IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- In our analysis of SpaceX's successful landings of first stage rockets:
  - We extracted SpaceX data through SpaceX's API using Python's request
  - We conducted exploratory data analysis and with Pandas and Sql Alchemy
  - We performed interactive visual analysis of the data with Folium, Plotly, and Dash
  - Finally, we predicted successful landings based on engineered features, using several machine learning algorithms from the Sklearn library:
    - We employed decision tree, K-nearest neighbors (KNN), support vector machine (SVM), and logistic regression classification models
    - We tuned model hyperparameters using a grid search and selected the algorithm with the best accuracy rate

- Our results may be summarized as follows:
  - Their CCASF LC-40 site was the most successful, both in number of successful launches and overall success rate (19 successful launches, and 73%, respectively)
  - The payload mass range associated with highest success rate was under 1,000 kg
  - The payload mass ranges associated with the lowest success rate were between 1,000 and 2,000kg, and around 9,000kg
  - The most successful version of the F9 booster was the v1.1
  - The types of orbits with higher success rates were ES_L1, GEO, HEO, and SSO

Introduction

- SpaceY (our client) intends to compete with SpaceX (the industry standard) in offering space launching services
- SpaceX value proposition is based on effective recovery of the first stage rocket (which lets them charge 1/3 of the price of their competitors)
- SpaceY required us to analyze which factors contribute to SpaceX's success in first stage recovery
- To analyze the relative importance of these factors
  - We obtained freely available data on SpaceX's launches
  - We employed several data preprocessing, evaluation, and machine learning classification techniques
- The main question to answer is:
  - Based on SpaceX's data, Is there a combination of launch site, payload weight, orbit, or booster version that offers better probability of a successful landing of the first stage rocket?

Section 1

Methodology

Methodology

Executive Summary

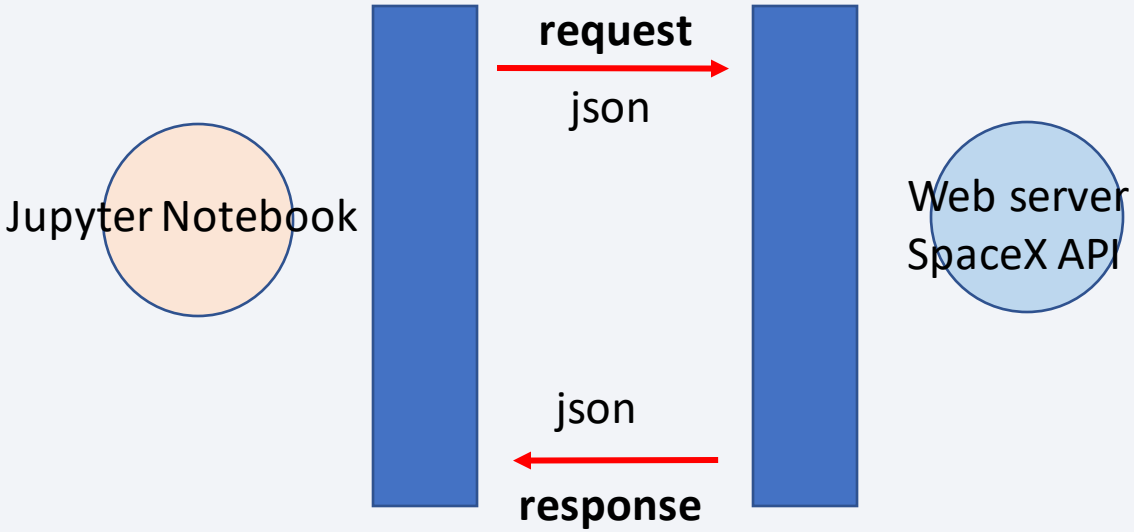
- Data collection methodology:
  - We extracted SpaceX data through SpaceX's API using python request
- Perform data wrangling
  - We used Pandas to clean the data, filtered only records employing F9 boosters, and dealt with null and missing values.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We employed decision trees, KNN, SVM, and logistic regression models
  - We tuned model hyperparameters using GridsearchCV and selected the algorithm with the best accuracy rate

Data Collection

- We used GET requests to extract a dataset from the SpaceX URL using the SpaceX API
- We defined a series of helper functions to call the API and append data about booster versions, launch sites, payload weights, orbits, and other core data to different lists.
- We combined the lists into a dictionary, and used the dictionary to create a Pandas dataframe.
- We filtered the dataframe to only include a type of booster, the Falcon 9.
- Finally, we dealt with null and missing values in the payload column, replacing them with the payload mean.

Data Collection – SpaceX API

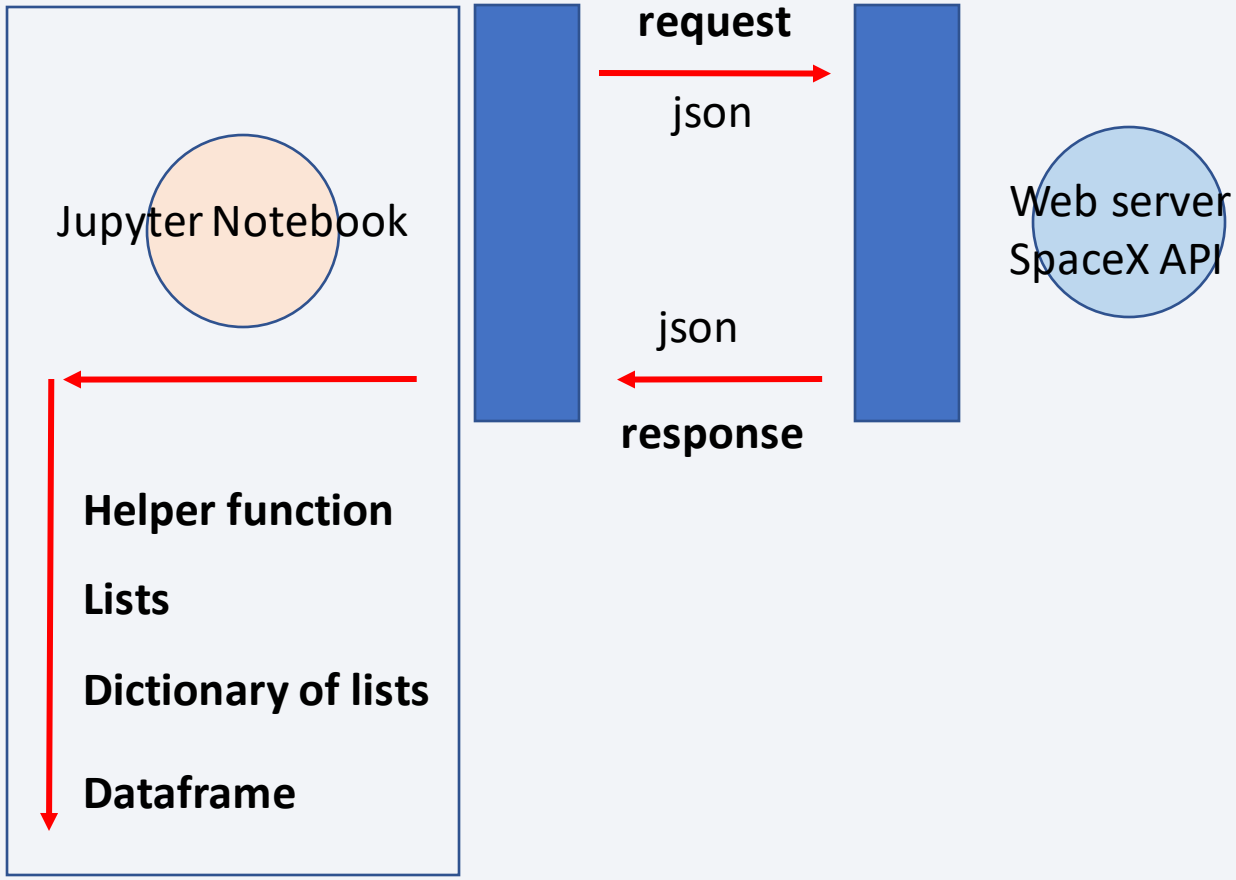
- Identify SpaceX URL
- Send GET request to SpaceX API
- Decode the response content as a Json and turn it into a Pandas dataframe



•External reference: https://github.com/hmadinaveitia/SpaceY_Capstone_Project/blob/main/SpaceY_Capstone_Project1.ipynb

Data Collection - Scraping

- Run helper functions that use the API again to get information about the launches using the IDs given for each launch, and store data about launch sites, payloads, boosters, and cores, and store the data in lists using utility functions
- Combine lists into a dictionary
- Create a Pandas data frame from the dictionary



External reference: https://github.com/hmadinaveitia/SpaceY_Capstone_Project/blob/main/SpaceY_Capstone_Project1.ipynb

Data Wrangling

- Filter the dataframe using the BoosterVersion column to only keep the Falcon 9 launches.
- Save the filtered data to a new dataframe called data_falcon9.
- Check the new dataframe for null or missing values.
- Calculate the mean for the PayloadMass. Then replace empty or null values in the data with the calculated mean.
- Verify we have no missing values in our dataset except for in LandingPad.
- Export the dataframe to a CSV file

External reference: https://github.com/hmadinaveitia/SpaceY_Capstone_Project/blob/main/SpaceY_Capstone_Project1.ipynb

EDA with Data Visualization

- We visualized the relationship between pairs of potential explanatory features through scatter plots. The plots helped visualize how the value of one feature changed as the value of the other one changed.
- We ran scatter plots of the relationship between (a) flight number and payload, (b)launch site and success ratio, (c) flight number and launch site,and (d) launch site and payload.
- We used a bar chart to determine the success ratio per type of orbit.
- Finally, we used a line plot with launch year on the x-axis and average success rate on the y-axis, to get the average launch success trend over time.

