IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jacob Martin
14 January 2025

Outline

EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

RESULTS

CONCLUSION

APPENDIX

Executive Summary - Methodologies

Data Collection:

Retrieved initial rocket launch data via SpaceX API, with supplementary launch details from web scraping the *"List of Falcon 9 and Falcon Heavy launches"* Wikipedia page (updated June 9, 2021).

Data Preprocessing:

Replaced missing values with column means (except for Landing Pad data left as NaN).

Filtered results to include only Falcon 9 launches.

Exploratory Data Analysis (EDA):

Examined relationships between Flight Number, Payload Mass, launch sites, orbit types, and landing outcomes.

Machine Learning:

Standardized key features and used Grid Search Cross Validation with four models: Logistic Regression, Support Vector Machines, Decision Tree, and k-Nearest Neighbors.

Executive Summary - Results

Key EDA Insights:

- Higher Flight Numbers correlate with greater landing success rates.
- VAFB SLC 4E and KSC LC 39A sites have higher success rates compared to CCAFS SLC 40.
- Orbit types ES-L1, GEO, HEO, and SSO achieved 100% success rates.
- Success rate plateaued around 85% since 2020.

Model Performance:

- Logistic Regression, Support Vector Machines, and k-Nearest Neighbors achieved similar test accuracies.
- Best performing model: Logistic Regression (83.33% accuracy).
- Worst performing model: Decision Tree (77.78% accuracy).

Introduction

SpaceX advertises Falcon 9 rocket launches on its website with a cost of $62 million; other providers cost upward of $165 million each, much of the savings is because SpaceX can reuse the first stage.

Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

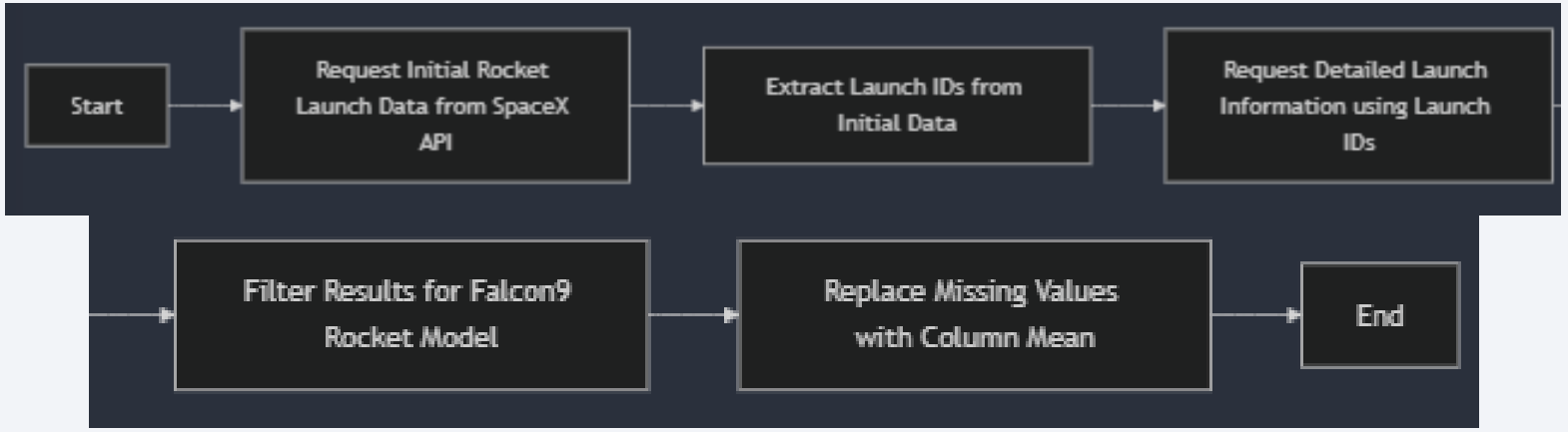
Data Collection - SpaceX API

The process begins by making an API call to the SpaceX API to request initial rocket launch data. Next, the launch IDs obtained from the initial API call are extracted for further use. Separate API calls are made using the previously extracted launch IDs to get detailed information about each launch. The results are then filtered to only include those pertaining to the Falcon 9 rocket model.

All missing (NaN) values are replaced with the mean of that column, except for the missing values related to the Landing Pad which are left as NaN.

Data Collection – SpaceX API

- https://github.com/jacobmartin26/IBM-Data-Science/blob/main/Capstone/Lab%201%20Collecting%20the%20data.ipynb

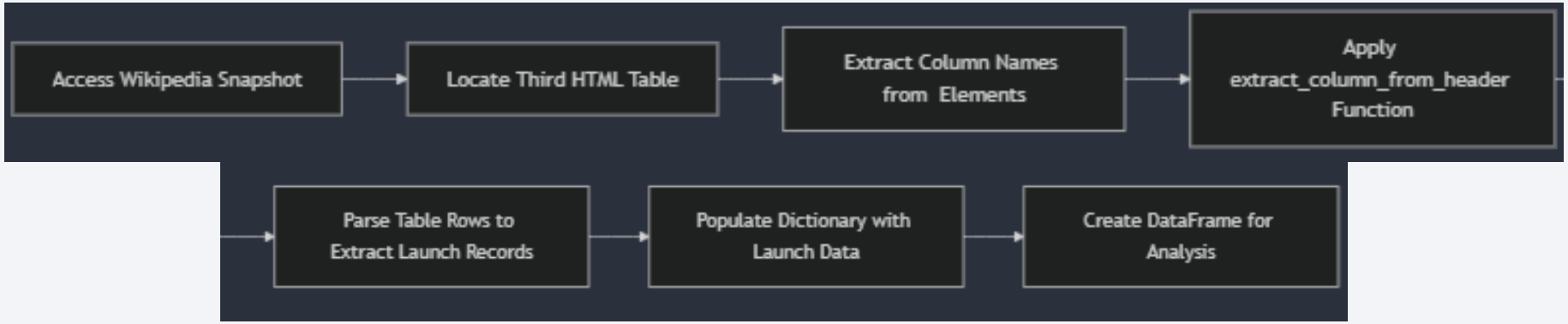


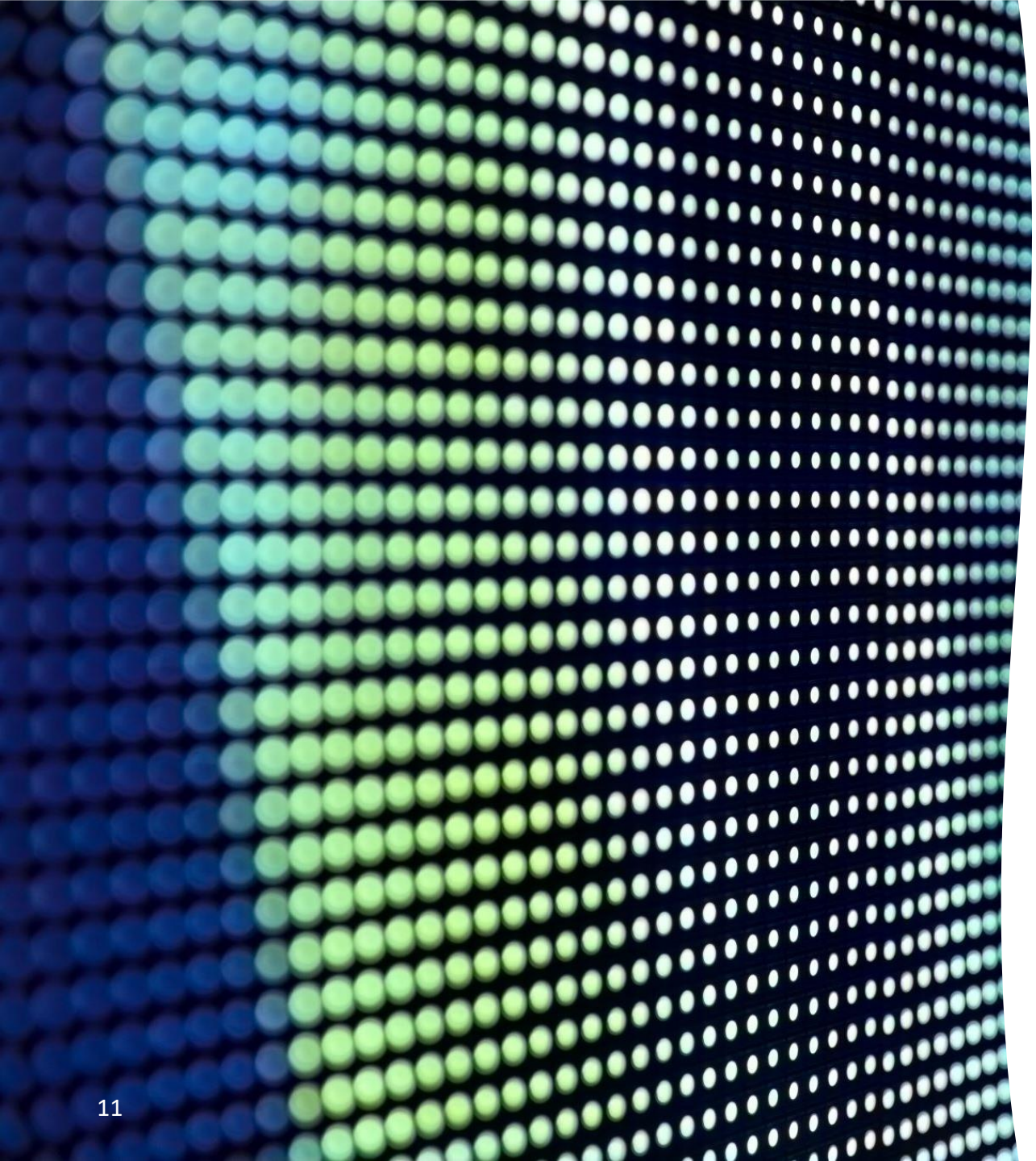
Data Collection - Scraping

- Webpage Selection:
  - A snapshot of the *"List of Falcon 9 and Falcon Heavy launches"* Wikipedia page (updated on June 9, 2021) was used as the data source.
- Target Table Identification:
  - The third HTML table on the page was identified as containing the relevant launch records.
- Column Name Extraction:
  - Extracted all column/variable names by iterating through the <th> elements in the table header using the extract_column_from_header() function.
- Row Parsing:
  - Parsed each row of the HTML table to extract launch records.
- Data Frame Creation:
  - Populated a dictionary (launch_dict) with the extracted records and transformed it into a structured DataFrame for analysis.

