IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Aaron McGinnity
29/01/2022

Executive Summary

- Several methodologies were employed across this analysis from data collection to predictive modelling:
  (1) Data collection through API and webscraping.
  (2) Exploratory data analysis using general Python modules, but also through SQL, with visualisations.
  (3) Interactive visual analytics through Folium and Plotly dashboard.
  (4) Predictive model evaluation and implementation.
- This analysis points out several key findings, including the high success rate of SpaceX, its behaviour and preferences of its space flight strategy, and in what areas it does well.

Executive Summary

- SpaceX is a private aerospace manufacturer, designing and constructing rockets.
- Its signature design, the Falcon 9 rocket, is capable of launching into orbit and returning to land. This makes it a reusable rocket and can contribute to massive savings on the vast costs of spaceflight.
- The viability of the savings depends on whether the Falcon 9 can reliably land, and this information is useful in determining the cost of a launch and helps inform competitors.
- This project investigates the likelihood of successful landings by the Falcon 9 rocket by analysing data of past launches and building a model to predict future successes.
- The models built in this analysis show that future Falcon 9 landings have a strong probability of successful landings, based on historical experience.

Section 1

Methodology

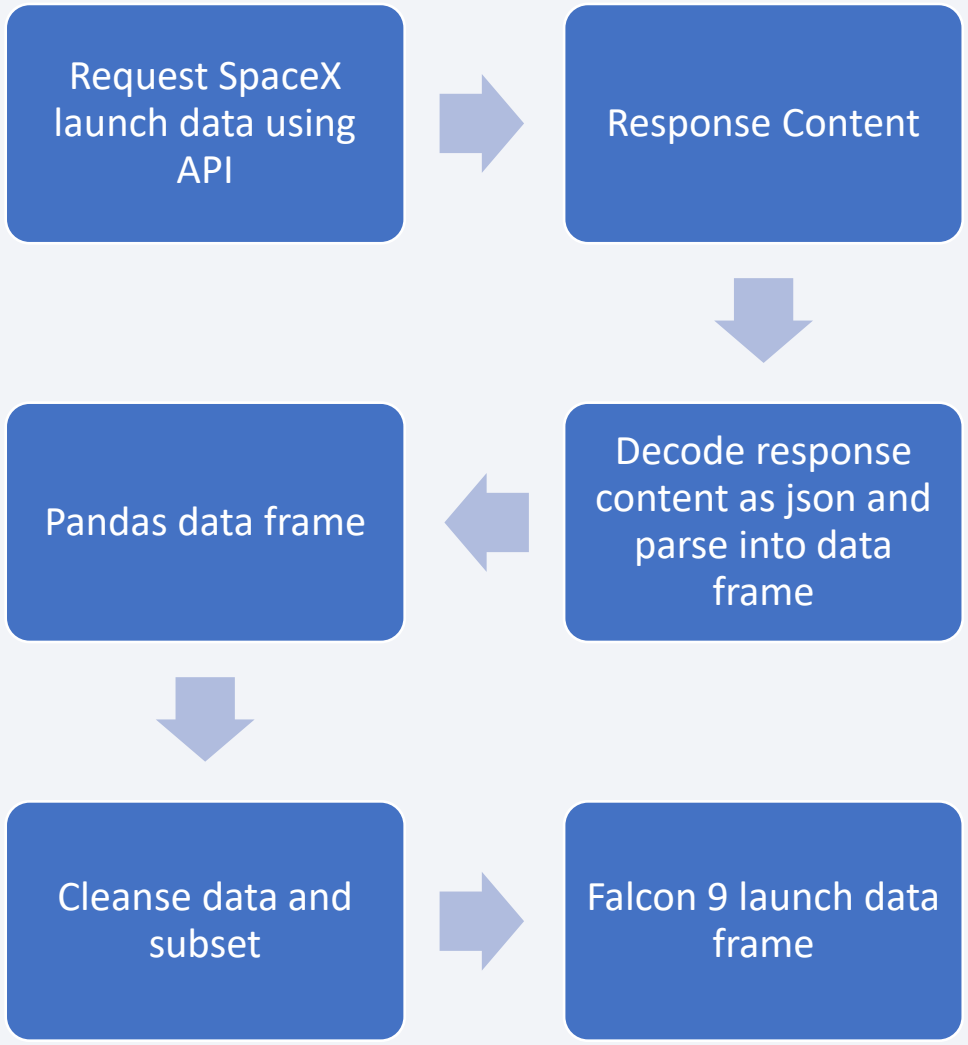
Methodology

Executive Summary

- Data collection methodology
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

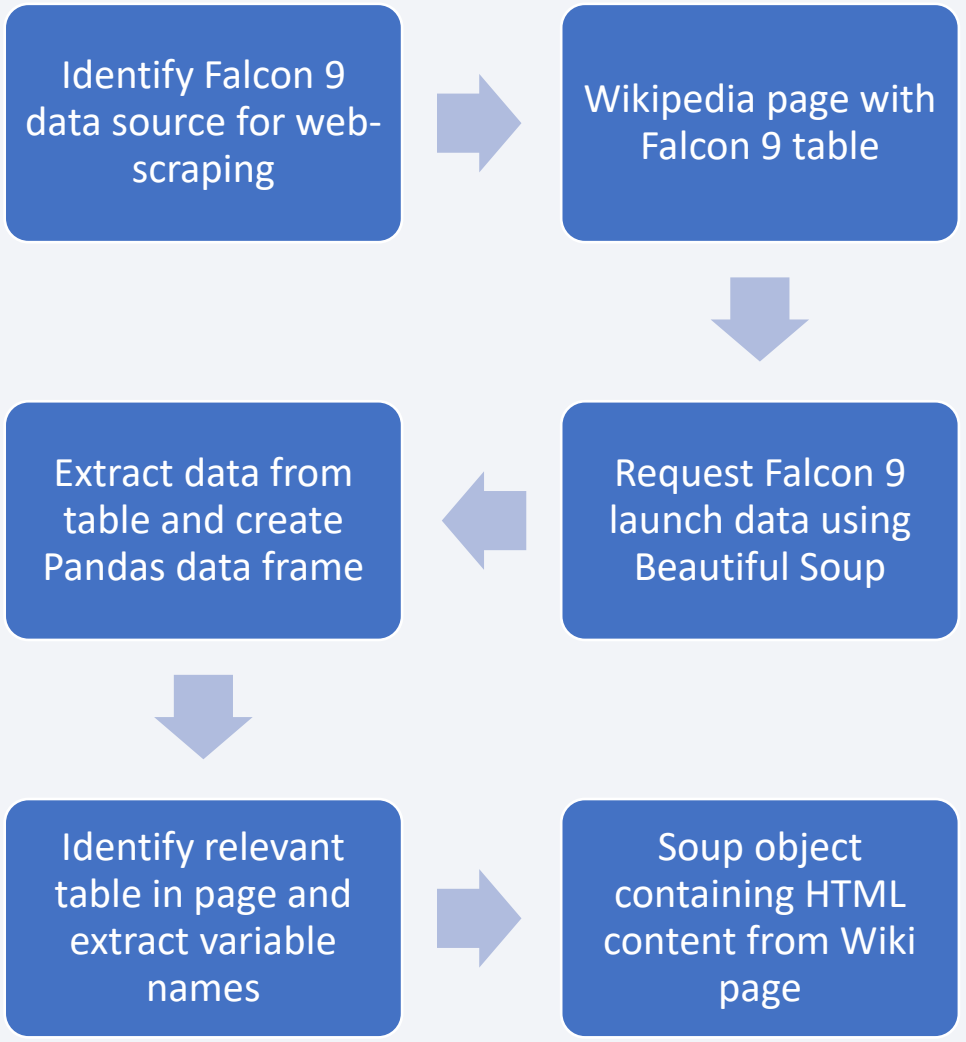
Data Collection – SpaceX API

- The SpaceX data was collected via an API according to the flow chart presented to the right.
- Github url: https://github.com/AaronMcGinnity/Python_Capstone_Project/blob/cfc388efbe36a071d6c42ab363b78143034e2838/jupyter-labs-spacex-data-collection-api.ipynb.



