IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Christopher Ferguson
July 16, 2025

Outline

Executive Summary

- **The goal of this project is to predict the likelihood of SpaceX successfully landing in order to provide a competitive advantage to our new company: SpaceY. The analysis that follows aims to acquire data from existing SpaceX data resources, identify key trends and variables that contribute to landing success, and identify a classifier model that can accuractely predict when a rocket will be reused.**
- **Summary of methodologies**
  - Data on SpaceX Falcon 9 launches was extracted via the SpaceX API and webscraping from Wikipedia and then wrangled using Pandas.
  - Exploratory data analysis was performed using graphing techniques with the Seaborn package as well as SQL queries to understand the data structure and variables that influence the outcome.
  - A map of relevant launch sites and success rates was generated using Folium and a dashboard was established using Plotly Dash to assess the relationship between launch site, payload, and success rate.
  - Finally, multiple models were trained to determine how accurately we can predict the likelihood of a successful Falcon 9 rocket launch.
- **Summary of all results**
  - EDA demonstrated that there were clear relationships between Launch Number, Site, Payload, and Orbit type with success rate – lending support to the hypothesis that a model can predict success.
  - Map analysis demonstrated that launch sites are established near common access points like highways, coastlines, and railways.
  - PLotly dashboard exploration identified numerous relationships between payload, booster version, launch site, and success rate.
  - Multiple machine learning models were able to accurate predict success rate with about 83% accuracy, with false positives being the main weak point.

Introduction

- The following slides describe the efforts of SpaceY to use data science and machine learning understand the drivers of rocket reuse success by SpaceX in order to out compete them.
- Key problems of interest to this study include:
  - Are any key rocket features associated with likelihood of successful rocket resuse, such as booster version, payload, launch date, etc.?
  - Are there aspects of rocket launch sites that may influence the likelihood of a successful launch?
  - What model best predicts whether or not a launch will be successful?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected through use of the SpaceX Data API and webscraping from the Falcon 9 Wikipedia page
- Perform data wrangling
  - Data was loaded from CSV and read into a Pandas DataFrame,
  - Quality checks were done to confirm appropriate data types in each column, absence of null values, and assess frequency of different launch properties
  - Text-based Outcomes data was then converted into a binary Landing Outcomes column
- 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 was split into train and test sets and then multiple models were trains using GridSearch and assessed for predictive accuracy

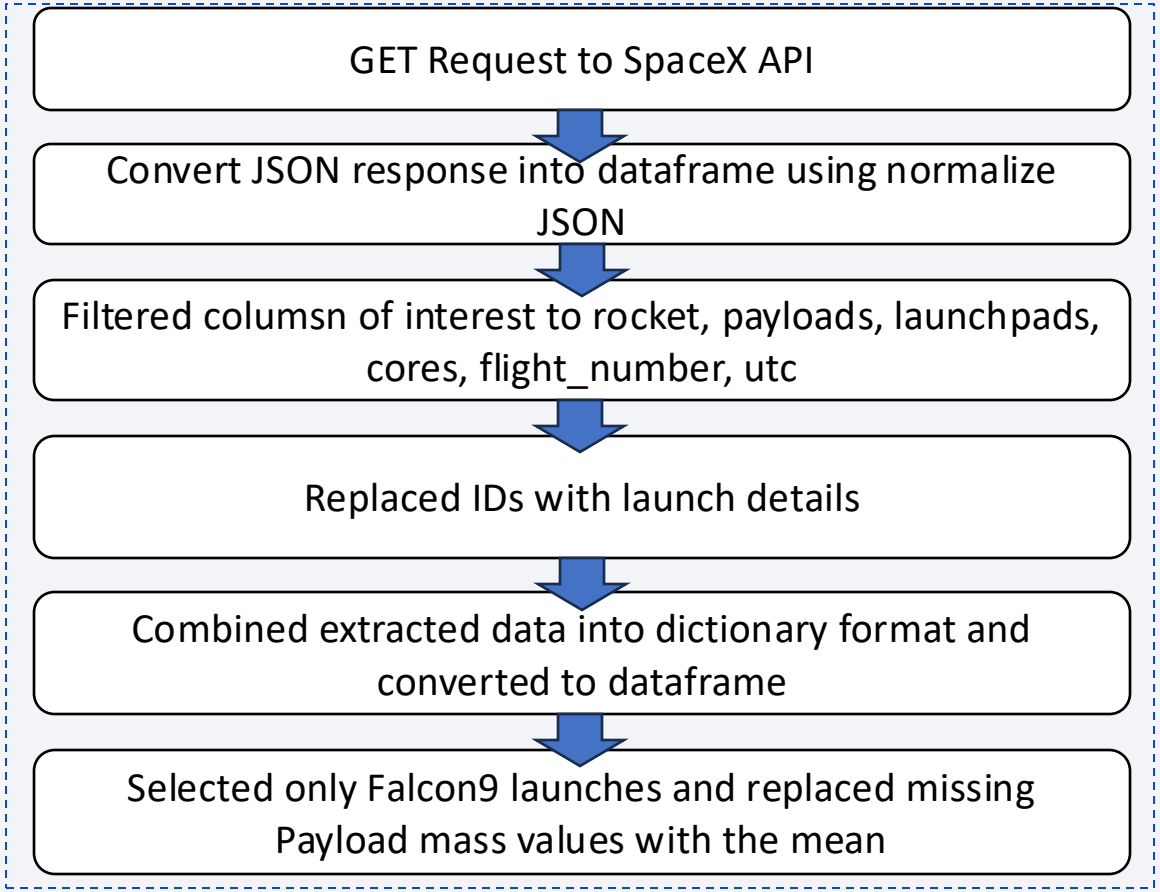
Data Collection

- Datasets were collected via two methods:
  - First, by using GET requests from the SpaceX data API for past launches: https://api.spacexdata.com/v4/launches/past
  - Second, by webscraping table data off of the Falcon9 Wikipedia page: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Both methods yielded slightly different information associated with individual Falcon9 rocket launches and are applied to later aspects of the analysis and model building process.

Data Collection – SpaceX API

- Key data values for Falcon 9 rocket launches including rocket identifier, payload mass, and launch pad identity, were collected using the methods outlined to the right from the SpaceX API at the following link: https://api.spacexdata.com/v4/launches/past

- The Jupyter Notebook can be found at: https://github.com/cfergu11/IBMDataScienceCapstop/blob/main/jupyter-labs-spacex-data-collection-api.ipynb
- The exported CSV can be found at: https://github.com/cfergu11/IBMDataScienceCapstop/blob/main/dataset_part1.csv