Data Collection - Scraping

- https://github.com/jacobmartin26/IBM-Data-Science/blob/main/Capstone/Lab%201.5%20Web%20scraping%20Falcon%209%20and%20Falcon%20Heavy%20Launches%20Records%20from%20Wikipedia.ipynb





Data Wrangling

Missing Values Analysis:

- Identified and calculated the percentage of missing values in each attribute.

Data Type Identification:

- Classified attributes into numerical and categorical columns for further processing.

Launch Site Analysis:

- CCAFS SLC 40: 55 launches
- KSC LC 39A: 22 launches
- VAFB SLC 4E: 13 launches

Orbit Analysis:

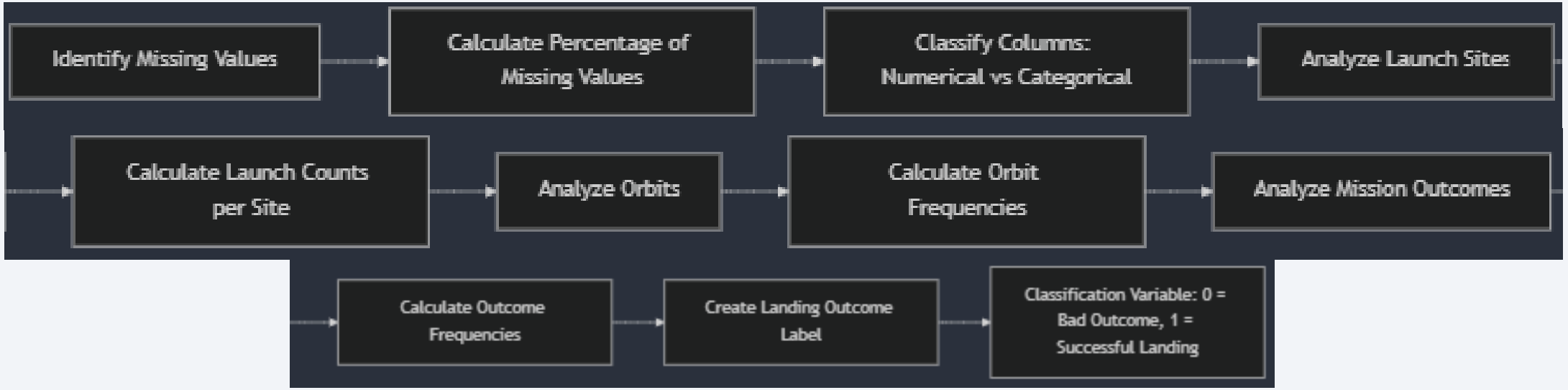
- GTO: 27 missions
- ISS: 21 missions
- VLEO: 14 missions
- Other orbit types had smaller frequencies, such as PO (9) and LEO (7).

Created a binary Landing Outcome Label based on the Outcome column:

- Value 0 for bad outcomes where the first stage did not land successfully.
- Value 1 for successful outcomes where the first stage landed successfully.

Data Collection - Scraping

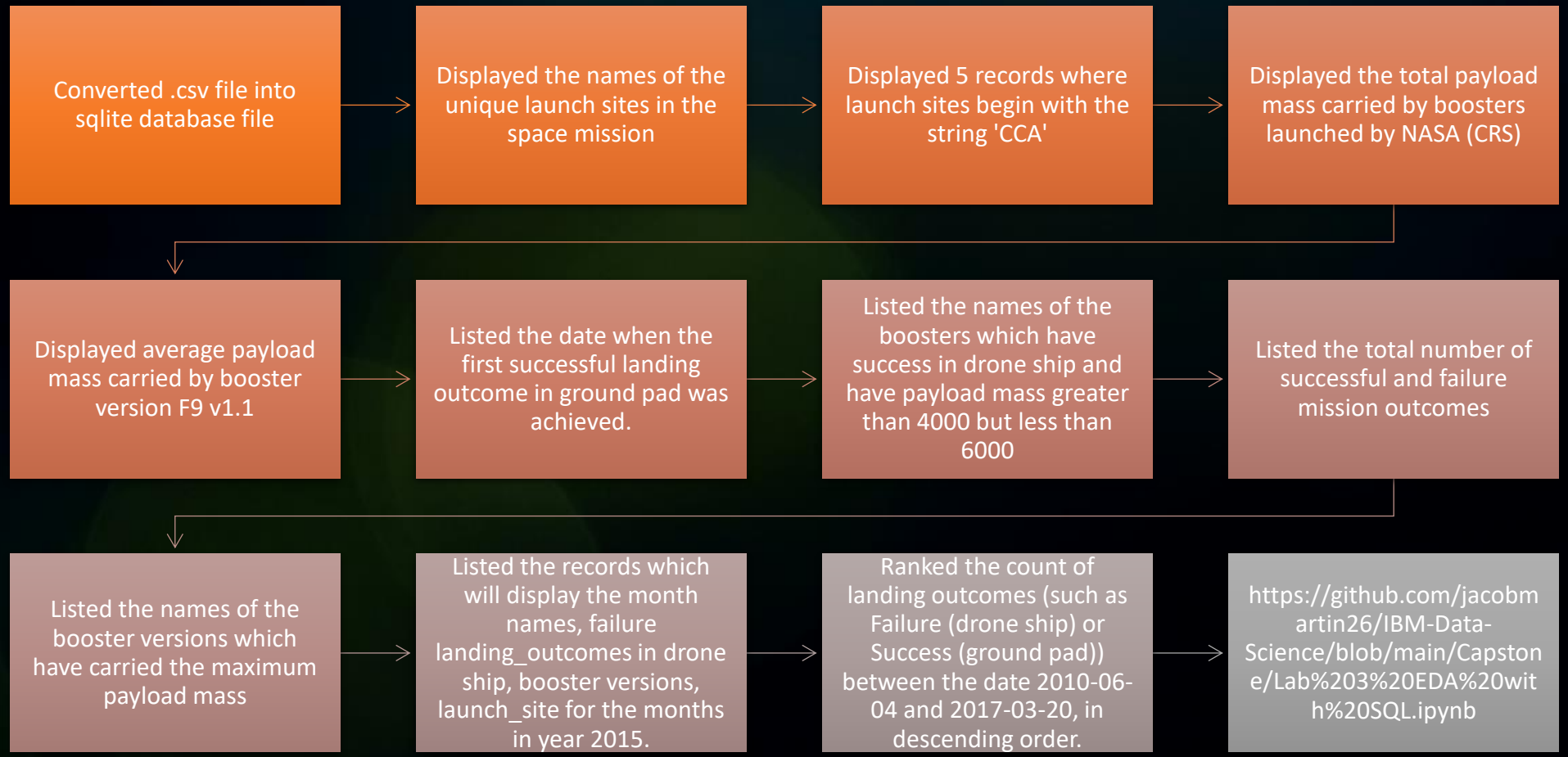
- https://github.com/jacobmartin26/IBM-Data-Science/blob/main/Capstone/Lab%202%20Data%20Wrangling.ipynb



EDA with Data Visualization

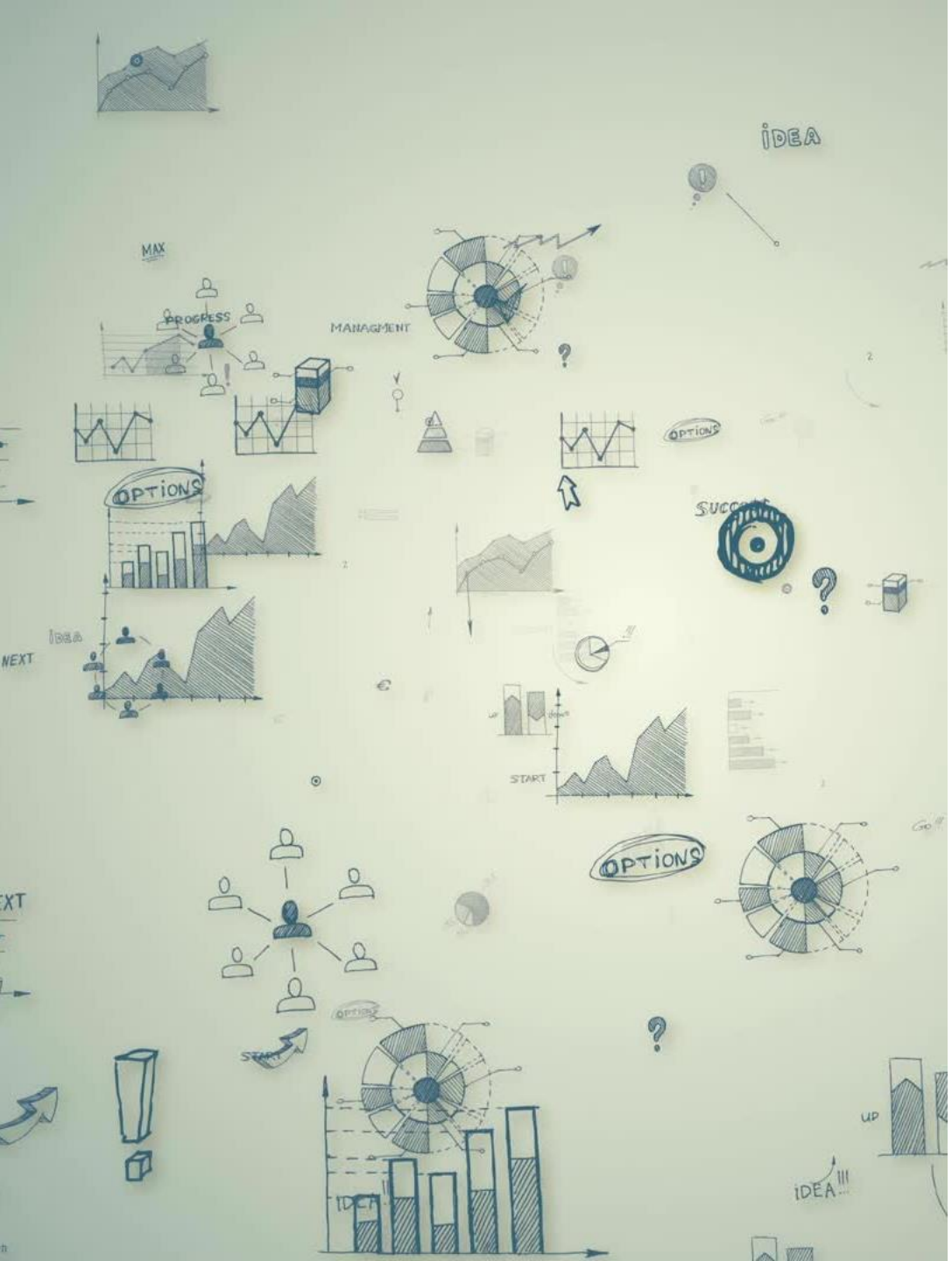
- The first relationship examined was how flight number and payload mass contribute to landing success rates. A scatterplot was used to easily show the correlation between the two features and our target variable. Scatterplots were also used to show the relationships between flight number vs launch site, payload mass vs launch site, flight number vs orbit type, and payload mass vs orbit type.
- A bar graph was used to show the success rates by each distinct orbit type.
- Lastly, a line graph was used to show how the overall success rate of launches changed year over year
- https://github.com/jacobmartin26/IBM-Data-Science/blob/main/Capstone/Lab%203.5%20EDA%20with%20Visualization%20Lab.ipynb