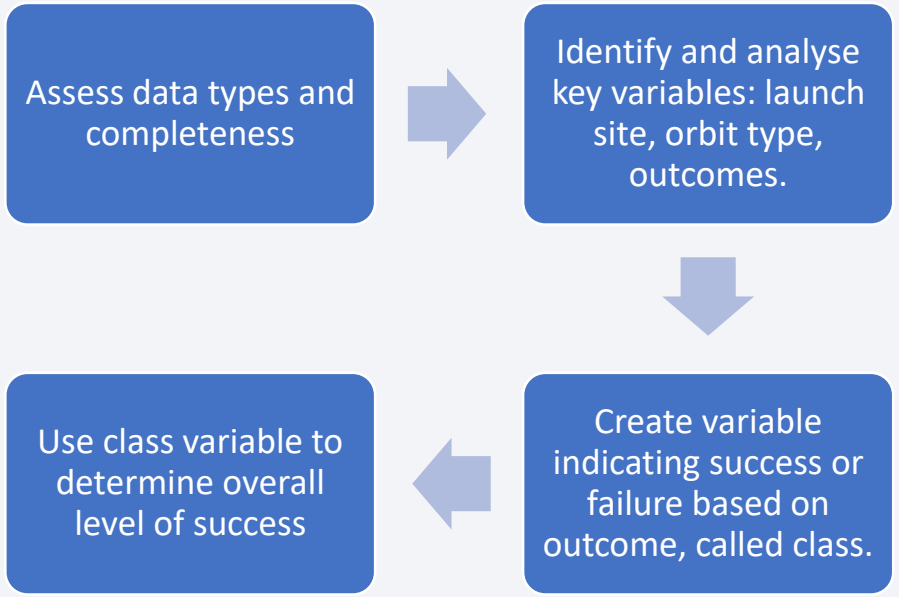
Data Collection - Scraping

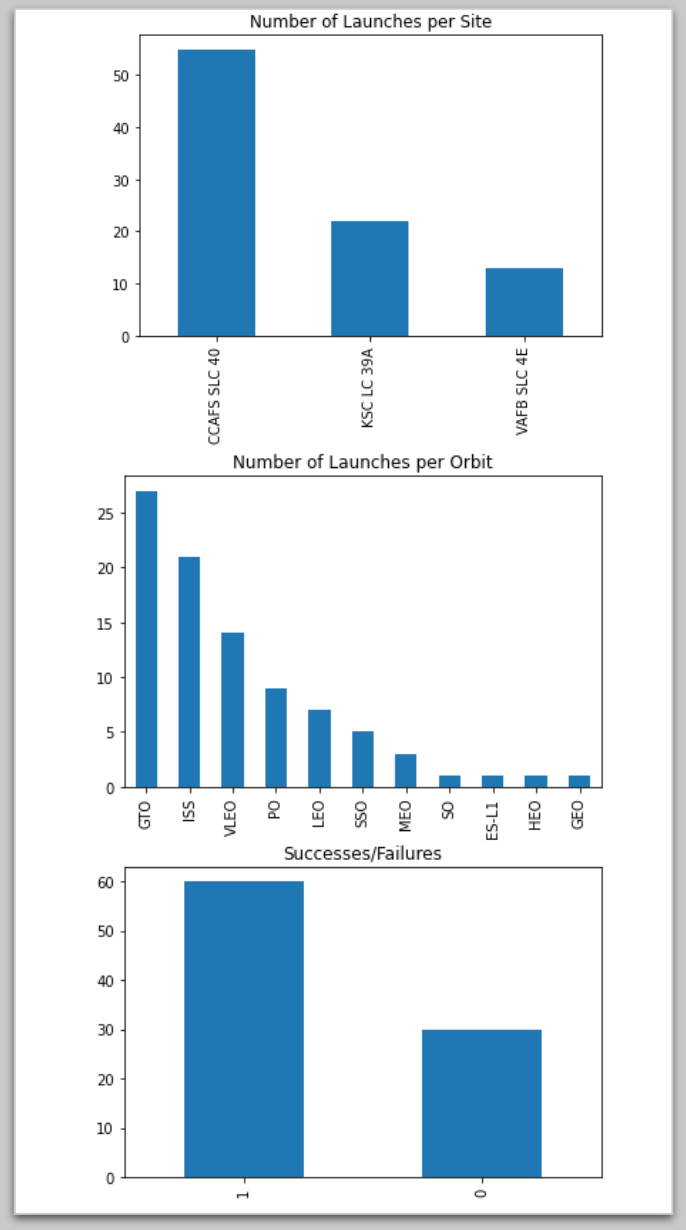
- Data was scraped from the Wikipedia webpage on the Falcon 9 launches.
- Using Beautiful Soup, the data was extracted and turned into a data frame.
- Github: https://github.com/AaronMcGinnity/Python_Capstone_Project/blob/cfc388efbe36a071d6c42ab363b78143034e2838/jupyter-labs-webscraping.ipynb



Data Wrangling

- Once the data was collected, exploratory data analysis was conducted to discover what the data showed and meant.
- Analysis showed three key variables: launch site, orbit type and mission outcome.
- Three launch sites, eleven orbit types and eight mission outcomes were identified.
- Github: https://github.com/AaronMcGinnity/Python_Capstone_Project/blob/cfc388efbe36a071d6c42ab363b78143034e2838/labs-jupyter-spacex-Data%20wrangling.ipynb





EDA with Data Visualization

- Exploratory data analysis was performed on several variables: Launch Sites, Orbit types and Number of missions per orbit.
- This EDA shows that the CCAFS launch site is most used, that the geosynchronous orbit is the most common and that there is a 60% launch success rate.

EDA with SQL

- SQL was used to perform several queries to better understand the data. This approach allowed granular EDA of such information as the date of the first successful landing, the list of boosters that have a payload of between 4,000kg and 6,000kg.
- Github: https://github.com/AaronMcGinnity/Python_Capstone_Project/blob/cfc388efbe36a071d6c42ab363b78143034e2838/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- The Folium module was used to visualize geographically the sites and launches.
- Launch sites were marked, with the number of launches labelled. Further granularity was added by adding cluster label to indicate the successes and failures.
- Github: https://github.com/AaronMcGinnity/Python_Capstone_Project/blob/cfc388efbe36a071d6c42ab363b78143034e2838/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

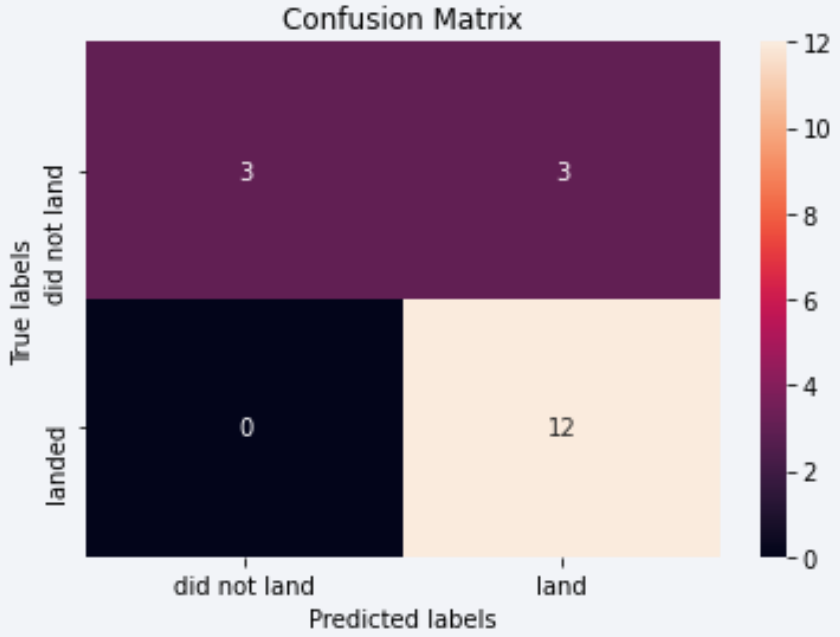
- Using Plotly Dash, a dashboard was created to graphically illustrate the distribution of successful launches across all sites, the success-failure ratio for specific sites and the successes by payload.
- Pie charts and scatterplots were built into the dashboard and a drop-down menu allowed selection of specific sites.
- The pie charts were suitable for displaying the distribution of successful launches across the sites. In this way, the most successful launch sites could be easily identified.
- The scatterplots were useful in showing the distribution of successes and failures across a payload range, with an added view of booster versions.
- GitHub URL: https://github.com/AaronMcGinnity/Python_Capstone_Project/blob/4e3c62a1b79aea78c7847578ede2ac2300055adf/spacex_dash_app.py

Predictive Analysis (Classification)

- To ensure a suitable survey of possible approaches was taken to determine how to calculate predictions, four models were evaluated. They were:
  - Logistic regression
  - Support Vector Machine
  - Decision Tree
  - K-Neighbours
- All four models showed equal levels of accuracy.
- GitHub URL: https://github.com/AaronMcGinnity/Python_Capstone_Project/blob/285e2da9b084f03cca7dd400ed26b58f823374c0/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