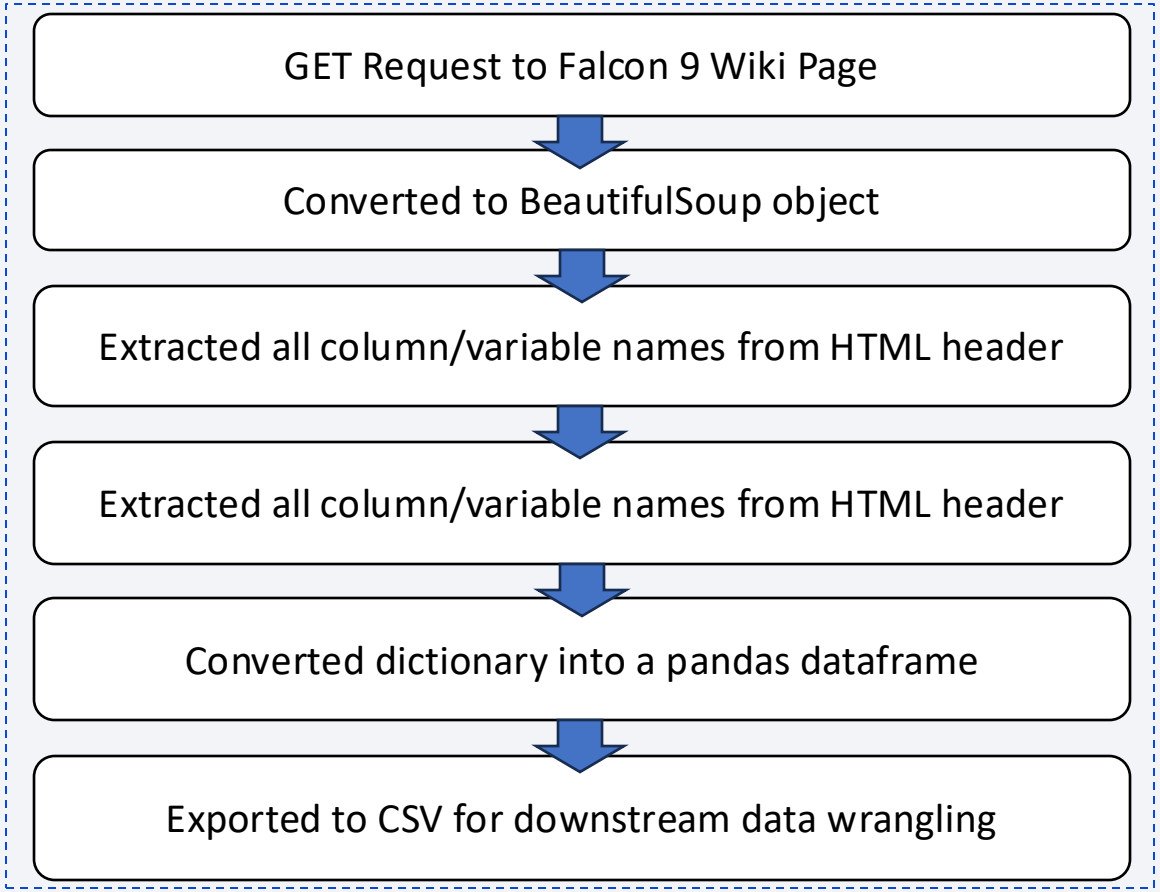


Data Collection - Scraping

- The secondform of data collection involved scraping data from tables included in the Falcon 9 wikipedia page located here: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"

- The Jupyter Notebook can be found at: https://github.com/cfergu11/IBMDataScienceCapstop/blob/main/jupyter-labs-webscraping.ipynb

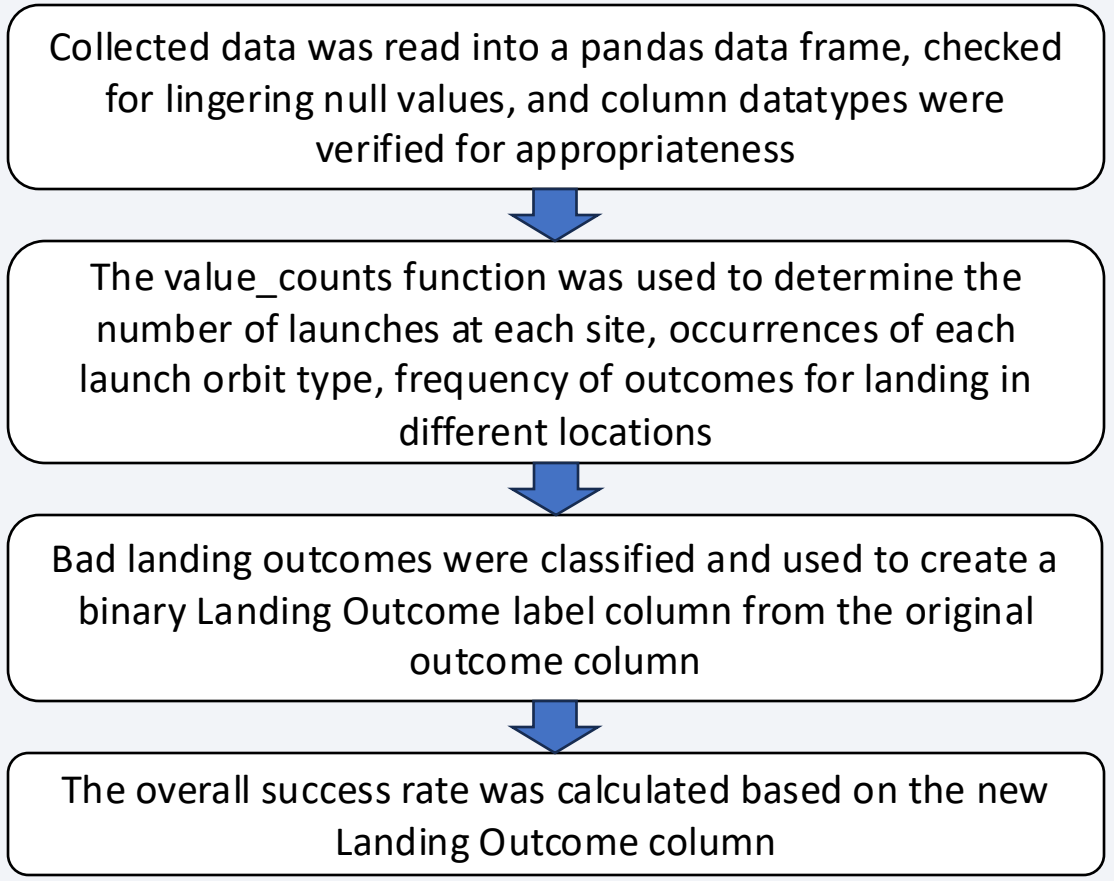
- The exported CSV can be found at: https://github.com/cfergu11/IBMDataScienceCapstop/blob/main/spacex_web_scraped.csv



Data Wrangling

- Once data was collected, it was read into a Pandas dataframe for additional processing and quality control as descibed in the flow chart to the right:

- The Jupyter Notebook can be found at: https://github.com/cfergu11/IBMDataScienceCapstop/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb
- The exported CSV can be found at: https://github.com/cfergu11/IBMDataScienceCapstop/blob/main/dataset_part_2.csv



EDA with Data Visualization

- Exploray data analysis was performed using a combination of pandas, matplotlib and seaborn packages.
- A combination of scatter plots, bar charts, and lineplots were used to identify potential pattersn in the data and potential variables of relevance to the target outcome of successful landings.
- Based on these visualizations, a subset of important variables were downselected for further analysis and model building.
- Details of individual visualizations and insights can be found in the results section, and the associated Jupyter Notebook can be found here: https://github.com/cfergu11/IBMDataScienceCapstop/blob/main/edadataviz.ipynb
- The resulting CSV file can be found here: https://github.com/cfergu11/IBMDataScienceCapstop/blob/main/dataset_part_3.csv

EDA with SQL

- As part of exploratory data analysis (EDA) collected data was loaded into a SQL data base using SQLite3 and queried in Python using sqalchemy
- EDA explored aspects of site naming, payload mass volumes and distribution, launch success and failure characteristics, and booster version influence.
- A final query assessed the ranked frequency of landing outcomes across a range of dates.
- The full list of queries and results can be found at the GitHub Link here: https://github.com/cfergu11/IBMDataScienceCapstop/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

