IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Steven Cheng
September 2025

Outline

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

Executive Summary

- Data was collected from the SpaceX API and by web scraping
- Exploratory data analysis using visualizations and SQL were used to identify variables that could predict launch success outcome
- Machine learning was used to train models and determine a method that could best predict launch success outcome
- Summary of all results
  - Orbit type, launch site, and payload mass are important variables that influence rocket launch outcome
  - The KNN tree classification model can predict launch success outcome with 83% accuracy

<This is original work>
<This is original work>

Introduction

- SpaceY is an aerospace company, focused on commercial space travel
- Rocket launches are expensive, estimated to cost upwards of $165 million per launch; SpaceX, a competitor, advertises that the Falcon 9 rocket launch can cost $62 million
- SpaceX can reduce costs in the rocket launch because the Falcon 9 rocket can recover the first stage of a rocket and be reused, however, the first stage of the rocket will not always successfully land or be reused due to mission parameters
- Goal: predict the price of a rocket launch, and use public information and machine learning to predict whether SpaceX will reuse the first stage of a rocket launch based on mission parameters

<This is original work>
<This is original work>

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX public launch data is extracted using the SpaceX REST API and web scraping methods
- Perform data wrangling

  - Data was processed using Python to extract out attributes that will help determine the outcome of rocket launch reuse
- Perform exploratory data analysis (EDA) using visualization and SQL

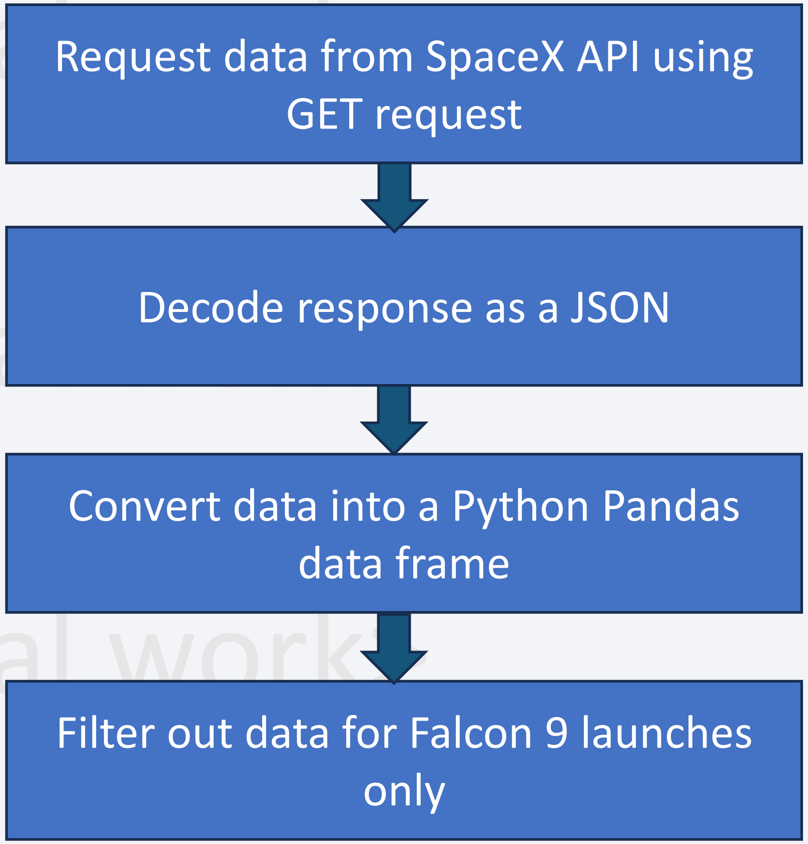
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Variables were chosen for classification models and the models were trained and tested for accuracy

Data Collection

- Datasets were collected from the SpaceX API and by scraping data from the web
- The data was analyzed and organized into data frames that were cleaned and processed for data analysis

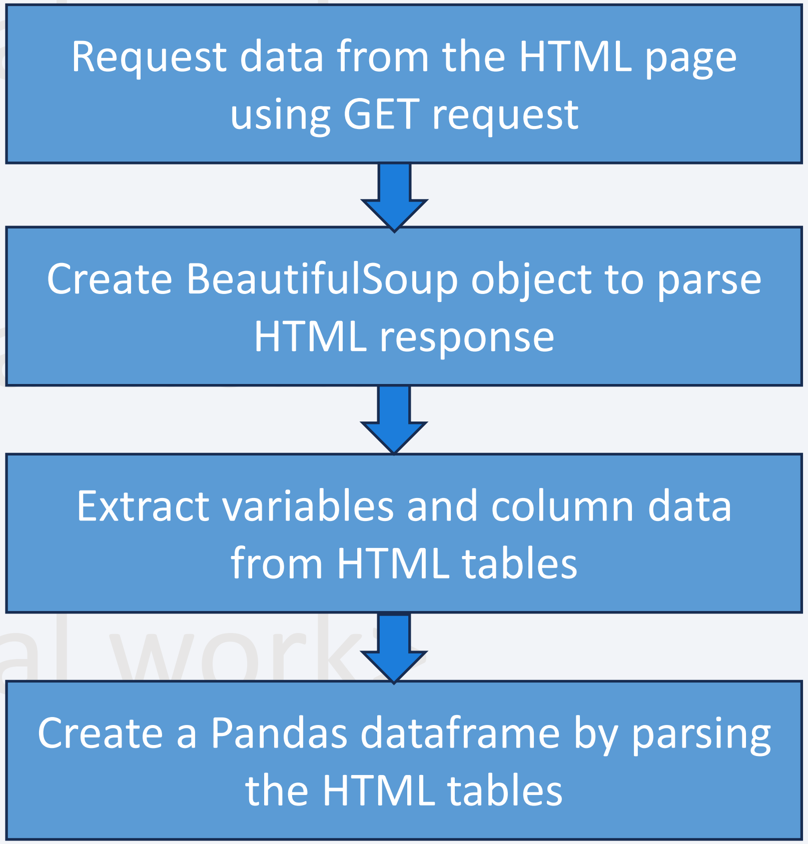
Data Collection – SpaceX API

- SpaceX REST API contains data detailing the launch history of all of the SpaceX rockets
- Clean the requested data to extract Falcon 9 launch data and retrieve parameters that can determine mission and rocket reuse outcome
- See https://github.com/syc9/data/blob/main/final/jupyter-labs-spacex-data-collection-api.ipynb for complete data collection



