IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Stijn van der Lippe
2024-04-05

Outline

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

Executive Summary

- In this work, we aim to predict if the first stage of a SpaceX rocket will be reused based on historical launch data, using a variety of classification algorithms.
- Specifically, we compare the performance of logistic regression, k-nearest neighbours, decision trees, and support vector machines.

- Our findings show that for this problem all classification algorithms have the same classification accuracy.
- Since the performance is the same across all methods, we decide to use a decision tree, since the classification process is easy to interpret for a human.

Introduction

- SpaceX is an aerospace company that can achieve much cheaper rocket launches by reusing the first stage of their rocket.
- For the reuse to be possible the first stage must land successfully, for instance on a drone ship at sea.

- We aim to predict if the first stage of the SpaceX rocket will land successfully, based on historical launch data.
- Additionally, we aim to get insight into the historical launch data. For instance, we want to identify if there is a relationship between a variety of features and successful landings (e.g., payload mass, launch site, booster version, etc.).

Section 1

Methodology

Methodology

Executive Summary

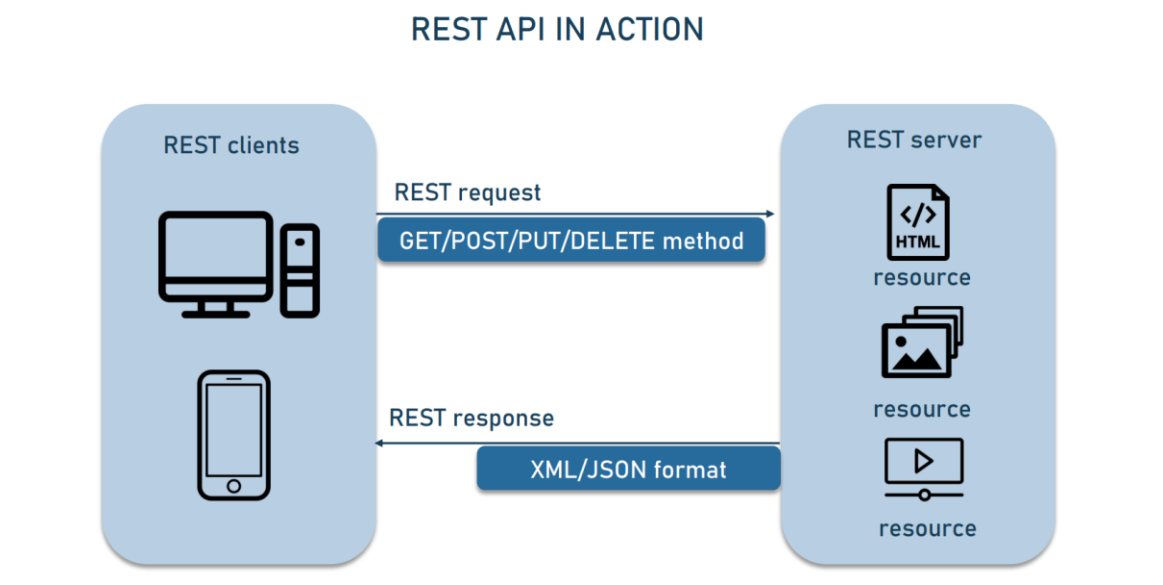
- Data collection methodology:
  - The historical launch data was gathered from a variety of sources.
  - This includes web scraping to get a list of historical launches and REST APIs to get specific details of each launch.
- Perform data wrangling
  - Before we can get insights into the gathered data, we need to process it. This includes removing null values, removing unnecessary data, and normalizing the data.
- 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 compare a range of algorithms, which we optimize to find the best algorithm.

Data Collection

- The historical launch data was gathered from a variety of sources.
- This includes REST APIs and web scraping, to get a list of launches and the details of each launch.
- All this data is necessary to answer the questions we posed earlier.

Data Collection – SpaceX API

- As a first step, we use the SpaceX API to get information about each launch.
- For this, we follow the procedure in the flowchart shown to the right.
- The result of our request is a .json file, which we normalize and output into a Pandas DataFrame.
- The source code for the collection of data using the SpaceX API can be found here.



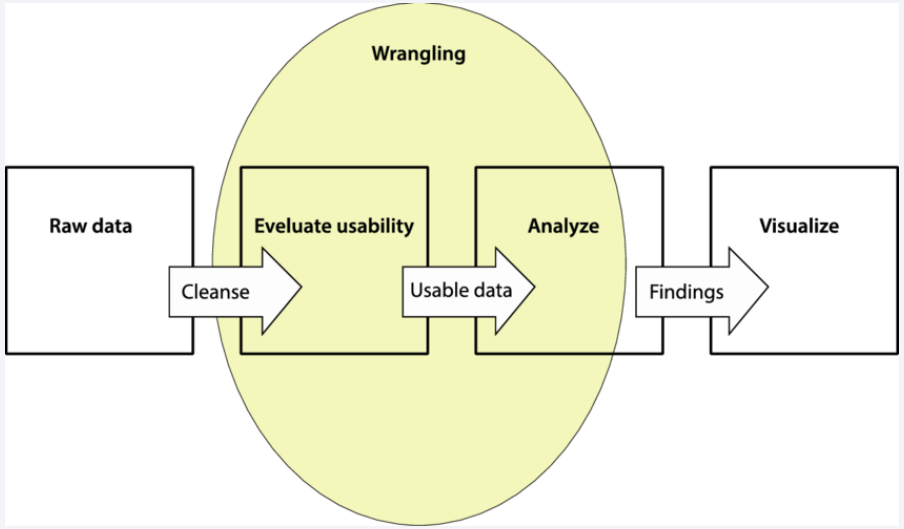
Data Collection - Scraping

- For the second step, we consult the web to gather more information about each launch.
- To automate this, we use web scraping to scrape Wikipedia tables, following the procedure shown to the right.
- The HTML tables are processed using the BeautifulSoup library in Python.

- The source code for the web scraping can be found here.

Data Wrangling

- Before we can analyze the gathered data, we need to process it.
- These steps include removing null data, removing unnecessary information and creating the desired outcome of our classification algorithms.



- The source code for the data wrangling process can be found here.

EDA with Data Visualization

- We gather initial insights after the data wrangling process. This allows us to get an initial idea of how the data is distributed, and what features may be related to the successful landing of the first stage rocket.

- The source code for the exploratory data analysis can be found here.

EDA with SQL

- We use SQL to gather additional insights into our data. This includes:
  - We find the total payload mass of boosters launched by NASA (CRS).
  - We find the average payload mass carried by booster F9 v1.1.
  - We find the date of the first successful landing on a ground pad.
  - We find the names of the boosters which have successfully landed on a drone ship and have payload mass greater than 4000 kg but less than 6000 kg.
  - We find the total number of successful and failure mission outcomes.
  - We find the names of the booster versions which have carried the maximum payload mass.
  - We find the month names, booster versions and launch sites for failed landings on a drone ship in the year 2015.
  - We find the ranking 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.
- The source code for EDA with SQL can be found here.

