IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yohanna H.
9/3/2021

Outline

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

Executive Summary

- Data for all SpaceX launches was analyzed for first stage landing success. Data was imported using the SpaceX API, and machine learning techniques (including SVM, Classification Trees, and Logistic Regression) in order to predict whether or not a launch would result in a successful landing
- Most machine learning models resulted in a 83.3% success rate of predicting first stage landings, when examining predictions on the test set. This was also the highest success rate achieved.

Introduction

- SpaceX is a rocket launch company that offers launches at a cost of $62 million. Compared to competitors costs of $165 million and above, this is a major cost savings for customers. It does this by landing the first stage of the rocket and reusing it for subsequent launches.
- We want to know if we can predict when a launch will be successful or not based on a range of launch parameters.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collection was done using the SpaceX API
- Perform data wrangling
  - Data was saved in a pandas data frame and Falcon 1 data was removed, NULL values were replaced, and a Binary “Outcome” column was generated
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, SVM, Decision Tree, and K nearest neighbor models were built and tuned using a grid search technique and a training set of data. Models were evaluated using a test data set.

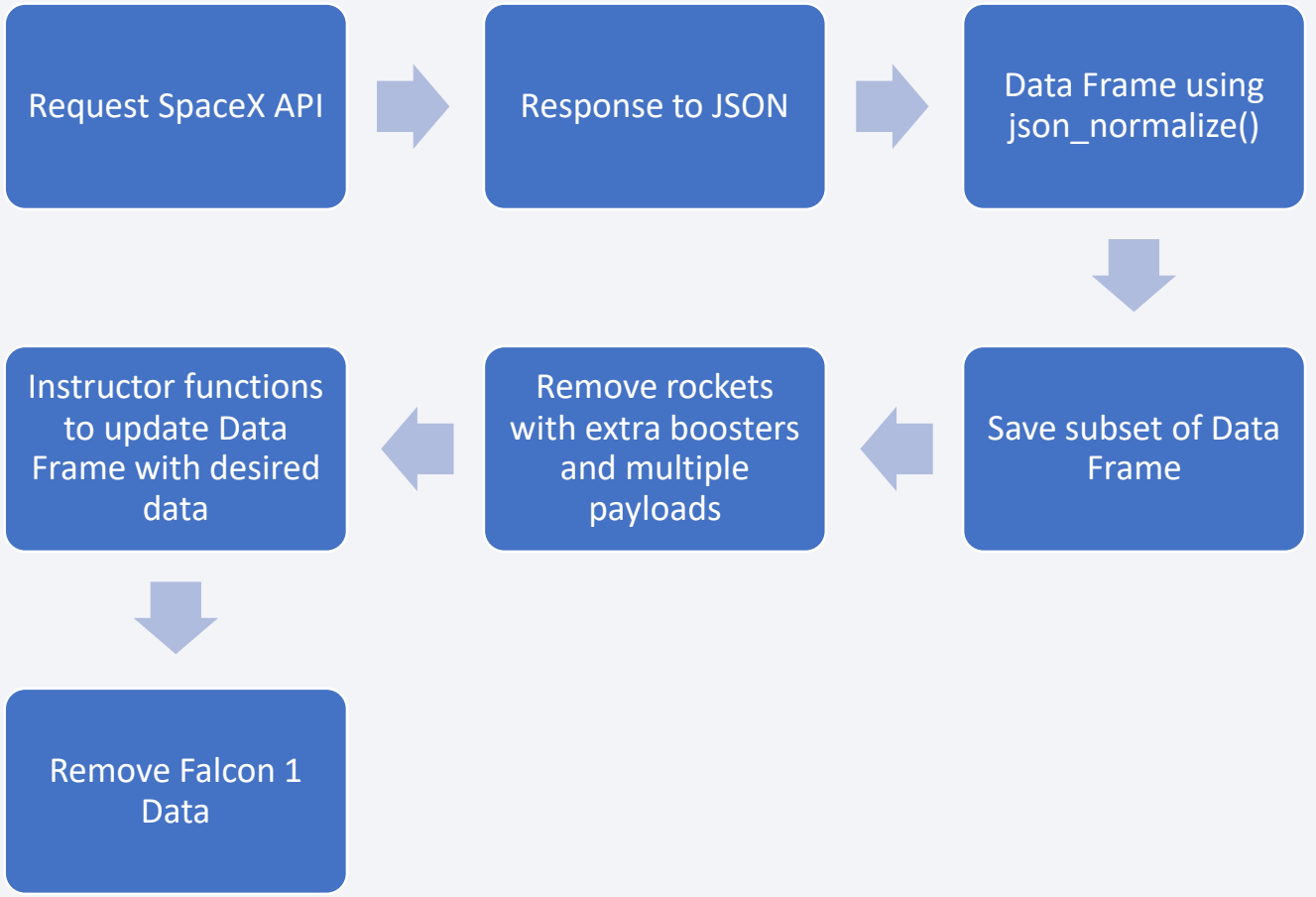
Data Collection

- Data was collected using the SpaceX API and the requests python library
- This response was parsed into JSON using the .json() method, which was then turned into a pandas data from using .json_normalize()
- This data frame was reduced to a subset of just Rockets, Payloads, Launchpad, Cores, Flight Number, and Date
  - Rockets with extra boosters and multiple payloads were removed
- Using functions written by the instructors, the data frame was updated to have the columns: Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Grid Fins, Reused, Legs, Block, Reused Count, Serial, Longitude, Latitude
- This was then reduced to just Falcon 9 launches

Data Collection - SpaceX API

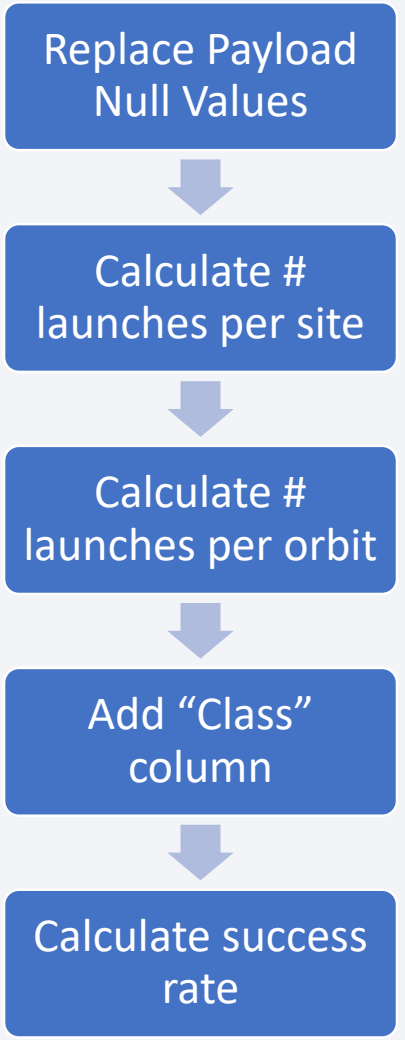
- The method shown the previous slide is visually shown as a flow chart to the right

- GitHub URL of the completed SpaceX API calls notebook: https://github.com/yohanna3/SpaceX/blob/eb2bc5b713b72b457c98a5daa58d23e1bda99513/Data%20Collection%20Lab.ipynb



Data Wrangling

- Both Payload Mass and Landing Pad had NULL Values
  - Payload Mass NULL values were replaced by the mean payload mass of all launches
  - Landing Pad values were left as None to indicate when a landing pad was not used
- Number of launches for each site was calculated
- Number of launches for each orbit type was calculated
- Number of each landing outcome was calculated
- Column named “Class” added to the data frame with 0 representing a failed landing and 1 representing a successful landing
- Success rated calculated as the mean value of “Class”
- Github URL: https://github.com/yohanna3/SpaceX/blob/eb2bc5b713b72b457c98a5daa58d23e1bda99513/Data%20Wrangling.ipynb