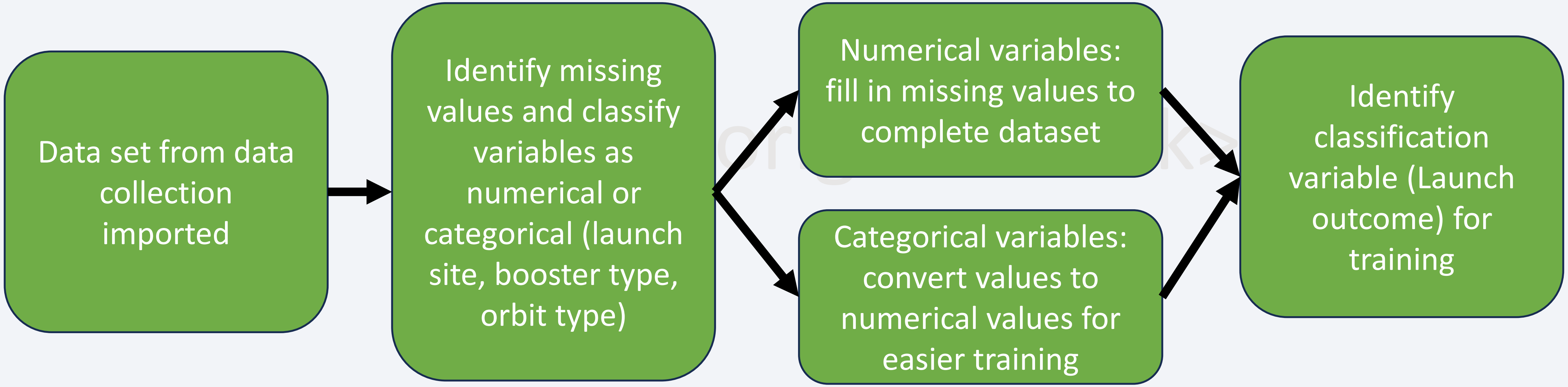
Data Collection - Scraping

- Web scraping was performed to extract Falcon 9 historical launch records from a Wikipedia page “List of Falcon 9 and Falcon Heavy launches
- Data was parsed from the table and extracted to a dataframe for cleaning and processing
- See https://github.com/syc9/data/blob/main/final/jupyter-labs-webscraping.ipynb for complete scraping



Data Wrangling

- Before Exploratory Data Analysis, missing values and incomplete values need to be resolved
- Data types and parameters need to be classified to be used to determine training labels for further analysis and machine learning use
- See https://github.com/syc9/data/blob/main/final/labs-jupyter-spacex-Data%20wrangling.ipynb for complete data wrangling process



EDA with Data Visualization

- Bar charts, line charts, and scatter charts were used to better visualize the relationship between various parameters, including flight number, payload mass, orbit type, and launch site

- Scatter charts were used to try and determine relationships between flight number, payload mass, and launch sites
- Bar charts were used to try and determine relationships between flight number and orbit type

- Line charts were used to visualize the success rate of launches year by year
- See https://github.com/syc9/data/blob/main/final/edadataviz.ipynb for full data visualization notebook

EDA with SQL

- With SQL, the following data was extracted:
    - Unique launch sites in the space mission
    - The total payload mass carried by boosters launched by NASA
    - Average payload mass carried by booster version F9 v1.1
    - The first successful landing outcome in ground pad was achieved
    - The names of the boosters which have success in drone ship and had payload mass between 4000 and 6000 kg
    - The total number of successful and failure mission outcomes
    - The booster versions that carried the maximum payload mass
    - The rank of landing outcomes
- See https://github.com/syc9/data/blob/main/final/jupyter-labs-eda-sql-coursera_sqllite.ipynb for full SQL notebook

<This is original work>

<This is original work>

Build an Interactive Map with Folium

- On the Folium map, the launch site and markers to indicate failed and successful rocket launches were added
- Lines and markers to indicate distances to certain features were added as well to highlight the strategic placement of the rocket launch sites
- See https://github.com/syc9/data/blob/main/final/lab_jupyter_launch_site_location.ipynb for full notebook

<This is original work>

<This is original work>

Build a Dashboard with Plotly Dash

- A SpaceX Launch Records dashboard was built using dash
- A pie chart showing the success launch rate by site and a scatter plot showing the correlation between payload and success rate by launch site is displayed

- The two charts give an interactive visual showing how launch site and payload range affect launch success rate
- See https://github.com/syc9/data/blob/main/final/spacex-dash-app.py for the code to run the Dash application

Predictive Analysis (Classification)

- The dataset was created by identifying the classification variable, transforming the variables and standardizing the data, and splitting the data to training data and testing data

- The best hyperparmeters were determined for the SVM, classification tree, and logistic regression model by training the model on the training dataset,

- The hyperparmeters were used to evaluating the performance of the model by running the model on the testing dataset and determining its accuracy

- See https://github.com/syc9/data/blob/main/final/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb for the model development and results