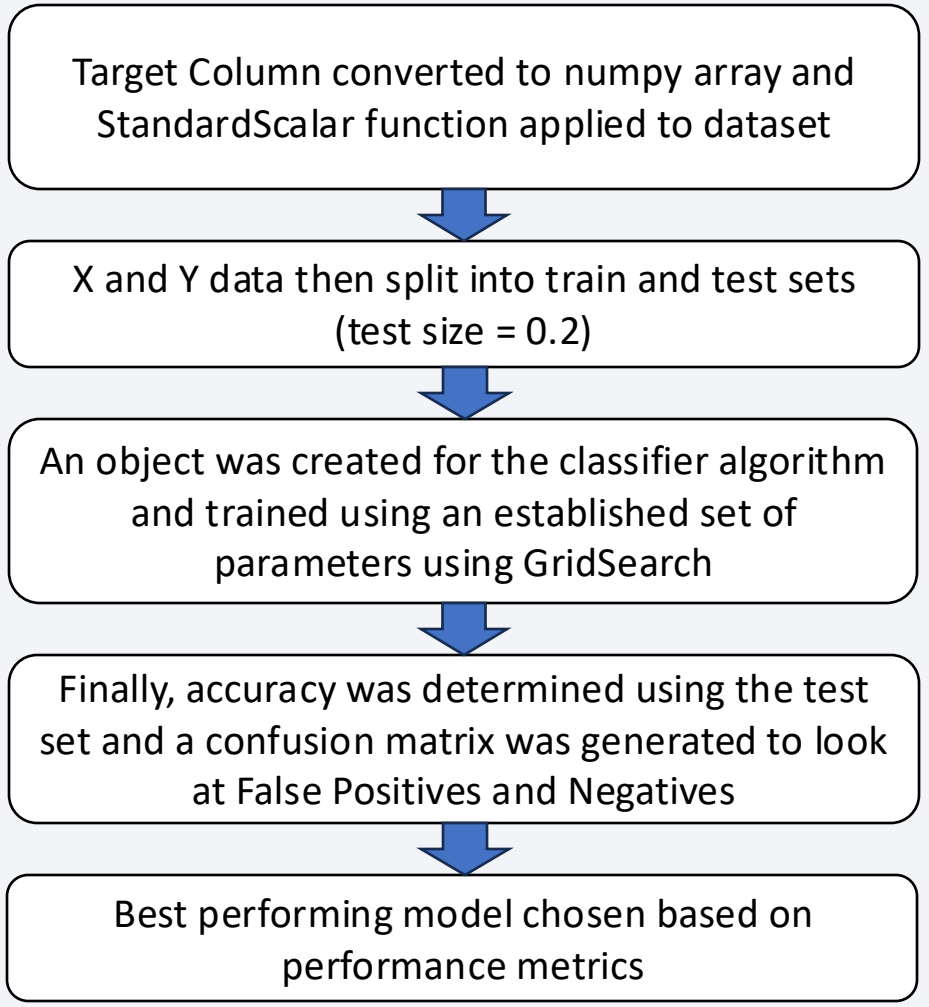
- Circle markers were used to indicate the 4 main launch sites described in the analysis. This allows us to see where these sites are with respect to each other, nearby landmarks, and the rest of the world.
- Next, for each site I created a marker cluster object and added colored markers to each site that correspond with an individual launch attempt as well as landing outcome indicated by color – Green = success, red =failure.
- Polyline markers and distance measurements were added between the VAFB site and key proximal entities including the coastline, railway, highway, and city to identify any potential trends that could be explained by launch requirements.
- The Jupyter Notebook can be located at the GitHub link here: https://github.com/cfergu11/IBMDataScienceCapstop/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The dashboard was designed to include the following:
  - A Dropdown menu that allows the user to select from either "All Launch Sites" or one of the four individual launch sites. This allows users to get a broad overview of Launch characteristics across all sites and then drill down into individual sites of interest to better understand things like success rate, payload, and Booster Use.
  - A pie chart that for comparison of overall success rates between sites or Launch Outcomesfor an individual site.
  - A slider that allows users to select a range of Payload Mass to understand the potentialimpact on success rate.
  - A scatter plot of success rate (class) against payload mass (kg) colored by Booster Version, allows users to drill down into the relationship between payload mass and booster version on class across all sites for individual sites selected with the dropdown.
- The full code for the Plotly Dash script can be found on GitHub here: https://github.com/cfergu11/IBMDataScienceCapstop/blob/main/spacex-dash-app.py

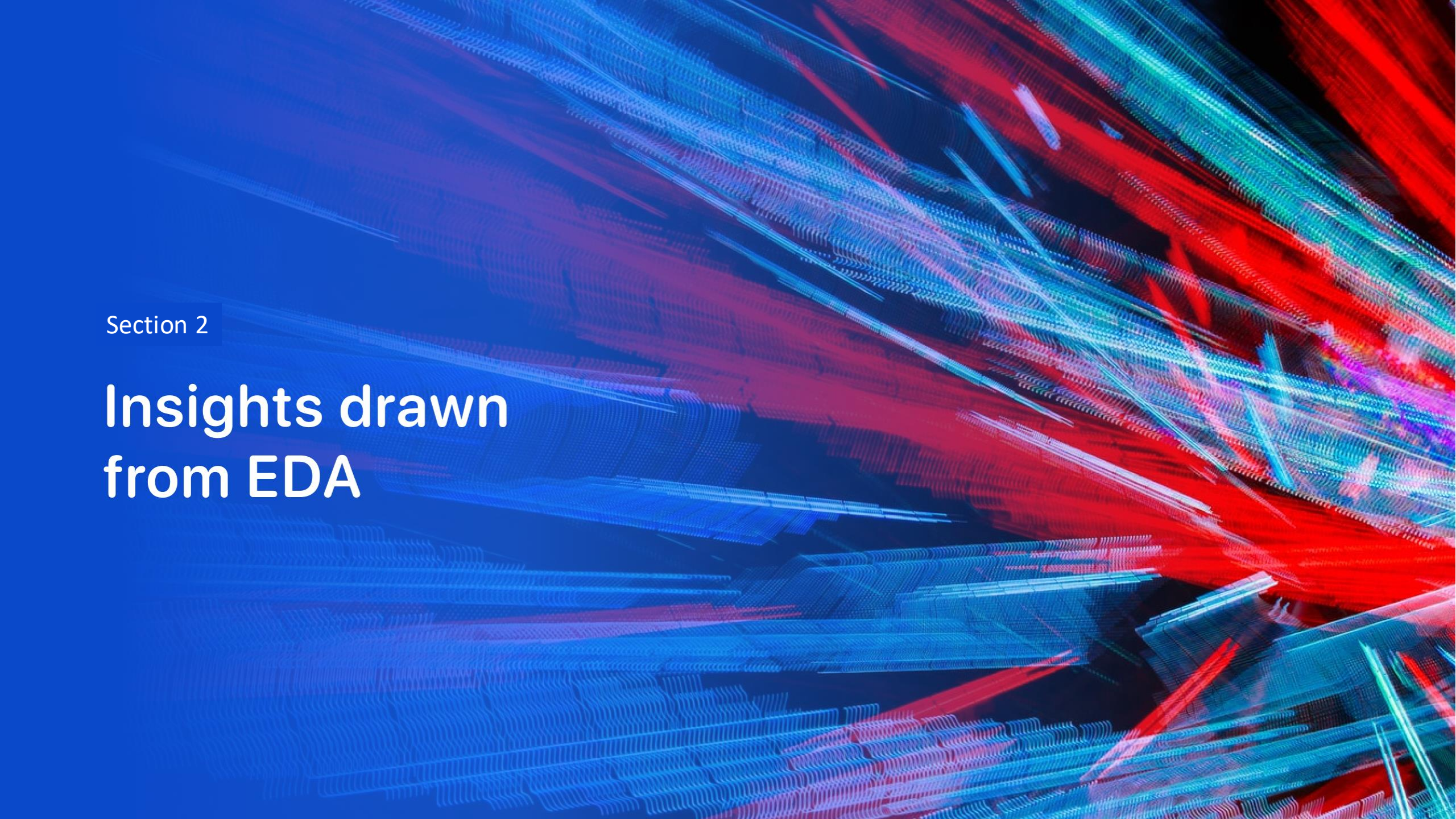
Predictive Analysis (Classification)

- Classification accuracy was tested for four different models, including logistic regression, support vector machine, decision tree classifier, and K-nearest neighbors based on the flow chart to the right.
- Performance was then compared between all four to identify the best performing model.
- The full Jupyter Notebook can be found at the following GitHub link: https://github.com/cfergu11/IBMDataScienceCapstop/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