EDA with SQL


Converted .csv file into sqlite database file
Displayed the names of the unique launch sites in the space mission
Displayed 5 records where launch sites begin with the string 'CCA'
Displayed the total payload mass carried by boosters launched by NASA (CRS)
Displayed average payload mass carried by booster version F9 v1.1
Listed the date when the first successful landing outcome in ground pad was achieved.
Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
Listed the total number of successful and failure mission outcomes
Listed the names of the booster versions which have carried the maximum payload mass
Listed the records which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015.
Ranked 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.
https://github.com/jacobmartin26/IBM-Data-Science/blob/main/Capstone/Lab%203%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Added a Folium circle and marker to each launch site on the site map to mark and label where each launch site is located on the US.
- Added a marker cluster that holds markers of both successful and failed launches for each unique launch site on the map. Successful launches are green while failed ones are red. This makes it quicker to see which sites are more popular and/or more successful with its launches.
- Added a line between the CCAFS SLC-40 launch site and its nearest coastline, highway, railway, and city to see if launch sites were near certain structures.
- https://github.com/jacobmartin26/IBM-Data-Science/blob/main/Capstone/Lab%204%20Visual%20Analytics%20with%20Folium.ipynb

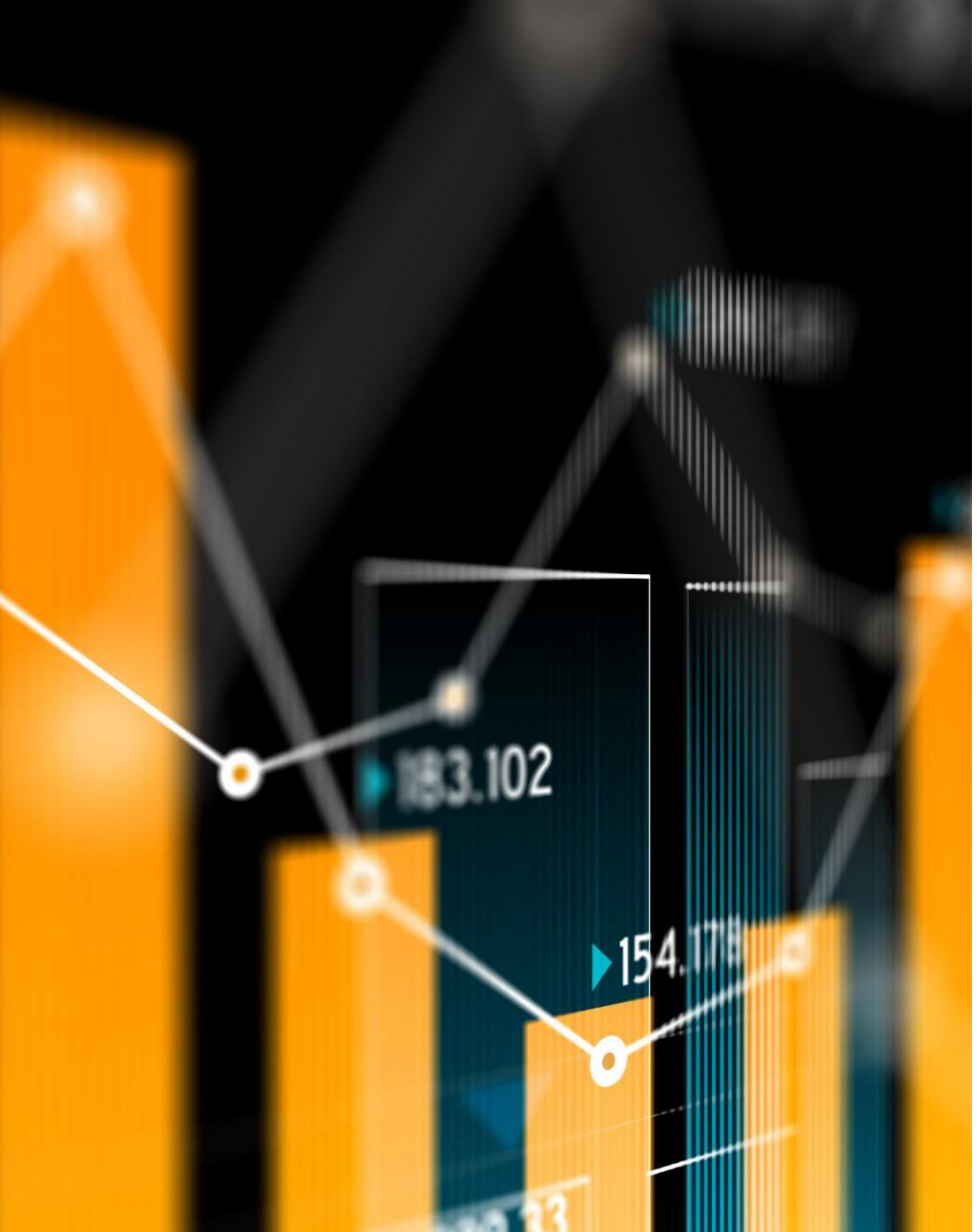


Build a Dashboard with Plotly Dash

- Added a dropdown list with all unique launch sites to enable launch site selection, defaults to all sites. The selected site will then change the data displayed in a pie chart that was added which will show the Success vs. Failed counts for the site. If all sites are selected, then it will instead show the total amount of successful launches for all sites.
- Added a slider to select a payload range (0 to 10000kg) for launches. This slider will update a scatter plot that was also added which shows the correlation between payload mass and launch success based on the selected launch site from the dropdown list.
- Overall, these features should make it quick and easy for the user to view key features for all launch sites or a single launch site.
- https://github.com/jacobmartin26/IBM-Data-Science/blob/main/Capstone/Lab%204.5%20Interactive%20Dashboard%20with%20Ploty%20Dash.py

Predictive Analysis (Classification)

- Feature Standardization:
  - Applied a Standard Scaler to standardize features assigned to X for consistency in model performance.
- Data Splitting:
  - Divided the dataset into training and testing subsets to evaluate model performance effectively.
- Model Selection and Evaluation:
  - Evaluated 4 classification models:
    - Logistic Regression
    - Support Vector Machines (SVM)
    - Decision Trees
    - K-Nearest Neighbors (KNN)
- Grid Search Cross Validation (CV):
  - Used Grid Search CV with a 10-fold cross-validation to optimize hyperparameters for all models and maximize accuracy.