Results

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

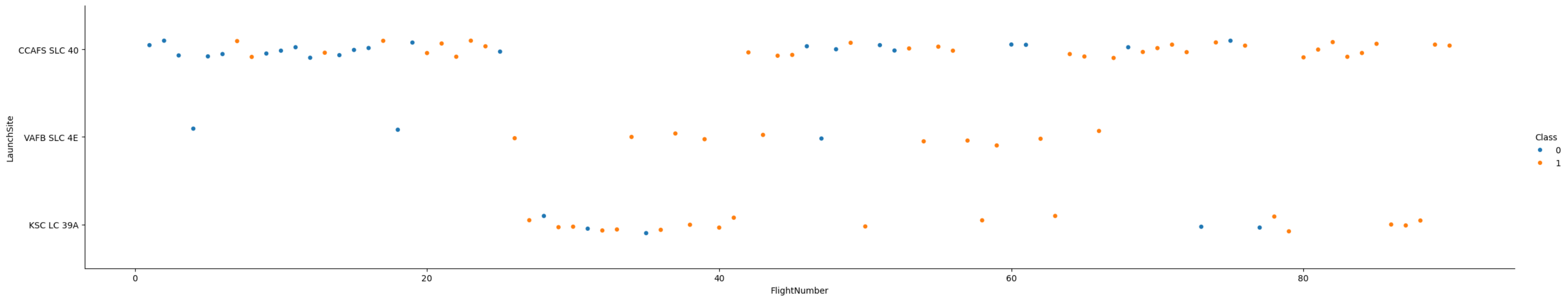
Section 2

Insights drawn from EDA

<This is original work>

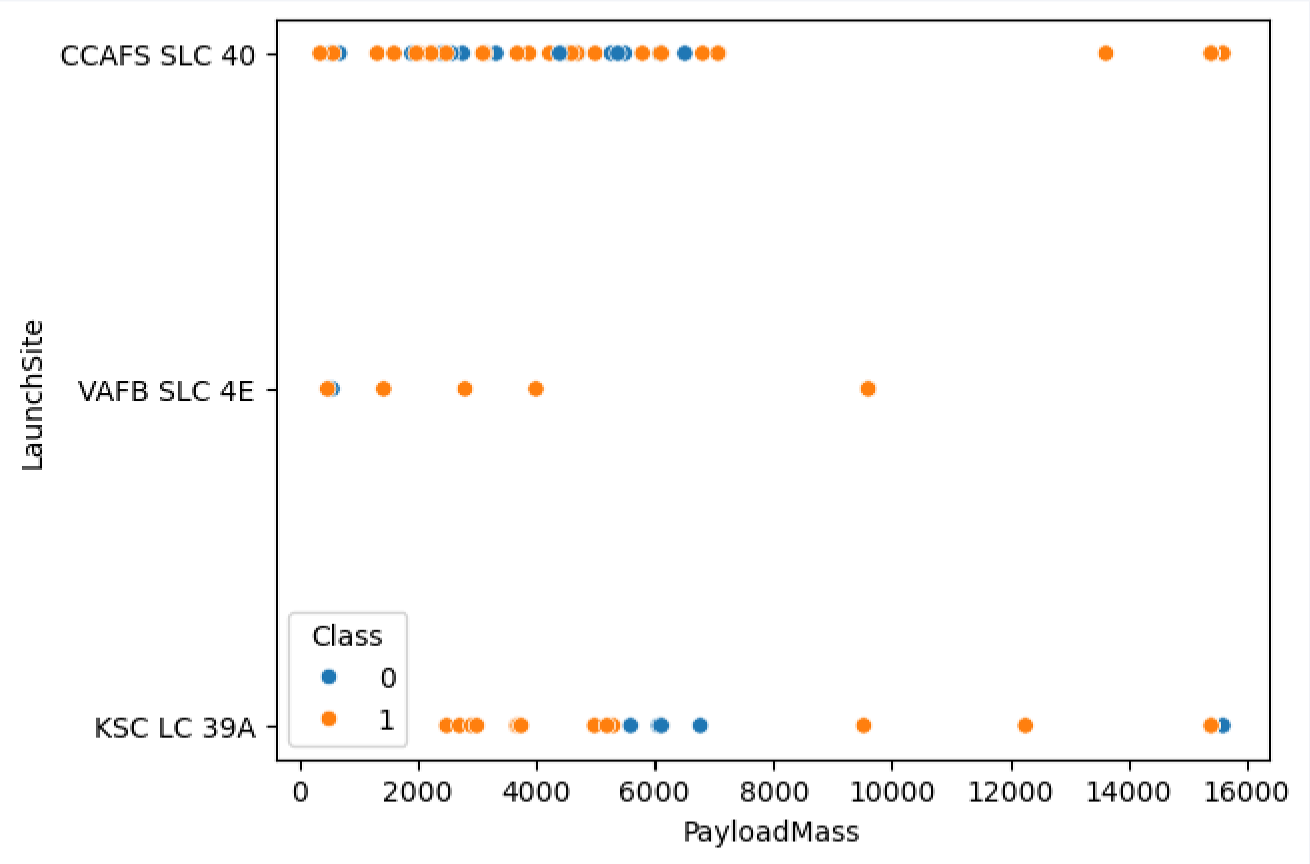
<This is original work>

Flight Number vs. Launch Site



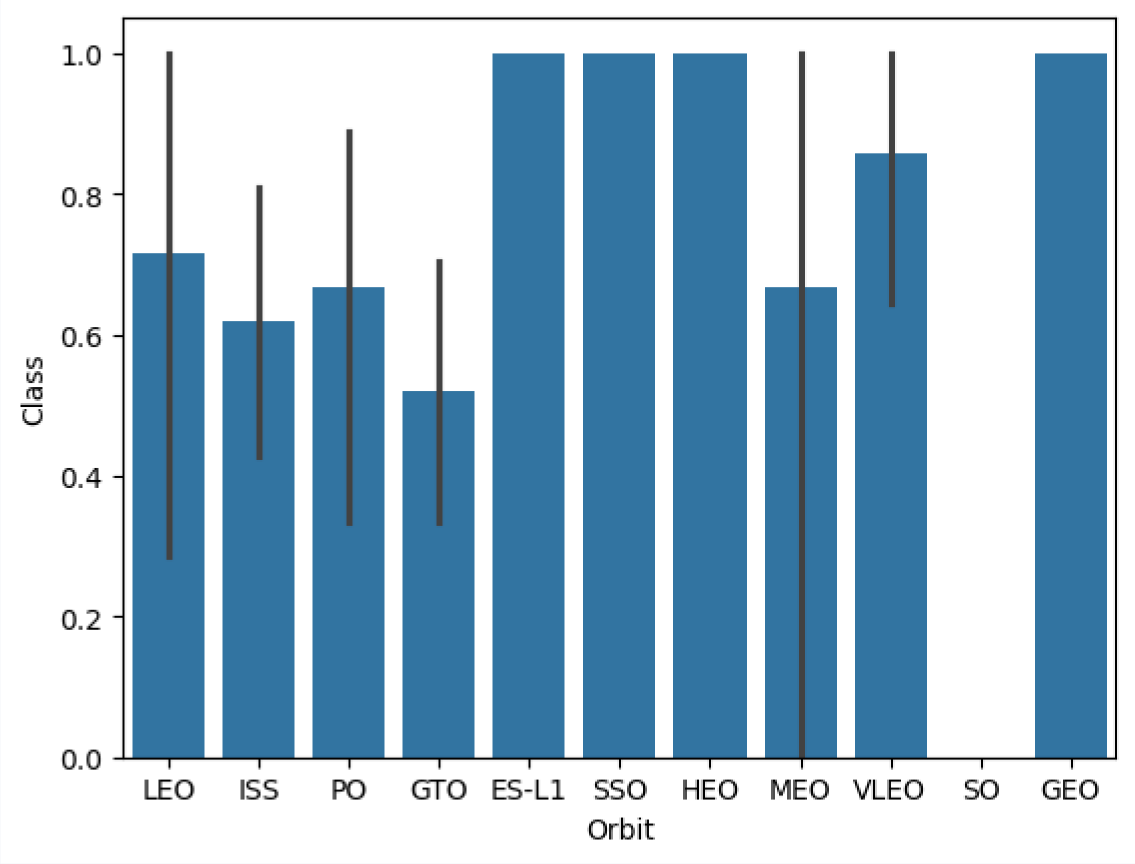
Success outcomes (orange) are more frequent at all launch sites as more flights were deployed (right), suggesting a higher success rate with more launches at all launch sites

Payload vs. Launch Site



- No clear trend is found with success rate at any of the launch sites, suggesting that payload mass is not a strong indicator of launch success outcome

