IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Dan K.>
<30 JUL 24>

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - This project used web scraping and APIs to gather data on Spacex's space program to evaluate its overall success and characteristics.
  - Data were wrangled and cleaned using pandas and python.
  - EDA was performed on the cleaned data, which allowed the development of predictive models that aimed to predict the success of a Spacex mission's landing based on characteristics within out dataset.
  - Supervised machine learning methods (Logistic Regression, SVM, Decision Trees, and K Nearest Neighbors) were the predictive models used for this project.
- Summary of all results
  - Spacex is able to recover 50% of their boosters that land on drone ships according to this data.
  - The decision tree model was able to correctly classify more outcomes than the other models (logistic regression, SVM, and K Nearest Neighbors).

Introduction

- This project aims to help Spacex predict successful landings based on salient characteristics in their data like launch site, orbit, booster version and payload.
- Essentially, we want to increase the number of boosters that Spacex can recover, thus helping them to keep their overall costs down by allowing them to recover more boosters.

Section 1

Methodology

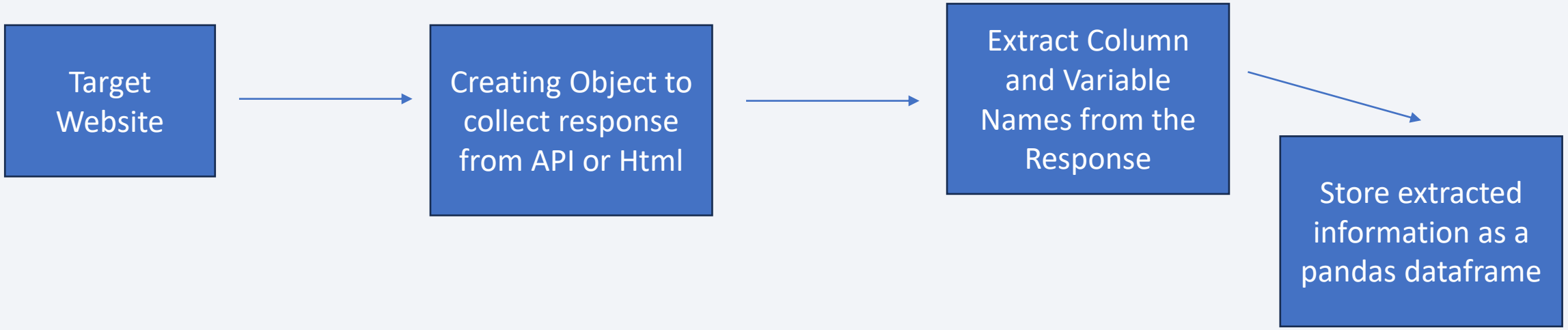
Methodology

Executive Summary

- Data collection methodology:
  - Data were collected using APIs and webscraping from public facing websites.
- Perform data wrangling
  - Missing values were filled in, or dropped in the dataset. Landing outcomes were created, a “class” column was created to help facilitate modeling.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Models were tuned using hyperparameters and evaluated primarily on their accuracy of classification of our target variable (class).

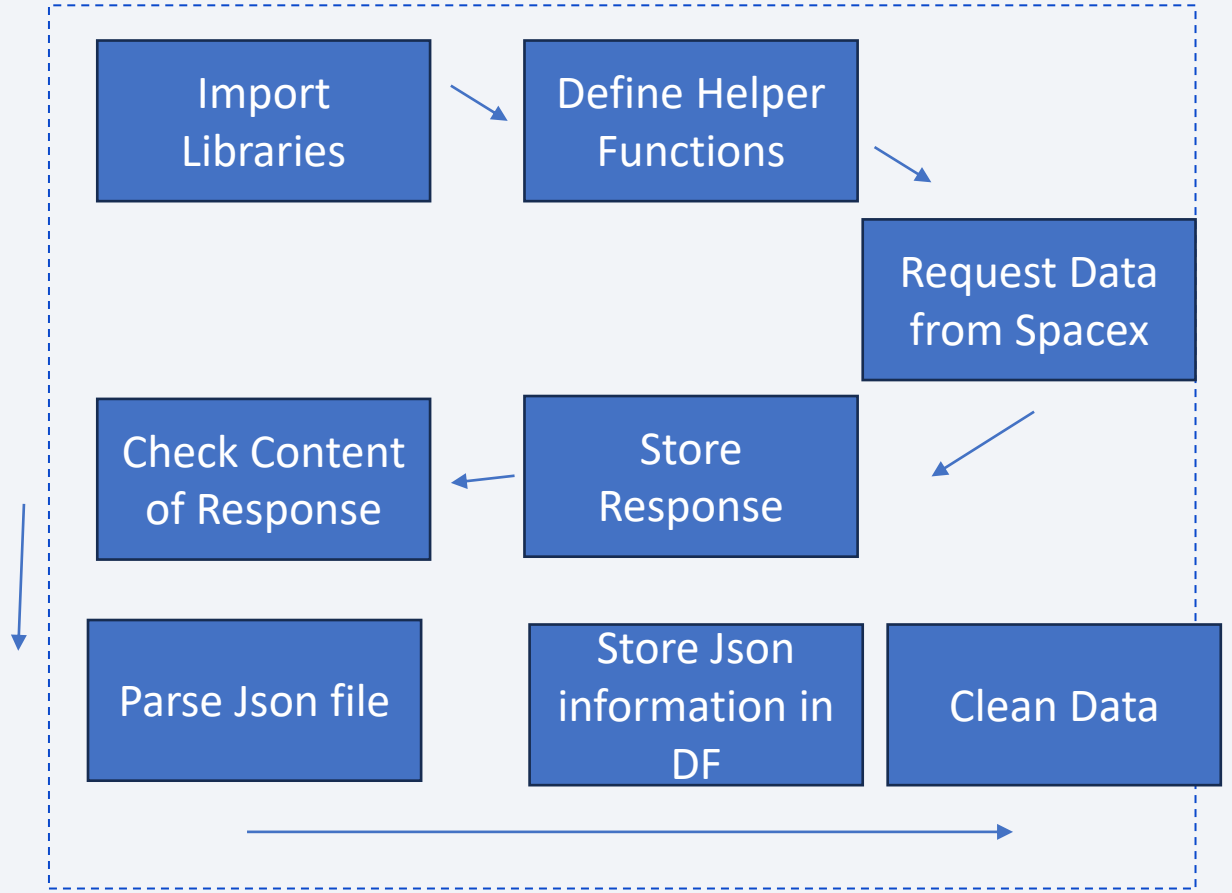
Data Collection

- Data were collected using webscaping techniques in python, notably the beautifulsoup library and by using Spacex APIs.
- You need to present your data collection process use key phrases and flowcharts



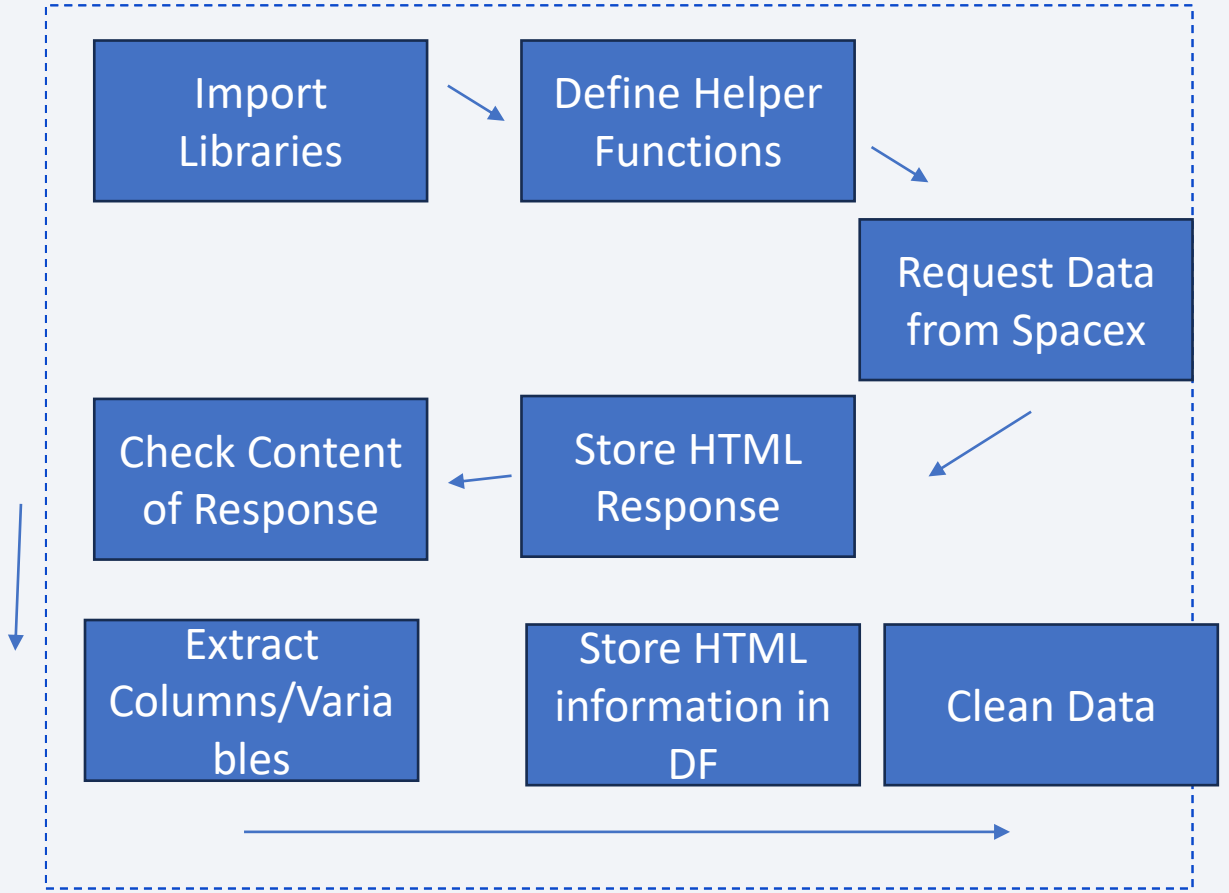
Data Collection – SpaceX API

- Git Hub Link for Code:
- https://github.com/dkeen1/DS_Cert/blob/main/jupyter-labs-spacex-data-collection-api%20-%20dan%20-%20part%201.ipynb