EDA with Data Visualization

- Scatter plot of Payload Mass vs Flight Number colored by Landing Success was used to determine the relationship between Payload Mass and Flight Number
- Scatter plot of Launch Site vs Flight Number colored by Landing Success was used to determine the relationship between Launch Site and Flight Number
- Scatter plot of Launch Site vs Payload Mass colored by Landing Success was used to determine the relationship between Payload Mass and the Launch Site
- Bar Chart comparing the Landing Success with different Orbits was used to determine if Orbit effects Landing Success
- Scatter plot of Orbit vs Flight Number colored by Landing Success was used to determine the relationship between Orbit and Flight Number
- Scatter plot of Orbit vs Payload Mass colored by Landing Success was used to determine the relationship between Orbit and Payload Mass
- Line plot of Landing Success vs Year was used to determine the trend of success with time
- GitHub URL: https://github.com/yohanna3/SpaceX/blob/eb2bc5b713b72b457c98a5daa58d23e1bda99513/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- Display all unique Launch Sites
- Display 5 records of Launch Sites beginning with “CCA”
- Display total Payload Mass carried by boosters launched by NASA
- Display average Payload Mass carried by booster version F9 v1.1
- Display when the first successful landing on a ground pad was achieved
- List names of the boosters which achieved successful landing on a drone ship and had a payload between 4000 and 6000 kg
- List the total number of successful and failed missions
- List the names of the Booster Versions which carried the maximum Payload Mass
- List the failed outcomes on a drone ship, their Booster Versions and Launch Sites
- Rank the count of landing outcomes between 06/04/2010 and 03/20/2017
- Using bullet point format, summarize the SQL queries you performed
- GitHub URL: https://github.com/yohanna3/SpaceX/blob/2601b9c0c8c276ae79491fe17a3326b44cb9dee8/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

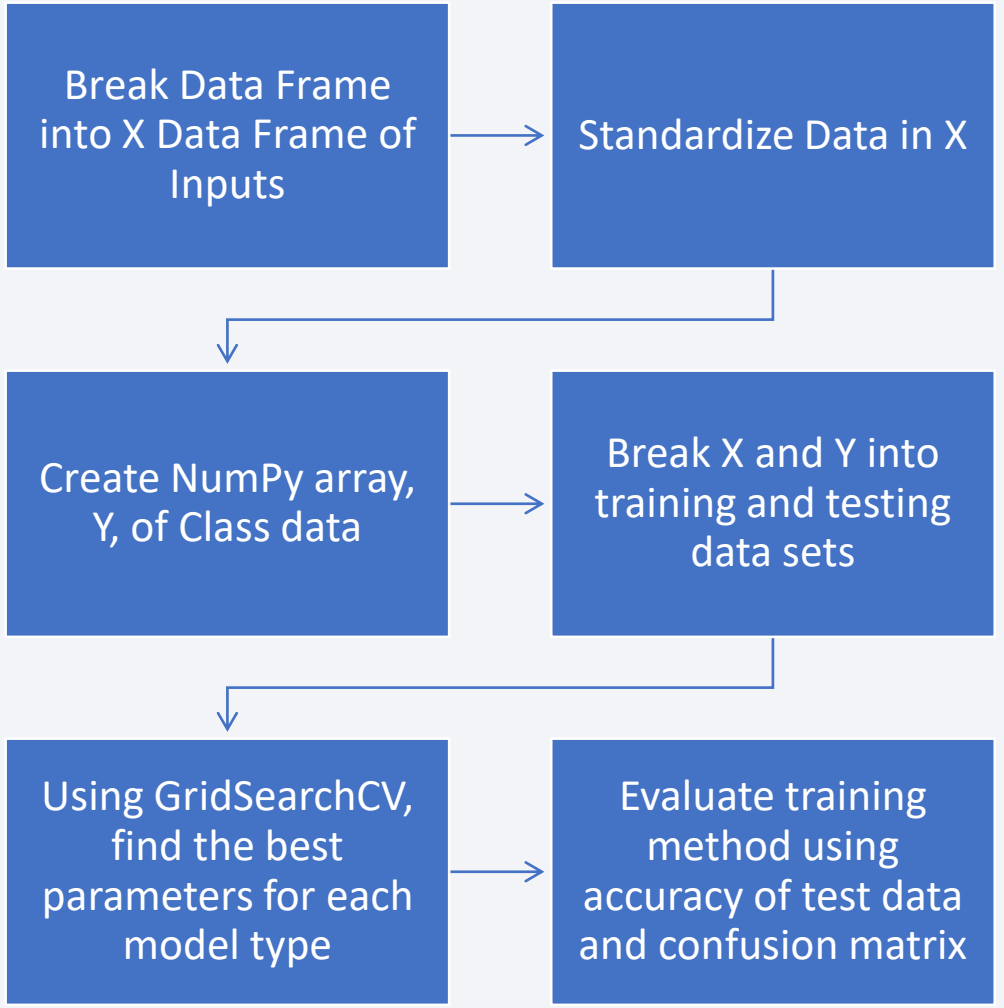
- Circles and Markers were placed to show launch sites and label them respectively
  - This so we could get an idea of the relative position of the launch sites to one another and the equator
- MarkerClusters were used to mark locations of launches cleanly. Markers for each launch were color code by whether they landed successfully
- The distance from the Vandenberg launch site and a rail line was plotted using PolyLine and a Marker was used show the calculated distance. This was to get an idea of how far the launch sites are from land marks
- Note that folium maps do not render on GitHub
- GitHub URL: https://github.com/yohanna3/SpaceX/blob/34e1f119ae9519b5ac38643a3c21e9322fcb667d/Data%20Visualization%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- If All Sites selected:
  - Pie graph generated showing the percentage of successful landings by Launch Site
  - Scatter plot of Launch success vs Payload Mass colored by Booster Version within the selected Payload Mass Range from the slider for all sites
- If a specific site is selected:
  - Pie graph generated showing the percentage of successful landings and failures for the selected Launch Site
  - Scatter plot of Launch success vs Payload Mass colored by Booster Version within the selected Payload Mass Range from the slider for the selected site
- These plots were hoping to show how landing success varies by site and how the payload effects the launch success
- Add the GitHub URL: https://github.com/yohanna3/SpaceX/blob/303a185937df9b8b6627f95768d00fbe992f8d0b/spacex_dash_app.py

Predictive Analysis (Classification)

- Used the flow chart to the right to find the best parameters for Logistic Regression, Support Vector Machine (SVM), Tree Classification, and k Nearest Neighbor methods
- For each method, a dictionary of parameters were fed into the GridSearchCV object
- Methods assessed relative to one another based on their best parameters
- GitHub URL: https://github.com/yohanna3/SpaceX/blob/303a185937df9b8b6627f95768d00fbe992f8d0b/Machine%20Learning%20Prediction.ipynb