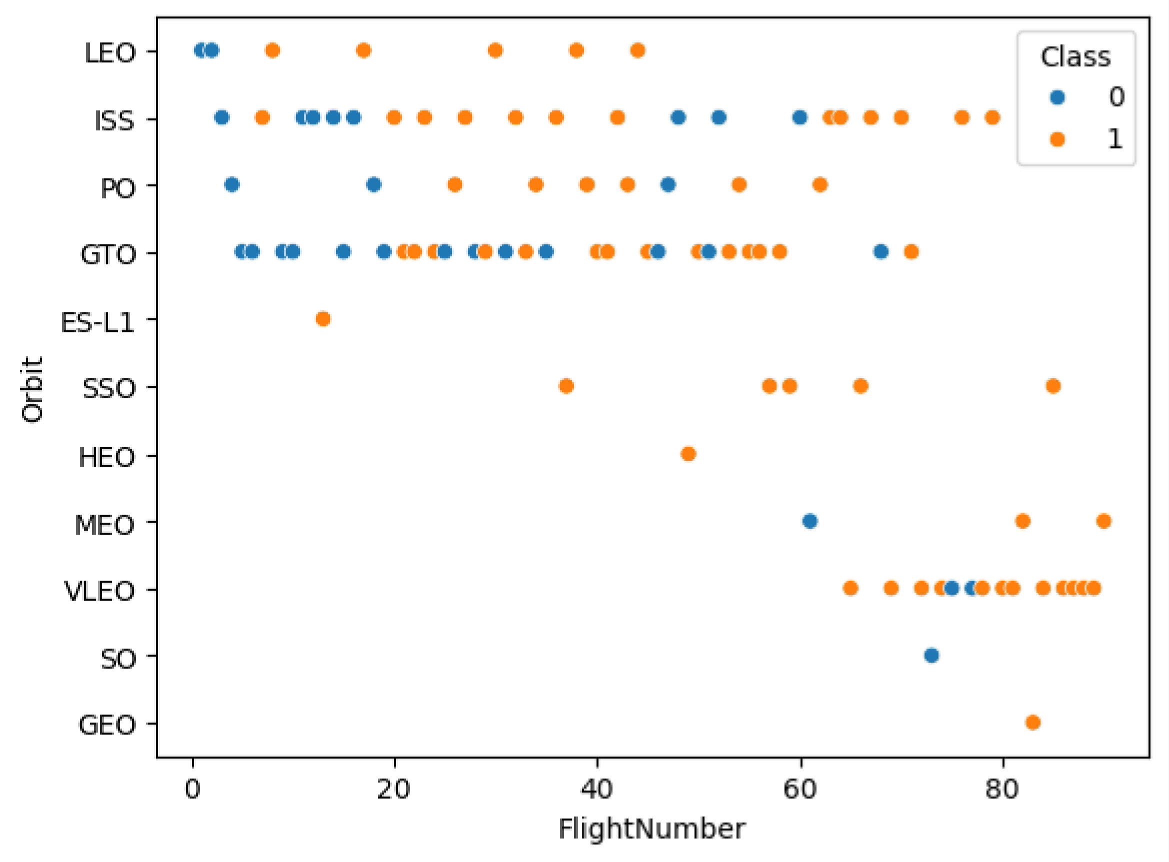
Success Rate vs. Orbit Type



- The ES-L1, SSO, HEO, and GEO orbit types have the highest success rate

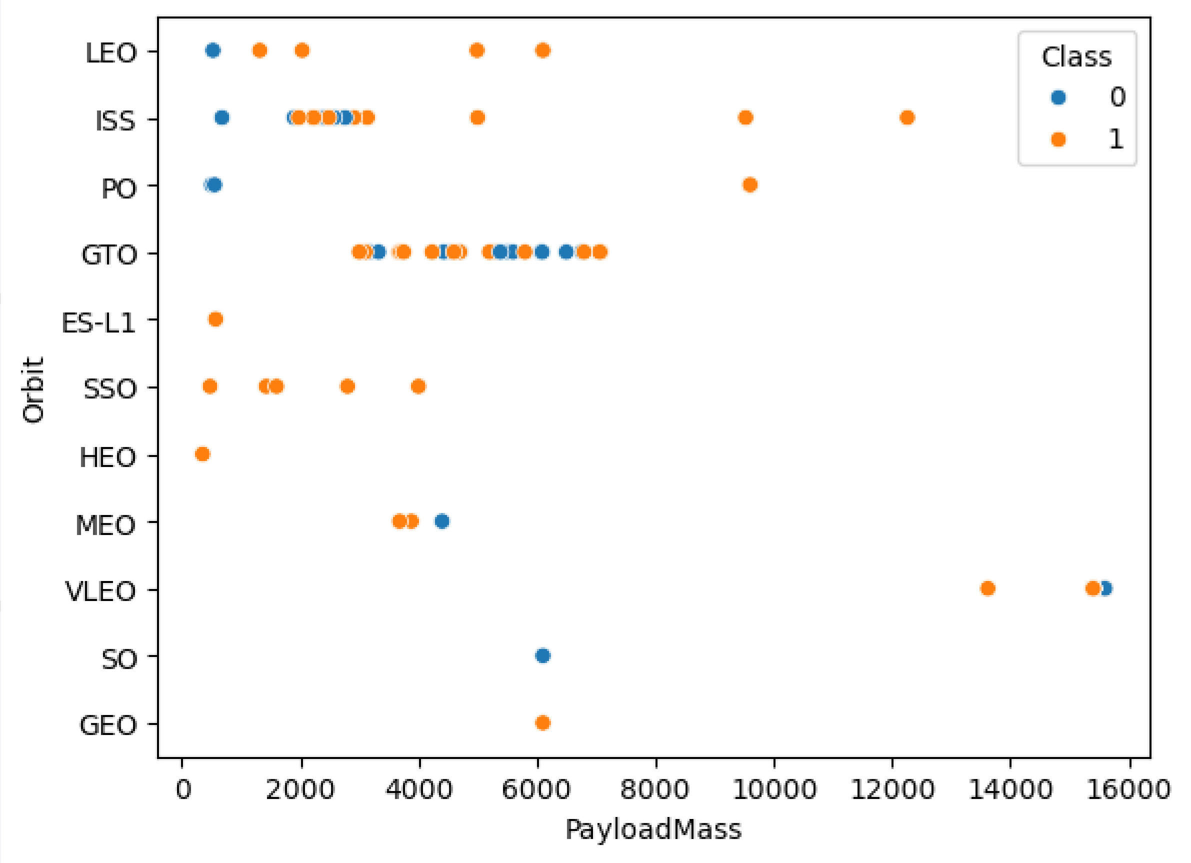
Flight Number vs. Orbit Type

- In the LEO orbit, success sems to be related to the number of flights
- In the GTO orbit, no relationship appears to exist between flight number and success



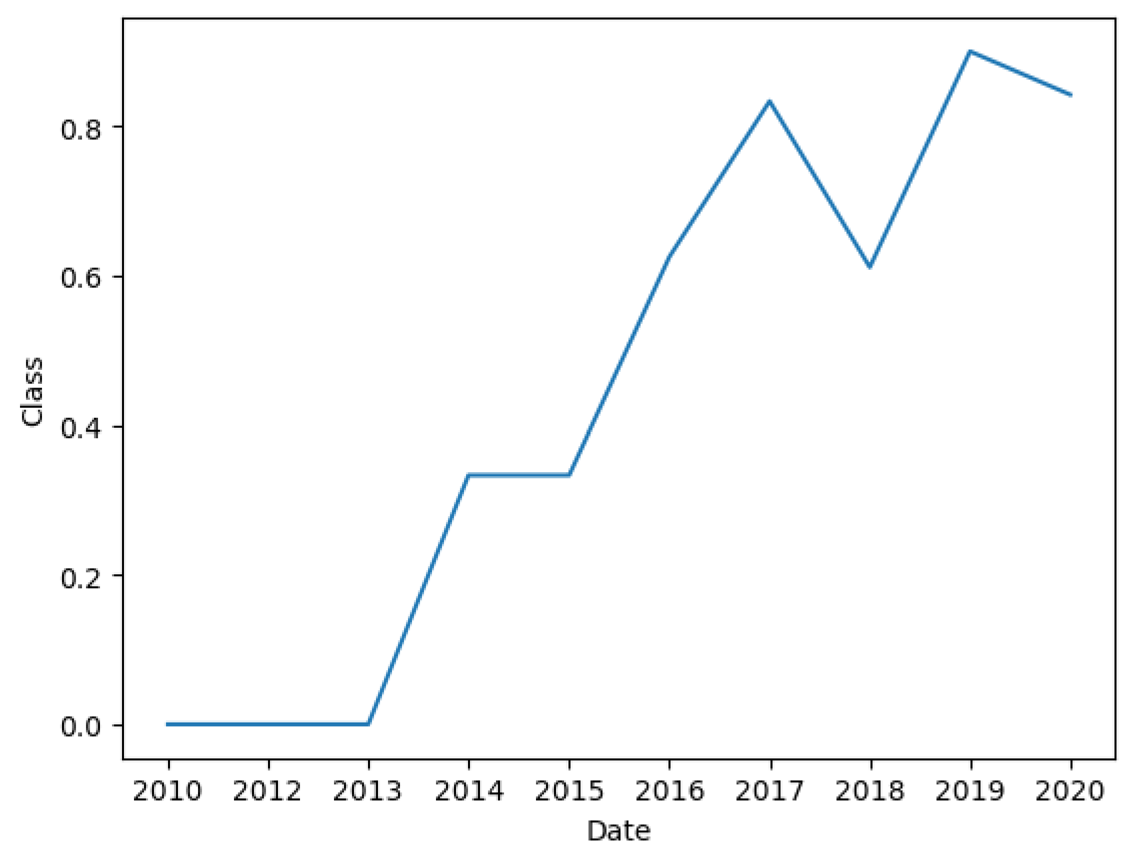
Payload vs. Orbit Type

- With heavy payloads, success landings or positive landing rate are more for PO, LEO and ISS orbit types
- For GTO, it is difficult to distinguish successful and unsuccessful landings with payload mass



Launch Success Yearly Trend

- The success rate of launches since 2013 has increased until 2020



All Launch Site Names

- From the table, display the names of the unique launch sites used in the space missions
- 4 unique launch sites were used, listed below

<This is original work>

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

<This is original work>

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA’

- The first 5 missions all showed successful mission outcomes from 2 different customers, but thee landing outcome was either not attempted or failed

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

- Calculate the total payload carried by boosters from NASA
- The total payload mass carried by all the boosters from NASA is 45596 kg

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- The average payload mass in kg carried by the F9 v1.1 booster was 2928.4

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- The first successful landing outcome on ground pad was on December 22, 2015

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- 4 Booster versions had successful landings on drone ship and had payload mass between 4000 and 6000 kg

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Out of the 101 total missions considered, 100 had successful outcomes, and 1 failed in flight

<This is original work>

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- The max payload mass carried is 15600 kg, and the following booster versions all carried the max payload mass.