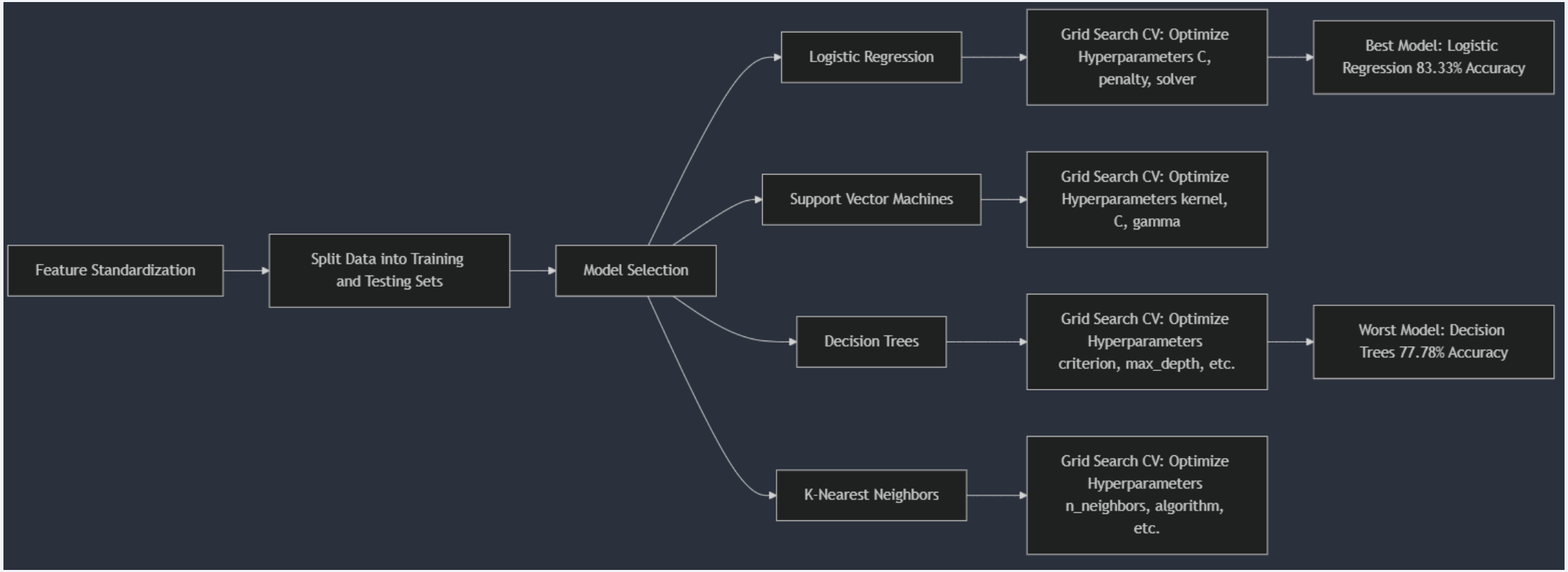


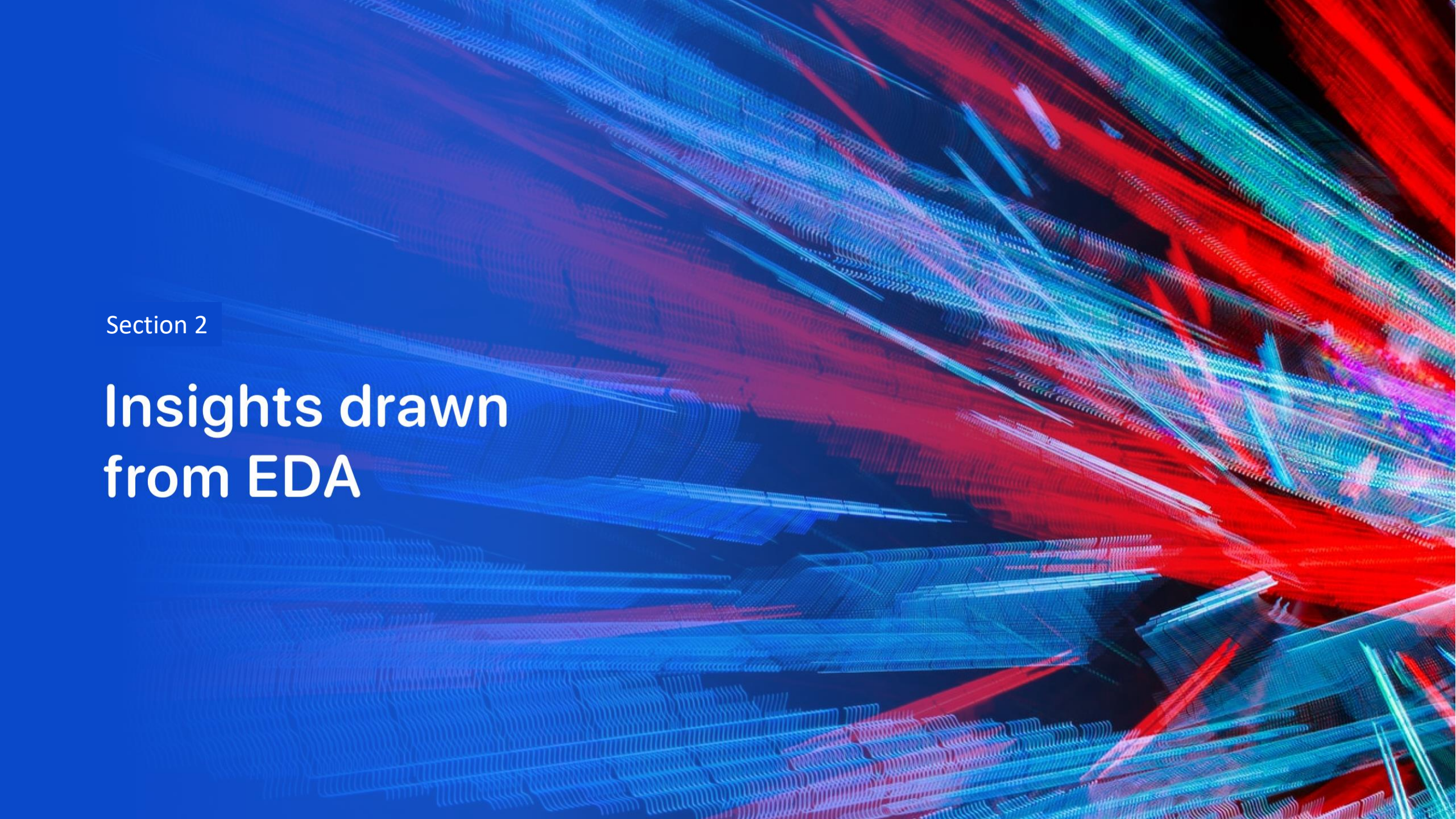
Predictive Analysis (Classification) Continued

- Hyperparameter Tuning:
  - Logistic Regression: Tuned C, penalty, and solver parameters.
  - SVM: Tuned kernel, C, and gamma parameters.
  - Decision Trees: Tuned criterion, splitter, max_depth, and other parameters.
  - KNN: Tuned n_neighbors, algorithm, and p parameters.
- Best-Performing Model:
  - Logistic Regression was the best-performing model with a test accuracy of 83.33%.
  - Decision Trees performed the worst with an accuracy of 77.78%.

Predictive Analysis (Classification) Continued

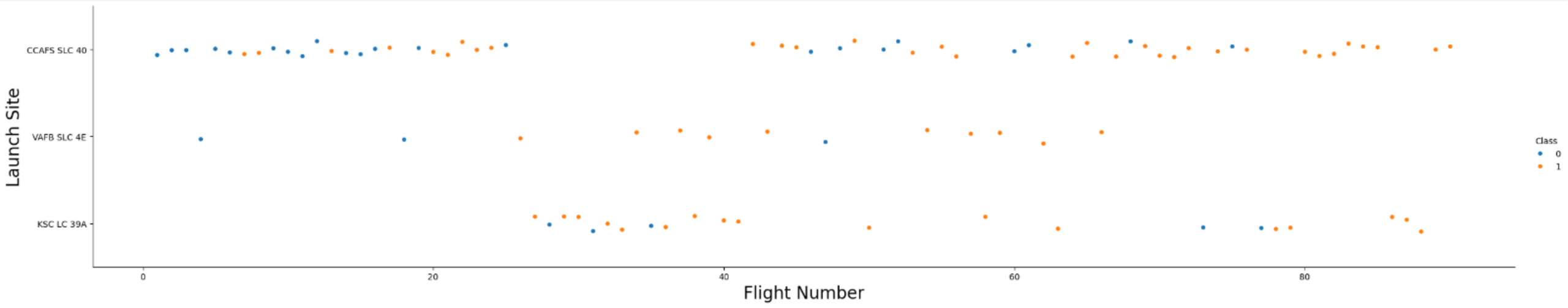
- https://github.com/jacobmartin26/IBM-Data-Science/blob/main/Capstone/Lab%205%20Machine%20Learning%20Prediction.ipynb