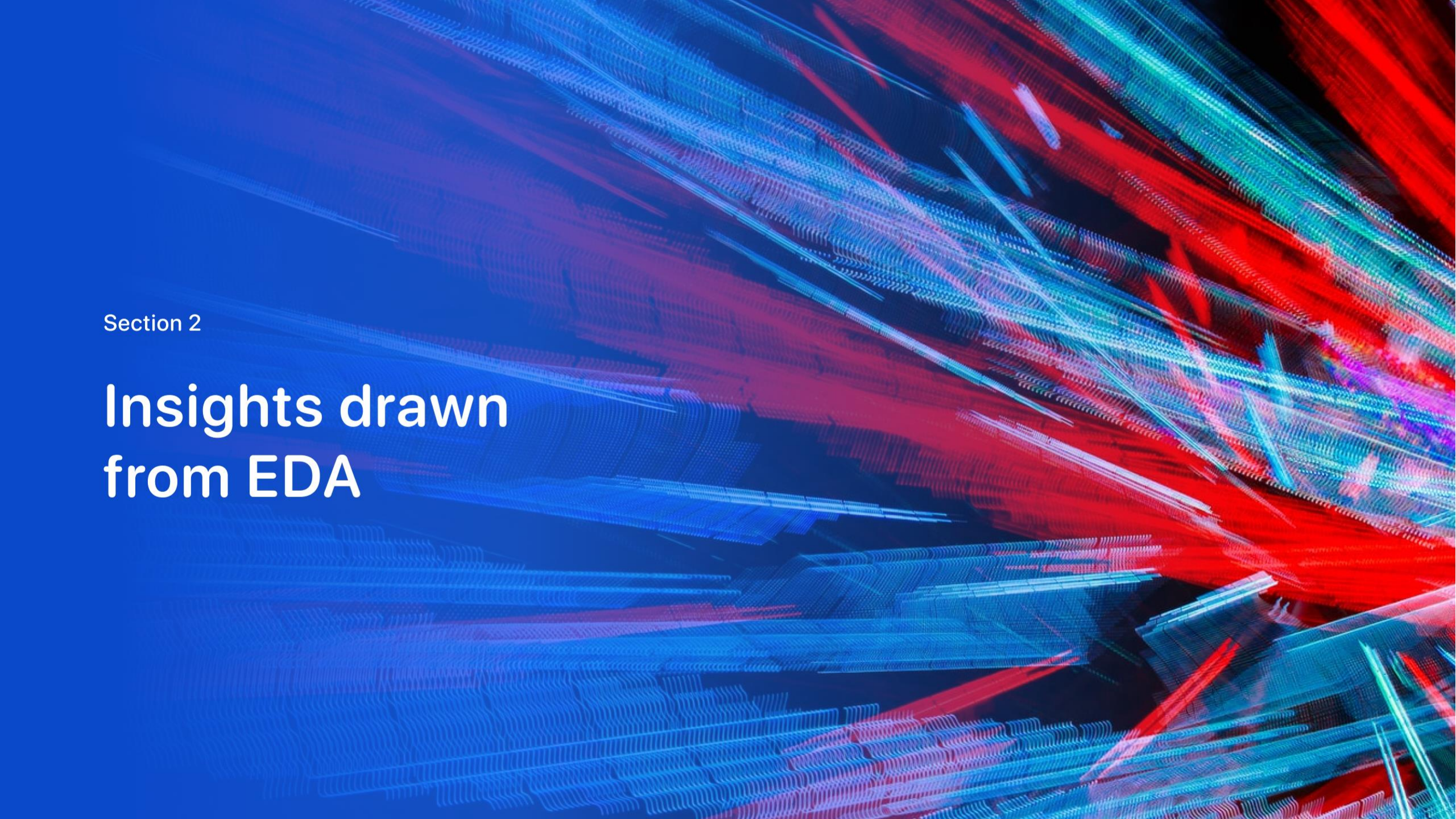


Results

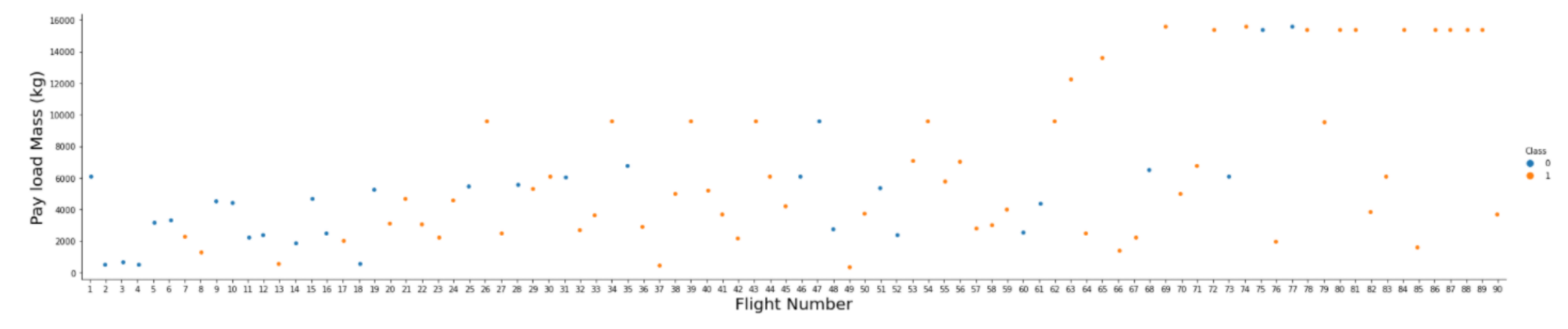
- Exploratory data analysis (EDA) showed that likelihood of successful landing increased with as time went on (i.e. the technology likely matured)
- EDA was also able to quantify the total payload that made it to space
- EDA visualization showed launch sites and tracked their successful launches and allowed for additional insights
- Interactive analytics demo plotted the success rates as well as a scatter plot of success vs payload for all and individual sites
- Predictive analysis showed no individual method outperformed the others in predicting landing success