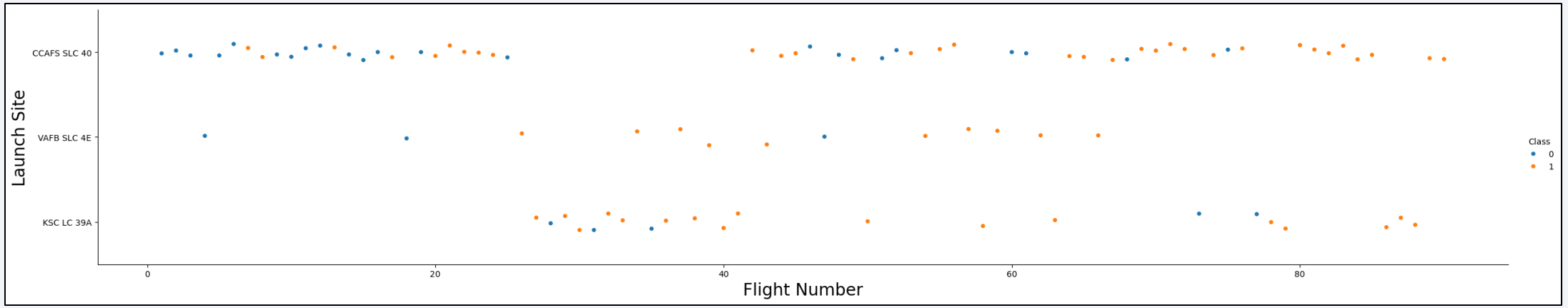
Results

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


Section 2
Insights drawn from EDA

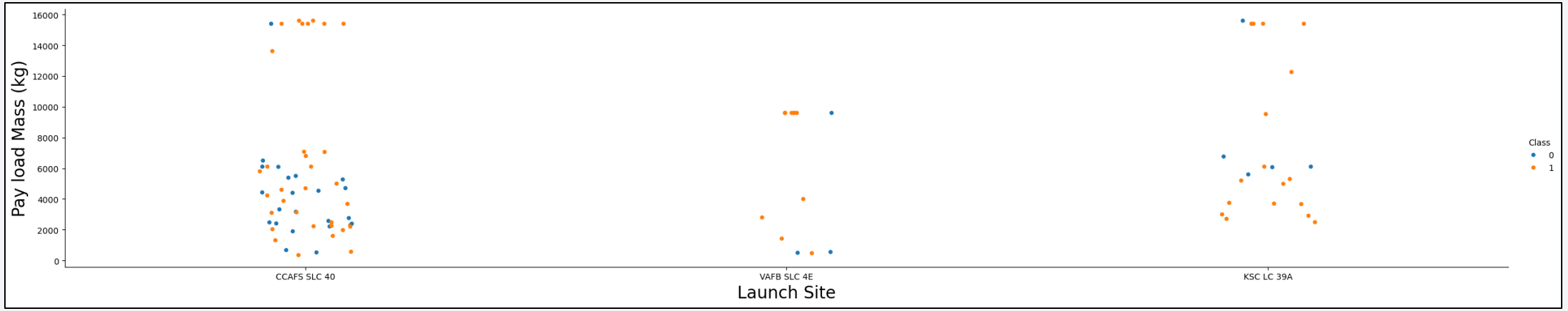
Flight Number vs. Launch Site

- Looking at trends associated with Launch Site and Flight Number on success rate using a scatterplot yielding the following observations:
  - Failures, indicated in blue, appear to be more common in earlier launches, with a higher rate of success in later launches across all three sites. This would made sense as more is learned from early failures, and success would be expected to improve over time..
  - Additionally, The majority of the launches came through CCAFS SLC 40 and the fewest launches came from VAFB SLC 4E, potentially related to resources at each of those sites.



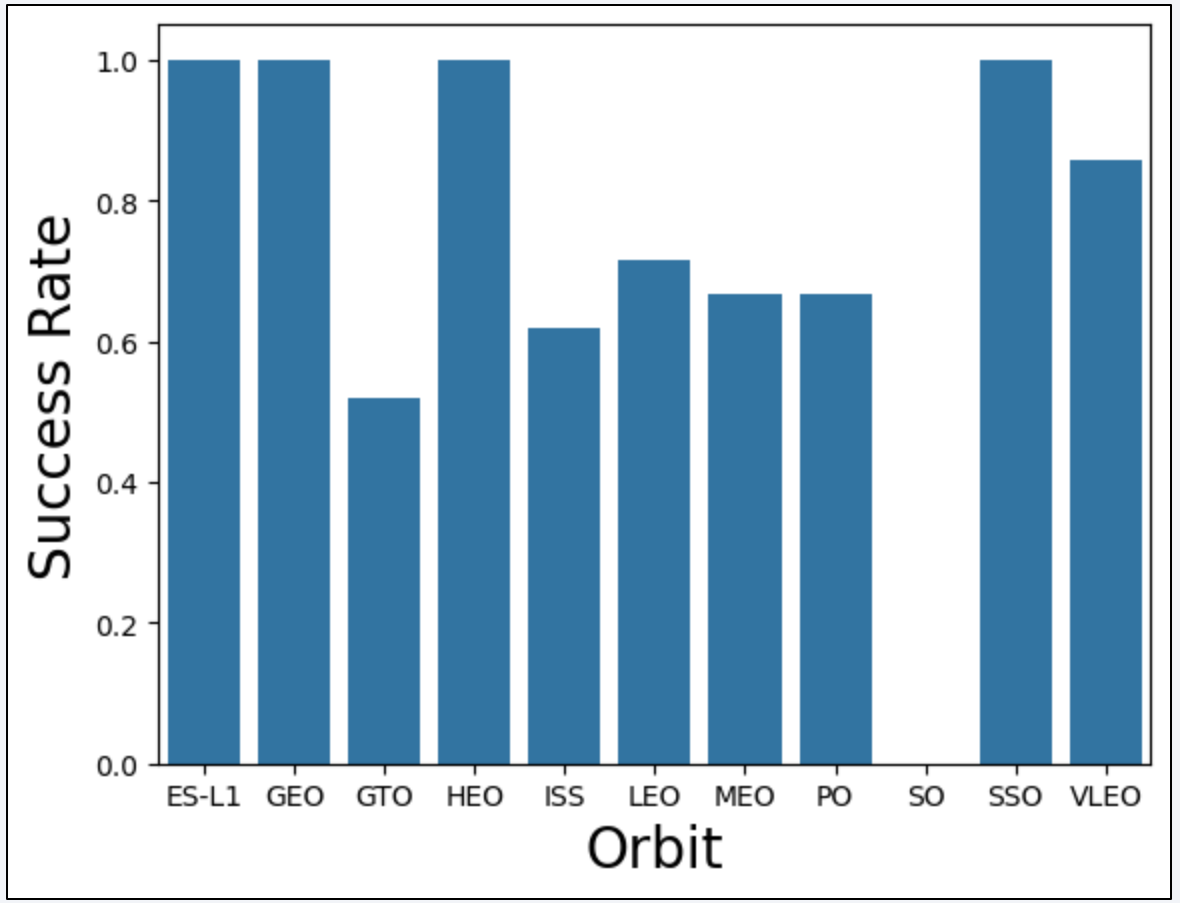
Payload vs. Launch Site

- Using scatterplots to look at association between site, payload mass, and success rate indicated the following:
  - For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000)
  - The highest payloads had the fewest failures, perhaps indicating a higher loss associated with failure
  - Additionally, there is a big jump between the low Pay Load Masses and the maximum, coupled with the previous observation this suggests low payloads may be associated with earlier testing



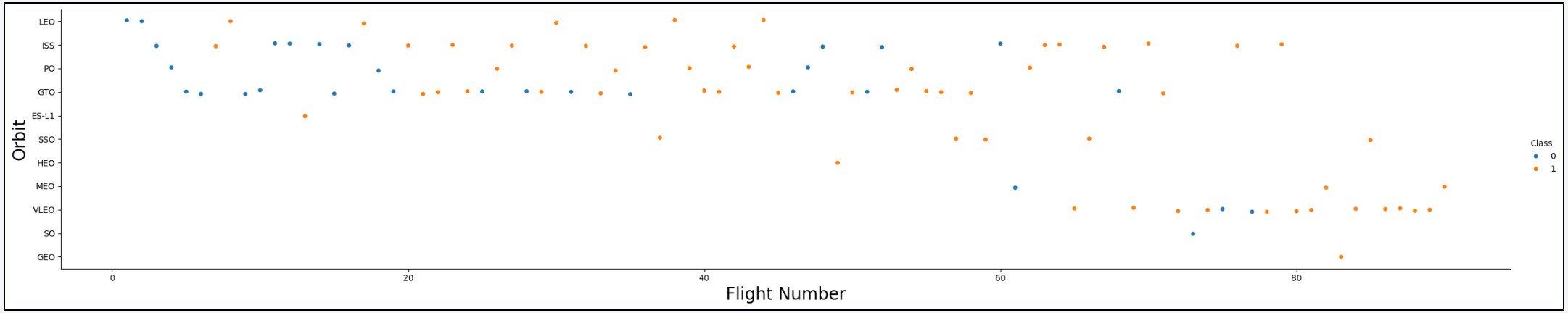
Success Rate vs. Orbit Type

- The barchart comparing success rate across multiple launch orbits shows that:
  - ES-L1, GEO, HEO, and SSO orbits all have the highest success rate at 100%.
  - Most other orbits fall between 50 and 70% success except for the SO orbit, with no successful missions



