Winning Space Race with Data Science

Michael Kleen

12 May 2023

Outline

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

Executive Summary

- Utilized the following data analysis methodologies:
  - Data collection with web scraping and SpaceX API.
  - Exploratory Data Analysis (EDA), which includes data wrangling, interactive visual analytics, and data visualization.
  - Machine learning prediction.
- Summary of all results
  - Public souces provided valuable data.
  - EDA identified which features are the best to predict launching success.
  - Machine Learning Prediction used the best model to predict which characteristics are important to drive effective decisionmaking using all collected data.

Introduction

- To evaluate the viability of Space Y to compete with Space X
- Problems:
    - What is the best way to estimate total launch cost by predicting successful landings of the first stage of rockets?
    - Where is the best launch site?

Section 1

Methodology

Methodology

Executive Summary

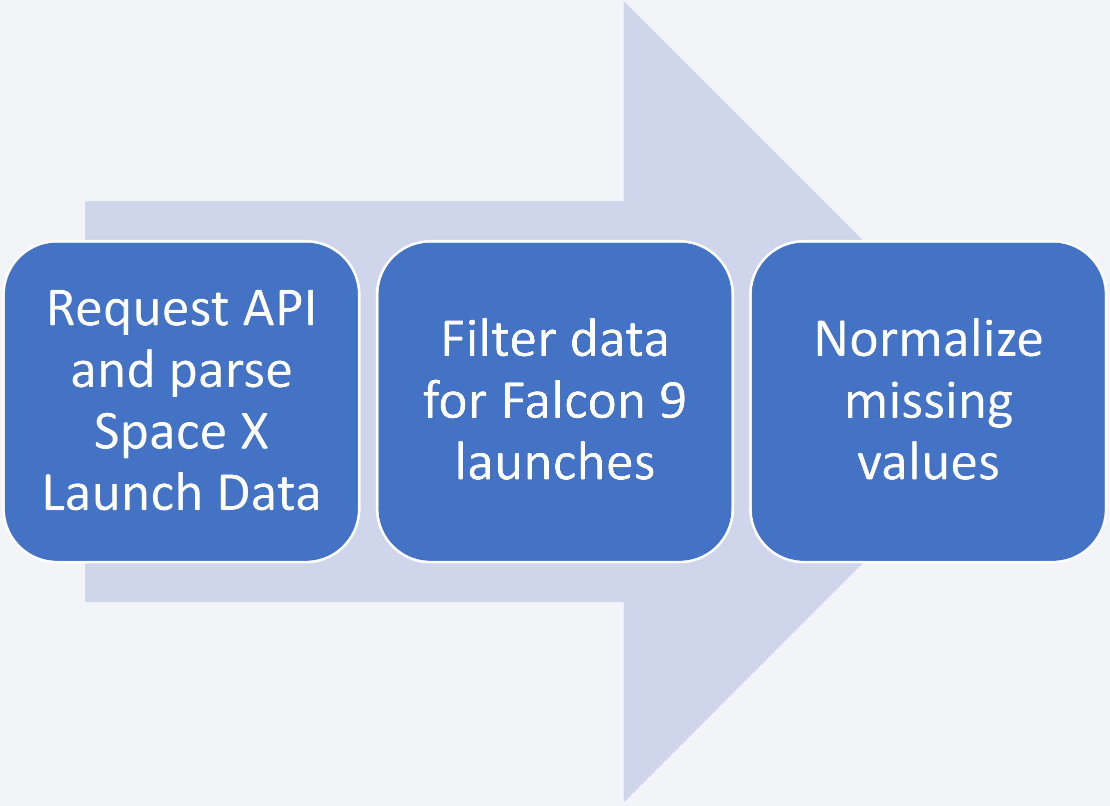
- Data collection methodology:
  - Data from Space X was obtained from the following sources:
    - Webscraping https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
    - Space X API
- Perform data wrangling
  - Enriched collected data by creating a landing outcome label based on outcome data after summarizing and analyzing features.
- 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 normalized, divided into training and test data sets, and evaluated by four different classification models. The accuracy of each model was evaluated using different combinations of parameters.

Data Collection

- **Space X API**
- **Webscraping**

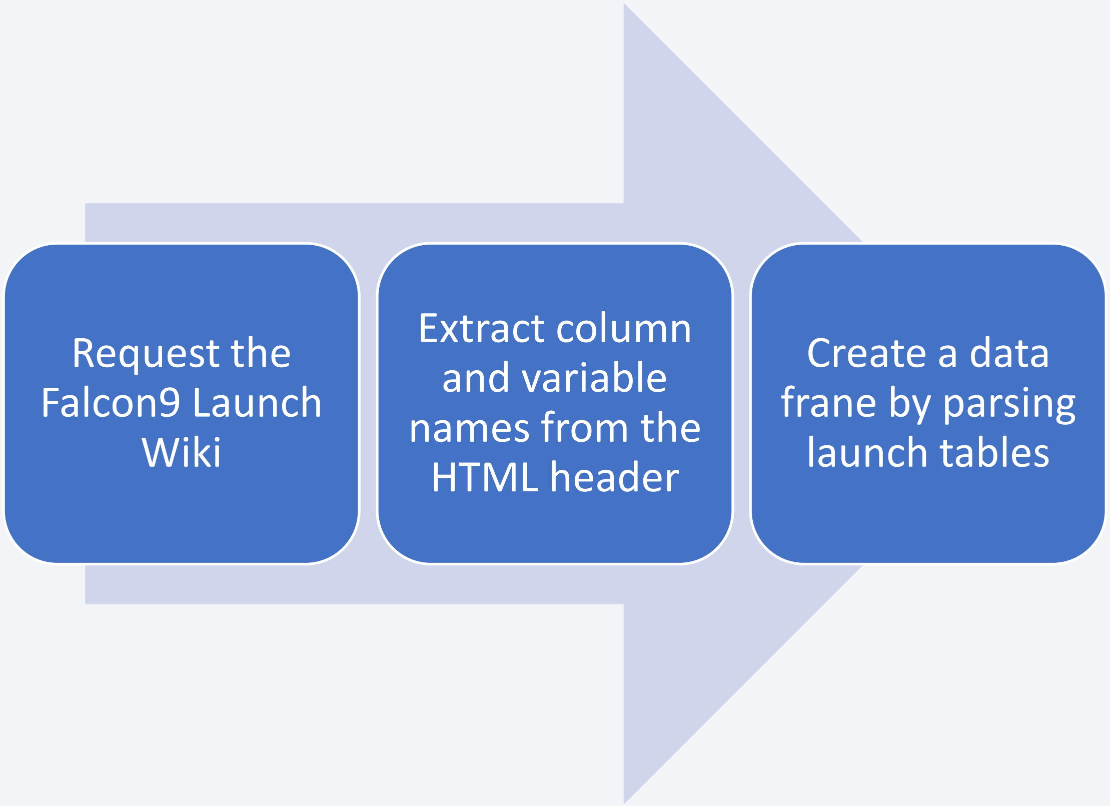
Data Collection – SpaceX API

- SpaceX offers a public API to allow public acess to data
- The accompanying flow chart shows how we obtained this data using API
- https://github.com/makleen81/Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