Section 2
Insights drawn from EDA

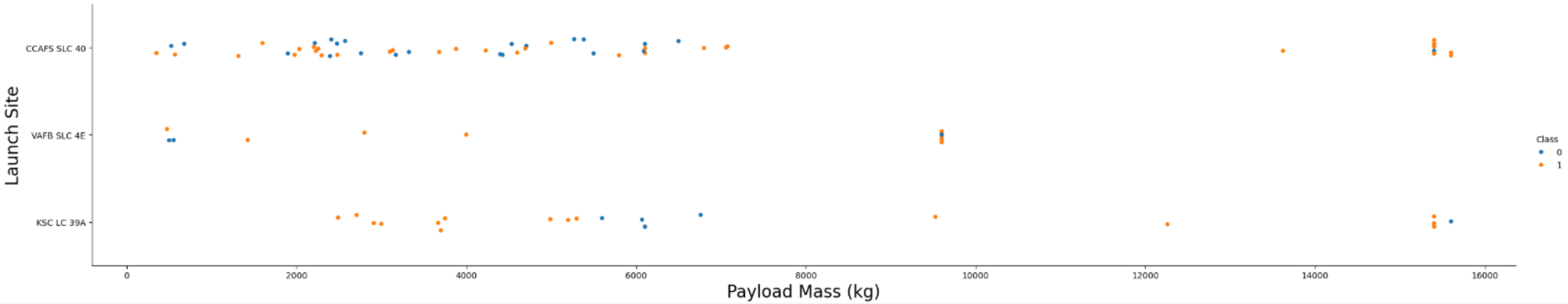
Flight Number vs. Launch Site

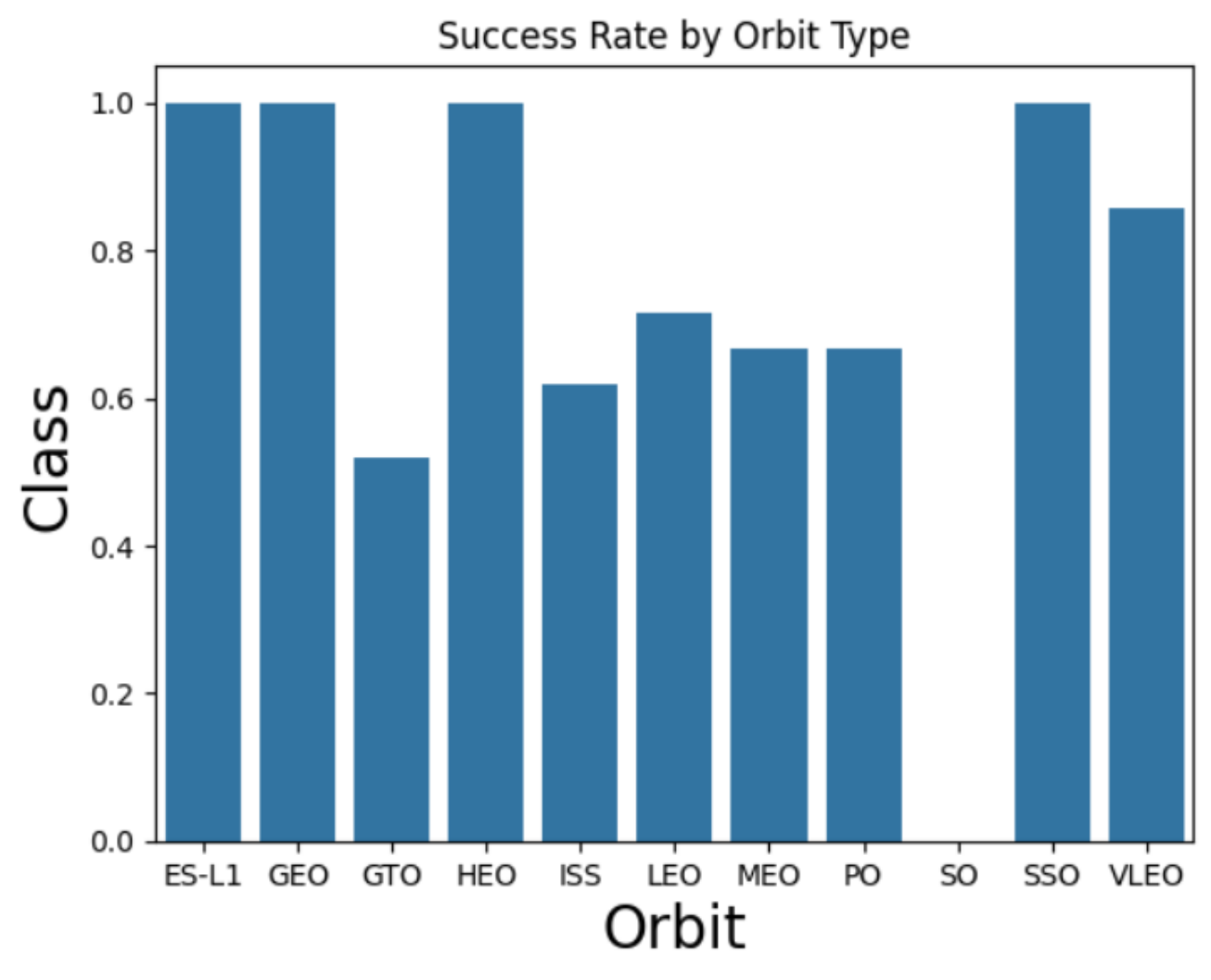
- CCAFS SLC 40 site appears to be the most popular site to launch from, but the other sites have a better success rate for their launches. However, the success rate for all sites has improved as number of flights has increased over time.



Payload vs. Launch Site

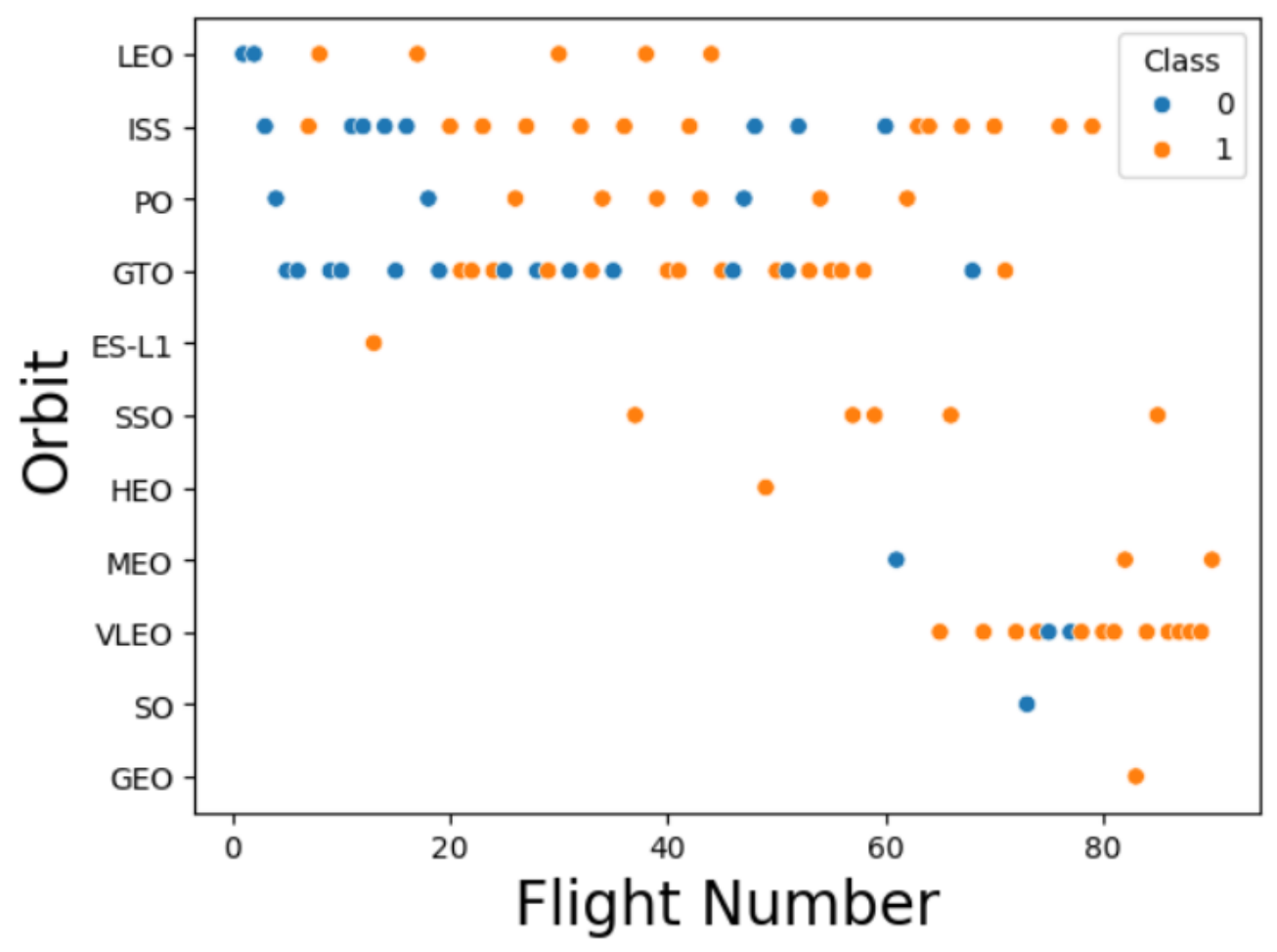
- All launch sites tend to have launches with a payload mass less than 10000kg with the VAFB SLC 4E having no launches with a mass greater than 10000kg. The other two sites do have some launches with a mass greater than 15000kg, but most are less than 8000kg.





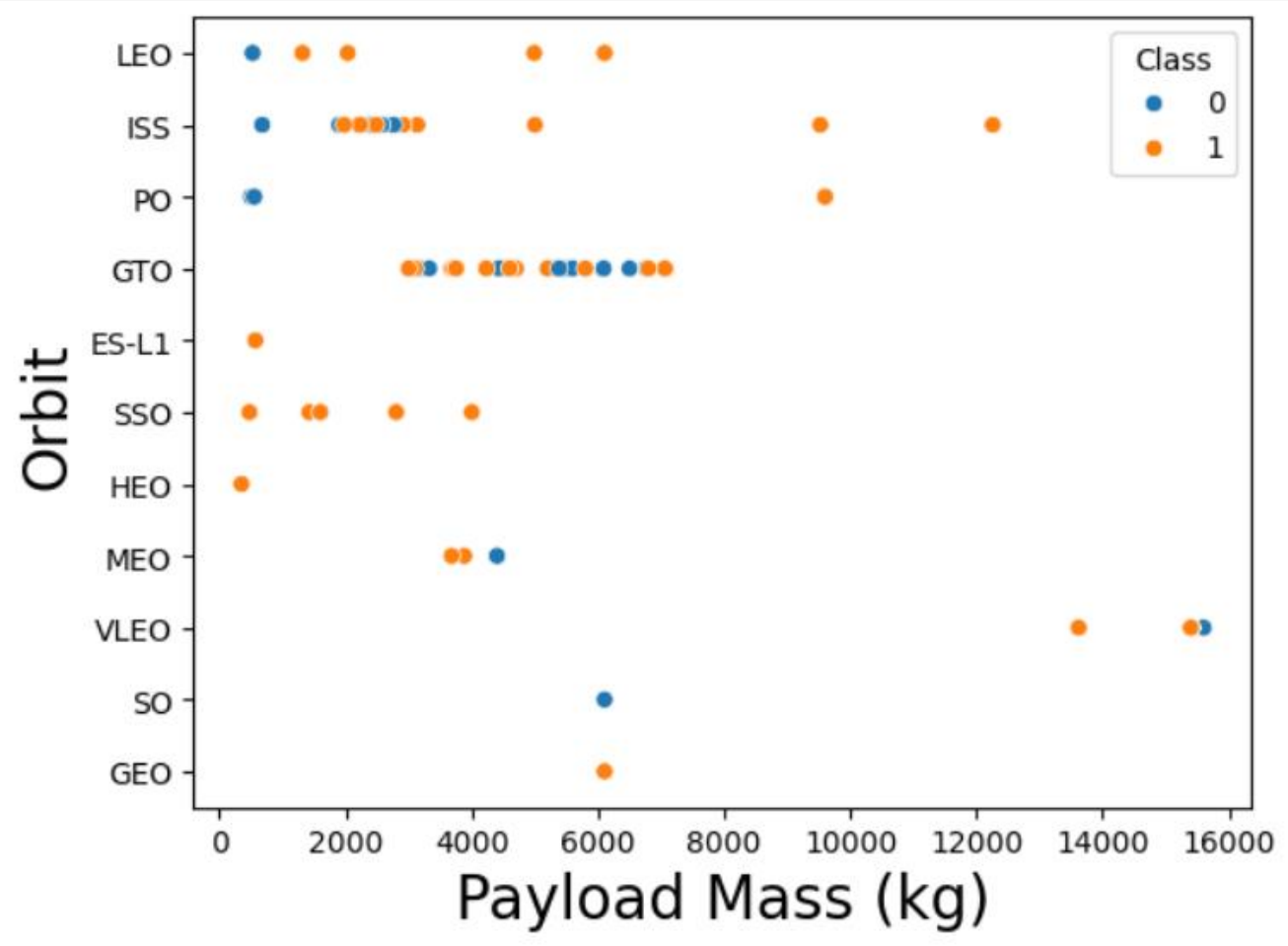
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO orbits have a 100% success rate while other orbit types have a rate between 50% to 85%. The worst performing orbit type was SO with a 0% success rate.