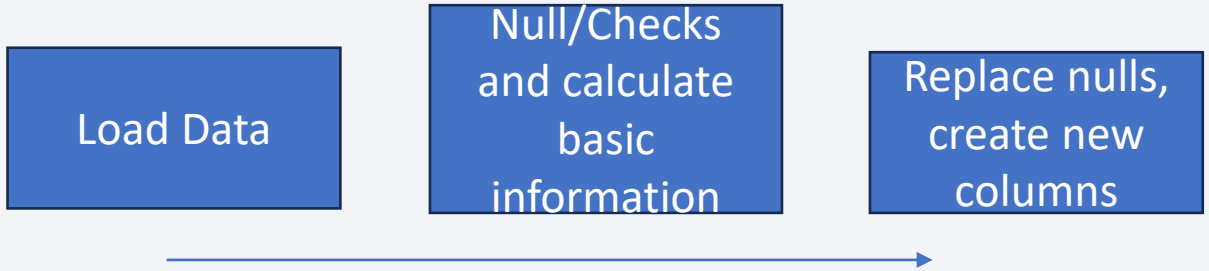
Data Collection - Scraping

- Web Scraping Git Hub:
- https://github.com/dkeen1/DS_Cert/blob/main/jupyter-labs-webscraping%20-%20dan%20-%20part2.ipynb



Data Wrangling

- Data were processed using standard methods, primarily using numpy and pandas in python.
- Data were loaded into a dataframe and checked for null values, data types were checked, basic calculations were performed to ensure outputs made sense (checking for outliers), created a set of outcomes indicating successful landing outcomes (to facilitate later analysis)



- Data Wrangling Git Hub Link - https://github.com/dkeen1/DS_Cert/blob/main/labs-jupyter-spacex-Data%20wrangling%20-%20Dan.ipynb

EDA with Data Visualization

- Various charts were used for this analysis:
  - Scatter plots – to see the correlation between different variables and their potential impact on successful landings (to help identify variables to use in our predictive models)
  - Bar Charts – to help understand success rates with our categorical variables, like orbit type or launch site. Again, to identify variables for our models.
  - Line Chart – to help display success over time, to help identify overall characteristics of the data and Spacex's program. Variable's predictive ability could change over time based on overall program trends.
- EDA With visualization Git Hub: https://github.com/dkeen1/DS_Cert/blob/main/jupyter-labs-eda-dataviz%20-%20dan.ipynb

EDA with SQL

- Performed queries looking at:
  - Distinct landing types
  - Distinct launch sites
  - Records where launch site started with CCA
  - Sum of payload mass from NASA missions
  - Average payload by the F9 V1.1 rocket
  - First successful landing date
  - Boosters that successfully landed on a drone ship
  - Total number of successful and unsuccessful mission
  - Boosters carrying spacex's max payload
  - Unsuccessful landings in 2015 on a drone ship
  - Count of landing outcomes by landing outcome type
- SQL EDA Git Hub Link - https://github.com/dkeen1/DS_Cert/blob/main/jupyter-labs-eda-sql-coursera_sqllite_dan.ipynb

Build an Interactive Map with Folium

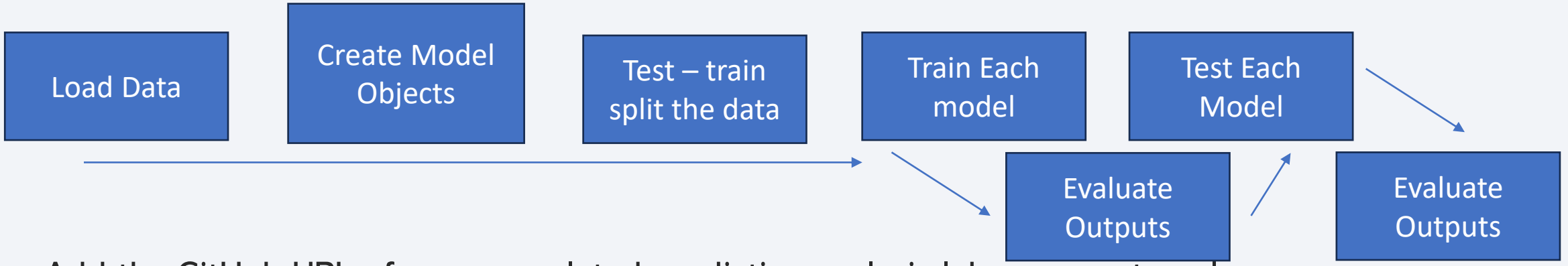
- Lines, circles, and points were added to the map.
- Objects were added to help identify where Spacex landing sites were in relation to surrounding infrastructure (points and lines) as well as where successful launches were located (circles).
- Folium Git Hub Link: https://github.com/dkeen1/DS_Cert/blob/main/Captsone_Dan-Folium.ipynb

Build a Dashboard with Plotly Dash

- Created a dashboard that showed successful launch percentages by site (in a pie chart) as well as the overall number of successful launches shared by each site. Also created a scatter plot to show the relationships between launch site, payload, and success.
- These features were added to help determine if launch site and payload had an impact on landing success and to quickly summarize the data in the dataset.
- Dash Git Hub Link: https://github.com/dkeen1/DS_Cert/blob/main/Dash_29%20JUL%20Progress.py

Predictive Analysis (Classification)

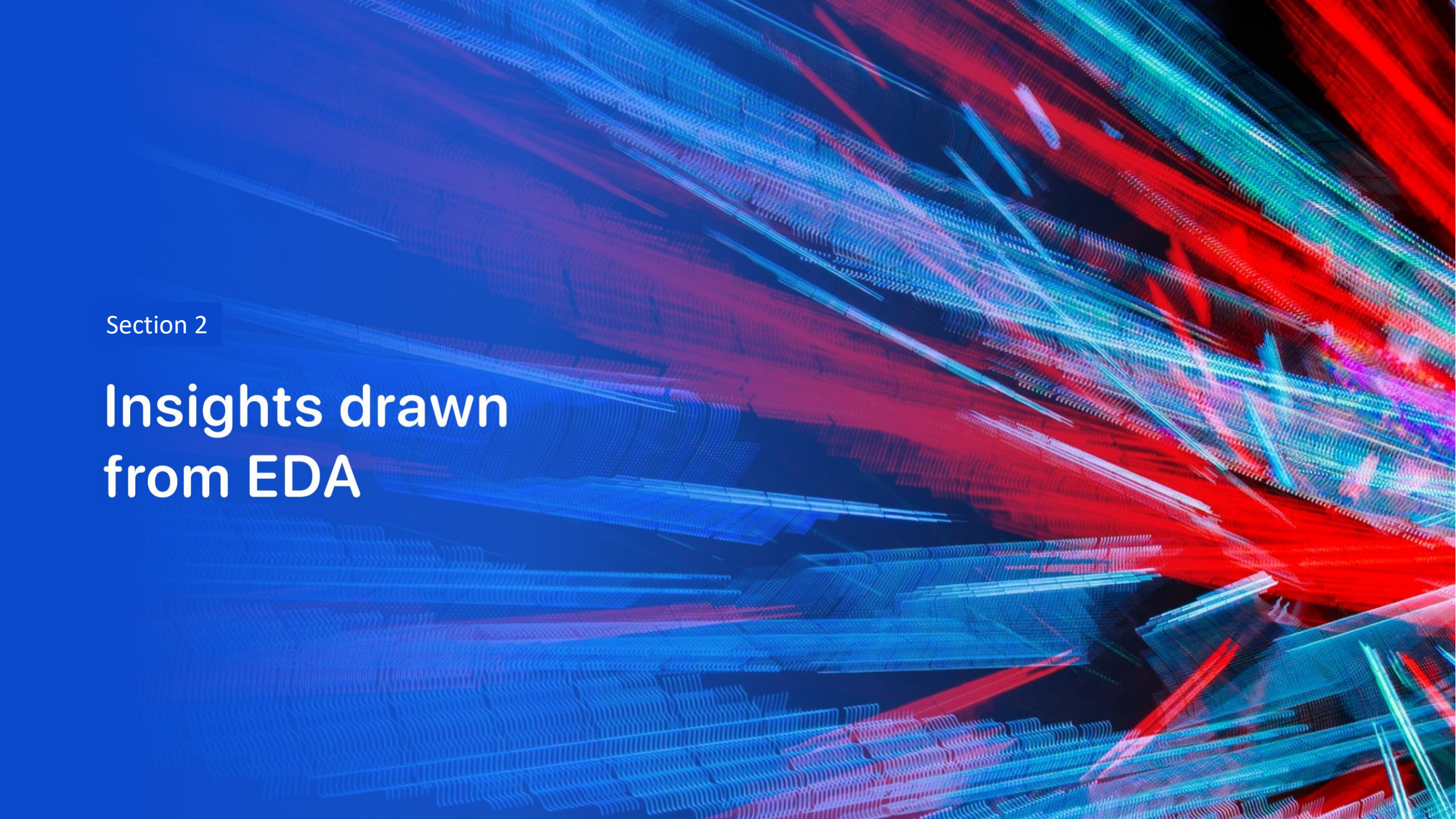
- Models were built using classification models, such as KNN, SVM, Logistic Regression and a Decision Tree. Models were trained on the most relevant variables identified during the EDA process. Models then had their hyperparameters tuned to chose the best variables for each one. The trained and tuned models were then tested on the test dataset and evaluated for their accuracy and recall.



- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

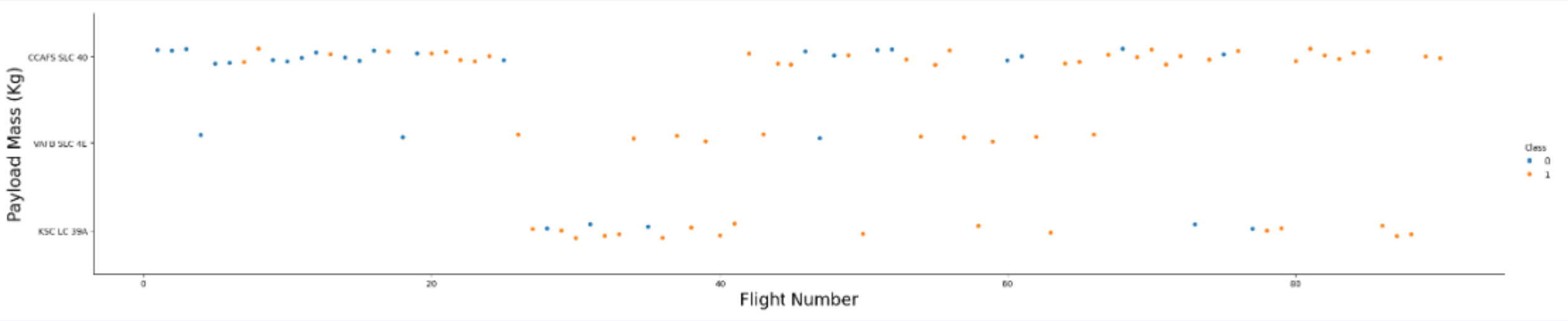
Results

- Results are presented in the following sections.


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Flight Numbers vs. Launch Sites (Hue = Success Rates)

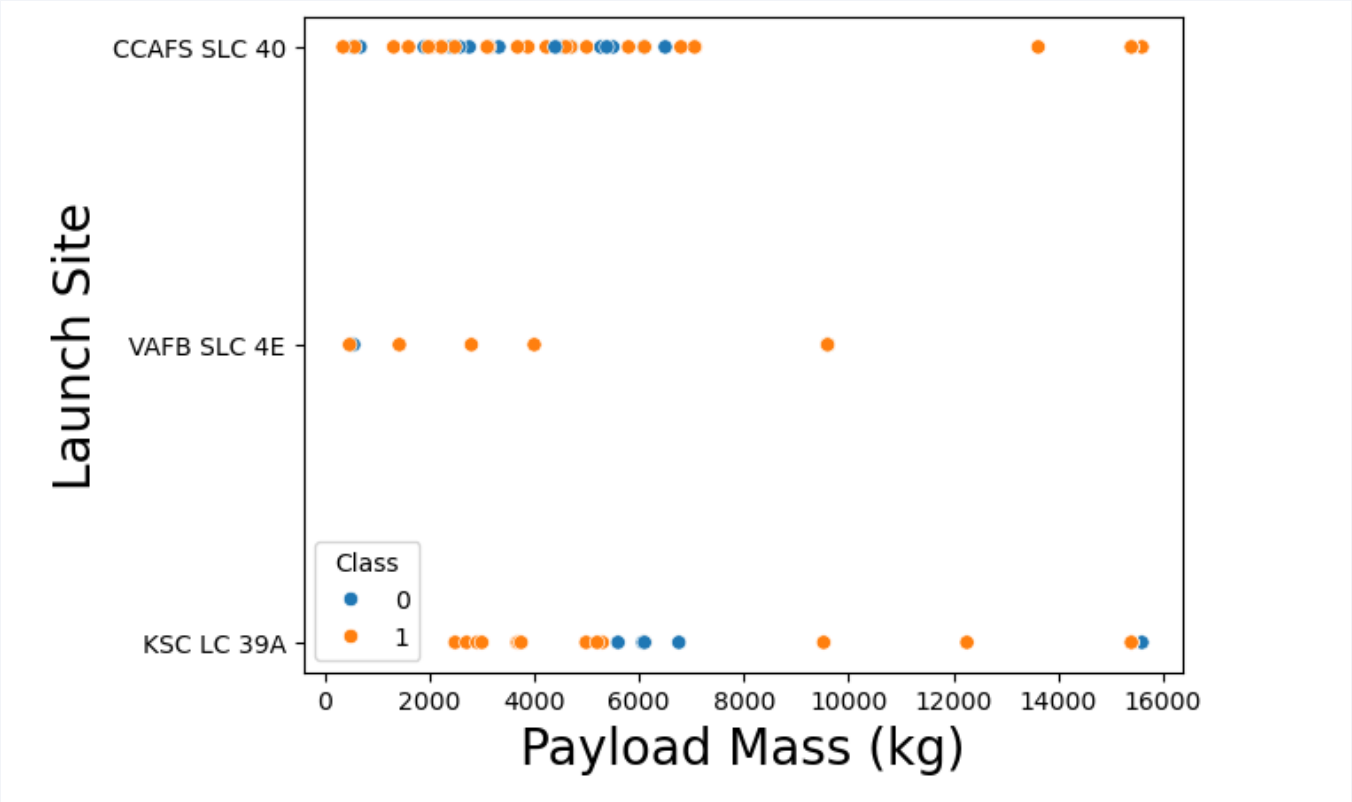


- This scatter plot shows us that Spacex started using different facilities the later they were in their program.
- The KSC LC 39A site has a fairly high success rate, as do all of spacex launches the higher the flight numbers are.

Payload vs. Launch Site

- Payload vs. Launch Site

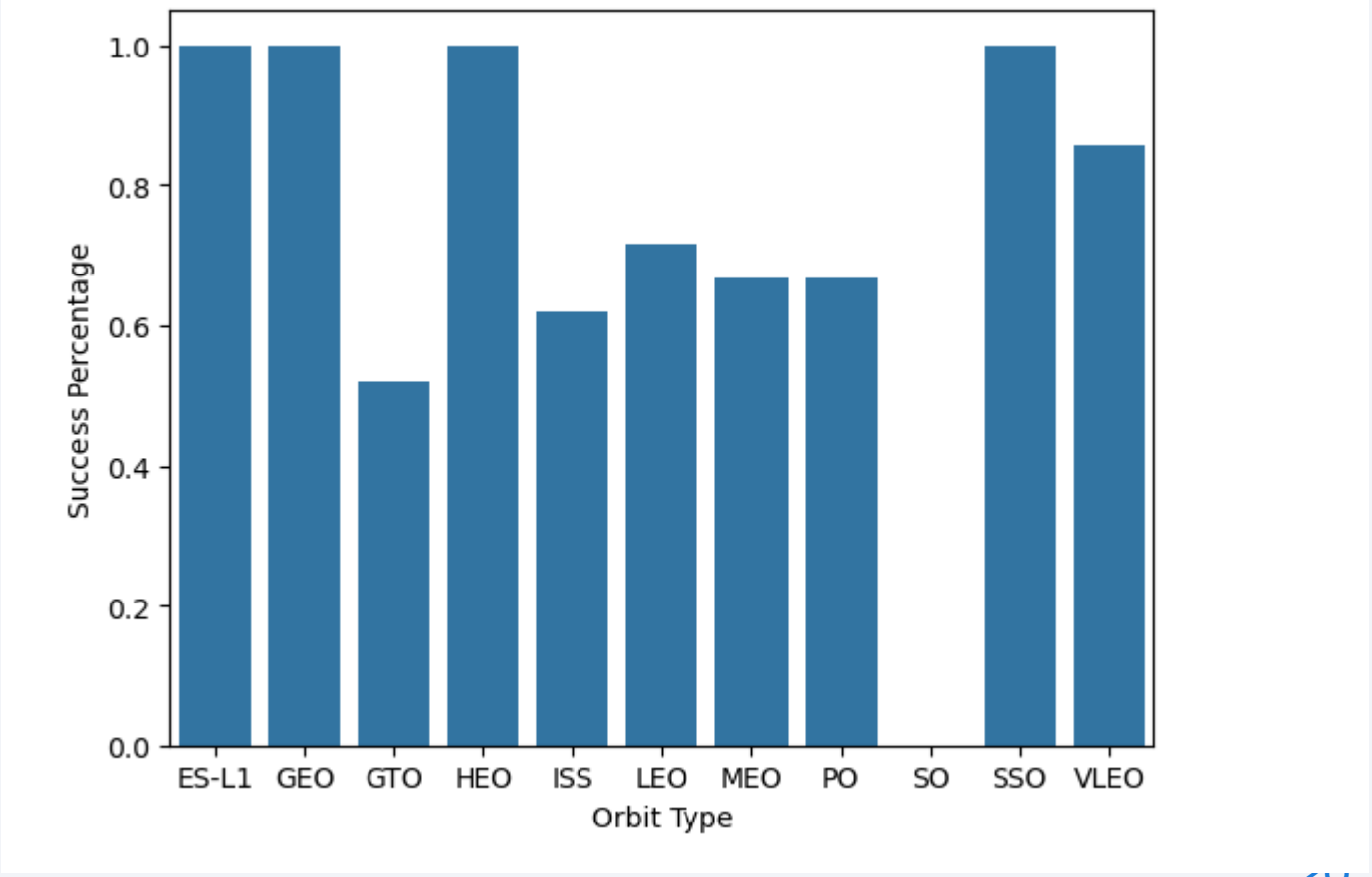
- This screenshot demonstrates heavier payloads seem to have higher percentages of success (but there are also fewer launches)
- The VAFB site also does not have any launches with payloads over 10k kg.