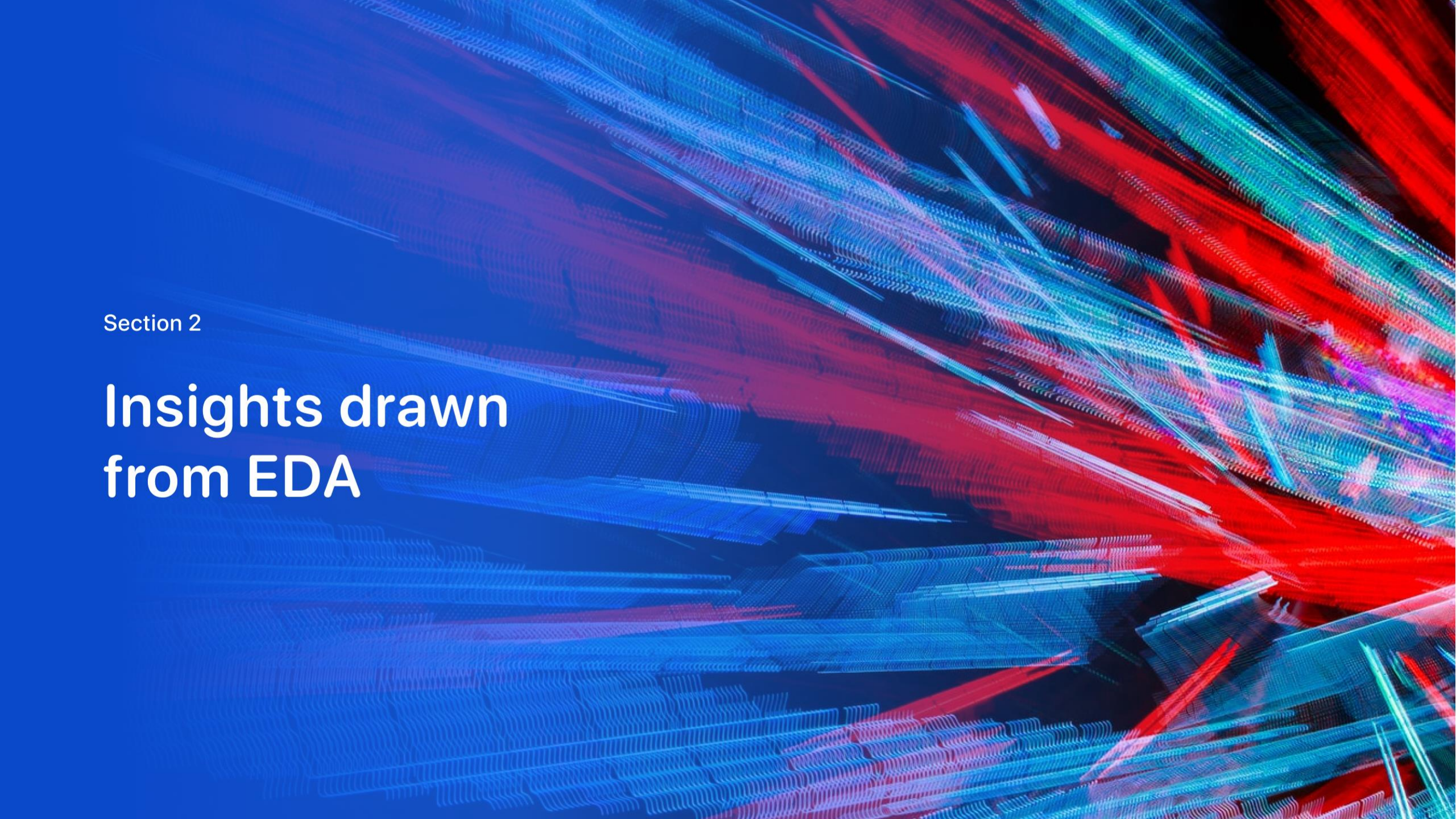
Results

- EDA has identified a suitable variable on which to base the model, i.e. mission success, which could be grouped by launch site, orbit type, etc.
- Each model that was tested is equally accurate with no false positives though a few false negatives, as seen in the confusion matrix below (which is applicable to all models).



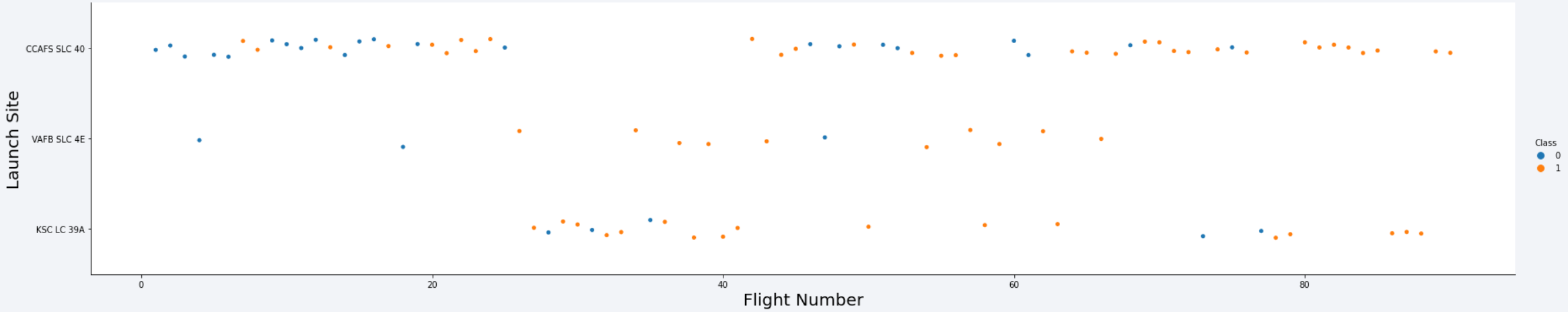
Section 2

Insights drawn from EDA



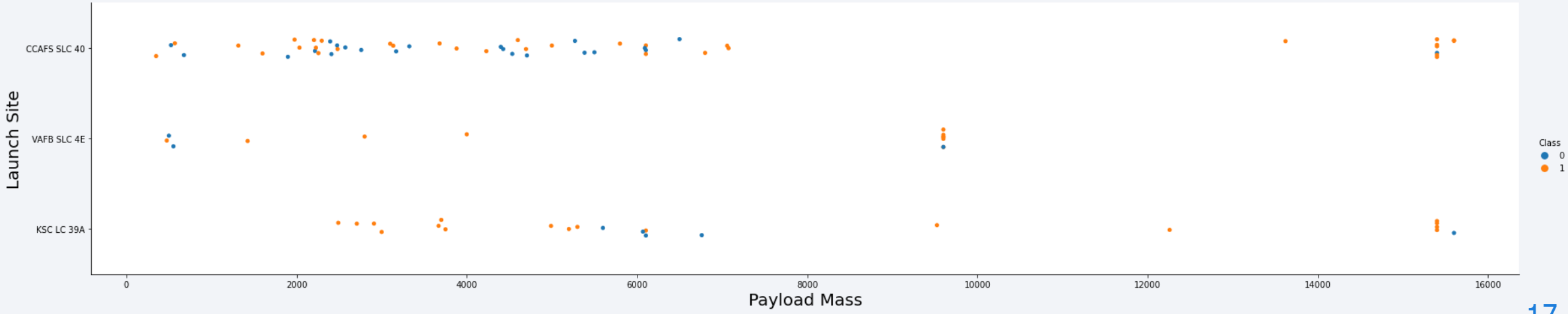
Flight Number vs. Launch Site

- Below is a scatter plot between the flight number and the launch sites.
- This plot shows that site CCAFS SLC 40 has been used throughout nearly all flights, beginning with a low success rate, but generally improving. There is a gap between 25 and 40, around the point where KSC LC 39A appears.
- Site VAFB SLC 4E has had the fewest flights, and three failures.
- Site KSC LC 39A only started flights around flight 25 and have continued sporadically over time.



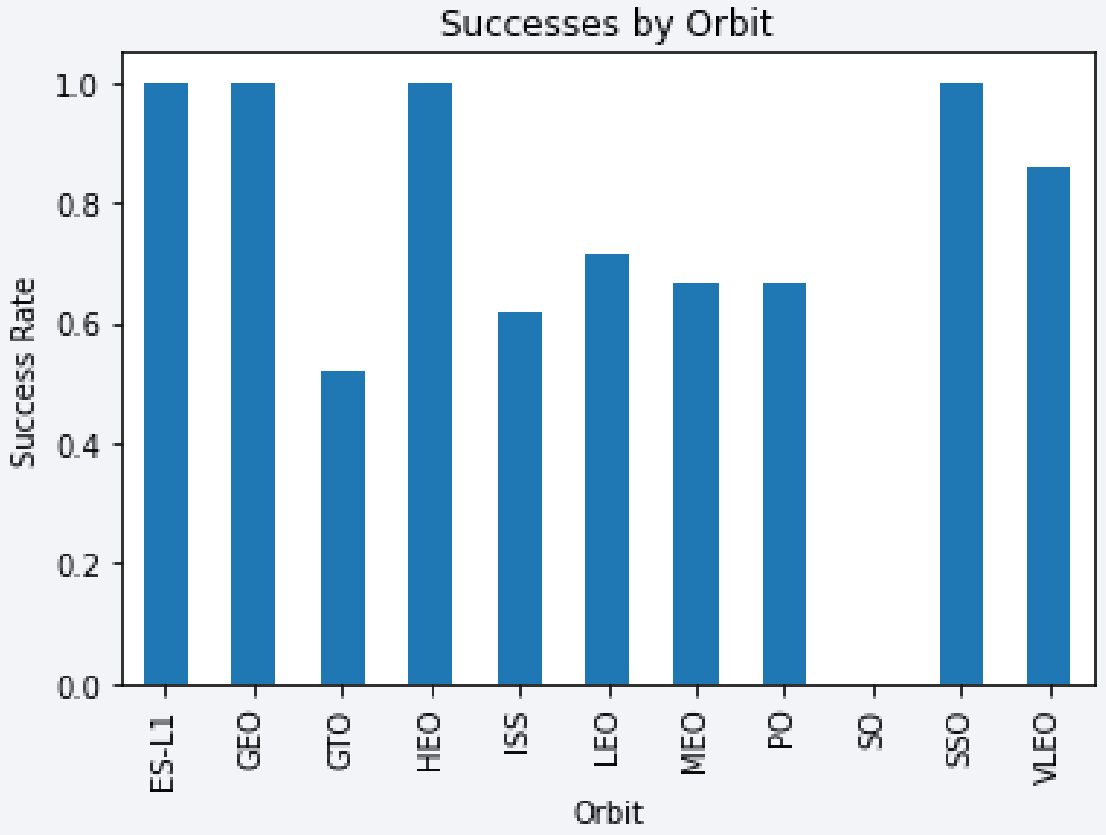
Payload vs. Launch Site

- The scatterplot below shows the flight number against the payload mass.
- Both CCAFS SLC 40 and KSC LC 39A have lower payload masses but also show a cluster of high-mass payloads. In contrast VAFB SLC 4E has launched lower to medium payloads, with none over 10,000 kg.
- Observing the hue of the observations, success is low for CCAFS SLC 40, though the few high-mass payloads have a good success rate. In comparison, the other sites seem to be more successful.