Section 2

Insights drawn from EDA

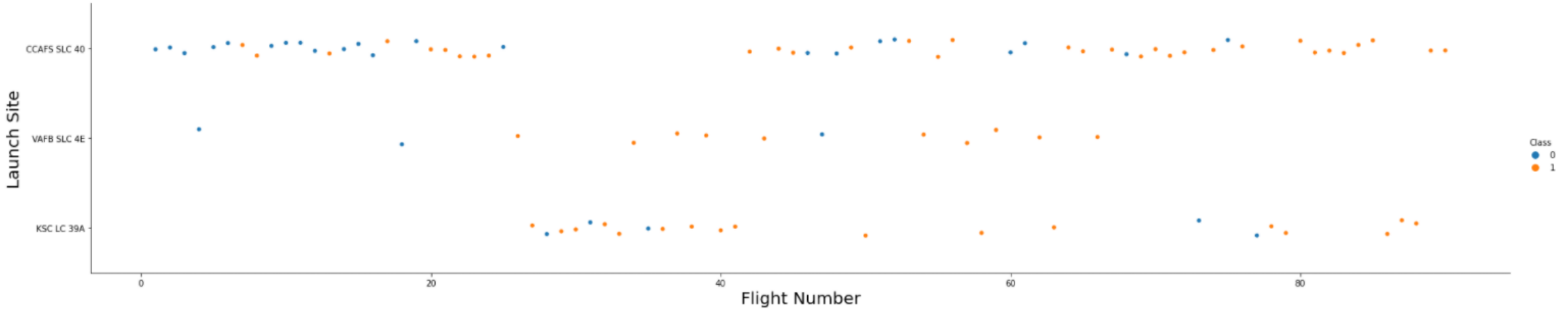


Flight Number vs. Payload



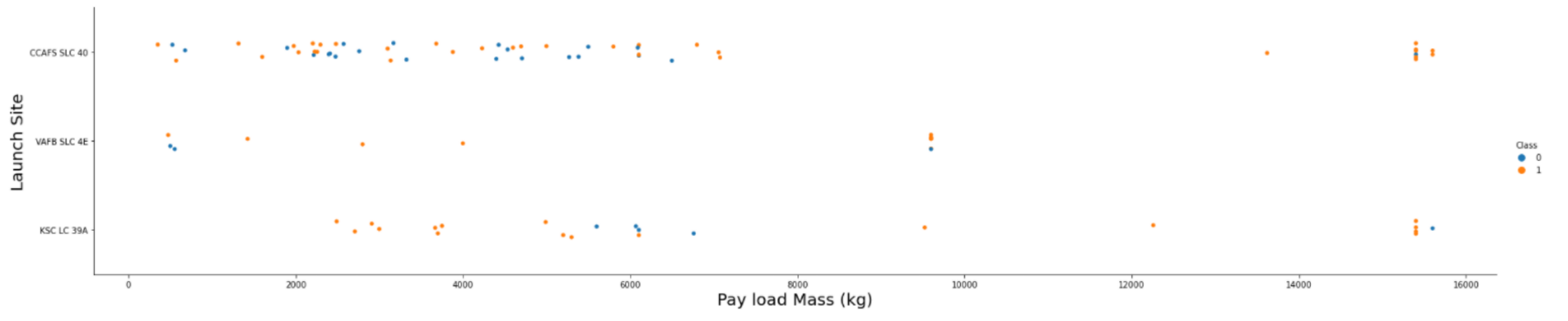
- Early flight numbers had more failures with landing the first stage
- As flight number increases, the payload also increases on average
  - More successful landings as well

Flight Number vs. Launch Site



- Early flight numbers mainly launched from Cape Canaveral SLC 40 and were unsuccessful for the most part
- As flight number increases, launches spread to Vandenberg Air Force Base and Kennedy Space Center, but mainly still Cape Canaveral
  - More successful landings as well

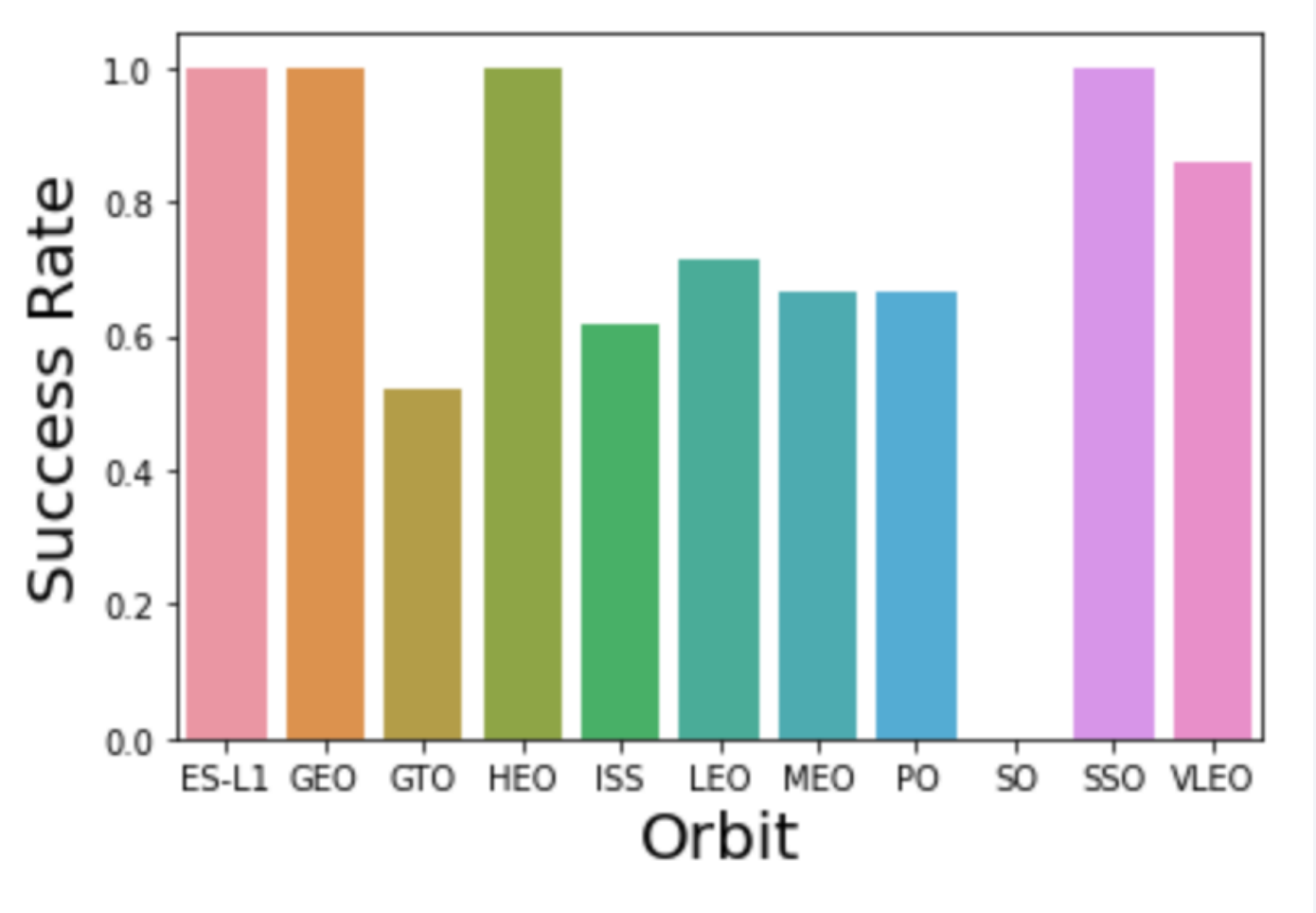
Payload vs. Launch Site



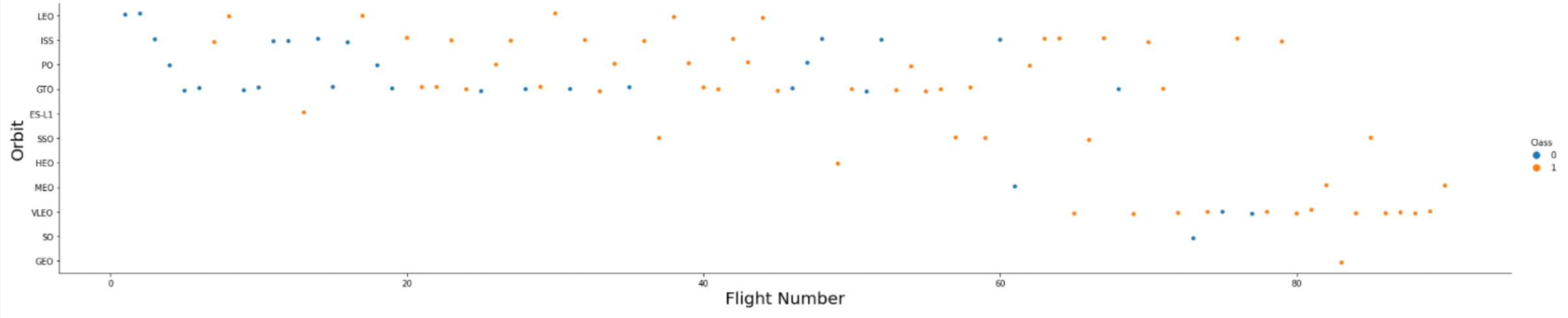
- No real trends between payload and launch sites
- KSC had no launches below 2,000 kg
- VAFB had no launches above 10,000 kg
- CCAFS had no launches between 8,000 and 12,000 kg