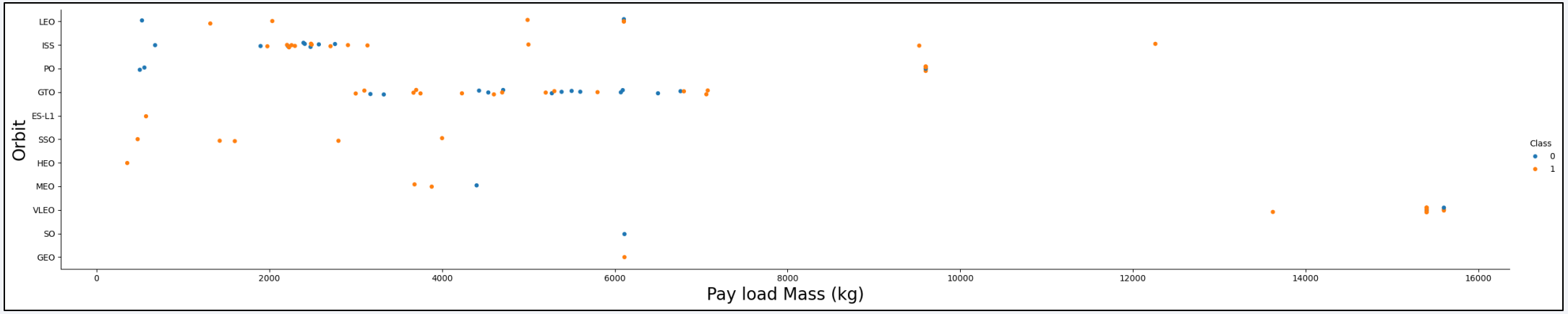
Flight Number vs. Orbit Type

- A comparison of Flight Number to Orbit Category shaded by outcome shows that:
  - With 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.
  - Finally, the later flight numbers see a large number of VLEO orbits, relative to earlier flights clustered from LEO to GTO orbits



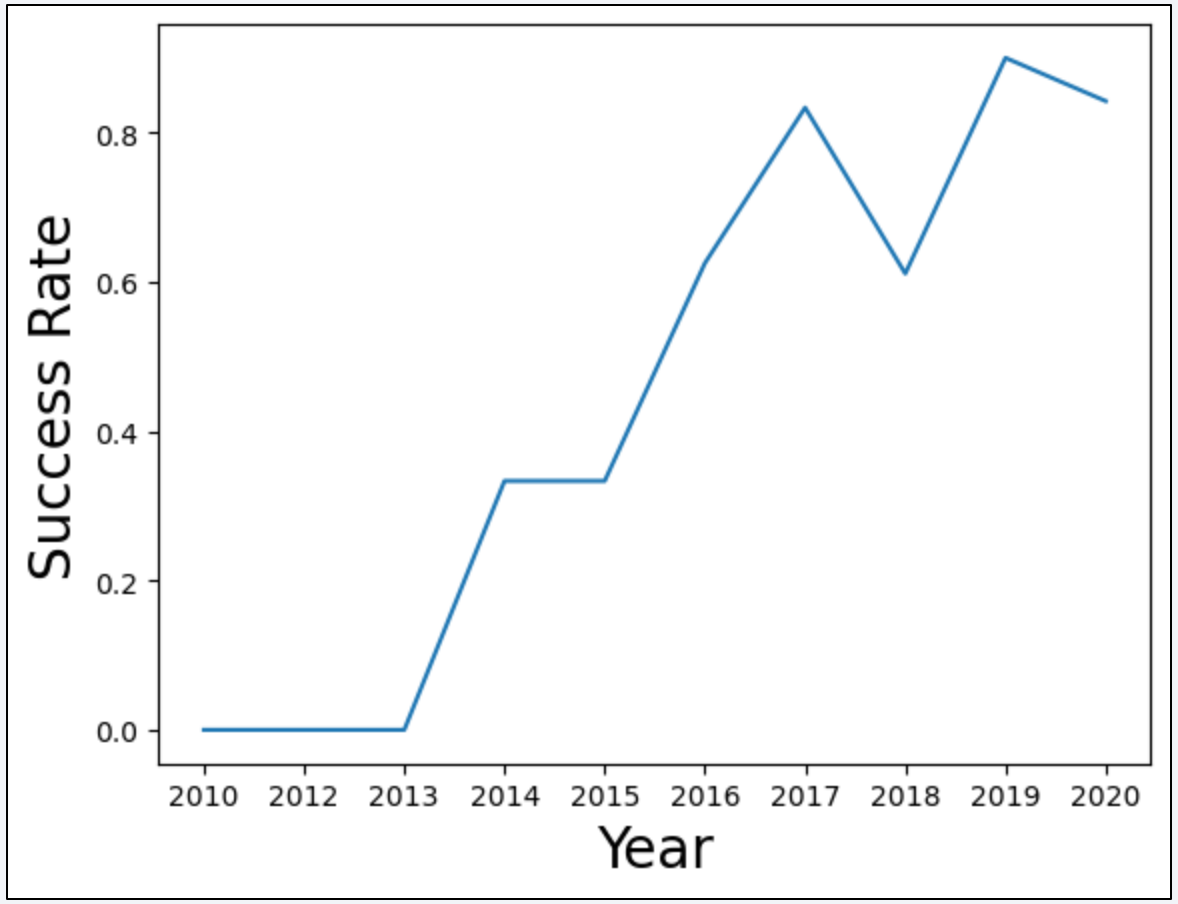
Payload vs. Orbit Type

- Comparing Pay load mass and Orbit category we can see that:
  - With heavy payloads the successful landing or positive landing rate is relatively high for Polar, LEO and ISS, compared to lower payloads in other orbits
  - However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present in relatively high abundance across the full range of payloads



Launch Success Yearly Trend

- Looking at a line plot for success rate as a function of launch year, we can see that success rate steadily increases after 2013to around 80-90% in 2019 and 2020
  - This aligns with a previous observation that later launches by number tended to have higher success rates



All Launch Site Names

- After doing some initial visualization of any relationships between variables and outcome, further EDA was performed using SQL analysis of the data.
- Here we can see that there are four unique sites across all the launches, which correlates with our findings from the visualization analysis.

```
%sql select distinct(Launch_Site) from SPACEXTABLE
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

- A query subsetting "CCA" sites demonstrates that there is a wide variety of Booster Versions, Payloads, and Landing Outcomes across the same launch site

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

```
 * sqlite:///my_data1.db
Done.
```

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

- The following query demonstrates that boosters launched by NASA (CRS) carried nearly 46,000 KG overall, which is significant compared to the number of individual launches that were less than 10,000 KG

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer IS 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- A SQL query to calculate the average payload mass carried by booster version F9 v1.1 yielded an average of just under 3,000 KG, supporting that on average most the payloads tend to be below 10,000 KG

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version IS 'F9 v1.1'
--- note that this does not include all F9 v1.1 B... versions ---
```

```
 * sqlite:///my_data1.db
Done.
```

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The first launch with a success landing on a ground pad occurred on 12-22-2015, more than two years after the first set of launches in 2013. This is in line with the line plot in the slides above suggests there was a signficant lag time between initial launches and successful landings.

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

```
 * sqlite:///my_data1.db
Done.
```

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- There were only 4 boosters that successfully landed on a drone ship with payload masses between 4000 and 6000 KG
  - This suggests that significant booster tuning was necessary to achieve success and that it might be somewhat dependent on payload

```
[17]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome is 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
 * sqlite:///my_data1.db
Done.
```

[17]:

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

Total Number of Successful and Failure Mission Outcomes

- The following query demonstrates a total of 100 successful missions and 1 failed mission. This is clearly distinct from landing success rate, which saw significantly more failures than the mission outcome.
- It is also worth noting that there are multiple subcategories for Mission_Outcome and further analysis would need to be standardized.

```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) as 'Total Number' FROM SPACEXTABLE GROUP BY Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

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

Boosters Carried Maximum Payload

- The query below demonstrated that the maximum payload carried by any single booster is 15,600 KG and it was carried by 12 distinct boosters.