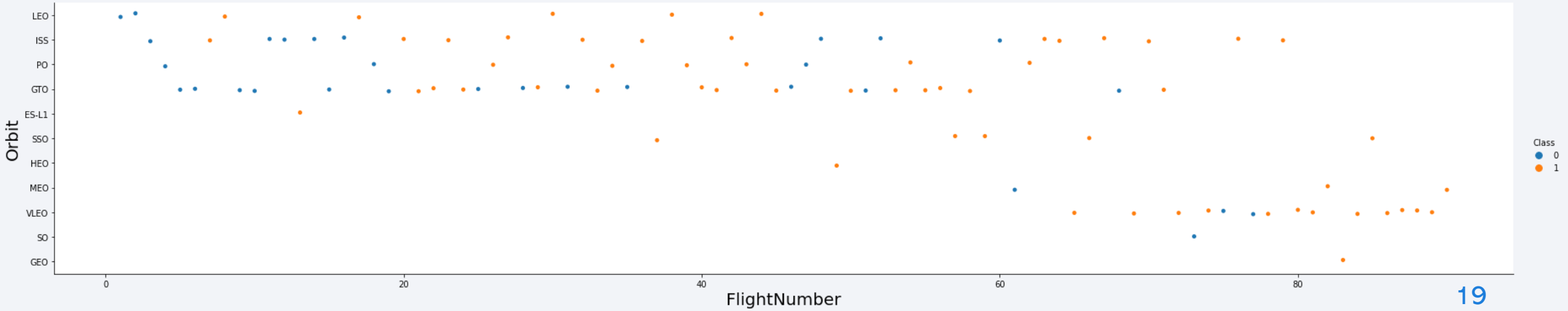
Success Rate vs. Orbit Type

- This bar chart shows the average success rate of launches by each orbit type.
- ES-L1, GEO, HEO and SSO orbits show a perfect record, with a 100% success rate.
- The SO orbit is a complete failure with a 0% success rate.
- The rest of the orbits range between middling to high success rates.



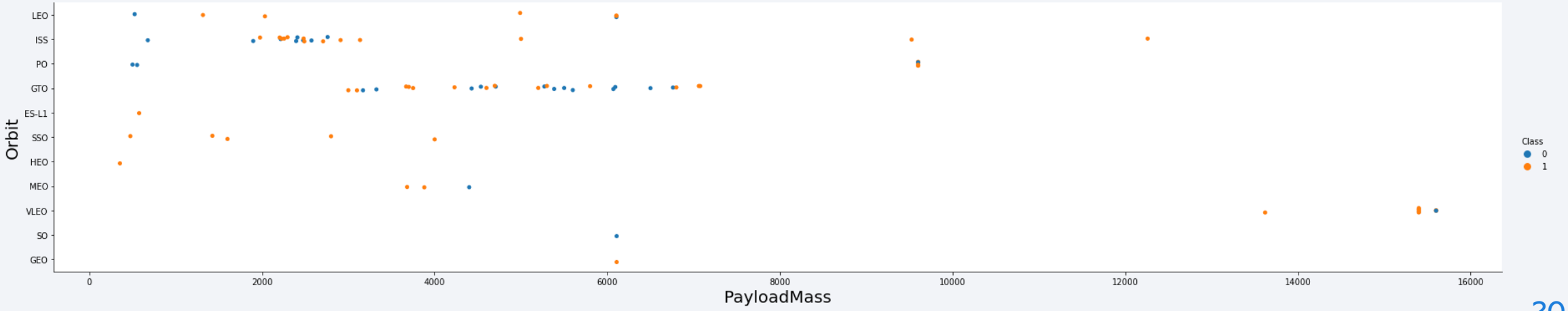
Flight Number vs. Orbit Type

- This scatterplots breaks down the bar chart from the previous slide by the flight number. Looking at this, many of the failures can be attributed to the early flights. Over time, success rates improve and flights begin to be launched to other orbits.



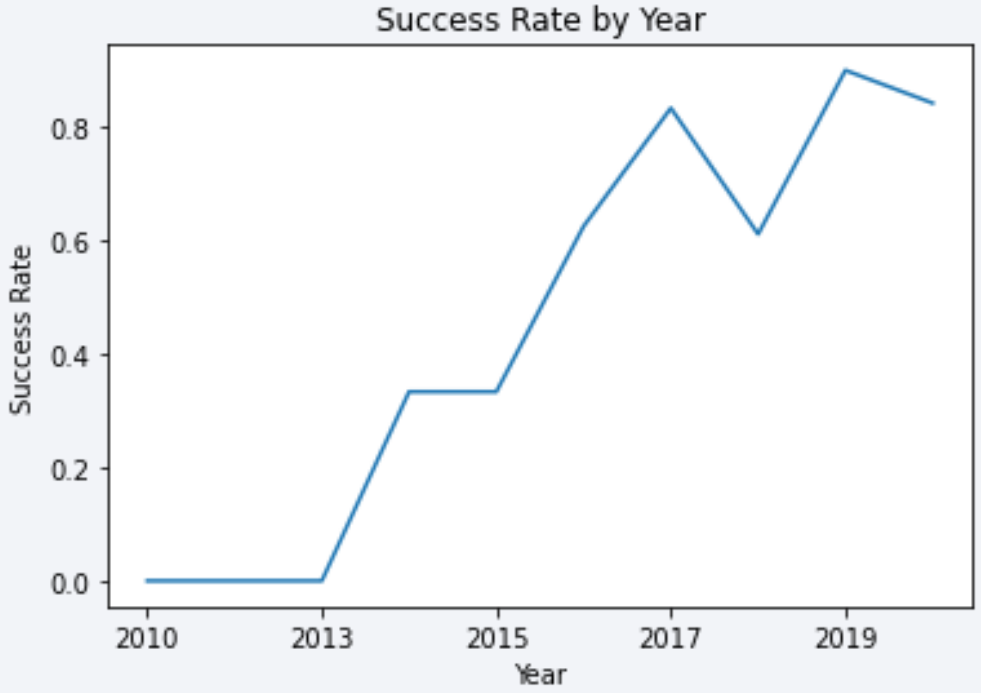
Payload vs. Orbit Type

- The scatterplot below shows the payload mass against the orbit.



Launch Success Yearly Trend

- This line chart graphs the success rate by year from 2010 to 2020.
- It shows a clear increasing trend over time, peaking at around 85% in 2019.



All Launch Site Names

- The name of the launch sites are in the table to the right.
- Only four launch sites have been used for the Falcon 9 rocket.

Launch Sites
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- 5 records with a launch site beginning with CCA were selected from the data, with the resulting table copied below.

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

Total Payload Mass

- The total payload mass was calculated carried by all boosters launched by NASA. The total was 45,596 kg.
- This makes NASA the second largest customer by payload amount, behind only SpaceX itself.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was calculated, coming to 2,279 kg.

First Successful Ground Landing Date

- The first successful landing occurred on 12 December 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- There were seven different boosters that had a successful landing with a payload between 4,000kg and 6,000kg.
- The list of boosters are shown below:

boosters
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1
F9 FT B1029.2
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1038.1

Total Number of Successful and Failure Mission Outcomes

- The number of accounts split by failure and successful outcomes is shown in the table below. There is only one failure and 100 successes.

mission_outcome	Number of Accounts
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The table to the right shows the booster versions which have carried the maximum payload mass of 15,600kg.
- There are 12 different boosters, all version F9 B5, with one launch each. All had a successful launch, but two had a failed landing.

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

2015 Launch Records

- The launches with failed landings in 2015 are shown in the table below. There are only two observations.

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

- A table ranking the landing outcomes by number of launches is shown to the right.
- The highest ranking outcome is ‘No attempt’.
- Failure and success for drone ships are tied in second place.