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO orbits successfully landed all first stages
- SO orbit never successfully landed any rocket components
- All other orbits had mixed landing success
- Data may be skewed due to the number or launches for each orbit type

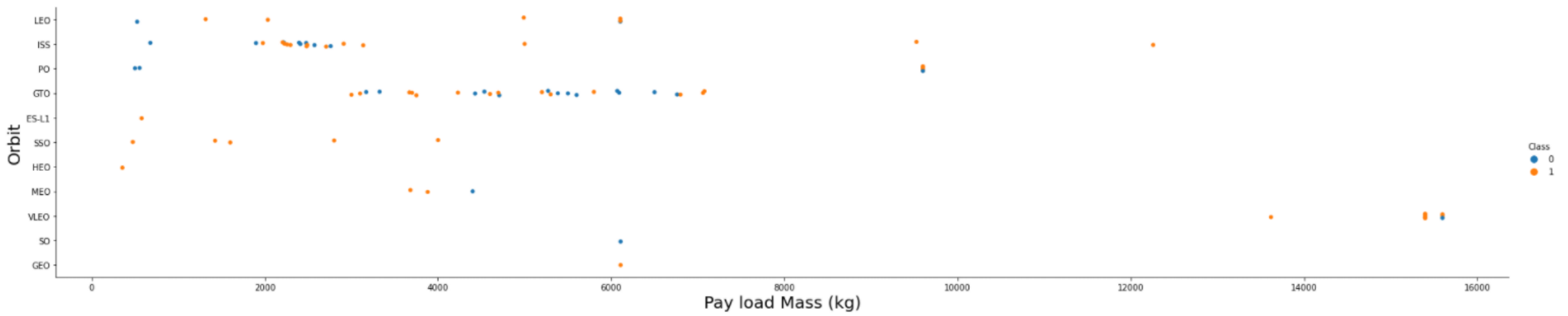


Flight Number vs. Orbit Type



- No real trends between flight number and orbit type
- SSO, HEO, MEO, VLEO, SO, and GEO orbits not attempted until much later flight numbers

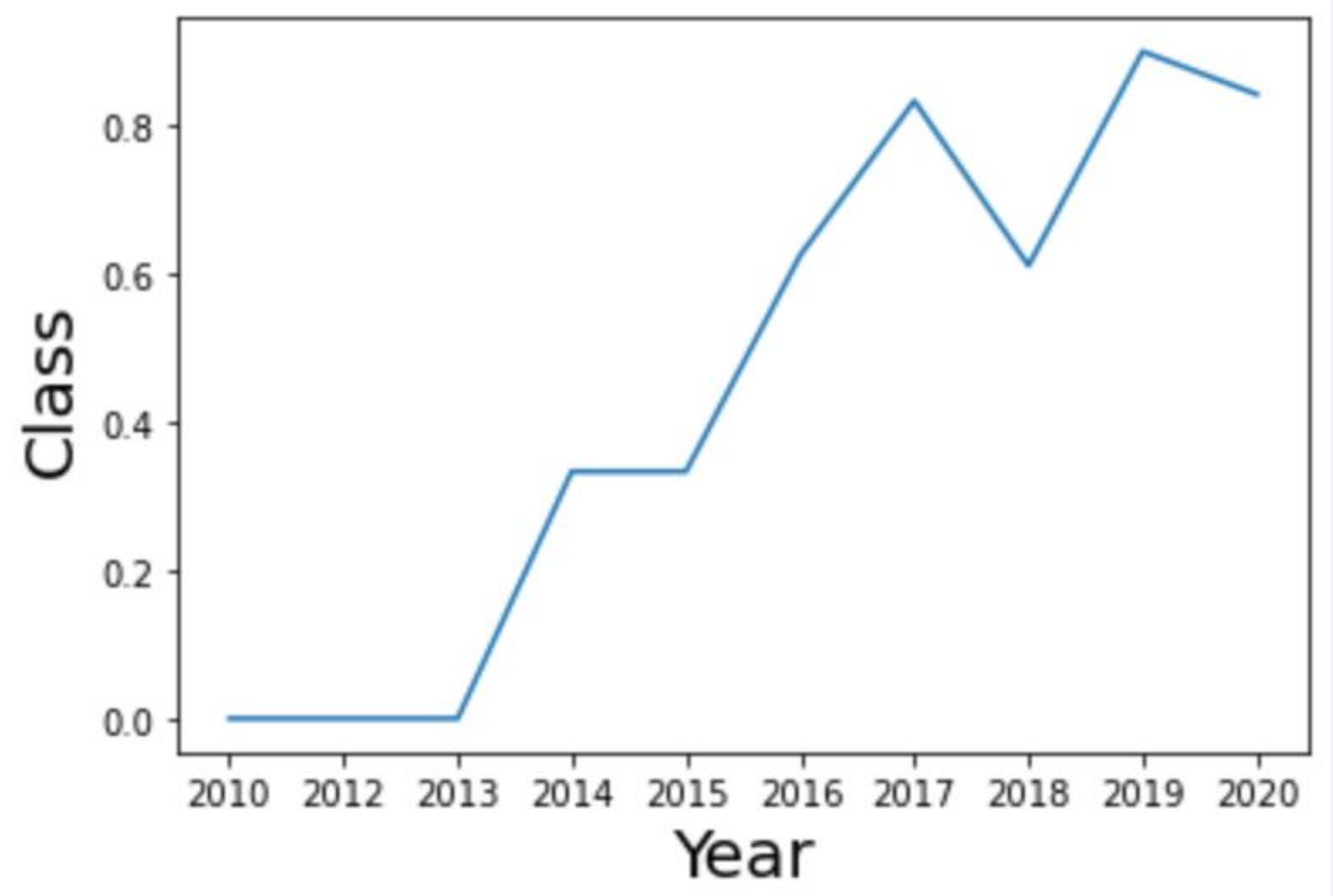
Payload vs. Orbit Type



- No real trends between payload and orbit type
- VLEO orbit only for very heavy payloads
- Only payloads around 6,000 kg launched into SO and GEO orbits

Launch Success Yearly Trend

- There were no successful landings till 2014
- The rate of success increased for the most part except for decreases in 2018 and 2020



All Launch Site Names

- The data frame was probed for all the unique launch site names, presented to the right
- Note, Cape Canaveral is split into LC-40 and SLC-40
  - This is likely a name change of the facility later in history

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

Launch Site Names Begin with 'CCA'

- 5 records were pulled from the data frame where the launch site began with CCA
- All data frame columns are presented
- Most of the customers are NASA

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The data frame was probed for the total payload mass carried to space, where NASA was the launch customer (image of query result presented to the right)
- The total mass was 99,980 kg

1
99980

Average Payload Mass by F9 v1.1

- The data frame was probed for the average payload mass launched by F9 v1.1 boosters (result presented to the right)
- The average mass was 2,928.4 kg

1
2928.400000

First Successful Ground Landing Date

- The data frame was probed for the first successful landing on the ground pad (result presented to the right)
- This was on December 22, 2015

	1
	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The data frame was probed for the booster versions where there was a successful landing on a drone ship and the payload was between 4,000 and 6,000 kg(result presented below)
- All the boosters were derivatives of the F9 FT booster

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 data frame was probed for the total number of successful and failure mission outcomes (result presented to the right)
- There were 61 successes
- There were 10 failures

success	failure
61	10

Boosters Carried Maximum Payload

- The data frame was probed for the booster versions which carried the maximum payload mass (result presented to the right)
- All the boosters were derivatives of the F9 B5 booster