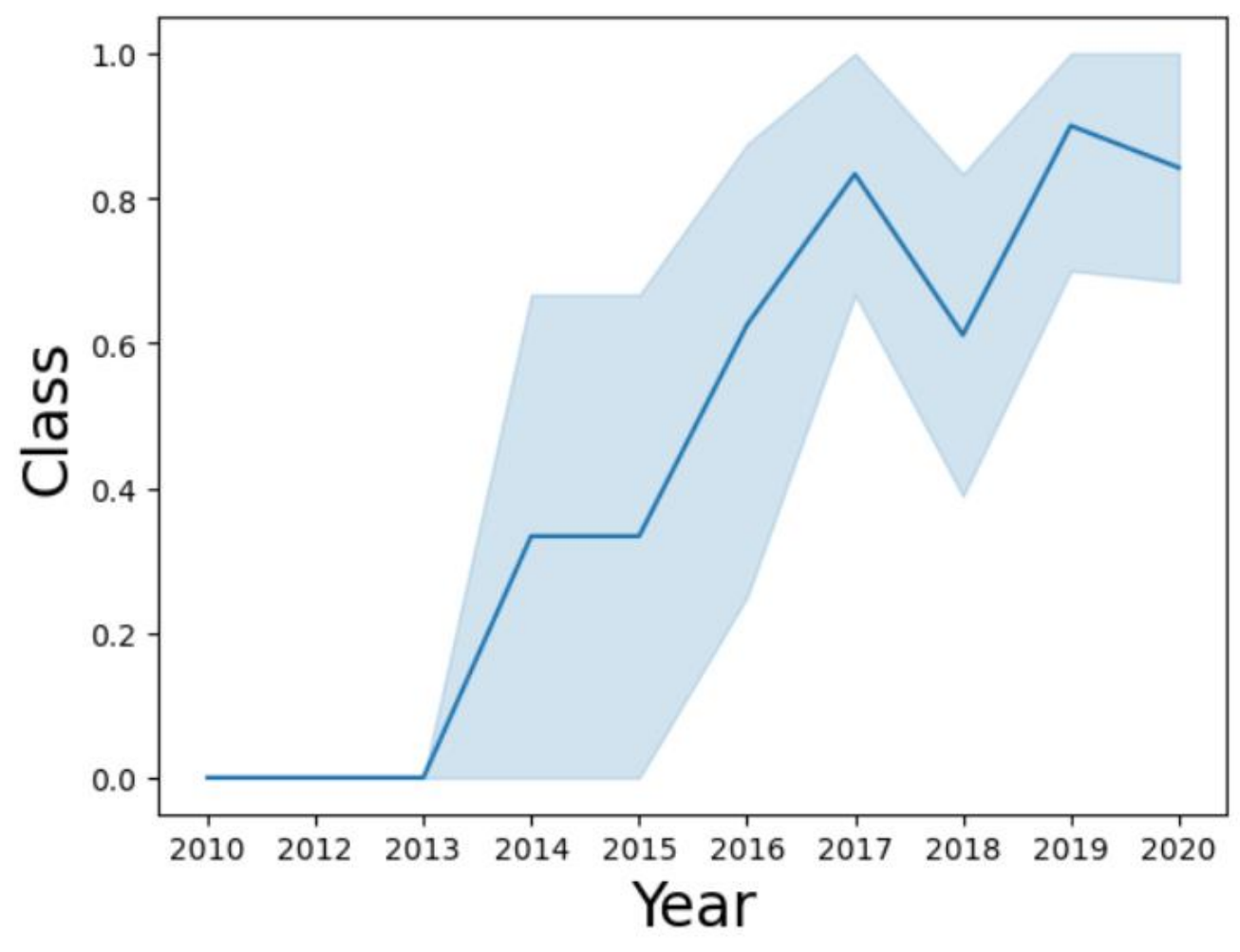
Flight Number vs. Orbit Type

- In the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success. Also, GTO orbit has become less popular as flight number increased, while the opposite is true for VLEO orbit.



Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for SSO, LEO, and ISS. However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.



Launch Success Yearly Trend

- Average success rate across all launches has steadily increased from 2013 to 2017. There was a drop in 2018, but it was improved in 2019. Success rate has since slowed down around the 85% mark.

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

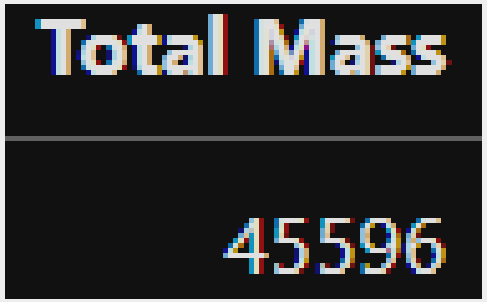
Launch Site Names

In the dataset, there are a total of four unique launch sites used for SpaceX launches.

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

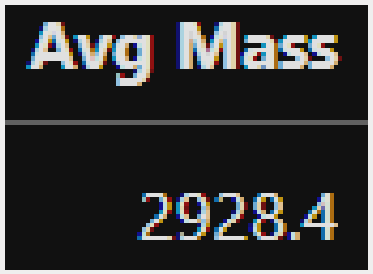
Launch Site Names Beginning with 'CCA'

- The first five launches done at this site had successful mission outcomes, but failed landings or they did not attempt a landing. The rockets all had a very low payload mass or no payload mass at all and were in LEO orbit.



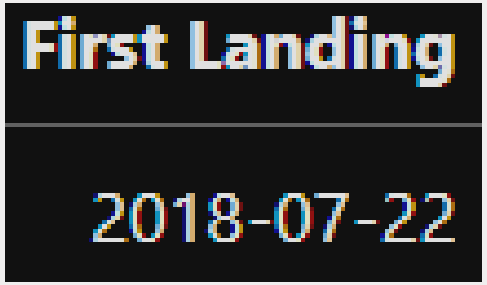
Total Payload Mass

The total payload mass carried by boosters launched by NASA (CRS) was 45596kg.



Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was 2928.4kg which is relatively light when compared to later launches with payload masses above 10000kg.



First Successful Ground Landing Date

The date of the first successful ground landing was 22 July 2018.

Successful Drone Ship Landing with Payload between 4000 and 6000

- There were eight different booster versions that had successful drone ship landings with a payload mass between 4000kg and 6000kg. However, only 3/8 launches had a payload mass above 5000kg.

Booster_Version	Landing_Outcome	PAYLOAD_MASS__KG_
F9 B5 B1046.2	Success	5800
F9 B5 B1047.2	Success	5300
F9 B5 B1046.3	Success	4000
F9 B5 B1048.3	Success	4850
F9 B5 B1051.2	Success	4200
F9 B5B1060.1	Success	4311
F9 B5 B1058.2	Success	5500
F9 B5B1062.1	Success	4311

Total Number of Successful and Failure Mission Outcomes

- While landing outcomes typically had a wider range of successful or failed outcomes, the mission outcomes were almost always successful with a success rate of 99% overall.

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

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

Boosters Carried Maximum Payload

- Many different booster versions have carried the maximum payload mass (156000kg). However, most launches will not have a payload this heavy. Also, all these boosters come from the F9 B5 generation of boosters.

2015 Launch Records

- There were two failed drone ship landings in 2015, both occurring at the CCAFS LC-40 launch site and both a part of the F9 v1.1 generation. In 2016, there were more failed landings at more sites and occurring with more booster generations.

Month	Year	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
01	2016	Failure (drone ship)	F9 v1.1 B1017	VAFB SLC-4E
03	2016	Failure (drone ship)	F9 FT B1020	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
06	2016	Failure (drone ship)	F9 FT B1024	CCAFS LC-40

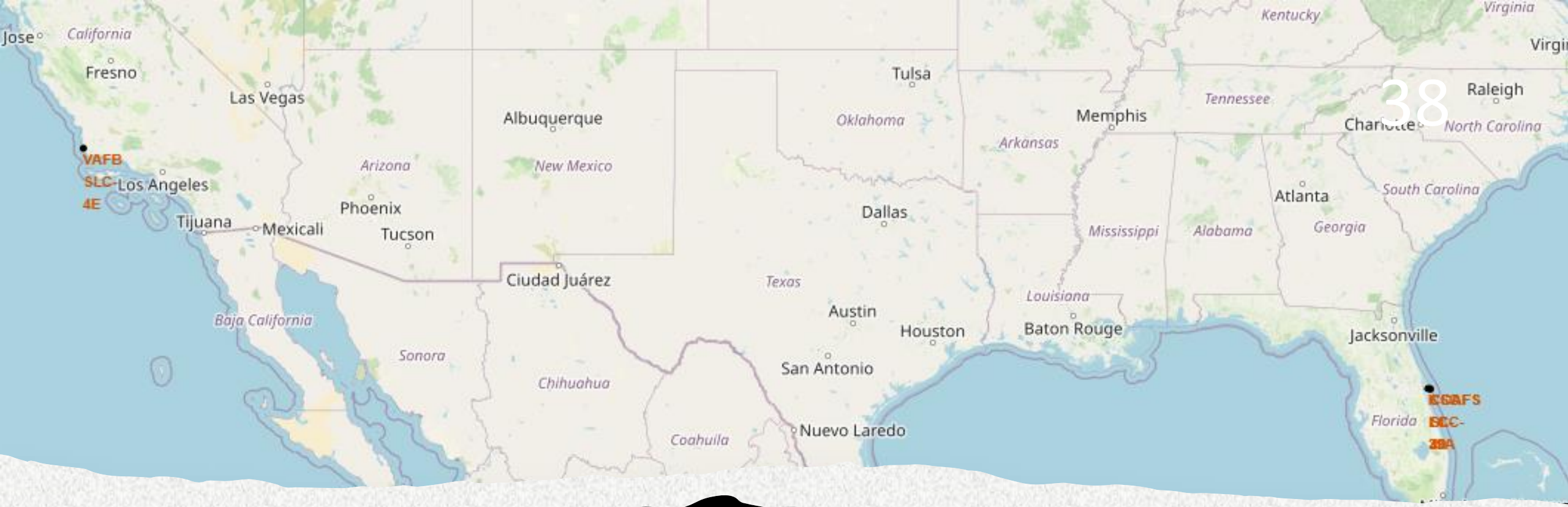
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- The success rate of landing outcomes increased as time went on. Failed attempts were much more common in the earlier years. Also, the variety of landings increased as time went on as well, such as drone ship, ground pad, and oceanic landings.