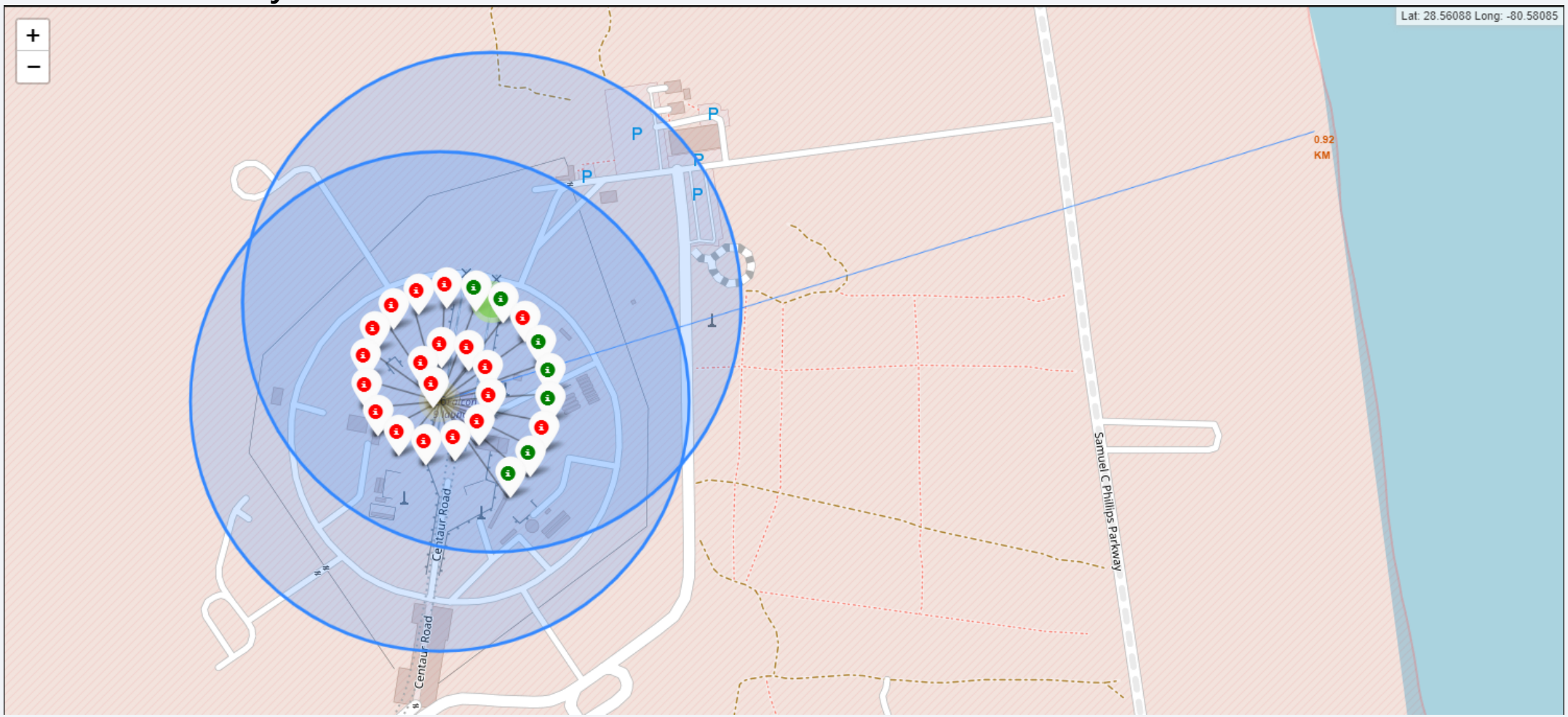
Build an Interactive Map with Folium

- To get insight into the launch sites of the rockets, we create a map using Folium.
- This map marks the launch site locations to see their geographical location.
- Furthermore, we map successful and failed launches per launch site in a marker cluster. This allows us to easily see the success rate per launch site.
- Lastly, we mark the distance to nearby geographical features (coastlines, train lines, highways), to find if launch sites are a close to such locations, or if they need to keep a certain distance away from such locations.

- The source code for the Folium map can be found here.

Build an Interactive Map with Folium

- For instance, we see that launch site CCAFS LCS-40 has mostly unsuccessful launches and is 0.92 km away from a coastline.



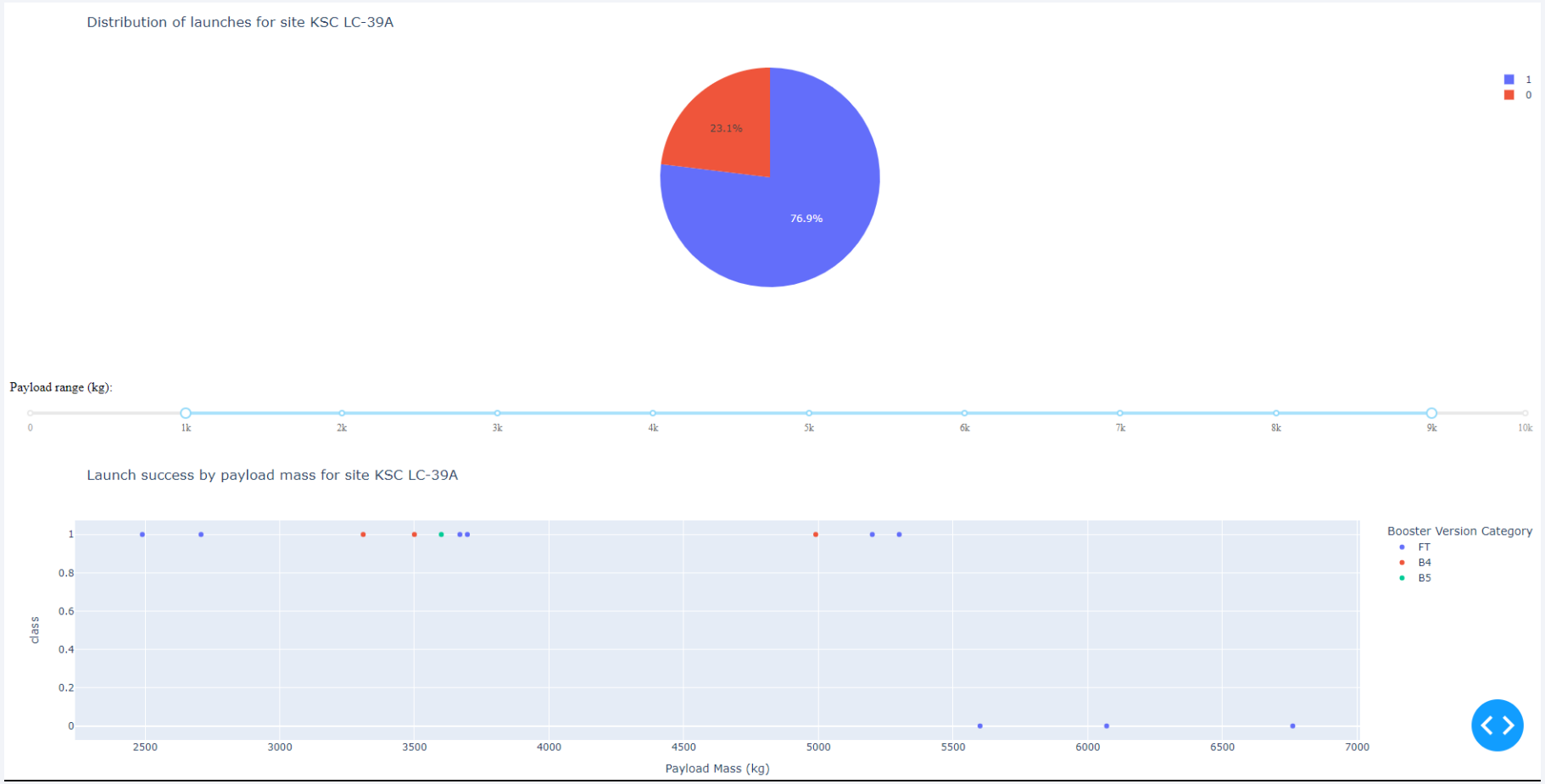
Build a Dashboard with Plotly Dash

- A dashboard allows us to interact with the data in a way that static visualizations do not allow for.
- Hence, we developed a dashboard that allows us to get insight into the success rate of launches across all sites, or only a specific site.
- Furthermore, we can investigate how the payload mass affects the success rate of a launch, where we also distinguish between booster versions.
- This is valuable insight for the development of a classification algorithm.

- The source code for the Plotly Dash dashboard can be found here.

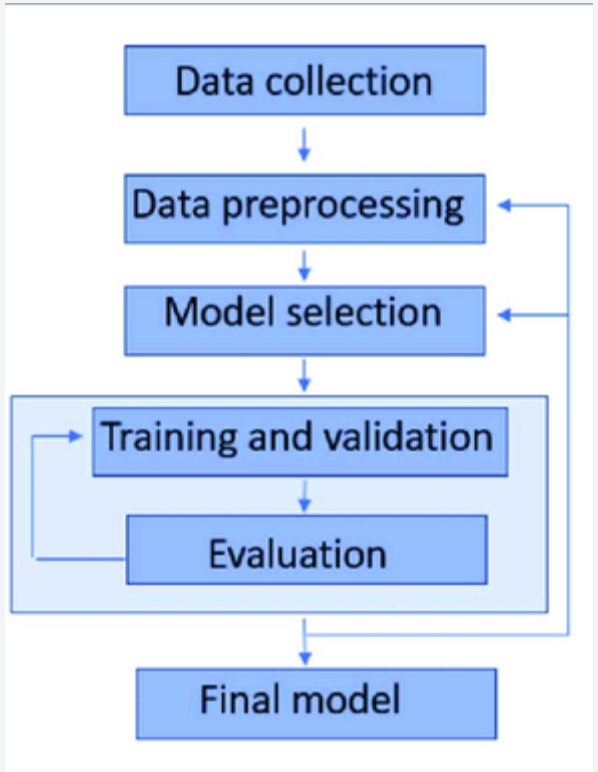
Build a Dashboard with Plotly Dash

This is screenshot of the dashboard, where we see the distribution of successful (1) and unsuccessful (0) launches in a pie chart. Furthermore, we see the relationship between payload mass and successful landing, distinguishing also by booster version.



Predictive Analysis (Classification)