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- The month of January and April had a failed landing outcome on the drone ship, both at CCAFS LC-40 launch site

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

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

- Rank 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

- In terms of landing outcomes, after no landing attempt, drone ship was the next most used landing outcome

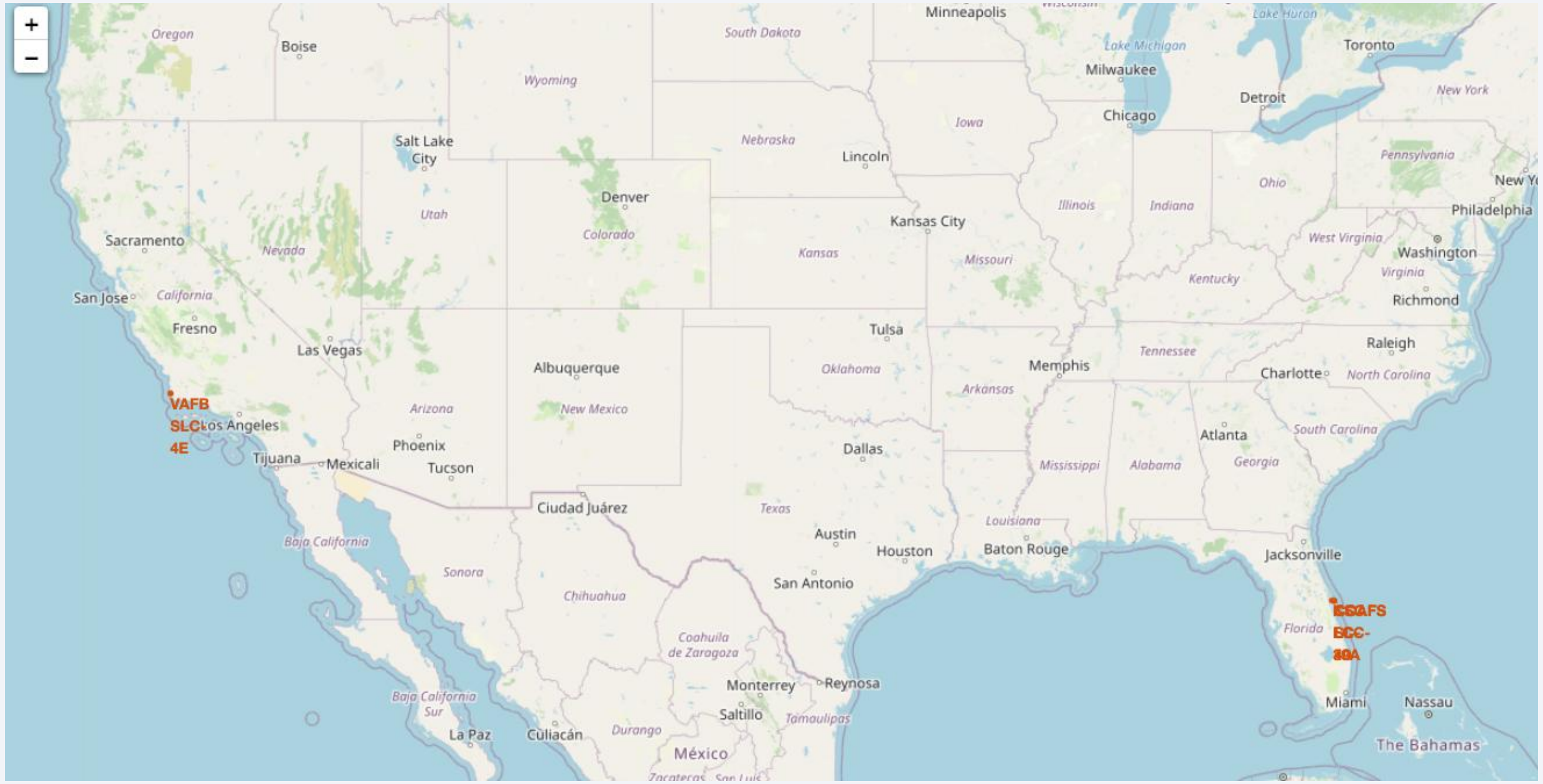
Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

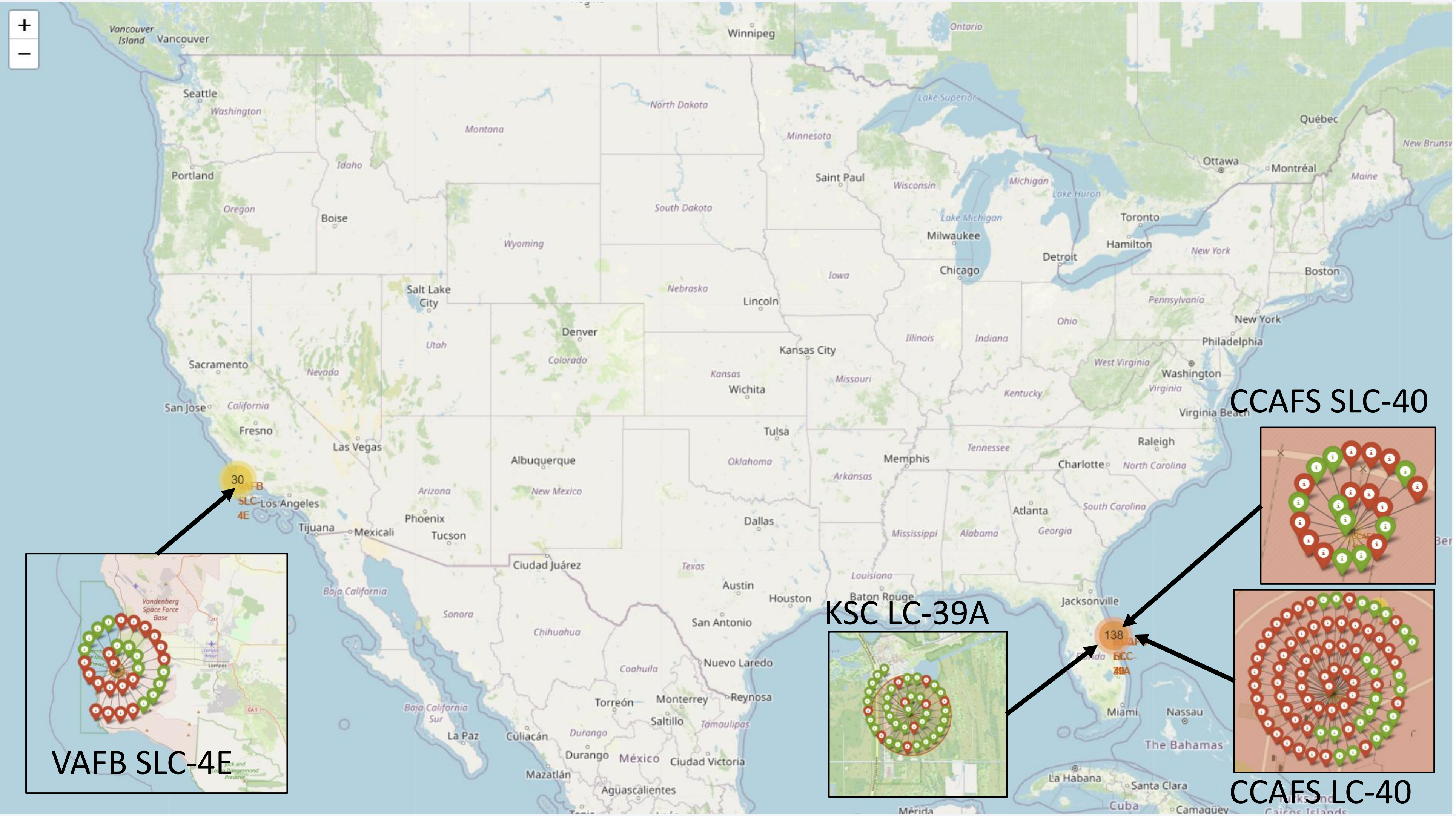
<This is original work>
<This is original work>