Date	Landing_Outcome	Count
2016-04-08	Success (drone ship)	5
2015-12-22	Success (ground pad)	3
2015-06-28	Precluded (drone ship)	1
2015-01-10	Failure (drone ship)	5
2014-04-18	Controlled (ocean)	3
2013-09-29	Uncontrolled (ocean)	2
2012-05-22	No attempt	10
2010-06-04	Failure (parachute)	2

Section 3

Launch Sites Proximities Analysis

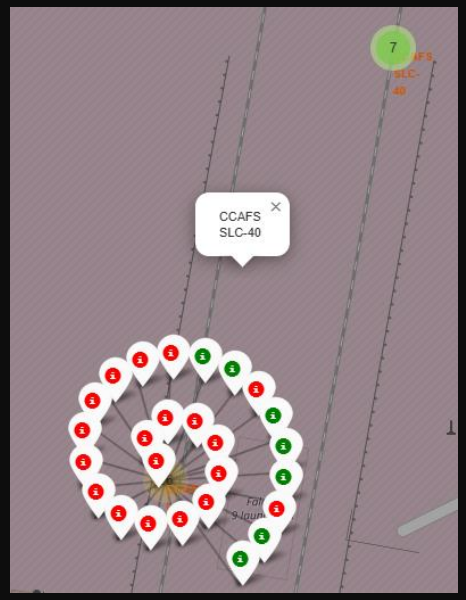


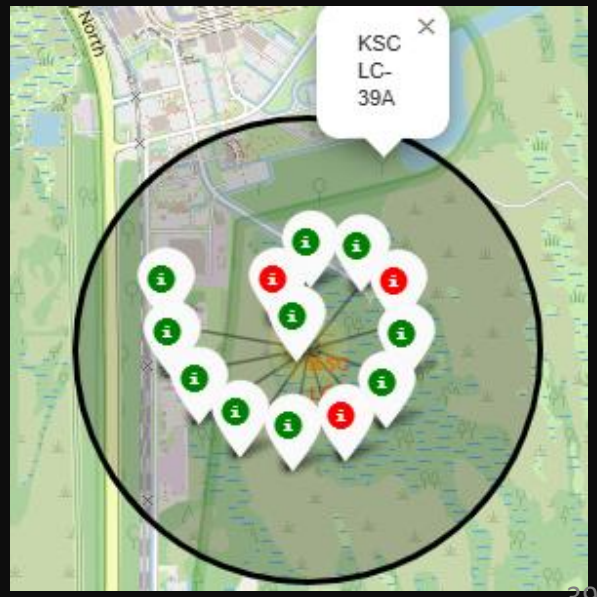
Launch Site Locations on Map

- The launch sites all tend to be near their respective coastlines, but do not seem to have any significant relationship with the equator. However, all sites are on the southern half of the US.

Launch Outcomes by Site

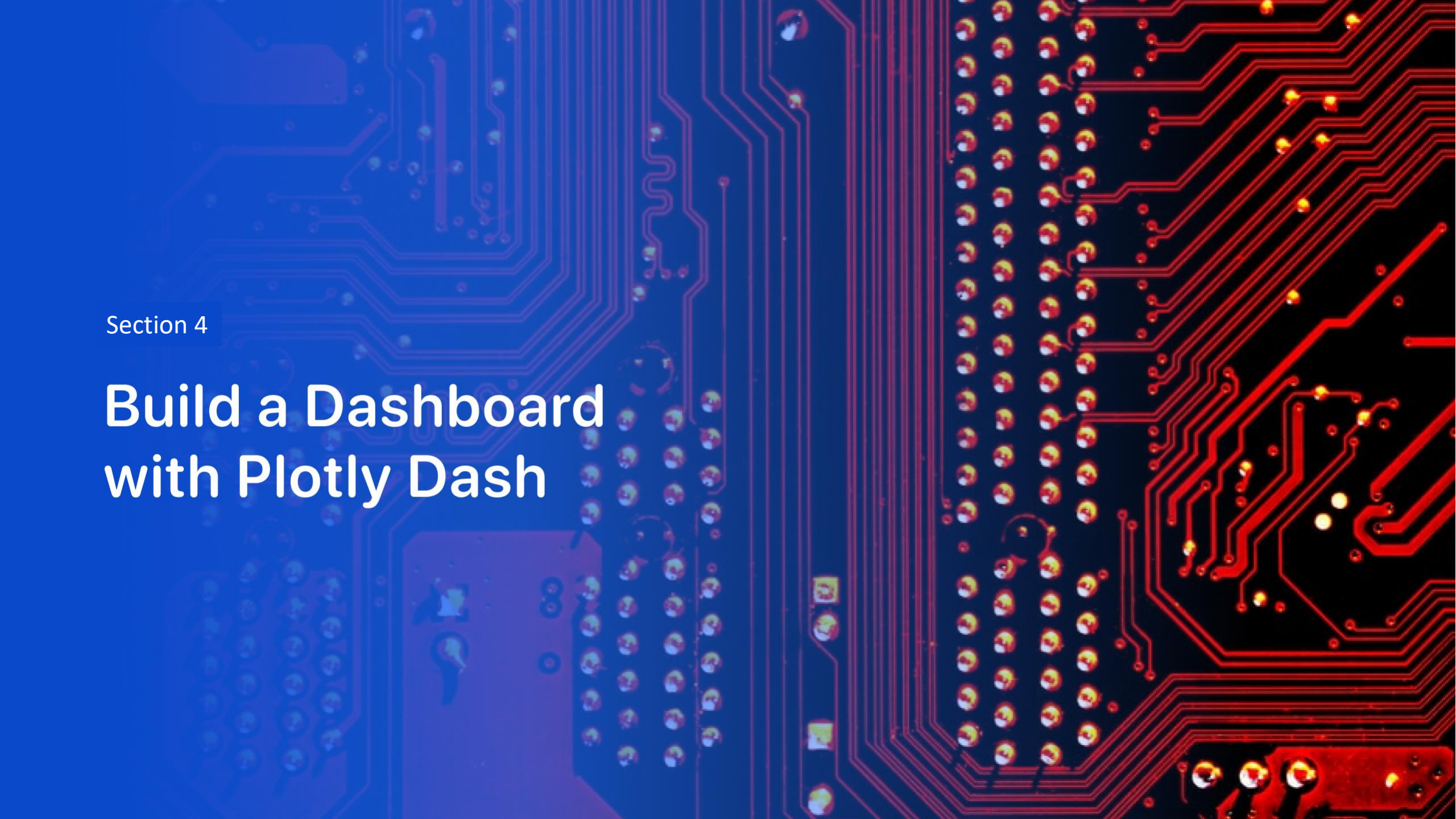
- CCAFS SLC-40 site tended to have more unsuccessful outcomes than other sites while KSC LC-39A had the most successful outcomes.

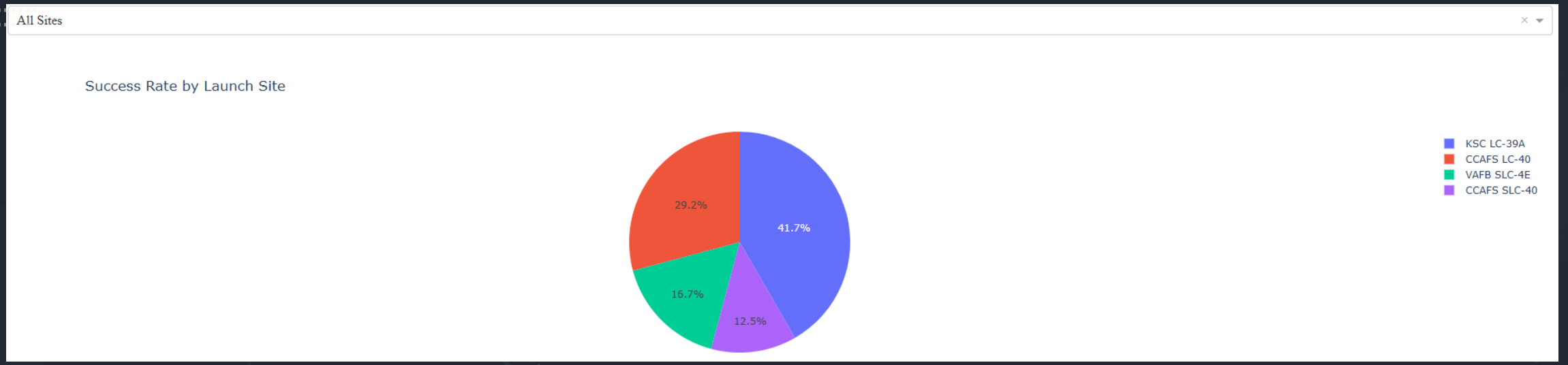




Distance of Launch Site to Proximities

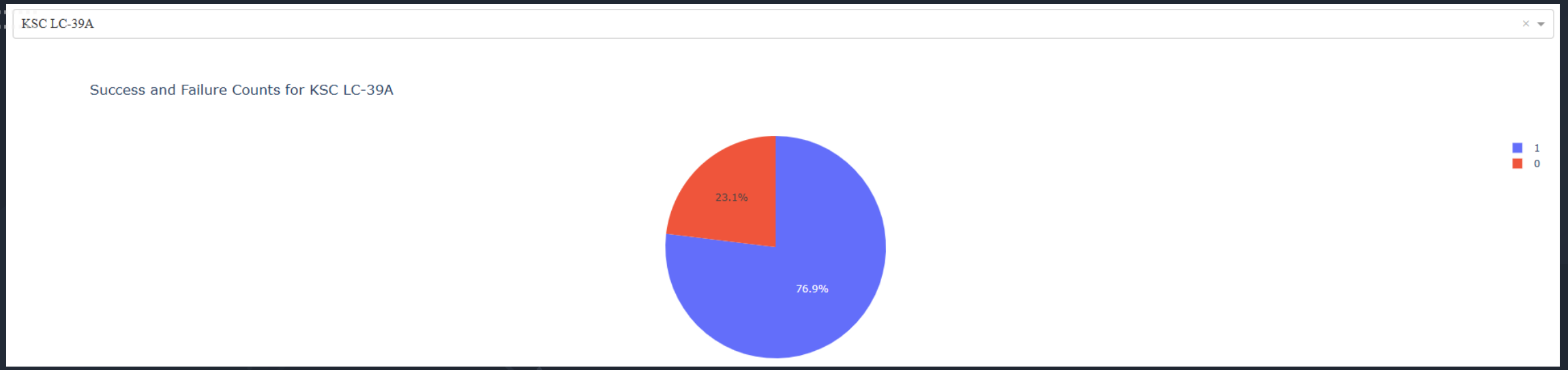
- Launch sites were typically less than a kilometer away from the nearest coastline while also being very close to some railways and highways (likely due to ease of transporting materials and equipment). However, sites were usually not close to any major cities.