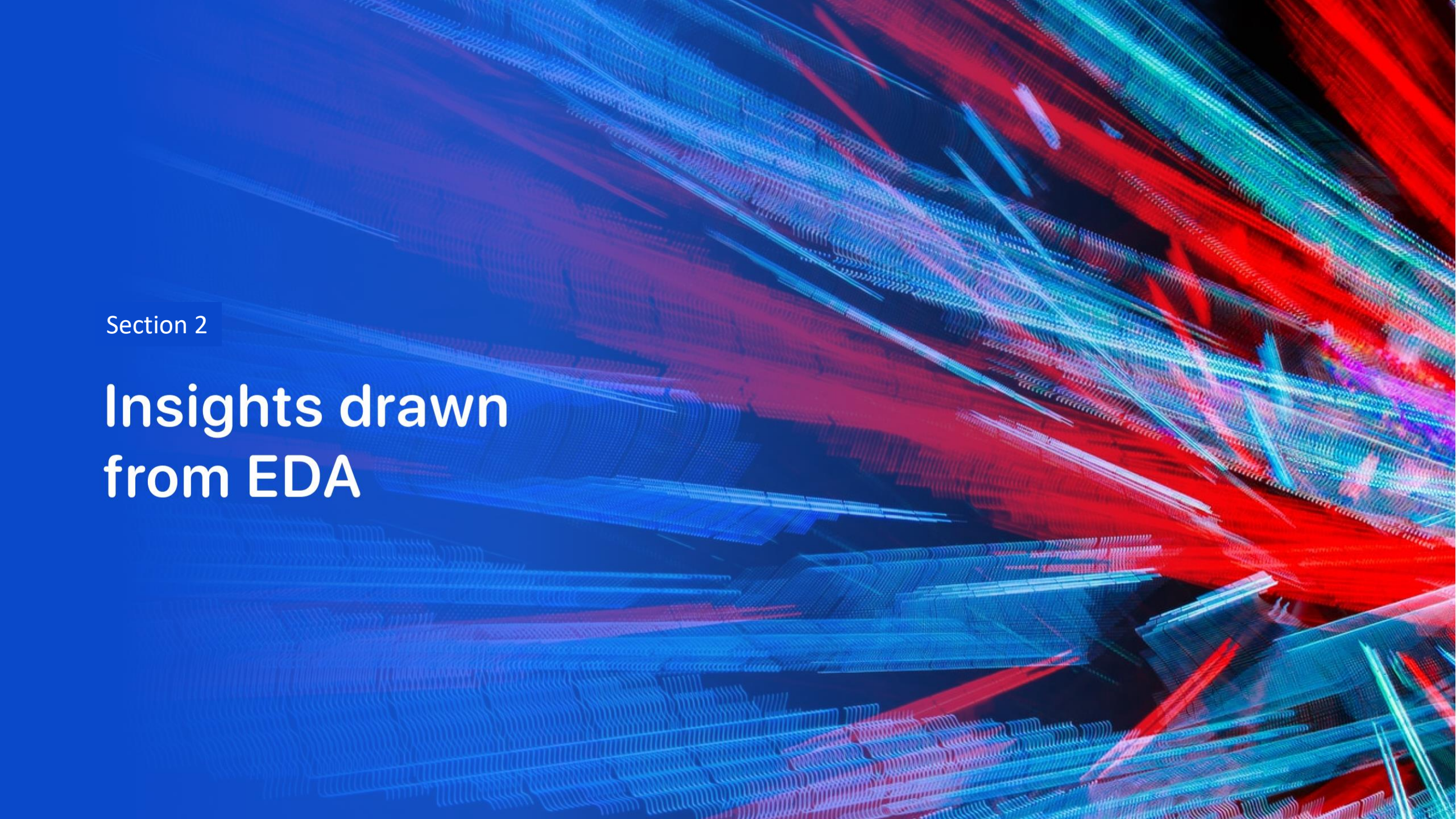
- Now that the data has been collected and we have gained initial insights, we want to use the historical launch data to predict if a future launch will be successful.
- This is a classification problem, and we built, optimize and evaluate various algorithms to find which works best for this problem.



- The source code for the model training and evaluation can be found here.

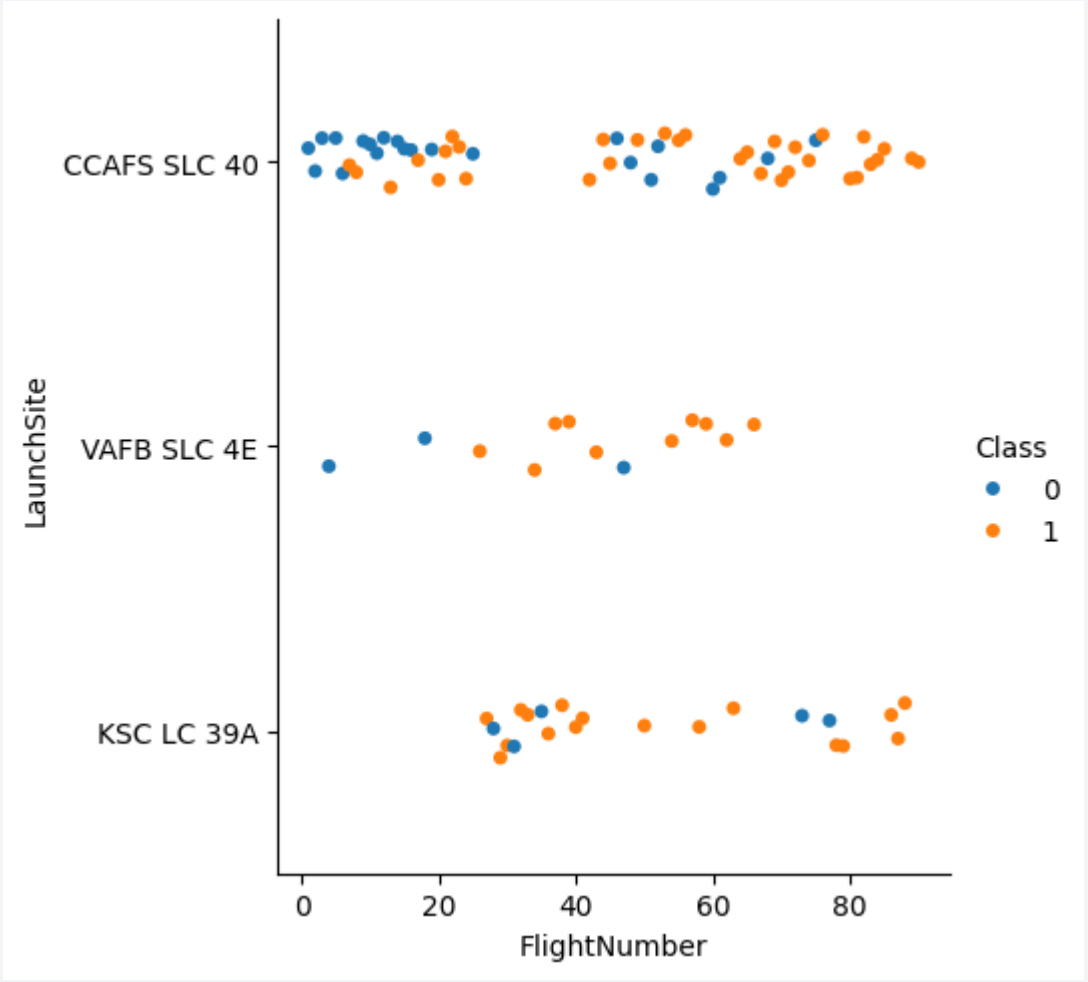
Results

- We will now present the results of the exploratory data analysis.
- Furthermore, we will show how we can interact with our data, using a dashboard that we developed.
- Finally, we show the results of our predictive analysis using classification algorithms.


Section 2
Insights drawn from EDA

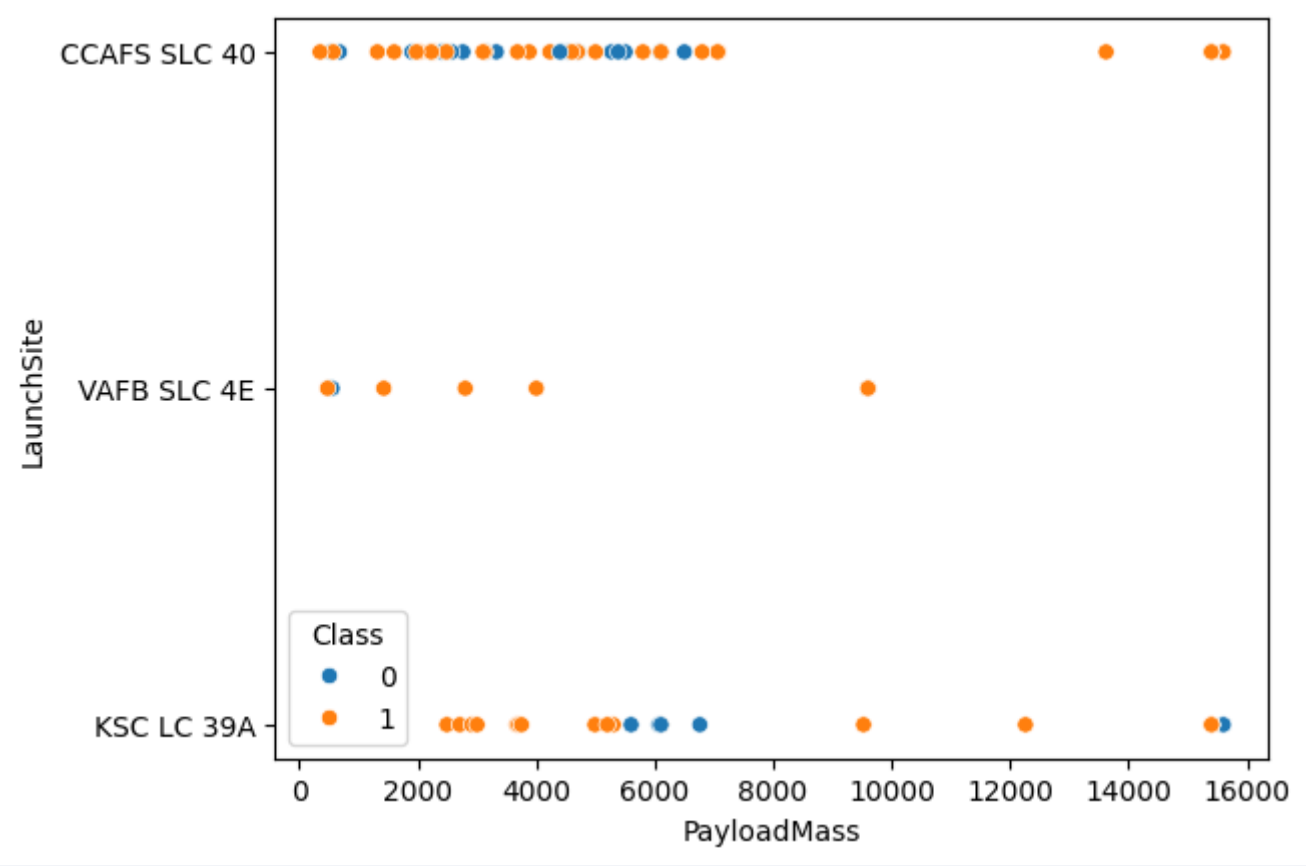
Flight Number vs. Launch Site

- We see that CCAFS SLC-40 has the largest number of launches. Additionally, there is no clear majority of successful or unsuccessful launches at this site.
- In contrast, VAFB SLC 4E and KSC LC 39A seem to have a majority of successful launches.