Location of Launch Sites used for Rocket Launches



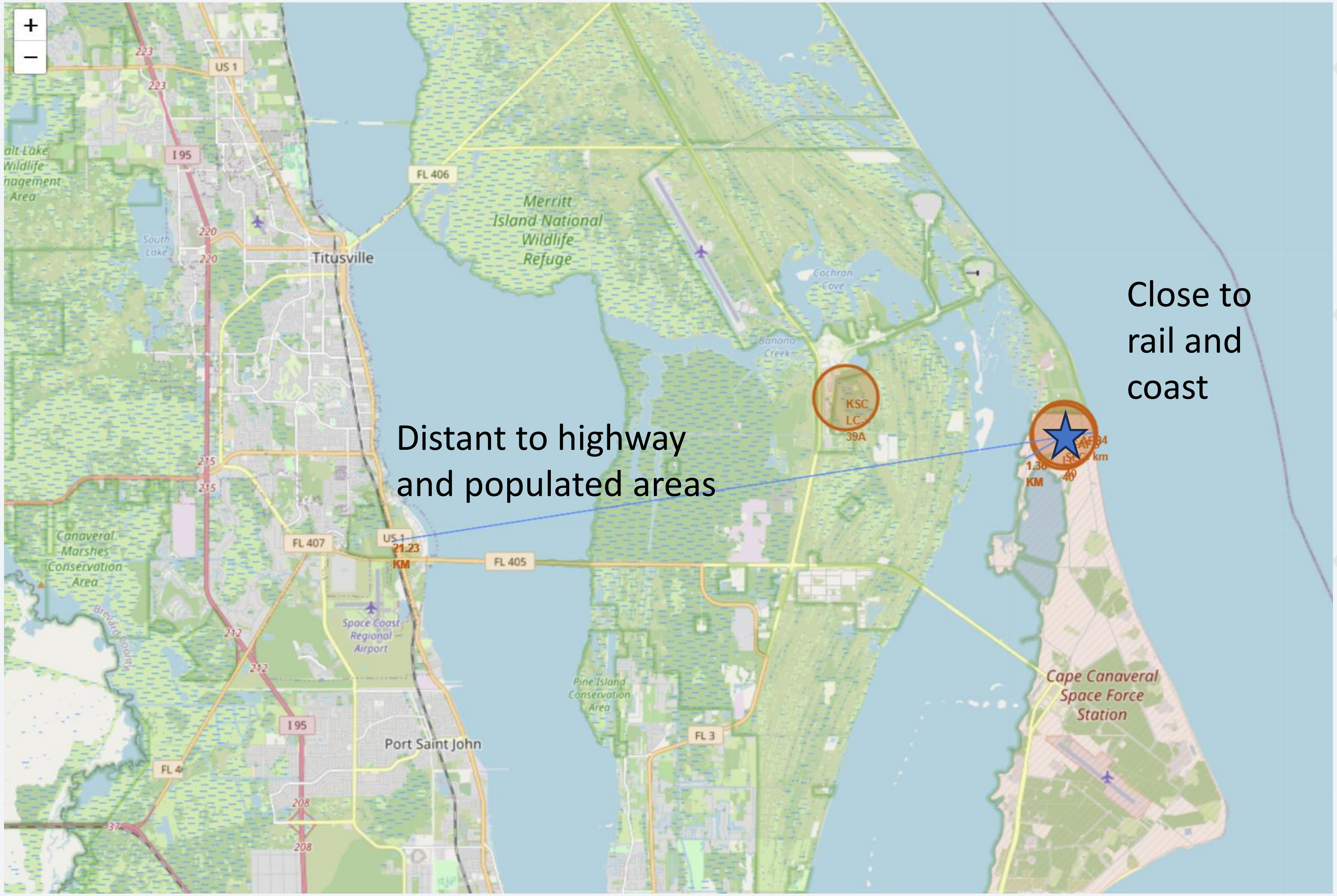
- 4 different launch sites were used, 3 located in Florida and the other one located in Southern California
- Launch sites are all located on established sites and are located close to large bodies of water

Launch Outcomes based on Launch Site



- At the 4 different Launch Sites, the mission outcomes are highlighted (green = success, red = failure)
- The KSC LC-39A site has the highest percentage of successful launch outcomes

Launch Site Proximity to Important Features



- At the CCAF launch site, the <1km distance to the railroad and coast is important, as the railroad allows for easier transport of the rocket and coast gives clear and open space for rocket launches
- The longer distance to highway and city, shown by the 21km to the closest populated area is also intended to be a safety precaution