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

- The data frame was probed for the booster versions and launch sites for failed landing outcomes on drone ships in the year 2015 (result presented below)
- Both launched from Cape Canaveral and had similar boosters (but not the same)

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- The data frame was probed for the count of landing outcomes ranked in descending order for launches between June 4, 2010 and March 20, 2017 (result presented to the right)
- In this time frame, it was mostly successes

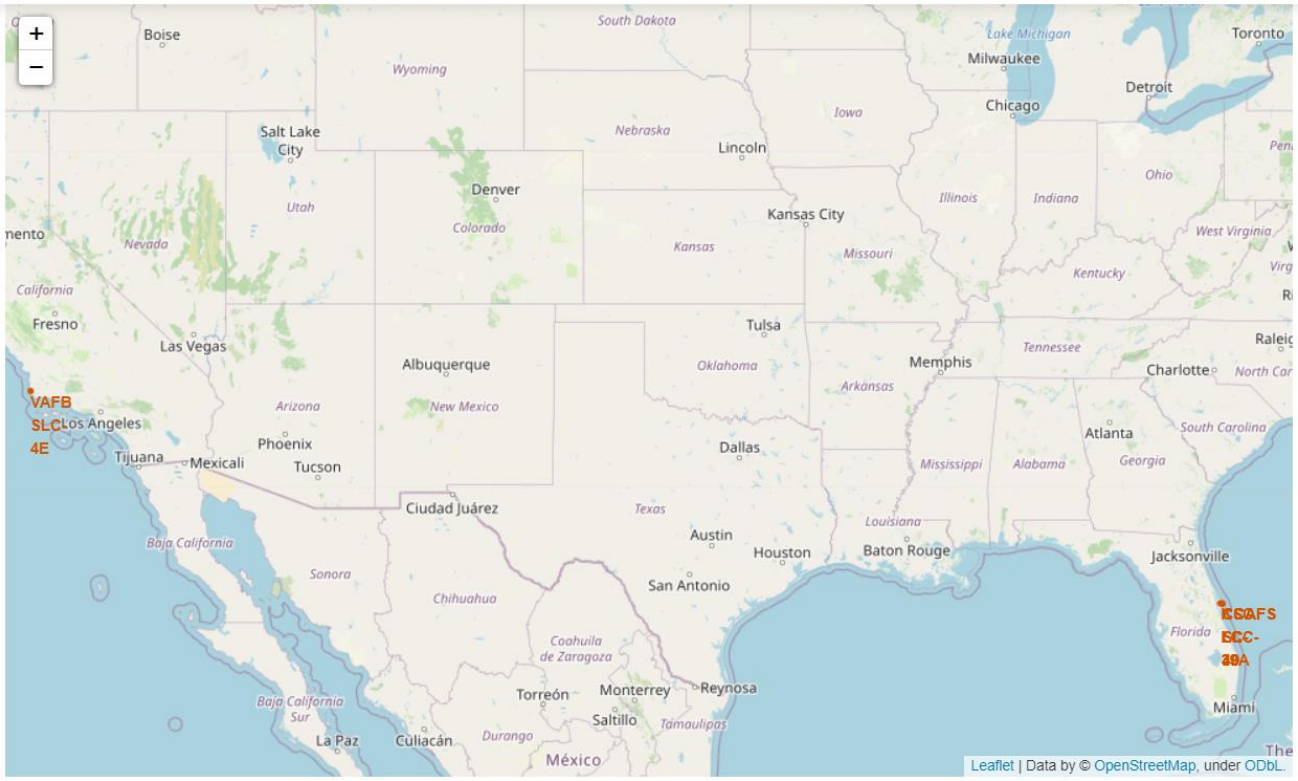
landing__outcome	2
Success	38
No attempt	22
Success (drone ship)	14
Success (ground pad)	9
Controlled (ocean)	5
Failure (drone ship)	5
Failure	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

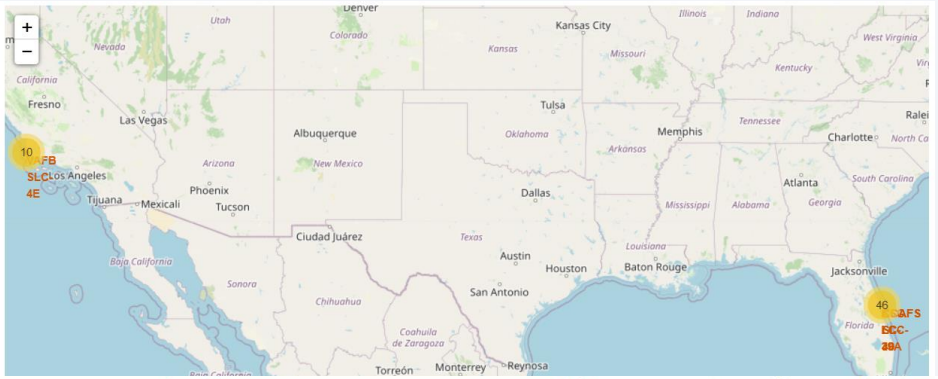
Map of Launch Sites

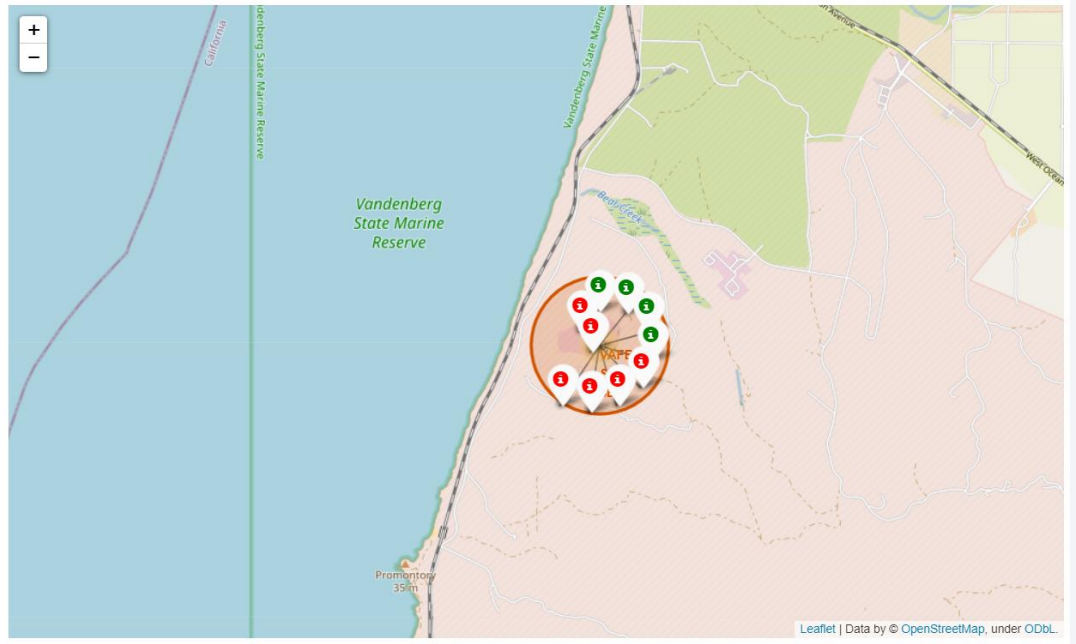
- All Launch sites appear close to one another in terms of latitude
- These are Southern points in the continental United States
    - This is likely to place the rocket close to the equator in order to give it a boost on launch