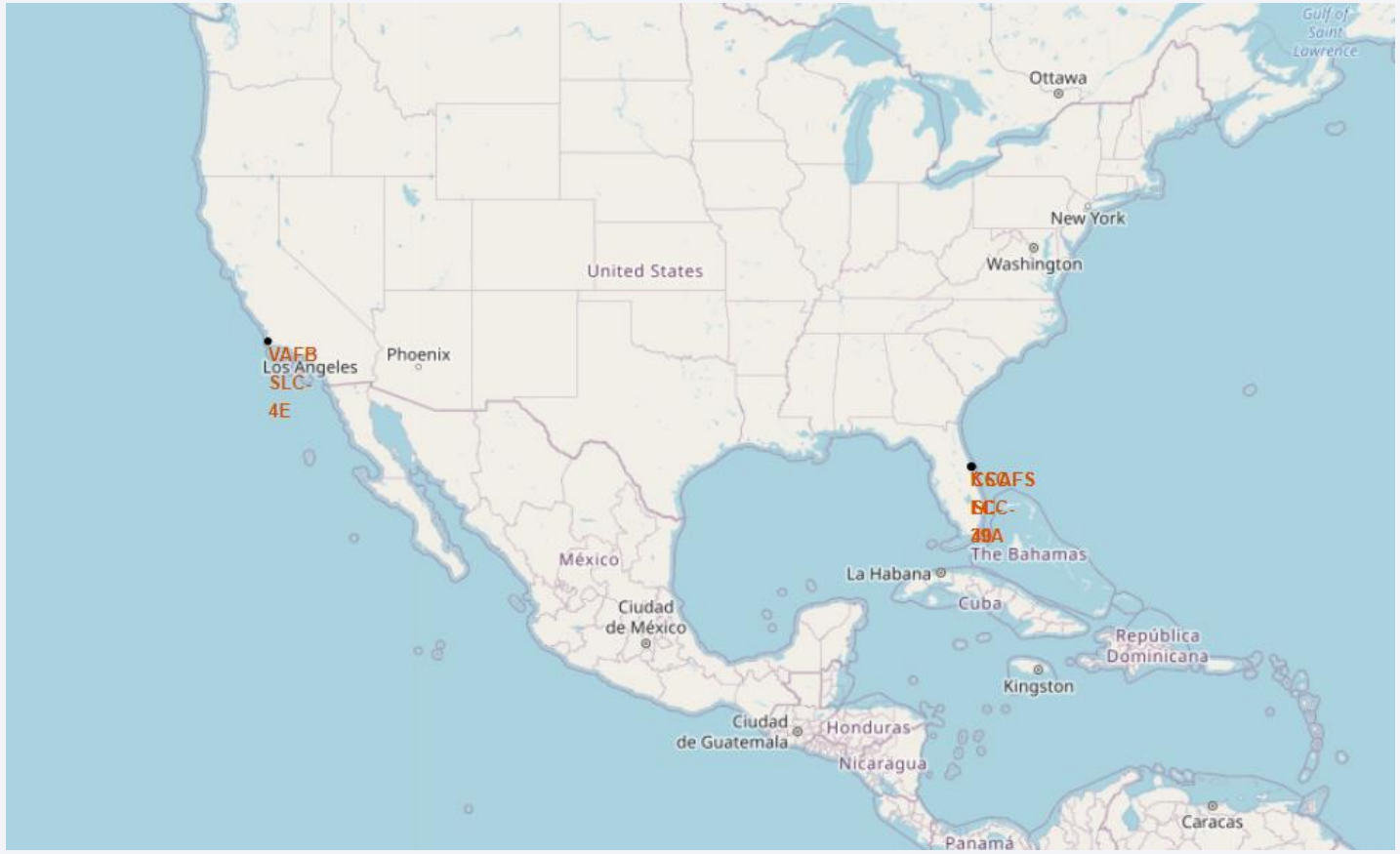
landing__outcome	total_number
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

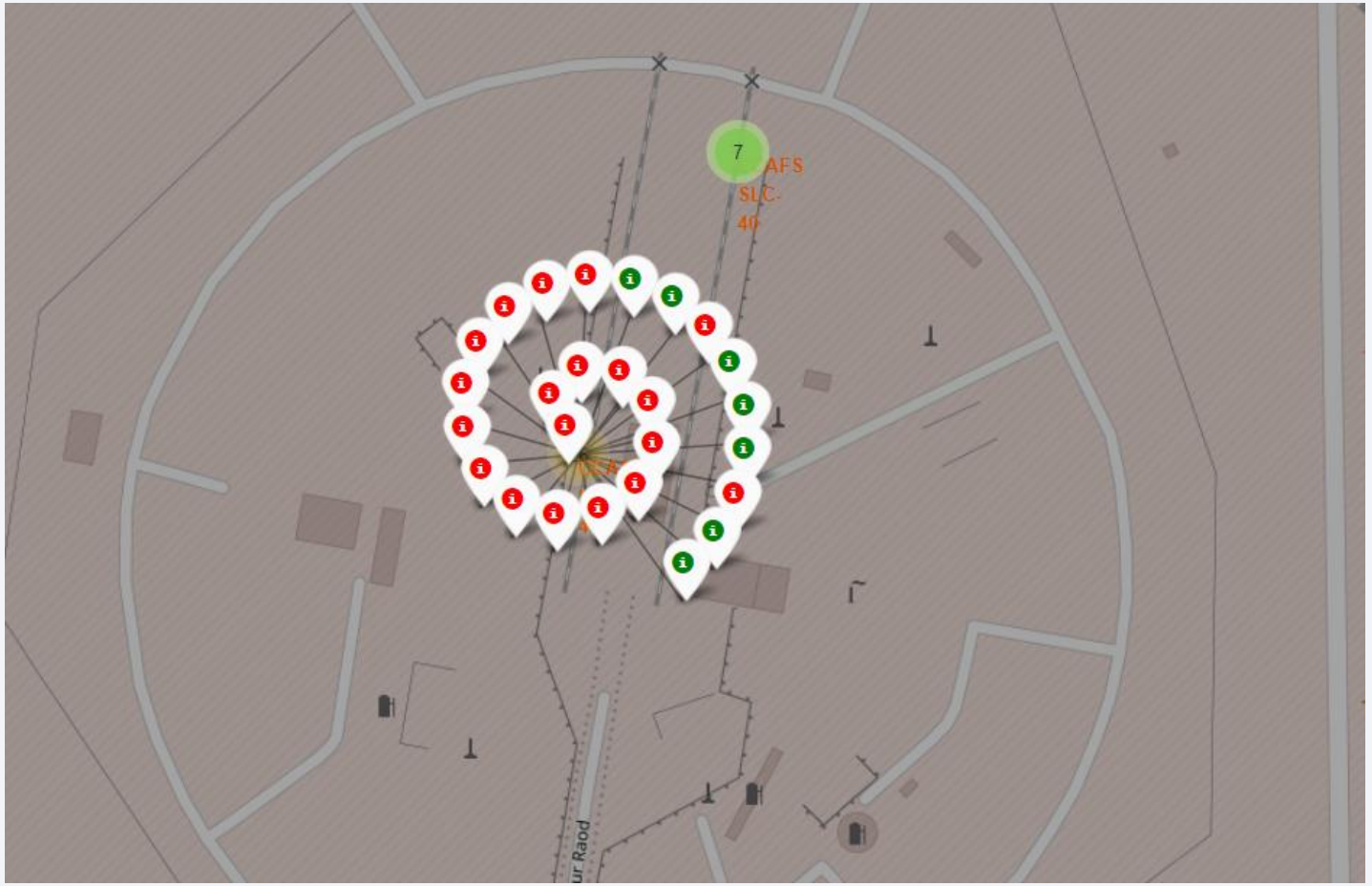
Launch Sites

- The Folium module has been used to create a map to geographically represent where the launch sites used by SpaceX are.
- There are two locations marked. One near Los Angeles, California.
- Two are in Florida. All are located very close to the coast.



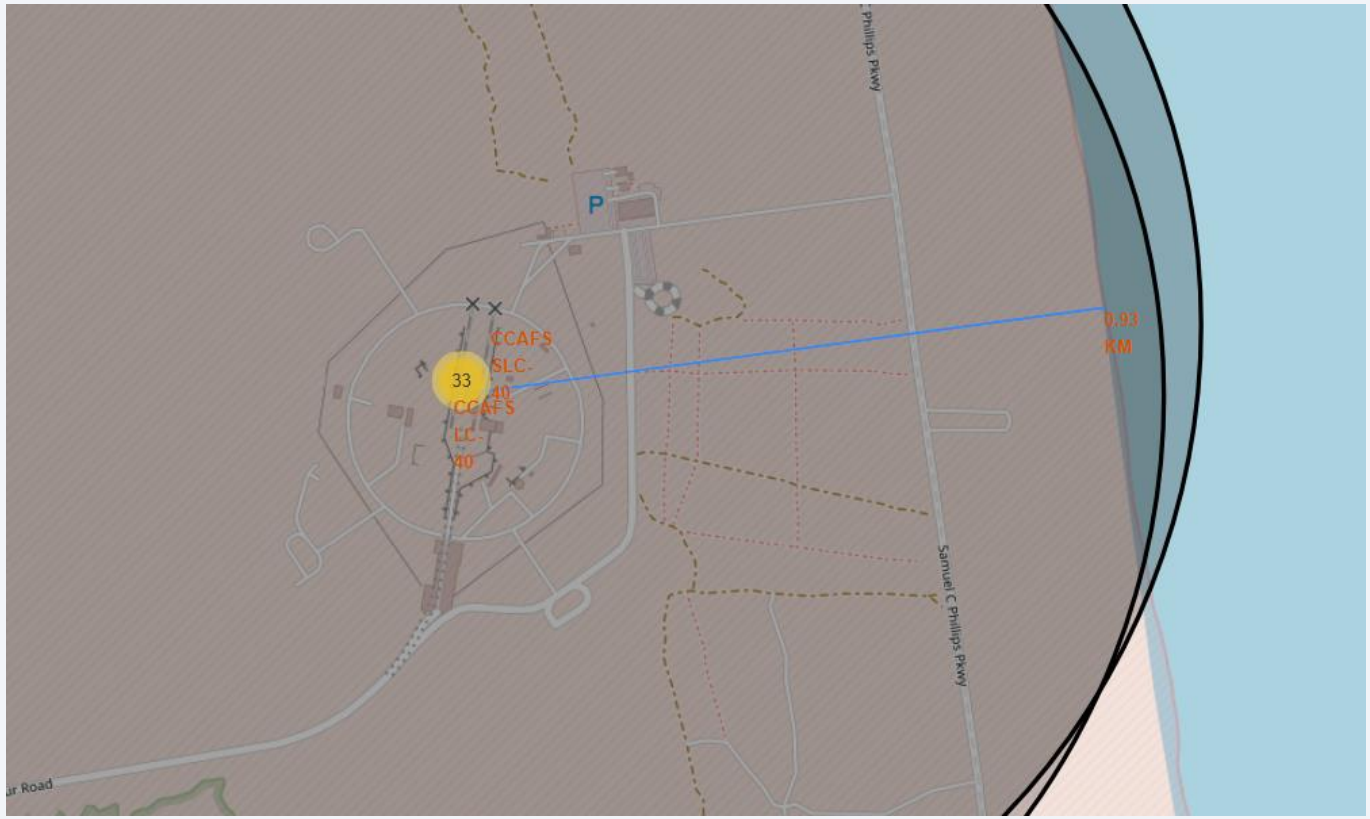
Launch Outcomes

- The map now shows a zoomed-in look at the Florida launch site, CCAFS LC-40.
- The Folium module has been used to cluster labels to visualize the successes and failures of the site.
- It shows many failures have occurred at this site.