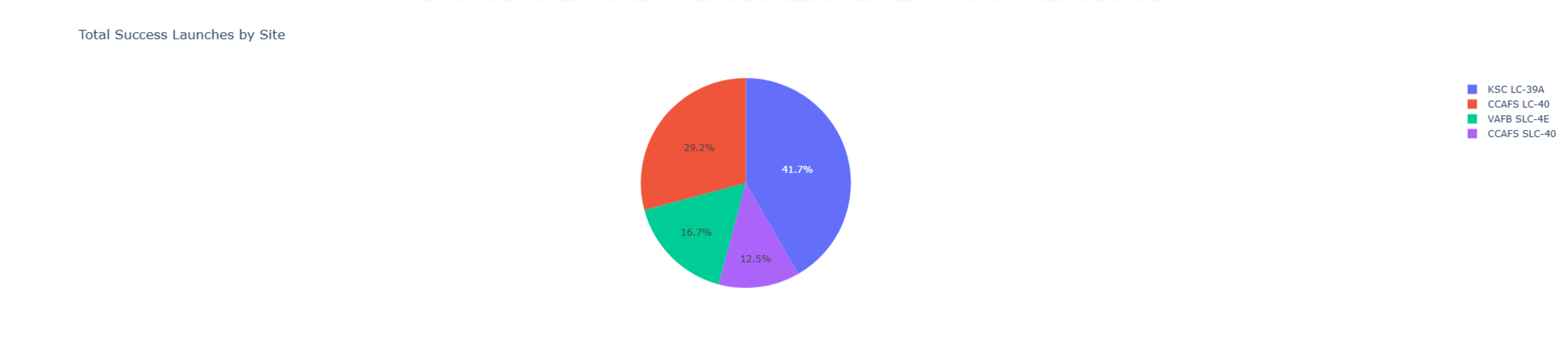
Section 4

Build a Dashboard with Plotly Dash

<This is original work>

<This is original work>

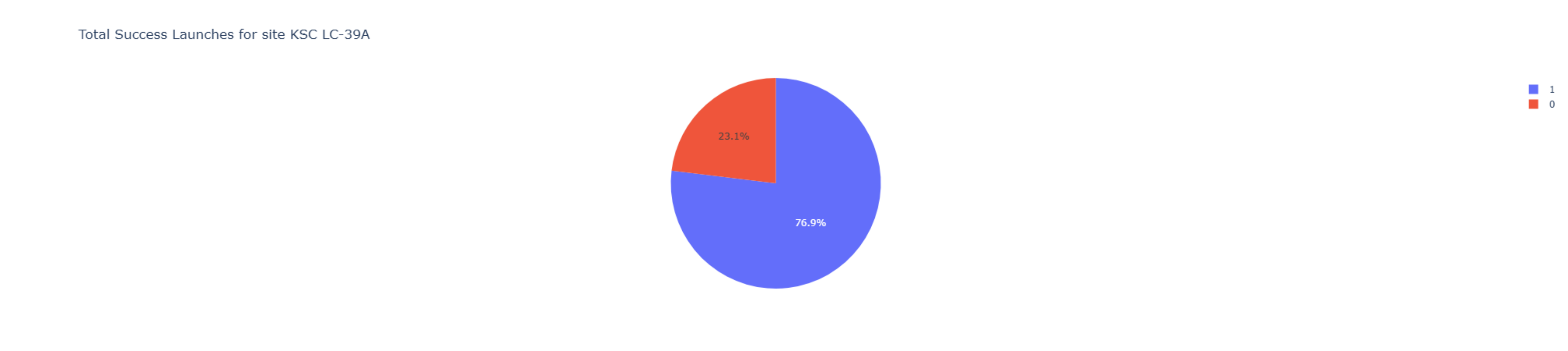
Total Success Launch Rate by Site



- This plot shows the distribution of successful launches by the four various launch sites
- The KSC LC-39A site had the largest share of successful launches

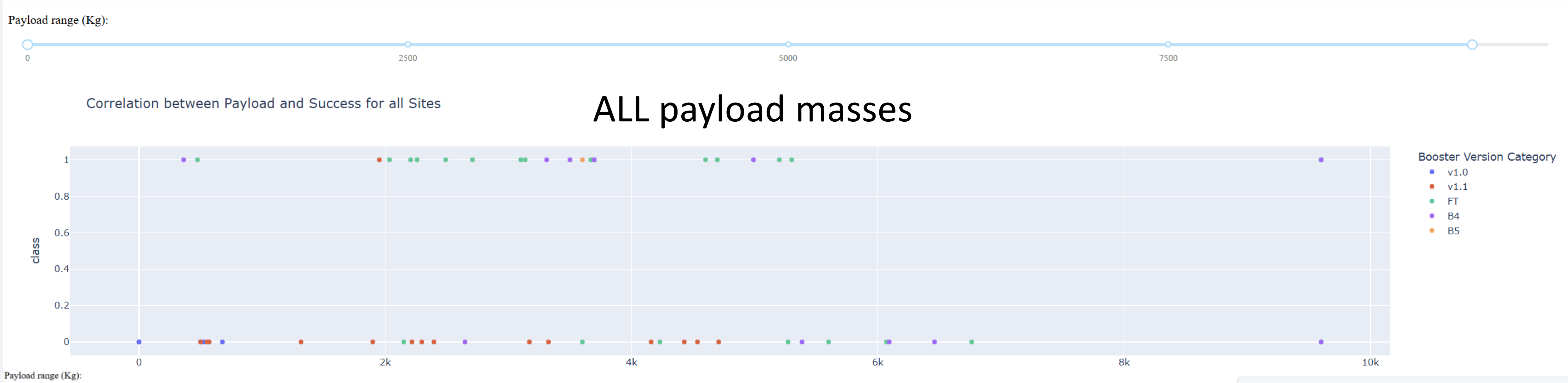
<This is original work>

KSC LC-39A Launch Site Success Rate

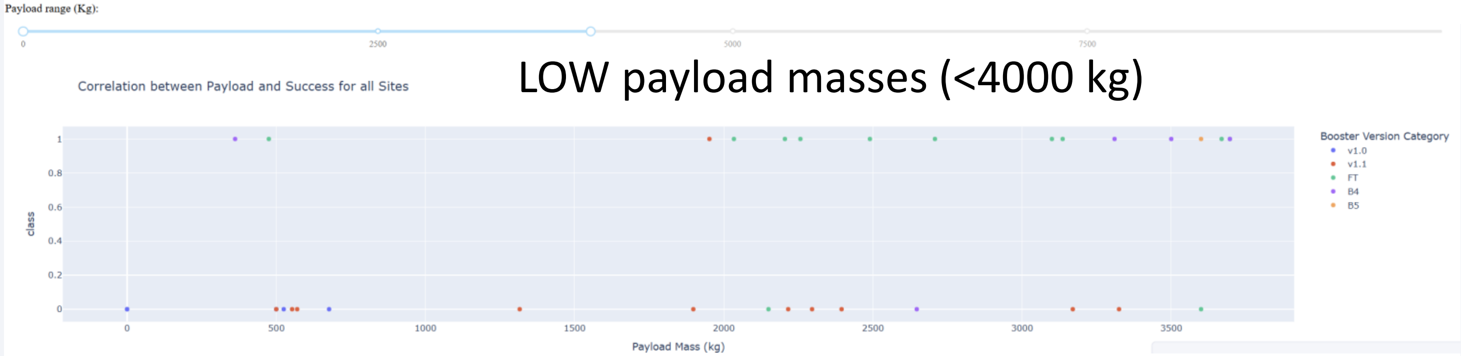


- The KSC LC-39A had the highest launch site success rate, with 76.9% of all launches deemed successful

Launch Success with Various Booster Versions



At all payload masses, the FT booster had the highest success rate.



At payload masses < 4000 kg, the FT and B4 booster had higher success rates and the v1.1 booster had lower success rates, especially at low payload masses.

At payload masses > 4000 kg, booster version and success rate are not strongly correlated.