```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ IS (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

```
 * sqlite:///my_data1.db
Done.
```

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

2015 Launch Records

- There were two launches that failed to land on drone ships in 2015. Both launched from the CCAFS LC-40 site, one in January and one in April, and they were booster F9 v1.1 B1012 and F9 v1.1B1015, respectively.

```
%sql SELECT substr(Date,6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(Date,0,5)='2015' AND Landing_Outcome is 'Failure (drone
```

```
 * sqlite:///my_data1.db
Done.
```

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

- Between 2010-06-04 and 2017-03-20 there were a total of 31 launches with different outcomes. Of those, the most common landing outcome was "No Attempt" and the least prevalent was Precluded (drone ship).

```
%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUNT(Landing_Outco
 * sqlite:///my_data1.db
Done.
```

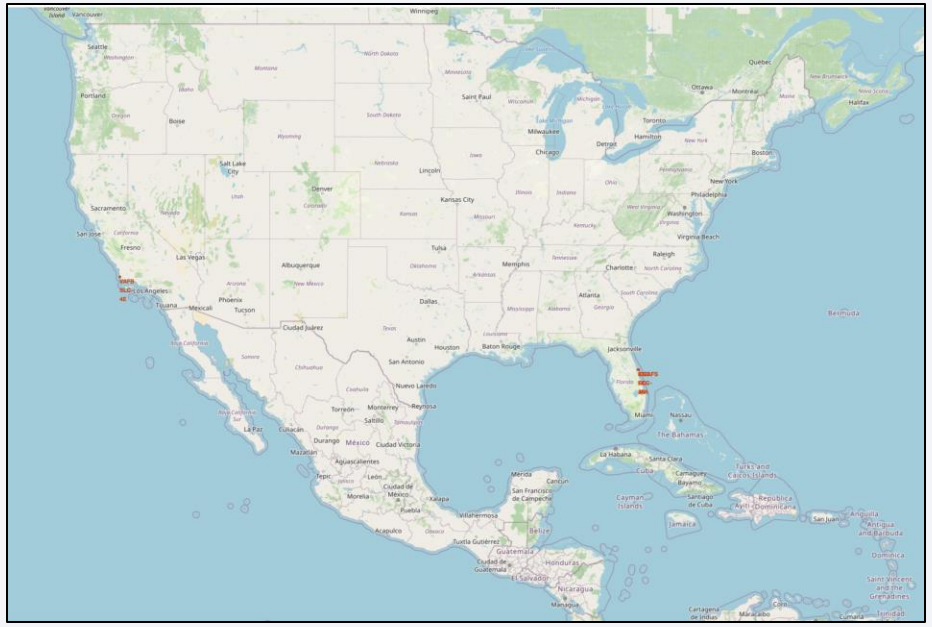
Landing_Outcome	COUNT(Landing_Outcome)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

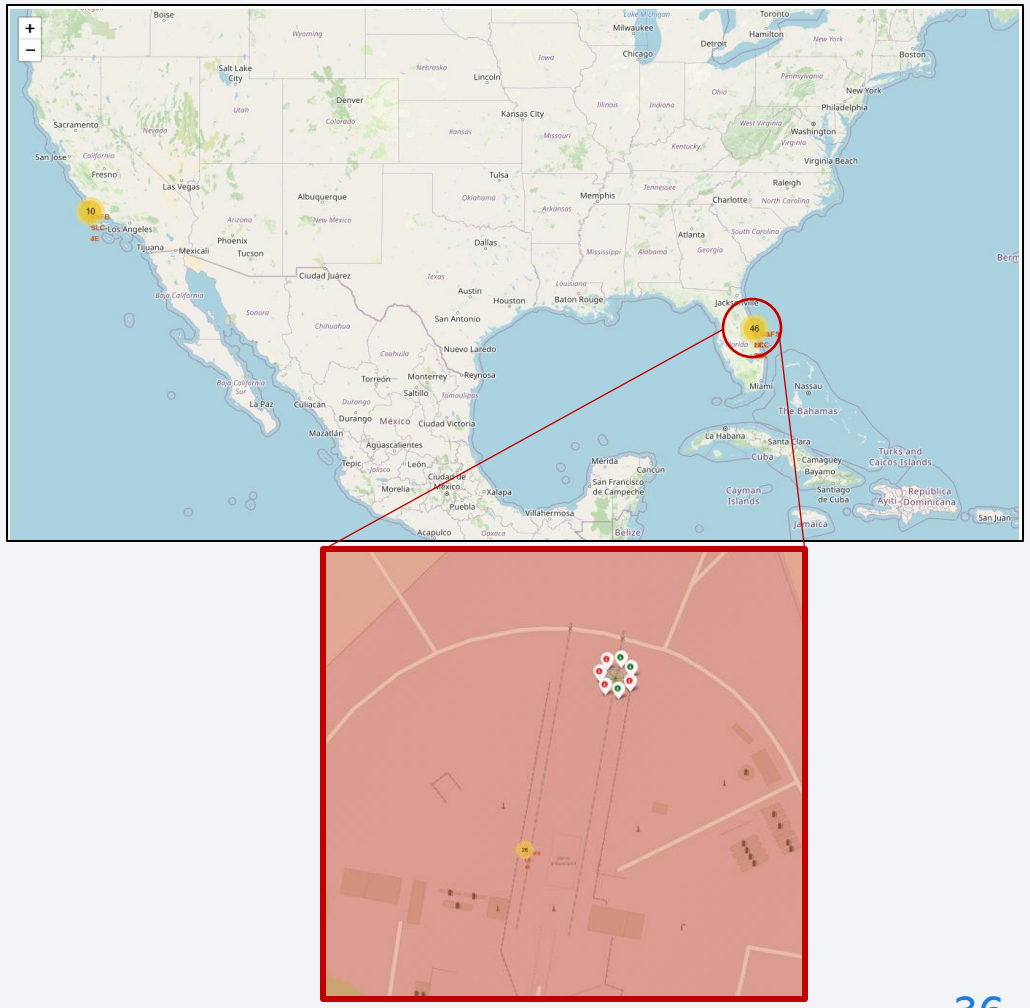
Global Map of Launch Site Locations

- The four sites are restricted to parts of North America, in line with SpaceX as a company.
- All 4 sites are located close to coastlines in the southern United States, with 1 in California and 3 in Florida, but none are particularly close to the equator.
- The clustering of 3/4 sites in Florida suggests there is a preference or increased density of mission-related resources in that area.



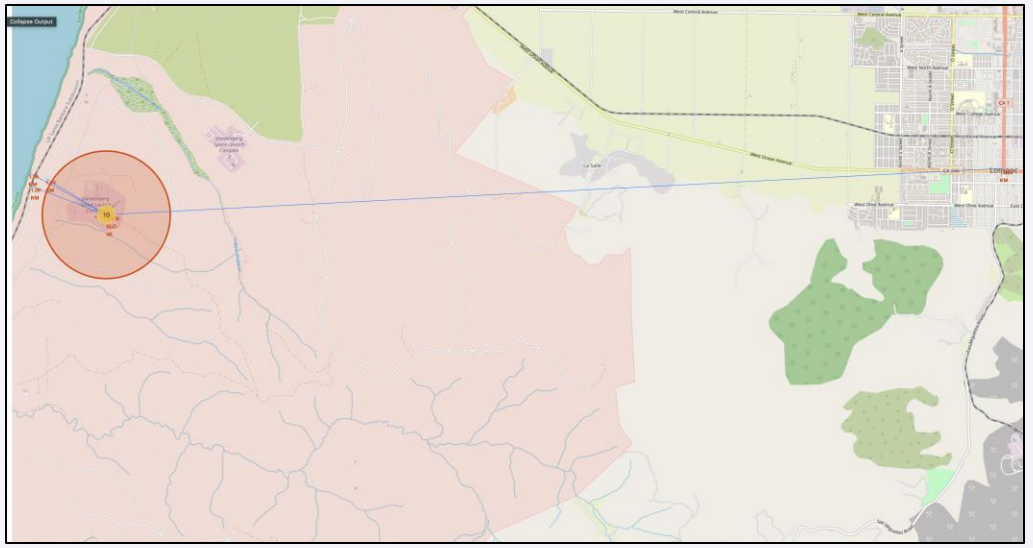
Global Map of Launches and Outcome by Site