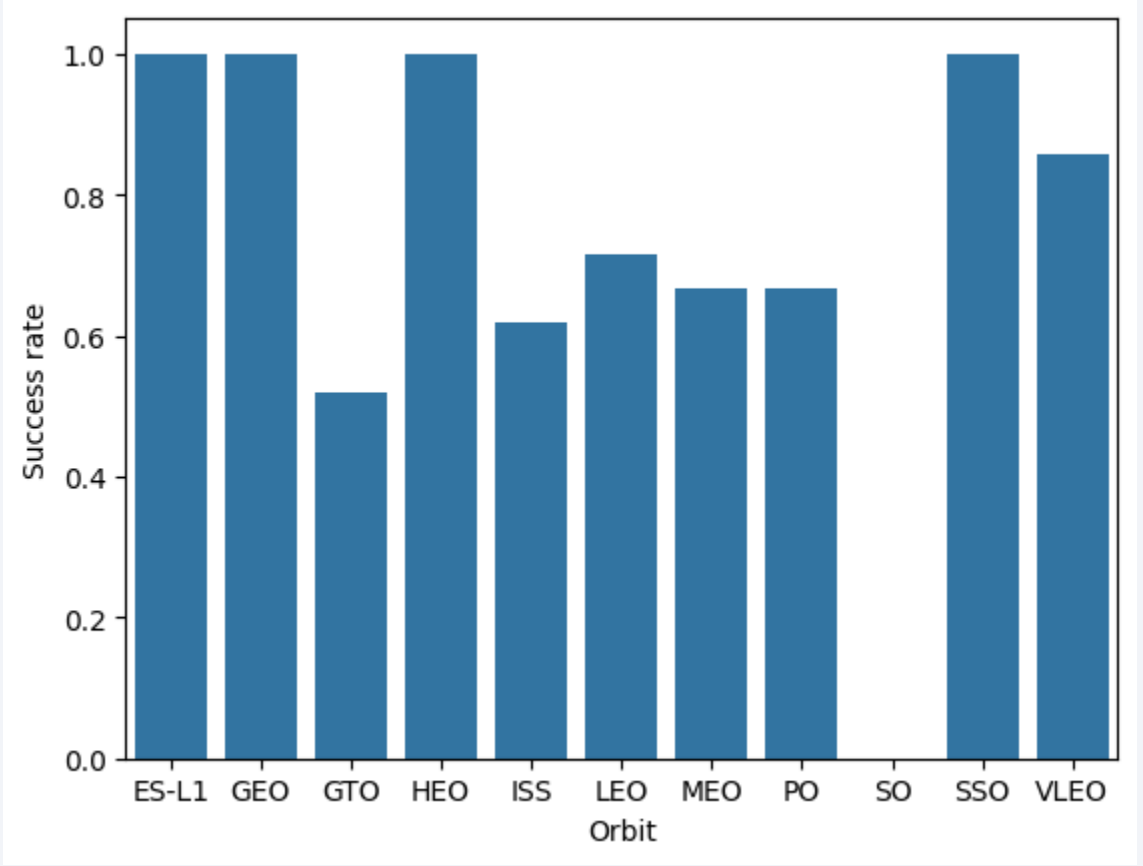
Payload vs. Launch Site

- In this plot, we find that CCAFS SLC 40 and KSC LC 39A are the sites where the largest payloads are launched from.
- Additionally, we see that large payloads typically have a high success rate.
- In contrast, low payloads have a mixed success rate.



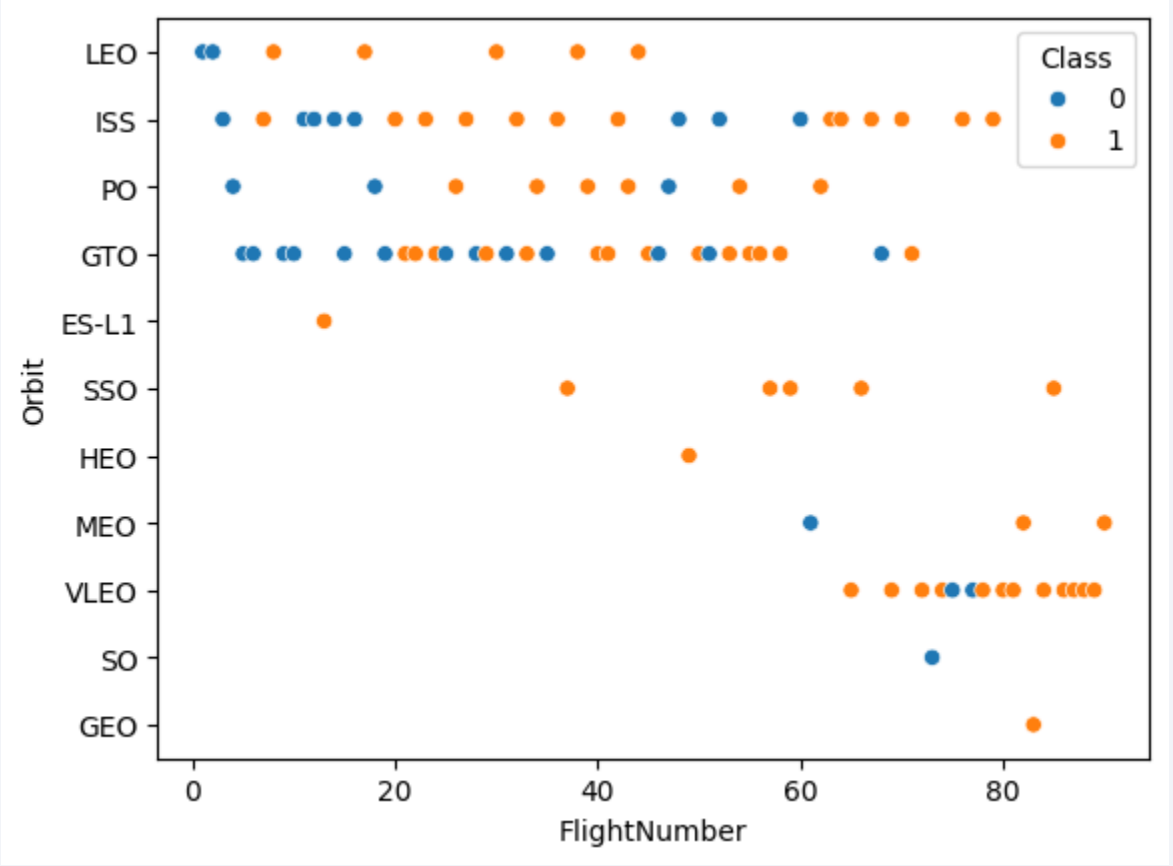
Success Rate vs. Orbit Type

- Here we see the success rate per orbit type.
- ES-L1, GEO, HEO and SSO are among the highest success rates.
- In contrast, GTO and ISS have some of the lowest success rates.