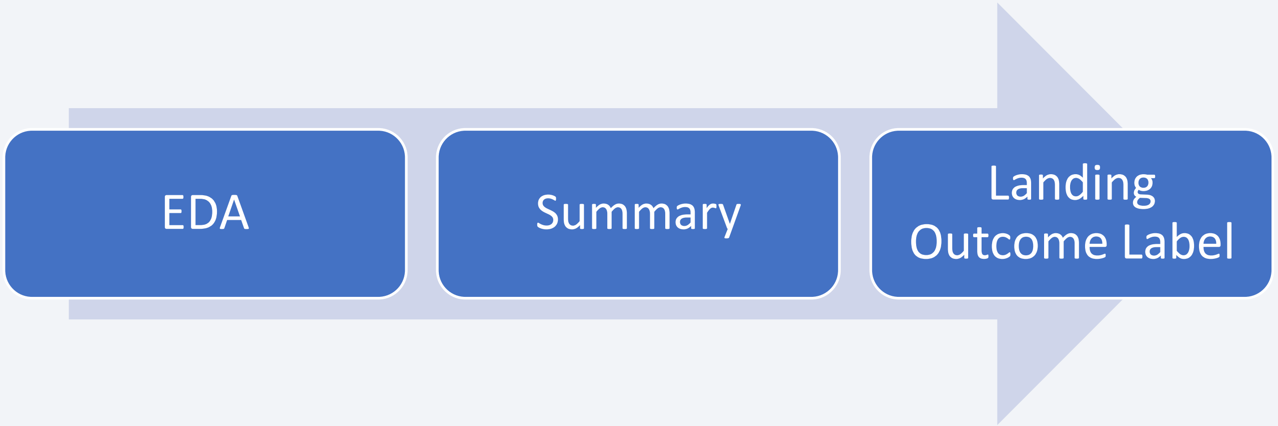
Data Collection - Scraping

- Data on SpaceX launches can be obtained from Wikipedia using scraping
- Data is obtained from Wikipedia via the process depicted on the right
- https://github.com/makleen81/Capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

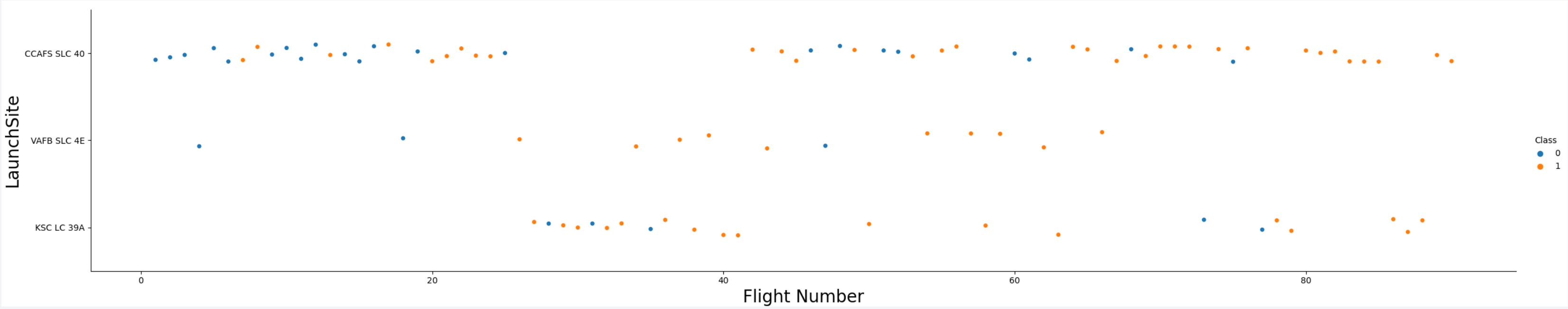
- Exploratory Data Analysis (EDA) was performed on the dataset.
- Then we calculated the launches per site, occurrences of each orbit, and occurrences of mission outcome per orbit type.
- Finally, we created a landing outcome label from the Outcome column.



- https://github.com/makleen81/Capstone/blob/main/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- Scatterplots and barplots were used to visualize the relationship between features:
- Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit



- https://github.com/makleen81/Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

The following SQL queries were performed:

- Names of the unique launch sites in the space mission;
- Top 5 launch sites whose name begin with the string 'CCA';
- Total payload mass carried by boosters launched by NASA (CRS);
- Average payload mass carried by booster version F9 v1.1;
- Date when the first successful landing outcome in ground pad was achieved;
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
- Total number of successful and failure mission outcomes;
- Names of the booster versions which have carried the maximum payload mass;
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015; and
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

- https://github.com/makleen81/Capstone/blob/main/jupyter-labs-EDA.ipynb

Build an Interactive Map with Folium

- Folium Maps utilized markers, circles, lines and marker clusters.
- Markers indicated points like launch sites.
- Circles indicated highlighted areas around specific coordinates.
- Marker clusters indicated groups of events in each coordinate.
- Lines were used to indicate distances between two coordinates.

- https://github.com/makleen81/Capstone/blob/main/jupyter-labs-Interactive-Visual-Analytics-with-Folium.ipynb

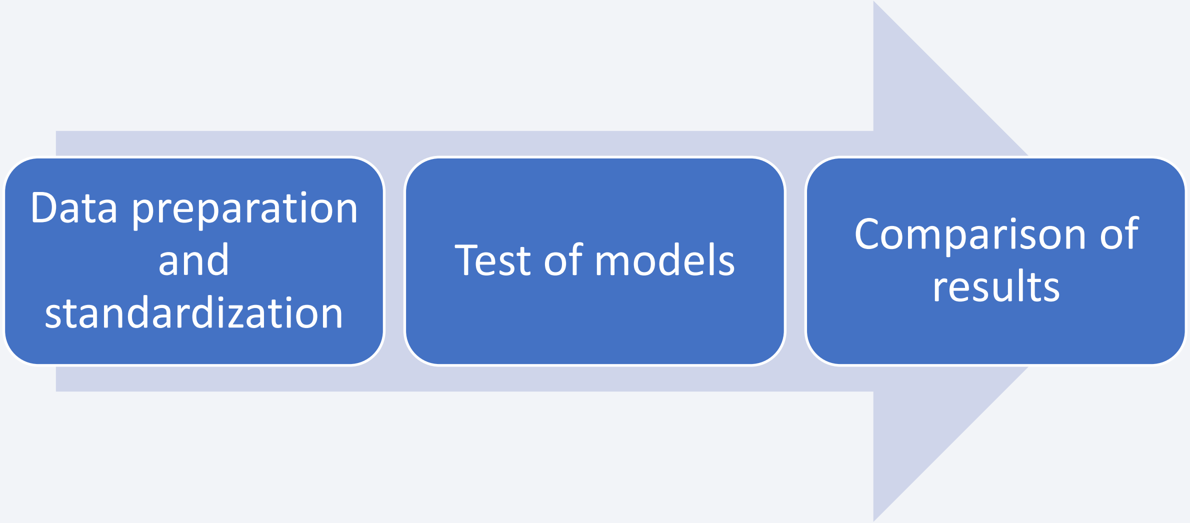
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data:
  - Percentage of launches by site
  - Payload range
- This combination allowed us to quickly analyze the relation between payloads and launch sites, which helped to identify the best place to launch according to payloads.