External reference: https://github.com/hmadinaveitia/SpaceY_Capstone_Project/blob/main/SpaceY_EDA_with_Pandas.ipynb

EDA with SQL

We employed sql queries to:

- Identify the unique launch sites in the space mission
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display the average payload mass carried by booster version F9 V1.1
- Find out when the first successful landing outcome in ground pad was achieved
- Identify the boosters with successful drone ship landings carrying payloads between 4,000 and 6,000 kg
- Calculate the number of successes and failures in mission outcomes
- List the booster versions with maximum payload mass
- List the failed landing outcomes in drone ship, their booster versions, and launch site names for in 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

External reference: https://github.com/hmadinaveitia/SpaceY_Capstone_Project/blob/main/SpaceY_EDA_with_SQL.ipynb

Build an Interactive Map with Folium

Among other factors, the launch success rate may depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories.

To help visualize the characteristics of each launch site:

- We created a Folium Map object, with an initial center location to be NASA Johnson Space Center at Houston, Texas.
- We added a circle for each of the four launch sites used by SpaceX.
- We added market clusters, to show the launch outcomes for each site, and see which sites have high success rates (marker clusters can be a good way to simplify a map containing many markers having the same coordinate).
- Finally, we added a MousePosition on the map to (I) get coordinates for a mouse over a point on the map, and (ii) calculate the distances between each launch site and different points of interests, using a given function.

External reference: https://github.com/hmadinaveitia/SpaceY_Capstone_Project/blob/main/SpaceY_Interactive_Visual_Analytics.ipynb

Build a Dashboard with Plotly Dash

To visualize the success rate of each launch site, and the relationship between booster versions and success rates for different payload ranges, we build a dashboard that contained the following components:

- A dropdown list that let the user see results of all launch sites, or a selected launch site
- A pie chart showing the breakdown of successes and failures per selected launch site
- A slider that would let the user choose the minimum and maximum weight for a given payload range
- A scatter plot showing the relationship between booster versions and launch outcomes, for the payload range chosen in the slider.

External reference: https://github.com/hmadinaveitia/SpaceY_Capstone_Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Using Pandas and different modules of python's Sklearn package:

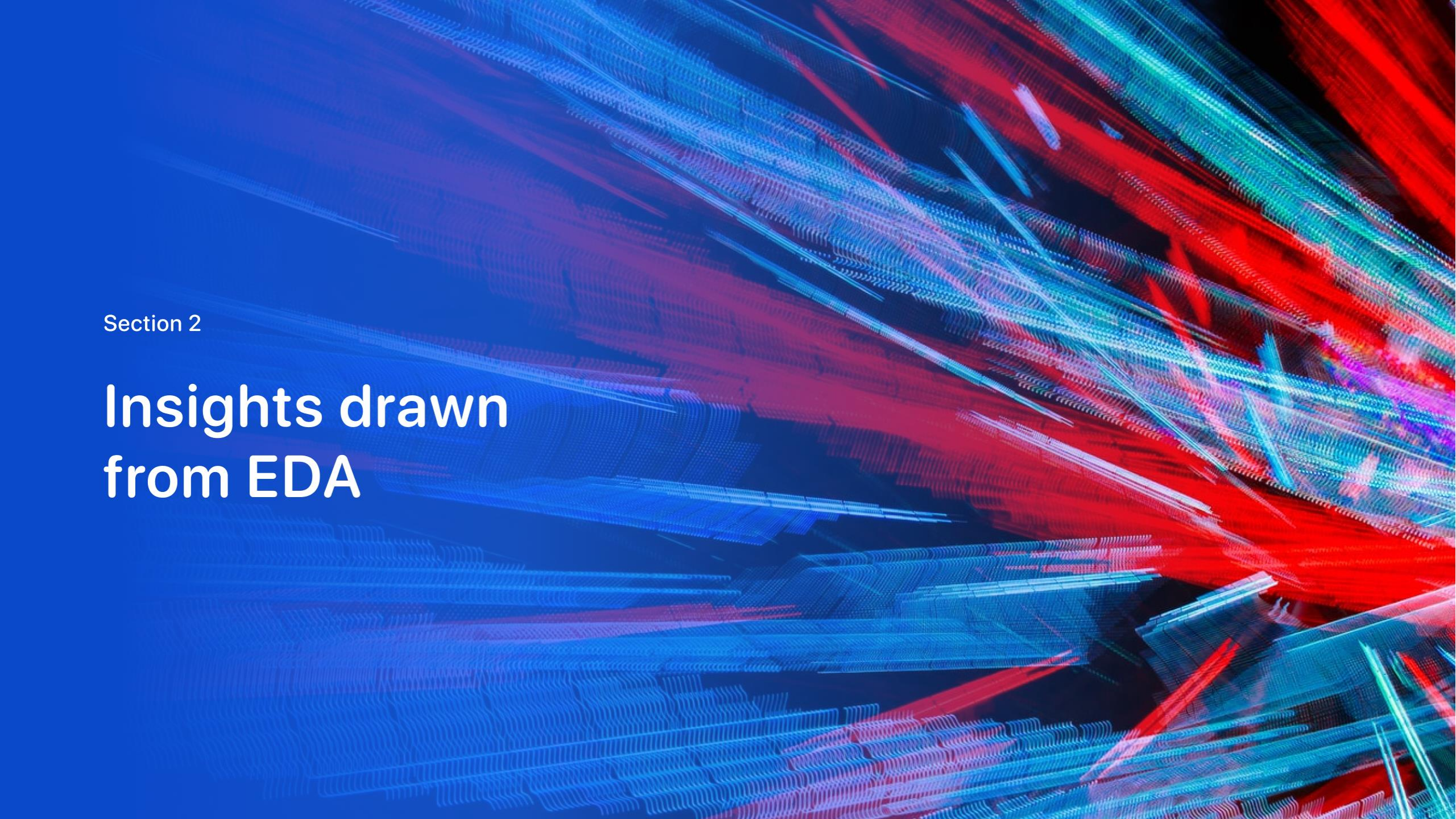
- We loaded mission data from a CSV file into a Pandas dataframe, separated the label column (Y series) from the feature column (X dataframe).
- We used Sklearn Standardscaler to standardize the features in the X dataframe, and Sklearn Train_test_split to separate the data into training and testing sets.
- We trained several machine learning classifying algorithms from the Sklearn library: decision trees, K-near neighbors (KNN), support vector machine (SVM), and logistic regression.
- We tuned the models' hyperparameters using Sklearn GridsearchCV, visualized the results using a confusion matrix, and selected the algorithm with the best accuracy rate

External reference: https://github.com/hmadinaveitia/SpaceY_Capstone_Project/blob/main/SpaceY_Predictive_Analysis.ipynb

Results

The following sections reflect the following take outs:

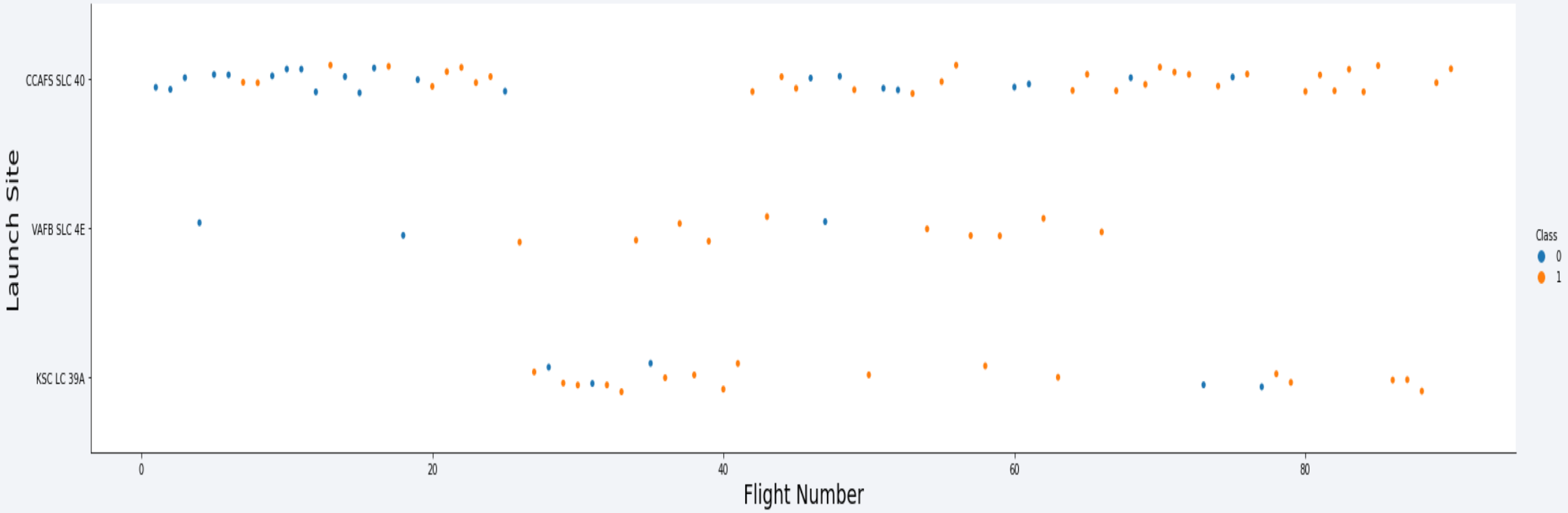
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

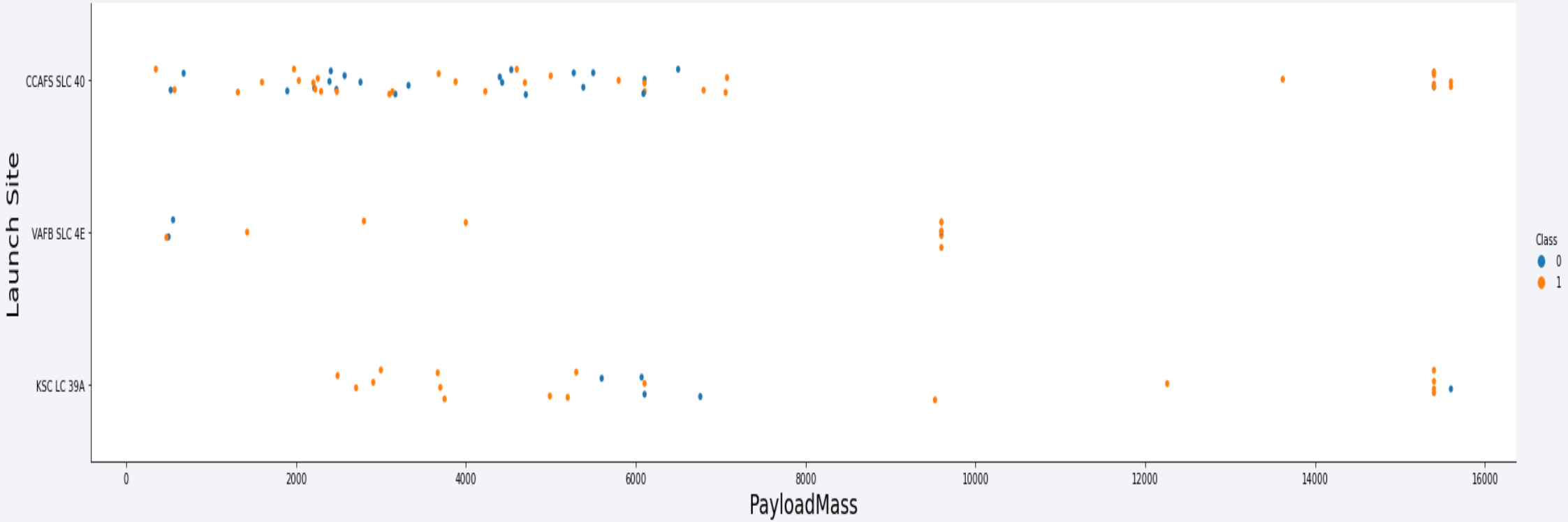
Insights drawn from EDA

Flight Number vs. Launch Site



- We see that, for all launch sites, as the flight number (the continuous launch attempts) increases, the first stage is more likely to land successfully.