Section 5

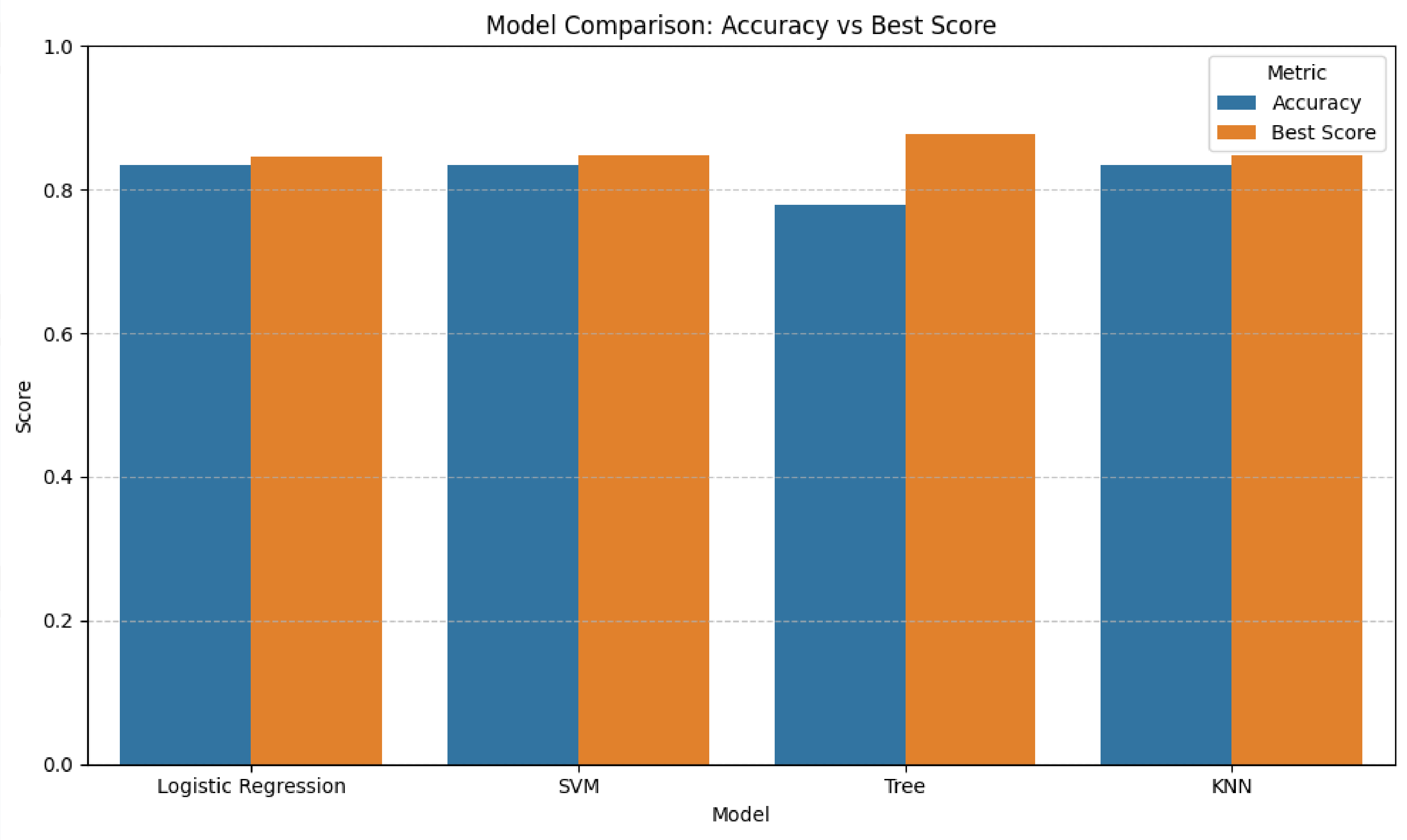
Predictive Analysis (Classification)

<This is original work>

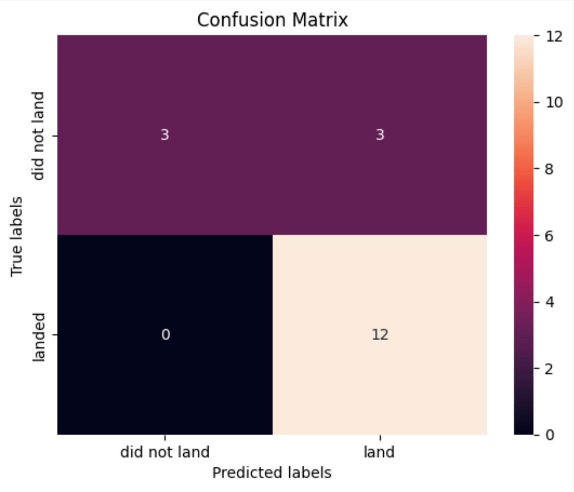
<This is original work>

Classification Accuracy

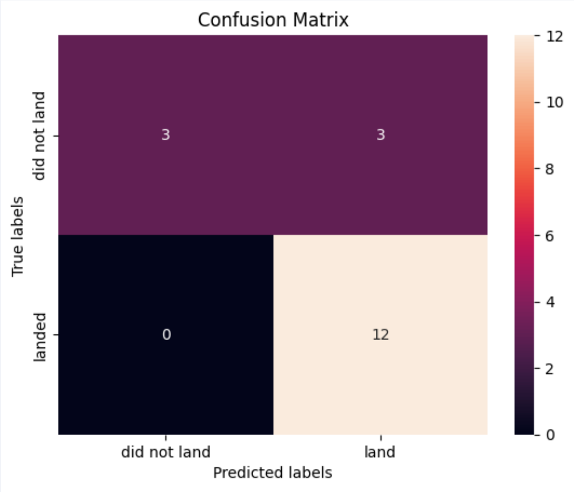
- The Logistic Regression, SVM, and kNN models have the best accuracy based on the testing dataset to predict landing outcomes
- The kNN model had the best score during training



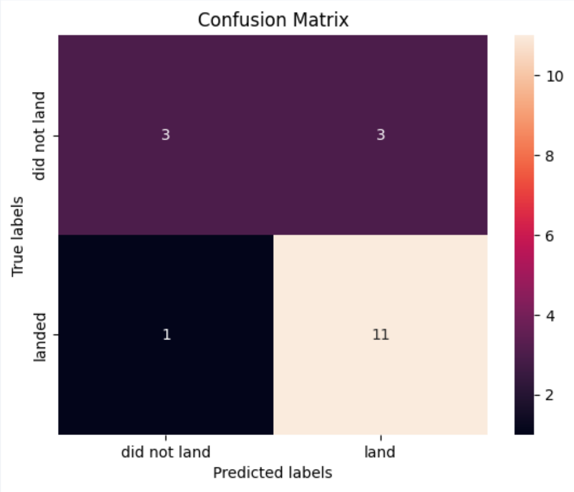
Confusion Matrix



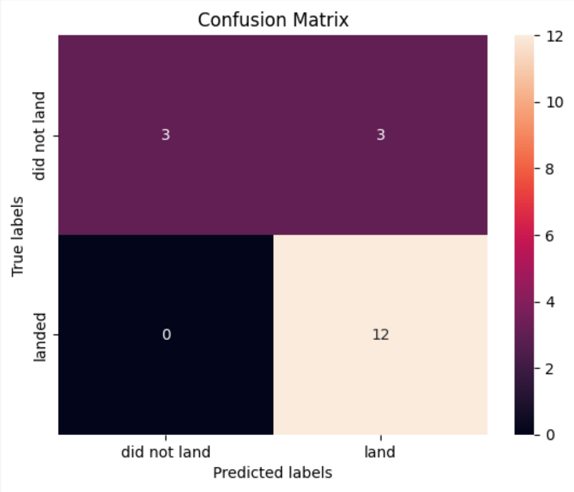
Logistic regression



SVM



tree



kNN

- The confusion matrix shows the number of true positives and false positives
- In the logistic regression, SVM, and kNN models, 12/18 tested data were identified as true positives
- A larger testing dataset could better distinguish the performance of the classification models

Conclusions

- From visualization, success outcomes increased as more flights were deployed
- Payload mass and orbit type showed strong correlation with success outcomes
- Launch Site proximity showed the importance of proximity to coastline and railroad and distance from populated areas for rocket launches
- The KNN classification tree model can predict launch outcome with the highest score and testing accuracy of 83.33%
- A larger training and testing dataset could better distinguish the performance of the training models

Thank you!

<This is original work>
<This is original work>