- https://github.com/makleen81/Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- We compared four classification models: logistic regression, support vector machine, decision tree, and k nearest neighbors.



- https://github.com/makleen81/Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

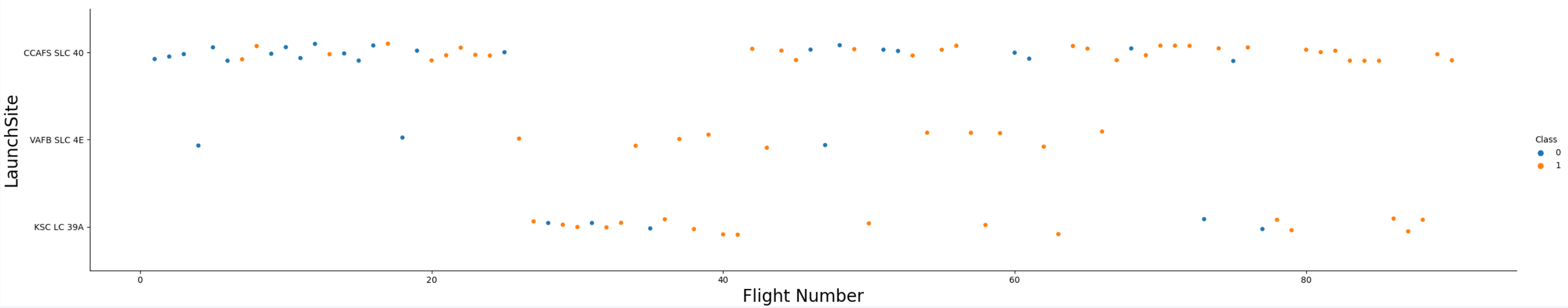
- Space X uses 4 different launch sites.
- The first launches were done to Space X itself and NASA.
- The average payload of F9 v1.1 booster is 2,928 kg.
- The first success landing outcome happened in 2015, five years after the first launch.
- Falcon 9 booster versions were successful at landing in drone ships having payload above the average,
- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015,
- Successful landings increased year over year.

Section 2

Insights drawn from EDA

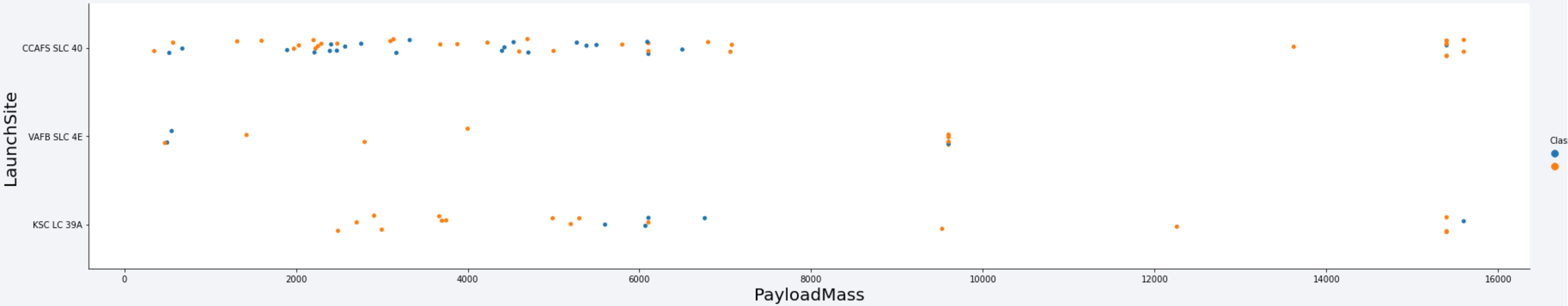
Flight Number vs. Launch Site

- The plot demonstrates that the best launch site is CCAF5 SLC 40, where a majority of recent launches were successful.
- KSC LC 39A is the launch site least conductive to success.
- We can see that the general success rate improved over time.

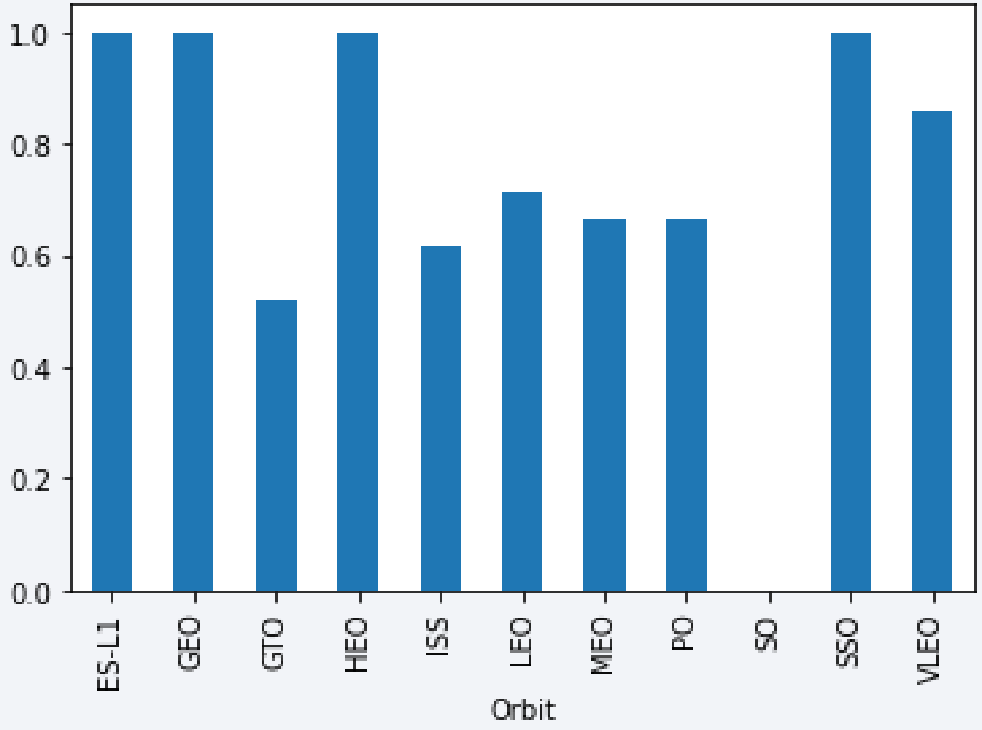


Payload vs. Launch Site

- Payloads over 9,000kg have a high success rate.
- Payloads over 12,000kg are mainly successful on CCAFS SLC 40 and KSC LC 39A launch sites.