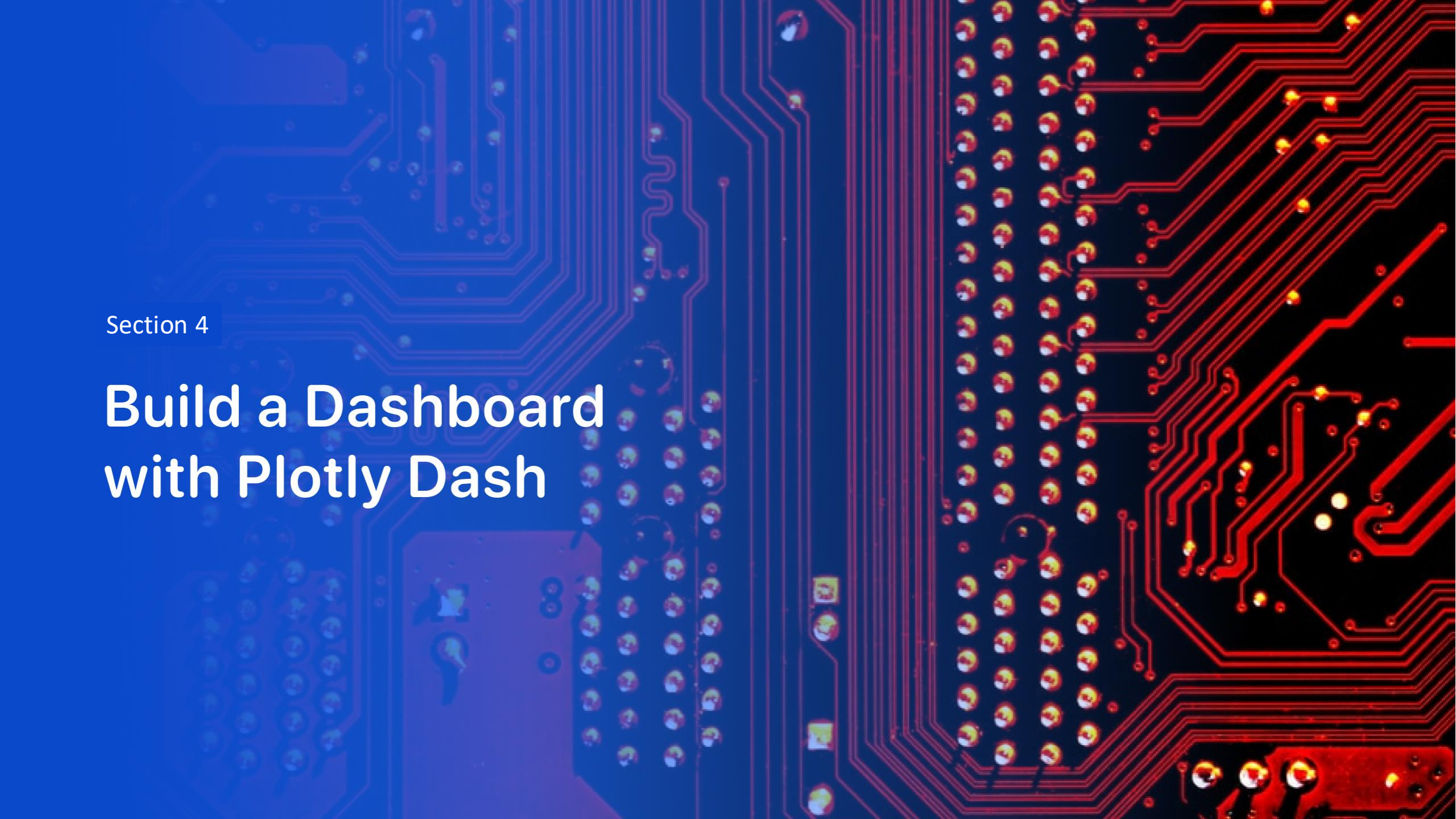
- The map to the right shows the number of launches at each of the 4 sites. Zooming in displays the outcome for each launch with a colored icon indicating success (green) or failure (red).
- CCASFS SLC-40 was less than 50% successful, but it was still more successful than CCASFS LC-40, which only had 3 successes out of ~26.
- KSC LC 39A was quite successful, with only 3 failures out of 13 attempts.
- VAFB SLC-4E had 4 failures out of 10 tries.



Distance from VAFB SLC-4E Launch Site and its Proximities

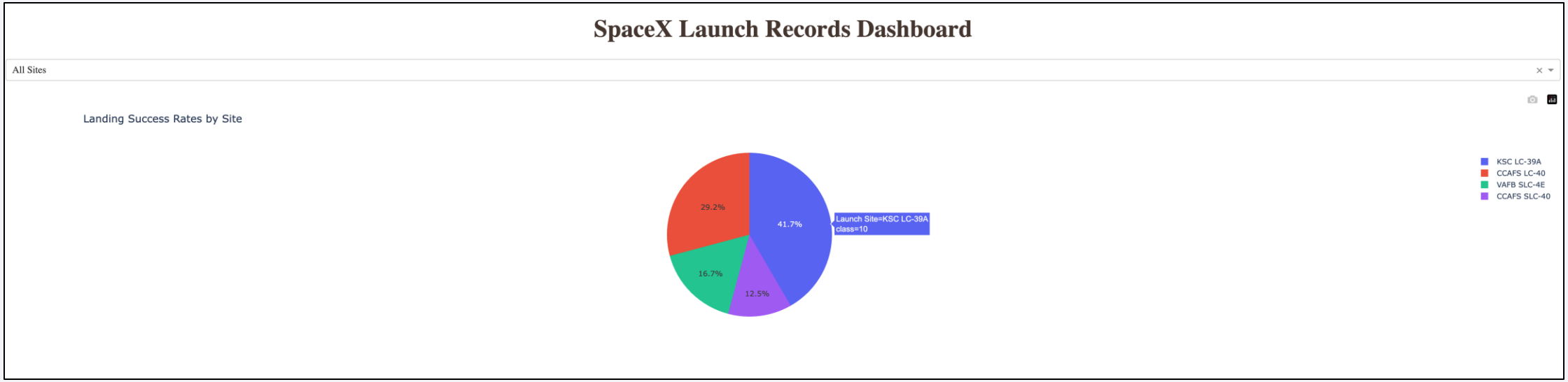
- The map to the right shows the VAFB SLC-4E site and its proximity to key points of interest:
  - The VAFB SLC 4E site is 1.38KM from the coast, 1.26KM from the nearestrailway, 1.10KM from the nearest highway, and 13.95KM from the closestcity.
- The close proximity to the city, highway, and railways makes sense for transporting the necessary materials and personnel to and around the launch site. However, the distance from the nearest city is likely for safety reasons, as a significant number of launches fail and could put nearby residents at risk otherwise.




Section 4
Build a Dashboard with Plotly Dash

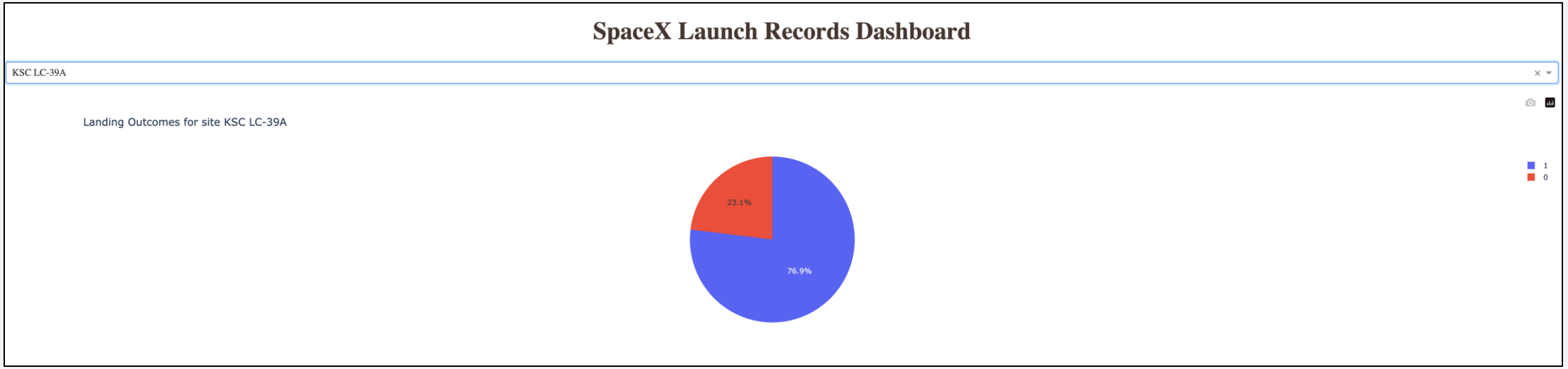
Pie Chart of Landing Successes by Site

- From the chart below of all site launches, we can see that the KSC LC-39A site had the most successful launches of all 4 sites with 10 total successes.
- However, this is only a measure of the number of successful launches and does not account for success rate, which can be better assessed by delving into the individual site data.