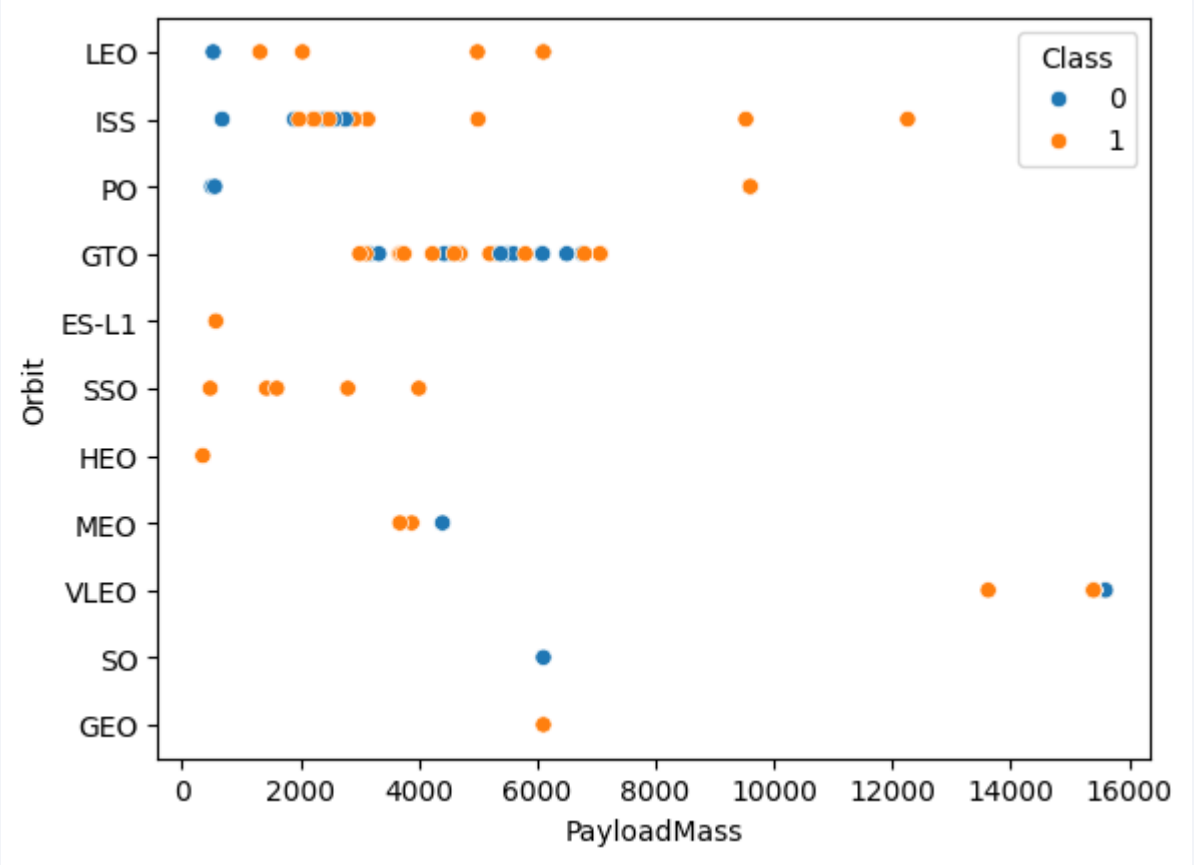
Flight Number vs. Orbit Type

- Recent flights have mostly transitioned to VLEO and ISS orbits, as indicated by the flight number vs orbit plot.
- Additionally, these recent flights have a higher success rate, compared to the earlier flights mostly to the LEO, ISS, PO and GTO orbits.



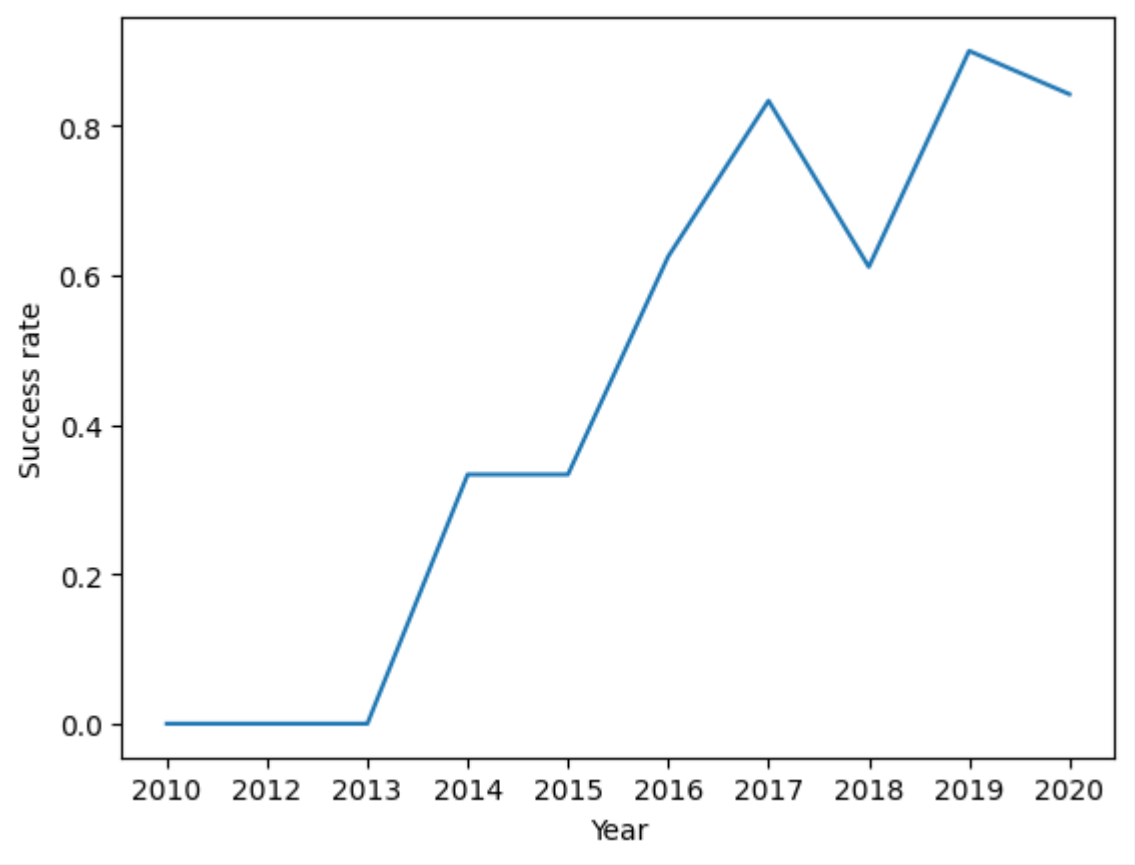
Payload vs. Orbit Type

- With heavy payloads the success rates are larger for Polar, LEO and ISS orbits
- However, for GTO there is a mixture of success rates, and no clear preference can be identified.



Launch Success Yearly Trend

- We observe that in recent years, the success rate of landing has increased.



All Launch Site Names

- %sql SELECT DISTINCT(`Launch_Site`) FROM SPACEXTABLE

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

- This query shows us the unique names of all launch sites.

Launch Site Names Begin with 'CCA'

- %sql SELECT * FROM SPACEXTABLE WHERE `Launch_Site` LIKE 'CCA%' LIMIT 5

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

- This shows us 5 records where the launch site starts with ‘CCA’.

Total Payload Mass

- %sql SELECT SUM(`PAYLOAD_MASS__KG_`) FROM SPACEXTABLE WHERE CUSTOMER = 'NASA (CRS)'

SUM(PAYLOAD_MASS__KG_)
45596

- This allows us to calculate the total payload that NASA (CRS) has carried on their launches.

Average Payload Mass by F9 v1.1

- %sql SELECT AVG(`PAYLOAD_MASS__KG_`) FROM SPACEXTABLE WHERE `Booster_Version` LIKE 'F9 v1.1%'

AVG(`PAYLOAD_MASS__KG_`)
2534.6666666666665

- This allows us to find the average payload mass carried by a specific booster version, in this case the F9 v1.1.

First Successful Ground Landing Date

- %sql SELECT MIN(DATE) FROM SPACEXTABLE WHERE `Landing_Outcome` = 'Success (ground pad)'

MIN(DATE)
2015-12-22

- This allows us to find the first date that a successful landing happened on a ground pad, which happens to be in December 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql SELECT `Booster_Version` FROM SPACEXTABLE WHERE `Landing_Outcome` = 'Success (drone ship)' AND (`PAYLOAD_MASS__KG_` > 4000 AND `PAYLOAD_MASS__KG_` < 6000)

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- Here we restrict the landings to successful landings on a drone ship, and those with payloads between 4000 kg and 6000 kg. The booster version is F9 FT.

Total Number of Successful and Failure Mission Outcomes

- %sql SELECT `Mission_Outcome`, COUNT(`Mission_Outcome`) FROM SPACEXTABLE GROUP BY `Mission_Outcome`

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

- We group the results by mission outcome and find that most missions are successful. Only 1 launch failed in flight.

Boosters Carried Maximum Payload

- %sql SELECT `Booster_Version` FROM SPACEXTABLE WHERE `PAYLOAD_MASS__KG_` = (SELECT MAX(`PAYLOAD_MASS__KG_`) FROM SPACEXTABLE)

- Here we find the booster versions that carried the maximum payloads using a subquery. This is always booster version F9 B5.

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

- %%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'

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

We find that in 2015 only 2 failures occurred, both on an attempted landing on a drone ship.

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

- %%sql SELECT `Landing_Outcome`, COUNT(`Landing_Outcome`)

FROM SPACEXTABLE

WHERE `Date` > '2010-06-04' AND `Date` < '2017-03-20'