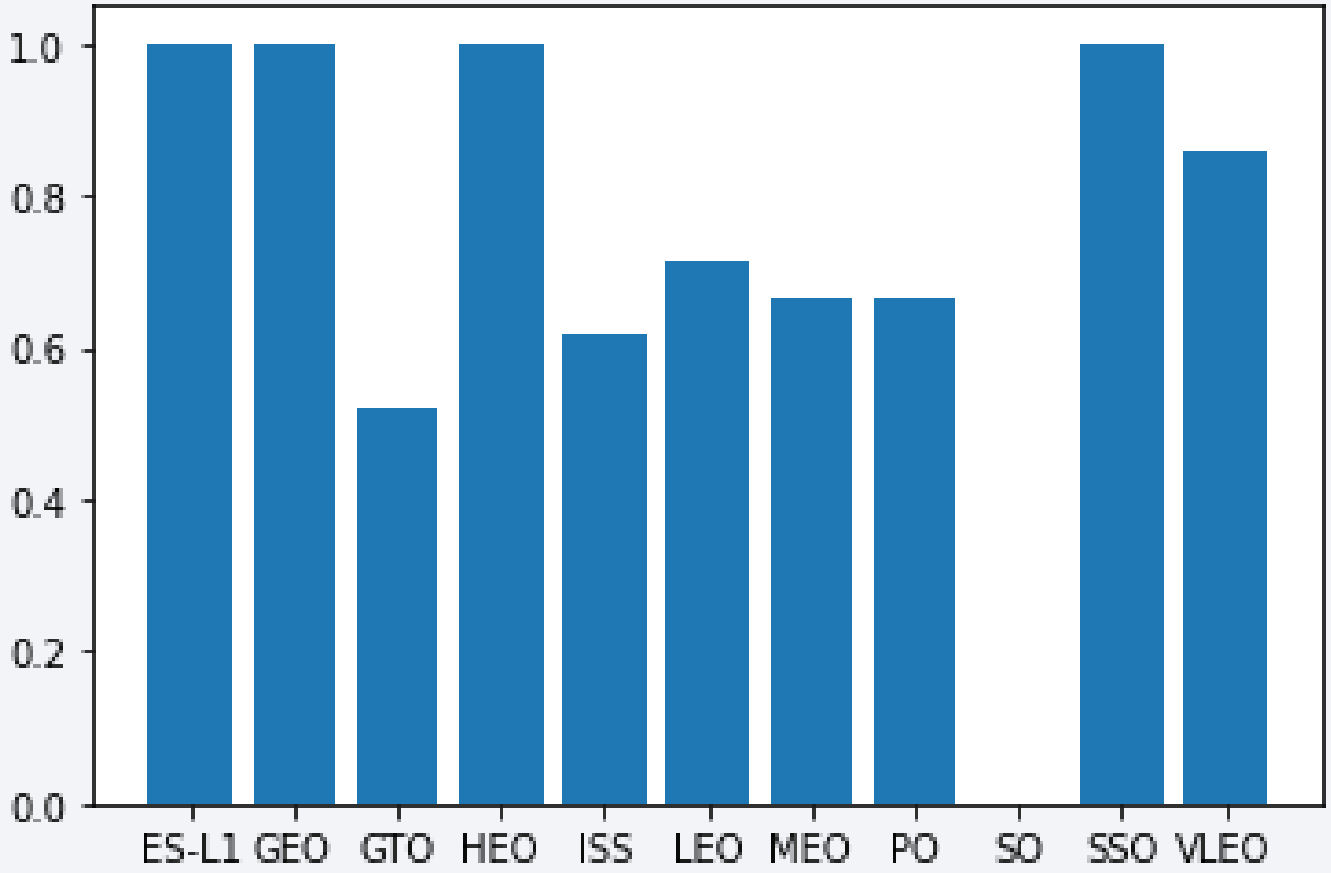
Payload vs. Launch Site



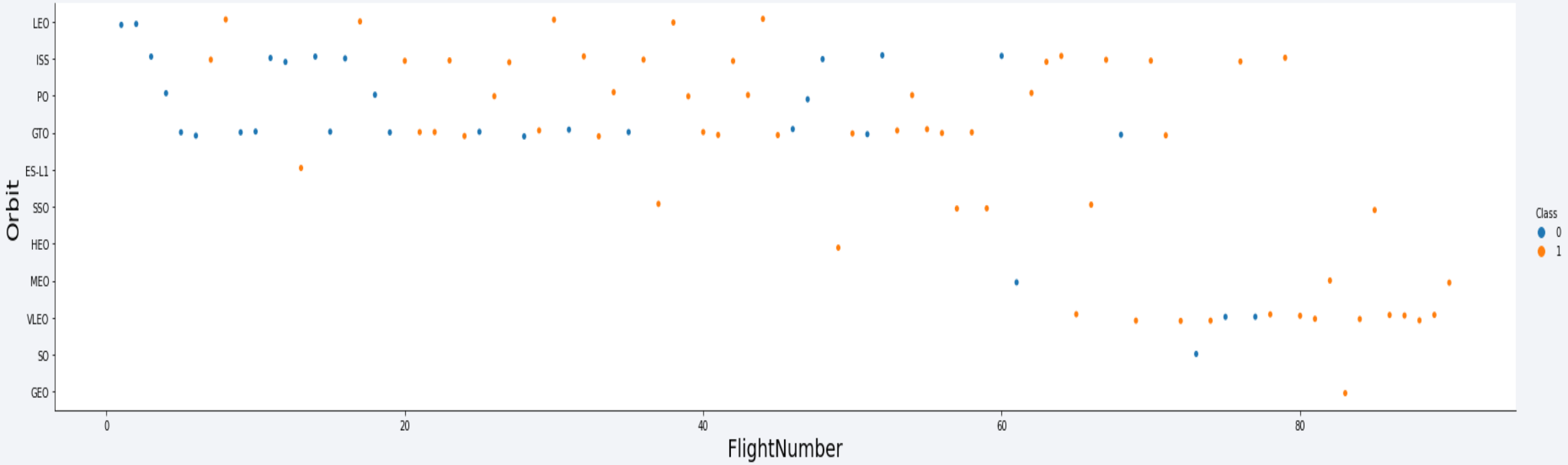
- For the CCAFS LC-40 and KSC LC-39A launch sites, heavier payloads seem to achieve successful landings.
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type

- The ES-L1, GEO, HEO, and SSO orbits show the highest success rate (100%), with the VLEO close behind, at 90%.
- The GTO orbit shows the lowest success rate (barely 50%).

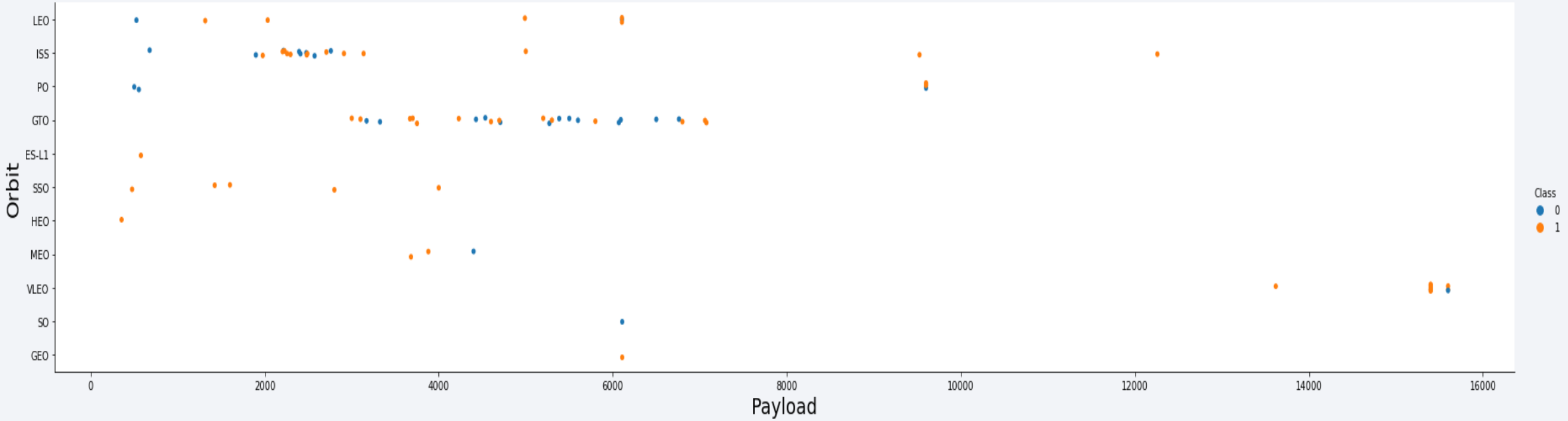


Flight Number vs. Orbit Type



- In the 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.

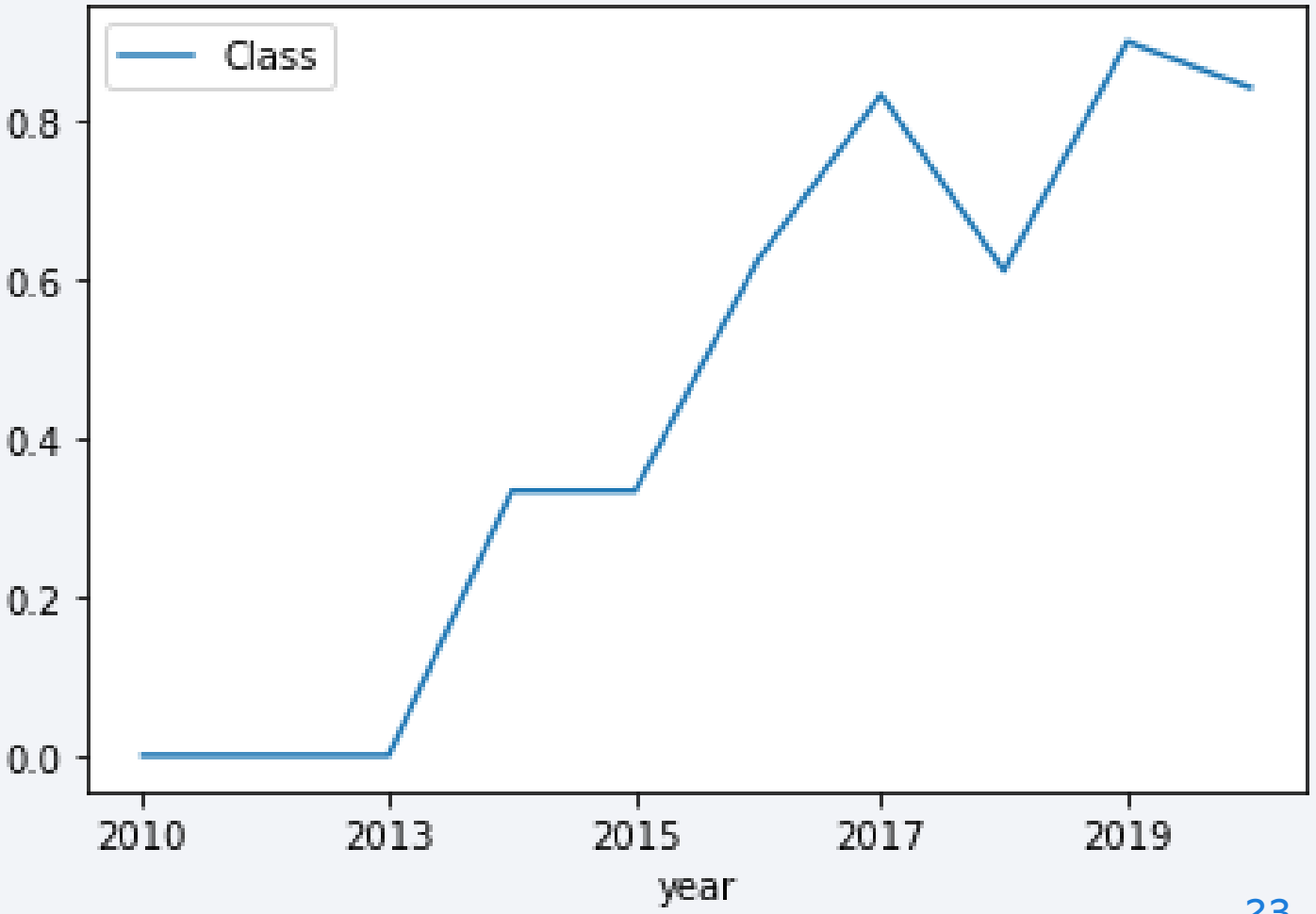
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS orbits
- However for GTO orbits we cannot establish a correlation between payload and orbit, as both successful and unsuccessful landings occur with the same payload weights.