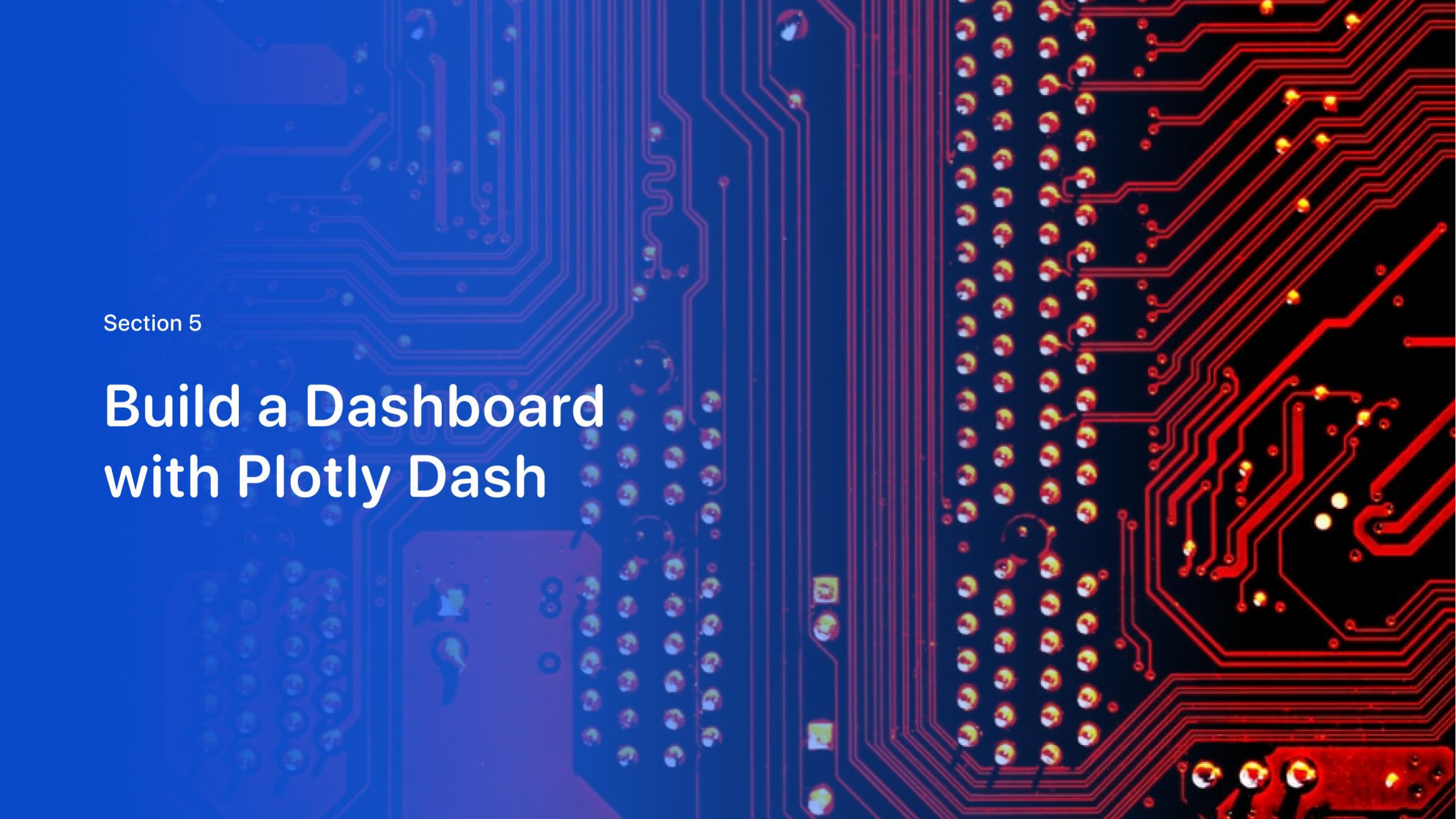
Launch Site Proximities

- The same location is now shown with the distance drawn between the launch site and the coast.
- This shows that the launch site is within 1km of the coast. The proximity to the coast therefore seems a significant factor in the launches.