GROUP BY `Landing_Outcome`

ORDER BY COUNT(`Landing_Outcome`) DESC

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

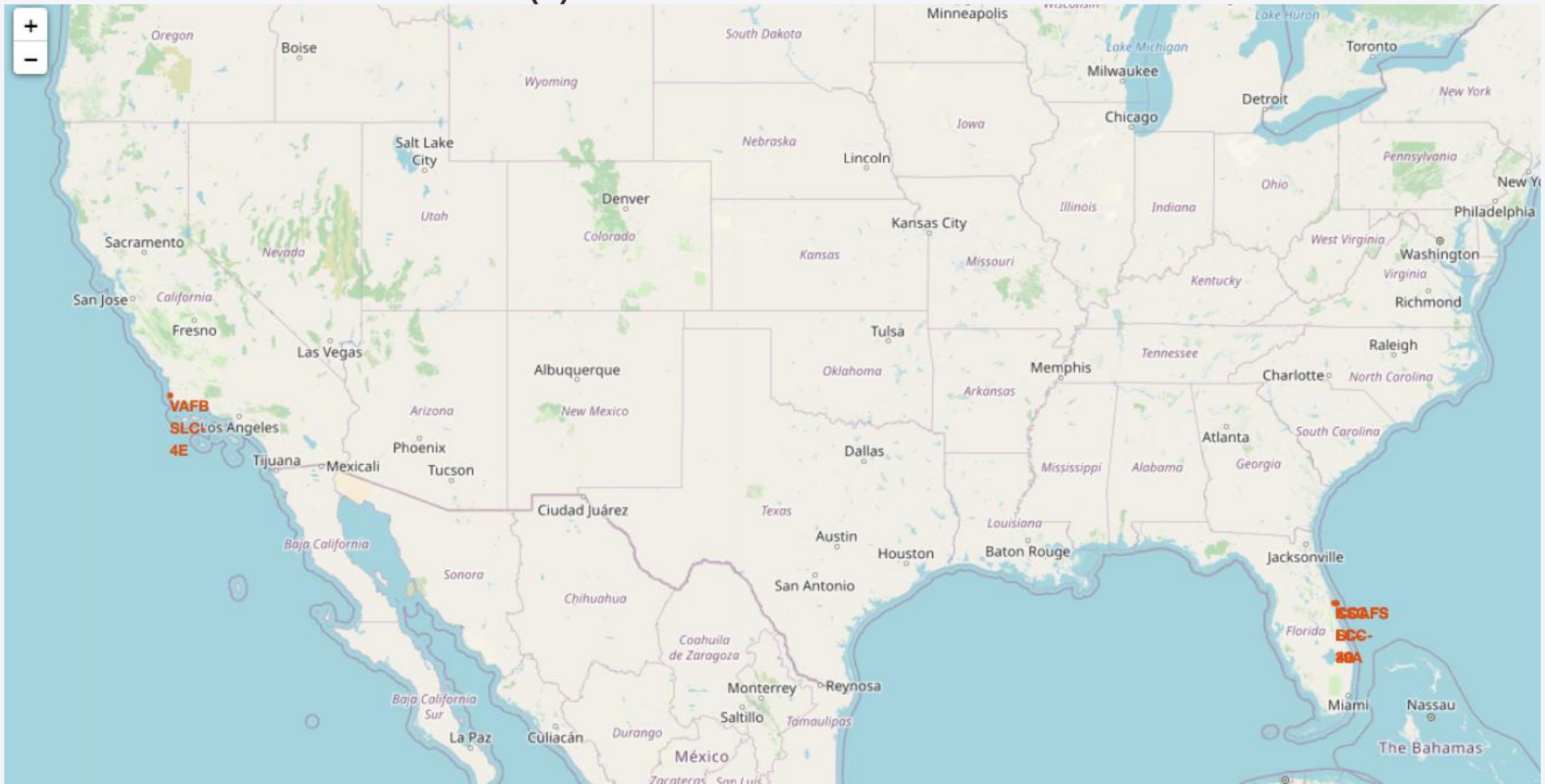
We order the results in descending order to see what the most common occurrences of landing outcomes are. This is no attempt, followed by a successful landing on a drone ship.

Section 3

Launch Sites Proximities Analysis

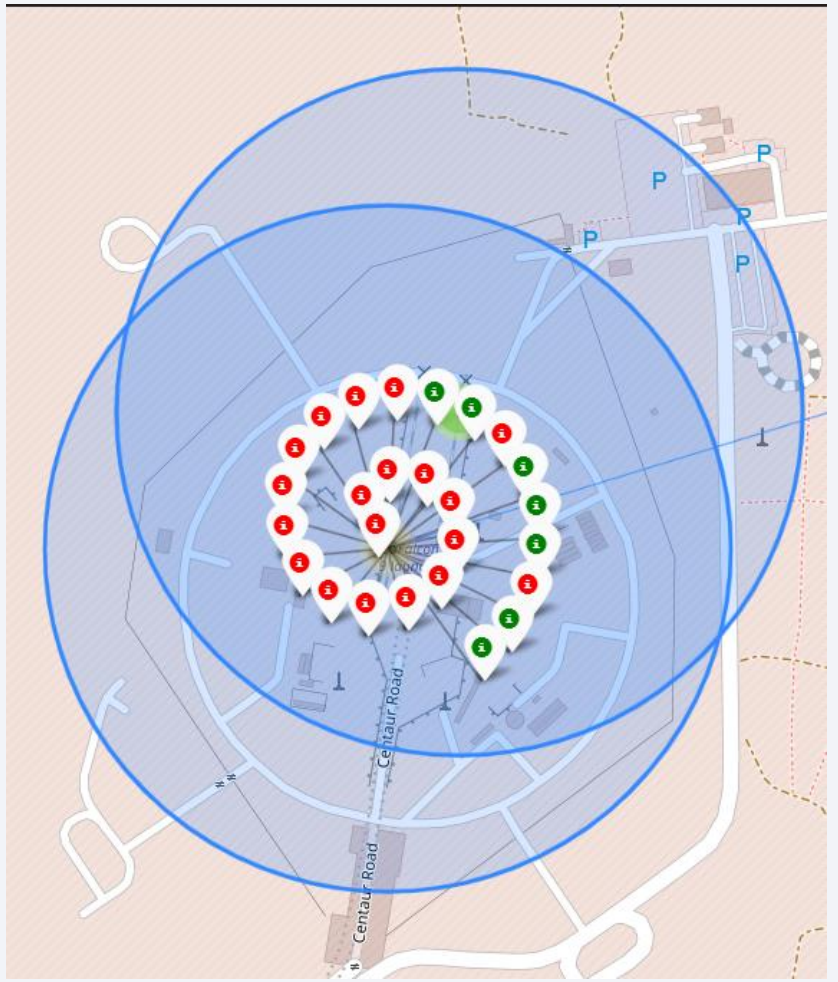
Launch site locations across the US

- Here we map the locations of the launch sites across the US.
- We observe that all launch sites are close to the coastline. VAFB SLC 4E is close to the west coast, while KSC LC-39A and CCAFS (S)LC-40 are close to the east coast.