Success Rate vs. Orbit Type

- This chart shows that some orbits, like SSO, HEO, GEO, and ES-L1 have perfect success rates (however it is possible that these orbits have lower case numbers).
- GTO orbits have the lowest success rate.
- LEO orbits have success rates of around 70%.

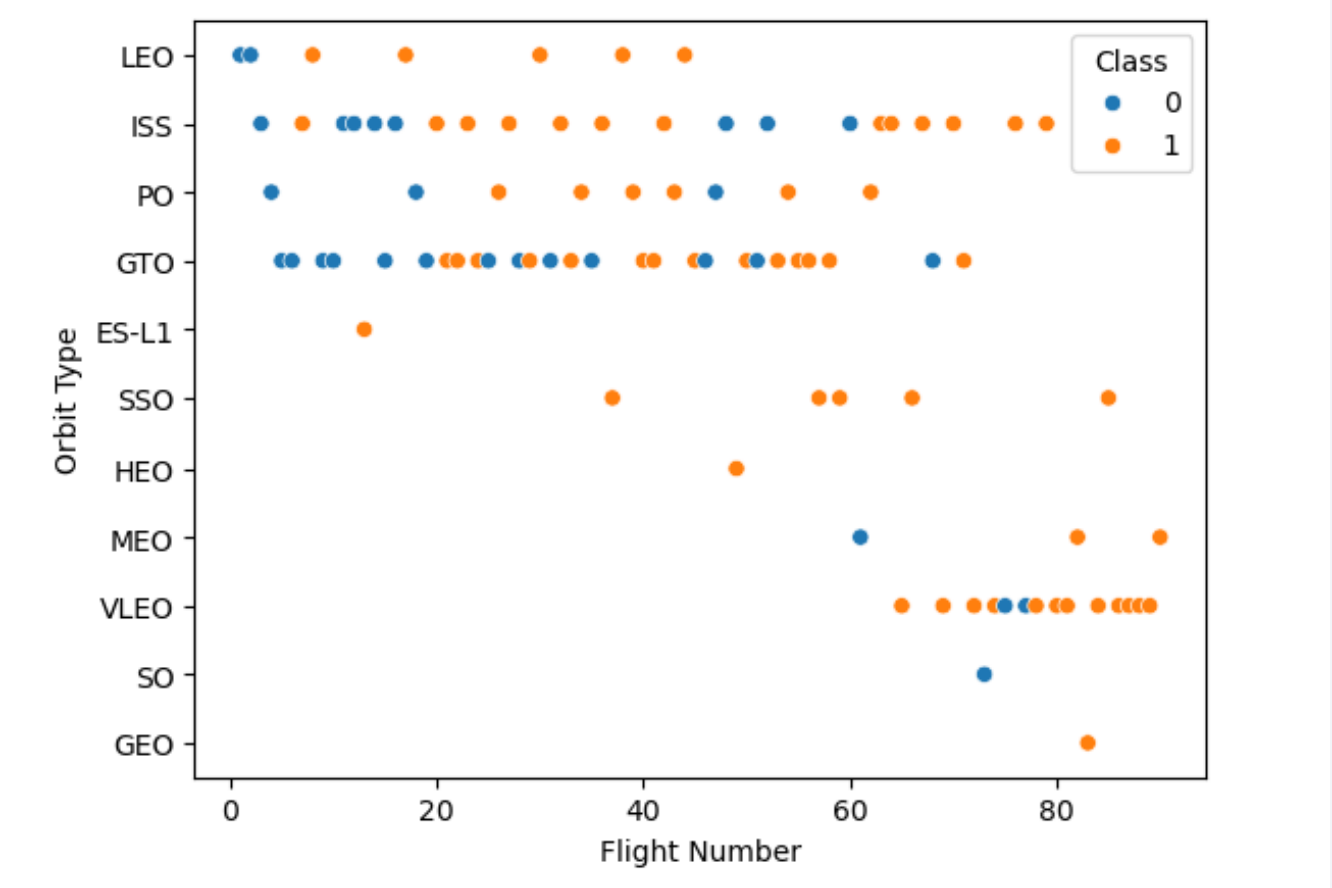
Success Percentage by Orbit Type



Flight Number vs. Orbit Type

- This plot shows that some of our very high success rate orbits, like SSO and ES-L1 actually do have low volume of launches. Meaning we should be cautious about drawing conclusions based on this low volume.
- The GTO Orbit has a higher volume of launches, and a lower success percentage.
- We generally see a trend of more successful launches the higher the flight number, and a shift in the orbit type of the launches as flight numbers get higher.

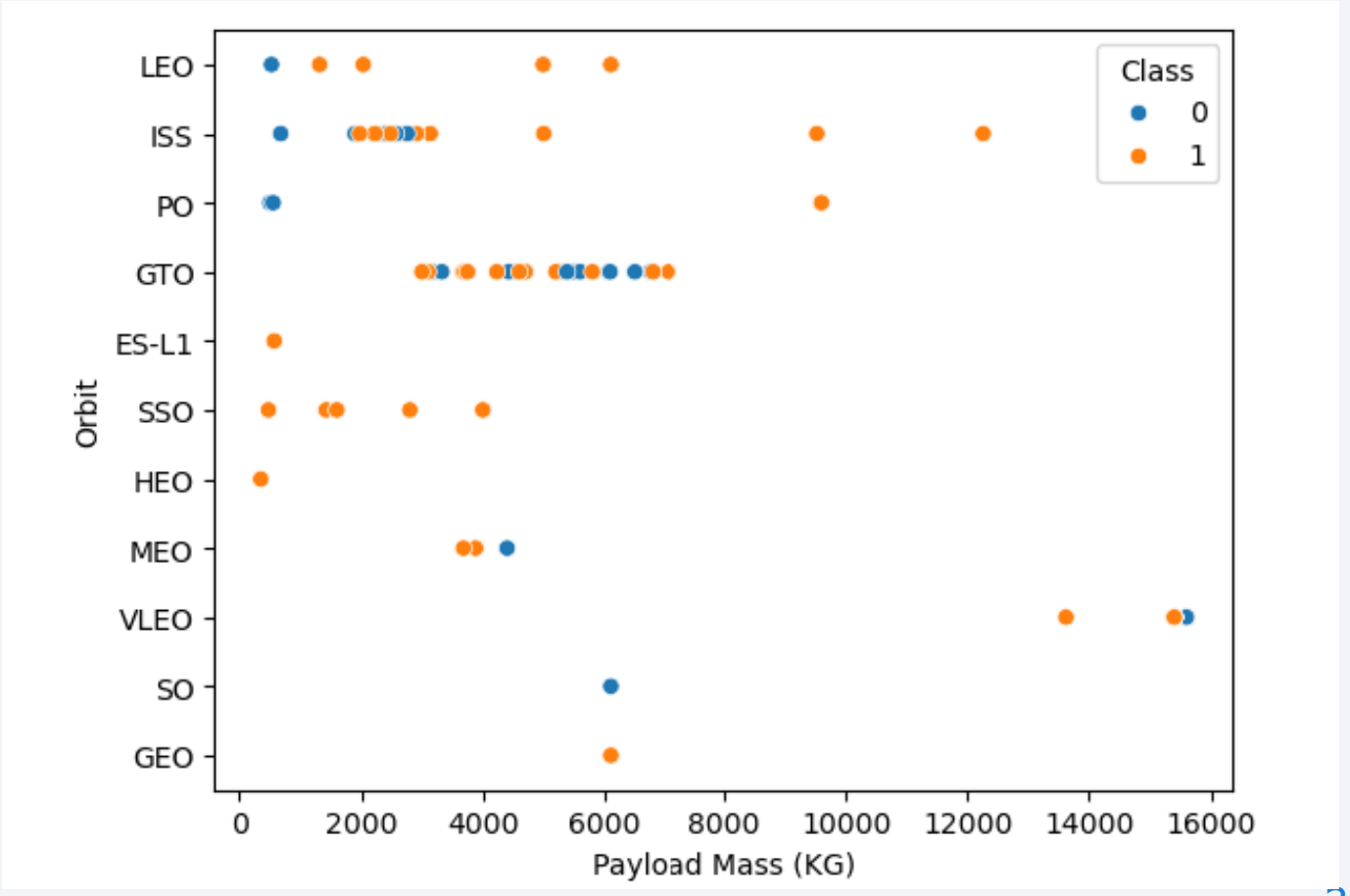
Flight number by Orbit Type



Payload vs. Orbit Type

- The heaviest payload are only sent into Very Low Earth Orbit (likely due to the higher fuel costs of getting them into the atmosphere).
- Lighter payloads are put into various different kinds of orbits.