Section 4
Build a Dashboard with Plotly Dash



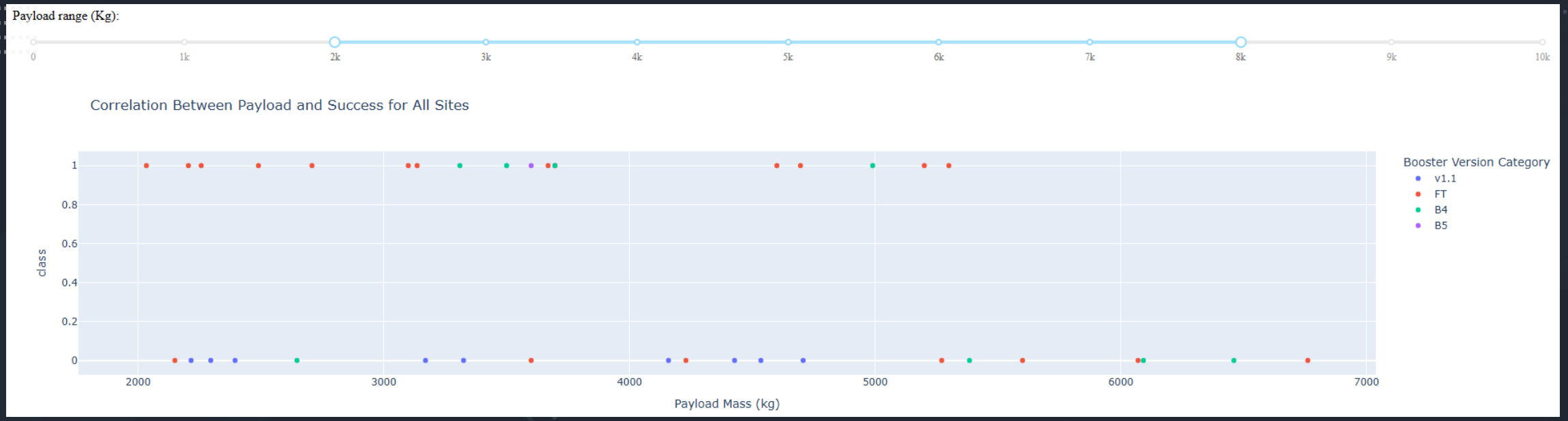
Launch Success Count for All Sites

- KSC LC-39A and CCAFS LC-40 both had a success count of 41.7% (CCAFS LC-40 and CCAFS SLC-40 both refer to same launch pad)



Highest Launch Site Success Ratio

- KSC LC-39A had the best success ratio (76.9 : 23.1) with ten successful launches and three failed launches. Additionally, no site had more failed launches than successful ones.

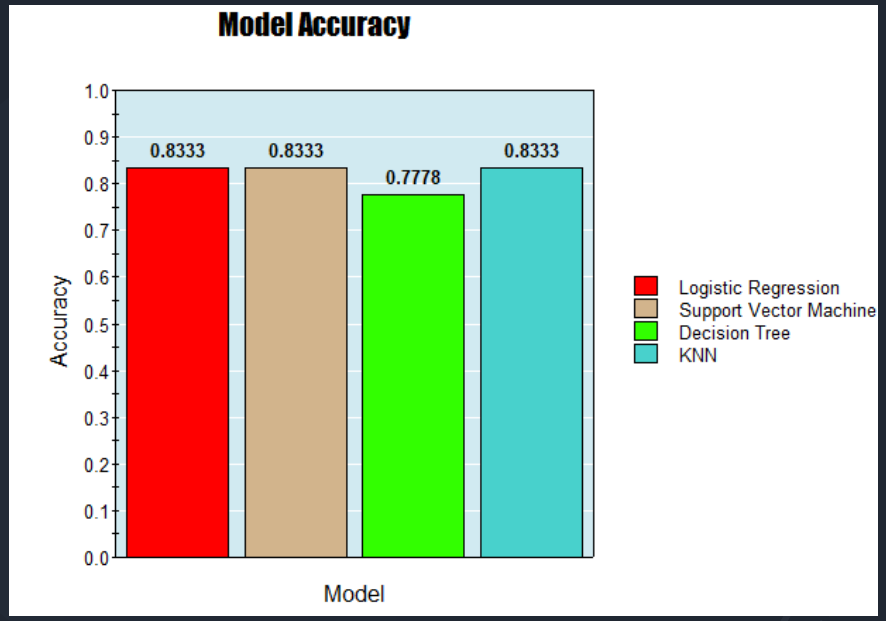


Payload vs Launch Outcome

- Payloads below 5500kg tend to have higher success rates than heavier payloads. Furthermore, newer booster versions (FT, B4, and B5) have higher success rates than older versions (v1.0 and v1.1).

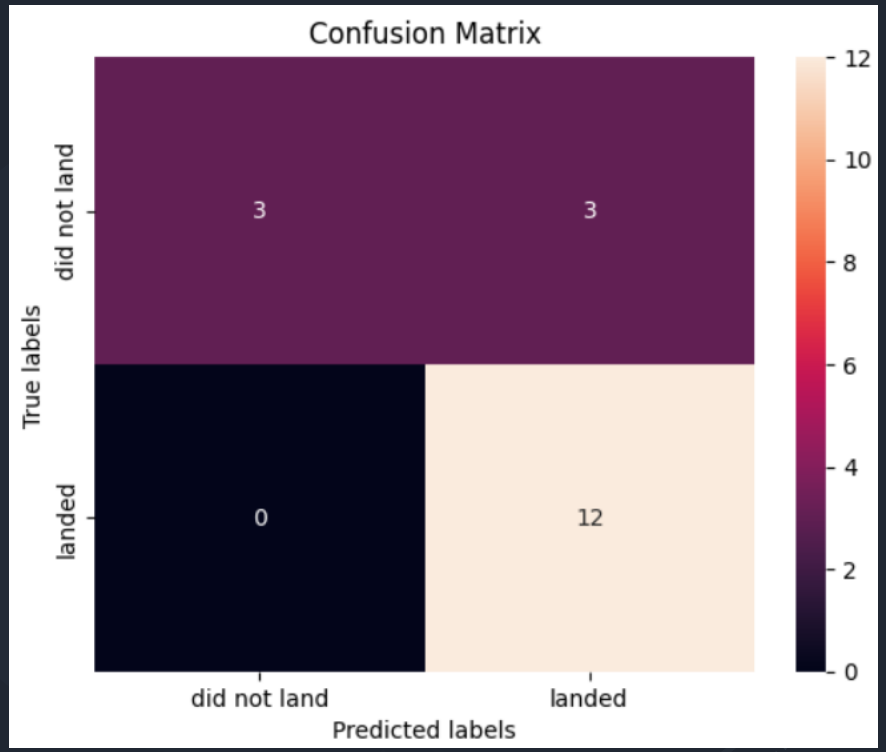
Section 5

Predictive Analysis (Classification)



Classification Accuracy

- Logistic Regression, SVMs, and KNN all had the same accuracy in my testing. However, due to Logistic Regression's speed and simplicity compared to other models, it is the best overall.



Confusion Matrix

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. The model appears to have some issues with false positives.

- True Positive - 12 (True label is landed, Predicted label is also landed)
- False Positive - 3 (True label is not landed, Predicted label is landed)

Conclusions

Findings:

Flight Number and Payload Mass strongly influence landing success.

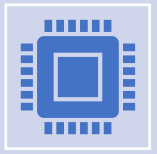
Certain launch sites and orbit types have higher success rates.

Logistic Regression for the most part outperformed other models with an accuracy of 83.33%.

Implications and Limitations:

Findings could improve prediction of landing success rates for different payload and orbit combinations.

Limited to Falcon 9 data as of June 2021, using more recent data and other rocket models could improve results

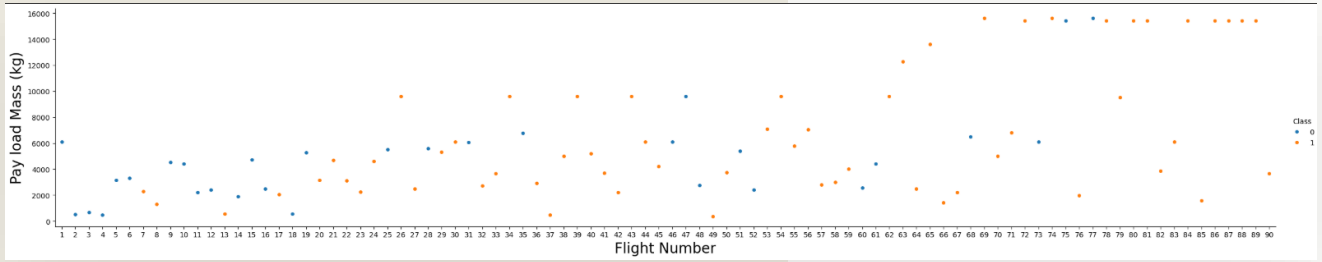


By leveraging data and machine learning, we have taken a step toward better understanding the key drivers of success for SpaceX missions. This approach can be expanded to unlock even greater insights in the future.

Appendix

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None
5	8	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None
6	10	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None
7	11	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None
8	12	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None

- Preview of the initial data:



- Relationship between Flight Number and Payload Mass:

Thank you!