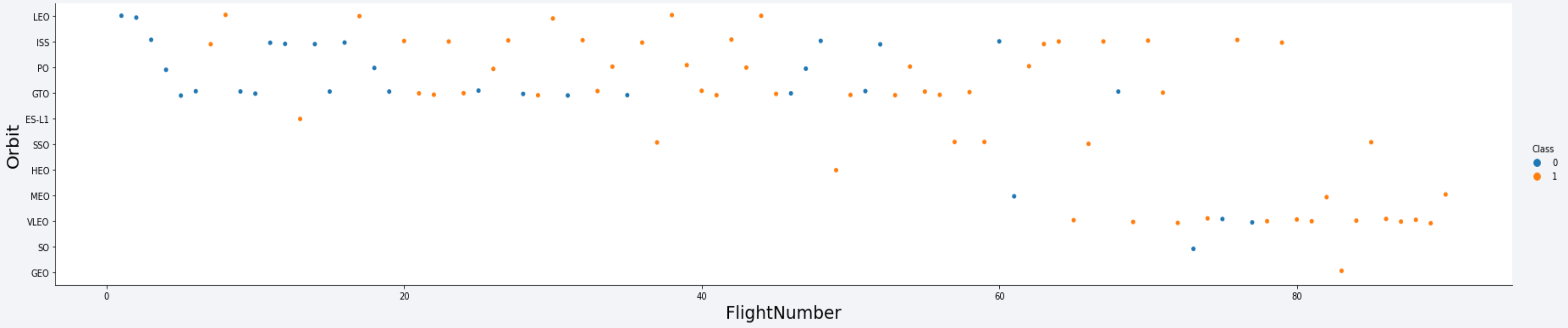
Success Rate vs. Orbit Type



- Success rates are highest for orbits ES-L1, GEO, HEO, and SSO.
- SO was a nonstarter, or experienced zero success.
- GTO was the second least-successful orbit.

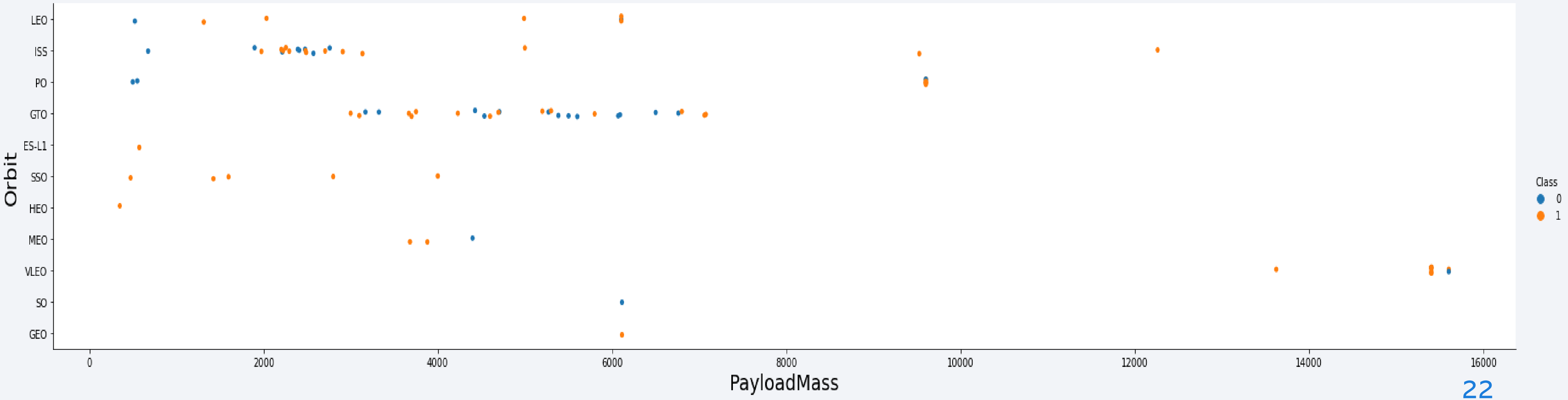
Flight Number vs. Orbit Type

- Success rates improved over time to all orbits
- VLEO represented a successful new orbit



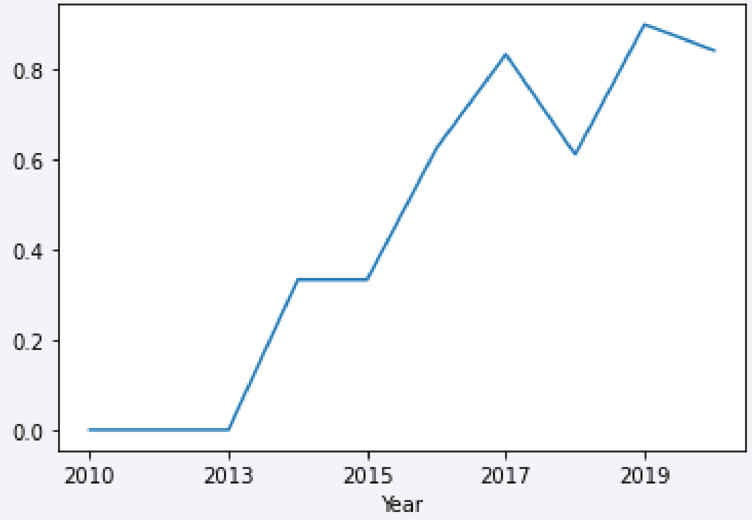
Payload vs. Orbit Type

- There is no relation between payload and success rate to orbit GTO.
- ISS orbit has the widest range of payload and a good rate of success.



Launch Success Yearly Trend

- Success rates increased in 2013 and continued to climb.
- Each successive year saw increases in technology and experience, leading to more positive outcomes.



All Launch Site Names

- There are four launch sites: CCAFS LC-40; CCAFS SLC-40; KSC LC-39A; and VAFB SLC-4E
- They were obtained by selecting unique occurrences of “launch_site” values from the dataset.

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

Launch Site Names Begin with 'CCA'

- Displayed are five records where launch sites begin with `CCA` - These are Cape Canaveral launches

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcor
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachu
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 (parachu
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attem
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attem
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attem

Total Payload Mass

- The total payload carried by boosters from NASA is 111,268
- Total payload was calculated by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

total_payload
111268

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2,928 kg
- This was obtained by filtering data by the booster version above and calculating the average payload mass.

avg_payload
2928

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was December 22, 2015.
- This was obtained by filtering the data by successful landing outcome on a ground pad and getting the minimum value for date.

first_success_gp
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

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

- This was obtained by selecting distinct booster versions according to the filters described above.

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes:

mission_outcome	qty
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- This was obtained by grouping mission outcomes and counting records for each group.

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass:

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- This was obtained by utilizing a subquery for boosters which have carried the maximum payload mass registered in the dataset.

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015:

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

- This was obtained by querying the booster version and launch site where the landing outcome = Failure, and the date.