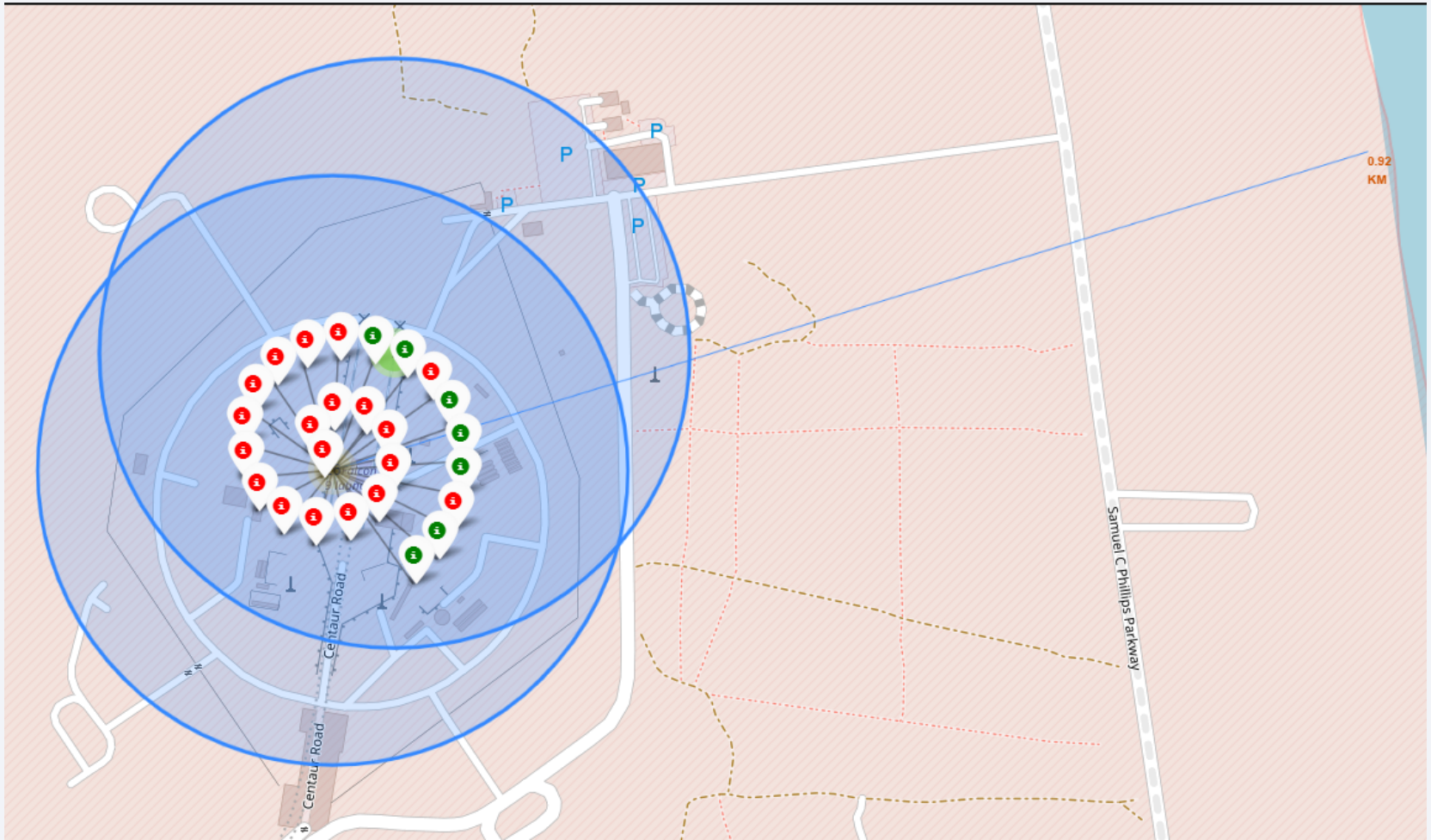
(Un)successful launches per launch site

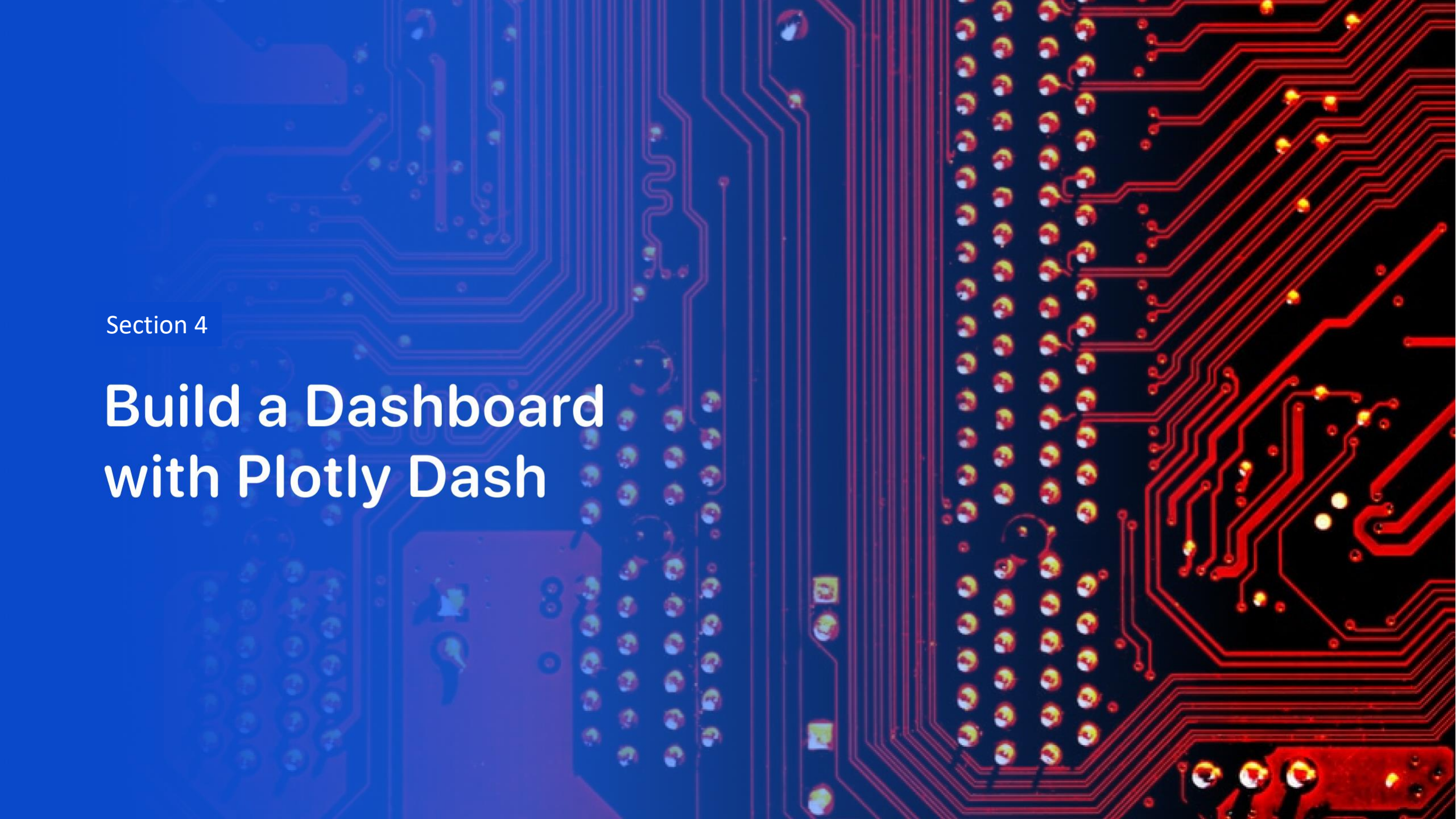
- We add markers and icons to each launch site based on if the launch was successful or not. This allows us to easily identify the success rate per launch site. For example, launch site CCAFS LC-40 has mostly unsuccessful launches.



Launch site proximities

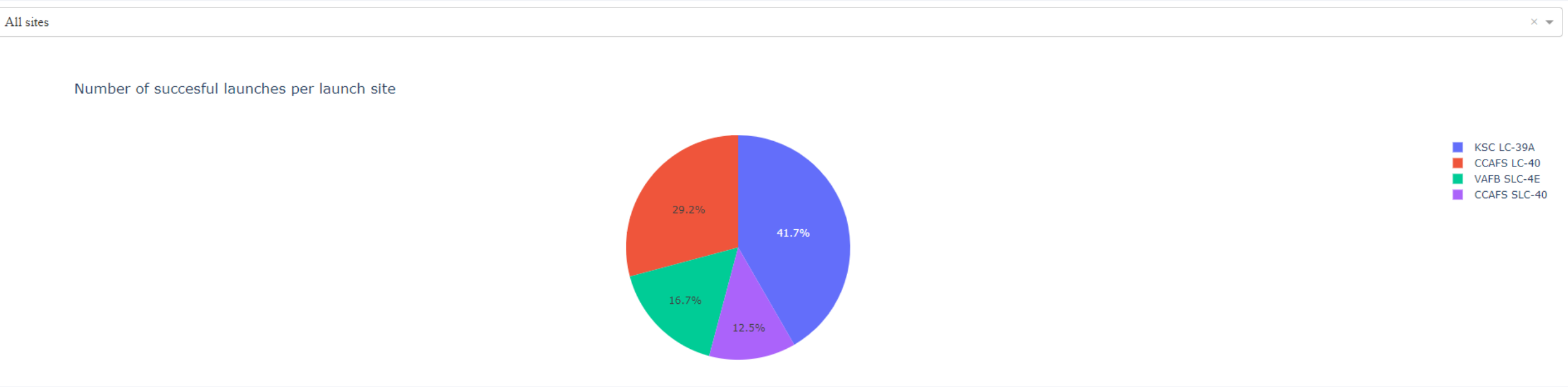
- To investigate the launch sites further, we identify the distance to nearby geographical features.
- For instance, the launch site CCAFS LC-40 is only 0.92 km away from the coastline, as can be seen in the map.




Section 4
Build a Dashboard with Plotly Dash

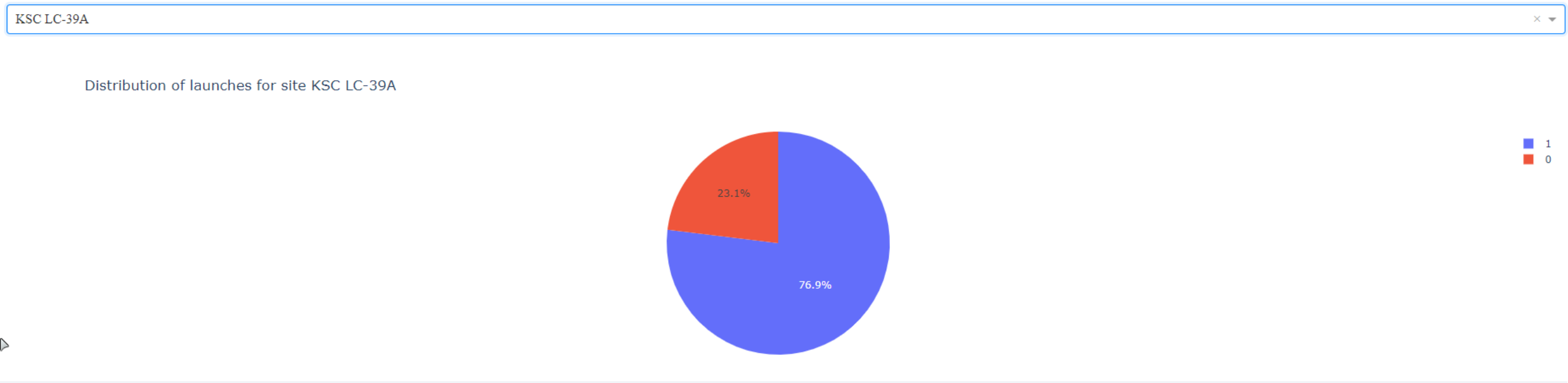
Successful launches per launch site

- This pie chart shows the distribution of successful launches across all launch sites. We observe that the largest share of successes is on site KSC LC-39A, and the least amount at CCAFS SLC-40.