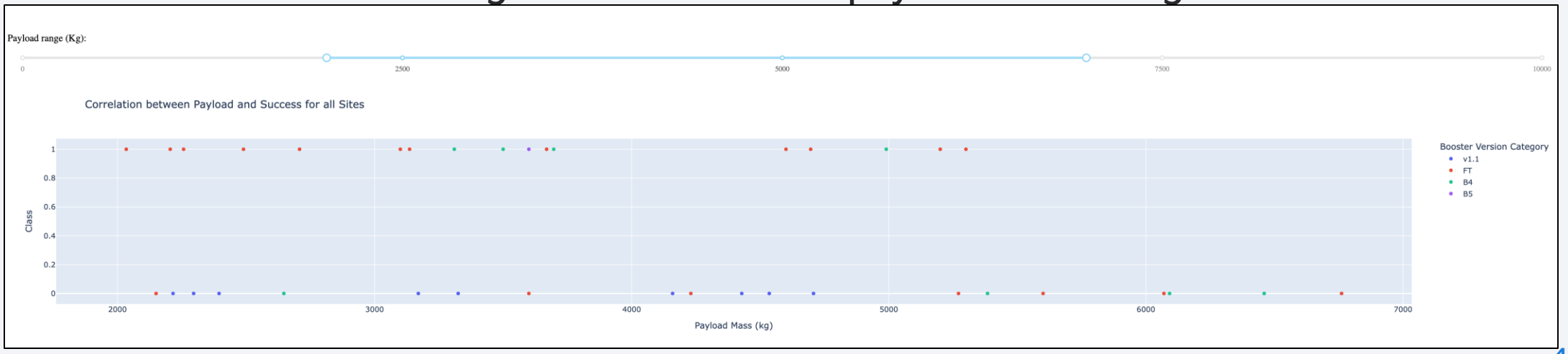
Landing Outcome at Site KSC LC-39A

- The highest site success ratio can be observed in the chart below, selected using the dropdown menu above the Pie Chart.
- KSC LC-39A Site has the highest rate of success at 76.9%, though this is closely followed by the CCAFS LC-40 site with 73.1%. This is significantly higher than the average success rate of around 66%



Impact of Payload on Success Rate

- The highest rate of success across all sites happens with payloads between 2000 and 6000KG
- Within that range, the FT Booster Version appears to have the highest rate of success while the v1.1 Booster Version appears to have the lowest rate of success.
- Note that this trend changes across different payload mass ranges.

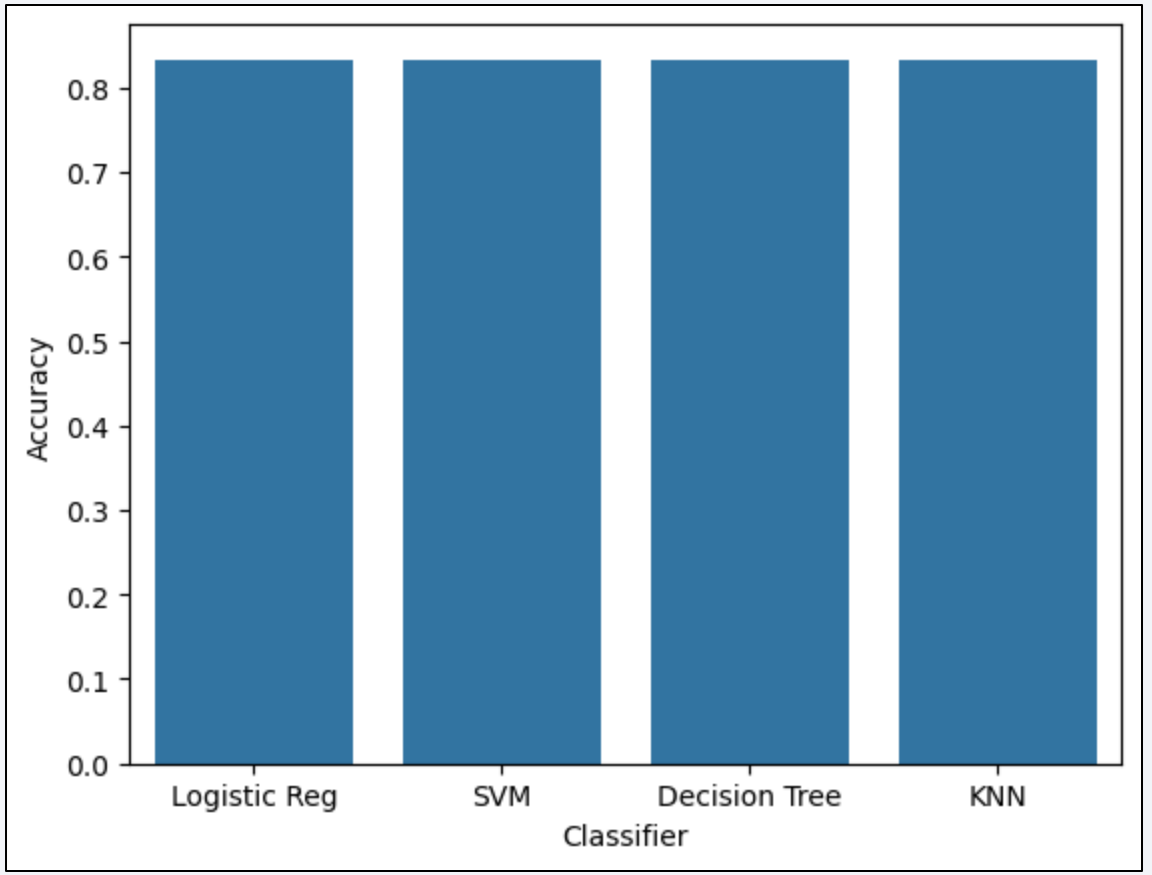


Section 5

Predictive Analysis (Classification)

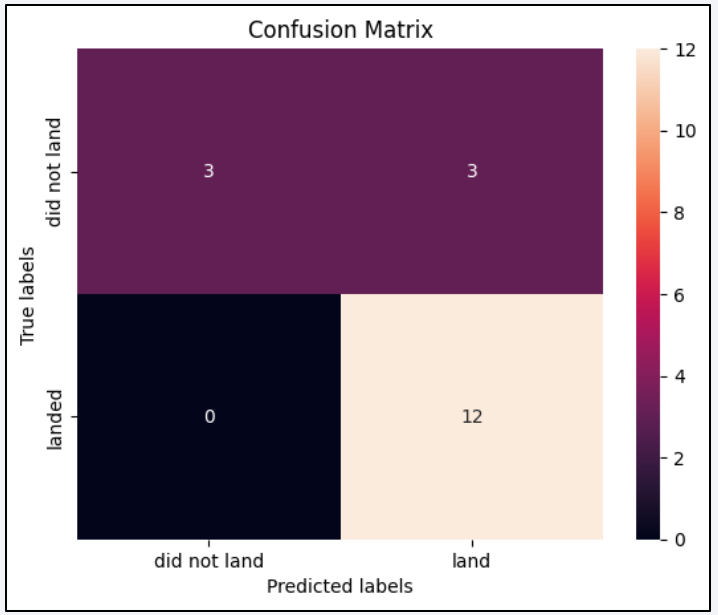
Classification Accuracy

- Based on the bar chart to the right, we can see that all four models have exactly the same performance, indicated by an accuracy of 83.33%
- No single model performed better than any other, so there is no best model based on accuracy.
- Typically, this would be very surprising but is likely attributable to the relatively small size of the overall dataset and test set.



Confusion Matrix

- Because model accuracy was the same for all four tested models, they all yielded the same confusion matrix seen to the right.
- From the confusion matrix we can see that the loss of accuracy was driven by 3 False Positve cases where the models predicted a successful landing that ended up being a failure.
- There were no False Negatives, but this could be because of a lack of balance in the dataset where positive outcomes outnumber negative outcomes 2:1.



Conclusions

- Data on SpaceX rocket launches was successfully collected using both API calls and webscraping
- Exploratory data analysis with SQL queries revealed key insights into the structure of the data including the number of unique sites, payload ranges, dates, and successful missions.
- Exploratory data analysis with visualizations highlighted relationships with key variables and the target outcome of successful landings, leading to down selection of variables for subsequent model building.
- Mapping of sites in Folium allowed extraction of key insights into global localization of major SpaceX sites and proximity to key points of interest like coast and highways.
- A Plotly Dash Dashboard revealed success rate trends between different sites and association with payloads and booster version.
- Finally, logistic regression, SVM, decision tree, and KNN classifiers all performed reasonable well at predicting success with 83.3% accuracy, but the value of these outside of this analysis is questionable given the low data set size.

Appendix

- All relevant code, notebooks, and outputs can be found at the following GitHub repo link: https://github.com/cfergu11/IBMDataScienceCapstop
- Datasets were collected from either:
  - GET requests from the SpaceX data API: https://api.spacexdata.com/v4/launches/past
  - Webscraping Falcon9 Wikipedia page table data: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Thank you!