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

- Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order:

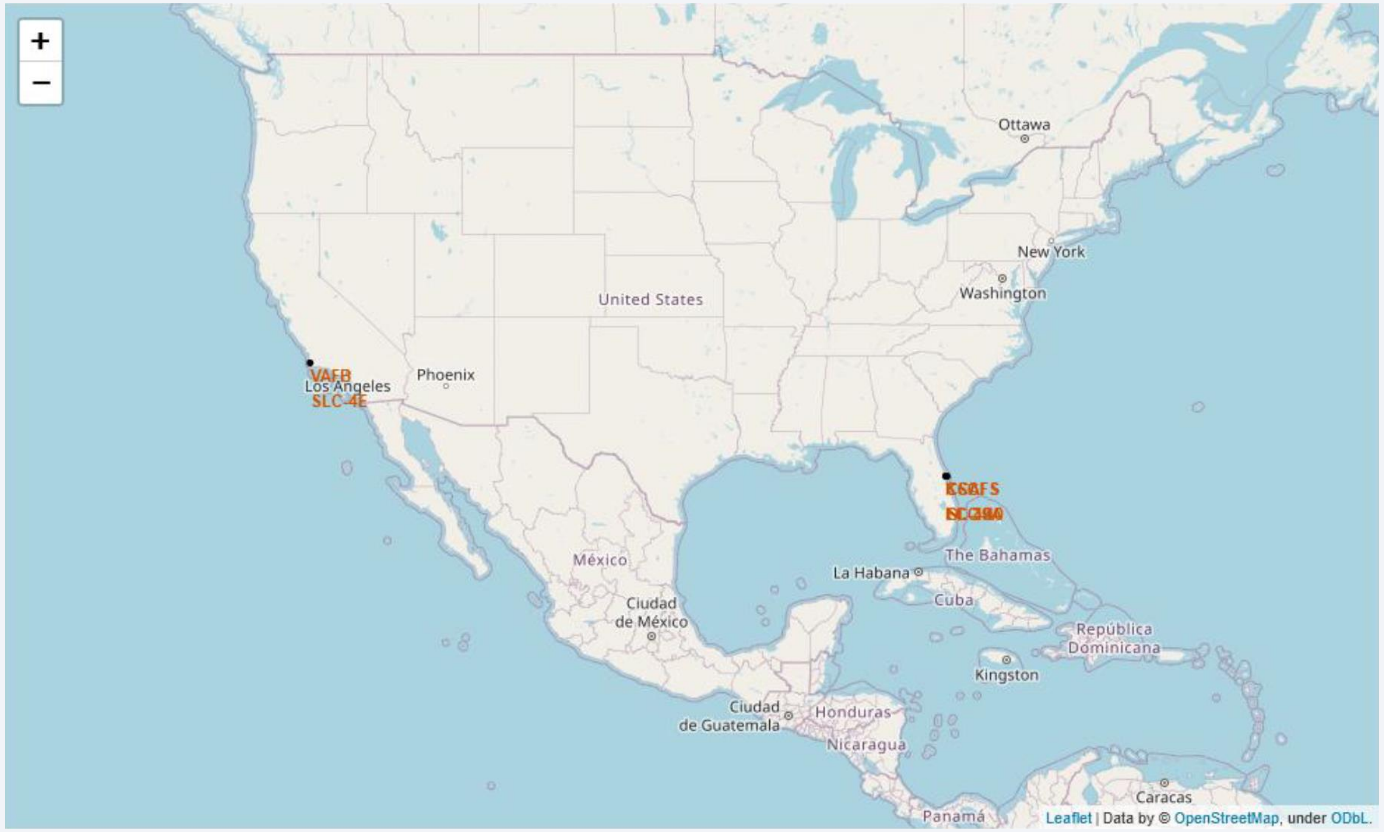
landing_outcome	qty
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- The data shows that “no attempt” accounts for a majority of the landing outcomes between those dates.