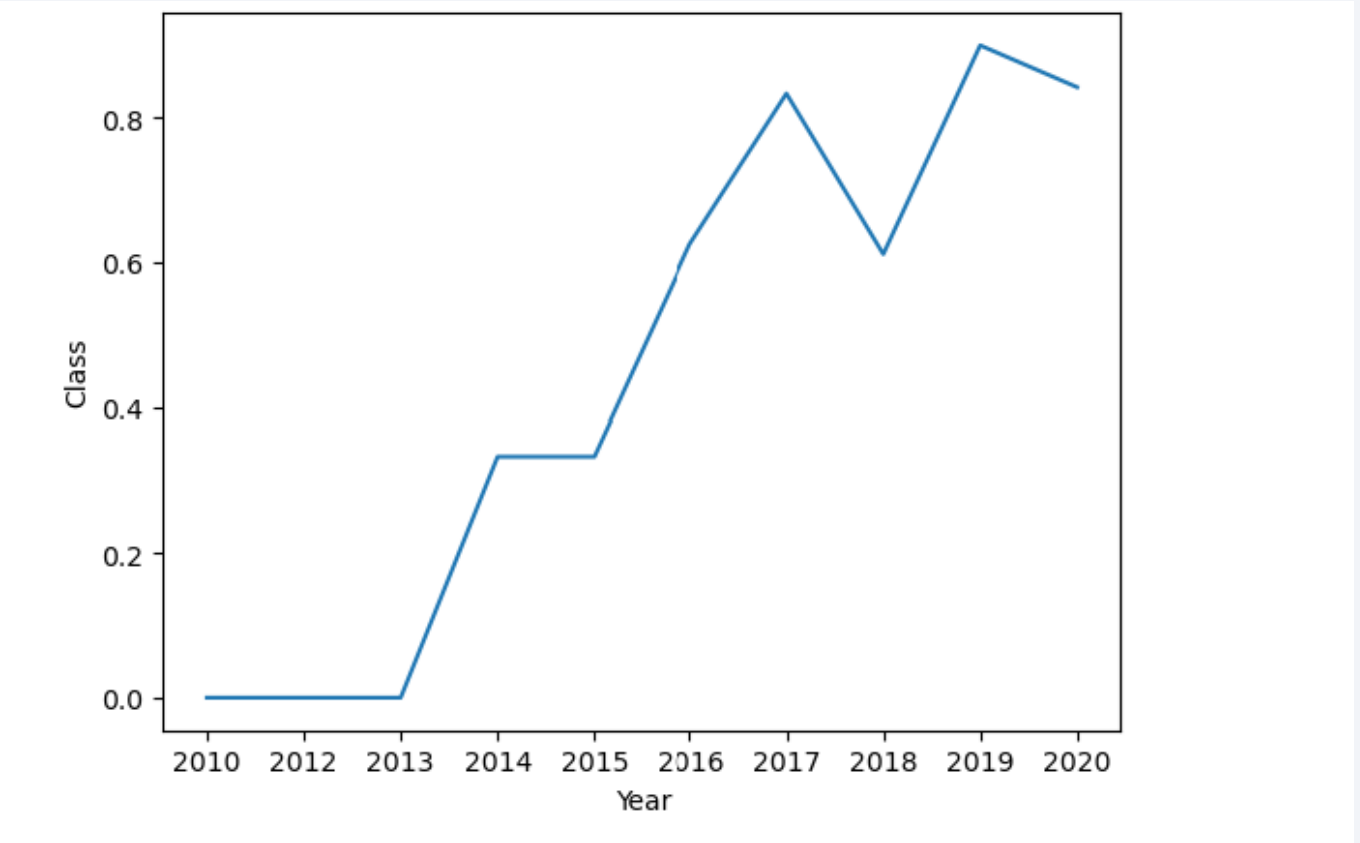
Payload Mass (KG) by Orbit Type



Launch Success Yearly Trend

- Broadly speaking, as Spacex advanced in their program, their success rates went up.
- Of course they had some regression from 2017 to 2018 but recovered after 2018.
- Rates peaked near 100% before declining to closer to 90% in 2020.

Average Launch Success Rate by Year



All Launch Site Names

- Spacex Unique Launch Sites are shown to the right.
- We can see Spacex utilizes four named sites over the course of our study period.

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

Launch Site Names Begin with 'CCA'

- This chart gives us an idea of the kinds of payloads, payload masses, customers, and orbits used by Sites beginning with the Spacex ‘CCA’ tag.

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 from launches where NASA was the customer equaled:
  - 45596 kg.
- This gives us an idea for how heavy all of the NASA payloads were, which gives us a general indication of NASA's demand for Spacex's services. Though there are likely better indicators of demand.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was:
  - 2928.4 kg
- This information gives us an indication of the general weight that the F9 booster is able to put into orbit.
- This would be an important planning consideration if we were planning a launch. It also gives us an indication for the booster's purpose which appears to be lighter loads (based on the information presented previously here)

First Successful Ground Landing Date

- The first successful Ground Pad Landing for Spacex occurred on:
- December 22, 2015
- This date gives us an idea of when their program started to become more successful, as landing on a ground pad was the first step in Spacex's overall development program which saw them land on increasingly smaller targets.

Successful Drone Ship Landing with Payload between 4000 and 6000

- The image on the right demonstrates that a lot of Spacex's booster types have been able to successfully land on a drone ship, which 7 distinct types doing so.

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1036.1
F9 B4 B1041.1
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The chart on the right shows that there are many more successful missions than unsuccessful ones. (Note, this is not referencing landing success)

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

Boosters Carried Maximum Payload

- The table on the right shows us that the F9 B5 booster, and its various versions are purpose built to carry Spacex's payloads.
  - We know they are purpose built as there are no other booster versions that carry Spacex's maximum payload.

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

- This table demonstrates that there were two unsuccessful drone ship landings in 2015.
- Both missions used the F9 v1.1 booster and both launched from the CCAFS LC-40 site.

month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- The highest volume of category was the “no landing attempt” version, which was likely used more frequently earlier in Spacex’s development.
- There are 5 successful drone ship landings and 5 failures on drone ships, illustrating that Spacex has not yet mastered drone ship landing. But, they are able to land successfully 50 percent of the time, which means they can recover 50 percent of their boosters and reuse them.

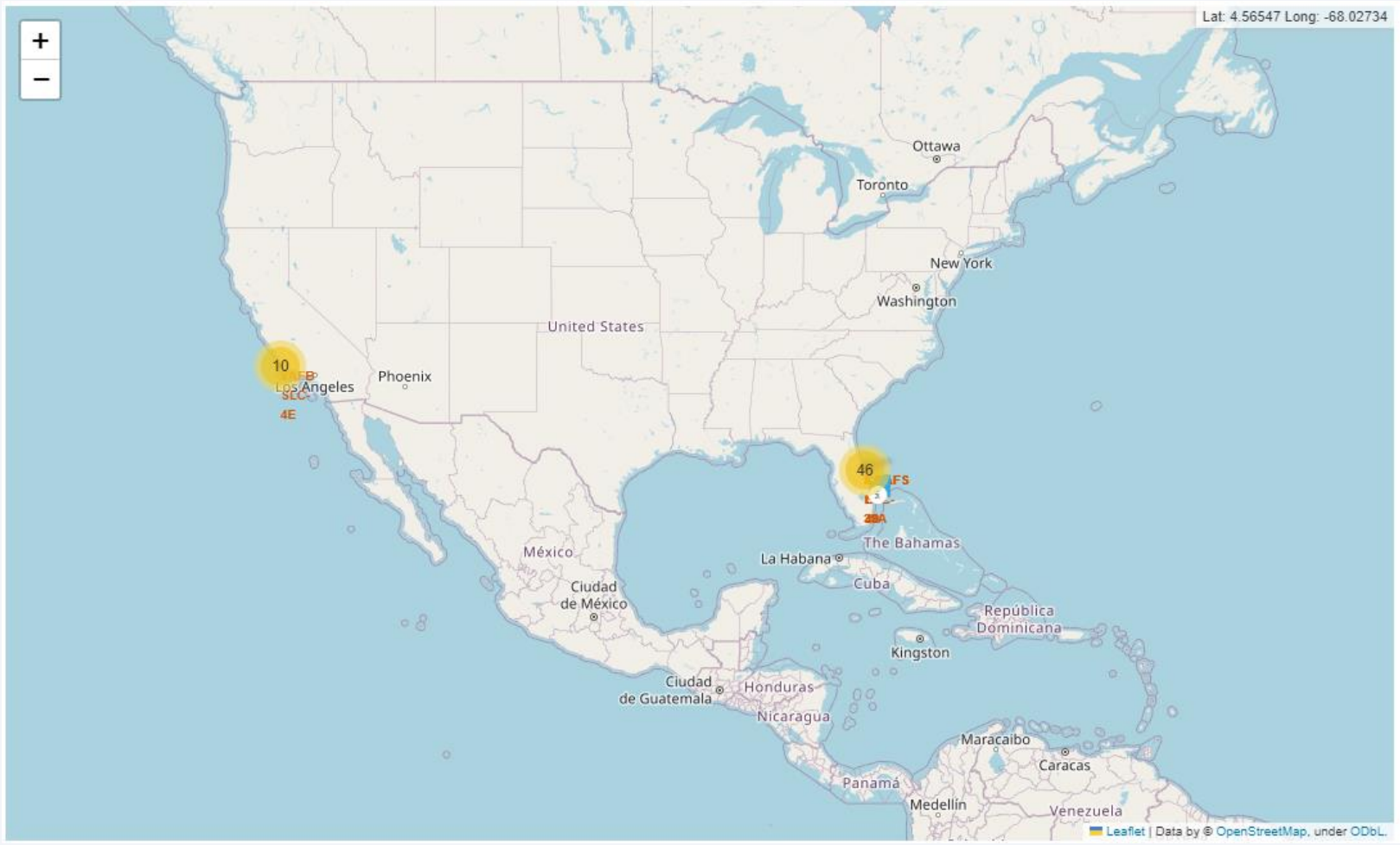
count	Landing_Outcome
10	No attempt
5	Success (drone ship)
5	Failure (drone ship)
3	Success (ground pad)
3	Controlled (ocean)
2	Uncontrolled (ocean)
2	Failure (parachute)
1	Precluded (drone ship)