Launch Success Yearly Trend

- The line chart shows a 3-year learning curve, with no successful landings, followed by an impressive increase in the success rate between 2013-2017 (from 0% to 80%)
- The success rate dropped somewhat in 2018, but recovered in 2019



All Launch Site Names

SQL Query	Result
%sql select distinct launch_site from spacextbl	launch_site CCAFS LC-40 CCAFS SLC-40 KSC LC-39A VAFB SLC-4
After loading the SQL extension (%load_ext sql), we selected all unique ('distinct') launch site names from the table spacextbl.	

Launch Site Names Begin with 'CCA'

SQL Query

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

Result

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

We obtained five launch sites which names begin with 'CCA' ("where launch_site like 'CCA%'") from table spacextbl

Total Payload Mass

SQL Query	Result
%sql select sum(payload_mass__kg_) as total_payload from spacextbl where customer='NASA (CRS)'	total_payload: 45,596

We calculated the total payload weight in kg ('sum(payload__mass__ksg_)') carried by boosters from NASA, from the table spacextbl

Average Payload Mass by F9 v1.1

SQL Query	Result
%sql select avg(payload_mass__kg_) as avg_payload from spacextbl where booster_version = 'F9 v1.1'	avg_payload: 2,928

We calculated the average payload mass in kg ('avg(payload_mass__kg_)') of launches that used the F9 v1.1 booster ("where booster_version = 'F9 v1.1'"), from table spacextbl.

.

First Successful Ground Landing Date

SQL Query	Result
%sql select min(date) as First_Date from spacextbl where landing__outcome = 'Success (ground pad)'	First_Date 2015-12-22

We obtained the first successful ground landing date ('min(date) where landing__outcome = 'Success (ground pad)') from table spacextbl

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL Query	Result
%sql select booster_version from spacextbl where payload_mass__kg_ between 4000 and 5999 and landing__outcome = 'Success (drone ship)'	booster_version F9 FT B1022 F9 FT B1026 F9 FT B1021.2 F9 FT B1031.2

We obtained the list of booster versions that carried a payload between 4,000kg and 6,000 kg ('where payload_mass__kg_ between 4000 and 5999') and had a successful landing on a drone ship ("and landing__outcome = 'Success (drone ship)'") from table spacextbl

Total Number of Successful and Failure Mission Outcomes

SQL Query	Result
%sql select mission_outcome, count(mission_outcome) as total from spacextbl group by mission_outcome	mission_outcome: Failure (in flight), total: 1 mission_outcome: Success, total: 99 mission_outcome: Success (payload status unclear), total: 1

We grouped the content of the table spacextbl by mission outcome ('group by mission_outcome) and calculated the number of each type of successful and failure missions ('count(mission_outcome')

Boosters Carried Maximum Payload

SQL Query	Result
%sql select distinct booster_version, payload_mass__kg_ from spacextbl where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextbl)	
We obtained the version number of those boosters that carried the maximum payload mass. We calculated the maximum payload mass in a subquery ('select max(payload_mass__kg_) from spacextbl'). Then, we used that figure to filter the names of the boosters that carried it ('select distinct booster_ version')	

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

2015 Launch Records

SQL Query	Result
%sql select date, landing__outcome, booster_version, launch_site from spacextbl where landing__outcome = 'Failure (drone ship)' and year(date) = 2015	
We identified the unsuccessful landings on drone ships ("where landing_outcome ='Failure (drone ship'") that took place in 2015 ('and year(date) =2015') from table spacextbl	

DATE	landing__ outcome	booster_ version	launch_site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

SQL Query

```
%sql select
date, landing__outcome, count(landing__outcome) as
total from spacextbl where date between date'2010-
06-04' and date'2017-03-20' group by
date, landing__outcome order by date desc
```

We grouped the data in table spacextbl by date and landing outcome, and obtained the total outcomes (by type of outcome) for each day between June 4, 2010, and March 20, 2017. We showed the results by date, in descending order (more recent results first). Only a partial list of the results shown here.

Result

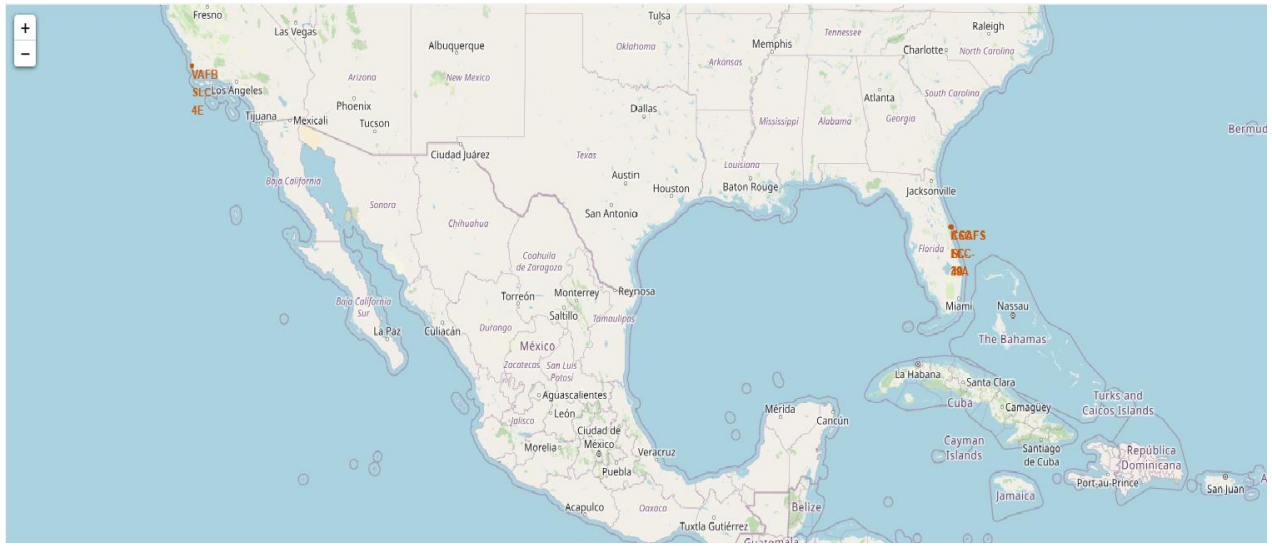
DATE	landing__outcome	total
2017-03-16	No attempt	1
2017-02-19	Success (ground pad)	1
2017-01-14	Success (drone ship)	1
2016-08-14	Success (drone ship)	1
2016-07-18	Success (ground pad)	1
2016-06-15	Failure (drone ship)	1

Section 4

Launch Sites Proximities Analysis

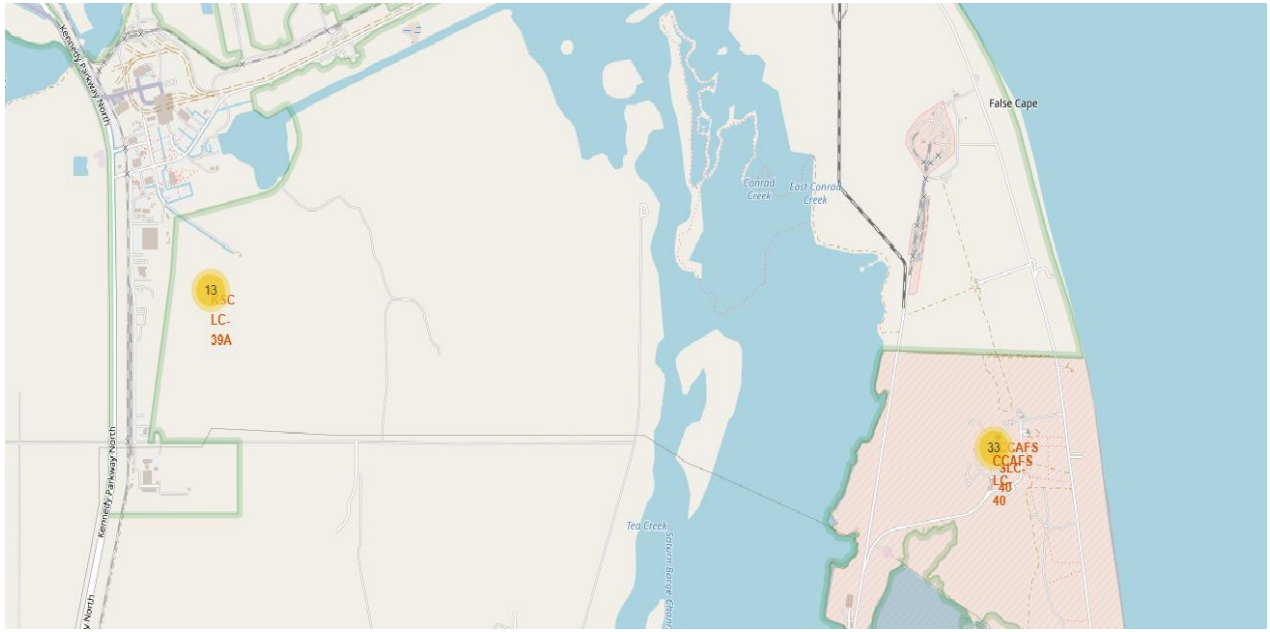
Folium Map – Launch Site Locations

- Using a Circle and a Marker, we can show the locations of all SpaceX launch sites on a map of the United States