Section 3

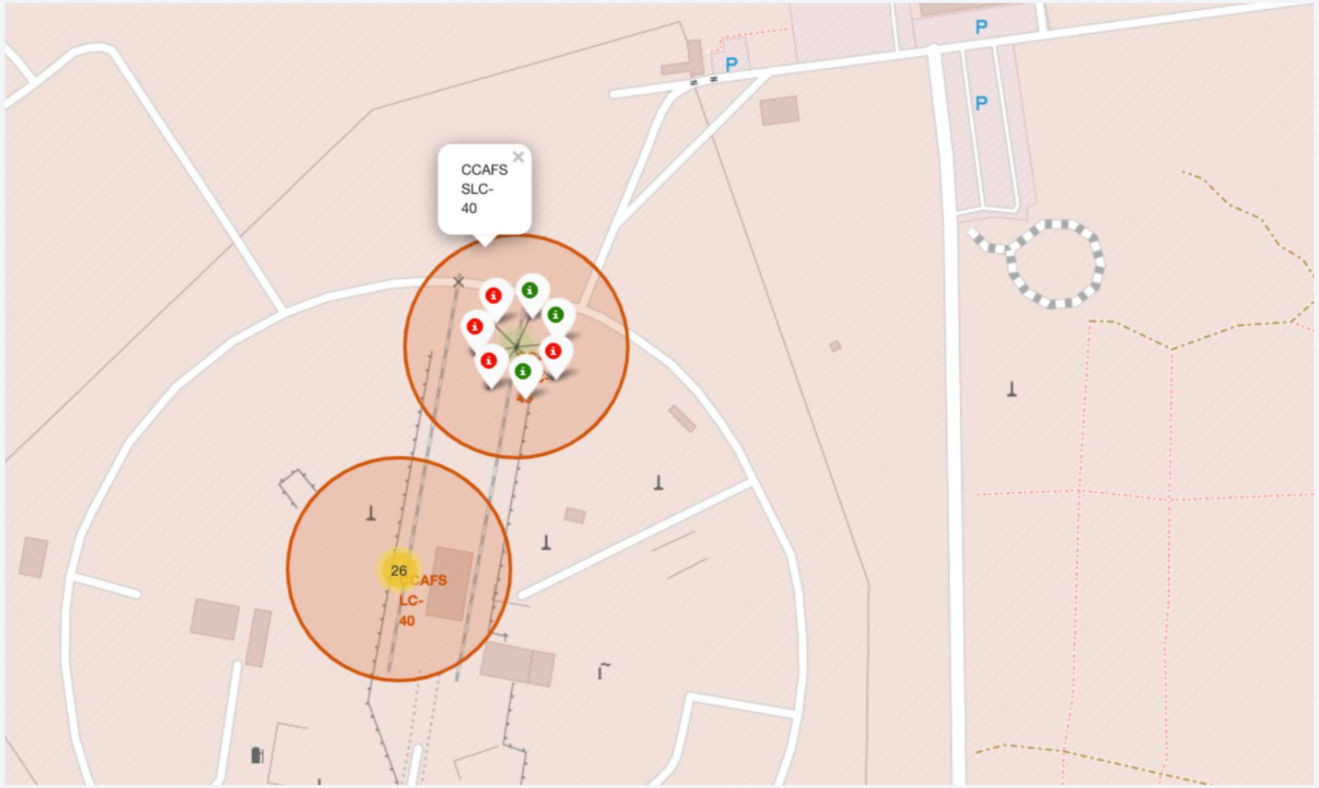
Launch Sites Proximities Analysis

All Launch Sites



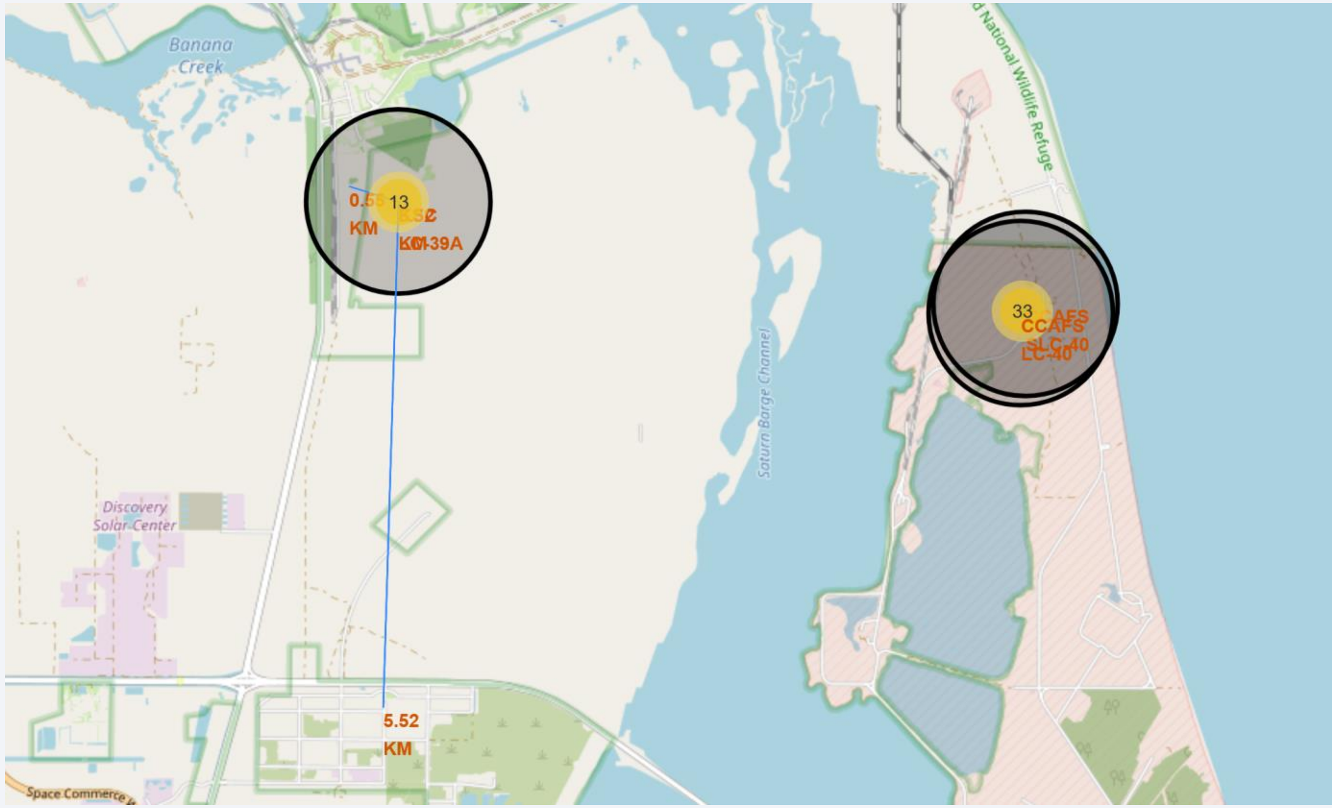
- The launch sites are located in two clusters, one on the west coast and one on the east coast, located near major infrastructure hubs.

Launch Outcomes by Site



- Green markers indicate a successful launch and red markers indicate an unsuccessful launch.

Infrastructure

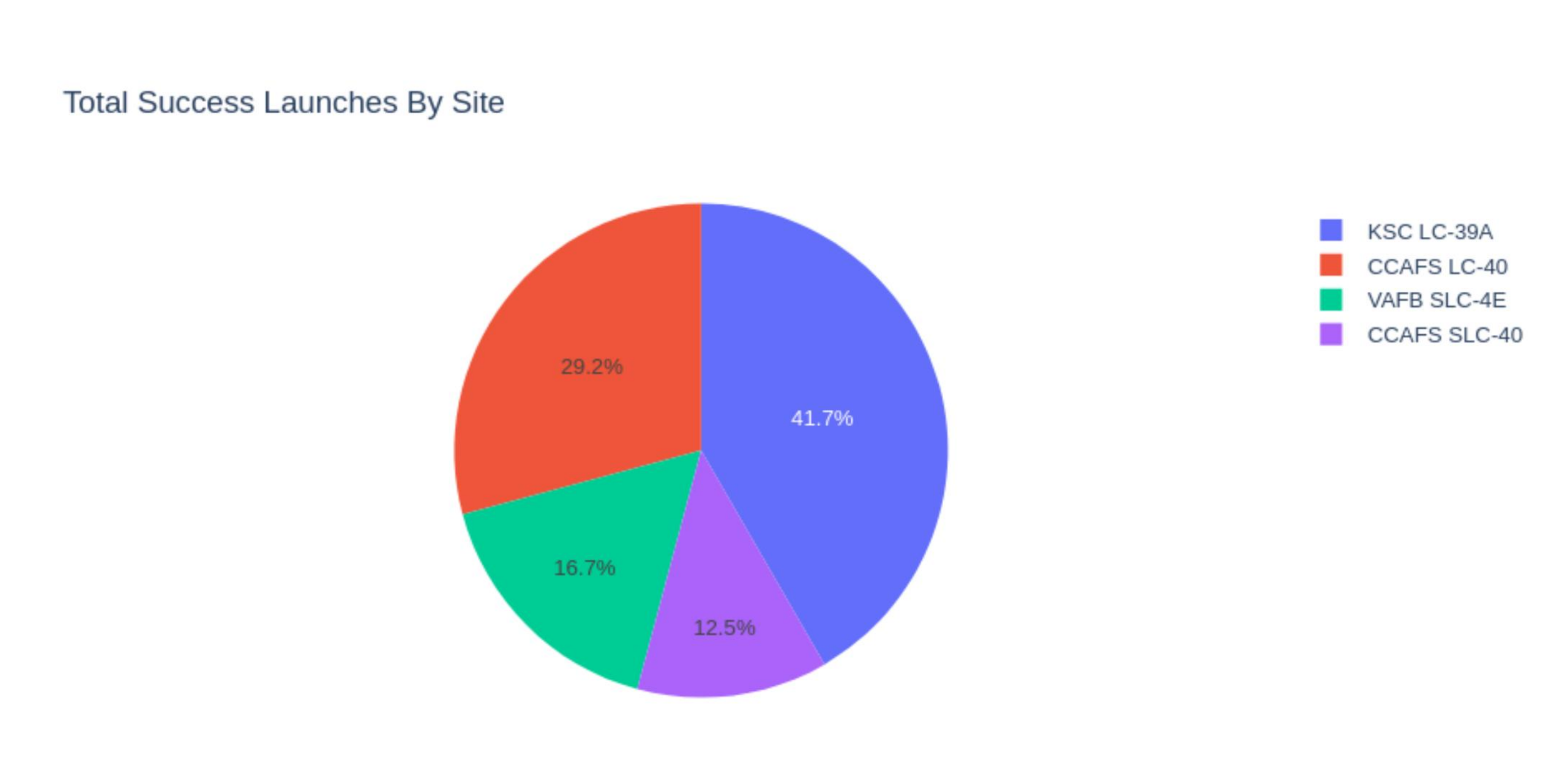


- Launch site KSC LC-39A has access to key infrastructure like railroads and highways. The closest populated area is 5.52km away.