Section 3

Launch Sites Proximities Analysis

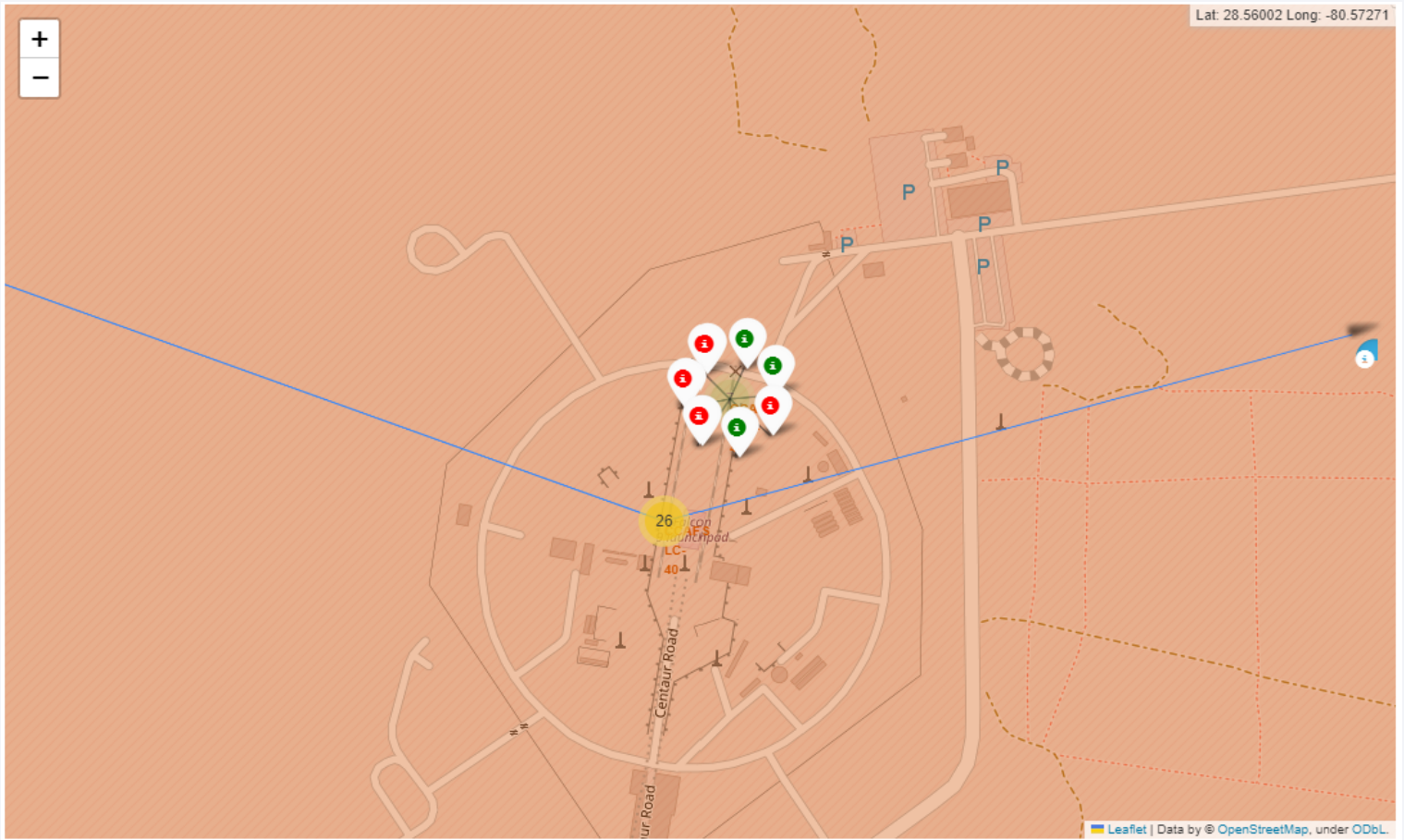
Spacex US Launch Sites

- This map demonstrates that Spacex uses both US coastlines for their launches.



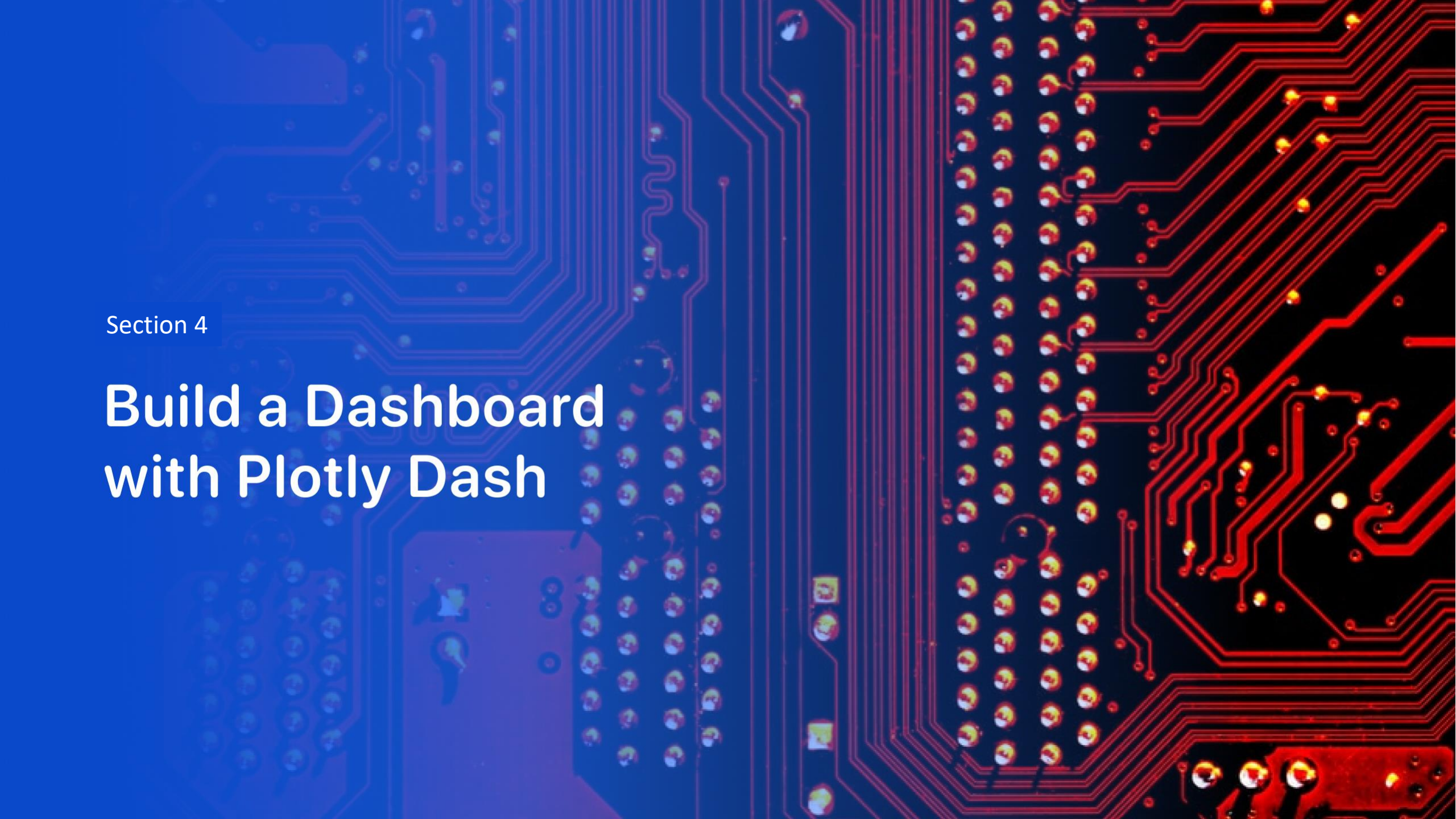
CCAFS-SLC 40 Map

- This map shows the number and success status of the launches from the CCAFS-SCL 40 site.
- The map shows us that 3/8 launches from this site ended successfully.