Clustering of Individual Launches

- All launches were plotted on the map
  - Clusters were made for comparable launch sites
  - Red indicates a failed landing
  - Green indicates a successful landing
- Most of the launches at Vandenberg, for example, were unsuccessful



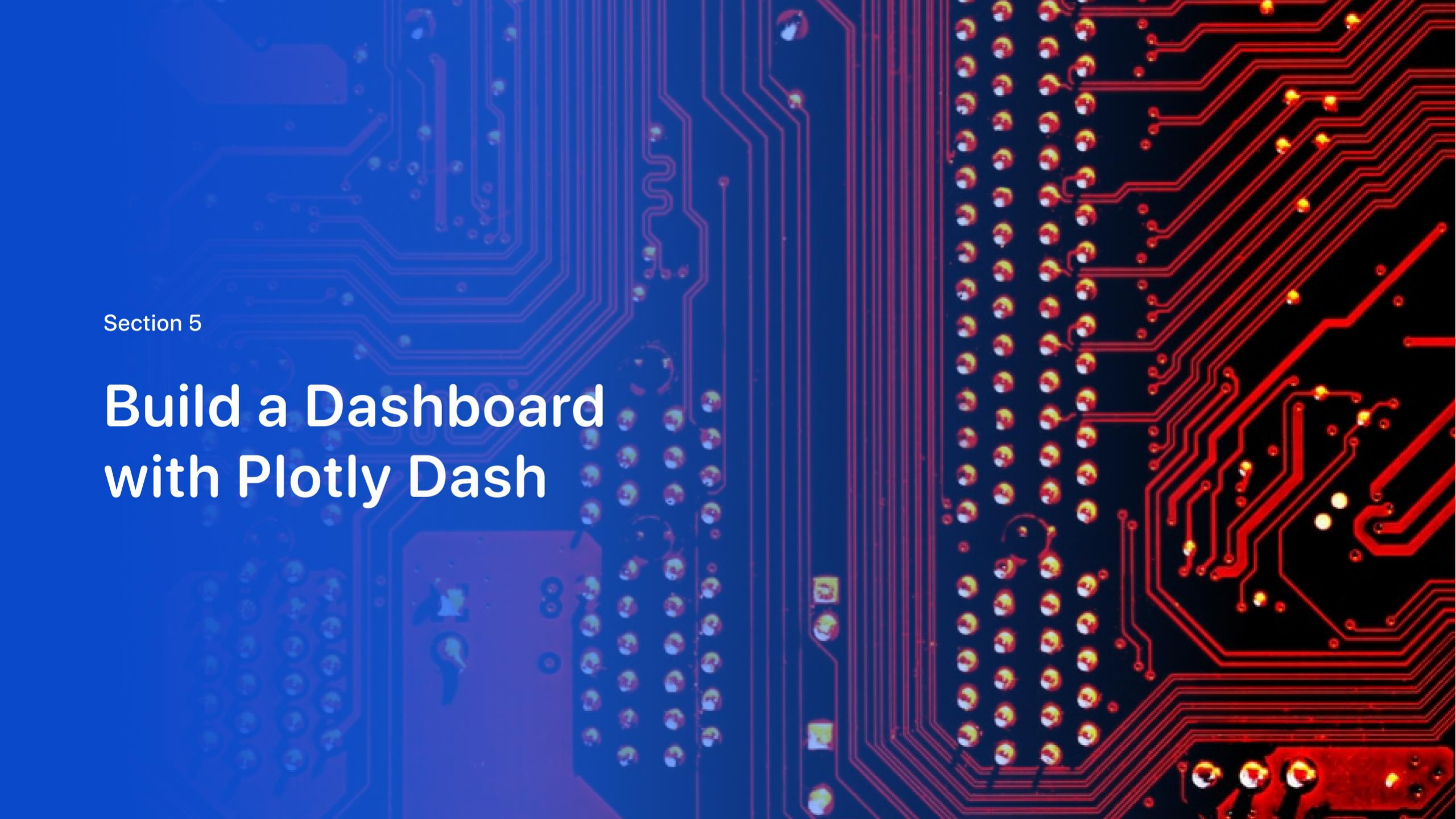


Launch Location Relative to Landmarks

- The yellow marker was placed in order to get the distance from a rail line to Vandenberg Air Force Base
- The distance was calculated to be a little over 1.34 km
    - The coastline is not further away
- This makes sense, so that the rocket can easily be transported to the launch site
- For safety reasons, it is a good idea to be close to an ocean so that if there is a malfunction, the rocket won't crash in a populated area