Section 4

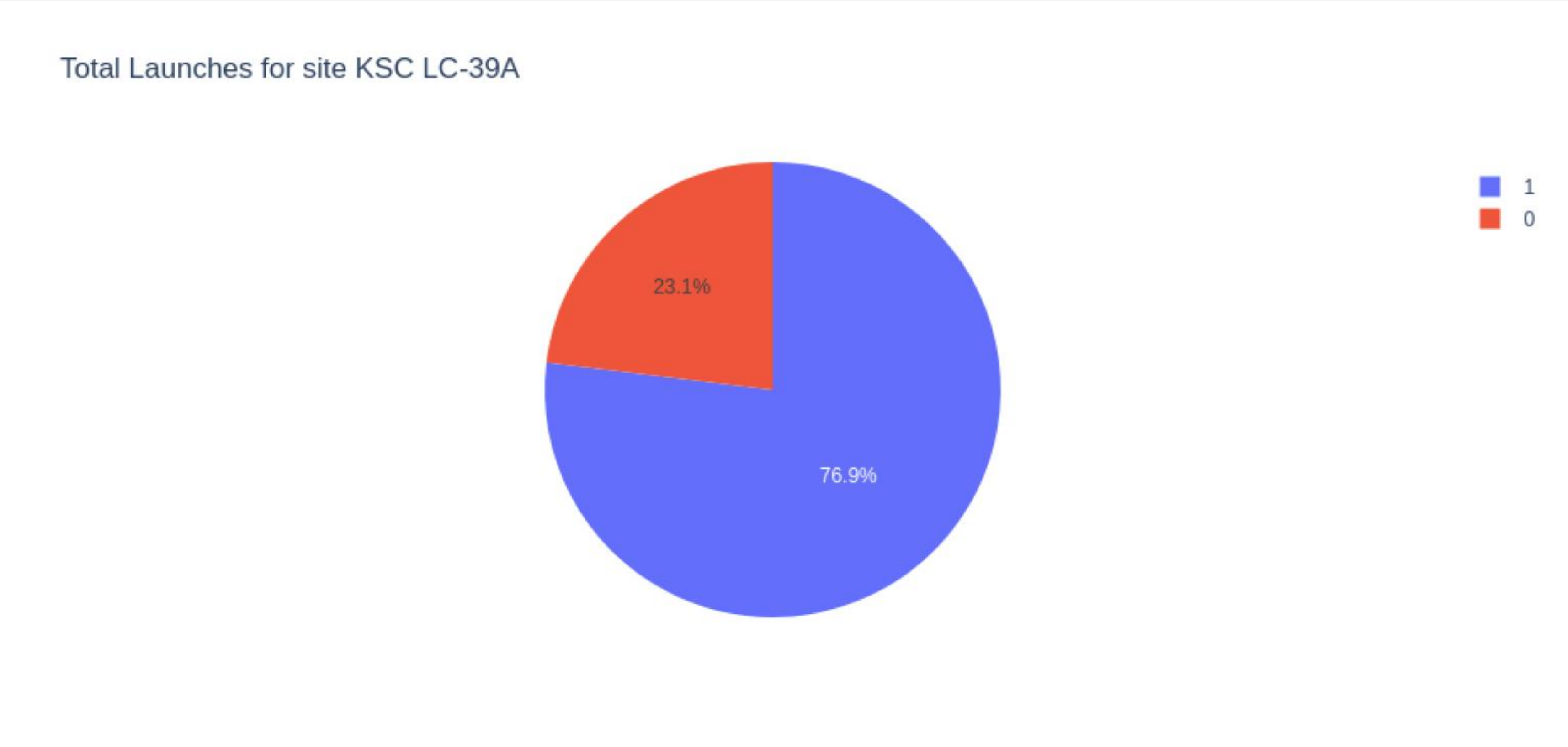
Build a Dashboard with Plotly Dash

Launch Success Count



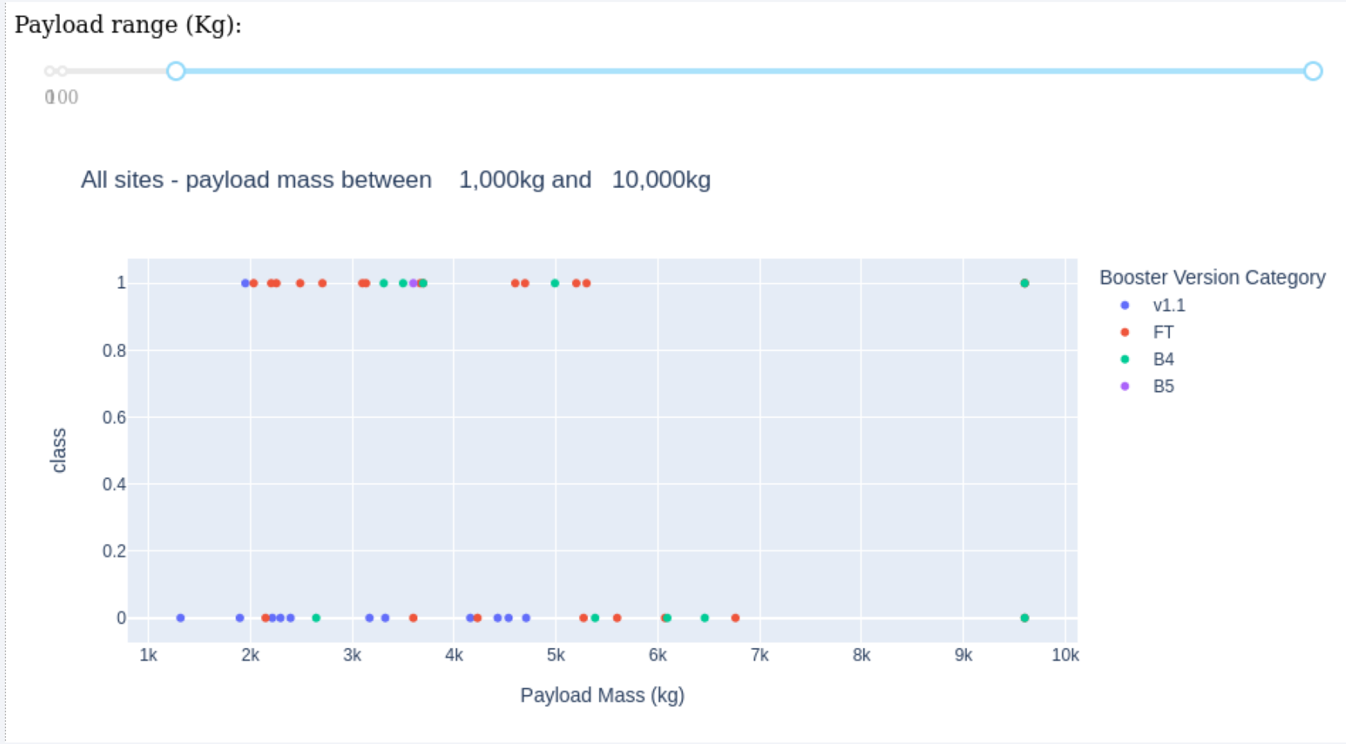
- This chart shows that launch site KSC LC-39A has the highest overall success rate, while launch site CCAFS SLC-40 is the least successful.