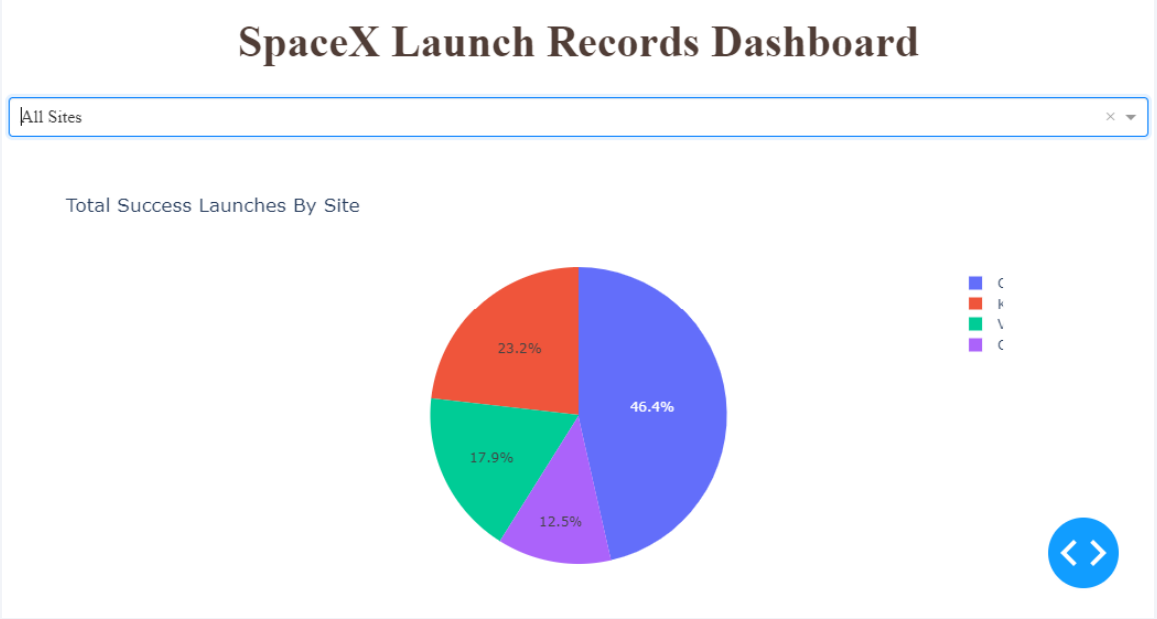
Map of the Florida Launch Sites

- Map of CCAFS-SLC 40 and LC-40 sites
- This map demonstrates that the launch site is close to the coastline and rail infrastructure.
- It is also far away from populated areas and a large highway.
- These factors taken together can help us infer that the sites are close to necessary infrastructure (rails) but far away from densely populated areas (for safety).


Section 4
Build a Dashboard with Plotly Dash

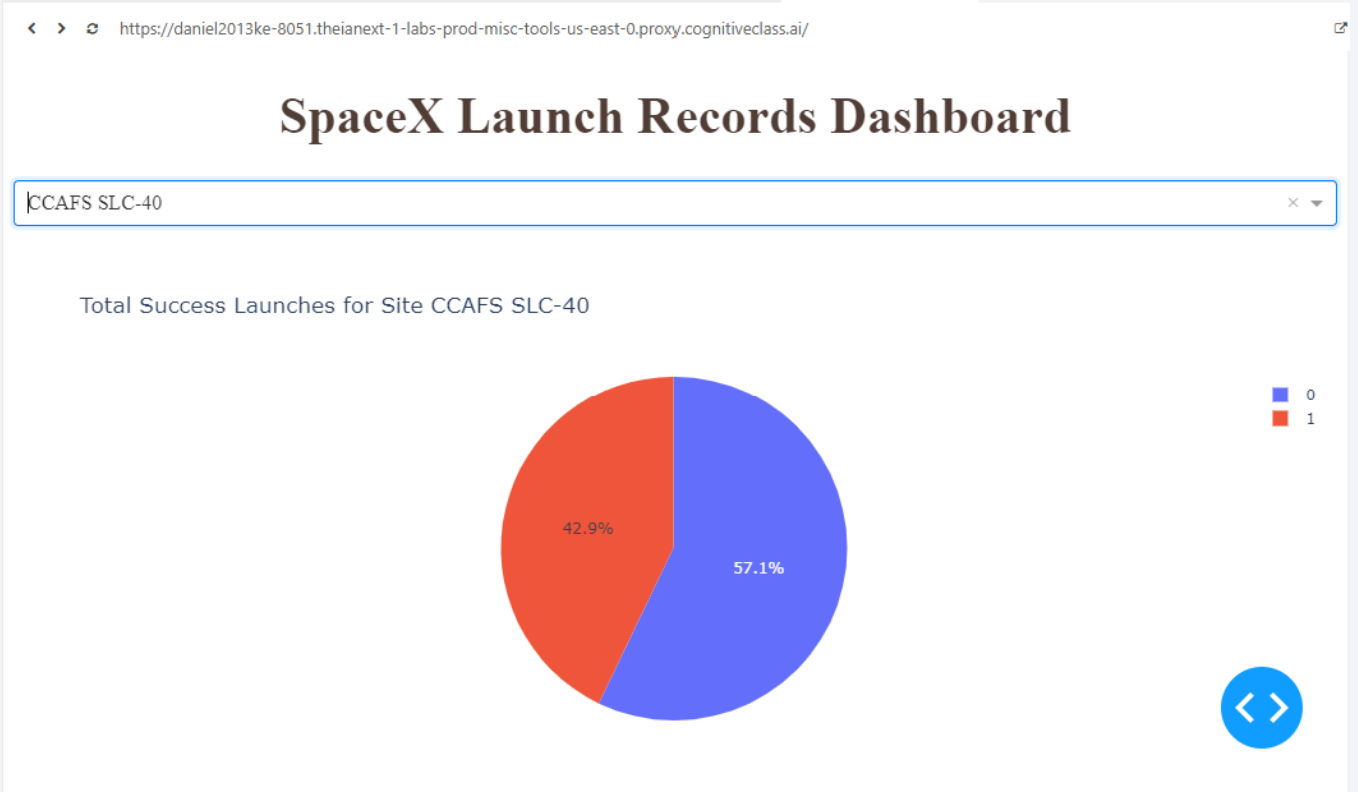
Percentage of Successful Launches by Site

- This chart shows the percentages of successful launches by site. The percentages are only factoring in the number of successful launches, and saying which site had a corresponding percentage of those successful launches.
- The CCAFS-SLC 40 Site had the greatest percentage of successful launches.



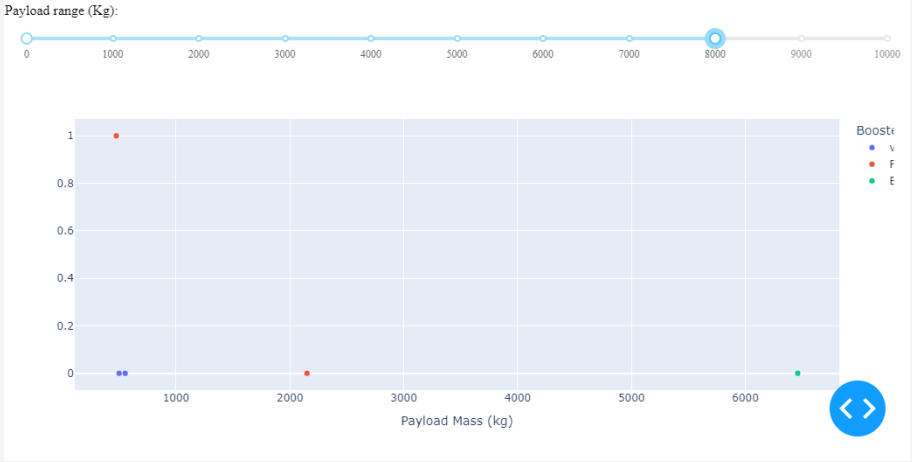
Launch Site – Greatest Success Percentage

- The site with the highest percentage of successful launches is the CCAFS SLC-40 launch site.
- This screenshot illustrates that 42.9% of the launches from the CCAFS SLC-40 site were successful.