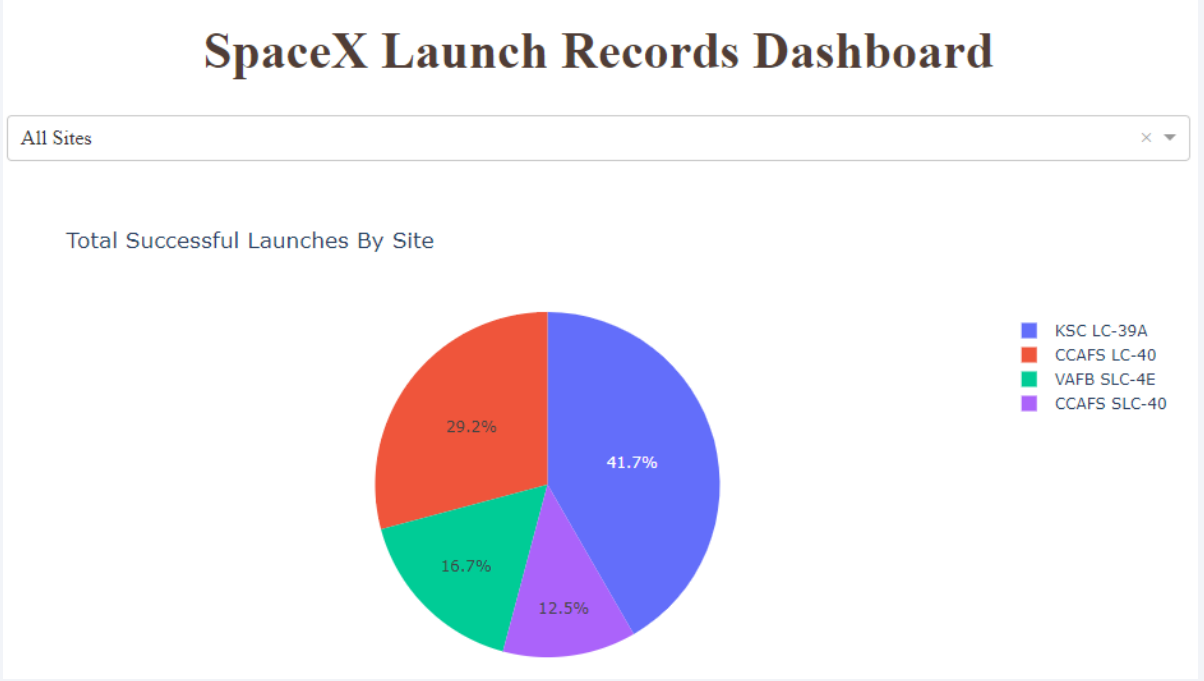
Section 5

Build a Dashboard with Plotly Dash



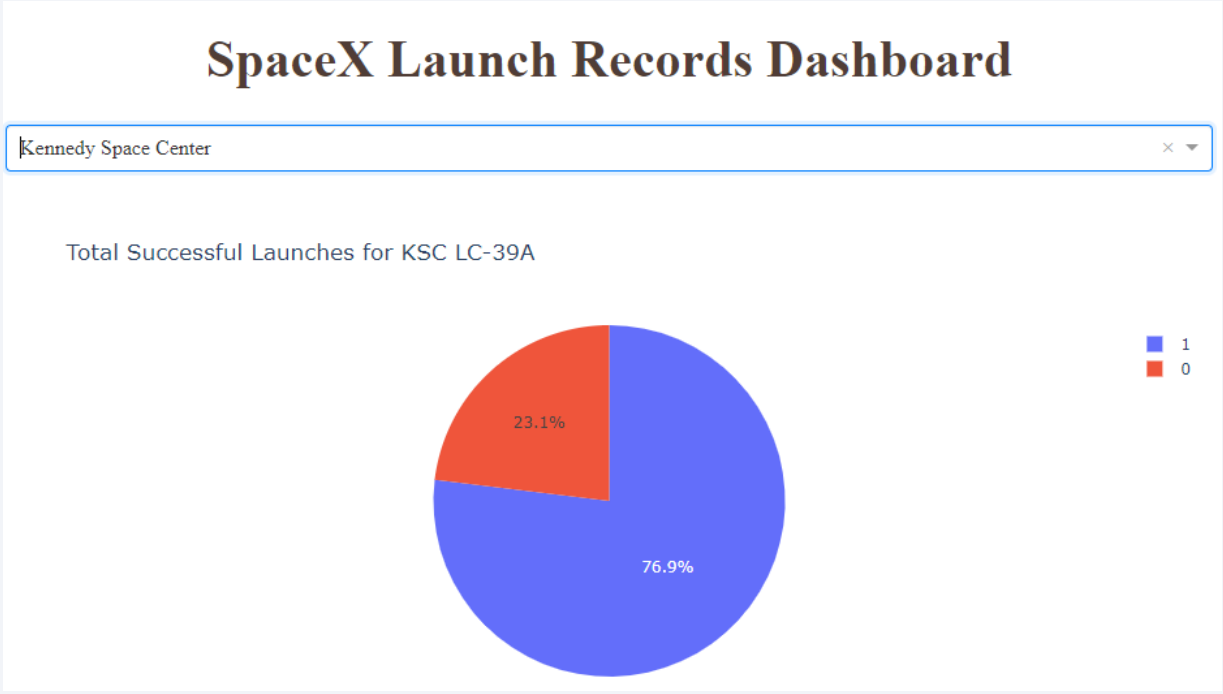
Successful Launches by Site

- The pie chart to the right takes all the successful launches and breaks it into the percentage by launch site
- Kennedy Space Center had 41.7% of all successful launches, which was the greatest
- Cape Canaveral SLC-40 had 12.5% of all successful launches, which was the least



Highest Launch Ratio

- The pie chart to the right looks at all the launches at Kennedy Space Center, and shows the number the percentage that were successful and unsuccessful
- Of all launch sites, KSC had the highest success to failure rate
- 76.9% of all launches at this site were successful



Payload and Launch Outcome

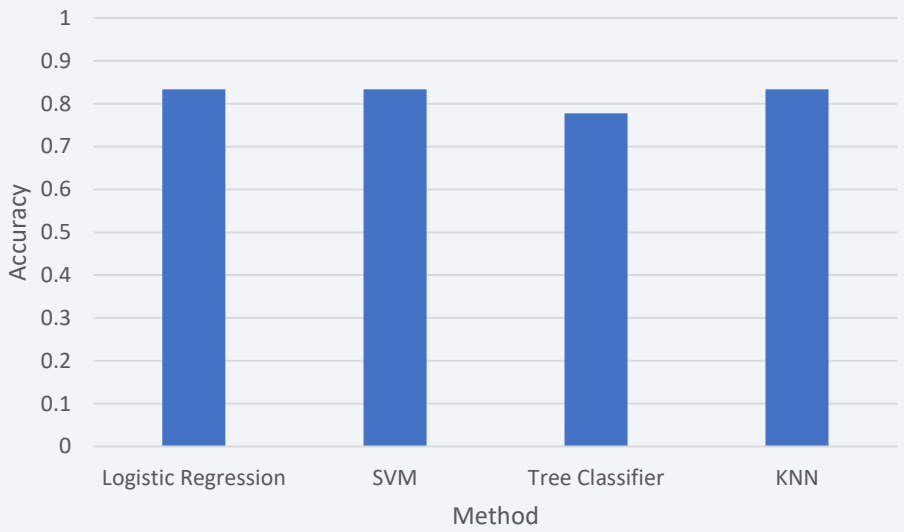
- The sliding bar allowed for custom payload ranges to be selected
- Data in scatter plots shown for all launch sites
- The heaviest launches tended to be successful
- Medium to light launches had no apparent trend to them

Section 6

Predictive Analysis (Classification)

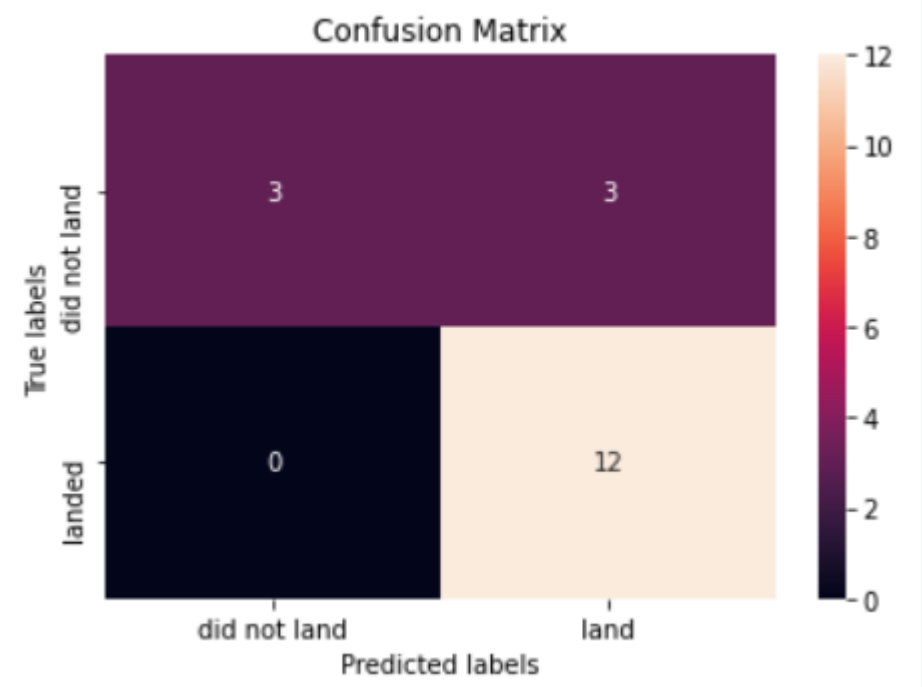
Classification Accuracy

- Bar chart to the right visualizes the accuracy of each model on predicting launch success for a test data set
- Logistic Regression, SVM, and KNN all had the same accuracy
- Tree Classifier underperformed the other methods
- Models should be reassessed in the future with additional data



Confusion Matrix

- Logistic Regression, SVM, and KNN all had the same confusion matrix
- All methods correctly labeled 15 out of 18 test cases
- It incorrectly gave 3 out of 18 false positives



Conclusions

- SpaceX has improved its launch success with time
- Larger payloads tend to launch successfully
  - Also tend to be launched later on
- No correlation of launch success with launch site
- Launch sites are as close to the equator as possible
- Some orbits tend to be more successful for launches
- Logistic Regression, SVM, and KNN all predicted labels for launches with 83% accuracy

Thank you!