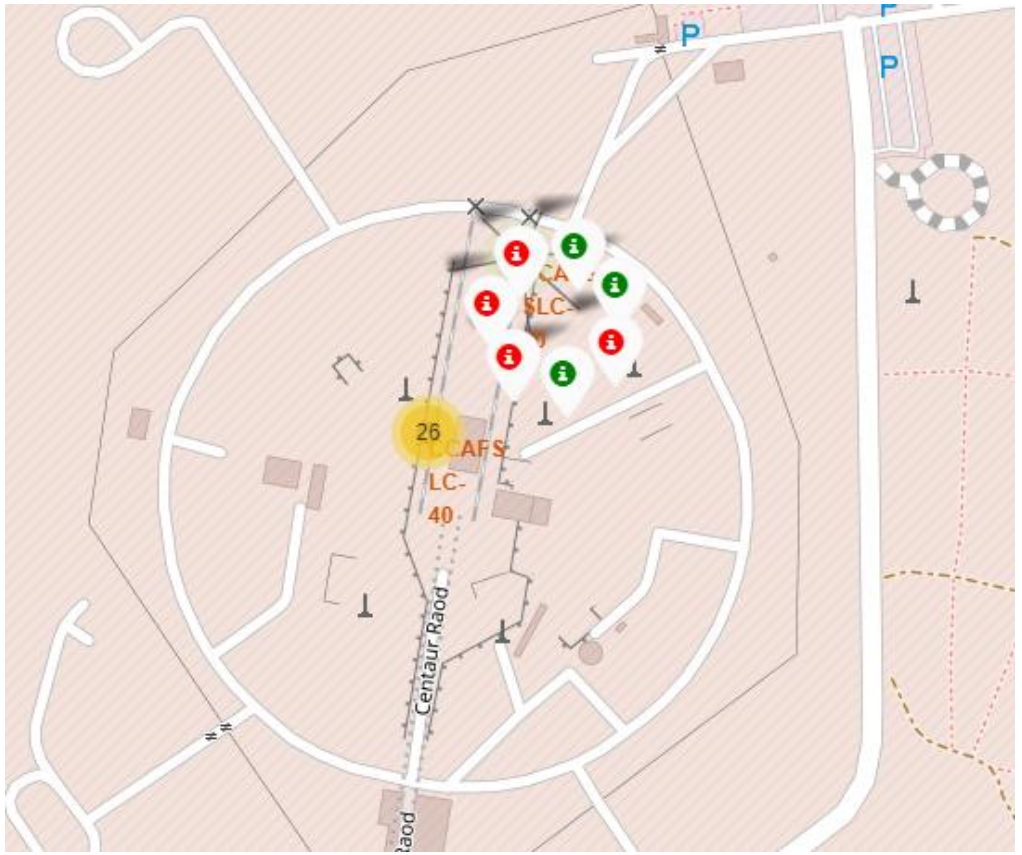
Folium Map – Launch Site MarkerClusters

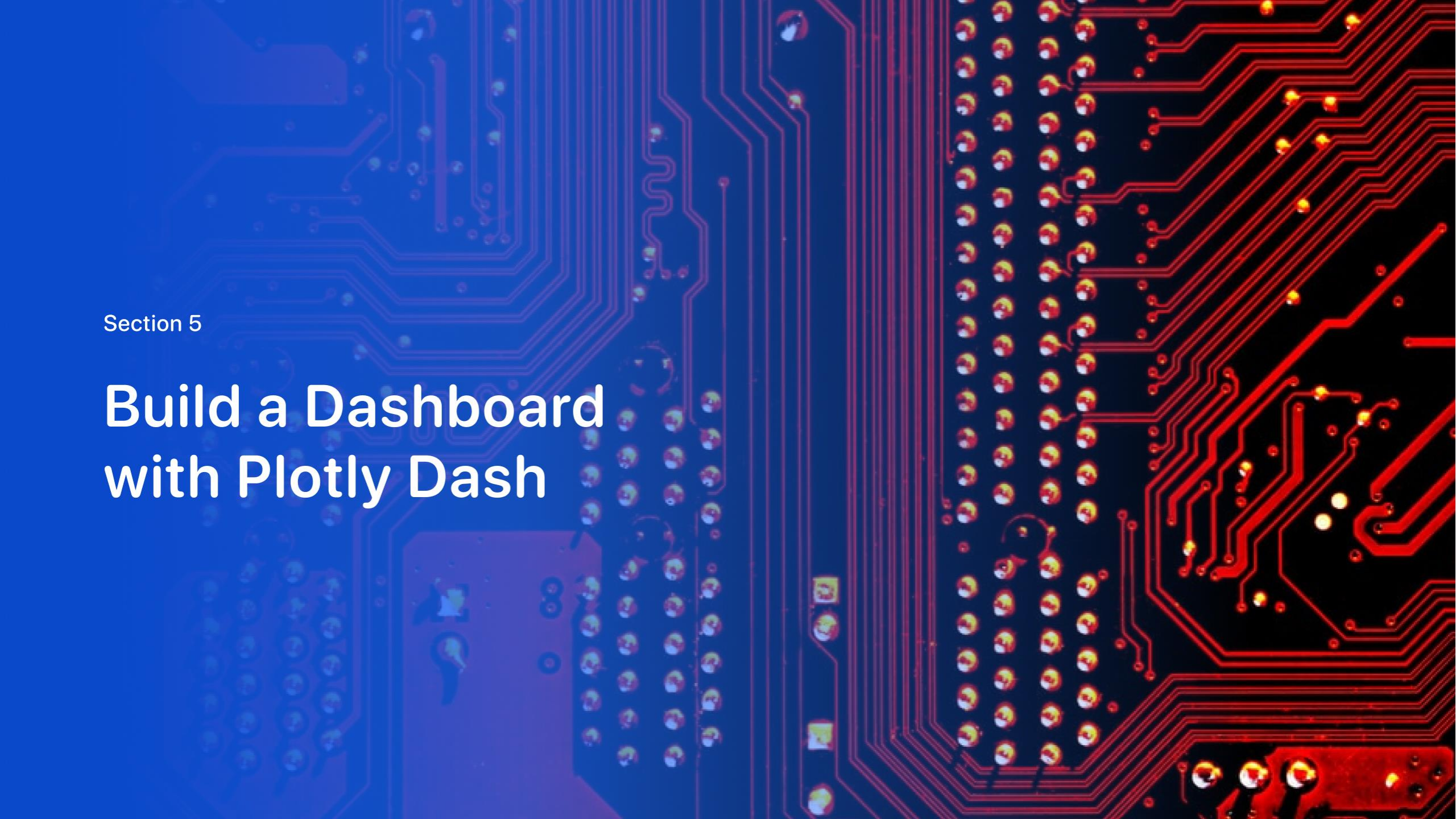
- By using MarkerClusters we can show the location of a launch site, and the number of relevant events that occurred at each site (in this case, launches)



Folium Map – Launch Sites/Launch Outcomes

- Using MarkerClusters allows us to show the location of launch site KSC LC-39A (left), and using Markers and DivIcons allow us to show the color-coded results of the 13 individual launches from the site (right), by clicking on the site.



