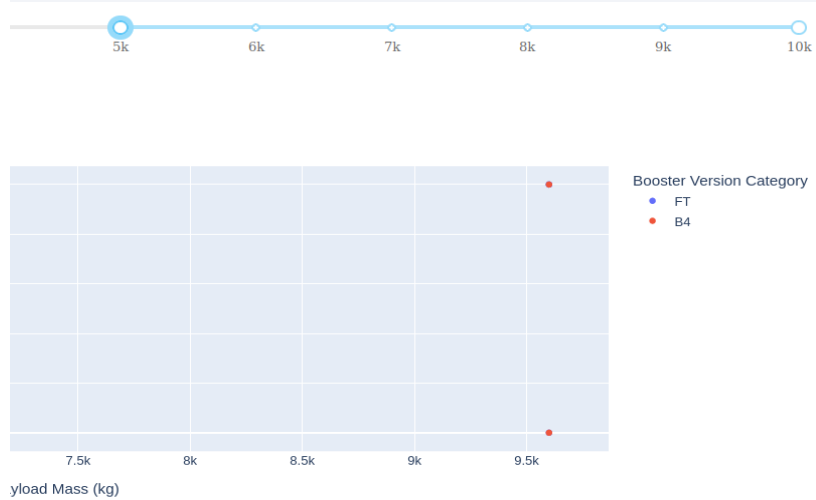
Success of Launches for Site KSC LC-39A

- KSC LC-39A not only had the most successes out of all 4 sites, it also had the highest success rate, with more than 3 out of 4 launches having successful first stage landings (76.9%).



Correlation Between Payload and Success for all Sites

- Launches carrying medium payload mass ranges had largest number of successes when FT and B4 booster versions were used. V1.1 versions were all unsuccessful when launching medium mass payloads.
- When large payloads were launched, most landings of the first stage were unsuccessful.

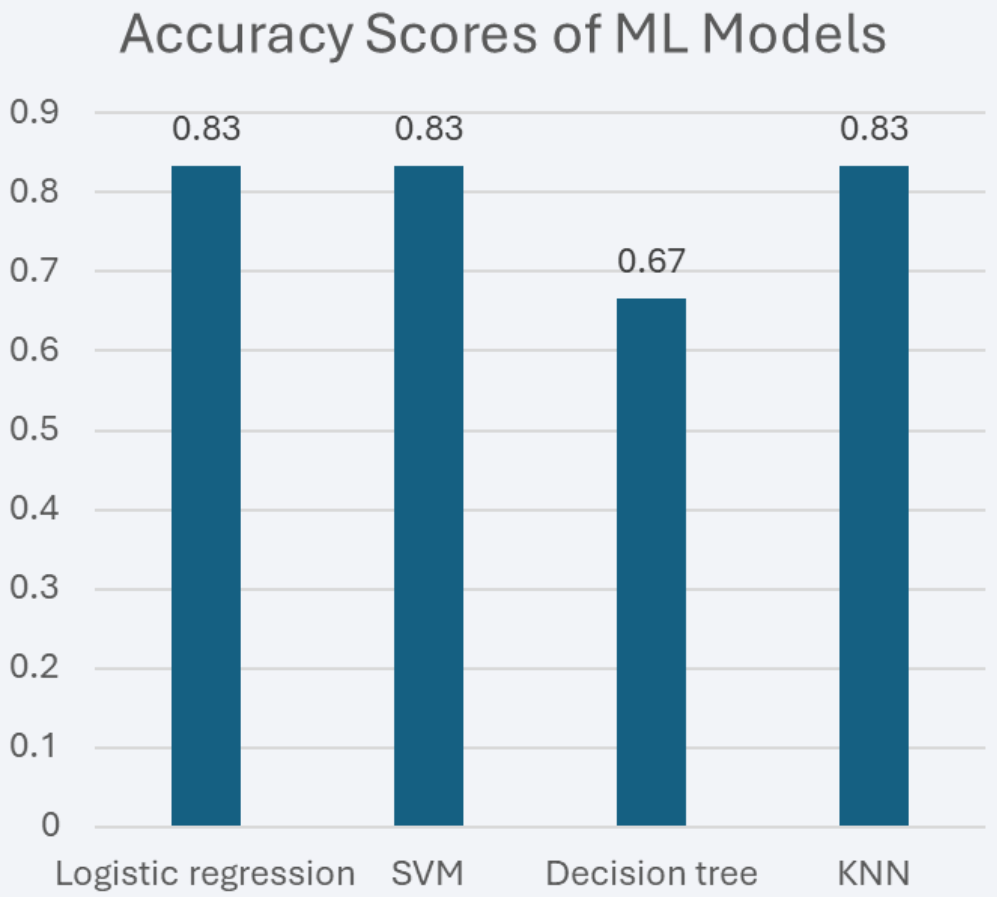


Section 5

Predictive Analysis (Classification)