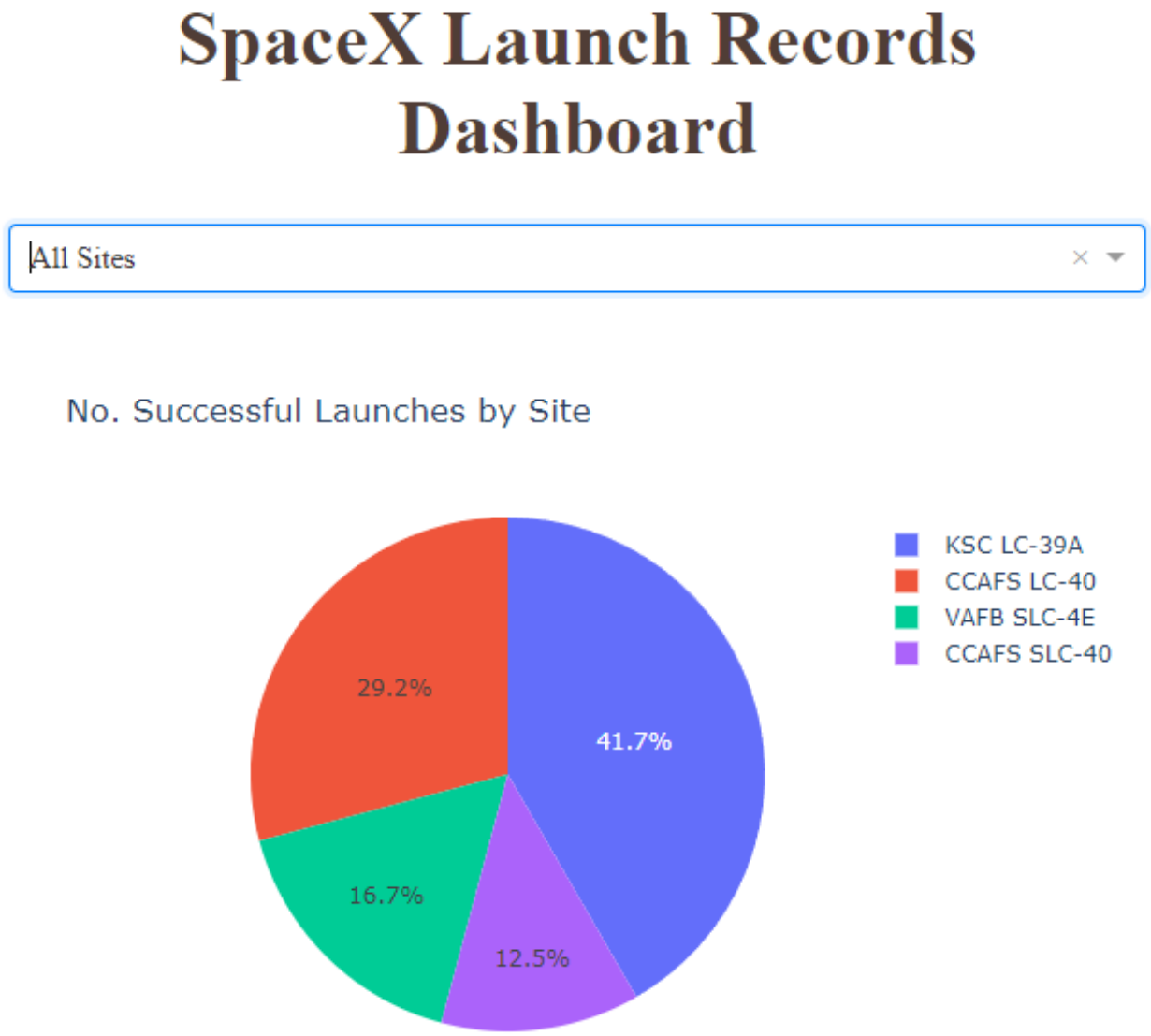
Section 5

Build a Dashboard with Plotly Dash



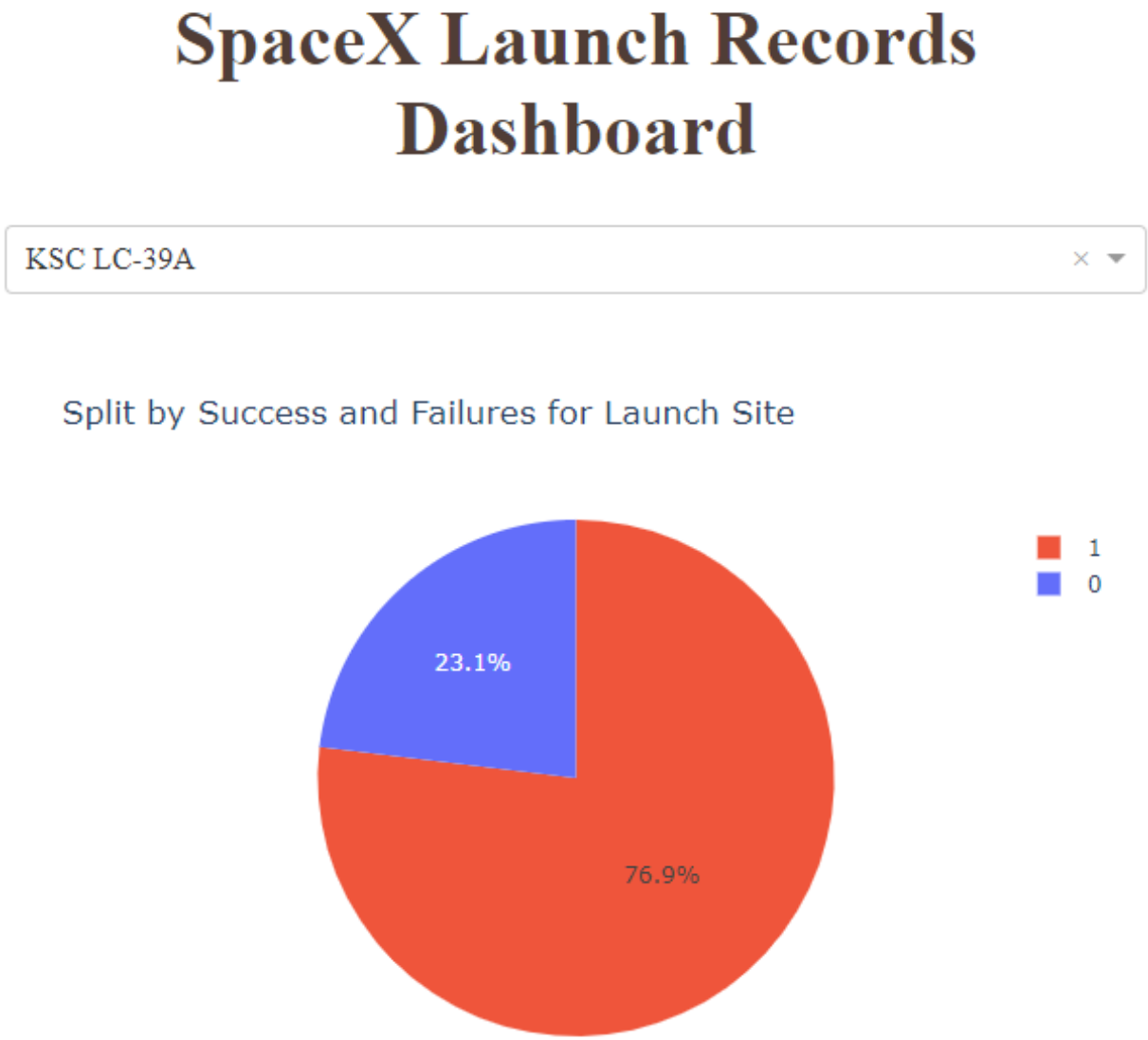
SpaceX Successful Launches by Site

- A pie chart with the proportional distribution of successful launches across the sites is shown.
- The KSC LC-39A site is shown to have the largest share of successful launches at 41.7%.
- The lowest share is held by CCAFS SLC-4 with 12.5%.



Most Successful Launch Site

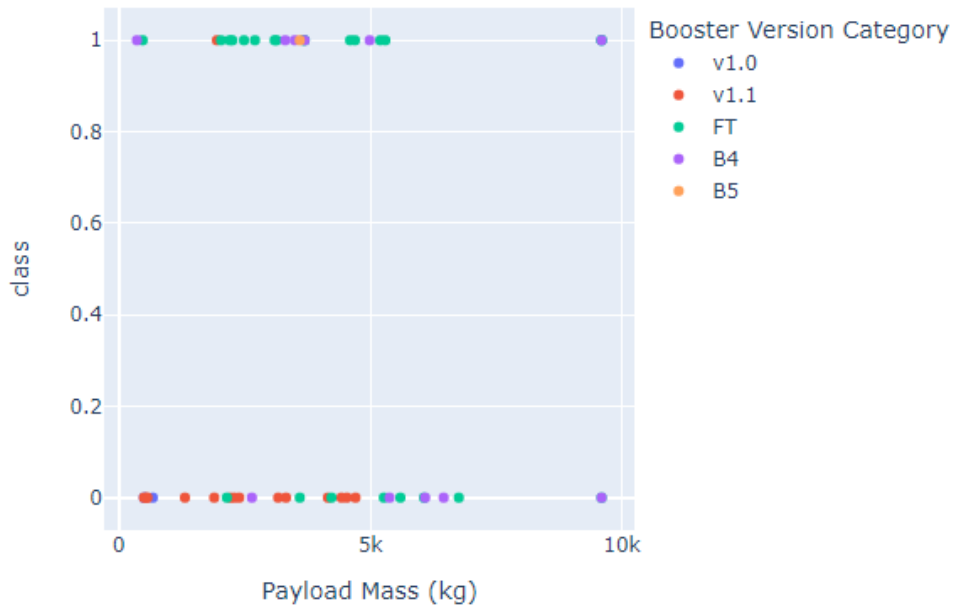
- The pie chart shows the distribution between successful and unsuccessful launches for the KSC LC-39A site.
- This is the most successful site, with 76.9% of its launches being successful.



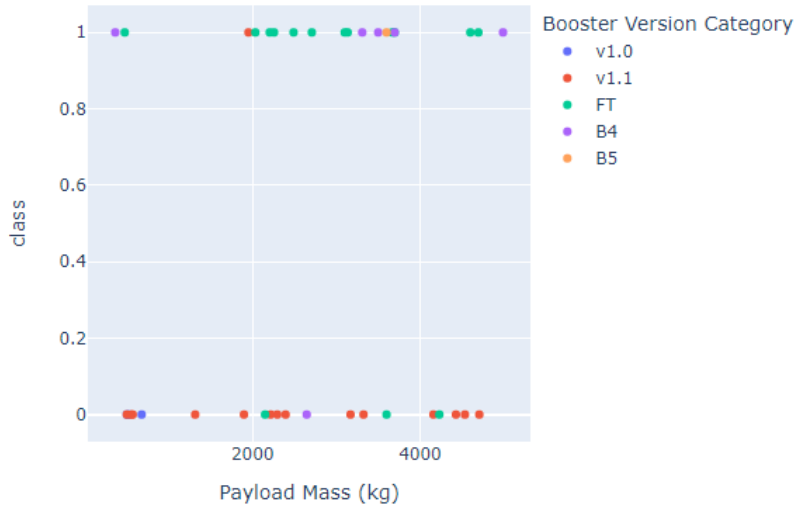
Payload vs. Launch Outcome Scatter Plot

- Scatterplots of the successful launches across all sites are shown below. Plot A shows the points across the full payload range, going from around 2,500kg to nearly 7,000kg. Plot B shows the points using the lower half of the range, while Plot C shows the points for the upper half of the range. The different booster versions are displayed in different colours.
- The FT booster appears to perform the best for all payloads while the v1.1 booster performs poorly for all payloads < 5,000kg; it was not used for higher-mass payloads.