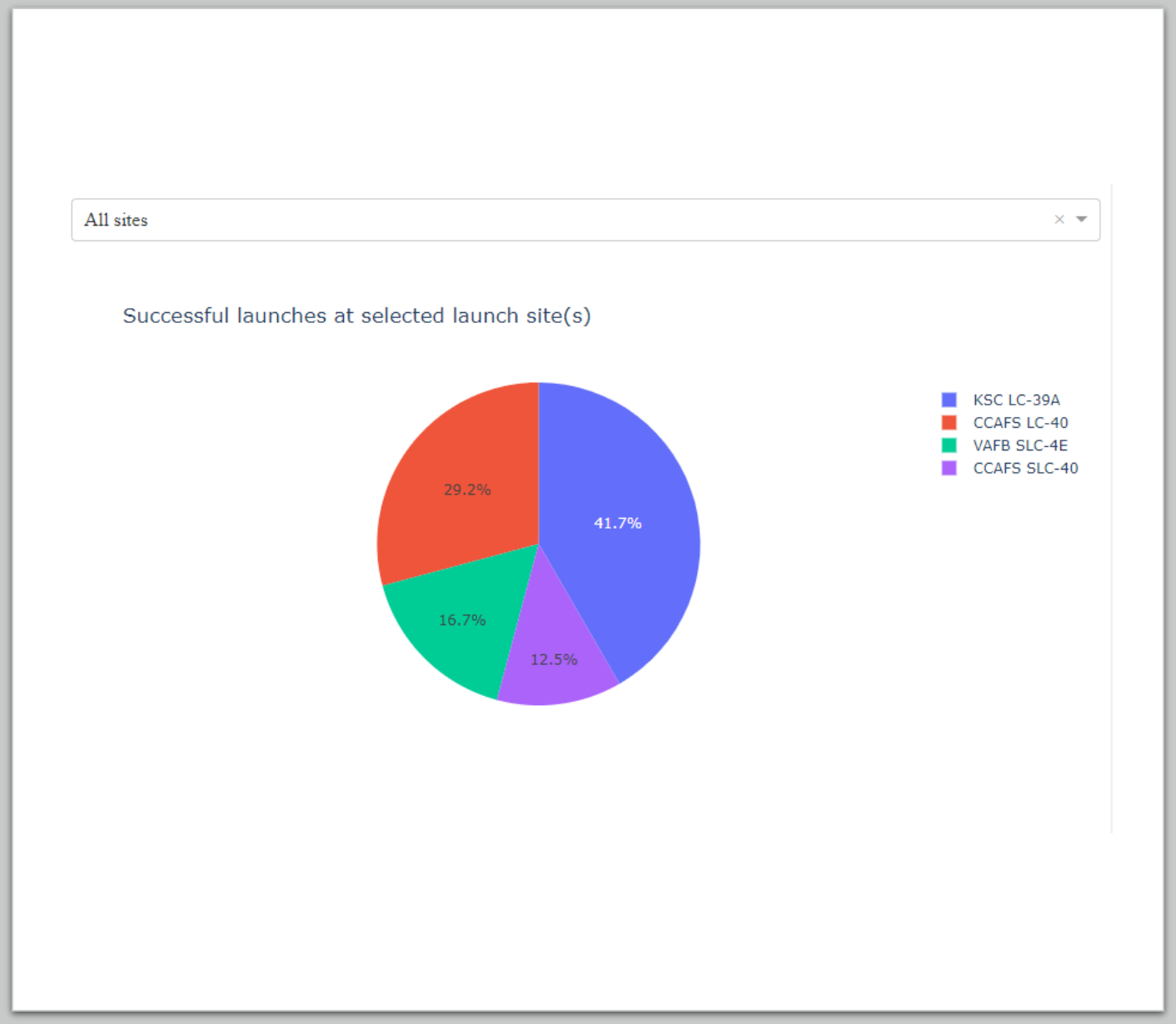


Section 5

Build a Dashboard with Plotly Dash

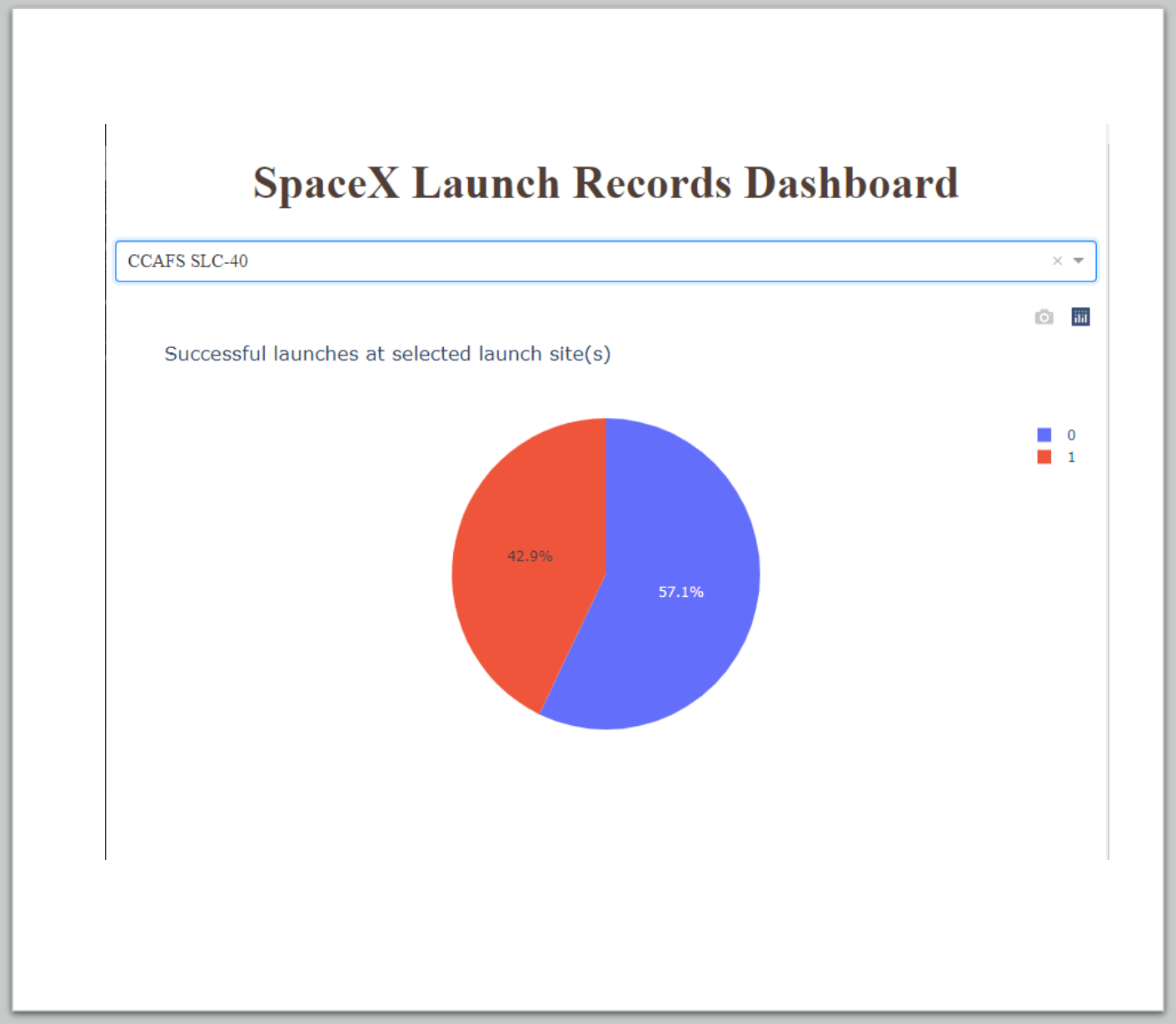
Dashboard – Successful Launches at all Sites

- The dropdown box at the top left of the image allows the user to select all launch sites (as in this case), or individual launch sites. The pie chart below reflects the percentage of successful launches per each of the four sites



Dashboard – Successful launches at a selected site

- Using the dropdown list, a user can select an individual site (in this case, KSC-LC-19A), and the pie chart below will show the percentage of successful launches at the selected site.