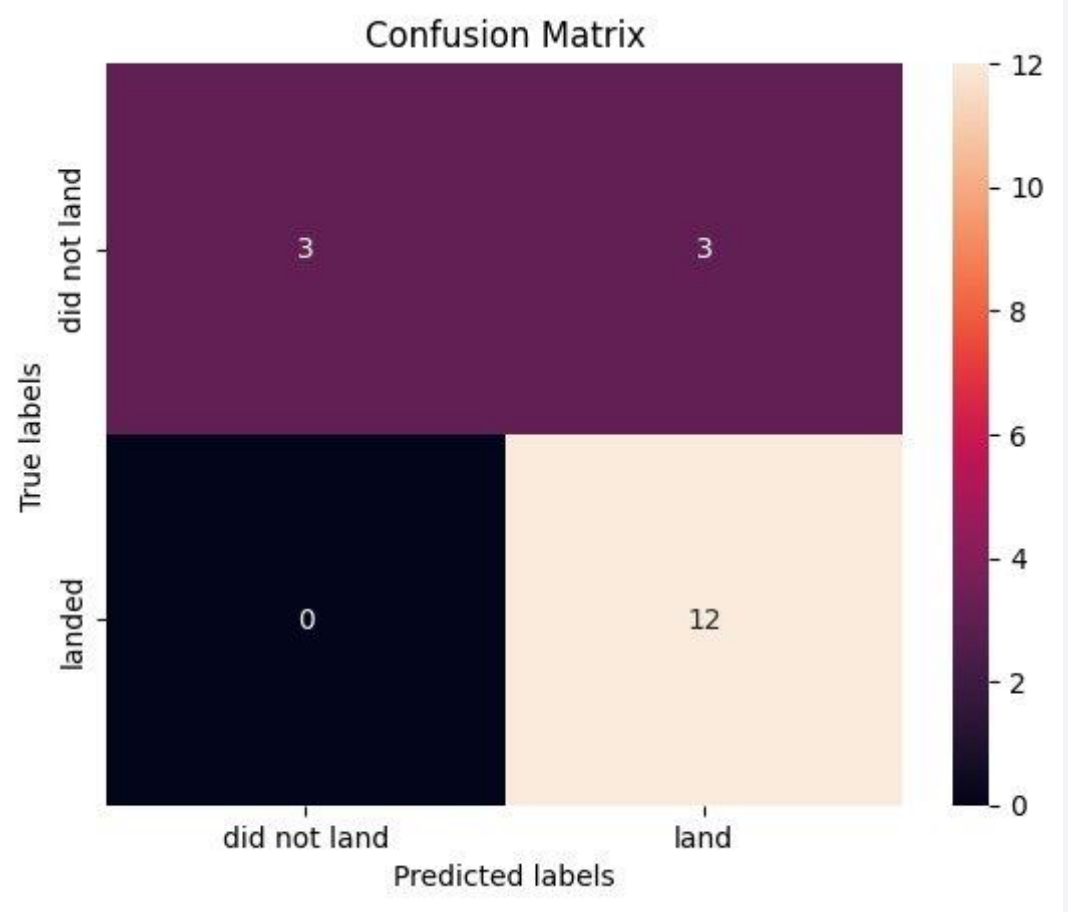
Classification Accuracy

- Logistic regression, SVM and KNN all had the same accuracy of 0.83, the highest obtained with the four tested models.



Confusion Matrix

- Logistic regression, SVM and KNN all had the same confusion matrix
- None of the three had false negatives, but all had 3 false positives, representing half of the failures



Conclusions

- Success rates increased in time, thus booster version categories are good predictors of success.
- Launch sites, orbits and payload mass are somewhat correlated with each other and also good predictors.
- The success or failure of first stage landings of Falcon 9 launches can be predicted with reasonable accuracy (0.83), knowing a few characteristics of the flight. Successes can be quite accurately determined, however failures are harder to predict. The tested models yield 50% false positives.
- Three classification models proved to be equaly suitable for the prediction: Logistic Regression, SVM and K-Nearest Neighbours.

Thank you!