Launch Site with Highest Launch Success



- This chart shows that site KSC LC-39A has a 76.9% launch success rate.

Payload vs. Launch Outcome



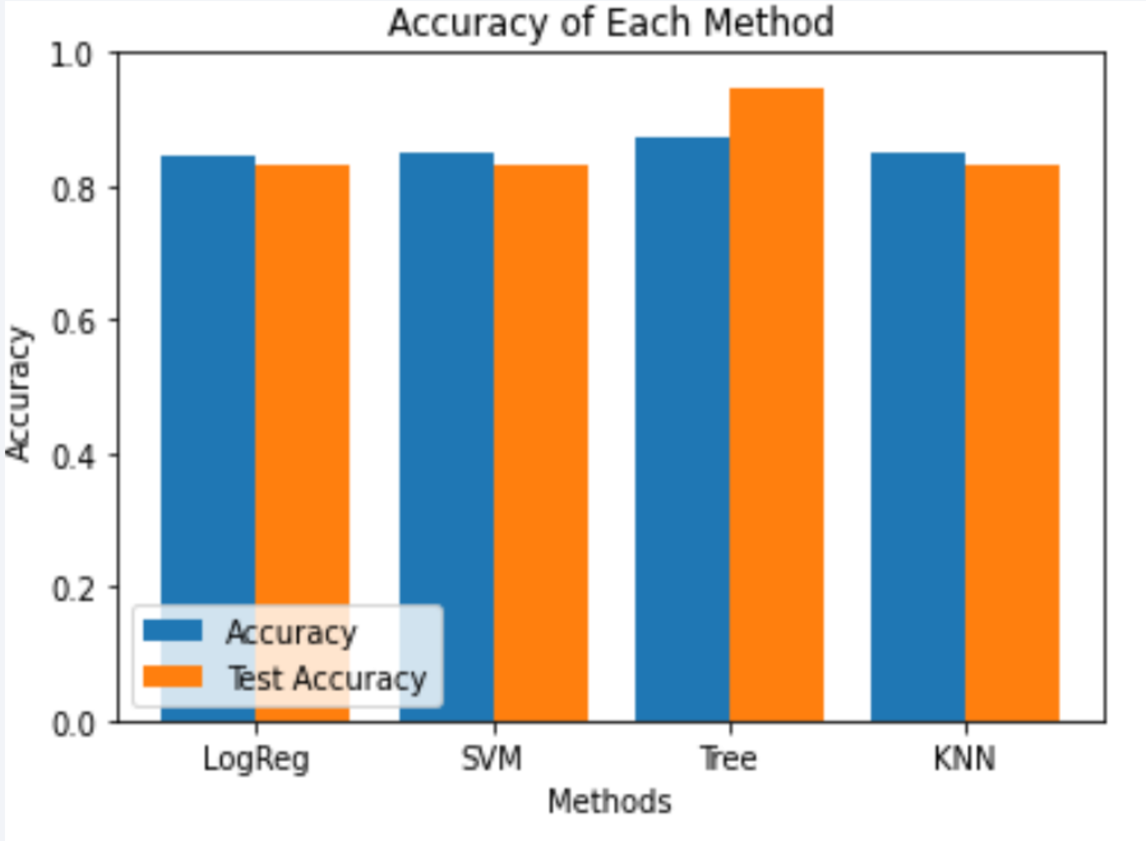
- This chart shows that the booster version category FT has the highest success rate, while v1.1 is the least successful.

Section 5

Predictive Analysis (Classification)

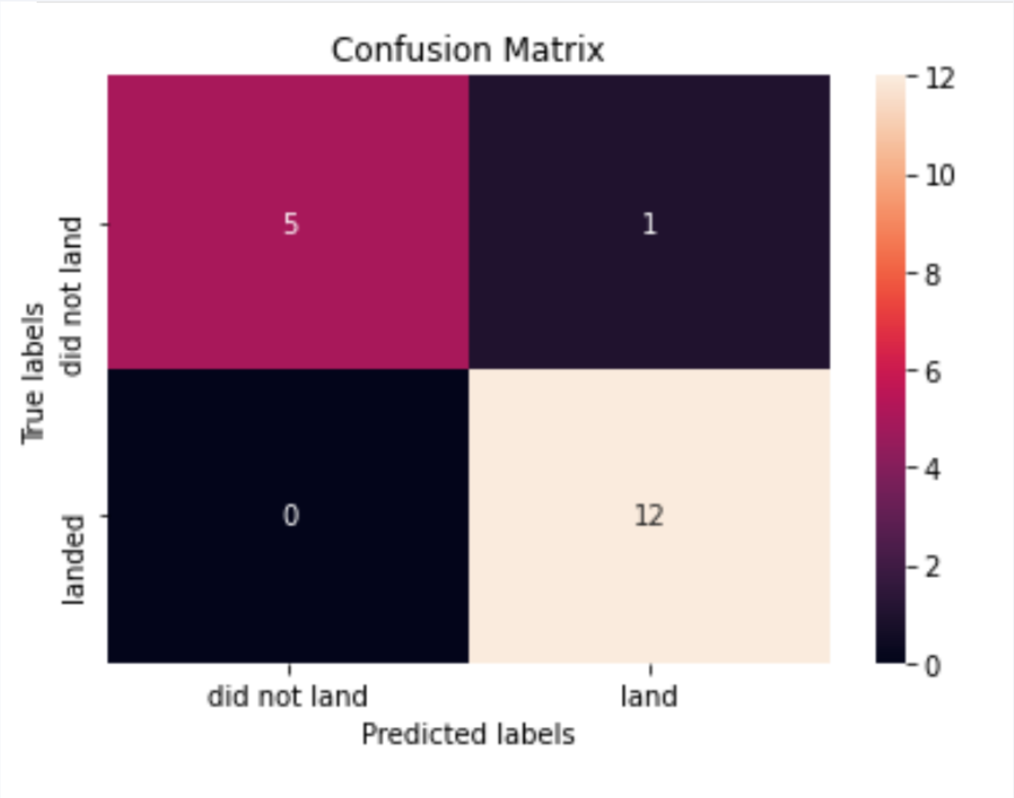
Classification Accuracy

- Of the four classification models, Tree had both the highest accuracy and test accuracy.
- LogReg, SVM, and KNN, all produced similar levels of accuracy.



Confusion Matrix

- According to the Confusion Matrix, the Decision Tree Classifier demonstrates its accuracy by showing the highest number of true positive and true negative results.



Conclusions

- In this analysis, we looked at different data sources and refined our conclusions along the way.
- We found that KSC LC-39A is the best launch site.
- Our analysis also found that launches above 7,000kg are less risky.
- Successful landing outcomes have improved over time with the evolution and refinement of technology.
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Thank you!