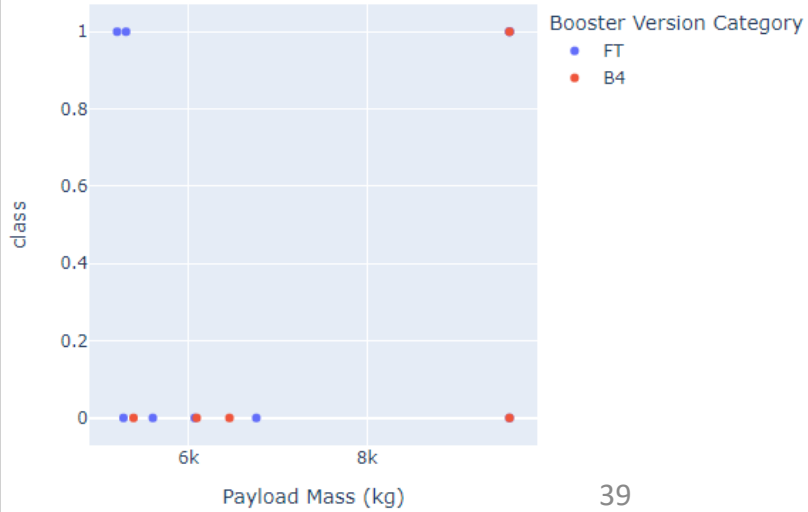
A) All Payloads



B) Payloads < 5,000 kg



C) Payloads > 5,000 kg

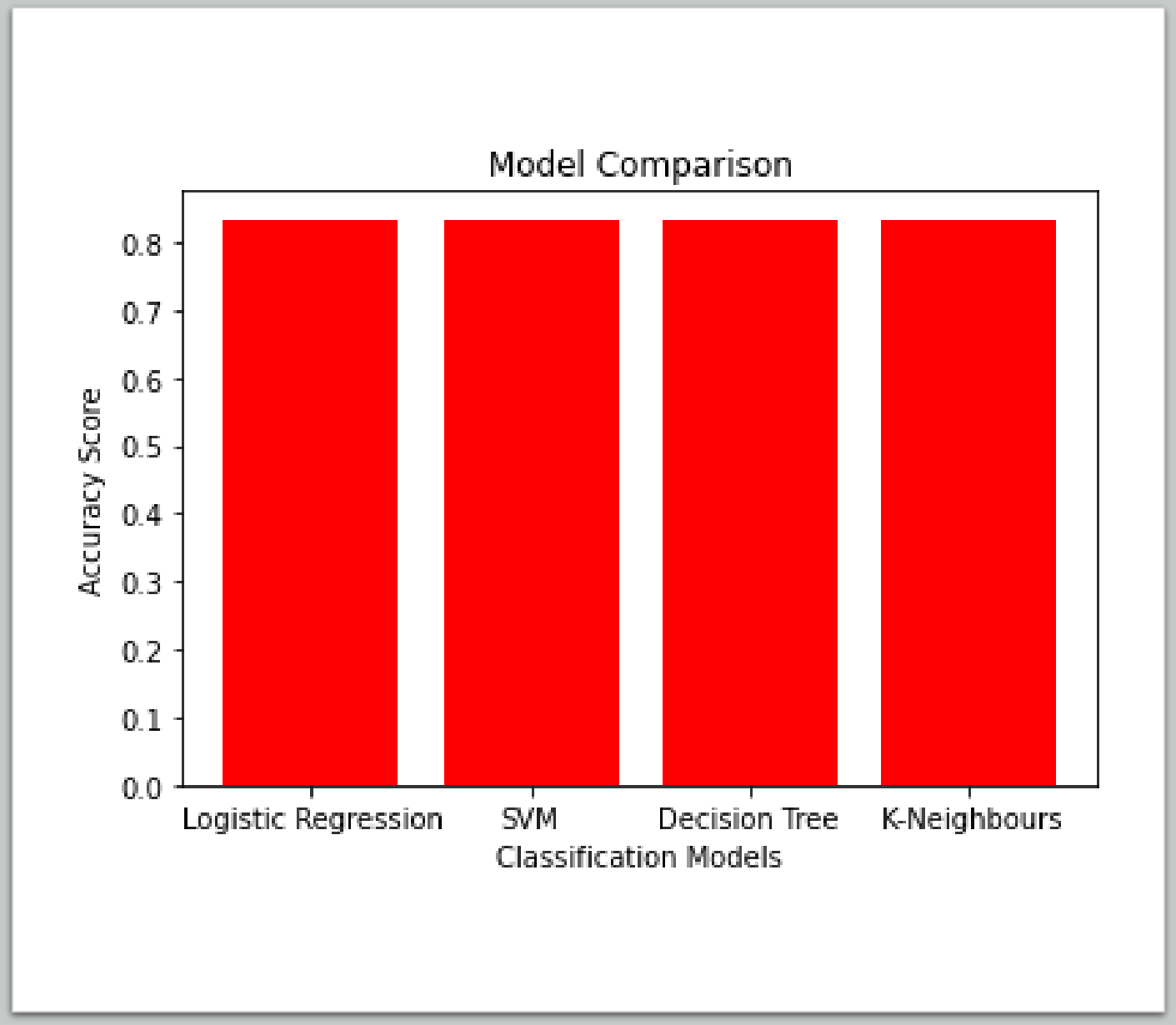


Section 6

Predictive Analysis (Classification)

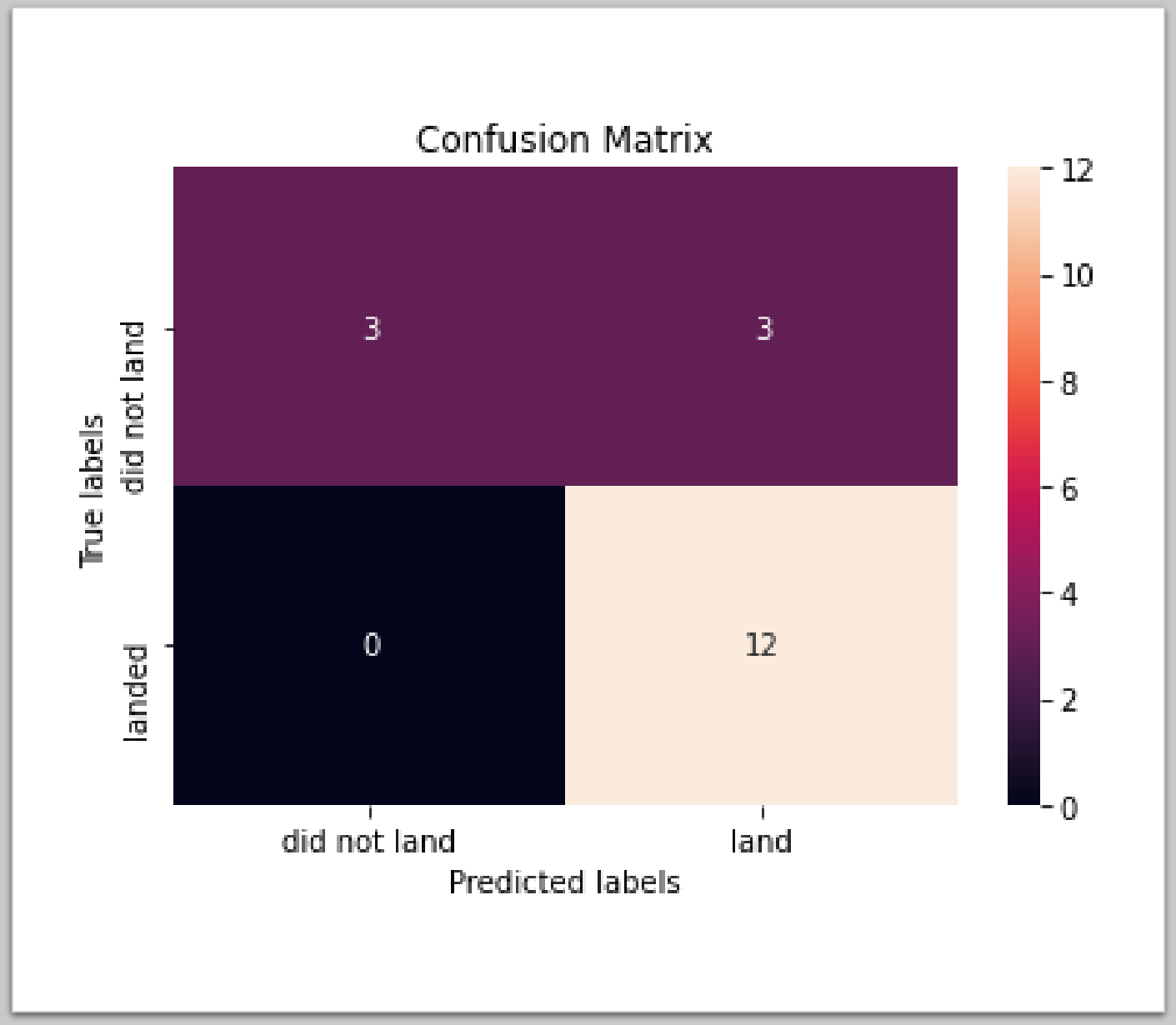
Classification Accuracy

- Four models were constructed to classify the land/did-not-land outcomes for the SpaceX flights: logistic regression, support vector machine, decision tree and K-Neighbours.
- Data was split into train and test subsets. Each of the models were trained and then their accuracy tested.
- Each model had the same accuracy score at 83%, as the bar chart shows.



Confusion Matrix

- The confusion matrix for the chosen model is shown to the right.
- The model is successfully at classifying the outcomes. However, there are three false positives
- Note that the number of observations is relatively low. Only 18 observations are in the test dataset. If one of these observations had a different outcome, the results here would look fairly different. Therefore, conclusions should be made with caution.



Conclusions

- From the analysis, SpaceX is shown to have become a very successful space flight organization, with a high success rate. Therefore, costs of spaceflight missions using the Falcon 9 rocket are very likely to be considerably reduced.
- Therefore, competitors can take note that the Falcon 9 design is successful, and that SpaceX's strategy in designing and building reusable rockets is viable.
- From the data, very predictive models can be built and with a 83% accuracy rate. Two important facts must be noted:
  (1) The population is very small. Only 90 observations were used, and these were split further between test and train datasets. Therefore, caution should be exercised when making any claims with this model.
  (2) The model is very simple with no segmentation and little granularity. In effect, it ignores various important factors such as payload mass, booster version. These could all play a role. Nevertheless, the model does give an effective overview of SpaceX success as a whole.

Thank you!