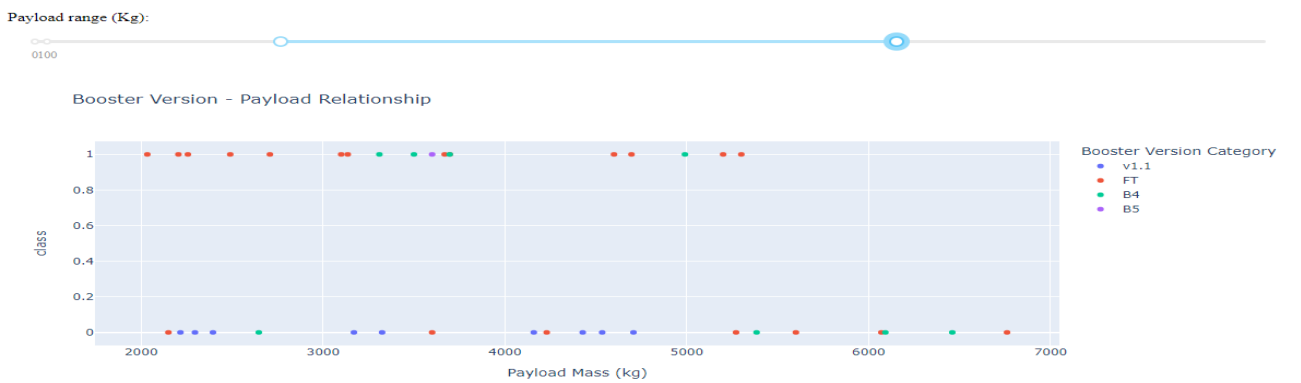


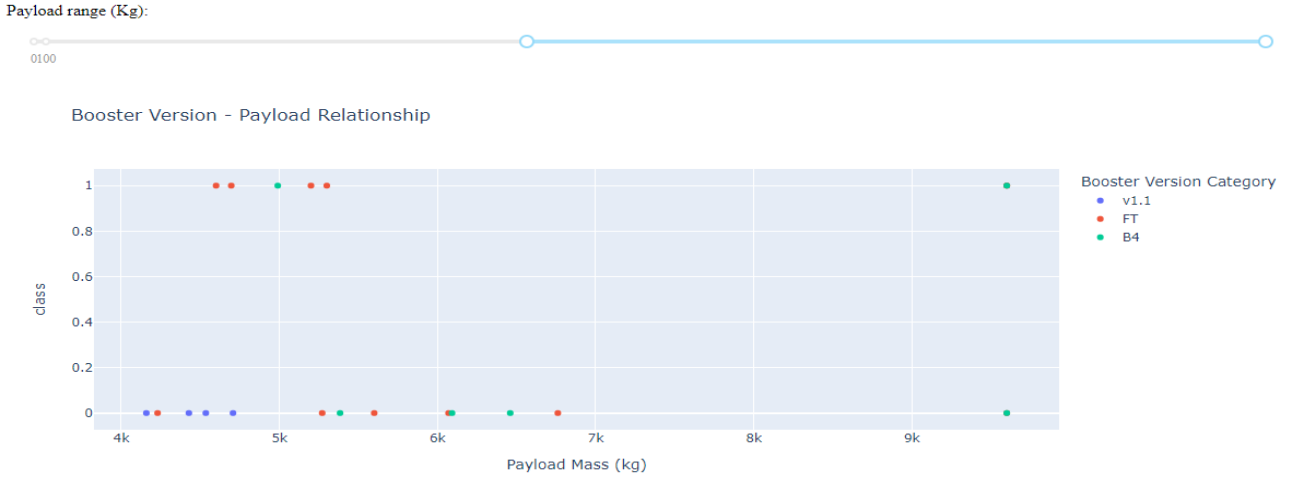
Payload Weight vs Success Rate per Booster

The slider allows the user to specify the range of payload weights to show in the scatter plot below, which relates payload mass to launch outcome, per each of the booster versions.

The screenshots show two different payload ranges (2,000 kg to 7,000 kg for the screenshot above, and 4,000 kg to 10,000 kg for the screenshot below)

The scatter plots show higher success rates at payload mass weights below 5,500 kg





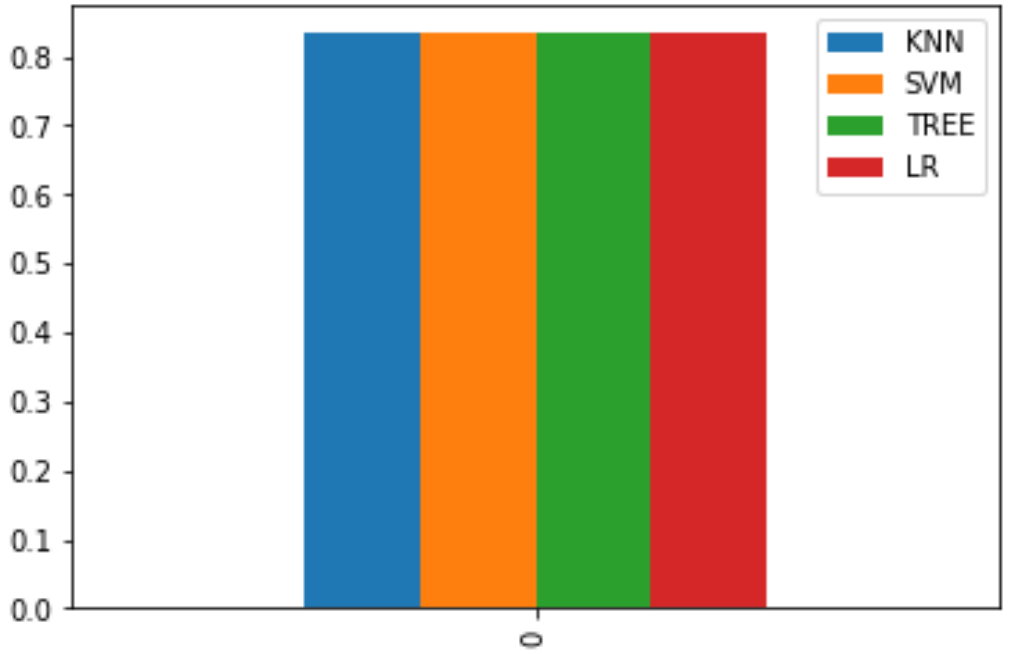
Section 6

Predictive Analysis (Classification)

Classification Accuracy

All four models show identical accuracy rate on the test data (0.83%).

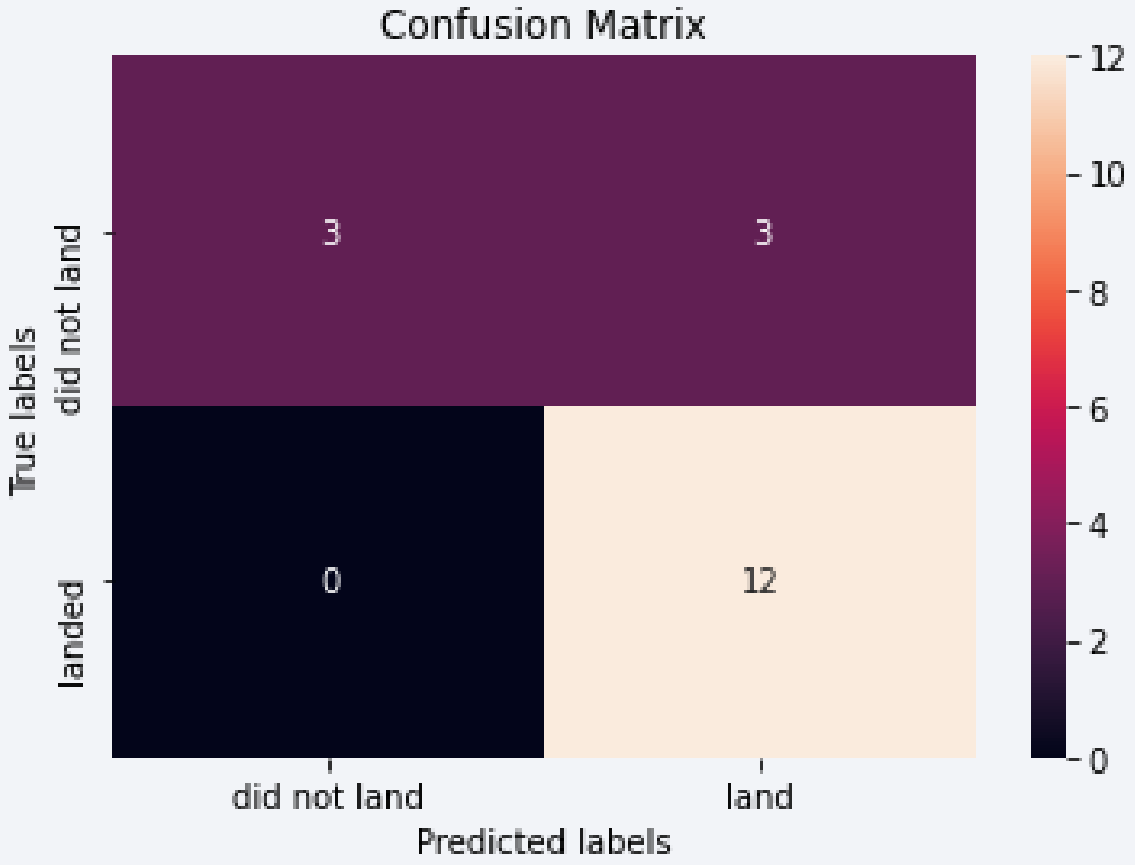
	KNN	SVM	TREE	LR
0	0.833333	0.833333	0.833333	0.833333



Confusion Matrix

The confusion matrix shows true values for successful (landed) and unsuccessful (did not land) launches in the rows, and predicted outcomes in the columns.

In the case of the KNN model, the model correctly predicted 12 out of the 12 successful launches, but it was less accurate in the prediction of unsuccessful launches: it predicted correctly 3 of the failed landings, but predicted the other 3 as successful.



Conclusions

Based on the data gathered from SpaceX website during the EDA:

- SpaceX CCASF LC-40 site was the most successful, both in number of successful launches and overall success rate (19 successful launches, and 73%, respectively)
- The payload mass range associated with highest success rate was under 1,000 kg
- The payload mass ranges associated with the lowest success rate were between 1,000 and 2,000kg, and around 9,000kg
- The most successful version of the F9 booster was the v1.1
- The types of orbits with higher success rates were ES_L1, GEO, HEO, and SSO

Thank you!