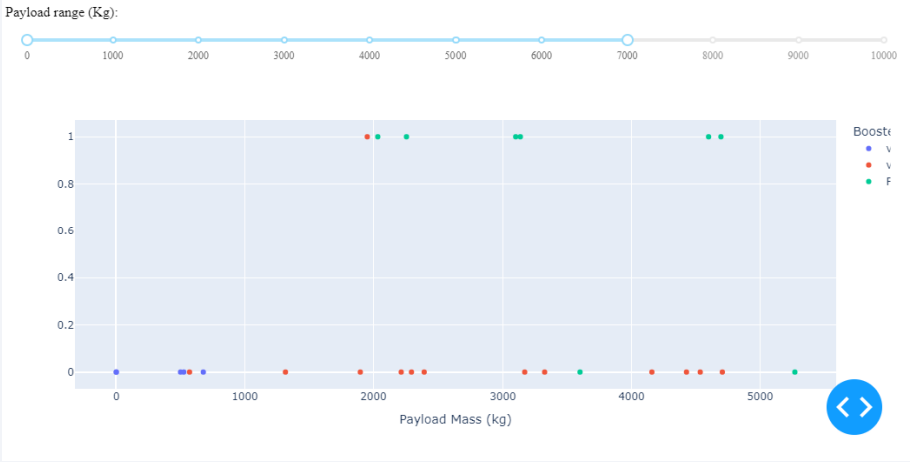


Payloads and Launch Outcomes

- The Key takeaways from this analysis are that:
  - The FT Booster has highest success rates
  - Heavier payloads (greater than 9000kg) have more success (proportionally, this may be affected by a lower sample size)

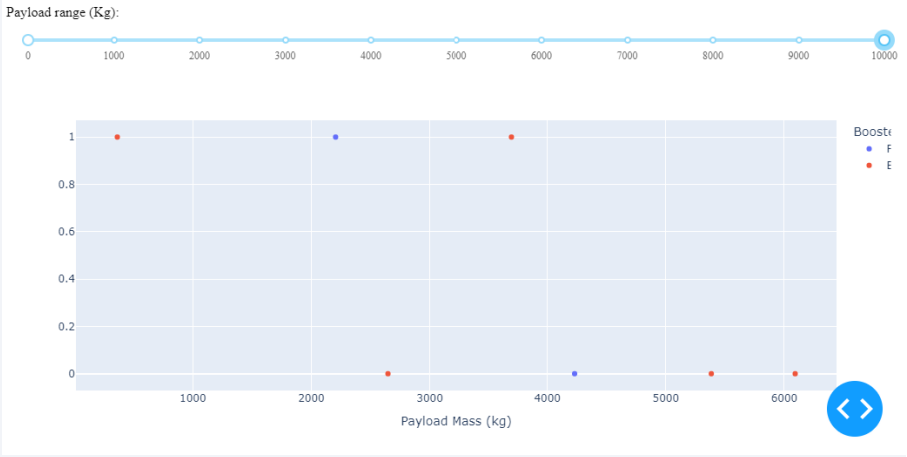


VAFB Payloads and Outcomes



CCAFS-LC 40 Payloads and Outcomes

KSC Payloads and Outcomes



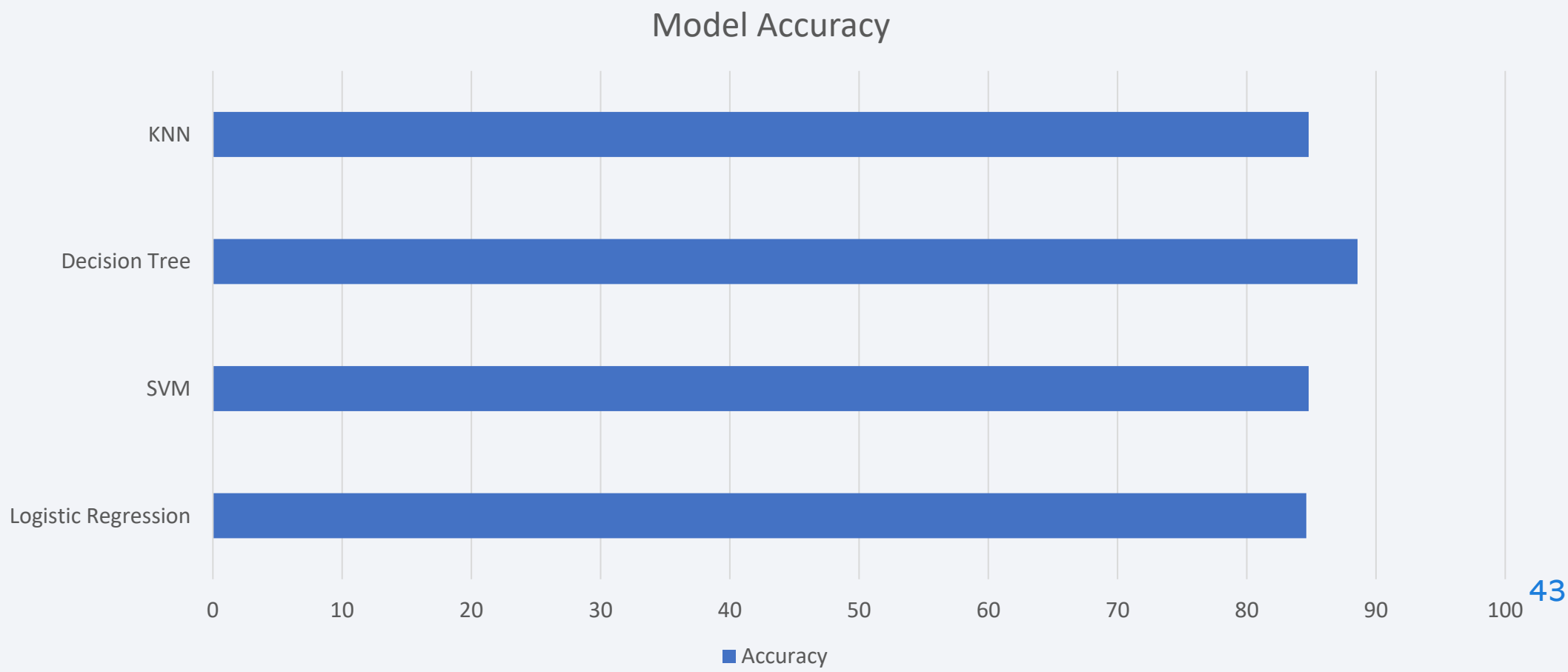
CCAFS-SLC 40 Payloads and Outcomes

Section 5

Predictive Analysis (Classification)

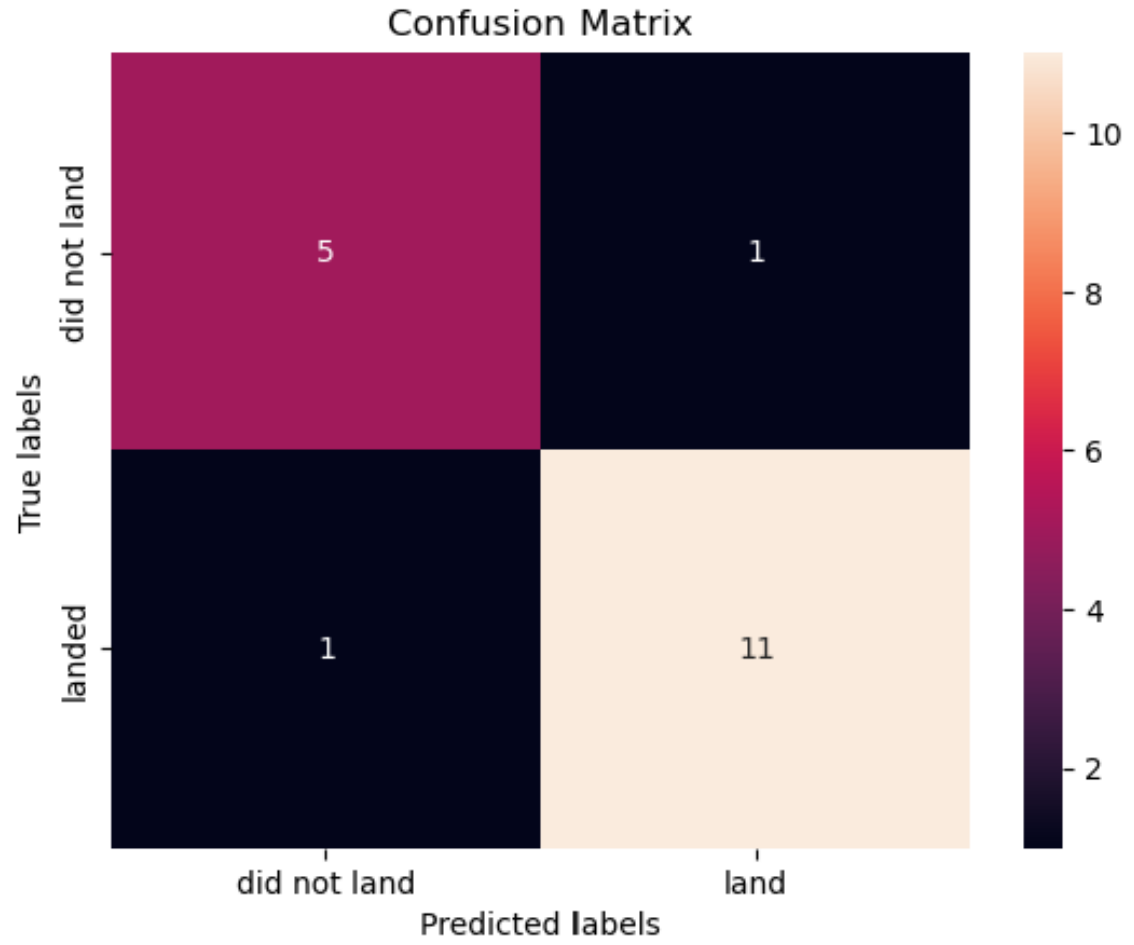
Classification Accuracy

- The Decision Tree model has the highest accuracy of all models tested



Confusion Matrix – Decision Tree

```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



- This is the confusion matrix output for the decision tree mode (the best performing model)
- This model had the best F1 score because it had the fewest misclassifications of the other models.
- It only had one false positive and one false negative.

Conclusions

- Point 1
- Spacex can land its booster successfully on a drone ship 50% of the time according to our dataset, this is an important fact for keeping their overall program costs low.
- The site with the highest percentage of successful launches is the CCAFS SLC-40 launch site.
- Based on the outputs of our data analysis, we would select the decision tree model to predict whether or not a landing will be successful for a spacex launch.
- …

Thank you!