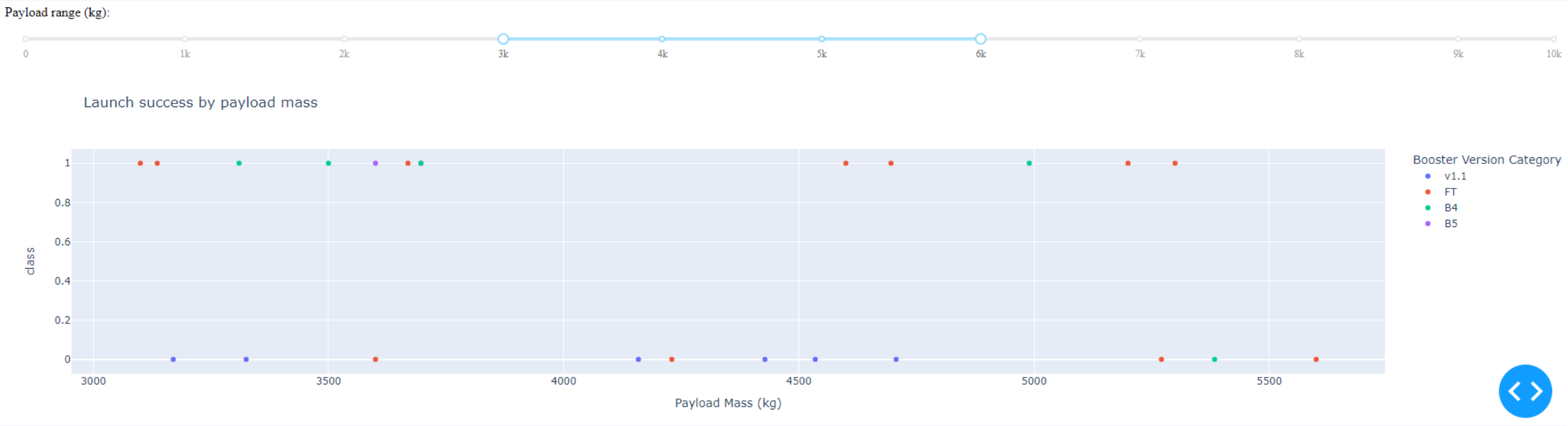
Success rate for KSC LC-39A

- Inspecting the launch site with the most successes (KSC LC-39A) further, we observe that nearly 77% of all launches are successful at this site.



Launch success by payload mass across all sites

- Here we restrict the payload mass to be between 3000 kg and 6000 kg. We investigate the success rate across all sites, distinguishing by booster version.
- We observe that there are approximately equal number of successful and unsuccessful launches in this payload range across all sites.
- The most successful booster in the payload range is the FT booster.

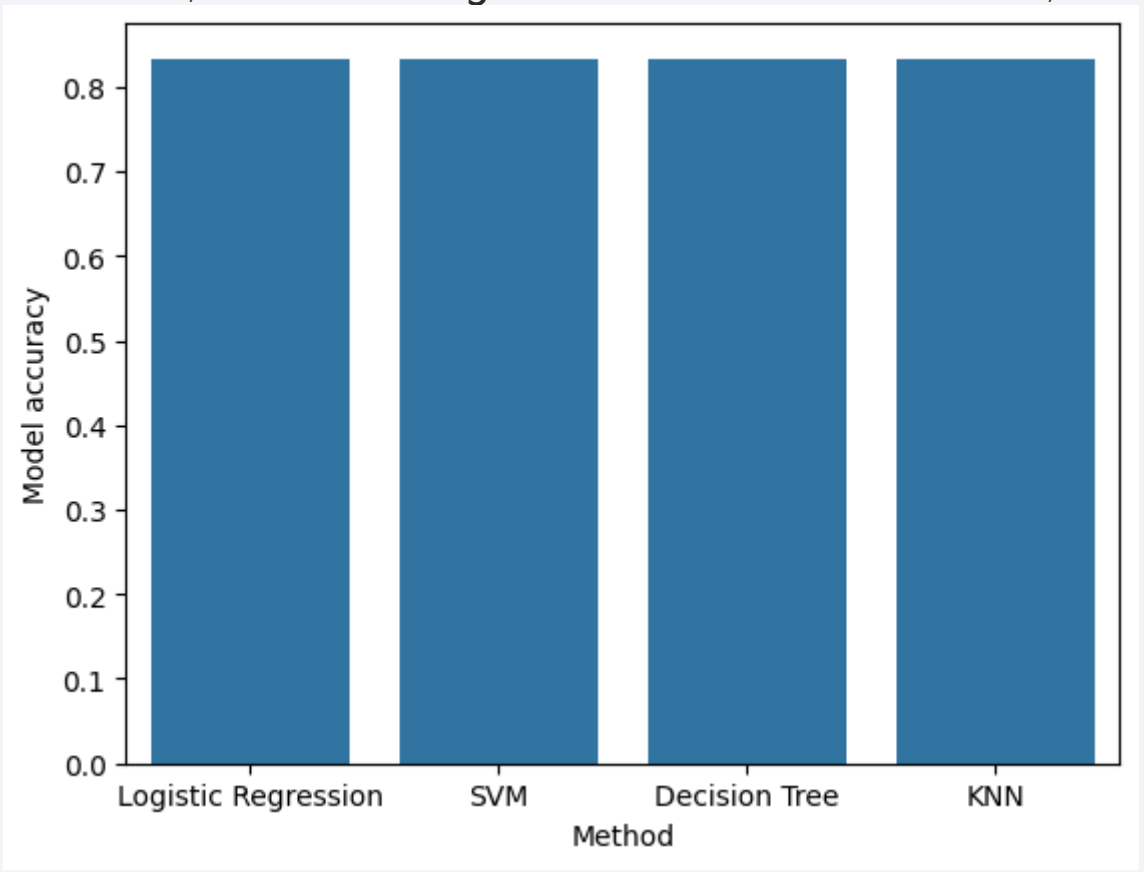


Section 5

Predictive Analysis (Classification)

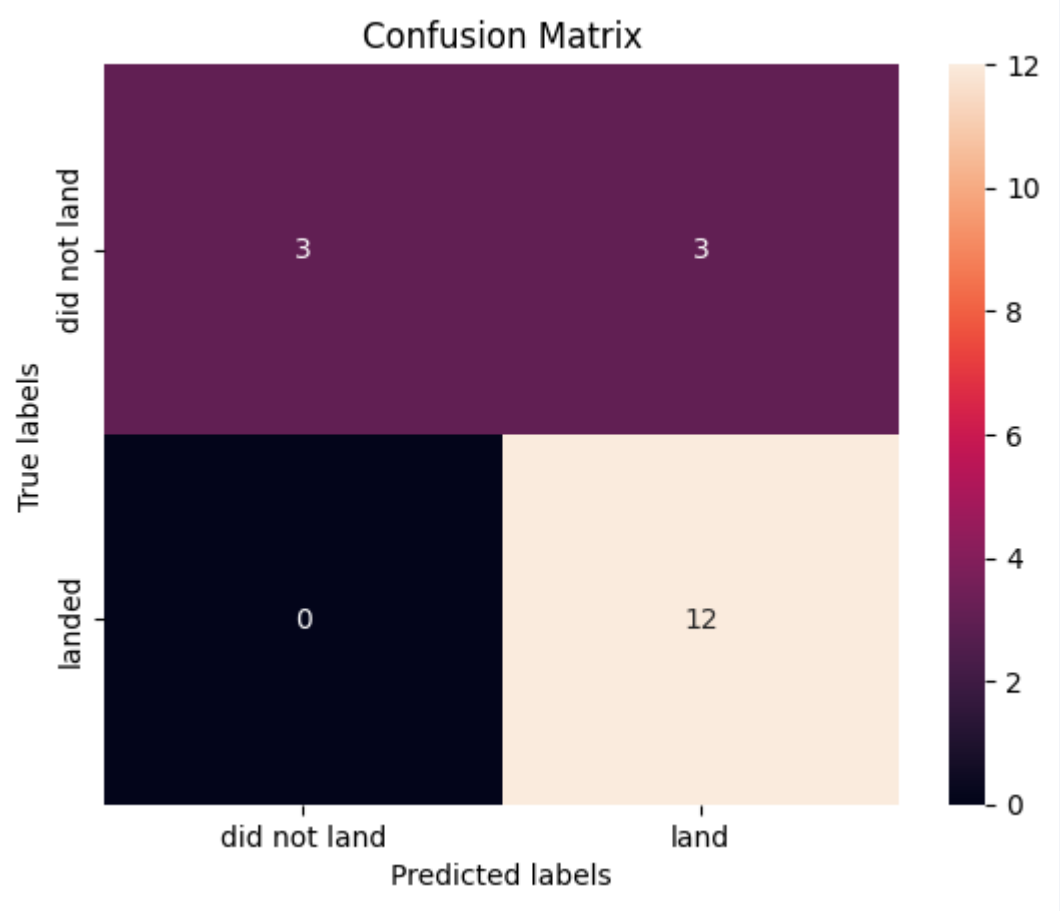
Classification Accuracy

- After building, training and evaluation all models, we compare the model performance on test data.
- We observe that all models achieve the same model accuracy.
- Since the model accuracy is identical, we choose to go with the decision tree model, since this is easiest to be interpreted by a human.



Confusion Matrix

- The confusion matrix of the best performing model (decision tree) is shown below.
- We observe that the model correctly predicts all true labels that are equal to ‘landed’, while it struggles with cases where the true label is equal to ‘did not land’. In other words, it has a high False Positive rate.



Conclusions

- We aimed to predict the successful landing of a booster, to predict if it can be reused for a future rocket launch. This significantly reduces the cost of a rocket launch.
- We gathered data using web scraping and REST APIs on historical launch data.
- After processing the gathered data, we observed trends in this data that gave us initial insights into the data.
- Finally, we built classification models that can predict if a landing will be successful based on historical launches.
- All models performed identical. We choose to go with a decision tree model since this is easiest to be interpreted by humans.

Appendix

- All code, figures and SQL queries can be found at the GitHub repo linked below.

- https://github.com/StijnvanderLippe/IBM-Data-Science-Professional-Certificate/tree/main/Module%2010%20Applied%20Data%20Science%20Capstone

Thank you!